Feature Selection with Distance correlation

Ranit Das
ranit@physics.rutgers.edu
RUTGERS

with
David Shih & Gregor Kasieczka

BOOST 2022

Date: 17/8/2022

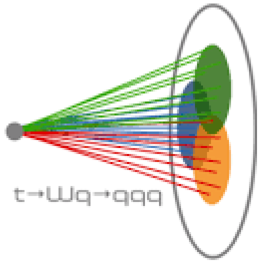
Outline

- 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

Uninterpretable Low level jet constituents



Interpretable High Level input features

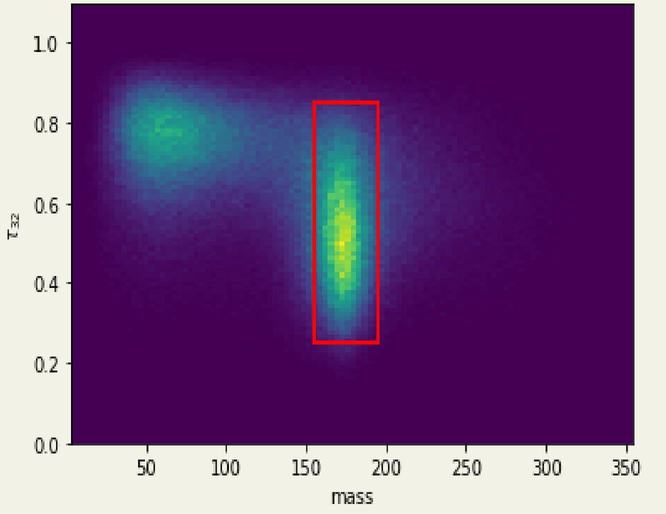
m

τ_{32}

A feature which probes 3 prong substructure

Model

Cuts on multiple high-level features



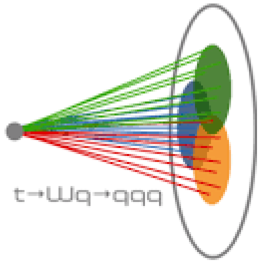
Output

It's a top quark!

History of Boosted object tagging

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

Uninterpretable Low level jet constituents

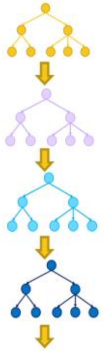


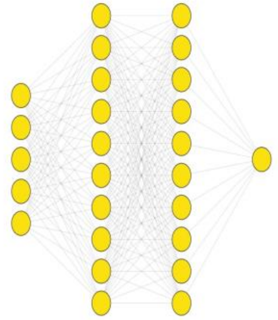
Interpretable High Level input features

m
p_T
τ_{32}
⋮

Models

BDT or DNN





Classifier output

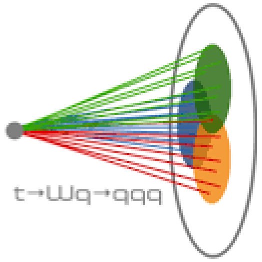
It's a top quark!

