Emergent Equivariance in Deep Ensembles

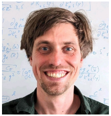
Jan E. Gerken*

CHALMERS UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

in collaboration with



Pan Kessel*

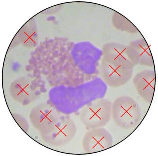
from Prescient Design A Genentech Accelerator

* Equal Contribution

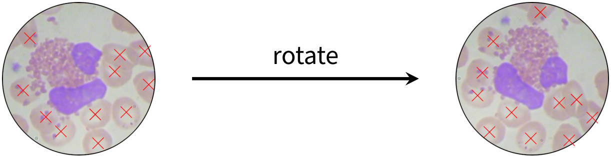
ICML International Conference On Machine Learning

Equivariance

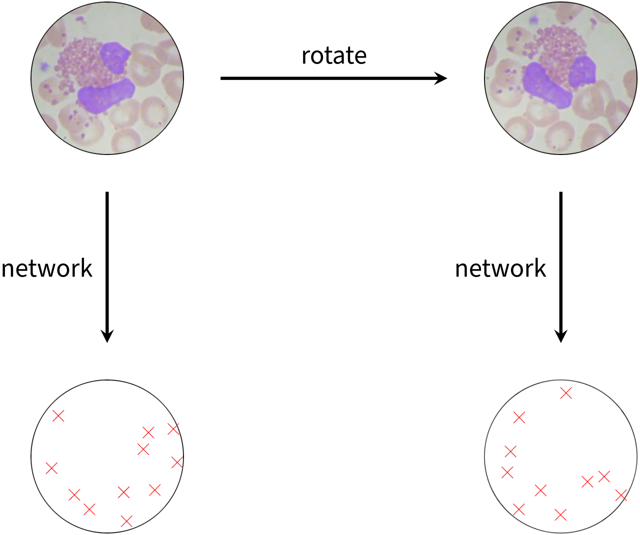
Equivariance



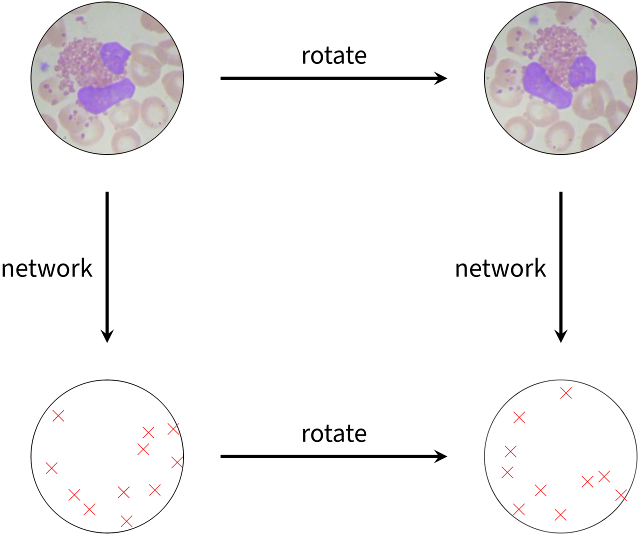
Equivariance



Equivariance

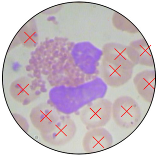


Equivariance

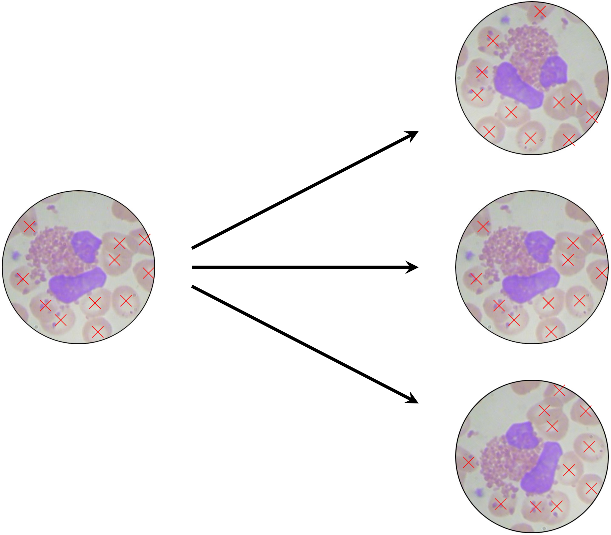


Data augmentation

Data augmentation



Data augmentation



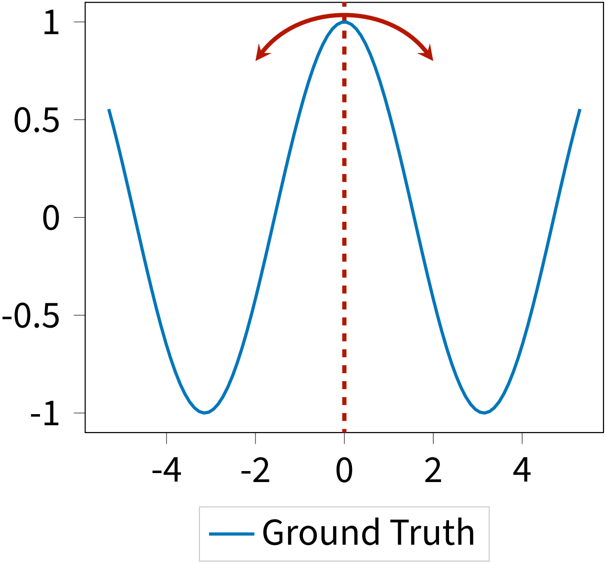
Data augmentation

- 👍 Easy to implement
- 👍 No specialized architecture necessary

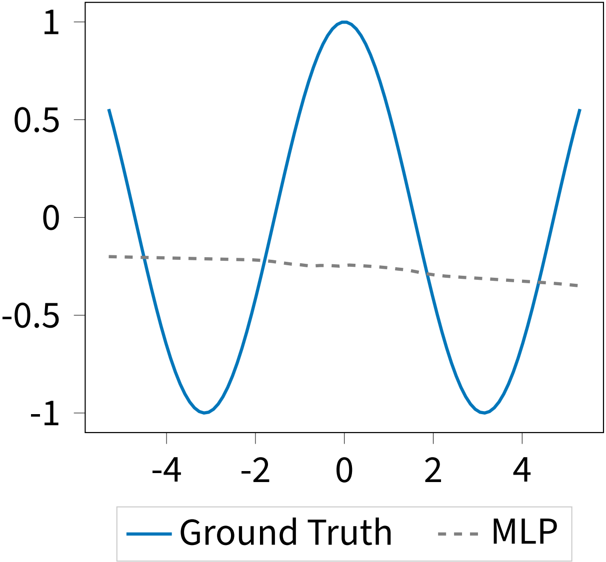
Data augmentation

- 👍 Easy to implement
- 👍 No specialized architecture necessary
- 👎 No exact equivariance

