2019 Data Processing

Norman Graf (SLAC)
HPS Summer Collaboration Meeting
June 25, 2021

Overview

- Detector performance with current software
  - Energy/Momentum Calibration
    - FEE
    - WAB
    - Three-prong Tridents
  - Tracking Performance
    - Track-finding efficiency
    - Momentum Reconstruction
- Data Reconstruction
  - Currently reconstructed data
  - Plans for the “Unblinded” sample
  - Plans for the full 2019 dataset

Je n’ai fait celle-ci plus longue que parce que je n’ai pas eu le loisir de la faire plus courte.

Blaise Pascal

Detector Calibration with Data

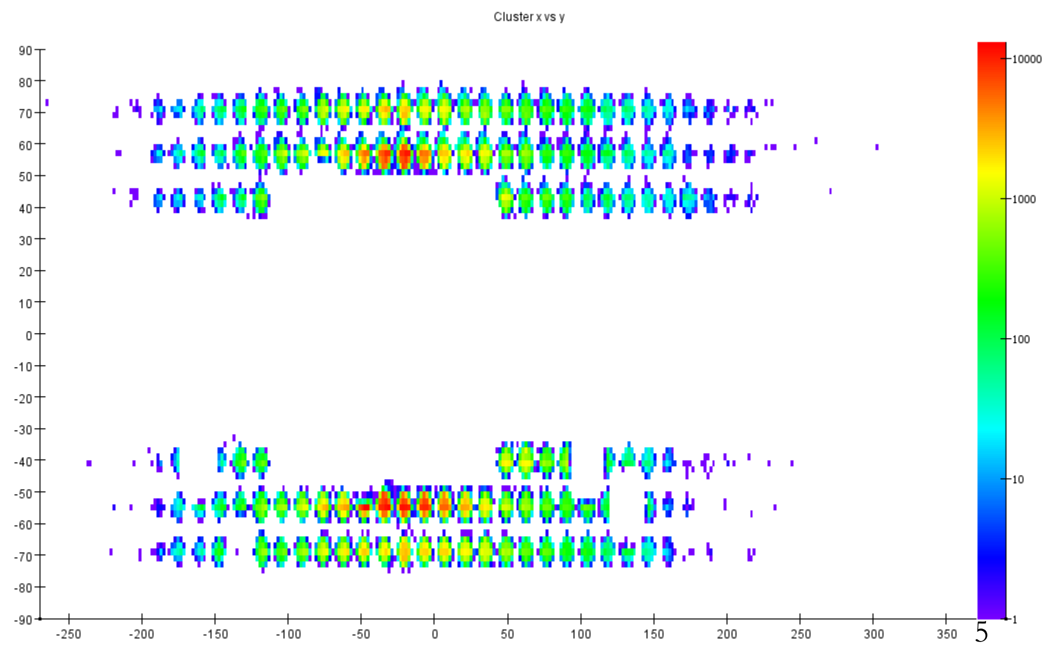
- Data Samples
  - Dedicated FEE runs 10103 & 10104
  - “Sample Partitions”
- Energy Calibration of the Ecal
- Momentum Calibration of the SVT
- Mass Calibration of the HPS detector

Energy Calibration of the Ecal

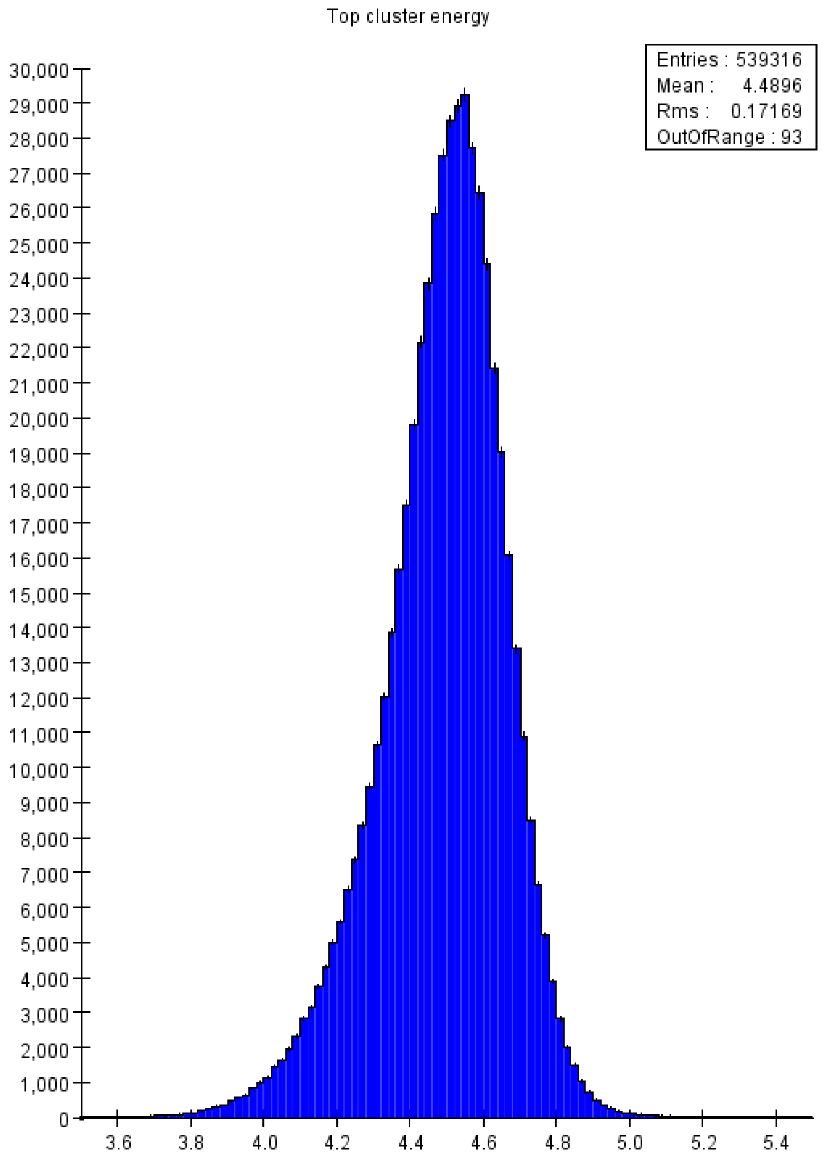
- Full Energy Electrons were used to calibrate the Ecal using an iterative crystal-by-crystal algorithm
- Single electron MC was used to derive "sampling fractions" *i.e.* energy lost in interstitial gaps.
- Check FEEs in data
  - Not an independent check since this data was used to establish the corrections, but useful in any case.
- Check Wide Angle Bremsstrahlung events
  - Sum of eγ energy should equal beam energy
- Check Fully-reconstructed Tridents
  - Sum of $e^+e^-e^-$ should equal beam energy

