

Task-structured Modularity Emerges in Artificial Networks and Aligns with Brain Architecture

Corresponding Author: Dr Shi Gu

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Thank you for the opportunity to review "Task-structured modularity emerges in artificial neural networks and aligns with brain architecture" by Wu and colleagues. The manuscript describes a study of how modularity emerges in response to multi task learning. The authors show that the demand to learn multitask representations induces modularity. They then show that this general phenomenon can be observed in a generative, growth-like model.

The manuscript is very strong and sure to be of broad interest to the community. The link between modularity and multitask learning is demonstrated very clearly and from multiple angles. The work is also very complete, in the sense that the last few sections (e.g. the generative model) help to round out the findings and position the work in a broader context. My suggestions are minor and should be taken as optional.

1. Most of the performance measures are shown without a sense of context. It would be interesting to know what is chance level, and how alternative types of networks (e.g. those with pronounced core-periphery organization) would perform.

2. Could the authors comment on differences between tasks, and the types of representations they elicit? Given the set of 20 tasks, are there broad patterns that emerge for say, tasks that tap more into the fading memory property or the separation property elicit greater modularity?

3. The discussion is fairly brief. The final result, showing that task-conditioned generative models yield more long-distance connections, is particularly interesting and the sort of thing that could perhaps be more emphasized.

Bratislav Misic, McGill University

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

This paper investigates whether modular structures emerge spontaneously within recurrent neural networks (RNNs) trained under different learning paradigms (single-task, simultaneous multi-task, and incremental multi-task), and whether these structures resemble brain networks. The authors have RNNs perform a series of cognitive tasks involving working memory, decision-making, classification, and inhibitory control, train them under different modes, and compare network structural features. The authors interpret their experimental results as evidence that when task load exceeds network capacity, multi-task training significantly increases the network's modularity, and performance remains relatively high even after randomly pruning part of the connections. They further report that these task-driven networks are closer to biological brain networks in certain topological measures, such as clustering coefficient and connection edge length distribution. Overall, the authors propose the conclusion that multi-task training induces stronger structural modularity in RNNs, and that such "brain-like" structures can be viewed as a result of neural systems adapting to complex cognitive tasks.

While the conclusions of this work appear substantial and intriguing, the paper is hampered by a lack of novelty,

methodological deficiencies (including the improper application of the modularity metric, which constitutes a key element of their conclusions), inadequately controlled experiments, and conclusions that are not fully substantiated by the presented results. All these issues prevent the paper from meeting the standards of Nature Machine Intelligence and it would be more appropriately suited to a specialized journal.

1. I find that the core methods and experimental ideas in this work are based on existing studies and lack substantial innovation. Multi-task learning, incremental training, modularity computation have all appeared in recent literature already. What I observe here is mainly a combination of these methods with minor adjustments in details, without introducing relatively new approaches in theoretical framework, algorithm design, or experimental validation.

2. The authors directly compare modularity values between networks of different scales and claim that smaller networks have higher modularity. This approach is theoretically problematic because modularity is strongly affected by the number of nodes, number of edges, and network density. Without normalization or comparison to an appropriate random graph baseline that matches size and degree distribution, such direct comparison is not meaningful. In addition, if modularity is to be used to indicate the presence of modular structure, it should be evaluated against an appropriate null model. I recommend that the authors refer to the discussion on modularity definitions and limitations in Community detection in graphs. Due to the significant misuse of the modularity index by the authors, the relevant results currently do not support their conclusions regarding modularity.

3. The authors attempt to assess whether modularity can serve as an indicator of model performance, and they observe a correlation between accuracy and modularity during the middle of training. However, this correlation disappears once the model converges. Correlation during mid-training can be caused by multiple training dynamics factors, making it difficult to directly extrapolate such a result to brain networks. In addition, for the conclusion that “task-driven adaptation enhances more brain-like structural organization,” the spatial+task method is configured such that two key parameters can both take non-zero values, whereas in the other two methods at least one of these parameters is fixed to zero. This setup inherently ensures that spatial+task has an upper bound that is not lower than the others, which weakens the rigor of this comparison.

4. The experiments are conducted exclusively on a single RNN architecture without evaluation on other neural network types such as convolutional networks, graph neural networks, or alternative RNN variants. Without cross-architecture validation, the generality of the conclusions is limited, and the results cannot be readily interpreted as evidence for brain-like mechanisms given the substantial differences between artificial RNNs and biological brain networks.

5. Some experiments do not control for key variables. For example, in comparing incremental multi-task and simultaneous multi-task training, it is not stated whether the total number of training samples, parameter update steps, and gradient trajectory characteristics are matched. Differences in these factors can affect the emergence of modularity and confound the interpretation of results.

6. The implementation details of the connection-incremental learning procedure are insufficiently described, including the order of adding connections, the selection strategy, and the degree of randomness. There is no control experiment to verify whether the observed increase in modularity is due to the incremental addition mechanism itself or simply the result of a different training path. In addition, no ablation or control experiments are provided to separate the effect of incremental connection addition from other potential influences such as altered training dynamics. I am concerned that the design of the connection-incremental learning may not effectively capture the phenomena it is intended to represent, which raises doubts about the validity of the conclusions drawn from it.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

Overall summary

In their paper “Task-structured Modularity Emerges in Artificial Networks and Aligns with Brain Architecture” the authors train recurrent neural networks of different sizes with different training schedules that vary in terms of how many tasks are learned and whether neuronal connections are gradually added throughout learning or all present at the start. The authors conduct their analyses with the motivation of studying network modularity. Their findings indicate that modularity in neural networks might be driven by multi-task learning, as has been hypothesized by prior literature, at least in networks of relatively small sizes. In related functional analyses they show how overlapping populations might be used for solving overlapping tasks and lastly integrate their multi task optimised network connectivity matrices into generative models to show that their multitask training complements prior metabolic constraints for generating more brain-like connectivity.

Overall comment on correction

I enjoyed reading the manuscript and think it tackles a topic that is of interest to the community. I think the general approach taken by the authors and many of the analyses conducted have a lot of potential, but I also think that a certain amount of additional work is required to get the manuscript in shape. Below I outline these in more detail – I hope authors find my input helpful for improving their work.

Major

- Interaction between size and tasks: Across the manuscript you find very interesting interaction effects between size of network and tasks and you hypothesize that modularity only develops when networks are at their 'representational capacity' – this is a key point across your manuscript as you obviously need to explain why modularity would only sometimes develop in the light of multi-task learning. Given how central this argument is to the paper, I would expect that you at least have one plot or analysis (e.g. via ANOVA-like analysis) that shows that this effect is actually true. For this you would need to directly measure the representational capacity -- as far as I can see you do not directly measure representational capacity here yet. I take it that you see number of neurons as a valid proxy for representational ability, but that is actually a strong assumption, given that the number of weights in an RNN obviously does not scale linearly with the size of the RNN. Maybe a measure like effective dimensionality of the weight matrix of latent activation might be more appropriate, or at the very least the number of parameters in the network.
- Optional: If you wanted to be 100% rigorous about this, you would not only try to quantify the representational capacity of the networks but also the representational complexity of the tasks, but as I myself do not have a suggestion for how exactly one would do that, I want to leave this as an optional comment. I would certainly welcome ideas from the authors. Maybe there is a measure in here that would be helpful <https://arxiv.org/abs/1808.03591> ?
- As a subpoint to the above, I then suggest that you use this representational capacity to color lines, such as in Fig.2 e-h, so that one can actually see how the representational capacity interact with the modularity. The bi-directional color scale with blues and greens makes these Figures quite difficult to interpret.
- Fig 3d: This is an extremely interesting analysis but I do not understand how you arrive at your conclusion "Together with the findings in Fig. 2, these results suggest that when representational capacity is constrained, modularization emerges as a compensatory procedure to improve the neural efficiency in representation." given the fact that the modularity falls below significant at the end of training for all network groups, supposedly when accuracy is the largest? So, would the appropriate conclusion to this analysis not be that modularity is not directly linked to accuracy but instead somehow to learning? That seems to be a very interesting new angle, as modularity and compositionality are obviously supposed to help with finding new solutions. Perhaps this could be cleared up if you correlated 'how well learning is going to go' (so end of training accuracy) with how much modularity was achieved during learning? You might not want to look into the learning effect (that is fine by me) but I do not see how your current claim comes out of your analyses.
- Line 235-238; "The network initialized with the community mask exhibited superior performance over those marked by the random mask (Fig. 4e). Notably, this performance advantage was especially pronounced in models with smaller capacity, consistent with earlier results shown in Fig. 1 and Fig. 2.". I do not agree with your interpretation, as the community mask only helped in smaller networks. Your wording "especially pronounced" would only be correct if all network sizes showed this effect, which does not seem to be true?
- I am not sure I fully agree with your framing of the prior literature on two accounts:
 - (1) I think you should make clearer that on the level of representations, the link between multi-task learning and modularity has been made by multiple prior papers (sometimes calling them composable, or abstract), such as Yang Nature Neuro 2019 (as you cite) but also more recently <https://www.nature.com/articles/s41467-023-36583-0> or <https://www.nature.com/articles/s41593-024-01668-6> -- you of course add the structural component, which is valuable in itself, but I think it would be fair to acknowledge this
 - (2) At two points (lines 43-46 and 271-273) in the manuscript you make it sound as if the prior literature suggested that modularity was solely driven by metabolic constraints and that you now add that it is driven by metabolic constraints and tasks (as discussed in section 241). I think this is a misrepresentation of prior research. Achterberg Nat Mach Int (as you cite) already specifically discusses the interaction of tasks + metabolic costs for driving modularity (note lower modularity for non-task solving networks on their paper) and also already discuss / hypothesise that this effect might be even larger in multi-task environments that allow for more modular learning (not tested by them). I think your analyses nicely confirm these hypothesised links, which is a very good contribution, but I think it would be better to acknowledge that they follow from prior work.
- Section starting line 208: This is an interesting analysis but I am a bit concerned that this is the same effect that Yang et al 2019 already show. They show their modular representation are achieved through a method which influences how much different synapses are updated as function of how important they are for a given task (first described in Zenke et al, "Continual learning through synaptic intelligence"). So while they do not add new connections, but they de facto end up training different groups of connections in the network over time, which intuitively seems similar. One of the complications with pinning down a possible link right now is that you do not discuss which optimiser you use. If you use a momentum based optimiser, depending on how you fix the recurrent weights, it might be that you carry the prior momentum forward across batches when new tasks are added and that might end up achieving an effect where specific tasks are optimised through specific weights, even if you do not hard code this. Another effect could be that the absolute weights of newly added neurons have a different magnitude than the prior ones, which will again influence how they are changed by the optimiser. I am not against this analysis per se, but I am somewhat concerned that what we are seeing here is perhaps a side effect of the optimiser and not directly of the adding weights, as you describe.
- Line 204; "In contrast, when the model's capacity is sufficiently large, the model exhibits a certain degree of over-parameterization, endowing it with inherent robustness. In the late iterations of training, further perturbations in the weights have minimal impact on task performance, leading to a weakened correlation between modularity and performance.": Do you have any analyses to back up your claim about perturbation robustness? This is the first time (and it seems to me, the only time) that you mention perturbations and you do not directly refer to any analyses here.

Minor

- Some important methods details, such as used optimiser (also see discussion of optimiser above), learning rate, initialisations, activation function (do you use Relu? In line 337 you just say "e.g. ReLU") are missing from the methods.

- Line 134-140 / Fig 2; e-h: I think you need to improve the writing to first clearly describe the results you observe and then include your more interpretative description afterwards. I find it confusing that you do not at all describe the results here (you seem to get quite an interesting interaction of number of tasks and model size?) and instead say "When the numbers of hidden neurons are sufficiently large (e.g., ≥ 32 for the current setup), it is almost possible to represent three tasks in separate, non-overlapping populations. Thus, incorporating more tasks may not result in a more significant modularization given Eqn. 2.". This is not a description of the results but instead a speculation about representational ability and the resulting modularity.
- Line 106/107, "Consistent with [9], we observe the similar phenomenon of co-embedding functionally similar tasks for reducing the demand of hidden nodes.": Where do you show this? This does not seem to link to a Figure? Do you just deduct this from the accuracy numbers, as I do not see any analysis of representations, as I believe ref.9 used to make that point?
- Line 144: You discuss co-embedding here but you have not yet studied the actual representations. Generally speaking, it would make it easier to follow your writing if you used less interpretative language and instead stated the variable names you have actually analysed so far — so not co-embedding but network size and number of tasks.
- I understand that it is difficult to fit methods details into the main text given the word limit, but I think it would be extremely helpful to know how many networks you train for which analysis, as part of the main text, e.g. sections starting line 92 and line 113 do not seem to mention that piece of information.
- Fig 2: Please use thicker legend lines, the colours are hard to read
- Line 177-179; "Inspired by representational similarity analysis (RSA) [26], we introduce a metric to quantify neuronal similarity based on activation patterns during task execution, thereby enabling a fine-grained assessment of shared versus task-specific neural involvement.": I think it is extreme to say that you introduce a new metric here, you just develop a specific way of analysing representations, as many electrophysiology and MechInterp papers do. So I would not say that you develop a new metric.
- Section starting line 175: So which specific network is analysed here and on which tasks is this trained? It might be interesting to understand how this developing modularity of representations develops alongside the task accuracy, but I am unsure which accuracy data would be linked to this specific network. I am actually a bit confused because assuming that what you show here is the neuron to neuron matrix of the recurrent units, these matrices seems to have 128 neurons even though before the largest one was 64? Maybe I am misunderstanding this, but if these are indeed larger networks, what does the modularity curve for these networks look like? Are they still modular?
- Fig 2 e-h: I feel the comparison across network sizes is difficult for the reader here as y-axis values are varying across plots
- Line 69: When you mention training mode III, I think it would be helpful to the reader to point to the fact that you only use this training method for the section starting line 208, if I am correct. I read the following section wondering whether the multitask networks are training mode II or III until that section later suggested that only this late section uses training mode III.
- Line 219; "iterations after 10000": Space missing before '10000'
- Line 251; "The 'task' models disregard the spatial constraints and define the feature kernel $K(u,v)$ using the functional connectivity matrices of the trained RNNs.": Could you define of which trained RNNs? Is this from the multi-task RNNs? if yes, which training plan? From your mention of this result in line 284 I assume it is the multi task connectivity matrix.

Conclusion

I want to thank the authors and editor for sending this interesting paper to me. I enjoyed reading this and hope my comments are helpful to improve this manuscript. I would be happy to review an improved version of this manuscript and consider it for publication in Nature Machine Intelligence.

(Remarks on code availability)

I have had a look at the Github repository and at the current stage the README file does not have enough clear instructions to easily reproduce the analyses.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have comprehensively addressed my suggestions and I recommend publication.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

I greatly appreciate the substantial efforts the authors have invested in revising the manuscript and preparing a comprehensive point-by-point response. The additional experiments (particularly the cross-architecture evaluations, expanded ablation studies, and a more detailed description of the connection-incremental procedure) have enhanced experimental clarity and transparency.

However, despite these valuable clarifications, the most critical methodological concern raised in the initial review remains

unresolved. In addition, although the authors have carried out substantial engineering work, the overall study does not align with the aims of Nature Machine Intelligence. The manuscript remains a mechanical combination of established methods rather than a genuinely novel paradigm, and several logical inconsistencies in the results further weaken the claim of conceptual innovation. The biological principles the authors highlight are effectively hard-coded into the system rather than emerging from a new framework with broader applicability.

As a consequence, the work does not meet the journal's expectation of delivering insights or methodologies that can meaningfully influence research beyond the immediate experimental setting. This prevents me from offering a positive assessment of the revised version -- I must emphasize that the manuscript has not yet met the publication standards of Nature Machine Intelligence. The specific comments are as follows.

1. In the initial review, I highlighted that the "spatial + task" model inherently possesses more degrees of freedom than either the spatial-only or task-only counterparts. This expanded flexibility enlarges its effective search space, meaning the model's superior fit is statistically expected rather than scientifically meaningful. In their response, the authors acknowledge this increased complexity but argue their goal was to illustrate the joint contribution of spatial and functional constraints, rather than to conduct model comparisons on an equal footing.

However, the core issue lies not in the interpretive framing but in the absence of a principled comparison that accounts for differences in model complexity. The manuscript still fails to employ any standard tools (e.g., akaike information criterion, bayesian information criterion, or cross-validation approaches) for comparing models with varying numbers of parameters. Without such critical controls, the improved fit of the combined model cannot be disentangled from the advantage conferred by additional degrees of freedom. Consequently, the evidence provided does not convincingly support the claim that the brain's topology requires both spatial and task-based constraints; instead, the observed results could simply reflect overfitting by a more flexible generative framework. This fundamental gap leaves the manuscript's key conclusions without sufficient empirical grounding.

2. While the introduction of the LME model attempts to statistically control for network size, it fundamentally fails to address the graph-theoretical core of my original concern. The issue is not merely that network size explains variance in modularity, but that the raw modularity score Q is mathematically biased by network density and sparsity. A sparse network will naturally exhibit higher raw modularity values purely as a mathematical artifact, independent of any biological or functional organization. The LME analysis attempts to regress out the effect of network size, but it does not normalize the metric against a topological null hypothesis. Consequently, the claim that smaller or incrementally trained networks are "more modular" remains theoretically unsubstantiated.

3. In the initial review, I emphasized that correlations between modularity and performance during mid-training are inherently difficult to interpret, as they may stem from optimization dynamics rather than meaningful structural properties of the model. While the revised manuscript does demonstrate more stable correlations under controlled conditions, it still fails to provide a principled demonstration that these relationships are structural in nature rather than artifacts of learning dynamics. Furthermore, there remains a lack of a theoretical framework explaining why mid-training modularity should bear a mechanistic relationship to learning efficiency—whether in artificial neural networks or biological systems. Although the authors have incorporated stronger experimental controls (e.g., fixed data streams and uniform initializations) in this revision, the core concern regarding the interpretability of these mid-training correlations persists unresolved.

4. The additional experiments across GRU, LSTM, CNN, and GNN architectures expand the empirical scope of the paper. However, the definition of modularity varies across architectures, and the manuscript does not provide a principled justification for treating these formulations as conceptually equivalent. In CNNs and GNNs especially, the computation of modularity via layer-derived feature matrices or bipartite graph constructions represents a fundamentally different notion of modular structure compared to modularity in recurrent weights. Without a unified theoretical framework explaining why these different quantities should be comparable or why multitask learning should induce modularity across such disparate architectures, the results appear more like a collection of case studies than evidence for a general principle.

5. The contributions, though well-executed from an engineering standpoint, do not introduce fundamentally new algorithms, new theoretical constructs, or a novel modeling paradigm that would meaningfully shift current understanding of neural modularity or task-driven structural emergence. The revised manuscript demonstrates the enhancements in experimental rigor and scope. Nevertheless, the core methodological concern identified in the initial review remains fundamentally unaddressed. Given that this issue directly undermines the validity of the manuscript's central conclusions, the current version cannot be accepted for publication in the current journal.

Additionally, lingering questions regarding the fairness of model comparisons and the interpretation of mid-training correlations further constrain the robustness of the paper's claims. For these reasons, the manuscript is better suited for a more specialized journal rather than Nature Machine Intelligence.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

Dear authors,

thank you so much for your thorough revisions of the manuscript. I appreciate the care you took and found this a very constructive reviewing process. I am happy to recommend the manuscript for publication in Nature Machine Intelligence.

(Remarks on code availability)

The README of the repository gives good instructions for running the code and having skimmed some of the code it seems to be generally well structured. I noticed however that some of the function scripts, like "math_utils.py" sometimes only have comments in the code in Chinese. I wondered whether English comments could be added for the international audience of NMI. I have not run the code myself.

