nature portfolio

Peer Review File

Biomimetic model of corticostriatal micro-assemblies discovers a neural code

Corresponding Author: Dr Richard Granger

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)
The authors have dealt with one of the prominent problems in neuroscience, namely, hierarchical modeling of cellular-to-cognitive functioning of the nervous system. They have attempted to reach this goal using a detailed cellular and connectivity model of the cortico-striato-thalamic system, for which there is a veritable mountain of neurobiological evidence. They have verified this model to account for cellular and behavioral evidence in the nonhuman primate. Most important for the present submission, the authors claim to have utilized the model to uncover a novel “neural code”. The highlighted features of the authors’ model is that it can reveal model outcomes that are, to use the authors’ terminology, “zero-trained” or “out-of-distribution inferences”. These are certainly marked strengths of the current paper, and in this sense, are certainly deserving of publication in Nature.
Despite strong support for the current submission, this reviewer finds several weaknesses that require revision. First is the authors’ justification of “lateral inhibition assemblies” as evidence for a “computational primitive” of the cortical-striatal circuit. Although this claim is certainly reasonable, lateral inhibition by itself is not terribly convincing as an emerging computational property. In addition, “winner-take-all” functions are hardly biologically reasonable. Aren’t there more complex/surprising and thus more interesting spatio-temporal patterns of activity emerging from the interconnected circuitry of neurons having the biophysical properties known for the modeled neurons? Given the known richness of voltage-dependent channels having differential dendritic and cellular distributions in cortical, striatal, nigral, and thalamic populations, is the proposed model really limited to age-old lateral inhibition? Although this question is open-ended and thus somewhat unfair to the authors, this reviewer finds that other spatio-temporal patterns must be evident and would represent a far stronger argument for the model.
The fact that the higher-level output of the model represents category learning is exciting, but it should be noted that the number of categories is small indeed, i.e., two. This is consistent with the emphasis on lateral inhibition: on or off. Perhaps this reviewer has missed something important about the category learning accomplished by the model, in which case whatever that is should be re-emphasized.
A second concern which is related to the first is the small number of neurons included in the model. The proposed model is composed of only hundreds of neurons (without compartments? see below). In these days of widespread computational power, an increased number of neurons per region in the model would seem more than reasonable. In addition, it would provide the basis for detection of more complex spatio-temporal patterns involved in coding schemes.
Third, an important issue is that the authors point out that both the Striatal and TAN subcircuits are not represented by sets of compartmental neurons, but instead as input/output elements. There is a rich literature documenting the strength of input/output models and their comparison to compartmental models, yet this literature is not cited or used to justify the authors’ approach.

This reviewer has several minor comments as follows: (1) is “lateral inhabitation” a useful, widely used term? vs lateral inhibition? (2) was the pars reticulata purposely left out of the model? (3) presynaptic inhibition also has been found among striatal neurons; was this form of inhibition also deliberately omitted? The authors cannot be expected to include every known function or feature in their model; simple justifications to these questions should suffice.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
In this paper, the authors constructed a biologically detailed model of the cortico-basal ganglia-thalamic networks of spiking neurons. Through numerical simulations of the model, the authors validated the model's neural activities using the data recorded from macaque monkeys. The authors showed that neuronal functional subtypes emerge to predict a failure behavioral choice (incongruent neurons) and claimed that the model is "zero trained". The two results sound interesting and constitute the major finding of this study. However, the present manuscript is hard to follow: it is written in small fonts and densely packed with many lines. It was also a labor to find the information necessary to clarify the details of the study. These style issues have nothing to do with the quality of the science reported, but they disturb the readers' understanding of the results. Partly due to these difficulties, I could not be fully convinced of the validity of the major claims. I feel that these claims are somewhat overstated.

Major comments:
1. The categorization task used should be explained more clearly with illustration (fig. s6). In the current manuscript, fig. s6 is referenced much later than where the task is first explained in ll. 53-55 at the beginning of the Result section. The reader can easily miss the figure and just find a short description of the task in the Methods section (without a reference to Fig. s6).

2. In ll. 34-36, the authors explained the "zero trained" condition as: " A far more rigorous standard of generalizability is to ask whether a model built from one kind of data can predict phenomena of a wholly different kind. This explanation confused me because the model does not seem to undergo any training on a different type of task before it is trained on the categorization task. The authors might have meant to say that they built the model solely based on anatomy and physiology data and tested its performance on learning behavioral data from the categorization task. If this is the case, however, I do not find a big surprise in that the model could learn the categorization task using its entire dataset. The authors should explain why this is non-trivial before they claim that the "zero trained" feature is an unexpected outcome of the model. The training conditions should also be explained more clearly.

3. Potentially related to the above comment, I am not convinced of the importance of the "primitive", which is a modular block of small numbers of excitatory and inhibitory neurons (fig. 1b). The computational benefit of this modular structure was not explored sufficiently in detail. The authors hypothesized that these primitives exist in the cortical layer 2/3 and they perform WTA competition via lateral inhibition. Are these hypotheses supported by experimental evidence or just theoretical assumptions?

4. In ll. 277-281, the authors claimed "Finally, we emphasize the broader point that it was the zero-trained nature of simulations based on biomimetic computation that enabled the emergence of these broad and unexpected findings. Whereas much computational work focuses on the statistical learning of specific data, the present work instead attempts to uncover how low-level physiological spiking, embedded in specific delimited sets of anatomical wiring organizations, can itself give rise to the forms of neural and behavioral results that brains produce." However, I doubt whether this finding is as surprising as the authors claim. As mentioned in my comment 3, the authors hypothesized WTA competition for the primitives, a basic computation essential for categorization tasks. Their large-scale model, therefore, involves a minimal circuit motif necessary to solve the task under consideration. When a large system contains a small system that can solve a task, I find no surprise that the large system can also solve the same task.

5. The A-correct, A-incorrect, B-correct, and B-incorrect neuronal responses should be defined explicitly. What exactly can A-incorrect and B-incorrect neurons tell about the categorization of input? Is it the probability that an input is unlikely to belong to A or B (i.e., the volume ratio between the pink shaded area vs its outside in Fig. S6)? Logically, X = not-A does not mean X=B, and similarly, X=not-B does not mean X=A. However, in the alternative choice task (e.g., succade to A or B), not-A may behaviorally imply B as the animal can choose one of the two options. How are the activities of the four functional subtypes combined to generate the model's decision? The computational implications and importance of incongruent activities depend on how the task was designed and how the neural responses were used to solve the task. I might find these details somewhere in the manuscript, but they should be explained more carefully.

6. What conditions ensure the emergence of incongruent responses in the model? Do incongruent neurons play any role in switching between exploitation and exploration? How crucial is having such a functional subtype for task performance?

7. I wonder whether references are cited correctly (for example, refs 16-18). Please check the references.

Minor comments:
8. How is the speed of learning compared with the experimental observation?

9. Are the responses of TANs crucial for switching exploitation and exploration?

10. In Fig. 2b, what <W> means? Does it refer to averaging over neurons? I guess some connections are strengthened, as shown in the figure, but the other connections are weakened even if their overall average increases during learning.

11. The limitations of the model should also be discussed briefly.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
This study proposes a computational model that can simulates multiple brain regions (cortex, striatum and thalamus) associated with perceptual decision-making and replicates both spikes and LFPs, which have been experimentally observed. As most brain areas do not work alone, it is imperative that we simulate interareal interactions. In this study, the authors propose to combine population models and spiking neural networks to reduce the complexity of multi-areal models, which is an interesting approach. As they constrain the model's structure using experimental data, their study may provide a good reference model for those who develop multi-areal models.

However, I would like to point out several concerns.

This model aims to replicate the recorded neural responses, while NHPs are performing a simple visual categorization task. Since the task is quite simple, no significant learning is necessary, but the authors implied that their model can provide insights into the learning process in the brain. In my opinion, the value of this model is that we could probe functional roles of brain areas on visual categorization using their model. Thus, I would like to ask the authors if they tried to quantitatively evaluate the influence of individual brain areas. If one of the areas is disabled, neural and behavioral responses may fail or differ from the observed responses. It may be possible to gain new insights into perceptual decision-making from these failure modes.

The authors claimed that their model can predict the existence of incongruent neurons, which are experimentally confirmed. However, I am concerned that this incongruent neuron could be just a byproduct of winner-take-all mechanisms. Let's consider three sets of neurons. Stimulus A, Stimulus B and none-of-the-above (NOA). Due to lateral inhibition, we expect only one of the three sets to be active at a time. In addition, we can further assume that Stimulus A or B is trained to become quiescent when an incongruent stimulus is presented. That is, when stimulus A is presented, Stimulus B neurons become silent, and vice versa. Then, if Stimulus A fails to activate Stimulus A neurons, NOA neurons will be active due to the lack of lateral inhibition, which will be consistent with incongruent neuron behaviors. Based on this line of reasoning, I am not sure what we can learn from their model's predictions.

In the Results section, the main cortical area of interest is "anterior cortex", but Methods section describes "prefrontal cortex", not anterior cortex. As the cortex model relies on a generic structure, such mix-up may not matter too much, but to maintain clarity of this study and not to confuse the readers, this mix-up should be fixed properly.

In lines 18-25, the author stated: "However, most of these models have been optimized for either mechanistic accuracy or to elicit emergent cognition, but not both. On the one hand, there are powerful models with biologically-based systems that reproduce extensive physiologies, from single-cell spiking through generation of oscillations and other critical emergent dynamics. Yet, while these can accurately simulate brain-wide phenomena (7) such as epileptic seizure genesis and propagation (8, 9), they are not designed to generate cognitive processes, such as learning. On the other hand, some of the most remarkable cognitive models are those grounded in artificial neural networks (10–15). Yet, while this is a class of models that effectively "learn," they do so without reliance on biologically realistic (biomimetic) mechanisms at the cellular scale."

I believe these statements are not accurate. A line of study uses biologically-realistic circuits to probe neural correlates of high-level cognitive functions such as decision-making, working memory and integration of sensory evidence. Also, all major simulation platforms, such as NEST and NEURON, support long-term synaptic plasticity and can be used to study 'learning' in the brain. The general learning algorithm for spiking neural networks, which corresponds to 'backpropagation' in deep learning, remains missing, but the lack of learning algorithm must not be confused with the lack of study of biological learning.

Selected References
Goldman, M. S., Compte, A., & Wang, X. J. (2009). Neural Integrator Models. In Encyclopedia of Neuroscience (pp. 165-178). Elsevier Ltd. https://doi.org/10.1016/B978-008045046-9.01434-0
Wang XJ. Theory of the Multiregional Neocortex: Large-Scale Neural Dynamics and Distributed Cognition. Annu Rev Neurosci. 2022 Jul 8;45:533-560. doi: 10.1146/annurev-neuro-110920-035434. PMID: 35803587.

Wang XJ. Decision making in recurrent neuronal circuits. Neuron. 2008 Oct 23;60(2):215-34. doi: 10.1016/j.neuron.2008.09.034. PMID: 18957215; PMCID: PMC2710297.

Hill S, Tononi G. Modeling sleep and wakefulness in the thalamocortical system. J Neurophysiol. 2005 Mar;93(3):1671-98. doi: 10.1152/jn.00915.2004. Epub 2004 Nov 10. PMID: 15537811.

