A Goal-Driven Approach to Systems Neuroscience

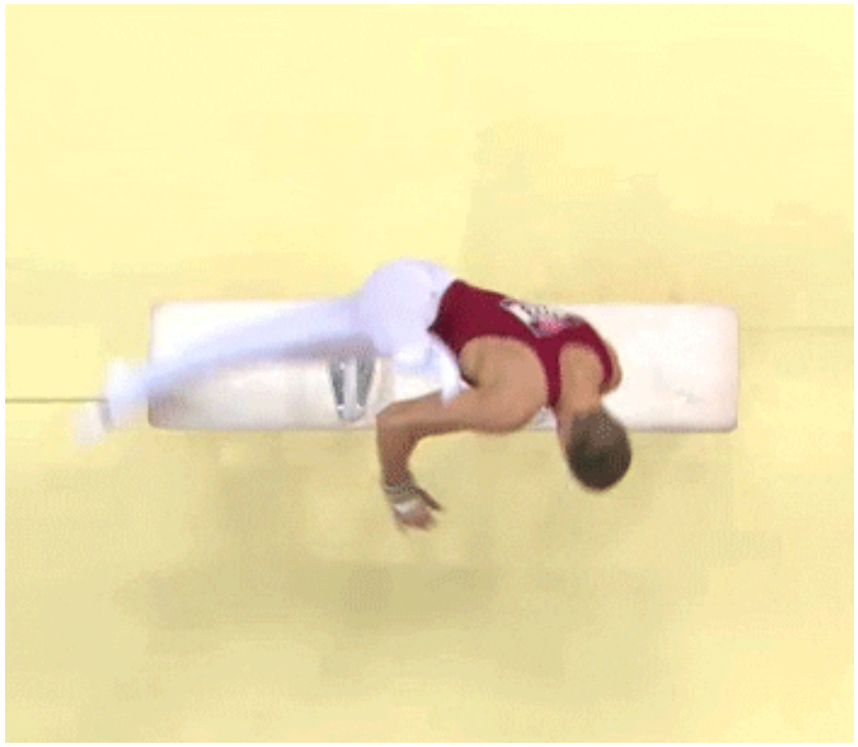
Aran Nayebi
Neurosciences PhD Candidate
Stanford University

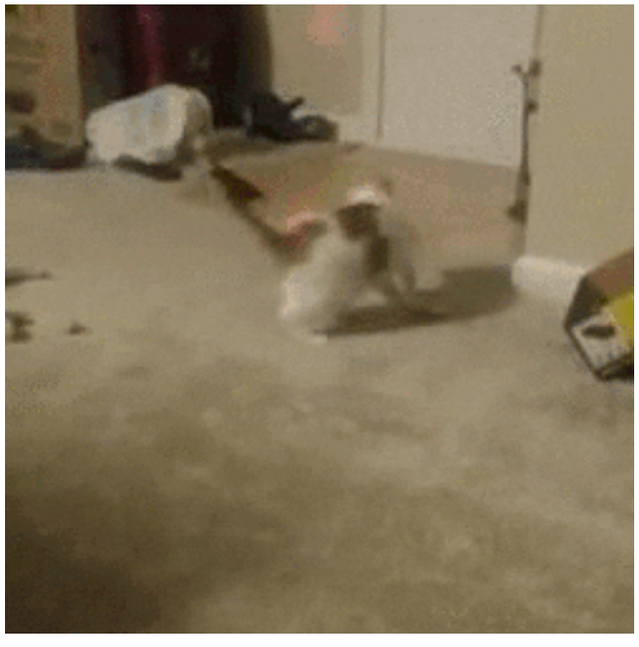
2022.03.15

Advisors:
Daniel L.K. Yamins
Surya Ganguli

Committee:
Lisa M. Giocomo (Defense Chair)
Stephen A. Baccus
Shaul Druckmann
David Sussillo

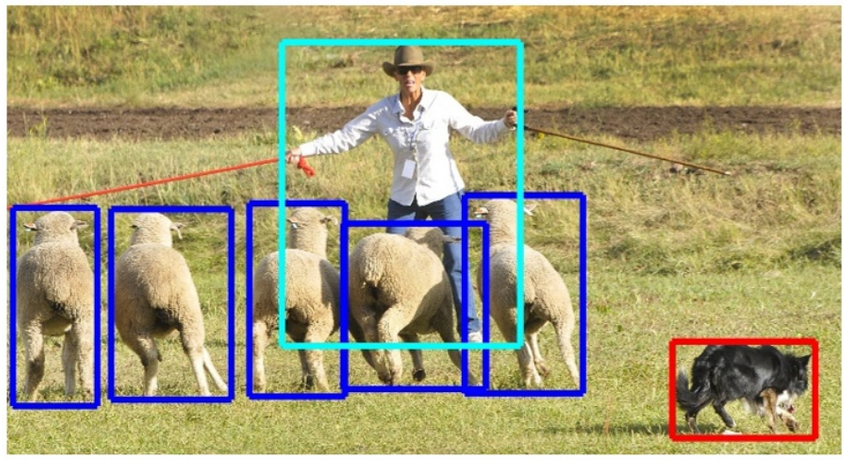
From Neurons to Behavior



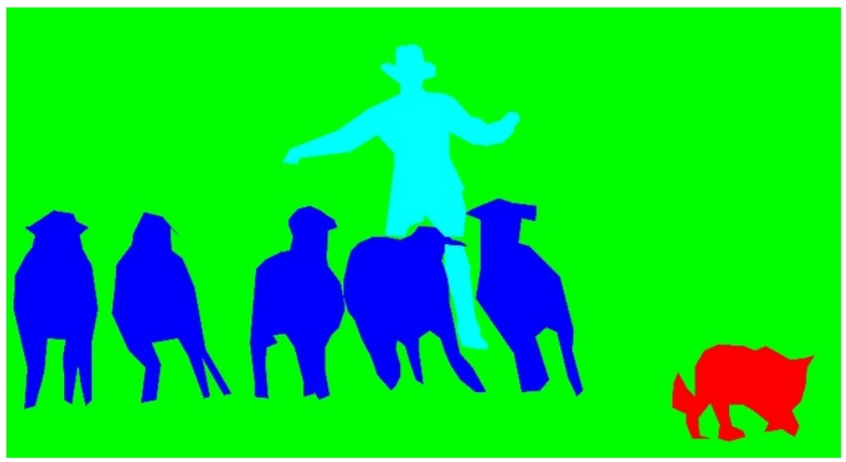


Classification, Segmentation, Localization, …

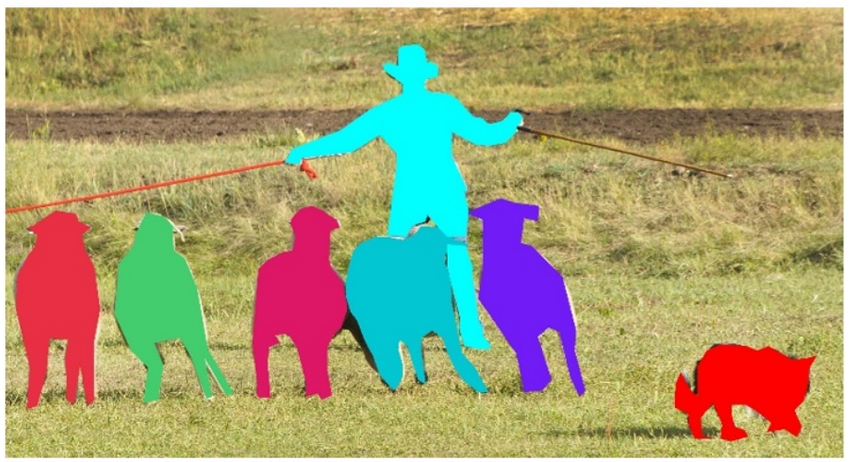
(a) **Image classification**



(b) **Object localization**



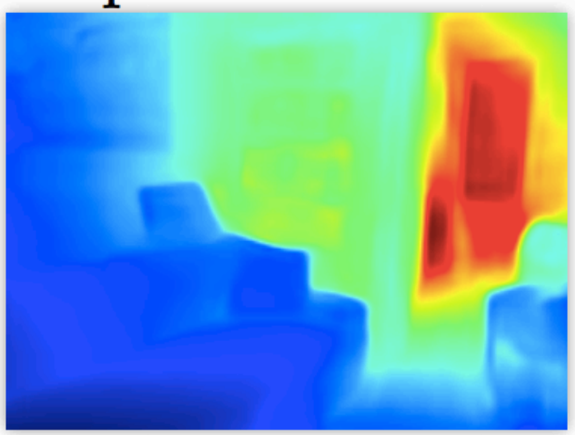
(c) **Semantic segmentation**



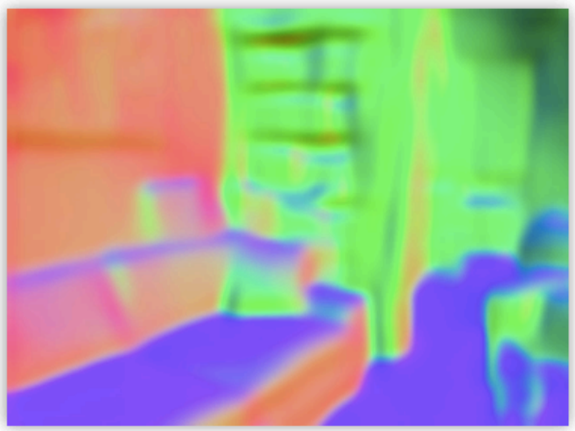
(d) **Instance segmentation**

Lin et al. 2014

Depth



Normals

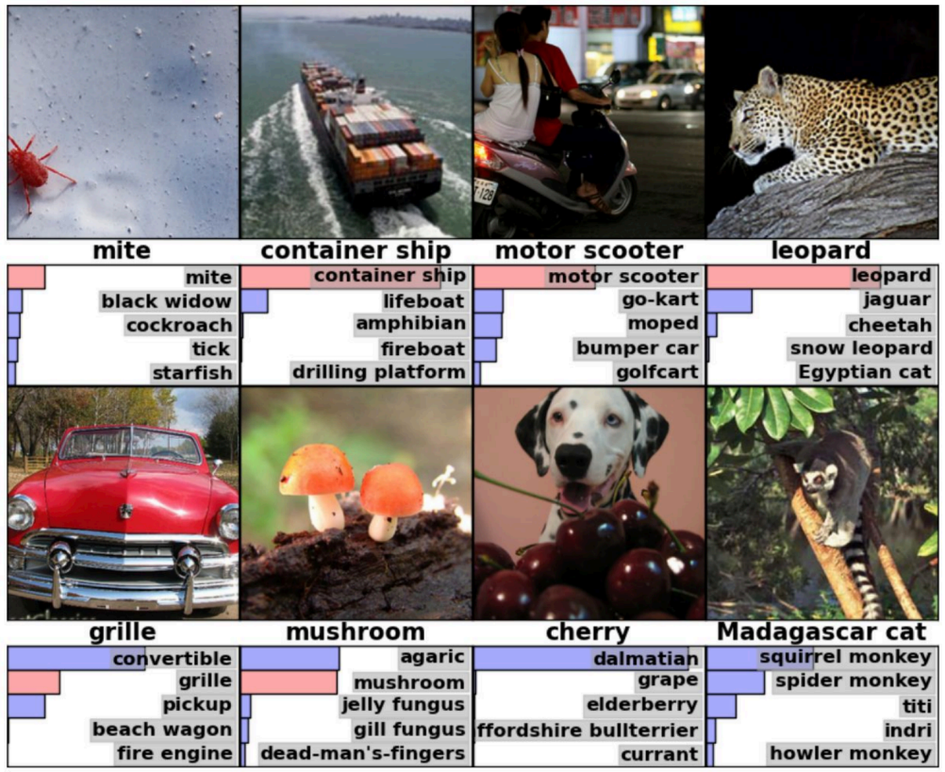


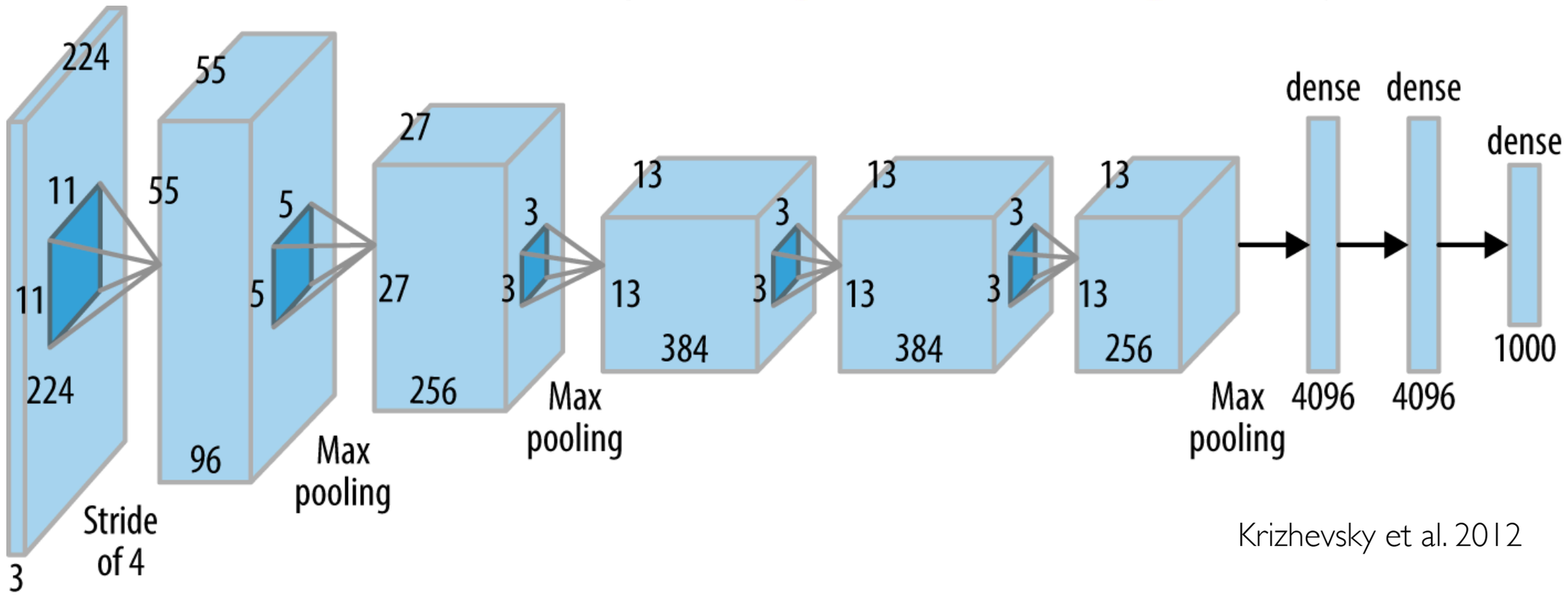
Eigen and Fergus 2015

Convolutional Neural Networks (CNNs)

ImageNet Challenge

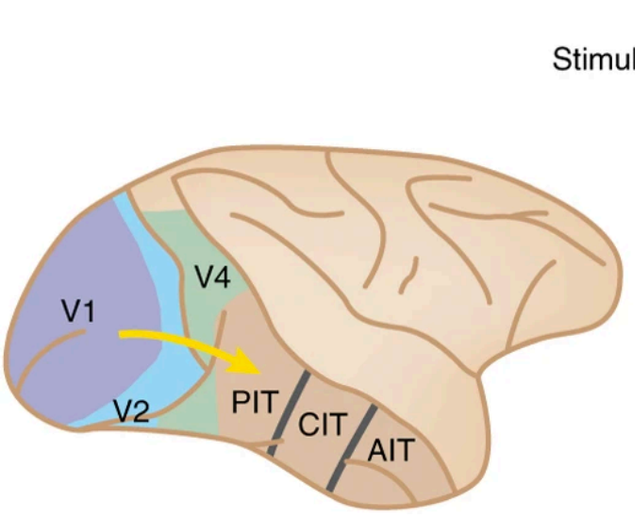
- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.

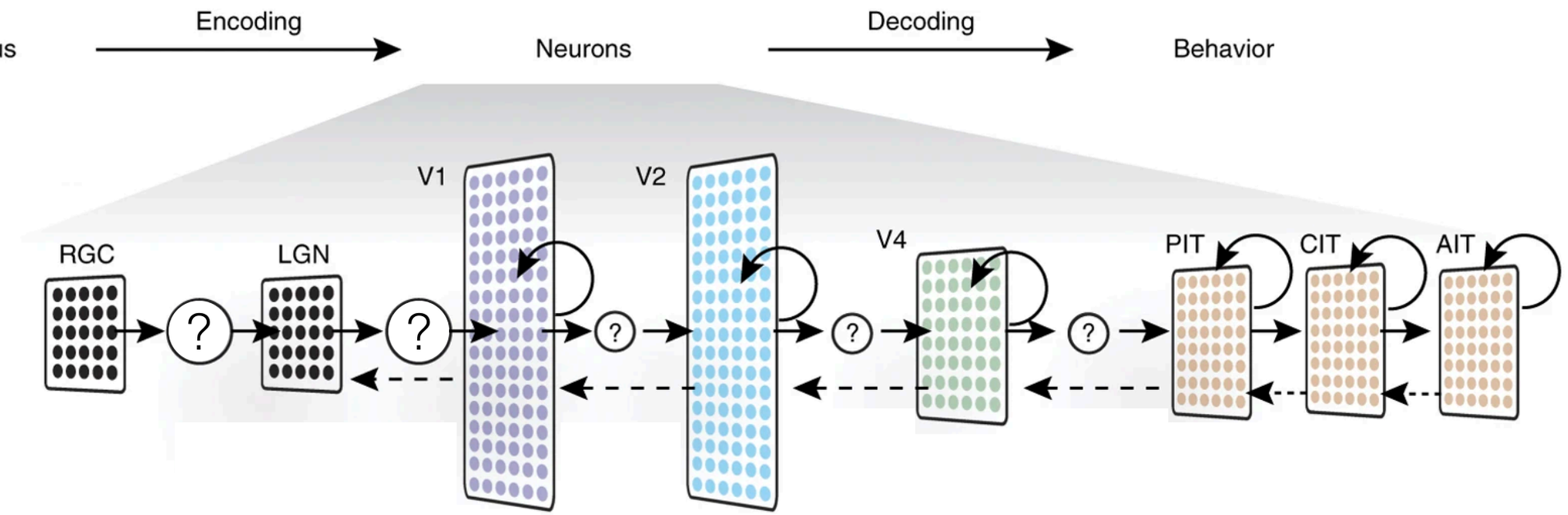




Krizhevsky et al. 2012

CNNs as Models of Object Recognition





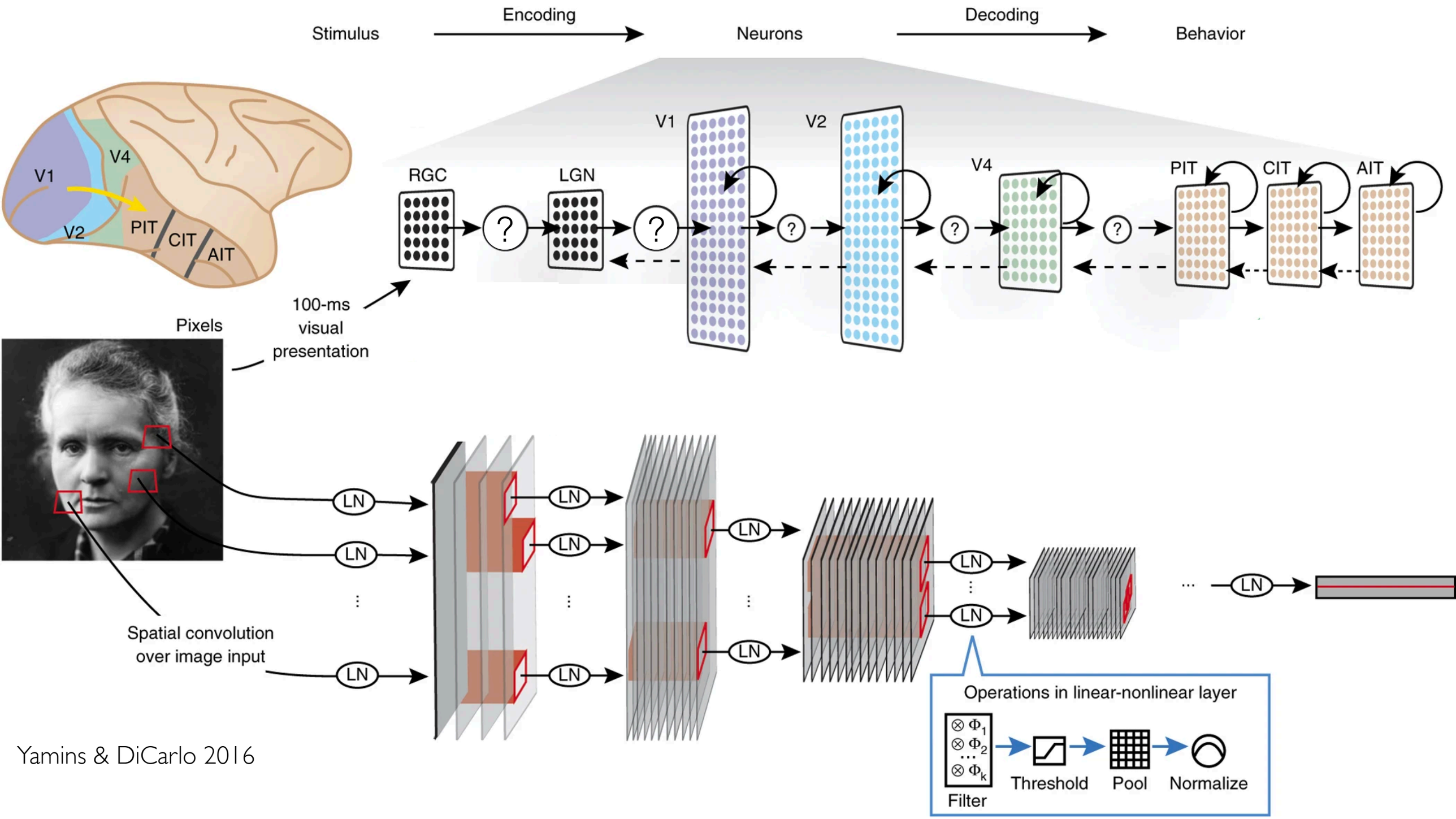
Convolutional Neural Networks (CNNs)

Fukushima, 1979; Lecun, 1995

CNNs are inspired by visual neuroscience:

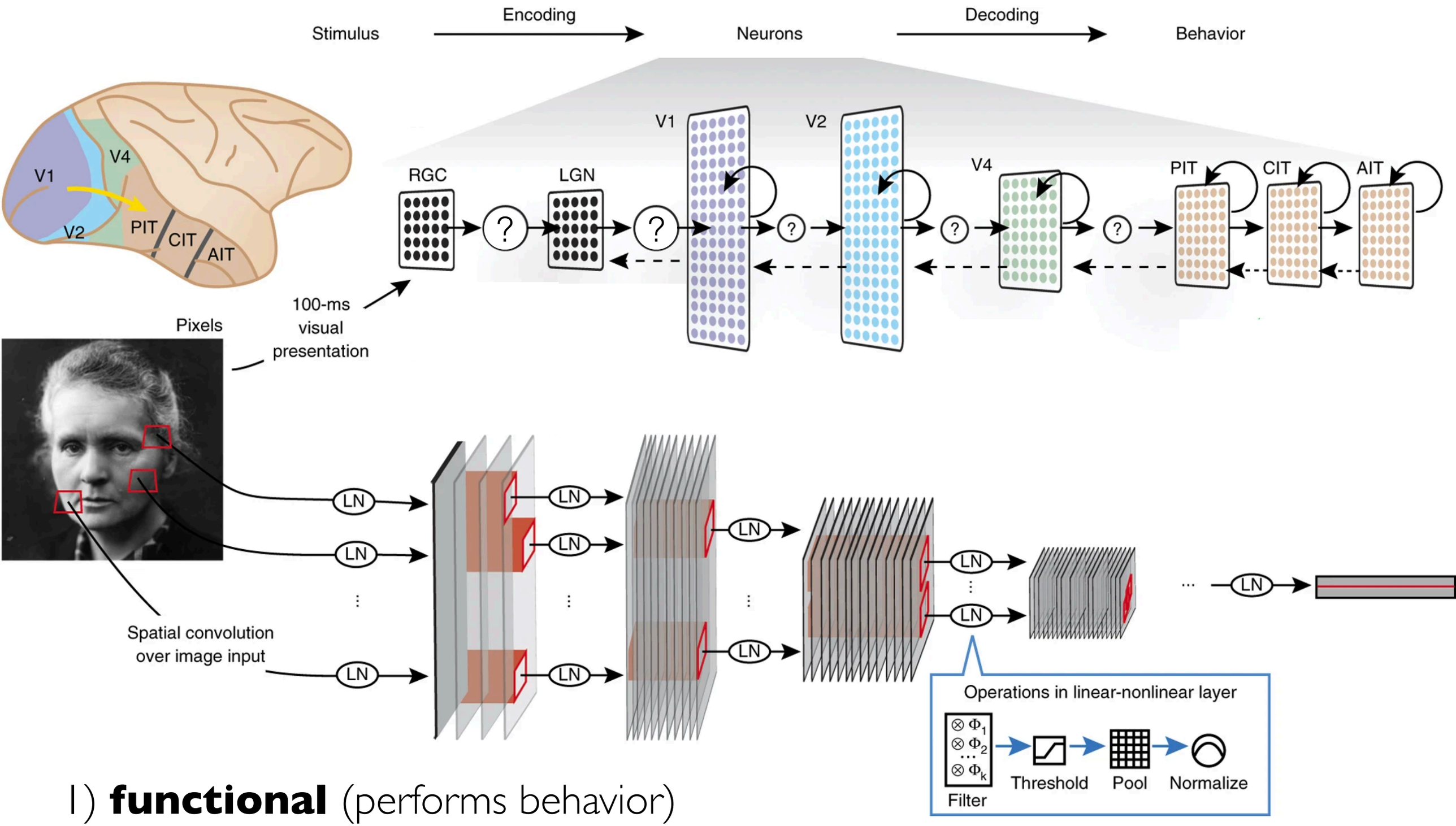
1) **hierarchy**
2) **retinotopy** (spatially tiled)

CNNs as Models of Object Recognition



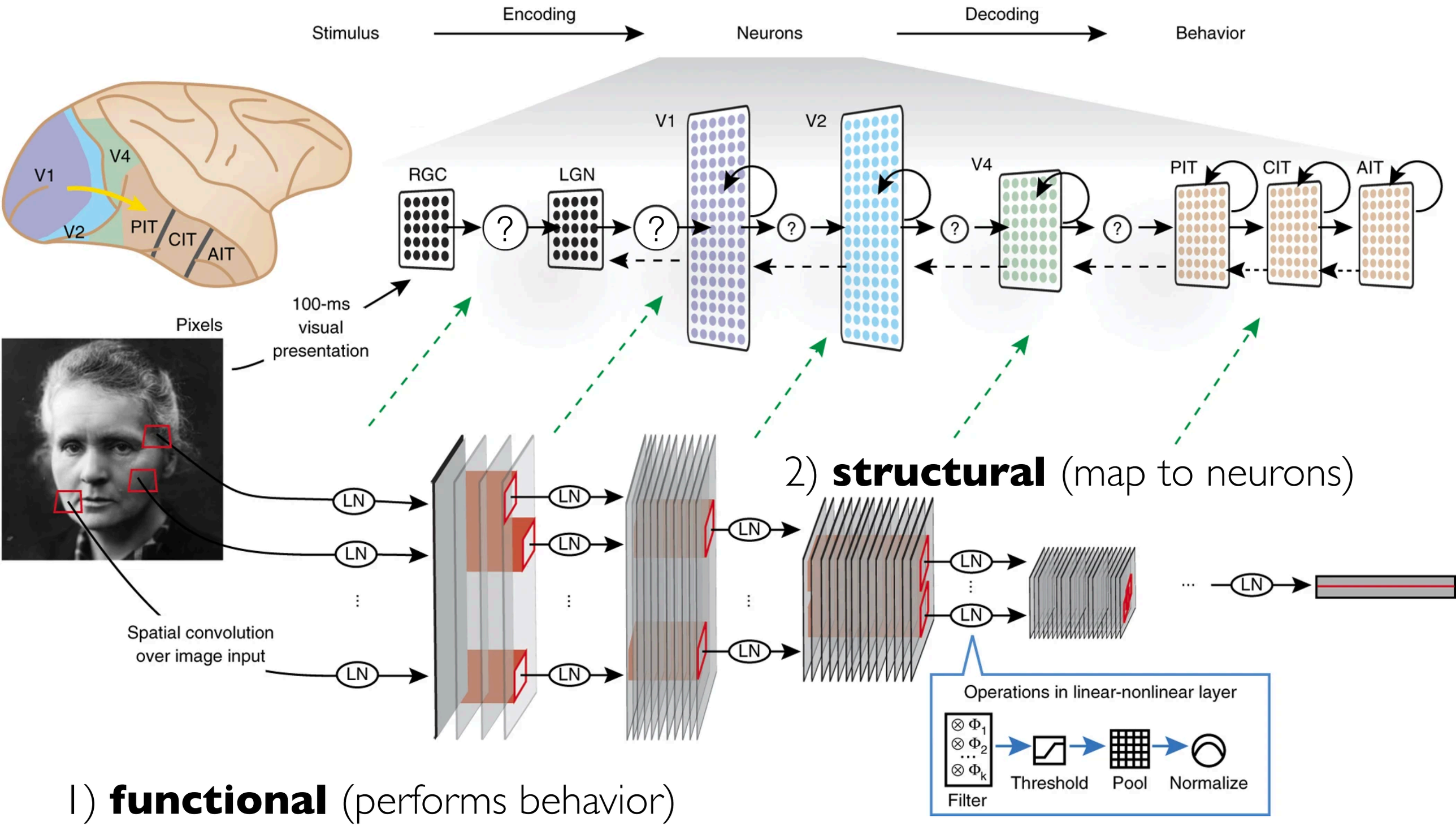
Yamins & DiCarlo 2016

CNNs as Models of Object Recognition



1) **functional** (performs behavior)

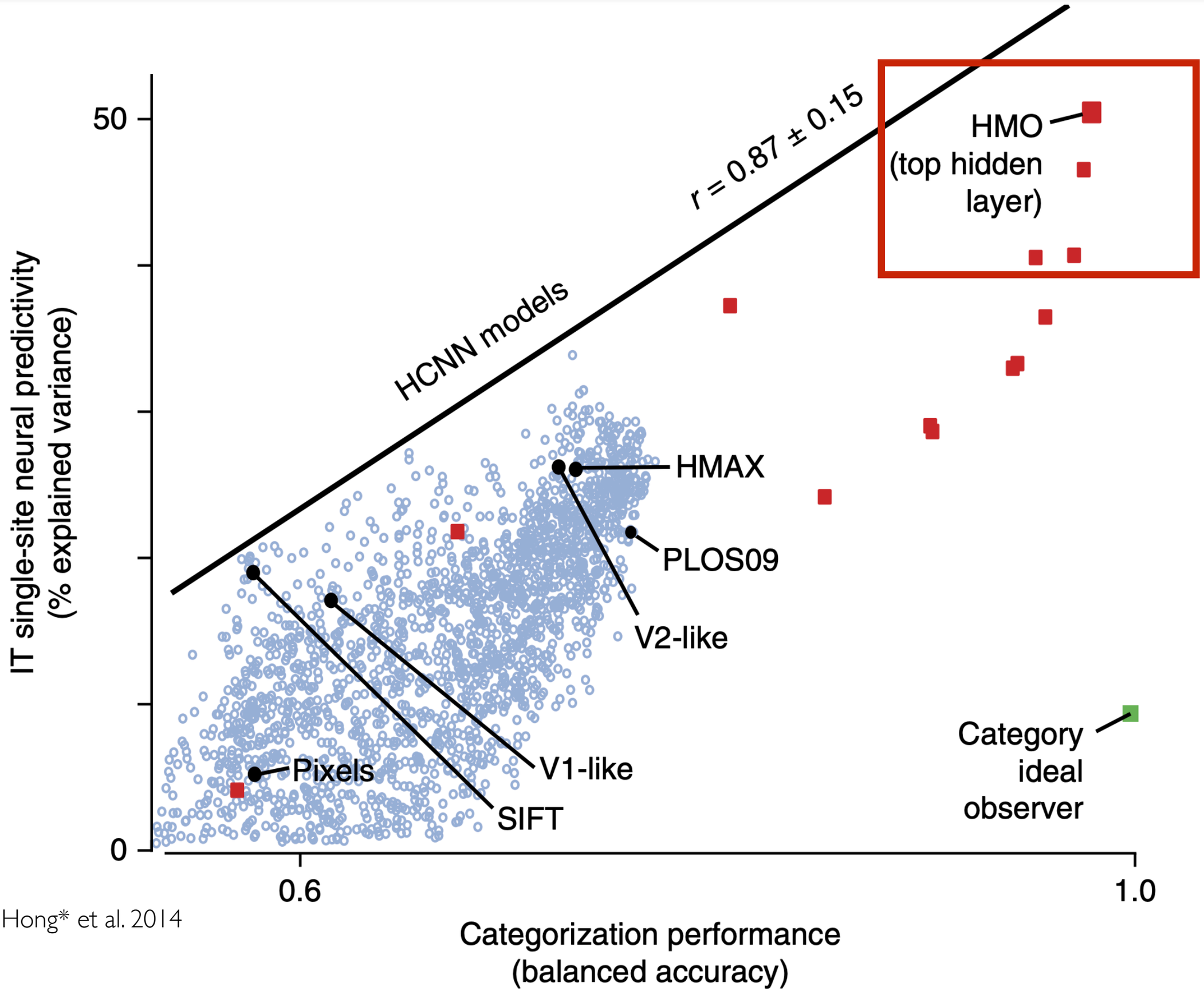
CNNs as Models of Object Recognition



2) **structural** (map to neurons)

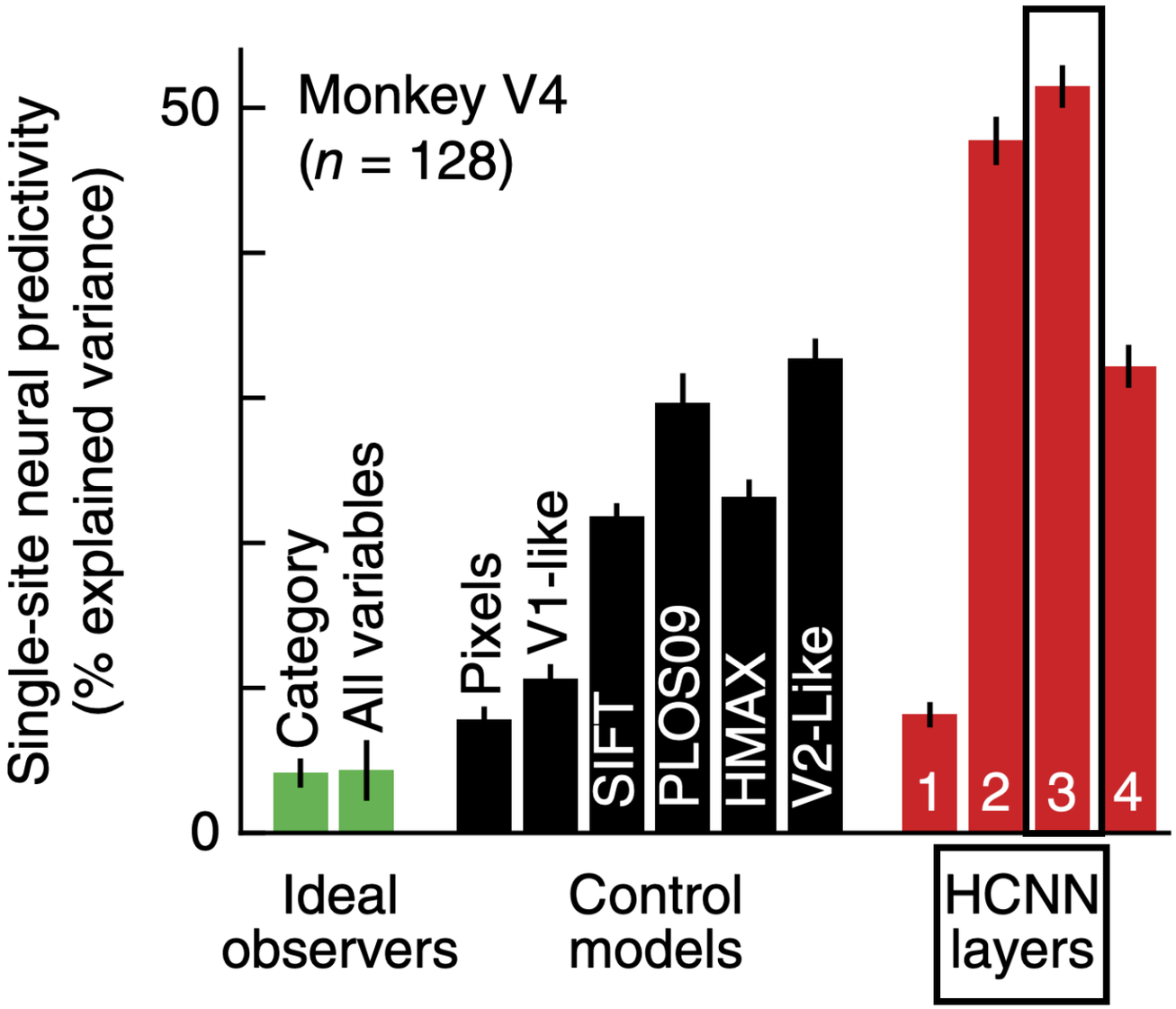
1) **functional** (performs behavior)

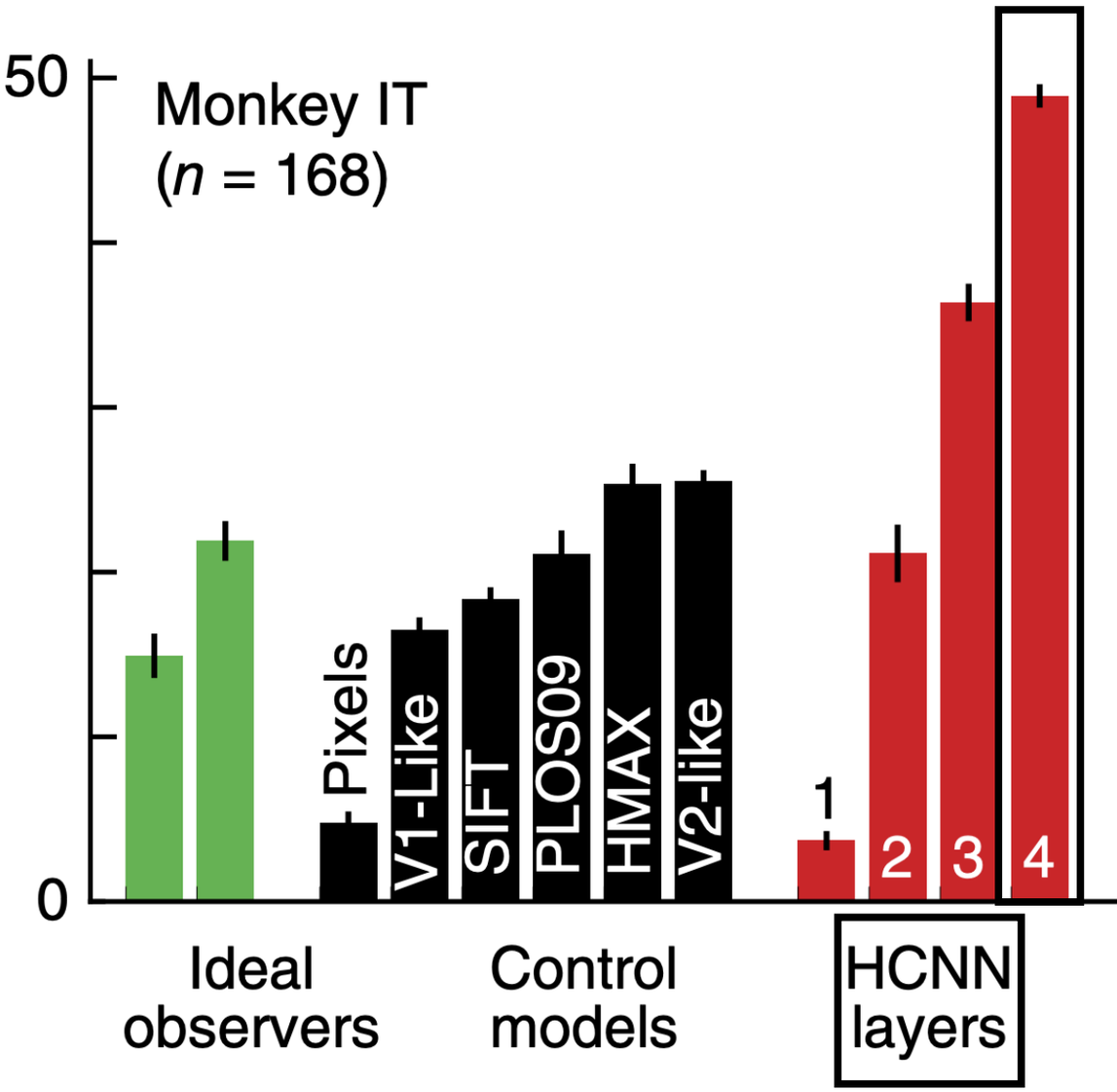
Categorization performance correlated with neural predictivity



Yamins*, Hong* et al. 2014

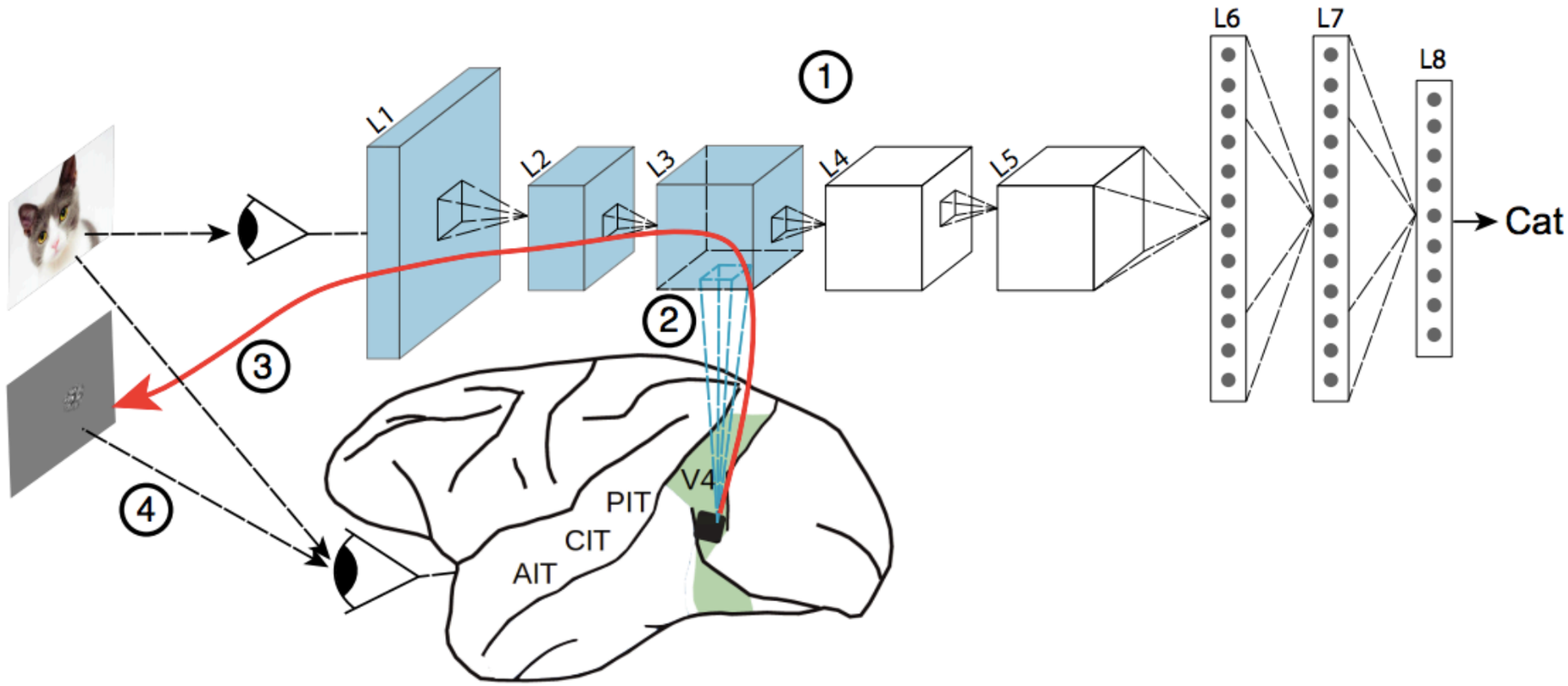
Hierarchy as a by-product of task optimization

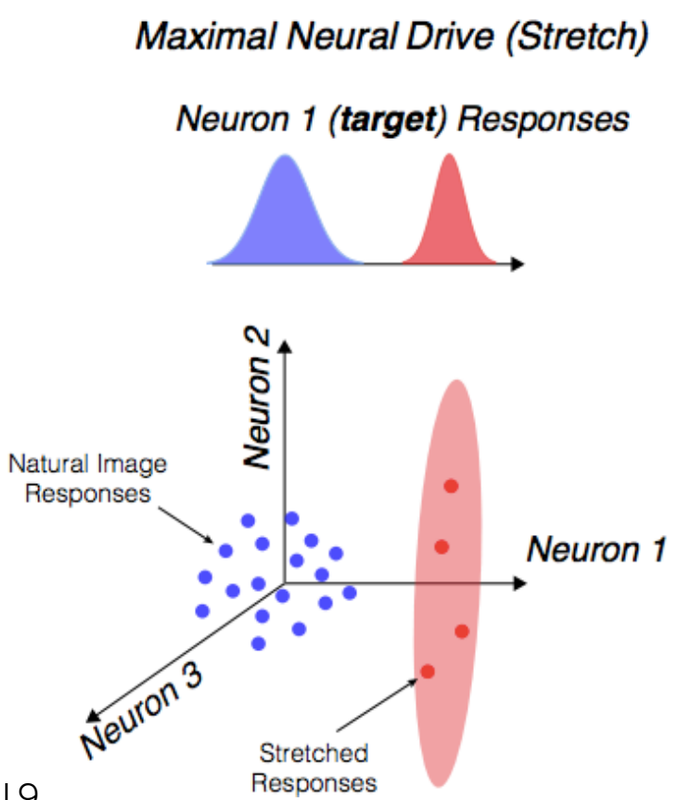


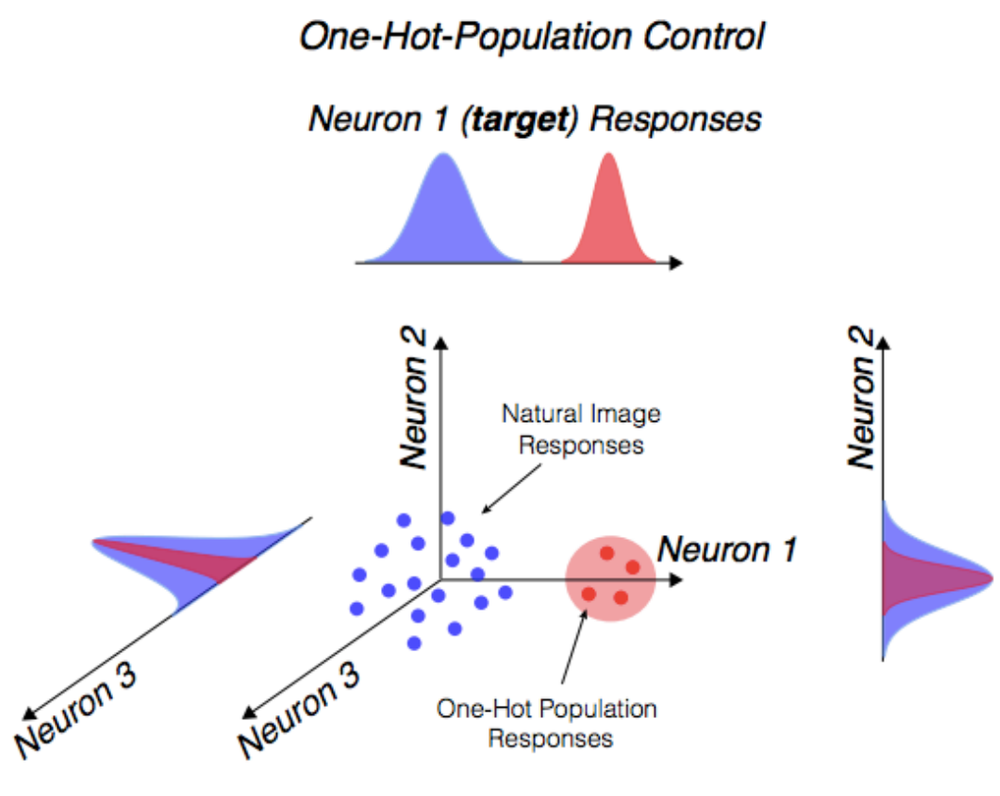


Yamins*, Hong* et al. 2014

Neural population control of intermediate areas

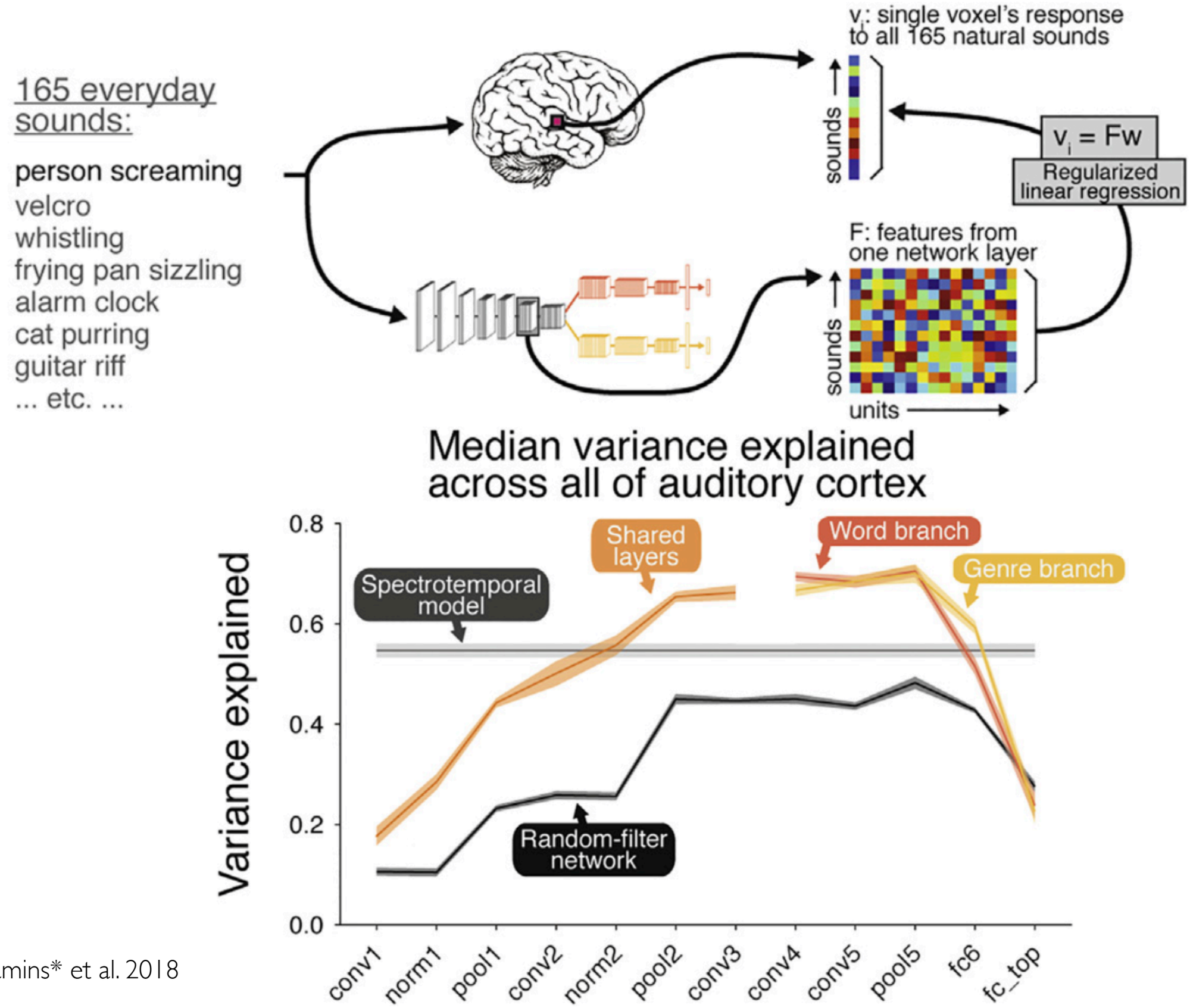






Bashivan et al. 2019

Not just limited to visual cortex, but also auditory cortex

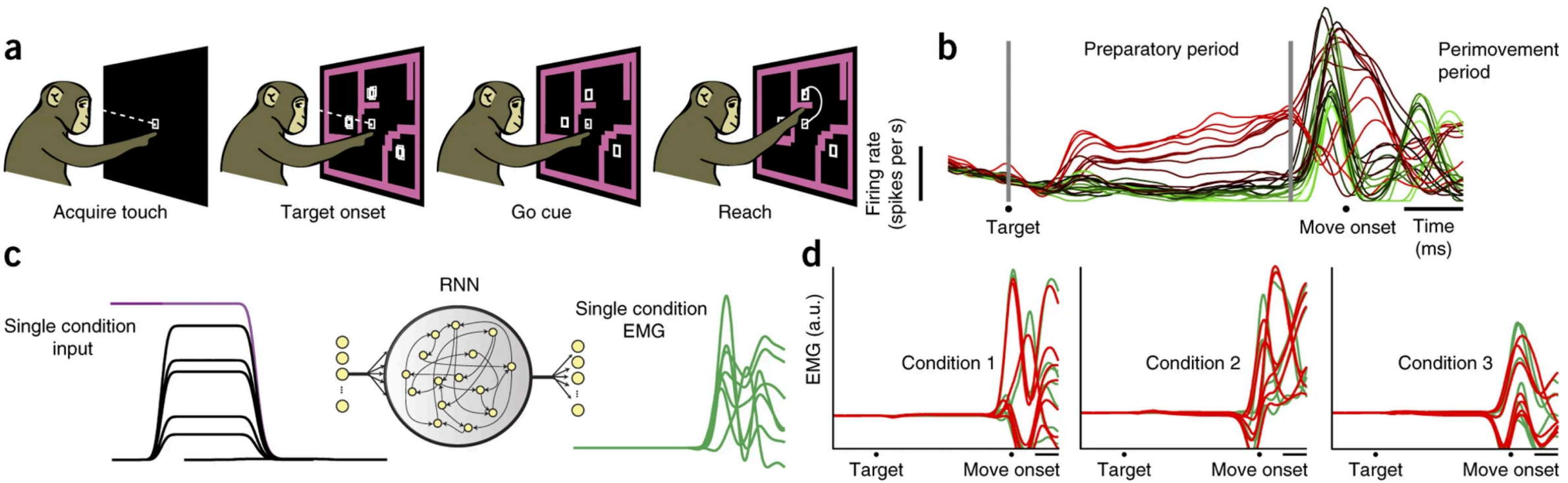


Kell*, Yamins* et al. 2018

Not just sensory, but applicable to motor areas

A neural network that finds a naturalistic solution for the production of muscle activity

David Sussillo ✉, Mark M Churchland, Matthew T Kaufman & Krishna V Shenoy



RNN trained to mimic muscle activities (EMG) as a function of condition

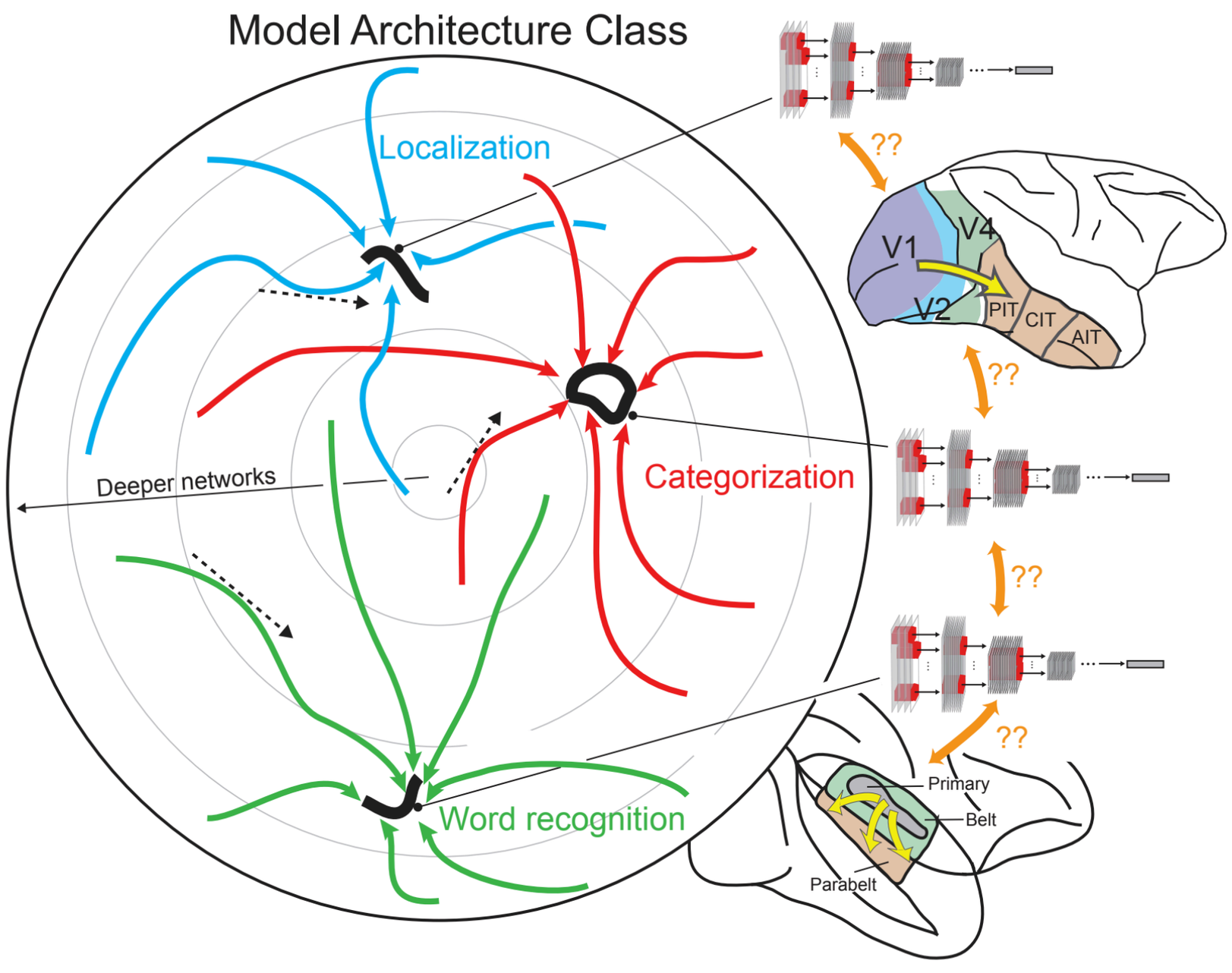
Goal-Driven Modeling (Sensory)

Sensory:

> Formulate comprehensive model class (**CNNs**)

> Choose challenging, ethologically-valid tasks (**categorization**)

> Implement generic learning rules (**gradient descent**)



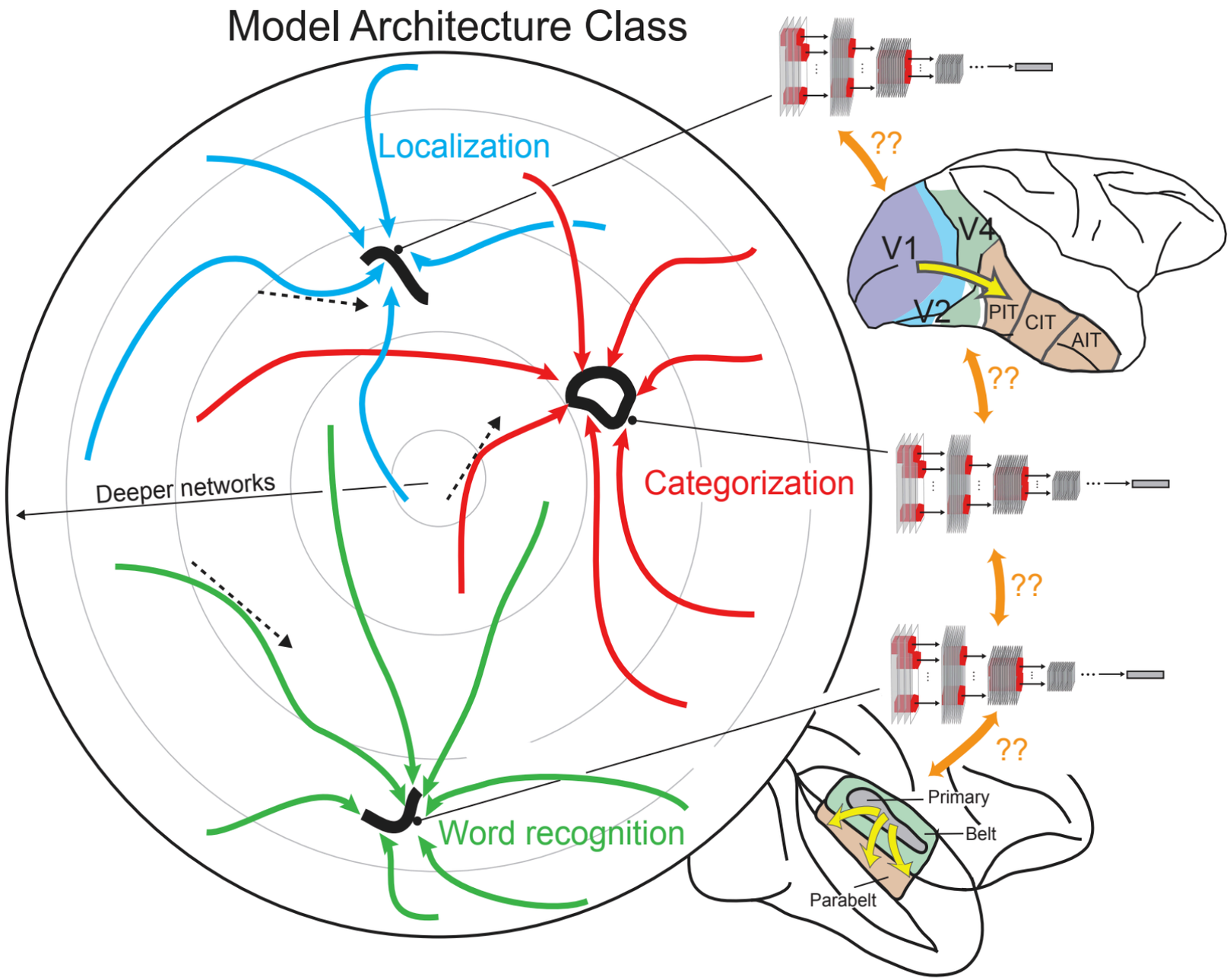
Goal-Driven Modeling (Motor)

Motor

> Formulate comprehensive model class (**RNNs**)

> Choose challenging, ethologically-valid tasks (**motion generation**)

> Implement generic learning rules (**gradient descent**)



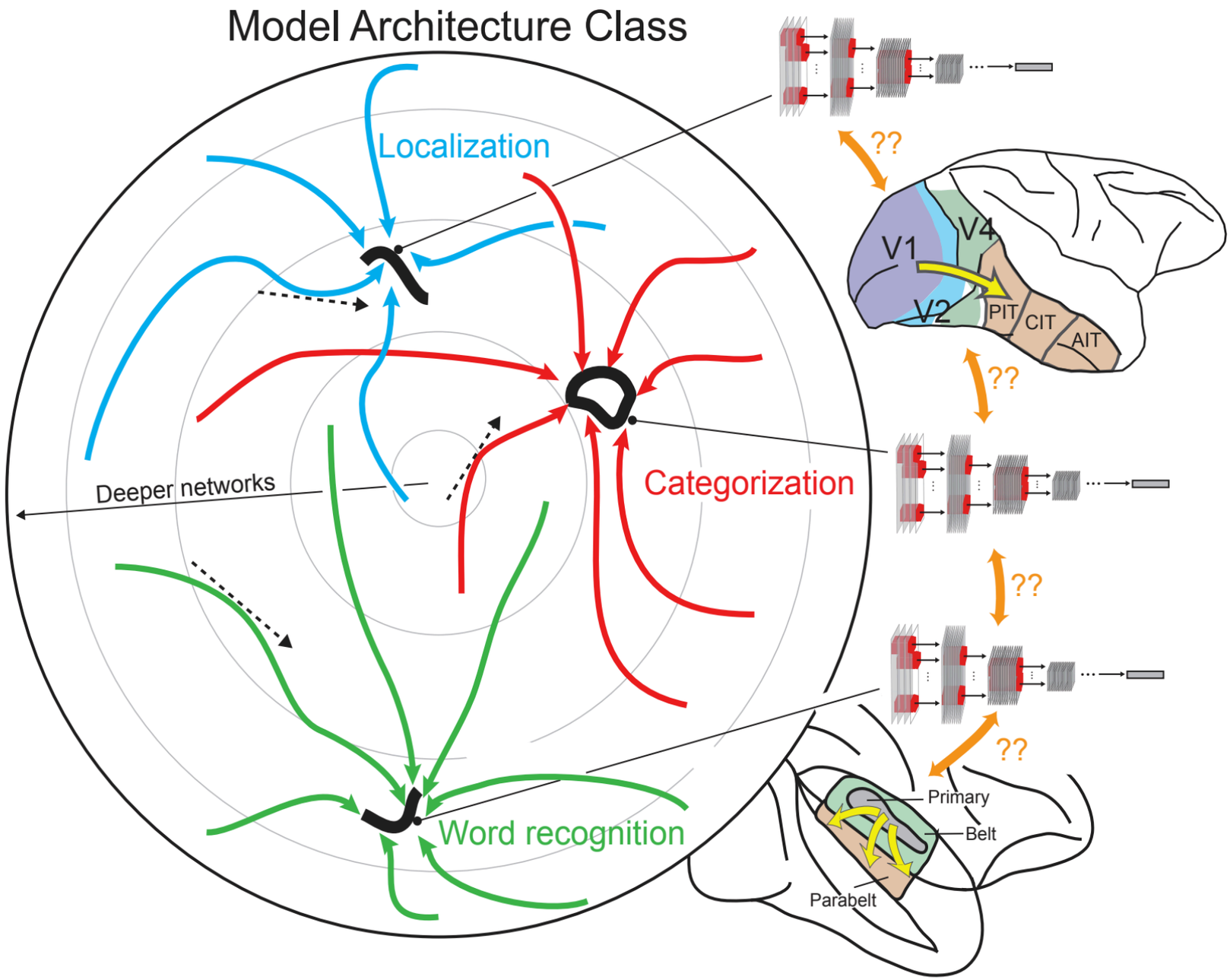
Goal-Driven Modeling

Motor

> Formulate comprehensive model class (**RNNs**)

> Choose challenging, ethologically-valid tasks (**motion generation**)

> Implement generic learning rules (**gradient descent**)



Similarity between sensory and motor:
Goal-driven optimization useful in both

Goal-Driven Modeling - Three Primary Components

1. *"circuit"*

A = *architecture class*

2. *"behavior"*

L = *loss function* ***D*** = *dataset*

3. *"plasticity mechanism"*

Learning Rule

$$\underset{a \in \mathcal{A}}{\operatorname{argmin}}[L(p_a^*)]$$

genetic algorithms

where p is result of*

backprop

$$\frac{dp_a}{dt} = -\lambda(t) \cdot \langle \nabla_{p_a} L(x) \rangle_{x \in \mathcal{D}}$$

Outline

- Recurrent Connections in the Primate Ventral Stream
- Goal-Driven Models of Mouse Visual Cortex
- Heterogeneity in Rodent Medial Entorhinal Cortex
- Building and Identifying Learning Rules

Outline

- Recurrent Connections in the Primate Ventral Stream
- Goal-Driven Models of Mouse Visual Cortex
- Heterogeneity in Rodent Medial Entorhinal Cortex
- Building and Identifying Learning Rules

Goal-Driven Modeling - Three Primary Components

"circuit"

1.

A = *architecture class*

A. Nayebi*, D. Bear*, J. Kubilius*, *et al.*
Task-Driven Convolutional Recurrent Models of the Visual System.
NeurIPS 2018

"behavior"

2.

L = *loss function* ***D*** = *dataset*

A. Nayebi, *et al.*
Recurrent Connections in the Primate Ventral Visual Stream Mediate a Tradeoff Between Task Performance and Network Size During Core Object Recognition. *Neural Computation 2022*

"plasticity mechanism"

3.

Learning Rule

$$\underset{a \in \mathcal{A}}{\operatorname{argmin}}[L(p_a^*)]$$

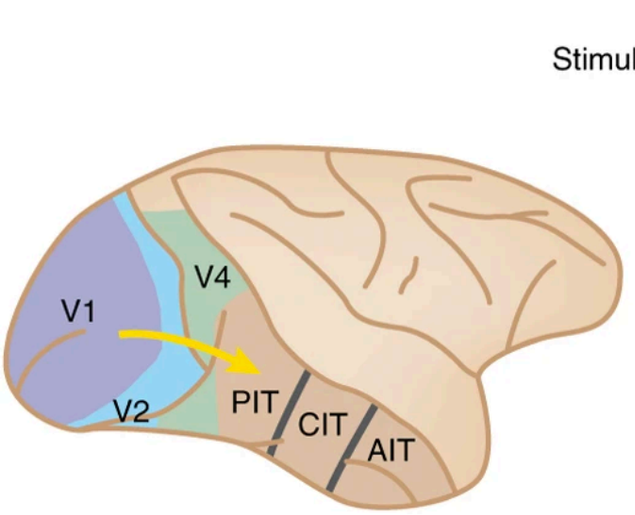
genetic algorithms

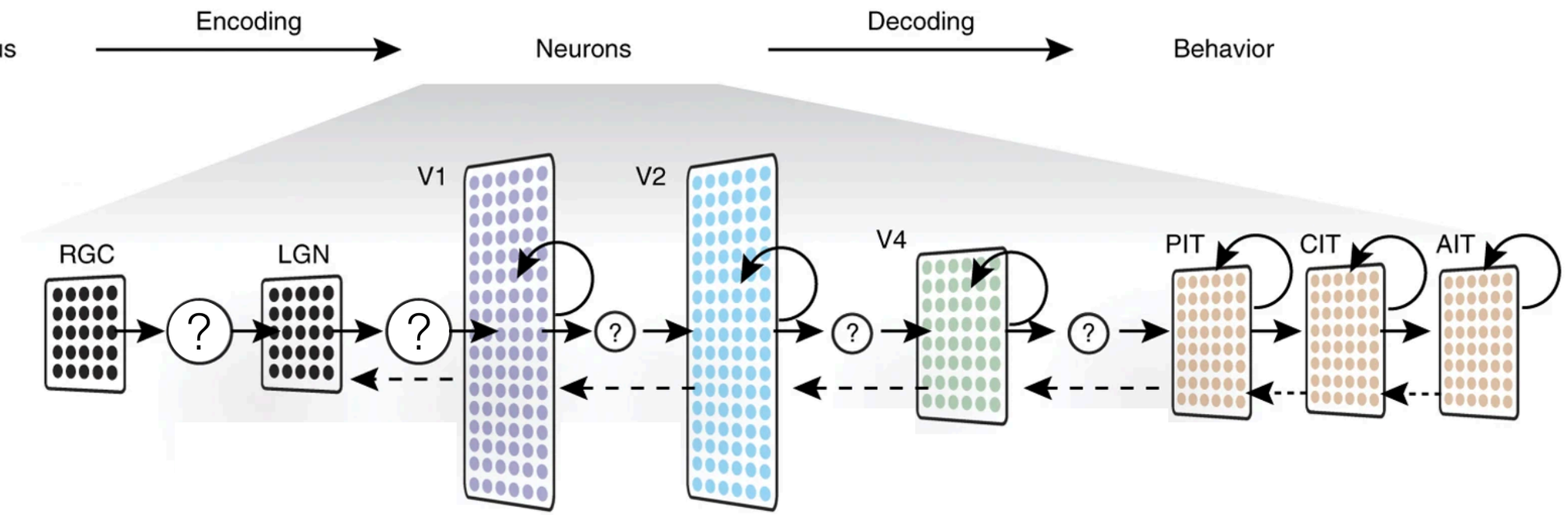
where p is result of*

backprop

$$\frac{dp_a}{dt} = -\lambda(t) \cdot \langle \nabla_{p_a} L(x) \rangle_{x \in \mathcal{D}}$$

CNNs as Models of Object Recognition





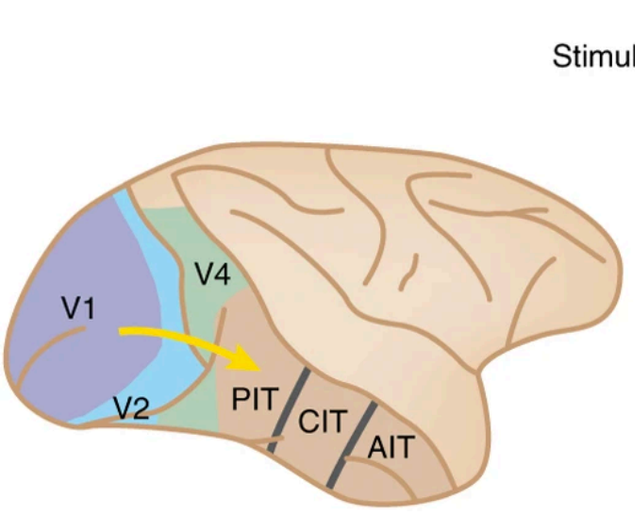
Convolutional Neural Networks (CNNs)

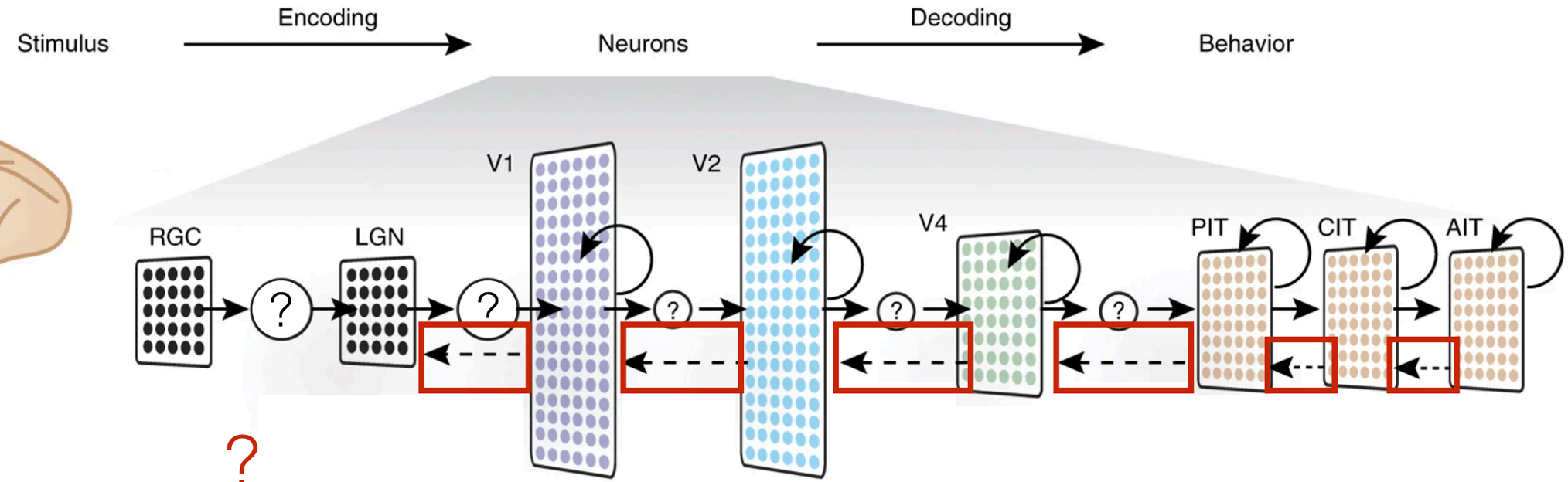
Fukushima, 1979; Lecun, 1995

CNNs are inspired by visual neuroscience:

1) **hierarchy**
2) **retinotopy** (spatially tiled)

…but lack feedback connections





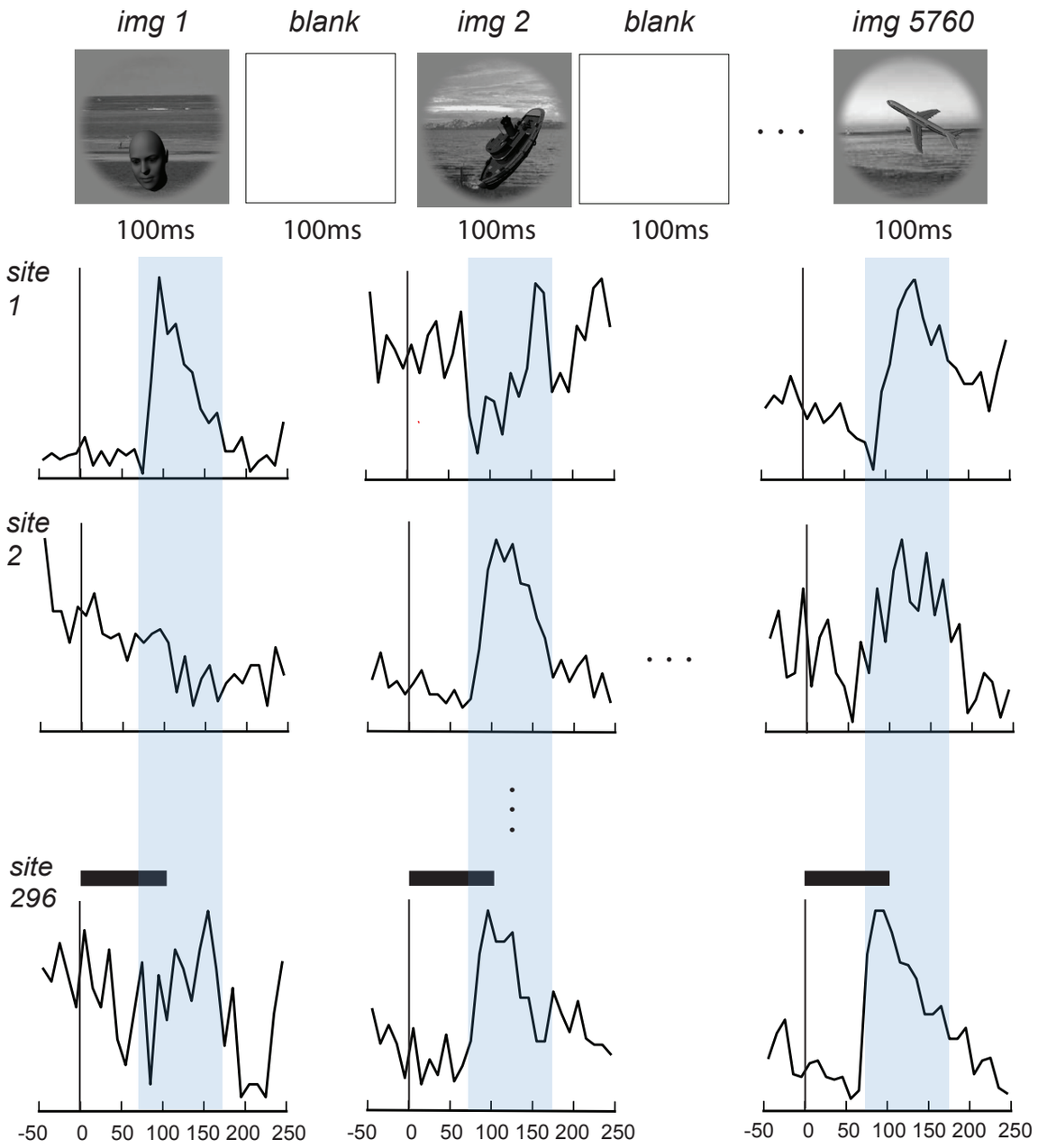
Convolutional Neural Networks (CNNs)

Fukushima, 1979; Lecun, 1995

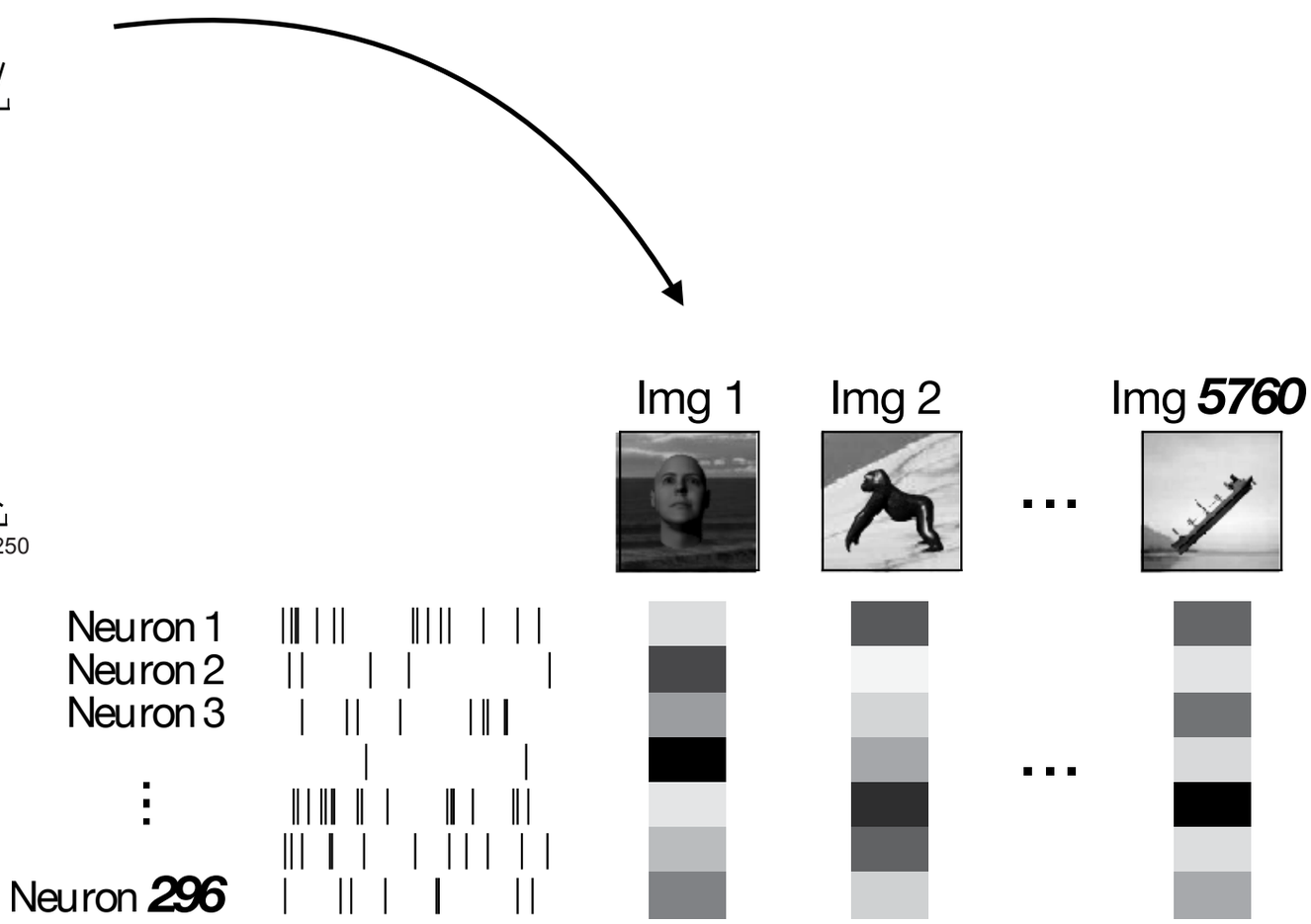
CNNs are inspired by visual neuroscience:

1) **hierarchy**
2) **retinotopy** (spatially tiled)

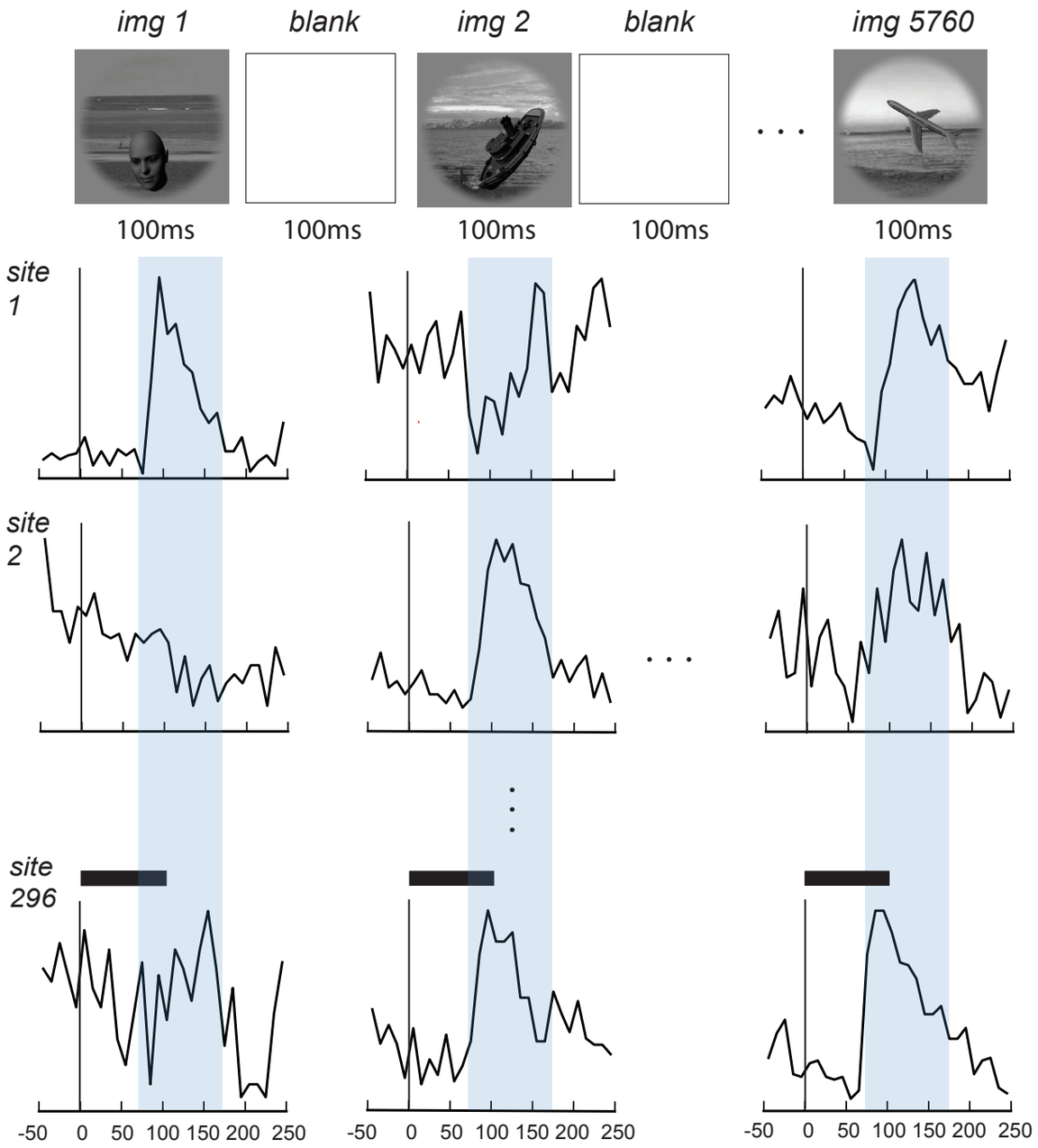
So far, only explaining temporal average of responses



e.g. Binned spike counts 70ms-170ms post stimulus presentation

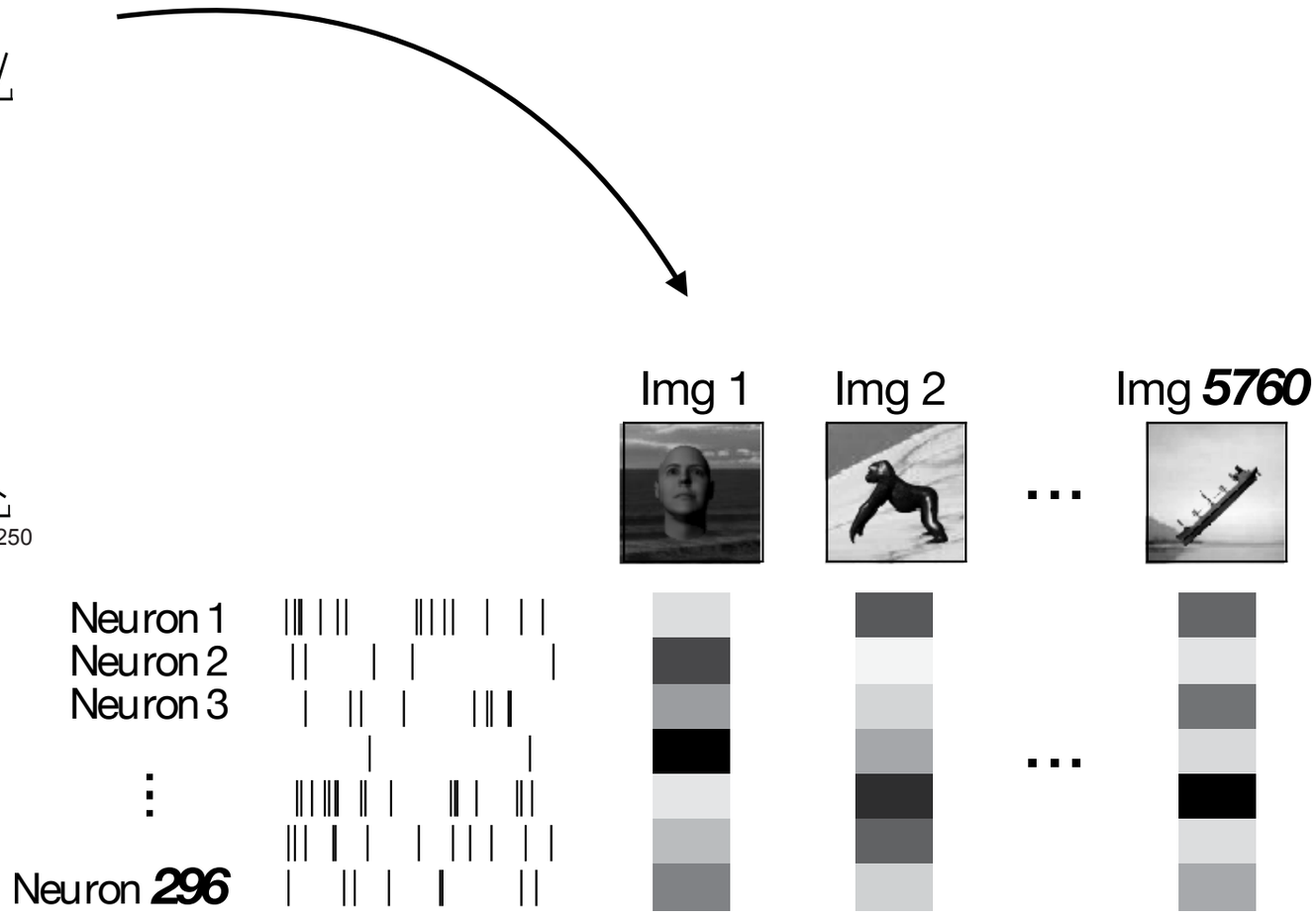


So far, only explaining temporal average of responses



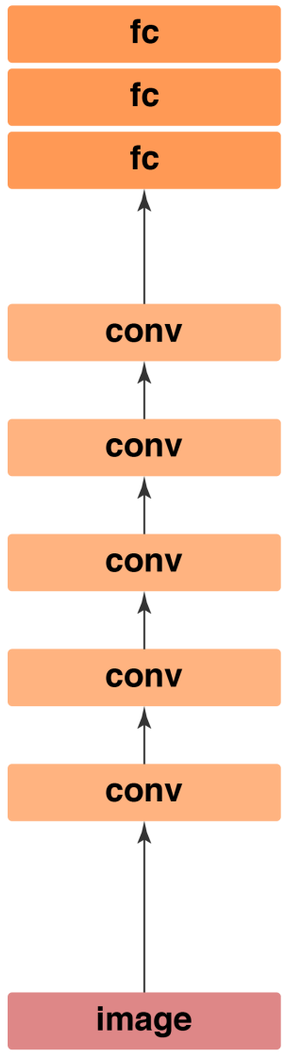
e.g. Binned spike counts 70ms-170ms post stimulus presentation

