Does Maximizing Neural Regression Scores Teach Us About The Brain?

Editors: List of editors' names

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 this methodology has become overused, and a more pluralistic approach is needed. Our view is 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 an egregious failure of the neural regressions methodology in which the best fitting 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.

Keywords: Computational neuroscience, brain-score, similarity metrics, neural alignment, neural network models of the brain, neural regressions methodology

1. Introduction

An influential methodology in computational neuroscience 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 (Yamins and DiCarlo, 2016a). The claim is that the ANN(s) with better performing neural regressions are more similar to the brain than alternative models. This approach has been widely used in vision, audition, language, and spatial navigation, most often with (regularized) linear models but occasionally with other metrics; due to limited space, we defer citations to Related Work (App. Sec. A).

In this opinion piece, we argue that computational neuroscience lacks sufficiently rich definitions of neural similarity, and such notions are likely context-dependent and difficult to construct. 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 models that win a selection process (e.g., on BrainScore) may do so more because of implicit biases of the proxy than because of meaningful relationships with the brain.

This perspective explains why, for example, the neural regressions methodology was confidently incorrect when applied to models of grid cells: linear regression has no interest

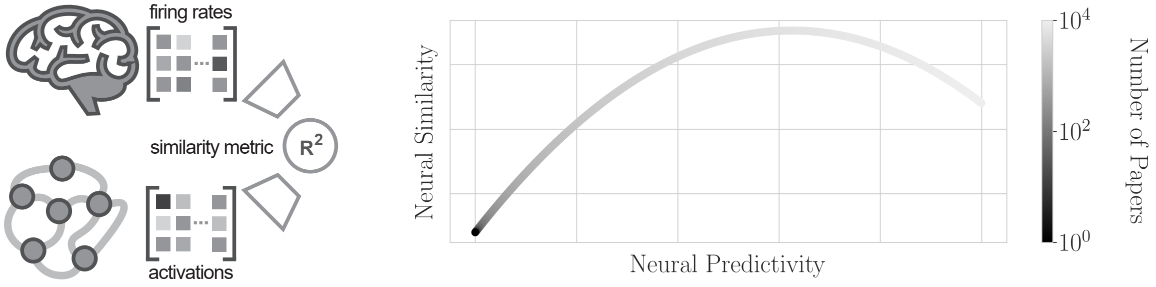

Figure 1: Computational neuroscience lacks canonical definitions of neural similarity, and rather than engaging with this difficult question, the neural regressions methodology simply devised a proxy – neural predictivity – then overfit to the proxy without verifying the extent to which the proxy agrees with neural similarity. Although we don't define neural similarity here, we emphasize that it is task, model, and question-dependent, and hence cannot always be neural predictivity.

in key criteria of neural similarity for grid cells: periodic tuning curves (Hafting et al., 2005), multiple grid modules with specific period ratios (Stensola et al., 2012), toroidal continuous attractor dynamics (Yoon et al., 2013; Gardner et al., 2022). This perspective also explains a finding by four independent research groups in different modalities, data, architectures and recording technologies (Schaeffer et al., 2022; Elmoznino and Bonner, 2024; Tuckute et al., 2023; Cheng and Antonello, 2024) of a quantitatively consistent relationship between test R^2 and effective dimensionality, that was mathematically refined and further empirically studied by Canatar et al. (2024): (linear) neural predictivity *is* (linear) regression, and (linear) regression has implicit 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 (Kriegeskorte et al., 2008b), CKA (Kornblith et al., 2019), SVCCA (Raghu et al., 2017), Procrustes (Williams et al., 2021), etc.) would not escape this critique; rather, they would simply change the implicit 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, 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. (2023) documented how artificial networks preferred by the neural regressions methodology lack or contradict properties of primate vision, and others have identified additional flaws (Mehrer et al., 2020; Xu and Vaziri-Pashkam, 2021; Han et al., 2023; Feather et al., 2023; Feghhi et al., 2024). 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 (Hafting et al., 2005) – differ from vision, audition and language in that humanity possesses scientific models (Fiete et al., 2008b; Burak and Fiete, 2009a, 2006; Sreenivasan and Fiete, 2011a) that have repeatedly proven predictive (Stensola et al. (2012); Yoon et al.

(2013); Gardner et al. (2022)), 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. In a domain we understand well, how did the regressions methodology do?

