A Model-Based Approach Towards Identifying the Brain's Learning Algorithms

MBCT Seminar

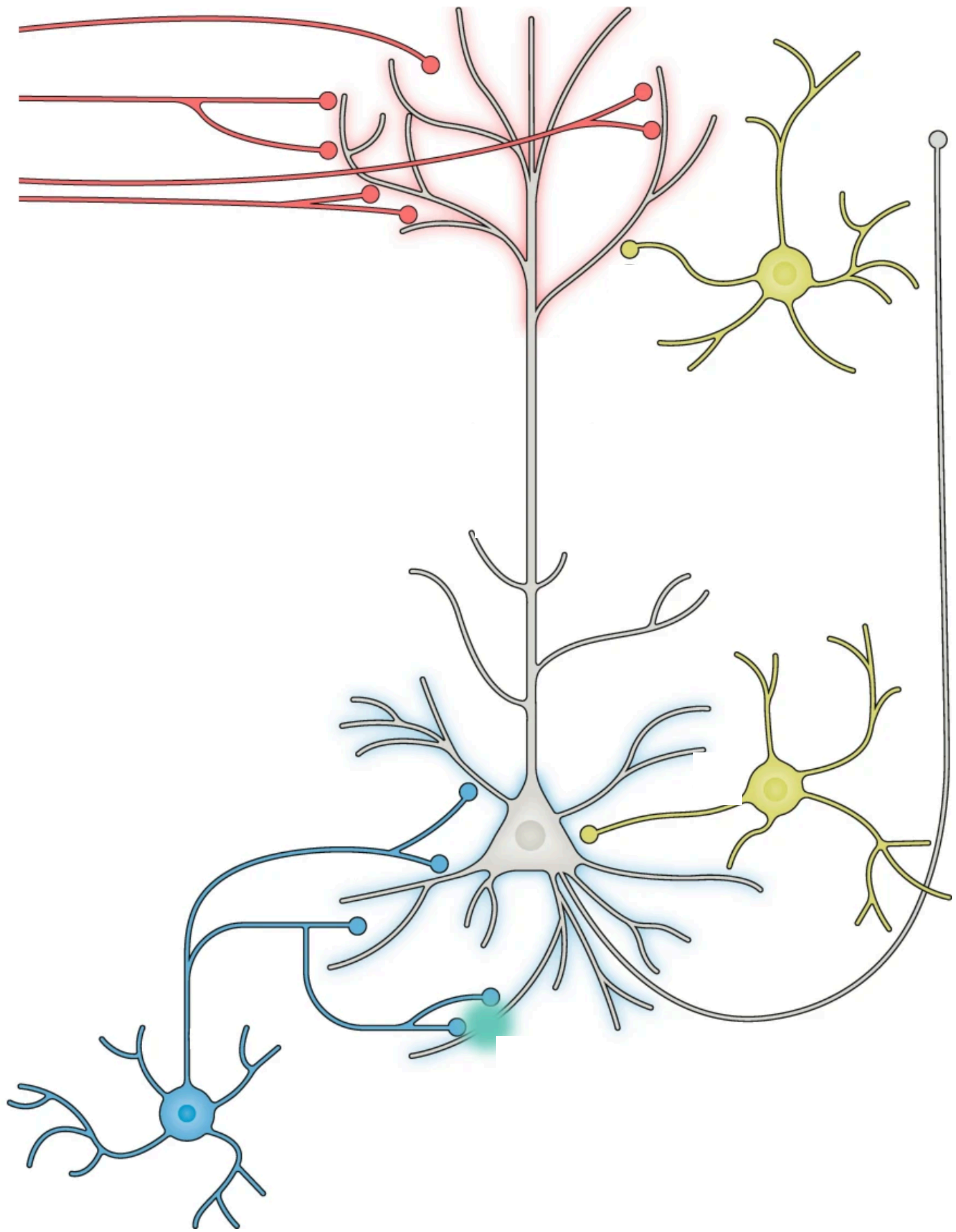
2021.01.25

Aran Nayebi

Neurosciences PhD Candidate

Stanford University

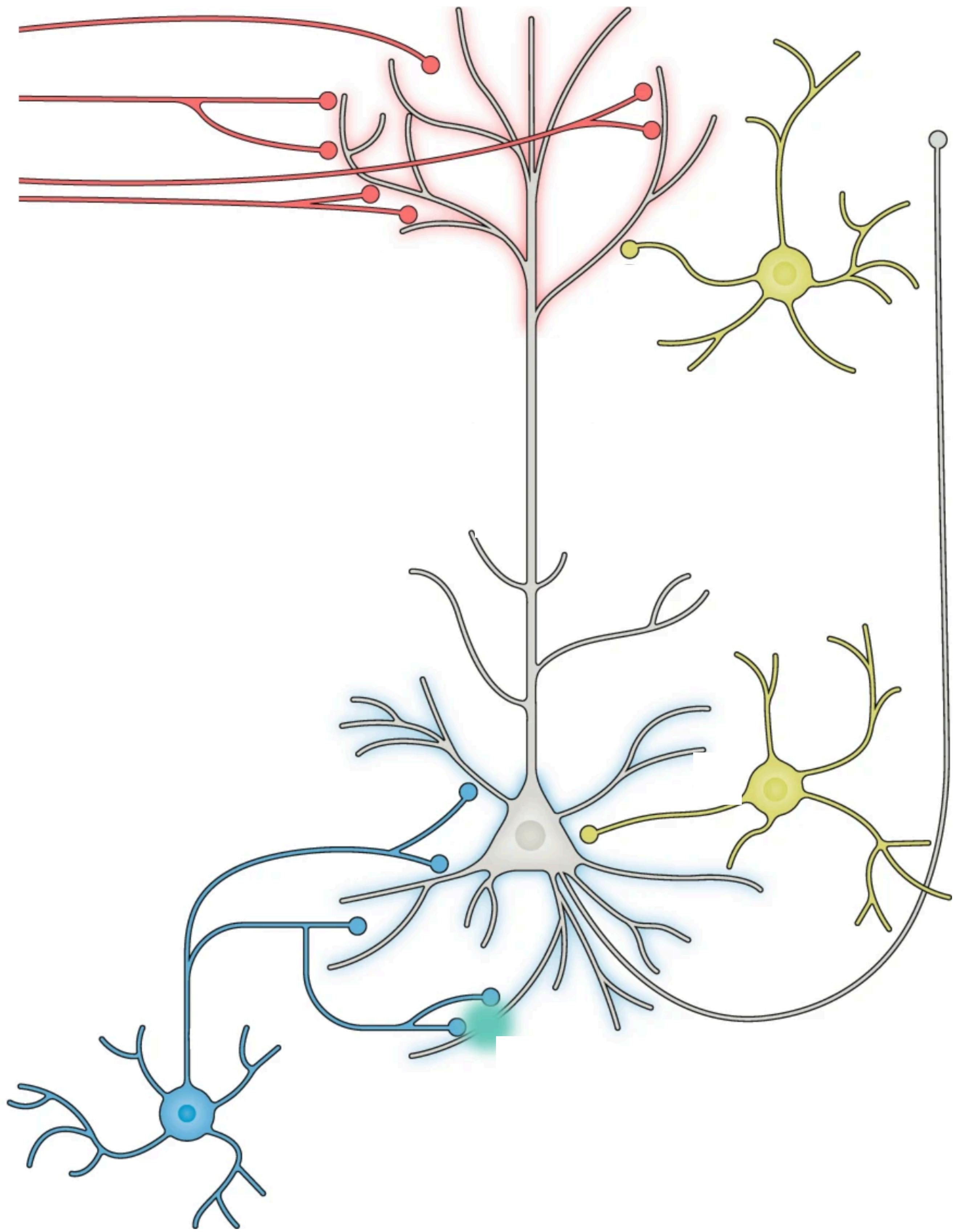
Problem set-up



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

Problem set-up

Post-synaptic activities?

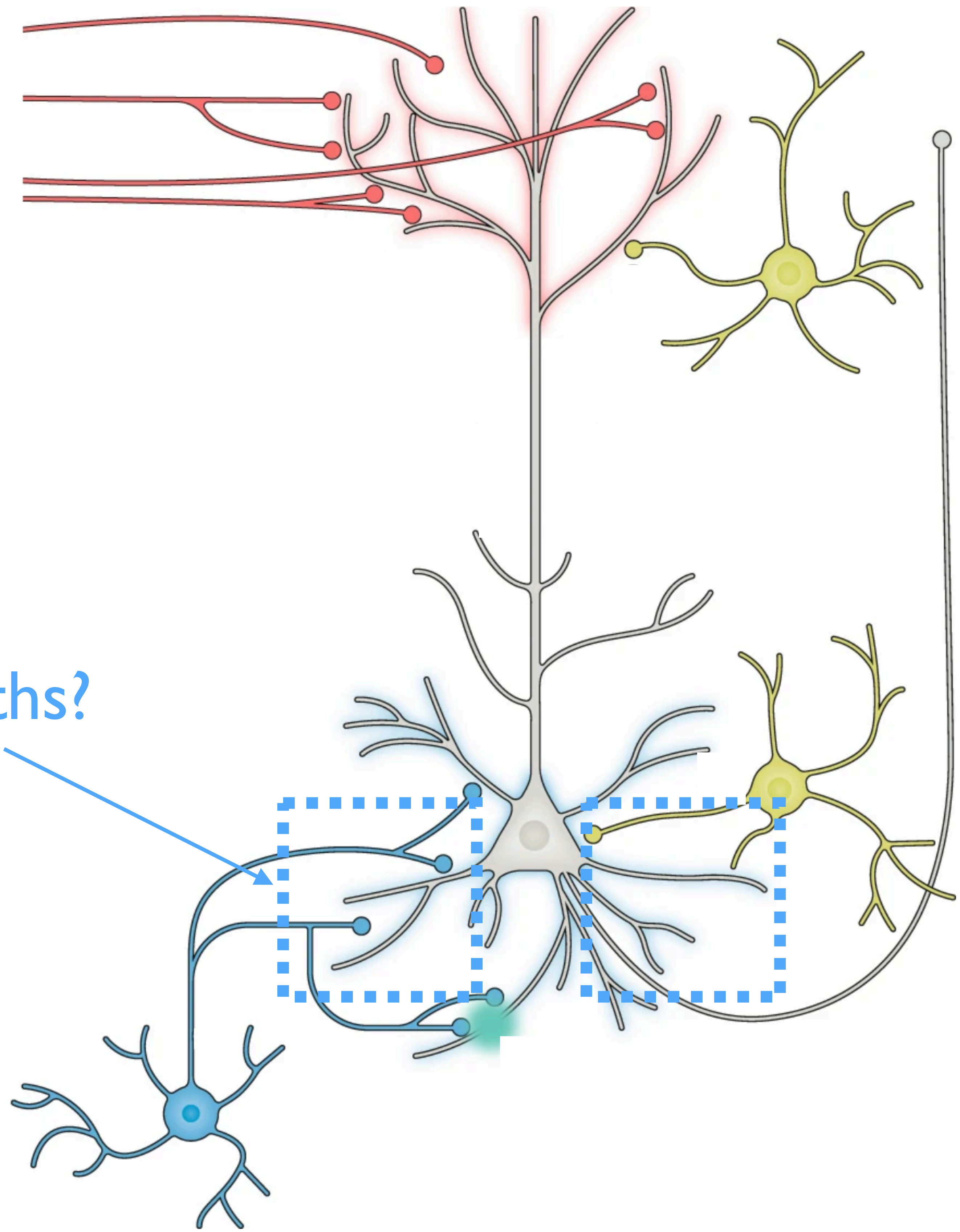


Lillicrap et al.
Nat. Rev. Neurosci. *(2020)*

Problem set-up

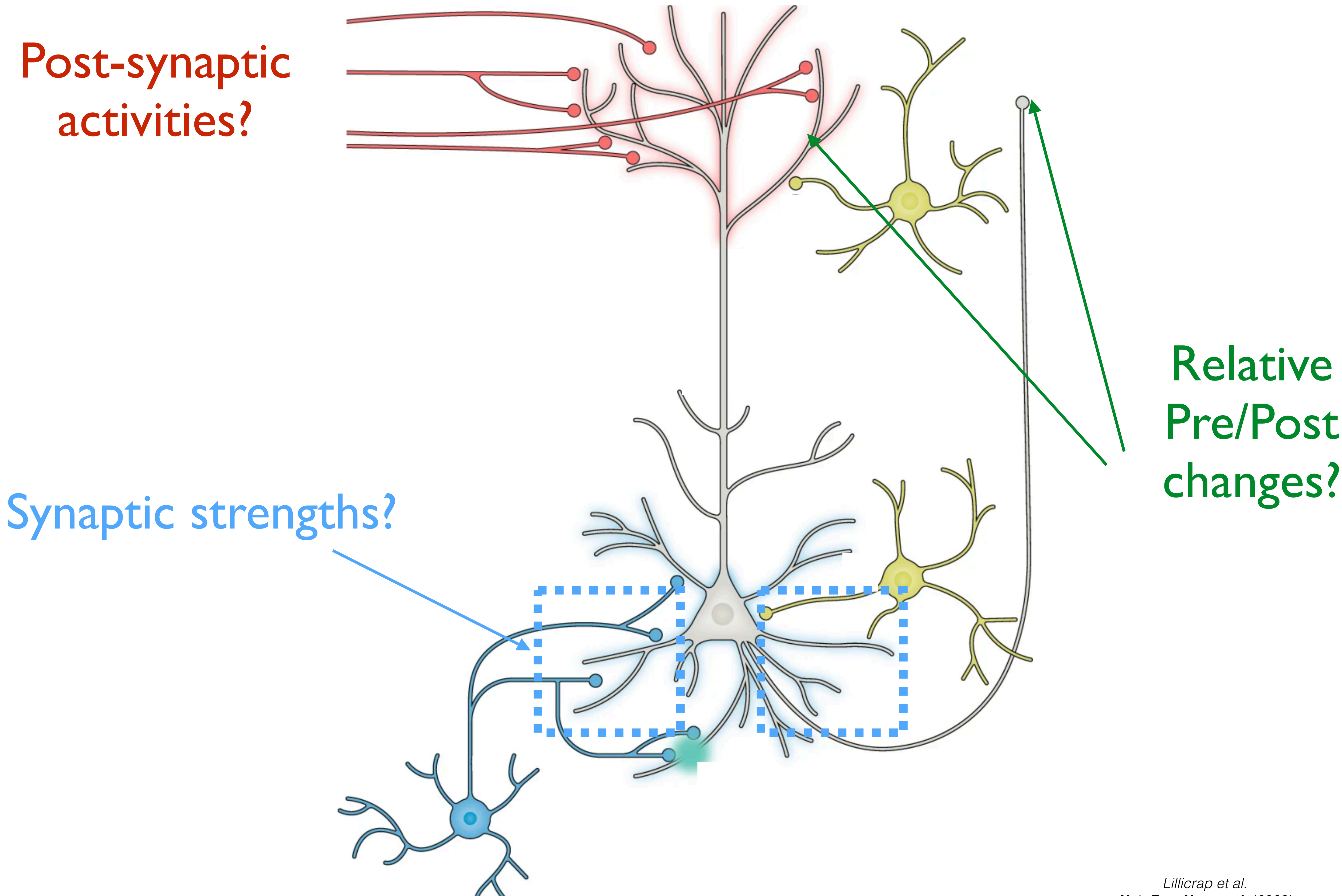
Post-synaptic activities?

Synaptic strengths?



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

Problem set-up



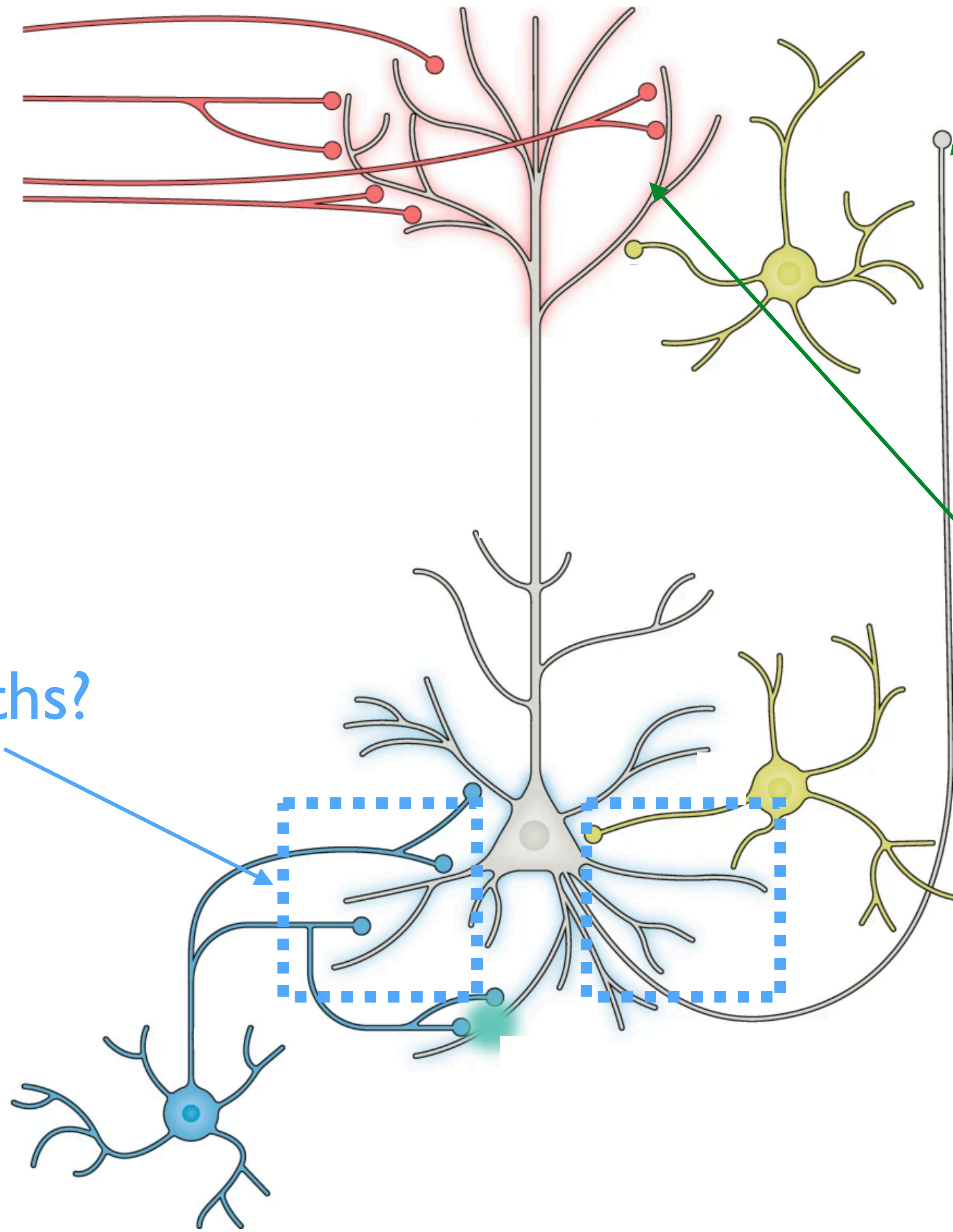
Lillicrap et al.
Nat. Rev. Neurosci. *(2020)*

Problem set-up

Post-synaptic activities?

Cortical area?

Synaptic strengths?