Wagatsuma N, Potjans TC, Diesmann M, Fukai T. Layer-Dependent Attentional Processing by Top-down Signals in a Visual Cortical Microcircuit Model. Front Comput Neurosci. 2011 Jul 8;5:31. doi: 10.3389/fncom.2011.00031. PMID: 21779240; PMCID: PMC3134838.

In lines 31-39, the authors stated: "Of course, in silico "discovery" of new mechanisms and treatments is possible only if models can apply their learned knowledge to novel or unexpected inputs. A computational model's ability to make predictions or draw conclusions about data that fall outside the distribution of its training set is referred to as out-of-

distribution inference. A standard training approach is to train a model from half the data of a particular kind and then test its ability to predict the other half. A far more rigorous standard of generalizability is to ask whether a model built from one kind of data can predict phenomena of a wholly different kind. These latter types of predictions are said to be zero-trained.

Here, we address these challenges by building a zero-trained computational model of the corticostriatal circuit designed to be mechanistically accurate across multiple scales, from spiking neurons to local field potentials."

I strongly recommend the authors to remove the statements regarding 'zero-trained' and clarify what their proposed model actually achieves. The authors are implying that their proposed model can perform zero-shot learning (zero-trained) or make predictions on novel inputs (out-of-distribution inference). While zero-shot learning and out-of-distribution are important topics in the field of deep learning, the authors do not provide any results relevant to zero-shot learning and out-of-distribution inference in their study.

Also, the authors need to clarify what 'multi-scale' means in their manuscript. Based on the phrase, "from spiking neurons to local field potentials", they appear to be referring to both structure (e.g., neurons or brain areas) and neural responses (e.g., spikes and LFPs), but to avoid confusion, they should clarify what they mean.

In line 106, the authors stated: "The arrangement used here is very distinct from ANN approaches"

ANN has not been defined. If ANN means artificial neural networks, the authors should define what kind of ANN approaches they are referring to.

In lines 126-128, the authors stated: "The interconnections between these subcircuits support the process of learning. As the model experiences many trials, there are changes to synaptic strength for cortico-cortical, cortico-striatal, and thalamo-cortical contacts that cause changes in behavior"

I think the authors should show how these interareal connections evolve over time. As the model depends on various cell assemblies, they should show the evolution of weights between cell assemblies.

In line 156, the authors stated: "For the simulation, we developed a similar metric called Local Spike Summation (LSS), a simple measure of simulated neuronal activity defined in Method"

Earlier works simulated LFPs in multiple ways, but LFPs are generally thought to reflect synaptic inputs, whereas LSS reflect outputs of neurons. The authors should, at least, justify why they used LSS to approximate LFPs. As their proposed model aims to produce multi-scale responses, it is important to simulate LFPs as accurately as possible.

In line 158, the authors stated: "one noticable difference is that NHP brains do not show as marked a decrease in spiking after stimulus offset as the model (Figure 2C)"

Can the authors provide any insights into this discrepancy? Alternatively, can they explain why it is not important?

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
This is an exciting paper, mainly because of the discovery in the model of a type of neural coding that was unknown before and is found in their data after its discovery in the model. This population of neurons codes for when the choice of response is going to be incorrect.

Since I took the theory track at Cornell, I never got diffEQ, so I did not check the equations carefully, but I will note that they switch in the article between W^{ij} meaning the weight from unit j to unit i (e.g., line 319) to vice versa (e.g., line 520) and back (line 99, suppl. mat.). Be consistent.

My main problem with the paper is that this type of model has been done before, perhaps not with such exacting precision with respect to the neurophysiological data, but the spirit is the same, by Randy O'Reilly and Yuko Munakata (2000 and subsequent versions), and there is no reference to this work. In that book, they simulate learning a go-no go task with a model basal ganglia using the Emergent simulator. Similar to the current model, the system is wired up in the same manner as the basal ganglia, and it is trained by the reinforcement learning system intrinsic to the model. Similarly, there's no reference to Chris Eliasmith's work with SPAUN.

In addition, you say "those models were designed to elicit a circumscribed family of behaviors centered on reinforcement learning and were not constructed in a manner that allows them to be easily extended beyond their original scope." This is certainly not true of Eliasmith's model, which shows the ability to learn seven different tasks. And while I haven't read all of them, O'Reilly, especially with Michael Frank, has published numerous papers on his model's explanation of working memory, etc.

I'm sure there are many important differences (for example, your model includes oscillatory behavior, crucial to some aspects of the model, which differs from O'Reilly's model if I recall correctly). However, it is still necessary to distinguish this model from those earlier works.

So, I believe that this work is important because of the discovery of a novel neural code that emerged from the simulations. However, I would like to see how it relates to earlier models in the same vein. If that issue were addressed, I think this rates as publishable in Nature Communications.

What would really nail the data would be if you can use your idea of promoting learning by stopping trials that the incongruent neurons predict will be an error in an actual monkey. I don't expect that, given the difficulty, expense, and time necessary, but wow, if you did that, it would be a paper for the ages.

More specific comments/wording suggestions/typos/notes:

Introduction: there is no reference to any of Randy O'Reilly's work??

"These latter types of predictions are said to be zero-trained."

I've never heard the concept of "zero-trained" - I've heard of zero-shot. If you are going to introduce new terminology, say instead "we call these models "zero-trained"".

However, that seems like an anomaly, since you do train it - by reinforcement learning, which is built into the system. I think a better description would be to call it "intrinsically-trained". But I will leave that up to the authors.

line 91: lateral inhibitation -> lateral inhibition

Fig1 caption: Detailed discussion in provided in the Methods
-> Detailed discussion is provided in the Methods

line 155: no need for scare quotes around local field potentials
line 160: no need for scare quotes around phase locking value.
There may be other examples of this.

line 161: reference 20: EG Antzoulatos, EK Miller (2014). Glancing at this paper, I take it that the strength of connections between the two areas (striatum and PFC) increase overall - i.e., both ways. In Figure 3b, you also show this connectivity increasing, but it isn't clear whether this is from stratum to your simulated frontal lobes or vice-versa. Can you be more specific than "portico-striatal"? Does this graph work both ways?

Line 186: Commprising -> Comprising

Figure 3A caption: highlighted in gray, not yellow. The colors of the lines are incorrectly specified (they are blue and green, not red and light red).

Figure 3B Figure: Correct me if I'm wrong, but you seem to have incorrect and correct flipped here.

Last line of Figure 4 caption: The sentence says "would improve" - so this is a prediction. If you were able to do this experiment, it would be especially confirmatory, although I know experiments with NHPs are expensive and time consuming.

Lines 251-252 - please remove the scare quotes.

line 269: NHPs are not "constructed"! Try "evolved".

Line 285 and others: You describe this as "working memory"? Can you elaborate on why you call it that?

In equations 30 and 31, I believe you have switched from W^{ij} meaning the weight from j to i (as in equation 3) to the reverse. Hopefully this is correct in the code!

It is unclear what is going on in equation 29, as there is only one index on the weight, that is, "i", which suggests this is a solely pre-synaptic learning rule, unlike the rule in equation 28. Explain this, or fix it.

Equation 33 reverts to the equation 3 definition of which way the weight W^{ij} goes.

Line 501: you need a reference for the BCM rule. Bienenstock, Cooper, and Munro 1982.

Line 502: typo: "formas" either this is supposed to be "form as" or "formulas"

In the Computational Methods section (starting line 547), you don't mention Chris Eliasmith's SPAUM model using the Semantic Pointer Architecture. How does NeuroBlox differ from his simulator, which seems to be addressing very similar concerns and uses actual spiking neurons. Or alternatively, how does it differ from Emergent (O'Reilly)?

Supplementary Material

Line 8: (NG-NMM) model, -> mass model (NG-NMM),

Line 24 Supplimental -> Supplemental

Figure S5 caption: "Striatal spiking shows a post-stimulus-onset silence that was subsequently found in empirical data;" This seems like a second discovery. Why don't you also emphasize this in the main text? Here, in terms of "silence", are you referring to the pause in spiking around 40-60ms? This appears to be borne out in the local spike summation, although the timing is not quite matched by the empirical data. The match between D and E in this figure is pretty cool; again, it seems like a second finding that you might push in the main text.

The mechanical explanation that you later derive in the supplemental material should be better reflected in the main text.

References
Eliasmith et al. (2012) A Large-Scale Model of the Functioning Brain. Science.

O'Reilly & Frank (2006) Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural Computation. [And many other papers with Michael Frank]

O'Reilly & Munakata (2000) Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain. Cambridge: MIT Press

Signed,

Gary Cottrell

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
I appreciate the authors' efforts to clarify my concerns, and they successfully addressed most of my concerns. Consequently, the clarity of the manuscript has been much improved, and now I have a much better understanding of the model and find some results, for instance, the finding of incongruent neurons, are interesting. Training a large-scale model of cortico-basal-ganglia-thalamic neural networks on a cognitive task (categorical learning) is a non-trivial problem, and the obtained results are consistent with experimental observations. Therefore, the manuscript may eventually be published. The model (and its future extensions) will also provide a platform to examine biological hypotheses in silico.

However, I strongly doubt whether the model achieved "zero-shot learning". Indeed, I wonder whether the zero-shot learning claimed in the manuscript is a well-defined theoretical concept. I feel the claim only causes unnecessary confusion. The authors said that the model has not been pre-trained on a similar task and hence achieves zero-shot learning. However, cortico-striatal and thalamo-cortical synapses of the model undergo long-term modifications implementing reinforcement learning. I feel that the authors' thought behind the "zero-shot learning" is that the model should be able to perform the same cognitive task as the brain can do so, if the model copies the biological circuits accurately enough. This naive view may hold in principle. However, the binary categorization task demonstrated in the paper is just a simple example of various cognitive tasks, and the model contains the crucial component (e.g., lateral inhibition between "primitives") for solving the particular task. The authors' claim about "zero-shot learning" should be removed from the manuscript before acceptance.

In addition, whether "primitives" (i.e., small-scale cortical functional units) are necessary for solving the task or improving the model's performance has not been addressed. What would happen if the cortical network model did not have the subnetwork structure? Although I do not insist that clarifying this point is necessary for the acceptance of the manuscript, it is regrettable that we cannot see whether the complex model describes the minimal network structure required for solving cognitive tasks (or more specifically, for generating incongruent neurons).

(Remarks on code availability)
I do not have sufficient knowledge about code written in modern computer languages.

Reviewer #3

(Remarks to the Author)
The authors have significantly improved the manuscript, and I believe this model, as it is, can serve as a reference for future studies.

I do have one more suggestion. In the model, there is only one generic type of inhibitory neurons, but recent studies found multiple functional types of inhibitory neurons and cell-type specific connections in cortical areas (see references below, for instance). If the authors could discuss potential links between the model and the observed cell-type specific connectome, I believe it can further strengthen the study.