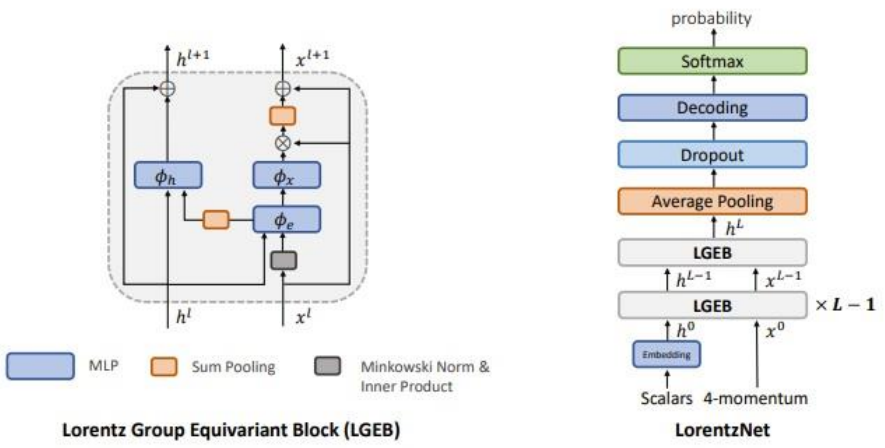
History of Boosted object tagging

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

State of the art Neural Networks

Uninterpretable Low level jet constituents





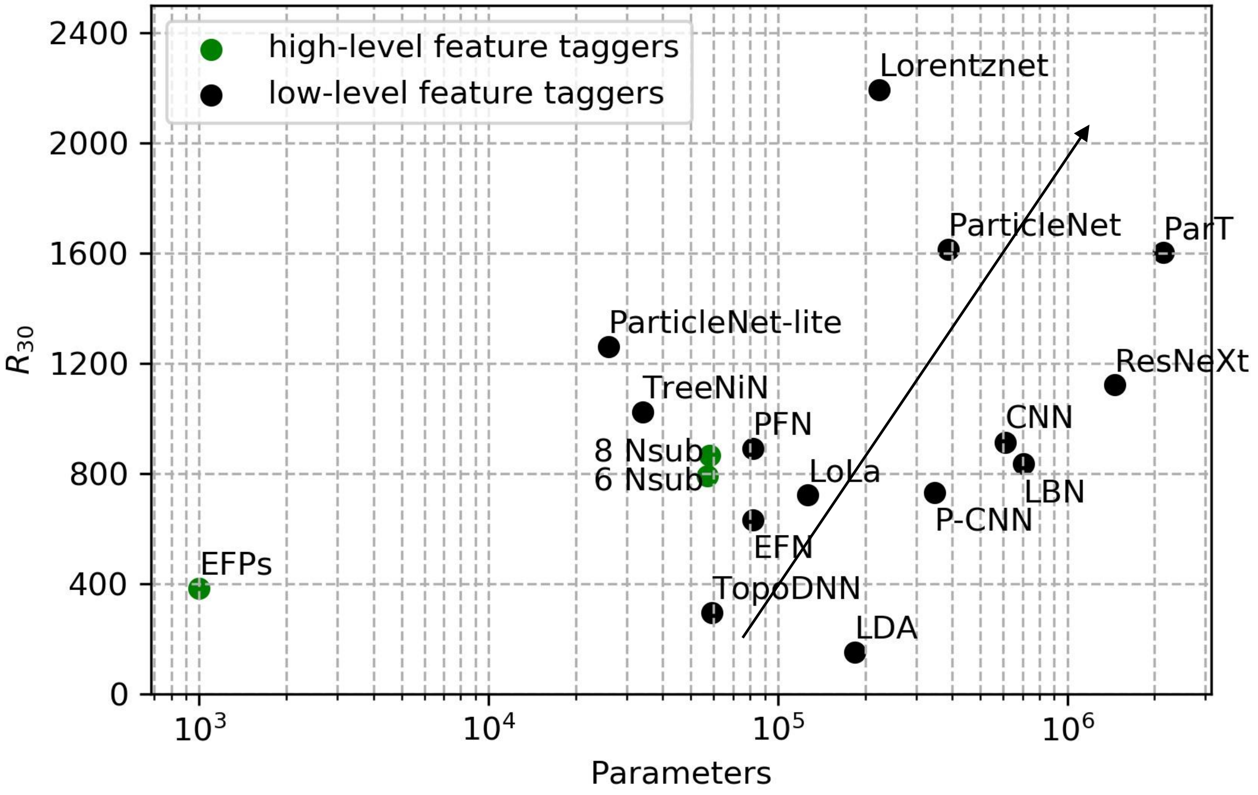
Classifier output

It's a top quark!

Previously on top tagging ....

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

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



The Machine Learning Landscape of Top Taggers: arXiv:1902.09914v3

Particle Transformer for Jet Tagging: arXiv:2202.03772

An Efficient Lorentz Equivariant Graph Neural Network for Jet Tagging: arXiv:2201.08187v5

ParticleNet: Jet Tagging via Particle Clouds: arXiv:1902.08570v3

Mapping Machine-Learned Physics into a Human-Readable Space arXiv:2010.11998

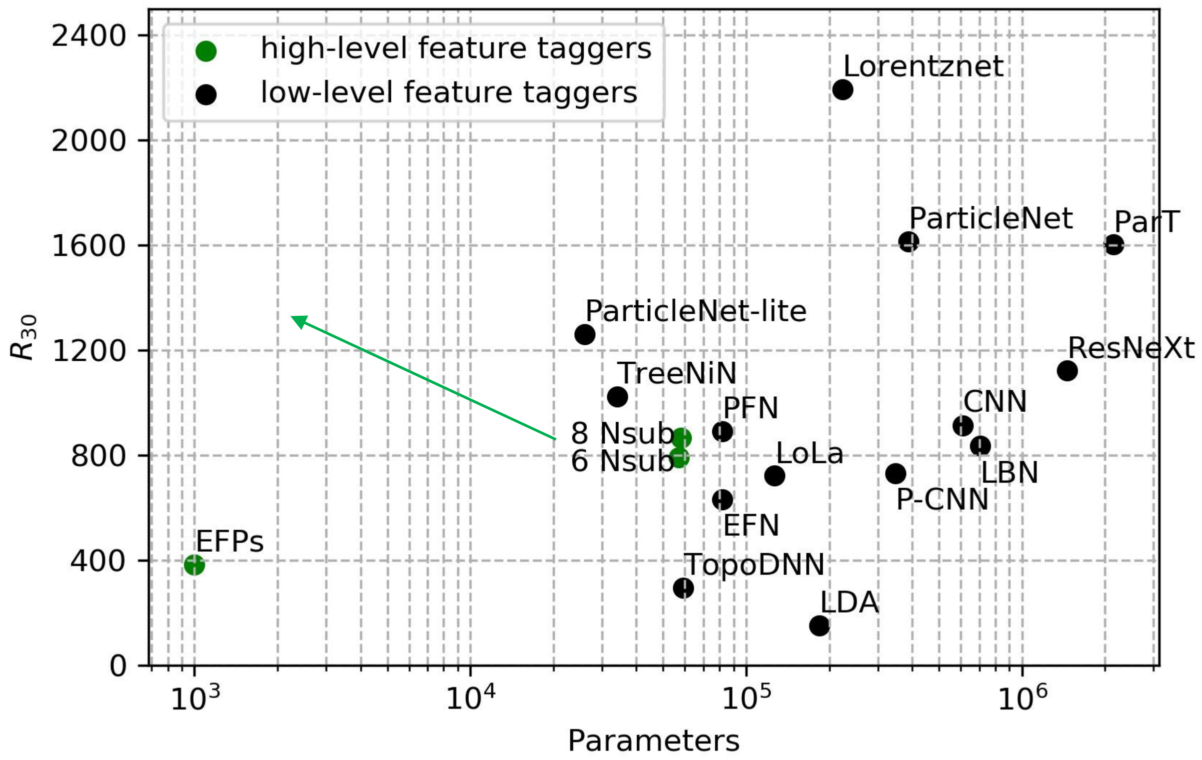
Reports of My Demise Are Greatly Exaggerated: N-subjettiness Taggers Take On Jet Images: arXiv:1807.04769

How Much Information is in a Jet?: arXiv:1704.08249v2

A complete linear basis for jet substructure: arXiv:1712.07124

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.

