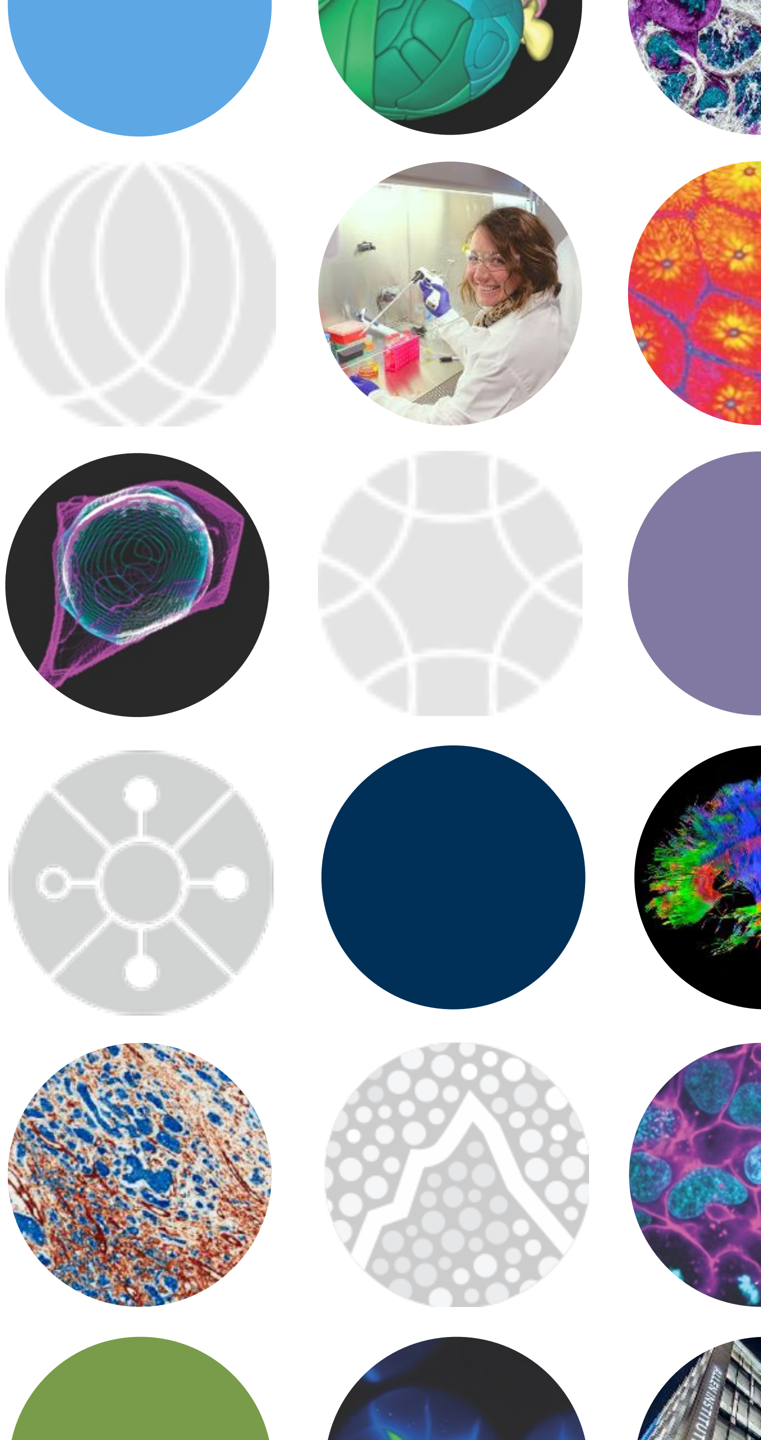
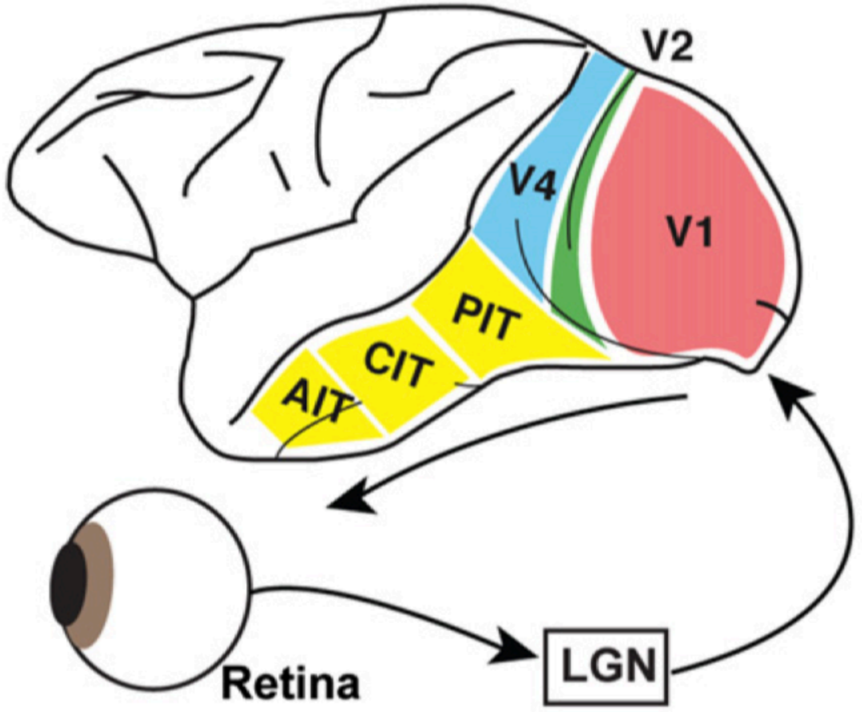
Learning Dynamics in Deep Networks with Multiple Pathways

