Continual learning under domain transfer with sparse synaptic bursting

Shawn L. Beaulieu[1] **Jeff Clune**[2] **Nicholas Cheney**[1]

[1]University of Vermont
[2]University of British Columbia

Abstract AI programs with the intelligence, resilience, and autonomy approaching that of biological systems must be capable of learning and retaining new information without arbitrarily frequent re-training. In this paper, we introduce a system that can learn sequentially over previously unseen datasets (ImageNet, CIFAR-100) with little forgetting over time. This is done by controlling the activity of weights in a convolutional neural network in a context-dependent manner using top-down regulation generated by a second feed-forward neural network. We find that our method learns continually under domain transfer to a new dataset with sparse bursts of heavy-tailed activity in weights that are recycled across tasks, rather than by maintaining task-specific modules. Sparse synaptic bursting is found to balance activity and suppression such that new functions can be learned without corrupting extant knowledge, perhaps mirroring the balance of order and disorder in systems poised at the edge of chaos. This behavior emerges during a prior pre-training (or "meta-learning") phase in which regulated synapses are selectively disinhibited, or grown, from an initial state of uniform suppression through prediction error minimization.

Catastrophic forgetting is the phenomenon wherein an artificial neural network trained over a sequence of inputs loses its ability to perform a function it acquired earlier in the sequence as new information is learned (French, 1999; Kirkpatrick et al., 2017). Forgetting is typically overcome by scrambling the temporal structure of learning with large batches of randomly ordered inputs that are stored at the programmer's discretion (IID training). However, it is unrealistic for most real-world scenarios to assume that all relevant data can be obtained prior to model deployment. Additionally, we may encounter constraints on time, storage, and compute that prevent us from scaling traditional solutions to catastrophic forgetting (Kirkpatrick et al., 2017; Kudithipudi et al., 2022; Hayes et al., 2019, 2021).

In this paper we present an algorithm for continual learning in a convolutional classifier, whose synapses are regulated in a context-dependent manner by a second neural network, which itself undergoes continual change. Prior to learning continually, synaptic regulation is first pre-trained via meta-learning (Finn et al., 2017; Javed and White, 2019). We find that this approach is most effective when meta-learning from an initial state of uniform suppression, where regulatory outputs are initialized to permit very little synaptic activity. For the system to acquire useful functions, the regulator must learn to *disinhibit* those synapses whose activation helps to identify relevant features of input. After meta-learning to grow sensors in the classifier, both the regulator and classifier are greedily updated over long sequences of input sampled from a previously unseen dataset (*domain transfer*). Under domain transfer from Omniglot (Lake et al., 2015) to ImageNet (Russakovsky et al., 2015) and CIFAR-100 (Krizhevsky and Hinton, 2009), we find that the regulator avoids forgetting by inducing sparse bursts of heavy-tailed activity in synapses that are recycled across tasks, rather than by maintaining task-specific modules, as in prior work on catastrophic forgetting (Kirkpatrick et al., 2017; Javed and White, 2019; Beaulieu et al., 2020; Ellefsen et al., 2015; Masse et al., 2018). Sparse synaptic bursting allows the regulator to control the *amount* of activity in

the classifier rather than its task-modular *location*, such that prior knowledge is protected without blocking adaptation to new inputs.

1 Tuning synapses via allostatic regulation.

The method we introduce [1] consists of two neural networks: (*i*) a convolutional classifier network, consisting of three convolutional layers and a linear class prediction layer; and (*ii*) a feed-forward regulatory network that takes in the same input image as the classifier network, but generates a real-valued mask on every weight of the classifier (sigmoidal output in the range [0, 1]). This mask affects both forward propogation of inputs and the backpropagation of error, enabling context-dependent selective plasticity (Masse et al., 2018; Beaulieu et al., 2020). Thus, our model has kinship with fast weight memory systems (Ashby, 1960; Schmidhuber, 1992; Ba et al., 2016; Tsuda et al., 2020) for which information is dynamically routed through a quickly evolving network by a supervising controller that adapts more slowly. However, we do not explicitly encode operations for maintaining a short-term memory, except insofar as previously learned weights are preserved and recruited by the regulator. Our model therefore belongs to the class of continual learning systems that learn how to learn according to empirical success (Stanley and Miikkulainen, 2002; Zenke et al., 2017; Finn et al., 2017; Pham et al., 2018) rather than manually designed rules (Isele and Cosgun, 2018; Fernando et al., 2017; French, 1992; Ellefsen et al., 2015; Kirkpatrick et al., 2017; LI et al., 2018; Zhang et al., 2017).

One of the chief contributions of this work is to demonstrate the relative advantage of systems that are forced to grow over time (Kauffman, 1986; Watts and Strogatz, 1998; Sporns et al., 2004; Bongard, 2011; Bernatskiy and Bongard, 2015; Neniskyte and Gross, 2017; Raghavan and Thomson, 2019) by meta-learning how to activate uniformly suppressed weights, against those that are forced to sculpt away components (LeCun et al., 1990; Tanaka et al., 2020; Evci et al., 2020; Bengio et al., 2013; Zhou et al., 2019; Wortsman et al., 2020) by meta-learning how to suppress uniformly active weights (Section 4).

2 Method for meta-learned sensor growth.

In selecting for the ability to learn without forgetting, we first pre-train TSAR (Tuning Synapses via Allostatic Regulation) using the Online-aware Meta-Learning (OML) algorithm (Javed and White, 2019). This involves an inner loop of sequential learning over training images sampled from a single Omniglot class (Lake et al., 2015), followed by an outer loop update on a batch of images sampled from K other random classes. The inner loop updates the classifier only, while the outer loop updates the initial weights of both the regulator and classifier, which are then used to start the next inner-loop. As a result, our model must learn how to learn over a sequence of Omniglot images, such that the corresponding updates do not corrupt the ability to classify previously seen Omniglot images. In addition to the baseline setting that uses 963 Omniglot classes ("data rich"), we also present a "data scarce" setting containing less than 3% of this data, or just 25 randomly sampled Omniglot classes.

Finally, we obtain regulators that "grow" connections in the classifier by varying the strength of the initial bias on regulation prior to meta-learning. In this way, we can cause regulation to be more or less *permissive* (≈ 1, unmasked) or *suppressant* (≈ 0, fully masked). We perform a sweep over initial biases covering the range of integers from $[-12, 12]$, and find that optimal performance is obtained with a bias of -8 (Fig. 1, C). This initialization we refer to as the canonical "Grow" condition, which we set against the canonical "Sculpt" condition, having the standard initial bias of 0. Sculpt begins meta-learning with weights in the classifier being uniformly active, and which may or may not be sculpted or pruned away over the course of meta-learning. For more detailed information, see Materials and Methods.

[1] Code available at `https://github.com/shawnbeaulieu/TSAR`

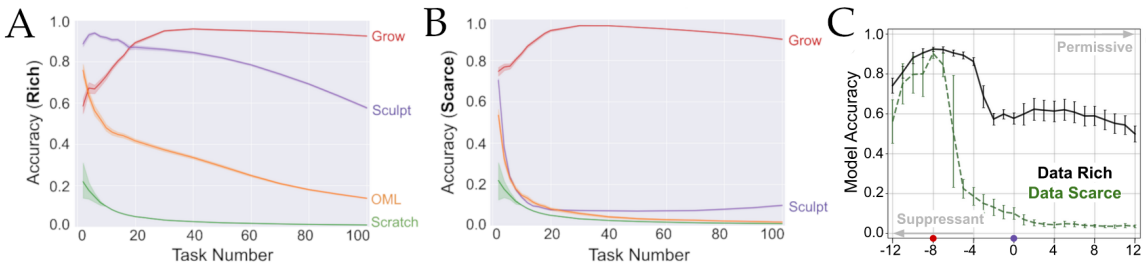

Figure 1: Continual learning under domain transfer to Imagenet after meta-learning on 100% (**A**) or less than 3% of Omniglot classes (**B**). The Grow treatment (initial regulatory bias=-8) outperforms the Sculpt treatment (bias=0) on the domain transfer task (**C**) and this difference is exacerbated with data limited meta-learning.

3 Performance under domain transfer.

Unlike prior work (Javed and White, 2019; Beaulieu et al., 2020) the ability to avoid catastrophic forgetting is evaluated under *domain transfer* to ImageNet (Russakovsky et al., 2015) and CIFAR-100 (Krizhevsky and Hinton, 2009) (see Appendix 5B)) after meta-learning on Omniglot (Lake et al., 2015). Doing this is analogous to transferring a robot from simulation to reality (Mouret and Chatzilygeroudis, 2017) or the introducing a novel stressor in biology (Emmons-Bell et al., 2019). Under domain transfer, we train on a random sequence of 100 previously unseen tasks, where each task contains 30 randomly sampled images from a pool of 600 possible images for a given class. Each run then consists in a total of 3,000 sequential gradient updates. Accuracy is measured as the degree to which prior inputs can be recalled. This captures the phenomenon of forgetting by quantifying how much of what was actually learned is remembered over time (Javed and White, 2019; Schmidhuber, 1992)(Fig. 1 A,B). We find that Grow (92.2%±1.1%) learns sequentially under domain transfer without significant loss in performance as new tasks are encountered—while competing methods, including different regulatory initializations, forget catastrophically (Sculpt: 57.5%±2.8%). Under the data scarce condition, performance for Grow modestly declines relative to the data rich condition, while Sculpt, OML, and Scratch forget catastrophically (Fig.1B). Similar results obtain for domain transfer to CIFAR-100 with minor variations in accuracy (Appendix 7A). To calibrate our understanding of what it means to perform well on this problem, we compare against the meta-learned algorithm *OML* (Javed and White, 2019), as well as our classification network trained from scratch without meta-learning or regulation (*Scratch*) both of which perform poorly, indicating the problem is non-trivial (OML: 13.87%±0.729%, Scratch: 0.971%±0.10%).

