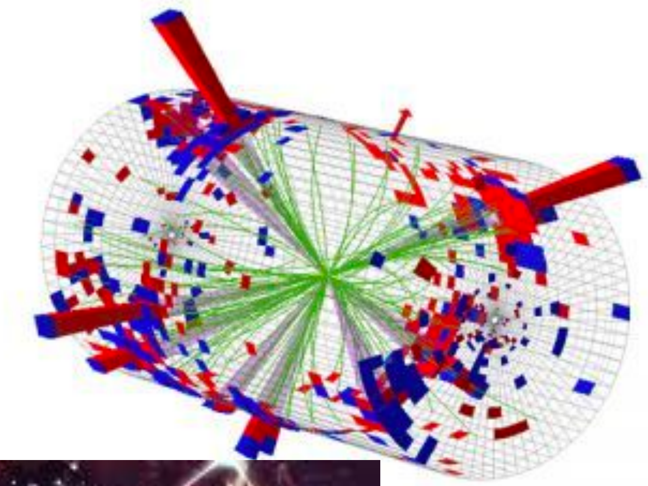
RUTGERS



DisCo Decorrelation:

Robustifying classifiers and automating the ABCD method

David Shih

4th IML Workshop
October 23, 2020

Work done in collaboration with

Motivation: beyond classification

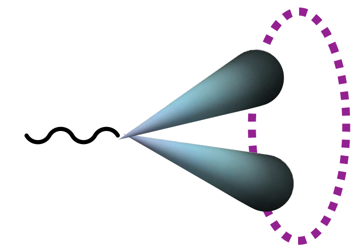
There has been enormous progress in the past ~5 years in improving jet classifiers with deep learning.

Now there is increasing interest in applications of deep learning to other issues necessary for realistic applications, beyond raw classification performance.

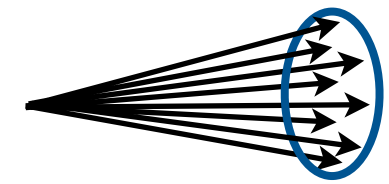
One important issue is the need for ***robust classifiers*** which are stable against variations in an auxiliary feature.

- data vs. simulation validation
- data-driven background estimation
- reducing systematic uncertainties

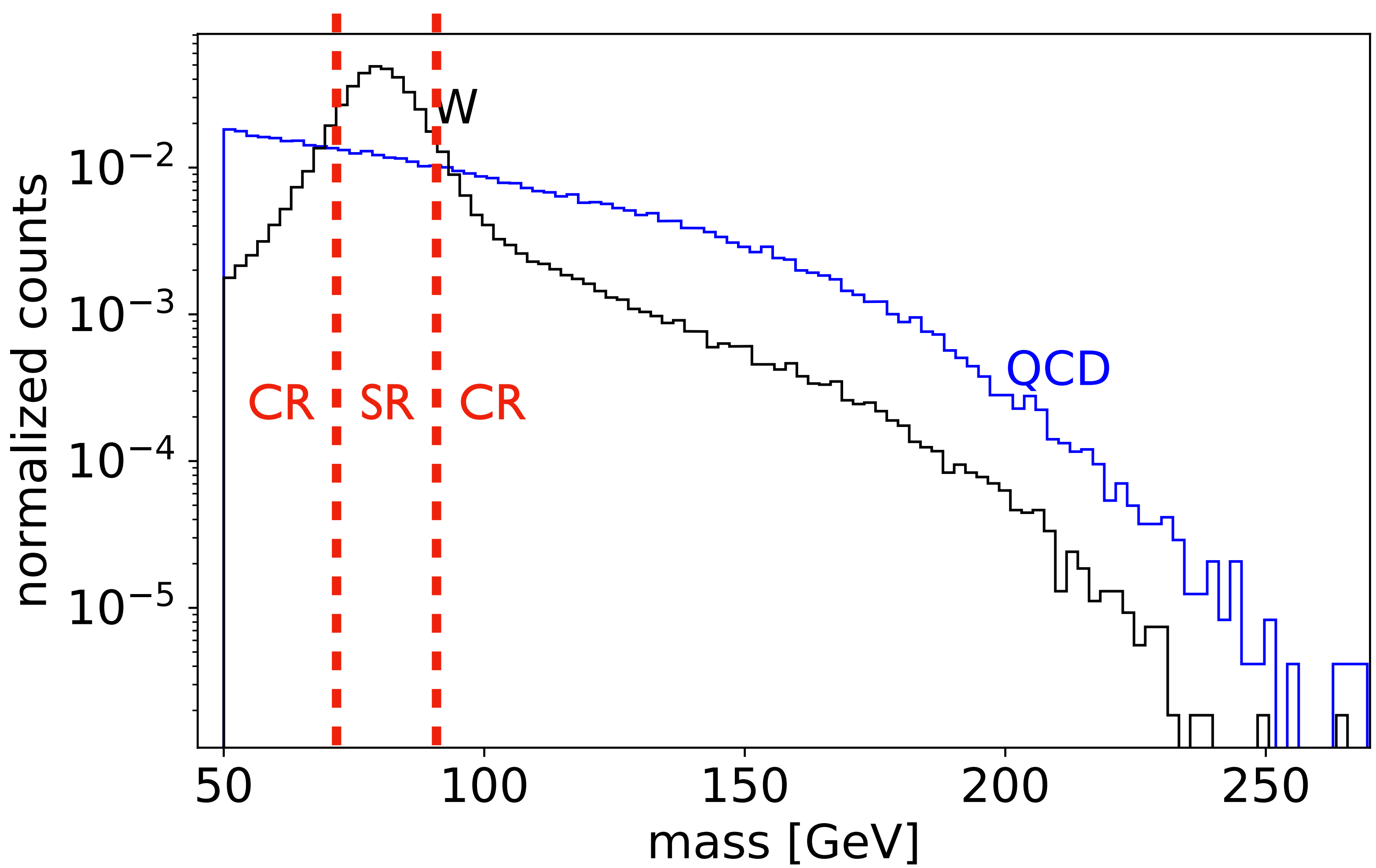
Motivation: beyond classification



vs.