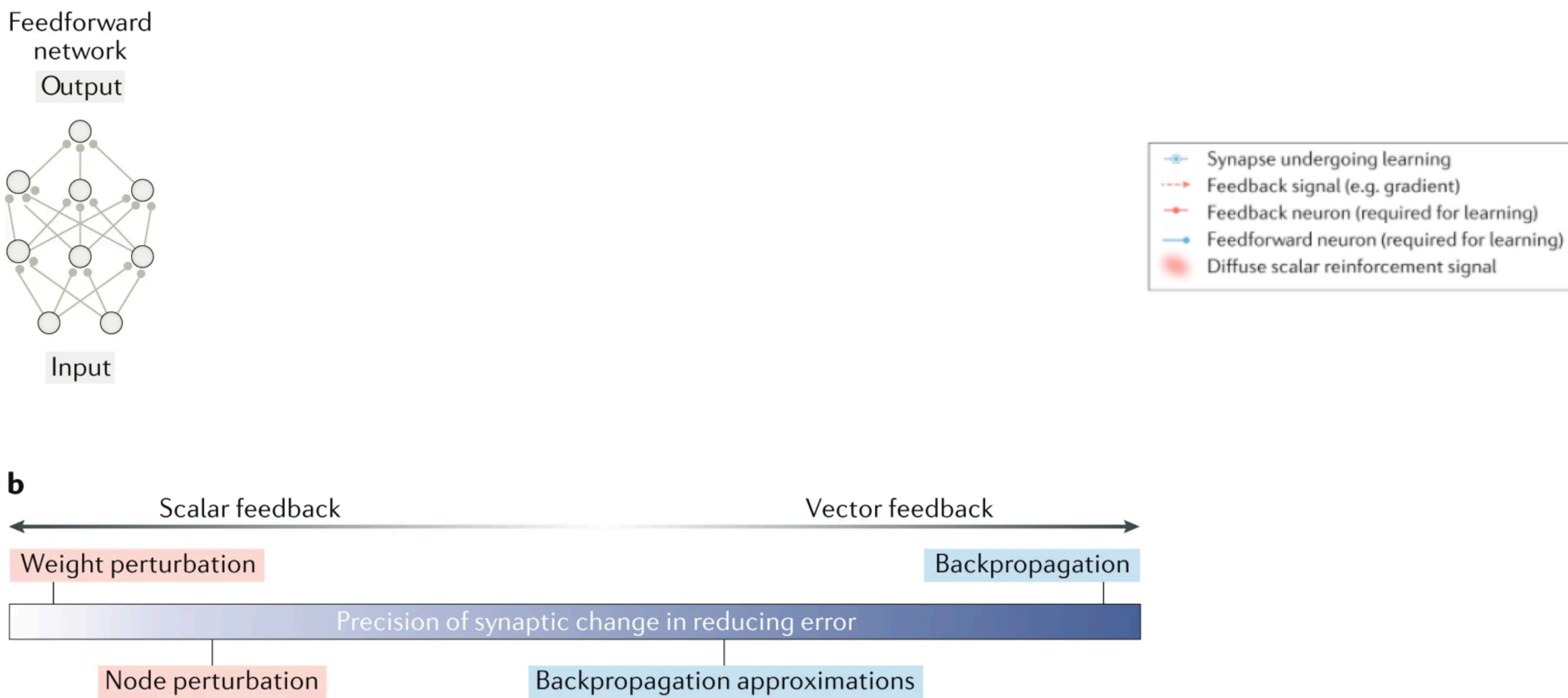
Relative Pre/Post changes?

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

Motivation

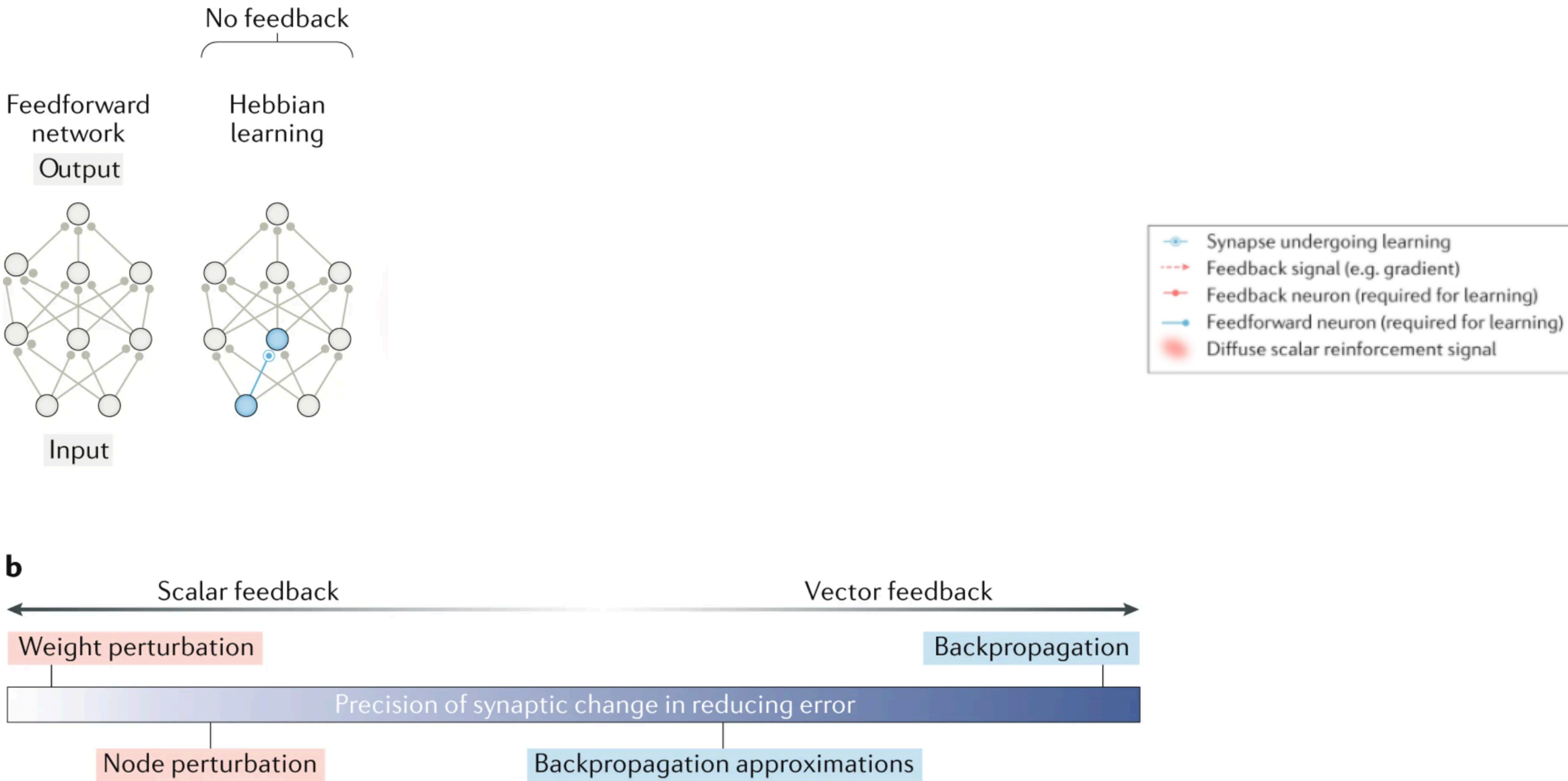
Why might this problem be worth considering?

Spectrum of learning strategies



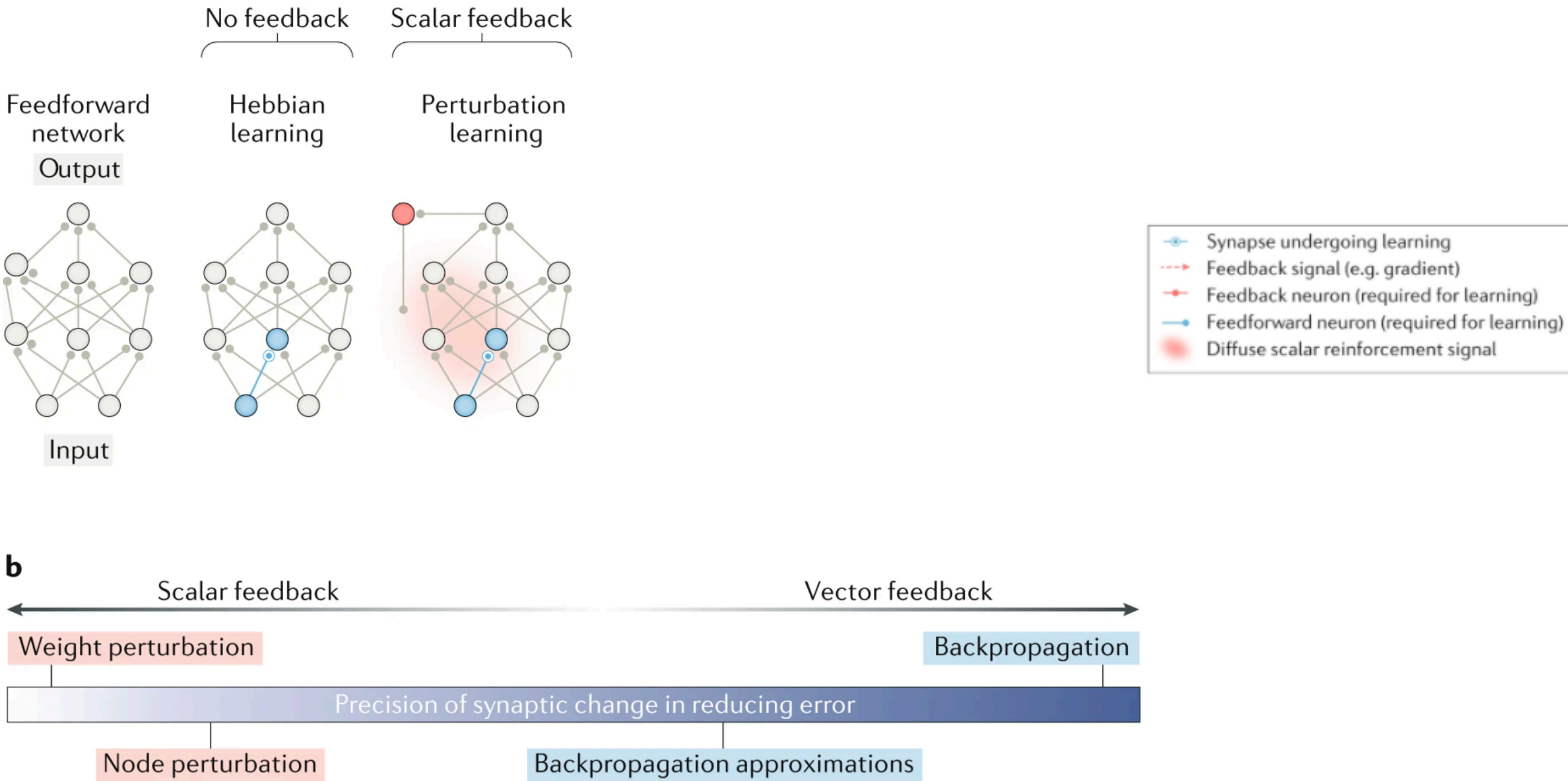
Lillicrap et al.
Nat. Rev. Neurosci. *(2020)*

Spectrum of learning strategies



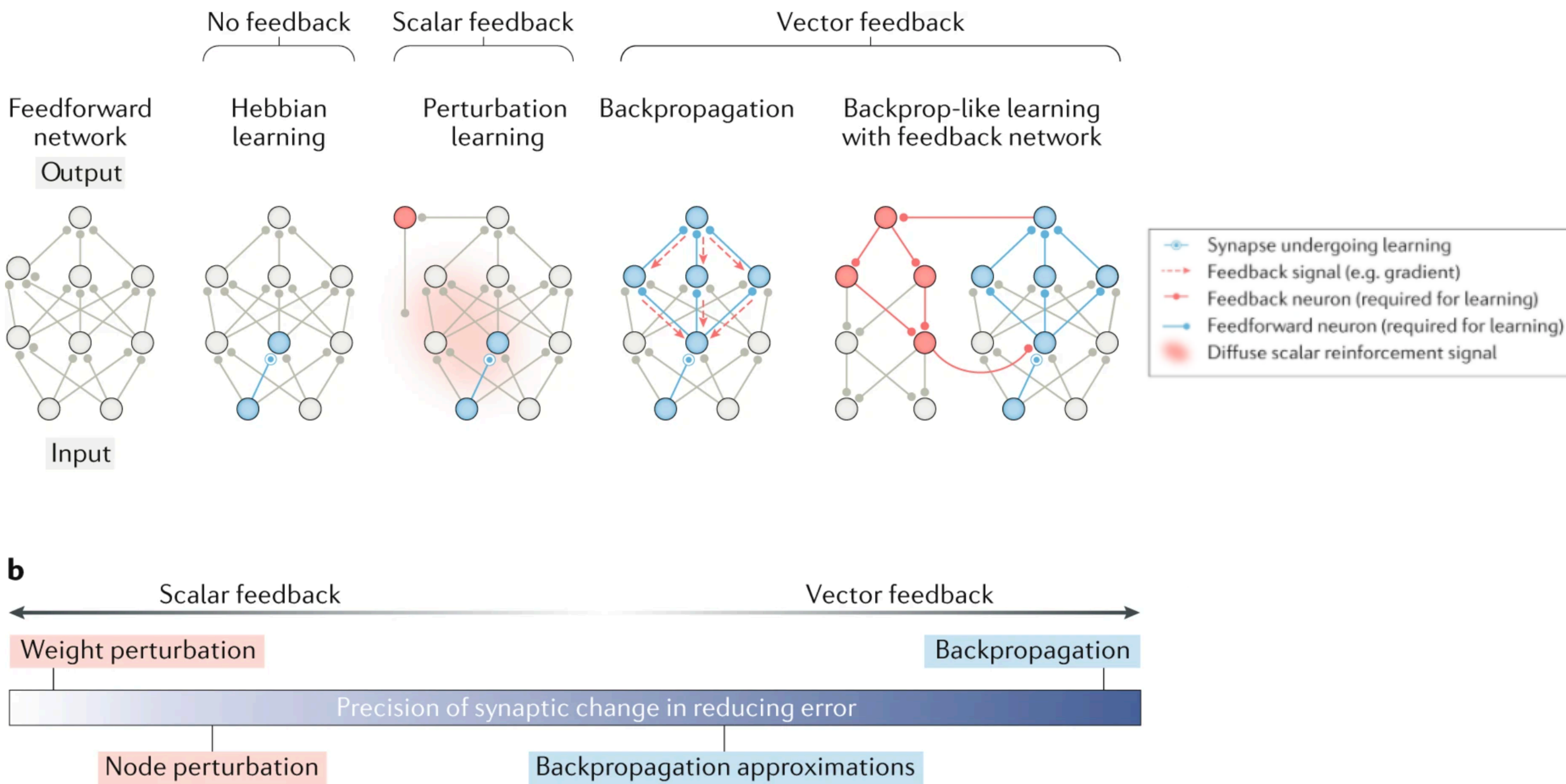
Lillicrap et al.
Nat. Rev. Neurosci. *(2020)*

Spectrum of learning strategies



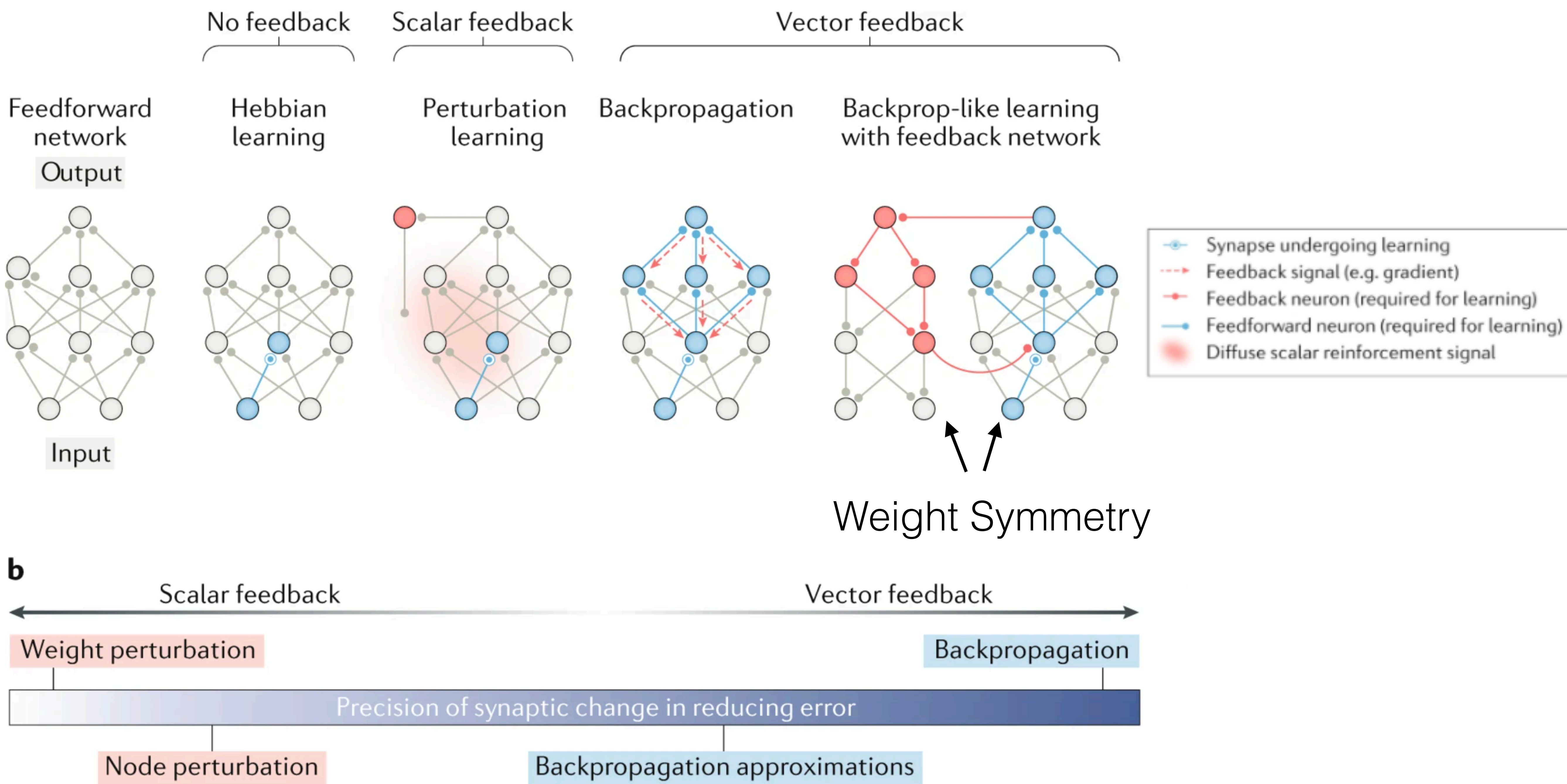
Lillicrap et al.
Nat. Rev. Neurosci. *(2020)*

Spectrum of learning strategies



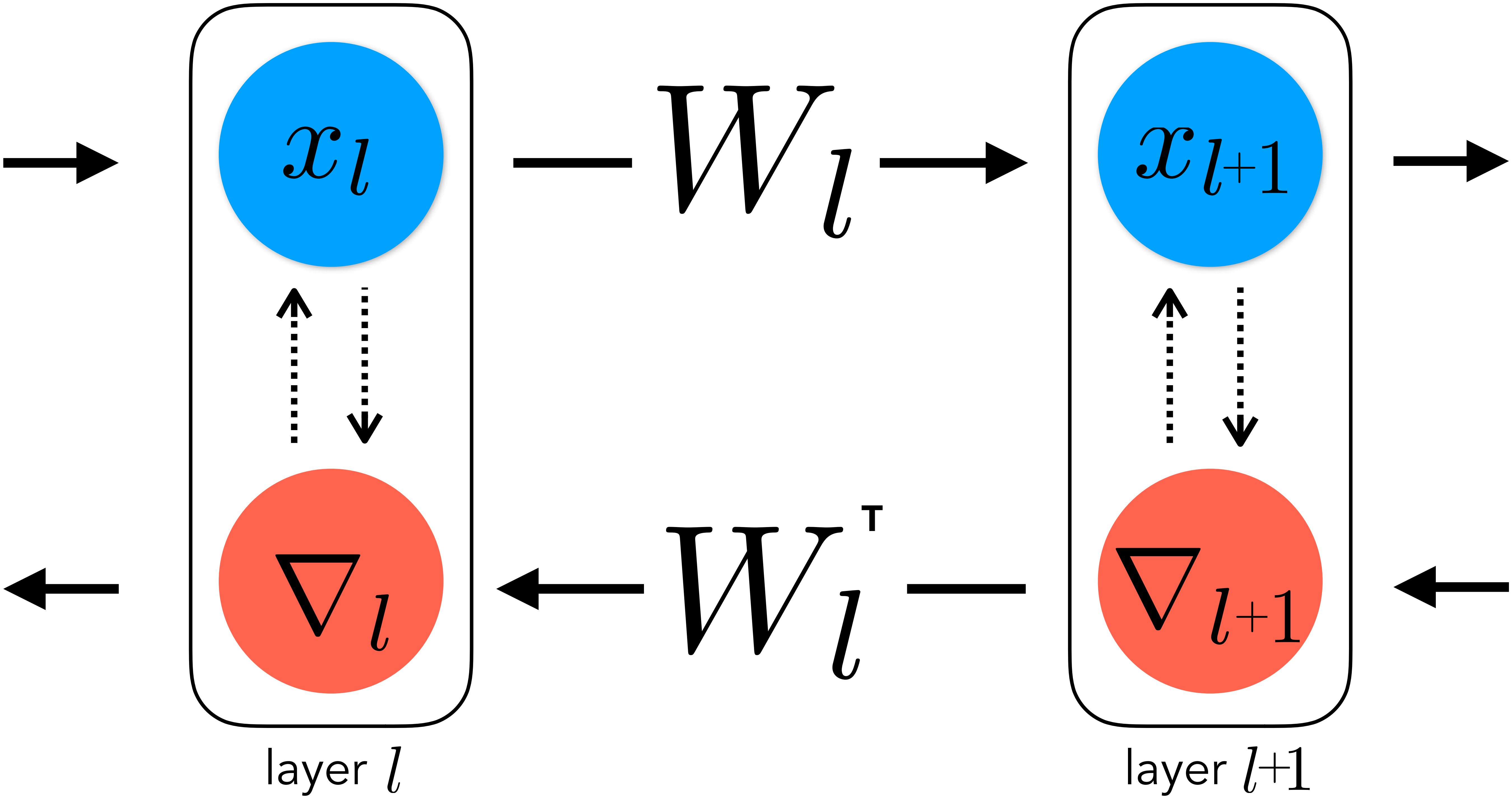
Lillicrap et al.
Nat. Rev. Neurosci. *(2020)*

