

Feature Selection with Distance correlation

Ranit Das

ranit@physics.rutgers.edu



with

David Shih & Gregor Kasieczka

Based on [arXiv:2212.00046](https://arxiv.org/abs/2212.00046)

Pheno 2023

Date: 05/08/2023

Outline

Motivation for feature selection

Feature selection algorithm using DisCo

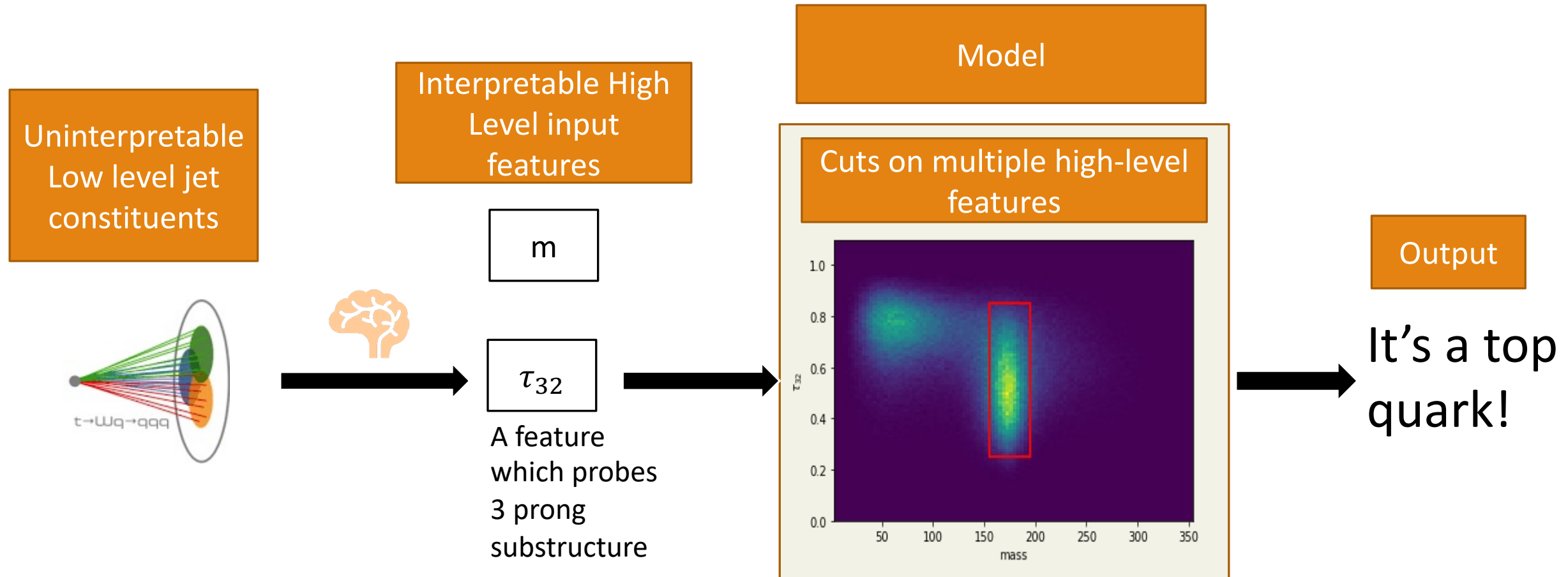
Application to Top Tagging

Results

Conclusion

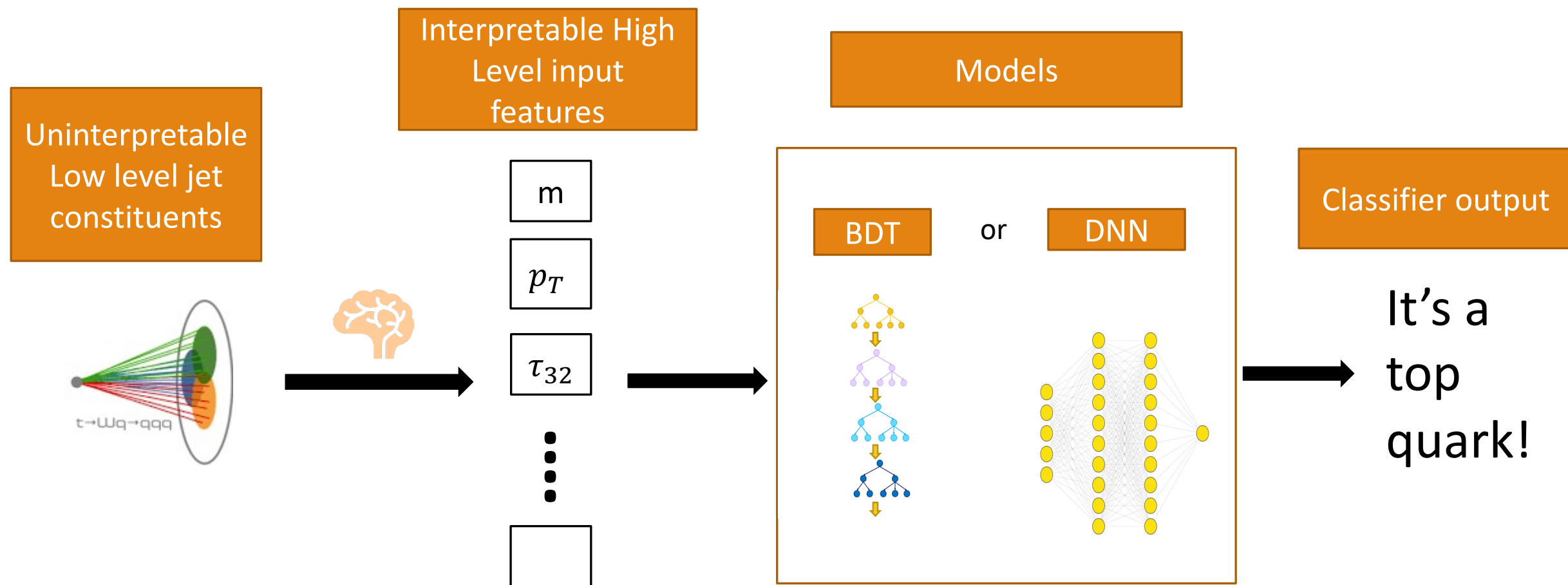
History of Boosted object tagging

1. Using cuts on multiple High-Level (HL) features



History of Boosted object tagging

2. Using a set of high-level features as inputs to BDT or DNN

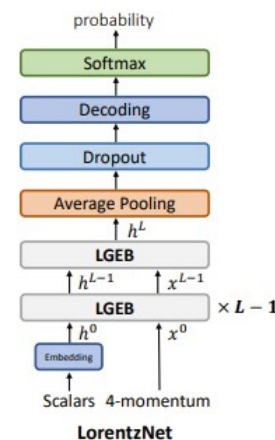
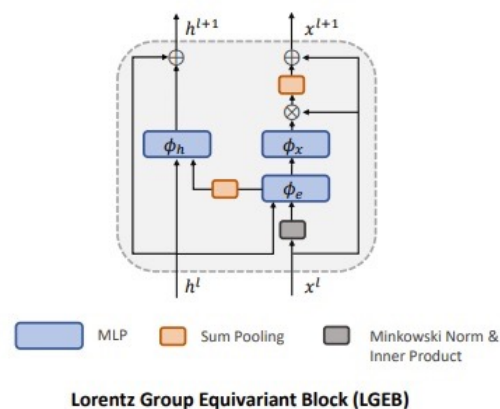
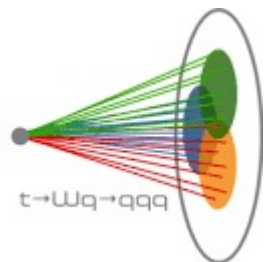