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

I appreciate the authors' detailed explanations. However, after reviewing the revised manuscript and response letter, I find that several core methodological and conceptual concerns from my previous review remain unaddressed. Consequently, the conclusions are still not sufficiently supported, and the work does not represent a significant advancement in the field. Furthermore, the manuscript lacks a unified theoretical framework to underpin its generalized claims. Given these limitations, I remain of the opinion that while this work represents an experimental case study, its inability to theoretically substantiate its generalized conclusions makes it fall short of the cutting-edge, broadly impactful research expected by Nature Machine Intelligence. Below are my comments.

Major:

1.
The current generative models suffer from a circular reasoning problem. Specifically, in constructing the 'task' and 'spatial+task' models, the authors define the feature kernel using functional connectivity derived from RNNs that have already been optimized for those tasks. While the study intends to demonstrate how task demands shape network topology, the generative model is essentially initialized with the functional correlations of a network that already possesses that exact topology. It is inherently expected that a generative model built using the "blueprint" of a task-optimized network will naturally output a task-optimized structure. Because the desired outcome is effectively baked into the starting conditions, this approach borders on circular reasoning. Consequently, the current empirical results do not fully support the assertive conclusions presented in the manuscript.

2.
While the addition of CNN and GNN experiments successfully expands the empirical scope, the mathematical definitions of modularity in these architectures (e.g., feature grouping in CNNs versus weight clustering in RNNs) appear to be distinct. The authors acknowledge in their rebuttal that a single mathematical framework for all these topologies is currently lacking. A unified theoretical framework is essential to support the broad conclusions of the paper; otherwise, the current findings would merely constitute a series of isolated case studies rather than a systematic set of conclusions.

3.
I previously noted the lack of a theoretical framework explaining why mid-training modularity bears a mechanistic relationship to learning efficiency. In the response letter, the authors provide hypotheses involving "structural gradient alignment," "gradient subspace separation," and orthogonalizing task gradients. While these offer helpful intuitive explanations, they do not yet constitute mathematically rigorous proofs or strict mechanistic demonstrations. Without a formal mechanistic derivation, the underlying reasons for why this dynamic emergence occurs remain somewhat speculative. Consequently, a certain disconnect persists between the theoretical framework and the conclusions presented.

Furthermore, the authors repeatedly claim to provide a "causal explanation" or "causal mechanism" for how functional demands shape biological brain network topology. I would suggest softening these claims. Artificial RNNs optimized via global backpropagation serve as a valuable functional analogy, but they do not fully replicate biological causal mechanisms. Because the results remain primarily algorithmic case studies and artificial abstractions without a unified theoretical framework, the broad conclusions claimed by the authors are not sufficiently supported. Therefore, I stand by my assessment that this work is better suited for a specialized technical or engineering-focused journal, rather than providing the broad conceptual innovations expected for Nature Machine Intelligence.

Minor:

The authors state in their rebuttal that they have calculated AIC and BIC to account for model complexity. While I acknowledge that these scores were calculated using a 'Synthetic Likelihood' approach in the supplementary materials, this critical discussion is currently absent from the main text and must be integrated to support the study's conclusions.

(Remarks on code availability)

Version 3:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I think that the authors have convincingly addressed the remaining concerns raised by myself and other reviewers and I recommend publication.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

I appreciate the authors' detailed response letter and the extensive revisions made to the manuscript, including the softening of causal claims and the addition of the new theoretical section. However, after a thorough evaluation of the revised manuscript and the rebuttal letter, I find that the fundamental methodological and conceptual limitations I raised previously have not been resolved. Instead of providing robust solutions to the foundational theoretical gaps, the revisions largely consist of retreating from the original broad claims and applying restrictive caveats. Consequently, the conclusions are not sufficiently supported for Nature Machine Intelligence, as the work does not represent a significant advancement in the field.

Below are my comments.

1. The newly introduced theoretical framework (Supplementary Note 7) fails to provide a valid mechanistic proof due to a severe model mismatch. To mathematically justify the emergence of modularity and its relationship to learning efficiency, the authors rely entirely on a "stable linearized multitask RNN approximation". However, the fundamental effectiveness of modern machine learning (specifically the continuous-time RNNs simulated in this study) is unequivocally rooted in non-linear dynamics. The actual models implemented and trained by the authors are governed by highly non-linear activation functions (ReLU or tanh) and are subjected to complex, dynamic rule inputs and continuous-time integration with noise. Non-linearity is not a peripheral feature; it is the core computational mechanism that allows artificial neural networks to dynamically route information, suppress interference via activation thresholding, and carve out distinct, robust representational manifolds for multiple tasks. By adopting a linearized assumption, the authors fundamentally violate this principle, stripping away the exact non-linear mechanisms responsible for the emergent topological phenomena they are trying to study. Consequently, the theoretical explanation remains fundamentally disconnected from the actual physical evolution of the network's hidden states.

2. The manuscript still lacks a unified framework to underpin its generalized claims across different architectures. In the response letter, the authors explicitly admit that providing a unified mathematical framework for CNNs, GNNs, and RNNs "is a big research program rather than a revision item". To circumvent this, they have explicitly downgraded the CNN and GNN results to mere "corroborating empirical evidence" and "robustness checks". This concession directly validates my previous concern: by stripping these architectures of their theoretical standing, the paper confirms that it is presenting a series of isolated, architecture-specific algorithmic case studies rather than uncovering a universal, generalized principle of intelligence.

3. The circular reasoning embedded within the generative modeling analysis has not been resolved, despite the authors' defense. The authors argue that because the generative model is evaluated against the independent empirical human connectome, the logic is sound. However, this conflates the fitting target with the generation mechanism. The generative model's wiring probability is explicitly governed by the feature kernel $K(u,v)$, which is defined directly by the functional connectivity of the task-optimized RNNs. Because the generation rule itself is heavily biased by a matrix that already possesses a task-optimized topology, it is mathematically tautological that the resulting synthetic network will exhibit task-optimized structural features. Fitting this synthetic network to human brain data merely confirms that the human brain exhibits task-optimized features; it does not mechanically prove that functional task demands are what causally induced this topology from scratch.

Given these persistent limitations, the authors' strategy of softening their claims through numerous caveats severely dilutes the conceptual impact of the study. While this work remains a thorough experimental case study of RNN algorithmic properties, the fundamental disconnect between its theoretical framework and empirical models, along with the unresolved circularity in its generative approach, prevents it from delivering the broad conceptual innovations it originally promised. Ultimately, the inability to theoretically substantiate its generalized conclusions renders the manuscript more appropriate for a specialized machine learning or engineering venue, rather than meeting the cutting-edge, interdisciplinary standards expected by Nature Machine Intelligence.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Response Letter

Reviewer 1:

We are grateful for your encouraging comments on our topic and methodological approach of understanding modularity. Your suggestions on the clarity and discussion are very beneficial for us to further improve the quality of the manuscript. We believe that these revisions, now highlighted in *blue* within the document, will comprehensively address these highlighted points. We will respond to the comments one by one in the following context.

Q: Most of the performance measures are shown without a sense of context. It would be interesting to know what is chance level, and how alternative types of networks (e.g. those with pronounced core-periphery organization) would perform.

A: We are grateful for your very helpful suggestion. We have updated our manuscript with a revised Figure 2a,b to provide the necessary context. The chance level is now indicated with a dashed line in the figure, which is approximately at 0.15.

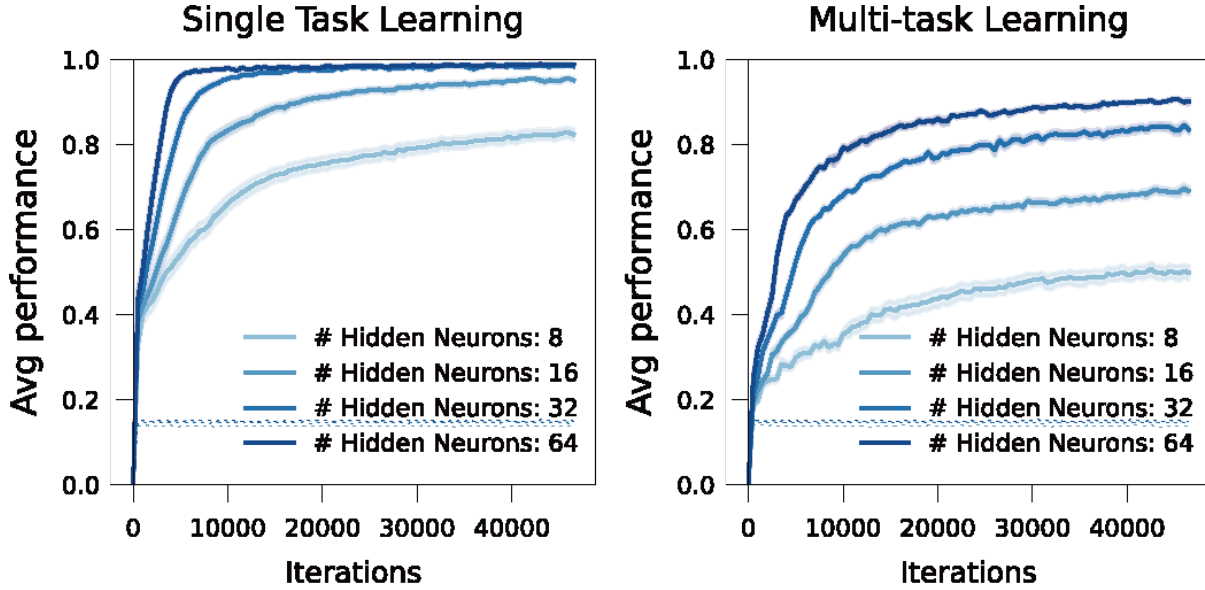


Figure 1. Performance of models trained on single task and multiple tasks. The dashed lines here in a, b indicate the chance level, which is approximately at 0.15.

Regarding the performance of alternative network types with pronounced core-periphery organization, we have conducted an additional experiment shown as follows:

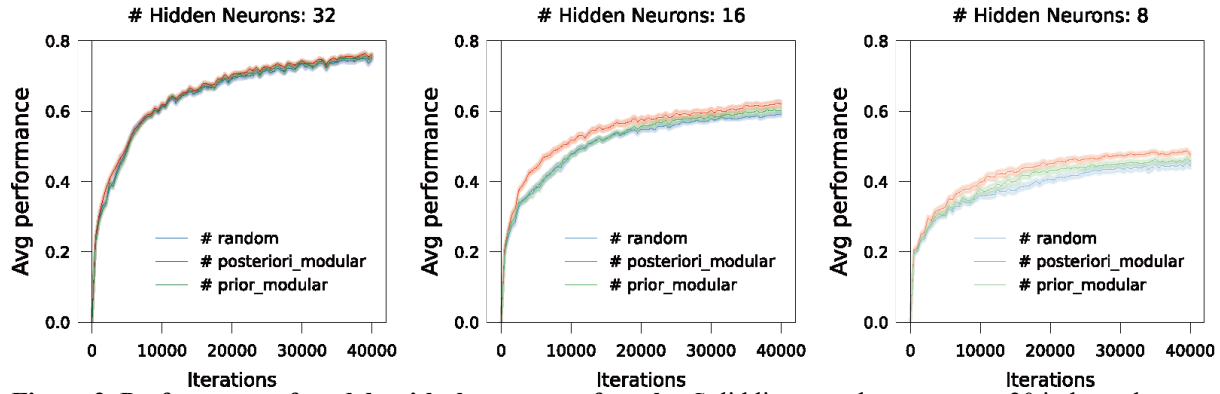


Figure 2. Performance of models with three types of masks. Solid lines are the mean over 20 independent runs. Shaded areas indicate the standard error.

We introduced a binary mask (0/1) to the recurrent connections of the RNN, where a value of 1 indicates a selected connection that participates in the computation, and 0 indicates a zeroed-out connection.

We defined three distinct groups/masks, all with the same number of connections:

- 1) the *prior_modular* mask: This mask was manually generated based on a core-periphery organization. Connections were sampled from a distribution based on edge type (e.g., intra_core_prob = 0.8, inter_core = 0.1, core_periphery_prob: 0.08, intra_periphery_prob: 0.02).
- 2) the *posterior_modular* mask: This mask was derived from the model after multi-task learning. We identified the emergent modular structure with the maximum modularity during training, and then we preserved all the intra-community connections and the top 5% high-magnitude inter-community connections, thereby maintaining the interactions between different modules.
- 3) the *random* mask: Connections were sampled uniformly at random.

We re-initialized the model and trained it from scratch using each of the three fixed masks. We found that:

The model with *posterior_modular* mask showed a significantly better performance compared to the ones trained with the other two masks. This suggests that the learned, emergent core-periphery organization provides a substantial advantage for network learning. While the manually generated prior_modular mask did show some improvement over the random mask, it was inferior to the emergent "winning lottery ticket" structure. We have updated Figure 4e in the revised manuscript to reflect these results.

Q: Could the authors comment on differences between tasks, and the types of representations they elicit? Given the set of 20 tasks, are there broad patterns that emerge for say, tasks that tap more into the fading memory property or the separation property elicit greater modularity?

A: Thanks for your insightful questions. For the first question, we have added a table to the Supplementary Note 1 introducing the functional roles of the different tasks.

Table 1 Cognitive task requirements and elicited representations

Task Family	Task Member	Core Computational Primitive	Required Representation Type
Go Family	fdgo, reactgo, delaygo	Sensorimotor Mapping, Inhibitory Control (in `fdgo`), Working Memory (in `delaygo`)	Sensorimotor & Memory: Representations must map a sensory location to a motor plan. `delaygo` specifically requires maintaining this location trace over a delay (fading memory). `fdgo` requires inhibiting a prepotent response.
Anti Family	fdanti, reactanti, delayanti	Flexible Sensorimotor Re-mapping, Inhibitory Control, Working Memory (in `delayanti`)	Rule-based Transformation: Representations must apply a contextual "Anti" rule to transform the sensory location into a new, opposite motor plan, requiring high separation of sensory input from motor output.
Decision Making (DM) Family	dm1, dm2, contextdm1, contextdm2, multidm	Spatial Evidence Integration, Selective Attention/Gating (context tasks)	Integrator/Accumulator: Representations must simultaneously compare and accumulate analog evidence (stimulus strength) from two competing spatial locations. Context tasks require this integration to be gated by a rule.
Delayed DM Family	delaydm1, delaydm2, contextdelaydm1, contextdelaydm2, multidelaydm	Temporal Evidence Integration, Analog Working Memory, Selective Attention/Gating	Analog Working Memory: Representations must store an *analog* value (strength of stim 1) over a delay and then compare it to a second sequential input. This demands robust memory of a continuous variable.
Matching Family	dmsgo, dmsnogo, dmcgo, dmcnogo	Feature-based Working Memory, Categorization (dmc tasks), Comparison	Abstract/Categorical: Representations must store discrete features (location/modality) in `dms`. For `dmc`, they must further abstract a continuous location into a binary category, maintain this abstract label, and compare it to a subsequent stimulus category.

In response to the second question, we have performed an analysis where we individually calculate the mean modularity during the training process for each of the 20 tasks, and the results are as follows:

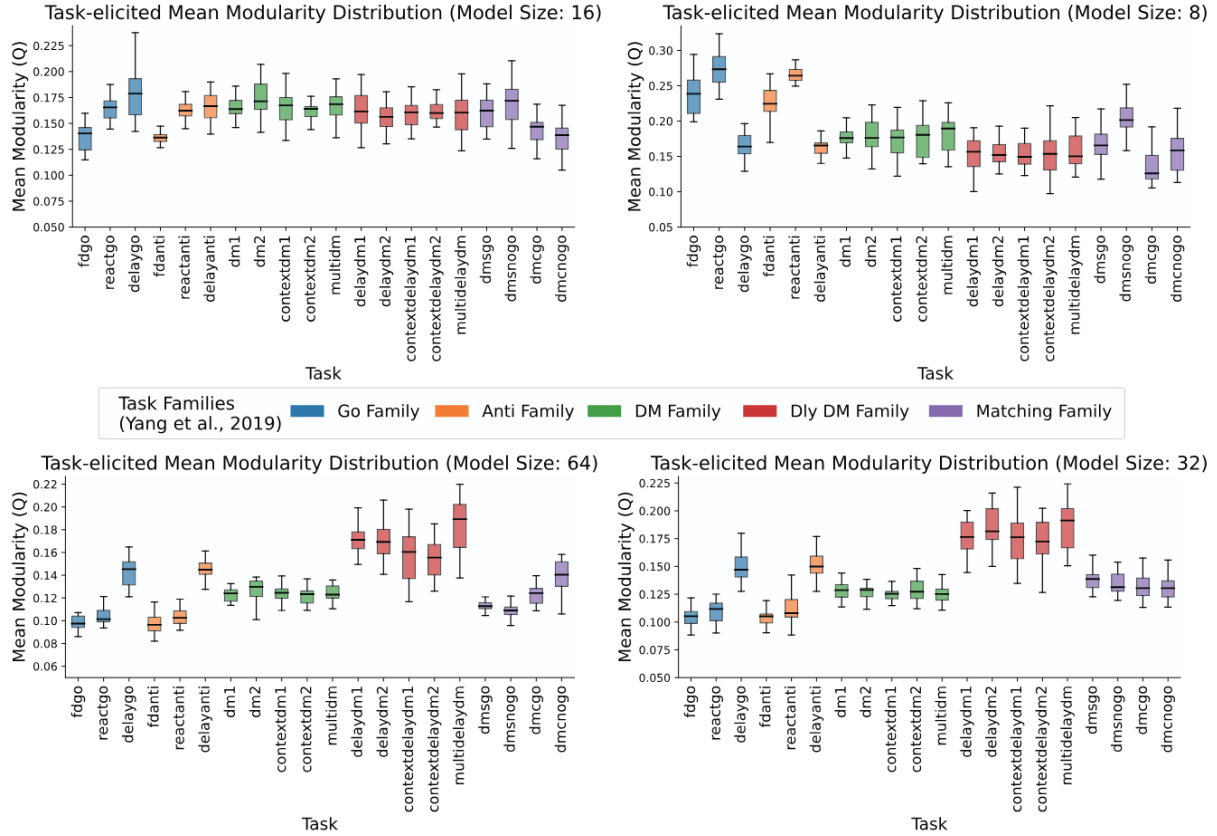


Figure 3. Task-specific modularity across all 20 cognitive tasks. Mean modularity (Q) of networks trained on each of the 20 single tasks, grouped by model size (N=8, 16, 32, 64). Tasks are color-coded by their functional family. Box plots show the distribution over 20 independent training runs.

The figure above (which we have included in the Supplementary Information) provides a detailed breakdown of the modularity elicited by each of the 20 tasks when trained individually. This analysis reveals broad, consistent patterns that directly link a task's computational requirements to the network's emergent structure, while also highlighting a crucial interaction with model capacity.

In larger models (Model size: 64 and 32), a clear hierarchy emerges:

- The Delayed Decision Making (Dly DM) family (red) consistently yields the highest median modularity. These tasks are the most computationally complex, as they require the integration of context, evidence accumulation, and working memory over a delay, thus tapping heavily into the need for sustained and segregated activity.
- Conversely, tasks requiring simpler, immediate sensorimotor mapping (the fdgo, reactgo, fdanti, and reactanti tasks) produce the lowest median modularity.
- The Decision Making (DM) and Matching families fall between these two extremes.

Impact of the 'Delay' Component: A "delay" component consistently induces greater modularization than its non-delay counterpart. This is seen in delaygo versus fdgo, delayanti versus fdanti, and the Dly DM family versus the DM family. This indicates that the demand for

working memory—which engages the "fading memory" and "separation" properties — introduces significant computational pressure that drives structural segregation.

Subtle Control Requirements: Within the *Go* and *Anti* families, tasks requiring reactive control (reactgo, reactanti) tend to elicit slightly higher modularity than their fixed-duration counterparts (*fdgo*, *fdanti*). This subtle increase suggests that the necessity for precise, reaction-time-based control over a fixed internal temporal mapping also contributes marginally to the pressure for structural segregation.

