Position: Maximizing Neural Regression Scores May Not Identify Good Models of the Brain

Rylan Schaeffer
Computer Science
Stanford University
rschaef@cs.stanford.edu

Mikail Khona
Physics
MIT
mikail@mit.edu

Sarthak Chandra
Brain & Cognitive Sciences
MIT
sarthakc@mit.edu

Mitchell Ostrow
Brain & Cognitive Sciences
MIT
ostrow@mit.edu

Brando Miranda
Computer Science
Stanford University
brando9@cs.stanford.edu

Sanmi Koyejo
Computer Science
Stanford University
sanmi@cs.stanford.edu

Abstract

A prominent methodology in computational neuroscience posits that the brain can be understood by identifying which artificial neural network models most accurately predict biological neural activations, measured according to regression test error or other similar metrics. In this opinion piece, we argue that the field lacks a canonical definition of model goodness, and rather than engaging with this difficult question, the neural regressions methodology simply asserted a proxy – neural predictivity – then overfit to this proxy. We begin with a notable failure of the neural regressions methodology in which the most predictive models disagreed with key properties of the neural circuit. Next, we highlight converging empirical and mathematical evidence that explains the disconnect: (linear) neural regressions are simply discovering the implicit biases of (linear) regression, which may not appropriately identify models that are actually brain-like. This is an instance of Goodhart's law: by selecting neural network models that optimize (linear) neural predictivity, the field's results have devolved into re-discovering general properties of (linear) regression, rather than furthering our understanding of the brain. These insights suggest that the neural regressions methodology may be insufficient for understanding the brain, and we call for a critical reevaluation of this methodology in computational neuroscience.

1 Introduction

An influential methodology in neuroscience-inspired artificial intelligence argues that task-optimized deep artificial neural networks (ANNs) should be considered good models of the brain if they capture a large fraction of variance in neural population recordings assessed via regressions of ANN unit activity onto biological neural responses [89]. The claim is that the ANN(s) with better performing neural regressions are more similar to the brain than alternative models [71]. This approach has been widely used in vision [90, 24, 43, 50, 72, 92, 40, 88, 79, 64, 18, 45], audition [46, 82, 55, 80], language [62, 39, 73, 3, 61, 15, 16, 33, 5, 2, 38, 58, 17, 44, 4, 81, 56, 37], and spatial navigation [57], most often with (regularized) linear models, but occasionally with non-linear models.

In this position piece, we argue that Neuro-AI lacks sufficiently rich definitions of neural similarity, and such notions are context-dependent and difficult to quantify. The neural regressions methodology sidesteps these challenges by defining a proxy – for instance, the test R^2 of linear regression between biological recordings and model activations – and then choosing models based on this proxy. The

Published at Unifying Representations in Neural Models (UniReps) Workshop at NeurIPS 2024.

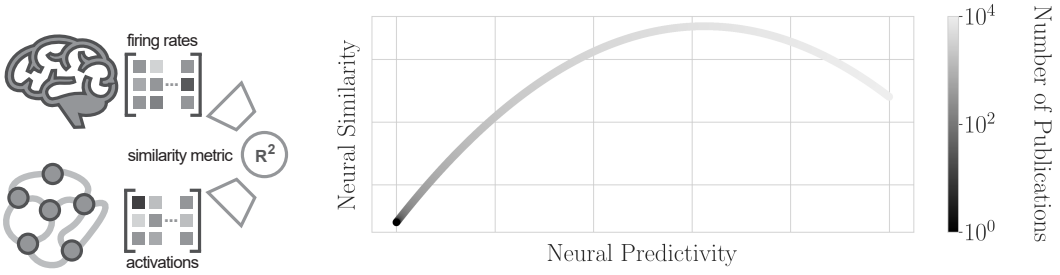

Figure 1: **Schematic.** Left: The neural regressions methodology posits a proxy – neural predictivity – of how similar model(s) are to a neural system of interest without verifying the extent to which the proxy agrees with neural similarity. Right: Overfitting to the proxy leads to mismatches with neural similarity. Although we do not define neural similarity here, we emphasize that it is task, neural-system, and question-dependent, and hence likely cannot always be neural predictivity. For a system-specific example of neural similarity, we offer criteria for grid cells in Appendix Sec. A.

models that win a selection process (e.g., on BrainScore [71]) may do so more because of inductive biases of the proxy, independent of any meaningful relationships with the brain (Fig. 1).

