Cosmogenic Rejection Studies
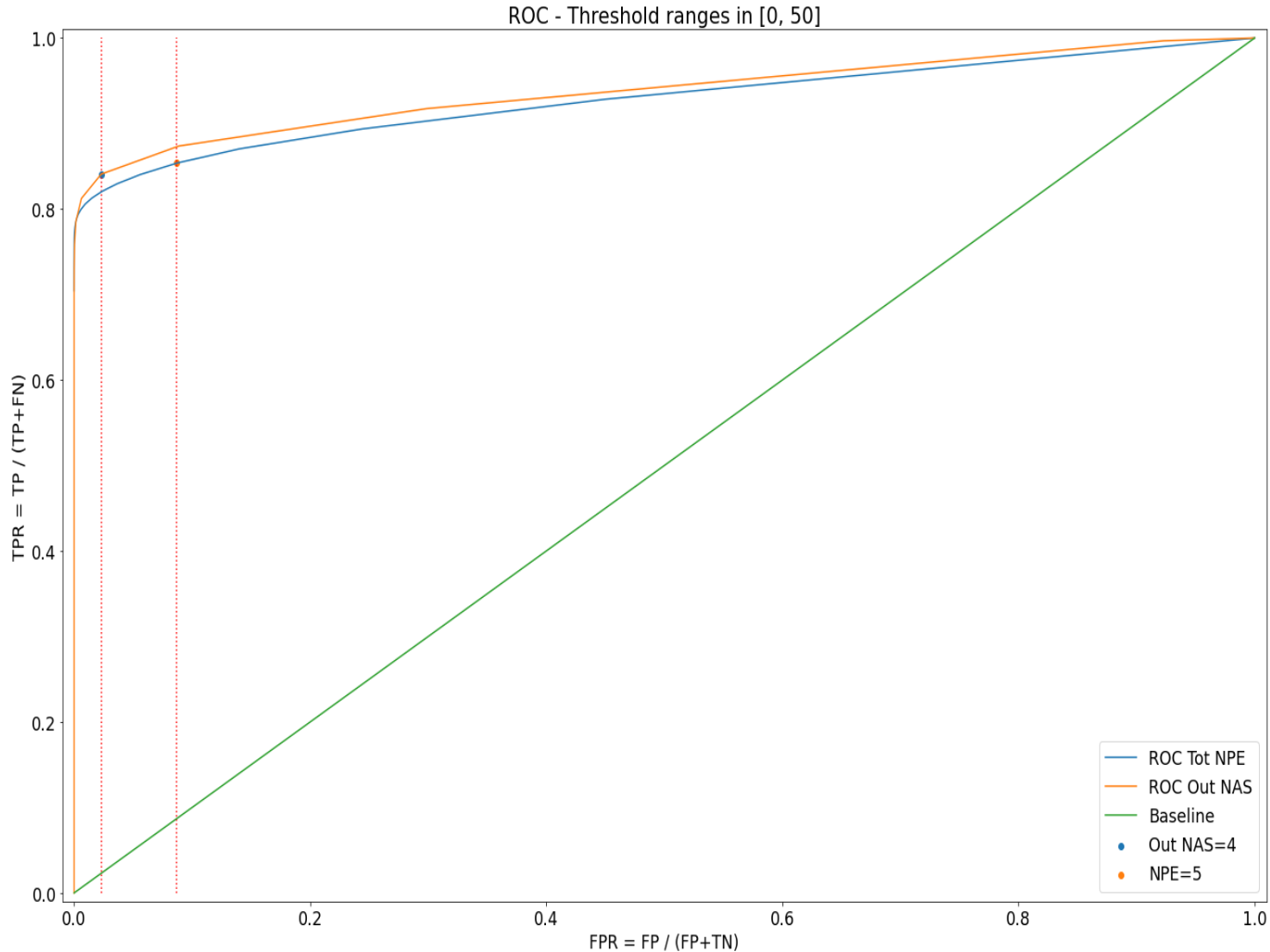
Update 11.10.2020

Recap

- Previous Approach:
  - Preliminary cut: everything < 5 PE detected => Ar39
  - Further Selection: CNN Model

- Criticities:
  - Interpretability of CNN Model: it results as a black-box
  - Feasibility of implementation: it would need a cluster

  Can we find a "simpler" model to achieve comparable performance?