Modulation by Model Capacity

This pattern is strongly modulated by model capacity:

- In the 16-neuron models, the trend largely holds, where delay tasks still induce higher modularity than their non-delay counterparts.
- However, this phenomenon dramatically inverts in the most severely constrained models (Model size: 8). Here, the simpler *fdgo* and *fdanti* tasks produce higher modularity than their more complex delaygo and delayanti counterparts.
- This inversion suggests a threshold effect: when network capacity is too limited, it cannot form the necessary complex, modular solution for the delay tasks. The network can, however, still adequately solve the simpler tasks, for which a moderately modular structure is sufficient and achievable.

Taken together, these results support our central hypothesis: Modularity is not a default property but an emergent structural adaptation to specific computational demands. The strongest demand is the need to maintain information and segregate computational processes over time—as seen in all delay-based tasks—driving the network to find more modular solutions, provided it has sufficient representational capacity.

Q: The discussion is fairly brief. The final result, showing that task-conditioned generative models yield more long-distance connections, is particularly interesting and the sort of thing that could perhaps be more emphasized.

A: Thank you for your suggestion. Based on your suggestion as well as one from Reviewer 3, we added an additional discussion paragraph in the revised manuscript that elaborates a bit more on this finding, especially with the potential application to network-level coupling research in the future. For the ease of review, we paste the context below.

Building on this trajectory, we further characterize the formation of modularized representation and examine the connection between network topology and modularity. Our findings extend the hypothesis in \cite{achterberg2023spatially} by demonstrating that functional pressure from task co-embedding acts as a key driver for this topology, interacting with, and at times opposing, metabolic costs. Moreover, we observe that the task-induced network yields more long-distance connections, which could result from the trade-off between task representation and metabolic efficiency. This pattern of findings suggests that long-range connections in the brain might support not only small-world organization \cite{bassett2017small} but also functional integration demanded by task load \cite{vatansever2015default,nour2019task}. Ideally, the optimization of metabolic efficiency would favor more homogenous connectivity

\cite{bullmore2012economy}, while the representation efficiency prioritizes more on the heterogeneity \cite{frankle2018lottery}. Thus, the divergence of these two optimization goals offers a mechanistic explanation for the brain's modularization, distinct from models that consider only spatial constraints or representational demands in isolation. This perspective may also provide tools for quantifying the network-level coupling of structure and function, extending from communication dynamics \cite{avena2018communication,suarez2020linking} to communication representation.

Reviewer 2:

We sincerely thank you for your comprehensive review and constructive criticism. Your detailed feedback, particularly regarding the experimental rigor, variable controls, and the generalizability of our findings across different architectures, has been extremely helpful in strengthening this manuscript. Your commentary inspired us to conduct extensive new analyses, including cross-architecture validation (GRU, LSTM, CNN, GNN) and strictly controlled ablation studies. These additional experiments have significantly bolstered the reliability of our reported results. We have carefully addressed each of your points below.

Q: I find that the core methods and experimental ideas in this work are based on existing studies and lack substantial innovation. Multi-task learning, incremental training, modularity computation have all appeared in recent literature already. What I observe here is mainly a combination of these methods with minor adjustments in details, without introducing relatively new approaches in theoretical framework, algorithm design, or experimental validation.

A: We appreciate your critical perspective on the novelty of our work, and we welcome the opportunity to clarify our specific contributions.

We fully agree that the individual methodological components, including multi-task learning, incremental training, and modularity analysis, are established techniques. However, we respectfully submit that the substantial innovation of this work lies not in the invention of these tools, but in their novel synthesis to resolve a fundamental biological question: *how functional demands causally shape network topology*. While previous studies have looked at these elements in isolation, our work introduces a new theoretical and experimental paradigm that bridges the gap between dynamical systems (physical constraints) and information processing systems (computational utility). Specifically, our innovation is threefold:

1. **Scientific Discovery:** As emphasized in the *Nature Machine Intelligence* Aims & Scope, machine intelligence is a powerful tool for scientific discovery. Our work contributes a function-to-structure perspective that complements traditional structure-to-function approaches, offering a new mechanism to explain how brains evolve in adaption to task loading.
2. **NeuroAI Paradigm:** We define a specific experimental protocol that mimics evolutionary/developmental constraints. This allows us to validate that *connection-*

incremental learning yields structures that are statistically more aligned with biological brain networks than those generated by spatial constraints alone.

3. **Causal Framework:** Unlike traditional neuroimaging studies which are often correlational, we established a generative framework that provides a causal explanation for the emergence of modularity.

In summary, while the building blocks are known, the architecture we have constructed with them provides new mechanistic insights into neural organization that were previously inaccessible. We believe this integration represents a significant contribution to the emerging field of NeuroAI.

Q: The authors directly compare modularity values between networks of different scales and claim that smaller networks have higher modularity. This approach is theoretically problematic because modularity is strongly affected by the number of nodes, number of edges, and network density. Without normalization or comparison to an appropriate random graph baseline that matches size and degree distribution, such direct comparison is not meaningful. In addition, if modularity is to be used to indicate the presence of modular structure, it should be evaluated against an appropriate null model. I recommend that the authors refer to the discussion on modularity definitions and limitations in Community detection in graphs. Due to the significant misuse of the modularity index by the authors, the relevant results currently do not support their conclusions regarding modularity.

A: We appreciate your rigorous attention to the theoretical nuances of comparing modularity Q across networks of varying scales. However, we wish to clarify a misunderstanding regarding our analytical approach.

In our original manuscript, we did not intend to compare absolute modularity values across different network sizes to claim that smaller networks are inherently "more modular" in a static sense. Rather, our objective was to characterize the relative change in modularization as a function of task demand relative to network capacity. We acknowledge that the standard modularity metric is normalized by the total weight of edges $2m$, which renders it independent of the scale of edge weights, though we agree that the number of nodes remains a confounding factor in direct comparisons.

To resolve this ambiguity and rigorously validate our conclusions—specifically that modularity emerges when functional demands strain capacity—we moved beyond direct comparisons and implemented a Linear Mixed-Effects (LME) analysis. This approach allows us to statistically verify the relationship while explicitly controlling for network scale:

1. **Controlled Statistical Model:** We treated Model Size as a random effect (random intercept) to account for baseline differences in modularity caused by the number of nodes and edges. We then treated the Task-to-Dimensionality Ratio (TDR) as the fixed effect.

2. **Robust Confirmation:** This analysis revealed a significant positive relationship between TDR and modularity ($\beta = 0.027, p < 0.001$), with the random effect test confirming that model size groupings explain significant variance.

This result statistically confirms that the phenomenon we observed is **not** an artifact of comparing different network scales. Instead, it provides robust evidence that modularization is an adaptive response driven by the interaction between task load and representational capacity, regardless of the network's absolute size. We have updated the manuscript to present this rigorous verification.

Q: The authors attempt to assess whether modularity can serve as an indicator of model performance, and they observe a correlation between accuracy and modularity during the middle of training. However, this correlation disappears once the model converges. Correlation during mid-training can be caused by multiple training dynamics factors, making it difficult to directly extrapolate such a result to brain networks. In addition, for the conclusion that “task-driven adaptation enhances more brain-like structural organization,” the spatial+task method is configured such that two key parameters can both take non-zero values, whereas in the other two methods at least one of these parameters is fixed to zero. This setup inherently ensures that spatial+task has an upper bound that is not lower than the others, which weakens the rigor of this comparison.

A: We appreciate your insightful critique regarding the interpretability of mid-training correlations. We agree that stochastic training dynamics can introduce significant noise, potentially confounding the relationship between modularity and performance. To address this, we rigorously refined our experimental setup to eliminate excessive variance caused by data shuffling and initialization. Specifically, we implemented two strict controls:

Standardized Input Stream: We fixed the data generation seed to ensure all models process the training data in an identical sequence, removing divergence caused by curriculum variations.

Neutralized Initialization: We replaced the random weight initialization with a uniform matrix of ones, scaled to 0.01. This effectively neutralizes any inherent initial modular structure, ensuring that all observed modularity is purely an emergent property of the learning process.

Following these strict controls, we observed that the correlation between modularity and performance became robust and statistically significant, particularly in capacity-limited models. Furthermore, inspired by Reviewer 3’s suggestion to link modularity to the learning process rather than just the final state, we analyzed the training speed (see rightmost subplot). We found that maximum modularity strongly predicts:

Learning Efficiency: A significant negative correlation ($r = -0.28, p < 10^{-5}$) with the number of iterations required to reach 30% accuracy.

Final Performance: A strong positive correlation ($r = 0.38, p < 10^{-9}$) with final accuracy.

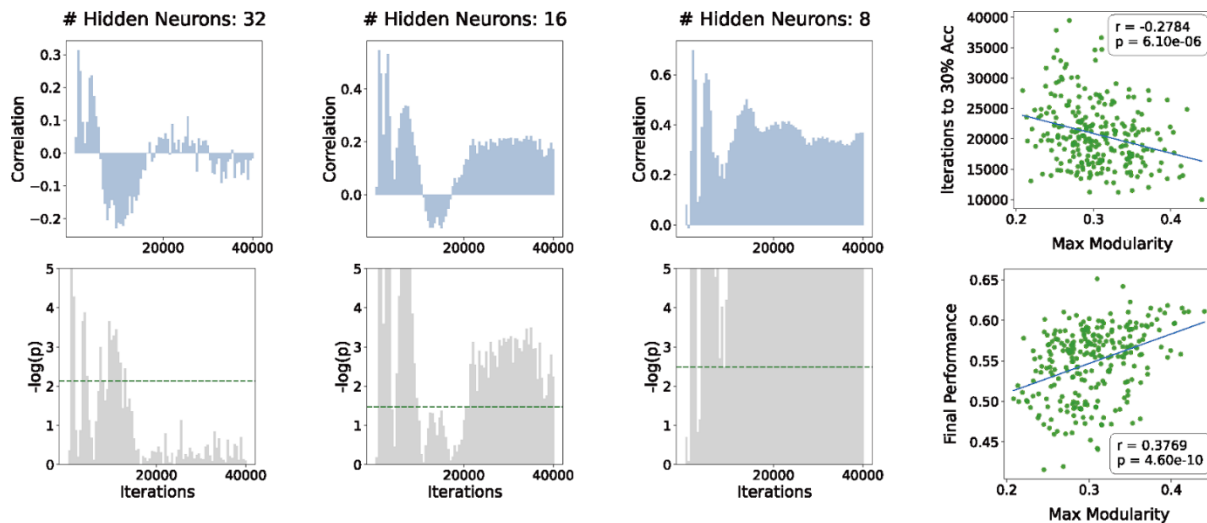


Figure 4. Modularity as an indicator of Model Performance. A significant correlation between modularity and task accuracy is observed in capacity-limited models. Each green dot represents one of 256 model instances under different seeds. The green dashed line denotes the False Discovery Rate-adjusted p-value threshold of 0.05, on a log scale.

These results suggest that modularization is not merely a transient phase but a structural adaptation that facilitates faster and more effective learning.

Regarding the comparison between the "spatial+task" model and the single-constraint models, we appreciate you raising the important point regarding the degrees of freedom. We wish to clarify two aspects of our comparison:

Normalization of Density: All generative models are constrained to produce networks with the exact same number of edges. Therefore, the "spatial+task" model does not achieve higher fitness simply by adding more connections.

Complementary Mechanisms: We acknowledge your point that the "spatial+task" model, having non-zero values for both parameters, theoretically possesses a search space that encompasses the other two models. However, our objective was not to prove that a two-parameter model fits data better than a one-parameter model (which is expected), but to demonstrate that biological brain networks cannot be fully explained by spatial constraints alone.

The significant improvement in fit provided by the "spatial+task" model confirms that the brain's topology reflects a trade-off between metabolic cost (spatial) and functional utility (task). While the "spatial+task" formulation inherently has a higher theoretical upper bound, the fact that the generated networks statistically align with empirical brain data confirms that both constraints are necessary to reproduce biological topology. We also performed a statistical test in the manuscript to verify that these differences are significant and not due to stochastic fluctuation.

Q: The experiments are conducted exclusively on a single RNN architecture without evaluation on other neural network types such as convolutional networks, graph neural networks, or alternative RNN variants. Without cross-architecture validation, the generality of the conclusions

is limited, and the results cannot be readily interpreted as evidence for brain-like mechanisms given the substantial differences between artificial RNNs and biological brain networks.

A: We appreciate this constructive suggestion regarding cross-architecture validation. To address the concern about generality, we have expanded our experimental scope and added comprehensive results involving alternative RNN variants (GRU, LSTM), as well as CNNs and GNNs.

Alternative Recurrent Architectures (GRU and LSTM). We first examined Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) networks. As these architectures share the fundamental recurrent nature of the RNNs used in our main text, they serve as a direct test of robustness. Consistent with our original findings, both GRU and LSTM models exhibited significantly induced modular structures under multi-task learning settings compared to single-task baselines. These results corroborate the phenomena observed in standard RNNs and further support the conclusions drawn in the main text.

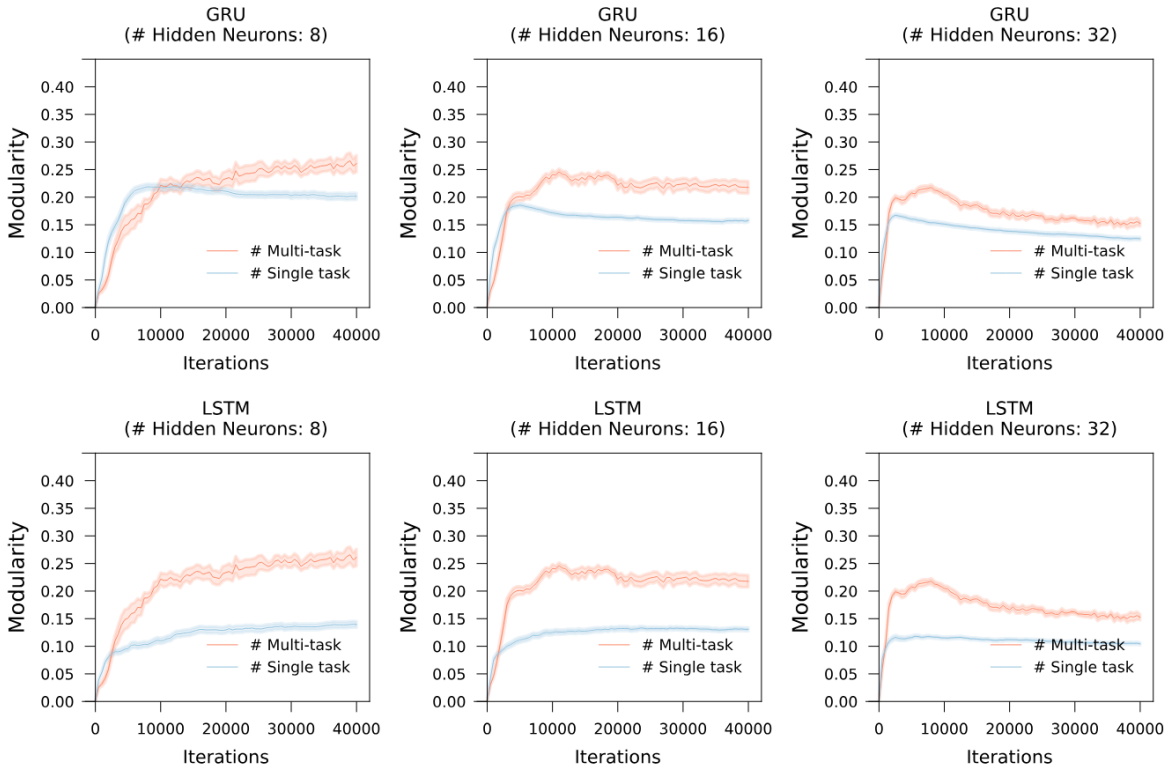


Figure 5. Community detection is applied in the recurrent layer of GRU, LSTM models. Multi-task learning can induce more modular structures than single task learning in these RNN variants models, consistent with the results of the vanilla RNN.

Feed-Forward Architectures (CNN and GNN) Regarding CNNs and GNNs, we note that their reliance on feed-forward weight structures differs substantially from the recurrent dynamics central to our study. While defining "modularity" for feed-forward weights is technically beyond the scope of this work, we found the inquiry theoretically intriguing. To facilitate a meaningful discussion, we explored two potential definitions for modularity in these architectures:

- Method A: Applying the weighted directed bipartite graph community detection algorithm, DIRT_LPA_wb_plus.

- Method B: Constructing feature matrices derived from the connectivity between convolutional or fully connected layers.

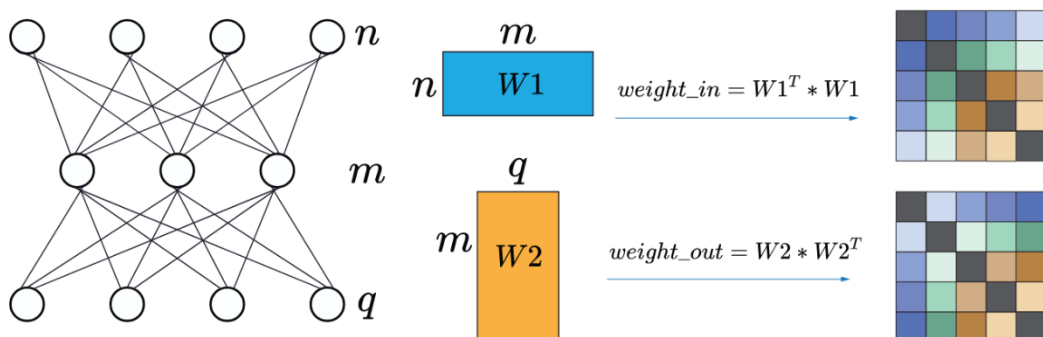


Figure 6. Schematic of constructing feature matrix (Method B). Feature matrices are derived from the connectivity between convolutional or fully connected layers. Then the community detection technique is applied to the obtained feature matrices.

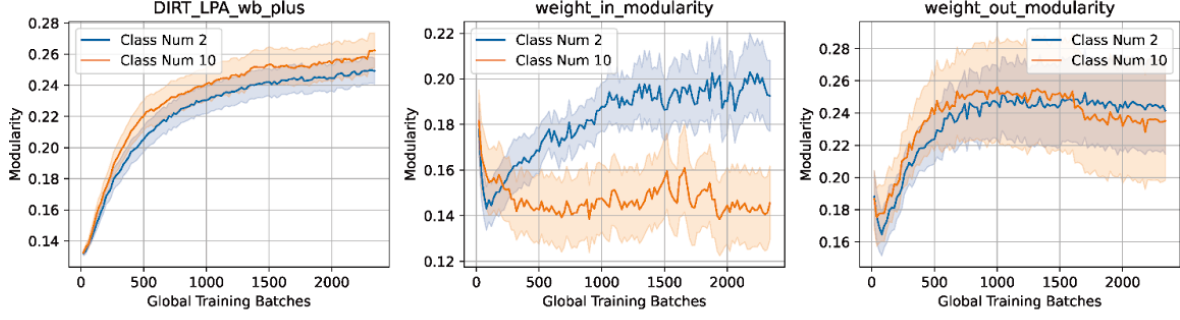
CNN Experiments: We conducted contrastive learning experiments using the CIFAR-10 dataset. To vary task complexity, we trained Siamese networks to distinguish whether two images belong to the same class, using subsets of either 2 classes (low complexity) or 10 classes (high complexity).

Analysis: We applied *Method A* (DIRT_LPA_wb_plus) to the second convolutional layer to obtain modularity values (shown in the first column of the figure below). Additionally, treating the first and second convolutional layers as W_1 and W_2 , we calculated the corresponding *weight_in* and *weight_out* matrices using *Method B* and performed community detection to derive their respective modularity scores (shown in the right columns of the Figure 7 below).