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 of grid cells are modeling their dynamics and the evolutionary causes for their existence. Previous and now near-normative models of grid cells 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 (Burak and Fiete, 2009b; Khona et al., 2022); and that multiple grid modules played key roles in disambiguating position over large ranges and in error correction (Fiete et al., 2008c; Sreenivasan and Fiete, 2011b). Later, models based on deep recurrent networks trained in a supervised manner to path integrate were shown to learn grid-like units (Banino et al., 2018; Cueva and Wei, 2018; Sorscher et al., 2019), and neural-regressions based work (Nayebi et al., 2021) showed that these supervised deep recurrent 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 (Schaeffer et al., 2023a); (2) The grid-like units lack key properties of real grid cells: properly periodic triangular tuning curves, multiple discrete grid modules, and specific ratios between grid modules (Schaeffer et al., 2022, 2023b); (3) the artificial grid units in some works were statistically indistinguishable from low pass-filtered noise (Sorscher et al., 2019, 2023). (4) In terms of evolutionary origins, the path integration objective of high-R^2 networks is not a sufficient objective for grid cells, as shown in Kanitscheider and Fiete (2017b,a); Schaeffer et al. (2023b), argued by prior theoretical work (Fiete et al., 2008a; Sreenivasan and Fiete, 2011a; Mathis et al., 2012; Wei et al., 2015), and shown by newer deep learning models (Gao et al., 2018; Xu et al., 2022; Dorrell et al., 2023; Schaeffer et al., 2024; Xu et al., 2024a,b). For common criteria of neural similarity to grid cells, see App. Sec. B

Why, then, did the neural regressions methodology so strongly support deep path integrators despite their discrepancies with known important properties of the neural circuit?

3. The Neural Regressions Methodology Reveals Insights Into Regression, Not Insights Into the Brain

Schaeffer et al. (2022) were unable to obtain the networks or neural recordings of mouse medial entorhinal cortex to investigate this question, but made a rough conjecture: “different [models] achieve different neural predictivity scores because different models learn different intrinsic dimensionalities that then provide richer/poorer bases for linear regressions.” As evidence, the authors trained the same networks studied by Nayebi et al. (2021) and showed that reported test Pearson correlations exhibit an approximately linear-log relationship with a measure of effective dimensionality of networks' representations called *participation ratio* (Fig 2a). To be more mathematically precise, 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 $\mathbf{Y} \in \mathbb{R}^{P \times N}$. The authors fit linear models using a training set of size $p < P$:

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

Letting $\mathbf{X}\mathbf{X}^T = \sum_{i=1}^P \lambda_i \mathbf{v}_i \mathbf{v}_i^T$, Schaeffer et al. (2022) found that approximately:

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

Concurrent and subsequent work found quantitatively similar results across species, modalities, brain circuits and neural recording technologies: Elmoznino and Bonner (2024) in deep convolutional networks trained on vision tasks to predict macaque IT cortex (Fig 2b), Tuckute et al. (2023) in deep auditory networks to predict human brain-wide fMRI responses (Fig 2c), and Cheng and Antonello (2024) in language models to predict human brain-wide fMRI responses (Fig. 2d). This finding by four independent research groups across different data modalities, training tasks, ANN architectures, species and neural 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 that was refined into a more descriptive spectral theory of the neural regressions methodology; specifically, Canatar et al. (2024) showed the normalized error $E_g(p)$ of any linear model $\hat{\mathbf{Y}}(p) \overset{\text{def}}{=} \mathbf{X}\hat{\beta}(p)$ is given as:

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

where $\kappa = \alpha_{\text{reg}} + \kappa \sum_{i=1}^P \frac{\lambda_i}{p\lambda_i + \kappa}$, $\gamma = \sum_{i=1}^P \frac{p\lambda_i^2}{(p\lambda_i+\kappa)^2}$. This result reveals higher dimensionality *can* reduce prediction error, but not always, and that a full characterization depends on the interplay between eigenvalues, eigenvectors and regression targets. Importantly, note that this theory of neural predictivity 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 (Bordelon et al., 2020; Canatar et al., 2021), this theory is fundamentally *a description of linear regression* (Schaeffer et al., 2023c). This leads to the 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.*

Due to space limitations, we defer our Future Outlook to App. Sec. D.

References

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. Brain-like language processing via a shallow untrained multihead attention network. *arXiv preprint arXiv:2406.15109*, 2024.

Richard Antonello, Javier S Turek, Vy Vo, and Alexander 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.

Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 36, 2024.

