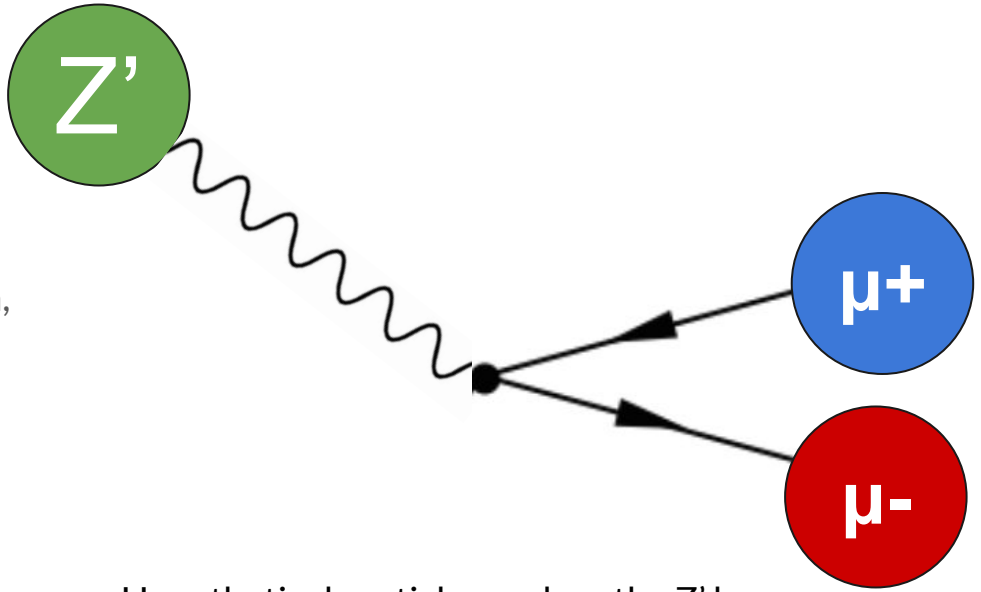
Learning to Identify Muons in Data

Ed Witkowski, Daniel Whiteson
University of California, Irvine

In collaboration with Ben Nachman
LBNL

Introduction

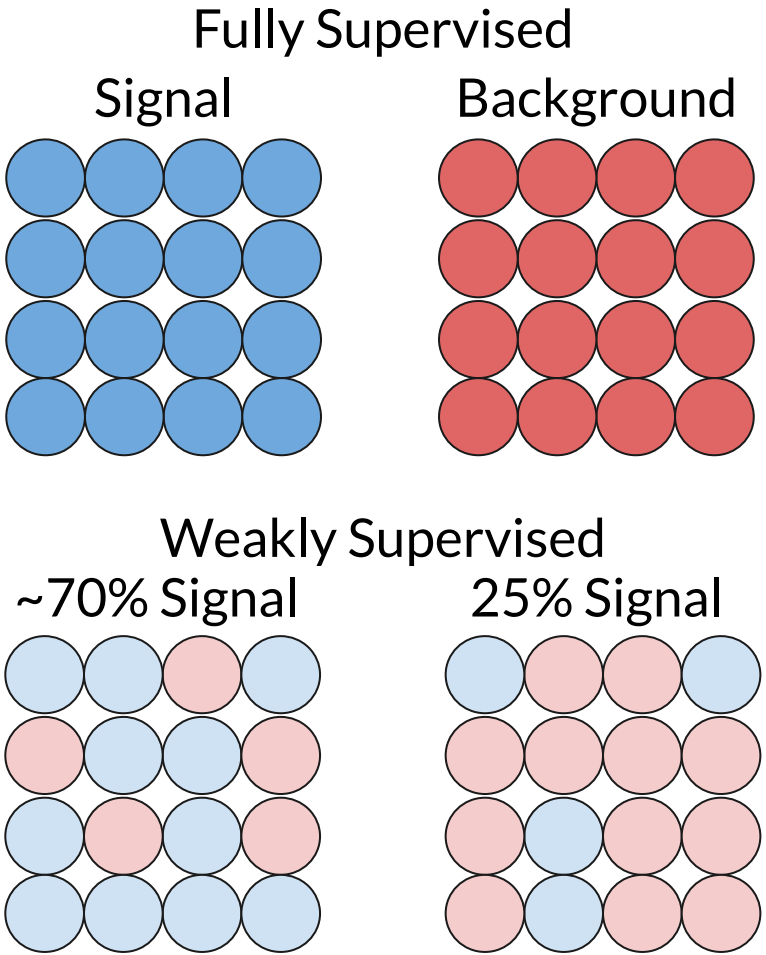
- Identifying prompt muons from heavy boson decay (Z, W, etc.) is important for the discovery of new physics
- Low level particle data is reduced to a scalar, isolation[1-3] - information loss[4]
- Use neural networks to learn, in real data, to identify prompt muons from non-prompt background
- Identify set of interpretable high level observables which yield similar NN performance to low level data



Hypothetical particles such as the Z’ boson might be identified through muon decay products

