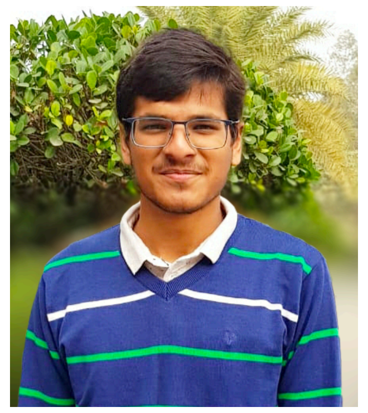
TECHNION Israel Institute of Technology

Delhi Technological University

Carnegie Mellon University

Measures of Information Reflect Memorization Patterns

Rachit Bansal, Danish Pruthi, Yonatan Belinkov



https://information-measures.cs.technion.ac.il

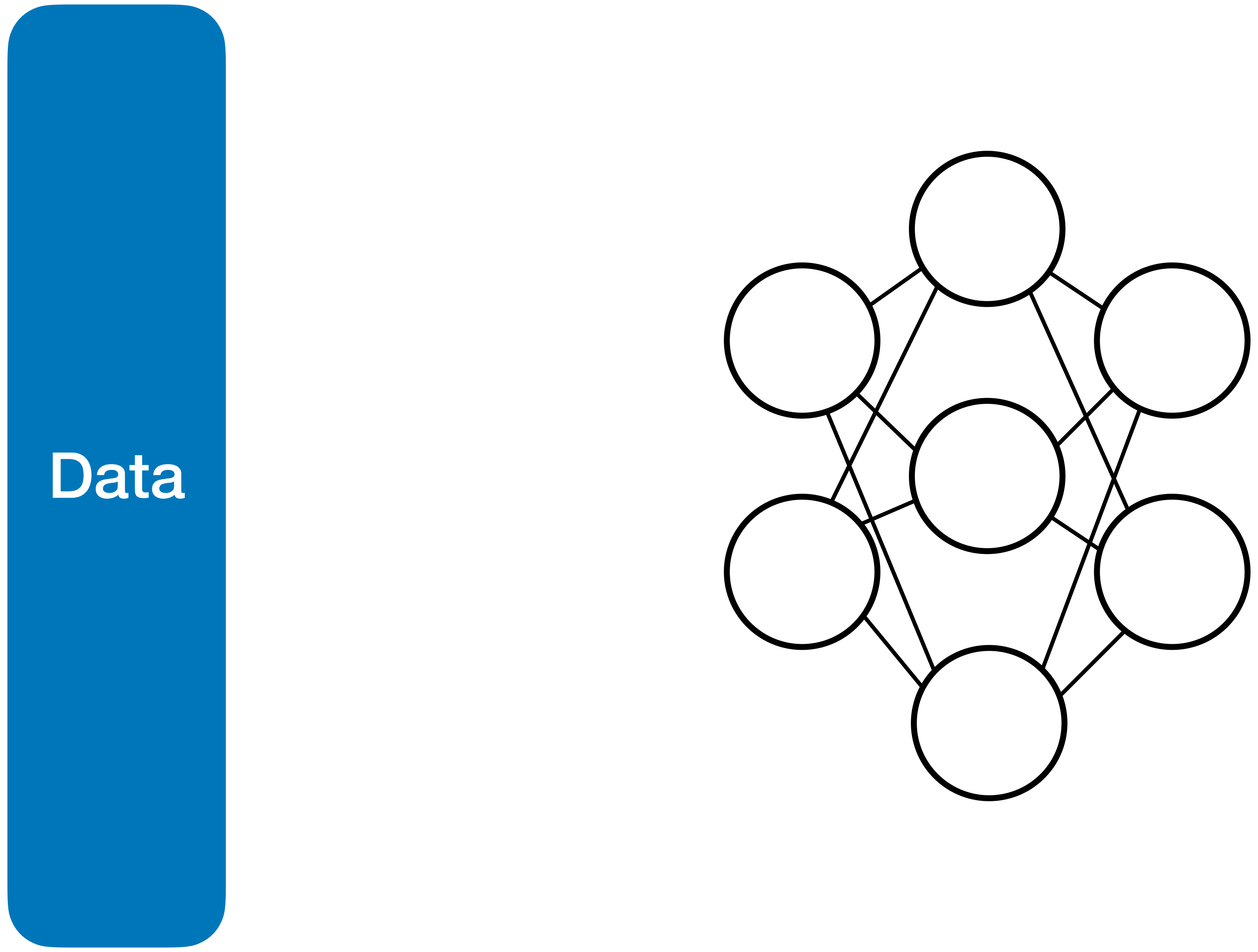
Correspondence: racbansa@gmail.com

NEURAL INFORMATION PROCESSING SYSTEMS

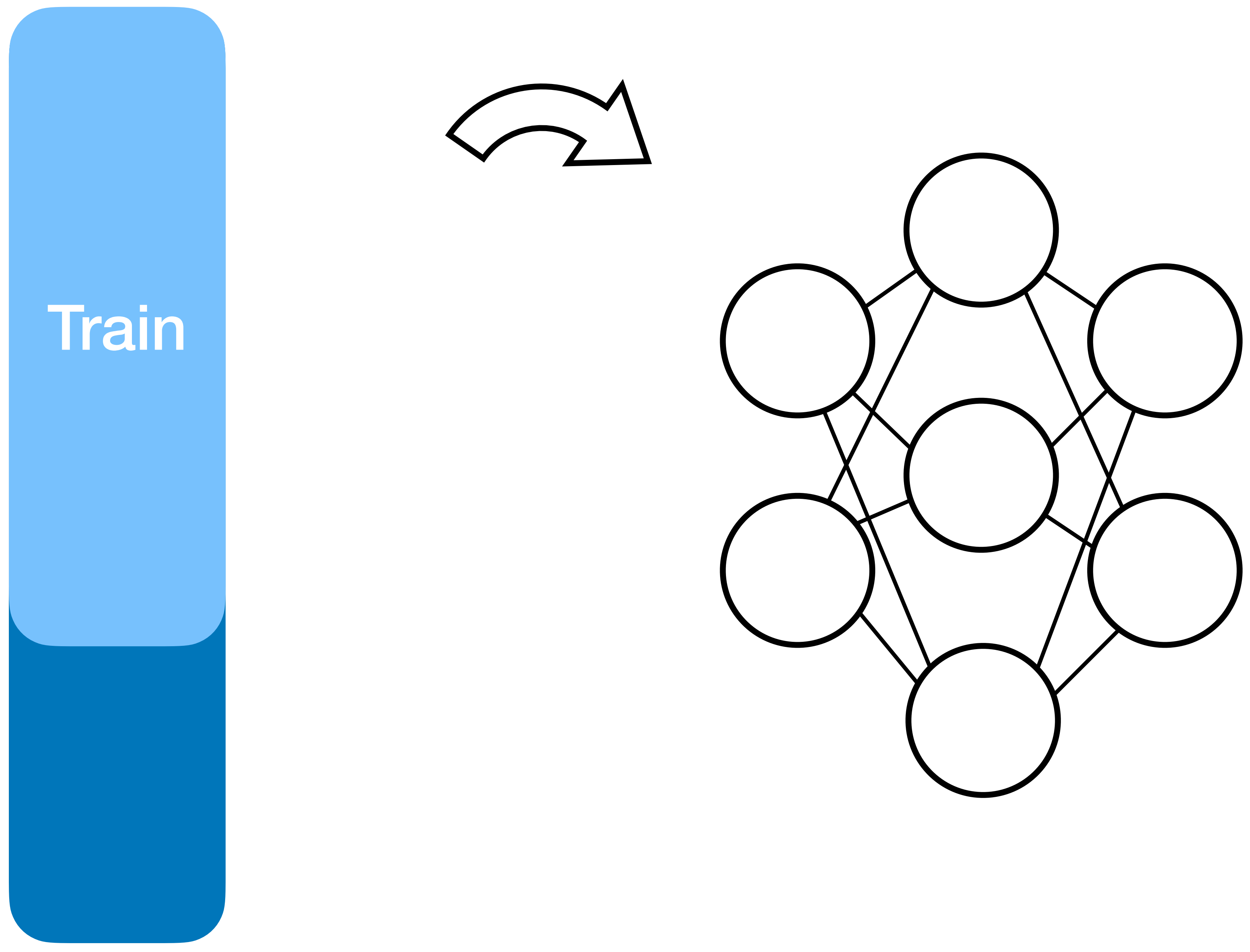
“Evaluating model generalization
using intrinsic,
information-theoretic metrics.”

“*Evaluating* model *generalization* using *intrinsic*, *information-theoretic* metrics.”

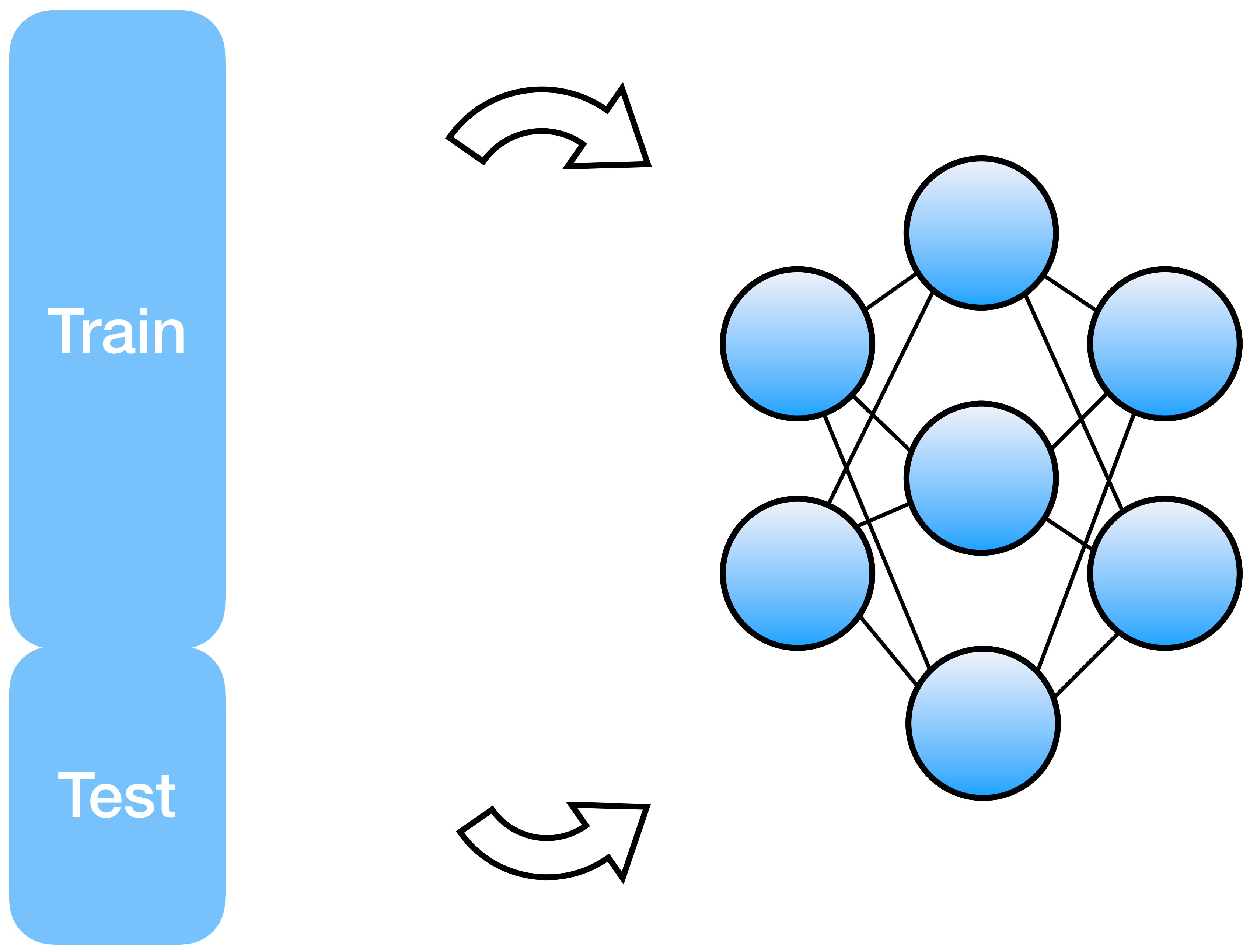
“*Evaluating* model *generalization* using *intrinsic*, *information-theoretic* metrics.”



