Resolving Extreme Jet Substructure

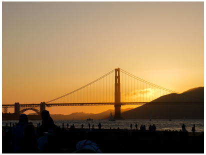
Boost 2023

arXiv:2202.00723

Yadong Lu, Alexis Romero, **Michael James Fenton**, Daniel Whiteson, Pierre Baldi

University of California, Irvine
mjfenton@uci.edu

August 2nd 2023
Berkeley, California



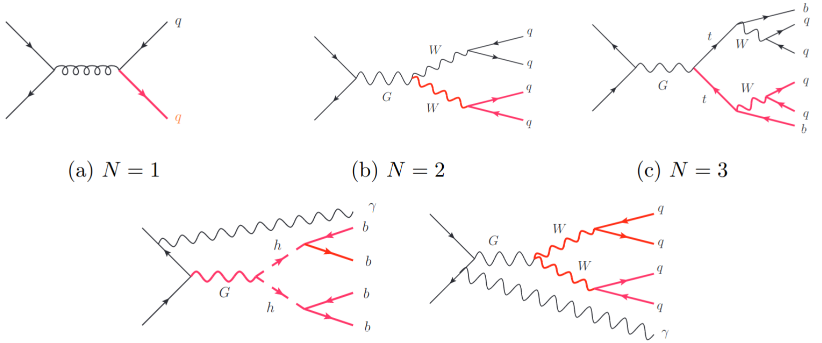
Motivation

- We know that jet taggers based on low level info usually can outperform high level features, at least for W/Z/top jets
  - But these low level taggers are hard to calibrate and use in practice
- Most large radius jet tagging studies have focused on those with $N = 2, 3$ hard subjets (ie W/Z/H or top)
  - are existing methods enough in extreme ($N > 3$) conditions to do jet tagging?
  - can we learn from the "low level" taggers, which directly use jet constituents, some "high level" features that we can use?

I consider myself something of a moral relativist.

- We compare the performance of various taggers, using either using high level observables or low level constituents, and then attempt to bridge the gap between them
- This work contributes to a growing body of work in HEP that uses ML to teach ourselves something, instead of the more usual inverse

Datasets

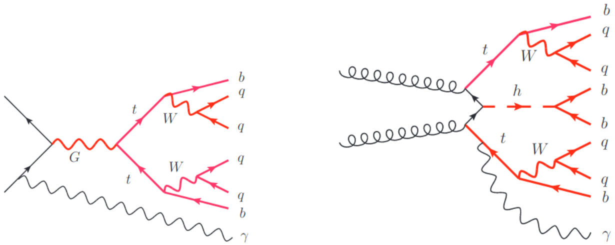


(a) $N = 1$ (b) $N = 2$ (c) $N = 3$

(d) $N = 4b$ (e) $N = 4q$

- We generate 7 classes of dataset that we classify by their ‘pronginess’ (how many hard sub-jets there are) using aMC@NLO+Pythia8+Delphes
  - We use a uniform granularity of 0.0125 in η and ϕ for the calorimeter
- We have two classes for N=4, which we call N=4q and N=4b, to investigate dependence on heavy flavor (more on this later)

Datasets

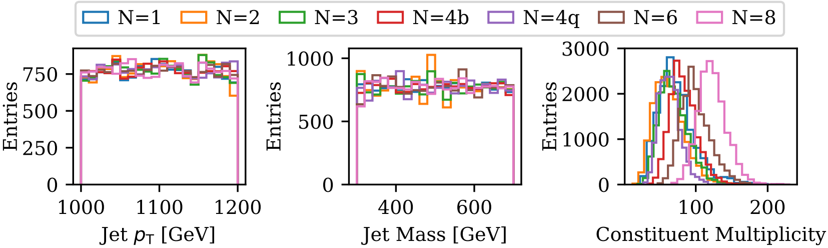


(f) $N = 6$ (g) $N = 8$

- We require a high p_T photon to recoil off the system to give it sufficient boost to be fully captured by a single R=1.2 anti-kT jet
- We also require that all quarks are ΔR matched to the jet
- We use the first 230 input constituents, more than enough for all of the jets, and zero-pad as necessary