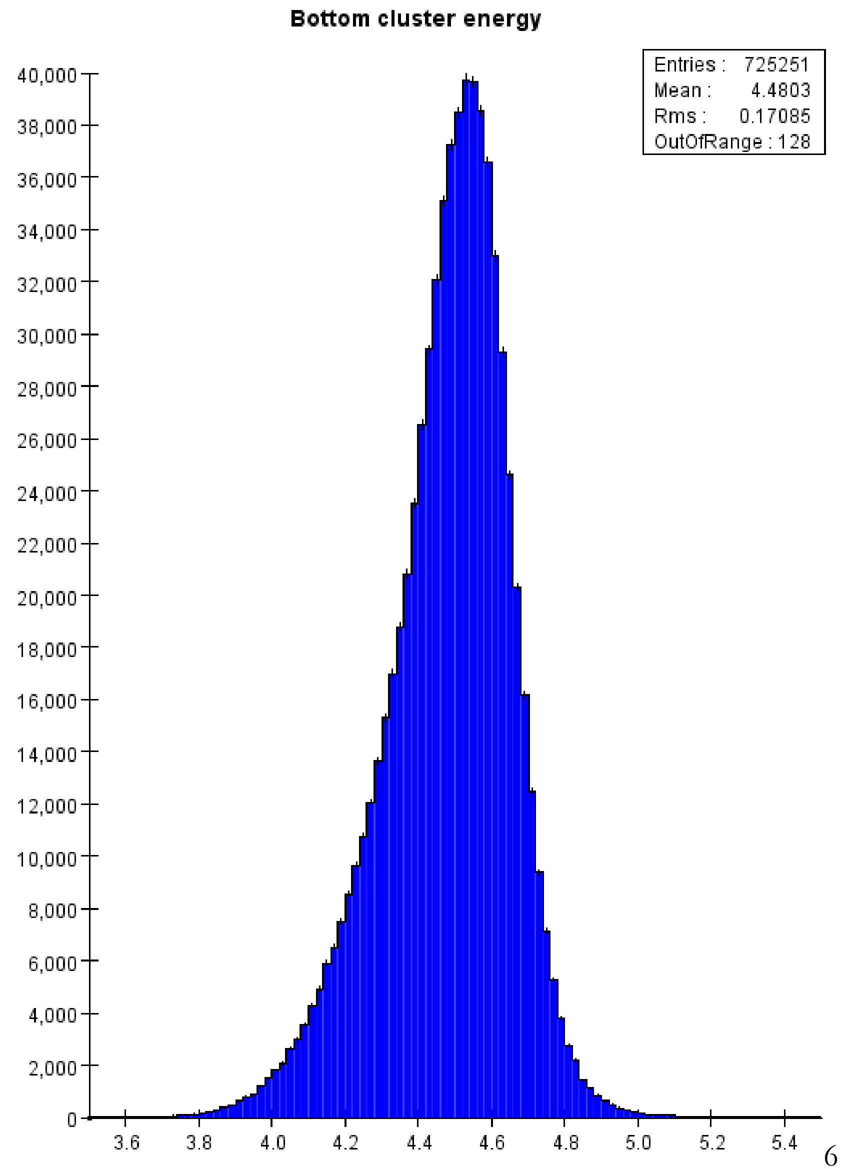
FEE Peak Calibration

- Test FEE peak by selecting with tight calorimeter-only cuts
  - One and only one cluster in the Ecal
  - “Fiducial clusters” : Cluster seed crystal not on edge
  - Seed crystal energy > 3GeV



FEE Single-cluster Energy



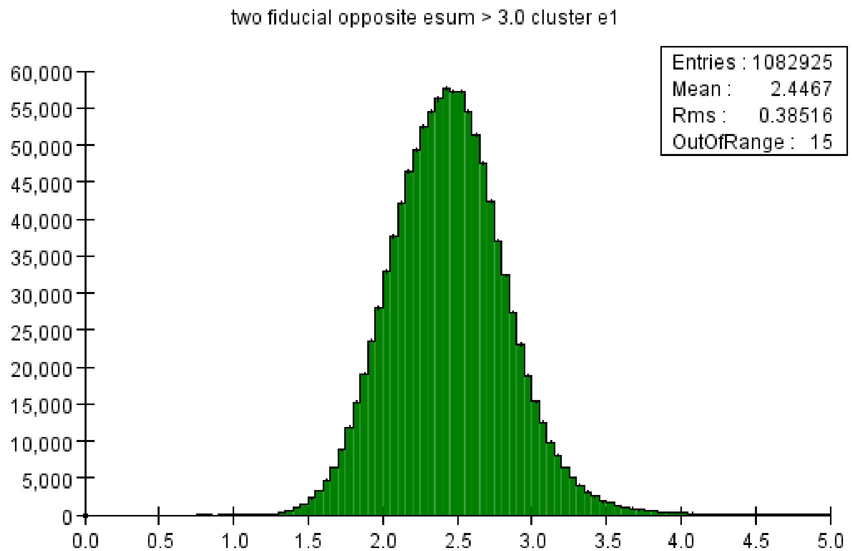


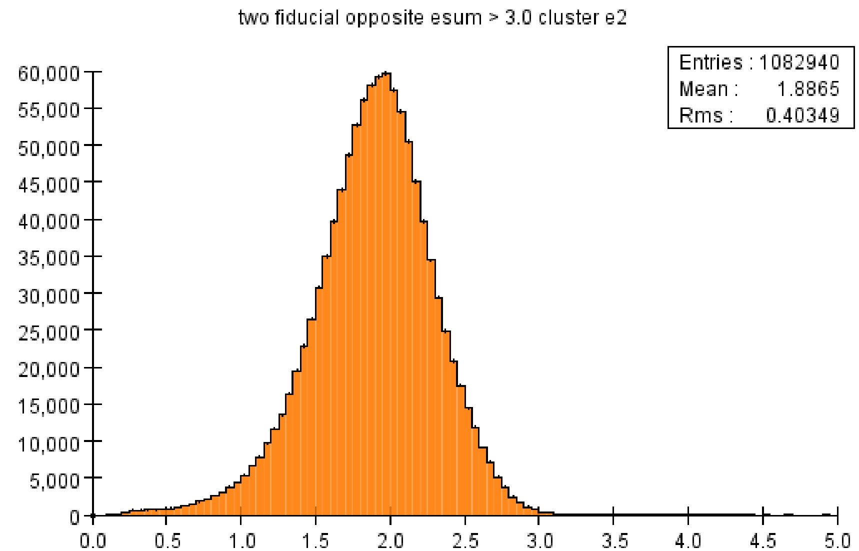
WAB Peak Calibration

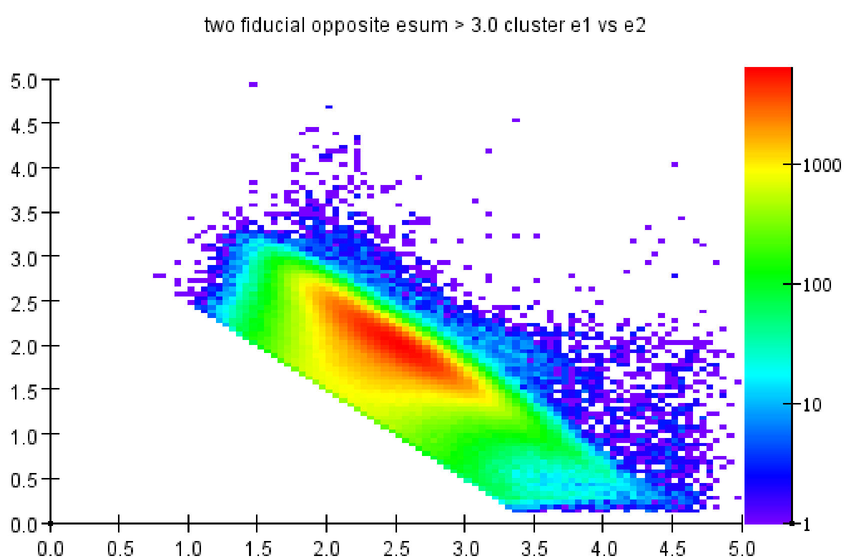
- Two and only two clusters in the event
- Clusters in diagonally opposite quadrants
- Both clusters in fiducial region
- Cluster times within 2ns of each other

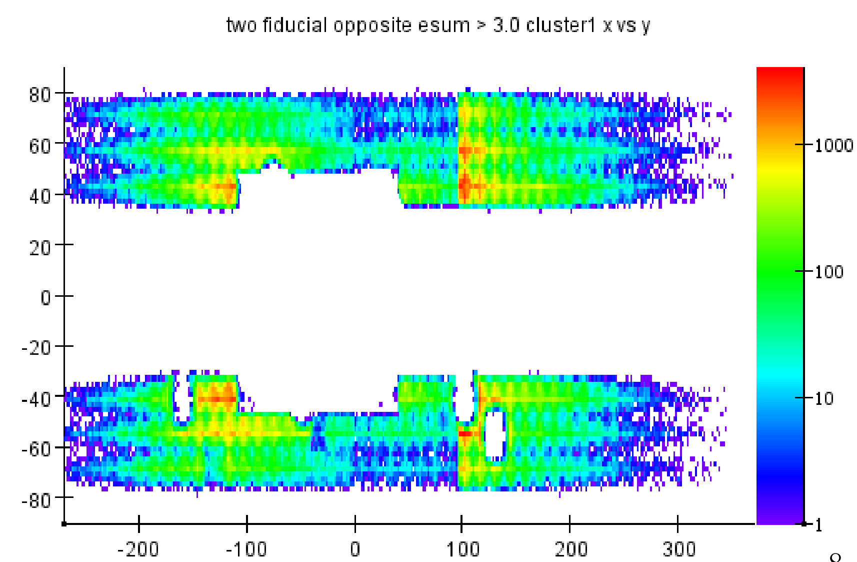
- Extends check of energy calibration to lower energies and broader coverage of ECal