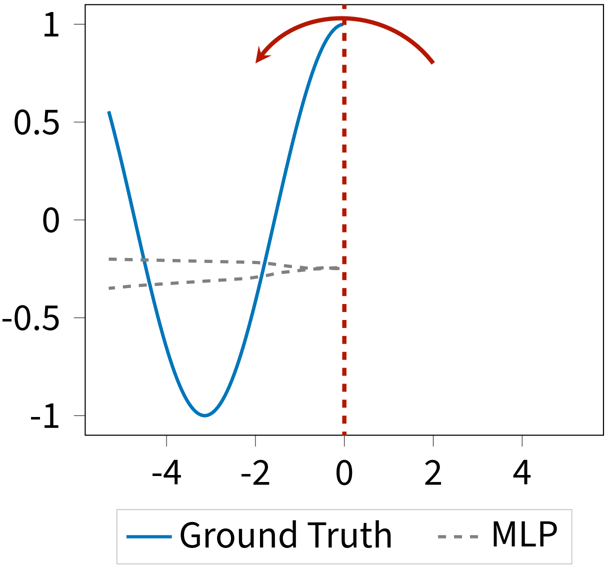
Toy example

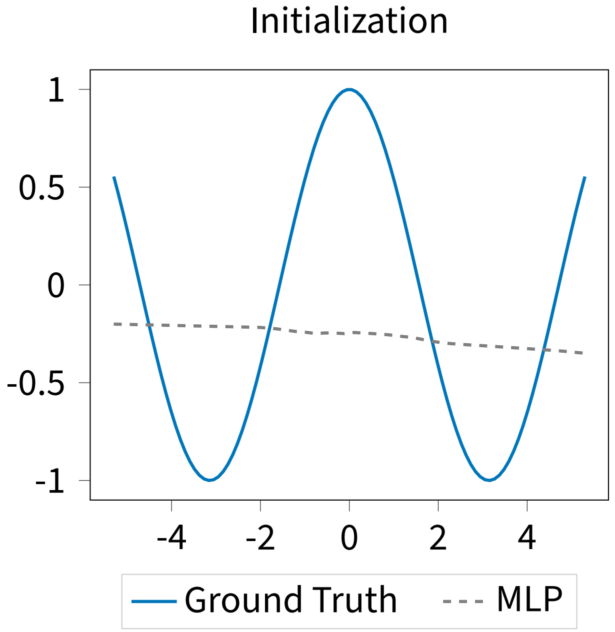

1
0.5
0
-0.5
-1
-4
-2
0
2
4
Ground Truth

Initialization

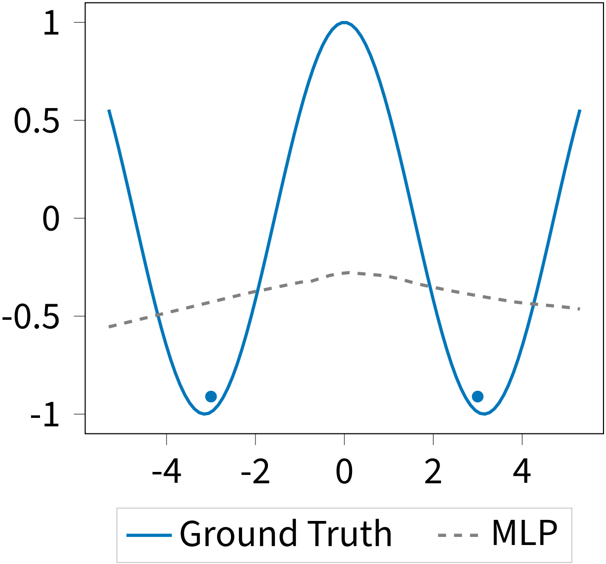


Initialization

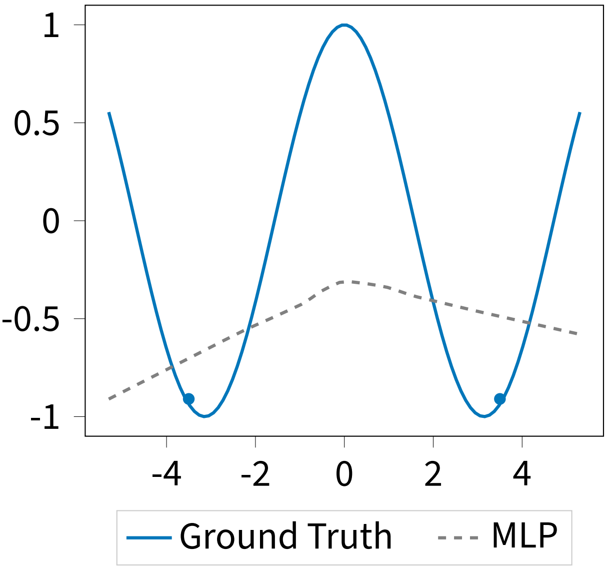



Initialization
1
0.5
0
-0.5
-1
-4
-2
0
2
4
Ground Truth
MLP

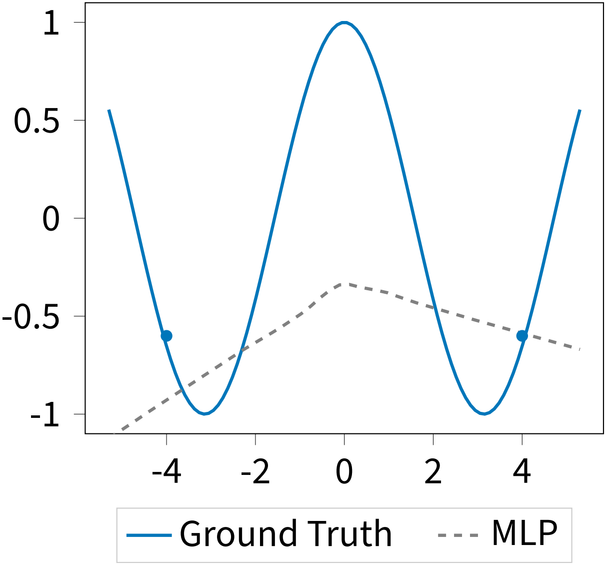
After 1 Training Step



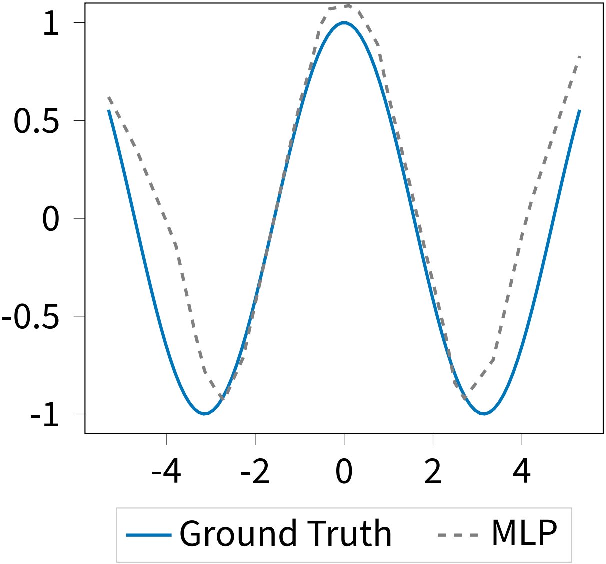
After 2 Training Steps



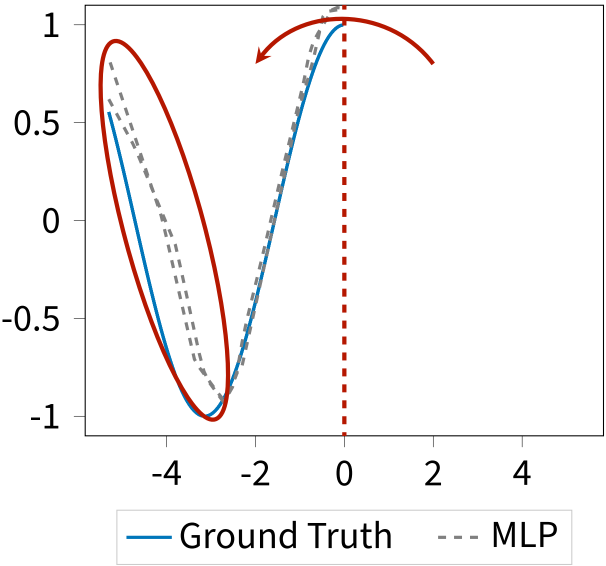
After 3 Training Steps



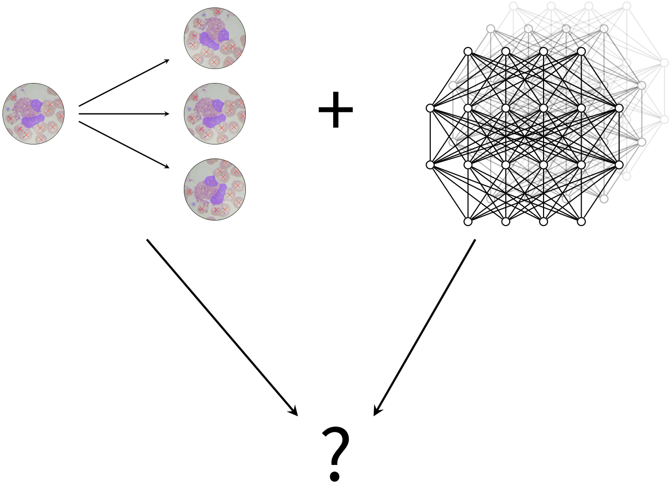
After 2000 Training Steps



After 2000 Training Steps



Can ensembles help?



Main conclusion

Deep ensembles trained with data augmentation are equivariant.

Main conclusion

Deep ensembles trained with data augmentation are equivariant.

✓ Proof of exact equivariance for
  - full data augmentation
  - infinite ensembles
  - at infinite width

Main conclusion

Deep ensembles trained with data augmentation are equivariant.

- ✓ Proof of exact equivariance for
  - full data augmentation
