RESEARCH STATEMENT

Meenakshi Khosla

Last updated: Jan'22

Introduction

Deep neural networks have emerged as powerful models of computations underlying sensory processing. While there has been remarkable recent success in predictive modeling of neural responses under naturalistic stimulation, driven largely by the architectural class of convolutional neural networks trained on behaviorally relevant tasks, there is still a large portion of these brain signals that remain unaccounted for by existing models and we need advances in computational modeling to better explain brain activity. Further, current state-of-the-art models of the visual cortex also remain biologically implausible, particularly because they learn in a manner highly inconsistent with our developmental experience and lack critical circuit motifs believed to play a key role in biological networks. As the first goal of our research (**Aim 1**), we ask the following question: how can we further improve DNNs as models of the human visual cortex and human behavior? We aim to investigate this through the lens of learning rules, visual diet and architectural constraints with the goal of developing DNN models that learn in a manner more aligned with development and that contain architectural motifs reminiscent of the brain. The former requirement also inherently constrains the data used to train any representational model with the type (e.g., unlabelled data, naturalistic input statistics) and quantity of inputs that infants might receive during development. **Aim 2** seeks to address statistical challenges associated with comparing representations in the brain and in neural models. **Aims 3 & 4** address the aspect of interpretability in deep neural network models of human vision and the development of validation techniques for the utility of studying biological systems using computational models. With creative model interpretability methods applied to deep neural network models of human vision, we hope to realize their full potential in characterizing the precise function of different brain areas, their input-output relationships when exposed to arbitrary stimuli and the underlying computational mechanisms. The goal is to move beyond the tradition of experimental reductionist approaches in cognitive neuroscience to using a combination of computational models, particularly deep learning models, and ethologically relevant naturalistic stimuli in order to understand the neural encoding of sensory information. Finally (**Aim 5**), we hope to tackle a fundamental question at the intersection of computational neuroscience and Artificial Intelligence (AI), i.e., how can we leverage neuroscientific insights or neural data to inform and inspire the next generation of models in AI?

1 Aim 1: Systematic comparison of learning rules, visual diet and architectures

The primary goal is to perform large-scale comparisons of deep learning models against neural datasets by looking through the lenses of learning rules, visual diet and architectures as described in detail in the following subsections.

1.1 Learning rules

Understanding the general principles of brain function necessitates understanding the goals of the system (as inspiringly formulated by David Marr) and a fundamental question still remains unanswered: what computational objectives give rise to cortical representations? In other words, why do specific patterns of brain activity occur reliably across trials and human participants in response to sensory stimuli in the first place? The striking similarity between representations in DNNs trained on visual recognition tasks and the representations in the primate ventral visual stream has led to the speculation that the primate ventral visual stream may have evolved to be an optimal system for object recognition in the same way that DNNs are identifying optimal computational systems for specific tasks [Khaligh-Razavi and Kriegeskorte, 2014, Yamins et al., 2014, Yamins and DiCarlo, 2016]. This raises an intriguing possibility for further investigation that may now be tractable: what objective functions or compu-

tational goals may explain neural representations in other parts of the brain, i.e., what is the task of the brain in other parts of the cortex? Different parts of the brain might be optimized for different objective functions. Several candidate goals or objective functions have been proposed before, including disentanglement of latent, interpretable factors in the case of vision [Higgins et al., 2020], predictive processing in the case of both language and vision [Schrimpf et al., 2020a, Bakhtiari et al., 2021], uncertainty estimation or redundancy reduction (efficient coding). A large-scale comparison of these and other novel hypothesis about the goals of biological systems will likely not only elevate our understanding of the brain but it could also provide a rich source of inspiration for the field of AI.