Data

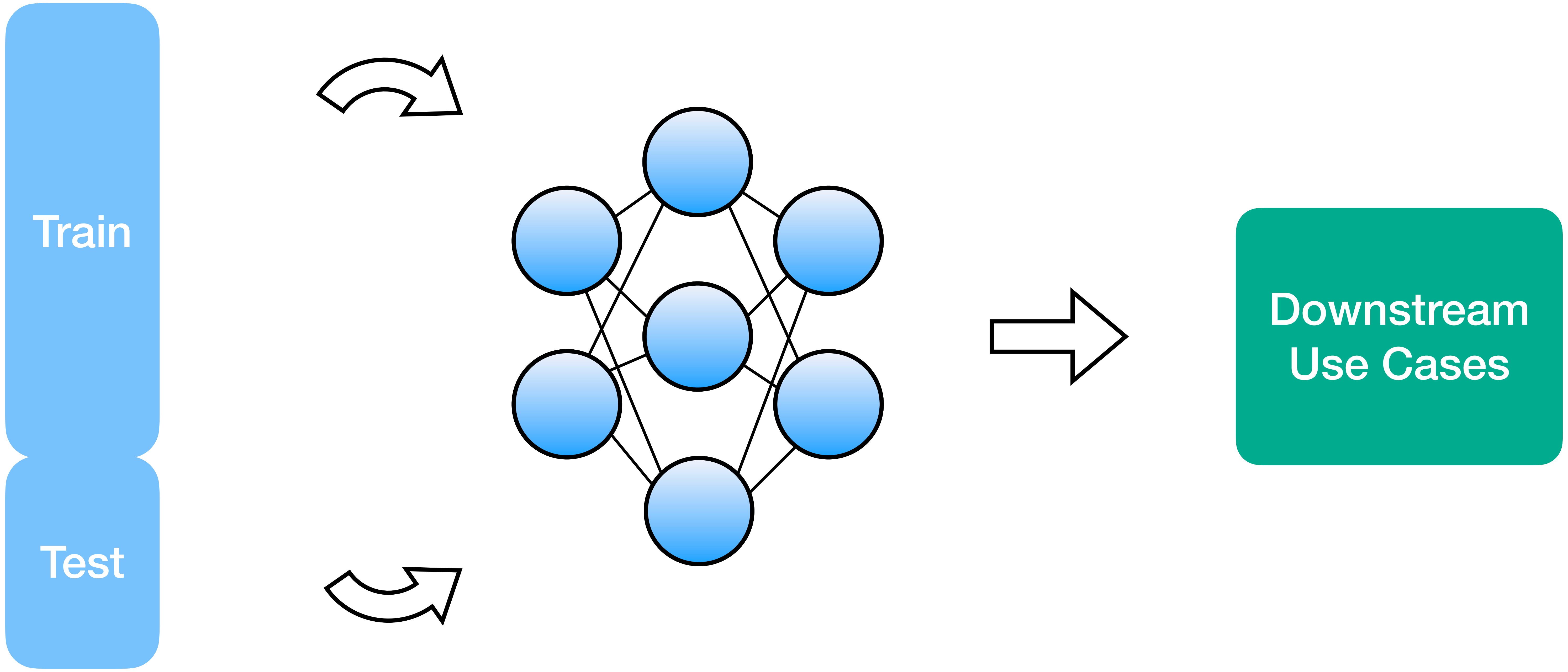


Train



Train

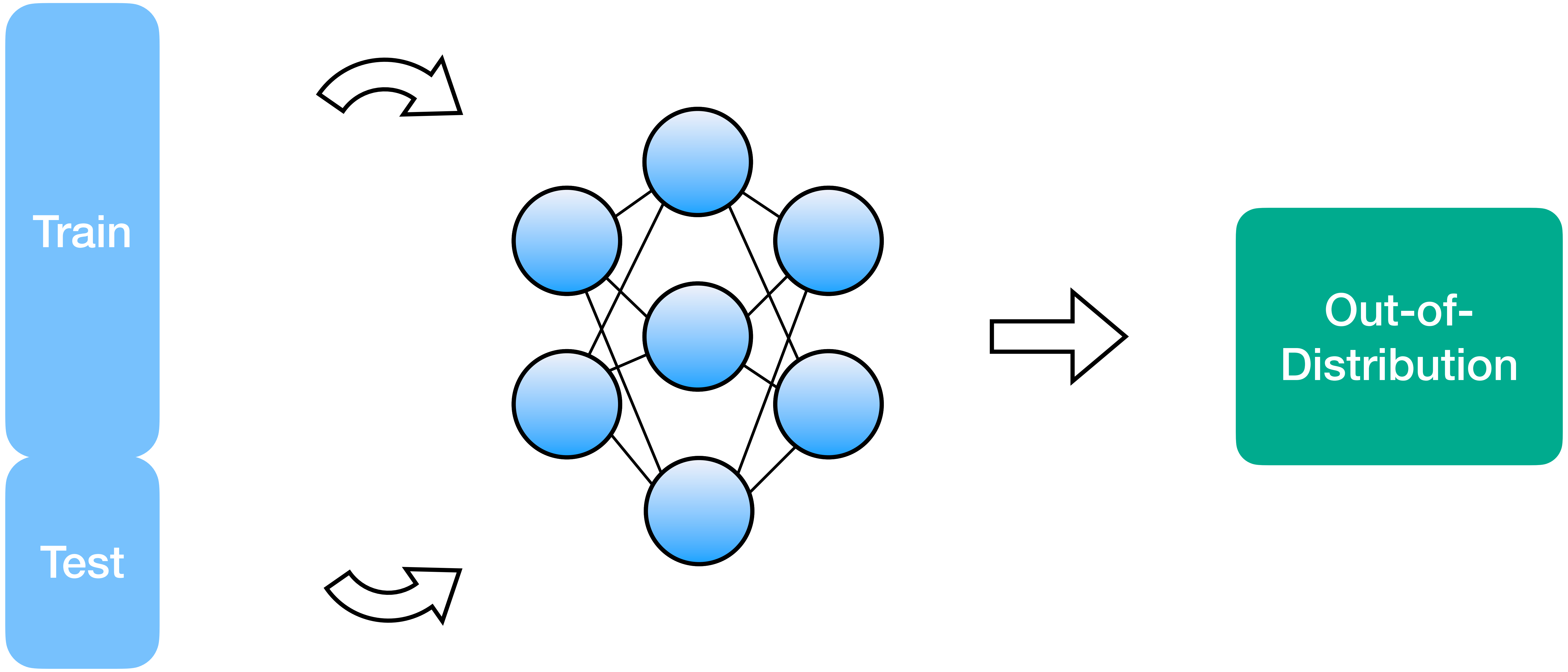
Test



Train

Test

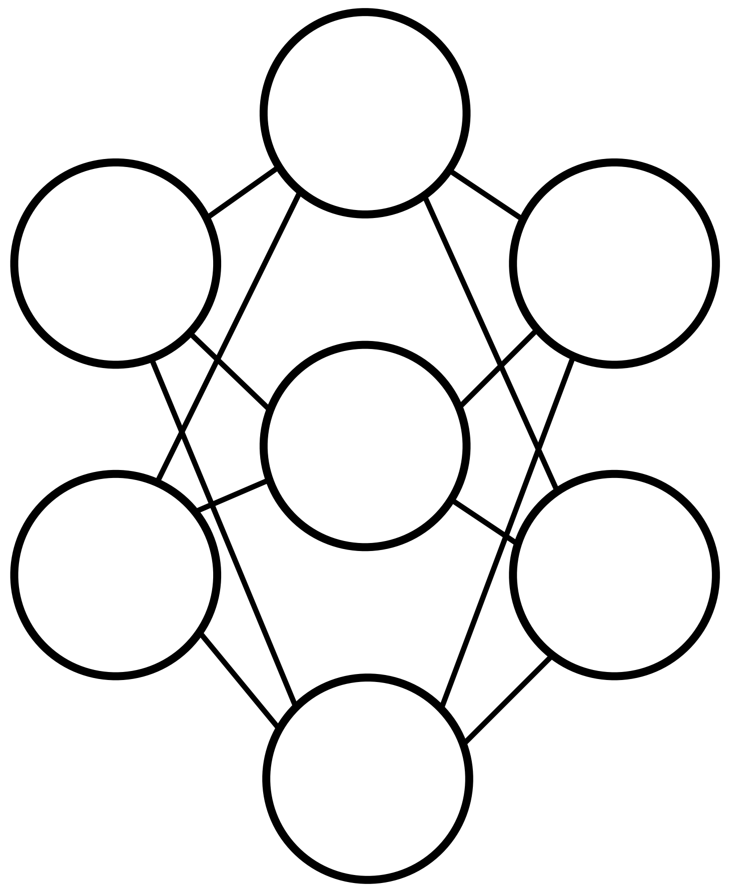
Downstream Use Cases

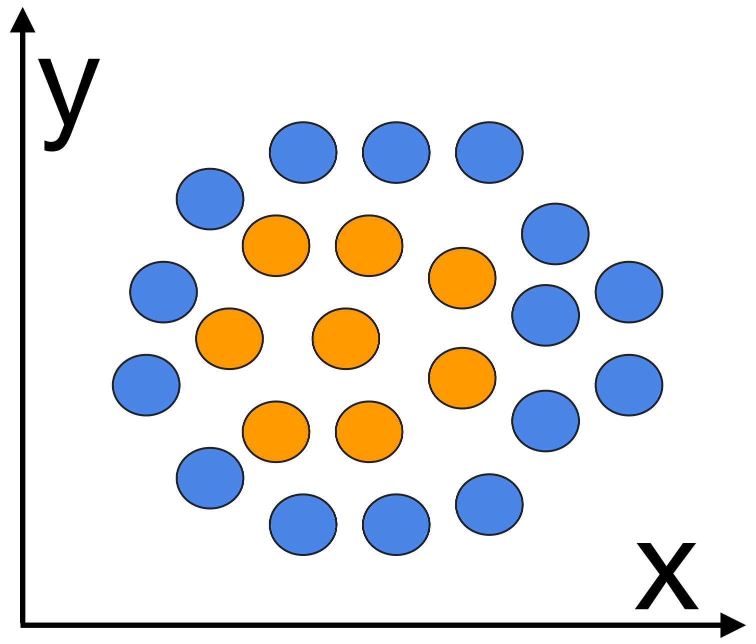


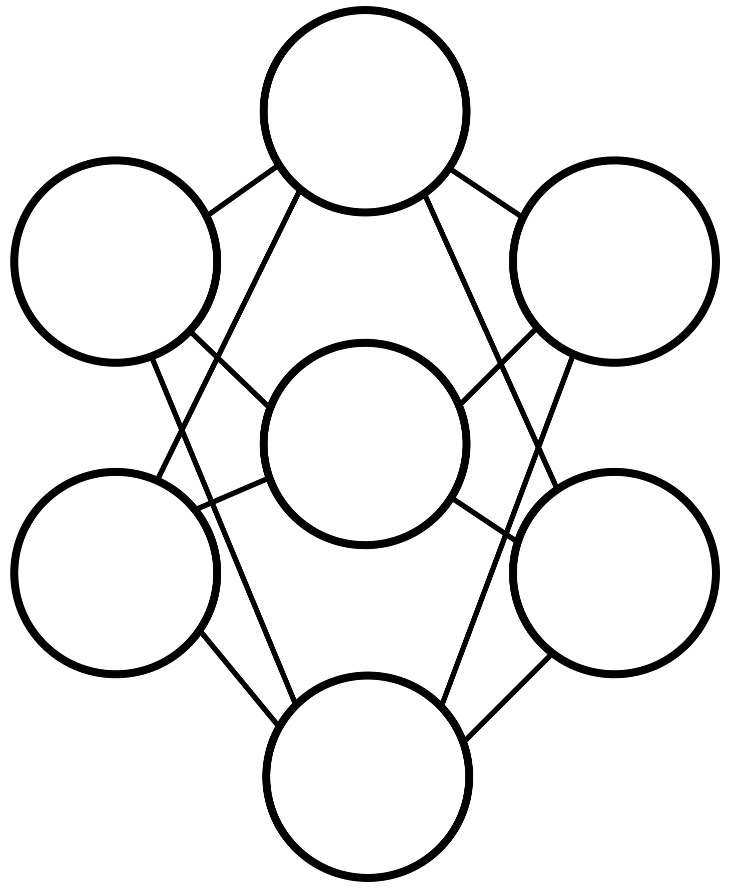
Train

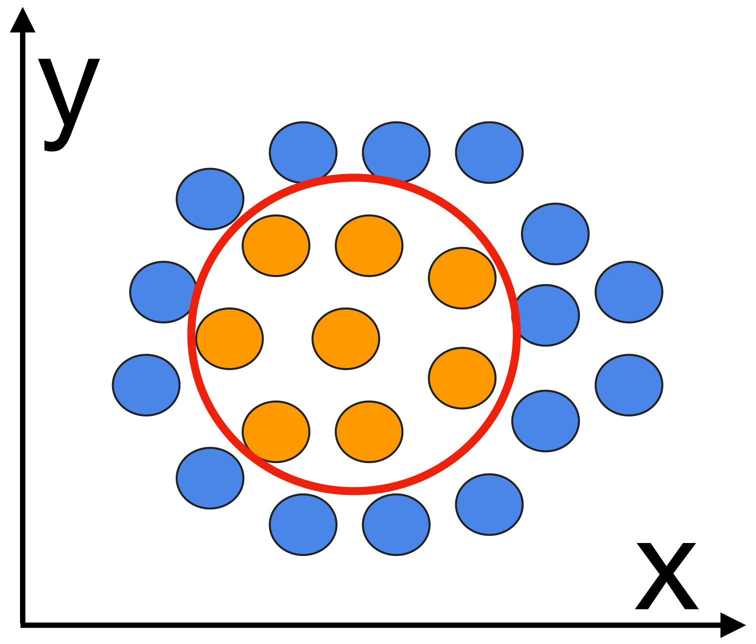
Test

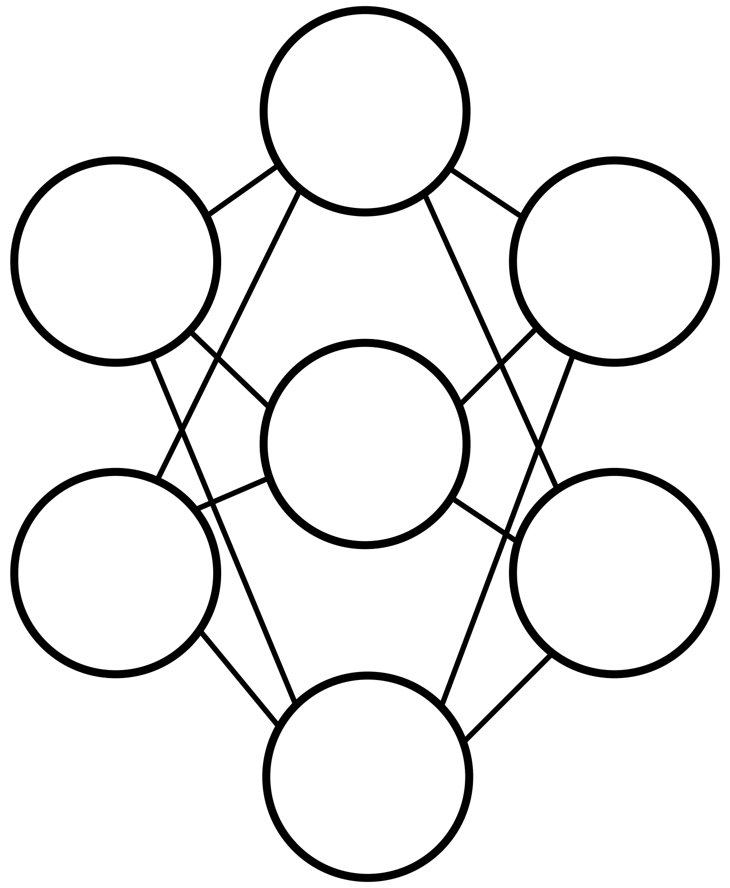
Out-of-Distribution

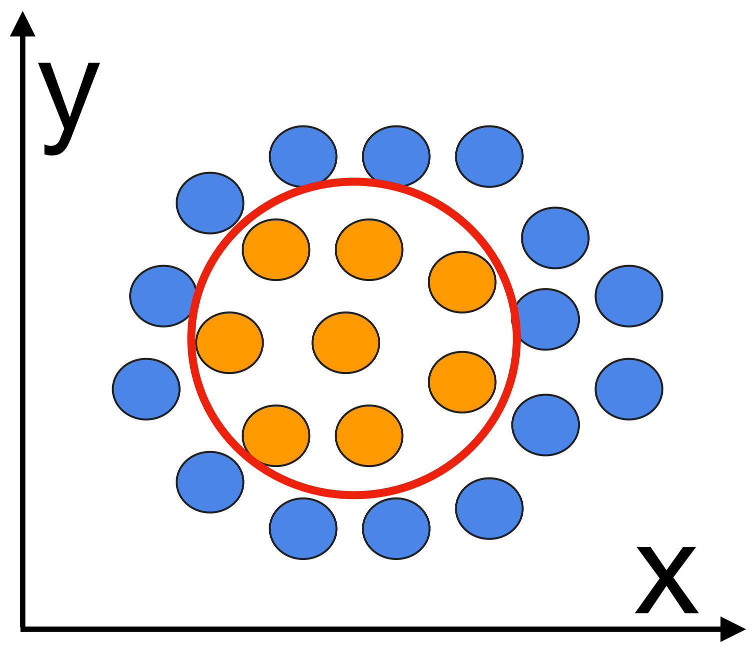


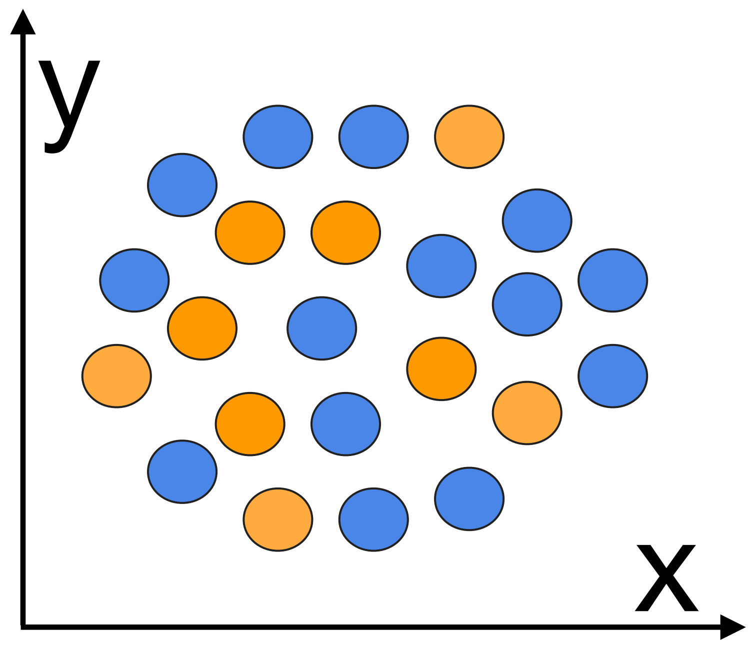

y
x

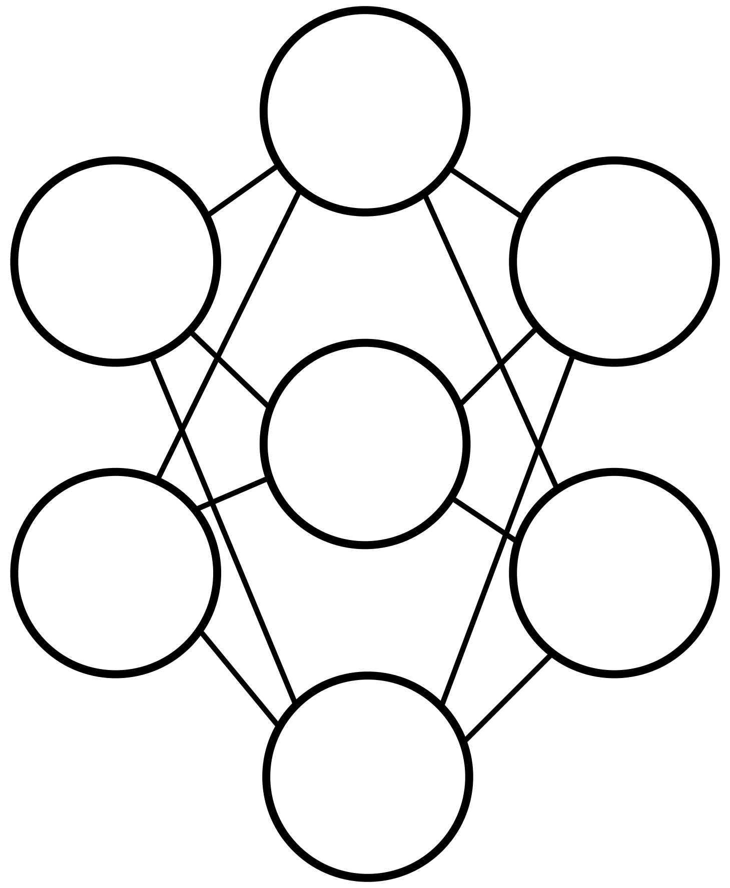


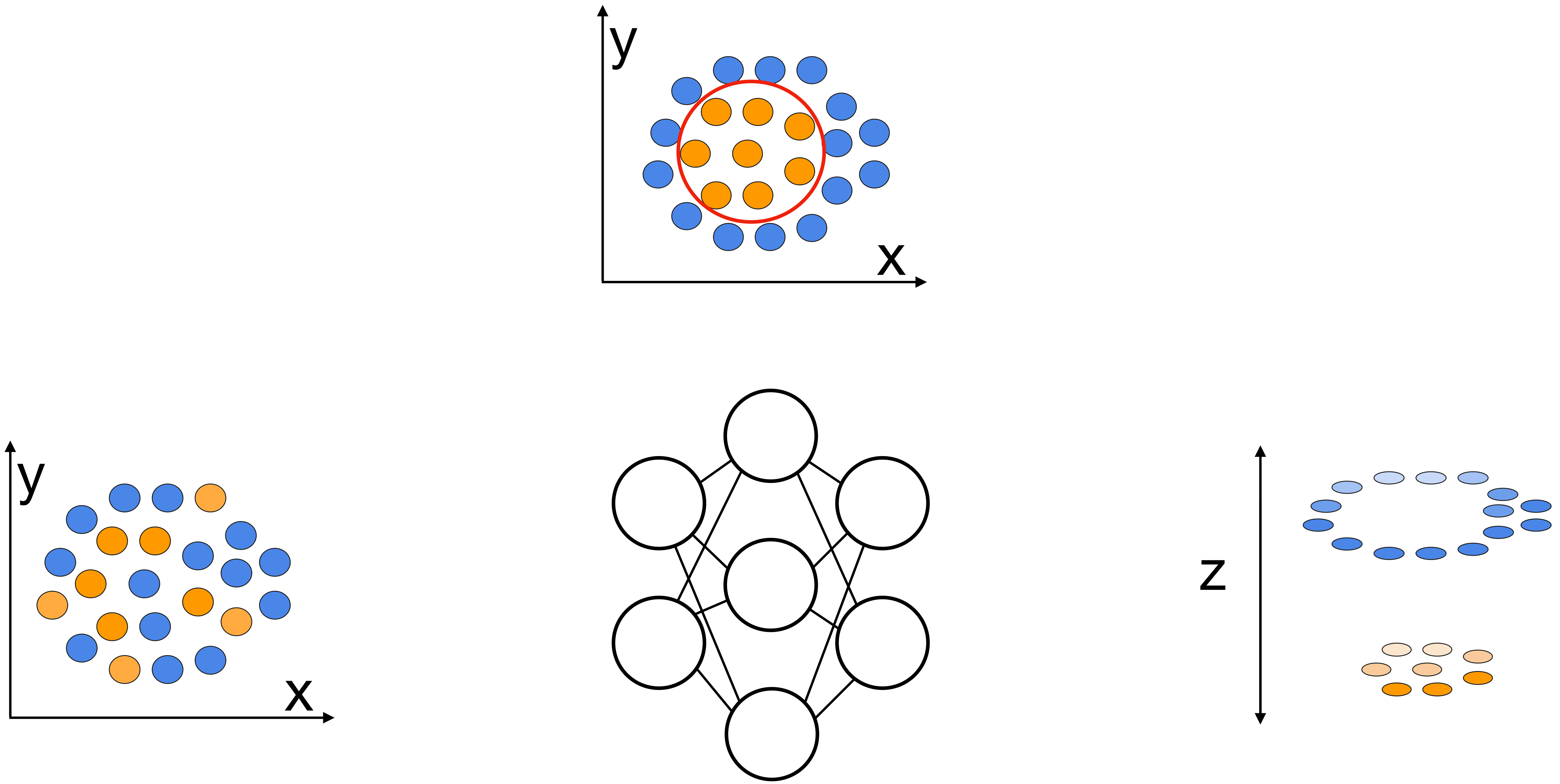

y
x

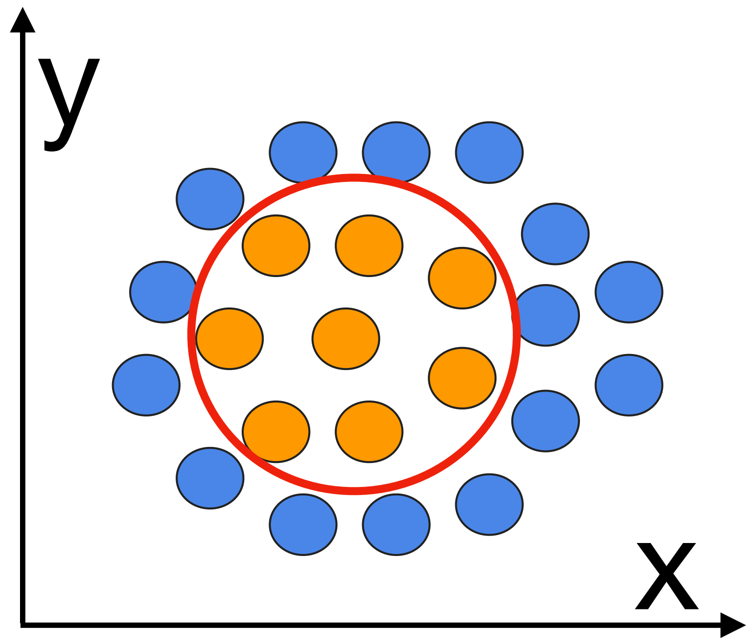


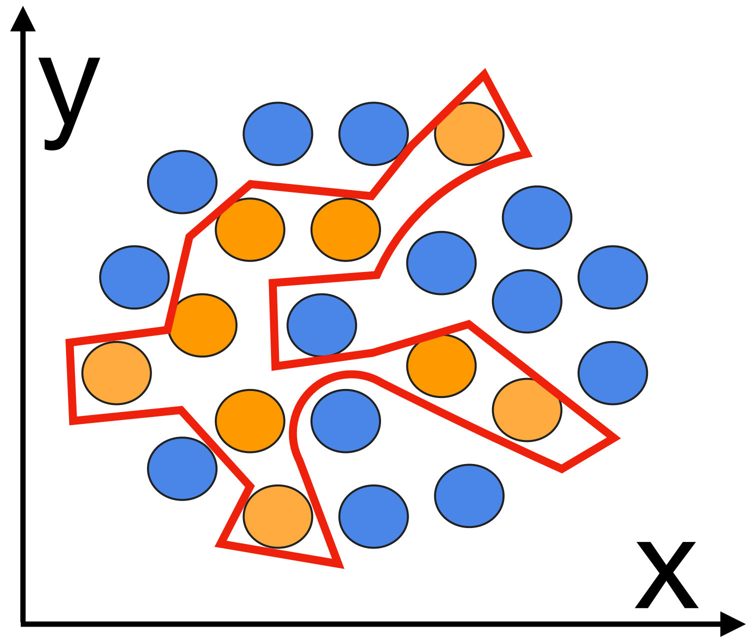

y
x

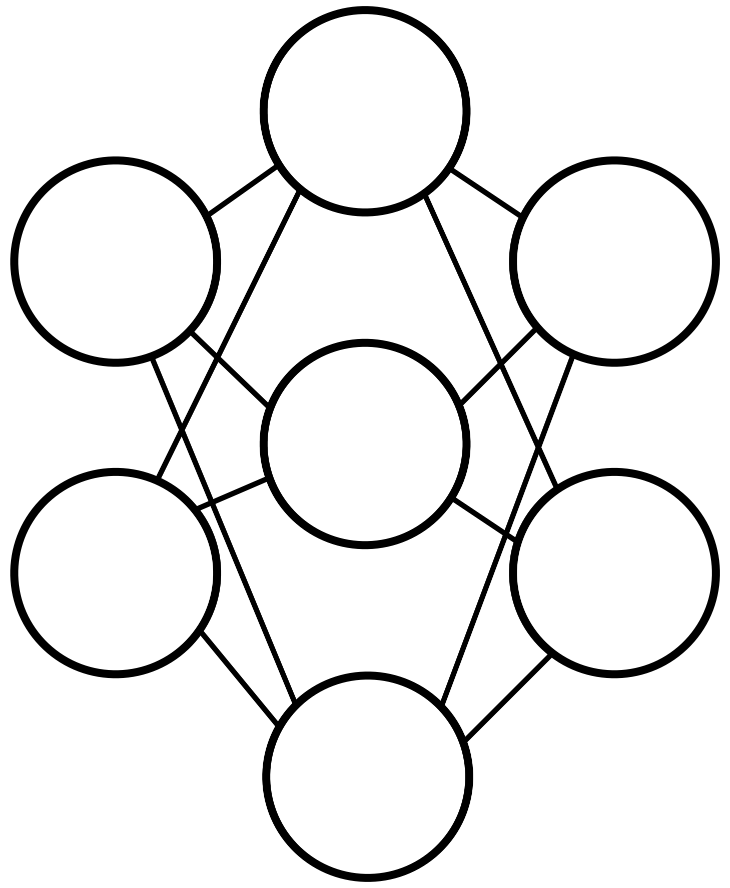

y
x

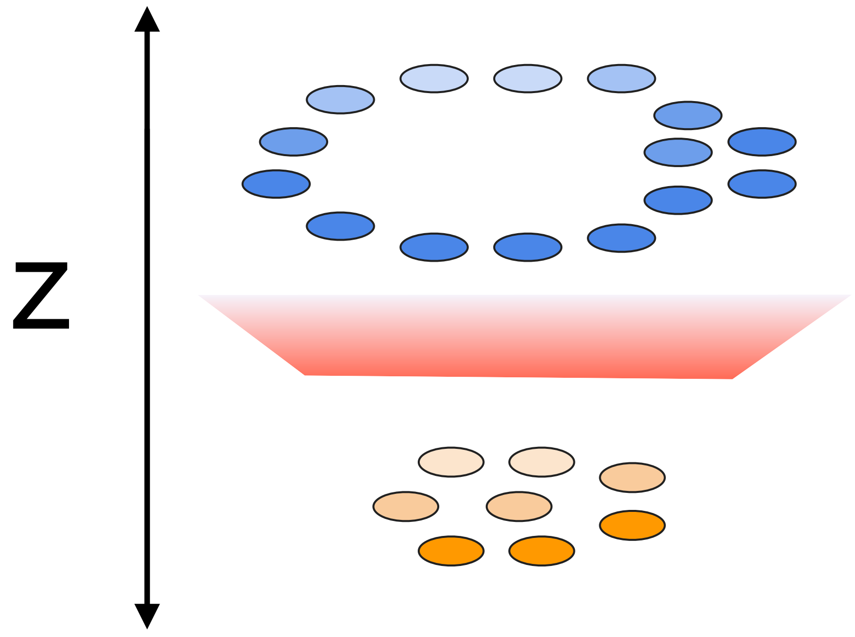


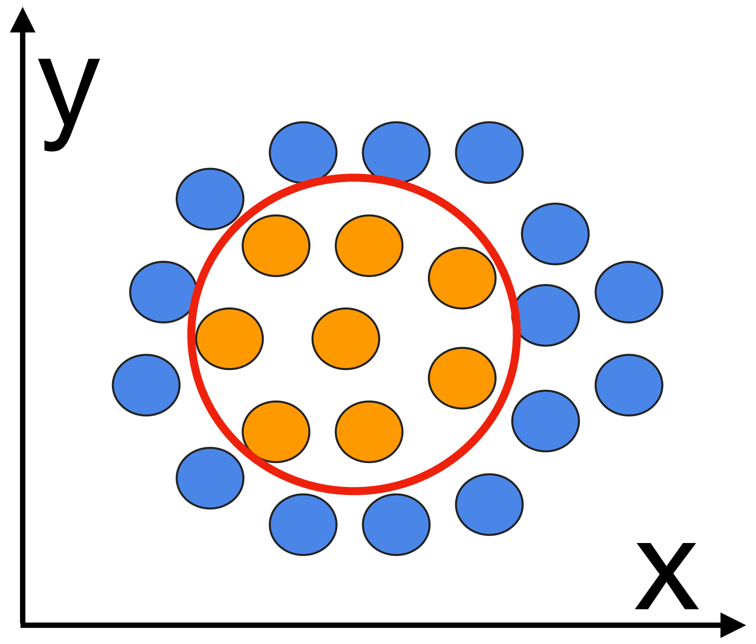

y
x
y
x
z

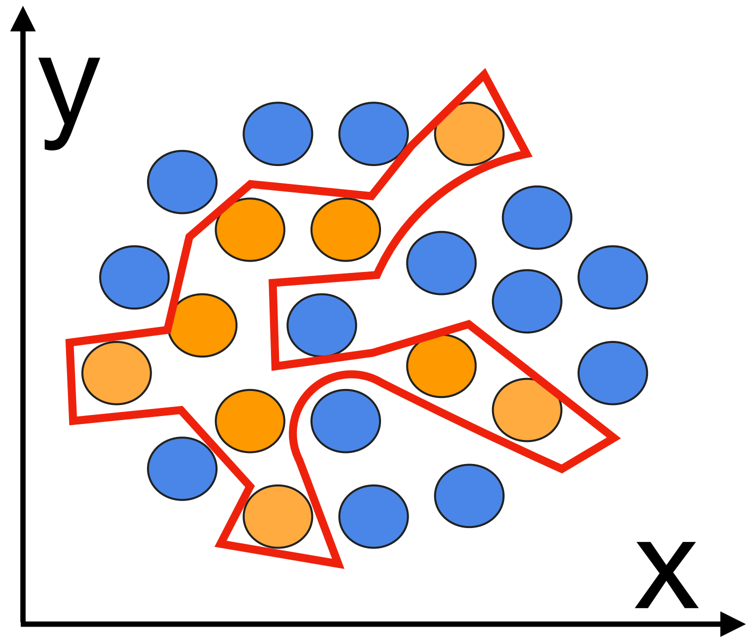