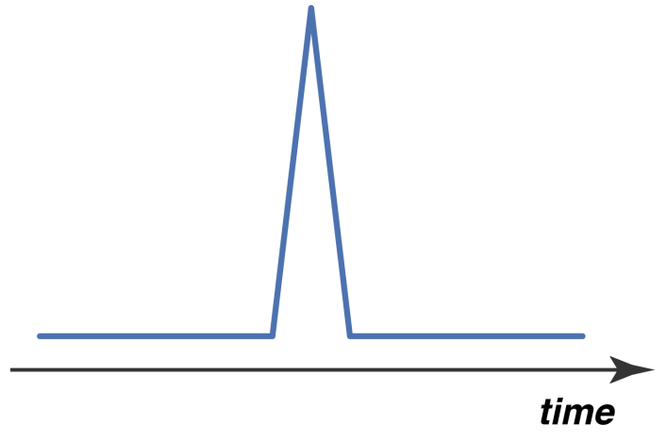
but actually the data is highly reliable at much finer grain —10ms bins



Trajectory Possibilities

Simple feedforward networks simple dynamics:

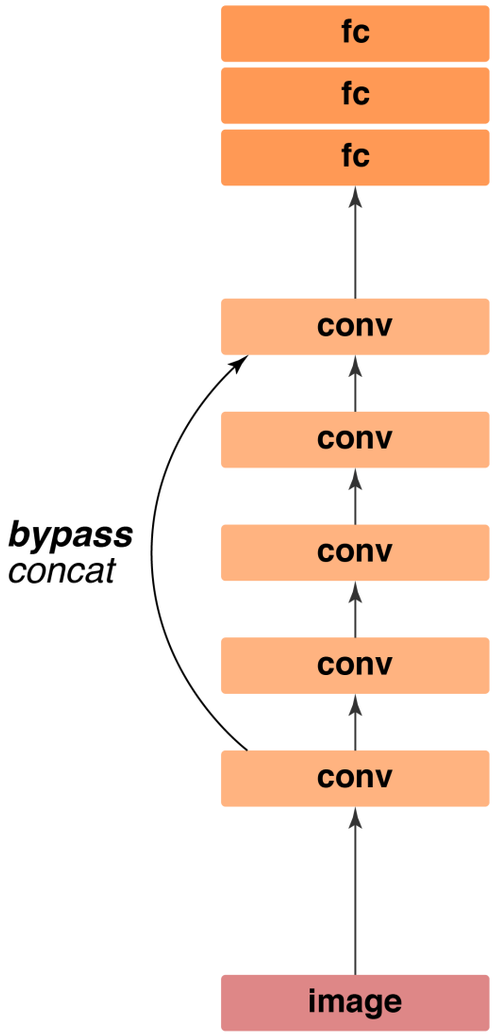


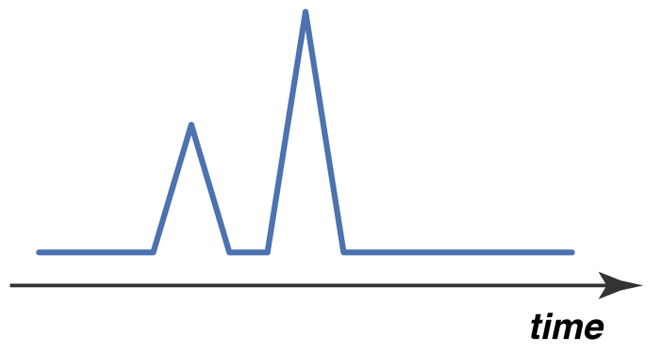


courtesy Jonas Kubilius

Trajectory Possibilities

Dynamics more interesting with bypasses:

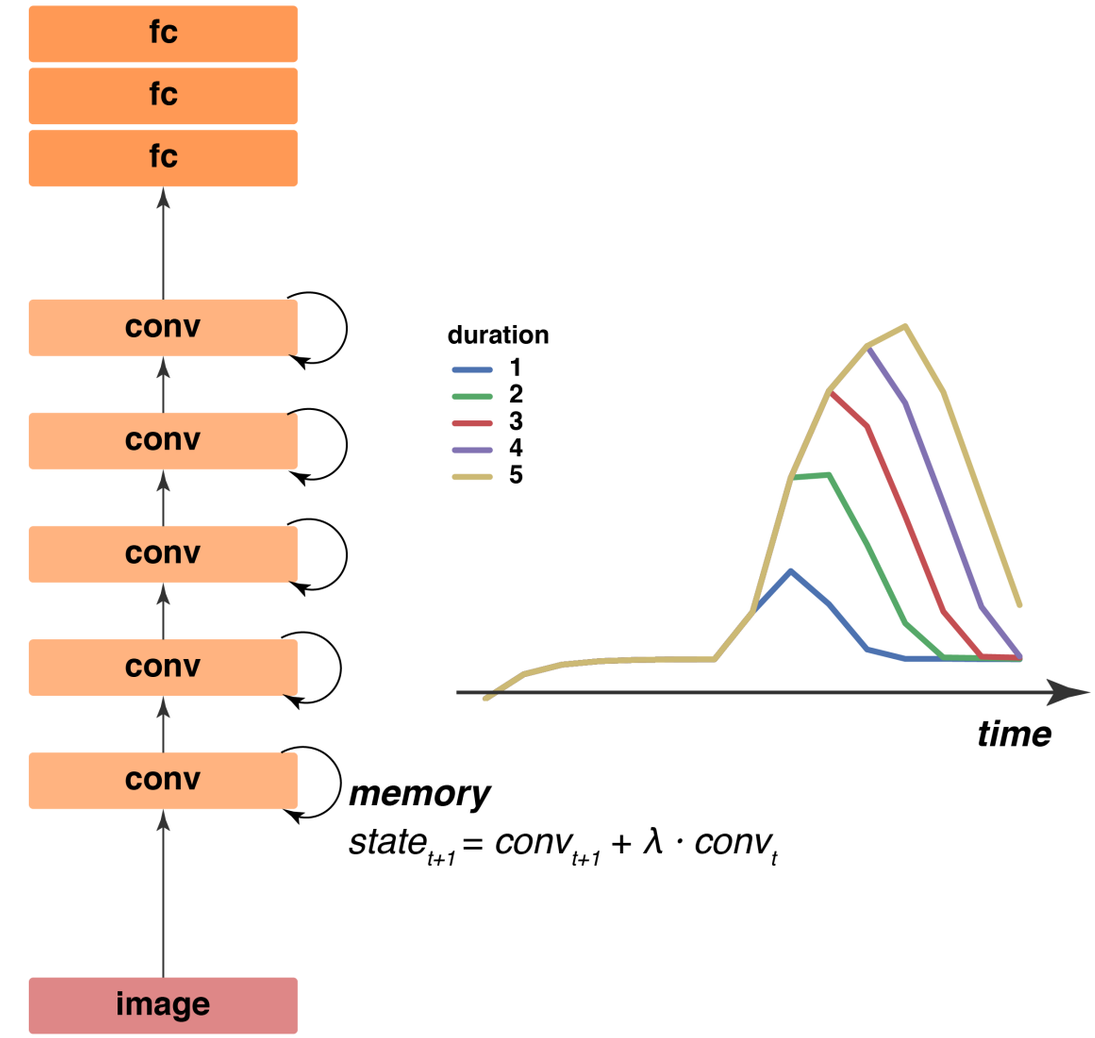




courtesy Jonas Kubilius

Trajectory Possibilities

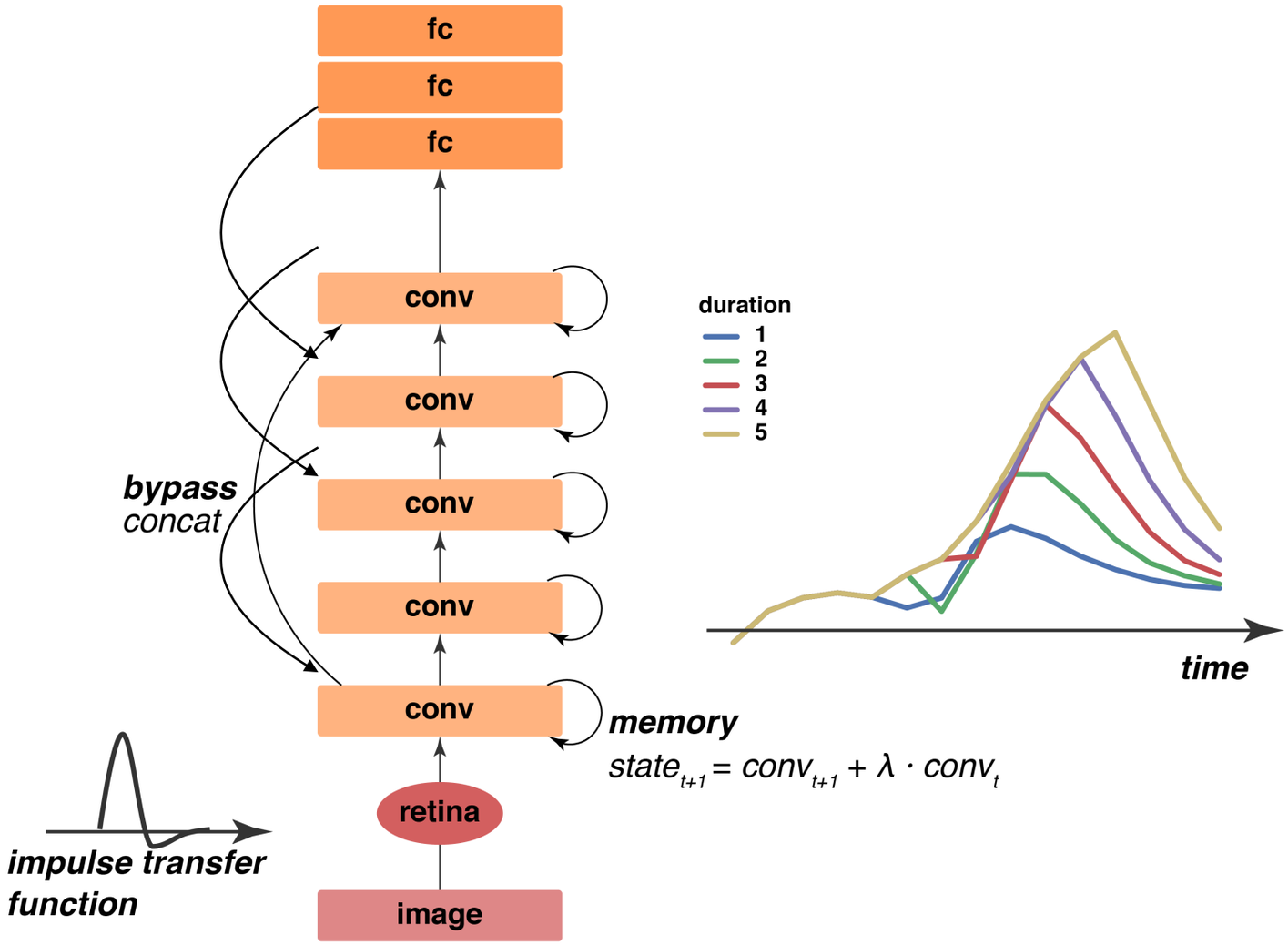
Dynamics more interesting with bypasses, local recurrence:



courtesy Jonas Kubilius

Trajectory Possibilities

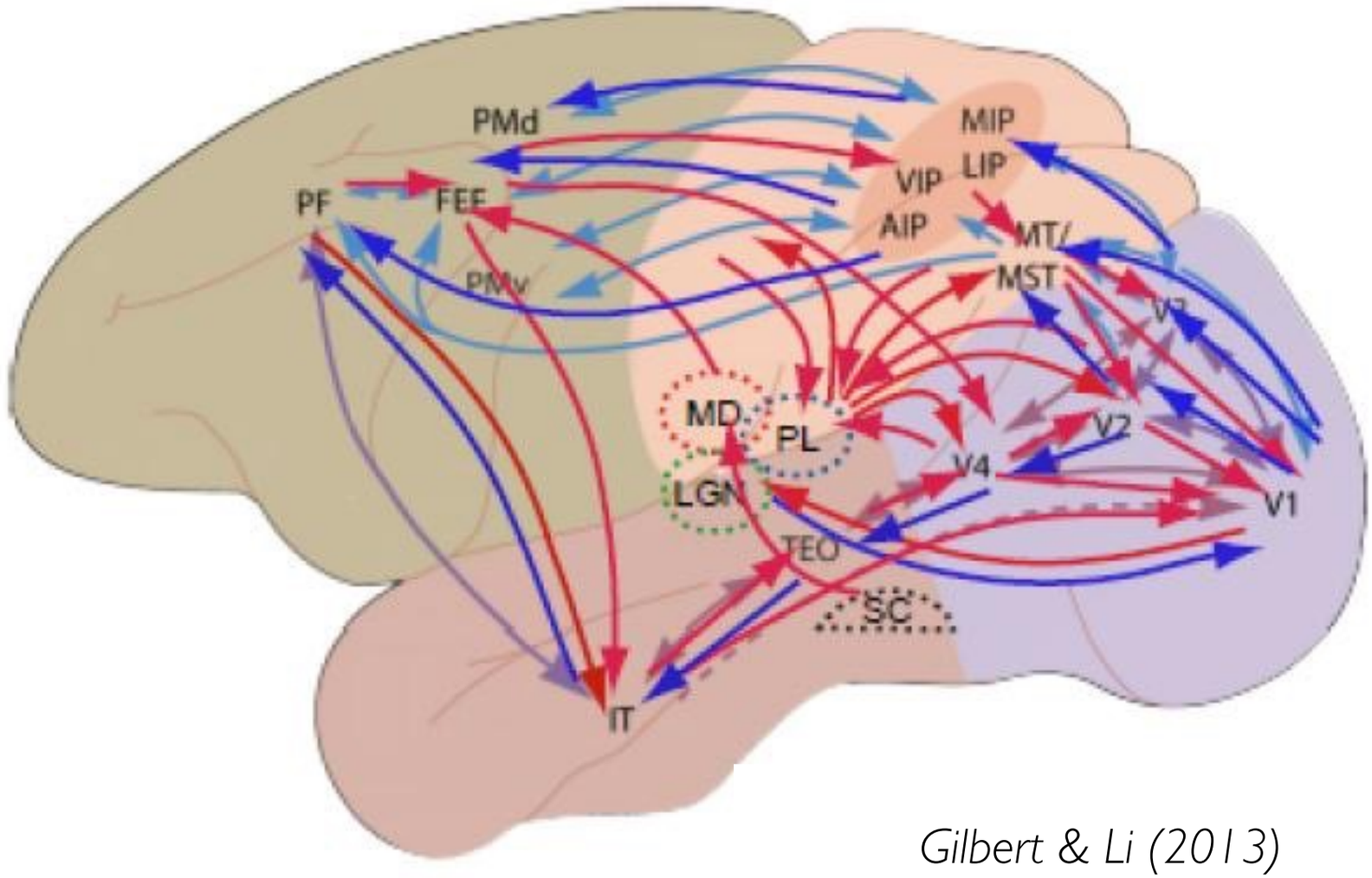
Dynamics more interesting with bypasses, local recurrence, long-range feedback:



courtesy Jonas Kubilius

Feedback connections are ubiquitous

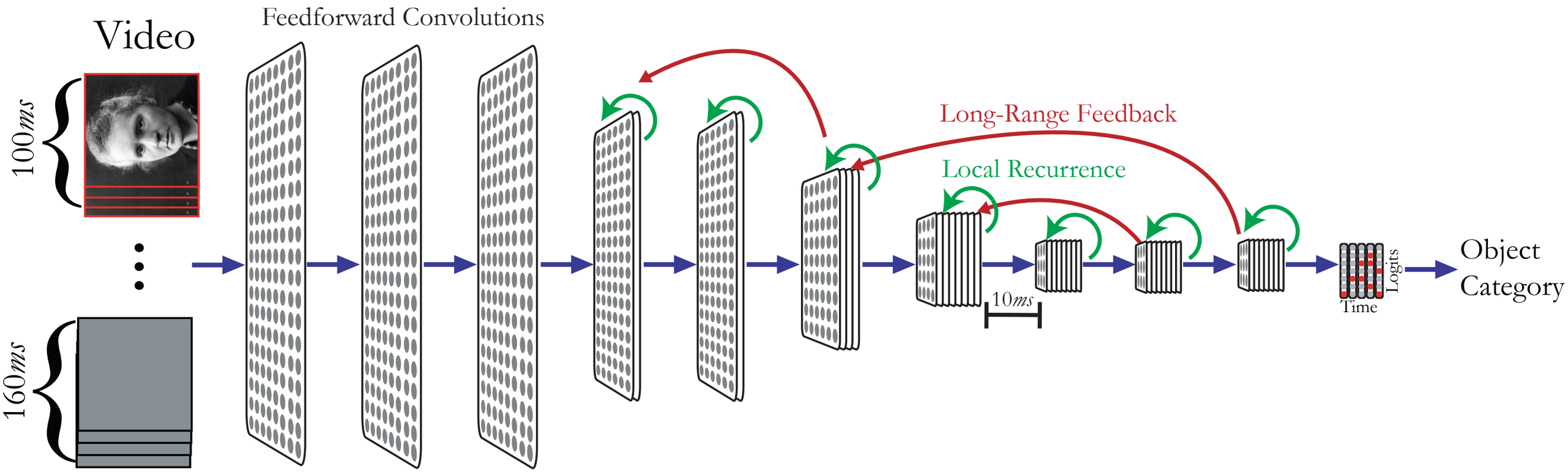
Feedbacks are everywhere anatomically:



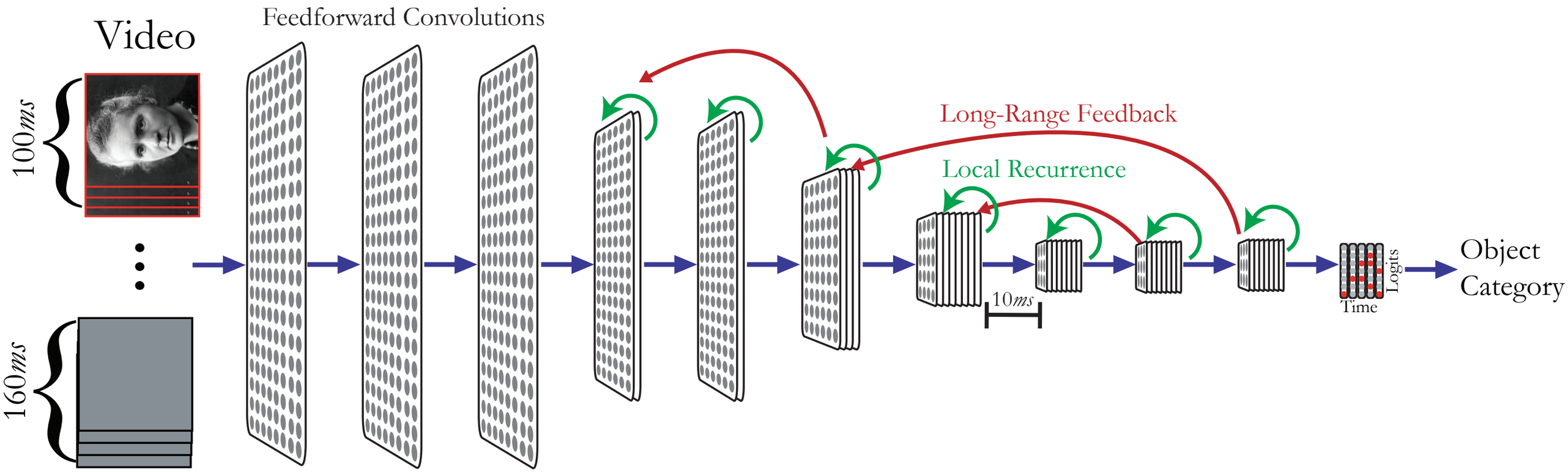
Gilbert & Li (2013)

. . . but what are they for?

Convolutional Recurrent Networks (ConvRNNs)

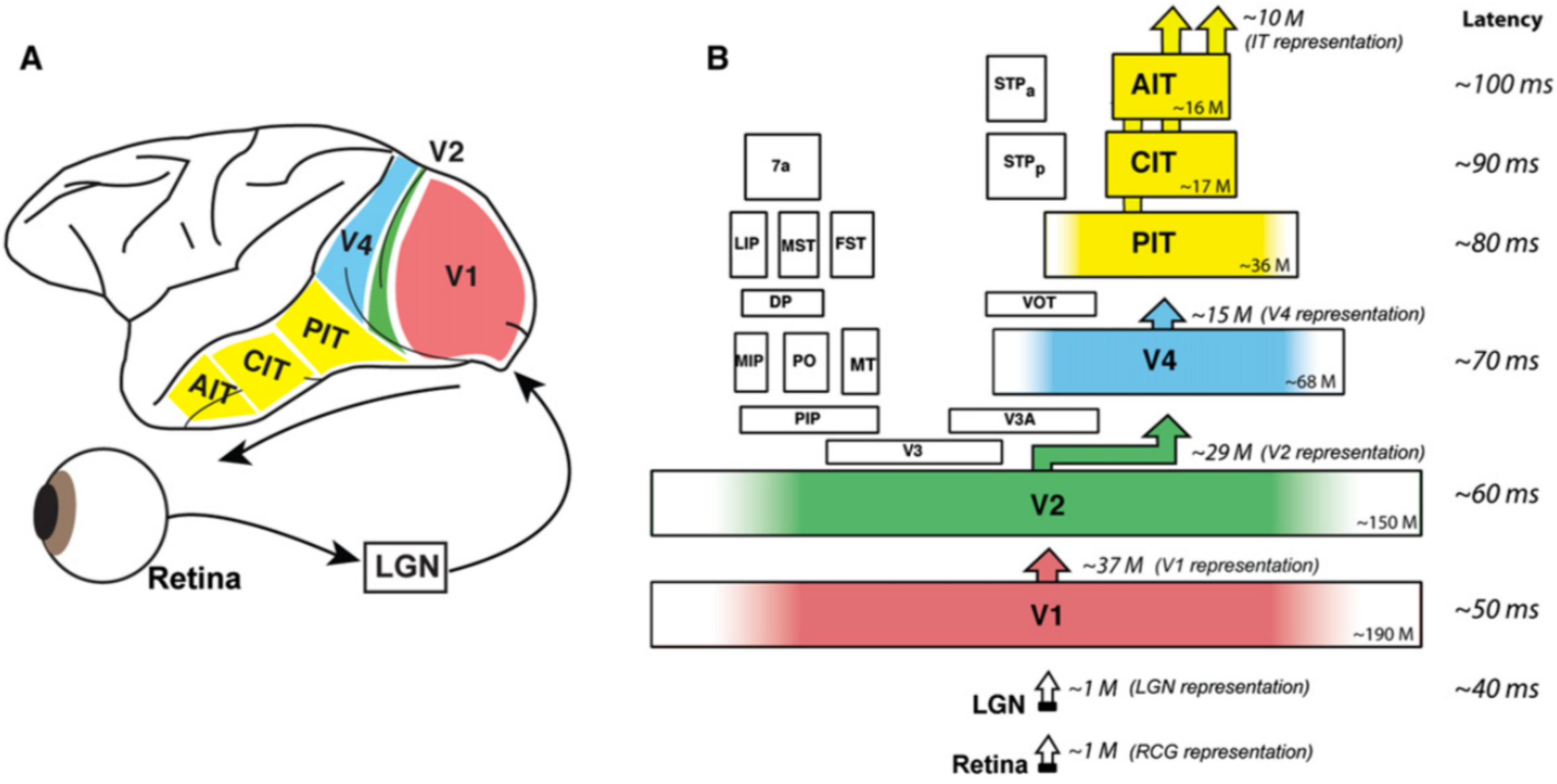


Convolutional Recurrent Networks (ConvRNNs)



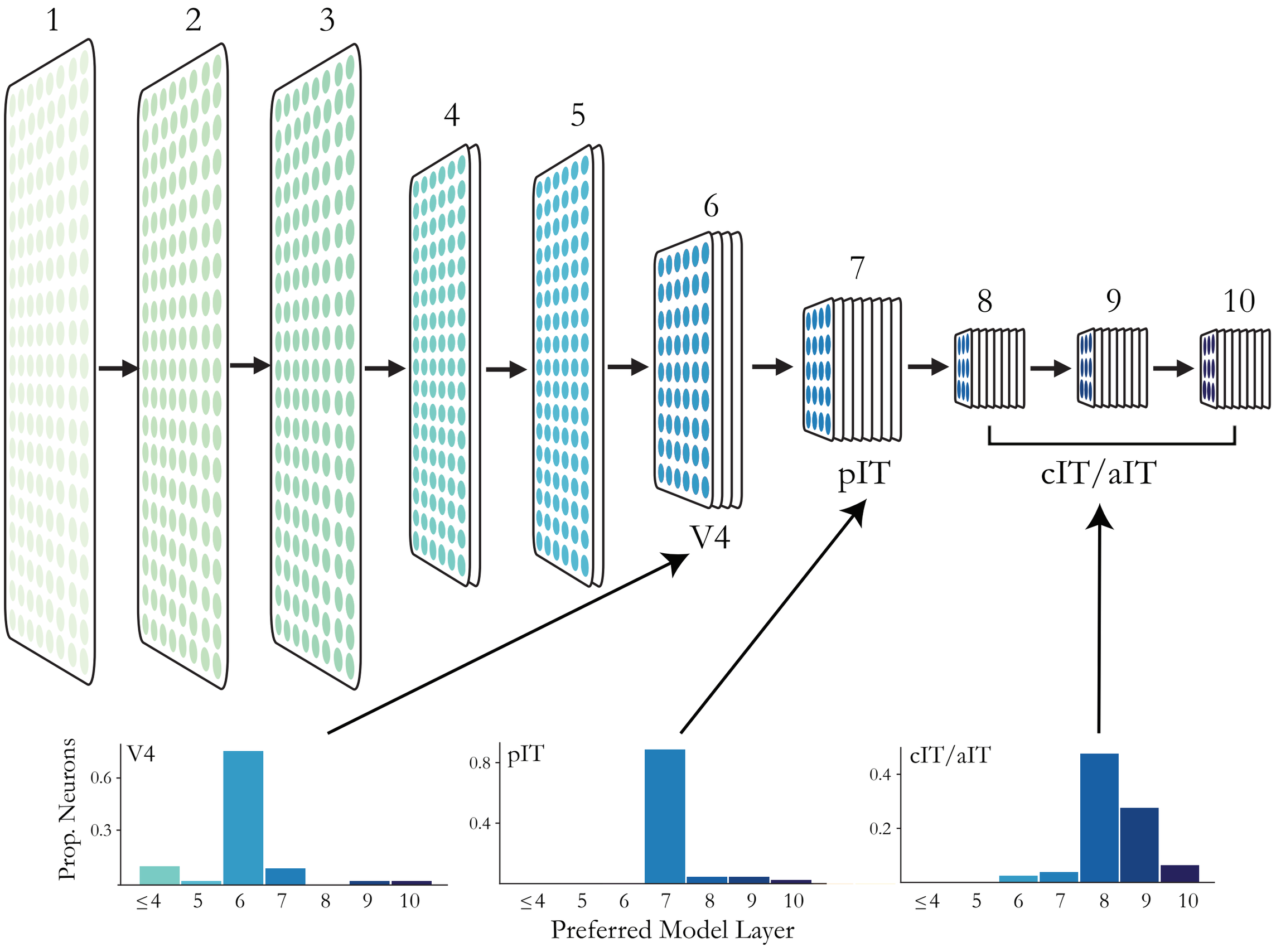
Each time-step (10 ms) is treated equally
— including feedforward steps

~10-12 "Layers" Plausible based on anatomy and timing



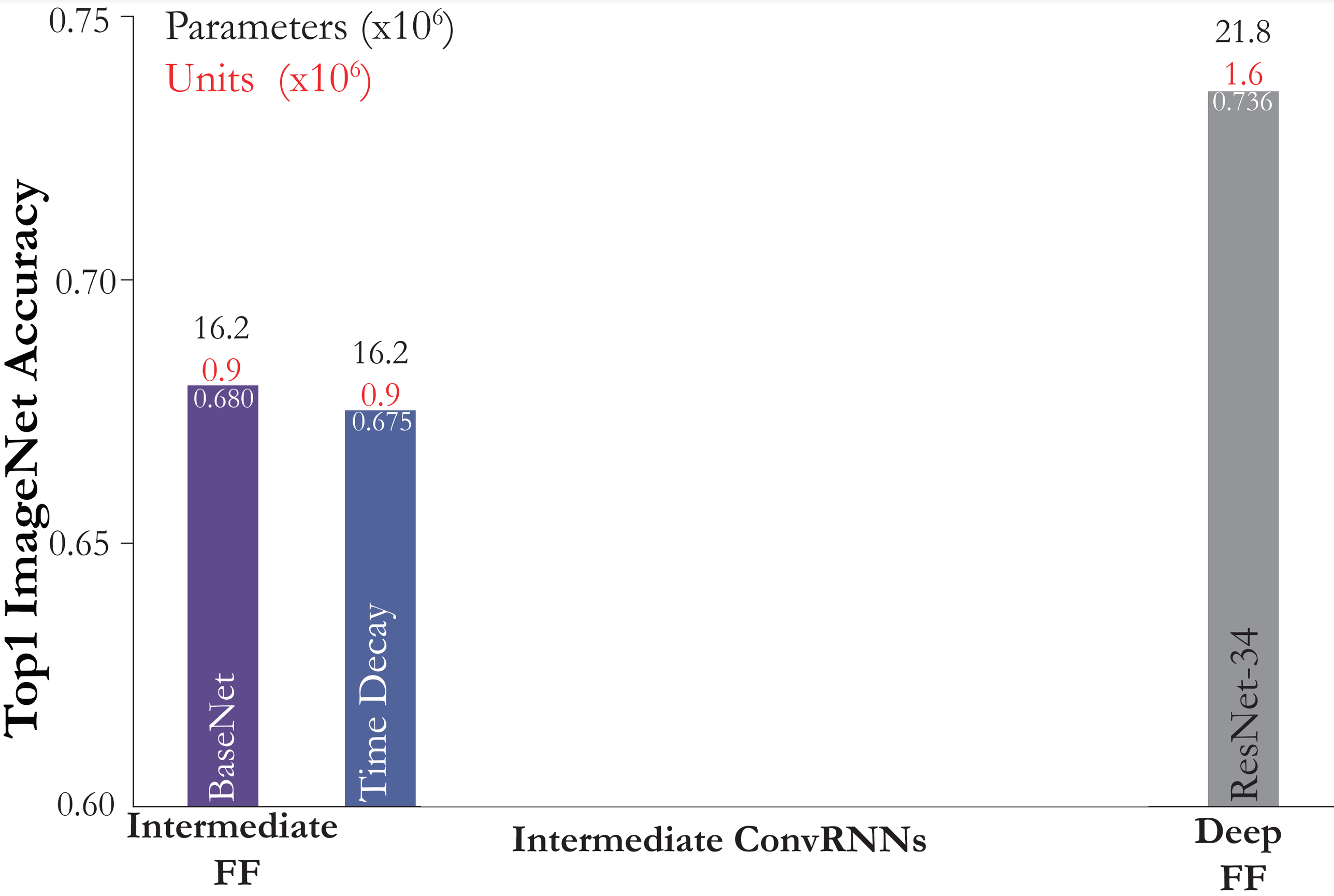
DiCarlo, Zoccolan, & Rust 2012

~10-12 "Layers" Plausible based on anatomy and timing

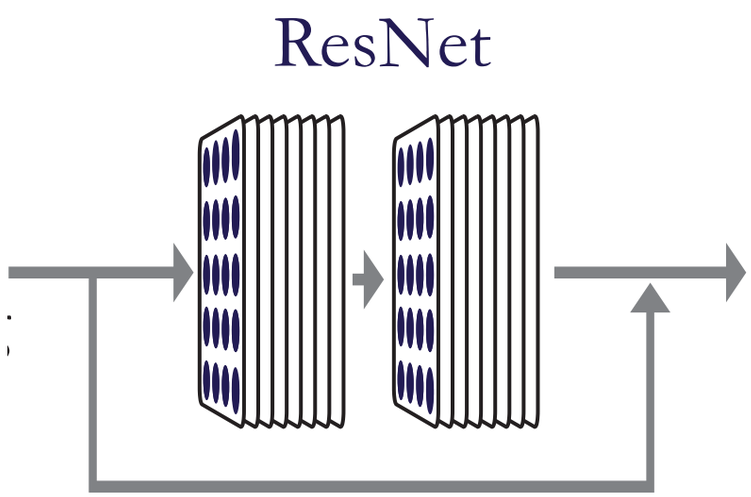


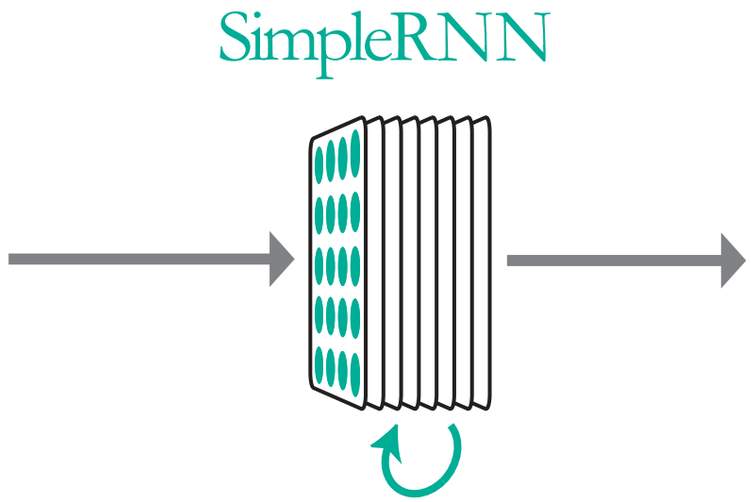
…But, such networks are not the most performant

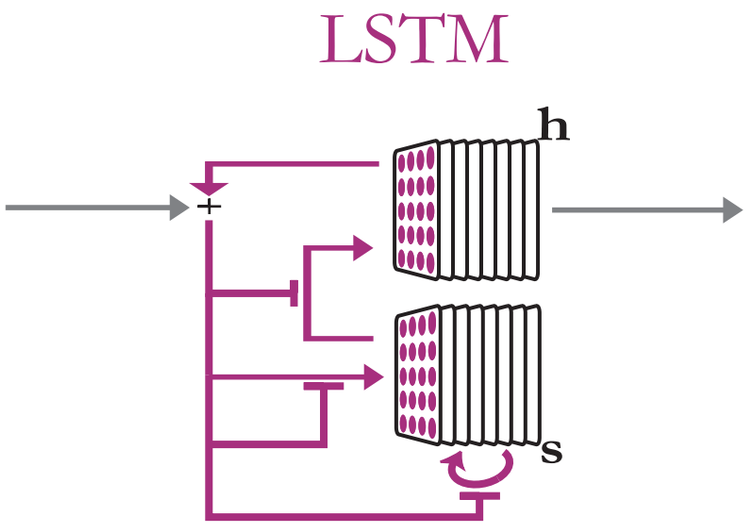
…adding simple recurrence doesn't work either



Many Choices of Local Recurrence







Circuit Diagram

Principles of Local Recurrence

Passthrough Mechanism

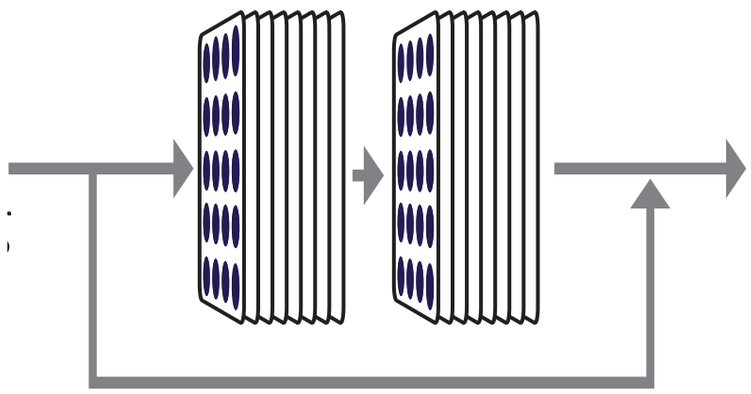
Input x_t → Output $f(W*x_t+b)$

State = 0:

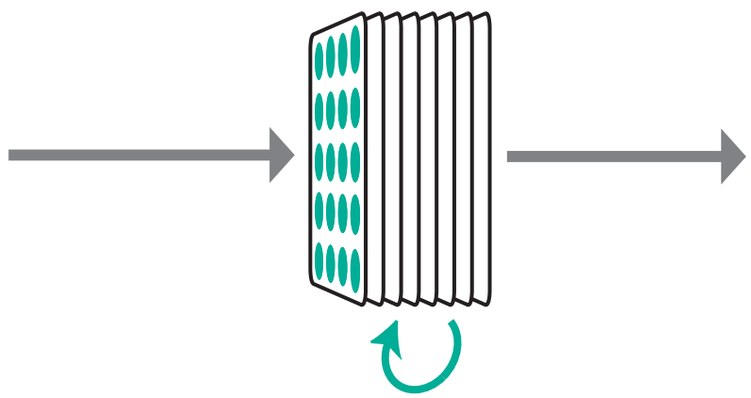
Gating Mechanism

Input x_t → Output $f(W*x_t+b) \circ g_t$, with g_t feeding back

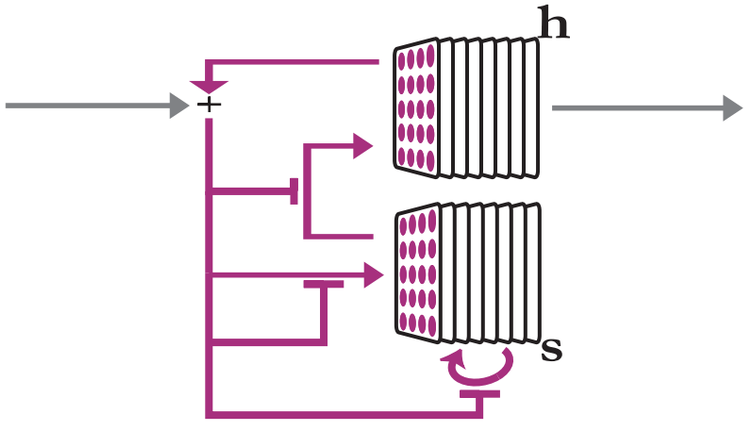
ResNet



SimpleRNN

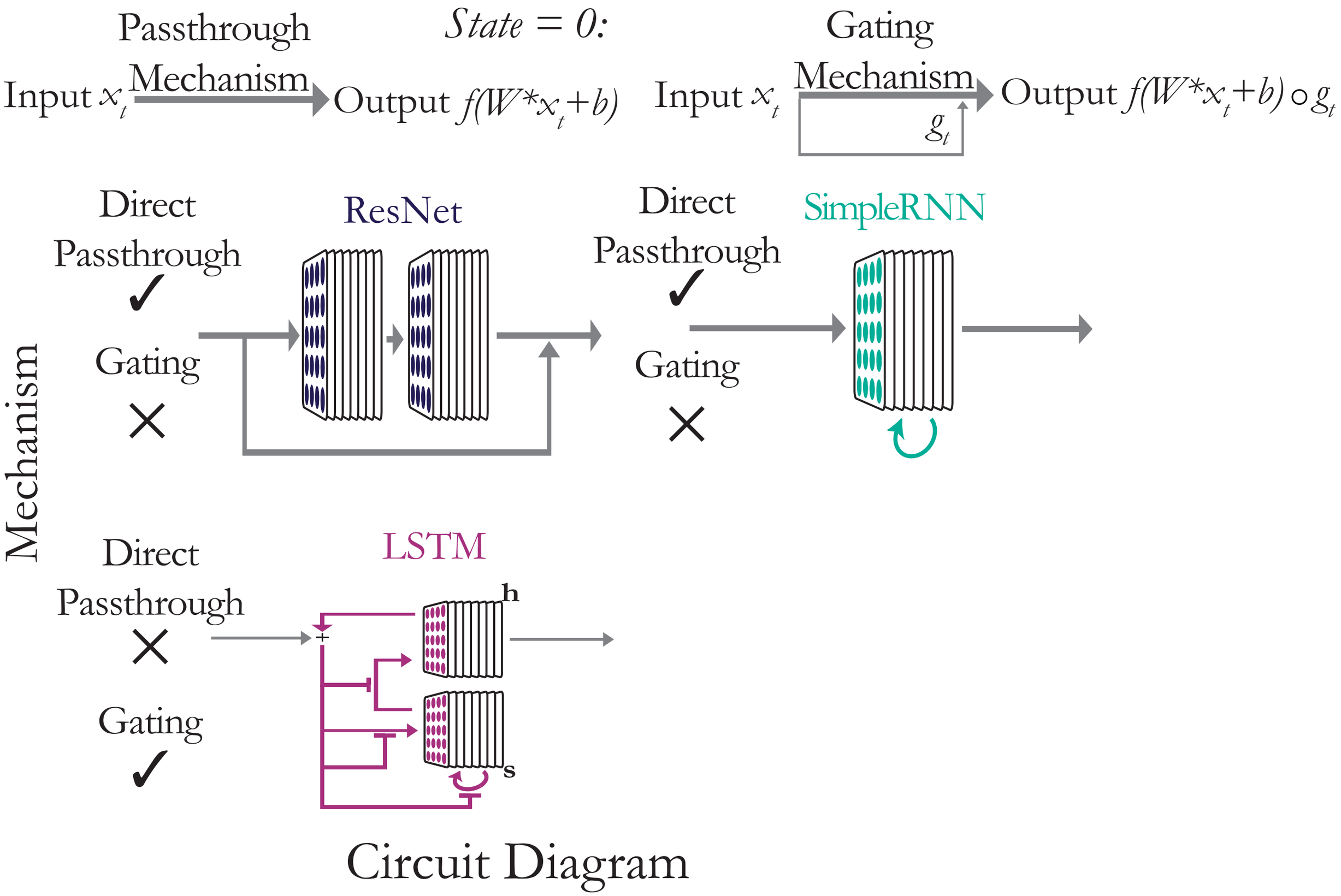


LSTM

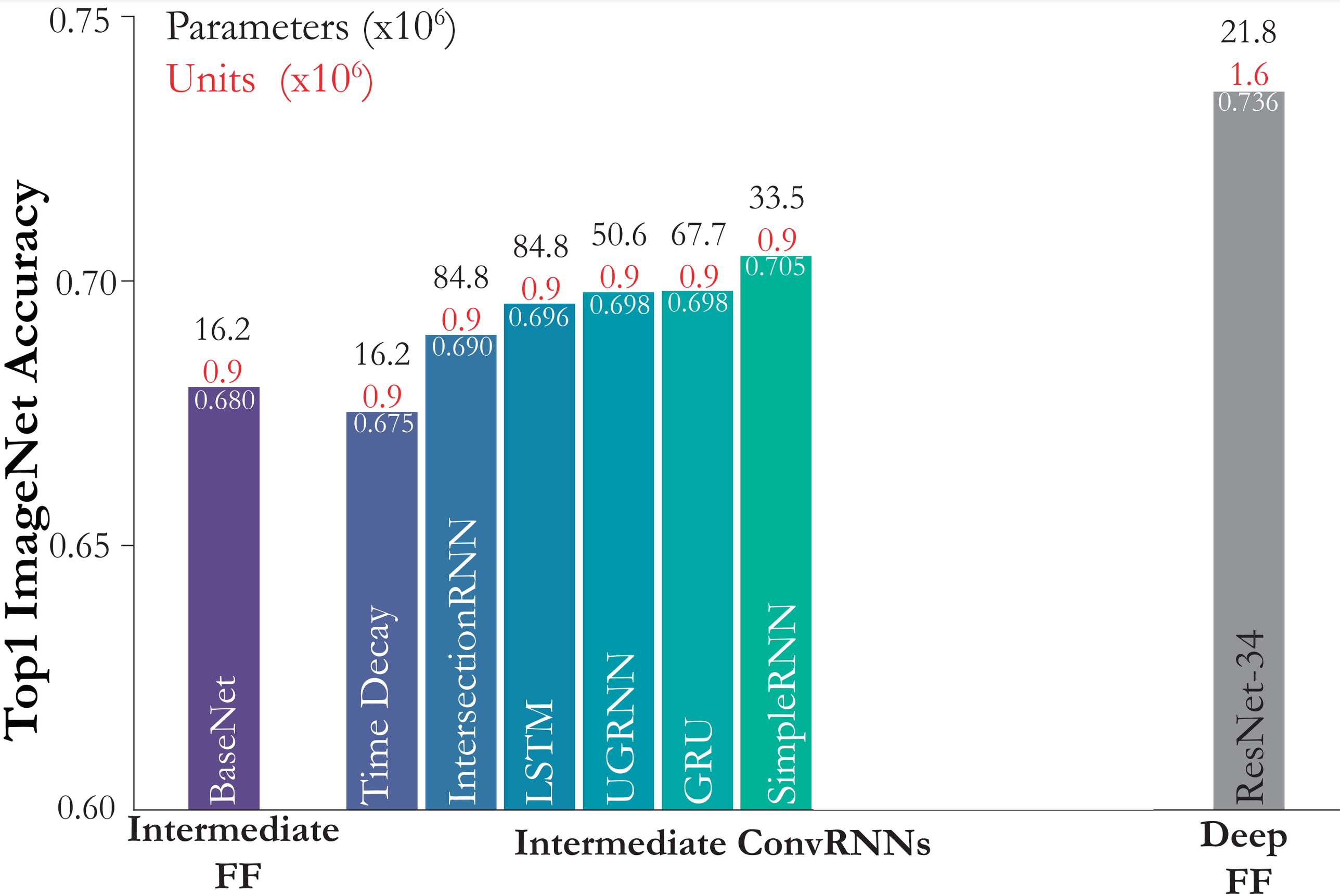


Circuit Diagram

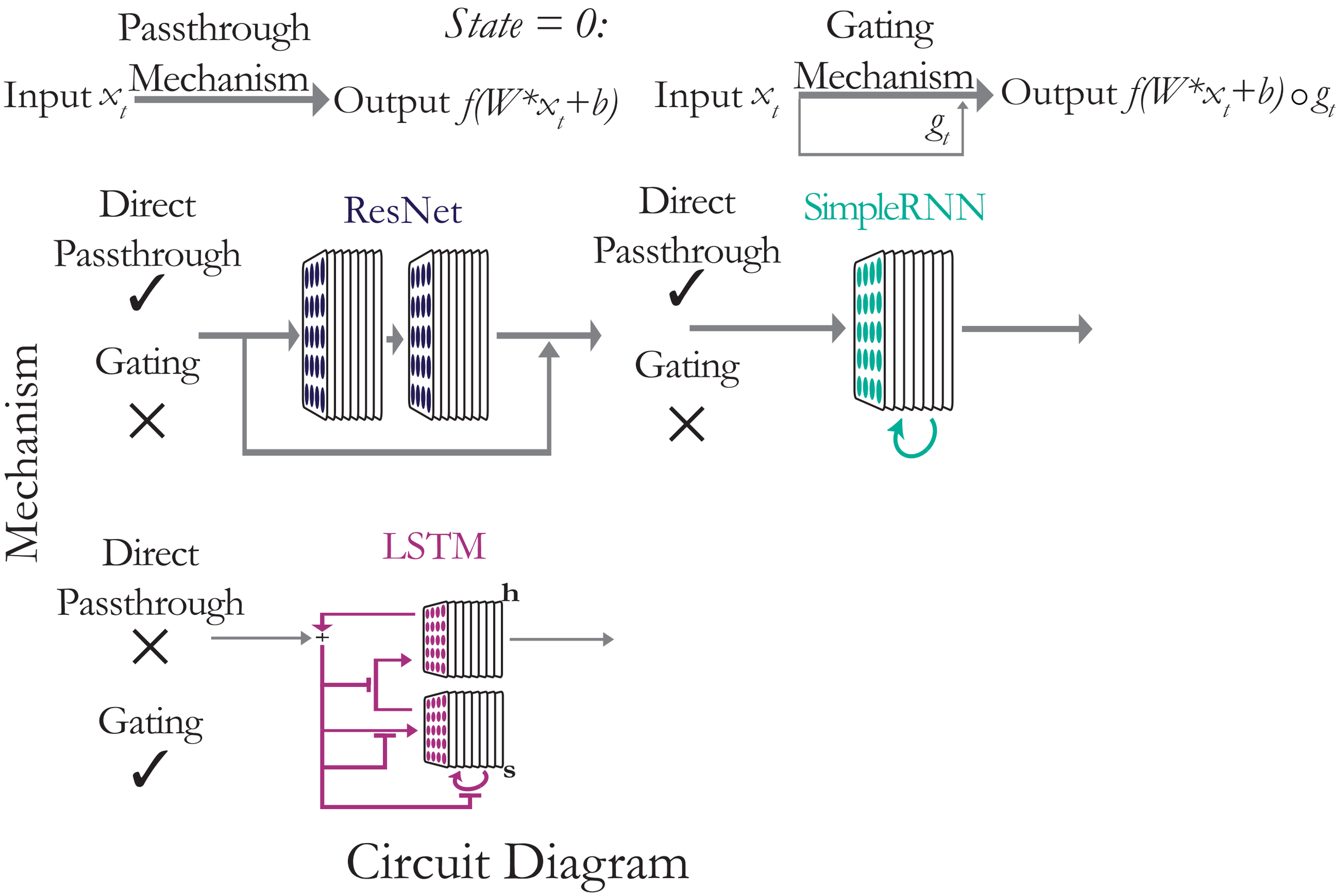
Principles of Local Recurrence



Adding these helps somewhat



Principles of Local Recurrence



Principles of Local Recurrence

Passthrough Mechanism

Input x_t → Output $f(W*x_t+b)$

State = 0:

Gating Mechanism

Input x_t → Output $f(W*x_t+b) \circ g_t$, with g_t

Mechanism

Direct Passthrough	ResNet	Direct Passthrough	SimpleRNN
✓		✓	
Gating ×		Gating ×	

Direct Passthrough	LSTM	Direct Passthrough	Reciprocal Gated Circuit
×		✓	
Gating ✓		Gating ✓	

Circuit Diagram

Principles of Local Recurrence

Passthrough Mechanism

Input x_t → Output $f(W*x_t+b)$

State = 0:

Gating Mechanism

Input x_t → Output $f(W*x_t+b) \circ g_t$, with g_t

Mechanism

Direct Passthrough	Gating	ResNet	LSTM	Reciprocal Gated Circuit
✓	×			
×	✓			
✓	✓			

Direct Passthrough ✓ / Gating × → ResNet

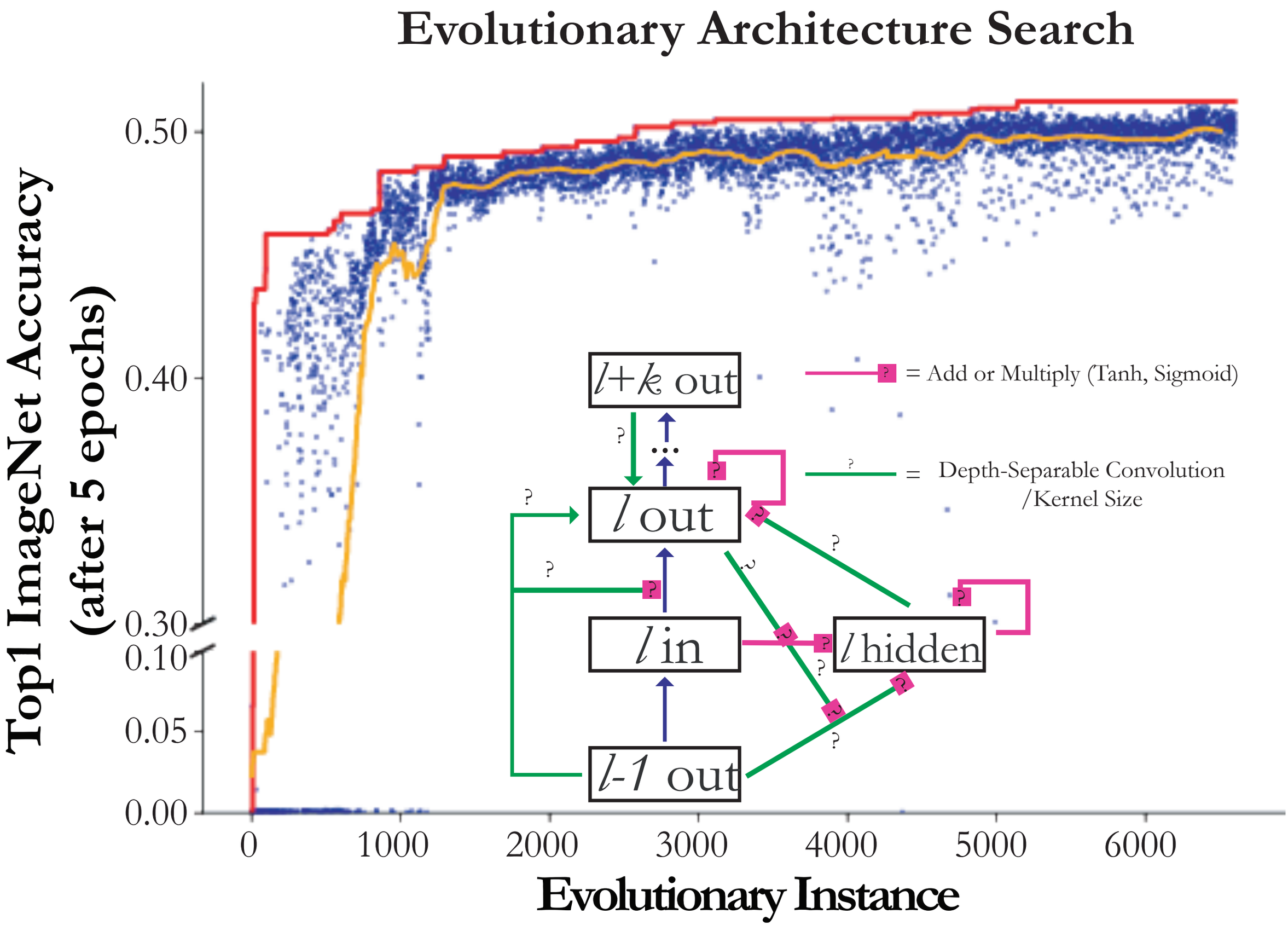
Direct Passthrough ✓ / Gating × → SimpleRNN

Direct Passthrough × / Gating ✓ → LSTM (h, s)

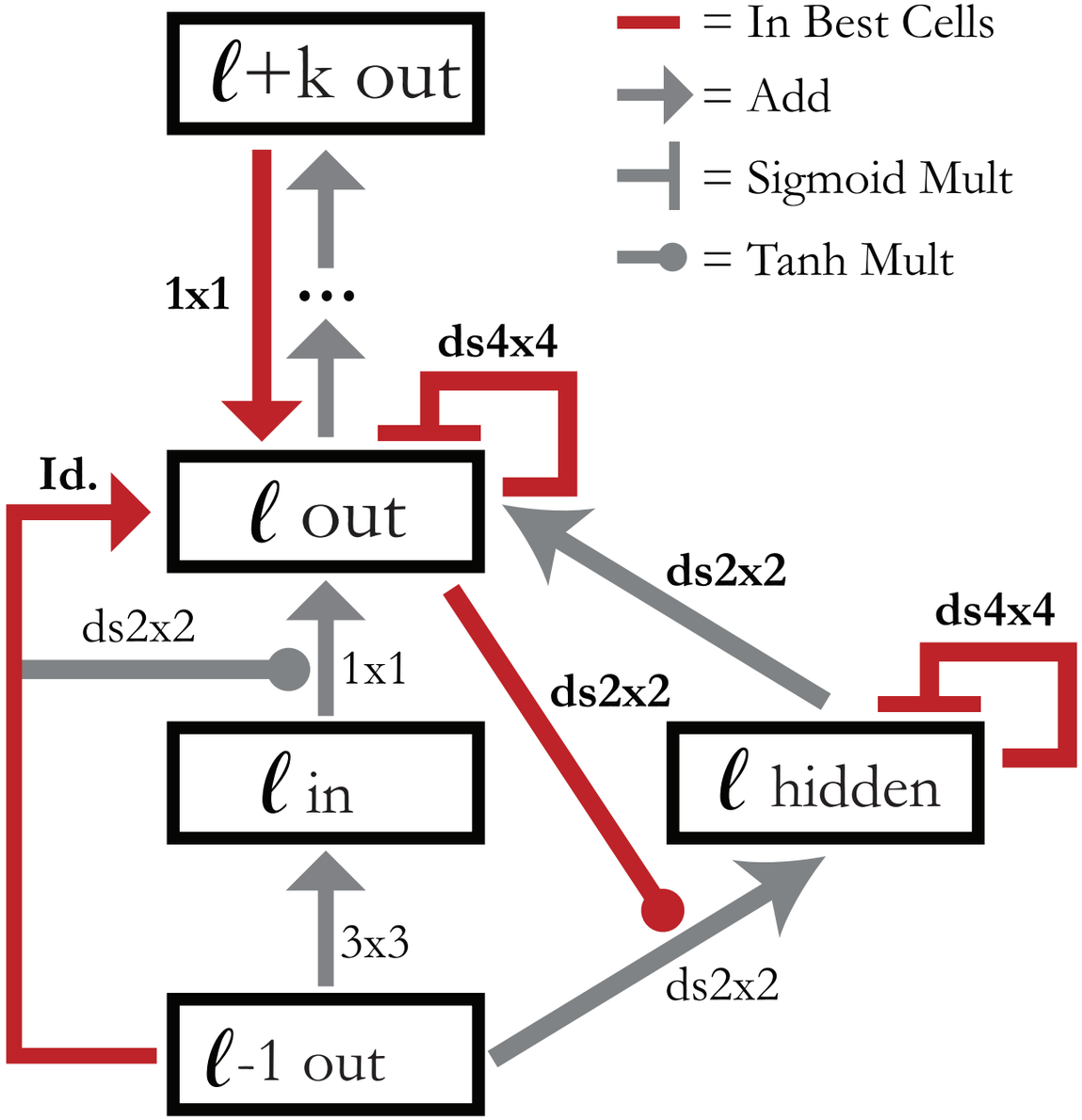
Direct Passthrough ✓ / Gating ✓ → Reciprocal Gated Circuit (h, s)

Circuit Diagram

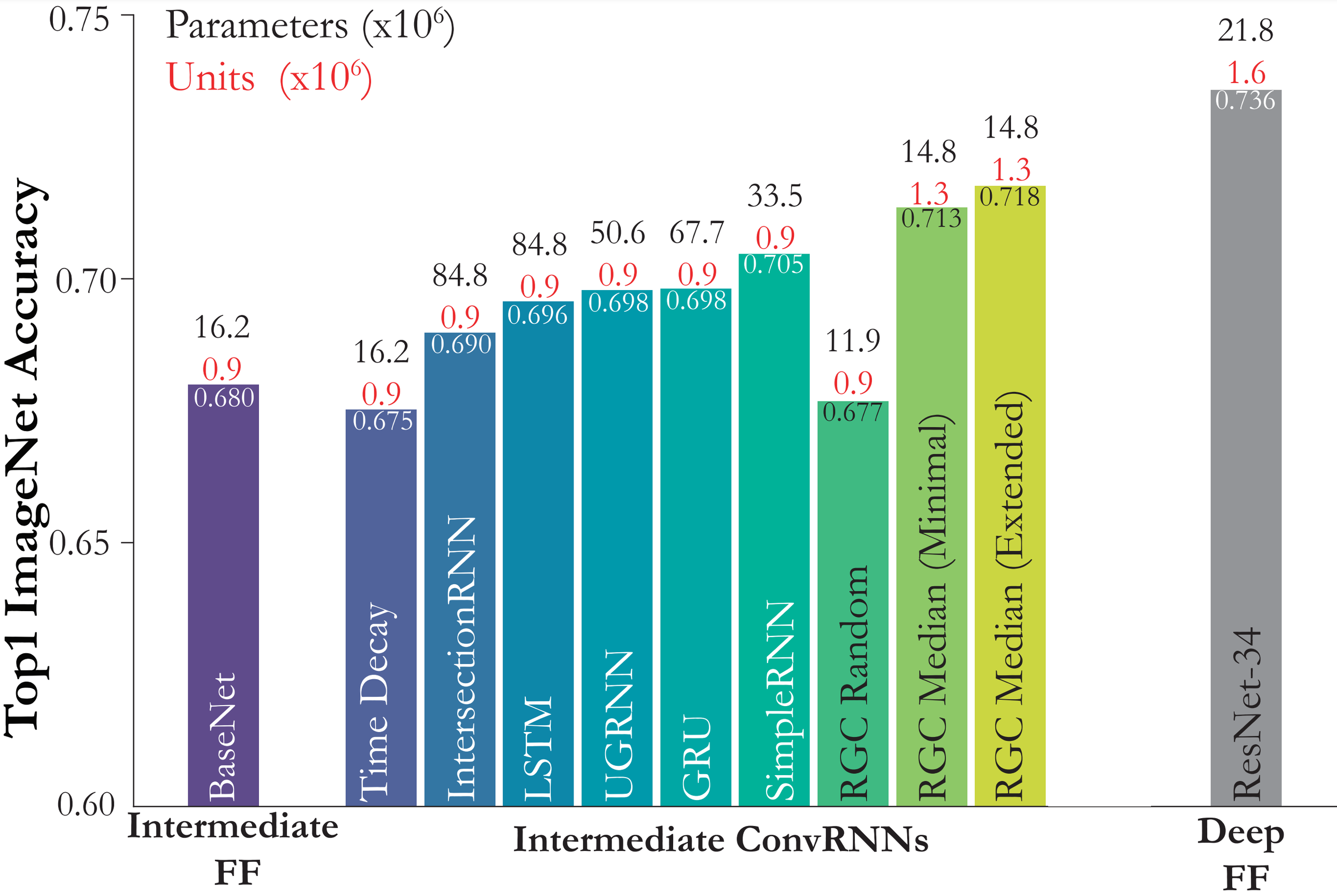
Search over local and global recurrence



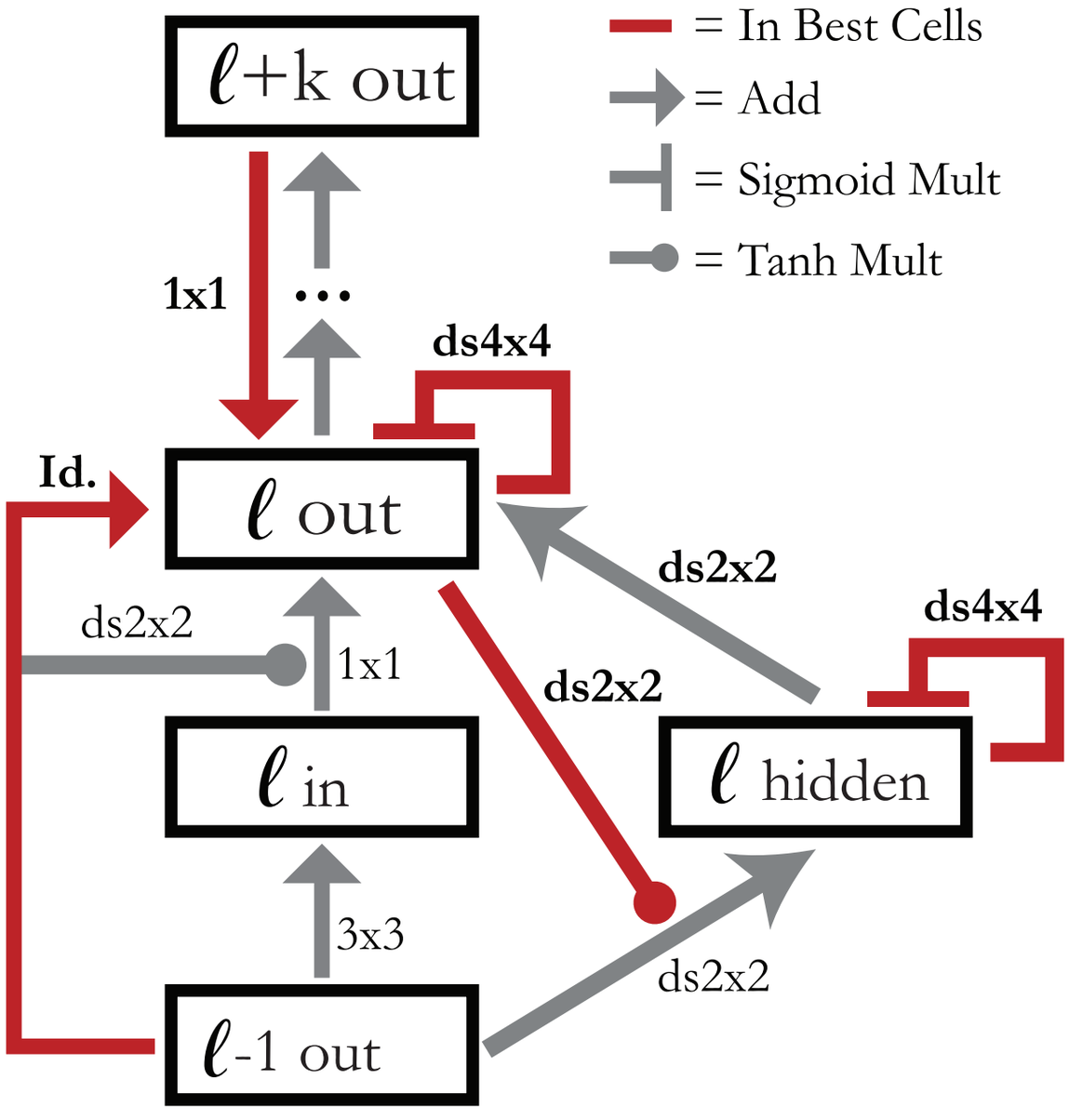
Emergent Local Connectivity Patterns

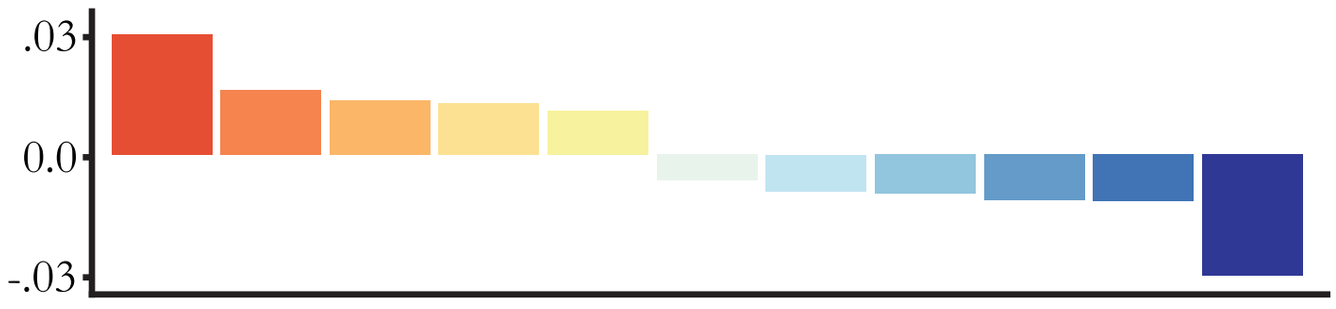


Evolutionary search yields improved performance



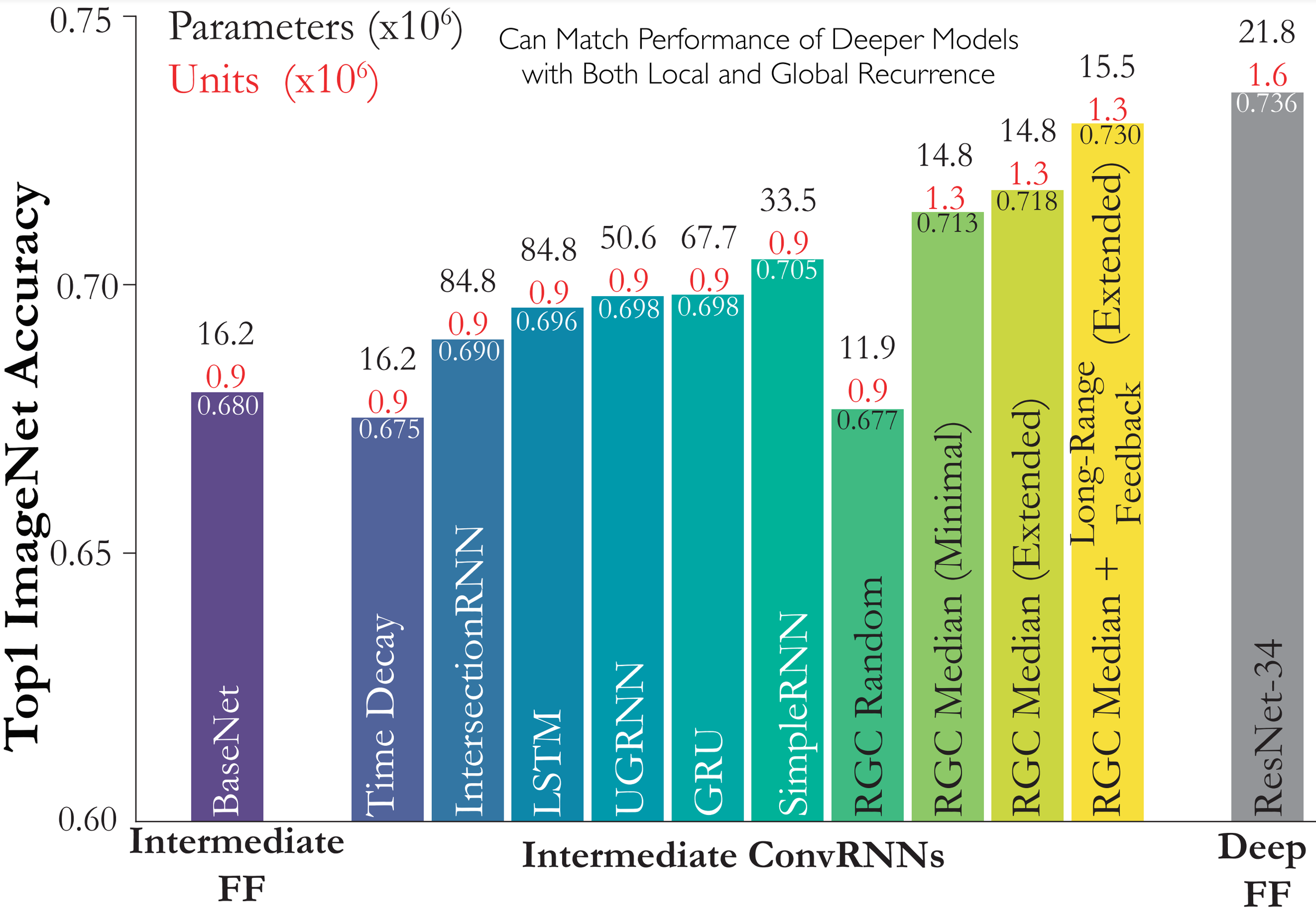
Emergent Local and Global Connectivity Patterns



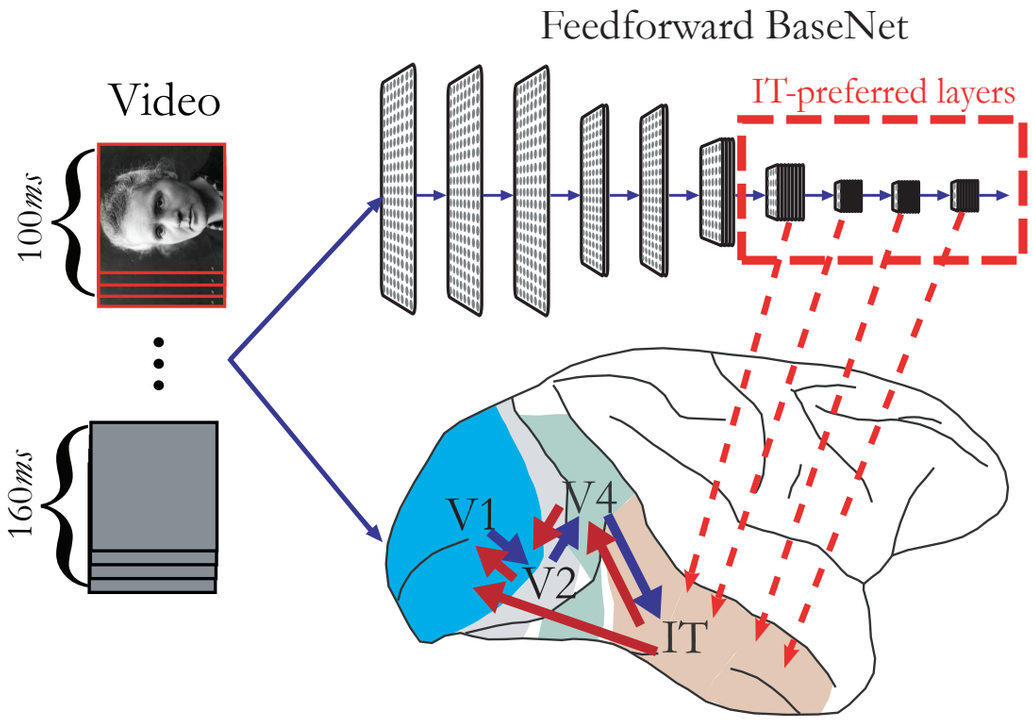


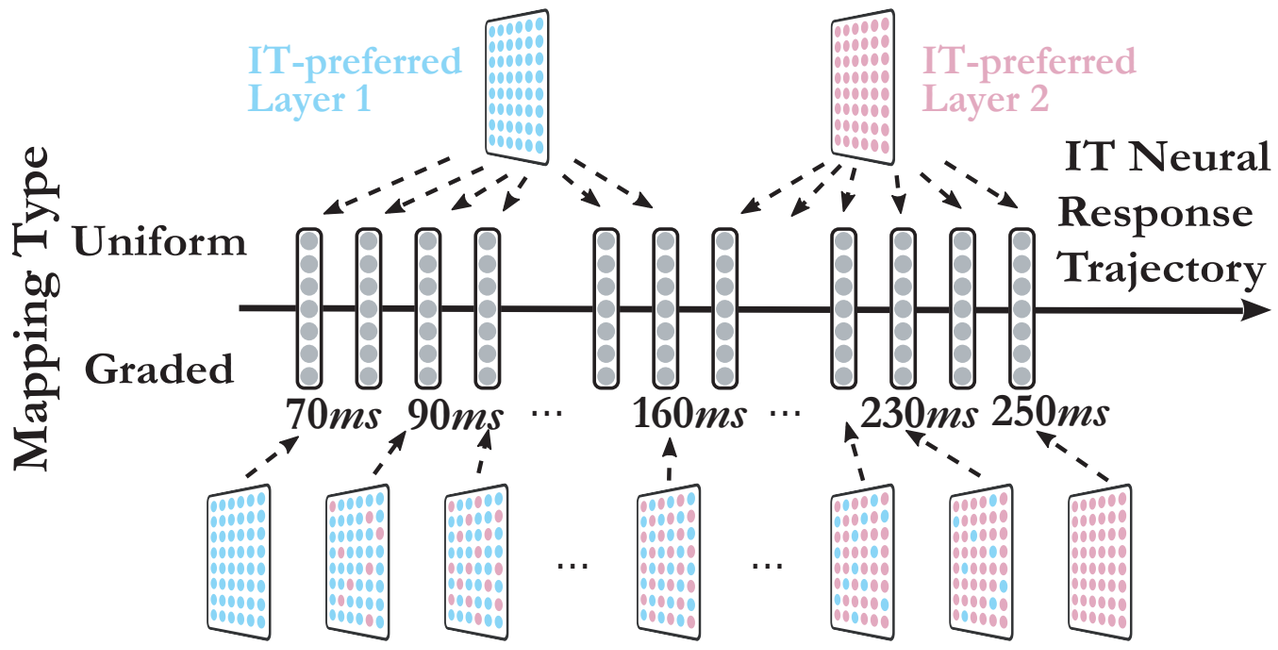
Top1 Gain from Feedback (Best to Worst)

Global Feedback Connections Matter



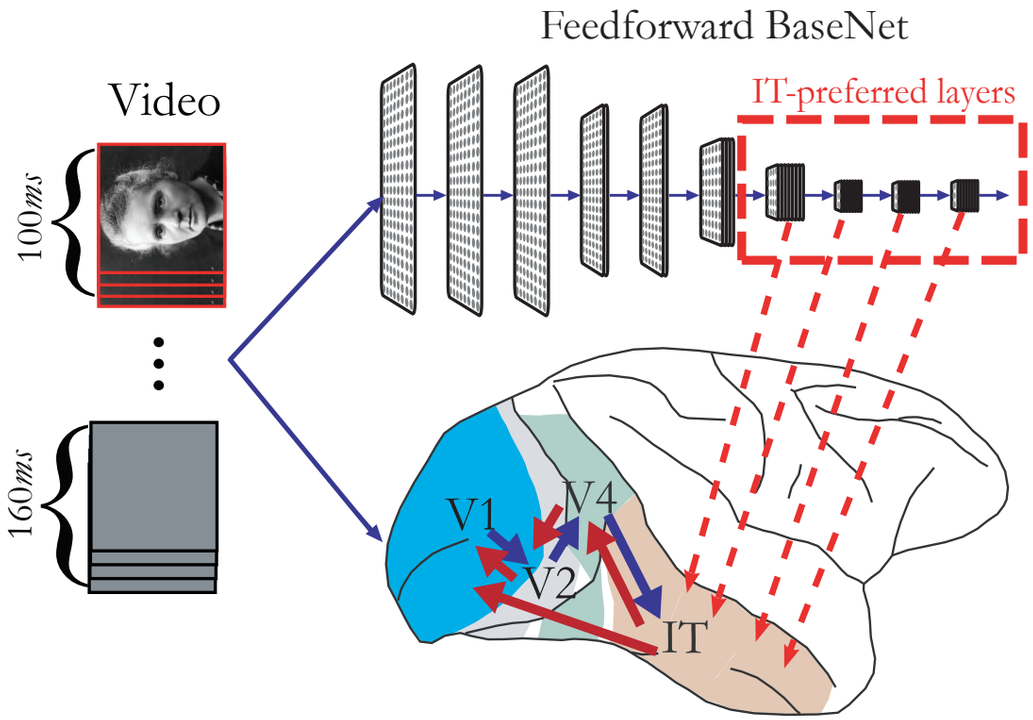
Comparing to Primate Object Solution Times (OSTs)

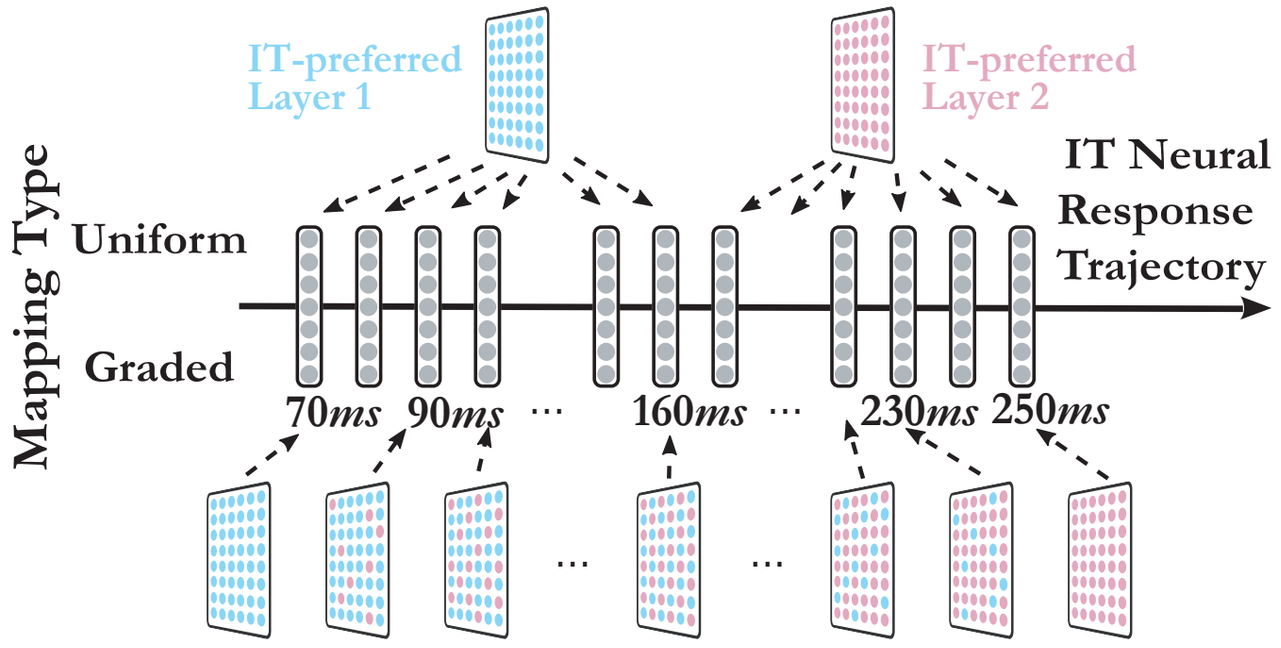


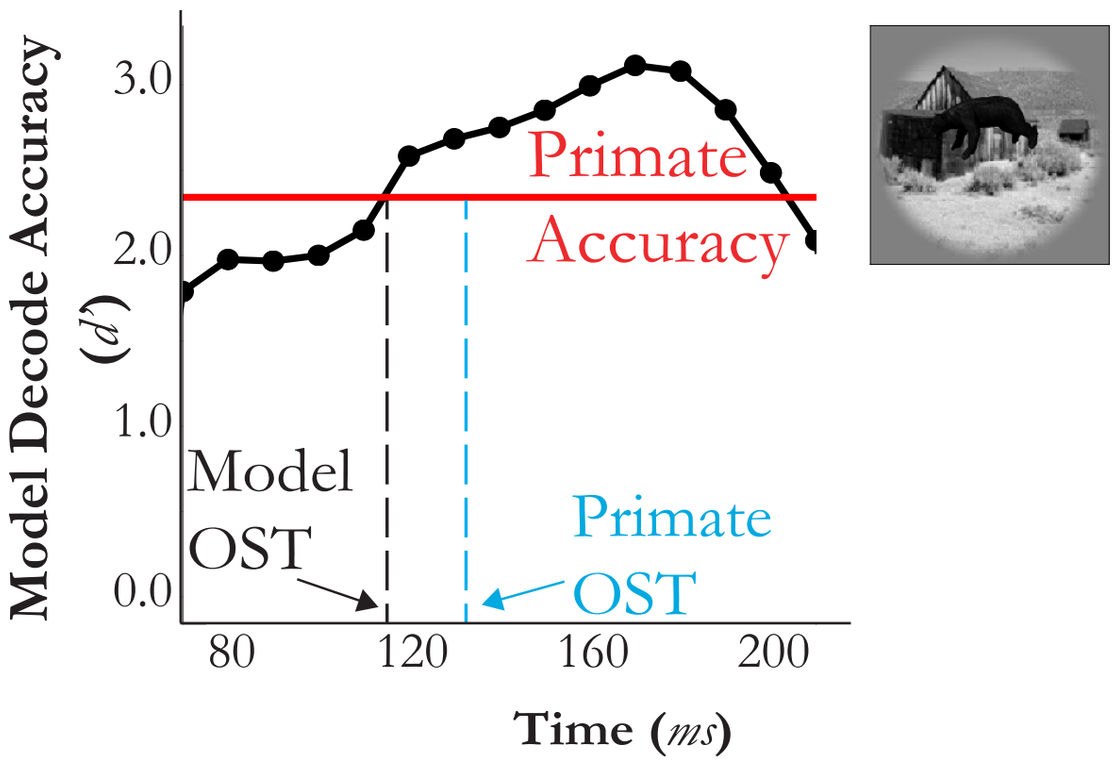


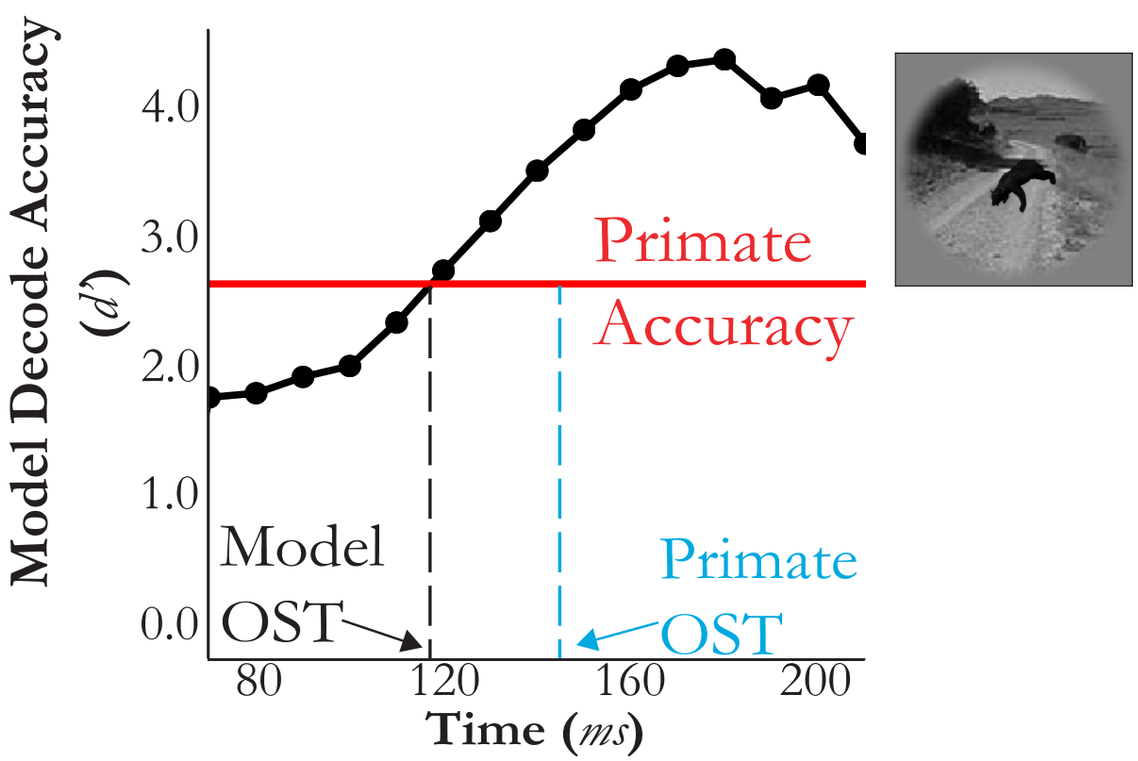
Both ConvRNNs and CNNs can be compared on this metric!