Spectrum of learning strategies

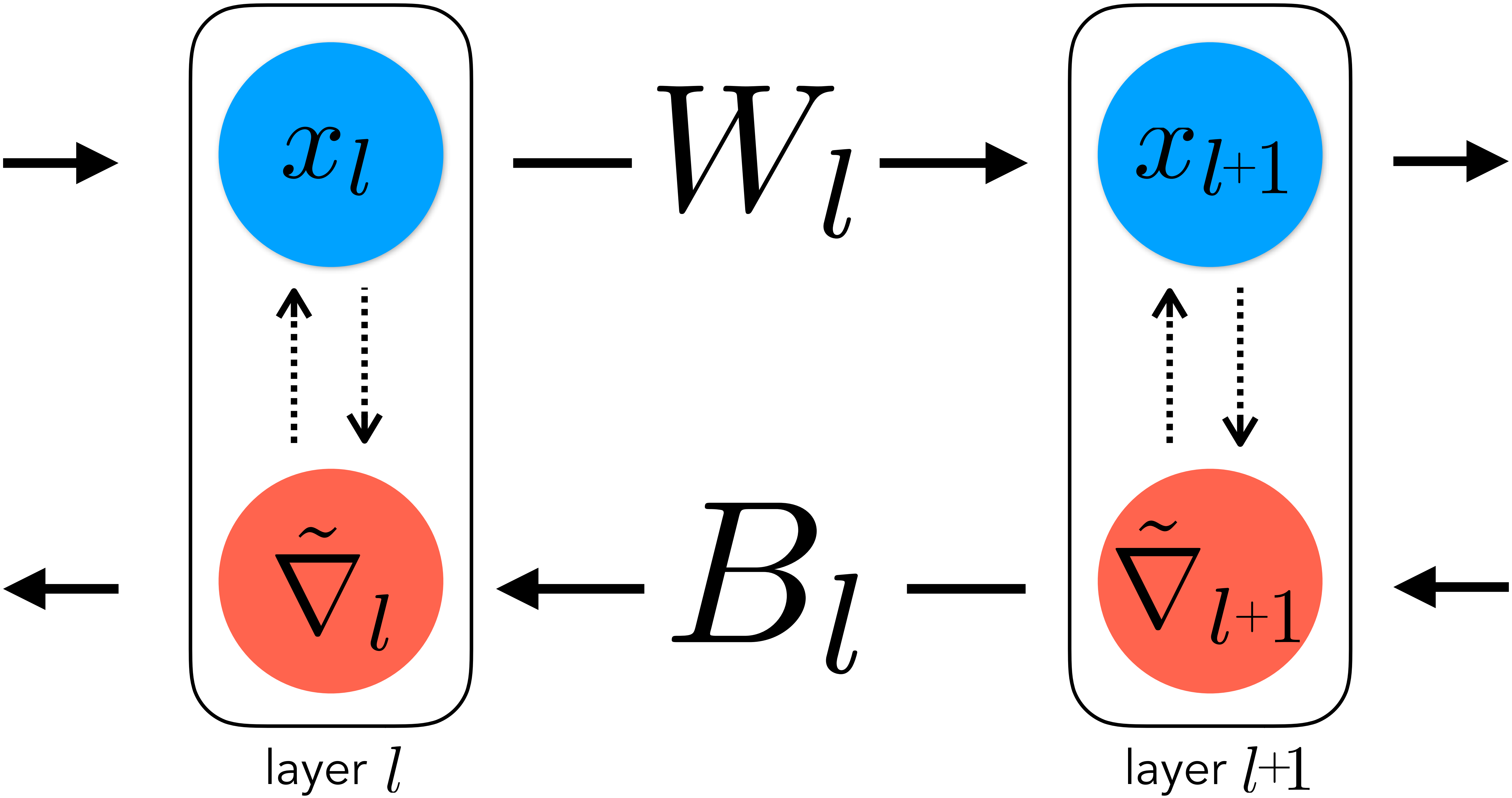


Lillicrap et al.
Nat. Rev. Neurosci. *(2020)*

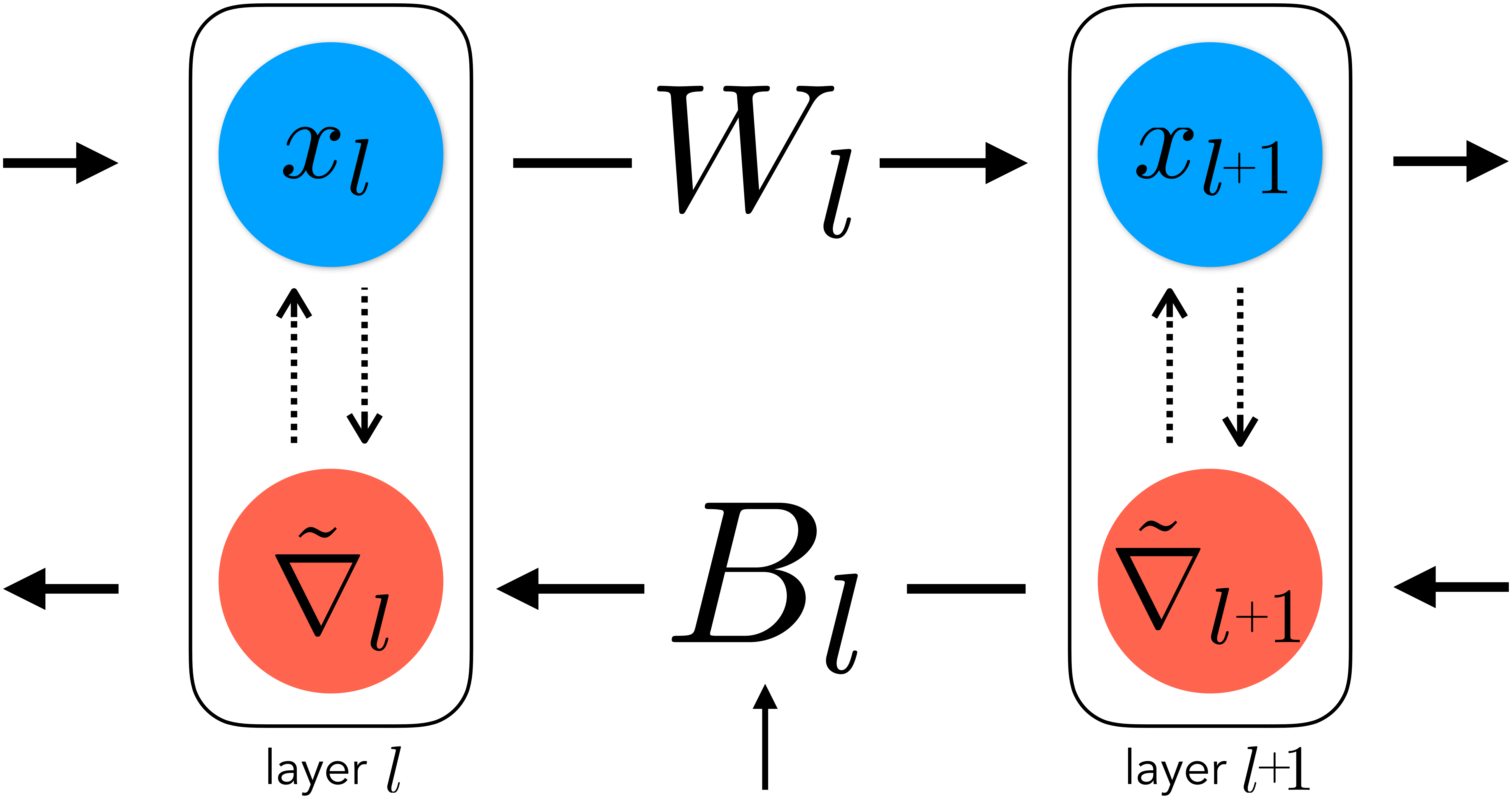
Relaxing the weight symmetry requirement



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

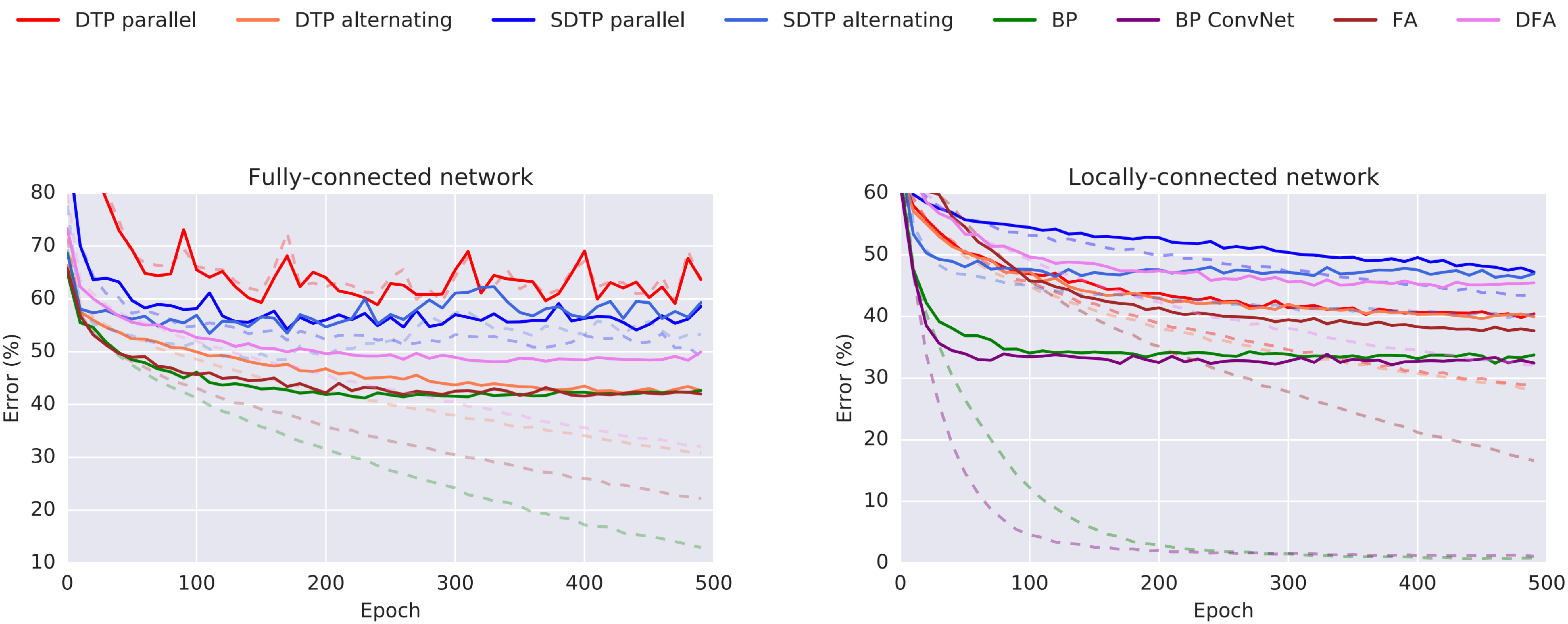


Figure 2: Train (dashed) and test (solid) classification errors on CIFAR.

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

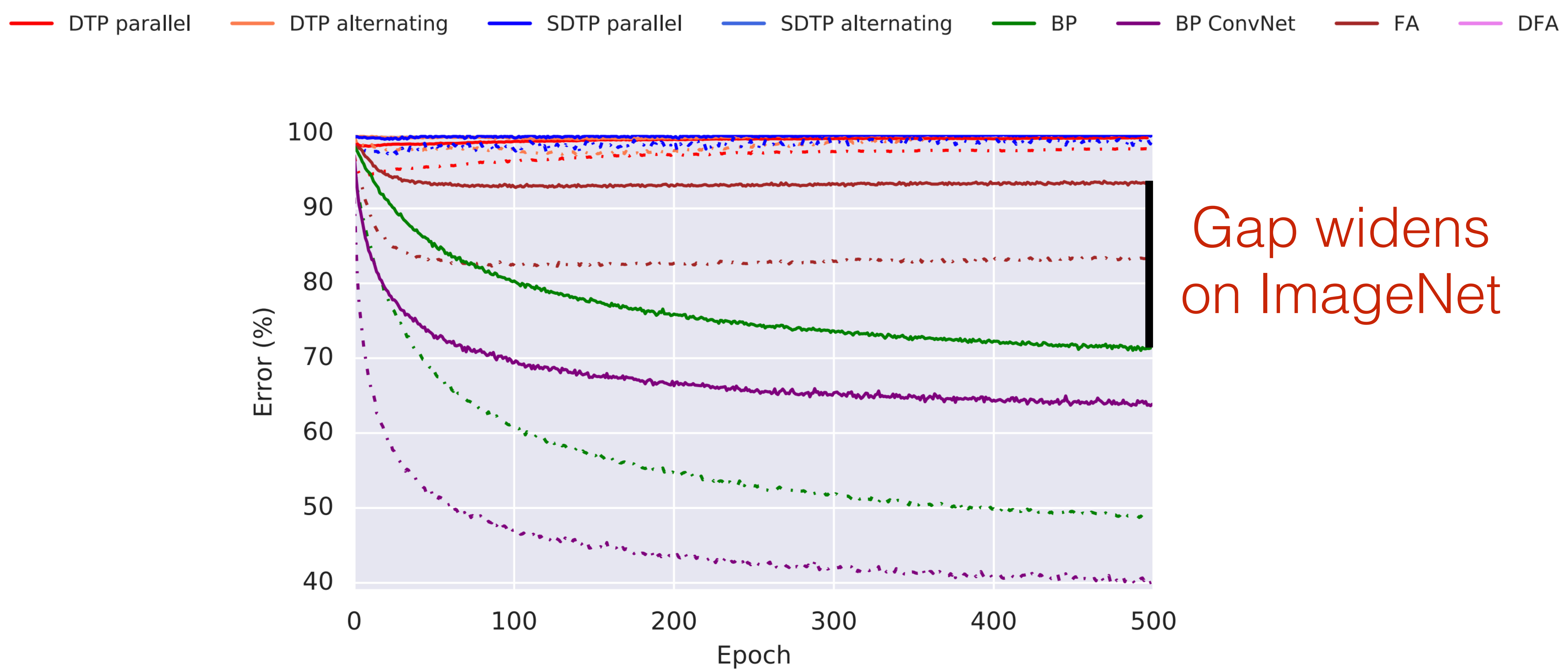
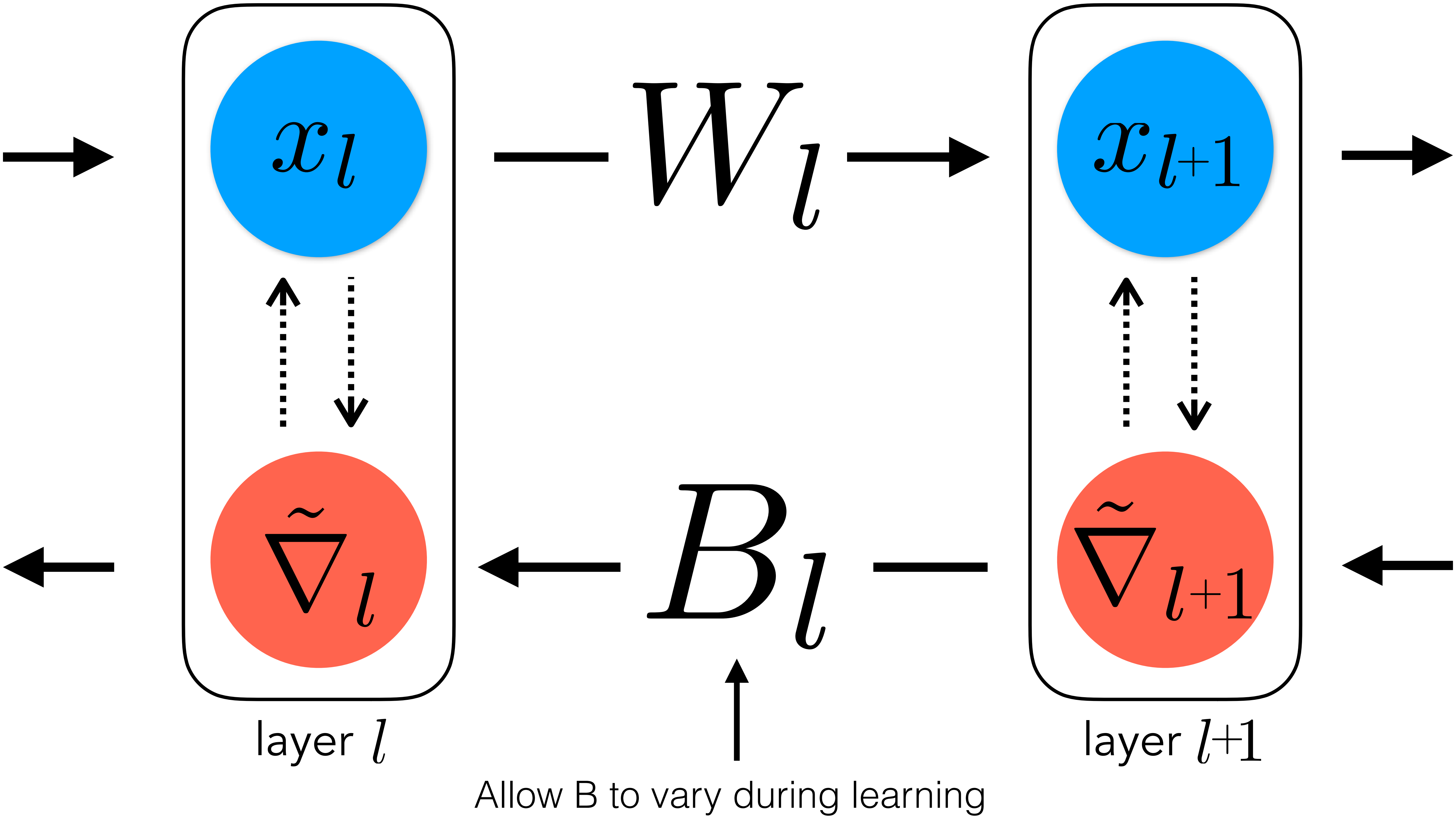


