Make Haste Slowly: A Theory of Emergent Structured Mixed Selectivity in Feature Learning ReLU Networks

Devon Jarvis, Richard Klein, Benjamin Rosman & Andrew Saxe

ICLR 2025

WITS UNIVERSITY

MIND
Machine Intelligence and Neural Discovery Institute

UCL

Sainsbury Wellcome Centre

GATSBY

Research Aim

To obtain analytical equations for the training dynamics of finite, feature learning ReLU neural networks.

Key Properties

- We motivate our new paradigm by the specific properties it can handle:

Key Properties

- We motivate our new paradigm by the specific properties it can handle:
    1. Finite (separate from NTK [1])

[1] Jacot, Arthur, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks." *Advances in neural information processing systems* 31 (2018).

Key Properties

- We motivate our new paradigm by the specific properties it can handle:
    1. Finite (separate from NTK [1])
    2. Feature Learning on structured data (separate from statistical physics [2])

[1] Jacot, Arthur, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks." *Advances in neural information processing systems* 31 (2018).

[2] Goldt, Sebastian, et al. "Modeling the influence of data structure on learning in neural networks: The hidden manifold model." *Physical Review X* 10.4 (2020): 041044.

Key Properties

- We motivate our new paradigm by the specific properties it can handle:
  1. Finite (separate from NTK [1])
  2. Feature Learning on structured data (separate from statistical physics [2])
  3. ReLU Networks (separate from previous linear dynamics [3])

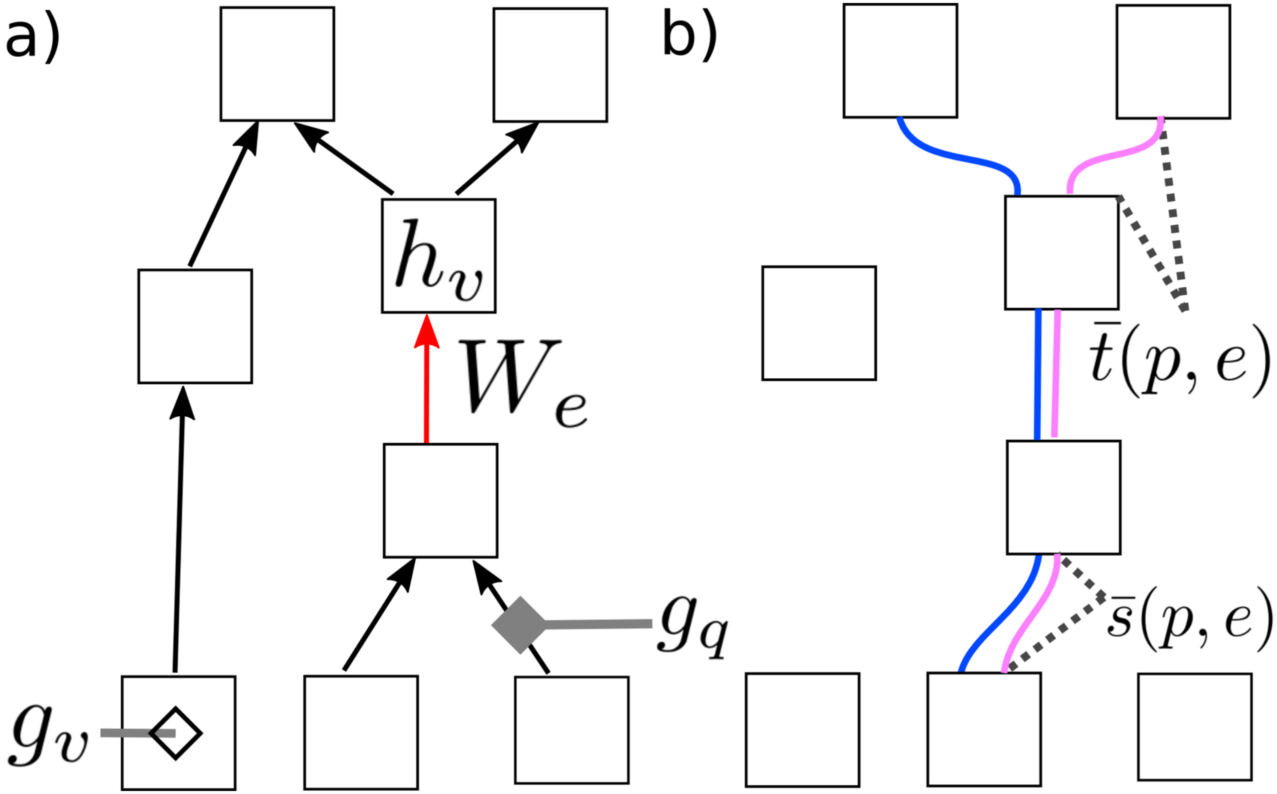
[1] Jacot, Arthur, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks." *Advances in neural information processing systems* 31 (2018).

[2] Goldt, Sebastian, et al. "Modeling the influence of data structure on learning in neural networks: The hidden manifold model." *Physical Review X* 10.4 (2020): 041044.

[3] Saxe, Andrew, Shagun Sodhani, and Sam Jay Lewallen. "The neural race reduction: Dynamics of abstraction in gated networks." *International Conference on Machine Learning*. PMLR, 2022.

Main Idea

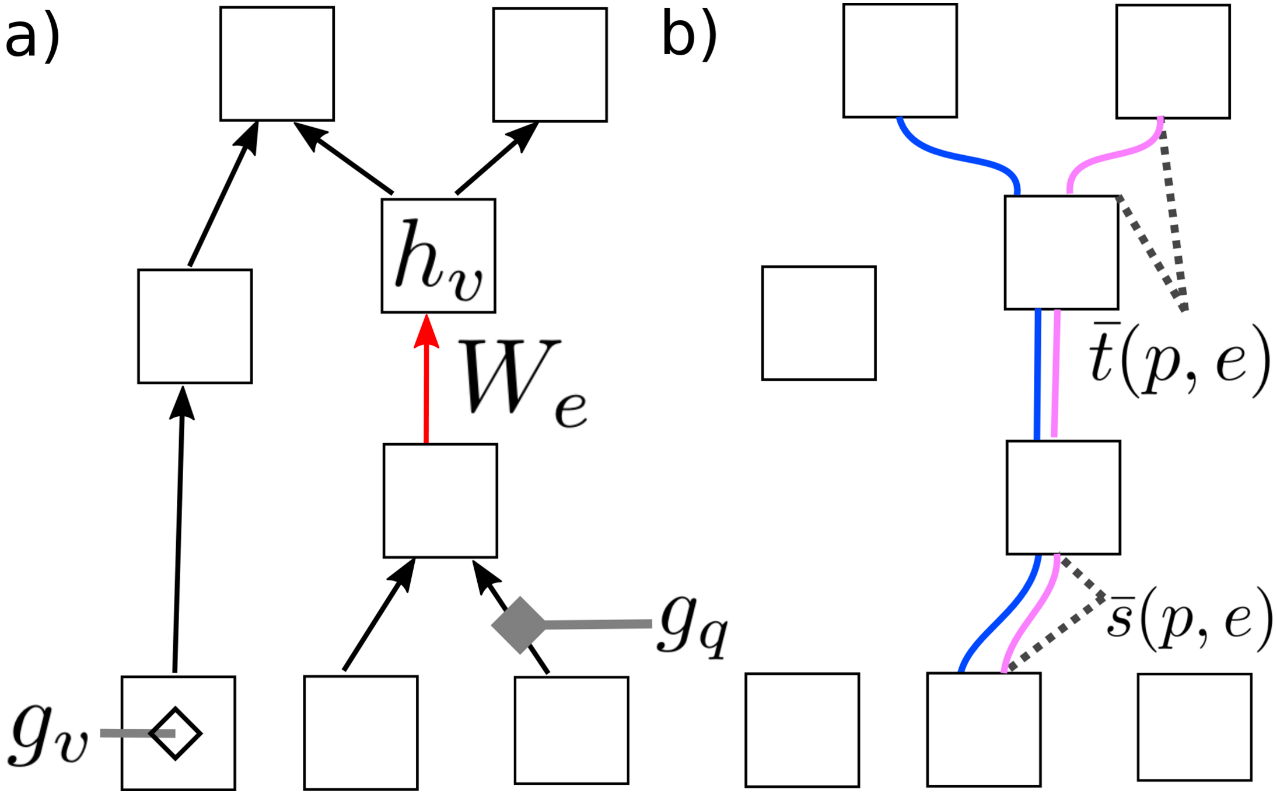
- Represent the ReLU network as a Gated Deep Linear Network [3].



[3] Saxe, Andrew, Shagun Sodhani, and Sam Jay Lewallen. "The neural race reduction: Dynamics of abstraction in gated networks." *International Conference on Machine Learning*. PMLR, 2022.

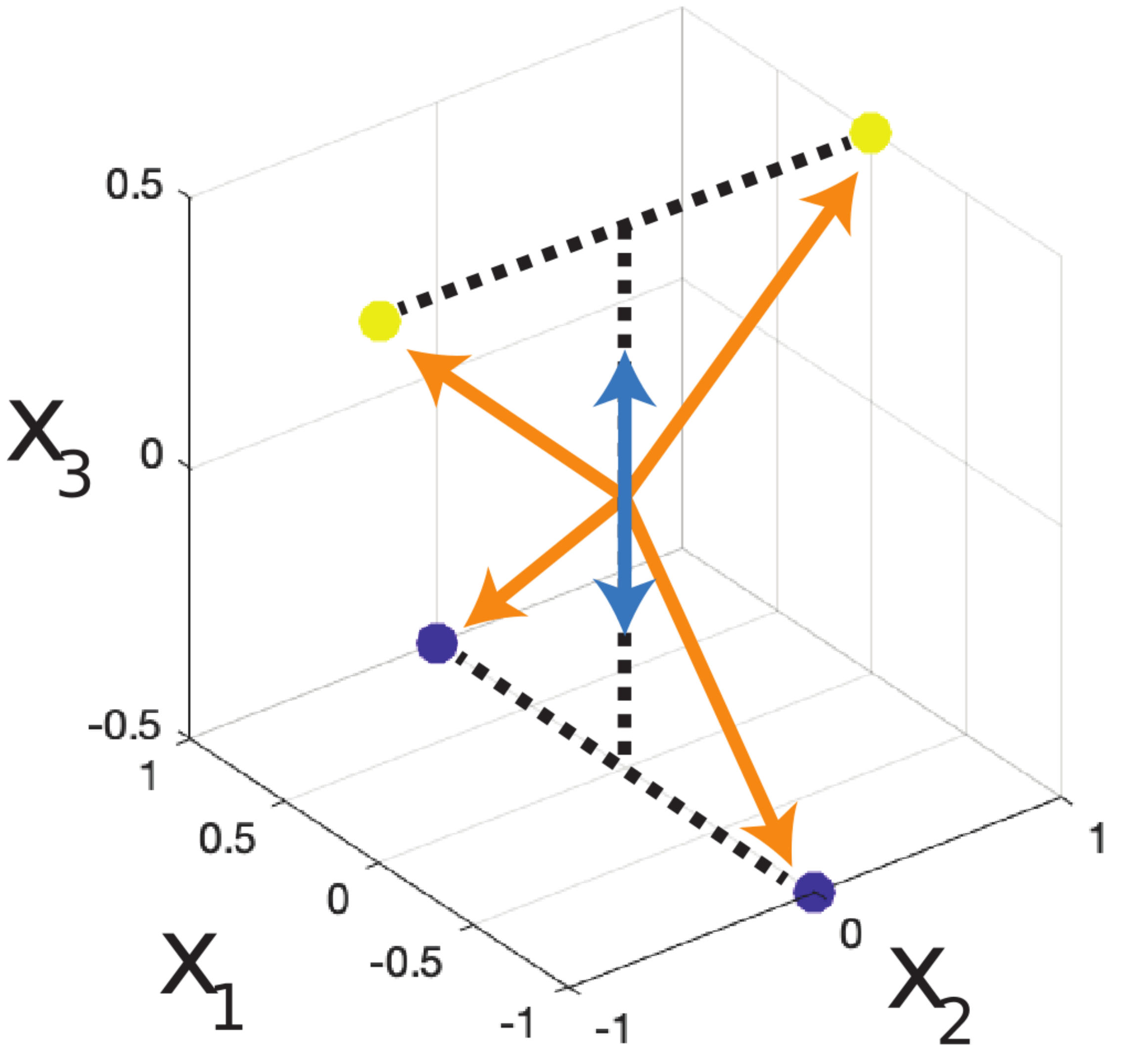
Main Idea

- Represent the ReLU network as a Gated Deep Linear Network [3].
- We prove that a mapping always exists.

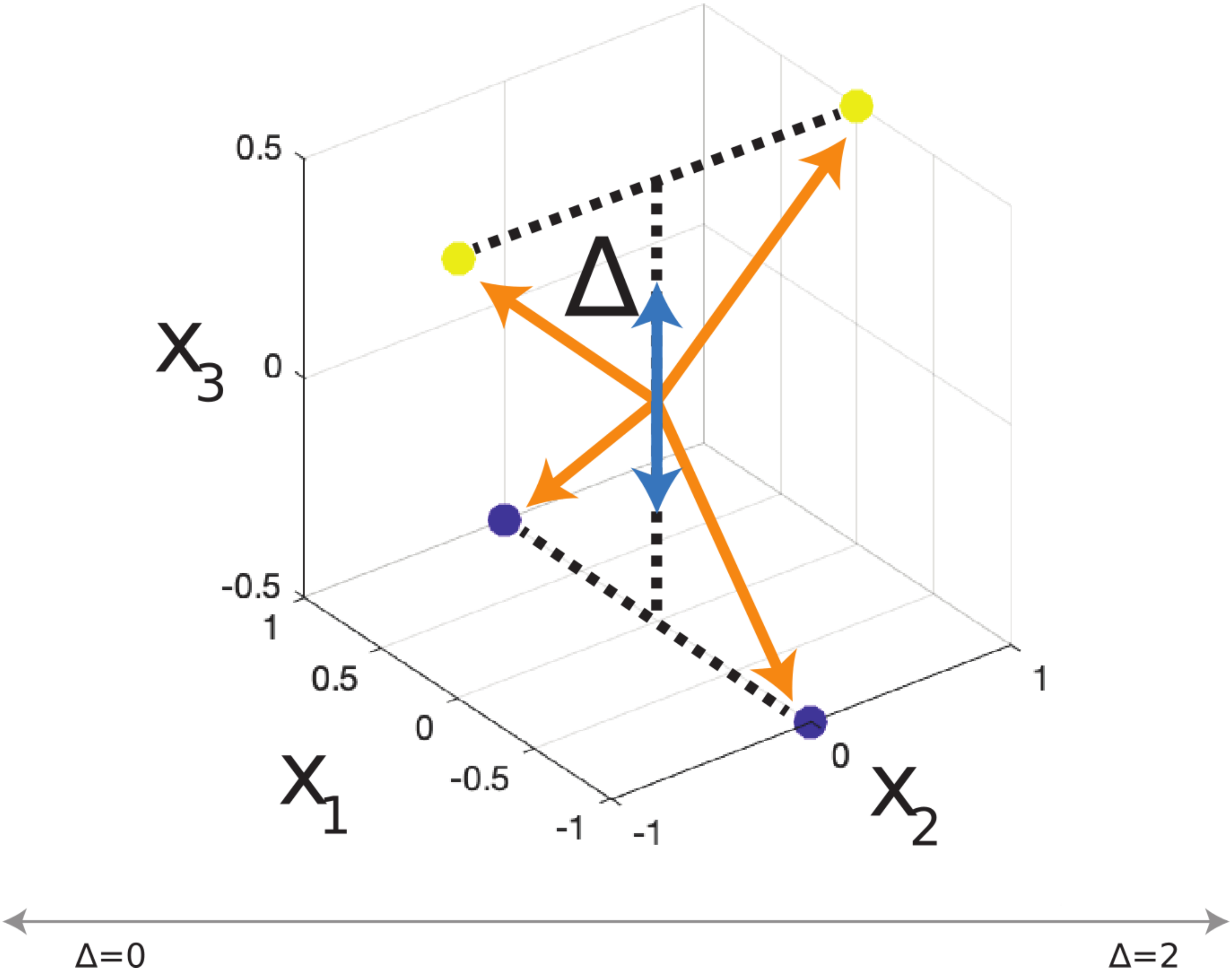


[3] Saxe, Andrew, Shagun Sodhani, and Sam Jay Lewallen. "The neural race reduction: Dynamics of abstraction in gated networks." *International Conference on Machine Learning*. PMLR, 2022.