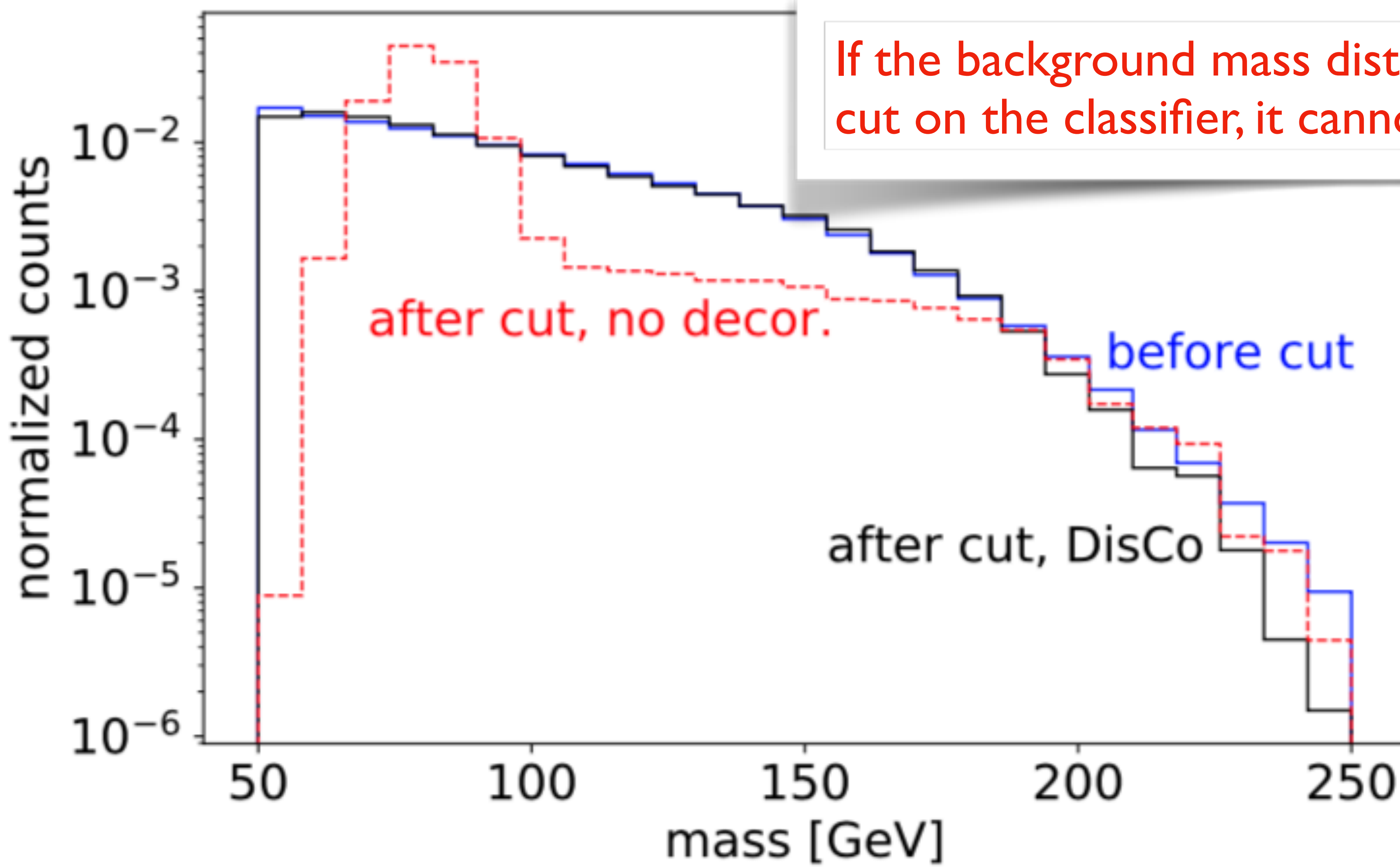


Example: boosted hadronic W tagging



Typically use jet mass to define signal and control regions in data, for validation and/or background estimation.

Motivation: beyond classification



If the background mass distribution is sculpted by a cut on the classifier, it cannot be used for sidebanding.

Would like a classifier that doesn't "learn" the jet mass, i.e. is statistically independent from it.

Challenging, because many of the input features to the classifier are highly correlated with mass.

State of the art in *mass decorrelation methods* was studied by ATLAS for boosted W-tagging in ATL-PHYS-PUB-2018-014

Variable	Type	Reference
$C_2,\ D_2$	Energy correlation ratios	[38]
τ_{21}	N-subjettiness	[41]
R_2^{FW}	Fox–Wolfram moment	[42]
$\mathcal{P}$	Planar flow	[43]
a_3	Angularity	[44]
A	Aplanarity	[45]
$Z_{\mathrm{cut}},\ \sqrt{d_{12}}$	Splitting scales	[46, 47]
$KtDR$	k_t-subjet ΔR	[48]

Cut-based

or

BDT

or

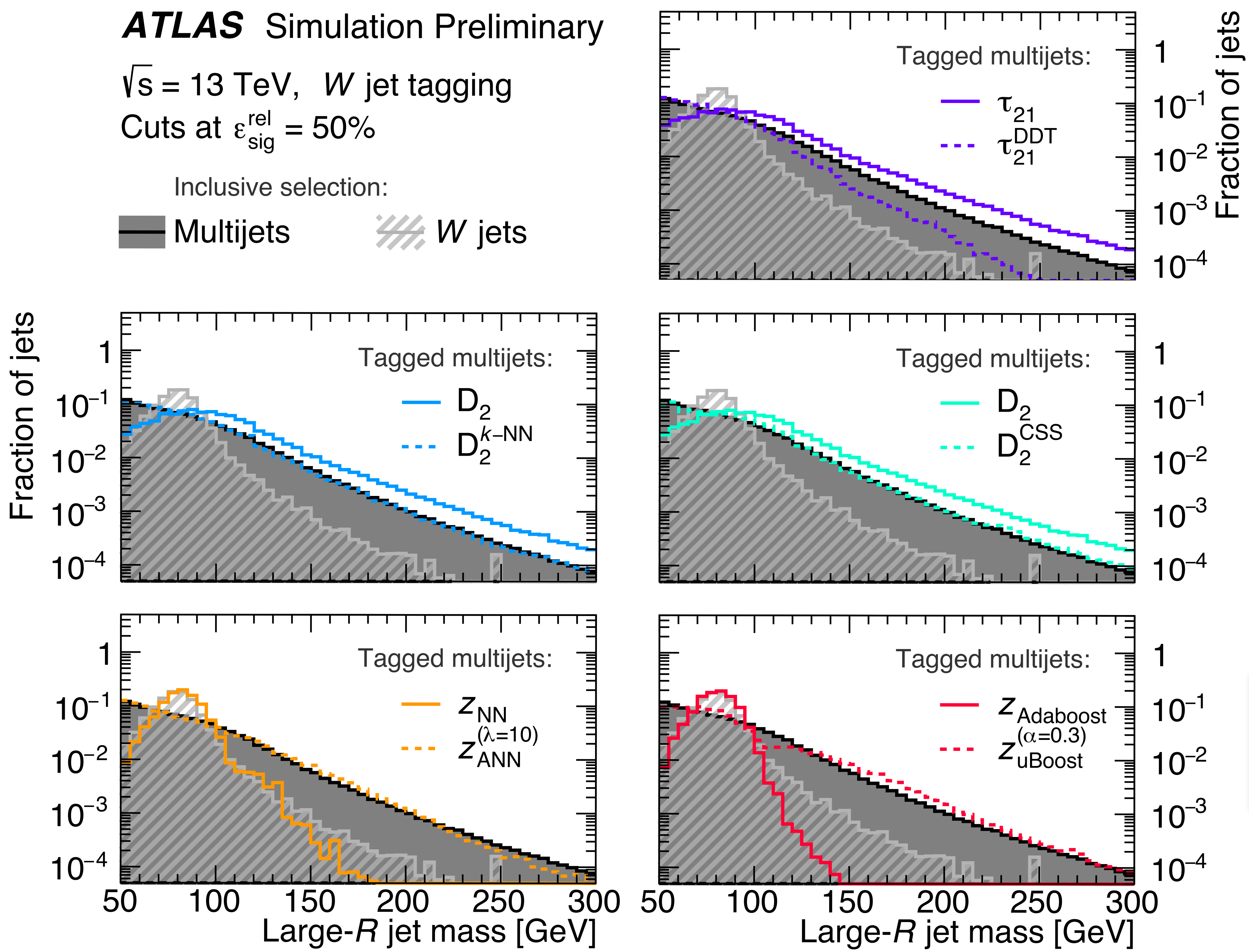
Dense NN (DNN)

Performance metrics:

R50: 1/[background efficiency (false positive rate) at 50% signal efficiency]

JSD50: Jensen-Shannon Divergence between bg mass histogram (all) and bg mass histogram (passing sig=50% cut)

ATL-PHYS-PUB-2018-014



Analytical / single-variable

Multivariate (MVA)

