# Estimating Epistemic and Aleatoric Uncertainty with a Single Model

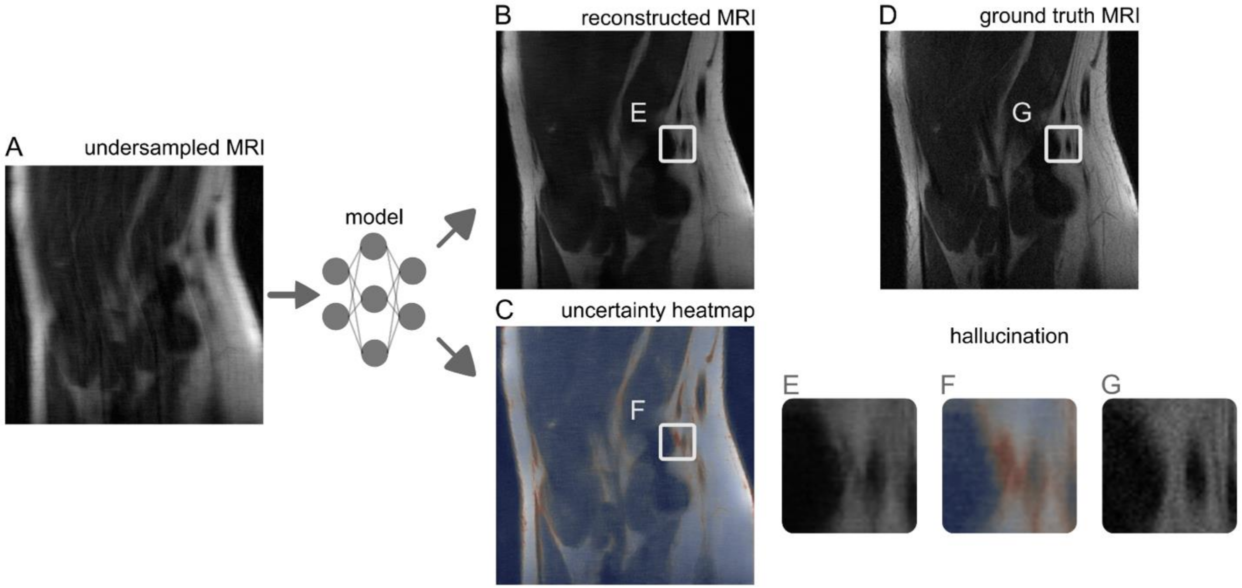
Matthew Chan, Maria Molina, Christopher Metzler

# Topics

- Motivation
  - What is uncertainty?
  - Why is uncertainty useful?
  - Aleatoric vs. epistemic uncertainty
- Problem Definition
  - Uncertainty estimation using generative models
  - Uncertainty estimation using deep ensembles
  - Drawbacks of existing methods
- Hyper-Diffusion Models
  - Proof-of-concept
  - Experiment results
  - Analysis against baselines

# Motivation

*Uncertainty* provides valuable insights into how confident a model's predictions are.

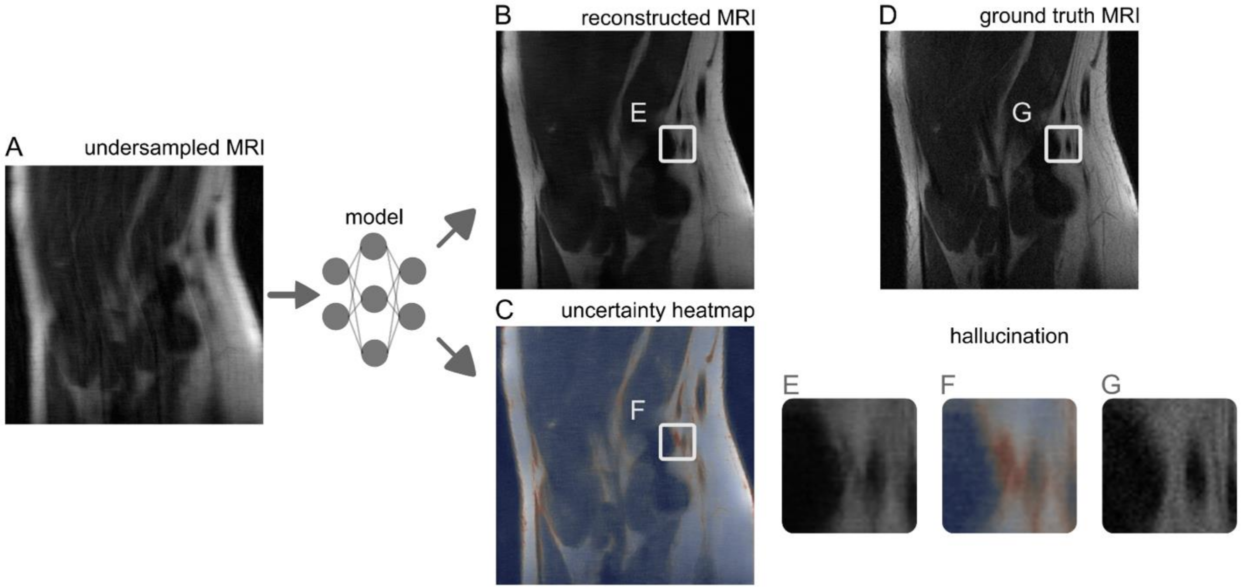



Angelopoulos, Anastasios N., et al. "Image-to-image regression with distribution-free uncertainty quantification and applications in imaging." International Conference on Machine Learning. PMLR, 2022.

# Motivation

*Uncertainty* provides valuable insights into how confident a model's predictions are.