This perspective explains why, for example, the neural regressions methodology was confidently incorrect when applied to models of grid cells: linear regression does not capture in key criteria of neural similarity for grid cells (e.g., periodic tuning curves [34], multiple grid modules with specific period ratios [77], toroidal continuous attractor dynamics [91, 32]; see Appendix Sec. A for a detailed list). This perspective also explains a finding by four independent research groups in different modalities, data, architectures and recording technologies [66, 25, 80, 17] of a quantitatively consistent relationship between test R^2 and effective dimensionality, that was mathematically corrected and further empirically studied by Canatar et al. [14]: (linear) neural predictivity *is* (linear) regression, and (linear) regression has inductive biases, irrespective of the underlying neuroscience. We focus on linear regression because of its ubiquity in the literature, but other preference functions (e.g., RSA [49], CKA [48], SVCCA [63], Procrustes [84], etc.) would not escape this critique; rather, they would simply change the inductive biases of the chosen preference function.

Together, these insights suggest that the neural regression methodology, and more broadly the idea that a uniform set of metrics can automate model selection, may be fundamentally flawed by overfitting to those metrics rather than advancing our understanding of the brain. We conclude by suggesting a re-evaluation of such methodologies.

2 Neural Regressions Can Reach Incorrect Conclusions with High Confidence

In vision, Bowers et al. [9] documented how artificial networks preferred by the neural regressions methodology lack or contradict properties of primate vision, and others have identified additional flaws [54, 88, 20, 35, 26, 27, 23, 36, 52]. Here, we chose to focus on the clearest example of a failure of the neural regressions methodology: grid cells. Why focus on grid cells? Grid cells – a surprising and important Nobel Prize-winning discovery [34] – differ from vision, audition and language in that humanity possesses scientific models [29, 11, 10, 76] that have repeatedly proven predictive ([77, 91, 32]), not in the regressions sense but in the sense of exhibiting fundamental properties, e.g., localization of each module to a two-dimensional subspace, quantization of grid module periods, preserved low-dimensional dynamics across waking and sleep that were subsequently validated. In a domain we understand well, how did the regressions methodology fare?

When applied to a specific neural circuit (grid cells) that humanity possesses near-normative models of, the neural regressions methodology preferred incorrect models with high confidence.

As context, the key research questions about grid cells are modeling their dynamics and the evolutionary causes for their existence. Previous and now near-normative models showed how strong recurrent interactions leading to pattern formation, coupled with a way for movement inputs to shift the pattern phase and thus perform path integration, could generate grid cell dynamics [11, 47]; and

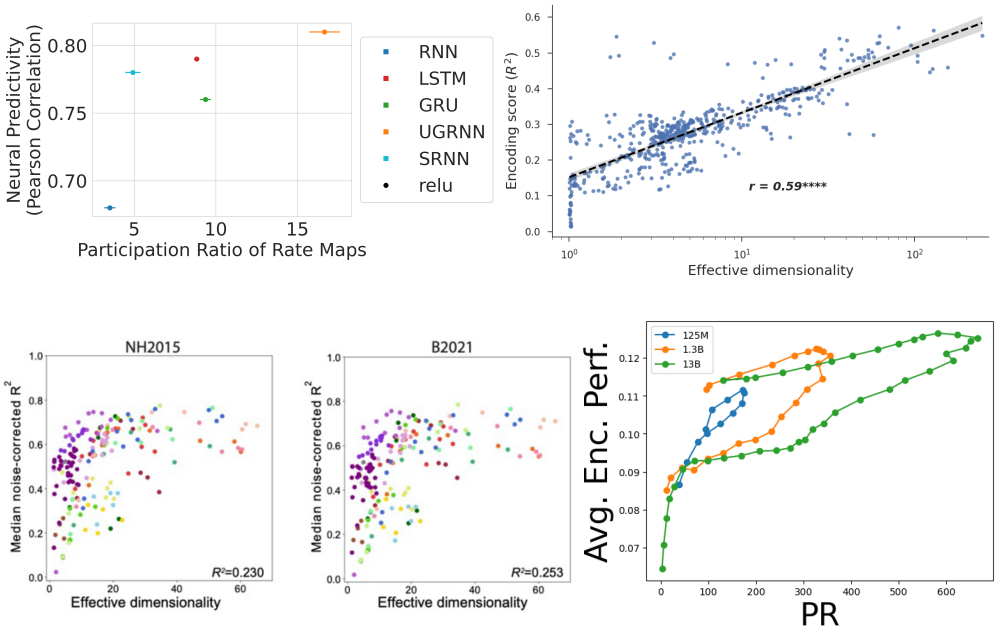

Figure 2: Four independent publications studying four different modalities and brain circuits in three different species found a consistent quantitative heuristic: Test R^2 is an affine transformation of the log participation ratio (Eqn. 2). Figures from Spatial Navigation in Mouse Medial Entorhinal Cortex [66], Vision in Macaque IT Cortex [25], Audition in Human Cortex. [80], Language in Human Cortex [17]. Later work [14] provided a spectral theory of the neural regressions methodology, which reveals results like these are attributable to *general properties of linear regression*, not the brain.

that multiple grid modules played key roles in disambiguating position over large ranges and in error correction [29, 76]. Later, deep recurrent networks trained in a supervised manner to path integrate were shown to learn grid-like units [7, 19, 74], and neural-regressions based work [57] showed that these supervised deep path integrators achieved the best performance possible at predicting recordings from mouse medial entorhinal cortex, leading the authors to call for better neural data.

