# Using Dropout to Capture Uncertainty

**Binbin Dong**[1,2] (binbin.dong@cern.ch)
[1] Shanghai Jiao Tong University
[2] Oklahoma State University

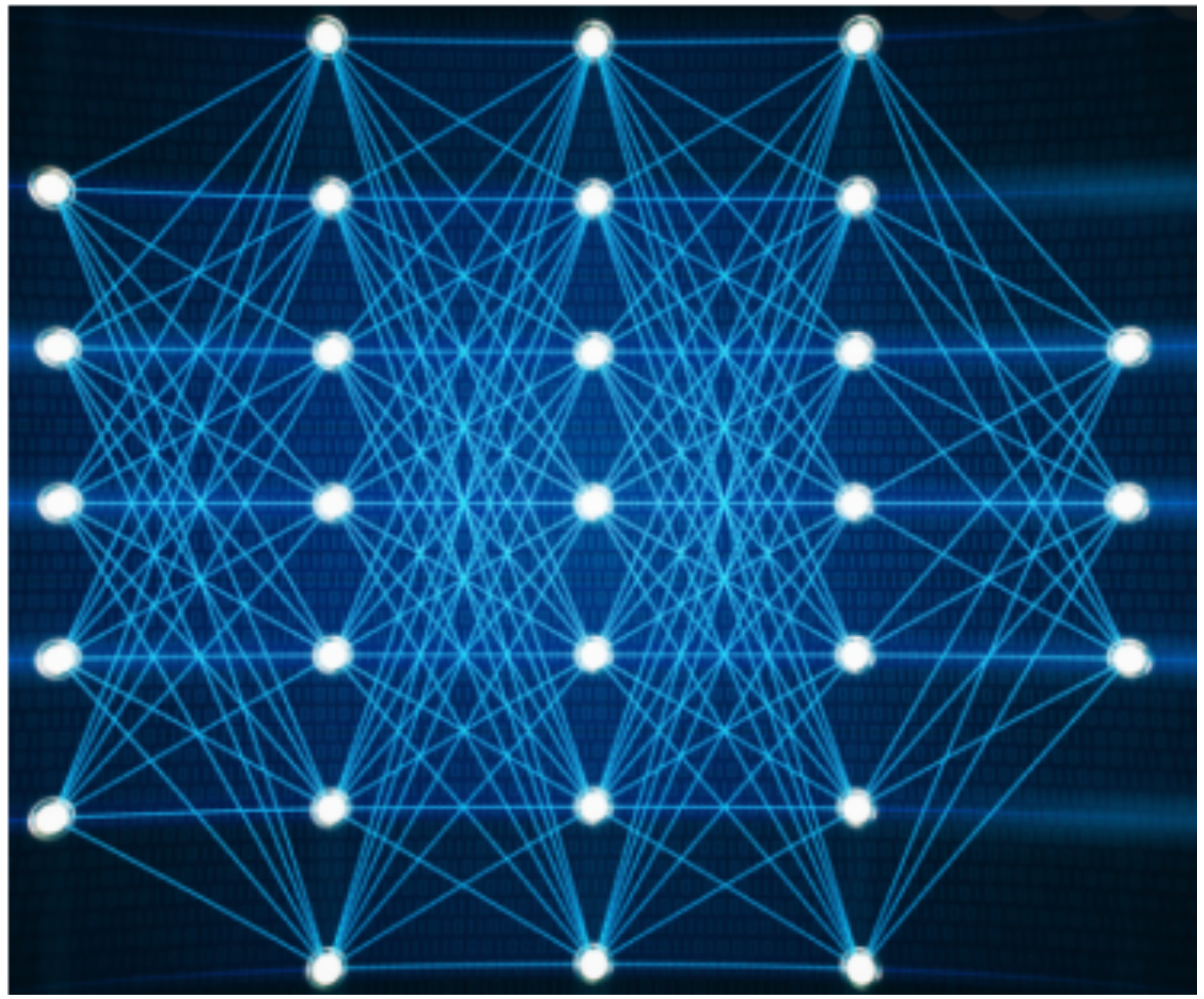
**Joint work with:**
**Yiming Abulaiti, Tyler Burch, Alexander Khanov, Jeremy Love, Flera Rizatdinova, Ning Zhou**

**Argonne AI & HPC seminar**
**August 27, 2021**

# Uncertainty quantification in DL

- ✤ Deep learning has gained tremendous attention in many field

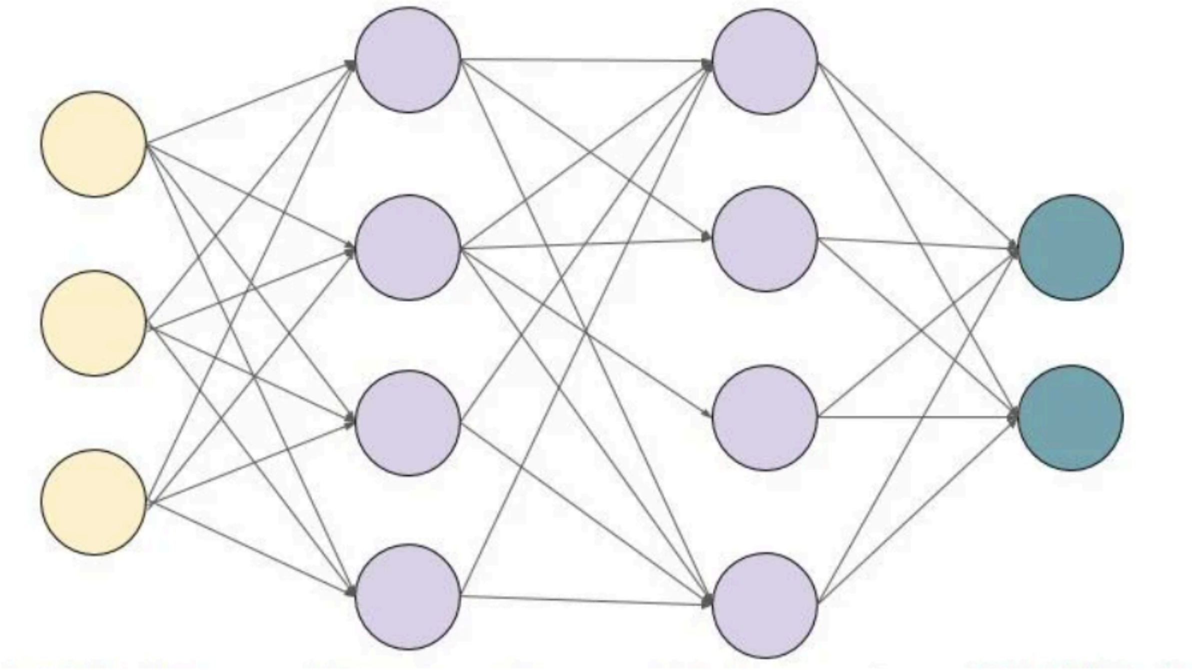
- ✤ Deep neural network model:
  - What does the output “probabilities” tell us?
  - How to tell if the model is making sensible predictions or giving random answers?
  - Does the model know what it doesn’t know?



- ✤ Uncertainty quantification can help us understand if our model is confident

# Out of distribution data

- Train: cats vs dogs images

Cat ?
Dog?

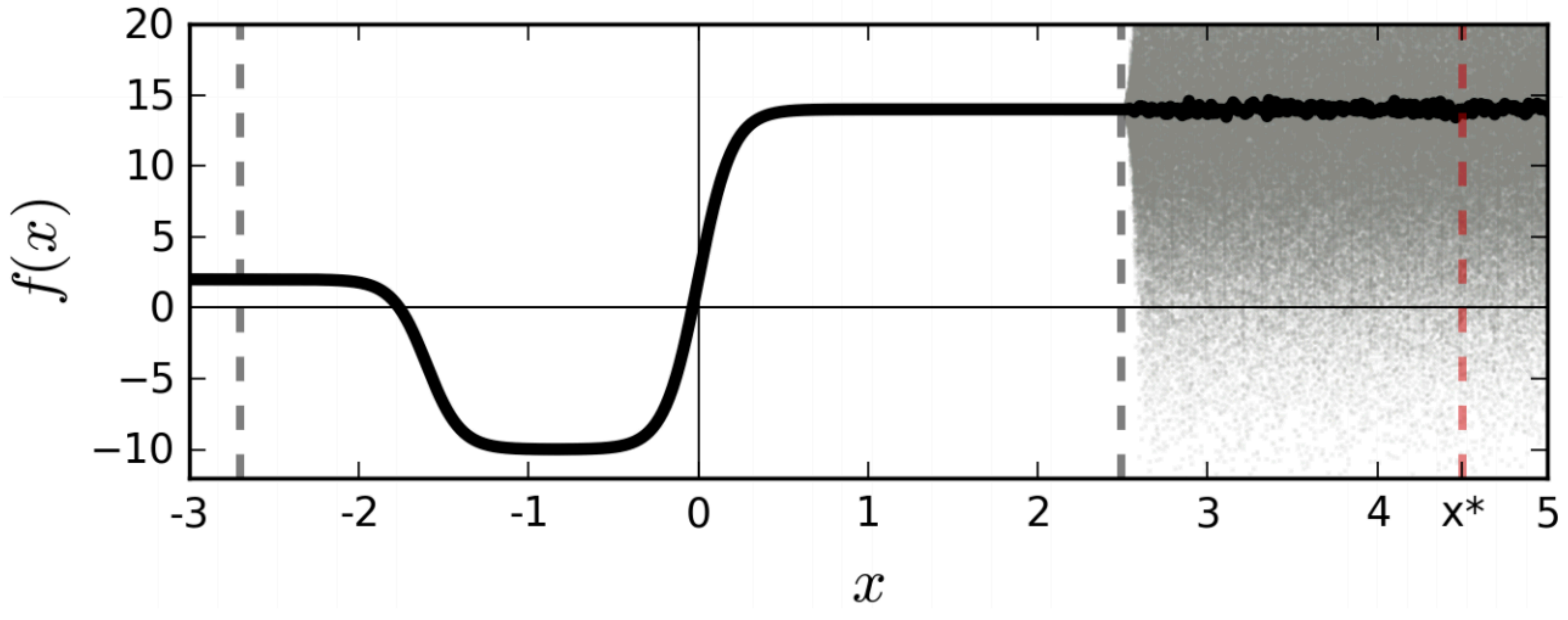
- During testing, a bird image enters
  - What would the model tell us?

?

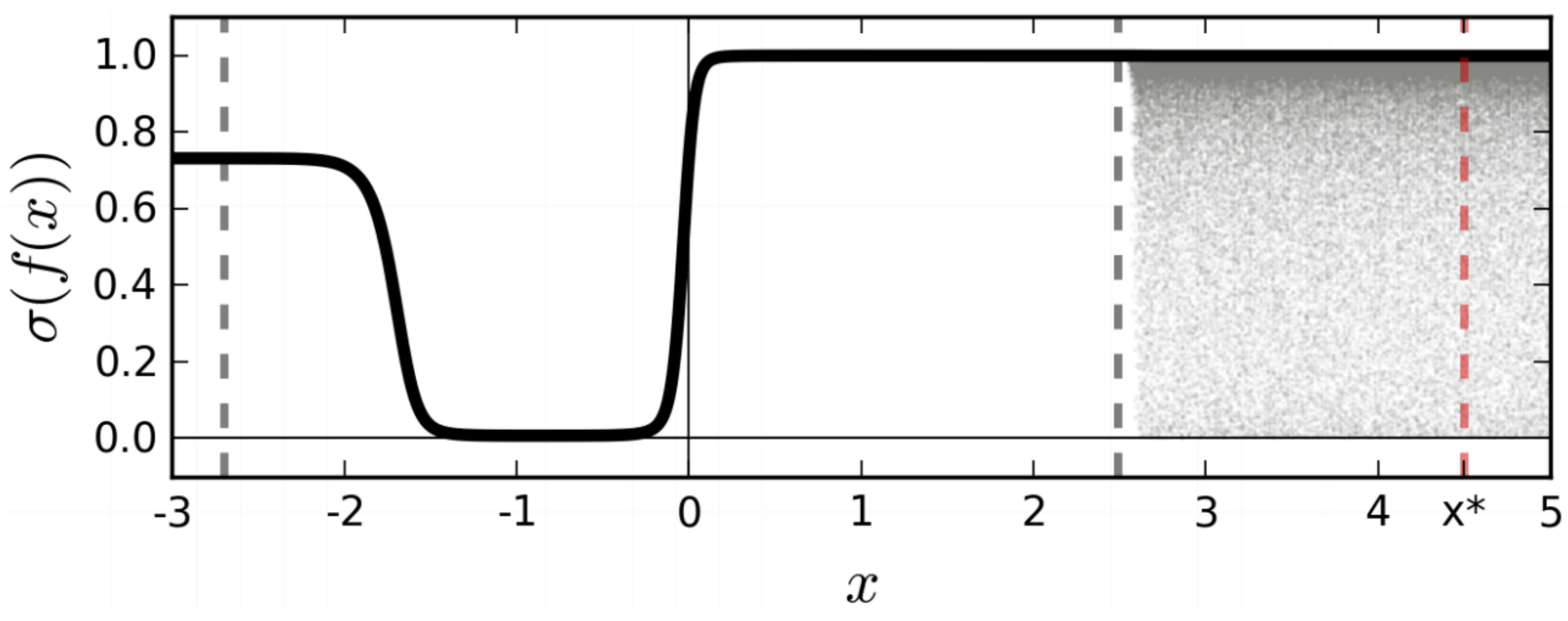
# Out of distribution data

- A sketch of softmax input and output for an idealized binary classification problem
  - Training data is given between the dashed grey lines
  - Function point estimate is the solid black line
  - Dashed red line is a point far from the training data
- Without uncertainty, a bird image can be classified as cat/dog with probability 1

Figure from paper: arxiv 1506.02142




(a) Arbitrary function $f(\mathbf{x})$ as a function of data $\mathbf{x}$ (softmax *input*)

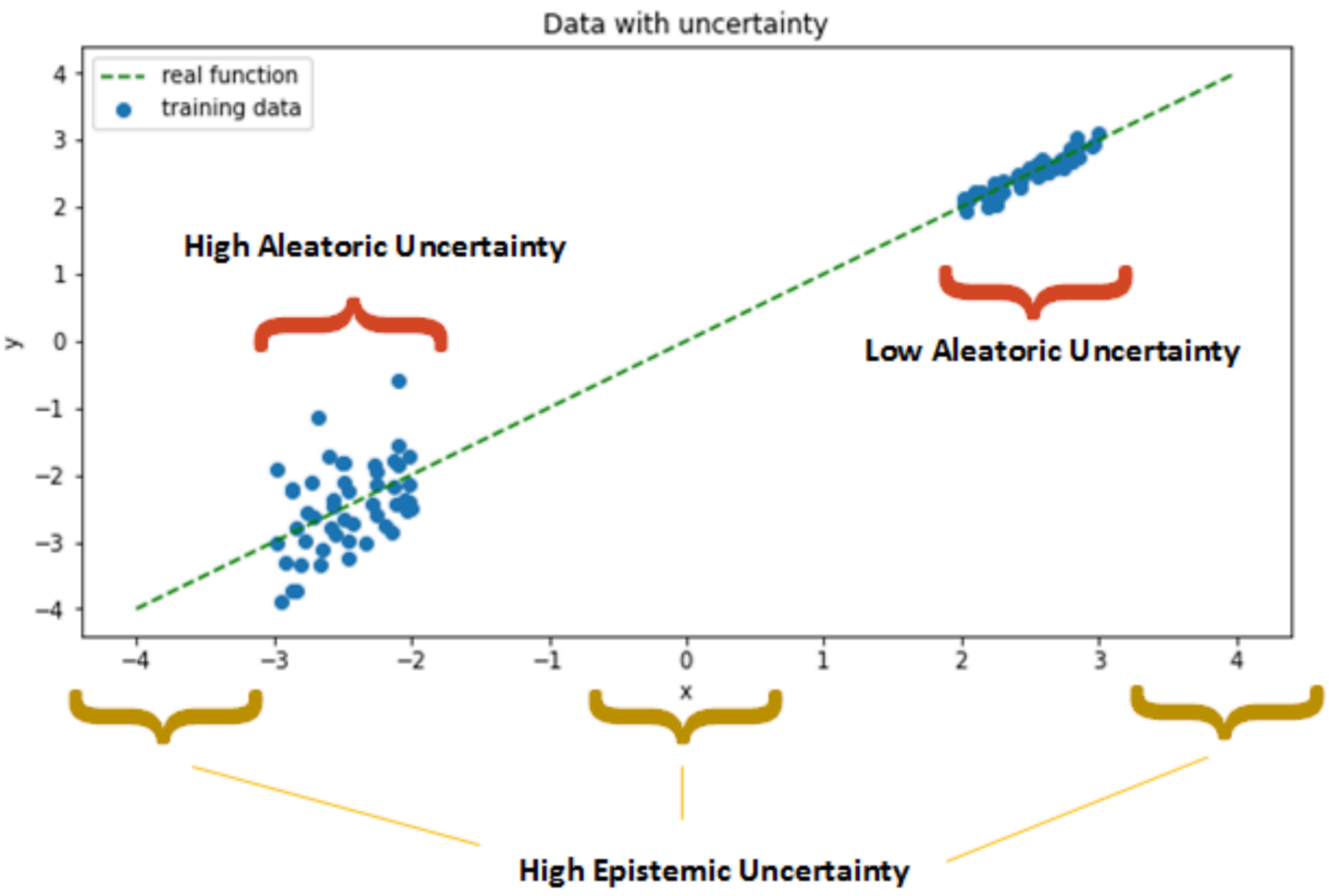



(b) $\sigma(f(\mathbf{x}))$ as a function of data $\mathbf{x}$ (softmax *output*)

# Types of uncertainties

- Epistemic uncertainty (also referred to as model uncertainty):
  - Describes what the model doesn't know due to limited data and knowledge on model parameters
  - Reduces when having more data

- Aleatoric uncertainty:
  - Raises from the natural stochasticity of observations
  - Non-reducible



# Types of uncertainties

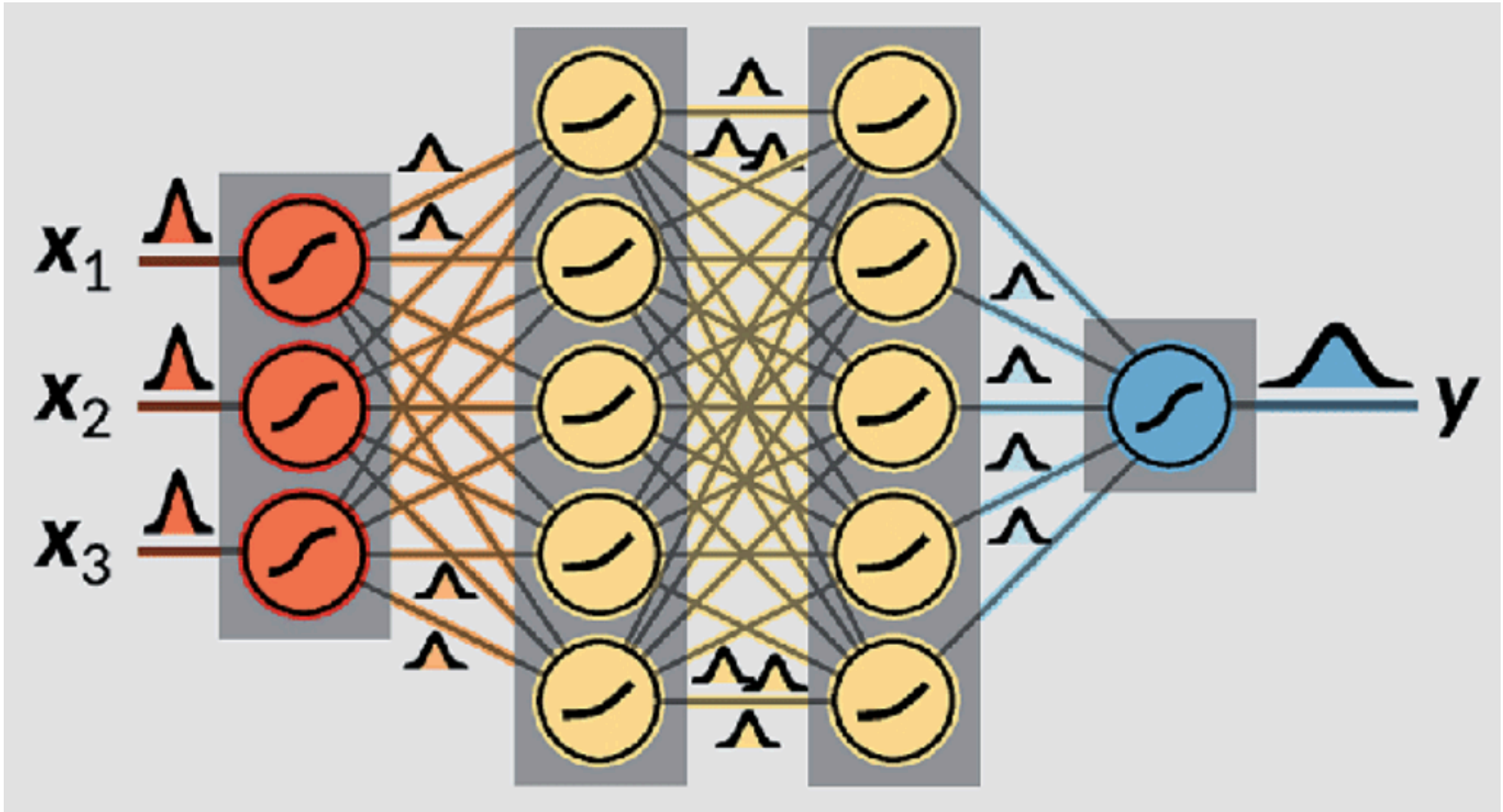
- Epistemic uncertainty (also referred to as model uncertainty):
  - Describes what the model doesn't know due to limited data and knowledge on model parameters
  - Reduces when having more data
- Aleatoric uncertainty:
  - Raises from the natural stochasticity of observations
  - Non-reducible