Dataset Balancing

- In order to keep performance ∼flat against mass and p_T, we selectively reject events until we have ∼flat distributions
- This ensures that the ML methods do not learn residual effects due to these kinematics



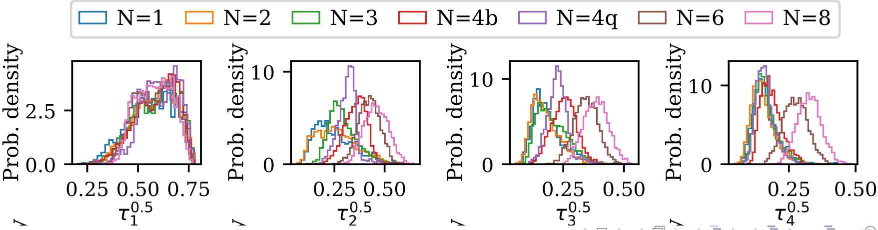
High Level Observables

- We use N-subjettiness ▸ Thaler et al

$$\tau_N = \frac{1}{d_0} \sum_k p_{T,k} \min\left\{\Delta R_{1,k}, \Delta R_{2,k}, \cdots, \Delta R_{N,k}\right\}^{\beta}$$

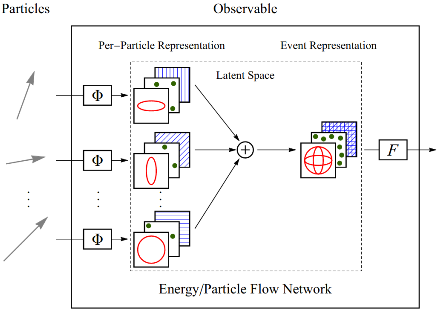
  as a basis of easily interpretable & familiar substructure variables that forms our baseline

- We input a total of 135 of these variables, with $N = 1, ..., 45$, $\beta = \{1/2, 1, 2\}$, as well as jet mass, into a fully connected dense NN, which we label DNN_{136}



Particle Flow Networks ▸ Komiske et al

$$\text{PFN}: F\left(\sum_{i\in\text{jet}} \Phi(p_{\mathrm{T}i}, \eta_i, \phi_i)\right)$$



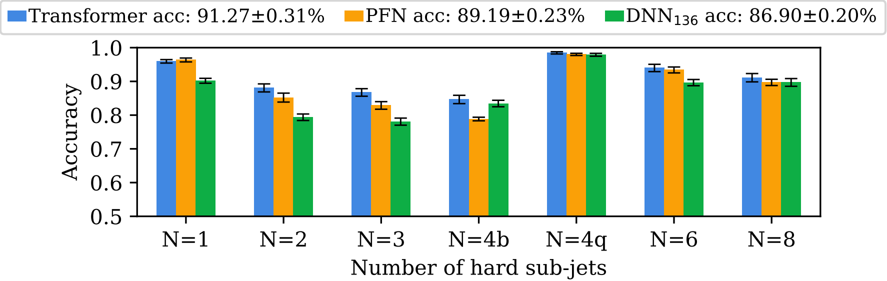
where ϕ is the latent space acting on the constituent 3-vectors and F the jet level latent space

- Based on DeepSets (▸ Zaheer et al), PFNs operate directly on **jet constituents**
- Naturally permutation invariant and capable of handling variable-length sets

