Combining ID & AD: simultaneous particle identification and anomaly detection at collider experiments using Bayesian neural networks

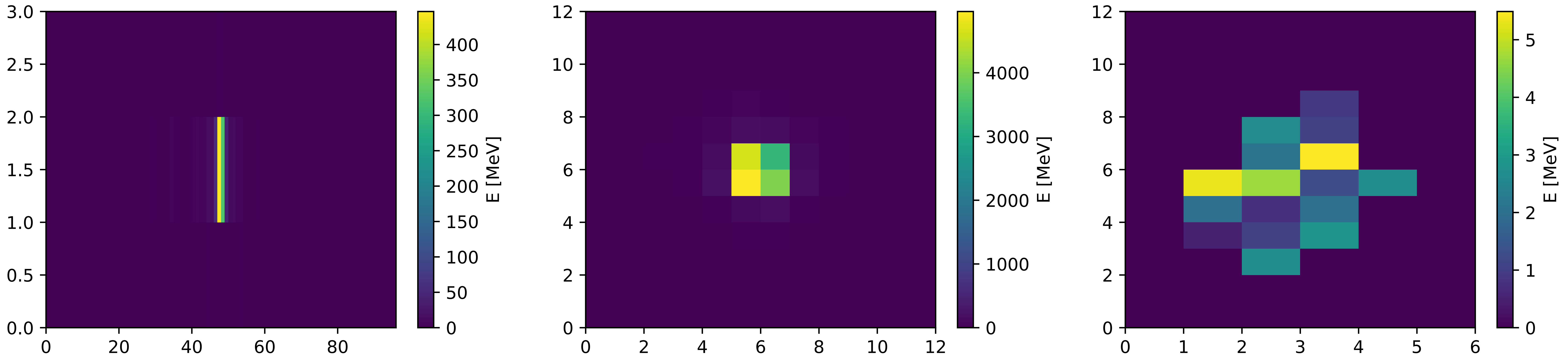
preliminary

Wiehl 2022

Johannes Erdmann, Burim Ramosaj, **Daniel Wall**

Outline

- Bayesian vs. Deterministic NNs
- Our example use case:
  - Multiclass classification of EM calorimeter images
- Classification Performance
- Anomaly Detection
- Conclusions



Bayesian Neural Networks

- Basic idea: learn the whole distribution over the NN weights w given training data {X,Y}
- *Which parameters w of a function f are likely to have generated Y from X ?*

$$p(w|X,Y) = \frac{p(X,Y|w)p(w)}{p(X,Y)} = \frac{p(Y|X,w)\cancel{p(X|w)}p(w)}{p(Y|X)\cancel{p(X)}} = \frac{p(Y|X,w)p(w)}{\int p(Y|X,w)p(w)\mathrm{d}w}$$

with for example $p\big(y=d|\mathbf{x},\boldsymbol{\omega}\big) = \frac{\exp(f_d^{\boldsymbol{\omega}}(\mathbf{x}))}{\sum_{d'}\exp(f_{d'}^{\boldsymbol{\omega}}(\mathbf{x}))}$

- Often intractable → approximate with simpler function $q_\theta(w)$ minimizing KL divergence

$$\mathrm{KL}(q_\theta(\boldsymbol{\omega}) \mid\mid p(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})) = \int q_\theta(\boldsymbol{\omega}) \log \frac{q_\theta(\boldsymbol{\omega})}{p(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})}\mathrm{d}\boldsymbol{\omega} = -\int q_\theta(\boldsymbol{\omega}) \log p(\mathbf{Y}|\mathbf{X},\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega} + \mathrm{KL}(q_\theta(\boldsymbol{\omega})||p(\boldsymbol{\omega})) + \mathrm{const}$$

- Sampling from optimal $q_\theta^*(w)$ → distribution of predictions instead of a point estimate

$$p(\mathbf{y}^*|\mathbf{x}^*,\mathbf{X},\mathbf{Y}) \approx \int p(\mathbf{y}^*|\mathbf{x}^*,\boldsymbol{\omega})q_\theta^*(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}$$

Description from Yarin Gal, Uncertainty in Deep Learning, PhD thesis, Cambridge, 2016 and arXiv:0712.4042.

Bayesian Neural Networks

- Basic idea: learn the whole distribution over the NN weights w given training data {X,Y}
- *Which parameters w of a function f are likely to have generated Y from X ?*

$$p(w|X,Y) = \frac{p(X,Y|w)p(w)}{p(X,Y)} = \frac{p(Y|X,w)\cancel{p(X|w)}p(w)}{p(Y|X)\cancel{p(X)}} = \frac{p(Y|X,w)p(w)}{\int p(Y|X,w)p(w)\mathrm{d}w}$$

with for example $p(y=d|\mathbf{x},\boldsymbol{\omega}) = \frac{\exp(f_d^{\boldsymbol{\omega}}(\mathbf{x}))}{\sum_{d'}\exp(f_{d'}^{\boldsymbol{\omega}}(\mathbf{x}))}$

- Often intractable → approximate with simpler function $q_\theta(w)$ minimizing KL divergence

$$\mathrm{KL}(q_\theta(\boldsymbol{\omega}) \,||\, p(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})) = \int q_\theta(\boldsymbol{\omega}) \log \frac{q_\theta(\boldsymbol{\omega})}{p(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})}\mathrm{d}\boldsymbol{\omega} = -\int q_\theta(\boldsymbol{\omega}) \log p(\mathbf{Y}|\mathbf{X},\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega} + \mathrm{KL}(q_\theta(\boldsymbol{\omega})||p(\boldsymbol{\omega})) + \mathrm{const}$$

- Sampling from optimal $q_\theta^*(w)$ → distribution of predictions instead of a point estimate

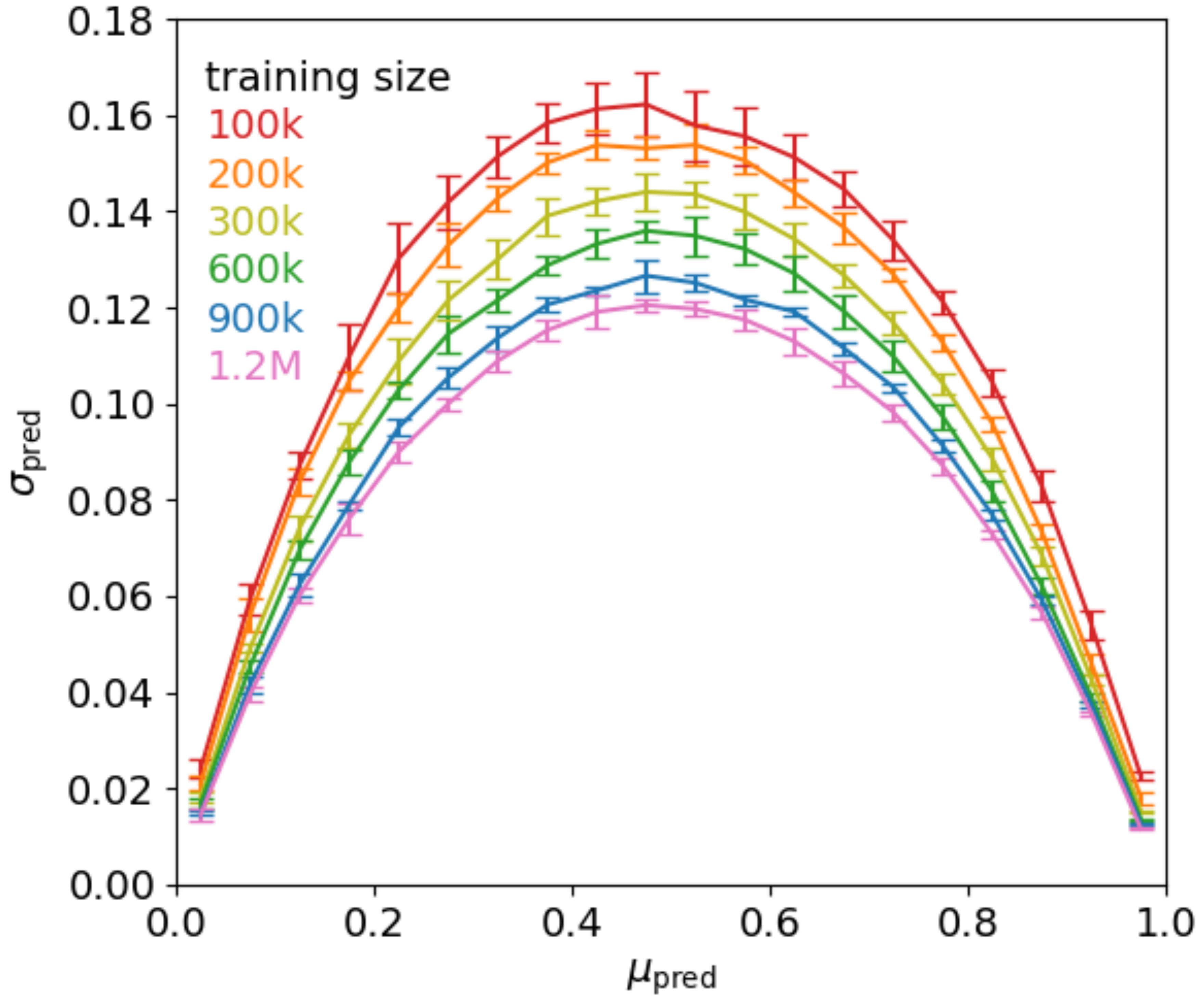
$$p(\mathbf{y}^*|\mathbf{x}^*,\mathbf{X},\mathbf{Y}) \approx \int p(\mathbf{y}^*|\mathbf{x}^*,\boldsymbol{\omega})q_\theta^*(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}$$

Description from Yarin Gal, Uncertainty in Deep Learning, PhD thesis, Cambridge, 2016 and arXiv:0712.4042.

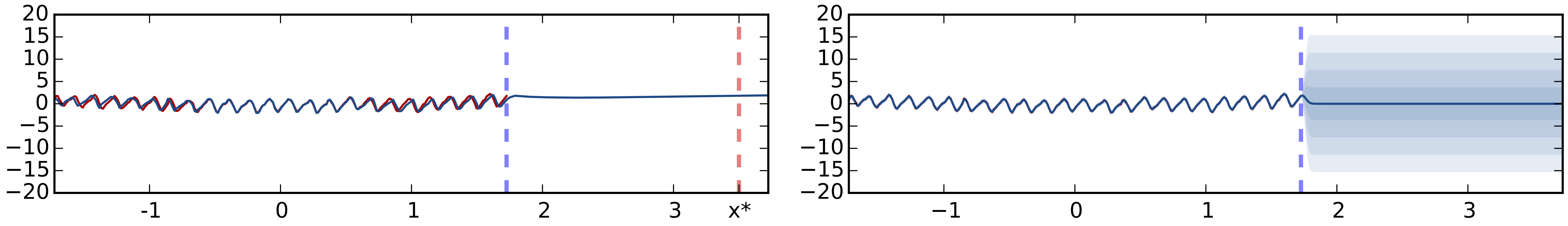
Bayesian Neural Networks

- Small training datasets lead to larger uncertainties in the BNN predictions (example from top-tagging in 1904.10004)

- BNNs predictions for out-of-distribution test samples can have large uncertainties



