ENSEMBLES AND UNCERTAINTY QUANTIFICATION

IBRAHIM ELSHARKAWY

THIS MATERIAL IS BASED UPON WORK SUPPORTED BY THE U.S. DEPARTMENT OF ENERGY, OFFICE OF SCIENCE, OFFICE OF HIGH ENERGY PHYSICS, UNDER AWARD NUMBER DE-SC0023704.

ABOUT ME

Undergrad at Rice University

- Studied Physics, Applied Math, Philosophy

Second-Year Grad at University of Illinois at Urbana-Champaign

- Advisors: Prof. Yoni Kahn and Prof.Ben Hooberman

Research + Work:

- Two Research Projects **at CMS**
- Few Years of Interning as a **Research Geophysicist**
- **Physics for AI** work with Advisors
- Higgs Uncertainty Challenge

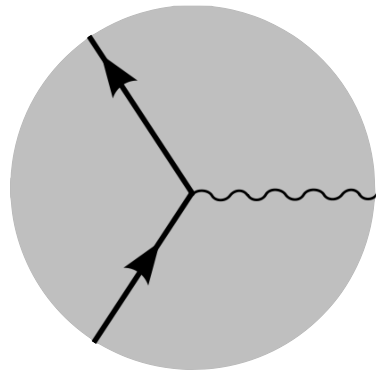
Physics
+
Machine Learning

BY THE END OF THE TALK

1. **Higgs Uncertainty Challenge:** Ensembles of Normalizing Flows, Systemic Uncertainty Robust Classifiers, Nuisance Parameters Estimation

2. **Uncertainty Quantifying From Scaling Laws:** Field Theory for NN, Infinite Width NN, and NN Scaling Laws

TABLE OF CONTENTS



01

Higgs Uncertainty Challenge

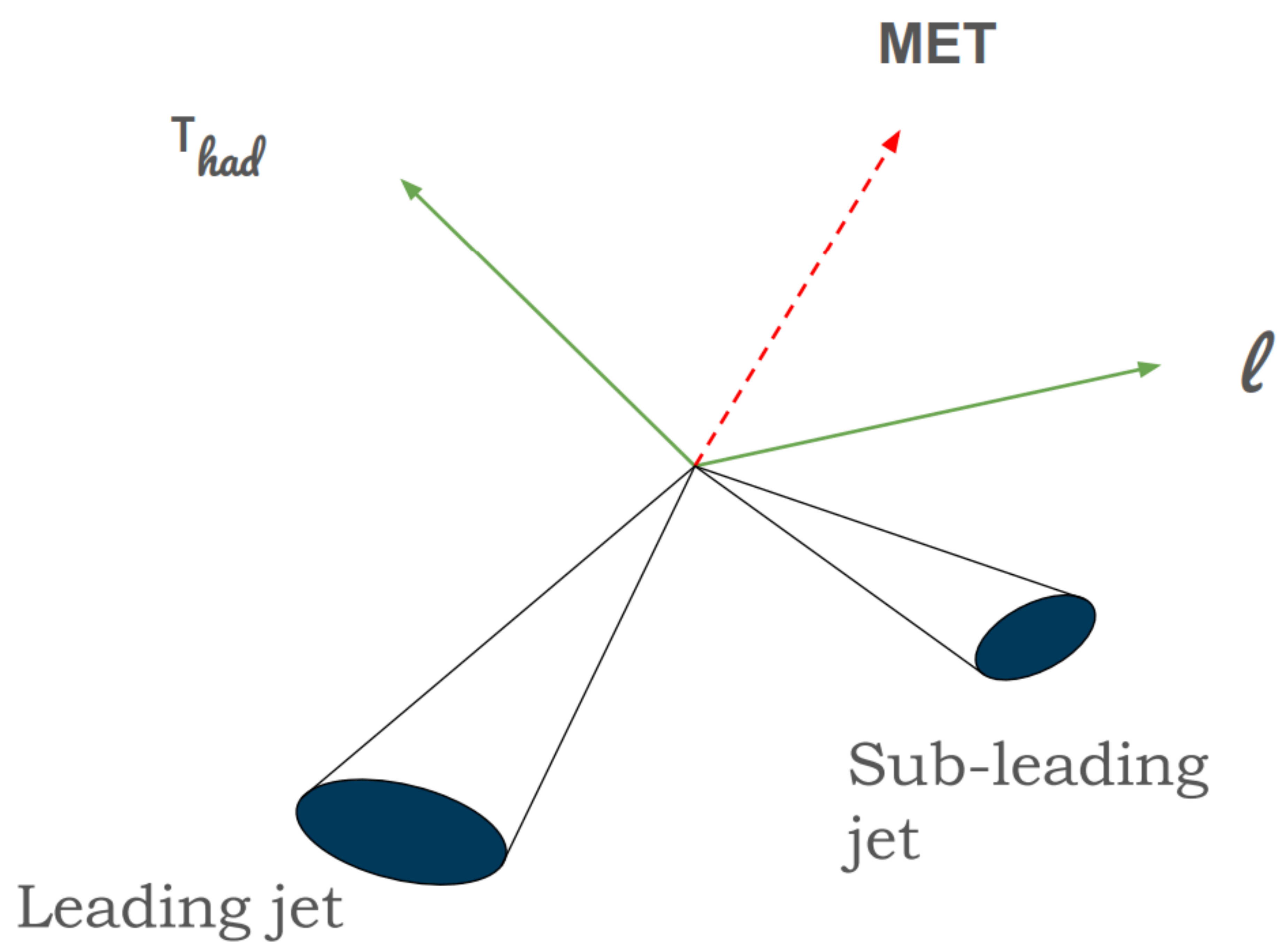
An Introduction to the Challenge,
and the First Iteration

Introduction: The Higgs Uncertainty Challenge

The Goal:

1) **Measure** the signal strength $\mu = \dfrac{\text{Observed Higgs}}{\text{Expected Higgs}}$

2) **Give** correct and small 68% CI on the measurement



The **signal process** is $H \to \tau\tau$

Data: **28 Input Features**

https://arxiv.org/abs/2001.08361

Introduction: The Higgs Uncertainty Challenge

Variable	Mean	Sigma	Range
α_{tes}	1.	0.01	[0.9, 1.1]
α_{jes}	1.	0.01	[0.9, 1.1]
$\alpha_{\text{soft\_met}}$	0.	3.	[0., +∞]
$\alpha_{\text{ttbar\_scale}}$	1.	0.25	[0., +∞]
$\alpha_{\text{diboson\_scale}}$	1.	0.025	[0., +∞]
$\alpha_{\text{bkg\_scale}}$	1.	0.01	[0., +∞]

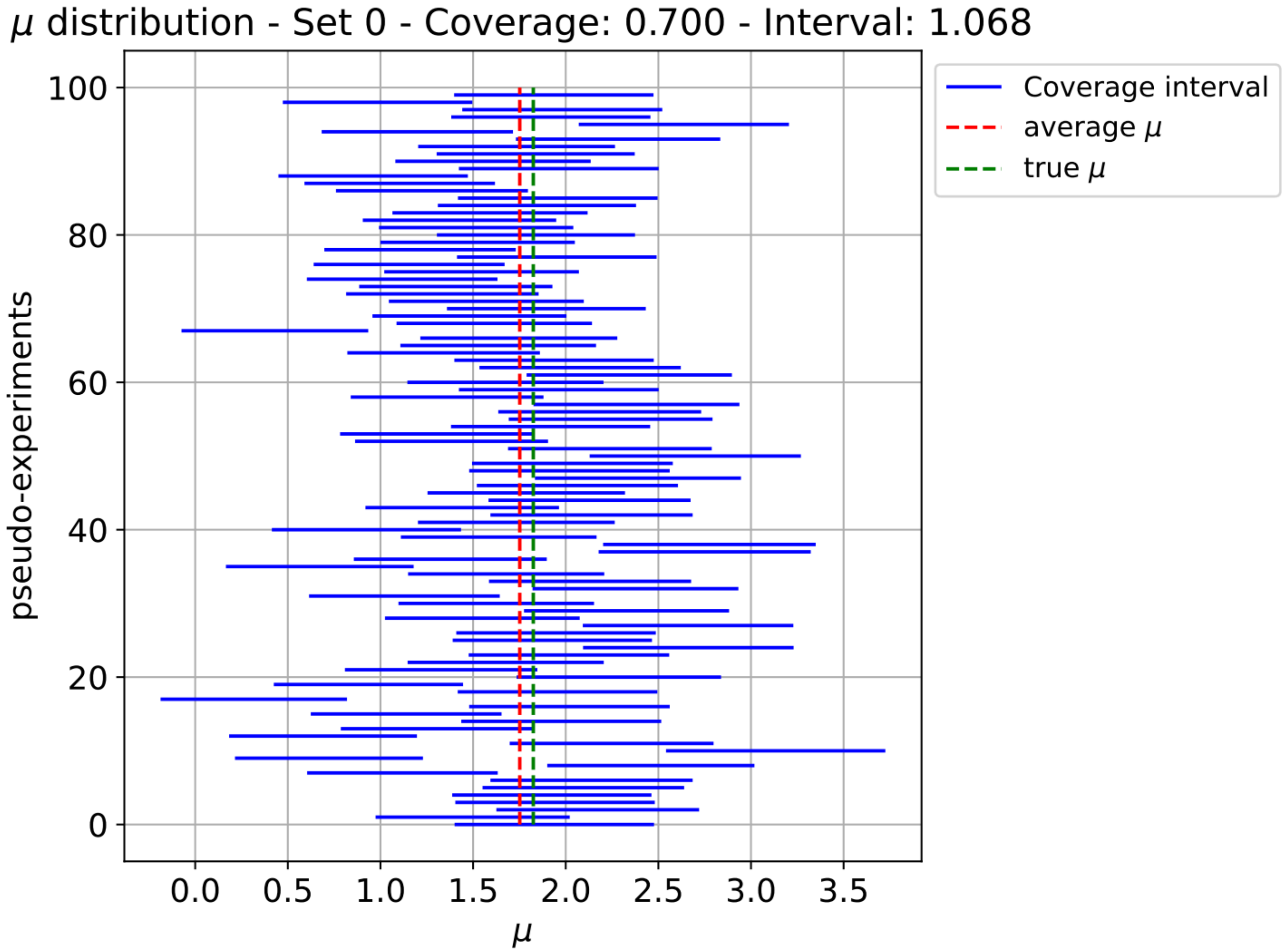
Six nuances parameters

Distorts the 28 features in a unknown nonlinear way

Method is evaluated by:

- Running 100 pseudo-experiments with **different nuisance parameters**
- On 10 **different values** of $\mu = [.1, 3]$

CI must be correct ~68% of the time and are **rewarded with smaller intervals**



First Iteration: A Bayesian Approach

$$\mathscr{P}(\mu \mid \{x\}) \propto \mathscr{P}(\{x\} \mid \mu, \theta) = \prod_{j}^{n} \mathscr{P}(x_i \mid \mu, \theta)$$

$$= \prod_{i}^{N} \left(\frac{\mu}{N} \mathscr{P}(x_i^{s} \mid \theta) + \frac{N-\mu}{N} \mathscr{P}(x_i^{bg} \mid \theta) \right)$$

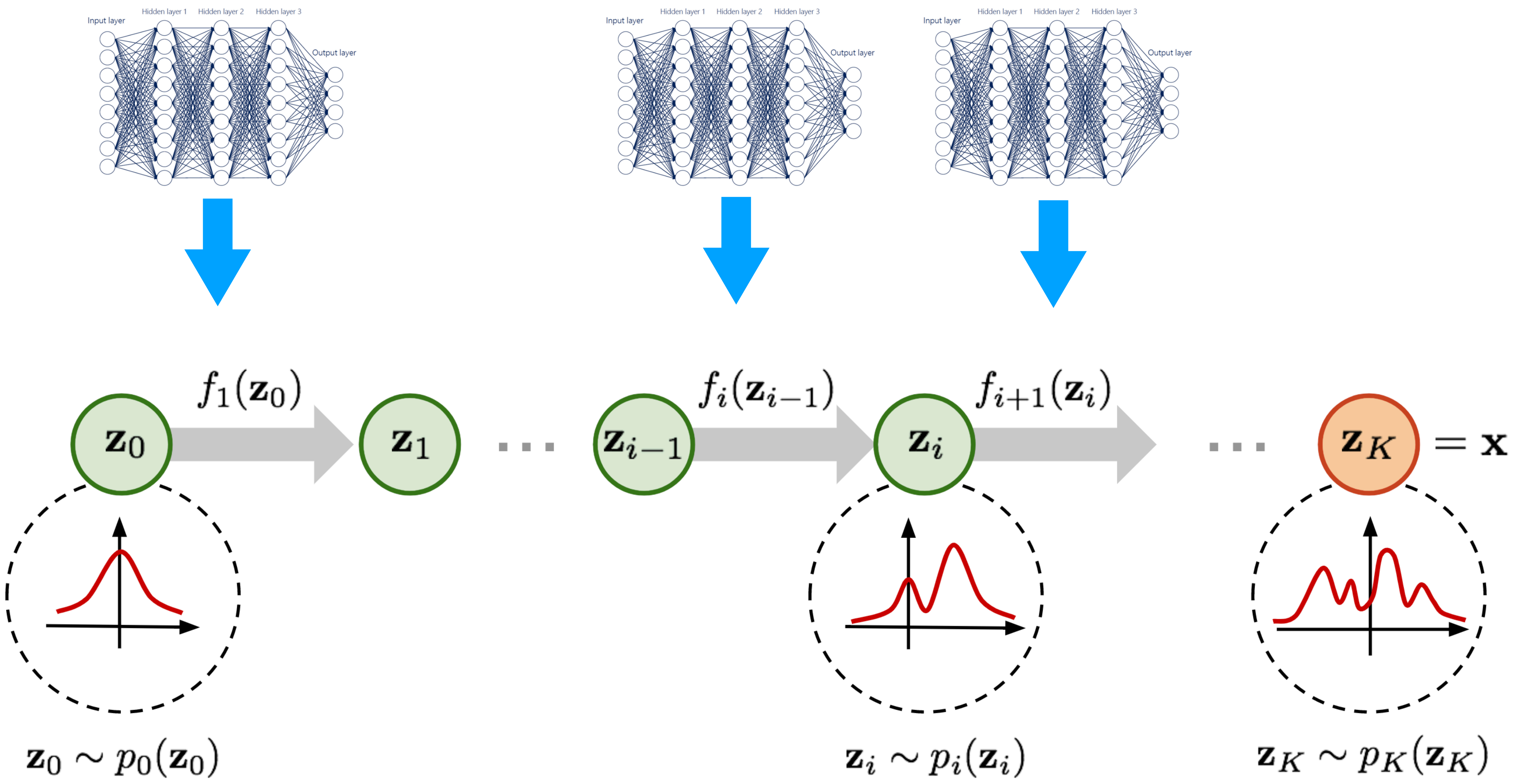
How do we estimate these?