4 Task-specific modularity.

All analysis henceforth concerns regulation of the third convolutional layer of the classification network (C3) unless otherwise stated. Qualitatively similar results were obtained for the regulation of other convolutional layers (Appendix 9-17)). Past efforts to solve catastrophic forgetting have relied on task-specific modules for conditionally modifying disjoint sets of weights (Ellefsen et al., 2015; Kirkpatrick et al., 2017; Javed and White, 2019; Beaulieu et al., 2020). To determine whether regulation in TSAR is calibrated for such task-specific modularity, we reason that task-specific modules should be characterized by heightened synaptic activity within a given task but reduced activity for all other tasks. Task-Specific activity is computed as the average regulatory signal for a given synapse over all instances of a given task, C. Task-Agnostic activity is computed as the average Task-Specific activity over all tasks other than C. Task-specific modularity, or its absence, can then be visualized by plotting ranked Task-Specific activity against ranked Task-Agnostic activity for each task under domain transfer (Fig.2A). We find that the synapses which are the most

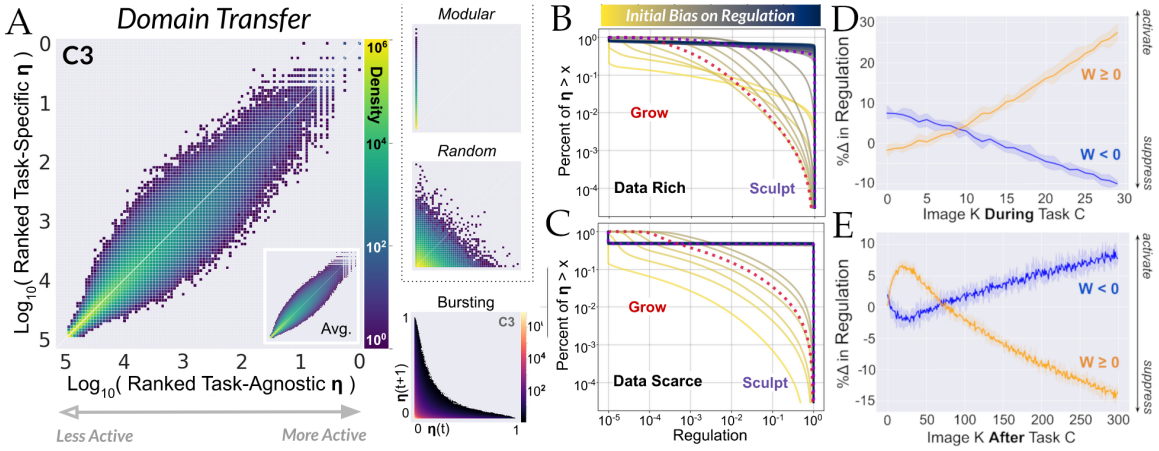

Figure 2: Regulation of synapses under domain transfer does not show evidence of task-specific modularity (see hypothetical "Modular") but is largely task-agnostic (**A**). Synapses are likely to be suppressed even if they were highly active on the previous image from the same class ("Bursting"; see Appendix 2C). Increasingly scale-free cumulative distribution functions for regulation obtain under domain transfer for increasingly successful regulators after data-scarce (**C**) and, especially, data-rich (**B**) meta-learning. Within each task, positive weights leading into the correct output node become increasingly active and thus increasingly plastic (**D**), and then transition over the next 10 tasks to become increasingly dormant and thus increasingly protected from future weight changes that may lead to forgetting (**E**).

active on a given task are among the most active over all tasks. Thus, we do not see evidence for task-specific weights or modules in the output of the regulatory network.

In the absence of task-specific modularity, we observe that synapses are conditionally recruited for task-agnostic bursting, in which the same set of weights over all tasks are together activated and suppressed in a alternating fashion, even within tasks (Appendix 2C). This produces waves of *enhanced* and *diminished* processing of inputs, in which some images strongly use and modify many weights of the network, while other images weakly use and modify very little of the network. Analysis shows that better Grow models exhibit regulation under domain transfer that is increasingly *scale-free* (Stumpf and Porter, 2012; Kauffman and Johnsen, 1991; Kauffman et al., 2004; Shew et al., 2011; Bertschinger and Natschläger, 2004; Rämö et al., 2007; Sporns et al., 2004; Mora and Bialek, 2011). Here scale-free regulation means that inputs that use (unmask) an exponentially larger amount of the network are exponentially more rare. We therefore claim that performant regulation is calibrated to control the *amount* of sensory processing in the classifier, rather than the task-modular *location* where such processing occurs. This is accomplished by maintaining an apposite balance of activity and suppression—and, thus, an apposite balance of plasticity and stability. Comparatively poor regulation, exhibited by regulators initialized to be more permissive (i.e. Sculpt), does not show this behavior. While many models of neural processes have suggested the benefits of burst-dependent plasticity (Lisman, 1997; Hahn et al., 2019; Park et al., 2019; Payeur et al., 2021) bursting itself it is not yet an established component in theories of how to avoid catastrophic forgetting, as we propose here. However, we note a possible connection to work in biological systems on rhythmic sampling (Fiebelkorn et al., 2013; Fiebelkorn and Kastner, 2019) and distributed robustness (Wagner, 2005; Anderson, 2014; Bruineberg and Rietveld, 2019; Maass and Markram, 2004; Palmigiano et al., 2017; Stringer et al., 2019) as our analysis reveals a positive bias in the degree of correlation among synapse pairs, but an absence of strong synchronicity or asynchronicity (Appendix 4C). This means that the same recurring set of synapses is being recruited to burst, but that burst composition is heterogeneous. Such weak synchronization may

reflect the creation and storage of robustly distributed memories (Wagner, 2005), as no single weight, or set of weights, encodes a unique function (Anderson, 2014; Bruineberg and Rietveld, 2019; Maass and Markram, 2004; Palmigiano et al., 2017; Stringer et al., 2019) but instead performs multiple functions depending on the context in which it is active.

If upstream sensory processing is characterized by task-agnostic synaptic bursting, how is it that correct class predictions are acquired and maintained by the network? First, we find that positive and negative weights in the class prediction (CP) layer are differentially correlated with upstream activity, such that negative weights are synchronized with upstream bursting, but positive weights are *inversely* synchronized (Appendix 3E). This manner of coordination across layers suggests that different modes of behavior are associated with enhanced and diminished processing (Golding et al., 2002; Fiebelkorn et al., 2013; Fiebelkorn and Kastner, 2019). Next, we find that positive weights in CP for a given class node become relatively active when that class is encountered, and relatively suppressed when the task changes. The inverse process occurs for the negative weights of the same class node. This both enables and protects task-specific predictions (Fig.2D,E). However, we find that optimizing the regulation of CP under domain transfer is not, on its own, sufficient to achieve good performance: when regulatory output layers that control convolutional weights are fixed, but regulation of CP is made trainable, performance drops relative to the condition where regulatory output for all layers is trainable(Appendix 3C).

5 Discussion.

In the preceding sections we presented an algorithm for regulating the synapses of a convolutional neural network for continual learning over sequences of previously unseen datasets. This occurs by meta-learning to disinhibit synapses initialized to a state of uniform suppression.

Under domain transfer, we found that regulation does not elicit task-specific modularity, but instead induces sparse bursts of activity in weights that are recycled across tasks. Sparse synaptic bursting is found to be heavy-tailed in both its total output (Fig.2B,C) and in the number of synapses it causes to burst image-to-image (Appendix 2D). Balancing contrasting states, like activity/suppression and order/disorder, has previously been implicated in the discovery of optimal communication and memory in simulated networks (Kauffman et al., 2004; Shew et al., 2011; Bertschinger and Natschläger, 2004; Rämö et al., 2007), and is believed to play a role in the adaptive behavior of living systems (Sporns et al., 2004; Mora and Bialek, 2011). We hypothesize that the initial biases on meta-learned regulation that achieve domain transfer corresponds to a critical range in which regulation learns to balance activity and suppression, thereby enabling adaptation to new inputs without corrupting extant functions. This resolves the core dilemma of continual learning: too much change causes forgetting, while too little change induces brittleness.

6 Broader Impact and Limitations.

The ethical implications involved in the development of machine learning models include things like their potential for biased decision-making, but also extend beyond any specific application. These algorithms—like any object or process that is, or poised to become, socially embedded—have the capacity to influence our perception and understanding of the world—shaping not only our social interactions but also our decision-making processes. Insofar as they become constitutive elements of social life, they have the potential to affect our judgments and preferences, much like any other social practice or institution. For our purposes, the essential question is whether the work we have produced poses unique dangers that aren't associated with other work in this field. Given that the central contributions of our paper include mitigating catastrophic failure modes in neural networks, reducing the volume of data needed to achieve this effect, and offering a means to intervene on the typical operations of a classifier, we do not believe this to be the case. After careful reflection, we have determined that this work presents no notable negative impacts to society or the environment.

References

Anderson, M. (2014). *After Phrenology: Neural Reuse and the Interactive Brain.* MIT Press.

Ashby, W. R. (1960). *Design for a brain: The origin of adaptive behaviour (2nd ed. rev.).* Chapman & Hall.

Ba, J., Hinton, G. E., Mnih, V., Leibo, J. Z., and Ionescu, C. (2016). Using fast weights to attend to the recent past. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Beaulieu, S., Frati, L., Miconi, T., Lehman, J., Stanley, K. O., Clune, J., and Cheney, N. (2020). Learning to continually learn. In *Proceedings of the 24th European Conference on Artificial Intelligence*.

Bengio, Y., Léonard, N., and Courville, A. C. (2013). Estimating or propagating gradients through stochastic neurons for conditional computation. *ArXiv*, abs/1308.3432.

Bernatskiy, A. and Bongard, J. C. (2015). Exploiting the relationship between structural modularity and sparsity for faster network evolution. In *Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation*. ACM.

Bertschinger, N. and Natschläger, T. (2004). Real-time computation at the edge of chaos in recurrent neural networks. *Neural Computation*, 16(7):1413–1436.

Bongard, J. (2011). Morphological change in machines accelerates the evolution of robust behavior. *Proceedings of the National Academy of Sciences*, 108(4):1234–1239.

Bruineberg, J. and Rietveld, E. (2019). What's inside your head once you've figured out what your head's inside of. *Ecological Psychology*, 31(3):198–217.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Ellefsen, K. O., Mouret, J.-B., and Clune, J. (2015). Neural modularity helps organisms evolve to learn new skills without forgetting old skills. *PLOS Computational Biology*, 11(4):e1004128.

Emmons-Bell, M., Durant, F., Tung, A., Pietak, A., Miller, K., Kane, A., Martyniuk, C. J., Davidian, D., Morokuma, J., and Levin, M. (2019). Regenerative adaptation to electrochemical perturbation in planaria: A molecular analysis of physiological plasticity. *iScience*, 22:147–165.

Evci, U., Ioannou, Y. A., Keskin, C., and Dauphin, Y. (2020). Gradient flow in sparse neural networks and how lottery tickets win.

Fernando, C., Banarse, D., Blundell, C., Zwols, Y., Ha, D. R., Rusu, A. A., Pritzel, A., and Wierstra, D. (2017). Pathnet: Evolution channels gradient descent in super neural networks. *ArXiv*, abs/1701.08734.

Fiebelkorn, I. C. and Kastner, S. (2019). A rhythmic theory of attention. *Trends in Cognitive Sciences*, 23(2):87–101.

Fiebelkorn, I. C., Saalmann, Y. B., and Kastner, S. (2013). Rhythmic sampling within and between objects despite sustained attention at a cued location. *Current Biology*, 23(24):2553–2558.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia. PMLR.

French, M. R. (1992). Semi-distributed representations and catastrophic forgetting in connectionist networks. *Connection Science*, 4(3-4):365–377.

French, M. R. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135.

Golding, N. L., Staff, N. P., and Spruston, N. (2002). Dendritic spikes as a mechanism for cooperative long-term potentiation. *Nature*, 418(6895):326–331.