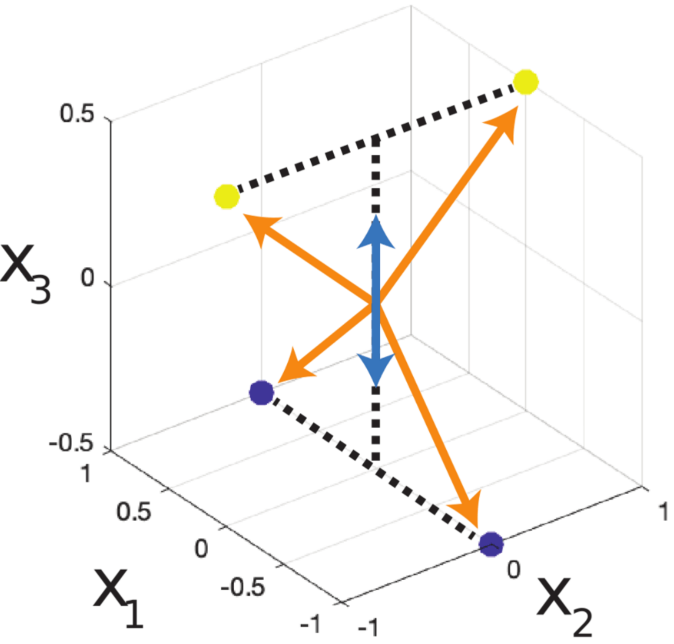
Setting 1: Extended XoR



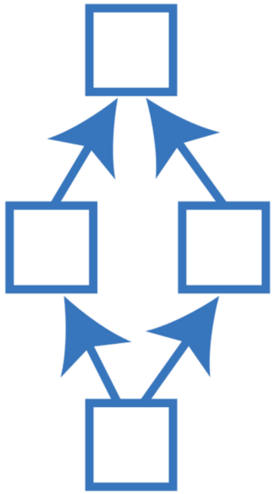
Setting 1: Extended XoR



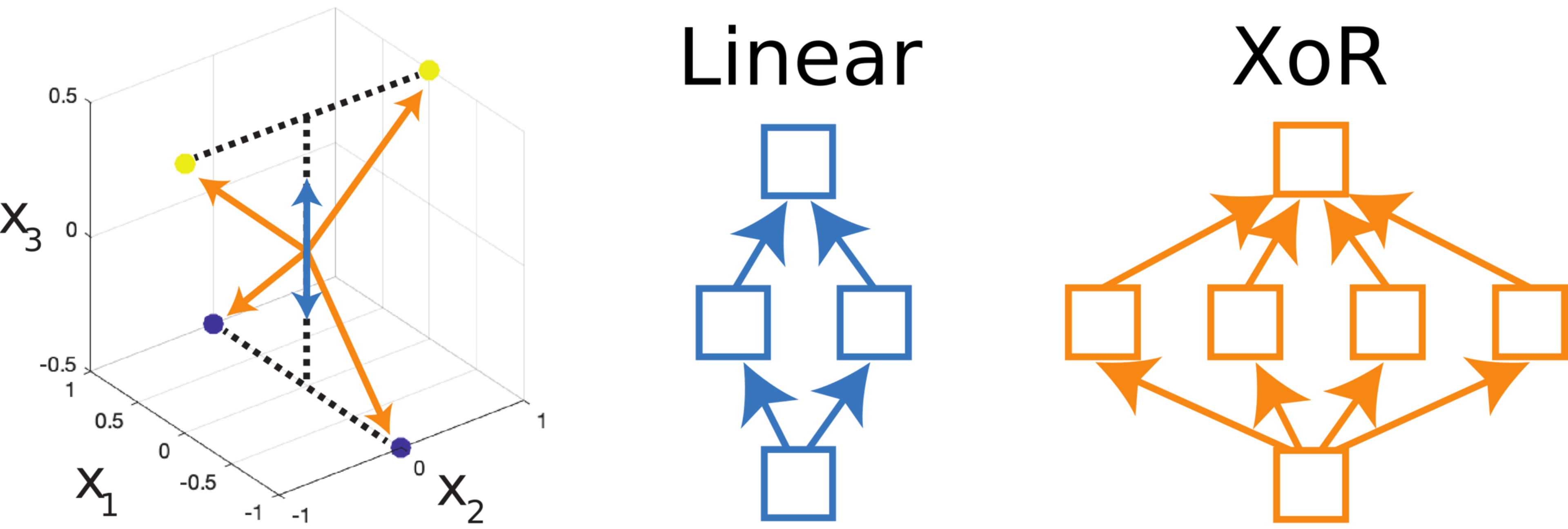
Setting 1: Extended XoR



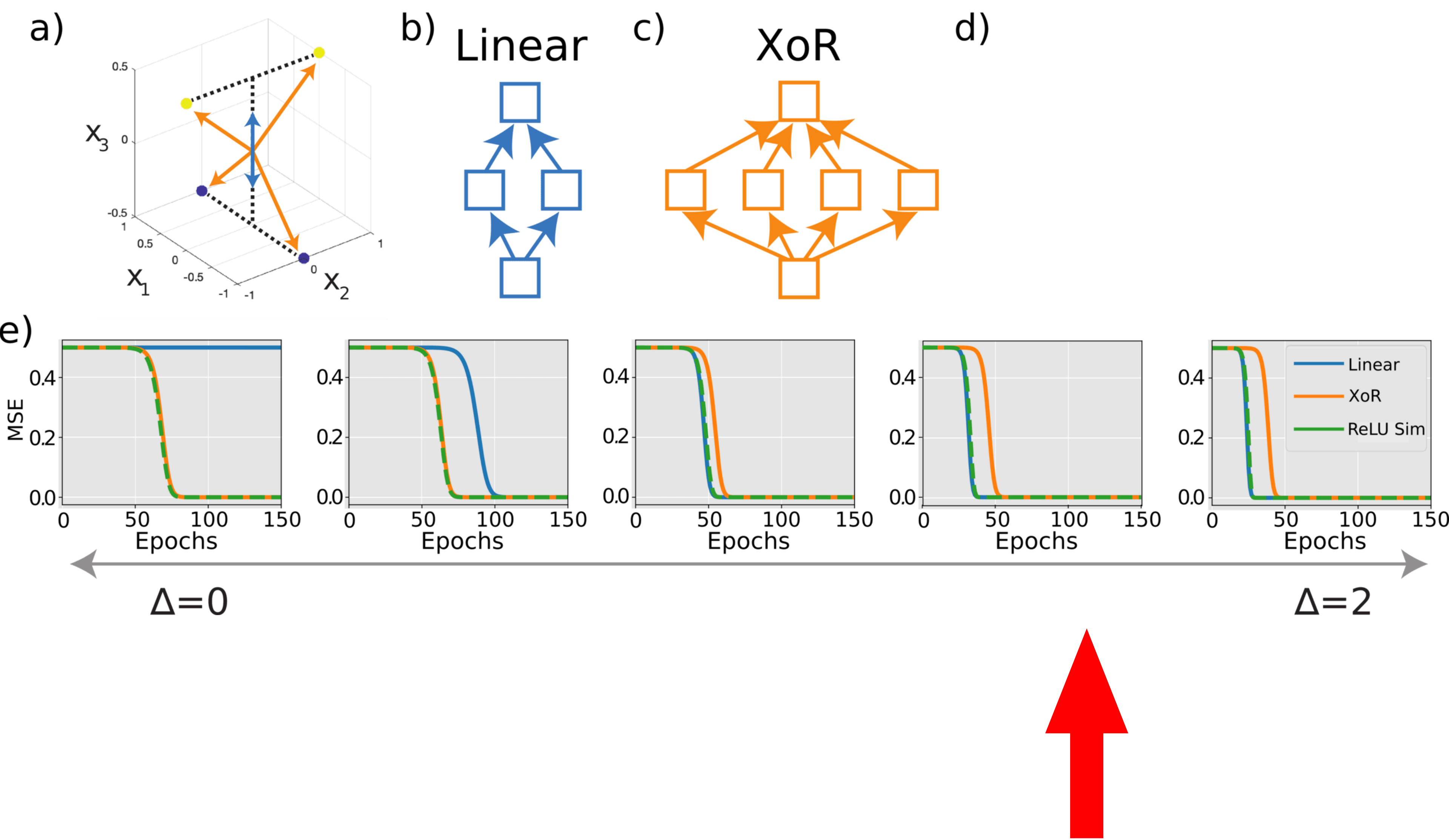
Linear



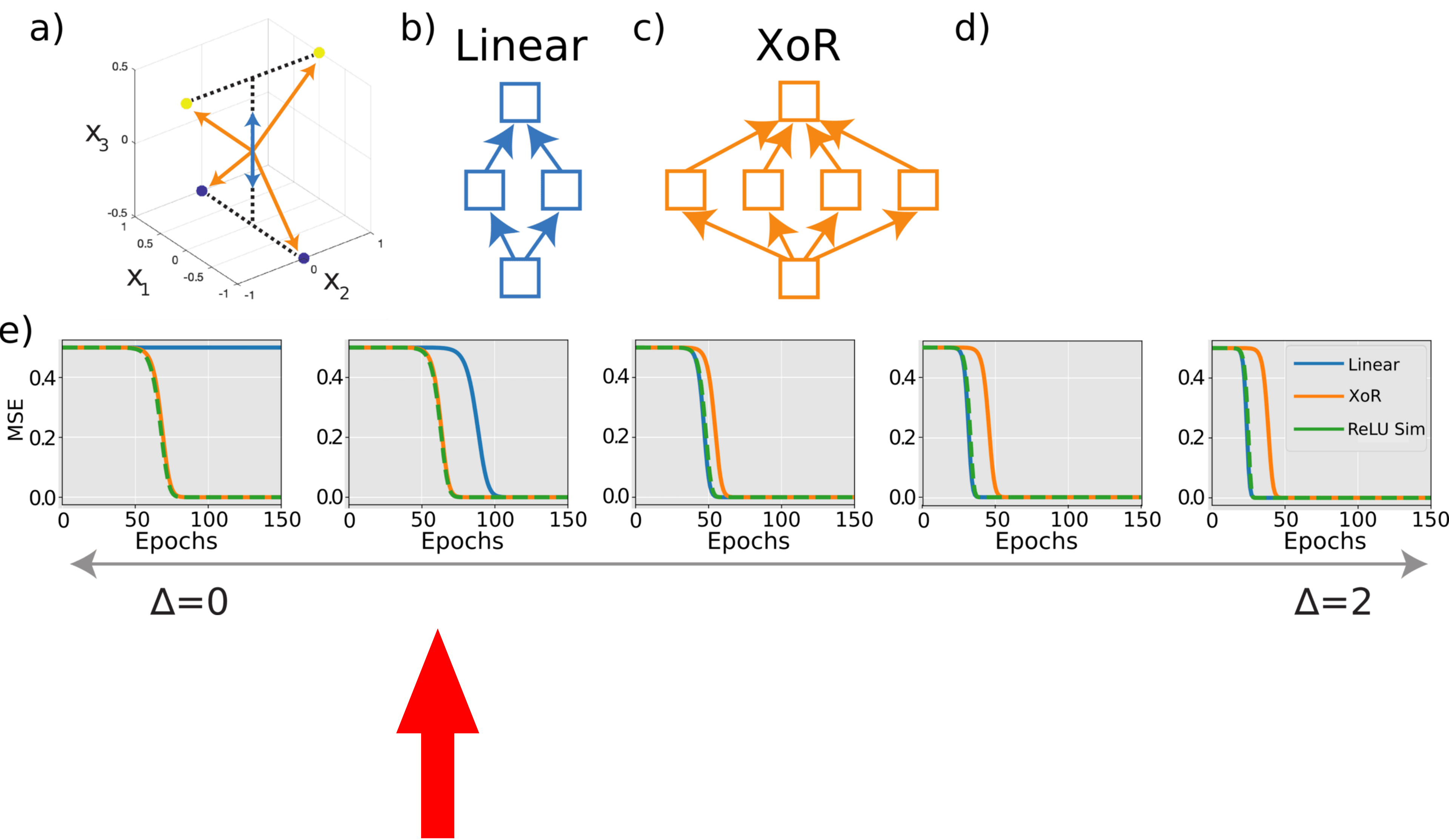
Setting 1: Extended XoR



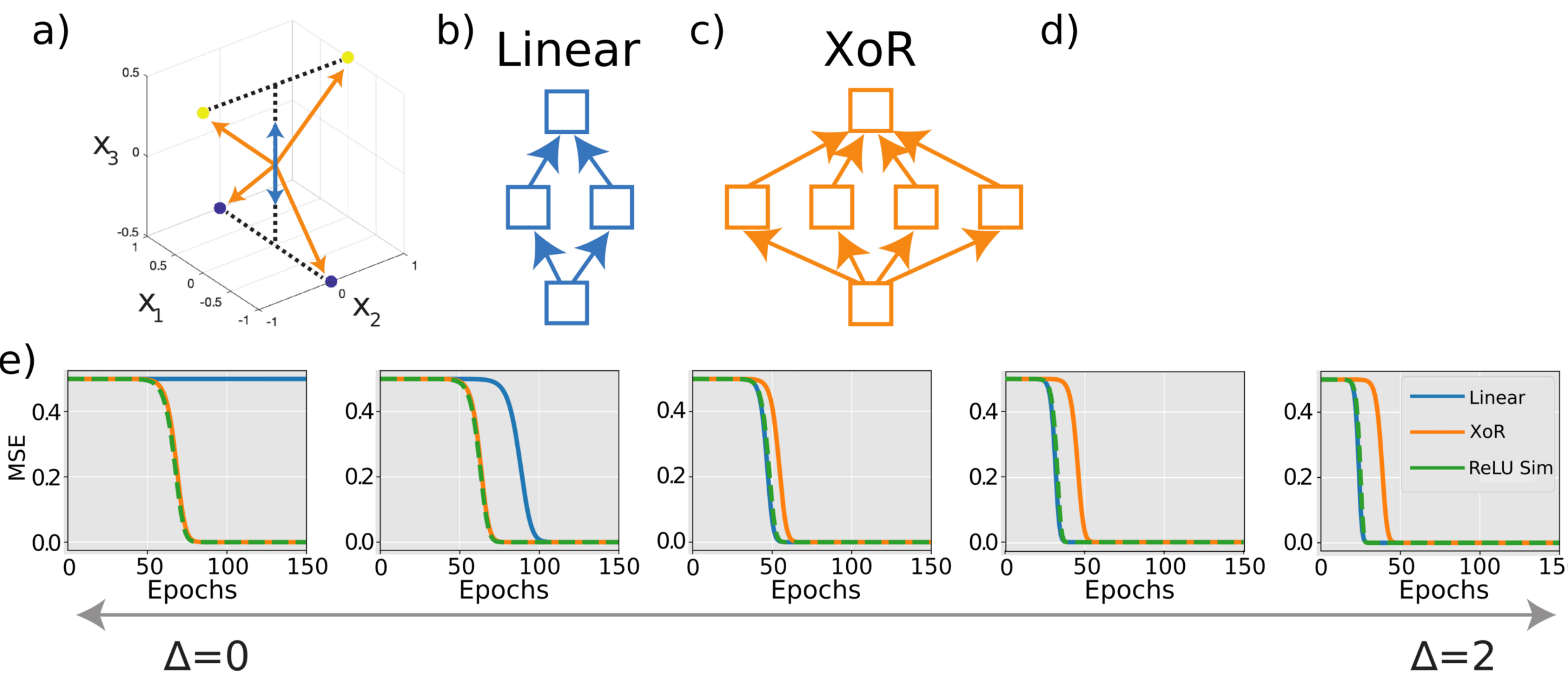
Setting 1: Extended XoR



Setting 1: Extended XoR



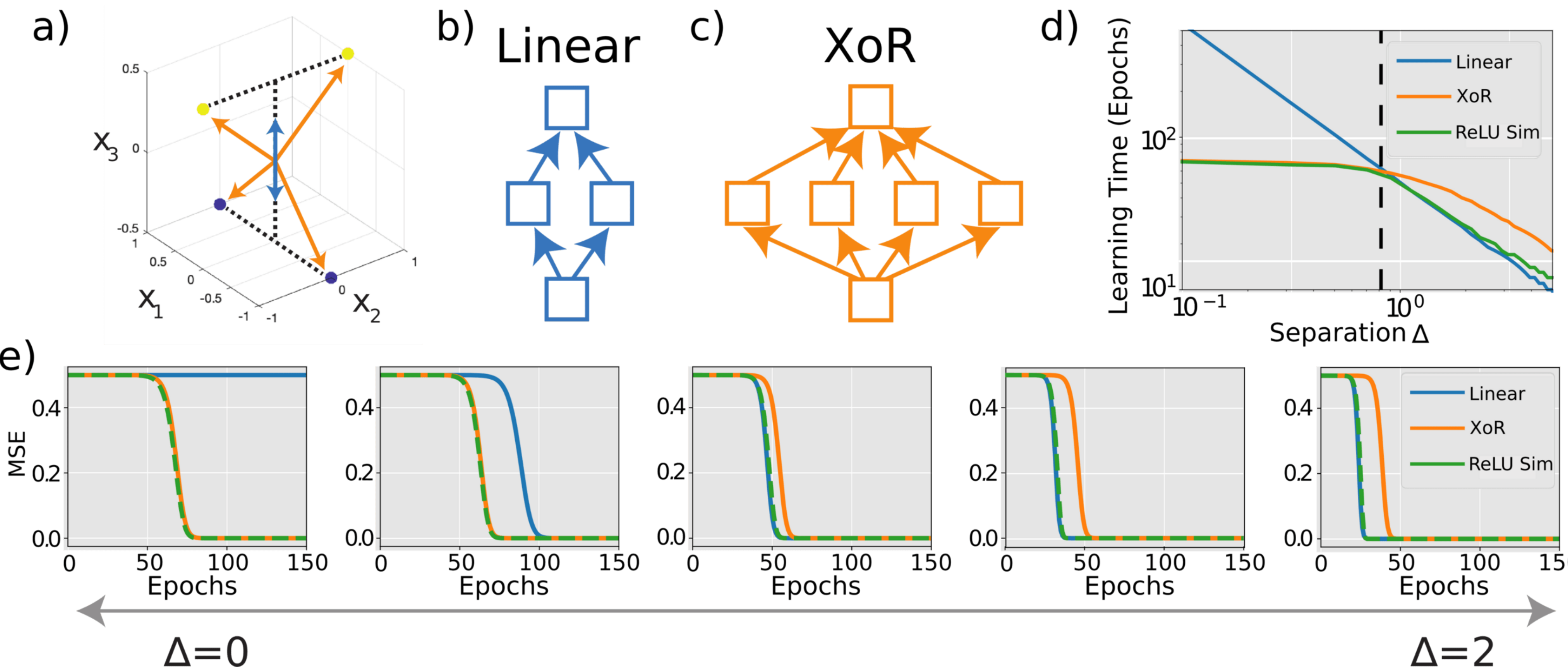
Setting 1: Extended XoR