Pfeffer, C., Xue, M., He, M. et al. Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. Nat Neurosci 16, 1068–1076 (2013). https://doi.org/10.1038/nn.3446

Jiang X, Shen S, Cadwell CR, Berens P, Sinz F, Ecker AS, Patel S, Tolias AS. Principles of connectivity among morphologically defined cell types in adult neocortex. Science. 2015 Nov 27;350(6264):aac9462. doi: 10.1126/science.aac9462. PMID: 26612957; PMCID: PMC4809866.

There is a typo in line 279; “Eventhough”.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
This is an exciting paper, mainly because of the discovery in the model of a type of neural coding that was unknown before and is found in their data after its discovery in the model. This population of neurons codes for when the choice of response is going to be incorrect.

The paper now reviews SPAUN and Emergent, so they responded to that critique.

So, I believe that this work is important because of the discovery of a novel neural code that emerged from the simulations. I noted in my previous review that of the issue of related models were addressed, then I think this rates as publishable in Nature. So, I recommend its publication.

More specific comments/wording suggestions/typos/notes:

This time around, I found the description of the striatal sub circuit and the TAN sub circuit to be confusing. The TAN is just one blob in the striatal sub circuit - wouldn’t it be better to simply refer to it as part (in the model, if not in reality) of the striatal sub circuit. And why not just call this the Basal Ganglia?

Line 24: “predetermined mathematical systems” What do you mean by “predetermined”? There must be some other word to describe your intent here.

Figure 1. You define all of the different acronyms except MSN (medium spiny neurons) which doesn’t appear until line 135. Define it in the caption.

line 118: You mention something will be illustrated by voltage/time graphs, but don’t give a pointer - I believe you’re referring to Figure S1A?

line 279: Eventhough -> Even though

Line 310: arecharacterized -> are characterized

Line 355: backprop is used throughout neural network land, including self-supervised or unsupervised learning; it isn’t only used for supervised learning.

Line 356: The sentence starting with “Confusingly…” is a non sequitur. Either provide some linking sentence or (better) start a new paragraph.

Line 388: physiolgy -> physiology

In the Methods section, you describe some properties of the circuits without giving any references for their origin. Since this is all supposed to be based on evidence, be sure to cite the evidence. This should be true anywhere you describe a gross circuit physiology. For example, at the end of the paragraph from 486-491, you mention that “the connection probability for cortico-cortical connection is 8%.” What evidence is this based on? Other examples include the stipulation that every neuron between the SVCR and the AC has the same number of outgoing as incoming connections, some properties of the striatum, etc.

Again, I did not check the equations carefully; my previous comments about these were responded to.

Signed,
Gary Cottrell

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
The revised manuscript has clarified my concerns. I am pleased to see that the authors have withdrawn their misleading claim of "zero-shot learning". I also agree that the role of subnetwork structure requires an extensive study and a separate publication. I recommend the publication of this manuscript.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors addressed all of my concerns.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
Review of Pathak, et al.

This is an exciting paper, mainly because of the discovery in the model of a type of neural coding that was unknown before and is found in their data after its discovery in the model. This population of neurons codes for when the choice of response is going to be incorrect.

So, I believe that this work is important because of the discovery of a novel neural code that emerged from the simulations. I recommend its publication.

As I have been in favor of publication all along, this review is mostly about style, rather than substance. In particular, the new version does have at least one issue, which should be addressed; basically, the authors go overboard with the fact that the model is not trained. The paragraph describing that is very redundant, and should be shortened considerably. There are some smaller issues, which are addressed in my detailed comments below. For example, there are way too many scare quotes throughout the article.

By the way, in your response to Review 2, you state: “Readers may have the same question as the reviewer. Thus, we added text to the revised manuscript (lines 121-131) that raises these questions and how they can be addressed in future work.”

Lines 121-131 do not do this at all. They are about something else completely. After reading the entire article (I did not read the supplementary material), I still did not find anything resembling this in the text.

More specific comments/wording suggestions/typos/notes:

There are so many of these! Sigh.

Lines 31-48: this is way too much! I would remove completely most of lines 42-47. Here is a rewrite of part of the paragraph (and the beginning of the next), starting with “The model…” on line 37:

“The model is then presented with an experimental task involving visual stimuli and rewards. Separately, non-human primates (NHPs) have also been exposed to this same experimental task. However, the model has not been built based on any of the animals’ physiological or behavioral data. Instead, the model is presented with the same visual stimuli as the animals, as well as simulated rewards. In response, the model learns the same task, at about the same rate, as the NHPs. In doing so, the model produces extended simulated physiological cell firing and synaptic change activity, across multiple separate anatomically modeled brain regions. This activity, which we term “physiological computation”, is then directly

compared with corresponding physiological measures from the monkey experiments.

In sum, we built a physiological computational model of the corticostriatal circuit designed to be mechanistically accurate across multiple scales, from spiking neurons to local field potentials, informed solely by the physiological literature. This circuit…"

If you don't like my rewrite, at least make these paragraphs less repetitive.

You have a great definition of "physiological computation" in the discussion, on lines 377-379. It would be great to move that into the paragraphs above, and then stop using scare quotes around it afterwards.

Similar to my comments above, for the sentence starting on line 67:
Change this:
We developed a generalizable, tractable, biomimetic, multi-scale brain model that categorizes visual stimuli after a delay. The model can be described as carrying out "physiological computation", because it was not directly designed for this task, nor was it "trained" with any empirical data from any target datasets, as per modeling approaches based on artificial neural networks, nor was it designed from pre-existing dynamical systems or other equations or algorithms.

to this:
We developed a generalizable, tractable, biomimetic, multi-scale brain model that categorizes visual stimuli after a delay. The model can be described as carrying out "physiological computation", because it was not directly designed for this task, but only uses bottom-up mechanistic information-processing operations that arise directly from the physiological steps taken by cells within well-defined anatomical circuitry from distinct brain regions.
[I.e., this is one possible place for the definition in the discussion.]

On the issue of scare quotes. Please remove these in lines:
80, 88, 90, 91, 92, 101, 121, 123, 130, 131, 168, 170, 171, 215 (suggest replacing "held" with "maintained", w/o quotes), 230, 234, 268, 273 (here, I suggest deleting the parenthetical "activity" altogether), 305 (here I suggest replacing "discovering" with "predicting" (no quotes)), 359, 395, 396, 758.

Similarly, it seems unnecessary to use italics in lines 361-364, and line 371 (I suggest replacing "back-propagation, or backprop" with one word: backpropagation, not in italics.

Small typos/syntax, etc., issues:

line 90: i.e., "physiological computation." -> i.e., their physiological computation.

line 113: you call it L-FLIC, later, line 121, LFLIC. Pick one, and be consistent for the rest of the paper.

Lines 110-112 are nearly completely repeated in lines 126-128. Condense or reword these two paragraphs, as they are pretty redundant.

Line 121: delete the parenthetical remark.

lines 194-195: What is the point of this sentence fragment starting with "but it is…"? Is this about using rule-based computations? You could just delete it.

paragraph 196-221 is very long; consider breaking it into smaller paragraphs (e.g., maybe start a new one at line 213 with "The activity during.." Up to you.

Figure 2 caption: "reach plateau performance by about trial 300". It looks much more like trial 450 or 500. It certainly doesn't plateau at 300.

Lines 242-243. "these changes in the model are compared directly against empirical changes in nonhuman primates (NHPs) (Figure 2b and 3b)."
Figure 2B does not compare them directly, nor does 3B. What are you referring to here?

Figure 3A: I note here that in the simulation, the cyan line is not above 0 in the stimulus on case.

Line 281: delete "very" in "very significantly".

Line 347: et. al. -> et al. (Typo Detection is the all-around *worst* superpower to have…)

Line 386: not -> are not. Also, the traditional schema for this kind of list of examples seems to call for three examples, not just two. It would be good to add a third here.

Lines 402-403: I think you're pointing to this idea, but I think it might be good to be more explicit about how ICN's could be important if contingencies change? Up to you.

Line 409: “new and never previously presented” -> novel.

Lines 422-423: Change:

other studies strongly suggest feedback inhibition to be performed by SOM cells (86), whereas still other work suggests SOM martinotti cells to cause feedforward inhibition (42, 84).

To:
other studies strongly suggest feedback inhibition is performed by SOM cells (86). Other work suggests SOM martinotti cells cause feedforward inhibition (42, 84).

Line 427: “as seen through” -> “consistent with”

Line 448: Saying your simulation encompasses sensory encoding is going too far. There isn’t anything in your simulations about actual sensory encoding. For example, see your own text on lines 771-772.

Methods section: There are issues here with the format (e.g., compare line 470 with lines 480-481), but I assume those will be taken care of at the page proof stage.

Line 518: evidences -> evidence

Lines 522-523: This is the first mention, I believe, of cortical columns. Your model doesn’t have cortical columns, as far as I can tell. I would delete mention of cortical columns.

Line 717: factors e.g. the -> factors, e.g., the

Lines 742-743. You say that “they are also not typically constructed to generate extended cognitive operations such as the behavioral perception, learning, working-memory, and decision task presented herein.” This is exactly what SPAUN and Emergent do. So this is an incorrect statement.

Line 773: smoothening -> smoothing

Line 835: it’s -> its

signed,
Gary Cottrell

Signed,
Gary Cottrell

(Remarks on code availability)

Reviewer Responses

for Pathak et al.,

"Biomimetic model of corticostriatal micro-assemblies discovers new neural code".

Reviewer #1 (Remarks to the Author):

Reviewer 1:
The authors have dealt with one of the prominent problems in neuroscience, namely, hierarchical modeling of cellular-to-cognitive functioning of the nervous system. They have attempted to reach this goal using a detailed cellular and connectivity model of the cortico-striato-thalamic system, for which there is a veritable mountain of neurobiological evidence. They have verified this model to account for cellular and behavioral evidence in the nonhuman primate. Most important for the present submission, the authors claim to have utilized the model to uncover a novel "neural code". The highlighted features of the authors' model is that it can reveal model outcomes that are, to use the authors' terminology, "zero-trained" or "out-of-distribution inferences".

These are certainly marked strengths of the current paper, and in this sense, are certainly deserving of publication in Nature.

Our reply:
We greatly appreciate these comments from the Reviewer. We have now responded to all reviewer comments, as well as substantial updates to the manuscript, all of which are noted and pointed to within our responses below.

Reviewer 1:
First is the authors' justification of "lateral inhibition assemblies" as evidence for a "computational primitive" of the cortical-striatal circuit. Although this claim is certainly reasonable, lateral inhibition by itself is not terribly convincing as an emerging computational property. In addition, "winner-take-all" functions are hardly biologically reasonable.

Our reply:
This is a very helpful point. These terms "winner take all" and "lateral inhibition assemblies" have been used in the literature to refer to multiple very different proposed mechanisms. The reviewer is correct, some of them are not biologically reasonable. We meant to refer to the set of specific mechanisms that are biologically plausible. We used "winner-take-all" because it is easy-to-understand concept that is familiar to many.

We actually used a set of mechanisms based on well-established anatomical and physiological relations among excitatory pyramidal and local feedback inhibitory cells. The time course of their activity (rapid excitation followed by slower and longer-duration feedback inhibition) and nonlinear computational functions are based on empirical observations (e.g., Rudy et al., 2011; Tremblay et al., 2016). Further, studies show that

this anatomical arrangement is conserved across cortical regions (e.g., Pfeffer et al., 2013; Campagnola et al., 2022).