Hahn, G., Ponce-Alvarez, A., Deco, G., Aertsen, A., and Kumar, A. (2019). Portraits of communication in neuronal networks. *Nature Reviews Neuroscience*, 20(2):117–127.

Hayes, T. L., Cahill, N. D., and Kanan, C. (2019). Memory efficient experience replay for streaming learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9769–9776. IEEE.

Hayes, T. L., Krishnan, G. P., Bazhenov, M., Siegelmann, H. T., Sejnowski, T. J., and Kanan, C. (2021). Replay in deep learning: Current approaches and missing biological elements. *Neural computation*, 33(11):2908–2950.

Isele, D. and Cosgun, A. (2018). Selective experience replay for lifelong learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Javed, K. and White, M. (2019). Meta-learning representations for continual learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 1820–1830. Curran Associates, Inc.

Kauffman, S., Peterson, C., Samuelsson, B., and Troein, C. (2004). Genetic networks with canalyzing boolean rules are always stable. *Proceedings of the National Academy of Sciences*, 101(49):17102–17107.

Kauffman, S. A. (1986). Autocatalytic sets of proteins. *Journal of Theoretical Biology*, 119(1):1–24.

Kauffman, S. A. and Johnsen, S. (1991). Coevolution to the edge of chaos: Coupled fitness landscapes, poised states, and coevolutionary avalanches. *Journal of Theoretical Biology*, 149(4):467–505.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images.

Kudithipudi, D., Aguilar-Simon, M., Babb, J., Bazhenov, M., Blackiston, D., Bongard, J., Brna, A. P., Chakravarthi Raja, S., Cheney, N., Clune, J., et al. (2022). Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, 4(3):196–210.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

LeCun, Y., Denker, J. S., and Solla, S. A. (1990). Optimal brain damage. In *Advances in Neural Information Processing Systems*, pages 598–605. Morgan Kaufmann.

LI, X., Grandvalet, Y., and Davoine, F. (2018). Explicit inductive bias for transfer learning with convolutional networks. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2825–2834, Stockholmsmässan, Stockholm Sweden. PMLR.

Lisman, J. E. (1997). Bursts as a unit of neural information: making unreliable synapses reliable. *Trends in neurosciences*, 20(1):38–43.

Maass, W. and Markram, H. (2004). On the computational power of circuits of spiking neurons. *Journal of Computer and System Sciences*, 69(4):593–616.

Masse, N. Y., Grant, G. D., and Freedman, D. J. (2018). Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proceedings of the National Academy of Sciences*, 115(44):E10467–E10475.

Mora, T. and Bialek, W. (2011). Are biological systems poised at criticality? *Journal of Statistical Physics*, 144(2):268–302.

Mouret, J.-B. and Chatzilygeroudis, K. (2017). 20 years of reality gap. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. ACM.

Neniskyte, U. and Gross, C. T. (2017). Errant gardeners: glial-cell-dependent synaptic pruning and neurodevelopmental disorders. *Nature Reviews Neuroscience*, 18(11):658–670.

Palmigiano, A., Geisel, T., Wolf, F., and Battaglia, D. (2017). Flexible information routing by transient synchrony. *Nature Neuroscience*, 20(7):1014–1022.

Park, S., Kim, S., Choe, H., and Yoon, S. (2019). Fast and efficient information transmission with burst spikes in deep spiking neural networks. In *Proceedings of the 56th Annual Design Automation Conference 2019*, pages 1–6.

Payeur, A., Guerguiev, J., Zenke, F., Richards, B. A., and Naud, R. (2021). Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. *Nature neuroscience*, 24(7):1010–1019.

Pham, H., Guan, M., Zoph, B., Le, Q., and Dean, J. (2018). Efficient neural architecture search via parameters sharing. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4095–4104, Stockholmsmässan, Stockholm Sweden. PMLR.

Raghavan, G. and Thomson, M. (2019). Neural networks grown and self-organized by noise. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Rämö, P., Kauffman, S., Kesseli, J., and Yli-Harja, O. (2007). Measures for information propagation in boolean networks. *Physica D: Nonlinear Phenomena*, 227(1):100–104.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Schmidhuber, J. (1992). Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139.

Shew, W. L., Yang, H., Yu, S., Roy, R., and Plenz, D. (2011). Information capacity and transmission are maximized in balanced cortical networks with neuronal avalanches. *Journal of Neuroscience*, 31(1):55–63.

Sporns, O., Chialvo, D. R., Kaiser, M., and Hilgetag, C. C. (2004). Organization, development and function of complex brain networks. *Trends in Cognitive Sciences*, 8(9):418–425.

Stanley, K. O. and Miikkulainen, R. (2002). Evolving neural networks through augmenting topologies. *Evolutionary Computation*, 10(2):99–127.

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., and Harris, K. D. (2019). High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365.

Stumpf, M. P. H. and Porter, M. A. (2012). Critical truths about power laws. *Science*, 335(6069):665–666.

Tanaka, H., Kunin, D., Yamins, D. L. K., and Ganguli, S. (2020). Pruning neural networks without any data by iteratively conserving synaptic flow. *CoRR*, abs/2006.05467.

Tsuda, B., Tye, K. M., Siegelmann, H. T., and Sejnowski, T. J. (2020). A modeling framework for adaptive lifelong learning with transfer and savings through gating in the prefrontal cortex. *Proceedings of the National Academy of Sciences*, 117(47):29872–29882.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Wagner, A. (2005). Distributed robustness versus redundancy as causes of mutational robustness. *BioEssays*, 27(2):176–188.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Wortsman, M., Ramanujan, V., Liu, R., Kembhavi, A., Rastegari, M., Yosinski, J., and Farhadi, A. (2020). Supermasks in superposition. *Advances in Neural Information Processing Systems*, 33:15173–15184.

Zenke, F., Poole, B., and Ganguli, S. (2017). Continual learning through synaptic intelligence. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3987–3995, International Convention Centre, Sydney, Australia. PMLR.

Zhang, J., Bargal, S. A., Lin, Z., Brandt, J., Shen, X., and Sclaroff, S. (2017). Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102.

Zhou, H., Lan, J., Liu, R., and Yosinski, J. (2019). Deconstructing lottery tickets: Zeros, signs, and the supermask. *Advances in neural information processing systems*, 32.

Meta-Learning Algorithm

We use the model-agnostic Online Meta-Learning algorithm (Javed and White, 2019) for learning how to learn over the background partition of the Omniglot image dataset (Lake et al., 2015). This dataset contains 963 hand-written character classes taken from various alphabets. Each character class contains 20 unique instances. All images are resized to 3x28x28 (channels, height, width). A single iteration of meta-learning consists of (i) an inner loop of sequential learning over the training images for a single character class $c_m \sim \mathcal{C}$; followed by (ii) an outer loop that computes accuracy on the inner loop class and a random batch of 64 validation images sampled from the combined set of all other meta-learning classes.

Prediction error on the outer loop following inner loop learning for network Θ_K^m is back-propagated through the entire model to update the weights used at the beginning of the inner loop sequence, Θ_0^m. These new initial weights are then carried over into the next meta-iteration, $\Theta_0^{m+1} \propto \Theta_0^m$ + update. Learning within each inner loop is discarded: only the initial weights are optimized over the outer loop as per (Finn et al., 2017). Successful meta-learners will have learned to learn over the inner loop without corrupting the classification of outer loop images, thus optimizing for the ability to learn without forgetting or interference (Javed and White, 2019).

Treatments presented in this paper differ both in their architecture, and in which layers are trainable during inner loop meta-learning. Because of this, we make no claims as to the inherent superiority of one model over another. Each treatment was meta-learned for 25,000 meta-iterations on a single NVIDIA Tesla V100 GPU for 25 independent runs having different random seeds. For more details regarding the meta-learning protocol see (Javed and White, 2019; Beaulieu et al., 2020; Finn et al., 2017).

Data Rich/Scarce Meta-Learning

Data rich meta-learning uses the full background set of Omniglot images (963 character classes) for the OML protocol. The data scarce condition instead uses less than 3% of the meta-learning data used in the data rich condition (25 randomly sampled character classes).

Domain Transfer

After executing the meta-learning protocol, each model is trained over a sequence of 100 classes from a previously unseen dataset (ImageNet or CIFAR-100). Classes contain 600 total images, from which 30 are randomly sampled without replacement for each class. This results in 3000 (100*30) sequential iterations of gradient descent. All images are resized to 3x28x28 (channels, height, width). Every run uses a different random seed, a randomized class order, and a random batch of 30 images per class. We applied a grid search over learning rates, and selected the highest performing setting over all subsequent runs.

OML

See (Javed and White, 2019; Beaulieu et al., 2020) for more details regarding theoretical motivation and network architecture. No changes were made to the meta-learning protocol, except to implement a post-publication correction for computing second-order gradients. The OML treatment, which is distinct from the OML *protocol*, consists of two neural networks: a representation learning network (RLN) composed of five convolutional layers, and a prediction learning network (PLN) composed of two linear layers that takes as input the output of the RLN. During meta-learning, the RLN is updated over the outer loop only, while the PLN is updated in both the inner and outer loops. Under domain transfer, the RLN is fixed over the course of training, while the PLN is fully trainable. Prior to domain transfer, the class prediction layer (CP) of the PLN is randomly reset as per (Javed and White, 2019).

TSAR

TSAR consists of two neural networks: a regulatory network, and a classifier network. The regulatory network consists of (*i*) a perception module; and (*ii*) a regulatory output layer. The regulatory perception module contains 3 convolutional layers, each containing 192 channels (window size=(3,3), stride=1, padding=0). The output of each convolutional layer is followed by instance normalization and a ReLU non-linear activation function. Max-pooling layers (stride=2, kernel size=2) are placed after the first two convolutional layers of the perception module to create an encoding layer of size 1728 from which all regulatory output is generated.

The regulatory output layer consists of four weight matrices, each of which produces a set of regulatory outputs that govern a specific layer in the classification network. A sigmoid activation

function is used on regulatory output before modulating synapses in the classifier via multiplicative gating ($\eta^{\ell} \odot W^{\ell}$, for layer ℓ in the classifier).

The classification network is made up of three convolutional layers (112 channels, window size=(3,3), stride=1, padding=0) and a linear output, or class prediction, layer (CP). Max-pooling layers (stride=2, kernel size=2) are placed after each convolutional layer and are followed by instance normalization and a ReLU non-linearity. Class predictions are computed with a softmax function applied to the raw output of the modulated CP layer. During meta-learning, the regulatory network is fixed during inner loop learning, but the prediction network is trainable. All parameters receive updates in the outer loop.

Under domain transfer, all layers in the prediction network are trainable. The regulatory output layer is also trainable. Only the convolutional layers, or perception module, of the regulatory network are fixed. As with OML, the CP layer, and the regulatory output that flows to this layer, is reset before domain transfer.