Comparing to Primate Object Solution Times (OSTs)

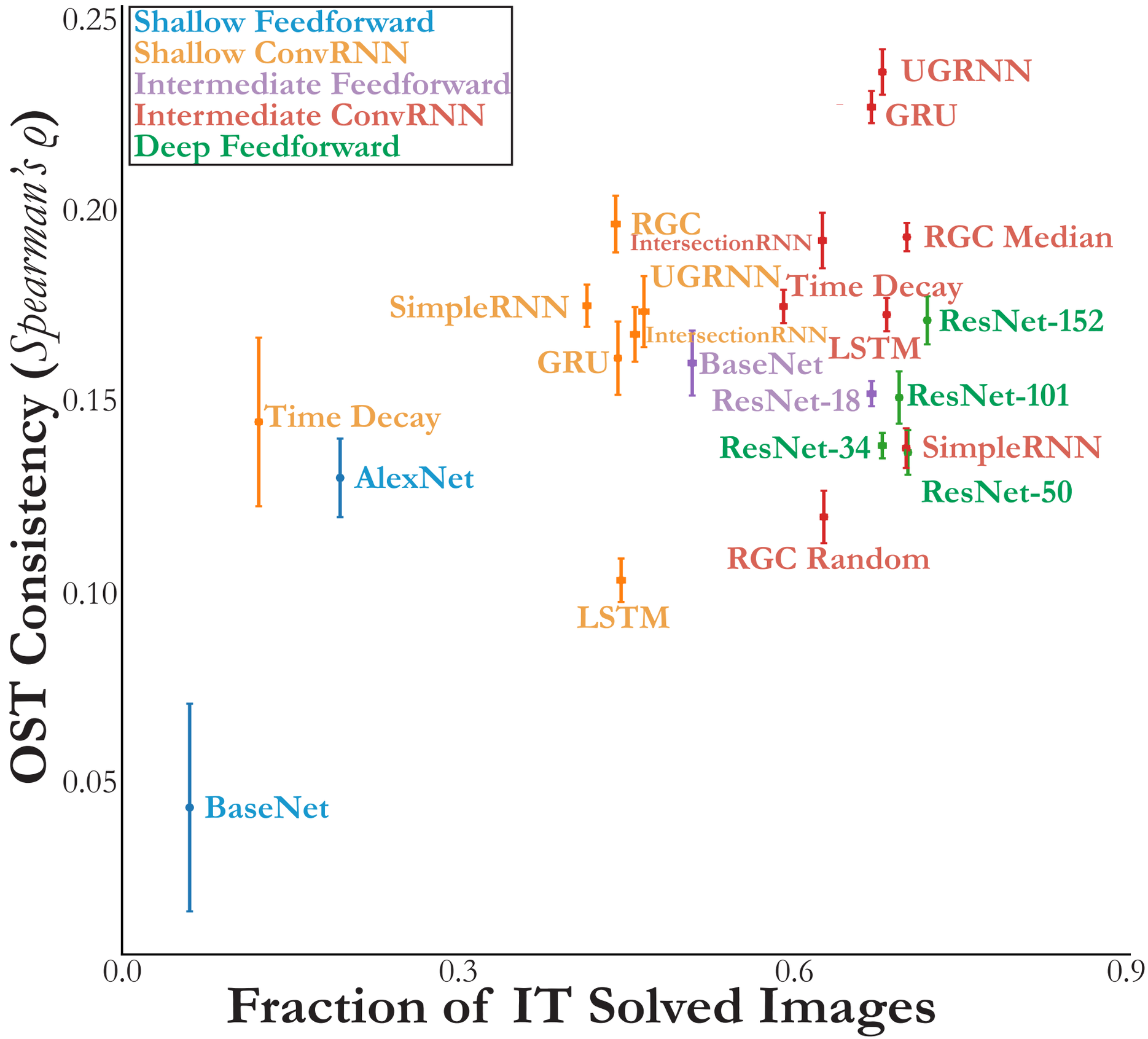




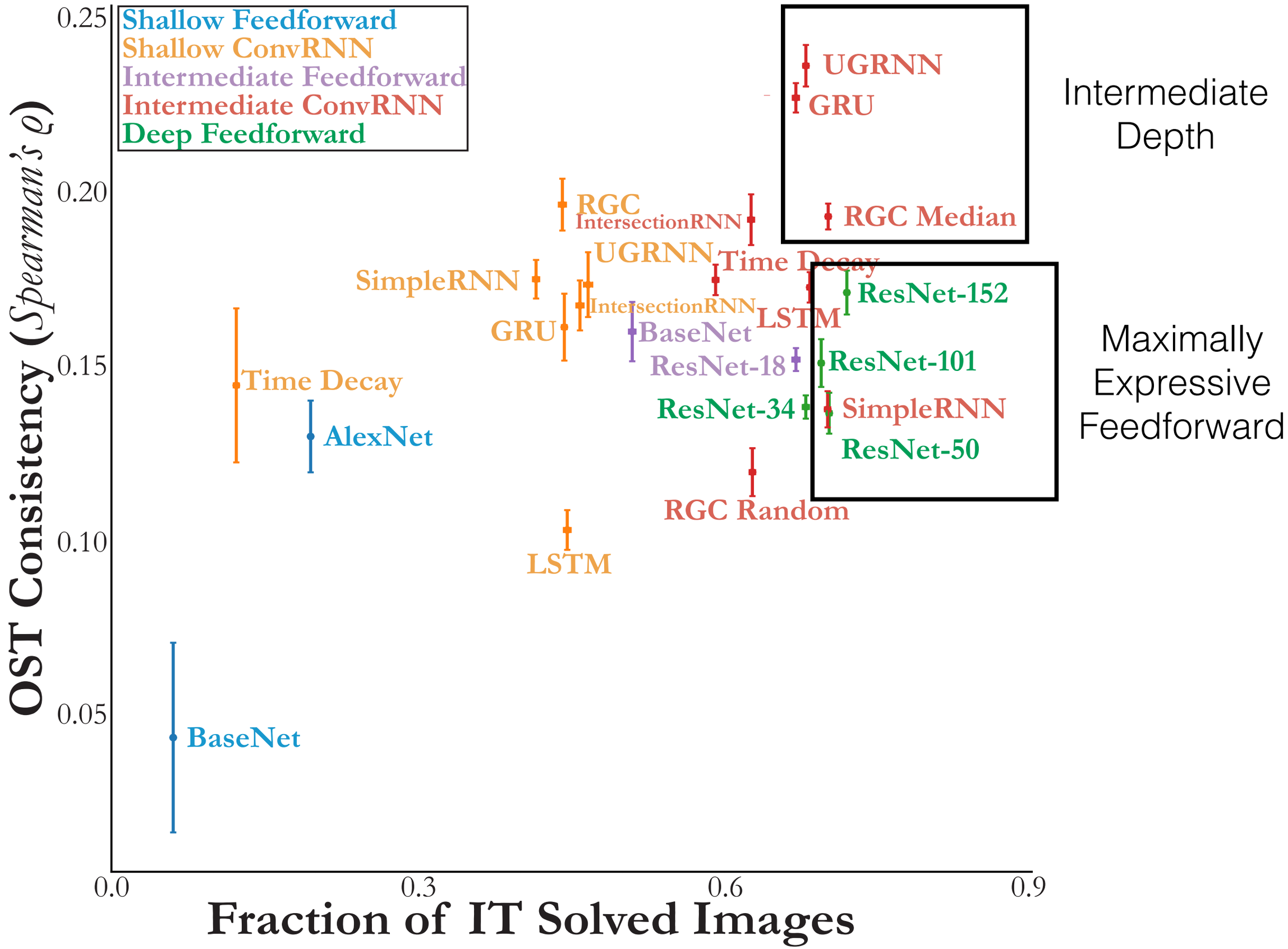




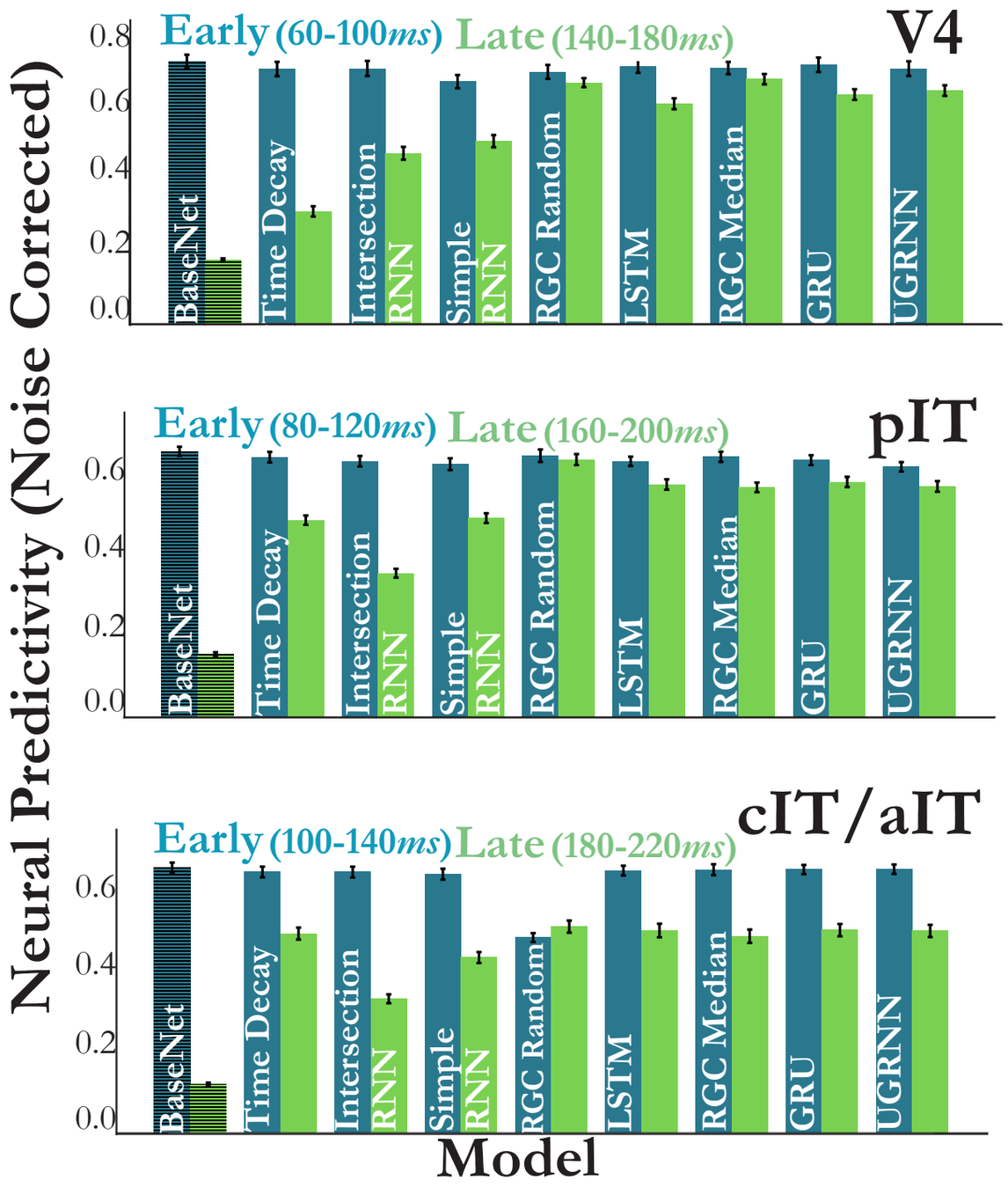
Intermediate ConvRNNs best match OSTs

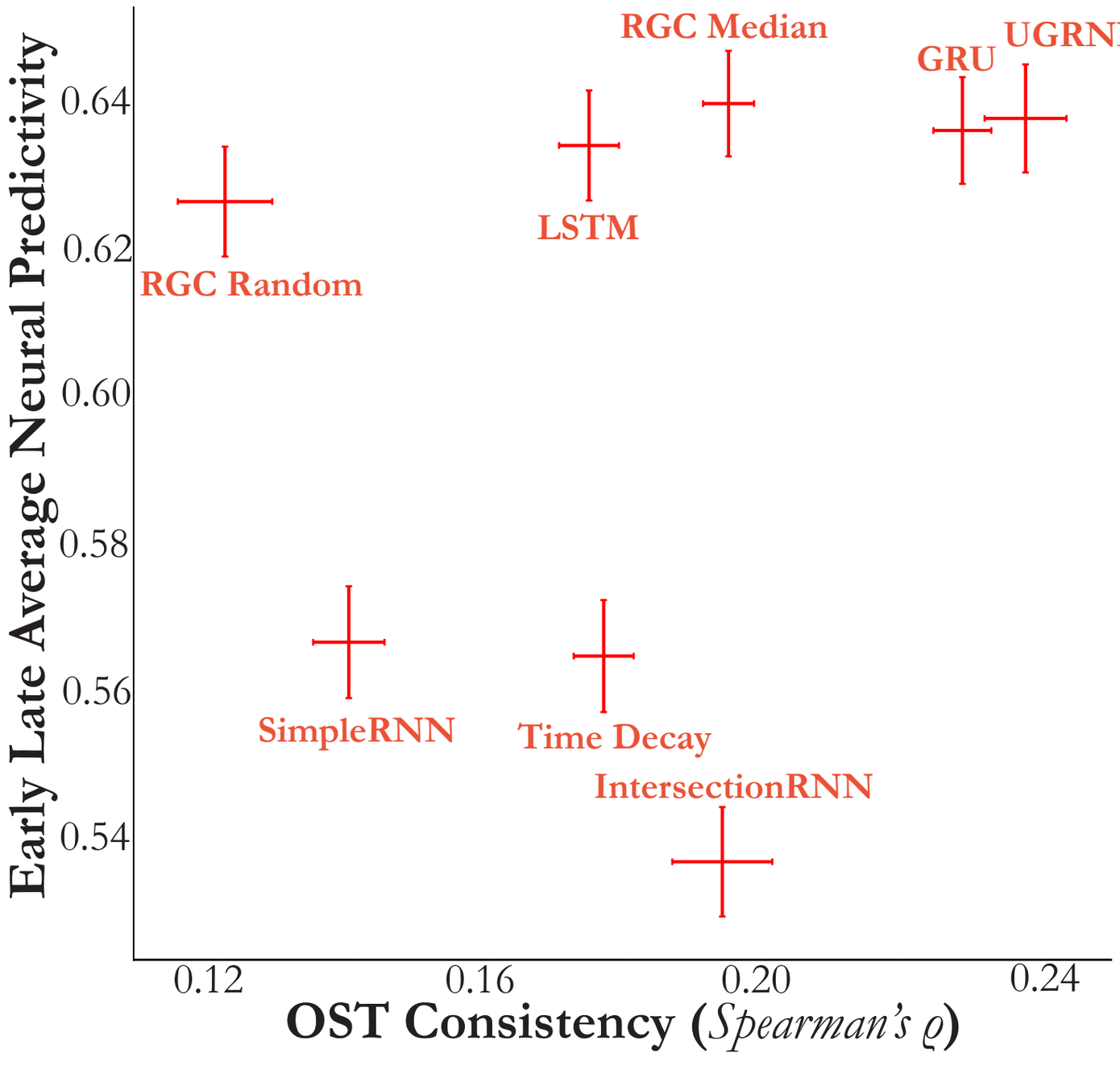


Intermediate ConvRNNs best match OSTs

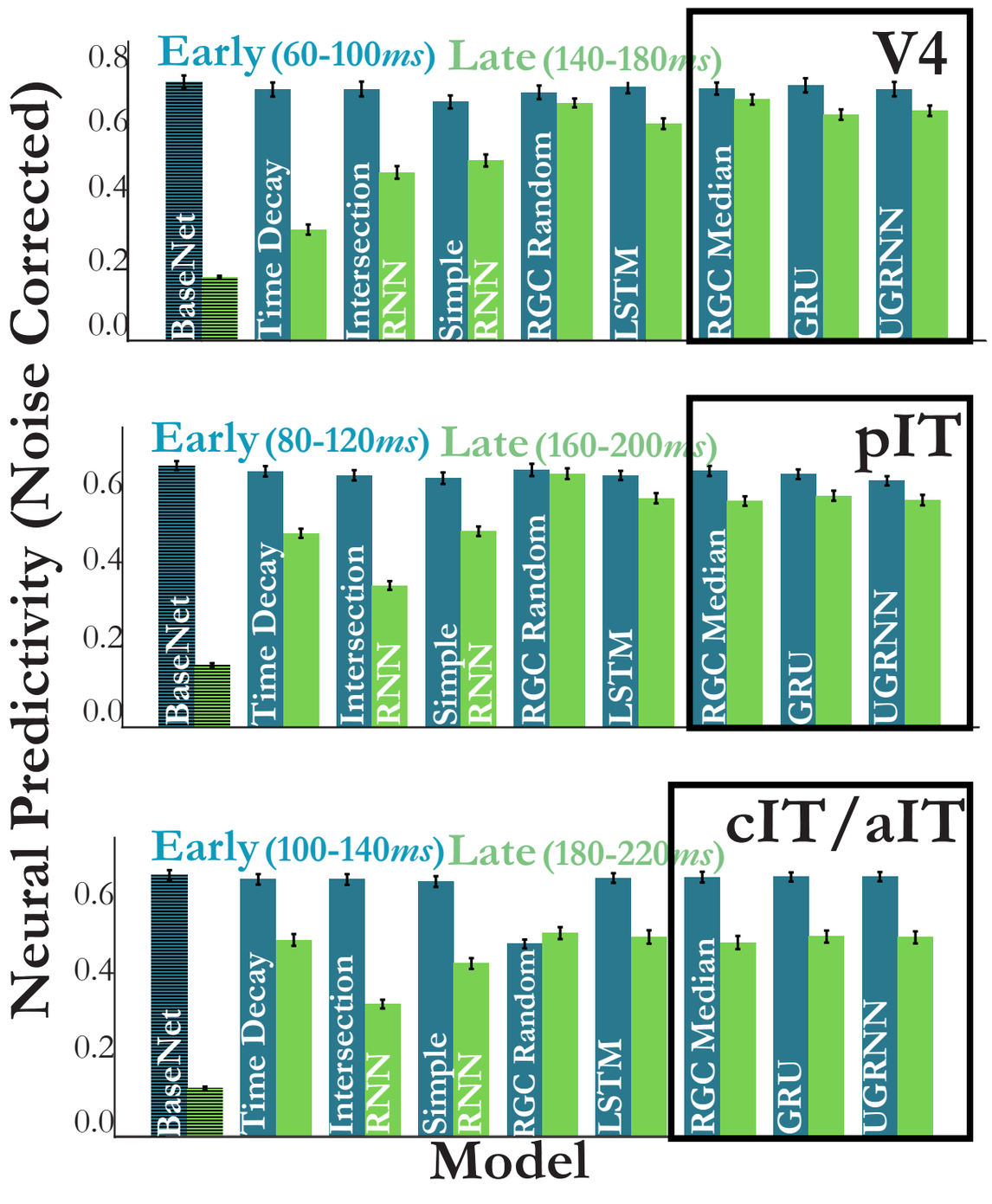


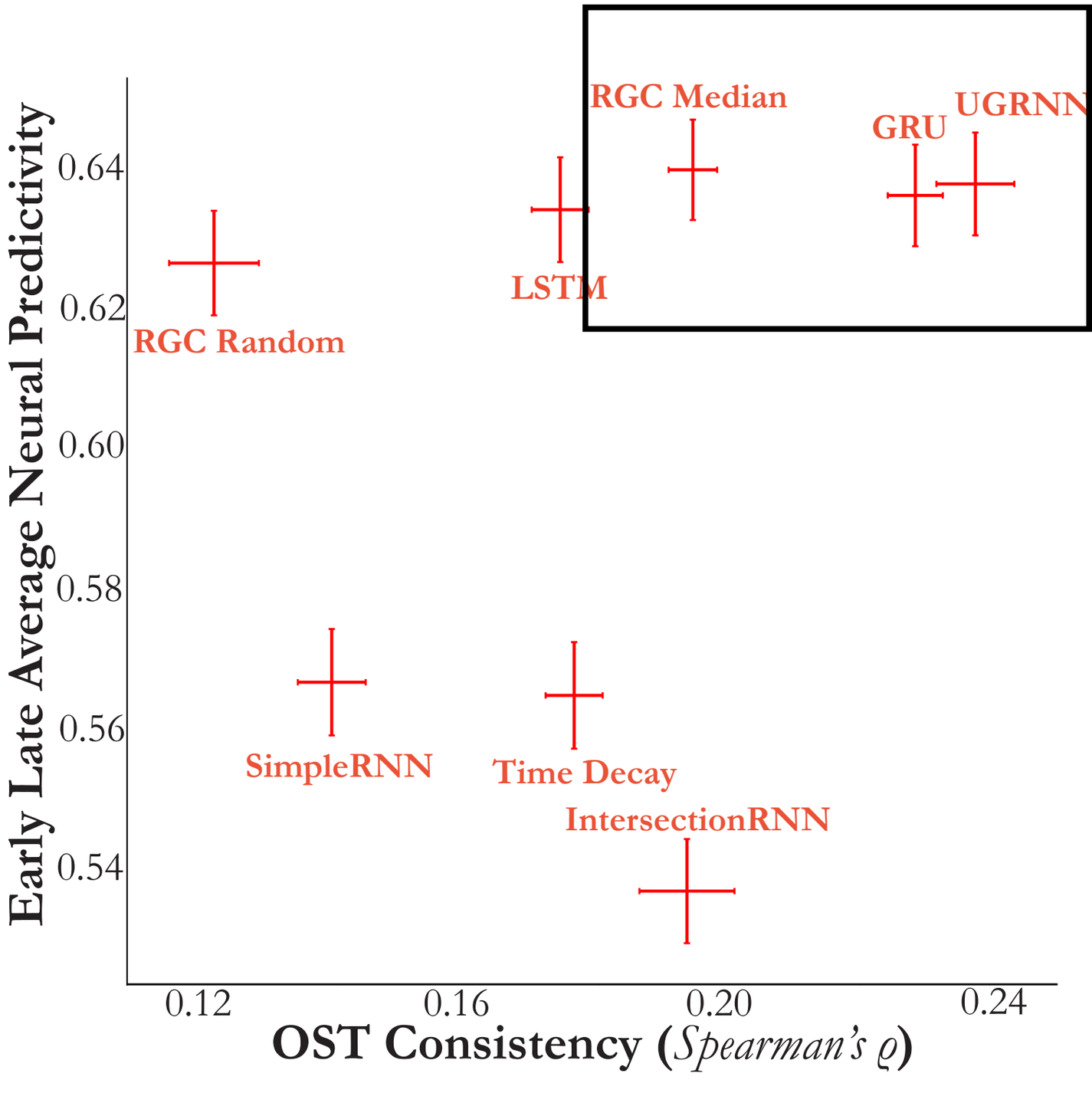
Intermediate ConvRNNs best match neural responses



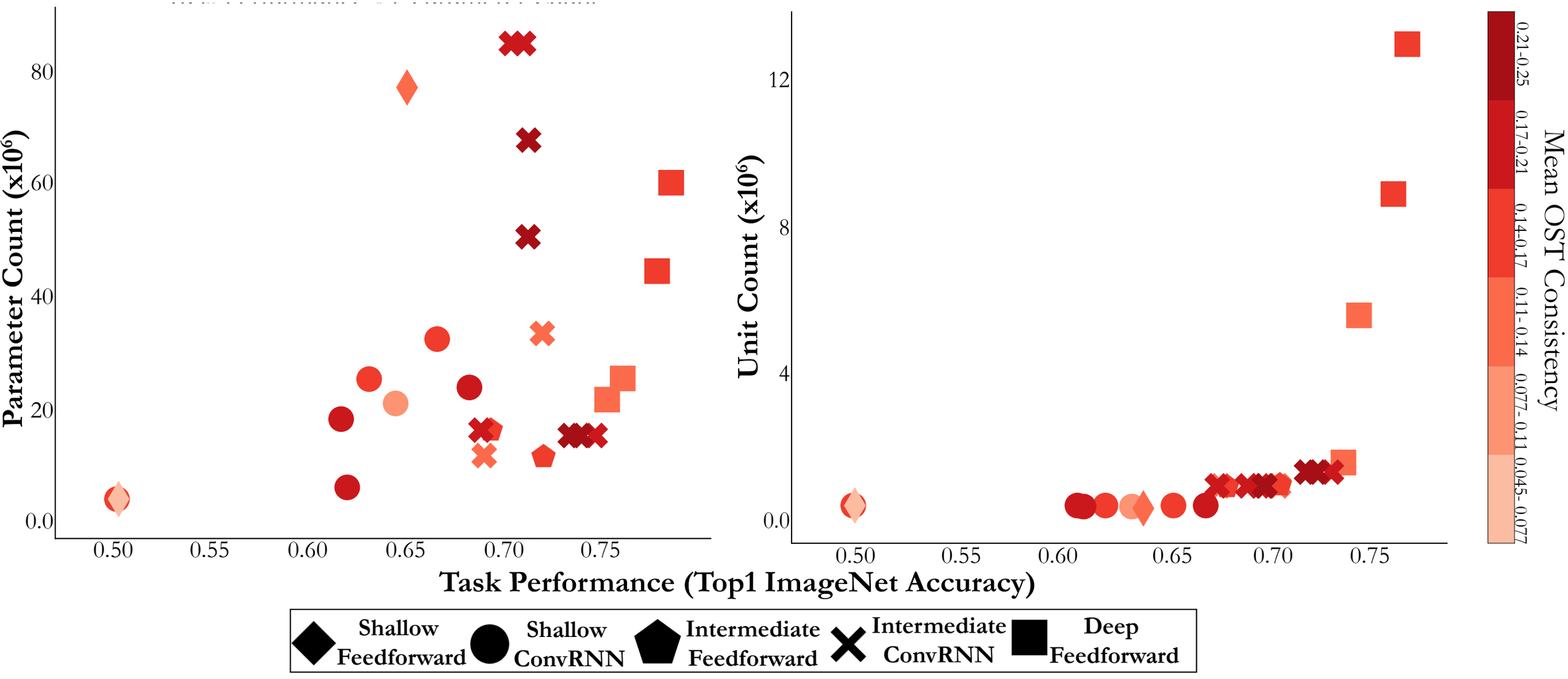


Intermediate ConvRNNs best match neural responses

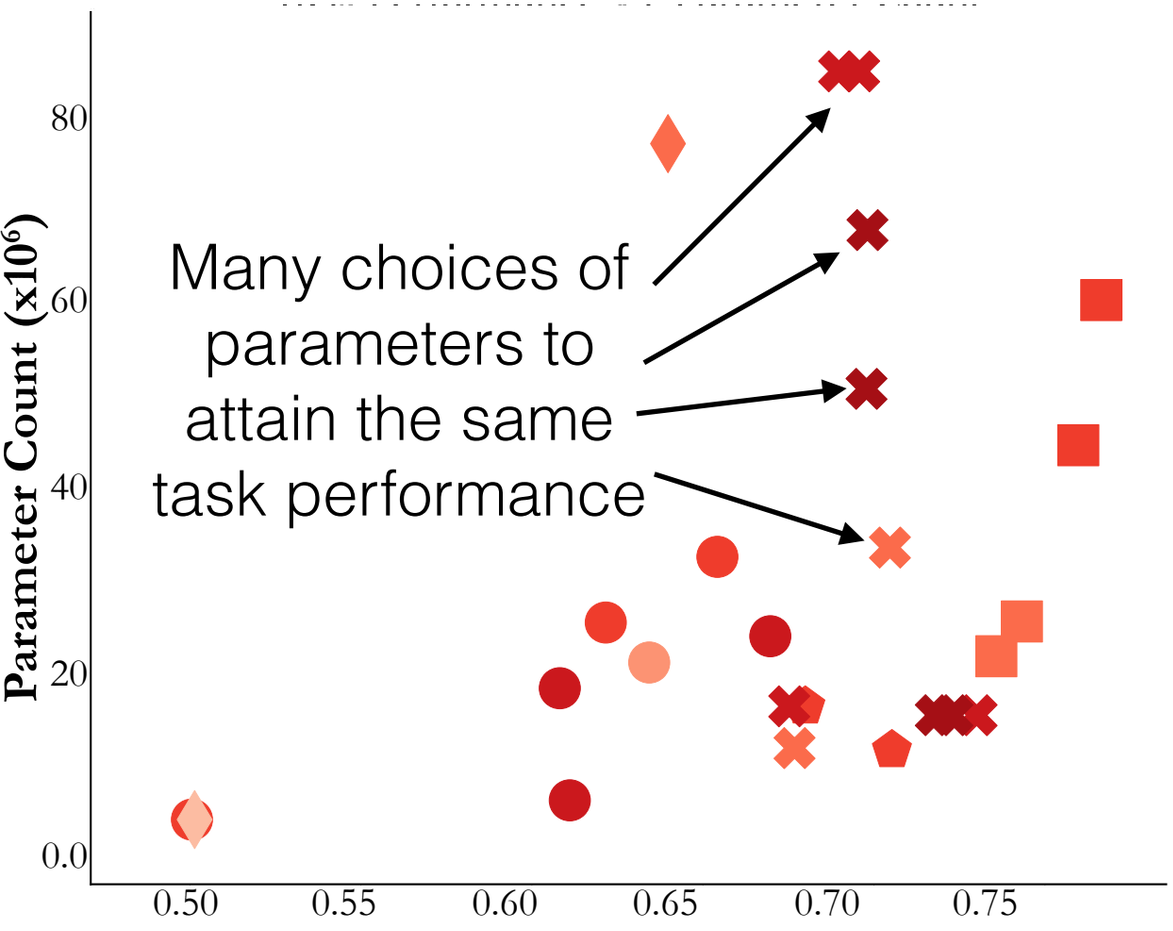


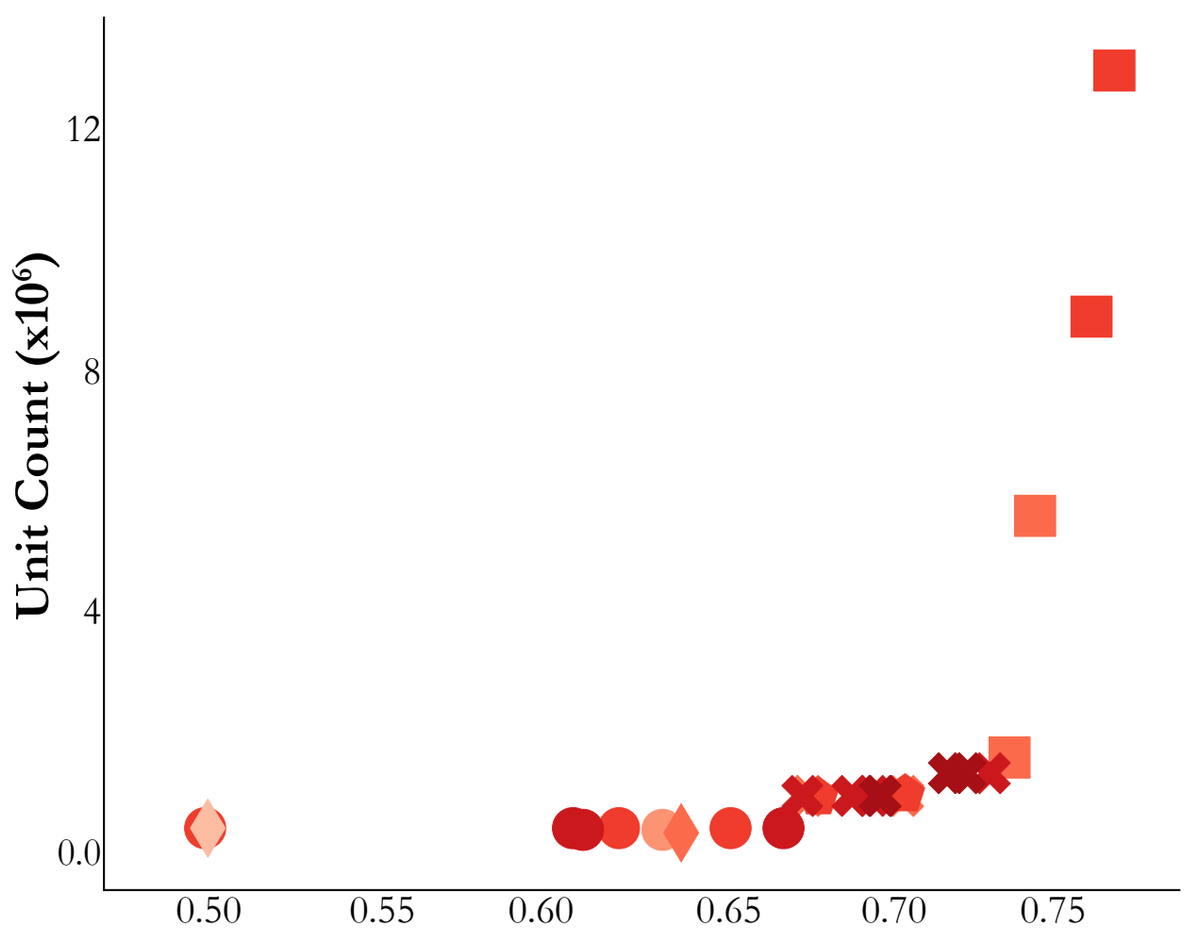


Conservation on network size + performance best matches OST



Conservation on network size + performance best matches OST

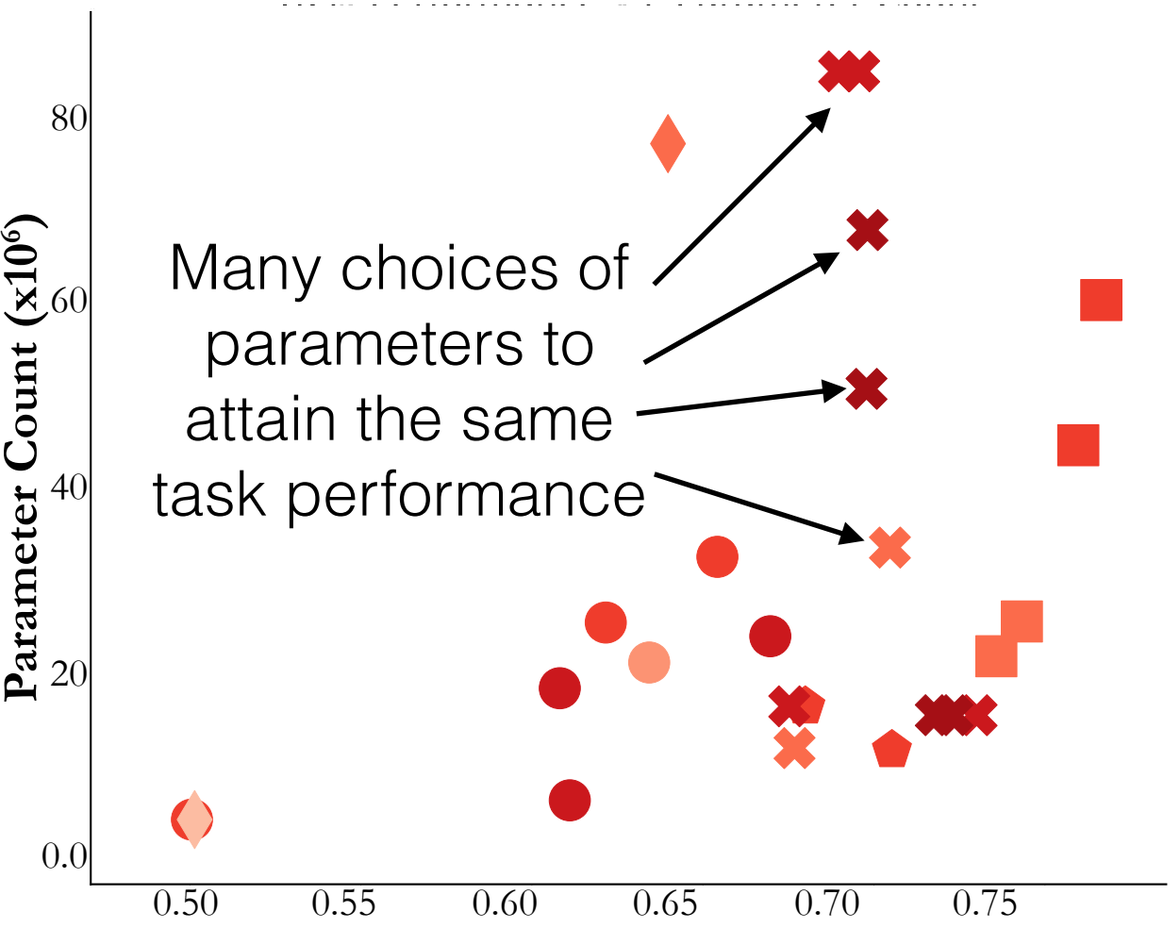


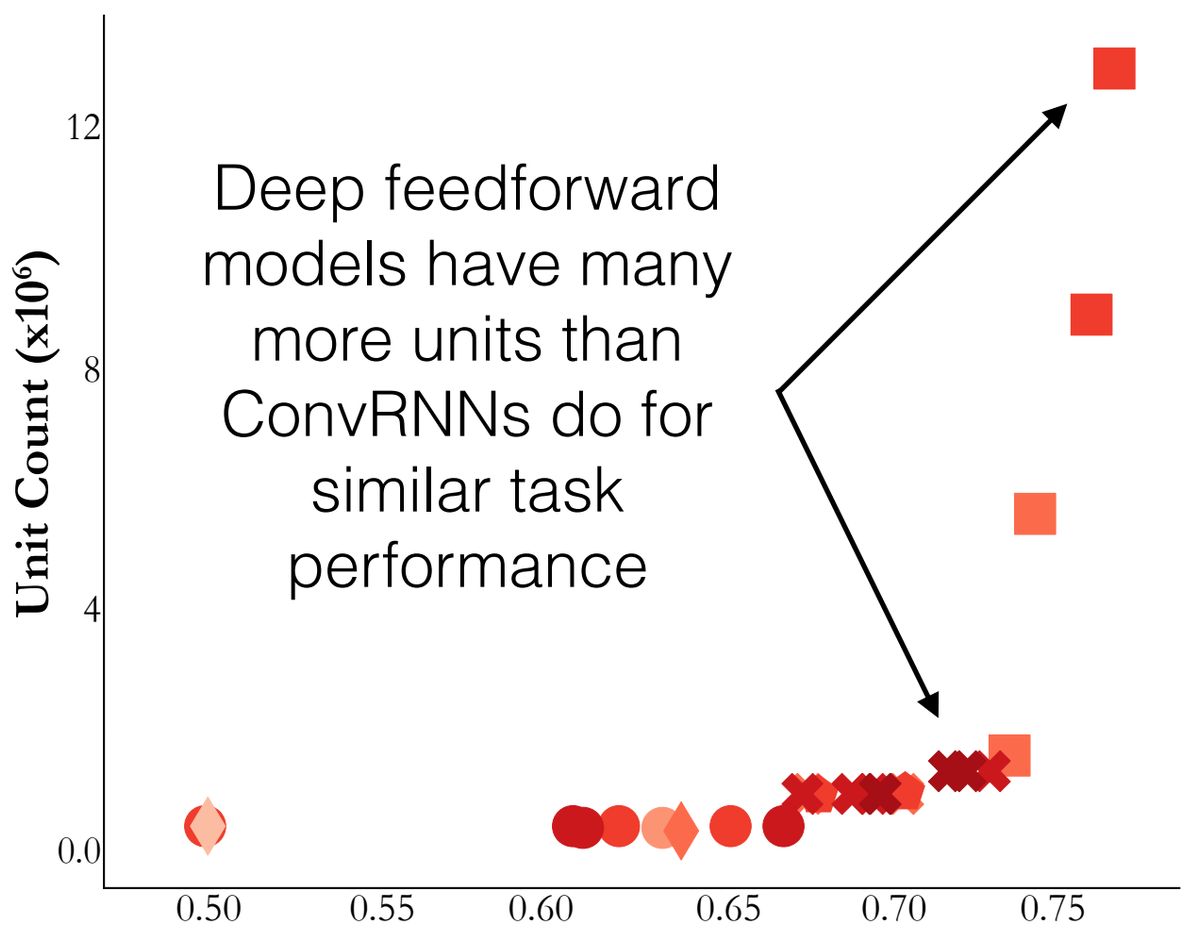


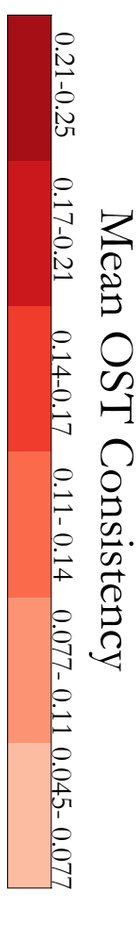
Task Performance (Top1 ImageNet Accuracy)

Shallow Feedforward | Shallow ConvRNN | Intermediate Feedforward | Intermediate ConvRNN | Deep Feedforward

Conservation on network size + performance best matches OST







Task Performance (Top1 ImageNet Accuracy)

Takeaways

- Recurrent motifs that incorporate specific mechanisms of "direct passthrough" and "gating" lead to matching the task performance of much deeper feedforward CNNs with fewer units and parameters.

Takeaways

- Recurrent motifs that incorporate specific mechanisms of "direct passthrough" and "gating" lead to matching the task performance of much deeper feedforward CNNs with fewer units and parameters.

- Unlike very deep feedforward CNNs, the mapping from the early, intermediate, and higher layers of these ConvRNNs to corresponding cortical areas is neuroanatomically consistent and reproduces prior quantitative properties of the ventral stream.

Takeaways

- Recurrent motifs that incorporate specific mechanisms of "direct passthrough" and "gating" lead to matching the task performance of much deeper feedforward CNNs with fewer units and parameters.

- Unlike very deep feedforward CNNs, the mapping from the early, intermediate, and higher layers of these ConvRNNs to corresponding cortical areas is neuroanatomically consistent and reproduces prior quantitative properties of the ventral stream.

- In fact, ConvRNNs with high task performance but small network size (as measured by number of neurons rather than synapses) are most consistent with the temporal evolution of primate IT object identity solutions.

Takeaways

- Recurrent motifs that incorporate specific mechanisms of "direct passthrough" and "gating" lead to matching the task performance of much deeper feedforward CNNs with fewer units and parameters.

- Unlike very deep feedforward CNNs, the mapping from the early, intermediate, and higher layers of these ConvRNNs to corresponding cortical areas is neuroanatomically consistent and reproduces prior quantitative properties of the ventral stream.

- In fact, ConvRNNs with high task performance but small network size (as measured by number of neurons rather than synapses) are most consistent with the temporal evolution of primate IT object identity solutions.

- Taken together, suggests that recurrence in the ventral stream extends feedforward computations by mediating a tradeoff between task performance and neuron count during core object recognition.

Outline

- Recurrent Connections in the Primate Ventral Stream
- Goal-Driven Models of Mouse Visual Cortex
- Heterogeneity in Rodent Medial Entorhinal Cortex
- Building and Identifying Learning Rules

Goal-Driven Modeling - Three Primary Components

"circuit"

1.

A = *architecture class*

"behavior"

2.

L = *loss function* ***D*** = *dataset*

A. Nayebi*, N. Kong* *et al.*
Shallow Unsupervised Models Best Predict Neural Responses in Mouse Visual Cortex.
bioRxiv (2021)

"plasticity mechanism"

3.

Learning Rule

$$\underset{a \in \mathcal{A}}{\operatorname{argmin}}[L(p_a^*)]$$

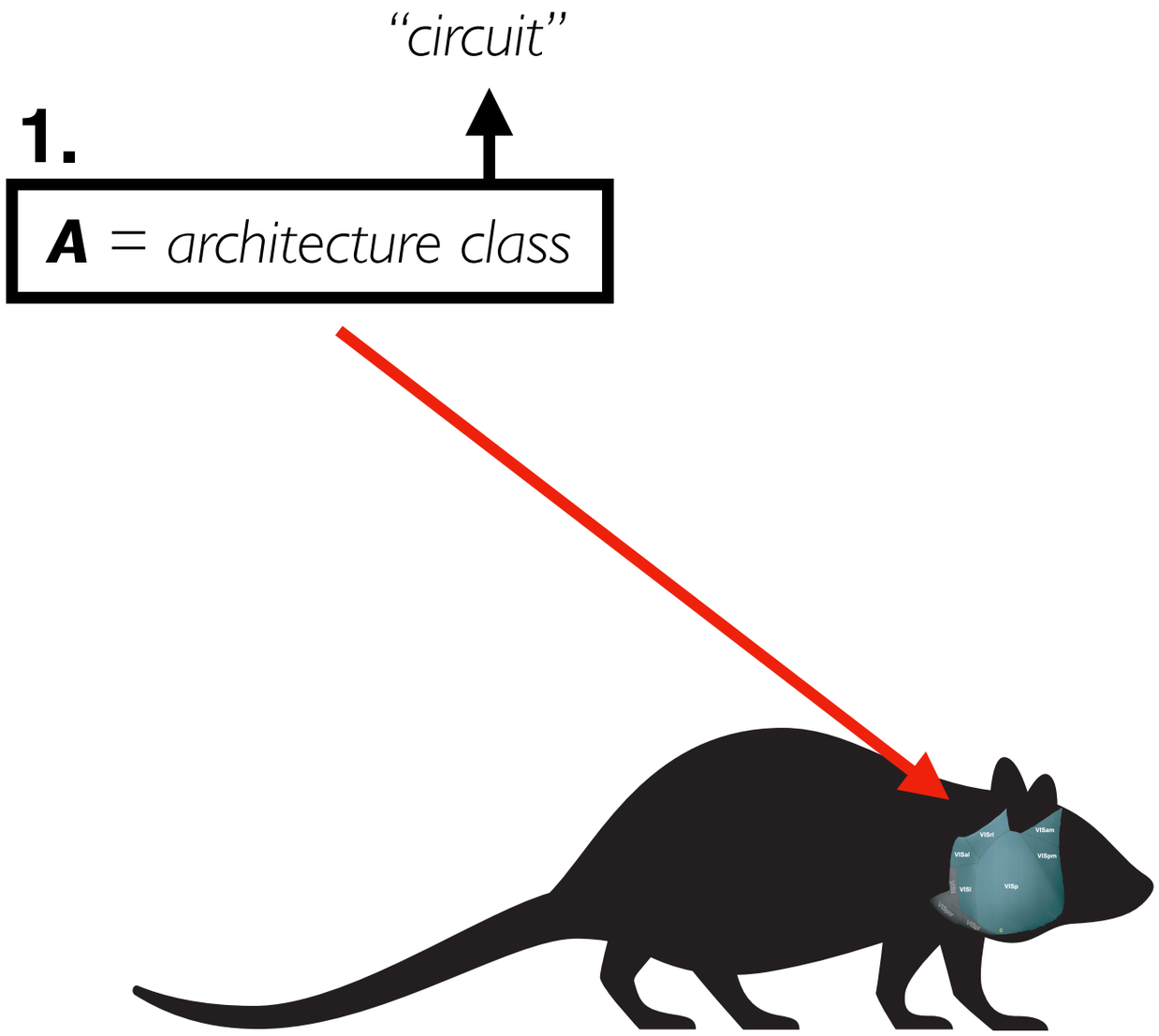
genetic algorithms

where p is result of*

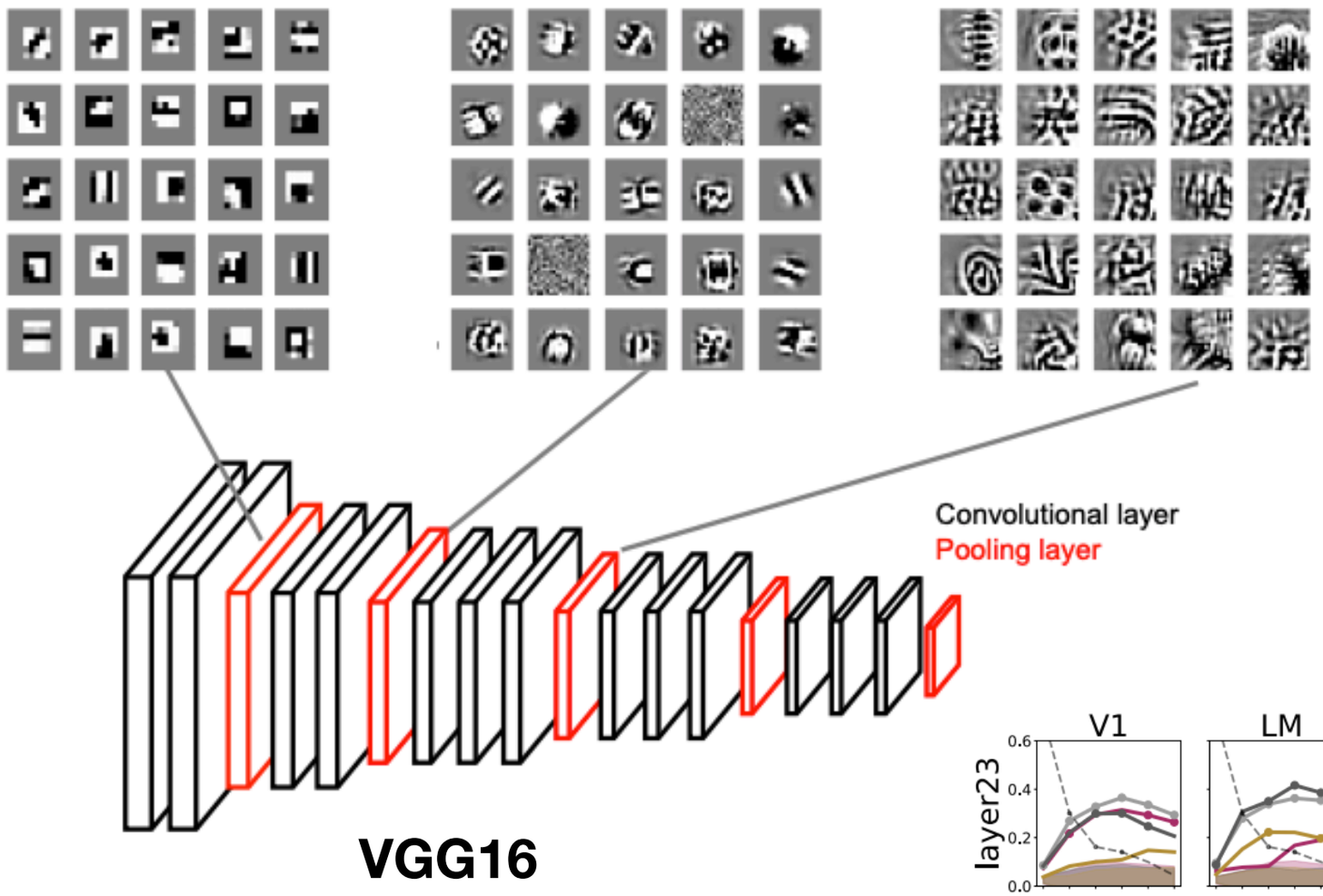
backprop

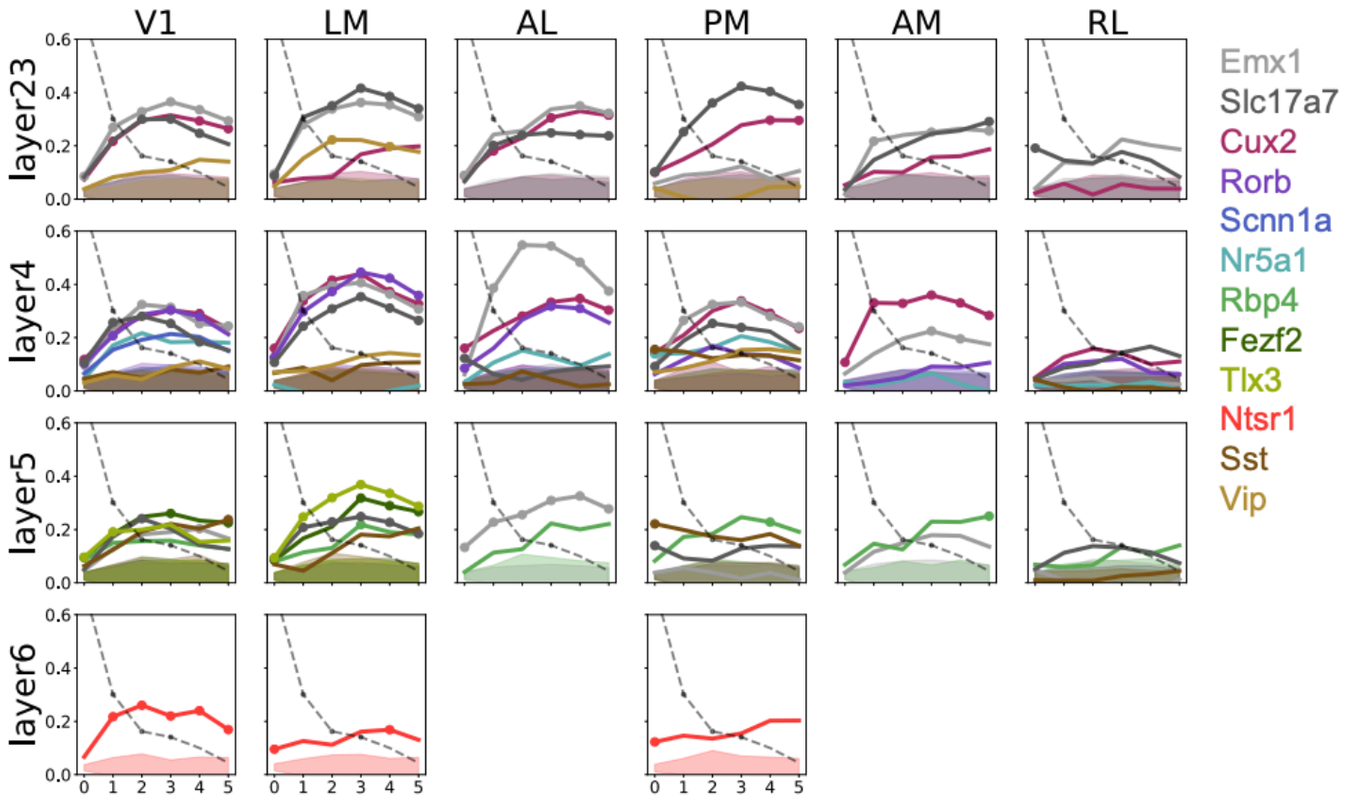
$$\frac{dp_a}{dt} = -\lambda(t) \cdot \langle \nabla_{p_a} L(x) \rangle_{x \in \mathcal{D}}$$

Distilling Constraints: Circuit Architecture



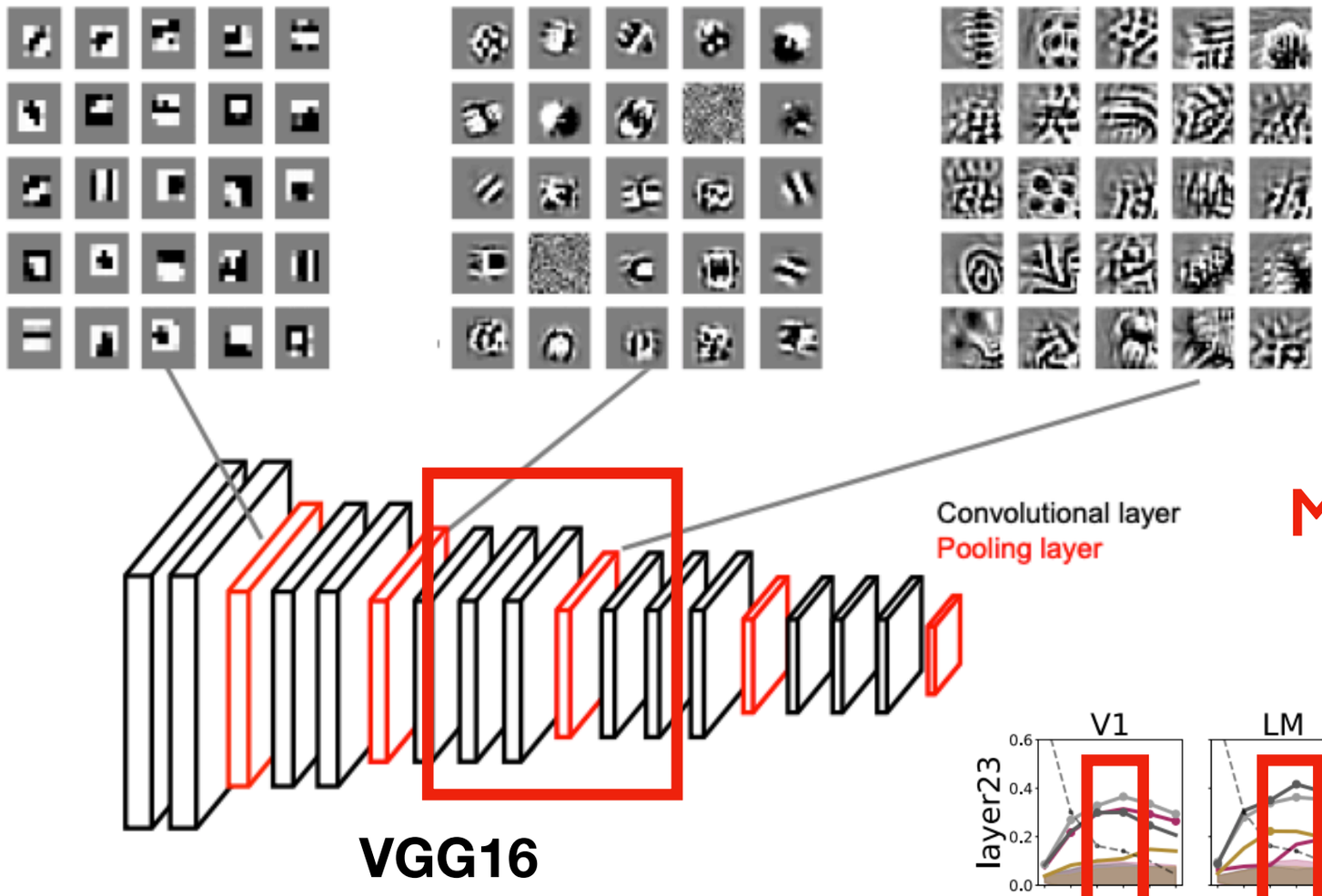
Deep models suggest mouse vision is representationally deep





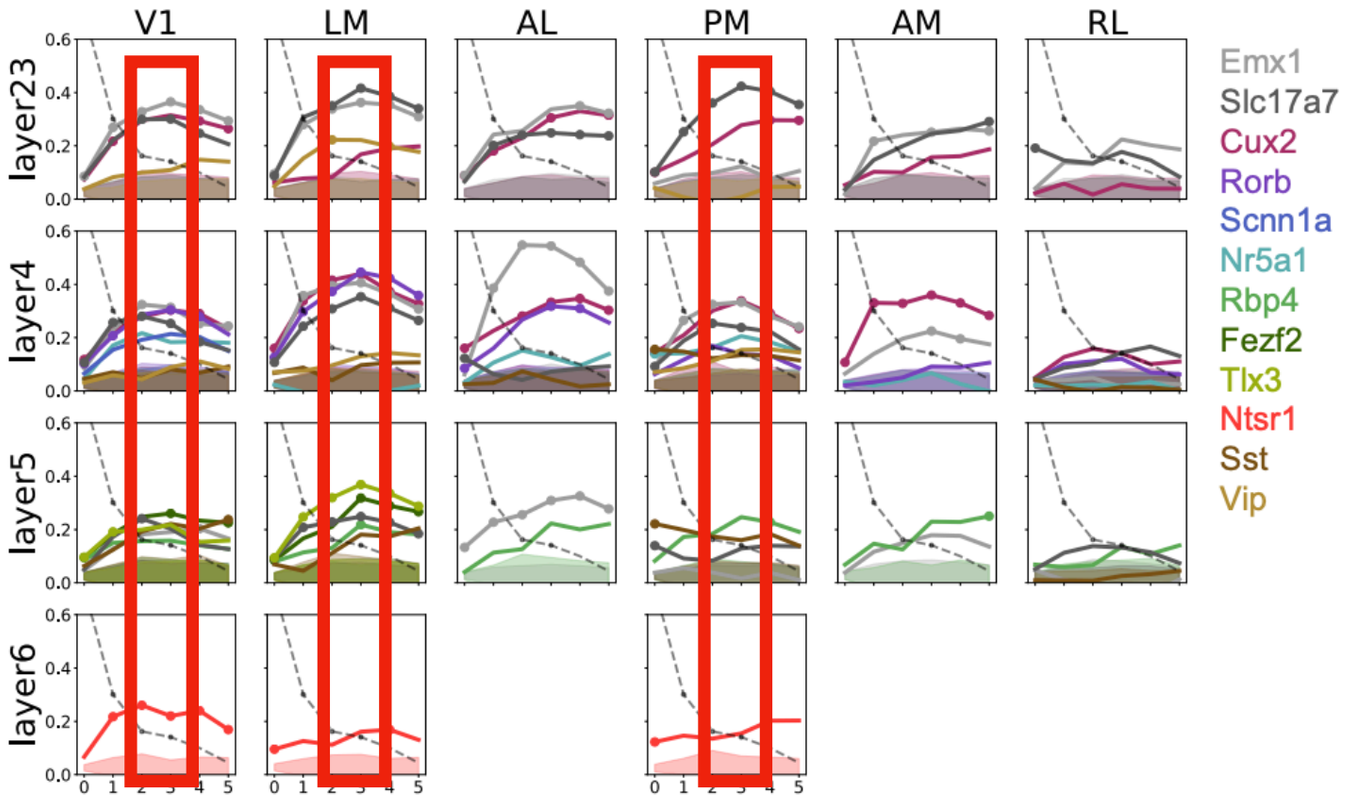
de Vries et al. 2020

Deep models suggest mouse vision is representationally deep



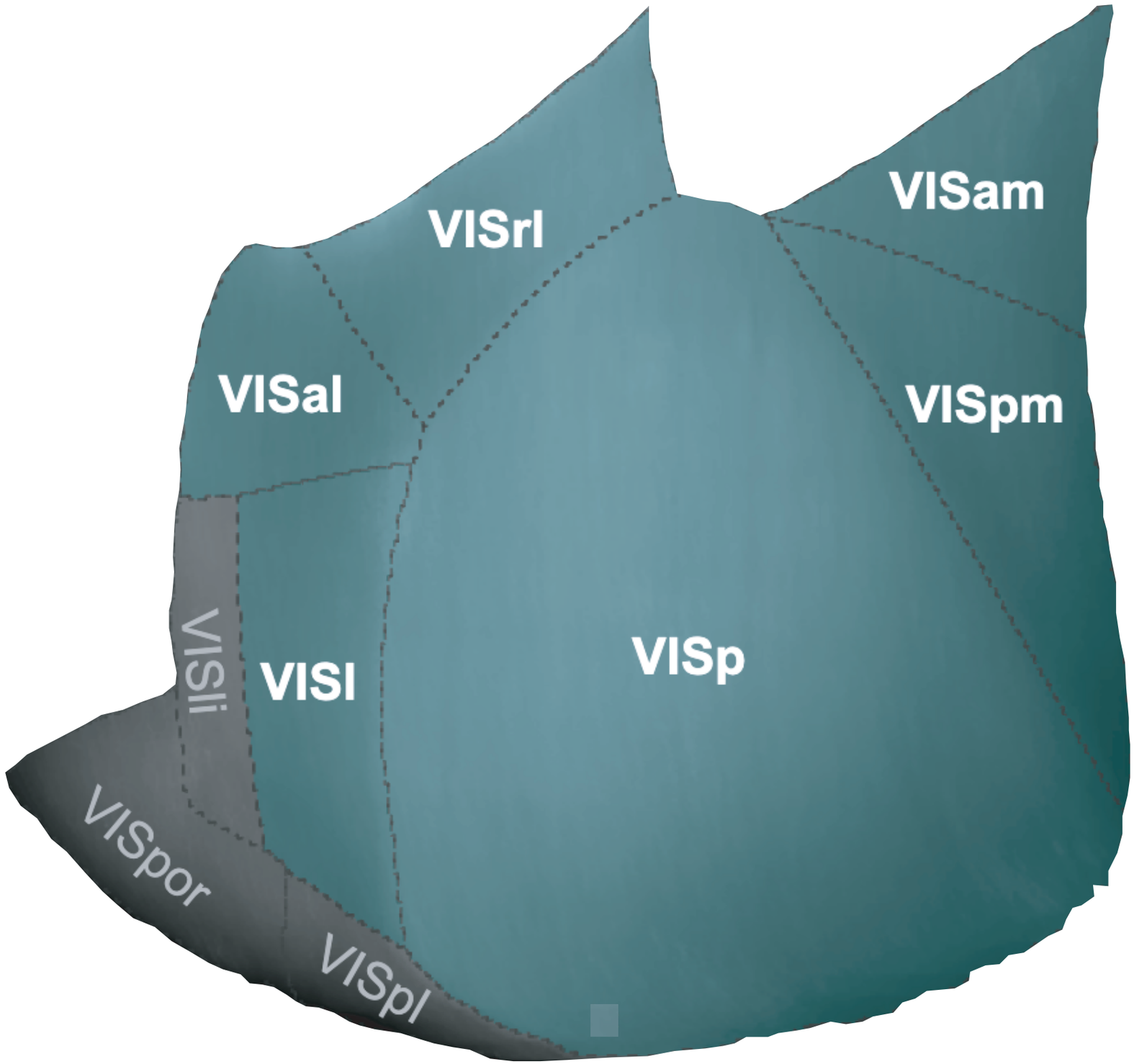
VGG16

Mid-levels best predict visual areas



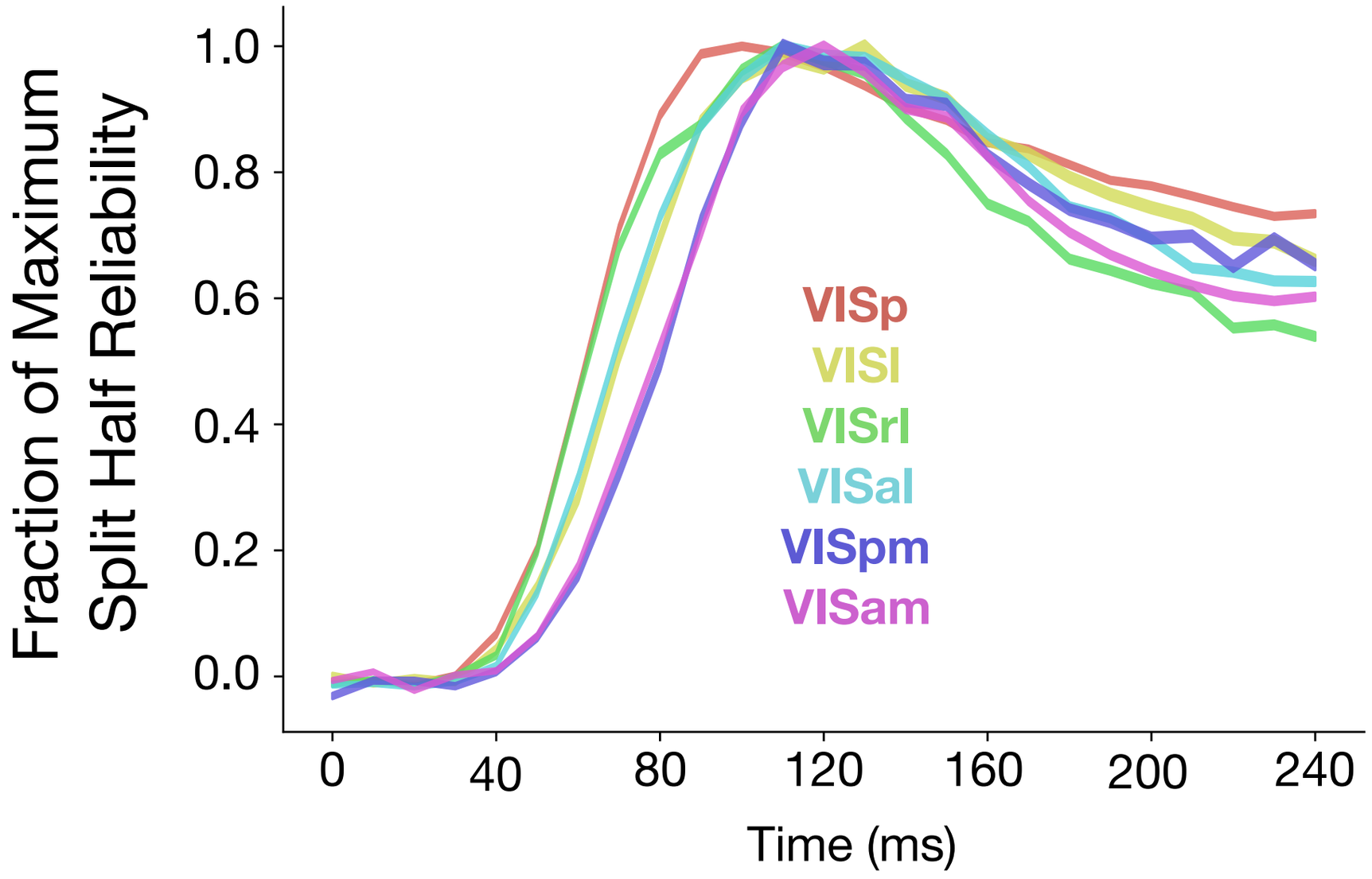
de Vries et al. 2020

But mouse visual cortex is anatomically shallow!

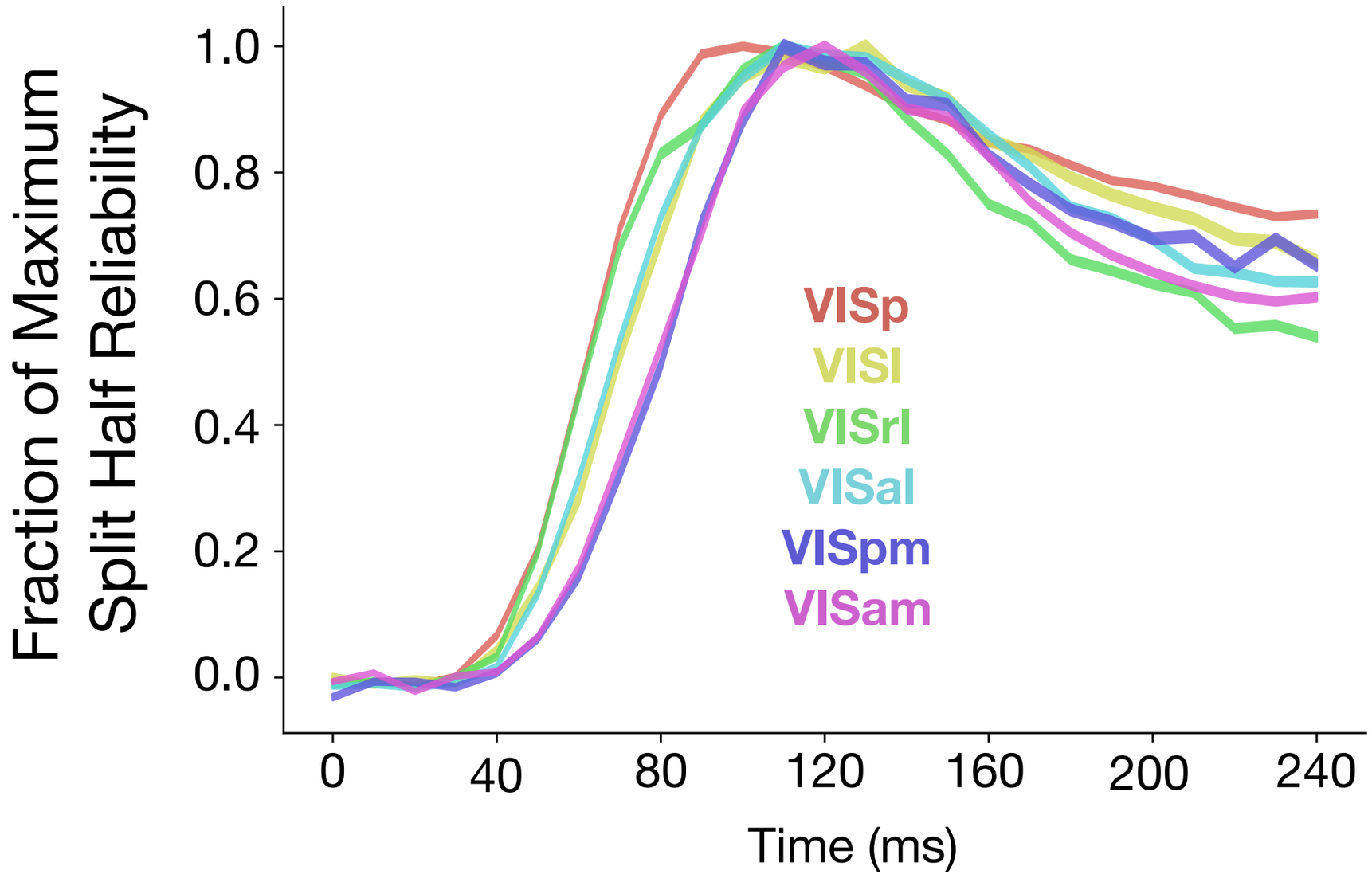


Allen Mouse Brain Atlas

But mouse visual cortex is anatomically shallow!

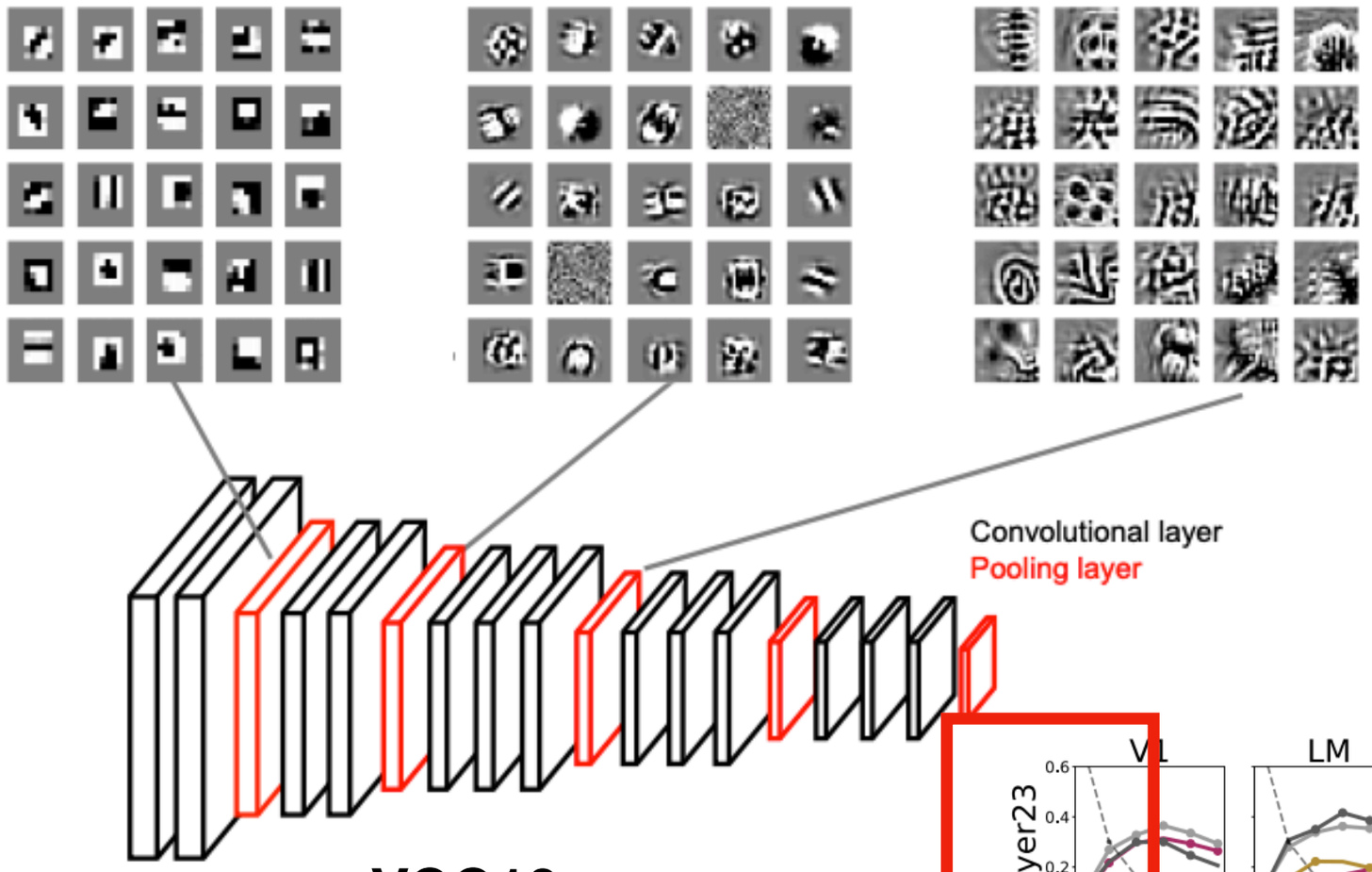


But mouse visual cortex is anatomically shallow!



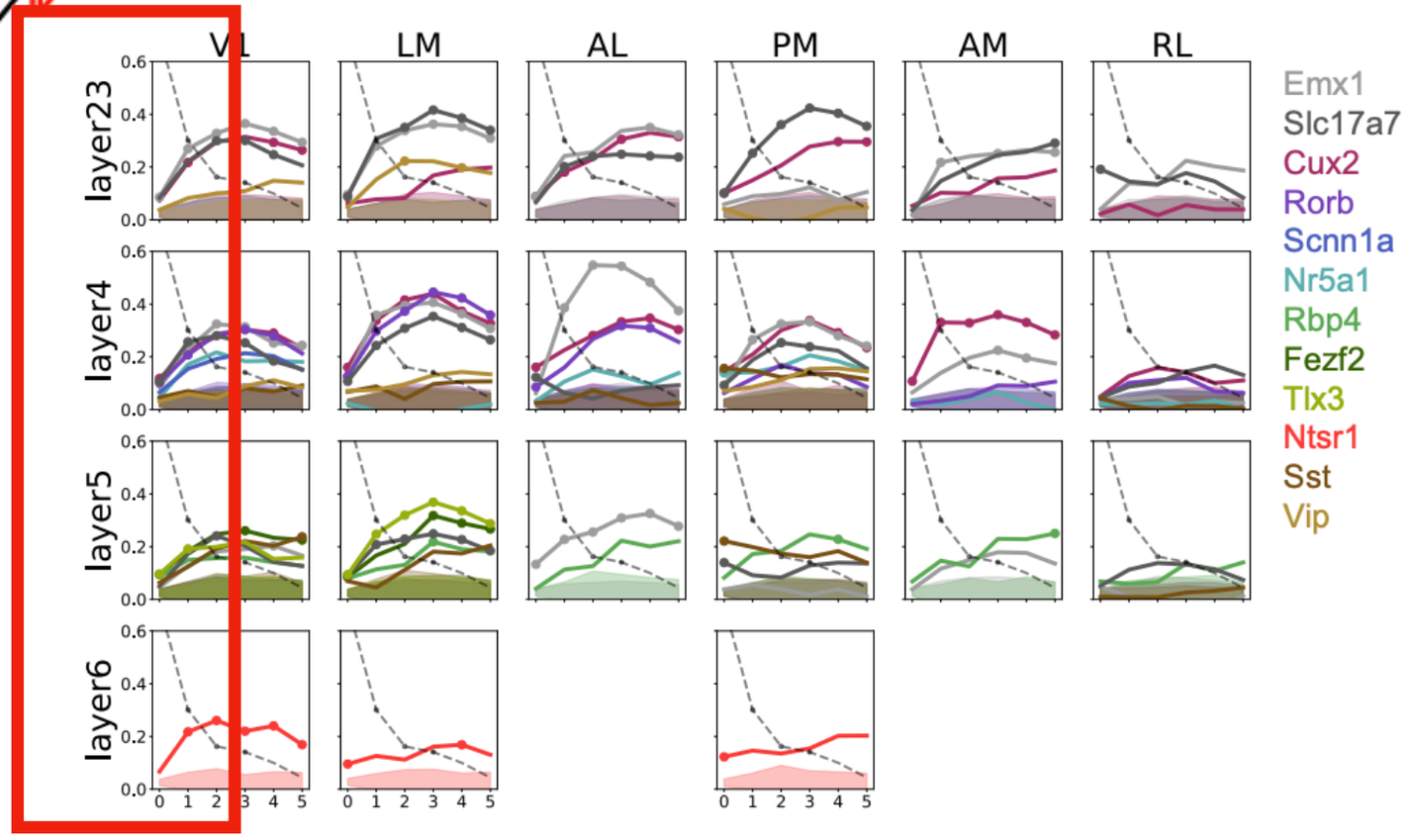
Suggests roughly 3-4 stages of processing based on reliability timing

Deep models are also a poor match to responses



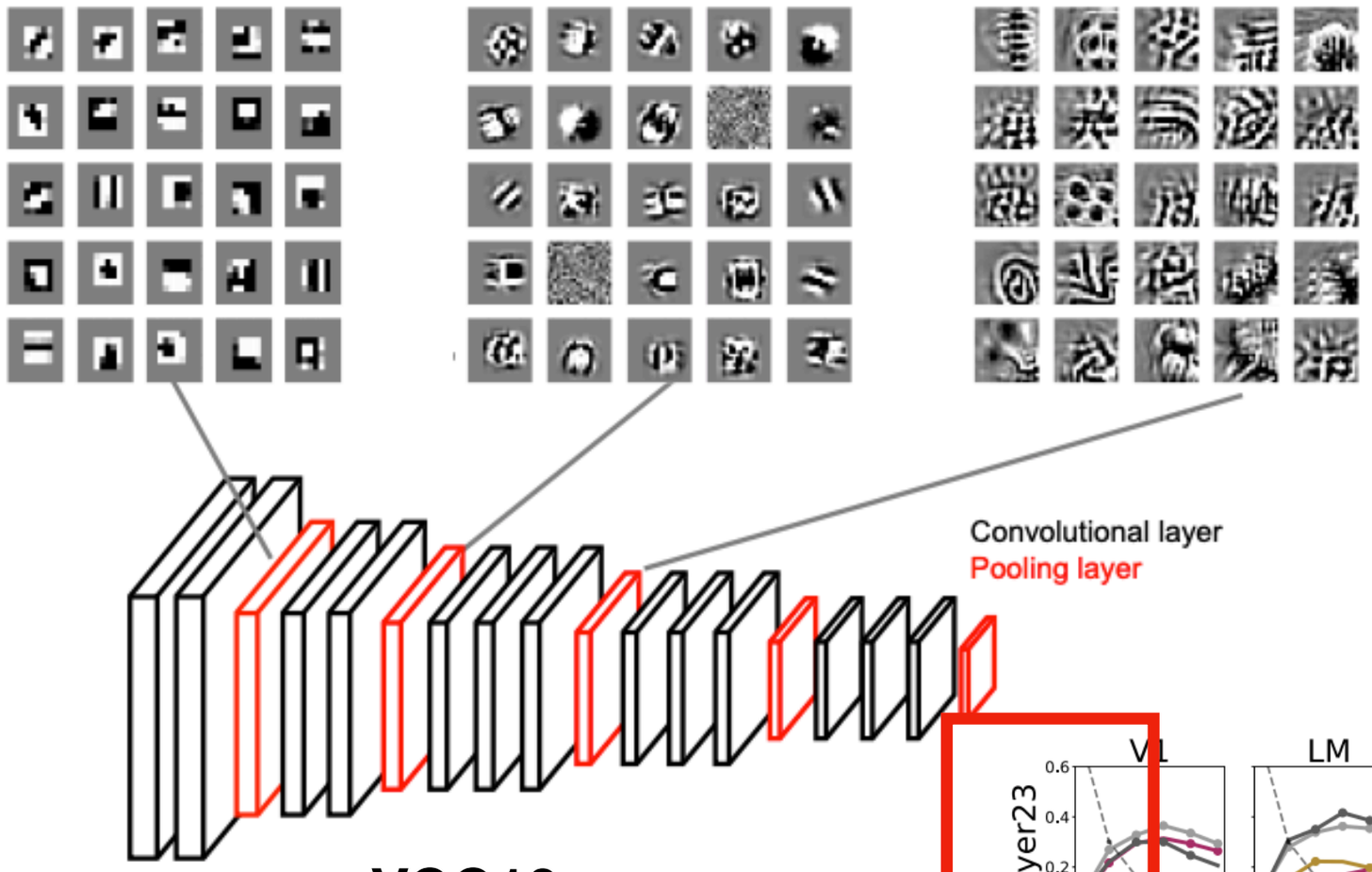
VGG16

Low predictivity between models layers and mouse visual areas!



de Vries et al., 2020

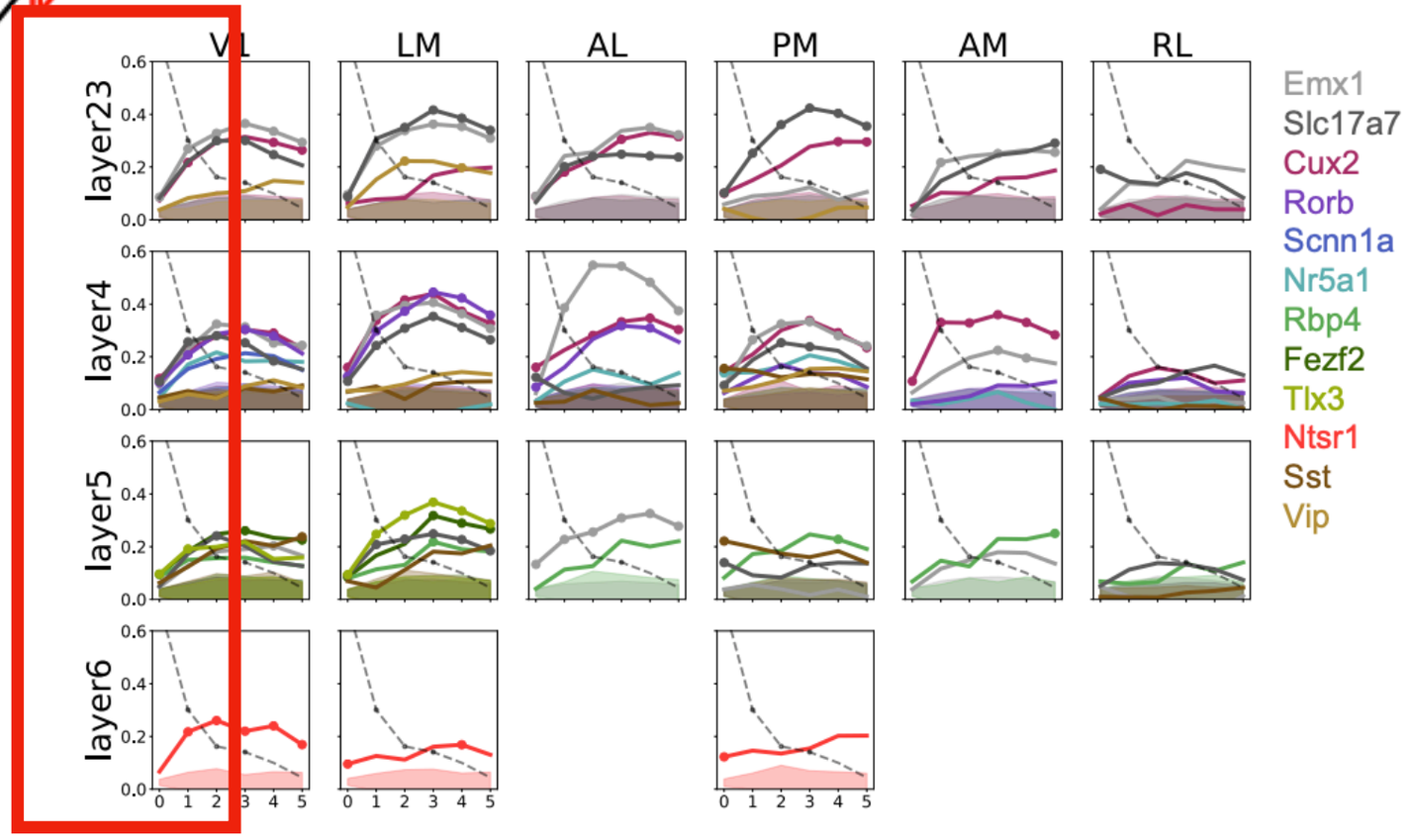
Deep models are also a poor match to responses



VGG16

Low predictivity between models layers and mouse visual areas!

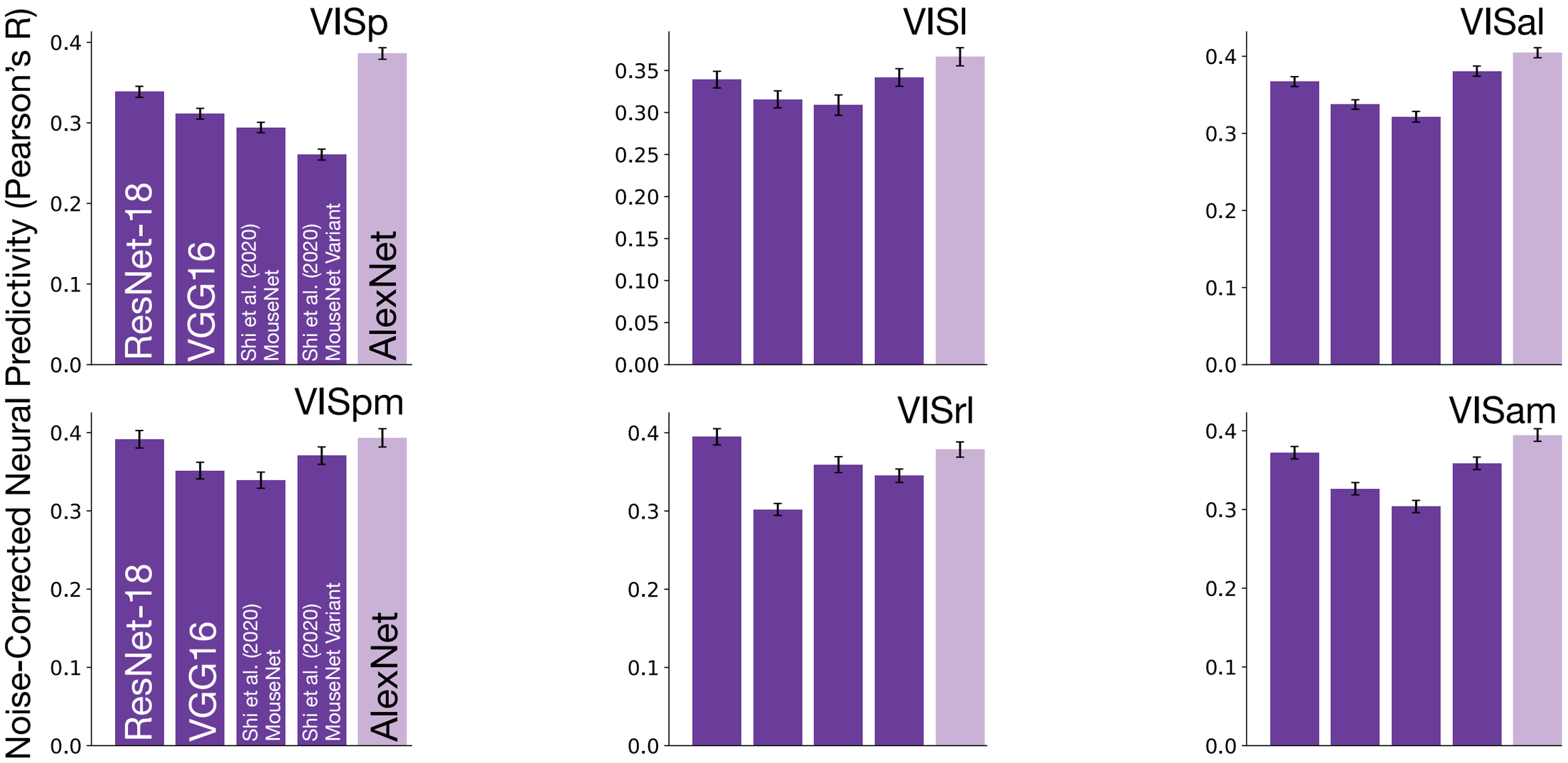
Can we improve this correspondence with a shallower model?



de Vries et al., 2020

Shallow architectures better predict mouse visual responses than deep architectures

Deeper Models **Shallower Models**

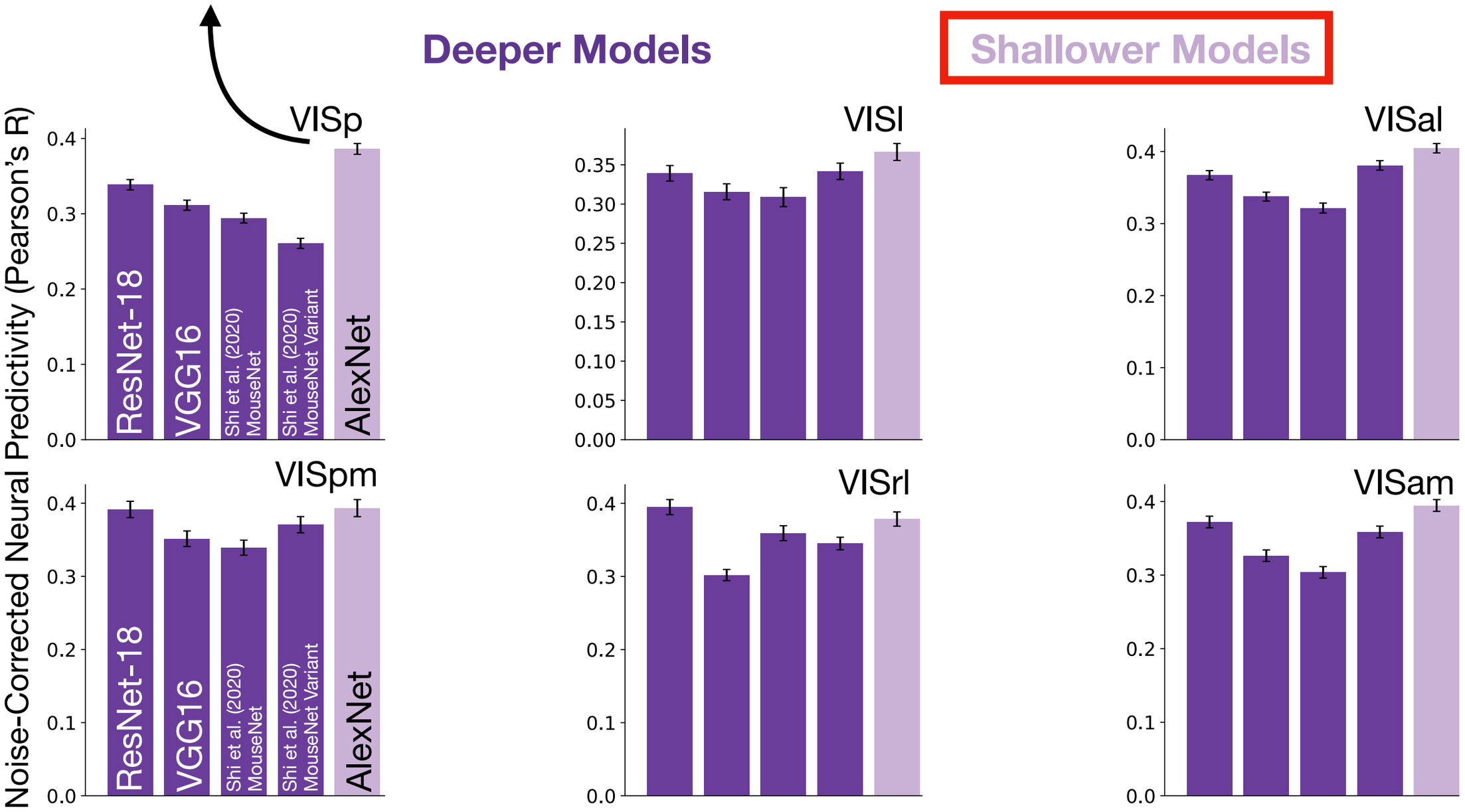


Shallow architectures better predict mouse visual responses than deep architectures

AlexNet is best among CNNs typically used for modeling primate vision

Deeper Models

Shallower Models

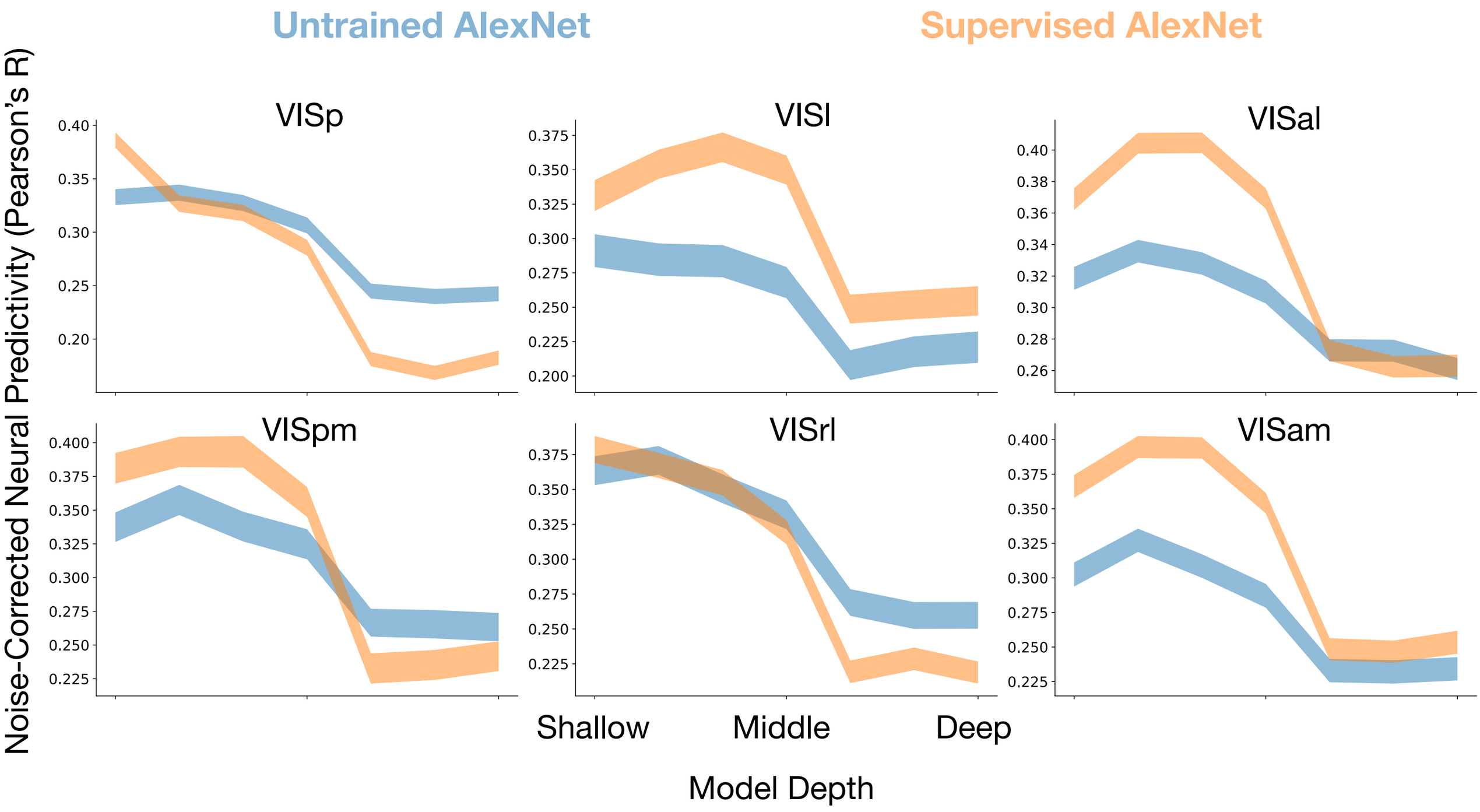


Even shallower models?