1904.10004

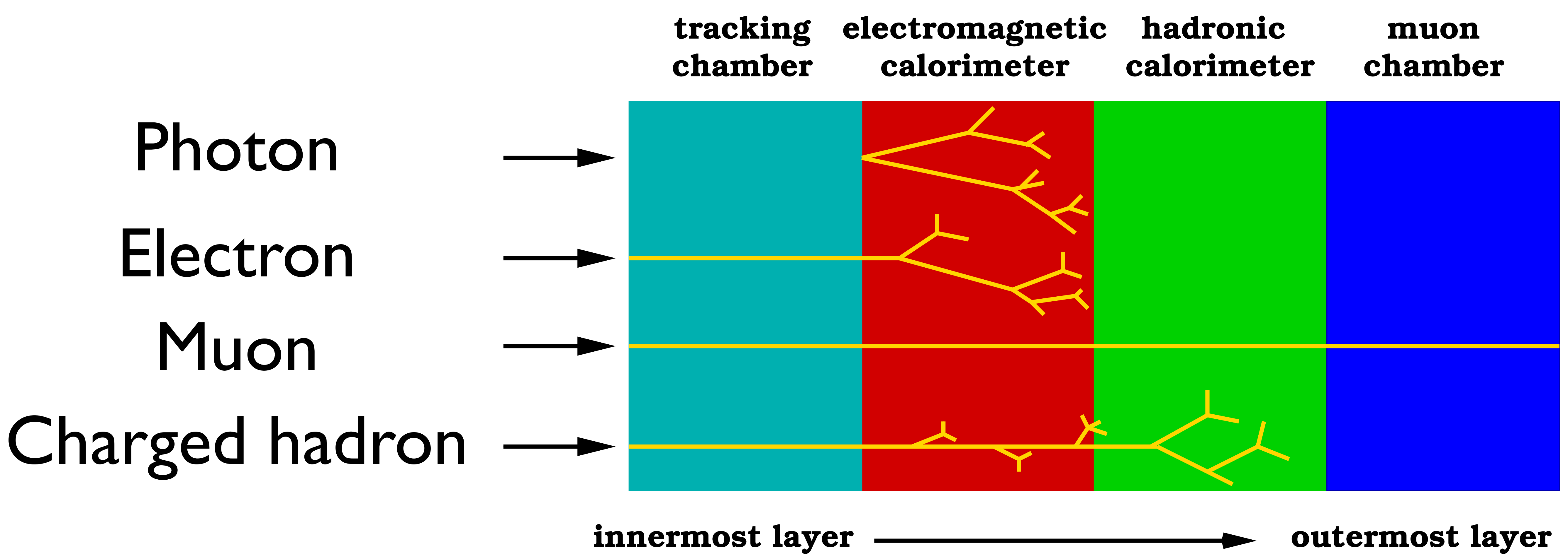


(a) Standard deep learning model (b) Probabilistic model

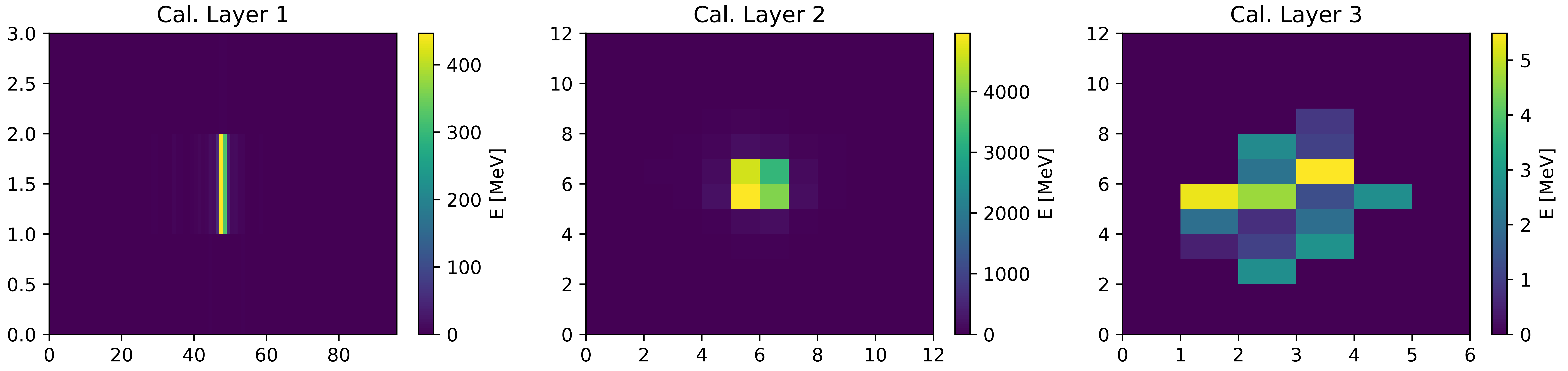
Y. Gal, PhD thesis, Cambridge

Our Example Use Case

- Classification of images in EM calorimeters = photon identification
- Main background:
  - High-energy $\pi^0 \rightarrow \gamma\gamma$
- Toy EM calorimeter à la 1712.10321
  - ATLAS-like: 3 layers of LAr+Pb
  - 1 m from Geant4 particle gun

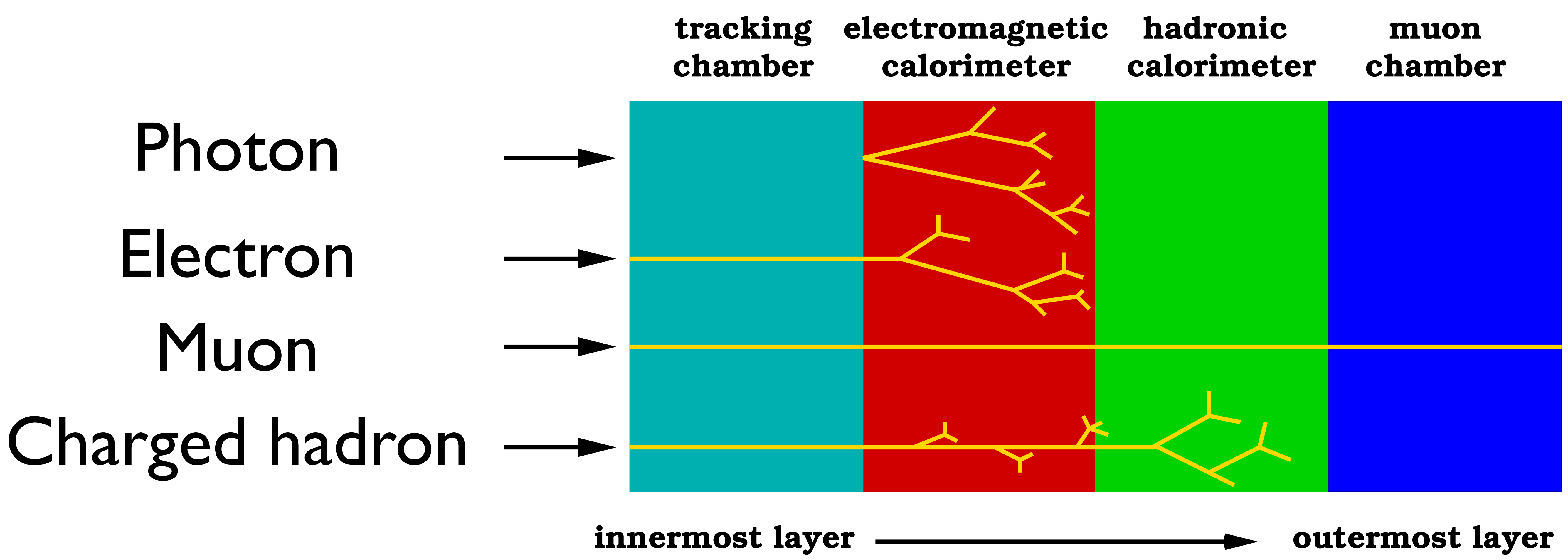


Photon

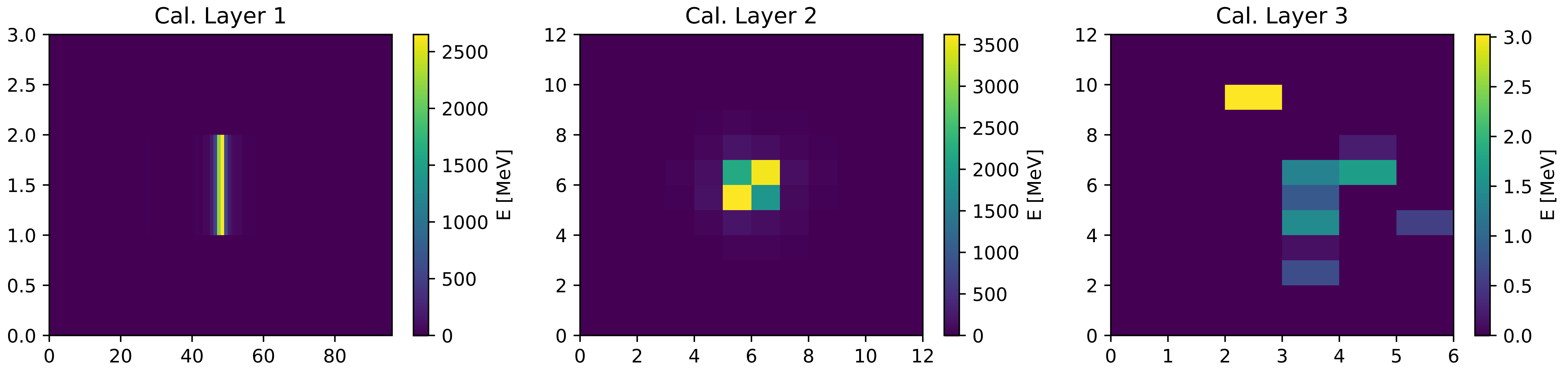


Our Example Use Case

- Classification of images in EM calorimeters = photon identification
- Main background:
  - High-energy $\pi^0 \to \gamma\gamma$
- Toy EM calorimeter à la 1712.10321
  - ATLAS-like: 3 layers of LAr+Pb
  - 1 m from Geant4 particle gun

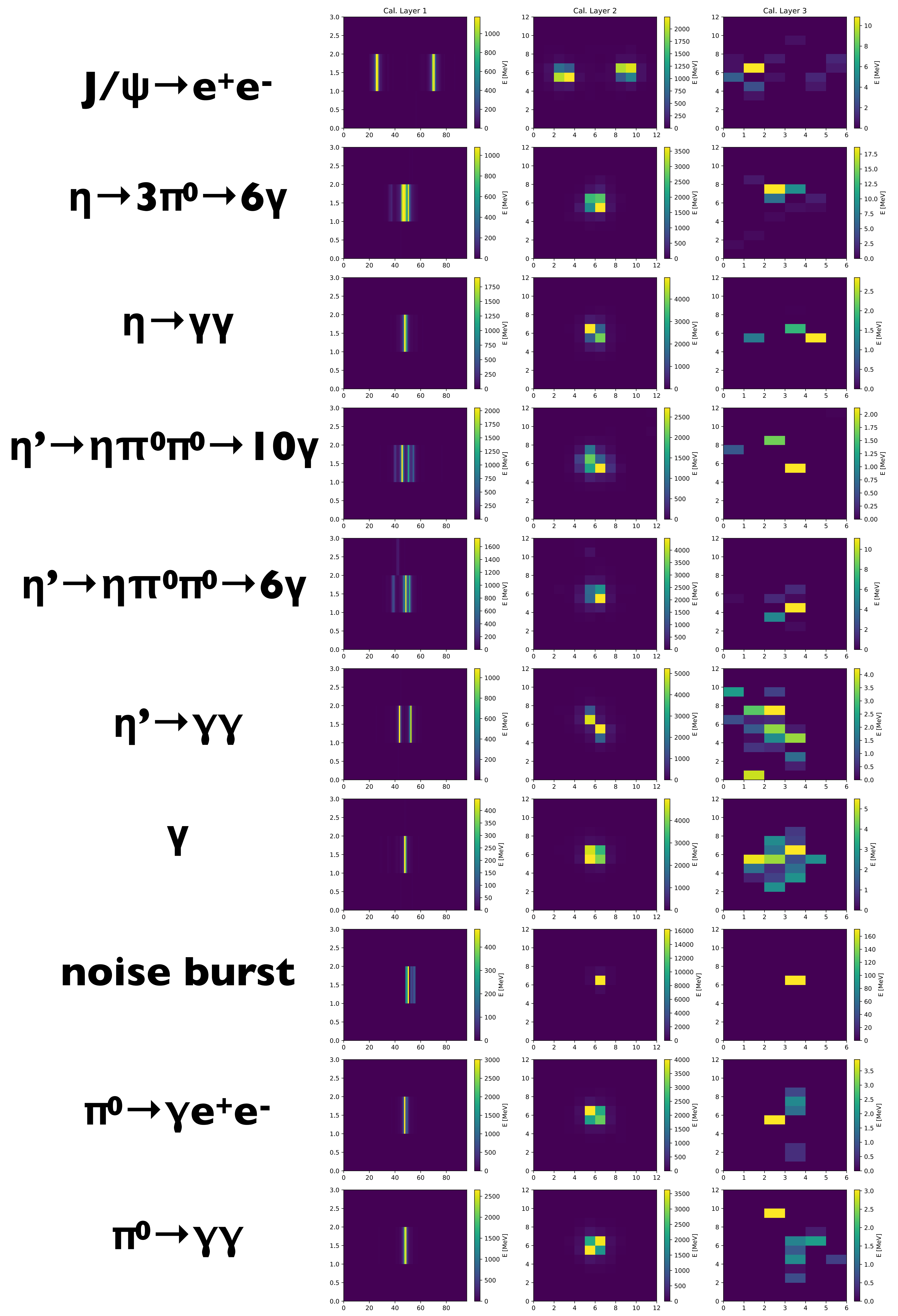


$\pi^0 \to \gamma\gamma$



Our Example Use Case

- Particle gun kinetic energy: 20 GeV
- Single photon (signal)
- 8 physical background classes:
  - Different purely EM decays of mesons with different masses → different opening angles
- + noise background class (noise burst in 2nd layer with 1% cross talk to neighbouring cells)
- Photon ID algorithm at LHC would be trained on effective mixture via parton shower programs