WAB Cluster Energies

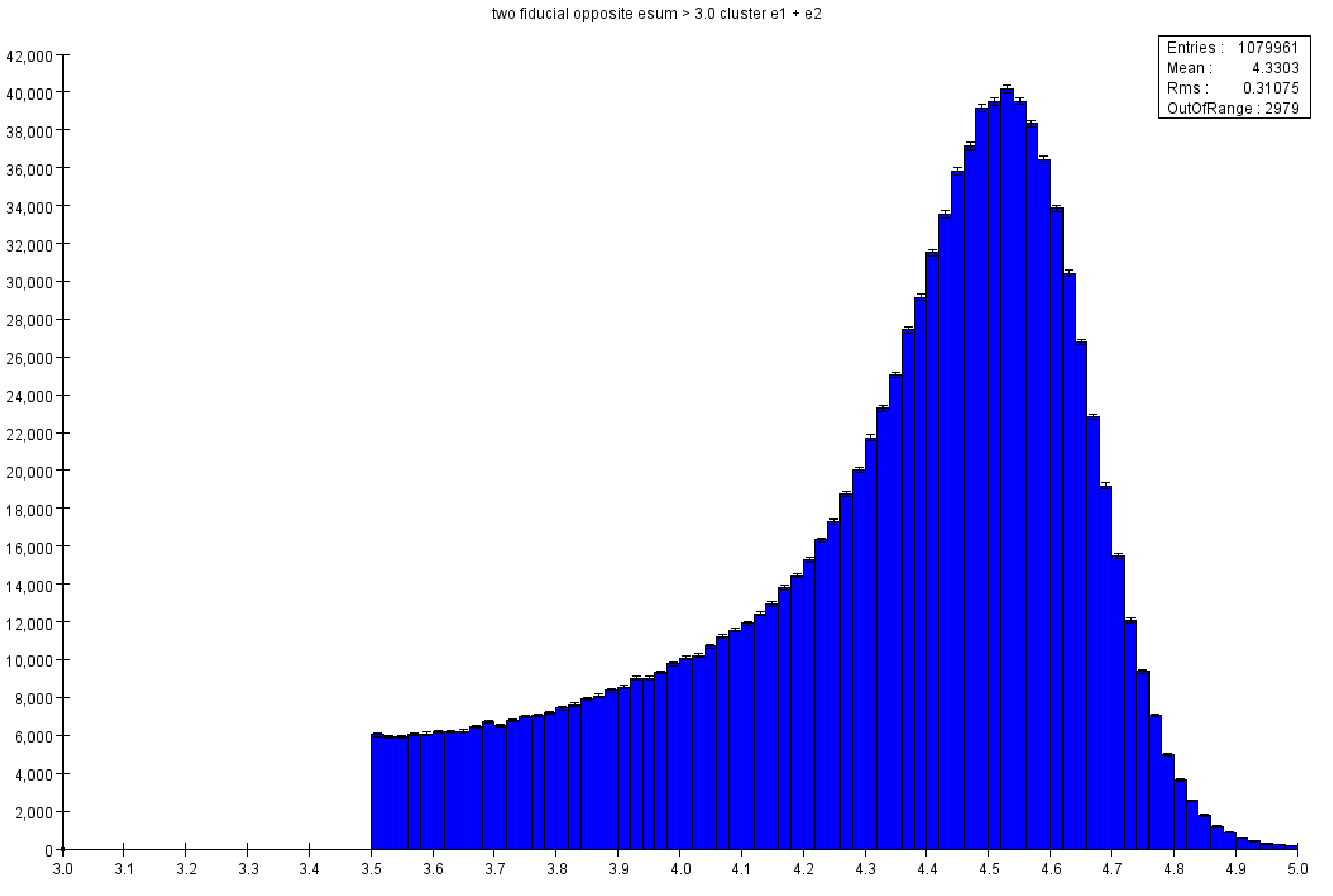








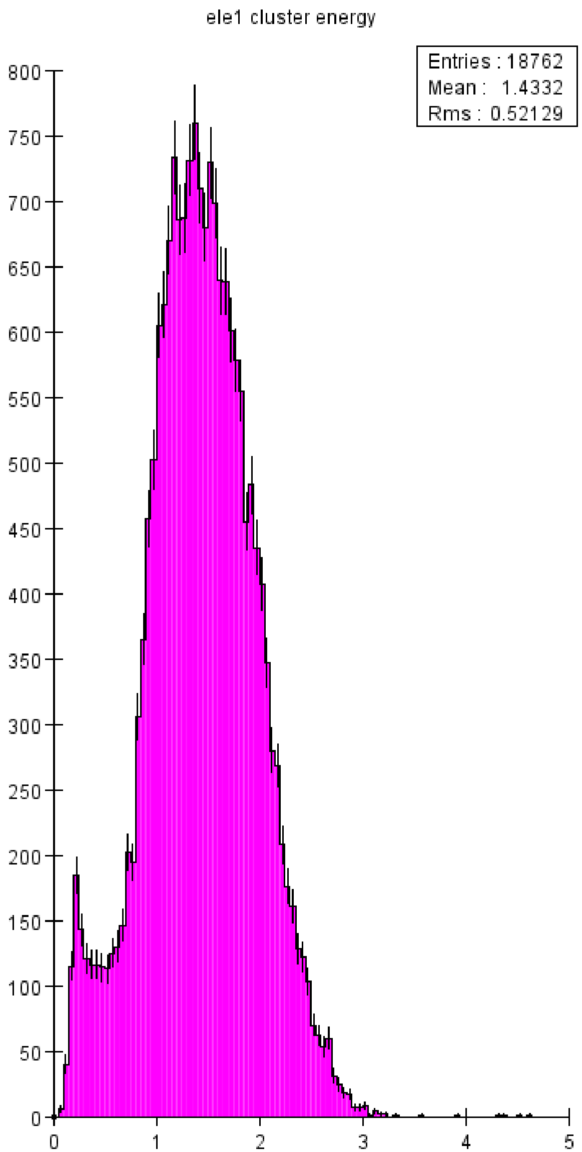
WAB Two-cluster Energy Sum

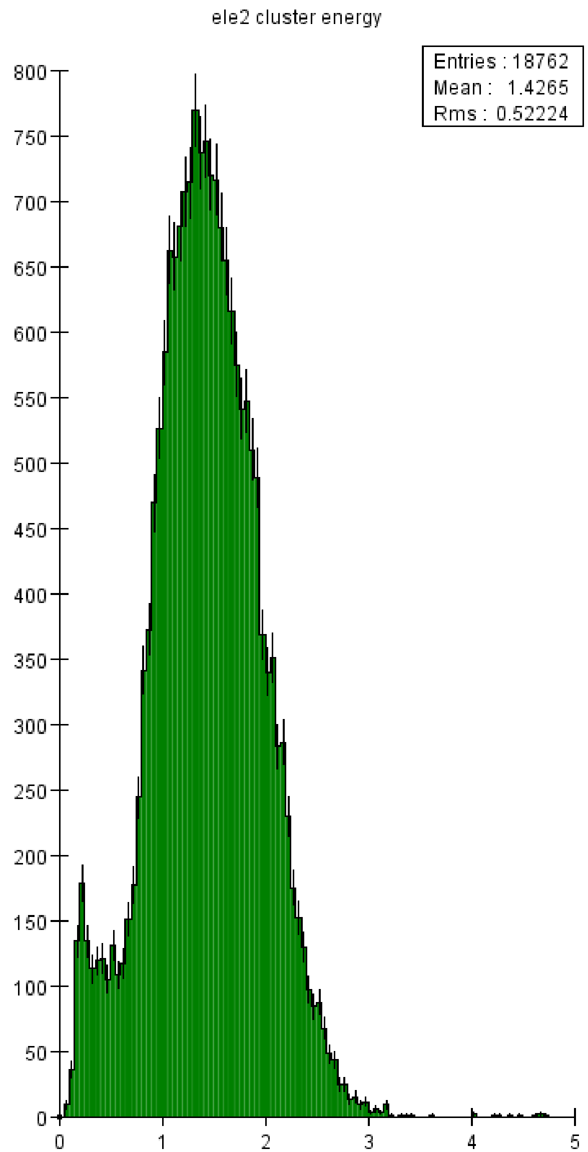


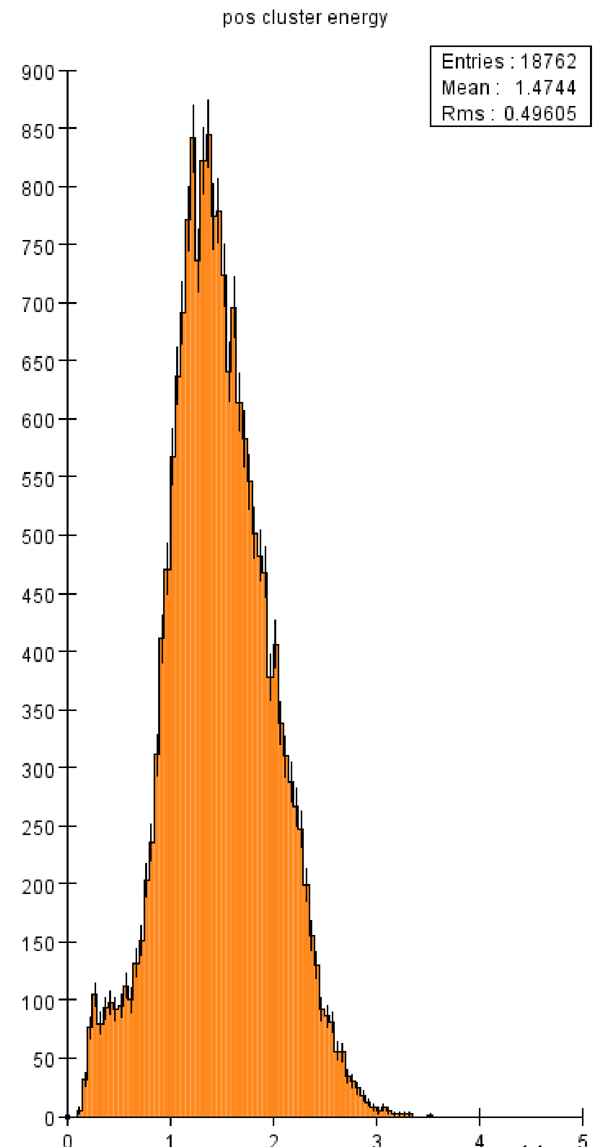
Trident Peak Calibration

- Select events with one reconstructed positron and two reconstructed electrons
  - note that this is the first time we have used tracking, so systematics are somewhat skewed by tracking efficiencies
- Require all three ReconstructedParticles to have an associate Ecal cluster
- All three clusters within 2ns of each other.