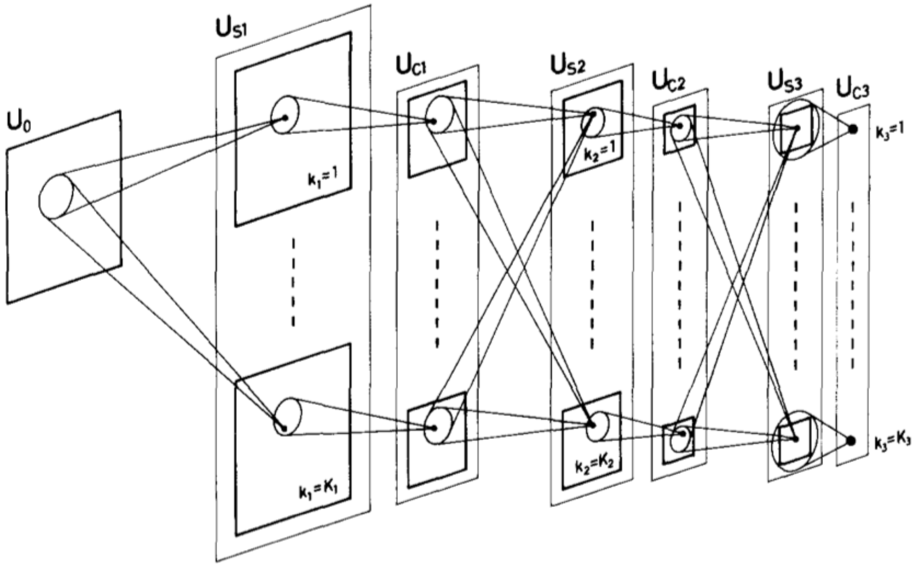
Jianghong Shi
Eric Shea-Brown
Michael A. Buice



Hierarchical architectures have computational advantages



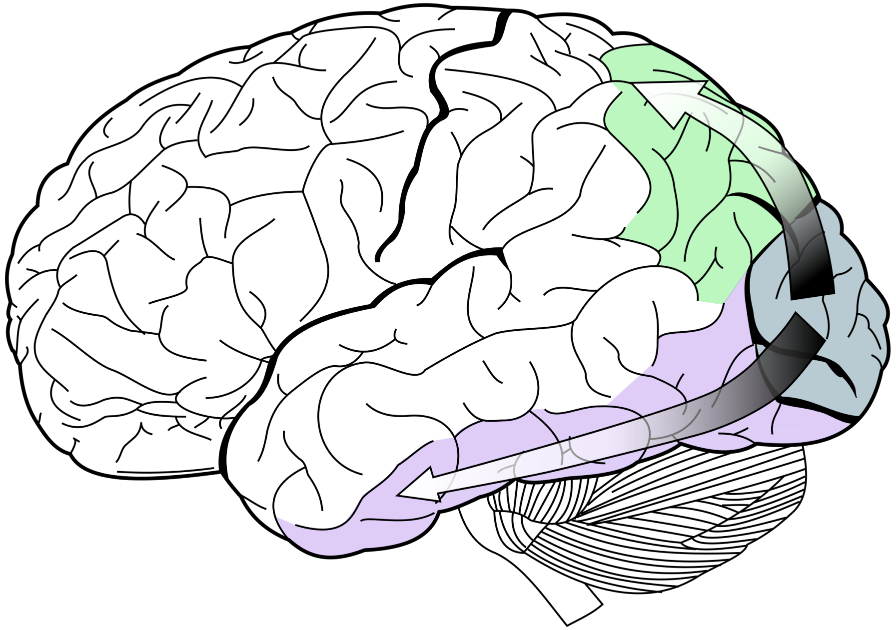
DiCarlo, Zoccolan, Rust 2012



Fukushima 1980

Biological systems have multiple computational pathways

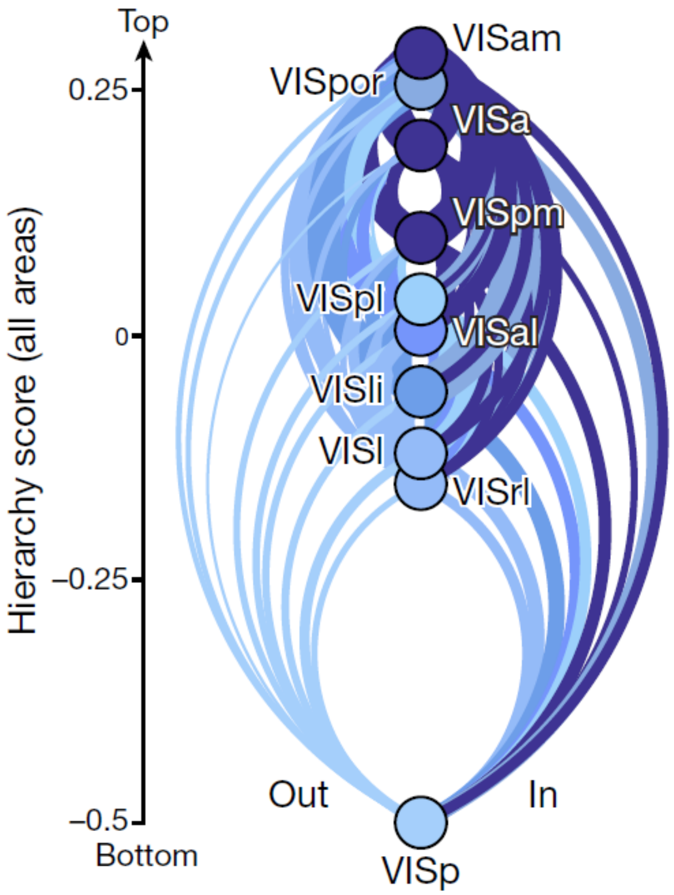
Dorsal - “spatial awareness”