Preliminary Cuts on Mu vs 1Ar39

- Comparison:
  - Cut on NPE (tot)
  - Cut on Nr Outer Slices wt >=1 PE (Nr Active Slices)

- Comparable Efficiency:
  - ~85%

- **Reduction** of **FPR** on 1 Ar39:
  - From 8.7% to 2.3%

Random Forest

- Train a RF model:
  - Signal Class: muons instances
  - Background Class: pileup of 1, ..., 5 Ar39

  - Hyperparameters:
    - 10 DTree estimators:
      - max depth = 5
      - min samples split = 100
      - min samples leaf = 100

- Why RF?
  - decision rules = if-then-else
  - interpretability = understand its logic
  - ensemble = parallelization

Performance RF on Validation Set:
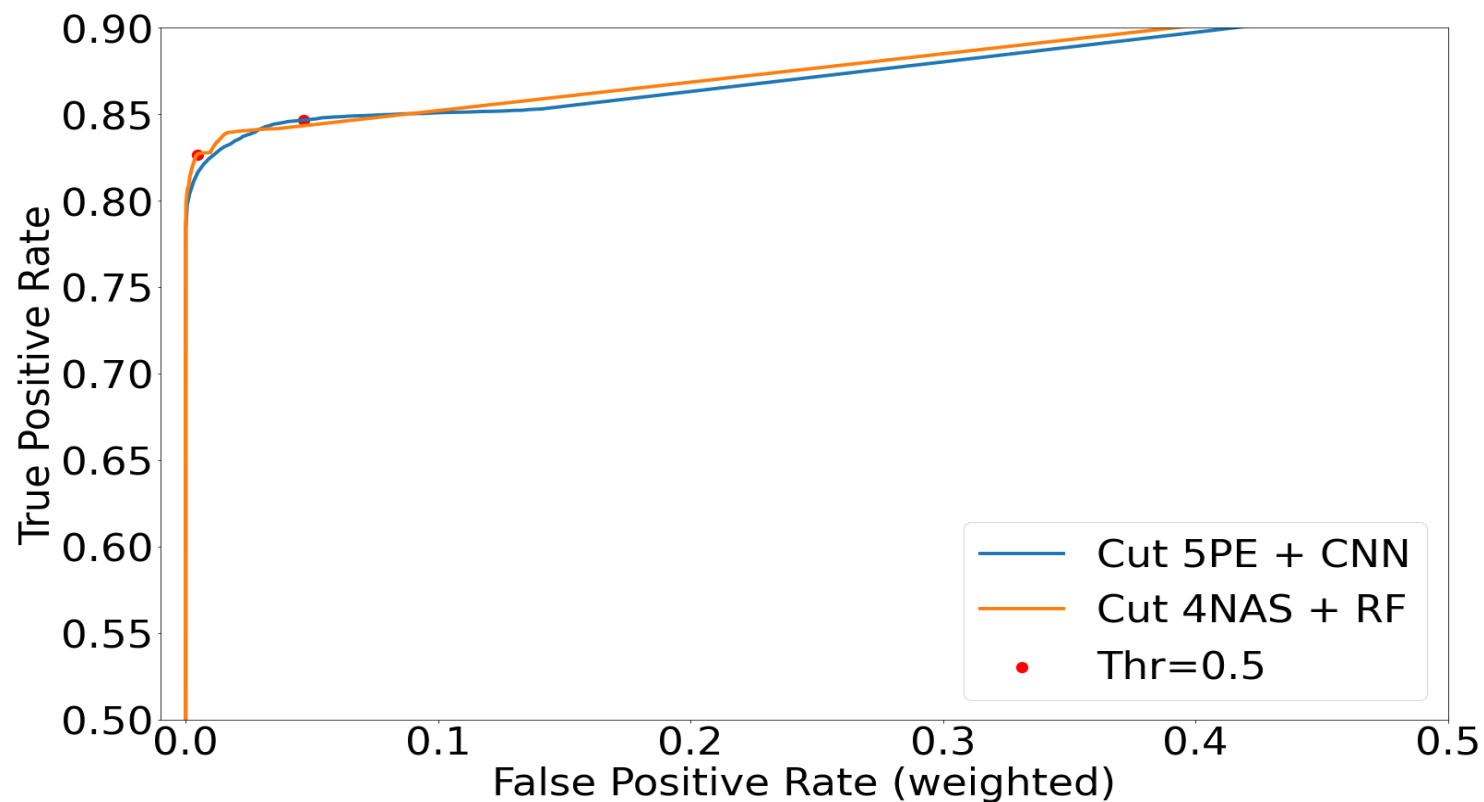Per-Class "Muon" Classification Rate
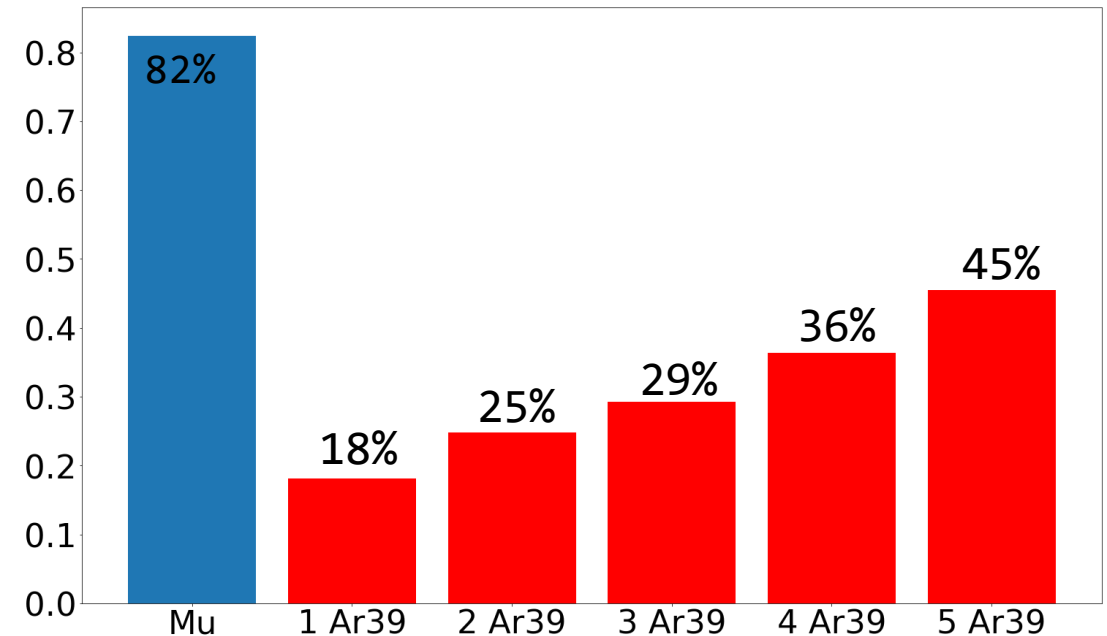
Note: we are not considering the combination with the preliminary cut