y
x

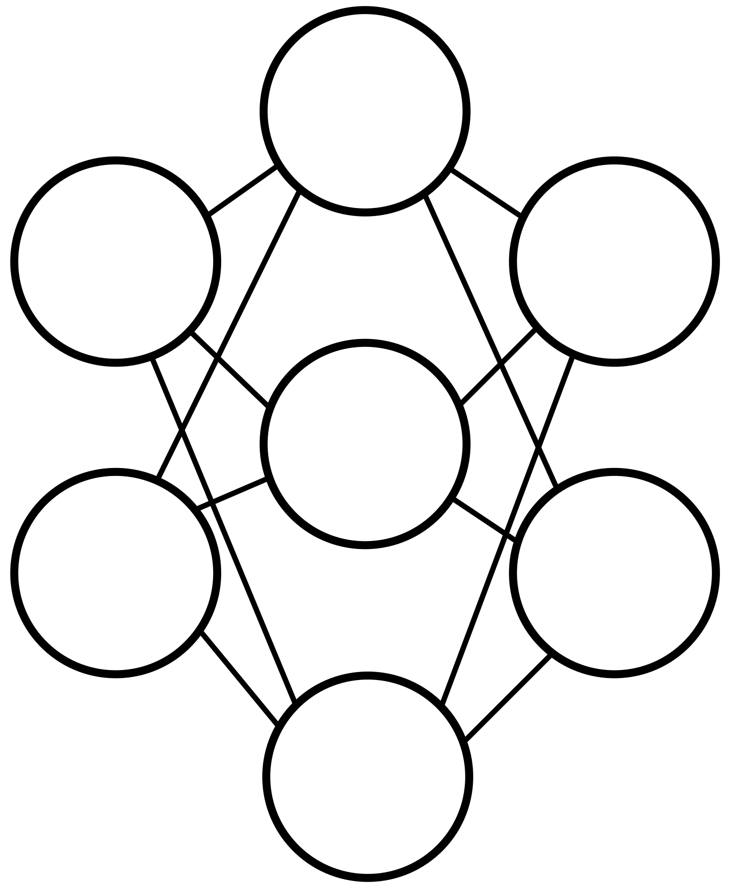

y
x

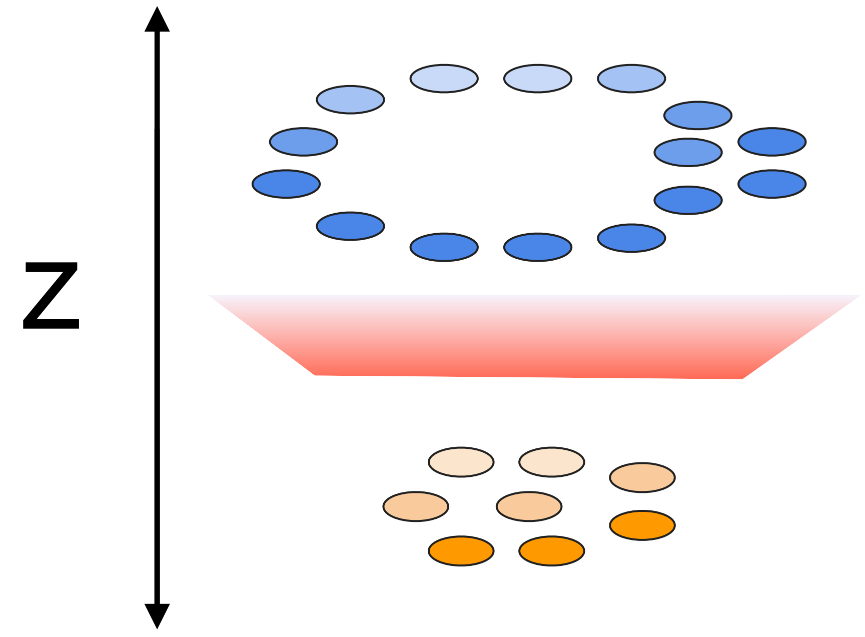


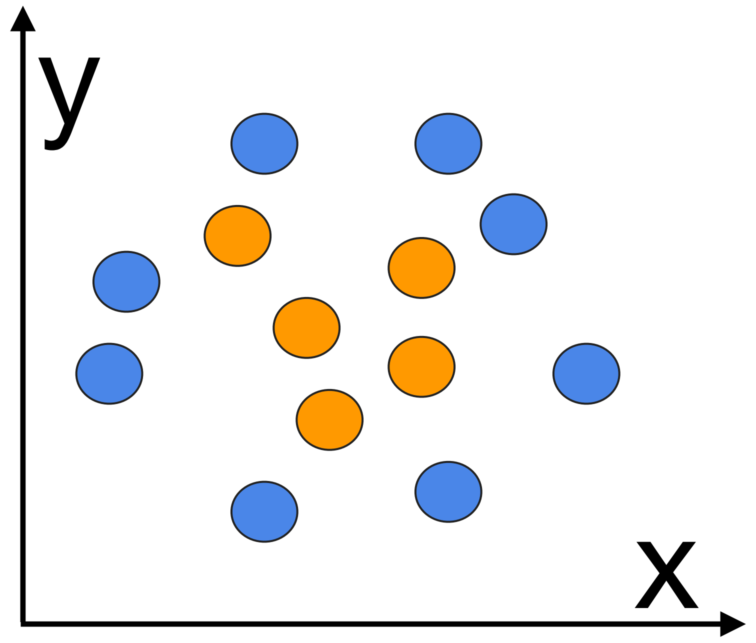

z



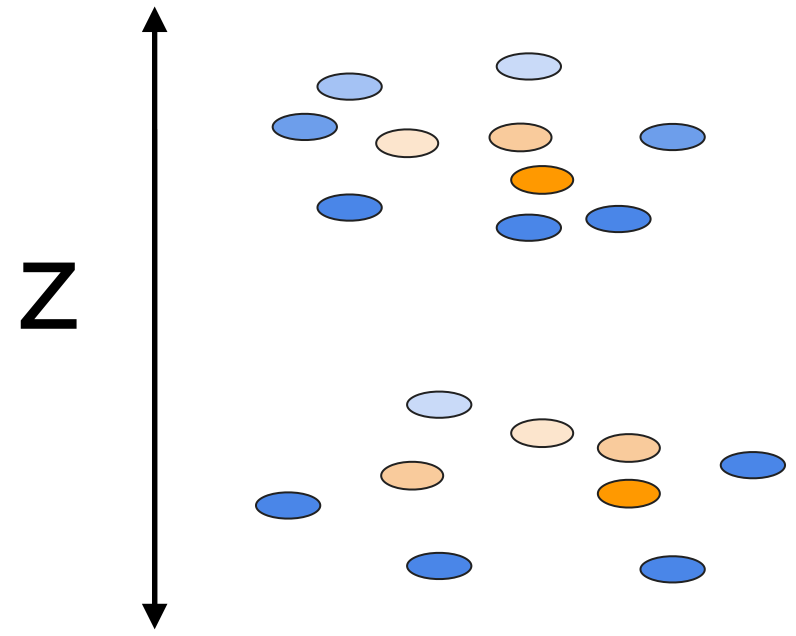






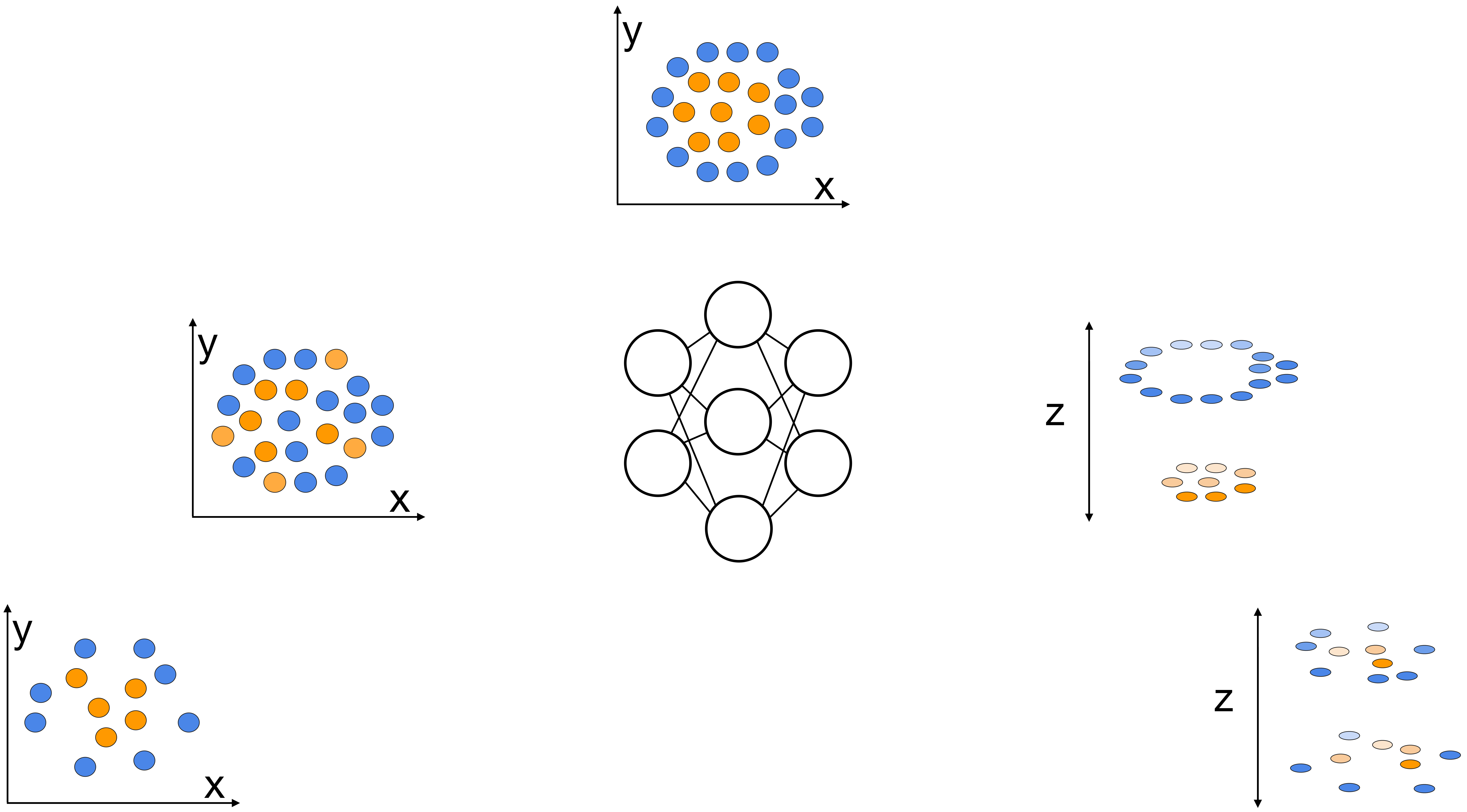


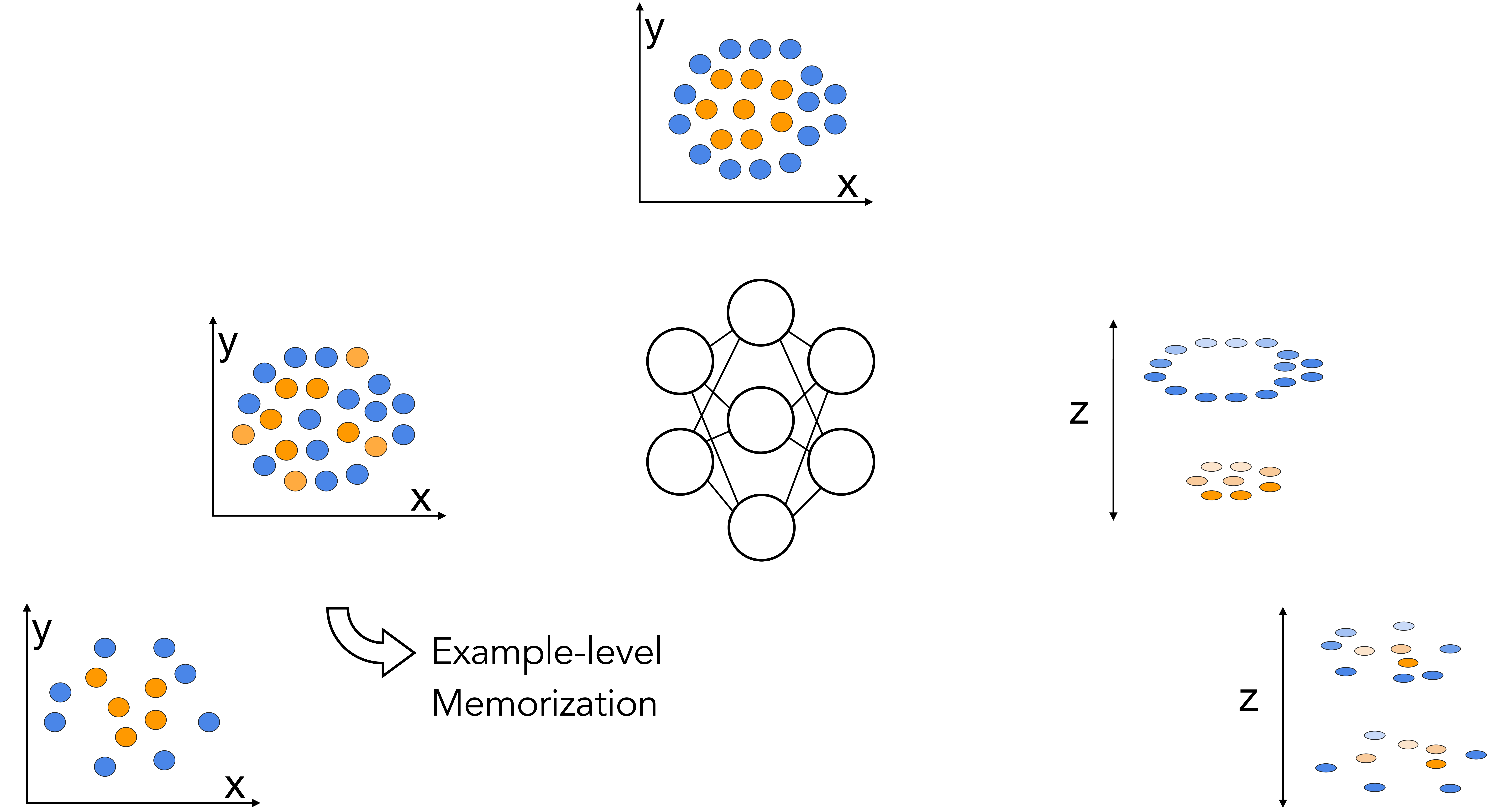
Specialised sets are required
to identify these cases of memorization.



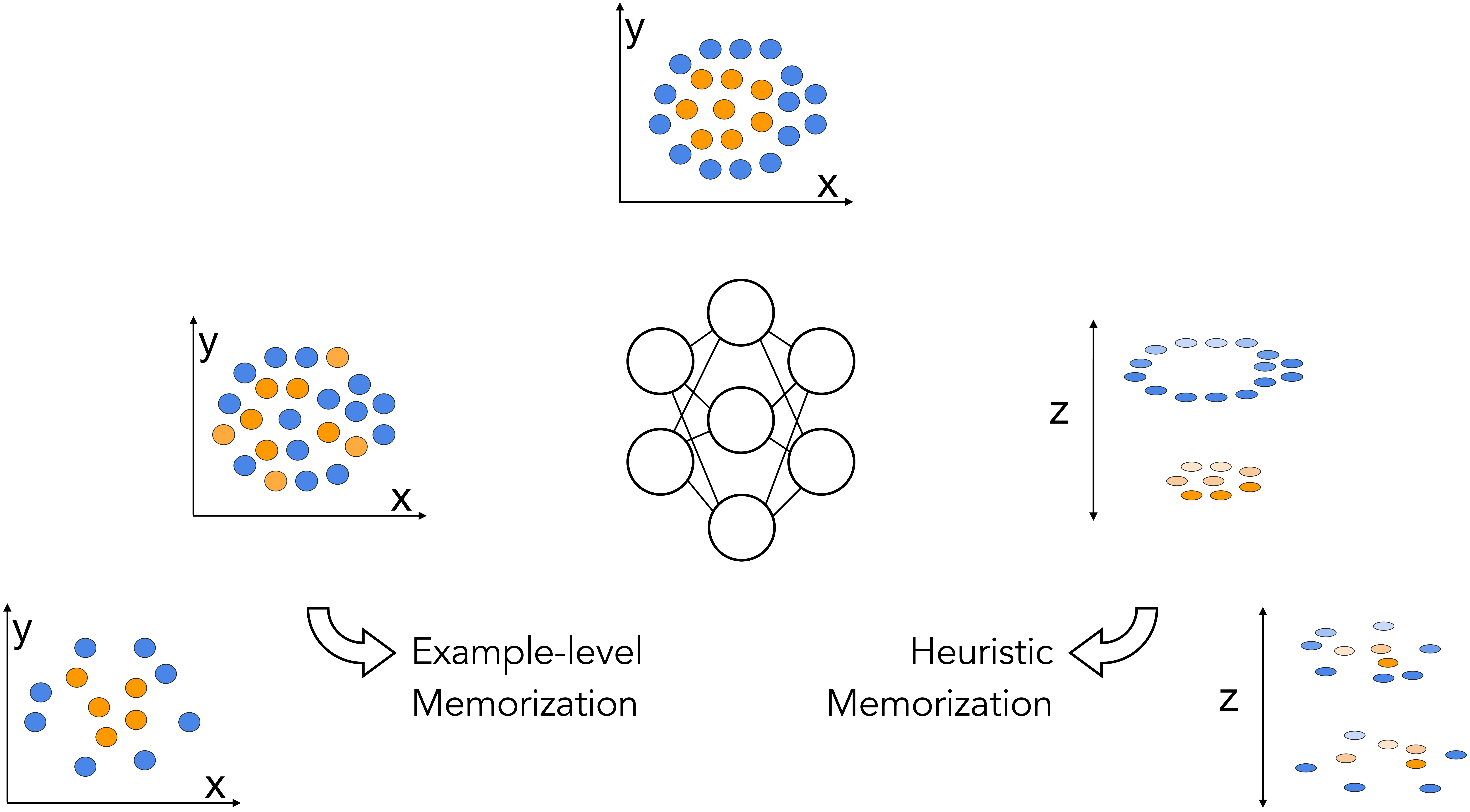
“*Evaluating* model *generalization* using *intrinsic*, *information-theoretic* metrics.”

“*Evaluating* model *generalization* using *intrinsic*, *information-theoretic* metrics.”


y
x
y
x
z
y
x
z



Example-level Memorization

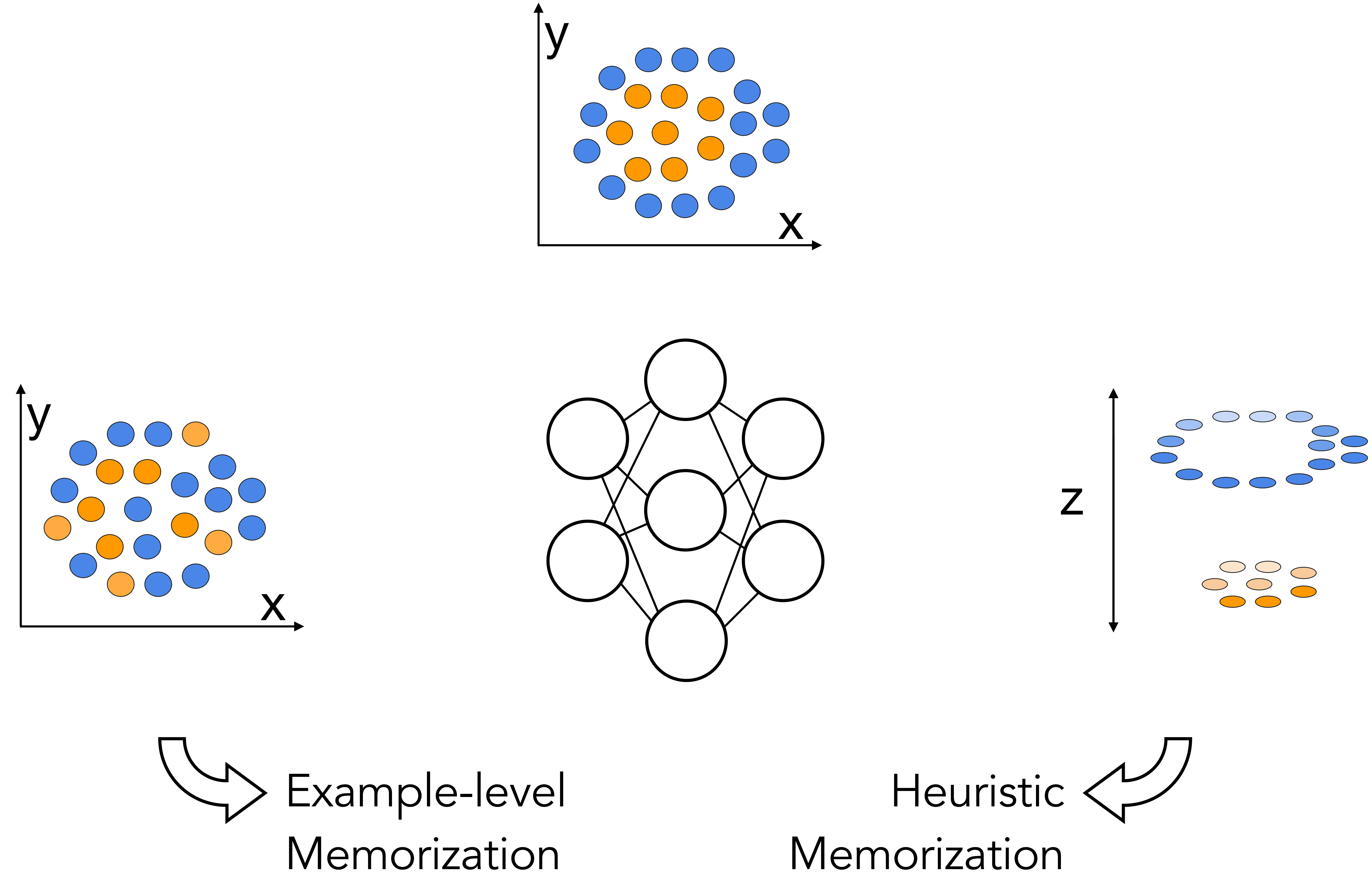


Example-level Memorization

Heuristic Memorization

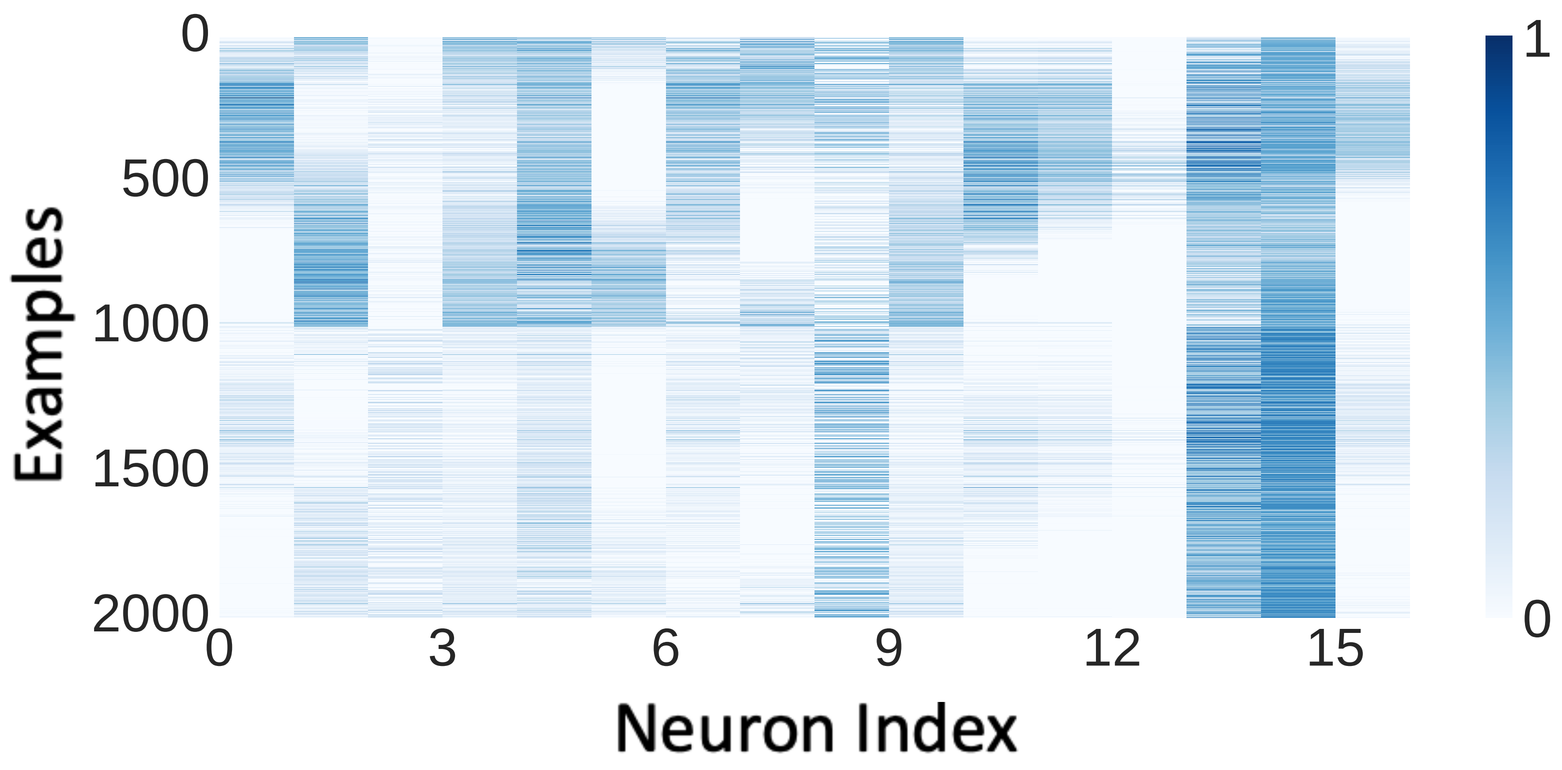
"*Evaluating* model *generalization*
using *intrinsic*,
information-theoretic metrics."