Furthermore, even supervised learning rules like object categorization, which previously offered the normative perspective that representations in DNNs and the primate ventral visual stream may be aligned with common computational goals, are biologically implausible, as they require large amounts of high-level, semantic labels during learning. More recent forms of self-supervised learning rules, like contrastive objective functions, have emerged as promising alternative candidate learning rules because they do not require category labels to train and are thus at least more biologically plausible than supervised learning objectives. But all contrastive learning methods implicitly learn a set representational invariances depending on the data augmentations they exploit during training. Importantly, since all contrastive learning methods employ data augmentations to bake in the desired invariances into the model, an important question arises: what makes for good views for contrastive learning? Contrastive learning is based on implicit knowledge of downstream task invariances. Can we compare constrastive learning objectives with different invariances to better understand the invariances in different parts of the visual cortex? Following a similar line of investigation, we can also systematically compare modern contrastive self-supervised representation learning methods like SimCLR, Instance-Discrimination, Momentum Contrast (MoCo) and Contrastive Predictive Coding (CPC) against non-contrastive unsupervised learning rules that are based on pretext tasks like rotation, colorization or context prediction. The latter aim to recover transformations between different views (i.e., augmentations) of the same data point, whereas the former approach is based on learning to be invariant to those transformations while remaining discriminative with respect to other data samples. Testing these two class of objectives against recordings from the human brain can sheld light on what kinds of unsupervised learning objectives might be operative in the brain.

Even among supervised objective functions, the precise task can vary in the degree of ecological relevance. For instance, one aspect of the visual categorization task that is straightforward to explore is task granularity. Do fine-grained recognition tasks lead to better representations by forcing the network to 'work harder' and thereby avoid reliance on spurious or trivial correlations when learning input-output relationships? Or do they lack 'behavioral relevance' and thus lead to a worse match to neural representations? One could also focus on learning trajectories and ask when in the course of learning a task, do representations emerge that match those of the brain? For example, is it when category structure is too explicit with fine grained categories well separated? Or does the learning proceed by first separating course categories, which are then tuned to more fine-grained distinctions and it is the intermediate stage with coarse categorical separation that is good enough (or perhaps even better) to explain neural representational geometry?

A thorough, systematic comparison of different supervised and unsupervised objective functions thus seems like a promising future direction to understand the high-level principles governing neural computations.

1.1.1 Discovering human understandable properties of CNNs that explain variance in neural predictivity

Previously, foundational work on the goal-driven framework for systems neuroscience revealed a strong correlation between object recognition performance and neural predictivity, indicating that task-optimization is a useful means to achieve brain-like models. A natural question thus arises: might we discover other human understandable properties in DNNs that correlate well with neural predictivity, providing novel computational theories of the mind? What other, possibly orthogonal, interpretable functional property of DNNs might explain the variance in their neural predictivity? One approach is to control different abstract, interpretable properties in DNNs using regularization or additional loss functions and study the causal relationship between the property of interest and the resulting similarity of intermediate DNN activations to neural representations. For instance, recent work explored the relationship between class selectivity (the interpretable property of interest) and visual categorization behavior by directly regularizing for or against selectivity [Leavitt and Morcos, 2020]. Another approach is based on randomly drawing several DNN models from the vast parameter space of DNNs and correlating the property of interest in these models against their respective neural predictivity. Some example properties that can be explored in this prospective study include, for instance, one-shot or few-shot learning performance of representational models, their degree of adversarial robustness or extent of shape-texture bias, as well as certain measures of disentanglement.

1.2 Visual diet

Here, we wish to investigate DNN models of human vision from the perspective of their training environment and we ask how does the training regimen affect the emergent representations? Ideally, we desire that these representational models be trained using a visual diet mimicking the inputs model organisms might receive over the course of development. The ImageNet dataset, upon which most SOTA vision models are built, is clearly not reflective of early visual experience as it contains stereotypical images of a thousand object categories, many of which are not relevant to humans while missing entire categories that represent behaviorally crucial everyday entities (e.g., the person category). A recent study already demonstrated that training datasets biased towards more ecologically valid input statistics, may already lead to the emergence of more brain-like representations in DNN models [Mehrer et al., 2021]. Building upon this work, we aim to systematically study the influence of the pre-training data domain on emergent DNN representations.

One can also ask the following question: how does selective deprivation in the visual diet influence the resulting representations? These selective deficiencies in the visual diet can also provide important clues into how response selectivity and other properties of the neural responses might arise in the first place. For instance, we can ask whether training a DNN on non-face-specific object recognition tasks by depriving the visual diet of faces altogether, can replicate face-selectivity and other response properties of the FFA. This could shed light on the role of visual experience with the preferred category for the development of category-selectivity in the brain.