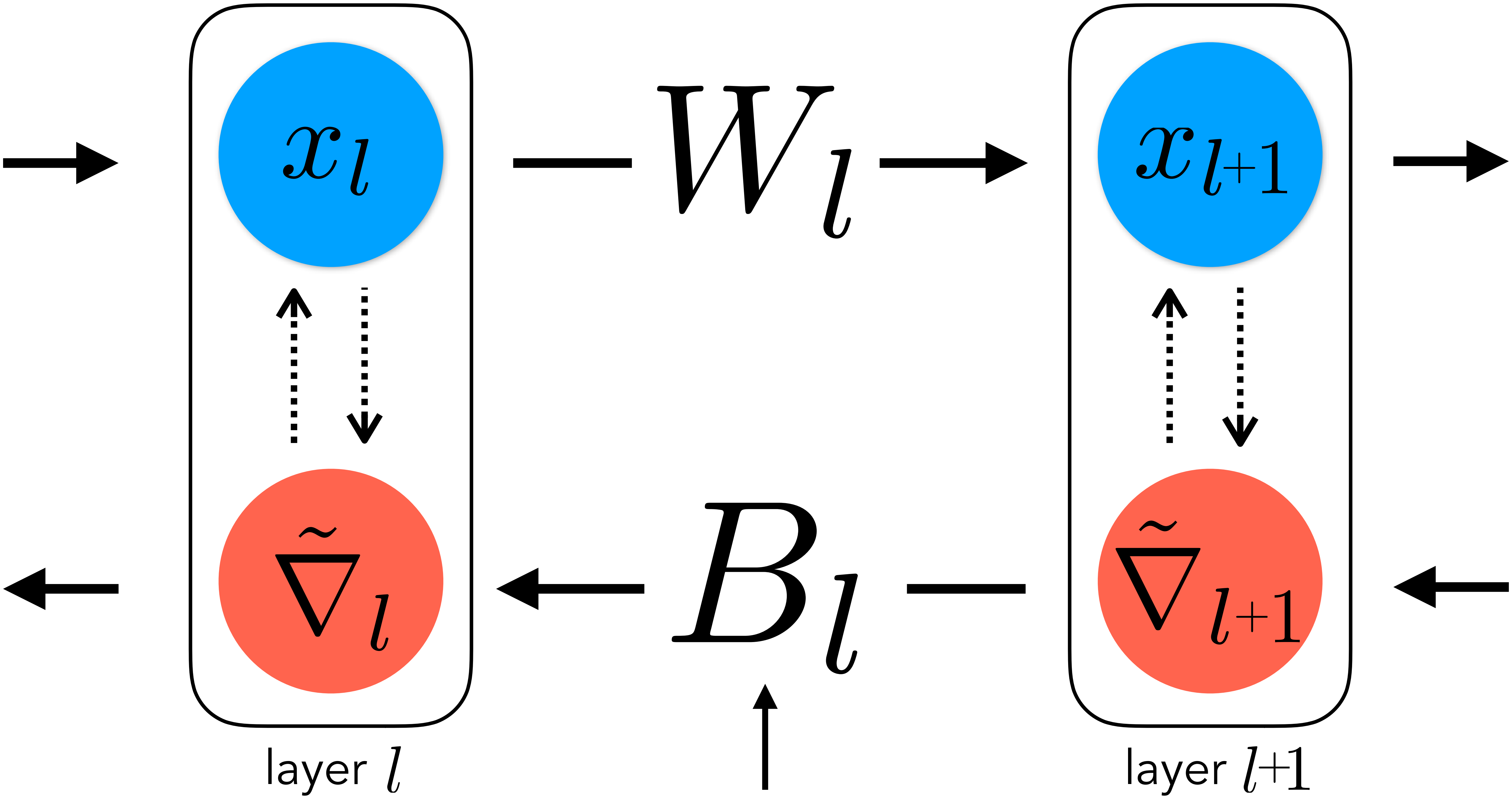
Figure 3: Top-1 (solid) and Top-5 (dotted) test errors on ImageNet. Color legend is the same as for figure 2.

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

Imposing dynamics on the backward weights



Weight Mirror (WM)

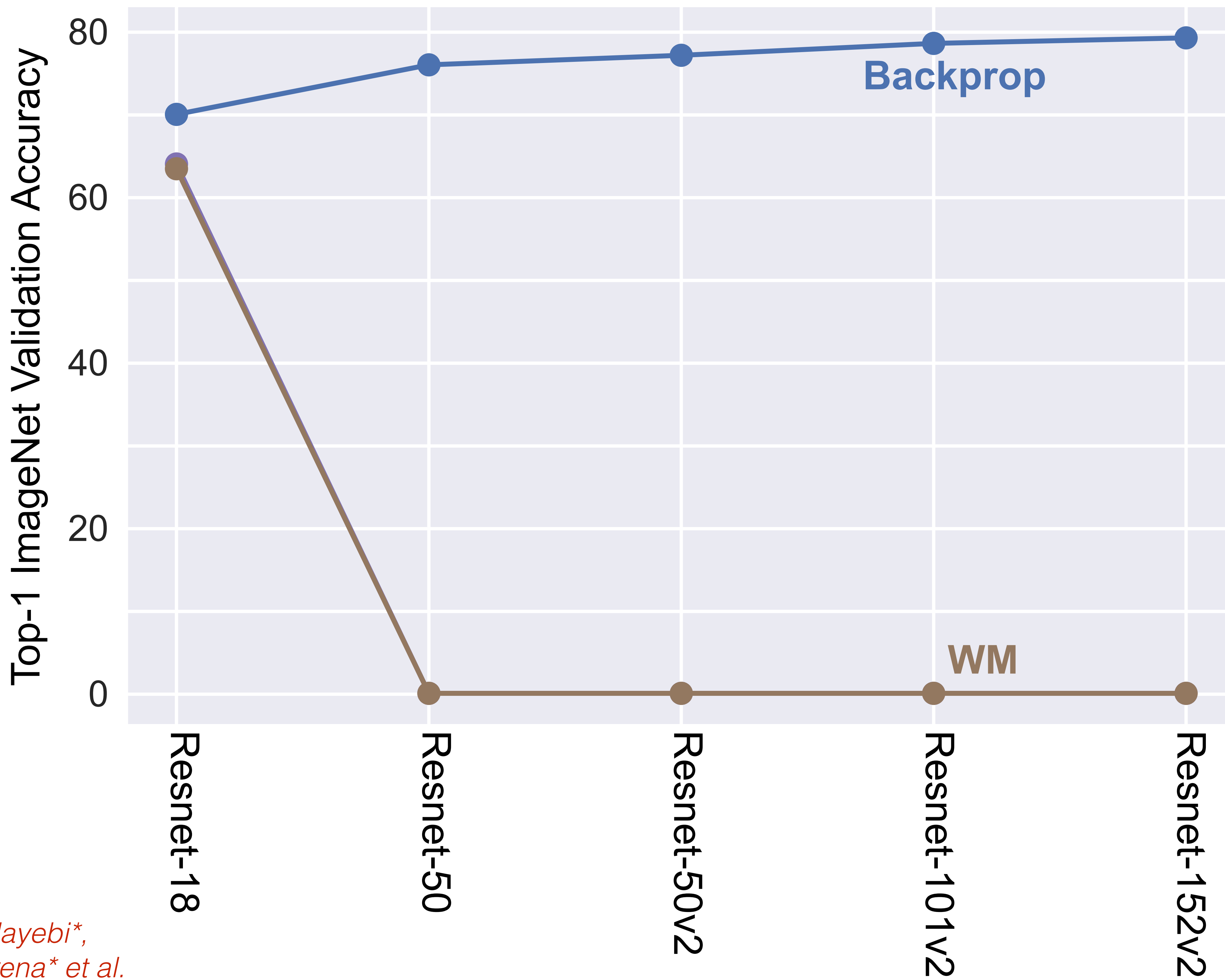


Weight Mirror (WM): B gradually aligns with W

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

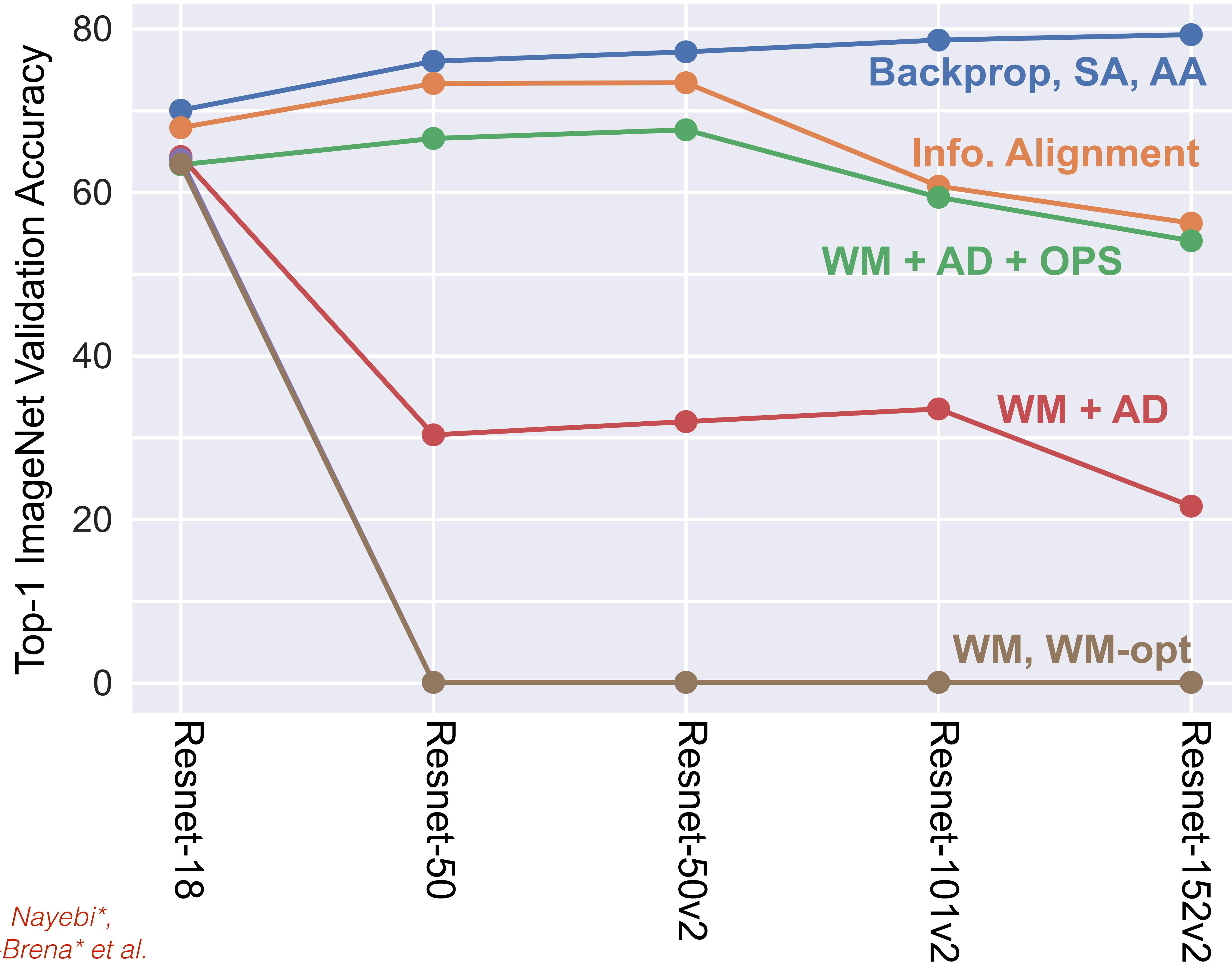
Akrout et al.
NeurIPS *(2019)*

Weight Mirror does not transfer across architectures



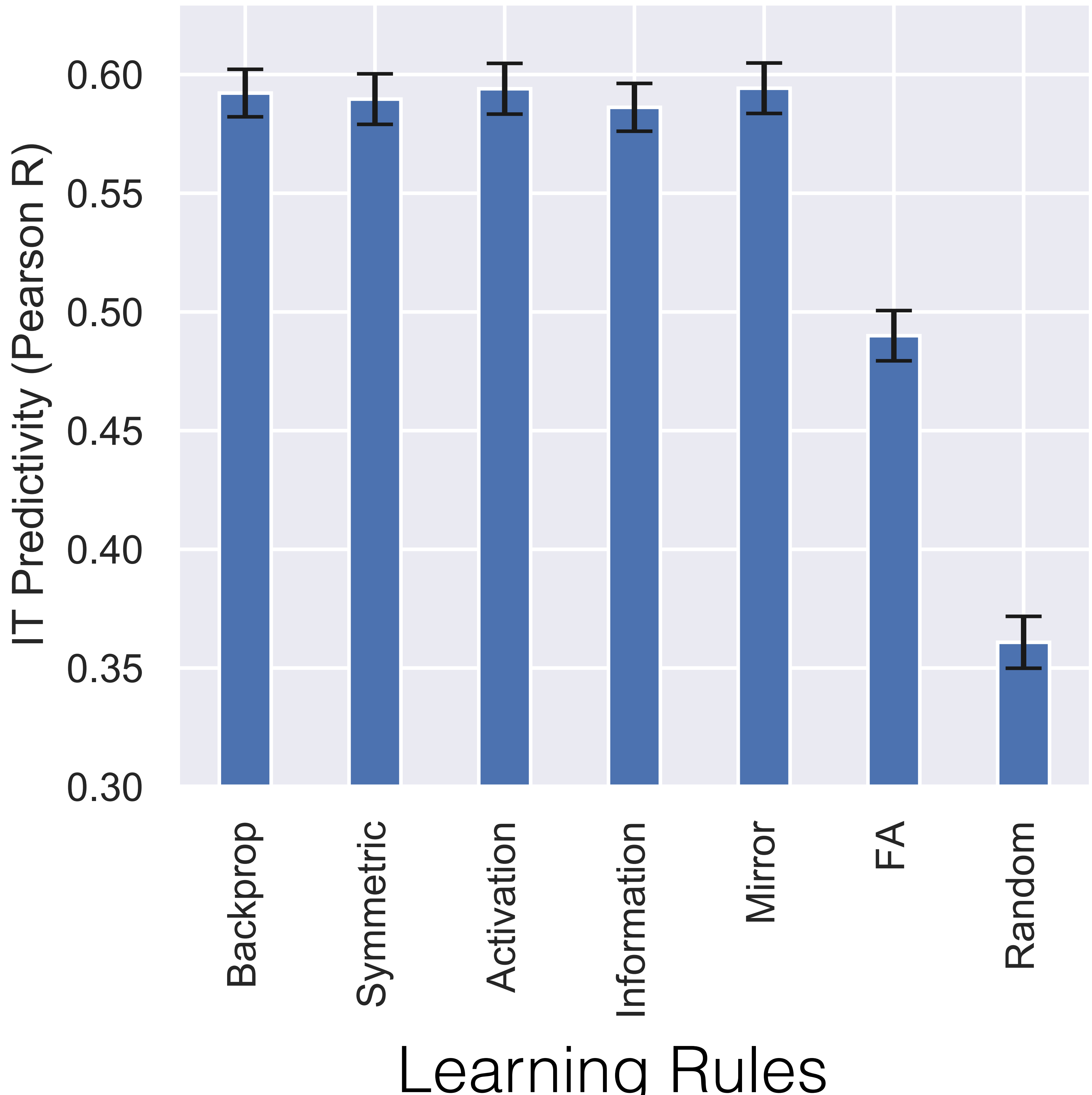
Kunin, Nayebi*, Sagastuy-Brena* et al.*
***ICML** (2020)*

Searching alternatives to Backprop scales across architectures



Kunin, Nayebi*, Sagastuy-Brena* et al.* ***ICML*** *(2020)*

Current neural data is insufficient to separate these alternatives



Kunin, Nayebi*, Sagastuy-Brena* et al.*
***ICML** (2020)*

Can fit single learning rule to post-synaptic activities alone

Can fit single learning rule to post-synaptic activities alone

nature neuroscience

Article | Published: 02 November 2015

Inferring learning rules from distributions of firing rates in cortical neurons

Sukbin Lim, Jillian L McKee, Luke Woloszyn, Yali Amit, David J Freedman, David L Sheinberg & Nicolas Brunel

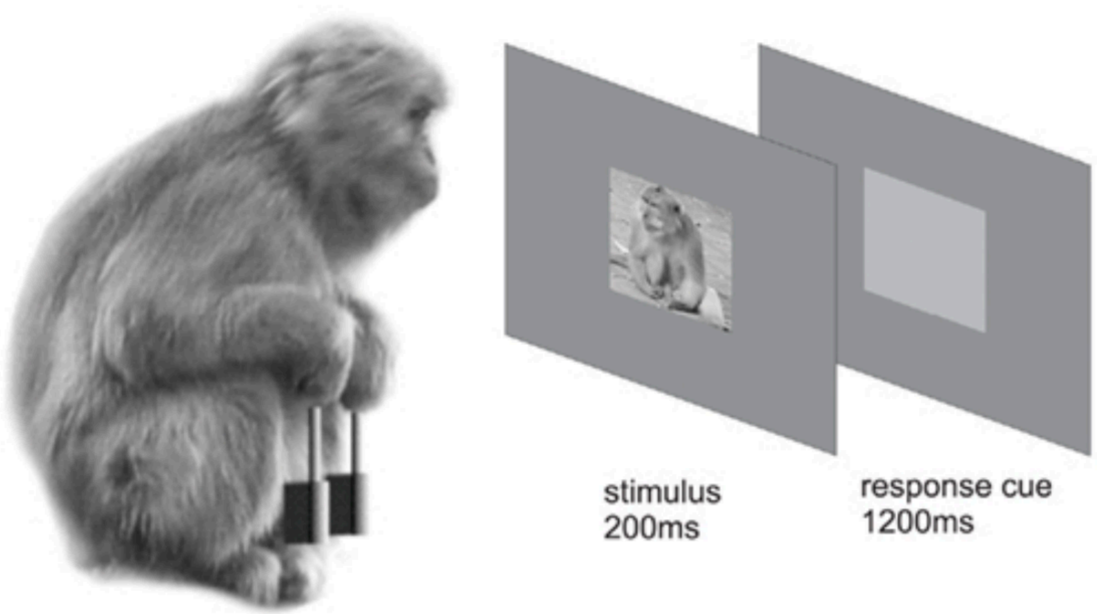
Can fit single learning rule to post-synaptic activities alone

nature neuroscience

Article | Published: 02 November 2015

Inferring learning rules from distributions of firing rates in cortical neurons

Sukbin Lim, Jillian L McKee, Luke Woloszyn, Yali Amit, David J Freedman, David L Sheinberg & Nicolas Brunel



- Each session uses 125 novel & 125 familiar stimuli; macaque IT

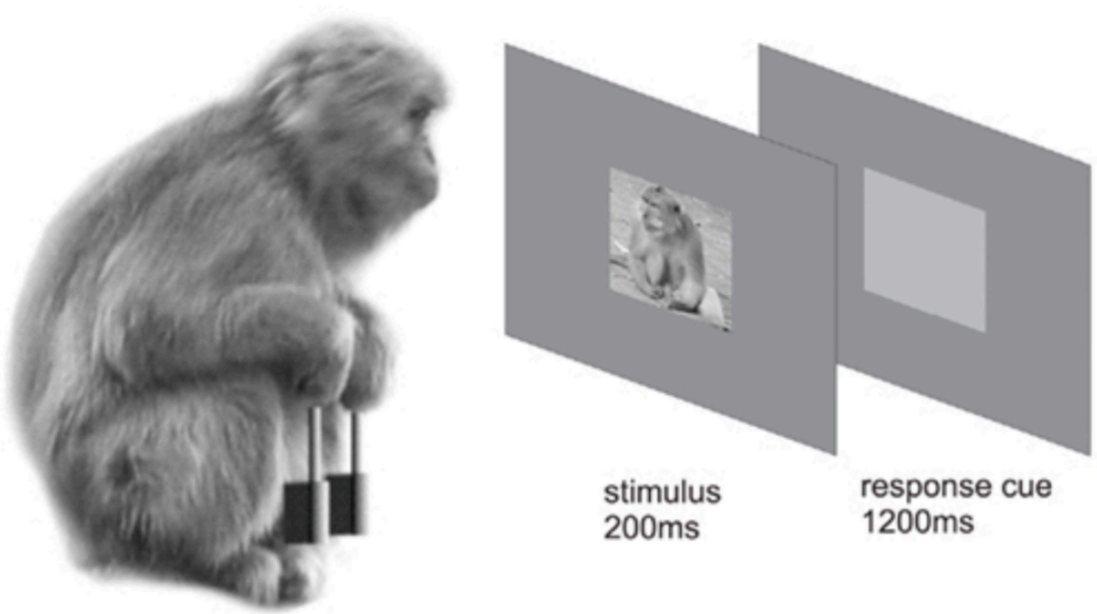
Can fit single learning rule to post-synaptic activities alone

nature neuroscience

Article | Published: 02 November 2015

Inferring learning rules from distributions of firing rates in cortical neurons

Sukbin Lim, Jillian L McKee, Luke Woloszyn, Yali Amit, David J Freedman, David L Sheinberg & Nicolas Brunel