Khai Loong Aw and Mariya Toneva. Training language models to summarize narratives improves brain alignment. In *The Eleventh International Conference on Learning Representations*, 2023.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J. Chadwick, Thomas Degris, Joseph Modayil, Greg Wayne, Hubert Soyer, Fabio Viola, Brian Zhang, Ross Goroshin, Neil Rabinowitz, Razvan Pascanu, Charlie Beattie, Stig Petersen, Amir Sadik, Stephen Gaffney, Helen King, Koray Kavukcuoglu, Demis Hassabis, Raia Hadsell, and Dharshan 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.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.

Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E Hummel, Rachel F Heaton, et al. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46:e385, 2023.

Yoram Burak and Ila Fiete. Do We Understand the Emergent Dynamics of Grid Cell Activity? *Journal of Neuroscience*, 26(37):9352–9354, September 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.

Yoram Burak and Ila R. Fiete. Accurate Path Integration in Continuous Attractor Network Models of Grid Cells. *PLOS Computational Biology*, 5(2):e1000291, February 2009a. 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.

Yoram Burak and Ila R. Fiete. Accurate Path Integration in Continuous Attractor Network Models of Grid Cells. *PLOS Computational Biology*, 5(2):e1000291, February 2009b. 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.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):2914, 2021.

Abdulkadir Canatar, Jenelle Feather, Albert Wakhloo, and SueYeon 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.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134, 2022.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3): 430–441, 2023.

Emily Cheng and Richard J Antonello. Evidence from fmri supports a two-phase abstraction process in language models. *arXiv preprint arXiv:2409.05771*, 2024.

Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia 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.

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

Will Dorrell, Peter E Latham, Timothy EJ Behrens, and James CR Whittington. Actionable neural representations: Grid cells from minimal constraints. In *The Eleventh International Conference on Learning Representations*, 2023.

Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

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

Jenelle Feather, Guillaume Leclerc, Aleksander Madry, and Josh H McDermott. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, 26(11):2017–2034, 2023.

Ebrahim Feghhi, Nima Hadidi, Bryan Song, Idan A Blank, and Jonathan 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.

Ila R Fiete, Yoram Burak, and Ted Brookings. What grid cells convey about rat location. *Journal of Neuroscience*, 28(27):6858–6871, 2008a.

Ila R. Fiete, Yoram Burak, and Ted Brookings. What Grid Cells Convey about Rat Location. *Journal of Neuroscience*, 28(27):6858–6871, July 2008b. 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.

Ila R. Fiete, Yoram Burak, and Ted Brookings. What Grid Cells Convey about Rat Location. *Journal of Neuroscience*, 28(27):6858–6871, July 2008c. 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.

Ruiqi Gao, Jianwen Xie, Song-Chun Zhu, and Ying Nian Wu. Learning grid cells as vector representation of self-position coupled with matrix representation of self-motion. *arXiv preprint arXiv:1810.05597*, 2018.

Richard J. Gardner, Erik Hermansen, Marius Pachitariu, Yoram Burak, Nils A. Baas, Benjamin A. Dunn, May-Britt Moser, and Edvard I. Moser. Toroidal topology of population activity in grid cells. *Nature*, 602(7895):123–128, February 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.

Ariel Goldstein, Eric Ham, Mariano Schain, Samuel Nastase, Zaid Zada, Avigail Dabush, Bobbi Aubrey, Harshvardhan Gazula, Amir Feder, Werner 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.

Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I. Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, August 2005. ISSN 1476-4687. doi: 10.1038/nature03721. URL https://www.nature.com/articles/nature03721. Number: 7052 Publisher: Nature Publishing Group.

Yena Han, Tomaso A Poggio, and Brian 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.

Zhuoqiao Hong, Haocheng Wang, Zaid Zada, Harshvardhan Gazula, David Turner, Bobbi Aubrey, Leonard Niekerken, Werner Doyle, Sasha Devore, Patricia 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.

Eghbal A Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina 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.

Shailee Jain, Vy Vo, Shivangi Mahto, Amanda LeBel, Javier S Turek, and Alexander Huth. Interpretable multi-timescale models for predicting fmri responses to continuous natural speech. *Advances in Neural Information Processing Systems*, 33:13738–13749, 2020.

Hojin Jang, Devin McCormack, and Frank Tong. Noise-trained deep neural networks effectively predict human vision and its neural responses to challenging images. *PLoS biology*, 19(12):e3001418, 2021.