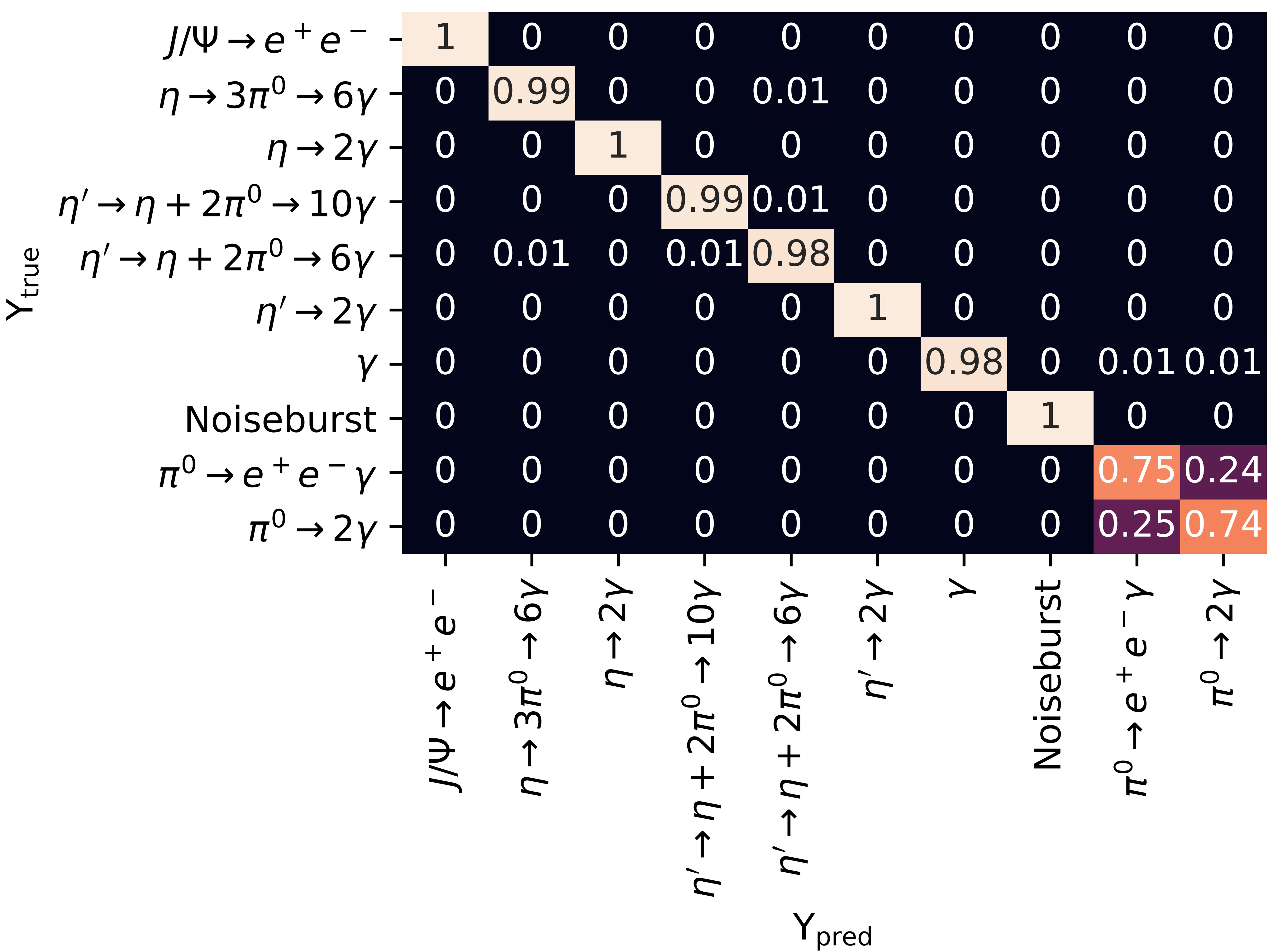
Classification Performance

- Setup: 2D CNN with 8 filters and 3x3 kernel size for each calorimeter layer

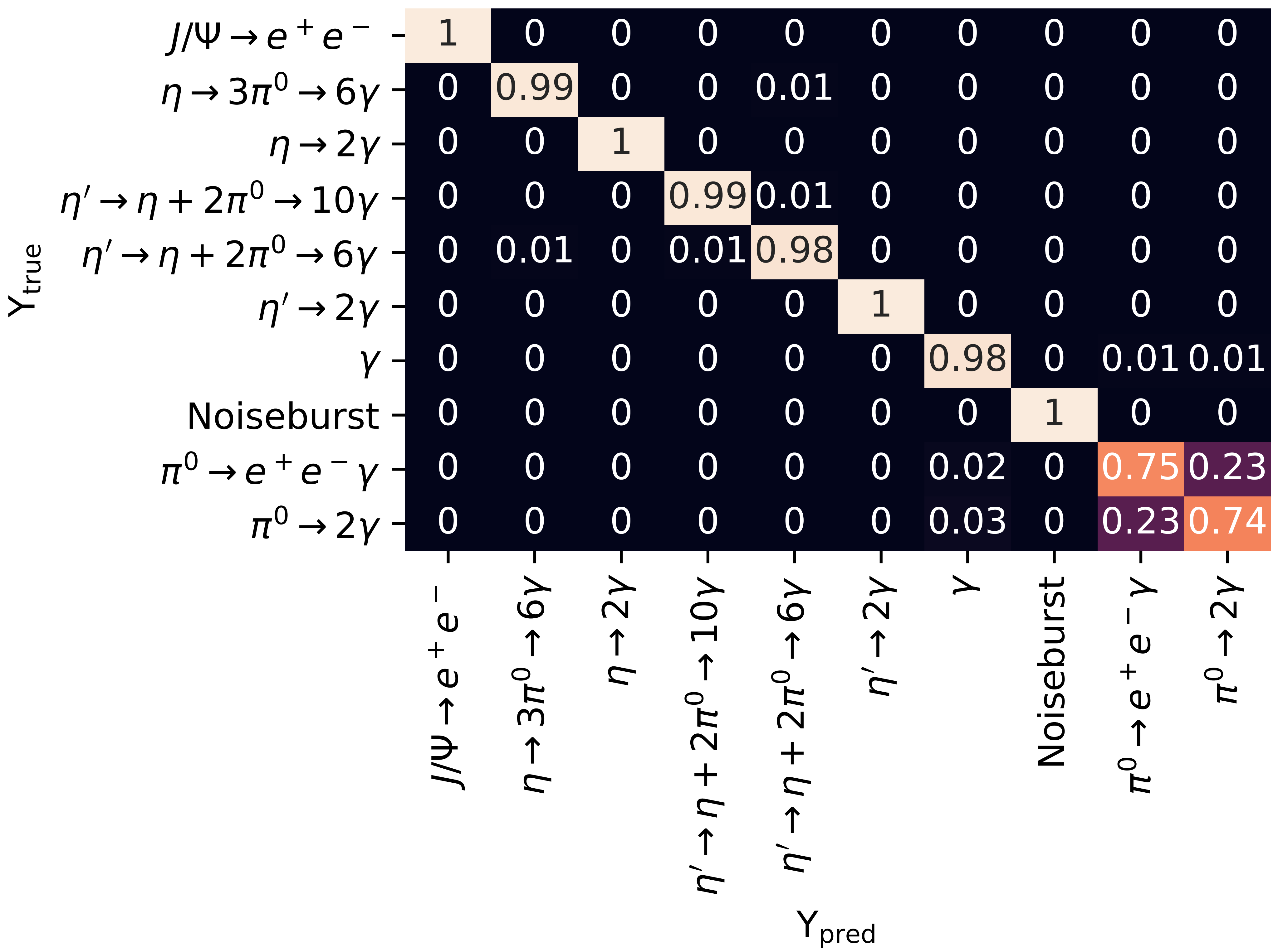
  + 10 output nodes all using Flipout (1803.04386)

- Assume weights to be Gaussian distributed, uncorrelated and with Gaussian priors
- Very similar performance to deterministic NN

Deterministic NN



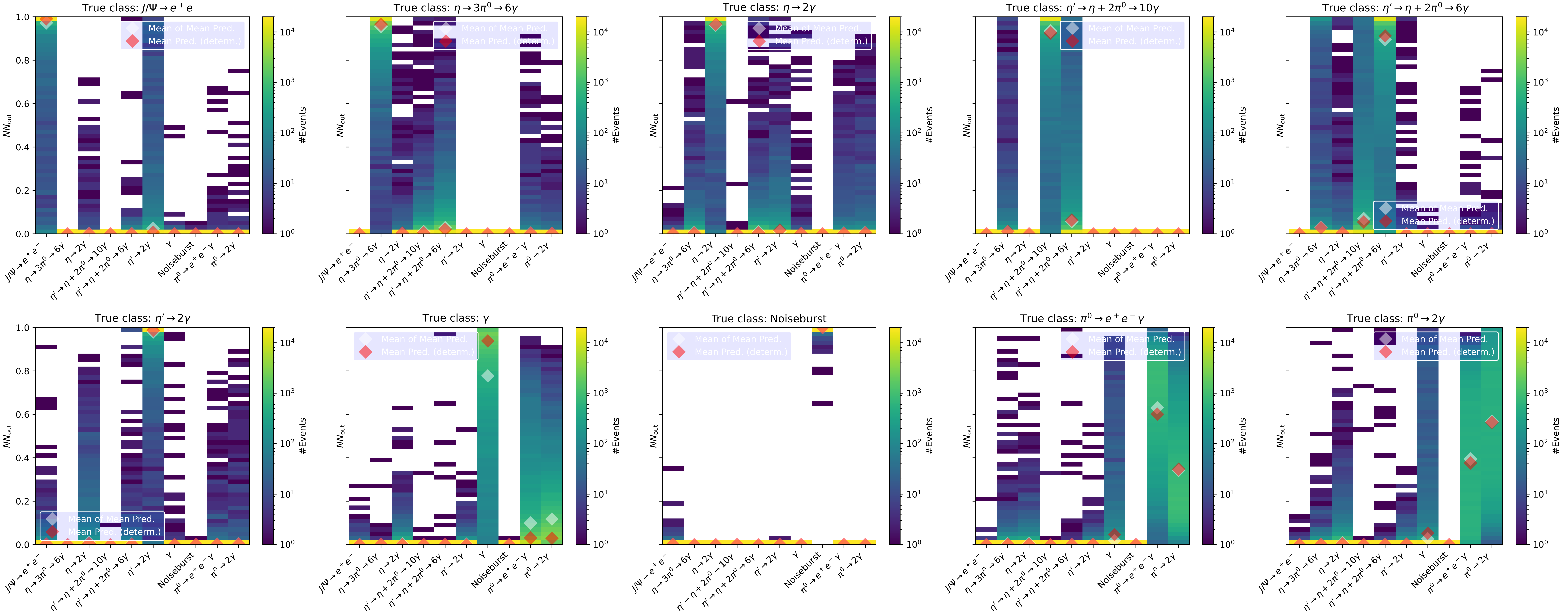
Bayesian NN



Results

- For each image, sample from the weight distributions → mean & variance per image

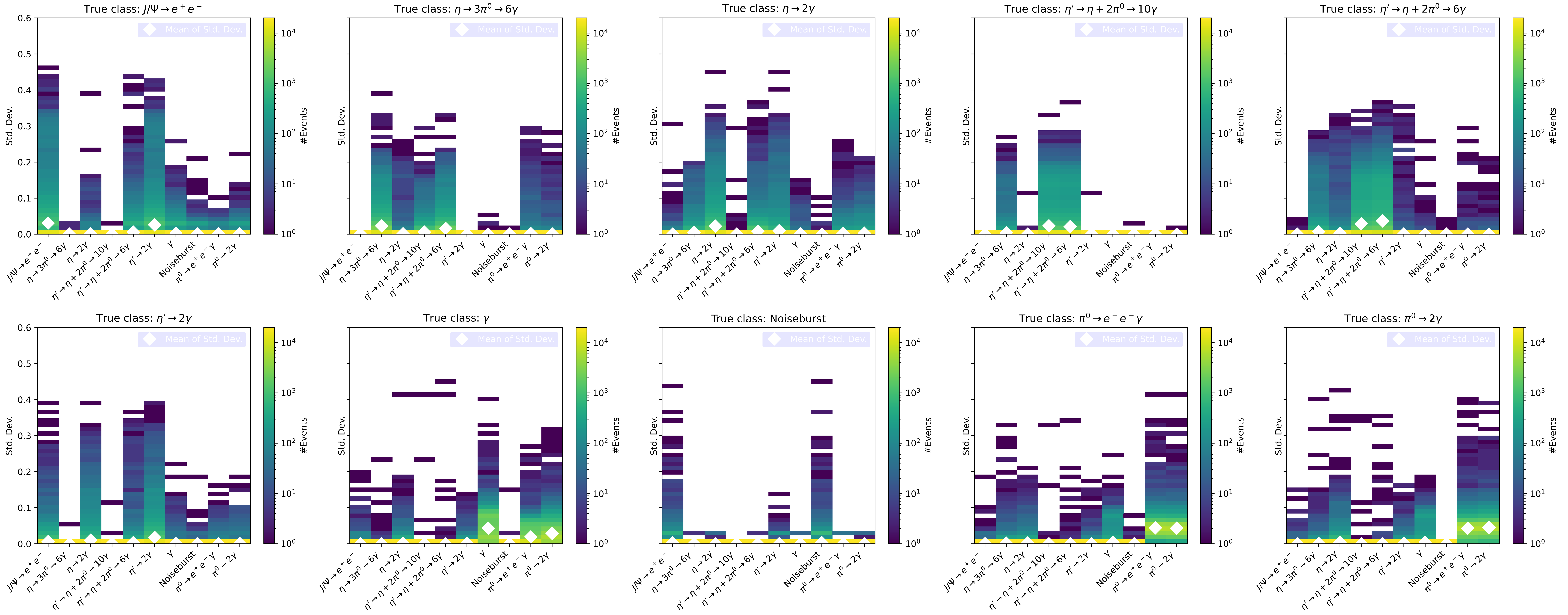
Mean



Results

- For each image, sample from the weight distributions → mean & variance per image

