Combine & Conquer: Event Reconstruction with Bayesian Ensemble Neural Networks

Jack Y. Araz
Institute For Particle Physics Phenomenology
Durham University

based on JHEP 04 (2021) 296, arXiv: 2102.01078 [hep-ph]
with Michael Spannowsky

Jets and their substructure from LHC data
June 3rd, 2021

Outline

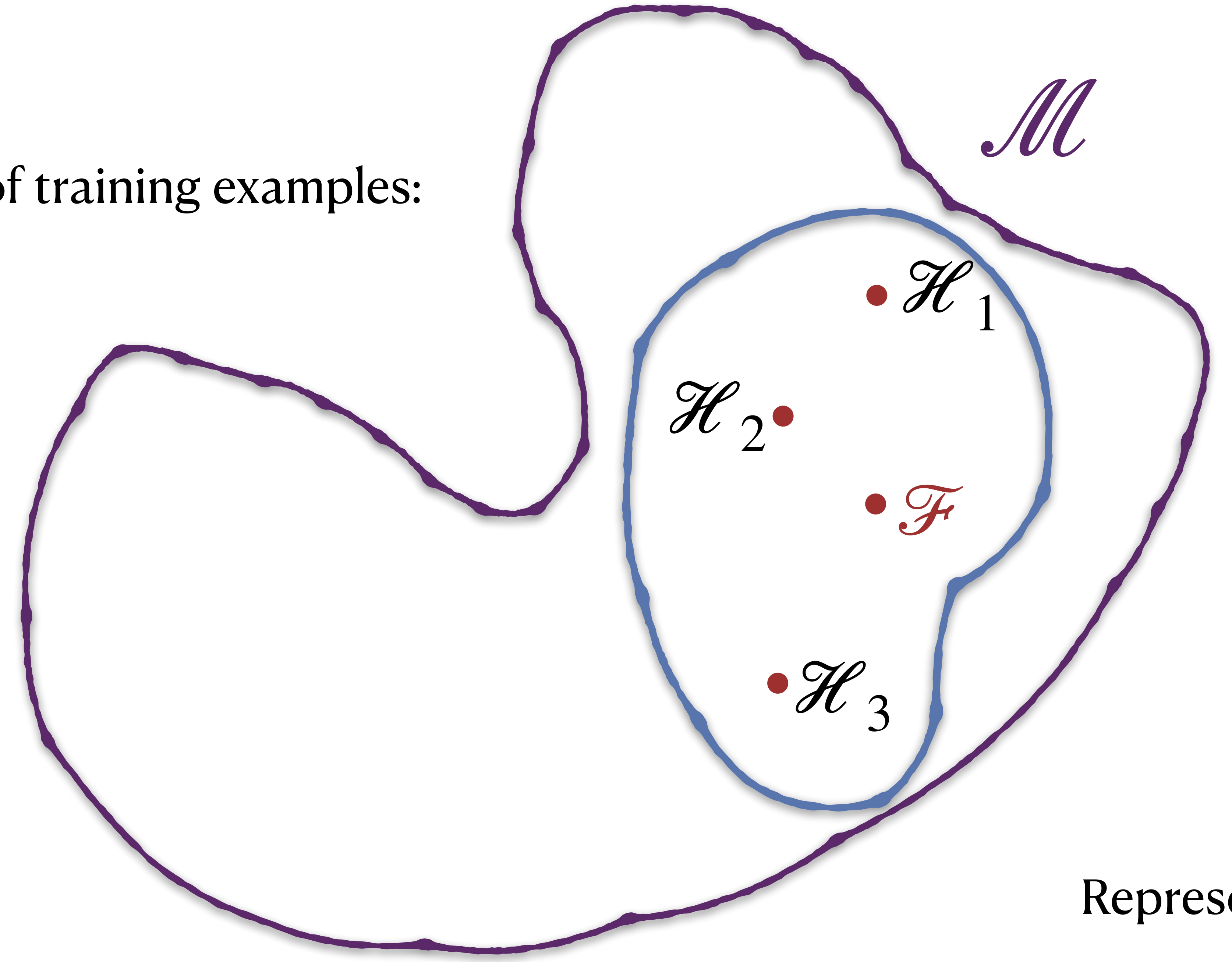
- Introduction
  - ❖ How and why to combine networks
- Top Tagging through Ensemble Learning
  - ❖ Preprocessing & Results
- Improving uncertainties with Ensemble networks
- Conclusion

Combine
“~~Divide~~ and conquer”
Gaius Julius Caesar

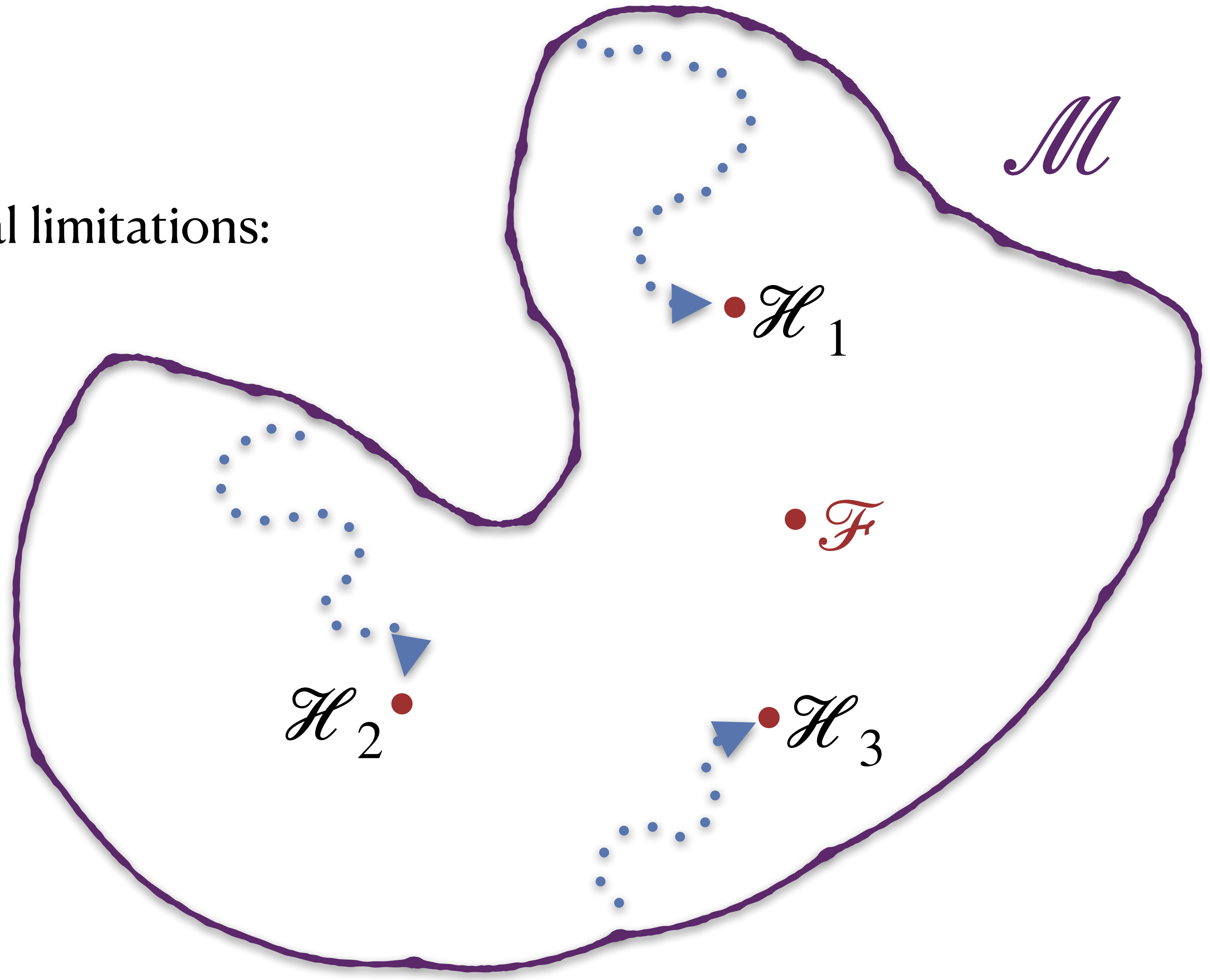
Introduction

Introduction: Why ensemble networks may work?

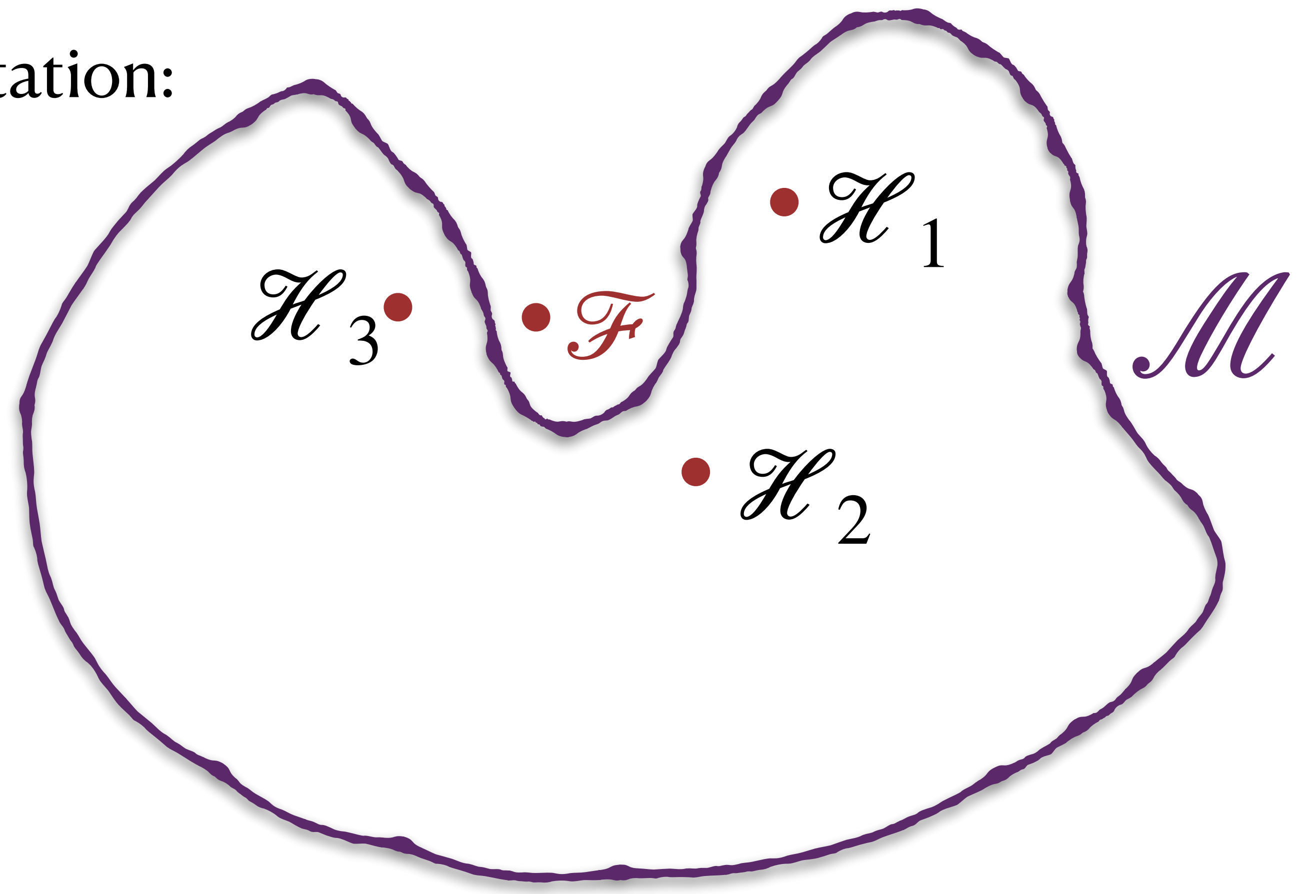
Lack of training examples:



Computational limitations:

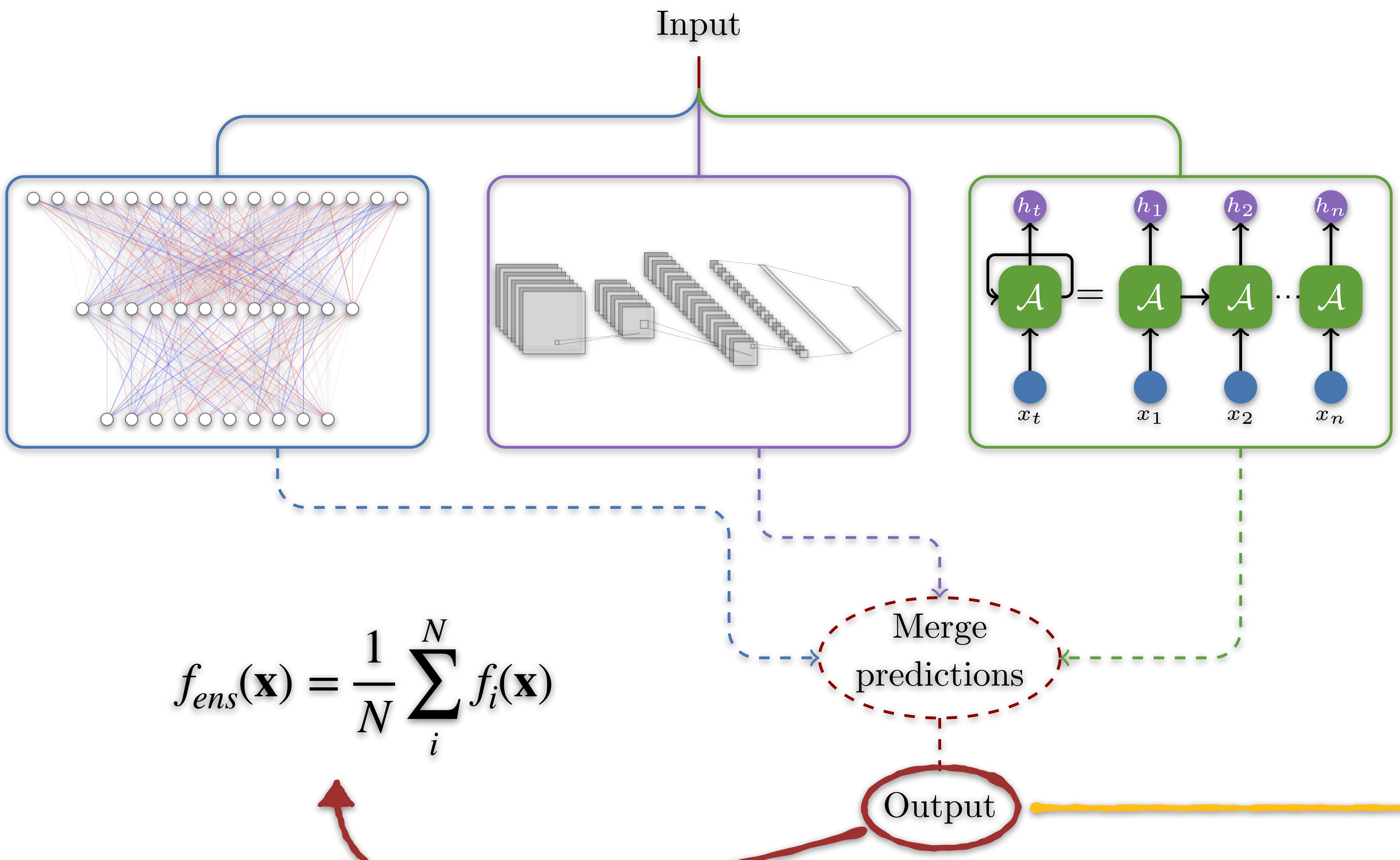


Representation:



Dietterich, 2000

Introduction: Why ensemble networks may work?