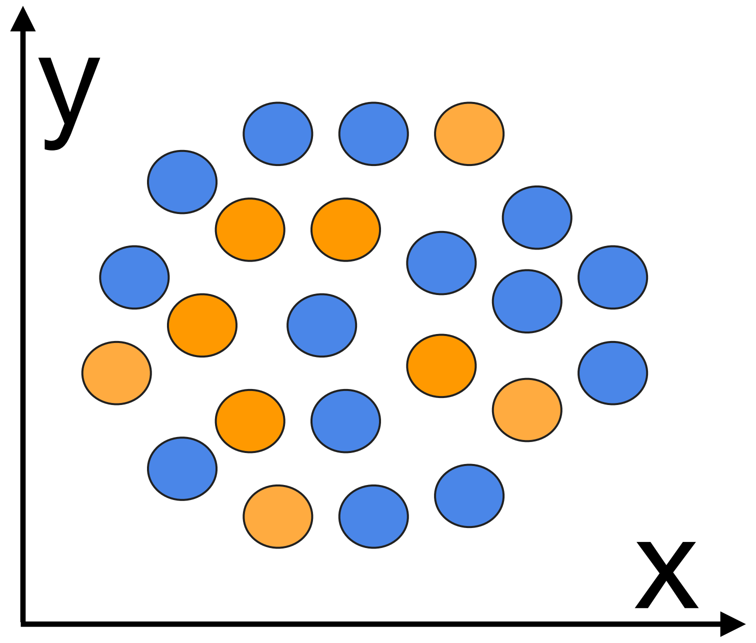
"*Evaluating* model *generalization*
using *intrinsic*,
information-theoretic metrics."

