DisCo Fever: Robust Networks Through Distance Correlation

Gregor Kasieczka, David Shih
(gregor.kasieczka@uni-hamburg.de)

ML4Jets 2020 - NYU

2020-01-16

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

Bundesministerium
für Bildung
und Forschung

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

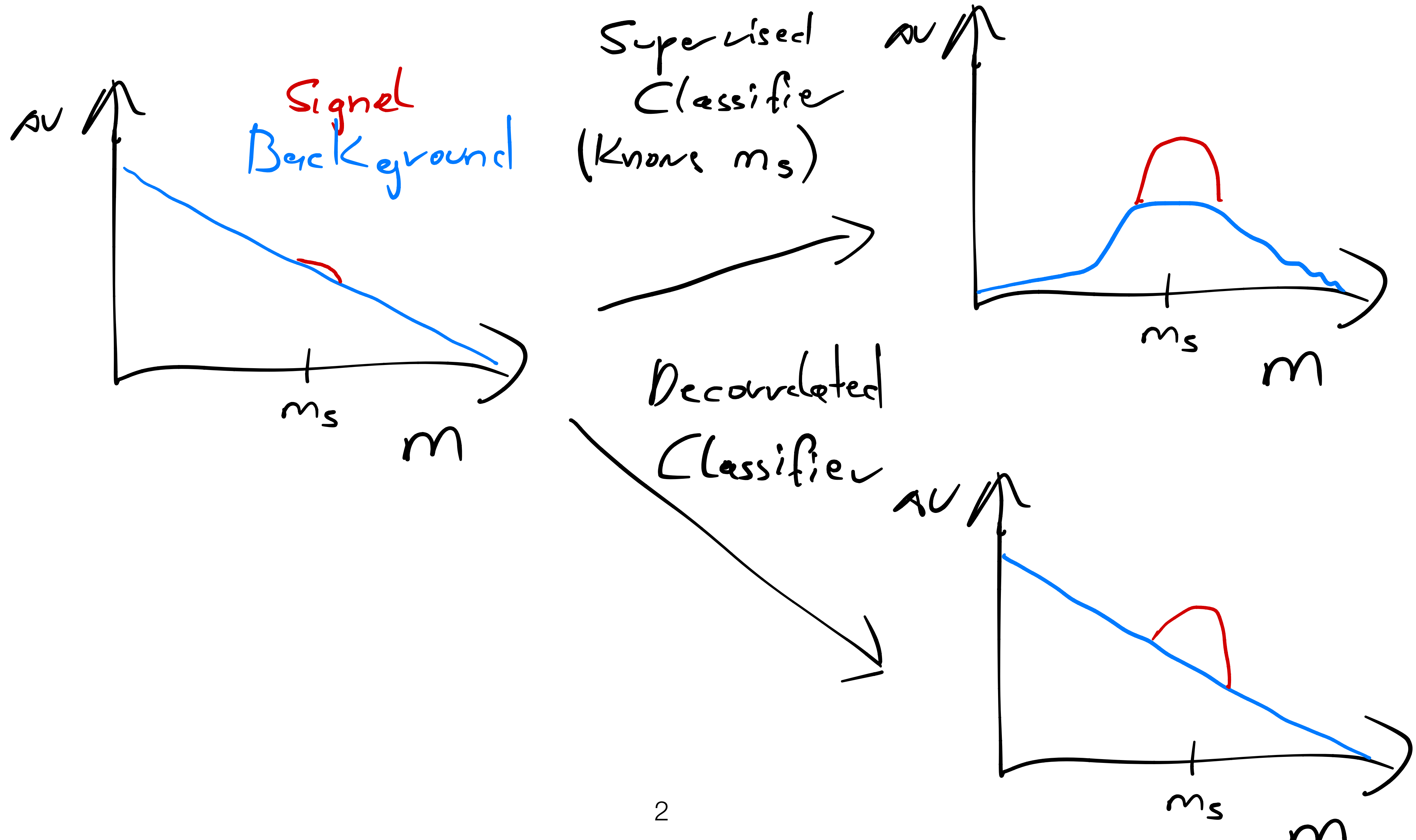
Emmy Noether-Programm
Deutsche Forschungsgemeinschaft
DFG

BERKELEY LAB

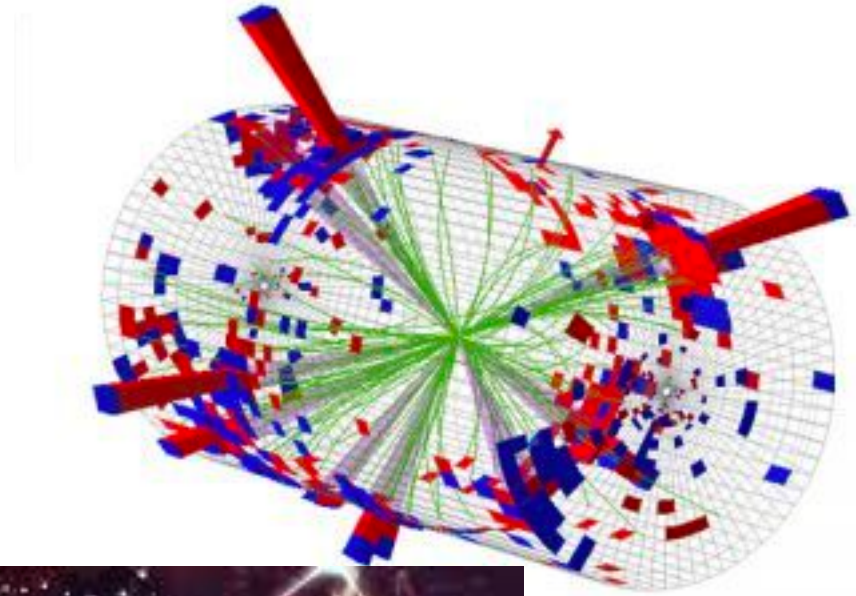
THE UNIVERSITY OF CALIFORNIA BERKELEY 1868

Motivation

- Reduce impact of other variables on analysis result
- Either remove correlation of classifier output with a systematic uncertainty or another variable



Overview



- Brief review of decorrelation Tools
- Recasting ATLAS
- Enter Distance Correlation (*DisCo*)
- Results

Simple approaches

- Obscurity:
  - Do not give mass [will be using this as stand-in for any variable we want to decorrelate agains] as input
  - Simple, does not work

- Data *planing*
  (old idea, studied and named in 1709.10106, 1908.08959):
  - Reweight input distributions to be flat

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

  - Can be powerful, but no guarantee - depending on type of correlation
    Good baseline method

Simple approaches contd.

- Designing Decorrelated Taggers - DDT (1603.00027):
    - Linearly transform output to be stable for one working point by subtracting for each bin

$$y' = y - M \cdot (x - O)$$

- Non-linear subtraction using regression

$$y^{\text{k-NN}} = y - y^{(P\ \%)}(x, x')$$

- Modified weighting for uniformity in BDT - uBoost (1305.7248)

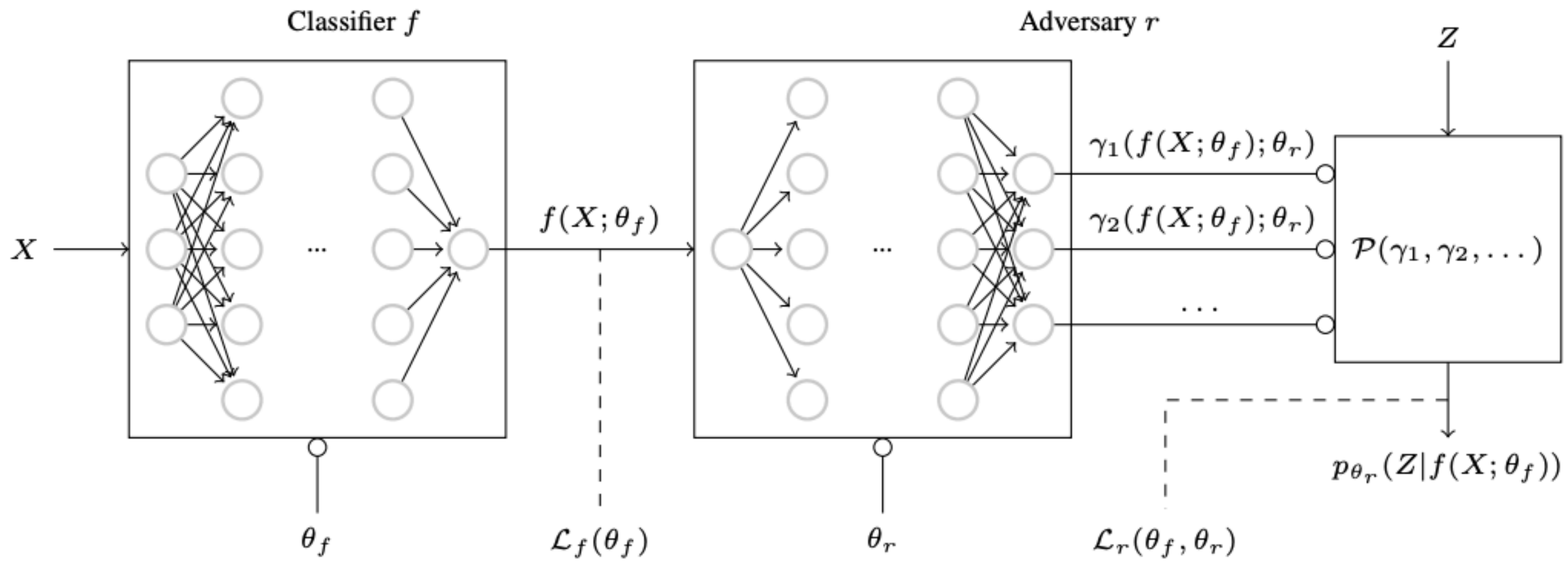
- Convolved substructure - CSS (1710.06859)
  Convolve with variable with shape function
  (not studied here)

