

Using Dropout to Capture Uncertainty: A Novel Application to B-tagging

Binbin Dong^{1,2}

¹ Shanghai Jiao Tong University

² Oklahoma State University

DPF21
July 12-14, 2021



Motivation

- ▶ Deep learning techniques have gained tremendous attention, however such techniques
 - Are not 100% accurate
 - Typically do not capture model uncertainty
- ▶ Bayesian models offer a mathematically grounded framework to quantify model uncertainty, but usually come with a prohibitive computational cost.
- ▶ In 2016, Yarin Gal brought up an idea of using Dropout as a Bayesian approximation to represent model uncertainty ([paper](#)) .

Dropout Uncertainty Quantification (DUQ)

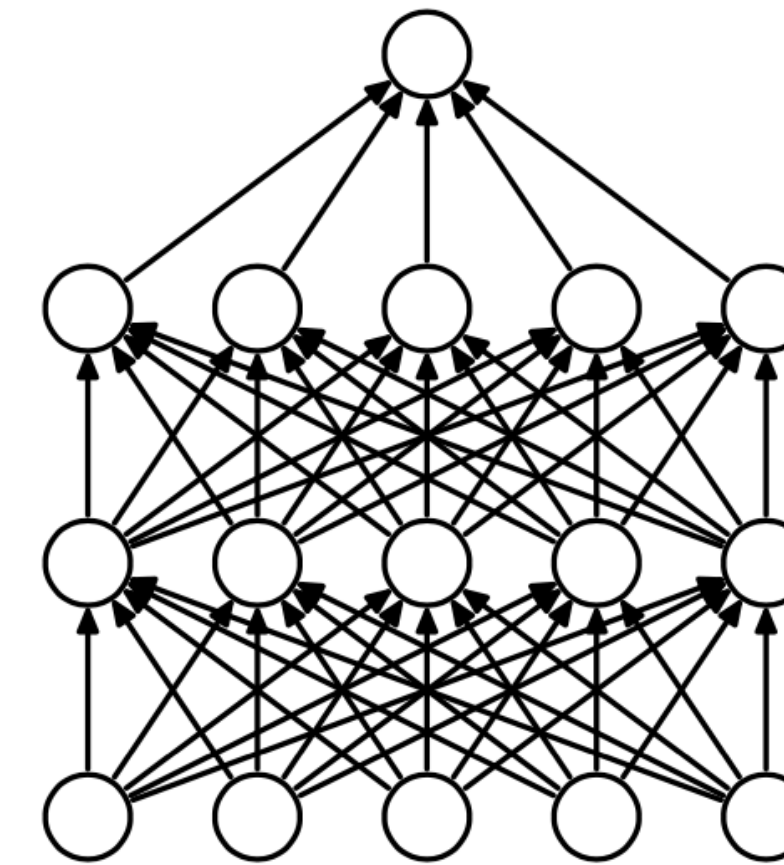
▶ Dropout:

- A standard technique for training neural networks.
- Avoids over-fitting by randomly deactivating connections between nodes of a neural networks during the training process.

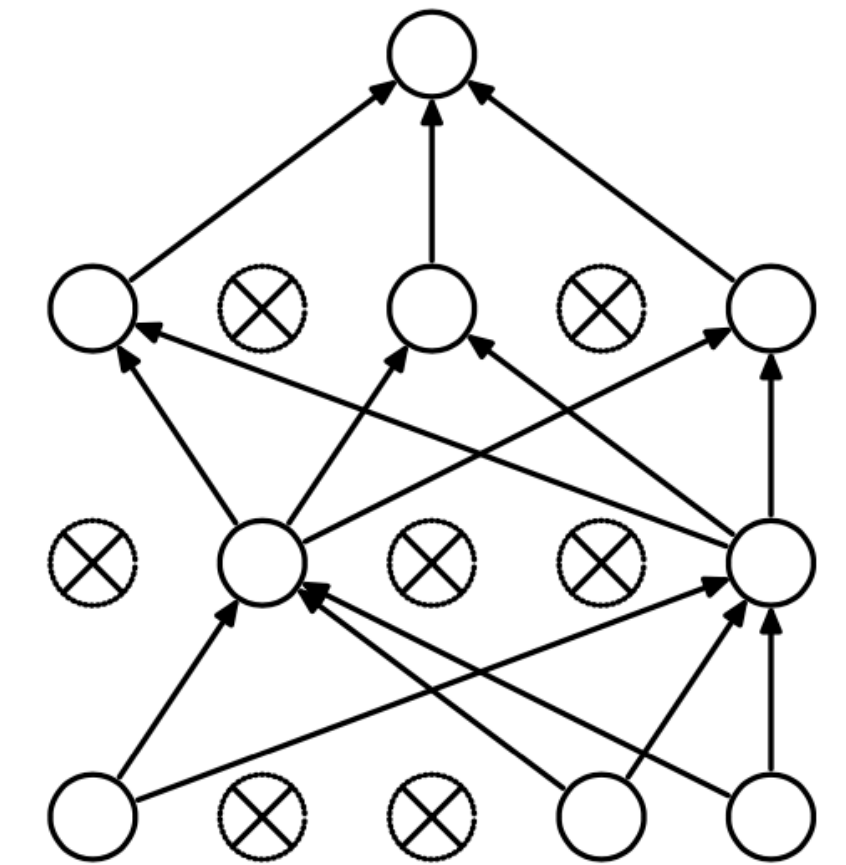
▶ Using Dropout to capture uncertainty proposed in 2016

- Claim by enabling Dropout during testing provides an approximation of the network's posterior probability distribution
 - But this claim is not rigorously justified nor systematically evaluated
 - Proposed to validate the method using a standard ML database

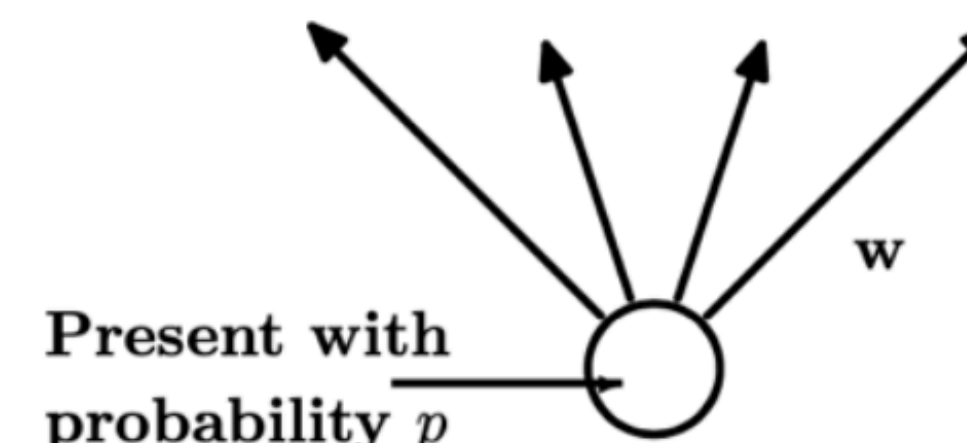
Figures from Dropout [paper](#)



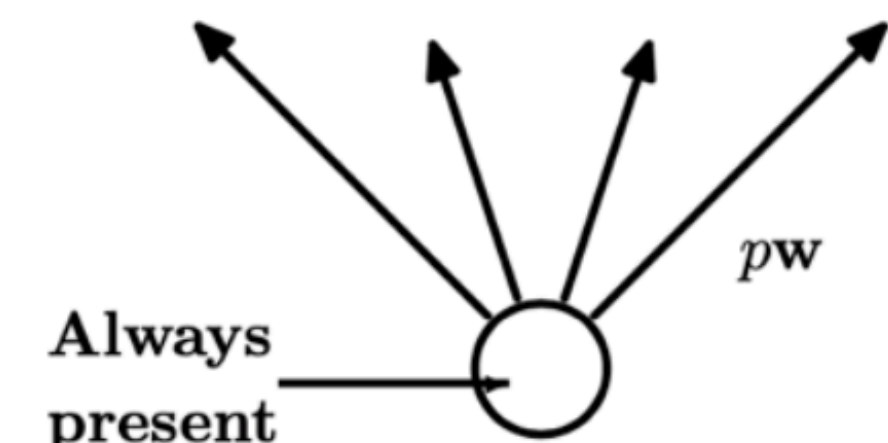
(a) Standard Neural Net



(b) After applying dropout.