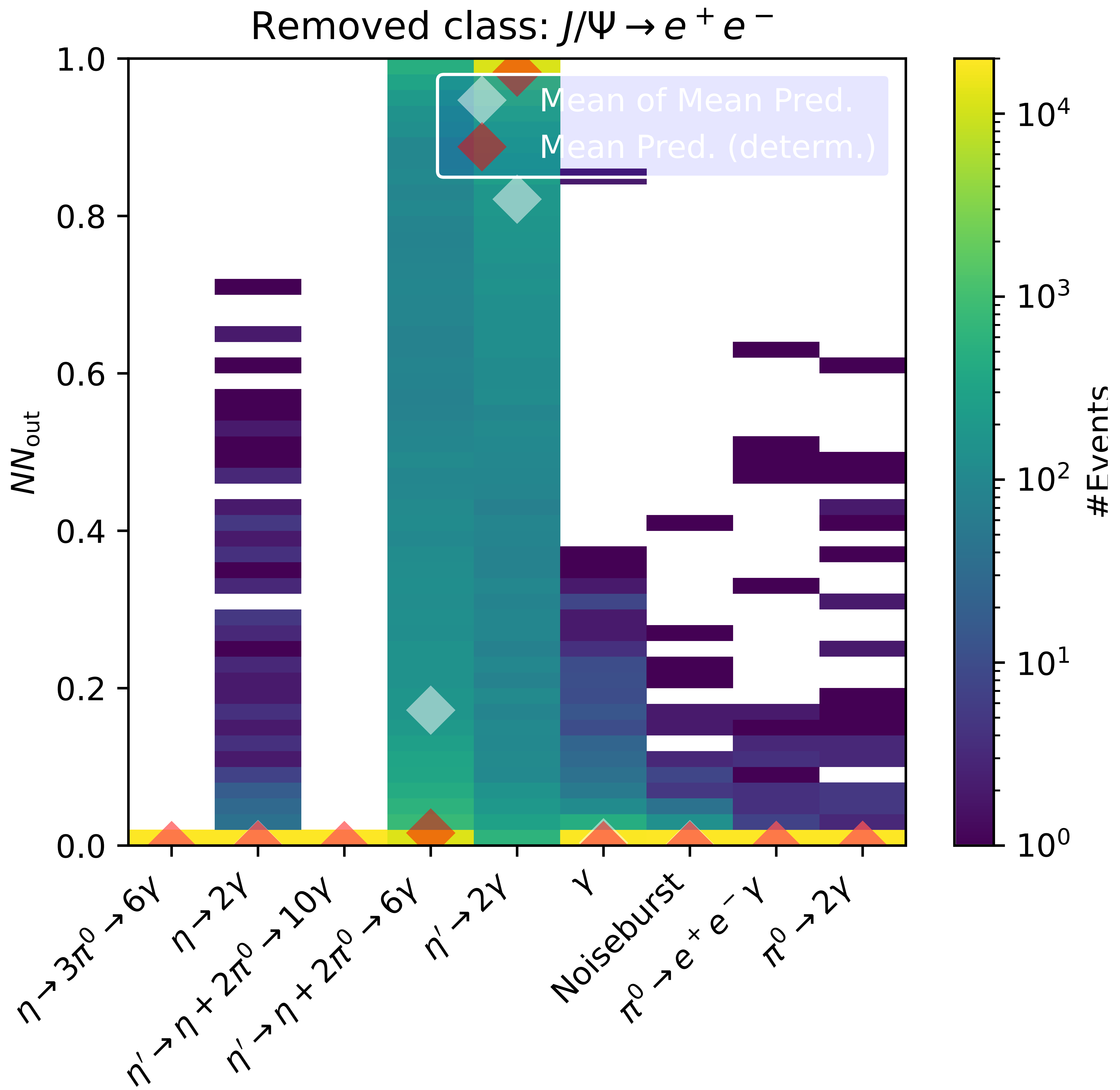
Std. Dev.



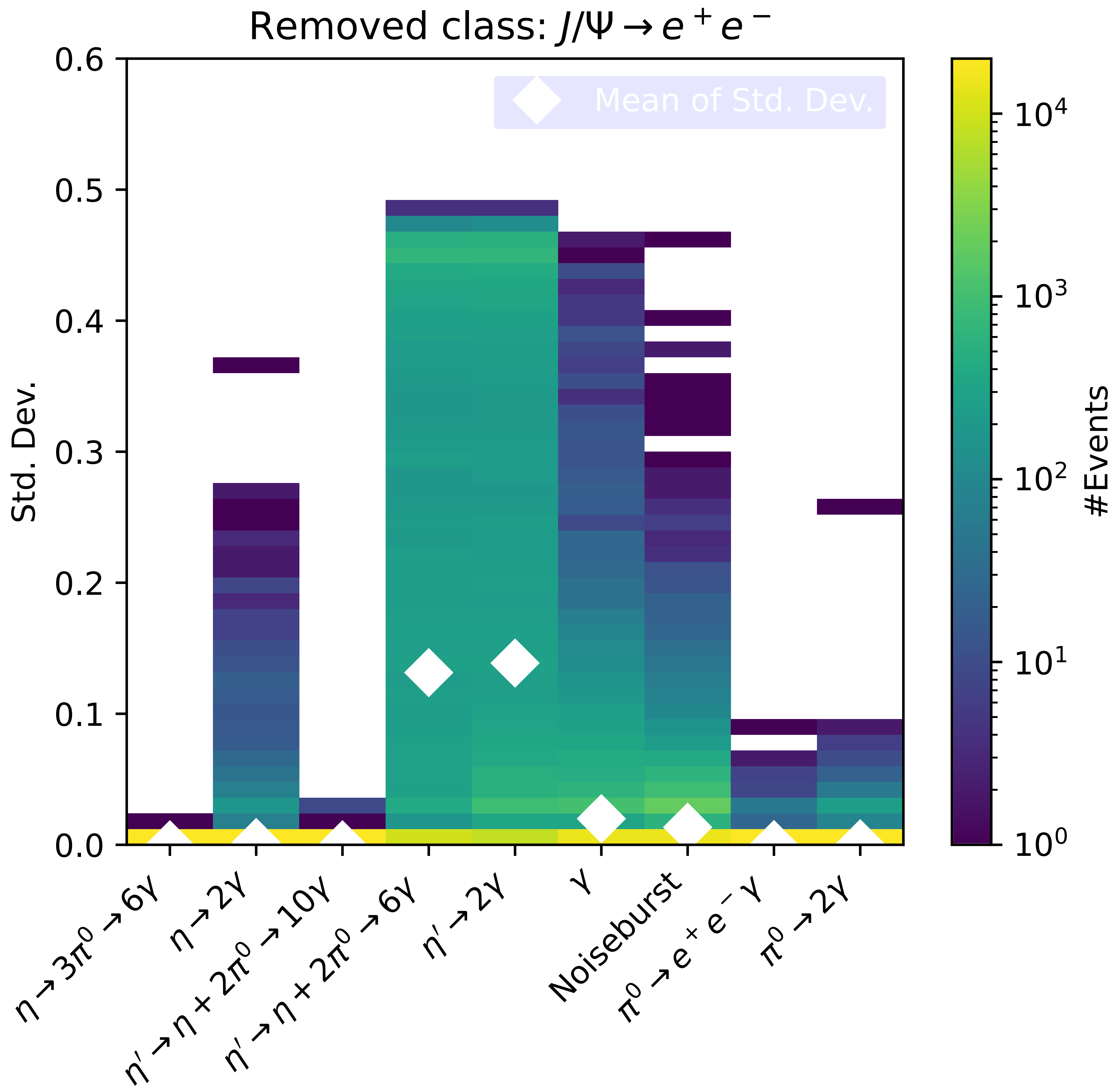
Results

- Now: remove one class at a time during training = anomaly (here: $J/\psi \rightarrow e^+e^-$)

Mean



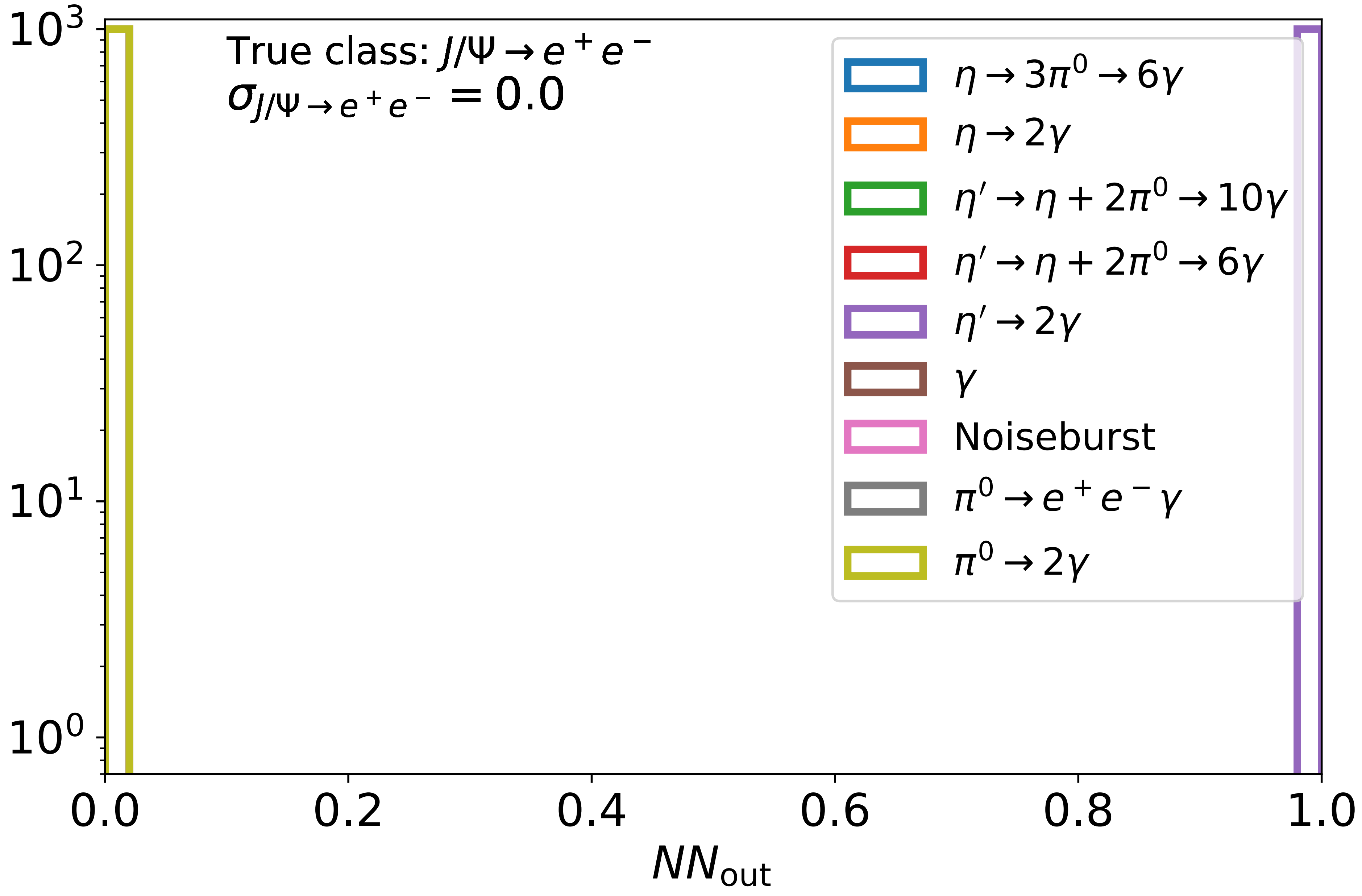
Std. Dev.