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

$$F_{\text{CSS}}(x|\alpha, \Omega_D) = \left(\frac{\alpha}{\Omega_D}\right)^{\alpha} \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{\alpha x}{\Omega_D}}$$

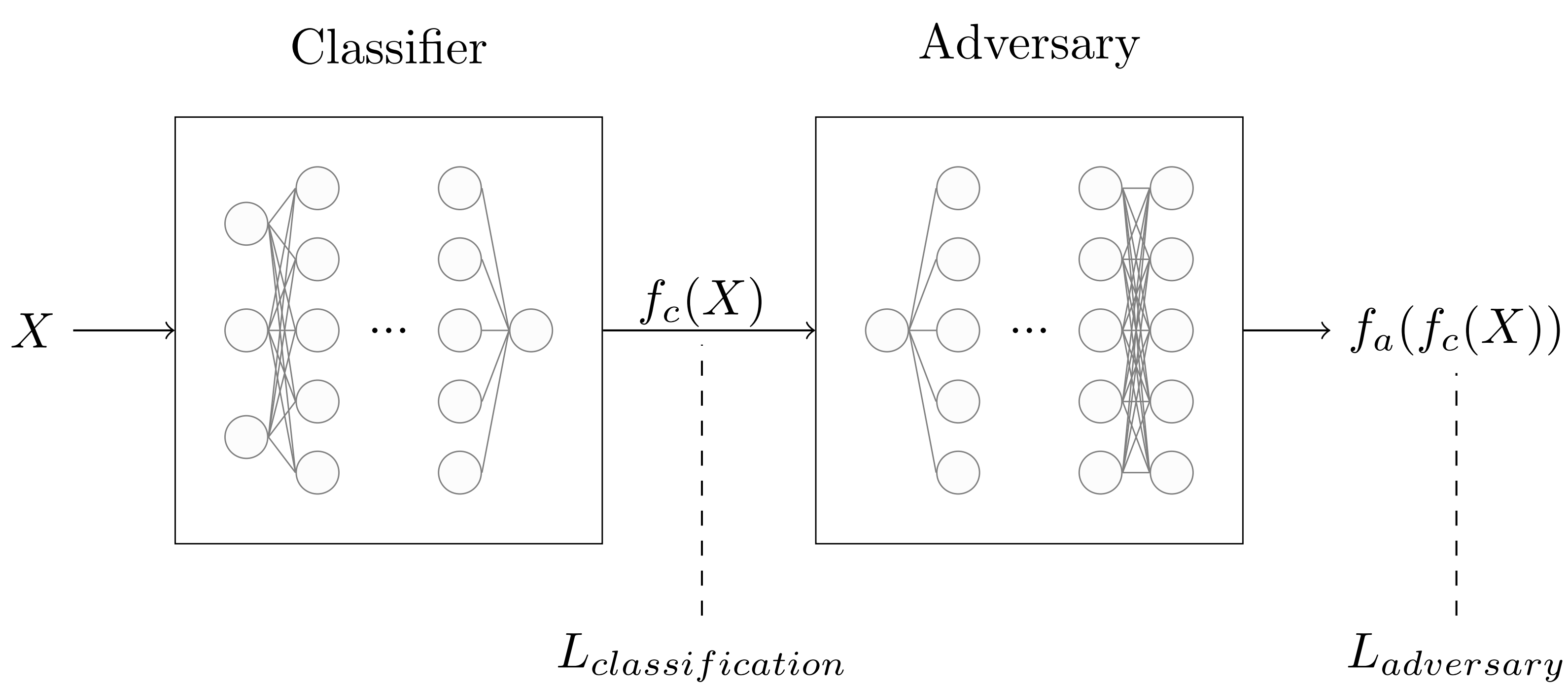
Complex Solutions

Learning to Pivot



1611.01046
Learn a probability distribution via Gaussian mixture model

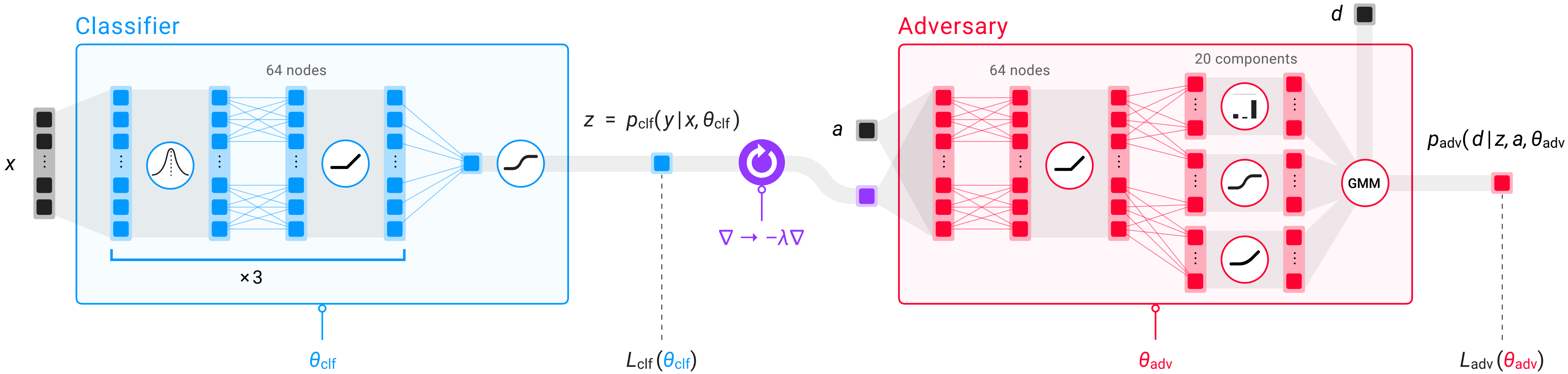
Decorrelated Tagging



1703.03507
Learn to predict the mass and minimise categorical cross entropy

Basic idea: If adversary can infer mass from classifier output, the output is not decorrelated

