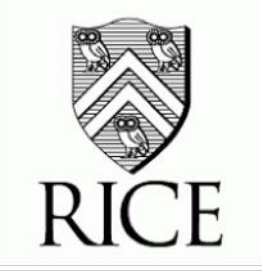



# EMTF pT Training using k-NN

Wei Shi

weishi@rice.edu

CMS@Rice

# Introduction

- Why are we doing pT training?

EMTF assign pT before global muon trigger(GMT), important for Level 1 trigger to reduce large amount of data in collisions at the LHC

Need precise pT assignment

- Why machine learning(ML)?

Muon tracks non-analytic, ML better than humans eye

- Why KNN?

Simple, intuitive method for machine learning, tuning parameters is easy

# k-nearest-neighbor(KNN)

x: input variables(dimension d) of train event
y: input variables of test event
$w_i$: weight from variable scaling
k: number of nearest neighbors
f: polynomial kernel weight function
$w_j$: weight of j in train sample
$t_j$: train target value
t: test target value

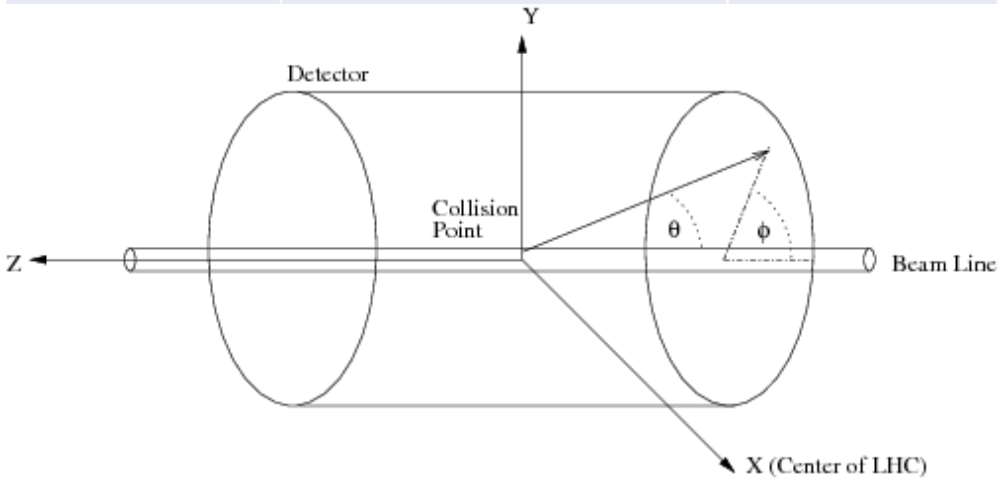
$$R_{\text{rescaled}} = \left( \sum_{i=1}^{d} \frac{1}{w_i^2} |x_i - y_i|^2 \right)^{\frac{1}{2}}$$

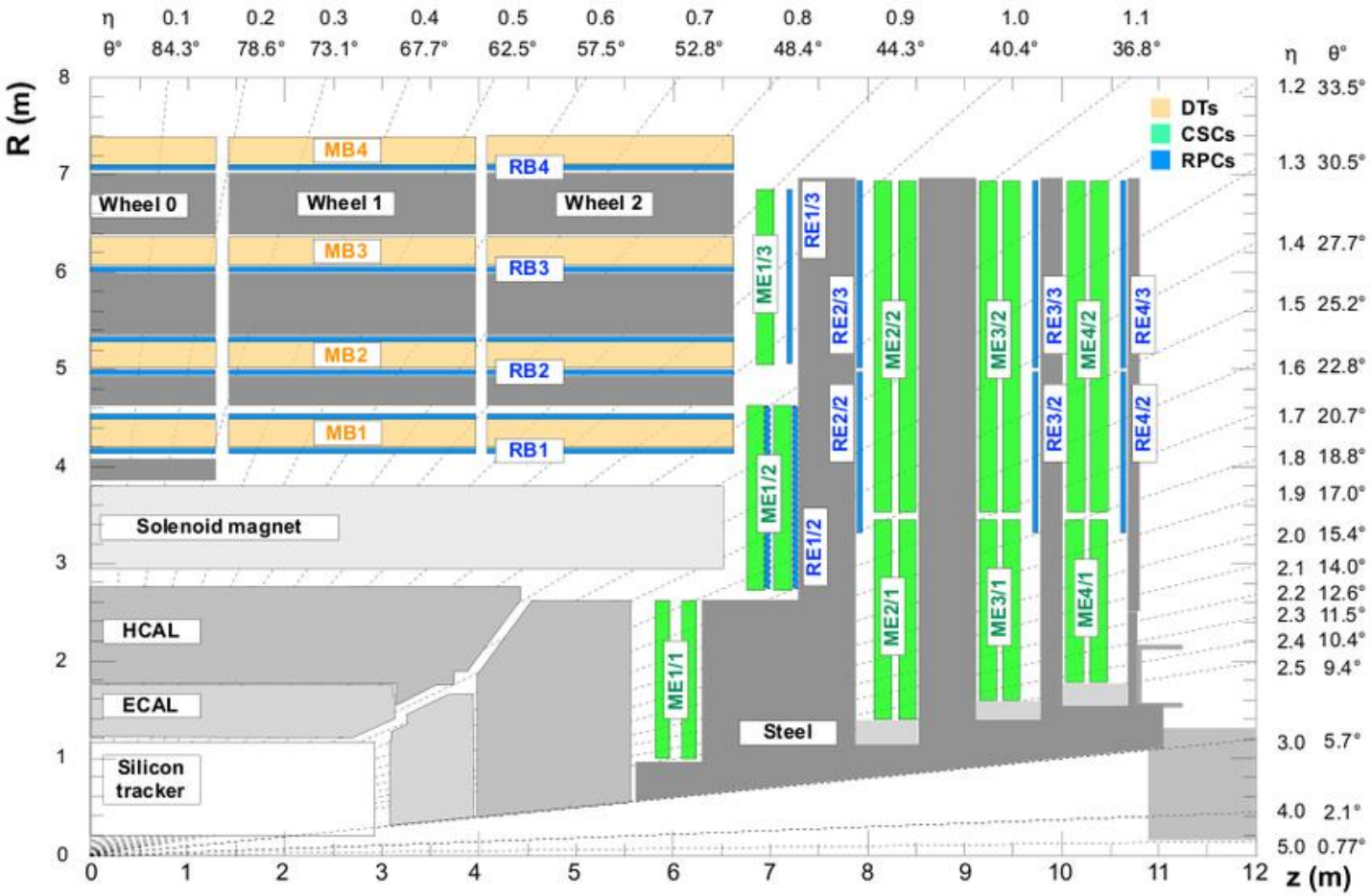
$$\langle t(i, V) \rangle = \frac{\sum_{j \in V} w_j t_j f(\text{dis}(i,j))}{\sum_{j \in V} w_j f(\text{dis}(i,j))}$$

More about k-NN:
http://cs231n.github.io/classification/#nn
kd-tree sorting:
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.160.335&rep=rep1&type=pdf

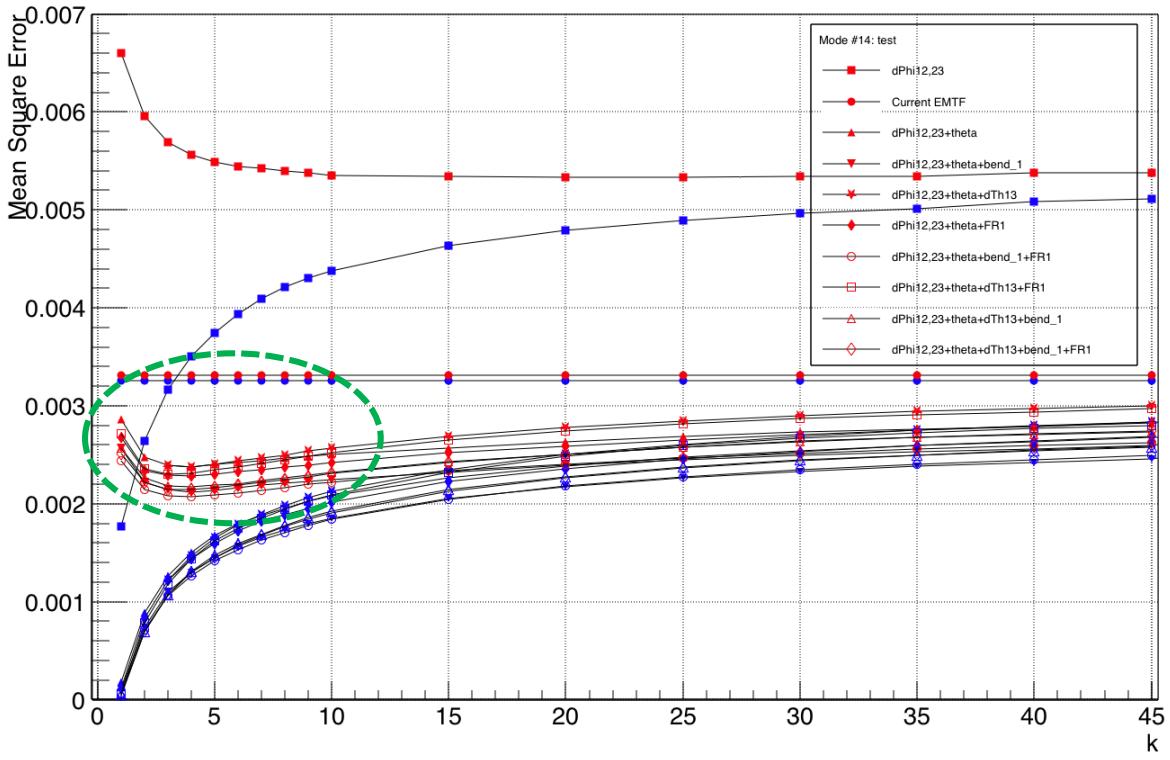
# EMTF modes

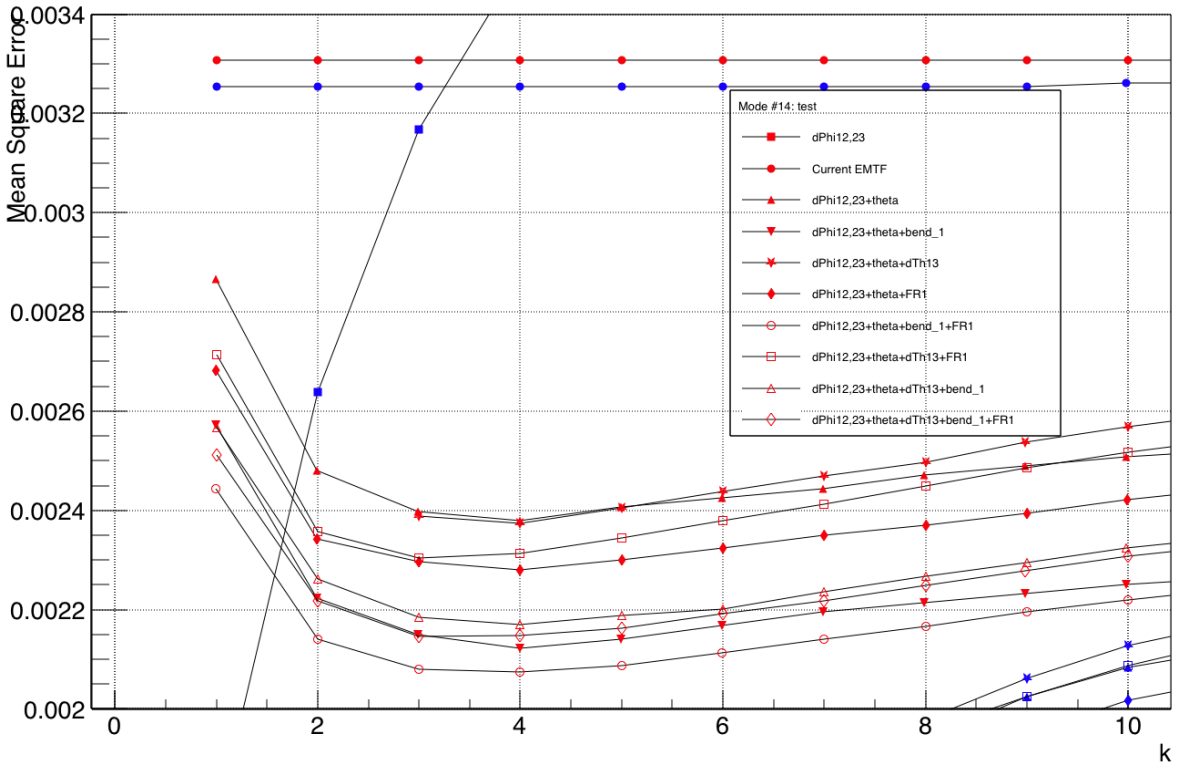
| Mode # | Definiition in code | Stations |
|---|---|---|
| 15 | 1+2+4+8 | 1,2,3,4 |
| 14 | 2+4+8 | 1,2,3 |
| 13 | 1+4+8 | 1,2,4 |
| 12 | 4+8 | 1,2 |
| 11 | 1+2+8 | 1,3,4 |
| 10 | 2+8 | 1,3 |
| 9 | 1+8 | 1,4 |