Here
μ = Observed Higgs

First Iteration: Normalizing Flows

$\mathcal{P}(x_i)$

$\sim \mathcal{N}(0,1)$



$\mathcal{P}(x_i^{bg})$

OR

$\mathcal{P}(x_i^{s})$

$f_\theta(x)$

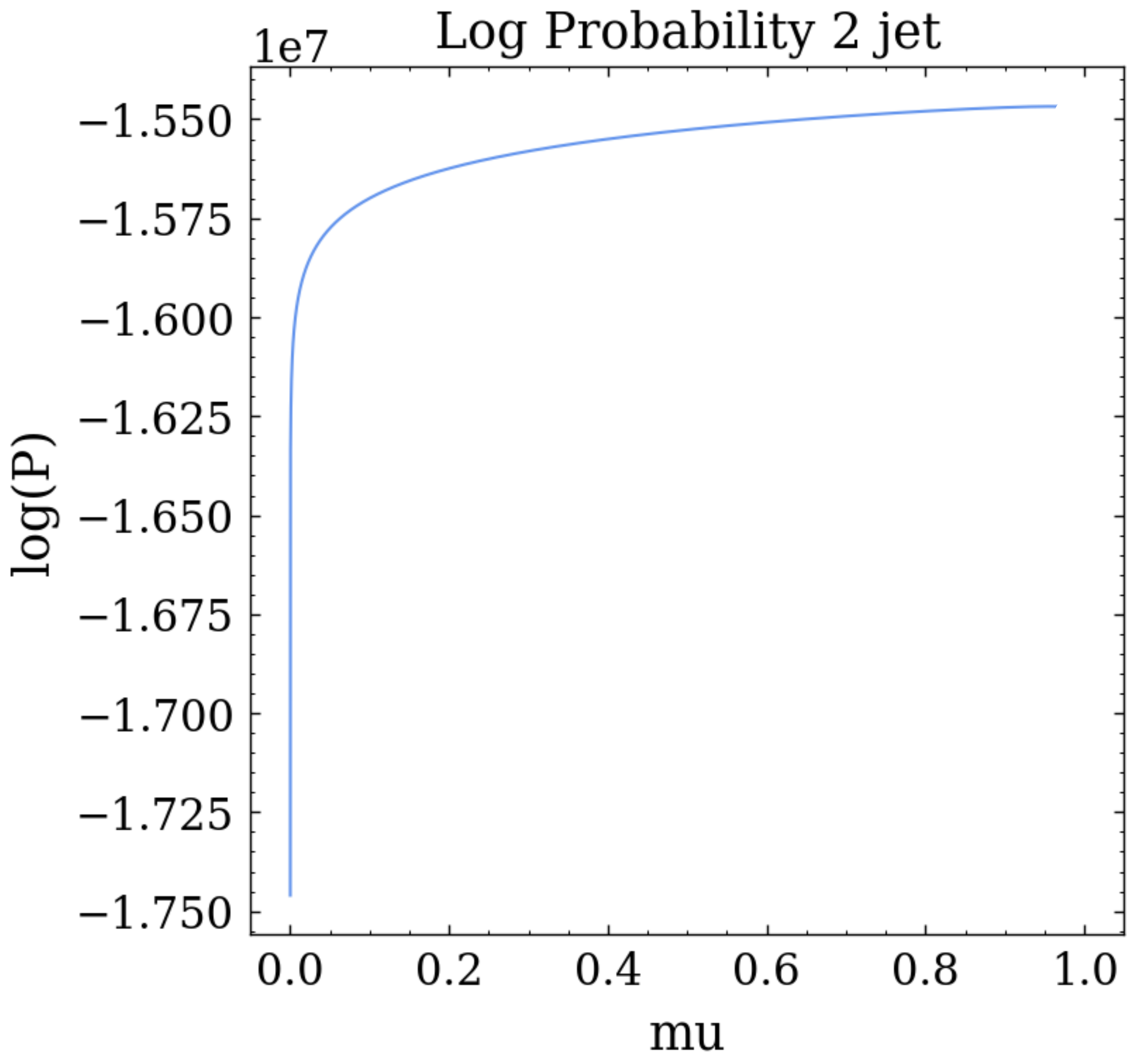
$$\mathcal{L}(\theta) = -\mathbb{E}\left[\log p_Z\left(f_\theta(x)\right) + \log\left|\det\left(\frac{\partial f_\theta}{\partial x}\right)\right|\right]$$

First Iteration: One Problem

$$\mathscr{P}(\{x\}\,|\,\mu,\theta) = \prod_j^n \mathscr{P}(x_i\,|\,\mu,\theta) = \prod_i^N \left(\frac{\mu}{N}\mathscr{P}(x_i^s\,|\,\theta) + \frac{N-\mu}{N}\mathscr{P}(x_i^{bg}\,|\,\theta) \right)$$

Not Enough

Information in the NF-Likelihood!



First Iteration: Adversarial Loss

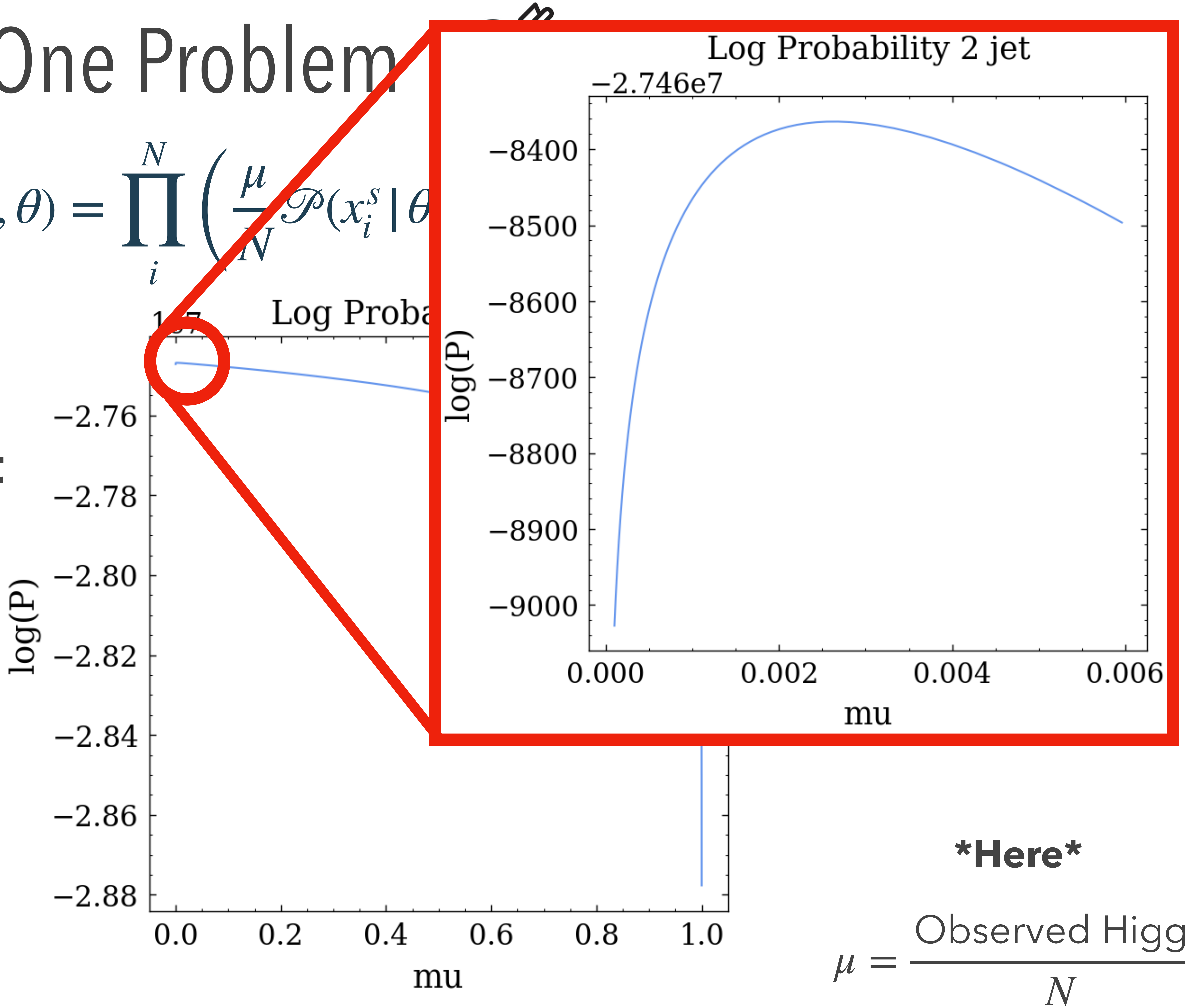
$$\mathscr{L}(\theta) = -\,\mathbb{E}\left[\log p_Z\left(f_\theta(x)\right) + \log\left|\det\left(\frac{\partial f_\theta}{\partial x}\right)\right|\right]$$

$$\mathscr{L}(\theta) = -\,\mathbb{E}\left[c\cdot\left(\log p_Z\left(f_\theta(x)\right) + \log\left|\det\left(\frac{\partial f_\theta}{\partial x}\right)\right|\right) - \log p_Z\left(f_\theta(y)\right) - \log\left|\det\left(\frac{\partial f_\theta}{\partial y}\right)\right|\right]$$

For Example: x = signal events, y = background events

First Iteration: One Problem

$$\mathscr{P}(\{x\}\,|\,\mu,\theta) = \prod_{j}^{n} \mathscr{P}(x_i\,|\,\mu,\theta) = \prod_{i}^{N} \Big(\frac{\mu}{N} \mathscr{P}(x_i^s\,|\,\theta$$



Peaks at close to the right value of mu!