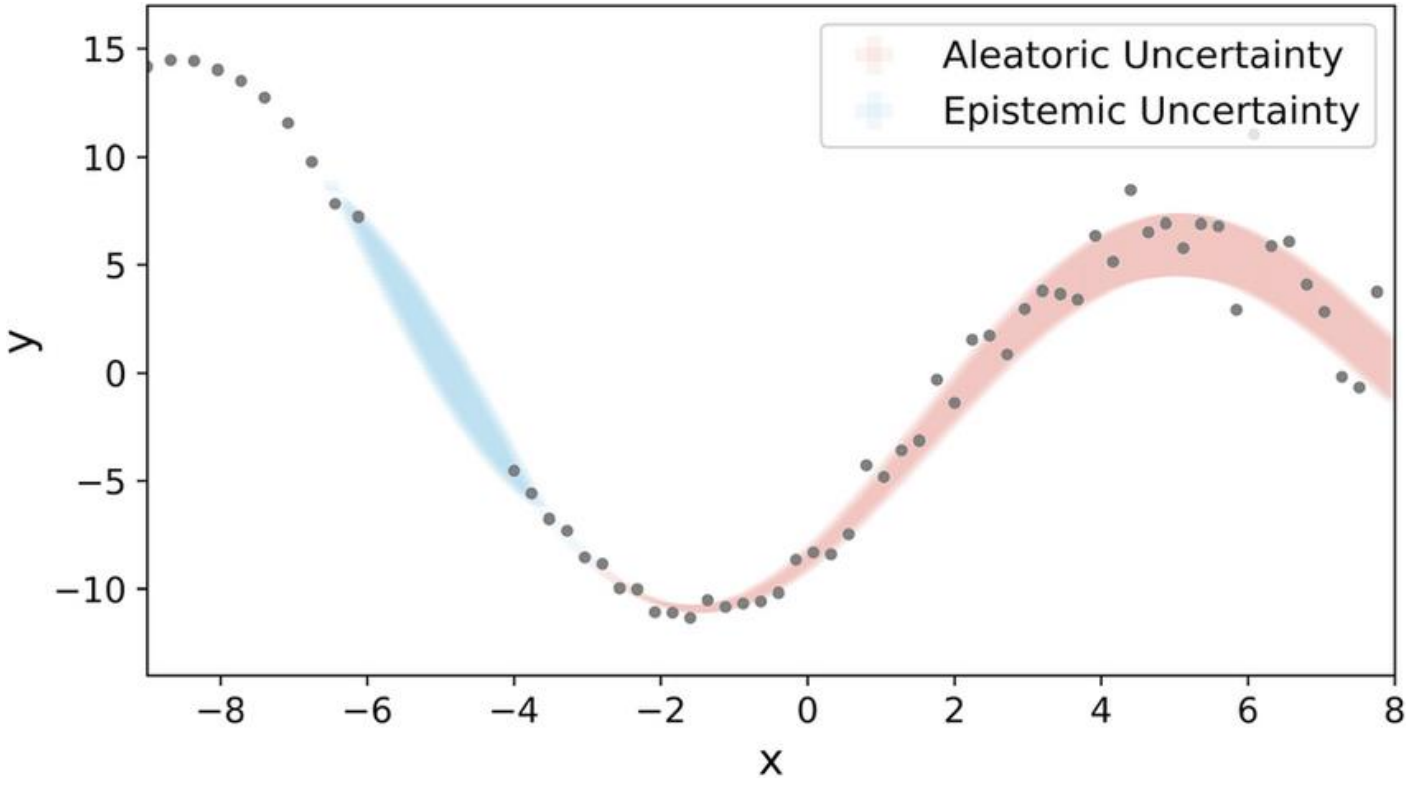
For high-stakes applications like MRI / CT reconstruction, uncertainty serves as a key indicator for rejection verification (i.e., whether model predictions should be verified by a human expert).




Angelopoulos, Anastasios N., et al. "Image-to-image regression with distribution-free uncertainty quantification and applications in imaging." International Conference on Machine Learning. PMLR, 2022.

# Objective

We seek to quantify two types of uncertainty:

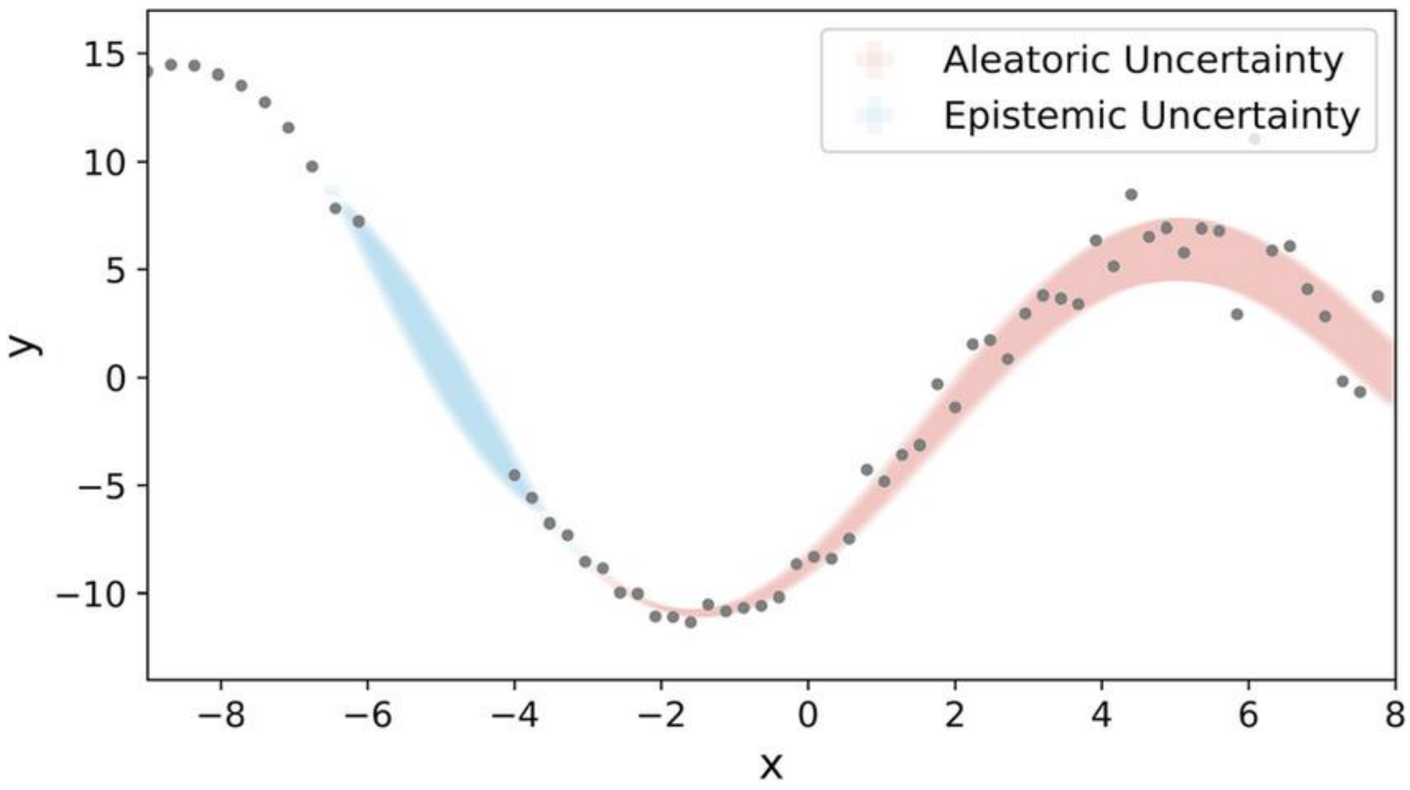



Yang, Chu-I & Li, Yi-Pei. (2023). Explainable uncertainty quantifications for deep learning-based molecular property prediction. Journal of Cheminformatics. 15. 10.1186/s13321-023-00682-3.

# Objective

We seek to quantify two types of uncertainty:

- **Aleatoric uncertainty**, which is *irreducible*, stems from inherent variability and randomness in the problem.