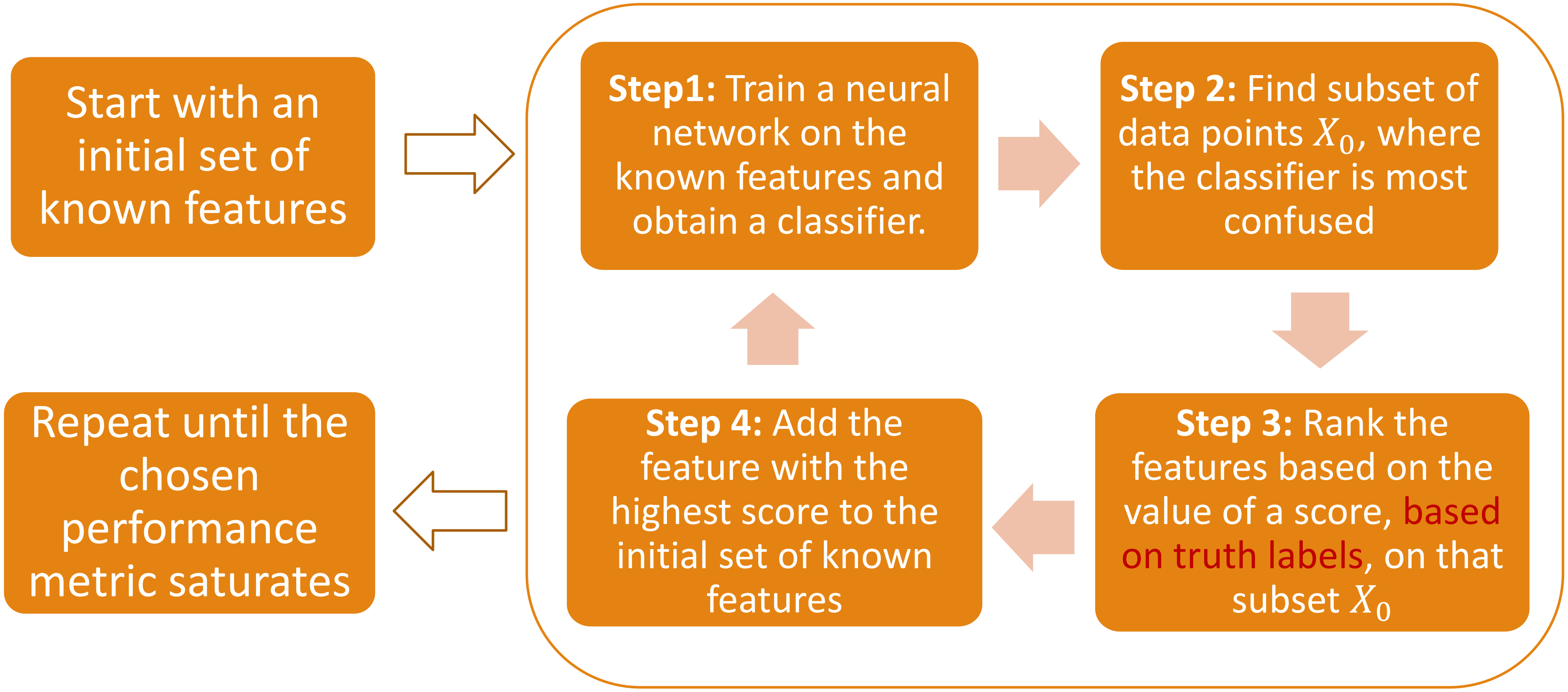
How to do Feature Selection?

- 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 **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 feature selection algorithm which relies only on truth labels

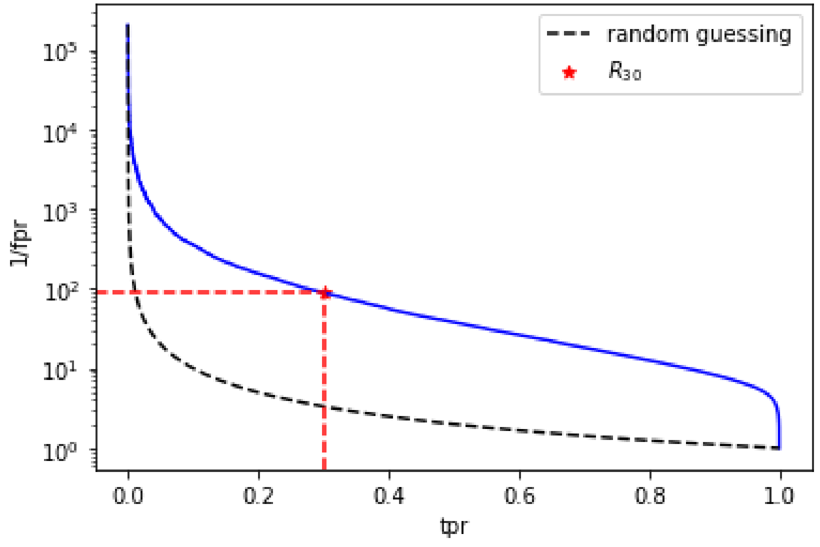


Application of the algorithm to top tagging

- **Data set:** The Machine Learning Landscape of Top Taggers (arXiv:1902.09914v3). (10.5281/zenodo.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:

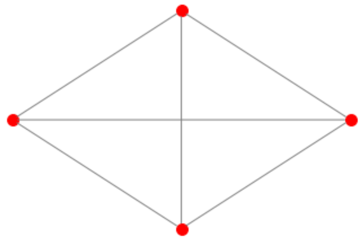
- z_a : The momentum fraction of in a calorimeter cell a
- θ_{ab}: Angular separation between calorimeter cells a and b

$$z_a^{(\kappa)} = \left(\frac{p_{T_a}}{\Sigma_b p_{T_b}}\right)^{\kappa}$$

$$\theta^{(\beta)} = \left(\Delta\,\eta_{ab}^2 + \Delta\phi_{ab}^2\right)^{\frac{\beta}{2}}$$

Energy flow polynomials: A complete linear basis for jet substructure: arXiv:1712.07124
ADO method: Mapping Machine-Learned Physics into a Human-Readable Space arXiv:2010.11998

Features: Energy Flow Polynomials (EFPs)