$$\mu_{real} \approx 0.0026$$

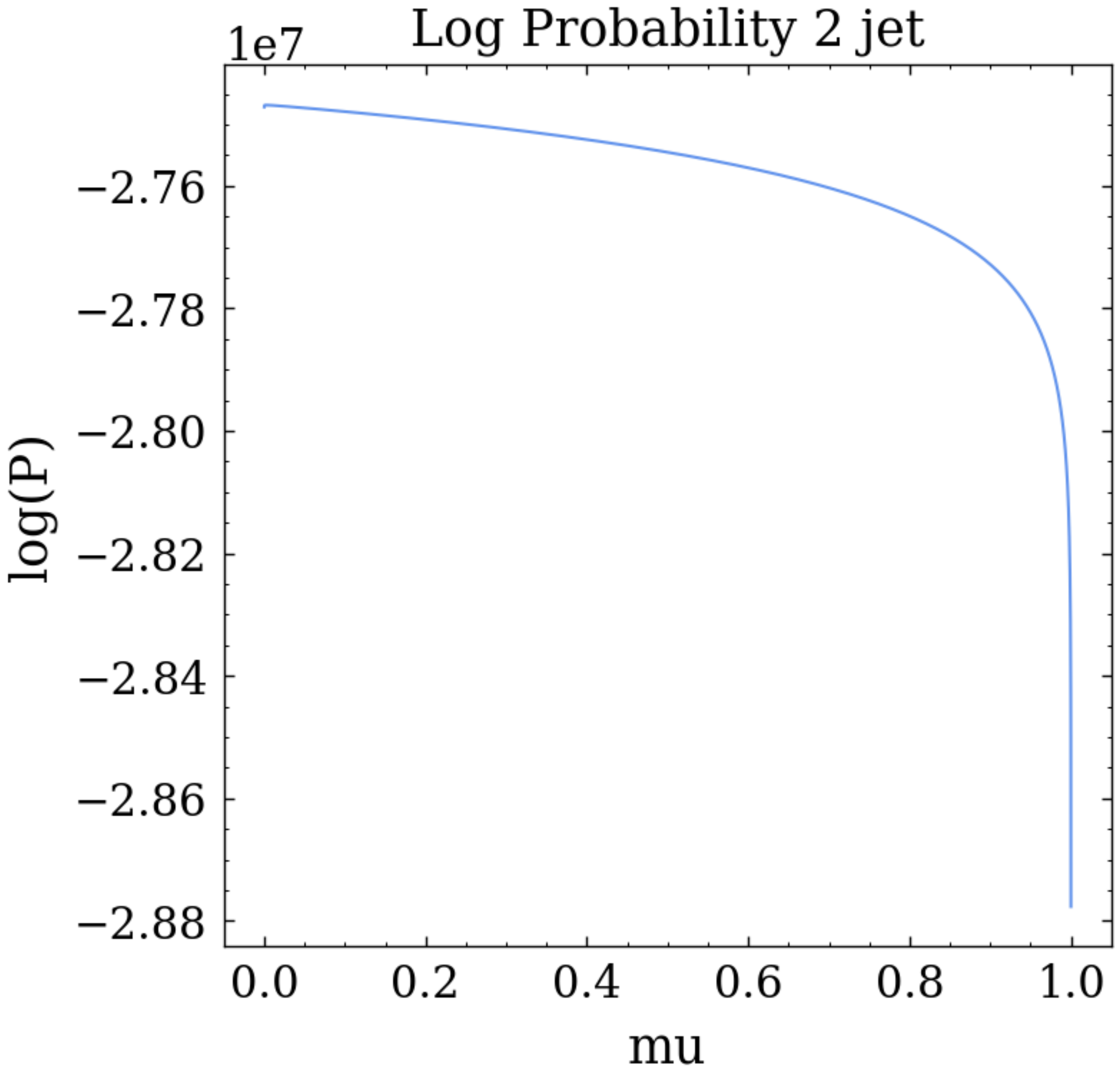
$$\mu_{peak} \approx 0.002$$

Here

$$\mu = \frac{\text{Observed Higgs}}{N}$$

First Iteration: Not Sufficient

$$\mathscr{P}(\{x\} \mid \mu, \theta) = \prod_{j}^{n} \mathscr{P}(x_i \mid \mu, \theta) = \prod_{i}^{N} \left(\frac{\mu}{N} \mathscr{P}(x_i^s \mid \theta) + \frac{N - \mu}{N} \mathscr{P}(x_i^{bg} \mid \theta) \right)$$



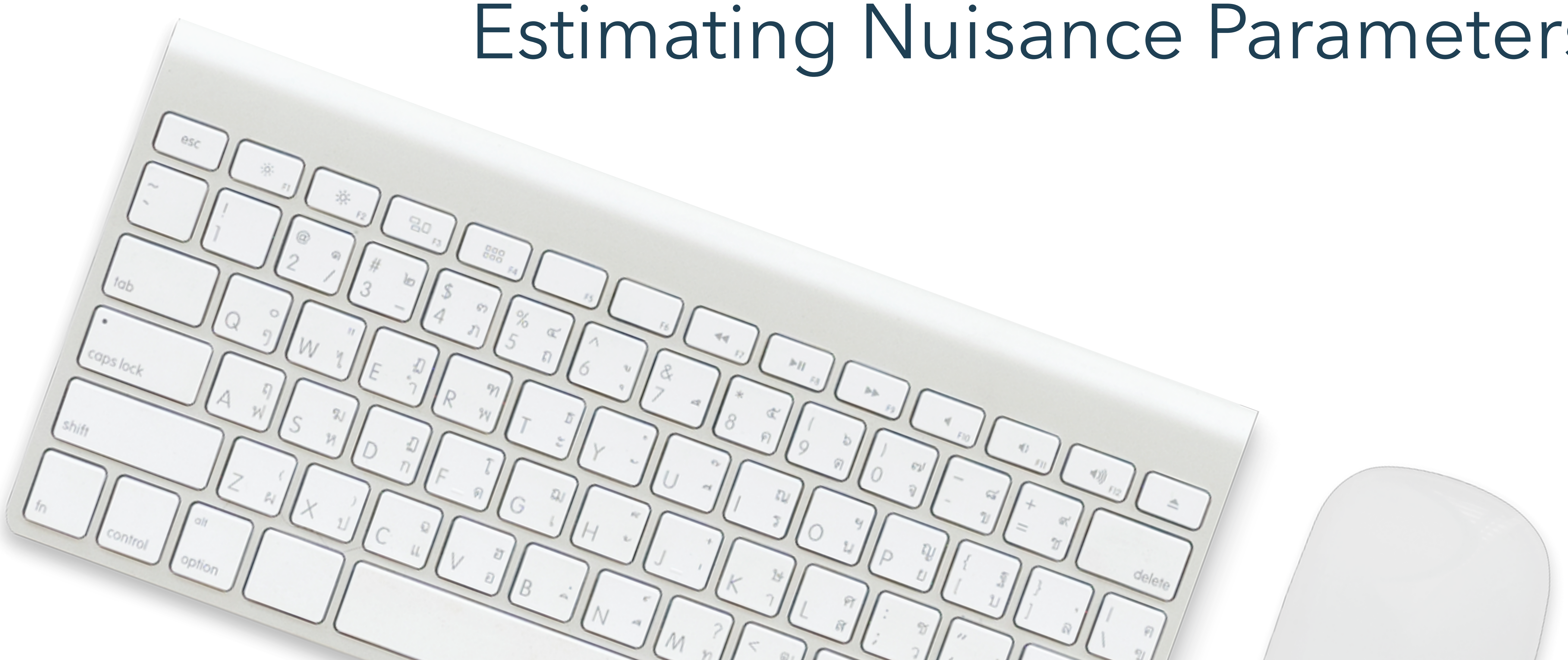
Issue: Not robust to:

1) **Systematics**
   - Very incorrect predictions
2) **Changes in μ**
   - "Sticky" peak and no sensitivity for $\mu < .5$

The Solution

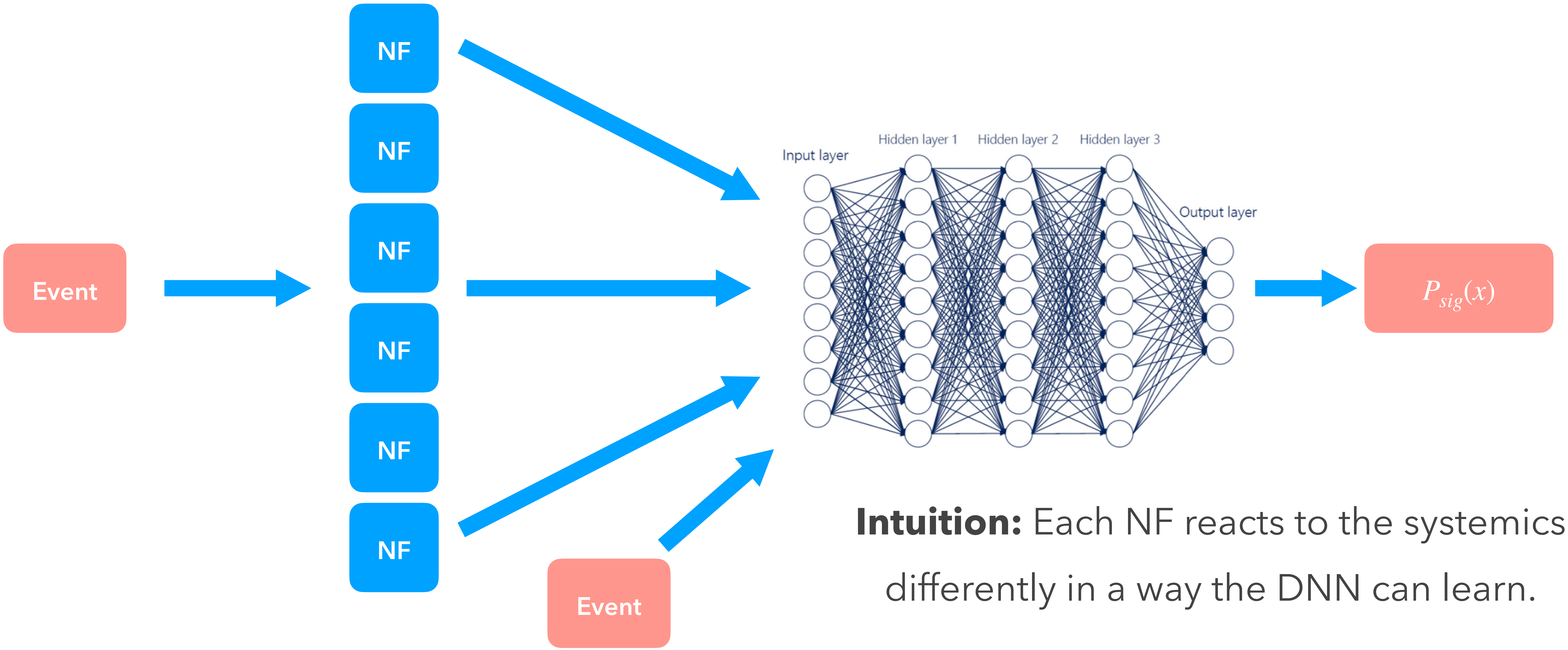
02

An Overview of the Final Iteration, involving NF Ensembles, Classifiers, and Estimating Nuisance Parameters



Final Iteration: Back to Classifiers

Idea: Use Ensemble of NF Likelihoods of as input to a simple DNN classifier

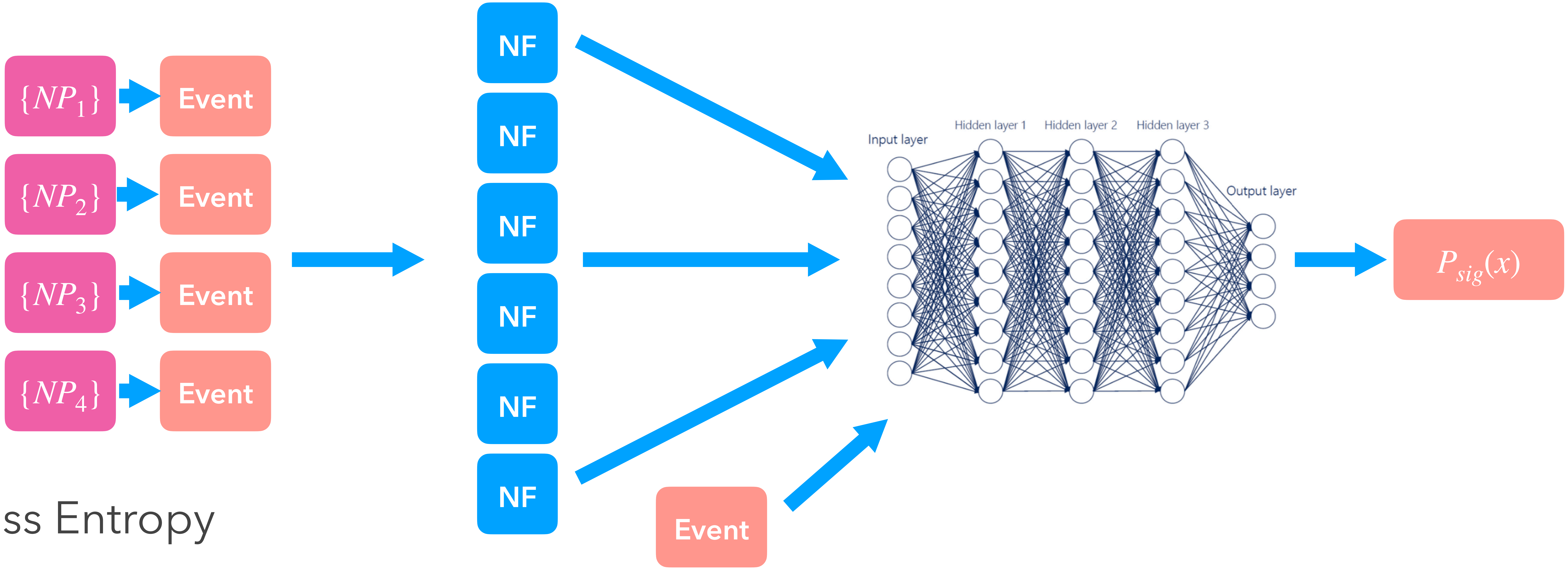


Intuition: Each NF reacts to the systemics differently in a way the DNN can learn.

Final Iteration: Classifier Training

Systematic Robust Training:

Train the same DNN with event data perturbed with a wide variety of nuisance parameters.