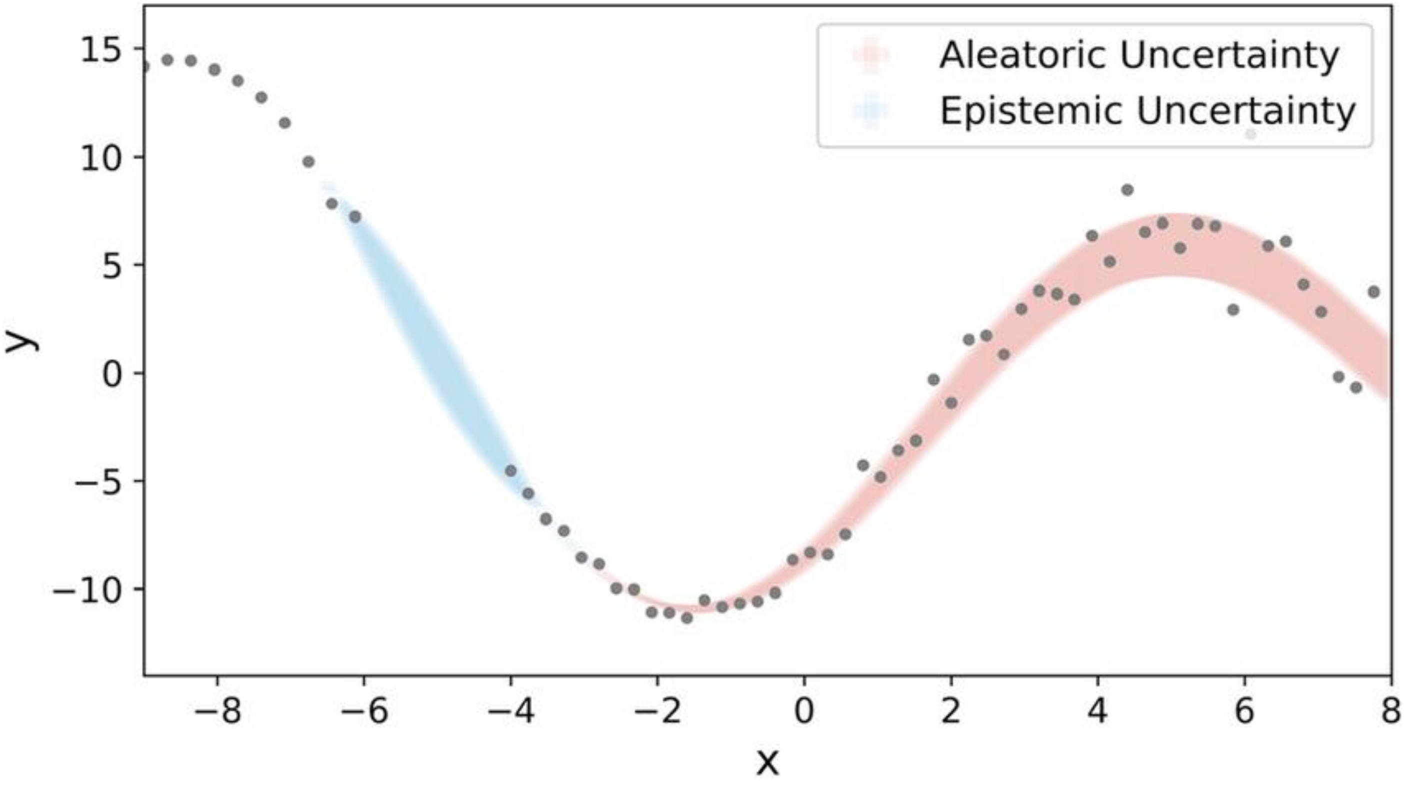



Yang, Chu-I & Li, Yi-Pei. (2023). Explainable uncertainty quantifications for deep learning-based molecular property prediction. Journal of Cheminformatics. 15. 10.1186/s13321-023-00682-3.

# Objective

We seek to quantify two types of uncertainty:

- **Aleatoric uncertainty**, which is *irreducible*, stems from inherent variability and randomness in the problem.
- **Epistemic uncertainty** relates to a lack of knowledge and is *reducible* with more training data.




Yang, Chu-I & Li, Yi-Pei. (2023). Explainable uncertainty quantifications for deep learning-based molecular property prediction. Journal of Cheminformatics. 15. 10.1186/s13321-023-00682-3.

# Problem Definition

Bayesian inference models a distribution of network predictions as the product between a likelihood (i.e., aleatoric) function and a posterior weight (i.e., epistemic) distribution:

$$p(x|y, \mathcal{D}) = \int \underbrace{p(x|y, \phi)}_{\text{aleatoric}} \underbrace{p(\phi|\mathcal{D})}_{\text{epistemic}} \, d\phi.$$

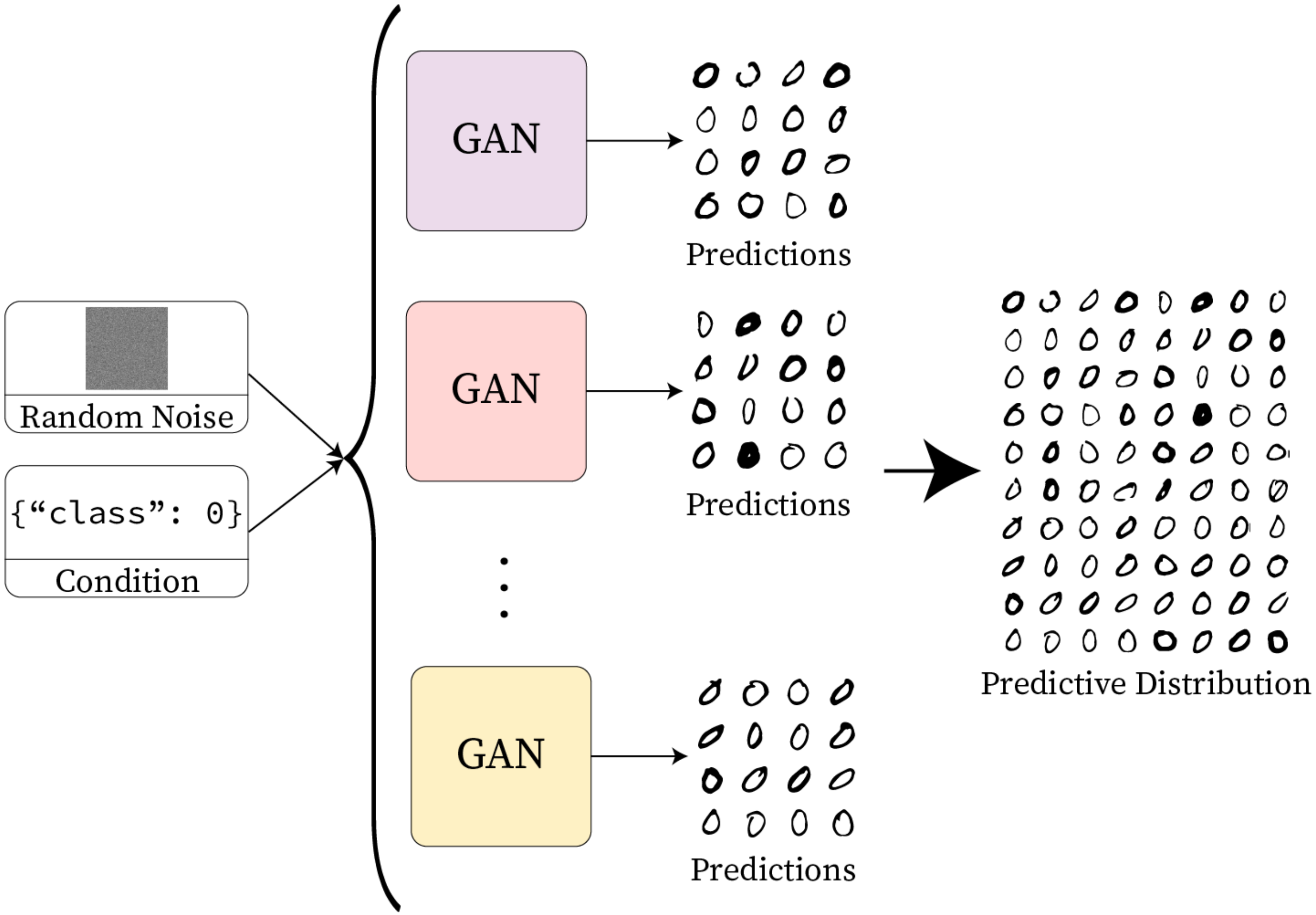
| Symbol | Meaning |
|---|---|
| $x$ | Signal to recover |
| $y$ | Observed measurement |
| $\phi$ | Model parameters |
| $\mathcal{D}$ | Training dataset |

# Building a Predictive Distribution

To build the predictive distribution

$$p(x|y,\mathcal{D}) = \int p(x|y,\phi)p(\phi|\mathcal{D})d\phi$$

we can train an *ensemble* of *generative models*.