Grow and Sculpt

For *Grow*, all biases in the regulatory output layer are initialized to a value of -8 before executing the meta-learning protocol. Due to the sigmoid activation, this results in initially high levels of synaptic suppression. For *Sculpt*, all biases in the regulatory output layer are initialized to the standard bias of 0 before executing the meta-learning protocol. Thus, all things being equal, synapses in the classification network of Sculpt begin meta-learning with their functional values approximately halved (sigmoid(0.0)=0.5). Under domain transfer, regulatory output to CP is reset with a bias of -2. This value was obtained using a grid search over CP biases and learning rates on domain transfer accuracy for all TSAR models.

Scratch

This treatment uses the classification network architecture described for TSAR, but with regulation disabled. Scratch is subject to domain transfer starting from a random initialization with no meta-learning.

No Regulation

This treatment uses the classification network architecture described for TSAR, but with regulation disabled. No Regulation is meta-learned on Omniglot using the convolutional layers as the RLN and the class prediction layer as the PLN. Under domain transfer, the PLN is re-initialized and trained over domain transfer sequences, while the RLN is fixed.

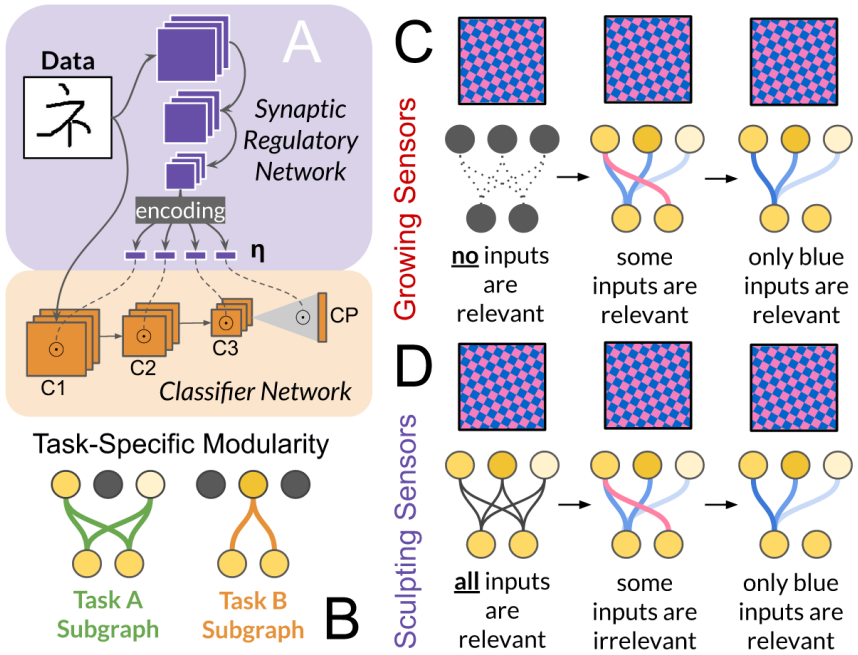

Figure 3: (**A**) Architecture for Tuning Synapses via Allostatic Regulation ("TSAR", Materials and Methods). (**B**) Task-specific modularity has traditionally be used to overcome catastrophic forgetting. Disjoint subgraphs, or "modules", prevent prediction interference and unwanted weight changes (**C**) By initializing meta-learned regulation to be highly suppressant ("Grow", Materials and Methods) we establish a prior on regulation such that no input features are relevant to the recruitment of synapses in the classifier. For performant behavior to emerge, the regulator must learn to disinhibit or "grow" synapses which minimize prediction error. Conversely, (**D**) if regulation is initialized to be highly permissive ("Sculpt", Materials and Methods) then all input features initially are specified as relevant.

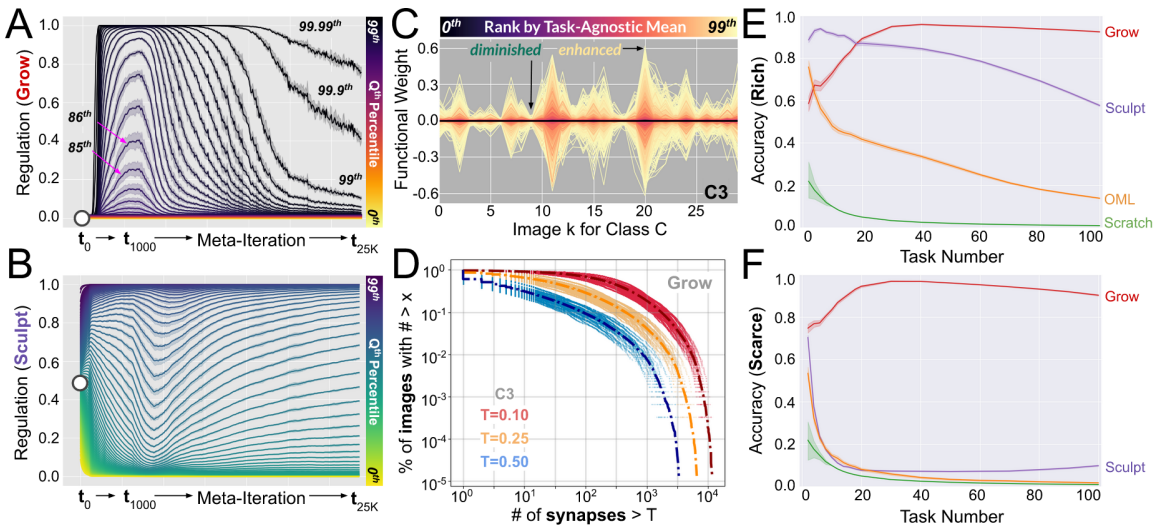

Figure 4: Regulation of layer C3 during meta-learning of Grow (**A**) and Prune (**B**). We report the Q^{th} percentile of regulation across 250 randomly sampled meta-learning classes. Confidence intervals are computed across runs (lower bound=20^{th} percentile; upper bound=80^{th} percentile). (**C**) Randomly sampled window of synaptic activity (post-masking) under domain transfer for Grow. (**D**) Complementary cumulative distribution function (CCDF) for the percent of images in ImageNet that cause a given number of synapses to receive regulation above a threshold (T). We report individual runs (dot) and the mean across runs (dash-dot). Continual learning under domain transfer to Imagenet after meta-learning on 100% (**A**) or less than 3% of Omniglot classes (**B**). The Grow treatment (initial regulatory bias=-8) outperforms the Sculpt treatment (bias=0) on the domain transfer task (**C**) and this difference is exacerbated with data limited meta-learning.

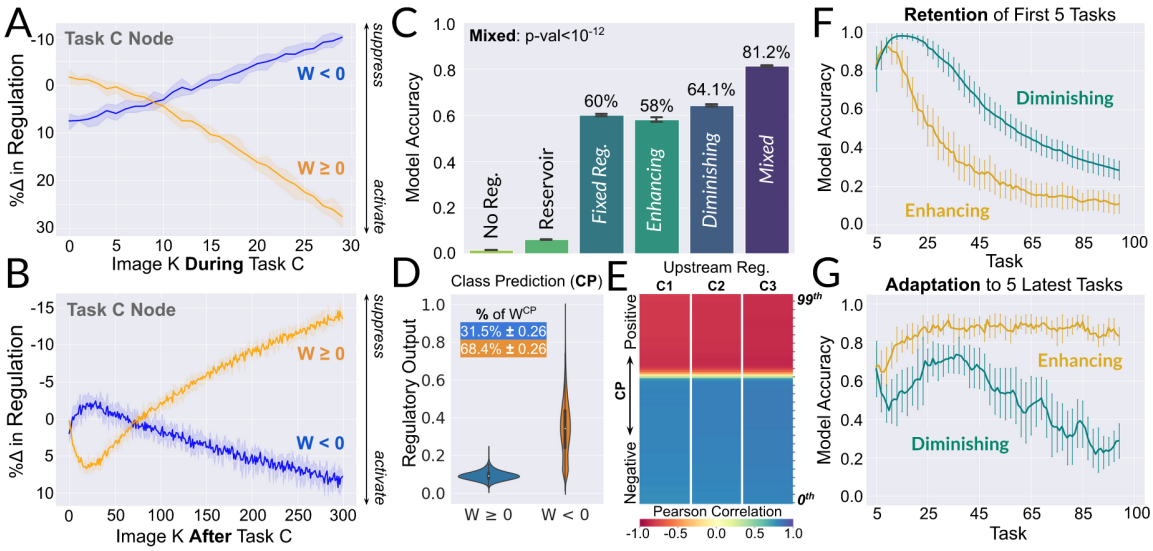

Figure 5: Within each task, positive weights leading into the correct output node become increasingly active and thus increasingly plastic (**A**), and then transition over the next 10 tasks to become increasingly dormant and thus increasingly protected from future weight changes that may lead to forgetting (**B**). (**C**) For each class in ImageNet, "enhancing" images are defined to be the top-30 images for that class ranked by mean regulatory output, while "diminishing" images are defined to be the bottom-30 images for that class ranked by mean regulatory output. Training over a sequence of 100 classes populated exclusively by "enhancing" or "diminishing" images results in significantly lower performance than training over 100 consecutive classes with randomly ordered but even mixtures of enhancing and diminishing images. *Reservoir* refers to the treatment which uses untrained random regulation. *No Reg.* refers to the treatment where the classifier of TSAR is meta-learned in the normal way without regulation. (**D**) In the class prediction layer (CP) positive weights receive regulation that is more suppressant and lower variance, while negative weights receive regulation that is more permissive and high variance. (**E**) Positive weights in CP are negatively correlated with upstream regulation, such that when upstream spiking occurs, positive weights are suppressed. Conversely, negative weights in CP are positively correlated with upstream regulation. Pearson correlation is computed with respect to first differences for (i) *upstream activity*, defined as the mean regulation for the top 10% of weights ranked by regulation; and (ii) the mean regulation governing weights in CP ranked by weight value and grouped centile. The top row of then reports the Pearson correlation between the mean regulation going to the top-10% of weights in C1, C2, or C3 and the mean regulation going to the 99^{th} percentile of weights in CP. We plot the mean Pearson correlation per centile across all 25 models. Diminished processing is found to improve performance *retention* (**F**) while enhanced processing is found to facilitate adaptation to new inputs (**G**). Error bars report standard deviation across all 25 models.