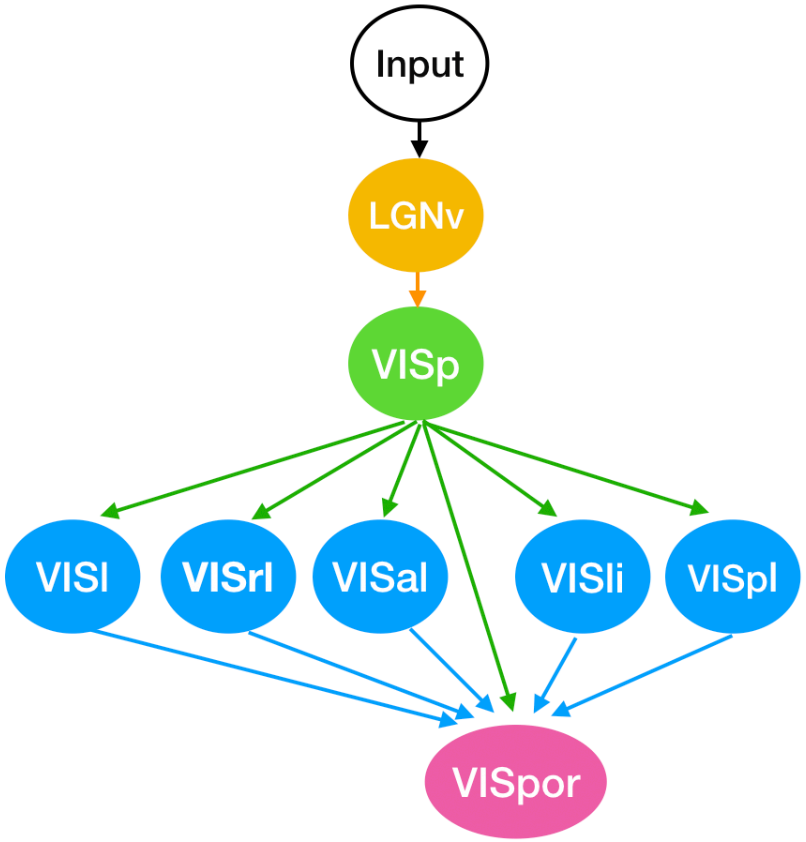
Ventral - “object properties”

Also: “Local” vs “Global” motion across various species

The architecture of the mouse visual system is parallel



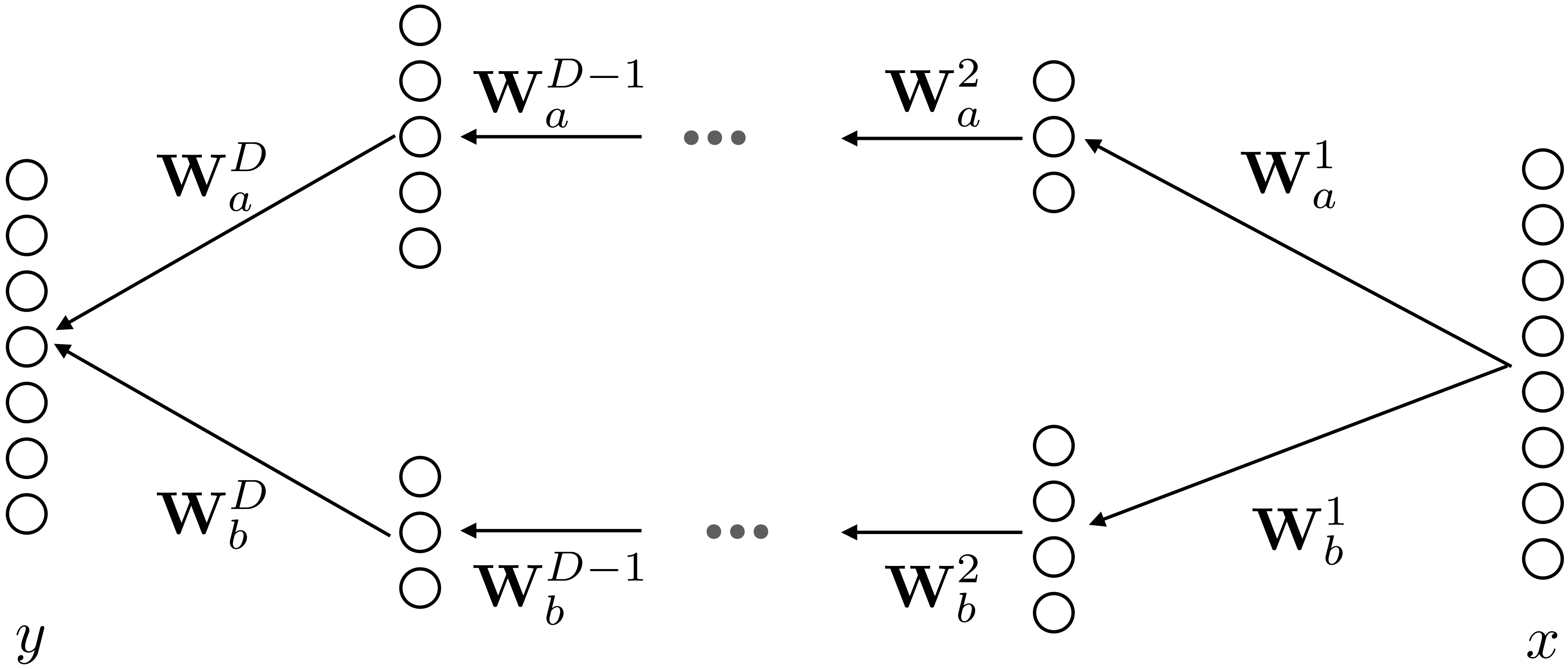
Harris, Mihalas, et al 2019



MouseNet

Shi, et al PLoS Comp Bio 2022

How does learning work in parallel pathways?



Shi, et al, NeurIPS 2022

What does this network learn?