- Each session uses 125 novel & 125 familiar stimuli; macaque IT
- Infer transfer function (nonlinearity) from neuron's inputs to rates, assuming single architecture

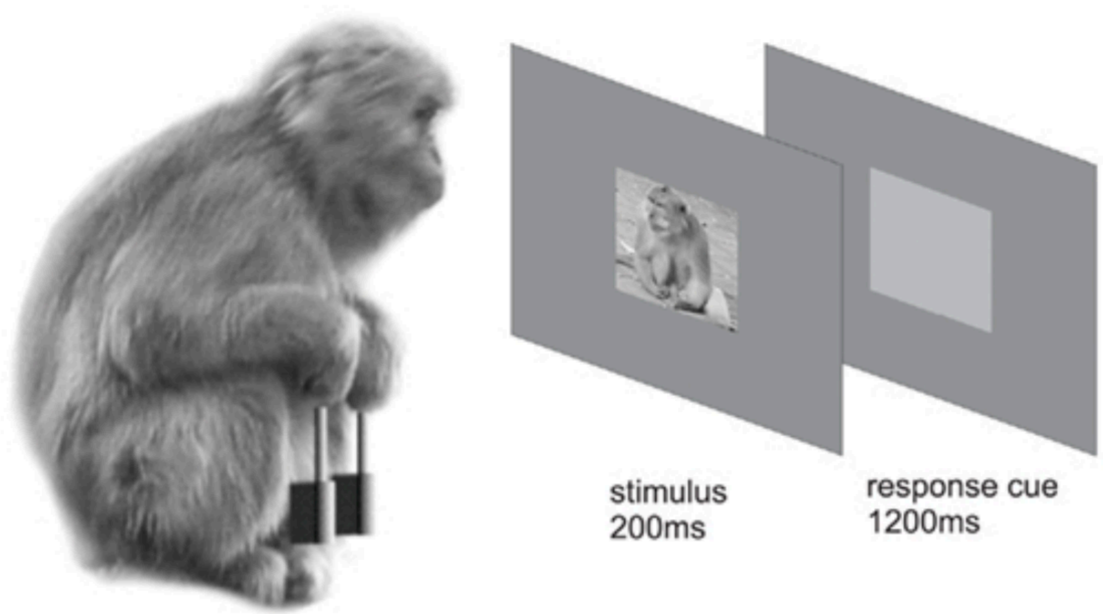
Can fit single learning rule to post-synaptic activities alone

nature
neuroscience

Article | Published: 02 November 2015

Inferring learning rules from distributions of firing rates in cortical neurons

Sukbin Lim, Jillian L McKee, Luke Woloszyn, Yali Amit, David J Freedman, David L Sheinberg & Nicolas Brunel



- Each session uses 125 novel & 125 familiar stimuli; macaque IT
- Infer transfer function (nonlinearity) from neuron's inputs to rates, assuming single architecture
- Fit a Hebbian learning rule by assuming separable pre- and post-synaptic activities (e.g. Hebbian assumption)

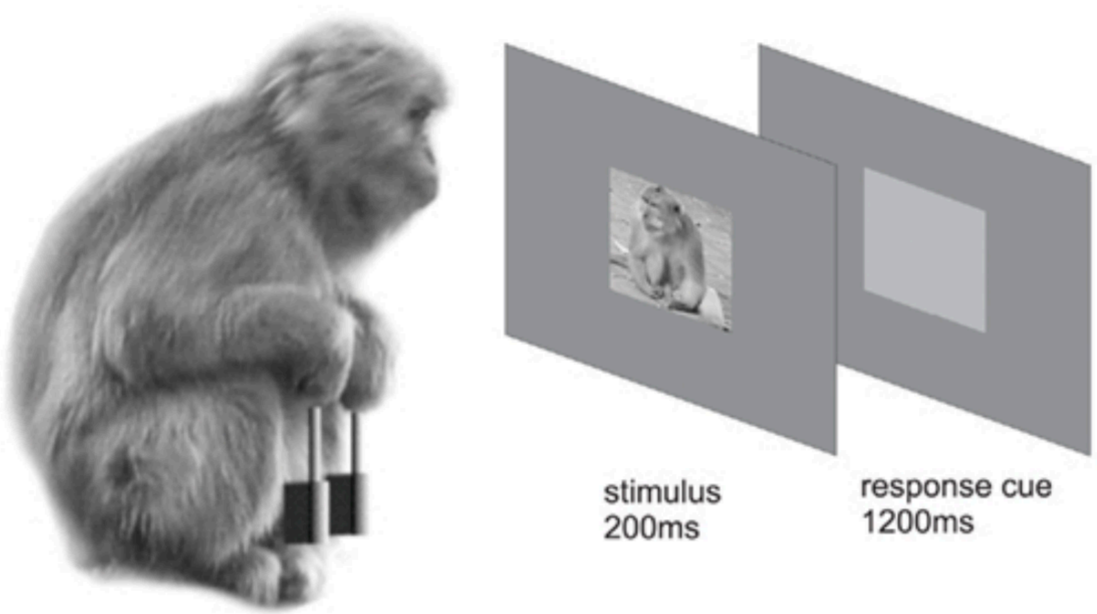
Can fit single learning rule to post-synaptic activities alone

nature neuroscience

Article | Published: 02 November 2015

Inferring learning rules from distributions of firing rates in cortical neurons

Sukbin Lim, Jillian L McKee, Luke Woloszyn, Yali Amit, David J Freedman, David L Sheinberg & Nicolas Brunel



- Each session uses 125 novel & 125 familiar stimuli; macaque IT
- Infer transfer function (nonlinearity) from neuron's inputs to rates, assuming single architecture
- Fit a Hebbian learning rule by assuming separable pre- and post-synaptic activities (e.g. Hebbian assumption)

Fits a *single* learning rule class (Hebbian) to data

"Virtual Experimental" Approach

What would you need to measure
to reliably <u>distinguish</u> *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

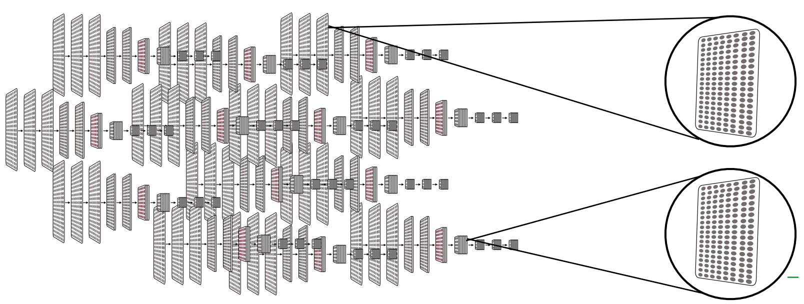
“Virtual Experimental” Approach — Generate Predictions

What would you need to measure
to reliably distinguish *classes* 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

Generating a large-scale dataset

Learning Rules

SGD+Momentum (SGDM)

Adam

Information Alignment (IA)

Feedback Alignment (FA)

Tasks

ImageNet (supervised)

SimCLR (self-supervised)

Word-Speaker-Noise (supervised)

CIFAR-10 (supervised)

Architectures

ResNet-34v2

ResNet-34

ResNet-18v2

ResNet-18

AlexNet

AlexNet-LRN

KNet4

KNet4-LRN

KNet5

KNet5-LRN

Hyperparameters

Batch size (128, 256, 512)

Model seed (None, 0)

Dataset seed (None, 0)

Generating a large-scale dataset

Learning Rules

SGD+Momentum (SGDM)

Adam

Information Alignment (IA)

Feedback Alignment (FA)

Tasks

ImageNet (supervised)

SimCLR (self-supervised)

Word-Speaker-Noise (supervised)

CIFAR-10 (supervised)

Architectures

ResNet-34v2

ResNet-34

ResNet-18v2

ResNet-18

AlexNet

AlexNet-LRN

KNet4

KNet4-LRN

KNet5

KNet5-LRN

Hyperparameters

Batch size (128, 256, 512)

Model seed (None, 0)

Dataset seed (None, 0)

Generating a large-scale dataset

Learning Rules

SGD+Momentum (SGDM)

Adam

Information Alignment (IA)

Feedback Alignment (FA)

Tasks

ImageNet (supervised)

SimCLR (self-supervised)

Word-Speaker-Noise (supervised)

CIFAR-10 (supervised)

Architectures

ResNet-34v2

ResNet-34

ResNet-18v2

ResNet-18

AlexNet

AlexNet-LRN

KNet4

KNet4-LRN

KNet5

KNet5-LRN

Hyperparameters

Batch size (128, 256, 512)

Model seed (None, 0)

Dataset seed (None, 0)

Generating a large-scale dataset

Learning Rules	Tasks	Architectures	Hyperparameters
SGD+Momentum (SGDM)	ImageNet (supervised)	ResNet-34v2	Batch size (128, 256, 512)
Adam	SimCLR (self-supervised)	ResNet-34	Model seed (None, 0)
Information Alignment (IA)	Word-Speaker-Noise (supervised)	ResNet-18v2	Dataset seed (None, 0)
Feedback Alignment (FA)	CIFAR-10 (supervised)	ResNet-18	
		AlexNet	
		AlexNet-LRN	
		KNet4	
		KNet4-LRN	
		KNet5	
		KNet5-LRN	

Generating a large-scale dataset

Learning Rules

SGD+Momentum (SGDM)

Adam

Information Alignment (IA)

Feedback Alignment (FA)

Tasks

ImageNet (supervised)

SimCLR (self-supervised)

Word-Speaker-Noise (supervised)

CIFAR-10 (supervised)

Architectures

ResNet-34v2

ResNet-34

ResNet-18v2

ResNet-18

AlexNet

AlexNet-LRN

KNet4

KNet4-LRN

KNet5

KNet5-LRN

Hyperparameters

Batch size (128, 256, 512)

Model seed (None, 0)

Dataset seed (None, 0)

Generating a large-scale dataset

Learning Rules

SGD+Momentum (SGDM)

Adam

Information Alignment (IA)

Feedback Alignment (FA)

Tasks

ImageNet (supervised)

SimCLR (self-supervised)

Word-Speaker-Noise (supervised)

CIFAR-10 (supervised)

Architectures

ResNet-34v2

ResNet-34

ResNet-18v2

ResNet-18

AlexNet

AlexNet-LRN

KNet4

KNet4-LRN

KNet5

KNet5-LRN

Hyperparameters

Batch size (128, 256, 512)

Model seed (None, 0)

Dataset seed (None, 0)

Generating a large-scale dataset

Learning Rules

SGD+Momentum (SGDM)

Adam

Information Alignment (IA)

Feedback Alignment (FA)

Tasks

ImageNet (supervised)

SimCLR (self-supervised)

Word-Speaker-Noise (supervised)

CIFAR-10 (supervised)

Architectures

ResNet-34v2

ResNet-34

ResNet-18v2

ResNet-18

AlexNet

AlexNet-LRN

KNet4

KNet4-LRN

KNet5

KNet5-LRN

Hyperparameters

Batch size (128, 256, 512)

Model seed (None, 0)

Dataset seed (None, 0)

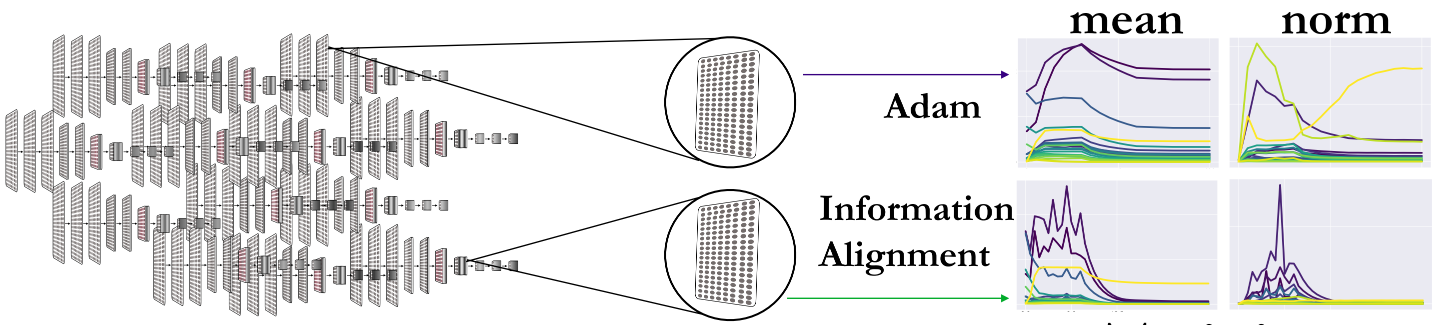
Generating a large-scale dataset

Learning Rules	Tasks	Architectures	Hyperparameters
SGD+Momentum (SGDM)	ImageNet (supervised)	ResNet-34v2	Batch size (128, 256, 512)
Adam	SimCLR (self-supervised)	ResNet-34	Model seed (None, 0)
Information Alignment (IA)	Word-Speaker-Noise (supervised)	ResNet-18v2	Dataset seed (None, 0)
Feedback Alignment (FA)	CIFAR-10 (supervised)	ResNet-18	
		AlexNet	
		AlexNet-LRN	
		KNet4	
		KNet4-LRN	
		KNet5	
		KNet5-LRN	

"Virtual Experimental" Approach

Data Generation

Observable Statistics



Training Epochs

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

Defining observable statistics

Weights

Proxy for synaptic strengths

Activations

Proxy for post-synaptic activities

Layer-wise Activity Changes

Proxy for relative change between pre- and post-synaptic activations

Defining observable statistics

Weights | Activations | Layer-wise Activity Changes

×

Identity | Square | Abs val

Defining observable statistics

Weights | Activations | Layer-wise Activity Changes

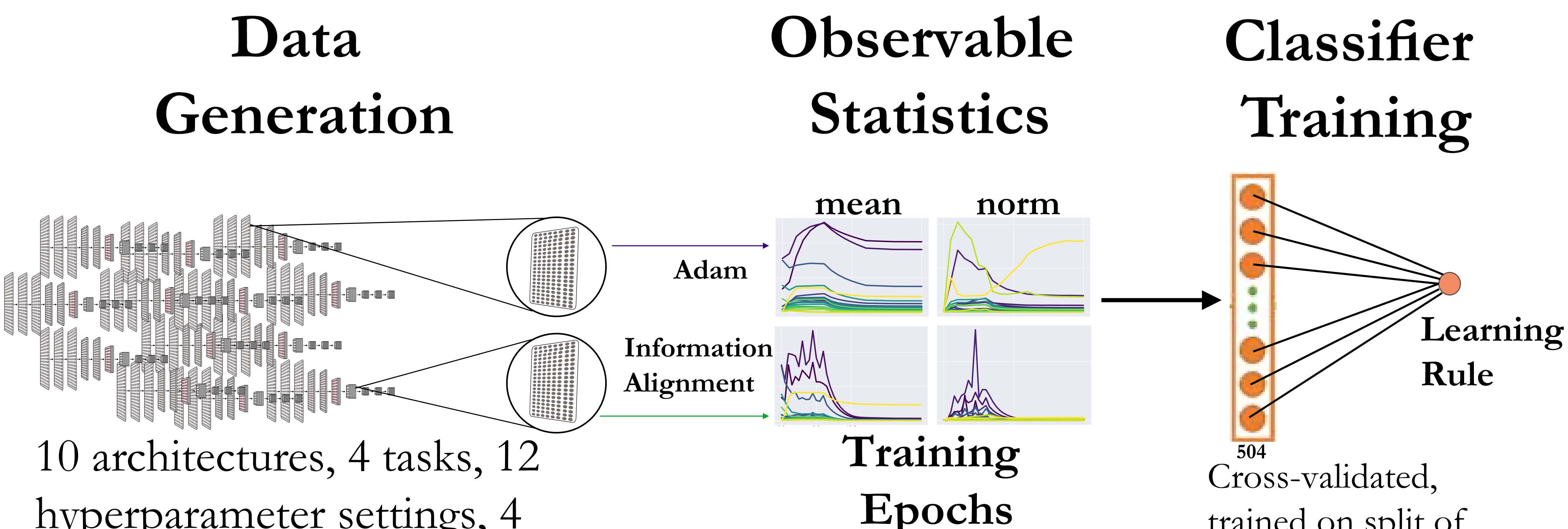
×

Identity | Square | Abs val

×

Norm | Mean | Median | 3rd quartile | Variance | Skew | Kurtosis

“Virtual Experimental” Approach

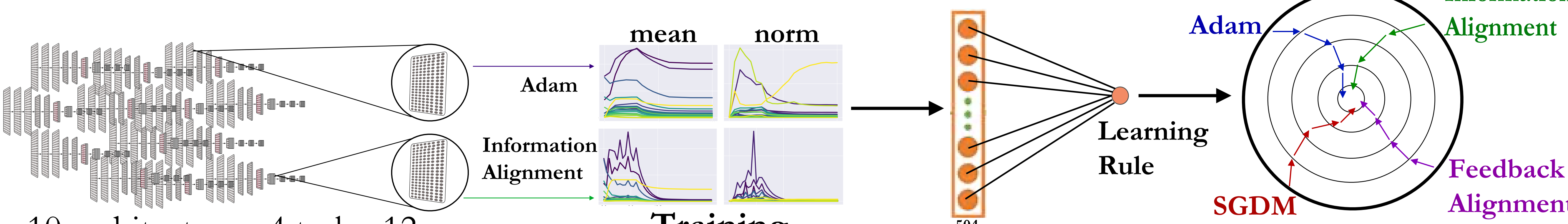


10 architectures, 4 tasks, 12 hyperparameter settings, 4 learning rules

Cross-validated, trained on split of observable trajectories

“Virtual Experimental” Approach

Data Generation | **Observable Statistics** | **Classifier Training** | **Classifier Testing**



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

Training Epochs

Cross-validated, trained on split of observable trajectories

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

Is this problem even tractable?