# Whole pT range, weighted 1/pT, mode 14

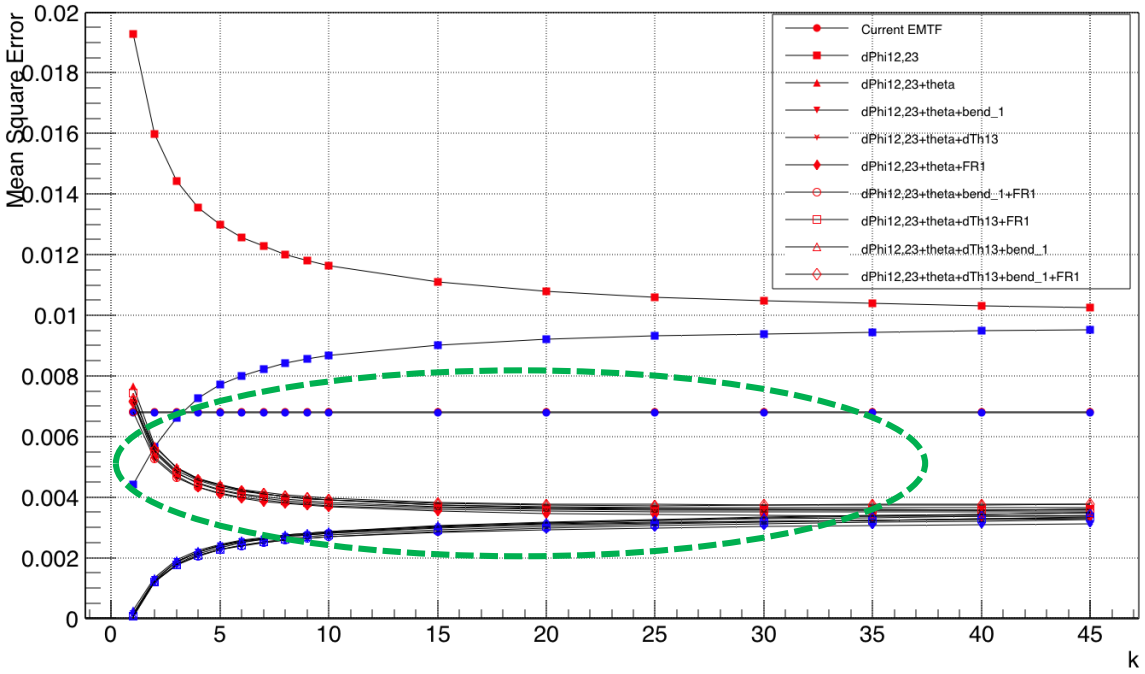


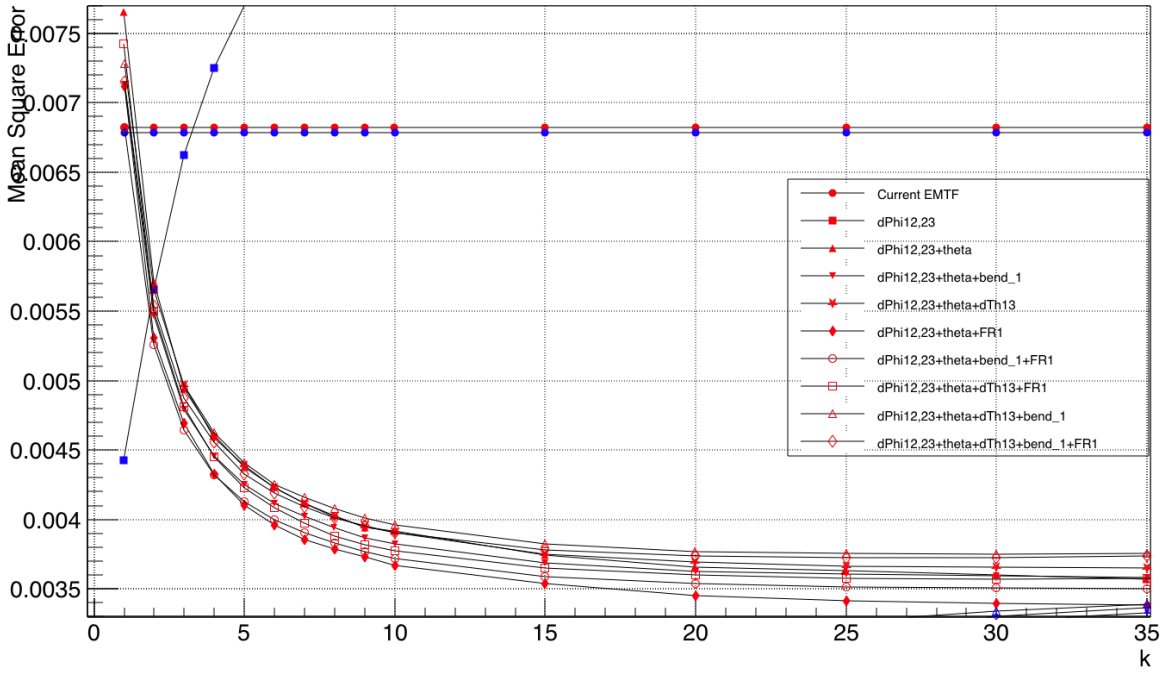




Zoom in of green dash circle in left plot

- Add theta: significant drop of Mean Square Error (MSE = (test pT – true pT)^2/# of test events)
- dPhis+theta+bend_1+FR1 give best performance on all pT range 1-1000 GeV
- Need to look at low/high pT performance separately

# pT 1-8GeV

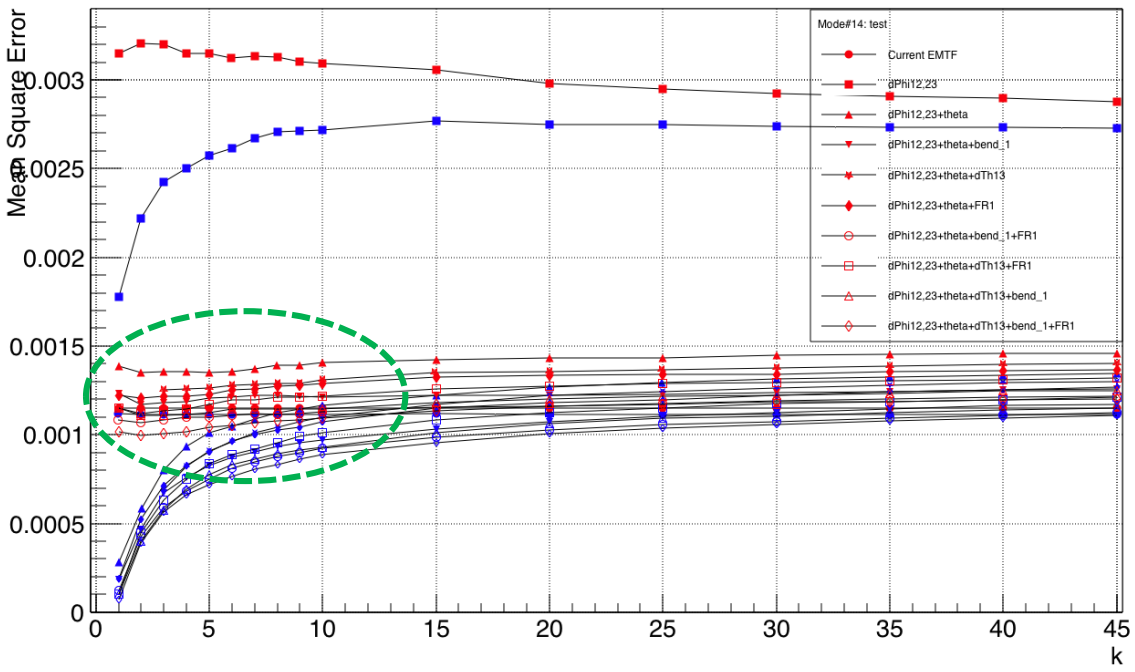


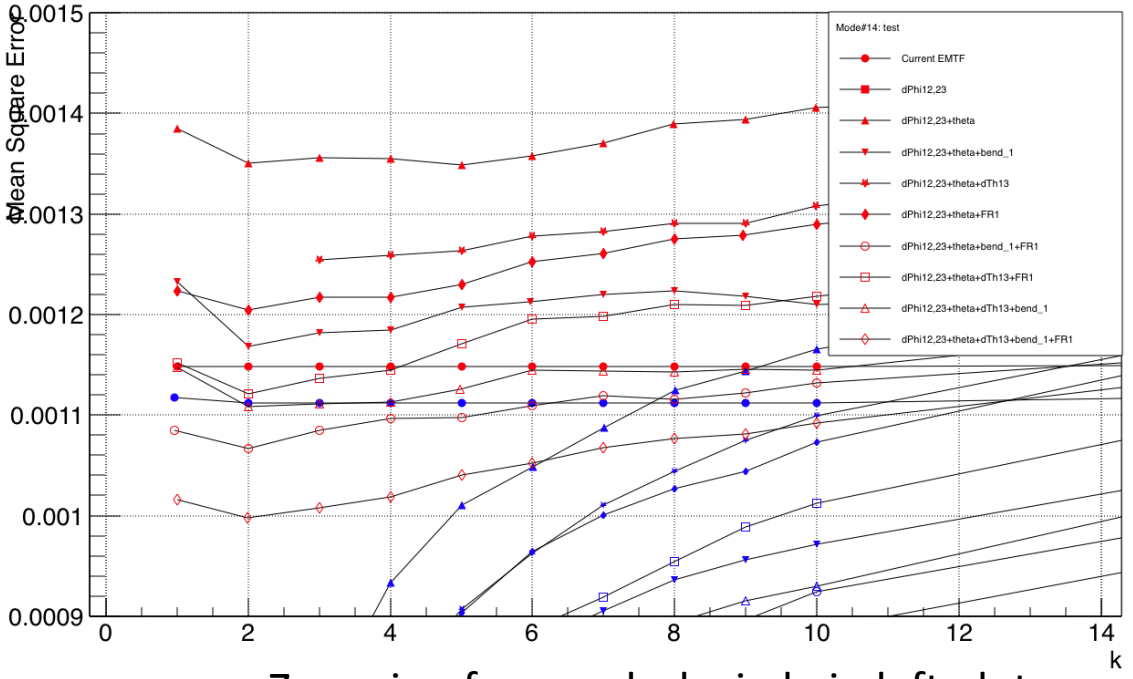




Zoom in of green dash circle in left plot

- Over fit starts when k<10, main contribution to over training in whole pT range plot
- Perform better at larger k