Can be used to induce predictive uncertainty:
- the confidence we have in a prediction

# Uncertainty quantification methods

Bayesian Neural Network

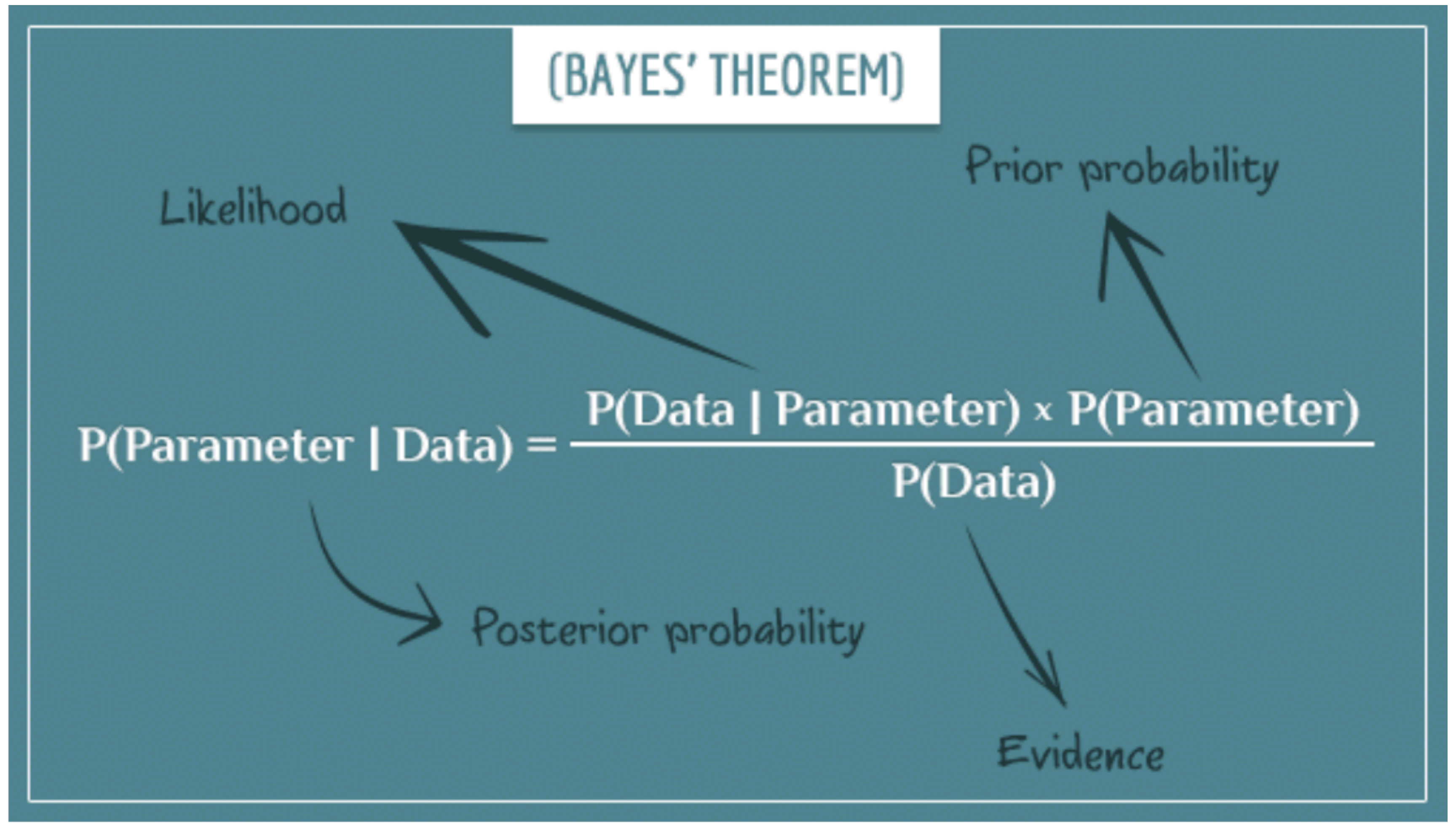
- Each weight in the neural net is given a prior and a Gaussian uncertainty
- Fit both weights and model uncertainty
- Posterior will be driven over model parameters




BNN models offer a mathematically grounded framework to quantify model uncertainty, and have been referred as a gold standard.

However the models:

- Double the number of parameters in a network, need more time for training
- Cost a prohibit computational resources
- Difficult to use

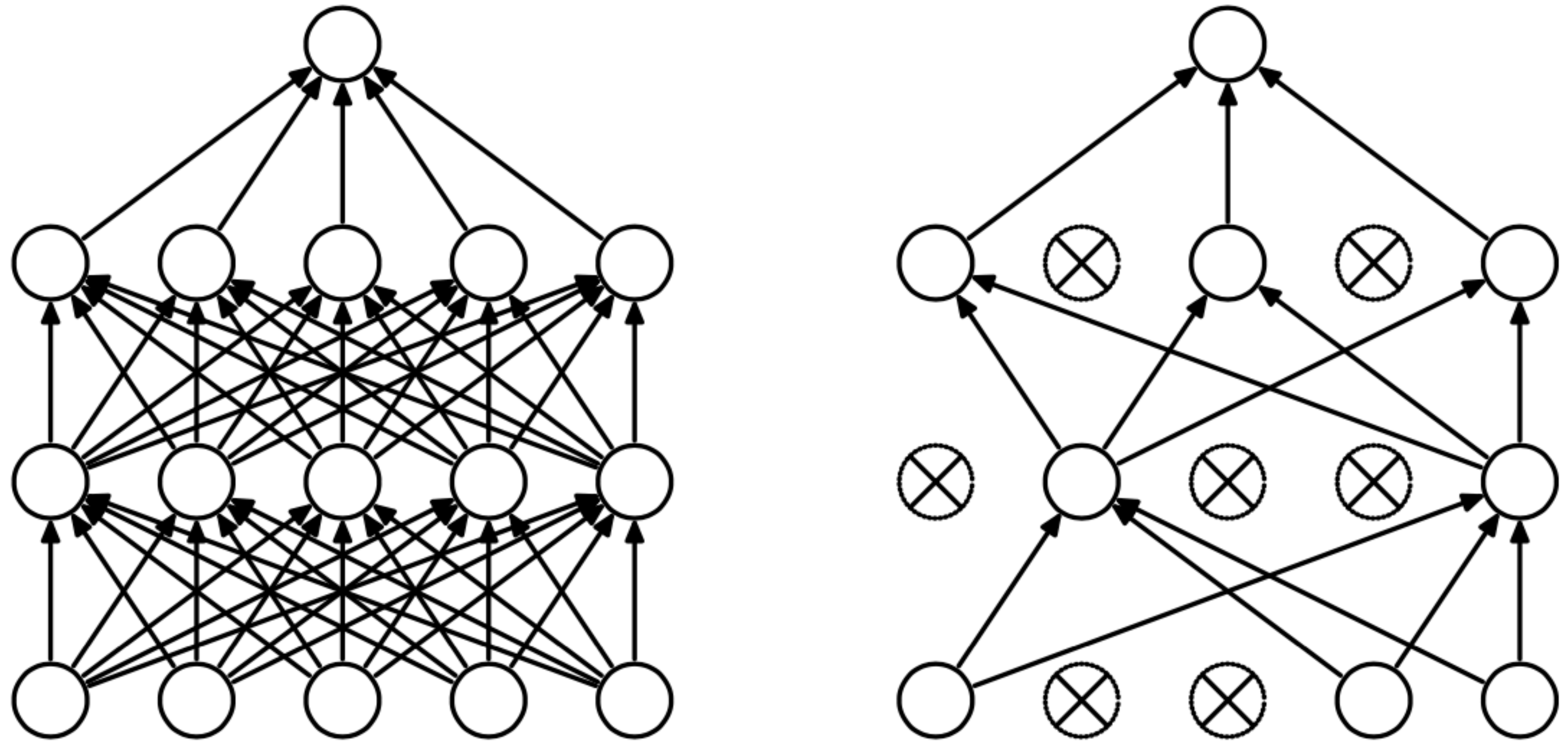


# Uncertainty quantification methods

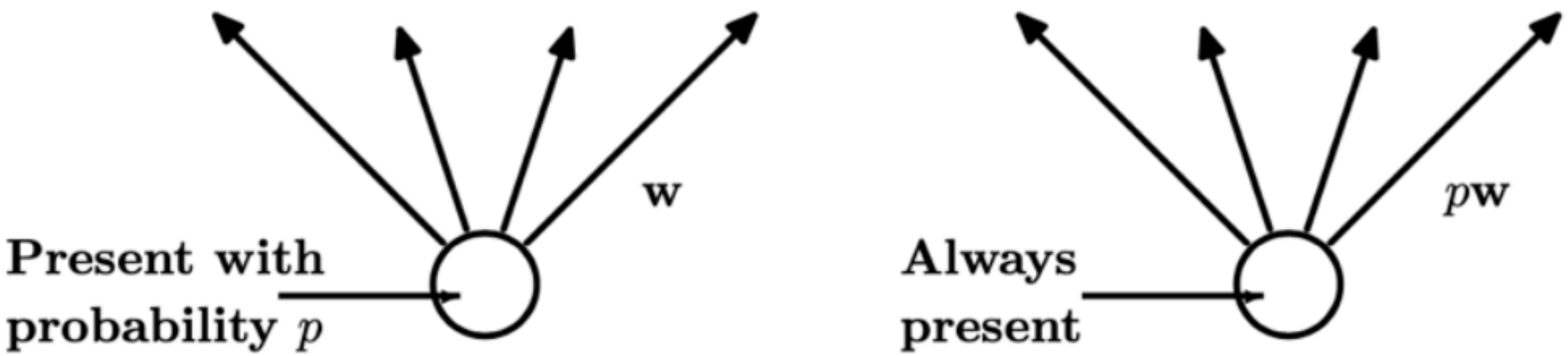
arxiv 1506.02142

**MC Dropout uncertainty quantification (DUQ) method**

- Dropout
  - A standard technique for training neural networks
  - Avoids over-fitting by randomly deactivating connections between nodes of neural network during the training process
  - All nodes exist during testing

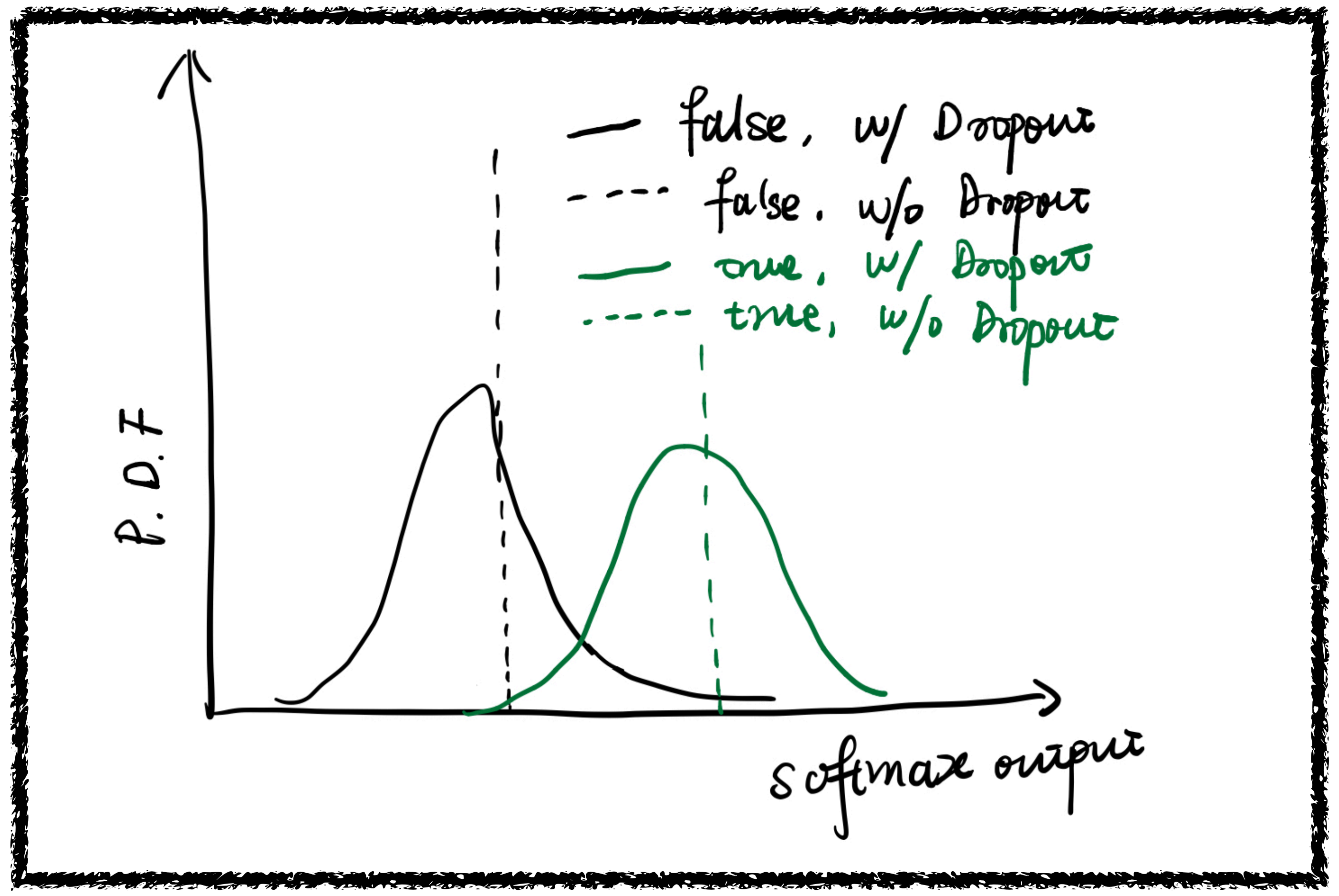


(a) Standard Neural Net (b) After applying dropout.




(a) At training time (b) At test time

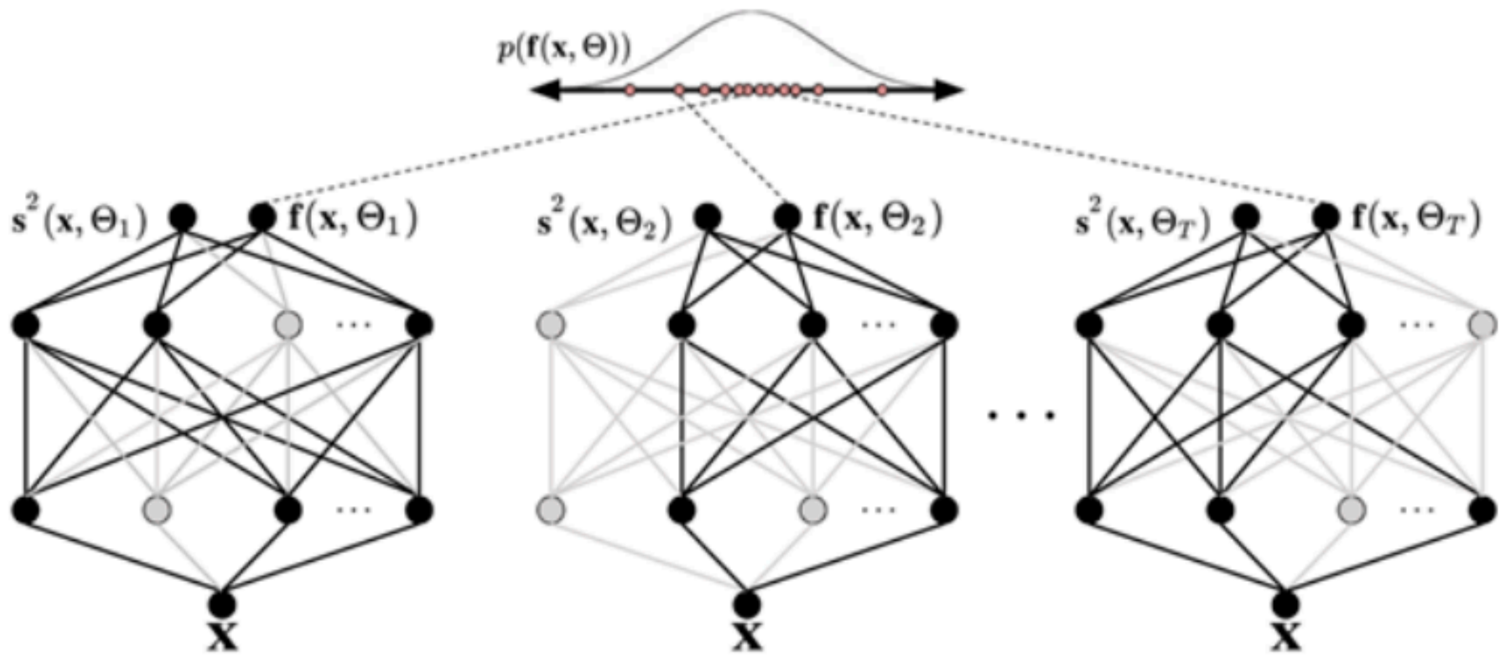
# Uncertainty quantification methods




**MC Dropout uncertainty quantification (DUQ) method**

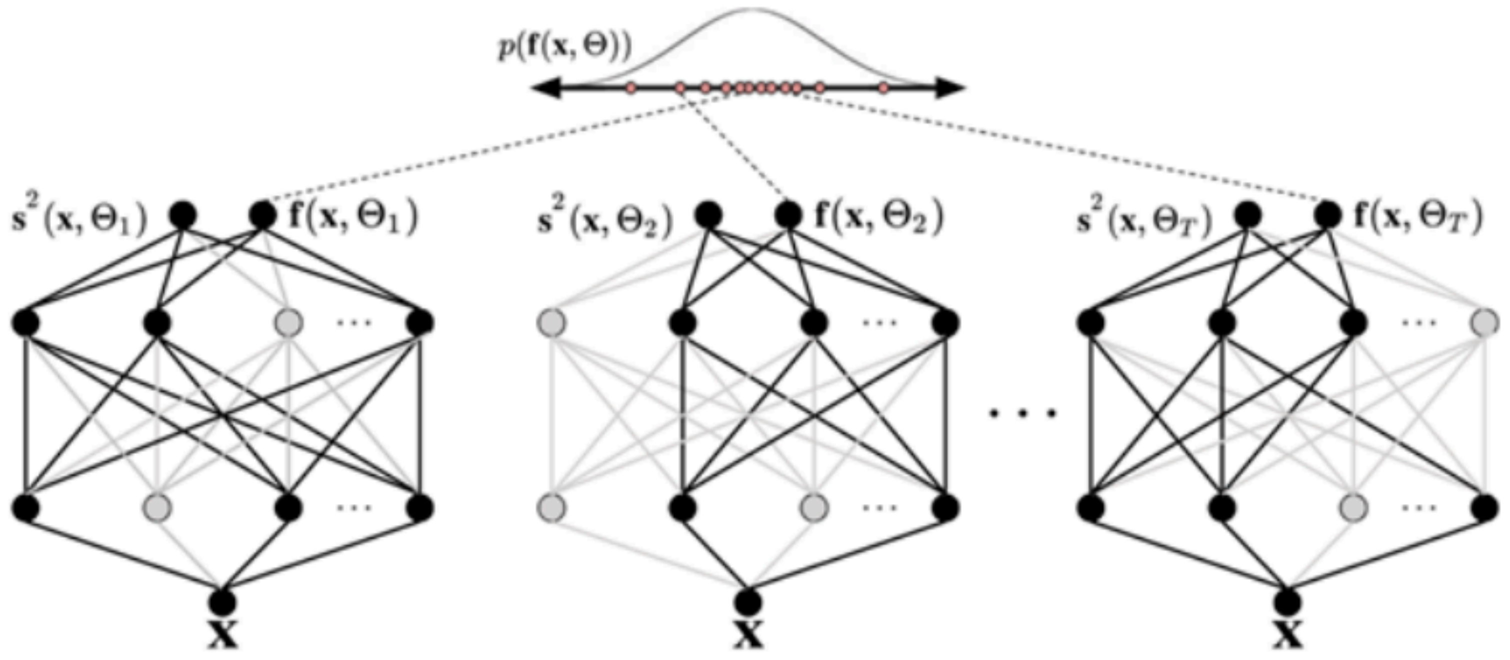
- **No change of either the training or the model**
- **No extra cost except to enable Dropout during testing**