Learning from static images may have fundamental limitations, and it may be the case that the only way to develop better models of human vision is to learn from visual dynamics. Several different pretraining datasets can be explored in this regard; one interesting dataset is the large SAYcam dataset that comprises naturalistic videos recorded from an egocentric perspective, intended to mimic an infant's experience during development [Sullivan et al., 2020]. Capturing structure within dynamic visual scenes might also benefit from novel objective functions and architectures, again indicating a strong interplay among these different axes of exploration.

1.3 Architecture

It is important to acknowledge the crucial ways in which the current models diverge from biological networks. State-of-the-art visual categorization networks are almost all dominantly feed-forward and fail to capture known properties of biological networks, such as local recurrence; however, they have been found to be useful for modelling neural activity across different sensory systems. Although the functional significance of intra-regional recurrent circuits in core object recognition is still a matter of vigorous debate, mounting evidence suggests that they may be subserving recognition under challenging conditions [Kar et al., 2019]. As future work, I would like to delve into the investigation of more neurobiologically plausible models of the cortex that innately model intra-regional recurrent computations, especially in relation to their role in visual recognition and neural response prediction. Interestingly, previous work has also shown that introducing anatomical constraints in DNN models of the visual system in the form of biological resource limitations as imposed by the physical bottleneck of the optic nerve, leads to the emergence of network properties that can simultaneously explain retinal and early visual cortical representations in a single task-optimized architecture [Lindsey et al., 2019]. More recently, a dual-pathway network (with *built-in* segregation) trained with self-supervised objective functions was shown to simultaneously predict both dorsal and ventral visual stream responses, and used to explain why functional specialization of the 'what' and 'where' pathways might be seen in the brain [Bakhtiari et al., 2021]. In the future, designing more neurobiologically plausible architectures using anatomical or functional constraints can be a satisfying and fruitful research endeavor. Powerful architectural priors can also serve as training data 'in disguise' and reduce the need for large amounts of training data to achieve a sufficient level of performance. This also suggests that perhaps the different axes of architecture, visual diet and learning rules may be best studied in tandem, rather than in isolation.

We can also compare different hypotheses about what role these different architectural motifs might play in the brain by evaluating the performance of models with different architectural inductive biases under diverse train-test scenarios. Or perhaps more intriguingly, specific architectural motifs might appear spontaneously in trained representational models without being implanted as built-in structures into DNNs, and their similarity with biological architectures (if it exists) might provide computational accounts of the functional architecture of the brain.

Connections with specialized cognitive processes

Over the course of the last decade, computer vision and AI researchers have realized the importance of incorporating modules that mimic cognitive and perceptual processes, such as working memory and attention, from both task-optimization and neural encoding perspectives. For instance, previous research with animal models and human fMRI shows that models that can efficiently store and access information over longer spans, such as Recurrent Neural Net-

works (RNNs) with sophisticated gating mechanisms, are much more suitable for modeling neural computations that unfold over time (as in stimulation with natural videos) in comparison to non-recurrent approaches [Sinz et al., 2018]. Since activations of units within RNNs depend not only on the incoming stimulus, but also on the "current" state of the network as influenced by past stimuli, they are capable of holding short-term events into memory. Adding the RNN module can thus be viewed as augmenting the encoding models with working memory. Further, in another study on attentional modulation, the phenomenon of *attention* was modelled as a novel module that selects certain portions of visual stimuli, the so-called attention "spotlights", for subsequent processing at the expense of others and this was shown to substantially improve neural response prediction [Khosla et al., 2020]. Further, in dynamic visual perception, the influence of feedback or top down signals in modulating early visual areas may be even more prominent and the bidirectional flow of information (bottom-up and top-down) as facilitated by the feedback connections between distinct anatomical areas in the brain may have important functional consequences. A complete biologically plausible computational model of attention would capture both bottom-up and top-down influences of working memory and context as they may ultimately constrain which spatial locations or events are selected for further processing. In the future, it would be interesting to draw upon the visual attention and saliency literature to develop models that mimic biological attentional mechanisms more closely and use it to dynamically combine bottom-up and top-down representations.