The GDLN which imitates the ReLU network is called the

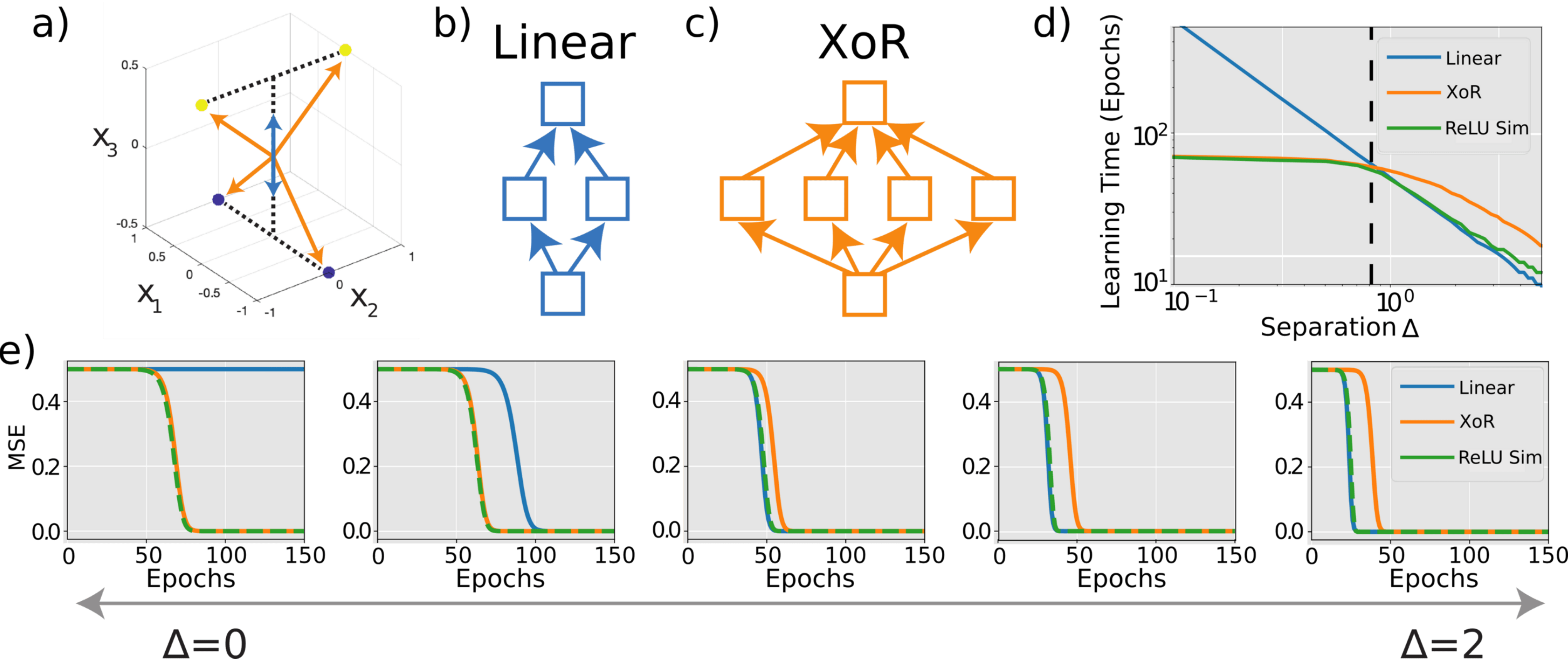
Rectified Linear Network (ReLN)

Setting 1: Extended XoR



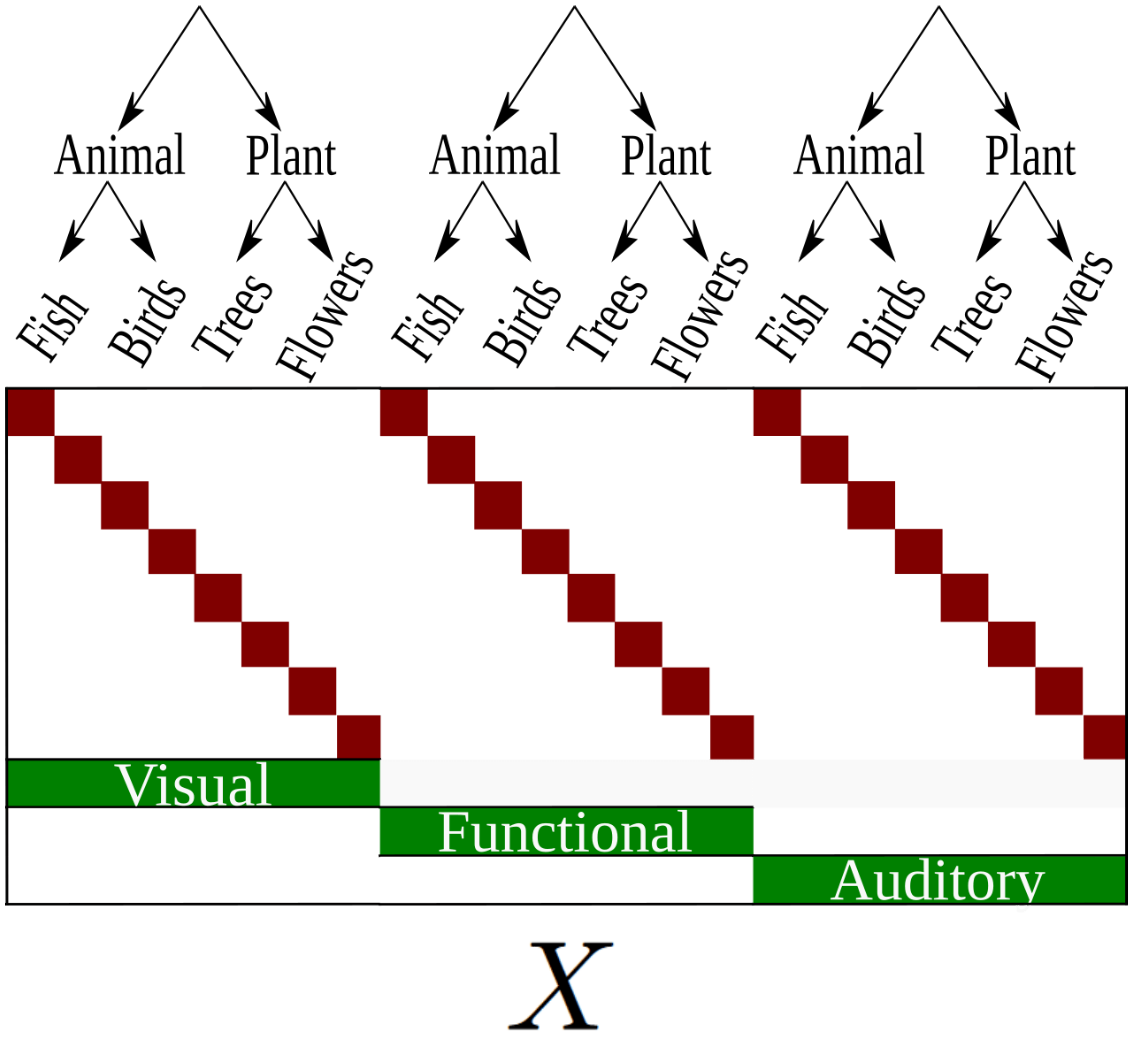
The GDLN which imitates the ReLU network is called the **Rectified Linear Network (ReLN)**

Setting 1: Extended XoR

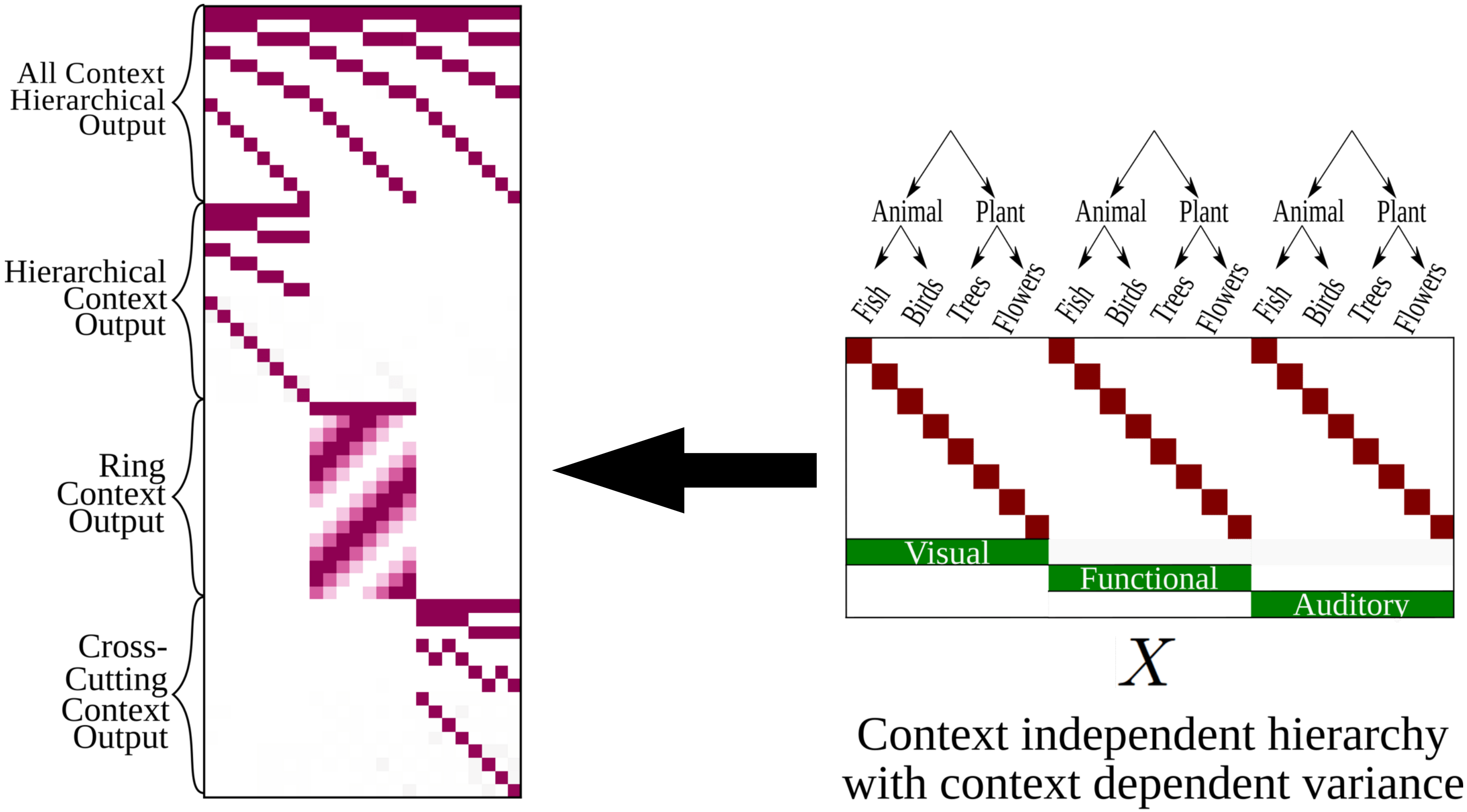


Finding 1: ReLU networks will sometimes choose nonlinear solution even when a linear option is possible but always favours the fastest learner.