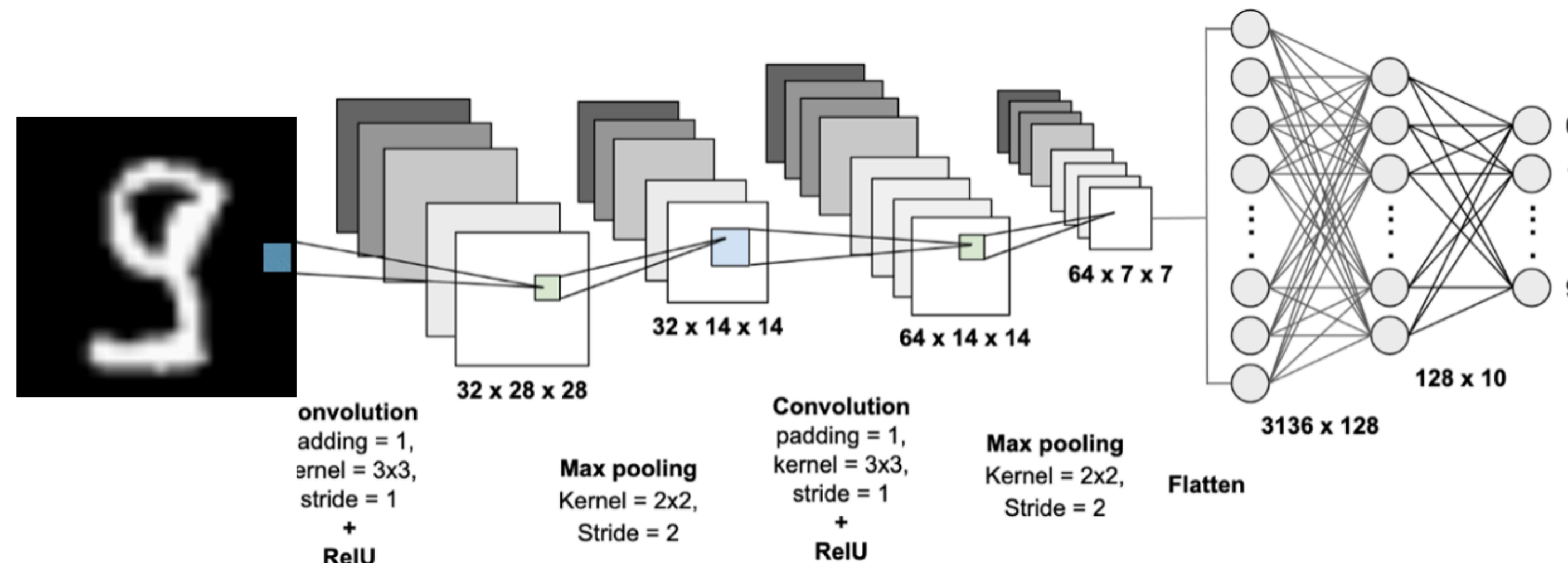
(a) At training time



(b) At test time

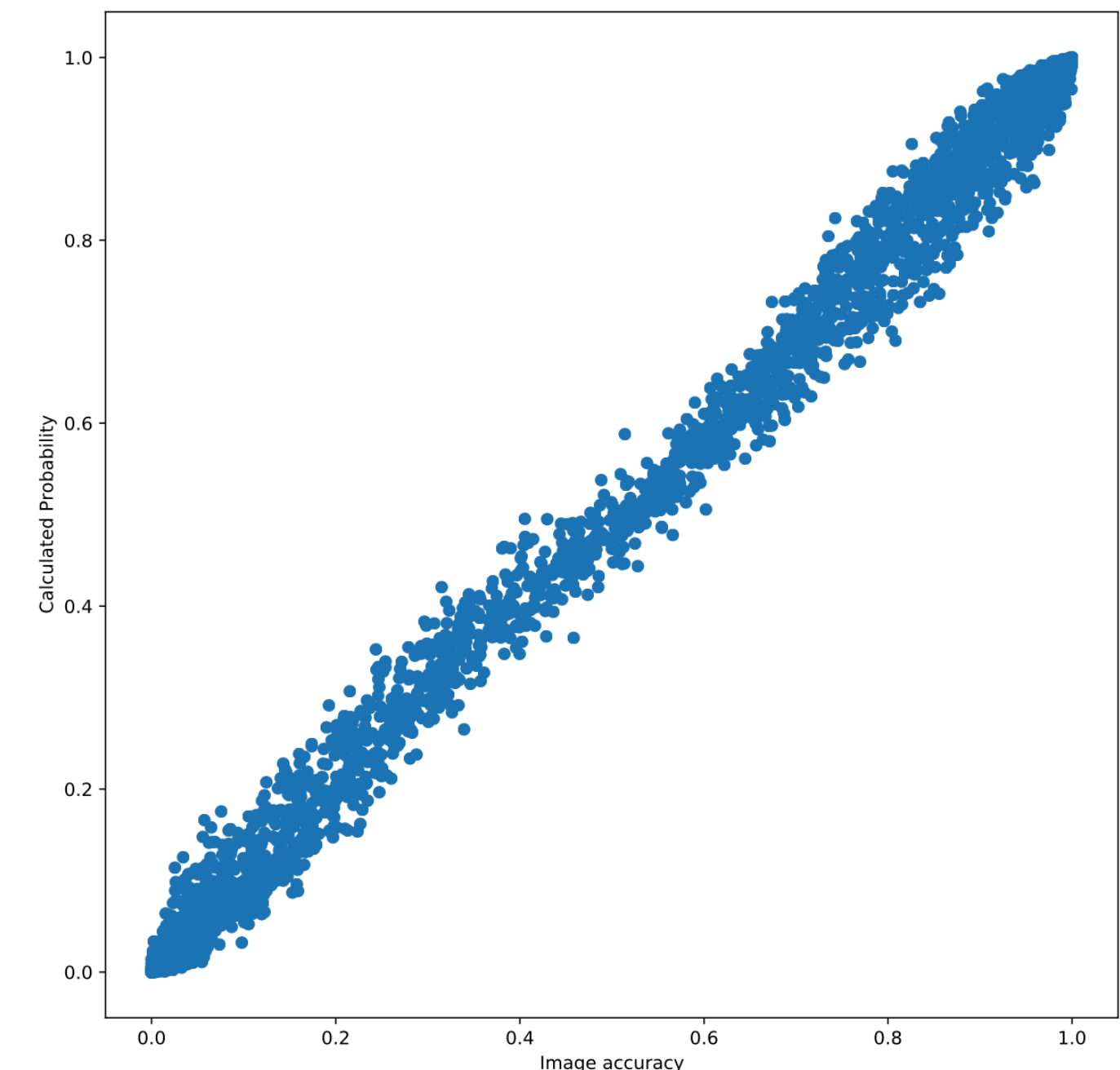
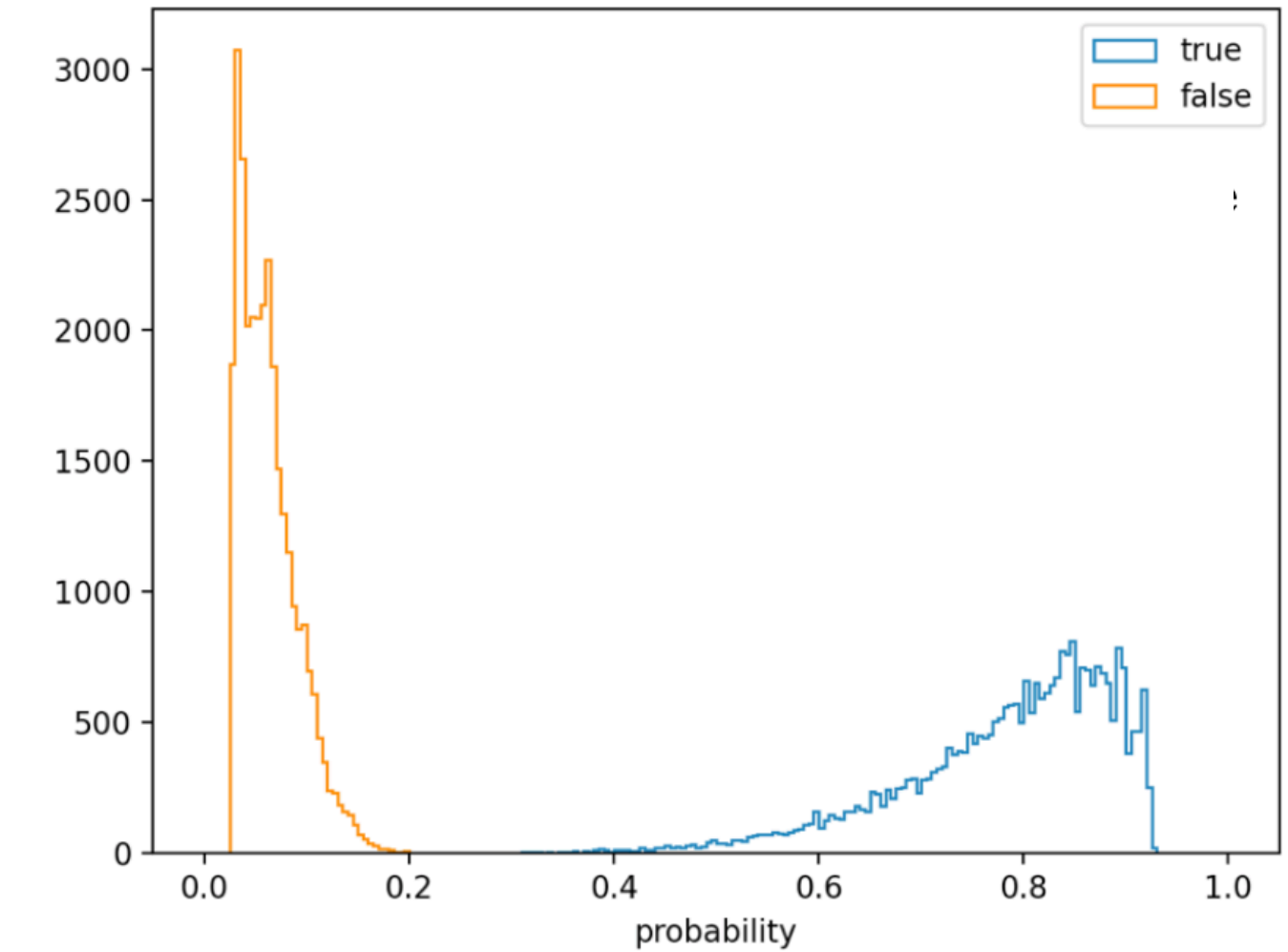
The MNIST Database