History of Boosted object tagging

3. Use low-level features directly as inputs to neural networks

State of the art Neural Networks

Uninterpretable
Low level jet
constituents



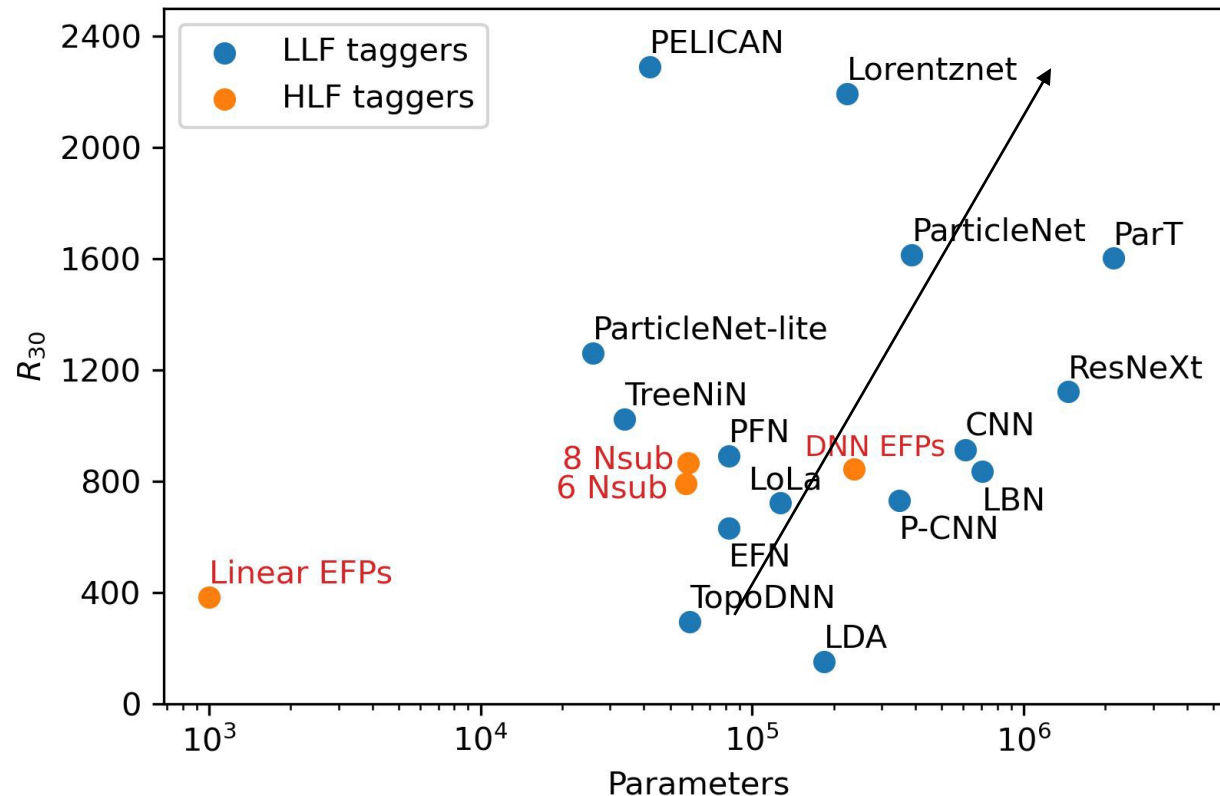
Classifier output

It's a top quark!

Previously on top tagging

HL feature taggers haven't been able to keep up with low-level feature taggers

R_{30}
(Rejection
factor at
30% true
positive
rate)



The Machine Learning Landscape of Top Taggers: [arXiv:1902.09914v3](https://arxiv.org/abs/1902.09914v3)

Particle Transformer for Jet Tagging: [arXiv:2202.03772](https://arxiv.org/abs/2202.03772)

An Efficient Lorentz Equivariant Graph Neural Network for Jet Tagging: [arXiv:2201.08187v5](https://arxiv.org/abs/2201.08187v5)

ParticleNet: Jet Tagging via Particle Clouds: [arXiv:1902.08570v3](https://arxiv.org/abs/1902.08570v3)

Mapping Machine-Learned Physics into a Human-Readable Space [arXiv:2010.11998](https://arxiv.org/abs/2010.11998)

Reports of My Demise Are Greatly Exaggerated: N-subjettiness Taggers Take On Jet Images: [arXiv:1807.04769](https://arxiv.org/abs/1807.04769)

How Much Information is in a Jet?: [arXiv:1704.08249v2](https://arxiv.org/abs/1704.08249v2)

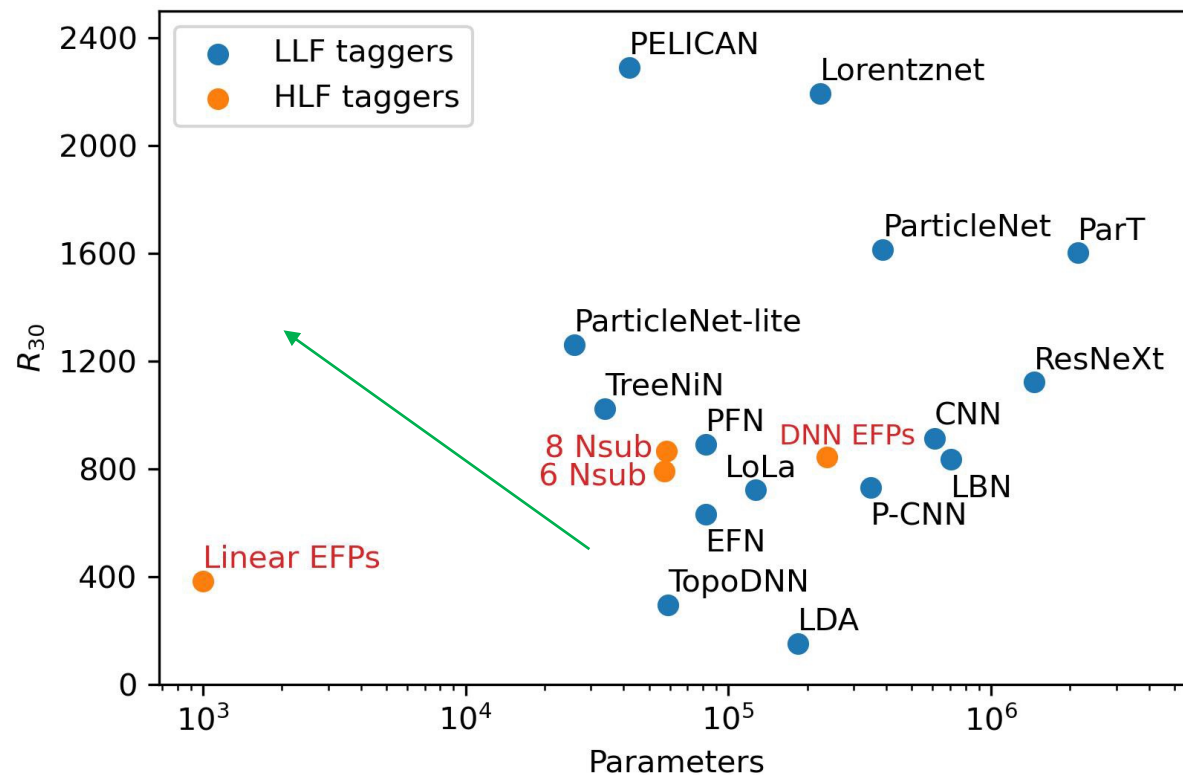
A complete linear basis for jet substructure: [arXiv:1712.07124](https://arxiv.org/abs/1712.07124)

PELICAN: Permutation Equivariant and Lorentz Invariant or Covariant Aggregator Network for Particle

[arXiv:2211.00454](https://arxiv.org/abs/2211.00454)

Why should we go back to high-level (HL) features?

Can build a more efficient model with less parameters



- High-level features are more interpretable.
- Faster evaluation
- More resource efficient
- Features can be more robust and easier to calibrate and validate between simulated and experimental data.

Feature Selection

is the process of selecting a subset of useful features to use in model construction/training.

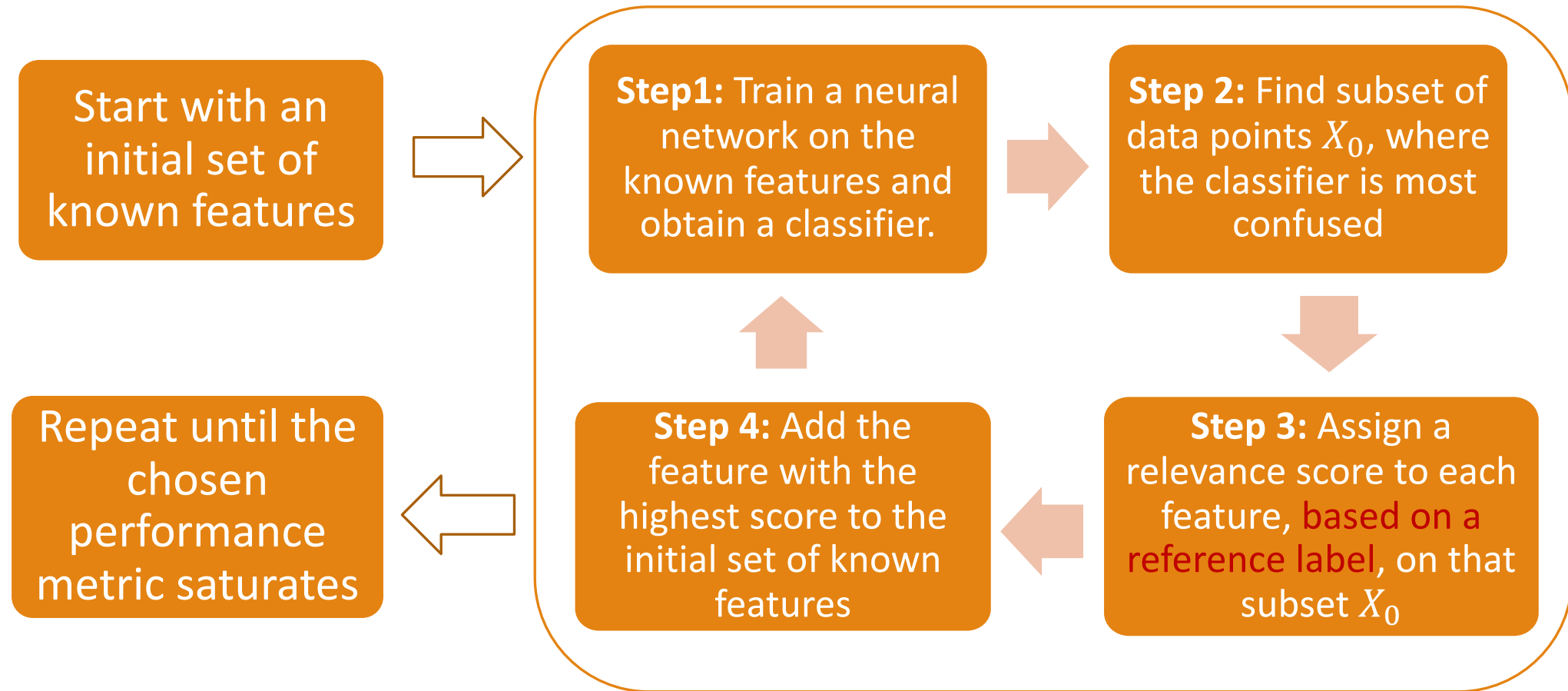
How to do Feature Selection?