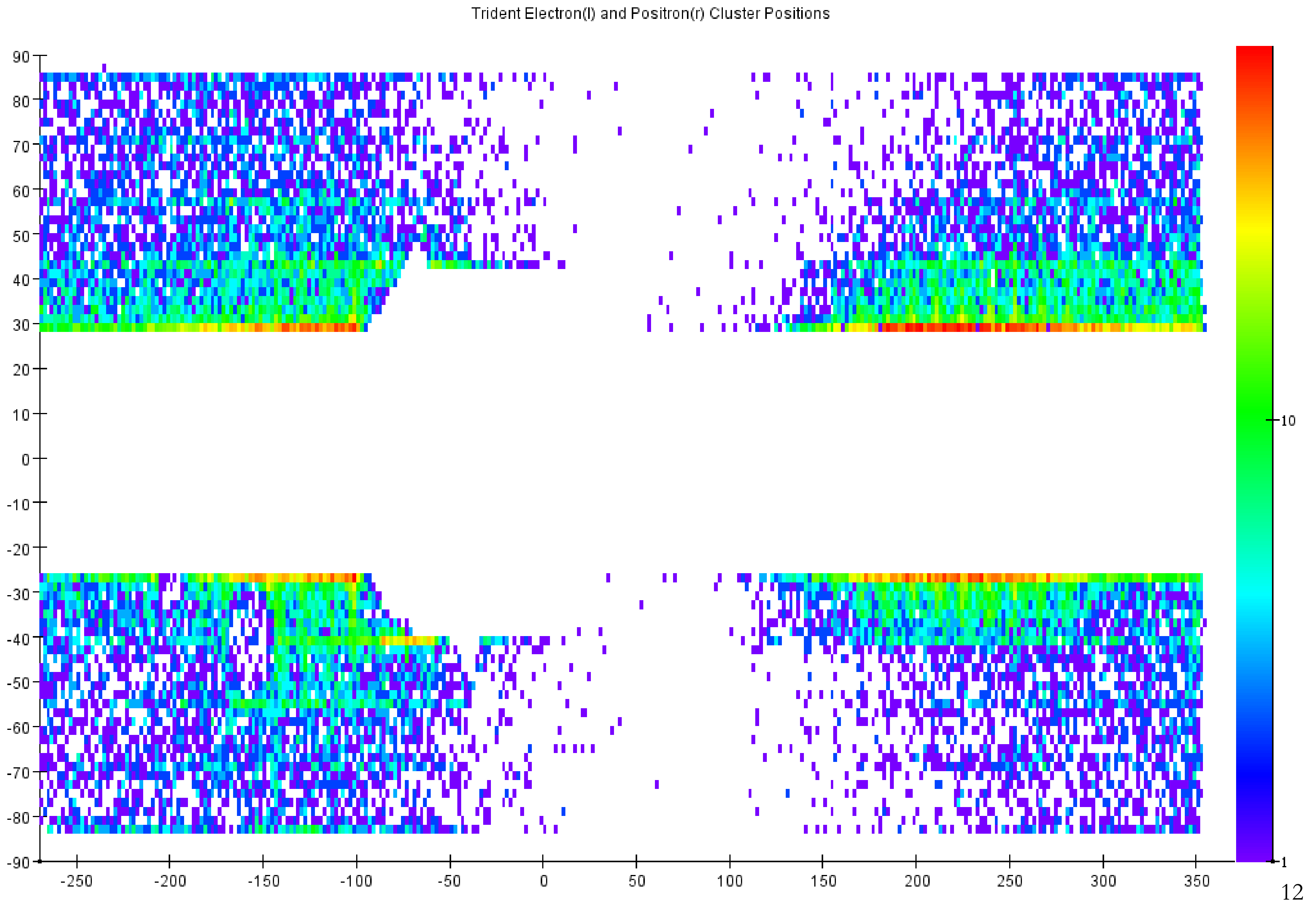
Trident Cluster Energies



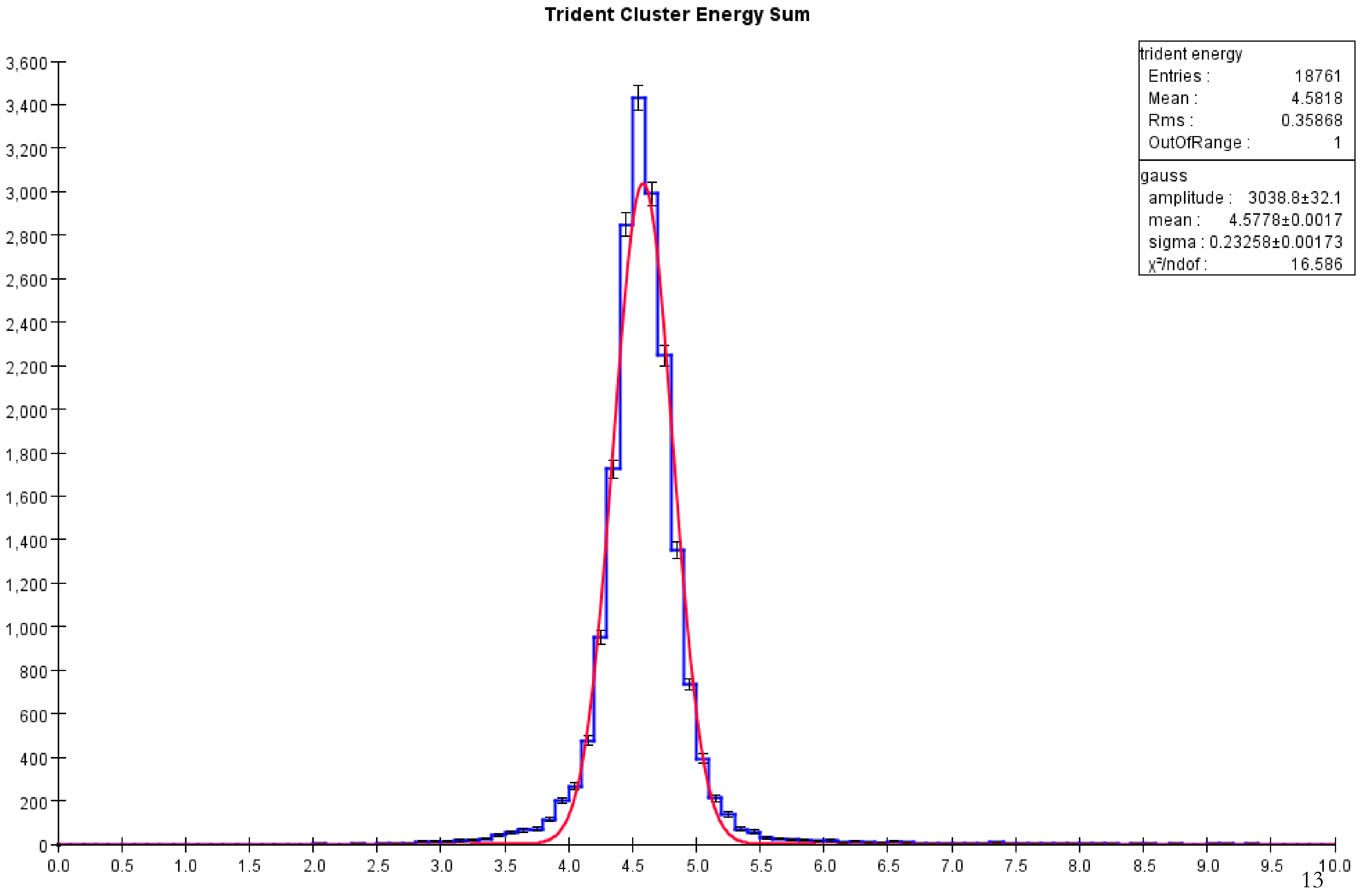




Trident Cluster Positions



2019 Trident Cluster Energy Sum



Energy Calibration

- Clean samples of FEEs, WABs and tridents can be isolated in the 2019 data with a few simple selection cuts.
- These allow us to check the Ecal cluster energy calibrations over almost the whole range of energies
  - FEE: 4.55GeV
  - WAB: ~1.5 – 3 GeV
  - Tridents: ~0.5 – 2GeV
- Distributions look good “by eye”, but analyses need to be quantified and compared to MC.