# pT 8-30GeV

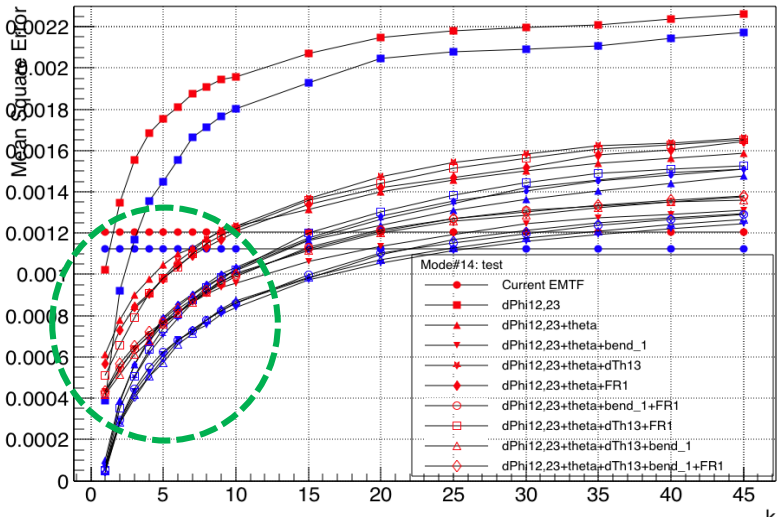


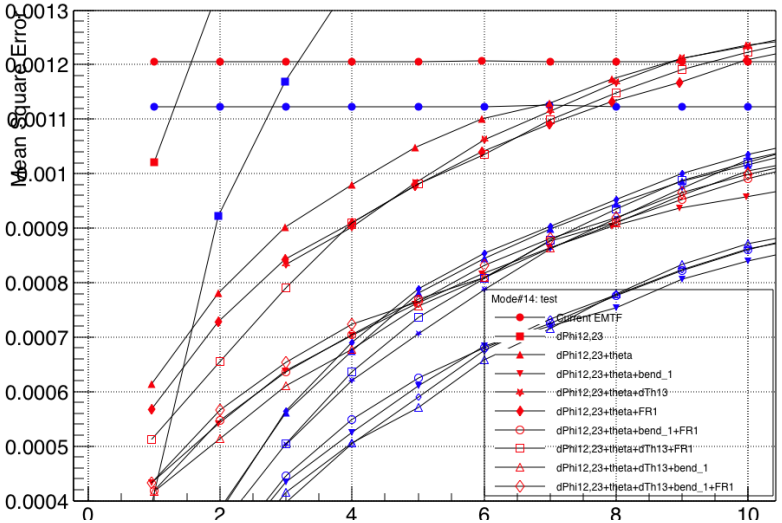



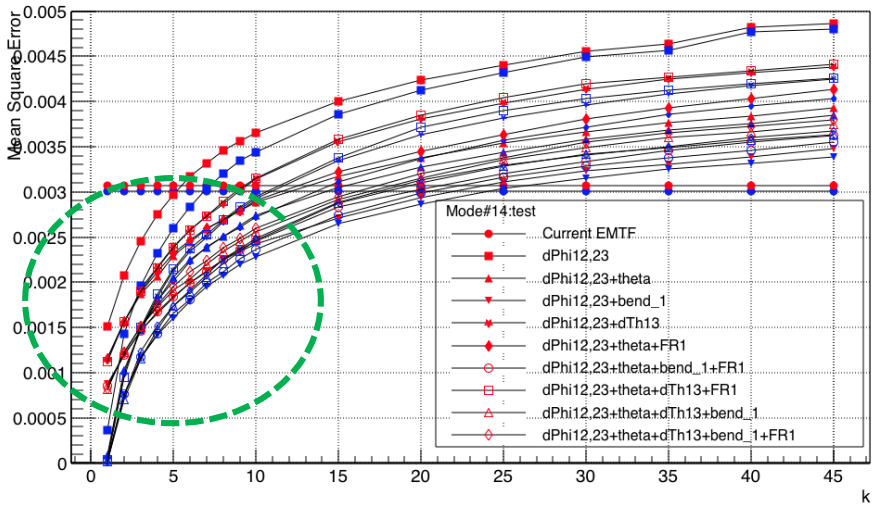

Zoom in of green dash circle in left plot

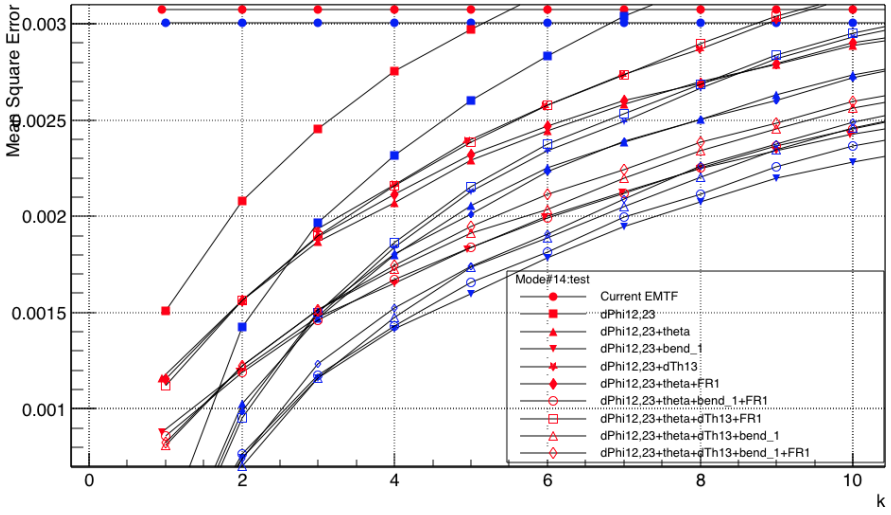
- MSE minima at very low k

# pT 30-120 GeV(upper plots), 120-1000 GeV(lower plots)



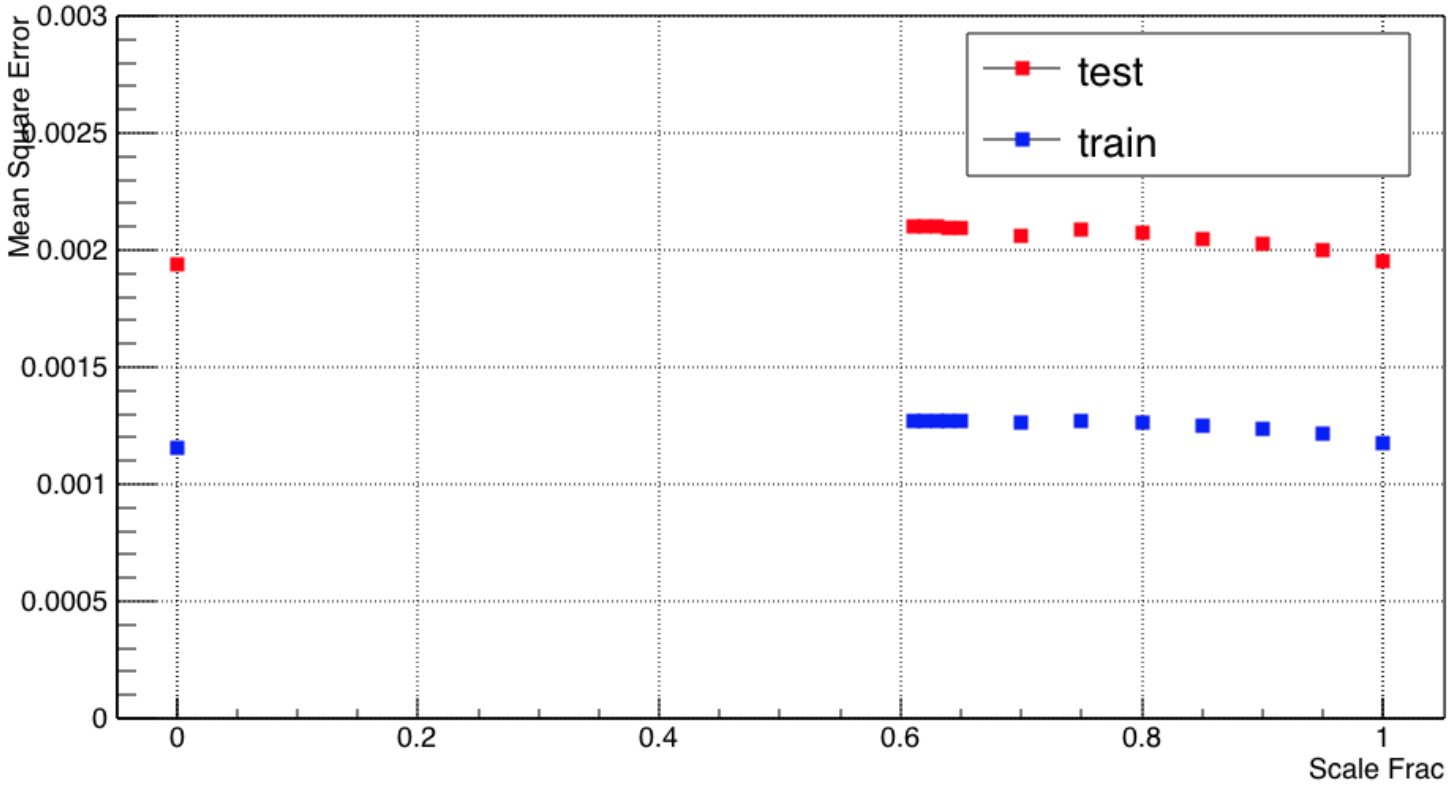




- High pT favors smaller k

# Scale input variables

- Input variables have different distribution
- e.g. θ has larger distribution than Δϕ
- Need to standardize variables when calculating distance

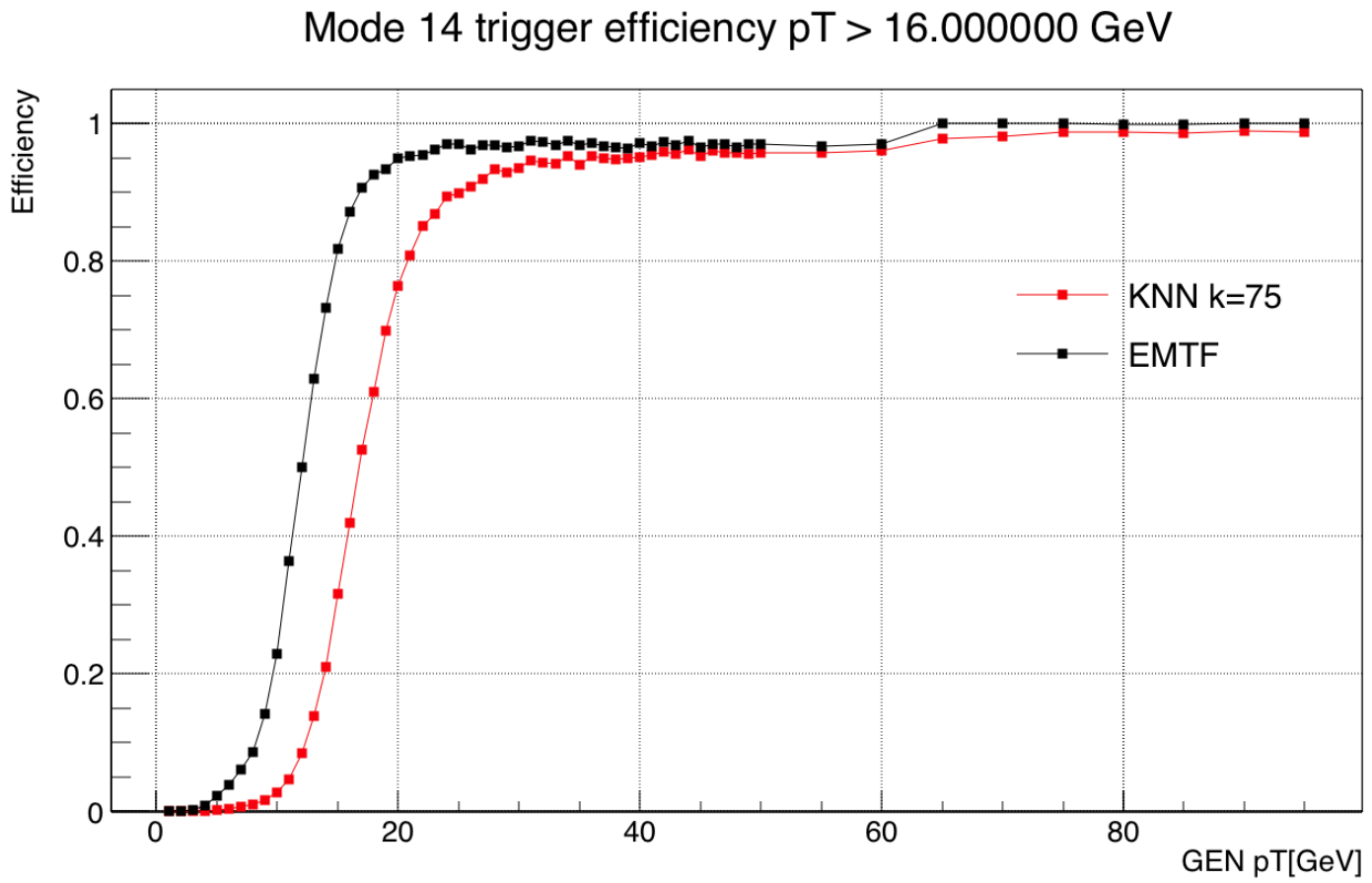



- Scale Frac = 1 gives best performance, but no big difference
- Scale Frac = 0 means turning off scale
- Scale Frac = (0, 0.61), TMVA fails

# What we know so far…

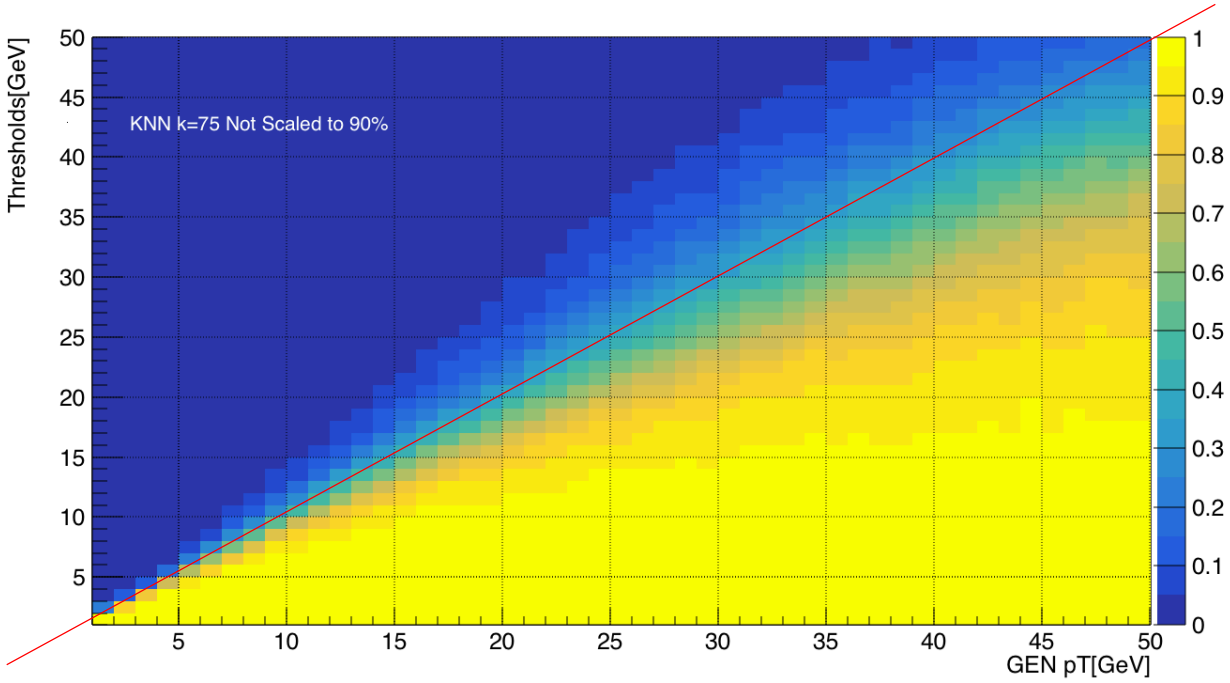
- Significant improve of MSE when adding theta
- As long as theta is added in input variables, MSE won't change too much
- Over training at small k contributed from low pT 1-8GeV, low pT favors larger k
- High pT favors smaller k
- Standardize input variables won't change performance much
- Choose dPhis+theta+bend_1+FR1, scale frac=1, vary k for new metric: efficiency/rate