Setting 2: Contextual Task

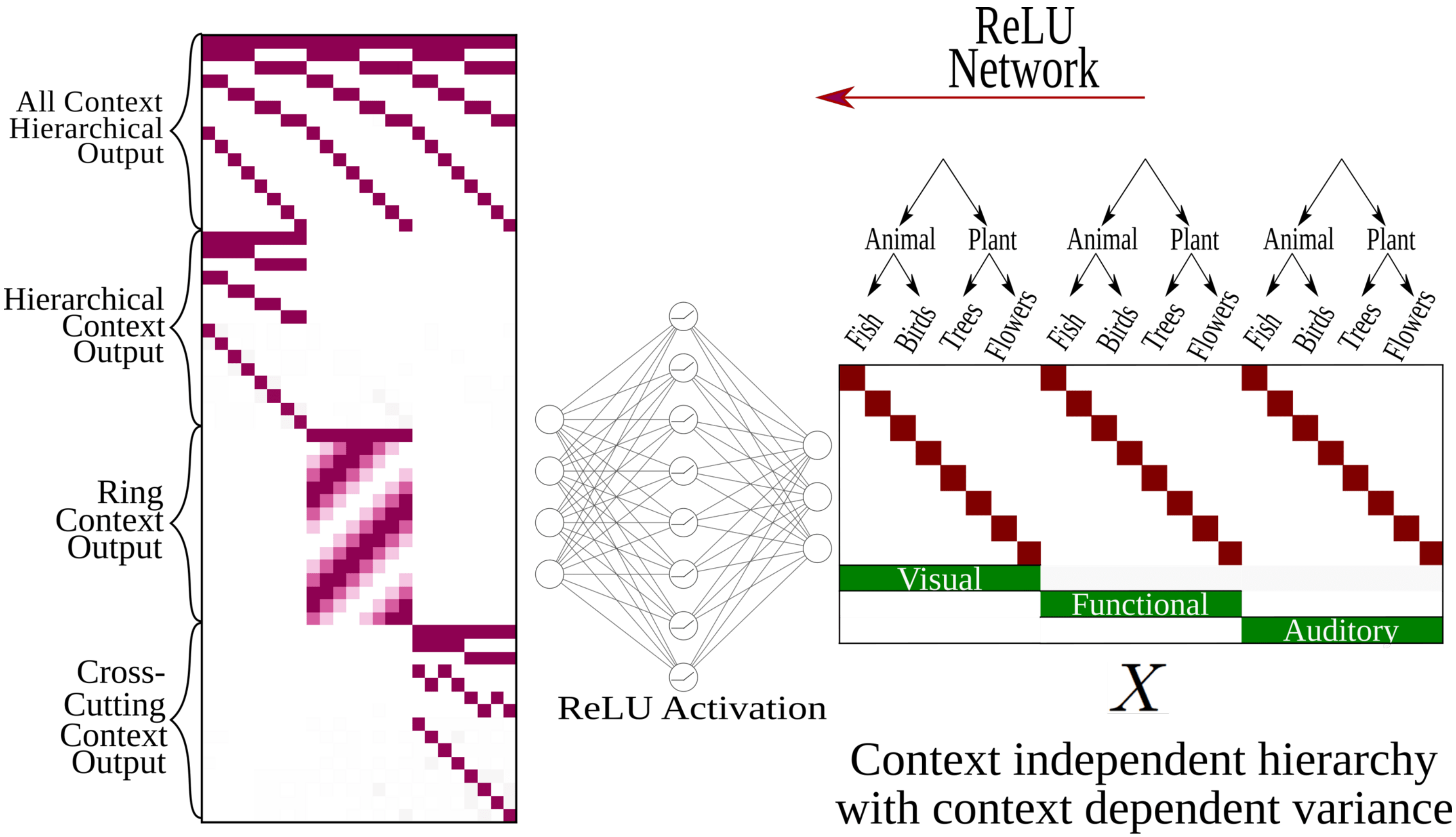


Setting 2: Contextual Task



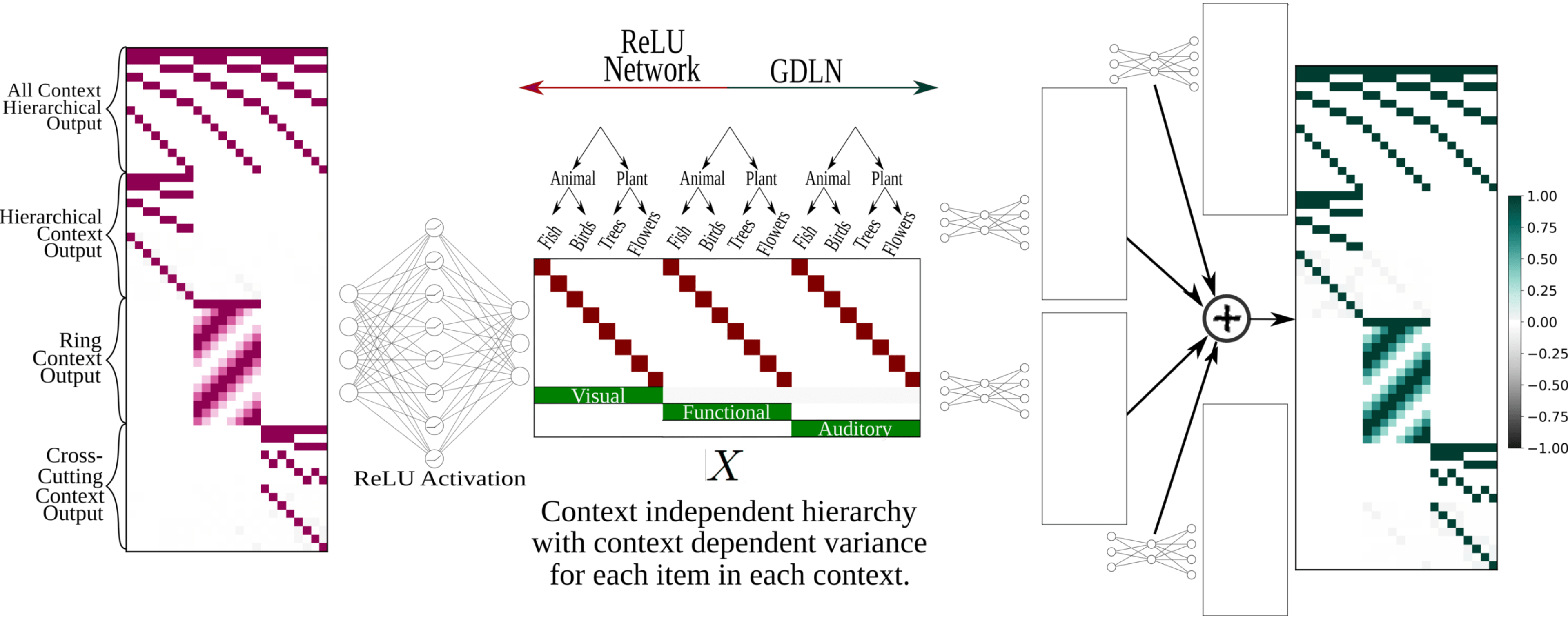
Context independent hierarchy with context dependent variance for each item in each context.

Setting 2: Contextual Task

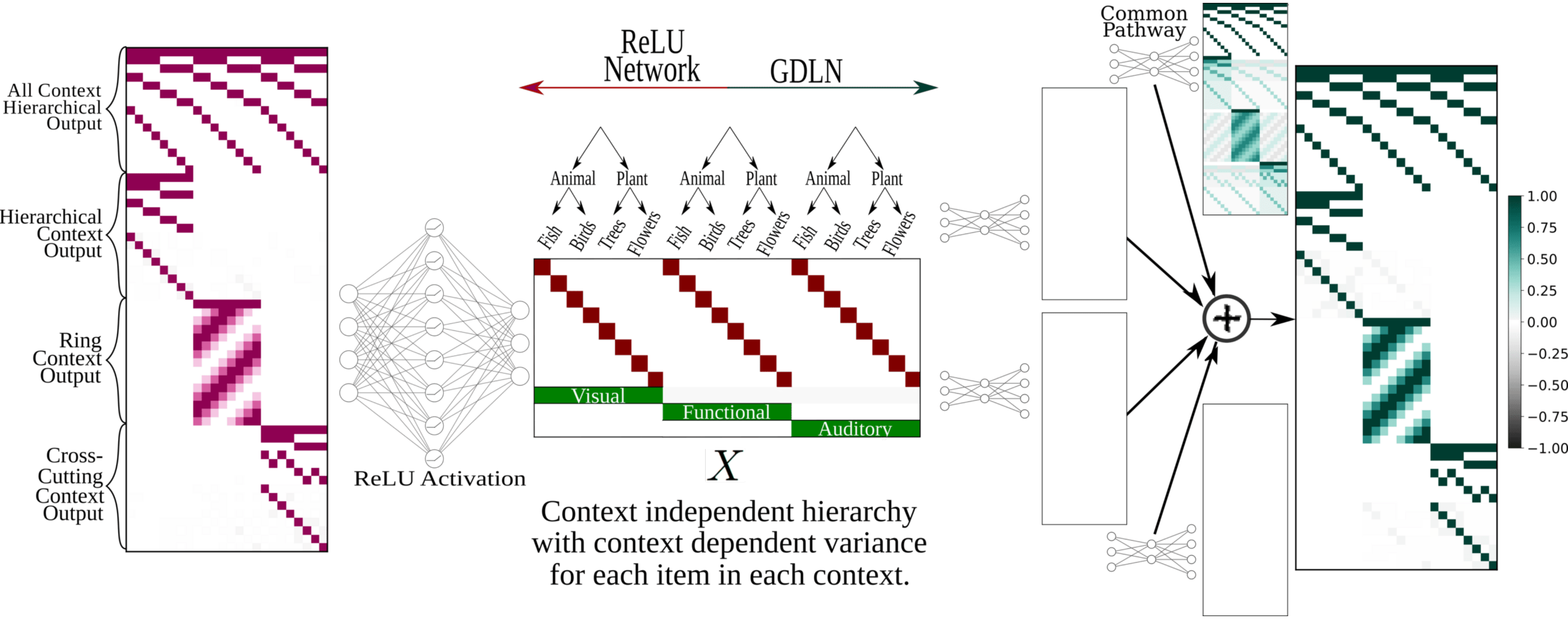


Context independent hierarchy with context dependent variance for each item in each context.