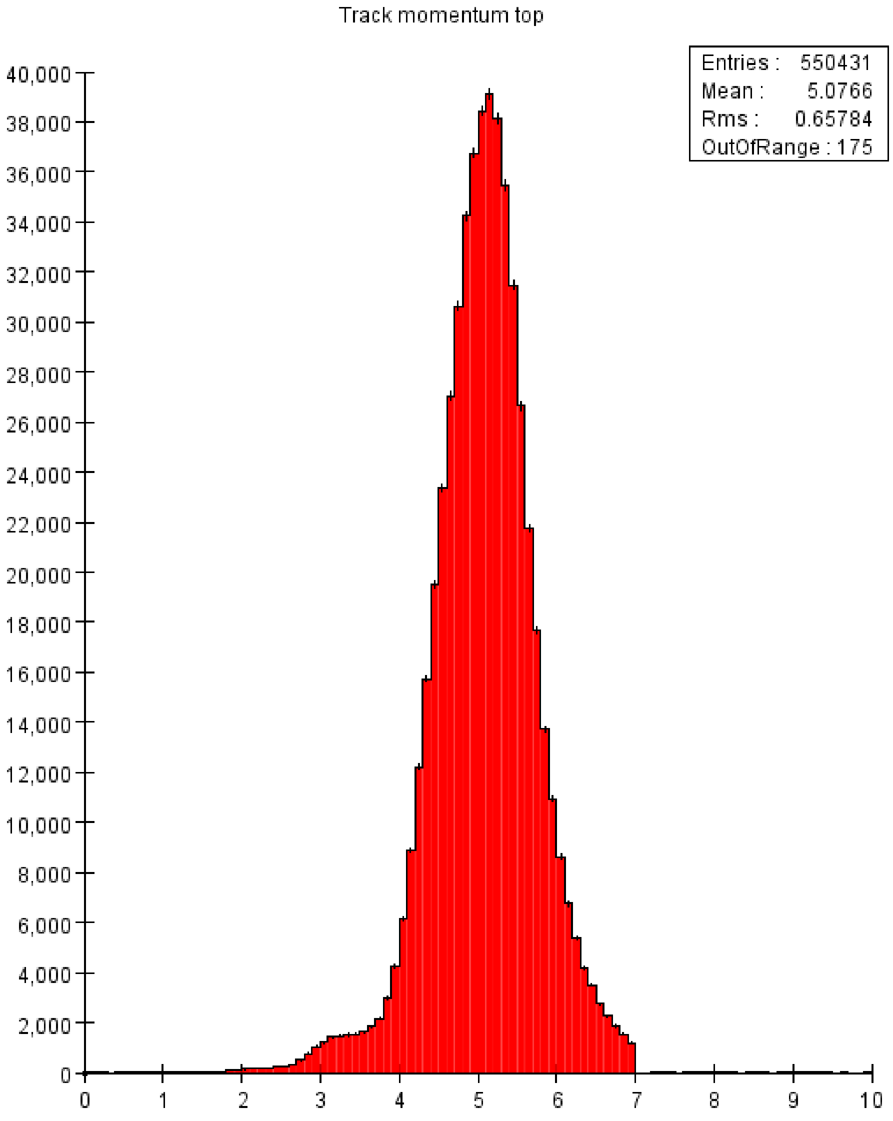
Momentum Calibration

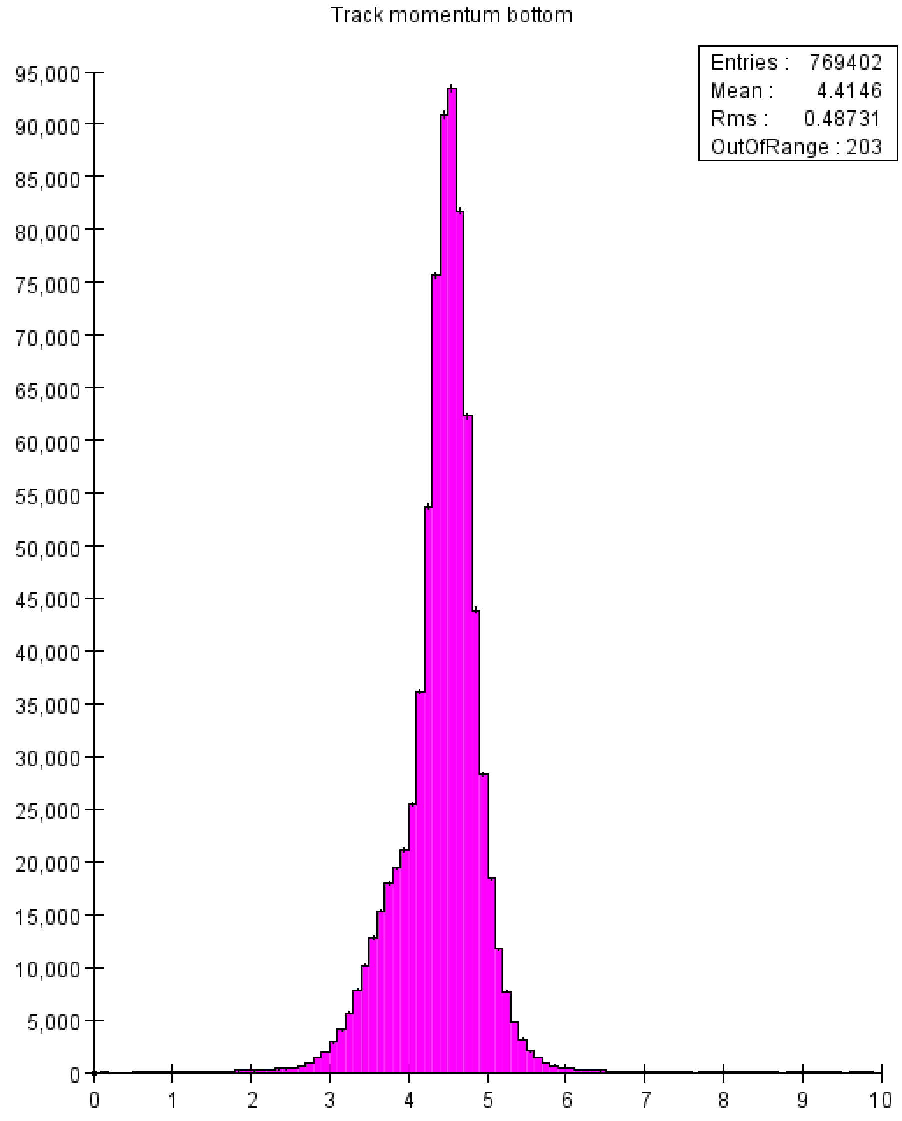
- Momentum calibration is inextricably intertwined with the alignment of the SVT.
  - See the enormous body of work done by PF!
- Can use FEEs and impose a momentum-constraint of 4.55 GeV on the alignment procedure.
- Once we trust the energy calibration of the Ecal, we can use tracks which have been associated with clusters to impose a momentum-constraint on the alignment procedure.
- Can also extend geometric coverage beyond that available from FEEs.
- I will not be talking about using momentum-constraints in the alignment.
- I will only be testing the calibration (alignment) by comparing energy to momentum of ReconstructedParticles.