Visualizing observables on ImageNet

Visualizing observables on ImageNet

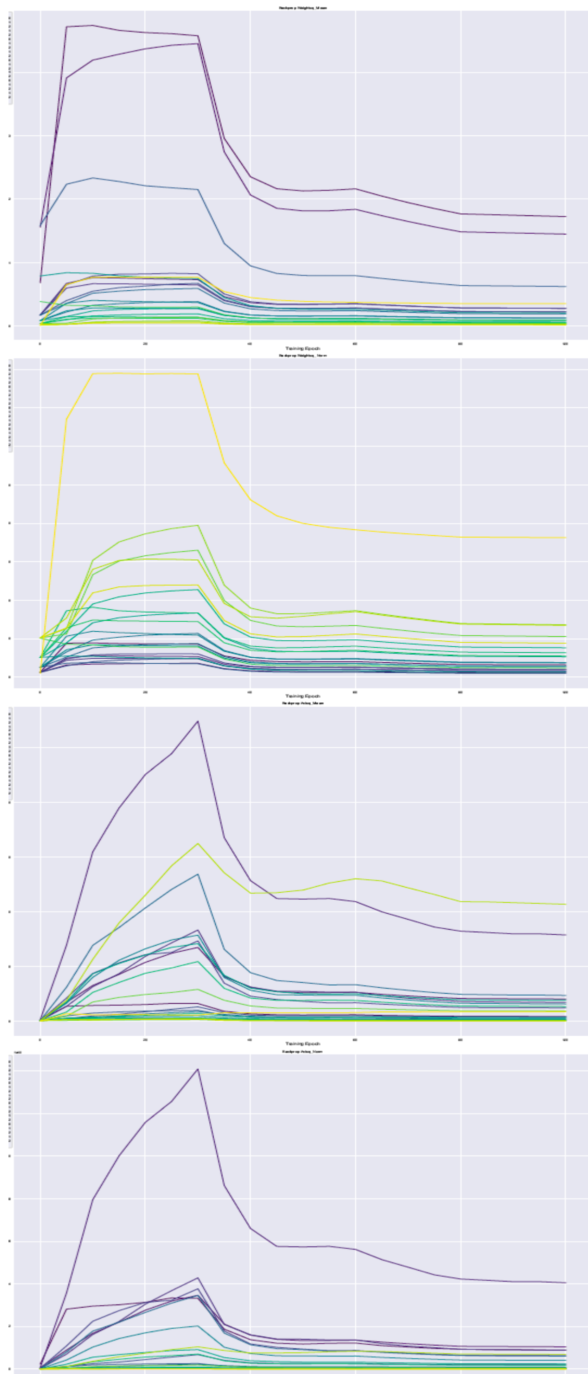
SGDM

weight-square-mean

weight-square-norm

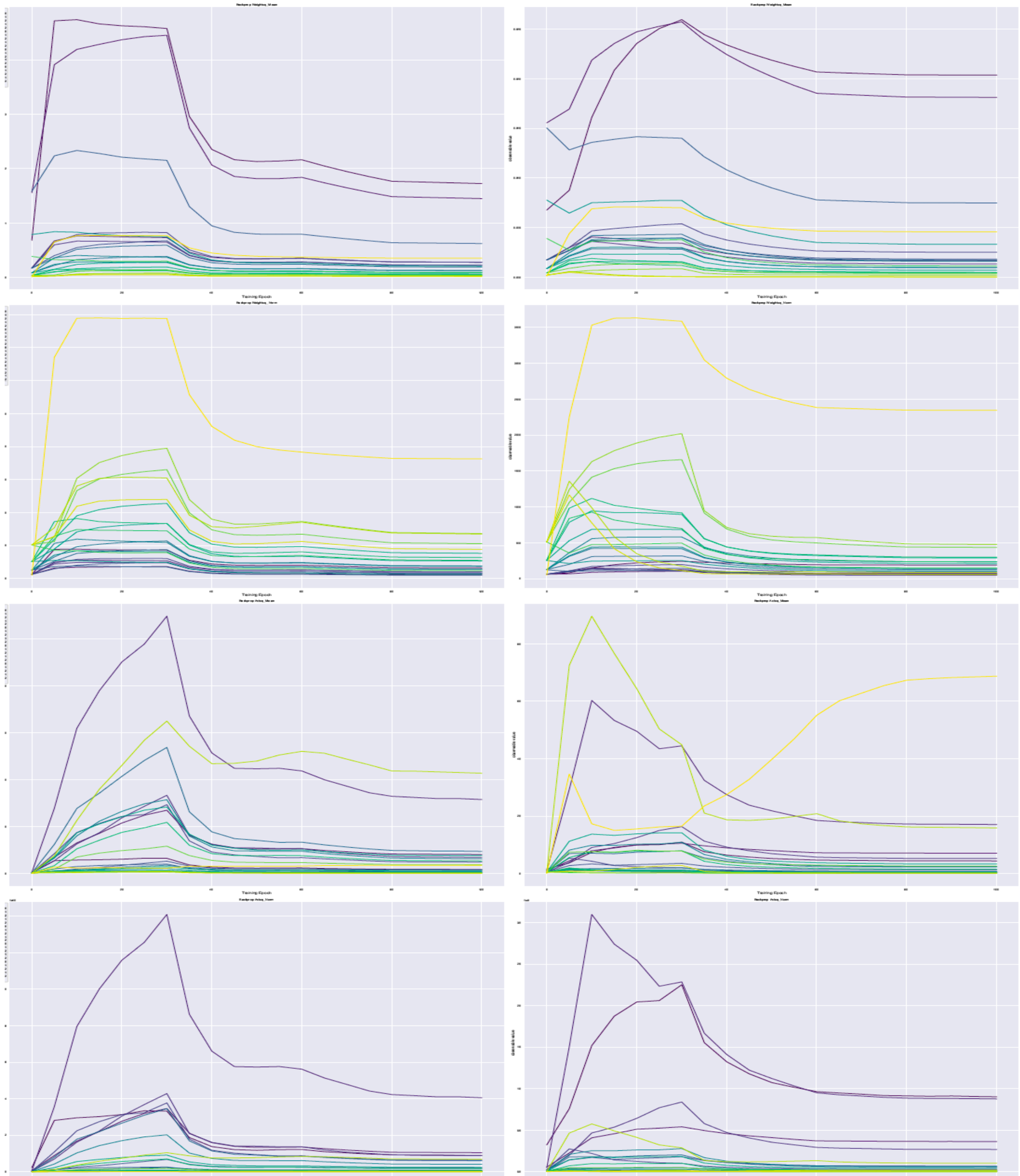
activation-square-mean

activation-square-norm

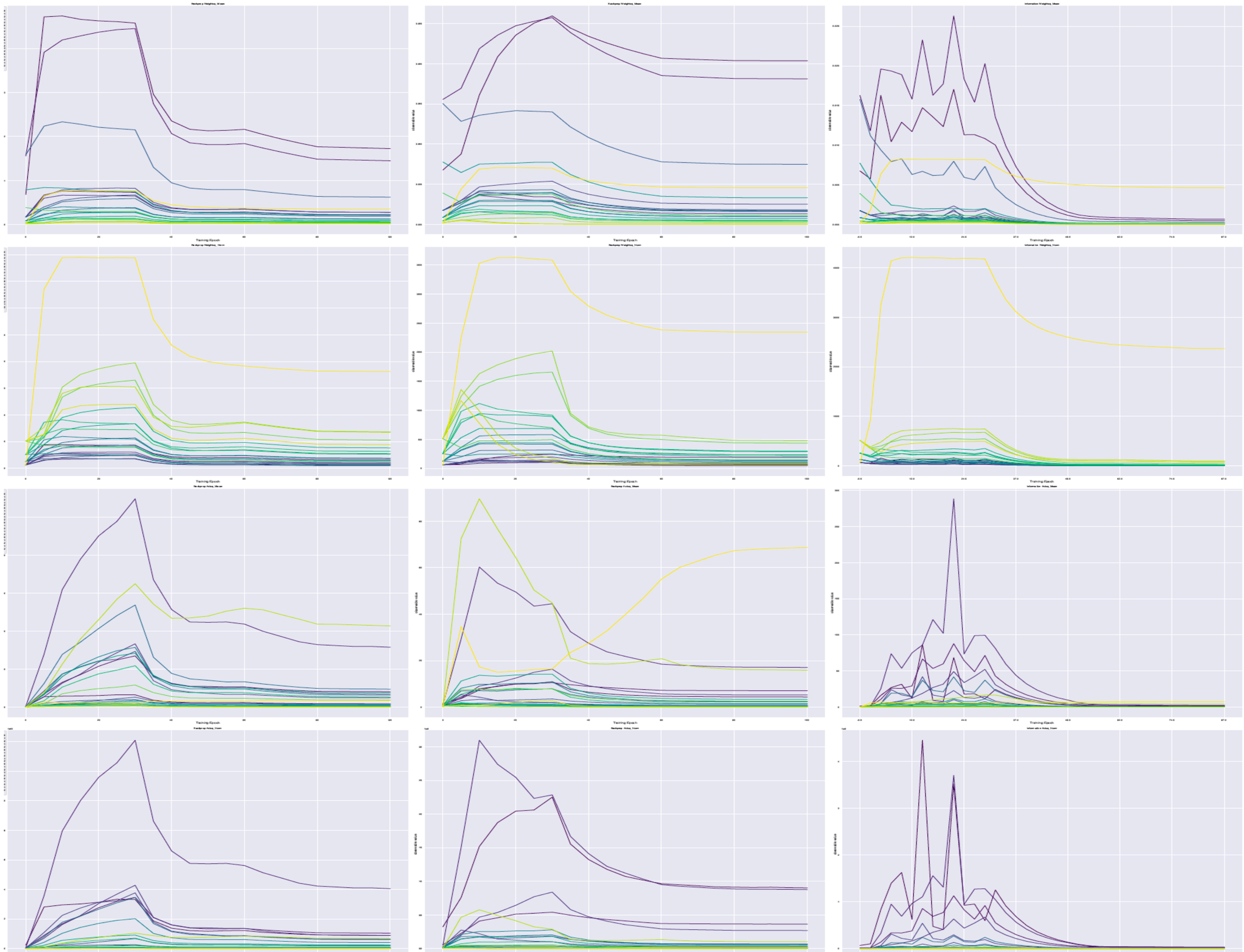


Training Steps

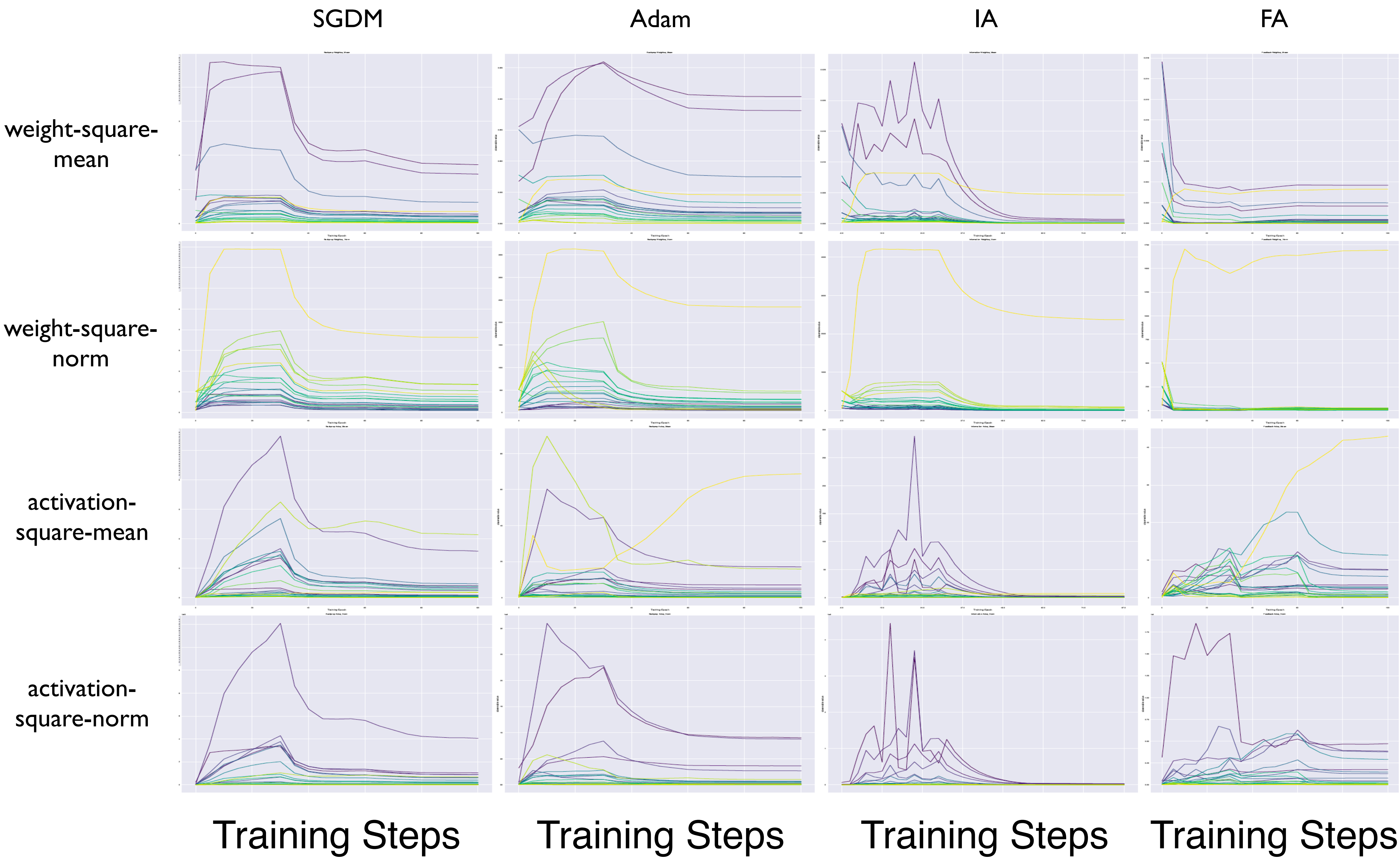
Visualizing observables on ImageNet



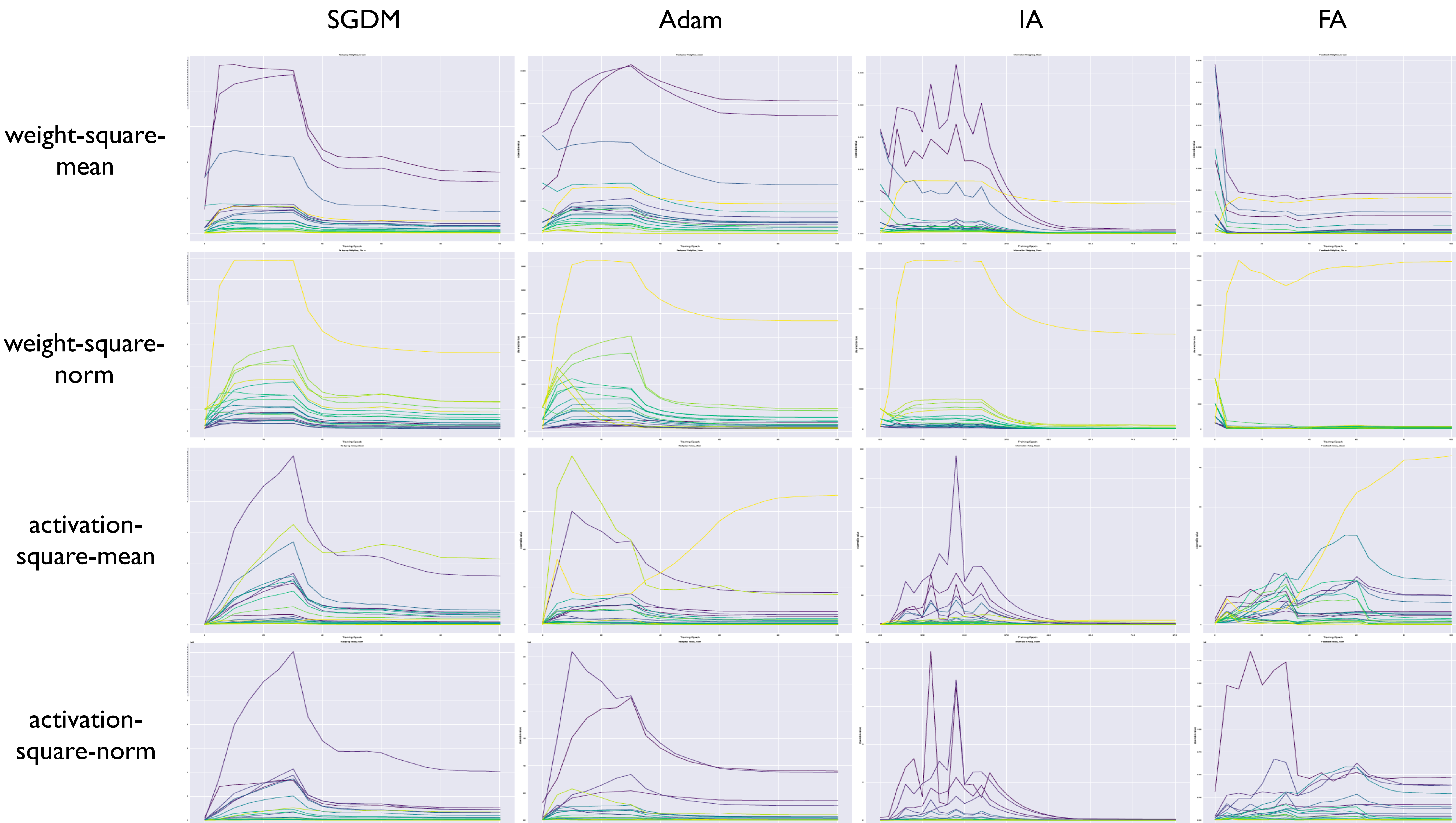
Visualizing observables on ImageNet



Visualizing observables on ImageNet



Visualizing observables on ImageNet

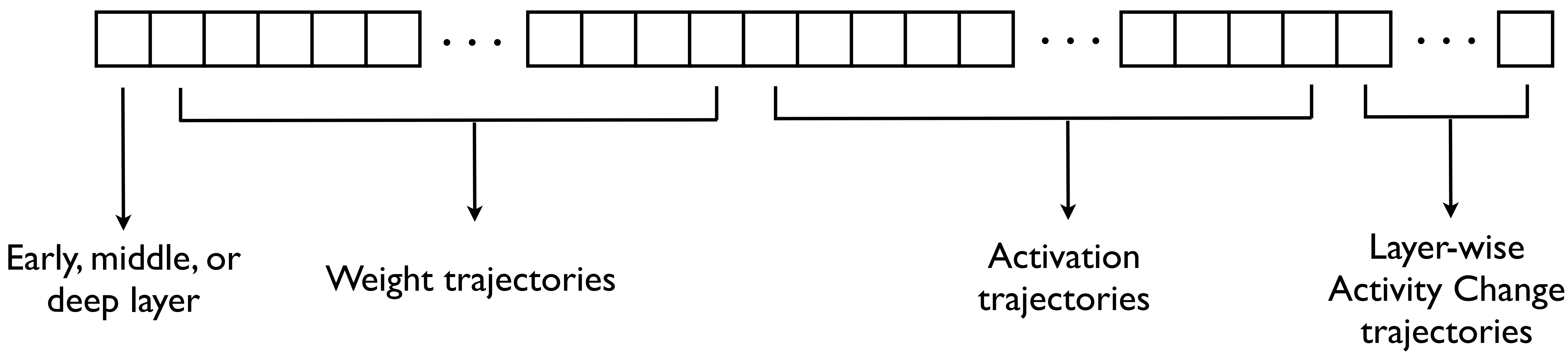


Trajectories across network training appear highly distinctive

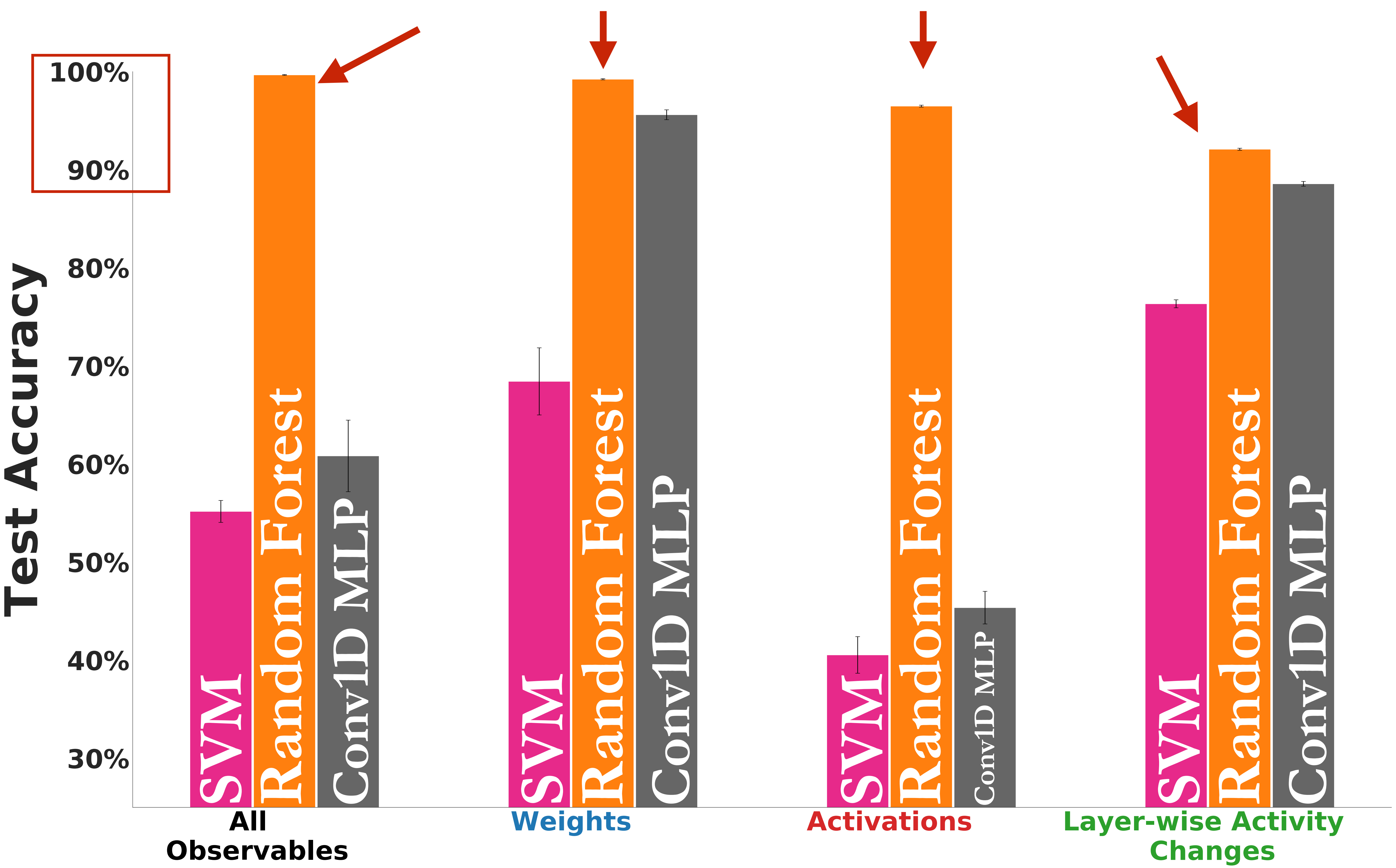
Framing it as a classification problem

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

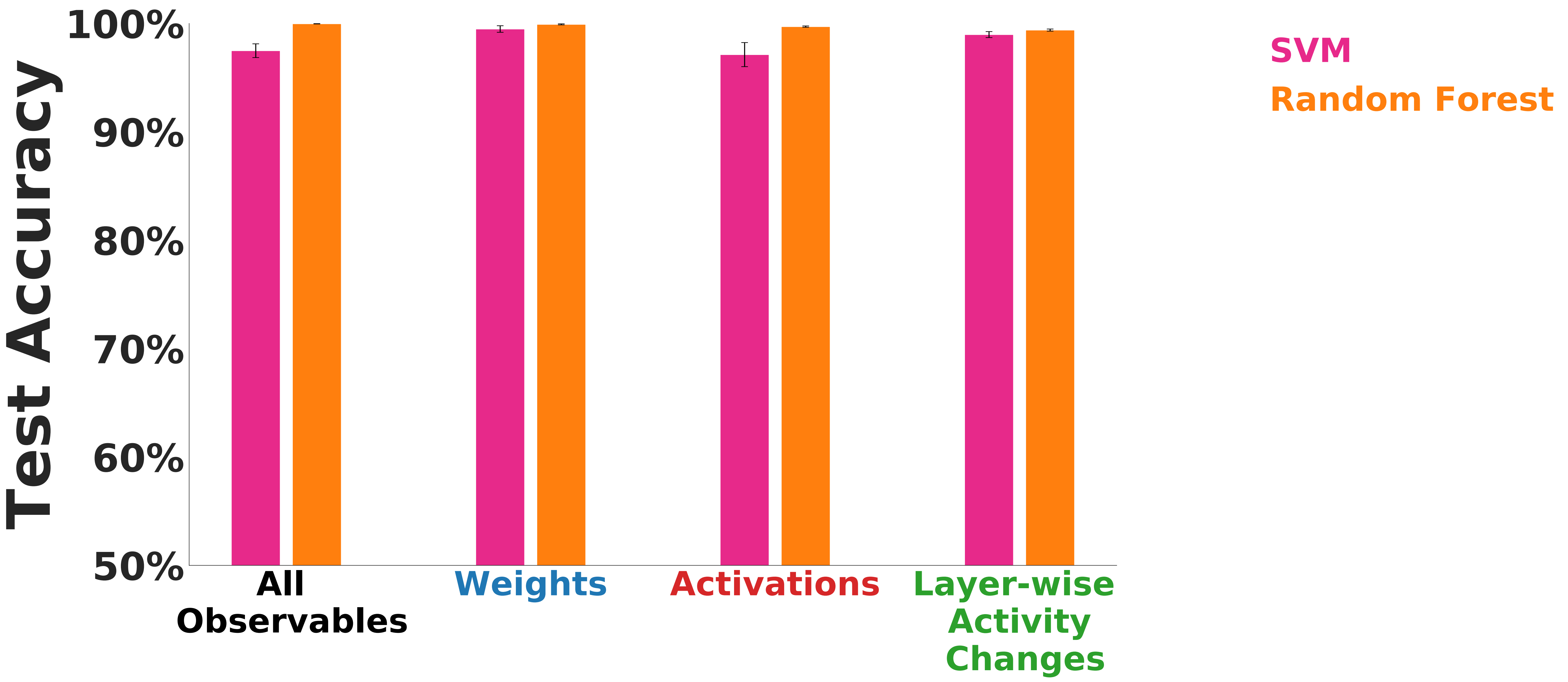
Sample is constructed from one layer of a trained network



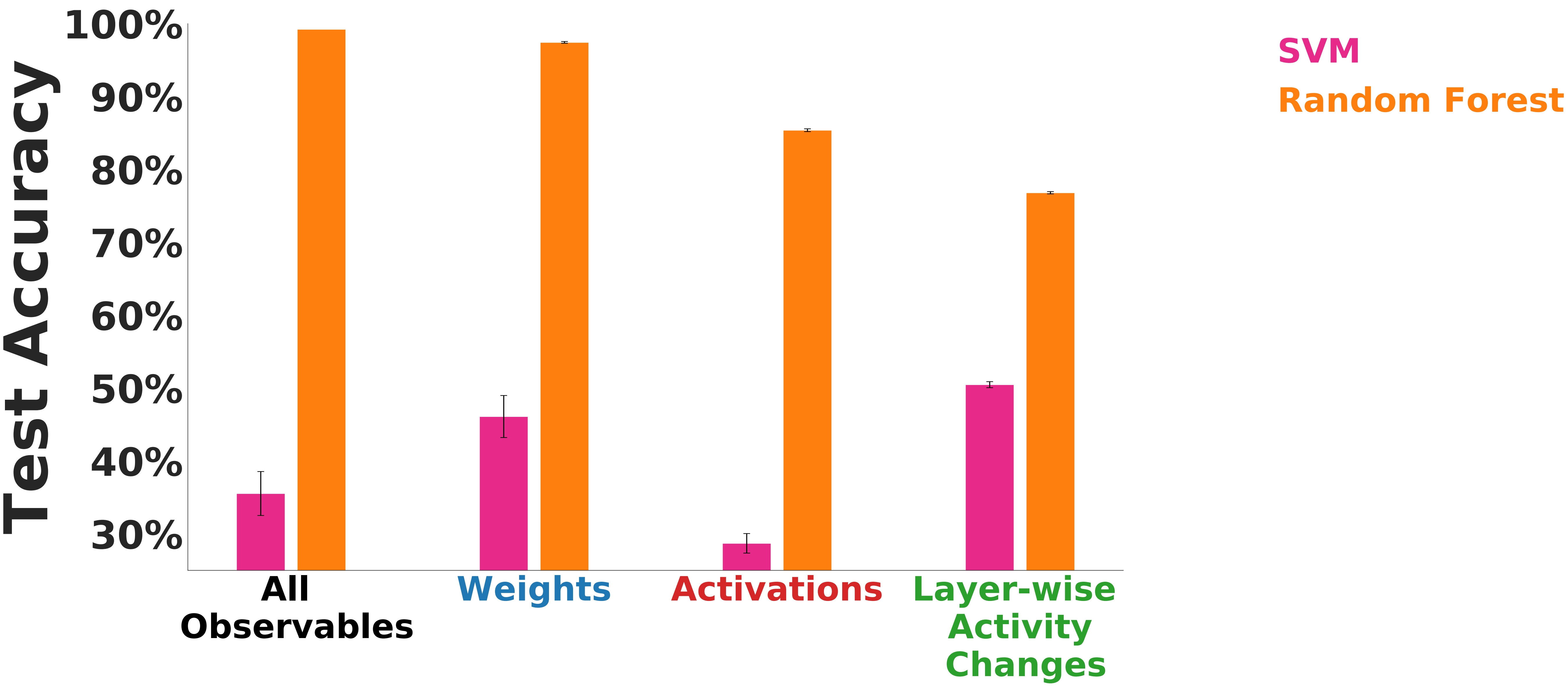