However, multiple independent lines of evidence demonstrated that these high R^2 deep learning models are worse models of grid cells: (1) The required supervised targets, putative place cells, contradict known biological properties of place cells at both the single cell and population levels [67]; (2) The grid-like units lack key properties of real grid cells: periodic triangular tuning curves, multiple discrete grid modules, and specific ratios between grid modules [66, 68]; (3) the artificial grid units in some works were statistically indistinguishable from low pass-filtered noise [74, 75]. (4) In terms of evolutionary origins, the path integration objective of high-R^2 networks is not a sufficient objective for grid cells, as demonstrated in empirical deep neural network work [42, 41, 68], argued by prior theoretical work [28, 76, 53, 83], and shown by newer deep learning models [31, 85, 22, 70, 86, 87].

To summarize, the neural regressions methodology strongly supported deep learning-based path integrators because the networks achieved high neural predictivity scores, despite their discrepancies with multiple key criteria of neural similarity (listed in Appendix Sec. A). Why?

3 The Neural Regressions Methodology Reveals the Implicit Biases of Regression, Not Which Candidate Networks Are Similar to the Brain

Schaeffer et al. [66] made a conjecture: “different [models] achieve different neural predictivity scores because they learn different intrinsic dimensionalities, that then provide richer/poorer bases for linear regressions." Larger models simply provide more basis features for regression, and thus can provide better predictions independent of the similarity with the brain. To test their conjecture, the authors trained the same networks studied by Nayebi et al. [57] and empirically discovered that

reported test Pearson correlations exhibit an approximately linear-log relationship with a widely-used measure of effective dimensionality called participation ratio (PR) [21] (Fig 2a).

More precisely, consider P stimuli, and denote artificial activations with M units as $\mathbf{X} \in \mathbb{R}^{P \times M}$ and biological responses with N neurons as $\mathbf{Y} \in \mathbb{R}^{P \times N}$. We fit linear models using $p < P$ data:

$$\hat{\beta}(p) \stackrel{\text{def}}{=} \underset{\beta \in \mathbb{R}^{M \times N}}{\arg\min} ||\mathbf{X}_{1:p}\,\beta - \mathbf{Y}_{1:p}||_F^2 + \alpha_{\text{reg}}||\beta||_F^2 \tag{1}$$

Letting $\mathbf{X}\mathbf{X}^T = \sum_{i=1}^{P} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$, Schaeffer et al. [66] empirically found that approximately:

$$\text{Test } R^2 \sim \alpha \log(\text{Participation Ratio}) + \beta \quad ; \quad \text{Participation Ratio} \stackrel{\text{def}}{=} \frac{(\sum_{i=1}^{P} \lambda_i)^2}{\sum_{i=1}^{P} \lambda_i^2} \tag{2}$$

Participation ratio (PR) is a linear geometric measure of effective dimensionality: for uniform eigenvalues, the PR is the ambient dimensionality, whereas for a single non-zero eigenvalue, the PR is 1. Concurrent and subsequent work found quantitatively similar results across species, modalities, brain circuits and recording technologies: Elmoznino and Bonner [25] in deep convolutional networks trained on vision tasks to predict macaque IT cortex (Fig 2b), Tuckute et al. [80] in deep auditory networks to predict human brain-wide fMRI responses (Fig 2c), and Cheng and Antonello [17] in language models to predict human brain-wide fMRI responses (Fig. 2d). This finding by four independent research groups across different data modalities, tasks, architectures, species and recording technologies is puzzling. Are these results indicative of some deeper scientific insight into the brain?

In our view, no. *This pattern is attributable to the neural regressions methodology, not the brain.* Participation ratio (PR) was a reasonable first guess but an imprecise one that was subsequently refined into a more complete spectral theory of the regressions methodology. Canatar et al. [14] showed the normalized error $E_g(p)$ of *any* linear model $\hat{\mathbf{Y}}(p) \stackrel{\text{def}}{=} \mathbf{X}\,\hat{\beta}(p)$ is given as:

$$E_g(p) \stackrel{\text{def}}{=} \frac{||\hat{\mathbf{Y}}(p) - \mathbf{Y}||_F^2}{||\mathbf{Y}||_F^2} = \sum_{i=1}^{P} \frac{||\mathbf{Y}^T \mathbf{v}_i||_2^2}{||\mathbf{Y}||_F^2} \cdot \frac{\kappa^2}{1-\gamma} \frac{1}{(p\lambda_i + \kappa)^2}, \tag{3}$$

where $\gamma = \sum_{i=1}^{P} \frac{p\lambda_i^2}{(p\lambda_i+\kappa)^2}$ and $\kappa = \alpha_{\text{reg}} + \kappa \sum_{i=1}^{P} \frac{\lambda_i}{p\lambda_i+\kappa}$ must be solved self-consistently. This result says that the focus on PR by previous work was incomplete: PR partially captures the dimensionality of the learnable subspace, but the error *also* depends on the terms $\{||Y^T v_i||_2^2 / ||Y||_F^2\}_i$, which express whether the target Y lies in that subspace. Thus, higher PR can be beneficial to express the task fully, but can also be harmful by being too expressive and then harming sample complexity. This partially explains why ZCA whitening to maximize participation ratio did not achieve exceptional neural predictivity, why increasing the number of covariates does not necessarily increase neural predictivity [25], and how randomly initialized networks can achieve high neural predictivity [45, 2].