- Minimize the mean squared error:

$$L = \frac{1}{2} \sum_{i=1}^{P} ||y^i - \mathbf{\Omega} x^i||^2$$

$$\mathbf{\Omega} = \sum_{a=1}^{M} \mathbf{\Omega}_a$$

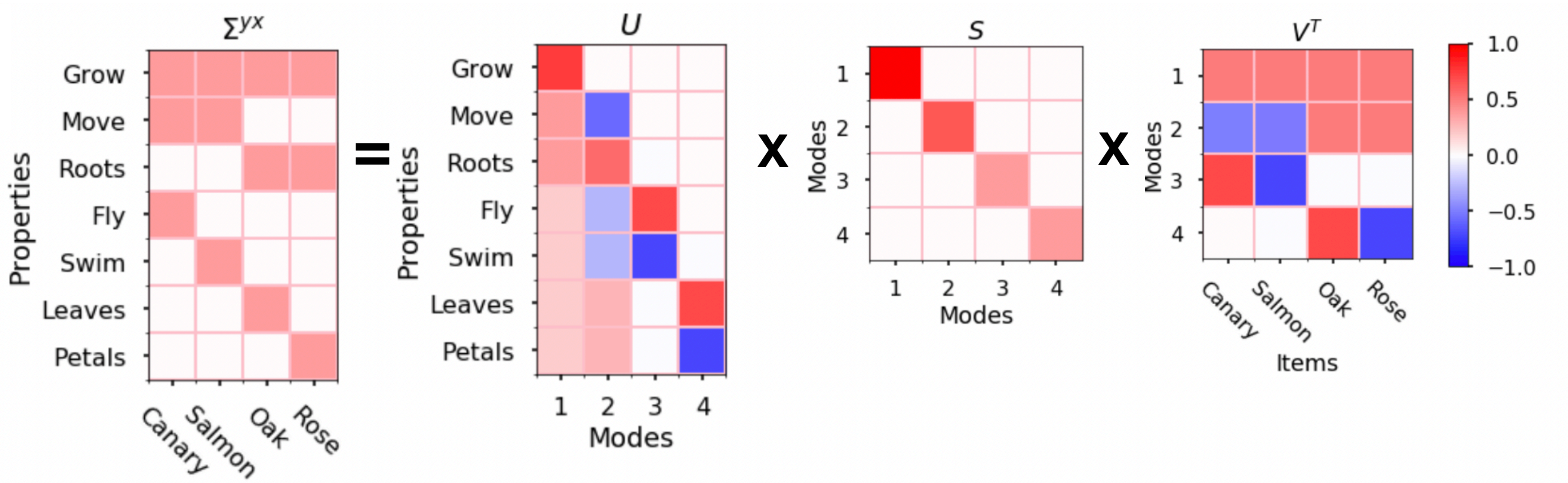
- Omega is the effective weight matrix that defines the computation in the network. After learning it will satisfy:

$$\mathbf{\Sigma}^{yx} = \mathbf{\Omega}$$

$$(\text{with } \mathbf{\Sigma}^{x} = \mathbf{I})$$

Shi, et al, NeurIPS 2022

Features relating input to output are captured in the singular vectors.

