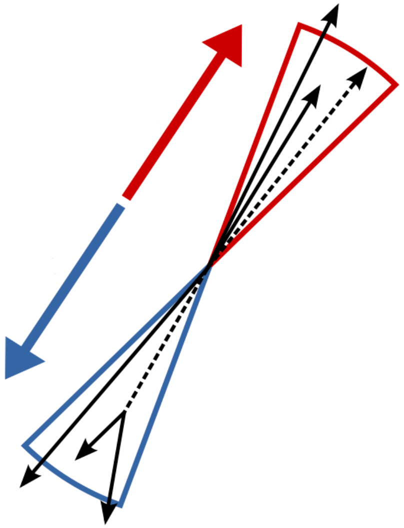
Improved continuum suppression using deep neural network with low-level input

Ori Fogel, Abner Soffer, Ran Gilad-Bachrach, Ofir Barogel
Tel Aviv University

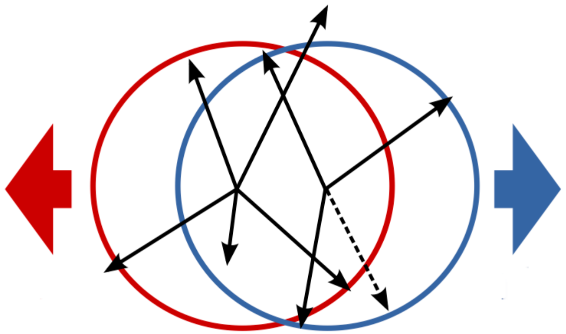
22 February 2024

Continuum Supression

- Continuum background ($ee \rightarrow q\bar{q}$)
- Event-shape variables are regularly used to suppress $q\bar{q}$ background.
- Combined with BDT/NN algorithms, e.g. FastBDT.
- These algorithms use high-level variables.



$q\bar{q}$



$B^{-}B^{+}$

Continuum Supression using high-level variables

https://software.belle2.org/development/sphinx/online_book/basf2/cs.html

- **Parameters usually used at Belle II:**
  - Ratio of the second and zeroth Fox-Wolfram moment: $R_2 = \frac{H_2}{H_0}$
  - Total thrust magnitude of both B candidate and ROE
  - $cos\theta_{B0}$ angle b/w thrust axes of B candidate and ROE
  - $cos\theta_p$ polar angle of thrust axis of B candidate
  - CLEO cones
  - KSFW variables
  - **All these variables aggregate particle momenta.**

Variable	Abbreviation
CleoConeCS(5)	CleoC1
KSFWVariables(hoo3)	KSFWV1
CleoConeCS(7)	CleoC2
KSFWVariables(hso14)	KSFWV2
CleoConeCS(6)	CleoC3
CleoConeCS(8)	CleoC4
CleoConeCS(4)	CleoC5
KSFWVariables(hoo1)	KSFWV3
CleoConeCS(9)	CleoC6
KSFWVariables(hoo4)	KSFWV4
KSFWVariables(hso04)	KSFWV5
KSFWVariables(mm2)	KSFWV6
KSFWVariables(hso24)	KSFWV7
KSFWVariables(hso20)	KSFWV8
KSFWVariables(hso00)	KSFWV9
thrustOm	thrus1
KSFWVariables(hoo0)	KSFWV10
KSFWVariables(et)	KSFWV11
CleoConeCS(3)	CleoC7
thrustBm	thrus2
KSFWVariables(hso22)	KSFWV12
KSFWVariables(hoo2)	KSFWV13
CleoConeCS(1)	CleoC8
CleoConeCS(2)	CleoC9
KSFWVariables(hso02)	KSFWV14
KSFWVariables(hso12)	KSFWV15
cosTBz	cosTB1
KSFWVariables(hso10)	KSFWV16
R2	R2
cosTBTO	cosTB2

Continuum Supression using low-level variables

- Aggregating information into high-level variables results in loss of information contained in the low-level variables (particle momenta, etc.)

- Different approach: use low-level variables of each particle as an input, and let the algorithm figure out how best to use them.

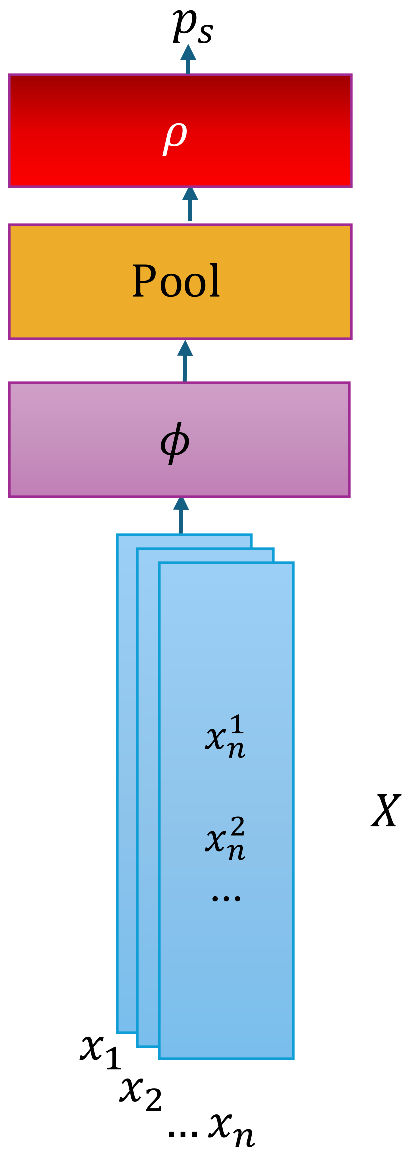
- **The problem**: BDT and various types of NN cannot incorporate a different number of inputs (particles) in each event.

Continuum Supression using Deep Sets

- **What is Deep Sets?**

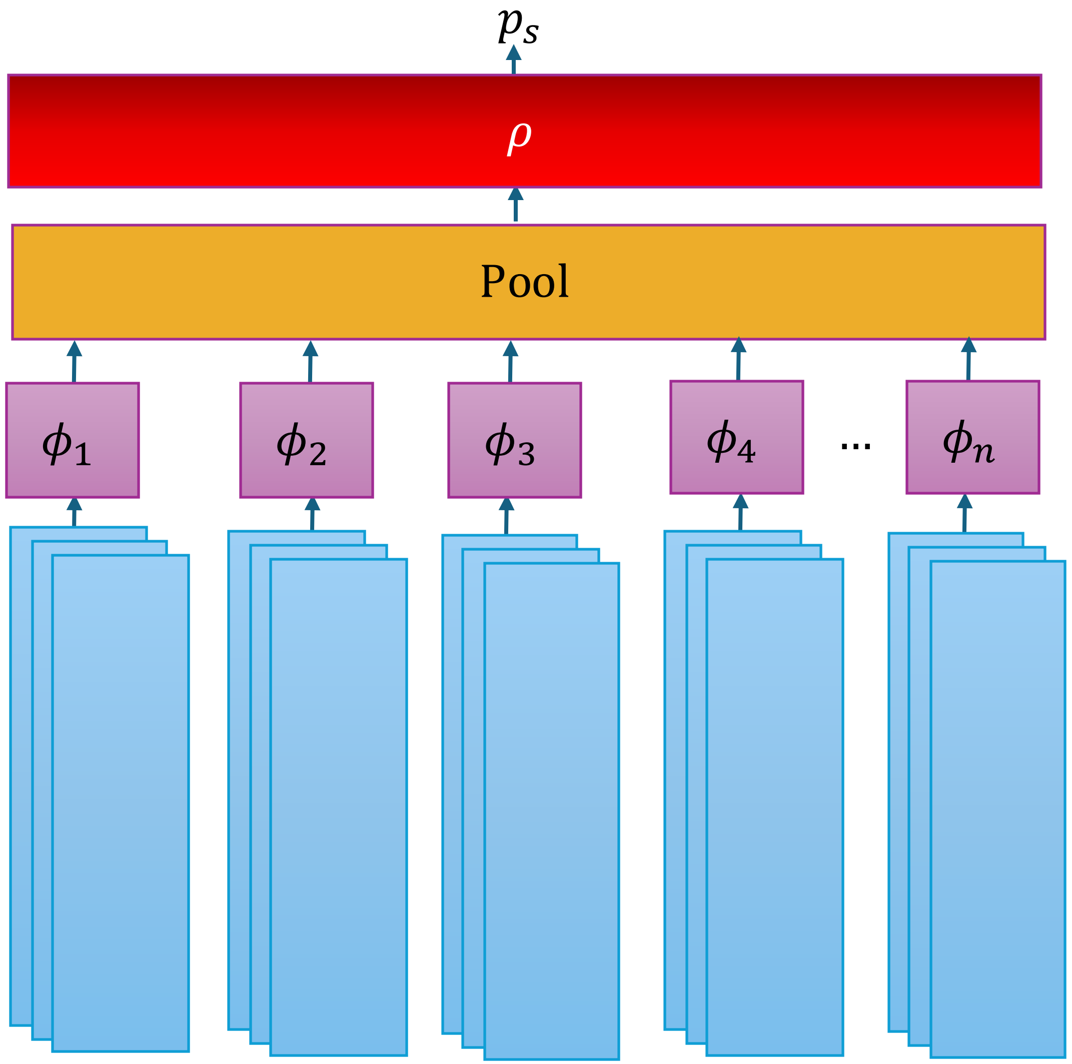
  https://arxiv.org/abs/1703.06114

  - DeepSets is a NN architecture that takes as input an unordered set $X = \{x_i\},\ i = \{1 \dots n\}$ with varying size n, where each element has features x_i^j.
  - Each element is fed into the same NN ϕ.
  - The n outputs of ϕ are aggregated with a permutation-invariant pooling operation (Sum/Mean/Max).
  - The aggregated representation is passed to another neural network ρ to produce the final output p_s.