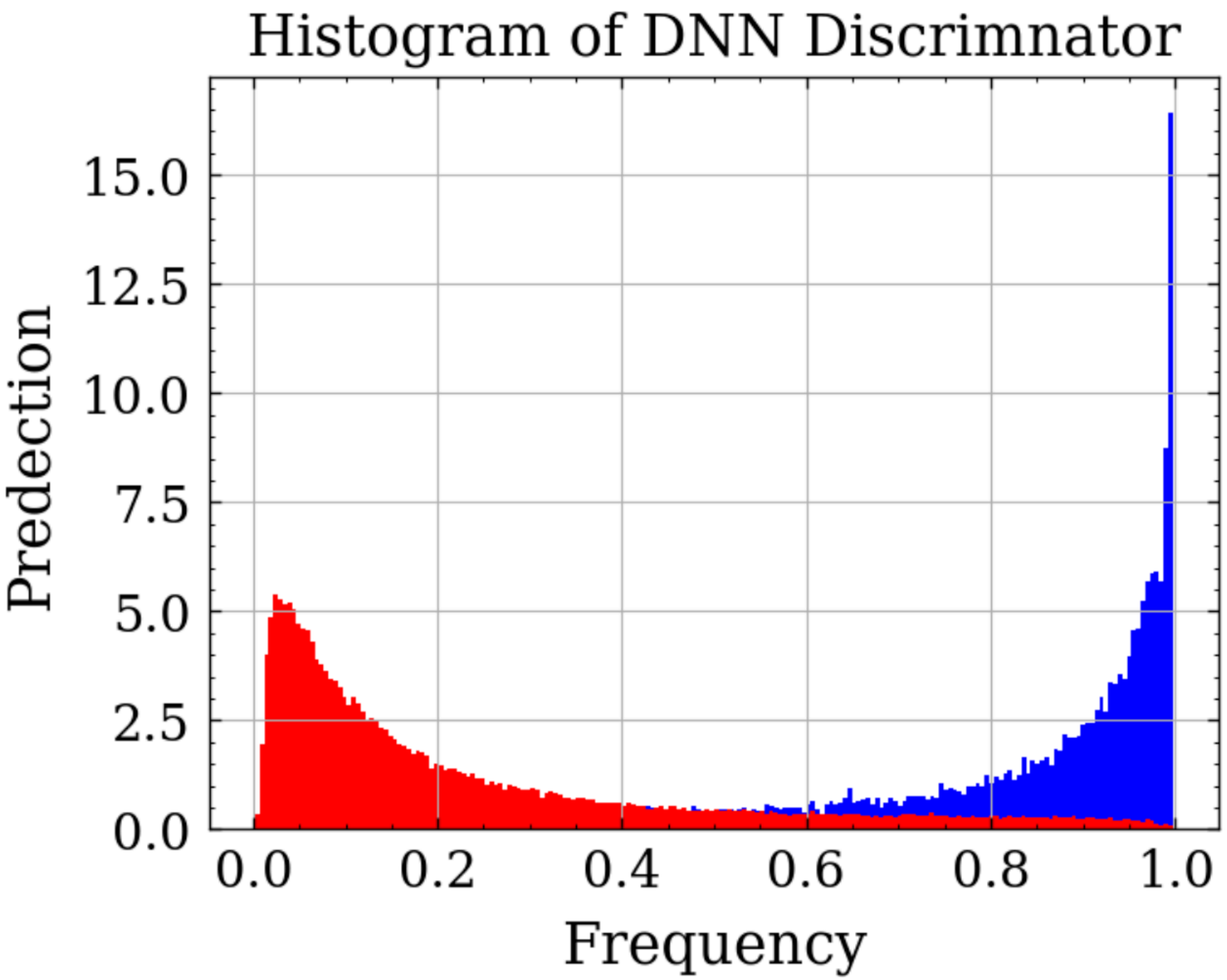
Loss: Binary Cross Entropy

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)\right]$$

Final Iteration: Binned Analysis

Here

$$\mu = \frac{\text{Observed Higgs}}{N}$$



Let k_i be the true counts of events in bin i:

$$E[k_i] = f_i \cdot \mathbf{E}[S_i] + b_i$$

The **Poisson Likelihood** is then given by:

$$P(\{f_i\} \mid \{k_i\}) \propto \mathscr{L}(\{k_i\} \mid \{f_i\}) = \prod_{i=1}^{N} \frac{(f_i S_i + b_i)^{k_i} e^{-(f_i S_i + b_i)}}{k_i!}$$

Then:

$$\mu \equiv \frac{\sum_i f_i \cdot \mathbb{E}[S_i])}{\sum_i (b_i + f_i \cdot \mathbb{E}[S_i])}$$

Final Iteration: Parameter Estimation

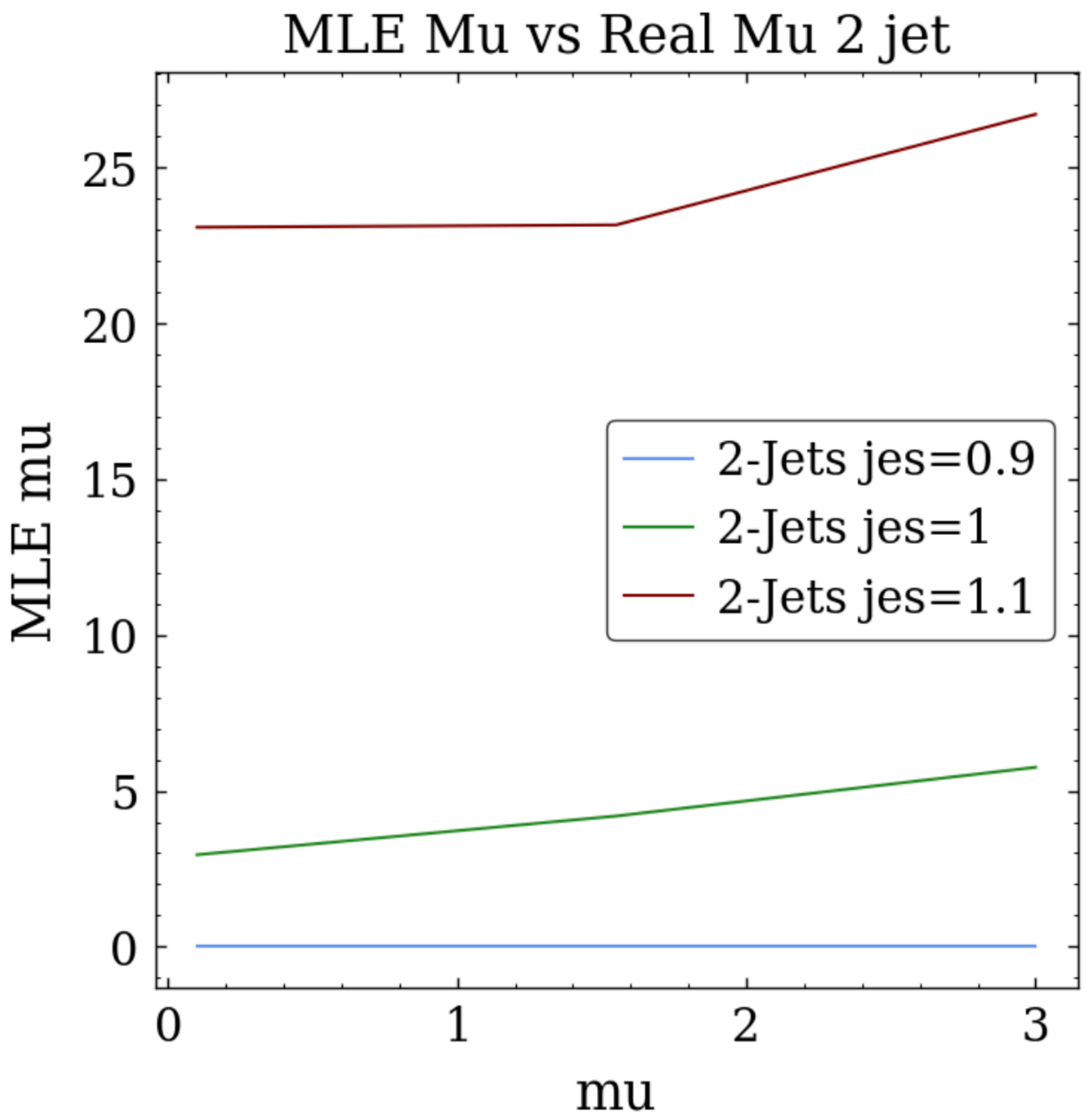
To deal with the **"worst" nuisance parameter** (jes) we can do the same procedure but compute instead:

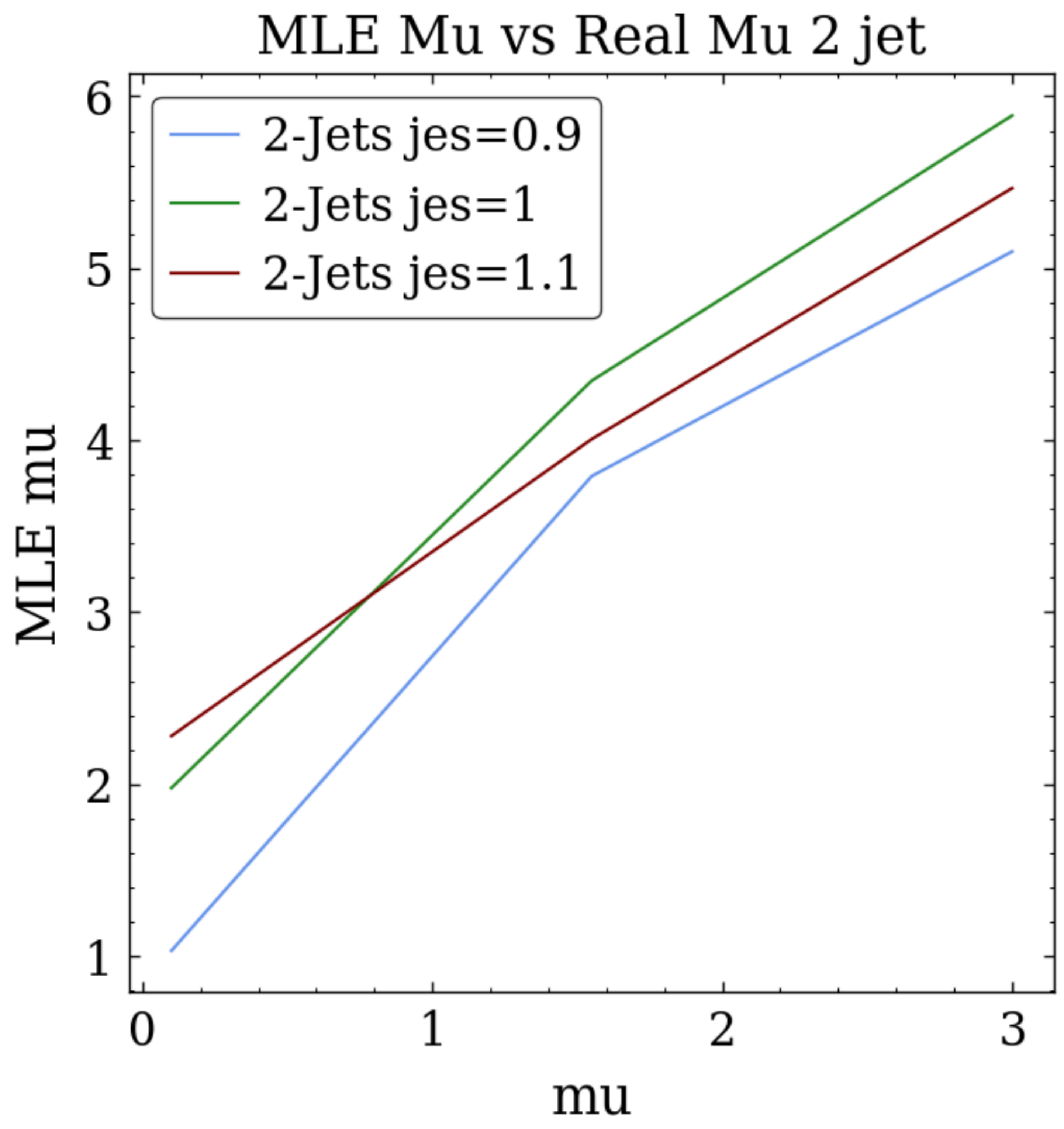
Do MLE on both θ and f_i!

$$P(\{f_i\} \mid \{k_i\}, \theta) \propto \mathscr{L}(\{k_i\} \mid \{f_i\}, \theta) = \prod_{i=1}^{N} \frac{(f_i \mathbf{E}[S_i \mid \theta] + b_i)^{k_i} e^{-(f_i \mathbf{E}[S_i|\theta] + b_i)}}{k_i!} \cdot P(\theta)$$