# Efficiency on test sample

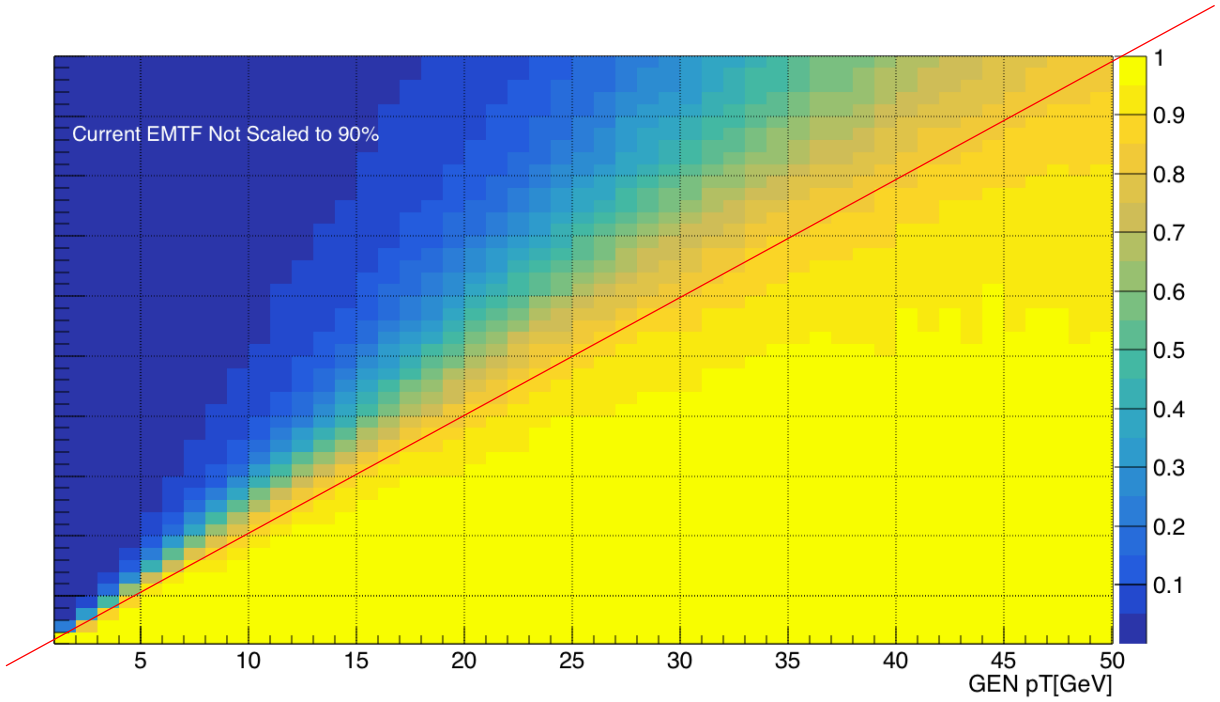



- Efficiency: muons with true pT = X (the “threshold”) are assigned a pT > X by the trigger(KNN/EMTF)

# Trigger efficiency(k=75)



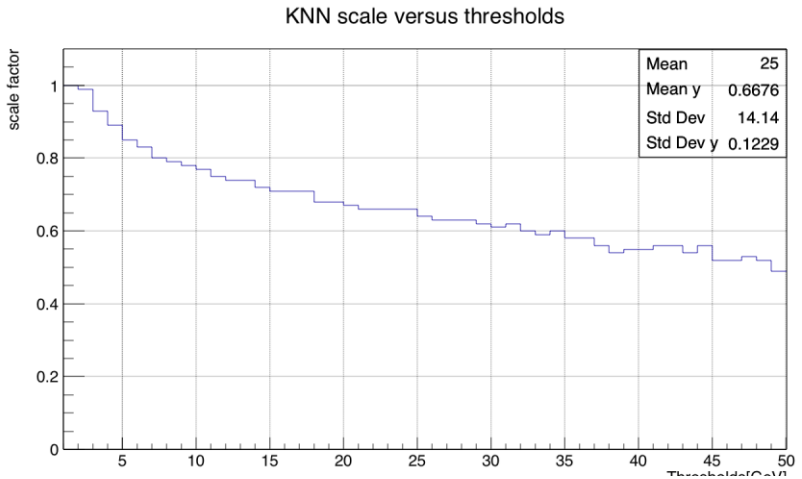

KNN (Not scaled)

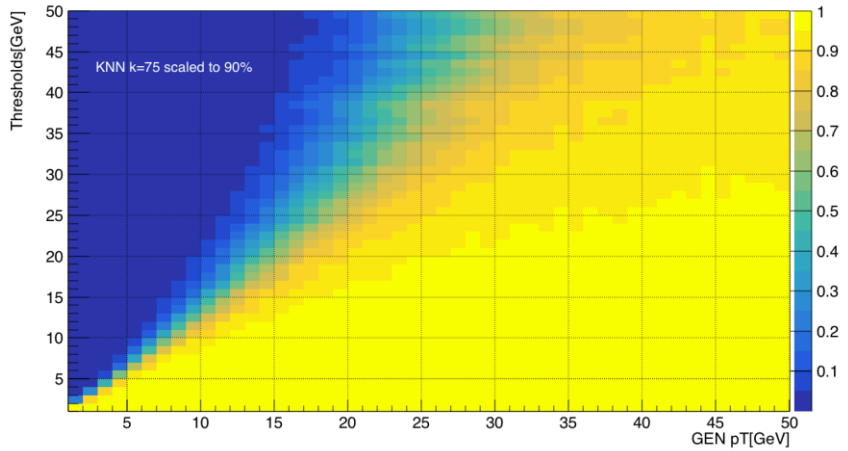



EMTF(Current was scaled to ~80%)

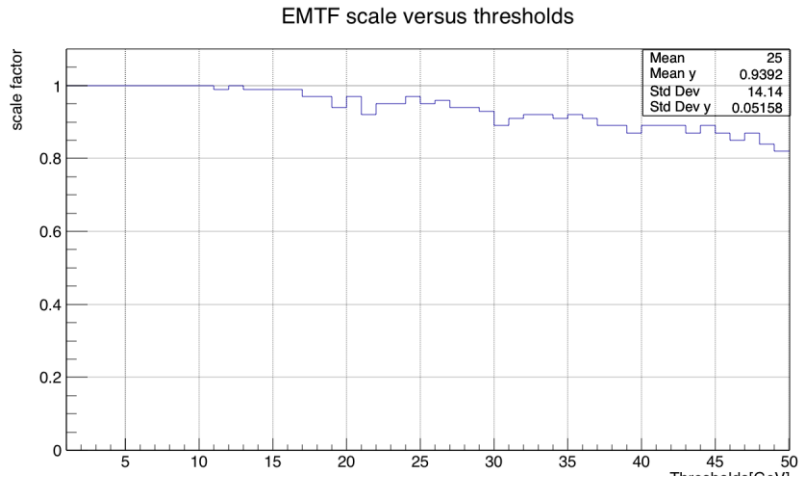
- Acceptable efficiency: 90% at pT=X, 95%-100% at pT>>X, “plateau efficiency”

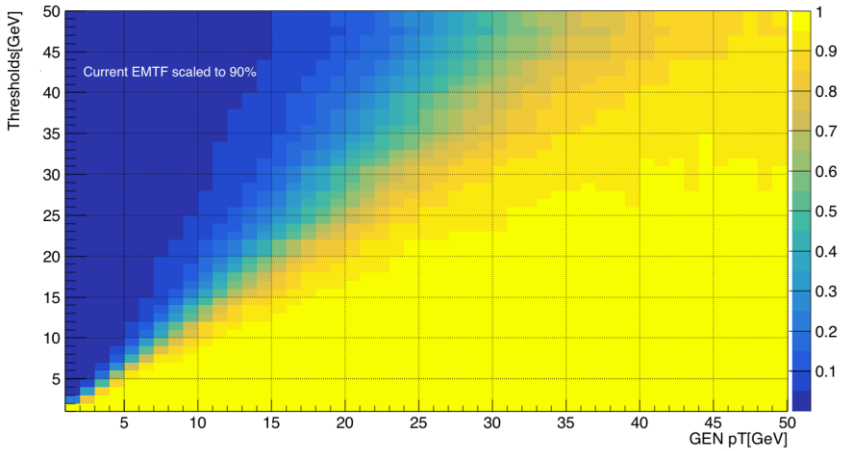
# Trigger efficiency(k=75)




KNN (Scaled)




EMTF(Scaled)

- Scale KNN efficiency to 90% at each threshold by multiplying a factor

# Training & evaluation

- SingleMu MC sample for training

Half of the muons below 32GeV, half are above

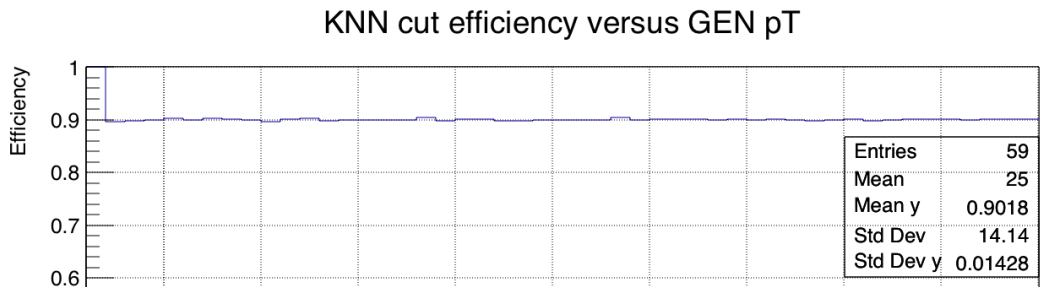
- ZeroBias dataset for evaluation

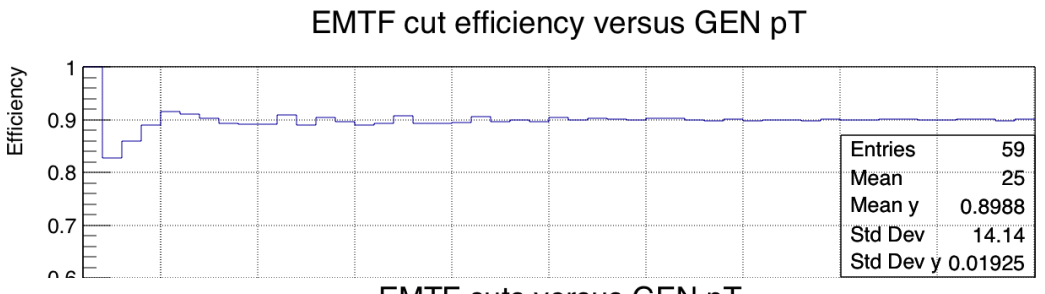
Average of the true data the trigger will actually see, for muons above 32 GeV, more like 1 - 1 million

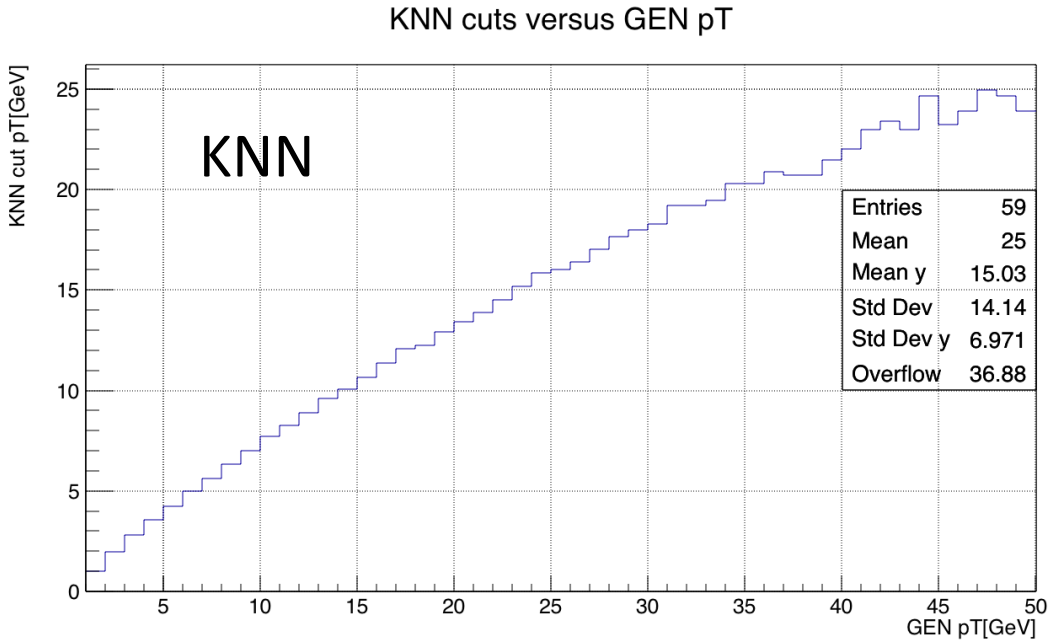
Really see hundreds muons above 32 GeV, pT assignment not good without the tracker, low pT assigned high pT