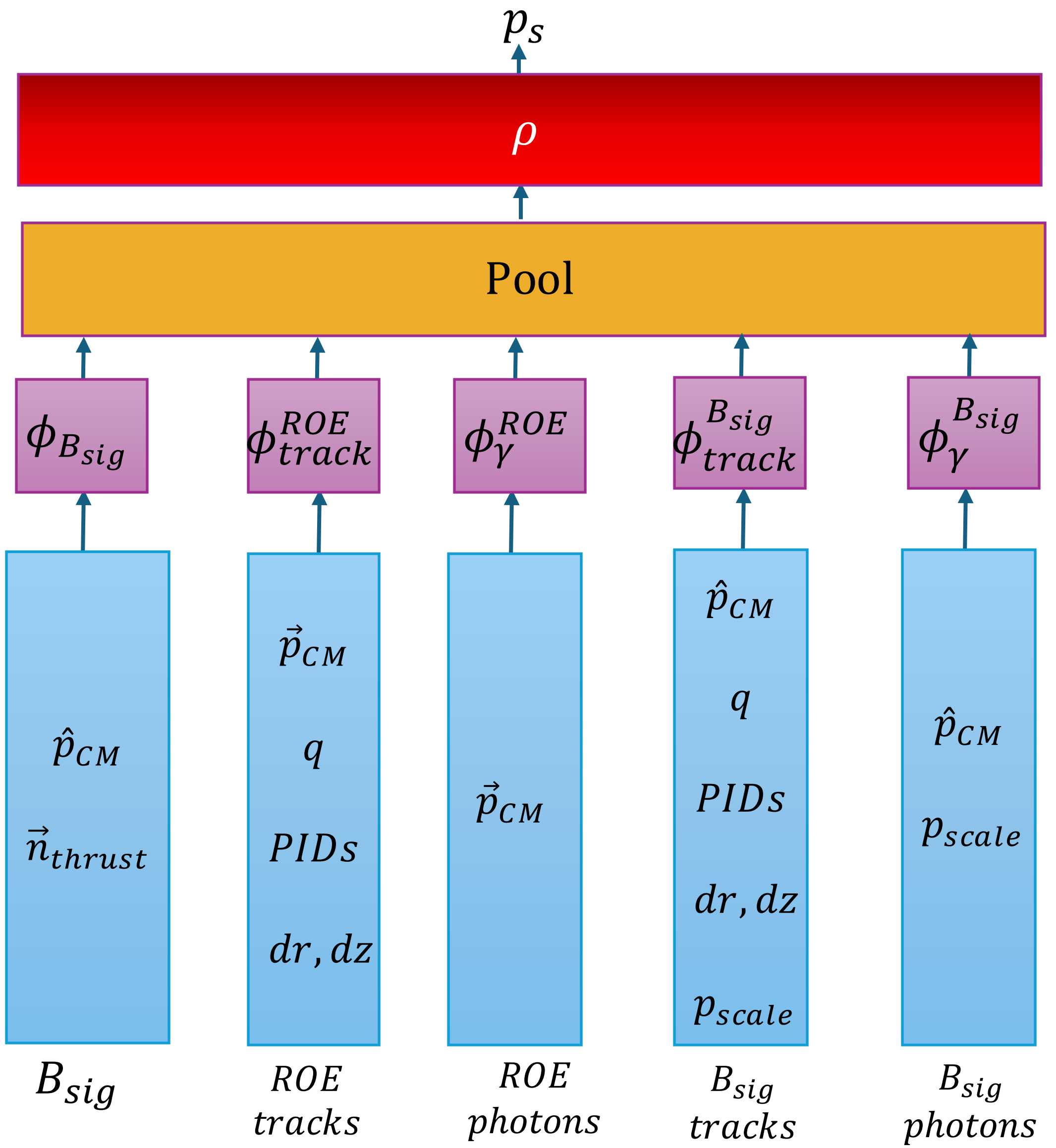
Continuum Supression using Deep Sets

- **<u>MultiDeepSets (MDS)</u>**
  - Our data contains different types of objects (tracks, photons).
  - MDS is a modification of Deep Sets (developed mostly by Roy Hircsh and Emilie Bertholet) in an X(3872) analysis.
  - Can deal with multiple sets.
  - Uses multiple ϕ_i NNs, each of which gets a different set as an input.
  - Pooling all ϕ_i output and proceeds the same as DeepSets.



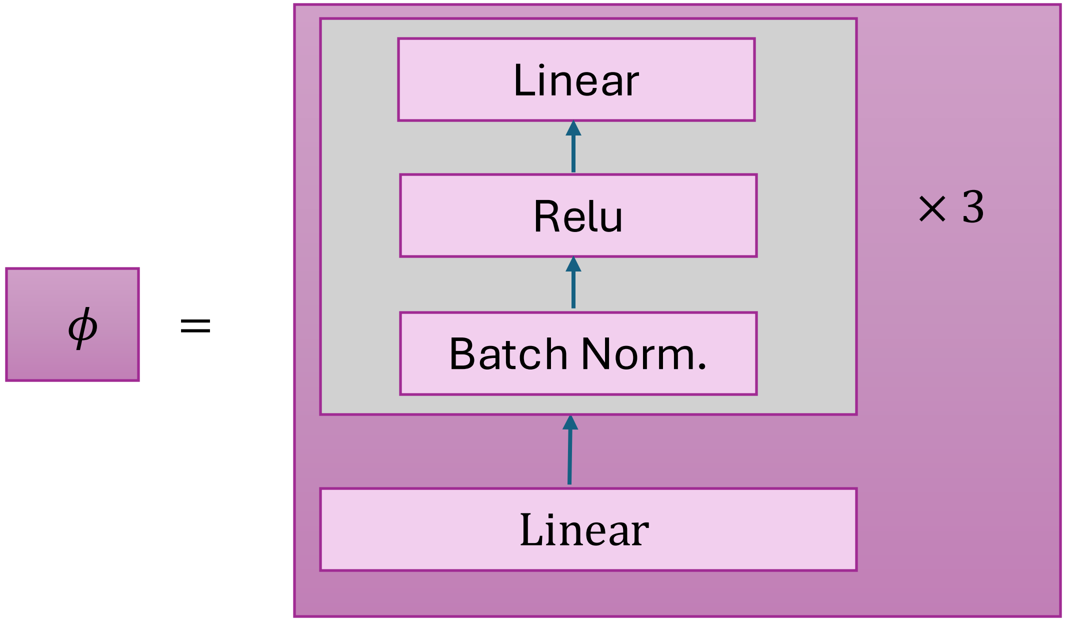
Continuum Supression using Deep Sets

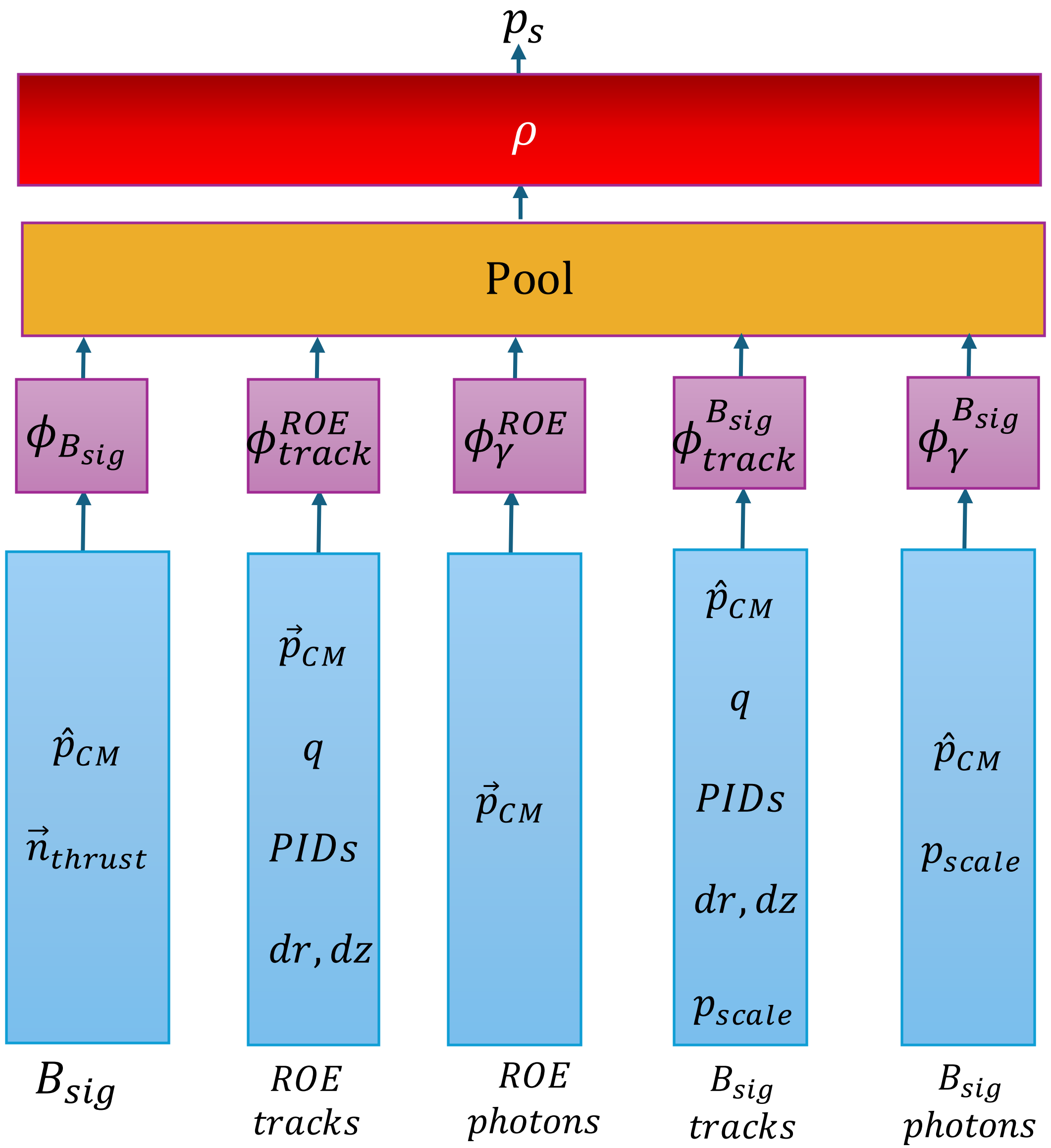
- **Architecture:**
  - 5 sets are fed as input: B^{sig} , tracks, γ-s (sig and ROE)
  - Each set includes arrays of inputs for each particle i:
    - $\boldsymbol{B^{sig}}$ (1 arrays)
    - **tracks** $(i = 1,2 \ldots \leq 10)$
    - **photons** $(i = 1,2 \ldots \leq 20)$
  - To avoid correlation with M_{bc} and ΔE, for the B_{sig} particles we use only:
    - $\hat{p}$: θ and ϕ angles.
    - p_{scale}: normalized momenta of tracks and gammas relative to the highest momentum.



Continuum Supression using Deep Sets

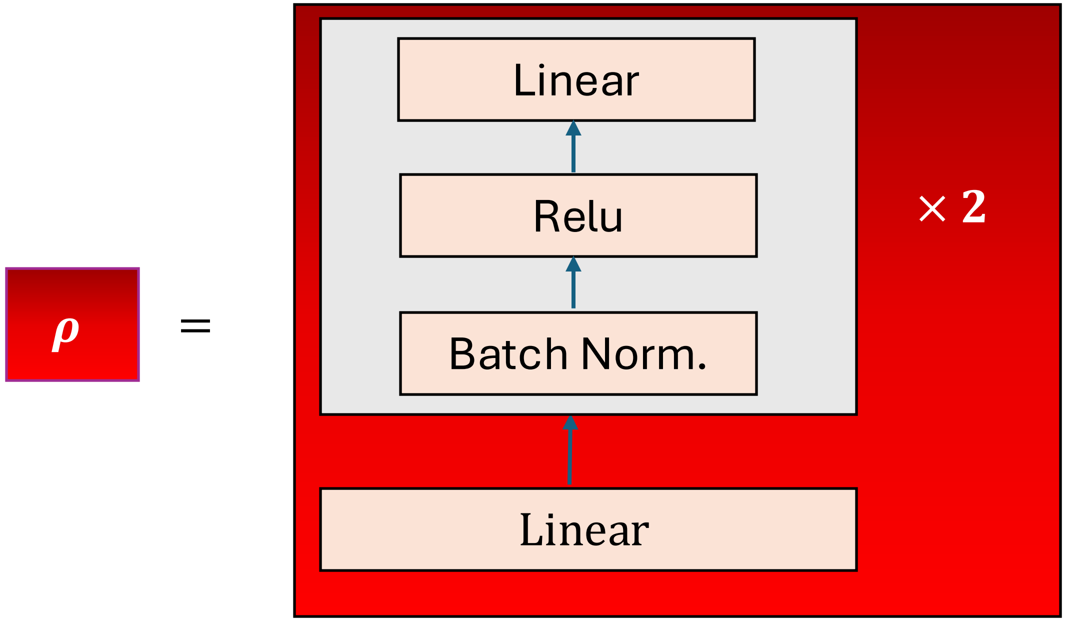
- **Architecture:**
  - Feed inputs into MLP ϕ (independently for each set).
  - Repeated block sizes: [20,50,100]
  - Calculate mean value of the ϕ outputs (size 150 each).