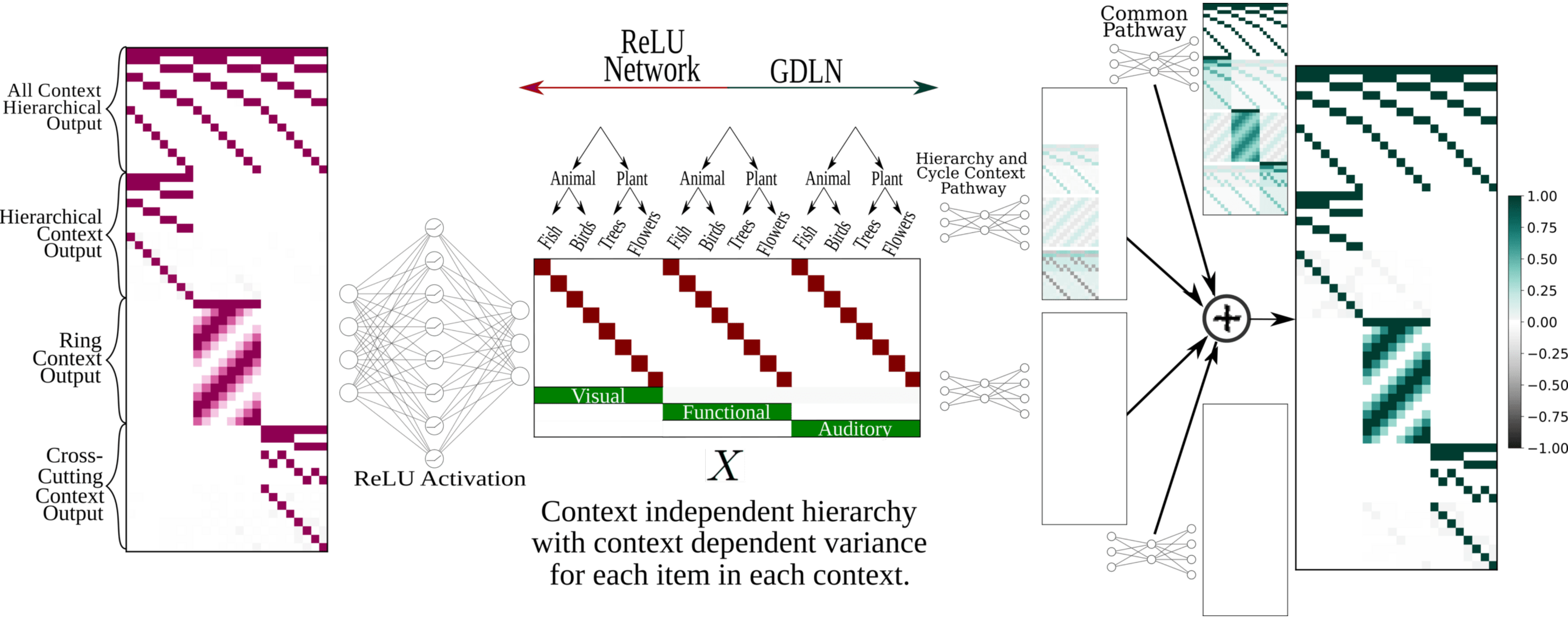
Setting 2: Contextual Task



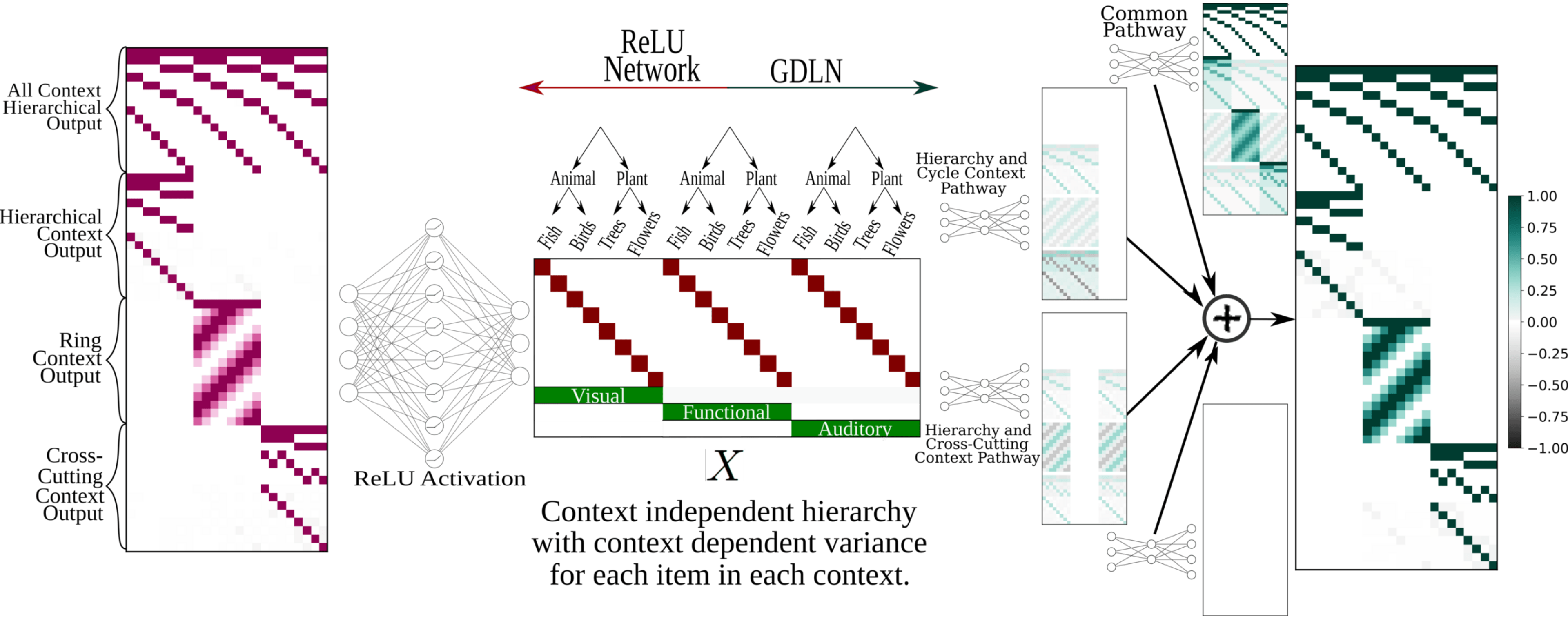
Setting 2: Contextual Task



Setting 2: Contextual Task

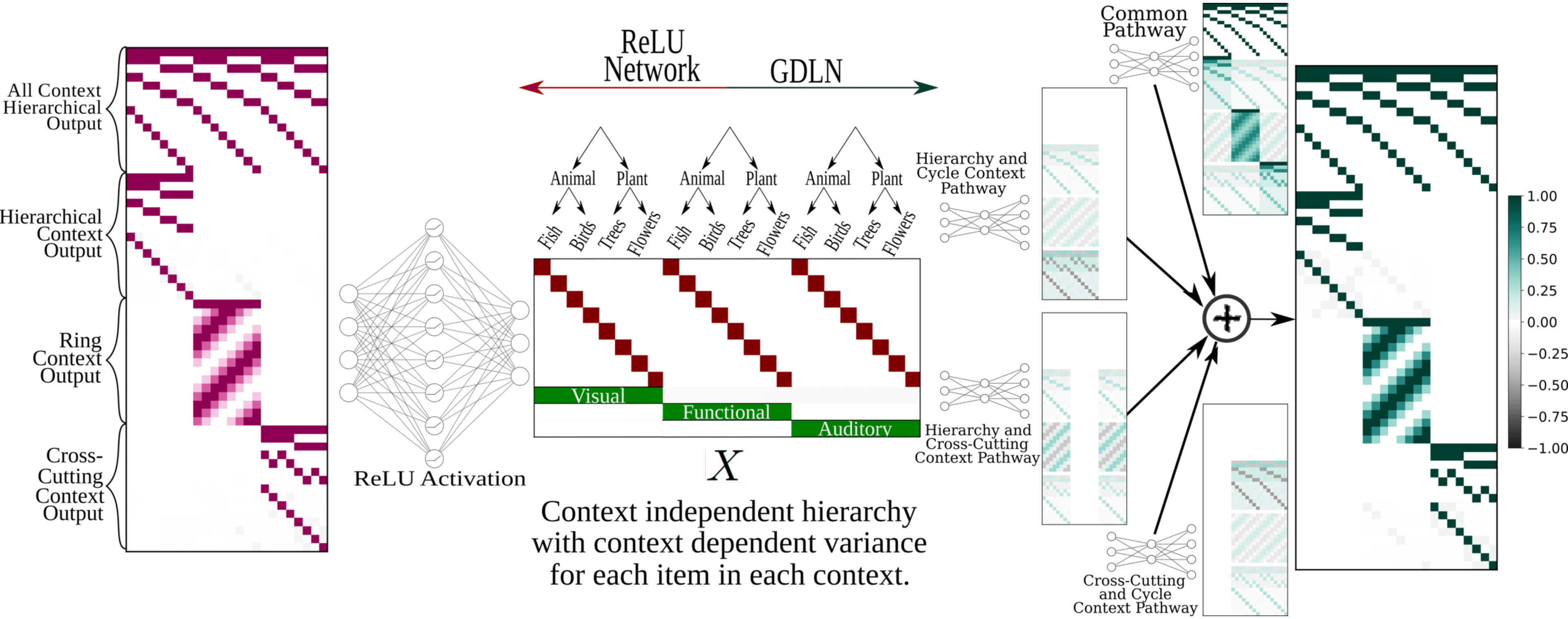


Setting 2: Contextual Task

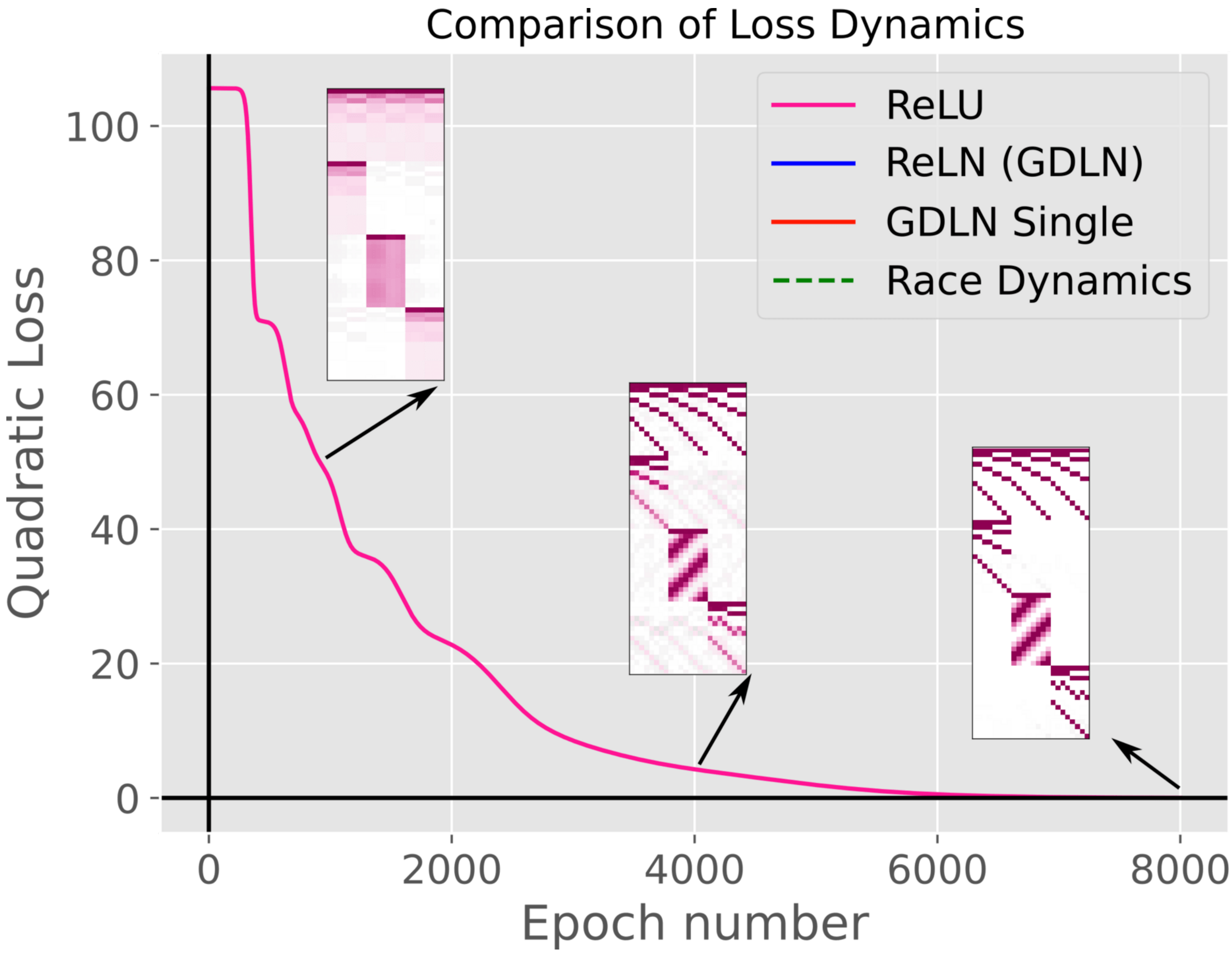


Context independent hierarchy with context dependent variance for each item in each context.

