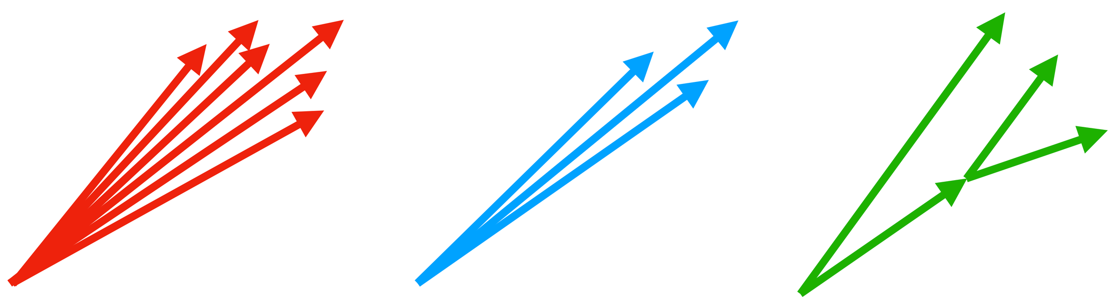
The Pareto Frontier of Resilient Jet Tagging

Rikab Gambhir (Cincinnati/IAIFI), Matt LeBlanc (Brown/IAIFI), Yuanchen Zhou (Brown)

BROWN Department of Physics

https://arxiv.org/abs/2509.19431, https://zenodo.org/records/16986897

Jet Tagging at the LHC



- **Jets** are collimated showers of particles from strongly-interacting quarks and gluons produced in high-energy particle collisions at colliders like Large Hadron Collider (LHC).
- **Jet tagging is the task to identify the particle that initiated the corresponding jet.** It is widely performed in LHC searches and measurements using AI/ML methods, but it is also challenging due to differences in jet fragmentation modeling.
- **We explore trade-offs between tagger classification power and model-dependence, and advocate for a holistic approach to algorithm design for collider physics tasks.**

Resilient jet tagging

- **Jet Taggers:**
  - AI/ML models are trained on Monte Carlo (MC) simulated events. (*e.g.* PYTHIA, HERWIG)
  - The accuracy is evaluated by the AUC value.
- **Resiliency:**
  - Classifier performance depends on the model used to generate the training set.
  - In addition to the accuracy, the uncertainty across different simulations of a model also worth the attention.

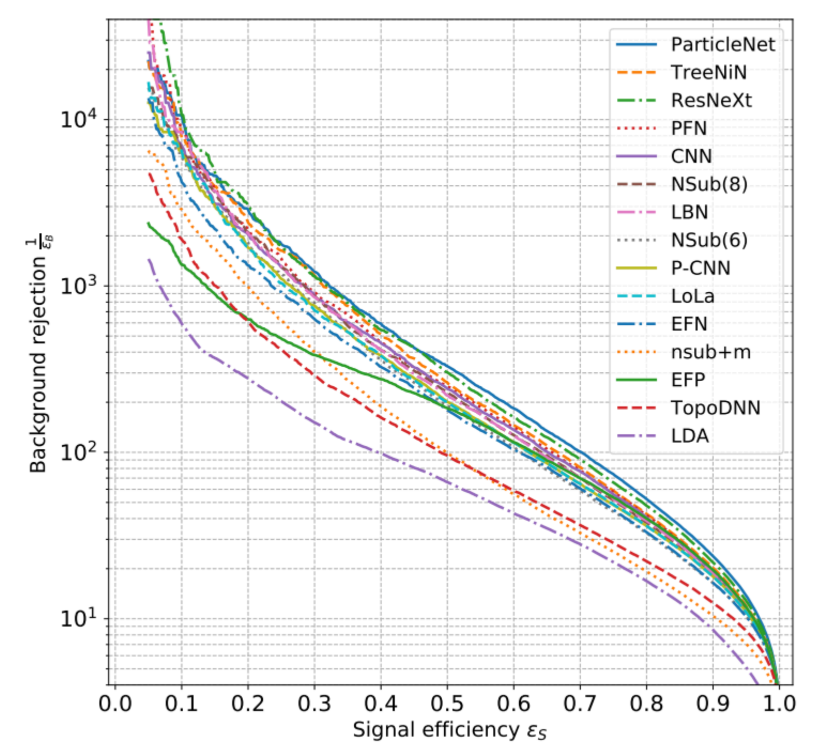
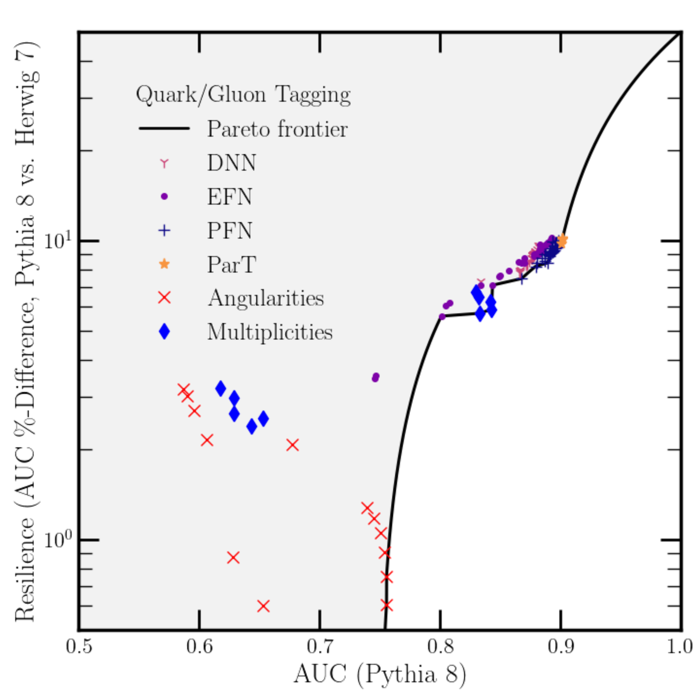