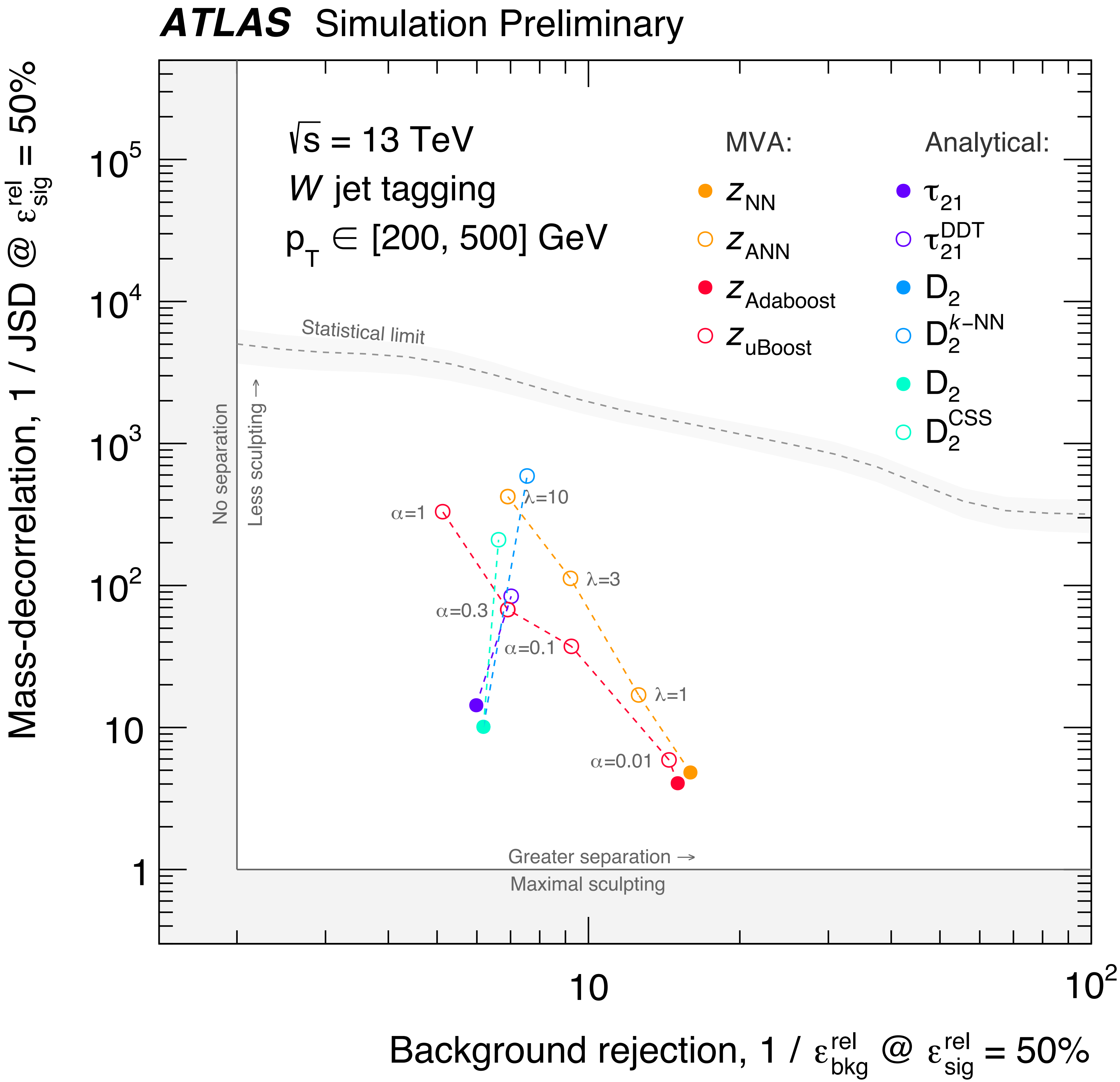
DDT

k-NN

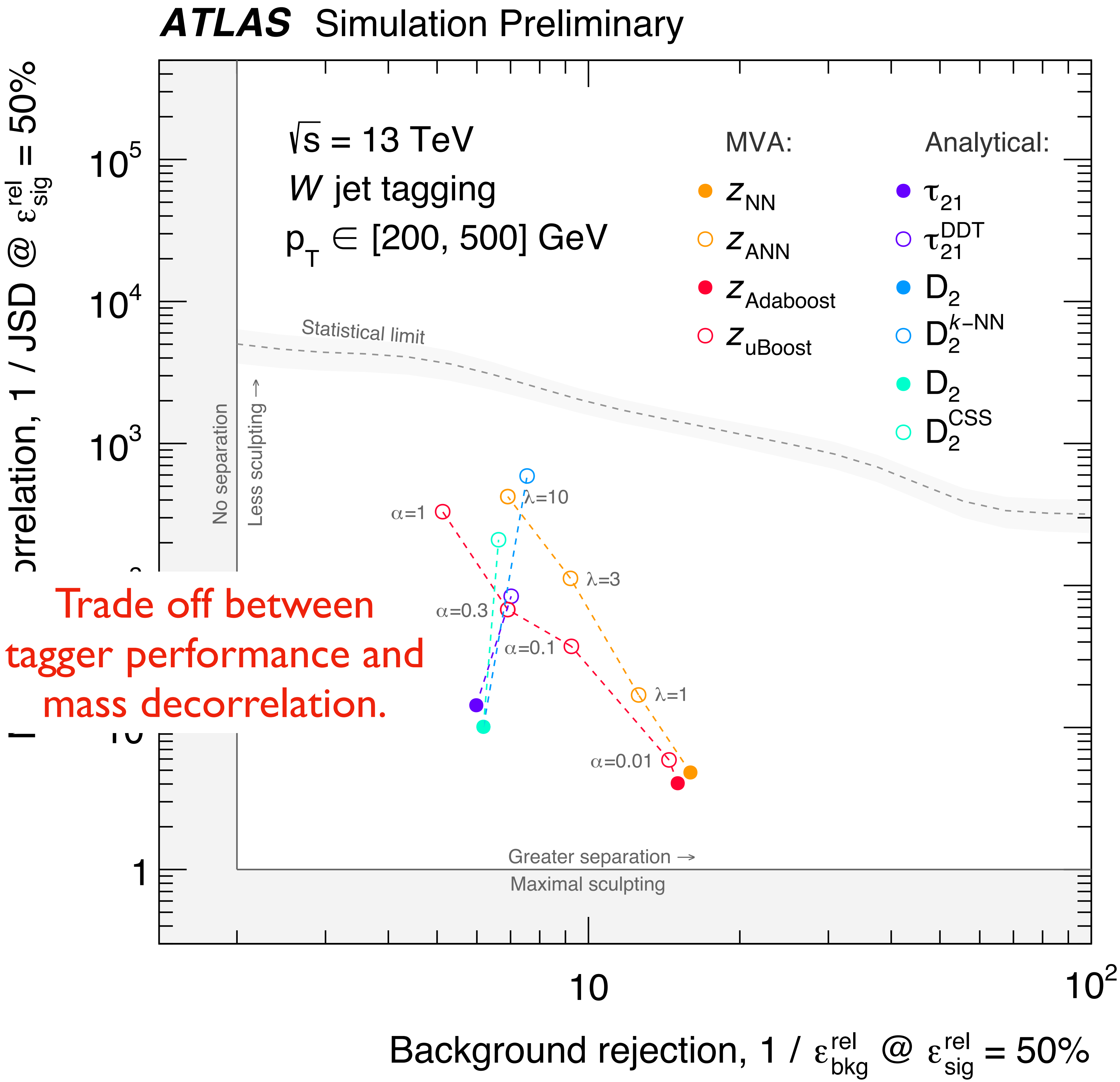
CSS

ANN

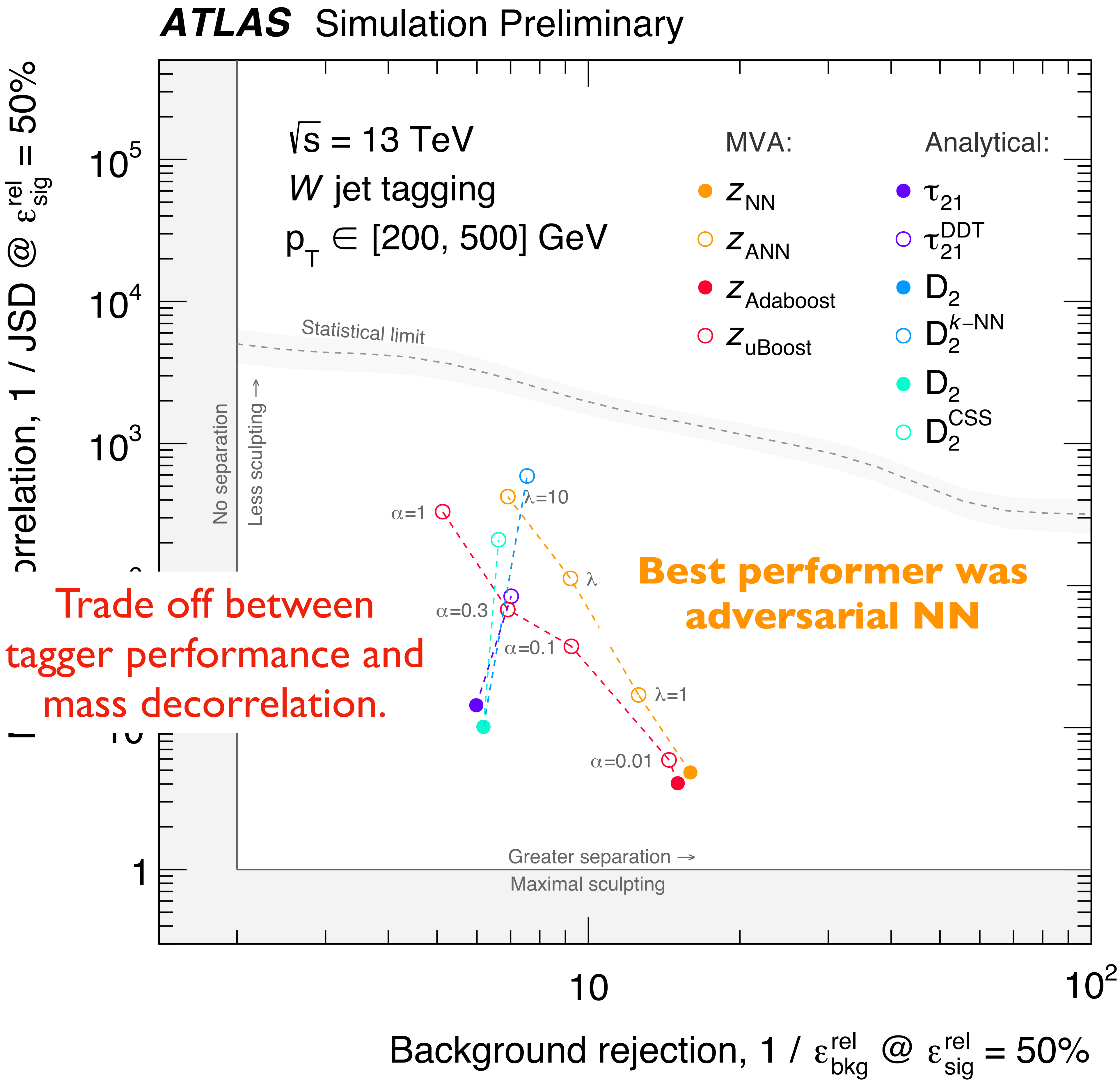
uBoost



ATLAS Simulation Preliminary
√s = 13 TeV
W jet tagging
p_T ∈ [200, 500] GeV
MVA:
z_NN
z_ANN
z_Adaboost
z_uBoost
Analytical:
τ_21
τ_21^DDT
D_2
D_2^k-NN
D_2
D_2^CSS
Statistical limit
No separation
Less sculpting →