ATLAS implementation



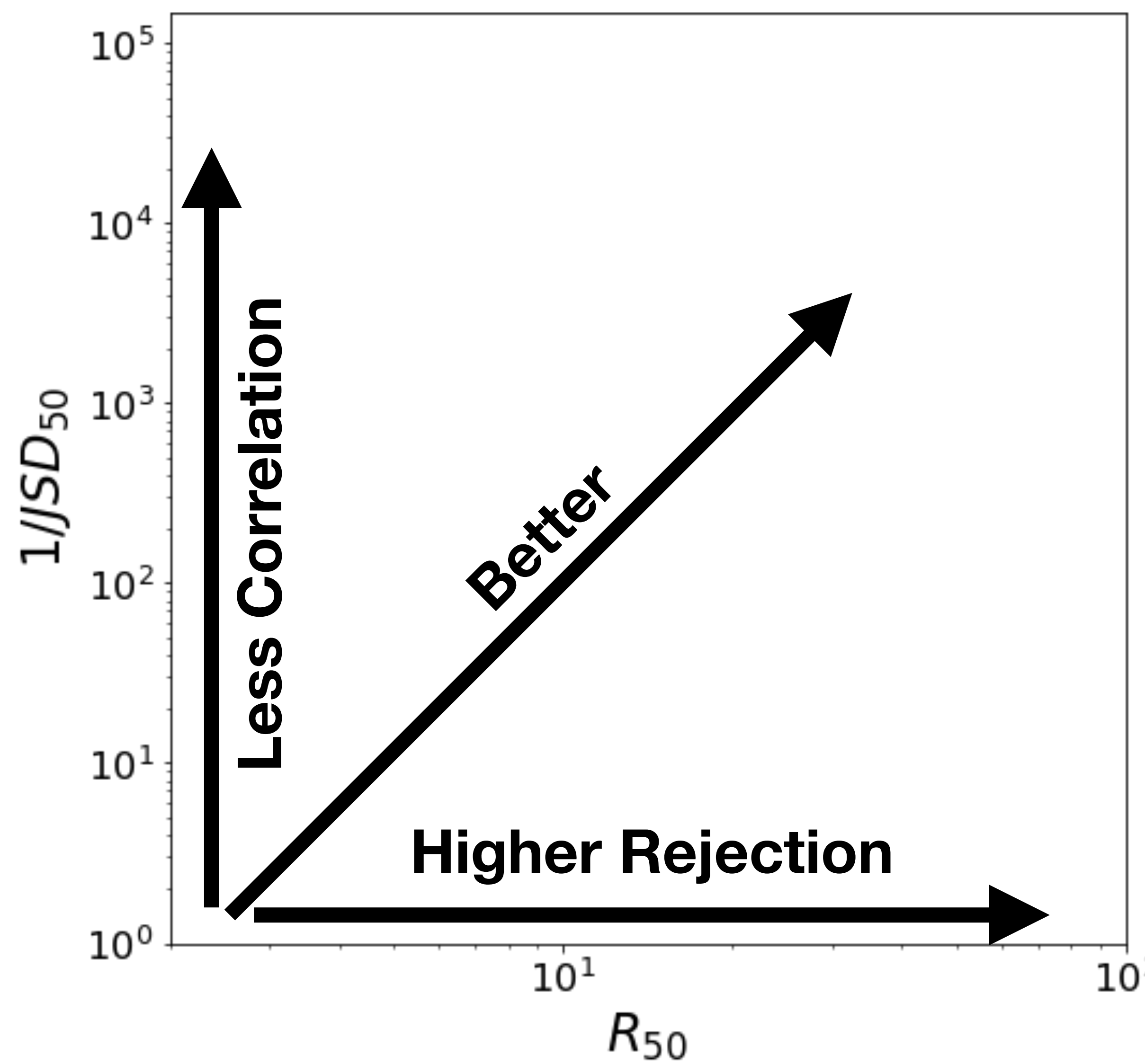
ATL-PHYS-PUB-2018-014

- Similar to learning to pivot, uses gradient reversal
- Classifier: fully connected NN with high-level jet variables

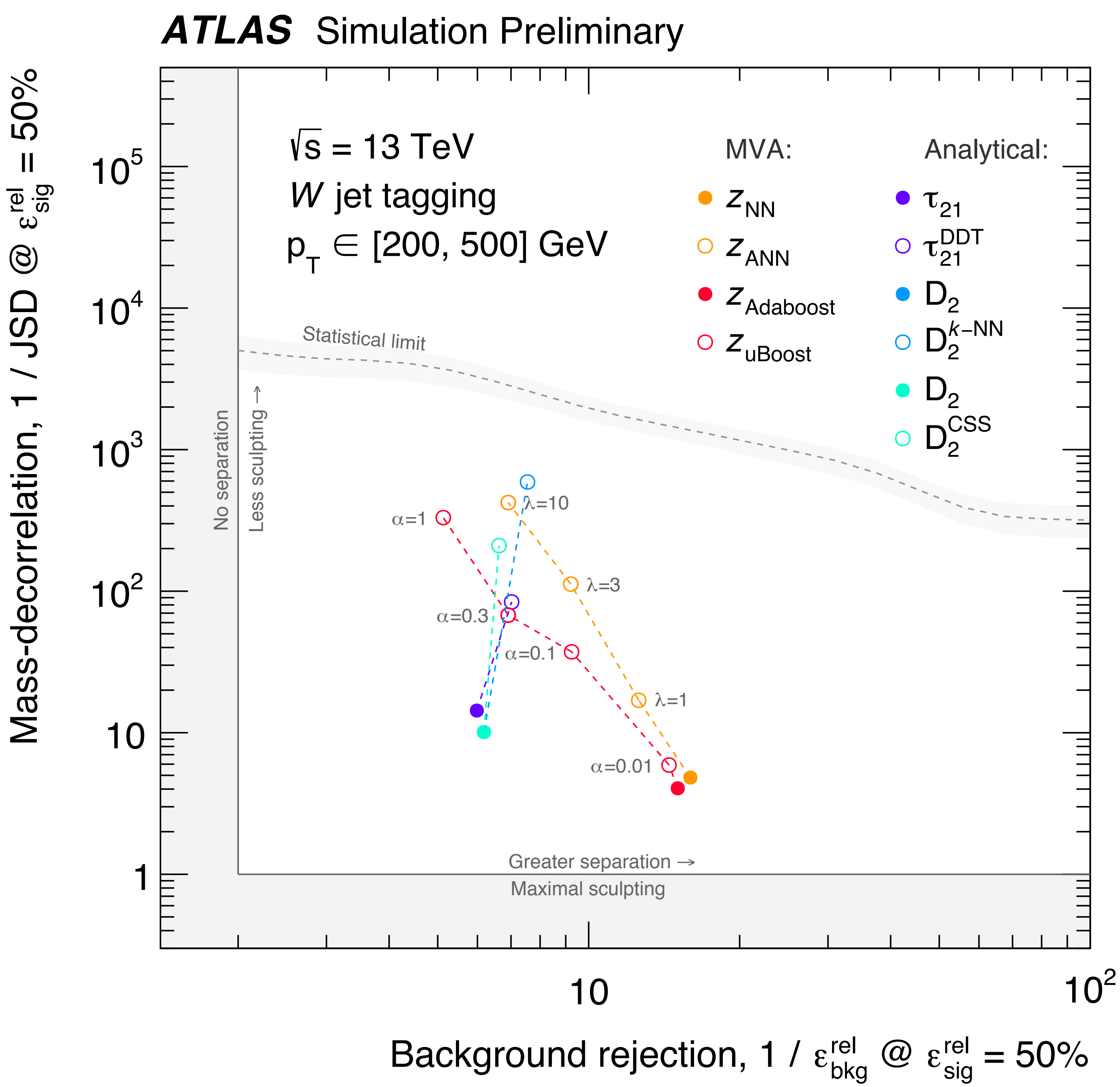
Variable	Type	Reference
$C_2,\ D_2$	Energy correlation ratios	[38]
τ_{21}	N-subjettiness	[41]
R_2^{FW}	Fox–Wolfram moment	[42]
$\mathcal{P}$	Planar flow	[43]
a_3	Angularity	[44]
A	Aplanarity	[45]
$Z_{\mathrm{cut}},\ \sqrt{d_{12}}$	Splitting scales	[46, 47]
$KtDR$	k_t-subjet ΔR	[48]

Performance metrics

- Following ATLAS we look at performance for **50% signs efficiency**
  - **R50:** background rejection
    (1 / background efficiency)
    - *Higher = better rejection*
  - JSD50 is Jensen-Shannon Divergence between:
    background(all) and background(pass cut)
  - **1/JSD50**
    - *Higher = better decorrelation*
- *Expect trade off between these two measures*