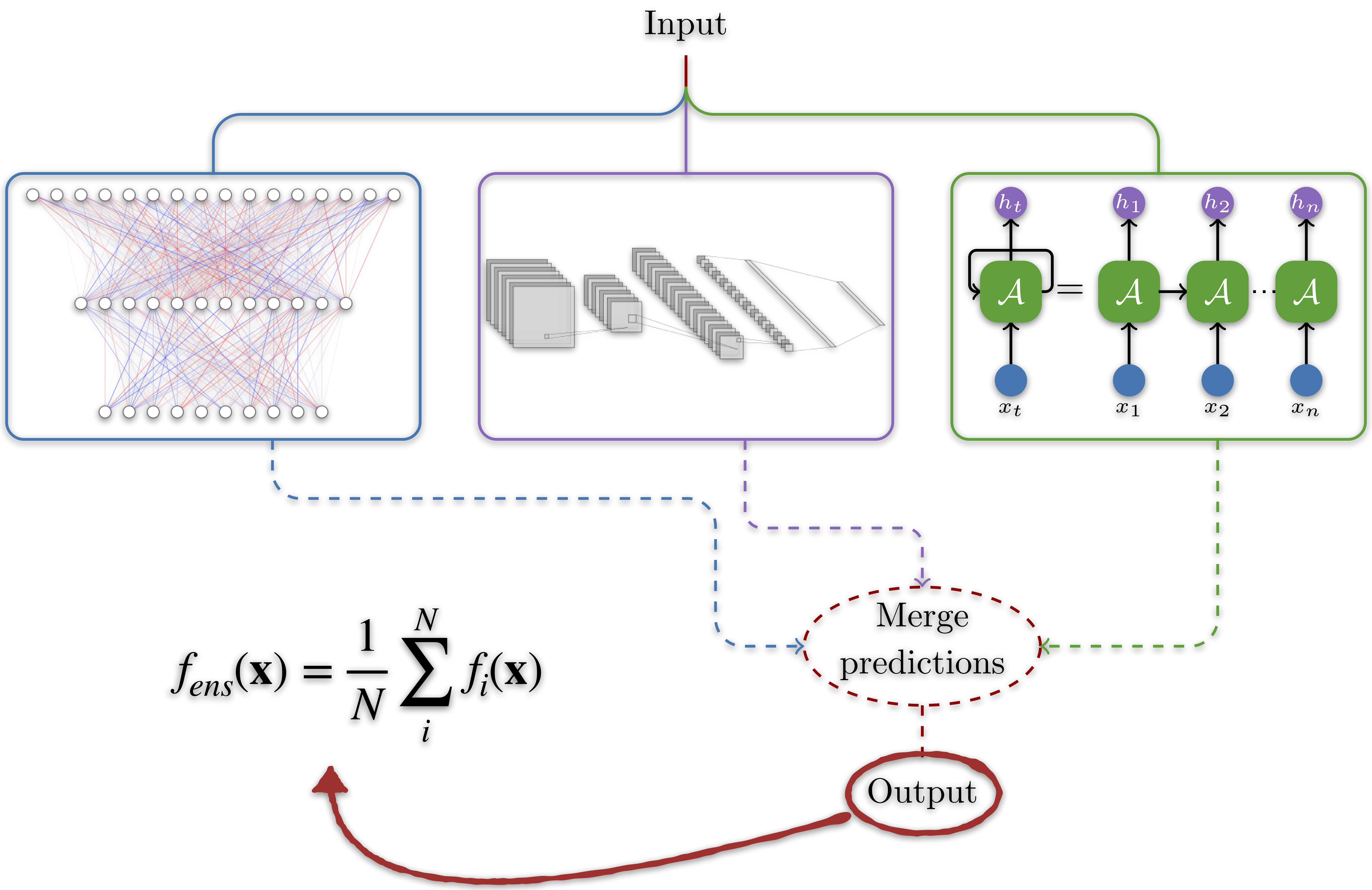


$$f_{ens}(\mathbf{x}) = \frac{1}{N}\sum_{i}^{N} f_i(\mathbf{x})$$

$$\mathrm{Err}(f_{ens}) = \mathrm{Err}\left\{\frac{1}{N}\overline{\mathrm{Var}}(\mathbf{x}) + \left(1 - \frac{1}{N}\right)\overline{\mathrm{Cov}}(\mathbf{x}) + \overline{\mathrm{Bias}}(\mathbf{x})^2\right\}$$

If the component networks are negatively correlated, the mean prediction will decrease the generalization error.

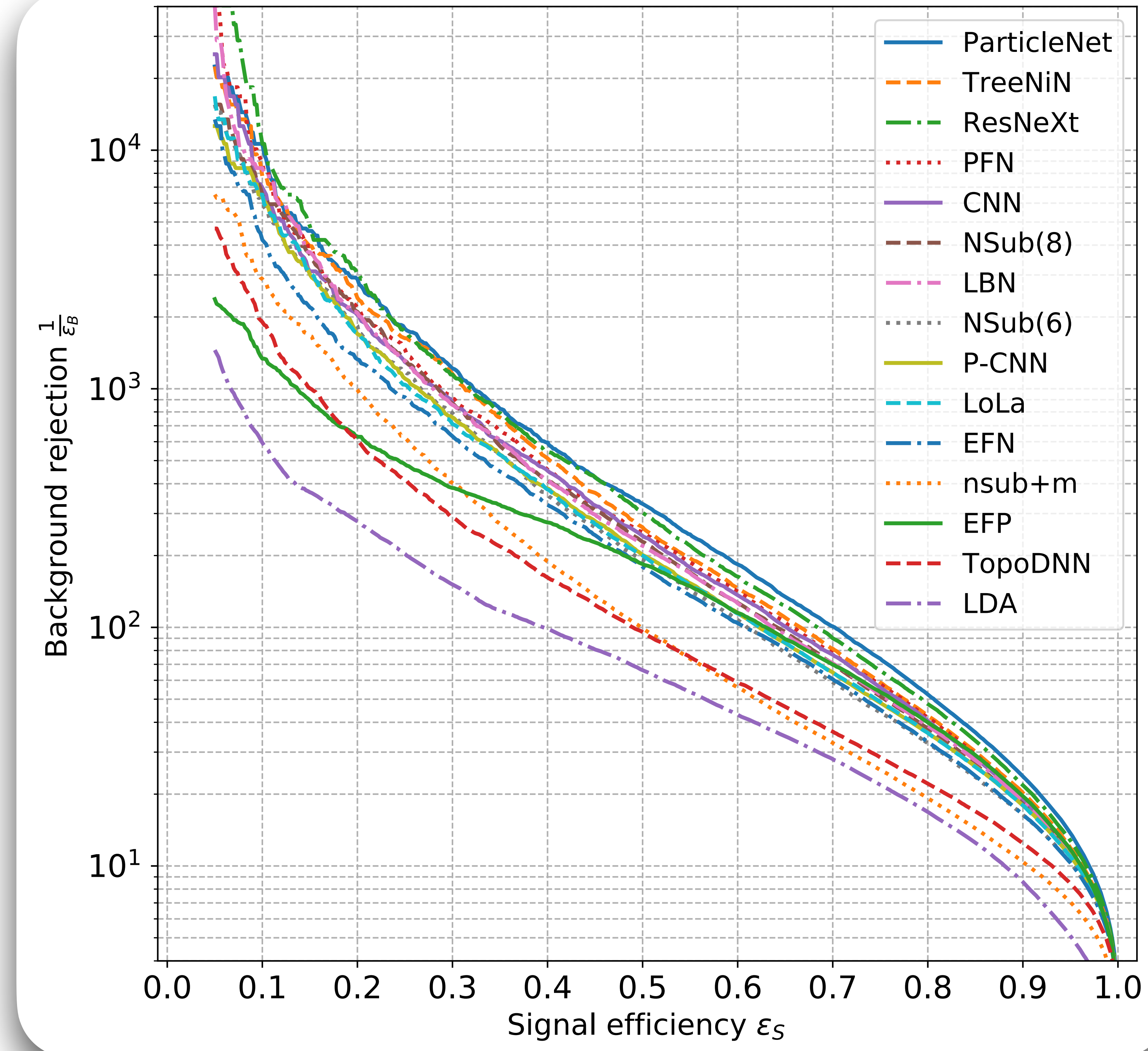
Introduction: Why ensemble networks may work?



$$f_{ens}(\mathbf{x}) = \frac{1}{N}\sum_{i}^{N} f_i(\mathbf{x})$$

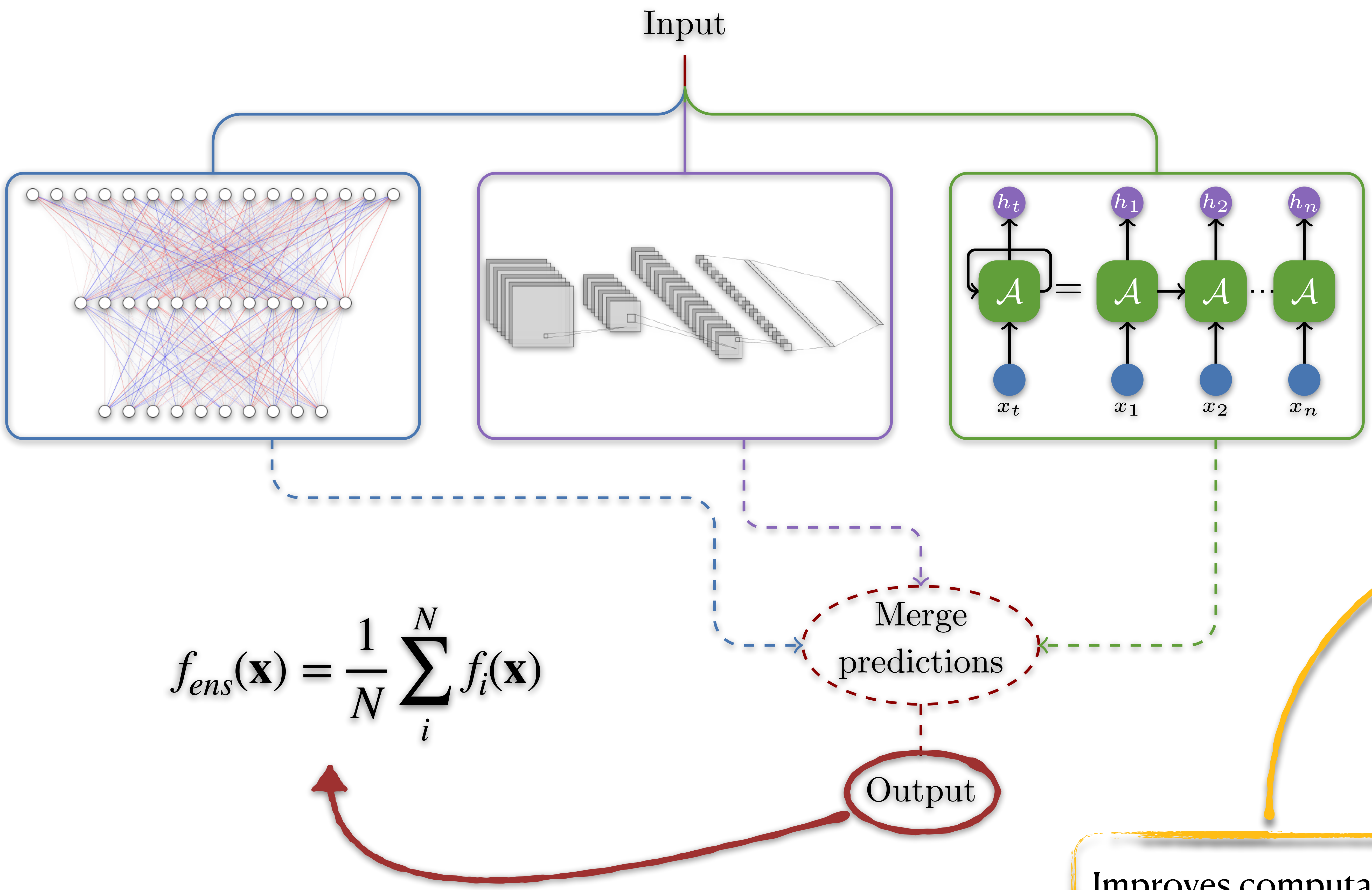
Kasieczka et. al. SciPost'19

Gregor's talk on Tuesday



Taking the mean of the same architecture which initialized multiple times can lead up to 15% improvement over component networks

Introduction: Why ensemble networks may work?

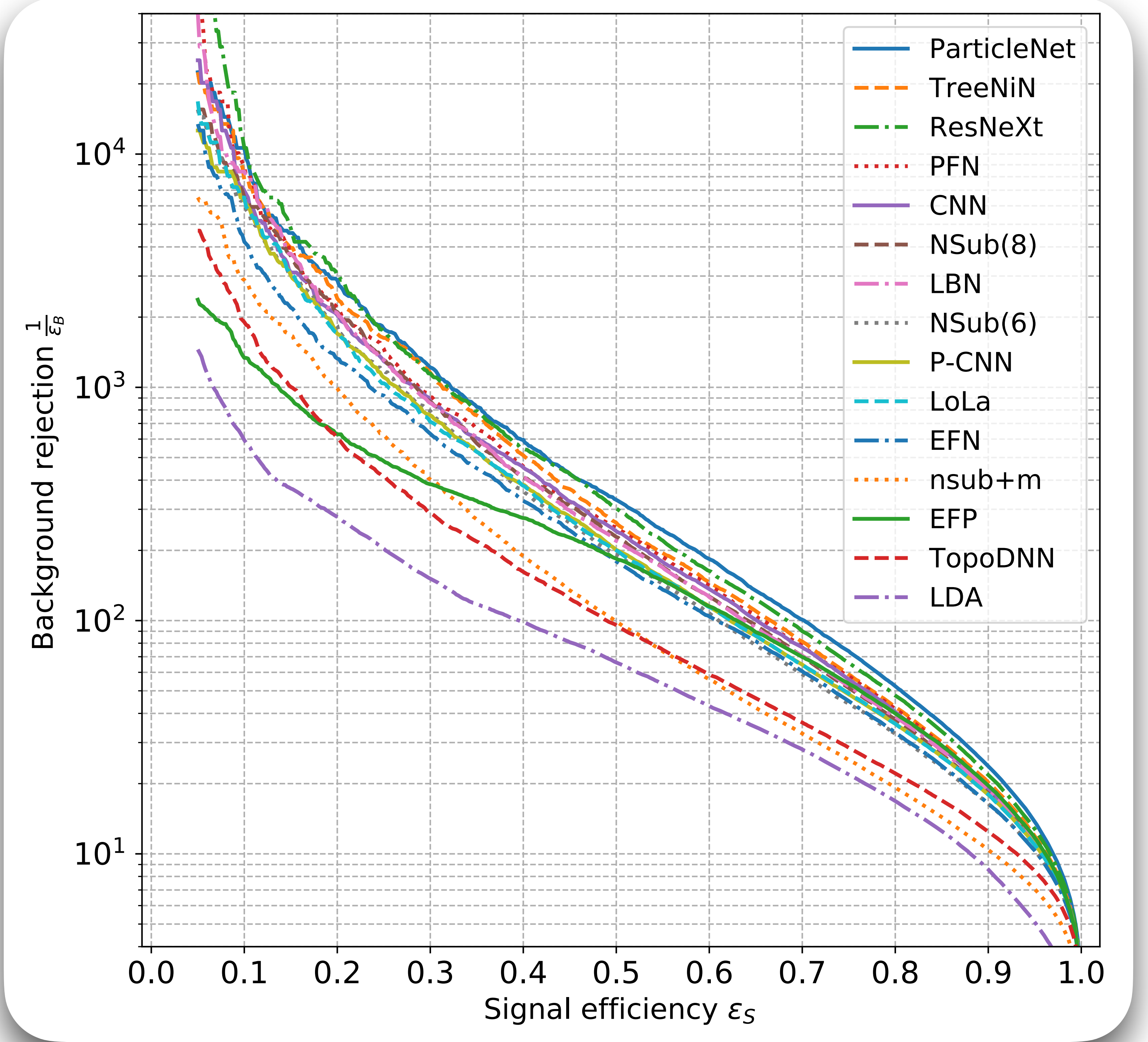


$$f_{ens}(\mathbf{x}) = \frac{1}{N}\sum_{i}^{N} f_i(\mathbf{x})$$

Improves computational ineficientes!!