Training steps




Uncertainty estimation

Apply dropout in prediction steps



# How do we measure the quality of uncertainty?

- **Multiple evaluations on each object 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:**

$$significance = \frac{\mu_{true} - \mu_{false}}{\sqrt{(\mu_{true} - CI_j)^2 + (\mu_{false} - CI_j)^2}}$$

  - **Image's probability which correspond to a correct categorization is calculated using the cumulative probability distribution over the calculated significance**

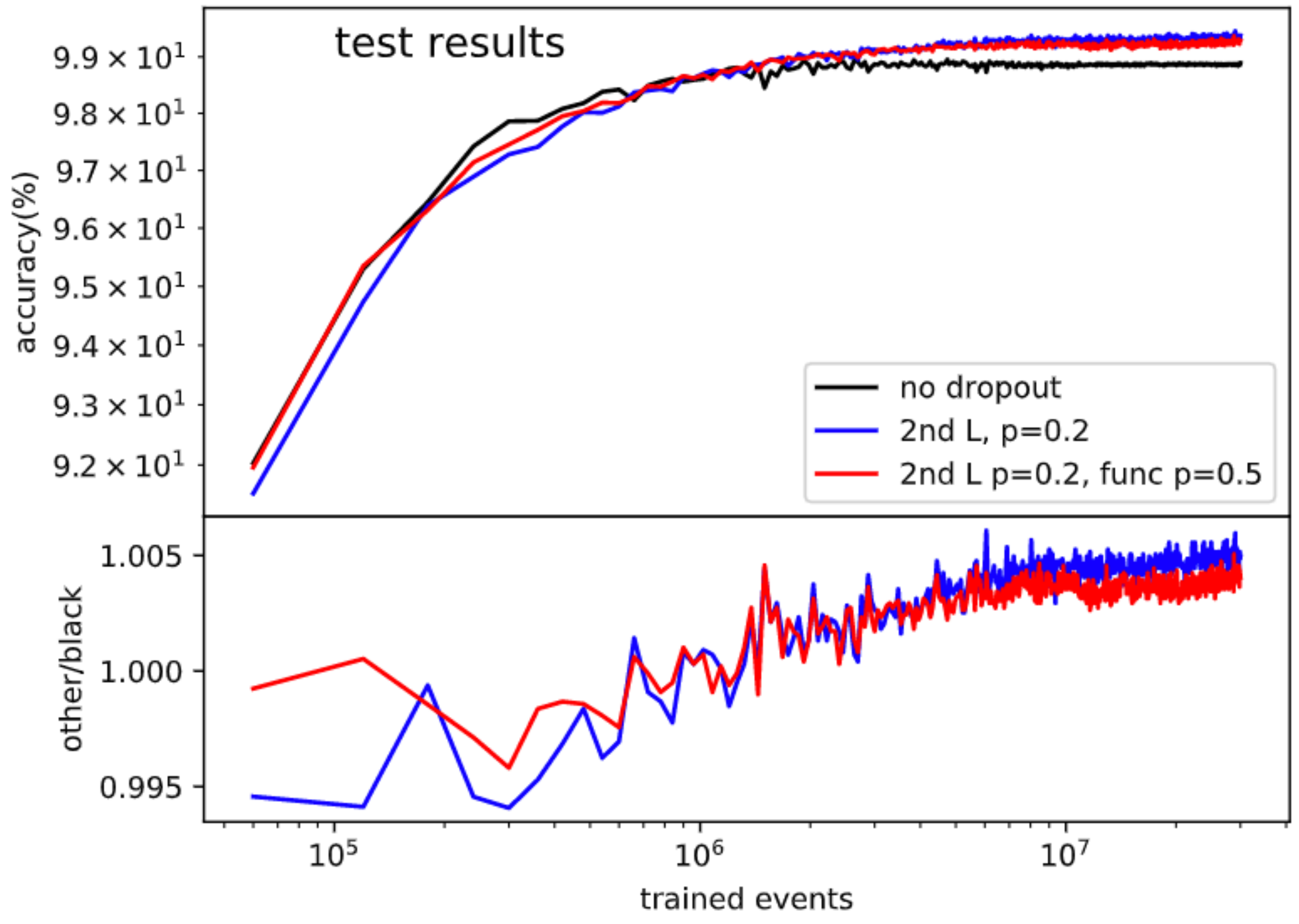
# The MNIST database

**The MNIST database:**

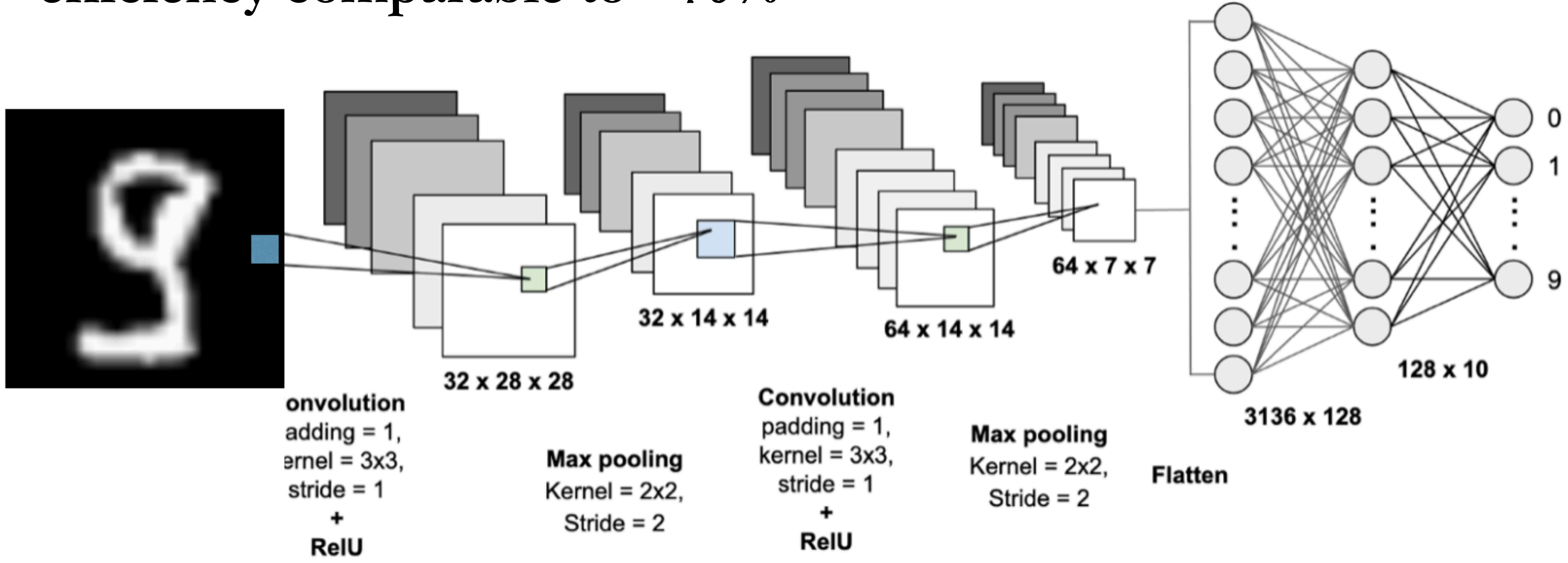
- **A database of handwritten, black and white digits from 0-9**
- **Has a training set of 60k images, and a testing set of 10k images**
- **All grey images are normalized to fit into a 28 x 28 pixel box**

# Training

- Trained the MNIST database for multi-classification studies
  - With a Convolutional Neural Network (CNN) which contains 2 hidden-layers with Dropout enabled

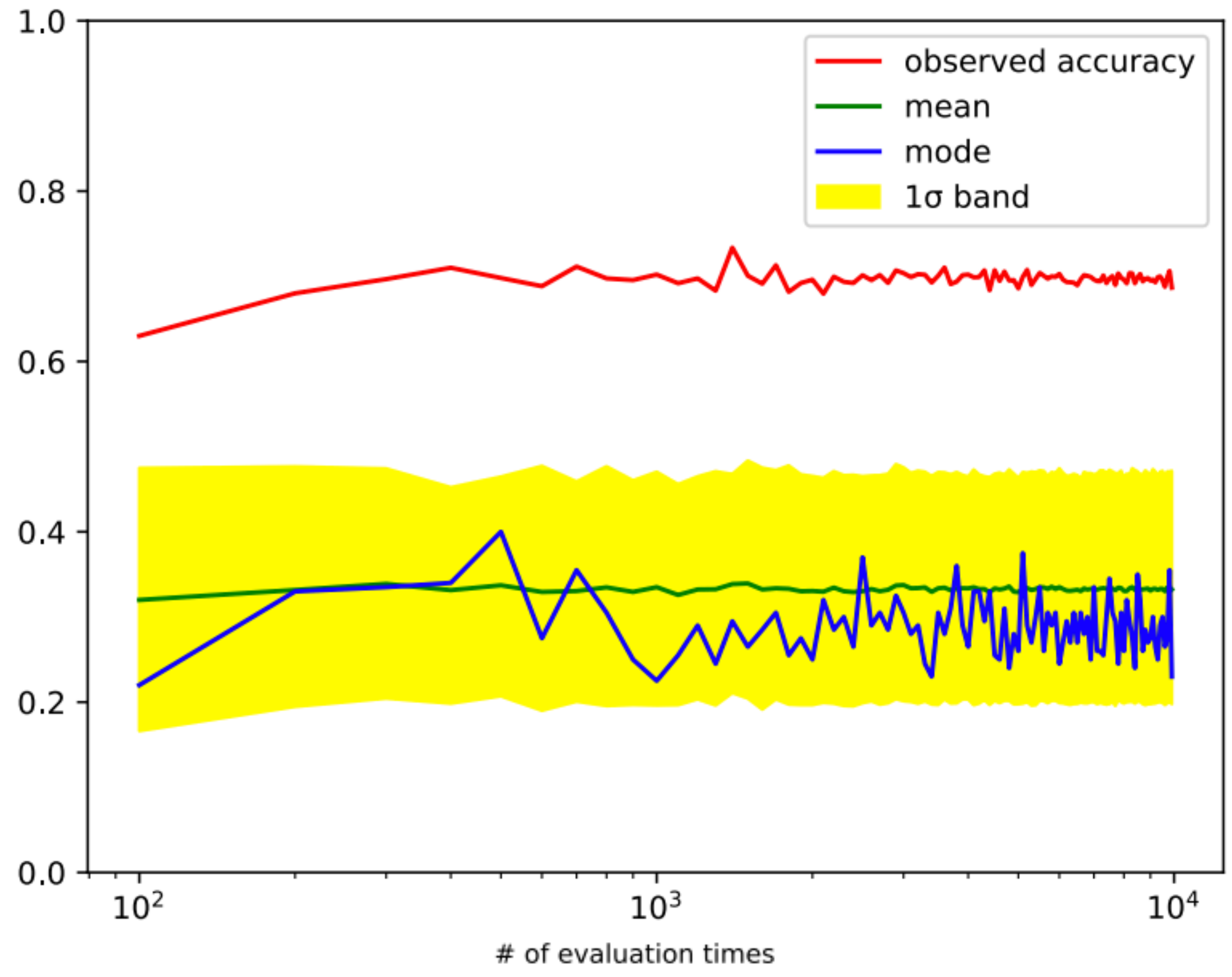


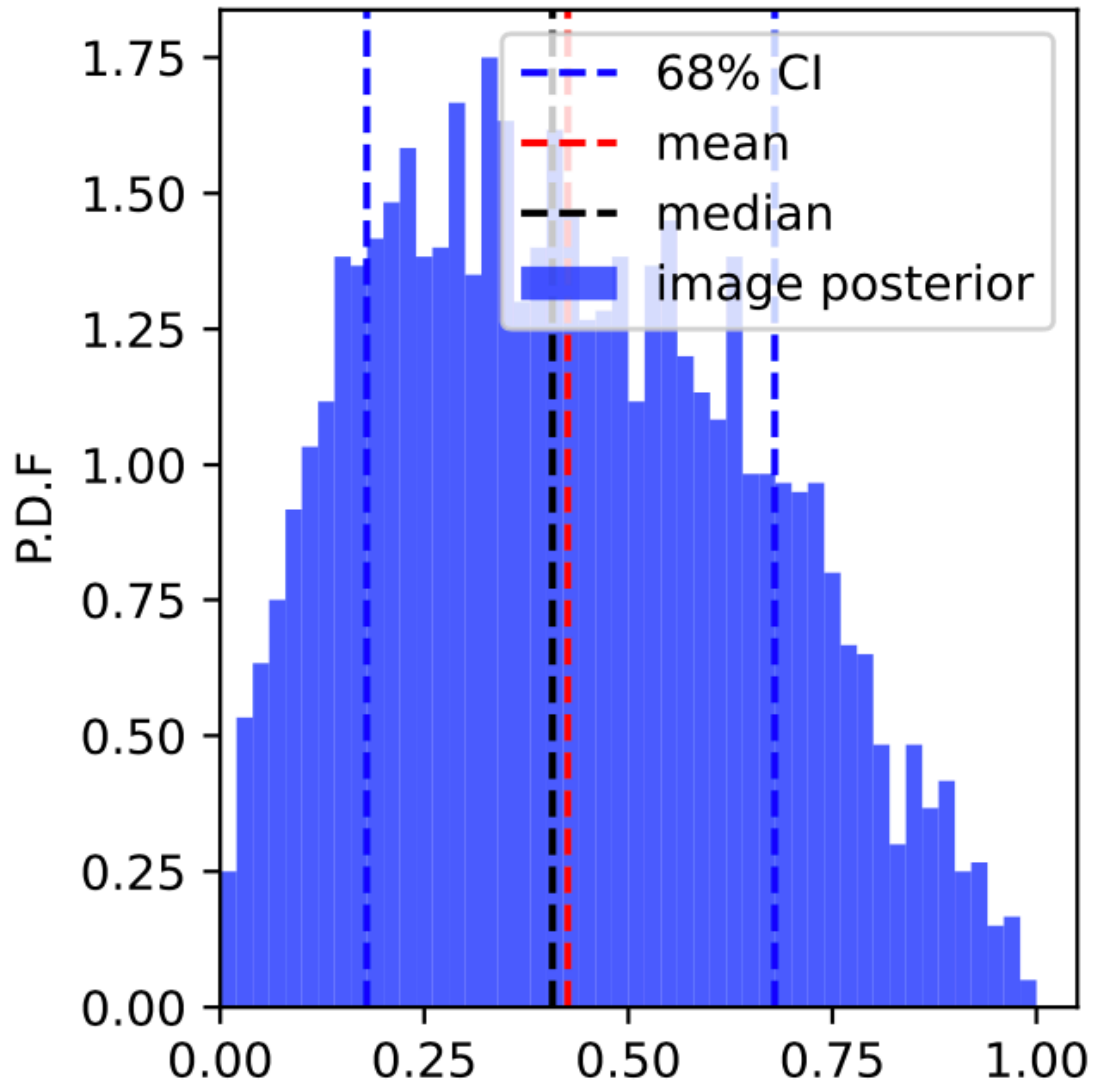

- Modern CNN can easily achieve > 99% accuracy
  - Great for postal mail sorting and bank check processing but not very interesting for uncertainty quantification studies
- Use simplified network structure and stopped training at efficiency comparable to ~70%



# 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 all the image accuracies, mean/median/mode values are stable
- Mode can be somewhat fluctuated for some images, while mean/median have similar value and reaches the stable point with 3k evaluations



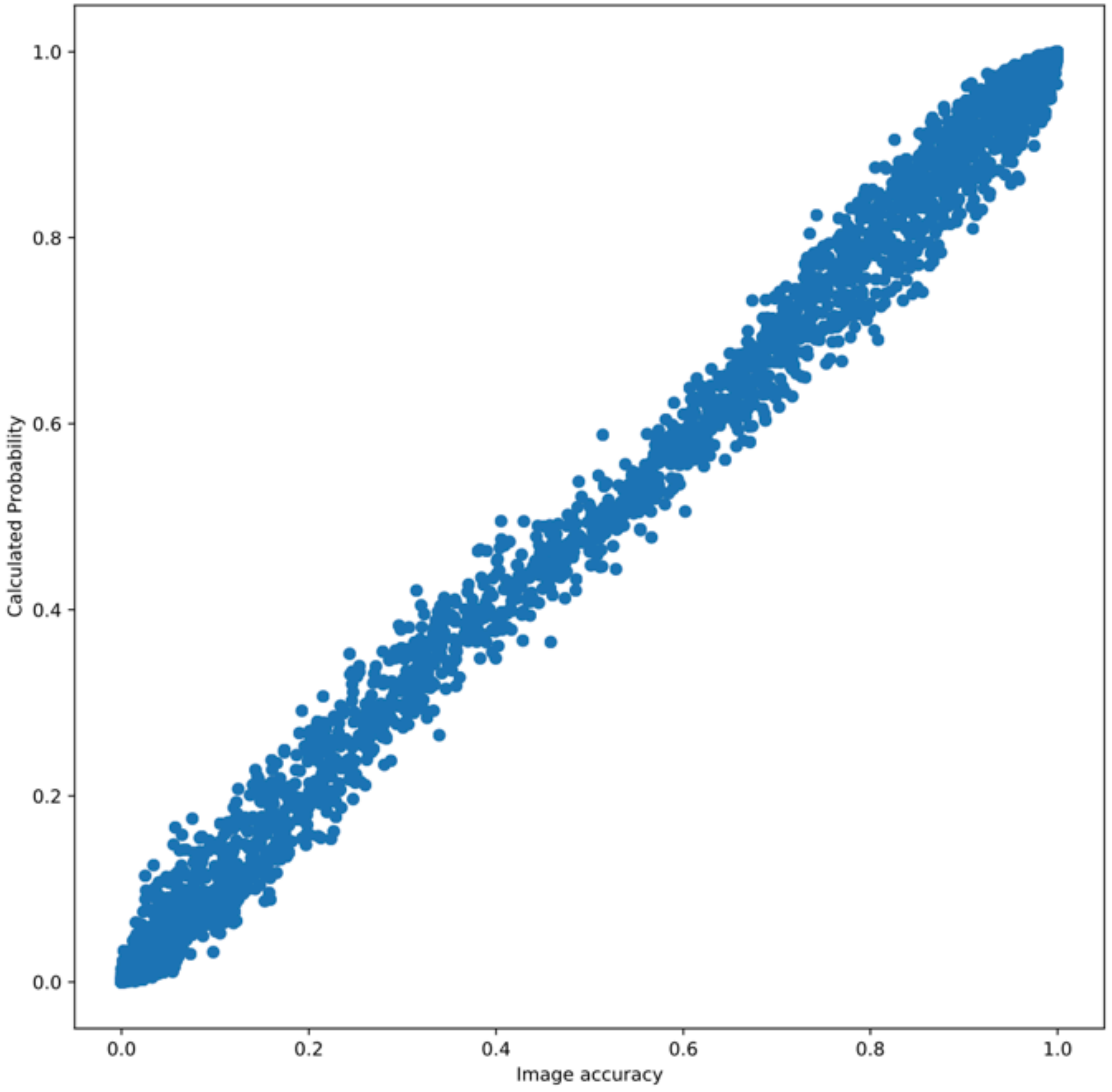



# Closure test

- For CNN with 2x2 convolution layers, the calculated probability accurately reflects how likely the predict if going to be correctly
- Across the full range of images, DUQ method captures the uncertainty well
- Small difference noticed at sample accuracy level
  - Observed 52.4% vs calculated 52.5%