Ingmar Kanitscheider and Ila Fiete. Emergence of dynamically reconfigurable hippocampal responses by learning to perform probabilistic spatial reasoning. *bioRxiv*, page 231159, 2017a.

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

Kohitij Kar, Jonas Kubilius, Kailyn Schmidt, Elias B Issa, and James 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.

Carina Kauf, Greta Tuckute, Roger Levy, Jacob Andreas, and Evelina 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.

Atlas Kazemian, Eric Elmoznino, and Michael F Bonner. Convolutional architectures are cortex-aligned de novo. *bioRxiv*, pages 2024–05, 2024.

Alexander J. E. Kell, Daniel L. K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh 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.

Mikail Khona, Sarthak Chandra, and Ila R Fiete. From smooth cortical gradients to discrete modules: spontaneous and topologically robust emergence of modularity in grid cells. *bioRxiv*, pages 2021–10, 2022.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 2008a. ISSN 1662-5137. URL https://www.frontiersin.org/articles/10.3389/neuro.06.004.2008.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008b.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, 32, 2019.

Alexander Mathis, Andreas VM Herz, and Martin Stemmler. Optimal population codes for space: grid cells outperform place cells. *Neural computation*, 24(9):2280–2317, 2012.

Johannes Mehrer, Courtney J Spoerer, Nikolaus Kriegeskorte, and Tim C Kietzmann. Individual differences among deep neural network models. *Nature communications*, 11(1): 5725, 2020.

Juliette Millet, Charlotte Caucheteux, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, Jean-Remi 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.

Gavin Mischler, Yinghao Aaron Li, Stephan Bickel, Ashesh D Mehta, and Nima Mesgarani. Contextual feature extraction hierarchies converge in large language models and the brain. *arXiv preprint arXiv:2401.17671*, 2024.

Aran Nayebi, Alexander Attinger, Malcolm Campbell, Kiah Hardcastle, Isabel Low, Caitlin S Mallory, Gabriel Mel, Ben Sorscher, Alex H Williams, Surya Ganguli, Lisa Giocomo, and Dan 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.

SubbaReddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Mitchell Ostrow, Adam Eisen, Leo Kozachkov, and Ila Fiete. Beyond geometry: Comparing the temporal structure of computation in neural circuits with dynamical similarity analysis. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 33824–33837. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6ac807c9b296964409b277369e55621a-Paper-Conference.pdf.

Alexandre Pasquiou, Yair Lakretz, John Hale, Bertrand Thirion, and Christophe Pallier. Neural language models are not born equal to fit brain data, but training helps. *arXiv preprint arXiv:2207.03380*, 2022.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):963, 2018.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

N Apurva Ratan Murty, Pouya Bashivan, Alex Abate, James J DiCarlo, and Nancy Kanwisher. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature communications*, 12(1):5540, 2021.

Rylan Schaeffer, Mikail Khona, and Ila 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.

Rylan Schaeffer, Mikail Khona, Adrian Bertagnoli, Sanmi Koyejo, and Ila Rani Fiete. Testing assumptions underlying a unified theory for the origin of grid cells. *arXiv preprint arXiv:2311.16295*, 2023a.

Rylan Schaeffer, Mikail Khona, Sanmi Koyejo, and Ila Rani Fiete. Disentangling fact from grid cell fiction in trained deep path integrators. *ArXiv*, 2023b.

Rylan Schaeffer, Mikail Khona, Zachary Robertson, Akhilan Boopathy, Kateryna Pistunova, Jason W Rocks, Ila Rani Fiete, and Oluwasanmi Koyejo. Double descent demystified: Identifying, interpreting & ablating the sources of a deep learning puzzle. *arXiv preprint arXiv:2303.14151*, 2023c.

Rylan Schaeffer, Mikail Khona, Tzuhsuan Ma, Cristobal Eyzaguirre, Sanmi Koyejo, and Ila Fiete. Self-supervised learning of representations for space generates multi-modular grid cells. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, November 2021. doi: 10.1073/pnas.2105646118. URL https://www.pnas.org/doi/10.1073/pnas.2105646118. Publisher: Proceedings of the National Academy of Sciences.

Ben Sorscher, Gabriel C Mel, Surya Ganguli, and Samuel 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.

Ben Sorscher, Gabriel C Mel, Samuel A Ocko, Lisa M Giocomo, and Surya Ganguli. A unified theory for the computational and mechanistic origins of grid cells. *Neuron*, 111 (1):121–137, 2023.