- ▶ The MNIST database:
 - A database of handwritten, black and white digits from 0-9
 - Has a training set of 60k images, and a test set of 10k images
 - All images are normalized to fit into a 28 x 28 pixel box
- ▶ Trained on the MNIST dataset for multi-classification studies
 - With a Convolutional Neural Network which contains 2 hidden-layers. For the 2nd hidden layer, Dropout was enabled with probability $p = 0.2$



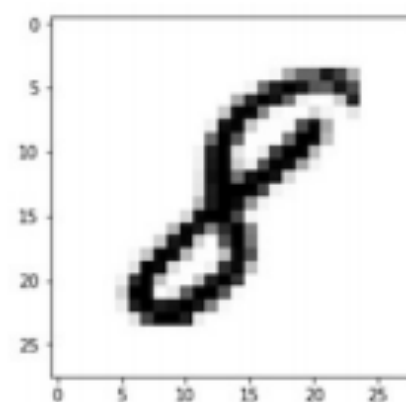
Network Output Processing

- ▶ Multiple evaluations on each image with Dropout enabled to get image posterior probability distribution
 - Calculate mean and asymmetric 68% Confidence Interval (CI)
- ▶ Perform a closure test by comparing the probability to the accuracy of correctly classify an image
 - Significance calculation:
$$\text{significance} = \frac{\mu_{true} - \mu_{false}}{\sqrt{(\mu_{true} - CI_j)^2 + (\mu_{false} - CI_j)^2}}$$
 - Images' probability which correspond to a correct categorization is calculated as the cumulative probability distribution of the calculated significance
 - Observe 52.4% sample accuracy vs predicted 52.5% sample accuracy