General separability problem *is* tractable



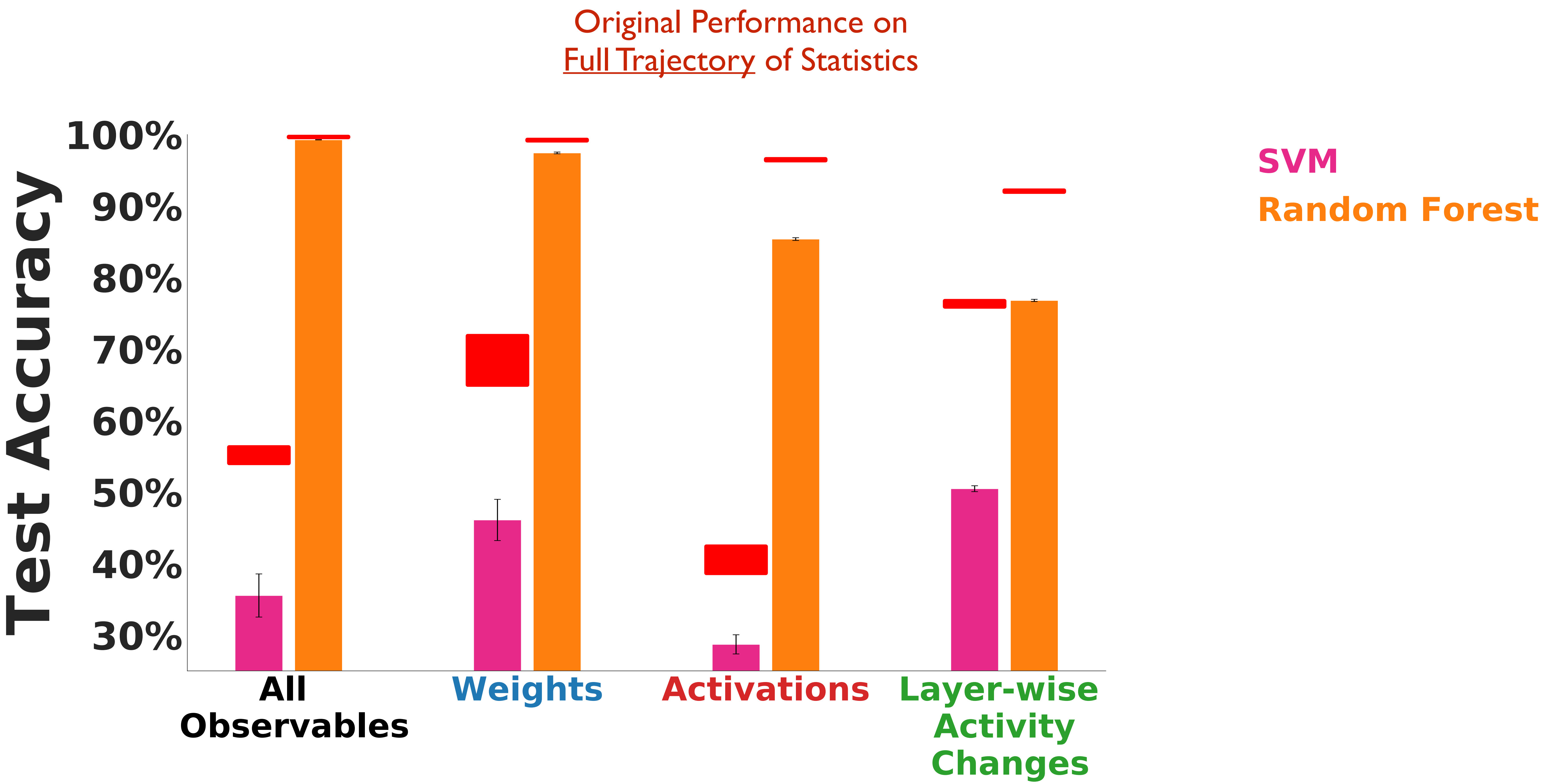
Not driven by task performance (where definable)



Scale of observable statistics is *not* sufficient in most cases



Scale of observable statistics is *not* sufficient in most cases



Adding Experimental Realism

Removing certain “animals” or “training curricula”: holdouts of entire input classes

Access to only portions of the learning trajectory: subsampling observable trajectories

Incomplete and noisy measurements: subsampling units and Gaussian noise before collecting observables

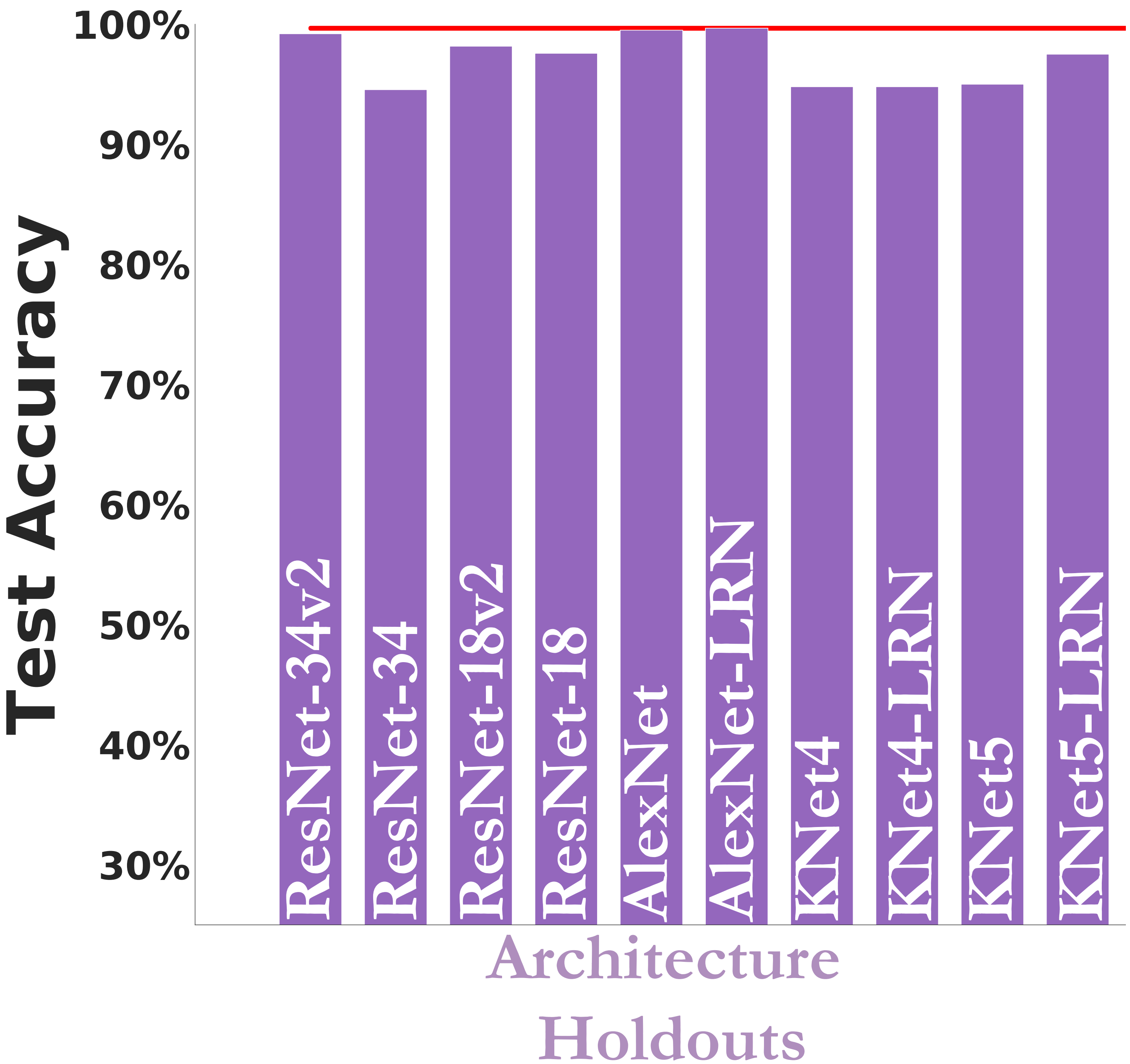
Adding Experimental Realism

Removing certain "animals" or "training curricula": holdouts of entire input classes

Access to only portions of the learning trajectory: subsampling observable trajectories

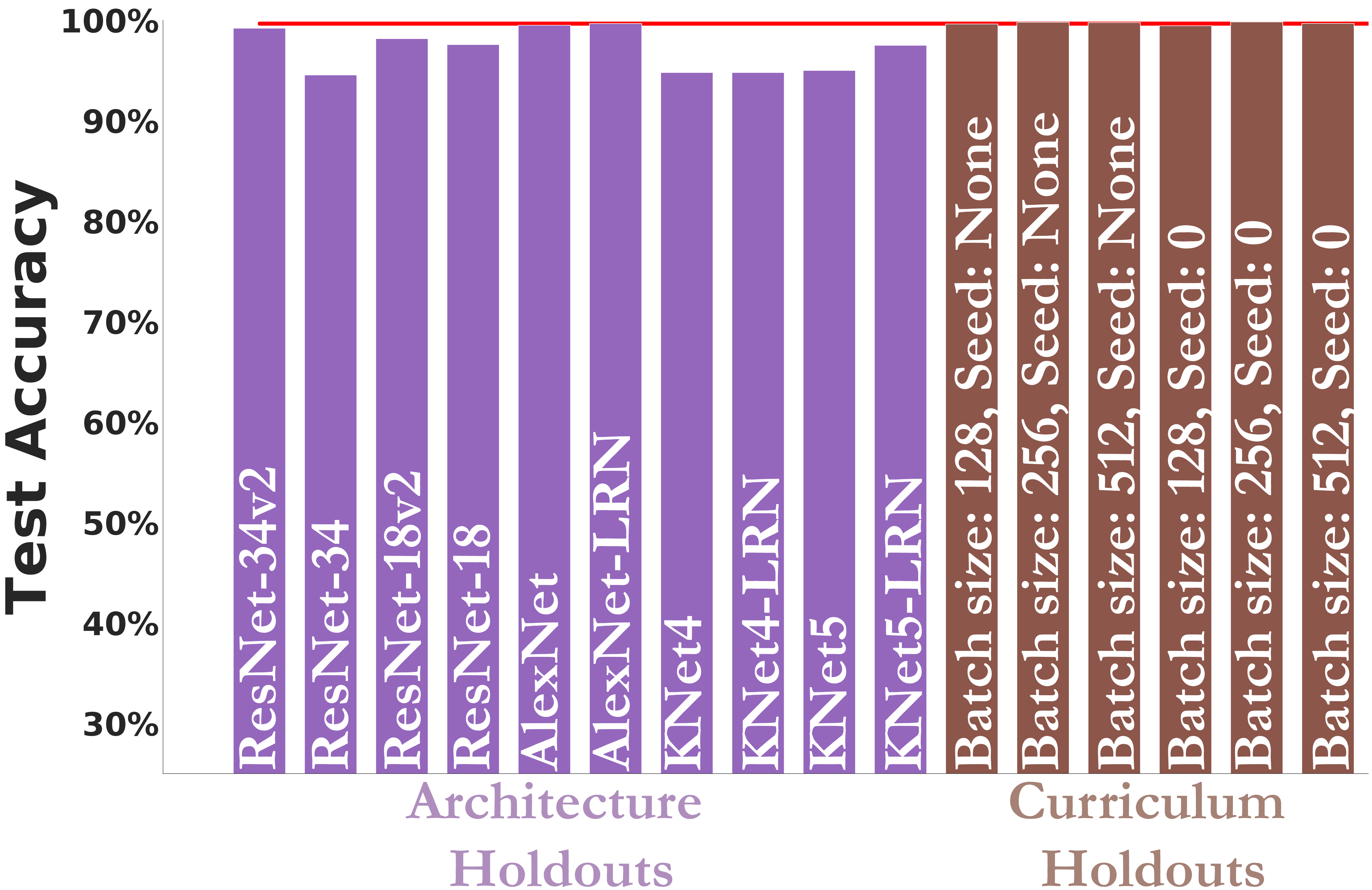
Incomplete and noisy measurements: subsampling units and Gaussian noise before collecting observables