# Building a Predictive Distribution

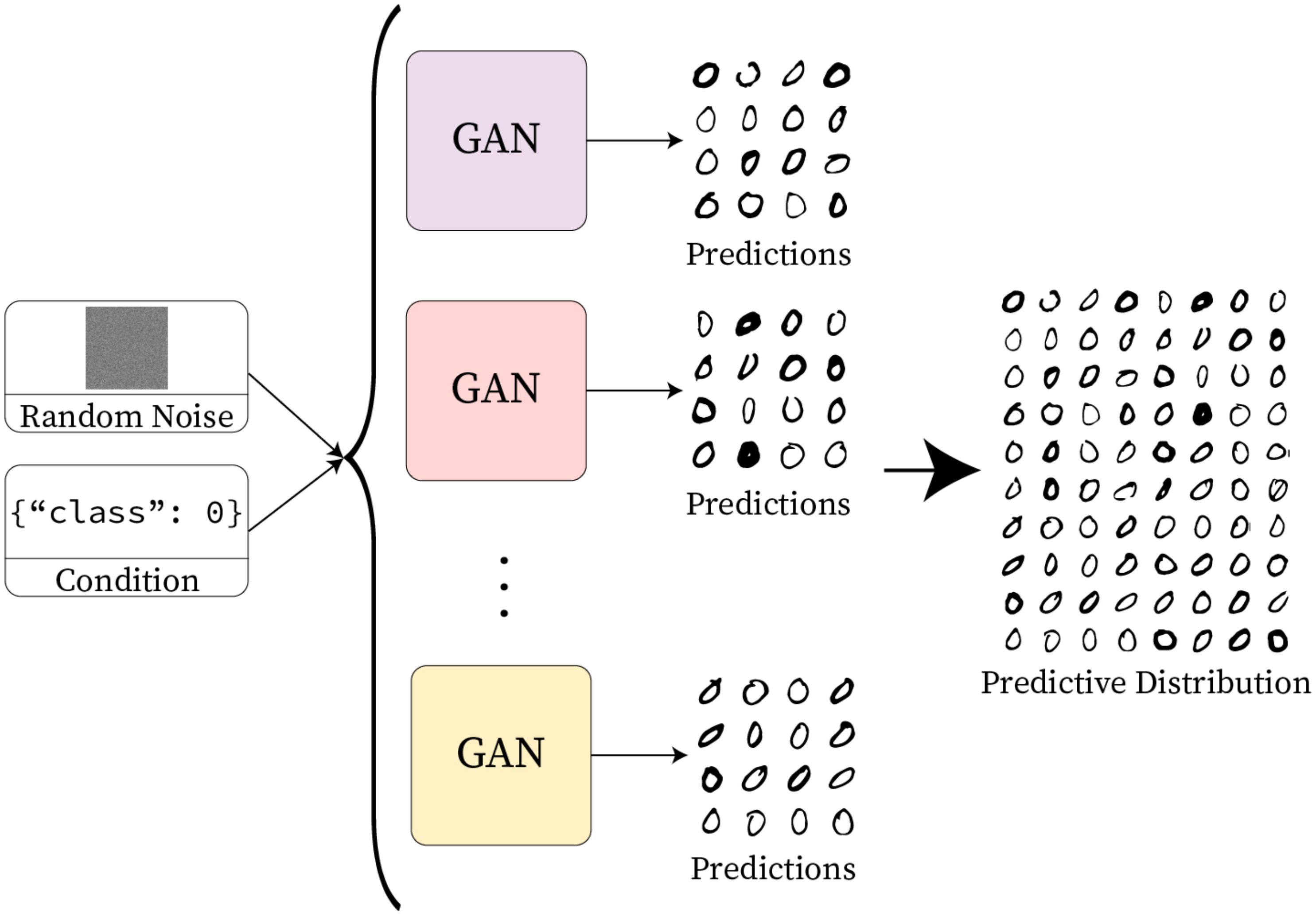
To build the predictive distribution

$$p(x|y,\mathcal{D}) = \int p(x|y,\phi)p(\phi|\mathcal{D})d\phi$$

we can train an *ensemble* of *generative models*.

We can decompose the predictive distribution into aleatoric and epistemic uncertainty, using the law of total variance:

- $\mathrm{AU} = \mathbb{E}_{\phi\sim p(\phi|\mathcal{D})}\left[\mathrm{Var}_{\hat{X}\sim p(x|y,\phi)}\left[\hat{X}\right]\right]$
- $\mathrm{EU} = \mathrm{Var}_{\phi\sim p(\phi|\mathcal{D})}\left[\mathbb{E}_{\hat{X}\sim p(x|y,\phi)}\left[\hat{X}\right]\right]$