- Know which features are useful!
- Use a **feature selection algorithm**.

Feature selection Algorithm

- Given a large number of features, a feature selection algorithm can select a few useful features based on a **relevance score** assigned to each feature. We use our score as a measure of correlation between each of our features and truth labels.
- **The score ranks features which are more useful than the others !**

Overview of a Forward Feature Selection (FFS) algorithm

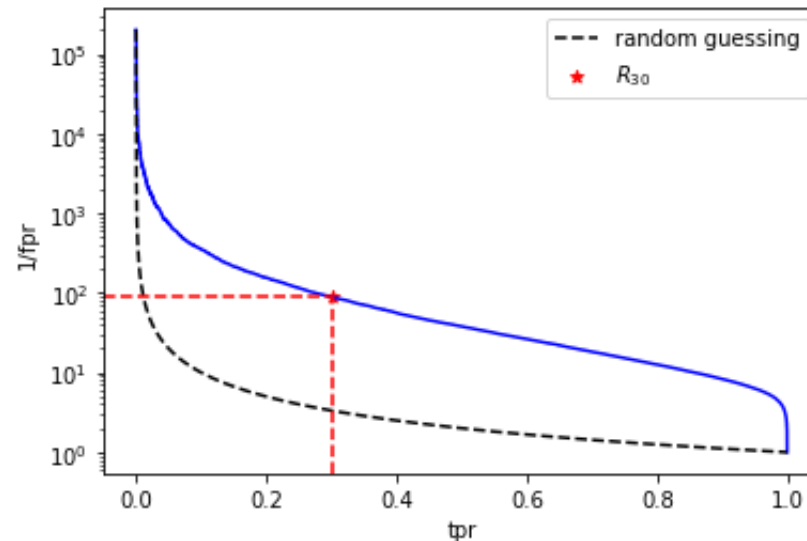


Application of the algorithm to top tagging

- **Data set:** The Machine Learning Landscape of Top Taggers ([arXiv:1902.09914v3](https://arxiv.org/abs/1902.09914v3)). ([10.5281/zenodo.2603255](https://zenodo.org/record/2603255))
- **2M jets:** Signal and Background, with only Energy-momentum four vectors.
- Training set (1.2 M), validation set (400k), and test set (400k)
- The algorithm is applied to the combined training and validation set, and the metric is evaluated on the test set.

Application of the algorithm to top tagging

- **Metric used:** R_{30} (Rejection factor at 30% true positive rate) is evaluated on a test set (400k events)



- **Initial set of features:** m_J , p_{T_J} , $m_{W-candidate}$

Features: Energy Flow Polynomials (EFPs)

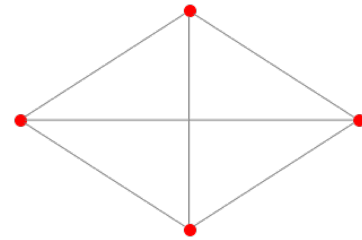
with $d \leq 7$, with $\kappa = \left[-1, 0, \frac{1}{2}, 1, 2\right]$ and $\beta = \left[\frac{1}{2}, 1, 2\right]$, 7350 features

Large set of features, which are functions of:

- z_a : The momentum fraction of jet constituent a
- θ_{ab} : Angular separation between jet constituents a and b

$$z_a^{(\kappa)} = \left(\frac{p_{T_a}}{\sum_b p_{T_b}} \right)^\kappa \qquad \theta^{(\beta)} = \left(\Delta \eta_{ab}^2 + \Delta \phi_{ab}^2 \right)^{\frac{\beta}{2}}$$

Features: Energy Flow Polynomials (EFPs)


$$= \sum_a z_a \sum_b z_b \sum_c z_c \sum_d z_d \theta_{ab} \theta_{ac} \theta_{ad} \theta_{bc} \theta_{bd} \theta_{cd}$$

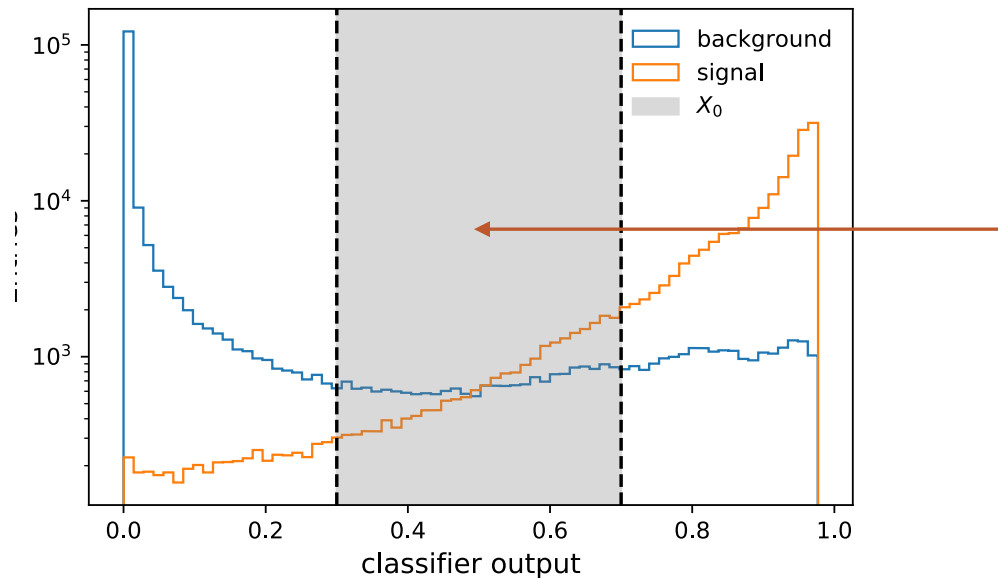
- Each node : $\sum_a z_a$
- Each edge : θ_{ab}

Step1: Train a neural network on the known features and obtain a classifier.

- We train a Neural network with an initial set of features:
 $F_{initial} = \{m_J, p_{T_J}, m_{W-candidate}\}$

Step 2: Find a subset X_0 , with data points where the classifier is most confused

- We select data points with a specific window around classifier output value 0.5, as points where the classifier is most confused. (**we call X_0 our confusion set**)



Confusion set X_0

Data points where the classifier most confused

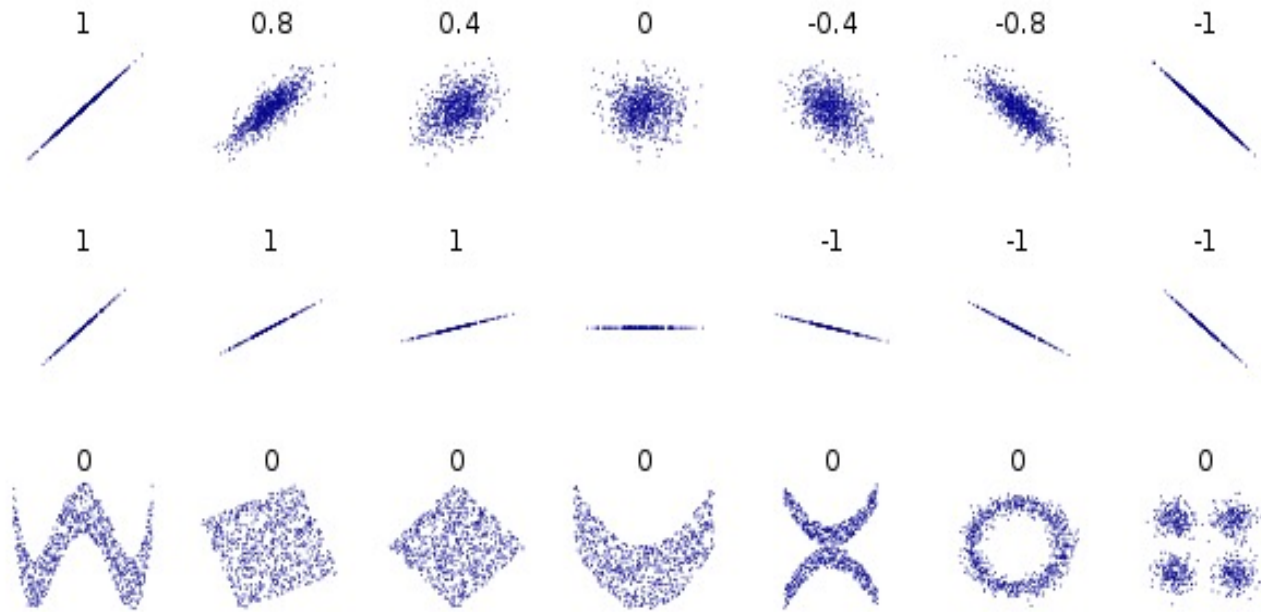
Step 3: Assign a relevance score to each feature, based on reference label, on that subset X_0

- On X_0 we evaluate:
 $DisCo(y_{ref}, [known\ variables, new\ feature])$ for each feature in the feature subspace.