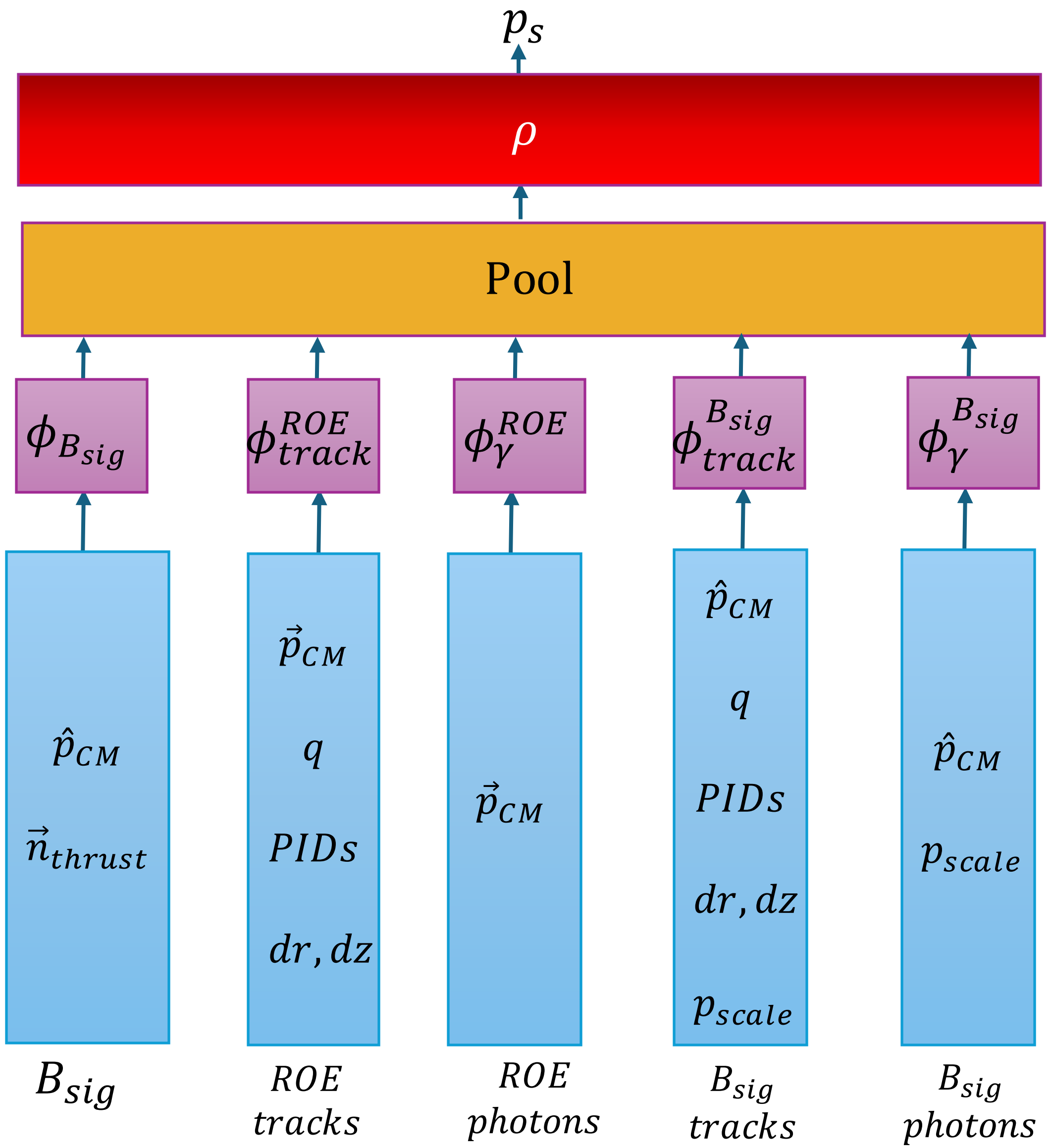




Continuum Supression using Deep Sets

- **Architecture:**
  - Feed mean (size 150) into MLP ρ, which gives output p_s.
  - Repeated block sizes: [100,50]





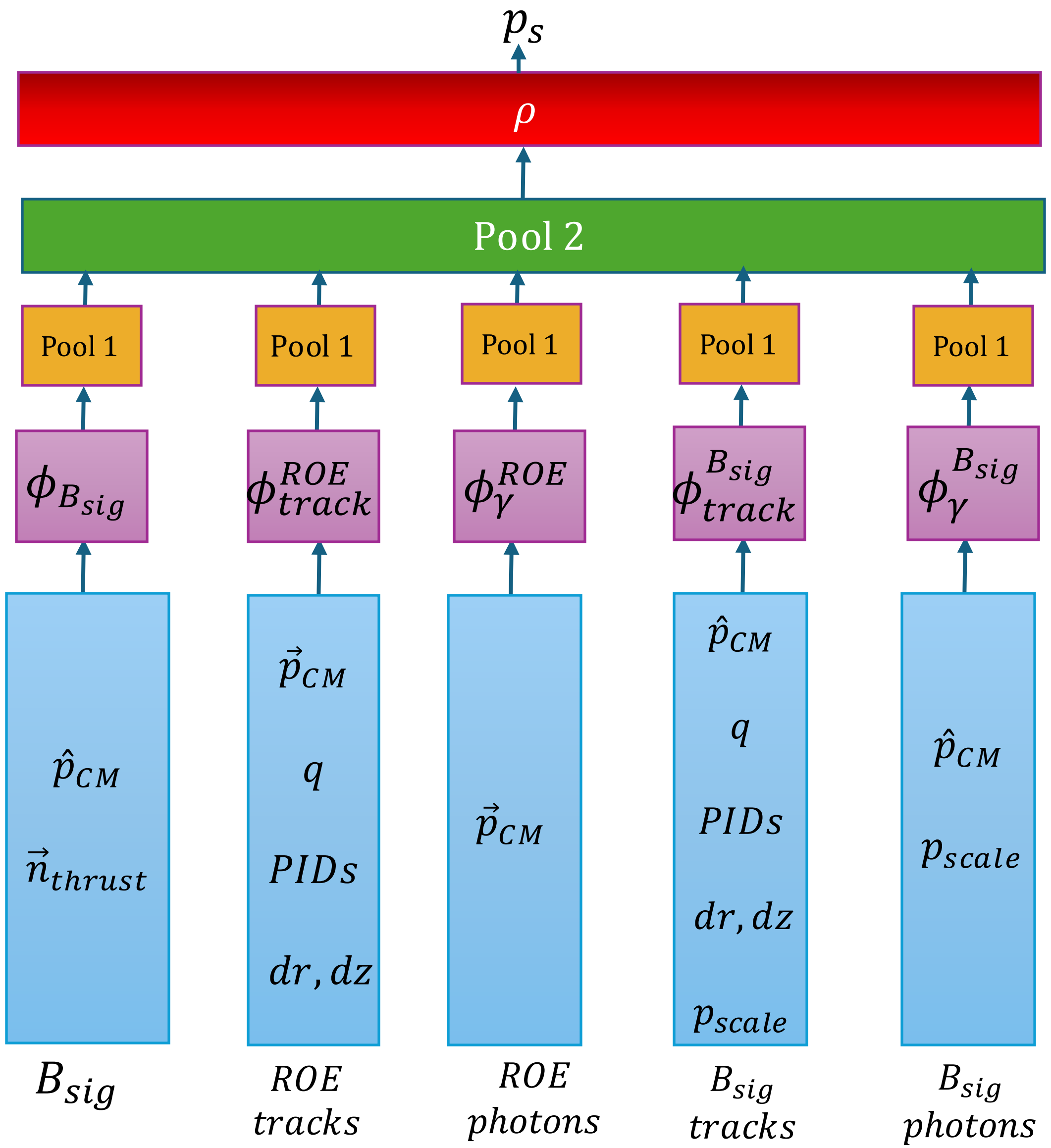
Other possible architectures

- The architecture I just presented uses **early fusion**, as it combines information from all sets immediately after the ϕ transformation.

- We tried different approaches:

Late Fusion

- **Architecture:**
  - Pooling each set separately.
  - Combining information from all the sets only after the first pooling.
  - Pool 1 $\neq$ Pool 2