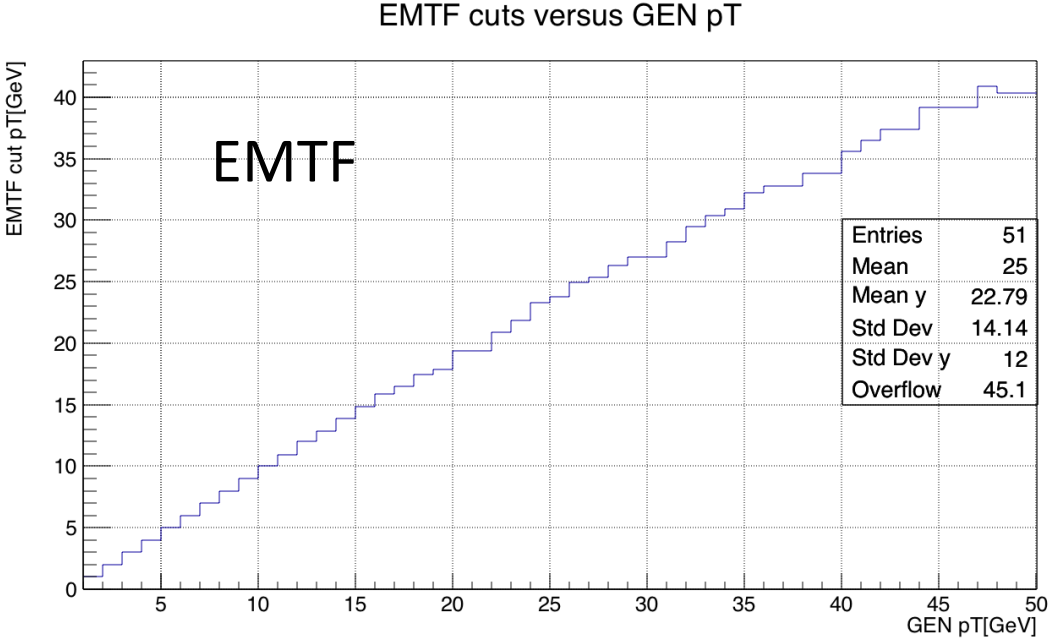
Tell us the true rate of muons for a given pT cut, the lower the better for the same efficiency

# Find 90% efficiency thresholds for rate computation



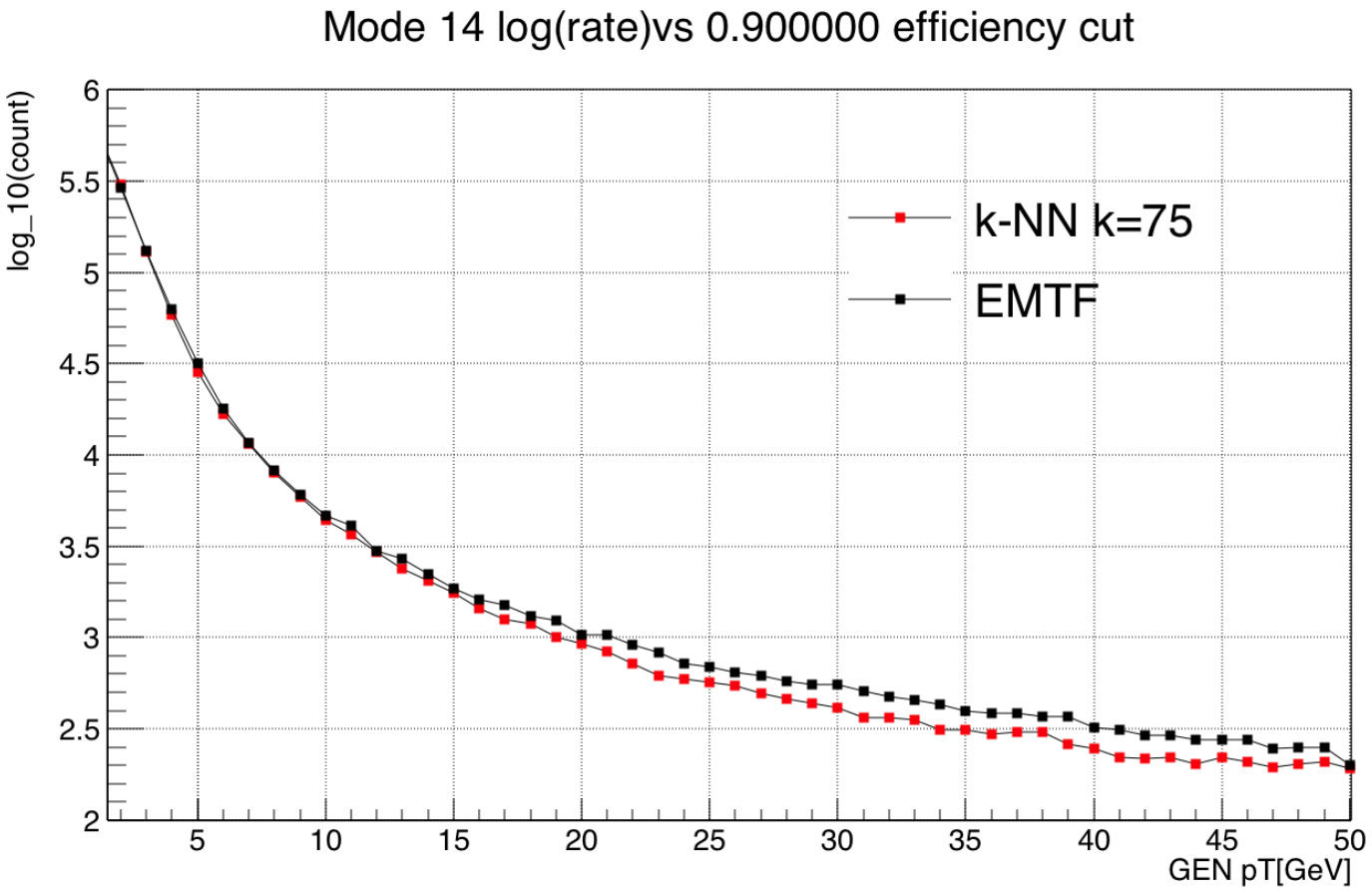




- Reduce the rate for a given threshold efficiency (90% here), without significantly lowering the plateau efficiency

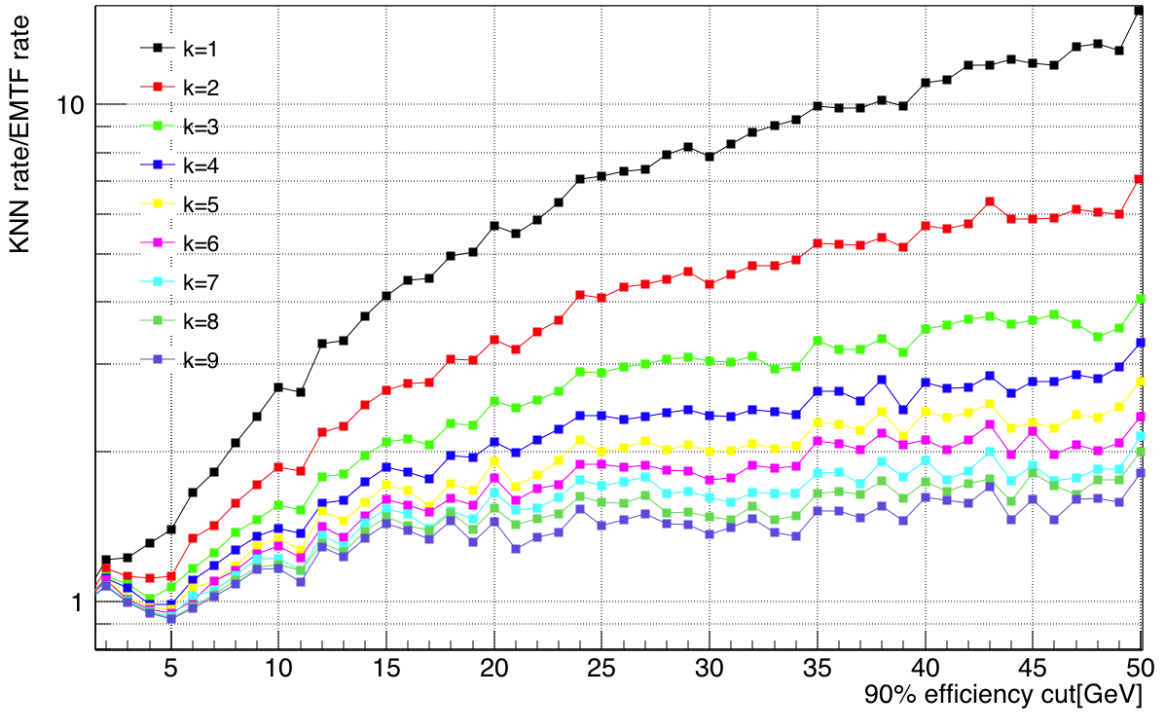
# Rate @ k=75

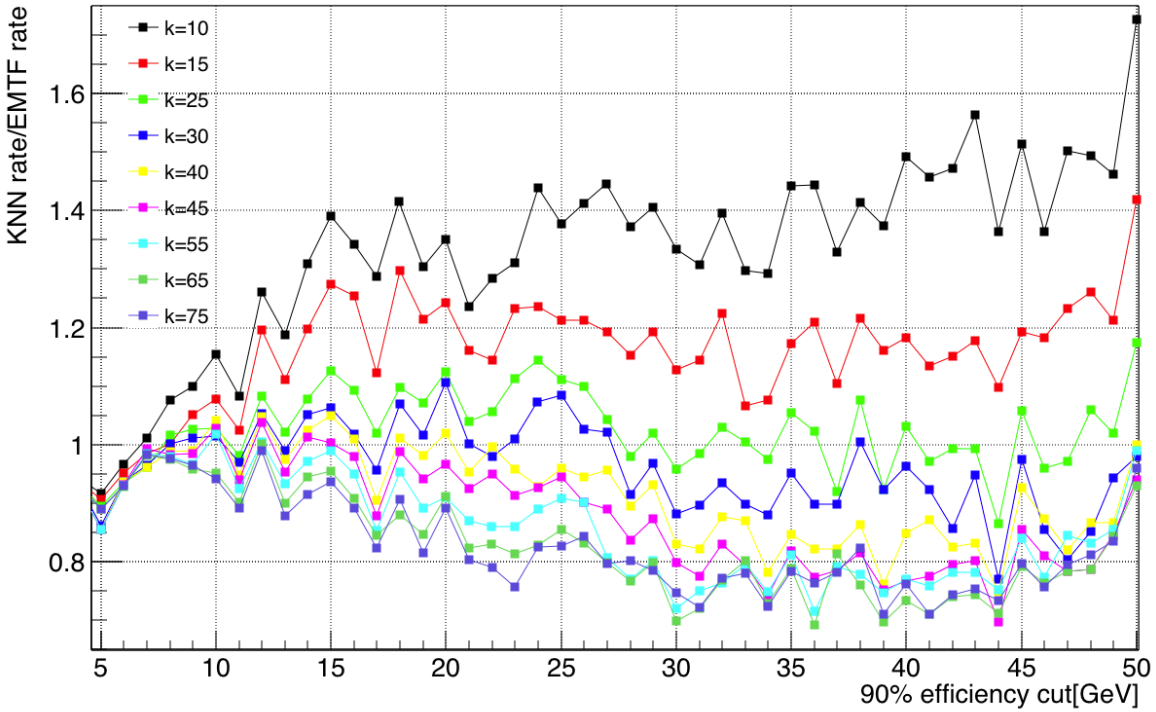



- Rate vs. pT cut at a given efficiency in ZeroBias data

For a given efficiency cut, compute the rate at various thresholds by counting ZeroBias events with pT > X