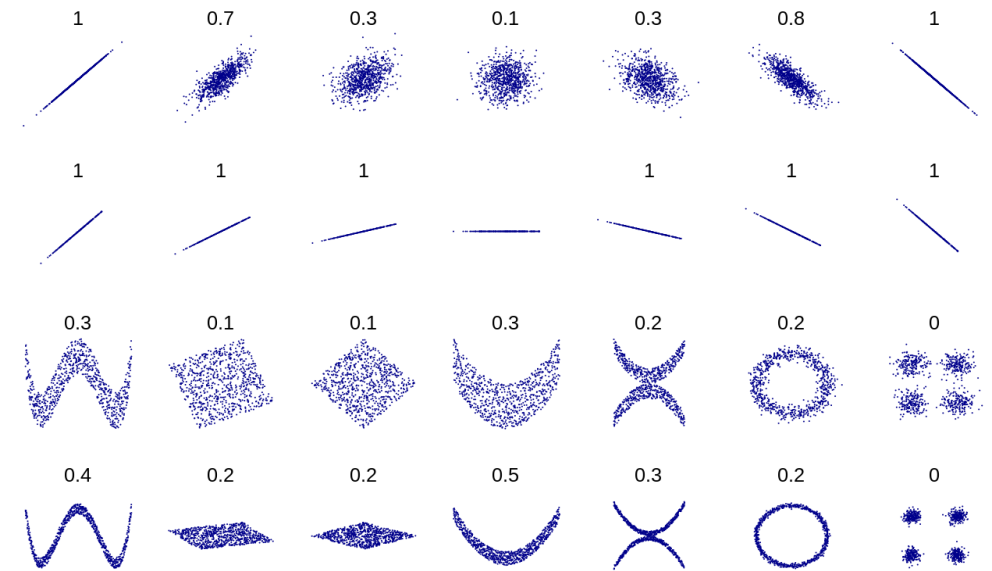
Relevance Score : Distance Correlation (DisCo)

- DisCo is used to find value of **non-linear correlations** of the EFPs with the reference label.
- Very powerful since we can quantify correlations between reference labels and multiple features.

Relevance Score : Distance Correlation (DisCo)



Pearson Correlation



DisCo

Reference label: Truth label or state-of the art model

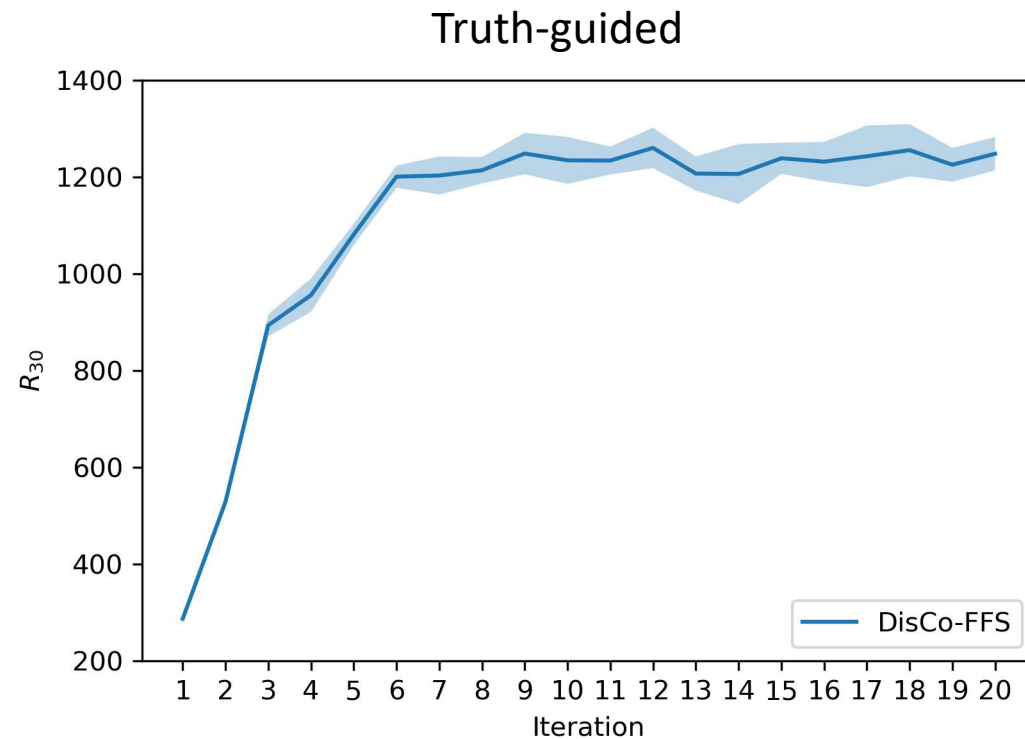
- In a truth-guided approach, the truth labels are used as the Reference label get the best possible tagger.
- In a state-of the art model-guided approach we use the **LorentzNet** (one of the highest performing top-taggers with a R_{30} of 2195) as the reference label. The features selected can be used to explain “What the machine learned?”

Step 4: Add the feature with the highest score to the initial set of known features

- The feature with the highest DisCo value is added to the list of known features, and a new Neural Network is trained using the new set of features.

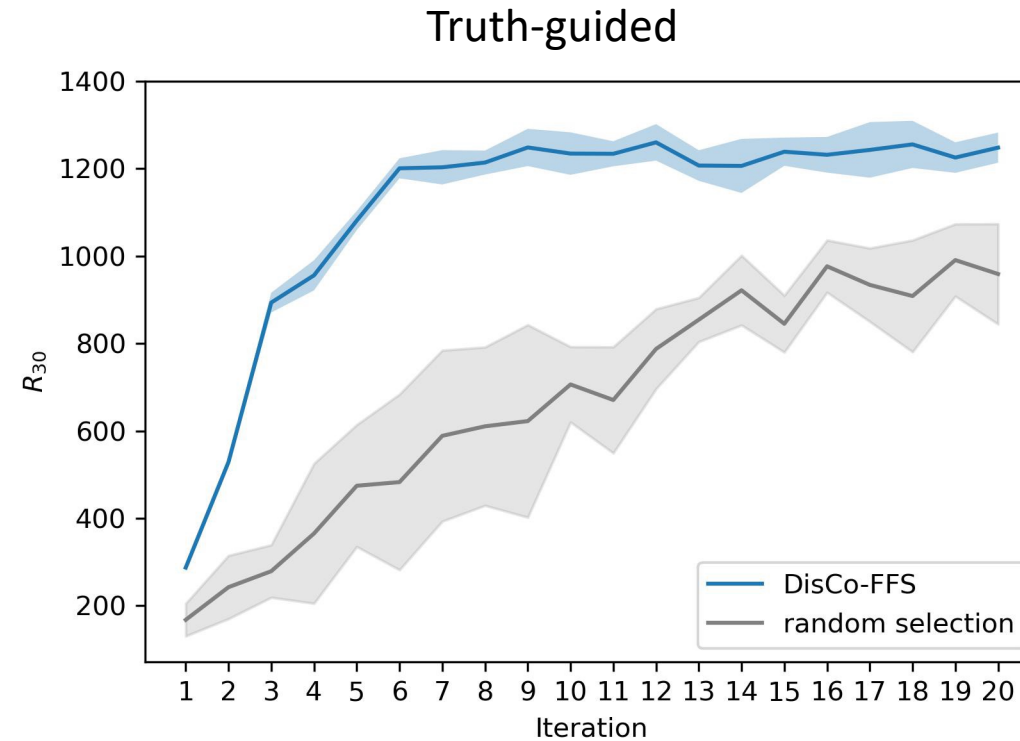
Performance after addition of new EFPs using feature selection algorithm

- Variance for each method is obtained by training each network 10 times.
- Our method can obtain an R_{30} of 1249 ± 43 , after 9 features.