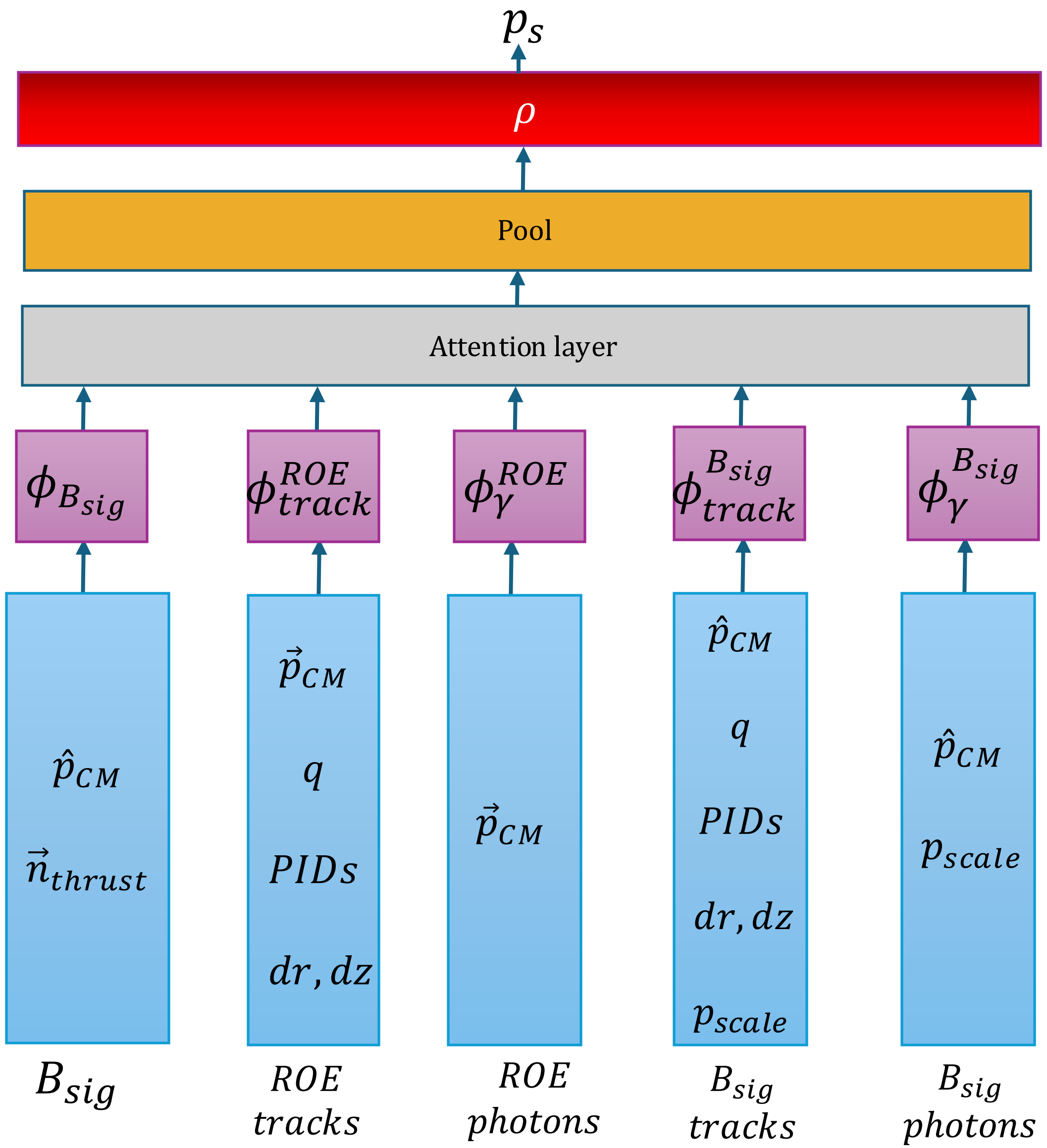


Early Fusion with an Attention Layer

- **Architecture:**

  - An attention layer enables the model to focus on relevant interactions between elements.

  - Here it captures the interactions of all the particles, as we give as an input all the different sets together.



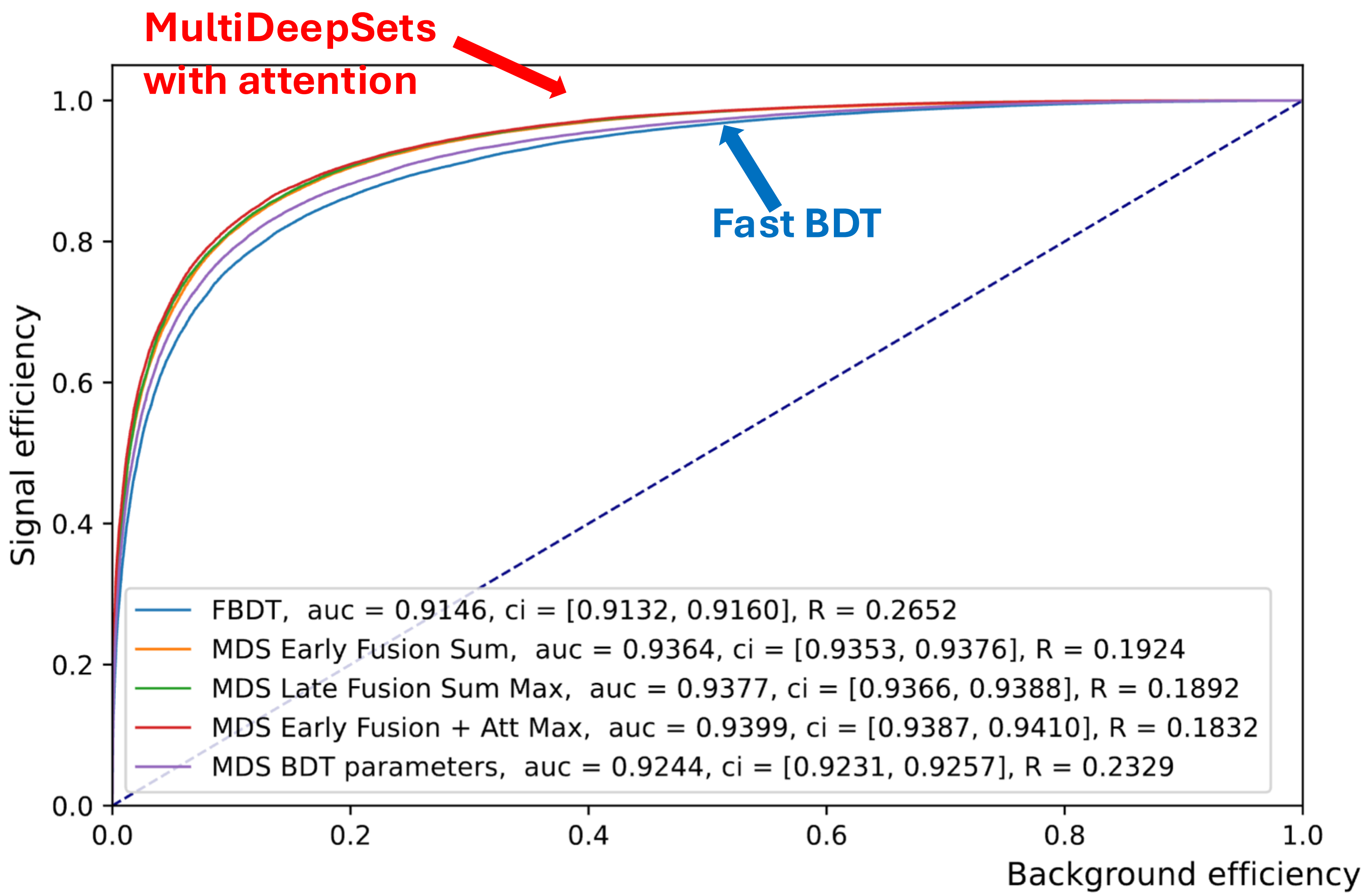
Sample used

- FEI hadronic skims, for convenience and to have a variety of decay modes (although missing 2-body charmless decays)

	$L[ab^{-1}]$	# evnets
B^+B^-	0.01215	752038
$u\bar{u}$	0.00259	296246
$d\bar{d}$	0.00259	73450
$c\bar{c}$	0.00259	327108
$s\bar{s}$	0.00259	53648

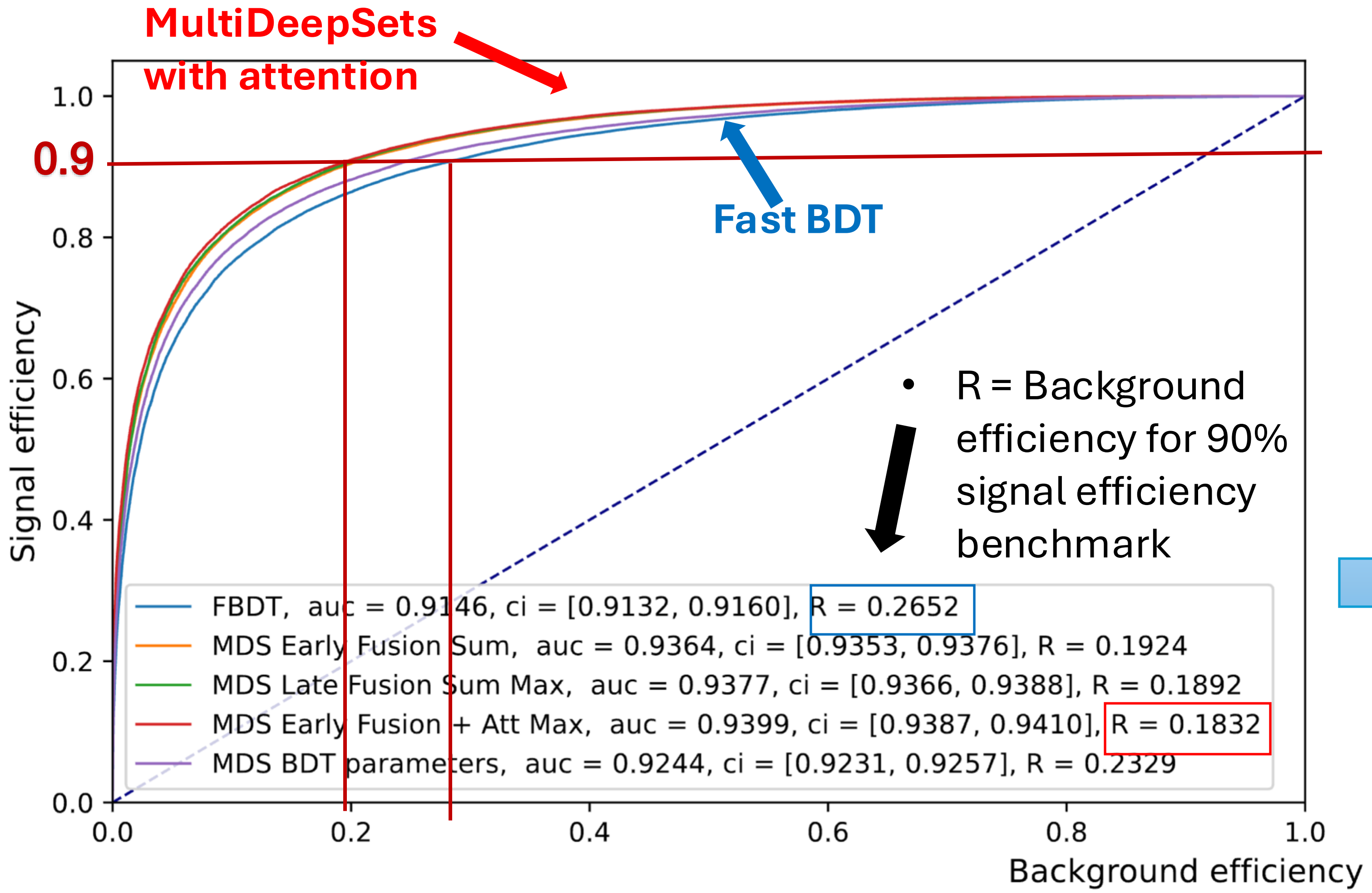
- **For training and validation**: ~750K signal events and ~750K bg events.
- **For inference:** independent 75K signal events and 75K bg events.
- B^{sig} is reconstructed, and we keep up to 10 tracks and 20 photons (for B^{sig} and ROE).