While it may be tempting to think that these empirical results and this spectral theory of neural predictivity have taught us about the brain, note that this theory makes no assumptions about a neural, behavioral, biological, ethological or otherwise meaningful relationship between $\mathbf{X}$ and $\mathbf{Y}$. Rather, as its origin makes clear [8, 13, 12], this theory is fundamentally *a description of linear regression* [69]. Consequently, this leads to the following realization:

> *Taken to its extreme, the neural regressions methodology has taught us the implicit biases of our chosen proxy function (e.g., test R^2 of linear regression), not which candidate artificial neural networks are actually similar to the brain.*

4 Discussion

To summarize, NeuroAI lacks canonical definitions of neural similarity, and such notions are likely task-, system-, and question-dependent, as well as difficult to quantify. Rather than facing these challenges, the neural regressions methodology sidesteps them by defining a proxy – for instance, the test R^2 of linear regressions fit between biological recordings and artificial activations – and then choosing networks based on this proxy. The networks that win a selection process do so because of the proxy's implicit biases, independent of any meaningful relationship with the brain.

To explain with an analogy, in the field of language modeling, researchers want language models to generate responses preferred by humans. However, collecting human preferences is slow, costly and noisy, so researchers instead train modified language models called reward models to serve as proxies of human preferences. These reward models are a proxy for what we actually care about – human preferences – but the field is willing to use these proxies because the reward models are directly trained to emulate human preferences and are correlated with human preferences empirically [93, 78, 6, 60, 51]; even so, overfitting to the reward models at the expense of human preferences is still a commonly encountered problem [78, 30, 1].

In comparison, in computational neuroscience, researchers want models that are most similar to brain system(s) of interest. However, interacting with neural systems and running experiments is slow, costly and noisy, so researchers instead fit neural regressions to serve as proxies of neural similarity. These regressions are a proxy for what we actually care about – neural similarity – but in contrast with reward models, neural regressions are not trained to emulate neural similarity and have no known relationship with neural similarity.

To reiterate an earlier point, we focused on linear regression because of its ubiquity in the literature, but other proxies of neural similarity (e.g., RSA [49], CKA [48], SVCCA [63], Procrustes [84]) would not escape this critique; rather, *other proxies would simply change the pertinent implicit biases*.

Altogether, these insights suggest that the neural regressions methodology may be flawed, teaching us about the preferences that we as researchers implicitly chose instead of advancing humanity's understanding of the brain. We conclude by calling for a critical and careful re-evaluation in computational neuroscience of how the neural regressions methodology is used and interpreted. For a Future Outlook, please see Appendix Sec. B.

References

[1] A. M. Ahmed, R. Rafailov, S. Sharkov, X. Li, and S. Koyejo. Scalable ensembling for mitigating reward overoptimisation. *arXiv preprint arXiv:2406.01013*, 2024.

[2] B. AlKhamissi, G. Tuckute, A. Bosselut, and M. Schrimpf. Brain-like language processing via a shallow untrained multihead attention network. *arXiv preprint arXiv:2406.15109*, 2024.

[3] R. Antonello, J. S. Turek, V. Vo, and A. Huth. Low-dimensional structure in the space of language representations is reflected in brain responses. *Advances in neural information processing systems*, 34:8332–8344, 2021.

[4] R. Antonello, A. Vaidya, and A. Huth. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 36, 2024.

[5] K. L. Aw and M. Toneva. Training language models to summarize narratives improves brain alignment. In *The Eleventh International Conference on Learning Representations*, 2023.

[6] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[7] A. Banino, C. Barry, B. Uria, C. Blundell, T. Lillicrap, P. Mirowski, A. Pritzel, M. J. Chadwick, T. Degris, J. Modayil, G. Wayne, H. Soyer, F. Viola, B. Zhang, R. Goroshin, N. Rabinowitz, R. Pascanu, C. Beattie, S. Petersen, A. Sadik, S. Gaffney, H. King, K. Kavukcuoglu, D. Hassabis, R. Hadsell, and D. Kumaran. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, May 2018. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-018-0102-6. URL http://www.nature.com/articles/s41586-018-0102-6.

[8] B. Bordelon, A. Canatar, and C. Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.