# Building a Predictive Distribution

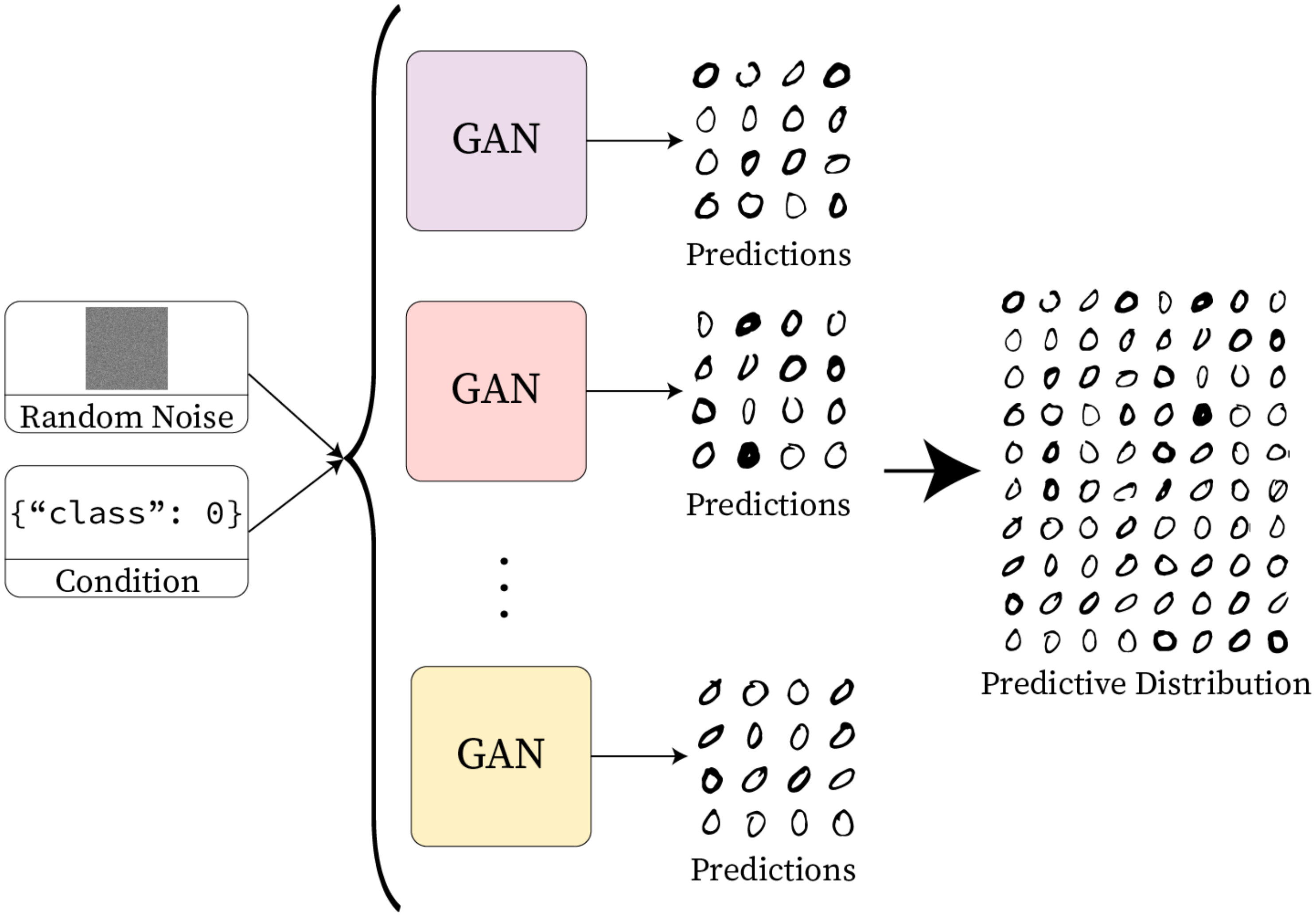
To build the predictive distribution

$$p(x|y,\mathcal{D}) = \int p(x|y,\phi)p(\phi|\mathcal{D})d\phi$$

we can train an *ensemble* of *generative models*.

We can decompose the predictive distribution into aleatoric and epistemic uncertainty, using the law of total variance:

- $\mathrm{AU} = \mathbb{E}_{\phi\sim p(\phi|\mathcal{D})}\left[\mathrm{Var}_{\hat{X}\sim p(x|y,\phi)}\left[\hat{X}\right]\right]$
- $\mathrm{EU} = \mathrm{Var}_{\phi\sim p(\phi|\mathcal{D})}\left[\mathbb{E}_{\hat{X}\sim p(x|y,\phi)}\left[\hat{X}\right]\right]$




**The computational cost of training large ensembles is prohibitively expensive!**

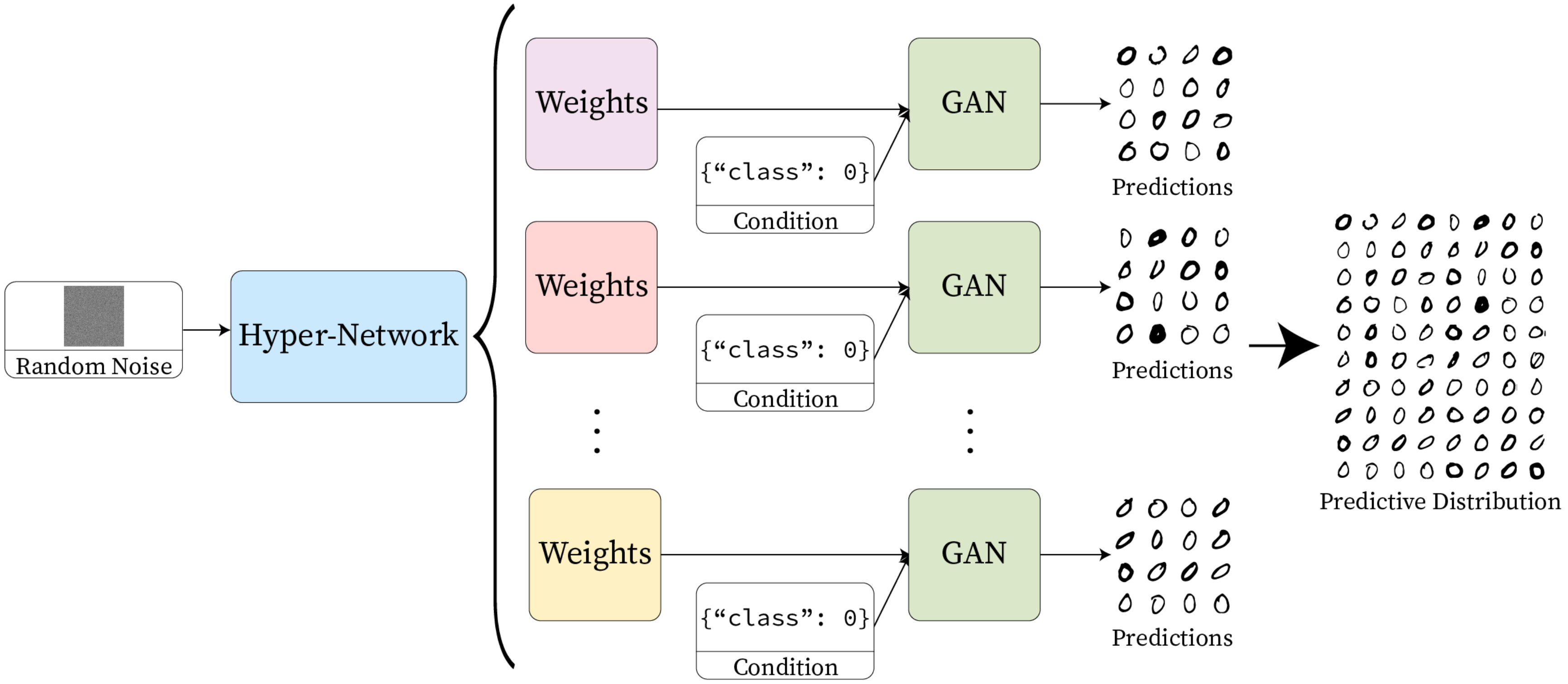
# Our Solution: Hyper-Networks

# Our Solution: Hyper-Networks

Hyper-networks are networks that generate weights for another “primary” network.