Look at neural predictivity across model layer

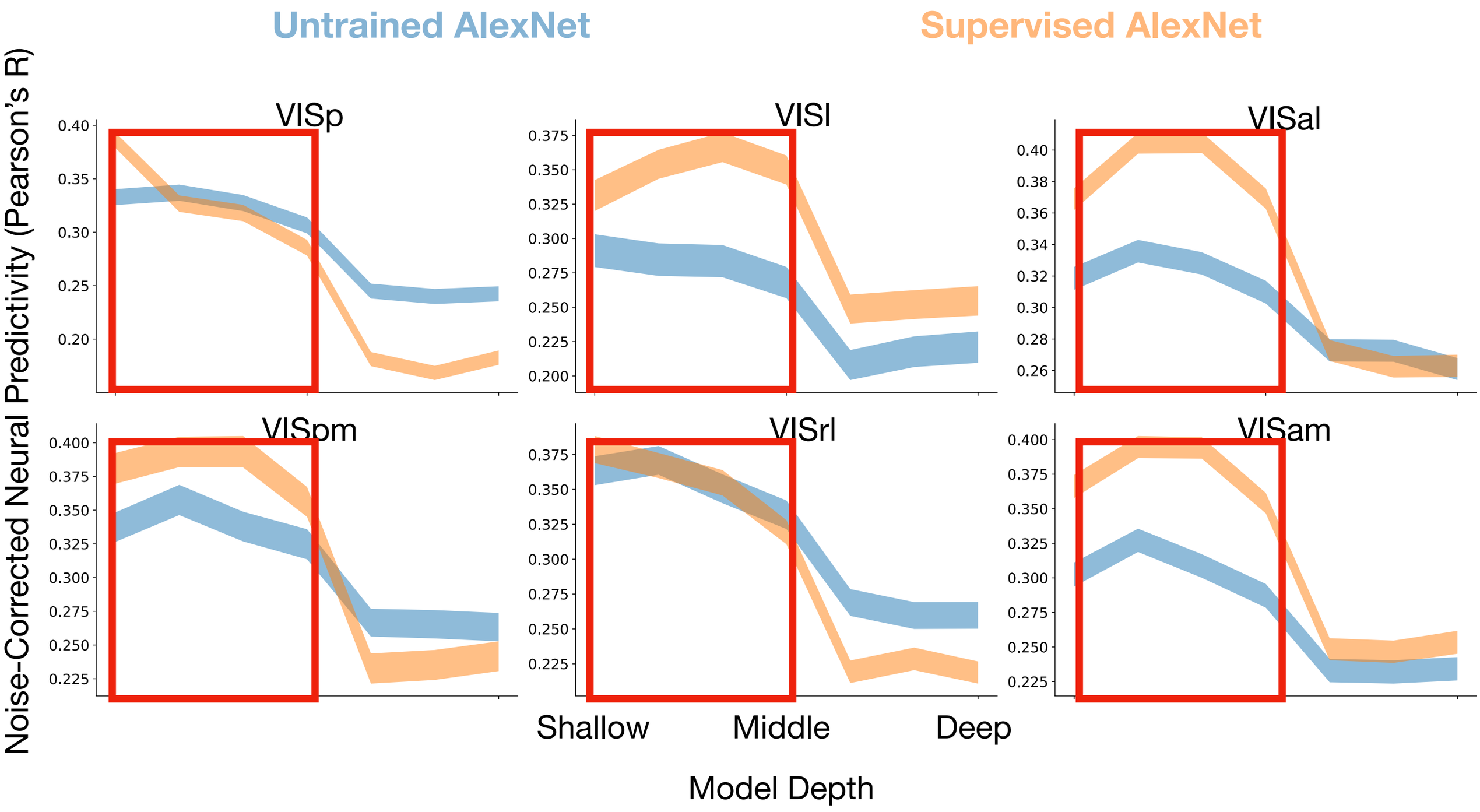
Even shallower models?

Look at neural predictivity across model layer

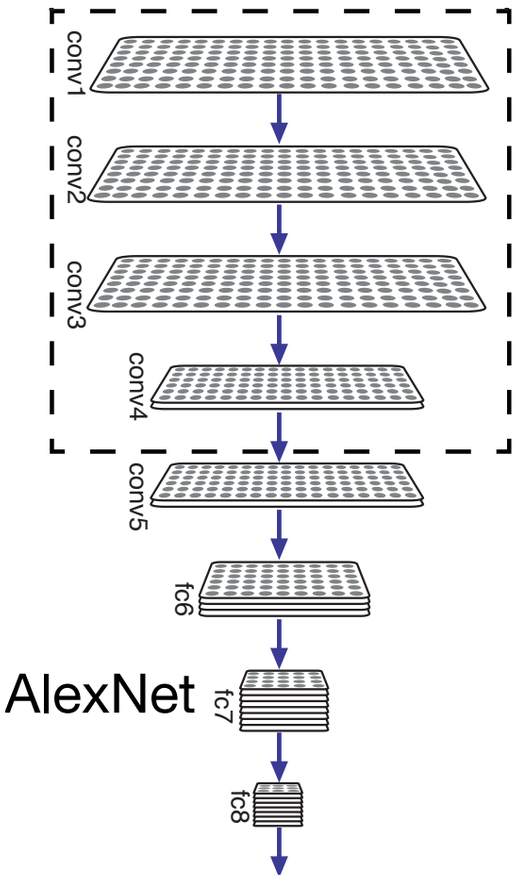


Even shallower models?

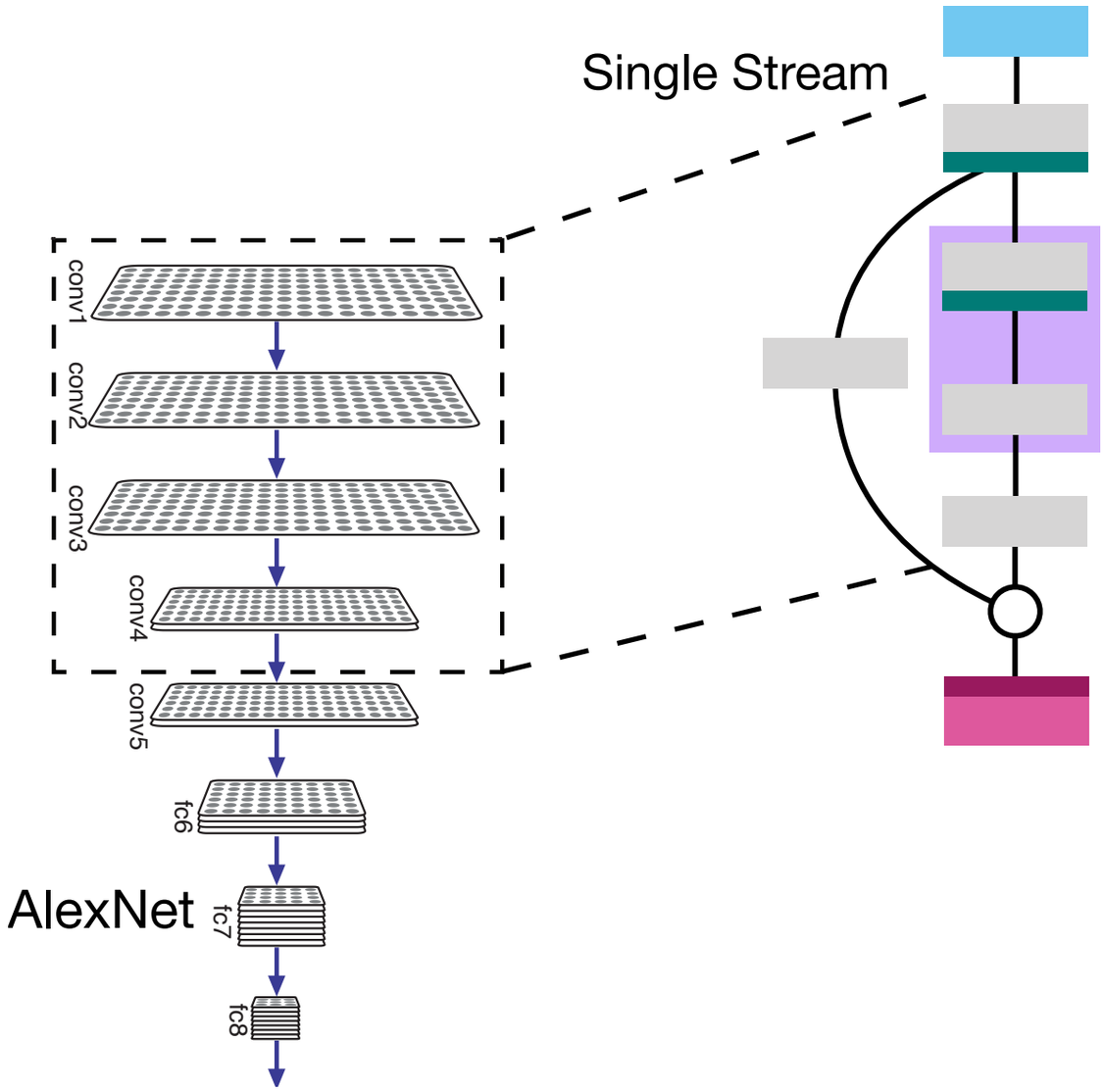
Look at neural predictivity across model layer



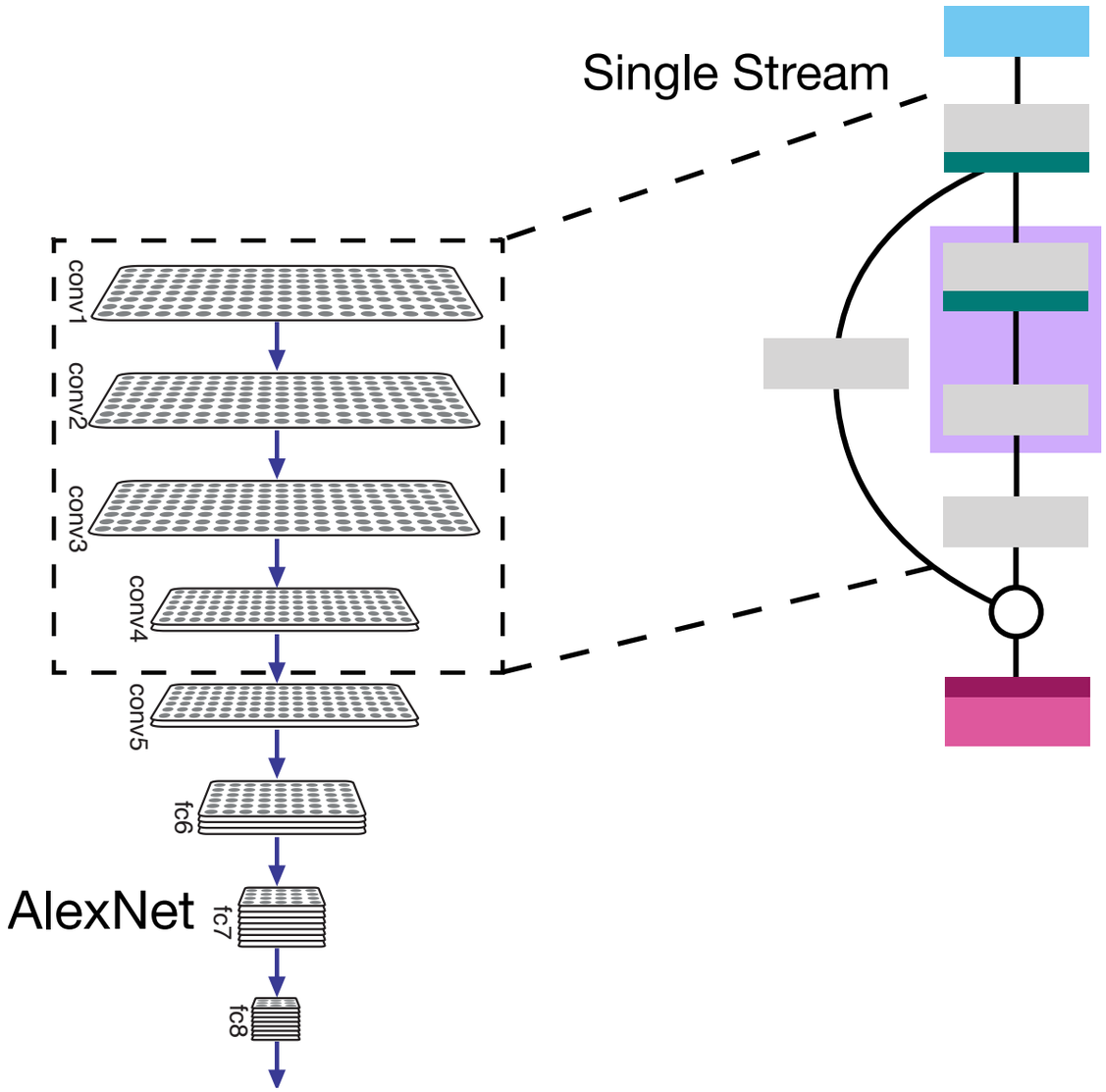
Even shallower models?

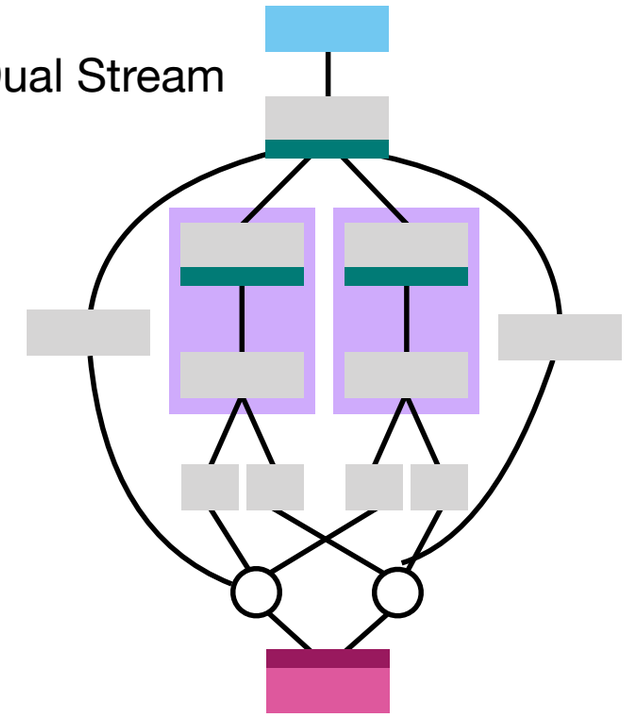


Even shallower models?

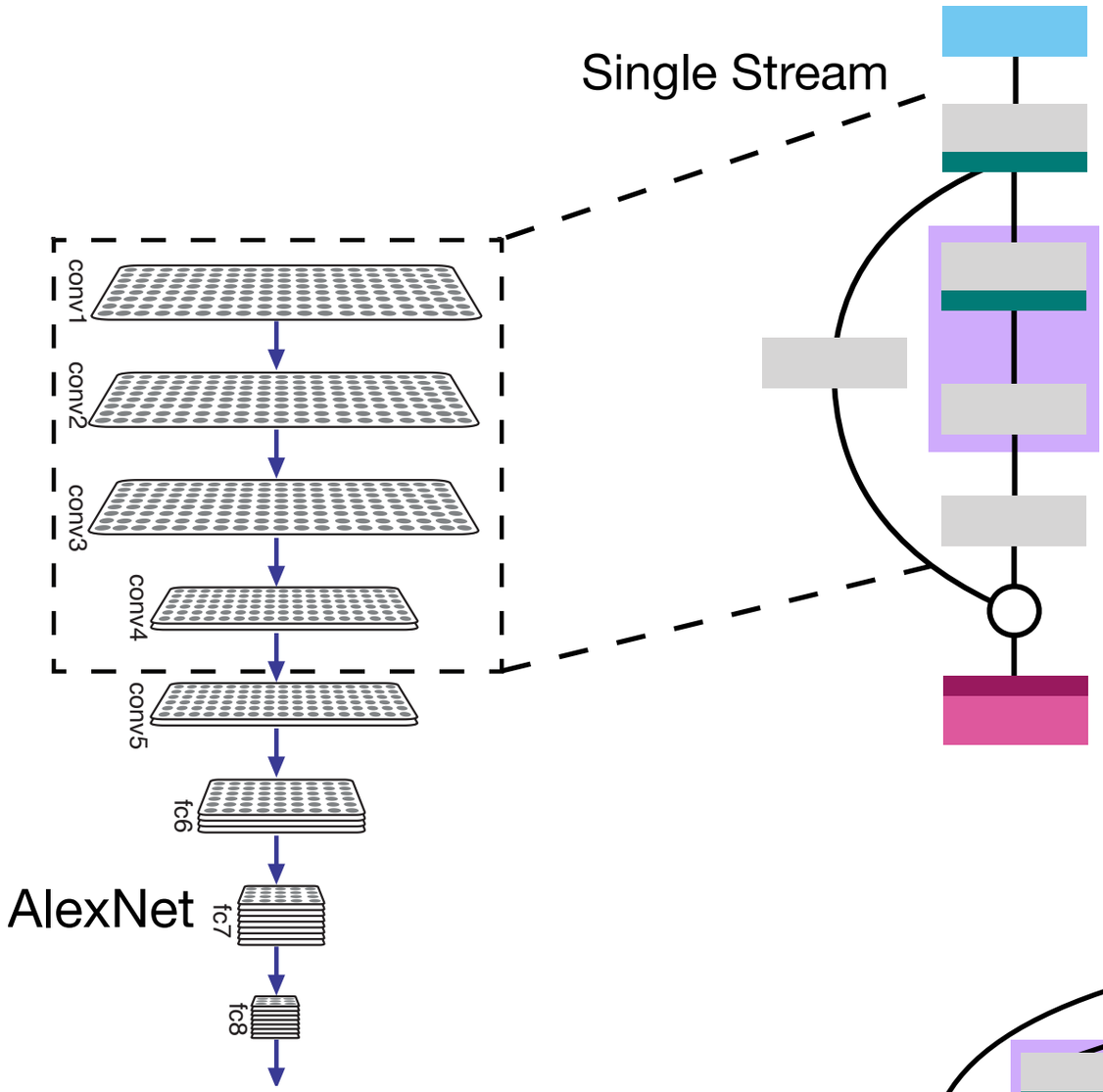


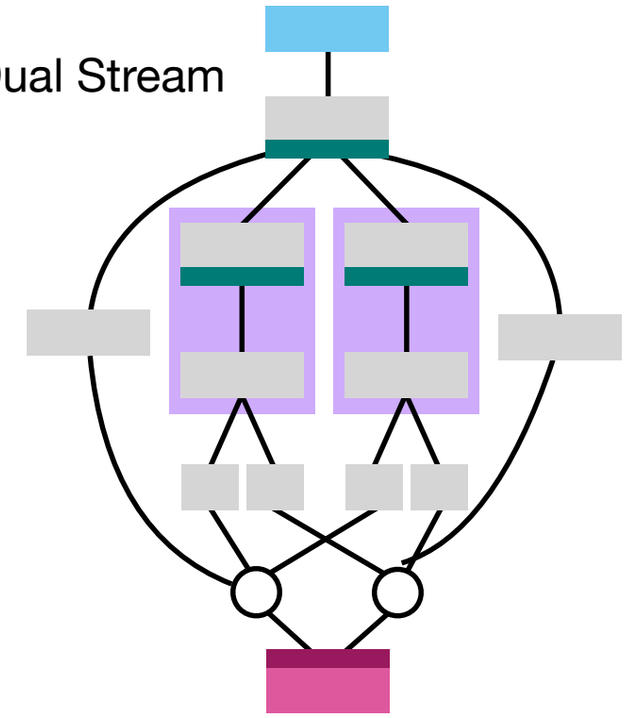
Even shallower models?

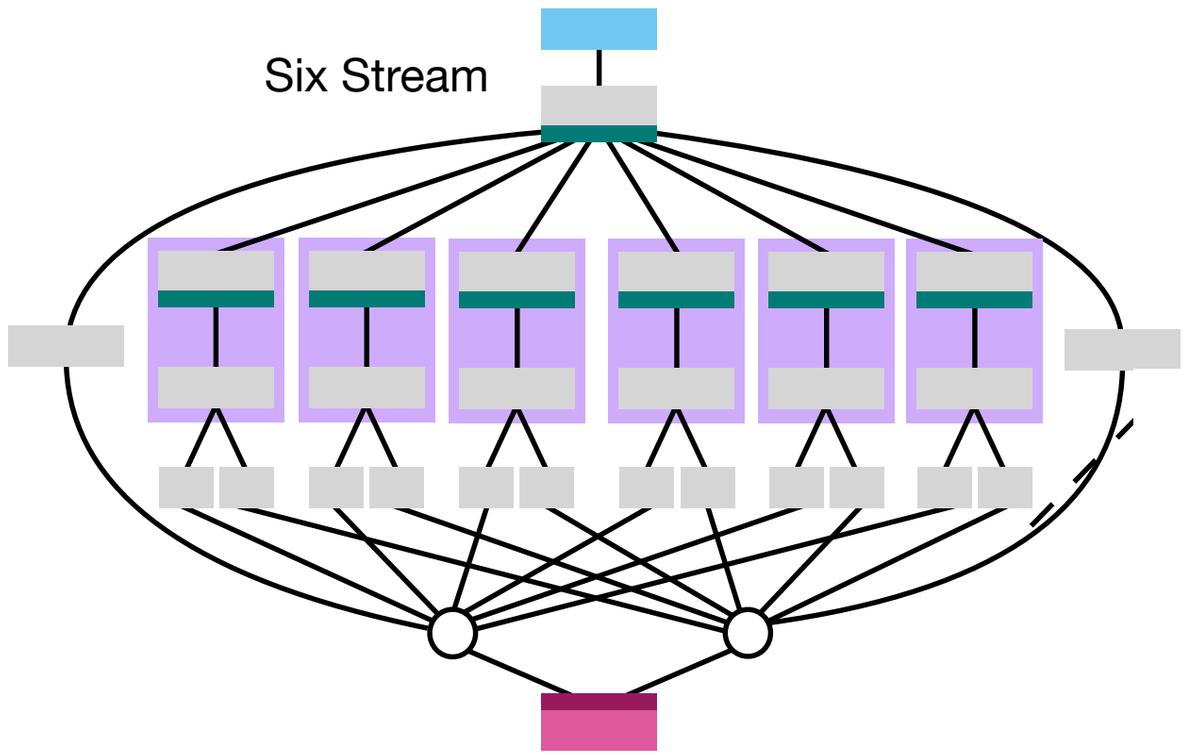




Even shallower models?

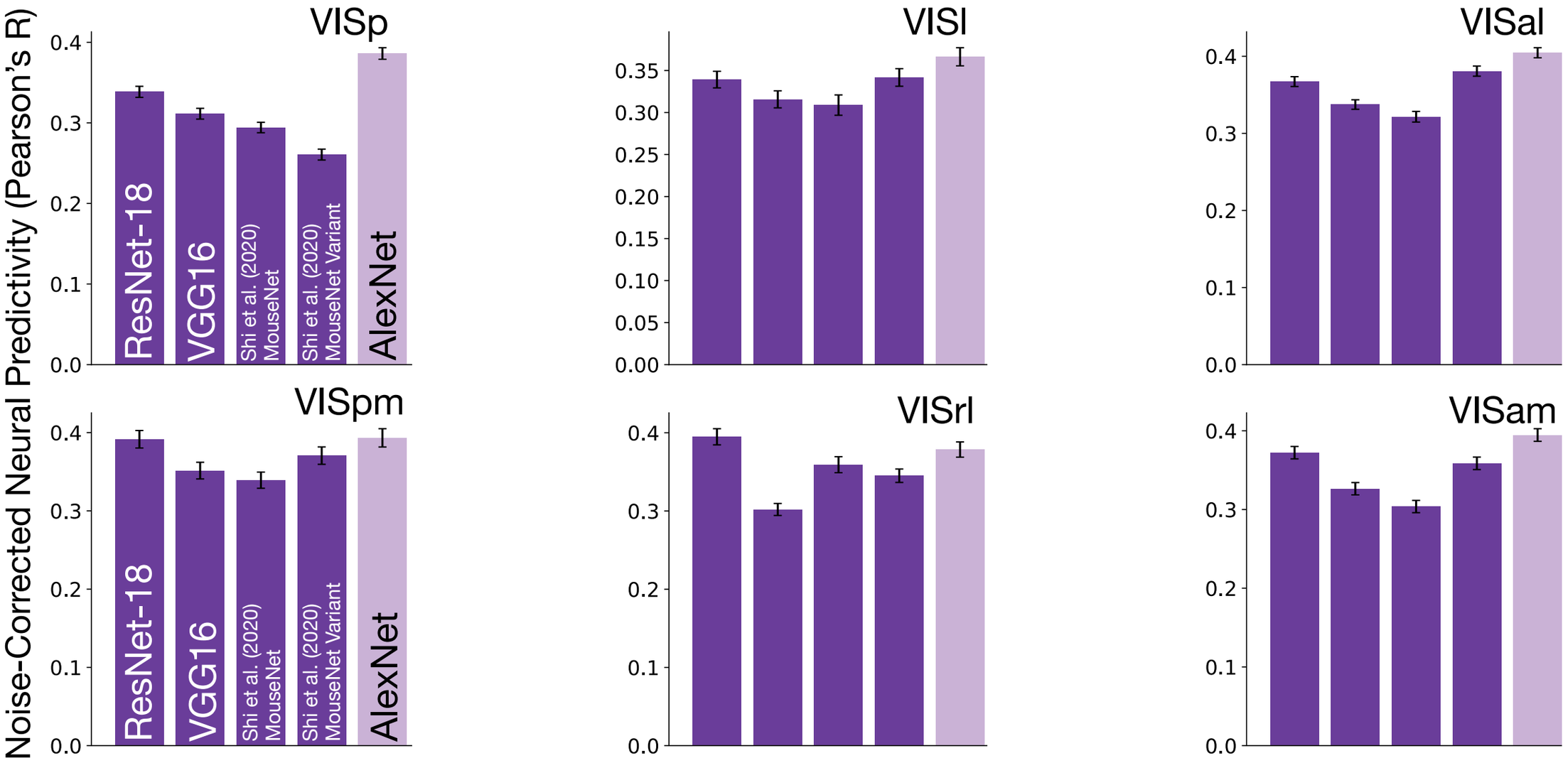




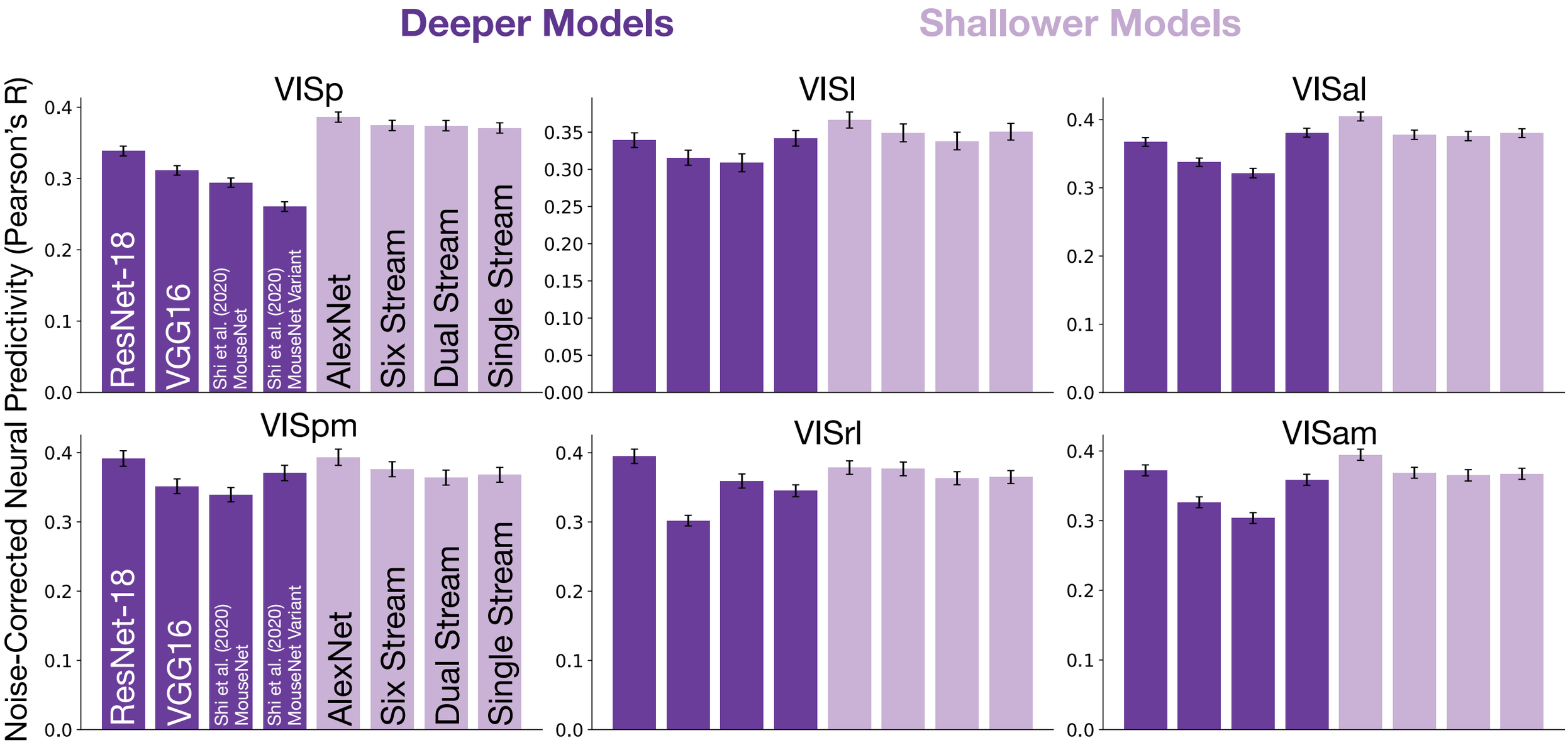


Shallow architectures better predict mouse visual responses than deep architectures

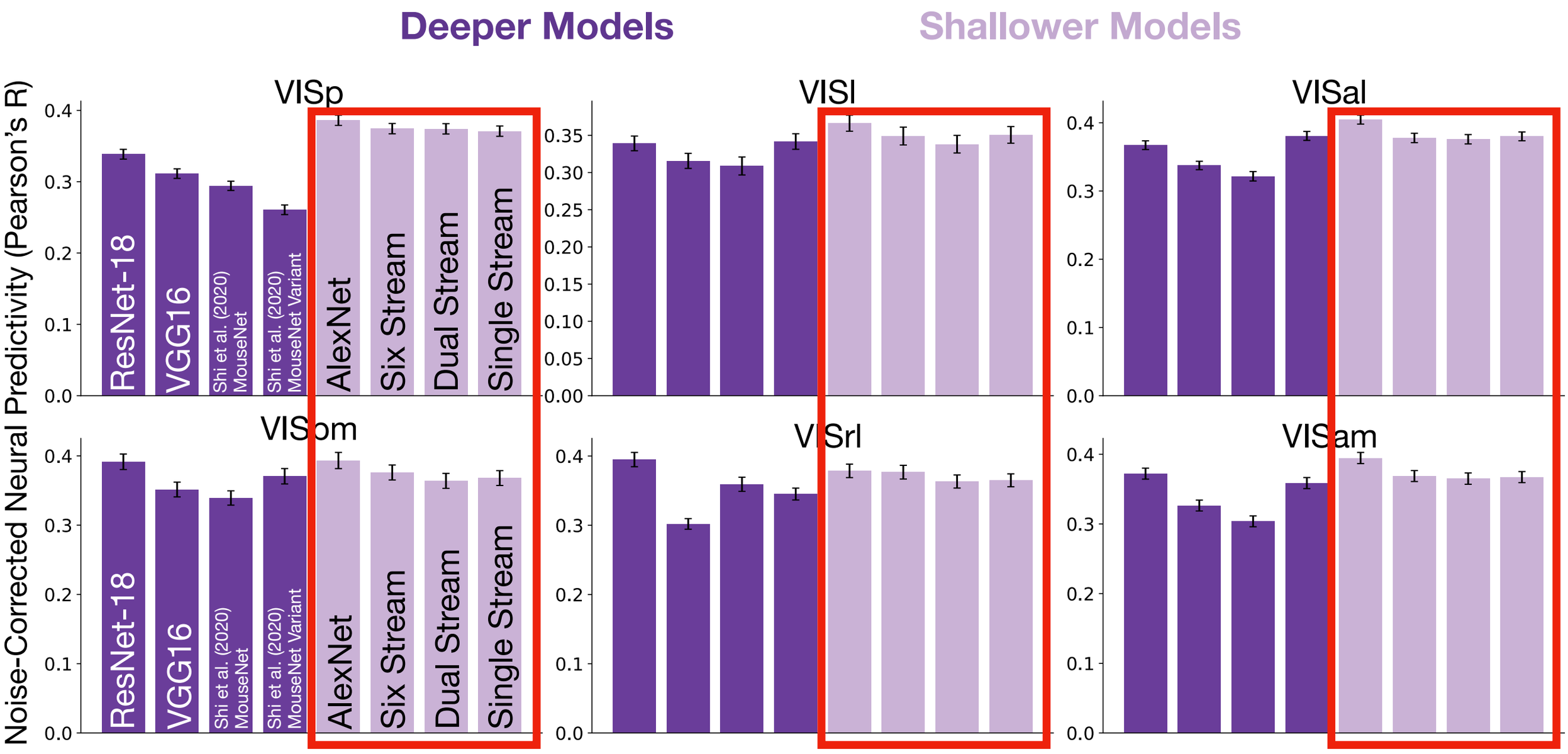
Deeper Models **Shallower Models**



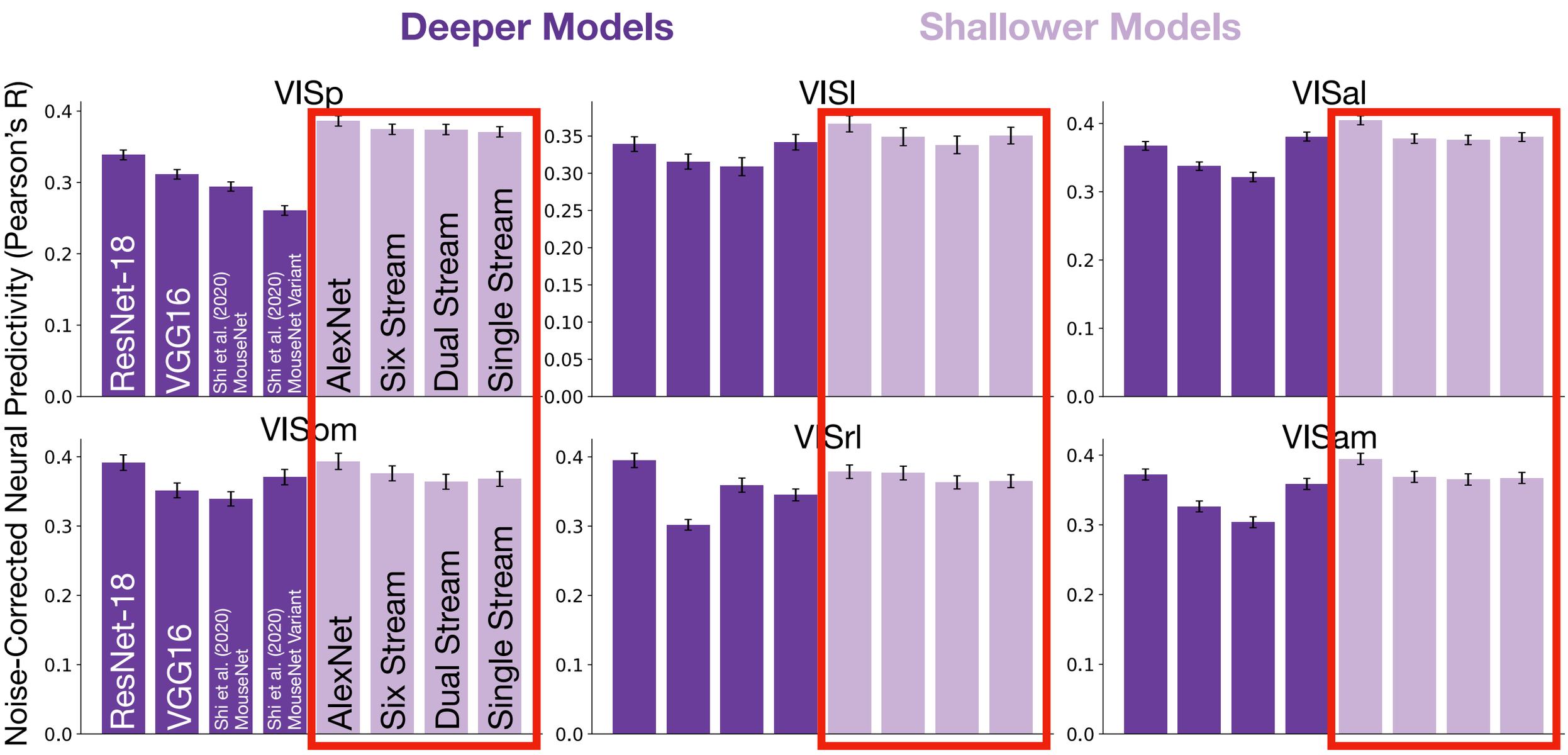
Shallow architectures better predict mouse visual responses than deep architectures



Shallow architectures better predict mouse visual responses than deep architectures

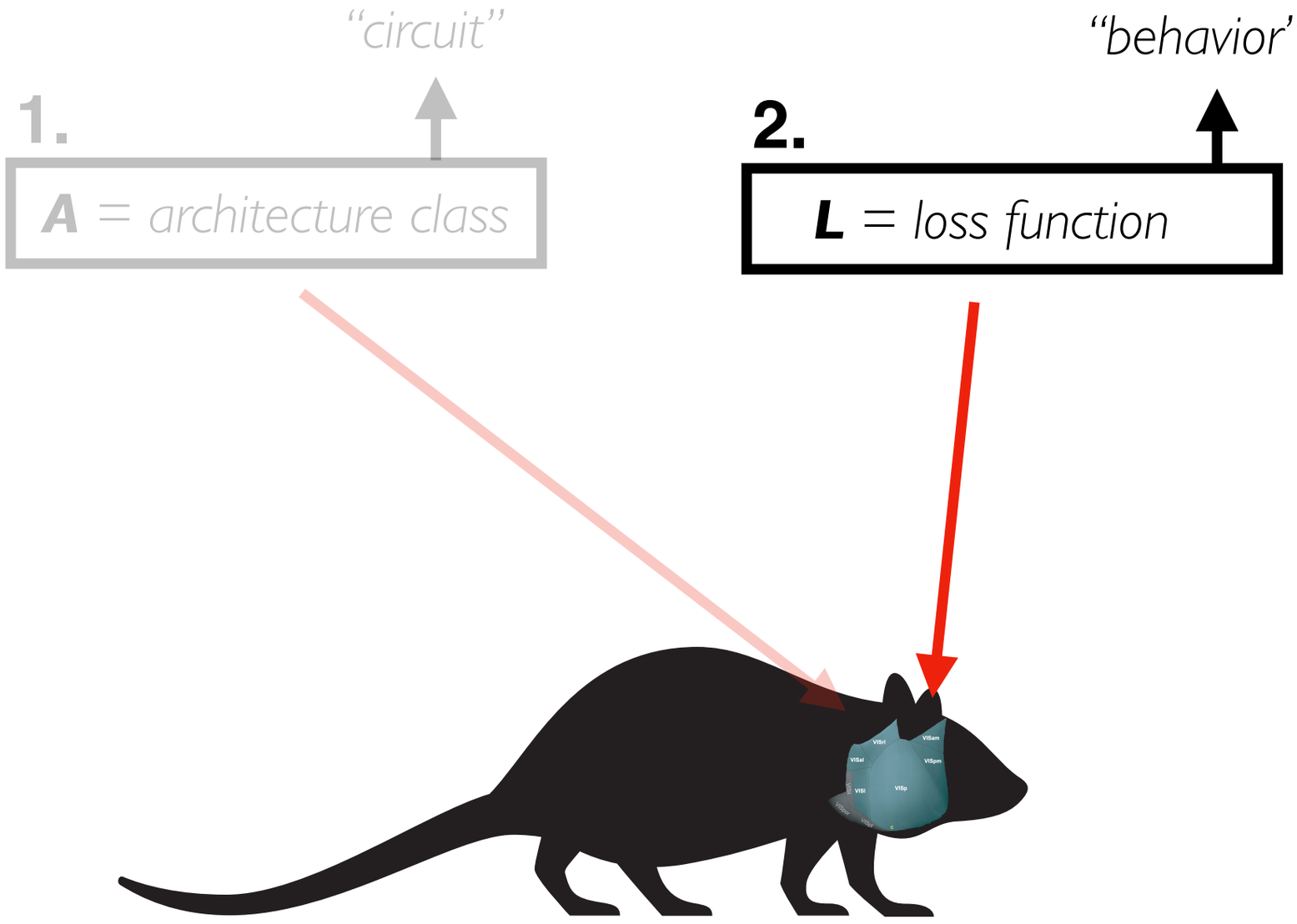


Shallow architectures better predict mouse visual responses than deep architectures



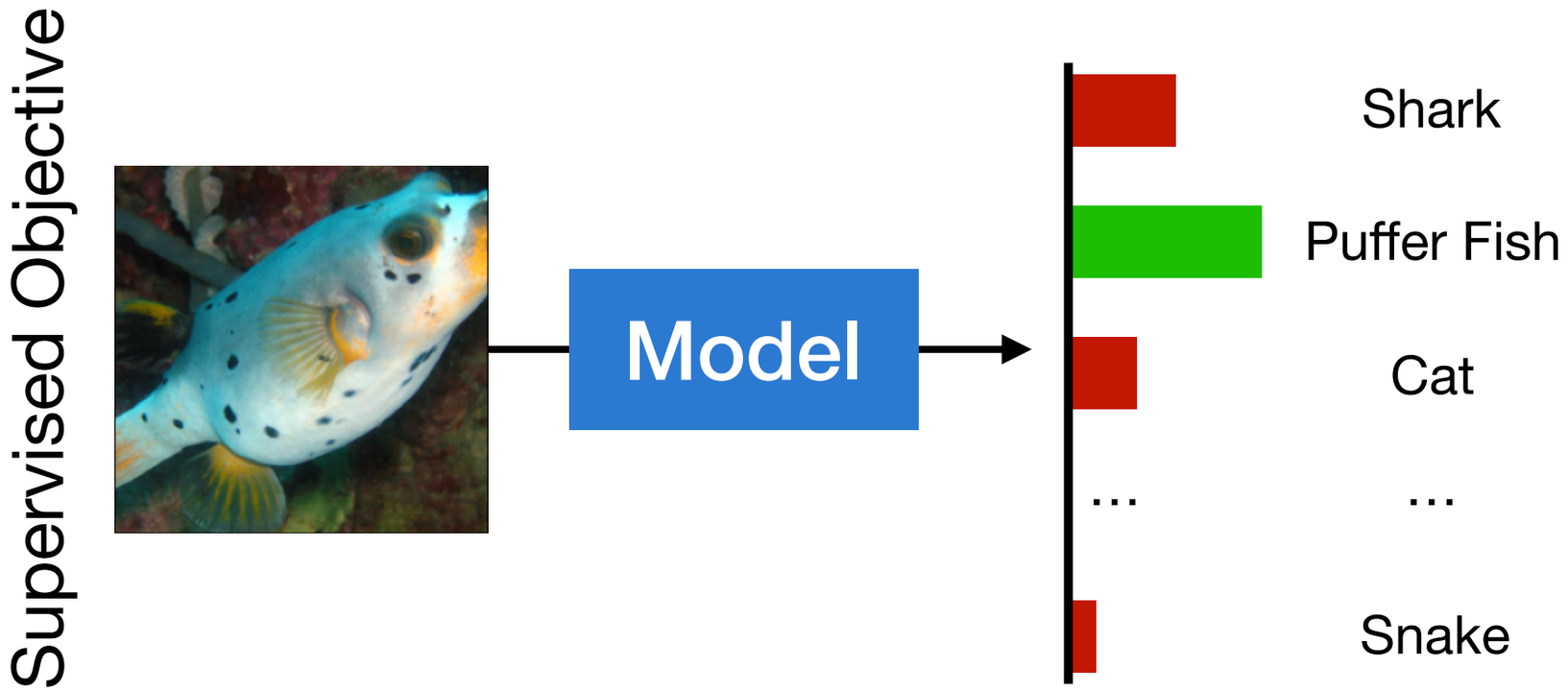
Comparable neural predictivity with a shallower model!

Distilling Constraints: Behavioral Goals



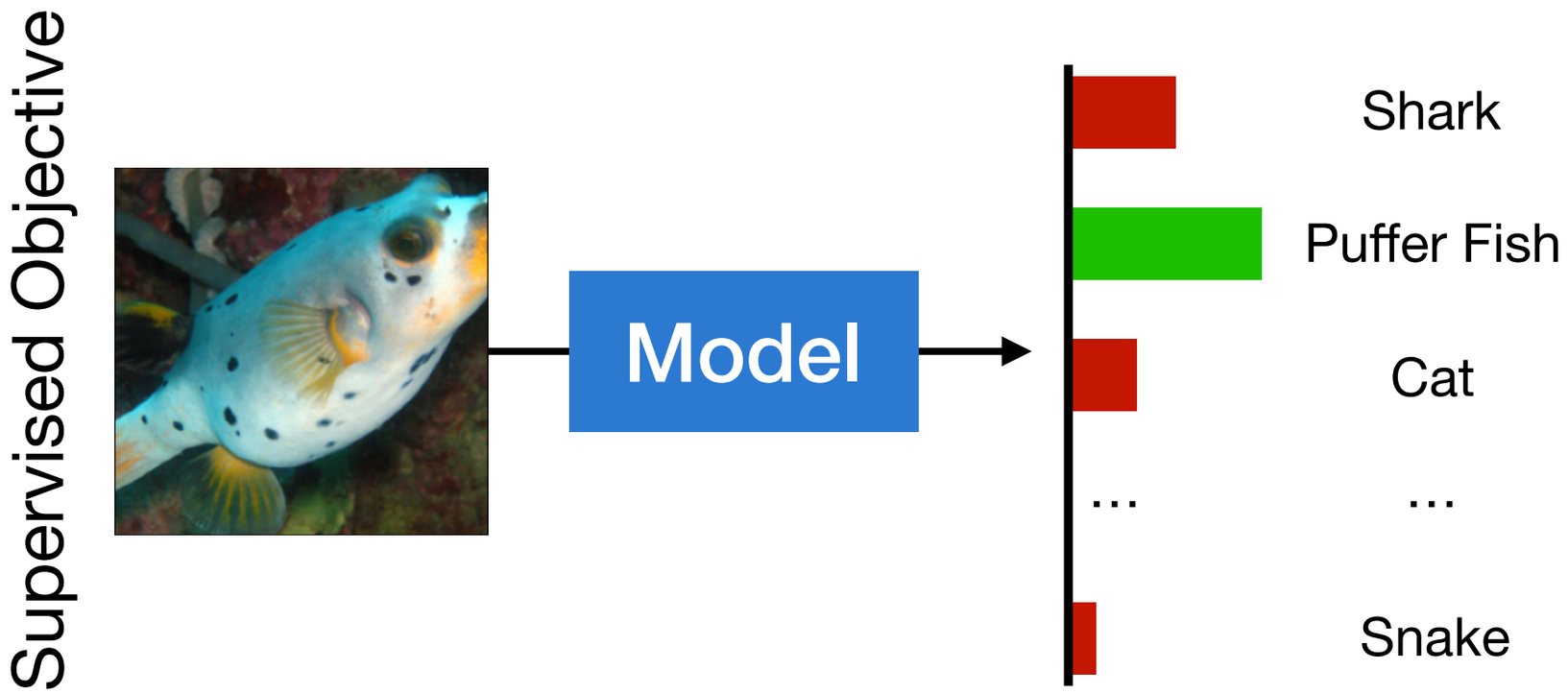
Distilling Constraints: Behavioral Goals

Typical setting: supervision with (1000) category labels



Distilling Constraints: Behavioral Goals

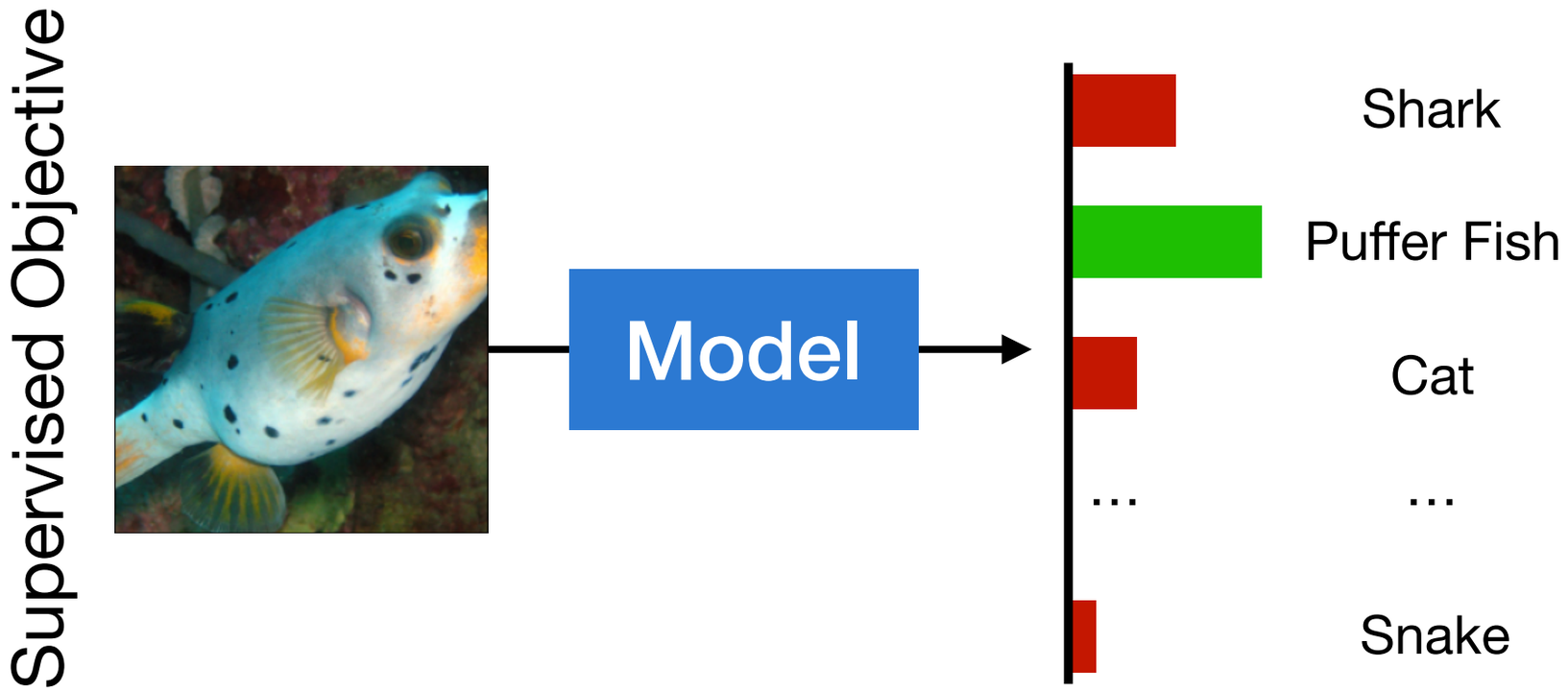
Typical setting: supervision with (1000) category labels
…but is very "unnatural" for mice!



Distilling Constraints: Behavioral Goals

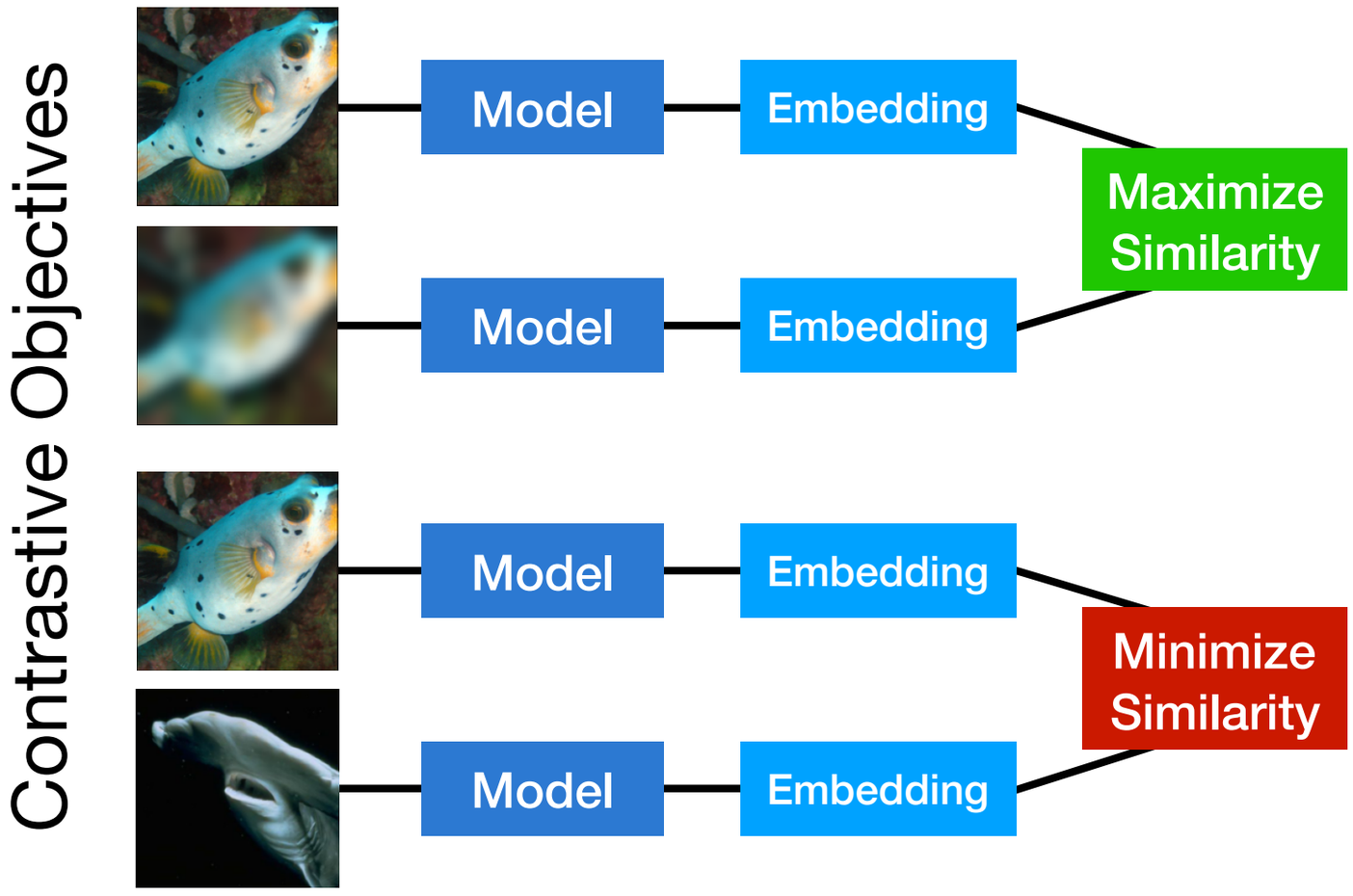
Typical setting: supervision with (1000) category labels
…but is very "unnatural" for mice!

Both the type and number of categories is unrealistic for mice

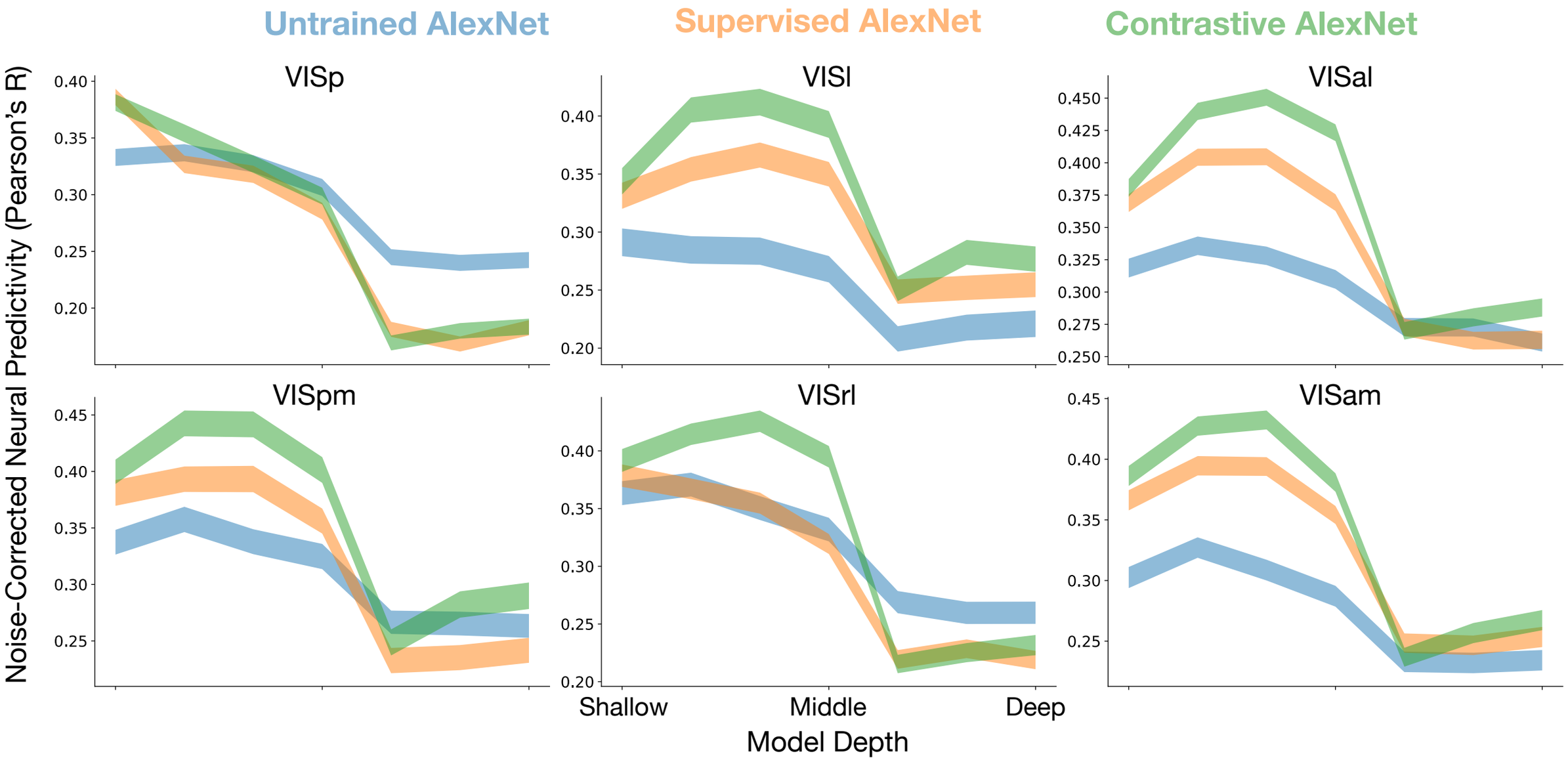


Distilling Constraints: Behavioral Goals

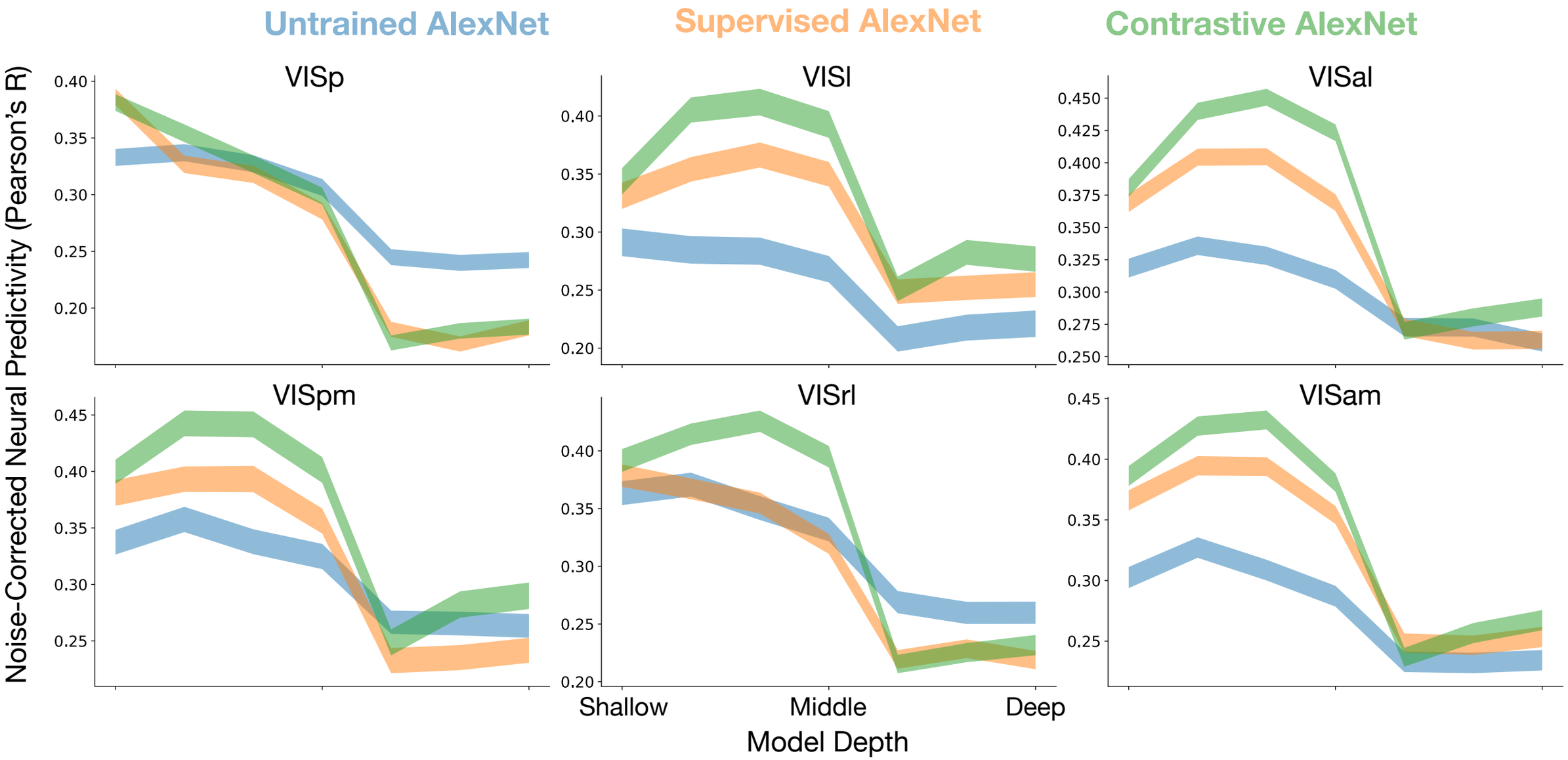
Consider unsupervised objectives, most notably "contrastive" objectives



Distilling Constraints: Behavioral Goals



Distilling Constraints: Behavioral Goals



Using an unsupervised, contrastive objective function further improves neural predictivity!

Distilling Constraints: Inputs

"circuit"

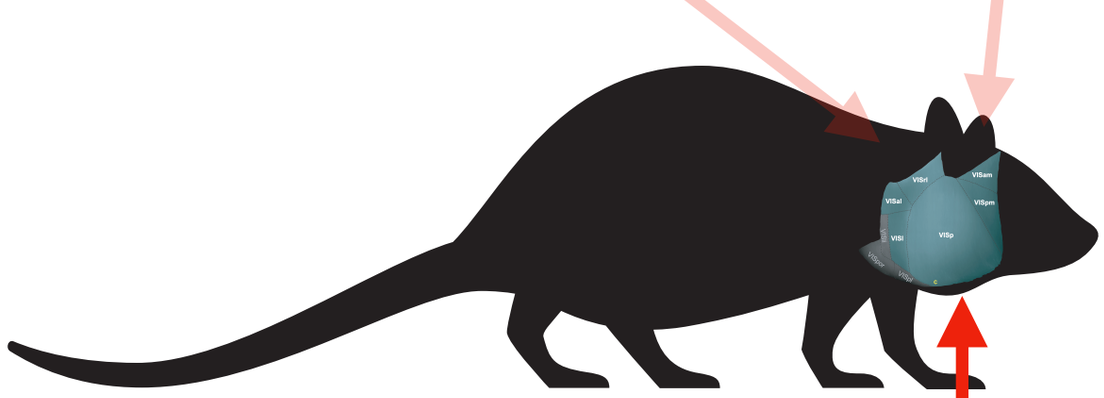
"behavior"

1.

A = *architecture class*

2.

L = *loss function*



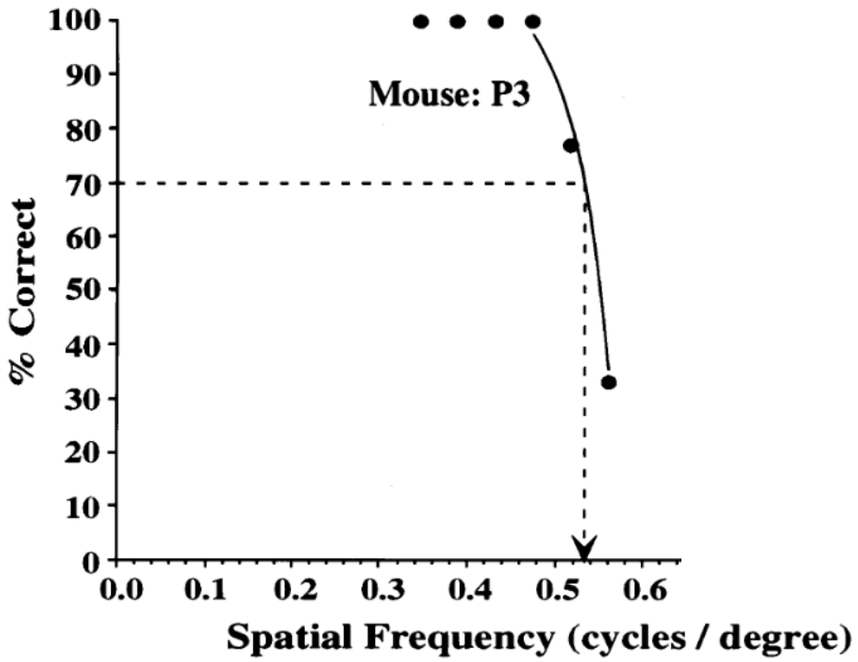
3.

D = *data stream*

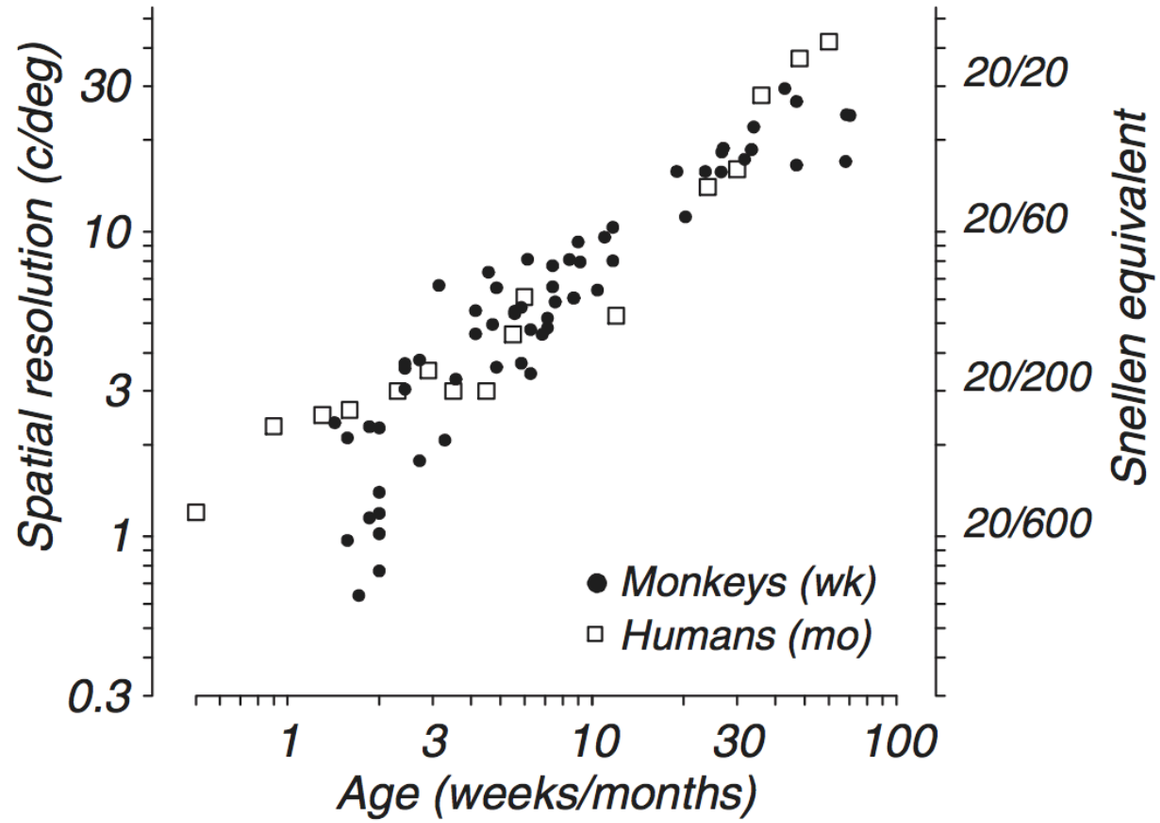
"inputs"