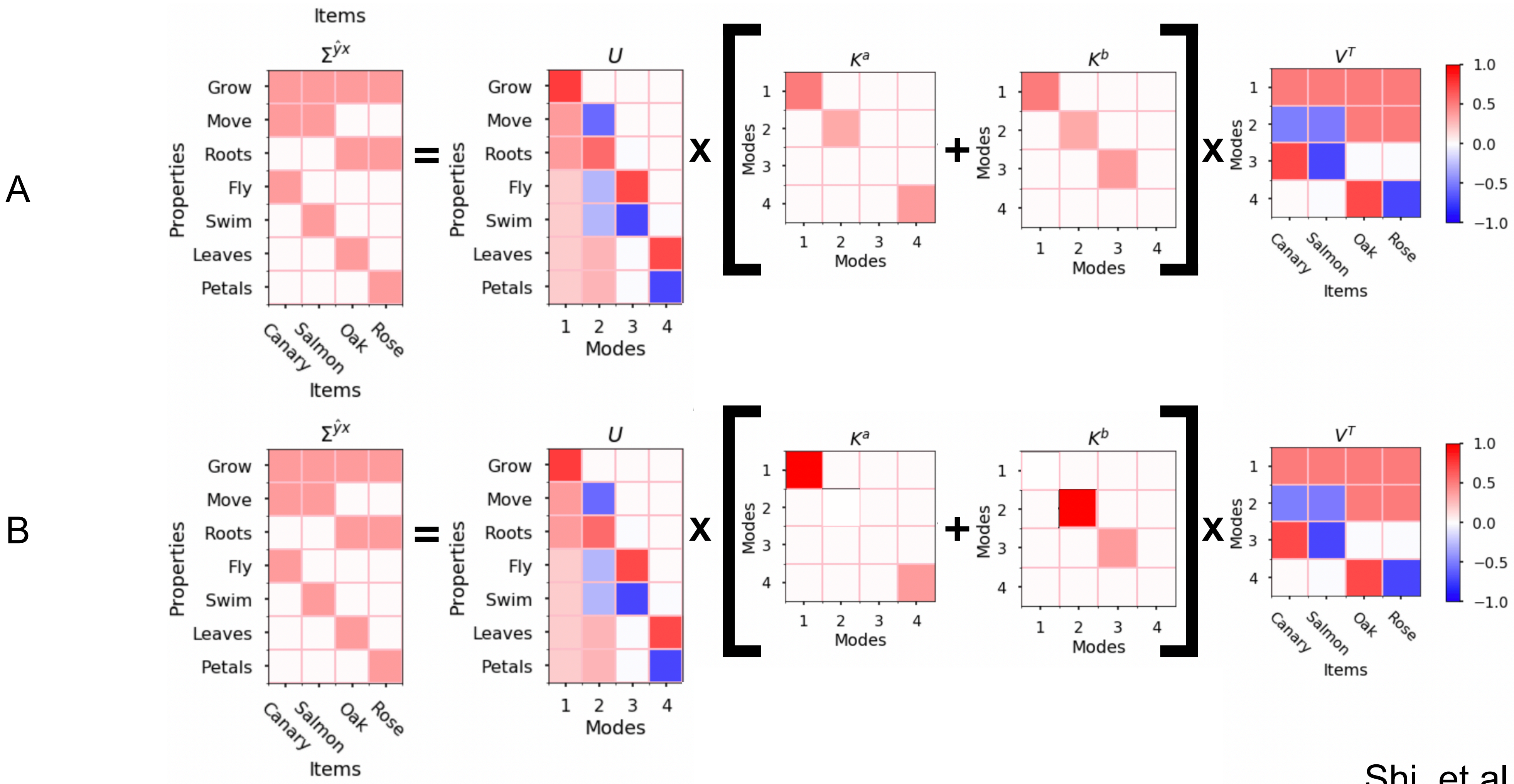


$$\Sigma^{yx} = \Omega$$

Saxe, et al 2019

Shi, et al, NeurIPS 2022

How do singular vectors distribute across pathways?



Shi, et al, NeurIPS 2022

What are the dynamics of learning in this network?

- Assume a *linearized framework* (e.g. Saxe, et al 2014)

- Minimize the mean squared error:

$$L = \frac{1}{2}\sum_{i=1}^{P} ||y^i - \mathbf{\Omega} x^i||^2$$

$$\mathbf{\Omega} = \sum_{a=1}^{M} \mathbf{\Omega}_a \equiv \sum_{a=1}^{M} \prod_{d=1}^{D_a} \mathbf{W}_a^d$$

- Consider Gradient Descent:

$$\tau \frac{d}{dt}\mathbf{W}_a^d = \left(\prod_{i=d+1}^{D_a} \mathbf{W}_a^i\right)^T [\mathbf{\Sigma}^{yx} - \mathbf{\Omega}\mathbf{\Sigma}^{x}] \left(\prod_{i=1}^{d-1} \mathbf{W}_a^i\right)^T$$

Shi, et al, NeurIPS 2022

Assumptions

- Each layer has a large number of units (i.e. “infinitely wide”).
- Initial weight values are drawn from a zero mean Gaussian distribution with variance $\frac{\sigma^2}{N}$

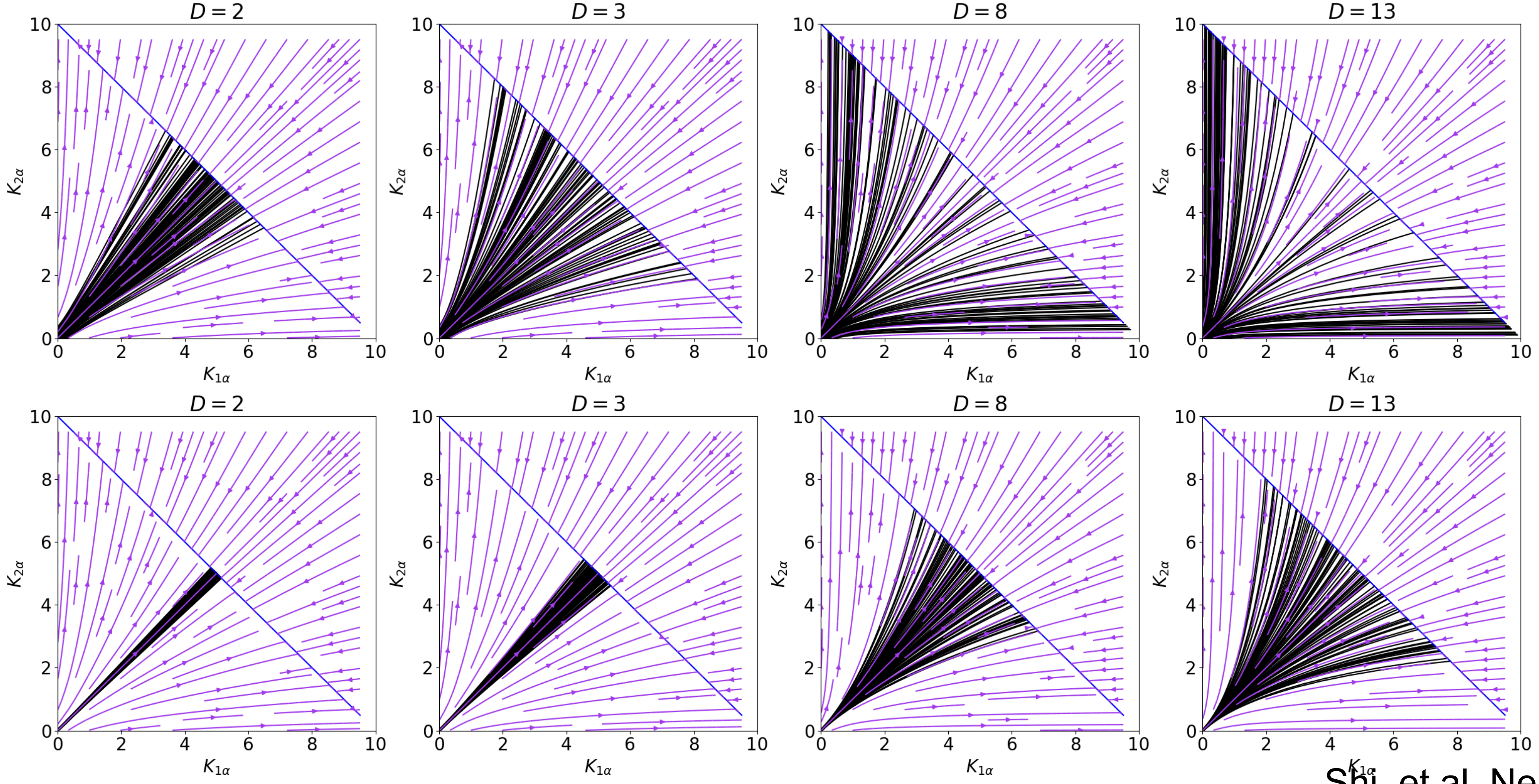
We can define coordinates in singular vector space for each layer in each channel.