# Our Solution: Hyper-Networks

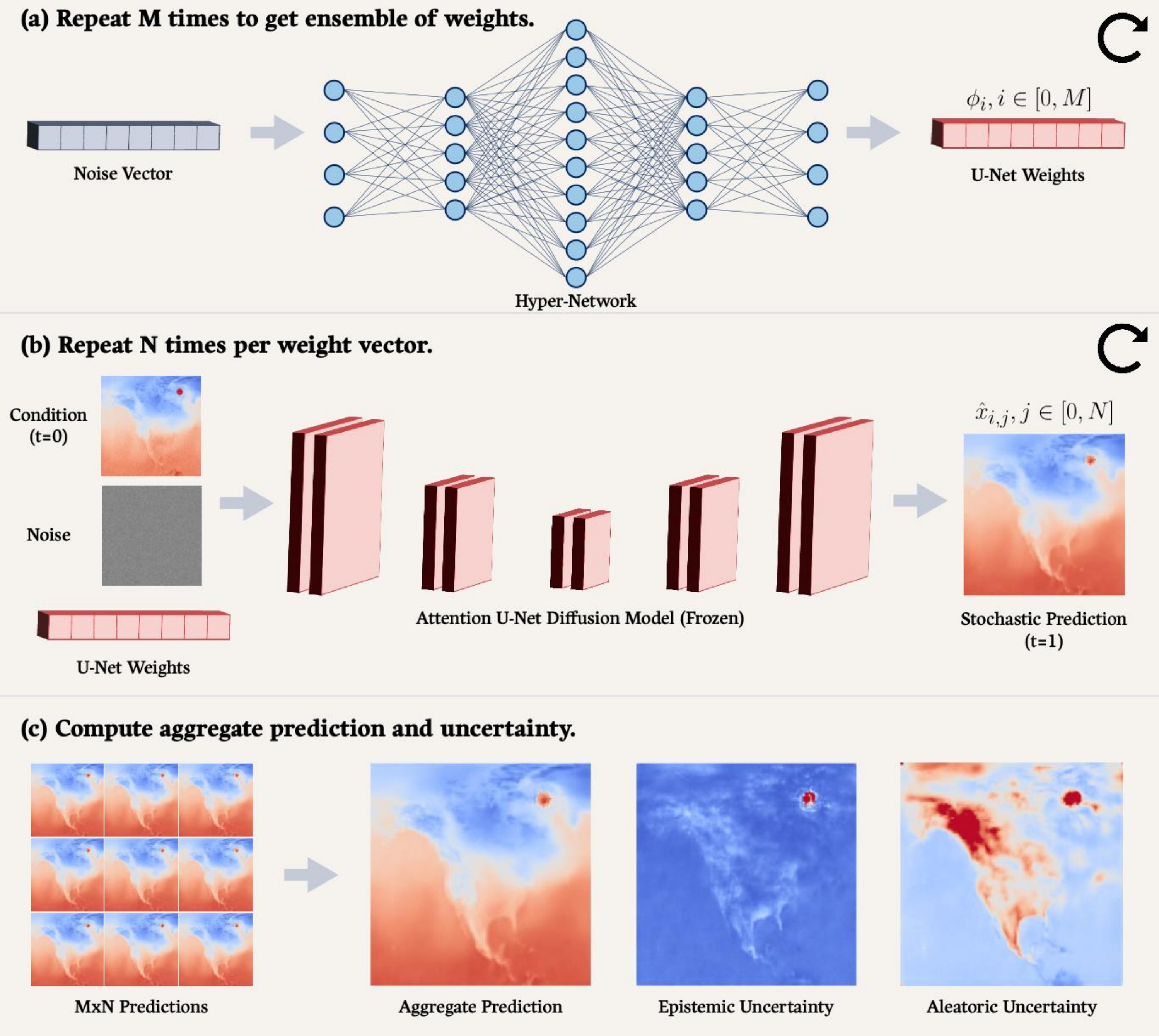
Hyper-networks are networks that generate weights for another “primary” network. They can approximate a deep ensemble, at a significantly reduced computational cost.



# Hyper-Diffusion Models

We combine *hyper-networks* with generative models (i.e., *diffusion models*) to build a predictive distribution and estimate uncertainty.

We validate our method, **hyper-diffusion models (HyperDM)**, on a toy problem and then apply it on weather forecasting and CT reconstruction tasks.



# Validation: Toy Problem

We generate training datasets with controlled uncertainty using

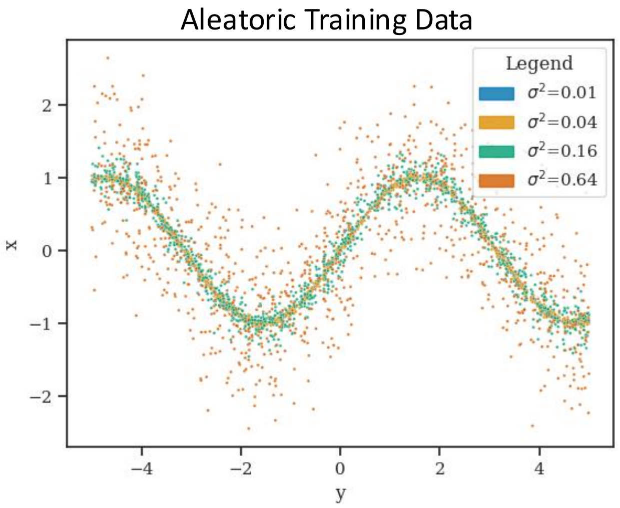
$$y = \sin(x) + \eta,\ \eta \sim \mathcal{N}(0, \sigma^2).$$

# Validation: Toy Problem

We generate training datasets with controlled uncertainty using

$$y = \sin(x) + \eta,\ \ \eta \sim \mathcal{N}(0, \sigma^2).$$

- Strength of the white noise controls *aleatoric uncertainty*.

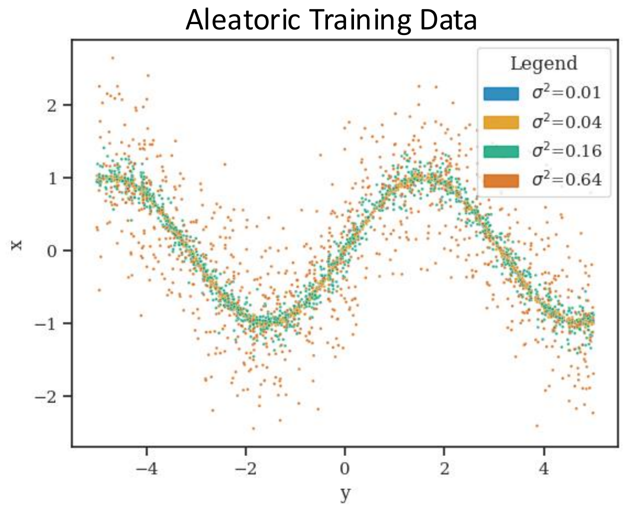


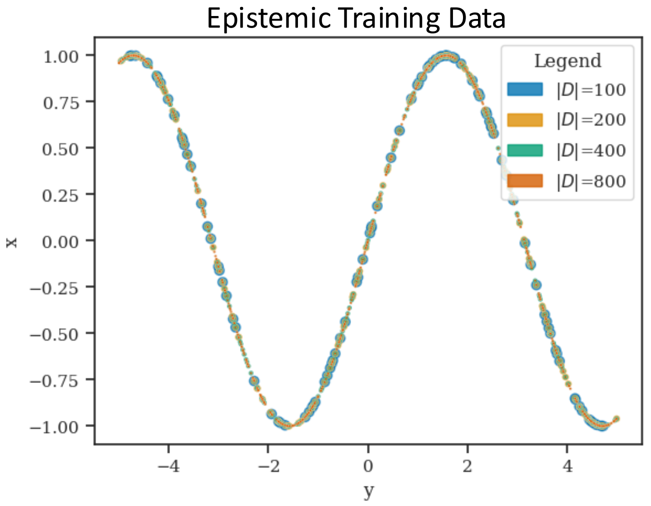
# Validation: Toy Problem

We generate training datasets with controlled uncertainty using

$$y = \sin(x) + \eta,\ \eta \sim \mathcal{N}(0, \sigma^2).$$

- Strength of the white noise controls *aleatoric uncertainty*.
- Size of the dataset controls *epistemic uncertainty*.