The revised manuscript now describes these mechanisms and their empirical support in greater detail (lines 93-114, Fig 1b). We now refer to these mechanisms as local-feedback lateral inhibition circuits (L-FLIC).

REFs

Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. Nature neuroscience, 16(8):1068–1076, 2013.

Luke Campagnola, Stephanie C Seeman, Thomas Chartrand, Lisa Kim, Alex Hoggarth, Clare Gam- lin, Shinya Ito, Jessica Trinh, Pasha Davoudian, Cristina Radaelli, et al. Local connectivity and synaptic dynamics in mouse and human neocortex. Science, 375(6585):eabj5861, 2022.

Rudy B, Fishell G, Lee S, Hjerling-Leffler J. Three groups of interneurons account for nearly 100% of neocortical gabaergic neurons. Developmental neurobiology, 71(1):45–61, 2011.

Tremblay R, Lee S, Rudy B. Gabaergic interneurons in the neocortex: from cellular properties to circuits. Neuron, 91(2):260– 292, 2016.

Douglas R, Martin K. Neuronal circuits of the neocortex. Ann. Rev. Neurosci. 2004. 27:419–51.

Reviewer 1:

Aren't there more complex/surprising and thus more interesting spatio-temporal patterns of activity emerging from the interconnected circuitry of neurons having the biophysical properties known for the modeled neurons?

Given the known richness of voltage-dependent channels having differential dendritic and cellular distributions in cortical, striatal, nigral, and thalamic populations, is the proposed model really limited to age-old lateral inhibition? Although this question is open-ended and thus somewhat unfair to the authors, this reviewer finds that other spatio-temporal patterns must be evident and would represent a far stronger argument for the model.

Our reply:

The reviewer's instincts are correct—it is not merely "age-old" lateral inhibition. The model generates spatiotemporal patterns, and we have revised the manuscript to make this explicit.

The model exhibits distinct cortico-striatal behaviors over time, including varying patterns of synchrony and phase locking directly linked to the behavior of the model (or animal) [see Figure 2d, 3a; revised manuscript lines 189-208].

Cortico-striatal activity patterns evolve due to distinct mechanisms in the cortex and striatum, as supported by nonhuman primate (NHP) data [see Figure 2b, 3b, Supplementary Figure S11; revised manuscript lines 170-173, 208-212, 647-705].

Crucially, the model was not trained on NHP neurophysiological data, yet it independently produced spatiotemporal patterns that closely align with NHP data across multiple levels (cells, assemblies, local fields, inter-region synchrony, and phase).

We agree that this strongly supports the model. These points should have been clearer in the original draft, and we have now addressed them.

Reviewer 1:
The fact that the higher-level output of the model represents category learning is exciting, but it should be noted that the number of categories is small indeed, i.e., two. This is consistent with the emphasis on lateral inhibition: on or off. Perhaps this reviewer has missed something important about the category learning accomplished by the model, in which case whatever that is should be re-emphasized.

Our reply:
We chose a two-alternative category decision task because it connects a large body of prior work that has used two-alternative forced choice paradigms. Future work will increase the model's abilities.

Importantly, this was not a simple two-alternative task. The nature of the category learning made it complex. The number of distinct exemplars that the model (and the NHPs) learn is substantial. There were 8 separate prototypes, each with ~1000 exemplars. We have revised the manuscript to make this explicit (see revised manuscript lines 756-758).

It may be useful to also note that the cortical inhibitory feedback mechanism (L-FLIC) that we employ is inherently capable of learning multiple categories, not two. The superficial-layer cell population within a cortical column contains multiple L-FLIC circuit assemblies and comprises hundreds of thousands of cells. Each dendrite in a cortical region has a unique distribution of synapses and strengths, leading to differential activation speeds (faster or slower voltage rise) in response to diverse input patterns, such as dot stimuli.

In short, the same mechanism applied to this architecture can learn many categories, not just two. This is now further explained in lines 491-497 in the revised manuscript.

Reviewer 1:
A second concern which is related to the first is the small number of neurons included in the

model. The proposed model is composed of only hundreds of neurons (without compartments? see below). In these days of widespread computational power, an increased number of neurons per region in the model would seem more than reasonable. In addition, it would provide the basis for detection of more complex spatio-temporal patterns involved in coding schemes.

Our reply:
We again very much agree. Work is under way to run substantially larger numbers of neurons as well as multi-compartment neural renderings. We anticipate that, as the reviewer notes, these will add to the model's abilities to produce more complex spatio-temporal patterns, and potential further neural codes.

We chose the initial size of the model to allow the model's operation to be sufficiently straightforward, transparent, and to explain the desired phenomena in a fully understandable and analyzable way, with no "black boxes" of training to intervene. In other words, we aimed to make the model as simple as possible, but no simpler.

We continue to be excited and surprised at the initial success of even this modest sized model to account for substantial complex physiological and behavioral phenomena, and as the reviewer appropriately points out, we are actively pursuing ongoing study of models that are larger and more complex.

We revised the manuscript to make these points (see manuscript lines 387-403).

Reviewer 1:
Third, an important issue is that the authors point out that both the Striatal and TAN subcircuits are not represented by sets of compartmental neurons, but instead as input/output elements. There is a rich literature documenting the strength of input/output models and their comparison to compartmental models, yet this literature is not cited or used to justify the authors' approach.

Our reply:
We agree. Current work is in process to expand the striatal, pallidal, and TAN regions to supplant simplified input/output entities with physiologically based medium spiny neurons (MSNs) and detailed tonically active neurons (TANs).

As the reviewer points out, modeling MSNs, TANs, and SNc as input/output elements is well supported by extensive studies of dendritic arbors (Carandini 1996) relating synaptic input characteristics and output spiking characteristics (Bernander 1994). Multiple physiological mechanisms have been proposed to incorporate active dendrites and ion channel dynamics into input/output computational units (Silver 2010; Cash 1998; 1999; Tamas 2002; Gasparini 2006).

This has been added to the revised manuscript (lines 140-142, 508-510).

References:
M Carandini, F Mechler, CS Leonard, JA Movshon, Spike train encoding by regular-spiking cells of the visual cortex. J. Neurophysiol.974
(1996).975

S Cash, R Yuste, Input Summation by Cultured Pyramidal Neurons Is Linear and Position-Independent. J. Neurosci. 18, 10–15976
(1998).977

S Cash, R Yuste, Linear Summation of Excitatory Inputs by CA1 Pyramidal Neurons. Neuron 22, 383–394 (1999).978

G Tamás, J Szabadics, P Somogyi, Cell Type- and Subcellular Position-Dependent Summation of Unitary Postsynaptic Potentials979
in Neocortical Neurons. J. Neurosci. 22, 740–747 (2002).980

S Gasparini, JC Magee, State-Dependent Dendritic Computation in Hippocampal CA1 Pyramidal Neurons. J. Neurosci. 26,981
2088–2100 (2006).982

RA Silver, Neuronal arithmetic. Nat. Rev. Neurosci. 11, 474–489 (2010).983

O Bernander, C Koch, RJ Douglas, Amplification and linearization of distal synaptic input to cortical pyramidal cells. J. Neurophysiol.985
72, 2743–2753 (1994)

Minor comments
Reviewer 1:
(1) is “lateral inhabitation” a useful, widely used term? vs lateral inhibition?

Our reply:
We thank the reviewer for catching this typographical error. Lateral inhibition has now been replaced by L-FLIC (revised manuscript lines 106-107)

Reviewer 1:
was the pars reticulata purposely left out of the model?

Our reply:
Yes, SNr is intentionally absent only for this initial model for the sake of simplicity. But Per the reviewer’s comments, we now added a discussion of the SNr to the revised manuscript (manuscript lines 579-585) with the following citations.

REFS
J.S. Schneider, U.E. Olazabal,
Behaviorally specific limb use deficits following globus pallidus lesions in rats,

Brain Research,
Volume 308, Issue 2, 1984, Pages 341-346,

W Hauber, S Lutz, M Münkle,
The effects of globus pallidus lesions on dopamine-dependent motor behaviour in rats,
Neuroscience,
Volume 86, Issue 1, 1998, Pages 147-157

Lai YY, Kodama T, Hsieh KC, Nguyen D, Siegel JM. Substantia nigra pars reticulata-mediated sleep and motor activity regulation. Sleep. 2021 Jan 21;44(1):zsaa151. doi: 10.1093/sleep/zsaa151. PMID: 32808987; PMCID: PMC7819837.,

Reviewer 1:
presynaptic inhibition also has been found among striatal neurons; was this form of inhibition also deliberately omitted? The authors cannot be expected to include every known function or feature in their model; simple justifications to these questions should suffice.

Our reply:
We are incorporating this characteristic into new versions of the model in ongoing work. The existing lateral inhibition between segments of the striatal complex may be producing some of the effects that otherwise would arise from presynaptic inhibition. We now discuss this in the revised manuscript (lines 620-623).

Reviewer #2 (Remarks to the Author):

In this paper, the authors constructed a biologically detailed model of the cortico-basal ganglia-thalamic networks of spiking neurons. Through numerical simulations of the model, the authors validated the model's neural activities using the data recorded from macaque monkeys. The authors showed that neuronal functional subtypes emerge to predict a failure behavioral choice (incongruent neurons) and claimed that the model is "zero trained". The two results sound interesting and constitute the major finding of this study.

However, the present manuscript is hard to follow: it is written in small fonts and densely packed with many lines. It was also a labor to find the information necessary to clarify the details of the study.

These style issues have nothing to do with the quality of the science reported, but they disturb the readers' understanding of the results. Partly due to these difficulties, I could not be fully convinced of the validity of the major claims. I feel that these claims are somewhat overstated.

Our reply
We apologize for our lack of clarity. We have revised the manuscript accordingly

Major comments:
1. The categorization task used should be explained more clearly with illustration (fig. s6). In the current manuscript, fig. s6 is referenced much later than where the task is first explained in ll. 53-55 at the beginning of the Result section. The reader can easily miss the figure and just find a short description of the task in the Methods section (without a reference to Fig. s6).

Our reply
These are very helpful comments and suggestions.

Per the reviewer's suggestion, the references to Figure S6 now appear very early, in revised manuscript line 65, as well as in the Methods section (revised manuscript line 730)

2. In ll. 34-36, the authors explained the "zero trained" condition as: " A far more rigorous standard of generalizability is to ask whether a model built from one kind of data can predict phenomena of a wholly different kind. This explanation confused me because the model does not seem to undergo any training on a different type of task before it is trained on the categorization task.

Our reply:
We have added explanatory text in the revised manuscript lines 36-42 and 405-415 to clarify the concepts and the terminology of "zero-trained", and lines 363-369 and 387-403 to elucidate the non-trivial accomplishments of the model.

The reviewer is correct that a common test of generalizability is training a model on one set of data and then tested on another dataset. Our test was far more rigorous. *Our model was not trained on any data.* Instead, the model was trained on the same behavioral task that the NHPs learned. Thus, the model's properties closely match real neurophysiological data even though no data was ever presented to the model.

We should have made this clearer. This is an important point because it shows that our model is biologically plausible. It can replicate observed neurophysiological data without being trained or tuned by data. We have revised the manuscript accordingly.