Distilling Constraints: Inputs

Mice



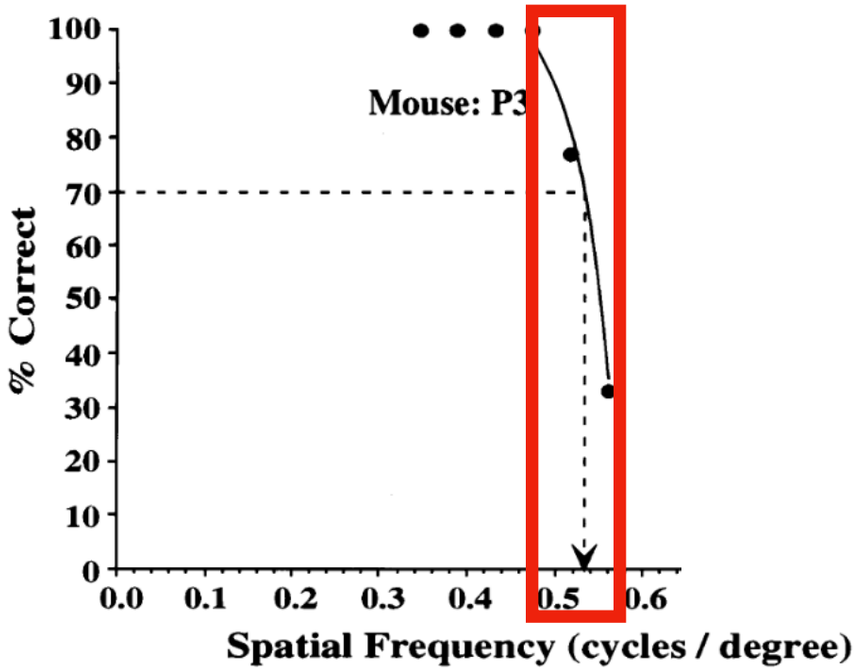
Primates



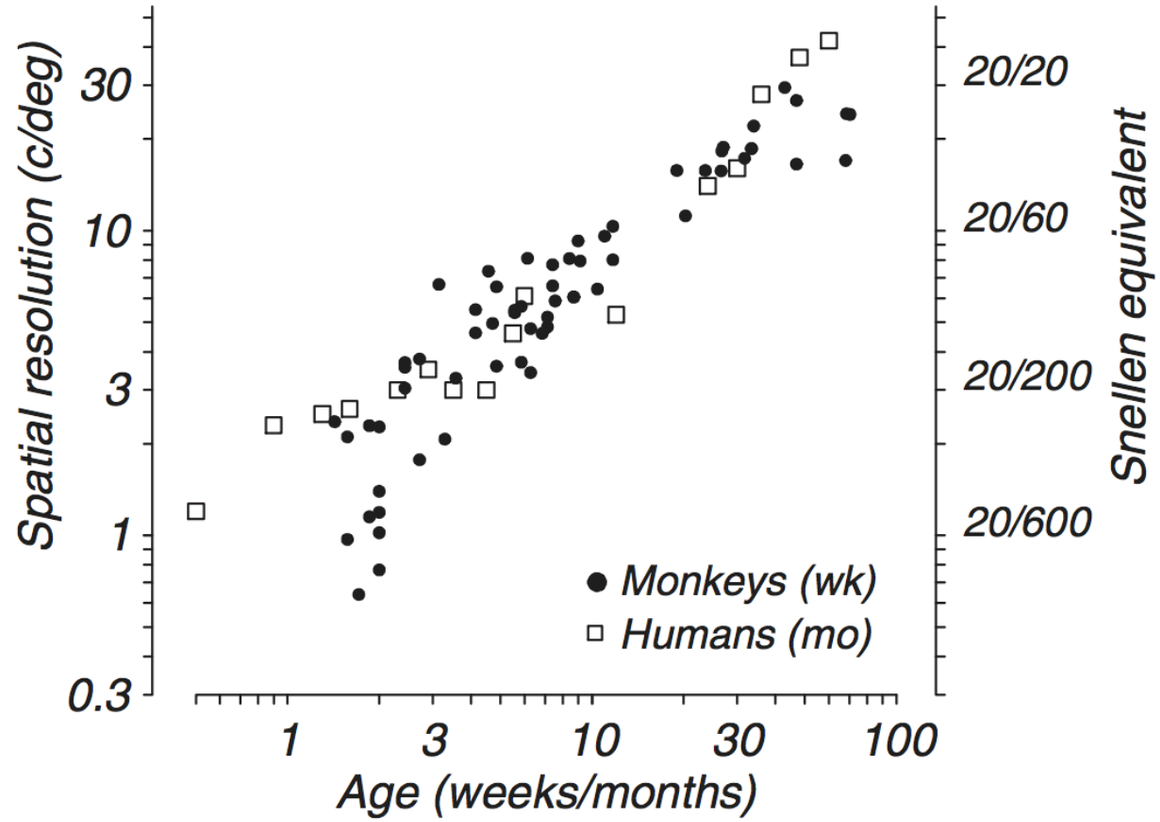
Prusky et al., 2000; Kiorpes, 2019

Distilling Constraints: Inputs

Mice



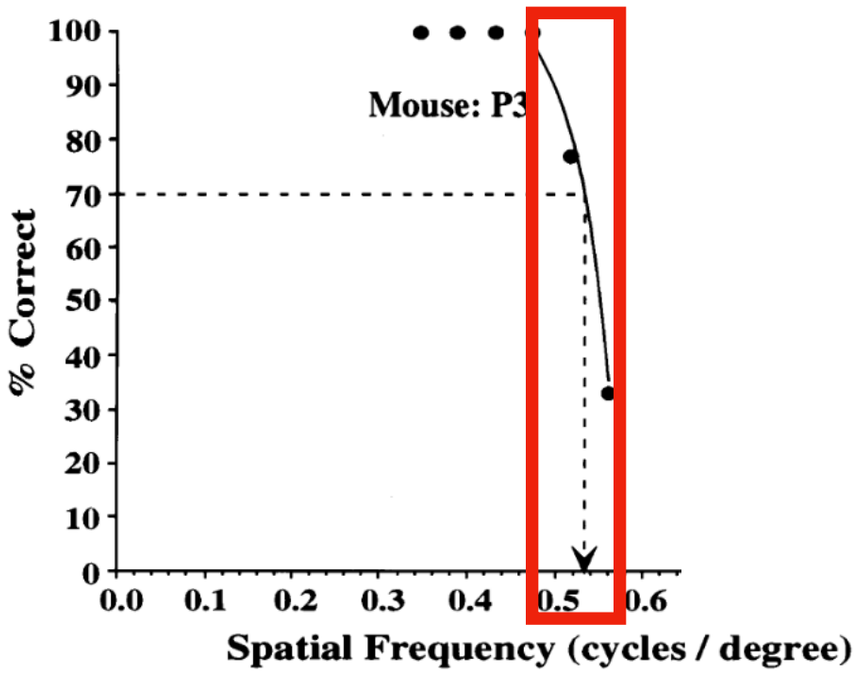
Primates



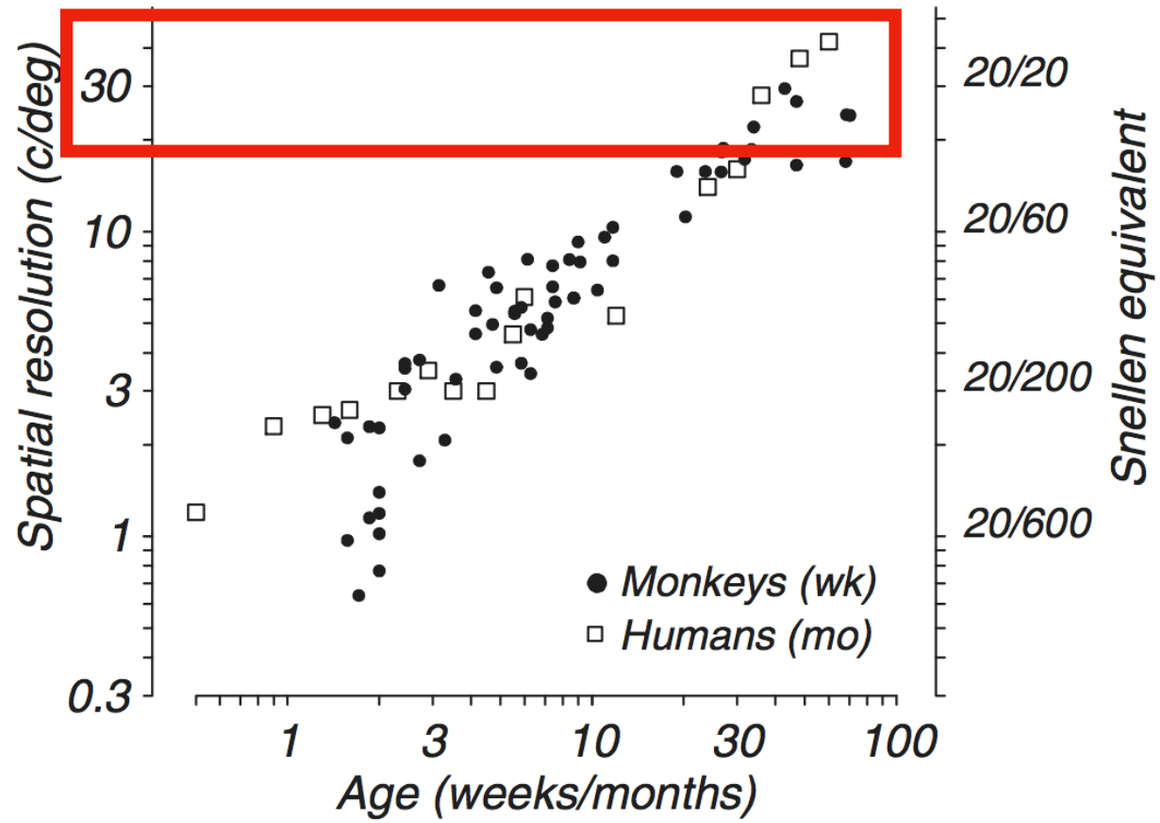
Prusky et al., 2000; Kiorpes, 2019

Distilling Constraints: Inputs

Mice

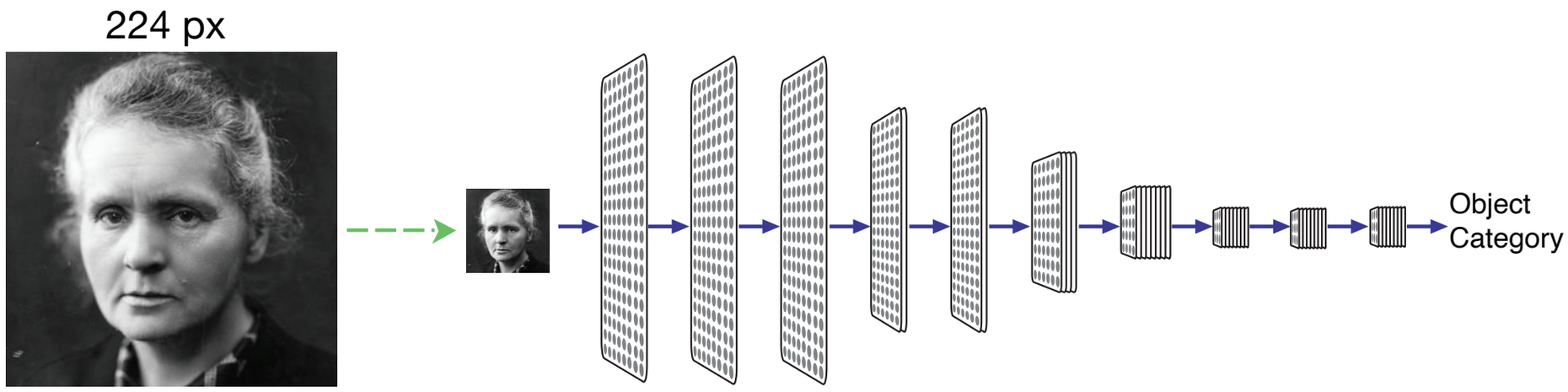


Primates

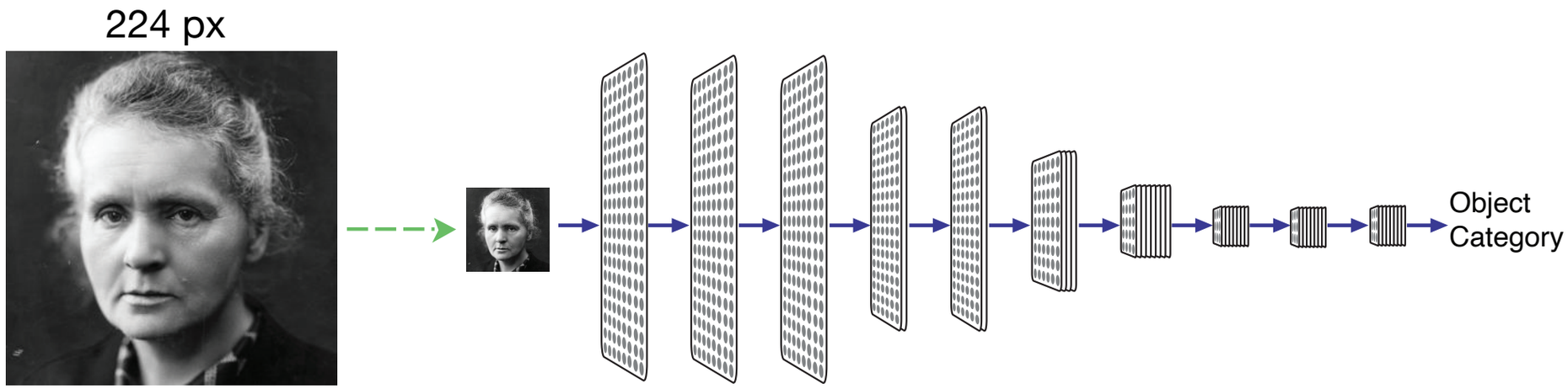


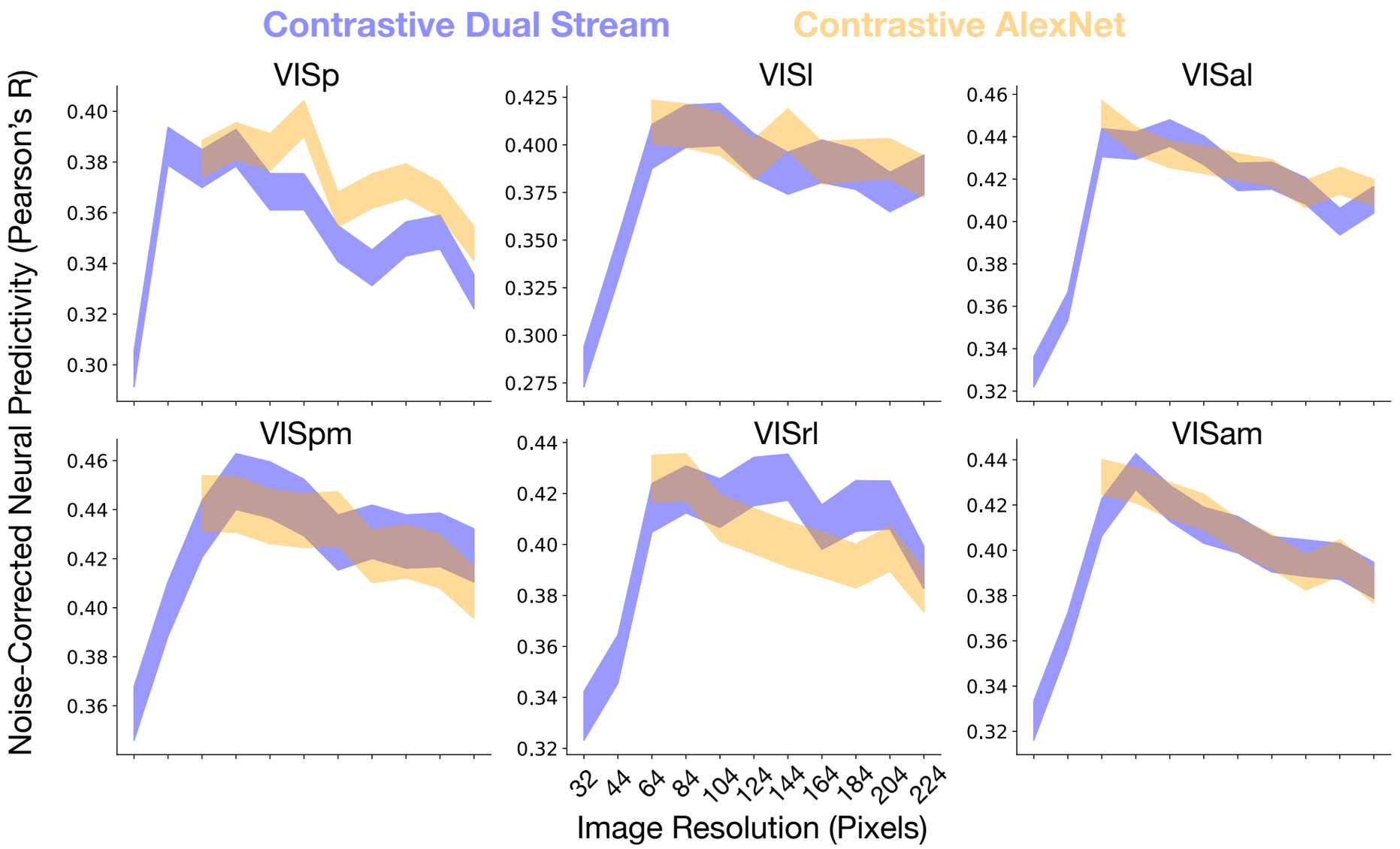
Prusky et al., 2000; Kiorpes, 2019

Distilling Constraints: Inputs

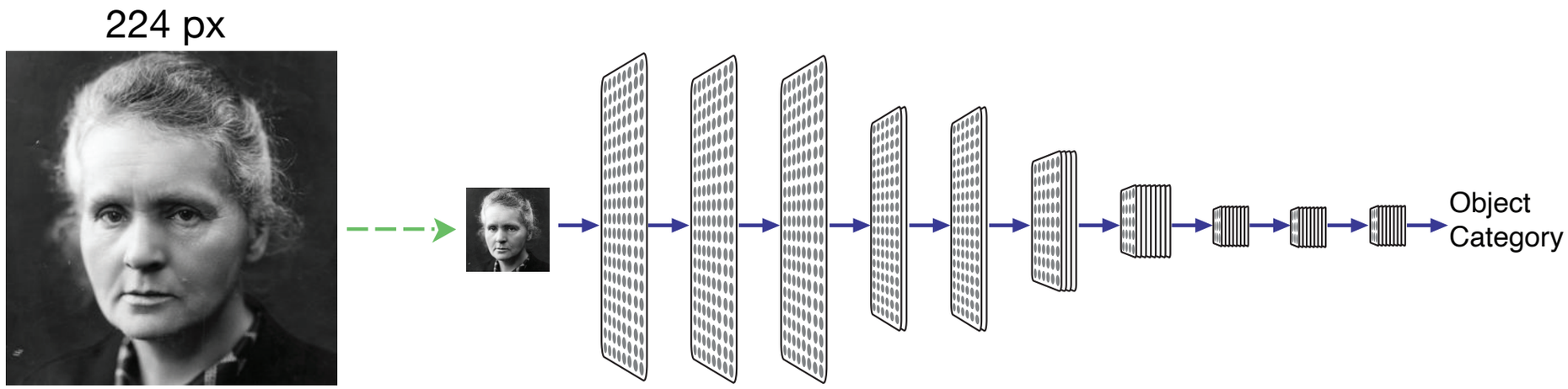


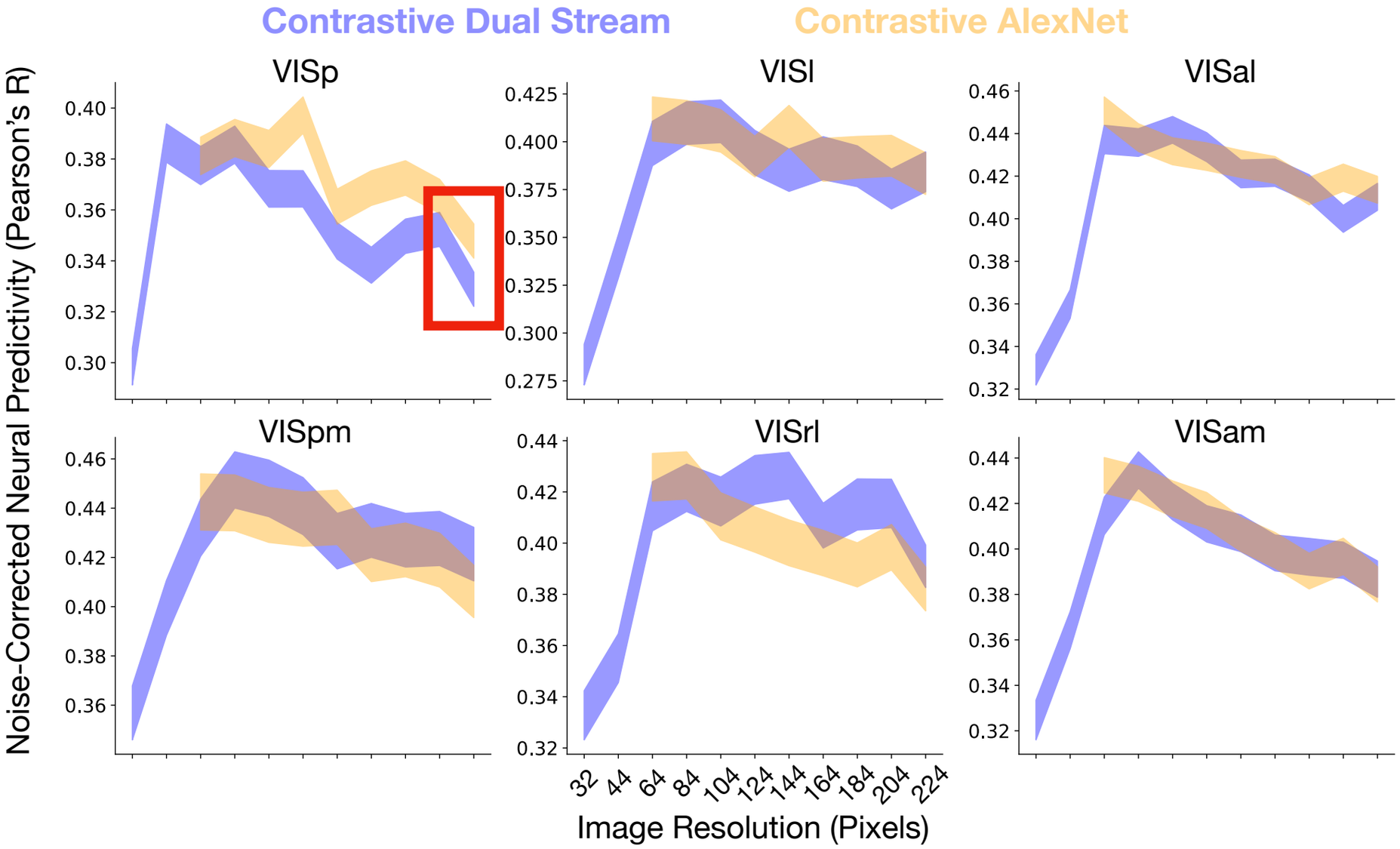
Distilling Constraints: Inputs



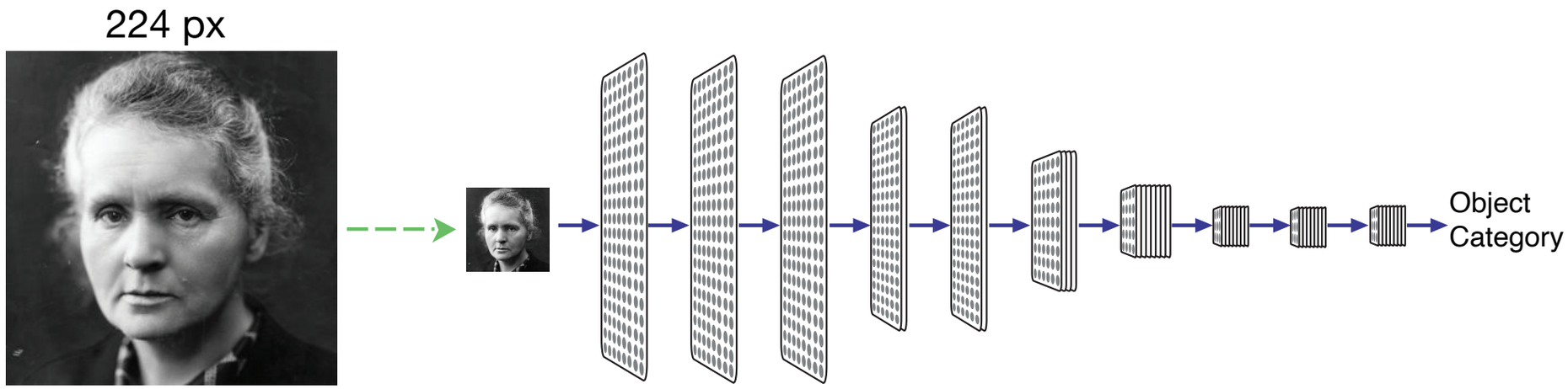


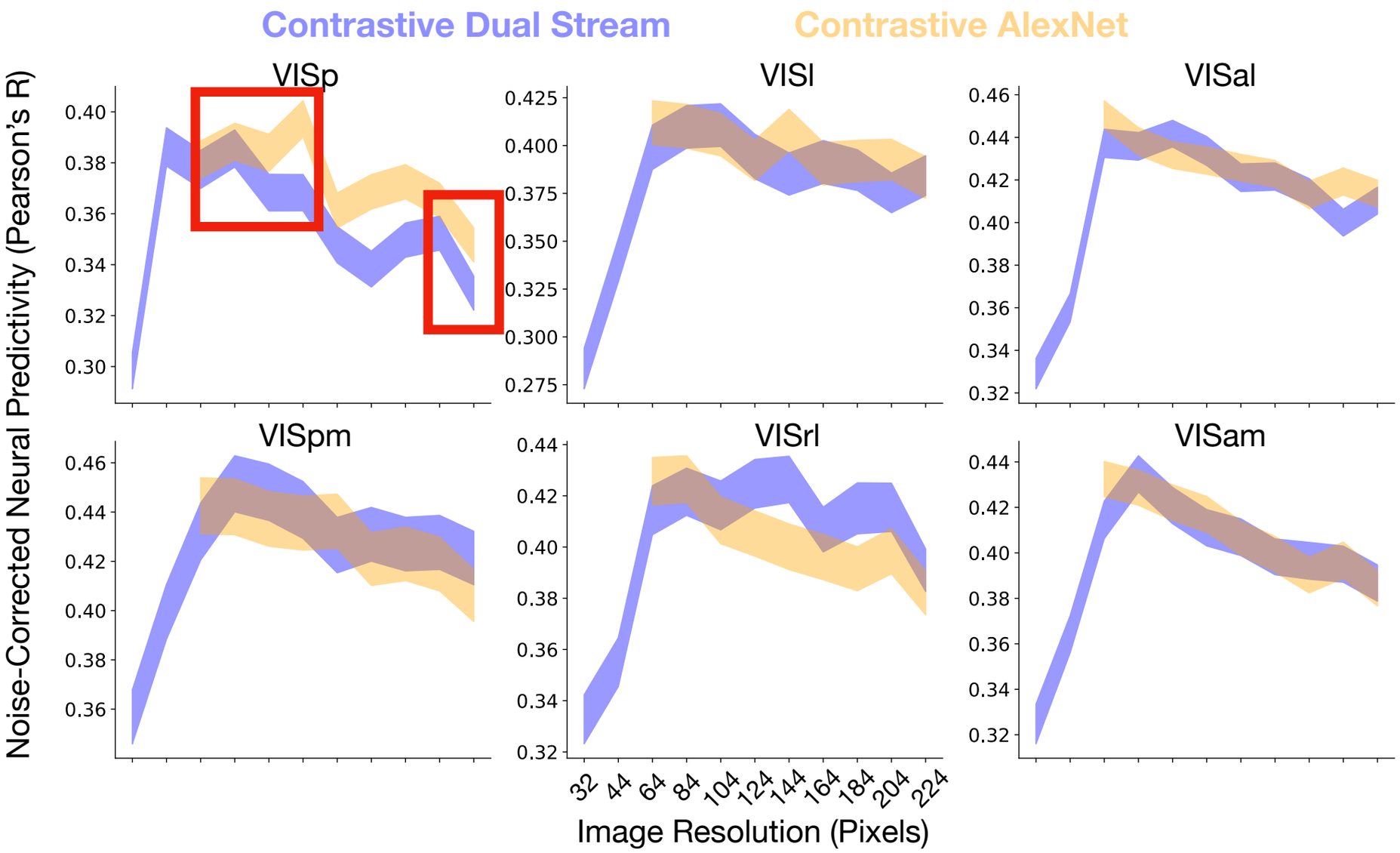
Distilling Constraints: Inputs



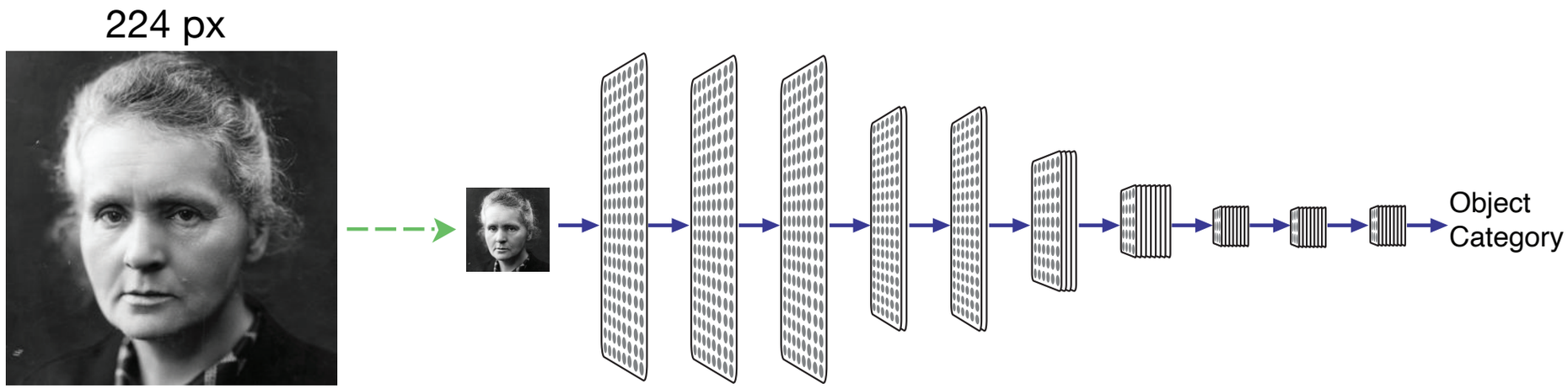


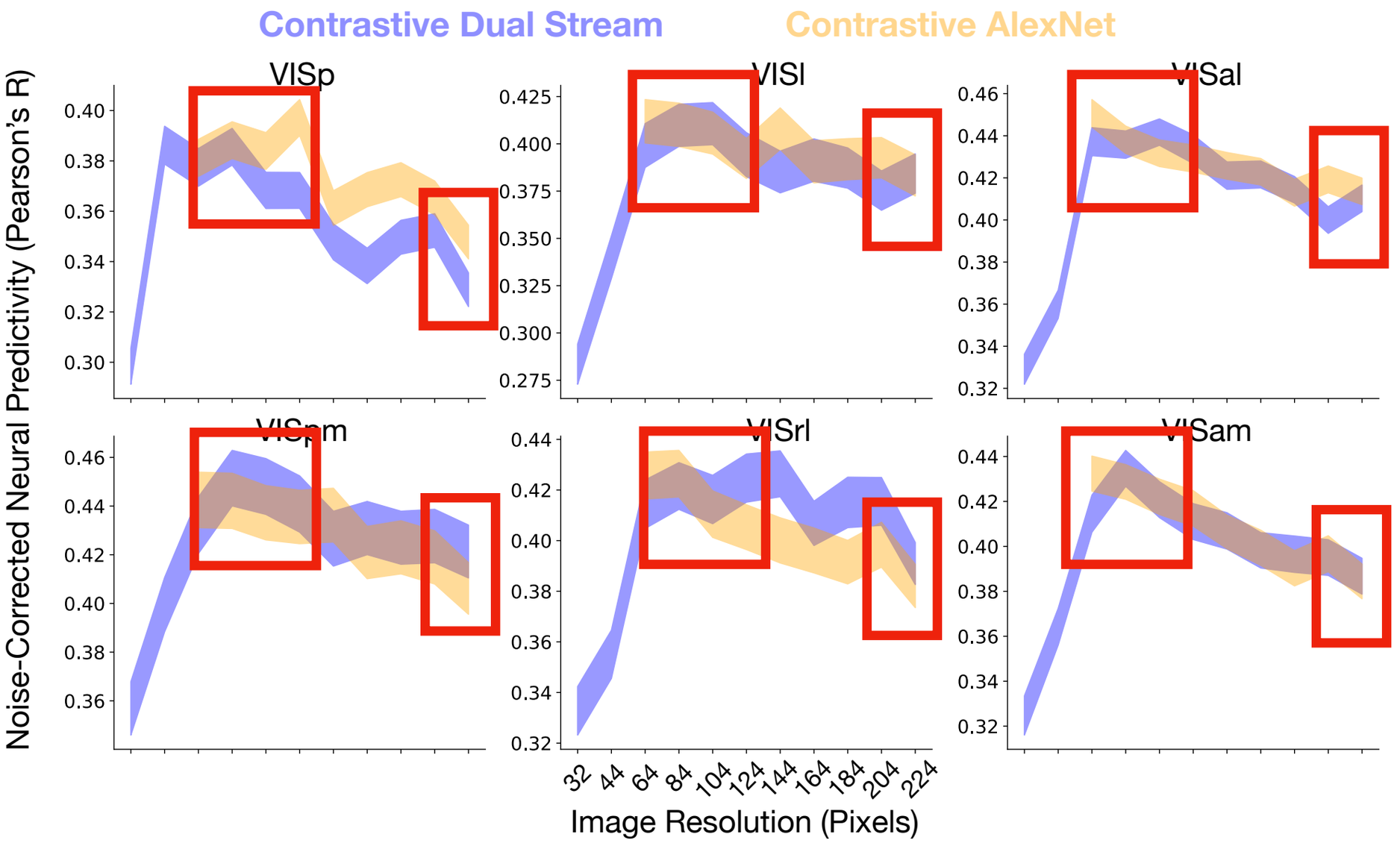
Distilling Constraints: Inputs



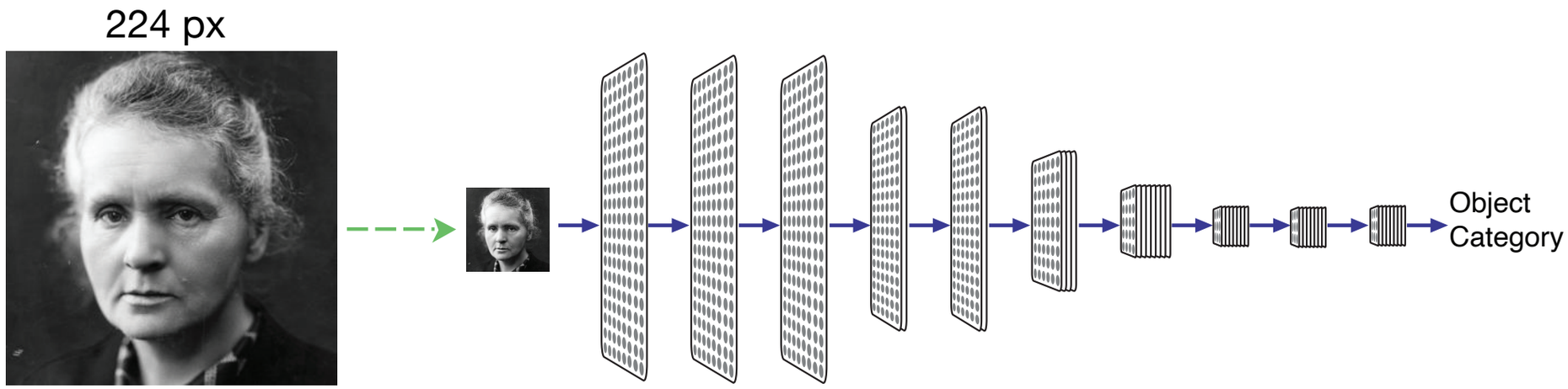


Distilling Constraints: Inputs

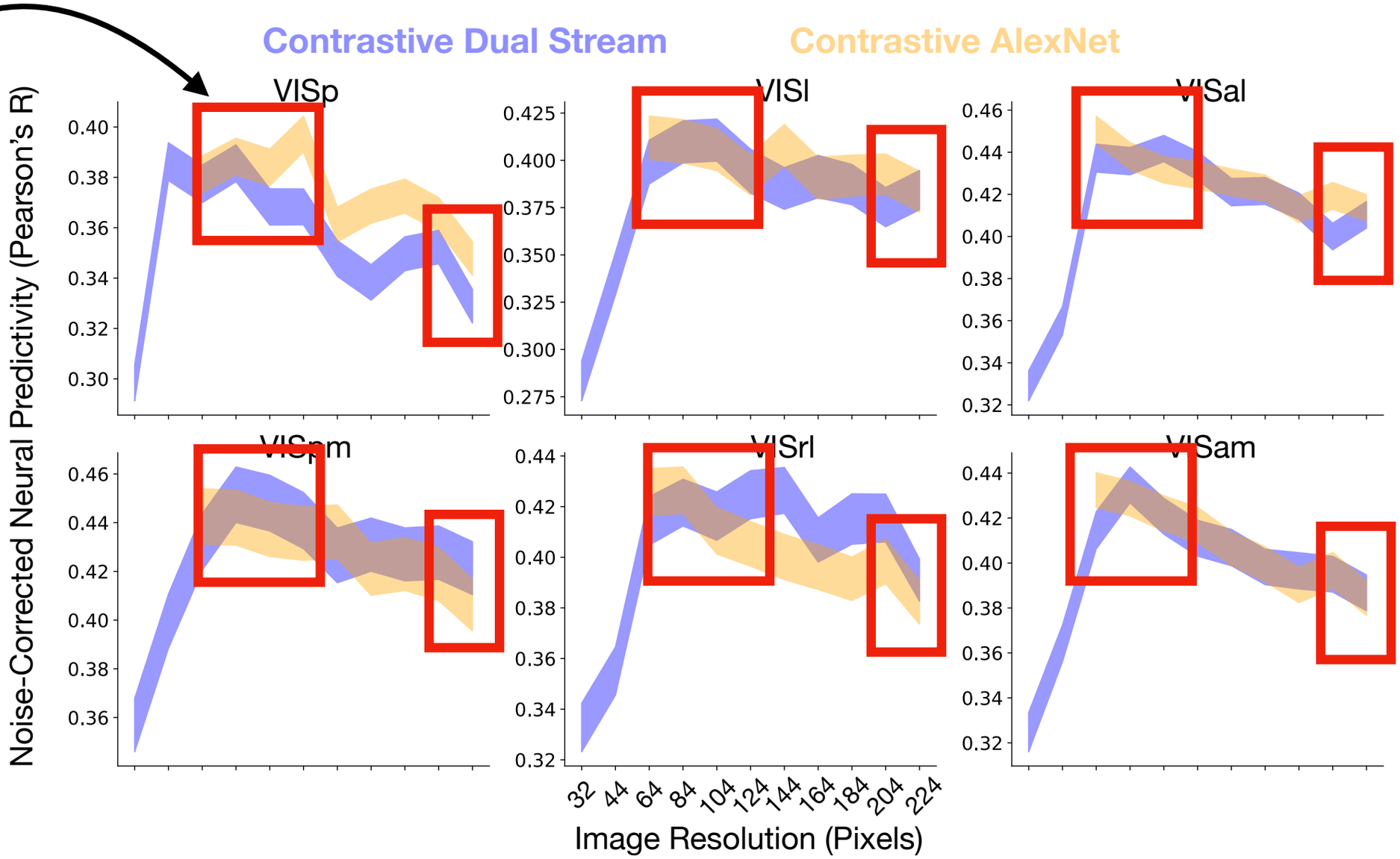




Distilling Constraints: Inputs



Reducing image resolution during training improves neural predictivity across all visual areas!



Putting it all together: Circuit, Behavior, Input

"circuit"

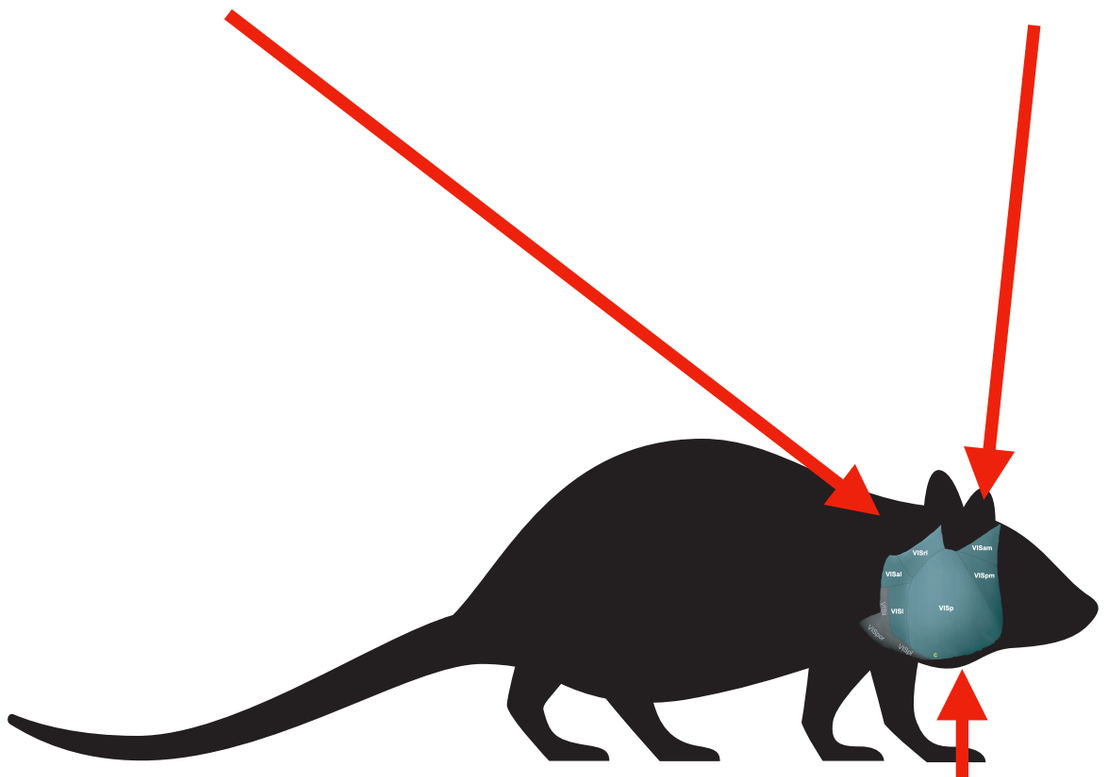
1.

A = *architecture class*

"behavior"

2.

L = *loss function*

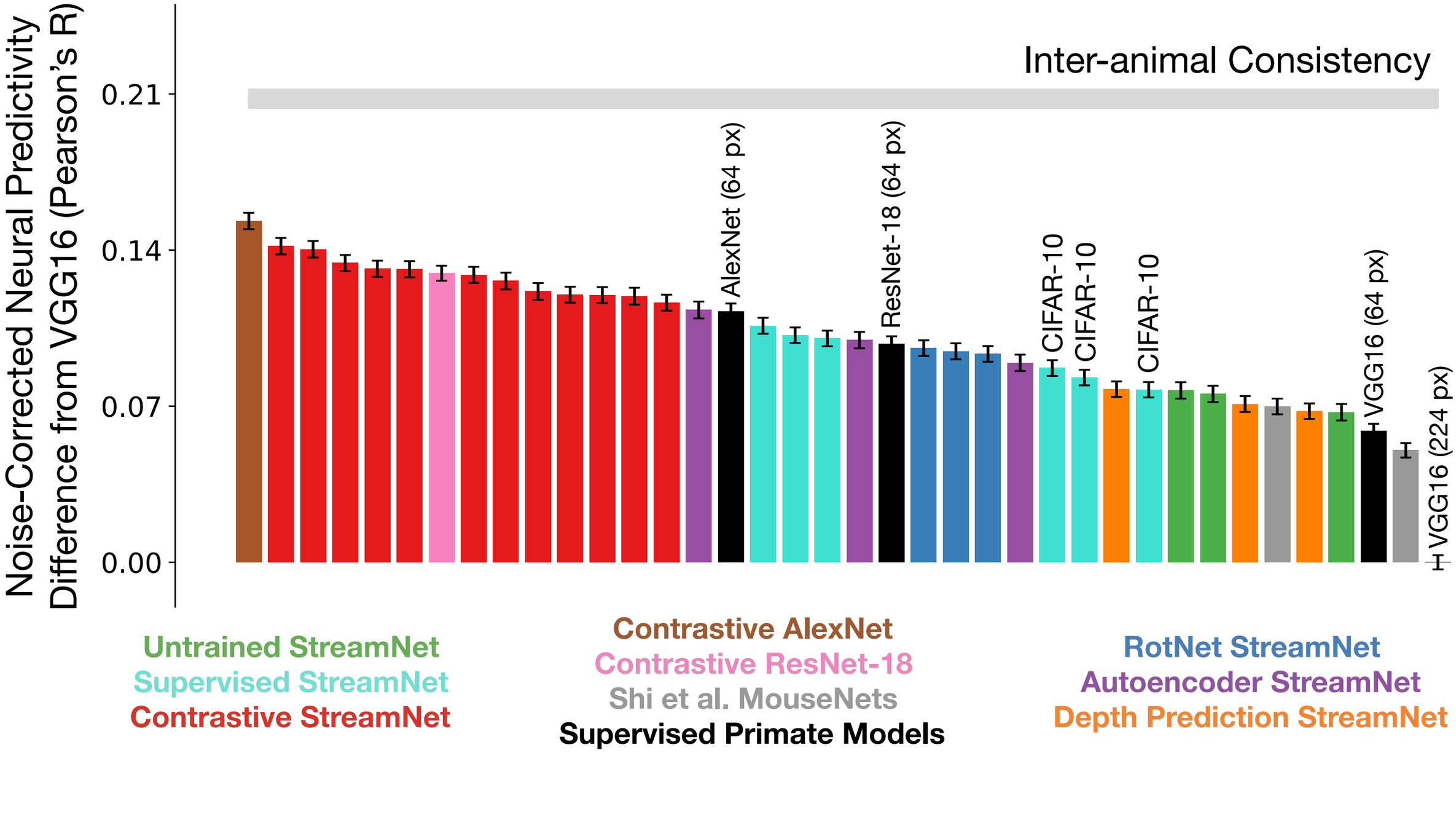


3.

D = *data stream*

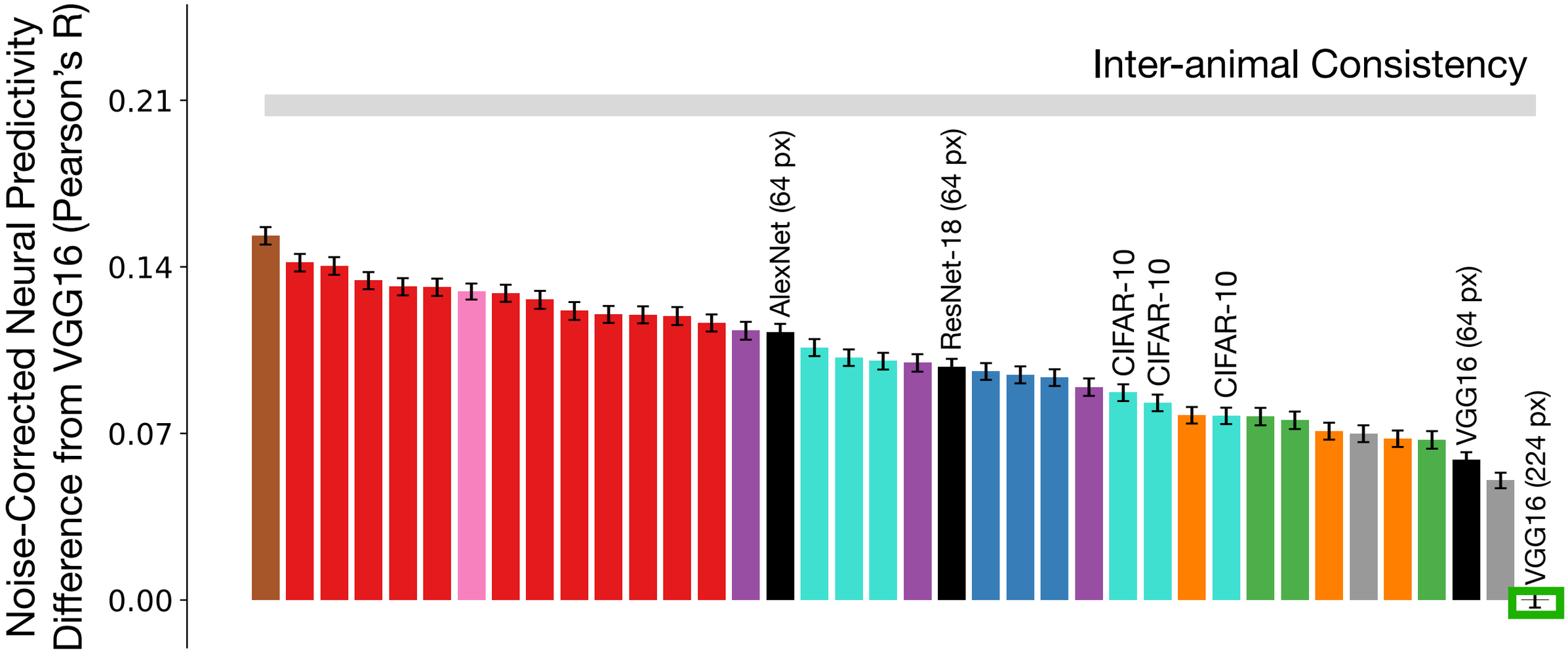
"inputs"

Putting it all together: Circuit, Behavior, Input



Putting it all together: Circuit, Behavior, Input

Previous baseline (deep supervised primate model)



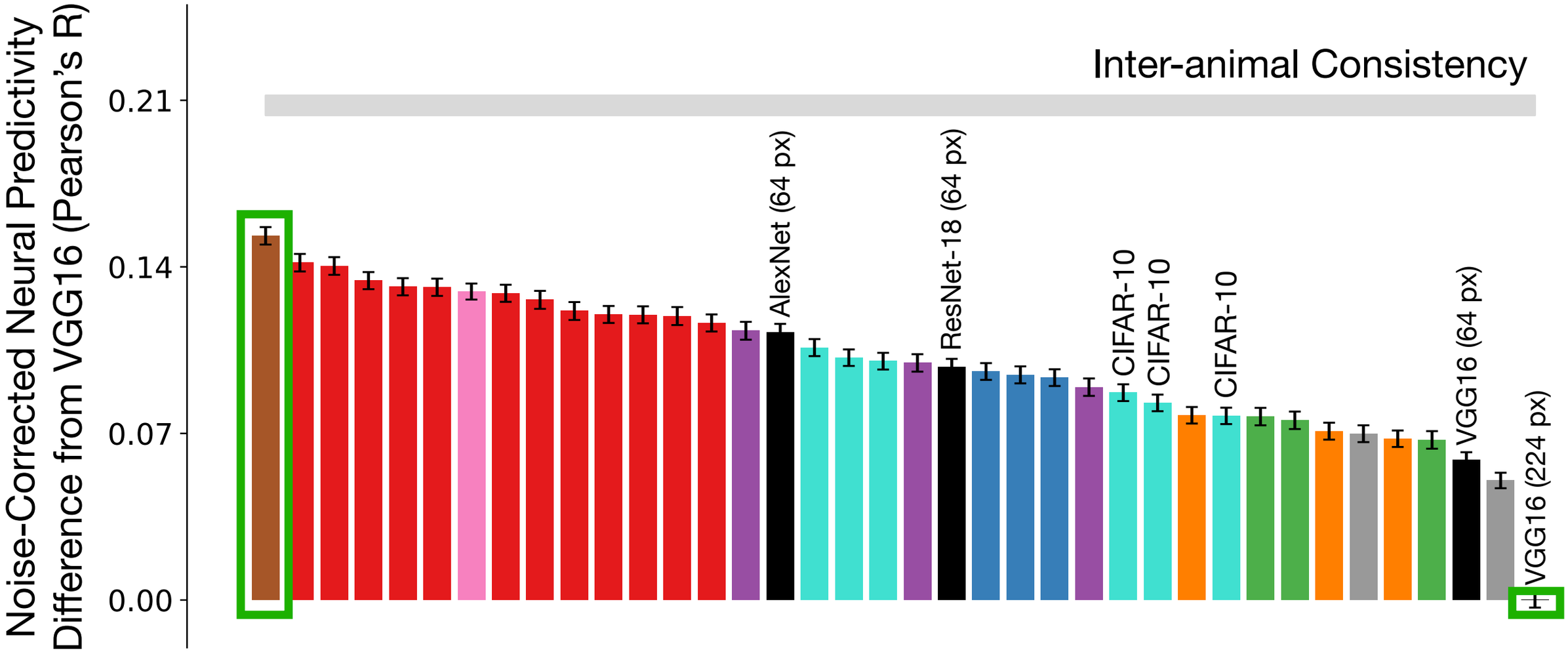
Untrained StreamNet
Supervised StreamNet
Contrastive StreamNet

Contrastive AlexNet
Contrastive ResNet-18
Shi et al. MouseNets
Supervised Primate Models

RotNet StreamNet
Autoencoder StreamNet
Depth Prediction StreamNet

Putting it all together: Circuit, Behavior, Input

Previous baseline (deep supervised primate model)
Shallow low resolution unsupervised model

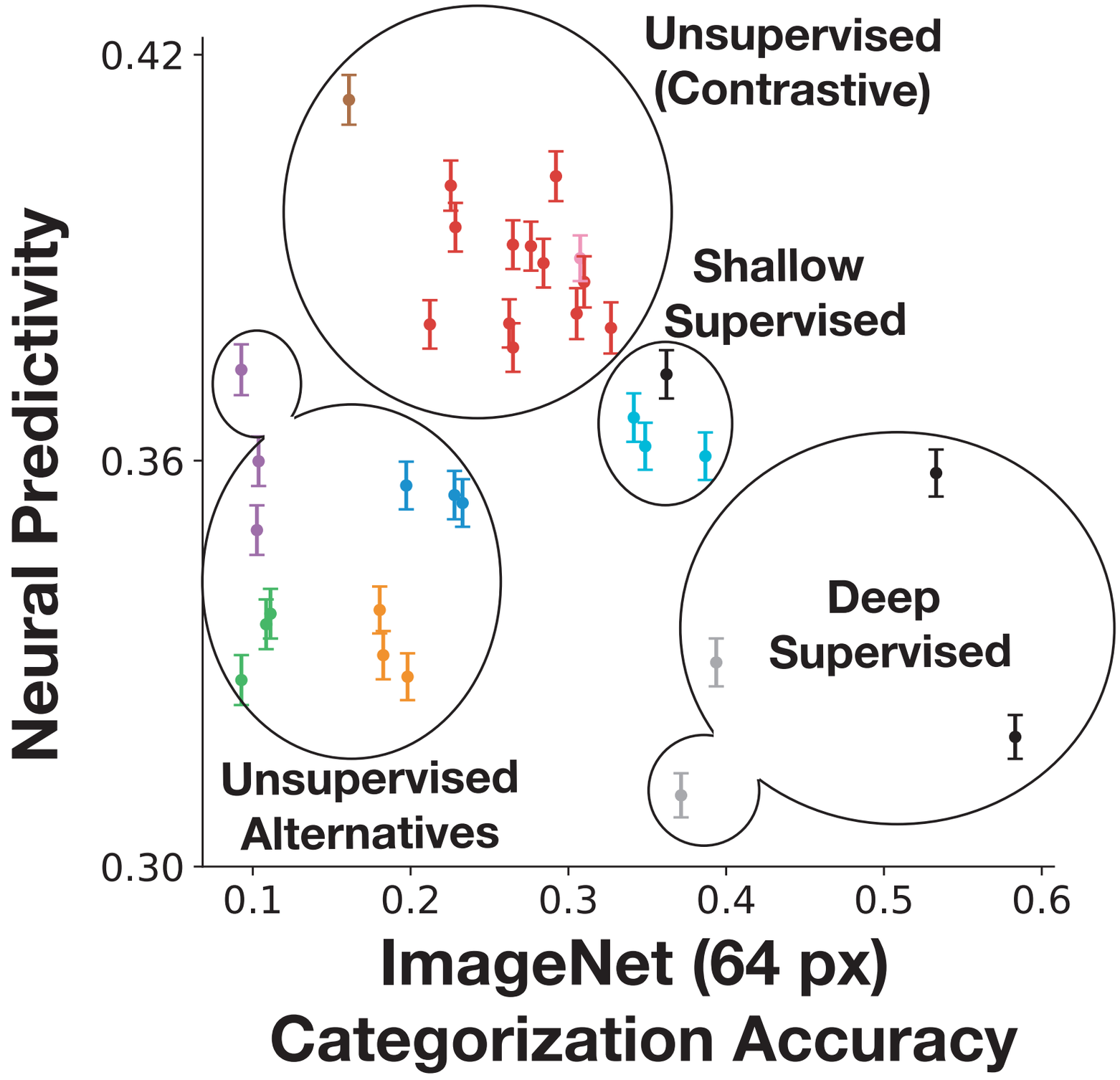


Untrained StreamNet
Supervised StreamNet
Contrastive StreamNet

Contrastive AlexNet
Contrastive ResNet-18
Shi et al. MouseNets
Supervised Primate Models

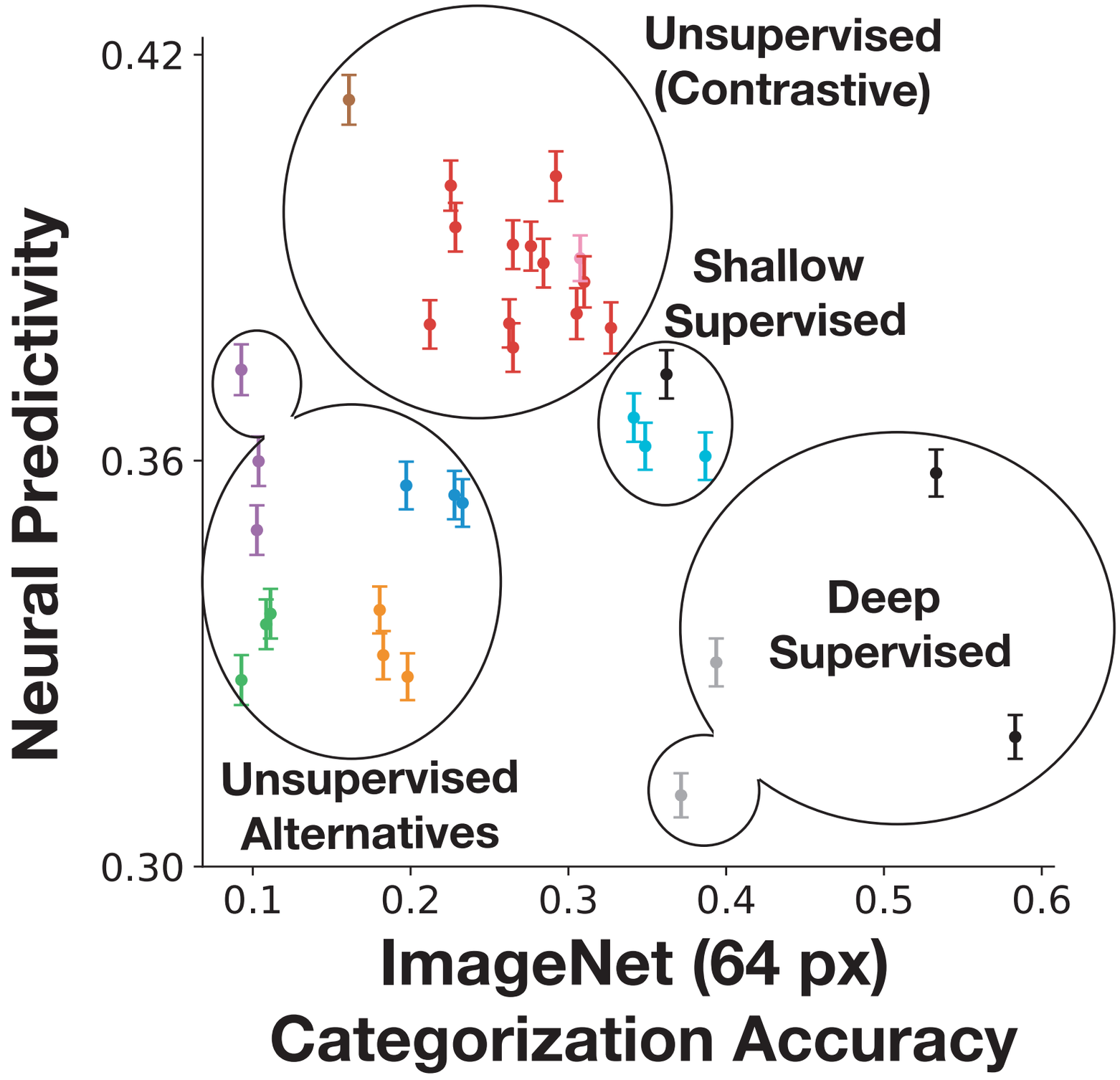
RotNet StreamNet
Autoencoder StreamNet
Depth Prediction StreamNet

Putting it all together: Circuit, Behavior, Input



Putting it all together: Circuit, Behavior, Input

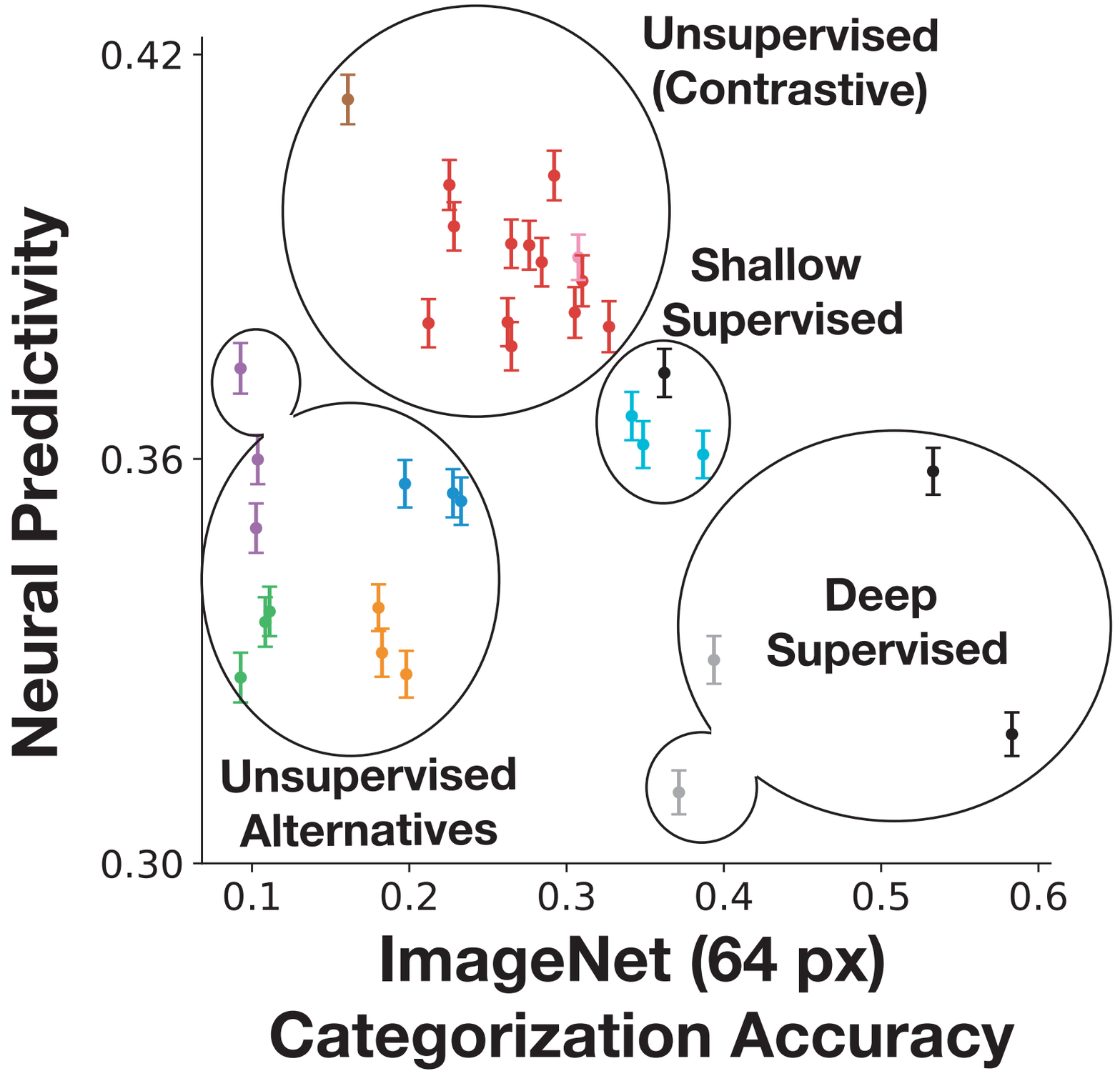
Not all unsupervised algorithms improve neural predictivity.



Putting it all together: Circuit, Behavior, Input

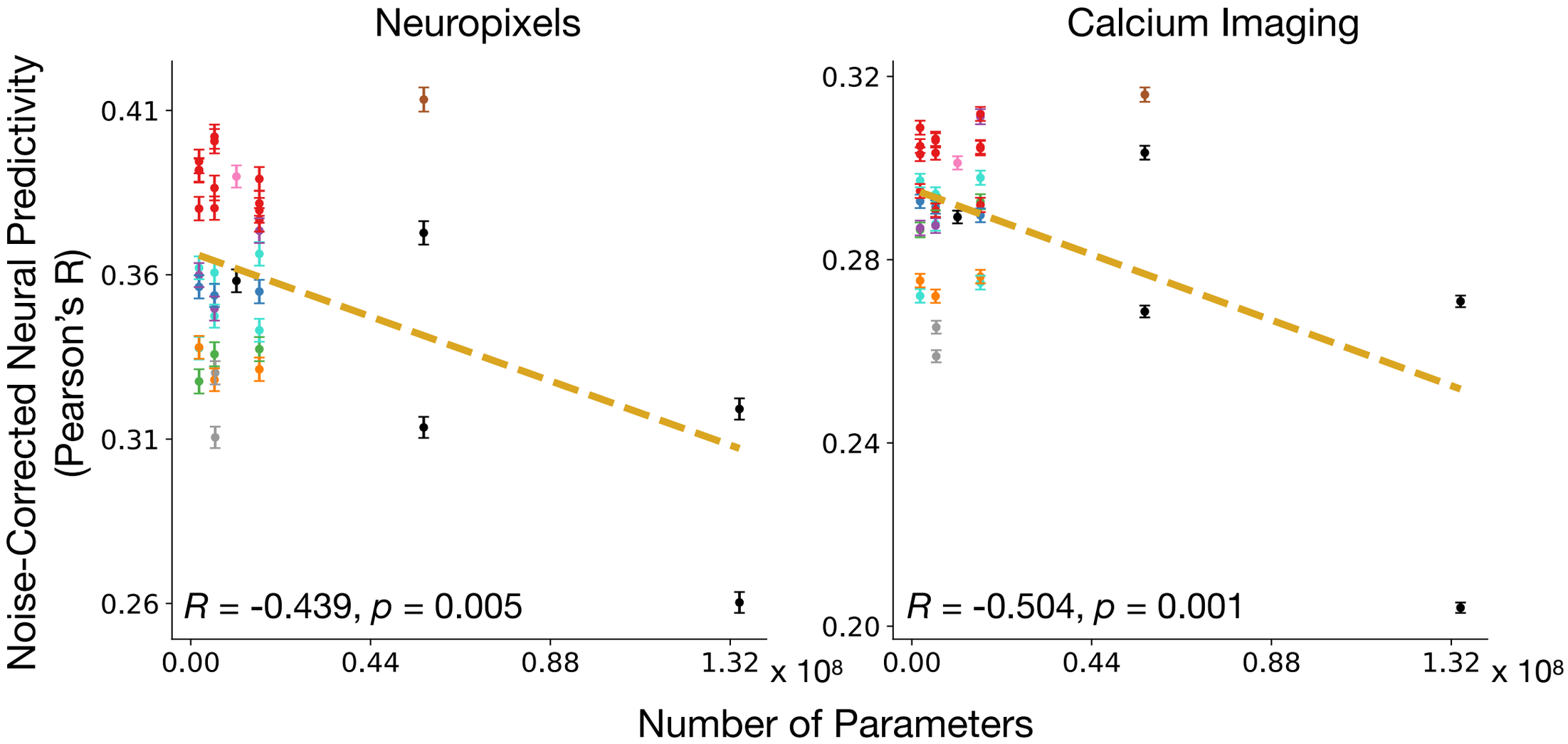
Not all unsupervised algorithms improve neural predictivity.

Higher categorization accuracy also does not improve neural predictivity.



Putting it all together: Circuit, Behavior, Input

Increasing number of parameters is also associated with worse neural predictivity



Takeaways

- Mouse visual cortex is best captured (so far) by three ingredients:

Takeaways

- Mouse visual cortex is best captured (so far) by three ingredients:
  - **Circuit**: shallow architecture
  - **Behavior**: contrastive (unsupervised)
  - **Input**: low resolution

Takeaways

- Mouse visual cortex is best captured (so far) by three ingredients:
  - **Circuit**: shallow architecture
  - **Behavior**: contrastive (unsupervised)
  - **Input**: low resolution

- Mouse visual cortex is a general-purpose machine utilizing its limited resources to perform a variety of visual tasks

Takeaways

- Mouse visual cortex is best captured (so far) by three ingredients:
  - **Circuit**: shallow architecture
  - **Behavior**: contrastive (unsupervised)
  - **Input**: low resolution

- Mouse visual cortex is a general-purpose machine utilizing its limited resources to perform a variety of visual tasks

- This is all in contrast to the deep, high-resolution, and task-specific visual system of the primate

Takeaways

- Mouse visual cortex is best captured (so far) by three ingredients:
  - **Circuit**: shallow architecture
  - **Behavior**: contrastive (unsupervised)
  - **Input**: low resolution

- Mouse visual cortex is a general-purpose machine utilizing its limited resources to perform a variety of visual tasks
- This is all in contrast to the deep, high-resolution, and task-specific visual system of the primate
- Generic nature of the behavior could be used by other sensory systems

Outline

- Recurrent Connections in the Primate Ventral Stream
- Goal-Driven Models of Mouse Visual Cortex
- Heterogeneity in Rodent Medial Entorhinal Cortex
- Building and Identifying Learning Rules

Goal-Driven Modeling - Three Primary Components

"circuit"

1.

A = *architecture class*

"behavior"

2.

L = *loss function* ***D*** = *dataset*

A. Nayebi, *et al.*
Explaining Heterogeneity in Medial Entorhinal Cortex with Task-Driven Neural Networks.
NeurIPS 2021 (spotlight)

"plasticity mechanism"

3.

Learning Rule

$$\underset{a \in \mathcal{A}}{\operatorname{argmin}}[L(p_a^*)]$$

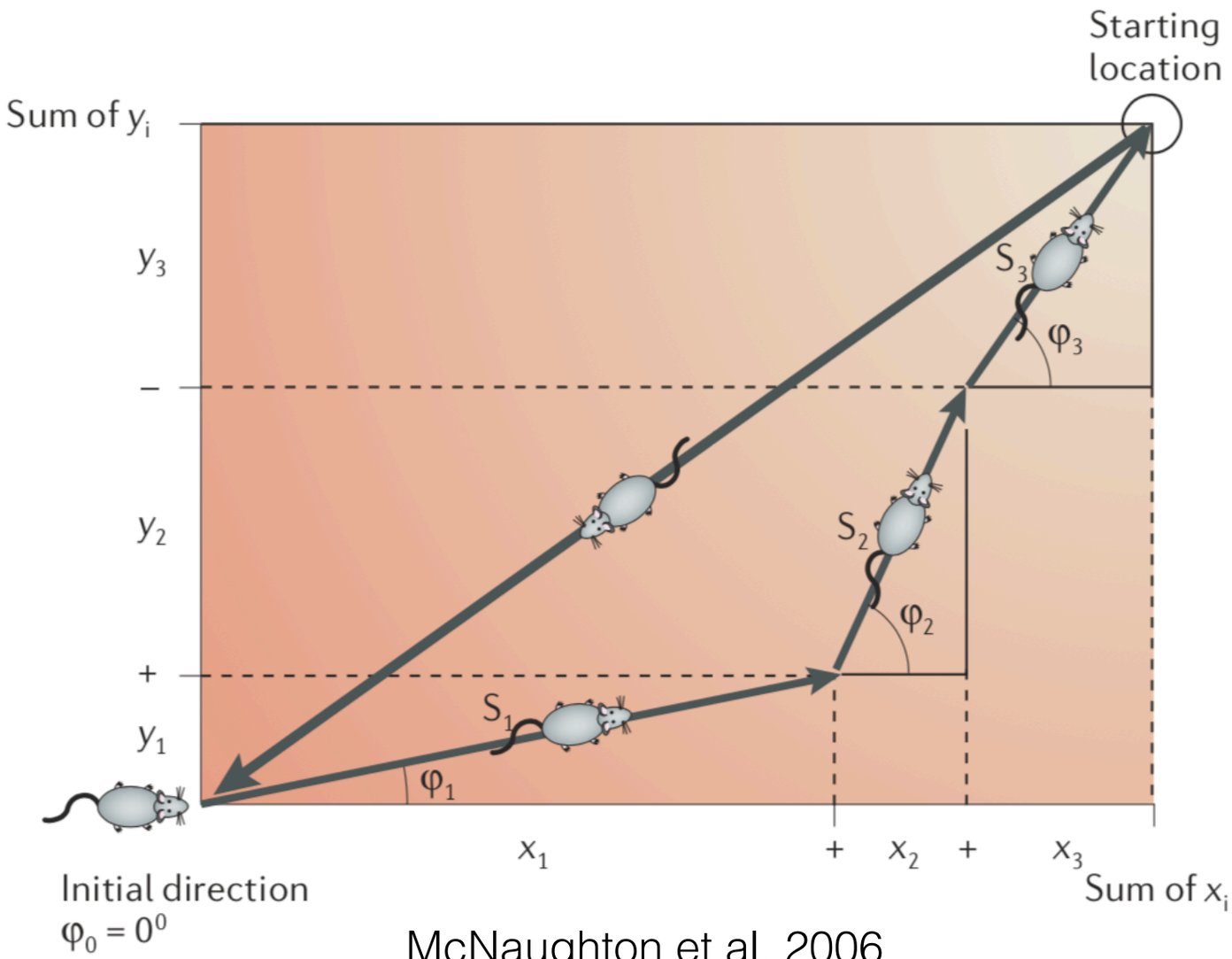
genetic algorithms

where p is result of*

backprop

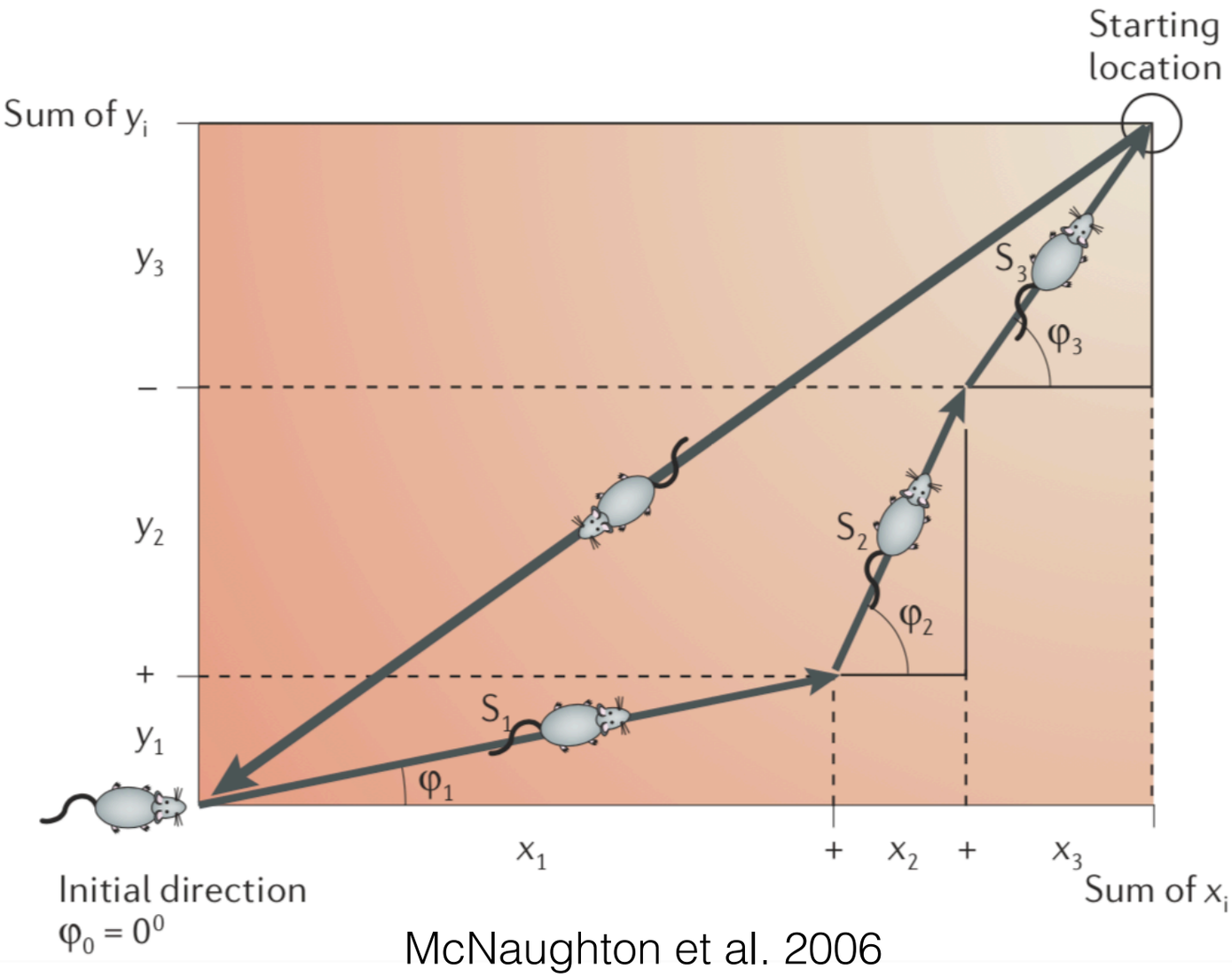
$$\frac{dp_a}{dt} = -\lambda(t) \cdot \langle \nabla_{p_a} L(x) \rangle_{x \in \mathcal{D}}$$

Hippocampal-Entorhinal Spatial Map



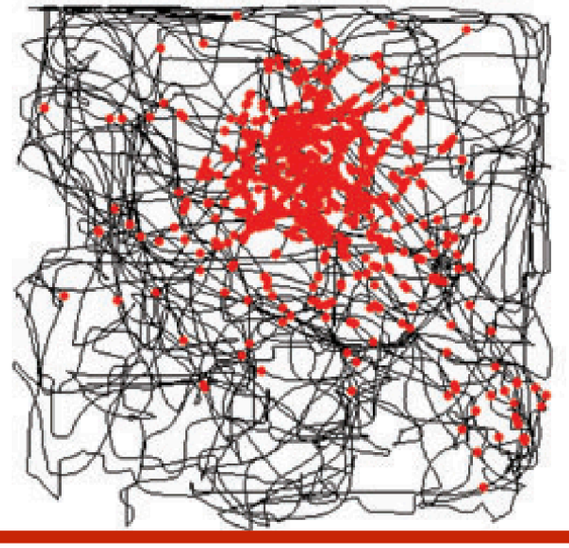
McNaughton et al. 2006

Hippocampal-Entorhinal Spatial Map

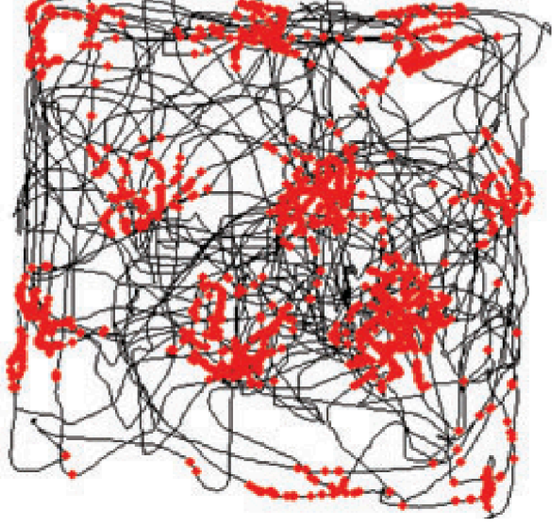


McNaughton et al. 2006

Place Cell (HPC)

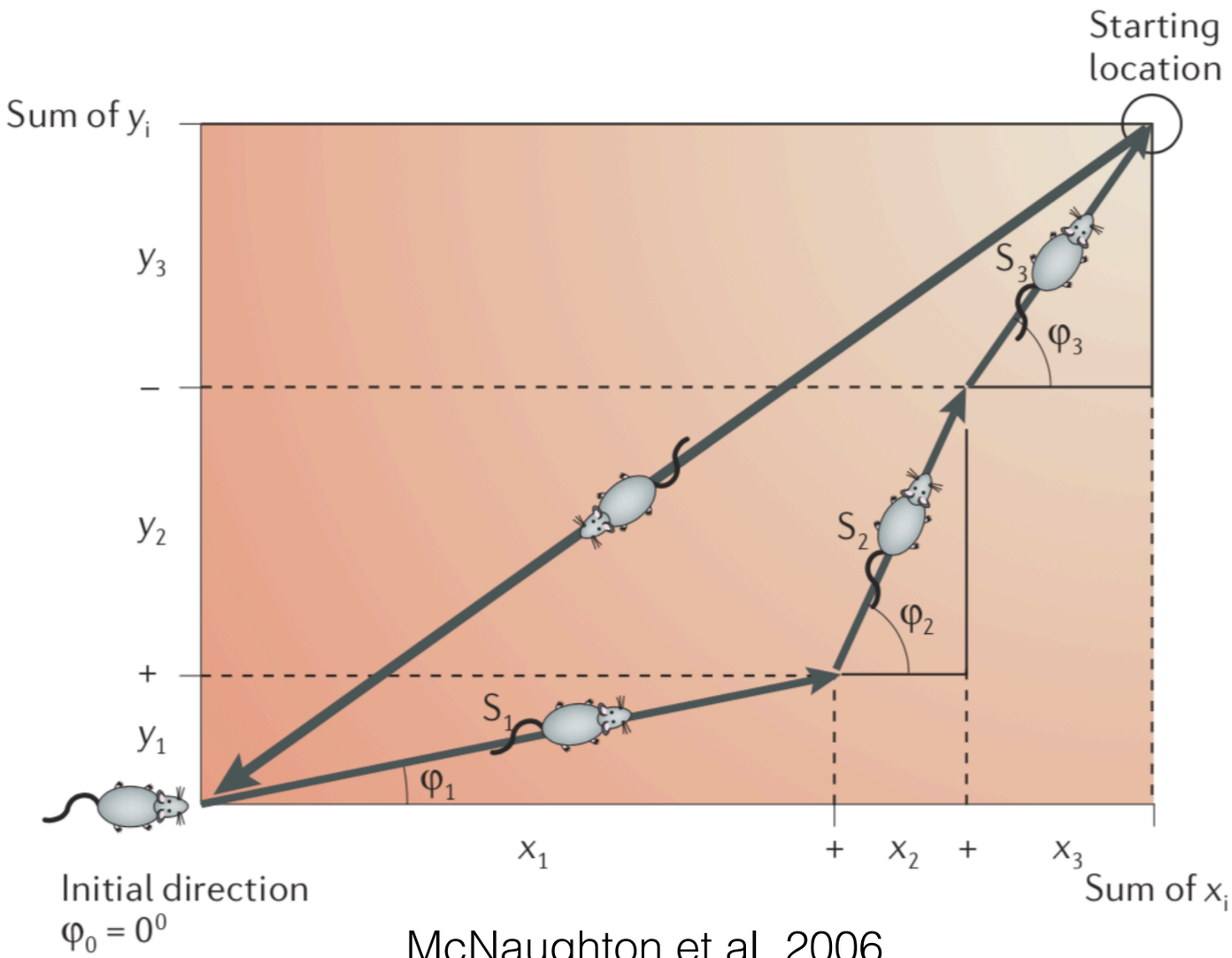


Grid Cell (MEC)



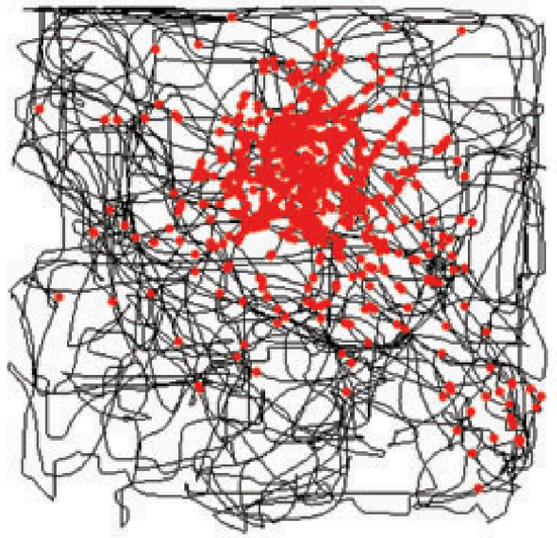
Moser et al. 2008

Hippocampal-Entorhinal Spatial Map

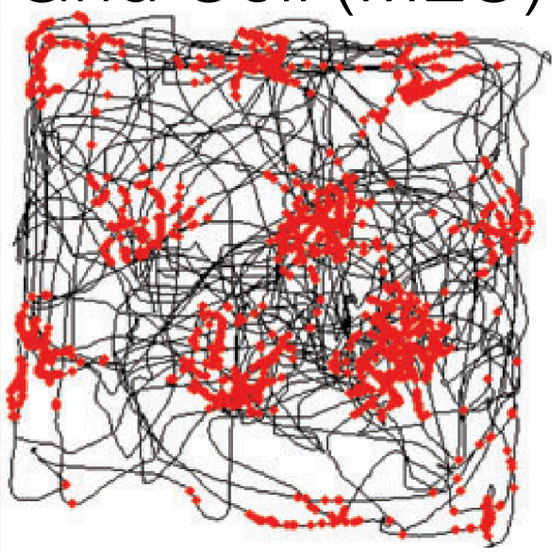


McNaughton et al. 2006

Place Cell (HPC)



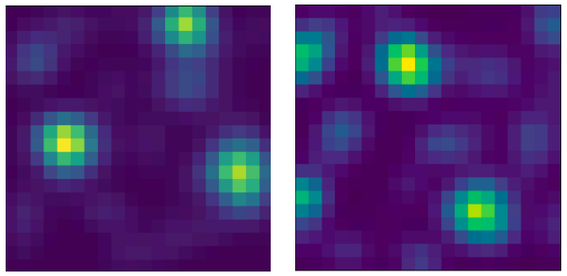
Grid Cell (MEC)



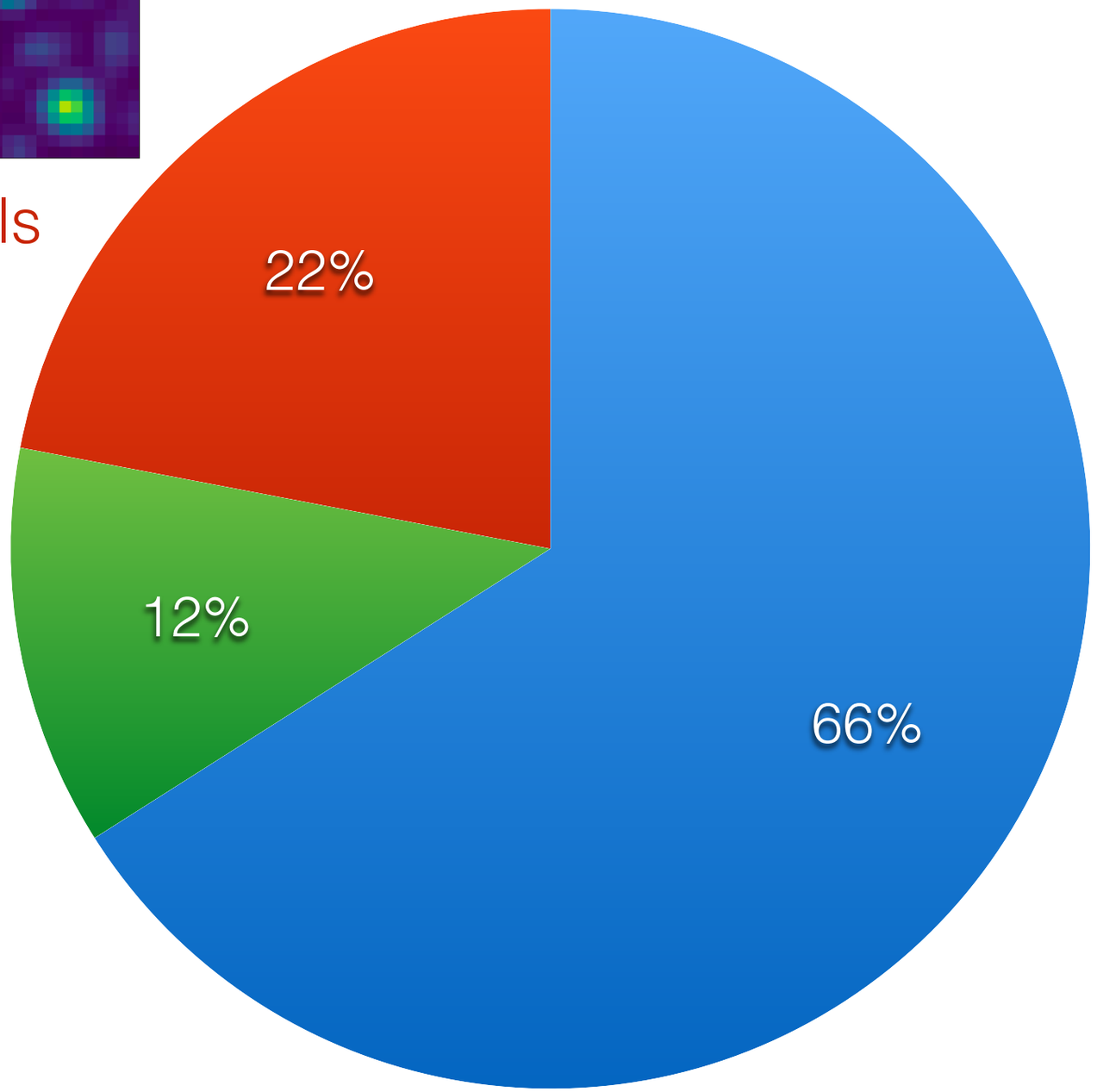
Moser et al. 2008

Accounting for heterogeneous code?

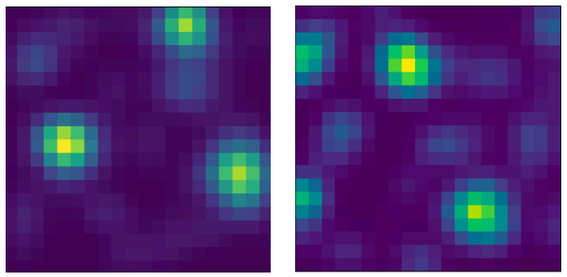
Accounting for heterogeneous code?



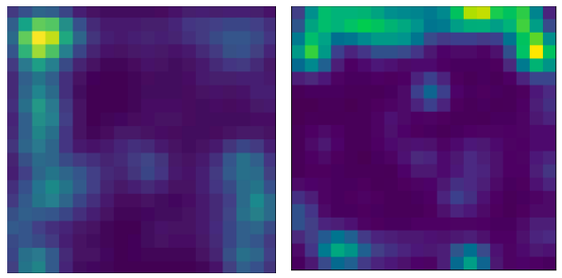
Grid Cells



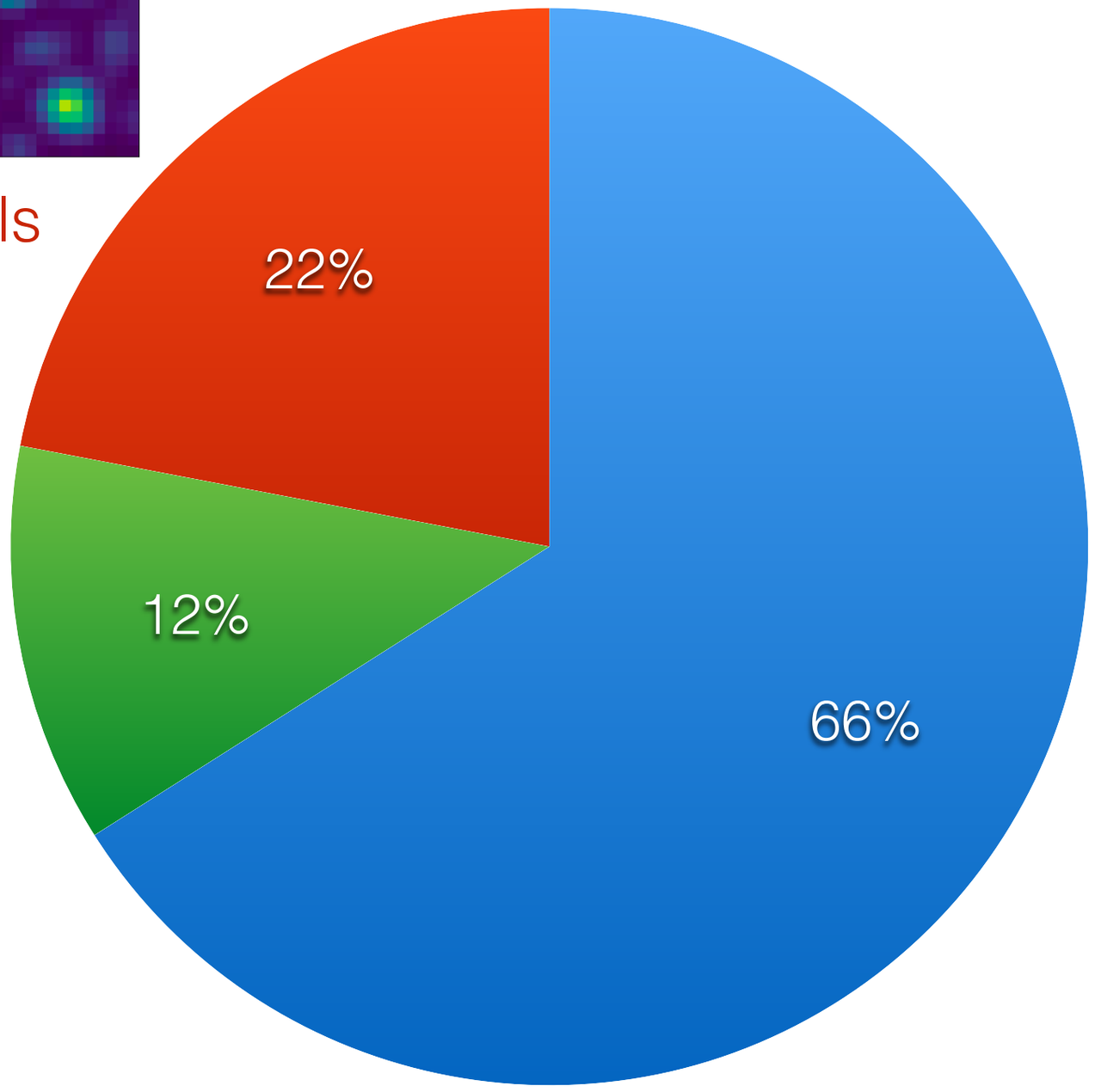
Accounting for heterogeneous code?



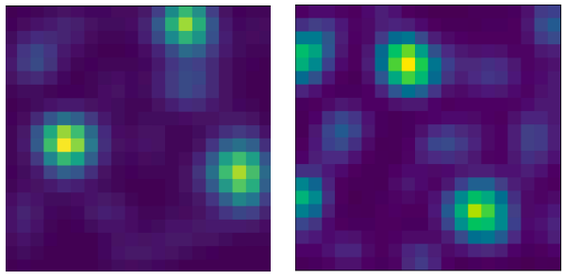
Grid Cells



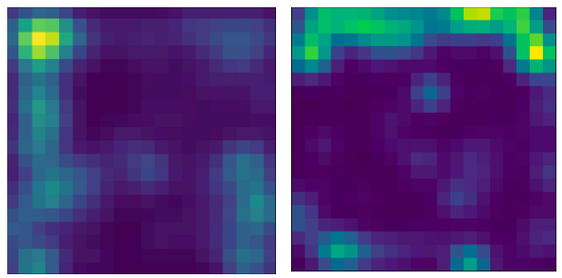
Border Cells



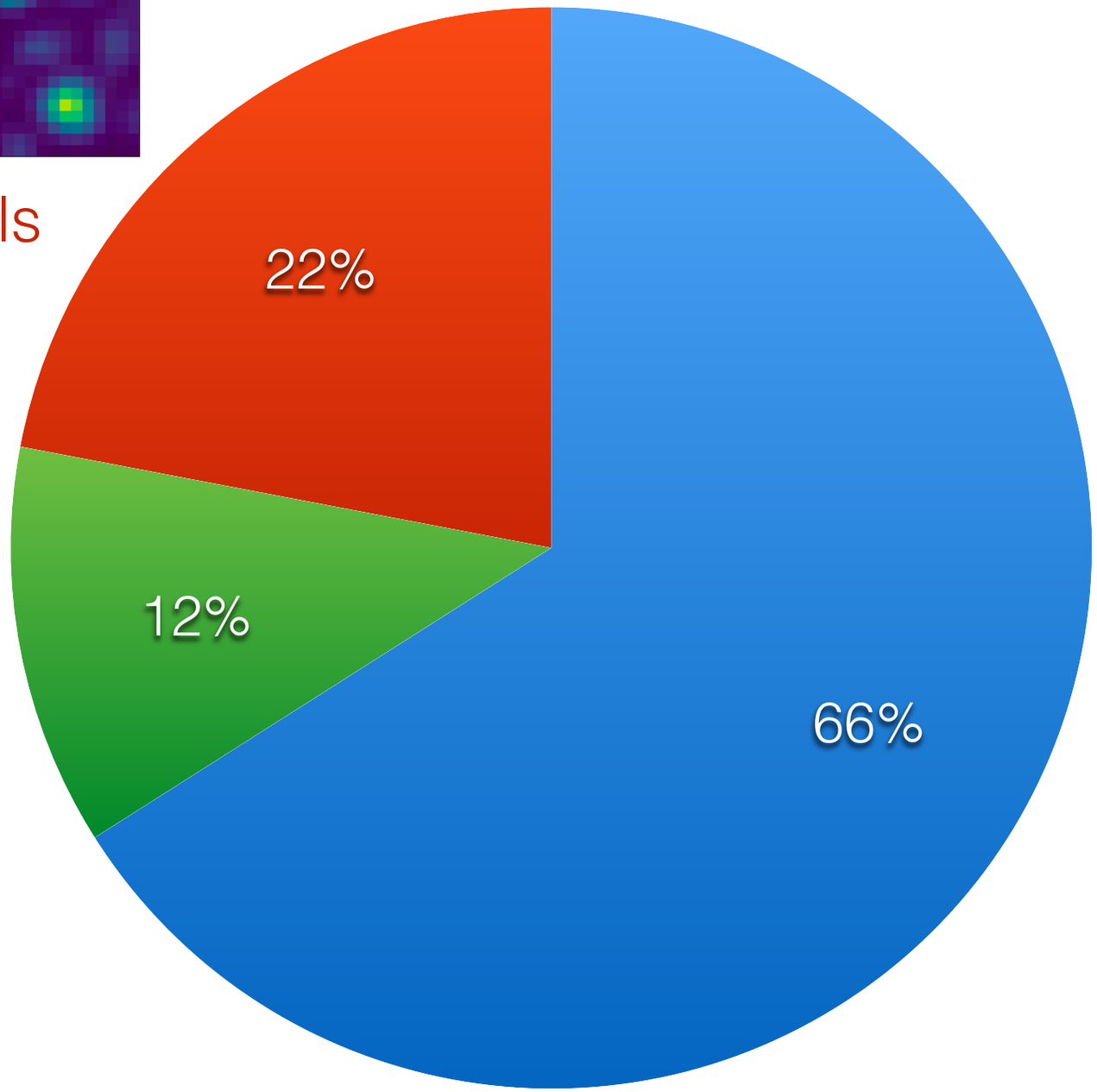
Accounting for heterogeneous code?



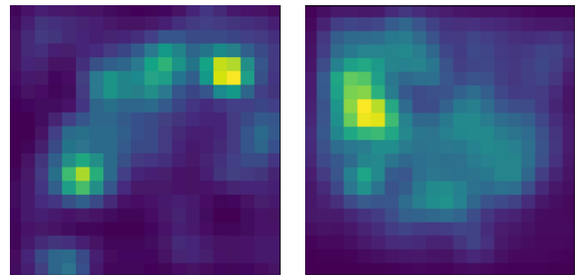
Grid Cells



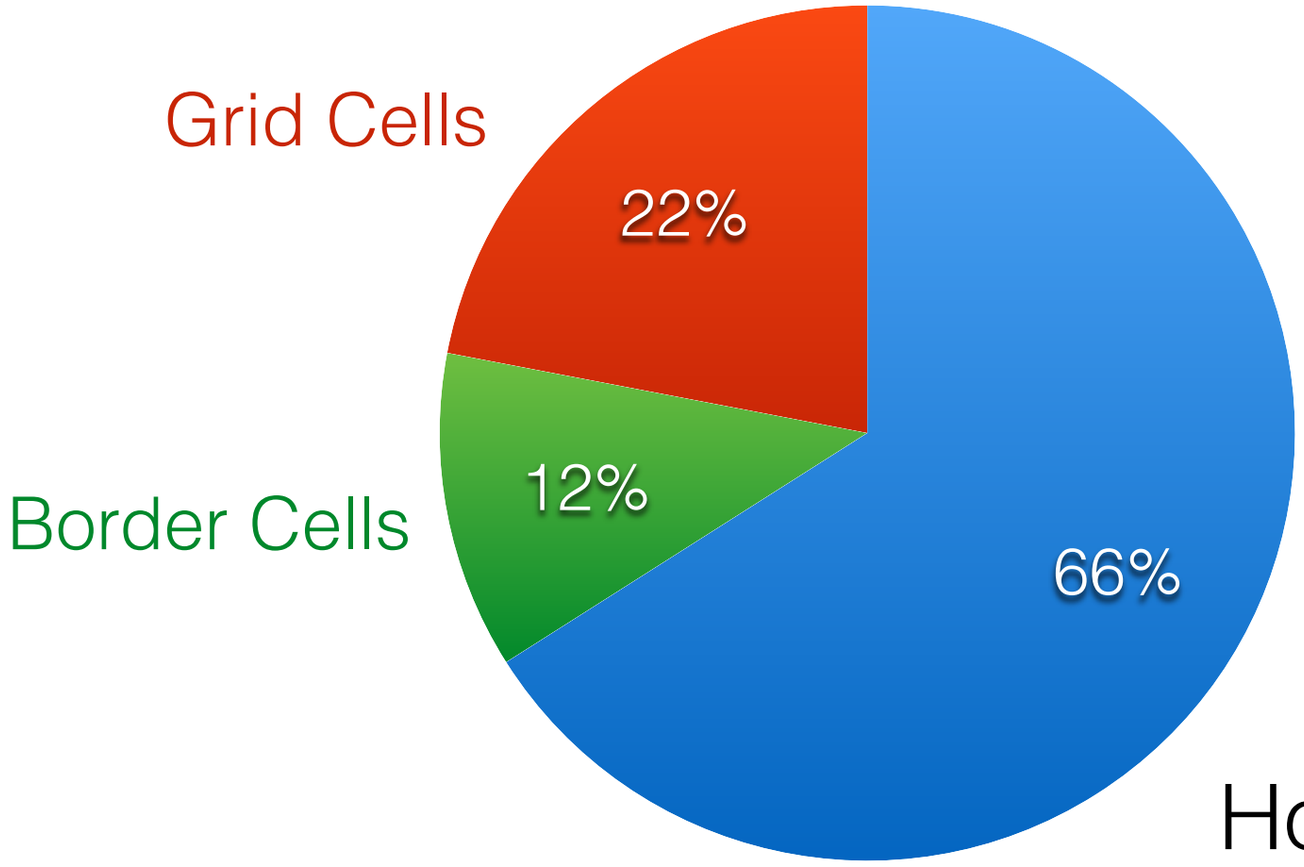
Border Cells



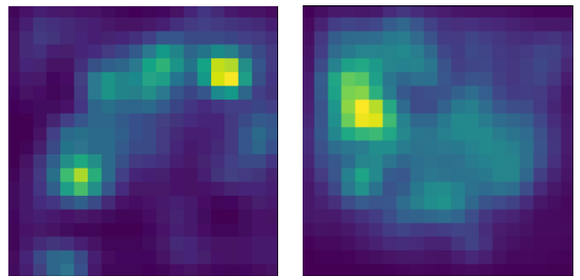
Heterogeneous Cells



Accounting for heterogeneous code?

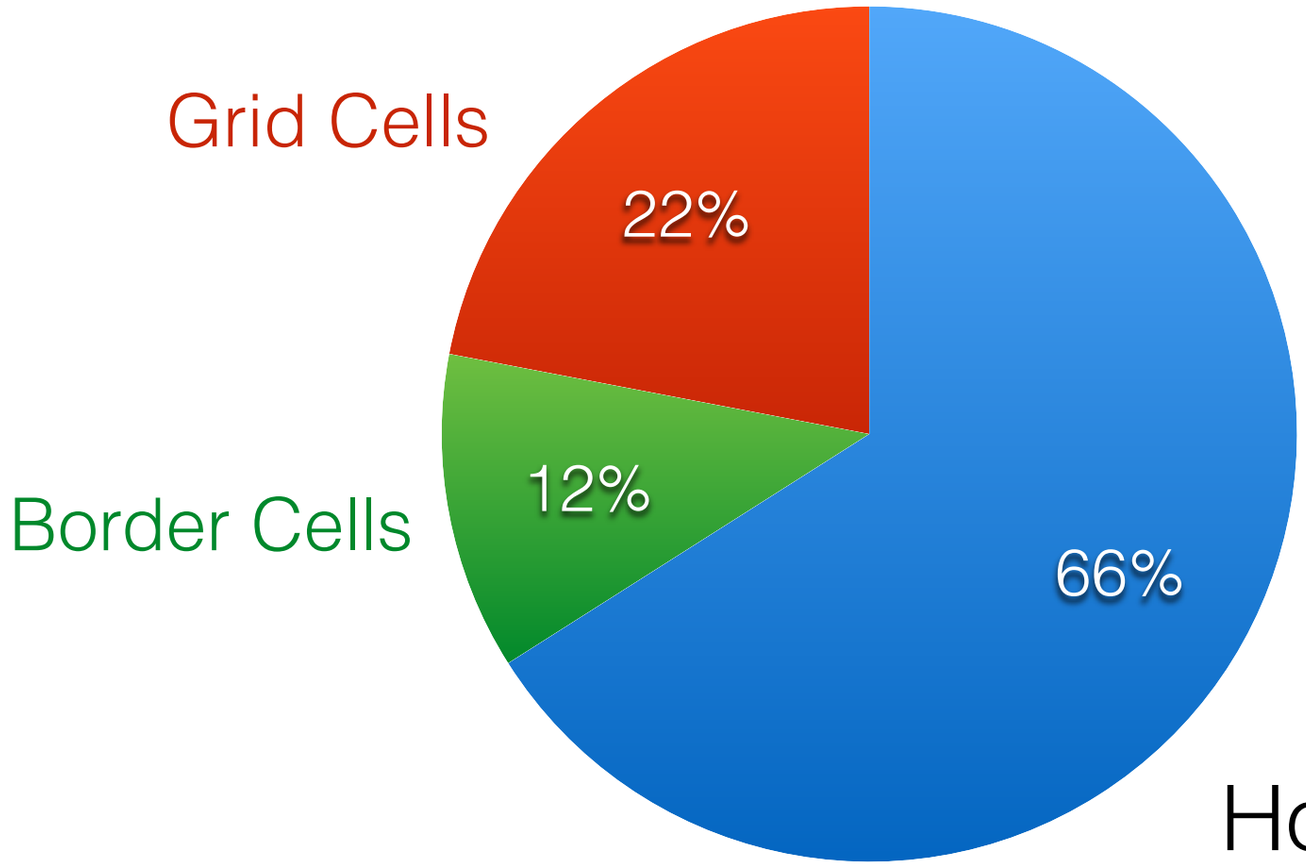


Heterogeneous Cells

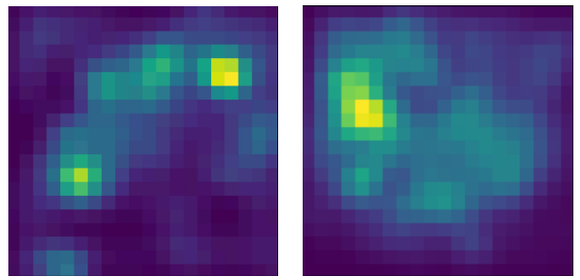


How might we characterize the response patterns of these heterogeneous cells?

Accounting for heterogeneous code?



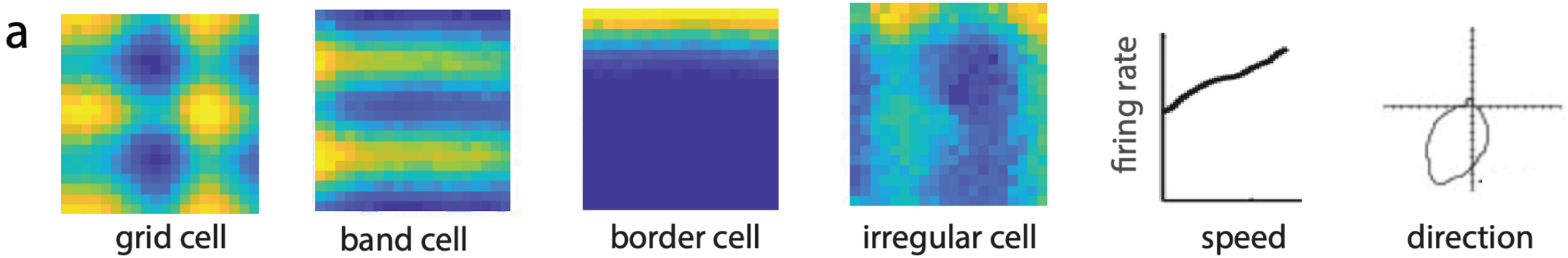
Heterogeneous Cells



How might we characterize the response patterns of these heterogeneous cells?

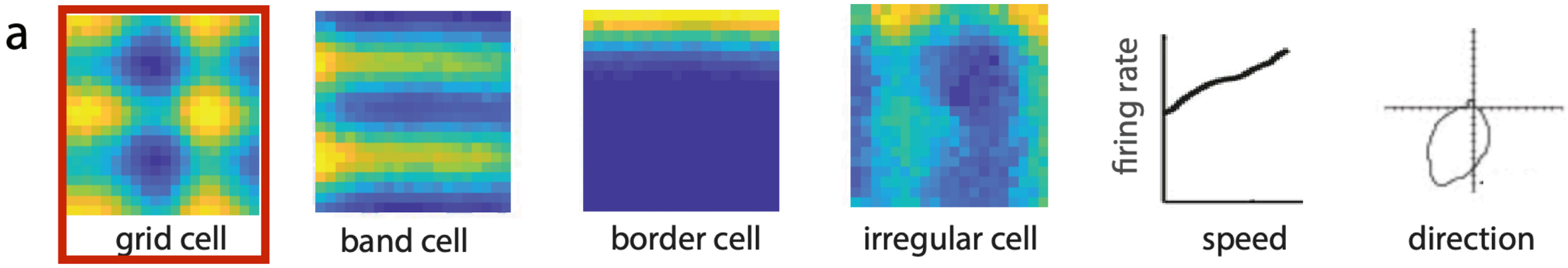
What functional role do these cells serve in the circuit, if any?