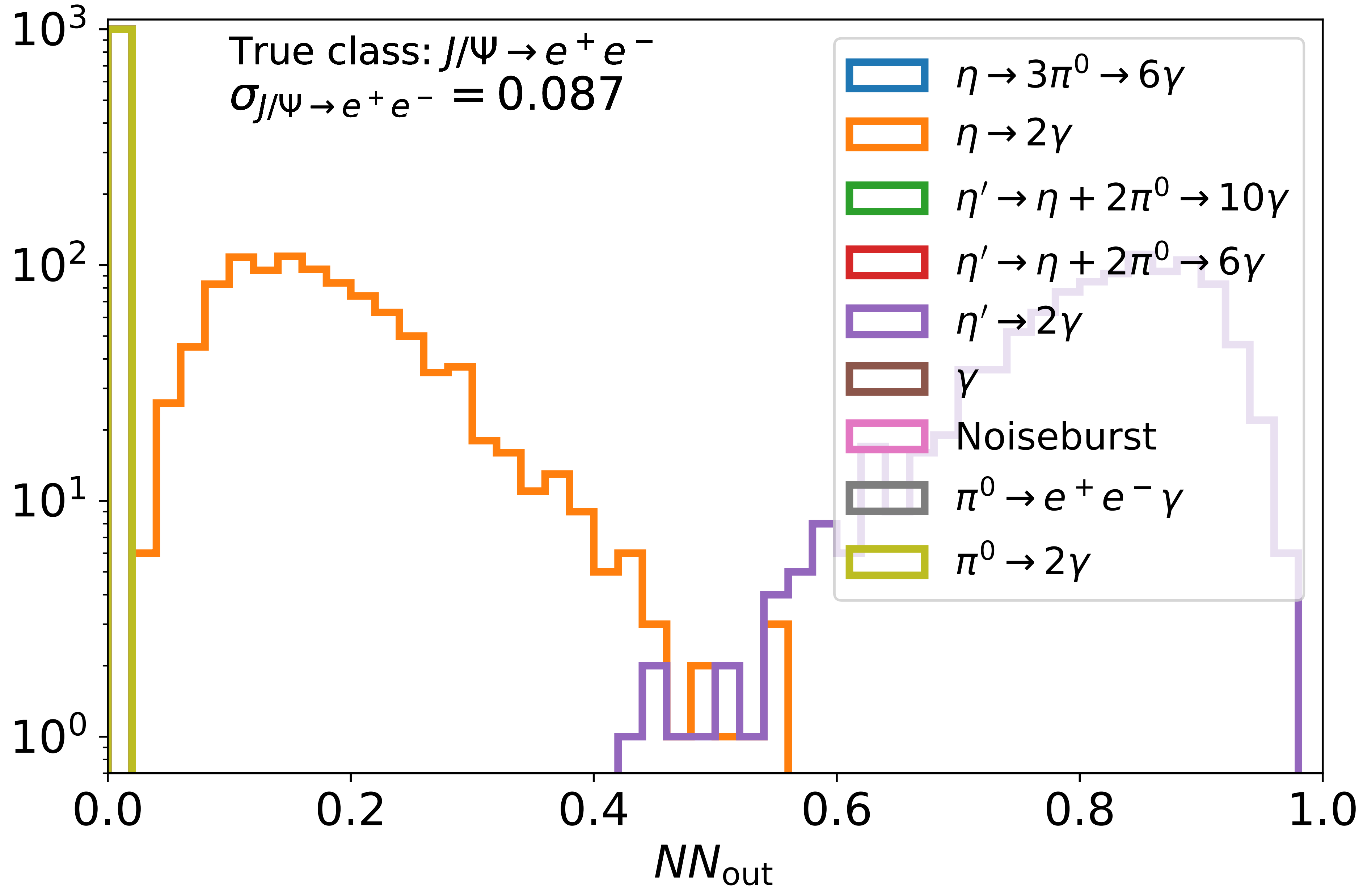
Results

- In general, larger uncertainty than before
- Stems from three different cases:
  A) Examples with ~0 variance
  B) Examples with >1 active output node
  C) Examples with "jumpy decisions"

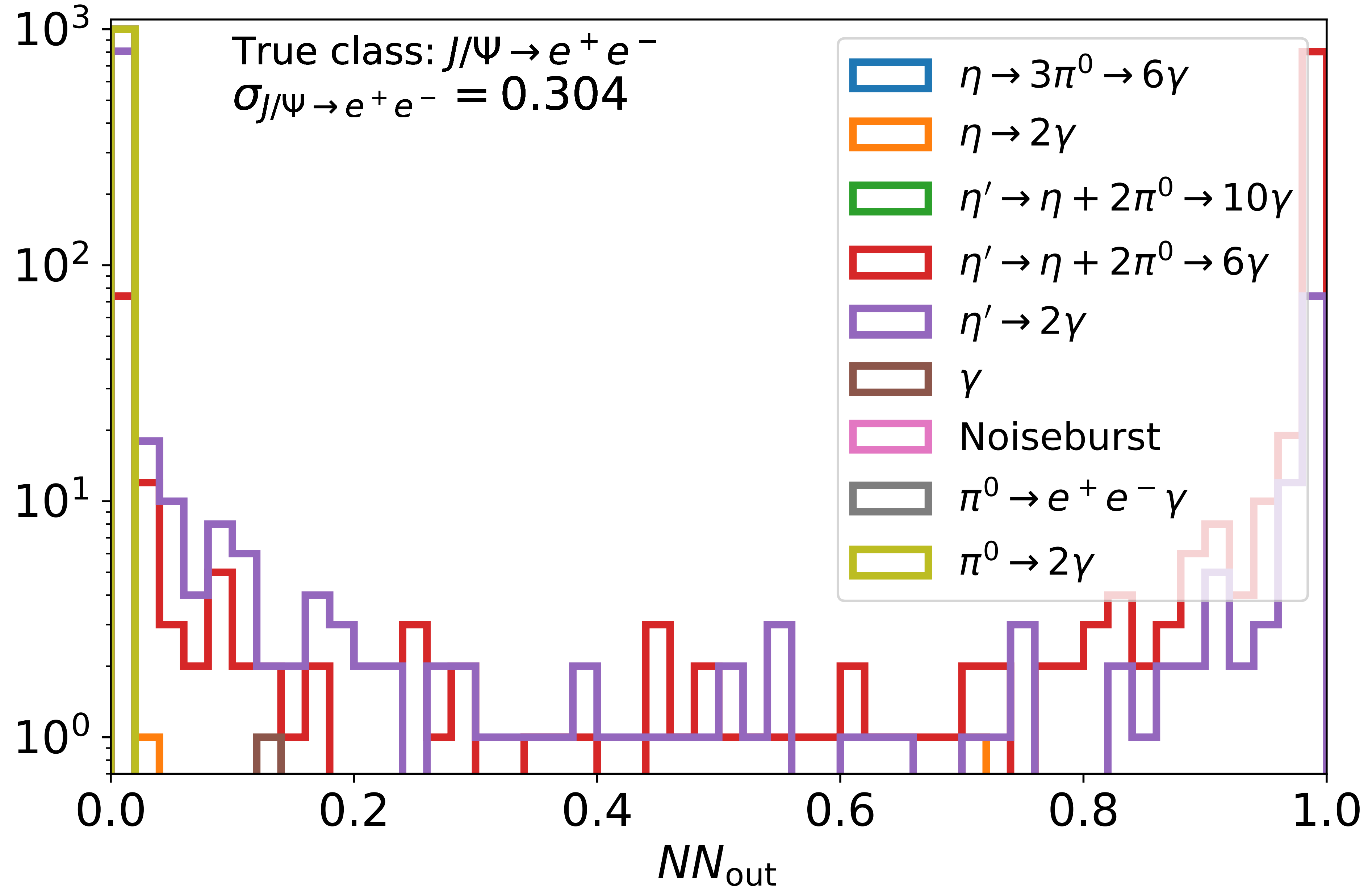
A)



B)



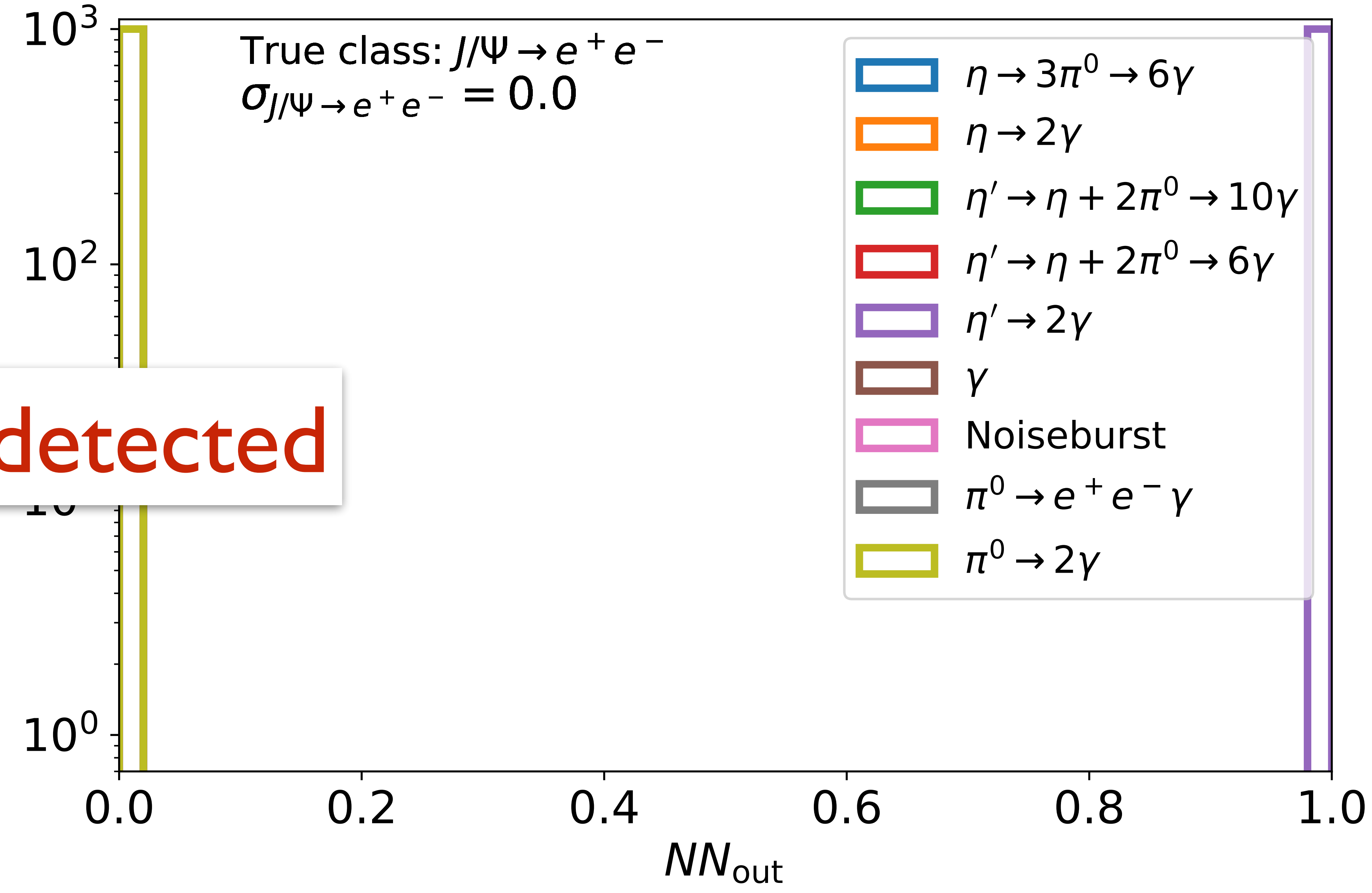
C)