Attention Is All You Need ▸ Vaswani et al

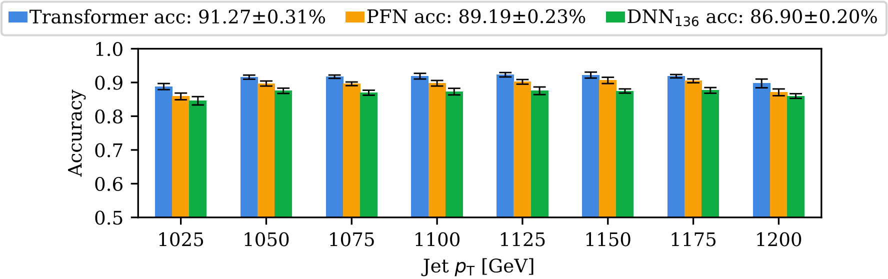
- In NLP, "Attention" mechanisms are now dominant, replacing RNNs / LSTMs in most applications
  - Process sentences as a sequence of words to perform translation, prediction, etc.
  - Reordering the inputs results in the same reordering of the attention matrices: **permutation invariant**
- In our case; input the full sequence of **jet constituents** to multiple self-attention heads in parallel
  - In principle these networks can become large, but we focus on architectures with similar numbers of free parameters ($\mathcal{O}(1M)$) as the other methods
- Transformer based architectures are now state of the art on many tasks in HEP, including event reconstruction ▸ Fenton et al, top-tagging ▸ Qu et al, and pileup mitigation ▸ Maier et al

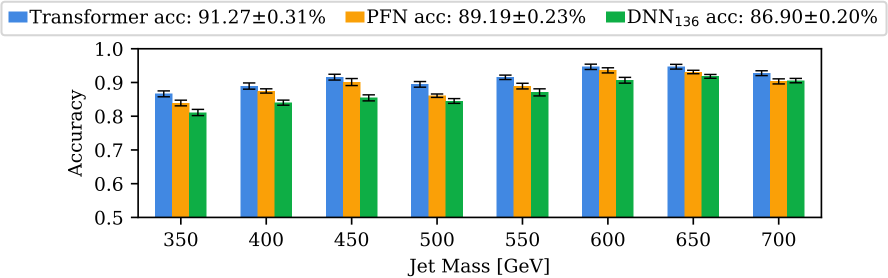
Initial Performance (1/3)



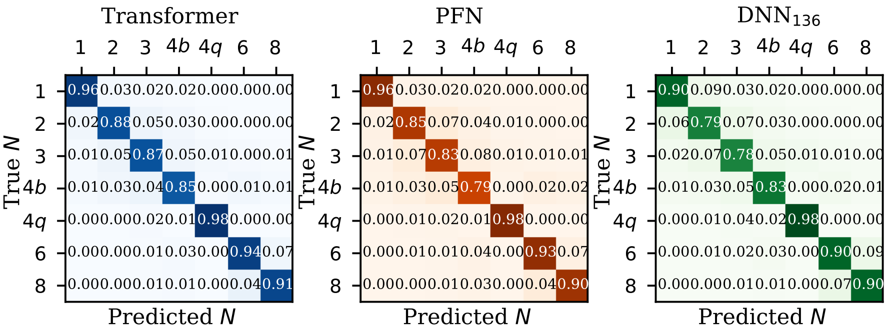
- We see good performance for all categories and a consistent story between architectures
- Transformer > PFN > DNN

Initial Performance (2/3)





Initial Performance (3/3)



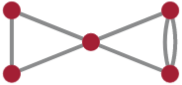
- Confusion matrices; ie which categories are mistaken for each other?
- Perhaps surprisingly, 4b and 4q are confused less often than 4b and 3, suggesting more is being learned than just "pronginess", even for DNN that takes τ variables as input
- Can we understand what information the networks are relying on?