Greater separation →
Maximal sculpting
α=1
α=0.3
α=0.1
α=0.01
λ=10
λ=3
λ=1
Mass-decorrelation, 1 / JSD @ ε_sig^rel = 50%
Background rejection, 1 / ε_bkg^rel @ ε_sig^rel = 50%
1
10
10^2
10^3
10^4
10^5
10
10^2



Trade off between tagger performance and mass decorrelation.

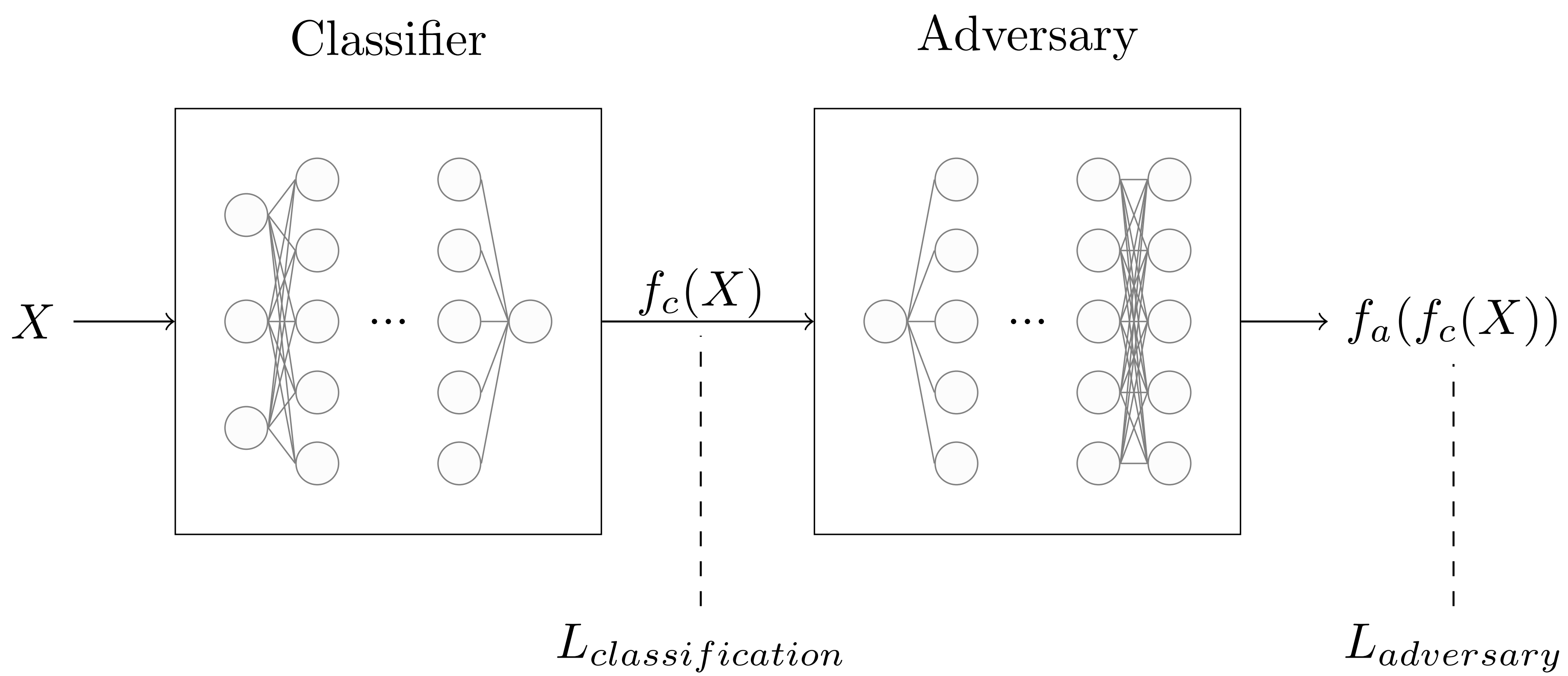


Trade off between tagger performance and mass decorrelation.