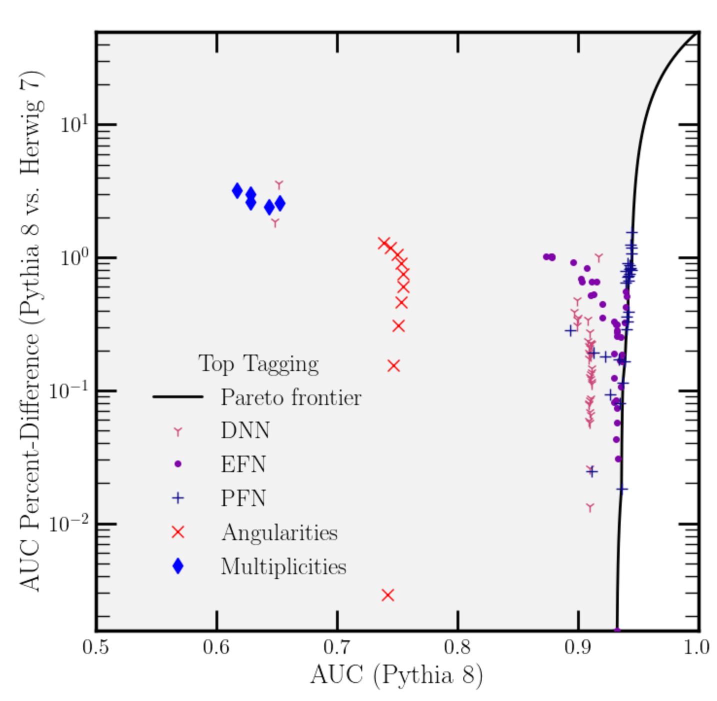
Figure from Ref. 3: Classifiers with larger AUC also tend to be more complex, and are found to be less resilient in this study.

  - Resilience, a proxy for model-dependent uncertainty, is evaluated as the percentage difference in AUC scores when predicting on an alternative MC simulation.
- **Tradeoff & Pareto Frontier**
  - More complex models tend to perform better in terms of accuracy, but will also be less resilient.
  - The Pareto frontier indicates the best combinations of uncertainty and accuracy a tagger can achieve.
- **Is it possible to land a model further to the bottom right, beating the current Pareto frontier?**