Results



MultiDeepSets
with attention
Fast BDT
Signal efficiency
Background efficiency
0.0
0.2
0.4
0.6
0.8
1.0
FBDT, auc = 0.9146, ci = [0.9132, 0.9160], R = 0.2652
MDS Early Fusion Sum, auc = 0.9364, ci = [0.9353, 0.9376], R = 0.1924
MDS Late Fusion Sum Max, auc = 0.9377, ci = [0.9366, 0.9388], R = 0.1892
MDS Early Fusion + Att Max, auc = 0.9399, ci = [0.9387, 0.9410], R = 0.1832
MDS BDT parameters, auc = 0.9244, ci = [0.9231, 0.9257], R = 0.2329

Results

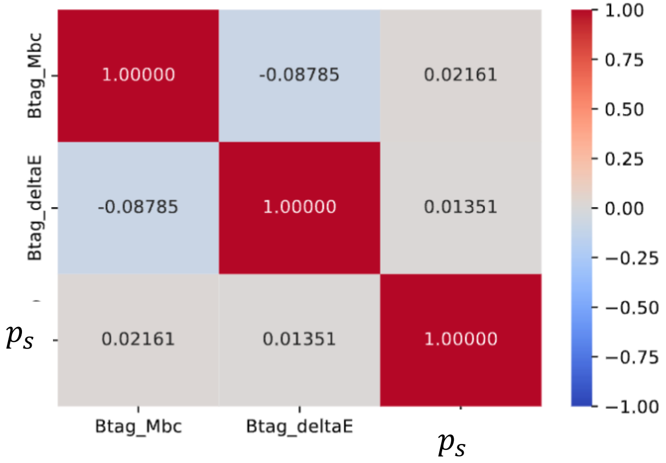


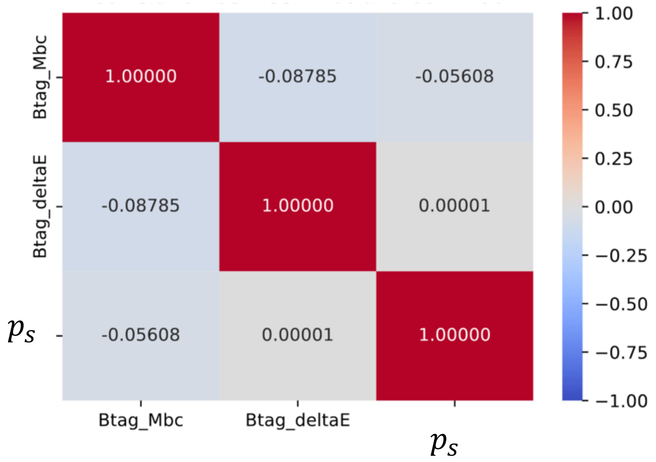
For benchmark of 90% signal efficiency, we reduce background by 31%!

Results - correlation

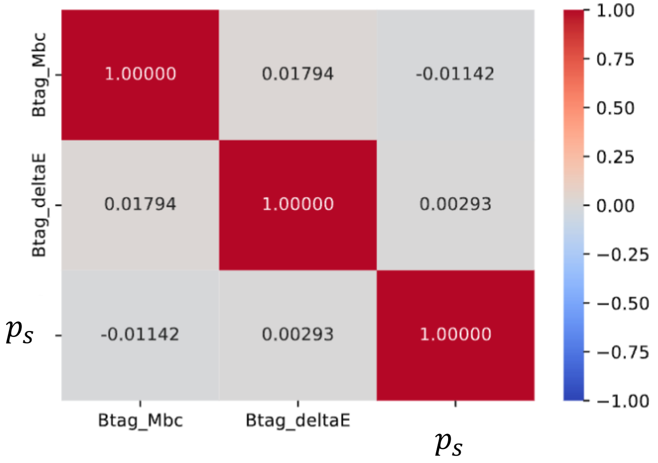
FastBDT | MultiDeepSets (attention)

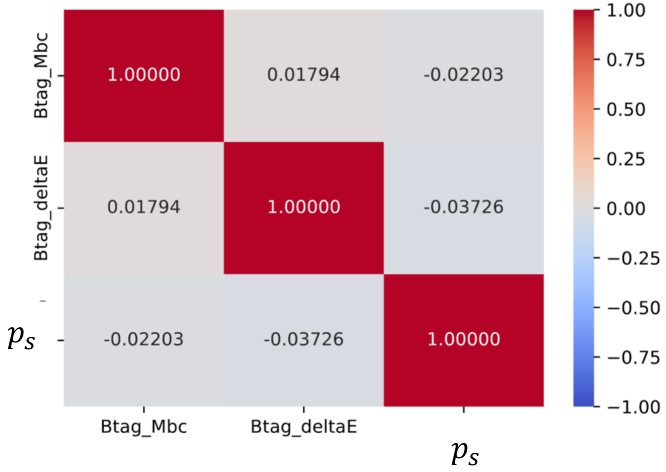
Continuum





Signal





FEI decay modes

Decay Mode	# events (ouf of 150K)	FastBDT		MultiDeepSets (with attention)		Background reduction for 90% signal
		AUC	R	AUC	R	
$B^+ \to \bar{D}^0\pi^+\pi^+\pi^-$	31196	0.9074	0.2946	0.9394	0.1863	36.76%
$B^+ \to \bar{D}^0\pi^+\pi^0$	30503	0.9237	0.2261	0.9360	0.1956	13.4%
$B^+ \to \bar{D}^0\pi^+$	17200	0.9271	0.2230	0.9452	0.1683	24.52%
$B^+ \to \bar{D}^0\pi^+\pi^+\pi^-\pi^0$	15112	0.8886	0.3541	0.9370	0.2008	43.29%
$B^+ \to \bar{D}^0\pi^+\pi^0\pi^0$	8208	0.9087	0.2821	0.9273	0.2295	18.64%
$B^+ \to \bar{D}^{*0}\pi^+\pi^+\pi^-$	8046	0.9060	0.2930	0.9274	0.2223	24.13%
$B^+ \to \bar{D}^{*0}\pi^0$	5568	0.9246	0.2166	0.9372	0.1895	12.51%
Other modes	34215	0.9094	0.2884	0.9411	0.1799	37.62%

Summary

- We wanted check if low-level variables would yield better continuum suppression.
- We used FEI to reconstruct many signal modes.
- We used all tracks and photons (in the ROE with $\vec{p}$ and in the signal B with $\hat{p}$) plus $\hat{p}$ and thrust axis of the signal B as input to a DeepSets-based NN.
- For a benchmark of 90% signal efficiency, the background efficiency is:
    - 26.5% with FastBDT
    - 18.3% with DeepSets

31% less background

- Multibody modes have a bigger improvement (because there is more information for the classifier to use).
- It highlights the importance of correct distributions in the MC, probably more so in the signal simulation.

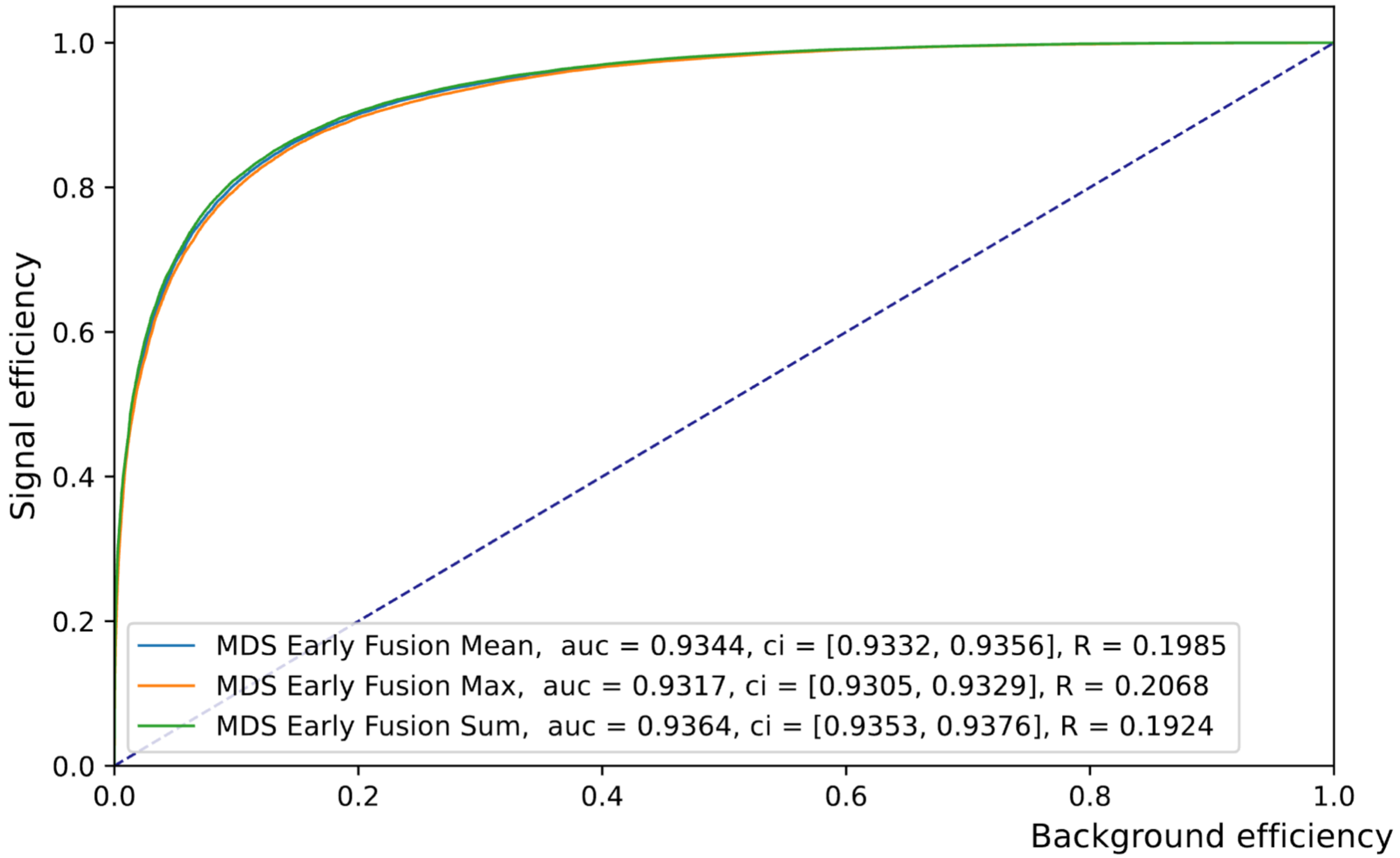
Next Steps

- Experiencing with a second attention layer.
- Study signal-mode dependence, particularly 2-body charmless signal.
- Check performances with and without retraining for the specific signal mode.
- Compare performance on data (using $B \to D^{*}\pi$ for signal and off-resonance data for continuum).
- Make the code available for Belle II use in basf2.