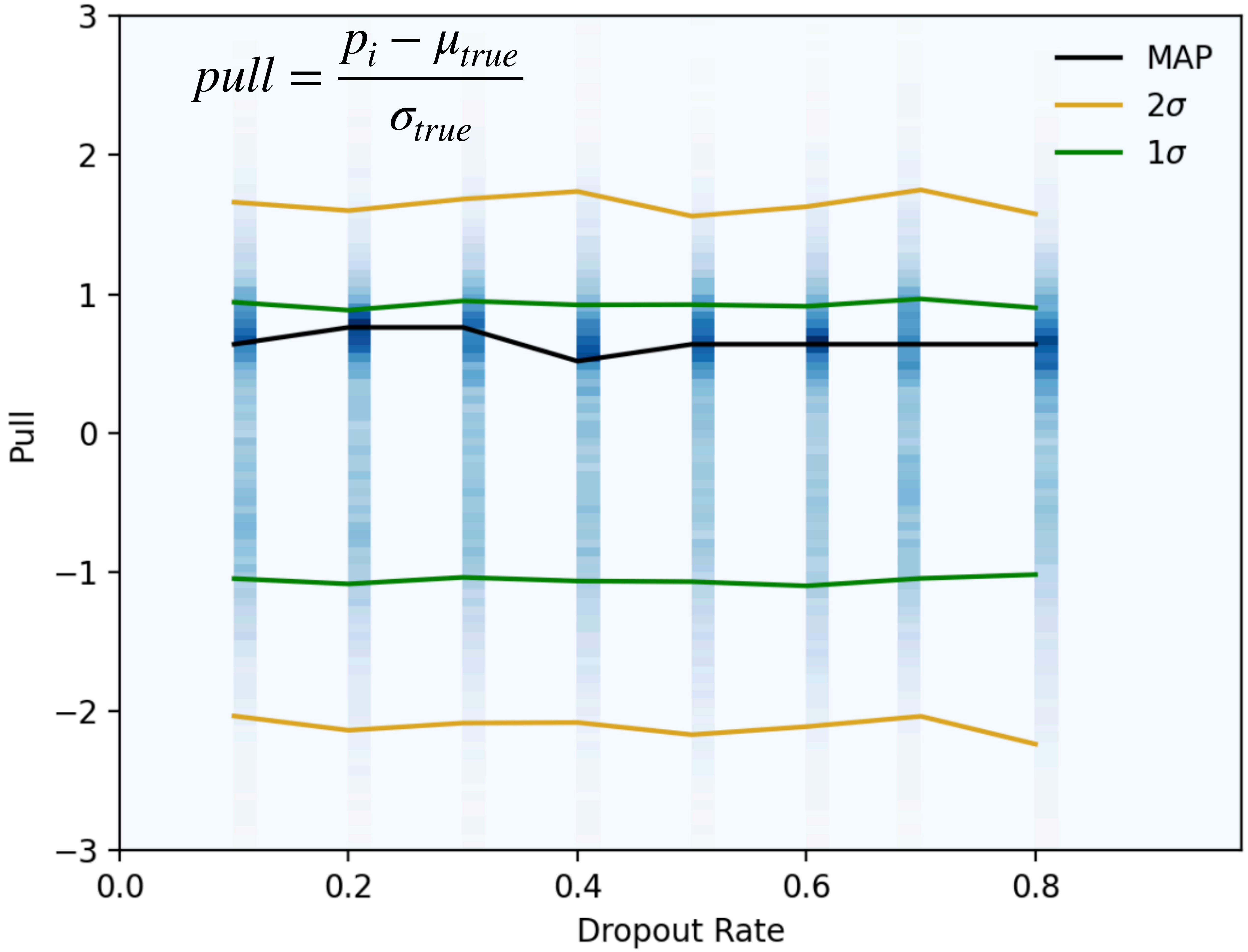
$$calculated\ probability = cdf(significance)$$

$$observed\ probability = \frac{\#\ correctly\ classified\ times}{\#\ of\ evaluations}$$



# Different dropout rates

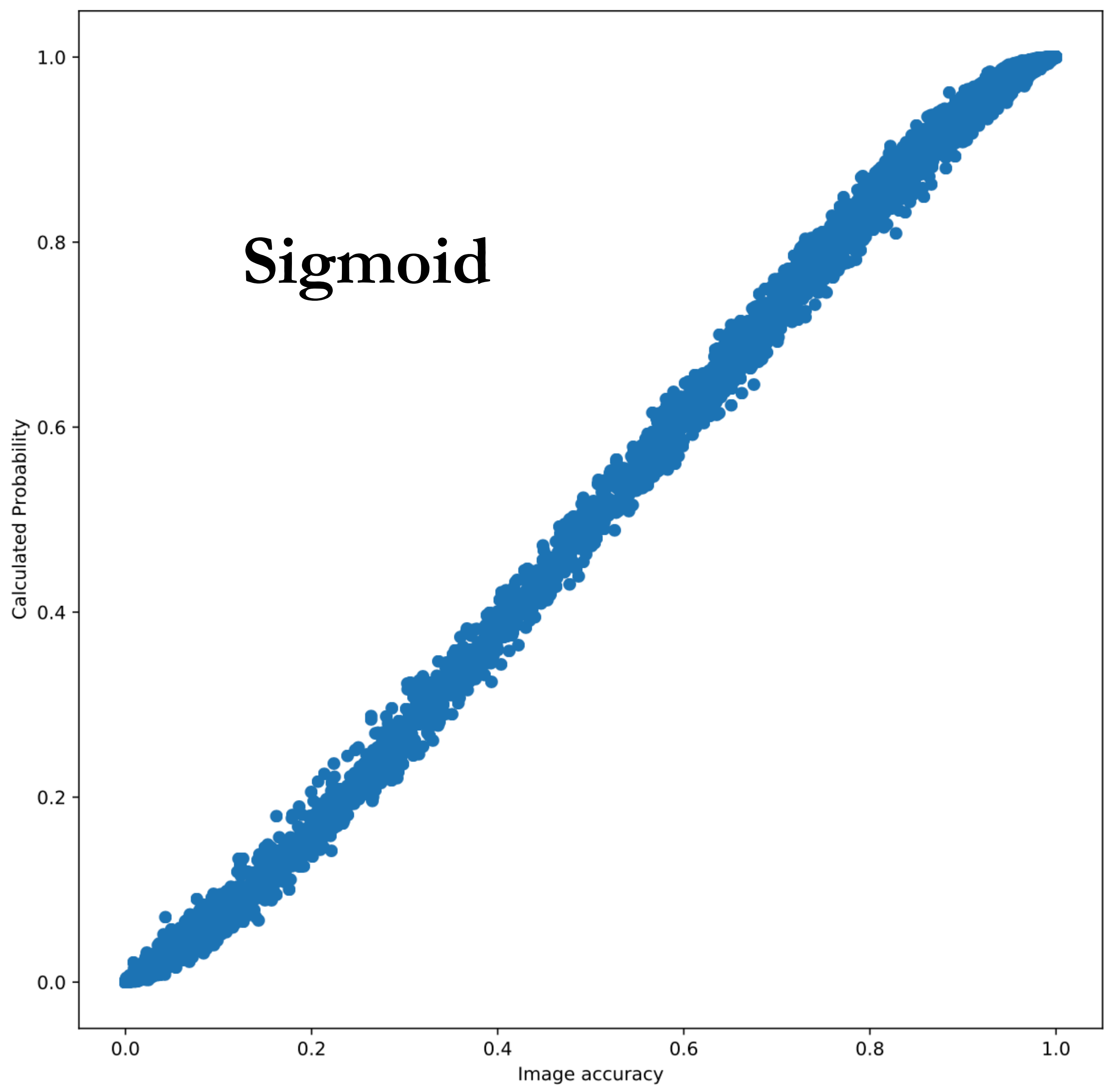
- How the posterior would differ with different dropout rates?
  - Varies dropout rate in the training
  - Set dropout rate in the testing same as in the training

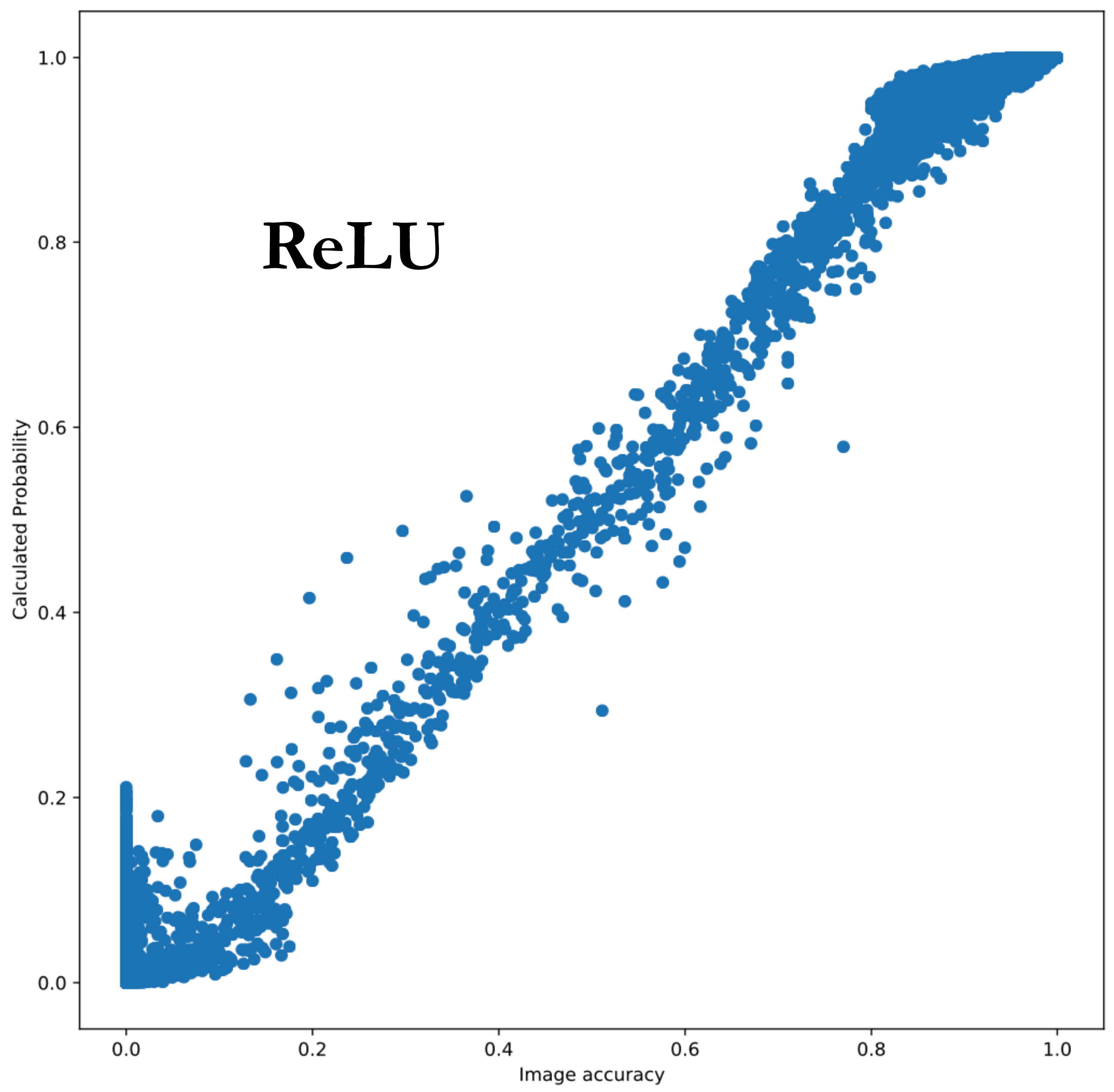


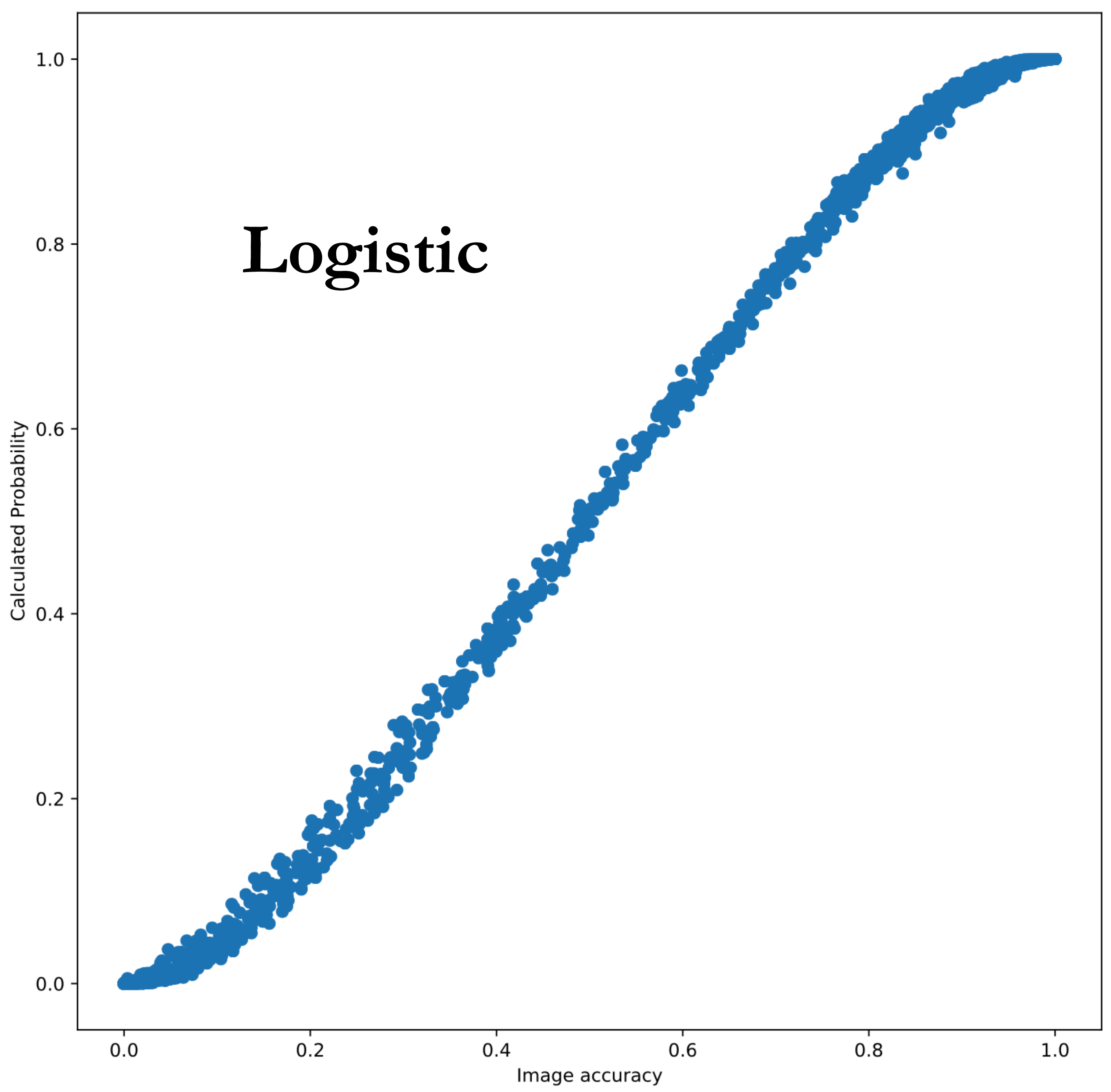

- Maximum a posteriori (MAP) varies below p=0.5, stay constant above
- $1\sigma$ and $2\sigma$ bands varies

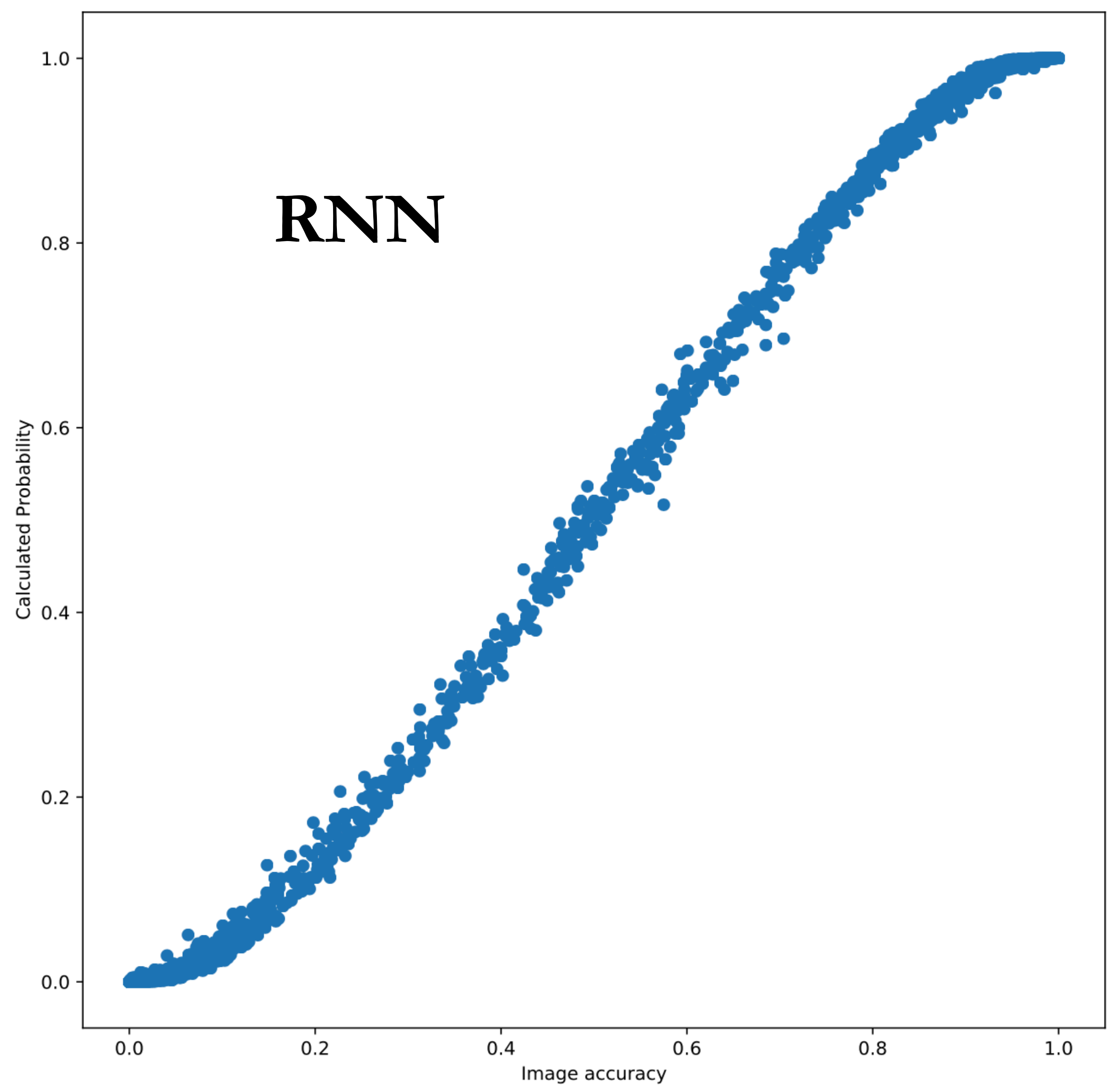
# Model dependency

- Tested the DUQ method on different NN models and activations
  - Some model dependency is observed
  - But worked at some level with all the models we tested