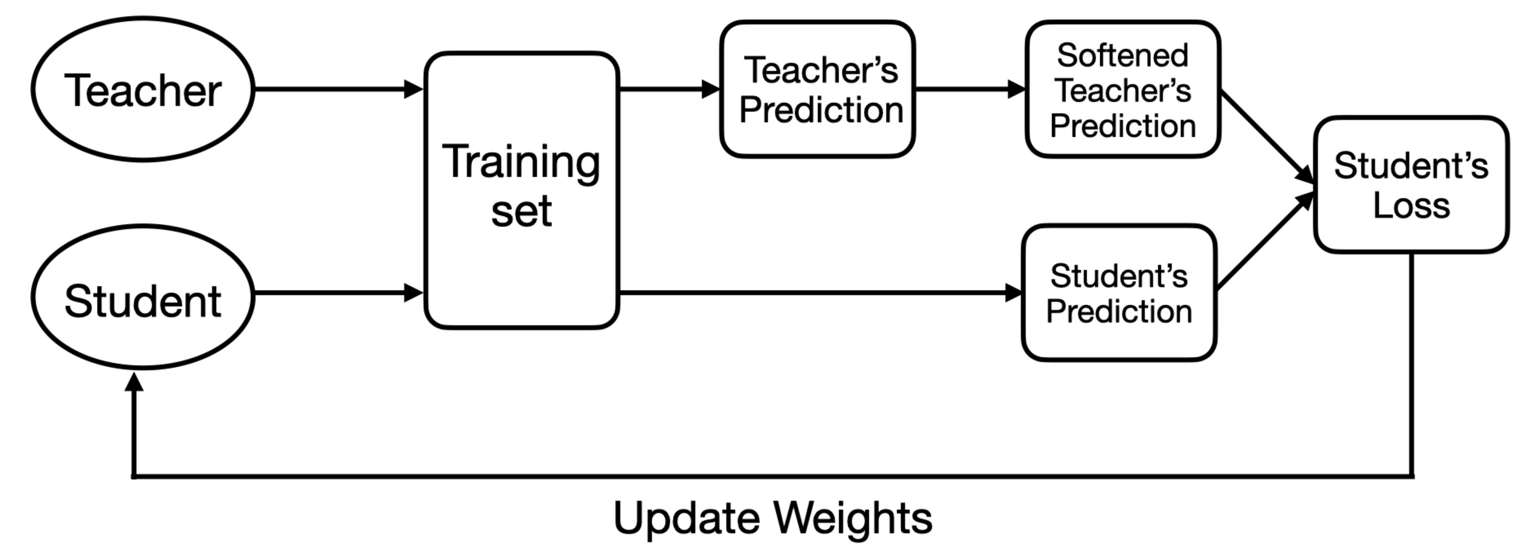
Q/G Tagging Frontier



Top Tagging Frontier

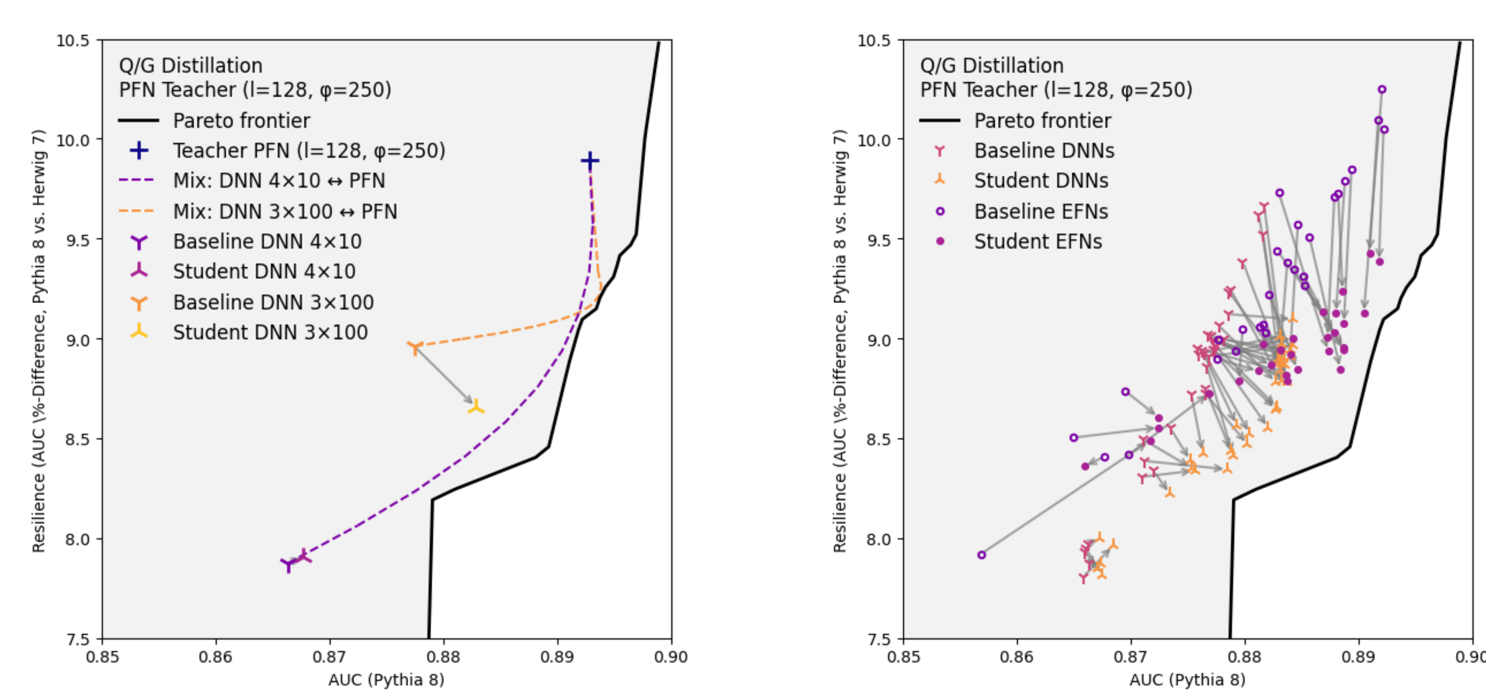


Knowledge Distillation



- **Knowledge Distillation (KD):**
  - A simpler 'student' model (DNN, EFN) is trained with the softened outputs of a pre-trained 'teacher' model that is more complex (PFN) instead of on the true labels.
- **Effects of KD:**
  - The student model is compared to the teacher model and a corresponding 'simple' model with the same structure as the student that is trained in a fully-supervised manner.
  - Most student models are more accurate and more more resilient than the corresponding simple models, approaching the Pareto frontier.
  - While many students improve, none surpass the frontier.
  - There may be other benefits from KD, *e.g.* using less complex models can reduce inference time [Ref. 4].

Students Compared to Simple & Teacher Models



Does using resilient networks matter?

Case Study: Determining Quark/Gluon Fraction

		Pythia 8	*Herwig 7*		
Classifier	**True κ**	**Inferred $\mathbb{E}[\hat{\kappa}]$**	**Inferred $\mathbb{E}[\hat{\kappa}]$**	**Calibrated $\mathbb{E}[\hat{\kappa}]$**	**Result**
Large PFN	0.50	0.490 ± 0.005	0.296 ± 0.007	0.529 ± 0.006	Biased ✗
	0.25	0.253 ± 0.005	0.125 ± 0.005	0.305 ± 0.006	Biased ✗
Small PFN	0.50	0.504 ± 0.013	0.336 ± 0.016	0.478 ± 0.017	Unbiased ✓
	0.25	0.258 ± 0.013	0.157 ± 0.014	0.268 ± 0.013	Unbiased ✓

- κ indicating the fraction of quark jets in the samples.
  - PYTHIA and HERWIG samples with 25% and 50% of quarks are construct.