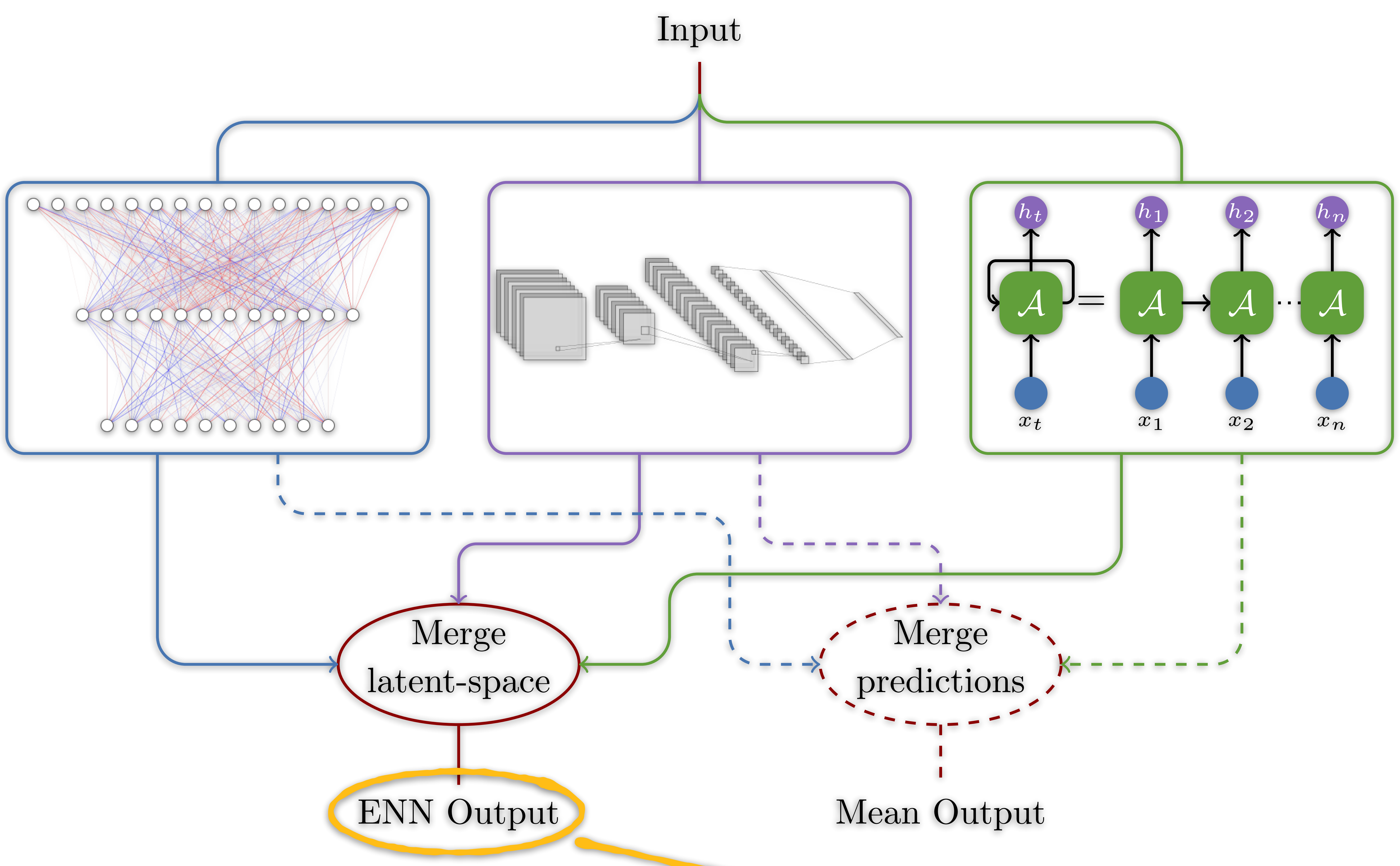
Kasieczka et. al. SciPost'19

Gregor's talk on Tuesday



Taking the mean of the same architecture which initialized multiple times can lead up to 15% improvement over component networks

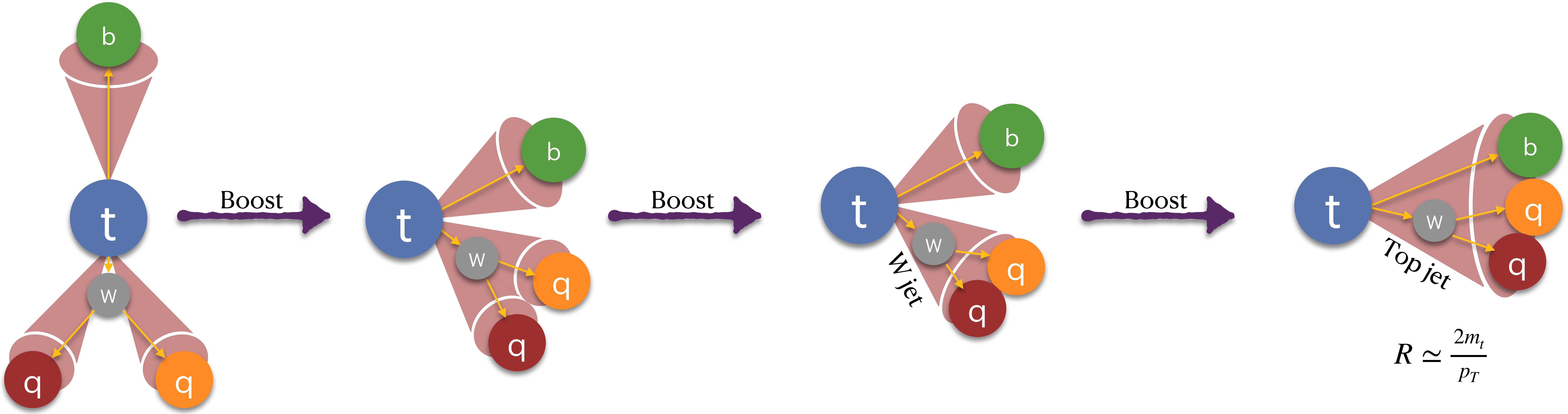
Introduction : Step further!



Instead of merging already trained output, merging latent-space during the training will allow two network to train with respect to each other. This will effectively expand the representation of the network hypothesis.

Top Tagging Through Ensemble Learning

Top Tagging Through Ensemble Learning



$$R \simeq \frac{2m_t}{p_T}$$

Traditional taggers

HEPTopTagger, Soft Drop Tagger, Mass Drop Tagger which are based on grooming, pruning and trimming techniques alongside with geometric substructure selection.

For an extensive review
Plehn, Spannowsky, IOP '12

ML Based Taggers

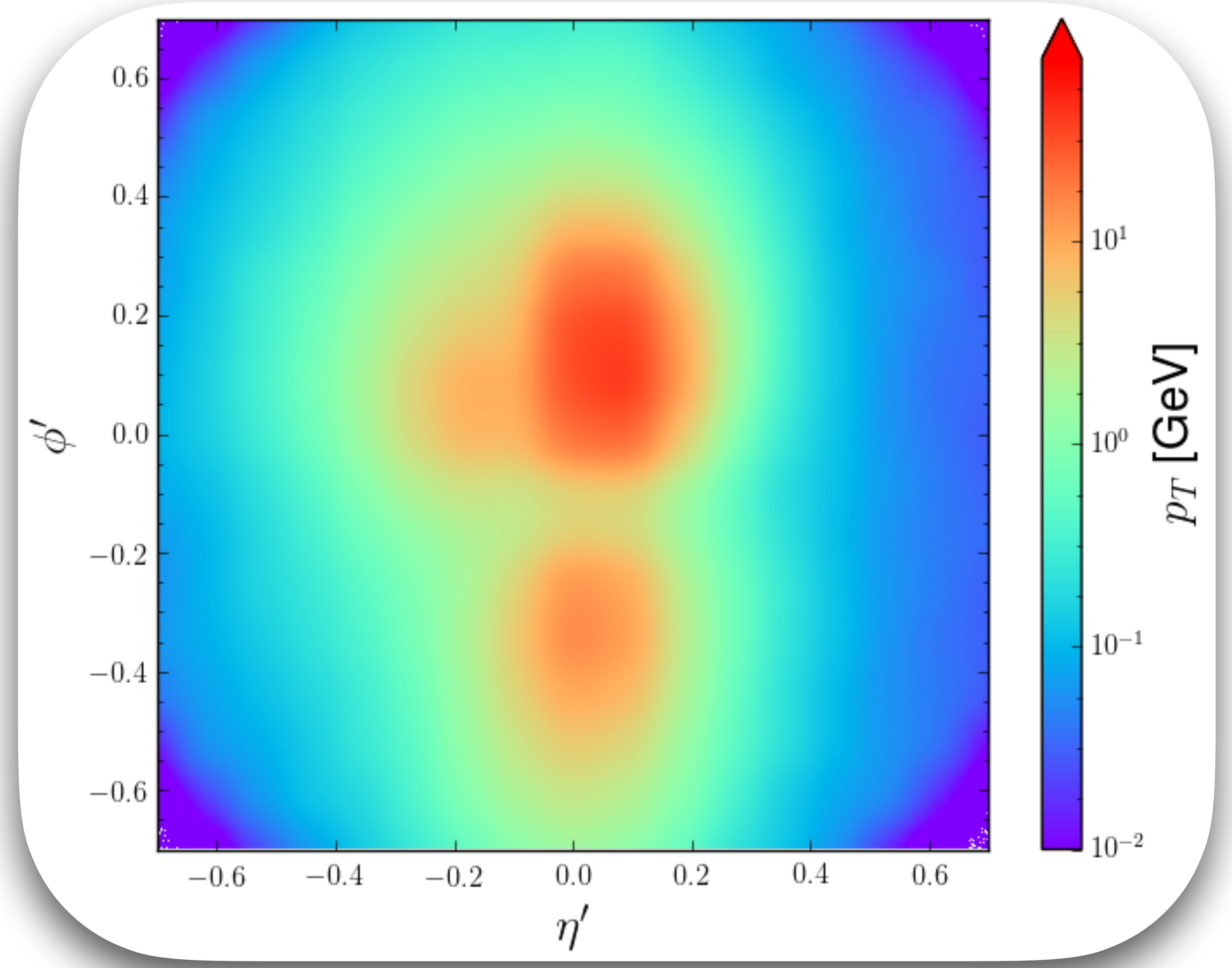
Several techniques have been presented so far with great success. In this talk we will look into Convolutional and Recurrent Networks.

Gregor's talk
on Tuesday

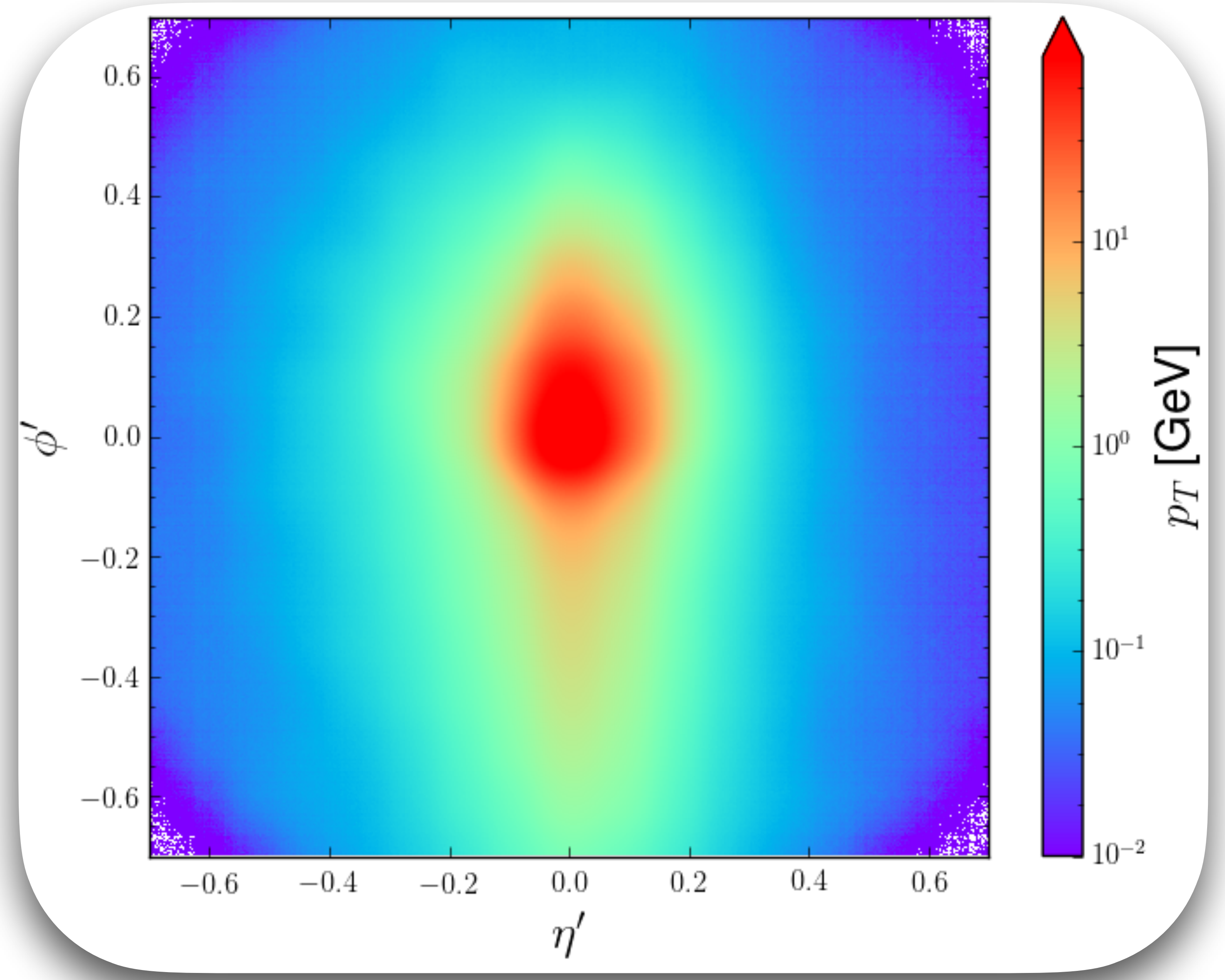
For an extensive review
Kasieczka et. al. SciPost'19

Top Tagging: CNN Preprocessing

Top Signal



QCD Background



- Leading FatJet Definition: anti-k_T algorithm with $R = 0.8, p_T \in [550,\ 650]$ GeV, $|\eta| < 2$
- Parton matching with $\Delta R(j, t_{truth}) < 0.8$
- Jets are centred with respect to p_T weighted centroid where jet vector is at $(\phi, \eta) = (0,0)$
- Principal axis has been rotated to $+\eta$ direction
- Energy deposits has been divided into 37×37 pixels which corresponds to η & $\phi \in [-1.5,\ 1.5]$.
- Image has been flipped to place the most energetic quadrant to the top right corner.