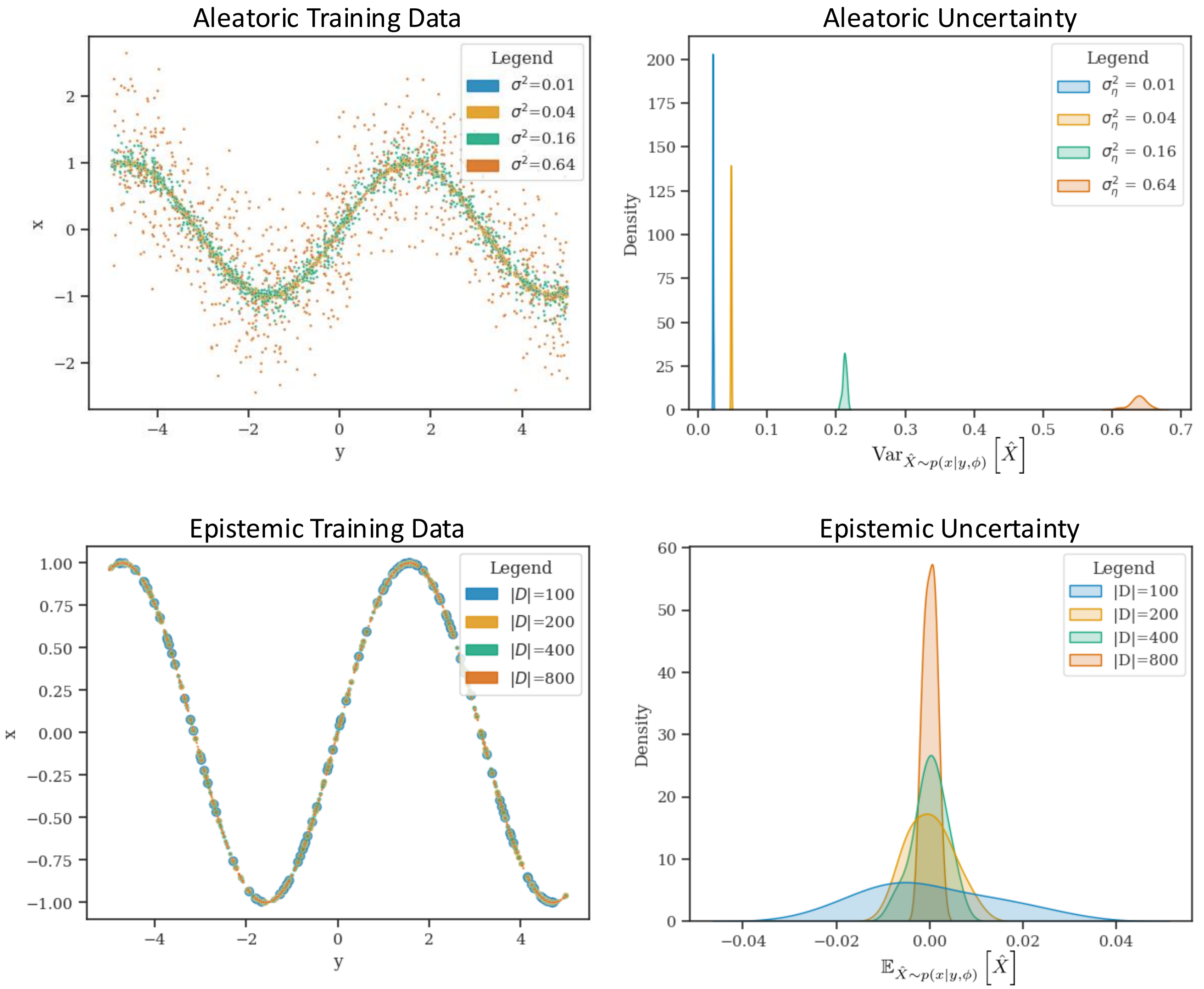
# Validation: Toy Problem

We generate training datasets with controlled uncertainty using

$$y = \sin(x) + \eta,\ \ \eta \sim \mathcal{N}(0, \sigma^2).$$

- Strength of the white noise controls *aleatoric uncertainty*.
- Size of the dataset controls *epistemic uncertainty*.

**Our estimates accurately predict the ground-truth uncertainty.**

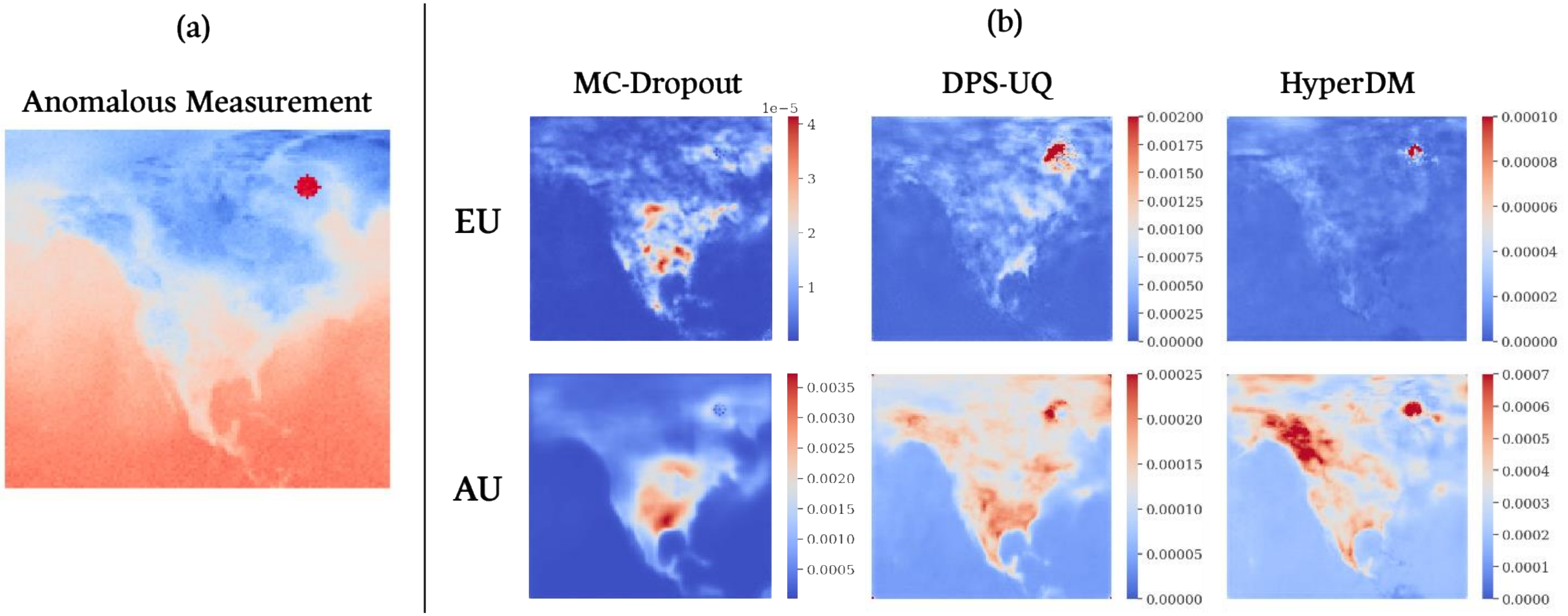


# Weather Forecasting

We use our method for out-of-distribution detection on the ERA5 dataset.