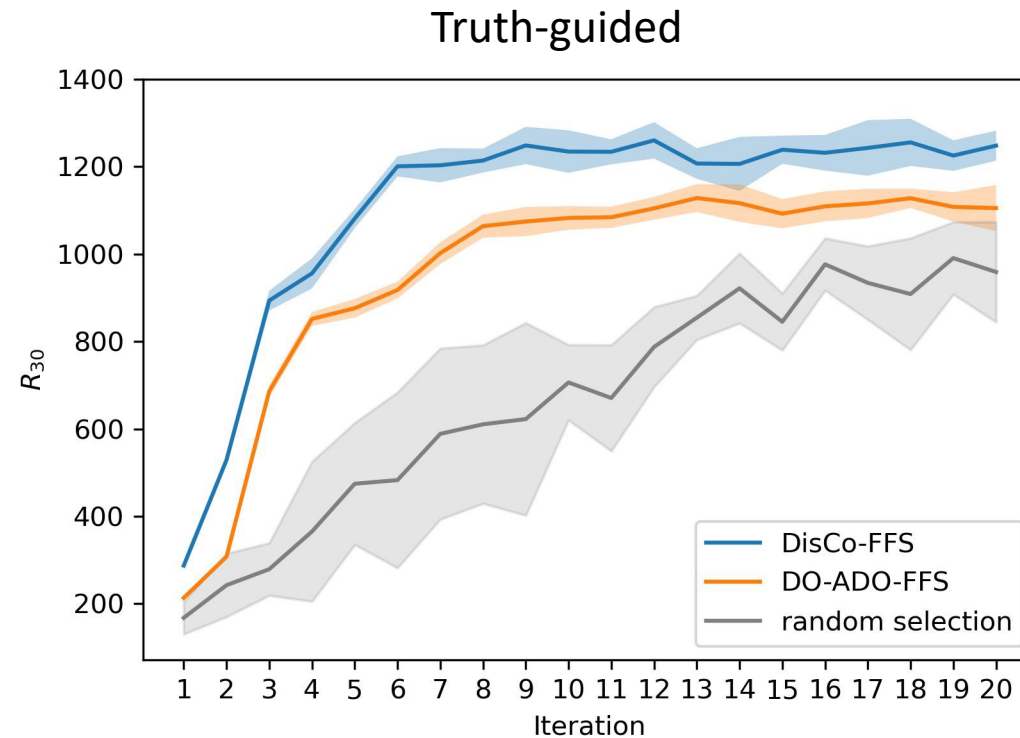
Baseline: Random selection of features

A feature selection algorithm should perform better than randomly selecting features.



Comparison to a previous feature selection algorithm DO-ADO (truth)

- A previous feature selection method, which relies on Decision ordering (DO) for finding subset of data where a classifier orders signal/background differently from the truth labels.
- Use Average Decision Ordering (ADO) between EFPs and the truth, as the score

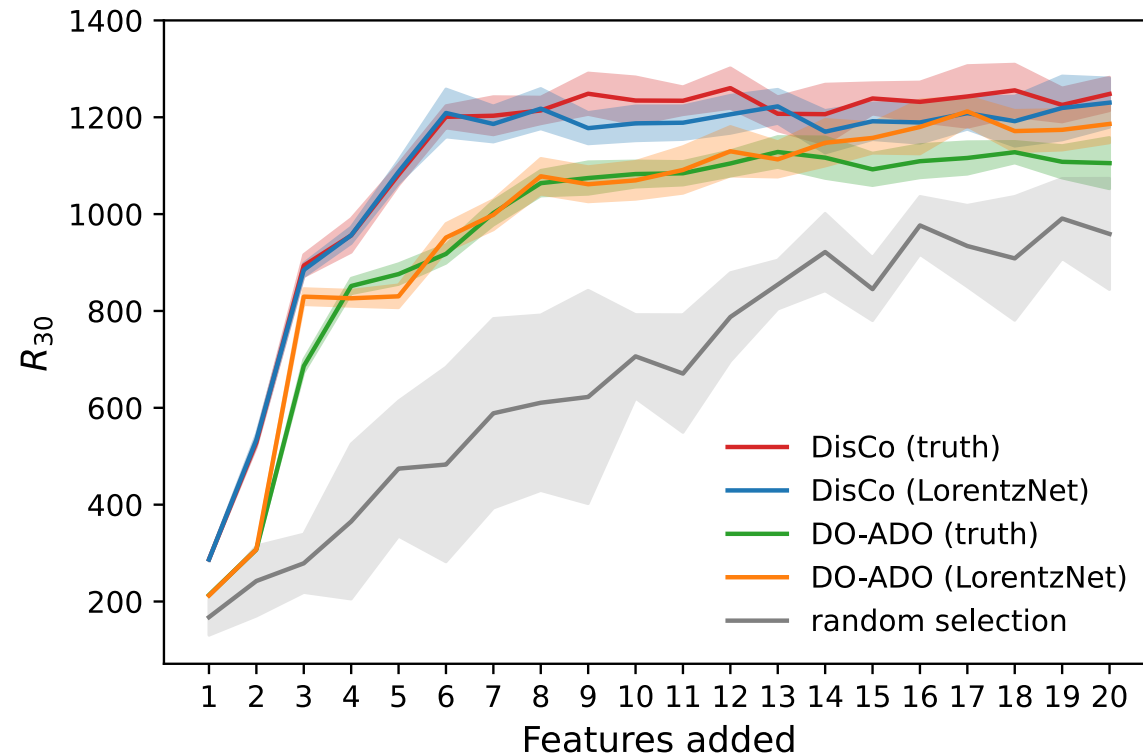


ADO method: Mapping Machine-Learned Physics into a Human-Readable Space [arXiv:2010.11998](https://arxiv.org/abs/2010.11998)

Comparison to LorentzNet guided feature selection

An Efficient Lorentz
Equivariant Graph
Neural Network for Jet
Tagging:
[arXiv:2201.08187v5](https://arxiv.org/abs/2201.08187v5)

DisCo-FFS has a similar performance for both the truth-guided and LorentzNet guided approach



DO-ADO has better performance for LorentzNet guided approach, as compared to truth guided approach (as noted in [arXiv:2010.11998](https://arxiv.org/abs/2010.11998))

Comparison to other top taggers

The Machine Learning Landscape of Top Taggers: [arXiv:1902.09914v3](https://arxiv.org/abs/1902.09914v3)

Particle Transformer for Jet Tagging: [arXiv:2202.03772](https://arxiv.org/abs/2202.03772)

An Efficient Lorentz Equivariant Graph Neural Network for Jet Tagging: [arXiv:2201.08187v5](https://arxiv.org/abs/2201.08187v5)

ParticleNet: Jet Tagging via Particle Clouds: [arXiv:1902.08570v3](https://arxiv.org/abs/1902.08570v3)

Mapping Machine-Learned Physics into a Human-Readable Space [arXiv:2010.11998](https://arxiv.org/abs/2010.11998)