Sameet Sreenivasan and Ila Fiete. Grid cells generate an analog error-correcting code for singularly precise neural computation. *Nature neuroscience*, 14(10):1330–1337, 2011a.

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

Hanne Stensola, Tor Stensola, Trygve Solstad, Kristian Frøland, May-Britt Moser, and Edvard I. Moser. The entorhinal grid map is discretized. *Nature*, 492(7427):72–78, December 2012. ISSN 1476-4687. doi: 10.1038/nature11649. URL https://www.nature.com/articles/nature11649. Number: 7427 Publisher: Nature Publishing Group.

Katherine R Storrs, Tim C Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus 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.

Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh 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.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024.

Aditya R. Vaidya, Shailee Jain, and Alexander 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].

Xue-Xin Wei, Jason Prentice, and Vijay Balasubramanian. A principle of economy predicts the functional architecture of grid cells. *Elife*, 4:e08362, 2015.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in Neural Information Processing Systems*, 34:4738–4750, 2021.

Dehong Xu, Ruiqi Gao, Wen-Hao Zhang, Xue-Xin Wei, and Ying Nian Wu. Conformal isometry of lie group representation in recurrent network of grid cells. *arXiv preprint arXiv:2210.02684*, 2022.

Dehong Xu, Ruiqi Gao, Wen-Hao Zhang, Xue-Xin Wei, and Ying Nian Wu. Emergence of grid-like representations by training recurrent networks with conformal normalization, 2024a. URL https://arxiv.org/abs/2310.19192.

Dehong Xu, Ruiqi Gao, Wen-Hao Zhang, Xue-Xin Wei, and Ying Nian Wu. An investigation of conformal isometry hypothesis for grid cells. *arXiv preprint arXiv:2405.16865*, 2024b.

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

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

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

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 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.

KiJung Yoon, Michael A Buice, Caswell Barry, Robin Hayman, Neil Burgess, and Ila R Fiete. Specific evidence of low-dimensional continuous attractor dynamics in grid cells. *Nature neuroscience*, 16(8):1077–1084, 2013.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L. K. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, January 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2014196118. URL https://pnas.org/doi/full/10.1073/pnas.2014196118.

Appendix A. Published Work Using the Neural Regressions Methodology

The neural regressions methodology has been widely used in vision (Yamins et al., 2014; Eickenberg et al., 2017; Schrimpf et al., 2018; Kar et al., 2019; Kubilius et al., 2019; Schrimpf et al., 2020; Zhuang et al., 2021; Jang et al., 2021; Xu and Vaziri-Pashkam, 2021; Storrs et al., 2021; Ratan Murty et al., 2021; Conwell et al., 2022; Kazemian et al., 2024), audition (Kell et al., 2018; Vaidya et al., 2022; Millet et al., 2022; Tuckute et al., 2023), language Pereira et al. (2018); Jain et al. (2020); Schrimpf et al. (2021); Antonello et al. (2021); Pasquiou et al. (2022); Caucheteux and King (2022); Caucheteux et al. (2023); Goldstein et al. (2023); Aw and Toneva (2023); AlKhamissi et al. (2024); Hosseini et al. (2024); Oota et al. (2024); Cheng and Antonello (2024); Kauf et al. (2024); Antonello et al. (2024); Tuckute et al. (2024); Mischler et al. (2024); Hong et al. (2024), and spatial navigation Nayebi et al. (2021), most often with (regularized) linear models, but occasionally with non-linear models. This list is not exhaustive and we welcome readers to contact us to suggest additional appropriate citations.

Appendix B. 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 - App. Sec. D), 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 cells. 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 is quantized
- The quantized periods of the modules exhibit precise ratios between adjacent periods
- The population of grid cells topologically lives on the cross product of multiple twisted tori, one per module
- That topological structure is a continuous attractor