$$\tau \frac{d}{dt} q_{a\alpha} = q_{a\alpha}^{D_a - 2} p_{a\alpha} \left[S_\alpha - \bar{\Omega}_\alpha \right]$$

$$\tau \frac{d}{dt} p_{a\alpha} = q_{a\alpha}^{D_a - 1} \left[S_\alpha - \bar{\Omega}_\alpha \right]$$

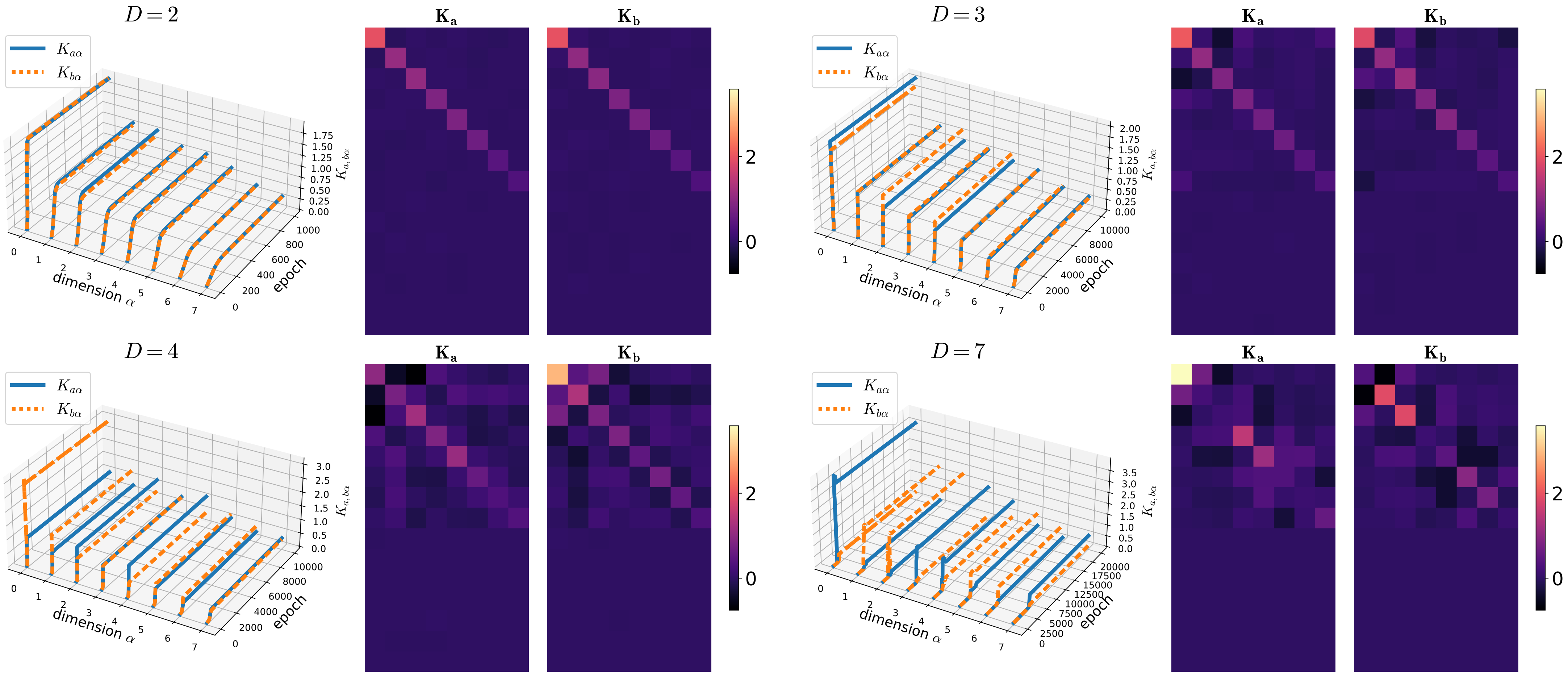
Shi, et al, NeurIPS 2022

Singular values concentrate more on one pathway with increasing depth.