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

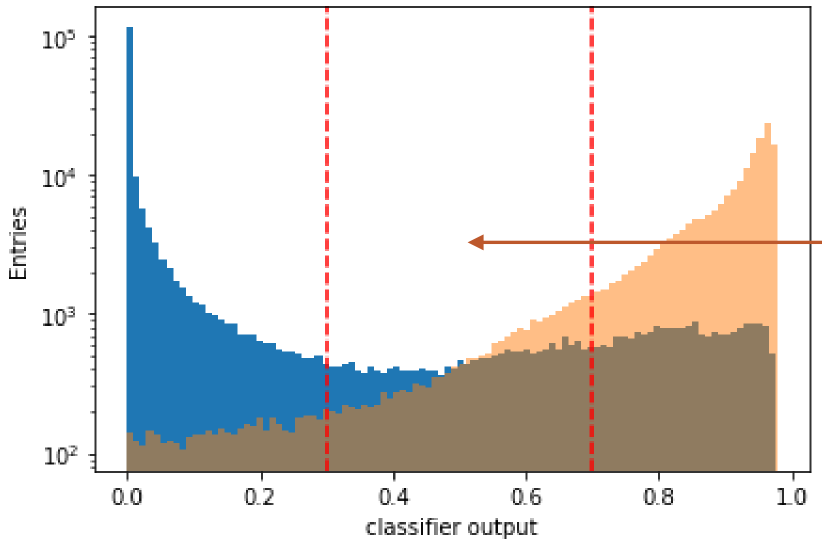
- Each node : $\sum_a z_\mathrm{a}$
- Each k-fold edge : θ_{ab}^{k}

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

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

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.



Data points where the classifier most confused

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

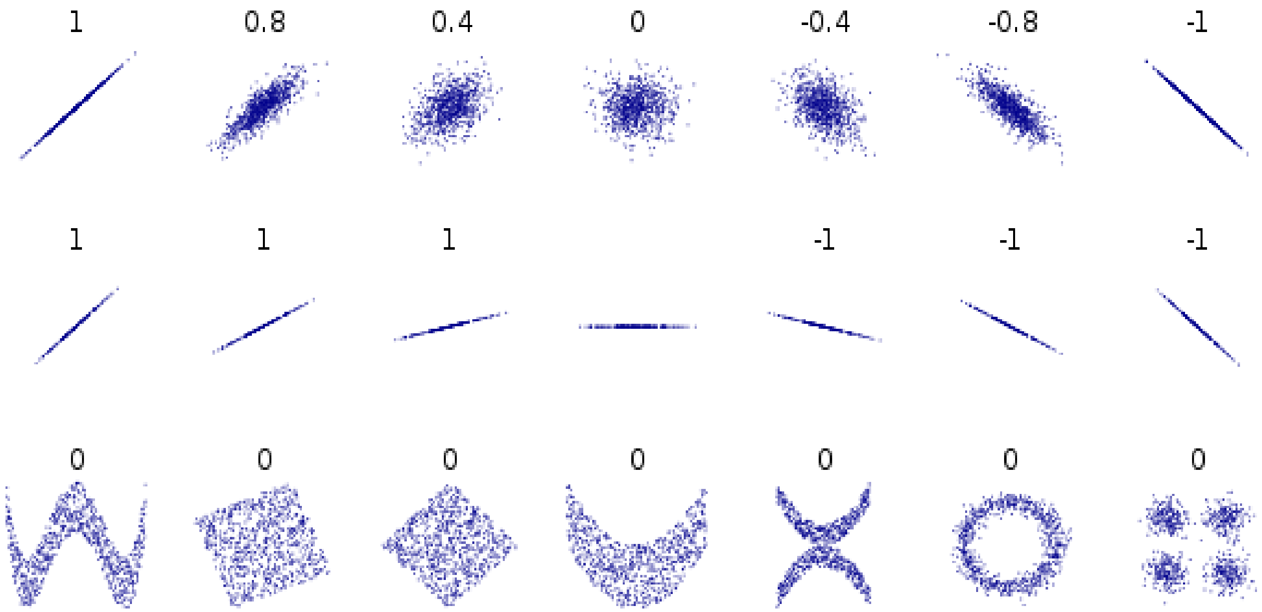
- On X_0 we evaluate: $DisCo(y^{truth}, [known\ variables, new\ feature])$ for each feature in the feature subspace.

Score to rank EFPs: Distance Correlation (DisCo)

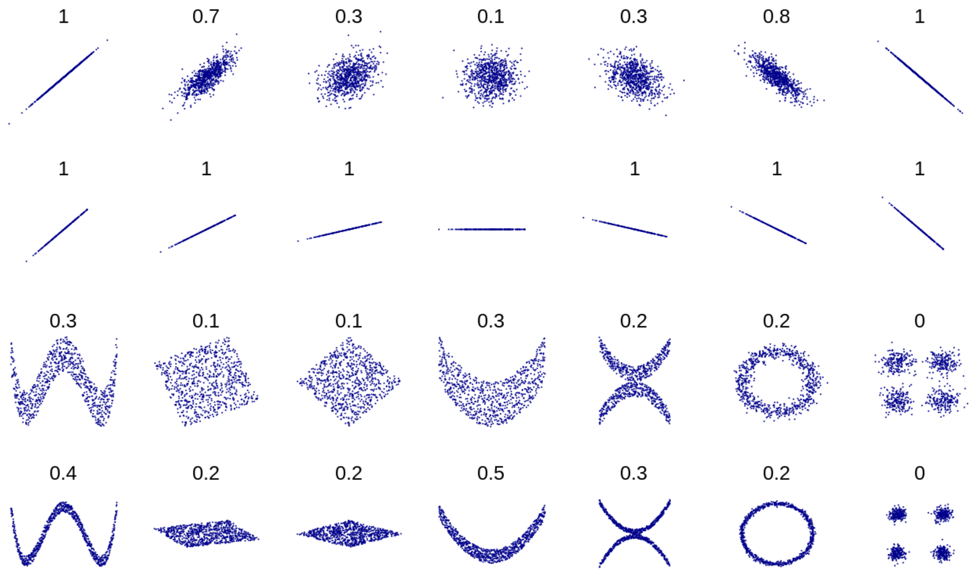
- DisCo is used to find value of **non-linear correlations** of the EFPs with the truth labels
- Very powerful since we can quantify correlations between truth labels and multiple features.