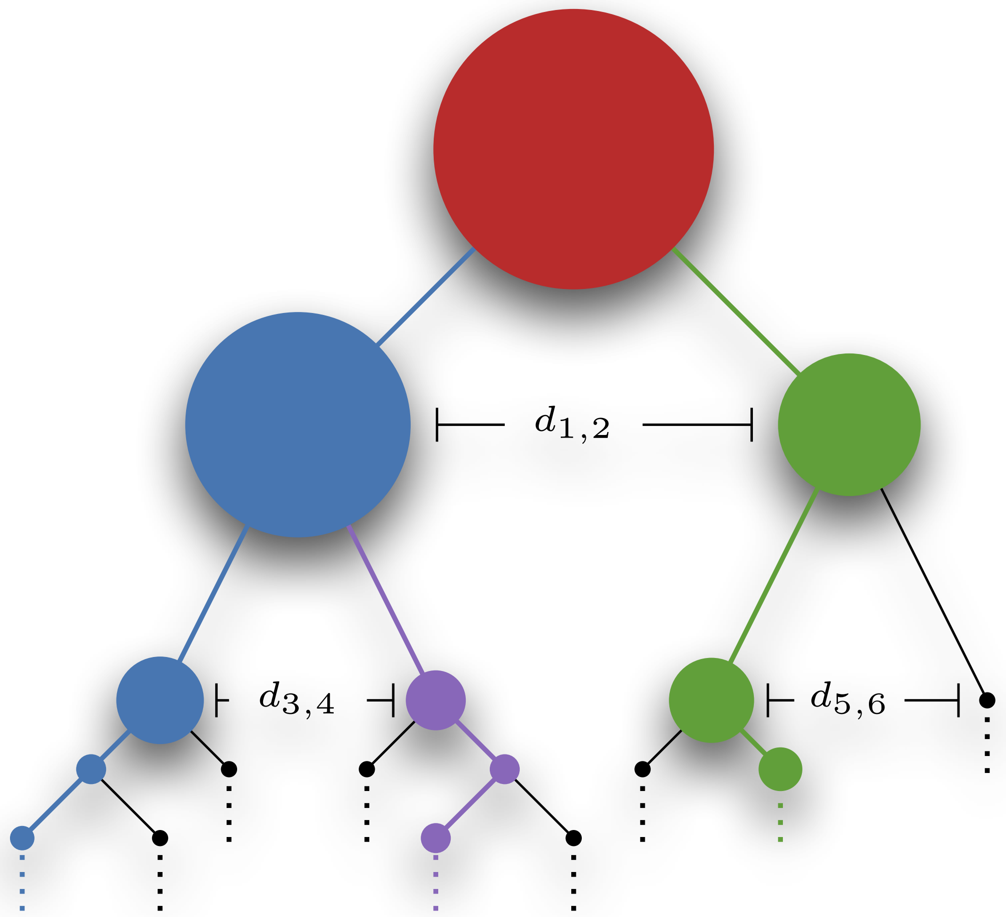
Similar studies:

Kasieczka, Plehn, Russell, Schell; JHEP '17

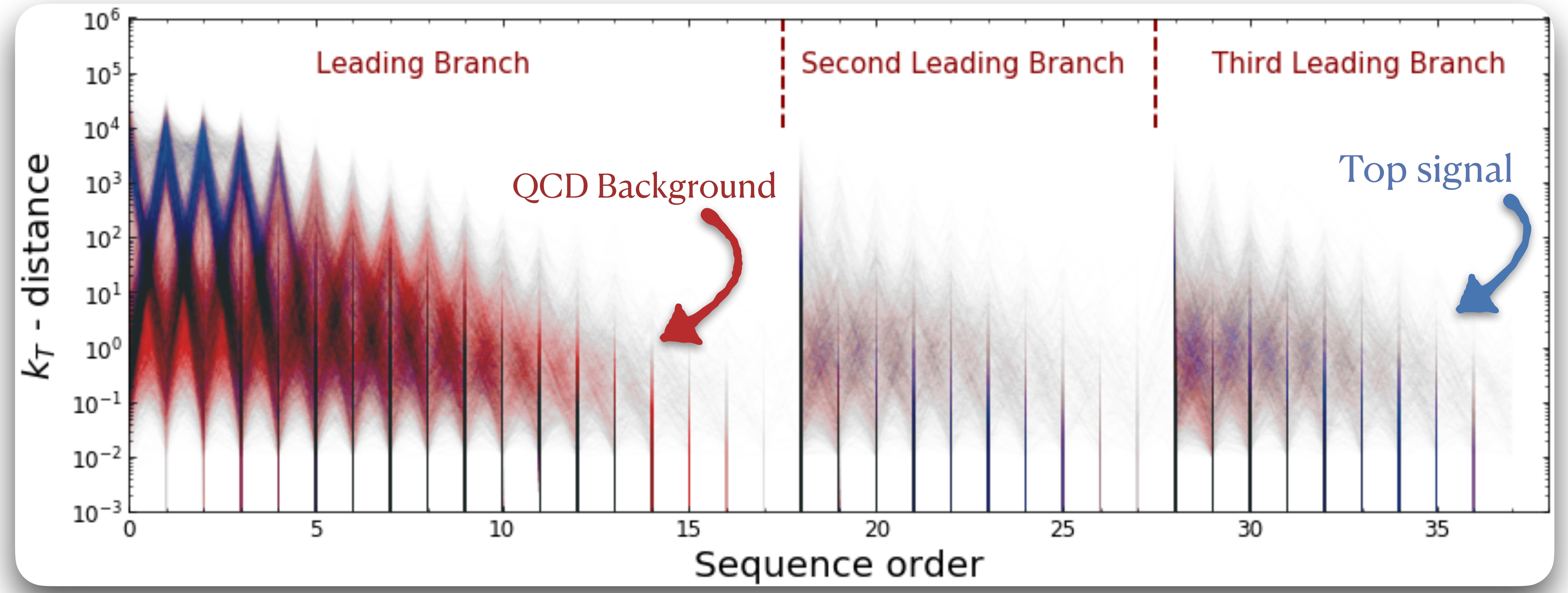
Macaluso, Shih; JHEP '18

Michael's talk this afternoon

Top Tagging: RNN Preprocessing



$$k_T - \text{distance} := min(p_{T,i}^2,\ p_{T,j}^2)\ \frac{\Delta R^2}{R}$$

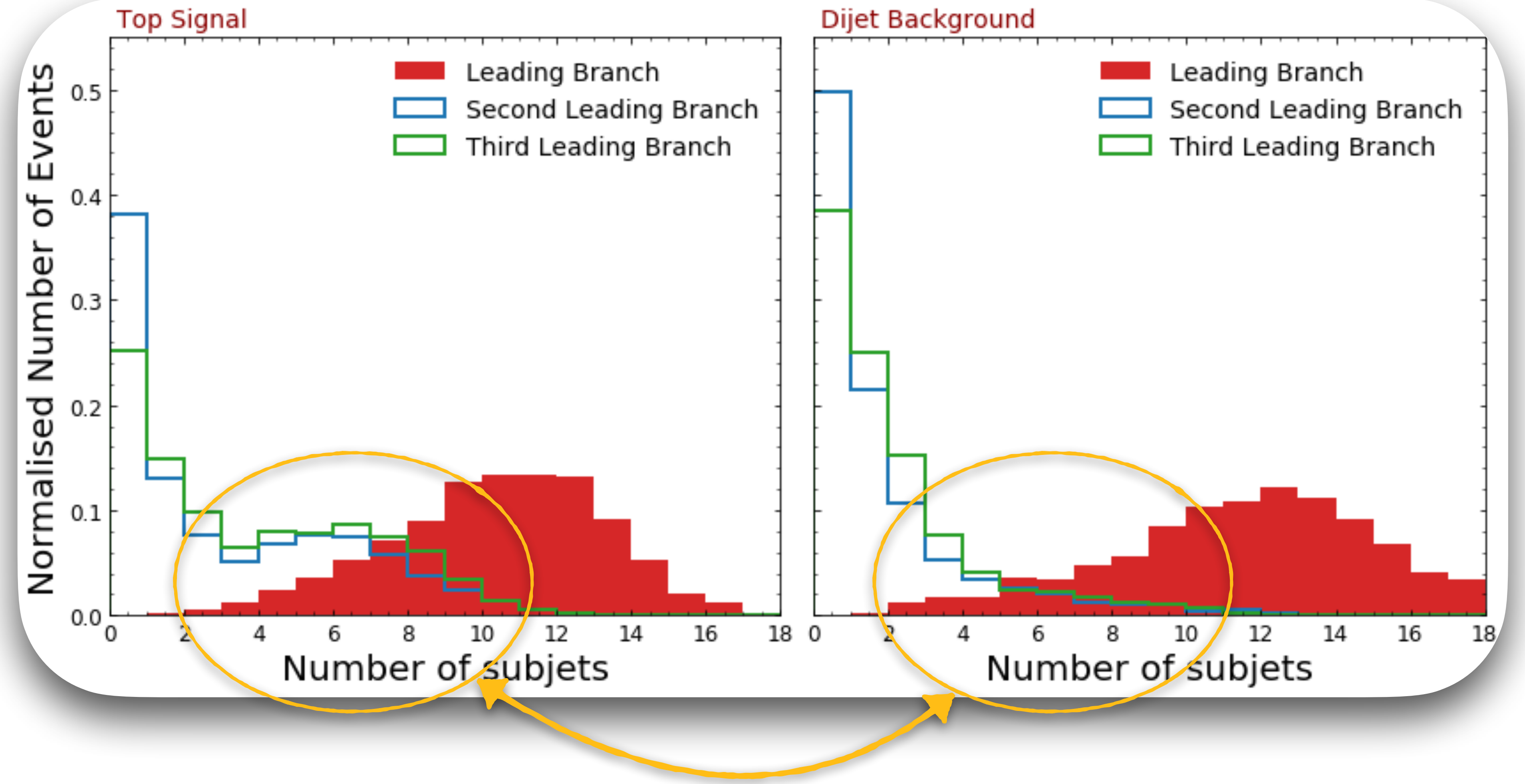


Similar studies:

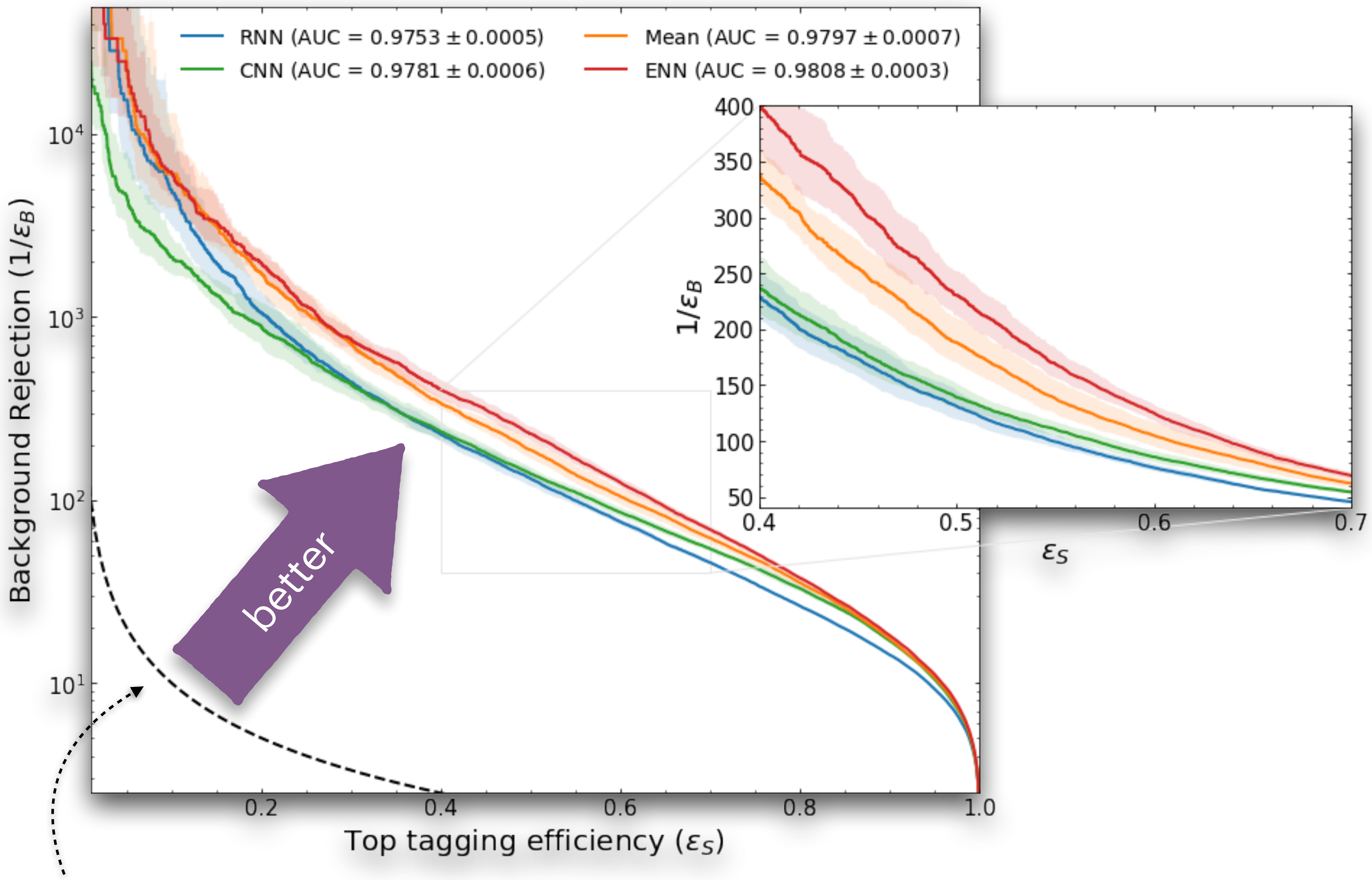
Egan, Fedorko, Lister, Pearkes, Gay '17

Louppe, Cho, Becot, Cranmer JHEP '19

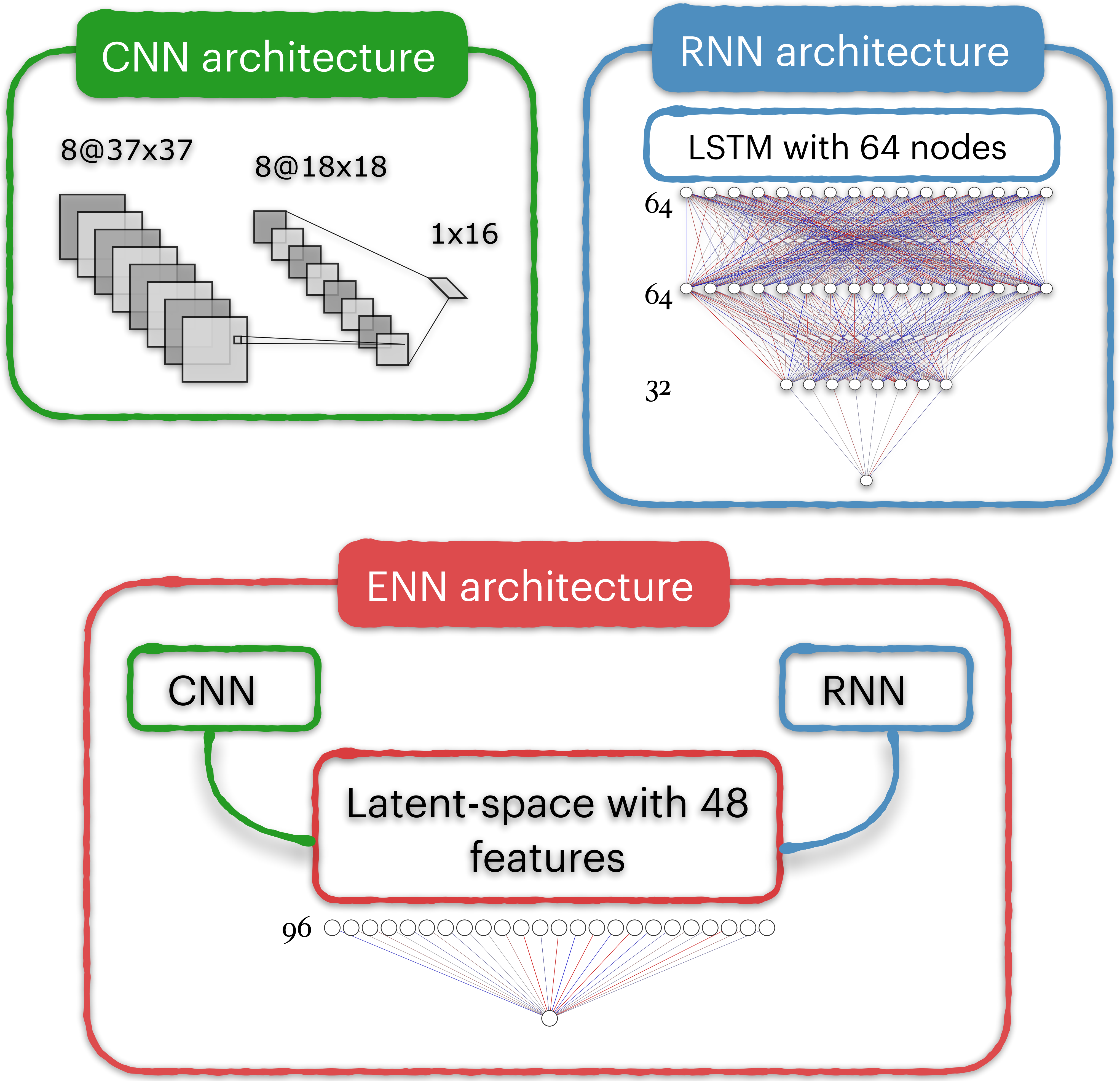
- Leading FatJet Definition: Cambridge algorithm with $R = 0.8, p_T \in [550,\ 650]$ GeV, $|\eta| < 2$
- Tree leading branches are selected alongside with fatjet mass and mass-drop tagger reconstructed fatjet mass.