Generalization to held-out "animals"



Animal "=" Architecture

Generalization to held-out "training curricula"



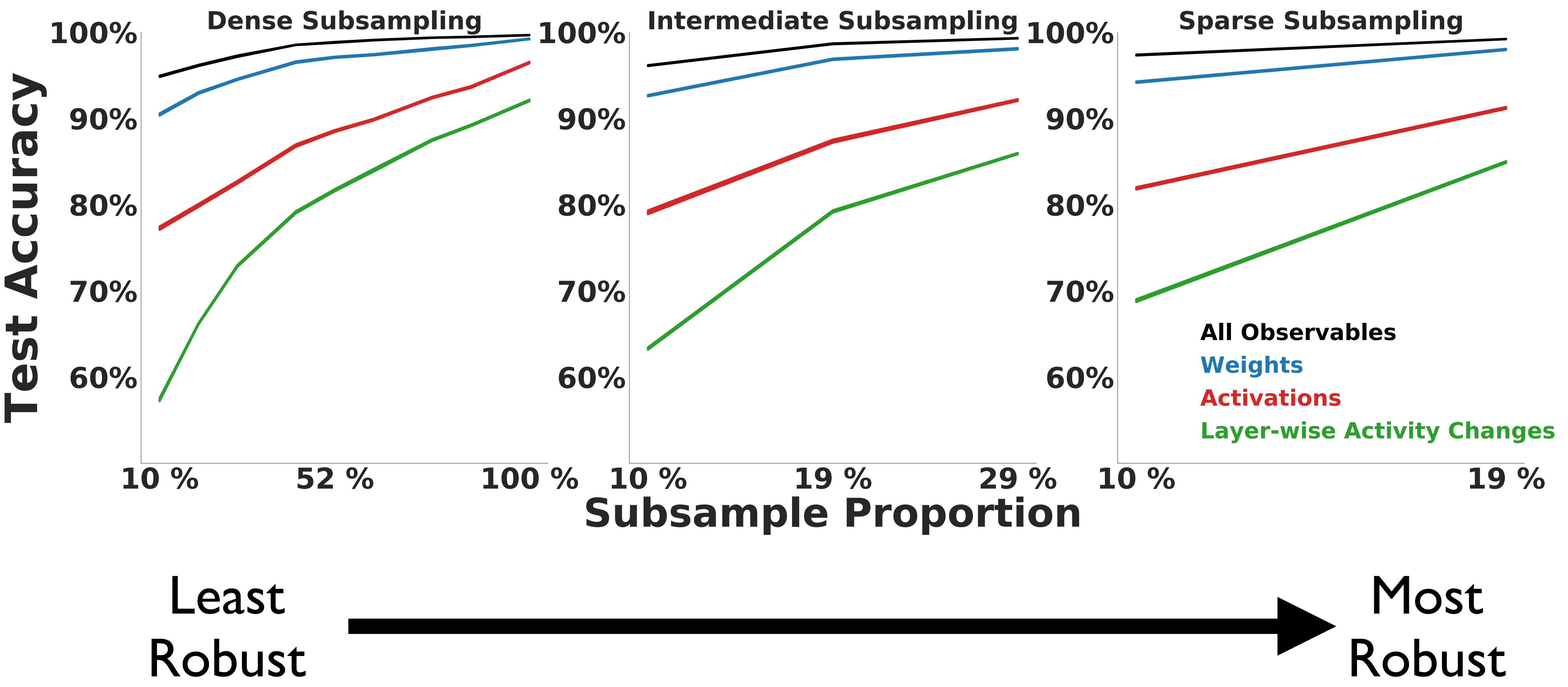
Adding Experimental Realism

Removing certain "animals" or "training curricula": holdouts of entire input classes

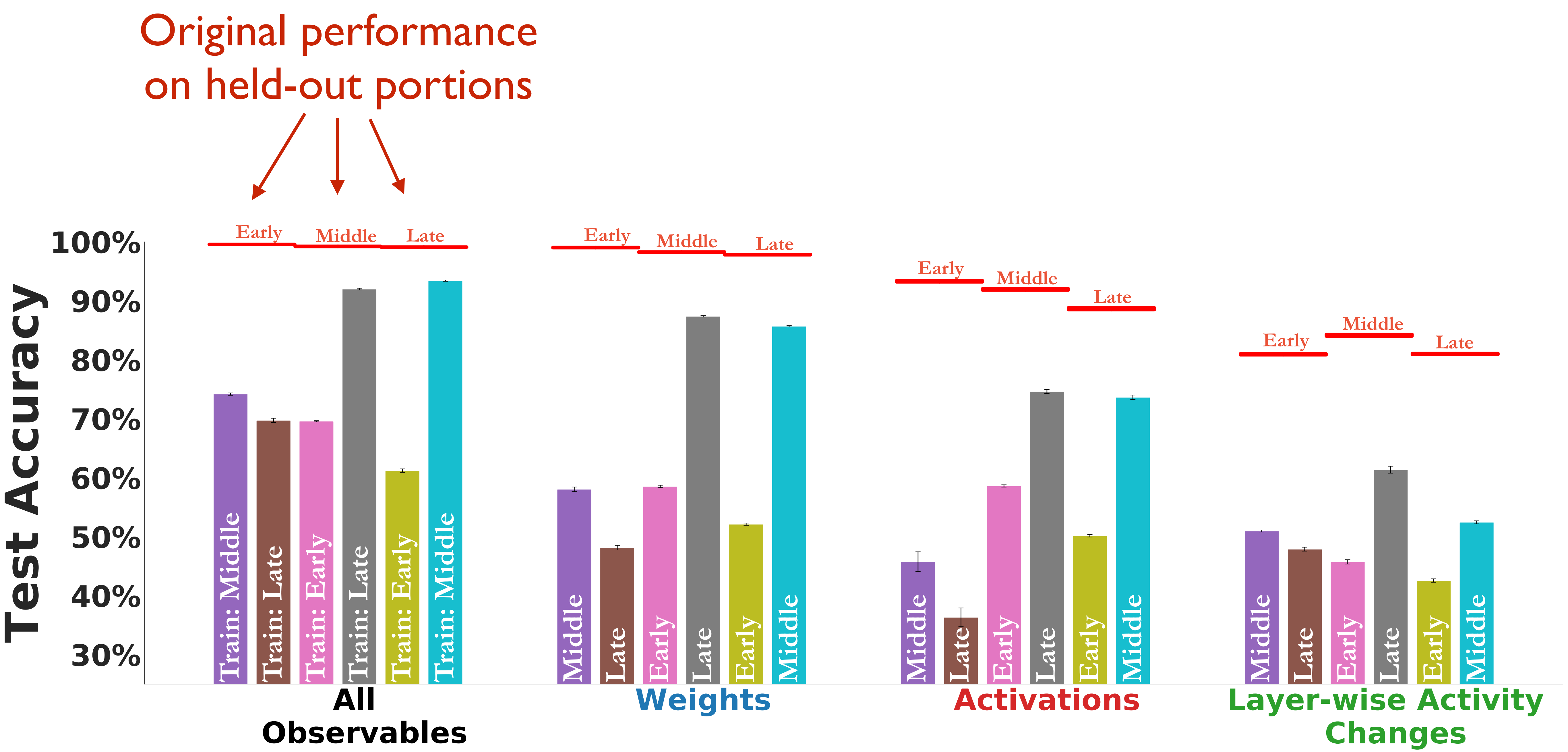
Access to only portions of the learning trajectory: subsampling observable trajectories

Incomplete and noisy measurements: subsampling units and Gaussian noise before collecting observables

Sparse subsampling *across* learning trajectory robust to trajectory undersampling



Sampling *across* learning trajectory is important for robustness to undersampling



Adding Experimental Realism

Removing certain "animals" or "training curricula": holdouts of entire input classes

Access to only portions of the learning trajectory: subsampling observable trajectories

Incomplete and noisy measurements: subsampling units and Gaussian noise before collecting observables

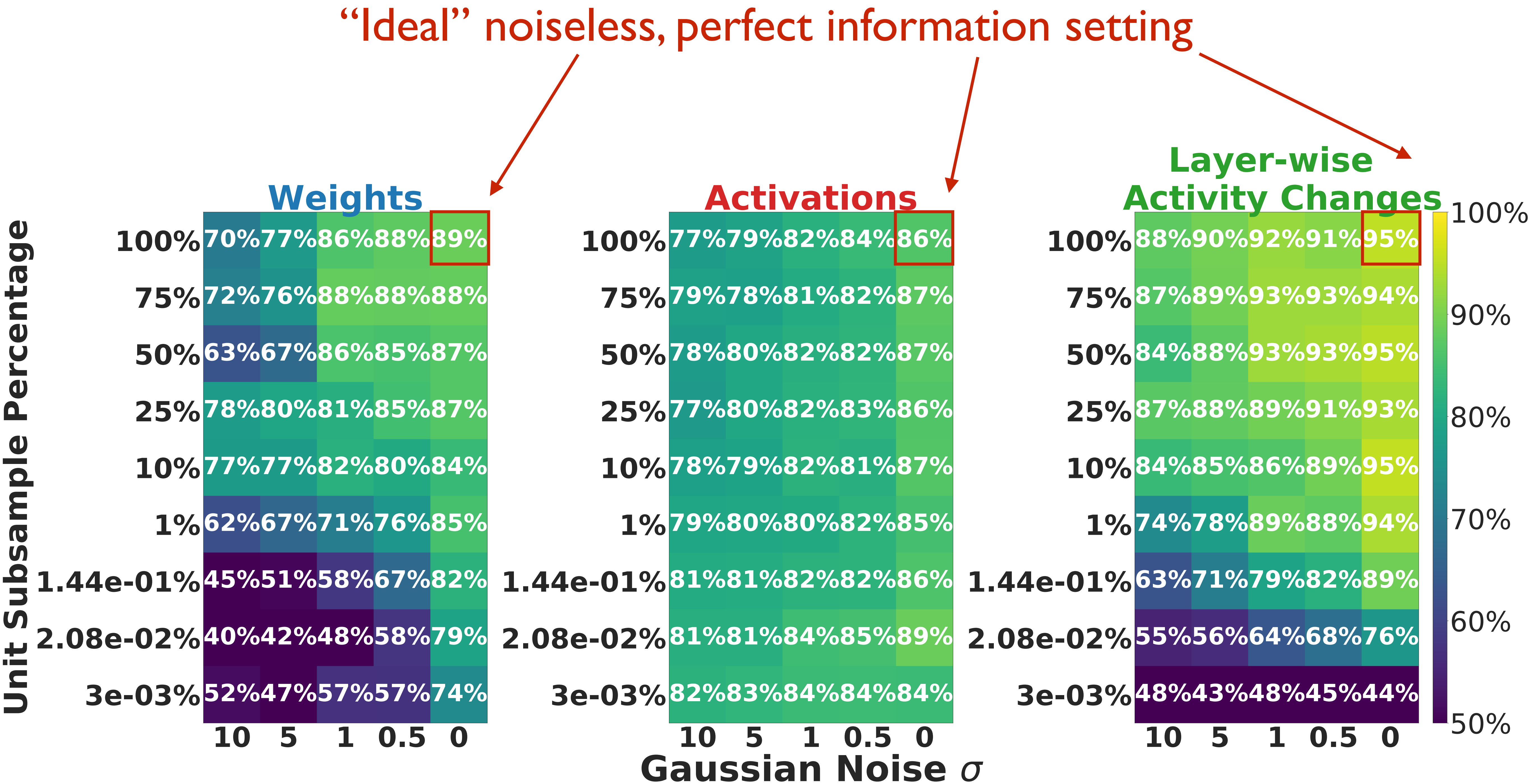
What insights could this approach potentially provide?

Different experimental tools have different limitations

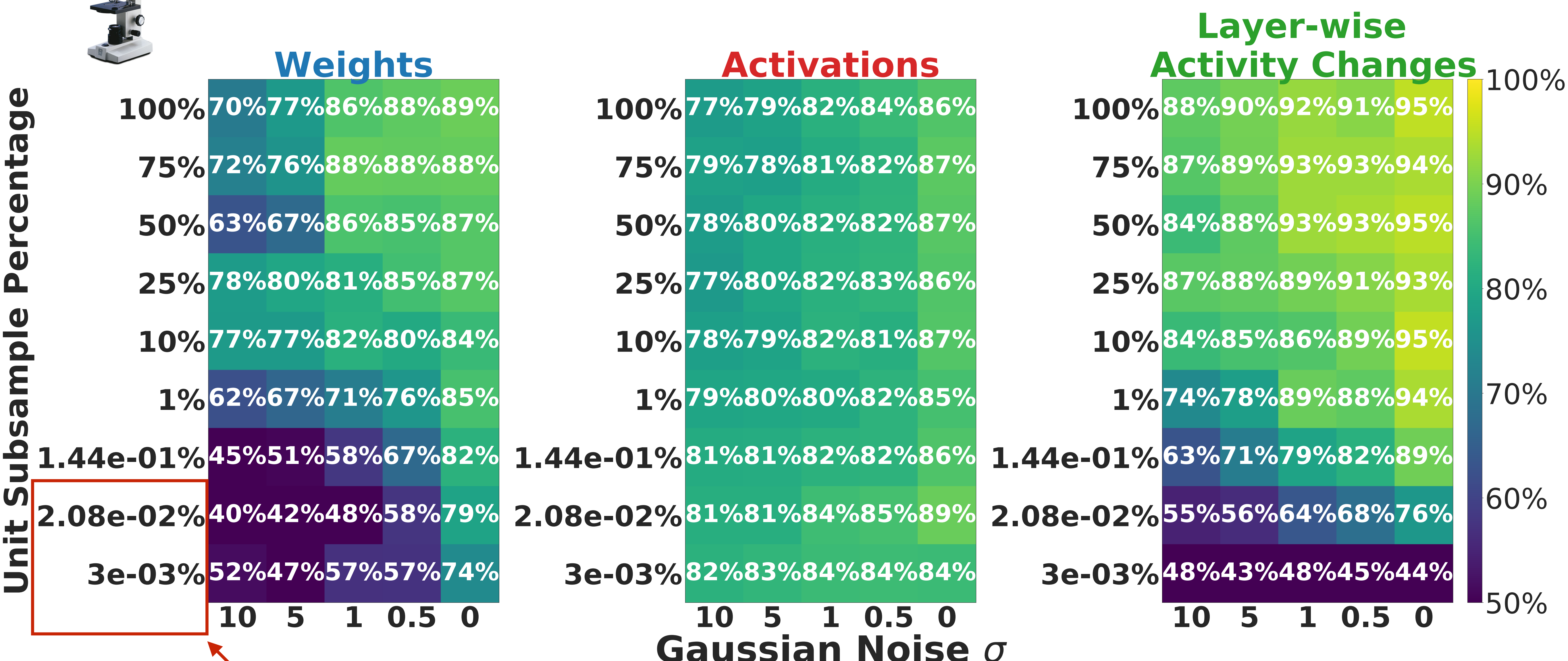
Optical imaging techniques usually give us simultaneous access to thousands of units but can have lower temporal resolution and signal-to-noise

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

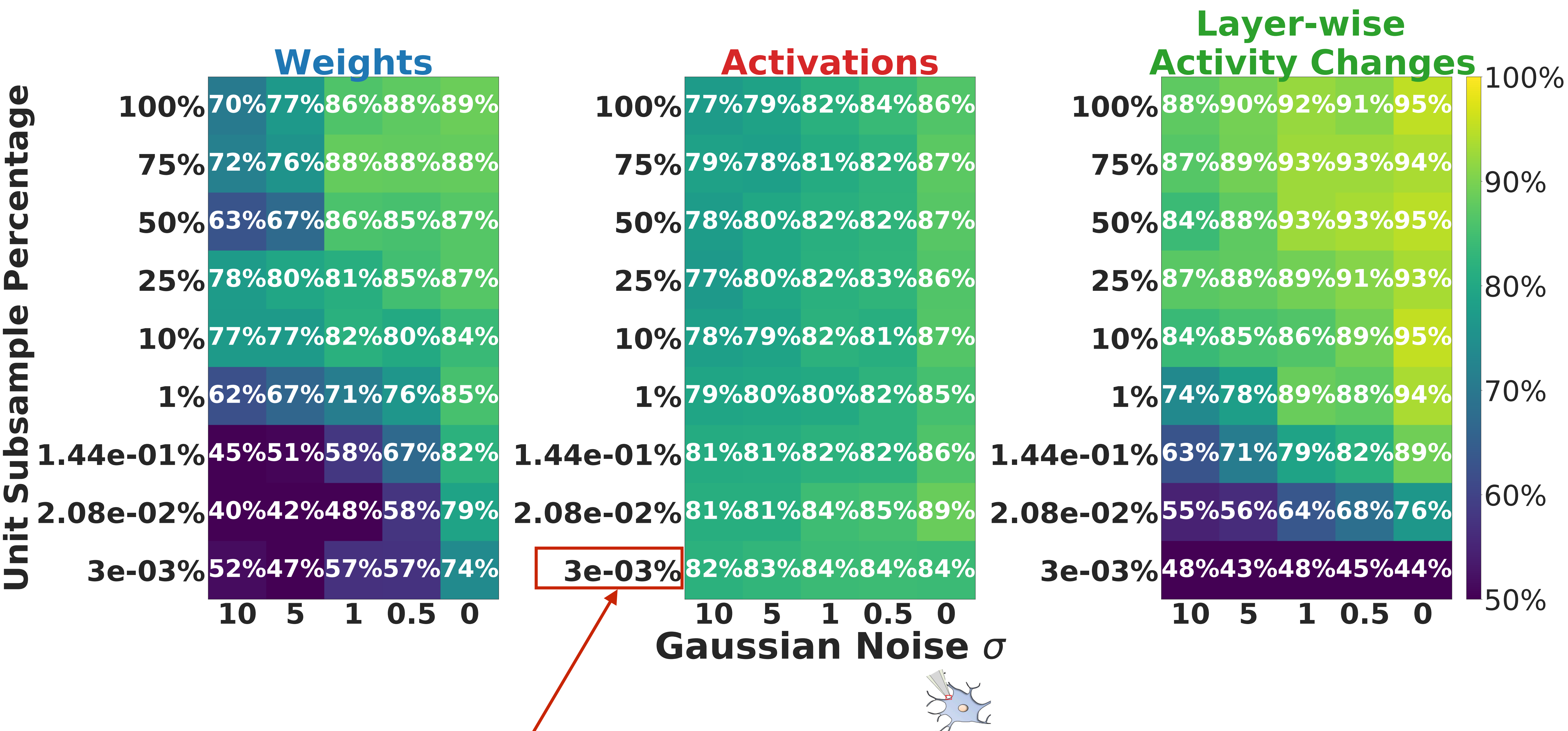


Weights are *not* robust to measurement noise and unit undersampling



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

Conclusions

Conclusions

Hypothesis: *in vivo* electrophysiological recordings of post-synaptic activities from a neural circuit on the order of several hundred units, frequently measured at wider intervals during the course of learning, may provide a good basis on which to identify learning rules

Conclusions

Hypothesis: *in vivo* electrophysiological recordings of post-synaptic activities from a neural circuit on the order of several hundred units, frequently measured at wider intervals during the course of learning, may provide a good basis on which to identify learning rules

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

Conclusions

Hypothesis: *in vivo* electrophysiological recordings of post-synaptic activities from a neural circuit on the order of several hundred units, frequently measured at wider intervals during the course of learning, may provide a good basis on which to identify learning rules

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

Acknowledgements

Thanks!


Sanjana Srivastava


Surya Ganguli


Daniel Yamins

Contact:

Email: anayebi@stanford.edu

Twitter: @aran_nayebi

NeurIPS 2020 Paper: https://arxiv.org/abs/2010.11765

Code & Dataset: https://github.com/neuroailab/lr-identify

Experimental collaborations welcome!