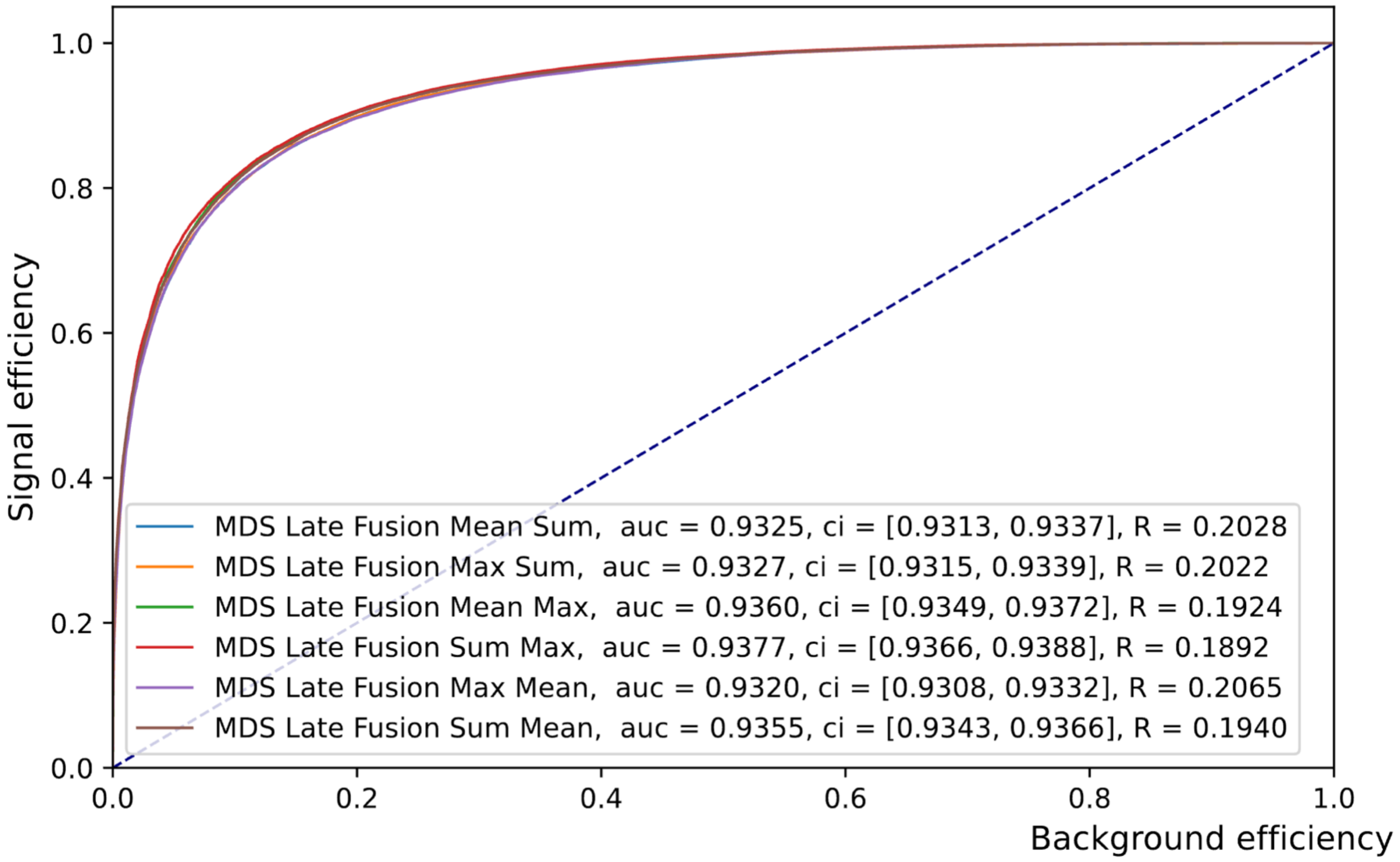
Thank You!