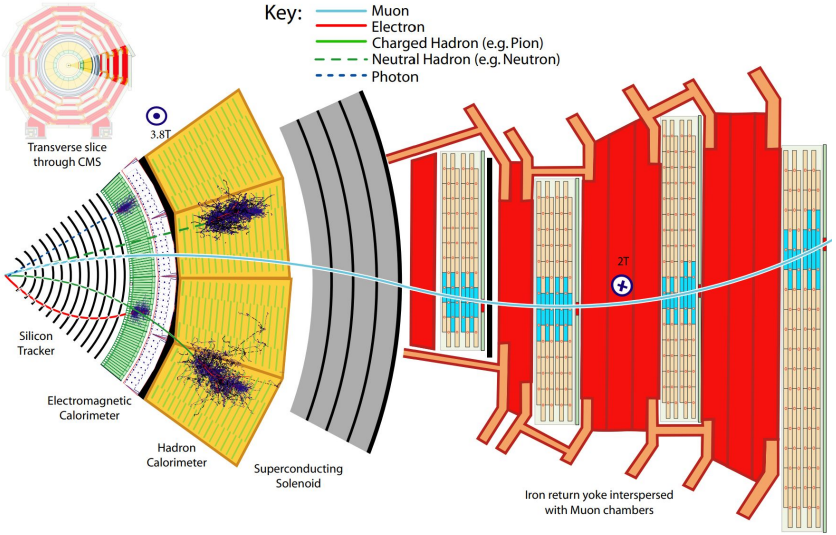
Introduction

- Unlike simulation, real data is unlabeled
- Fully supervised NN training techniques used on simulation won't directly work here
- Can determine overall sample composition -> weakly supervised learning



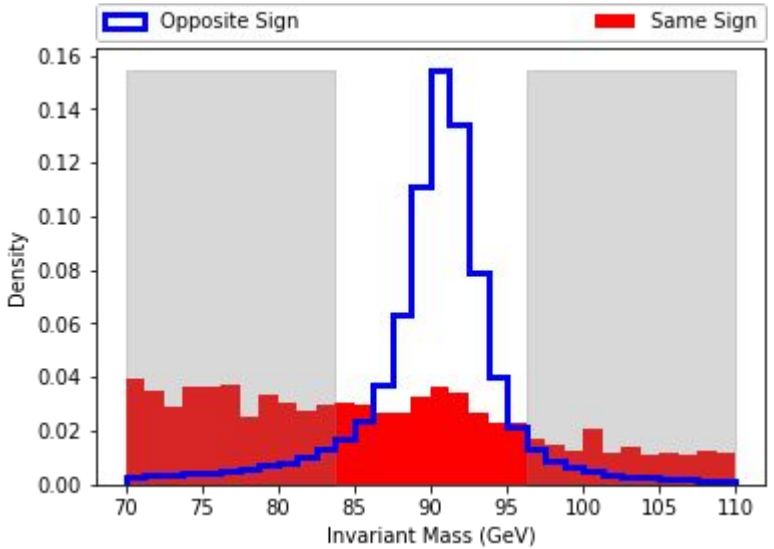
Dataset - Reconstruction

- Our dataset was obtained from CMS Open Data, collected during 2012 run[5]
- Data reconstructed with Particle Flow algorithm - calo + track information[6]
- Results in particle data objects with associated pT, eta, phi, charge
- Objects are categorized as:
  - Muons and electrons
  - Charged and neutral hadrons
  - Photons
  - Pileup



Dataset - Selection

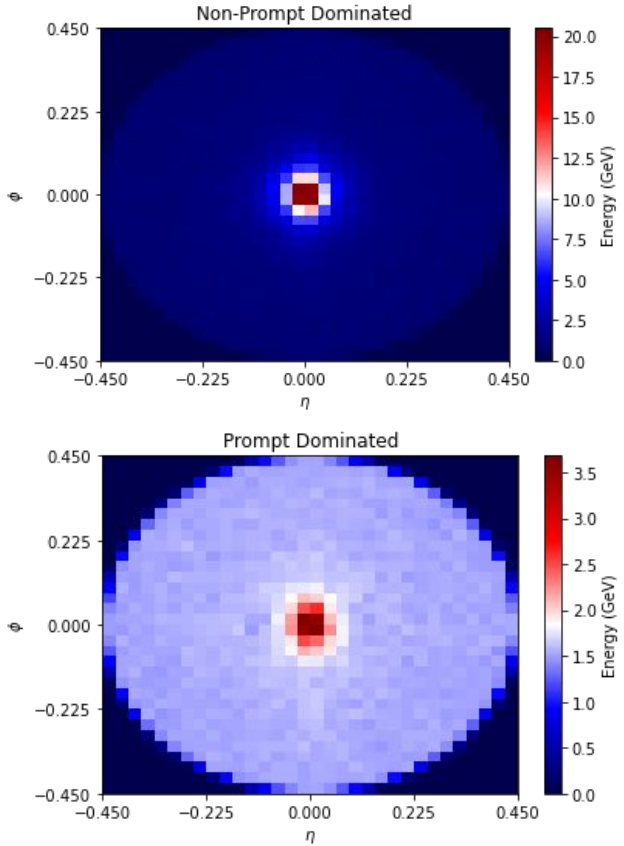
- Selection criteria:
  - Dimuon events
  - Both muon pTs > 25 GeV
  - Both muon |eta|s < 2.1
  - Invariant mass between 70 - 110 GeV
- Separate into 2 samples:
  - Prompt muon dominated (907488 events, 95.6% prompt)
  - Non-prompt muon dominated (171238 events, 6.83% prompt)



Non-Prompt dominated sample contains all similarly charged muon events + all from shaded regions, remaining events make up prompt dominated sample

Approach

- Compare performance of neural networks on:
  - Low level particle flow data
  - Isolation
- Low level performance provides benchmark
- How does isolation compare?
- Can we match low level performance with more interpretable set of high level observables?