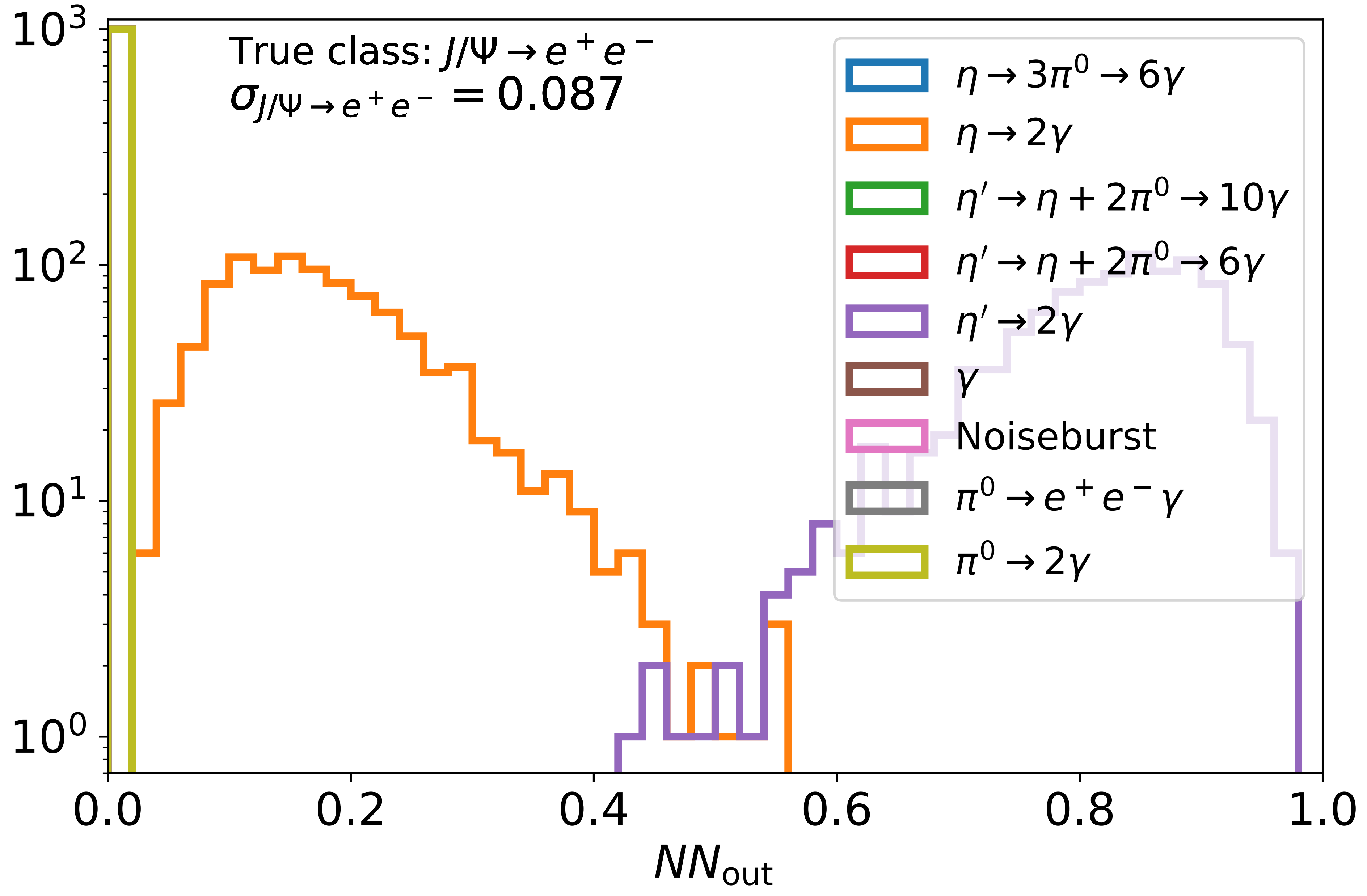
Results

- In general, larger uncertainty than before
- Stems from three different cases:
  - A) Examples with ~0 variance ← anomaly not detected
  - B) Examples with >1 active output node
  - C) Examples with "jumpy decisions"

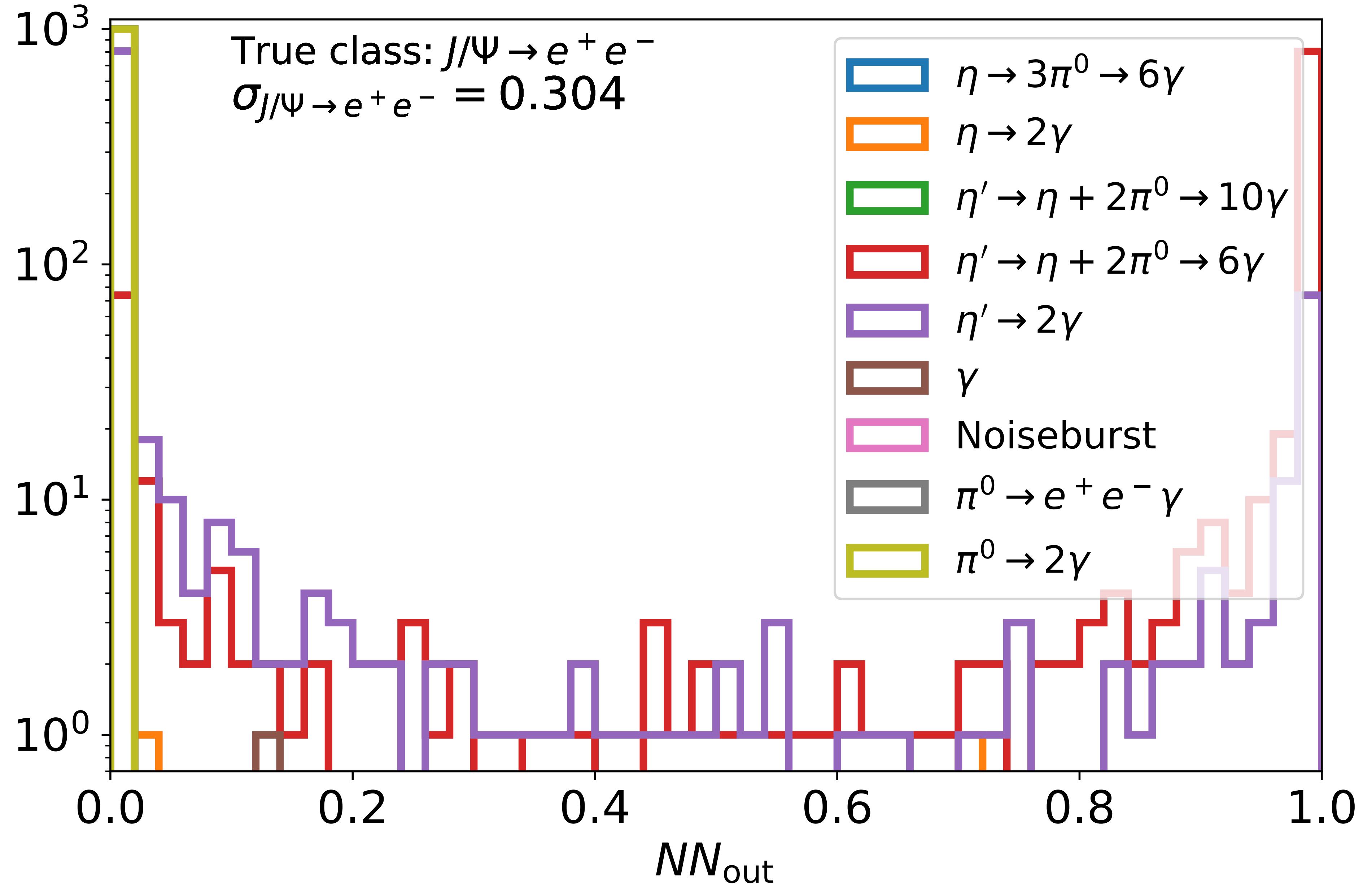
A)



B)



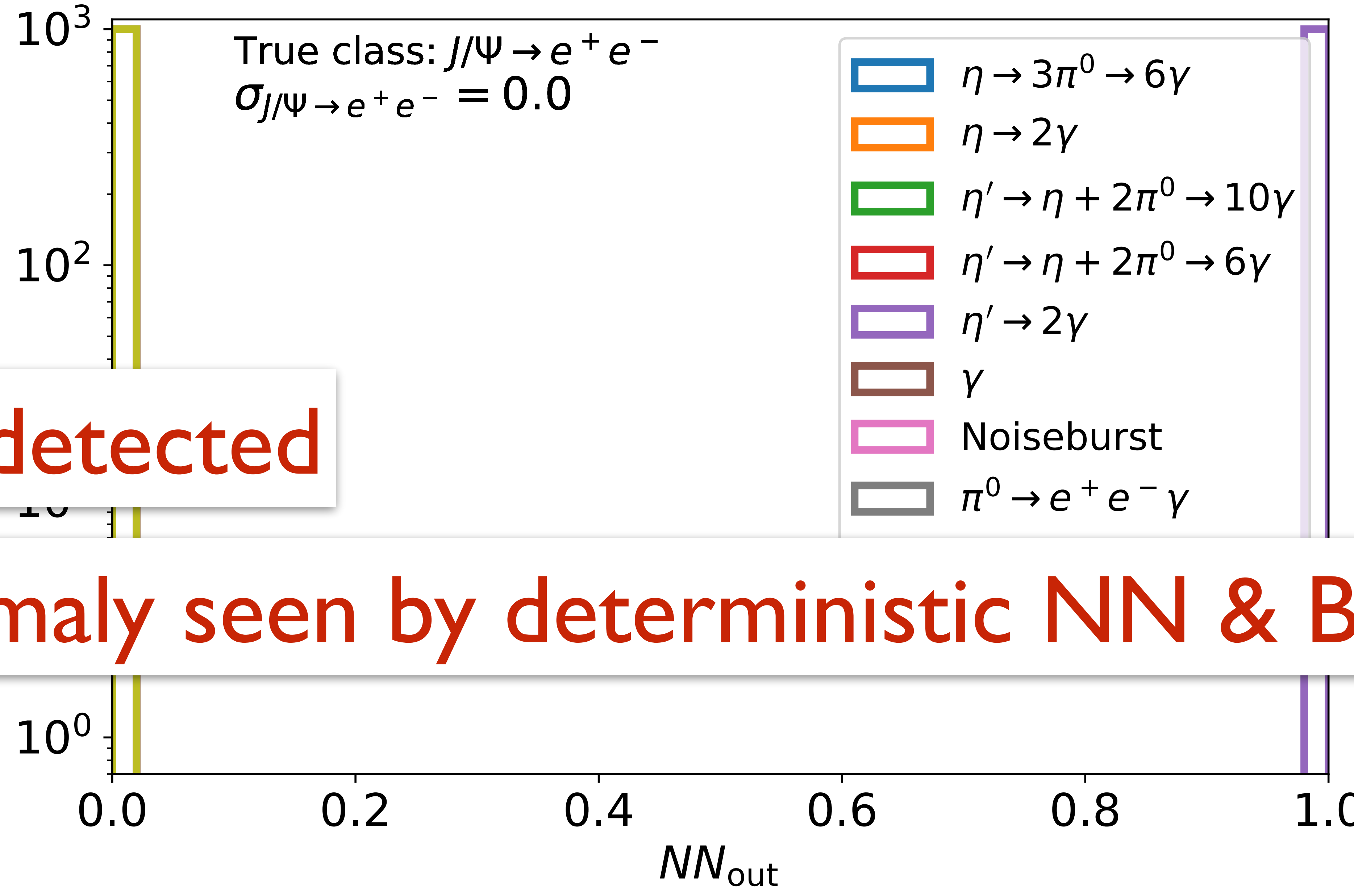
C)



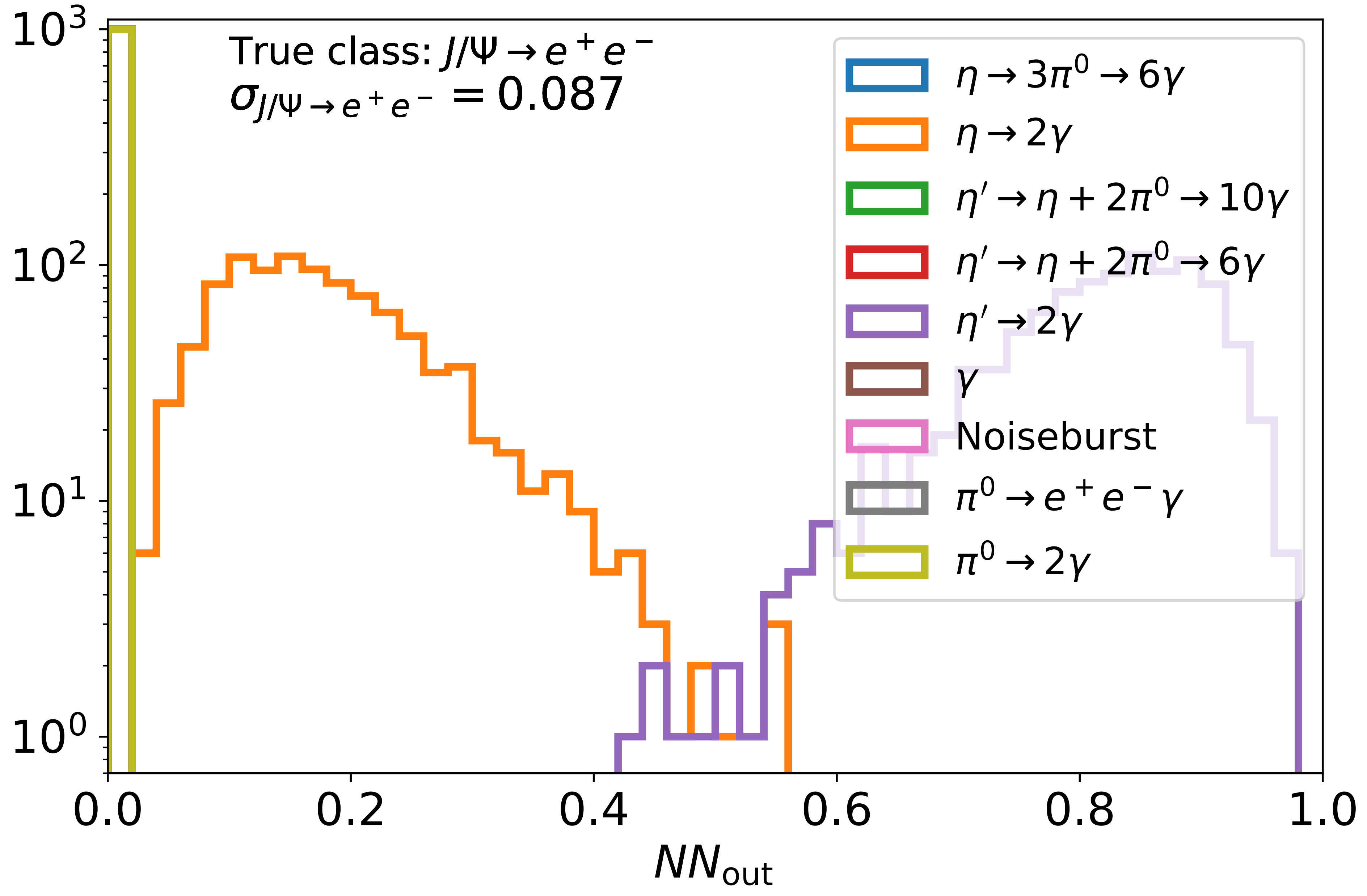
Results

- In general, larger uncertainty than before
- Stems from three different cases:
  - A) Examples with ~0 variance ← anomaly not detected
  - B) Examples with >1 active output node ← anomaly seen by deterministic NN & BNN
  - C) Examples with "jumpy decisions"