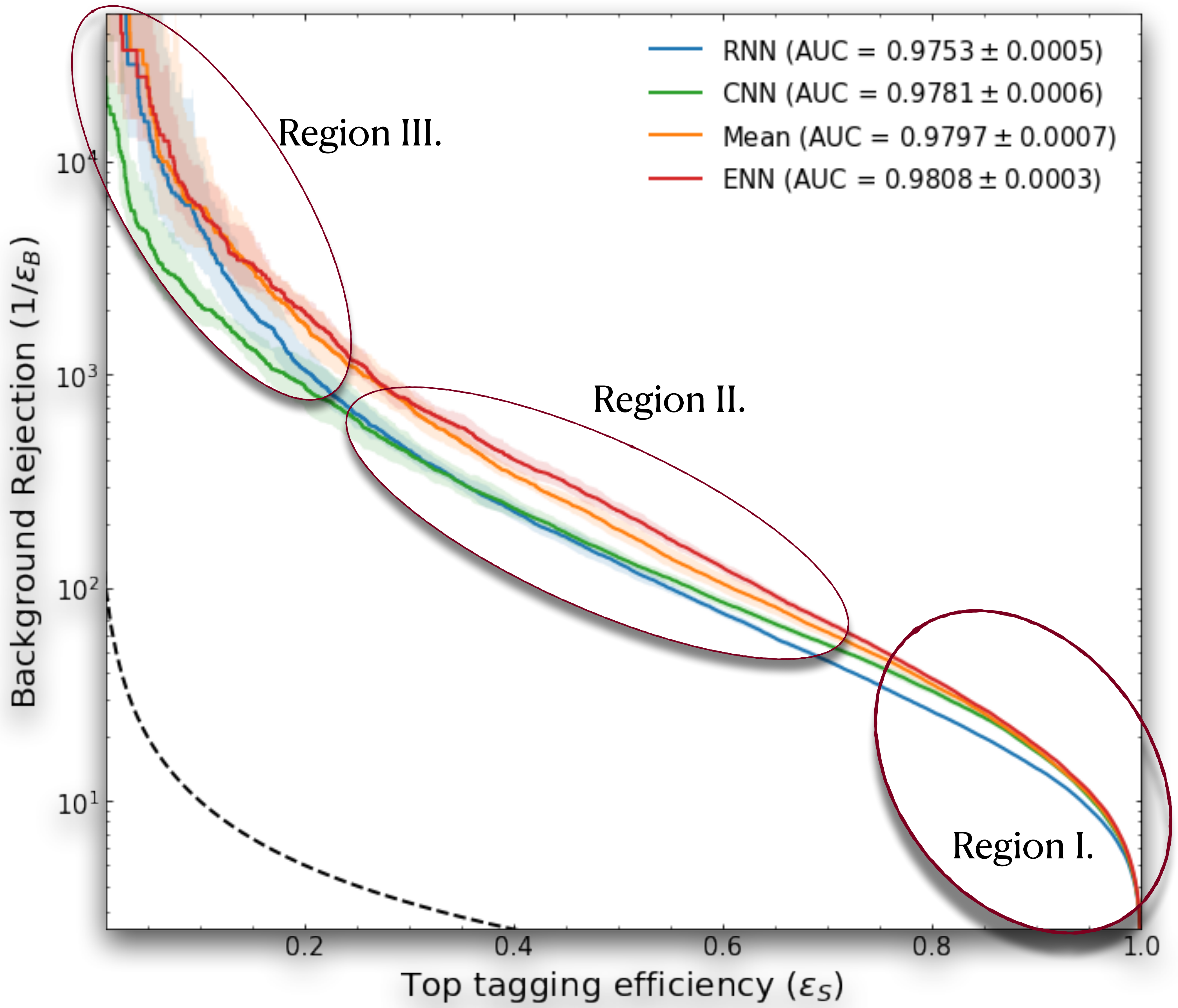
Top Tagging Through Ensemble Learning



Random choice

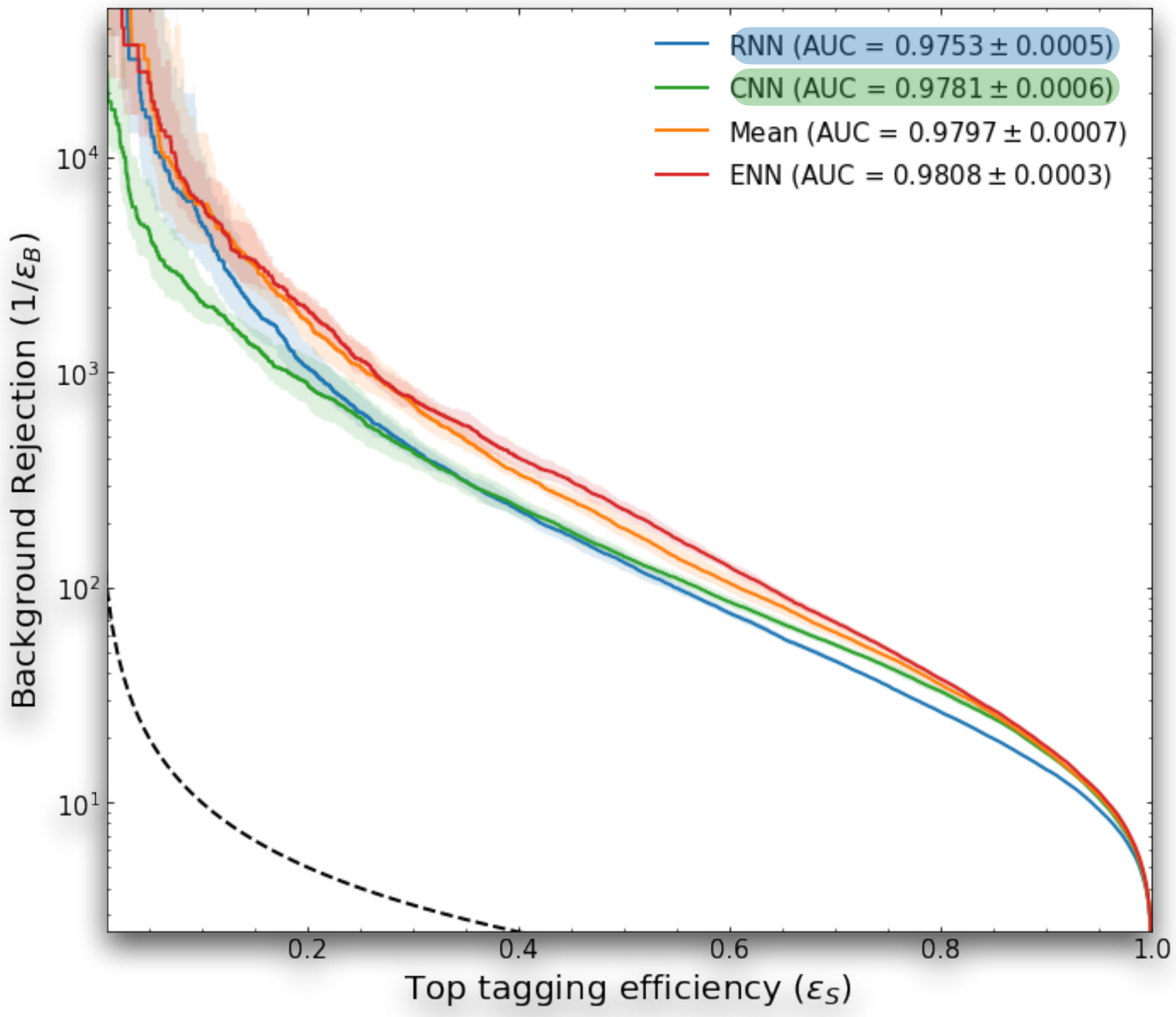


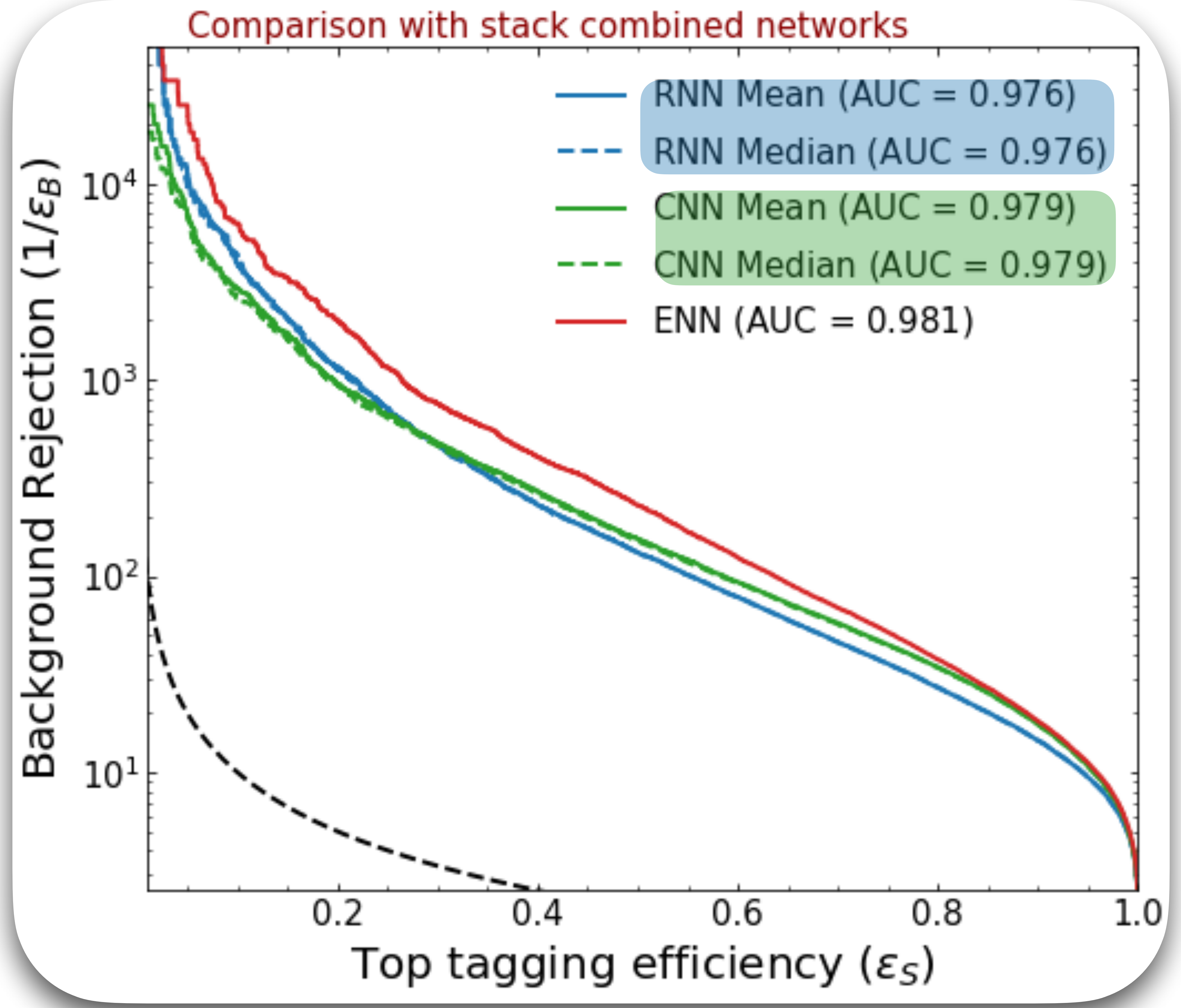
Top Tagging Through Ensemble Learning



- Region I: Both RNN & CNN captures 3-prong substructure
- Region III: RNN & CNN captures dipole type substructure
- Region II: Mixed information coming from both RNN & CNN.