Energy Flow Polynomials ▸ Komiske et al

- A complete basis for jet substructure, easily represented by simple graphs

$$\mathrm{EFP}_G = \sum_{i_1=1}^{M} \cdots \sum_{i_N=1}^{M} z_{i_1} \cdots z_{i_N} \prod_{(k,\ell)\in G} \theta_{i_k i_\ell},$$

ie



$$= \sum_{i_1=1}^{M}\sum_{i_2=1}^{M}\sum_{i_3=1}^{M}\sum_{i_4=1}^{M}\sum_{i_5=1}^{M} z_{i_1} z_{i_2} z_{i_3} z_{i_4} z_{i_5} \theta_{i_1 i_2}\theta_{i_2 i_3}\theta_{i_1 i_3}\theta_{i_1 i_4}\theta_{i_1 i_5}\theta_{i_4 i_5}^2.$$

$z_i = \frac{p_{T_i}}{\sum_i p_{T_J}}$, $\theta_{ij} = (\delta\eta_{ij}^2 + \delta\phi_{ij})$

We further introduce an angular weighting factor $\beta = \{1/2, 1, 2\}$ which we attach to the θ terms for each graph with N=5 or fewer nodes, giving us a total of 162 observables

Bridging the Gap with LASSO ▸ R. Tibshirani

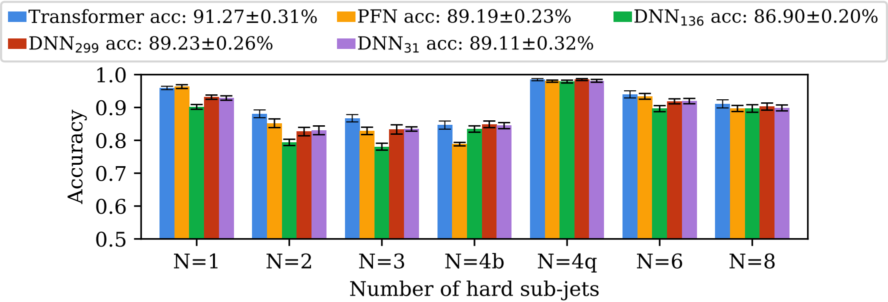
- We can then add these EFP observables to the DNN to see if we can bridge the gap to the PFN and Transformer models
- But this blunt force approach isn't that useful practically; what we want is the minimal set of observables that can match the performance

→ LASSO: "least absolute shrinkage and selection operator"

- Add a learnable parameter per input observable that shrinks to zero if the observable is not useful for classification
- Loss $L = -\log f(Y, Y_{\mathrm{pred}}) + \lambda \sum_i^{299} |g_i|$
  - where $-\log f(Y, Y_{\mathrm{pred}})$ is the negative log likelihood and λ is a hyperparameter of the network ($\lambda = 5$)

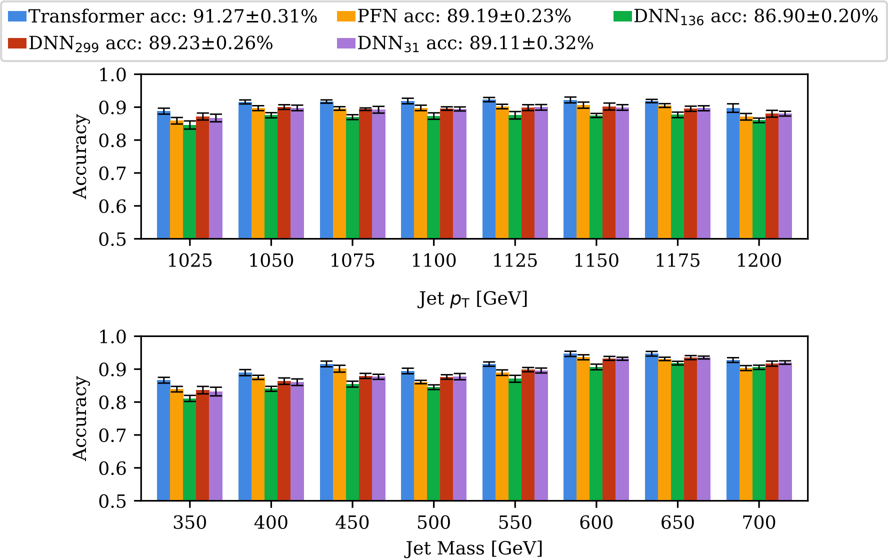
See also: Average Decision Ordering ▸ arXiv:2010.11998:

Bridging the Gap Results

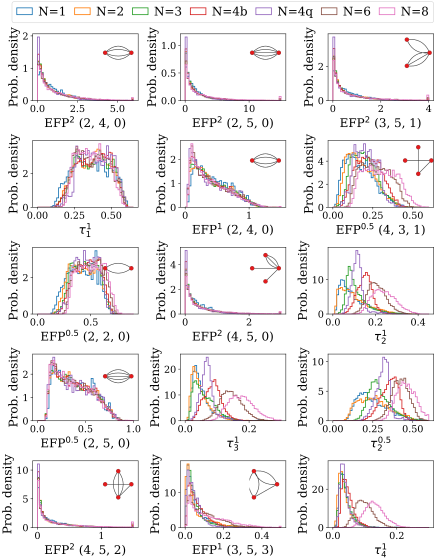


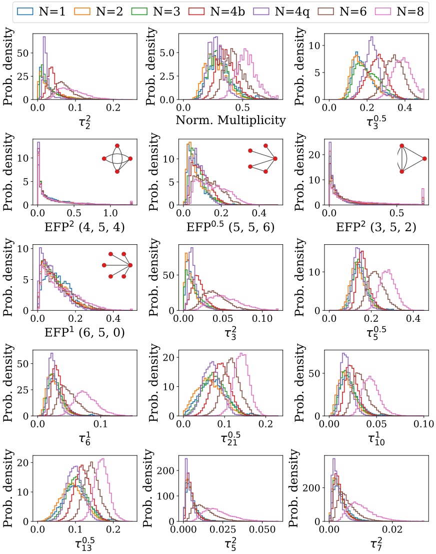
- All 299 variables achieve the ∼same performance as the PFN, but a gap to the Transformer model remains
- We find that with just 31 LASSO-selected variables, we can achieve the same overall performance as the larger model
- Note though that performance is not identical in each class!

Bridging the Gap Results (2)