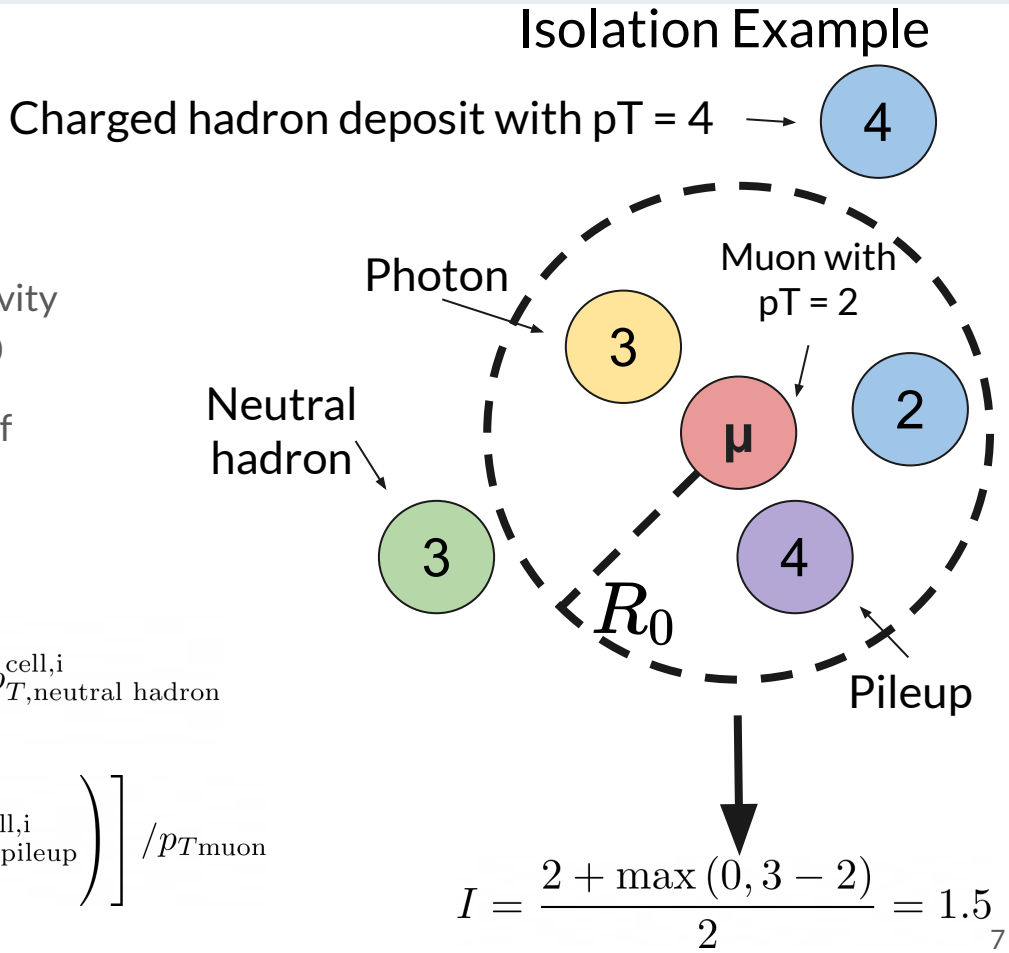


Visualization of low level data - pT deposits binned in (eta,phi) + averaged across samples

Isolation

- Isolation quantifies the amount of activity around a given muon with a radius R_0
- Other studies indicate using a radius of 0.3, which we include in our set of calculated isos[7]

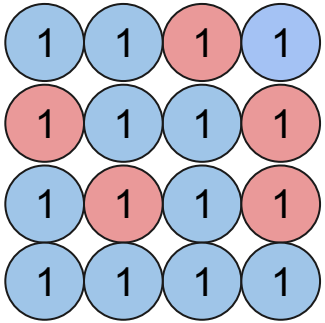
$$I_\mu(R_0) = \left[\sum_{i,R<R_0} p_{T,\text{charged hadron}}^{\text{cell,i}} + \max\left(0, \sum_{i,R<R_0} p_{T,\text{neutral hadron}}^{\text{cell,i}} + \sum_{i,R<R_0} p_{T,\text{photon}}^{\text{cell,i}} - \frac{1}{2} \sum_{i,R<R_0} p_{T,\text{pileup}}^{\text{cell,i}}\right)\right] / p_{T\text{muon}}$$



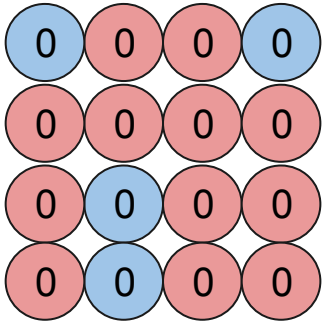
Training

- Optimal classifier for two mixed class samples can be shown to be equivalent to that for underlying classes - Classification Without Labels (CWoLa)[8]
- Separate events into two samples determined to have different class distributions
- Label events according to sample drawn from -> supervised learning

Mixed Sample 1
All labeled 1



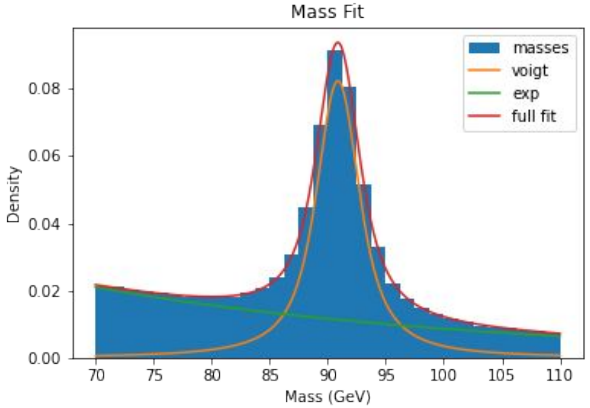
Mixed Sample 2
All labeled 0



Training on pure samples should yield equivalent classifier to training on mixed samples 1 & 2

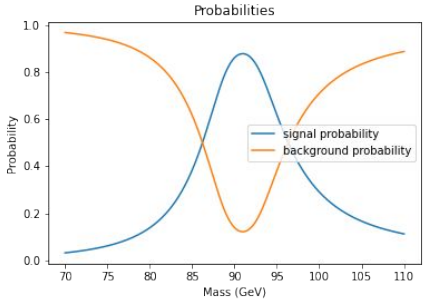
Performance Evaluation

- Want model performance at class level
- Typical AUC calculation would use true class labels
- Instead we use a reweighting method (sPlots)[9]
- Fit the masses, using exponential and voigt distributions as signal / bg components
- These can be used to calculate weight values