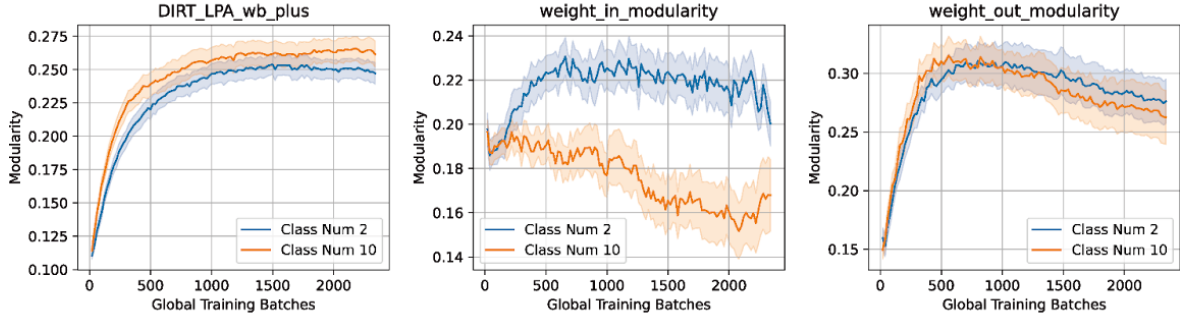
GNN Experiments: For the Graph Neural Network experiments, we utilized the Tox21 dataset, which comprises 12,060 training samples and 647 test samples of chemical compounds. Each sample is annotated with 12 binary labels indicating activity or inactivity across various toxicological assays.

Analysis: We compared **Single-Task Learning** models (predicting one toxicity label at a time) against **Multi-Task Learning** models (simultaneously predicting all 12 labels). The results, illustrated in Figure 8, clearly demonstrate the structural differences induced by these distinct learning paradigms.

Convolutional Neural Networks During Contrastive Learning
(# channel_in: 8, channel_out: 8)



Convolutional Neural Networks During Contrastive Learning
(# channel_in: 16, channel_out: 16)



Convolutional Neural Networks During Contrastive Learning
(# channel_in: 32, channel_out: 32)



Convolutional Neural Networks During Contrastive Learning
(# channel_in: 64, channel_out: 64)

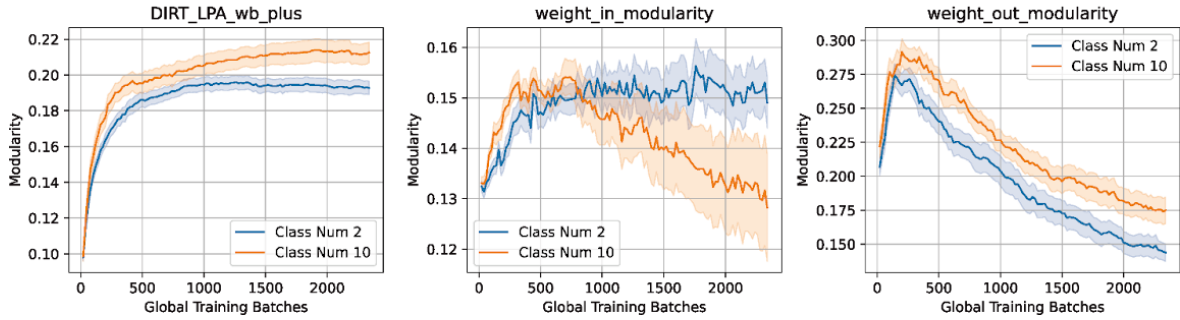
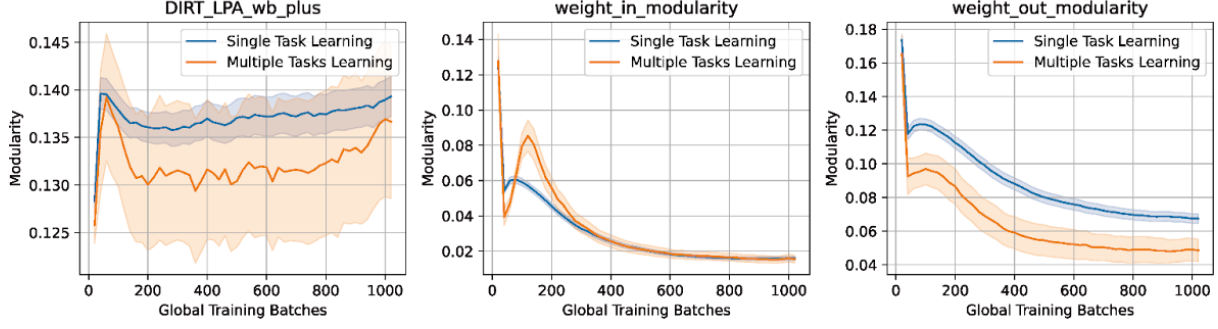
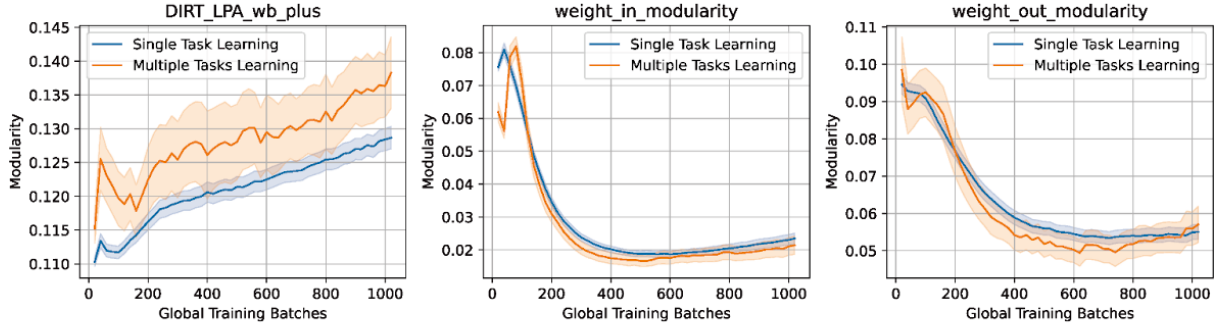


Figure 7. Community detection results of CNN models trained on contrastive learning tasks using the cifar10 dataset. Solid lines are the mean over 10 independent runs. Shaded areas indicate the standard error.

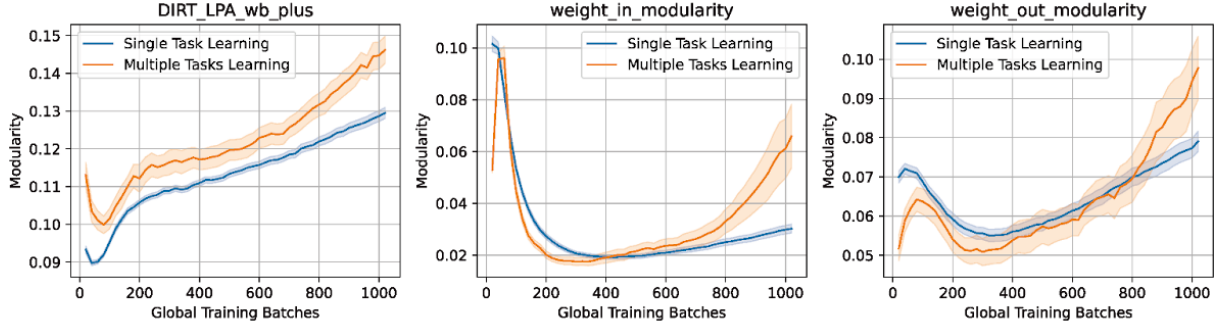
Graph Neural Networks During Training on Tox21 Dataset
(# hidden_channels: 16)



Graph Neural Networks During Training on Tox21 Dataset
(# hidden_channels: 32)



Graph Neural Networks During Training on Tox21 Dataset
(# hidden_channels: 64)



Graph Neural Networks During Training on Tox21 Dataset
(# hidden_channels: 128)

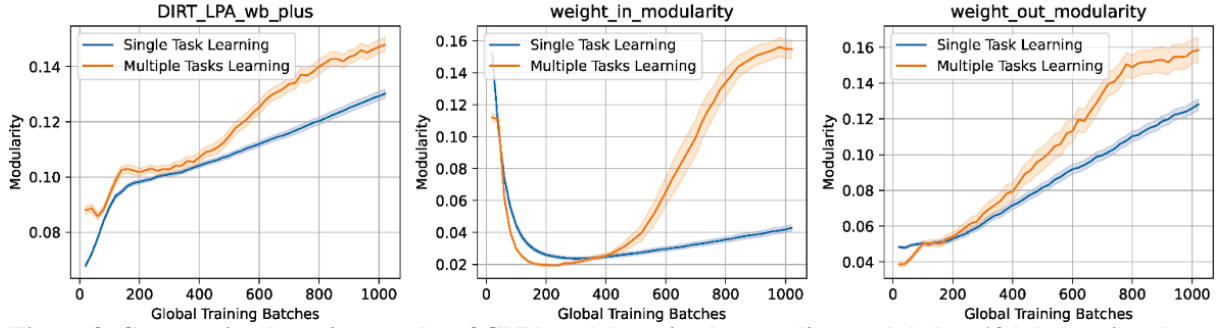


Figure 8. Community detection results of GNN models trained to predict one label or 12 labels using the Tox21 dataset. Solid lines are the mean over 10 independent runs. Shaded areas indicate the standard error.

We believe these additional experiments across diverse architectures significantly strengthen the generality of our findings.

Q: Some experiments do not control for key variables. For example, in comparing incremental multi-task and simultaneous multi-task training, it is not stated whether the total number of training samples, parameter update steps, and gradient trajectory characteristics are matched. Differences in these factors can affect the emergence of modularity and confound the interpretation of results.

A: We answer this question together with the next one.

Q: The implementation details of the connection-incremental learning procedure are insufficiently described, including the order of adding connections, the selection strategy, and the degree of randomness. There is no control experiment to verify whether the observed increase in modularity is due to the incremental addition mechanism itself or simply the result of a different training path. In addition, no ablation or control experiments are provided to separate the effect of incremental connection addition from other potential influences such as altered training dynamics. I am concerned that the design of the connection-incremental learning may not effectively capture the phenomena it is intended to represent, which raises doubts about the validity of the conclusions drawn from it.

A: We thank the reviewer for these critical questions. Since both Comment 5 and Comment 6 pertain to the experimental setup, variable controls, and implementation details regarding the connection-incremental learning in Fig. 4b, we here combined our response below to provide a comprehensive explanation.

First, we apologize for the insufficient detail regarding the implementation of these experiments in our initial manuscript. We have provided more details on the mask generation protocol in Supplementary Note 4:

To recapitulate the biological growth of brain networks, we leveraged empirical human brain data from the Human Connectome Project \cite{van2013wu} and selected a parcellation atlas of 84 cortical areas \cite{rolls2020automated}. We then extracted both the mean inter-regional distance matrix and the structural connectivity (SC) matrix. To determine the appropriate network sparsity, we analyzed binarized SC matrices from 758 participants, identifying an average of approximately 2000 structural connections. Accordingly, the final connections in the RNN recurrent layer were constrained by a 84×84 binary mask fixed at 2000 connections. The topology of this mask was derived from the inter-regional distance matrix, where the connection probability between nodes u and v is defined as $P(u,v) = E(u,v)^{-\eta}$, with $E(u,v)$ representing the Euclidean distance and η governing the decay of connection probability with distance. For the experiments presented in Fig. 4b, we set $\eta = -3.0$; additional control experiments examining connection counts, randomness levels, and addition schedules are detailed as follows.

In the original setup, we embedded a RNN with 84 recurrent nodes into a three-dimensional space. The connection selection strategy follows a probabilistic distribution where the probability of a connection forming between nodes u and v is inversely proportional to the power of the distance between them ($P_{uv} \propto d_{uv}^{-\eta}$) and for the experiments in Fig.4b, we set $\eta = -3.0$.

We fully acknowledge the reviewer's concern that the inherent differences between "incremental" and "simultaneous" training modes inevitably lead to distinct training trajectories. While it is challenging to completely eliminate the "result of a different training path," we agree that it is crucial to isolate the effect of the incremental mechanism from other confounding factors. Therefore, we have conducted a series of additional control experiments to rule out potential influences such as training dynamics and sample size.

1. **Robustness to Connection Schedules** In the main text, we added a total of 2,000 connections. To test if the specific order of addition mattered, we explored different schedules for adding connections while keeping the total number of added edges constant. We found no significant difference between the schedules:

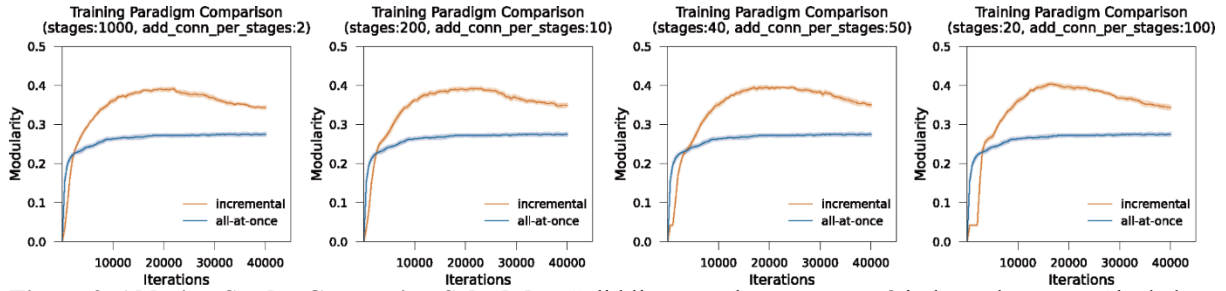


Figure 9. Ablation Study: Connection Schedules. Solid lines are the mean over 8 independent runs. Shaded areas indicate the standard error.

2. **Selection Strategy and Randomness** We investigated the connection selection strategy and the degree of randomness. Here, the parameter η controls the spatial constraint: $P_{uv} \propto (distance_{uv})^{-\eta}$. As shown in the figure below, the magnitude of η significantly influences the overall height of the modularity curves, confirming that stronger spatial constraints (lower η) correlate with higher modularity.

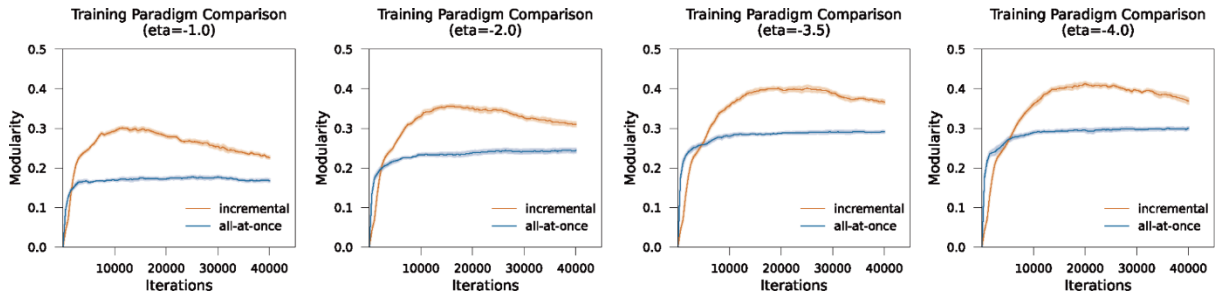


Figure 10. Ablation Study: Selection Strategy and Randomness. Solid lines are the mean over 8 independent runs. Shaded areas indicate the standard error.

3. **Impact of Total Connection Density** To verify the robustness of our conclusions, we examined the impact of the total number of connections (testing 2,000, 4,000, 5,000, 6,000, and the fully connected 7,056 edges) while holding other conditions (like η) constant. As the

total number of connections increases and the model transitions from sparse to dense, modularity decreases in both training modes. Notably, in the "simultaneous" (all-at-once) mode, modularity gradually falls below 0.1 as the network becomes fully connected. This aligns with our observation in Figure 2 for the "Model Size=64" case (which starts fully connected), effectively validating the consistency of our results.

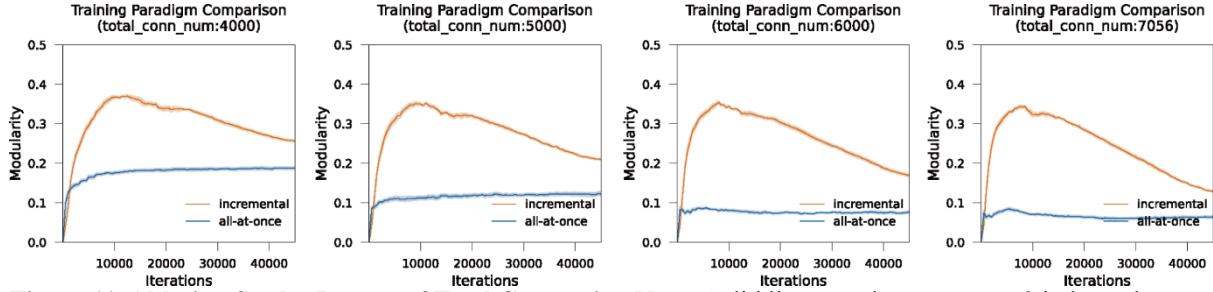


Figure 11. Ablation Study: Impact of Total Connection Num. Solid lines are the mean over 8 independent runs. Shaded areas indicate the standard error.

4. **Controlling for Training Samples and Update Steps** To address the specific concern in Question 5 regarding the matching of training samples and update steps, we performed a strictly controlled comparison. In the figure below, the vertical green dashed line marks a critical point: after this point, the models in both curves possess identical connectivity (the same number and exact set of connections). It can be seen that even when the topology and subsequent training plan are identical (post-green line), the "incremental" method maintains higher modularity. Although the gap tends to narrow over time—which is expected, as the structural connectivity are identical after the green line—the persistence of higher modularity demonstrates that the difference is inherited from the structural characteristics formed during the earlier incremental phase, rather than being a byproduct of mismatched sample sizes or update counts.

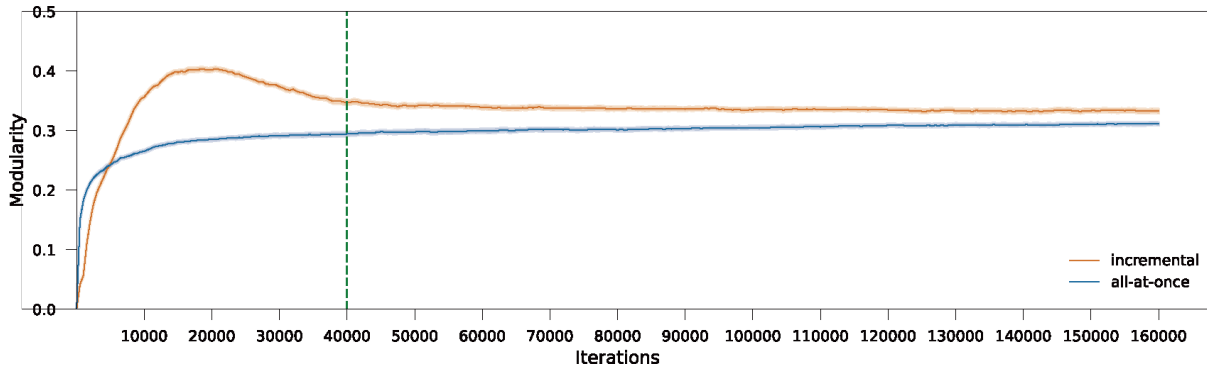


Figure 12. Ablation Study: Controlling for Training Samples and Update Steps. Solid lines are the mean over 8 independent runs. Shaded areas indicate the standard error.

These results have been added to the Supplementary Figure 8 to separate the effect of incremental connection addition from other potential influences for readers.

Reviewer 3:

Thank you for your careful review and constructive suggestions. We are greatly encouraged by your appreciation of our general approach, the potential of our analyses, and the relevance of the topic to the community. Your detailed comments, particularly those concerning the interaction between network sizes and tasks, the need to quantify representational capacity, the nuanced relationship between modularity and accuracy during learning, and your suggestions on framing prior literature, have profoundly inspired us to conduct deeper investigations and perform additional analyses. We believe this supplemental work will significantly improve the rigor and reliability of our reported results and lead to a more comprehensive manuscript. We will now address each of your specific concerns point-by-point below.