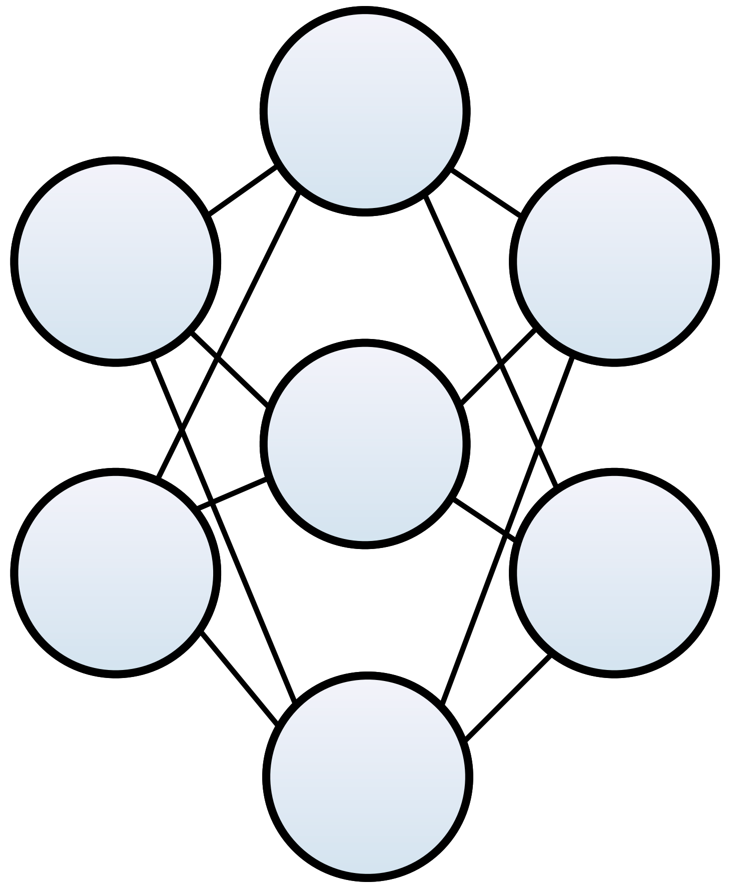


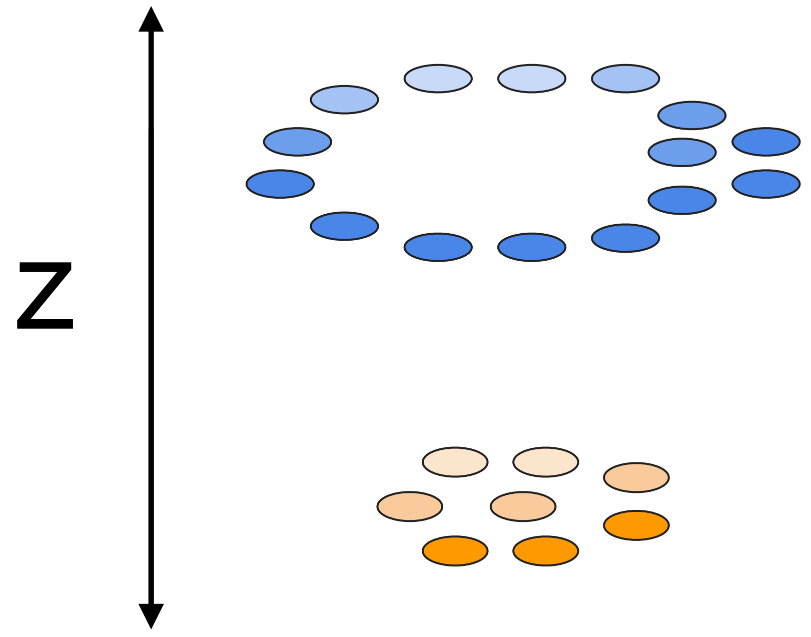
Example-level Memorization

Heuristic Memorization



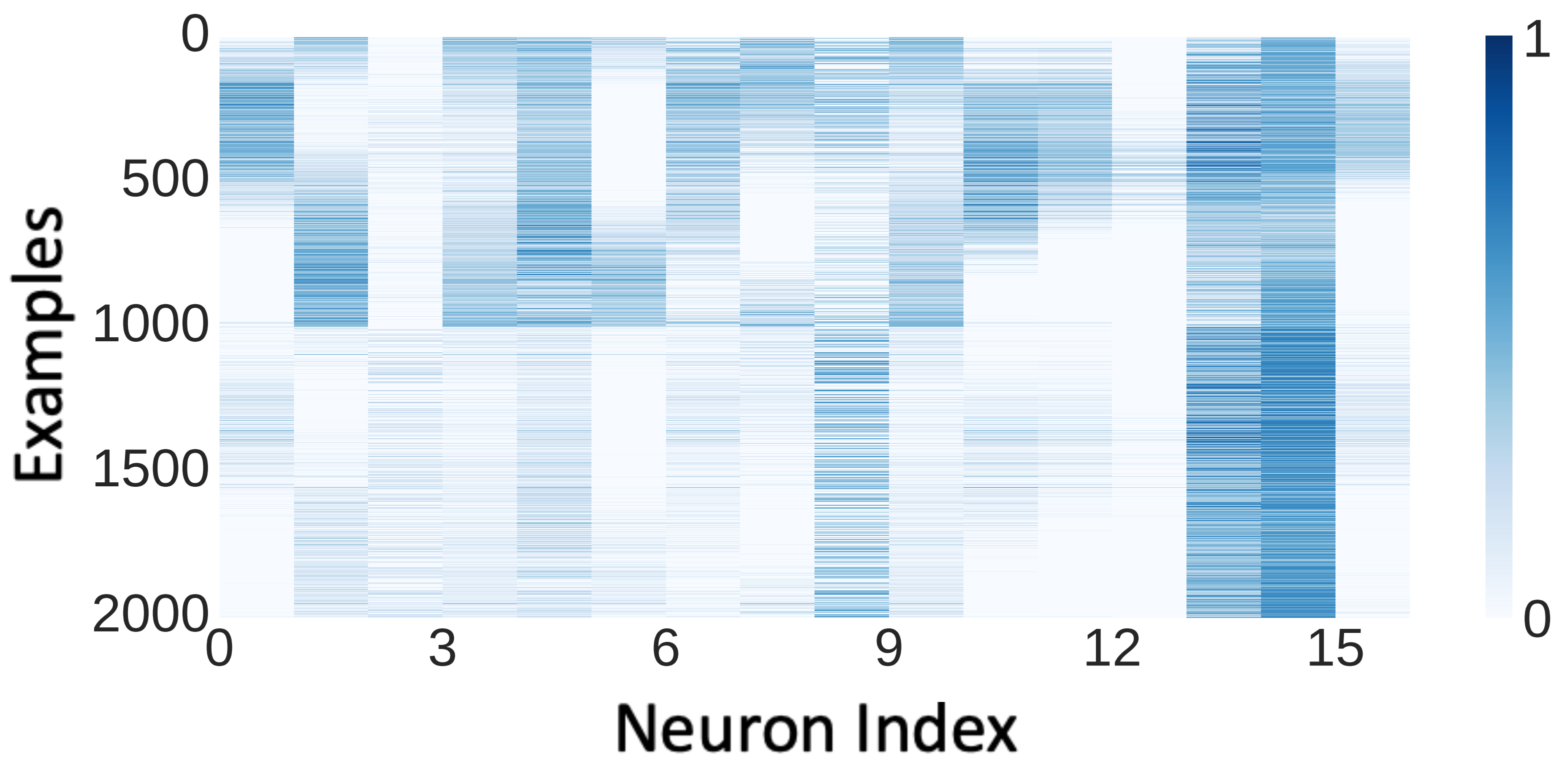


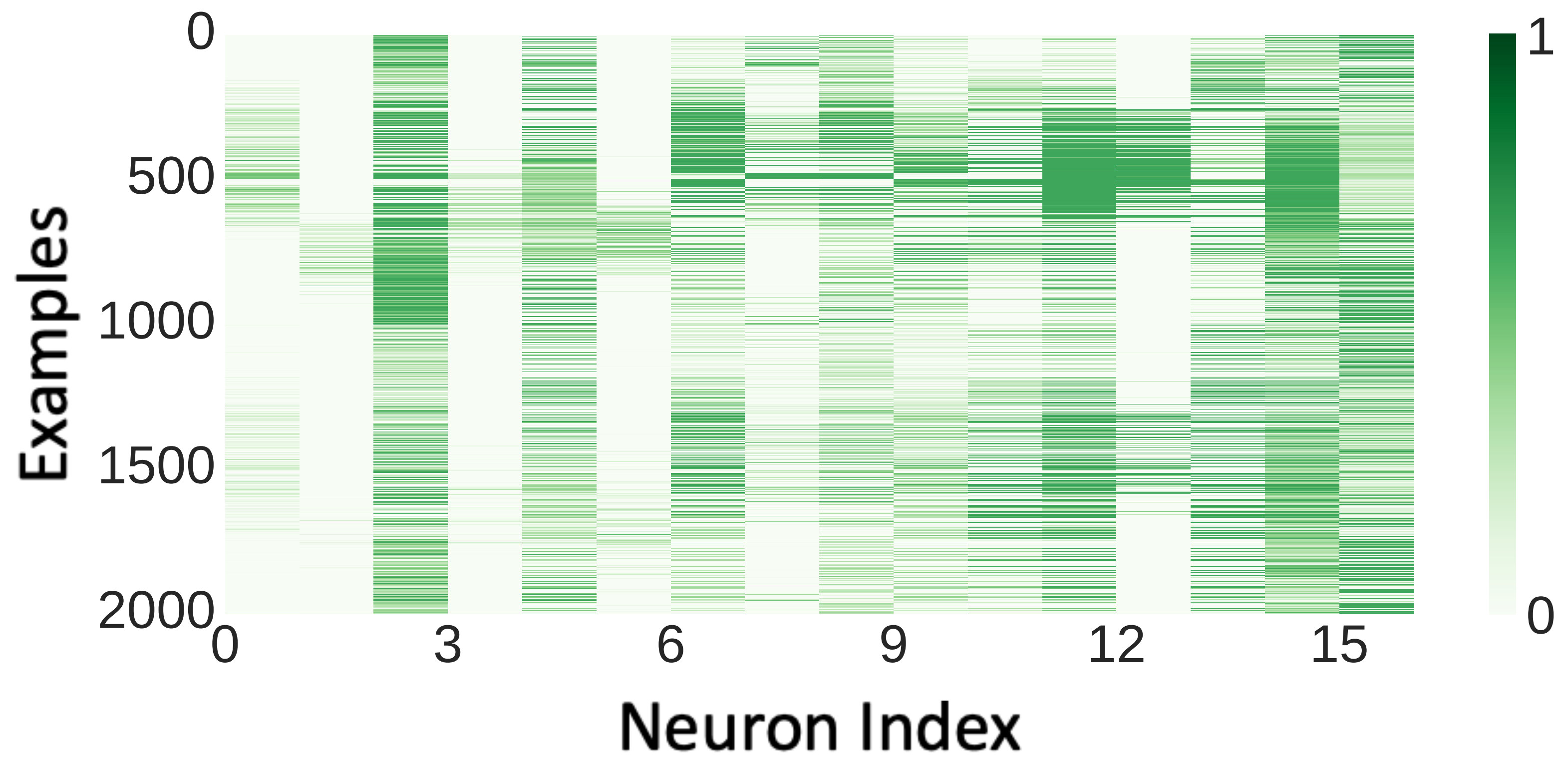


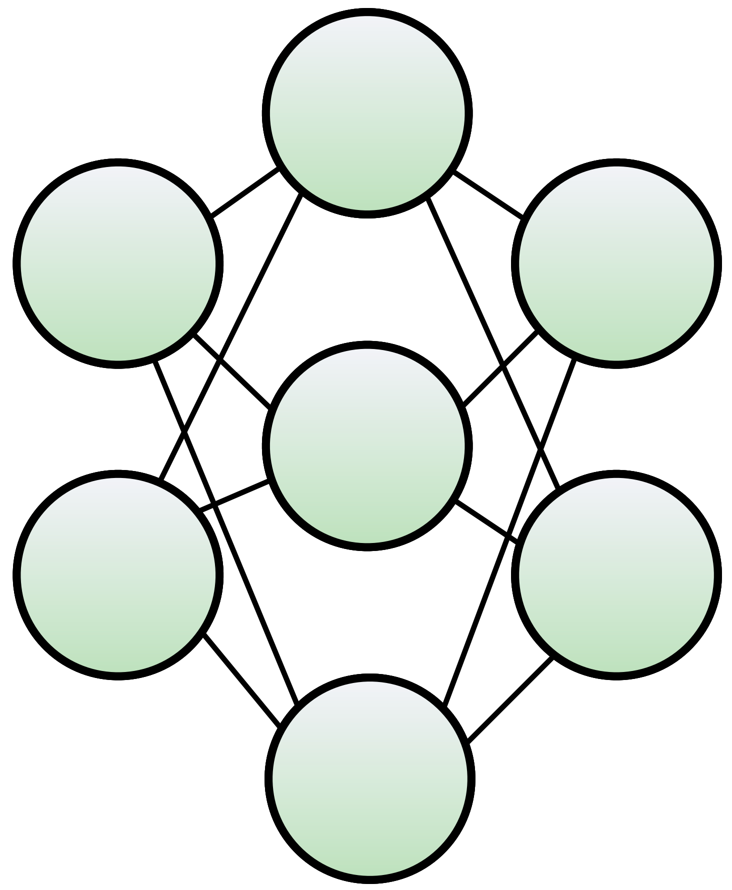


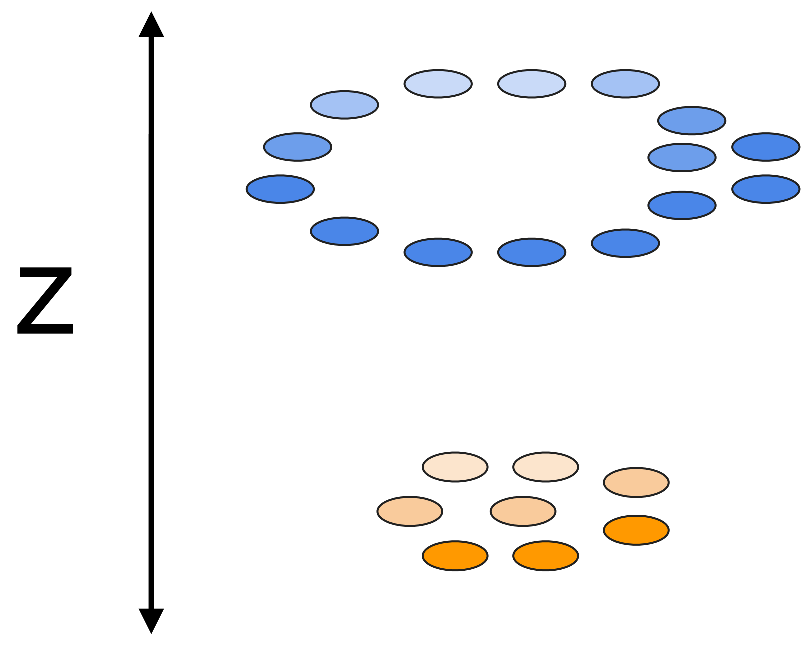
Example-level
Memorization

Heuristic
Memorization



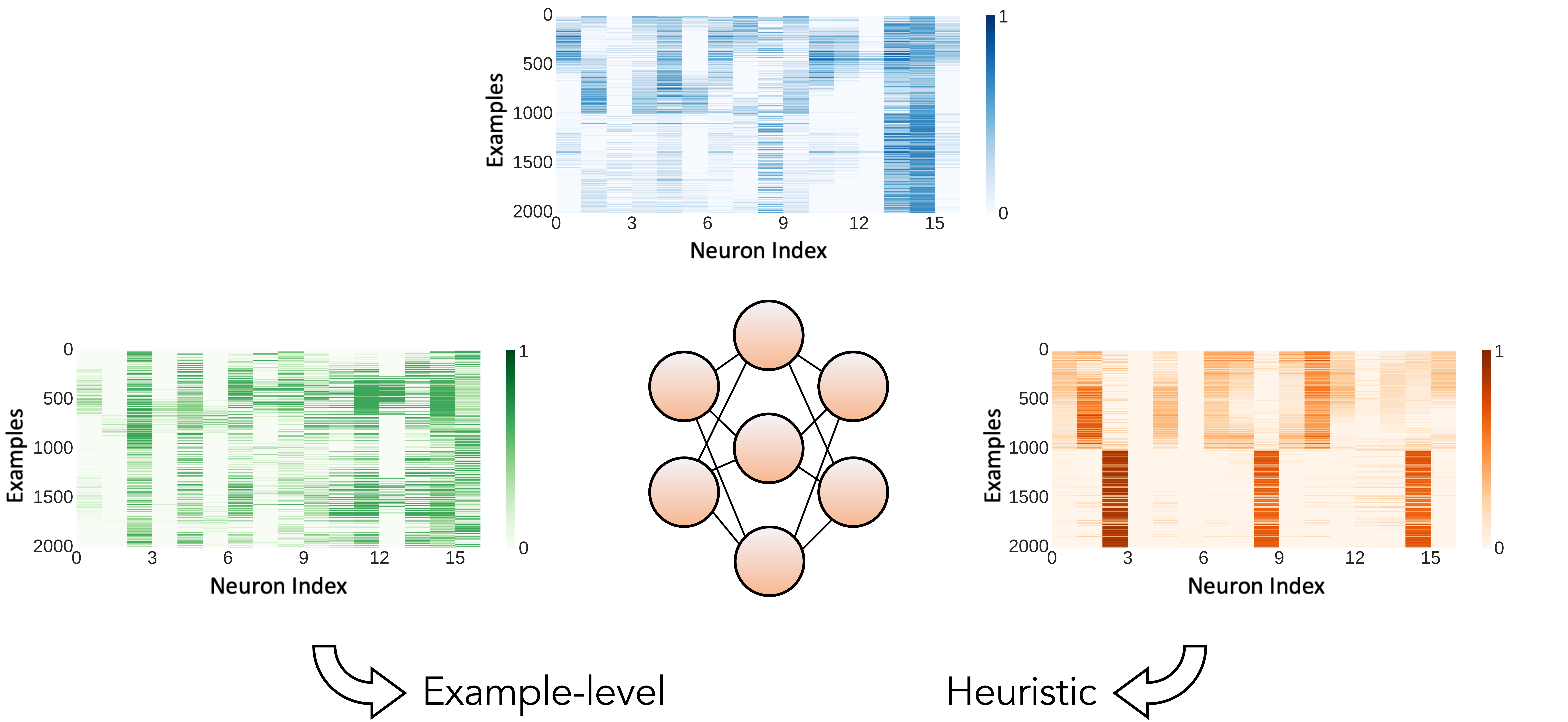






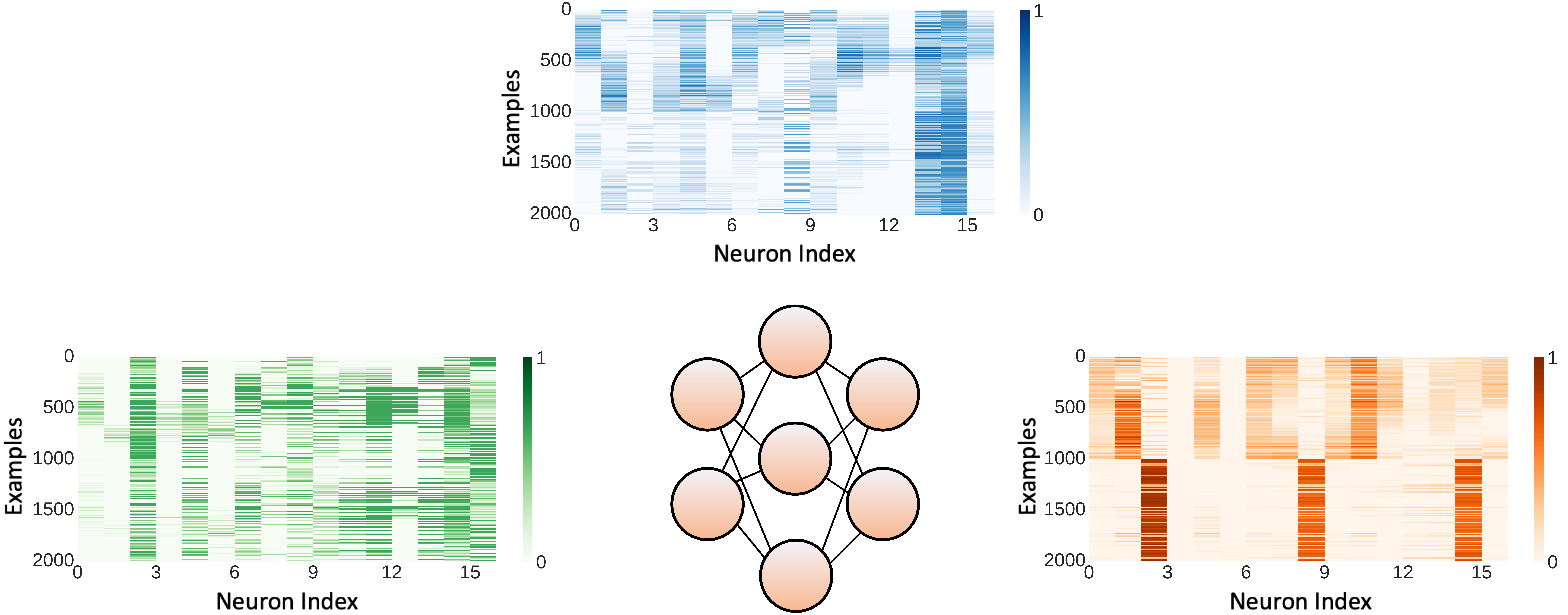
Example-level
Memorization

Heuristic
Memorization

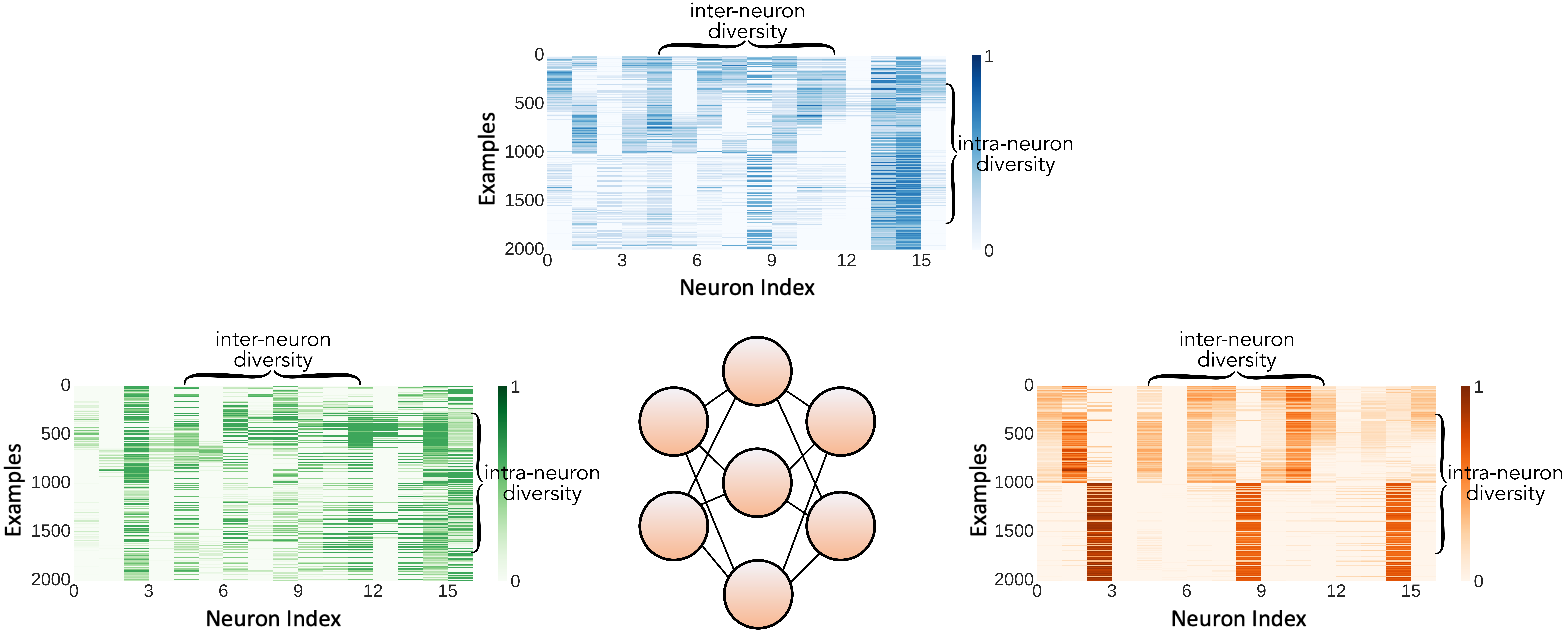


Example-level Memorization

Heuristic Memorization



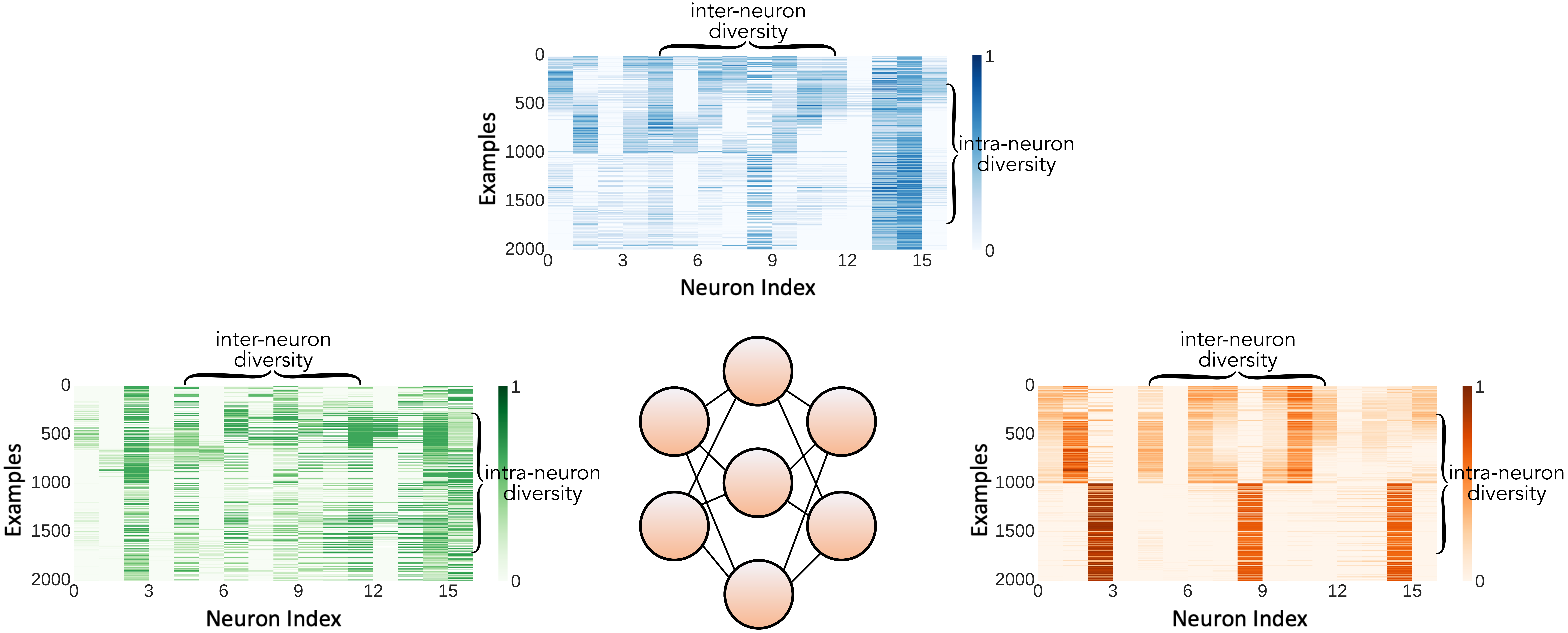
Hypothesis: Neuron diversity is reflective of model generalization



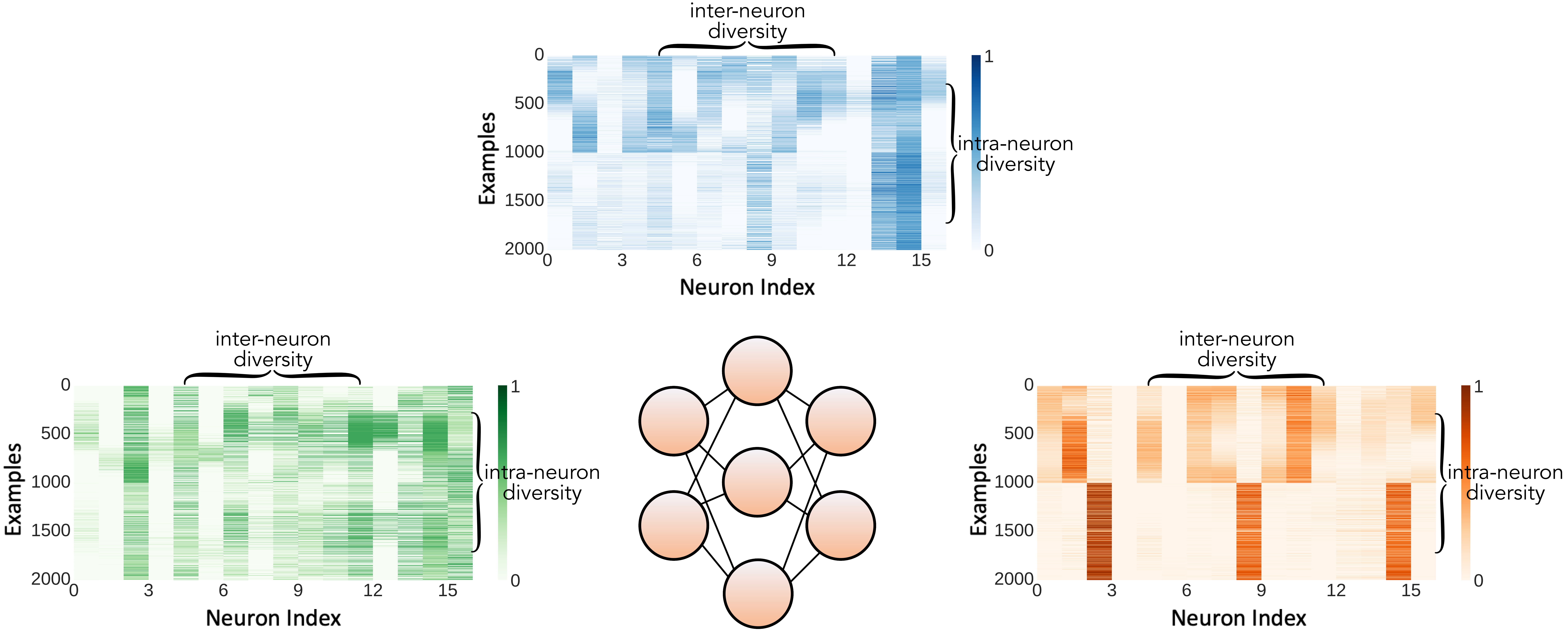
Hypothesis: Neuron diversity is reflective of model generalization

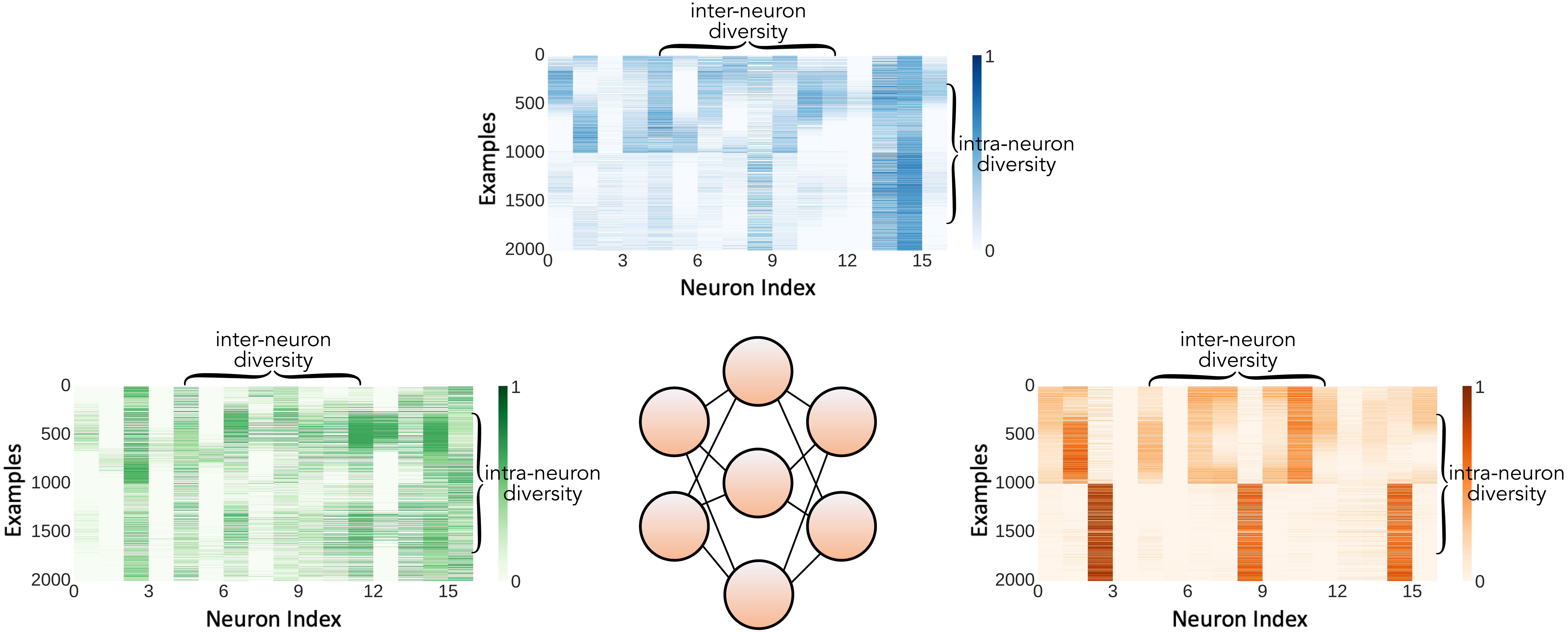
"*Evaluating* model *generalization*
using *intrinsic*,
information-theoretic metrics."

“*Evaluating* model *generalization* using *intrinsic*, *information-theoretic* metrics.”

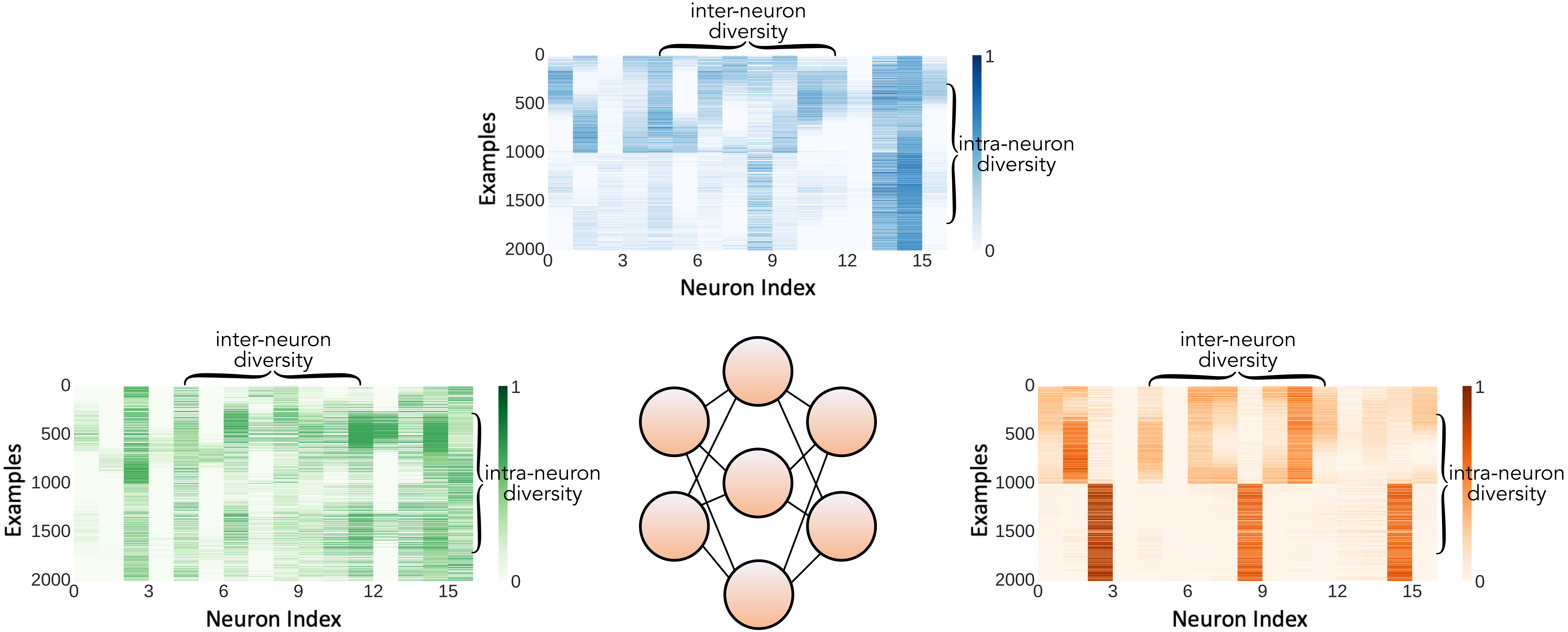


Hypothesis: Neuron diversity is reflective of model generalization


inter-neuron diversity
intra-neuron diversity
Examples
Neuron Index
0
500
1000
1500
2000
0
3
6
9
12
15
1
0
inter-neuron diversity
intra-neuron diversity
Examples
Neuron Index
inter-neuron diversity
intra-neuron diversity
Examples
Neuron Index



$$A_1, \ldots, A_N \leftarrow \{f(x_i)\}_{i=1}^{S}$$

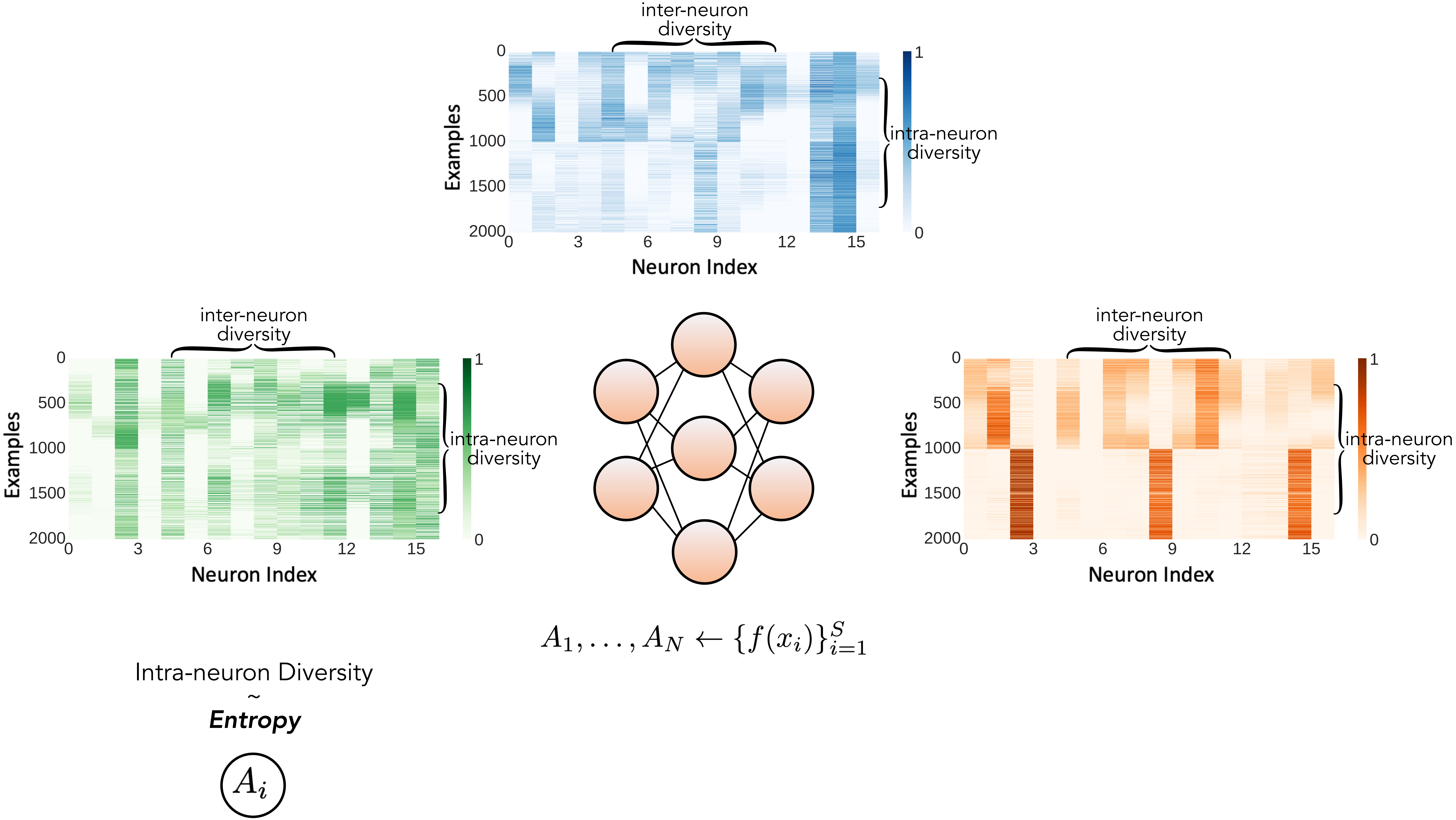


$$A_1, \ldots, A_N \leftarrow \{f(x_i)\}_{i=1}^{S}$$

Intra-neuron Diversity

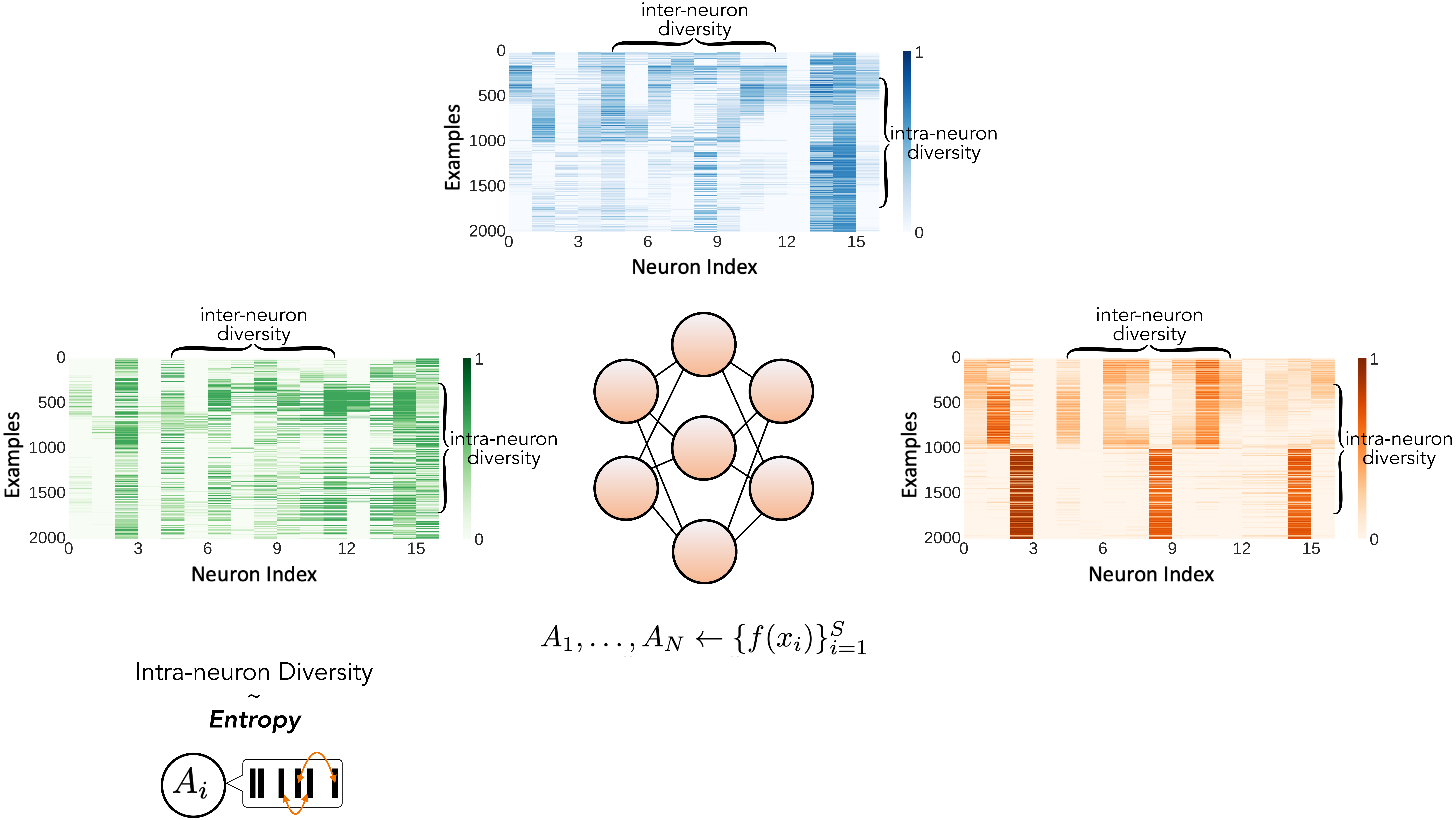
~

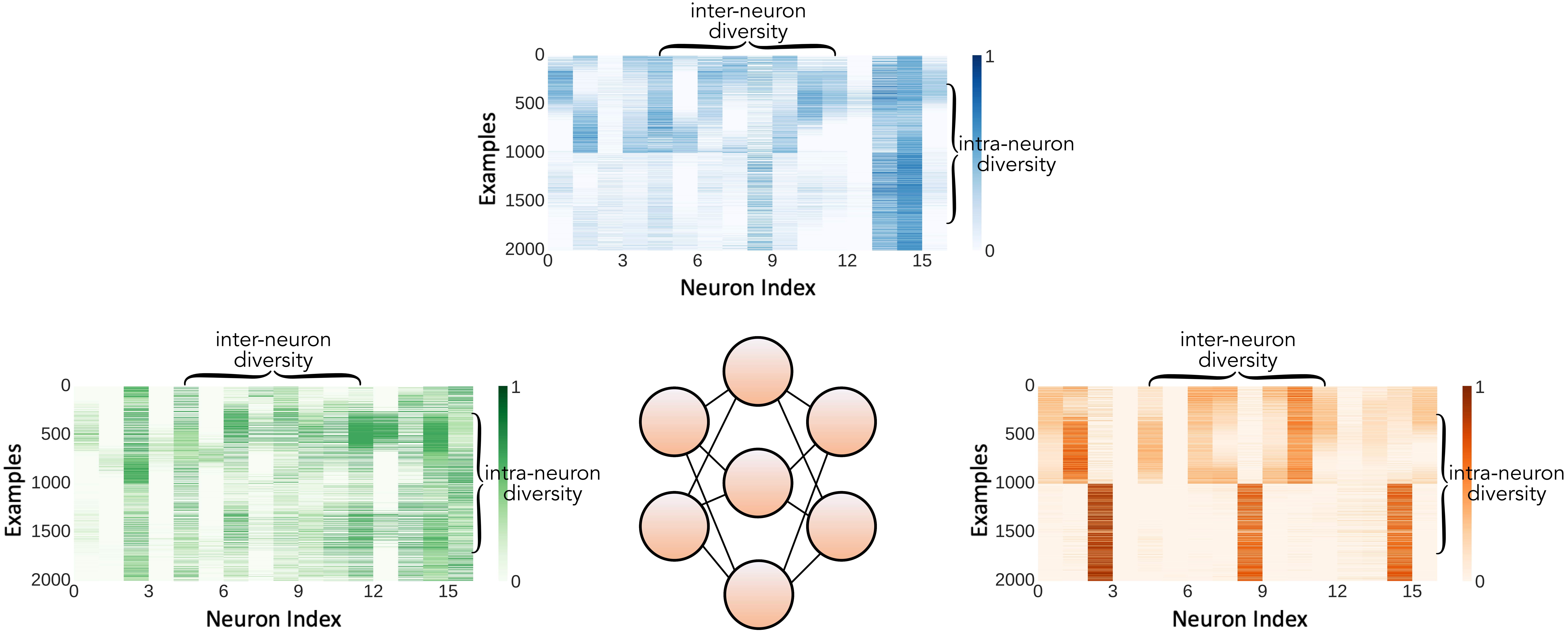
Entropy



$$A_1, \ldots, A_N \leftarrow \{f(x_i)\}_{i=1}^{S}$$

Intra-neuron Diversity
~
Entropy


inter-neuron diversity
intra-neuron diversity
Examples
Neuron Index
0
500
1000
1500
2000
0
3
6
9
12
15
1
0
$A_1, \ldots, A_N \leftarrow \{f(x_i)\}_{i=1}^{S}$
Intra-neuron Diversity
~
Entropy
A_i