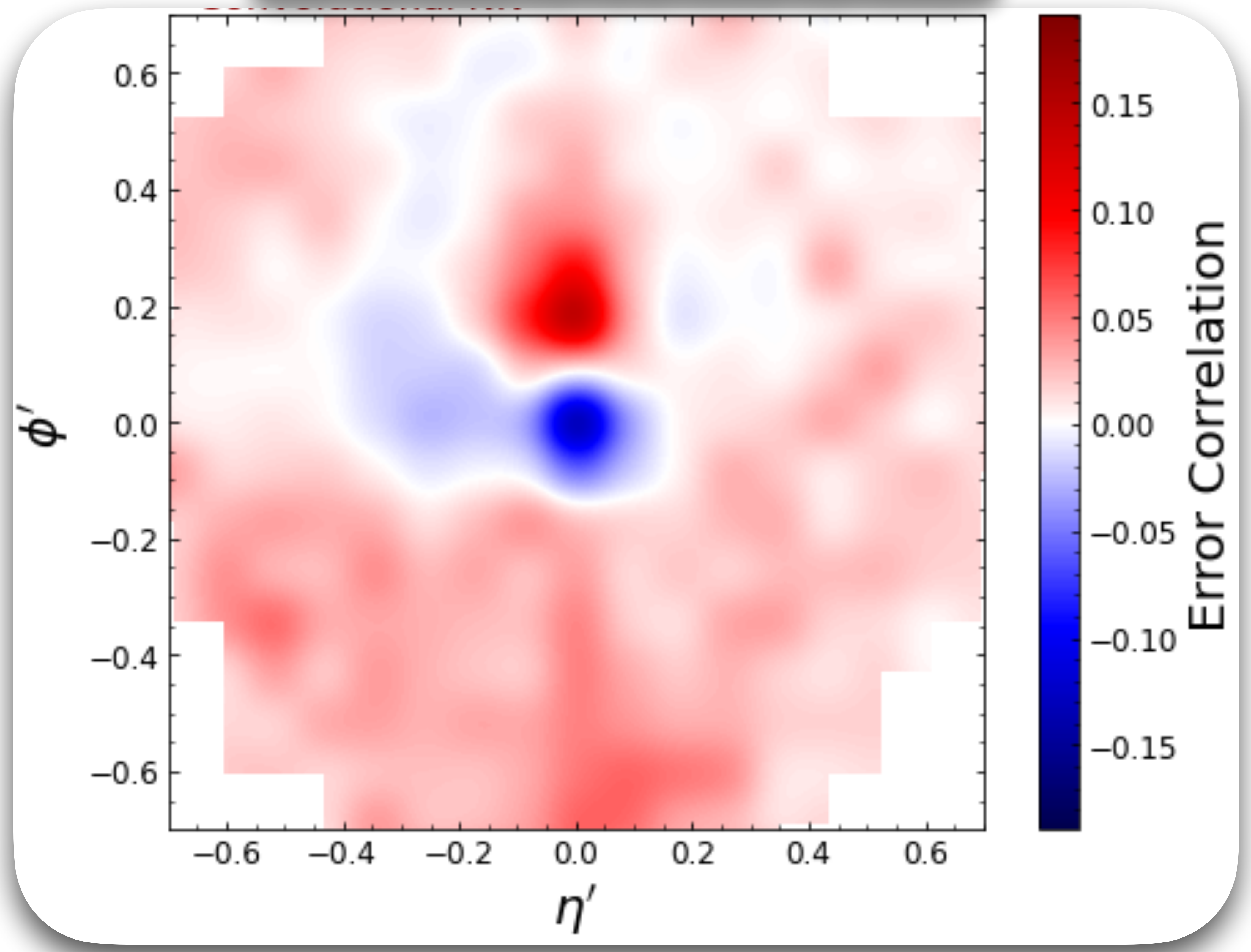
Top Tagging Through Ensemble Learning



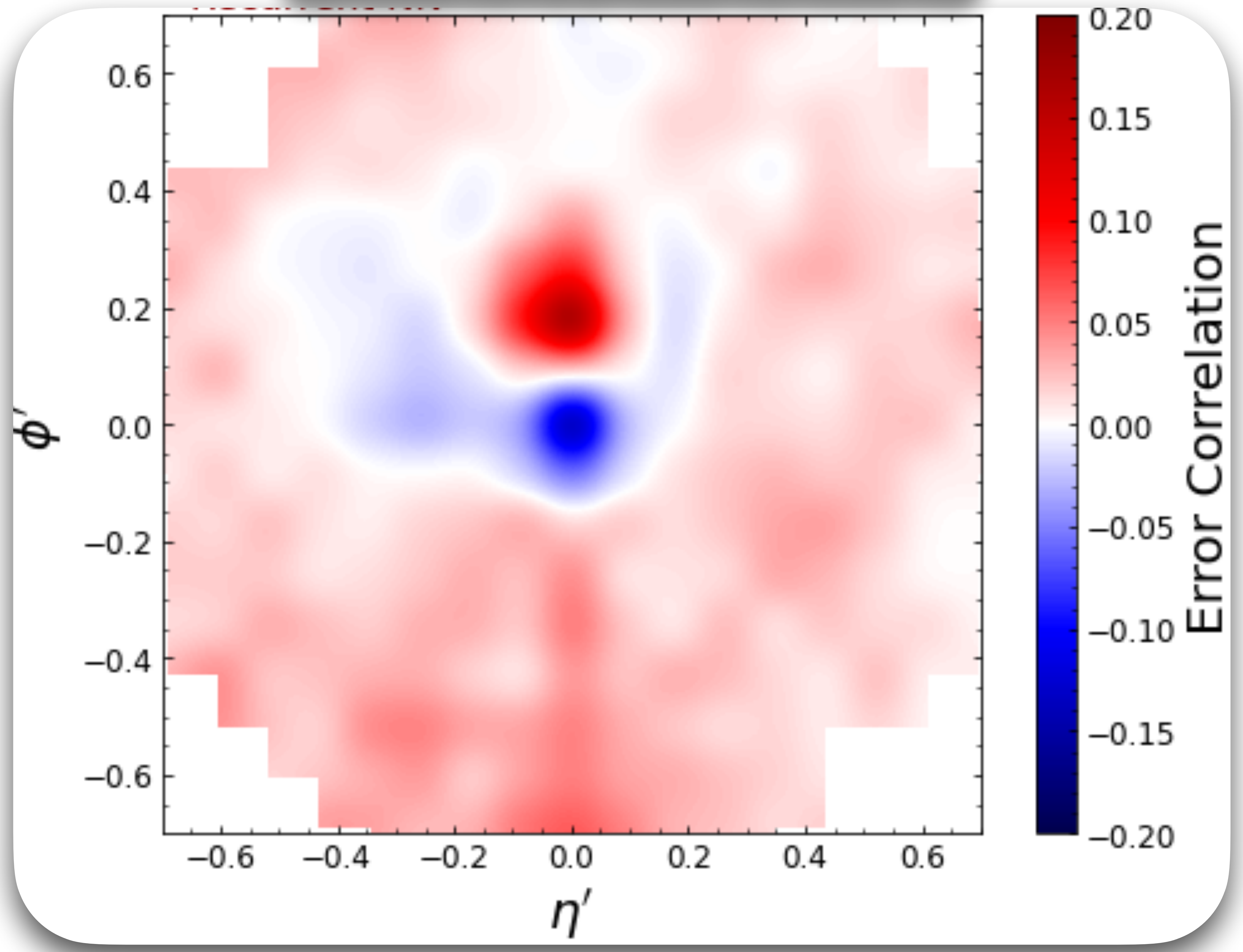


Top Tagging Through Ensemble Learning

Convolutional NN

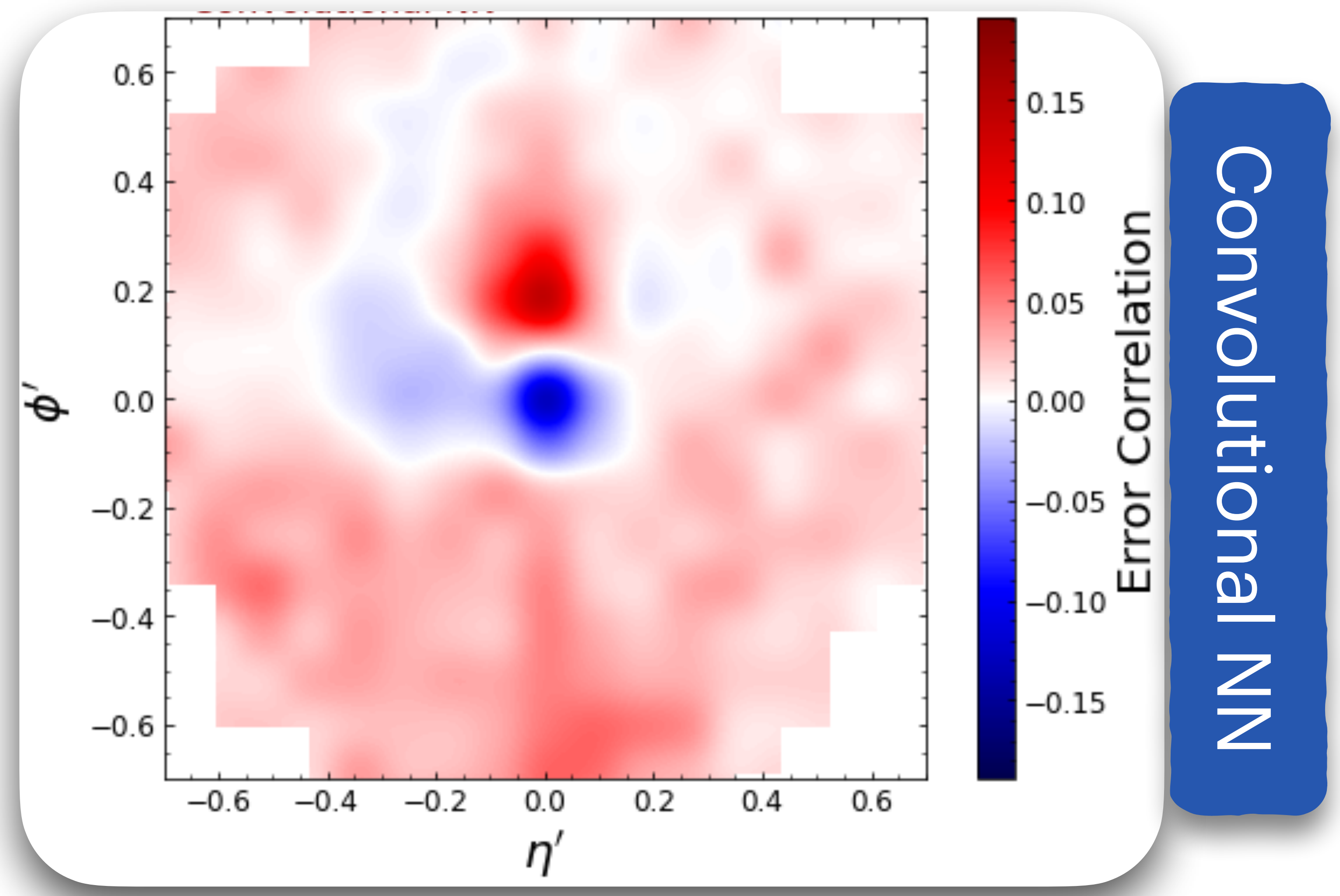


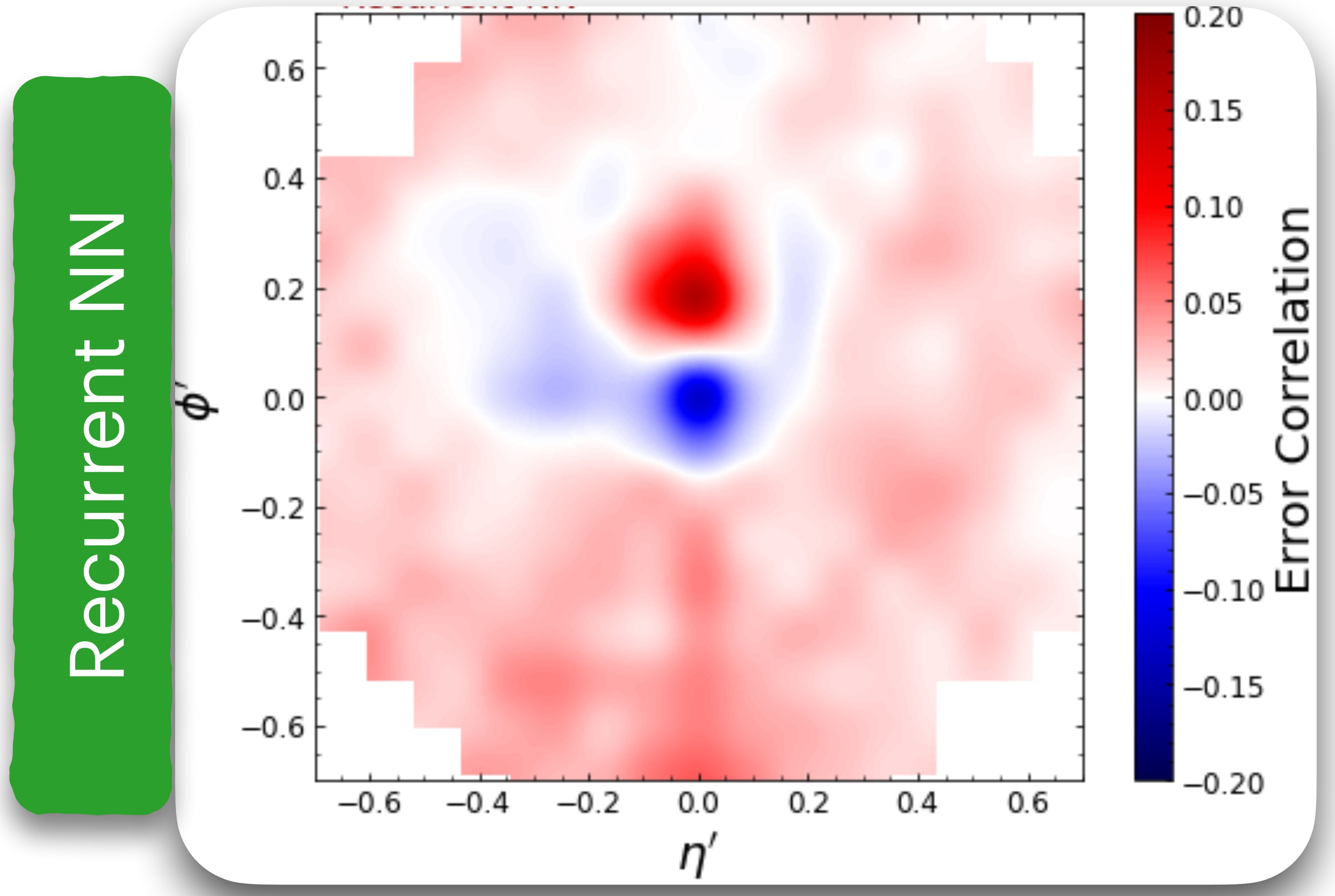
Recurrent NN