[9] J. S. Bowers, G. Malhotra, M. Dujmović, M. L. Montero, C. Tsvetkov, V. Biscione, G. Puebla, F. Adolfi, J. E. Hummel, R. F. Heaton, et al. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46:e385, 2023.

[10] Y. Burak and I. Fiete. Do We Understand the Emergent Dynamics of Grid Cell Activity? *Journal of Neuroscience*, 26(37):9352–9354, Sept. 2006. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.2857-06.2006. URL https://www.jneurosci.org/content/26/37/9352. Publisher: Society for Neuroscience Section: Journal Club.

[11] Y. Burak and I. R. Fiete. Accurate Path Integration in Continuous Attractor Network Models of Grid Cells. *PLOS Computational Biology*, 5(2):e1000291, Feb. 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000291. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000291. Publisher: Public Library of Science.

[12] A. Canatar, B. Bordelon, and C. Pehlevan. Out-of-distribution generalization in kernel regression. *Advances in Neural Information Processing Systems*, 34:12600–12612, 2021.

[13] A. Canatar, B. Bordelon, and C. Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):2914, 2021.

[14] A. Canatar, J. Feather, A. Wakhloo, and S. Chung. A spectral theory of neural prediction and alignment. *Advances in Neural Information Processing Systems*, 36, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9308d1b7d4ae2d3e2e67ae94b1078bf7-Abstract-Conference.html.

[15] C. Caucheteux and J.-R. King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134, 2022.

[16] C. Caucheteux, A. Gramfort, and J.-R. King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3):430–441, 2023.

[17] E. Cheng and R. J. Antonello. Evidence from fmri supports a two-phase abstraction process in language models. *arXiv preprint arXiv:2409.05771*, 2024.

[18] C. Conwell, J. S. Prince, K. N. Kay, G. A. Alvarez, and T. Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, pages 2022–03, 2022.

[19] C. J. Cueva and X.-X. Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. *International Conference on Learning Representations*, page 19, 2018.

[20] M. Davari, S. Horoi, A. Natik, G. Lajoie, G. Wolf, and E. Belilovsky. Reliability of cka as a similarity measure in deep learning. *arXiv preprint arXiv:2210.16156*, 2022.

[21] M. Del Giudice. Effective dimensionality: A tutorial. *Multivariate behavioral research*, 56(3): 527–542, 2021.

[22] W. Dorrell, P. E. Latham, T. E. Behrens, and J. C. Whittington. Actionable neural representations: Grid cells from minimal constraints. In *The Eleventh International Conference on Learning Representations*, 2023.

[23] M. Dujmovic, J. Bowers, F. Adolfi, and G. Malhotra. Inferring dnn-brain alignment using representational similarity analyses can be problematic. *ICLR 2024 Workshop on Representational Alignment*, 2024.

[24] M. Eickenberg, A. Gramfort, G. Varoquaux, and B. Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

[25] E. Elmoznino and M. F. Bonner. High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLOS Computational Biology*, 20(1):e1011792, 2024.

[26] J. Feather, G. Leclerc, A. Mądry, and J. H. McDermott. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, 26(11): 2017–2034, 2023.

[27] E. Feghhi, N. Hadidi, B. Song, I. A. Blank, and J. C. Kao. What are large language models mapping to in the brain? a case against over-reliance on brain scores. *arXiv preprint arXiv:2406.01538*, 2024.

[28] I. R. Fiete, Y. Burak, and T. Brookings. What grid cells convey about rat location. *Journal of Neuroscience*, 28(27):6858–6871, 2008.

[29] I. R. Fiete, Y. Burak, and T. Brookings. What Grid Cells Convey about Rat Location. *Journal of Neuroscience*, 28(27):6858–6871, July 2008. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.5684-07.2008. URL https://www.jneurosci.org/content/28/27/6858. Publisher: Society for Neuroscience Section: Articles.

[30] L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

[31] R. Gao, J. Xie, S.-C. Zhu, and Y. N. Wu. Learning grid cells as vector representation of self-position coupled with matrix representation of self-motion. *arXiv preprint arXiv:1810.05597*, 2018.

[32] R. J. Gardner, E. Hermansen, M. Pachitariu, Y. Burak, N. A. Baas, B. A. Dunn, M.-B. Moser, and E. I. Moser. Toroidal topology of population activity in grid cells. *Nature*, 602(7895): 123–128, Feb. 2022. ISSN 1476-4687. doi: 10.1038/s41586-021-04268-7. URL https://www.nature.com/articles/s41586-021-04268-7. Number: 7895 Publisher: Nature Publishing Group.

[33] A. Goldstein, E. Ham, M. Schain, S. Nastase, Z. Zada, A. Dabush, B. Aubrey, H. Gazula, A. Feder, W. K. Doyle, et al. The temporal structure of language processing in the human brain corresponds to the layered hierarchy of deep language models. *arXiv preprint arXiv:2310.07106*, 2023.