2 Aim 2: Comparing representations in deep neural networks and the brain with information-theoretic and sample complexity measures

An important application of encoding models is in comparing emergent representations in different layers of DNNs trained across different architectures, tasks or learning rules against representations in biological systems. One straightforward metric employed in these encoding-based studies is neural predictivity, which measures the agreement between predicted and measured responses, where the predicted response of every neuron or voxel for a particular image is expressed as a linear function of the corresponding DNN representation. However, there are several limitations of simple neural predictivity as an evaluation metric for different representational models, perhaps the most important one being that it completely disregards the complexity or statistical efficiency of learning the linear predictor on top of a representational model. Instead of looking at the neural predictivity alone which is the de facto approach, there is perhaps merit in also looking at the loss-data landscape, i.e., the trend of how much stimulus-response data we need to achieve a certain level of prediction performance. Sample complexity or learning speed of the linear predictor may be an indication that the inductive bias of the representational model was strong in the sense that it constrained the space of possible solutions in the "right manner" so that the error dropped significantly only after encountering a few samples. 'Best' representation would be the one that allows the most efficient learning of the downstream linear predictor. Importantly, restricting the size of the stimulus-response set when evaluating encoding models might also magnify the differences between different representational models or the difference between models pre and post training, highlighting the true benefit of 'learning'.

Neural predictivity is also confounded by the complexity and dimensionality of the representation, i.e., a higher dimensional representation also has added degrees of freedom. Information theoretic concepts like codelength can be useful in this aspect, because they weigh not only the final quality of predictions, but also the amount of effort required to achieve it. Thus, it might be beneficial to replace asymptotic neural predictivity as an evaluation metric for encoding models with an estimate of the codelength, which is a combination of the quality of the fit of the linear predictor as well as the cost of transmitting the model (linear predictor) itself. A systematic study comparing these different evaluation metrics, and proposing comprehensive metrics based on insights developed from the former, can thus be very useful to the study of DNNs as models of the brain.

3 Aim 3: Explainable deep learning models of the visual cortex

Broadly, the aim here is to apply model interpretation techniques to understand how and why deep neural network models might capture neural phenomena. Modern deep neural networks have been highly successful in capturing structure within large piles of data in rich and fruitful ways. Here, we wish to apply model compression techniques to distill the knowledge or the structure learned by these large-scale models of the visual cortex into smaller, tractable models that are more amenable to interpretation. There are several advantages to probing a computational model of the brain, rather than probing the brain using neural measurements directly, the most important being that the former allows complete access the network weights ('the connectome') and allows us access to the predicted neural responses to arbitrarily large probe datasets well beyond what realistic experimental procedures might allow due to time, budget or other constraints.

Here, we wish to go beyond simply maximizing the variance explained by our models and looking deeper into their hidden units and their tuning properties to gain conceptual insights into the mechanisms employed by these models to predict brain responses by adopting a 'predict, then simplify' approach. For this goal, model reduction and subsequent interpretation using *probe* datasets can be very powerful. A candidate region to apply such an approach to, could be the area V4 in the primate brain, whose response properties are still a matter of active debate and whose function has been variedly described in the literature so far, such as whether it detects shapes, textures, or both. Training deep CNN models to reproduce brain activity in this region can help us assess whether these models trivially learnt the input-to-output transformations for the regions or whether they learned plausible *mechanisms* as well, for e.g., do internal representations of these networks map onto sub-parts preceding V4 in the visual hierarchy (V1 or V2) and compose them to form more complex representations? We could also gain a circuit-based understanding of these networks using attribution [Cammarata et al., 2021] or network pruning and distillation techniques. The latter was recently shown to work very well in explaining retinal responses to a wide range of artificial stimuli [Maheswaranathan et al., 2018]. Another interesting hypothesis to test here is whether distilled models for a visual area all exhibit certain shared, universal properties. As an example, we could observe maximally activating images for different architectures and their distilled (pruned) versions and see if distillation of different models can reduce their individual differences. Previously, it was shown that the initial state of weights in a model can significantly impact the emergent representations in the DNN instance and subsequently, its ability to predict cortical representations [Mehrer et al., 2020]. Just as network regularization increased the consistency between models with different initializations in that scenario, distillation could potential have the same effect of increasing consistency among models. This could help answer questions like the following: do distillation procedures applied to these models of the cortex yield similar computational mechanisms across different initializations ? Are the reduced models more stable across initialization and architectures in comparison to messy, complex state-of-the-art models with over million parameters which are highly sensitive to initialization? We could also attempt to understand divergences between a complex model and a simpler model by synthesizing stimuli that elicit divergent responses following the approach of controversial stimuli [Golan et al., 2020] to assess the reducibility or distillation of encoding models. This could either validate the simpler model or simply provide justification for added model complexity.