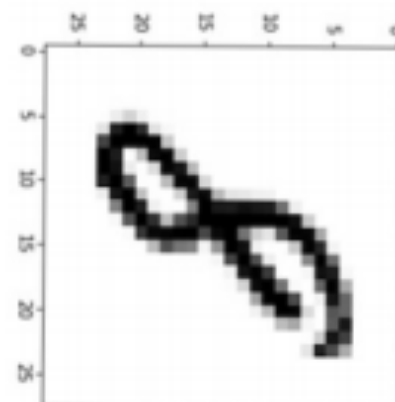


Bias Test

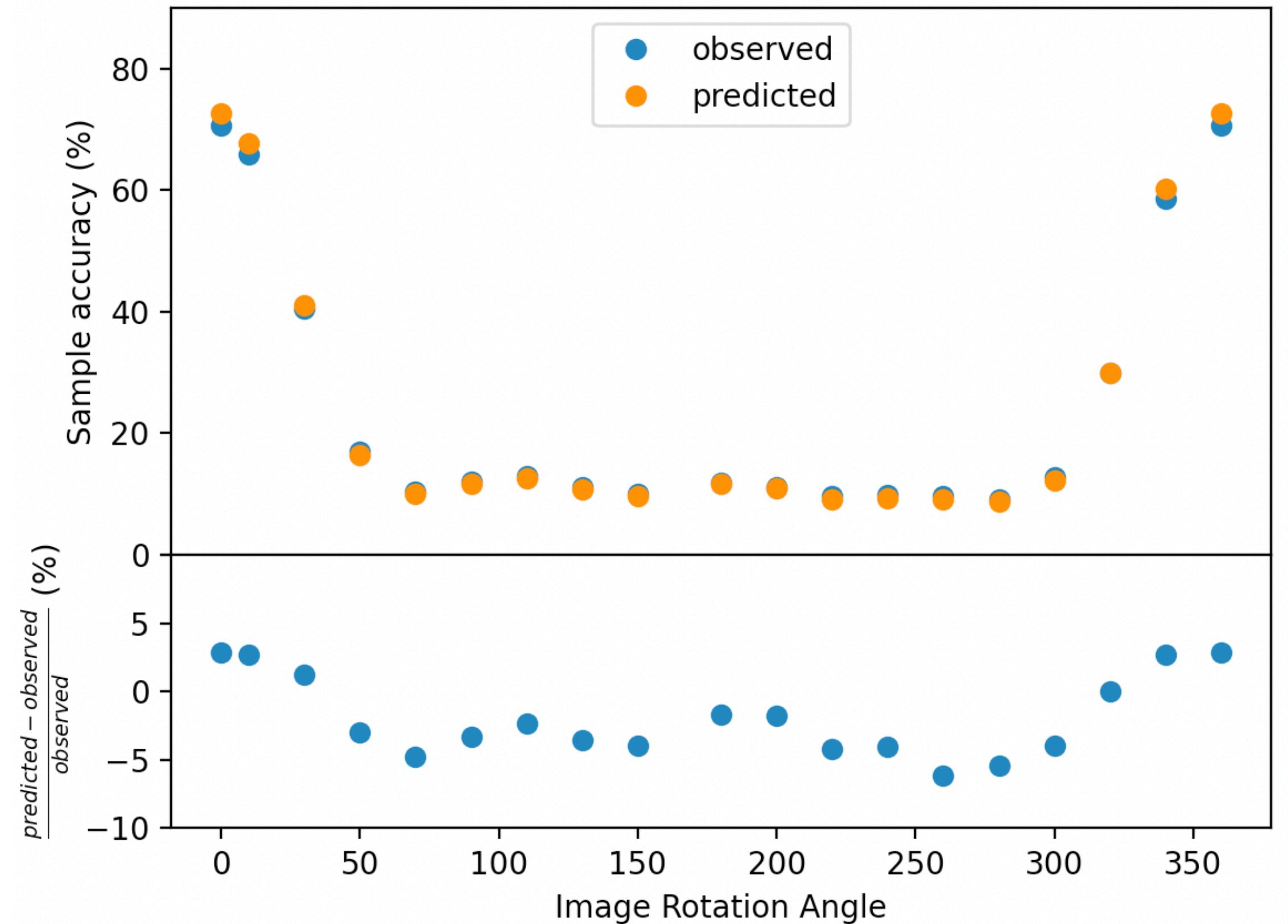
- ▶ Biases test performed by rotating images of the testing dataset from 0 to 360°
- ▶ Sample accuracy dropped fast with the rotation
 - But with DUQ method, the relative difference is within 5%
- ▶ With biases, DUQ method to capture uncertainties still work well