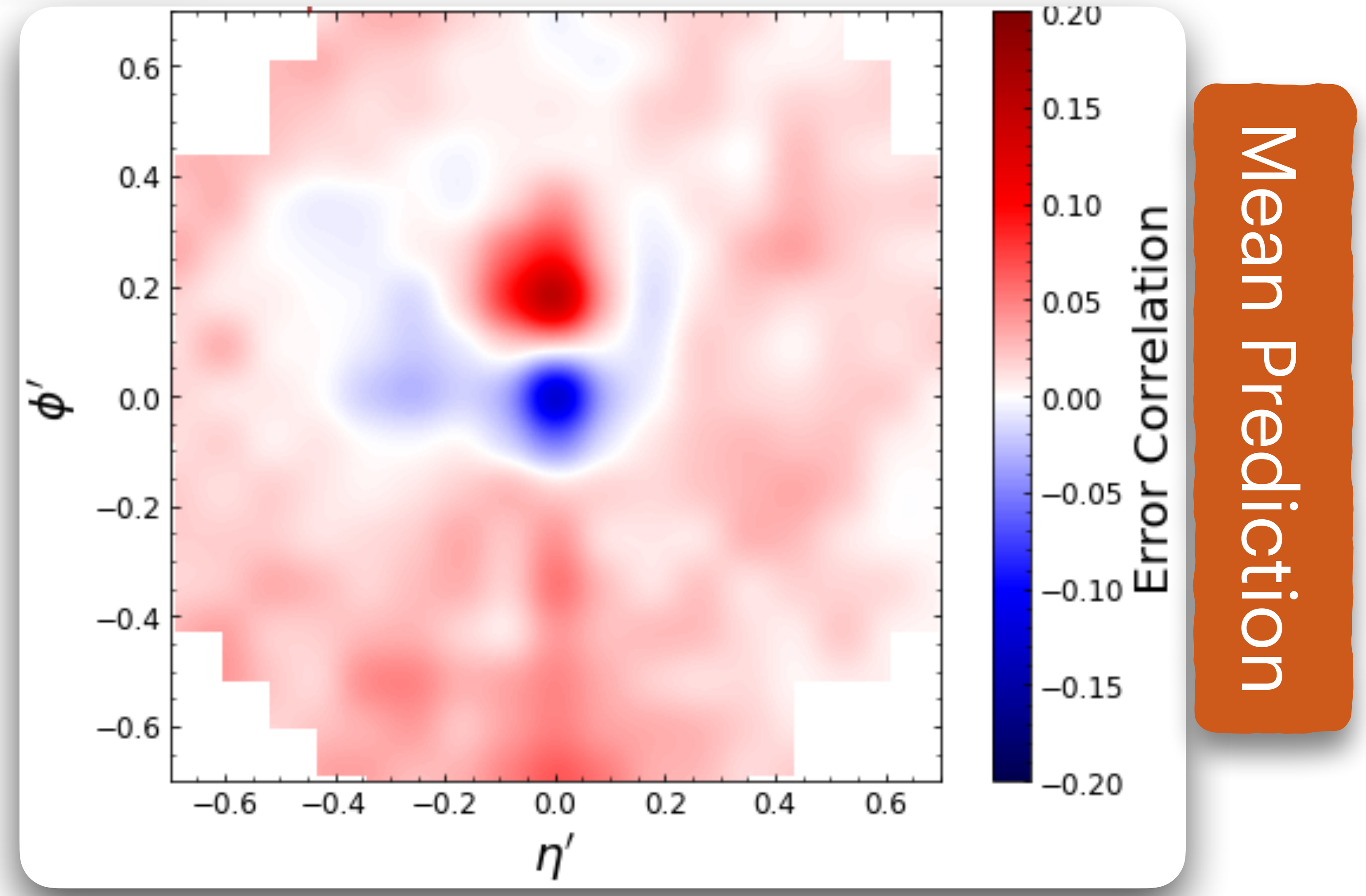


Each pixel's correlation with squared error: $(\hat{y} - y_{truth})^2$

Top Tagging Through Ensemble Learning

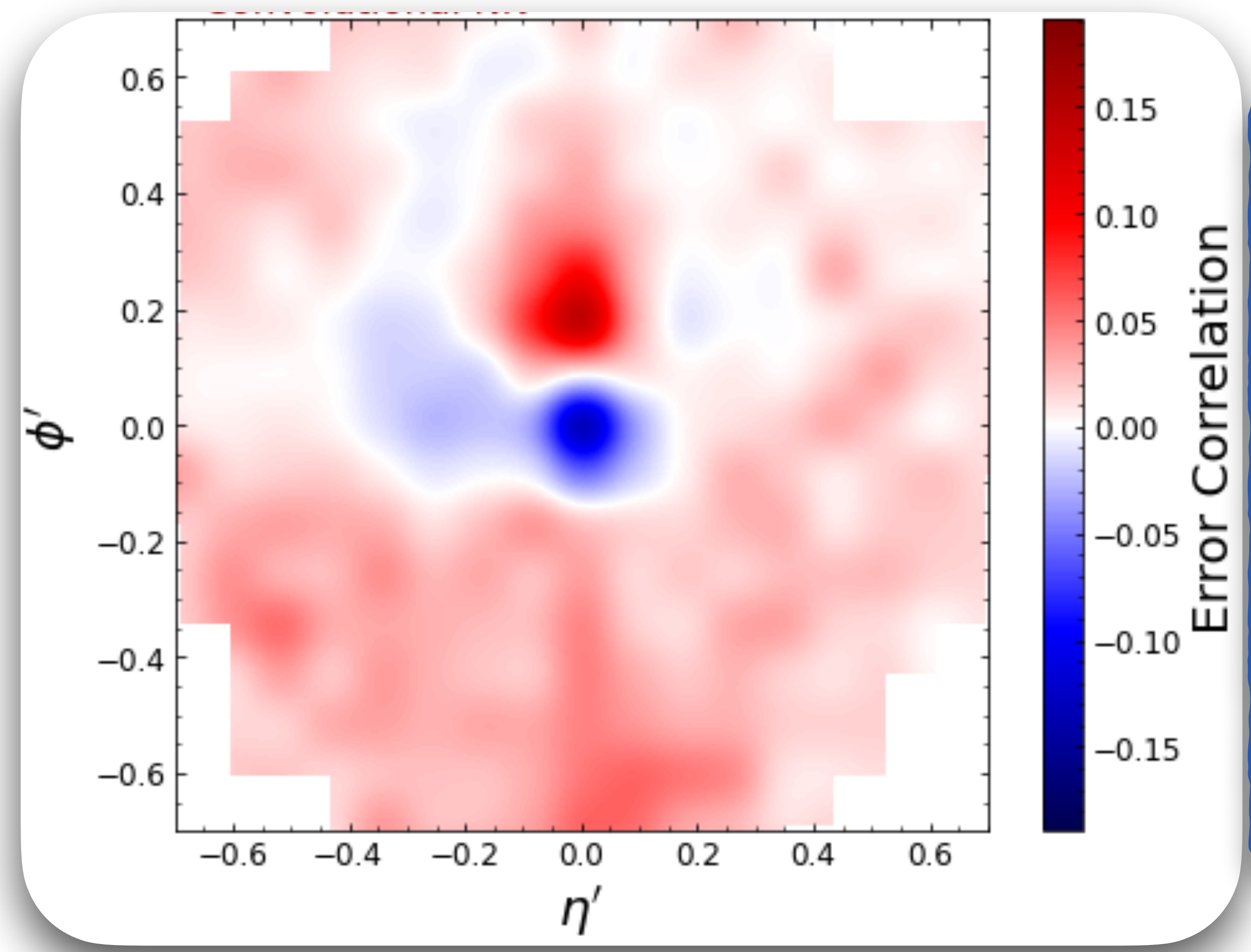






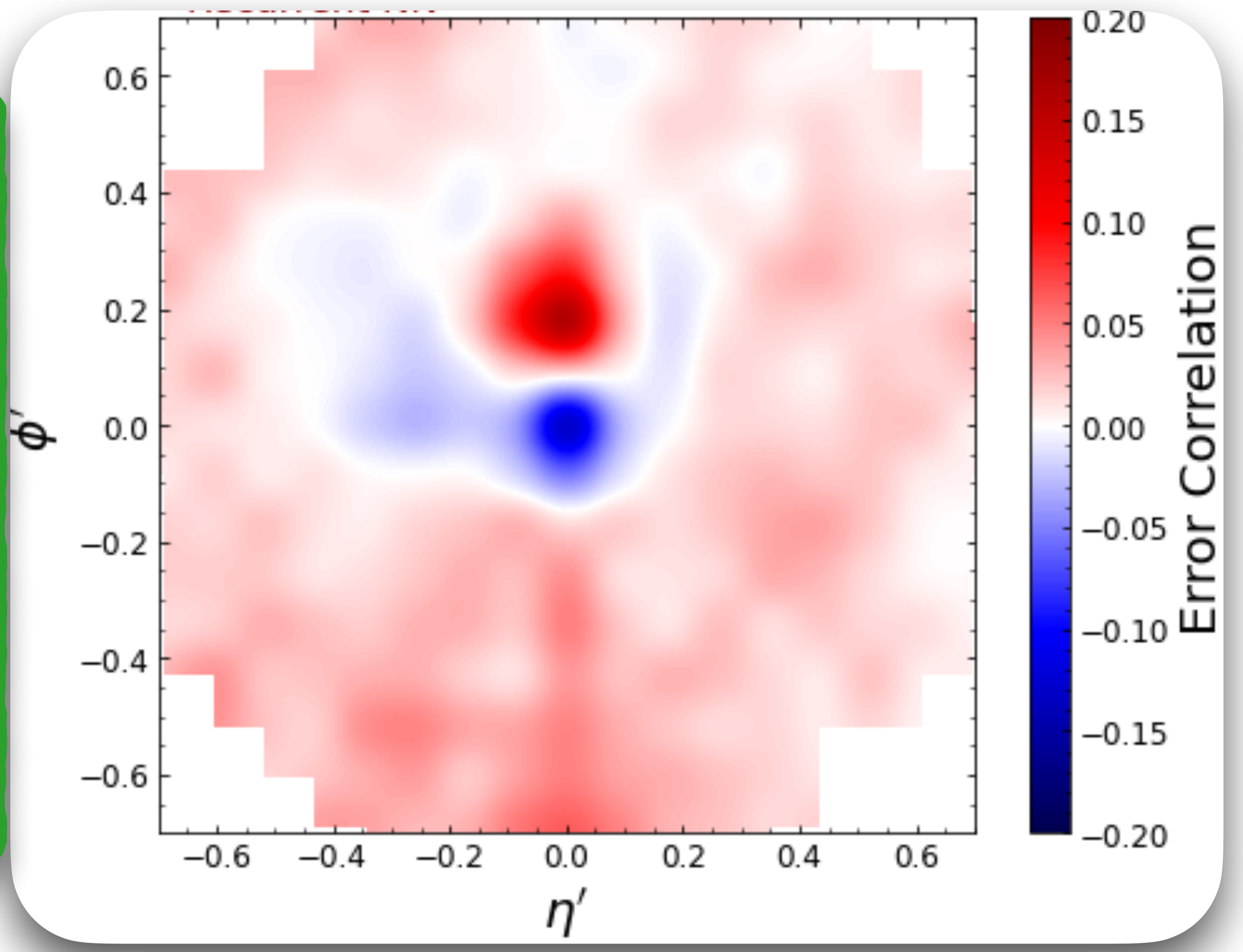
Each pixel's correlation with squared error: $(\hat{y} - y_{truth})^2$