Reviewer 2:
The authors might have meant to say that they built the model solely based on anatomy and physiology data and tested its performance on learning behavioral data from the categorization task. If this is the case, however, I do not find a big surprise in that the model could learn the categorization task using its entire dataset. The authors should explain why this is non-trivial before they claim that the "zero trained" feature is an unexpected outcome of the model. The training conditions should also be explained more clearly.

Our reply:
We apologize for not being clear. Our model was not trained on any data. Nonetheless, it matched real neurophysiological data. It even predicted a previously unknown neural property that we subsequently confirm in the data (see revised manuscript lines 370-386). It made a new discovery. We hope the reviewer agrees that this is non-trivial.

Reviewer 2:
3. Potentially related to the above comment, I am not convinced of the importance of the "primitive", which is a modular block of small numbers of excitatory and inhibitory neurons (fig. 1b). The computational benefit of this modular structure was not explored sufficiently in detail. The authors hypothesized that these primitives exist in the cortical layer 2/3 and they perform WTA competition via lateral inhibition. Are these hypotheses supported by experimental evidence or just theoretical assumptions?

Our reply:
We have clarified this and added details in the revised version of the manuscript. Our use of block of excitatory and inhibitory neurons is indeed supported by empirical evidence.

We used a set of mechanisms based on well-established anatomical and physiological relations among excitatory pyramidal and local feedback inhibitory cells. The time course of their activity (rapid excitation followed by slower and longer-duration feedback inhibition) and nonlinear computational functions are based on empirical observations (e.g., Rudy et al., 2011; Tremblay et al., 2016). Further, studies show that this anatomical arrangement is conserved across cortical regions (e.g., Pfeffer et al., 2013; Campagnola

et al., 2022). The revised manuscript now describes these mechanisms and their empirical support in greater detail (revised manuscript lines 93-114 and 491-497, Fig 1b)

Reviewer 2:
4. In ll. 277-281, the authors claimed "Finally, we emphasize the broader point that it was the zero-trained nature of simulations based on biomimetic computation that enabled the emergence of these broad and unexpected findings. Whereas much computational work focuses on the statistical learning of specific data, the present work instead attempts to uncover how low-level physiological spiking, embedded in specific delimited sets of anatomical wiring organizations, can itself give rise to the forms of neural and behavioral results that brains produce." However, I doubt whether this finding is as surprising as the authors claim. As mentioned in my comment 3, the authors hypothesized WTA competition for the primitives, a basic computation essential for categorization tasks. Their large-scale model, therefore, involves a minimal circuit motif necessary to solve the task under consideration. When a large system contains a small system that can solve a task, I find no surprise that the large system can also solve the same task.

Our reply:
We appreciate the opportunity to highlight what makes our model unique. We have clarified these key points in the revised manuscript.

First, our results extend beyond behavior. The model didn't just solve the task—it did so in a brain-like manner. Despite not being trained on neurophysiological data, it replicated such data and even led to a novel discovery (revised manuscript lines 370-386).

Notably, the model exhibits neural spiking patterns, local field potentials, inter-region synchrony, and phase relations—all closely matching monkey physiology—none of which were used to construct or train the model (revised manuscript lines 363–369).

The local-feedback lateral inhibition circuit (L-FLIC) is not a simple winner-take-all network. It is grounded in well-established neuroscience literature (revised manuscript lines 93-114).

We did not intend to suggest that the L-FLIC alone drives category learning. Other physiological and computational components contribute, including working memory mechanisms, local field potential effects, and cross-area synchrony (revised manuscript lines 123-128, 219–225).

Reviewer 2:
The A-correct, A-incorrect, B-correct, and B-incorrect neuronal responses should be defined explicitly. What exactly can A-incorrect and B-incorrect neurons tell about the categorization of input? Is it the probability that an input is unlikely to belong to A or B (i.e., the volume ratio

between the pink shaded area vs its outside in Fig. S6)? Logically, X = not-A does not mean X=B, and similarly, X=not-B does not mean X=A. However, in the alternative choice task (e.g., succade to A or B), not-A may behaviorally imply B as the animal can choose one of the two options. How are the activities of the four functional subtypes combined to generate the model's decision? The computational implications and importance of incongruent activities depend on how the task was designed and how the neural responses were used to solve the task. I might find these details somewhere in the manuscript, but they should be explained more carefully.

Our reply:
We appreciate the reviewer's insightful comments and have revised the manuscript accordingly. We now clarify the definitions of A-correct, A-incorrect, B-correct, and B-incorrect neurons, as well as the data analysis (revised manuscript lines 238–244, 278–285).

The prefix A or B denotes the stimulus category presented in a trial, while the suffix correct or incorrect indicates the trial outcome based on the subject's response. For example, A-incorrect means an A-type stimulus was presented, but the subject mistakenly responded with B instead of A.

Analysis of neural responses across these four trial types revealed distinct neurons selectively responding to one type but not the others—an emergent property of synaptic learning. These neurons are not biologically pre-categorized but develop their selectivity through experience.

Notably, incongruent neurons (ICNs), such as A-incorrect and B-incorrect, do not merely encode stimulus type but also reflect decision-making processes that unfold after stimulus presentation. Strikingly, ICNs spike selectively within the first 200 ms of a trial, effectively "predicting" an incorrect response before the decision is executed.

Further analysis confirms that ICN activity is not solely stimulus-driven; instead, these neurons are also tuned to the upcoming response outcome. While both congruent (A-correct, B-correct) and incongruent (A-incorrect, B-incorrect) neurons encode stimulus category (A vs. B), only ICNs exhibit selectivity for incorrect action selection in the striatum. These findings are detailed in Supplementary Figure S10.

Reviewer 2:
What conditions ensure the emergence of incongruent responses in the model? Do incongruent neurons play any role in switching between exploitation and exploration? How crucial is having such a functional subtype for task performance?

- What conditions ensure the emergence of incongruent responses in the model?

Our reply:

Incongruent neurons (ICNs) emerge under specific conditions in the model, responding when a stimulus from one category is about to be incorrectly classified as the other. In other words, they predict an upcoming incorrect choice long before it occurs within the trial.

Reviewer 2:
Do incongruent neurons play any role in switching between exploitation and exploration?

Our reply:
The reviewer's insights are correct.

ICNs predict an impending incorrect action, suggesting the animal is "choosing" a response that may not be in its best interest. However, these neurons could become valuable if environmental changes require a shift to exploration. In essence, they may serve as a reserve mechanism, allowing the network to escape local exploitation minima when exploration becomes advantageous.

We have added a brief discussion on this point in the revised manuscript and will explore it further in future work. (lines 379-386)

Reviewer 2:
How crucial is having such a functional subtype for task performance?

Our reply:
ICN activity corresponds with pronounced local field potential (LFP) effects, prominently observed in nonhuman primate (NHP) cortico-striatal synchrony (see Figure 3a). This suggests ICNs play a key role in large-scale cortical-striatal signaling.

We have included this in the revised manuscript (lines 372-386). As the reviewer noted, ICNs may also influence exploration vs. exploitation, a topic we will address in future work.

Reviewer 2:
I wonder whether references are cited correctly (for example, refs 16-18). Please check the references.

Our reply:
Thanks you for noting this. This has been corrected in the revised manuscript lines 326-332.

Reviewer 2 - Minor comments:
8. How is the speed of learning compared with the experimental observation?

Our reply:
As the reviewer suggests, the model's learning rate aligns with behavioral learning in macaques. It reaches 80% accuracy within 100–400 trials and maintains this level for the remainder of the session (add to manuscript lines 167–170).

9. Are the responses of TANs crucial for switching exploitation and exploration?

Our reply:
A very useful observation. TANs play a direct role in the model's shift from exploration to exploitation. Early in learning, TAN activity introduces stochasticity to medium spiny neurons (MSNs) in the striatum, promoting exploration. As learning progresses, striosomal (patch) MSNs increasingly inhibit TANs, reducing stochastic influence and making MSN activity more reliant on cortical input. This marks the transition to exploitation, where the model reliably associates input features with specific behaviors.

This progression is now detailed in Supplementary Figure S3 and manuscript lines 147–153.

10. In Fig. 2b, what <W> means? Does it refer to averaging over neurons? I guess some connections are strengthened, as shown in the figure, but the other connections are weakened even if their overall average increases during learning.

Our reply:
The <W> in Figure 2b represents the average synaptic weight across all connections between specified brain regions: cortico-cortical (top panel), cortico-striatal (second panel), and thalamo-cortical (third panel). This clarification has been added to the Figure 2b caption. The analysis functions as the reviewer suggested.

11. The limitations of the model should also be discussed briefly.

Our reply:
We now do. Our model is meant to be first steps in using biomimetic model by elimination artificial neural network data-learning, focusing instead on emergent physiological properties.

But as a first step, it needs further development in the future. This includes incorporating additional key structures, such as the amygdala, hippocampal circuits, and specific thalamic details. This discussion has been added in lines 387–403 of the revised manuscript.

Reviewer #3 (Remarks to the Author):

This study proposes a computational model that can simulate multiple brain regions (cortex, striatum and thalamus) associated with perceptual decision-making and replicates both spikes and LFPs, which have been experimentally observed. As most brain areas do not work alone, it is imperative that we simulate interareal interactions. In this study, the authors propose to combine population models and spiking neural networks to reduce the complexity of multi-areal models, which is an interesting approach. As they constrain the model's structure using experimental data, their study may provide a good reference model for those who develop multi-areal models.

This model aims to replicate the recorded neural responses, while NHPs are performing a simple visual categorization task. Since the task is quite simple, no significant learning is necessary, but the authors implied that their model can provide insights into the learning process in the brain.

In my opinion, the value of this model is that we could probe functional roles of brain areas on visual categorization using their model. Thus, I would like to ask the authors if they tried to quantitatively evaluate the influence of individual brain areas. If one of the areas is disabled, neural and behavioral responses may fail or differ from the observed responses. It may be possible to gain new insights into perceptual decision-making from these failure modes.

Our Reply:
We strongly agree with these comments. Adding experiments to test simulated lesions would significantly enhance the existing work, and we are already planning future studies to explore this in more detail.

For now, we have incorporated several new observations on model lesions based on the reviewer's suggestions (revised manuscript lines 215-218).

A new section (Section VI) has been added to the Supplemental Material, systematically analyzing the effects of disabling three key circuit elements: the SNc, striosomal components of the striatal complex, and TAN neurons. Their functional roles are complex, and our new analysis provides further insight. We have expanded Supplementary Figure S3, and added three new figures (S12, S13, and S14), to illustrate these findings.

Key Findings:
SNc Disablement: Corticostriatal synapses gradually weakened, leading MSNs to rely more on inputs from TANs rather than cortical inputs (Figure S12).

Striosomal Disablement: Learning no longer influenced TAN activity, causing perseverative behavior, where the model produced the same response regardless of

input (Figure S12). Striosomal neurons project to both TANs and SNc via inhibition. As previously shown, striosomal suppression of TANs shifts behavior from exploration to exploitation (Supplementary Figure S3, revised manuscript lines 149–153).