Backup

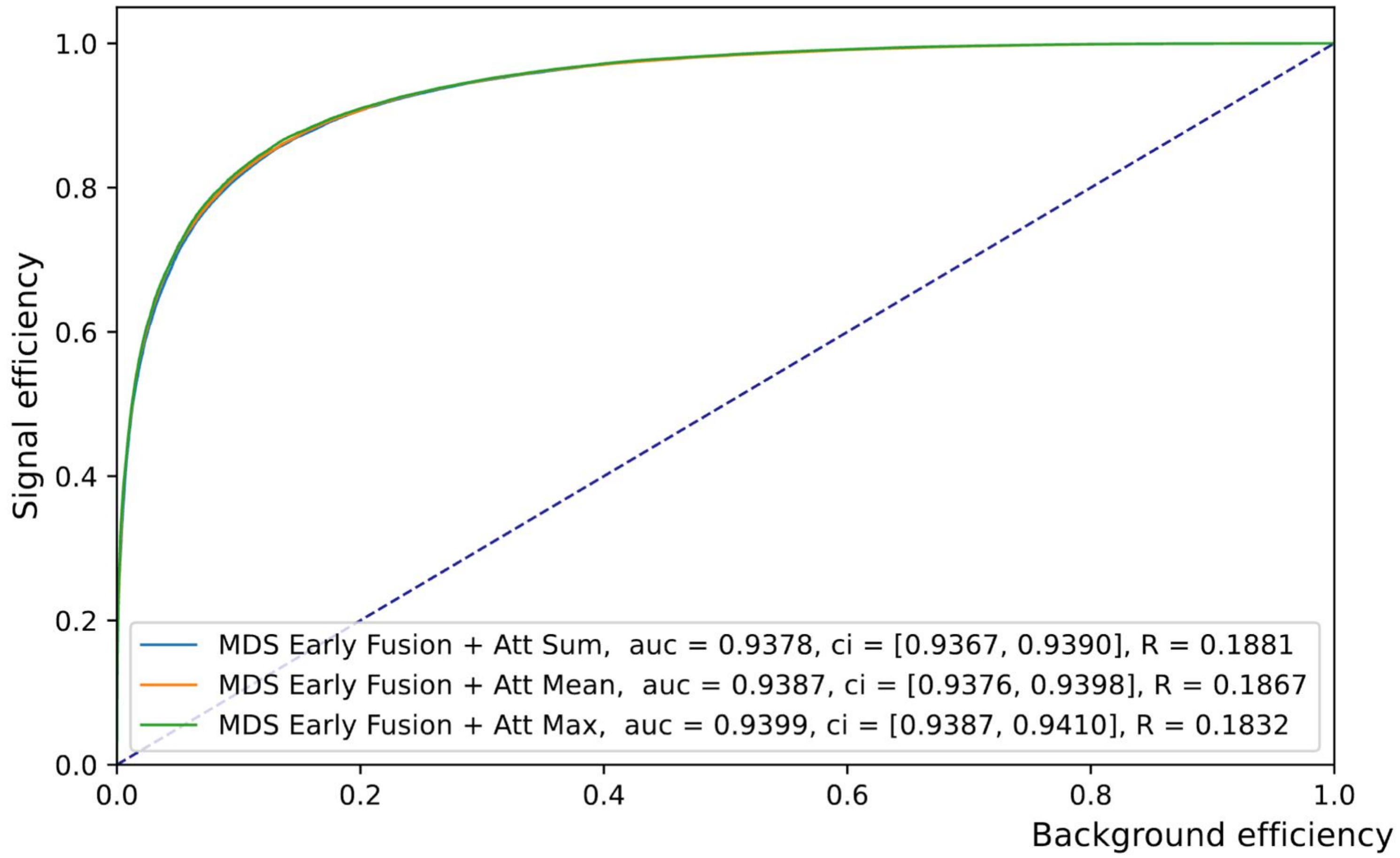
Early Fusion - Pooling Comparison



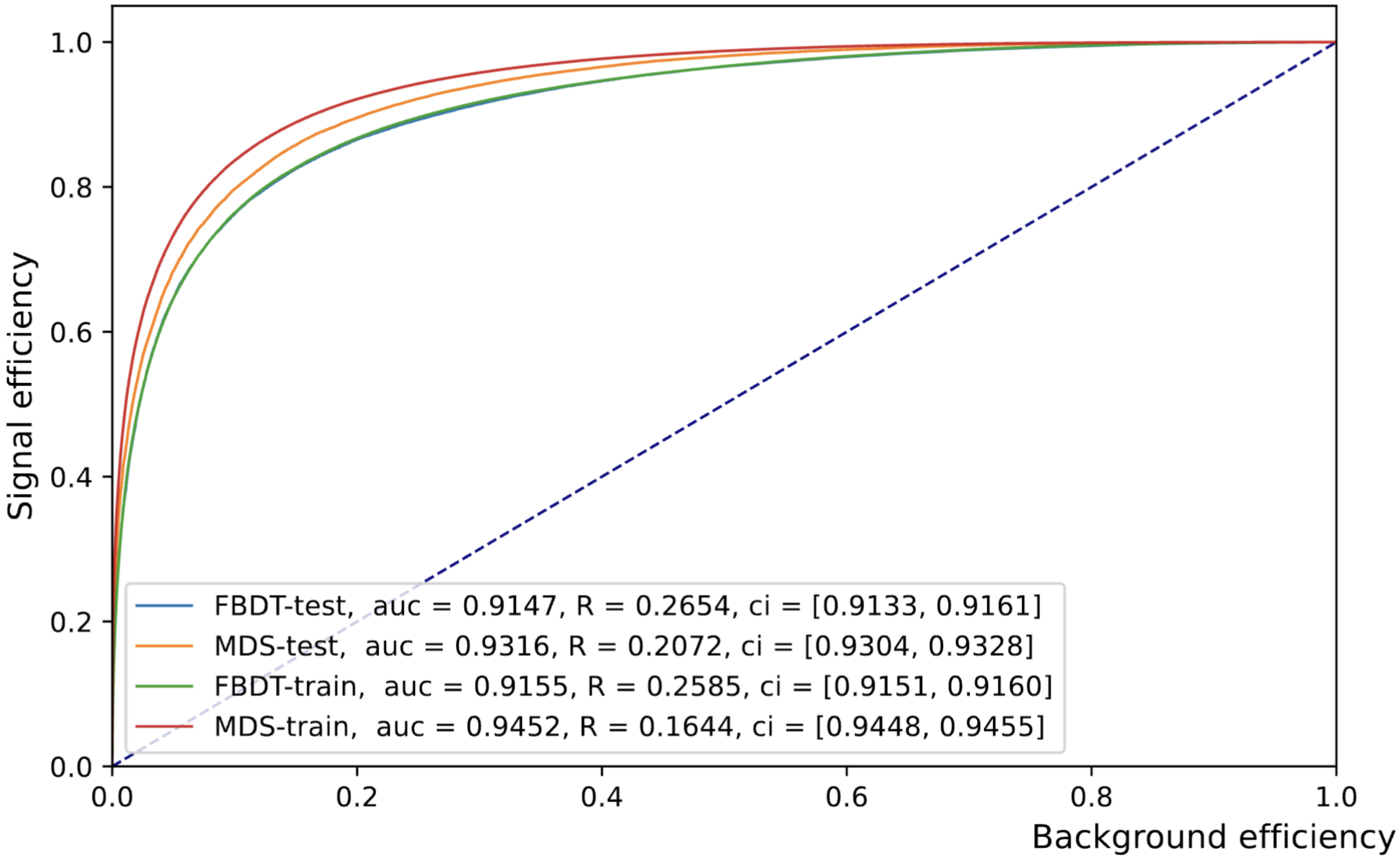
Late Fusion - Pooling Comparison



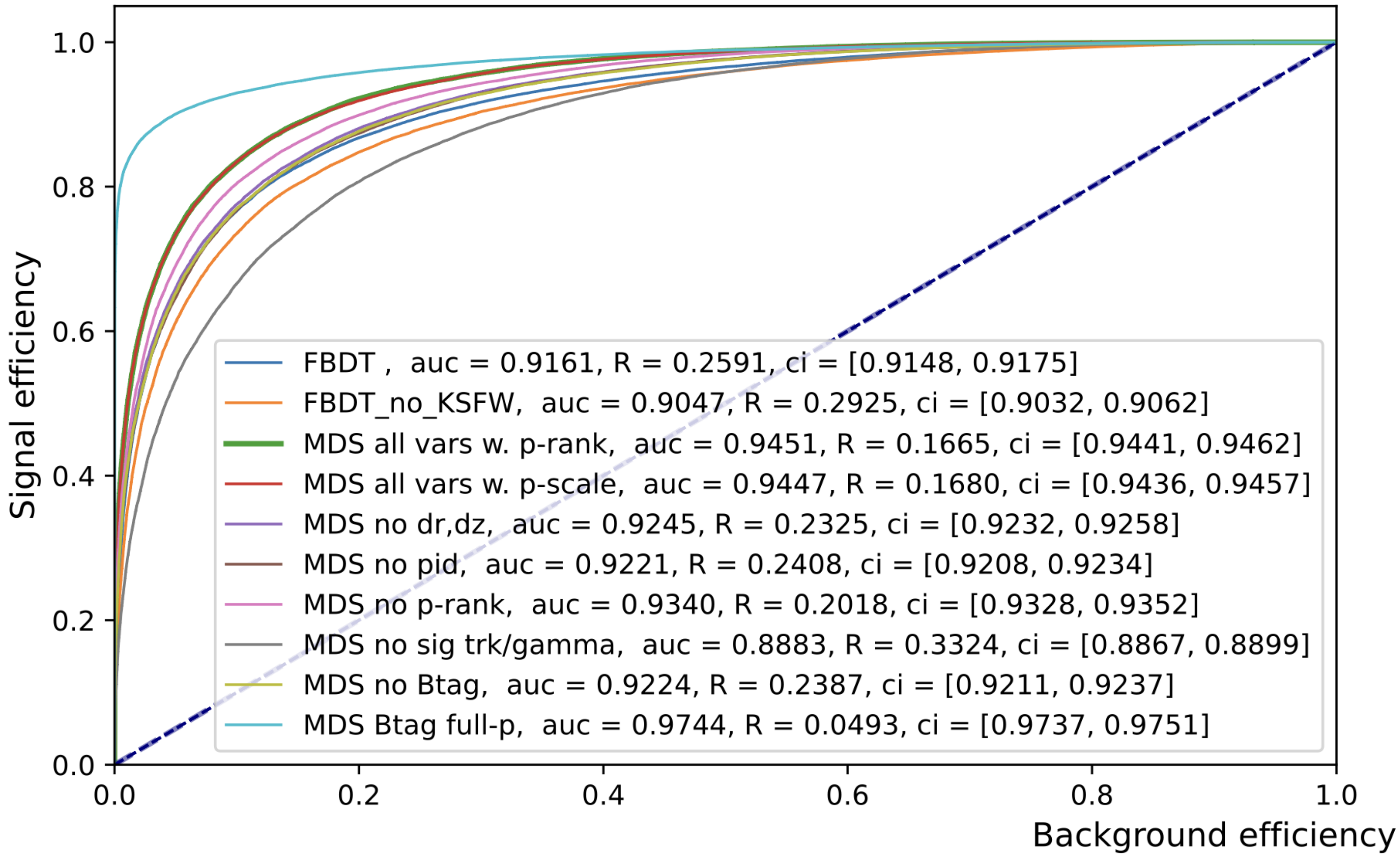
Early Fusion + Attention Fusion - Pooling Comparison



Training testing comparison (old numbers)