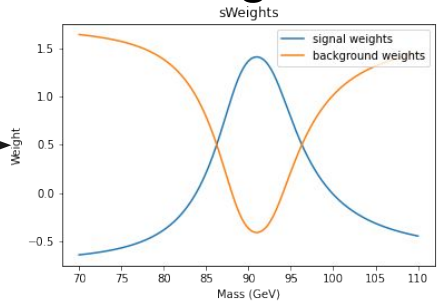


Mass Fit

Probabilities



Weights

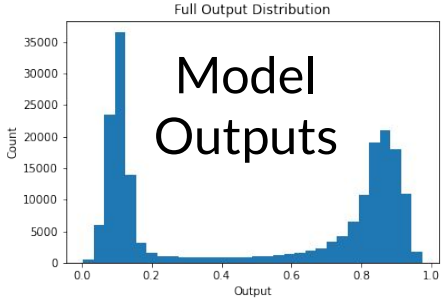


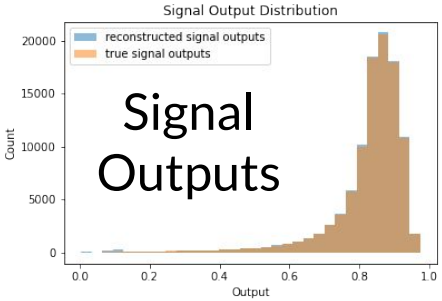
Weights are calculated such that they reduce the contribution of an unwanted component to a given histogram

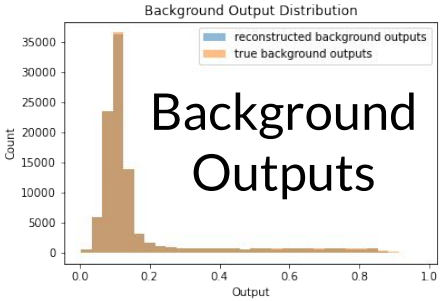
Performance Evaluation

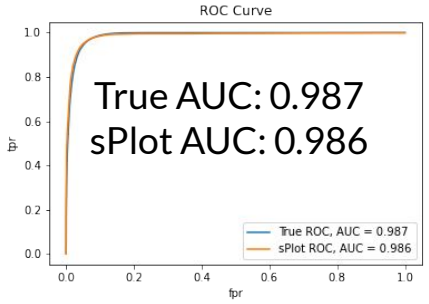
- The weights can be applied to histograms of other quantities
- Reconstructs signal / background components
- Apply to model outputs
- Use separated output distributions to calculate FPR / TPR -> AUC