TAN Disablement: Effects varied based on synaptic learning parameters. When learning parameters was reward-biased, TAN disablement led to perseverative behavior. However, disrupting TANs had no effect when learning parameters was punishment-biased. This suggests possible redundancy between the TAN and SNc systems, where SNc may compensate for TAN disruption to maintain learning (Supplementary Figure S13, 14 and Supplementary Materials lines 118–126).

These findings suggest new experimental tests, which we plan to explore in future work.

Reviewer 3:
The authors claimed that their model can predict the existence of incongruent neurons, which are experimentally confirmed. However, I am concerned that this incongruent neuron could be just a byproduct of winner-take-all mechanisms. Let's consider three sets of neurons. Stimulus A, Stimulus B and none-of-the-above (NOA). Due to lateral inhibition, we expect only one of the three sets to be active at a time. In addition, we can further assume that Stimulus A or B is trained to become quiescent when an incongruent stimulus is presented. That is, when stimulus A is presented, Stimulus B neurons become silent, and vice versa. Then, if Stimulus A fails to activate Stimulus A neurons, NOA neurons will be active due to the lack of lateral inhibition, which will be consistent with incongruent neuron behaviors. Based on this line of reasoning, I am not sure what we can learn from their model's predictions.

Our reply:
The reviewer raises an important thought experiment regarding incongruent neurons (ICNs). While ICNs could have arisen from lateral inhibition, their behavior extends beyond simply signaling the opposing category. They selectively activate when an incorrect behavioral response is about to occur.

In other words, ICNs do not merely represent a "wrong" or "leftover" category due to lateral inhibition—they actively predict an impending incorrect choice. This is a novel discovery, confirmed by neurophysiological recordings in nonhuman primates (NHPs).

Further analysis of model ICNs revealed that synaptic receptive fields in cortical cells undergo differential training via striatal-thalamo-cortical feedback, aligning correct cortical activity with incorrect striatal action-selection (MSN) responses.

This is now discussed in detail in Figures 4, Supplementary Figures S8–S10, and revised manuscript lines 226–289, 370–378. Additionally, as suggested by Reviewer 2, we propose that ICNs may also play a role in exploration vs. exploitation (manuscript lines 379–386).

Reviewer 3:
In the Results section, the main cortical area of interest is "anterior cortex", but Methods section describes "prefrontal cortex", not anterior cortex. As the cortex model relies on a generic structure, such mix-up may not matter too much, but to maintain clarity of this study and not to confuse the readers, this mix-up should be fixed properly.

Our reply:
The reviewer is providing a very useful comment, and we have carefully changed the revised manuscript to solely refer to 'anterior cortex' in the model.

Reviewer 3:
In lines 18-25, the author stated: "However, most of these models have been optimized for either mechanistic accuracy or to elicit emergent cognition, but not both. On the one hand, there are powerful models with biologically-based systems that reproduce extensive physiologies, from single-cell spiking through generation of oscillations and other critical emergent dynamics. Yet, while these can accurately simulate brain-wide phenomena (7) such as epileptic seizure genesis and propagation (8, 9), they are not designed to generate cognitive processes, such as learning. On the other hand, some of the most remarkable cognitive models are those grounded in artificial neural networks (10–15). Yet, while this is a class of models that effectively "learn," they do so without reliance on biologically realistic (biomimetic) mechanisms at the cellular scale."

I believe these statements are not accurate. A line of study uses biologically-realistic circuits to probe neural correlates of high-level cognitive functions such as decision-making, working memory and integration of sensory evidence. Also, all major simulation platforms, such as NEST and NEURON, support long-term synaptic plasticity and can be used to study 'learning' in the brain. The general learning algorithm for spiking neural networks, which corresponds to 'backpropagation' in deep learning, remains missing, but the lack of learning algorithm must not be confused with the lack of study of biological learning.

Selected References
Goldman, M. S., Compte, A., & Wang, X. J. (2009). Neural Integrator Models. In Encyclopedia of Neurosci (pp. 165-178). Elsevier Ltd. https://doi.org/10.1016/B978-008045046-9.01434-0
Wang XJ. Theory of the Multiregional Neocortex: Large-Scale Neural Dynamics and Distributed Cognition. Annu Rev Neurosci. 2022 Jul 8;45:533-560. doi: 10.1146/annurev-neuro-110920-035434. PMID: 35803587.
Wang XJ. Decision making in recurrent neuronal circuits. Neuron. 2008 Oct 23;60(2):215-34. doi: 10.1016/j.neuron.2008.09.034. PMID: 18957215; PMCID: PMC2710297.
Hill S, Tononi G. Modeling sleep and wakefulness in the thalamocortical system. J Neurophysiol. 2005 Mar;93(3):1671-98. doi: 10.1152/jn.00915.2004. Epub 2004 Nov 10. PMID: 15537811.

Wagatsuma N, Potjans TC, Diesmann M, Fukai T. Layer-Dependent Attentional Processing by Top-down Signals in a Visual Cortical Microcircuit Model. Front Comput Neurosci. 2011 Jul 8;5:31. doi: 10.3389/fncom.2011.00031. PMID: 21779240; PMCID: PMC3134838.

Our reply:
The reviewer's comments are highly valuable. We fully agree that various simulation platforms and studies have been effectively used to investigate synaptic plasticity, learning, and network dynamics.

We have revised the manuscript to acknowledge these and other earlier works, including all of these cited by the reviewer, which emphasize balancing mechanistic accuracy with emergent cognition (manuscript lines 25–28). Additionally, we have expanded the discussion on this topic (lines 326–340).

Our work specifically highlights the complete absence of artificial neural network learning (e.g., backpropagation), replacing it with a more biologically grounded approach, while acknowledging extensive work in the existing literature on biological learning. We now make this clear.

Reviewer 3:
I strongly recommend the authors to remove the statements regarding 'zero-trained' and clarify what their proposed model actually achieves. The authors are implying that their proposed model can perform zero-shot learning (zero-trained) or make predictions on novel inputs (out-of-distribution inference). While zero-shot learning and out-of-distribution are important topics in the field of deep learning, the authors do not provide any results relevant to zero-shot learning and out-of-distribution inference in their study.

Our reply:
We have revised the manuscript per the reviewer's suggestions.

To clarify, the manuscript demonstrates that the model can make predictions and perform "out-of-distribution" tasks despite having no training on animal data—literally zero training. The reviewer correctly notes that "zero-shot" and "zero-training" learning typically refer to deep learning. However, our model exhibits this behavior, as it can perform tasks without any prior training. We should have made this clearer in the original manuscript. Now, we do. We now state this explicitly (see revised manuscript lines 36–42 and 404–415).

Reviewer 3:
Also, the authors need to clarify what 'multi-scale' means in their manuscript. Based on the phrase, "from spiking neurons to local field potentials", they appear to be referring to both

structure (e.g., neurons or brain areas) and neural responses (e.g., spikes and LFPs), but to avoid confusion, they should clarify what they mean.

Our reply:
We appreciate the opportunity to clarify these statements. We refer to both structures and activities. The work spans multiple structural scales by anatomically incorporating different neuron types, wired into assemblies with distinct wiring diagrams. This, in turn, gives rise to spikes, local field potentials (measured across cell neighborhoods), regional synchrony, and phase relationships among regions—phenomena that can only be measured across multiple areas rather than locally. We have revised the manuscript (lines 74–79) to state this more clearly.

Reviewer 3:
In line 106, the authors stated: “The arrangement used here is very distinct from ANN approaches”

ANN has not been defined. If ANN means artificial neural networks, the authors should define what kind of ANN approaches they are referring to.

Our reply:
Thank you for noticing this. To avoid the unclear reference, the phrase that the reviewer cites has been changed to eliminate “ANN” ; it now reads: “The arrangement used here is notably distinct from many other approaches that have been used to generate lateral inhibition.” (revised manuscript line 96).

Reviewer 3:
In lines 126-128, the authors stated: “The interconnections between these subcircuits support the process of learning. As the model experiences many trials, there are changes to synaptic strength for cortico-cortical, cortico-striatal, and thalamo-cortical contacts that cause changes in behavior”

I think the authors should show how these interareal connections evolve over time. As the model depends on various cell assemblies, they should show the evolution of weights between cell assemblies.

Our reply:
We very much agree with this suggestion. We have now added a new supplemental figure S11, which gives extensive descriptions and illustration of the changing evolution of weights between cell assemblies.

Reviewer 3:
In line 156, the authors stated: “For the simulation, we developed a similar metric called Local Spike Summation (LSS), a simple measure of simulated neuronal activity defined in Method”

Earlier works simulated LFPs in multiple ways, but LFPs are generally thought to reflect synaptic inputs, whereas LSS reflect outputs of neurons. The authors should, at least, justify why they used LSS to approximate LFPs. As their proposed model aims to produce multi-scale responses, it is important to simulate LFPs as accurately as possible.

Our reply:
We agree and have clarified this further (manuscript lines 769–781) per the reviewer’s suggestion.

Many LFP approximation methods (e.g., Mazzoni 2015) convolve EPSPs over spatial filters. Our goal was to relate individual cell responses to LFPs and measure synchrony and phase relationships among fields. To achieve this, we used LSSs, which combine individual cell responses. To confirm their validity, we demonstrated similarities between model LSSs and empirical LFPs.

Reviewer 3:
In line 158, the authors stated: “one noticable difference is that NHP brains do not show as marked a decrease in spiking after stimulus offset as the model (Figure 2C)”

Can the authors provide any insights into this discrepancy? Alternatively, can they explain why it is not important?

Our reply:
This is a fair point. Models inevitably differ from the actual brain, and we emphasize these differences to highlight features that future work may need to address.

In this case, real PFC neurons receive substantial input from multiple cortical and subcortical regions, which are not modeled here. The model has no extraneous inputs or artificially injected noise, so some discrepancies between empirical and model data are expected. These differences may be meaningful, and further studies are underway.

That said, we observed notable correspondences in inter-area cortico-striatal synchronies and their changes in both the model and macaques. These points are discussed in manuscript lines 185–189.

Reviewer #4 (Remarks to the Author):

This is an exciting paper, mainly because of the discovery in the model of a type of neural coding that was unknown before and is found in their data after its discovery in the model. This population of neurons codes for when the choice of response is going to be incorrect.

Our reply:
We very much appreciate these positive statements from the reviewer.

Since I took the theory track at Cornell, I never got diffEQ, so I did not check the equations carefully, but I will note that they switch in the article between W^{ij} meaning the weight from unit j to unit i (e.g., line 319) to vice versa (e.g., line 520) and back (line 99, suppl. mat.). Be consistent.

Our reply:
We appreciate the reviewer catching this typo. It has now been fixed (see revised manuscript lines 675 and 679)

Reviewer 4:
My main problem with the paper is that this type of model has been done before, perhaps not with such exacting precision with respect to the neurophysiological data, but the spirit is the same, by Randy O'Reilly and Yuko Munakata (2000 and subsequent versions), and there is no reference to this work. In that book, they simulate learning a go-no go task with a model basal ganglia using the Emergent simulator. Similar to the current model, the system is wired up in the same manner as the basal ganglia, and it is trained by the reinforcement learning system intrinsic to the model. Similarly, there's no reference to Chris Eliasmith's work with SPAUN.