[34] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, Aug. 2005. ISSN 1476-4687. doi: 10.1038/nature03721. URL https://www.nature.com/articles/nature03721. Number: 7052 Publisher: Nature Publishing Group.

[35] Y. Han, T. A. Poggio, and B. Cheung. System identification of neural systems: If we got it right, would we know? In *International Conference on Machine Learning*, pages 12430–12444. PMLR, 2023.

[36] M. Helmer, S. Warrington, A.-R. Mohammadi-Nejad, J. L. Ji, A. Howell, B. Rosand, A. Anticevic, S. N. Sotiropoulos, and J. D. Murray. On the stability of canonical correlation analysis and partial least squares with application to brain-behavior associations. *Communications Biology*, 7(1):217, 2024.

[37] Z. Hong, H. Wang, Z. Zada, H. Gazula, D. Turner, B. Aubrey, L. Niekerken, W. Doyle, S. Devore, P. Dugan, et al. Scale matters: Large language models with billions (rather than millions) of parameters better match neural representations of natural language. *bioRxiv*, pages 2024–06, 2024.

[38] E. A. Hosseini, M. Schrimpf, Y. Zhang, S. Bowman, N. Zaslavsky, and E. Fedorenko. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, 5(1):43–63, 2024.

[39] S. Jain, V. Vo, S. Mahto, A. LeBel, J. S. Turek, and A. Huth. Interpretable multi-timescale models for predicting fmri responses to continuous natural speech. *Advances in Neural Information Processing Systems*, 33:13738–13749, 2020.

[40] H. Jang, D. McCormack, and F. Tong. Noise-trained deep neural networks effectively predict human vision and its neural responses to challenging images. *PLoS biology*, 19(12):e3001418, 2021.

[41] I. Kanitscheider and I. Fiete. Emergence of dynamically reconfigurable hippocampal responses by learning to perform probabilistic spatial reasoning. *bioRxiv*, page 231159, 2017.

[42] I. Kanitscheider and I. Fiete. Training recurrent networks to generate hypotheses about how the brain solves hard navigation problems. *Advances in Neural Information Processing Systems*, 30, 2017.

[43] K. Kar, J. Kubilius, K. Schmidt, E. B. Issa, and J. J. DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience*, 22(6):974–983, 2019.

[44] C. Kauf, G. Tuckute, R. Levy, J. Andreas, and E. Fedorenko. Lexical-semantic content, not syntactic structure, is the main contributor to ann-brain similarity of fmri responses in the language network. *Neurobiology of Language*, 5(1):7–42, 2024.

[45] A. Kazemian, E. Elmoznino, and M. F. Bonner. Convolutional architectures are cortex-aligned de novo. *bioRxiv*, pages 2024–05, 2024.

[46] A. J. E. Kell, D. L. K. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3):630–644.e16, May 2018. ISSN 1097-4199. doi: 10.1016/j.neuron.2018.03.044.

[47] M. Khona, S. Chandra, and I. R. Fiete. From smooth cortical gradients to discrete modules: spontaneous and topologically robust emergence of modularity in grid cells. *bioRxiv*, pages 2021–10, 2022.

[48] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.

[49] N. Kriegeskorte, M. Mur, and P. A. Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

[50] J. Kubilius, M. Schrimpf, K. Kar, R. Rajalingham, H. Hong, N. Majaj, E. Issa, P. Bashivan, J. Prescott-Roy, K. Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, 32, 2019.

[51] N. Lambert, V. Pyatkin, J. Morrison, L. Miranda, B. Y. Lin, K. Chandu, N. Dziri, S. Kumar, T. Zick, Y. Choi, N. A. Smith, and H. Hajishirzi. Rewardbench: Evaluating reward models for language modeling. `https://huggingface.co/spaces/allenai/reward-bench`, 2024.

[52] G. Malhotra and J. Bowers. Predicting brain activation does not license conclusions regarding DNN-brain alignment: The case of Brain-Score. In *Cognitive Computational Neuroscience Conference*, Johnson Ice Rink, August 2024. Poster B81.

[53] A. Mathis, A. V. Herz, and M. Stemmler. Optimal population codes for space: grid cells outperform place cells. *Neural computation*, 24(9):2280–2317, 2012.

[54] J. Mehrer, C. J. Spoerer, N. Kriegeskorte, and T. C. Kietzmann. Individual differences among deep neural network models. *Nature communications*, 11(1):5725, 2020.

[55] J. Millet, C. Caucheteux, Y. Boubenec, A. Gramfort, E. Dunbar, C. Pallier, J.-R. King, et al. Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems*, 35:33428–33443, 2022.

[56] G. Mischler, Y. A. Li, S. Bickel, A. D. Mehta, and N. Mesgarani. Contextual feature extraction hierarchies converge in large language models and the brain. *arXiv preprint arXiv:2401.17671*, 2024.