We could also apply network dissection techniques to understand the properties of simulated or predicted population responses in the 'no-man's land of the visual cortex', across thousands of images from rich naturalistic, probe datasets [Bau et al., 2017]. This is the part of the visual cortex not identified by any functional localizer and whose precise functional characterization remains elusive. Even though single neurons or single voxels don't exhibit high selectivity for object categories in this area of the brain, we can ask whether populations of neurons or voxels in these regions encode and represent human-understandable concepts using novel network interpretability techniques [Fong and Vedaldi, 2018].

Limitations and challenges

Probing computational models of the brain imposes several challenges that may lead to confounded conclusions and hypotheses about neural computations. The most important confound is that model predictions can deputize for experimental data only to the extent that they are 'accurate' and ultimately, all interpretative analyses and conclusions rest on the predictive accuracy of the model. It is further easy for models to learn trivial input-output dependencies without remaining faithful to the the mechanism. Despite these limitations, computational models can nonetheless provide novel *testable* hypotheses, which can be accepted or refuted with future experimentation. Even if the hypothesis is refuted, model failures can further drive model development and lead to the generation of improved hypothesis. In this manner, a tight loop between modeling and experiments ,beyond simple offline analyses of neural data, can expedite neuroscientific discovery.

4 Aim 4: Closed-loop experiments: Probing neural response selectivity and invariance

I am also interested in the design of closed loop experiments that exploit computational models of neural responses and focus on evolving stimuli in real time to analyse the preferred stimuli for individual neurons or voxels, the invariances near the most activating stimuli, shapes of the discovered tuning curves, and other response properties of interest. These closed loop experiments offer an interesting proof of concept for the utility of studying biological systems using computational models; their application is not simply limited to verification of the models but also beyond that to test specific hypotheses about representations in different regions, as well as in stimulating different brain regions beyond their typically observed levels of activation under naturalistic stimulation. Neuroscience is constrained by the the amount of stimuli that can be can be used to record brain responses from subjects due to the time consuming and expensive data acquisition process. Understanding response characteristics with a limited stimulus-response set alone can be both challenging and misleading. For instance, characterizing the invariances across the visual hierarchy

necessitates the search through a vast stimulus set with all the relevant transformations. Computational models coupled with high throughput stimulus optimization techniques can circumvent this issue and provide a means to efficiently probe selectivity and invariance across the sensory cortex.

5 Aim 5: Towards artifcial neural networks with brain-like inductive biases

Thus far, the aims of our future research have been focused on the use of artificial neural networks in modelling neural responses. One can also turn the question around and ask how neural data or neuroscientific insights can help inform and inspire the next generation of models in Artificial Intelligence (AI). Recently, there has been a growing interest in this direction and some recent studies have already laid the seed for cross-fertilization between machine learning and neuroscience. For instance, an integrative benchmark was recently proposed in [Schrimpf et al., 2020b] wherein they developed a comprehensive 'Brain-score' to measure the similarity of DNNs to brain's core object recognition using neural and behavioral measures. These benchmarks aim to guide model development in AI by adopting the reverse-engineering approach. Other studies have shown that explicitly encouraging neural networks to build representations that are similar to neural representations or principles can improve their robustness under stimulus noise and adversarial attacks ([Li et al., 2019, Dapello et al., 2020]). Attempts at aligning the activity of neural networks with brain-like representations are based on the hypothesis that networks favoring brain-like representations will demonstrate brain-like capabilities, for example, *stronger generalization* with limited supervision, *robustness* to stimulus noise, *lower sample complexity* in few-shot learning settings, *continual learning* without catastrophic forgetting etc. Further, beyond model development, we can get inspiration from neuroscience not only in the form of theories about how the brain functions but also in terms of toolkits and formalizations developed and adopted by neuroscience researchers to understand complex biological networks. This can, for instance, lead to novel methods for understanding single-unit and population responses of model neurons and lead to a better understanding of representations and computations in DNNs. I am interested in delving into this exciting topic with human fMRI and behavioral data in my research to study if increasing the brain-bias in computational models can indeed lend them brain-like capabilities, and in adapting existing formalization employed in neuroscience research to develop new methods for understanding DNNs. A natural approach to tackle the former goal is to constrain or regularize neural network models trained on computer vision tasks by large-scale neural activity data through additional loss functions that attempt to increase the similarity of model representations to human brain activity patterns. While capturing information processing mechanisms of the brain may not be a 'necessary' condition for designing systems that exhibit intelligent behavior, the brain is certainly a 'sufficient' general purpose intelligence system and currently the only proof that such a system can exist; therefore, understanding the brain can potentially guide principles on the basis of which machines with human like capabilities may be built, laying the foundation for bridging the gap between human and artificial intelligence.