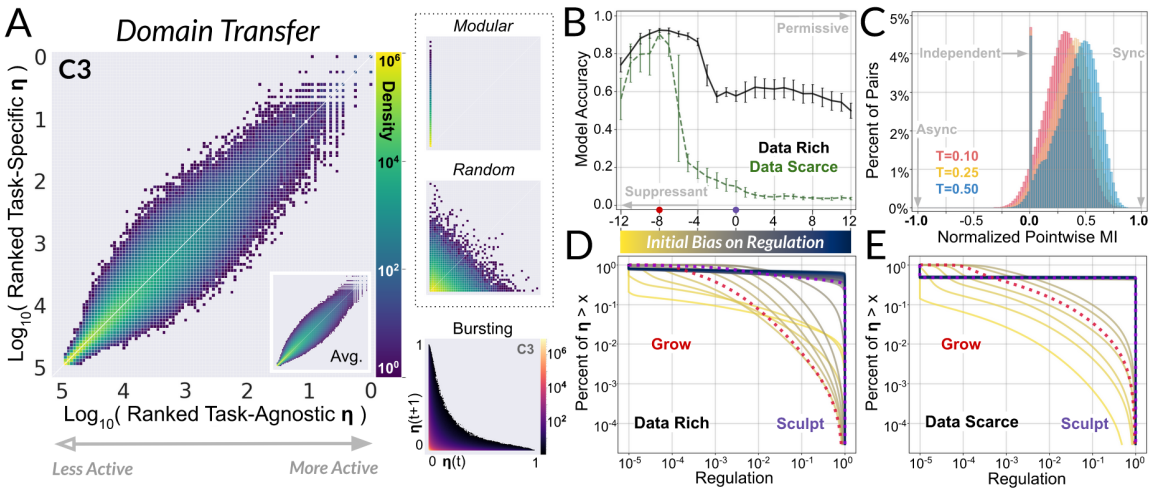

Figure 6: Regulation of synapses under domain transfer does not show evidence of task-specific modularity (see hypothetical "Modular") but is largely task-agnostic (**A**). Synapses are likely to be suppressed even if they were highly active on the previous image from the same class ("Bursting"). (**B**) Performance under domain transfer to ImageNet for the range of initial biases considered. Standard deviation is reported for error bars. (**C**) Normalized pointwise mutual information (Church and Hanks, 1990) for each pair of synapses receiving regulation above the given threshold for more than 1% of images. We find a positive bias in the degree of correlation in synapse pairs, but an absence of strong synchronicity or asynchronicity. This indicates that the coalition of weights used for a given image is relatively heterogeneous with respect to those used for other images. Increasingly scale-free cumulative distribution functions for regulation obtain under domain transfer for increasingly successful regulators after data-scarce (**E**) and, especially, data-rich (**D**) meta-learning. See Appendix 1 for domain transfer to CIFAR-100.

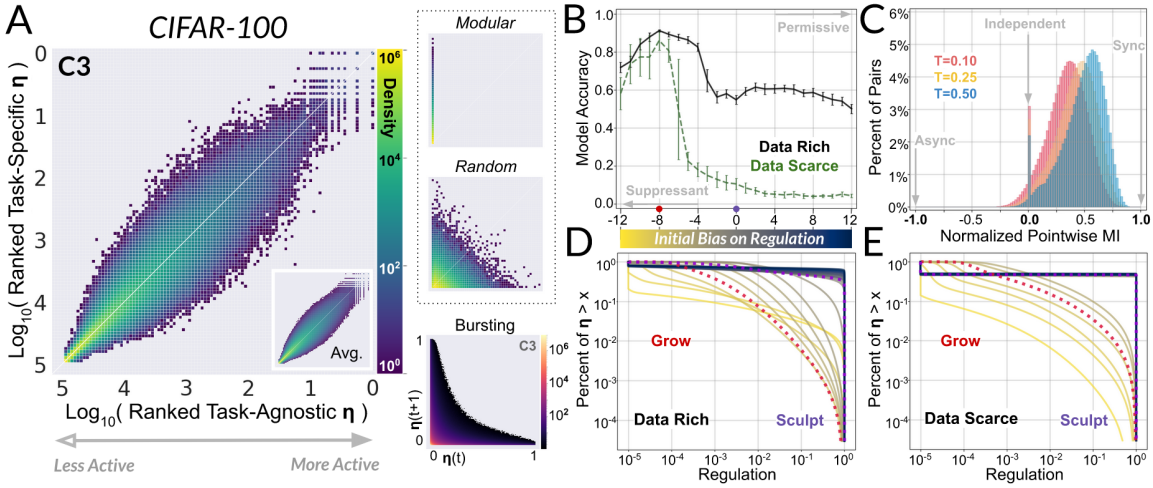

Figure 7: Domain transfer to CIFAR-100. Regulation of synapses under domain transfer does not show evidence of task-specific modularity (see hypothetical "Modular") but is largely task-agnostic (**A**). Synapses are likely to be suppressed even if they were highly active on the previous image from the same class ("Bursting"). (**B**) Performance under domain transfer to CIFAR-100 for the range of initial biases considered. Standard deviation is reported for error bars. (**C**) Normalized pointwise mutual information (Church and Hanks, 1990) for each pair of synapses receiving regulation above the given threshold for more than 1% of images. We find a positive bias in the degree of correlation in synapse pairs, but an absence of strong synchronicity or asynchronicity. This indicates that the coalition of weights used for a given image is relatively heterogeneous with respect to those used for other images. Increasingly scale-free cumulative distribution functions for regulation obtain under domain transfer for increasingly successful regulators after data-scarce (**E**) and, especially, data-rich (**D**) meta-learning.

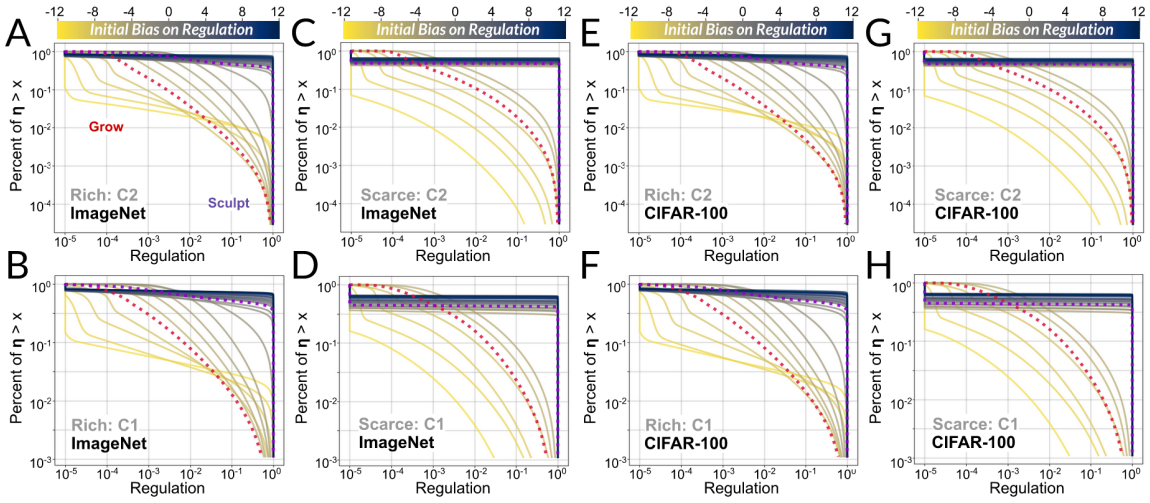

Figure 8: **A-H** For data rich and data scarce meta-learning, regulators that are increasingly successful under domain transfer exhibit regulatory distributions that are increasingly linear (scale-free) in the logarithmic CCDF (Stumpf and Porter, 2012). This trend holds across several orders of magnitude in all convolutional layers of the Grow condition. However, our results do not hinge on whether regulation is precisely power-law distributed, or merely log-normally distributed. These results shows that Grow has obtained a balance of activity and suppression that is analogous to the balance of order and disorder in critical systems at the edge of chaos (Kauffman and Johnsen, 1991). We present CCDF plots for the highest performing runs for each model.

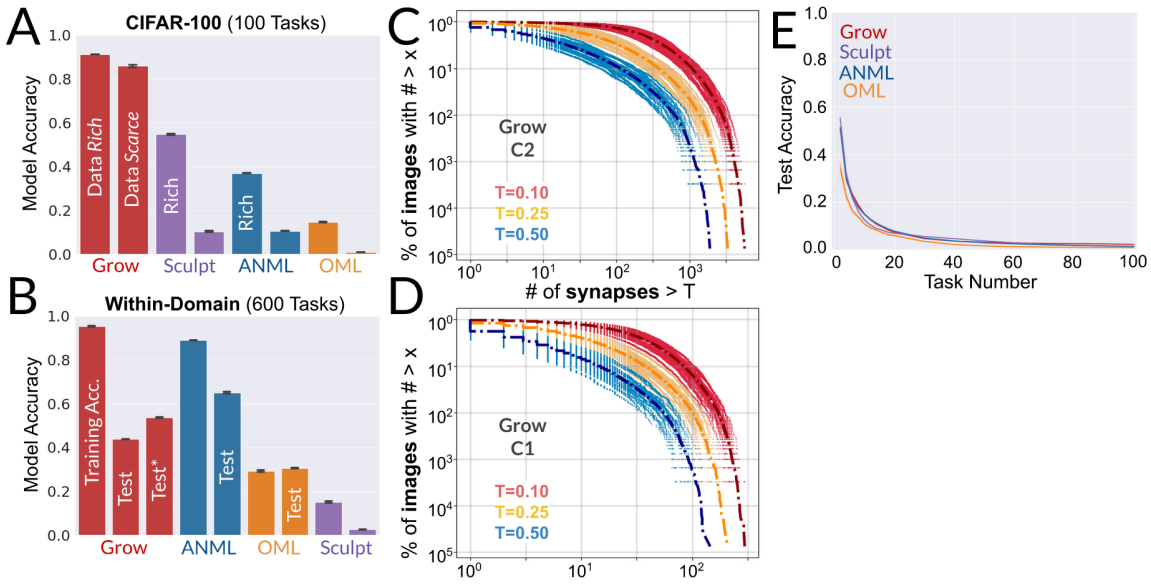

Figure 9: (**A**) Training accuracy under domain transfer to CIFAR-100. **Rich**: Grow=91.1% ±0.56% (SD), Sculpt=54.7% ±2.4%, ANML=37% ±1.70%, OML=14.6% ±2.6%. **Scarce**: Grow=85.9% ±5.12%, Sculpt=10.4% ±2.9%, ANML=10.57% ±0.73%, OML=1% ±0.10%. Accuracy for 600 previously unseen tasks learned sequentially from Omniglot. **Training**: Grow=95.2% ±1.4%, Sculpt=15.1% ±2.23%, ANML=88.9% ±0.59%, OML=29.3% ±2.23%. **Validation**: Grow=43.8.2% ±0.93%, Sculpt=2.6% ±0.28%, ANML=65.06% ±2.88%, OML=30.6% ±1.04%. (**B**) Test accuracy on held-out Omniglot images (600 classes; 15 training images/5 test images per class). Although our method under-performs ANML, only the final layer of the classification network in ANML is trainable. ANML also uses a different network architecture than TSAR, and so is not directly comparable. When the regulatory output layers that control C1, C2, and C3 in Grow are fixed, but the weights they regulate are trainable, test accuracy ("Grow", 3rd bar) increases to 53.75% ±1.45%. This final version of Grow corresponds to a relative drop of 15.5% compared to the IID setting (epochs=3: 63.5%). (**C**, **D**) Complementary cumulative distribution function (CCDF) for the percent of images in ImageNet that cause a given number of synapses to receive regulation above a threshold (T). We report individual runs (dot) and the mean across runs (dash-dot). (**E**) Test accuracy on ImageNet. These results indicate that the problem of forgetting may be orthogonal to the problem of generalization.