Q: Interaction between size and tasks: Across the manuscript you find very interesting interaction effects between size of network and tasks and you hypothesize that modularity only develops when networks are at their 'representational capacity' – this is a key point across your manuscript as you obviously need to explain why modularity would only sometimes develop in the light of multi-task learning. Given how central this argument is to the paper, I would expect that you at least have one plot or analysis (e.g. via ANOVA-like analysis) that shows that this effect is actually true. For this you would need to directly measure the representational capacity -- as far as I can see you do not directly measure representational capacity here yet. I take it that you see number of neurons as a valid proxy for representational ability, but that is actually a strong assumption, given that the number of weights in an RNN obviously does not scale linearly with the size of the RNN. Maybe a measure like effective dimensionality of the weight matrix of latent activation might be more appropriate, or at the very least the number of parameters in the network.

• Optional: If you wanted to be 100% rigorous about this, you would not only try to quantify the representational capacity of the networks but also the representational complexity of the tasks, but as I myself do not have a suggestion for how exactly one would do that, I want to leave this as an optional comment. I would certainly welcome ideas from the authors. Maybe there is a measure in here that would be helpful <https://arxiv.org/abs/1808.03591> ?

• As a subpoint to the above, I then suggest that you use this representational capacity to color lines, such as in Fig.2 e-h, so that one can actually see how the representational capacity interact with the modularity. The bi-directional color scale with blues and greens makes these Figures quite difficult to interpret.

A: We sincerely thank you for your insightful and constructive suggestions. Your comments correctly identified that the interaction between network size and task load is a central claim of our manuscript, and the critique was threefold:

1. We needed a more direct, statistical analysis to prove this interaction rather than just hypothesize it.
2. Our proxy for "representational capacity" (i.e., "# Hidden Neurons") was a strong assumption, and better measures like "number of parameters" or "effective dimensionality" should be used.
3. Our original Figure 2(e-h) was difficult to interpret.

These points were extremely well-taken and have led to a significant improvement in the paper.

Firstly, we replot our original Fig. 2(e-h) with a clearer color scale and add an annotation for recurrent layer parameters. Due to space limitations, we replaced one of the subplots in Fig. 2(e-h) with an ANOVA-like analysis that will be discussed below.

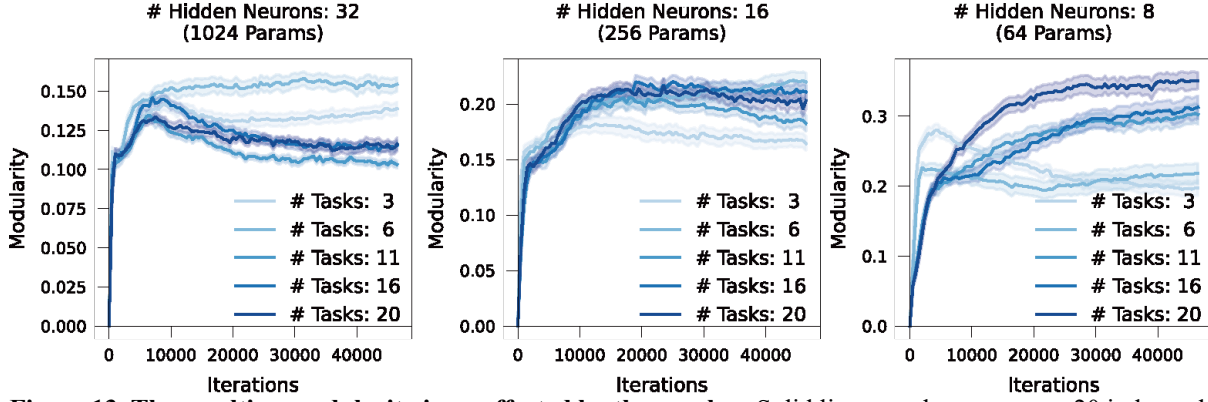


Figure 13. The resulting modularity is co-affected by the number. Solid lines are the mean over 20 independent runs. Shaded areas indicate the standard error.

Secondly, to provide a direct statistical validation of the interaction, we followed your sophisticated suggestion to measure "representational capacity" using the Effective Dimensionality (ED) of the latent activations calculated via Principal Component Analysis following the below steps:

1. For each of our 20 experimental conditions (e.g., 8-neuron/3-task, ..., 8-neuron/20-task, ..., 64-neuron/20-task), we ran the trained network on its respective tasks and collected its hidden state activations and organized them into a matrix.
2. Based on the above matrix, we applied the Principal Component Analysis.
3. We defined the Effective Dimensionality (ED) as $ED = (\sum_i \lambda_i)^2 / (\sum_i \lambda_i^2)$, where λ_i represents the eigenvalues of the covariance matrix derived from PCA.
4. We defined a new metric, "Task-to-Dimensionality Ratio" = (Number of Tasks / Effective Dimensionality).
5. Based on that, we created a new scatter plot of this Task-to-Dimensionality Ratio (X-axis) vs. Final Modularity (Y-axis). Each point on the plot is a single trained network (seed).

To validate this statistically and to account for the distinct clustering of different model sizes, we performed a Linear Mixed-Effects (LME) analysis using the lme4 package, treating the Task-to-Dimensionality Ratio as a fixed effect and the Model Size as a random effect (random intercepts). This allows us to test the relationship while controlling for intrinsic scale differences between networks. We found a significant positive relationship between Task-to-Dimensionality Ratio and Modularity ($\beta = 0.027, p < 0.001$). Furthermore, we utilized the *exactRLRT()* function from the RLRsim package to test the significance of the random effects, confirming that model size groupings explain significant variance ($p < 0.001$) and justifying the use of the mixed model. This analysis confirms our hypothesis: Modularity is not simply a function of fixed network size but emerges specifically when the task load approaches or exceeds the network's effective representational capacity.

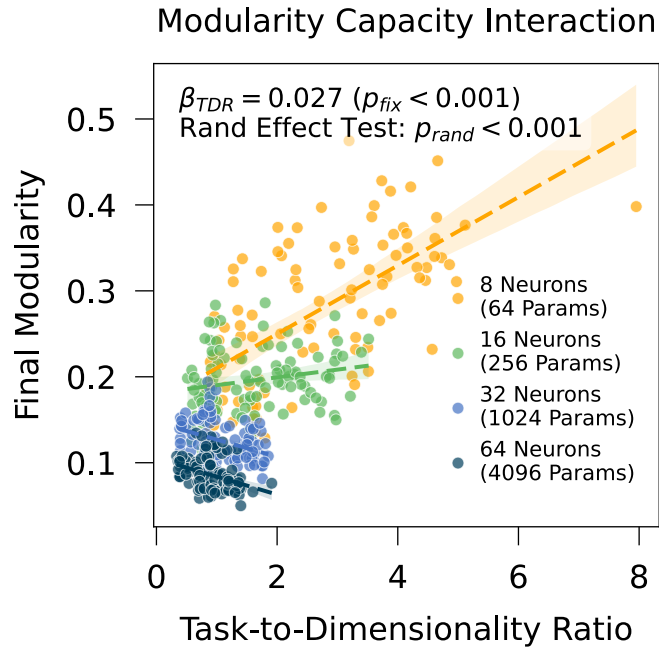


Figure 14. Modularity increases as the task-to-dimensionality ratio grows. Each dot represents one of the model instances of 4 capacity sizes and trained under 5 different number of tasks (e.g., 8-neuron/3-task, ..., 8-neuron/20-task, ..., 64-neuron/20-task).

Q: Fig 3d: This is an extremely interesting analysis but I do not understand how you arrive at your conclusion “Together with the findings in Fig. 2, these results suggest that when representational capacity is constrained, modularization emerges as a compensatory procedure to improve the neural efficiency in representation.” given the fact that the modularity falls below significant at the end of training for all network groups, supposedly when accuracy is the largest? So, would the appropriate conclusion to this analysis not be that modularity is not directly linked to accuracy but instead somehow to learning? That seems to be a very interesting new angle, as modularity and compositionality are obviously supposed to help with finding new solutions. Perhaps this could be cleared up if you correlated ‘how well learning is going to go’ (so end of training accuracy) with how much modularity was achieved during learning? You might not want to look into the learning effect (that is fine by me) but I do not see how your current claim comes out of your analyses.

A: We greatly appreciate your insightful critique regarding the interpretation of Fig. 3d. This inspired us to thoroughly re-examine and refined the experimental procedures for the data presented in Fig. 3d. Our initial analysis was indeed complicated by two major sources of high variance:

Stochastic Training Trajectories: The random generation of training data for each batch caused the training trajectories of different models to diverge significantly.

Initial Structural Bias: The use of random orthogonal initialization for the recurrent connections, (even when scaled down by 0.1), introduced an initial stochastic structure in the weight matrix, which potentially masked the modular structure we sought to study.

To mitigate these confounding factors and isolate the relationship between modularity and task learning, we implemented the following critical adjustments:

Fixed Input Sequence: We standardized the input stream, ensuring all models read the same dataset in a fixed sequence.

Neutralized Initial Structure: We replaced the random orthogonal initialization for recurrent connections with all-ones initialization, scaled to 0.01. This effectively eliminates any inherent initial modular structure, allowing the emergent, learned modularity to be clearly observed. We switched the activation function from ReLU to tanh due to numerical stability issues encountered with the all-ones initialization.

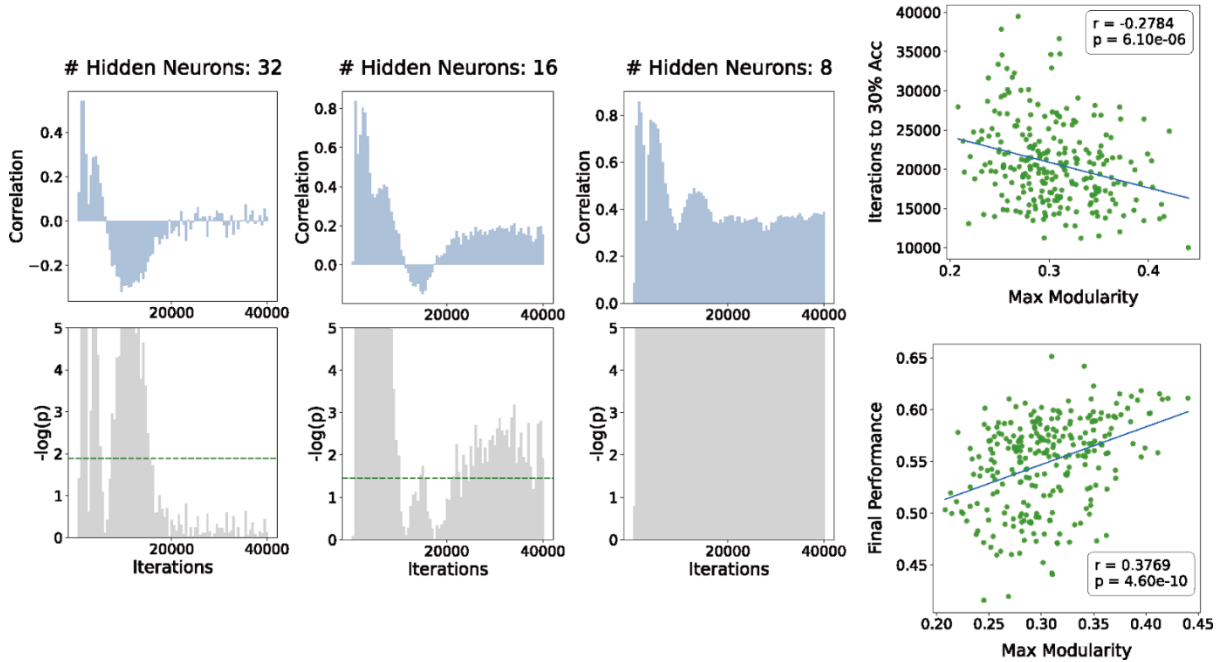


Figure 15. A significant correlation between modularity and task accuracy is observed in capacity-limited models. Each green dot represents one of 256 model instances under different seeds. The green dashed line denotes the False Discovery Rate-adjusted p-value threshold of 0.05, on a log scale.

Following these methodological improvements, we observed that the correlation between the network modularity and its task performance (test accuracy) improved significantly. Furthermore, **inspired by your hypothesis** that "modularity is somehow linked to learning", we perform the verification shown in the rightmost subplot (models with 8 hidden neurons). We found that the maximum modularity reached during training is significantly correlated with learning efficiency: it shows a negative correlation with the number of iterations required to reach 30% accuracy ($r = -0.28, p < 10^{-5}$), indicating faster acquisition of the tasks. Additionally, it exhibits a strong, positive correlation with the model's final performance ($r = 0.38, p < 10^{-9}$). These results strongly suggest a fundamental link between modularity and the overall learning trajectory. We have updated the discussion of Fig. 3d in the revised manuscript to reflect this robust interpretation.

Q: Line 235-238; “The network initialized with the community mask exhibited superior performance over those marked by the random mask (Fig. 4e). Notably, this performance advantage was especially pronounced in models with smaller capacity, consistent with earlier results shown in Fig. 1 and Fig. 2. “. I do not agree with your interpretation, as the community mask only helped in smaller networks. Your wording “especially pronounced” would only be correct if all network sizes showed this effect, which does not seem to be true?

A: We fully agree with your assessment and appreciate you correcting our interpretation. You are correct that the wording "especially pronounced" was imprecise, as it implied a universal benefit that merely varied in magnitude. As our results demonstrate, the performance advantage of the community mask is strictly tied to capacity constraints and diminishes in larger networks where over-parameterization provides sufficient robustness. We have revised the manuscript to explicitly state that this structural advantage is specific to capacity-limited settings.

Below, we detail the updated experimental setup and results included in the revised Fig. 4e.

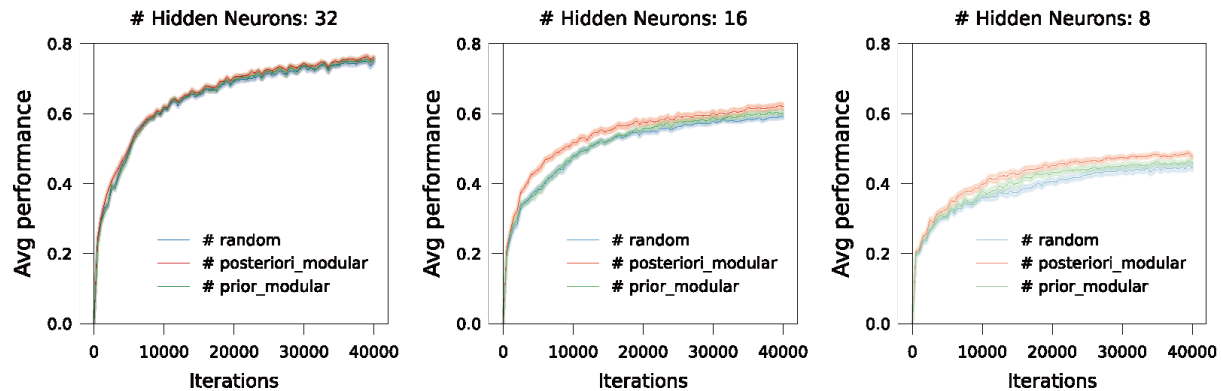


Figure 16. Performance of models with three types of masks. Solid lines are the mean over 20 independent runs. Shaded areas indicate the standard error.

Experimental Setup: We introduced a binary mask (0/1) to the recurrent connections of the RNN, where a value of 1 indicates a selected connection that participates in the computation, and 0 indicates a zeroed-out connection.

We defined three distinct groups/masks, all with the same number of connections:

- 1) The *prior modular* mask: This mask was manually generated based on a core-periphery organization. Connections were sampled from a distribution based on edge type (e.g., intra_core_prob = 0.8, inter_core = 0.1, core_periphery_prob: 0.08, intra_periphery_prob: 0.02).
- 2) The *posterior modular* mask: This mask was derived from the model after multi-task learning. We identified the emergent modular structure with the maximum modularity during training, and then we preserved all the intra-community connections and the top 5% high-magnitude inter-community connections, thereby maintaining the interactions between different modules.
- 3) The *random* mask: Connections were sampled uniformly at random.

As shown in the revised figure, the *posteriori*_modular mask (red curve) demonstrates a significant performance advantage over the *random* mask (blue curve) in the 8-neuron and 16-neuron models. However, consistent with your observation, this advantage disappears in the 32-neuron model, where the curves converge. This confirms that the emergent modular structure serves as a "winning lottery ticket" primarily when the network lacks the capacity to brute-force a solution. We have updated the text in the revised manuscript to accurately reflect this capacity-dependent phenomenon.

Q: I am not sure I fully agree with your framing of the prior literature on two accounts:

(1) I think you should make clearer that on the level of representations, the link between multi-task learning and modularity as been made by multiple prior papers (sometimes calling them composable, or abstract), such Yang Nature Neuro 2019 (as you cite) but also more recently <https://www.nature.com/articles/s41467-023-36583-0> or <https://www.nature.com/articles/s41593-024-01668-6> -- you of course add the structural component, which is valuable in itself, but I think it would be fair to acknowledge this
(2) At two points (lines 43-46 and 271-273) in the manuscript you make it sound as if the prior literature suggested that modularity was solely driven by metabolic constrains and that you now add that it is driven by metabolic constrains and tasks (as discussed in section 241). I think this is a misrepresentation of prior research. Achterberg Nat Mach Int (as you cite) already specifically discusses the interaction of tasks + metabolic costs for driving modularity (note lower modularity for non-task solving networks on their paper) and also already discuss / hypothesise that this effect might be even larger in multi-task environments that allow for more modular learning (not tested by them). I think your analyses nicely confirm these hypothesised links, which is a very good contribution, but I think it would be better to acknowledge that they follow from prior work.

A: Thank you for your suggestions. Combined with the comments from Reviewer 1, we add additional discussion on the relationship between the current work and related work. For the ease of review, we paste the content below and marked the related part in blue:

Combining such two perspectives sheds light on interpretable approaches for both biological and artificial neural networks. In the context of multi-task learning, both cognitive neuroscience \cite{cole2011rapid,cole2013multi} and RNN-based task modeling \cite{yang2019task} have recognized the compositional nature of task representation, suggesting the possibility of clustering wiring patterns when we follow the dogma from Donald Hebb that \emph{``Neurons that fire together, wire together''} \cite{lowel1992selection} or take it as a pattern that echoes the task-cortical mappings analogous to the well-studied retino-cortical map \cite{st2023brain,li2024enhancing}. Subsequent work has verified that the learning of multiple tasks can induce wiring phenotypes termed abstract representations \cite{johnston2023abstract} and shared dynamical motifs \cite{driscoll2024flexible}. Building on this trajectory, we further characterize the formation of modularized representation and examine the connection between network topology and modularity. Our findings extend the hypothesis in \cite{achterberg2023spatially} by demonstrating that functional pressure from task co-embedding acts as a key driver for this topology, interacting with, and at times opposing, metabolic costs. Moreover, we observe that the task-induced network yields more long-distance connections, which could result from the trade-off between task representation and metabolic

efficiency. This pattern of findings suggests that long-range connections in the brain might support not only small-world organization \cite{bassett2017small} but also functional integration demanded by task load \cite{vatansever2015default,nour2019task}. Ideally, the optimization of metabolic efficiency would favor more homogenous connectivity \cite{bullmore2012economy}, while the representation efficiency prioritizes more on the heterogeneity \cite{frankle2018lottery}. Thus, the divergence of these two optimization goals offers a mechanistic explanation for the brain's modularization, distinct from models that consider only spatial constraints or representational demands in isolation. This perspective may also provide tools for quantifying the network-level coupling of structure and function, extending from communication dynamics \cite{avena2018communication,suarez2020linking} to communication representation.