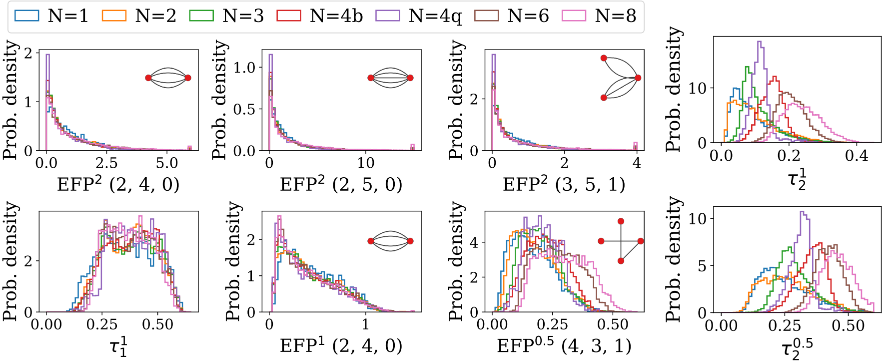


Selected Observables



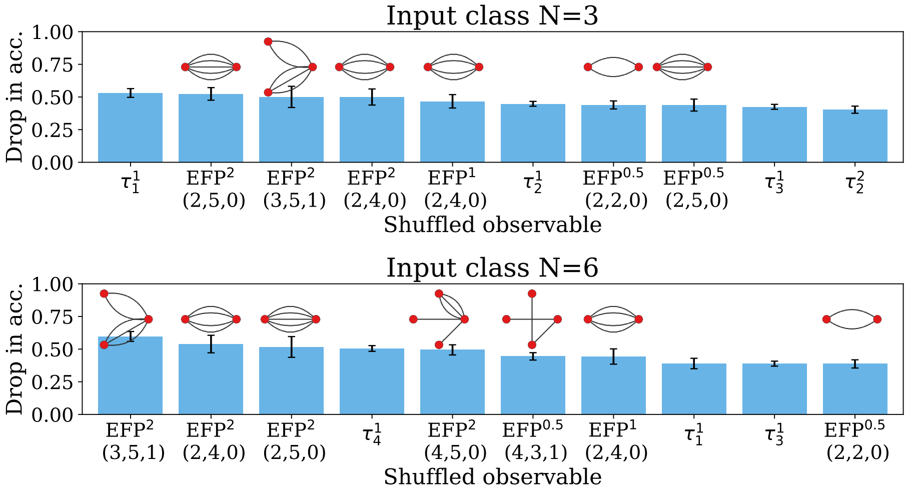


Selected Observables



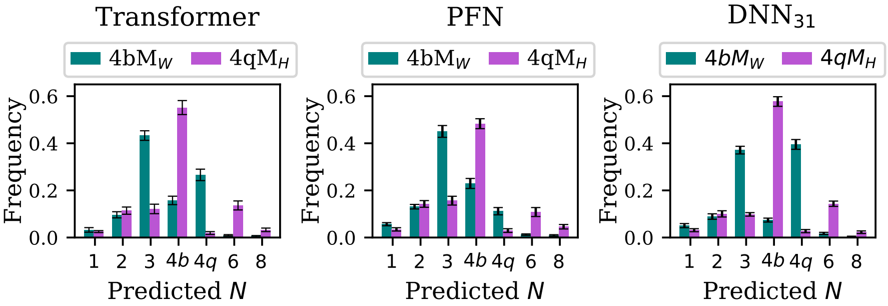
- Lots of $N = 2$ correlators
- Many EFP's do not, by themselves, show good separation between classes, yet are ranked higher than τ's which show more obvious differences
- Most important τ variable is τ_1, which is a direct measure of collimation
- Other highly ranked τ variables are mostly the "usual suspects", ie $N \leq 4$
- Lots of high angular weighting exponents

Breaking down the Selected Observables



- Ranking plot for specific classes (more in backup)

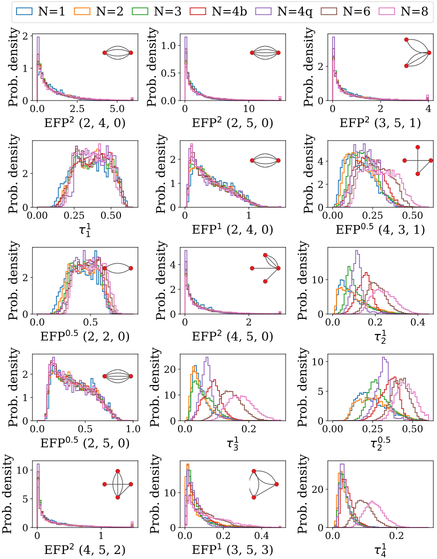
Zooming in on 4b vs 4q



- To understand the 4b vs 4q performance, we generate alternative $G \to HH$ / $G \to WW$ samples whereby the Higgs and W masses are swapped
- We see that $4\mathrm{qM}_H$ is usually classified as 4b, indicating the importance of the intermediate masses
- For $4\mathrm{bM}_W$, we find mixed results; the low-level networks often guess N=3, in which there is both a W and a b
    - Including b-tagging information directly in the networks may improve performance

Summary

- We have investigated the performance of both low level and high level jet taggers in extreme conditions, with up to $N = 8$ hard subjets
- The low level taggers typically out perform the "traditional" high level taggers
- Some of the gap can be filled by cleverly selecting/adding new variables, but still Transformer architectures maintain supremacy
- Performance is dependent on mass structures as well as heavy flavor content
- LASSO regulation a useful tool for automated feature selection