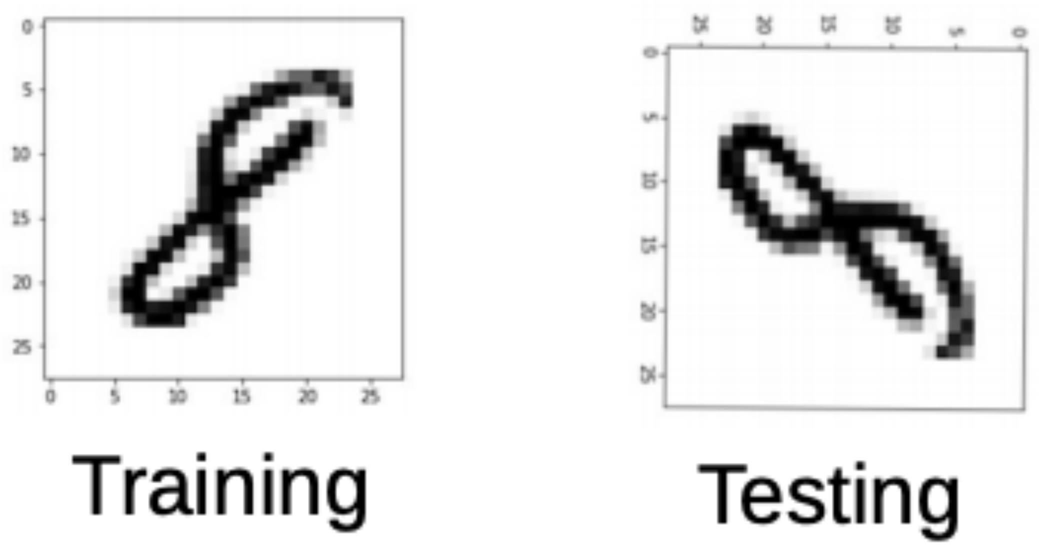



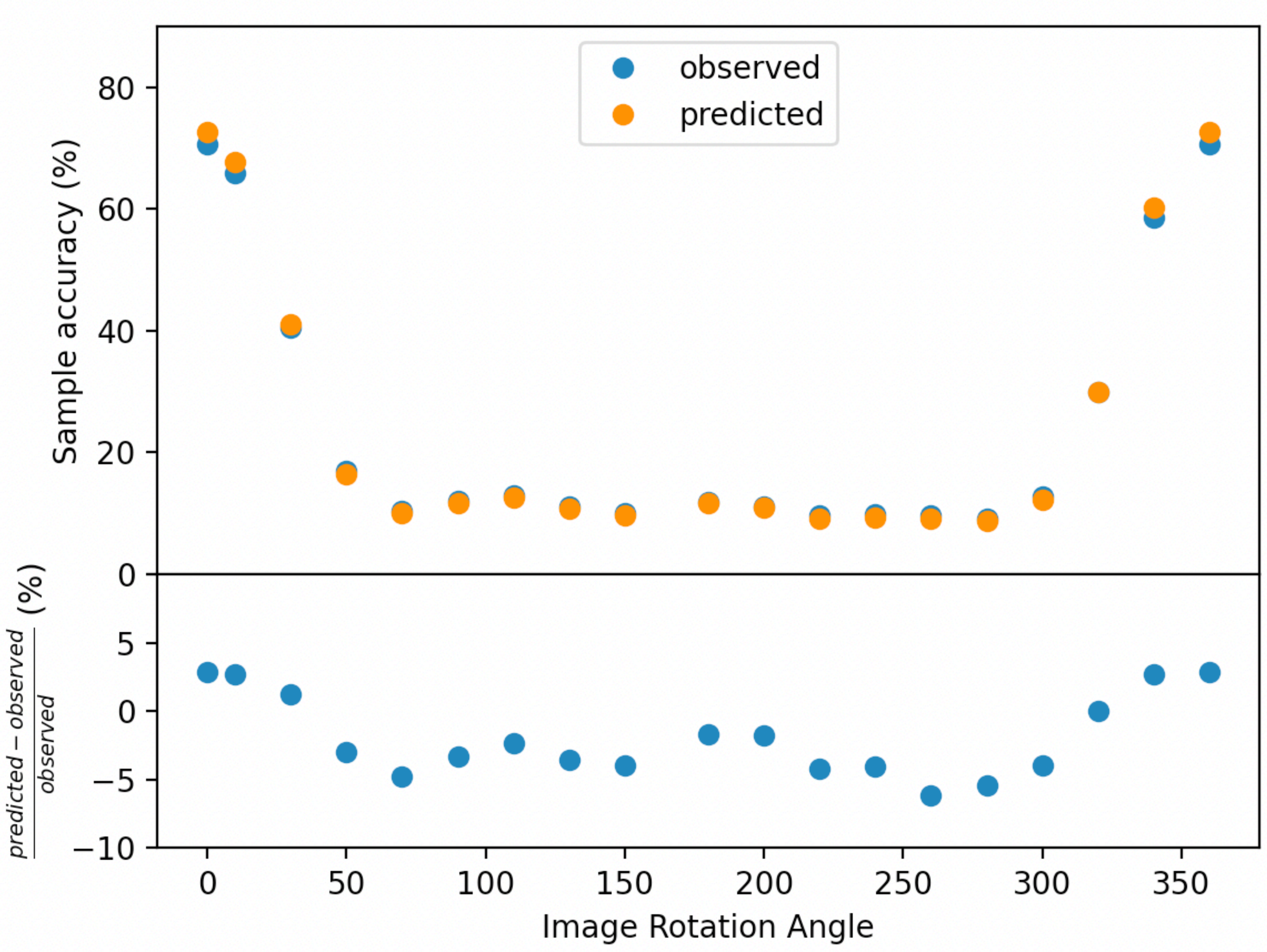



# Systematic mismodeling capture

- Trained a model on the nominal MNIST database
- Test performed by rotating images in the testing dataset by θ° ($\theta \in (0,360)$)

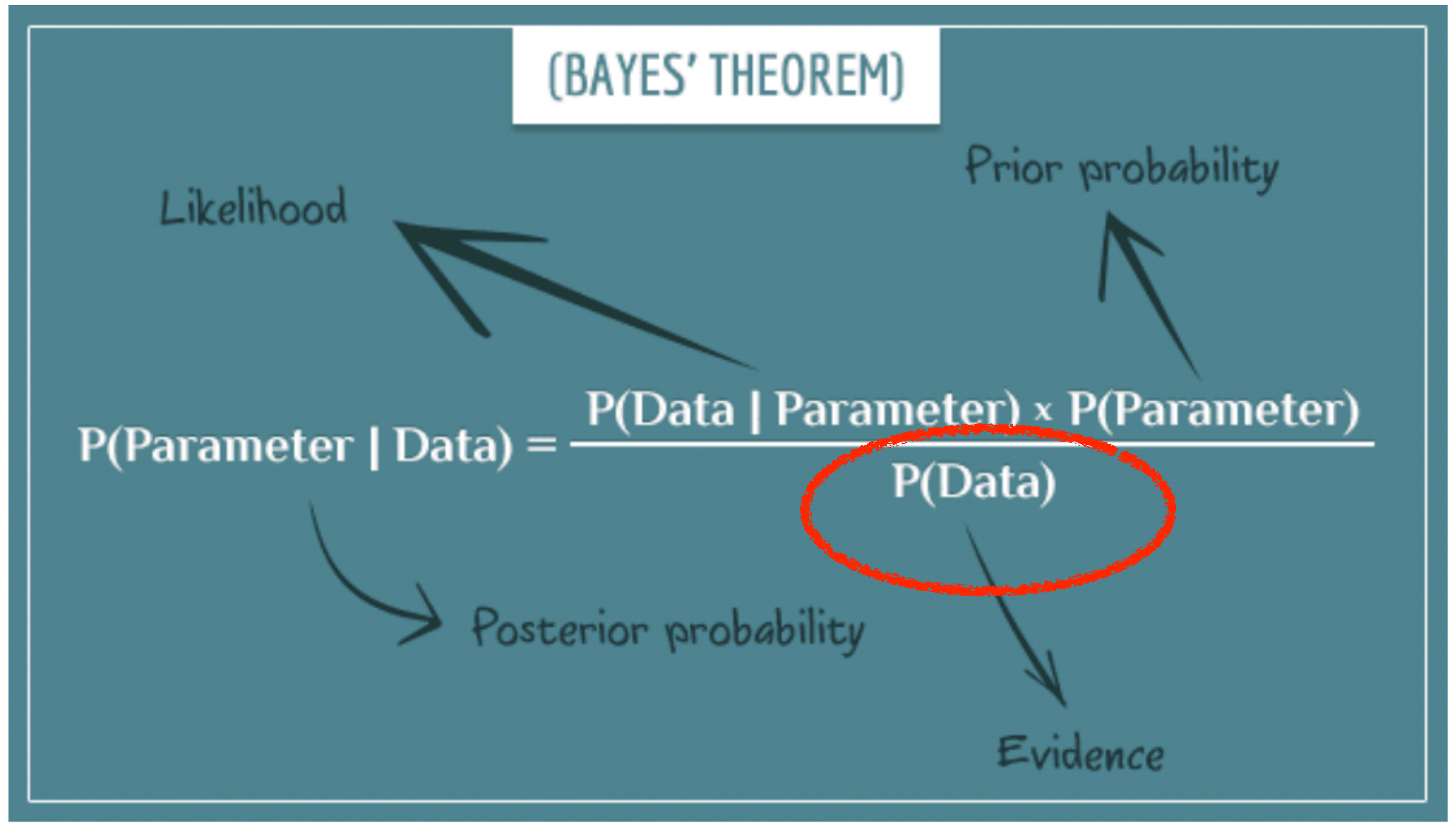



- Sample accuracy drops because of mismodeling
- But even with systematic mismodeling causing larger than 60% shift in sample accuracy, DUQ method still predict sample accuracy closes to its observed value.

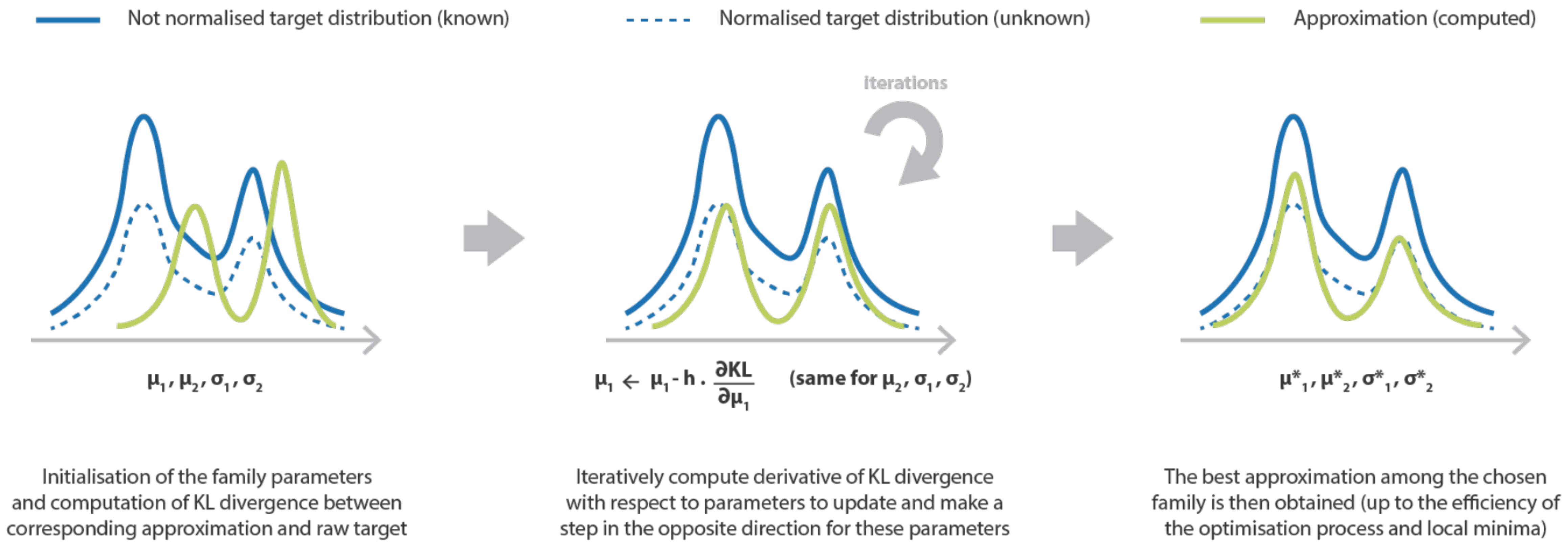
# BNN vs DUQ

- No statistical foundation for why DUQ should work
  - In order to “validate” the method, a comparison performed between DUQ and BNN



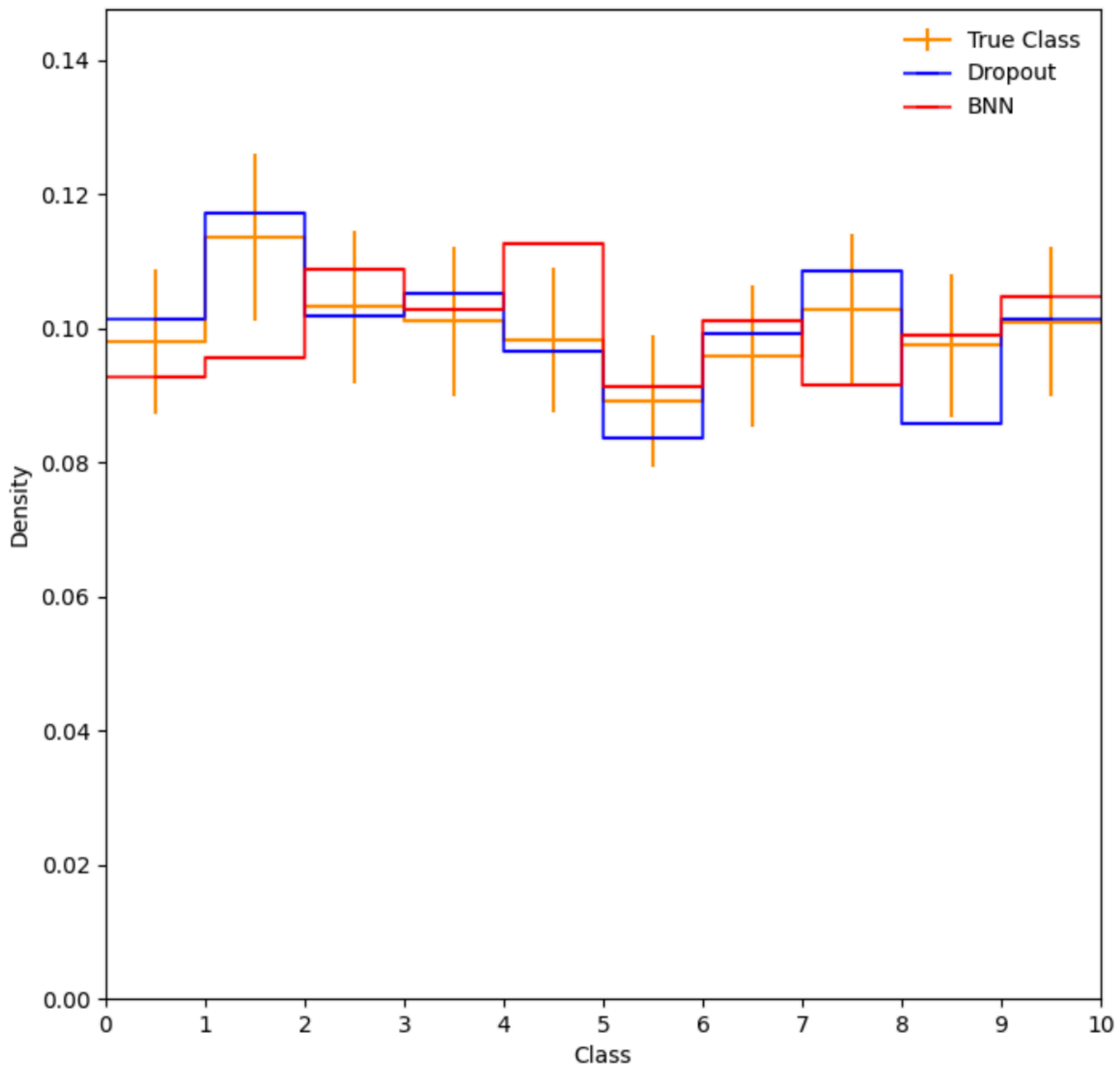

- BNN
  - $p(data)$ is computed via integrating over all possible parameter values: $$P(x) = \int_{\Theta} P(x, \theta) d\theta$$
  - Impossible in closed form for non-trivial problems, approximation needed - probabilistic programming

- Pyro used for BNN model training
  - Built on top of PyTorch
  - Scalable, flexible, universal
  - Has Stochastic Variational Inference



# BNN vs DQU

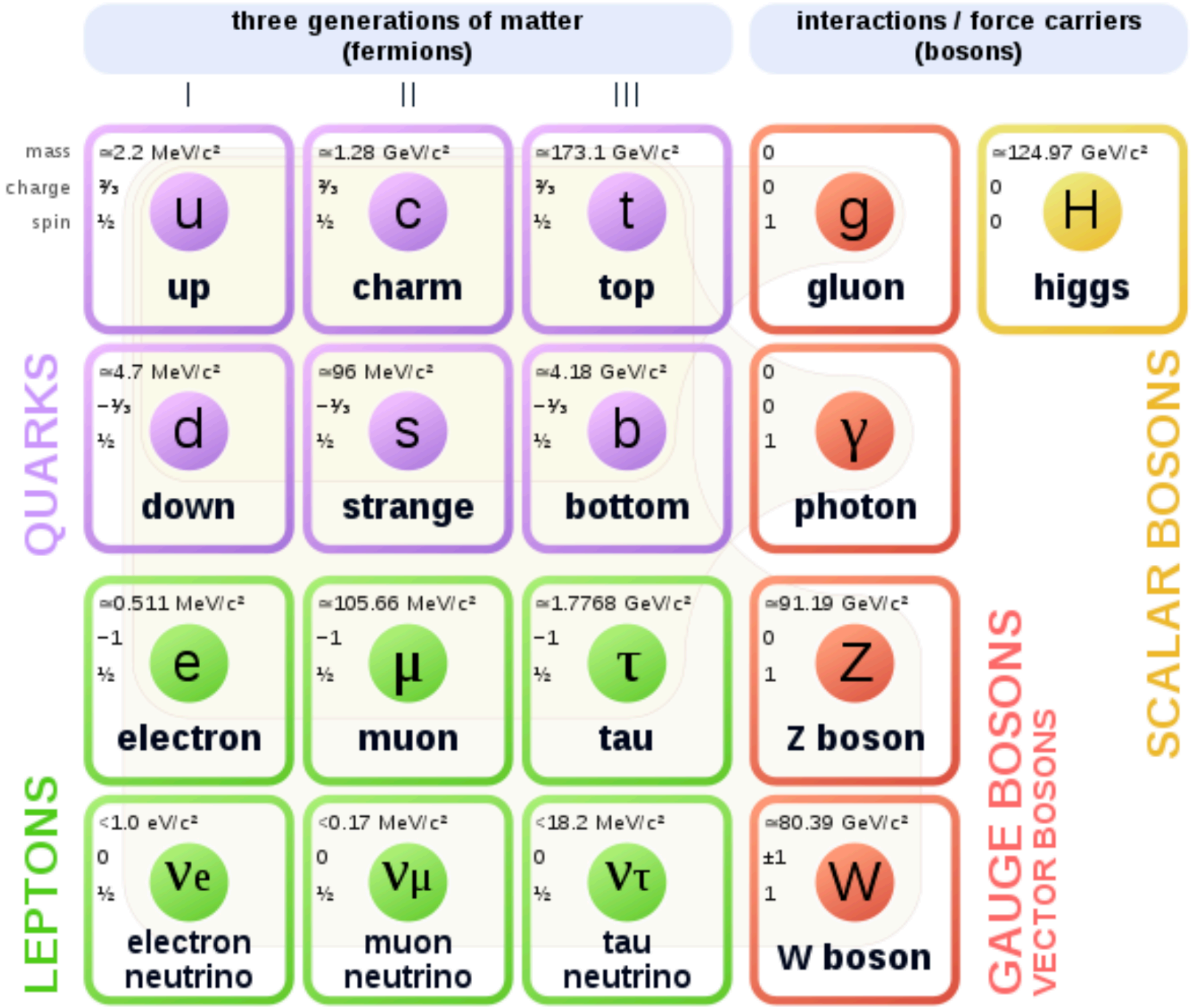
- Comparison done between BNN and DUQ
  - Models:
    - Same number of layers
    - Same number of nodes and dropout rate in each layer
    - Normal distribution is applied as prior on weight of each node in BNN model
    - Trained on same dataset with same epochs
  - Results
    - Poisson uncertainty added on truth class as error bar
    - DUQ prediction tend to have better agreement with the truth class