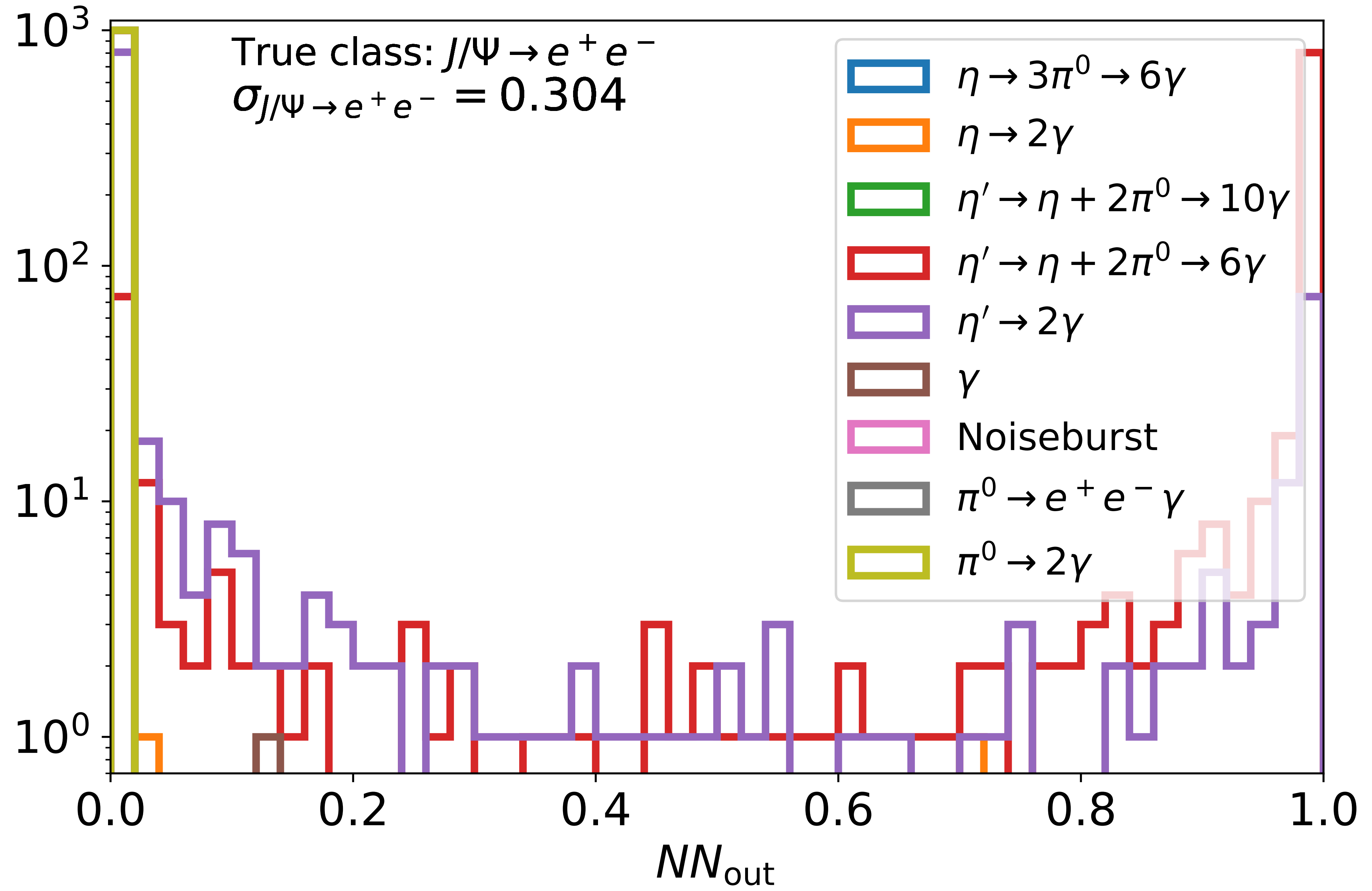
A)



B)



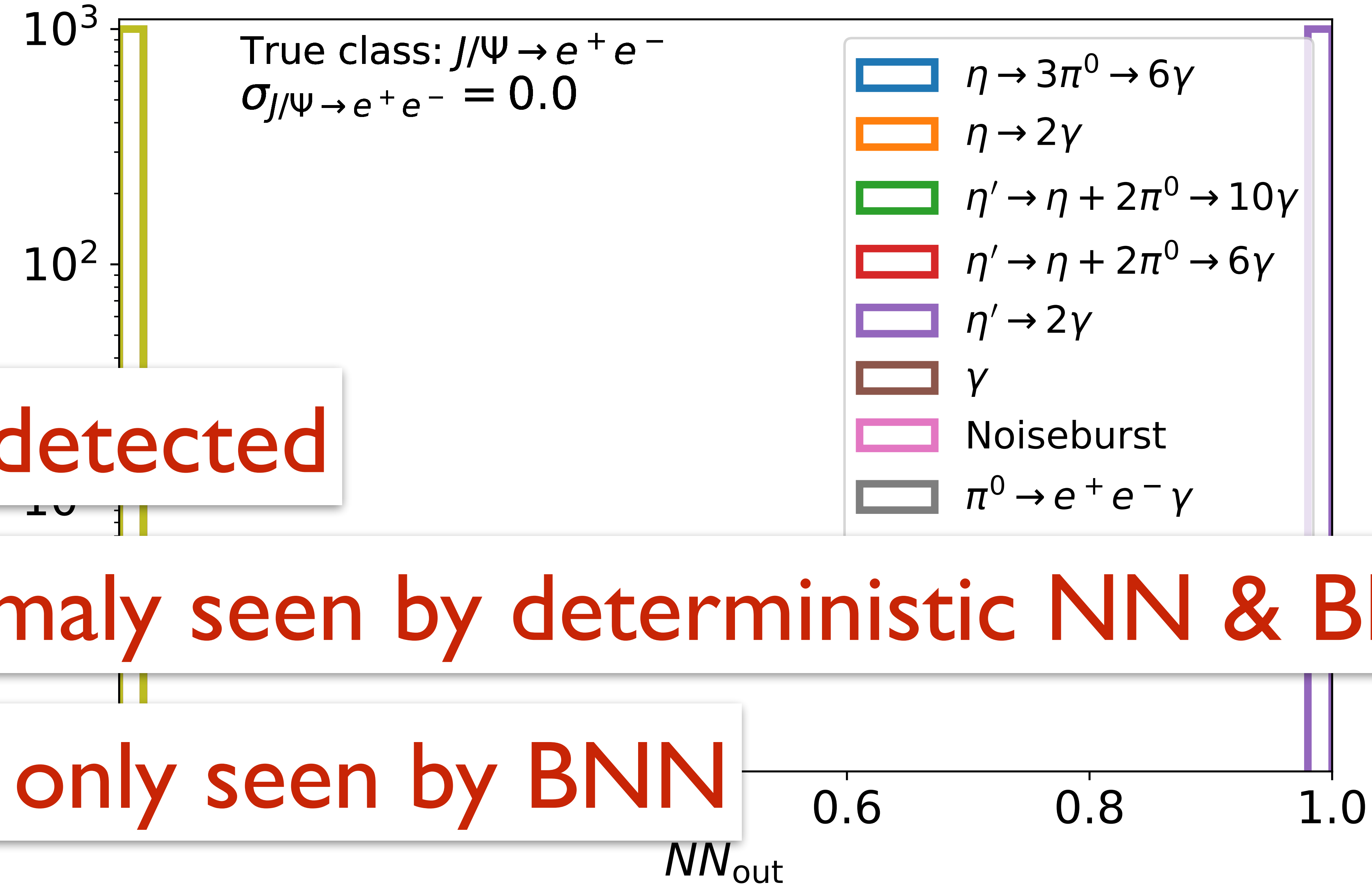
C)



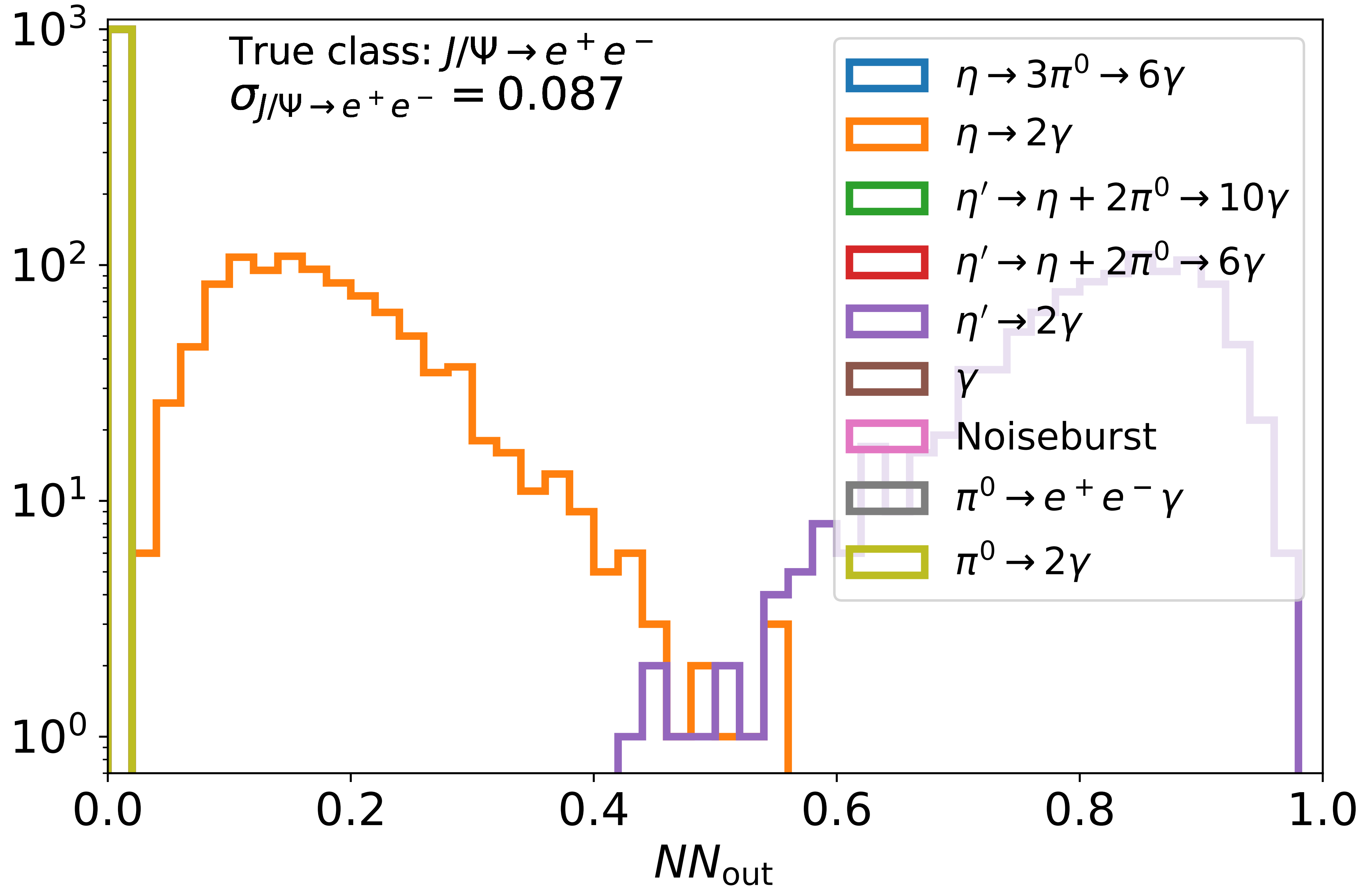
Results

- In general, larger uncertainty than before
- Stems from three different cases:
  - A) Examples with ~0 variance ← anomaly not detected
  - B) Examples with >1 active output node ← anomaly seen by deterministic NN & BNN
  - C) Examples with "jumpy decisions" ← anomaly only seen by BNN