Recasting ATLAS

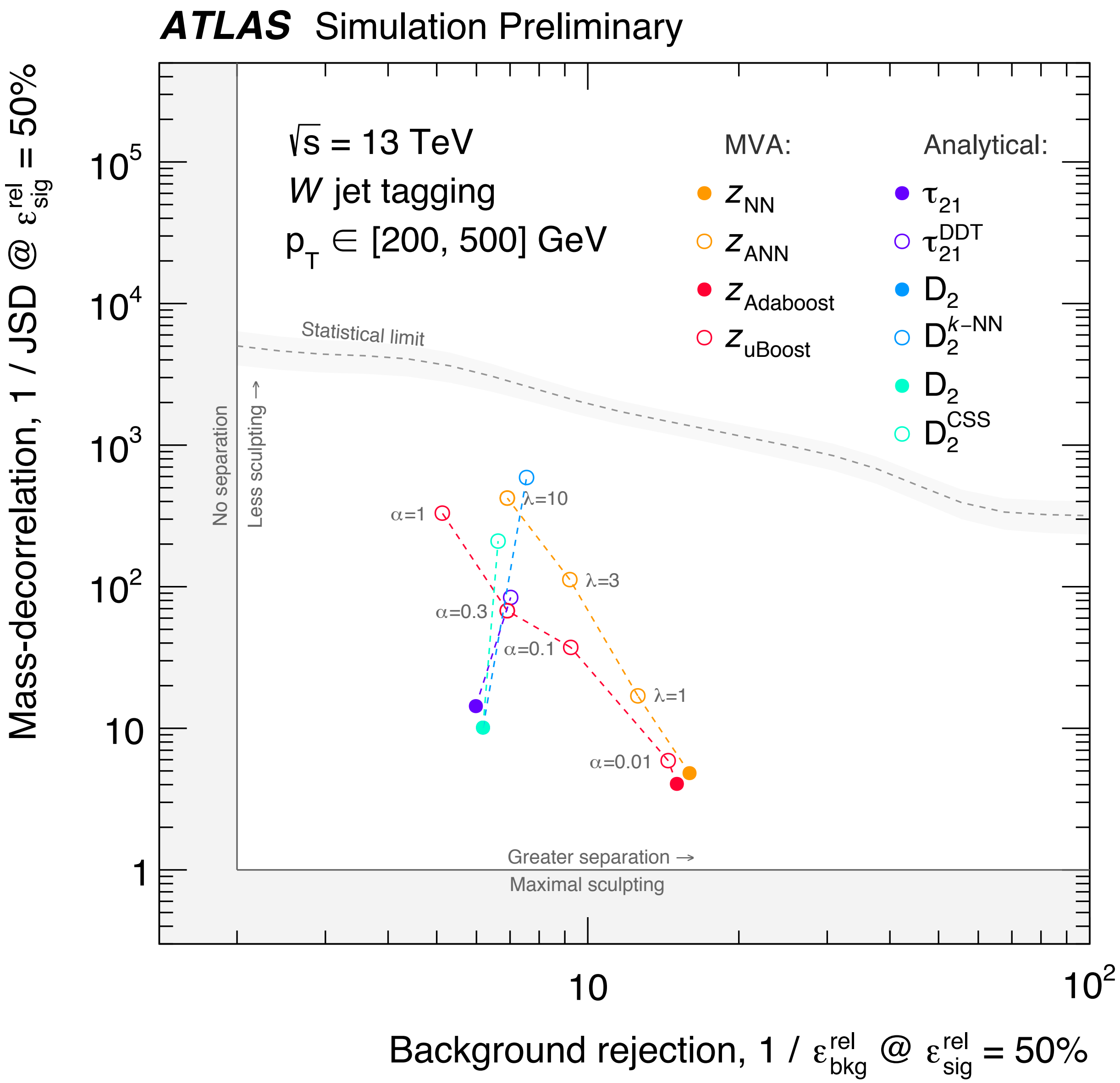


ATL-PHYS-PUB-2018-014

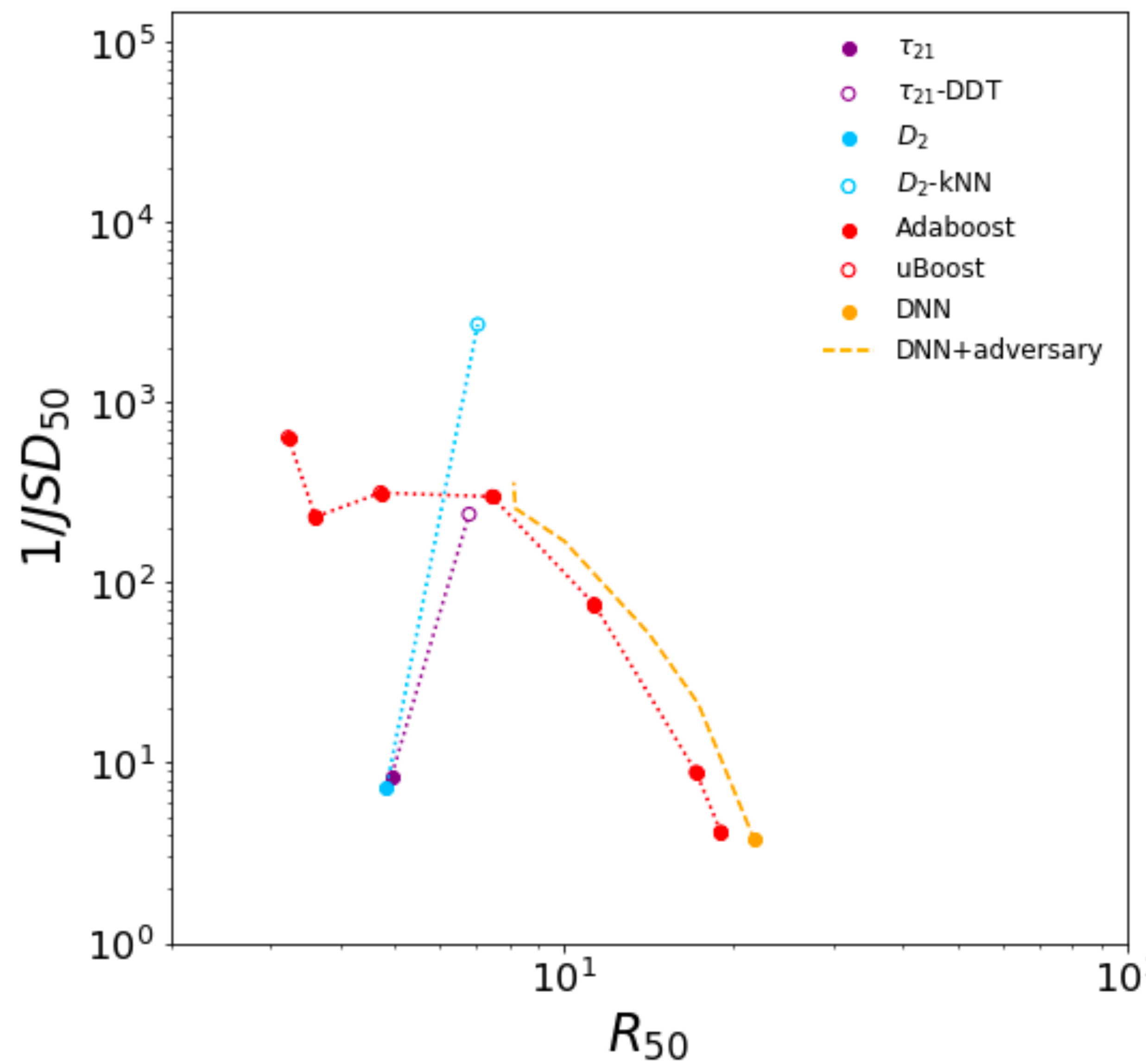
- Hadronic W tagging
  (vs light quark/gluon QCD jets)
- Anti-k_T, R=1.0 jets
  with p_T in [200, 2000] GeV and mass in [50,300] GeV
- Studied analytical and machine learning approaches
- Best performance-correlation trade-off: Adversarial NN

Recasting ATLAS

ATL-PHYS-PUB-2018-014



Our version



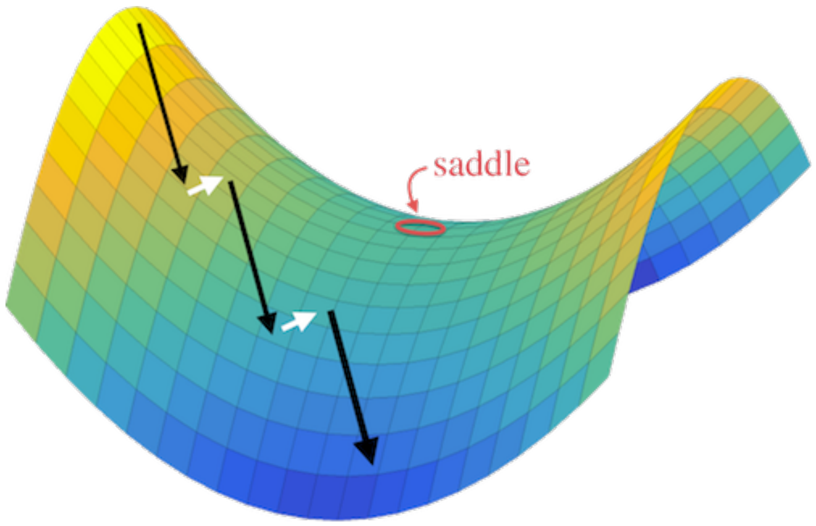
- Pythia + Delphes
- Limit to p_T in [300, 400] GeV

Key features qualitatively and quantitatively well reproduced!