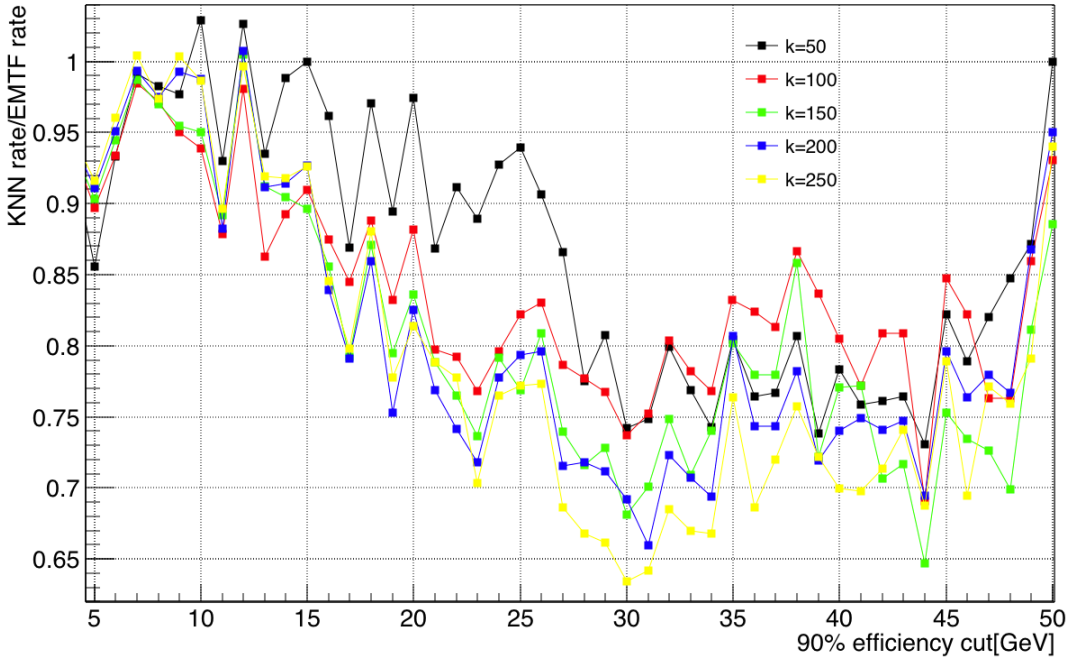
# Plot ratio: k-NN rate/EMTF rate

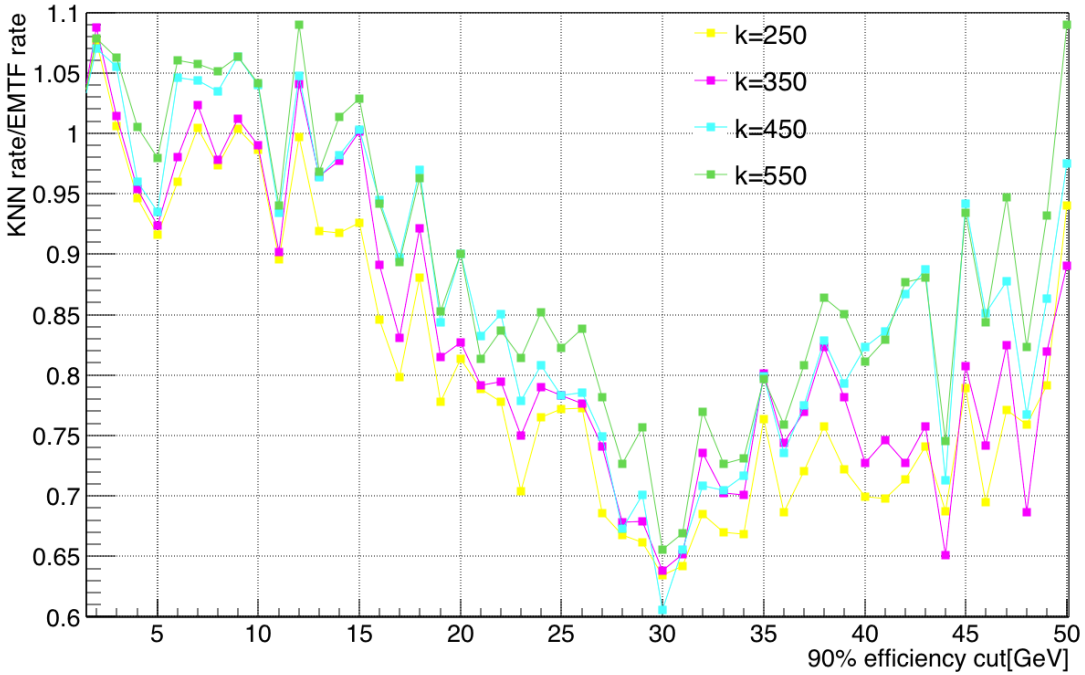



- 90% efficiency cut
- For k in [1, 75], performance goes better as k increases
- Approximate the current EMTF at k~25
- Good rate reduction performance in pT 5-50 GeV for k~75

# Large scan step: 50 and 100



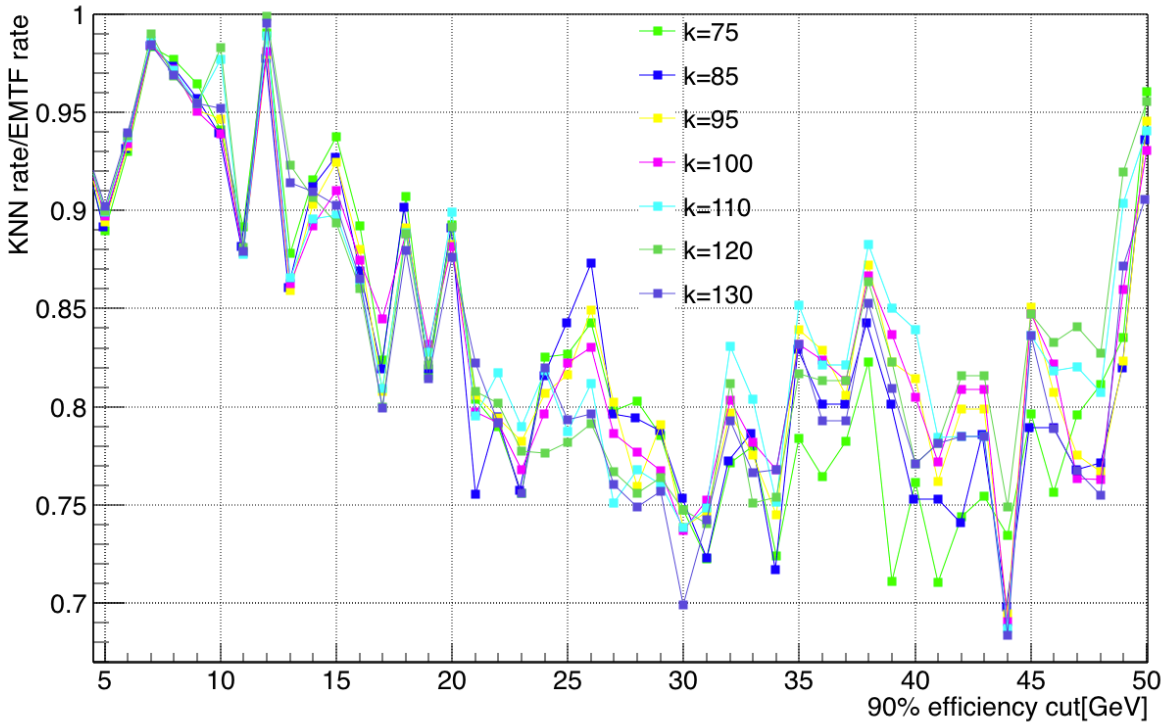



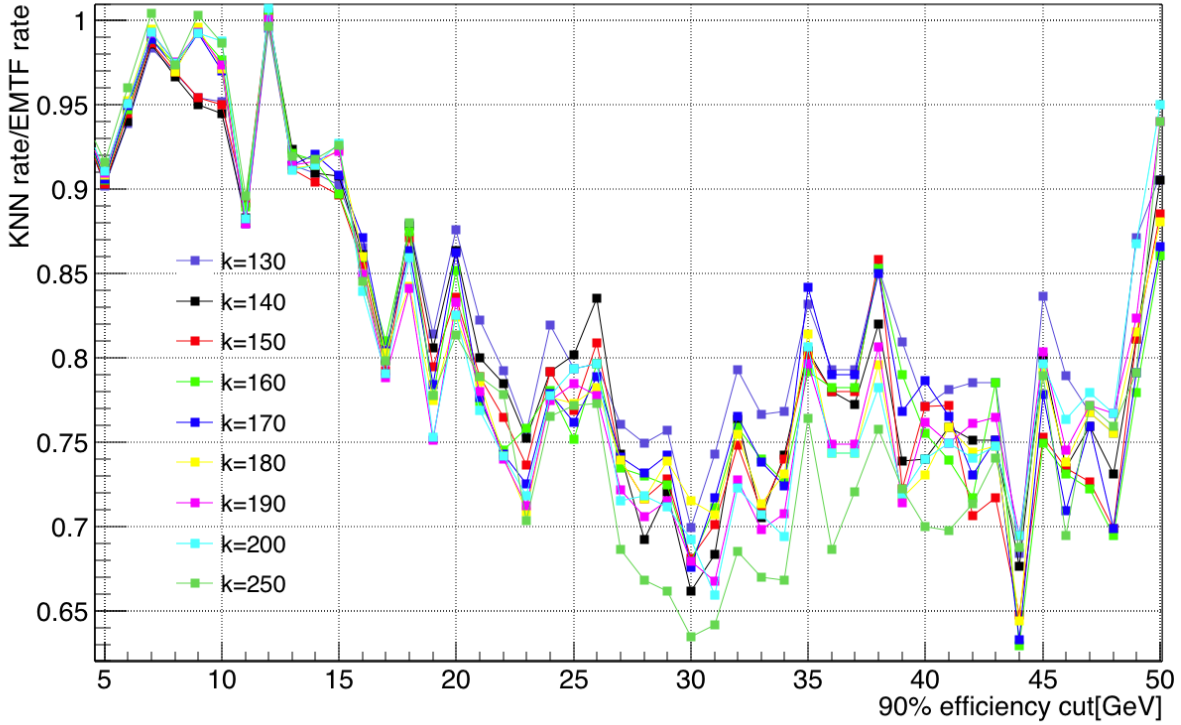

- Wide scan shows good rate reduction keeps until k~250
- Start to degrade the performance as k goes from 250 to 550
- A good stop at k~550, no need to go larger

Very long time for training/testing at k ~ O($10^3$), average running time of nearest neighbor search is O(log N)

Very large k is bad for high pT according to previous MSE results

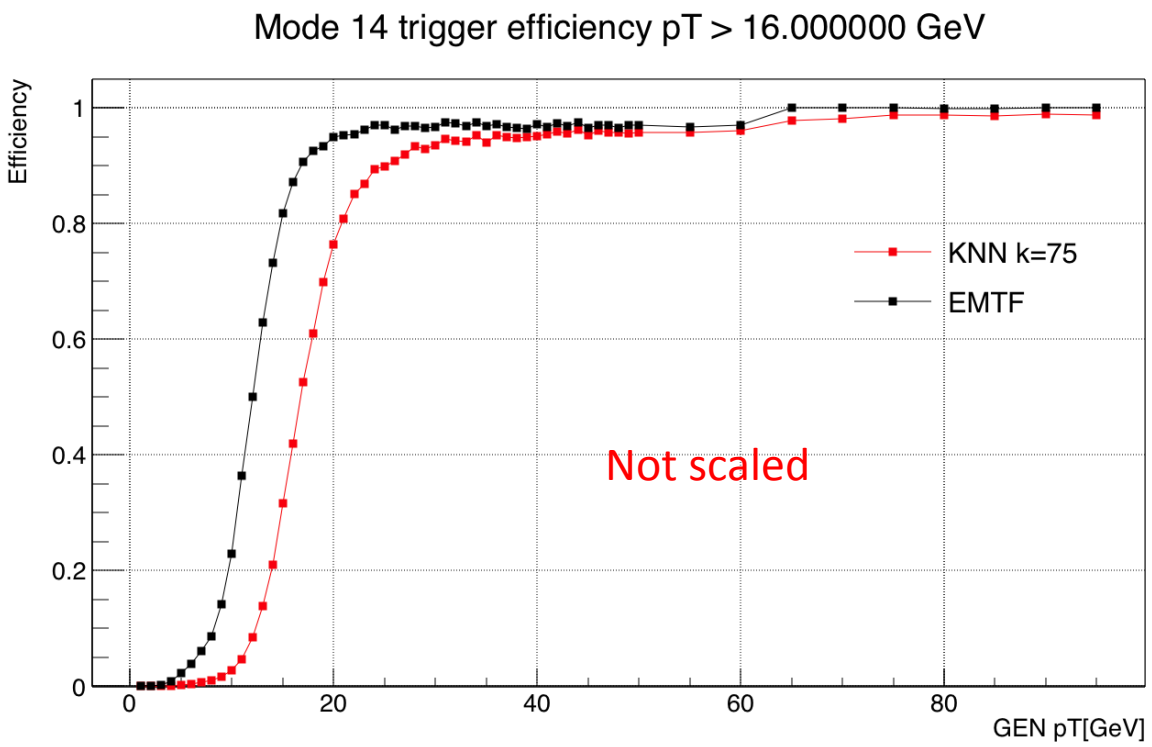
# k values between 75 and 250

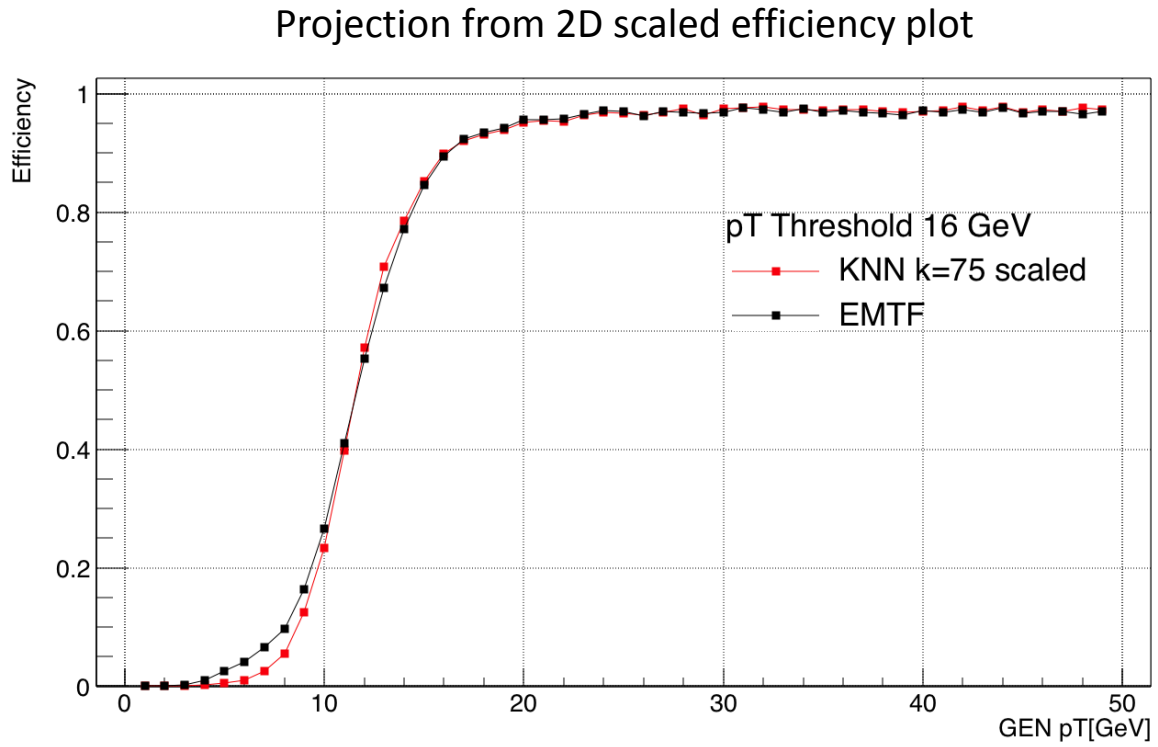



- General trend: performance is better as k goes from 75 to 250
- Best rate reduction depends on the pT cut
- No significant or overall improvement on pT range 5-50 GeV
- Recommend using k~75, rate reduction 15%-29% more than EMTF at pT 20-40 GeV