Training

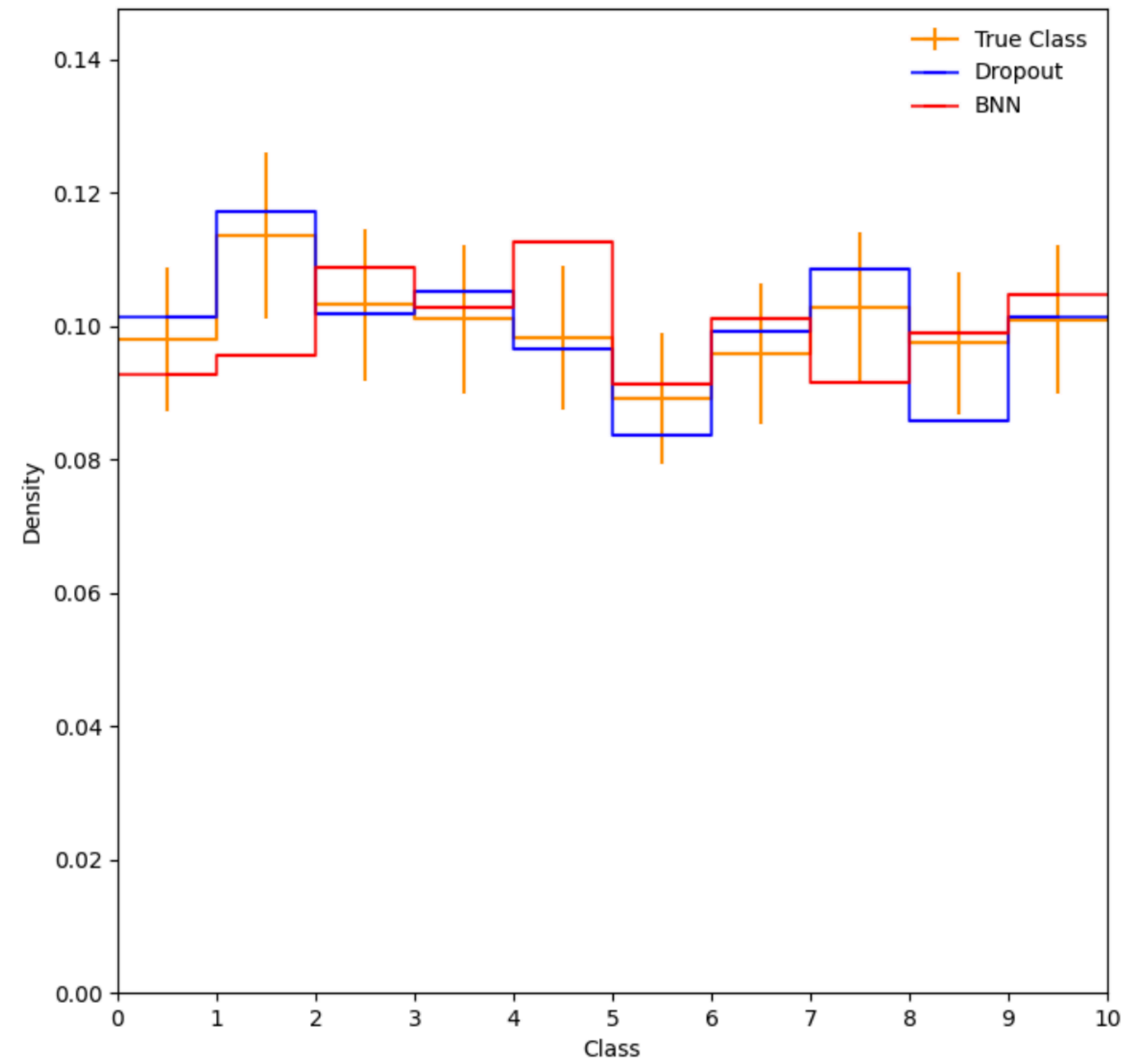


Testing



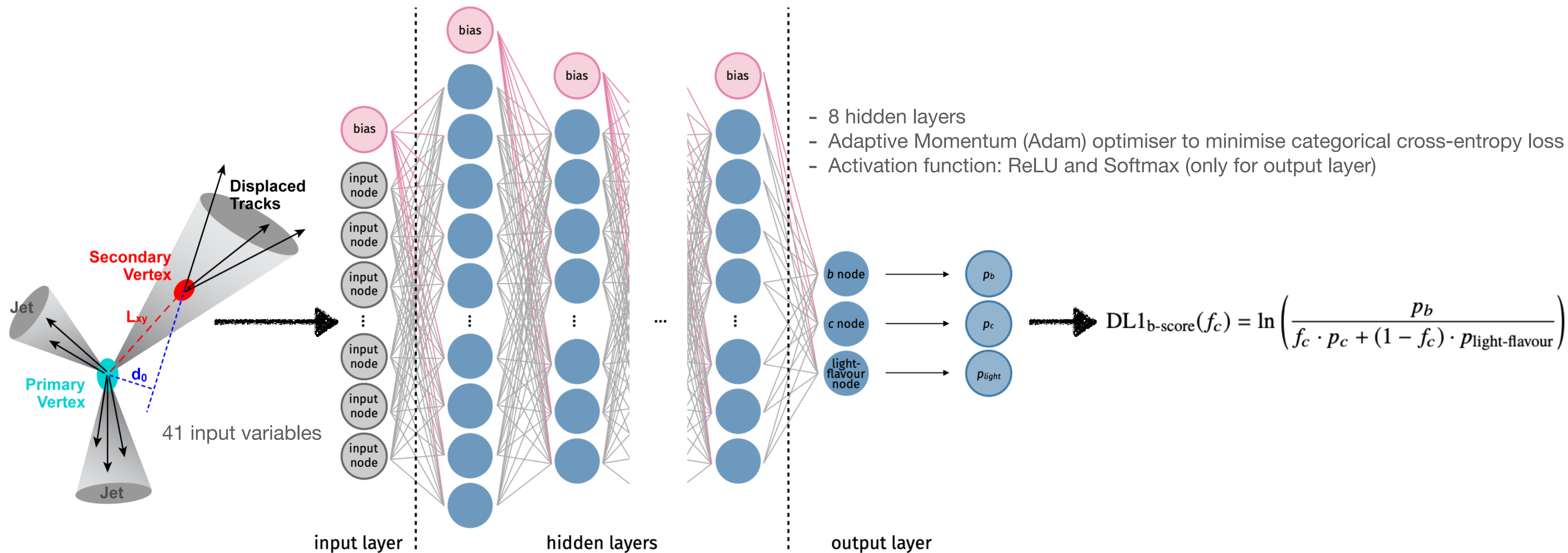
Comparison to Other Methods

- ▶ Bayesian Neural Networks offer a mathematically grounded framework to quantify model uncertainty
 - Each weight in neural net is given a prior and a Gaussian uncertainty
 - Fit both weights and model uncertainty
 - Downsides:
 - Doubles the number of parameters in a network
 - Comes with a prohibitive computational cost
- ▶ Performance between DUQ and BNN is comparable
 - Poisson uncertainties added on True probability
 - DUQ may provide a more accurate prediction