But more recently there are neural network models that "develop" these cells…



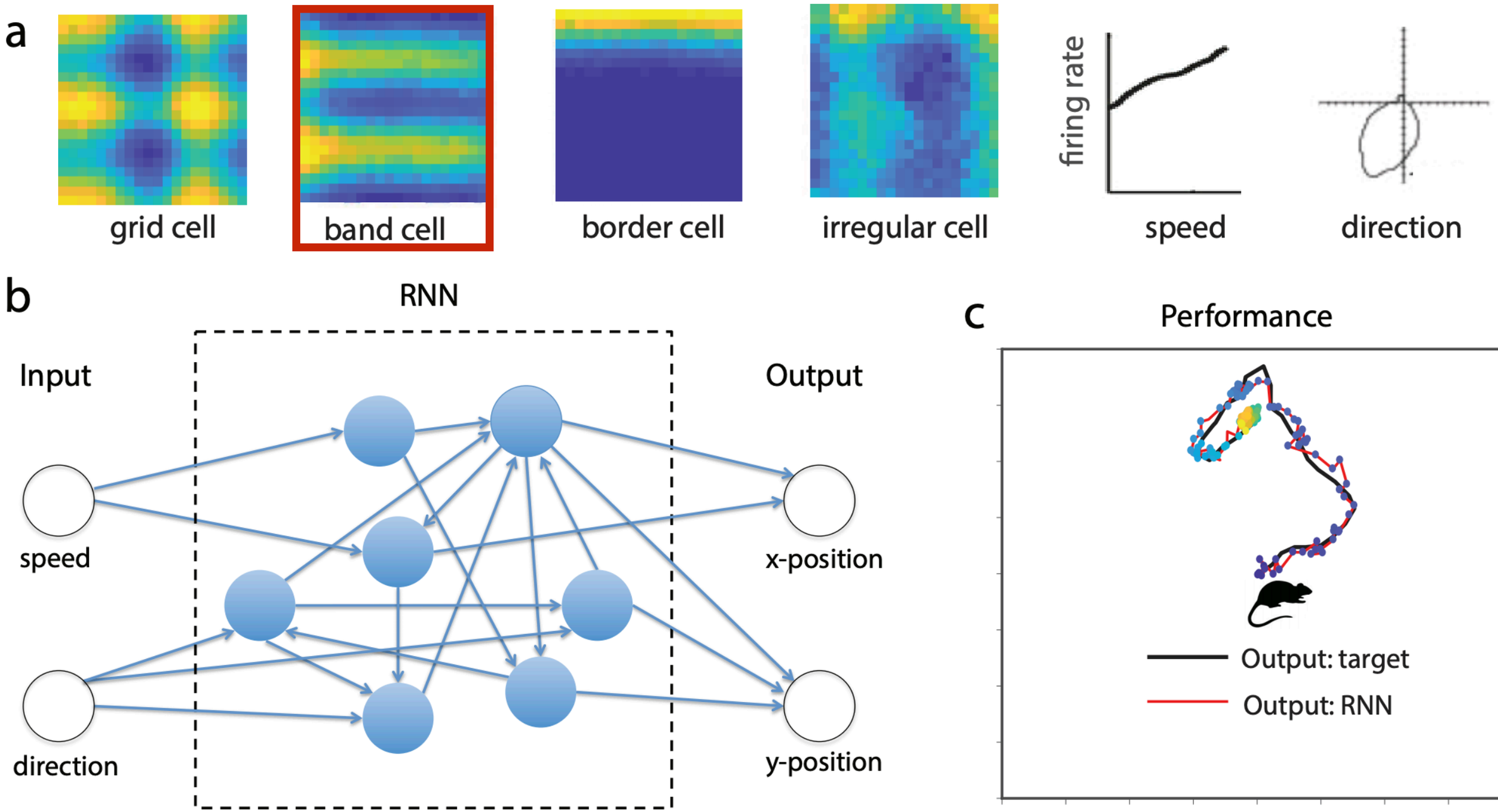
Cueva* & Wei* 2018

But more recently there are neural network models that "develop" these cells…



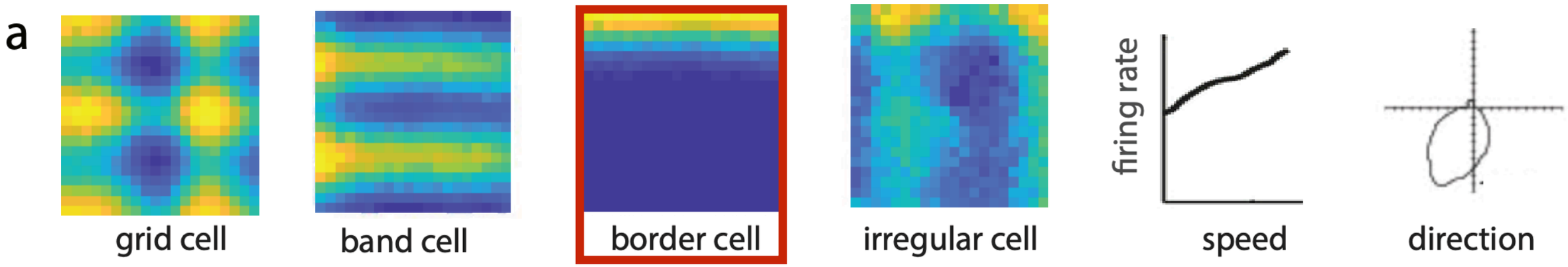
Cueva* & Wei* 2018

But more recently there are neural network models that "develop" these cells…



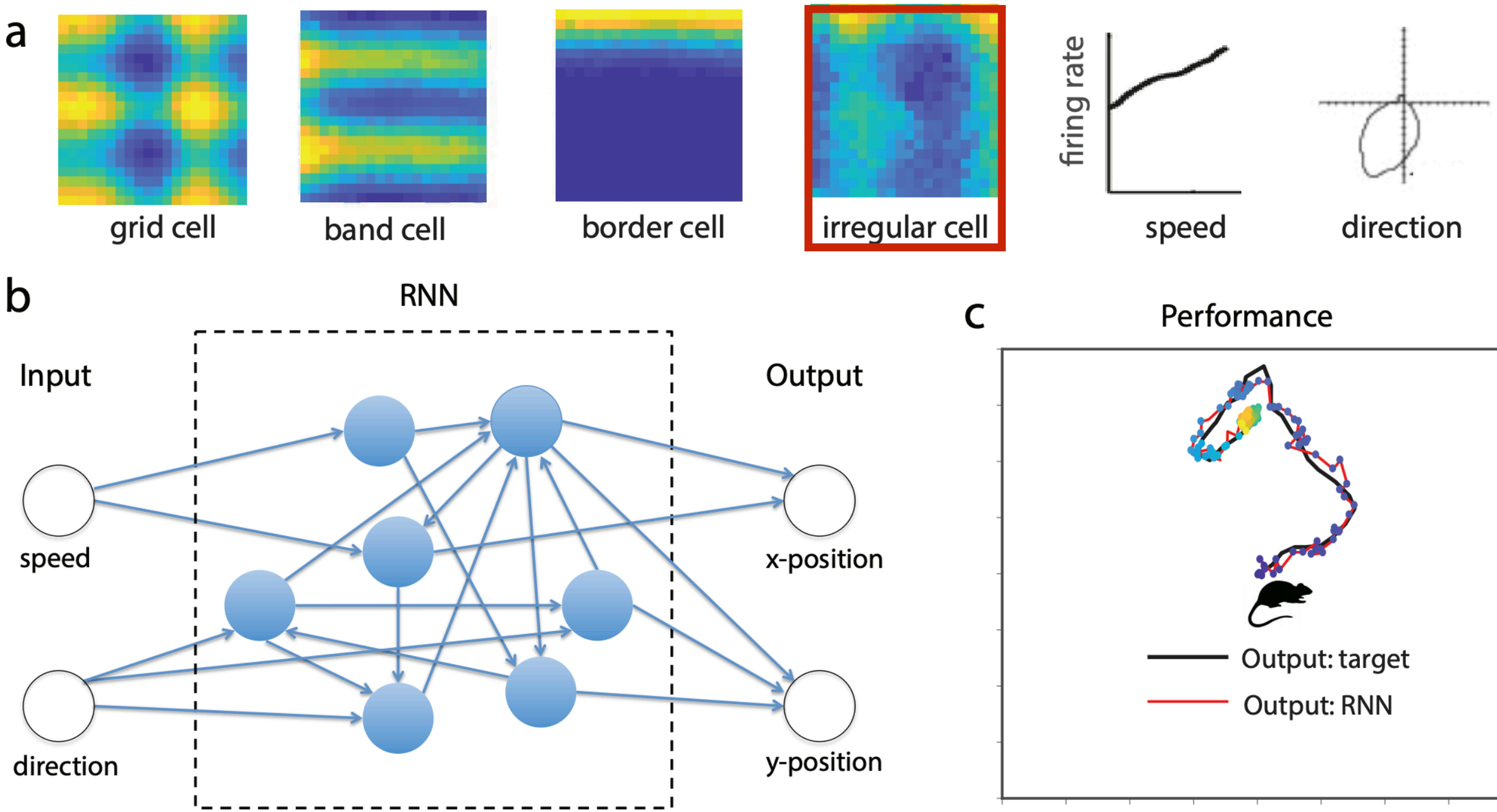
Cueva* & Wei* 2018

But more recently there are neural network models that "develop" these cells…



Cueva* & Wei* 2018

But more recently there are neural network models that "develop" these cells…

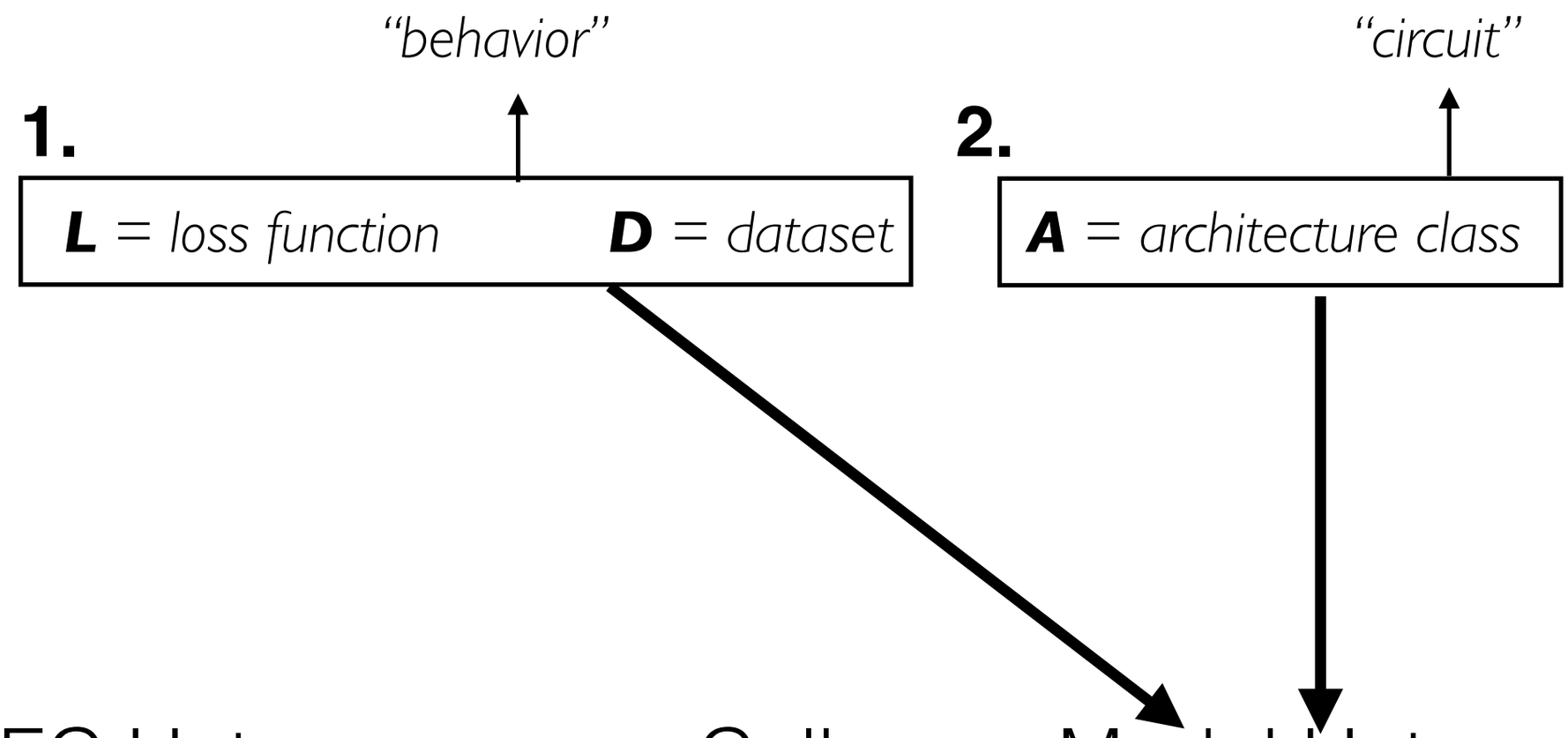


Cueva* & Wei* 2018

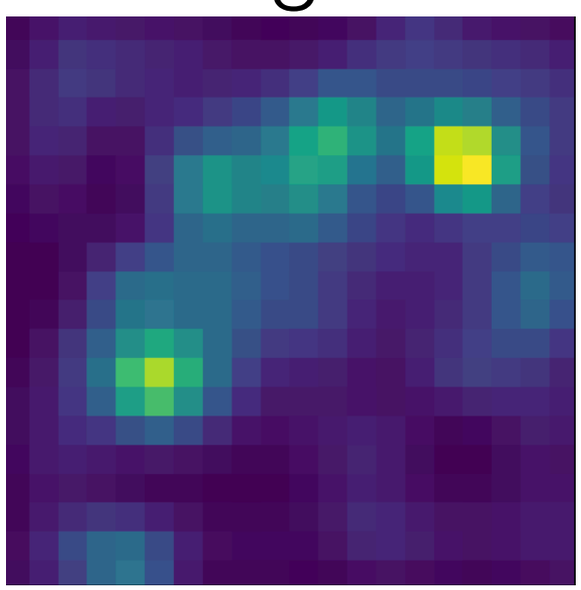
Main Questions

Can we use models to provide an *understanding* of the circuit?
What do we mean by “understanding”?

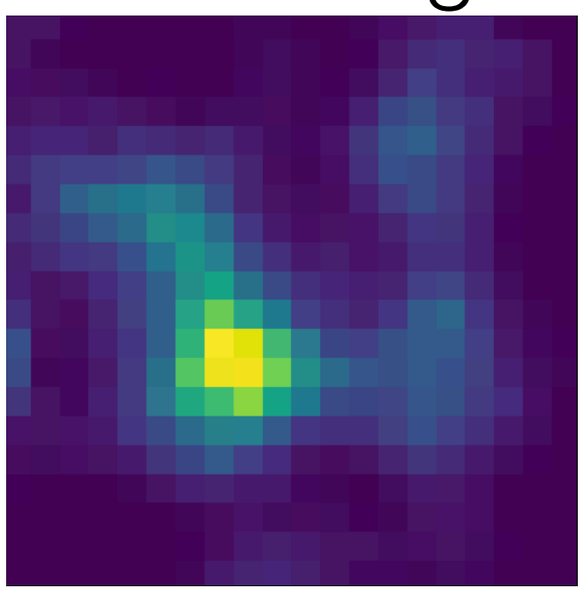
Goal-Driven Approach



MEC Heterogeneous Cell



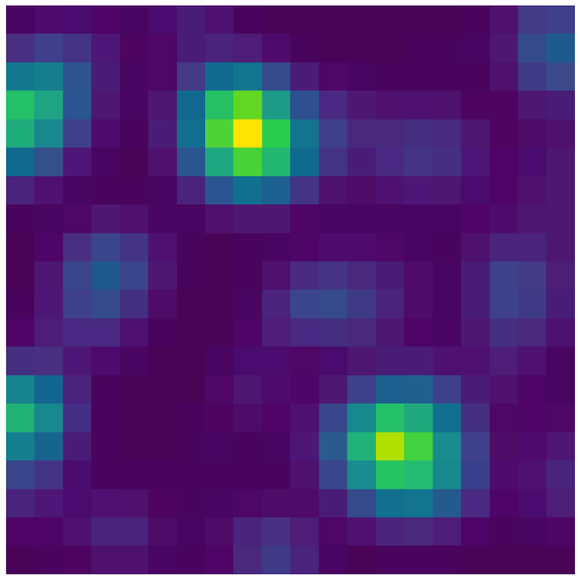
Model Heterogeneous Cell



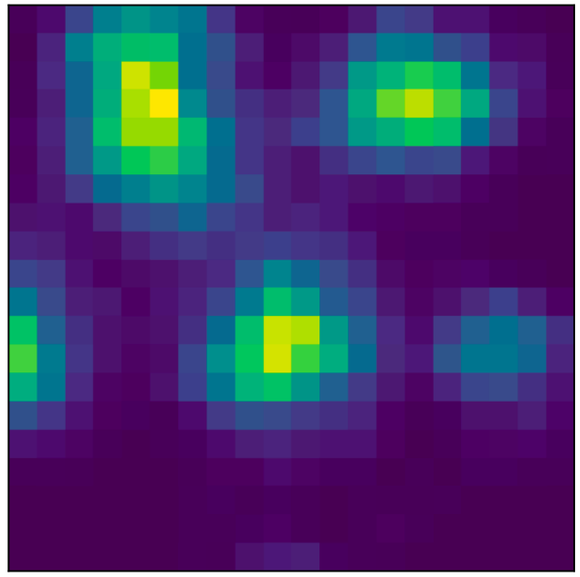
Main Questions

But are they a good ***quantitative*** model of these responses?

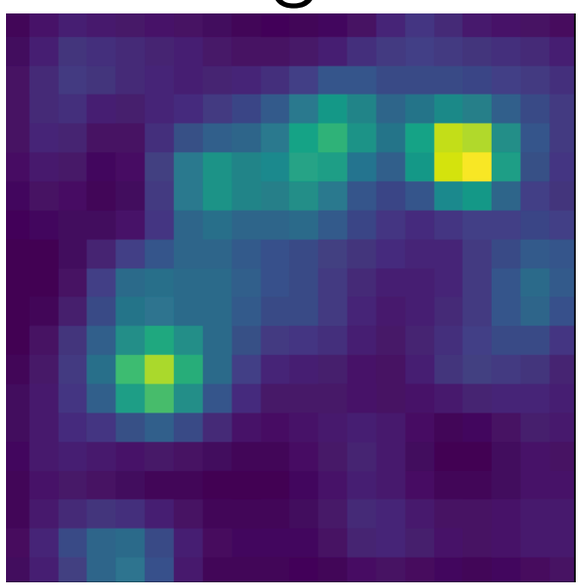
MEC Grid Cell



Model Grid Cell

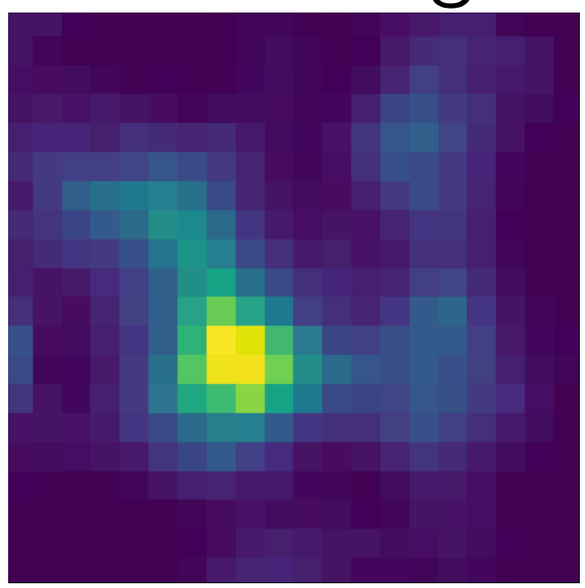


MEC Heterogeneous Cell



?

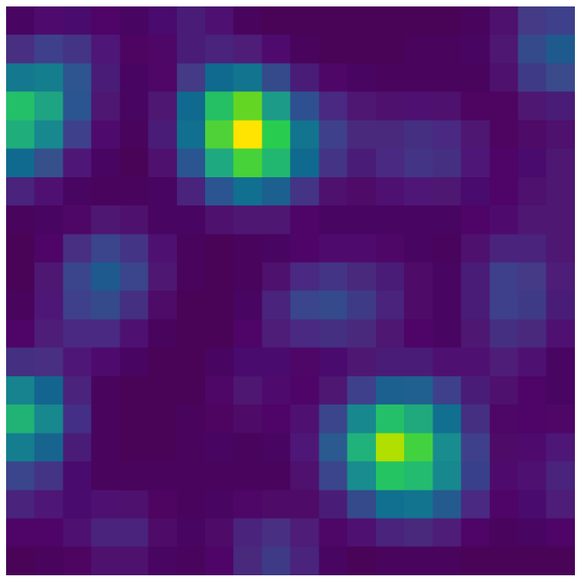
Model Heterogeneous Cell



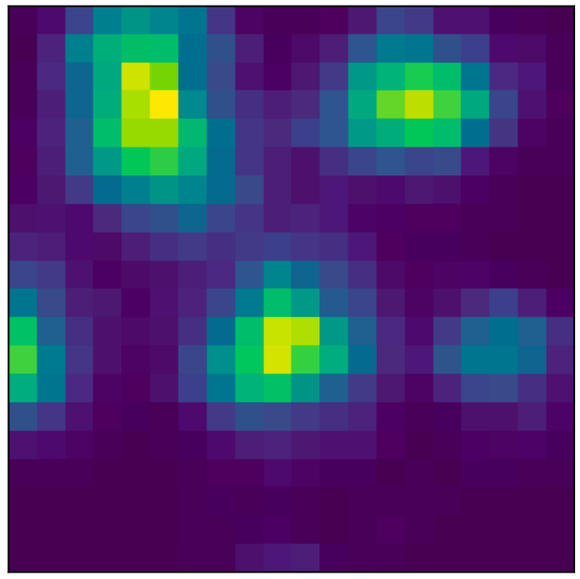
Main Questions

But are they a good ***quantitative*** model of these responses?

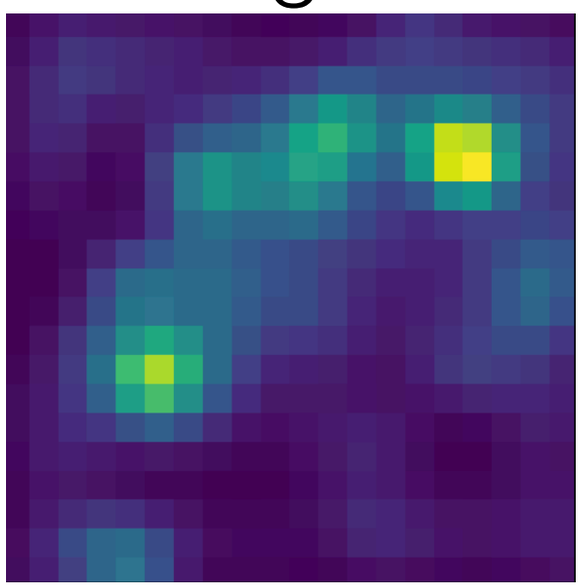
MEC Grid Cell



Model Grid Cell

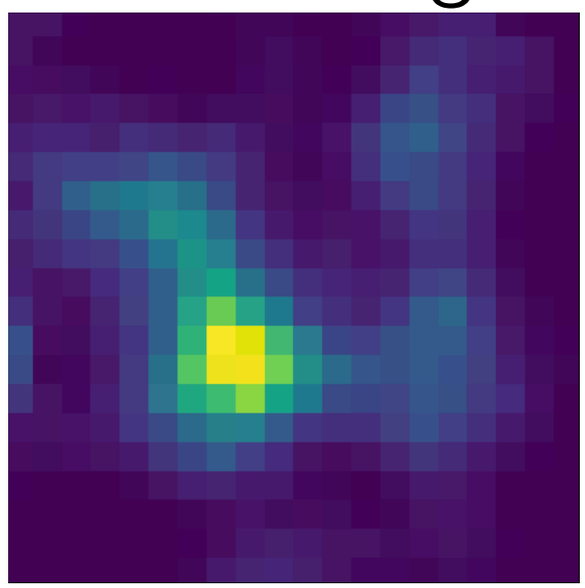


MEC Heterogeneous Cell



Not all models are equal!

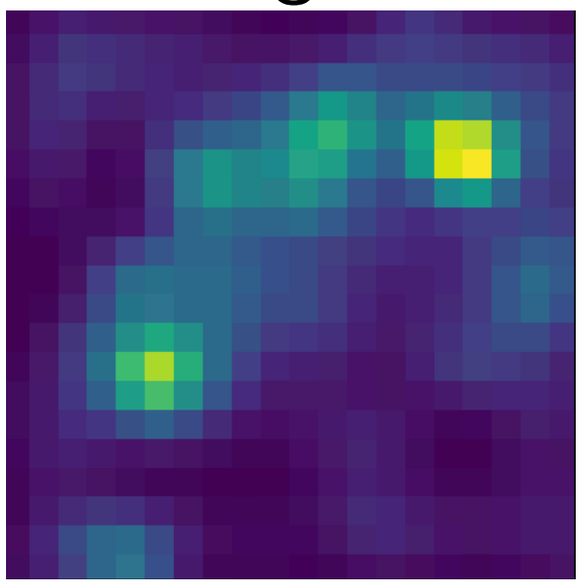
Model Heterogeneous Cell



Main Questions

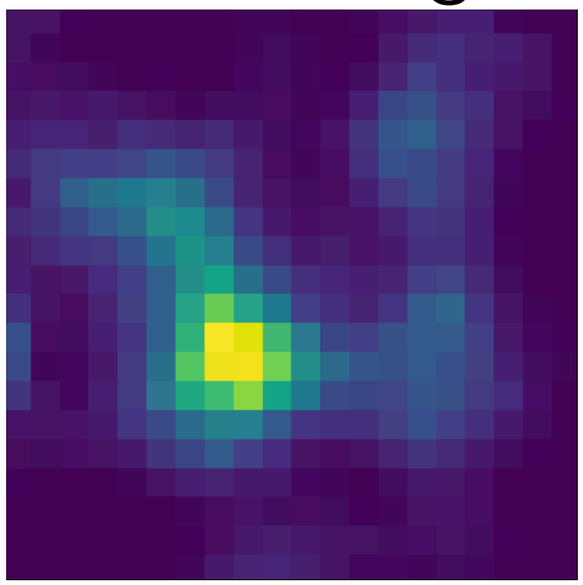
How do we define similarity between sets of heterogeneous responses we can't adequately describe in words?

MEC Heterogeneous Cell



?

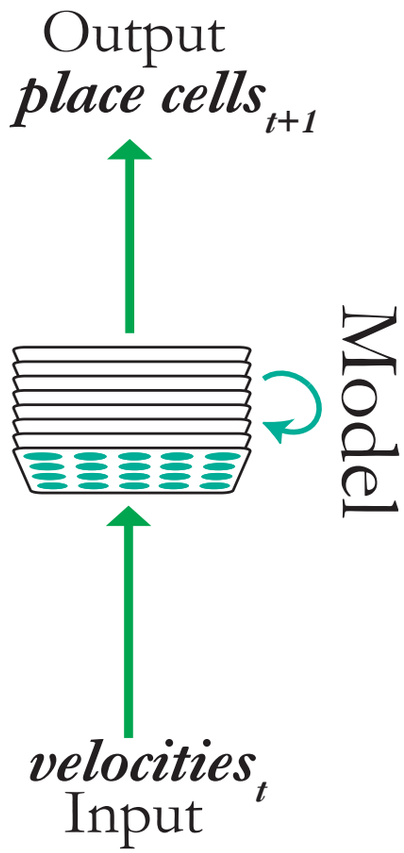
Model Heterogeneous Cell



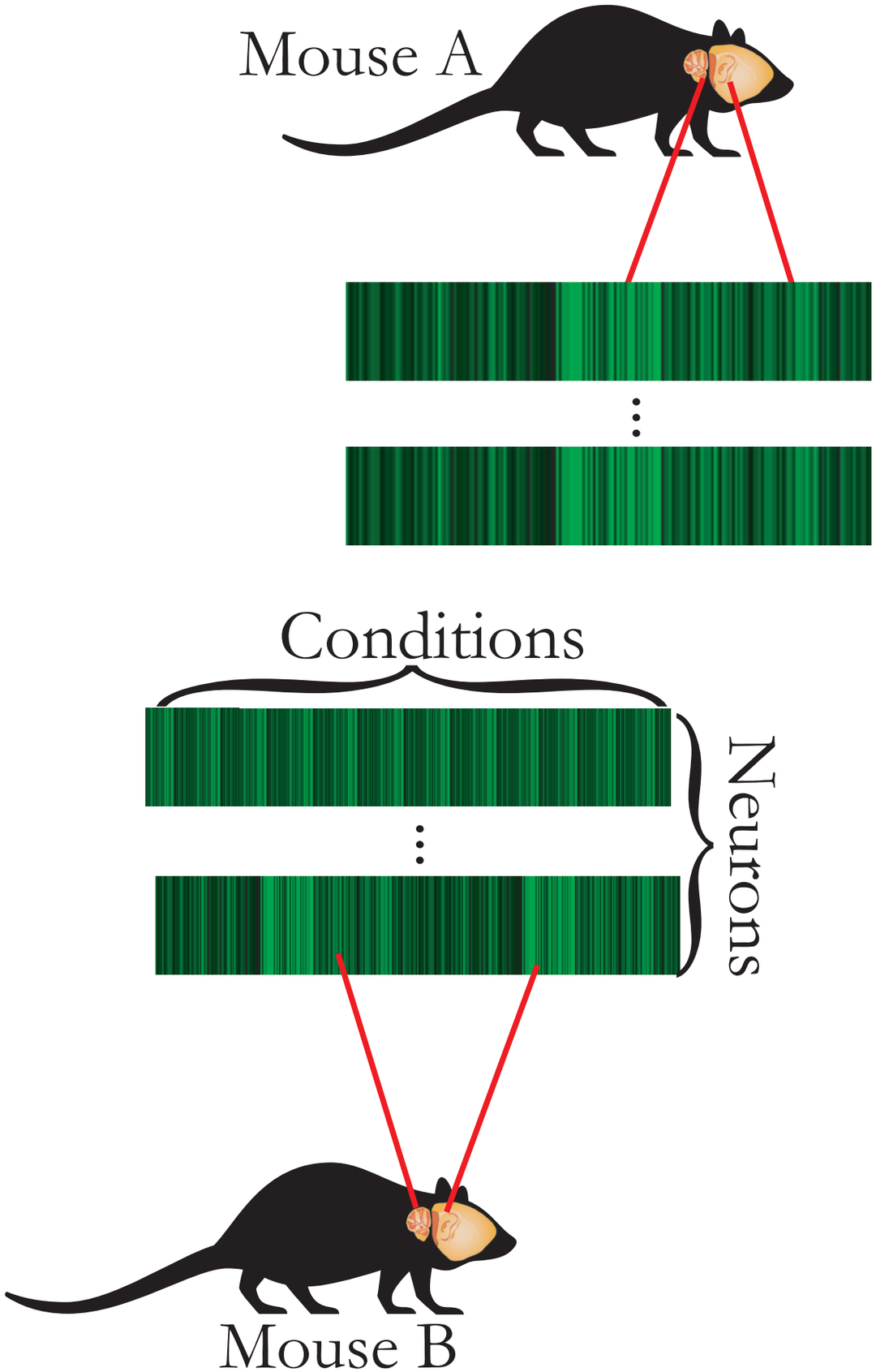
Overall procedure

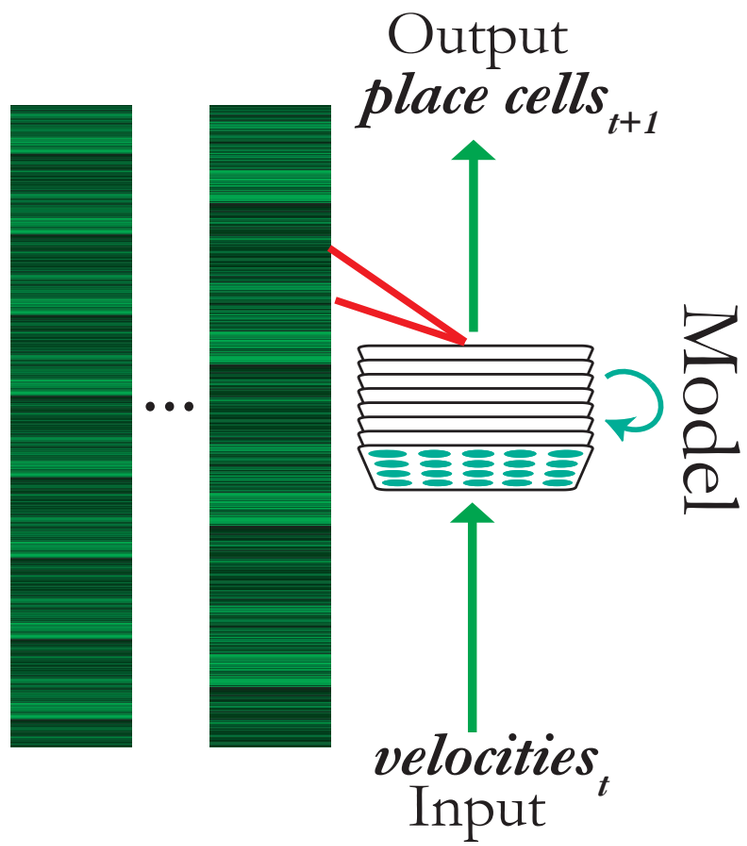
Our approach is that a model should be as similar to the system is unto itself.

Overall procedure

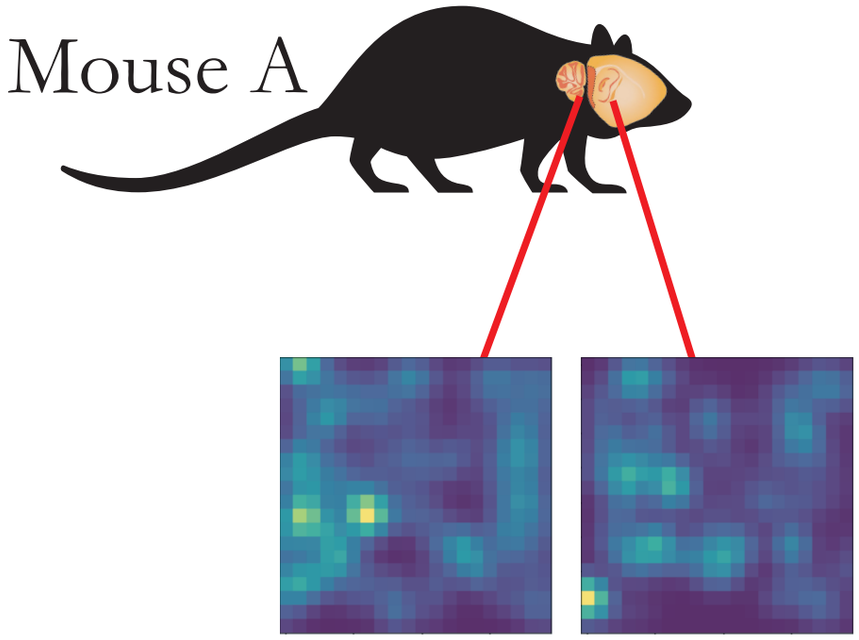


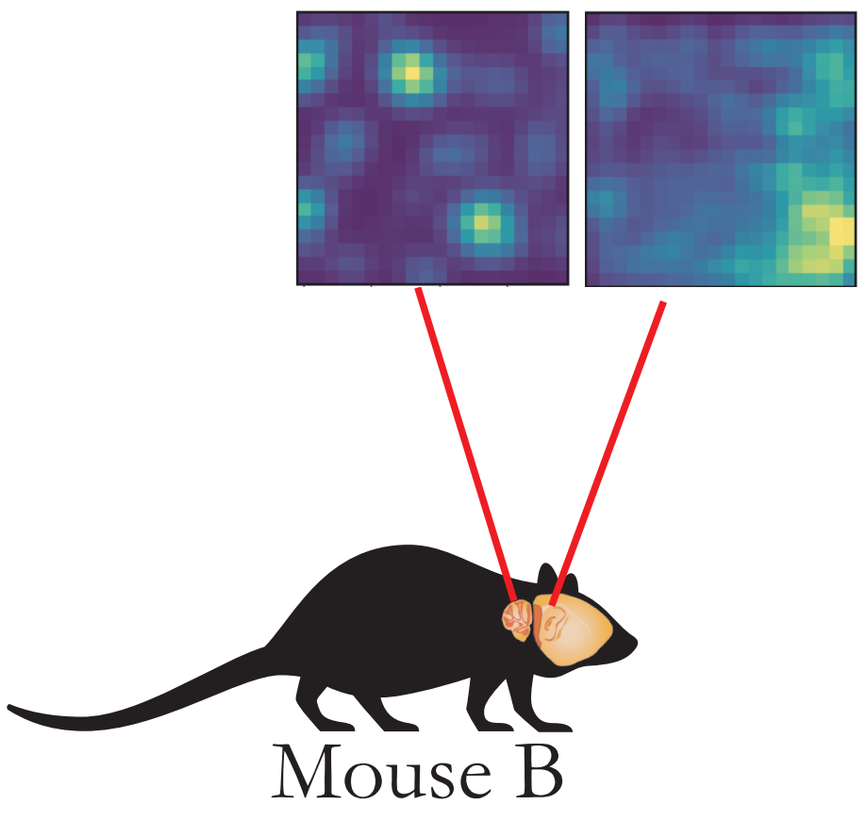
Overall procedure

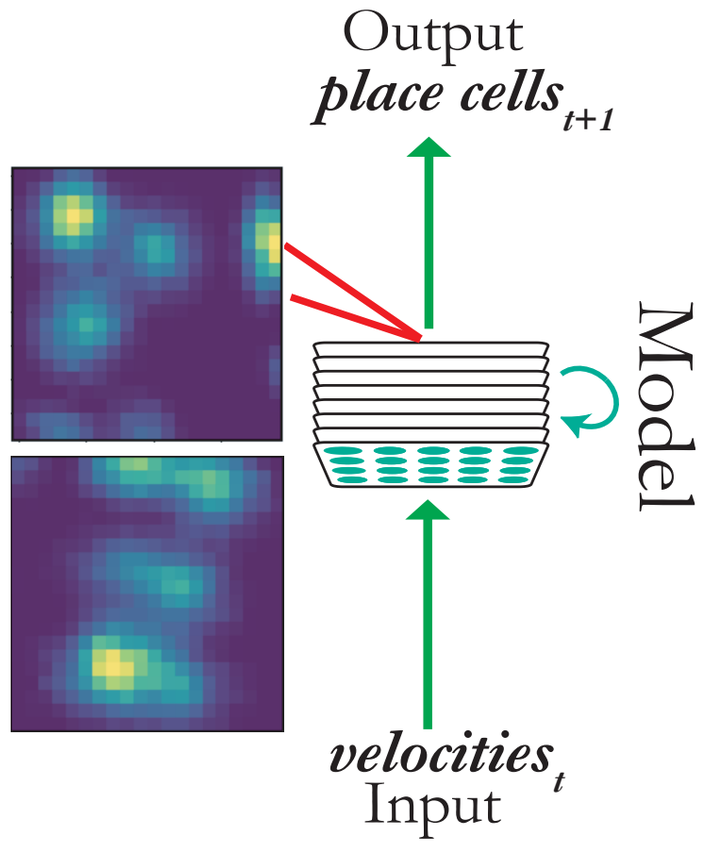




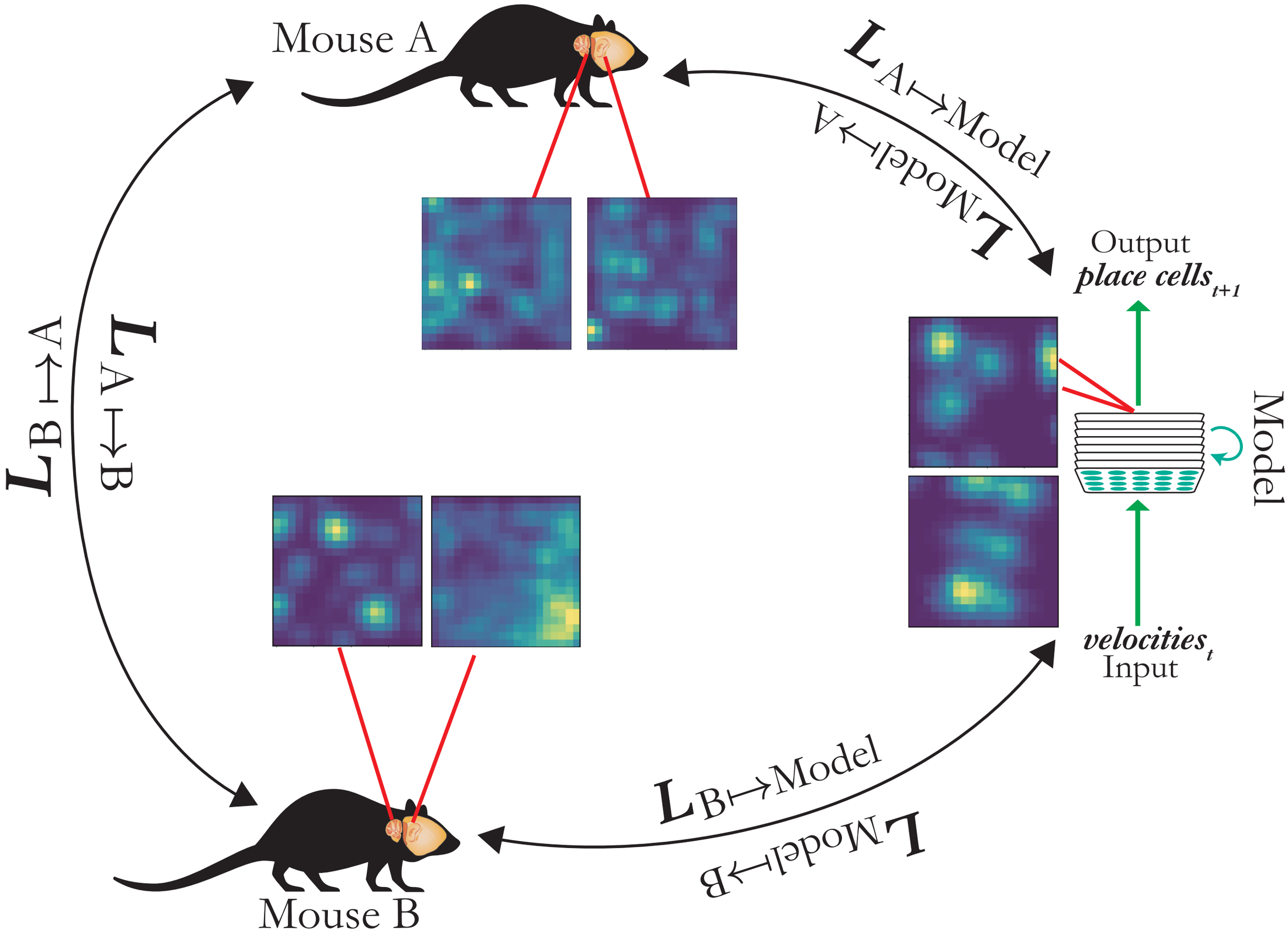
Overall procedure





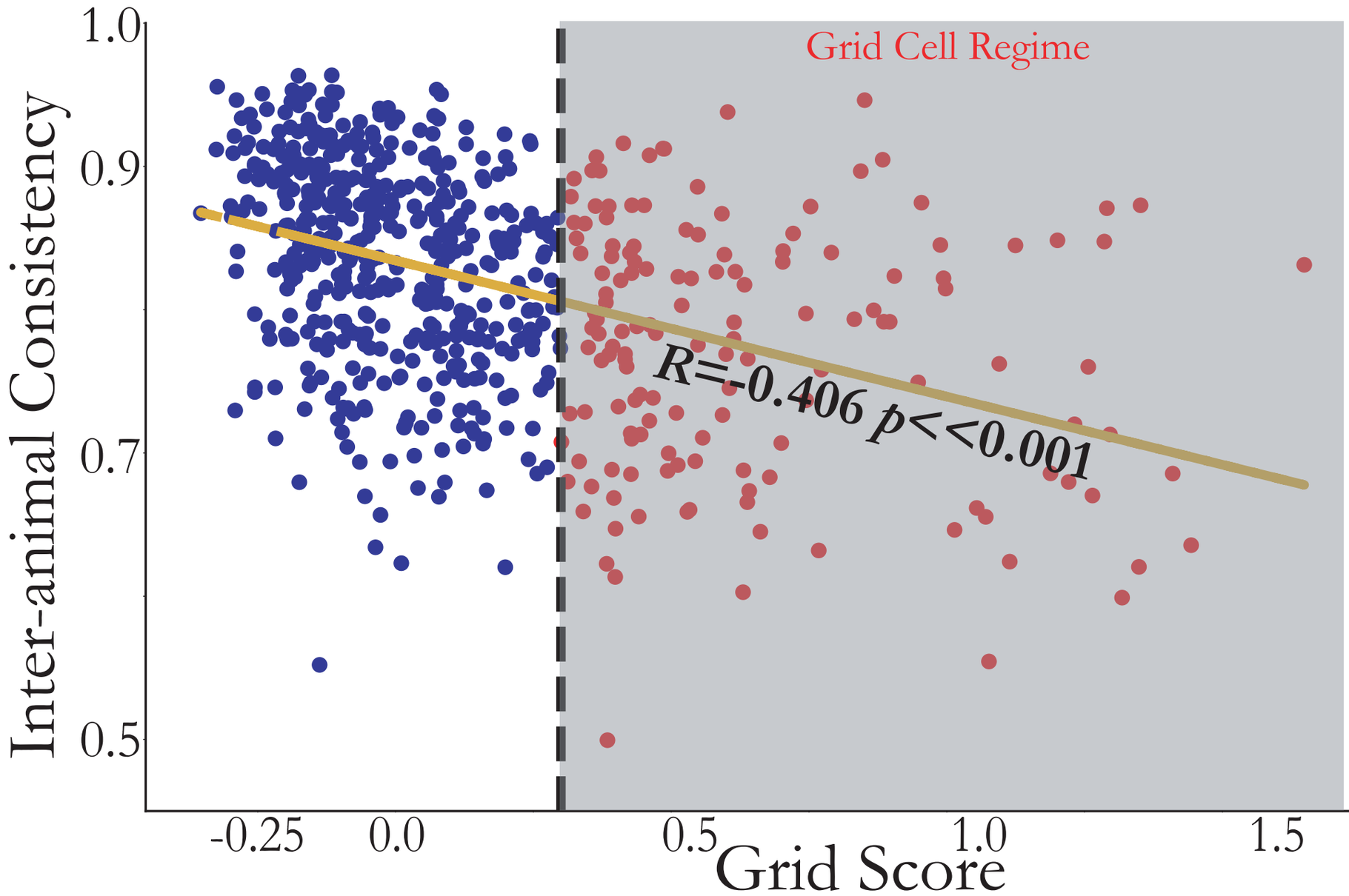


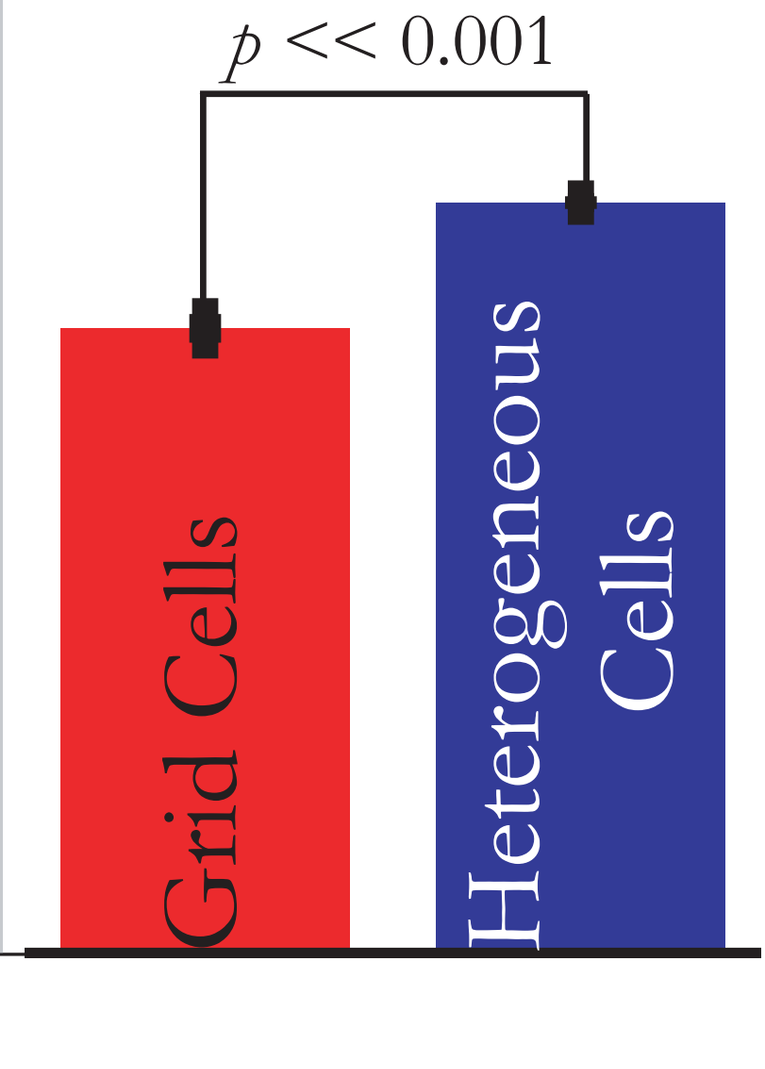
Overall procedure



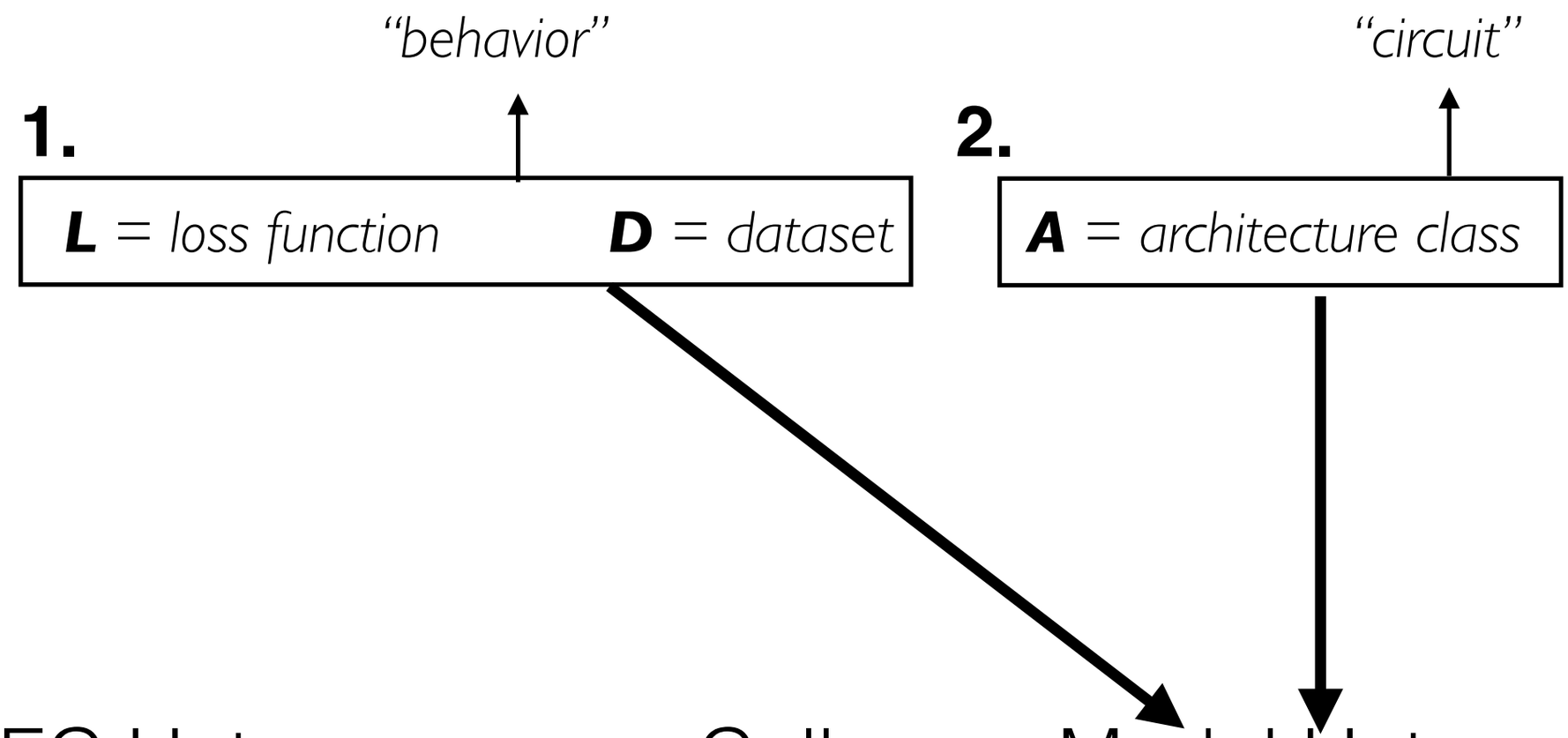
Heterogeneous cells are reliable targets of explanation

Consistent reliability *across* all cells

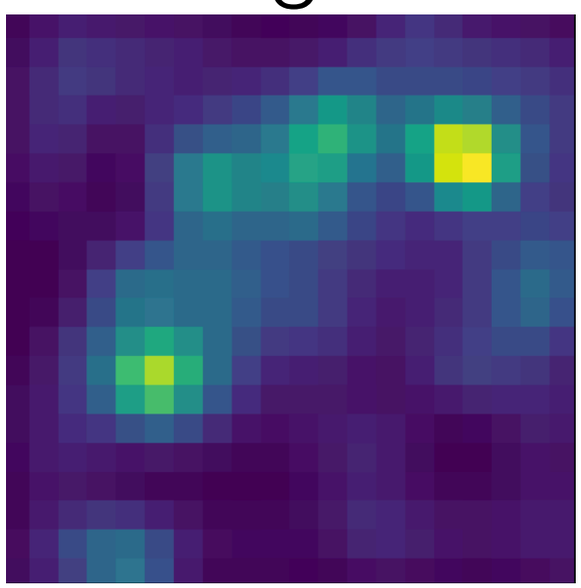




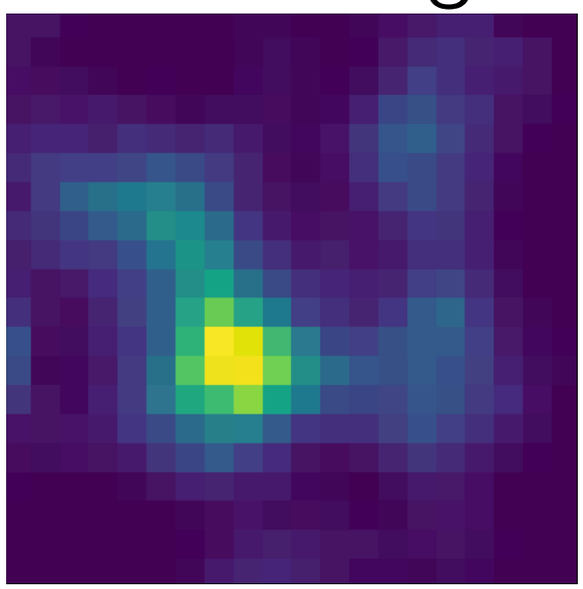
Goal-Driven Modeling - Primary Components



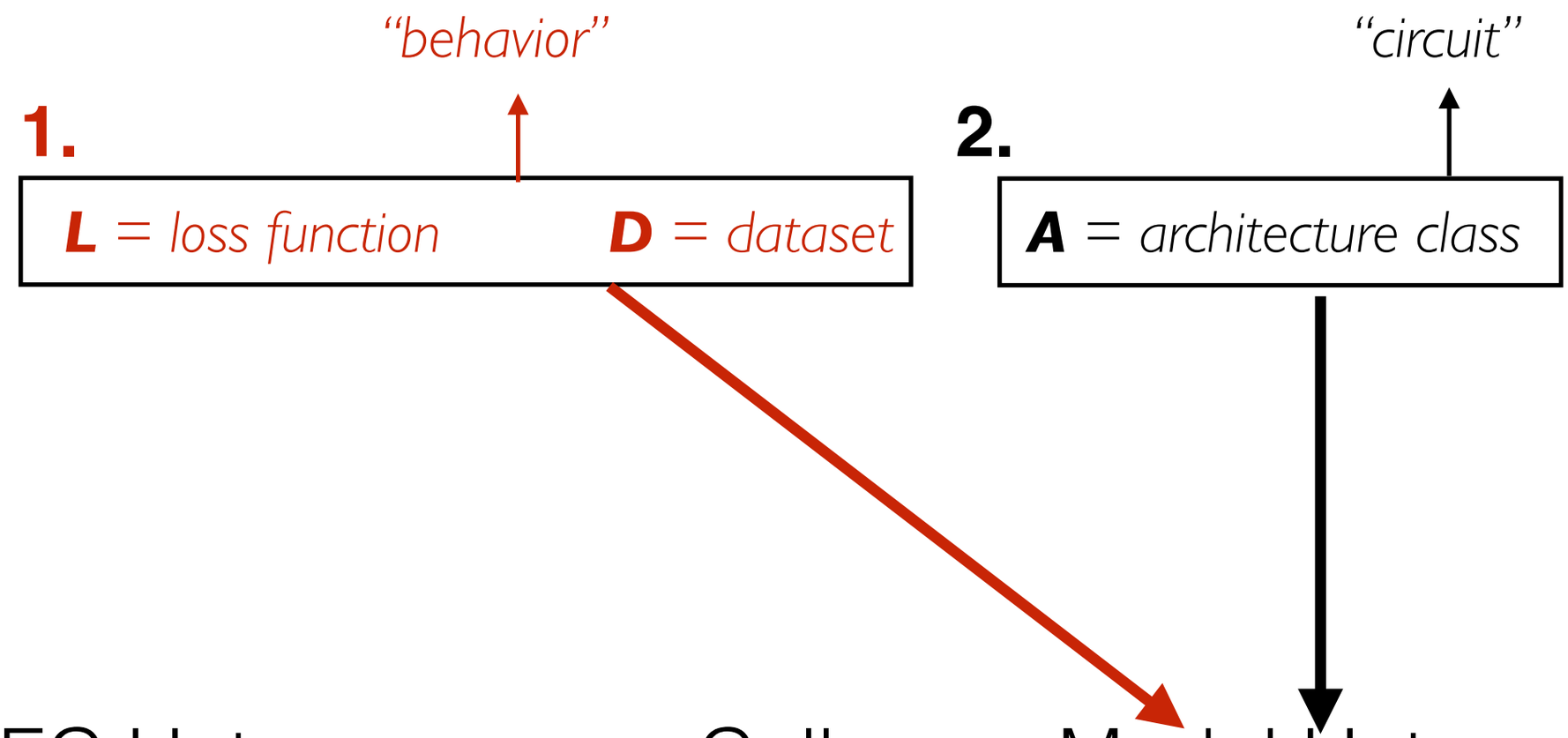
MEC Heterogeneous Cell



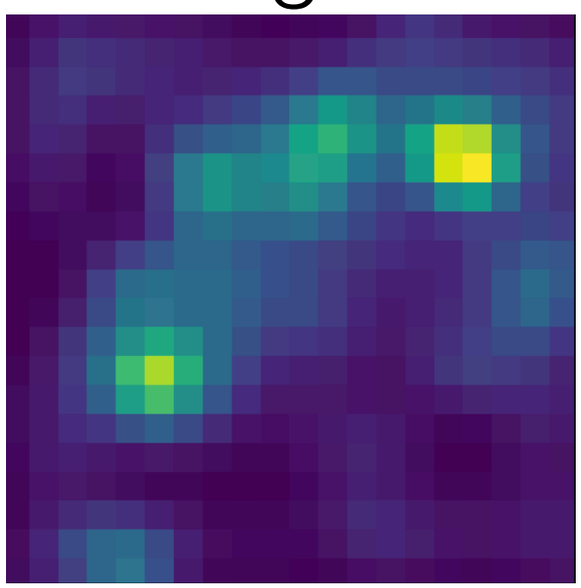
Model Heterogeneous Cell



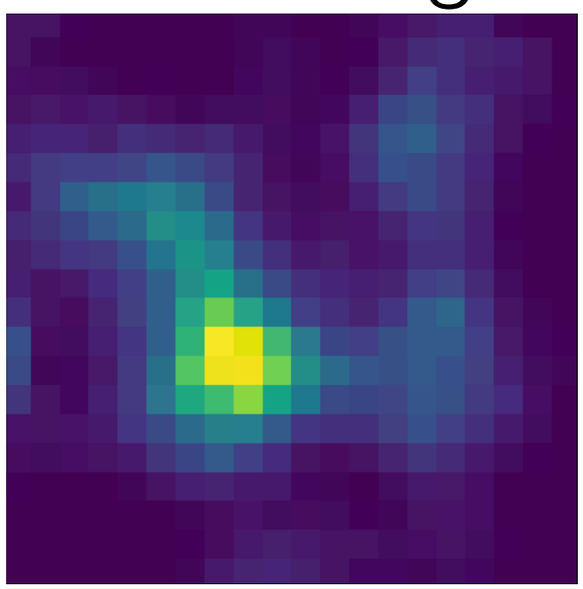
Goal-Driven Modeling - Primary Components



MEC Heterogeneous Cell

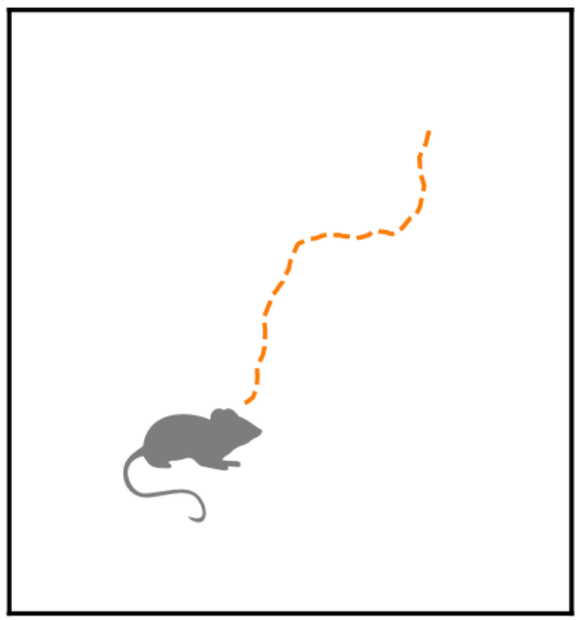


Model Heterogeneous Cell

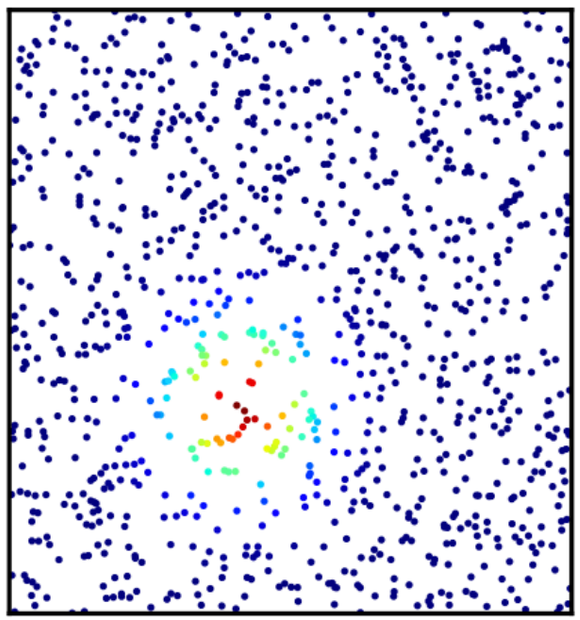


A spectrum of tasks

Simulated trajectory



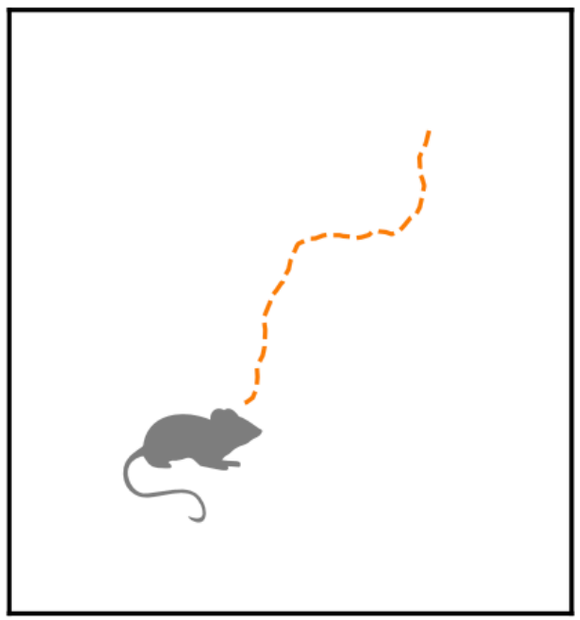
Place cell centers



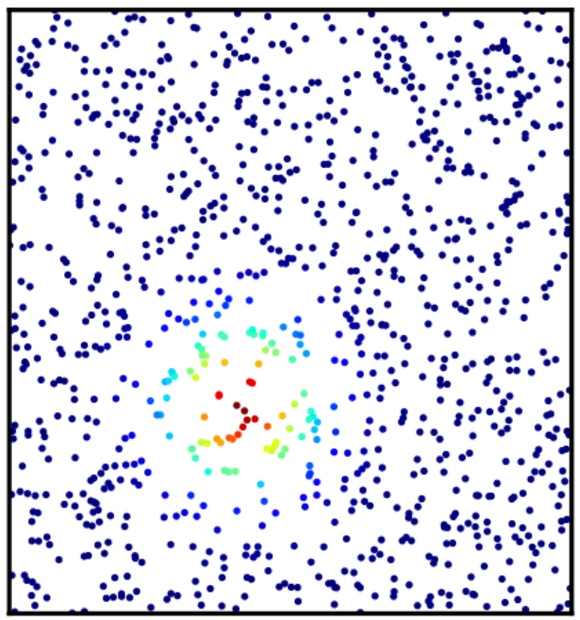
Sorscher, Mel*, Ganguli, Ocko*
***NeurIPS** (2019)*

A spectrum of tasks

Simulated trajectory



Place cell centers



Sorscher, Mel*, Ganguli, Ocko*
***NeurIPS** (2019)*

Velocity ⟶ MEC ⟶ Place Cells ⟶ Position (x,y)

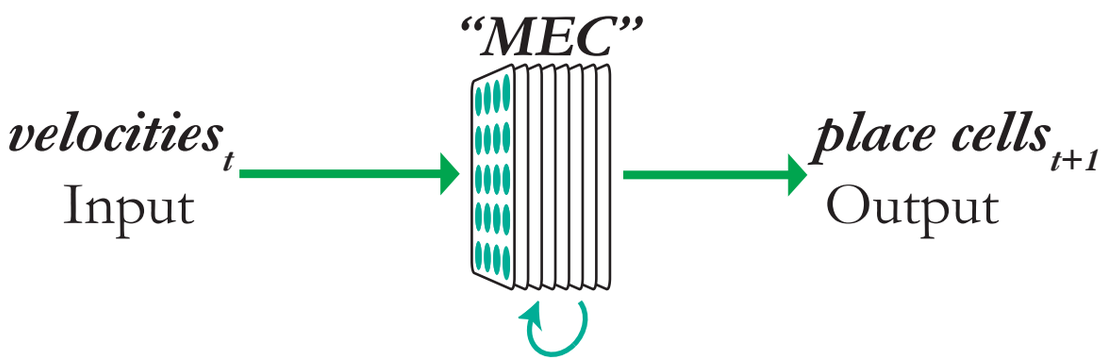
A spectrum of tasks

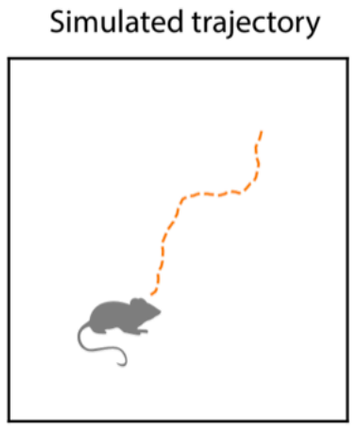
Simplest “model”

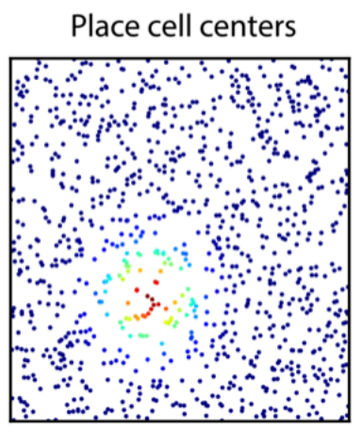
A spectrum of tasks

$$\mathcal{L}(\hat{p}, p) := -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_p} p_i^t \log \hat{p}_i^t$$

Banino*, Barry* *et al.* 2018





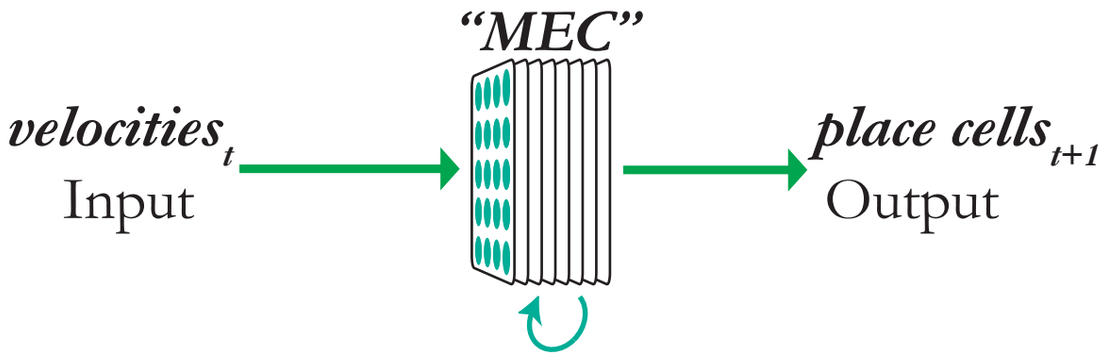


Velocity ⟶ MEC ⟶ Place Cells ⟶ Position (x,y)

A spectrum of tasks

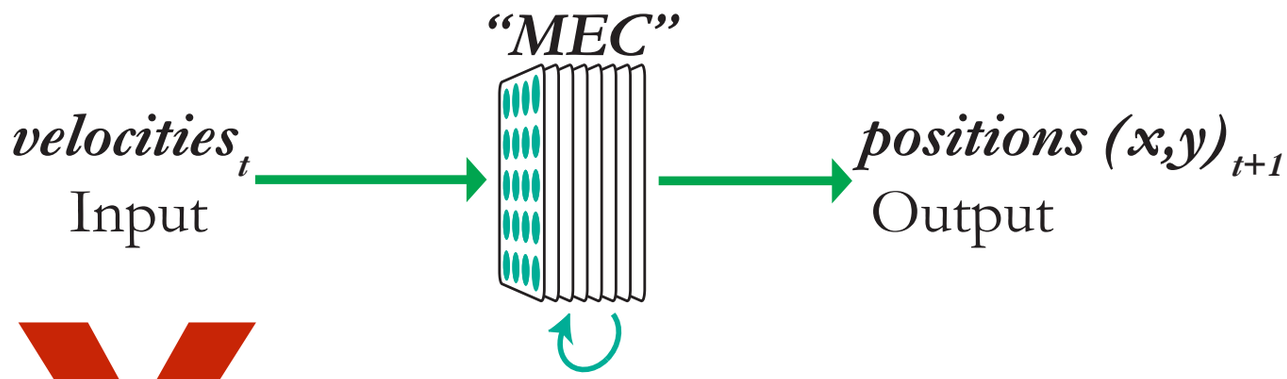
$$\mathcal{L}(\hat{p}, p) := -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_p} p_i^t \log \hat{p}_i^t$$

Banino*, Barry* *et al.* 2018



$$\mathcal{L}(\hat{p}, p) := \frac{1}{2}\frac{1}{T}\sum_{t=1}^{T}\left(\left(p_x^t - \hat{p}_x^t\right)^2 + \left(p_y^t - \hat{p}_y^t\right)^2\right)$$

Cueva* & Wei* 2018

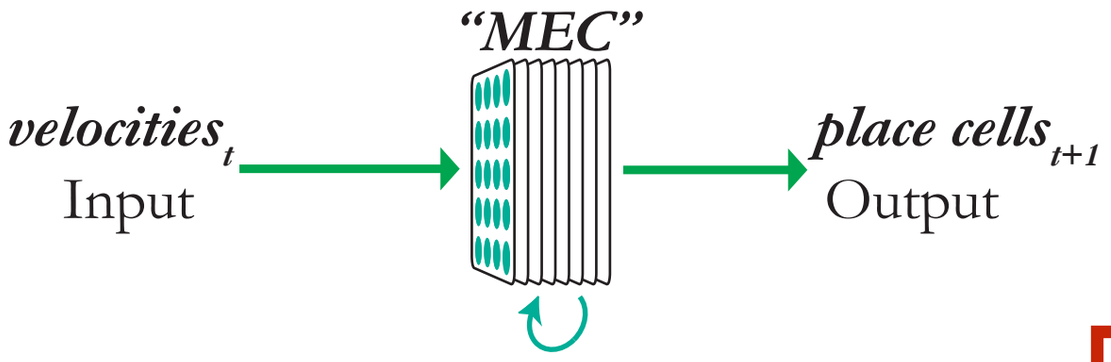


Velocity → MEC → ~~Place Cells~~ → Position (x,y)

A spectrum of tasks

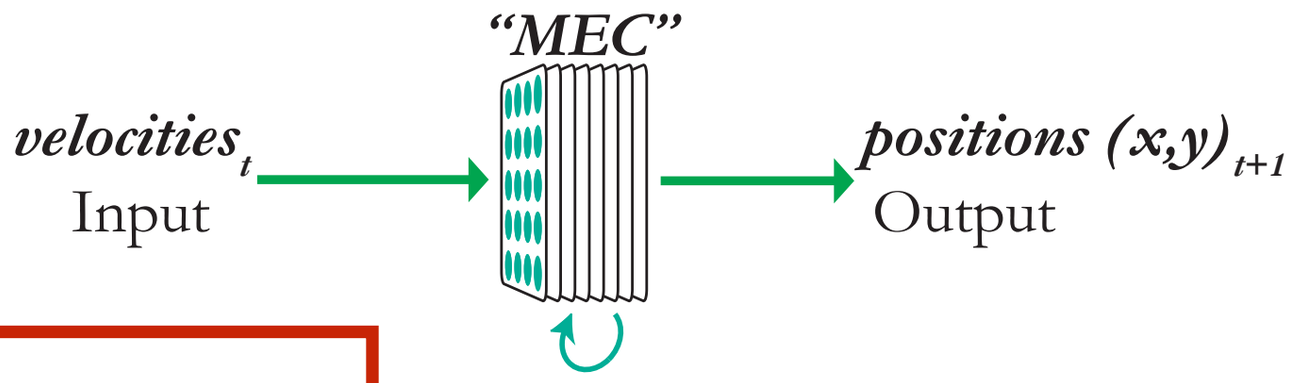
$$\mathcal{L}(\hat{p}, p) := -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_p} p_i^t \log \hat{p}_i^t$$

Banino*, Barry* *et al.* 2018



$$\mathcal{L}(\hat{p}, p) := \frac{1}{2}\frac{1}{T}\sum_{t=1}^{T}\left(\left(p_x^t - \hat{p}_x^t\right)^2 + \left(p_y^t - \hat{p}_y^t\right)^2\right)$$

Cueva* & Wei* 2018



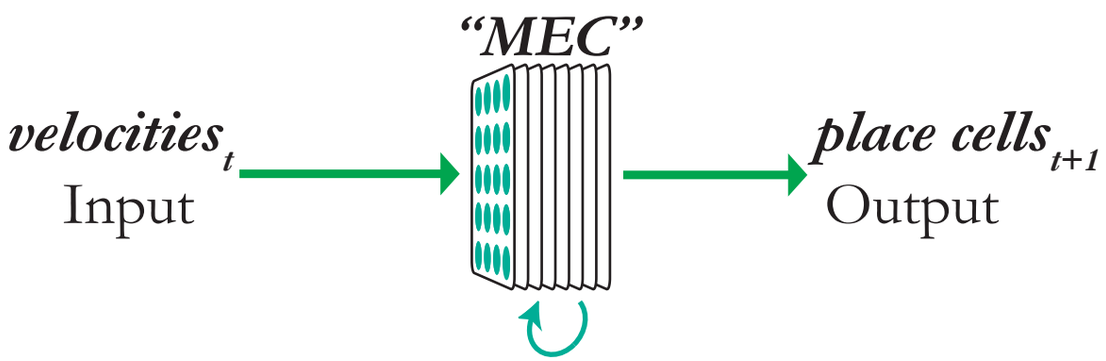
Velocity → MEC → Place Cells → Position (x,y)

Output-based models

A spectrum of tasks

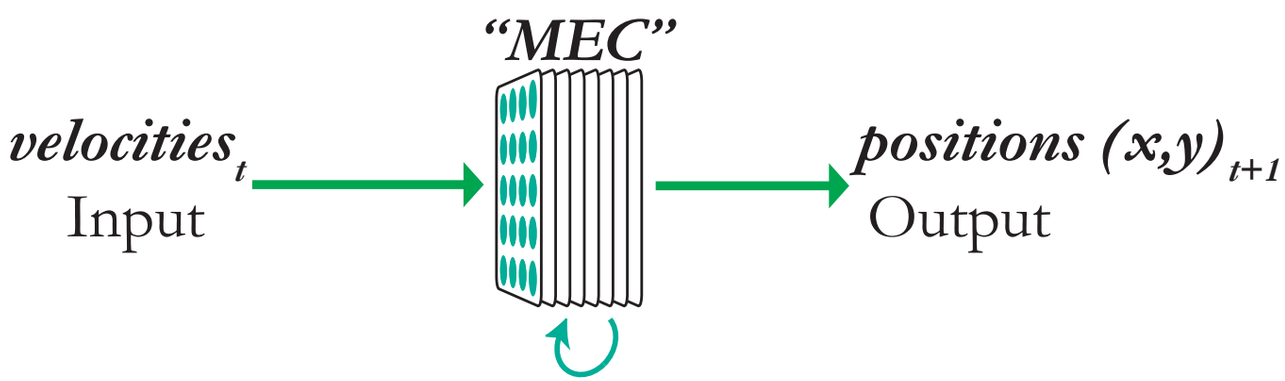
$$\mathcal{L}(\hat{p}, p) := -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_p} p_i^t \log \hat{p}_i^t$$

Banino*, Barry* *et al.* 2018



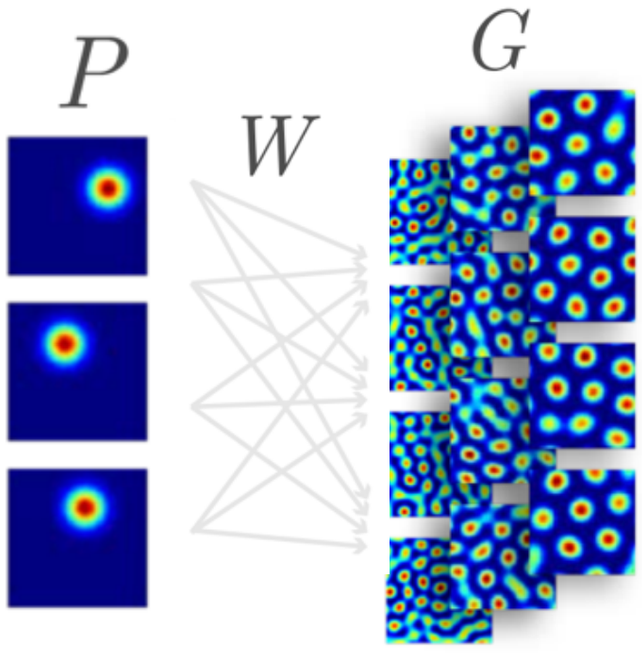
$$\mathcal{L}(\hat{p}, p) := \frac{1}{2}\frac{1}{T}\sum_{t=1}^{T}\left(\left(p_x^t - \widehat{p}_x^t\right)^2 + \left(p_y^t - \widehat{p}_y^t\right)^2\right)$$

Cueva* & Wei* 2018

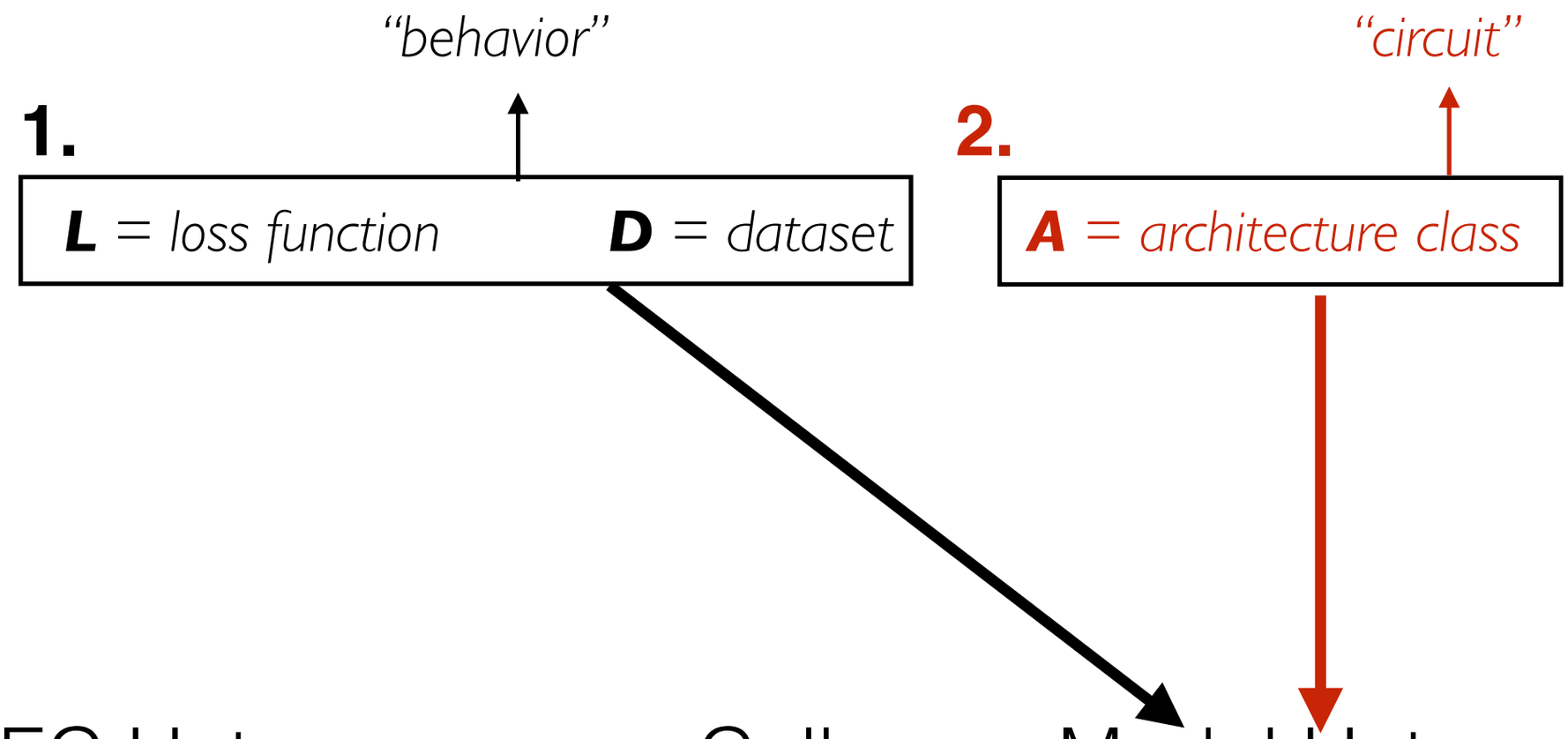


Velocity → MEC → Place Cells → Position (x,y)

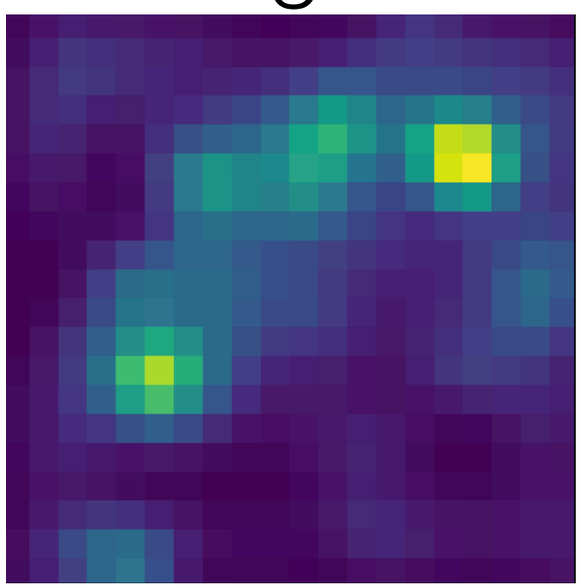
NMF
(Place Cell Input)



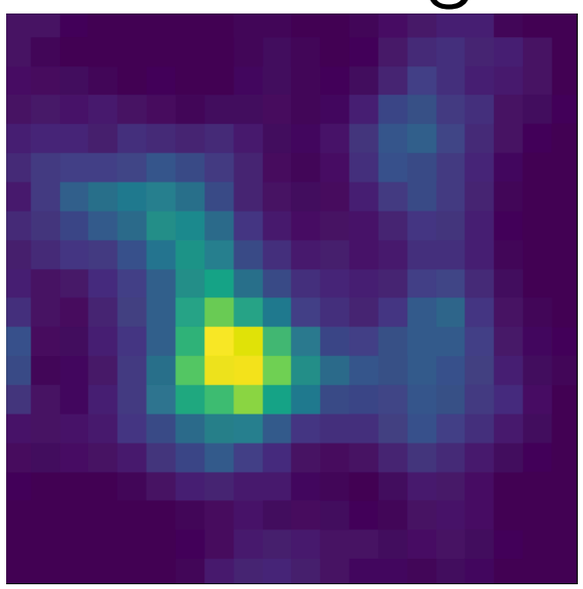
Goal-Driven Modeling - Primary Components



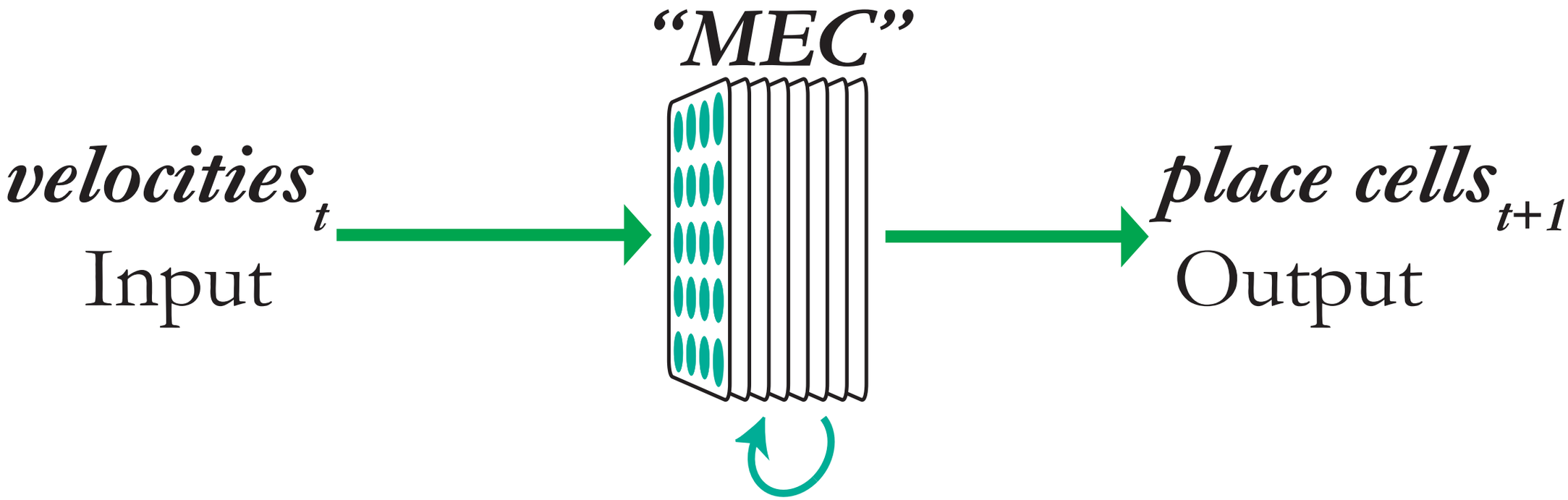
MEC Heterogeneous Cell



Model Heterogeneous Cell

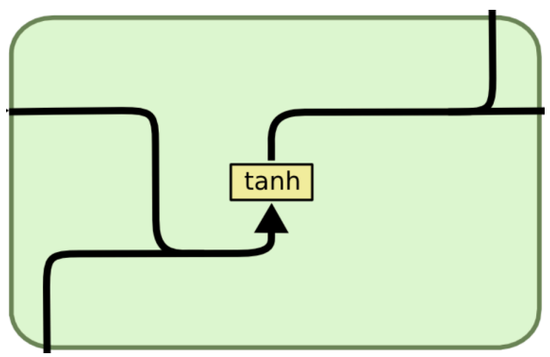


A spectrum of circuits

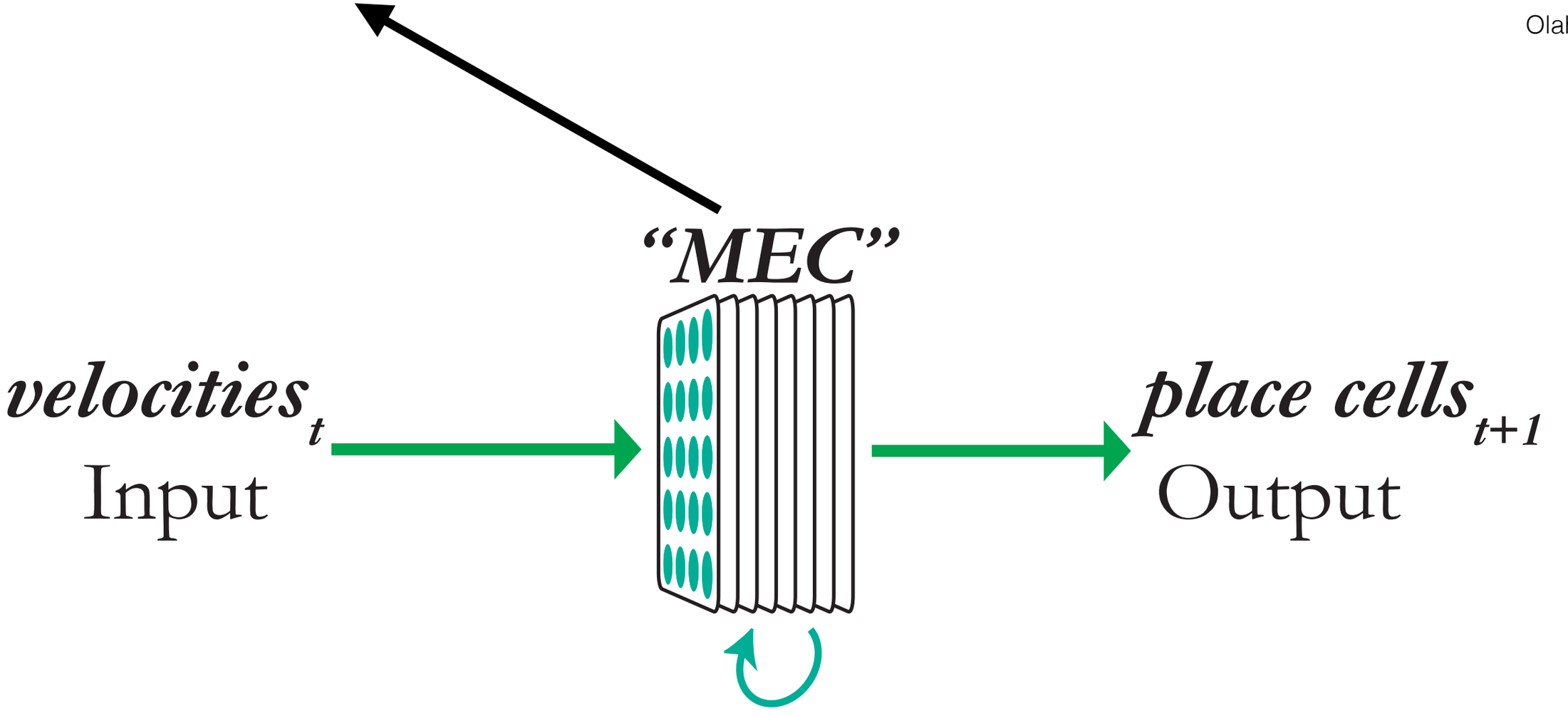


A spectrum of circuits

SimpleRNN

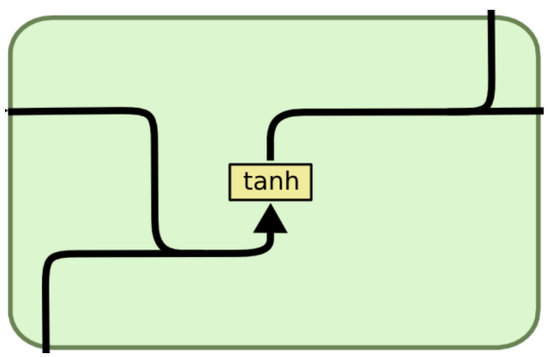


Olah 2015

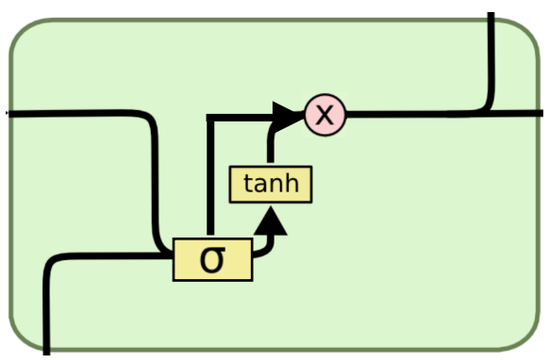


A spectrum of circuits — learnable modulation ("gating")

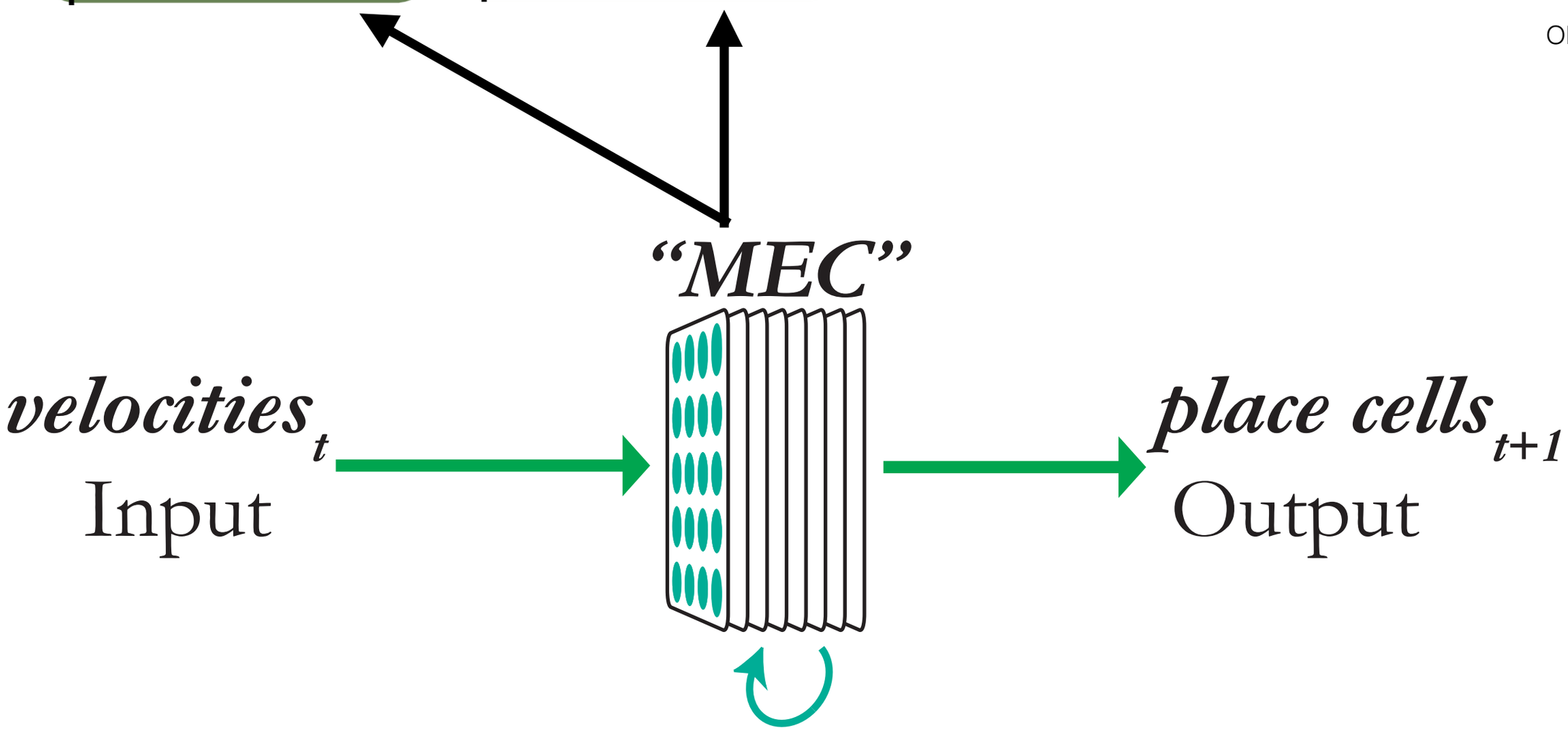
SimpleRNN



UGRNN

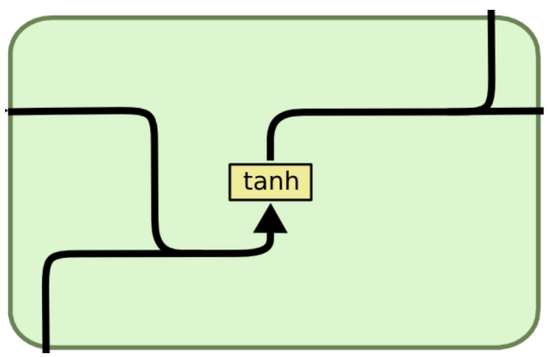


Olah 2015

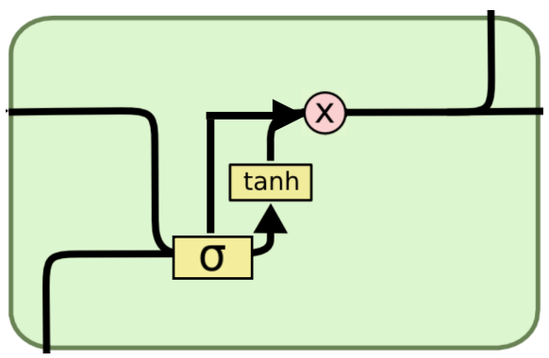


A spectrum of circuits — learnable modulation ("gating")