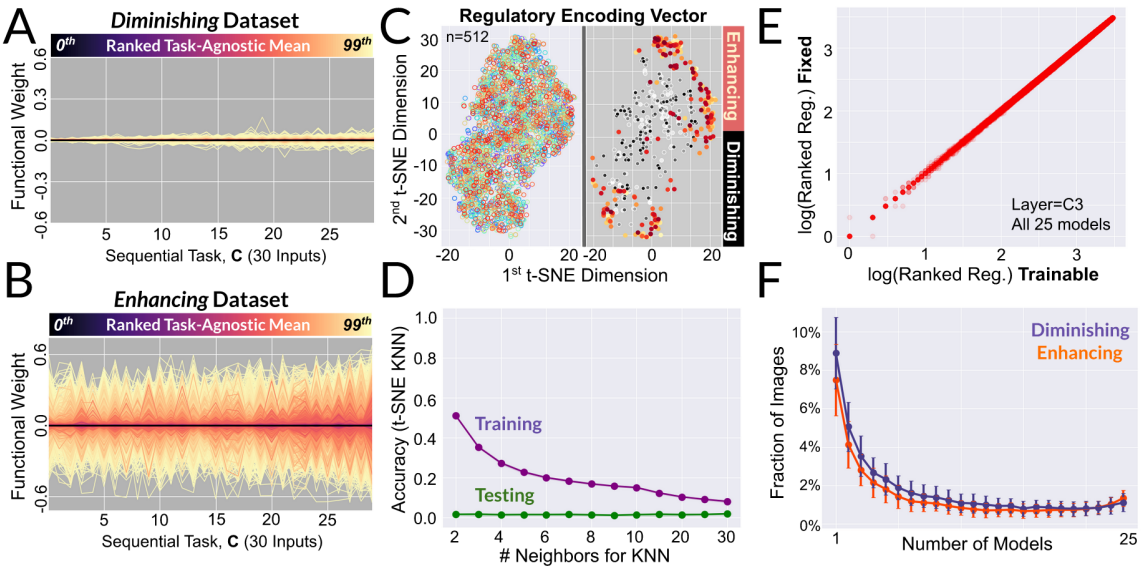

Figure 10: Randomly sampled window of synaptic activity (post-masking) under domain transfer for Grow in the Diminishing (**A**) and Enhancing (**B**) datasets. (**C**) The regulator toggles between two states, each with a similar presentation across tasks (*enhancing* and *diminishing*). Points in the right panel are dimensionally reduced regulator encodings for the images that are the most enhancing (top 256, yellow-red) and most diminishing (bottom 256, grey-black). Coloring by task identity (left panel) yields no discernable clusters, and for no run are tasks identifiable by their t-SNE projections using K-Nearest Neighbors. Dimensional reduction is obtained through a combination of principle component analysis (50 components) and the t-SNE algorithm (van der Maaten and Hinton, 2008). (**D**) Dimensionally reduced regulatory encoding vectors are not identifiable by their class ID. Following dimensional reduction (D=2) through a combination of PCA (50 components) and t-SNE (van der Maaten and Hinton, 2008), K-Nearest Neighbors for class prediction was performed. These results show that images belonging to a particular class are not represented by the regulator in a way that is more similar than images belonging to other classes. If they were, class prediction accuracy by KNN clustering would be high. Confidence intervals report standard deviation across all runs (n=25). (**E**) Log-ranked regulation for each image in ImageNet across 25 models of Grow when regulation is trainable (x-axis) and when it is fixed (y-axis). (**F**) For each class in Imagenet, all 600 images are ranked according to mean regulatory output for all 25 models of Grow. We report the mean fraction of images with membership in the top/bottom 100 images over all classes for the corresponding number of models (x-axis). Thus, the leftmost result reads that an average of 9% of images belong to the diminishing set (bottom 100) of just 1 model. Enhancing and diminishing sets used in the main text consider only the top 30 images per class. Error bars report standard deviation. (**C**).

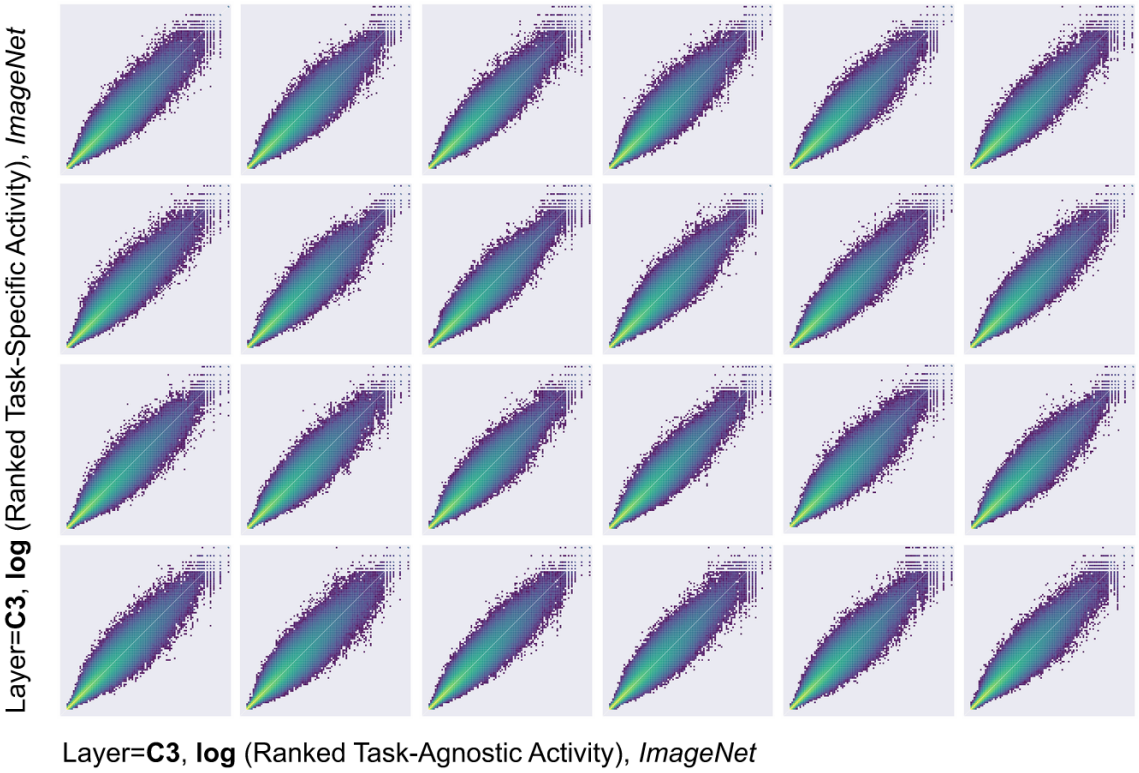

Figure 11: Regulation in C3 of the data rich condition under domain transfer to ImageNet does not elicit task-specific modularity. Instead, the most active weights on a given task are the most active weights over all tasks (*synaptic recycling*), and no synapse dramatically changes rank for any individual task. Here we present results for a single trial for each of the remaining 24 models. Finally, we note that the degree of context *sensitivity* is higher for layer C3 under domain transfer to ImageNet and CIFAR-100 than it is for layers C2 and C1. This is distinct from context *modularity*, and likely reflects the commonly observed property that downstream layers are dedicated to less generic features of the input.

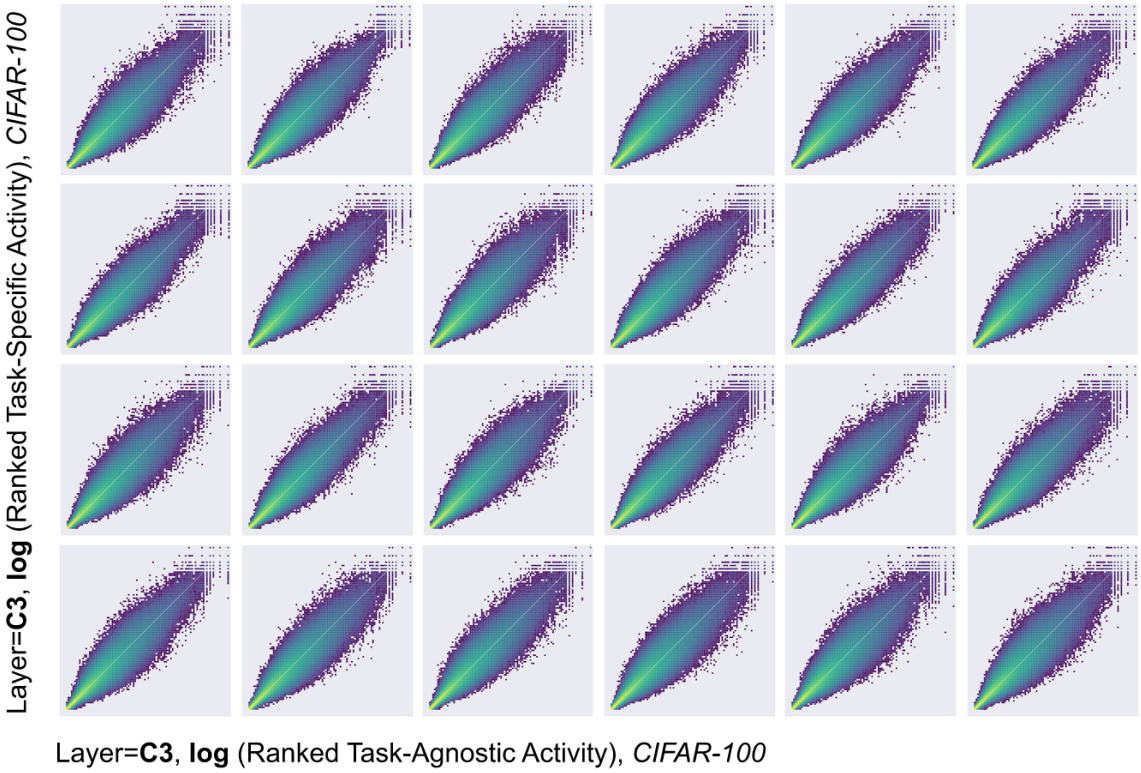

Figure 12: Regulation in C3 of the data rich condition under domain transfer to CIFAR-100 does not elicit task-specific modularity. Instead, the most active weights on a given task are the most active weights over all tasks (*synaptic recycling*), and no synapse dramatically changes rank for any individual task. Here we present results for a single trial for the remaining 24 independent models. Finally, we note that the degree of context *sensitivity* is higher for layer C3 under domain transfer to ImageNet and CIFAR-100 than it is for layers C2 and C1. This is distinct from context *modularity*, and likely reflects the commonly observed property that downstream layers are dedicated to less generic features of the input.