  - infinite ensembles
  - at infinite width
- ✓ Equivariance holds for all training times

Main conclusion

Deep ensembles trained with data augmentation are equivariant.

- ✓ Proof of exact equivariance for
  - full data augmentation
  - infinite ensembles
  - at infinite width
- ✓ Equivariance holds for all training times
- ✓ Equivariance holds away from the training data

Intuitive explanation

- ✓ Equivariance holds for all training times
- ✓ Equivariance holds away from the training data

Intuitive explanation

- ✓ Equivariance holds for all training times
- ✓ Equivariance holds away from the training data

ⓘ At infinite width, the mean output at initialization is zero everywhere.

Intuitive explanation

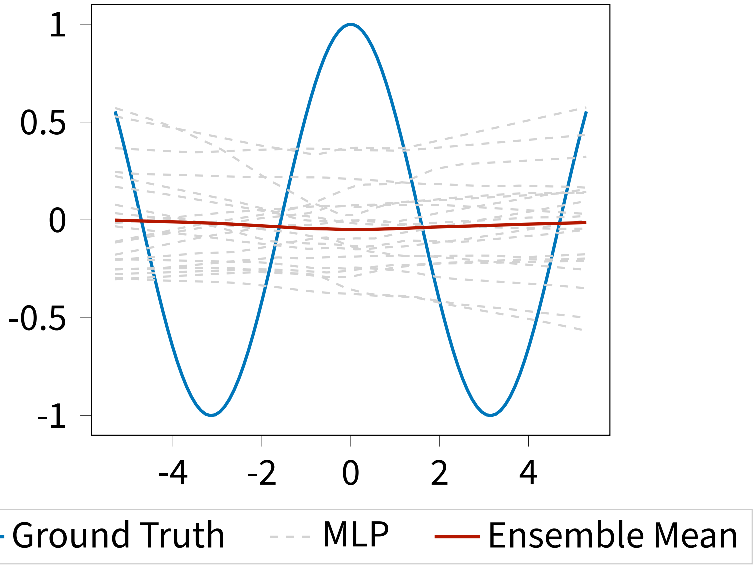
- ✓ Equivariance holds for all training times
- ✓ Equivariance holds away from the training data

ⓘ At infinite width, the mean output at initialization is zero everywhere.

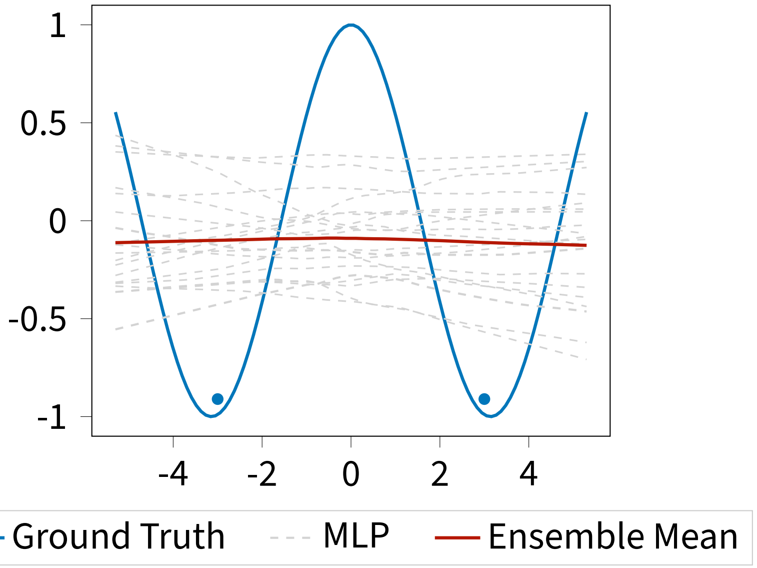
⇨ Training with full data augmentation leads to an equivariant function.