References

[Bakhtiari et al., 2021] Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C., and Richards, B. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning.

[Bau et al., 2017] Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549.

[Cammarata et al., 2021] Cammarata, N., Goh, G., Carter, S., Voss, C., Schubert, L., and Olah, C. (2021). Curve circuits. *Distill*, 6(1):e00024–006.

[Dapello et al., 2020] Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D. D., and DiCarlo, J. J. (2020). Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *bioRxiv*.

[Fong and Vedaldi, 2018] Fong, R. and Vedaldi, A. (2018). Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8730–8738.

[Golan et al., 2020] Golan, T., Raju, P. C., and Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117(47):29330–29337.

[Higgins et al., 2020] Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., and Botvinick, M. (2020). Unsupervised deep learning identifies semantic disentanglement in single inferotemporal neurons. *arXiv preprint arXiv:2006.14304*.

[Kar et al., 2019] Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., and DiCarlo, J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22:974–983.

[Khaligh-Razavi and Kriegeskorte, 2014] Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915.

[Khosla et al., 2020] Khosla, M., Ngo, G., Jamison, K., Kuceyeski, A., and Sabuncu, M. (2020). Neural encoding with visual attention. *Advances in Neural Information Processing Systems*, 33.

[Leavitt and Morcos, 2020] Leavitt, M. L. and Morcos, A. (2020). Selectivity considered harmful: evaluating the causal impact of class selectivity in dnns. *arXiv preprint arXiv:2003.01262*.

[Li et al., 2019] Li, Z., Brendel, W., Walker, E. Y., Cobos, E., Muhammad, T., Reimer, J., Bethge, M., Sinz, F. H., Pitkow, X., and Tolias, A. S. (2019). Learning from brains how to regularize machines. In *NeurIPS*.

[Lindsey et al., 2019] Lindsey, J., Ocko, S. A., Ganguli, S., and Deny, S. (2019). A unified theory of early visual representations from retina to cortex through anatomically constrained deep cnns. *arXiv preprint arXiv:1901.00945*.

[Maheswaranathan et al., 2018] Maheswaranathan, N., McIntosh, L., Kastner, D. B., Melander, J., Brezovec, L., Nayebi, A., Wang, J., Ganguli, S., and Baccus, S. A. (2018). Deep learning models reveal internal structure and diverse computations in the retina under natural scenes. *BioRxiv*, page 340943.

[Mehrer et al., 2021] Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., and Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8).

[Mehrer et al., 2020] Mehrer, J., Spoerer, C. J., Kriegeskorte, N., and Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature communications*, 11(1):1–12.

[Schrimpf et al., 2020a] Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J., and Fedorenko, E. (2020a). Artificial neural networks accurately predict language processing in the brain. *BioRxiv*.

[Schrimpf et al., 2020b] Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., and DiCarlo, J. J. (2020b). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*.

[Sinz et al., 2018] Sinz, F. H., Ecker, A. S., Fahey, P. G., Walker, E. Y., Cobos, E., Froudarakis, E., Yatsenko, D., Pitkow, X., Reimer, J., and Tolias, A. (2018). Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. *bioRxiv*.

[Sullivan et al., 2020] Sullivan, J., Mei, M., Perfors, A., Wojcik, E. H., and Frank, M. C. (2020). Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective.

[Yamins and DiCarlo, 2016] Yamins, D. L. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365.

[Yamins et al., 2014] Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624.