Adversarial Problems

- Adversarial training is inherently unstable (hard to set up and sensitive to hyper parameter changes)
  - Looking for a saddle point

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



  - Many hyper parameters (second network + fine tuning of learning rates)
- Find a regulariser term that fulfils the same goal but allows simple training to convergence

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

Distance Correlation

$$x_{jk} = |X_j - X_k|$$

Distances of all examples in batch for classifier output

$$y_{jk} = |Y_j - Y_k|$$

... for variable to decorrelate

$$\hat{x}_{jk} = x_{jk} - \overline{x}_{j\cdot} - \overline{x}_{\cdot k} + \overline{x}_{\cdot\cdot}$$

$$\hat{y}_{jk} = y_{jk} - \overline{y}_{j\cdot} - \overline{y}_{\cdot k} + \overline{y}_{\cdot\cdot}$$

Center distributions

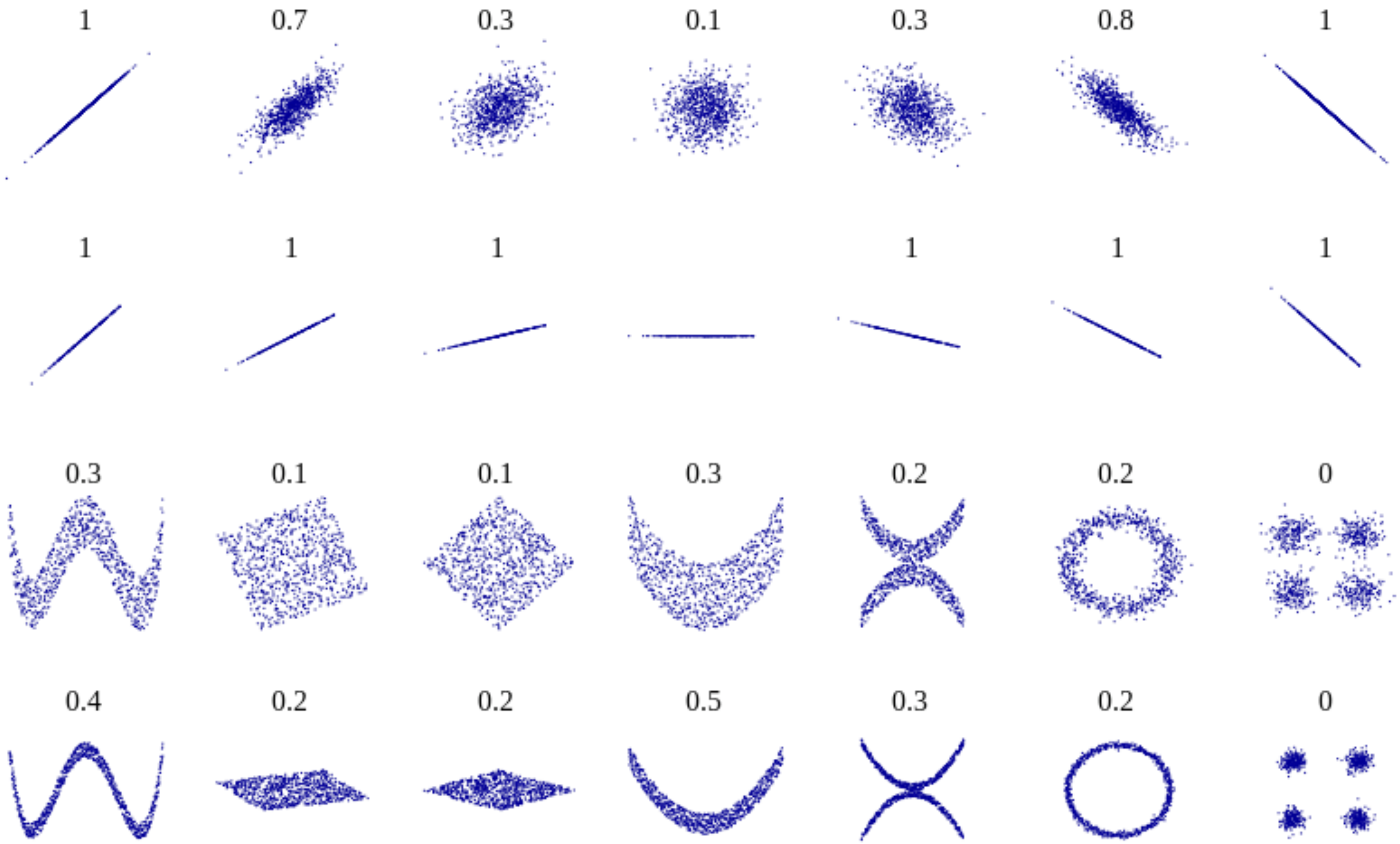
$$\mathrm{dCov}^2 = \frac{1}{n} \sum_j \sum_k \hat{x}_{jk} \hat{y}_{jk}$$

And calculate average product per batch

Some nice properties:

- Zero iff X,Y are independent; positive otherwise!
- Computationally tractable!
- Doesn't require binning!

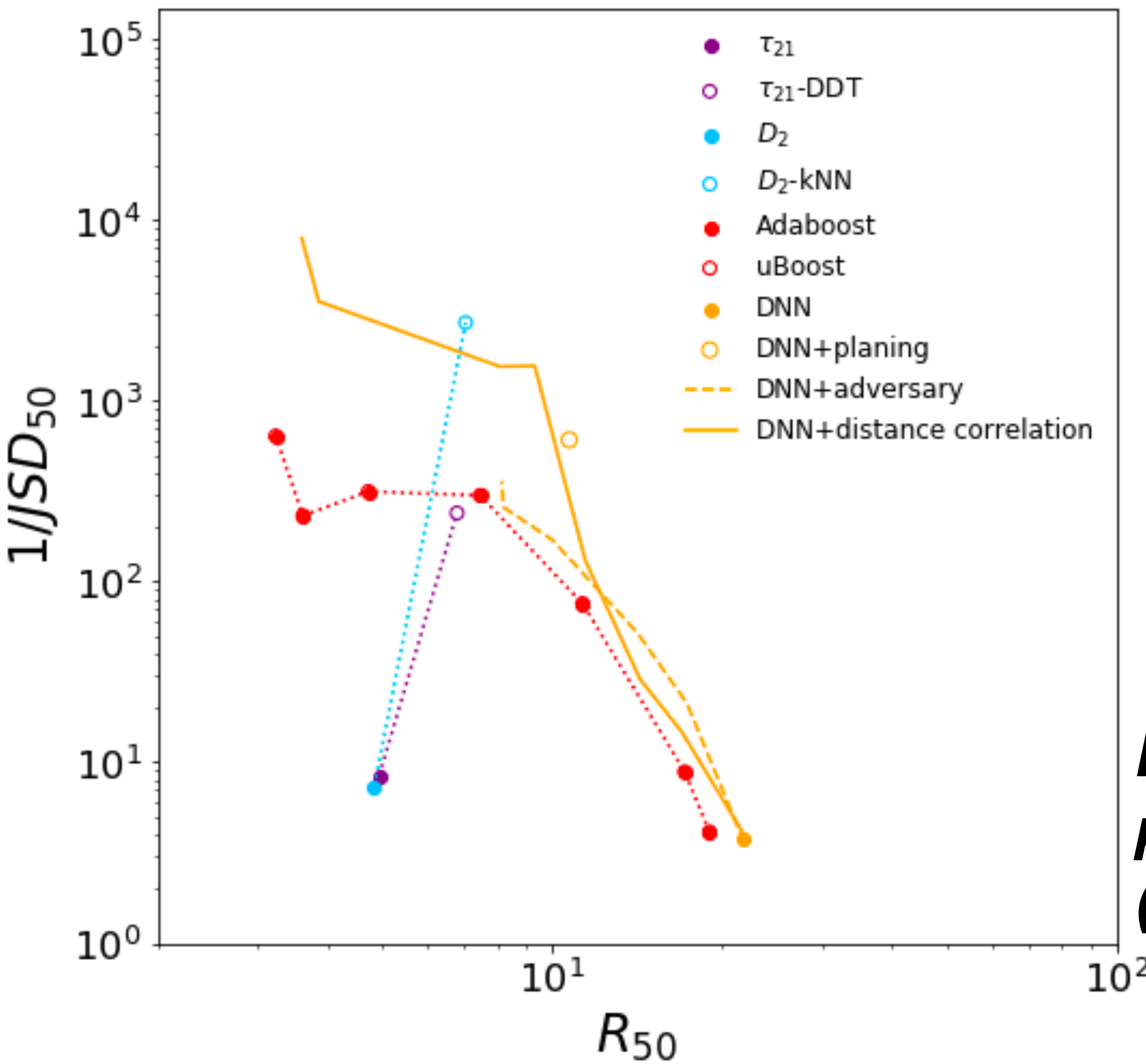
0803.4101, 1010.0297



(Wikipedia)