- A large PFN and a small PFN are both trained on PYTHIA jets, both taken from the Pareto frontier.

- **On PYTHIA Samples,** both are able to recover the mixture fraction within 2σ.
- **On HERWIG Samples**, both are biased.
- **On Calibrated HERWIG Samples,** the more resilient small PFN successfully recovers the mixture fraction while the large PFN is still biased.
  - The q/g tagger is calibrated by reweighing its output with a PYTHIA-HERWIG classifier PFN.
- **A less accurate but more resilient classifier can yield a less biased physics result in an actual analysis!**

References

1. Arianna Garcia Caffaro, Ian Moult, and Chase Shimmin. Energy-Energy Flow Networks. 2025. arXiv: 2510.06314 [hep-ph]. url: https://arxiv.org/abs/2510.06314.
2. ATLAS Collaboration, "Accuracy versus precision in boosted top tagging with the ATLAS detector". In: Journal of Instrumentation 19.08 (Aug. 2024), P08018. issn: 1748-0221. doi: 10.1088/1748-0221/19/08/p08018. url: http://dx.doi.org/10.1088/1748-0221/19/08/P08018.
3. Gregor Kasieczka et al. "The Machine Learning landscape of top taggers". In: SciPost Physics 7.1 (July 2019). issn: 2542-4653. doi: 10.21468/scipostphys.7.1.014. url: http://dx.doi.org/10.21468/SciPostPhys.7.1.014.
4. Vinicius Mikuni and Benjamin Nachman. "CaloScore v2: single-shot calorimeter shower simulation with diffusion models". In: Journal of Instrumentation 19.02 (Feb. 2024), P02001. issn: 1748-0221. doi: 10.1088/1748-0221/19/02/p02001. url: http://dx.doi.org/10.1088/1748-0221/19/02/P02001.

Preprint

Top Samples (New!)