Model outputs corresponding to above mass distribution





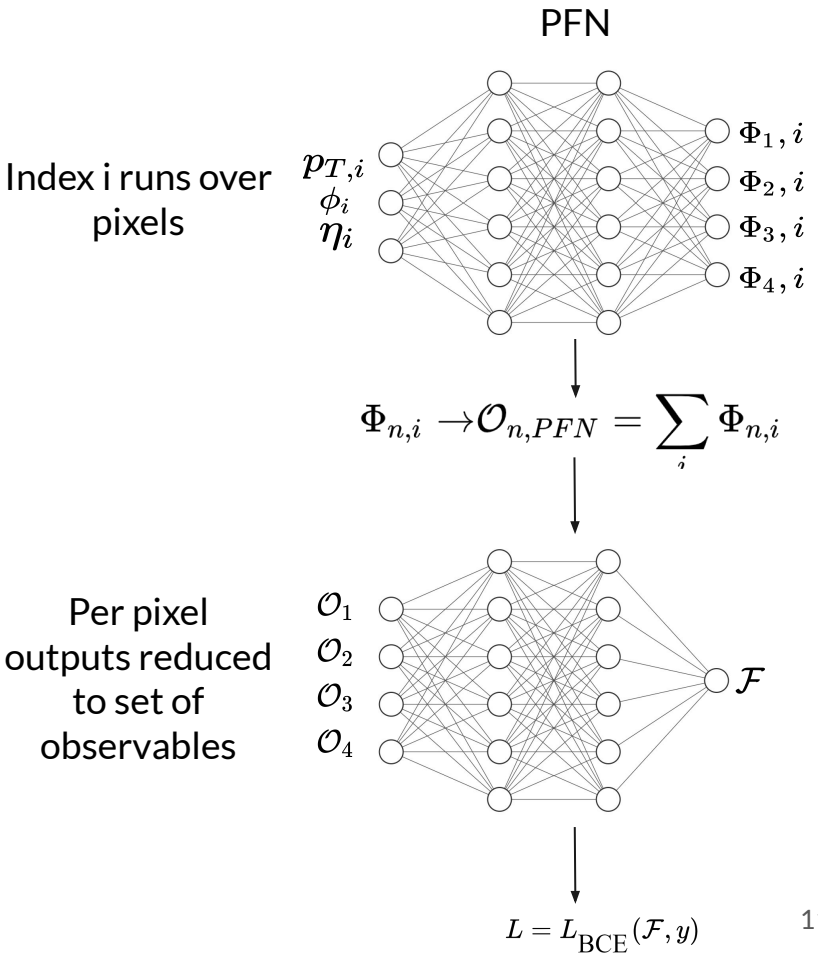




Example ROC curve, very close to truth

Networks

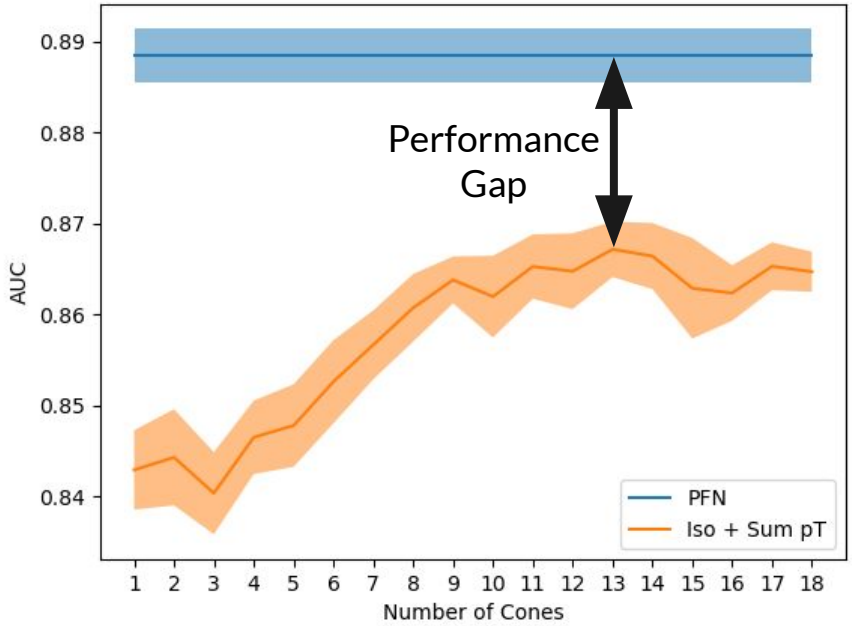
- Low level - Particle Flow Network (PFN)[10]
  - Two fully connected networks
  - Each (pT, eta, phi) in an event is fed individually into first network
  - Outputs are summed over event
  - Fed into second fully connected network for classification
- High level - Fully connected networks
  - Train with sets of isos of different radii + summed event pT



Isolation Performance

- Networks trained with sets of isolations and the summed event pTs as input
- Performance + error bands are calculated over 5 fold stratified cross validation
- Overall as cones are added, performance increases
- Improvements start to drop off at 9 isolations
- Performance never reaches that of the low level data

Preliminary Results



Performance as additional isolations are included in the feature set, starting with the largest and then adding subsequently smaller cones

EFP Selection

- Energy Flow Polynomials (EFPs) are parameterized functions which act over the (pT, eta, phi)'s in an event[11]
- Including these as features may recover information not captured by isos
- Average Decision Ordering (ADO) metric used to select EFP set[12]
- ADO measures probability two classifiers yield similarly ordered pair of outputs
- Iterative algorithm compares EFPs to PFN with ADO and selects optimal set

EFPs can be represented as graphs

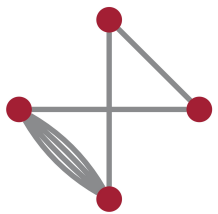
Nodes: $\sum_i^N z_i^\kappa$ Edges: θ_{ij}^β

Where:

$$z_i = \frac{p_{T,i}}{\sum_j p_{T,j}}$$

$$\theta_{ij} = \sqrt{\Delta\eta_{ij}^2 + \Delta\phi_{ij}^2}$$

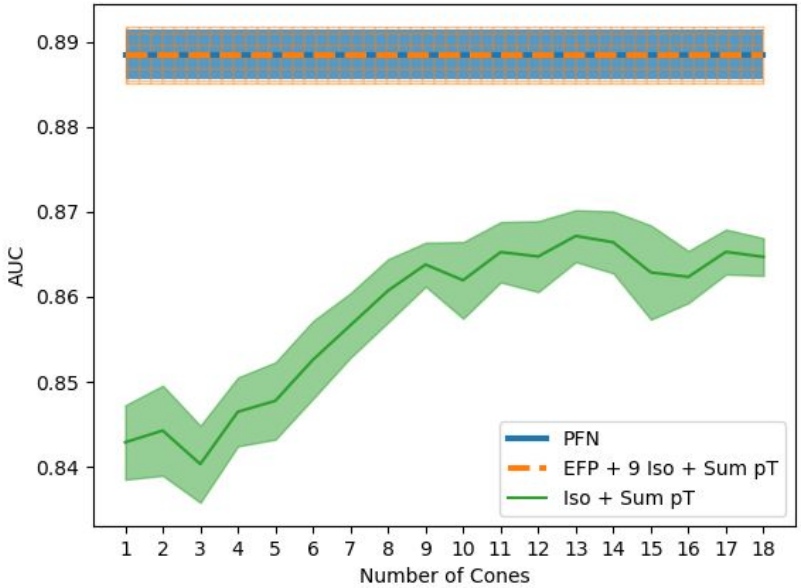
Search identifies this EFP as useful:



$$(\kappa = 1, \beta = \frac{1}{4}) = \sum_{a,b,c,d=1}^{N} z_a z_b z_c z_d \left(\theta_{ab}\theta_{ac}\theta_{bd}\theta_{cd}^4\right)^{1/4}$$

Results

- The first EFP identified by our search seems to close the performance gap between the high and low level data
- Minimum high level observables: 9 isos + summed pT + 1 EFP
- Previous results on simulation had identified 10 isos + summed pT + 5 EFPs to match PFN