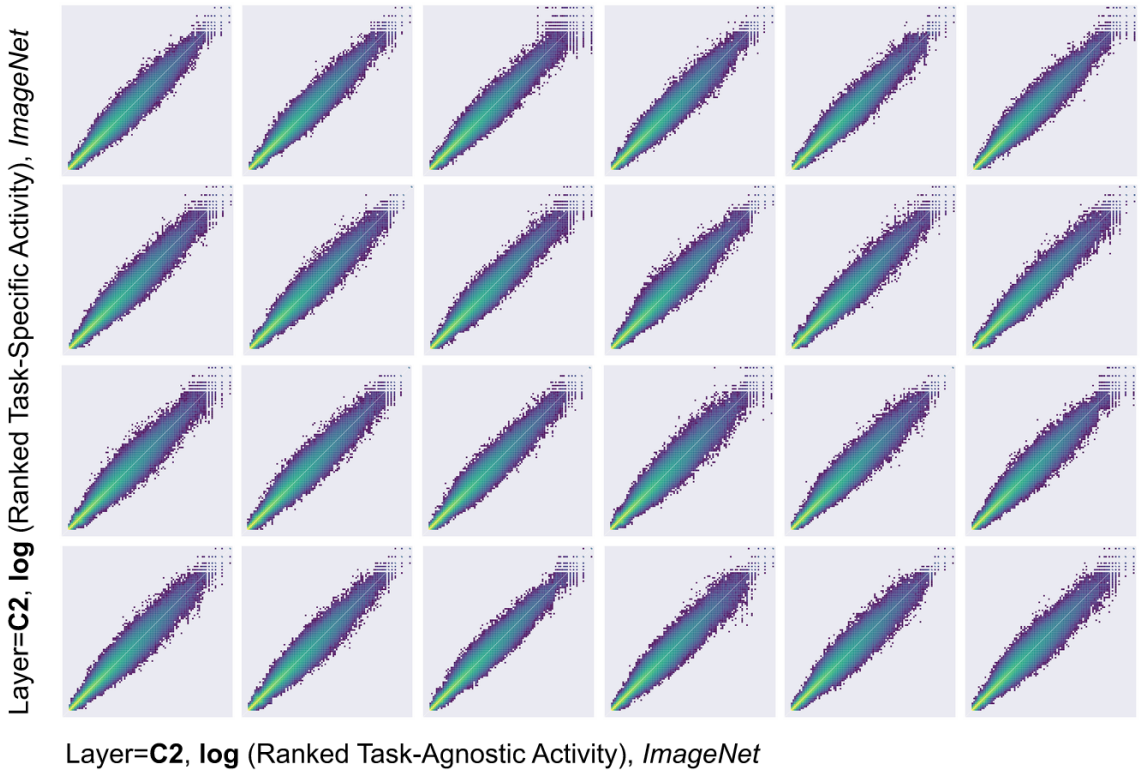

Figure 13: Regulation in C2 of the data rich condition under domain transfer to ImageNet does not elicit task-specific modularity. Instead, the most active weights on a given task are the most active weights over all tasks (*synaptic recycling*), and no synapse dramatically changes rank for any individual task. Here we present results for a single trial for 24 independent models. Compared to results for regulation of C3 under domain transfer to ImageNet, regulation of C2 is noticeably less context-sensitive. This may be attributed to the common observation that successive layers in neural networks attend to decreasingly generic properties of inputs. Nevertheless, context-dependent modules, which are an extreme form of context sensitivity, are not present in any of the convolutional layers.

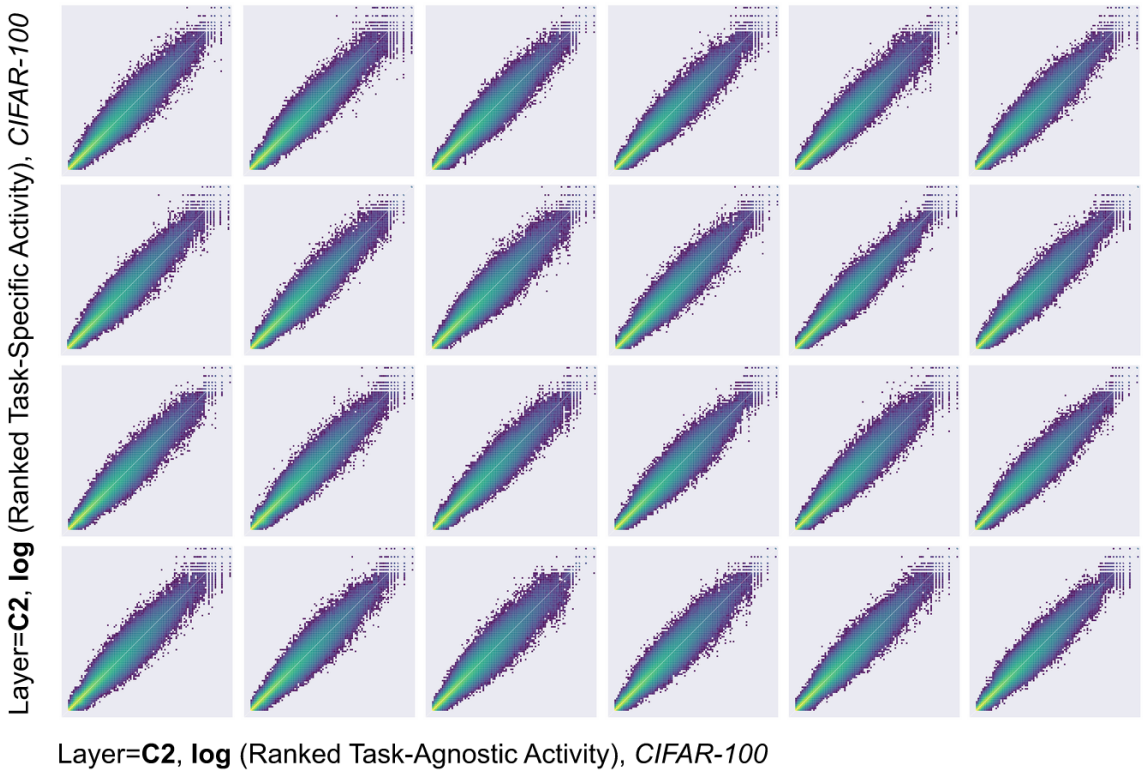

Figure 14: Regulation in C2 of the data rich condition under domain transfer to CIFAR-100 does not elicit task-specific modularity. Instead, the most active weights on a given task are the most active weights over all tasks (*synaptic recycling*), and no synapse dramatically changes rank for any individual task. Here we present results for a single trial for each of the remaining 24 models. Compared to results for regulation of C3 under domain transfer to CIFAR-100, regulation of C2 is noticeably less context-sensitive. This may be attributed to the common observation that successive layers in neural networks attend to decreasingly generic properties of inputs. Nevertheless, context-dependent modules, which are an extreme form of context sensitivity, are not present in any of the convolutional layers.

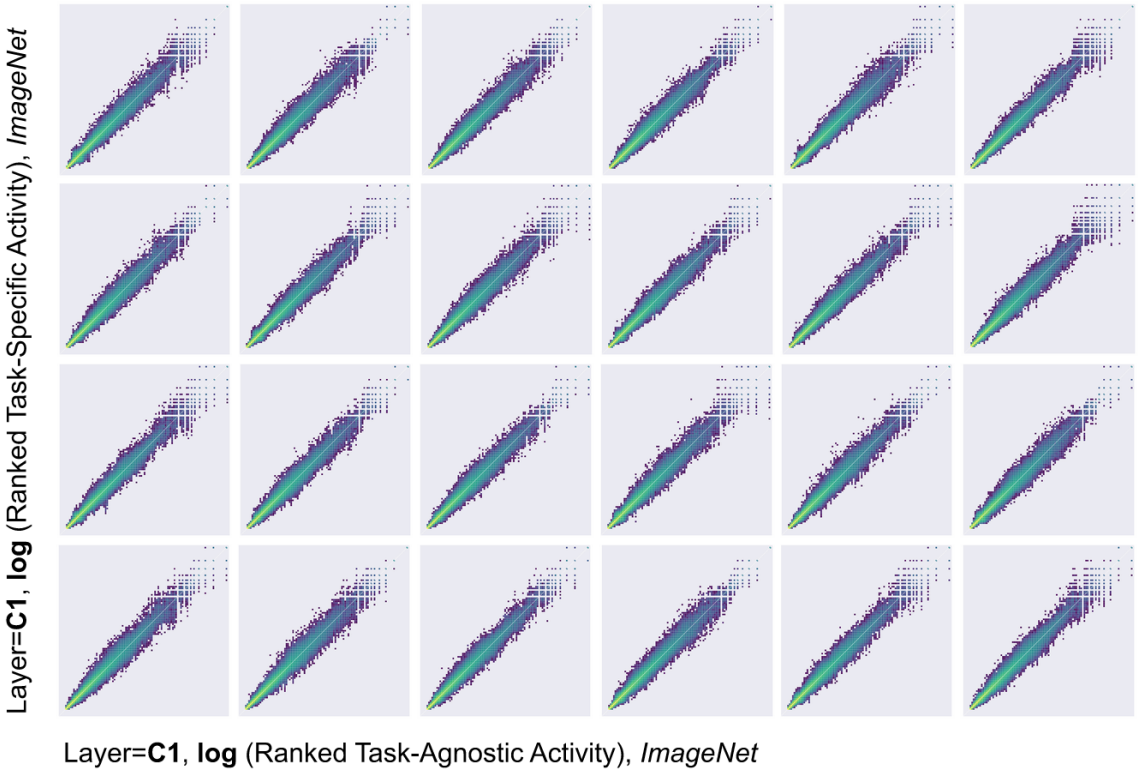

Figure 15: Regulation in C1 of the data rich condition under domain transfer to ImageNet does not elicit task-specific modularity. Instead, the most active weights on a given task are the most active weights over all tasks (*synaptic recycling*), and no synapse dramatically changes rank for any individual task. Here we present results for a single trial for each of the remaining 24 models. Compared to results for regulation of C3 and C2 under domain transfer to ImageNet, regulation of C1 is noticeably less context-sensitive. This may be attributed to the common observation that successive layers in neural networks attend to decreasingly generic properties of inputs. Nevertheless, context-dependent modules, which are an extreme form of context sensitivity, are not present in any of the convolutional layers of the classifier.

Figure 16: Regulation in C1 of the data rich condition under domain transfer to CIFAR-100 does not elicit task-specific modularity. Instead, the most active weights on a given task are the most active weights over all tasks (*synaptic recycling*), and no synapse dramatically changes rank for any individual task. Here we present results for a single trial for each of the remaining 24 models. Compared to results for regulation of C3 and C2 under domain transfer to CIFAR-100, regulation of C1 is noticeably less context-sensitive. This may be attributed to the common observation that successive layers in neural networks attend to decreasingly generic properties of inputs. Nevertheless, context-dependent modules, which are an extreme form of context sensitivity, are not present in any of the convolutional layers of the classifier.