In addition, you say "those models were designed to elicit a circumscribed family of behaviors centered on reinforcement learning and were not constructed in a manner that allows them to be easily extended beyond their original scope." This is certainly not true of Eliasmith's model, which shows the ability to learn seven different tasks. And while I haven't read all of them, O'Reilly, especially with Michael Frank, has published numerous papers on his model's explanation of working memory, etc.

I'm sure there are many important differences (for example, your model includes oscillatory behavior, crucial to some aspects of the model, which differs from O'Reilly's model if I recall correctly). However, it is still necessary to distinguish this model from those earlier works.

So, I believe that this work is important because of the discovery of a novel neural code that emerged from the simulations. However, I would like to see how it relates to earlier models in the same vein. If that issue were addressed, I think this rates as publishable in Nature Communications.

Our reply:
We fully agree. Incorporating O'Reilly's and Eliasmith's work is valuable, as our model builds on their contributions. Both Emergent and Spaun were groundbreaking in advancing computational principles in neural systems. The revised manuscript now references and discusses Emergent and Spaun and discusses how our model serves a bridging role (revised manuscript lines 25-26, 301–325).

Reviewer 4:
What would really nail the data would be if you can use your idea of promoting learning by stopping trials that the incongruent neurons predict will be an error in an actual monkey. I don't expect that, given the difficulty, expense, and time necessary, but wow, if you did that, it would be a paper for the ages.

Our reply:
We (of course) love this idea. In fact, we aim to do just that in future experiments.

Minor comments.
Reviewer 4:
More specific comments/wording suggestions/typos/notes:
Introduction: there is no reference to any of Randy O'Reilly's work??

Our reply:
Indeed, our mistake. O'Reilly's work is highly relevant. We have added a substantial discussion of O'Reilly's, as detailed above.

Reviewer 4:
"These latter types of predictions are said to be zero-trained."
I've never heard the concept of "zero-trained" - I've heard of zero-shot. If you are going to introduce new terminology, say instead "we call these models "zero-trained"".

However, that seems like an anomaly, since you do train it - by reinforcement learning, which is built into the system. I think a better description would be to call it "intrinsically-trained". But I will leave that up to the authors.

Our reply:
We agree with the reviewer and have added the phrase "we term these models 'zero-trained' " (e.g., revised manuscript line 37).

Zero-trained emphasizes a crucial point: the model was not trained on any neurophysiological data (i.e., zero), yet its activity closely matches neurophysiological recordings from NHPs. This contrasts with models trained on such data and tested on held-out samples. Our model goes beyond data-fitting approaches, by generating neurophysiological-like activity de novo.

We did not make this explicit in the original manuscript but have now done so (revised manuscript lines 36–42 and 404–415).

Reviewer 4:
line 91: lateral inhibitation -> lateral inhibition

Fig1 caption: Detailed discussion in provided in the Methods
-> Detailed discussion is provided in the Methods

Our reply:
Thank you: typo has been corrected.

Reviewer 4:
line 155: no need for scare quotes around local field potentials
line 160: no need for scare quotes around phase locking value.
There may be other examples of this.

Our reply:
We agree. Done.

Reviewer 4:
line 161: reference 20: EG Antzoulatos, EK Miller (2014). Glancing at this paper, I take it that the strength of connections between the two areas (striatum and PFC) increase overall - i.e., both ways. In Figure 3b, you also show this connectivity increasing, but it isn't clear whether this is from stratum to your simulated frontal lobes or vice-versa. Can you be more specific than "portico-striatal"? Does this graph work both ways?

Our reply:
This is a helpful clarifying question. In Figure 3b, we show the increase in synaptic weights from anterior cortical to striatal cells, separately displayed because they are modulated by reward-driven dopamine. Other connections also change, as the reviewer notes. Both cortico-cortical and thalamo-cortical synapses adapt via their internal plasticity rules, as shown in Figure 2b. Figure 3b specifically illustrates cortex-to-striatum connections—this is now clarified in the caption.

Reviewer 4:
Line 186: Commprising -> Comprising

Figure 3A caption: highlighted in gray, not yellow. The colors of the lines are incorrectly specified (they are blue and green, not red and light red).

Figure 3B Figure: Correct me if I'm wrong, but you seem to have incorrect and correct flipped here.

Our reply:
Thank you. The typos have been corrected.

Reviewer 4:
Last line of Figure 4 caption: The sentence says “would improve” - so this is a prediction. If you were able to do this experiment, it would be especially confirmatory, although I know experiments with NHPs are expensive and time consuming.

Our reply:
Indeed. That would be great. It is in our future plans after we secure funding for further experiments.

Reviewer 4:
Lines 251-252 - please remove the scare quotes.

line 269: NHPs are not “constructed”! Try “evolved”.

Our reply:
Done.

Reviewer 4:
Line 285 and others: You describe this as “working memory”? Can you elaborate on why you call it that?

Our reply:
We use working memory because the task includes a short (1000 ms) delay, requiring the NHP to retain cued information. This aligns with a large body of neurophysiological research on working memory, which we now clarify in the revised manuscript (see lines 193-203, 688-689).

In equations 30 and 31, I believe you have switched from W^{ij} meaning the weight from j to i (as in equation 3) to the reverse. Hopefully this is correct in the code!

Our reply:
This was indeed a typo (only in the text, not in the code). It has now been fixed. Thanks.

It is unclear what is going on in equation 29, as there is only one index on the weight, that is, "i", which suggests this is a solely pre-synaptic learning rule, unlike the rule in equation 28. Explain this, or fix it.

Our reply:
This does require further explanation: equation 29 is a specific form of Hebbian learning (derived from eq. 28). There are in fact two indices in eq.29; one is superscript "i" whereas the other is subscript "x". The "i" refers to the presynaptic "i-th" cortical neuron, while the "x" refers to the label of the target striatal block (the values of x in the current model are simply 1 or 2). We use a different index convention for striatum because the striatal block in the current simulation is implemented differently from cortex, as described in equations 12 and 17. We have added new additional text in the revised ms (lines 656-658) to clarify these points.

Equation 33 reverts to the equation 3 definition of which way the weight W^{ij} goes.
Our reply:
Yes, this is in fact the correct convention, and we now have corrected the other instances of this to conform, as mentioned in responses to previous comments above.

Line 501: you need a reference for the BCM rule. Bienenstock, Cooper, and Munro 1982.
Our reply:
We have added this reference – thank you (revised manuscript reference 98).

Line 502: typo: "formas" either this is supposed to be "form as" or "formulas"
Our reply:
Thank you. This typo has now been corrected. It should be "form as".

Reviewer 4:
In the Computational Methods section (starting line 547), you don't mention Chris Eliasmith's SPAUN model using the Semantic Pointer Architecture. How does NeuroBlox differ from his simulator, which seems to be addressing very similar concerns and uses actual spiking neurons. Or alternatively, how does it differ from Emergent (O'Reilly)?

Our reply:
Indeed. We now correct this by adding reference to both SPAUN and Emergent in Computational Methods section (revised manuscript line 710). As detailed above, we have also added a discussion of how our model bridges between these approaches and neurophysiological experiments (revised manuscript lines 301-325).

Reviewer 4
Supplementary Material

Line 8: (NG-NMM) model, -> mass model (NG-NMM),

Line 24 Supplimental -> Supplemental

Our reply:
Fixed. Thanks.

Reviewer 4:
Figure S5 caption: "Striatal spiking shows a post-stimulus-onset silence that was subsequently found in empirical data;" This seems like a second discovery. Why don't you also emphasize this in the main text? Here, in terms of "silence", are you referring to the pause in spiking around 40-60ms? This appears to be borne out in the local spike summation, although the timing is not quite matched by the empirical data. The match between D and E in this figure is pretty cool; again, it seems like a second finding that you might push in the main text.

The mechanical explanation that you later derive in the supplemental material should be better reflected in the main text.

Our reply:
We appreciate this. We have added this observation to the main text. (see manuscript lines 160-163).

Reviewer 4:
References

Eliasmith et al. (2012) A Large-Scale Model of the Functioning Brain. Science.
O'Reilly & Frank (2006) Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural Computation. [And many other papers with Michael Frank]
O'Reilly & Munakata (2000) Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain. Cambridge: MIT Press

Our reply:
Agreed – description and discussion of this work is included in the revised manuscript, and these citations are now included in our references.

Reviewer #2 (Remarks to the Author):

I appreciate the authors' efforts to clarify my concerns, and they successfully addressed most of my concerns. Consequently, the clarity of the manuscript has been much improved, and now I have a much better understanding of the model and find some results, for instance, the finding of incongruent neurons, are interesting. Training a large-scale model of cortico-basal-ganglia-thalamic neural networks on a cognitive task (categorical learning) is a non-trivial problem, and the obtained results are consistent with experimental observations. Therefore, the manuscript may eventually be published. The model (and its future extensions) will also provide a platform to examine biological hypotheses in silico.

Our reply:
We appreciate the positive comments and support.

Reviewer 2:
However, I strongly doubt whether the model achieved "zero-shot learning". Indeed, I wonder whether the zero-shot learning claimed in the manuscript is a well-defined theoretical concept. I feel the claim only causes unnecessary confusion. The authors said that the model has not been pre-trained on a similar task and hence achieves zero-shot learning. However, cortico-striatal and thalamo-cortical synapses of the model undergo long-term modifications implementing reinforcement learning. I feel that the authors' thought behind the "zero-shot learning" is that the model should be able to perform the same cognitive task as the brain can do so, if the model copies the biological circuits accurately enough. This naive view may hold in principle. However, the binary categorization task demonstrated in the paper is just a simple example of various cognitive tasks, and the model contains the crucial component (e.g., lateral inhibition between "primitives") for solving the particular task. The authors' claim about "zero-shot learning" should be removed from the manuscript before acceptance.

Our reply:
These are very helpful points. We have entirely removed any claims about "zero shot learning" from the manuscript.

Reviewer 2:
In addition, whether "primitives" (i.e., small-scale cortical functional units) are necessary for solving the task or improving the model's performance has not been addressed. What would happen if the cortical network model did not have the subnetwork structure? Although I do not insist that clarifying this point is necessary for the acceptance of the manuscript, it is regrettable that we cannot see whether the complex model describes the minimal network structure required for solving cognitive tasks (or more specifically, for generating incongruent neurons).

Our reply:
We agree with the reviewer that understanding how the subnet structure contributes to model function is an important question. Our ongoing work is addressing this by varying the contribution of different primitives and identifying the minimal required structure. However, the necessary analyses and figures are beyond the scope—and likely the word and figure limits—of this manuscript. This will be the focus of our next paper.

Readers may have the same question as the reviewer. Thus, we added text to the revised manuscript (lines 121-131) that raises these questions and how they can be addressed in future work.

Reviewer #3 (Remarks to the Author):

The authors have significantly improved the manuscript, and I believe this model, as it is, can serve as a reference for future studies.

Our reply:
We thank the reviewer for helpful comments. They improved our manuscript.