Preliminary Results

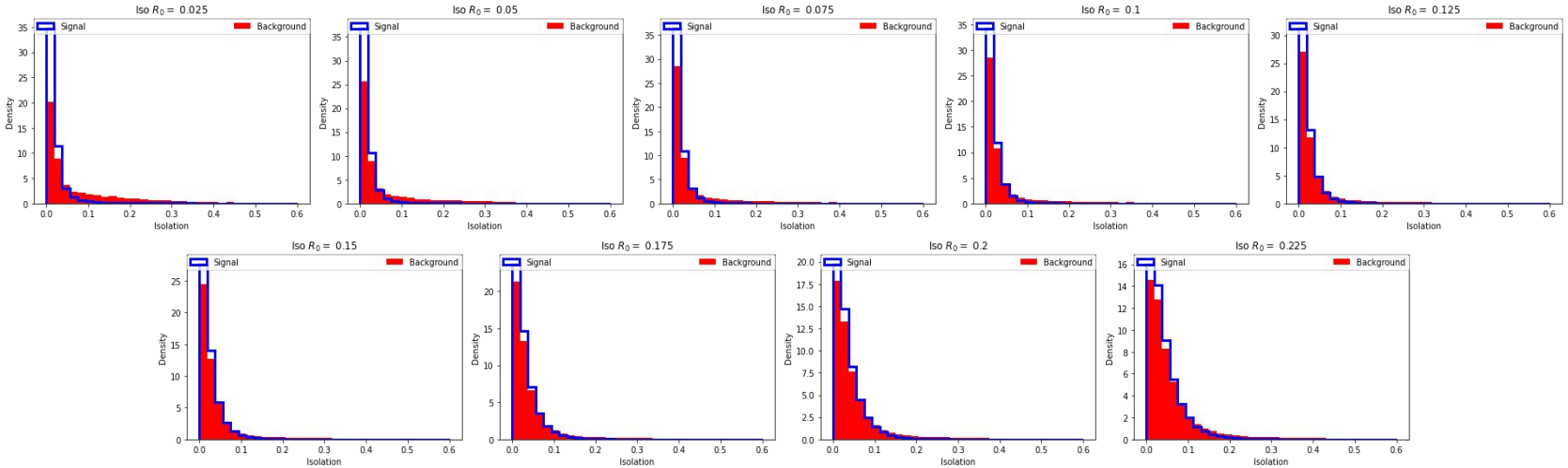
Method	AUC	σ
Single Iso Cone + $\sum p_T$	0.843	4.37×10^{-3}
9 Iso + $\sum p_T$	0.864	2.59×10^{-3}
9 Iso + $\sum p_T$ + EFP	0.888	3.30×10^{-3}
Particle-Flow Net	0.888	2.92×10^{-3}

Conclusions

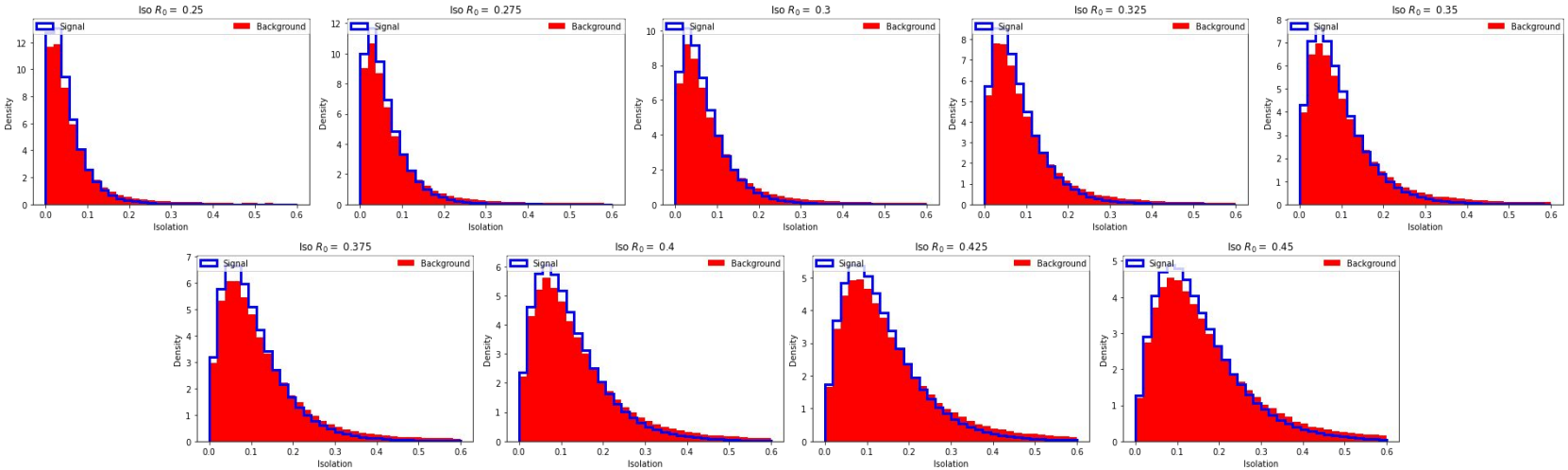
- We find that isolation does not capture all useful discriminating information present in real particle flow data
- We employ CWoLa, a weakly supervised learning strategy, to train on real data
- Notably previous studies of this kind have only been done on simulation, and CWoLa has only been used in bump hunt applications
- We identify a minimal set of interpretable high level observables that has similar discriminating power to the full low level data