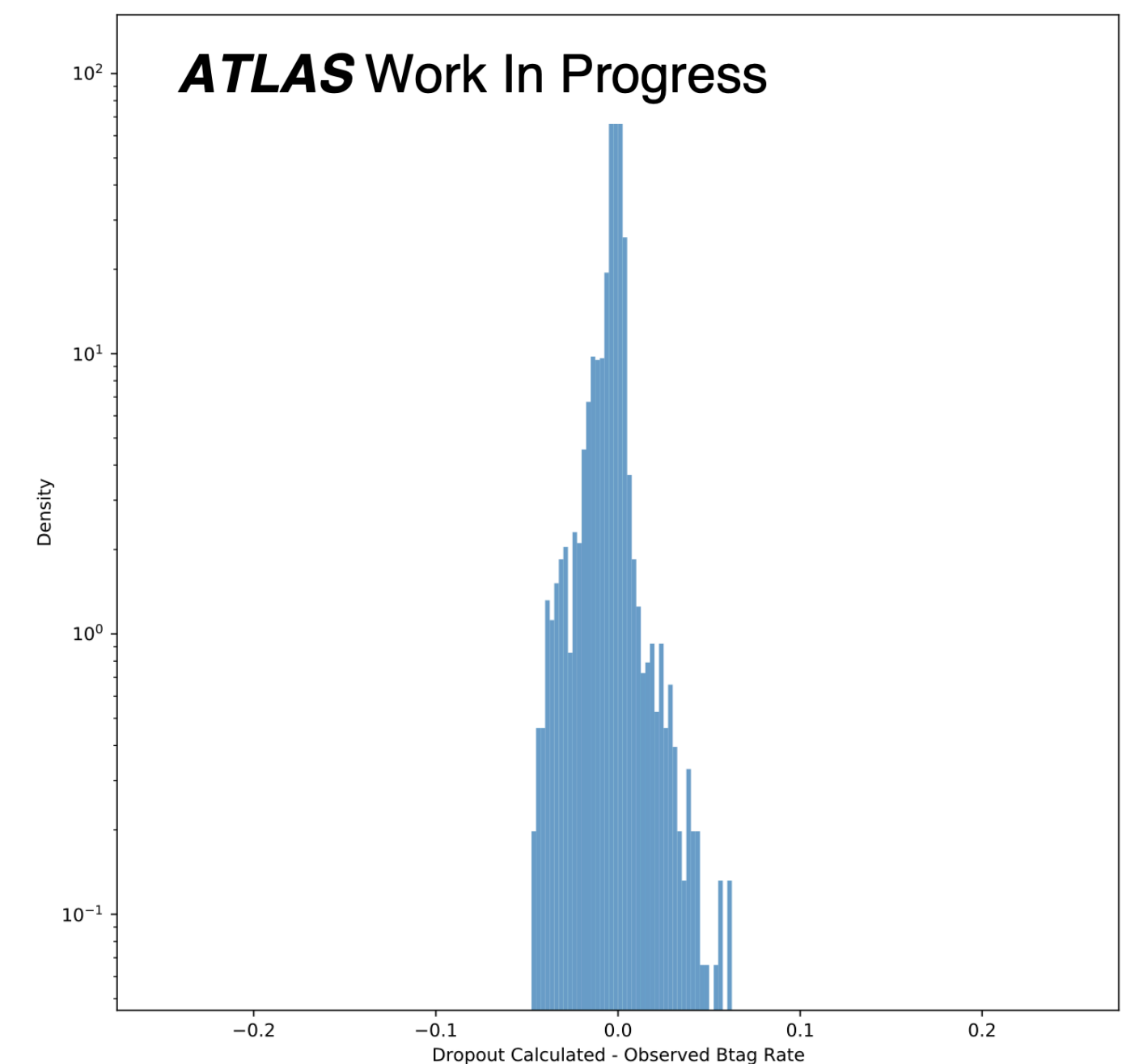
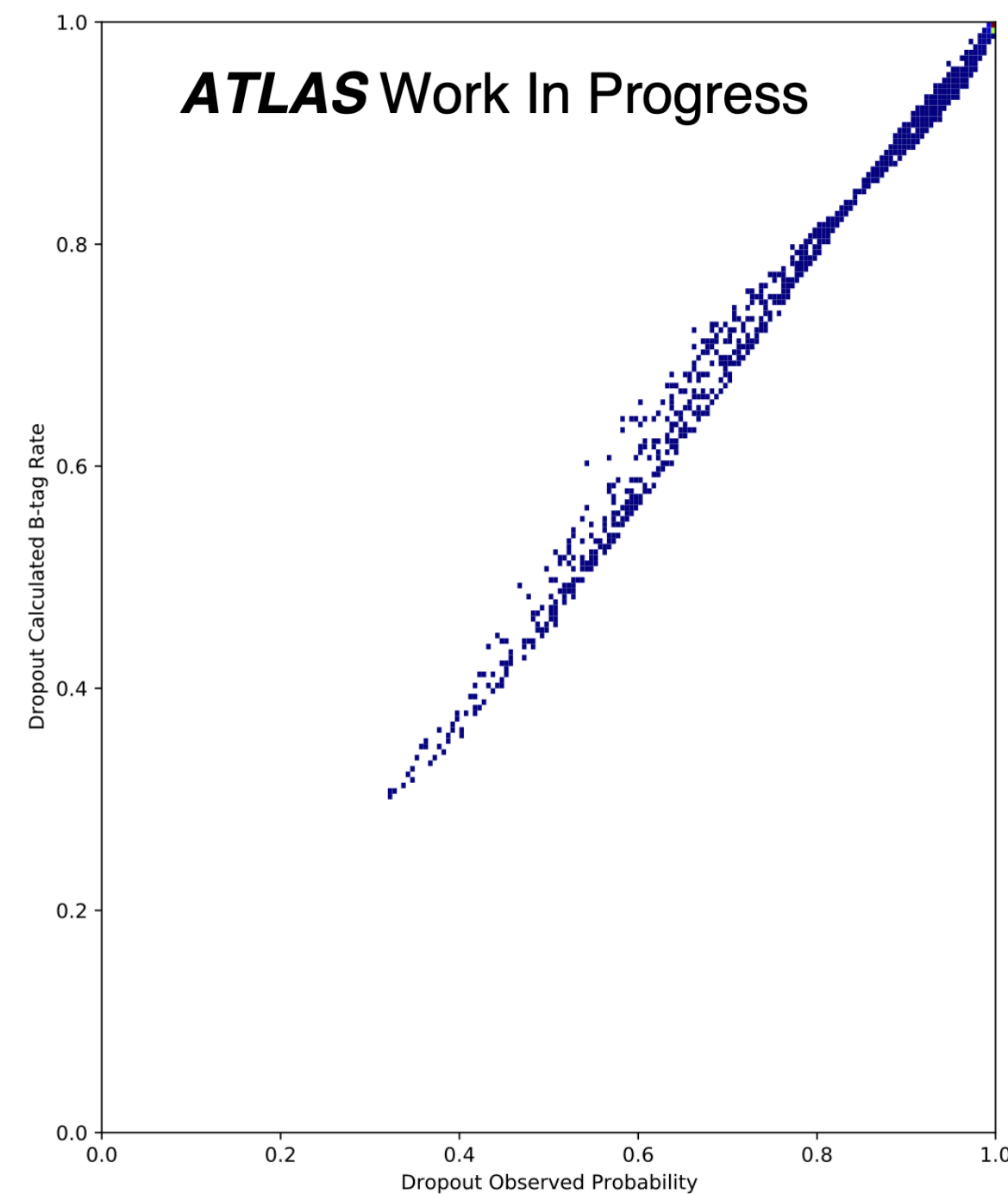
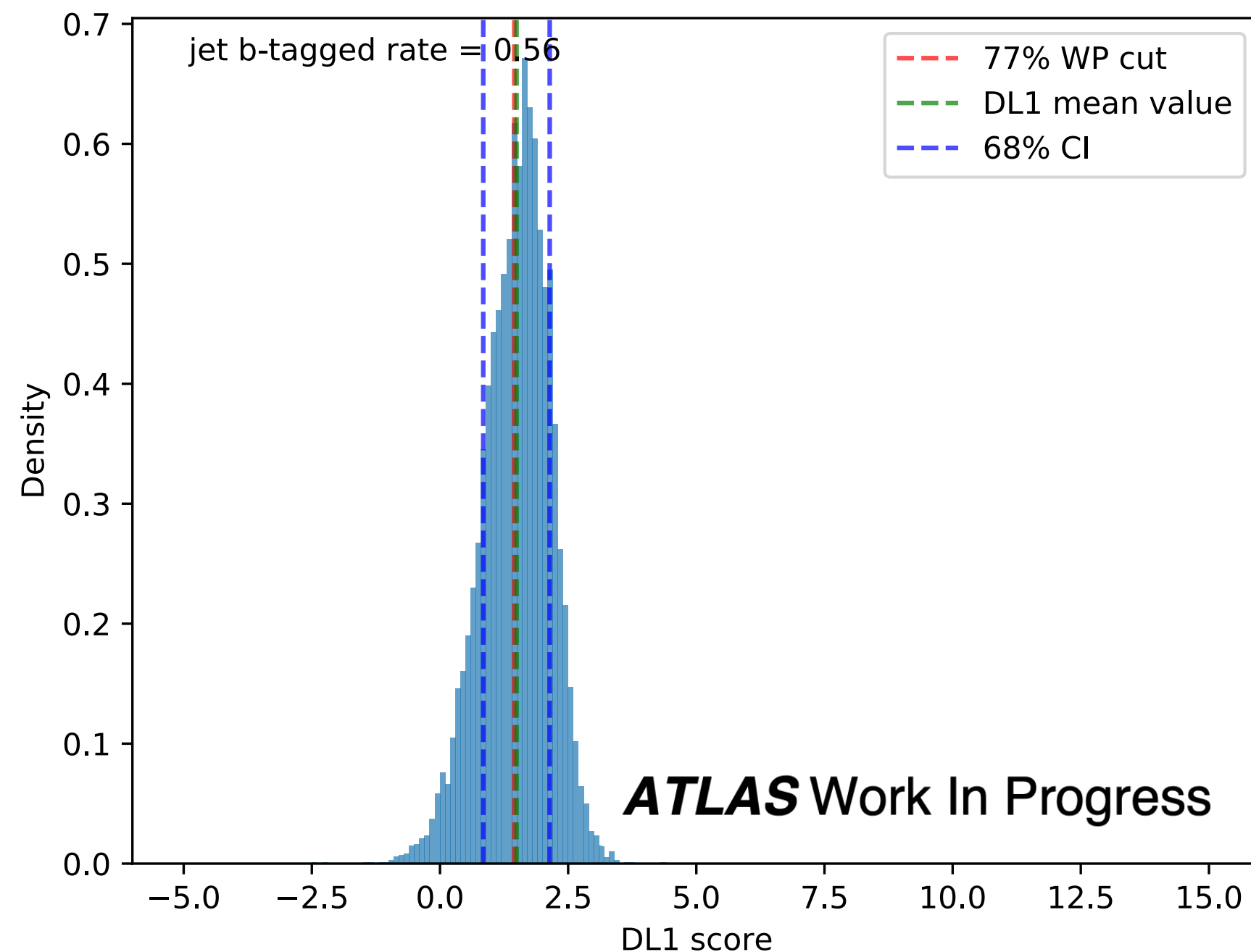
B-tagging in ATLAS

- ▶ Average b-hadron can travel ~mm before decaying
- ▶ Deep learning technique is used to identify jets containing b-hadrons