DisCo Fever: Robust Networks Through Distance Correlation: arXiv:2001.05310
Brownian distance covariance: arXiv:1010.0297

Score to rank EFPs: Distance Correlation (DisCo)



Pearson Correlation



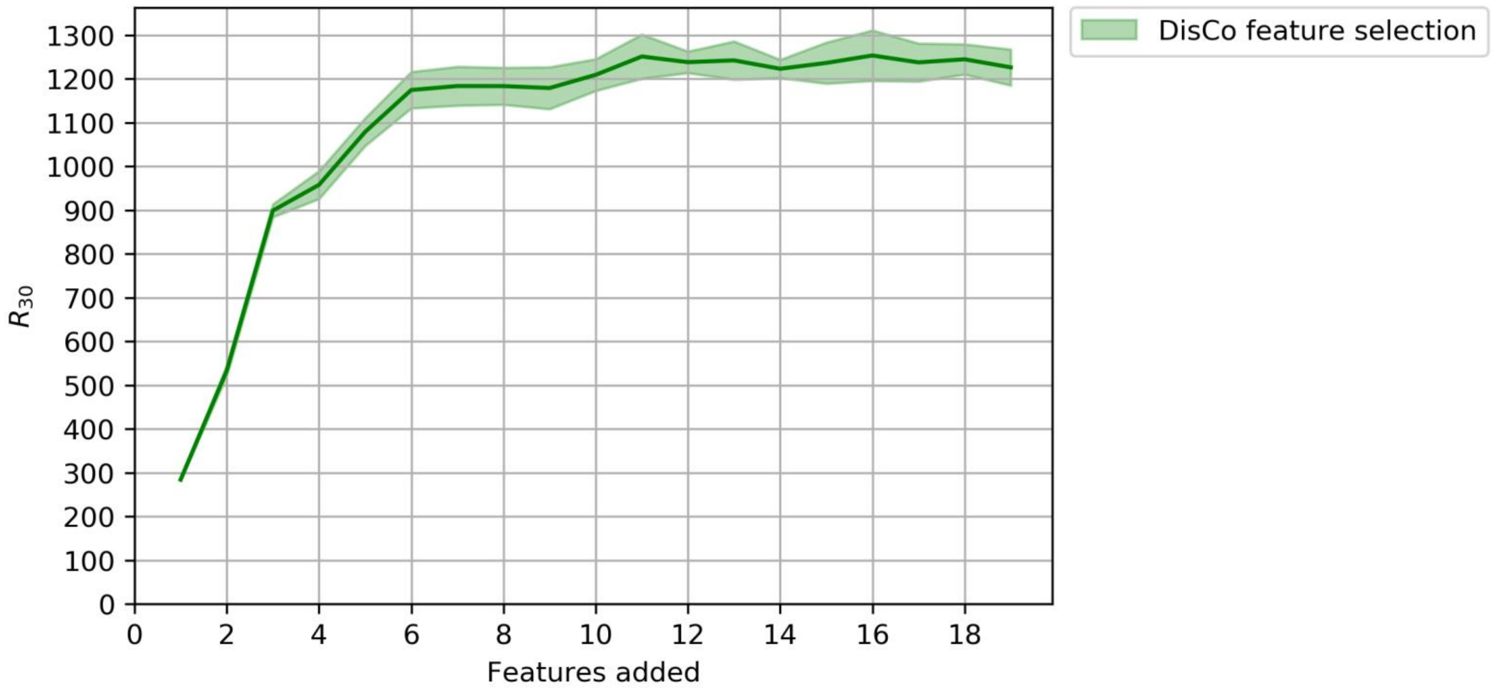
DisCo

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.

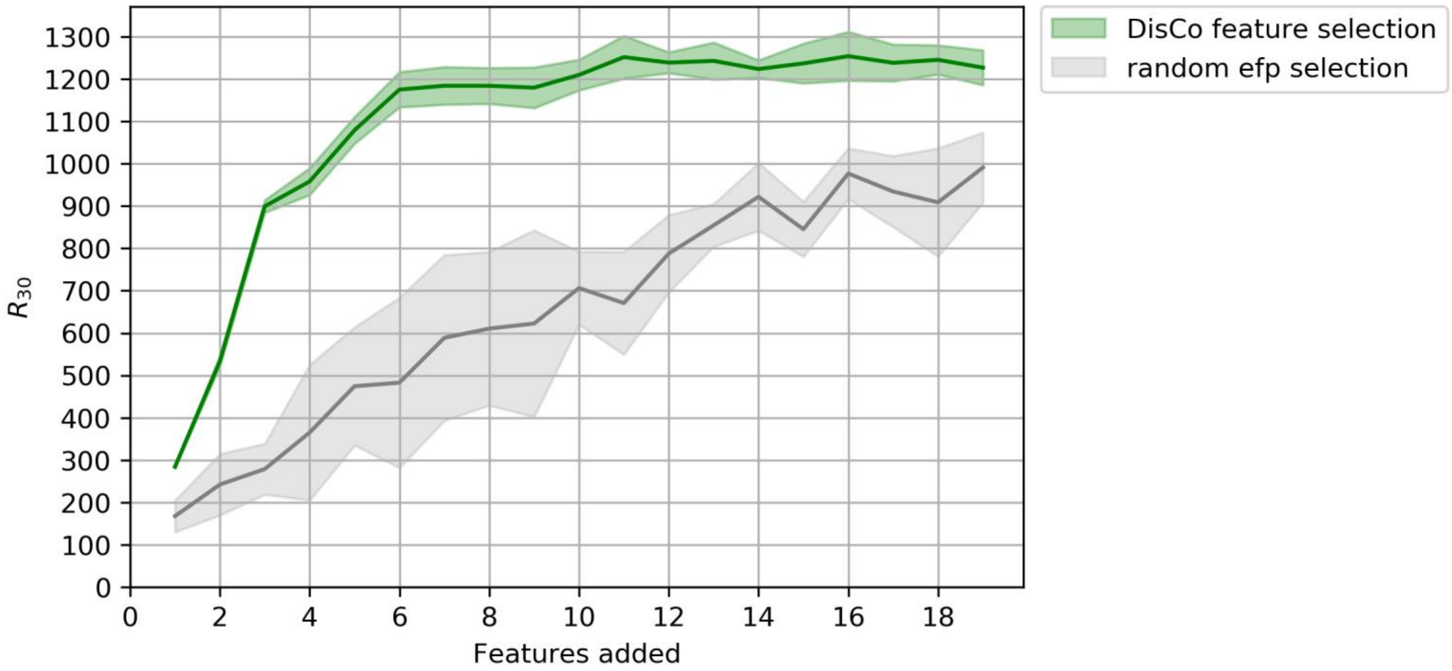
Performance after addition of new EFPs using feature selection algorithm

- Variance for each method is obtained by training each network 10 times.
- Our method can obtain an R_{30} of 1263 ± 50, after 11 features.