Toy example

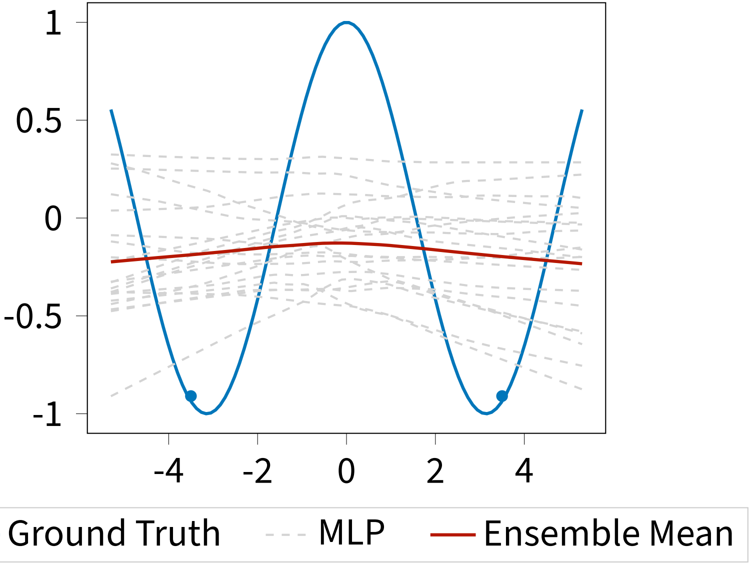
Initialization



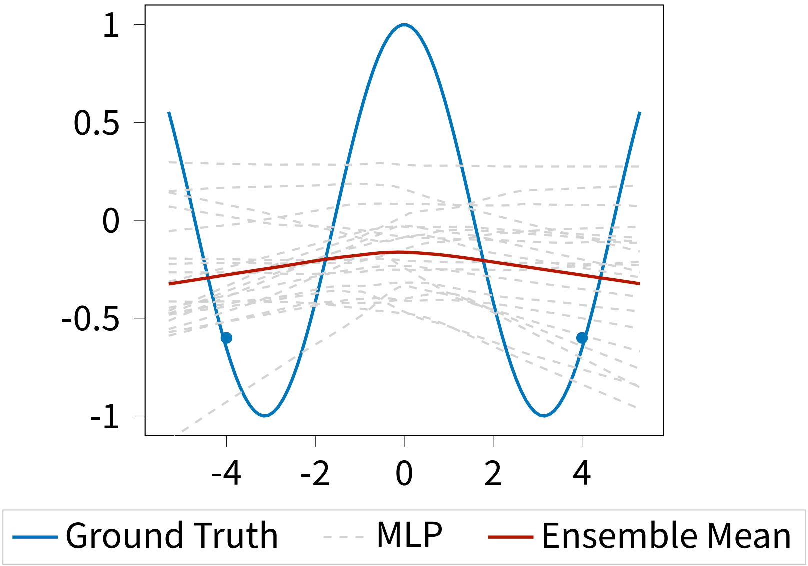
After 1 Training Step



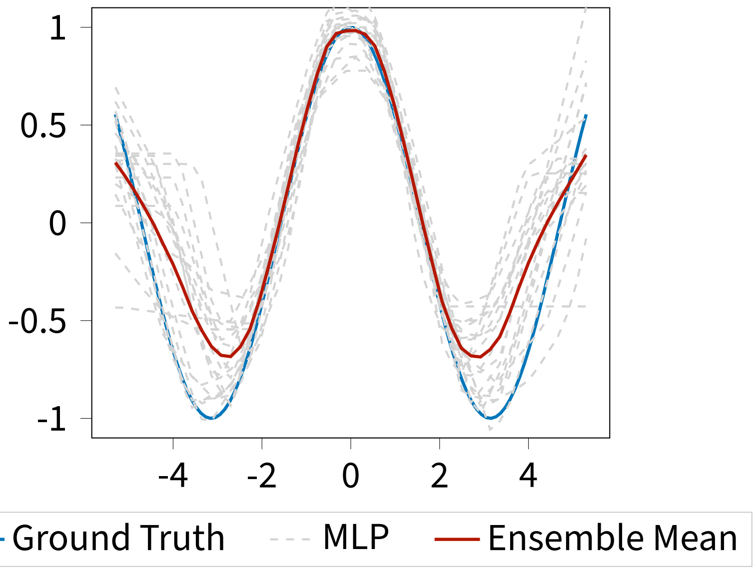
After 2 Training Steps



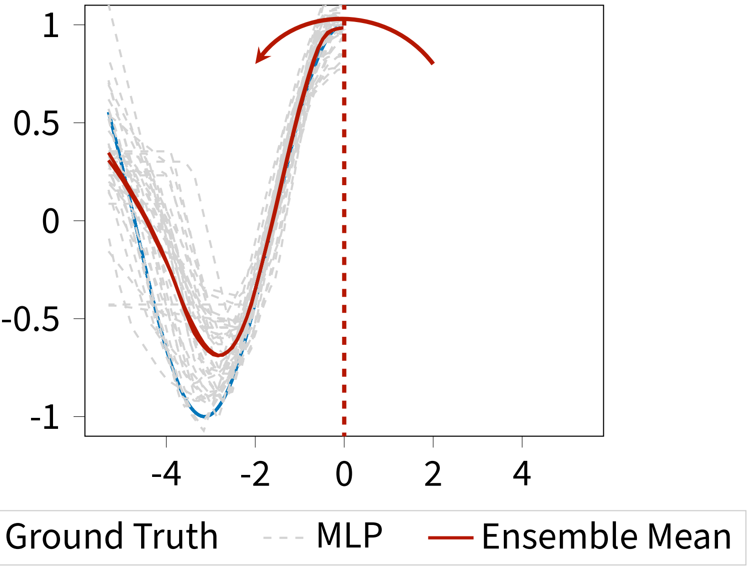
After 3 Training Steps



After 2000 Training Steps



After 2000 Training Steps



Proof idea

Mean prediction

test point

$$\underbrace{\frac{1}{n} \sum_{\theta_0=\text{init}_1}^{\text{init}_n} f_{\theta_t}(x)}_{\text{mean prediction of deep ensemble}}$$

Mean prediction

test point

$$\underbrace{\lim_{n\to\infty} \frac{1}{n} \sum_{\theta_0=\text{init}_1}^{\text{init}_n} f_{\theta_t}(x)}_{\text{mean prediction of deep ensemble}}$$

Mean prediction

test point

$$\underbrace{\lim_{n\to\infty} \frac{1}{n} \sum_{\theta_0=\text{init}_1}^{\text{init}_n} f_{\theta_t}(x)}_{\text{mean prediction of deep ensemble}} = \mathbb{E}_{\theta_0 \sim \text{initializations}}[f_{\theta_t}(x)]$$

Mean prediction

test point

$$\underbrace{\lim_{n\to\infty} \frac{1}{n} \sum_{\theta_0=\text{init}_1}^{\text{init}_n} f_{\theta_t}(x)}_{\text{mean prediction of deep ensemble}} = \mathbb{E}_{\theta_0 \sim \text{initializations}}[f_{\theta_t}(x)] = \mu_t(x)$$

Mean prediction from NTK

[Jacot et al. 2018]

At infinite width, the mean prediction is given in terms of the neural tangent kernel (NTK)

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

Mean prediction from NTK

[Jacot et al. 2018]

- At infinite width, the mean prediction is given in terms of the neural tangent kernel (NTK)

neural tangent kernel

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

Mean prediction from NTK

[Jacot et al. 2018]

At infinite width, the mean prediction is given in terms of the neural tangent kernel (NTK)

neural tangent kernel

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

train data

Mean prediction from NTK

[Jacot et al. 2018]

ⓘ At infinite width, the mean prediction is given in terms of the neural tangent kernel (NTK)

neural tangent kernel

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

learning rate

train data

Mean prediction from NTK

[Jacot et al. 2018]

- At infinite width, the mean prediction is given in terms of the neural tangent kernel (NTK)

neural tangent kernel

train labels

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

learning rate

train data

Proof idea

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

Proof idea

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

augmented data

augmented labels

Proof idea

group transformation

$$\mu_t(\rho(g)x) = \Theta(\rho(g)x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

augmented data

augmented labels

Proof idea

group transformation

for augmented data

$$\mu_t(\rho(g)x) = \Theta(\rho(g)x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

augmented data

augmented labels

Proof idea

group transformation

$$\mu_t(\rho(g)x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})\rho(g)Y$$

augmented data

augmented labels

Proof idea

group transformation

augmented labels

$$\mu_t(\rho(g)x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})\underbrace{\rho(g)Y}_{=Y}$$

for invariance

Proof idea

group transformation

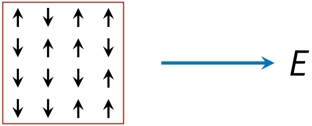
$$\mu_t(\rho(g)x) = \Theta(x,X)\Theta(X,X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})\underbrace{\rho(g)Y}_{=Y}$$