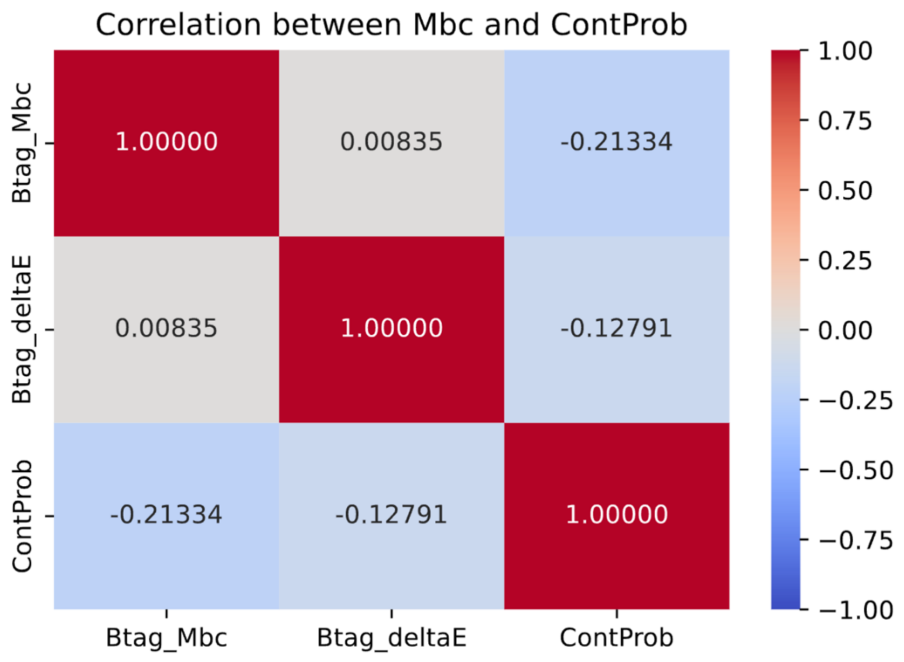


Roc curves for all tests

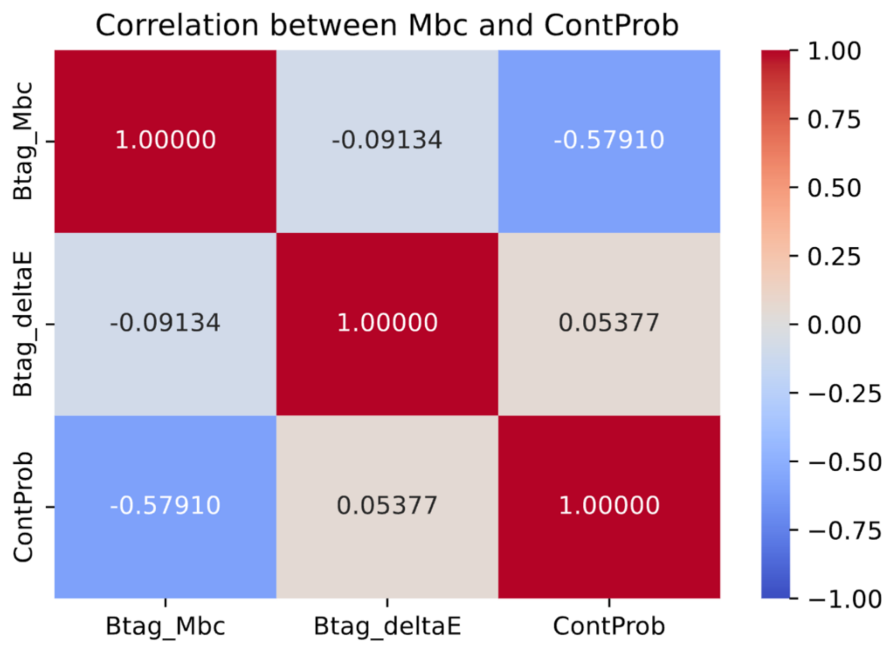


Correlation using B_{tag} momentum

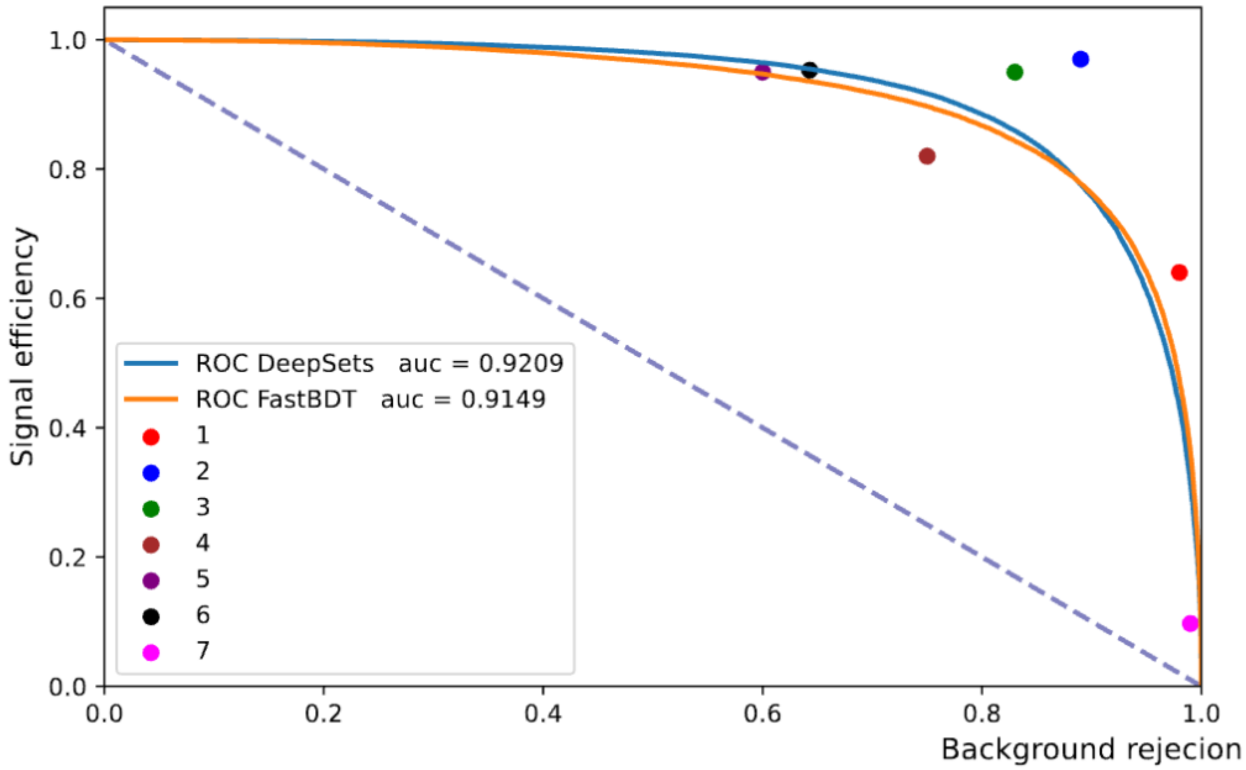
Signal



Background



FastBDT vs. DeepSets using FastBDT data



DeepSets 90% signal eff: 77.87% bg rejection

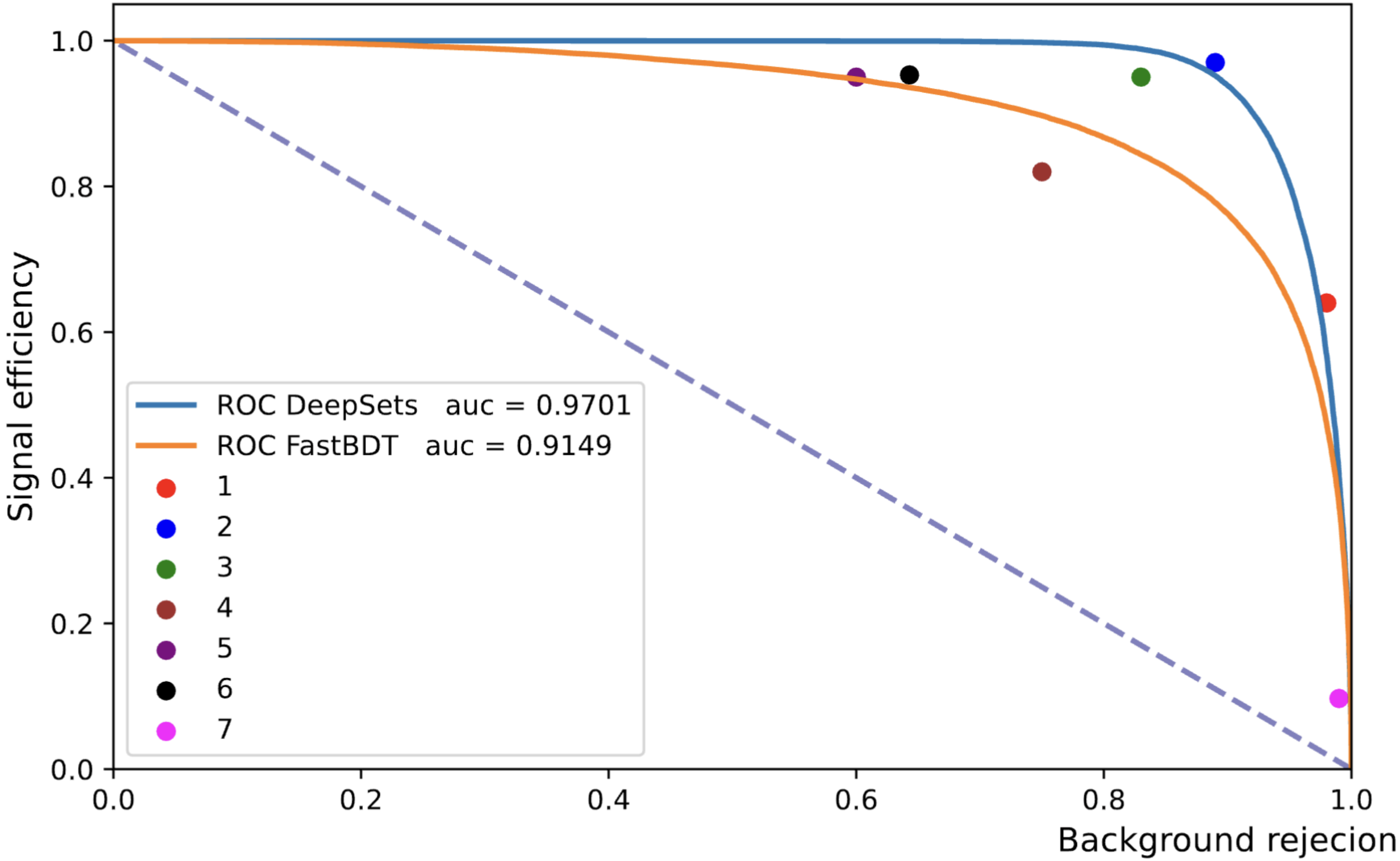
FastBDT 90% signal eff: 74.36% bg rejection

	Decay
1	$B^0 \to \pi^0\pi^0$
2	$B^0 \to J/\psi\pi^0$
3	$B^0 \to K^*(892)\gamma$
4	$B^0 \to \gamma\gamma$
5	$B^0 \to \eta' K_S^0$
6	$B^\pm \to DK^\pm$ $B^\pm \to D\pi^\pm$
7	$B^- \to D^0\rho^-$

Important note about previous results

- In the last talk, we presented almost perfect classifier.
- We noticed that we used $\vec{p}$ of B_{sig} instead of $\hat{p}$.
- Hence classifier output was highly correlated with M_{bc}
- We fixed that by using $\hat{p}$,but it required adding signal side information (similar to KSFW) due to decrease of AUC.

Old Results



	Decay
1	$B^0 \to \pi^0\pi^0$
2	$B^0 \to J/\psi\pi^0$
3	$B^0 \to K^*(892)\gamma$
4	$B^0 \to \gamma\gamma$
5	$B^0 \to \eta' K_S^0$
6	$B^\pm \to DK^\pm$ $B^\pm \to D\pi^\pm$
7	$B^- \to D^0\rho^-$

Fox-Wolfram moments

Fox-Wolfram moments are rotationally-invariant parametrisations of the distribution of particles in an event. They are defined by:

$$H_l = \sum_{i,j} \frac{|p_i||p_j|}{E_{\text{event}}^2} P_l(\cos\theta_{i,j})$$

with the momenta $p_{i,j}$, the angle $\theta_{i,j}$ between them, the total energy in the event E_{event} and the Legendre Polynomials P_l.

CLEO cones

- 9 variables corresponding to the momentum flow around the thrust axis of the B candidate, binned in nine cones of 10∘ around the thrust axis

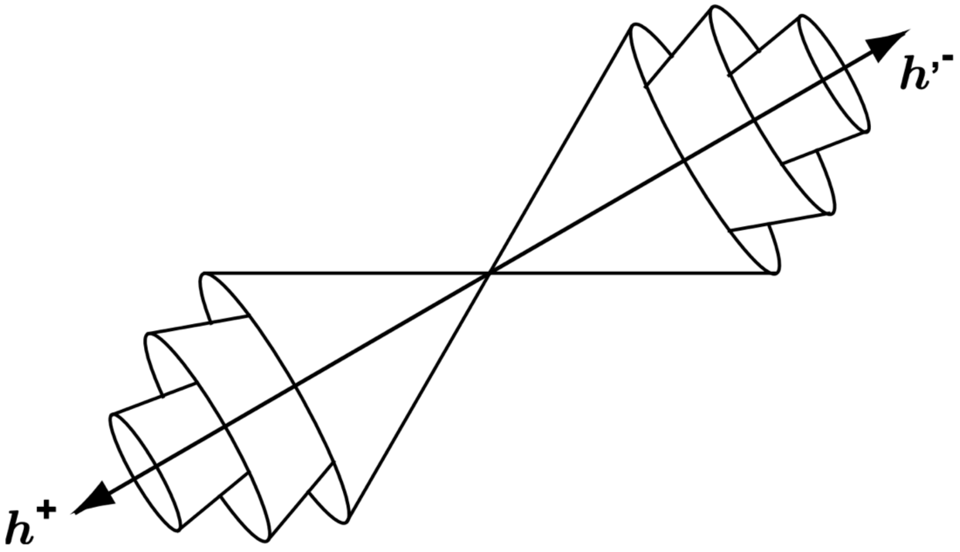


Figure 9.3.1. A graphical illustration of the CLEO Fisher discriminant, from (Asner et al., 1996). The h^+, h'^- arrows indicate the momenta of the two charged hadronic tracks in a $B^0 \to h^+h'^-$ candidate; the momentum of ROE particles within each cone (the first three cones around its thrust axis being drawn in the figure) are summed and combined to give the Fisher discriminant.

KSFW

9.5.2 *KSFW*

To further improve the continuum suppression, a second Fisher discriminant was developed by Belle:

$$KSFW = \sum_{l=0}^{4} R_l^{so} + \sum_{l=0}^{4} R_l^{oo} + \gamma \sum_{n=1}^{N_t} |(P_t)_n|, \quad (9.5.3)$$

where R_l^{so} and R_l^{oo} are modified Fox-Wolfram moments similar to h_l^{so} and h_l^{oo} in Eq. (9.5.2), respectively; the third term is the scalar sum of the transverse momentum of each particle multiplied by a free parameter γ and N_t is the total number of particles. The expressions of R_l^{so} and R_l^{oo} are described as follows:

– R_l^{so}
  In constructing R_l^{so}, the missing momentum of an event is treated as an additional particle and the moment is decomposed into three categories: a charged particle part (c), neutral particle part (n), and missing particle part (m). The variable R_l^{so} is expressed as

$$R_l^{so} = \frac{\alpha_{\mathrm{c}l} H_{\mathrm{c}l}^{so} + \alpha_{\mathrm{n}l} H_{\mathrm{n}l}^{so} + \alpha_{\mathrm{m}l} H_{\mathrm{m}l}^{so}}{E_{\mathrm{beam}}^{*} - \Delta E}. \quad (9.5.4)$$

  For odd l, we have

$$H_{\mathrm{n}l}^{so} = H_{\mathrm{m}l}^{so} = 0 \quad \text{and} \quad (9.5.5)$$

$$H_{\mathrm{c}l}^{so} = \sum_i \sum_{j\mathrm{x}} Q_i Q_{j\mathrm{x}} |p_{j\mathrm{x}}| P_l(\cos\theta_{i,j\mathrm{x}}), \quad (9.5.6)$$

  where i runs over the B daughters; $j\mathrm{x}$ indexes the ROE in the category x (x = c, n, m); Q_i and $Q_{j\mathrm{x}}$ are the charges of particle i and $j\mathrm{x}$, respectively; $p_{j\mathrm{x}}$ is the momentum of particle jx; and $P_l(\cos\theta_{i,j\mathrm{x}})$ is the l-th order Legendre polynomial of the cosine of the angle between particles i and $j\mathrm{x}$.
  For even l,

$$H_{\mathrm{x}l}^{so} = \sum_i \sum_{j\mathrm{x}} |p_{j\mathrm{x}}| P_l(\cos\theta_{i,j\mathrm{x}}), \quad (9.5.7)$$

  which is similar to Eq. (9.5.6) except for the charge factors. There are two free parameters for $l = 1, 3$ and nine (3×3) for $l = 0, 2, 4$.

– R_l^{oo}
  The definition of the second term of Eq. (9.5.3) is simpler.
  For odd l, we have

$$R_l^{oo} = \sum_j \sum_k \beta_l Q_j Q_k |p_j||p_k| P_l(\cos\theta_{j,k}), \quad (9.5.8)$$

  where j and k run over the ROE and other variables are the same as used in Eq. (9.5.6).
  For even l, we have

$$R_l^{oo} = \sum_j \sum_k \beta_l |p_j||p_k| P_l(\cos\theta_{j,k}). \quad (9.5.9)$$

FEI skims

Skim	Skim Code	Available MC Collections	Available Data Collections (362.2 fb^{-1} of on-resonance data)	Off-Resonance Data Collections
feiHadronic WITHOUT the ECL cut	11180500	All MC: /belle/collection/MC/11180500_MC15ri_noEcl (2.8 ab-1 of BB and 1 ab-1 of qqbar) Continuum only: /belle/collection/MC/11180500_MC15ri_continuum_noEcl (1 ab-1 of qqbar) Off-resonance: /belle/collection/MC/11180500_MC15ri_offres_noEcl	/belle/collection/Data/proc13prompt_skim_11180500_noEcl	/belle/collection/Data/proc13prompt_skim_11180500_noEcl_offres