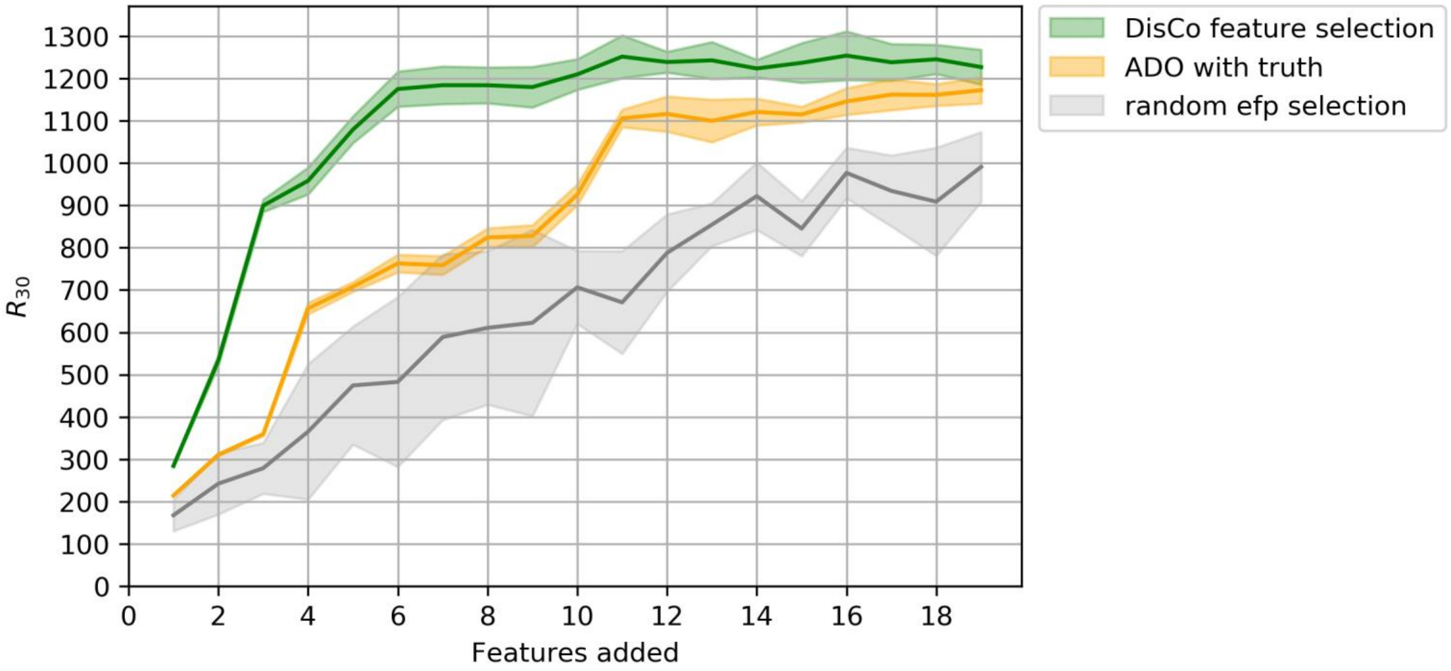
Baseline: Random selection of features

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



Comparison to a previous feature selection algorithm

- 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

*The ADO plot was made using my implementation of the ADO algorithm

Comparison to other top taggers

The Machine Learning Landscape of Top Taggers: arXiv:1902.09914v3

Particle Transformer for Jet Tagging: arXiv:2202.03772

An Efficient Lorentz Equivariant Graph Neural Network for Jet Tagging: arXiv:2201.08187v5