Best performer was adversarial NN

Adversarial decorrelation

Louppe et al 1611.01046, Shimmin et al 1703.03507

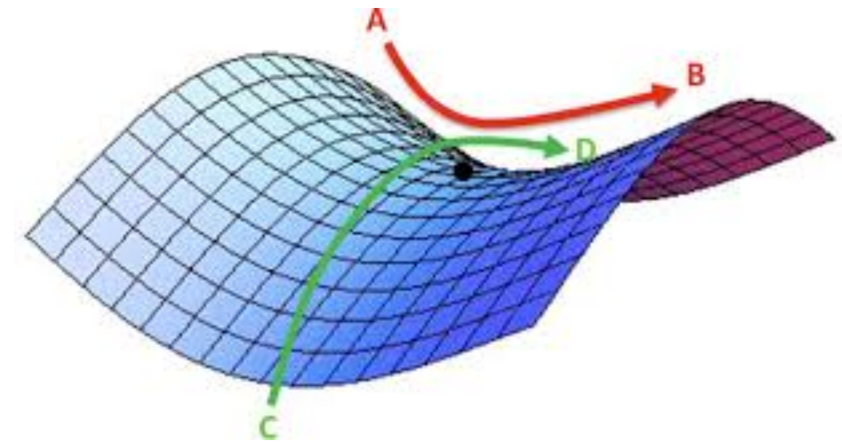


$$L_{\text{tagger}} = L_{\text{classification}} - \lambda L_{\text{adversary}}$$

Idea: train a second neural network (the "adversary") that attempts to predict the mass from the classifier output.

If classifier and mass are independent, adversary will fail.

Alternatives to adversaries



Adversaries are notoriously **tricky to train** — saddle point optimization

$$\min_{\theta_{\mathrm{clf}}} \max_{\theta_{\mathrm{adv}}} L_{\mathrm{clf}}(y(\theta_{\mathrm{clf}})) - \lambda L_{\mathrm{adv}}(y(\theta_{\mathrm{clf}}), m; \theta_{\mathrm{adv}})$$

Would be great if we could achieve the same performance but with a convex regularizer term

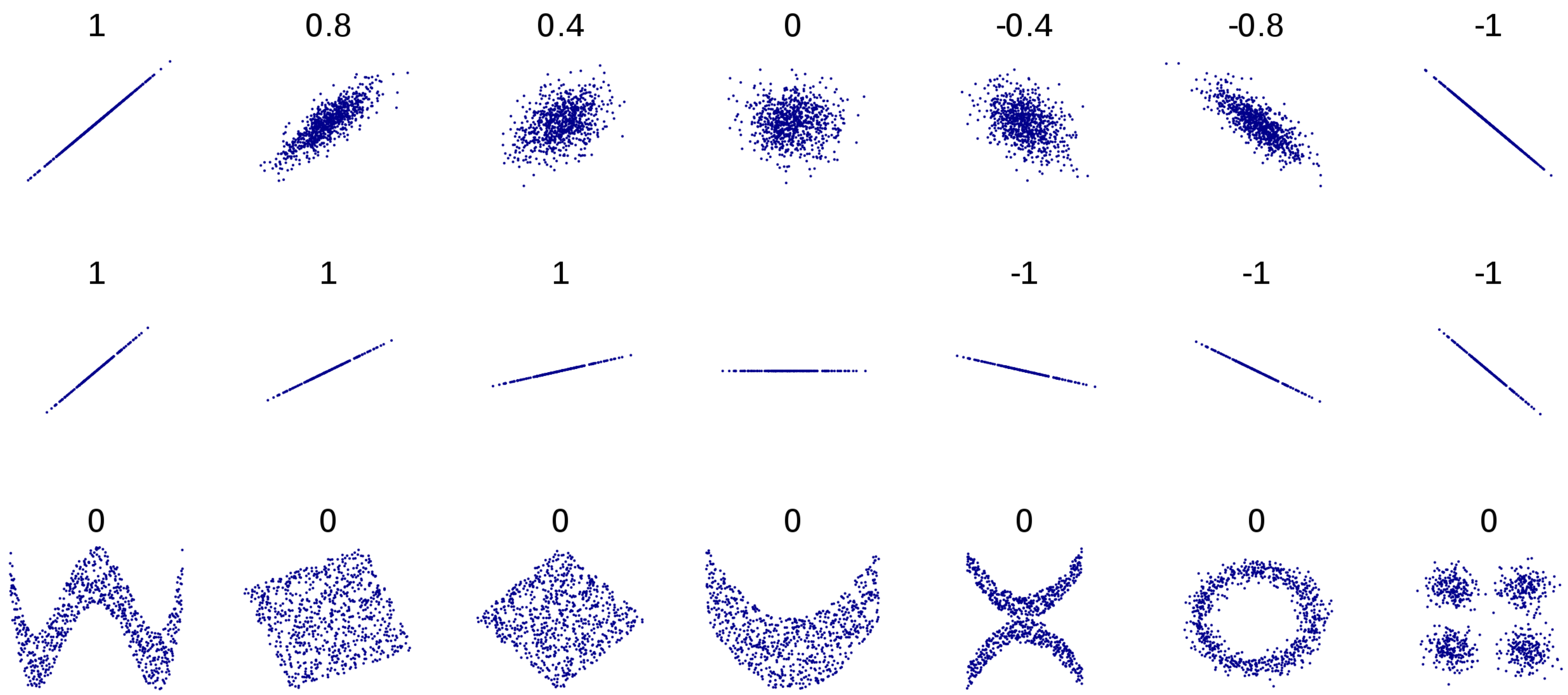
$$\min_{\theta_{\mathrm{clf}}} L_{\mathrm{clf}}(y(\theta_{\mathrm{clf}})) + \lambda C_{\mathrm{reg}}(y(\theta_{\mathrm{clf}}), m)$$

First idea: can we just use Pearson correlation coefficient?

$$C_{\mathrm{reg}} = R(y, m) \propto \sum_i y_i m_i$$

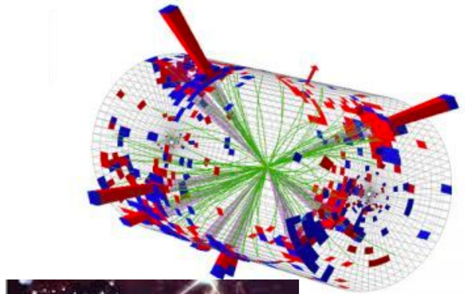
Problem: this only measures linear correlations

Pearson correlation



y and m can be highly correlated yet R=0

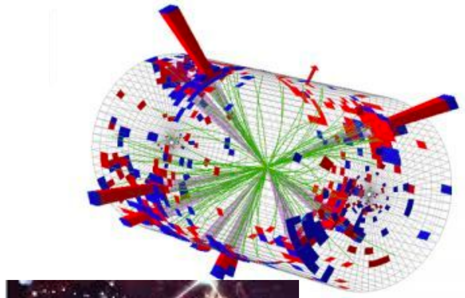
Distance correlation ("DisCo")



$$
\begin{aligned}
\mathrm{dCov}^2(X,Y) = \langle |X - X'||Y - Y'|\rangle \\
+ \langle |X - X'|\rangle\langle |Y - Y'|\rangle \\
- 2\langle |X - X'||Y - Y''|\rangle
\end{aligned}
$$

(Szekely, Rizzo, Bakirov 2007; Szekely & Rizzo 2009)

Distance correlation ("DisCo")



$$\mathrm{dCov}^2(X,Y) = \langle |X-X'||Y-Y'|\rangle + \langle |X-X'|\rangle\langle |Y-Y'|\rangle - 2\langle |X-X'||Y-Y''|\rangle$$