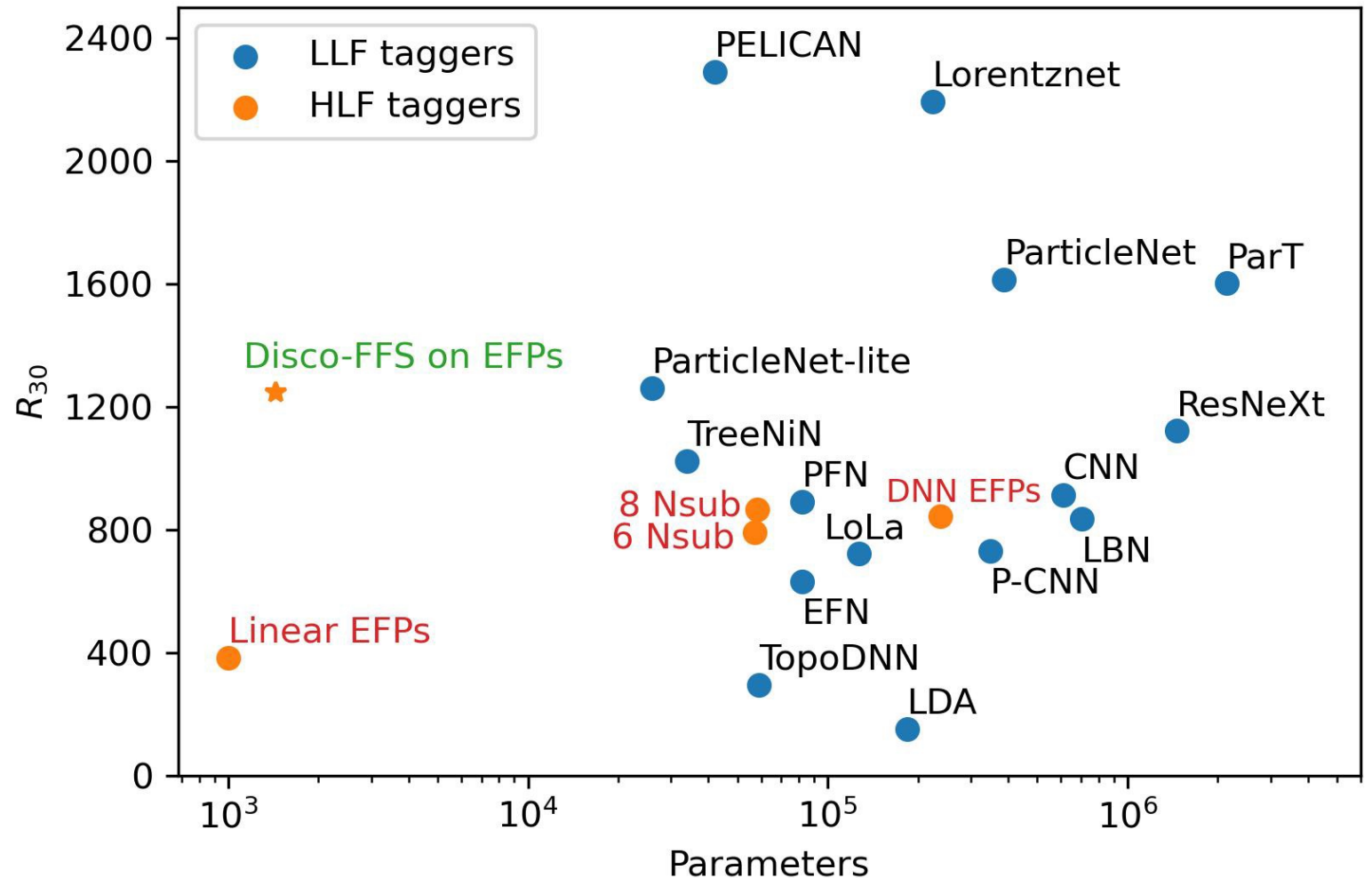
Reports of My Demise Are Greatly Exaggerated: N-subjettiness Taggers Take On Jet Images: [arXiv:1807.04769](https://arxiv.org/abs/1807.04769)

How Much Information is in a Jet?: [arXiv:1704.08249v2](https://arxiv.org/abs/1704.08249v2)

A complete linear basis for jet substructure: [arXiv:1712.07124](https://arxiv.org/abs/1712.07124)

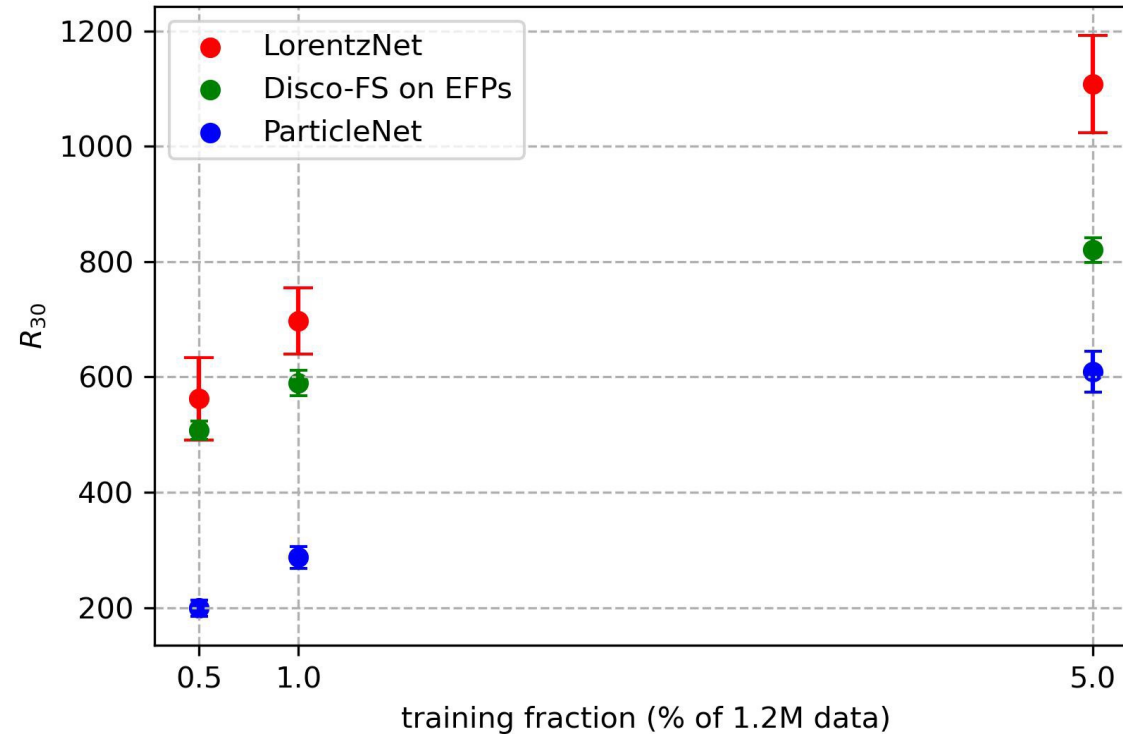
PELICAN: Permutation Equivariant and Lorentz Invariant or Covariant Aggregator Network for Particle Physics [arXiv:2211.00454](https://arxiv.org/abs/2211.00454)

Our method achieves state of the art performance with only a very small fraction of the parameters!



Sample Efficiency

Our feature selected model, outperforms the ParticleNet, and matches the LorentzNet, when trained on less training data.



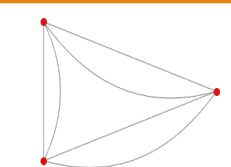
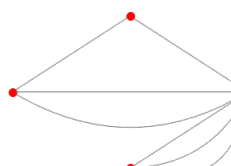

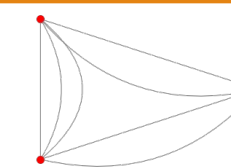
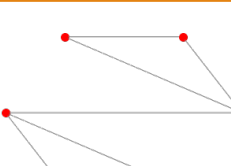
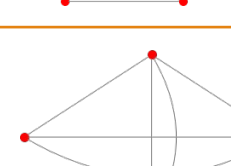
*We use the features, which were selected using the larger dataset.

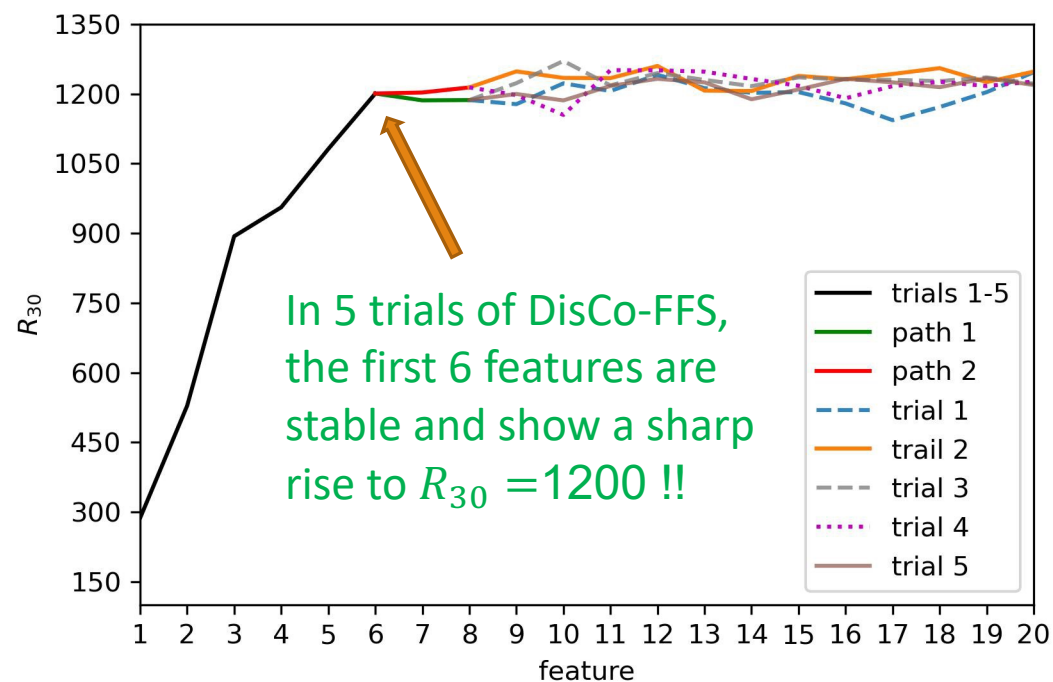
An Efficient Lorentz Equivariant Graph Neural Network for Jet Tagging: [arXiv:2201.08187v5](https://arxiv.org/abs/2201.08187v5)

ParticleNet: Jet Tagging via Particle Clouds: [arXiv:1902.08570v3](https://arxiv.org/abs/1902.08570v3)

Robustness of DisCo-FFS

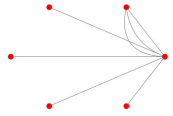
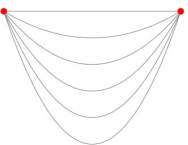
- On 5 independent trials of doing DisCo-FFS selects the same first 6 features in every trial.
- Chromatic number (c) is a proxy for number of prongs in a jet
- 5 of the first 6 EFPs have $c=3$, which means our algorithm selects features which probe the 3-prong substructure which is relevant for top-tagging.
- One of them is probe of 2-prong substructure.