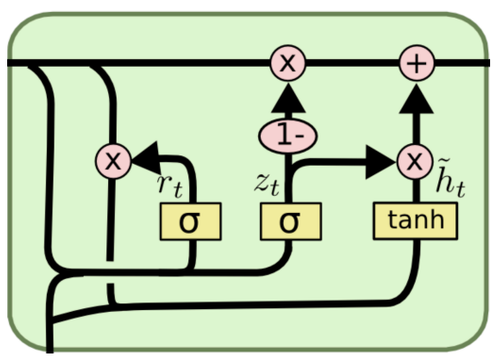
SimpleRNN



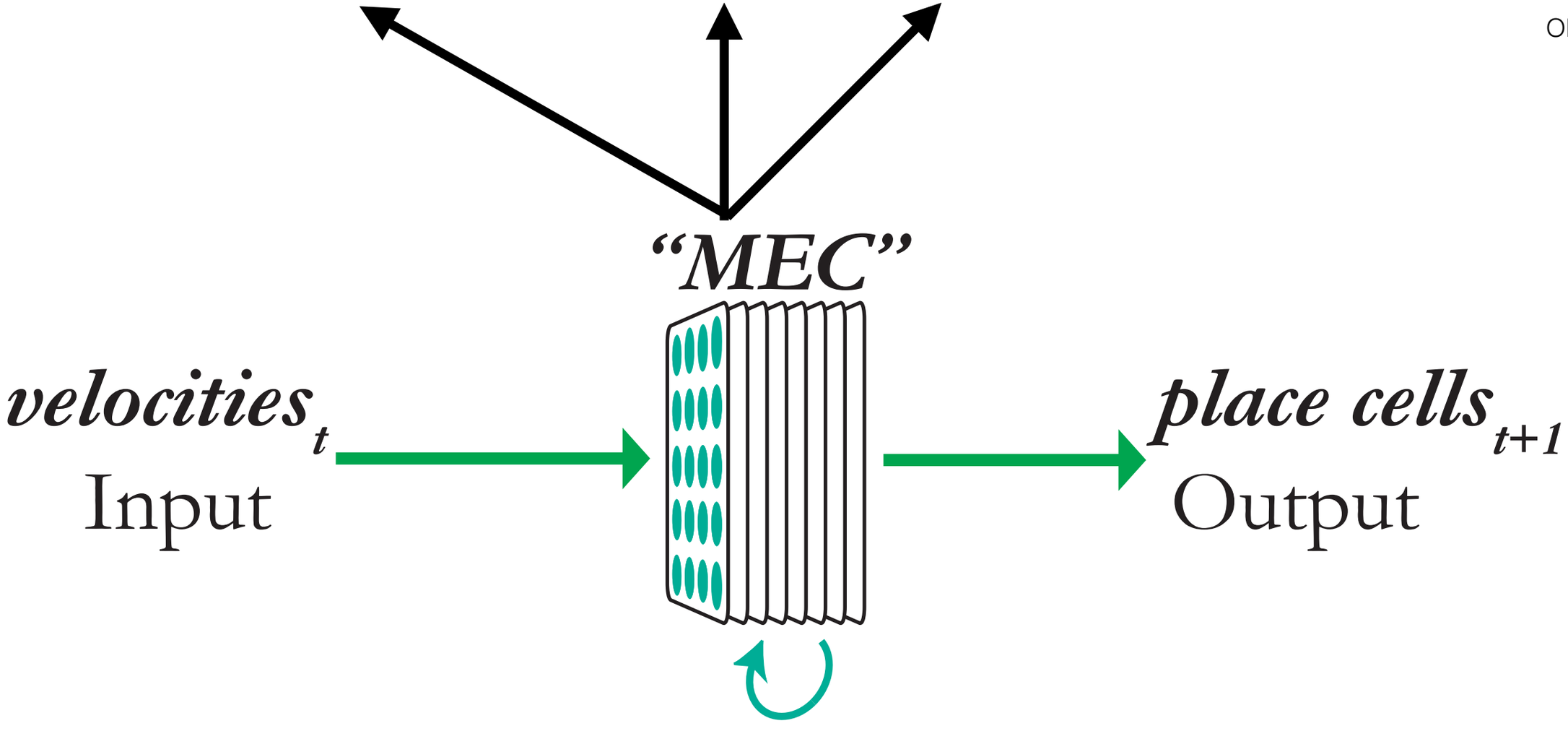
UGRNN



GRU

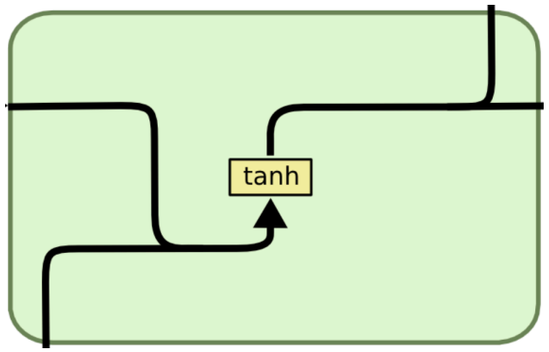


Olah 2015

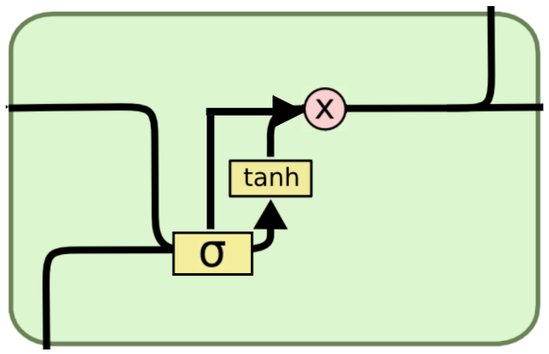


A spectrum of circuits — learnable modulation ("gating")

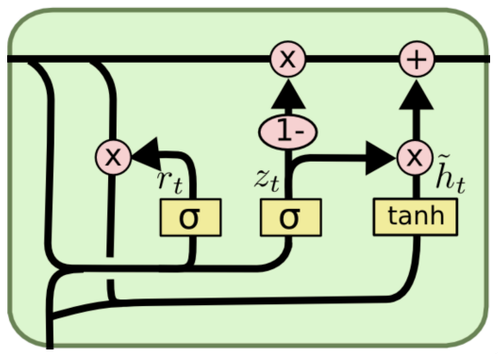
SimpleRNN



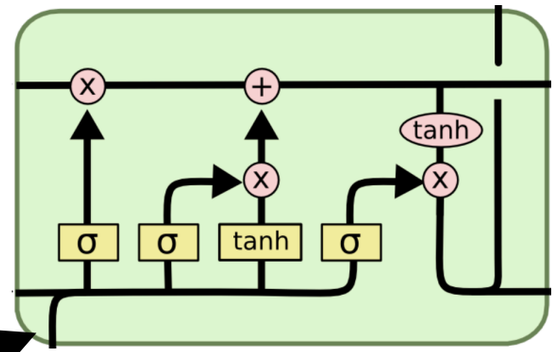
UGRNN



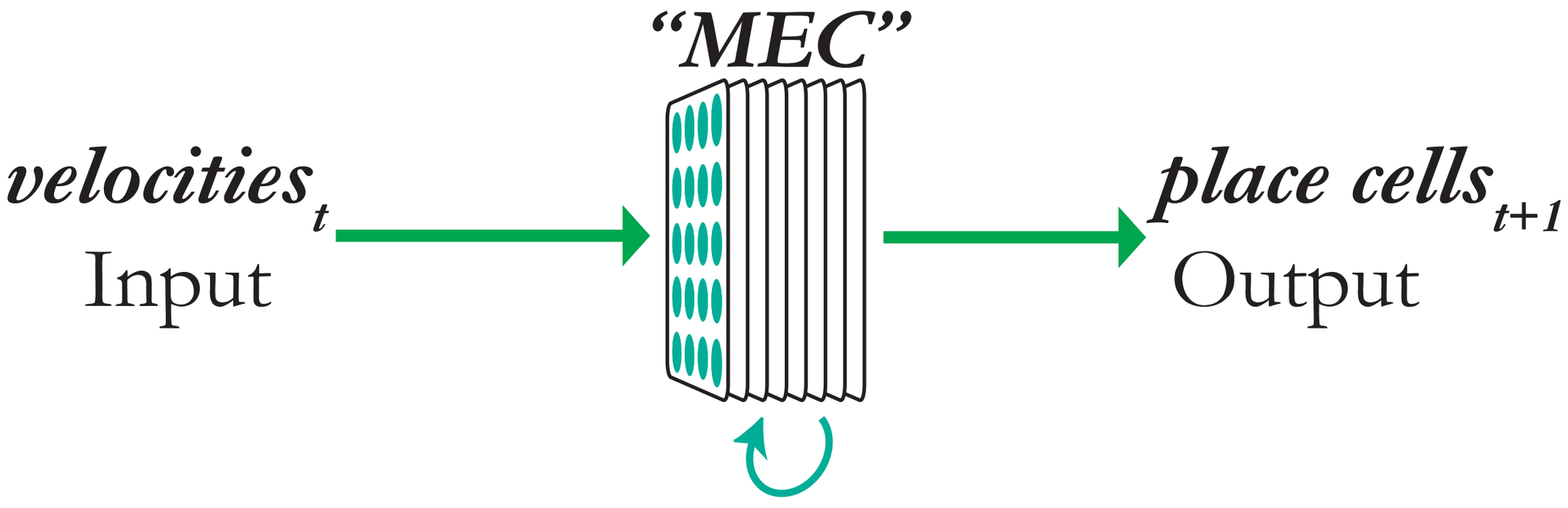
GRU



LSTM

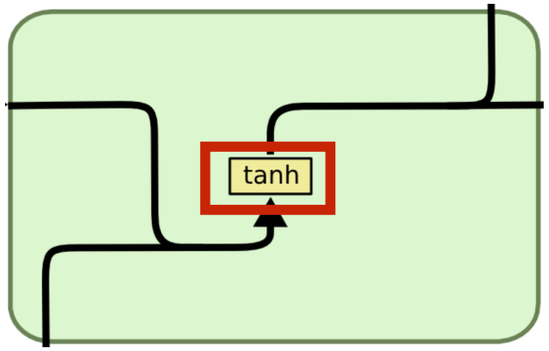


Olah 2015

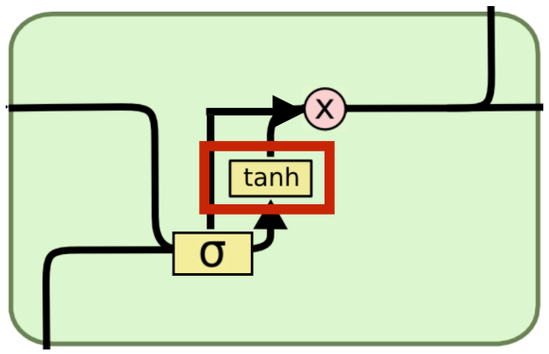


A spectrum of circuits — output nonlinearity

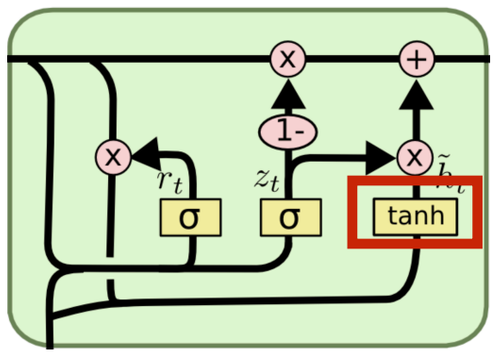
SimpleRNN



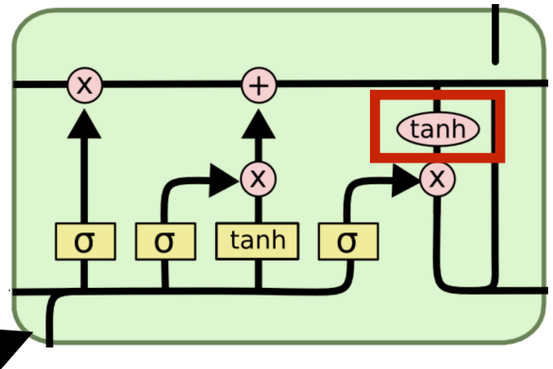
UGRNN



GRU

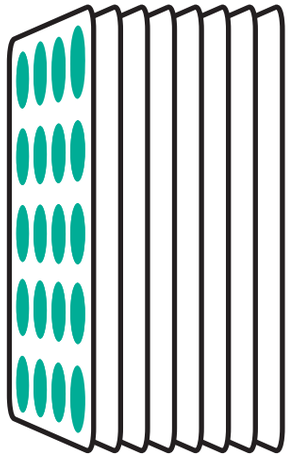


LSTM



Olah 2015

"MEC"

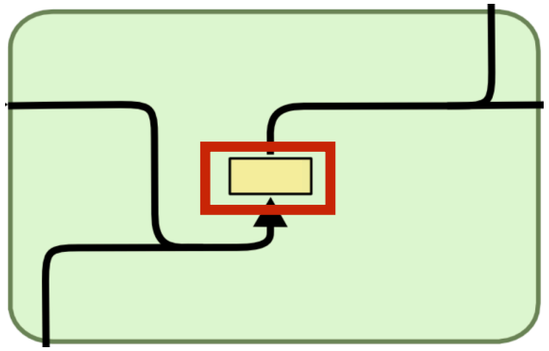


velocities$_t$
Input

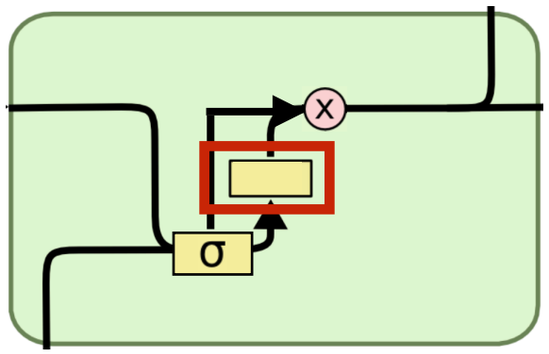
place cells$_{t+1}$
Output

A spectrum of circuits — output nonlinearity

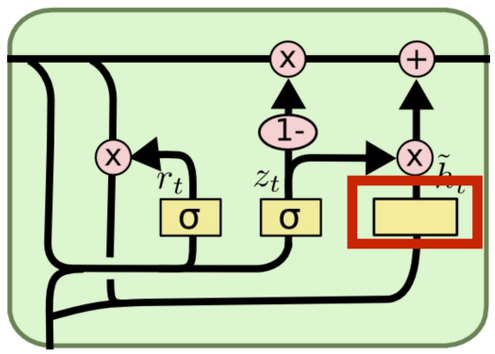
SimpleRNN



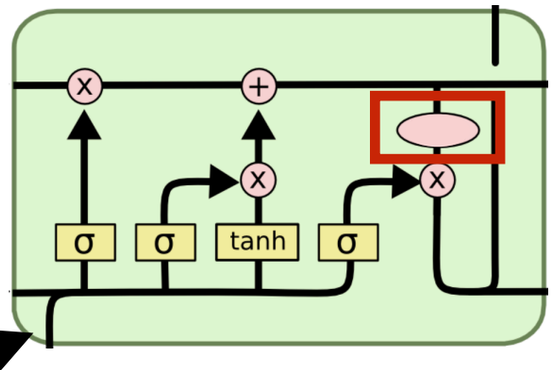
UGRNN



GRU



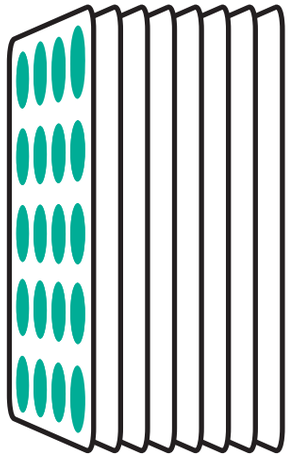
LSTM



Olah 2015

- Linear
- Tanh
- Sigmoid
- ReLU

"MEC"



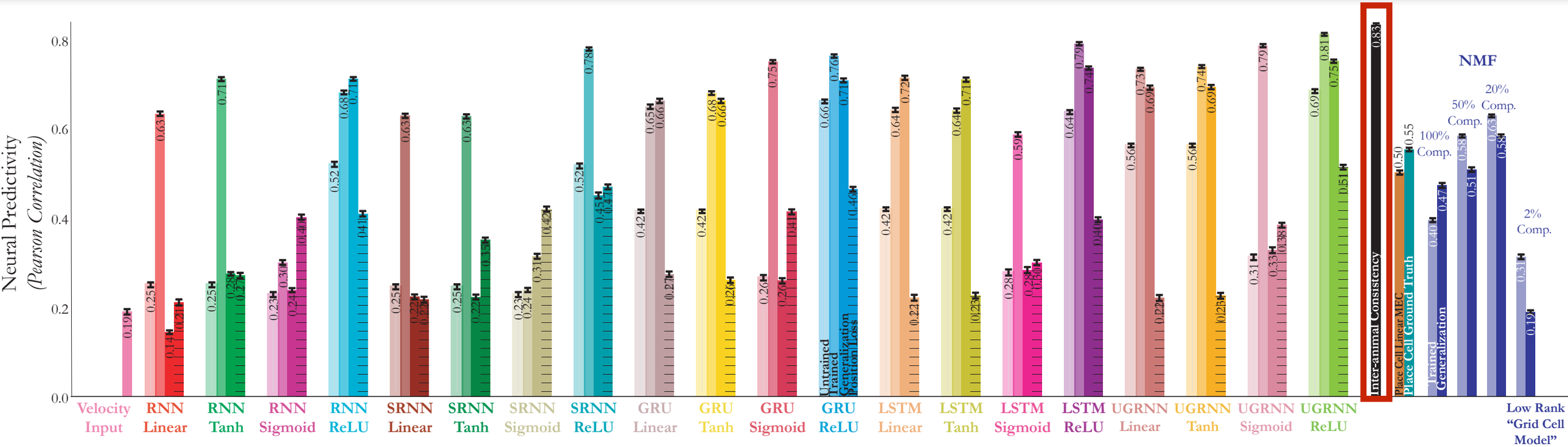
velocities$_t$

Input

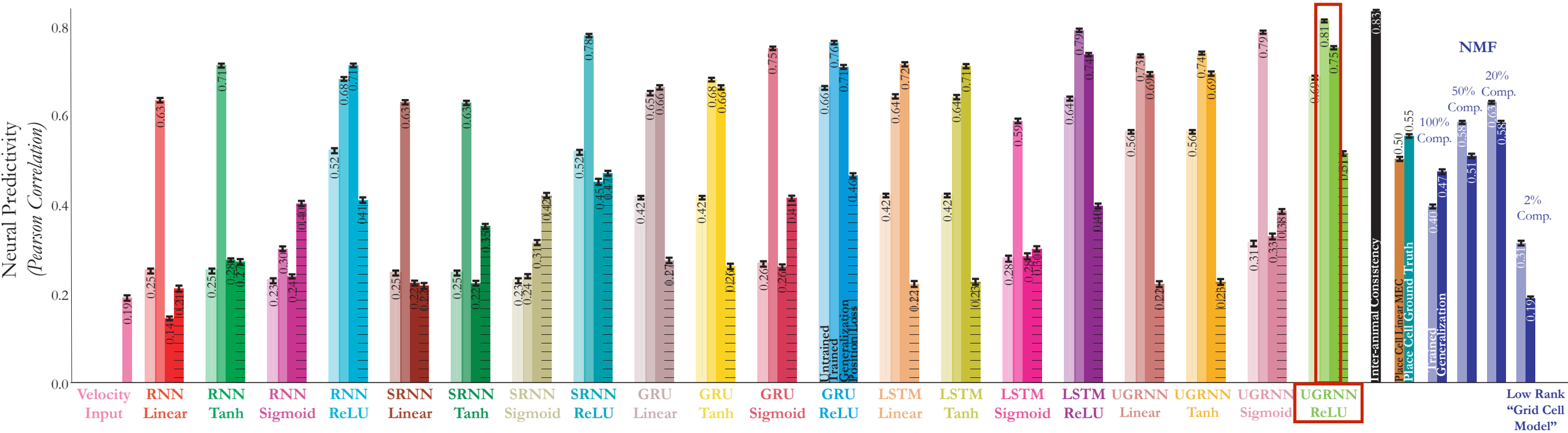
place cells$_{t+1}$

Output

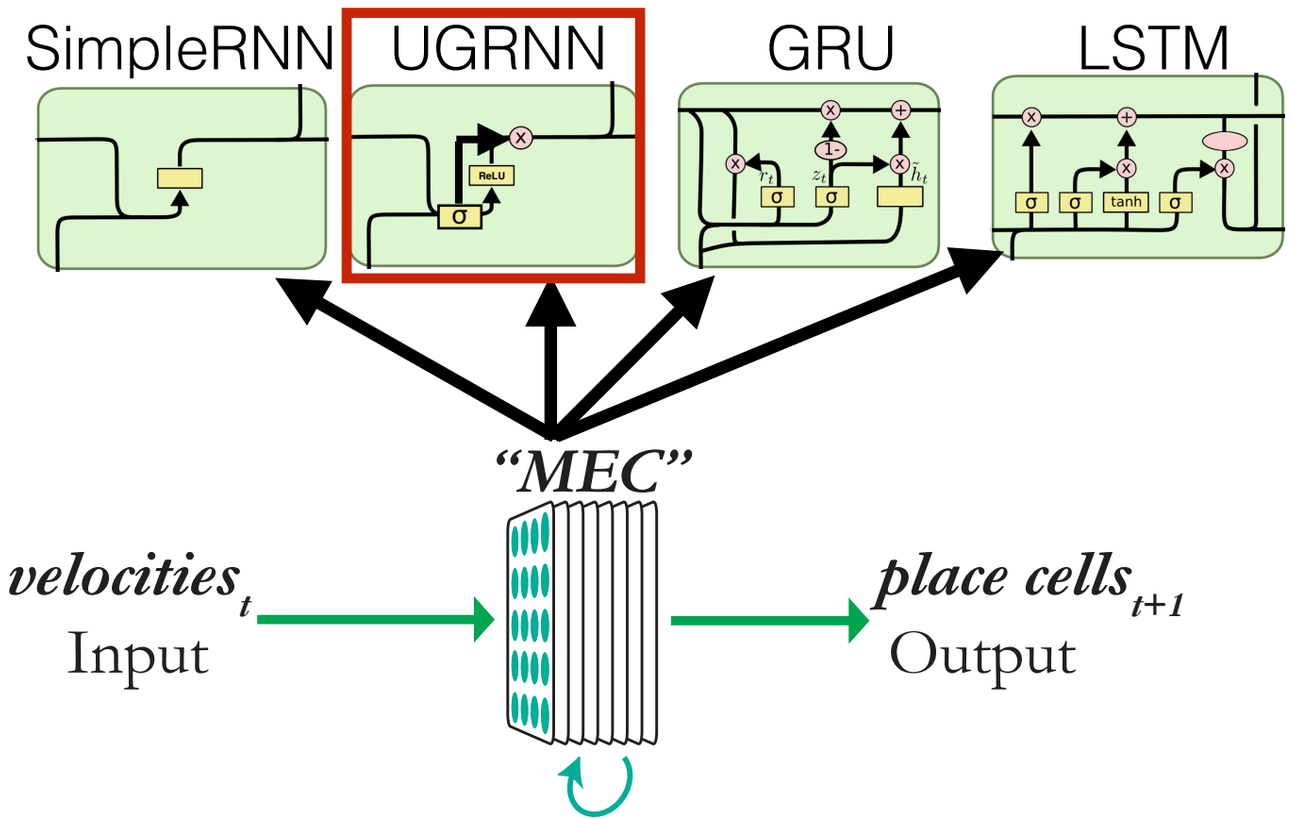
Benchmarking models with the same transform as between animals



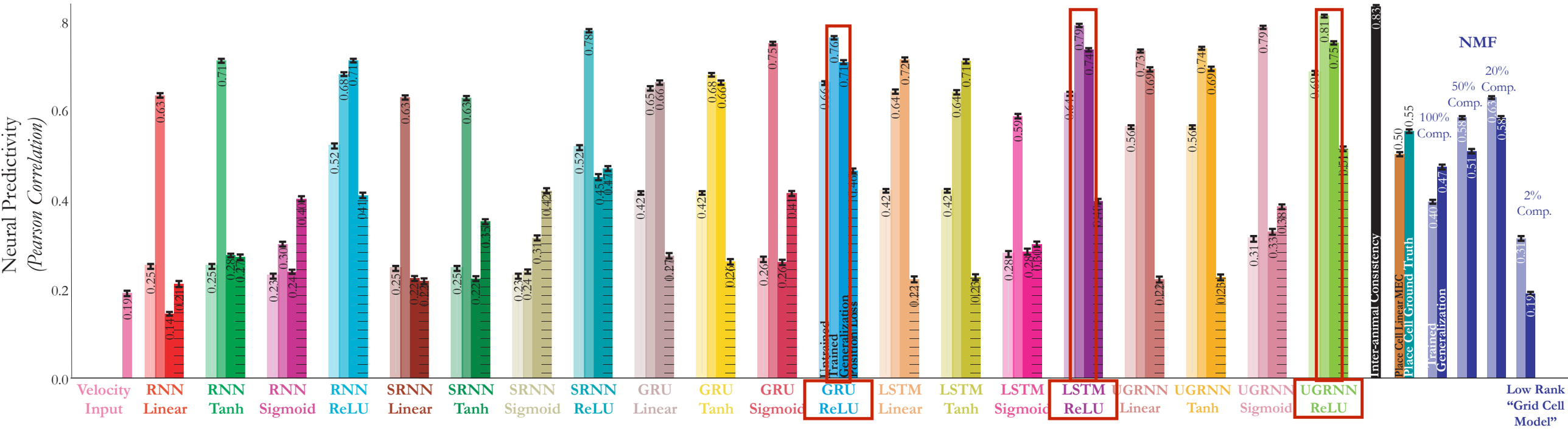
Task-optimized navigational models best predict the entire MEC population



Best task-optimized models "solve" the neurons

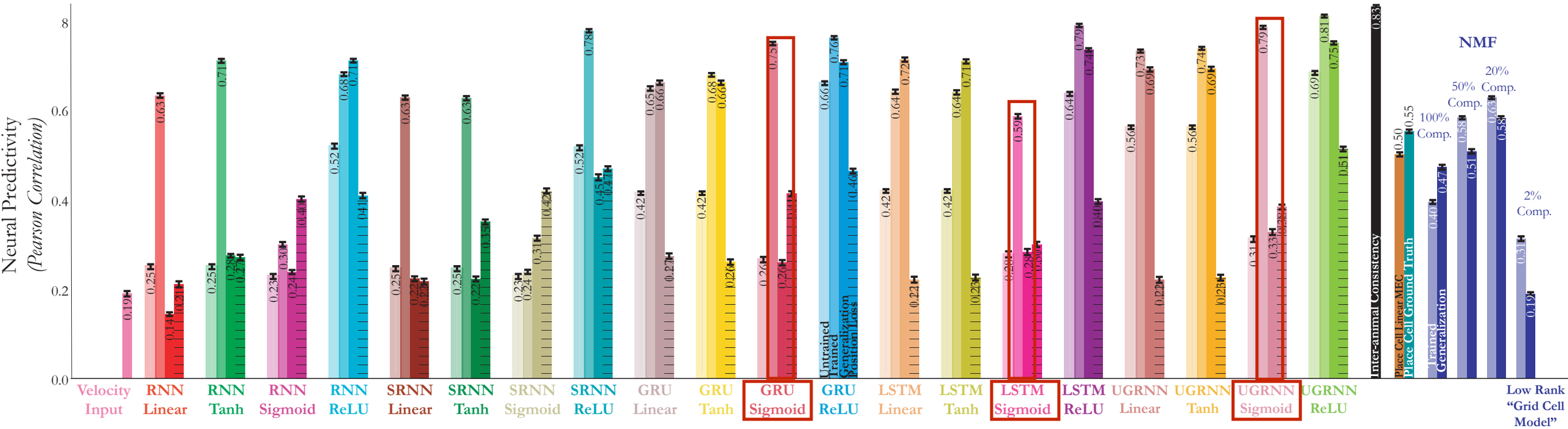


Nonlinearity affects generalization



Nonnegativity constraint + gating aids in generalization across environments

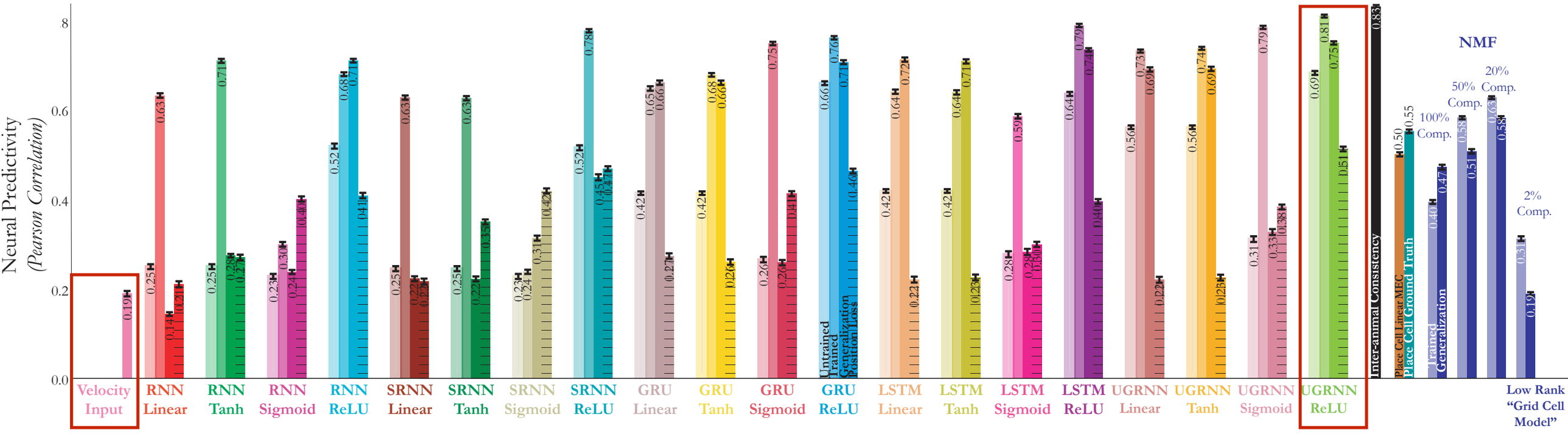
Nonlinearity affects generalization



Nonnegativity constraint + gating aids in generalization across environments

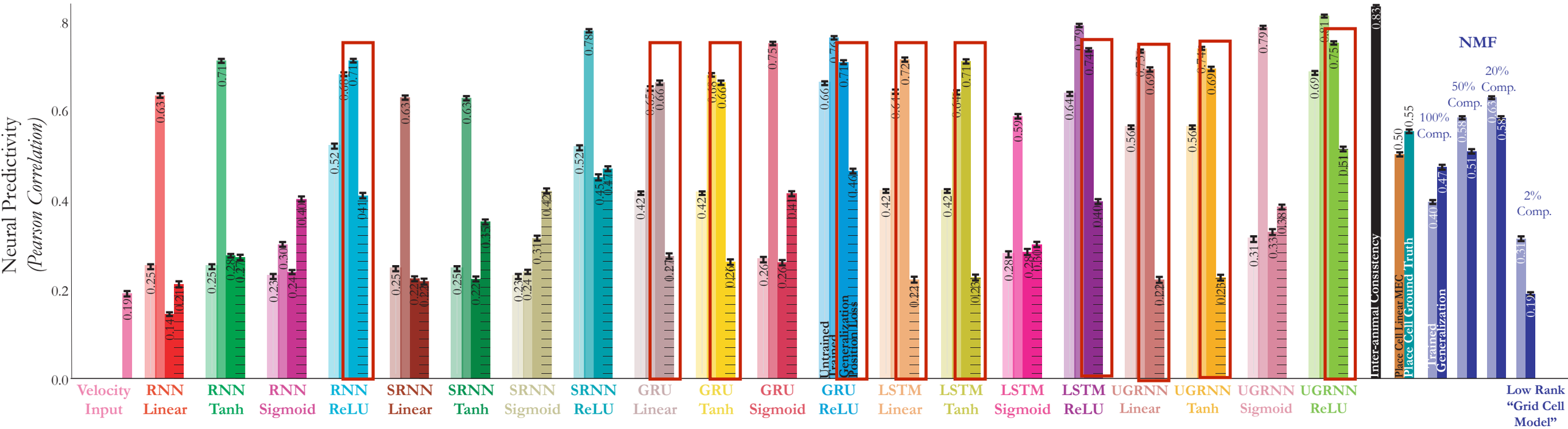
But this nonnegativity constraint must *not* saturate either!

Model input is a poor predictor of population



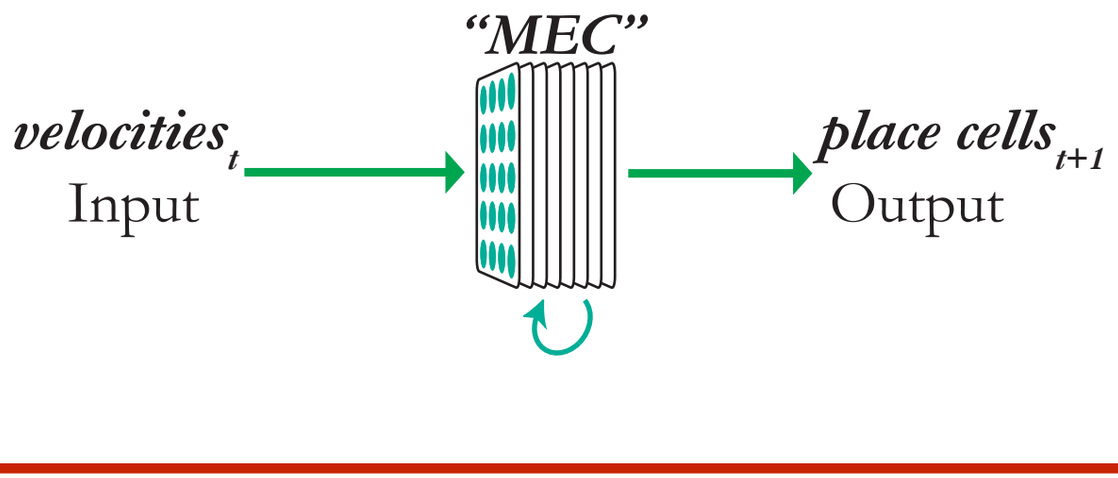
Models add a lot of predictive power to their velocity inputs

Directly supervising on Cartesian coordinates fails to generalize

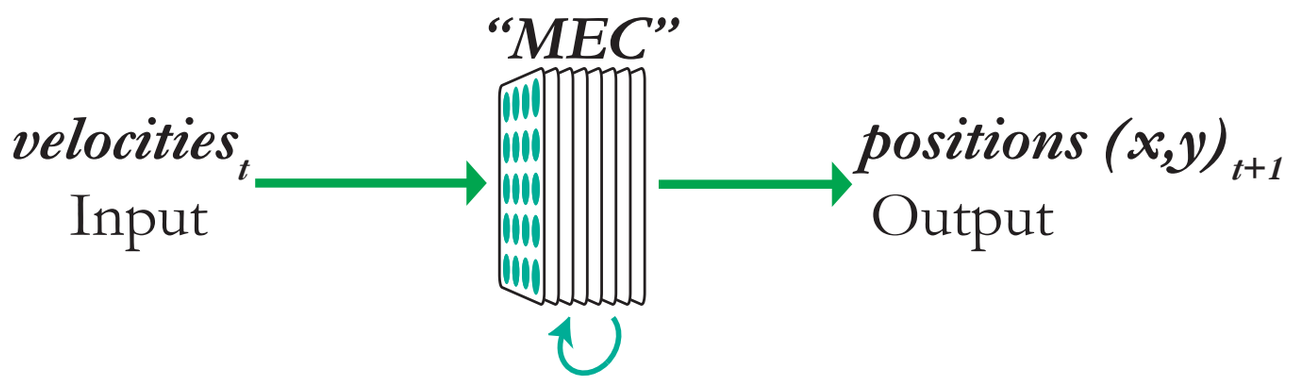


Output place cell supervision provides better generalization over direct supervision of position

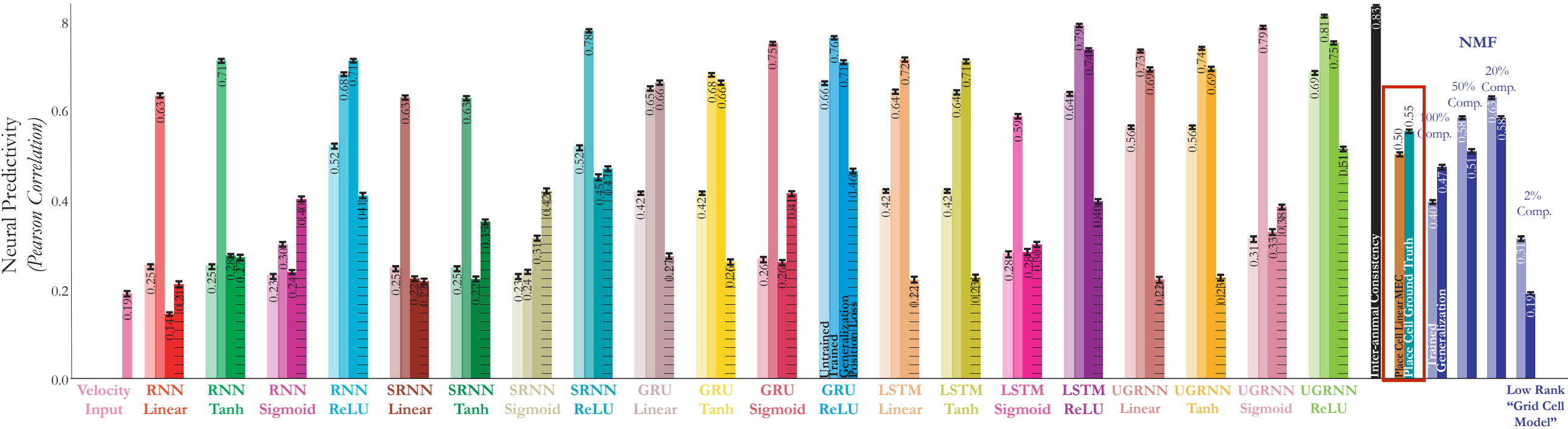
$$\mathcal{L}(\hat{p}, p) := -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_p} p_i^t \log \hat{p}_i^t$$



$$\mathcal{L}(\hat{p}, p) := \frac{1}{2}\frac{1}{T}\sum_{t=1}^{T}\left(\left(p_x^t - \hat{p}_x^t\right)^2 + \left(p_y^t - \hat{p}_y^t\right)^2\right)$$

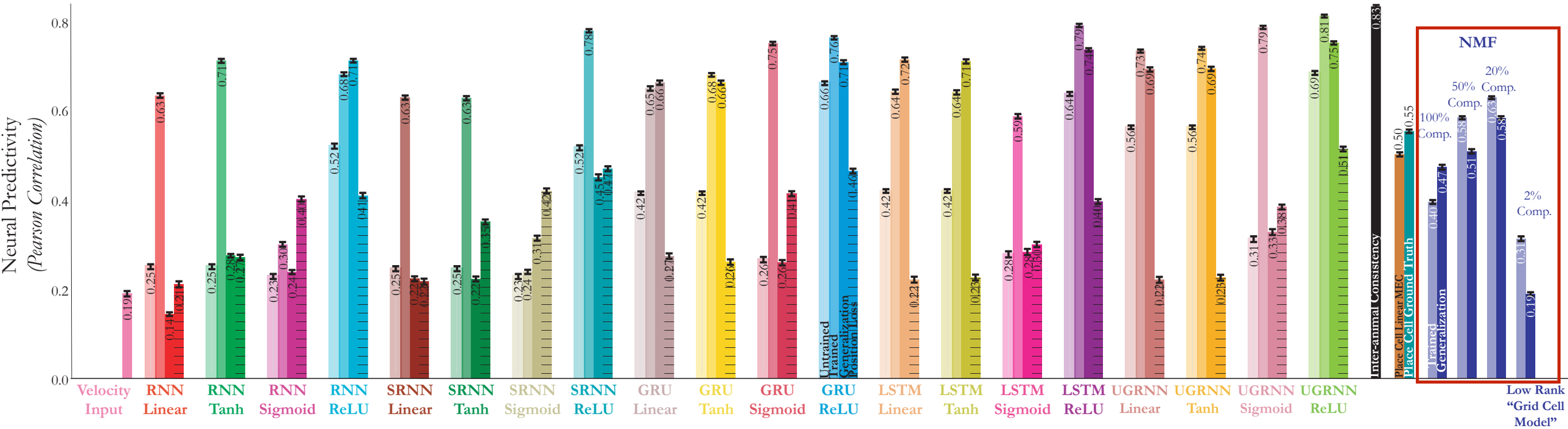


Place cells alone are a poor predictor



But place cells alone are not a good predictor of MEC (good!)

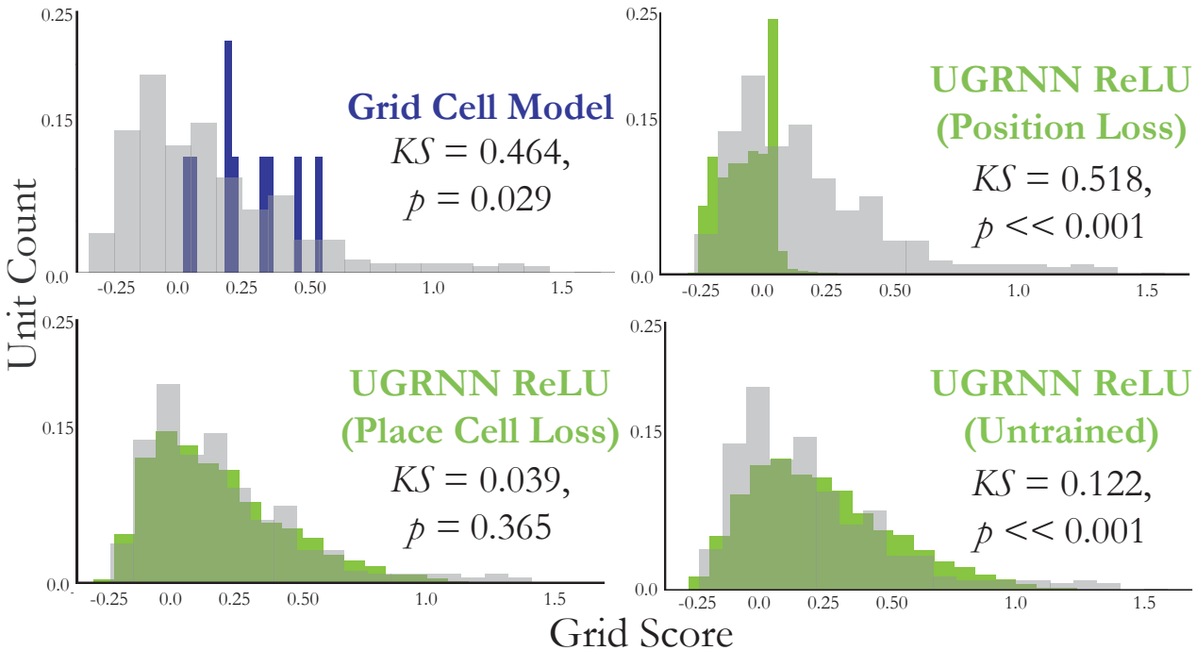
…as is NMF



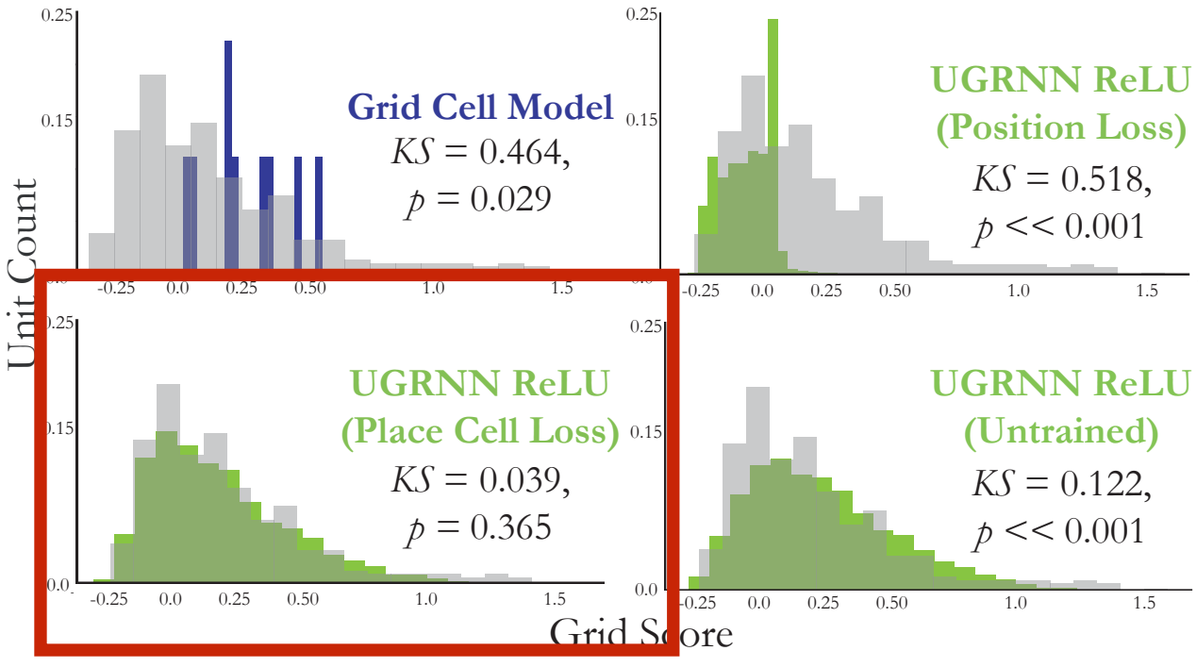
Dimensionality reduction on place cells is not a good predictor of MEC either

Task-optimized navigational models best predict the entire MEC population

Grid score distribution does not require any parameter fitting

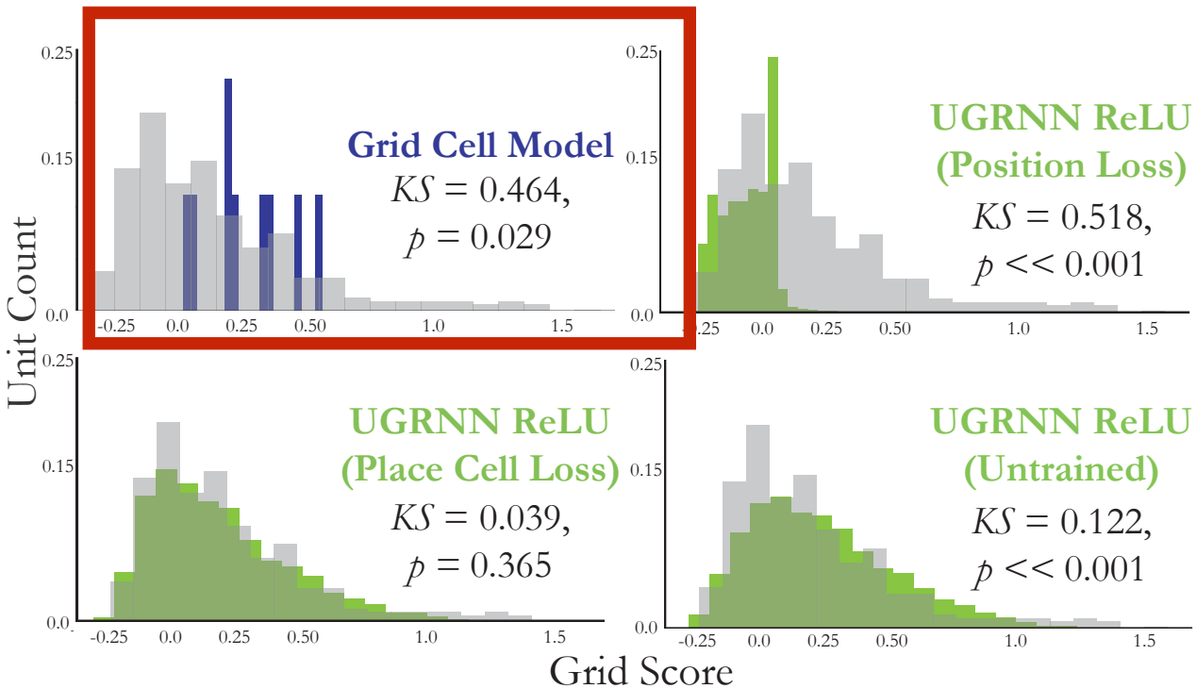


Task-optimized navigational models best predict the entire MEC population



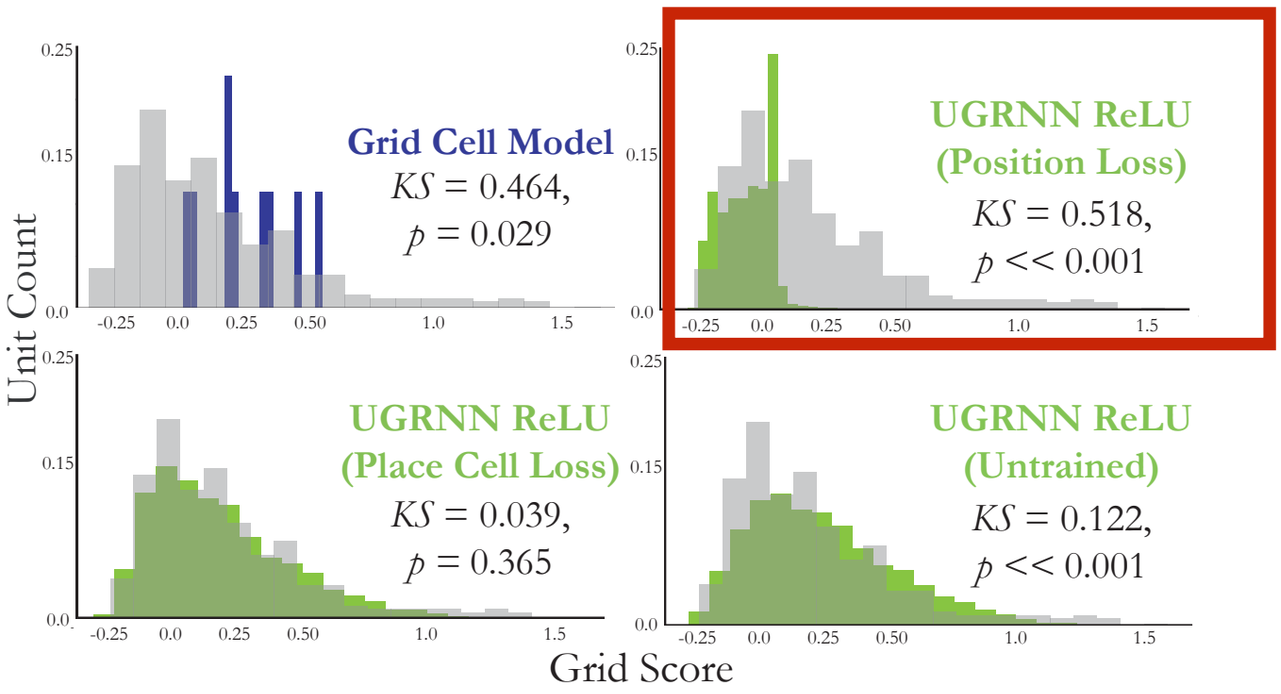
Best model class in terms of neural predictivity also matches grid score distribution in its own synthetic population

Task-optimized navigational models best predict the entire MEC population



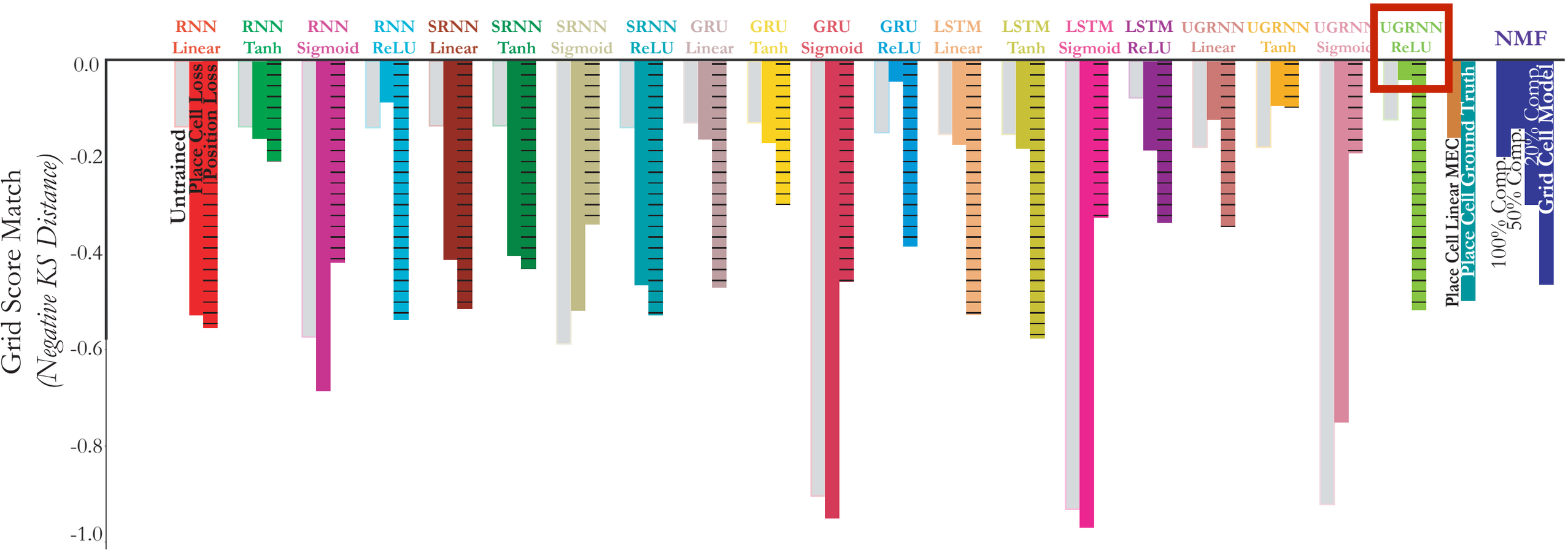
Low-rank model is too biased towards grid-like units

Task-optimized navigational models best predict the entire MEC population



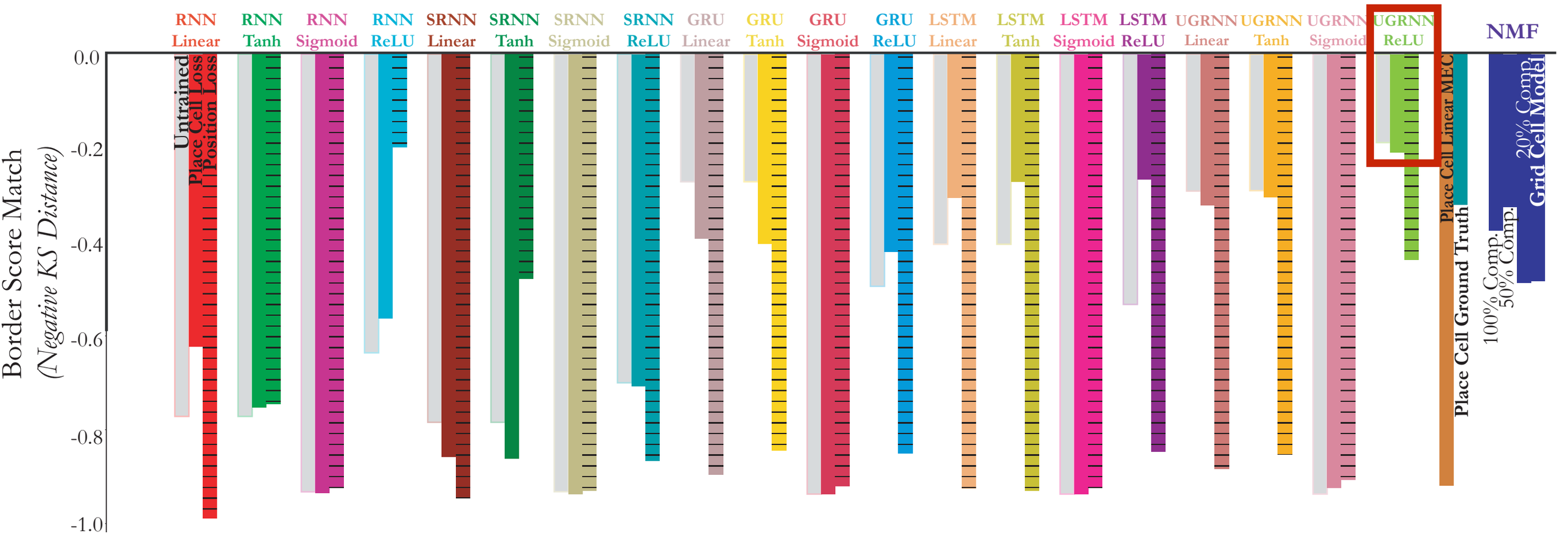
Without place cell representation, the model is too biased towards *non* grid-like units

More fine-grained unit matching metrics



Best model matches the data's grid score distribution in its own synthetic population

More fine-grained unit matching metrics

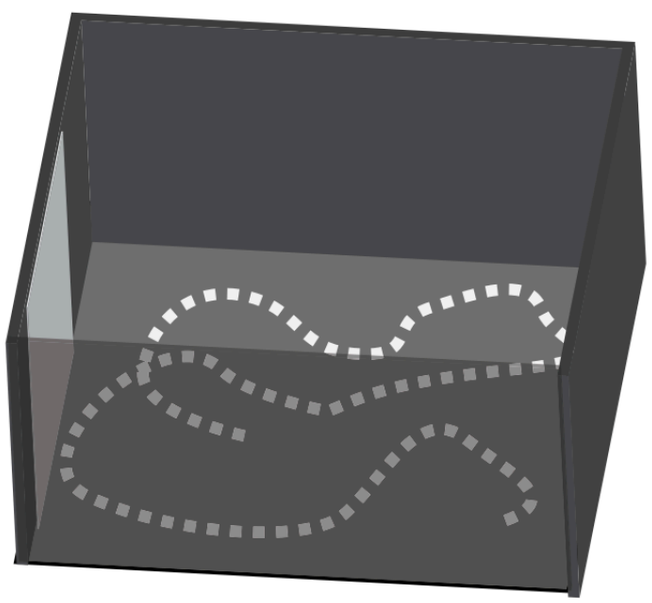


Best model also matches the data's *border* score distribution in its own synthetic population

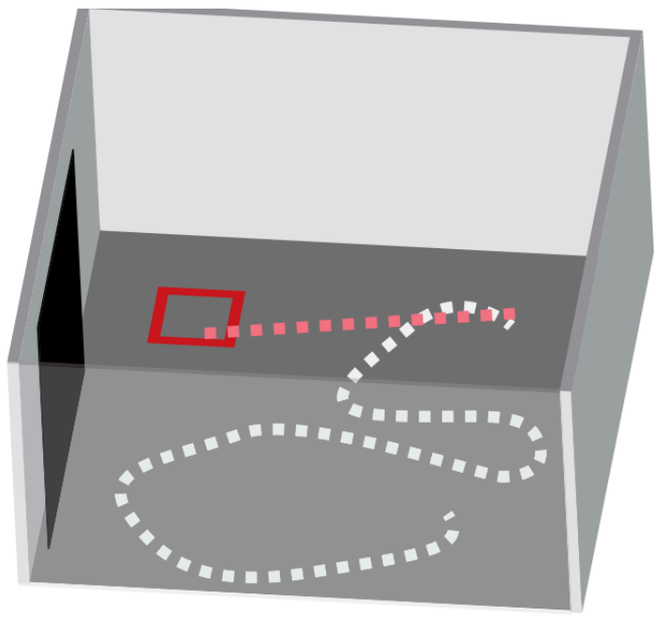
Remapping in the presence of reward

Remembered reward locations restructure entorhinal spatial maps

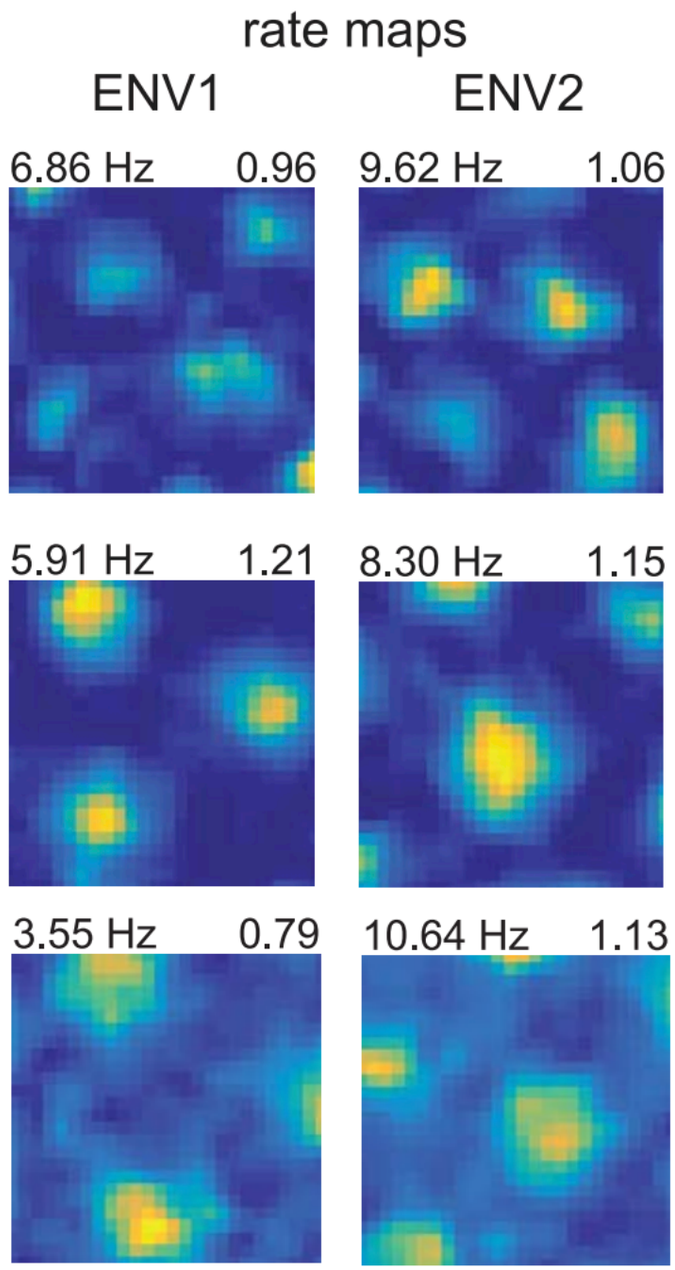
William N. Butler*, Kiah Hardcastle*, Lisa M. Giocomo†



free foraging (ENV1)



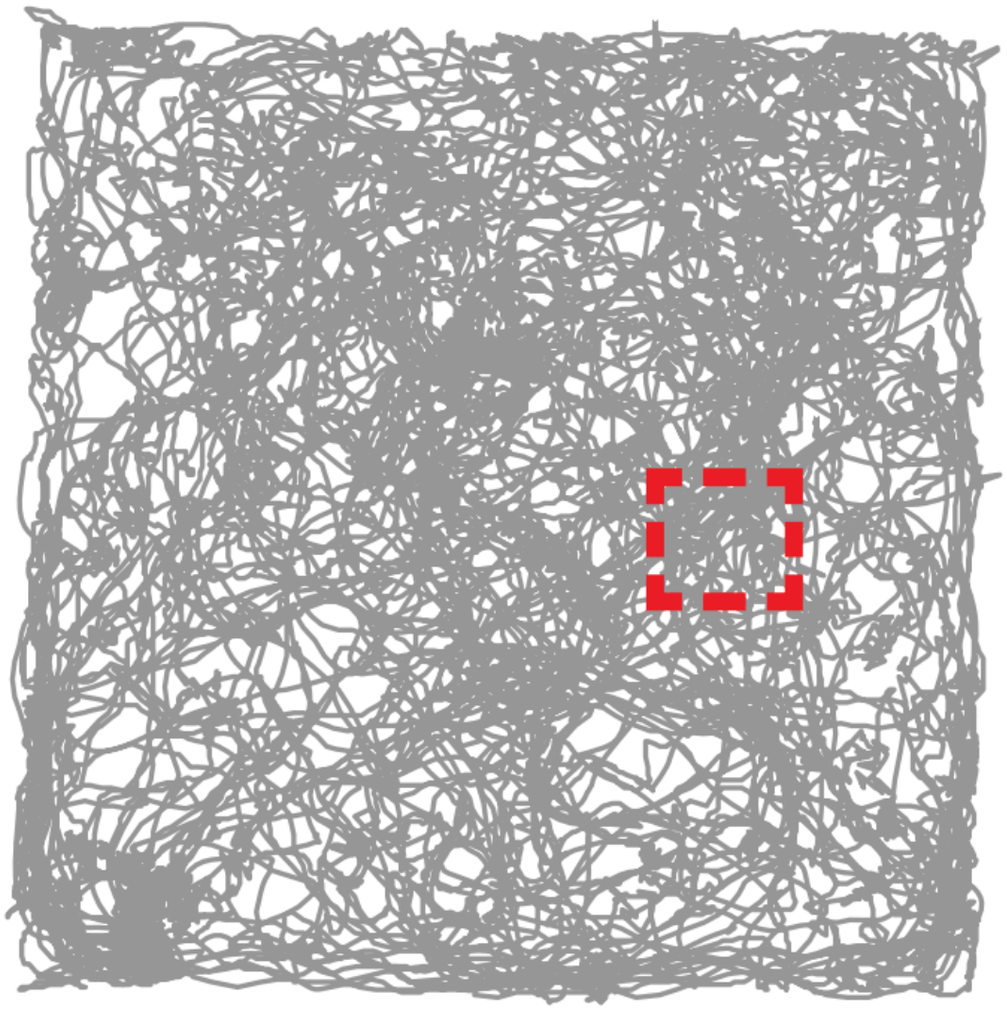
spatial task (ENV2)



Inspiration from animal behavior — rapid, direct paths

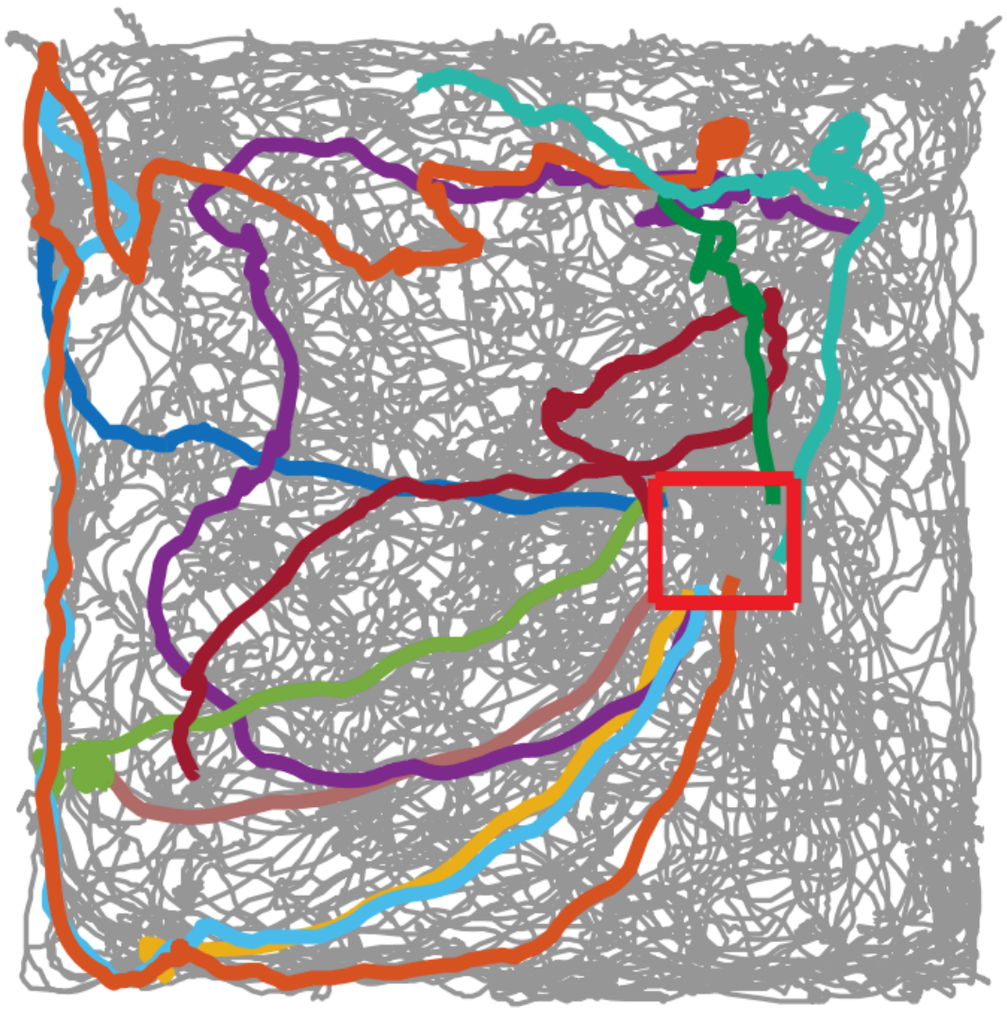
Animals tend to take rapid, direct paths to reward zone

ENV1



0.5m

ENV2

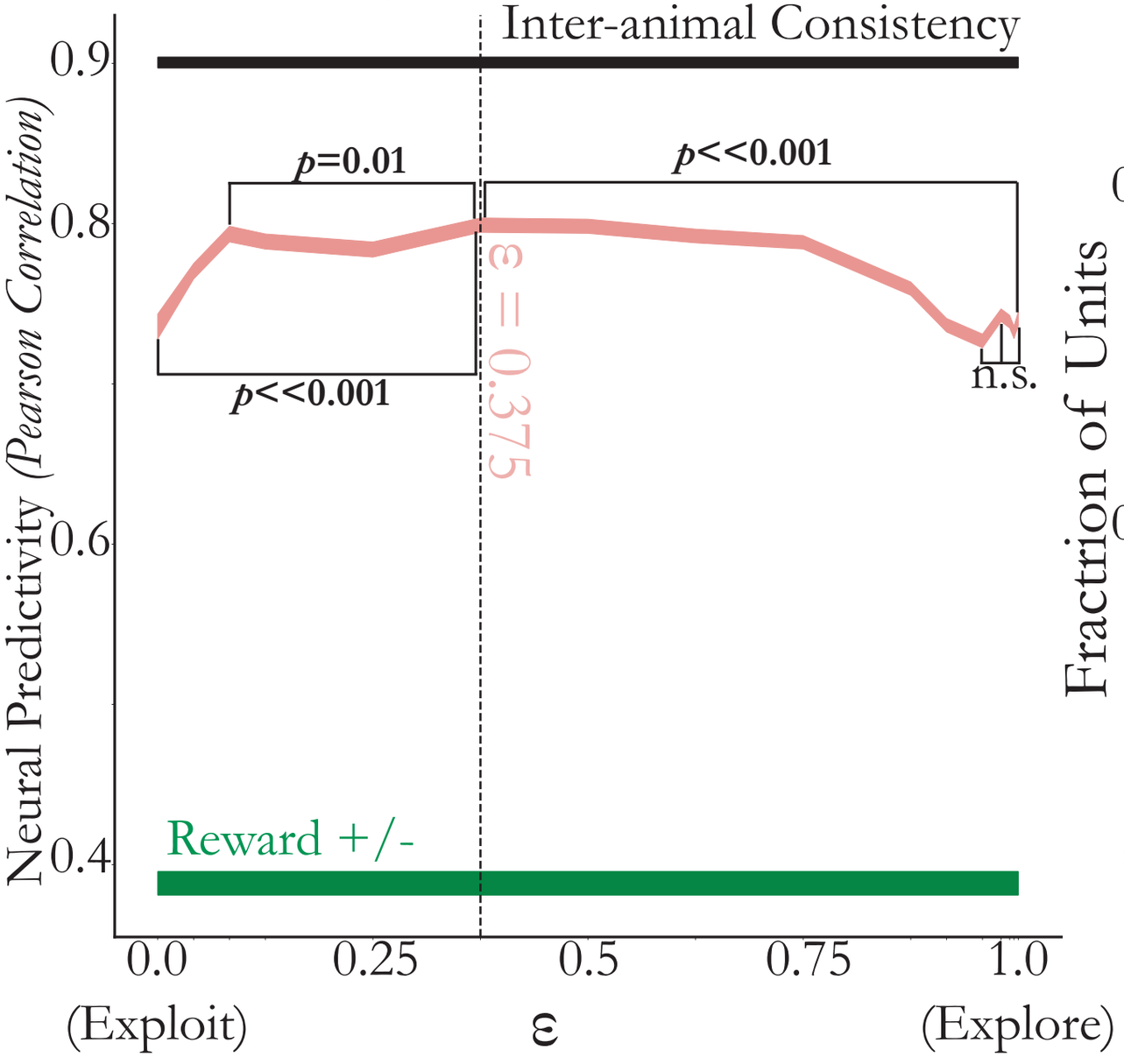


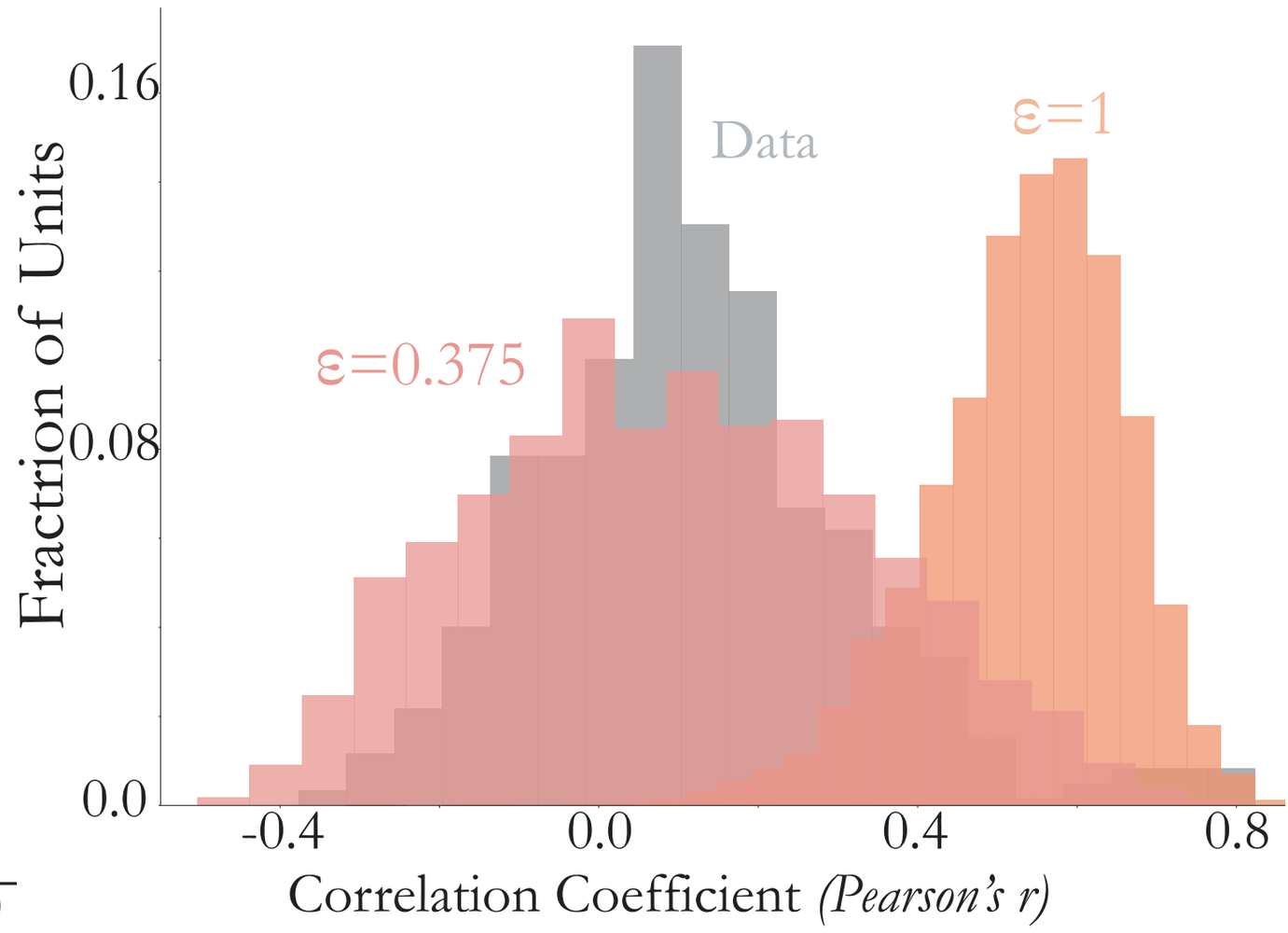
circuity = 0.42

time = 7.4 s

Reward-biased path integration captures remapping of responses in the presence of reward

Slight bias to exploitation preferred

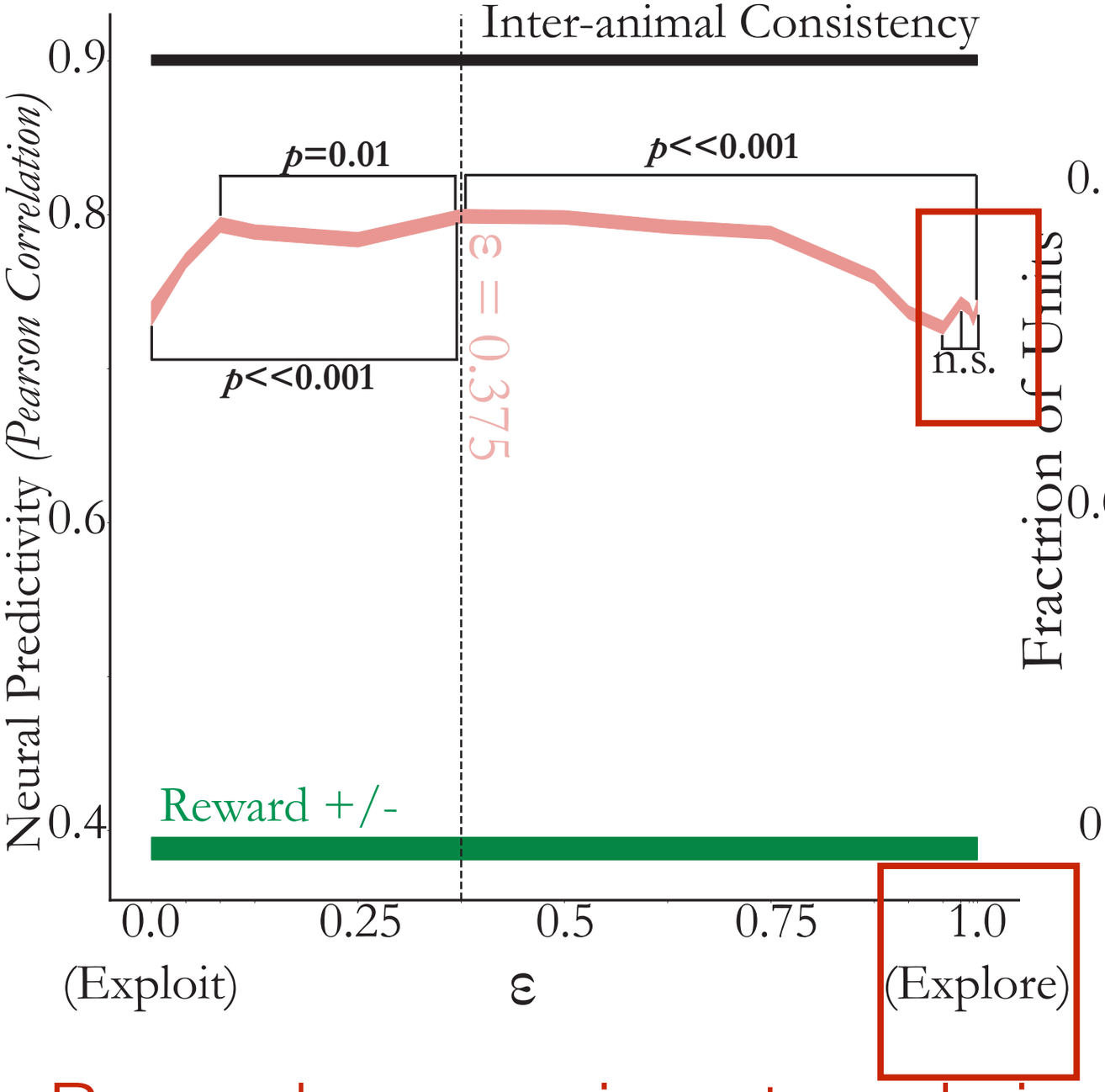


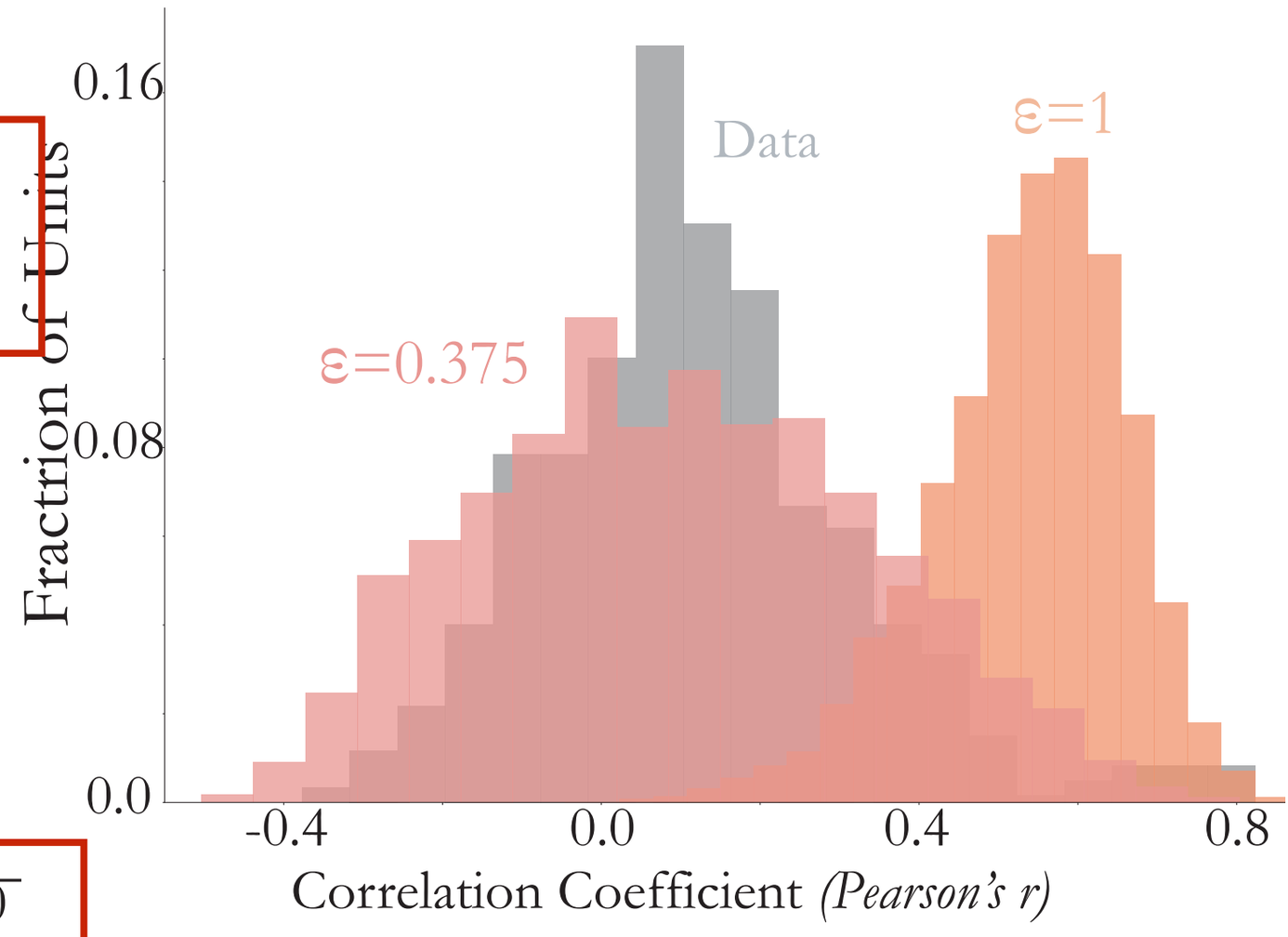


Reward-biased path integrator best captures remapping

Reward-biased path integration captures remapping of responses in the presence of reward

Slight bias to exploitation preferred





Reward remapping strongly input driven (but not completely)

Takeaways

Takeaways

1. **Heterogeneous cells are reliable:** Animals can explain each other quite well, but under a suitably chosen transform class

Takeaways

1. **Heterogeneous cells are reliable:** Animals can explain each other quite well, but under a suitably chosen transform class

Modeling conclusions (under transform class):

2. Classic theoretical model does not quantifiably explain all of the data: NMF, (dimensionality reduction on simulated place cells) is very far from the inter-animal consistency.

Takeaways

1. **Heterogeneous cells are reliable:** Animals can explain each other quite well, but under a suitably chosen transform class

Modeling conclusions (under transform class):

2. Classic theoretical model does not quantifiably explain all of the data: NMF, (dimensionality reduction on simulated place cells) is very far from the inter-animal consistency.

3. *Task Differentiation:* Navigational task training loss gives you higher correlation than NMF loss, especially for the non-grid like units. Intermediate Place Cell representation is important.

Takeaways

1. **Heterogeneous cells are reliable:** Animals can explain each other quite well, but under a suitably chosen transform class

Modeling conclusions (under transform class):

2. Classic theoretical model does not quantifiably explain all of the data: NMF, (dimensionality reduction on simulated place cells) is very far from the inter-animal consistency.

3. *Task Differentiation:* Navigational task training loss gives you higher correlation than NMF loss, especially for the non-grid like units. Intermediate Place Cell representation is important.

4. *Circuit Differentiation:* UGRNN ReLU gives the best match overall, and approaches the inter-animal consistency even when trained in a different environment.

Takeaways

1. **Heterogeneous cells are reliable:** Animals can explain each other quite well, but under a suitably chosen transform class

Modeling conclusions (under transform class):

2. Classic theoretical model does not quantifiably explain all of the data: NMF, (dimensionality reduction on simulated place cells) is very far from the inter-animal consistency.

3. *Task Differentiation:* Navigational task training loss gives you higher correlation than NMF loss, especially for the non-grid like units. Intermediate Place Cell representation is important.

4. *Circuit Differentiation:* UGRNN ReLU gives the best match overall, and approaches the inter-animal consistency even when trained in a different environment.

Overall Conclusion: A process of biological performance optimization directly shaped the neural mechanisms in MEC as a whole *(cell types form a unified continuum in the same path integrator network)*.