Backup

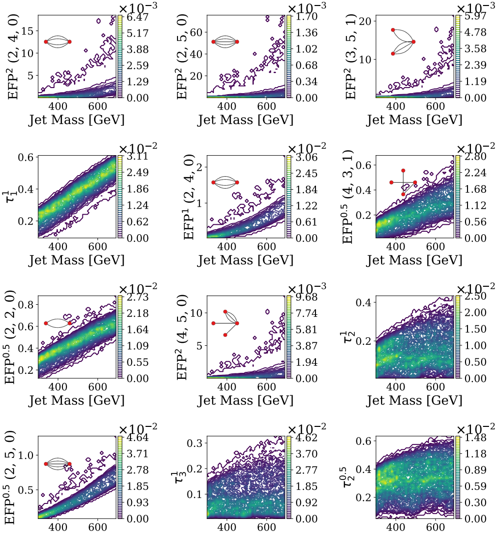
MC Details

N hard sub-jets	Process	M_W	M_h	M_t	M_G	requirements
1	$q\bar{q} \to q\bar{q}$					$p_T^q > 1000$
2	$q\bar{q} \to G \to W^+W^-$	80.4			2200	
		264.5			2200	
		440.8			2500	
		617.1			2800	
3	$q\bar{q} \to G \to t\bar{t}$			300	2200	
				500	2500	
				700	3000	
$4b$	$q\bar{q} \to \gamma G \to \gamma hh$				400	$p_T^\gamma > 1000$
					600	$p_T^\gamma > 1000$
					800	$p_T^\gamma > 1000$
$4q$	$q\bar{q} \to \gamma G \to \gamma W^+W^-$				400	$p_T^\gamma > 1000$
					600	$p_T^\gamma > 1000$
					800	$p_T^\gamma > 1000$
6	$q\bar{q} \to \gamma G \to \gamma t\bar{t}$				400	$p_T^\gamma > 1000$
					600	$p_T^\gamma > 1000$
					800	$p_T^\gamma > 1000$
8	$q\bar{q} \to \gamma t\bar{t}h$		100	125		$p_T^\gamma > 1000$
			125	175		$p_T^\gamma > 1000$

Network Details

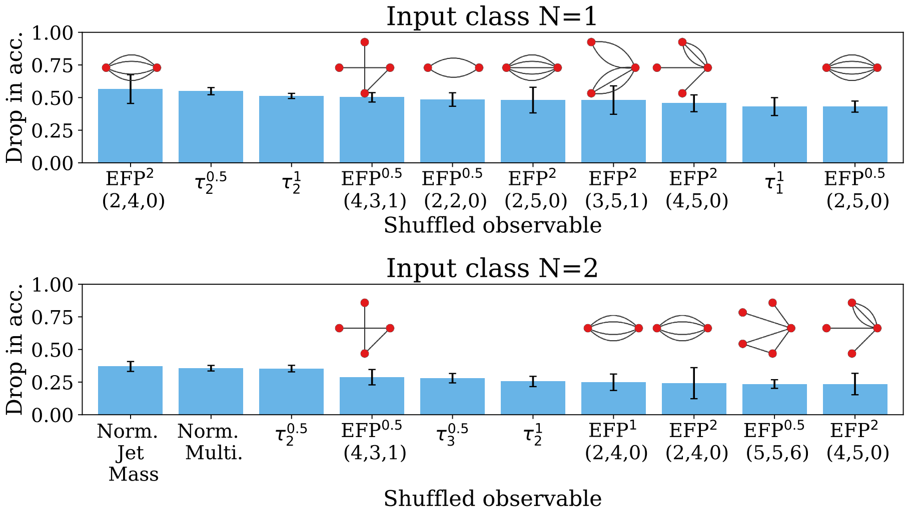
Model	Description	No. of Params.	Accuracy
Transformer	Transformer Network trained on the jet constituents.	1,388,807	91.27 ± 0.31 %
PFN	Particle-Flow Network trained on the jet constituents.	1,205,895	89.19 ± 0.23 %
DNN_{136}	Fully-connected neural network trained on the 135 N-subjettiness observables and the norm. jet mass.	2,732,519	86.90 ± 0.20 %
DNN_{299}	Fully-connected neural network trained on the 135 N-subjettiness, observables the normalized jet mass, and the full set of EFP observables.	2,862,919	89.23 ± 0.26 %
DNN_{31}	Fully-connected neural network trained on the 31 LASSO-selected observables.	2,622,663	89.11 ± 0.32 %

Why does performance increase as mass increases?