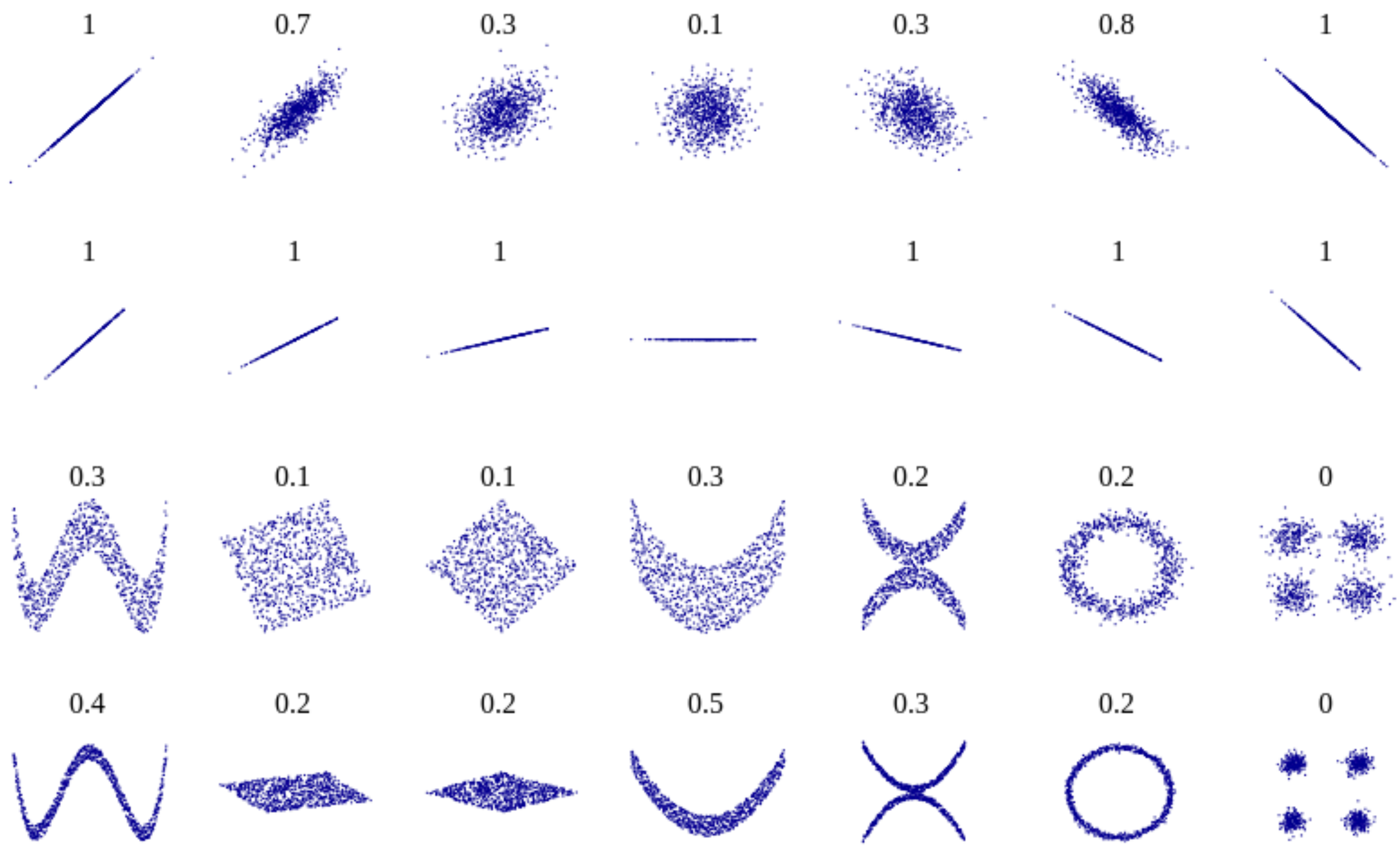
(Szekely, Rizzo, Bakirov 2007; Szekely & Rizzo 2009)

- Zero iff X,Y are statistically independent; positive otherwise
- Tractable, can be estimated from finite samples
- **Idea: Add DisCo to loss function during classifier training**

$$L = L_{classifier}(\vec{y}, \vec{y}_{true}) + \lambda\, \mathrm{dCorr}^2_{y_{true}=0}(\vec{m}, \vec{y})$$

Gregor Kasieczka & DS, PRL 125 (2020), 2001.05310

Distance correlation (“DisCo”)



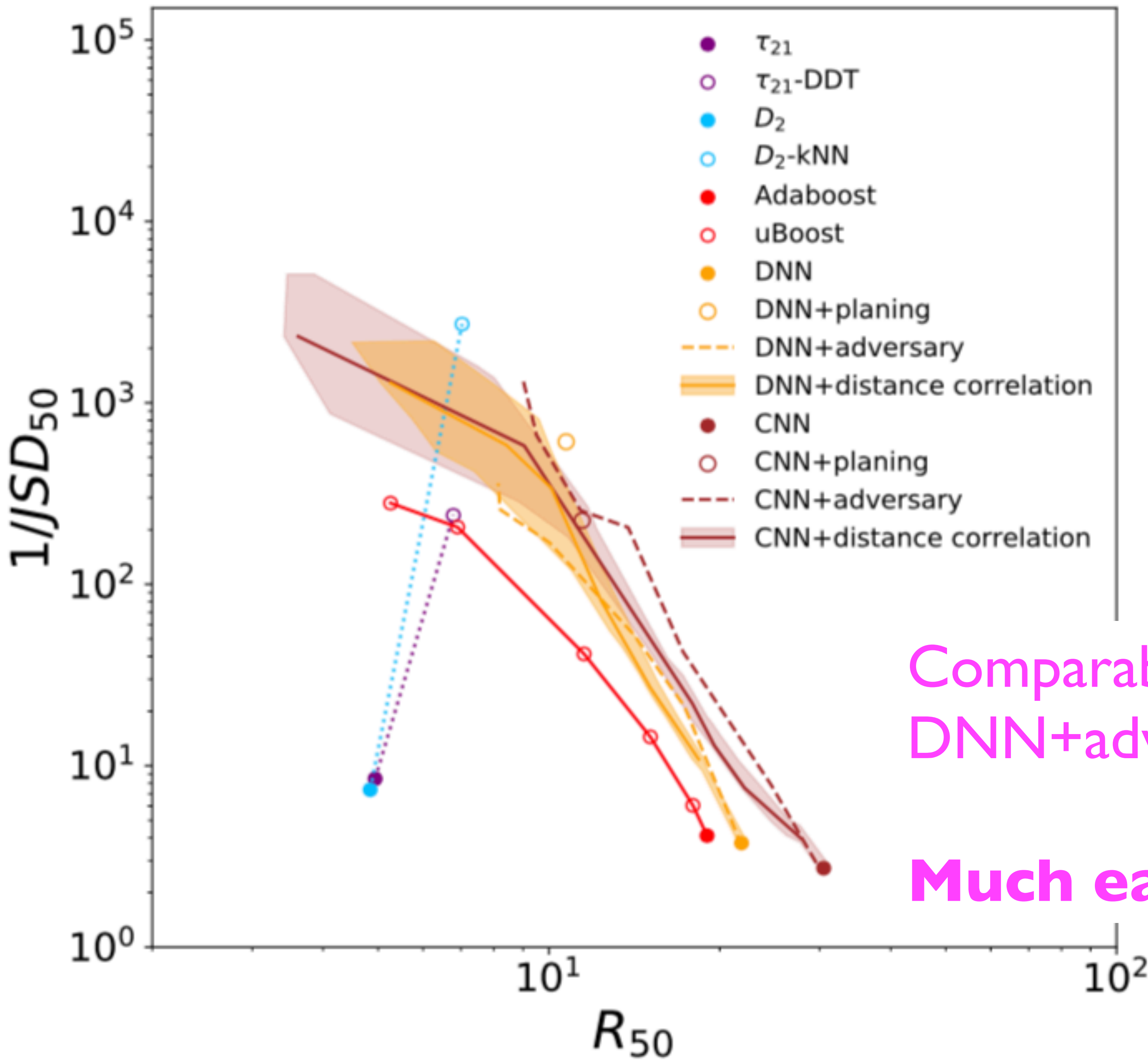
Disco is sensitive to nonlinear correlations!