# Efficiency of 16 GeV threshold at k=75




- k=75: 90% efficiency at 16 GeV threshold, larger than 95% for pT >>16 GeV

# k-NN performance summary

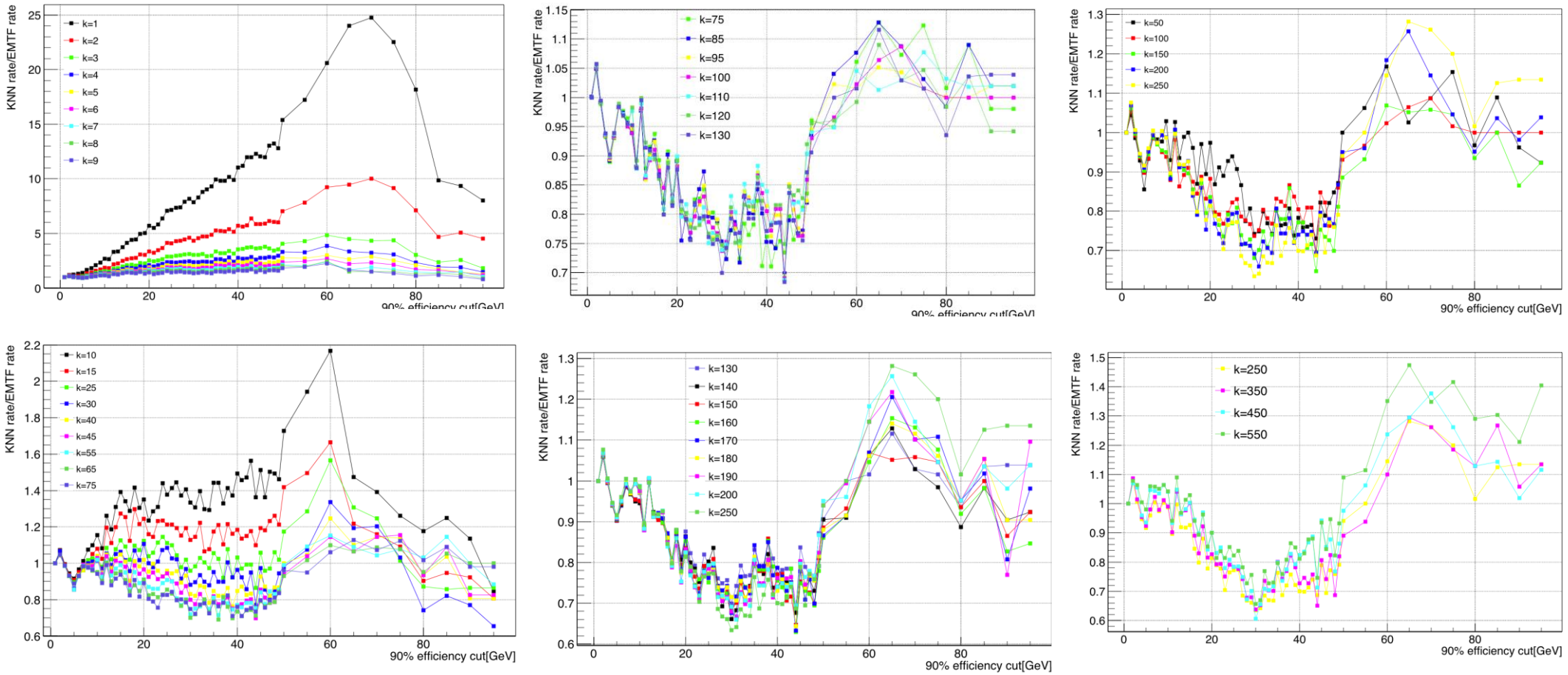
- k-NN rate reduction improves when k increases, consistent with MSE results for low pT
- k-NN rate reduction starts to degrade after k~250
- k = 75 is recommended for a good rate reduction for pT at 5-50 GeV
- k = 75 gives acceptable efficiency

Future work:

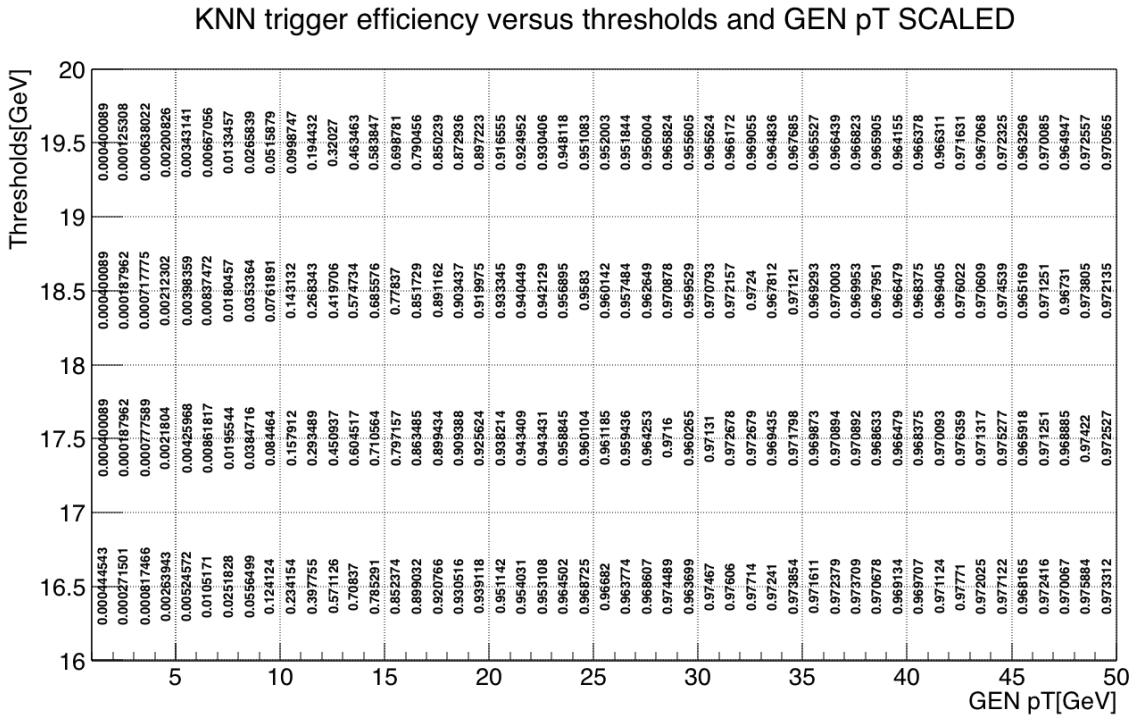
- Include RPC, truncate bits
- Compare different trigger algorithms (BDT, MLP, ANN) with k-NN using the 2017 LUT variables scheme for mode 14 (add FR2, ring # in station 1, # can be 1 or 2)
- Study other 3-station modes

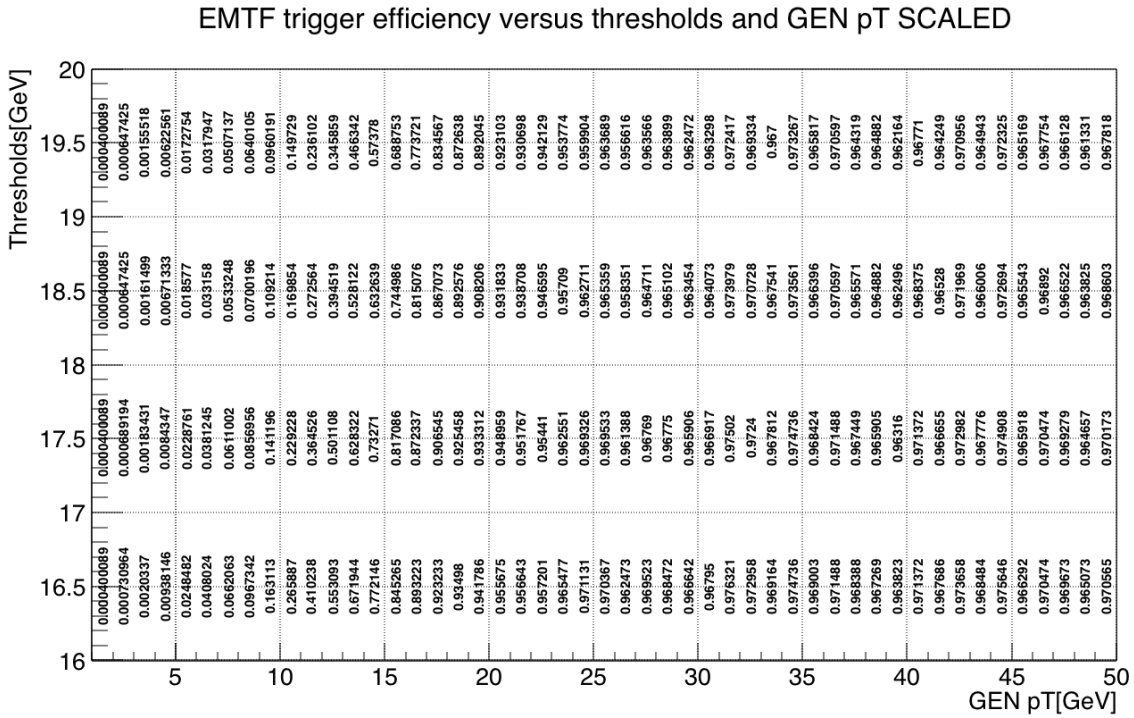
# BACK UP

# Rate tatio vs k, pT 1-99GeV



# Efficiency at k=75, threshold pT 16-20 GeV



# 2016 pT LUT based on BDT

| Track mode | Δφ | | | | | | | Δθ | | | | | | Bits |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1-2 | 1-3 | 1-4 | 2-3 | 2-4 | 3-4 | +/- | 1-2 | 1-3 | 1-4 | 2-3 | 2-4 | 3-4 | |
| 15 | 7 | | | 5 | | 6 | 2** | | | | | | | 20 |
| 14 | 7 | | | 5 | | | 2 | | 3 | | | | | 17 |
| 13 | 7 | | | | 5 | | 2 | | | 3 | | | | 17 |
| 12 | 9 | | | | | | 1 | 3 | | | | | | 13 |
| 11 | | 7 | | | | 5 | 2 | | | 3 | | | | 17 |
| 10 | | 9 | | | | | 1 | | 3 | | | | | 13 |
| 9 | | | 9 | | | | 1 | | | 3 | | | | 13 |
| 7 | | | | 7 | | 6 | 2 | | | | | 3 | | 18 |
| 6 | | | | 9 | | | 1 | | | | 3 | | | 13 |
| 5 | | | | | 9 | | 1 | | | | | 3 | | 13 |
| 3 | | | | | | 9 | 1 | | | | | | 3 | 13 |

| Track mode | Bend (CLCT) | | | | | FR | | | | θ | Md | Σ | Σ | Bits |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | +/- | 1 | 2 | 3 | 4 | | | | | |
| 15 | | | | | | 1 | | | | 5 | 4 | 10 | 20 | 30 |
| 14 | 2 | | | | 1 | 1 | | | | 5 | 4 | 13 | 17 | 30 |
| 13 | 2 | | | | 1 | 1 | | | | 5 | 4 | 13 | 17 | 30 |
| 12 | 2 | 2 | | | 2 | 1 | 1 | | | 5 | 4 | 17 | 13 | 30 |
| 11 | 2 | | | | 1 | 1 | | | | 5 | 4 | 13 | 17 | 30 |
| 10 | 2 | | 2 | | 2 | 1 | 1 | | | 5 | 4 | 17 | 13 | 30 |
| 9 | 2 | | | 2 | 2 | 1 | | | 1 | 5 | 4 | 17 | 13 | 30 |
| 7 | | 2 | | | 1 | | | | | 5 | 4 | 12 | 18 | 30 |
| 6 | | 2 | 2 | | 2 | | 1 | 1 | | 5 | 4 | 17 | 13 | 30 |
| 5 | | 2 | | 2 | 2 | | 1 | | 1 | 5 | 4 | 17 | 13 | 30 |
| 3 | | | 2 | 2 | 2 | | | 1 | 1 | 5 | 4 | 17 | 13 | 30 |