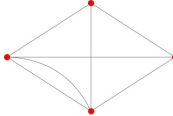
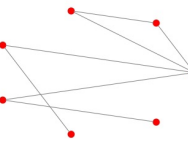
#	Graphs	c	κ	β
1		3	2	1
2		3	2	1
3		2	0	1
4		3	1	0.5
5		3	1	1
6		3	2	0.5

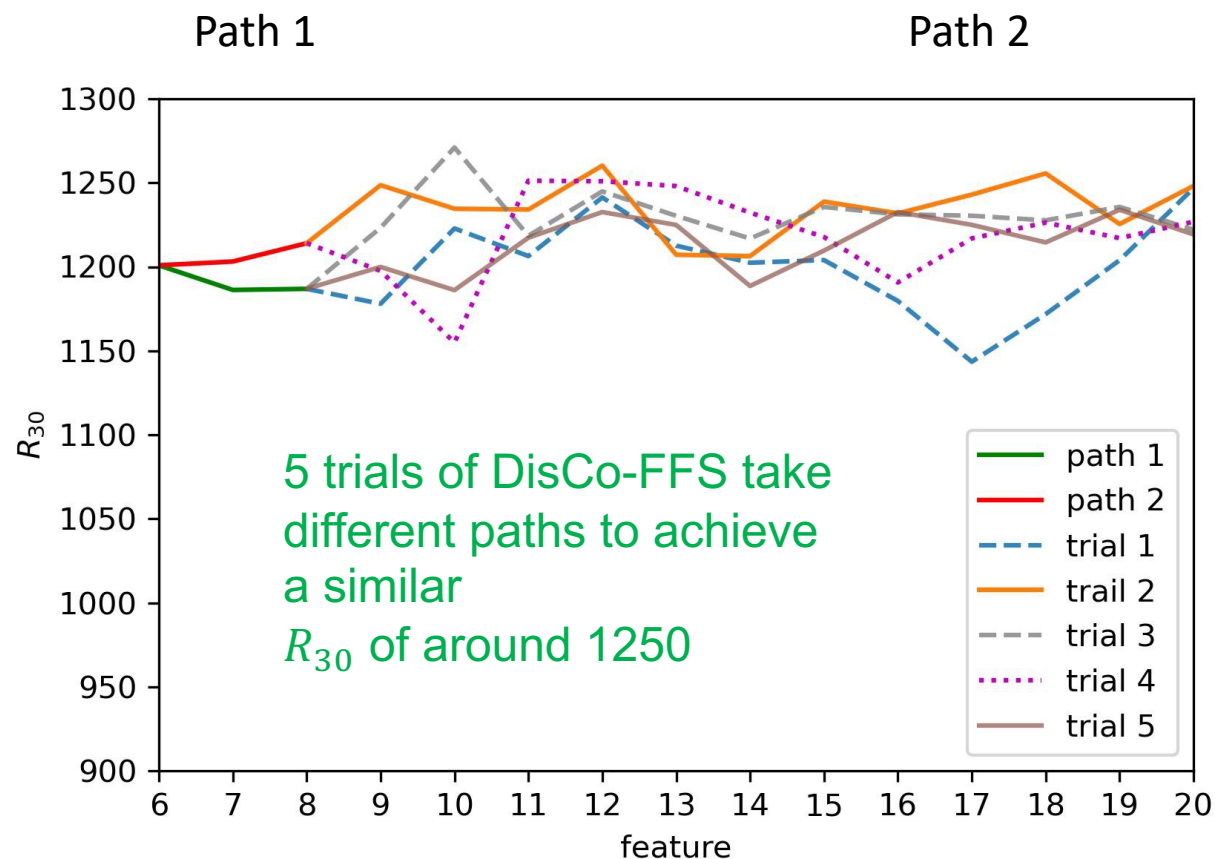


Robustness of DisCo-FFS

- After the 6th iteration, we see some degree of randomness, as we see two unique possible paths taken by DisCo-FFS in the 7th and 8th iteration, and after the 9th iteration it selects 5 different features.
- In Path 1, the first feature it selects probes 4-prong substructure, followed by a feature which probes 3-prong substructure
- In Path 2, it selects 2 features which probe 2-prong substructure.

#	Graphs	c	κ	β
7		2	0	0.5
8		2	2	2

#	Graphs	c	κ	β
7		4	0.5	0.5
8		3	1	1



Conclusion

- Using a Disco based feature selection for the case of top tagging, we were able to obtain a handful of input features, which gave a very competitive performance, given the number of parameters.
- EFPs selected could make for a very lightweight and performant top tagger, which could have important applications to triggering ([arXiv:1804.06913](https://arxiv.org/abs/1804.06913))

Possible reasons for not getting a better performance:

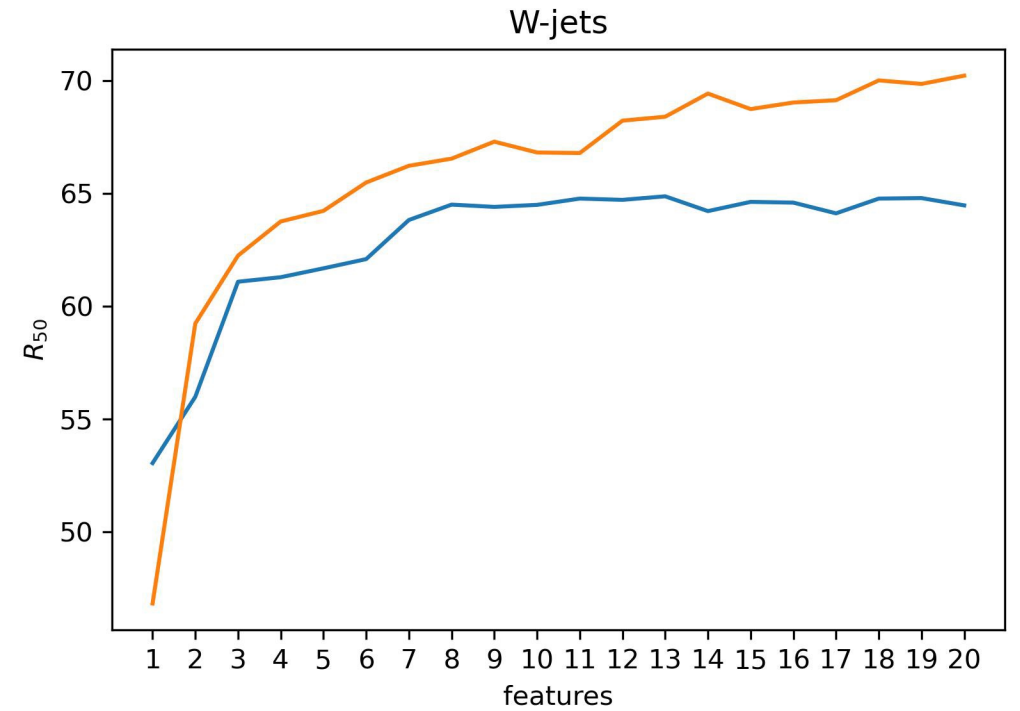
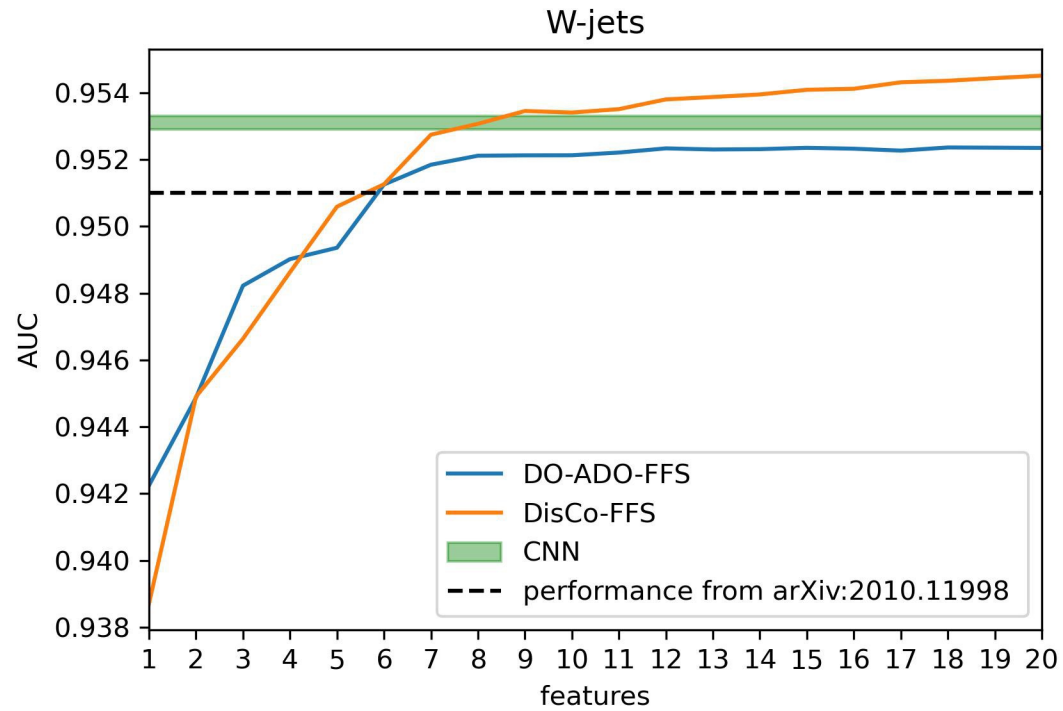
- The feature space considered could be insufficient for top tagging, which could explain our inability to close the gap with higher performing black box models.
- Need a better feature selection algorithm?