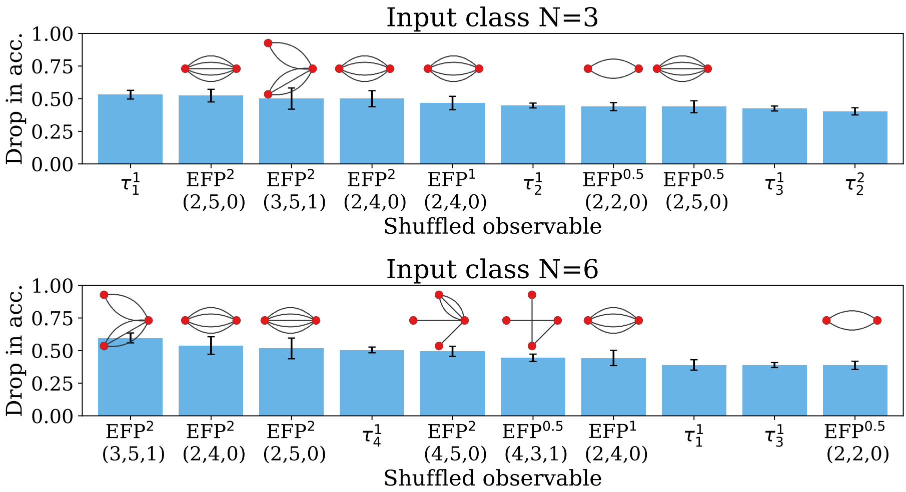


- We inspect the correlations of the highest ranked variables against the mass to understand why higher masses seem easier to classify
- τ_1, which is a measure of collimation along the jet axis, gives us a clue; this variable, ranked 4th, is highly correlated to mass
  - $\rightarrow$ more collimated jets are harder to classify, probably due to merging constituents
- Our results are consistent if we only use half of the overall (quite large) mass range

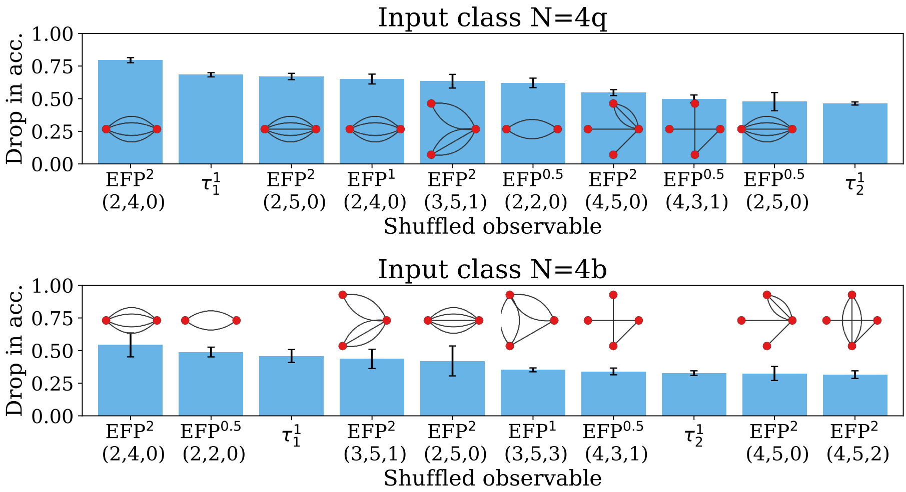
Breaking down the Selected Observables



Breaking down the Selected Observables



Breaking down the Selected Observables



Breaking down the Selected Observables