Shi, et al, NeurIPS 2022

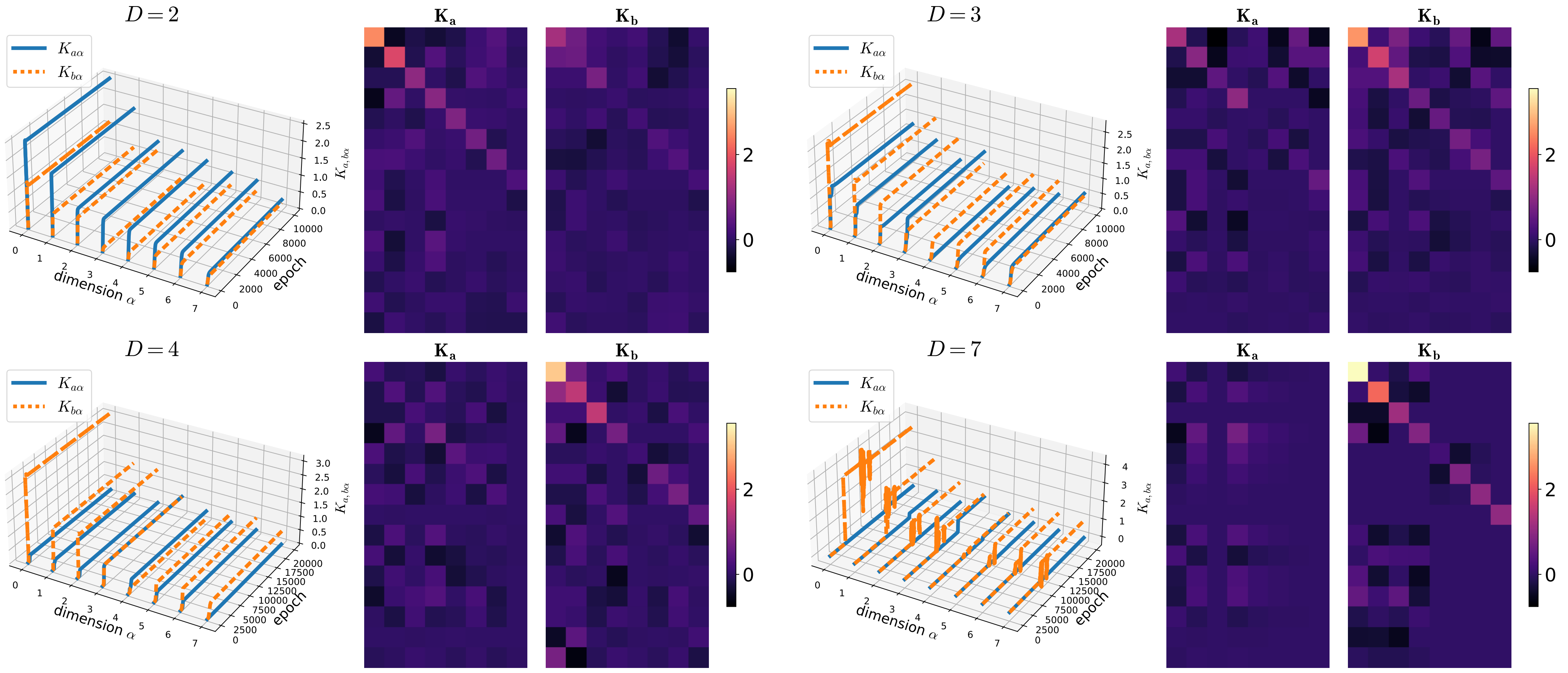
Simulations



Also holds true with nonlinearities! (tanh and ReLU)