Backup

CMS Isolation Distributions 0.025 - 0.225

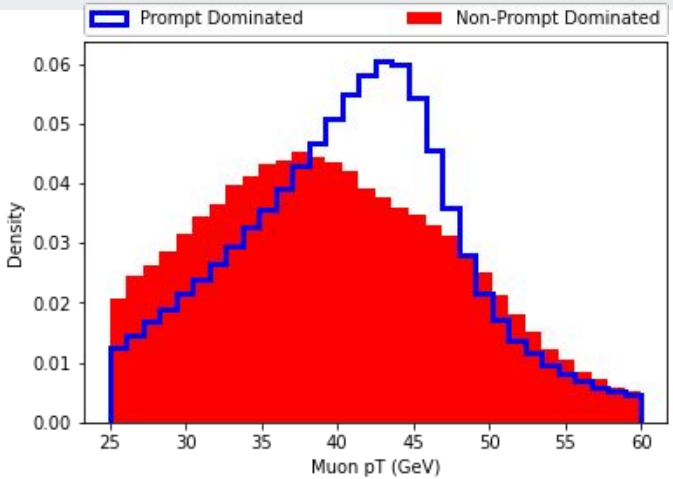


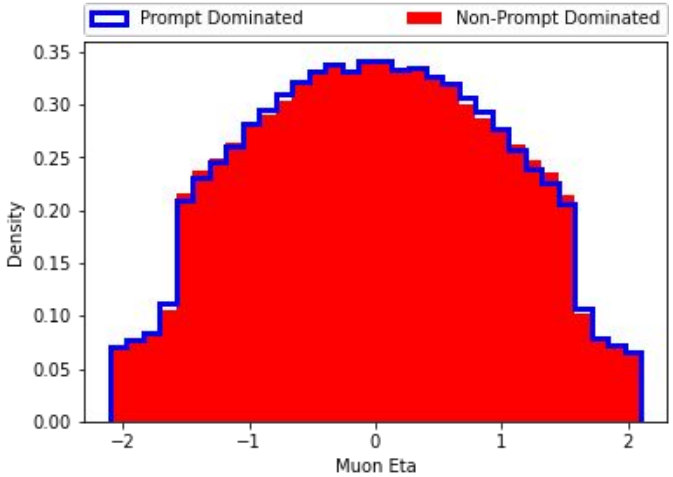
CMS Isolation Distributions 0.225 - 0.45



Muon pT / eta distributions

- Muon pT / eta distributions are shown for our separated prompt / non-prompt dominated samples
- These are weighted such that they match between the samples
- Weights are calculated using 2D gaussian KDE in the pT / eta dimensions, and applied to loss during training

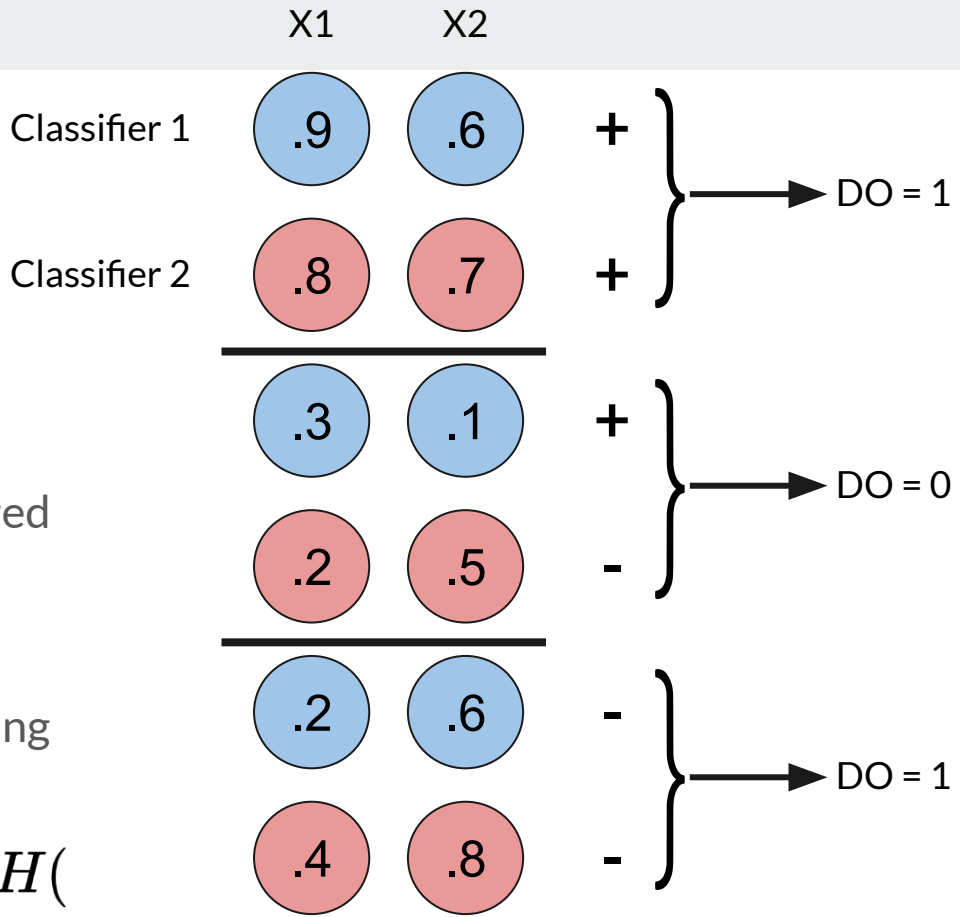




ADO[12]

- Measures probability two classifiers yield similarly ordered pair of outputs
- For classifiers **f** & **g**, classes distributed as **p_1** & **p_2** (**H** being Heaviside step)

$$\mathrm{ADO} = \int \mathrm{d}x\mathrm{d}x' p_1(x) p_2(x') H($$
$$[f(x) - f(x')][g(x) - g(x')])$$



Decision ordering calculated for individual pairs from two classifiers

Guided Search[12]

- Select random signal / bg pairs, compare on average how often two classifiers give similarly ordered output
- Select points which a high level network and the PFN order differently
- Compute ADO between PFN and a set of EFPs
- Select highest ADO EFP and include as input for high level network
- Iterate until ADO between the PFN and high level network stops improving