# Proposed 2017 pT LUT

| Track mode | Δφ | | | | | | | Δθ | | | | | | Bits |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1-2 | 1-3 | 1-4 | 2-3 | 2-4 | 3-4 | +/- | 1-2 | 1-3 | 1-4 | 2-3 | 2-4 | 3-4 | |
| 15 | 7 | | | 5 | | 4 | 2** | | | 2 | | | | 20 |
| 14 | 7 | | | 5 | | | 1** | | 3 | | | | | 16 |
| 13 | 7 | | | | 5 | | 1** | | | 3 | | | | 16 |
| 12 | 7 | | | | | | | 3 | | | | | | 10 |
| 11 | | 7 | | | | 5 | 1** | | | 3 | | | | 16 |
| 10 | | 7 | | | | | | | 3 | | | | | 10 |
| 9 | | | 7 | | | | | | | 3 | | | | 10 |
| 7 | | | | 7 | | 5 | 1** | | | | | 3 | | 16 |
| 6 | | | | 7 | | | | | | | 3 | | | 10 |
| 5 | | | | | 7 | | | | | | | 3 | | 10 |
| 3 | | | | | | 7 | | | | | | | 3 | 10 |

| Track mode | Bend + RPC | | | | | FR | | | | θ | Md | Σ | Σ | Bits |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | +/- | 1 | 2 | 3 | 4 | | | | | |
| 15 | 2** | 1** | 1** | 1** | | 1 | | | | 3** | 1 | 10 | 20 | 30 |
| 14 | 2 | 1** | 1** | | | 1 | 1 | | | 5 | 3 | 14 | 16 | 30 |
| 13 | 2 | 1** | | 1** | | 1 | 1 | | | 5 | 3 | 14 | 16 | 30 |
| 12 | 3 | 3 | | | | 1 | 1 | | | 5 | 7 | 20 | 10 | 30 |
| 11 | 2 | | 1** | 1** | | 1 | | 1 | | 5 | 3 | 14 | 16 | 30 |
| 10 | 3 | | 3 | | | 1 | 1 | | | 5 | 7 | 20 | 10 | 30 |
| 9 | 3 | | | 3 | | 1 | | | 1 | 5 | 7 | 20 | 10 | 30 |
| 7 | | 2 | 1** | 1** | | | 1 | | | 5 | 4 | 14 | 16 | 30 |
| 6 | | 3 | 3 | | | | 1 | 1 | | 5 | 7 | 20 | 10 | 30 |
| 5 | | 3 | | 3 | | | 1 | | 1 | 5 | 7 | 20 | 10 | 30 |
| 3 | | | 3 | 3 | | | | 1 | 1 | 5 | 7 | 20 | 10 | 30 |

# Mode 14: Station 1-2-3

Compare to last time:

- Much higher statistics(10M events, mode 14 has 787858 train and 788544 test), 42 * last time: 18746 train)
- 2016 LUT(dPhi12, 23; theta; CLCT1; dTheta13; FR 1)(mainly tune value k)
- Low/high pT divide up
- Weight 1/pT(unweight not plotted yet)
- Tune KNN scale frac
- Doing classification instead of regression?

# Steps

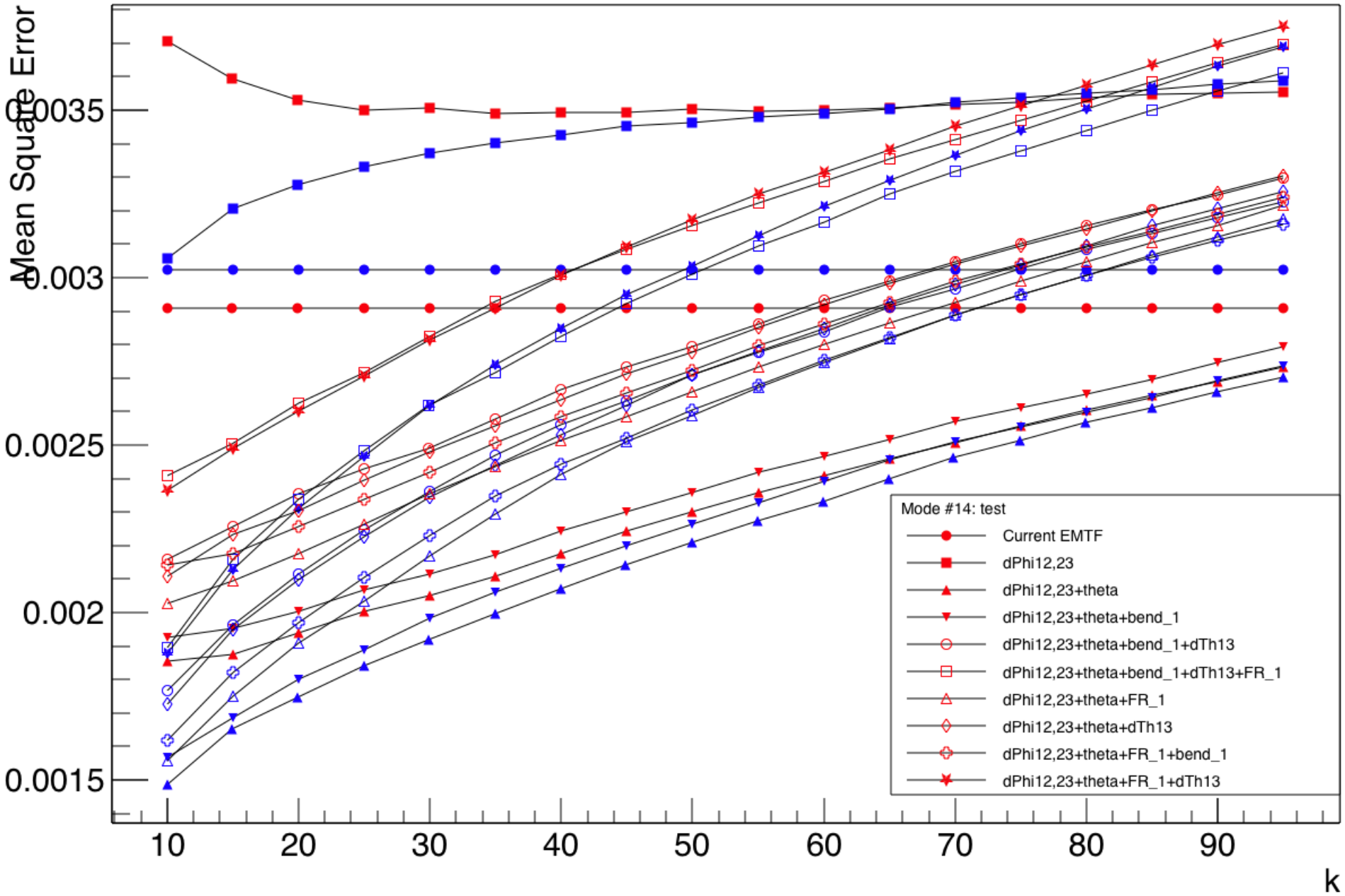
- Find optimized input/target/k/scalfrac for KNN for other modes
- Select the better method used for other modes: change inputs
- Include RPC hits and repeat 1-3 step
- Remove/truncate to fit in 29 bits
- Repeat for other modes
- Train charge assignment(classification)

# kNN

- Tune k and Scale frac
- Continuous/categorical variable: different metric(Euclidean/Hamming)
- When mixture of both kinds of variables, standardization or scale variables(called feature normalization)
- E.g. x'=(x-min)/(max-min); x'= (x-x_mean)/x_variance; assign weights to d(i,j)(TMVA adopts this)
- A non-parametric method
  Unlike other supervised learning algorithms, K-Nearest Neighbors doesn't learn an explicit mapping f from the training data
- Simply uses the training data at the test time to make predictions
- Need large dataset/cross validation dataset

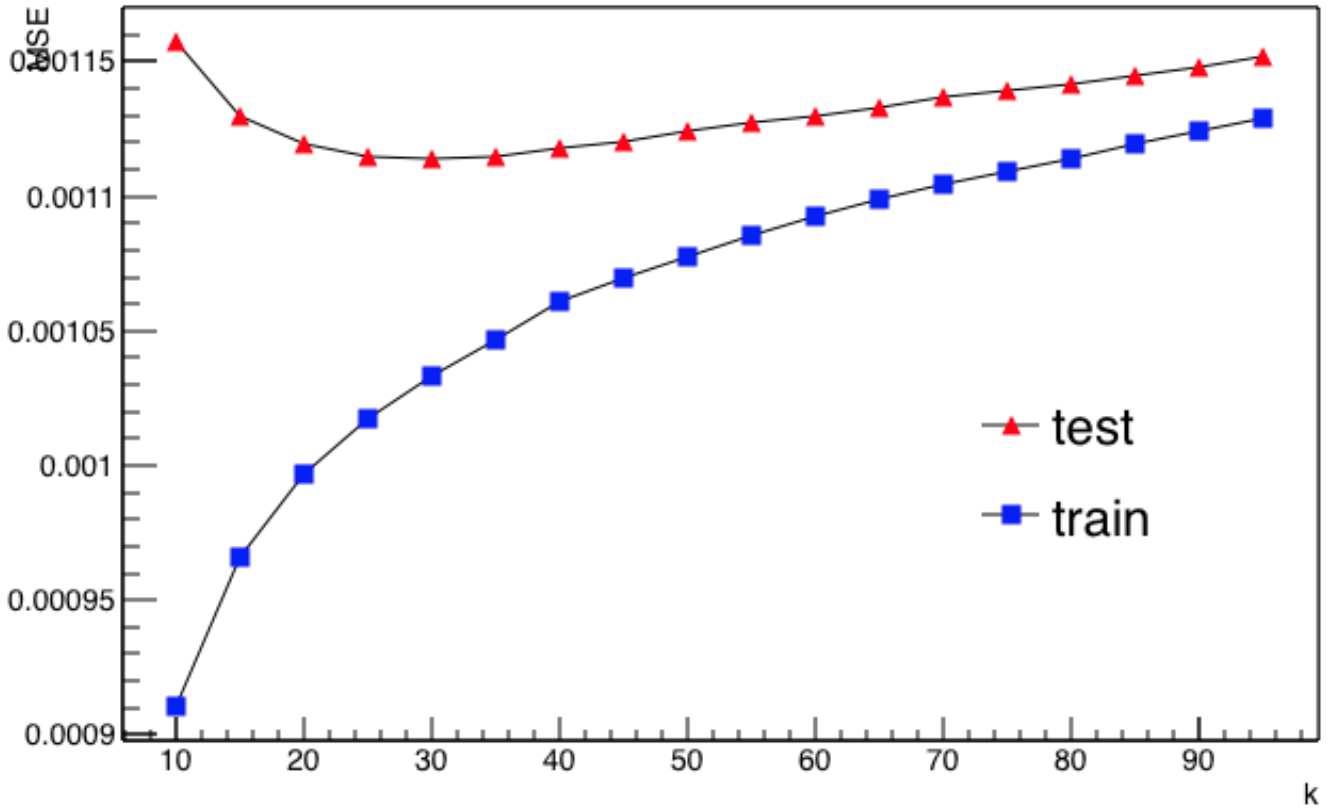
# KNN Regression

- For a test event, the algorithm finds the k-nearest neighbours using the input variables, where each training event contains a regression value. The predicted regression value for the test event is the weighted average of the regression values of the k-nearest neighbours
- The choice of the metric governs the performance of the nearest neighbour algorithm. When input variables have different units a variable that has a wider distribution contributes with a greater weight to the Euclidean metric. This feature is compensated by rescaling the variables using a scaling fraction determined by the option ScaleFrac.


Mean Square Error
0.0035
0.003
0.0025
0.002
0.0015
10
20
30
40
50
60
70
80
90
k
Mode #14: test
Current EMTF
dPhi12,23
dPhi12,23+theta
dPhi12,23+theta+bend_1
dPhi12,23+theta+bend_1+dTh13
dPhi12,23+theta+bend_1+dTh13+FR_1
dPhi12,23+theta+FR_1
dPhi12,23+theta+dTh13
dPhi12,23+theta+FR_1+bend_1
dPhi12,23+theta+FR_1+dTh13

# "nkNN=30:ScaleFrac=0.8:SigmaFact=1.0:Kernel=Gaus:UseKernel=F:UseWeight=T:!Trim"

MSE = Deviation^2/N




Mode: 15
Input variables: dphi12, dphi23, dphi34
Target variable: 1/Gen pT

/afs/cern.ch/work/a/abrinke1/public/EMTF/Analyzer/ntuples/EMTF_MC_NTuple_SingleMu_noRPC_300k.root

# Comments

- Maybe we don’t need to choose the best model, instead, we can use weighted models for final pT LUT
- Maybe we can use different model for different input variables, some model may perform better than others in certain parameter region(eta)

# BDT parameters(mode 15) from Andrew B.(check)

For mode 15

- 400 trees
- Depths: 5
- 1/pT weight(0-120 GeV), Log2(pT) is better target than 1/pT
- For very high pT, unweighted events better(>120 GeV)
- Input variables: FR bits bring significant improvement at low and high pT
- In addition to track theta, FR 1, and dPhi 1-2, 2-3, and 3-4(LUT v1), add combinations of dPhis, and ring number of station 1
- https://indico.cern.ch/event/608207/contributions/2451751/subcontributions/218758/attachments/1402616/2142649/2017_01_26_Mode_15_BDT.pdf

# BDT good/bad

- Good: little tuning required(simple)

- Bad:

1. will ignore non-discriminating variables as for each node splitting only the best discriminating variable is used
2. theoretically best performance on a given problem is generally inferior to other techniques like neural networks.

- See TMVA tutorial