Reviewer 3:
I do have one more suggestion. In the model, there is only one generic type of inhibitory neurons, but recent studies found multiple functional types of inhibitory neurons and cell-type specific connections in cortical areas (see references below, for instance). If the authors could discuss potential links between the model and the observed cell-type specific connectome, I believe it can further strengthen the study.

Pfeffer, C., Xue, M., He, M. et al. Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. Nat Neurosci 16, 1068–1076 (2013). https://doi.org/10.1038/nn.3446

Jiang X, Shen S, Cadwell CR, Berens P, Sinz F, Ecker AS, Patel S, Tolias AS. Principles of connectivity among morphologically defined cell types in adult neocortex. Science. 2015 Nov 27;350(6264):aac9462. doi: 10.1126/science.aac9462. PMID: 26612957; PMCID: PMC4809866.

Our reply:
This is very helpful. Per the reviewer's request, we have referenced this work and added a discussion of potential links between the model and the observed cell-type specific connectome (lines 415 -424).

Reviewer 3:
There is a typo in line 279; "Even though".

Our reply
Thank you. This has been corrected.

Reviewer #4 (Remarks to the Author):

This is an exciting paper, mainly because of the discovery in the model of a type of neural coding that was unknown before and is found in their data after its discovery in the model. This population of neurons codes for when the choice of response is going to be incorrect.

Our reply:
Thank you.

Reviewer 4:
The paper now reviews SPAUN and Emergent, so they responded to that critique.

So, I believe that this work is important because of the discovery of a novel neural code that

emerged from the simulations. I noted in my previous review that if the issue of related models were addressed, then I think this rates as publishable in Nature. So, I recommend its publication.

Our reply:
Thank you for the excellent suggestion.

Reviewer 4:
More specific comments/wording suggestions/typos/notes:

This time around, I found the description of the striatal sub circuit and the TAN sub circuit to be confusing. The TAN is just one blob in the striatal sub circuit - wouldn't it be better to simply refer to it as part (in the model, if not in reality) of the striatal sub circuit. And why not just call this the Basal Ganglia?

Our reply:
This is very helpful. Per the reviewer's suggestion, we now use the terms striatal sub circuit and Basal Ganglia to simplify the description. We agree that this is clearer.

Reviewer 4:
Line 24: "predetermined mathematical systems" What do you mean by "predetermined"? There must be some other word to describe your intent here.

Our reply:
This was indeed confusing. We now simply say "other related algorithms" (line 24).

Reviewer 4:
Figure 1. You define all of the different acronyms except MSN (medium spiny neurons) which doesn't appear until line 135. Define it in the caption.

Our reply:
We now defined it the first time we use the acronym. Thanks.

Reviewer 4:
line 118: You mention something will be illustrated by voltage/time graphs, but don't give a pointer - I believe you're referring to Figure S1A?

Our reply:
We fixed this typo. The voltage time graph actually refers to Fig 1 b.

Reviewer 4:
line 279: Eventhough -> Even though

Our reply:
Thank you. This has been corrected.

Reviewer 4:
Line 310: are characterized -> are characterized

Our reply:
Thank you. This has been corrected.

Reviewer 4:
Line 355: backprop is used throughout neural network land, including self-supervised or unsupervised learning; it isn't only used for supervised learning.

Our reply:
Thank you. This has been corrected.

Reviewer 4:
Line 356: The sentence starting with "Confusingly…" is a non sequitur. Either provide some linking sentence or (better) start a new paragraph.

Our reply:
Indeed. We started a new paragraph.

Reviewer 4:
Line 388: physiolgy -> physiology

Our reply:
Thank you. This has been corrected.

Reviewer 4:
In the Methods section, you describe some properties of the circuits without giving any references for their origin. Since this is all supposed to be based on evidence, be sure to cite the evidence. This should be true anywhere you describe a gross circuit physiology. For example, at the end of the paragraph from 486-491, you mention that "the connection probability for cortico-cortical connection is 8%." What evidence is this based on? Other examples include the stipulation that every neuron between the SVCR and the AC has the same number of outgoing as incoming connections, some properties of the striatum, etc.

Our reply:
Per the reviewer's suggestion, we have added specific references to support all of these statements (revised manuscript lines 517-519). The revised manuscript also includes Table A1 in the supplementary material. It contains a list of anatomical and physiological references guiding each of the circuit components.

This is an exciting paper, mainly because of the discovery in the model of a type of neural coding that was unknown before and is found in their data after its discovery in the model. This population of neurons codes for when the choice of response is going to be incorrect.

So, I believe that this work is important because of the discovery of a novel neural code that emerged from the simulations. I recommend its publication.

We thank the reviewer for their support for our work.

As I have been in favor of publication all along, this review is mostly about style, rather than substance. In particular, the new version does have at least one issue, which should be addressed; basically, the authors go overboard with the fact that the model is not trained. The paragraph describing that is very redundant, and should be shortened considerably. There are some smaller issues, which are addressed in my detailed comments below. For example, there are way too many scare quotes throughout the article.

By the way, in your response to Review 2, you state: “Readers may have the same question as the reviewer. Thus, we added text to the revised manuscript (lines 121-131) that raises these questions and how they can be addressed in future work.”

Lines 121-131 do not do this at all. They are about something else completely. After reading the entire article (I did not read the supplementary material), I still did not find anything resembling this in the text.

We erroneously had omitted this. We now have made a statement that clearly responds (corresponding to what we responded to reviewer 2)

More specific comments/wording suggestions/typos/notes:

Lines 31-48: this is way too much! I would remove completely most of lines 42-47. Here is a rewrite of part of the paragraph (and the beginning of the next), starting with “The model…” on line 37:

“The model is then presented with an experimental task involving visual stimuli and rewards. Separately, non-human primates (NHPs) have also been exposed to this same experimental task. However, the model has not been built based on any of the animals’ physiological or behavioral data. Instead, the model is presented with the same visual stimuli as the animals, as well as simulated rewards. In response, the model learns the same task, at about the same rate, as the NHPs. In doing so, the model produces extended simulated physiological cell firing and synaptic change activity, across multiple separate anatomically modeled brain regions. This activity, which we term “physiological computation”, is then directly compared with corresponding physiological measures from the monkey experiments.

In sum, we built a physiological computational model of the corticostriatal circuit designed to be mechanistically accurate across multiple scales, from spiking neurons to local field potentials, informed solely by the physiological literature. This circuit…”

If you don’t like my rewrite, at least make these paragraphs less repetitive.

Thanks for the suggested rewrite. We have incorporated a slightly edited version of it in lines 37-48.

You have a great definition of “physiological computation” in the discussion, on lines 377-379. It would be great to move that into the paragraphs above, and then stop using scare quotes around it afterwards.

Suggestion taken.

Similar to my comments above, for the sentence starting on line 67:
Change this:
We developed a generalizable, tractable, biomimetic, multi-scale brain model that categorizes visual stimuli after a delay. The model can be described as carrying out “physiological computation”, because it was not directly designed for this task, nor was it “trained” with any empirical data from any target datasets, as per modeling approaches based on artificial neural networks, nor was it designed from pre-existing dynamical systems or other equations or algorithms.

to this:
We developed a generalizable, tractable, biomimetic, multi-scale brain model that categorizes visual stimuli after a delay. The model can be described as carrying out “physiological computation”, because it was not directly designed for this task, but only uses bottom-up mechanistic information-processing operations that arise directly from the physiological steps taken by cells within well-defined anatomical circuitry from distinct brain regions.
[I.e., this is one possible place for the definition in the discussion.]

Thank you for this suggested rewrite. We have incorporated a slightly edited version of this in lines 67-69.

On the issue of scare quotes. Please remove these in lines:
80, 88, 90, 91, 92, 101, 121, 123, 130, 131, 168, 170, 171, 215 (suggest replacing “held” with “maintained”, w/o quotes), 230, 234, 268, 273 (here, I suggest deleting the parenthetical “activity” altogether), 305 (here I suggest replacing “discovering” with “predicting” (no quotes)), 359, 395, 396, 758.

Similarly, it seems unnecessary to use italics in lines 361-364, and line 371 (I suggest replacing “back-propagation, or backprop” with one word: backpropagation, not in italics.

All these suggestions have been incorporated.

Small typos/syntax, etc., issues:

line 90: i.e., “physiological computation.” -> i.e., their physiological computation.

line 113: you call it L-FLIC, later, line 121, LFLIC. Pick one, and be consistent for the rest of the paper.

Lines 110-112 are nearly completely repeated in lines 126-128. Condense or reword these two paragraphs, as they are pretty redundant.

Line 121: delete the parenthetical remark.

lines 194-195: What is the point of this sentence fragment starting with “but it is…”? Is this about

using rule-based computations? You could just delete it.

paragraph 196-221 is very long; consider breaking it into smaller paragraphs (e.g., maybe start a new one at line 213 with “The activity during..” Up to you.

All suggestions have been incorporated.

Figure 2 caption: “reach plateau performance by about trial 300”. It looks much more like trial 450 or 500. It certainly doesn’t plateau at 300.

This error has been corrected.

Lines 242-243. “these changes in the model are compared directly against empirical changes in nonhuman primates (NHPs) (Figure 2b and 3b).”
Figure 2B does not compare them directly, nor does 3B. What are you referring to here?

The figure reference has been corrected.

Figure 3A: I note here that in the simulation, the cyan line is not above 0 in the stimulus on case.

We have changed the caption to say “… Values that are significantly above zero…” This instance is not significantly different from zero.

Line 281: delete “very” in “very significantly”.

Line 347: et. al. -> et al. (Typo Detection is the all-around *worst* superpower to have…)

Line 386: not -> are not. Also, the traditional schema for this kind of list of examples seems to call for three examples, not just two. It would be good to add a third here.

Lines 402-403: I think you’re pointing to this idea, but I think it might be good to be more explicit about how ICN’s could be important if contingencies change? Up to you.

Line 409: “new and never previously presented” -> novel.

All changes incorporated.

Lines 422-423: Change:

other studies strongly suggest feedback inhibition to be performed by SOM cells (86), whereas still other work suggests SOM martinotti cells to cause feedforward inhibition (42, 84).

To:
other studies strongly suggest feedback inhibition is performed by SOM cells (86). Other work suggests SOM martinotti cells cause feedforward inhibition (42, 84).

Line 427: “as seen through” -> “consistent with”

Suggestion taken.

Line 448: Saying your simulation encompasses sensory encoding is going too far. There isn't anything in your simulations about actual sensory encoding. For example, see your own text on lines 771-772.

Point taken. We removed sensory encoding.

Methods section: There are issues here with the format (e.g., compare line 470 with lines 480-481), but I assume those will be taken care of at the page proof stage.

Thanks for pointing out. We corrected indentation issue.

Line 518: evidences -> evidence

Lines 522-523: This is the first mention, I believe, of cortical columns. Your model doesn't have cortical columns, as far as I can tell. I would delete mention of cortical columns.

Line 717: factors e.g. the -> factors, e.g., the

Corrections made.

Lines 742-743. You say that "they are also not typically constructed to generate extended cognitive operations such as the behavioral perception, learning, working-memory, and decision task presented herein." This is exactly what SPAUN and Emergent do. So this is an incorrect statement.

We have corrected the statement.

Line 773: smoothening -> smoothing

Line 835: it's -> its

Typos corrected.