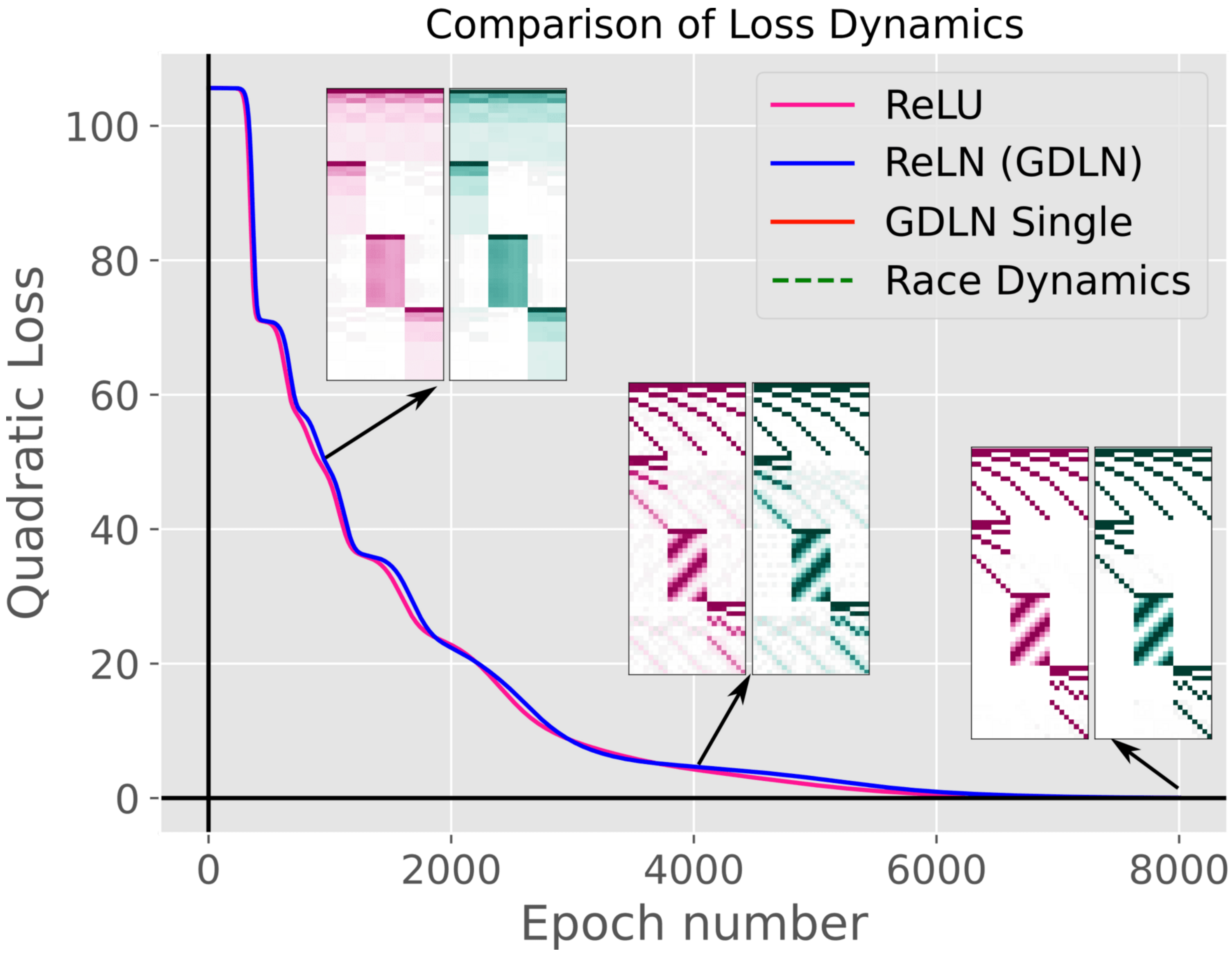
Setting 2: Contextual Task



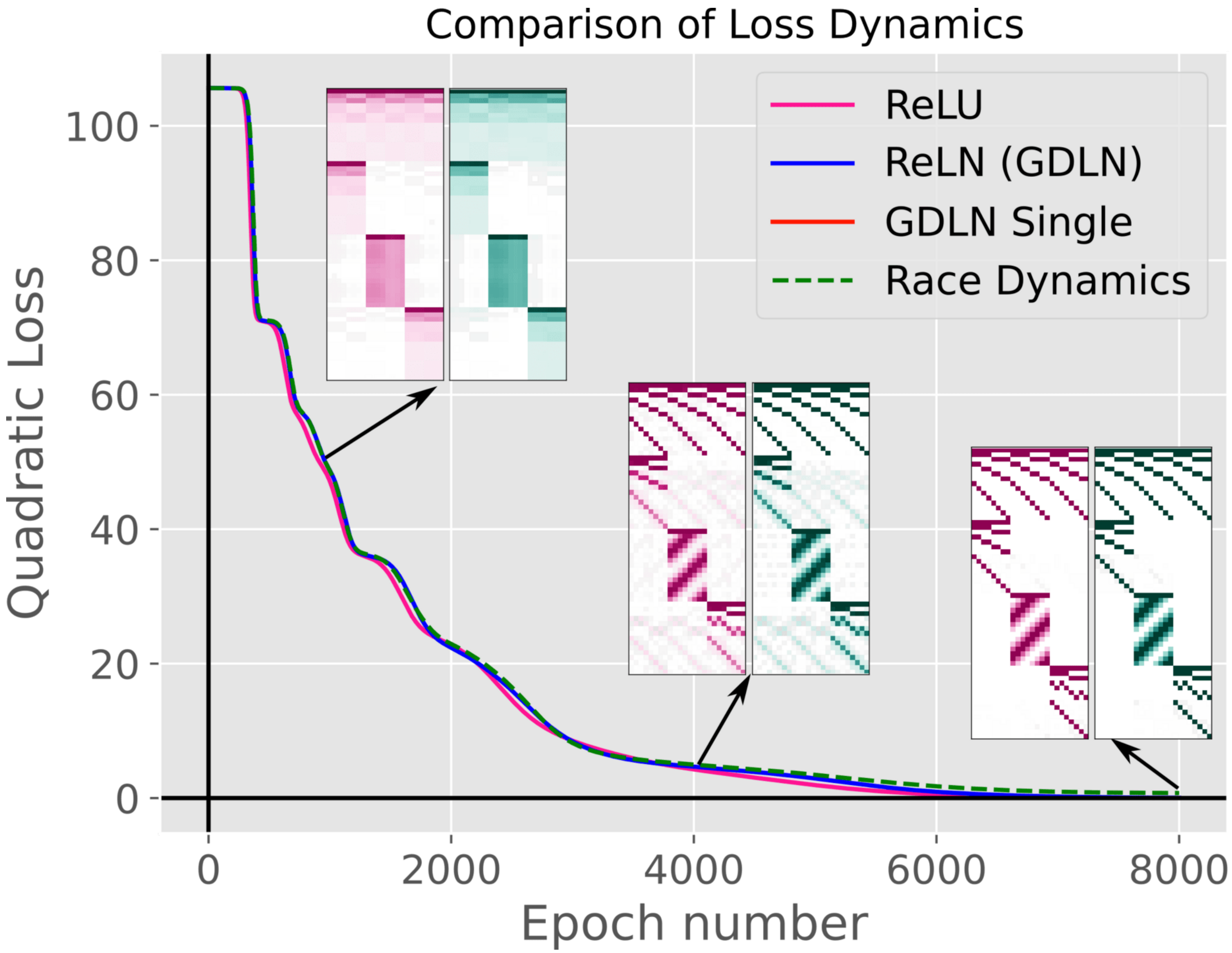
Setting 2: Contextual Task



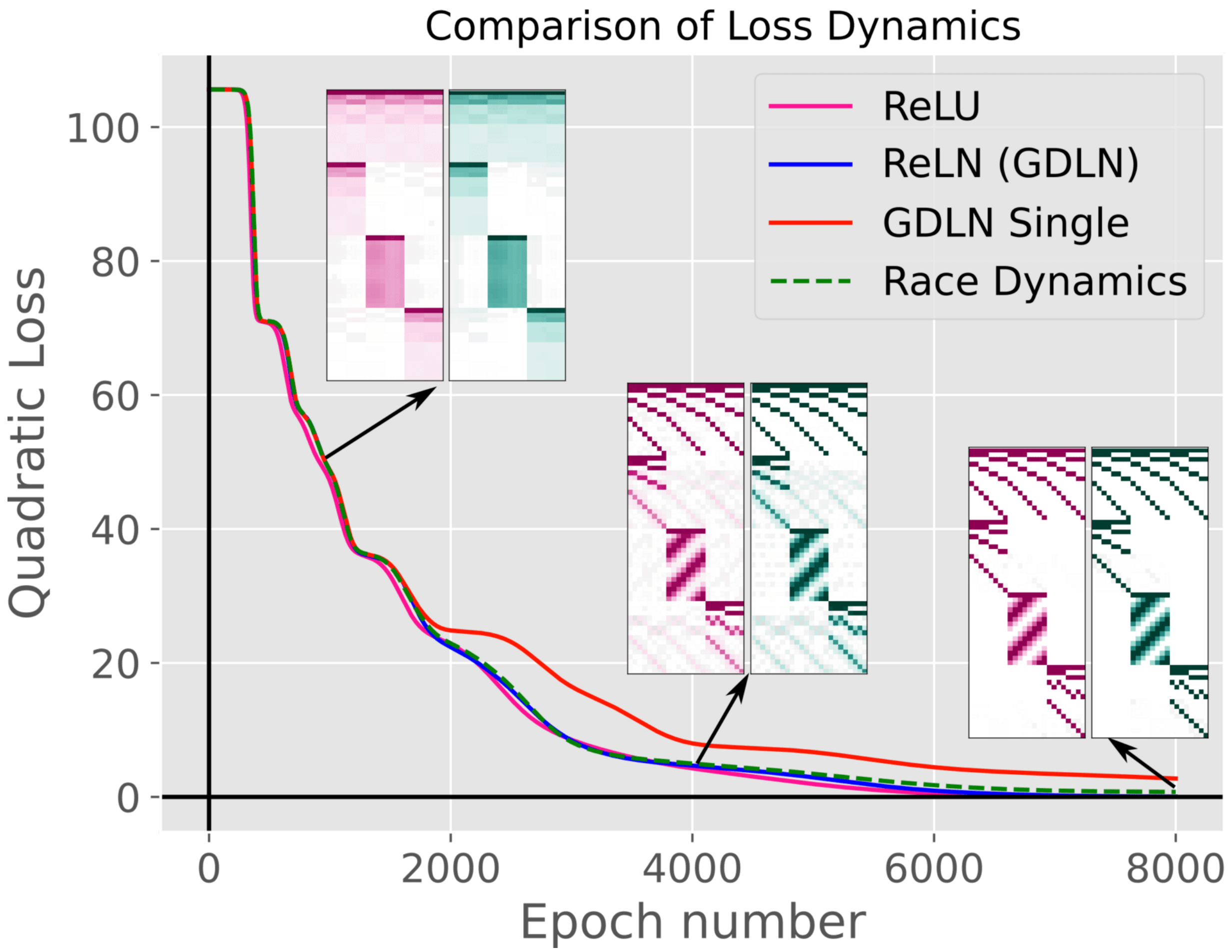
Setting 2: Contextual Task



Setting 2: Contextual Task

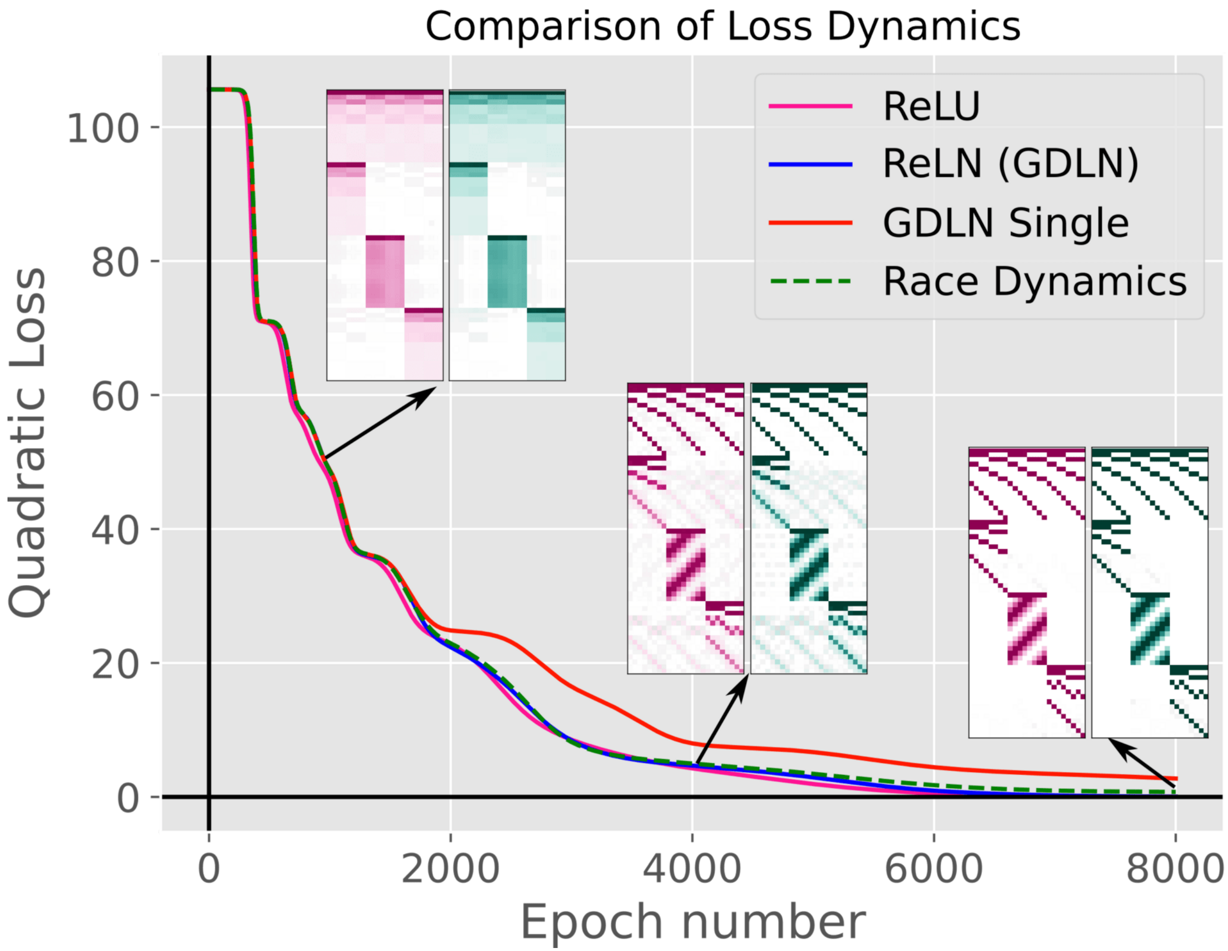


Setting 2: Contextual Task



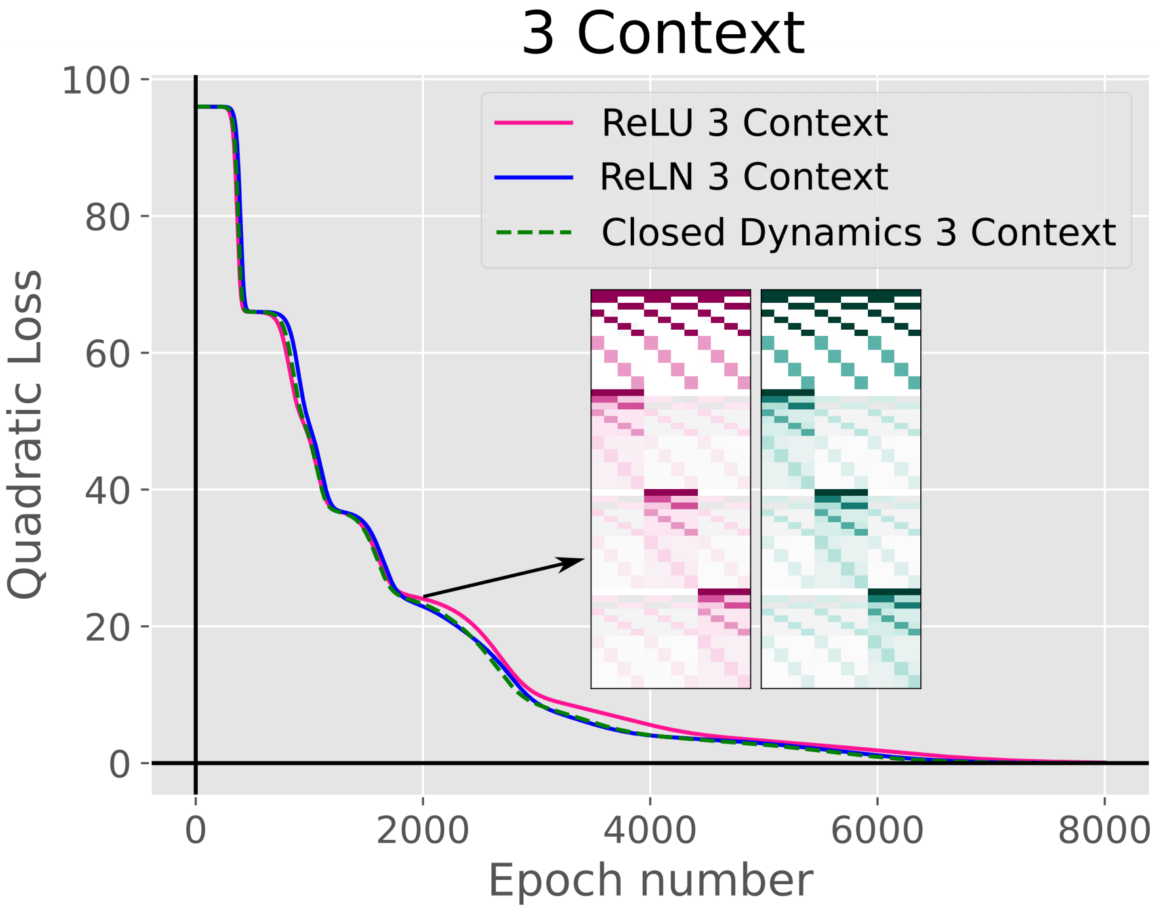
We **prove** the **uniqueness** of the **ReLN** we find in this setting.