Comparison Cut4NAS+RF vs Cut5PE+CNN

- Test Set:
  - Data from previous run (07.2020)
    - 1M Muons
    - 2M Ar39, then aggregated in pileups
  - 100% unseen data

- Evaluation: TPR, FPR, ROC

- Issue: FPR includes several Ar39 classes of events with different frequency

- FPR (weighted):
  - $w_1 * FPR_1 + w_2 * FPR_2 + ... + w_5 * FPR_5$
  - $w_i = pr(i) / sum(pr(j), j=1...5)$
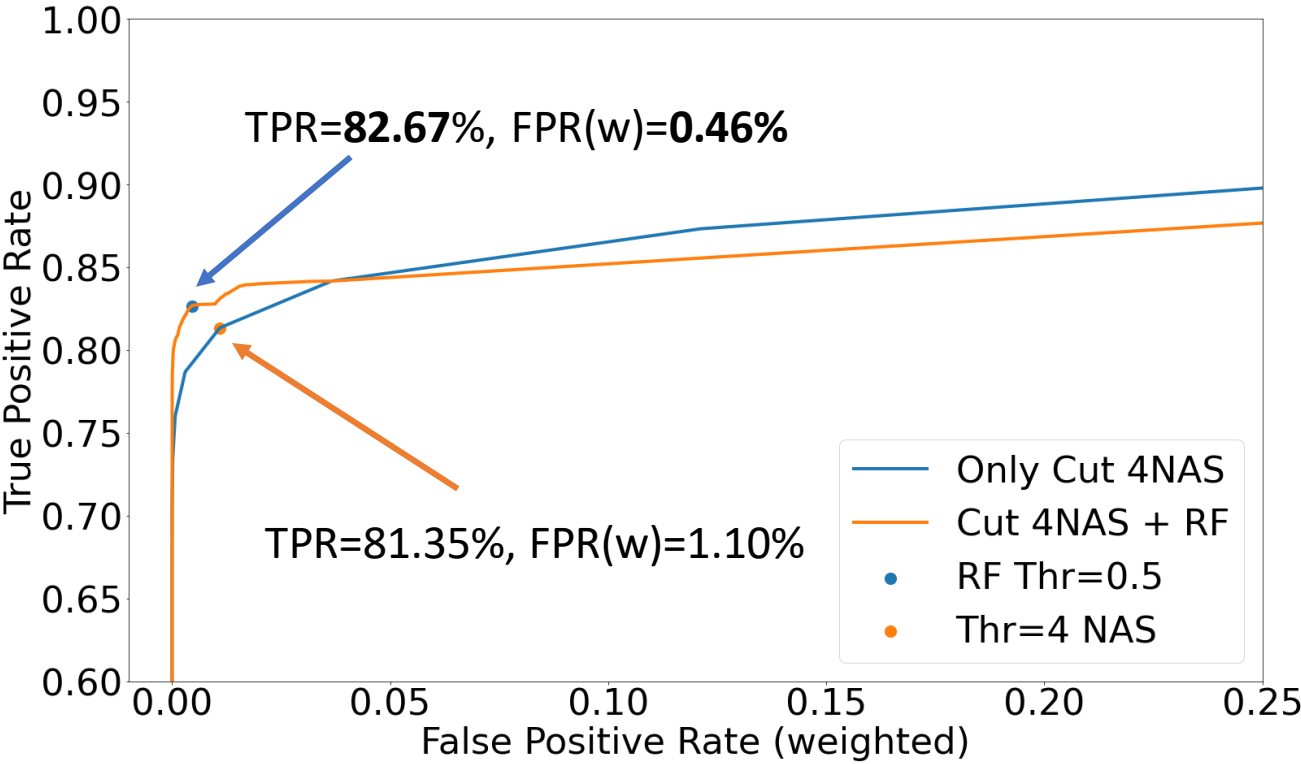 - $pr(i) = Poisson(i\ decays \mid T=10us)$