Top Tagging Through Ensemble Learning

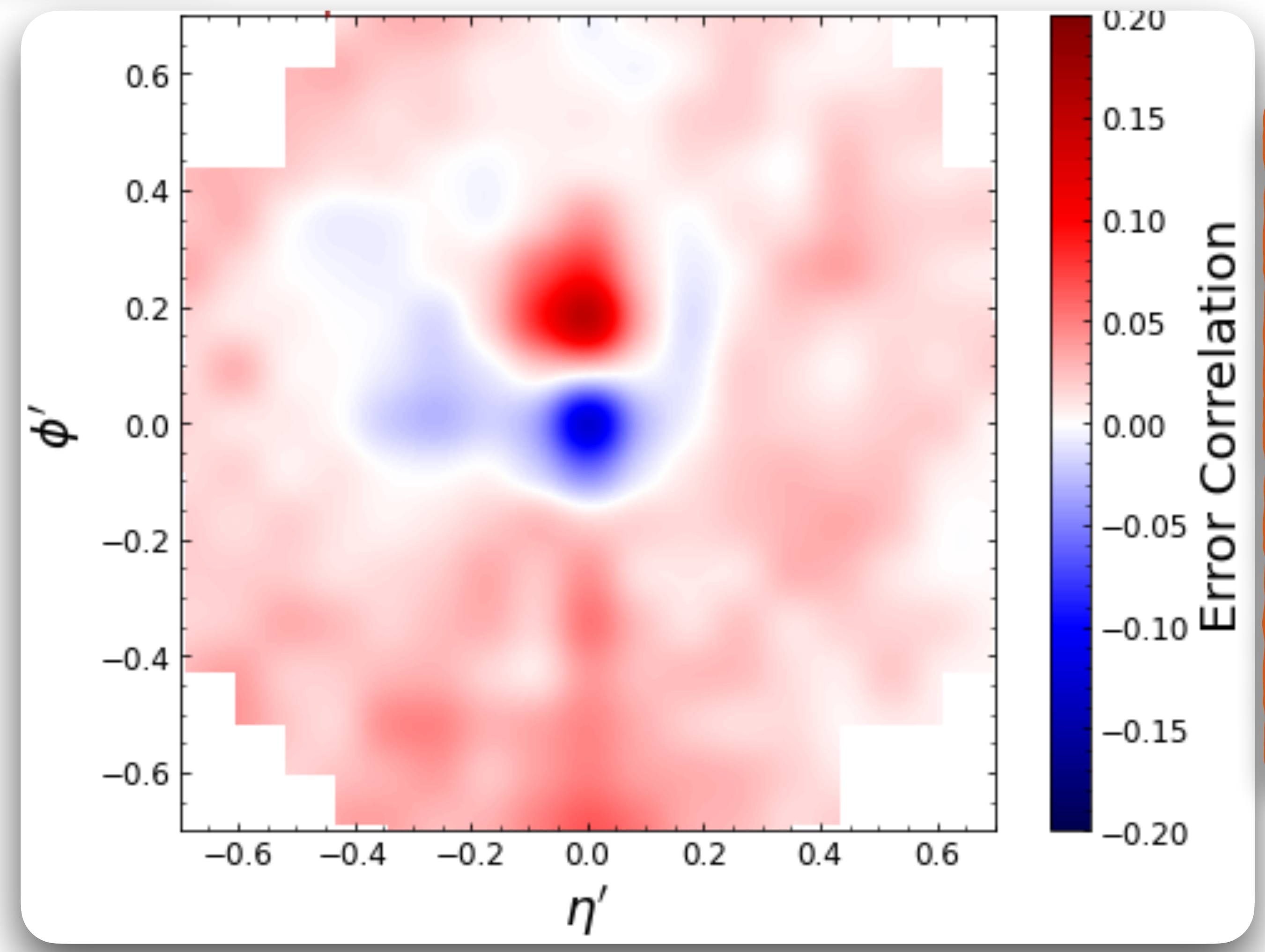


Convolutional NN

Recurrent NN

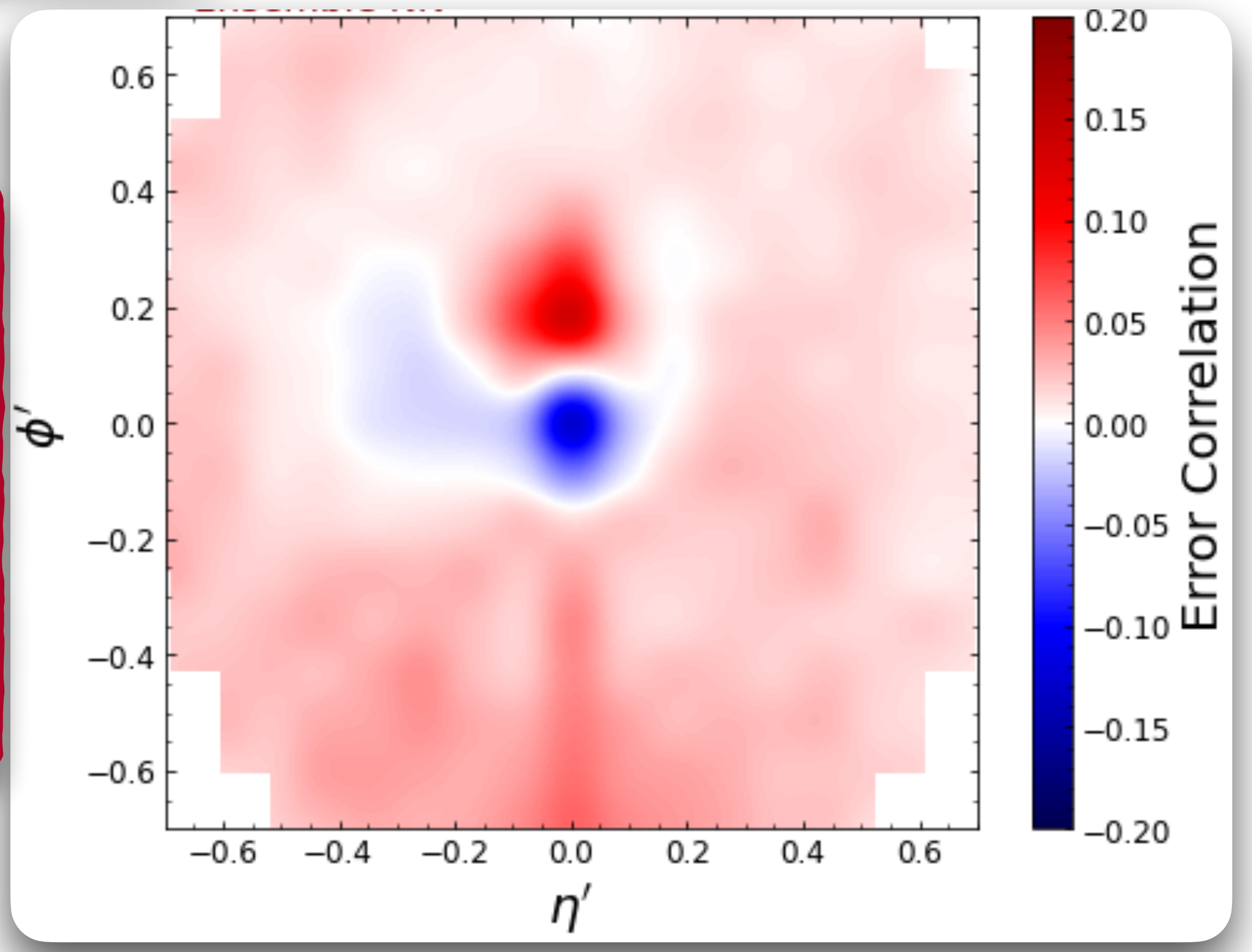


Each pixel's correlation with squared error: $(\hat{y} - y_{truth})^2$



Mean Prediction

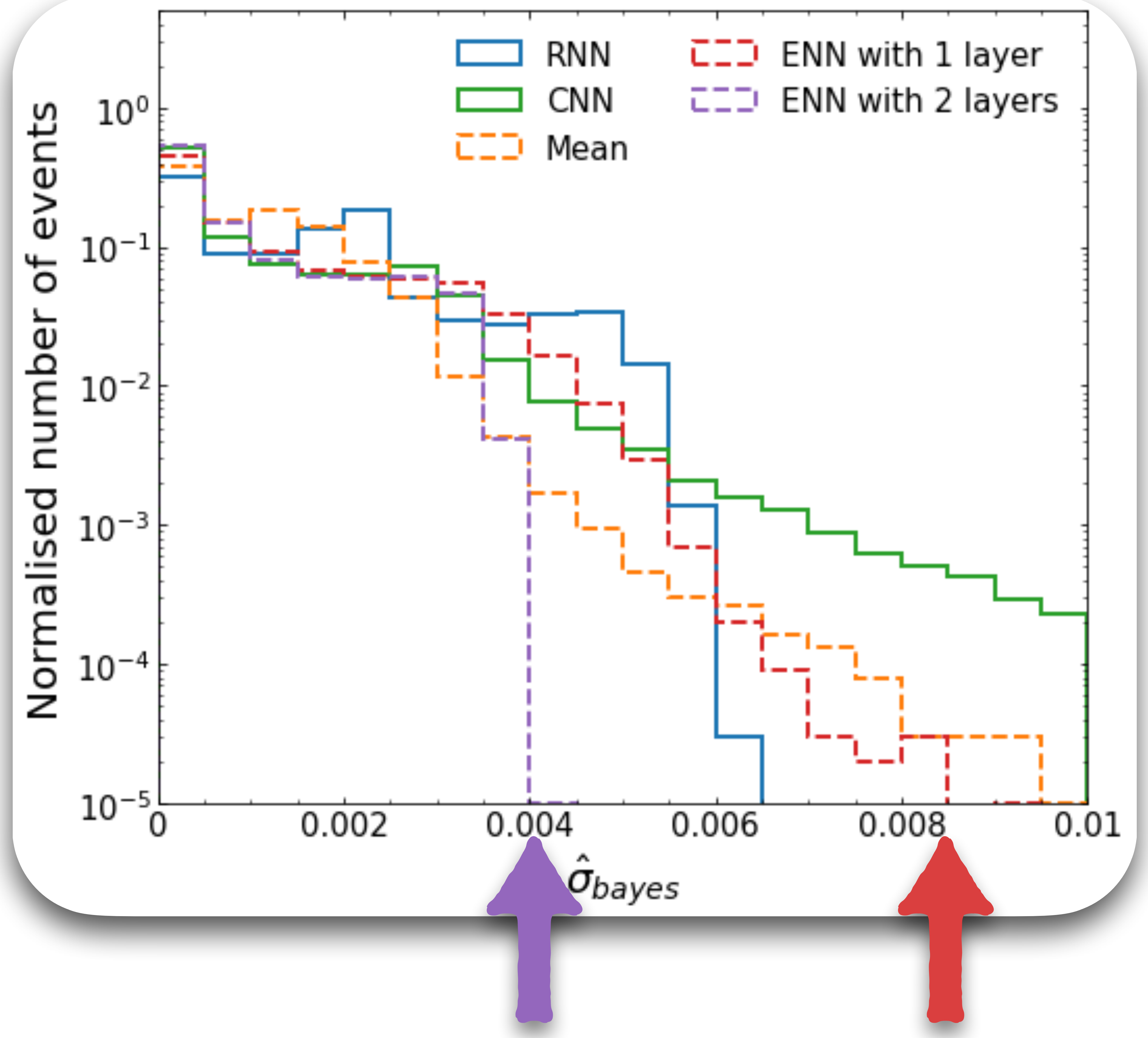
Ensemble NN



Improving Uncertainties with Ensemble Networks

Improving uncertainties with Ensemble Networks

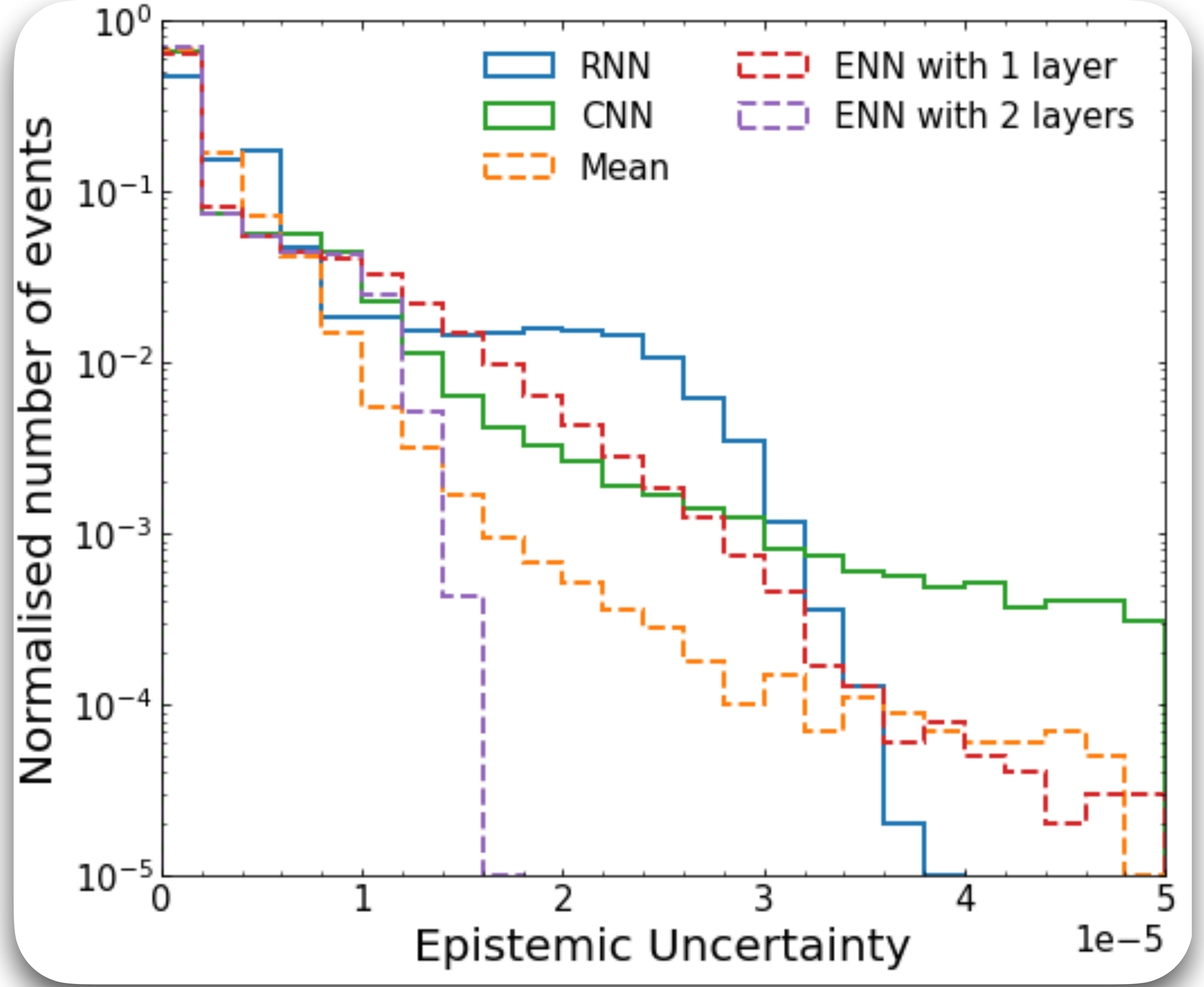
- Using Bayesian methods, one can estimate the deviation in the prediction of the network.
- The last layer of each network has been converted to a Bayesian layer to estimate the deviation in the prediction.
- Each network has been sampled 100 times with 100,000 test sample.



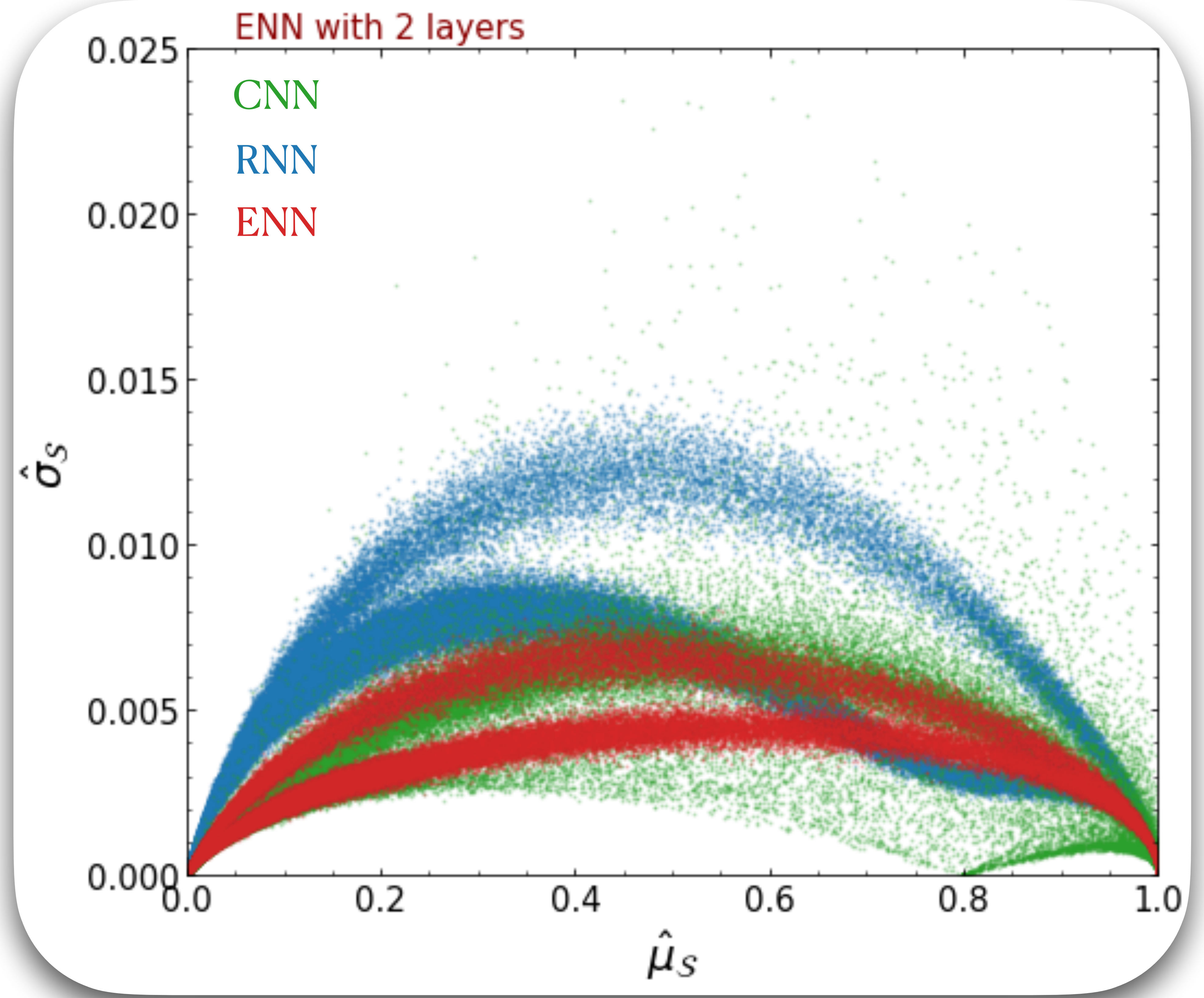
Improving uncertainties with Ensemble Networks

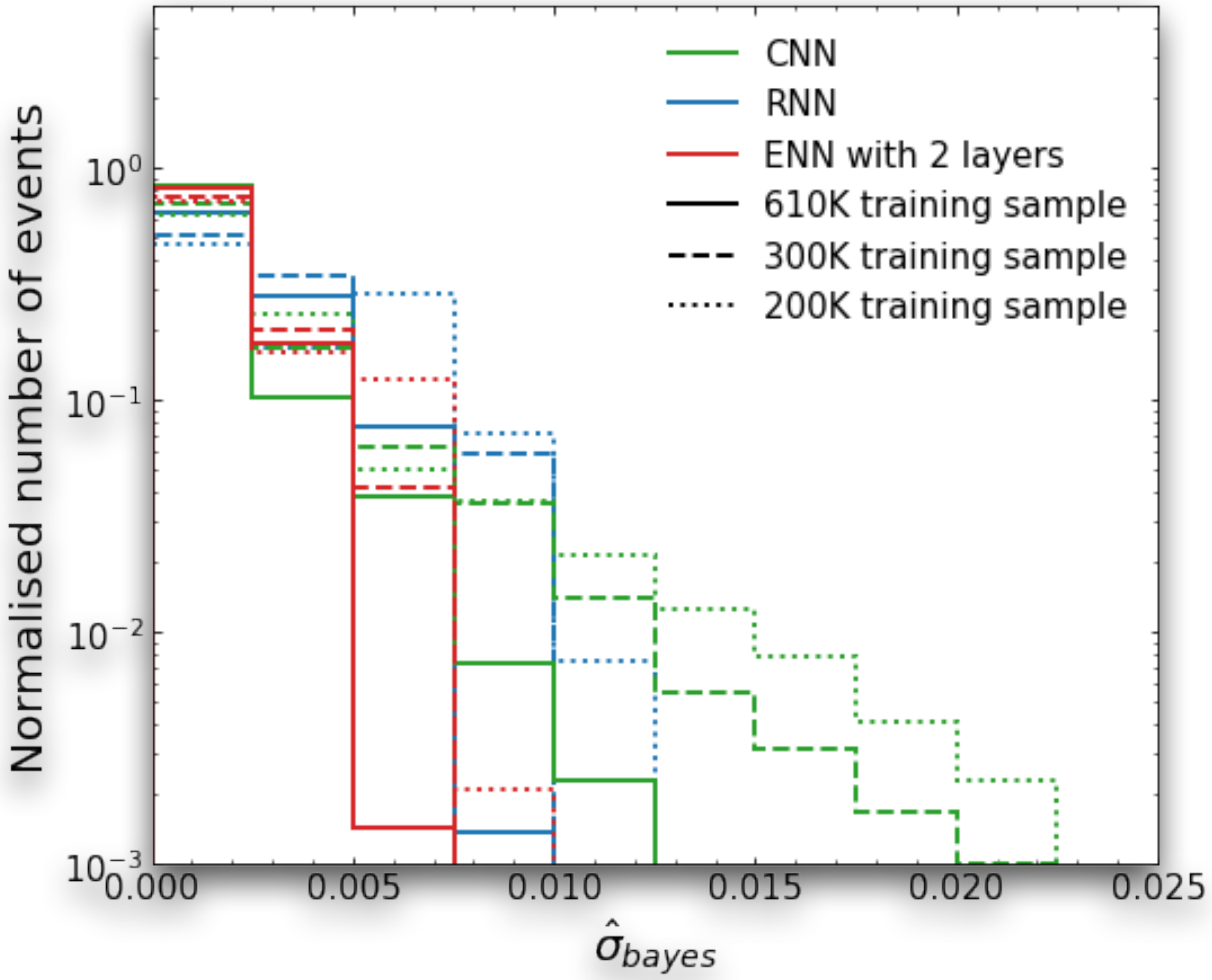
- Epistemic Uncertainties: The uncertainties intrinsic to the proposed hypothesis.
- Aleatoric Uncertainties: The irreducible noise in the observations.

$$Var(y) = \underbrace{\langle \hat{y}^2 \rangle - \langle \hat{y} \rangle^2}_{\text{epistemic}} + \underbrace{\langle \; \hat{y} \; (1 - \hat{y}) \; \rangle}_{\text{aleatoric}}$$



Improving uncertainties with Ensemble Networks





$$\mathcal{S} = -\left(\hat{y}\log_2(\hat{y}) + (1-\hat{y})\log_2(1-\hat{y})\right)$$

Conclusion

Conclusion

Summary

- Ensemble methods provide a framework to expand the representation of a network hypothesis and allow more robust classification.
- It improves the classification performance by combining component error correlations.
- It goes beyond stack combining techniques by optimizing over joint latent-space.
- It reduces the epistemic uncertainties and less susceptible to lack of training samples.

Discussion

- The parallel combining method does not render stack combining invalid. Each method is designed to solve a particular problem.
- ENN methods do not reduce bias or aleatoric uncertainties. Genetic-Algorithm-based Selective Ensembles might be the key to ultimate optimization. Zhou, Wu, Tang AI 2002

Conclusion

Summary

- Ensemble methods provide a framework to expand the representation of a network hypothesis and allow more robust classification.
- It improves the classification performance by combining component error correlations.
- It goes beyond stack combining techniques by optimizing over joint latent-space.
- It reduces the epistemic uncertainties and less susceptible to lack of training samples.

Discussion

- The parallel combining method does not render stack combining invalid. Each method is designed to solve a particular problem.
- ENN methods do not reduce bias or aleatoric uncertainties. Genetic-Algorithm-based Selective Ensembles might be the key to ultimate optimization. Zhou, Wu, Tang AI 2002