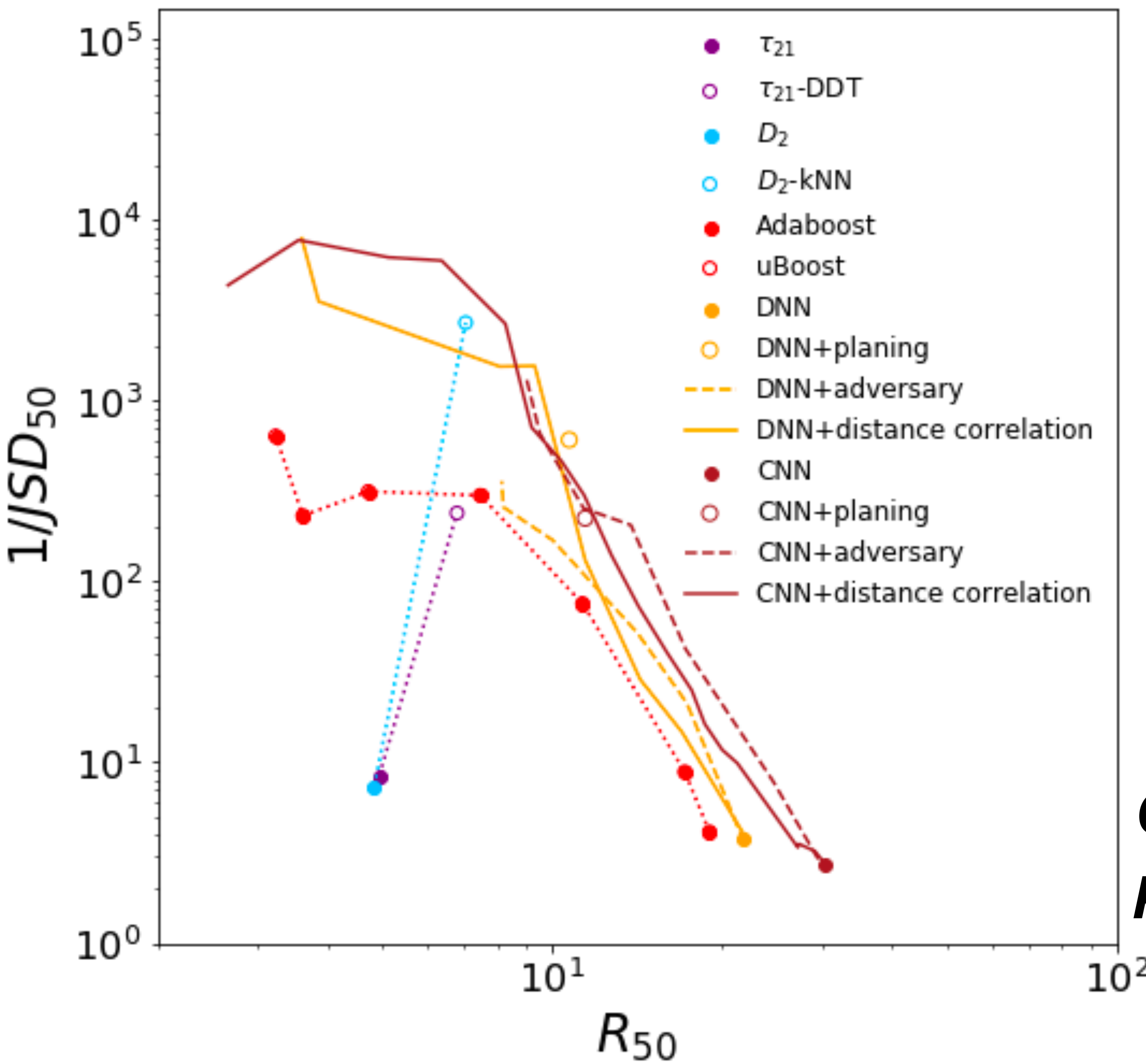
Results

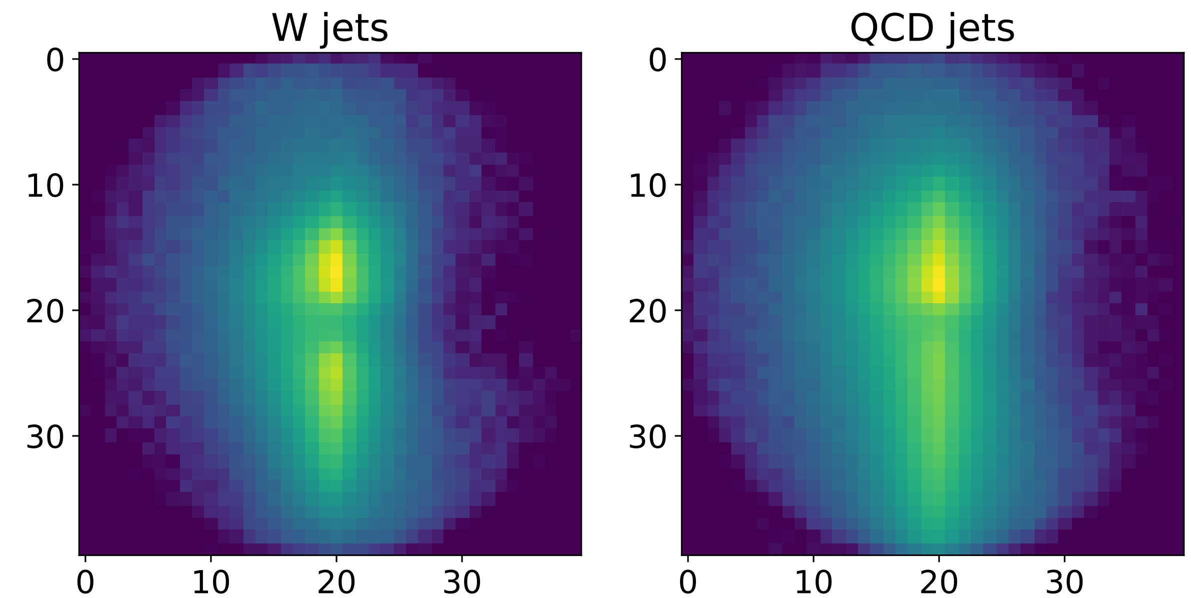
$$L = L_{classifier}(\vec{y}, \vec{y}_{true}) + \lambda \, \mathrm{dCorr}^2(\vec{m}, \vec{y})$$



DisCo achieves state-of-the-art performance
(with much simpler training)