# The Standard Model

- The Standard Model (SM) of particle physics
  - A mathematical framework which describes the strong, weak and electromagnetic forces
  - Incorporates all directly observed elementary particles to date

- Limitation of the SM
  - Dark Matter (DM)
  - Matter-antimatter asymmetry
  - …

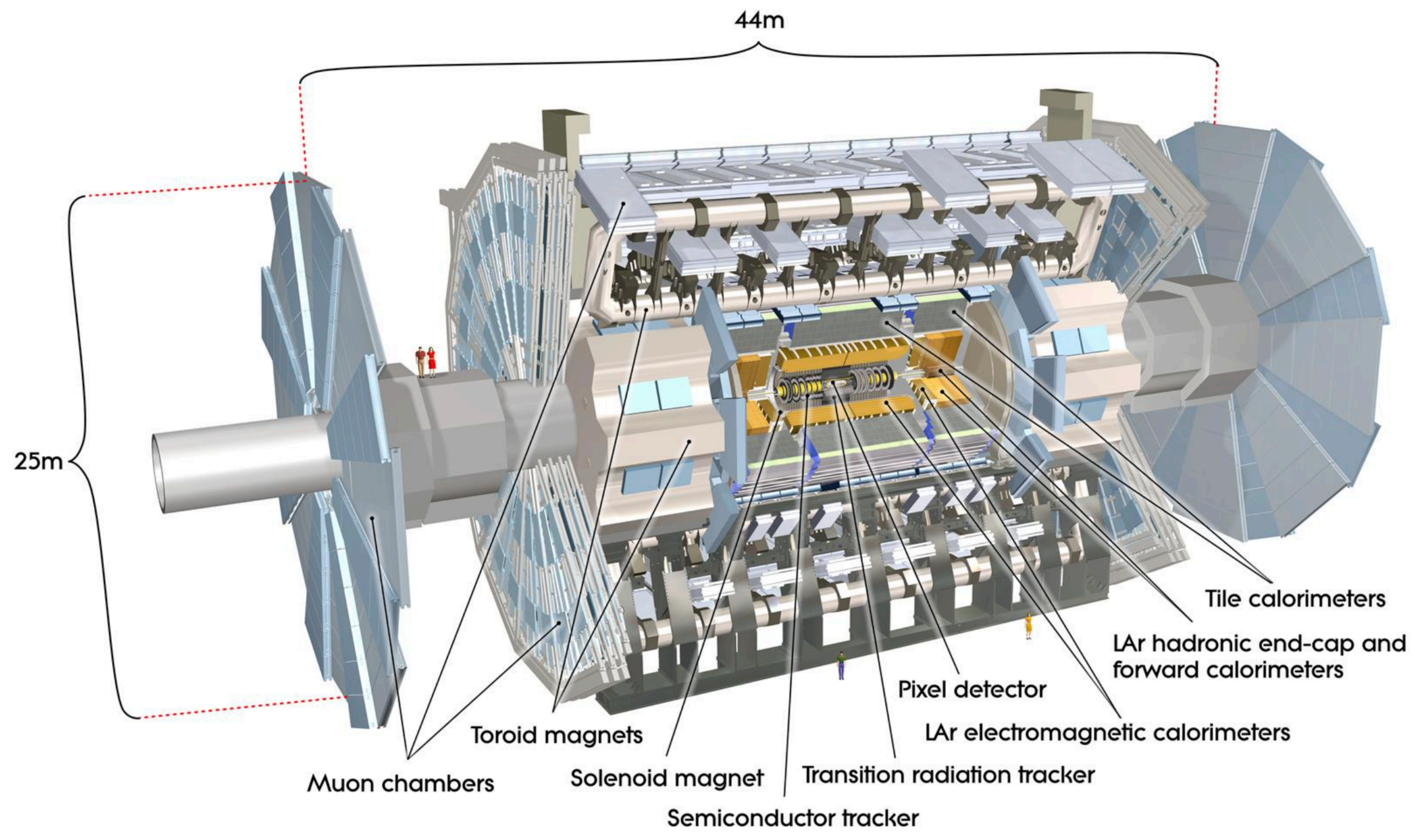


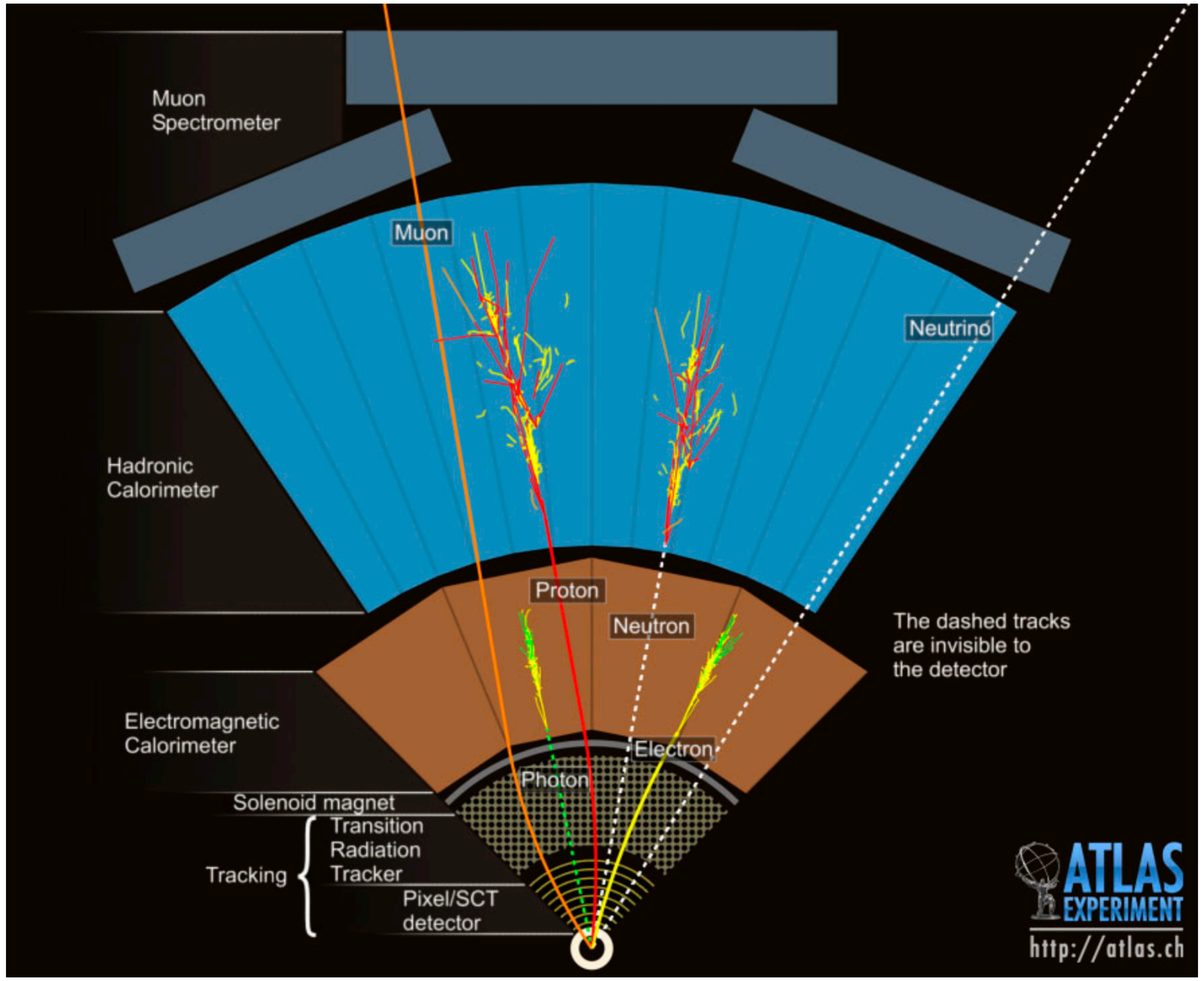
# The Large Hadron Collider

- Lies in a tunnel 27 kilometers in circumference, 175 meters beneath the France-Switzerland border
- Protons accelerated to 0.999999990 the speed of light
- Two opposing particle beams of protons at up to 6.5 tera electron volts (TeV) per nucleon, with center-of-mass energy at 13 TeV collision energy were smashed in LHC machine
- Collide at 4 primary points where detectors are situated

# The ATLAS detector

- A toroidal LHC Apparatus (ATLAS) is one of two general purpose detectors at LHC
- Aims to measure signals resulting from pp collision to cover vast range of analyses



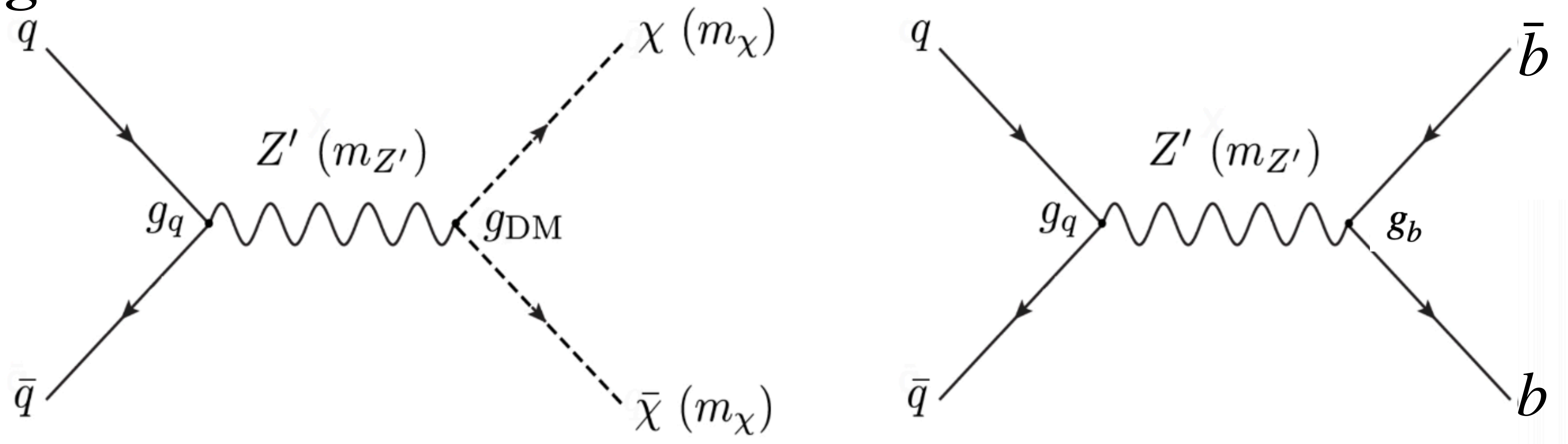




- ATLAS is a many-layered detector
  - Inner detector: describes charged particle trajectory through the detector and magnetic filed
  - Electromagnetic calorimeter: electromagnetic signatures (photons, electrons)
  - Hadronic calorimeter: particles that interact via the strong force (quarks, gluons)
  - Muon detector: dedicated subsystem for detecting muons

# B-tagging in ATLAS

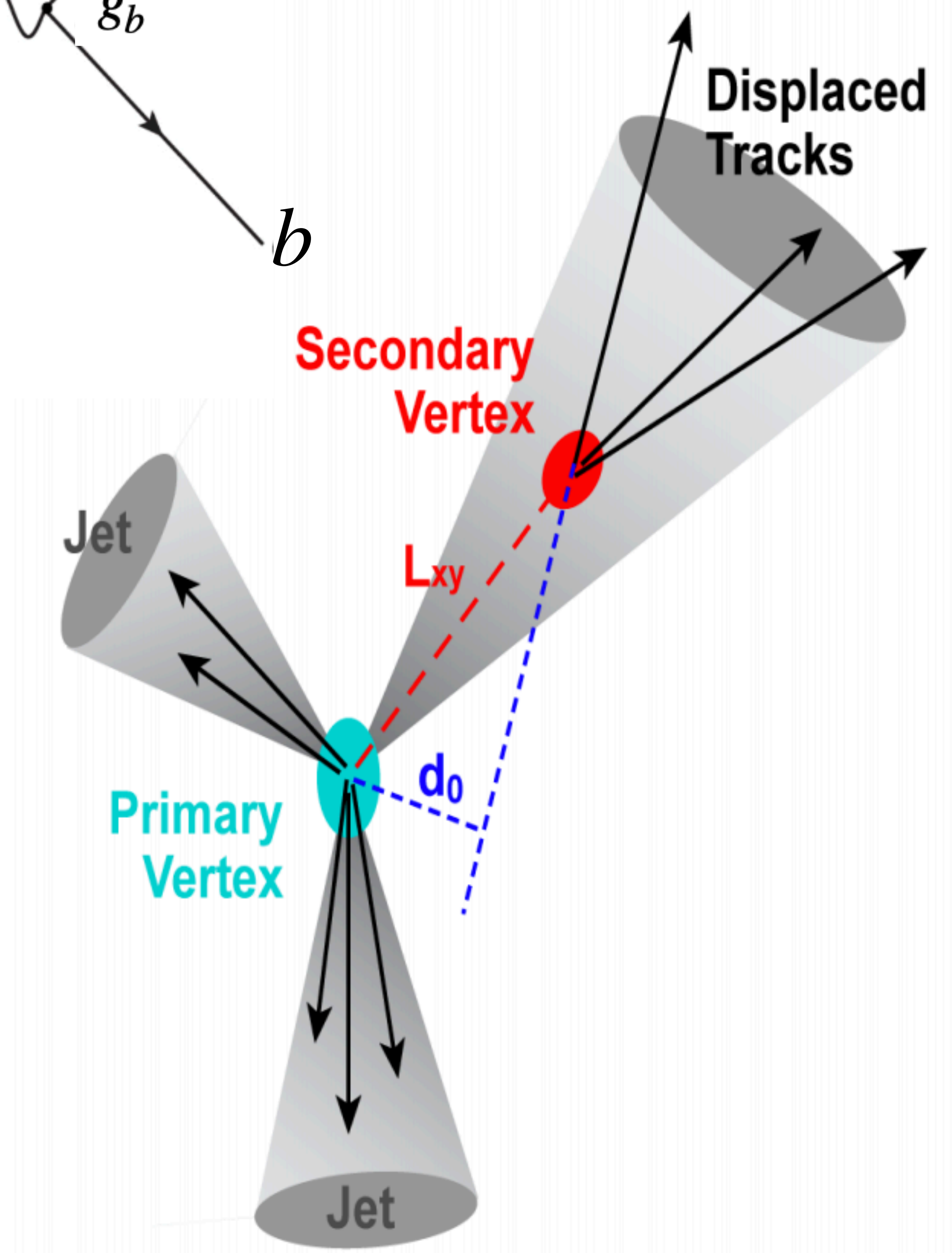
The identification of jets containing B-hadrons (b-tagging) is essential for many physics in ATLAS

- For example: searching for dark matter



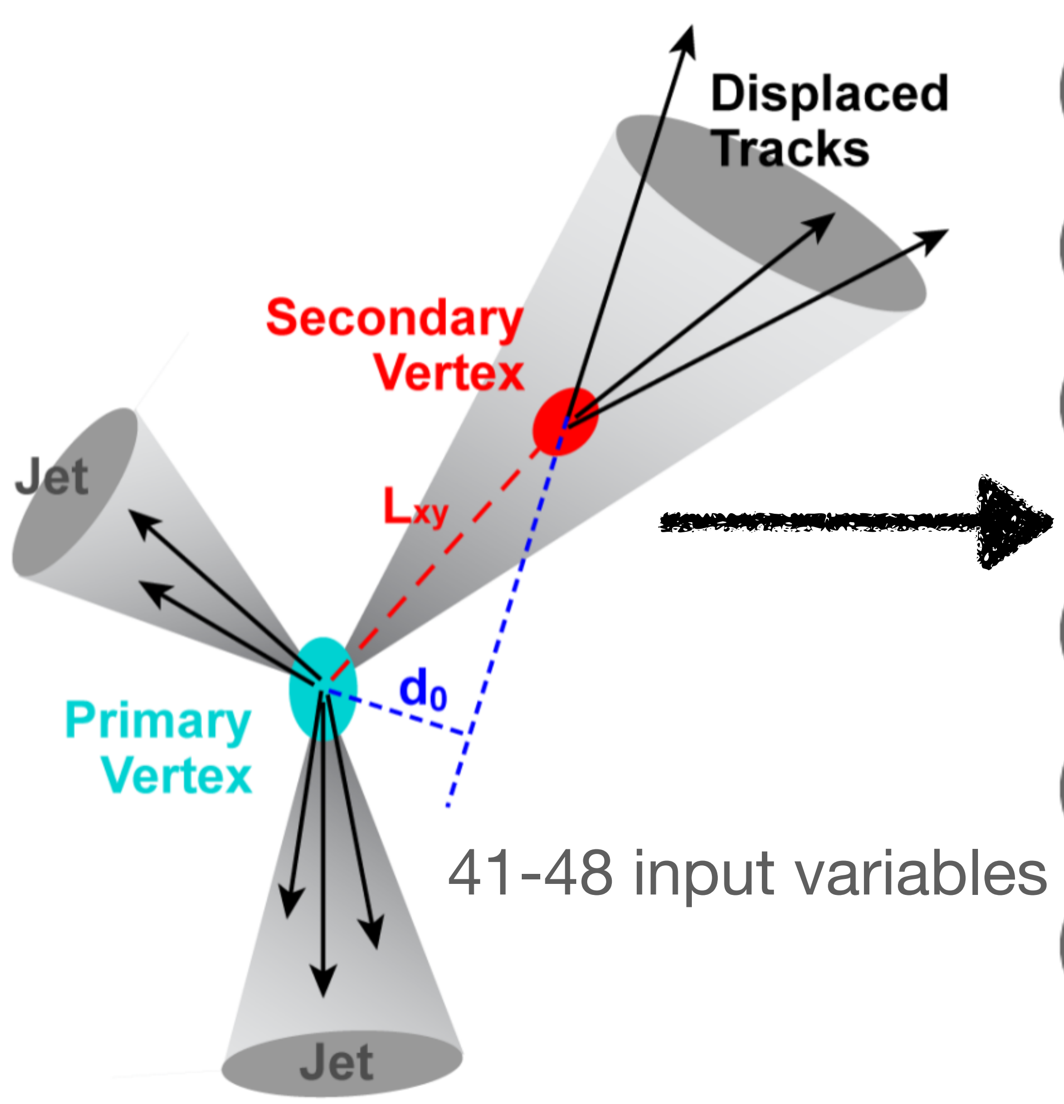

B-tagging rely on B-hadron properties:

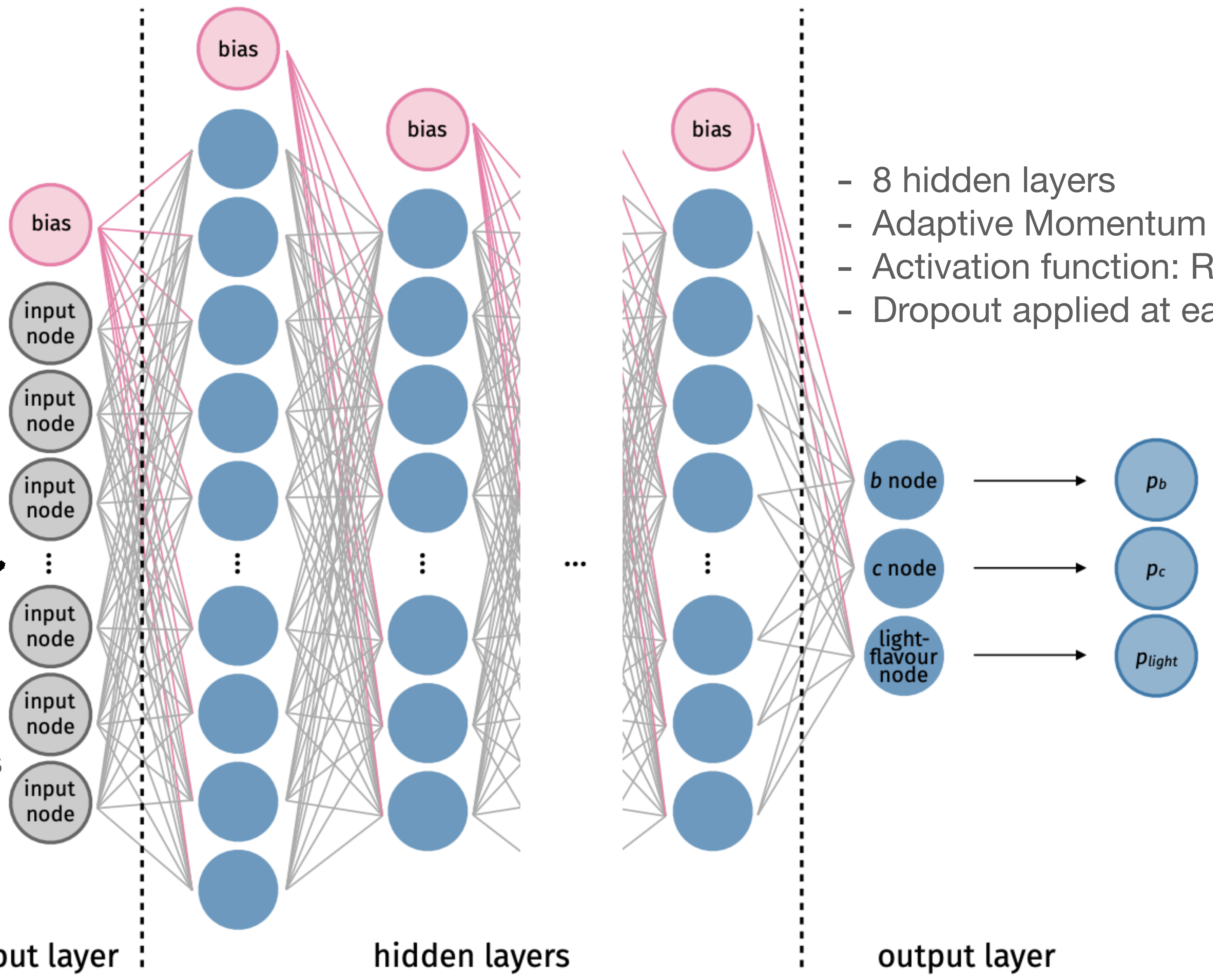
- Secondary vertex from primary vertex due to its long life time
- Large B-hadron mass
- Large impact parameter
- Semi-leptonic decay of B-hadron



# B-tagging algorithm

- Deep learning technique is used for b-tagging




- 8 hidden layers
- Adaptive Momentum (Adam) optimiser to minimise categorical cross-entropy loss
- Activation function: ReLU and Softmax (only for output layer)
- Dropout applied at each layer

$$\mathrm{DL1}_{\text{b-score}}(f_c) = \ln\left(\frac{p_b}{f_c \cdot p_c + (1 - f_c) \cdot p_{\text{light-flavour}}}\right)$$

# B-tagging uncertainties

- B-tagging calibrations obtained in the forms of data-to-simulation scale factors (SF)
- Uncertainties from data are added to the SFs

$$\mathrm{SF}_{b,w} = \frac{\epsilon^{\mathrm{data}}_{b,w}}{\epsilon^{\mathrm{MC}}_{b,w}}$$

  - At pT > 400 GeV
    - Not enough statistics in data
    - SF in this region defined as:

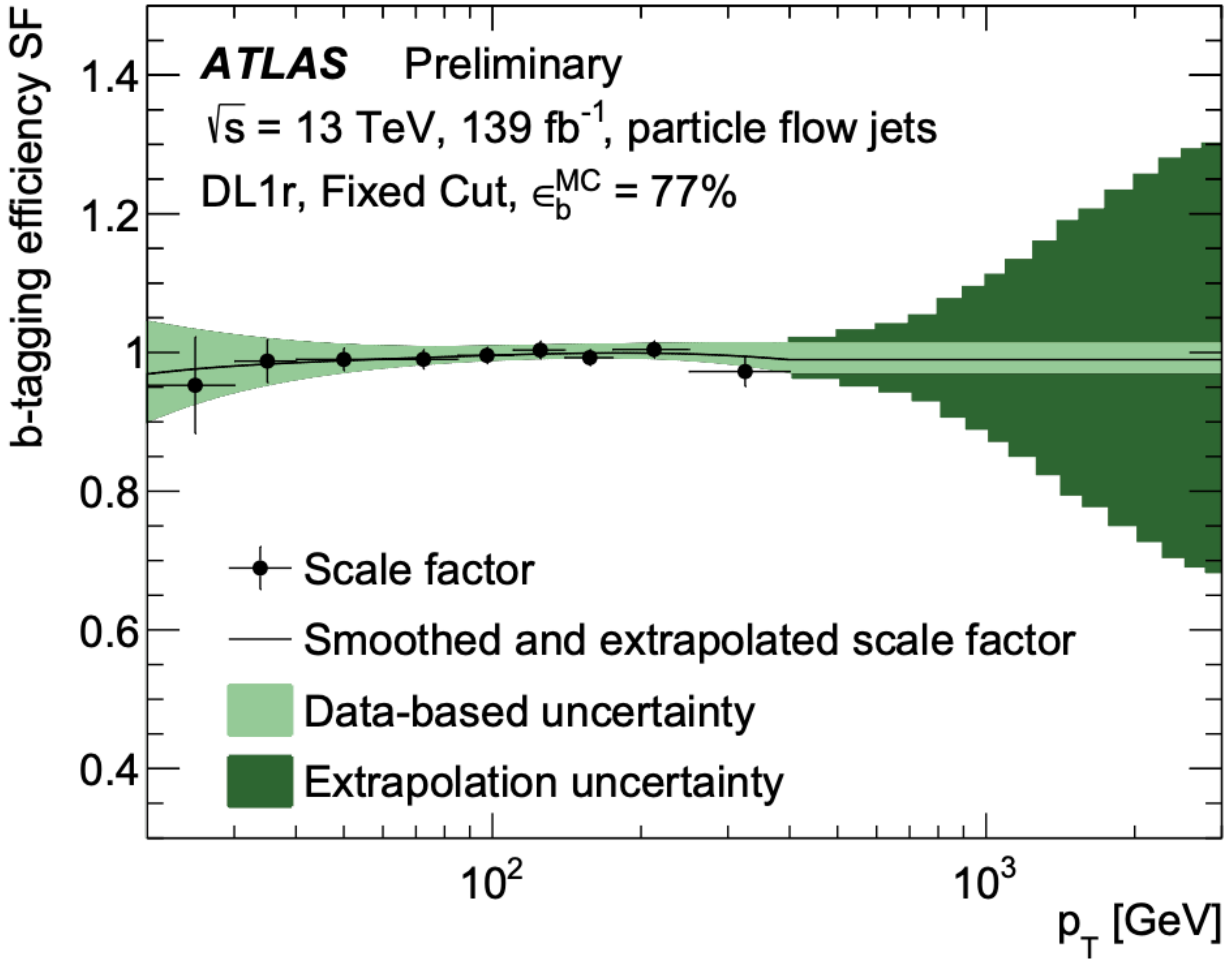
$$\mathrm{SF}_b(p_T) := \mathrm{SF}_b(p_{T,\mathrm{ref}}) \cdot \mathcal{R}^{\mathrm{MC}}_b(p_T; p_{T,\mathrm{ref}})$$

    - Uncertainties:

$$\begin{aligned}\sigma^2_{\mathrm{rel}}(\mathrm{SF}_b(p_T)) &= \sigma^2_{\mathrm{rel}}(\mathrm{SF}_b(p_{T,\mathrm{ref}})) + \sigma^2_{\mathrm{rel}}(\mathcal{R}^{\mathrm{MC}}_b(p_T; p_{T,\mathrm{ref}})) \\ &= \sigma^2_{\mathrm{rel}}(\mathrm{SF}_b(p_{T,\mathrm{ref}})) + \sigma^2_{\mathrm{extrap}}(p_T; p_{T,\mathrm{ref}})\end{aligned}$$

      - An additional extrapolation uncertainty determined by modifying DNN input variables is added - which explodes as pT increases

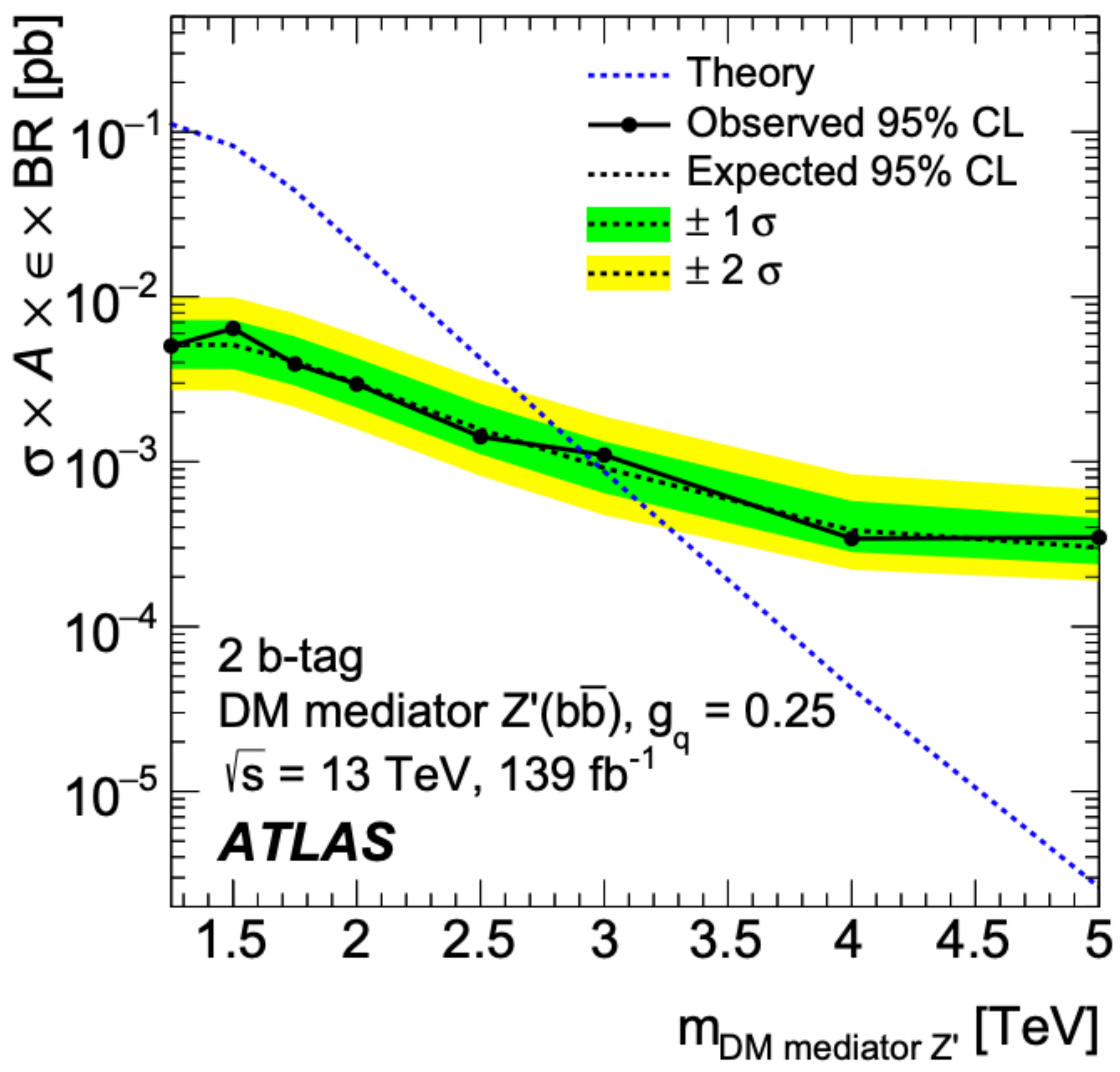
ATL-PHYS-PUB-2021-00



# B-tagging uncertainties

- **DUQ can be tried as a new approach, as the method**
  - **Can be potentially used to capture uncertainties in any classification case as long as Dropout is enabled in the training**
  - **Can capture uncertainties for each jet regardless of statistics**

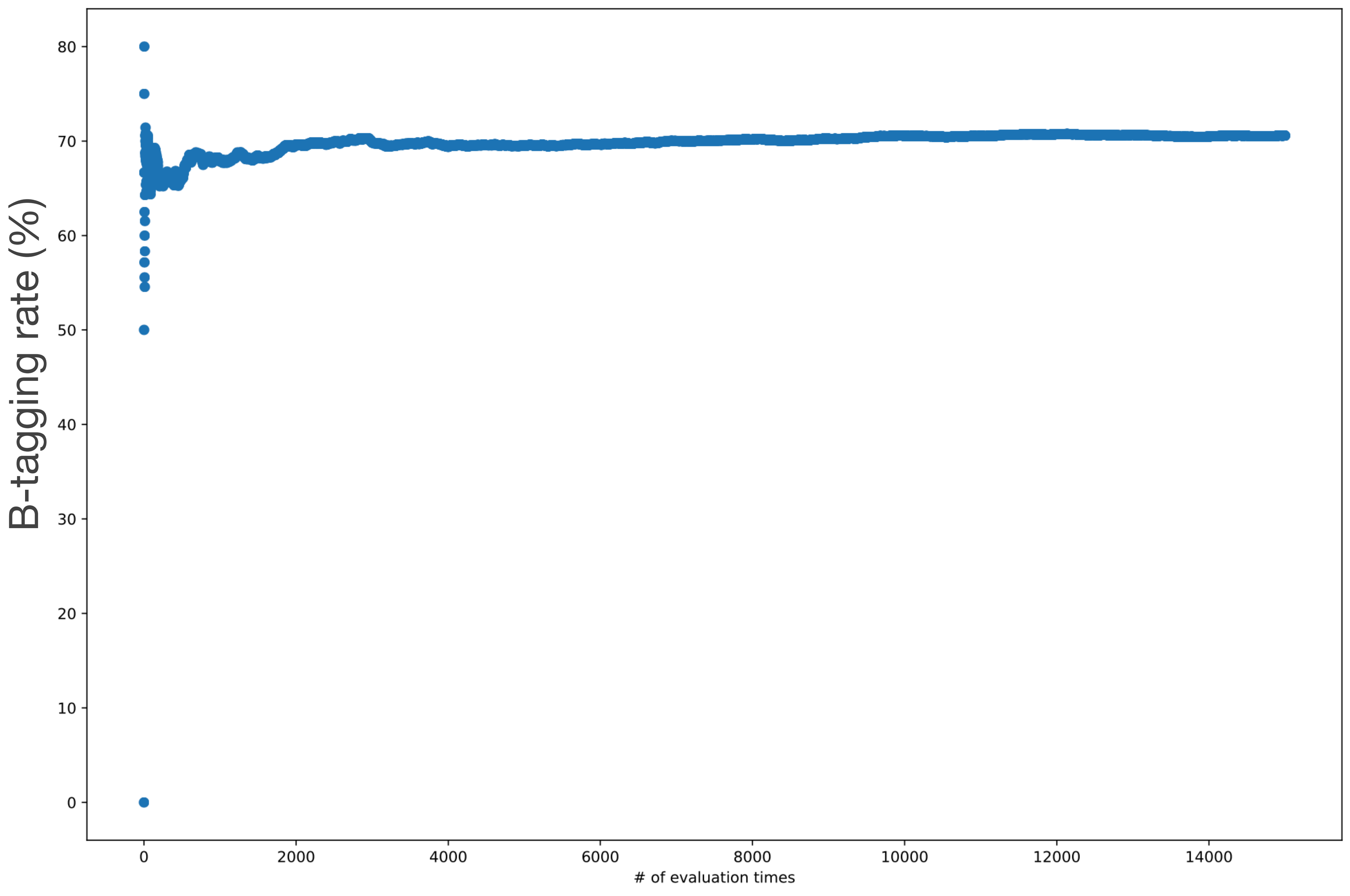
ATL-COM-PHYS-2019-21

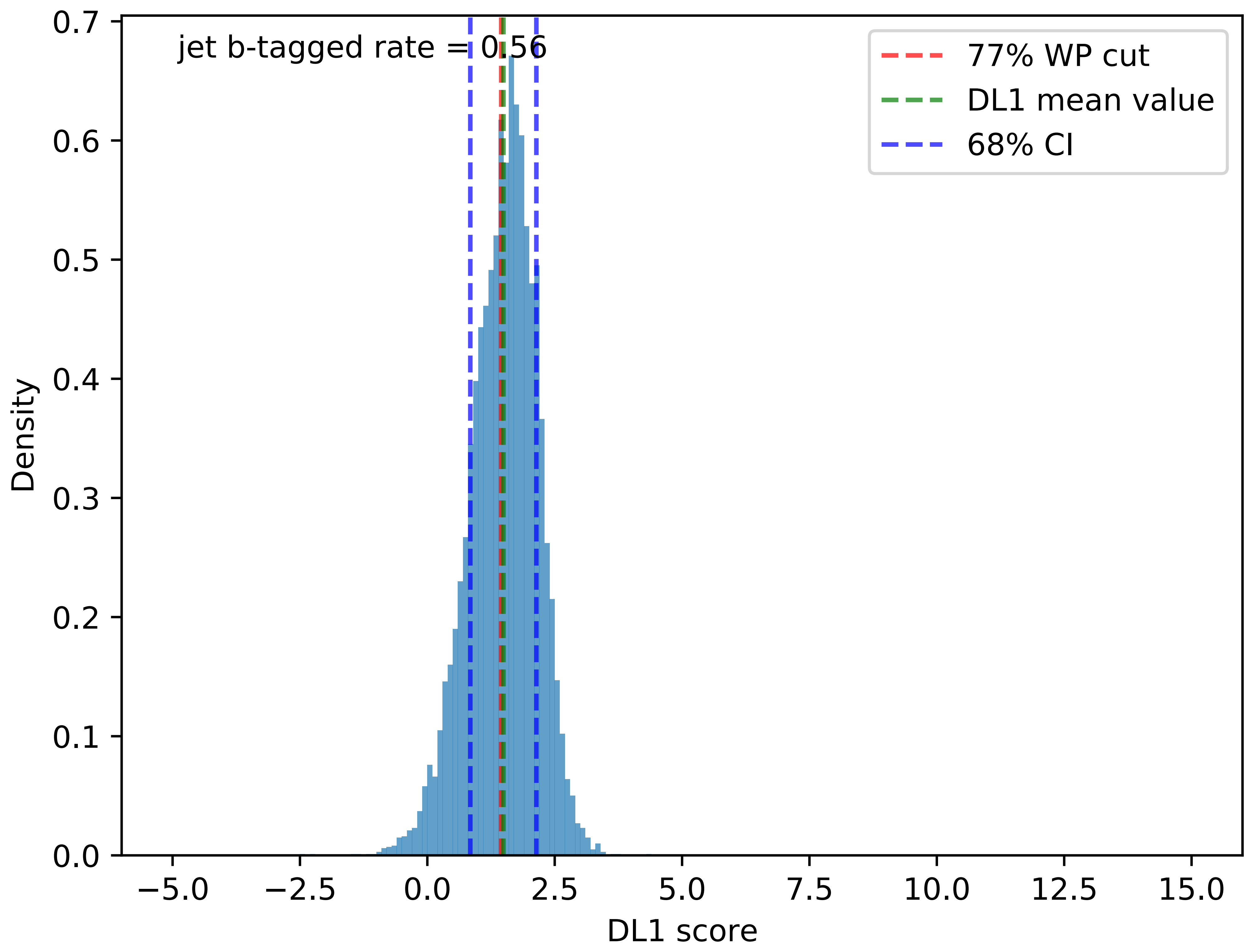



- **Physics analysis, for example searching for DM Z' decays to $b\bar{b}$ can directly benefit from reducing the b-tagging uncertainties**