Setting 2: Contextual Task

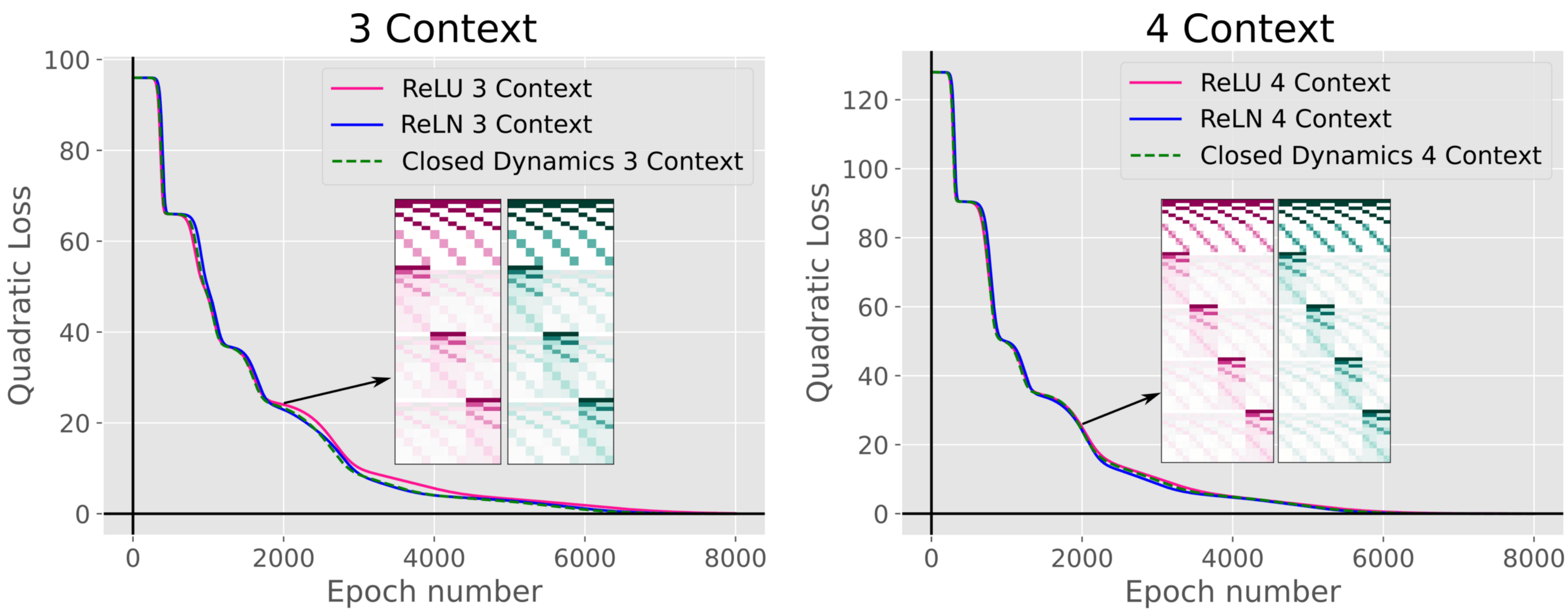


Finding 2: ReLU networks in this setting have a preference towards structured mixed-selectivity due to the learning speed boost.