Results

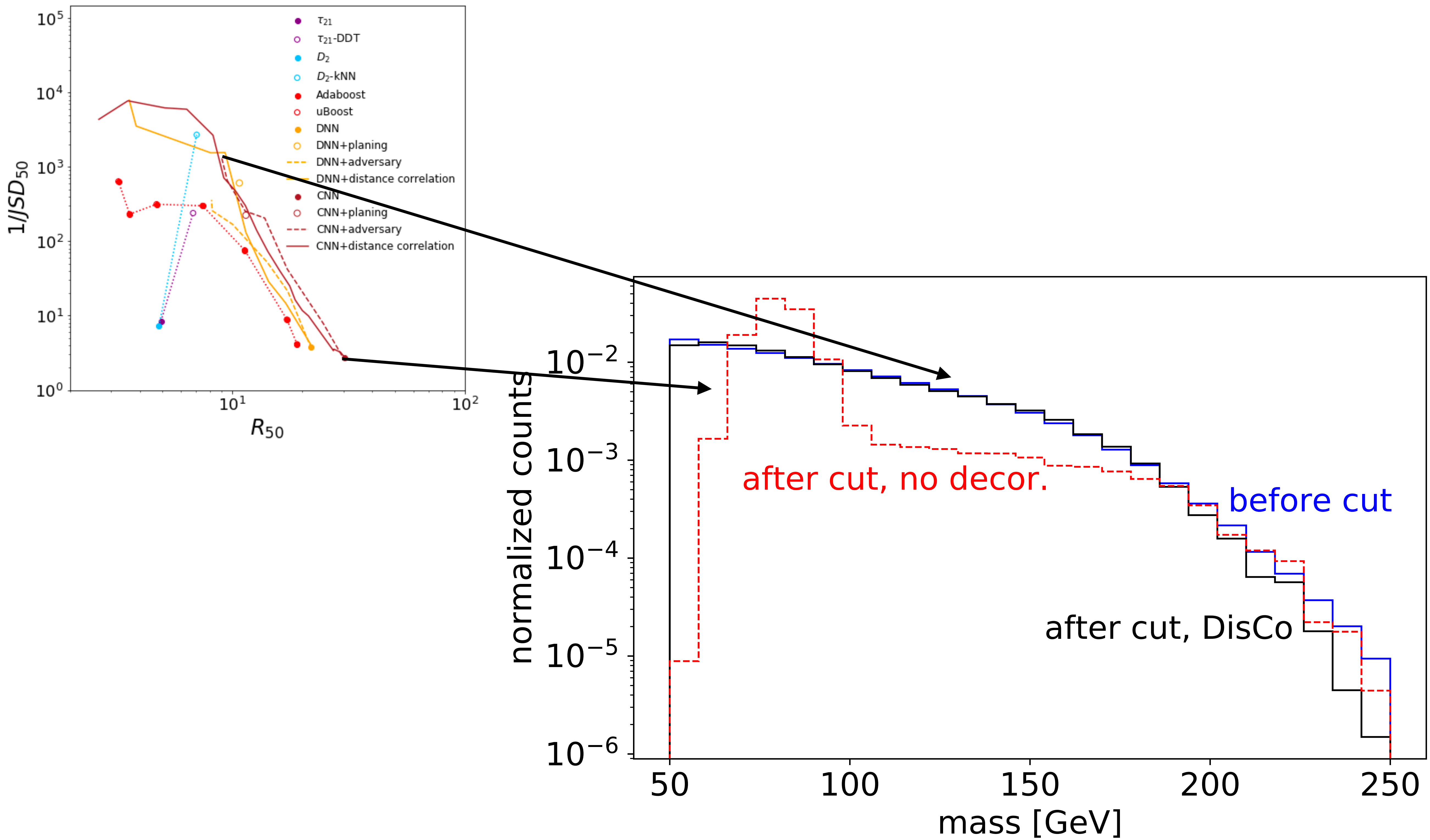




Overlay of 100k examples

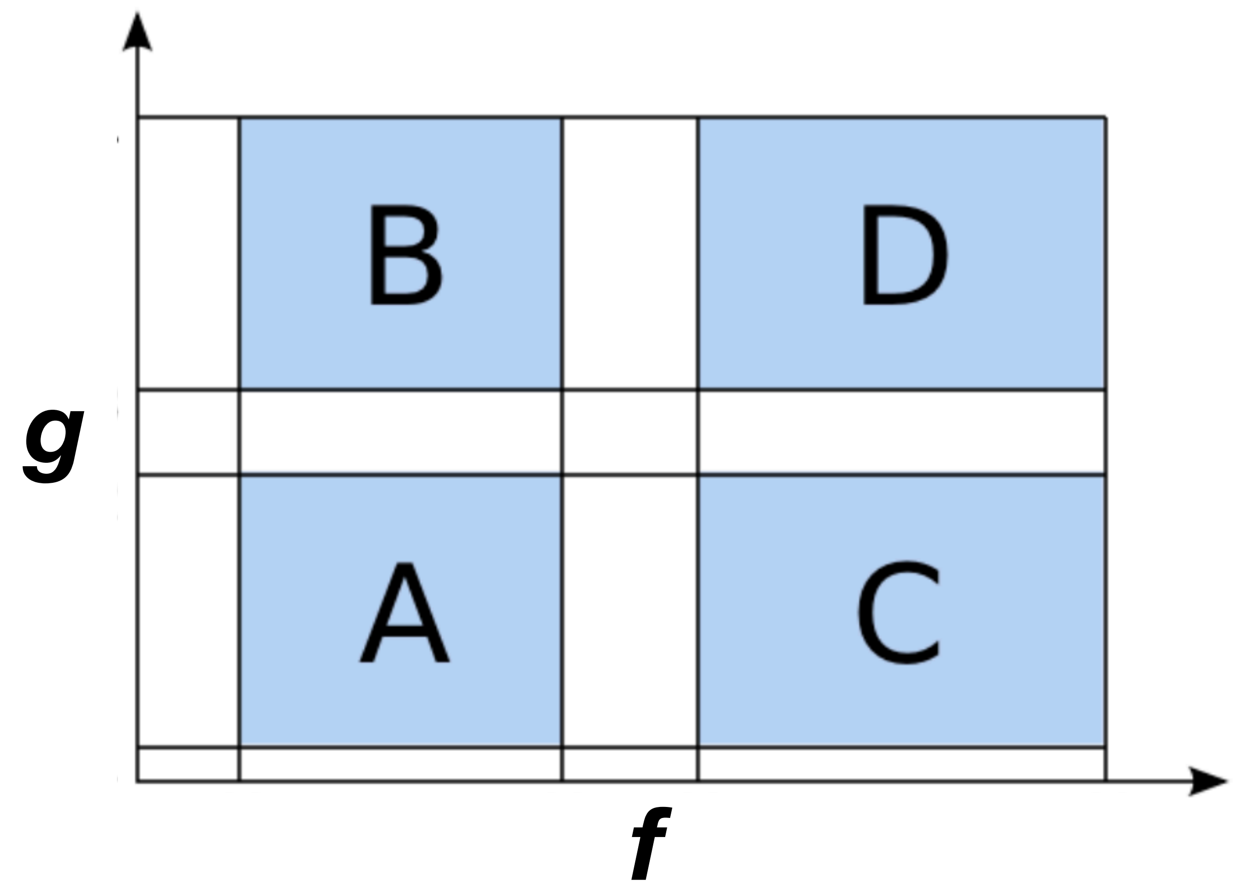
Can also decorrelate more powerful CNN on jet images

Mass shapes


τ_{21}
τ_{21}-DDT
D_2
D_2-kNN
Adaboost
uBoost
DNN
DNN+planing
DNN+adversary
DNN+distance correlation
CNN
CNN+planing
CNN+adversary
CNN+distance correlation
$1/JSD_{50}$
R_{50}
after cut, no decor.
before cut
after cut, DisCo
normalized counts
mass [GeV]

What's next?

- Can we find an optimal pair of variables for ABCD background estimation?
- Goal:
  - Two variables (f,g) with maximal signal/background discrimination and no correlation



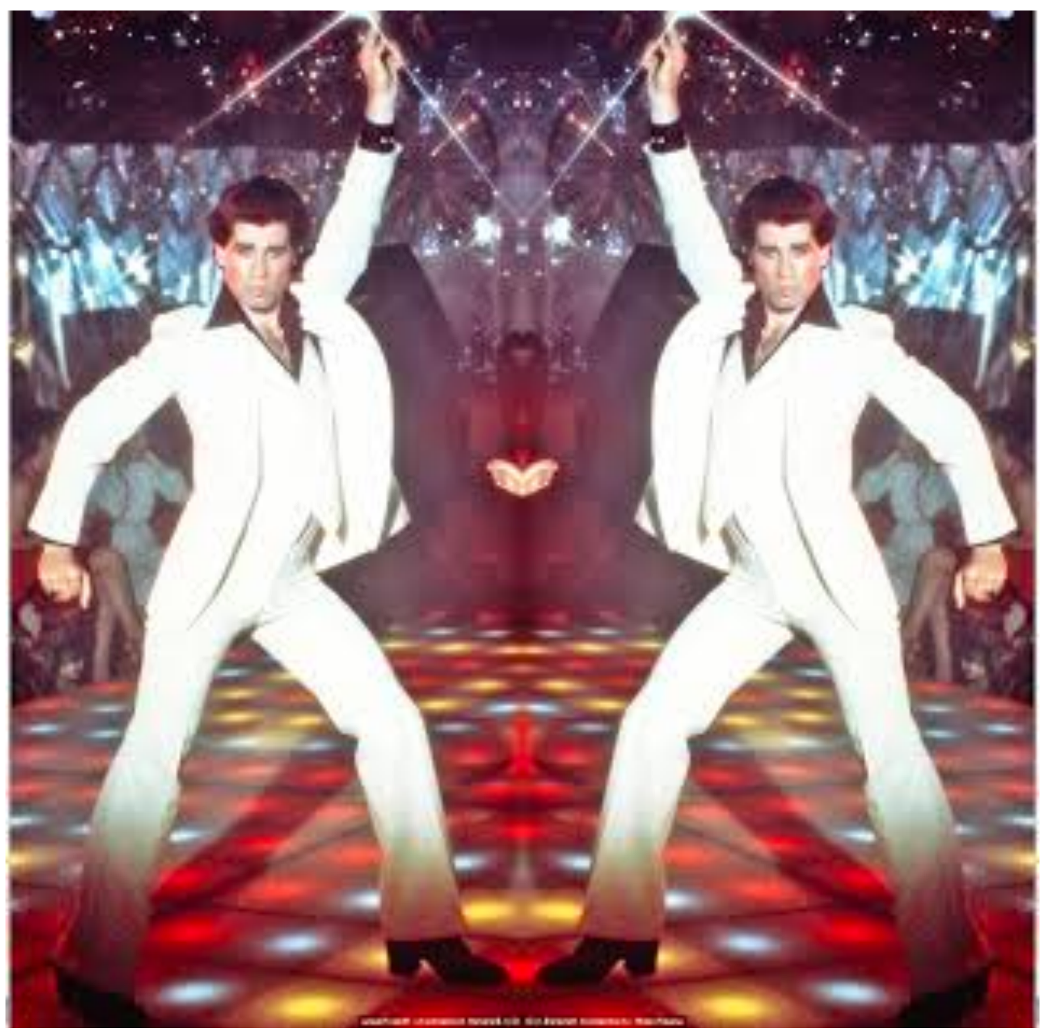
Kasieczka (2009)