Q: Section starting line 208: This is an interesting analysis but I am a bit concerned that this the same effect that Yang et al 2019 already show. They show their modular representation are achieved through a method which influences how much different synapses are updated as function of how important they are for a given task (first described in Zenke et al, "Continual learning through synaptic intelligence"). So while they do not add new connections, but they de facto end up training different groups of connections in the network over time, which intuitively seems similar. One of the complications with pinning down a possible link right now is that you do not discuss which optimiser you use. If you use a momentum based optimiser, depending on how you fix the recurrent weights, it might be that you carry the prior momentum forward across batches when new tasks are added and that might end up achieving an effect where specific tasks are optimised through specific weights, even if you do not hard code this. Another effect could be that the absolute weights of newly added neurons have a different magnitude than the prior ones, which will again influence how they are changed by the optimiser. I am not against this analysis per se, but I am somewhat concerned that what we are seeing here is perhaps a side effect of the optimiser and not directly of the adding weights, as you describe.

A: Thank you for this sophisticated and critical observation regarding the interplay between optimizer dynamics and structural evolution. We agree that momentum-based optimizers (like Adam) could theoretically mimic "synaptic intelligence" mechanisms by carrying historical gradients forward, potentially confounding the effects of adding connections. To address this, we performed specific control experiments to isolate the structural effect from optimizer artifacts and clarified the distinction between our approach and existing literature.

Controlling for Optimizer Dynamics (Momentum): You correctly noted that if the optimizer's state carries over when new connections are added, the effect might stem from momentum rather than the architecture itself. To test this, we attempted to switch to Stochastic Gradient Descent (SGD), which lacks momentum. However, as is common with complex recurrent architectures, the models failed to converge under pure SGD. To resolve this while retaining trainability, we implemented a "State-Reset" protocol using the Adam optimizer. Specifically, we completely re-instantiated the optimizer (resetting all internal momentum and variance states) at every stage where new connections were added. This approach effectively eliminates the "history" or "synaptic memory" carried by the optimizer.

As shown in Figure 17 as well as the response above to Reviewer 2, the connection-incremental method (orange curve) consistently achieves higher modularity than the all-at-once baseline (blue curve), even when the optimizer is reset repeatedly. Furthermore, this advantage remains robust across different addition schedules (adding connections every 20, 40, 200, or 1000 stages), proving that the modularization is a result of the structural growth strategy, not an artifact of the optimization schedule.

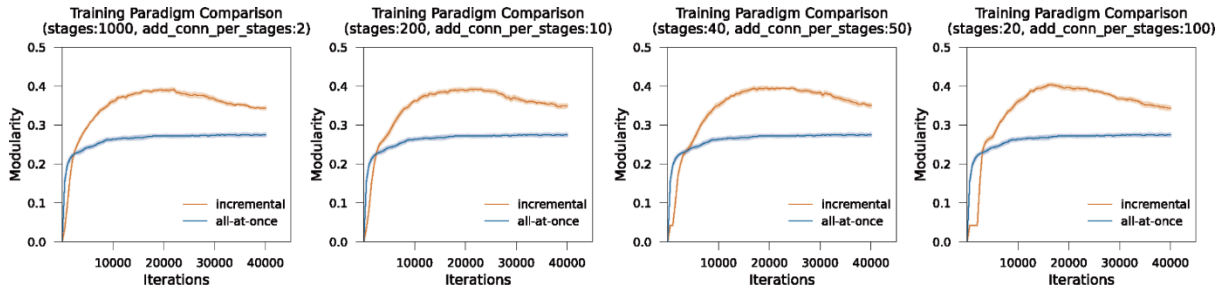


Figure 17. Ablation Study: Connection Schedules. Solid lines are the mean over 8 independent runs. Shaded areas indicate the standard error.

Distinction from Yang et al. and Zenke et al. We appreciate the suggestion to include the reference to Zenke et al. *Continual Learning Through Synaptic Intelligence*, which we have now added to the reference. However, we wish to clarify a fundamental distinction between our work and the frameworks of these two works. Their studies focus on continual learning, where tasks are introduced sequentially. Their methods protect important weights to prevent catastrophic forgetting of previous tasks. In this specific experiment, the tasks are learned simultaneously (Multi-task Learning) and the connectivity is incremental. Our results demonstrate that gradually introducing synapses drives the network toward a modular topology. This is a structural adaptation mechanism distinct from the synaptic consolidation weight-protection mechanisms described in prior literature. We have expanded the Discussion section to explicitly frame this contribution within the context of these important prior works.

Q: Line 204; “In contrast, when the model’s capacity is sufficiently large, the model exhibits a certain degree of over-parameterization, endowing it with inherent robustness. In the late iterations of training, further perturbations in the weights have minimal impact on task performance, leading to a weakened correlation between modularity and performance.”: Do you have any analyses to back up your claim about perturbation robustness? This is the first time (and it seems to me, the only time) that you mention perturbations and you do not directly refer to any analyses here.

A: Thank you for highlighting this point. The original statement in this section was indeed not sufficiently rigorous. We have now included a detailed perturbation analysis in the Supplementary Figure 9 to substantiate our claim. This analysis is specifically designed to quantify the weight perturbation robustness of the large-capacity (over-parameterized) model relative to smaller-capacity models.

Perturbation Analysis Setup:

1. **Perturbation Method:** We conducted the perturbation by adding Gaussian white noise ($\mathcal{N}$) to the recurrent connection weight matrix (W_{rec}), simulating the effect of various random errors or noise that the model might encounter.
2. **Strength Control (SNR):** To precisely control the perturbation intensity, we utilized the Signal-to-Noise Ratio (SNR), a scientific and standard technique in signal processing. This approach links the noise power to the power of the original weights (the signal), ensuring that the perturbation strength is independent of the model's intrinsic scale. This allows for a fair comparison of robustness across models of different capacities.
3. **Calculation:** We define the signal power (P_{signal}) as the total energy (sum of squares) of the original weight matrix, $P_{signal} = \sum W_{rec}^2$. For a given SNR (in dB), the required noise power P_{noise} is calculated as:

$$P_{noise} = P_{signal} / 10^{(SNR_{dB}/10)}$$

The random Gaussian noise is then scaled to possess exactly the required P_{noise} .

4. **Evaluation Metric:** We measure the incremental change in task loss ($\Delta\text{Loss} = \text{Loss}_{\text{perturbed}} - \text{Loss}_{\text{original}}$) after the perturbation. A smaller ΔLoss indicates greater robustness to that level of perturbation.

The figure below clearly demonstrates the following trend: for a given training step and SNR perturbation, models with larger capacity (larger N_{RNN}) show a significantly lower variance in ΔLoss compared to smaller-capacity models. Furthermore, the median ΔLoss is also lower in most cases for the larger models. This robustly supports our argument that a certain degree of over-parameterization endows the model with inherent robustness.

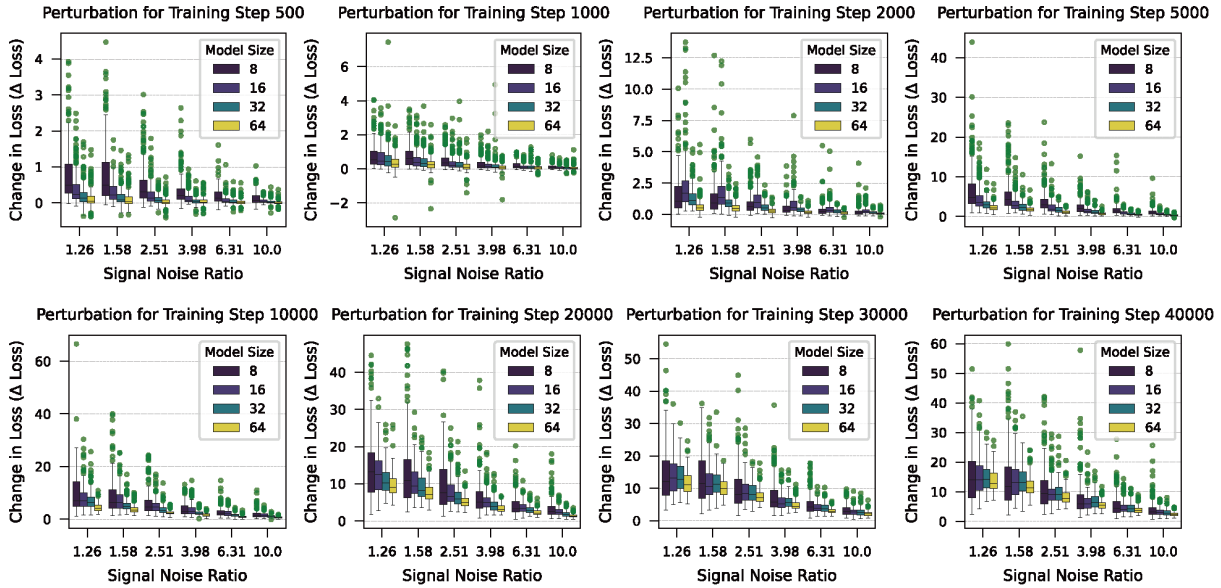


Figure 18. Larger models exhibit a certain degree of over-parameterization, endowing it with inherent robustness, where further perturbations in the weights have less impact on task performance.

Q: Some important methods details, such as used optimiser (also see discussion of optimiser above), learning rate, initialisations, activation function(do you use Relu? In line 337 you just say “e.g. ReLU”) are missing from the methods.

A: Thank you for your notice. We have added more details (highlighted in blue) about the training methods in the start of the Method section. For the ease of review, we paste it below:

Training and Initialization Protocol. We used the Adam optimizer with a learning rate of 0.001, maintaining default hyperparameters. Unless otherwise noted, recurrent weights were initialized orthogonally, while input and readout weights followed the default initialization. Crucially, to isolate the effects of task-driven plasticity from initial random connectivity, we scaled the initialized recurrent weights by a factor of 0.1. This scaling dampens the influence of the initial reservoir, ensuring that the final topology primarily reflects task-specific adaptation (see Supplementary Fig. 3 for control experiments). Specially for the experiment in Fig. 3d, we employed a constant initialization (all weights set to 1) for the recurrent layer, scaled by a factor of 0.01. Furthermore, to prevent activation saturation and numerical instability associated with unnormalized constant initialization in ReLU networks, we replaced the ReLU activation with the hyperbolic tangent (Tanh) function for these specific trials.

Q: Line 134-140 / Fig 2; e-h: I think you need to improve the writing to first clearly describe the results you observe and then include your more interpretative description afterwards. I find it confusing that you do not at all describe the results here (you seem to get quite an interesting interaction of number of tasks and model size?) and instead say “When the numbers of hidden neurons are sufficiently large (e.g., ≥ 32 for the current setup), it is almost possible to represent three tasks in separate, non-overlapping populations. Thus, incorporating more tasks may not result in a more significant modularization given Eqn. 2.”. This is not a description of the results but instead a speculation about representational ability and the resulting modularity.

A: Thank you for your suggestions on the writing. Based on your suggestions and the revised Figure 2, we update the referred paragraph as follows:

To obtain a more comprehensive understanding of modularization under different balances of model capacity and task load, we varied the parameters of task number and hidden neuron number at a finer granularity (Fig. 2e-g). We observed a distinct interaction between network capacity and task load: for capacity-constrained networks (e.g., 8 hidden neurons; Fig. 2g), increasing the task count from 3 to 20 resulted in a substantial increase in modularity. Conversely, larger networks (e.g., 32 hidden neurons; Fig. 2e) exhibited relatively stable modularity levels regardless of the number of tasks. To quantify this relationship, we measured the Task-to-Dimensionality Ratio (TDR), which is defined as the ratio of task number to the effective dimensionality of the latent space (Fig. 2h). We found a positive correlation between TDR and final modularity ($p < 1.0 \times 10^{-25}$), suggesting that modularization emerges as an adaptive response primarily when functional demands outstrip the network's available capacity. This finding implies that when hidden neurons are abundant, tasks can be represented in non-overlapping populations without necessitating structural segregation. However, as representational demand increases, the network tends to adopt a modular topology to efficiently manage multiple tasks.

Q: Line 106/107, “Consistent with [9], we observe the similar phenomenon of co-embedding functionally similar tasks for reducing the demand of hidden nodes.”: Where do you show this?

This does not seem to link to a Figure? Do you just deduct this from the accuracy numbers, as I do not see any analysis of representations, as I believe ref.9 used to make that point?

A: We thank the reviewer for highlighting this confusion. In this specific passage, we were inferring the presence of co-embedding based on the network's ability to maintain high performance despite a drastically reduced parameter count (efficiency). However, you are absolutely correct that direct evidence requires an analysis of the representations themselves.

To address this, we have rephrased the sentence to clarify that the efficiency results suggest this phenomenon, and we explicitly added a forward reference to Figure 3, where we provide the detailed representational analysis (task overlaps and RSA) that you were looking for.

We observe a reduction in the demand for hidden nodes, suggesting a co-embedding of functionally similar tasks. Indeed, a single model with a sufficient number of hidden units (e.g., # Hidden Neurons = 32) trained under multi-task learning (Fig. 2b) achieves comparable average performance to separate models, each with a small number of hidden units (e.g., # Hidden Neurons = 8 for 20 tasks, totaling 160 units) trained under single-task learning (Fig. 2a). This efficiency implies that the network compresses multiple tasks into shared low-dimensional manifolds. We thus provide a direct analysis of these shared representations and their geometric relationships in the subsequent section.

Q: Line 144: You discuss co-embedding here but you have not yet studied the actual representations. Generally speaking, it would make it easier to follow your writing if you used less interpretative language and instead stated the variable names you have actually analysed so far — so not co-embedding but network size and number of tasks.

A: We fully accept this constructive suggestion. We agree that referring to "co-embedding" before presenting the actual representational analysis could be confusing. In the revised manuscript, we have updated the text to strictly describe the observed empirical variables, network size and number of tasks, throughout the sections discussing Fig. 1 and Fig. 2. We now reserve the interpretive discussion of "co-embedding" for the subsequent section (Fig. 3), where we directly analyze the neuronal representations.

Accordingly, we have rewritten the specific sentence you highlighted to serve as a clear transition from the performance metrics to the representational analysis:

To further elucidate how task relationships underlie the interaction between task count and model capacity, we illustrate three typical phenotypes based on the overlap of embedding neurons in the hidden layer (Fig. 3a).

Q: I understand that it is difficult to fit methods details into the main text given the word limit, but I think it would be extremely helpful to know how many networks you train for which analysis, as part of the main text, e.g. sections starting line 92 and line 113 do not seem to mention that piece of information.

A: We appreciate you highlighting the need for explicit sample size details in the main text. You are absolutely correct that documenting the exact number of trained models is essential for reproducibility and interpretation. We have updated the relevant sections of the manuscript (specifically the paragraphs corresponding to lines 92 and 113) to explicitly state the sample sizes (N) for each analysis. For clarity, the breakdown is as follows:

Figure 2a (Single-Task Training): We independently trained 20 networks for each of the 20 tasks, across each of the 4 model sizes. This resulted in a total of $N = 20 \text{ tasks} \times 4 \text{ sizes} \times 20 \text{ networks} = 1,600 \text{ networks}$.

Figure 2b (Multi-Task Training): For simultaneous multi-task learning (all 20 tasks), we independently trained 20 networks for each of the 4 model sizes. This totaled $N = 4 \text{ sizes} \times 20 \text{ networks} = 80 \text{ networks}$.

Figure 2d: The data for this figure is drawn from the set of single-task and multi-task models generated for Fig. 2a and 2b, respectively.

Figures 2e-2h: In the original submission, we examined 5 different task sets. For each of these 5 task sets and across each of the 4 model sizes, we independently trained 20 networks. This resulted in $N = 5 \text{ sets} \times 4 \text{ sizes} \times 20 \text{ networks} = 400 \text{ networks}$. (Note: Since the all-20-tasks set used in Figure 2b is one of these 5 sets, this represents an addition of 4 new task sets to the analysis.)

These details are now fully integrated into the revised text to ensure the methodological foundation is transparent.

Q: Fig 2: Please use thicker legend lines, the colours are hard to read

A: Thanks for your helpful suggestion! We have updated the figures in the latest version.

Q: Line 177-179; “Inspired by representational similarity analysis (RSA) [26], we introduce a metric to quantify neuronal similarity based on activation patterns during task execution, thereby enabling a fine-grained assessment of shared versus task-specific neural involvement.”: I think it is extreme to say that you introduce a new metric here, you just develop a specific way of analysing representations, as many electrophysiology and MechInterp papers do. So I would not say that you develop a new metric.

A: Thank you for pointing out this. We have updated the sentence as

Inspired by representational similarity analysis (RSA), we develop a tailored version based on activation patterns during task execution, thereby enabling a fine-grained assessment of shared versus task-specific neural involvement.

Q: Section starting line 175: So which specific network is analysed here and on which tasks is this trained? It might be interesting to understand how this developing modularity of representations develops alongside the task accuracy, but I am unsure which accuracy data would be linked to this specific network. I am actually a bit confused because assuming that what you show here is the neuron to neuron matrix of the recurrent units, these matrices seems to have 128 neurons even though before the largest one was 64? Maybe I am misunderstanding this, but if

these are indeed larger networks, what does the modularity curve for these networks look like? Are they still modular?

A: We appreciate you raising this important question regarding the specific network configuration used in the representational analysis. Clarity here is vital to accurately link the visual evidence (matrices) to the quantitative results (accuracy/modularity curves).

The network analyzed in this section is a single model instance trained under the exact same protocol as Fig. 2b. Specifically, it was trained via simultaneous multi-task learning on all 20 cognitive tasks for over 40,000 iterations. You correctly identified that the matrix in the original figure displayed 128 neurons, despite our quantitative analyses focusing on a maximum size of 64. We apologize for this confusion. The 128-neuron model was originally selected for this visualization purely to provide a higher-resolution depiction of the cluster formation, avoiding the pixelation inherent in smaller matrices. To ensure strict consistency between our visual illustrations and our quantitative data, we have replaced the figure in the revised manuscript. The updated figure now displays the weight matrix of a 64-neuron model drawn directly from the pool of models analyzed in Fig. 2b. This ensures that the representational modularity shown corresponds directly to the accuracy and modularity curves presented earlier.

Q: Fig 2 e-h: I feel the comparison across network sizes is difficult for the reader here as y-axis values are varying across plots

A: We appreciate your suggestion regarding the harmonization of y-axes across Figure 2e–h. We carefully considered this approach; however, we opted to maintain variable scales to maximize the interpretability of the results for each distinct model size.

Our rationale is twofold: First, for the theoretical validity, as noted by Reviewer 2, raw modularity values are intrinsically dependent on network properties like node and edge count. Therefore, a direct comparison of absolute modularity magnitudes across different network sizes is not theoretically meaningful. Second, for the visual clarity, enforcing a unified, global y-axis would inevitably compress the modularity curves for larger models (e.g., $N_{hidden} = 64$) into a very narrow range (e.g., 0.1 – 0.2). This would render the fine-grained differences between task loads indistinguishable.

By allowing the axes to scale independently, we ensure that the relative hierarchy of modularity under different task loads, which is the primary focus of this analysis, remains clearly visible for every model size. This visualization strategy highlights the key finding: smaller models are highly sensitive to task count (large spread), while larger models remain robust (tight clustering), a trend that would be obscured by a unified scale.

Q: Line 69: When you mention training mode III, I think it would be helpful to the reader to point to the fact that you only use this training method for the section starting line 208, if I am correct. I read the following section wondering whether the multitask networks are training mode II or III until that section later suggested that only this late section uses training mode III.

A: We appreciate this excellent suggestion for improving readability. We agree that it is crucial to distinguish the training modes early to prevent confusion.

To address this without disrupting the flow of the Introduction, we have inserted an explicit clarification at the beginning of the Results section. This ensures the reader understands immediately that the initial multi-task analyses utilize Training Mode II, while Mode III is reserved for the specific section on connection-incremental learning. For the ease of review, we paste the content below:

The RNNs are then trained with noisy inputs (Fig. 1a) in three training modes (Fig. 1b) to evaluate all pertinent information. Unless specified as incremental, the default multi-task training refers to Training Mode-II in the following contexts.

Q: Line 219; “iterations after 10000”: Space missing before ‘10000’

A: We have corrected this typo. Thank you for your careful review!