FEE Track Momenta

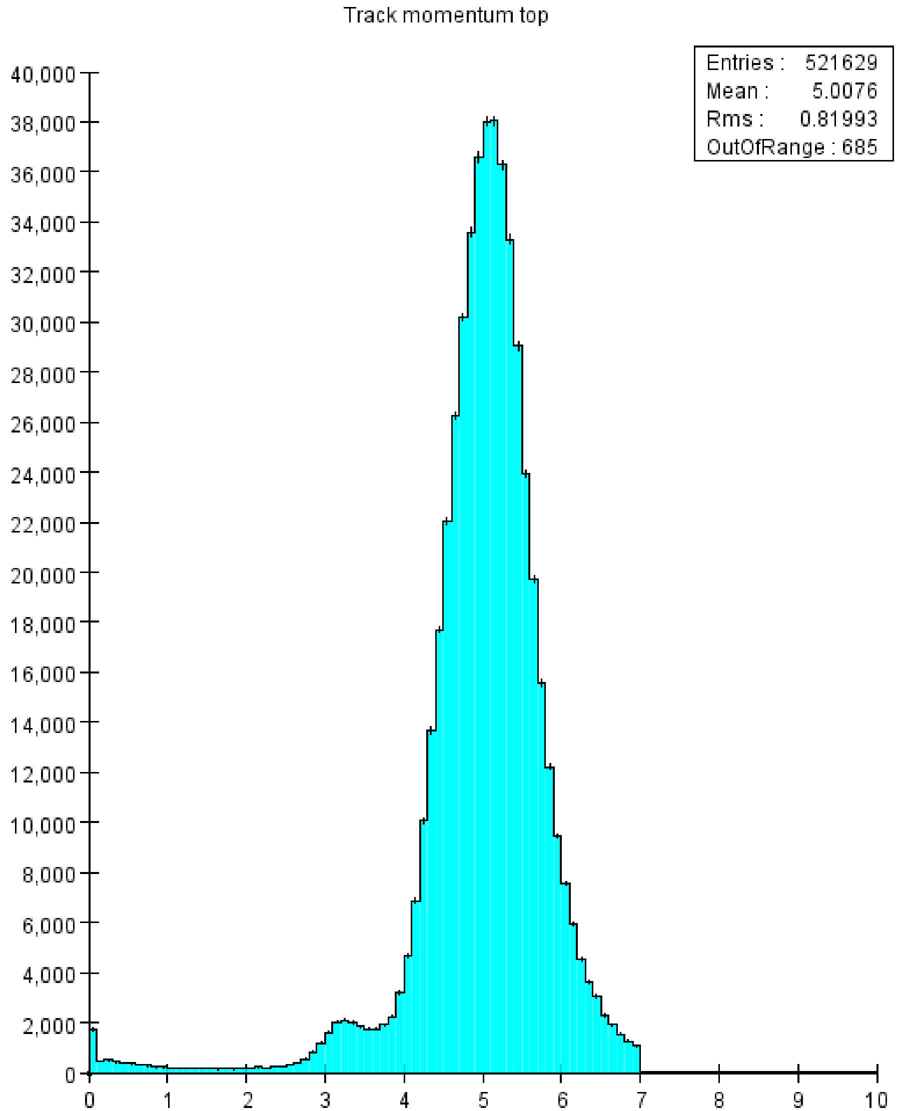
- For the time being we are comparing the legacy track finding & fitting using the Seed Tracker and General Broken Lines (GBL) with the Kalman Filter (KF)
- Full comparison of the two is beyond this talk, but some issues germane to calibration will be presented.
- Small diversion to touch on relative track-finding efficiencies
  - Using Ecal-only selection criteria, plot relative number of electrons vs photons. Note that we are only looking at events in the fiducial region of the calorimeter.