DisCo Decorrelation

Gregor Kasieczka & DS, PRL 125 (2020), 2001.05310

Our recast of the ATLAS study using Pythia8 + Delphes + FastJet plugins

Samples available on [Zenodo]

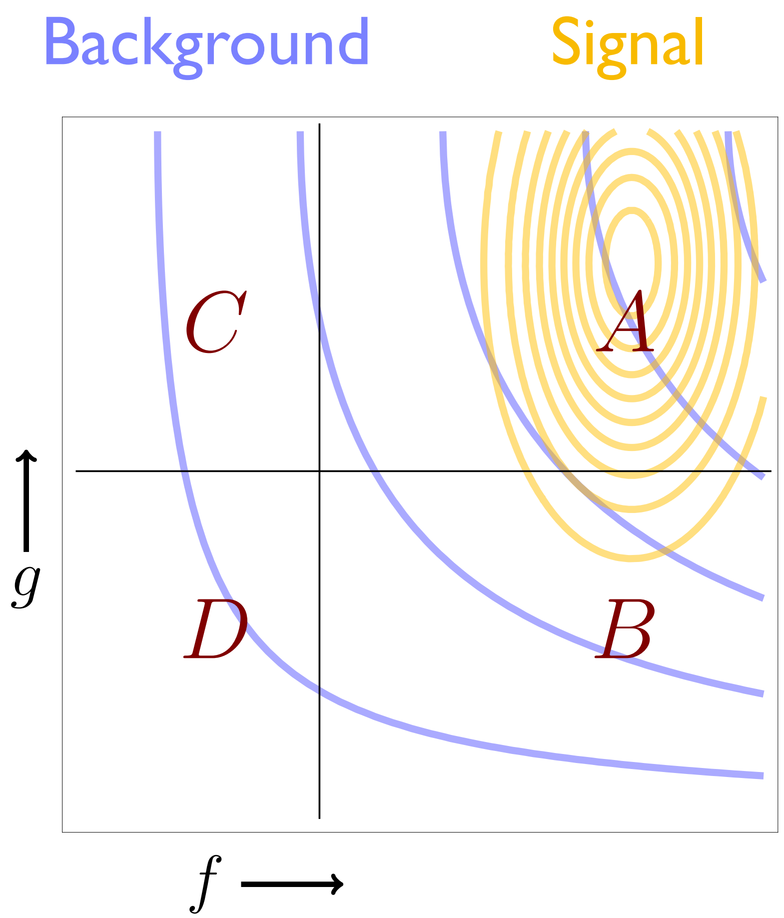


Comparable performance to DNN+adversary.

Much easier to train.

ABCD Method

Another place where statistically independent features are required is the widely used "ABCD method" for data-driven background estimation



If f and g are statistically independent for the background, then:

$$N_{A,bg} = \frac{N_{B,bg} N_{C,bg}}{N_{D,bg}}$$

Usually features f and g are simple kinematic quantities chosen "by-hand"...

ABCDisCo

Kasieczka, Nachman, Schwartz & DS 2007.14400

Idea: could construction of f and/or g be automated using NNs+DisCo?

Single ABCDisCo: decorrelate NN classifier against fixed feature (eg mass)

$$\mathcal{L}[f(X)] = \mathcal{L}_{\text{classifier}}[f(X), y] + \lambda\, \text{dCorr}^2_{y=0}[f(X), X_0]$$

Double ABCDisCo: decorrelate two NN classifiers against each other

$$\mathcal{L}[f, g] = \mathcal{L}_{\text{classifier}}[f(X), y] + \mathcal{L}_{\text{classifier}}[g(X), y] + \lambda\, \text{dCorr}^2_{y=0}[f(X), g(X)]$$

Role of signal contamination

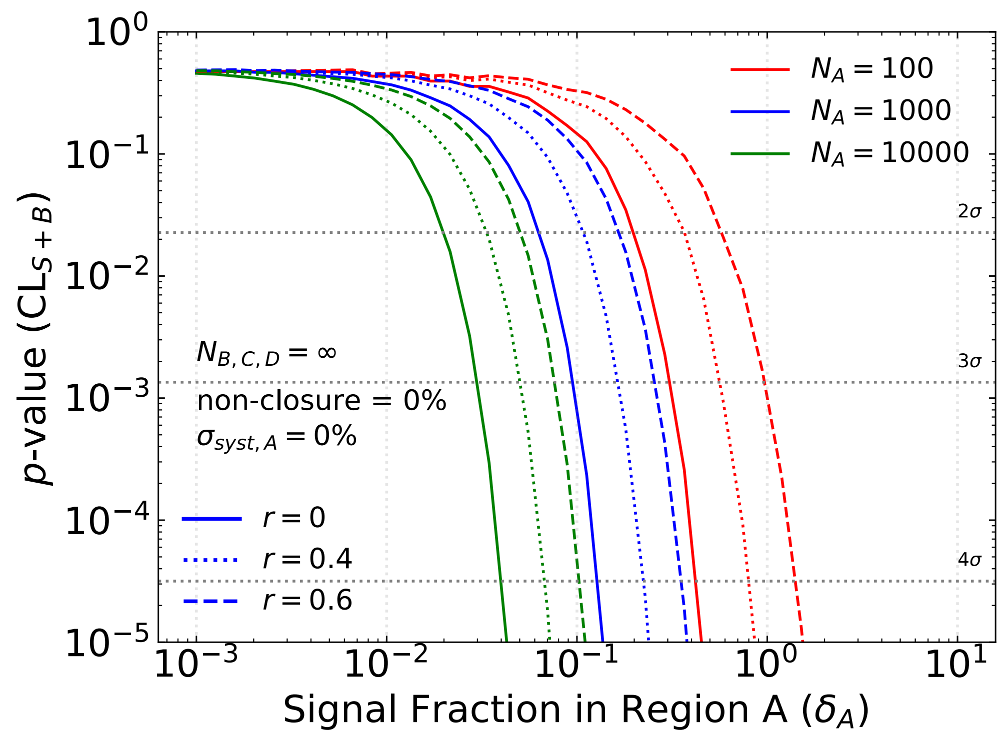
$$N_{A,bg} = \frac{N_{B,bg} N_{C,bg}}{N_{D,bg}} \quad \Rightarrow \quad N_{A,bg} \approx \frac{N_B N_C}{N_D}$$

Key point (neglected in previous analyses?): can only estimate background in A using ***data*** in B,C,D provided that the signal contamination in B,C,D is negligible *relative to the signal fraction in A.*

$$r = \left(\frac{N_{A,s}}{N_{A,b}}\right)^{-1} \left(\frac{N_{B,s}}{N_{B,b}} + \frac{N_{C,s}}{N_{C,b}} - \frac{N_{D,s}}{N_{D,b}}\right)$$