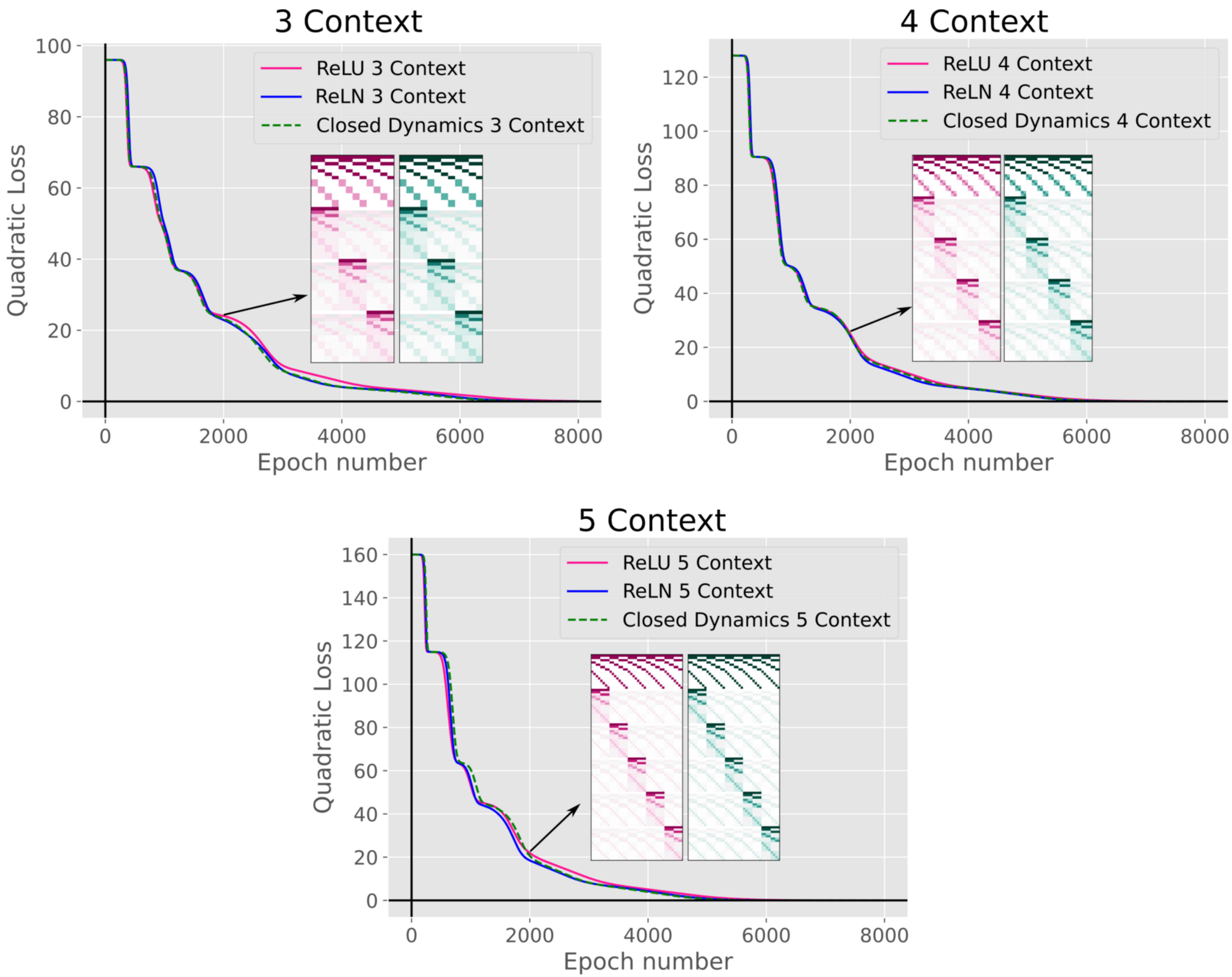
Setting 3: More Contexts



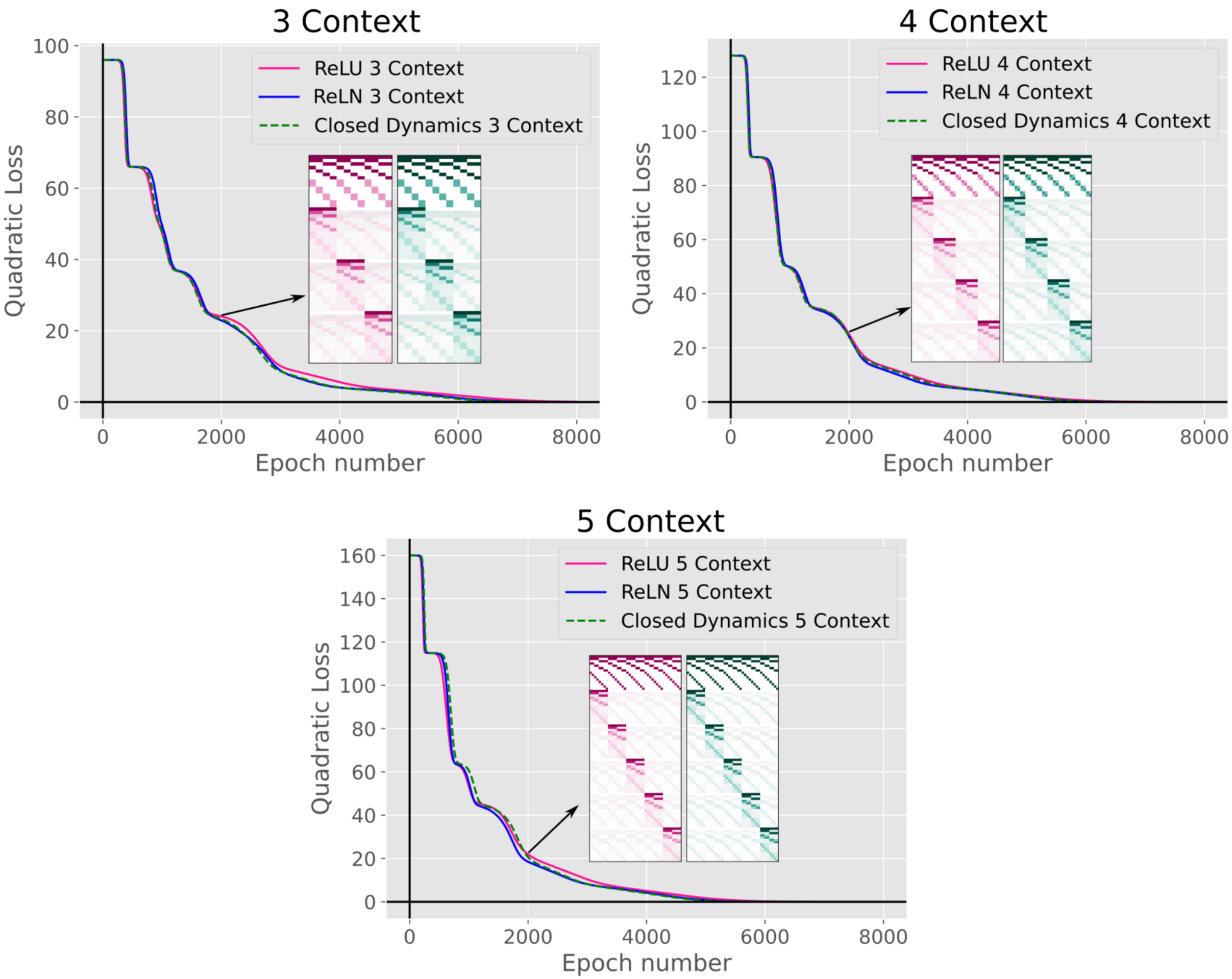
Setting 3: More Contexts



Setting 3: More Contexts

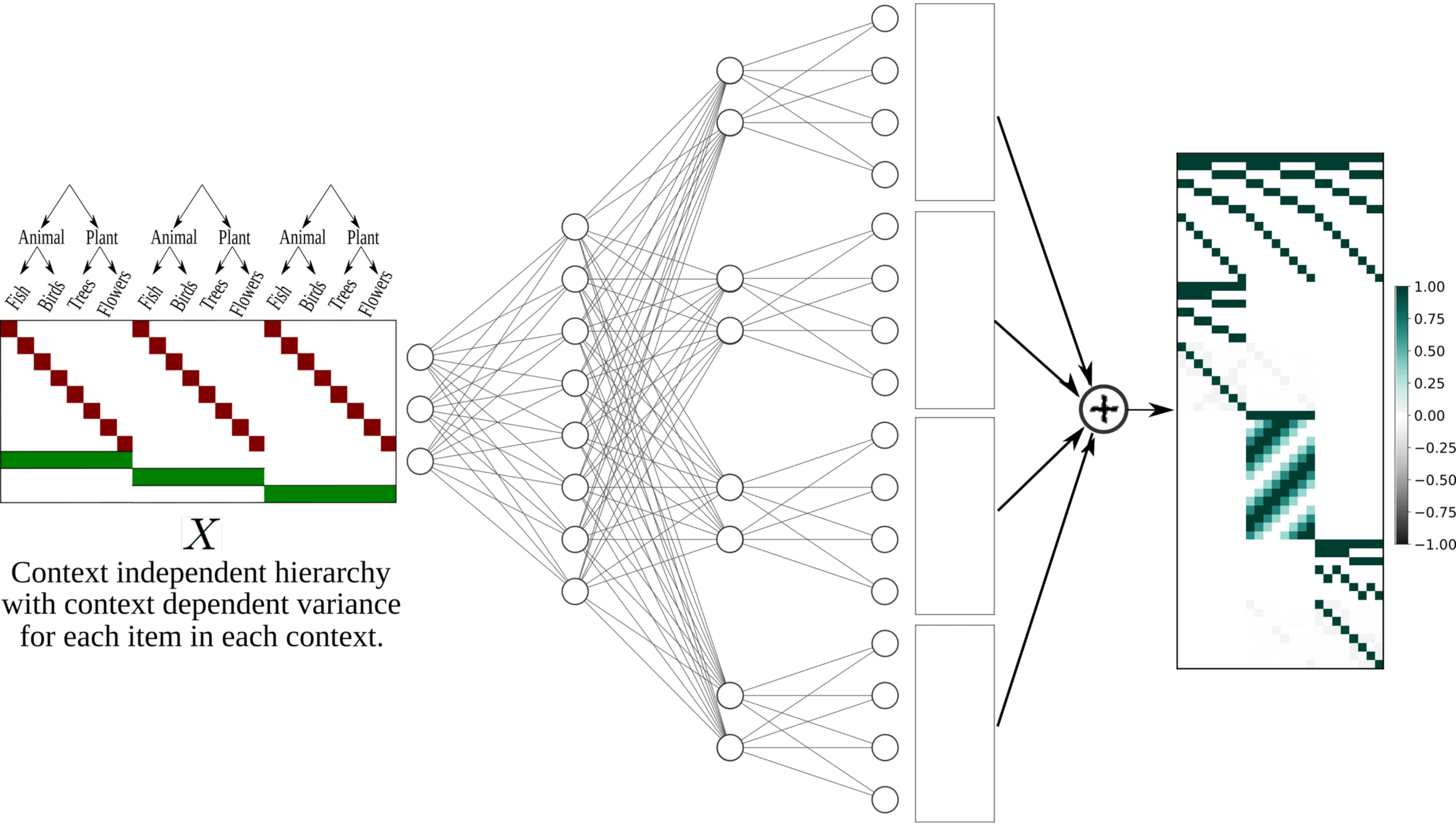


Setting 3: More Contexts



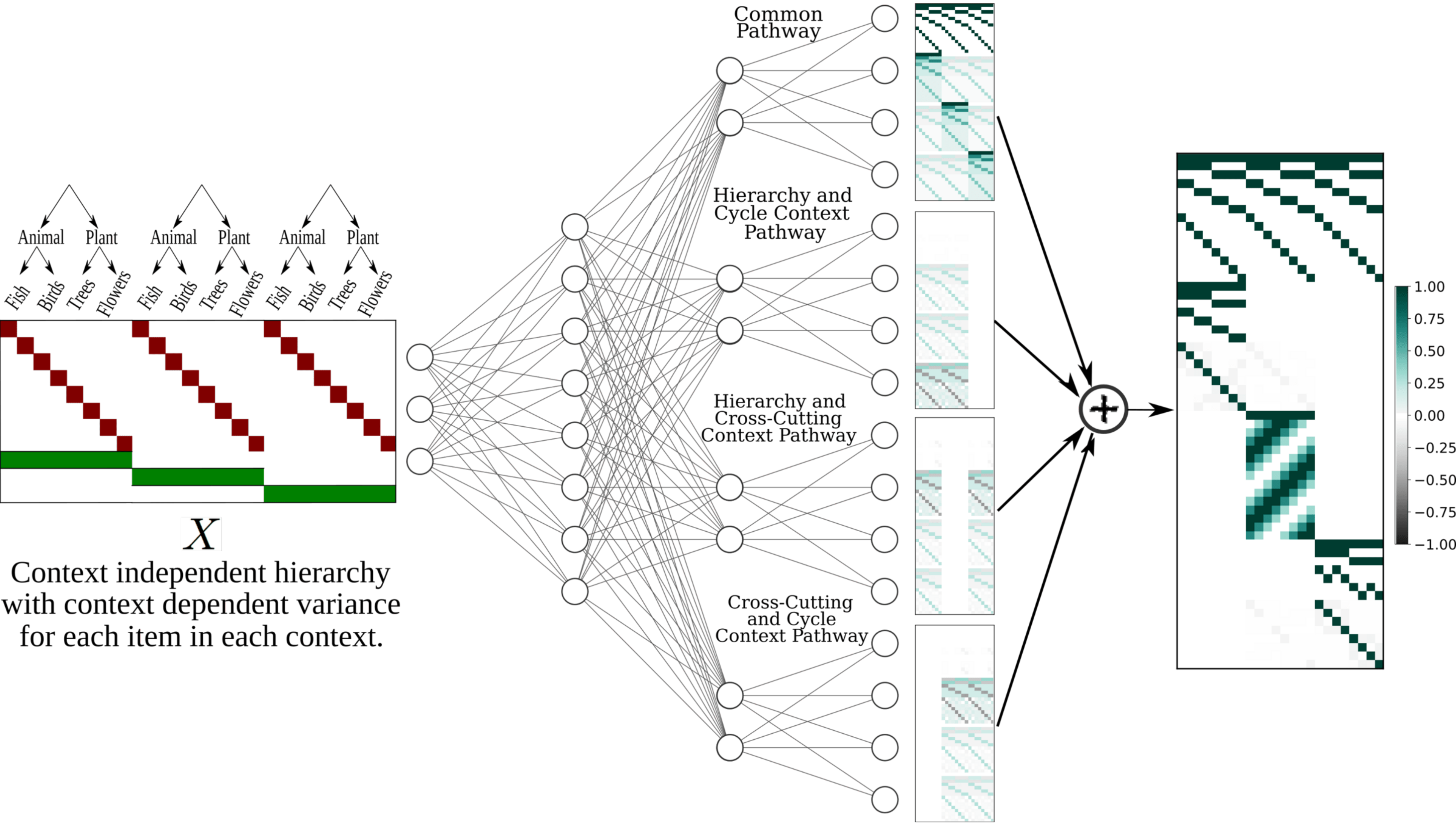
Finding 3: ReLU networks still prefer structured mixed-selective representations as the number of contexts grows.

Setting 4: Depth

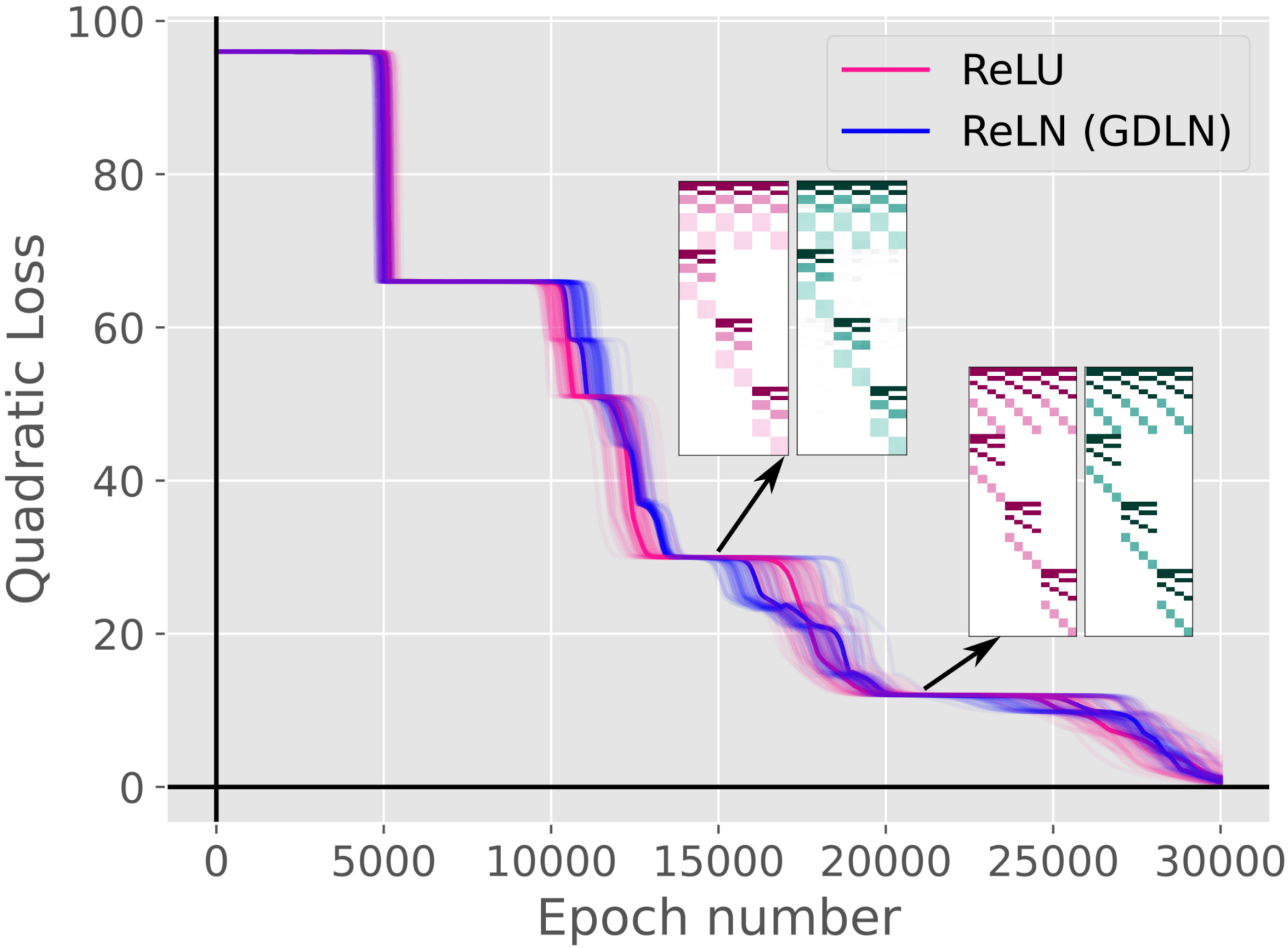


Context independent hierarchy with context dependent variance for each item in each context.

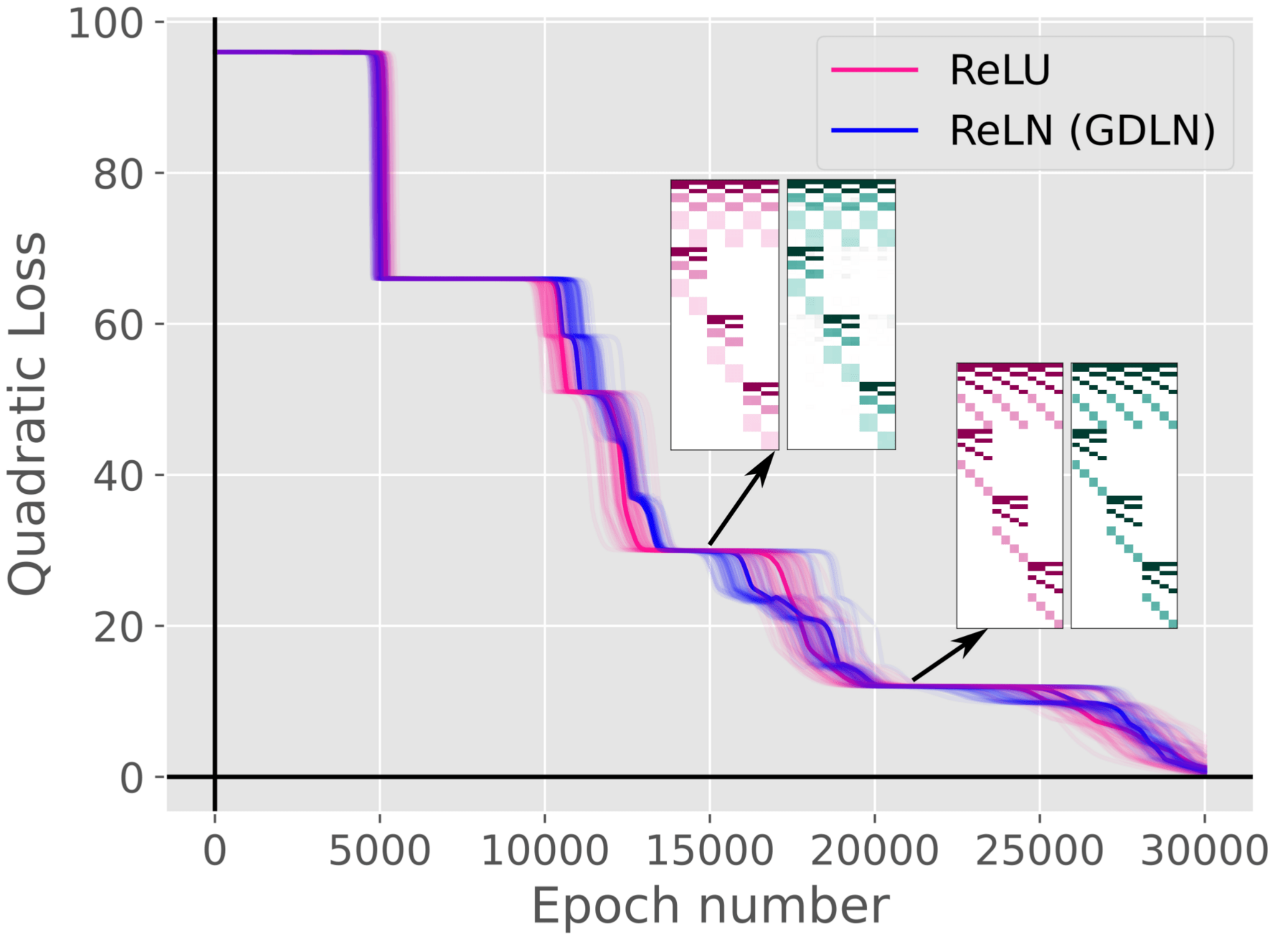
Setting 4: Depth



Setting 4: Depth



Setting 4: Depth



Finding 4: Additional hidden layers makes the network dynamics inconsistent. We can still design a GDLN which fits the distribution of dynamics.

Make Haste Slowly: A Theory of Emergent Structured Mixed Selectivity in Feature Learning ReLU Networks

Devon Jarvis, Richard Klein, Benjamin Rosman & Andrew Saxe

ICLR 2025

WITS UNIVERSITY

MIND
Machine Intelligence and
Neural Discovery Institute

UCL

Sainsbury Wellcome Centre

GATSBY

Finding the Gates

Algorithm 1 *A preliminary algorithm for finding a ReLN*. This follows a simple K-means clustering algorithm, but with samples taken throughout training such that it is easier to identify pathways through the network as they emerge.

Require: $num\_trainings > 0, num\_epochs > 0, \sigma > 0$,($X \in \mathbb{R}^{d \times N}, Y \in \mathbb{R}^{p \times N}$) (the dataset), $H \in \mathbb{Z}, K \in \mathbb{Z}$
Ensure: $\sigma < \epsilon$ for sufficiently small $\epsilon \in \mathbb{R}$
  for i in $num\_trainings$ **do**
    $\bar{W}_0 \in \mathbb{R}^{H \times d} \sim \mathcal{N}(0, \sigma), \bar{W}_1 \in \mathbb{R}^{p \times H} \sim \mathcal{N}(0, \sigma)$
    for j in num_epochs **do**
      $\{\bar{W}_0, \bar{W}_1\} \leftarrow$ gradient_descent($\{\bar{W}_0, \bar{W}_1\}, X$) ▷ Apply gradient descent update step
      if j mod 100 = 0 **then** ▷ Sample at different times to find different structure as it emerges
        sample = maximum($\bar{W}_0 X, 0$) ▷ Sample latent representations with ReLU activation
        sample_binary = *step*(sample) ▷ Threshold the sample to indicate if a neuron is active
        samples = vstack(samples,sample_binary) ▷ Stack binary latent representations
      end if ▷ Each sample appended vertically appears like a new neuron
    end for
  end for
  centroids = K-means(samples,K)
  return centroids

Finding the Gates

Algorithm 1 *A preliminary algorithm for finding a ReLN*. This follows a simple K-means clustering algorithm, but with samples taken throughout training such that it is easier to identify pathways through the network as they emerge.

Require: $num\_trainings > 0, num\_epochs > 0, \sigma > 0$,($X \in \mathbb{R}^{d\times N}, Y \in \mathbb{R}^{p\times N}$) (the dataset), $H \in \mathbb{Z}, K \in \mathbb{Z}$
Ensure: $\sigma < \epsilon$ for sufficiently small $\epsilon \in \mathbb{R}$
  for i in $num\_trainings$ **do**
    $\bar{W}_0 \in \mathbb{R}^{H\times d} \sim \mathcal{N}(0,\sigma), \bar{W}_1 \in \mathbb{R}^{p\times H} \sim \mathcal{N}(0,\sigma)$
    for j in num_epochs **do**
      $\{\bar{W}_0, \bar{W}_1\} \leftarrow$ gradient_descent($\{\bar{W}_0, \bar{W}_1\}, X$) ▷ Apply gradient descent update step
      if j mod 100 = 0 **then** ▷ Sample at different times to find different structure as it emerges
        sample = maximum($\bar{W}_0 X, 0$) ▷ Sample latent representations with ReLU activation
        sample_binary = *step*(sample) ▷ Threshold the sample to indicate if a neuron is active
        samples $=$ vstack(samples,sample_binary) ▷ Stack binary latent representations
      end if ▷ Each sample appended vertically appears like a new neuron
    end for
  end for
  centroids $=$ K-means(samples,K)
  return centroids

Finding 5: We provide a preliminary algorithm to identify ReLNs from ReLU Networks.