Prior on

Nuances Parameters

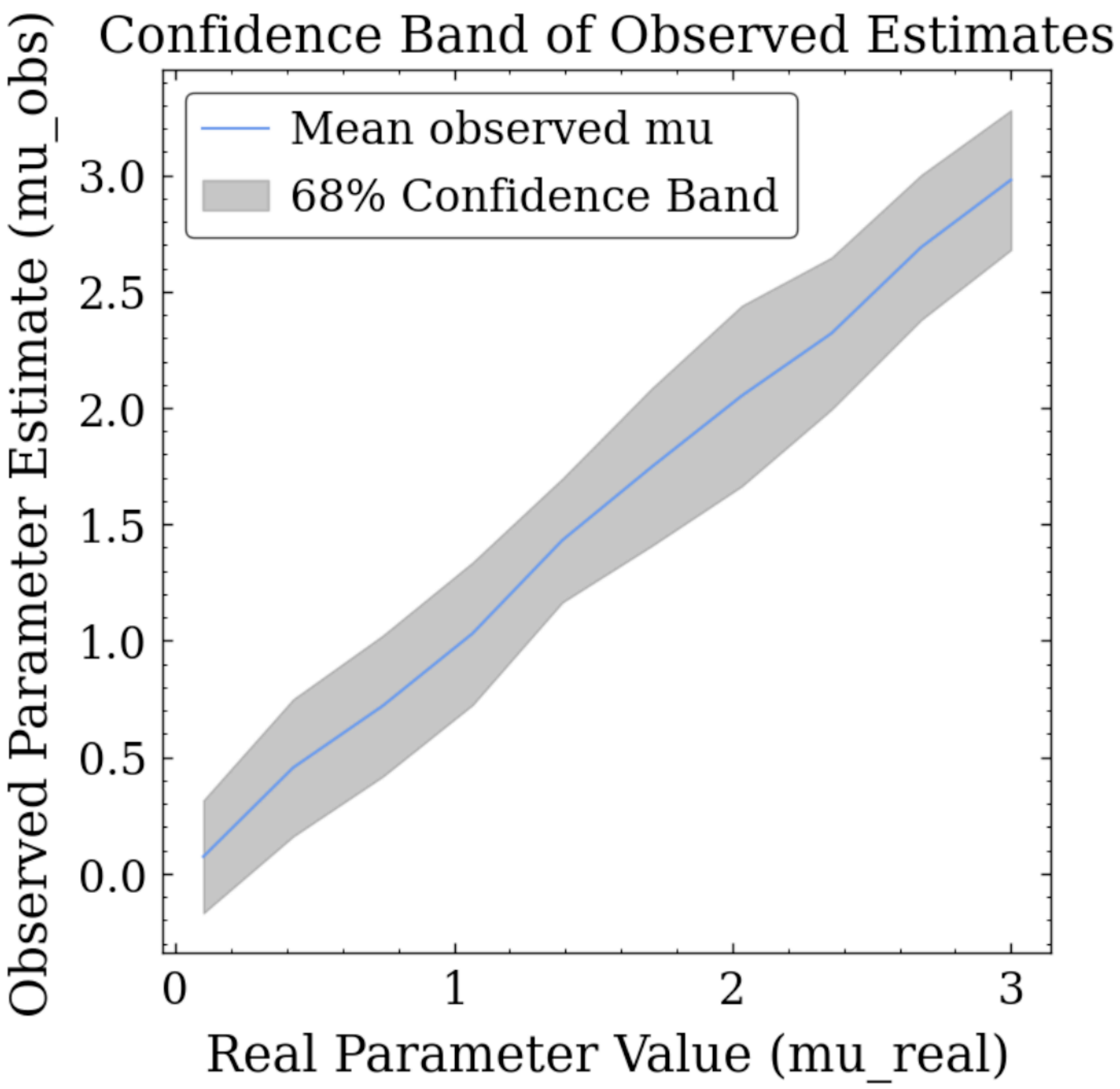




Final Iteration: Neyman Construction

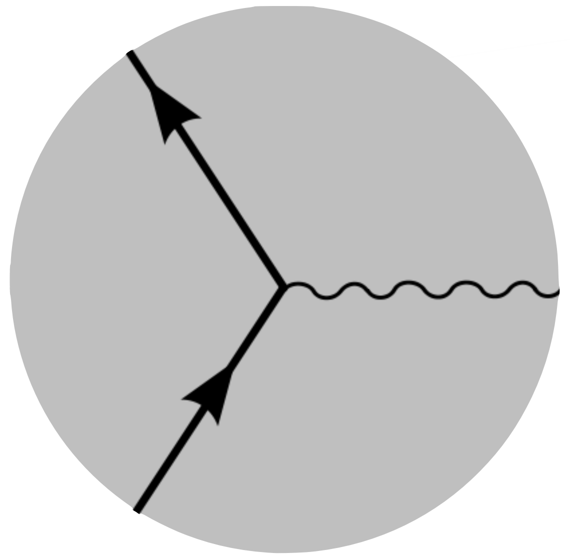
For Error Bars we use the **Neyman Construction**

1) **For each value of real mu** compute estimate ~100 time with a different draw of NP
2) With mean and std of the estimates:
   1) Apply a **Bias Correction** to the mean
   2) And use std as error bar estimate **for a given estimate of mu**