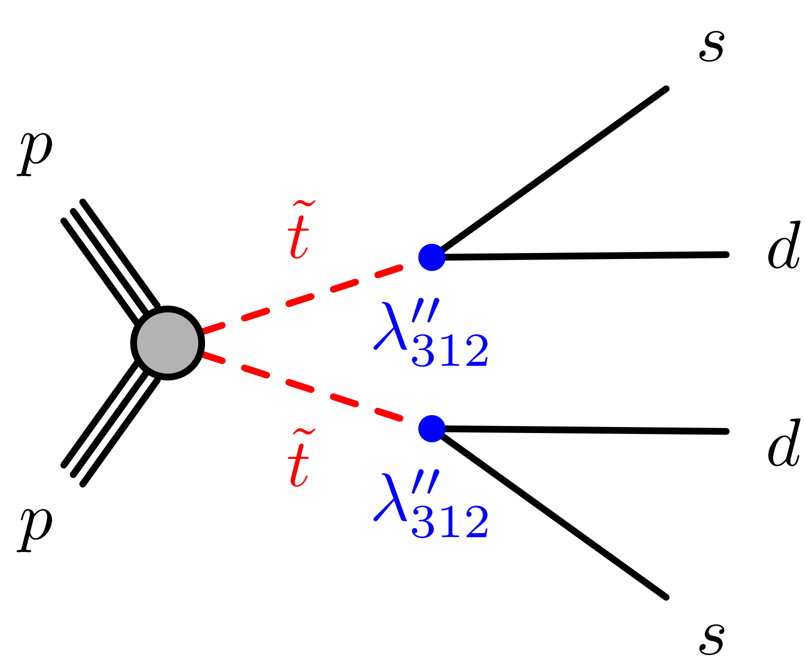
Need: $|r| \ll 1$

otherwise p-values are biased



ATLAS paired dijet resonance search

1710.07171, 13 TeV, 36/fb [see also CMS version 1808.03124]



Counting experiment in bins of

$$m_{\text{avg}} = \frac{1}{2}(m_{\text{dijet 1}} + m_{\text{dijet 2}})$$

Uses ABCD method for background estimation with features

$$A_{\text{mass}} = \frac{1}{m_{\text{avg}}}|m_{\text{dijet 1}} - m_{\text{dijet 2}}|$$

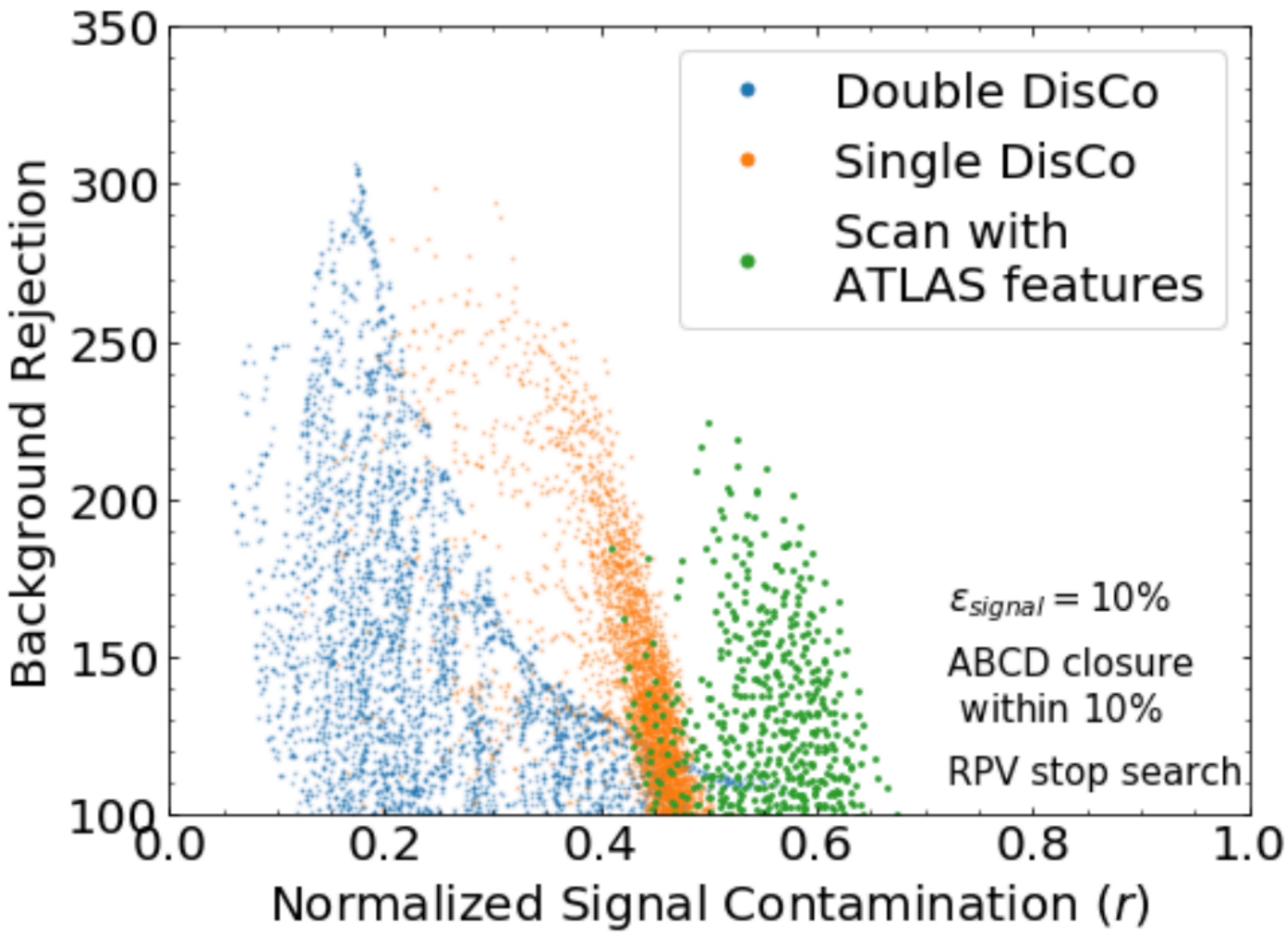
$|\cos\theta^*| =$ angle of squark with beamline in squark-squark rest frame

Simple DNN with input features

$\Delta R_{\text{min}},\ m_{\text{avg}},\ \cos\theta^*,\ A_{\text{mass}},\ z_{12},\ z_{34},\ \Delta R_{12},\ \Delta R_{34},\ m_{12},\ m_{34},\ \Delta\eta,\ \Delta\phi,\ p_{T,12},\ p_{T,34}$

ABCDisCo

Kasieczka, Nachman, Schwartz & DS 2007.14400



Can significantly reduce signal contamination and boost background rejection!

Conclusions

Decorrelating NNs against auxiliary features is a fascinating topic with many important applications to the LHC and beyond.

Adding a simple regularizer term based on Distance Correlation to the loss function achieves state-of-the-art performance for W-tagging.

DisCo can also be used to effectively automate feature construction for the ABCD method, simultaneously boosting background rejection and reducing signal contamination.

Stay tuned, more applications to come!

- Applications to real-world issues such as AI bias and algorithmic fairness?

Thanks for your attention!

Backup

Previous approaches

- Data "planing" [old idea, named and studied in 1709.10106, 1908.08959]

$$w_{i,C}\big|_{x_i \text{ in bin } j} = A_C \frac{1}{n_j}$$

  - reweight training data to flatten mass distribution
  - very simple and potentially powerful, but cannot guarantee full statistical independence

- Designed decorrelated taggers - DDT [1603.00027]

$$\tau_{21}^{DDT} = \tau_{21} - a \times \log \frac{m^2}{p_T \mu}$$

  - Removes most of the dependence of τ_{21} on mass

Previous approaches

- Nonlinear subtraction via kNN regression

$$D_2^{k\text{-NN}} = D_2 - D_2^{(16\%)}$$

  - Use kNN regression to remove dependence on mass and pT for a single cut efficiency

- Convolved SubStructure - CSS [1710.06859]

$$\frac{1}{\sigma}\frac{d\sigma}{dx} \mapsto \frac{1}{\sigma}\frac{d\sigma}{dx_{\text{CSS}}} = \frac{1}{\sigma}\frac{d\sigma}{dx} \otimes F_{\text{CSS}}(x|\alpha, \Omega_D),$$

  - Generalization of DDT
  - Convolve variable with shape function

- uBoost [1305.7248]

  - Modified BDT, adaptive boosting for classification performance and uniformity at fixed selection efficiency