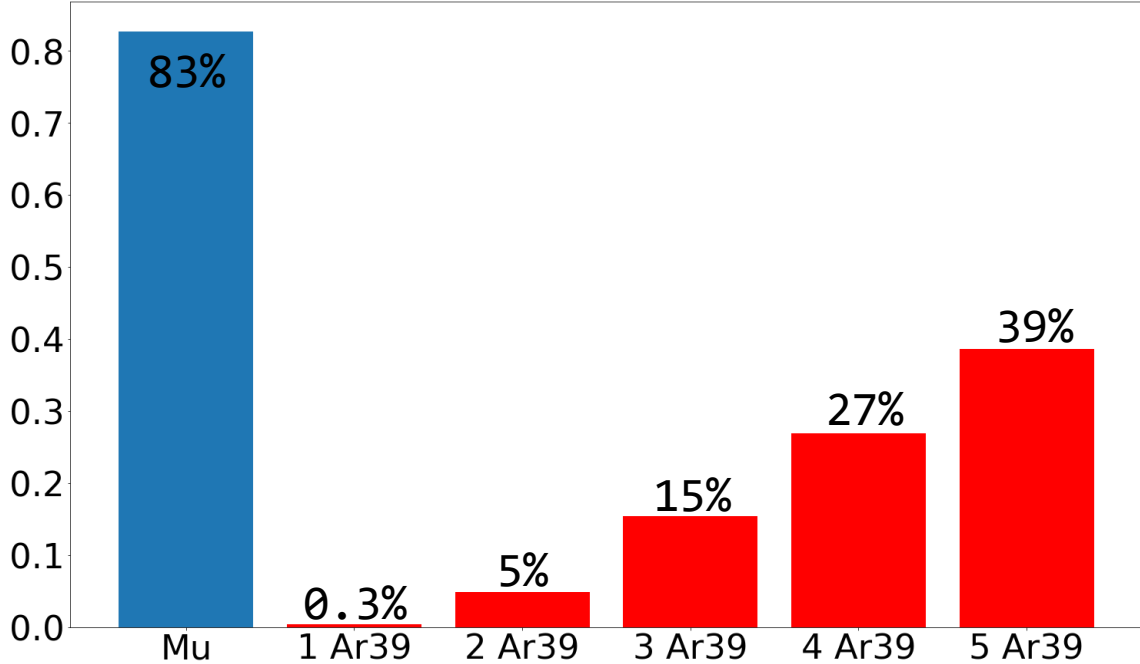# Cut4NAS+RF vs Cut5PE+CNN on Test Set

- Similar ROC Curve

- Threshold=0.50:
  - CNN+: TPR: 84.7%, FPR: 4.7%
  - RF+:   TPR: 82.7%, FPR: 0.5%

Evaluation Cut4NAS+RF on Test Set

Compare Only Cut vs Cut+RF
Test Data

TPR=**82.67**%, FPR(w)=**0.46%**

TPR=81.35%, FPR(w)=1.10%

Only Cut 4NAS
Cut 4NAS + RF
RF Thr=0.5
Thr=4 NAS

Performance Cut+RF on Test Set:
Per-Class "Muon" Classification Rate

83%   0.3%   5%   15%   27%   39%

Mu   1 Ar39   2 Ar39   3 Ar39   4 Ar39   5 Ar39

Conclusions

- The new approach results more effective w.r.t. Ar39 detection
- It is simpler to implement and its logic can be analysed

- Next Steps:
  - Optimization of the model (reduce nr features, reduce redundancy in dtrees)
  - Implementation in C code to test its performance