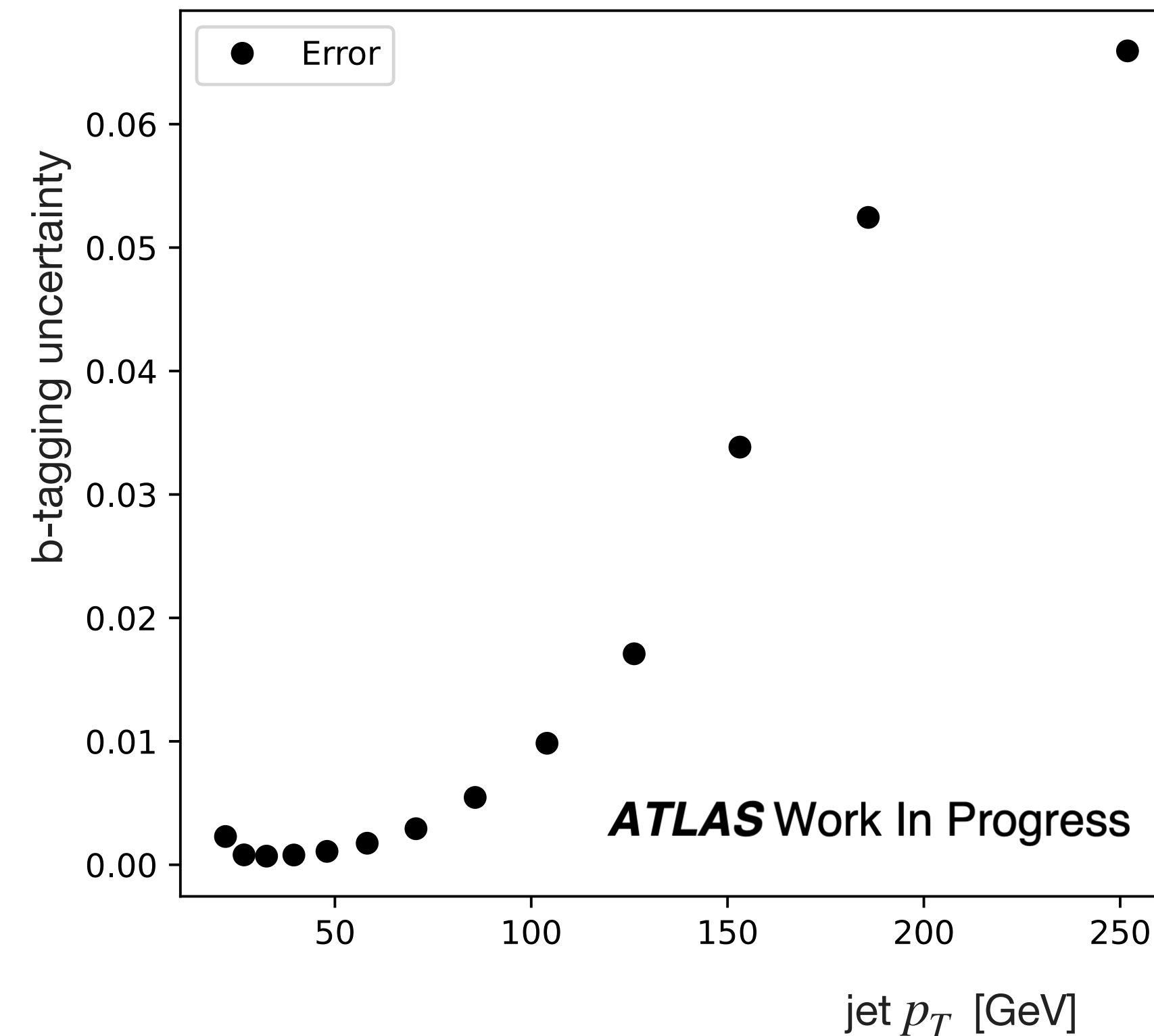
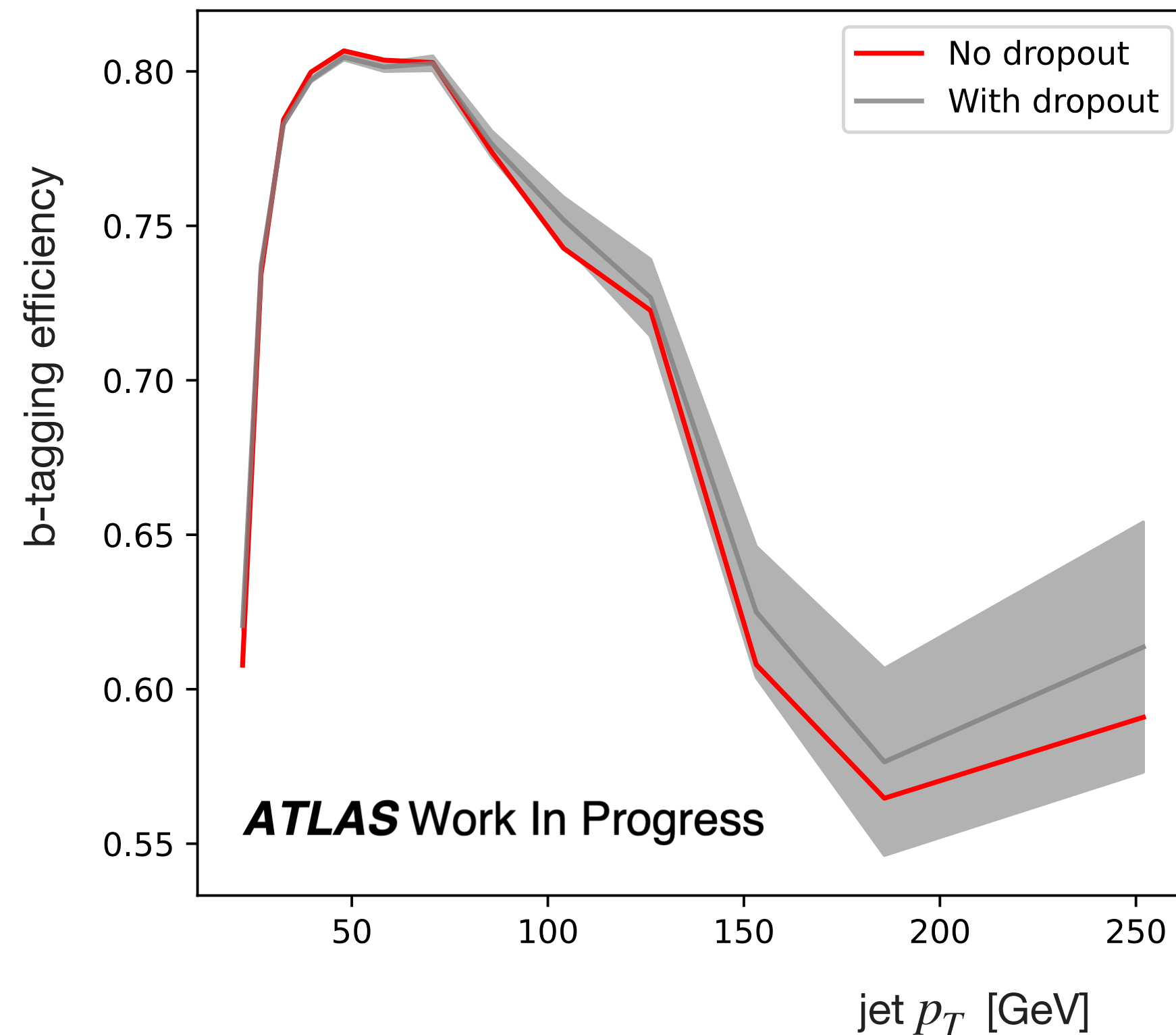
DUQ Application to B-tagging

- ▶ Repeat the MNIST procedure of calculating probability from significance with Dropout enabled during evaluation
- ▶ Calculated vs Observed
 - Quite diagonal, indicates calculated probability can well reflect image accuracy
 - The difference is centered at 0 with a width of 2%



DUQ Application to B-tagging

- ▶ DUQ method performed to get b-tagging efficiency as a function of jet p_T
- ▶ Sample jet p_T up to 250 GeV, within $\sim 7\%$ uncertainty noticed