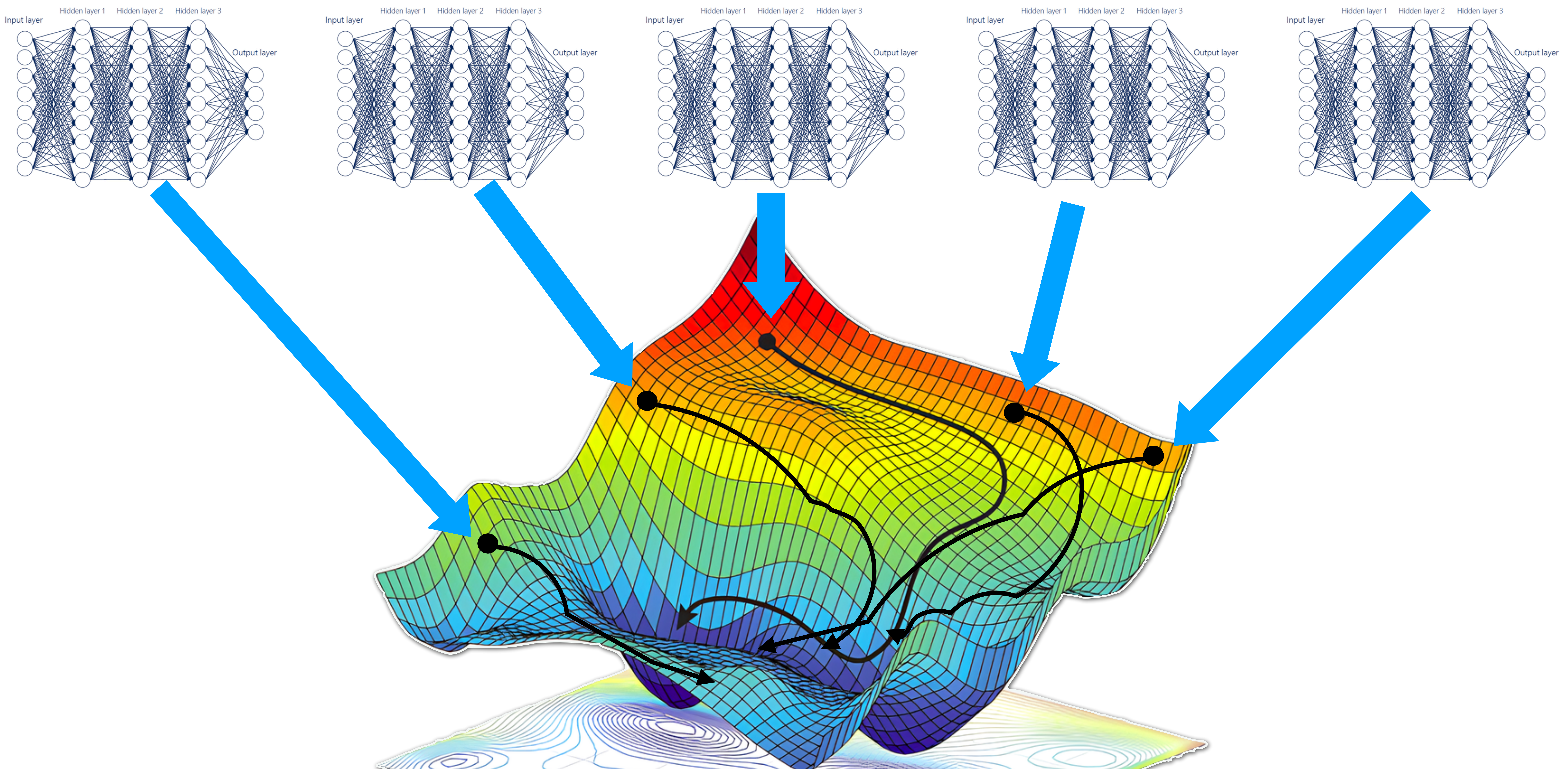
03

Uncertainty Quantifying From Scaling Laws

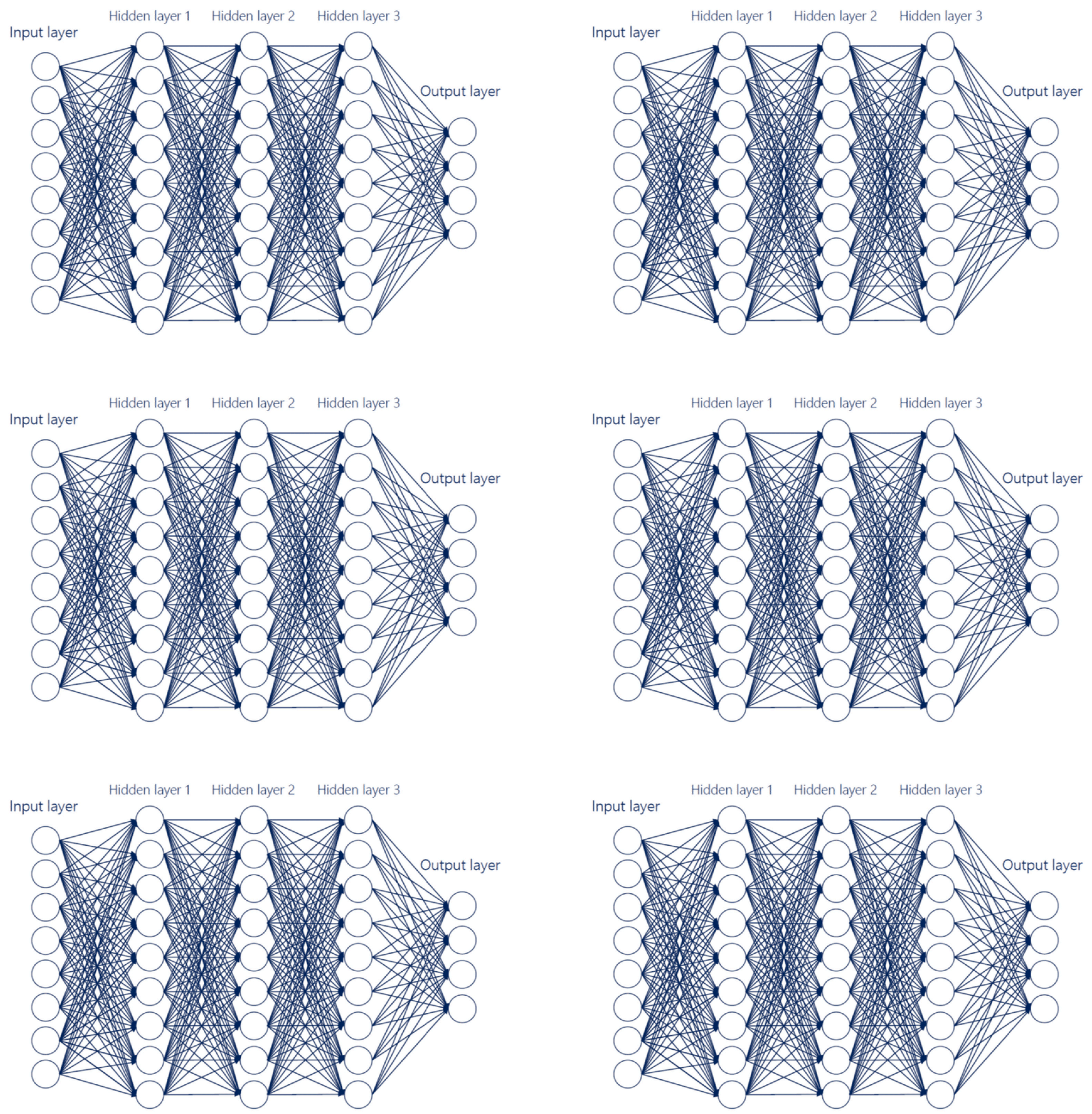


Empirical Results comparing theoretical results with empirical reality

Uncertainty Quantification: Ensemble Uncertainty

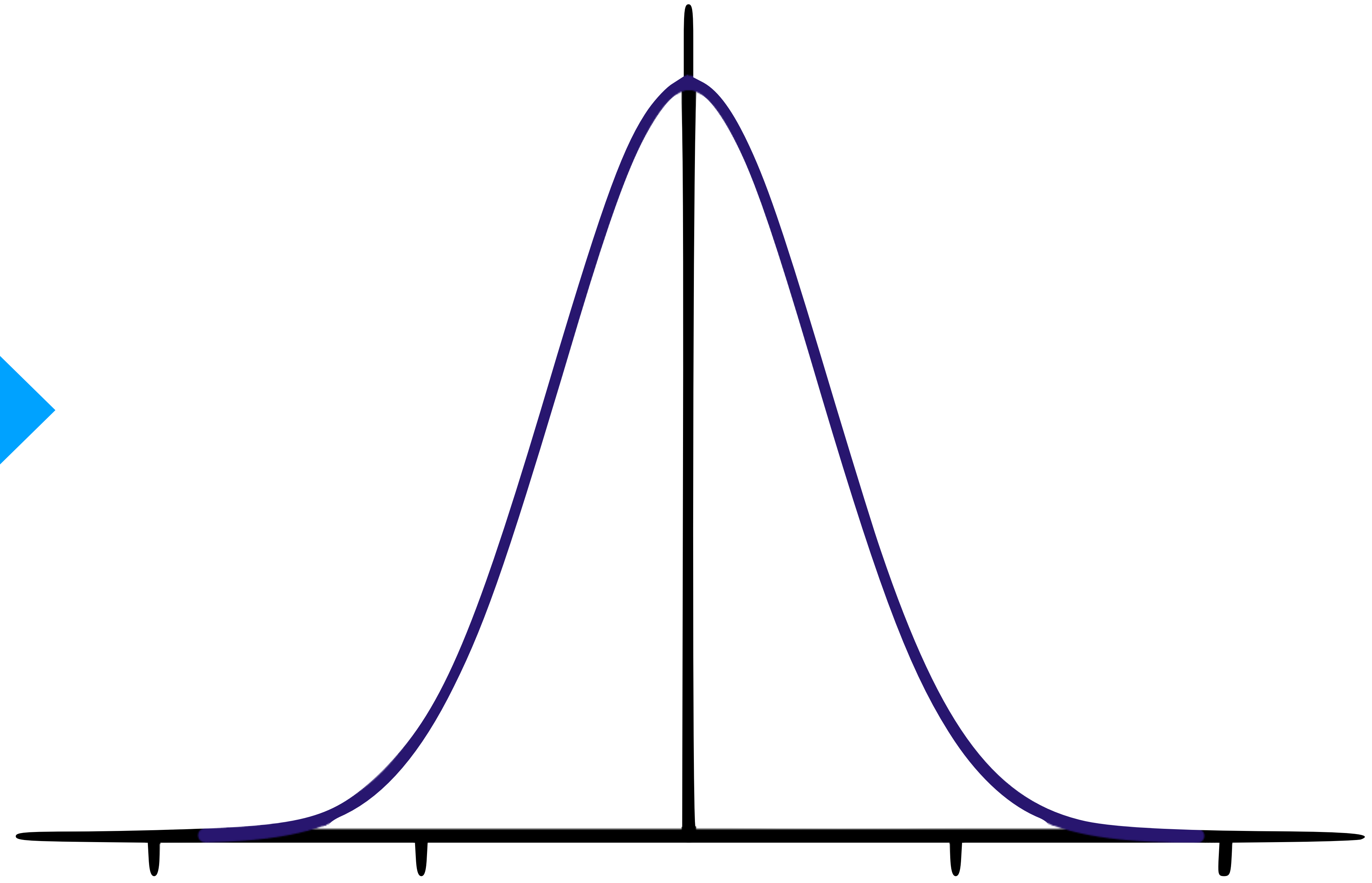


Uncertainty Quantification: Ensemble Uncertainty

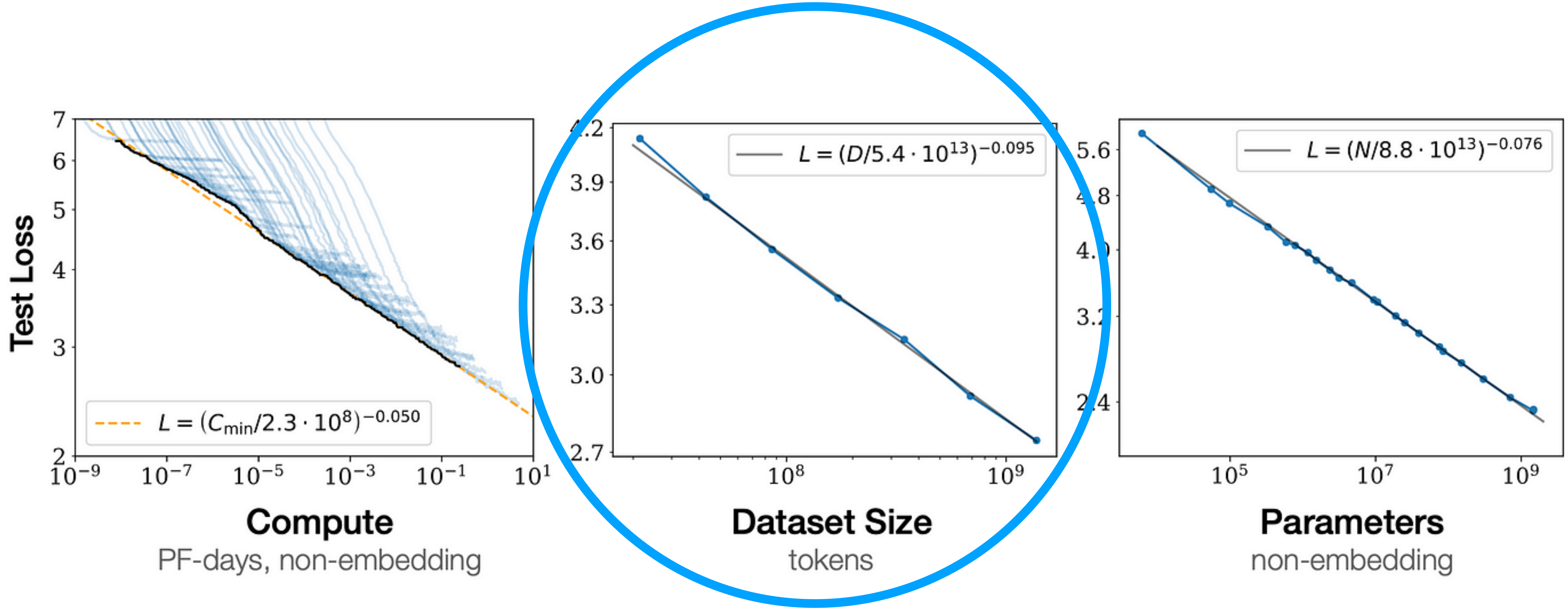


Our Goal: Compute the Variance of an Ensembles Prediction *without training an ensemble*

Central Limit Theorem

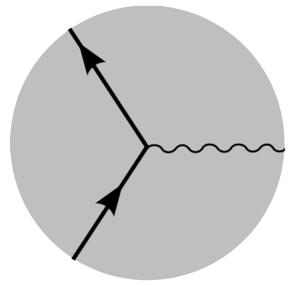


Neural Network Scaling Laws

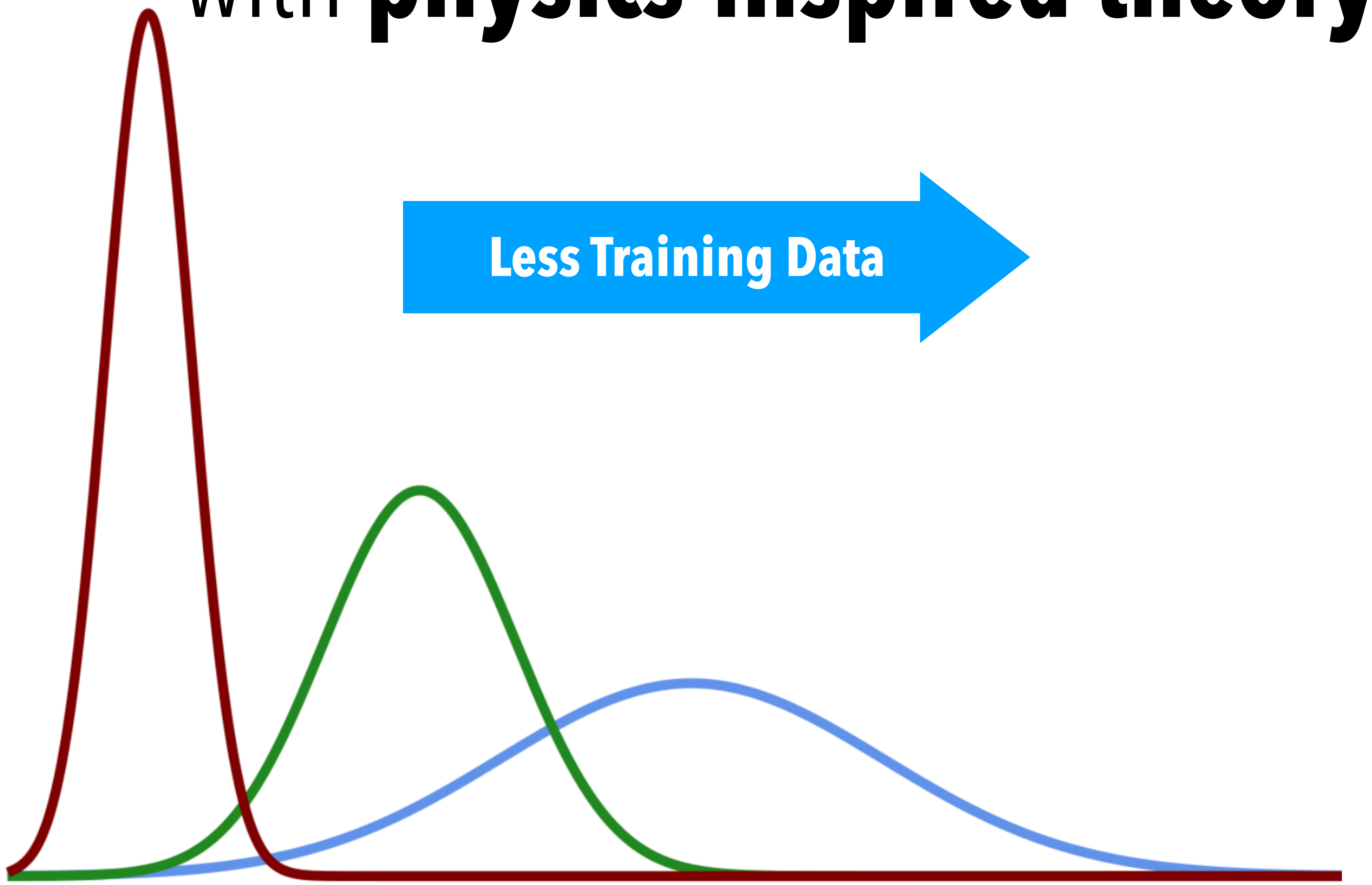


https://arxiv.org/abs/2001.08361

Neural Network Scaling Laws



Question: How does Ensemble Variance **scale** with Training Dataset Size? Can we predict this scaling with **physics-inspired theory**?



Physics-Inspired Theory: The NTK

NTK = Neural Tangent Kernel

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \left.\frac{\partial \mathcal{L}_A}{\partial \theta_\mu}\right|_{\theta_\mu = \theta_\mu(t)}$$

Taylor Expansion

$$\Delta \mathcal{L}_A = -\eta \sum_{\mu,\nu} \lambda_{\mu\nu} \frac{\partial \mathcal{L}_A}{\partial \theta_\mu} \frac{\partial \mathcal{L}_A}{\partial \theta_\nu} + O(\eta^2),$$

Chain Rule

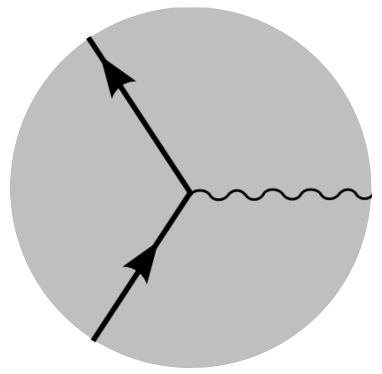
The Error Factor

The NTK!

$$\Delta \mathcal{L}_A = -\eta \sum_{i_1, i_2 = 1}^{n^{(L)}} \sum_{\tilde{\alpha}_1, \tilde{\alpha}_2 \in \mathcal{A}} \left[\frac{\partial \mathcal{L}_A}{\partial z_{i_1;\tilde{\alpha}_1}} \frac{\partial \mathcal{L}_A}{\partial z_{i_2;\tilde{\alpha}_2}} \right] \left[\sum_{\mu,\nu} \lambda_{\mu\nu} \frac{d z_{i_1;\tilde{\alpha}_1}}{d\theta_\mu} \frac{d z_{i_2;\tilde{\alpha}_2}}{d\theta_\nu} \right].$$

$$\frac{\partial \mathcal{L}_A}{\partial z_{i;\tilde{\alpha}}} = z_i(x_{\tilde{\alpha}}; \theta) - y_{i;\tilde{\alpha}}$$

Physics-Inspired Theory: The NTK



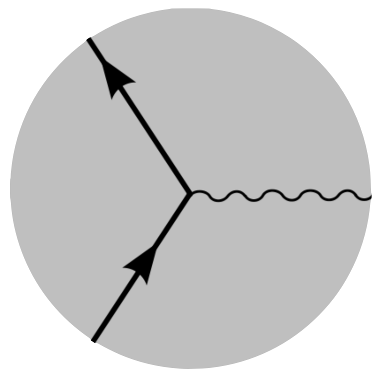
$$\Theta_{i_1 i_2;\tilde{\alpha}_1\tilde{\alpha}_2} \equiv \sum_{\mu,\nu} \lambda_{\mu\nu} \frac{dz_{i_1;\tilde{\alpha}_1}}{d\theta_\mu} \frac{dz_{i_2;\tilde{\alpha}_2}}{d\theta_\nu}$$