[57] A. Nayebi, A. Attinger, M. Campbell, K. Hardcastle, I. Low, C. S. Mallory, G. Mel, B. Sorscher, A. H. Williams, S. Ganguli, L. Giocomo, and D. Yamins. Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 12167–12179. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/656f0dbf9392657eed7feefc486781fb-Abstract.html`.

[58] S. Oota, M. Gupta, and M. Toneva. Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[59] M. Ostrow, A. Eisen, L. Kozachkov, and I. Fiete. Beyond geometry: Comparing the temporal structure of computation in neural circuits with dynamical similarity analysis. *Advances in Neural Information Processing Systems*, 36, 2024.

[60] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[61] A. Pasquiou, Y. Lakretz, J. Hale, B. Thirion, and C. Pallier. Neural language models are not born equal to fit brain data, but training helps. *arXiv preprint arXiv:2207.03380*, 2022.

[62] F. Pereira, B. Lou, B. Pritchett, S. Ritter, S. J. Gershman, N. Kanwisher, M. Botvinick, and E. Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):963, 2018.

[63] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

[64] N. A. Ratan Murty, P. Bashivan, A. Abate, J. J. DiCarlo, and N. Kanwisher. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature communications*, 12(1):5540, 2021.

[65] R. Schaeffer, M. Khona, L. Meshulam, I. Fiete, et al. Reverse-engineering recurrent neural network solutions to a hierarchical inference task for mice. *Advances in Neural Information Processing Systems*, 33:4584–4596, 2020.

[66] R. Schaeffer, M. Khona, and I. Fiete. No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. *Advances in neural information processing systems*, 35:16052–16067, 2022.

[67] R. Schaeffer, M. Khona, A. Bertagnoli, S. Koyejo, and I. R. Fiete. Testing assumptions underlying a unified theory for the origin of grid cells. *arXiv preprint arXiv:2311.16295*, 2023.

[68] R. Schaeffer, M. Khona, S. Koyejo, and I. R. Fiete. Disentangling fact from grid cell fiction in trained deep path integrators. *ArXiv*, 2023.

[69] R. Schaeffer, M. Khona, Z. Robertson, A. Boopathy, K. Pistunova, J. W. Rocks, I. R. Fiete, and O. Koyejo. Double descent demystified: Identifying, interpreting & ablating the sources of a deep learning puzzle. *arXiv preprint arXiv:2303.14151*, 2023.

[70] R. Schaeffer, M. Khona, T. Ma, C. Eyzaguirre, S. Koyejo, and I. Fiete. Self-supervised learning of representations for space generates multi-modular grid cells. *Advances in Neural Information Processing Systems*, 36, 2024.

[71] M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, F. Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.

[72] M. Schrimpf, J. Kubilius, M. J. Lee, N. A. R. Murty, R. Ajemian, and J. J. DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3): 413–423, 2020.

[73] M. Schrimpf, I. A. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118,

Nov. 2021. doi: 10.1073/pnas.2105646118. URL https://www.pnas.org/doi/10.1073/pnas.2105646118. Publisher: Proceedings of the National Academy of Sciences.

[74] B. Sorscher, G. C. Mel, S. Ganguli, and S. A. Ocko. A unified theory for the origin of grid cells through the lens of pattern formation. *Advances in Neural Information Processing Systems*, page 18, 2019.

[75] B. Sorscher, G. C. Mel, S. A. Ocko, L. M. Giocomo, and S. Ganguli. A unified theory for the computational and mechanistic origins of grid cells. *Neuron*, 111(1):121–137, 2023.

[76] S. Sreenivasan and I. Fiete. Grid cells generate an analog error-correcting code for singularly precise neural computation. *Nature Neuroscience*, 14(10):1330–1337, Oct. 2011. ISSN 1546-1726. doi: 10.1038/nn.2901. URL https://www.nature.com/articles/nn.2901. Number: 10 Publisher: Nature Publishing Group.

[77] H. Stensola, T. Stensola, T. Solstad, K. Frøland, M.-B. Moser, and E. I. Moser. The entorhinal grid map is discretized. *Nature*, 492(7427):72–78, Dec. 2012. ISSN 1476-4687. doi: 10.1038/nature11649. URL https://www.nature.com/articles/nature11649. Number: 7427 Publisher: Nature Publishing Group.

[78] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano. Learning to summarize from human feedback, 2022. URL https://arxiv.org/abs/2009.01325.

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

[80] G. Tuckute, J. Feather, D. Boebinger, and J. H. McDermott. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *Plos Biology*, 21(12):e3002366, 2023.

[81] G. Tuckute, A. Sathe, S. Srikant, M. Taliaferro, M. Wang, M. Schrimpf, K. Kay, and E. Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024.

[82] A. R. Vaidya, S. Jain, and A. G. Huth. Self-supervised models of audio effectively explain human cortical responses to speech, May 2022. URL http://arxiv.org/abs/2205.14252. arXiv:2205.14252 [cs].