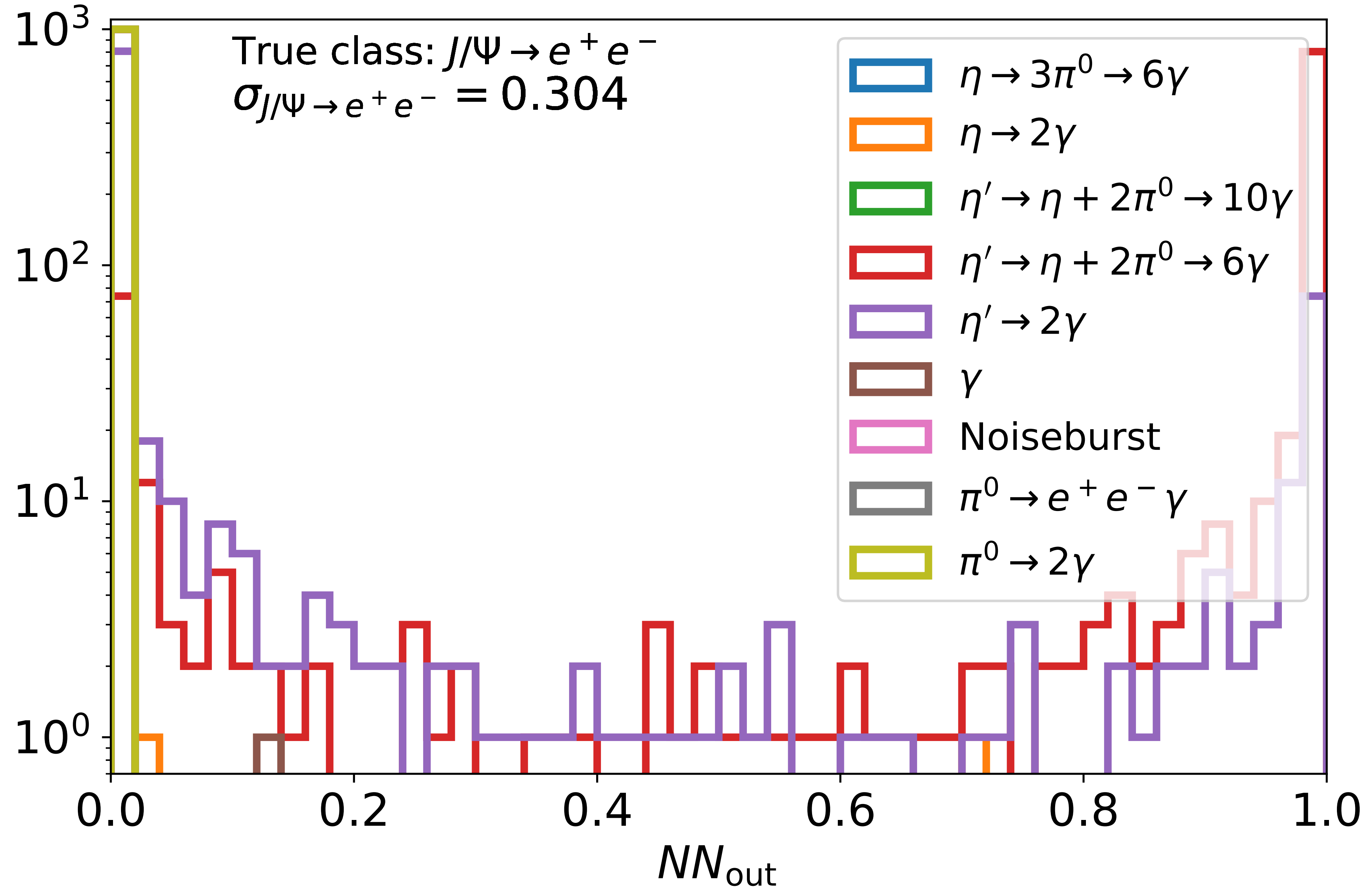
A)



B)

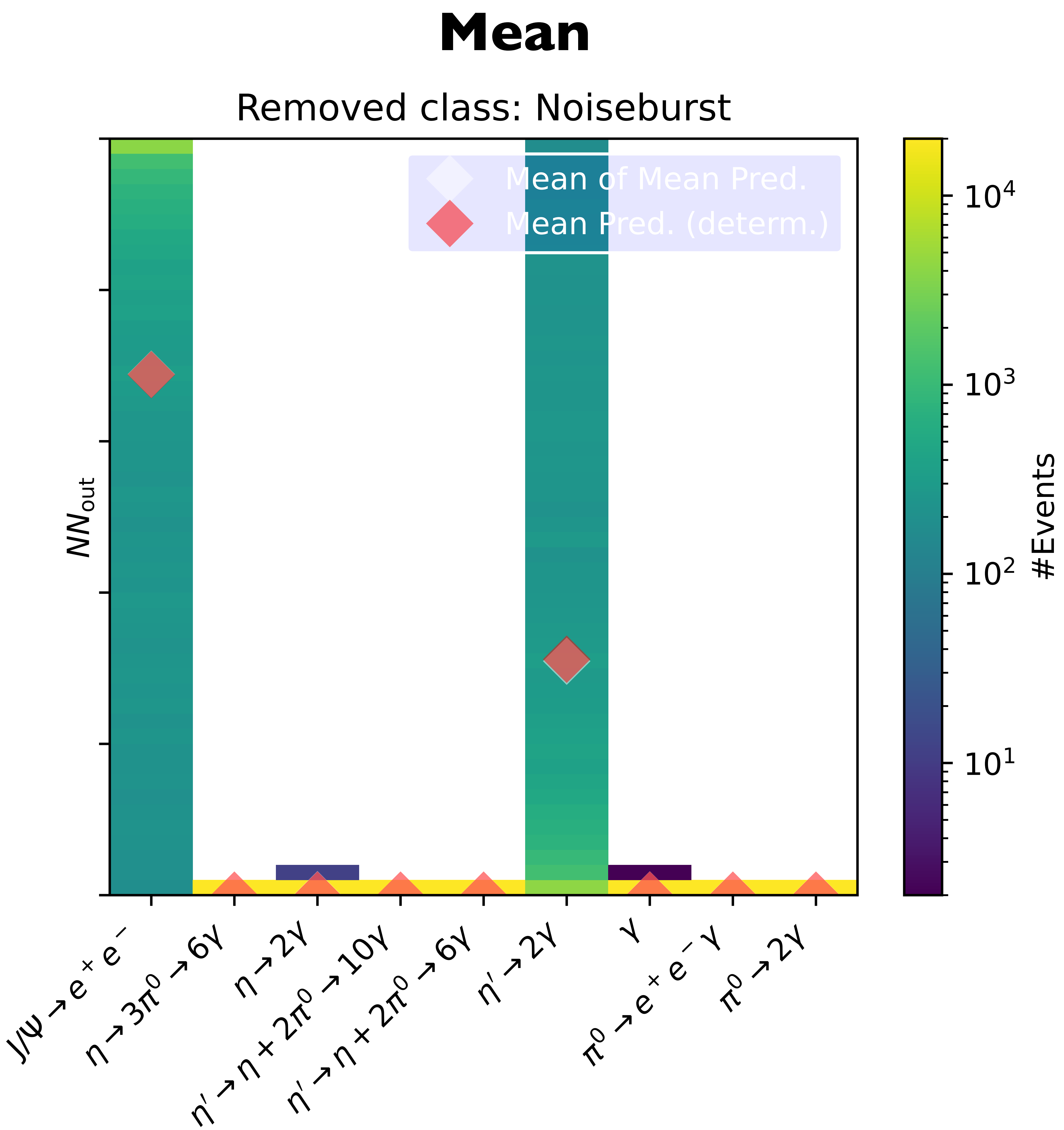


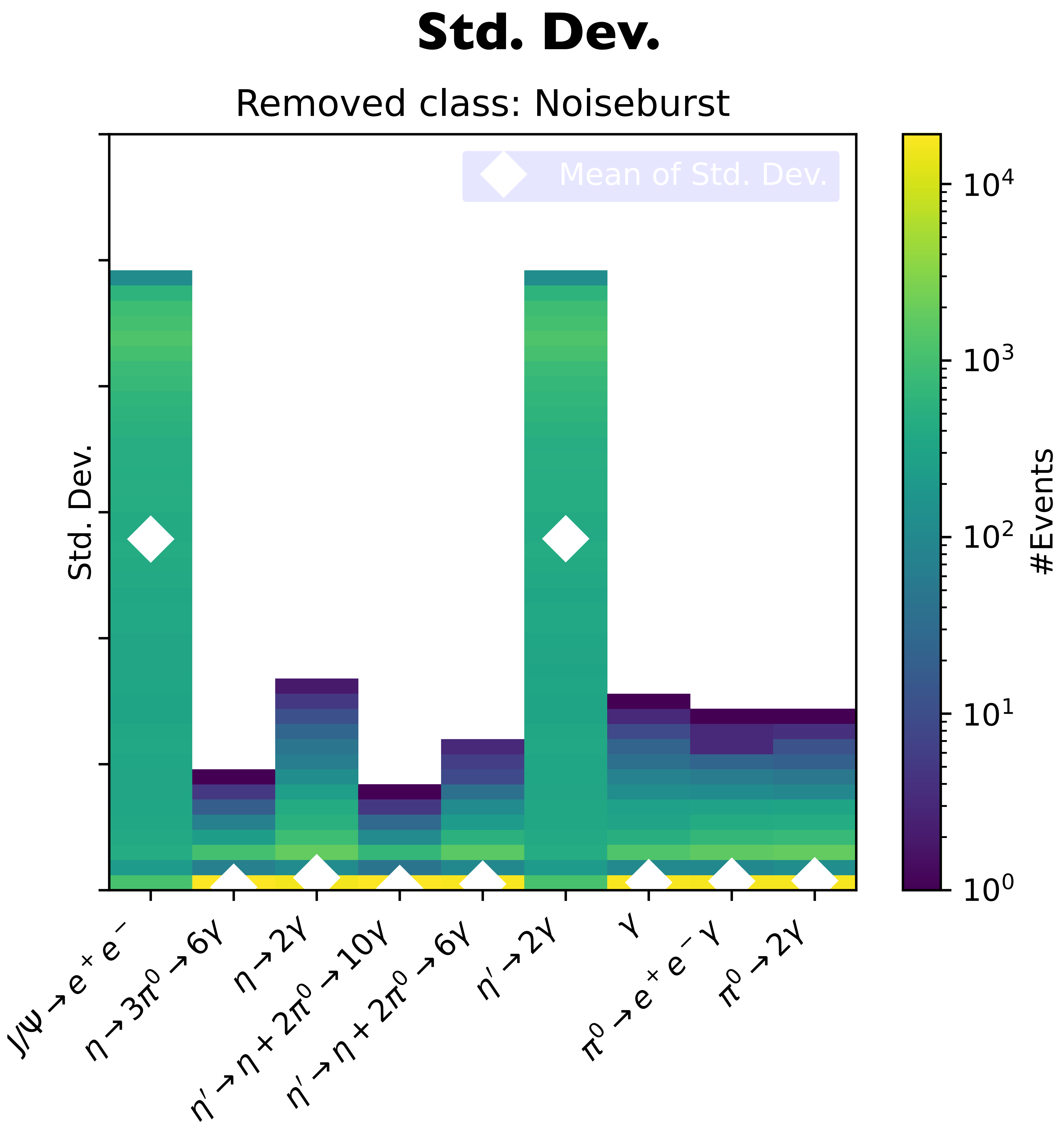
C)



Results

- Another example: noise burst as anomaly





Results

- Standard deviations when each class is removed one-by-one
- Some anomalies are easier to identify than others

Conclusions

- Bayesian network's uncertainty estimate may help to identify anomalies
- Semisupervised approach (LHC Olympics 2020, 2101.08320)

Section	Short Name	Method Type
3.1	VRNN	Unsupervised
3.2	ANODE	Unsupervised
3.3	BuHuLaSpa	Unsupervised
3.4	GAN-AE	Unsupervised
3.5	GIS	Unsupervised
3.6	LDA	Unsupervised
3.7	PGA	Unsupervised
3.8	Reg. Likelihoods	Unsupervised
3.9	UCluster	Unsupervised
4.1	CWoLa	Weakly Supervised
4.2	CWoLa AE Compare	Weakly/Unsupervised
4.3	Tag N' Train	Weakly Supervised
4.4	SALAD	Weakly Supervised
4.5	SA-CWoLa	Weakly Supervised
5.1	Deep Ensemble	Semisupervised
5.2	Factorized Topics	Semisupervised
5.3	QUAK	Semisupervised
5.4	LSTM	Semisupervised

- We don't know, yet, if it can compete with specialized AD algorithms
- One advantage:
  Provides AD capabilities as a sanitiy/DQ cross check for standard classification tasks (such as ID)
  - At some additional training and prediction costs