The NTK is the **main driver** of the **function-approximation** dynamics.*

Thus, by understanding the NTK and how it evolves we can directly understand the behavior of NNs

* for DNNs trained with full batch gradient descent

Physics-Inspired Theory: The NTK Perturbation



NTK dynamics are possible to understand **perturbatively and analytically**

$$\Theta^{(\ell)}_{\alpha_1\alpha_2} = \Theta^{\{0\}(\ell)}_{\alpha_1\alpha_2} + \underbrace{\frac{1}{n_\ell}\Theta^{\{1\}(\ell)}_{\alpha_1\alpha_2} + \frac{1}{n_\ell^2}\Theta^{\{2\}(\ell)}_{\alpha_1\alpha_2} + O\left(\frac{1}{n_\ell^3}\right)}$$

Infinte Width

Finite Width Corrections

Perturbing in the **width of the network**

Physics-Inspired Theory: RG-Flow of the NTK

How to compute the infinite width NTK? –> RG Flow

Recursively Repeat through each layer until **final layer** action is computed! This is **RG Flow in an DNN.**

$$\Theta^{\{0\}(\ell)}_{\alpha_1\alpha_2} \longrightarrow e^{-S_{eff}(z^{(\ell)})}$$

$$S_{full}(z^{(1)}, \ldots, z^{(L)}) = \sum_{\ell=1}^{L} S_M(z^{(\ell)}) + \sum_{\ell=1}^{L-1} S_I(z^{(\ell+1)} \,|\, z^{(\ell)})$$

Compute **effective action of a layer** ℓ by **marginalizing** over all pre-activations in layer $\ell - 1$

$$e^{-S_{eff}(z^{(L)})} = \int \left[\prod_{\ell=1}^{L-1} dz^{(\ell)}\right] e^{-S_{full}(z^{(1)},\ldots,z^{(L)})}$$

Physics-Inspired Theory: Infinite Width Predictions

With the Final Layer NTK, we can compute the **"end of training"** prediction:

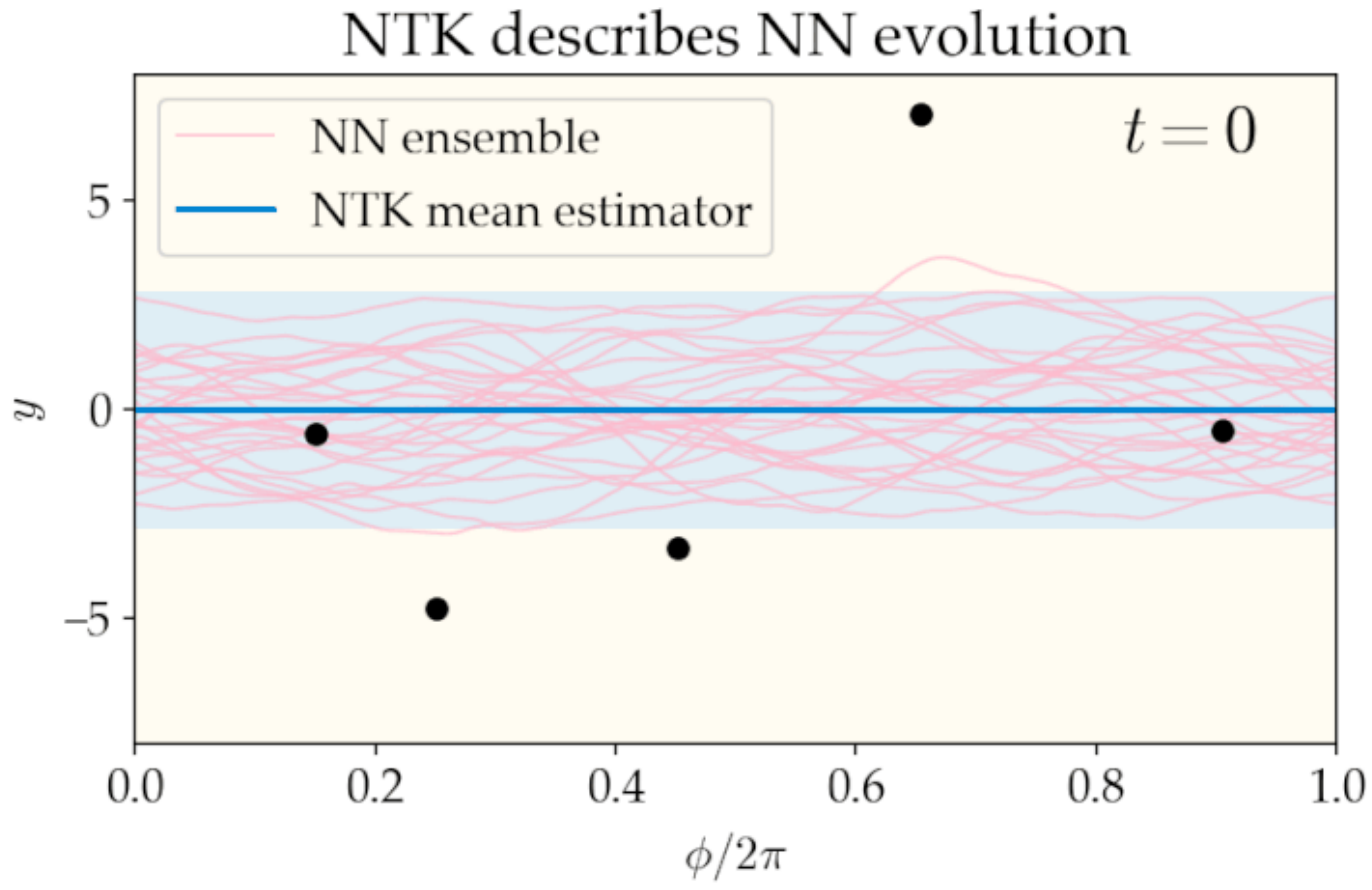
$$m^{\infty}_{i;\beta} \equiv \mathbb{E}\left[z^{(L)}_{i;\beta}(T)\right] = \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2 \in \mathcal{A}} \Theta^{(L)}_{\beta\tilde{\alpha}_1}(\Theta^{(L)}_{\tilde{\alpha}_1\tilde{\alpha}_2})^{-1} y_{i;\tilde{\alpha}_2}$$

And the **variance on that prediction:**

$$Cov\left[z^{(L)}_{i_1;\beta_1}(T), z^{(L)}_{i_2;\beta_2}(T)\right] = \mathbb{E}\left[z^{(L)}_{i_1;\beta_1}(T) z^{(L)}_{i_2;\beta_2}(T)\right] - m^{\infty}_{i_1;\beta_1} m^{\infty}_{i_2;\beta_2}$$

$$= \delta_{i_1 i_2} K^{(L)}_{\beta_1\beta_2} - \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2 \in \mathcal{A}} \Theta^{(L)}_{\beta_1\tilde{\alpha}_1} K^{(L)}_{\tilde{\alpha}_1\tilde{\alpha}_2} \Theta^{(L)}_{\tilde{\alpha}_2\beta_2} - \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2 \in \mathcal{A}} \Theta^{(L)}_{\beta_2\tilde{\alpha}_1} K^{(L)}_{\tilde{\alpha}_1\tilde{\alpha}_2} \Theta^{(L)}_{\tilde{\alpha}_2\beta_1} + \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4 \in \mathcal{A}} \Theta^{(L)}_{\beta_1\tilde{\alpha}_1} \Theta^{(L)}_{\tilde{\alpha}_1\tilde{\alpha}_2} K^{(L)}_{\tilde{\alpha}_2\tilde{\alpha}_3} \Theta^{(L)}_{\tilde{\alpha}_3\tilde{\alpha}_4} \Theta^{(L)}_{\tilde{\alpha}_4\beta_2}.$$

Physics-Inspired Theory: Infinite Width Predictions



https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf

04

Empirical Results

Empirical Results Comparing Theoretical Results with Empirical Reality

Empirical Results: Infinite Width Predictions

Are infinite width calculations **predictive** on **real Machine Learning** problems?

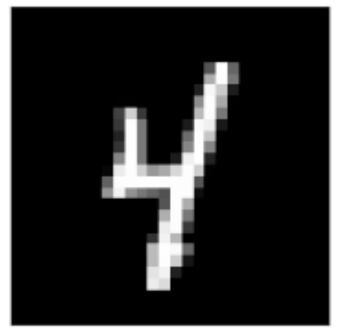
We test this by computing:

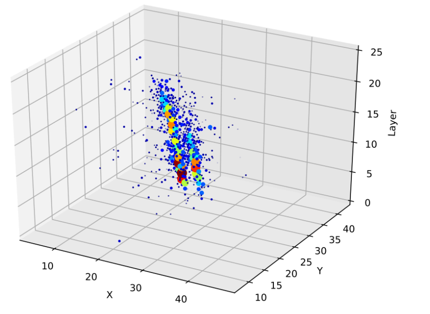
1) Infinte Width Prediection
2) A Trained Ensemble of DNNs (~150 networks)
   - Width-30, Early Stopping, Full Batch Gradient Decent

For a range of training set sizes

On three datasets:

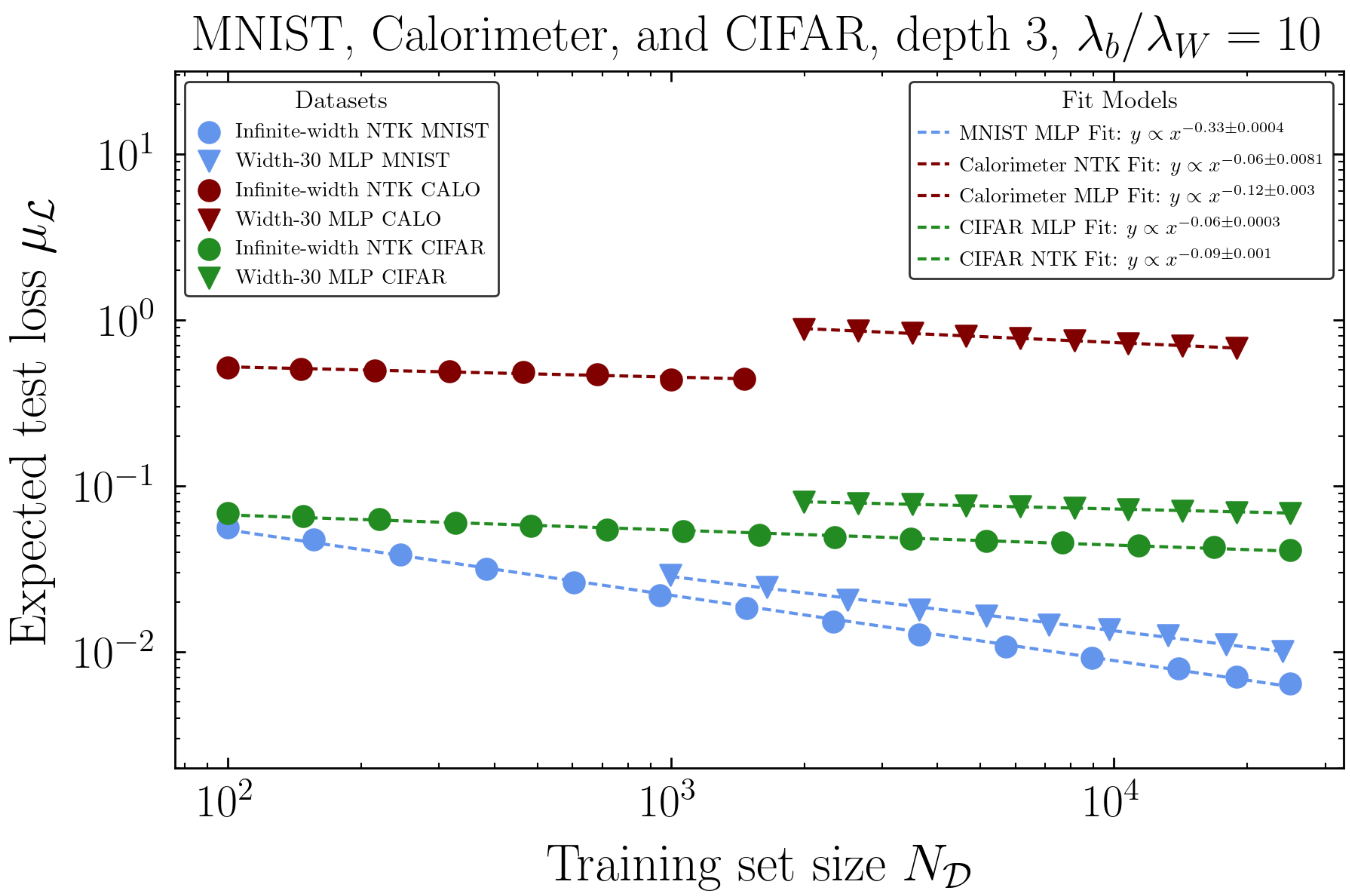
1) MNIST Image Classification
2) CIFAR Image Classification
3) A HEP Calorimeter Energy Regression Problem





https://link.springer.com/article/10.1140/epjc/s10052-020-8251-9

Empirical Results: Infinite Width Prediction Loss



The **mean test loss** for a trained ensemble of DNNs and Infinite Width Networks **for three datasets**