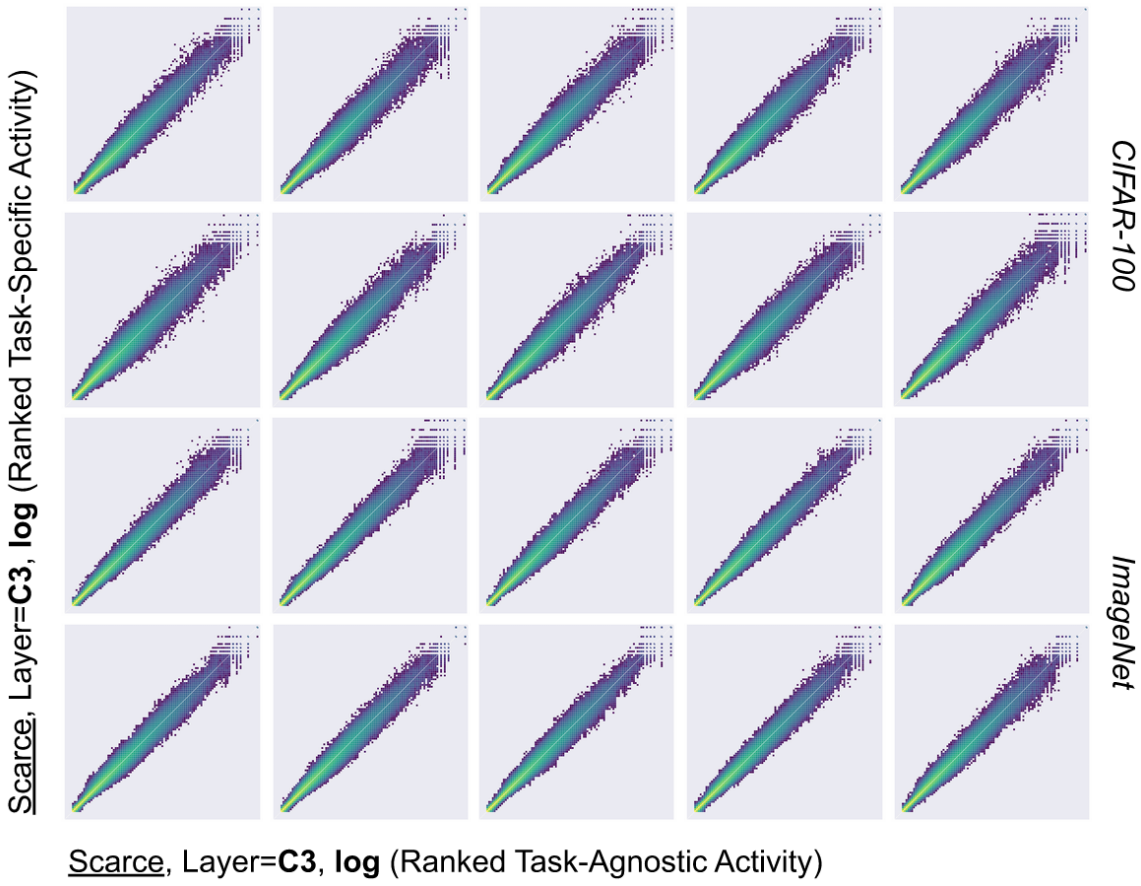

Figure 17: The lack of task-specific modularity observed in the regulation of C3 in the data rich condition is recapitulated in regulation of C3 in the data scarce condition. We also find that regulation in the data scarce condition is noticeably less context sensitive than in the data rich condition. Thus, we find that synaptic recycling in C3 is more strongly pronounced in the data scarce condition. We present results for 10 randomly sampled independent runs per dataset.

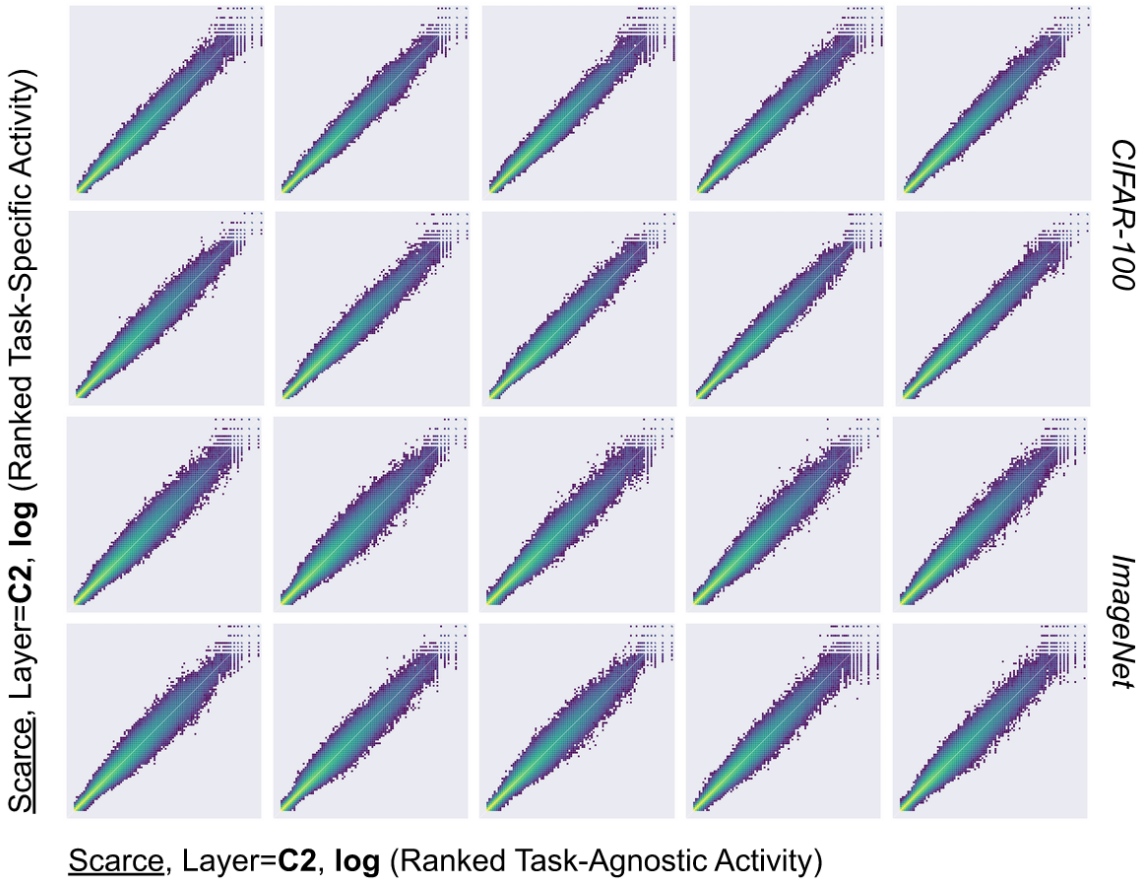

Figure 18: The lack of task-specific modularity observed in the regulation of C2 in the data rich condition is recapitulated in regulation of C2 in the data scarce condition. We also find that regulation in the data scarce condition is noticeably less context sensitive than in the data rich condition. Thus, we find that synaptic recycling in C2 is more strongly pronounced in the data scarce condition. We present results for 10 randomly sampled independent runs per dataset.

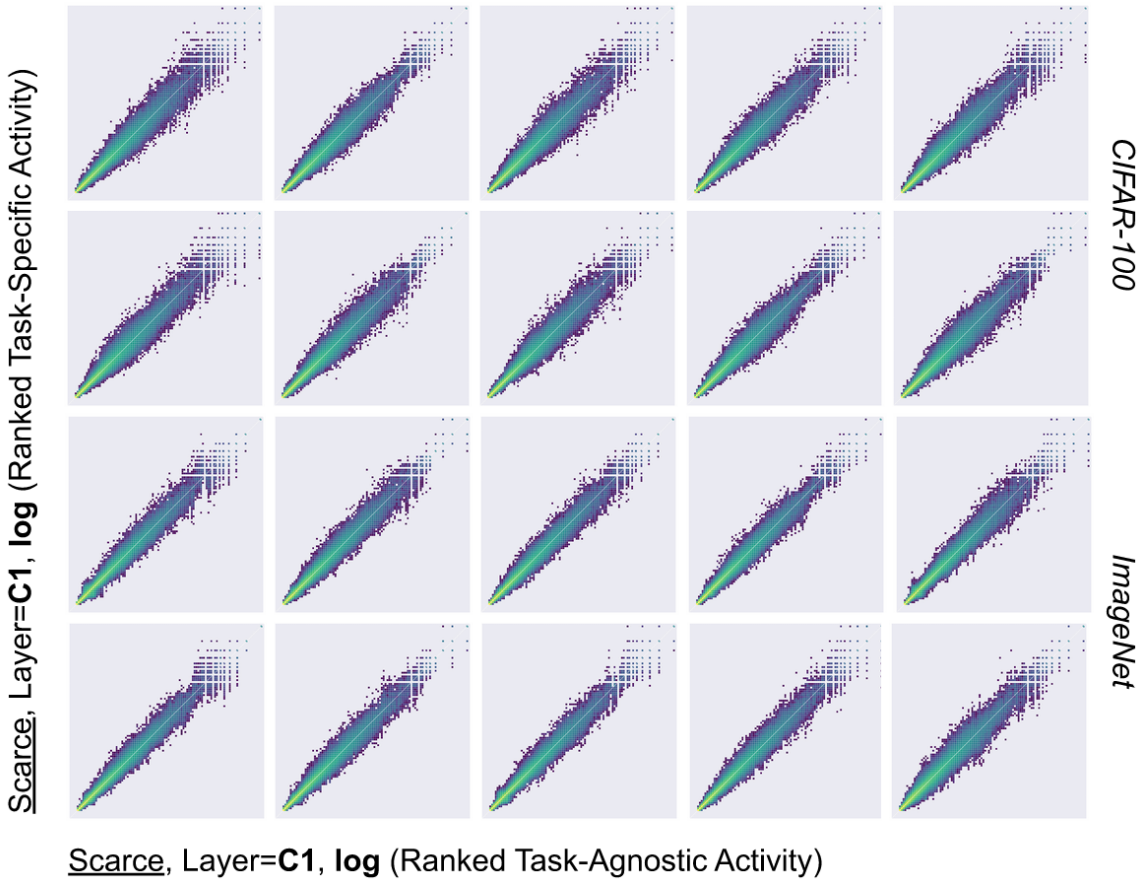

Figure 19: The lack of task-specific modularity observed in the regulation of C1 in the data rich condition is recapitulated in regulation of C1 in the data scarce condition. We also find that regulation in the data scarce condition is noticeably less context sensitive than in the data rich condition. Thus, we find that synaptic recycling in C1 is more strongly pronounced in the data scarce condition. We present results for 10 randomly sampled independent runs per dataset.