$$= \mu_t(x)$$

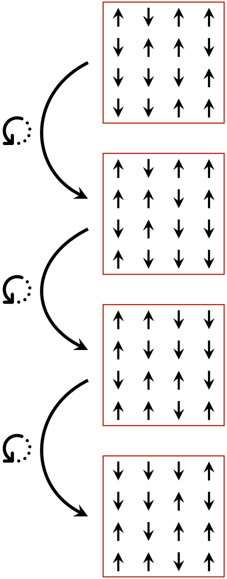
for invariance

Experiments

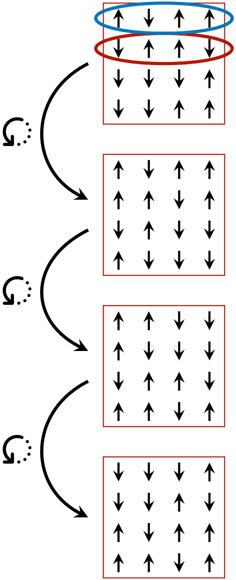
Ising model



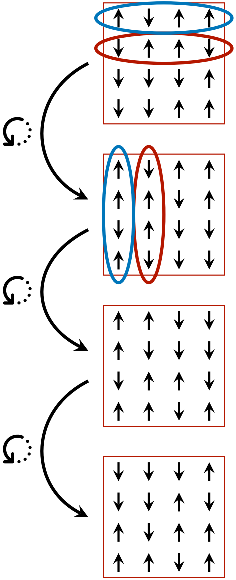
Ising model



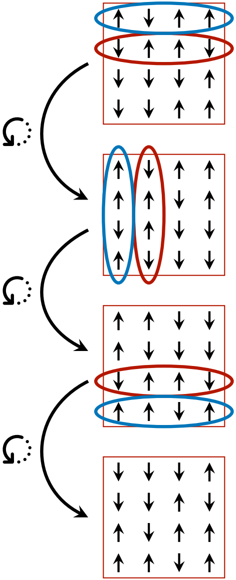
Ising model



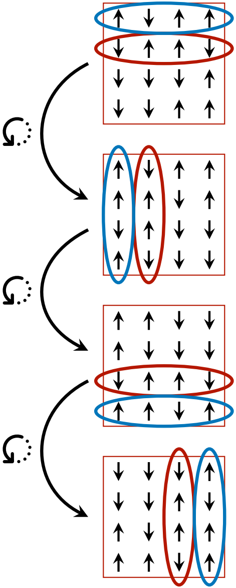
Ising model



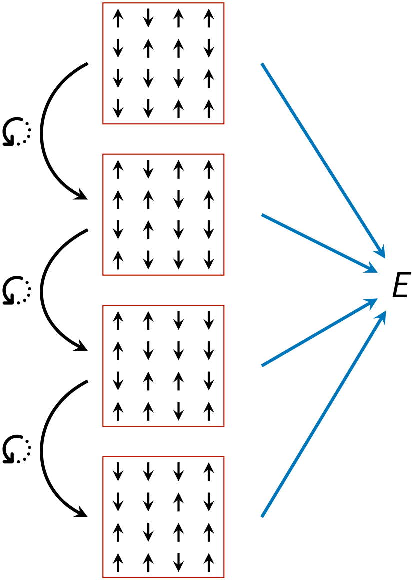
Ising model



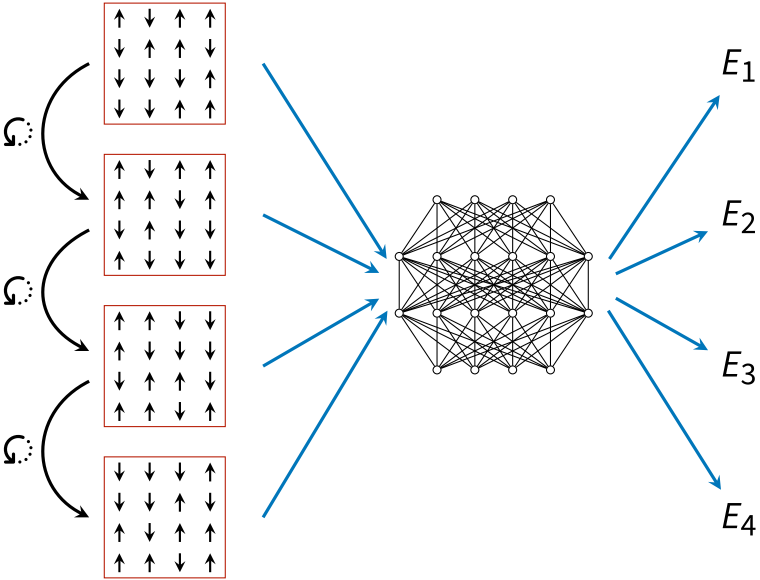
Ising model



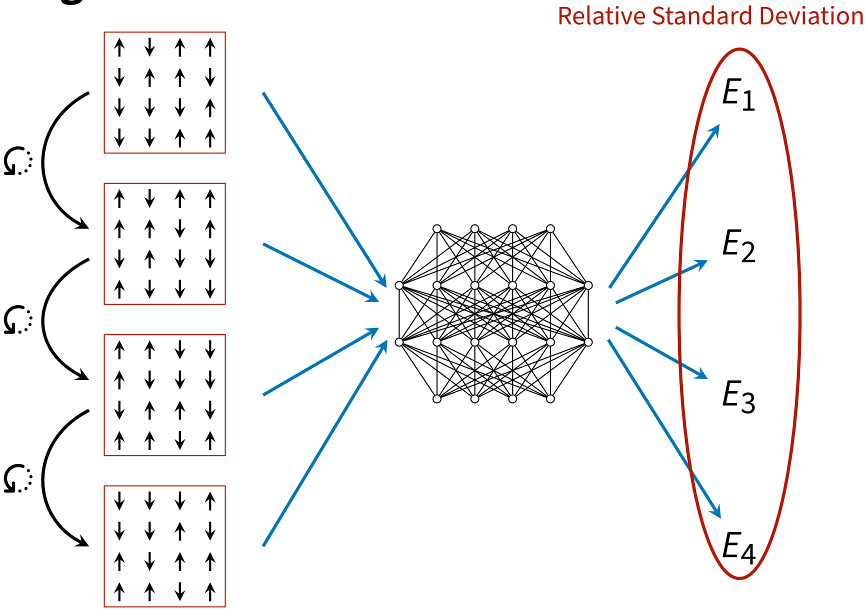
Ising model

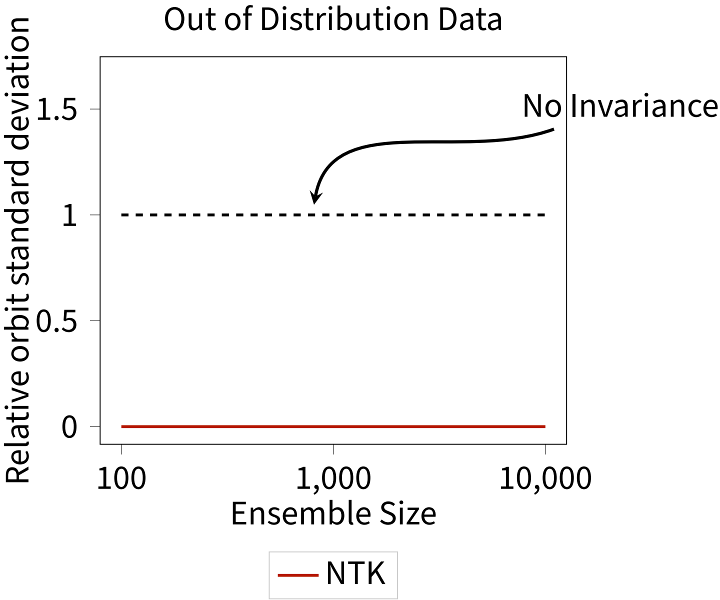


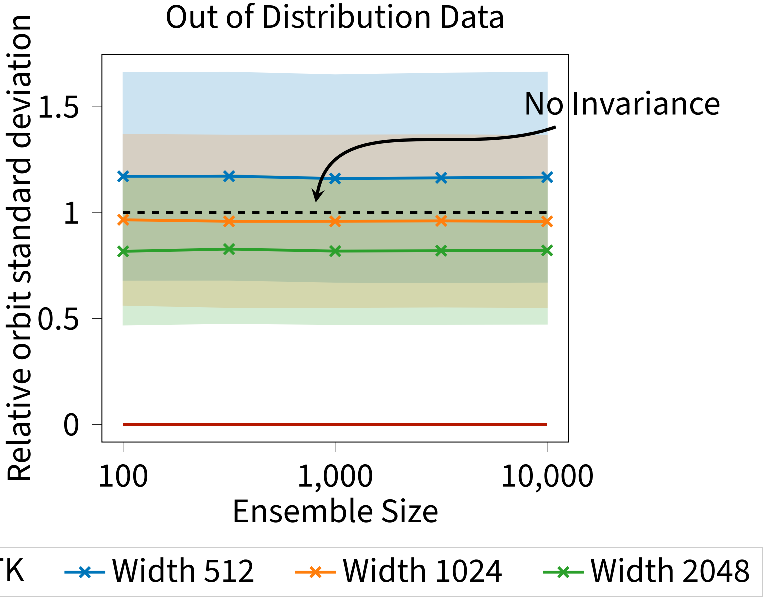
Ising model

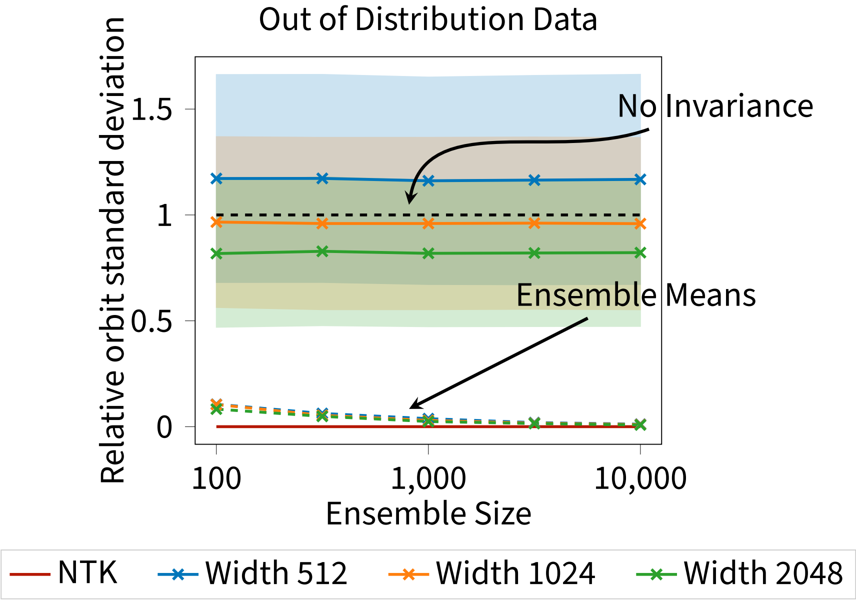


Ising model

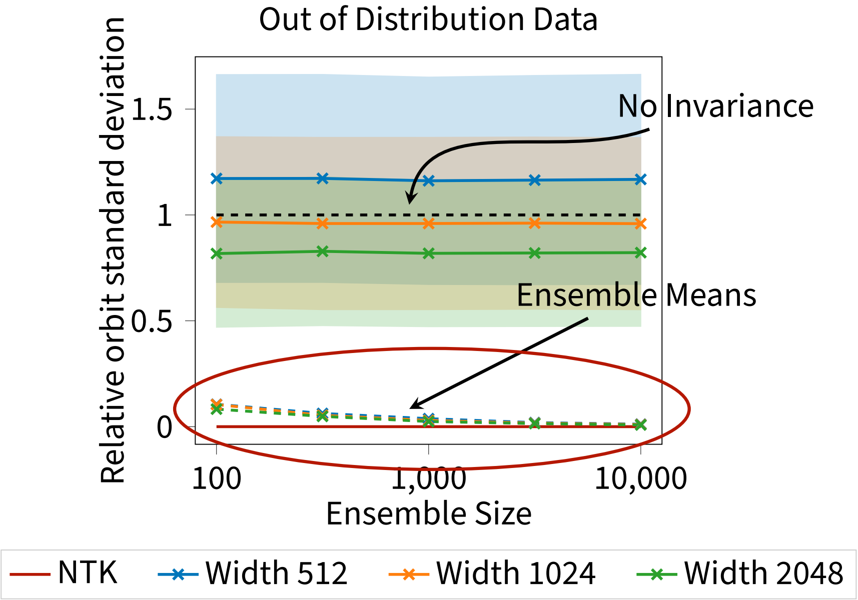


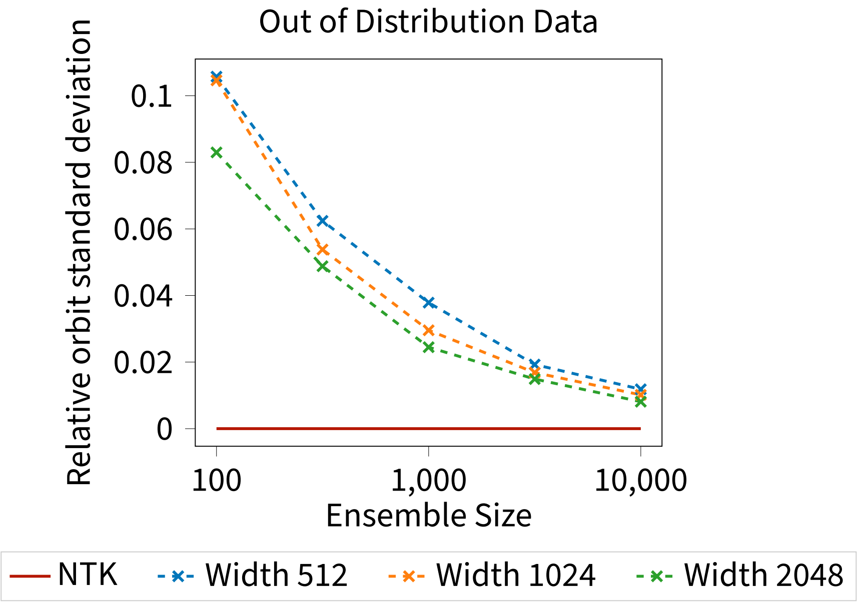

Out of Distribution Data
No Invariance
Relative orbit standard deviation
1.5
1
0.5
0
100
1,000
10,000
Ensemble Size
NTK


Out of Distribution Data
No Invariance
Relative orbit standard deviation
1.5
1
0.5
0
100
1,000
10,000
Ensemble Size
NTK
Width 512
Width 1024
Width 2048



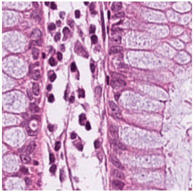
Out of Distribution Data
No Invariance
Ensemble Means
Relative orbit standard deviation
1.5
1
0.5
0
100
1,000
10,000
Ensemble Size
NTK
Width 512
Width 1024
Width 2048


Out of Distribution Data
No Invariance
Ensemble Means
Relative orbit standard deviation
1.5
1
0.5
0
100
1,000
10,000
Ensemble Size
NTK
Width 512
Width 1024
Width 2048


Out of Distribution Data
Relative orbit standard deviation
0.1
0.08
0.06
0.04
0.02
0
100
1,000
10,000
Ensemble Size
NTK
Width 512
Width 1024
Width 2048

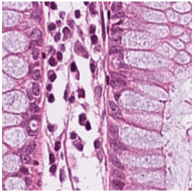
Histological slices

[Kather et al. 2018]