Q: Line 251; “The ‘task’ models disregard the spatial constraints and define the feature kernel $K(u, v)$ using the functional connectivity matrices of the trained RNNs.”: Could you define of which trained RNNs? Is this from the multi-task RNNs? if yes, which training plan? From your mention of this result in line 284 I assume it is the multi task connectivity matrix.

A: Thank you for your question. You are correct that the feature kernel $K(u, v)$ is derived from the multi-task RNNs. As detailed in the Materials and Methods (subsection Functional Connectivity of Network Activity), these RNNs were trained using Training Mode II, identical to the experiments presented in Fig. 2b. However, distinct from the standard performance benchmarks, the RNNs used for this specific generative analysis were trained with an additional communication cost regularization term. This term simulates the metabolic cost of information flow between neurons spatially embedded in a 3D Euclidean space. The resulting functional connectivity matrix, which reflects both task demands and spatial embedding, was then used to define the kernel $K(u, v)$.

Q: I have had a look at the Github repository and at the current stage the README file does not have enough clear instructions to easily reproduce the analyses.

A: We have updated the repository code and provided a significantly more detailed README file. This revised documentation now includes clear, step-by-step instructions for: environment setup and installation, training the models, executing the analysis and plotting the results. These updates should now allow for straightforward installation and easy reproduction of the key findings presented in the manuscript.

<https://github.com/brain-intelligence-lab/modular-brain-rnn>

Response Letter

Reviewer 2:

We thank the reviewer for their constructive assessment of our manuscript. We particularly appreciate the reviewer identifying the need for rigorous statistical controls regarding model complexity, which we have now implemented.

In this revision, we have addressed the reviewer's concern that the "Spatial+Task" model's performance was merely a result of increased degrees of freedom. We have now calculated the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for our generative models. These metrics explicitly penalize the additional parameters. Our new analysis confirms that the "Spatial+Task" model achieves a statistically superior fit, even after this strict penalization, demonstrating that the improvement is topological and not due to overfitting.

Regarding the scope and contribution, our study aligns with a growing body of research in *Nature Machine Intelligence* that utilizes artificial neural networks to bridge the gap between computational function and biological organization. Recent work has approached this convergence from complementary angles:

1. Gornet & Thomson (2024) demonstrated that functional objectives like visual predictive coding are sufficient to drive the automated emergence of cognitive maps, linking sensory tasks to structured representations [1].
2. Concurrently, Liu et al. (2023) applied topological data analysis (TDA) to reveal that the predictive capability of deep networks is intrinsically tied to complex, quantifiable topological structures [2].
3. Most directly related to our physical modeling approach, Achterberg et al. (2023) established that spatially embedded RNNs, when constrained by metabolic costs, naturally converge on brain-like structural and functional motifs [3].

Our work synthesizes these perspectives: by integrating the functional task drivers highlighted by Gornet et al. with the physical spatial constraints utilized by Achterberg et al., and introducing a novel incremental developmental pressure, we provide a causal mechanism for the emergence of the modular topology that underpins these biological intelligences."

[1] Gornet, J. & Thomson, M. Automated construction of cognitive maps with visual predictive coding. *Nat. Mach. Intell.* **6**, 820–833 (2024).

[2] Liu, M., Dey, T. K. & Gleich, D. Topological structure of complex predictions. *Nat. Mach. Intell.* **5**, 1369–1381 (2023).

[3] Achterberg, J., et al. Spatially embedded recurrent neural networks reveal widespread links between structural and functional neuroscience findings. *Nat. Mach. Intell.* **5**, 1214–1226 (2023).

Furthermore, we have revised the manuscript to explicitly address the reviewer's theoretical points and respond to each of the reviewer's concerns one by one below.

Q: In the initial review, I highlighted that the "spatial + task" model inherently possesses more degrees of freedom than either the spatial-only or task-only counterparts. This expanded

flexibility enlarges its effective search space, meaning the model's superior fit is statistically expected rather than scientifically meaningful. In their response, the authors acknowledge this increased complexity but argue their goal was to illustrate the joint contribution of spatial and functional constraints, rather than to conduct model comparisons on an equal footing. However, the core issue lies not in the interpretive framing but in the absence of a principled comparison that accounts for differences in model complexity. The manuscript still fails to employ any standard tools (e.g., akaike information criterion, bayesian information criterion, or cross-validation approaches) for comparing models with varying numbers of parameters. Without such critical controls, the improved fit of the combined model cannot be disentangled from the advantage conferred by additional degrees of freedom. Consequently, the evidence provided does not convincingly support the claim that the brain's topology requires both spatial and task-based constraints; instead, the observed results could simply reflect overfitting by a more flexible generative framework. This fundamental gap leaves the manuscript's key conclusions without sufficient empirical grounding.

A: We thank the reviewer for this helpful suggestion. We agree that accounting for model complexity is essential to distinguish genuine topological necessity from overfitting. To address this point, we have performed a formal model comparison using both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) across all three generative frameworks (Spatial, Task, and Spatial+Task).

Methodological Approach: Standard edge-wise likelihoods (e.g., Bernoulli probability of specific edges) are often ill-suited for Generative Network Models (GNMs), as these models aim to reproduce global topological distributions (e.g., degree, clustering) rather than the precise location of every edge. Therefore, to rigorously calculate the likelihood $\hat{\mathcal{L}}$ while respecting the GNM framework, we adopted a Synthetic Likelihood approach with the following steps:

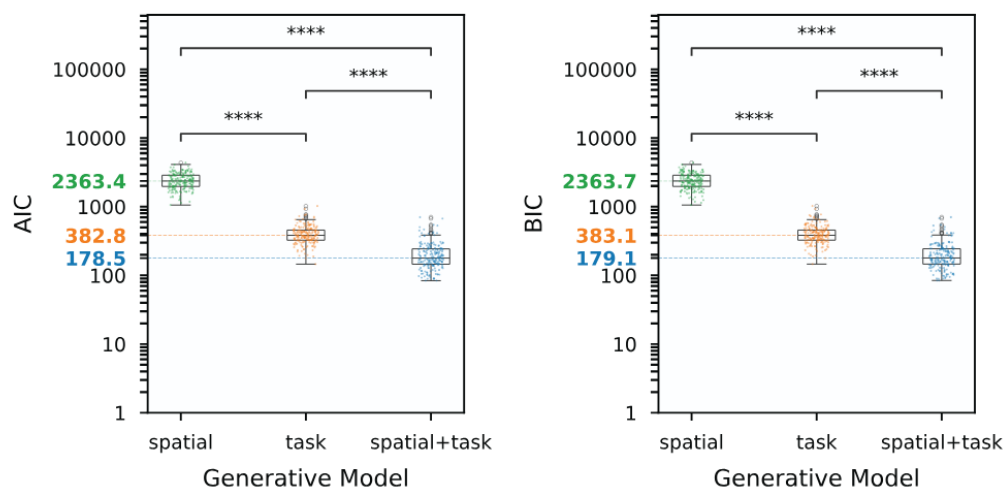
Step 1: We constructed a feature vector for each network by extracting summary statistics (mean, variance, and quantiles) from the distributions of degree, clustering coefficient, betweenness centrality, and edge length.

Step 2: To address collinearity, we applied Principal Component Analysis (PCA), retaining the top 10 components ($m = 10$) which explained over 98% of the variance.

Step 3: We then calculated the likelihood of the empirical data under the multivariate Gaussian distribution estimated from the model simulations (validated via the Henze-Zirkler normality test).

As shown in the new Supplementary Fig. 10, the "Spatial + Task" model achieves significantly lower AIC and BIC values compared to both spatial-only and task-only models. Even with the stricter penalty imposed by BIC for the additional parameter ($k = 2$ vs. $k = 1$), the combined model remains the strongest statistical fit. Specially, the AIC value for "Spatial + Task" is significantly lower than that for both "Spatial" ($t = 49.66, p < 10^{-5}$) and "Task" ($t = 17.61, p < 10^{-5}$). And the AIC value for "Spatial + Task" is significantly lower than that for both "Spatial" ($t = 49.65, p < 10^{-5}$) and "Task" ($t = 17.58, p < 10^{-5}$).

This confirms that the superior fit of the "Spatial + Task" model is not a trivial result of expanded degrees of freedom but reflects a necessary contribution of both constraints to explain the brain's complex topological organization. We have provided the full mathematical derivation and detailed results in Supplementary Note 6. For ease of review, we also provide the figure with caption below.



Supplementary Fig. 10. Model selection via Information Criteria. Comparison of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) across the three generative model families: Spatial-only (green), Task-only (orange), and Spatial+Task (blue). Lower values indicate a better trade-off between goodness-of-fit and model complexity. The combined "Spatial+Task" model consistently achieves the lowest scores, statistically outperforming single-constraint baselines even after penalizing for the additional parameter ($k = 2$ vs $k = 1$). Statistical significance was assessed via two-sample t-tests.

Q: While the introduction of the LME model attempts to statistically control for network size, it fundamentally fails to address the graph-theoretical core of my original concern. The issue is not merely that network size explains variance in modularity, but that the raw modularity score Q is mathematically biased by network density and sparsity. A sparse network will naturally exhibit higher raw modularity values purely as a mathematical artifact, independent of any biological or functional organization. The LME analysis attempts to regress out the effect of network size, but it does not normalize the metric against a topological null hypothesis. Consequently, the claim that smaller or incrementally trained networks are "more modular" remains theoretically unsubstantiated.

A: We appreciate the reviewer's theoretical concern that raw modularity scores Q are mathematically biased by network density and sparsity, and that comparing modularity across networks of different sizes must be done with care. As the reviewer noted, we previously utilized LME models to statistically account for these size variations. However, we acknowledge that regression controls may not fully neutralize the non-linear biases inherent in graph topology. For this very reason, we prioritized experimental designs that avoid cross-size comparisons entirely for our central claims.

Specifically, our primary finding that incremental learning drives superior modularity relies exclusively on the comparison shown in Figure 4b ("Incremental" vs. "All-at-once"). In this specific comparison, the networks are strictly density-matched: both groups are compared at the final iteration where they possess the identical number of nodes and edges (2,000 connections). Because the network size and density are identical between these two groups, the mathematical inflation of Q due to sparsity is constant for both. Therefore, the significantly higher modularity observed in the incremental group cannot be a mathematical artifact of density but must reflect a genuine topological difference driven by the learning curriculum. Moreover, for the results in Fig.2, what we demonstrated was the trend of modular change over training rather than claiming the relationship between network size and modularity. We have updated the text to explicitly state that this density-matched design serves as the primary control against the metric bias the reviewer correctly identified.

Q: In the initial review, I emphasized that correlations between modularity and performance during mid-training are inherently difficult to interpret, as they may stem from optimization dynamics rather than meaningful structural properties of the model. While the revised manuscript does demonstrate more stable correlations under controlled conditions, it still fails to provide a principled demonstration that these relationships are structural in nature rather than artifacts of learning dynamics. Furthermore, there remains a lack of a theoretical framework explaining why mid-training modularity should bear a mechanistic relationship to learning efficiency—whether in artificial neural networks or biological systems. Although the authors have incorporated stronger experimental controls (e.g., fixed data streams and uniform initializations) in this revision, the core concern regarding the interpretability of these mid-training correlations persists unresolved.

A: We appreciate the reviewer further clarifying their concerns regarding the nature of these correlations. To address the concern that these might be transient optimization artifacts, we implemented strict experimental controls in the revised manuscript.

Specifically, we standardized the initialization (neutralized weights) and fixed the random seeds across 256 independent model instances. As shown in Figure 3d, the strong positive correlation between modularity and performance ($p < 10^{-9}$) persists robustly under these controlled conditions. This observation demonstrates that the relationship is not a byproduct of random initialization variance or noisy gradients, but a consistent structural property: networks that successfully "find" modular configurations are consistently more efficient at multi-task learning.

To provide a theoretical grounding for these correlations, we may understand it from the perspective of structural gradient alignment. Specifically, we can posit the mid-stage modularity as a proxy for the degree of gradient subspace separation. In multi-task learning, the biggest issue is gradient interference (also known as "negative transfer"). This happens when the gradient for Task A updates a weight in a way that hurts Task B. Early in training, the 'Incremental' curriculum forces the network to allocate disjoint sets of weights to different tasks to avoid catastrophic interference (orthogonalizing the task gradients). High mid-stage modularity indicates that this subspace partitioning has been well achieved. Once established, this topology essentially diagonalizes the multi-task optimization problem, mathematically suggesting a smoother convergence trajectory towards a better final performance.

To support such theoretical view and disentangle structural advantages from learning dynamics, we point to our *Lottery Ticket* experiment (Figure 4e). Here, we extracted the emergent modular topology (mask) and re-initialized the network weights, thereby resetting all optimization momentum. The fact that this *posteriori modular* structure still outperformed random topologies, even after wiping the optimization history, demonstrates that the performance advantage is inherent to the topological structure itself, and not merely a transient artifact of the specific SGD trajectory that created it.

Q: The additional experiments across GRU, LSTM, CNN, and GNN architectures expand the empirical scope of the paper. However, the definition of modularity varies across architectures, and the manuscript does not provide a principled justification for treating these formulations as conceptually equivalent. In CNNs and GNNs especially, the computation of modularity via layer-derived feature matrices or bipartite graph constructions represents a fundamentally different notion of modular structure compared to modularity in recurrent weights. Without a unified theoretical framework explaining why these different quantities should be comparable or why multitask learning should induce modularity across such disparate architectures, the results appear more like a collection of case studies than evidence for a general principle.

A: We thank the reviewer for this helpful point. To provide some background motivation and context, we included the experiments on CNNs and GNNs to demonstrate the empirical robustness of the "task-drive $\rightarrow$ modularity" hypothesis beyond our primary model system. Our core scientific claim focuses on temporal/recurrent processing (RNNs) as a model for brain network topology. However, we believe it is scientifically valuable to report that similar modularization phenomena emerge in feedforward and graph-based architectures, even if the specific mathematical definition of "module" must be adapted to the architecture (e.g., feature grouping in CNNs vs. weight clustering in RNNs). We have revised the Discussion to clarify that these results serve as distinct, corroborating evidence for the universality of the principle, rather than implying that a single mathematical framework currently describes all deep learning topologies.

For the ease of review, we paste the content below as well.

It has been an emerging trend in both neuroscience and AI to model the brain's computational and cognitive function using recurrent neural networks (RNNs). The existing research approaches this bidirectionally. On one hand, RNNs are used either to replicate internal neural dynamics [27–29] and behavior [9, 30, 31], or as virtual platforms to assess whether certain brain mechanisms facilitate task learning [32, 33] or emerge from the physical constraints [6, 31]. Our work spans both perspectives, modeling internal dynamics while also demonstrating how adaptation to external task demands induces modular organization. While prior studies on image classification with network architecture search suggest that the feature representation alone may not give rise to the brain-like topology [34], our findings, alongside recent ones on familiarity detection [13], functional specialization [32], and spatial modularization [35], support a task-driven view of brain-like organization. Crucially, we observed that this structural emergence is not limited to recurrent systems but extends to feedforward (CNN) and graph-based (GNN) architectures. Although the mathematical definition of modularity differs across these

architectures, for example, as feature grouping in CNNs rather than the weight clustering in RNNs, the consistent emergence of modularization under multitasking pressure corroborates the universality of this principle beyond a single modeling framework. Our findings also align with recent perspectives in cognitive neuroscience that emphasize the role of task demands and generalization [30], including multitasking and continual learning [36].

Q: The contributions, though well-executed from an engineering standpoint, do not introduce fundamentally new algorithms, new theoretical constructs, or a novel modeling paradigm that would meaningfully shift current understanding of neural modularity or task-driven structural emergence. The revised manuscript demonstrates the enhancements in experimental rigor and scope. Nevertheless, the core methodological concern identified in the initial review remains fundamentally unaddressed. Given that this issue directly undermines the validity of the manuscript's central conclusions, the current version cannot be accepted for publication in the current journal.

A: We respectfully disagree with the assessment that this work represents a "mechanical combination" with "hard-coded" principles. We believe this view overlooks the core theoretical advance of the study: the identification of incremental development as the missing causal link between physical constraints and functional topology, offering a new mechanistic explanation for brain organization that neither pure spatial nor performance models can provide alone.

Why are our findings not “Hard-Coded” Principles?

The reviewer suggests that modularity is effectively hard-coded into our system. We define "hard-coded" as features explicitly enforced by the loss function or fixed architecture. We systematically ruled out this possibility through our experimental controls.

Alternative 1: Spatial Determinism. Our "All-at-once" control model (Fig. 4b) is subject to the exact same spatial embedding and metabolic costs as the main model. If space determined modularity, this control would be equally modular. It is not; it remains entangled and less modular.

Alternative 2: Functional Determinism. Non-modular networks can also solve the task (as seen in the "All-at-once" condition), albeit less efficiently. Thus, modularity is not a trivial requirement of the task itself. By eliminating these alternatives, we provide causal evidence that modularity is an emergent property of the interaction between task load and incremental growth. The system is not "told" to be modular; it "discovers" modularity as the energy-efficient solution to growing under pressure.

Alternative 3: Optimization Artifacts. We controlled for initialization (neutralized weights) and random seeds. Furthermore, our "Posteriori Modular" experiment (Fig 4e) shows that extracting the learned topology and re-initializing weights still confers a learning advantage. This isolates the topology itself as the causal factor for efficiency, independent of the optimization history.

How do our findings shift current understanding?

Briefly saying, our work meaningfully shifts the current understanding of neural modularity from "Static Optimization" to "Dynamic Emergence".

Current Paradigm: Existing theories generally attribute modularity either to static physical constraints (minimizing wiring cost, e.g., spatial generative models) or to functional necessities (minimizing interference, e.g., MoE architectures).

Our contribution: we demonstrate that neither is sufficient on its own. Physical constraints without development fail to capture functional long-range edges (as shown by our AIC analysis), and functional demands without development lead to entangled solutions. We introduce the theoretical construct of developmental energy minimization, shifting the paradigm to view brain topology not as a static endpoint, but as a trajectory defined by the process of learning.

Response Letter

Q: The current generative models suffer from a circular reasoning problem. Specifically, in constructing the 'task' and 'spatial+task' models, the authors define the feature kernel using functional connectivity derived from RNNs that have already been optimized for those tasks. While the study intends to demonstrate how task demands shape network topology, the generative model is essentially initialized with the functional correlations of a network that already possesses that exact topology. It is inherently expected that a generative model built using the "blueprint" of a task-optimized network will naturally output a task-optimized structure. Because the desired outcome is effectively baked into the starting conditions, this approach borders on circular reasoning. Consequently, the current empirical results do not fully support the assertive conclusions presented in the manuscript.

A: We respectfully but firmly disagree that the generative modeling analysis involves circular reasoning, and we wish to clarify the experimental logic in detail, as this point is critical.

The reviewer states that the generative model is "*initialized with the functional correlations of a network that already possesses that exact topology.*" This characterization conflates two entirely distinct networks. The generative model does not reconstruct the RNN. It reconstructs the empirical human structural connectome derived from the Human Connectome Project (Schaefer et al., 2018), a dataset that is never used during RNN training and shares no architectural correspondence with the RNN.

The experimental logic proceeds as follows:

- (1) An RNN is trained solely on cognitive tasks. Its training objective contains no information about the human connectome's topology. The functional connectivity that emerges in the RNN is purely a byproduct of task optimization.
- (2) This emergent FC is then used as a hypothesized wiring prior $K(u, v)$ in a generative network model (Eq. 4), which starts from an empty graph and stochastically grows edges.
- (3) The synthetic network produced by this generative model is evaluated against the independent empirical human connectome using Kolmogorov–Smirnov statistics on four topological properties (degree, clustering, betweenness, edge length).

Circularity would require that the outcome of step (3) is guaranteed by the inputs to step (1) — i.e., that using task-induced FC as a wiring prior necessarily reproduces the human connectome. This is not the case. The RNN's task-induced organization could easily have failed to explain the brain's structural topology. Indeed, this is precisely what happens for certain topological properties in the "task-only" model (e.g., node degree, Fig. 5c), demonstrating that the mapping is partial and non-trivial.