Outline

- Recurrent Connections in the Primate Ventral Stream
- Goal-Driven Models of Mouse Visual Cortex
- Heterogeneity in Rodent Medial Entorhinal Cortex
- Building and Identifying Learning Rules

Goal-Driven Modeling - Three Primary Components

1. *"circuit"*

A = *architecture class*

2. *"behavior"*

L = *loss function* ***D*** = *dataset*

3. *"plasticity mechanism"*

Learning Rule

$$\underset{a \in \mathcal{A}}{\operatorname{argmin}} [L(p_a^*)]$$

genetic algorithms

where p is result of*

backprop

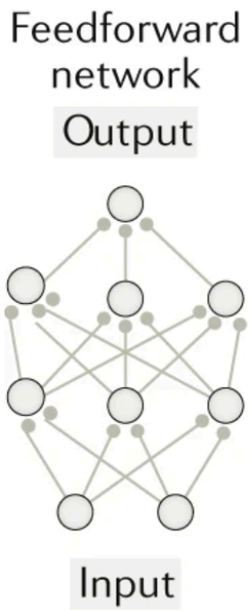
$$\frac{dp_a}{dt} = -\lambda(t) \cdot \langle \nabla_{p_a} L(x) \rangle_{x \in \mathcal{D}}$$

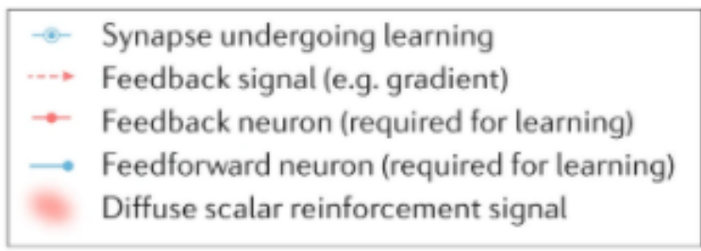
D. Kunin*, **A. Nayebi***, J. Sagastuy-Brena*, *et al.*
Two Routes to Scalable Credit Assignment without Weight Symmetry. *ICML 2020*

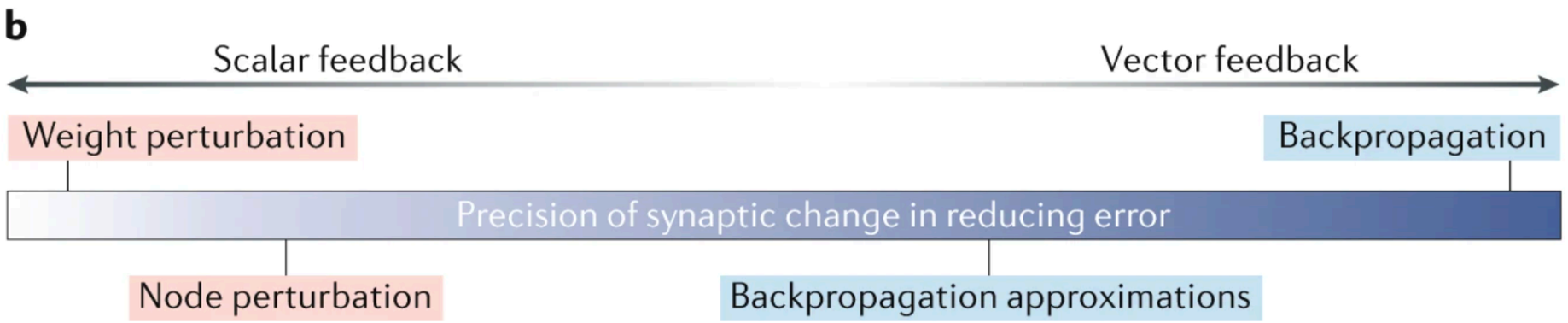
A. Nayebi*, S. Srivastava*, S. Ganguli, DLK Yamins. Identifying Learning Rules from Neural Network Observables. *NeurIPS 2020 (spotlight)*

J. Melander*, **A. Nayebi***, *et al.* Distinct *in vivo* Dynamics of Excitatory Synapses Onto Cortical Pyramidal Neurons and Inhibitory Interneurons. *Cell Reports 2021*

Spectrum of learning strategies

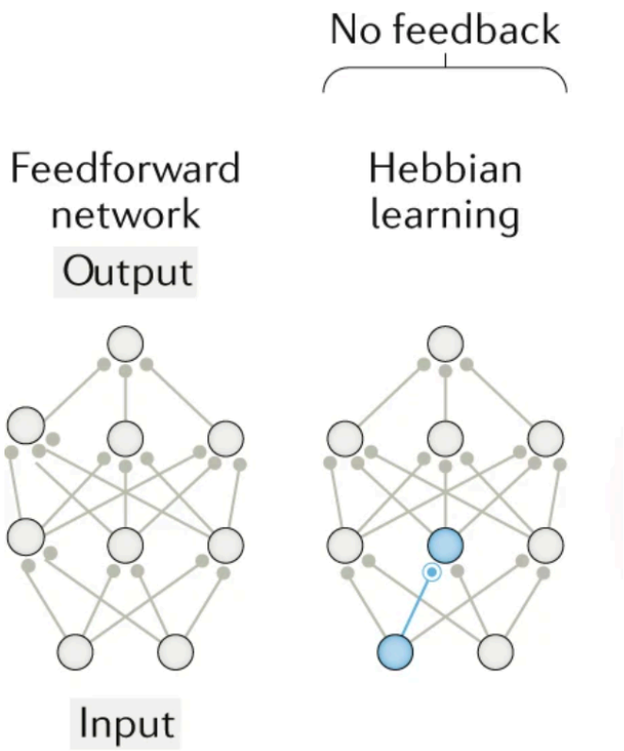


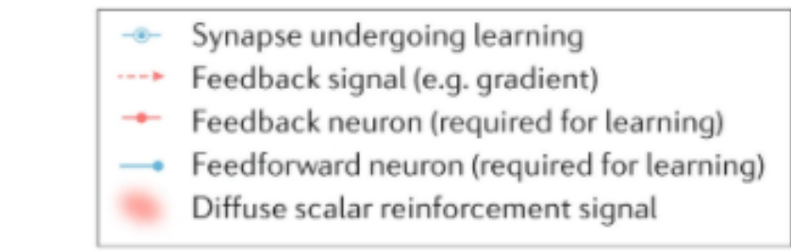


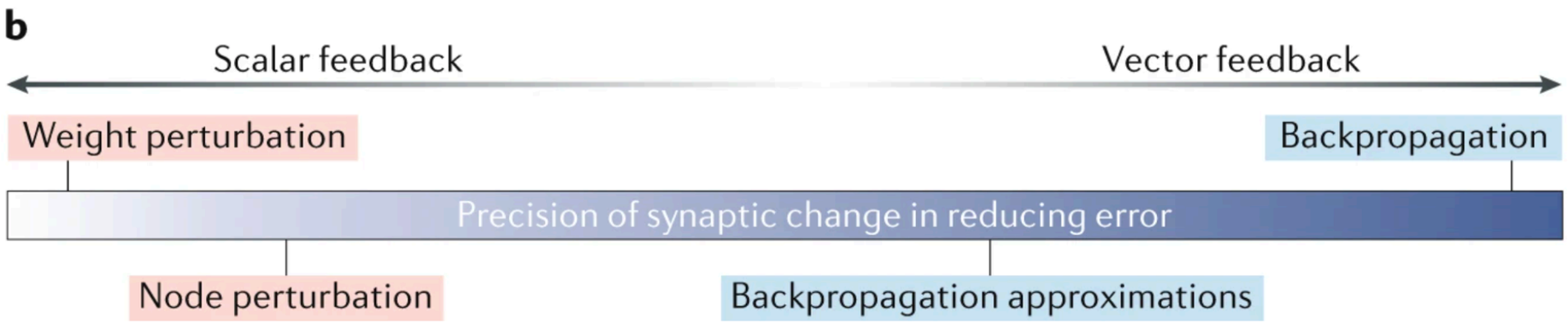


Lillicrap et al.
Nat. Rev. Neurosci. *(2020)*

Spectrum of learning strategies

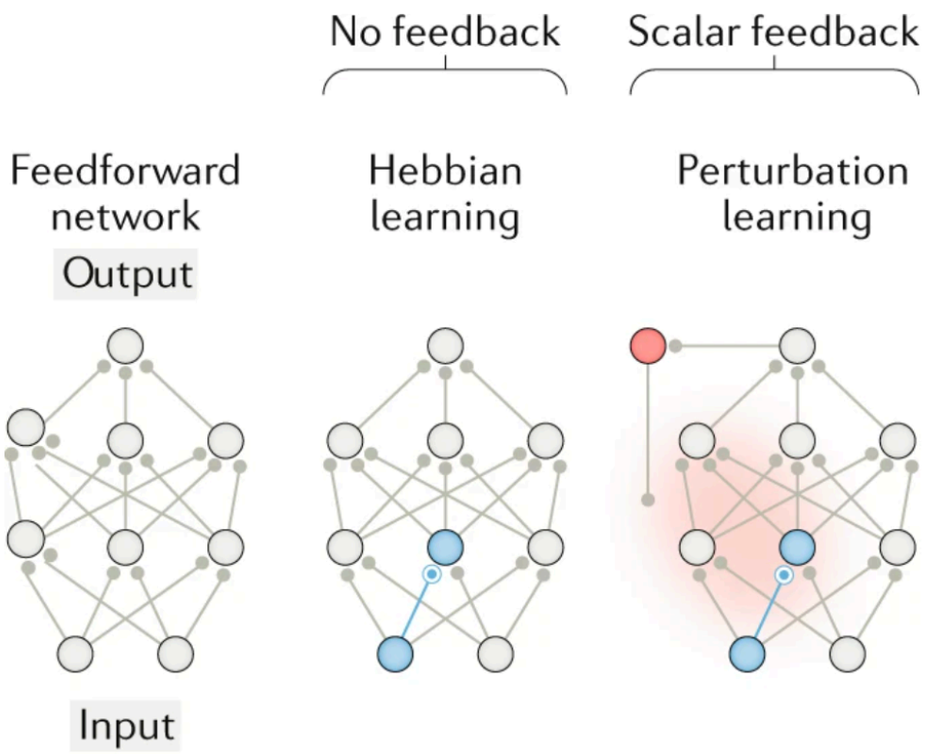


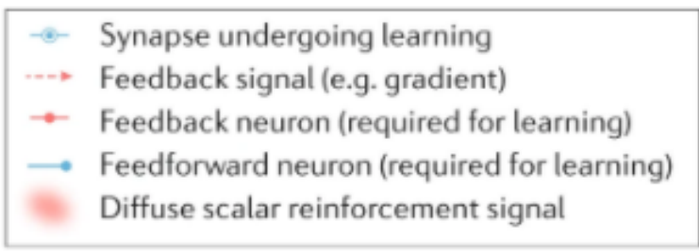


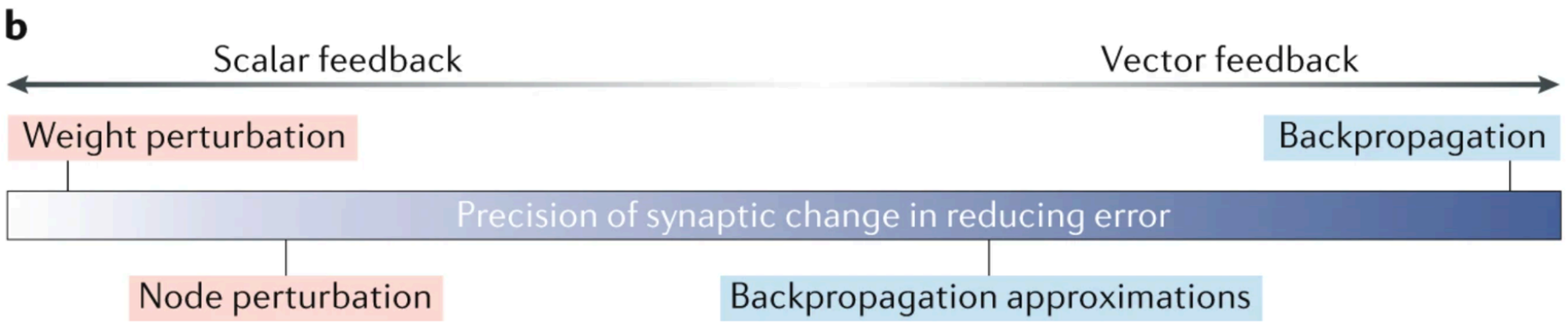


Lillicrap et al.
Nat. Rev. Neurosci. *(2020)*

Spectrum of learning strategies

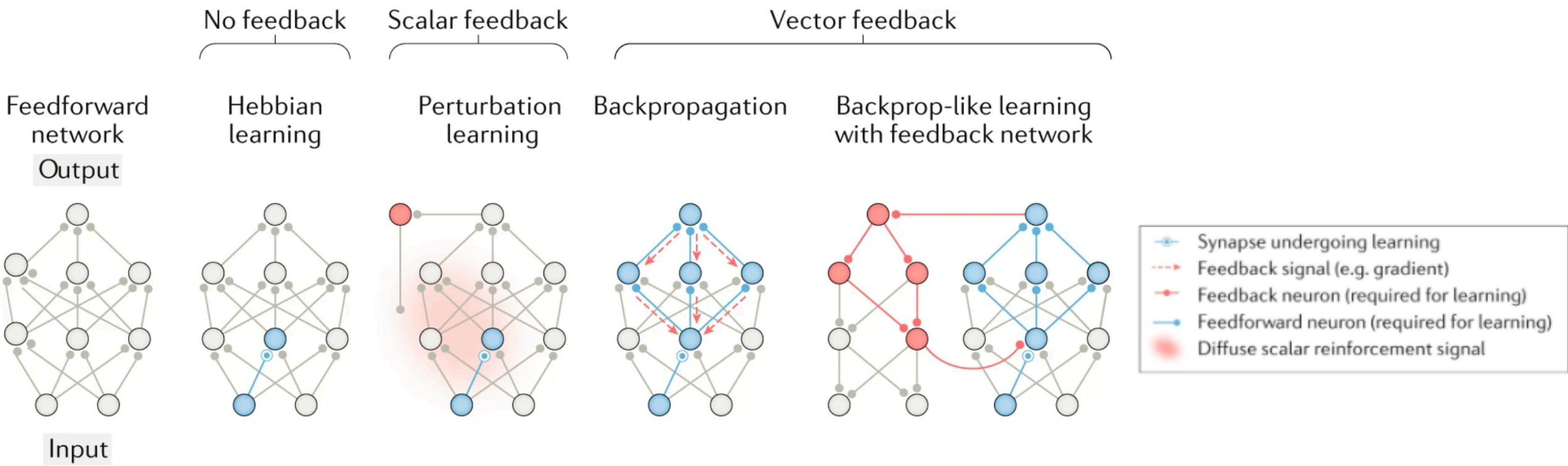


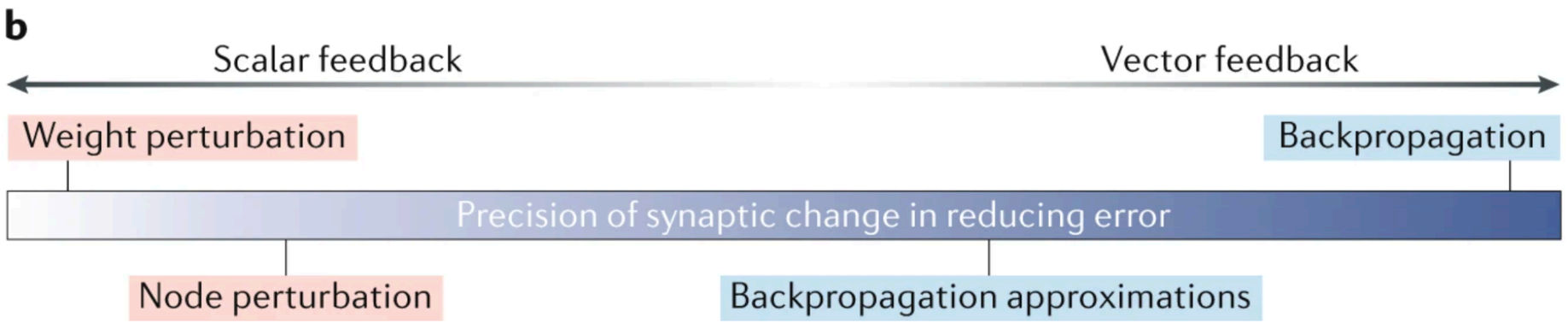




Lillicrap et al.
Nat. Rev. Neurosci. *(2020)*

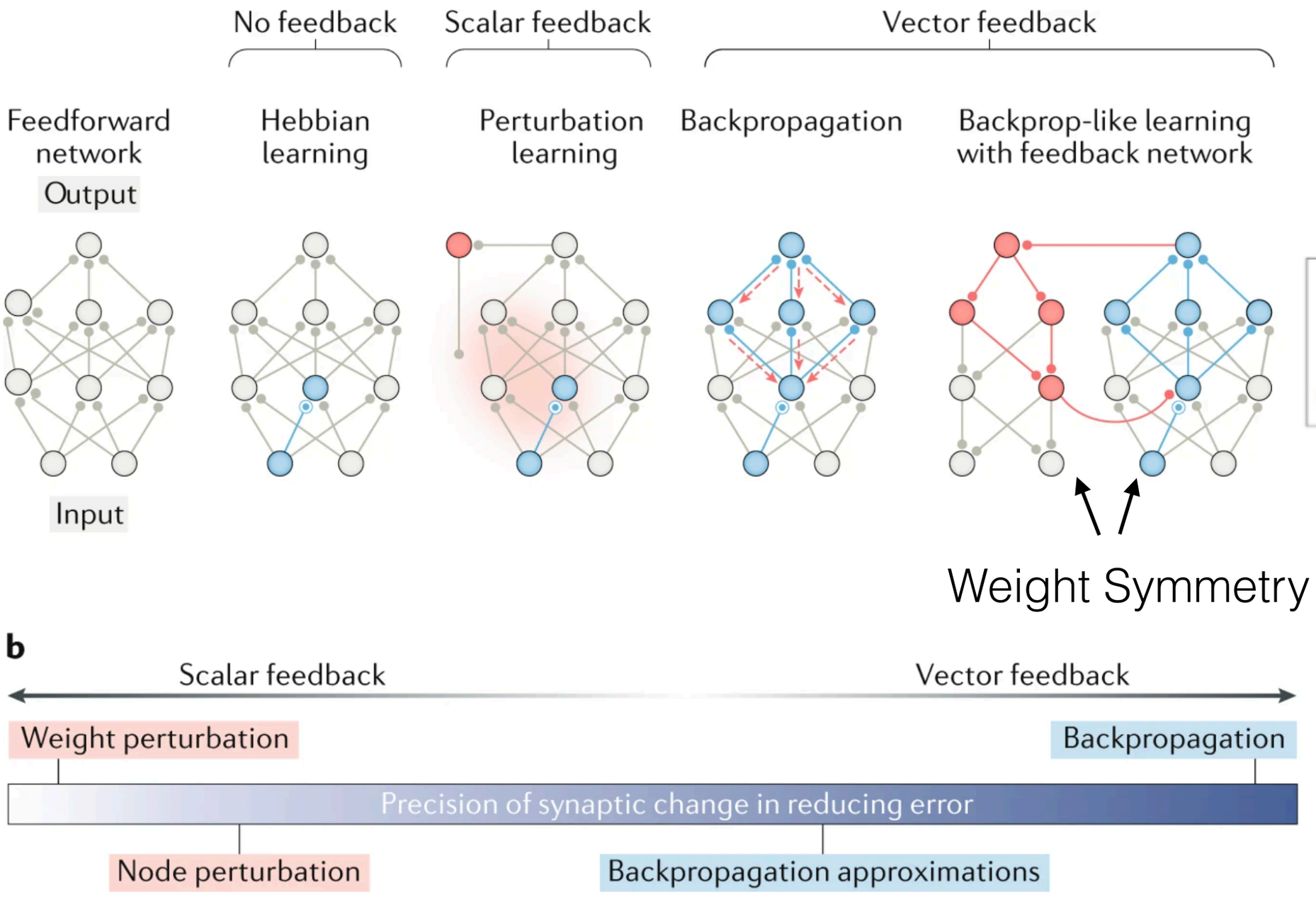
Spectrum of learning strategies





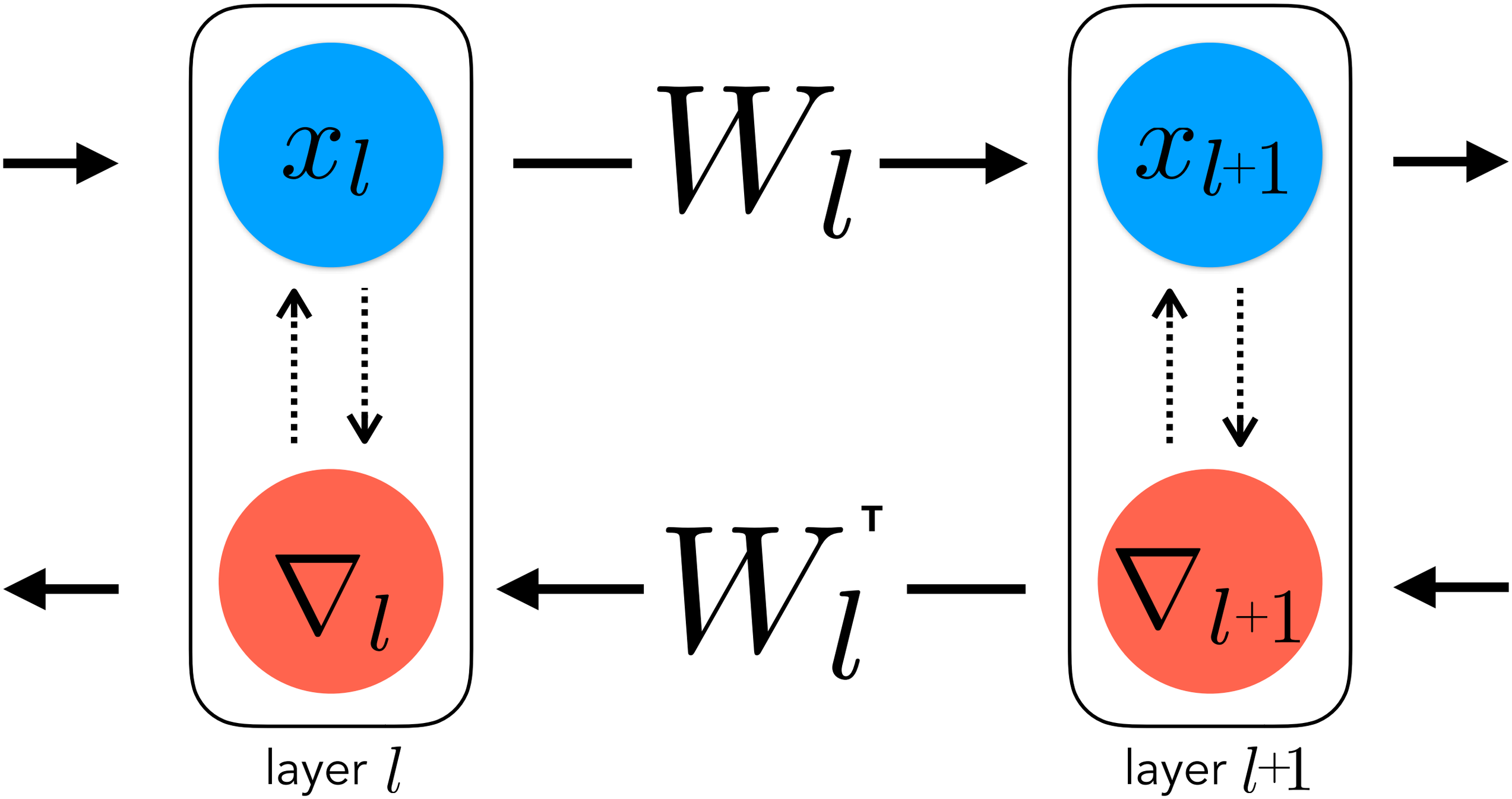
Lillicrap et al.
Nat. Rev. Neurosci. *(2020)*

Spectrum of learning strategies

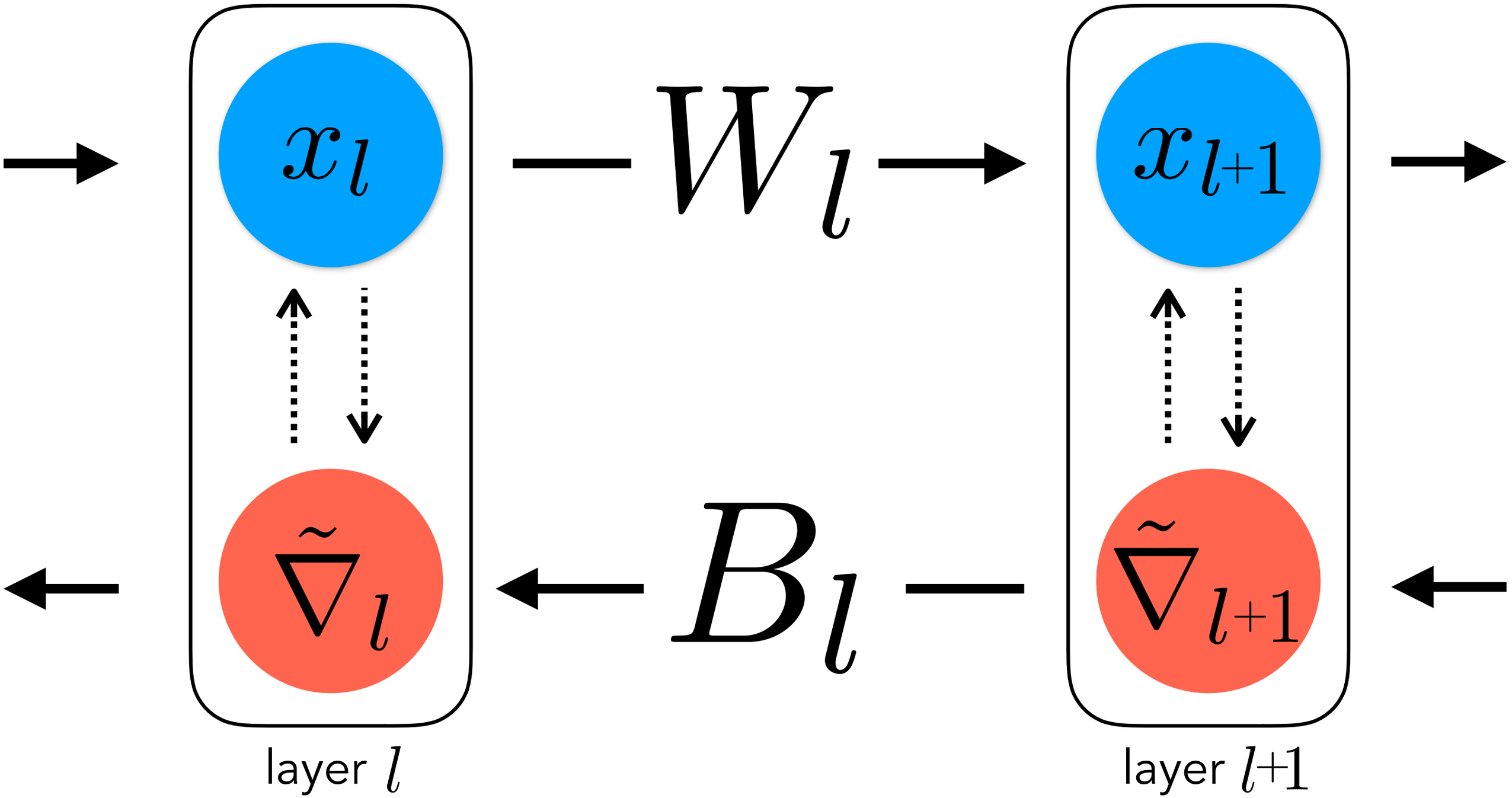


Lillicrap et al.
Nat. Rev. Neurosci. *(2020)*

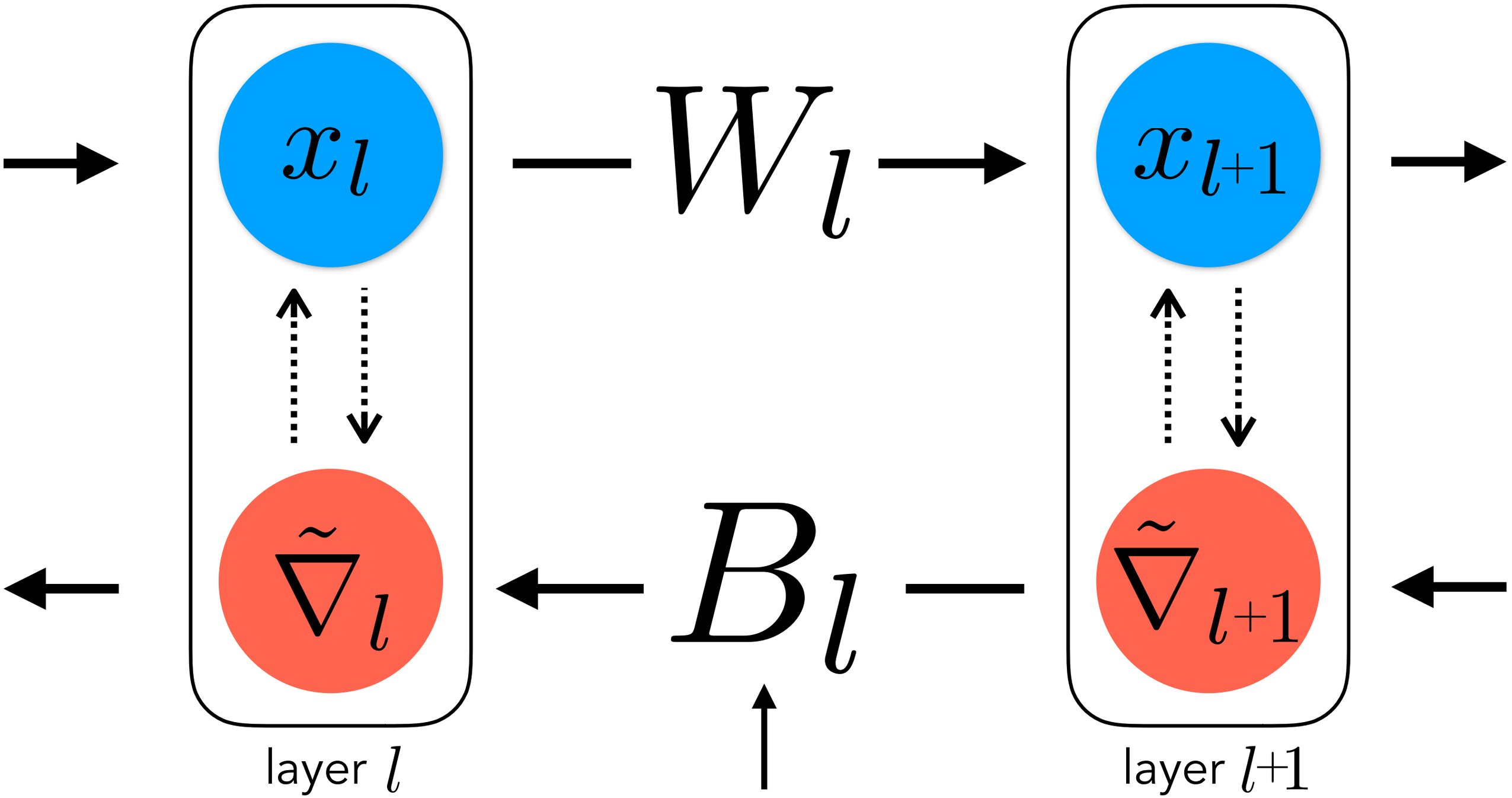
Relaxing the weight symmetry requirement



Relaxing the weight symmetry requirement



Feedback Alignment (FA)



Feedback Alignment (FA): B is random

Lillicrap et al.
Nat. Commun. *(2016)*

Comparing Feedback Alignment to Backprop

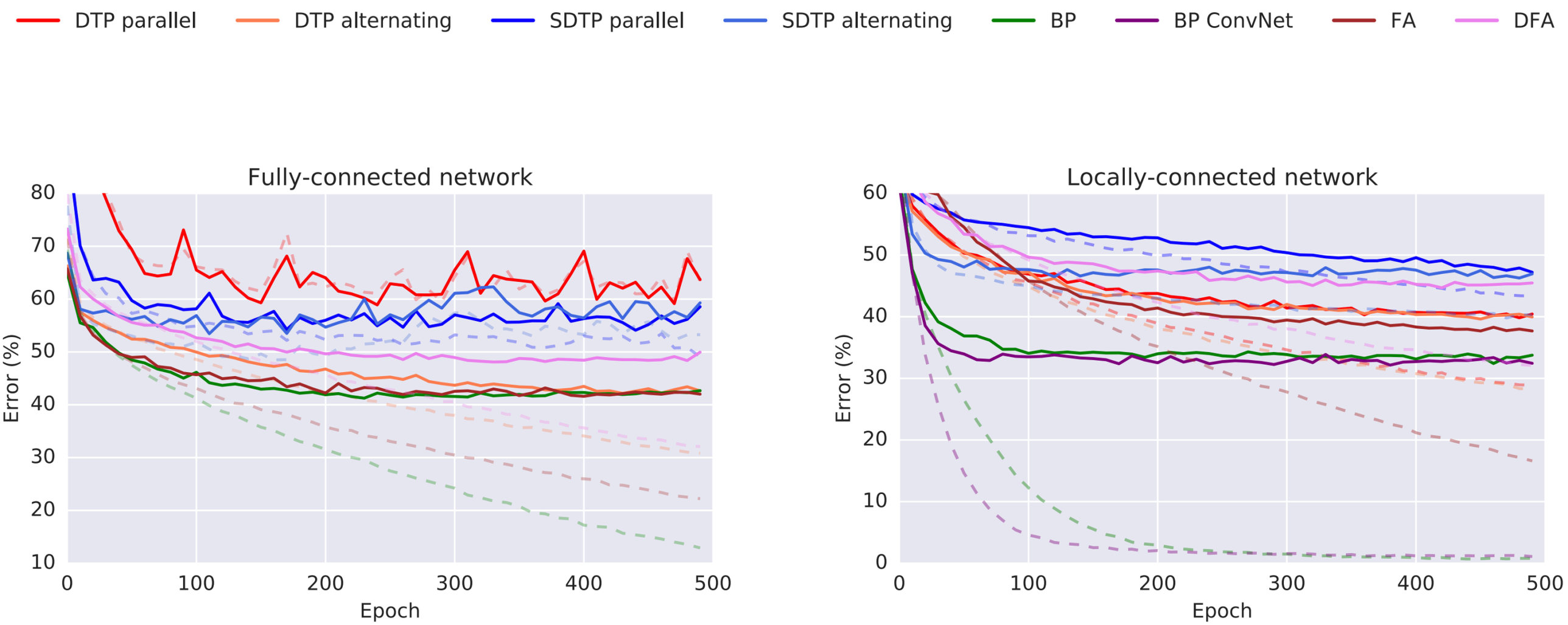


Figure 2: Train (dashed) and test (solid) classification errors on CIFAR.

Bartunov et al.
NeurIPS *(2018)*

Scales as Backprop does on simple tasks

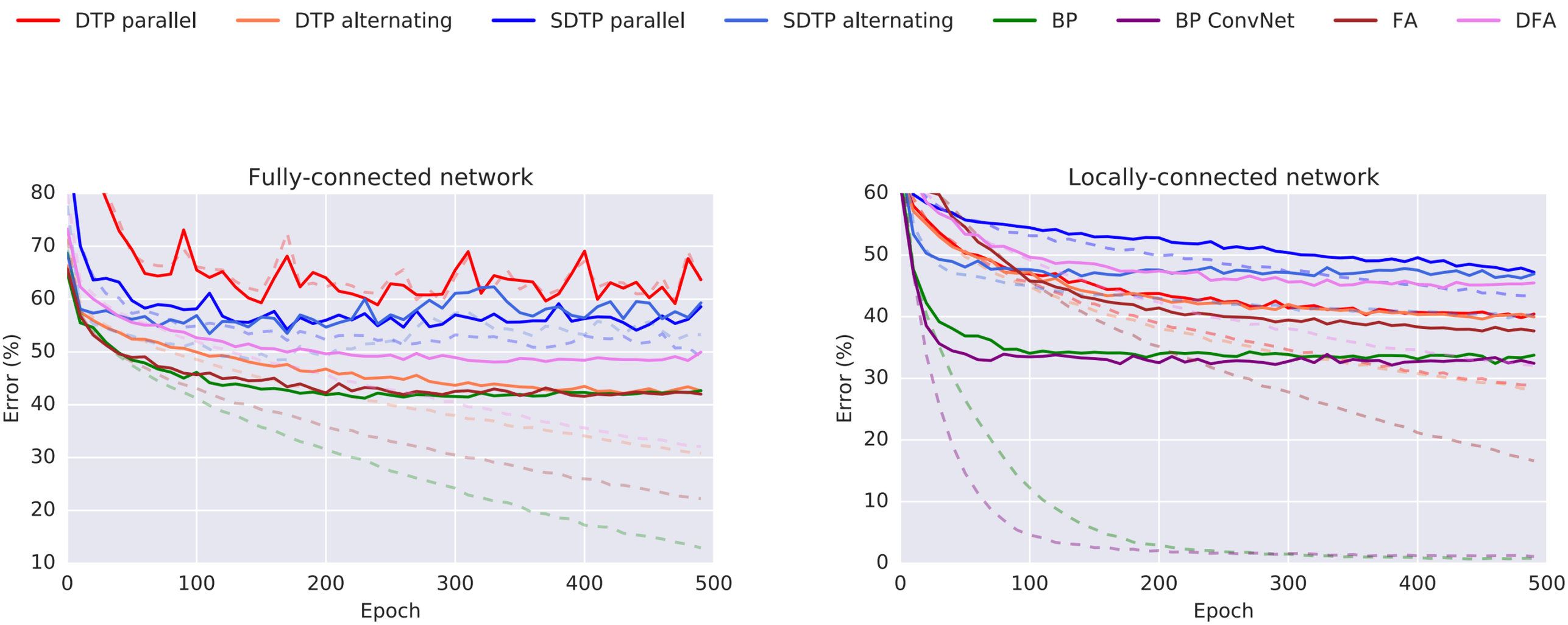
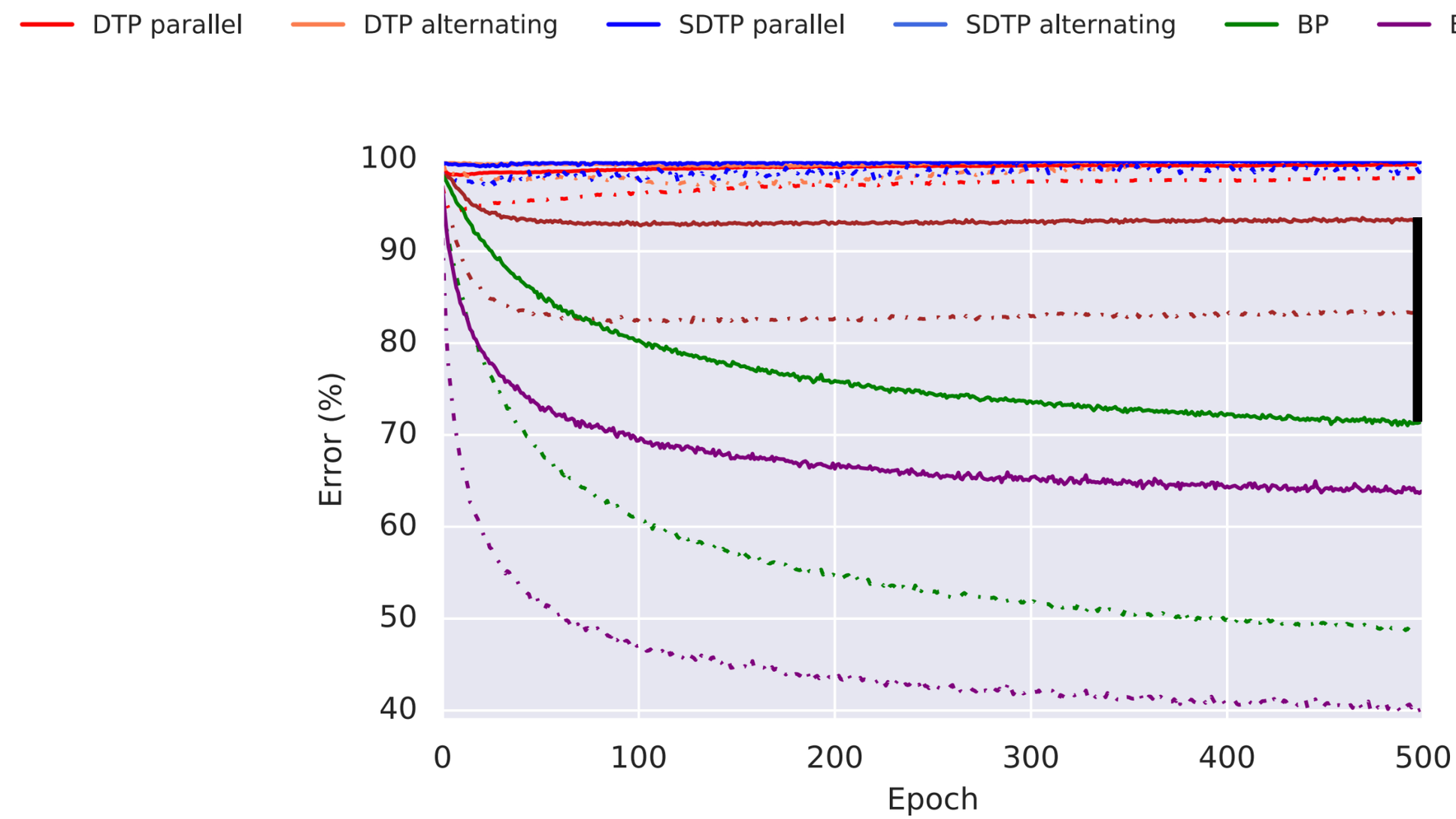


Figure 2: Train (dashed) and test (solid) classification errors on CIFAR.

Similar performance between FA and Backprop on small tasks.

Bartunov et al.
NeurIPS *(2018)*

Does *not* scale as Backprop does on harder tasks

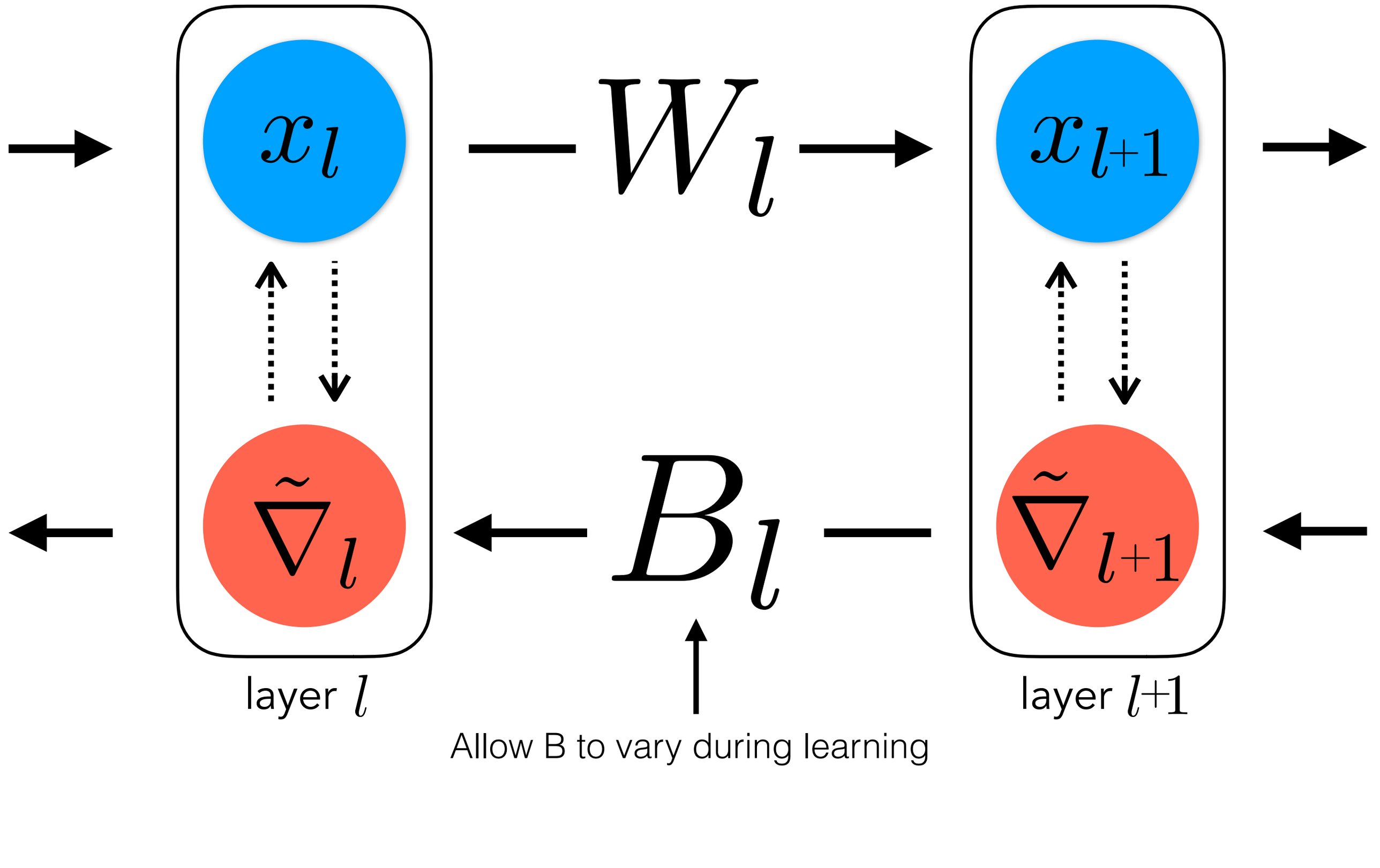


Gap widens on ImageNet

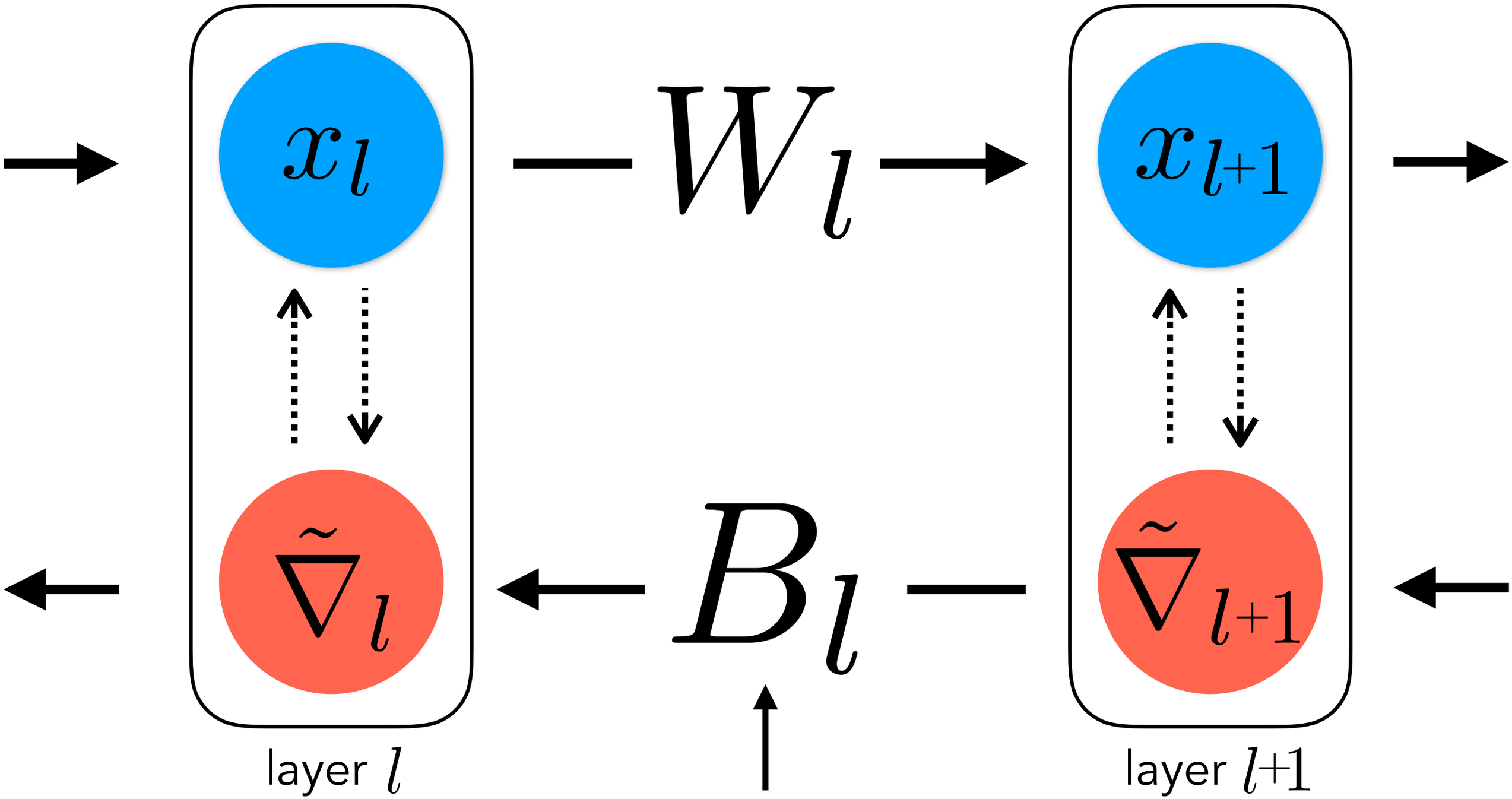
Figure 3: Top-1 (solid) and Top-5 (dotted) test errors on ImageNet. Color legend is the same as for figure 2.

Bartunov et al.
***NeurIPS** (2018)*

Imposing dynamics on the backward weights



Weight Mirror (WM)

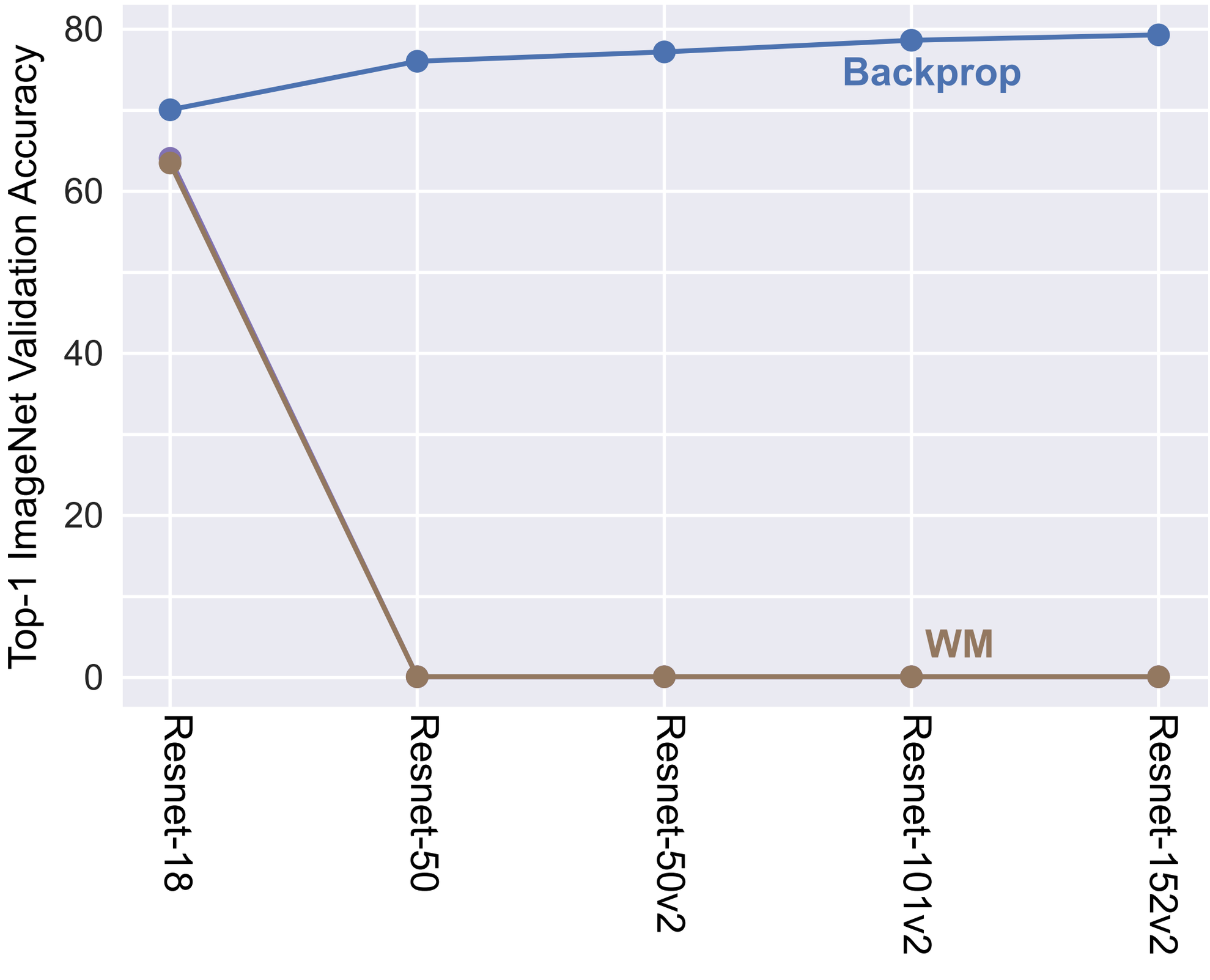


Weight Mirror (WM): B gradually aligns with W

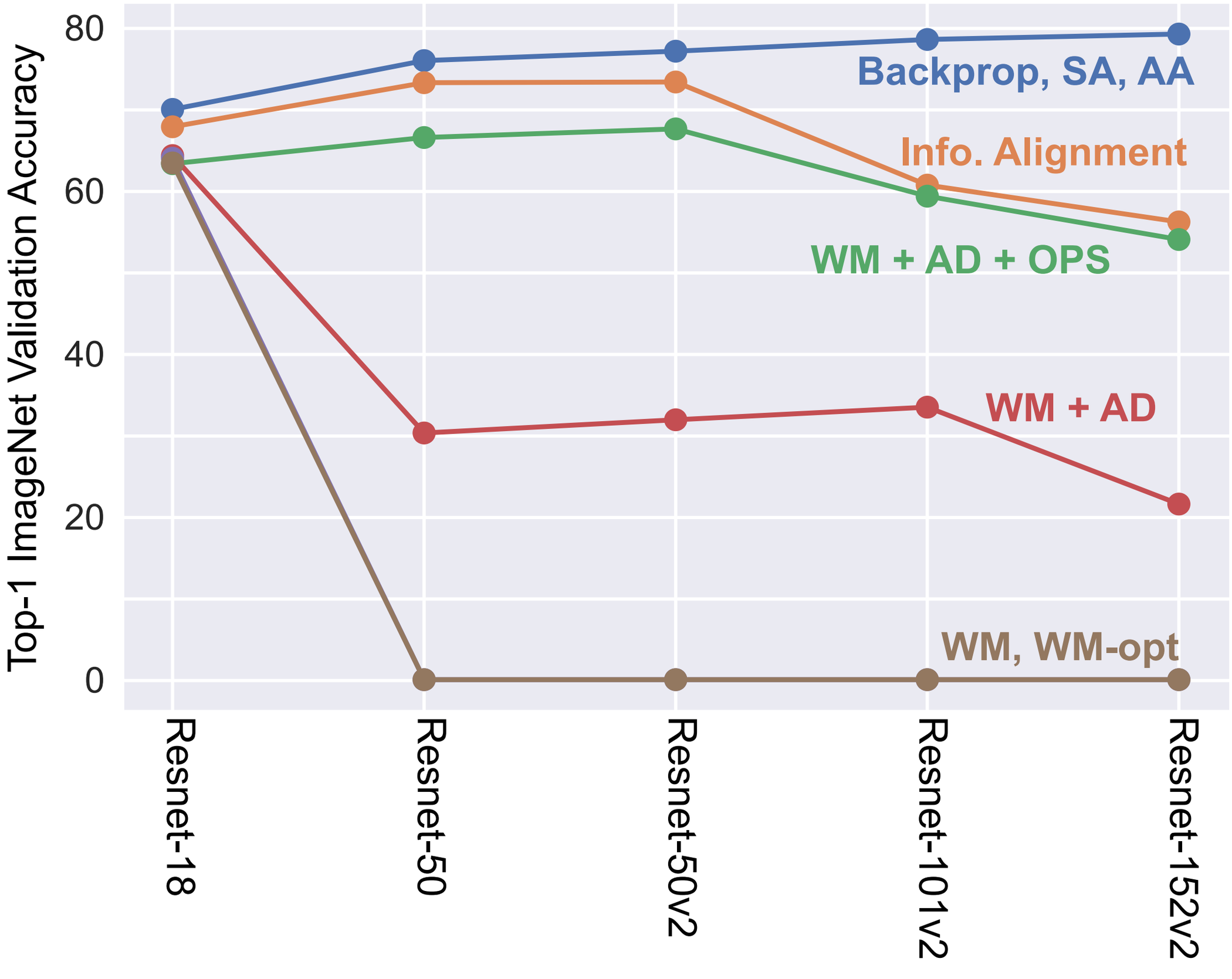
~~**Feedback Alignment (FA):** B is random~~

Akrout et al.
***NeurIPS** (2019)*

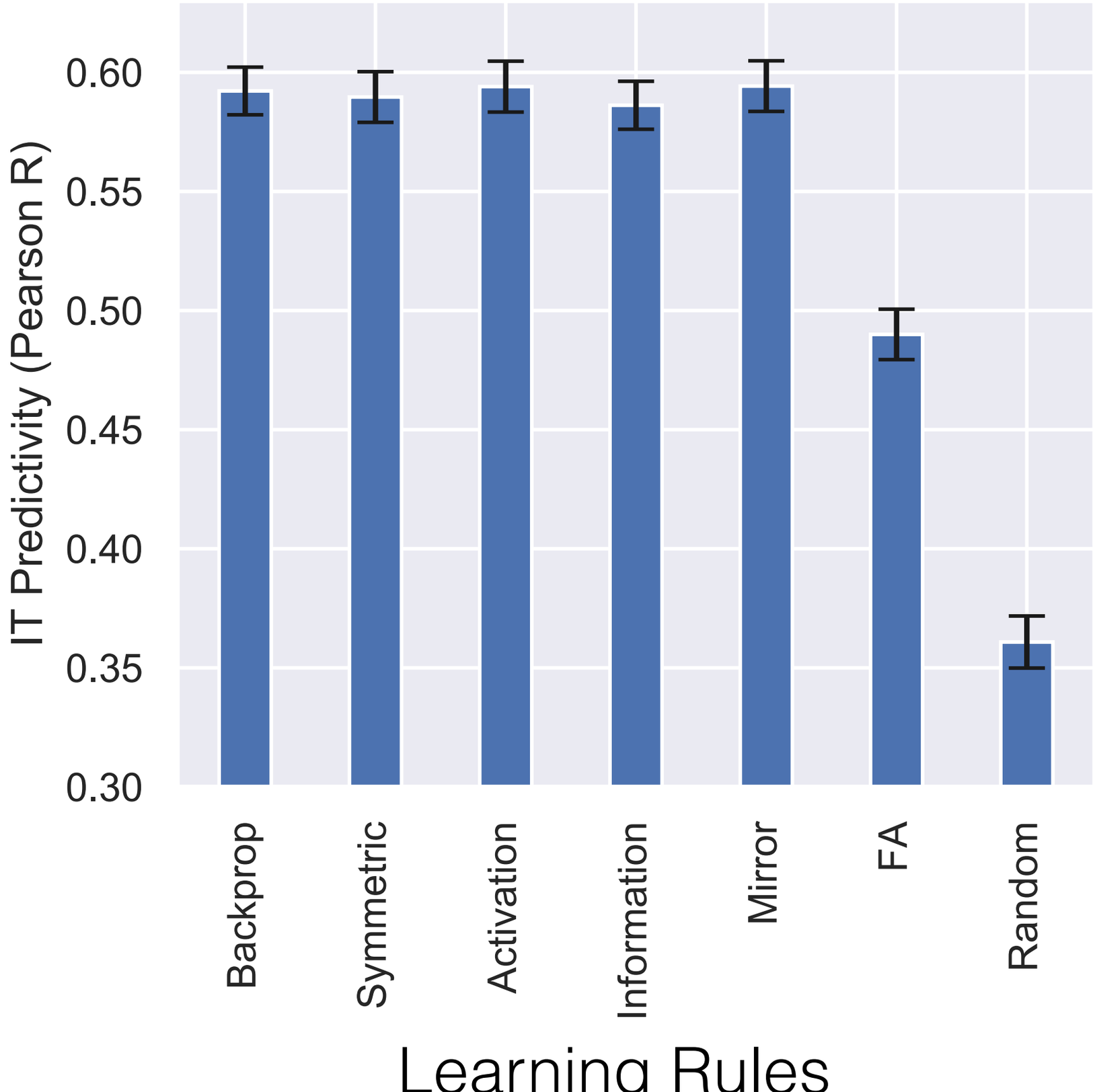
Weight Mirror does not transfer across architectures



Searching alternatives to Backprop scales across architectures

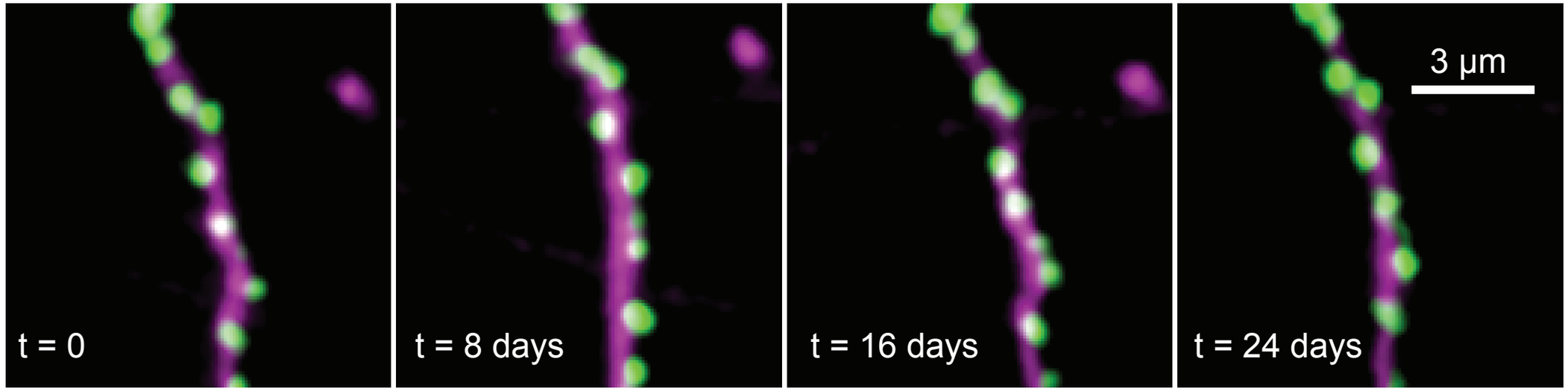


Current neural data is insufficient to separate these alternatives



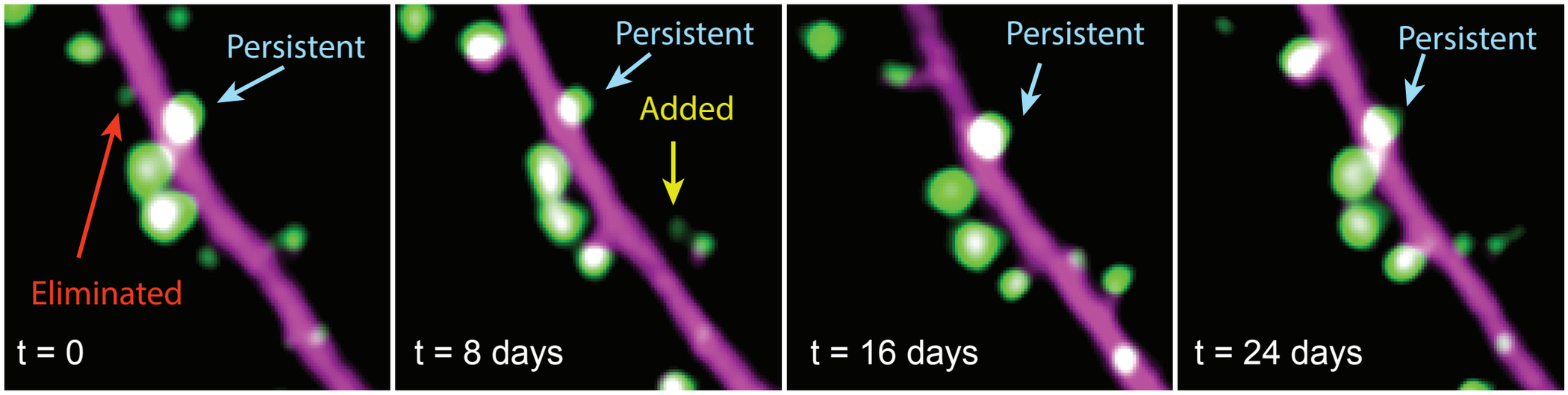
What can we measure?

PV



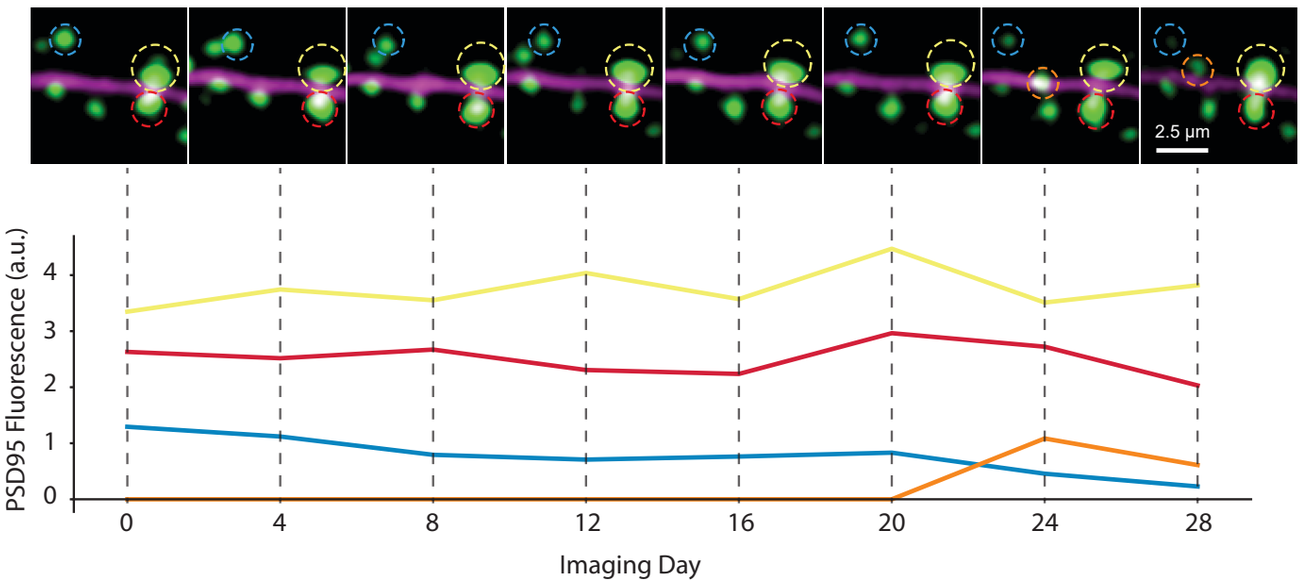
Joshua Melander

L2/3 Pyr

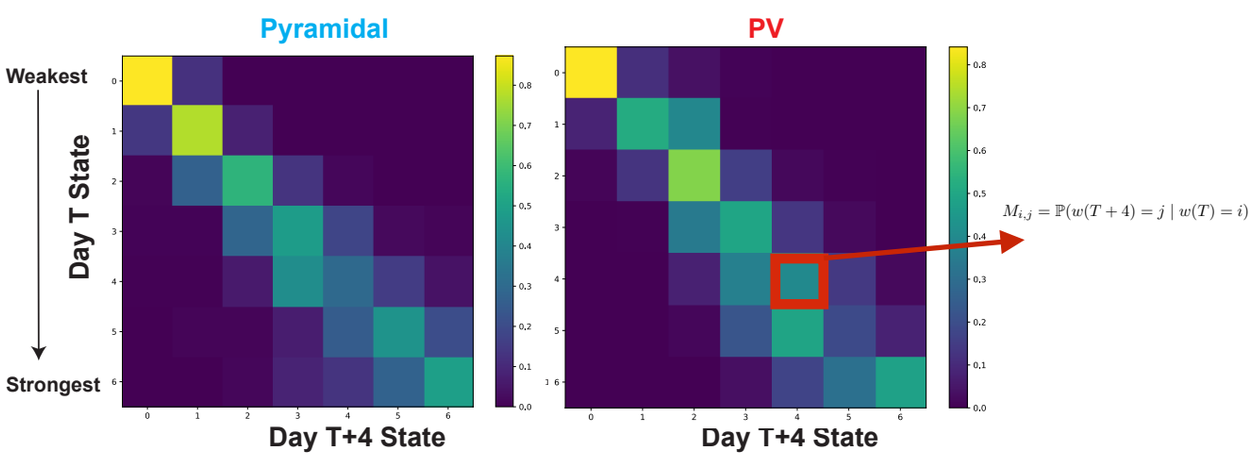


What can we measure?

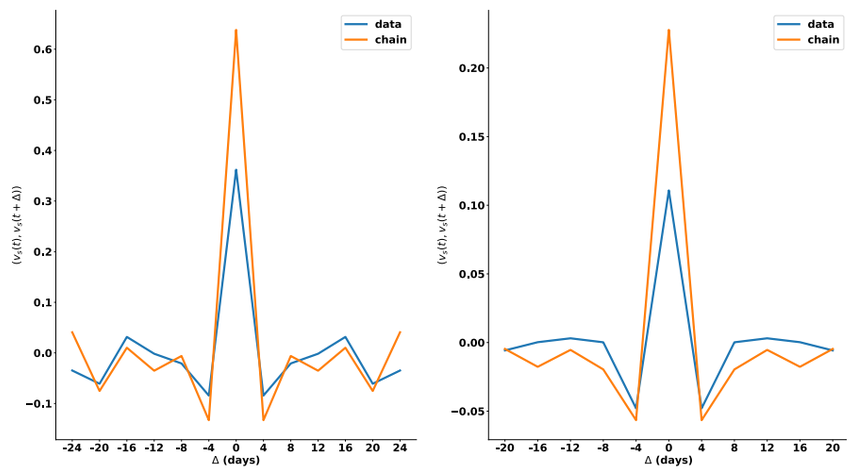
$$w(T+4) = w(T) \cdot (1+\varepsilon)$$

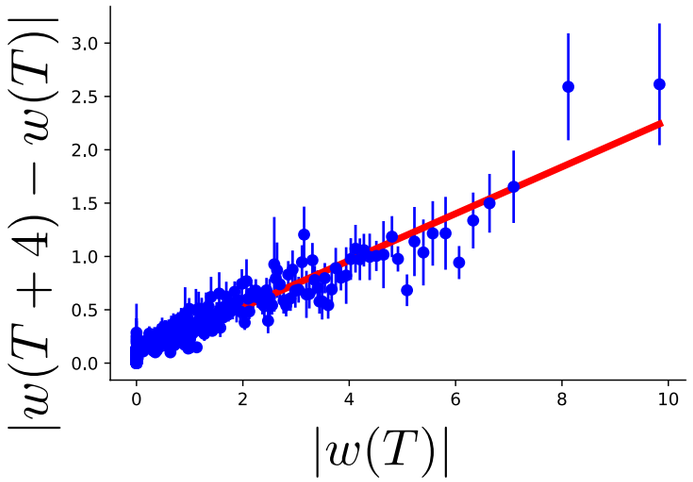


- Using a PSD-95 ENABLED strategy, month-long *in vivo* imaging of populations of synaptic strength onto both excitatory and inhibitory cortical cell types, hundreds of synapses



- Evidence of a stable baseline of multiplicative (Markovian) synaptic dynamics across excitatory synapses onto multiple cell types, but strong additive component inhibitory PV+ neurons





"Virtual Experimental" Approach

What would you need to measure to reliably distinguish *classes* of learning rules?

"Virtual Experimental" Approach — Generate Predictions

What would you need to measure to reliably distinguish *classes* of learning rules?

With artificial neural networks, we can measure anything we want & know the ground truth learning rule we trained the model with

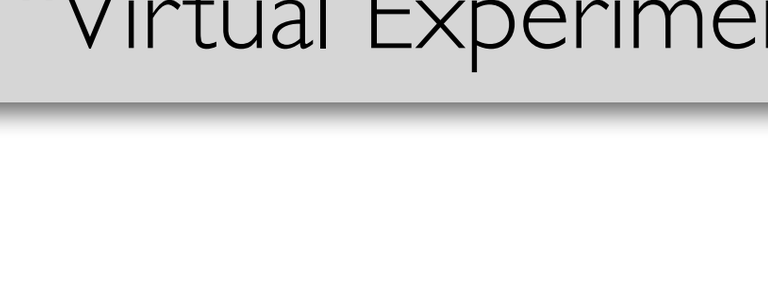
"Virtual Experimental" Approach — Generate Predictions

What would you need to measure to reliably distinguish *class*es of learning rules?

Hypothesis: measuring post-synaptic activities from a neural circuit on the order of several hundred units, may provide a good basis on which to identify learning rules.

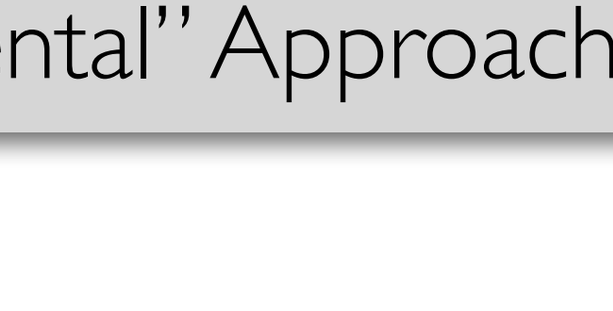
"Virtual Experimental" Approach

Data Generation



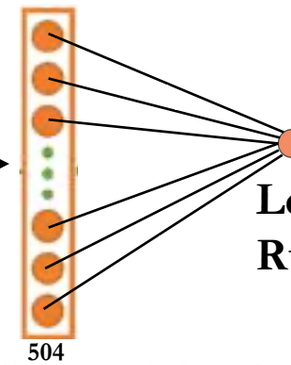
10 architectures, 4 tasks, 12 hyperparameter settings, 4 learning rules

Observable Statistics



Training Epochs

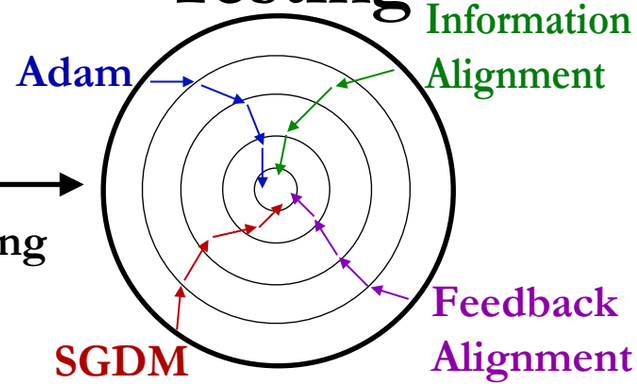
Classifier Training



Cross-validated, trained on split of observable trajectories

Learning Rule

Classifier Testing

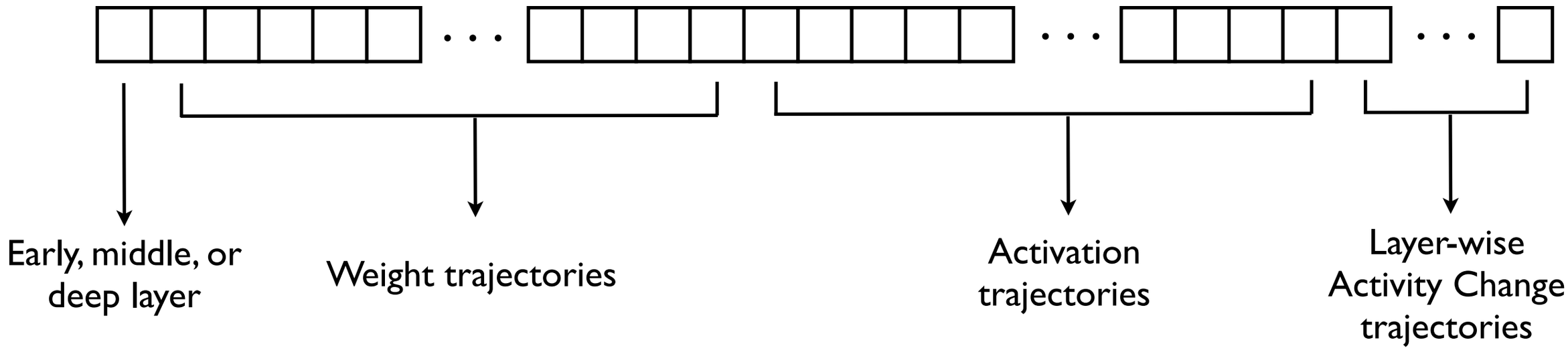


Tested on remaining data, held-out cases, etc.

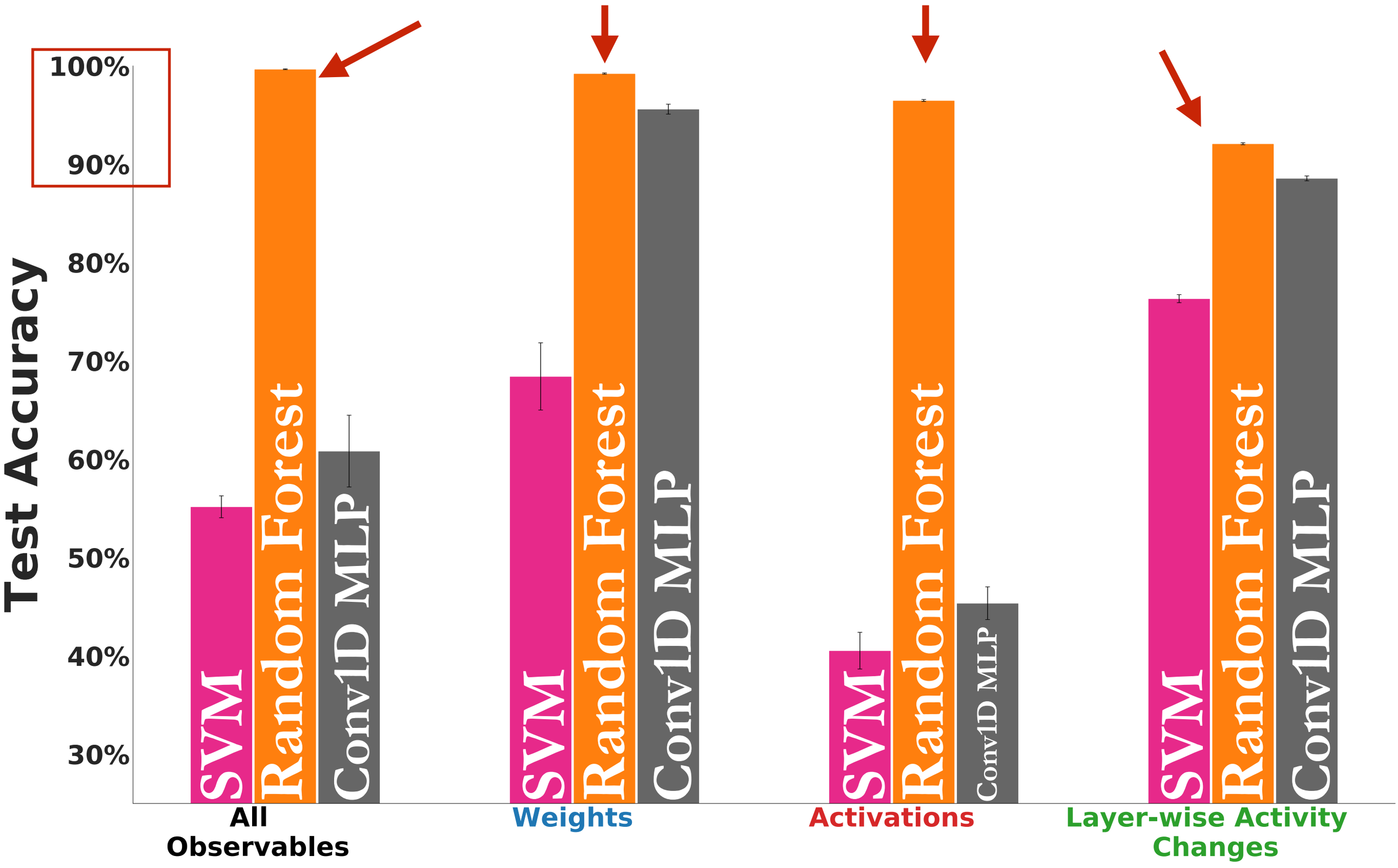
Framing it as a classification problem

How well can we do by framing it as a classification problem?

Sample is constructed from one layer of a trained network



General separability problem *is* tractable

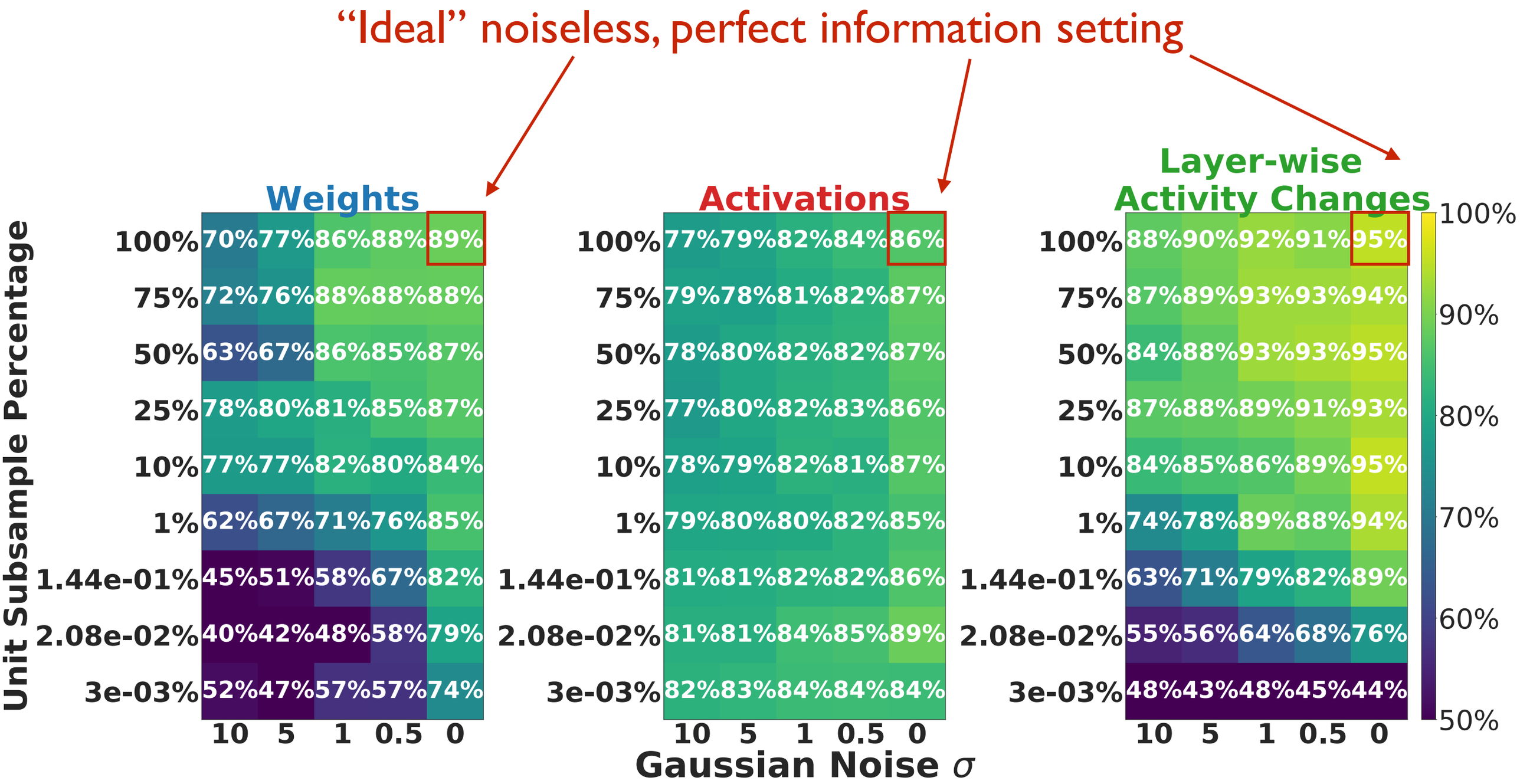


Different experimental tools have different limitations

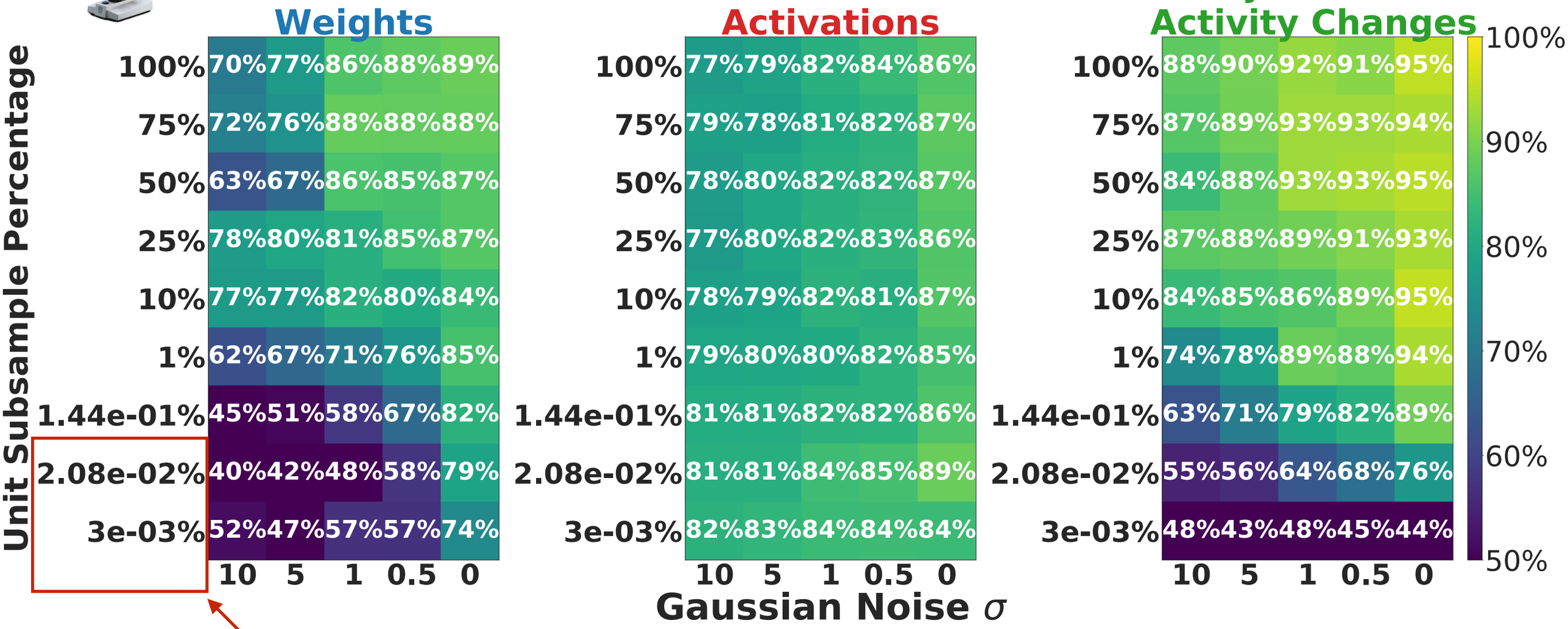
Optical imaging techniques usually give us simultaneous access to thousands of units but can have lower temporal resolution and signal-to-noise

Electrophysiological recordings can have higher signal-to-noise and better temporal resolution, but can lack the coverage to thousands of units

Modeling unit subsampling and measurement noise

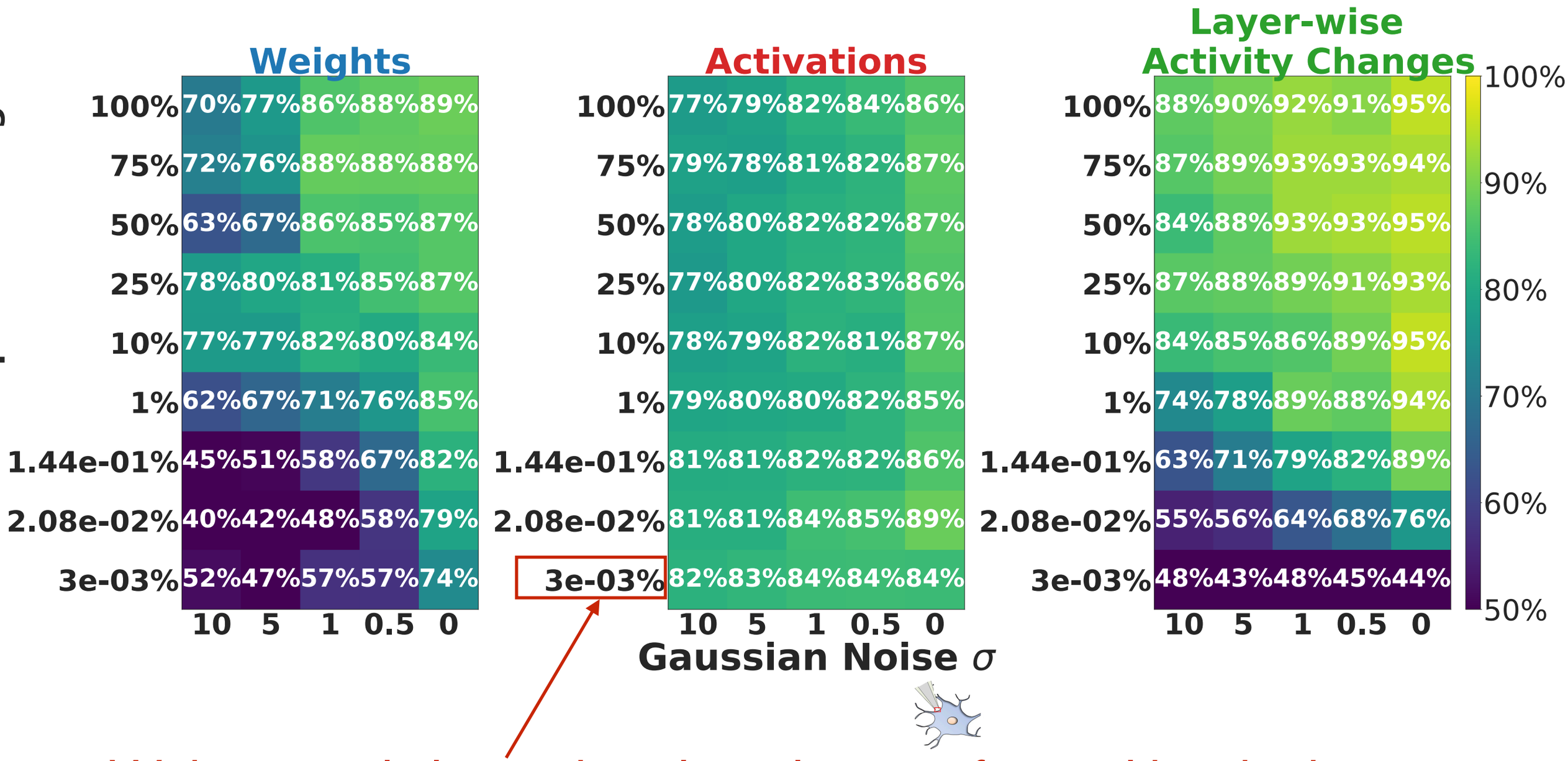


Weights are *not* robust to measurement noise and unit undersampling



Within typical imaging range of several hundred to several thousand synapses

Activations are the most robust to measurement noise and unit undersampling



Within typical electrophysiological range of several hundred units

Takeaways

Biologically implausible weight symmetry is not a strict requirement for effective learning — we can build performant relaxations with the appropriate layer-wise learning circuit motif

Takeaways

Biologically implausible weight symmetry is not a strict requirement for effective learning — we can build performant relaxations with the appropriate layer-wise learning circuit motif

We can identify learning rules *only* on the basis of aggregate statistics of observable measures: weights, activations, or layer-wise activity changes

Takeaways

Biologically implausible weight symmetry is not a strict requirement for effective learning — we can build performant relaxations with the appropriate layer-wise learning circuit motif

We can identify learning rules *only* on the basis of aggregate statistics of observable measures: weights, activations, or layer-wise activity changes

This observation holds across various scenarios of experimental realism of certain held-out input classes, trajectory undersampling, and unit undersampling & measurement noise, with *network activations being the most robust*

Broad Takeaways

- Recurrent Connections in the Primate Ventral Stream

Enable a high performing network to fit in cortex, attaining computational power through temporal rather than spatial complexity during core object recognition.

- Goal-Driven Models of Mouse Visual Cortex

Low-resolution, shallow network that makes best use of the mouse's limited resources to create a light-weight, general-purpose visual system.

- Heterogeneity in Rodent Medial Entorhinal Cortex

Heterogeneous cells are <u>not</u> functionally segregated from classic cell types, but rather form a continuum of cells within a single unified network that naturally encompasses grid, border, and heterogeneous cells.

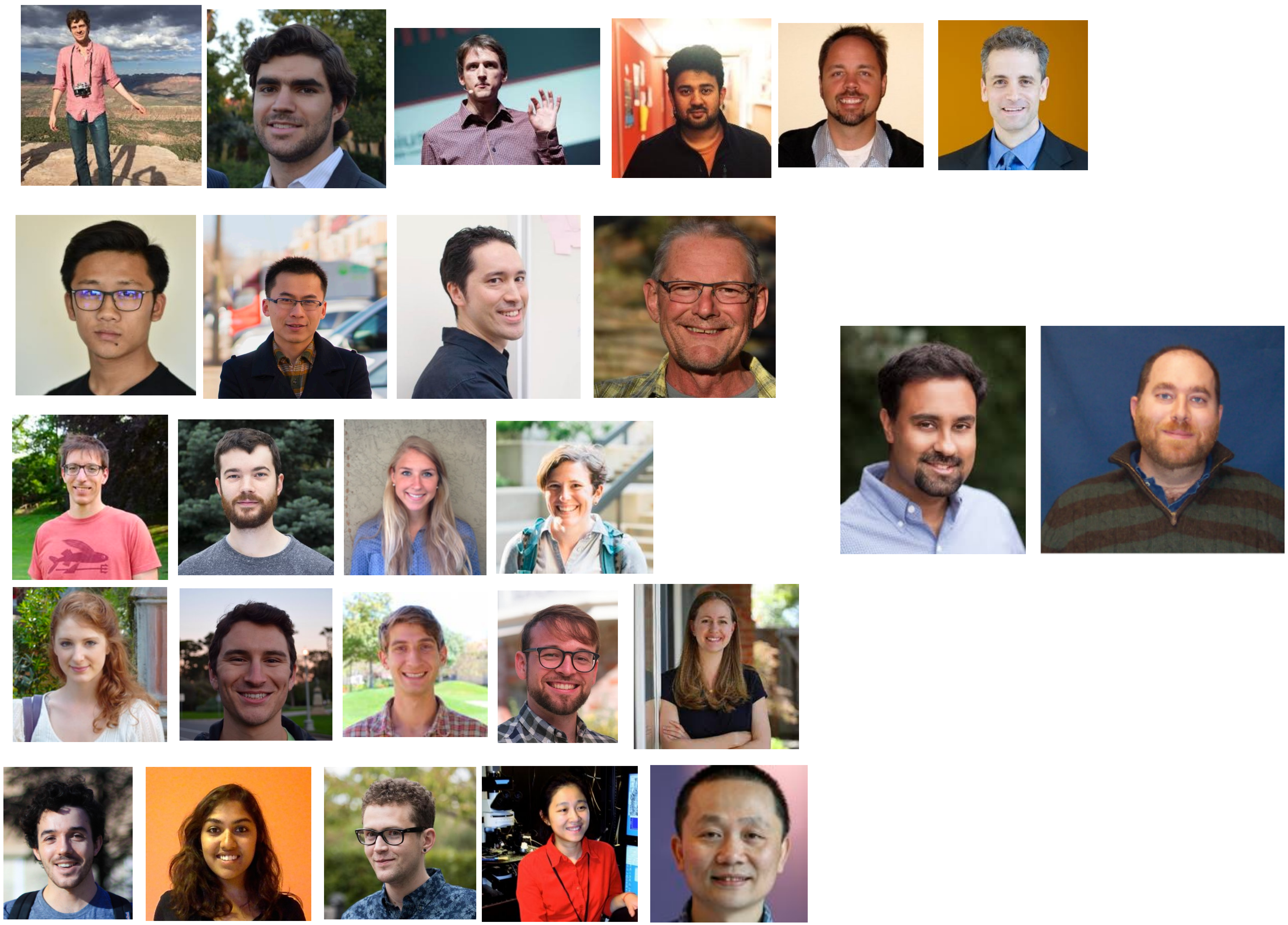
- Building and Identifying Learning Rules

Weight symmetry is <u>not</u> required for effective learning, and can be relaxed with the appropriate learning circuit. We can differentiate these hypotheses for learning circuits from (imperfect) recordings of post-synaptic activations.

Acknowledgements

Thanks to my awesome collaborators!

Contact:
anayebi@stanford.edu



Funding:

Neurosciences PhD Program

Stanford Mind, Brain, Computation and Technology Training Program,
Wu Tsai Neurosciences Institute