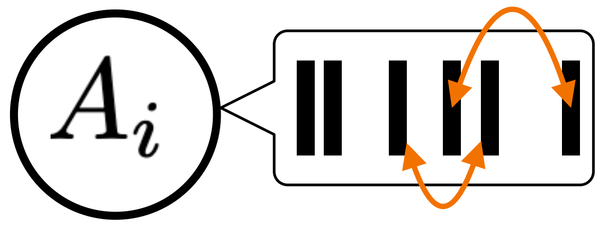


$$A_1, \ldots, A_N \leftarrow \{f(x_i)\}_{i=1}^{S}$$

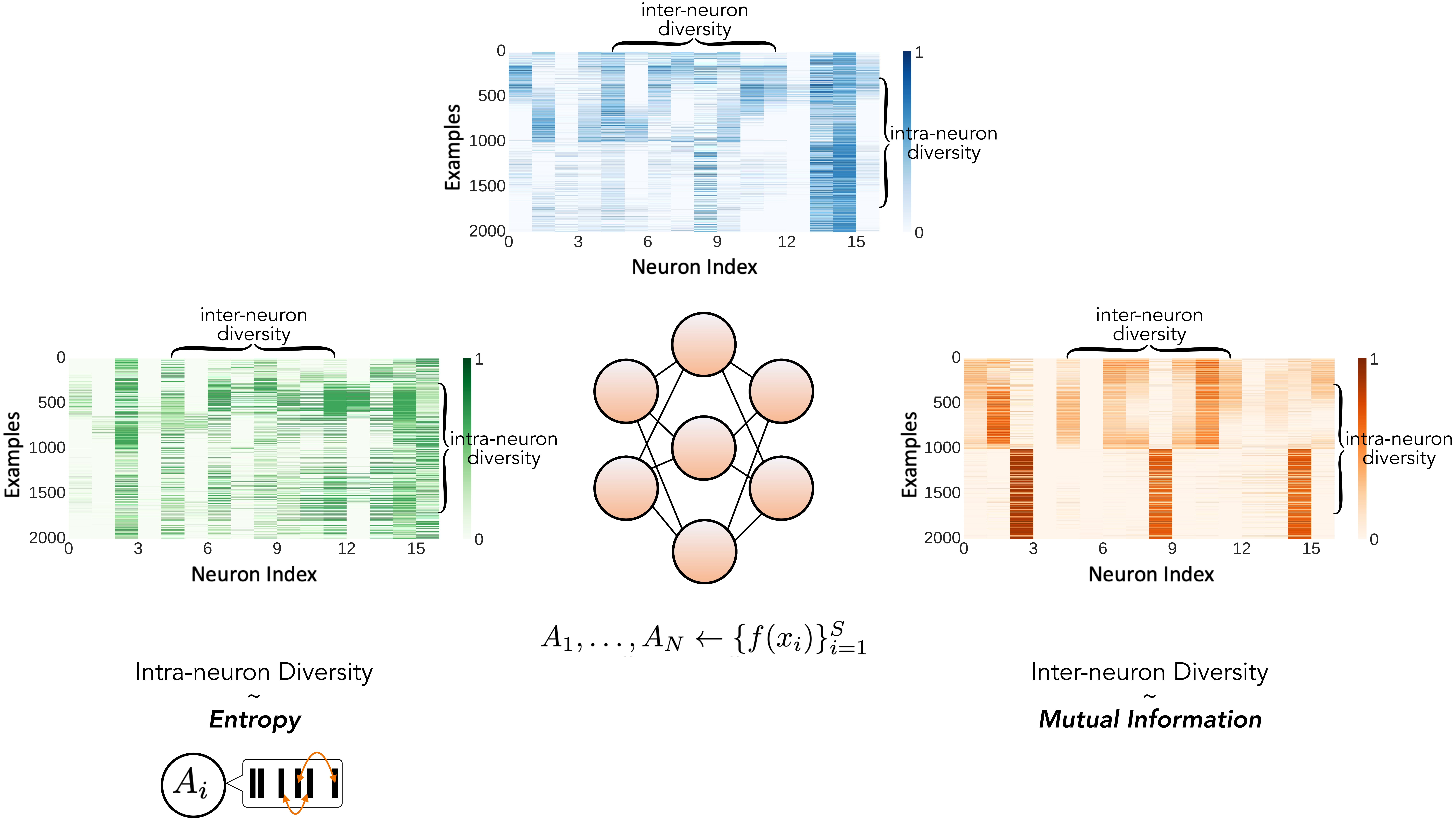
Intra-neuron Diversity

~

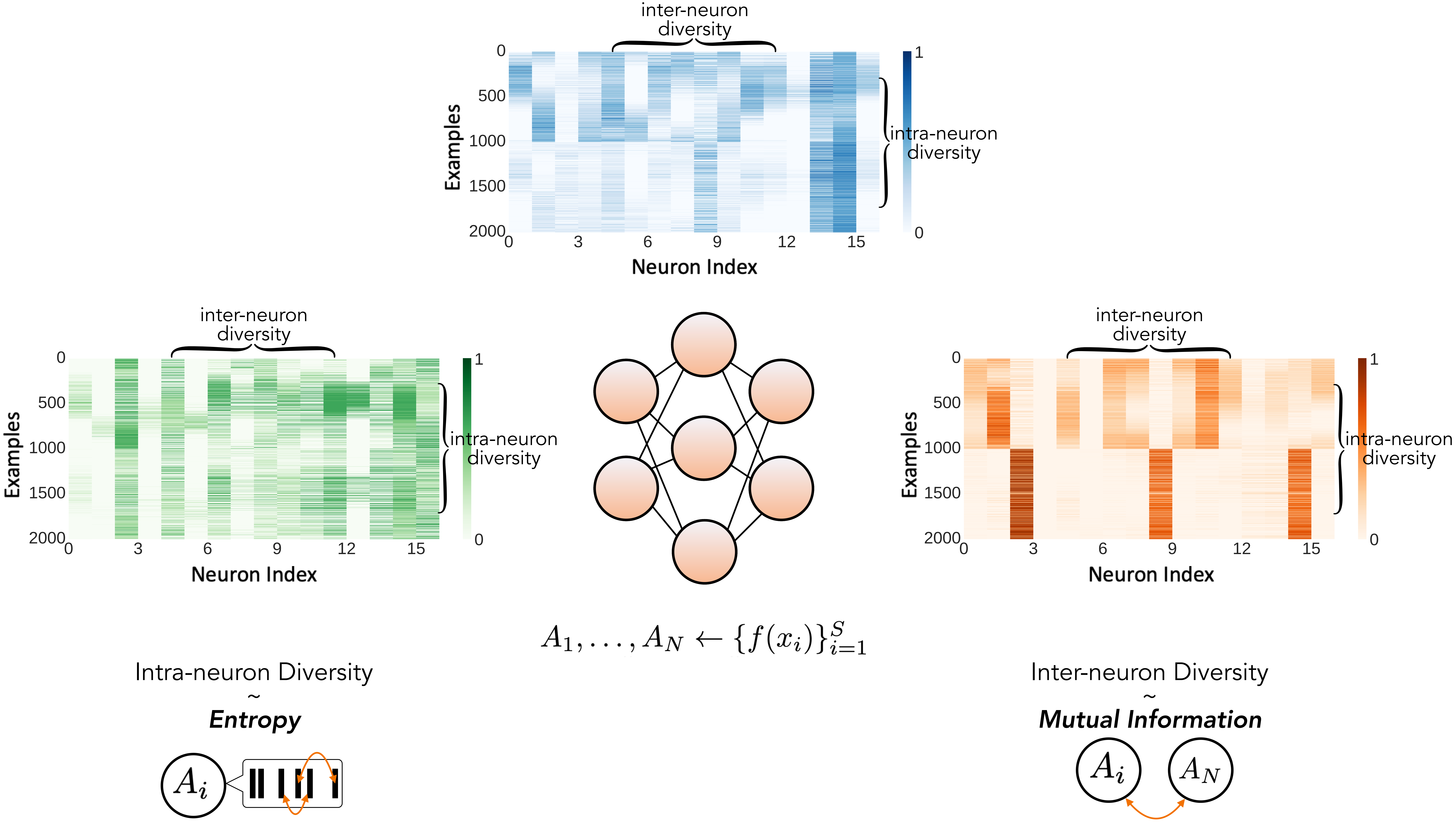
Entropy



$$H(A_i) = \mathop{\mathbb{E}}_{\hat{a}_i^s \in A_i}[h(\hat{a}_i^s)] = \sum_{j=1}^{N_{\text{bins}}} p(\hat{a}_i^j) \log(\frac{1}{p(\hat{a}_i^j)})$$



$$H(A_i) = \mathop{\mathbb{E}}_{\hat{a}_i^s \in A_i}[h(\hat{a}_i^s)] = \sum_{j=1}^{N_{\text{bins}}} p(\hat{a}_i^j) \log(\frac{1}{p(\hat{a}_i^j)})$$



$$A_1, \ldots, A_N \leftarrow \{f(x_i)\}_{i=1}^{S}$$

Intra-neuron Diversity

~

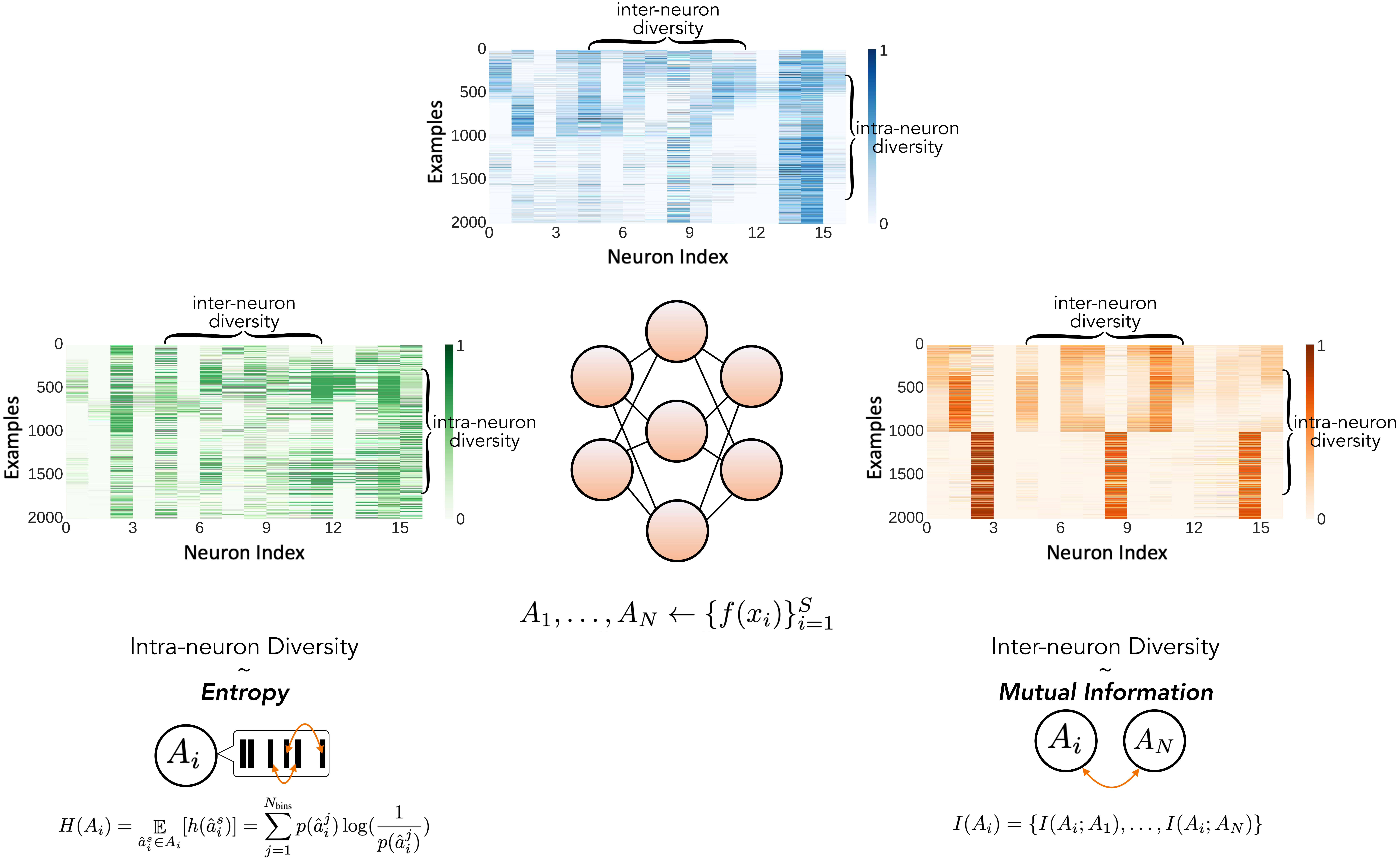
Entropy

Inter-neuron Diversity

~

Mutual Information

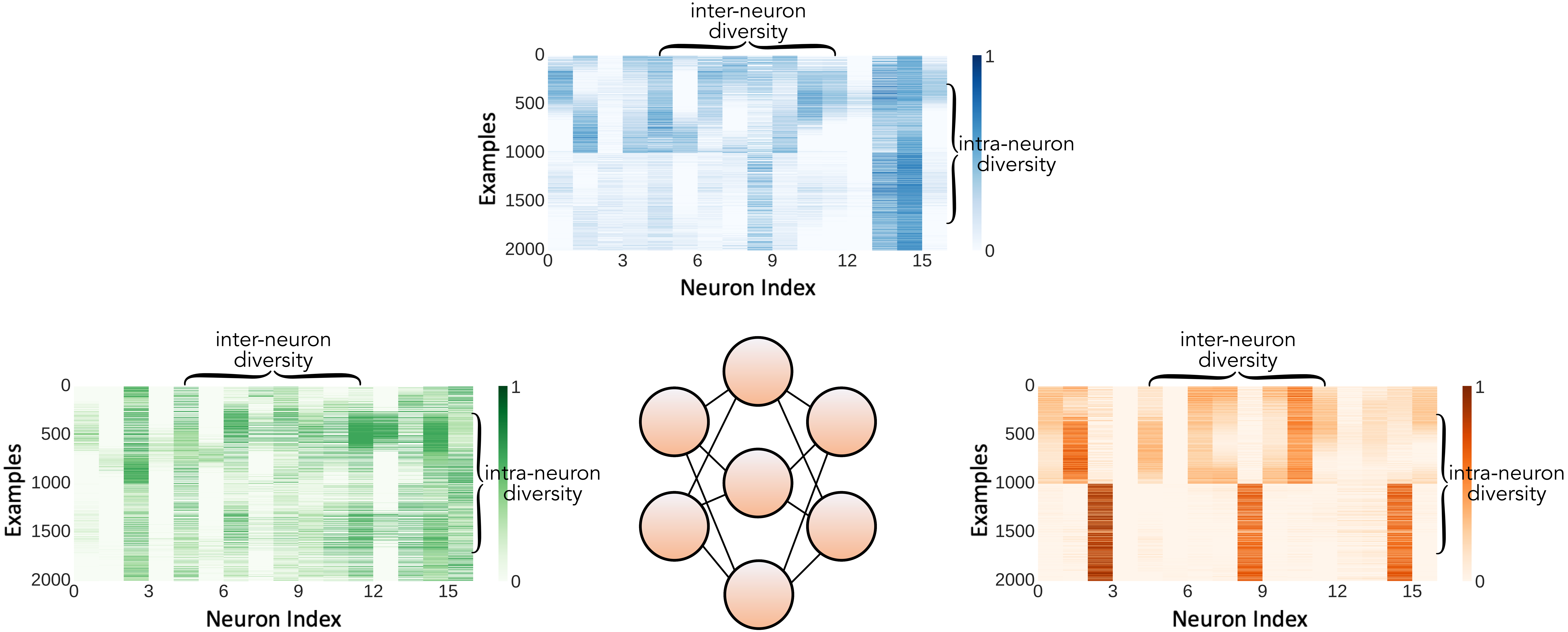
$$H(A_i) = \mathop{\mathbb{E}}_{\hat{a}_i^s \in A_i}[h(\hat{a}_i^s)] = \sum_{j=1}^{N_{\text{bins}}} p(\hat{a}_i^j) \log(\frac{1}{p(\hat{a}_i^j)})$$



$$A_1, \ldots, A_N \leftarrow \{f(x_i)\}_{i=1}^{S}$$

$$H(A_i) = \mathop{\mathbb{E}}_{\hat{a}_i^s \in A_i}[h(\hat{a}_i^s)] = \sum_{j=1}^{N_{\text{bins}}} p(\hat{a}_i^j) \log(\frac{1}{p(\hat{a}_i^j)})$$

$$I(A_i) = \{I(A_i; A_1), \ldots, I(A_i; A_N)\}$$

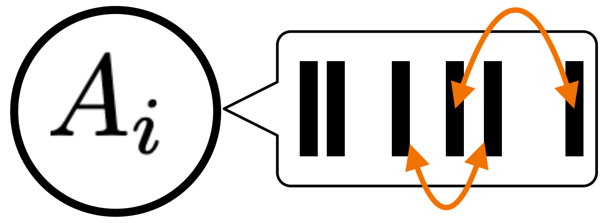


$$A_1, \ldots, A_N \leftarrow \{f(x_i)\}_{i=1}^{S}$$

Intra-neuron Diversity

~

Entropy



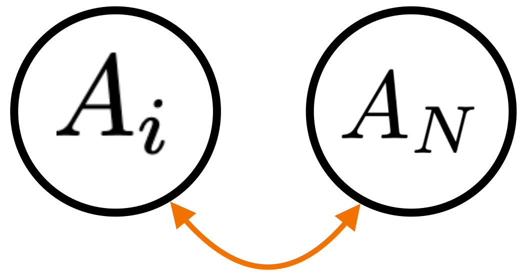
$$H(A_i) = \mathop{\mathbb{E}}_{\hat{a}_i^s \in A_i}[h(\hat{a}_i^s)] = \sum_{j=1}^{N_{\text{bins}}} p(\hat{a}_i^j) \log(\frac{1}{p(\hat{a}_i^j)})$$

Memorization	Diversity	
	Intra-neuron ($\propto$ Entropy)	Inter-neuron ($\propto$ MI^{-1})
Heuristic	↓	↓
Example-level	↑	↑

Inter-neuron Diversity

~

Mutual Information



$$I(A_i) = \{I(A_i; A_1), \ldots, I(A_i; A_N)\}$$

Some experimental results

3-layered MLP on Colored MNIST

Hypothesis:

↓ Entropy

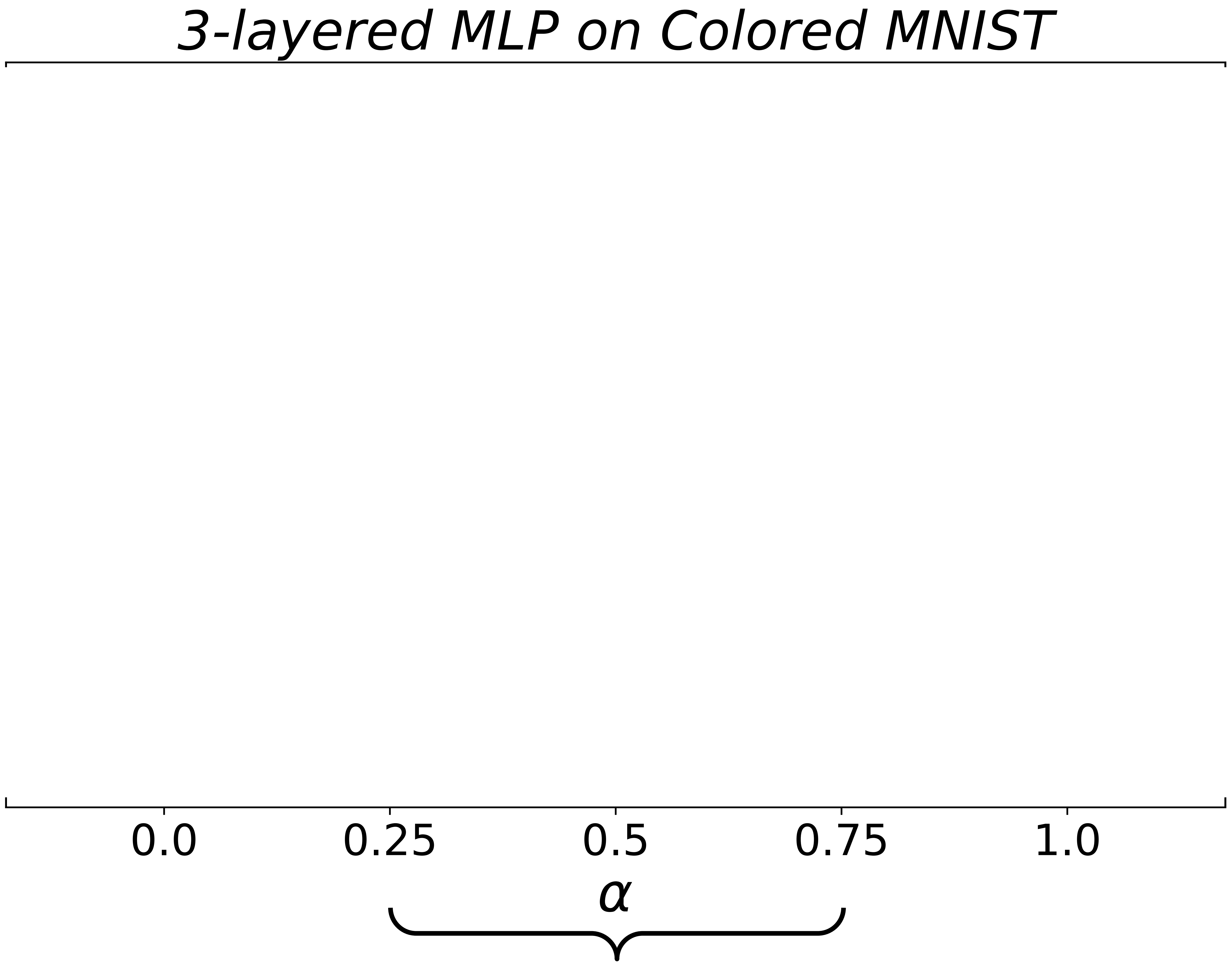
↑ MI

3-layered MLP on Colored MNIST

Hypothesis:

↓ Entropy

↑ MI

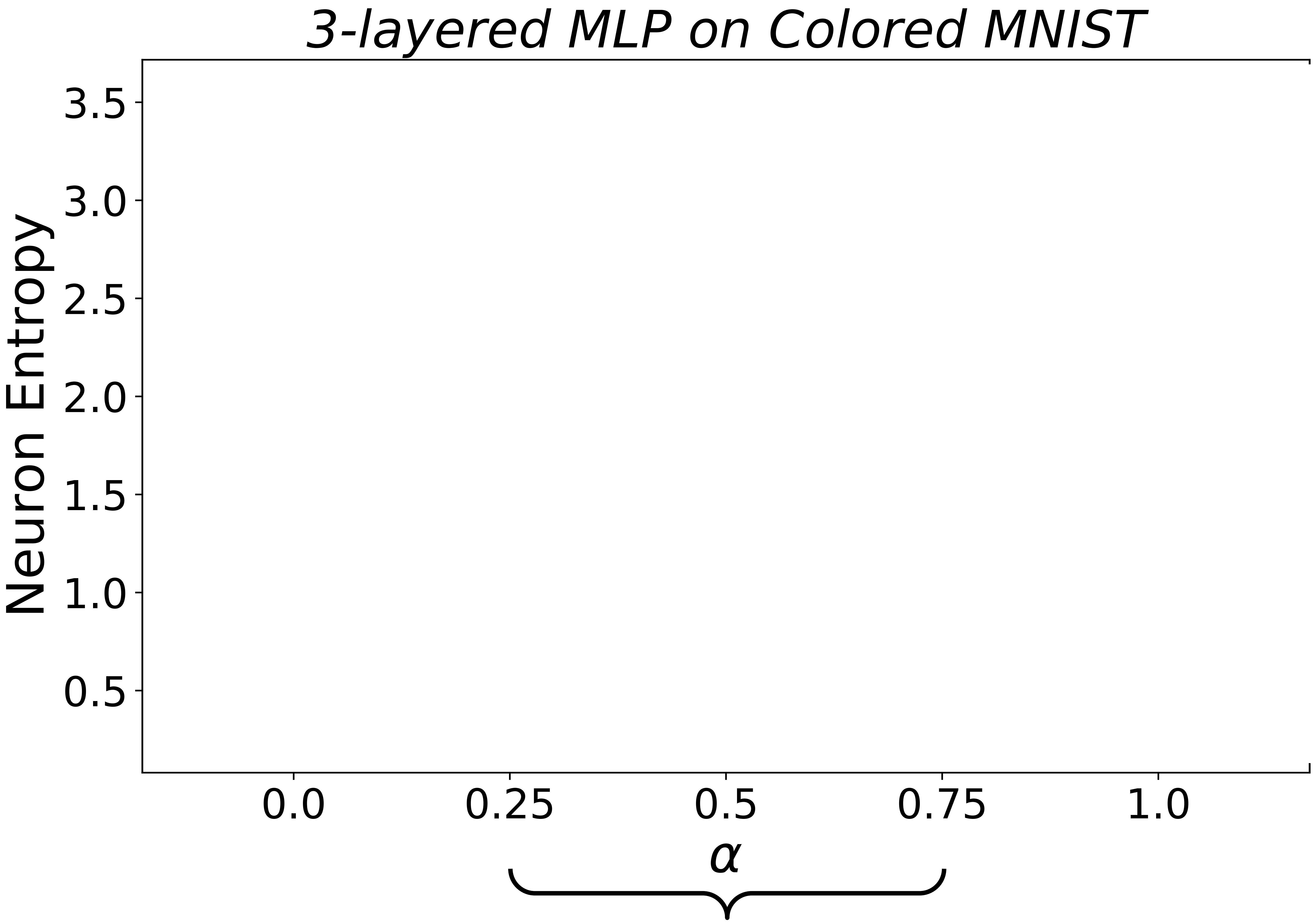


"proportion of the labels synthetically
changed to follow the spurious correlation"

Hypothesis:

↓ Entropy

↑ MI

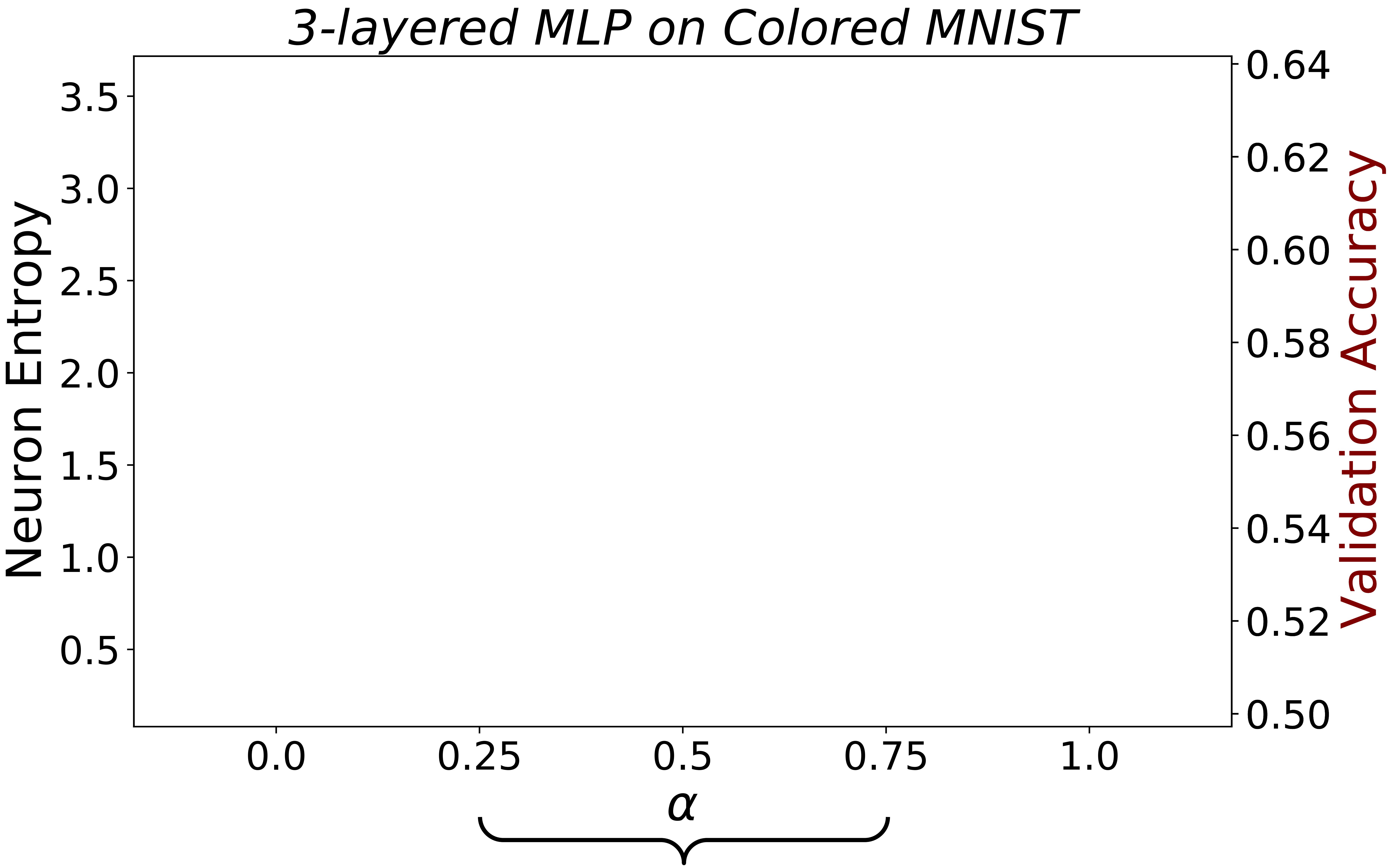


"proportion of the labels synthetically
changed to follow the spurious correlation"

Hypothesis:

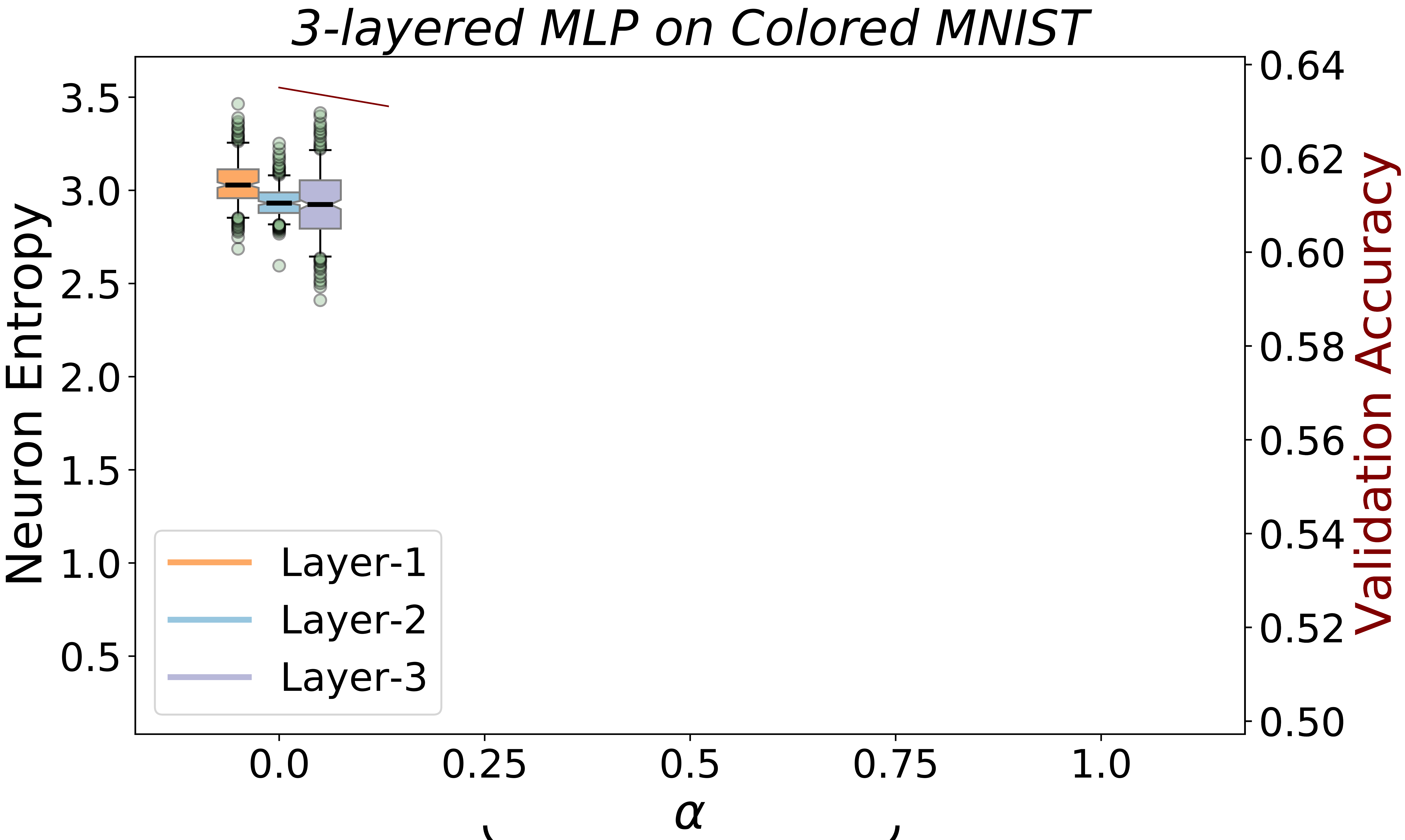
↓ Entropy

↑ MI



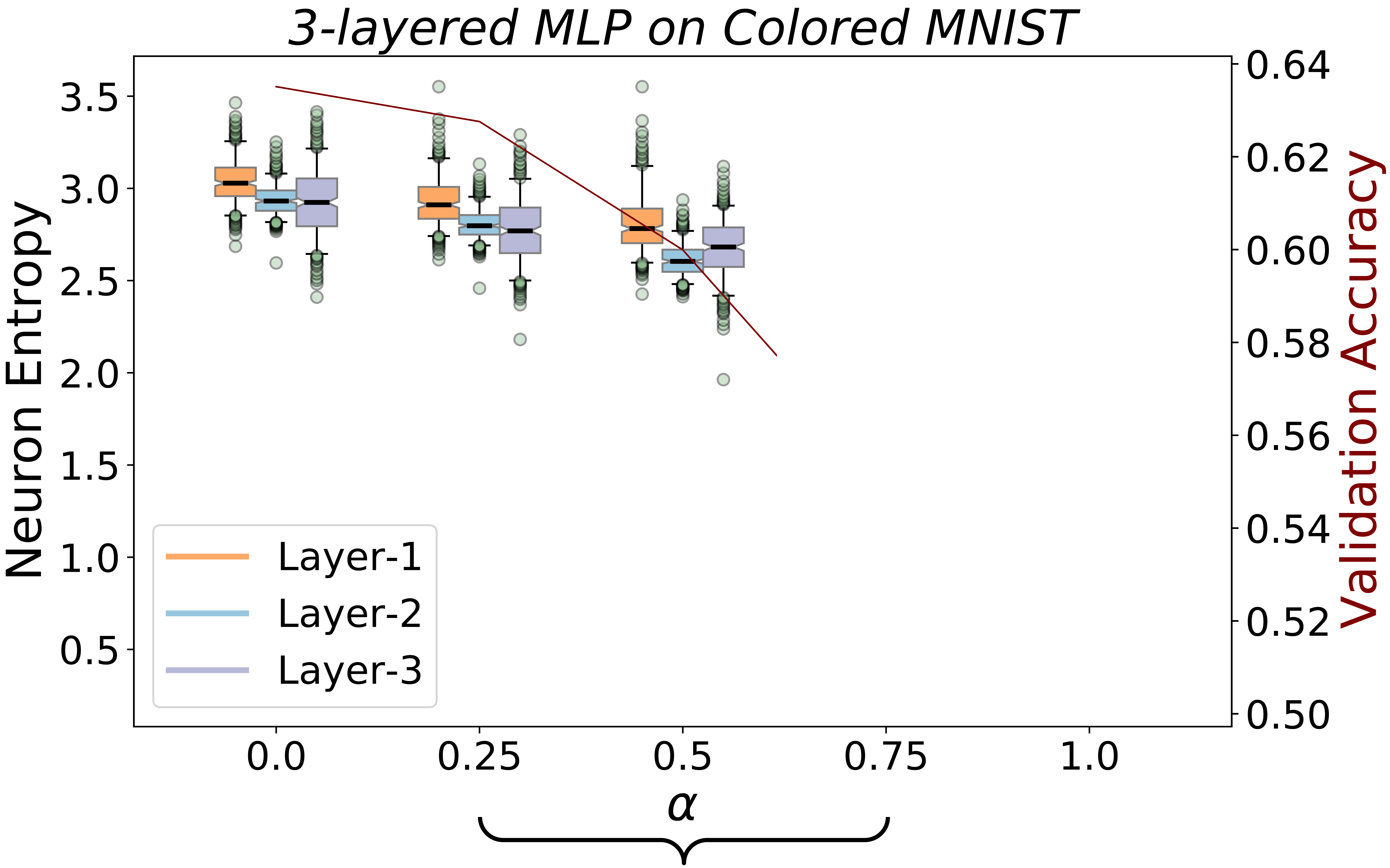
"proportion of the labels synthetically
changed to follow the spurious correlation"

Hypothesis:
↓ Entropy
↑ MI



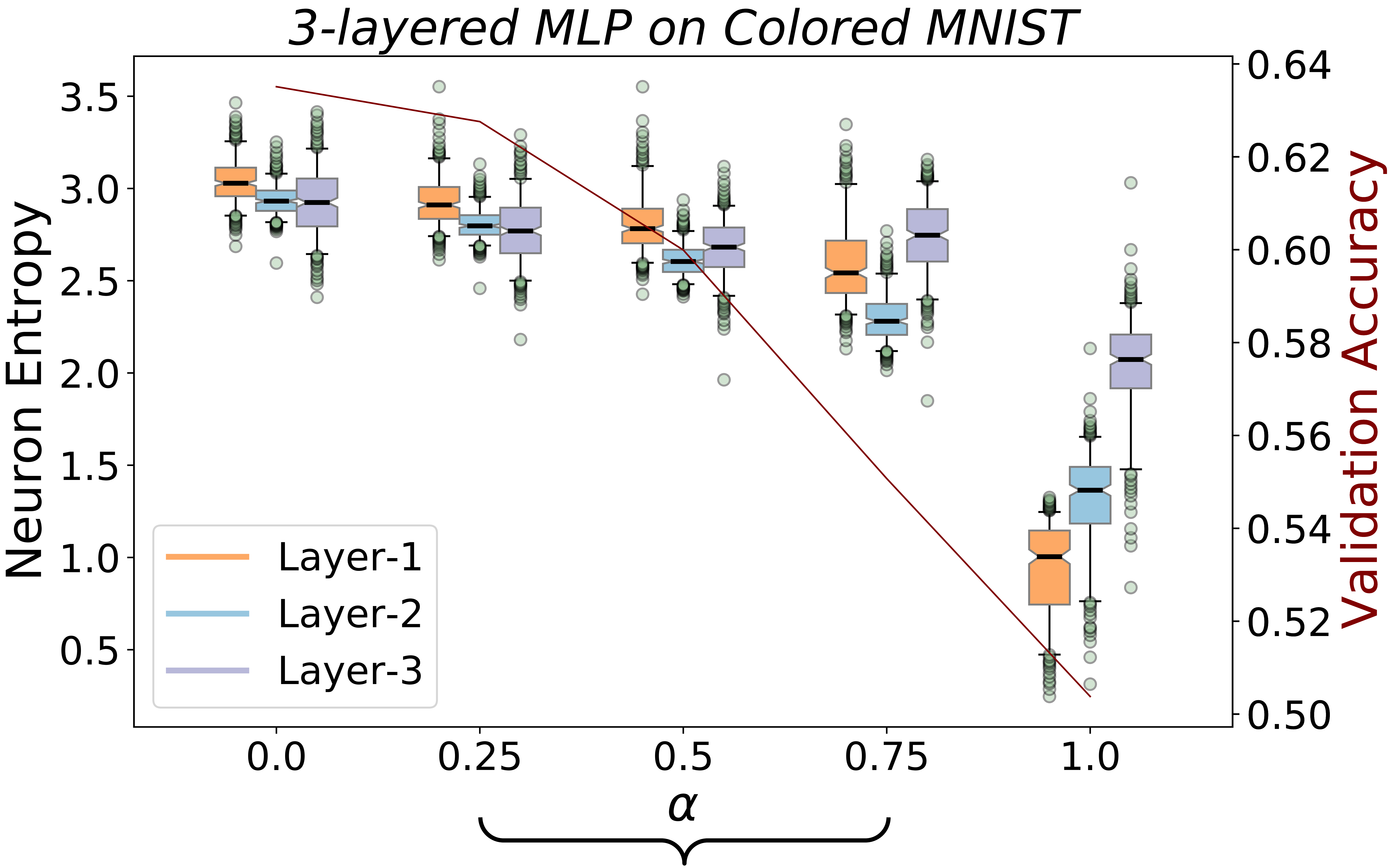
"proportion of the labels synthetically
changed to follow the spurious correlation"

Hypothesis:
↓ Entropy
↑ MI



"proportion of the labels synthetically
changed to follow the spurious correlation"

Hypothesis:
↓ Entropy
↑ MI



"proportion of the labels synthetically
changed to follow the spurious correlation"

Hypothesis:

↑ Entropy

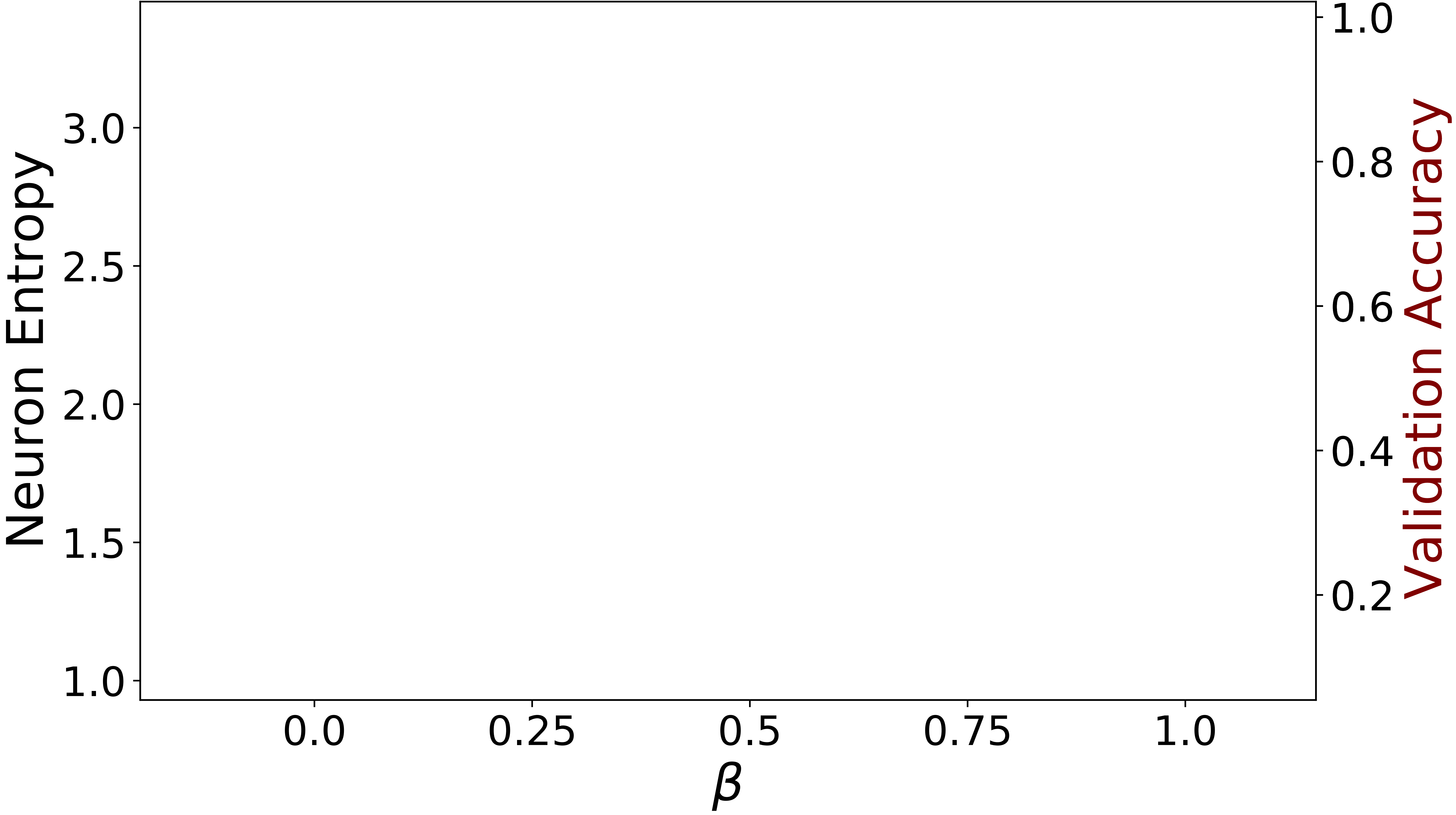
↓ MI

Hypothesis:

↑ Entropy

↓ MI

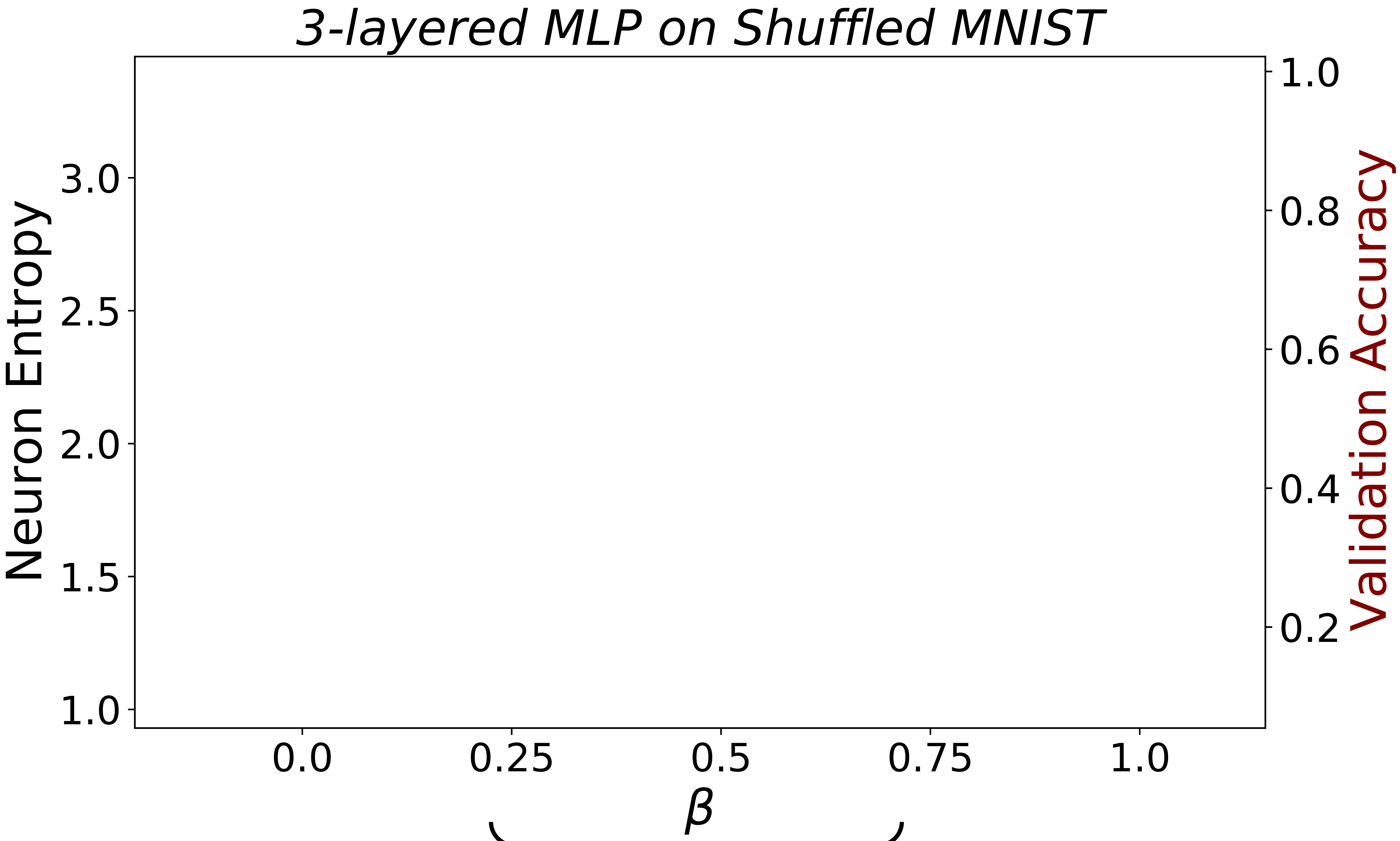
3-layered MLP on Shuffled MNIST



Hypothesis:

↑ Entropy

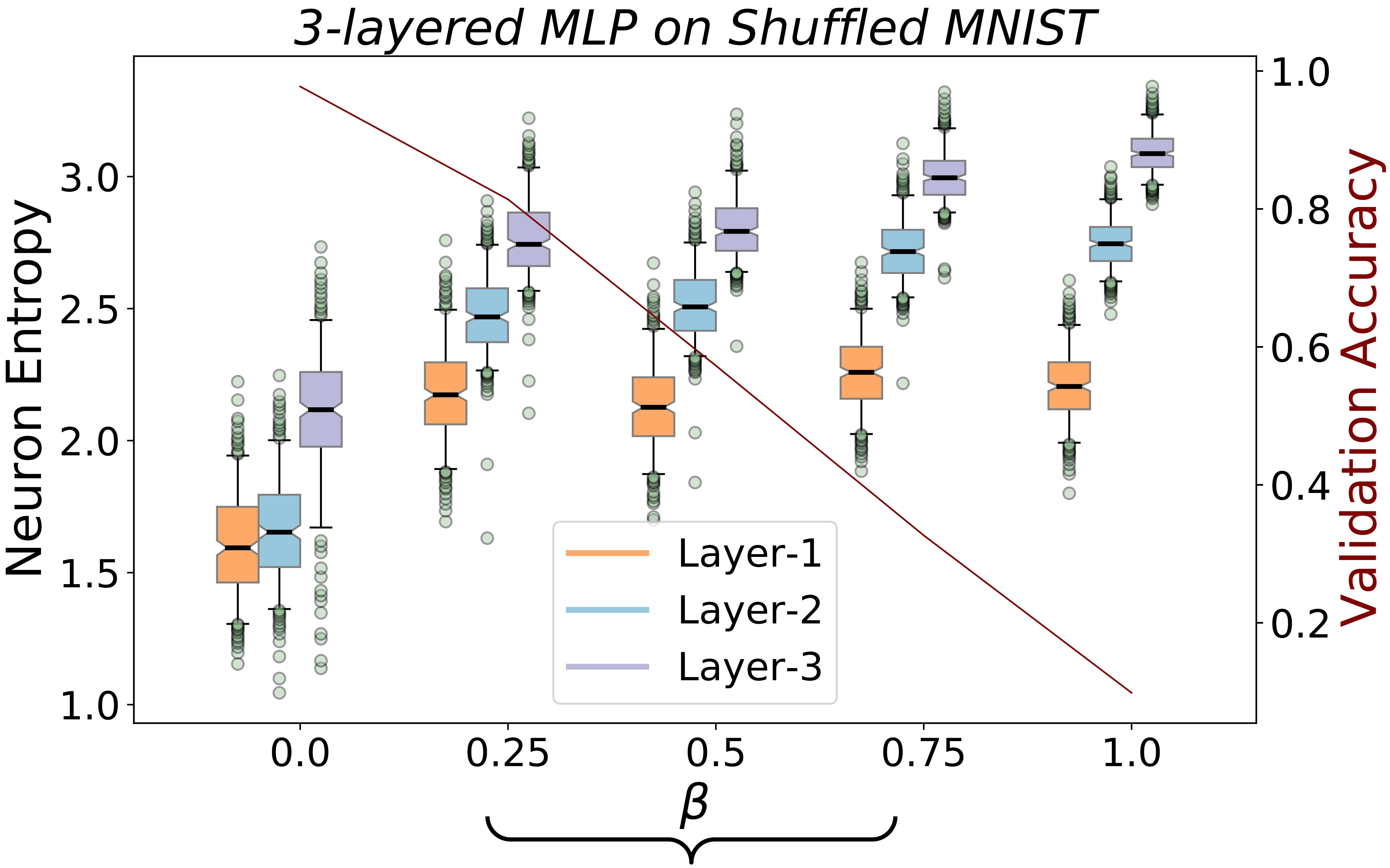
↓ MI



3-layered MLP on Shuffled MNIST

"proportion of the label noise
added to the training dataset”

Hypothesis:
↑ Entropy
↓ MI

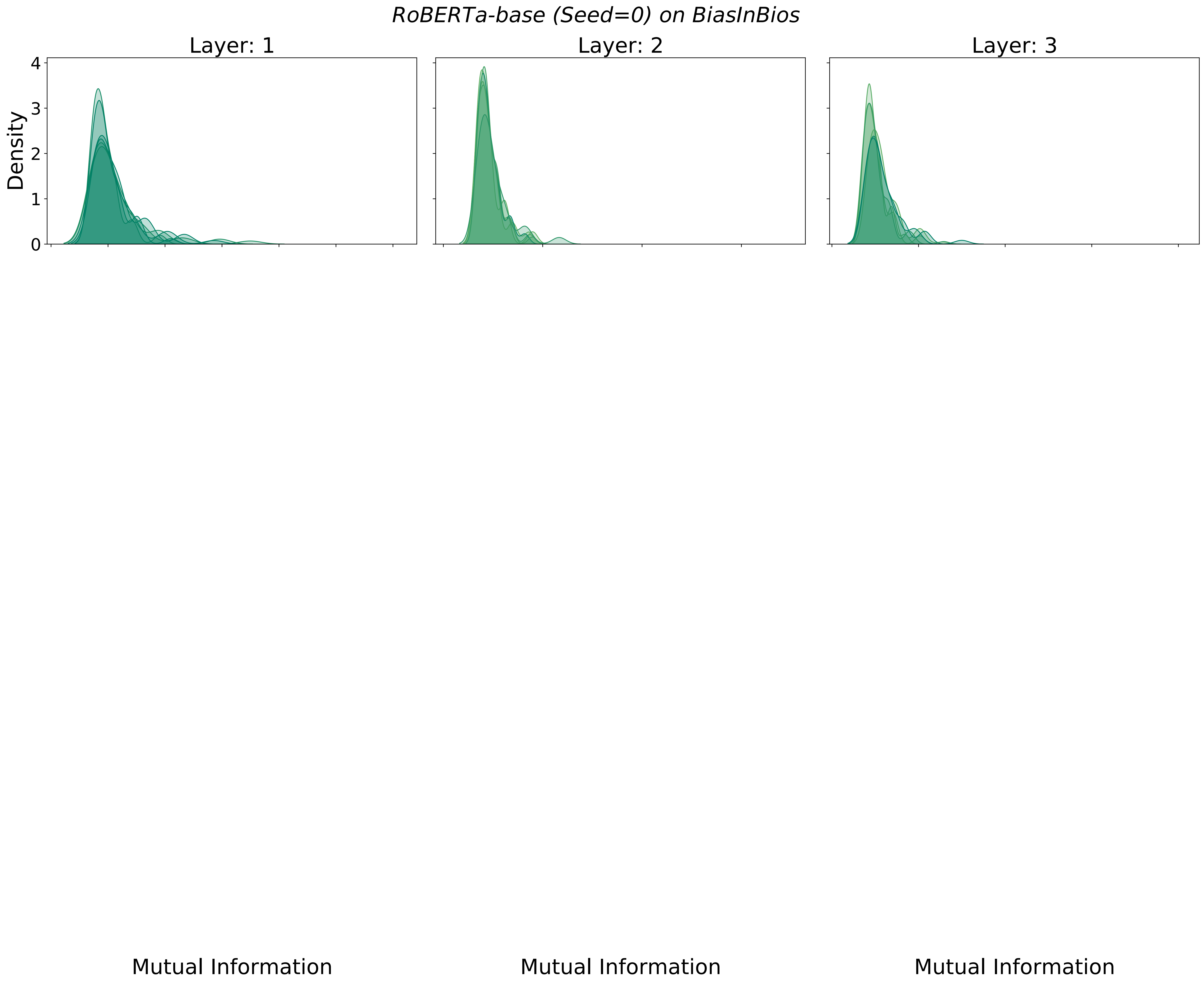


"proportion of the label noise added to the training dataset"

Take aways

Take aways

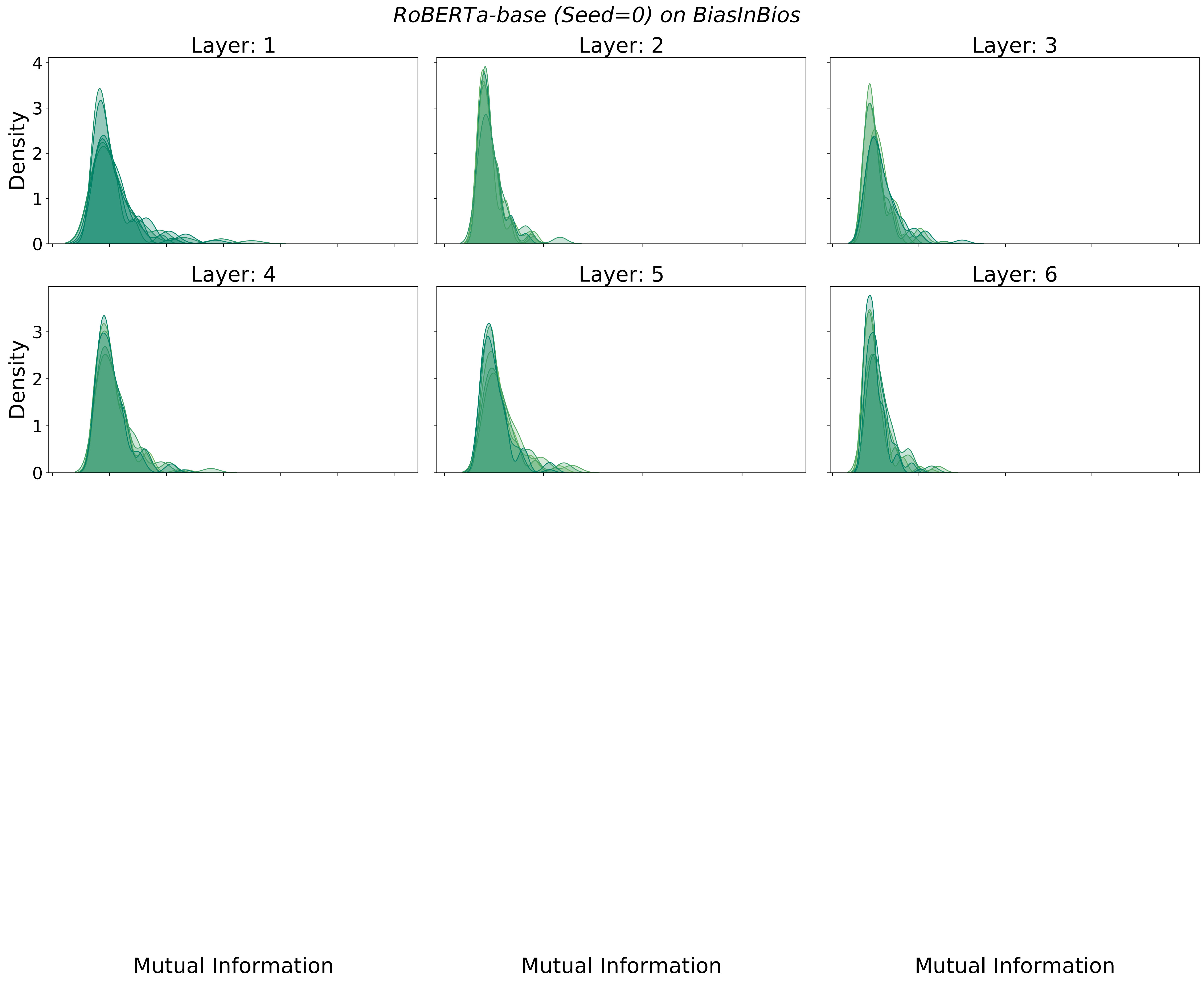
- Intrinsic activation patterns in neural networks are indicative of memorization.



Take aways

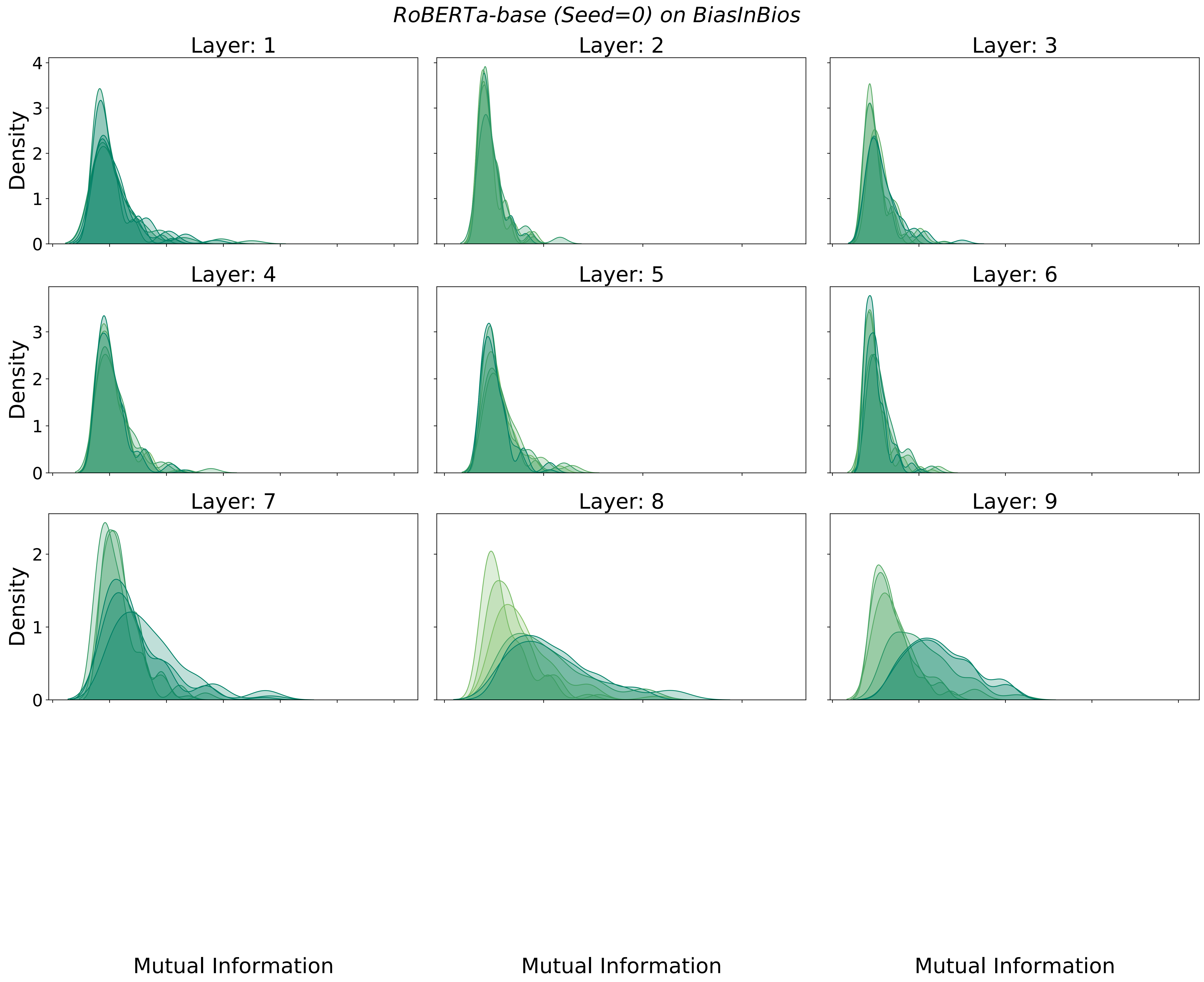
- Intrinsic activation patterns in neural networks are indicative of memorization.

- Neuron diversity (as a measure of information) correlates to generalization.



Take aways

- Intrinsic activation patterns in neural networks are indicative of memorization.
- Neuron diversity (as a measure of information) correlates to generalization.
- Useful for model selection. Potentially extensible for OOD detection, model regularization, and understanding training dynamics.

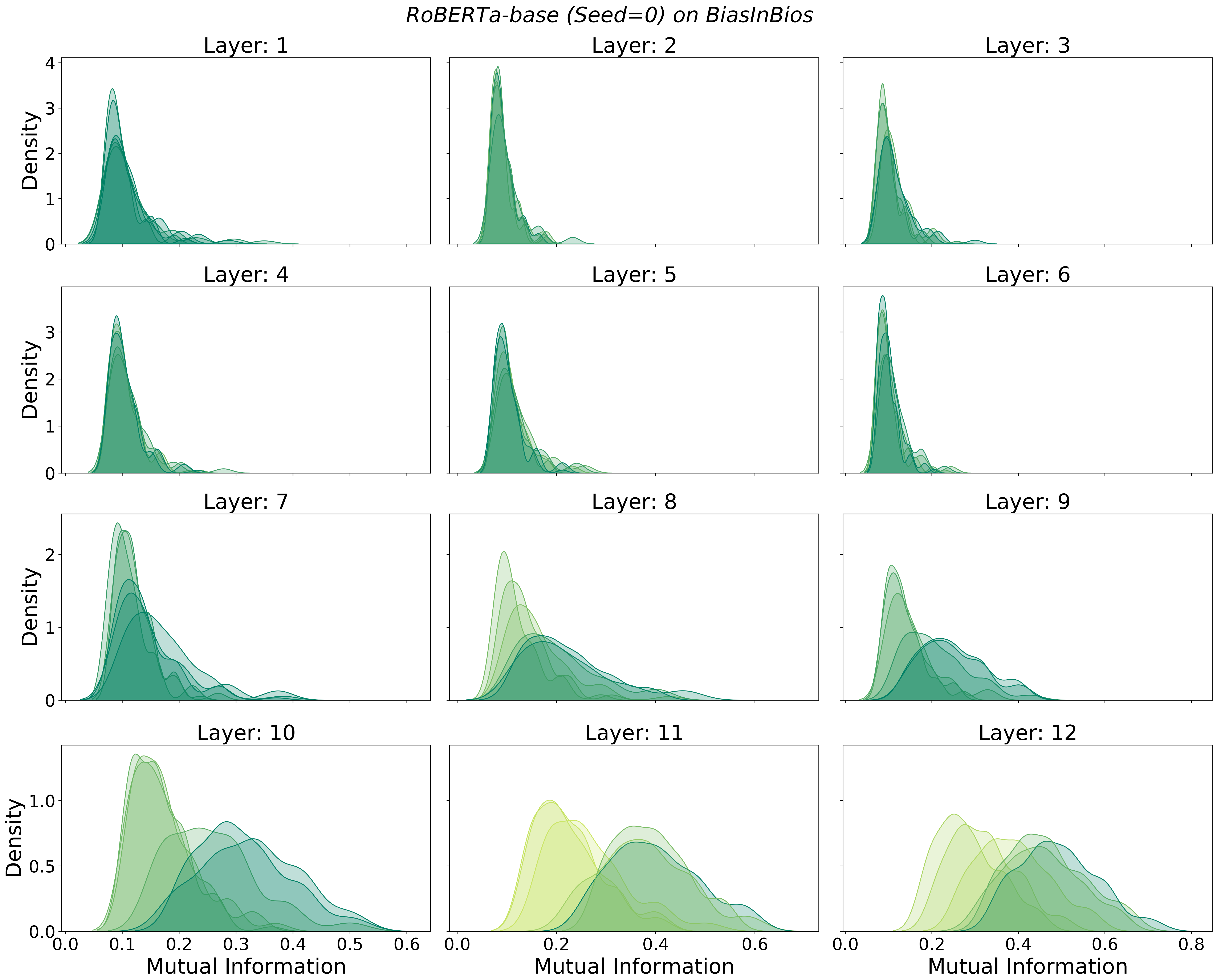


Take aways

- Intrinsic activation patterns in neural networks are indicative of memorization.

- Neuron diversity (as a measure of information) correlates to generalization.

- Useful for model selection. Potentially extensible for OOD detection, model regularization, and understanding training dynamics.


RoBERTa-base (Seed=0) on BiasInBios
Layer: 1
Layer: 2
Layer: 3
Layer: 4
Layer: 5
Layer: 6
Layer: 7
Layer: 8
Layer: 9
Layer: 10
Layer: 11
Layer: 12
Density
Mutual Information