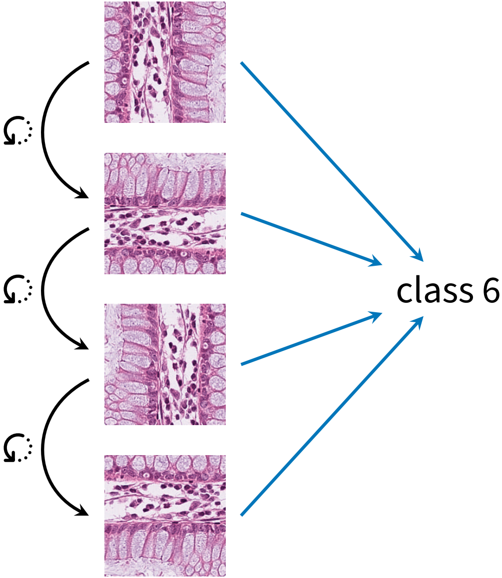
Histological slices

[Kather et al. 2018]

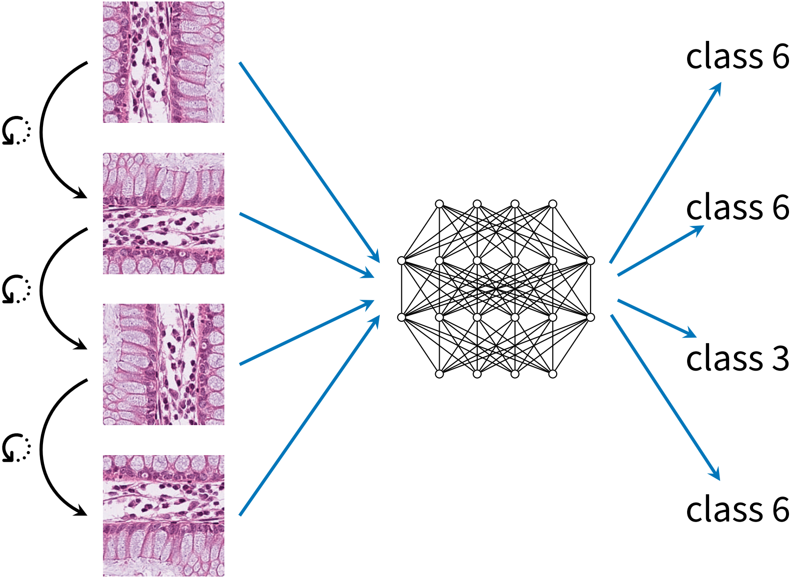


class 6

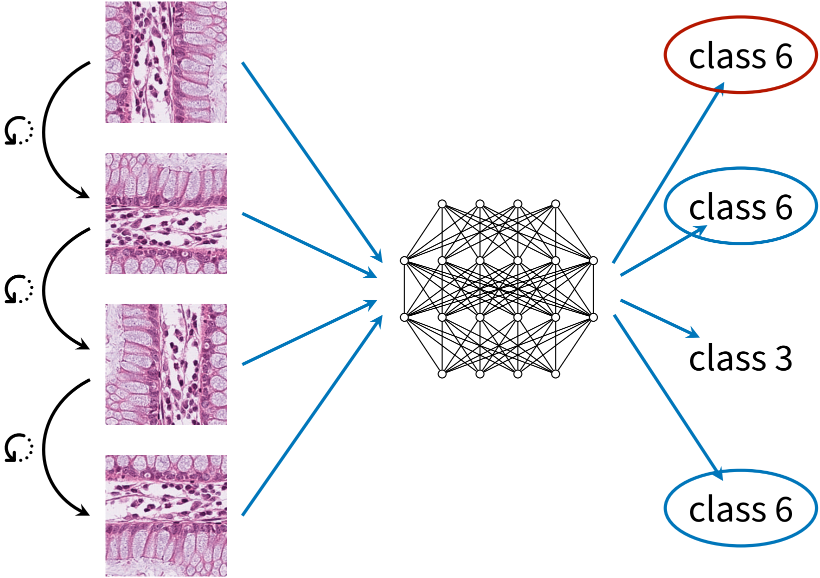
Histological slices



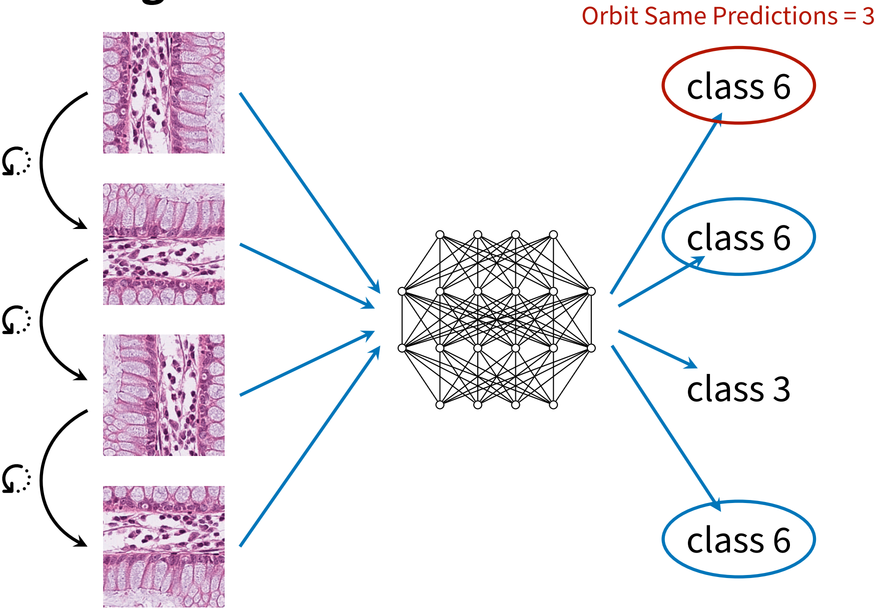
Histological slices



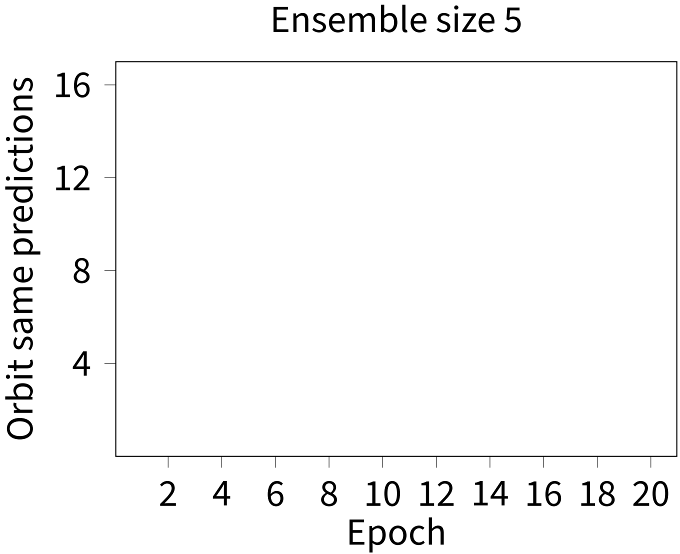
Histological slices



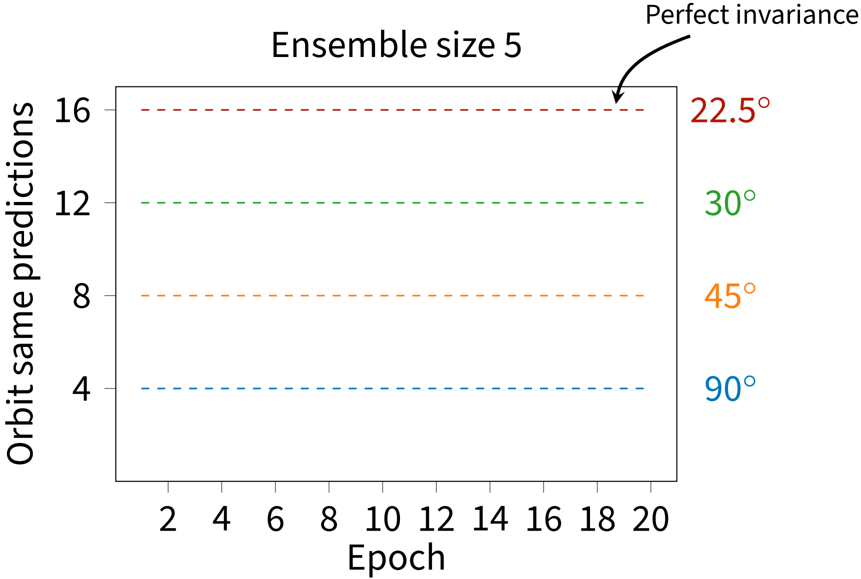
Histological slices



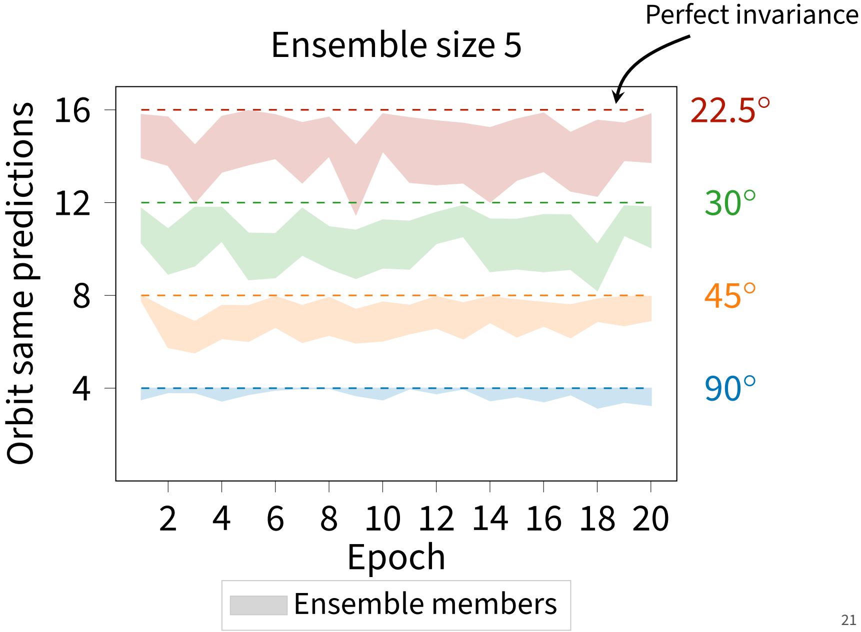
Out of distribution results



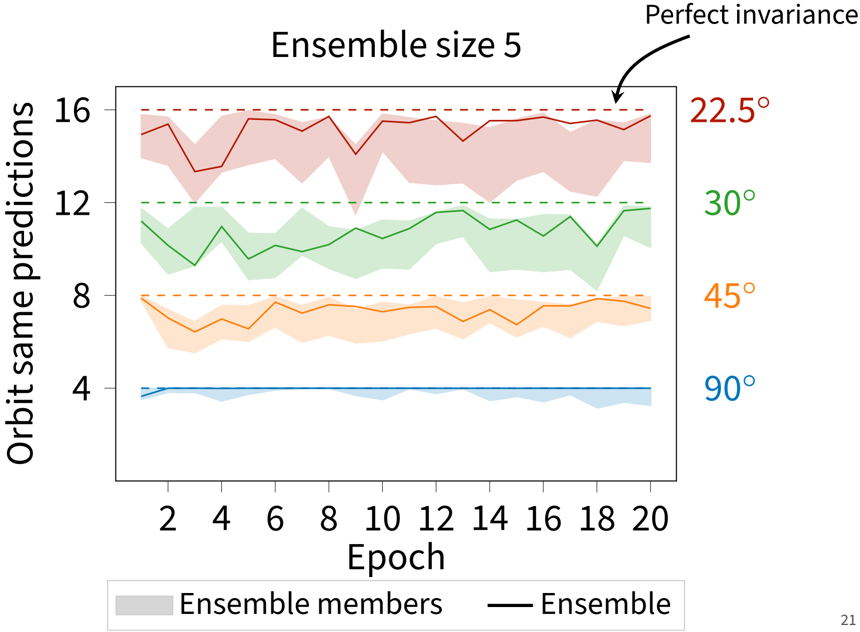
Out of distribution results



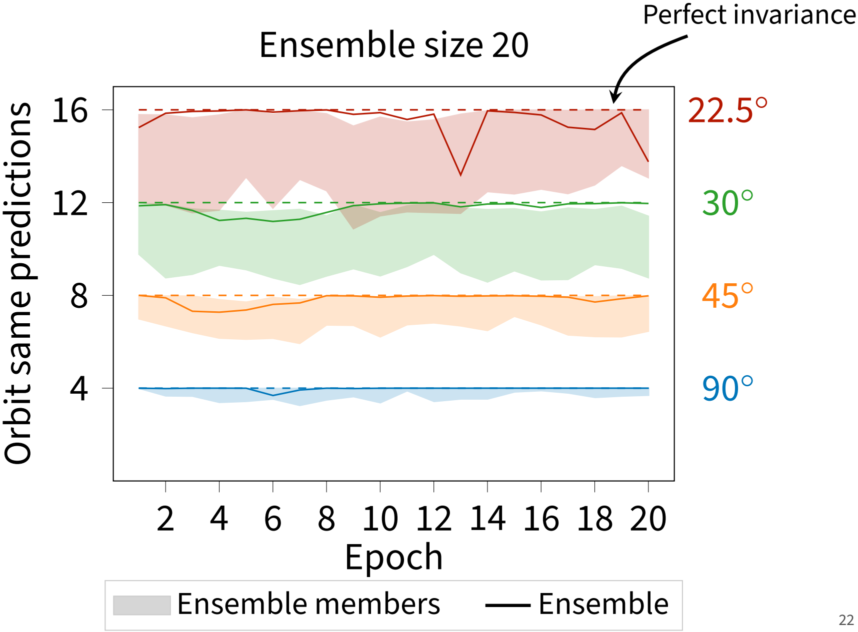
Out of distribution results



Out of distribution results



Out of distribution results



Comparison to other methods

Comparison to other methods

➪ Models trained on rotated FashionMNIST

Comparison to other methods

➩ Models trained on rotated FashionMNIST

Orbit same predictions out of distribution:

	C_4	C_8	C_{16}
DeepEns+DA	3.85±0.12	**7.72±0.34**	**15.24±0.69**
only DA	3.41±0.18	6.73±0.24	12.77±0.71
E2CNN[1]	**4±0.0**	**7.71±0.21**	**15.08±0.34**
Canon[2]	**4±0.0**	**7.45±0.14**	12.41±0.85

[1] [Weiler et al. 2019], [2] [Kaba et al. 2022]

Key takeaways

Key takeaways

If you need ensembles

- 👍 use data augmentation to obtain an equivariant model.

Key takeaways

If you need ensembles

- 👍 use data augmentation to obtain an equivariant model.

If you need data augmentation

- 👍 use an ensemble to boost the equivariance.

Poster

Thursday, 25 July 2024
11.30am – 1.00pm
Hall C 4-9
Poster 817