ParticleNet: Jet Tagging via Particle Clouds: arXiv:1902.08570v3

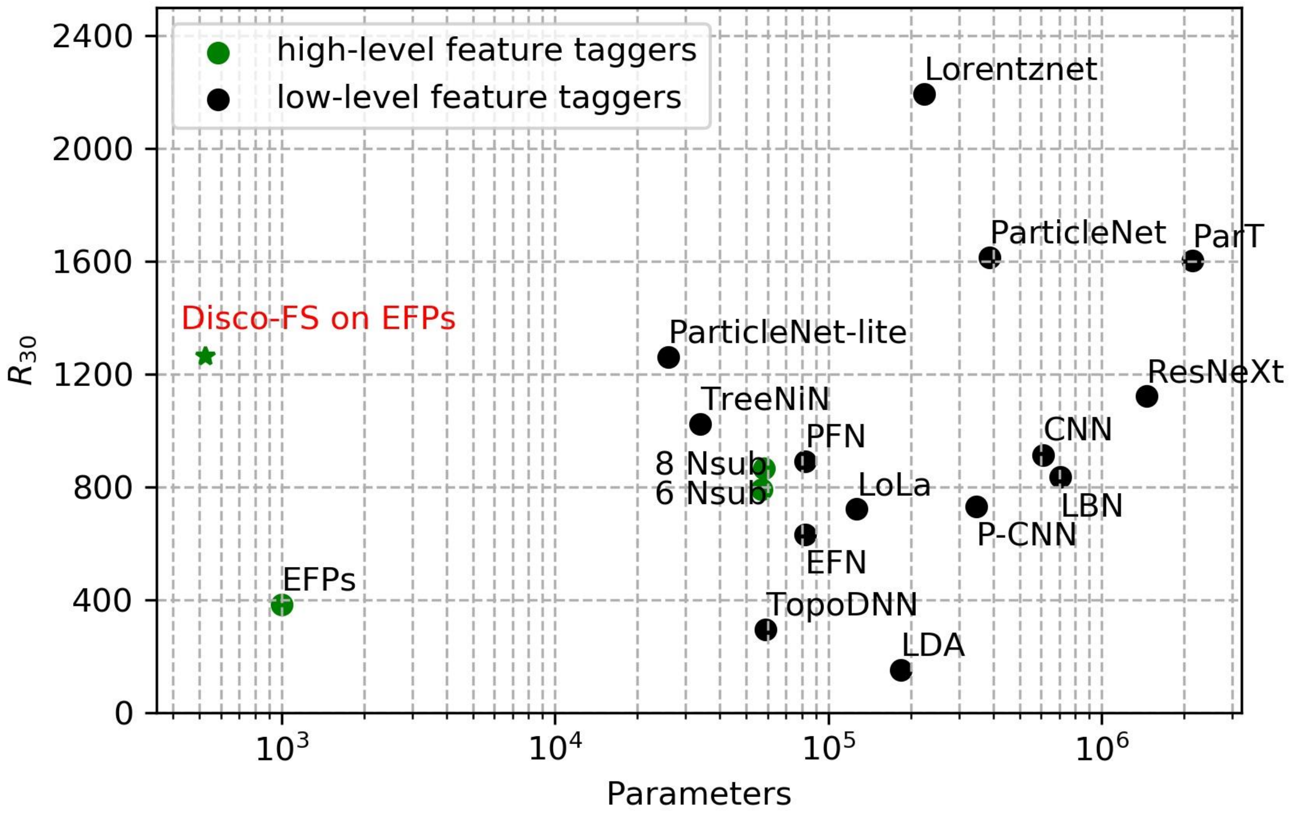
Mapping Machine-Learned Physics into a Human-Readable Space arXiv:2010.11998

Reports of My Demise Are Greatly Exaggerated: N-subjettiness Taggers Take On Jet Images: arXiv:1807.04769

How Much Information is in a Jet?: arXiv:1704.08249v2

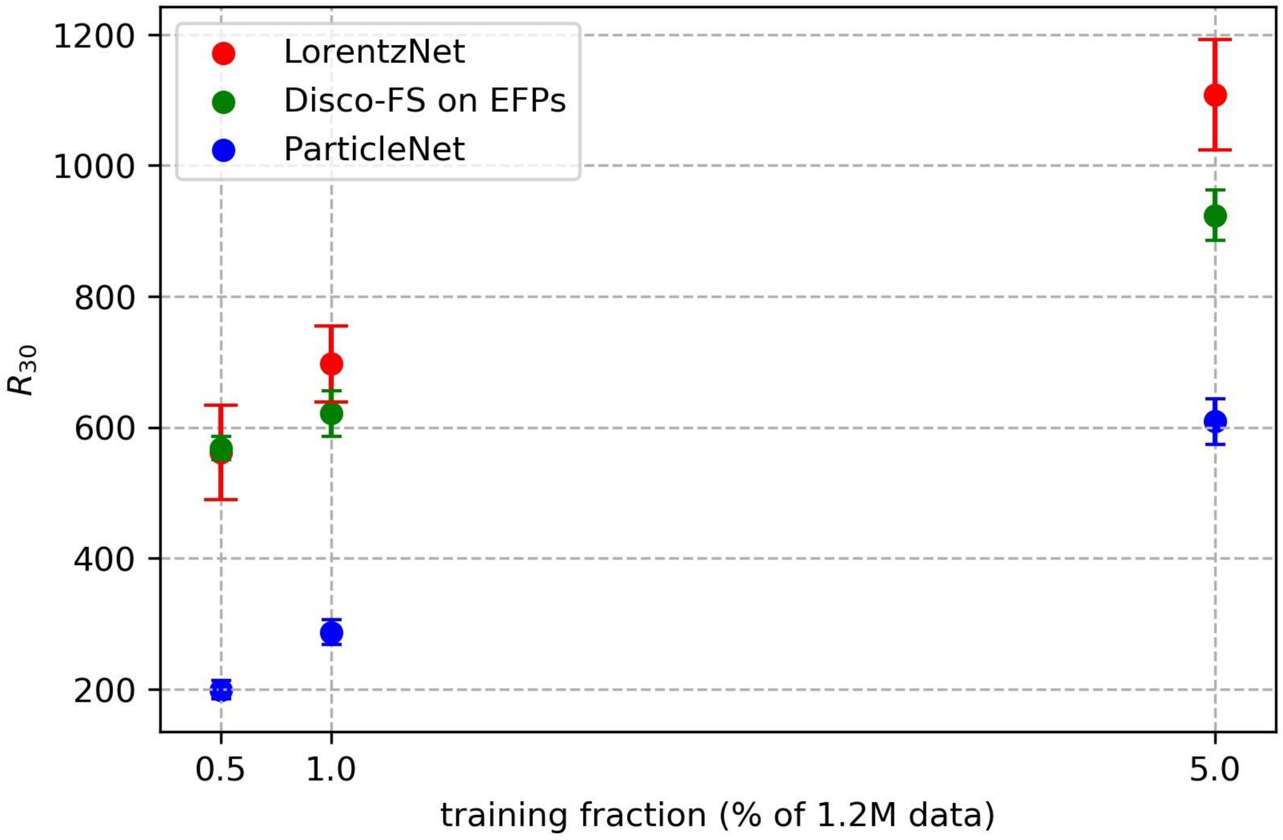
A complete linear basis for jet substructure: arXiv:1712.07124

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.