Appendix C. Test R^2 Versus Participation Ratio

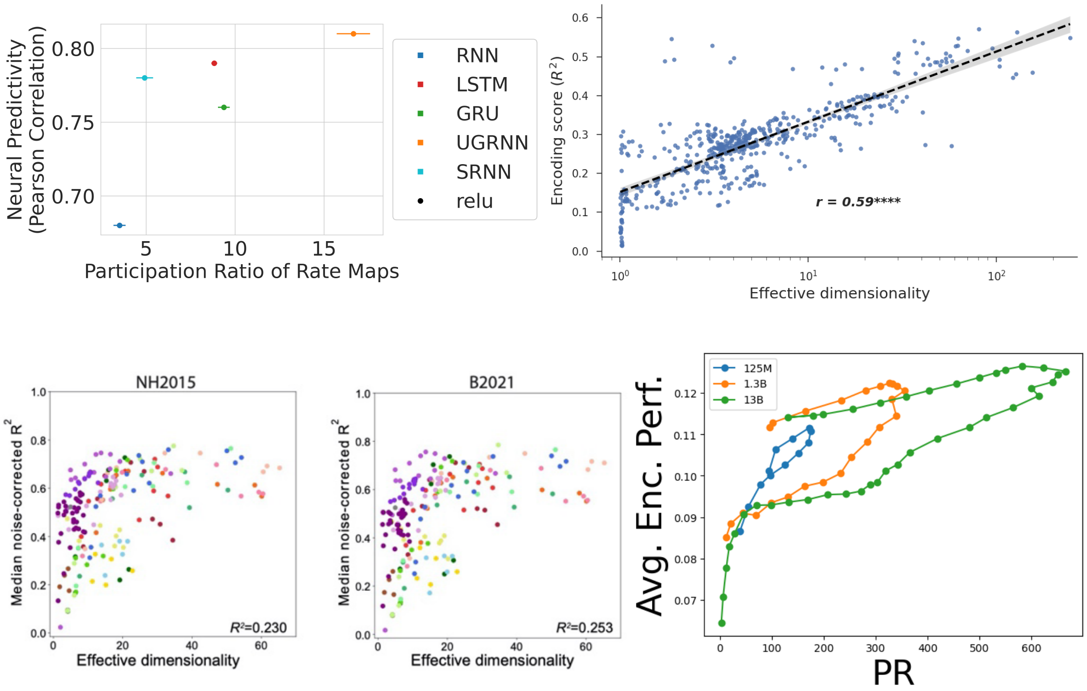

Figure 2: Four independent publications studying four different modalities and brain circuits in three different species found a consistent quantitative heuristic: Test R^2 is an affine transformation of the log participation ratio (Eqn. 2). *Note the log-X scaling in the top right.* Figures from Schaeffer et al. (2022); Elmoznino and Bonner (2024); Tuckute et al. (2023); Cheng and Antonello (2024). Canatar et al. (2024) later provided a more comprehensive spectral theory of the neural regressions methodology, which demonstrates results like these are attributable to general properties of linear regression.

Appendix D. Future Outlook: Methodological Pluralism

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

One short-term answer: 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 (Schrimpf et al. (2018); Yamins and DiCarlo (2016b)), neural dynamics on top of neural geometry (Ostrow et al., 2023), or using a variety of different metrics that have different biases (Han et al. (2023)). Beyond linear regression, computational neuroscience has introduced a number of other candidates into the literature, including RSA (Kriegeskorte et al. (2008a)), Procrustes (Williams et al. (2021)), CKA (Kornblith et al. (2019)), SVCCA (Raghu et al. (2017)) , and a number of variants of these metrics. These developments are promising, although it is worth noting that any single method alone can fall prey to Goodhart's law. It is also important to note that depending on the scientific question, the type of system feature being compared may change. All of the above metrics only seek to compare geometric features of neural activations. Recently proposed methods, such as Dynamical Similarity Analysis (DSA, Ostrow et al. (2023)) seek to compare different features of the system like dynamical structure. Using more types of comparison, both in terms of metrics **and** data, will mitigate the biases of individual comparisons, making Goodharting more challenging (but still possible) and resulting in more robust conclusions.

In the longer-term, beyond significantly increasing the number of types of comparisons being done, it is worth taking a step back and asking 'what do we mean by neural similarity'? We intentionally did not attempt to propose notions of neural similarity here, for two reasons. Firstly, 4 pages is too short to both critique the neural regressions methodology and propose and justify an alternative. Secondly, the "right" notions are likely (1) highly bespoke to the particular brain circuit and/or behavior being studied, (2) effortful to identify and quantify, (3) contentious in the community. Answering this question will likely warrant an entirely separate paper. To briefly sketch our view, neural similarity is almost certainly a function of the scientific question at hand. In some cases, similarity may be the geometry of neural activations, in which case the above family may be sufficient, provided a battery of metrics are used. In other cases, more care should be taken to define 'similarity' and identify modes of comparison that allow to draw real conclusions about the brain, not our metric.