Shi, et al, NeurIPS 2022

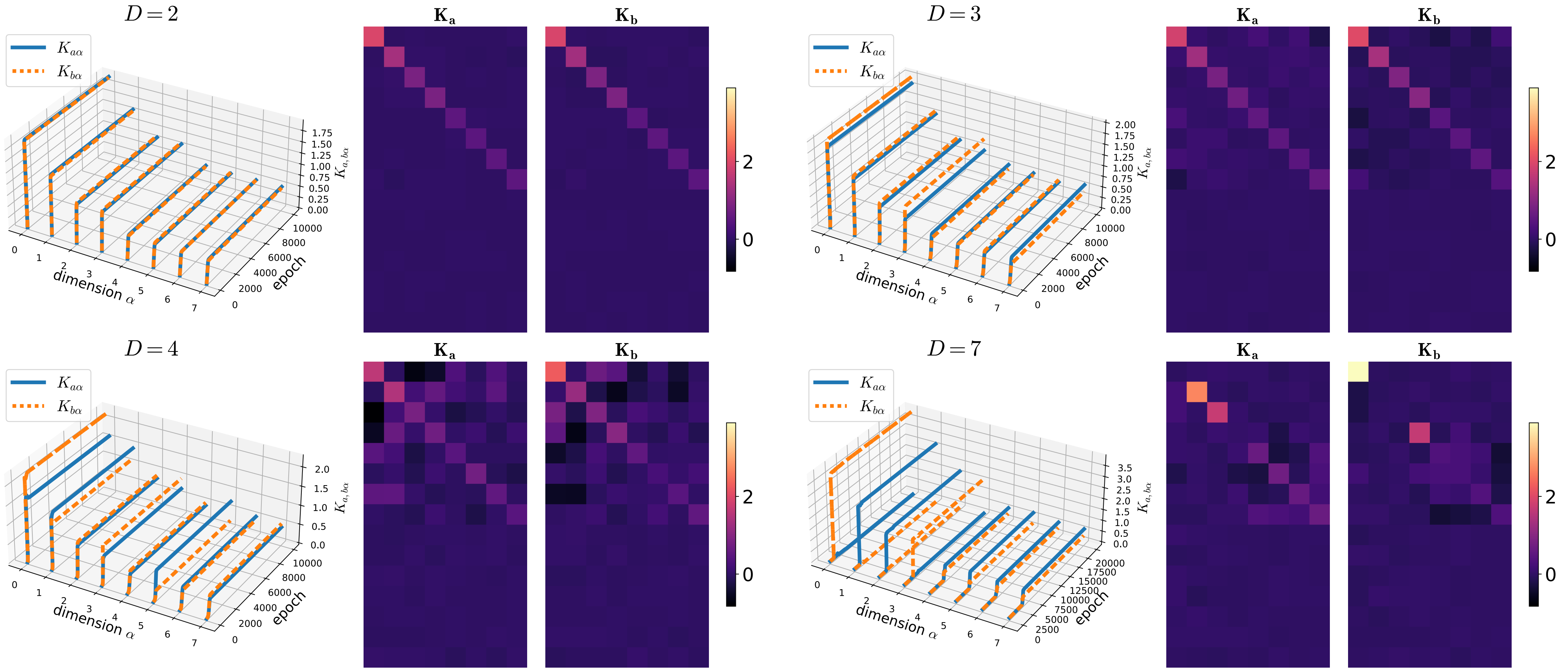
Simulations



ReLU

Shi, et al, NeurIPS 2022

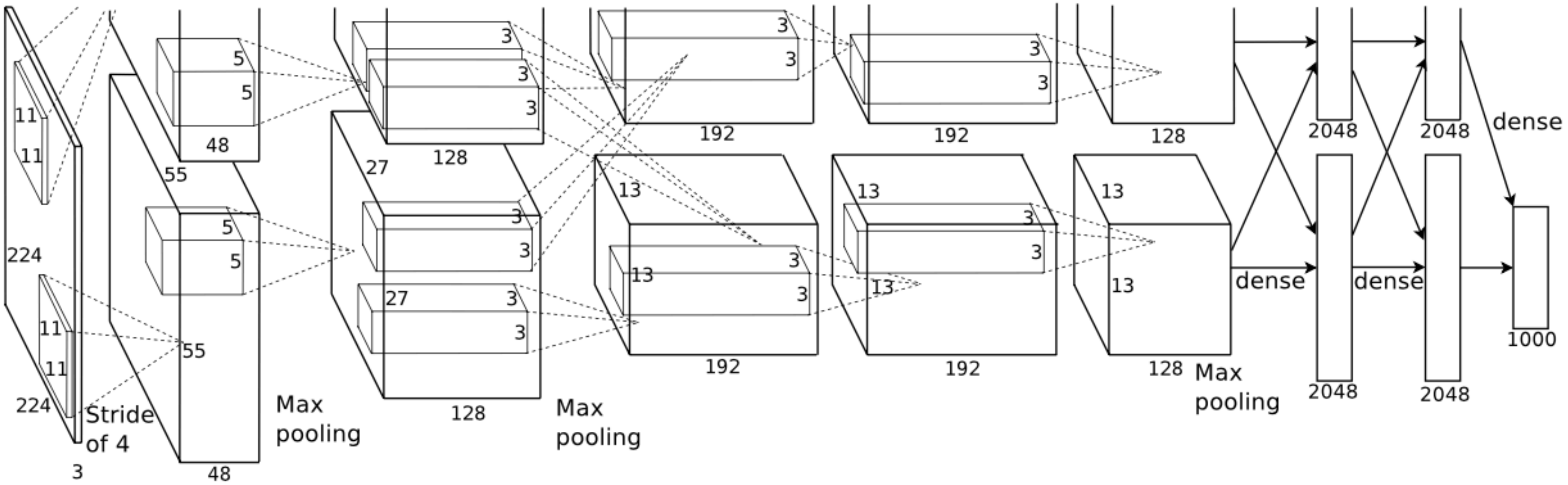
Simulations



Tanh

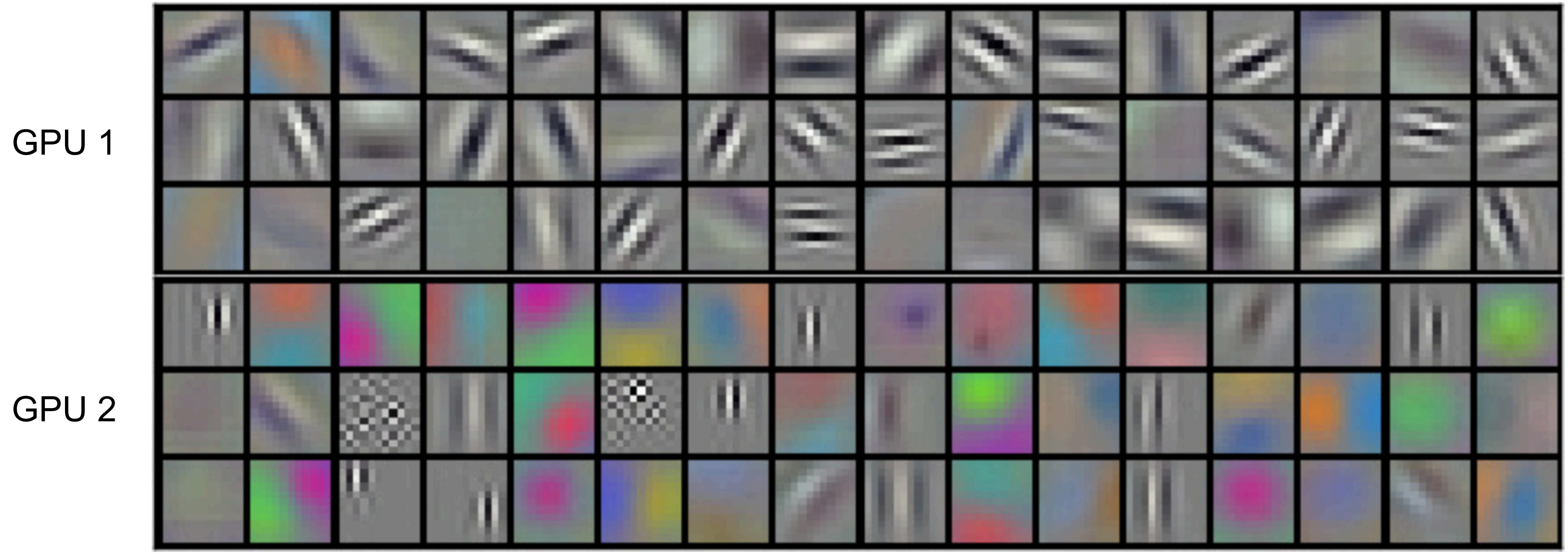
Shi, et al, NeurIPS 2022

AlexNet was constructed with multiple pathways



Krizhevsky, et al 2012

AlexNet learned different types of features in each pathway



Krizhevsky, et al 2012

Summary

- Using a linearized framework, we derive a set of coupled differential equations for the dynamics of gradient descent in networks with multiple pathways.

- The tendency of each singular vector to concentrate on a single pathway increases with depth.

- The results extend in simulations to networks with nonlinearities.

Shi, et al, NeurIPS 2022

THANK YOU

We wish to thank the Allen Institute founder, Paul G. Allen, for his vision, encouragement, and support.

brain-map.org
alleninstitute.org

ALLEN INSTITUTE | MindScope PROGRAM