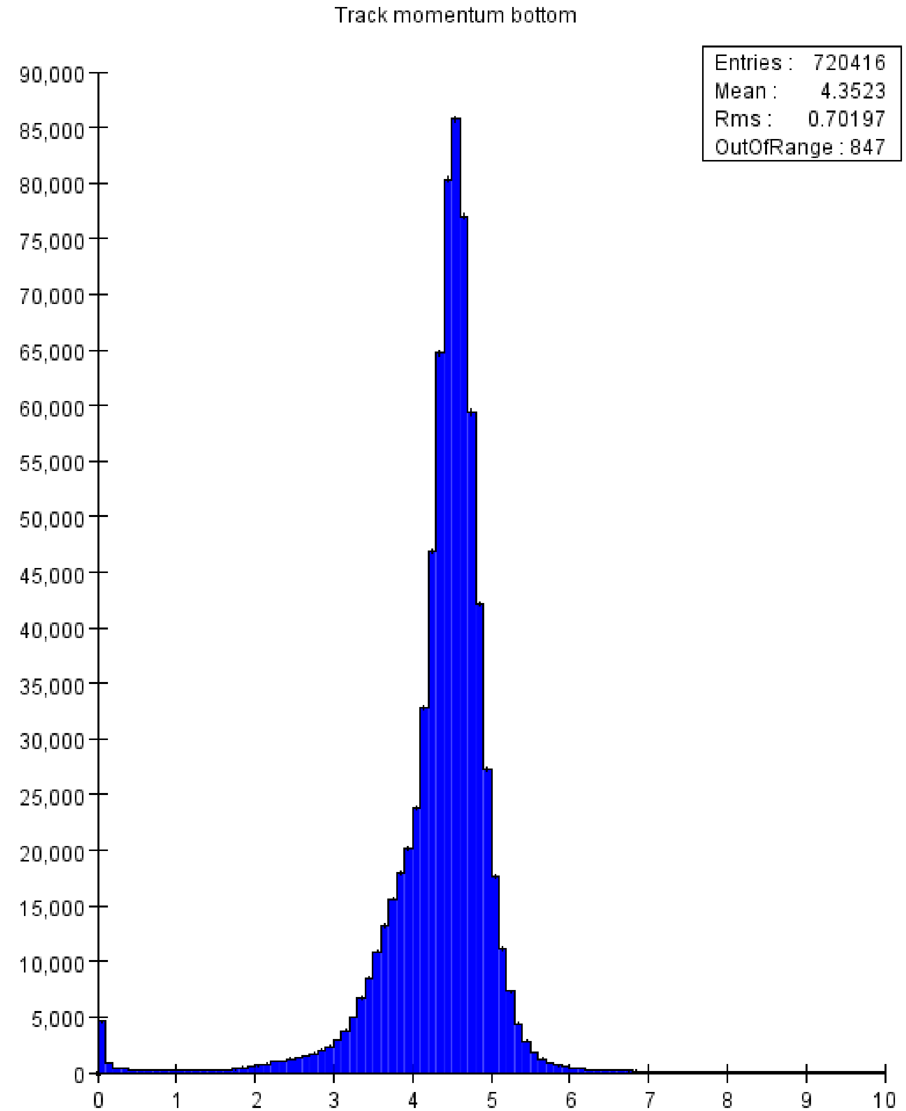
FEE Track Momenta GBL





FEE Track Momenta KF

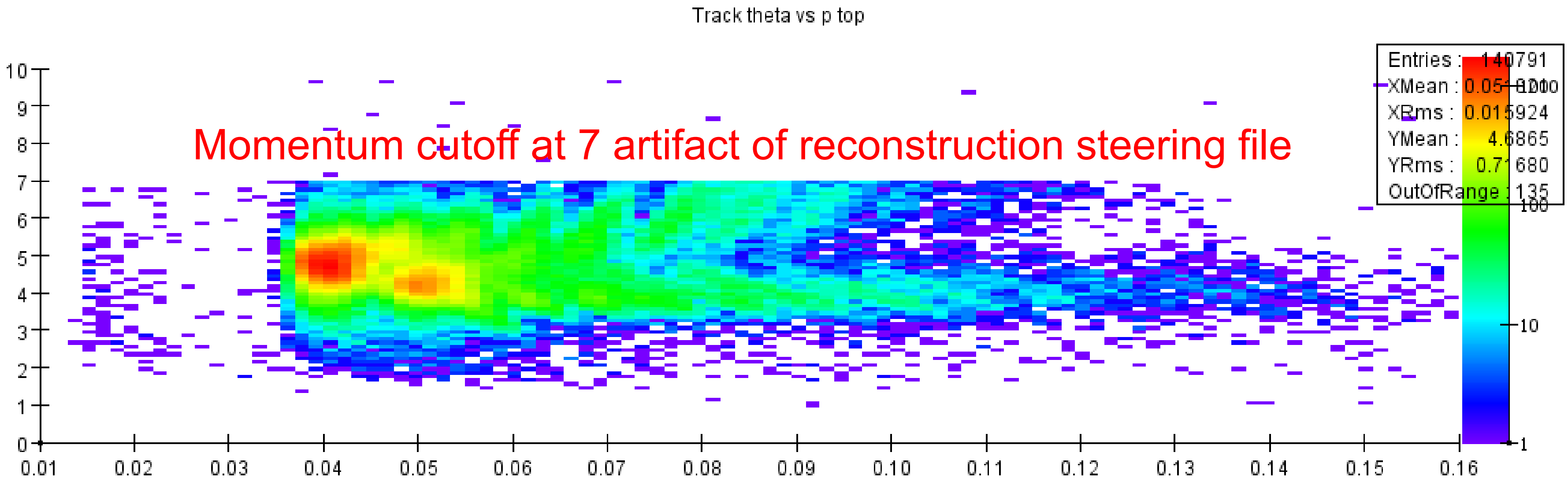


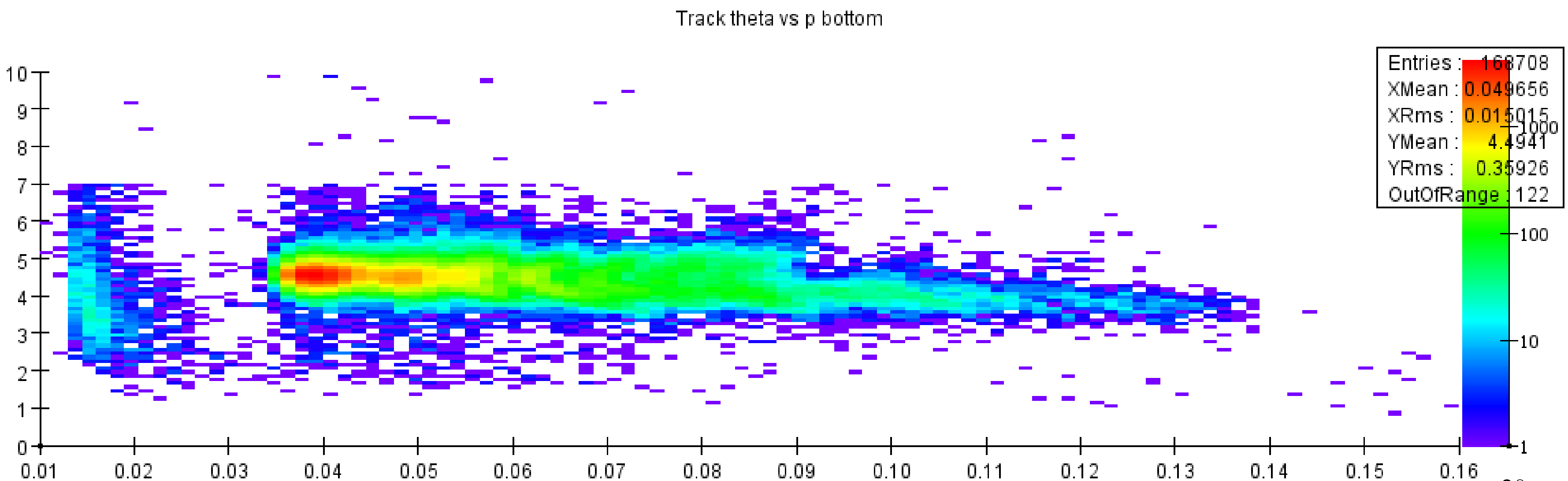


What's up with the top SVT?

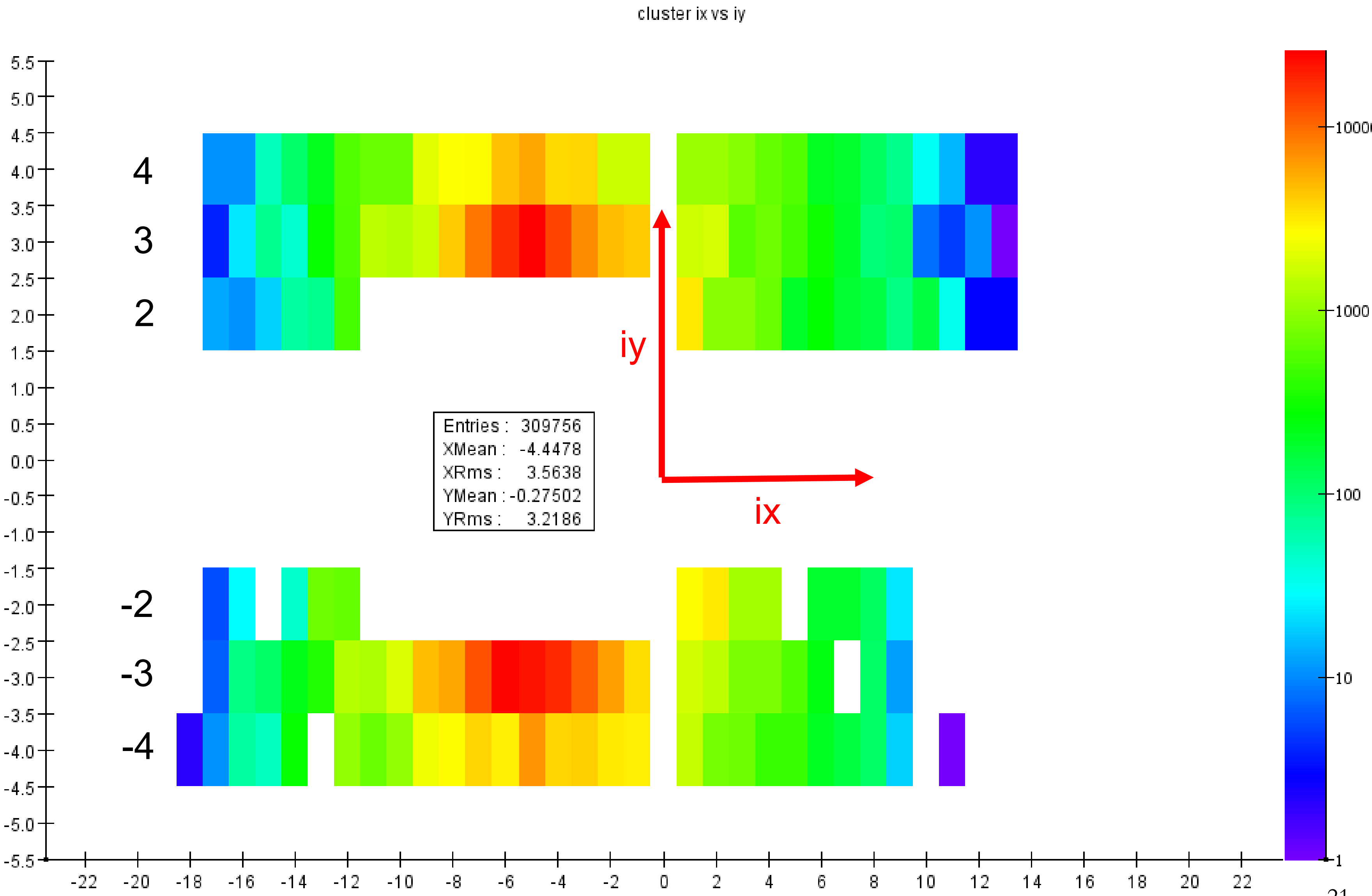
- SVT momentum is clearly being measured much more poorly in the top than the bottom.
- Are the sensors simply that much more misaligned?
- Are there some larger global issues that we are missing?
- Are there some systematics we are overlooking?