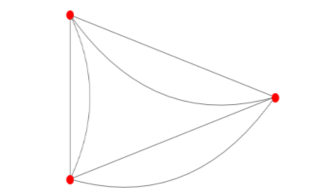
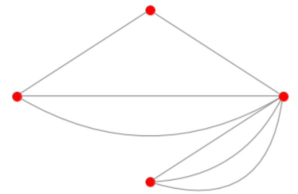
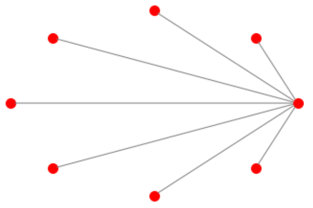
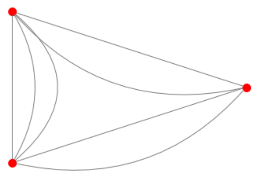
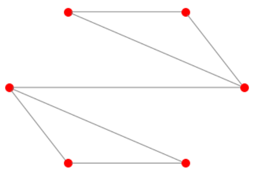
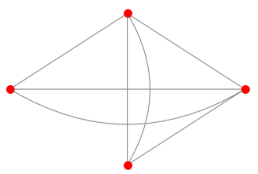
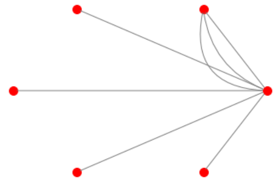
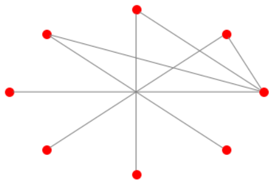
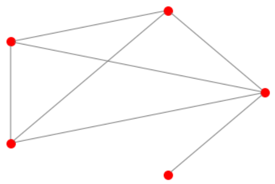
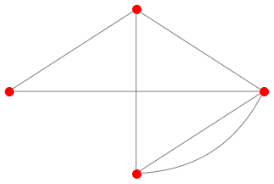
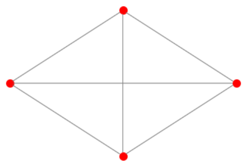
*We use the features, which were selected using the larger dataset.

An Efficient Lorentz Equivariant Graph Neural Network for Jet Tagging: arXiv:2201.08187v5

ParticleNet: Jet Tagging via Particle Clouds: arXiv:1902.08570v3

Selected EFPs

- EFPs with chromatic number c, probes the deviation from (c-1) prong substructure of a jet. The presence of 7 c=3 EFPs, of the 11 EFPs selected, emphasizes the importance of these EFPs for top-tagging.
- We also see the presence c=2, and c=4 EFPs, which shows that deviations from 1-prong and 3-prong substructure information can also be useful.
- EFPs with $\kappa \neq 1$ are IRC unsafe, which shows that IRC-unsafe information can also be useful.

#	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
7		2	0	0.5
8		2	1	1
9		4	2	0.5
10		3	2	0.5
11		4	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.

Possible reasons for not getting a better performance:

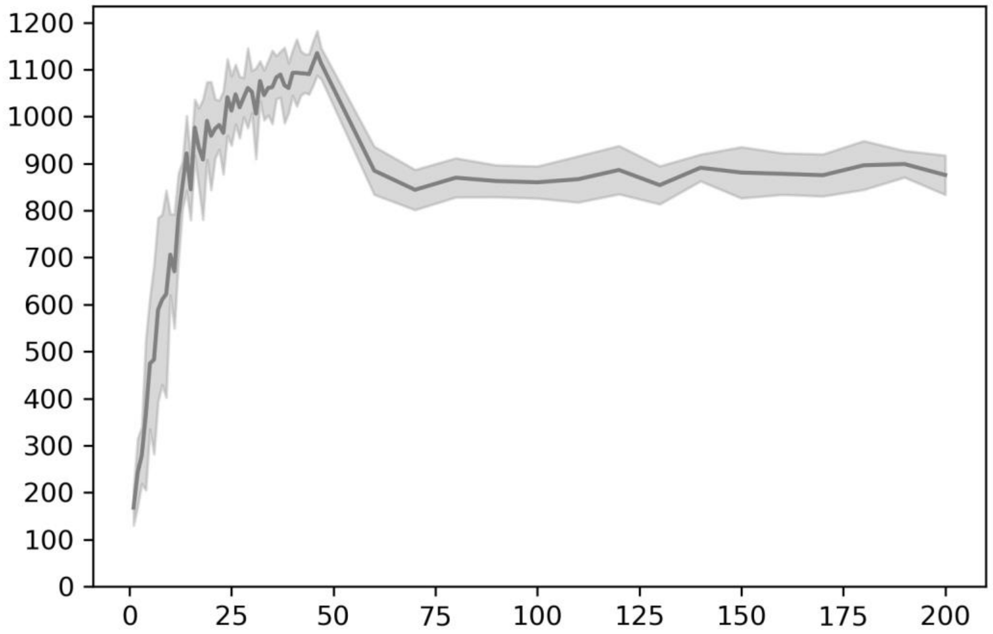
- The feature space considered could be insufficient for top tagging, which could explain our inability to close the gap with higher performing black box models.
- Need a better feature selection algorithm

Paper coming soon.

Thank You!

BACK UP SLIDES

Random Selection



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=1 . 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.

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

Measuring and testing dependence by correlation of distances: arXiv:0803.4101

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, $ADO(y^{truth/background}, new\ feature)$

Comparison to other top taggers

Taggers	R_{30}	Parameters
CNN	914±14	610k
ResNeXt	1122±47	1.46M
TopoDNN	295±5	59k
Multi-body N-subjettiness 6	792±18	57k
Multi-body N-subjettiness 8	867±15	58k
TreeNiN	1025±11	34k
P-CNN	732±24	348k
LBN	836±17	705k
LoLa	722±17	127k
LDA	151±0.4	184k
EFPs	384	1k
EFN	633±31	82k
PFN	891±18	82k
ParticleNet	1615 ± 93	366k
ParticleNet-Lite	1262 ± 49	26k

The Machine Learning Landscape of Top Taggers: arXiv:1902.09914v3
An Efficient Lorentz Equivariant Graph Neural Network for Jet Tagging: arXiv:2201.08187v5
ParticleNet: Jet Tagging via Particle Clouds: arXiv:1902.08570v3