[83] X.-X. Wei, J. Prentice, and V. Balasubramanian. A principle of economy predicts the functional architecture of grid cells. *Elife*, 4:e08362, 2015.

[84] A. H. Williams, E. Kunz, S. Kornblith, and S. Linderman. Generalized shape metrics on neural representations. *Advances in Neural Information Processing Systems*, 34:4738–4750, 2021.

[85] D. Xu, R. Gao, W.-H. Zhang, X.-X. Wei, and Y. N. Wu. Conformal isometry of lie group representation in recurrent network of grid cells. *arXiv preprint arXiv:2210.02684*, 2022.

[86] D. Xu, R. Gao, W.-H. Zhang, X.-X. Wei, and Y. N. Wu. Emergence of grid-like representations by training recurrent networks with conformal normalization, 2024. URL https://arxiv.org/abs/2310.19192.

[87] D. Xu, R. Gao, W.-H. Zhang, X.-X. Wei, and Y. N. Wu. An investigation of conformal isometry hypothesis for grid cells. *arXiv preprint arXiv:2405.16865*, 2024.

[88] Y. Xu and M. Vaziri-Pashkam. Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature communications*, 12(1):2065, 2021.

[89] D. L. K. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, Mar. 2016. ISSN 1546-1726. doi: 10.1038/nn.4244. URL https://www.nature.com/articles/nn.4244. Number: 3 Publisher: Nature Publishing Group.

[90] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, June 2014. doi: 10.1073/pnas.1403112111. URL https://www.pnas.org/doi/abs/10.1073/pnas.1403112111. Publisher: Proceedings of the National Academy of Sciences.

[91] K. Yoon, M. A. Buice, C. Barry, R. Hayman, N. Burgess, and I. R. Fiete. Specific evidence of low-dimensional continuous attractor dynamics in grid cells. *Nature neuroscience*, 16(8): 1077–1084, 2013.

[92] C. Zhuang, S. Yan, A. Nayebi, M. Schrimpf, M. C. Frank, J. J. DiCarlo, and D. L. K. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, Jan. 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2014196118. URL https://pnas.org/doi/full/10.1073/pnas.2014196118.

[93] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

A Example Criteria of Neural Similarity to Grid Cells

In this paper, we intentionally do not provide a general definition of “neural similarity" (see Future Outlook - Appendix Sec. B), in part because we feel such a definition is likely highly context dependent. But we can offer a constructive example in the narrow context of grid and place cells viewed at the level of their circuit dynamics. When considering models, researchers often consider the following (non-exhaustive) list of relevant criteria for evaluating whether a model is similar to the circuit:

- Individual neurons exhibit equilateral triangular periodic tuning curves
- In the population of grid cells, multiple grid periodicities exist
- The periodicities of the grid cells are quantized
- The quantized periods of the modules exhibit precise ratios between adjacent periods
- The population states of each grid module (subpopulation with common period) lie on the surface of a 2D torus.
- The cell-cell relationships of co-modular grid cells (and their toroidal population states) are invariant across spatial environments and behavioral states.
- In any environment, grid-like tuning is present from the first trajectory fragment.
- While grid cells remain invariant in their relationships across environments, place cells remap or scramble their relationships.

B Future Outlook

Despite our critiques of the neural regressions methodology, model-system comparison is a fundamental and necessary component of a modeling science. How, then, can we move beyond flaws arising as a consequence of emphasizing only a single metric?

One possibility is to use a number of different comparisons that emphasize different aspects of model and system. This may include comparing behavior on top of neural activations, as is already a feature of the Brain-Score platform ([89, 71]), neural dynamics on top of neural geometry [59], or using a variety of different metrics that have different biases ([35]). Beyond linear regression, computational neuroscience has introduced a number of other candidates into the literature, including RSA [49], Procrustes [84], CKA [48], SVCCA [63], and a number of variants of these metrics. All of the above metrics compare geometric features of neural activations. Recently proposed methods, such as Dynamical Similarity Analysis (DSA, [59]) compare different features of the system, like dynamical structure. Older work sought to study similarity using combined perspectives of behavior, representations, dynamics and circuit mechanisms, e.g., [65]. Using more types of comparison, both in terms of metrics and data, should help mitigate the biases of individual comparisons, making Goodharting more challenging. However, it is important to note that even combinations of such metrics are liable to fall prey to Goodhart’s law. Depending on the scientific question, the relevant quantity to be compared may change.

More generally, beyond significantly increasing the number of types of comparisons being done, it is worth taking a step back and asking what we mean by a ‘good model’. Although we do not define neural similarity here, in our view, neural similarity depends on the task, the neural system and the particular scientific question. Thus, in our opinion, neural similarity cannot always be neural predictivity. This perspective lays bare the difficulty and, in a sense incoherence, of seeking to globally define good models in terms of one or even a set of metrics.