HyperDM is trained to predict surface temperature at 6-hour intervals. We construct an anomalous hotspot over northeastern Canada and estimate uncertainty.

**Our method's epistemic uncertainty estimate highlights the out-of-distribution feature better than deep ensembles.**

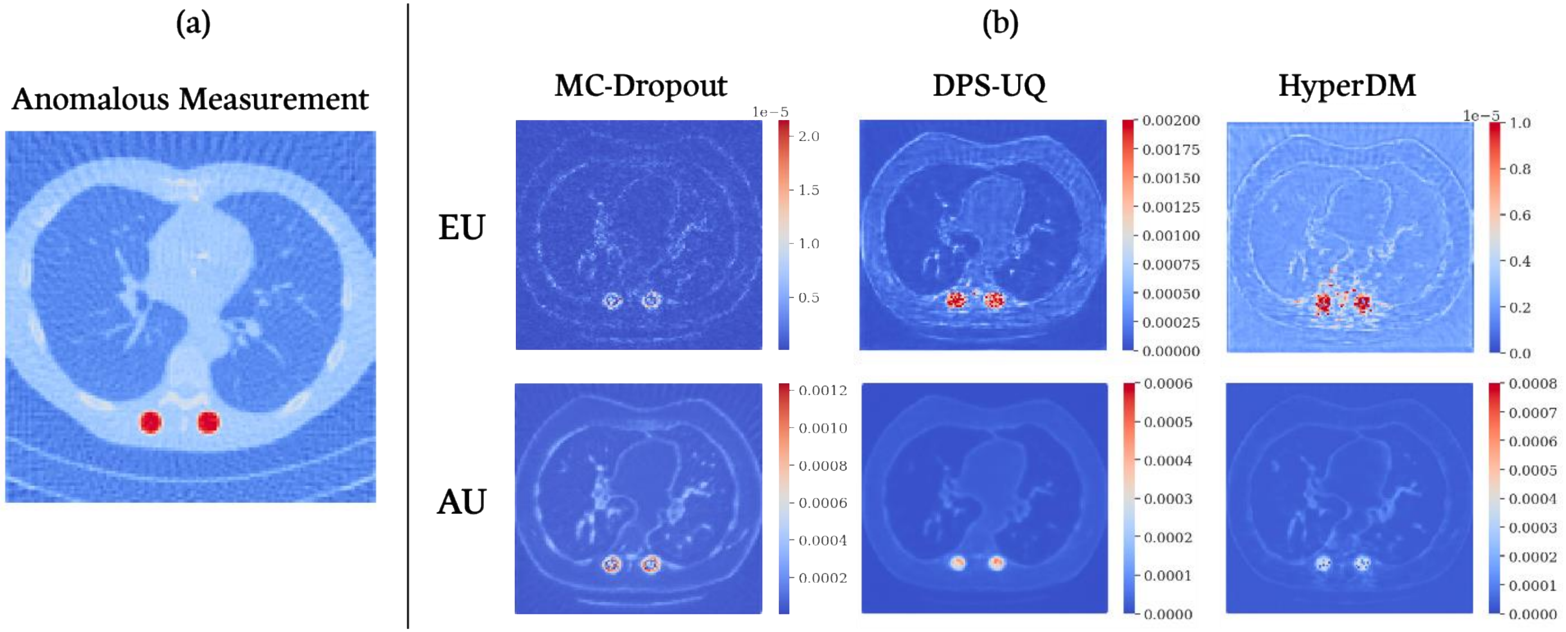


# Computed Tomography

We similarly test HyperDM on the LUNA16 dataset.

Our method is trained to recover high-quality CT scans from poor sinogram reconstructions. Out-of-distribution measurements are created by synthetically inserting metallic implants near the spine.

**Our method's epistemic uncertainty estimate highlights the abnormal feature similarly to a deep ensemble.**



# Prediction Quality

We evaluate the predictive distribution's accuracy on a hold-out test set using the *structural similarity index (SSIM)*, *peak signal-to-noise ratio (PSNR)* and *continuous ranked probability score (CRPS)*.

**HyperDM performs similarly to, if not better than, deep ensembles.**

Table 2: **Ensemble prediction quality on real-world data.** The image quality assessment metrics achieved by each method on a CT reconstruction dataset (i.e., LUNA) and a weather prediction dataset (i.e., ERA5) are reported below. Best scores are highlighted in red and second best scores are highlighted in blue.

| | LUNA | | | ERA5 | | |
|---|---|---|---|---|---|---|
| METHOD | SSIM ↑ | PSNR (DB) ↑ | CRPS ↓ | SSIM ↑ | PSNR (DB) ↑ | CRPS ↓ |
| MC-DROPOUT [16] | 0.77 | 30.25 | 0.023 | 0.93 | 31.34 | 0.034 |
| DPS-UQ [13] | 0.89 | 34.95 | 0.01 | 0.94 | 32.83 | 0.013 |
| HYPERDM | 0.87 | 35.16 | 0.01 | 0.95 | 33.15 | 0.012 |

# Prediction Quality

We evaluate the predictive distribution's accuracy on a hold-out test set using the *structural similarity index (SSIM)*, *peak signal-to-noise ratio (PSNR)* and *continuous ranked probability score (CRPS).*

**HyperDM performs similarly to, if not better than, deep ensembles.**

**Additionally, it has a significantly lower training cost due to the hyper-network.**

Table 2: **Ensemble prediction quality on real-world data.** The image quality assessment metrics achieved by each method on a CT reconstruction dataset (i.e., LUNA) and a weather prediction dataset (i.e., ERA5) are reported below. Best scores are highlighted in red and second best scores are highlighted in blue.

| | LUNA | | | ERA5 | | |
|---|---|---|---|---|---|---|
| METHOD | SSIM ↑ | PSNR (DB) ↑ | CRPS ↓ | SSIM ↑ | PSNR (DB) ↑ | CRPS ↓ |
| MC-DROPOUT [16] | 0.77 | 30.25 | 0.023 | 0.93 | 31.34 | 0.034 |
| DPS-UQ [13] | 0.89 | 34.95 | 0.01 | 0.94 | 32.83 | 0.013 |
| HYPERDM | 0.87 | 35.16 | 0.01 | 0.95 | 33.15 | 0.012 |

Table 1: **Comparison of training and inference times.** The time required to train an $M = 10$ member ensemble on the LUNA16 dataset is shown in the second column. The third column shows the time required to generate a predictive distribution of size $M \times N = 1000$ for a single input.

| METHOD | TRAINING TIME (MINUTES) | EVALUATION TIME (MINUTES) |
|---|---|---|
| MC-DROPOUT [16] | 47.03 | 3.70 |
| DPS-UQ [13] | 441.09 | 3.31 |
| HYPERDM | 48.53 | 3.18 |

# Summary

We propose HyperDM, a single-model method that can efficiently estimate both *aleatoric* and *epistemic* uncertainty.

- Advantages:
  - **vs. deep ensembles**: HyperDM offers comparable performance at a fraction of the computational training cost.
  - **vs. Monte-Carlo dropout:** HyperDM predictions and uncertainty estimates significantly outperform Monte-Carlo dropout.
  - **vs. Bayesian neural networks:** HyperDM training and inference is much faster than Bayesian neural networks because it doesn't require per-layer weight sampling.

- Future work:
  - **Scalability**: the number of hyper-network parameters scales proportionally with the size of the primary network.

# End