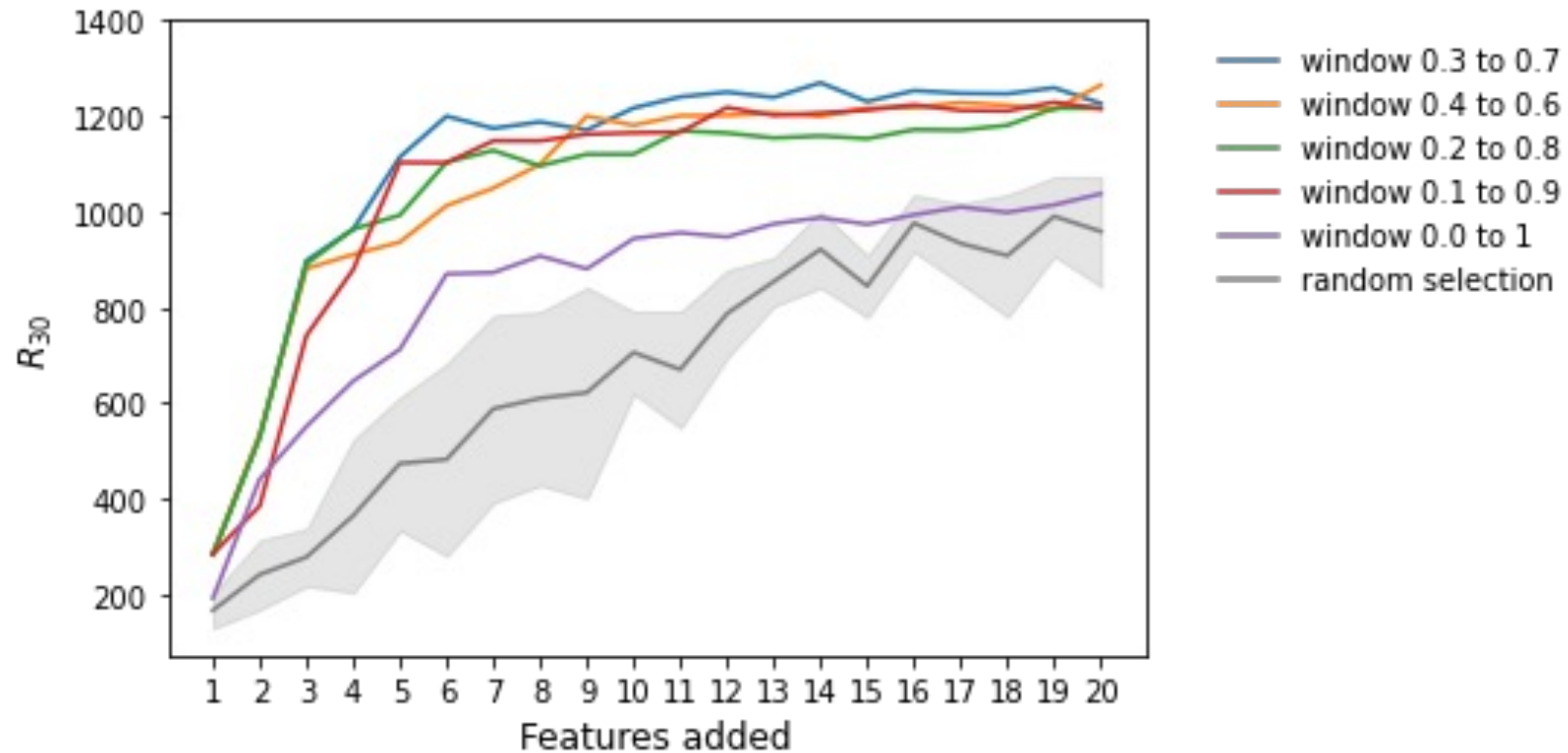
Thank You!

BACK UP SLIDES

W-jets validation



Instead of calculating score on full data, the selection of the confusion set improves the performance!



The classifier output window 0.3 to 0.7 was optimal for the case of top-tagging

DO-ADO

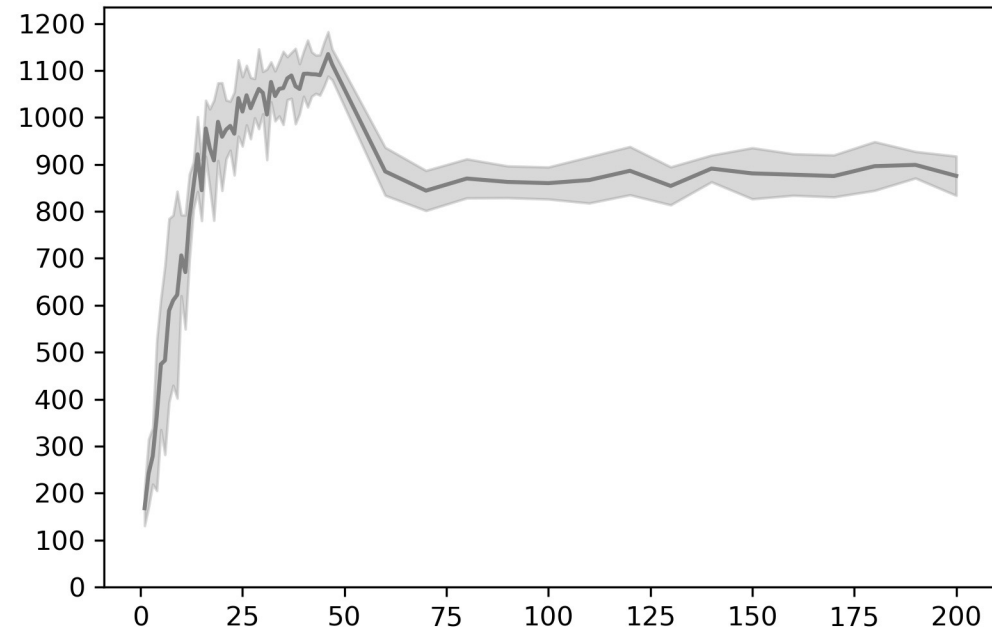
$DO(f(x), g(x)) = \Theta((f(x_s) - f(x_b))(g(x_s) - g(x_b)))$, where s refers to signal, and b refers to background.

DO is a measure of relative ordering $f(x)$ with respect to $g(x)$, for a single signal-background pair .

Same ordering gives $DO=1$, whereas different ordering leads to $DO=0$. Eg: $DO = 1$, if $f(x_s) > f(x_b)$ and $g(x_s) > g(x_b)$, whereas $DO = 0$, if $f(x_s) > f(x_b)$ and $g(x_s) < g(x_b)$

Average Decision Ordering (ADO) is the average value of DO over a sample of signal-background pairs.

Random Selection



Affine Invariant Distance Correlation (DisCo)

It has some nice properties:

Zero iff X, Y are independent, positive otherwise.

Can quantify non-linear correlations between 2 unequal sets of features X and Y .

Is invariant under linear rescaling of features in each set X and Y

Step 2: Find a subset X_0 , with data points where the classifier is most confused

Our method using
Distance
Correlation (DisCo)

- We select data points with a specific window around classifier output value 0.5, as points where the classifier is most confused.

DO-ADO method

- Selects a subsample of signal-background pairs with $DO(y, y^{truth/blackbox}) = 0$, i.e, signal-background pairs for which the classifier output, which is different relative to the truth labels (y^{truth}) or a blackbox classifier output ($y^{blackbox}$) with a high-performance score.

Step 3: Use a score to rank the features over the subset X_0

Our method using
Distance
Correlation (DisCo)

- On X_0 we evaluate, $DisCo(y^{truth}, [initial/known\ variables, new\ feature])$ for each feature in the feature subspace.

DO-ADO method

- On X_0 evaluate, $ADO(y^{truth/background}, new\ feature)$