We note that this experimental logic is identical to the standard methodology established by Betzel et al. (2016) and adopted by Achterberg et al. (2023). In the "spatial" model, $E(u, v)$ is derived from the brain's own geometry — the Euclidean distances between cortical region centers. This spatial prior is also "derived from" the brain, yet it is not considered circular because it serves as a hypothesized constraint tested for its ability to explain the connectivity pattern. The task-derived functional connectivity plays the identical logical role: it is a candidate wiring rule whose explanatory power is empirically evaluated against an independent target.

To summarize: the RNN training uses no brain connectivity information; the generative model's target is the human connectome; and the RNN's FC is a mediating hypothesis, not the evaluation target. There is no logical circularity.

Q: While the addition of CNN and GNN experiments successfully expands the empirical scope, the mathematical definitions of modularity in these architectures (e.g., feature grouping in CNNs versus weight clustering in RNNs) appear to be distinct. The authors acknowledge in their rebuttal that a single mathematical framework for all these topologies is currently lacking. A unified theoretical framework is essential to support the broad conclusions of the paper; otherwise, the current findings would merely constitute a series of isolated case studies rather than a systematic set of conclusions.

A: We thank the reviewer for clarifying this concern. We agree that a fully unified mathematical definition of modularity across recurrent, convolutional, and graph-based architectures would be valuable. However, such a framework is beyond the intended scope of the present study. Our main claims are tested within the RNN setting using a consistent recurrent-weight modularity definition. The CNN and GNN experiments are supplementary robustness checks, not foundational claims. Our central contribution—and the focus of the manuscript's main text—is the emergence of modularity in recurrent neural networks as models of brain-like computation, evaluated under a single, well-established definition of modularity (Newman, 2008; Leicht & Newman, 2008). All five of our principal claims (multi-

task learning increases modularity, capacity constraints amplify the effect, modularity correlates with performance under capacity limits, incremental connectivity strengthens modularity, and task-induced networks better resemble the human connectome) are made and tested within this single, internally consistent RNN framework. To make this scope explicit, we have revised the Discussion to clarify that the CNN/GNN results are presented as corroborating evidence for the empirical generality of the task-pressure-to-modularity relationship, not as part of a unified theorem.

A unified mathematical framework across all deep learning architectures is a big research program rather than a revision item. The reviewer's request would require defining a single quantity that meaningfully captures "modularity" across recurrent, convolutional, and graph-based computations—a problem that, to our knowledge, has not been solved in the broader machine learning literature. We do not believe its absence invalidates the well-defined RNN findings, just as the absence of a unified definition of "complexity" across physical systems does not invalidate domain-specific analyses of complexity in fluids, solids, or biological networks. The same plurality of definitions is accepted within neuroscience itself. Modularity in the brain is operationalized differently across structural connectivity (white matter tracts), functional connectivity (BOLD correlations), effective connectivity (causal influences), and gene co-expression networks. These definitions are not mathematically equivalent, yet researchers consistently treat modularity as a unifying organizational principle across these levels (Sporns & Betzel, 2016; Bassett & Sporns, 2017). Our cross-architecture results follow this same accepted scientific practice: distinct mathematical formalizations of a shared organizational concept, providing convergent empirical support without requiring formal equivalence.

Regarding the theoretical foundation for the central RNN claims, to address the reviewer's broader concern about theoretical grounding (also raised in Major 3), we have now developed a formal mechanistic theory specifically for the RNN setting that constitutes our main contribution. This derivation, presented in detail in Supplementary Note 7 and our response to Major 3 below, provides a closed analytical chain from task statistics to recurrent weight modularity to learning efficiency. We believe this targeted theoretical foundation—rigorous within its scope—serves the paper better than a speculative cross-architecture framework would.

For ease of review, we have updated the Discussion as follows:

Crucially, we observed that this structural emergence is not limited to recurrent systems but extends to feedforward (CNN) and graph-based (GNN) architectures. We note that the mathematical definition of modularity is necessarily adapted to each architecture (e.g., feature grouping in CNNs vs. weight clustering in RNNs), reflecting the same plurality of operationalizations found across neuroscience itself, where structural, functional, and gene co-expression modularity coexist as complementary measures of a shared organizational principle. The consistent emergence of modularization under multitasking pressure corroborates the universality of this principle beyond a single modeling framework. Accordingly, the CNN and GNN results are presented as corroborating empirical evidence for the generality of the task-pressure-to-modularization relationship, while our formal theoretical analysis (Supplementary Note 7) is targeted at the RNN setting that constitutes our main contribution and most direct analog to brain network organization. Our findings also align with recent perspectives in cognitive neuroscience that emphasize the role of task demands and generalization [30], including multitasking and continual learning [36].

Q: I previously noted the lack of a theoretical framework explaining why mid-training modularity bears a mechanistic relationship to learning efficiency. In the response letter, the authors provide hypotheses involving "structural gradient alignment," "gradient subspace separation," and orthogonalizing task gradients. While these offer helpful intuitive explanations, they do not yet constitute mathematically rigorous proofs or strict mechanistic demonstrations. Without a formal mechanistic derivation, the underlying reasons for why this dynamic emergence occurs remain somewhat speculative. Consequently, a certain disconnect persists between the theoretical framework and the conclusions presented.

A: We thank the reviewer for emphasizing the need for a more explicit theoretical framework. We agree that the previous version relied too heavily on intuitive descriptions, such as gradient subspace separation, task-gradient alignment, and reduced cross-task interference, without providing a sufficiently formal account of how task-induced modularity relates mechanistically to learning efficiency. To address this concern, we have added a new Supplementary Note 7, *Mechanistic Account of Task-Driven Modularity and Learning Efficiency*.

The new note provides an idealized but explicit theoretical framework for the recurrent multitask setting studied in the manuscript. It formalizes the following empirical chain observed in our experiments:

task statistics $\rightarrow$ family-structured hidden representations $\rightarrow$ block-biased recurrent updates $\rightarrow$ modular recurrent weights $\rightarrow$ reduced destructive cross-task interference $\rightarrow$ improved learning efficiency.

Importantly, we still want to clarify that we do not present this framework as a universal theorem for all nonlinear RNNs or all neural architectures. Rather, it is a mechanistic account under clearly stated idealized assumptions, intended to clarify why modularity emerges under multitask pressure and why it can be predictive of learning efficiency, especially in capacity-limited regimes.

First, hidden-state second moments under multitask dynamics. We derive the decomposition of task-conditioned hidden-state second moments under a stable linearized multitask RNN approximation. This shows how task-family structure can be reflected in hidden-layer functional covariance. We also clarify that the raw one-hot rule inputs are not assumed to be block-structured; rather, the block structure arises from learned task-conditioned hidden statistics.

Second, capacity-induced task overlap. We use a frame-potential bound to show that, when the number of tasks exceeds the latent representational dimension, task representations cannot remain mutually orthogonal. This provides a formal basis for why increasing task load relative to representational capacity favors structured reuse among compatible tasks. We now describe the empirical task-to-dimensionality ratio as a practical proxy for this latent capacity ratio, rather than as the exact theoretical variable.

Third, a bridge from representational modularity to recurrent weight modularity. We explicitly derive how family-structured hidden activity can induce block-biased recurrent weight updates. For a task belonging to family g , the hidden activity and backpropagated error are decomposed as

$$h_t^{(m)} = P_g z_t^{(m)} + r_t^{(m)}, \delta_{(t+1)}^{(m)} = Q_g e_t^{(m)} + s_t^{(m)}.$$

Since the recurrent gradient has the outer-product form

$$\nabla_{W_{rec}} \mathcal{L}_m = \Sigma_t \delta_{(t+1)}^{(m)} h_t^{(m)\top},$$

the dominant component of the expected update is block-supported:

$$Q_g \mathbb{E} \left[e_t^{(m)} z_t^{(m)\top} \right] P_g,$$

with cross-family contributions controlled by the residual terms. This fills the previous logical gap between representational modularity and structural modularity in the recurrent weight matrix.

Fourth, modular recurrent structure suppresses destructive multitask interference. In a steady-state linearized RNN, the gradient inner product between two tasks factorizes as

$$G_{ij} = \beta^2 (\delta_i^\top \delta_j) (r_i^\top r_j).$$

Thus, cross-task interference becomes small when either forward activity overlap or backward error overlap is small. For approximately modular recurrent weights,

$$W_{\text{rec}} = W_D + \gamma \widehat{W}_O,$$

a perturbation analysis gives

$$|G_{ij}| \leq C\gamma^2 + O(\gamma^3),$$

where γ measures the strength of inter-family recurrent coupling. This provides a quantitative mechanism linking modular recurrent structure to reduced destructive cross-task interference.

Fifth, reduced destructive interference improves learning efficiency. We connect the above interference reduction to both local descent and stochastic task-sampling variance. Reduced destructive cross-family interference decreases gradient cancellation and stabilizes multitask optimization. This provides a mechanistic explanation for why mid-training modularity can predict learning speed and final performance, especially in capacity-limited networks.

We also revised the interpretation of the theoretical framework to avoid overstating its scope. The new analysis is an idealized mechanistic account, not a complete proof for arbitrary nonlinear recurrent networks or a universal theorem across architectures. For CNN and GNN experiments, we now position them as empirical robustness analyses under architecture-specific interaction graphs, rather than as evidence for a fully unified architecture-agnostic mathematical theory.

Finally, we also added a new paragraph on this issue in the discussion as:

To clarify the mechanism linking modularity and learning efficiency, we provide an idealized theoretical account in Supplementary Note 7. Under a stable linearized multitask RNN approximation, task-conditioned hidden-state statistics can acquire family-wise block structure. When task number exceeds latent representational capacity, full task separation becomes impossible, favoring clustered reuse among compatible tasks. We further show that family-structured hidden activity and error propagation induce block-biased recurrent updates, and that approximately modular recurrent weights suppress destructive cross-family gradient interference. This provides a mechanistic explanation for why modularity is most predictive of learning efficiency under capacity constraints, while remaining an idealized account rather than a universal theorem for arbitrary nonlinear networks.

Q: Furthermore, the authors repeatedly claim to provide a "causal explanation" or "causal mechanism" for how functional demands shape biological brain network

topology. I would suggest softening these claims. Artificial RNNs optimized via global backpropagation serve as a valuable functional analogy, but they do not fully replicate biological causal mechanisms. Because the results remain primarily algorithmic case studies and artificial abstractions without a unified theoretical framework, the broad conclusions claimed by the authors are not sufficiently supported. Therefore, I stand by my assessment that this work is better suited for a specialized technical or engineering-focused journal, rather than providing the broad conceptual innovations expected for Nature Machine Intelligence.

A: We thank the reviewer for this suggestion and agree that the term "causal" warrants more precise usage. In the revised manuscript, we have carefully distinguished two distinct levels of claim and adjusted our language accordingly. We respectfully disagree, however, with the broader assessment that the work is unsuitable for Nature Machine Intelligence, and we wish to explain our reasoning.

We agree that our previous use of “causal explanation” and “causal mechanism” was too broad when referring to biological brain network topology. In the revised manuscript, we distinguish between causal evidence within the controlled computational model and biological causal claims. Within the artificial RNN model, we manipulate task regime, task load, capacity, and connection-growth schedule and observe systematic changes in topology. This supports a controlled computational claim: task demands and capacity constraints are sufficient to induce modular organization in artificial recurrent networks. Respectively, we have revised the manuscript at several locations to clarify the scope of our claims. Specifically, we softened the Abstract and Introduction to avoid claiming a biological causal mechanism; renamed the Results subsection previously titled “A Causal Explanation for the Emergence of Modularization”; added a pointer to Supplementary Note 7 in the Results; inserted an idealized mechanistic-account paragraph early in the Discussion; softened the interpretation of CNN/GNN experiments as empirical robustness checks rather than evidence for a unified architecture-agnostic theory; and revised the limitation paragraph to explicitly state that backpropagation-trained RNNs do not reproduce the biological developmental, evolutionary, or biophysical mechanisms of brain network formation. We also integrated the AIC/BIC model-complexity analysis into the Fig. 5 Results section and added a Methods pointer to Supplementary Note 6.

For the ease of review, we paste the related modified paragraphs below and the full change is marked in blue in the revised manuscript.

Old-1: *Thus, the divergence of these two optimization goals offers a mechanistic explanation for the brain's modularization, distinct from models that consider only spatial constraints or representational demands in isolation.*

New-1: *Thus, the divergence between metabolic and representational objectives offers a computational interpretation of why modular organization may arise when spatial constraints and task-driven representational demands act together, distinct from models that consider only spatial constraints or representational demands in isolation.*

Old-2: *Together, our results provide a causal structural explanation for the emergence and function of the brain's modular topology.*

New-2: *Together, our results provide a function-to-structure account in artificial networks, showing how task demands, representational capacity, and connection-growth constraints can induce modular organization and generate topological features that resemble aspects of brain network architecture.*

Old-3: *Despite providing a causal explanation for task-induced modularization, this study has limitations. The proposed protocol, while compared against generative models[7, 56], remains a simplified, non-fully-validated biological abstraction. Furthermore, while RNNs allow for tractable analysis, the complexity of both the human brain and modern foundation models suggests no single principle may explain intelligence. Future work should explore extending this investigation to explain other characteristics of brain function and yield generally efficient machine learning algorithms.*

New-3: *While our results provide a controlled computational account of task-induced modularization in artificial recurrent networks, the proposed protocol remains a simplified biological abstraction. RNNs optimized by backpropagation do not reproduce the full developmental, evolutionary, plasticity-based, or biophysical mechanisms that shape biological brain networks. Furthermore, while RNNs allow for tractable analysis, the complexity of both the human brain and modern foundation models suggests that no single principle is likely to explain intelligence. Future work should examine whether related function-to-structure principles hold in more biologically grounded learning systems, empirical longitudinal data, and broader classes of computational models.*

Regarding the fitness to Nature Machine Intelligence, we respectfully disagree that the revised work is merely an algorithmic case study. Nature Machine Intelligence

has published a substantial body of work using artificial neural networks to derive insights about biological neural organization, including the spatially embedded RNN framework of Achterberg et al. (2023, NMI), the cognitive map work of Gornet & Thomson (2024, NMI), and the topological analysis of Liu et al. (2023, NMI) cited in our last response. Our work directly extends this established line of research by introducing a third organizational principle (incremental developmental pressure) that complements the spatial and functional principles previously established in NMI. We respectfully suggest that a contribution which builds on, integrates, and extends three prior NMI papers in the same conceptual space is well-aligned with the journal's scope.

Furthermore, we also note that several findings in our manuscript are of direct relevance to broader questions in the NeuroAI community: the demonstration that task-induced wiring captures features of brain connectivity beyond spatial geometry, the lottery-ticket-style result showing that modular topology itself confers learning advantages independent of weight values, and the formal connection between modularity and gradient interference reduction. These contributions speak to readers across machine learning, computational neuroscience, and cognitive science, the cross-disciplinary audience that NMI is designed to serve.

We hope these revisions and clarifications address the reviewer's remaining concerns. We have carefully considered each point raised across both rounds of review, implemented concrete methodological improvements (AIC/BIC analysis, density-matched comparisons, formal theoretical derivation, softened causal language), and we believe the resulting manuscript is substantially strengthened.

Q: The authors state in their rebuttal that they have calculated AIC and BIC to account for model complexity. While I acknowledge that these scores were calculated using a 'Synthetic Likelihood' approach in the supplementary materials, this critical discussion is currently absent from the main text and must be integrated to support the study's conclusions.

A: We thank the reviewer for pointing this out. We agree that the information-criterion analysis should be discussed in the main text rather than only in the Supplementary Information. We have now added a paragraph to the Fig. 5 Results section explaining that we used a synthetic-likelihood-based AIC/BIC analysis to account for model complexity. This analysis shows that the Spatial+Task model retains the lowest AIC and BIC scores and significantly outperforms both the Spatial-only and Task-only models even after penalizing the additional model

parameter. We also added a pointer in the Methods section to Supplementary Note 6, where the full synthetic-likelihood procedure, feature construction, and statistical results are provided. For the ease of review, we copy the added context in the Results section as below:

To ensure that this improvement was not merely driven by the additional degree of freedom in the combined model, we further evaluated model complexity using a synthetic-likelihood-based information criterion analysis. Specifically, we computed AIC and BIC from the likelihood of empirical topological summary statistics under the simulated network distribution generated by each model family. The Spatial+Task model retained the lowest AIC and BIC scores, significantly outperforming both the Spatial-only and Task-only models even after penalizing the additional parameter (Supplementary Note 6; Supplementary Fig. 10; Supplementary Table 5). These results indicate that the improved fit of the Spatial+Task model reflects genuine explanatory value rather than overfitting due to model complexity.

Response Letter

Reviewer 2:

We thank Reviewer #2 for the careful re-evaluation of our manuscript and for acknowledging the substantial revisions. We recognize that the reviewer remains unconvinced, but respectfully disagree that the remaining concerns reveal foundational flaws in the study. In our view, these concerns assess individual components as though each were intended to support a substantially broader claim than the manuscript assigns to it. A more comprehensive nonlinear theory and a causal developmental account would be valuable extensions, but neither is necessary to establish the conclusions advanced here. The current conclusions are supported primarily by controlled empirical results, while the theoretical analysis provides a tractable mechanistic account and the generative analysis provides complementary comparative evidence. We therefore maintain that the explicit scope of these complementary analyses does not diminish either the validity or the conceptual significance of the central findings.

1. Regarding the theoretical analysis

We respectfully disagree that the use of a stable linearized approximation renders the theoretical analysis invalid. Linearization around a trajectory or operating regime is a well-established general strategy for analyzing nonlinear dynamical systems and trained RNNs. It does not replace the nonlinear model but provides a tractable means of isolating the mechanisms governing local activity and error propagation. A complete exact theory of learning and state evolution in noisy, input-driven nonlinear RNNs is not available in a generally applicable form. Indeed, requiring such a theory would constitute a substantially broader theoretical program rather than a necessary condition for the present mechanistic analysis. Importantly, our central findings are demonstrated directly in the full nonlinear RNNs. Supplementary Note 7 complements these experiments by explaining, under explicit assumptions, how family-structured representations induce block-biased updates and how modular recurrent organization reduces cross-task interference under capacity constraints. The approximation therefore limits the domain of the formal result, not its validity or relevance. We consider it a principled and appropriate theoretical treatment of the specific mechanism investigated here.

2. Regarding generalization across architectures

We also respectfully disagree that the absence of a single theorem spanning RNNs, CNNs, and GNNs reduces the results to unrelated case studies. The formal analysis is deliberately targeted to RNNs, which constitute the main model class and the most direct setting for studying recurrent network topology as in a bunch of recent work in top journals. The cross-architecture claim is empirical and more specific: within each architecture's native interaction graph, increased multitasking pressure is associated with increased modular organization. Generalizing the direction of an intervention–outcome relationship does not require the modularity measures to be mathematically identical across fundamentally different architectures. We therefore regard the CNN and GNN results as meaningful corroborating evidence for the robustness of the task-

pressure-to-modularization relationship, rather than as a claim of an architecture-invariant mathematical theorem.

3. Regarding the generative modeling analysis

The circularity argument conflates the use of a hypothesized generative rule with encoding the target topology into that rule. The feature kernel $K(u, v)$ is constructed from task-evoked activity correlations in trained RNNs; it is not derived from the empirical human structural adjacency matrix or from the graph statistics used to evaluate the generated networks. Any generative model is necessarily biased by its proposed wiring rule. This does not make its evaluation tautological. A task-derived functional kernel is not guaranteed to reproduce the human connectome's degree, clustering, betweenness, and edge-length distributions, nor is it guaranteed that the combined "*Spatial+Task*" model will outperform both component models after accounting for model complexity. The analysis therefore provides a nontrivial test of whether task-derived functional organization adds explanatory value beyond spatial geometry. It is not intended as, and is not required to be, a causal reconstruction of biological development.