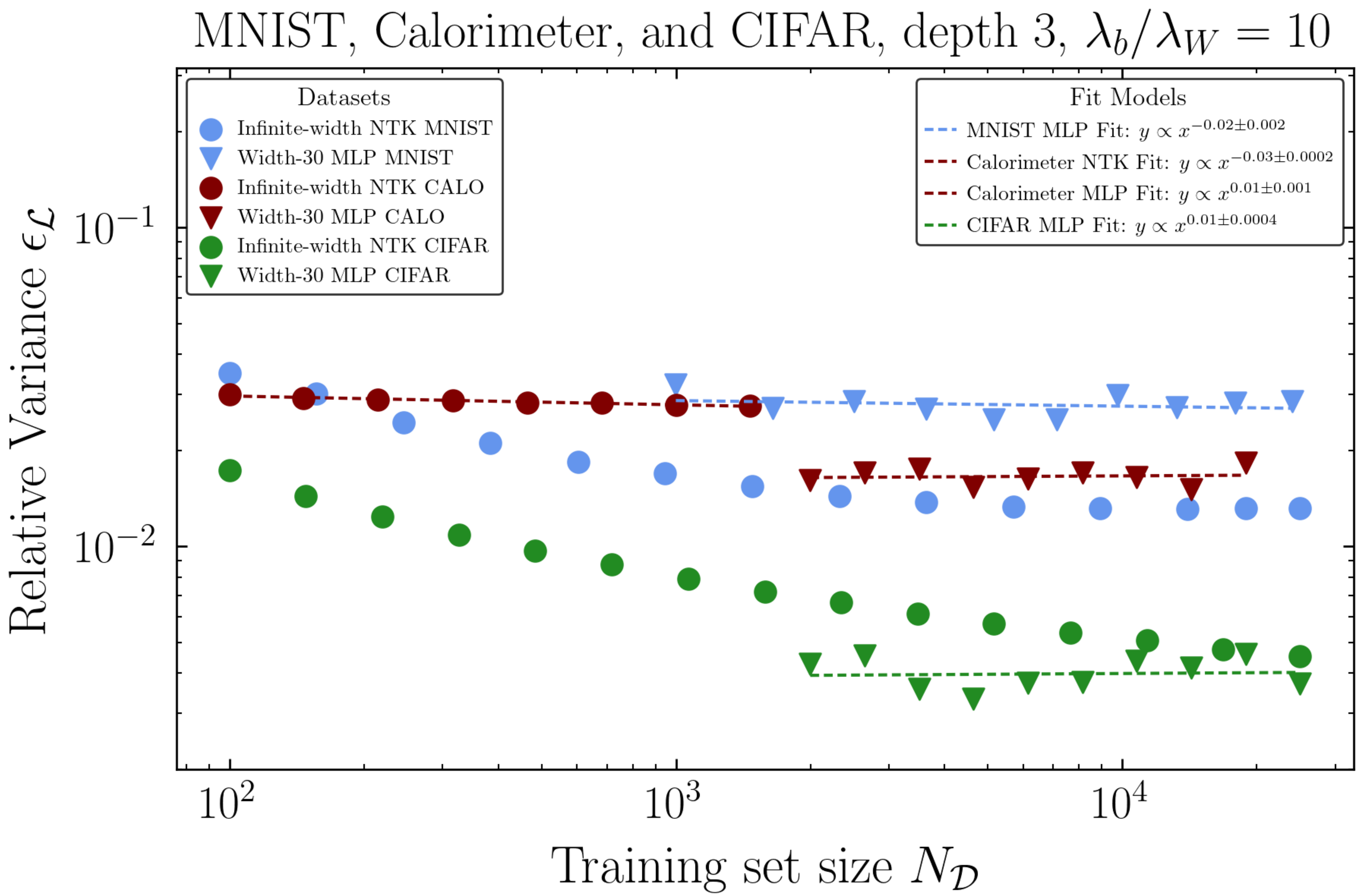
Empirical Results: Infinite Width Coefficient of Variation

Suppose we introduce the **Coefficient of Variation:**

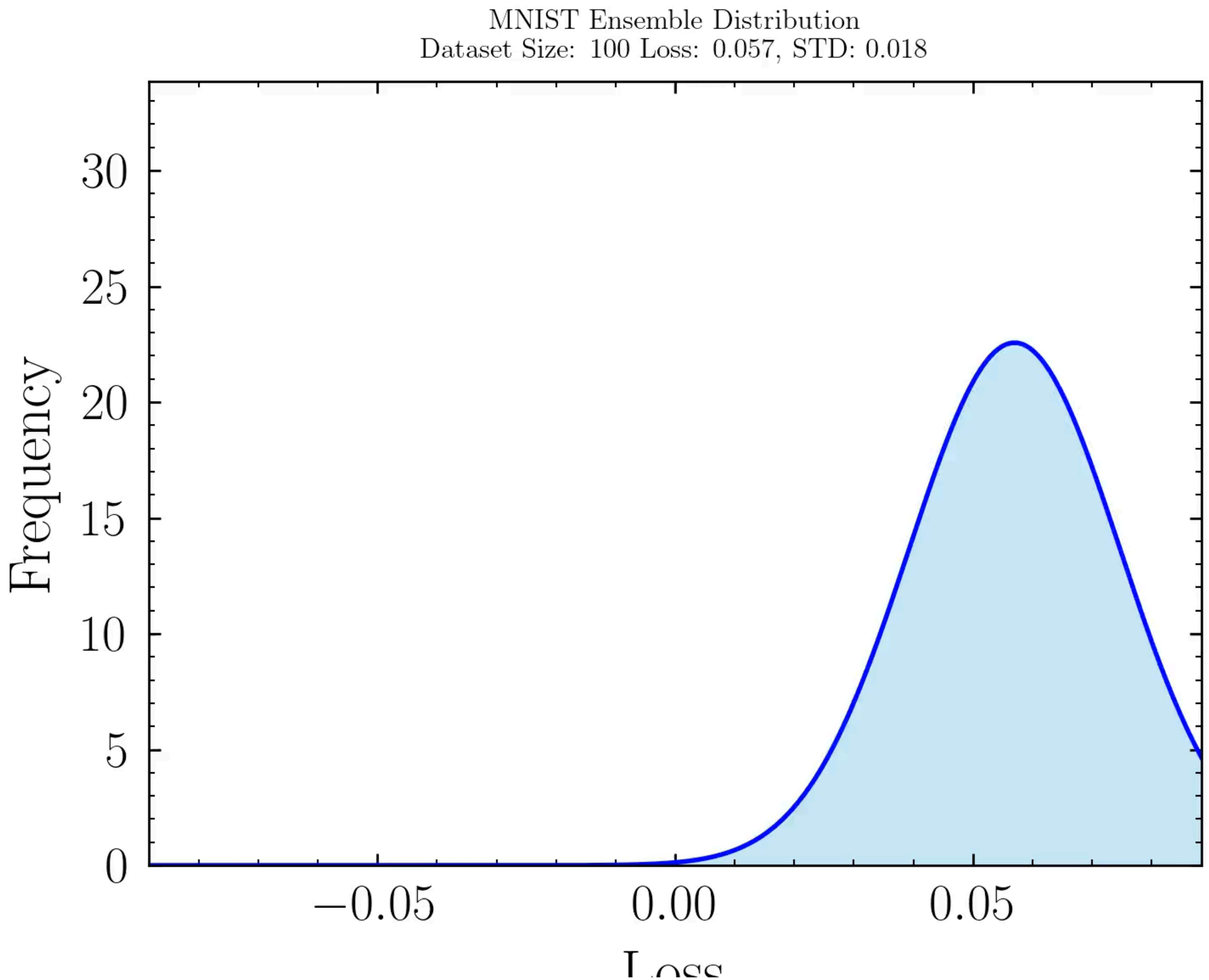
$$\epsilon_{\mathscr{L}} \equiv \frac{\sigma_{\mathscr{L}}}{\mu_{\mathscr{L}}}$$

We find:

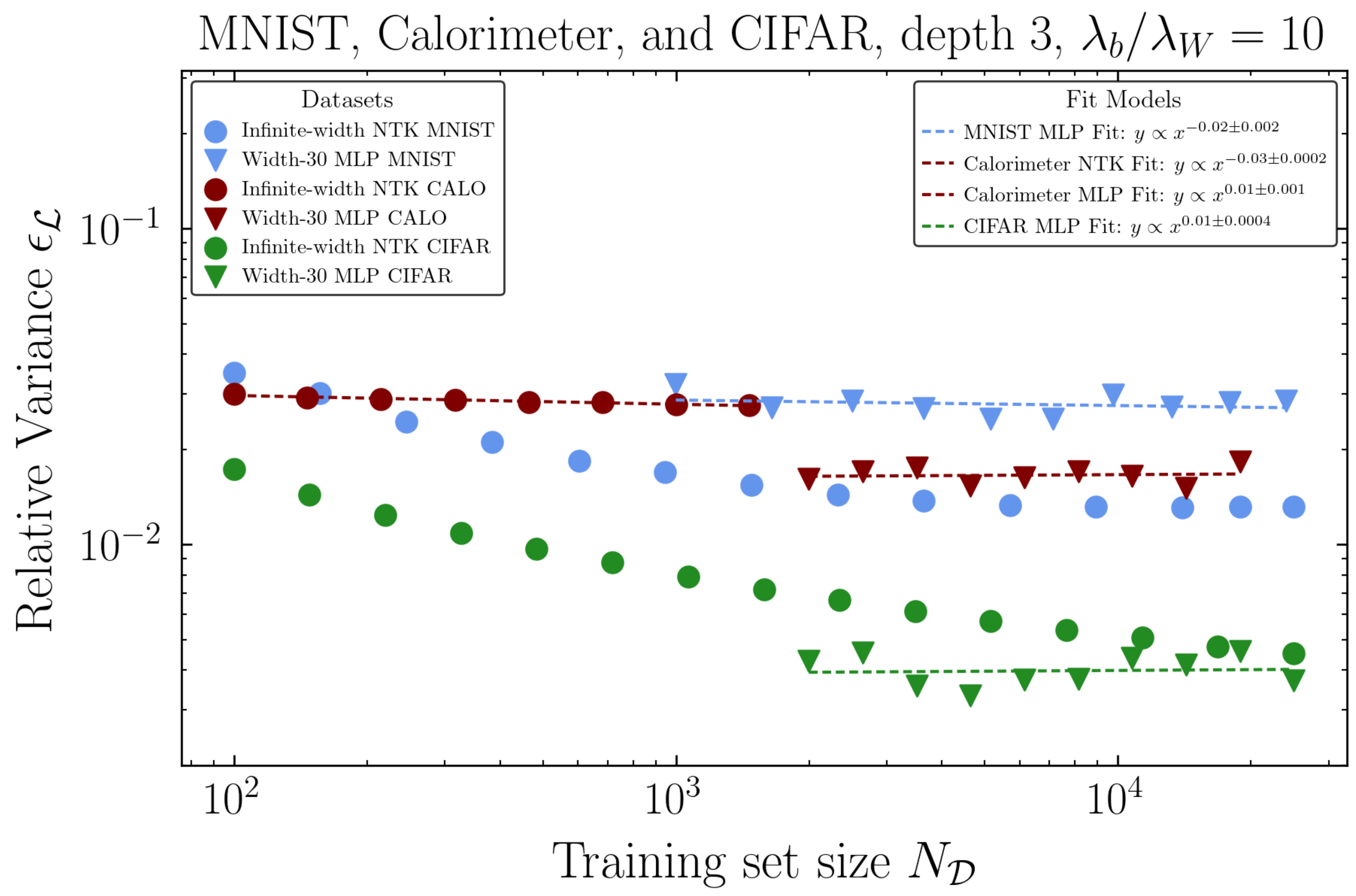
1) DNN $\epsilon_{\mathscr{L}}$ flat with dataset size!
2) Infinite width $\epsilon_{\mathscr{L}}$ asymptotes flat.



Empirical Results: Infinite Width Prediction Loss



Conclusion: Implications



Implications of our work:

1. We **find that** $\epsilon_{\mathscr{L}}$ **is small** ($\mu_{\mathscr{L}} > \sigma_{\mathscr{L}}$)
   - We can assign $\mu_{\mathscr{L}}$ as the systematic uncertainty due to the DNN

2. $\epsilon_{\mathscr{L}}$ **is flat *and similar to the infinite width value***, thus one can estimate $\epsilon_{\mathscr{L}}$ by either:
   - **Training an ensemble for small** $N_{\mathscr{D}}$ (cheap) and extrapolate $\epsilon_{\mathscr{L}}$ value to larger $N_{\mathscr{D}}$
   - Compute Infinite Width Value after $N_{\mathscr{D}}$ asymptotes (very cheap)

QUESTIONS?