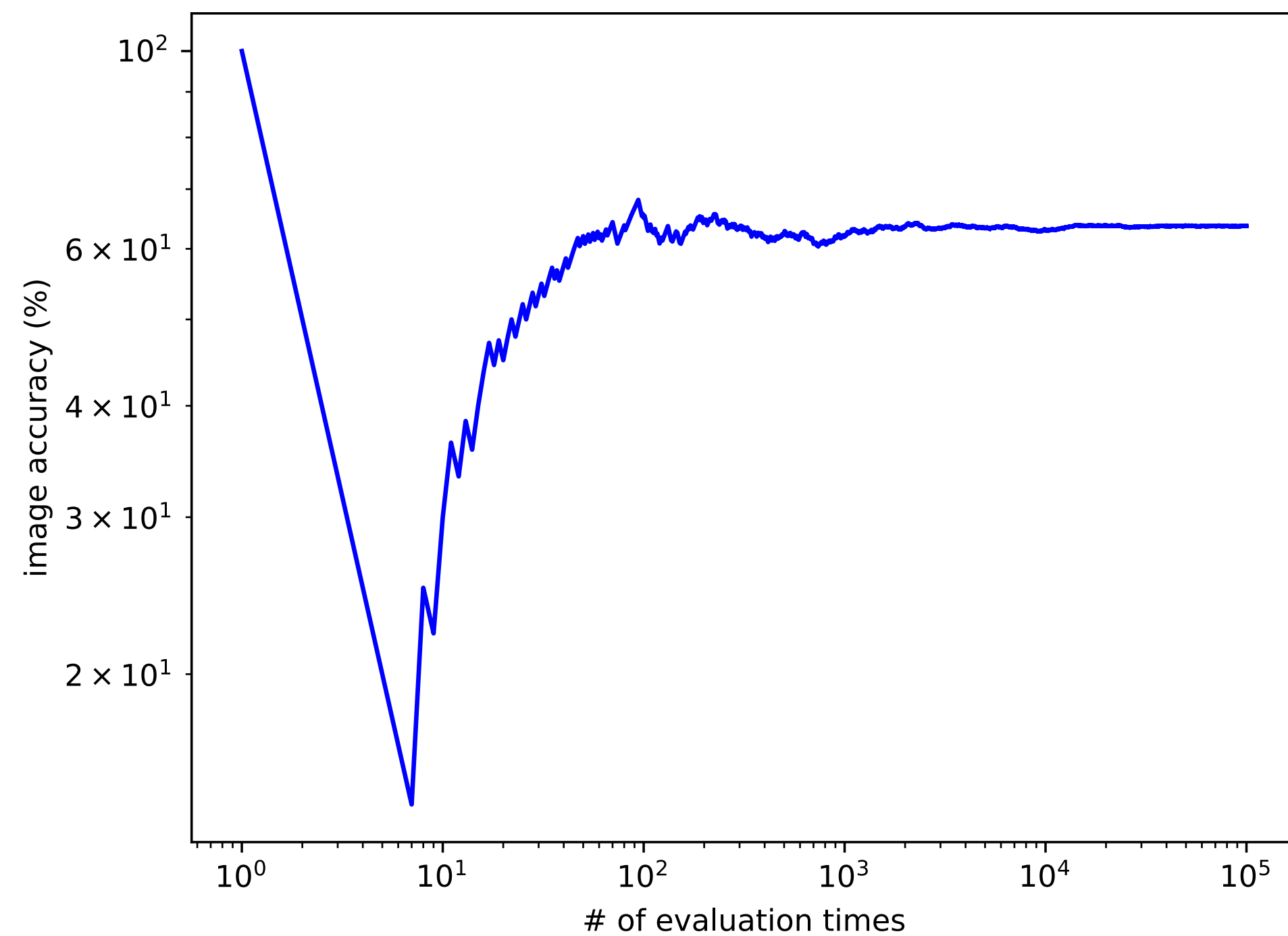
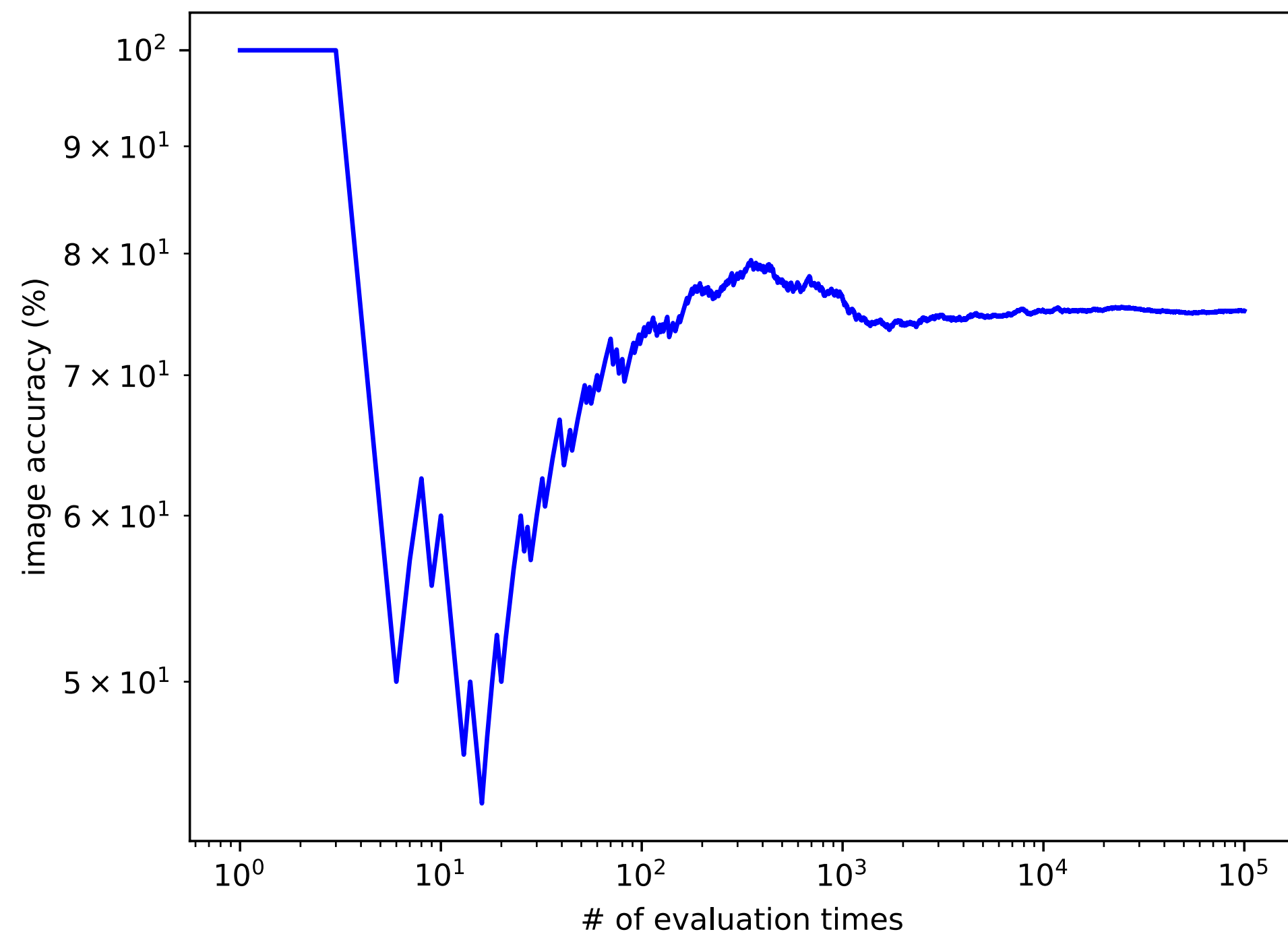
Summary

- ▶ Using Dropout to capture uncertainty:
 - Enabling Dropout during evaluation samples the posterior probability distribution
 - Calculate per object significance and categorization probability using the mean and asymmetric 68% confidence interval
- ▶ Method tested on the MNIST database
 - The probability accurately predicts image and sample accuracies
 - Bias tested performed to verify the method can accurately accounts for systematic mis-modeling
- ▶ Applied DUQ method to ATLAS b-tagging
 - Unbiased closure test accurate to $\sim 2\%$

Backup

Stability

- ▶ Number of evaluation times for each image are important for the method
 - In principle, the more the better. But it costs more computational resources
 - Find a point where the image accuracies are stable
 - 3k evaluations are enough in the case



Model Dependency Check

- ▶ Model dependency check
 - Tested the DUQ method on different NN models and activations (Sigmoid, ReLu, Logistic, RNN, etc.)
 - Some model dependence is observed
 - But still worked at some level with all the models we tested