# DUQ application to b-tagging

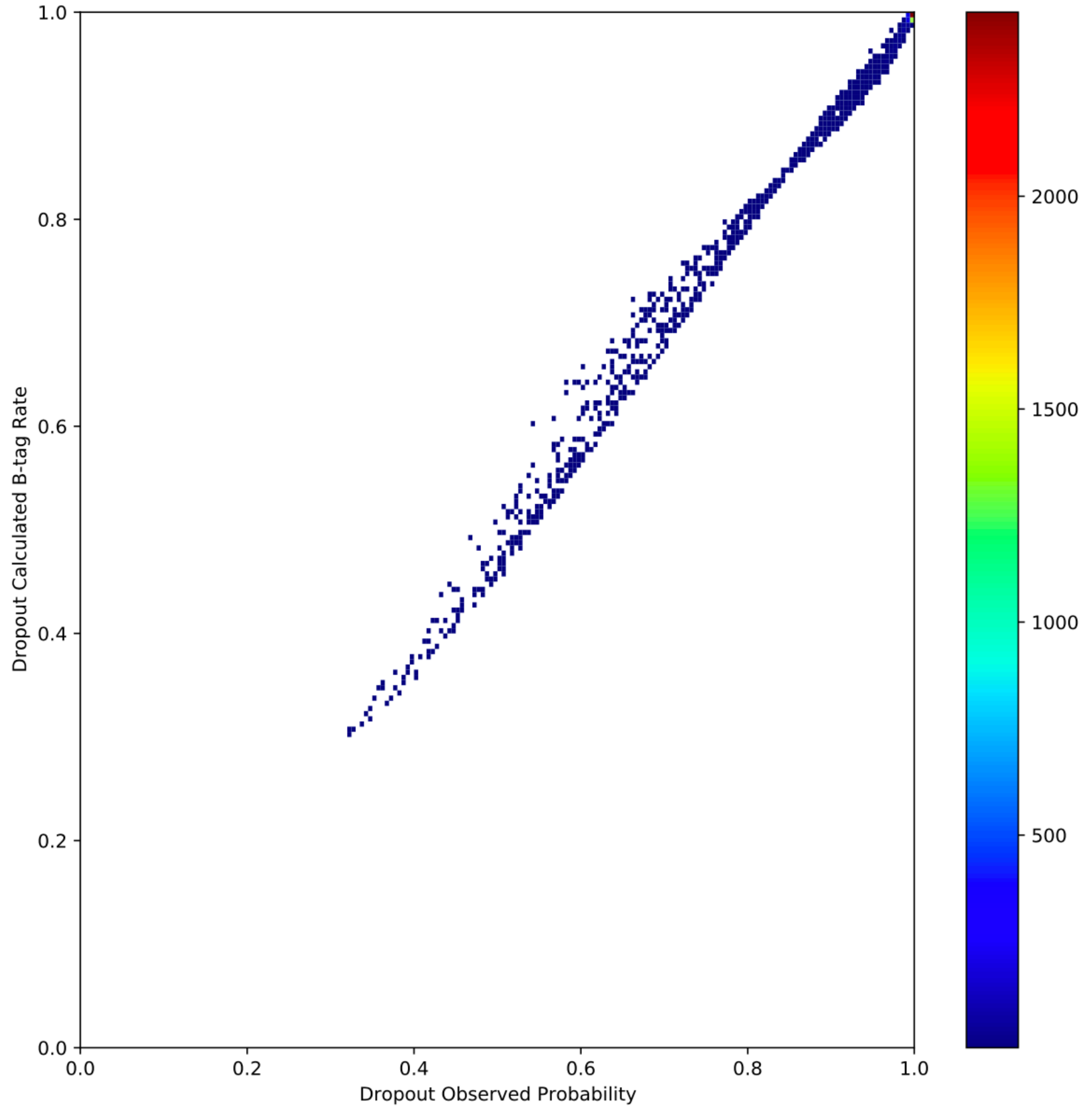
- Repeat the MNIST procedure of calculating probability from significance with Dropout enabled during evaluation
- Evaluated each jet multiple times
  - 10k evaluations for each are enough

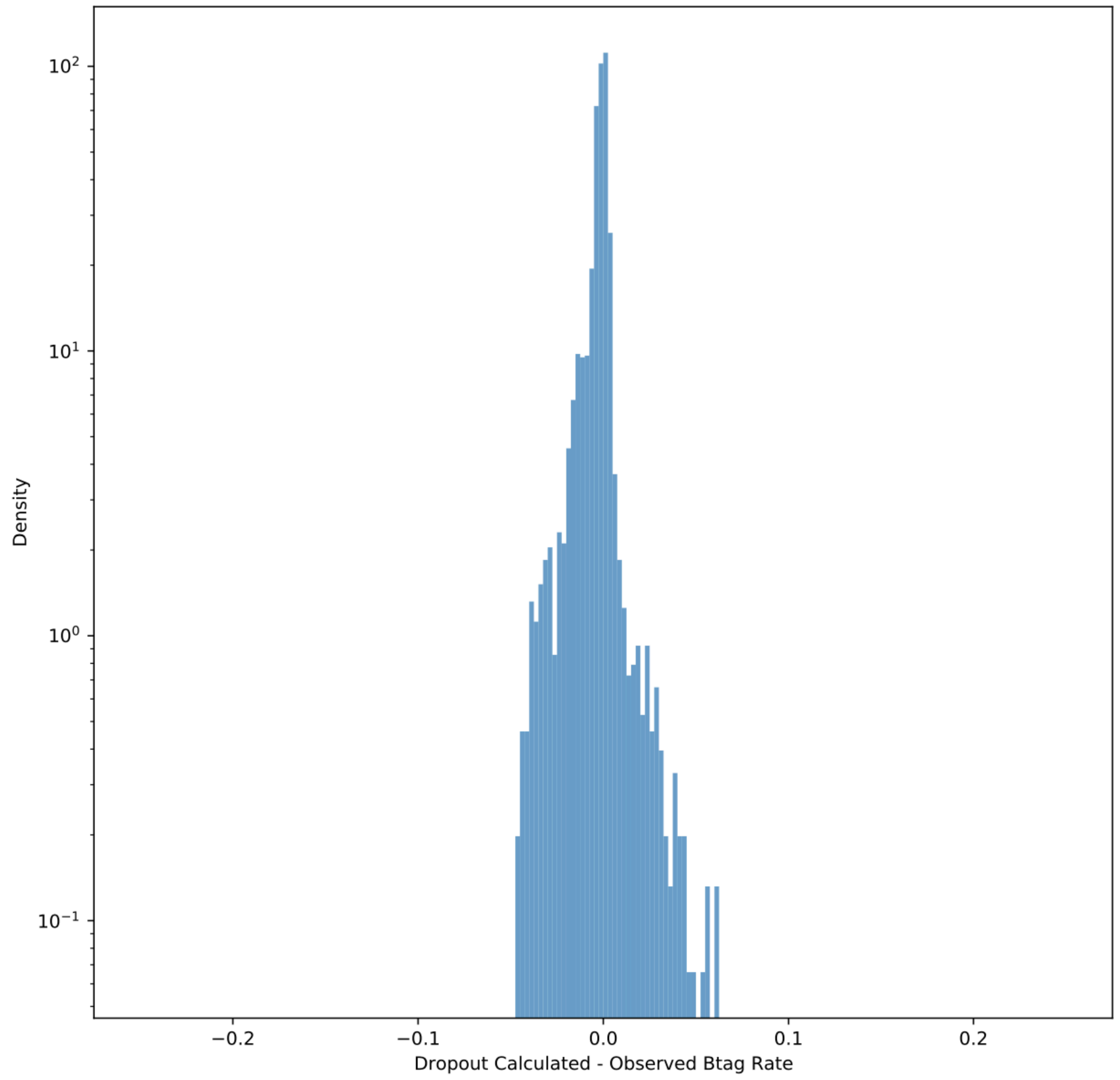


# DUQ application to b-tagging

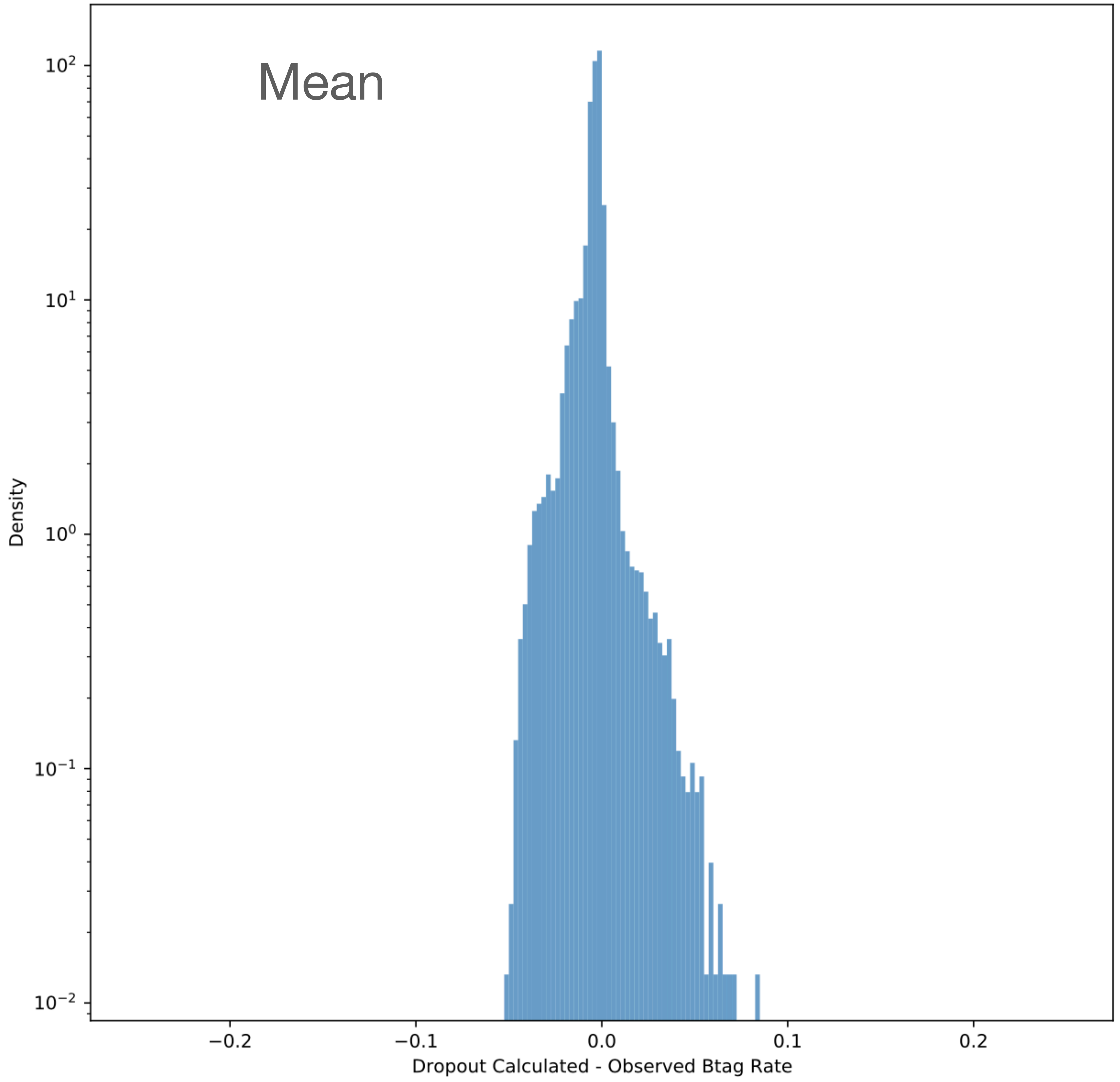
- Calculated vs Observed probability
  - Quite diagonal, indicates calculated probability well reflect jet accuracy
  - The difference is centered at 0 with a width of 2%

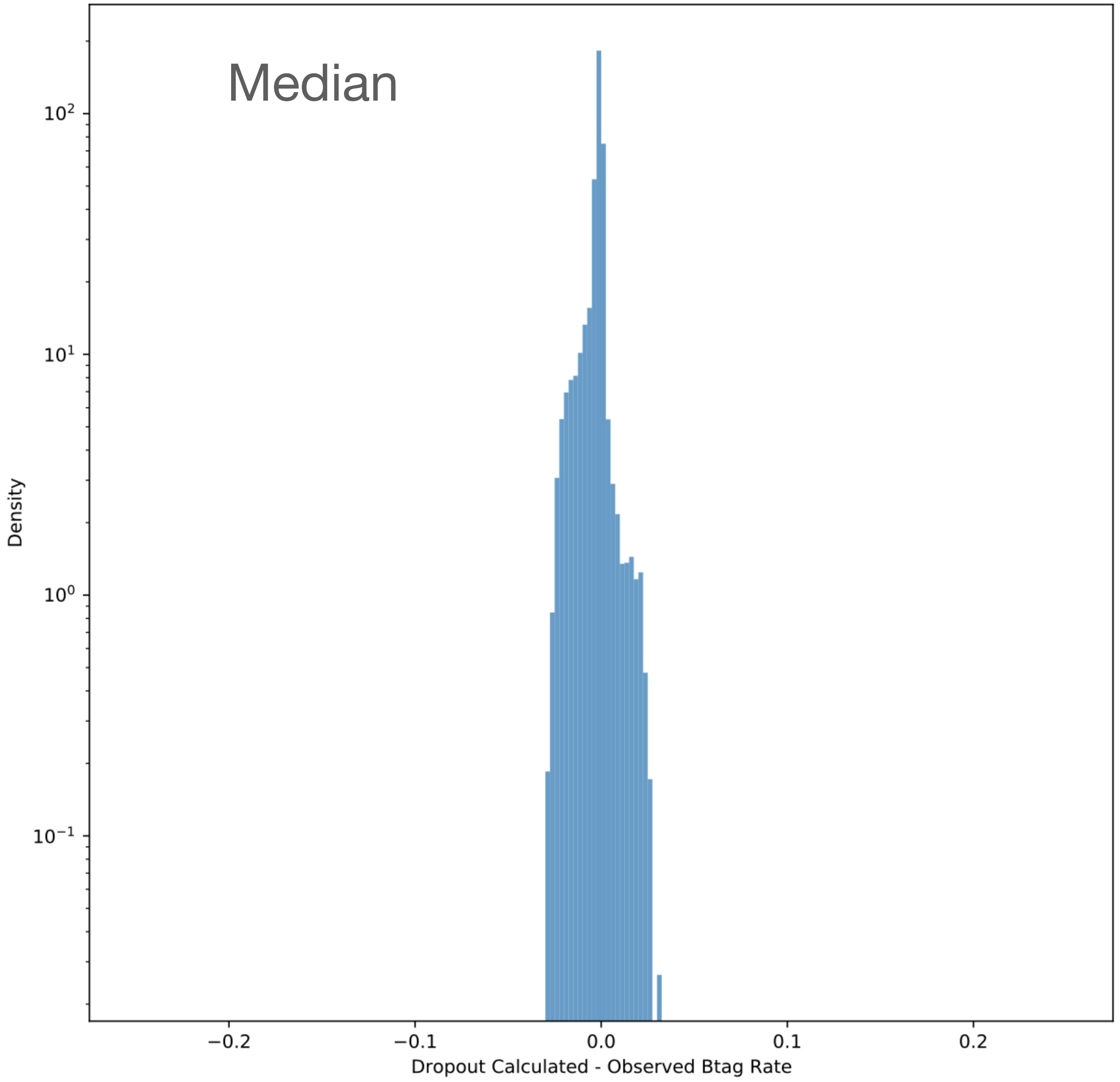


# Mean vs Median

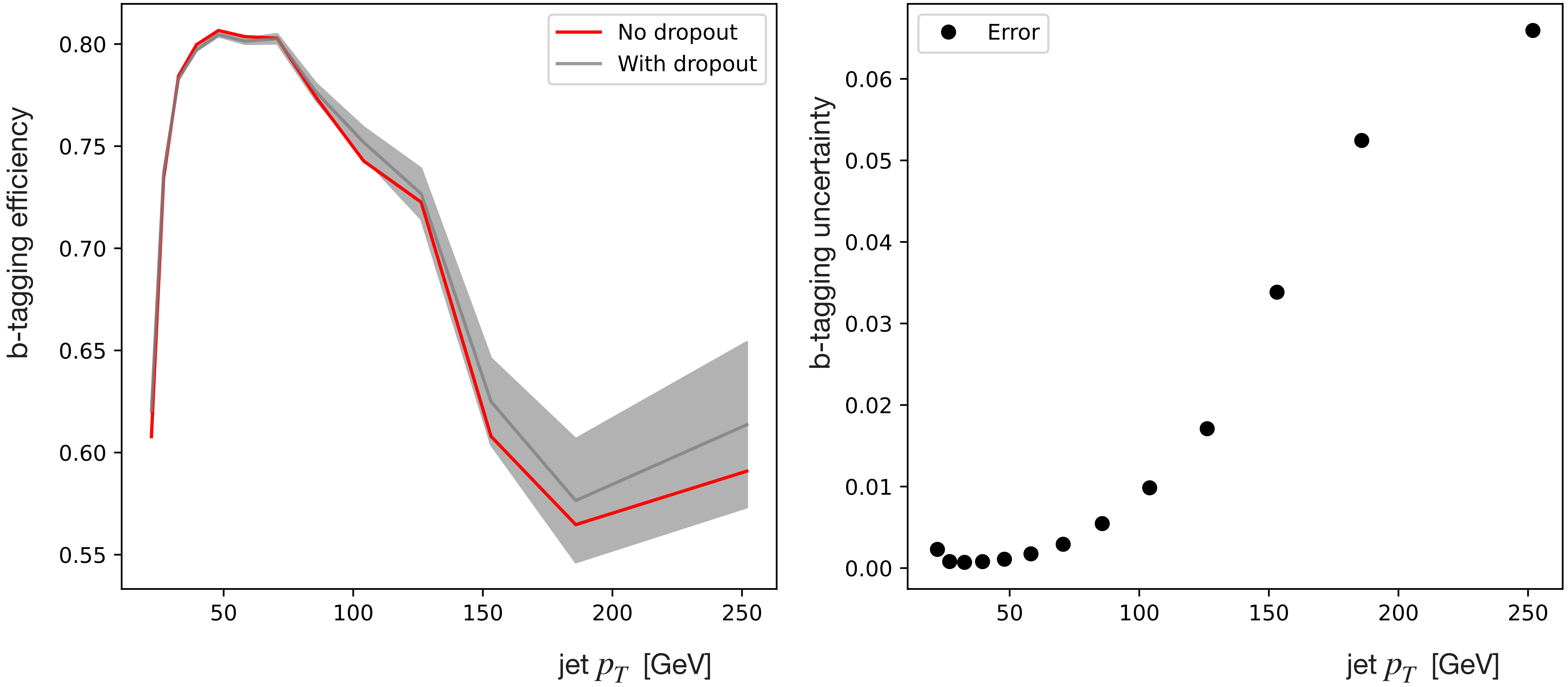
- Using median capture a better quality of the uncertainty than using mean value of the DL1 distribution for each jet



# DUQ application to b-tagging

- **DUQ method performed to get b-tagging efficiency as a function of jet transverse momentum**
- **Sample jet transverse momentum up to 250 GeV, within ~7% uncertainty noticed**



# Summary

- Using Dropout to capture uncertainty
  - Enabling Dropout during evaluation for multiple time samples the posterior probability distribution
  - Calculate per object significance and categorization probability using the median and asymmetric 68% confidence interval

- Method tested on the MNIST database
  - Calculated probability accurately predicts image and sample accuracies
  - Bias test performed to verify the method can also accurately accounts for systematic mismodeling

- Preliminary studies done on the application to ATLAS b-tagging
  - Promising uncertainty capture