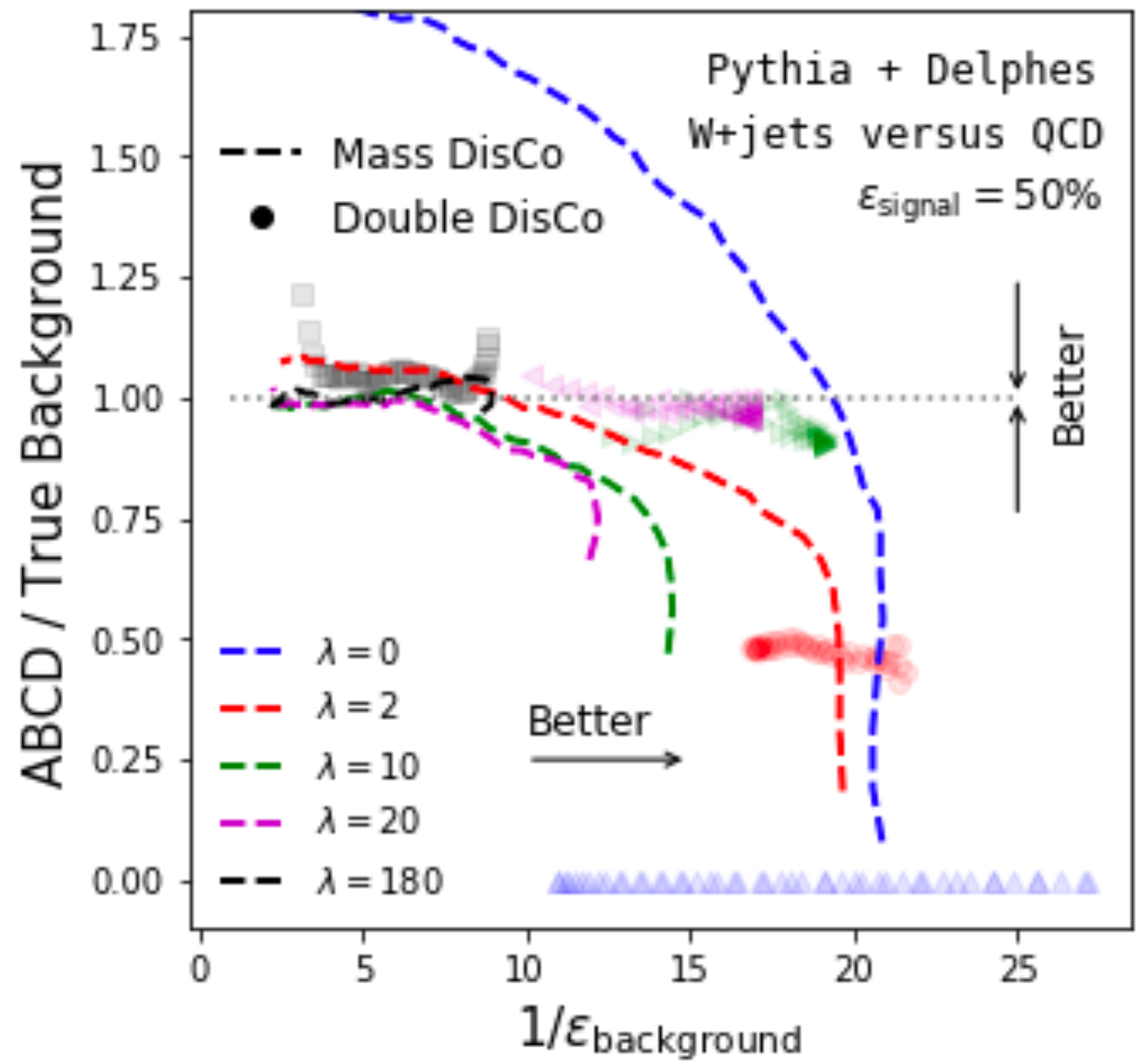
Work in progress with Ben Nachmann, Matt Schwartz & David Shih

Double Disco

$$\text{Loss} = (\text{Loss for } f) + (\text{Loss for } g) + \lambda(\text{Loss term to make } f \text{ and } g \text{ independent})$$

Usual Cross Entropy **DisCo(f,g)**





Work in progress with Ben Nachmann, Matt Schwartz & David Shih

x2 Improvement over mass

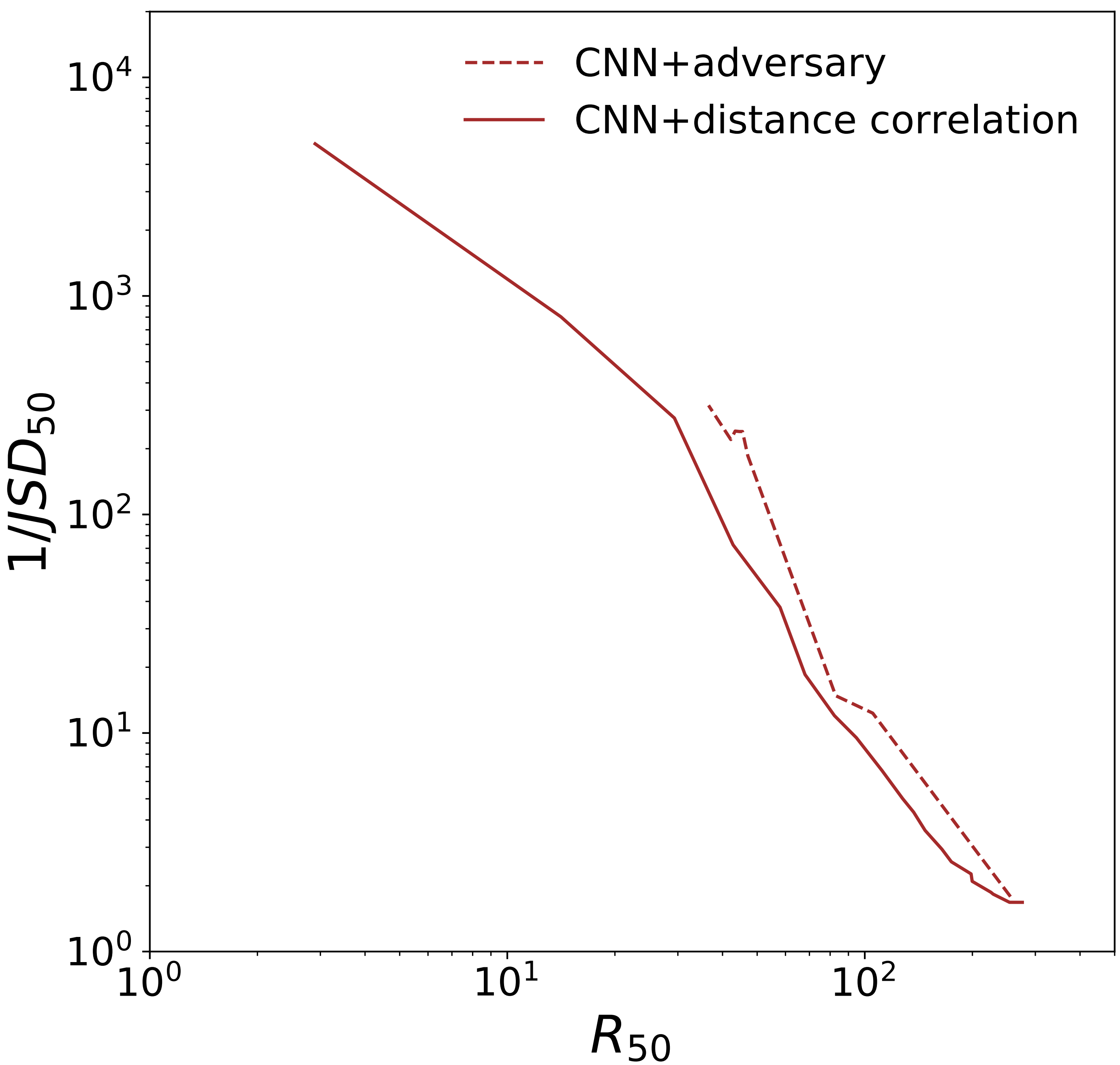
Conclusions

- Decorrelation of classifiers important for many applications
- Simple regulariser term based on distance correlation (DisCo) achieves state of the art performance for W tagging
    - Also decorrelates stronger CNN tagger
- Paper out 2001.05310
  Code here: https://github.com/gkasieczka/DisCo
- DisCo's not dead: more DisCo to come

Thank you!

Bonus Material

Top Tagging



Top images based on top tagging reference dataset