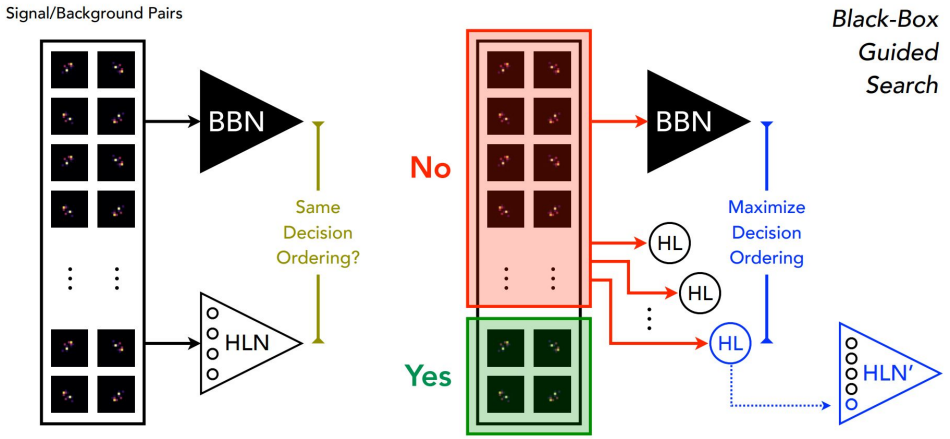


Figure from [12] Illustration of the guided search strategy. Here the PFN takes the place of the black box network, and the EFPs are used as the high level (HL) observables.

CMS Open Data Analysis - CWoLa Details[8]

- The likelihood ratio for two samples S and B defines the optimal classifier between them (letting p_S and p_B be pdfs)

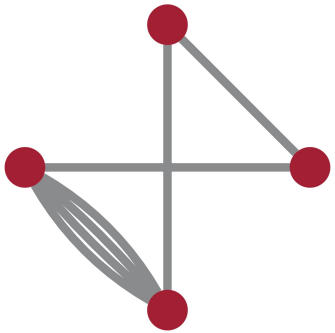
$$L_{S/B}(\vec{x}) = p_S(\vec{x})/p_B(\vec{x})$$

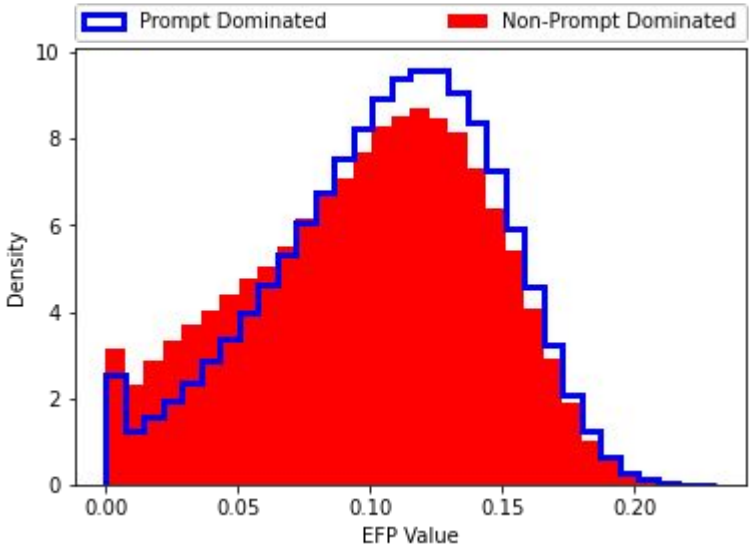
- Given mixed samples M1 and M2 with signal fractions f1 and f2, we can relate their likelihood ratio to that of the one for S and B

$$L_{M_1/M_2} = \frac{p_{M_1}}{p_{M_2}} = \frac{f_1\, p_S + (1-f_1)\, p_B}{f_2\, p_S + (1-f_2)\, p_B} = \frac{f_1\, L_{S/B} + (1-f_1)}{f_2\, L_{S/B} + (1-f_2)}$$

- We see that the likelihood ratio between M1 and M2 is a rescaling of the likelihood ratio for S and B
- Since rescaling doesn't change the decision ordering these two are equivalent at the classification level
- Training on our mixed samples should therefore yield a network that can classify between the signal and background that we're interested in

Selected EFP





$$(\kappa = 1, \beta = \frac{1}{4}) = \sum_{a,b,c,d=1}^{N} z_a z_b z_c z_d \left(\theta_{ab}\theta_{ac}\theta_{bd}\theta_{cd}^4\right)^{1/4}$$

Sources

[1] A. M. Sirunyan et al. (CMS), JINST 12, P10003 (2017), 1706.04965.

[2] J. Pata, J. Duarte, J.-R. Vlimant, M. Pierini, and M. Spiropulu (2021), 2101.08578.

[3] A. M. Sirunyan et al. (CMS), JINST 12, P10003 (2017), 1706.04965.

[4] J. Collado, K. Bauer, E. Witkowski, T. Faucett, D. Whiteson, and P. Baldi, JHEP 21, 200 (2020), 2102.02278

[5] CMS collaboration (2017). DoubleMuParked primary dataset in AOD format from Run of 2012 (/DoubleMuParked/Run2012C-22Jan2013-v1/AOD). CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.M5AD.Y3V3

Sources

[6] CMS Collaboration, JINST 12 (2017) P10003, 1706.04965

[7] CMS Collaboration, “Particle-flow reconstruction and global event description with the CMS detector”, JINST 12 (2017) P10003, doi:10.1088/1748-0221/12/10/P10003, arXiv:1706.04965.

[8] Eric M. Metodiev, Benjamin Nachman, Jesse Thaler, JHEP 10 (2017) 174, 1708.02949

[9] Schmelling, Michael, ”Using sWeights to disentangle signal and background”, lecture, PHYSTAT-Flavour2020, Oct 21, 2020

[10] P. T. Komiske, E. M. Metodiev, and J. Thaler, Journal of High Energy Physics 2019 (2019), ISSN 1029-8479

Sources

[11] P. T. Komiske, E. M. Metodiev, and J. Thaler, JHEP 04, 013 (2018), 1712.07124.]

[12] T. Faucett, J. Thaler, and D. Whiteson (2020), 2010.11998