Track Momentum vs Theta

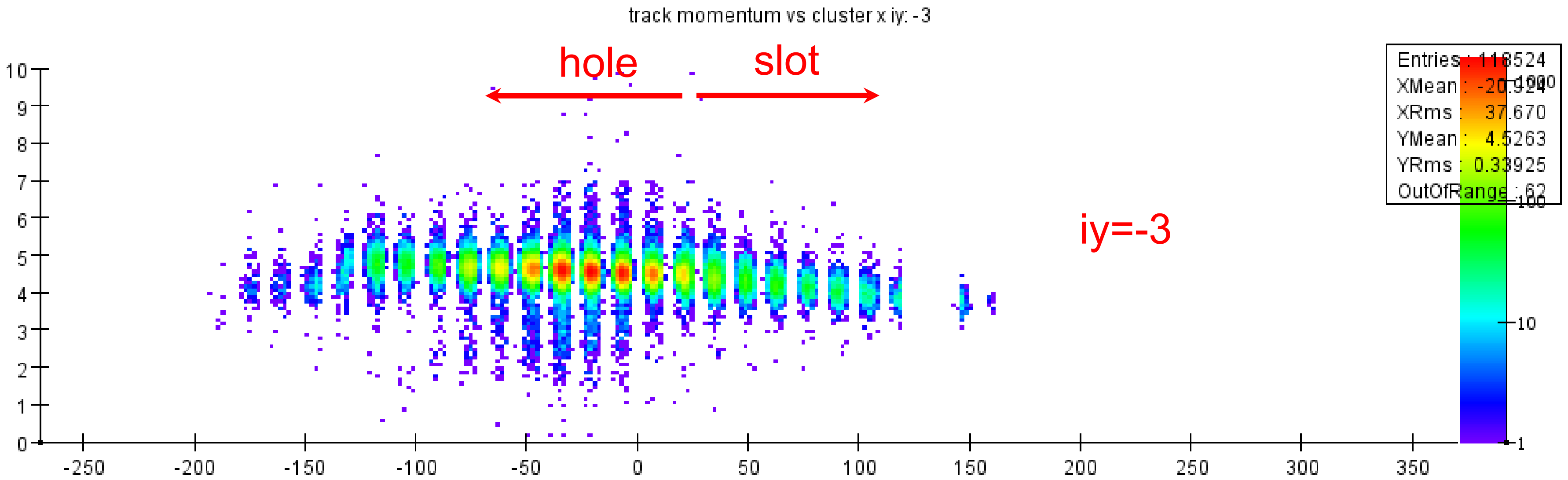


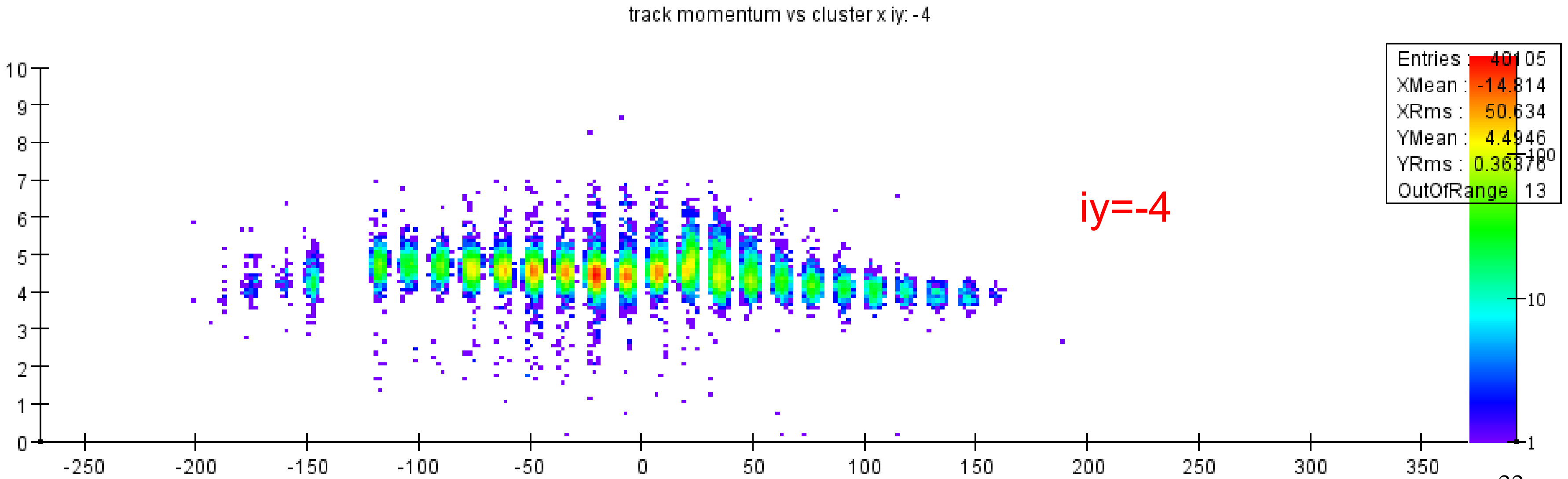


Clusters With Track Seed Crystal ID

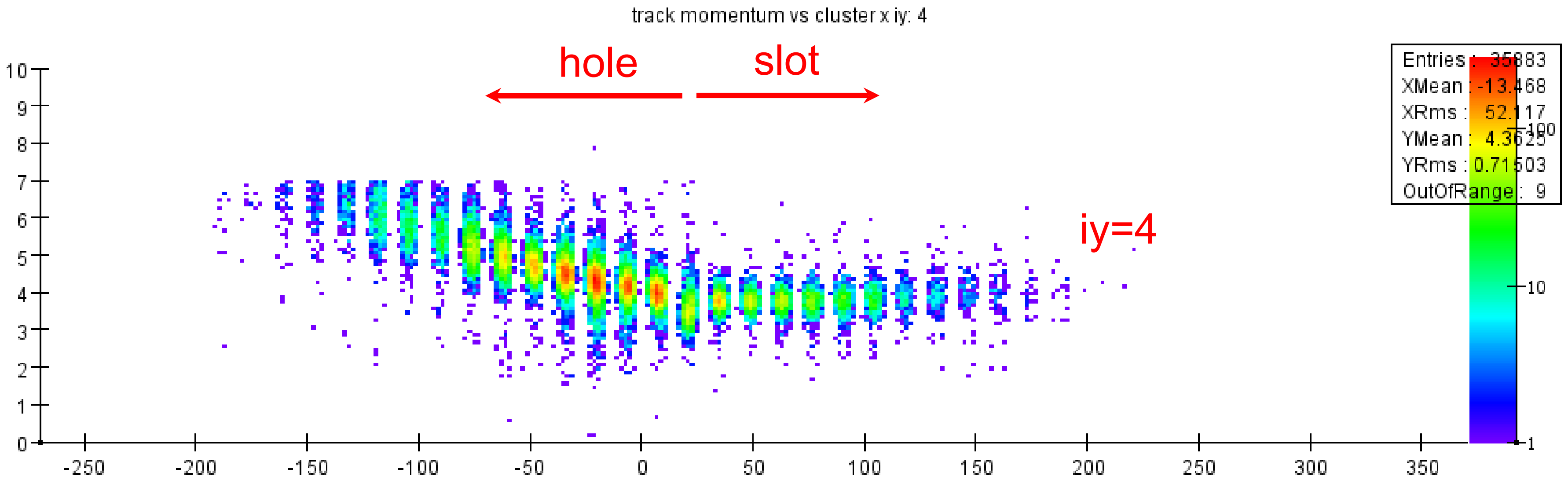


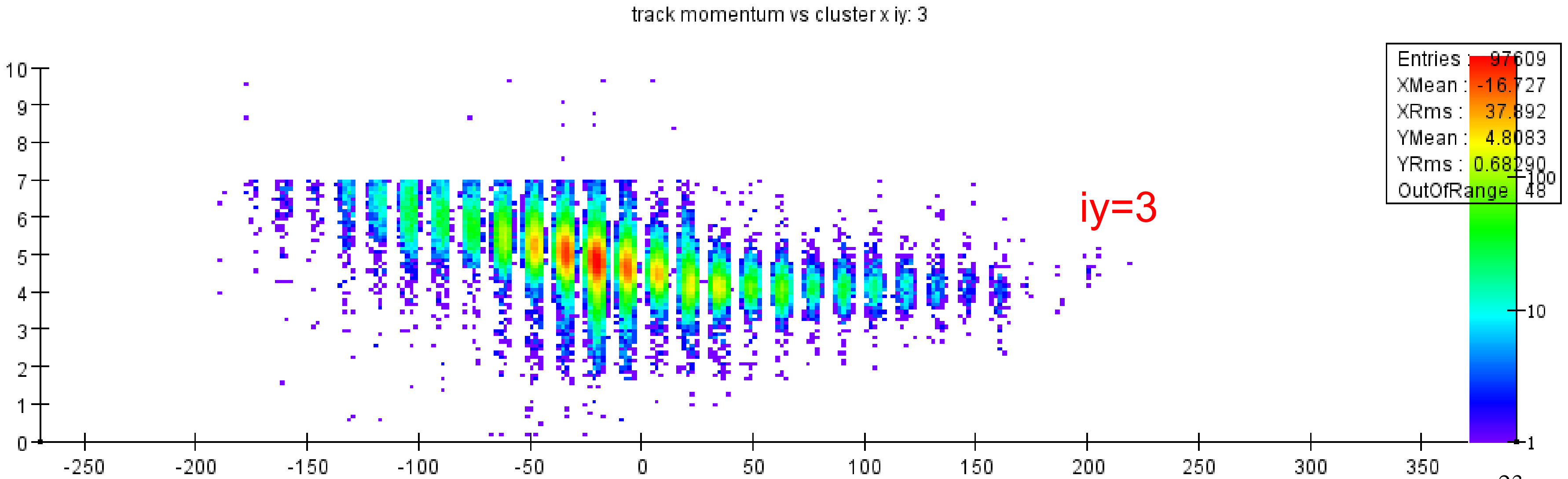
Bottom Track Momentum vs Cluster



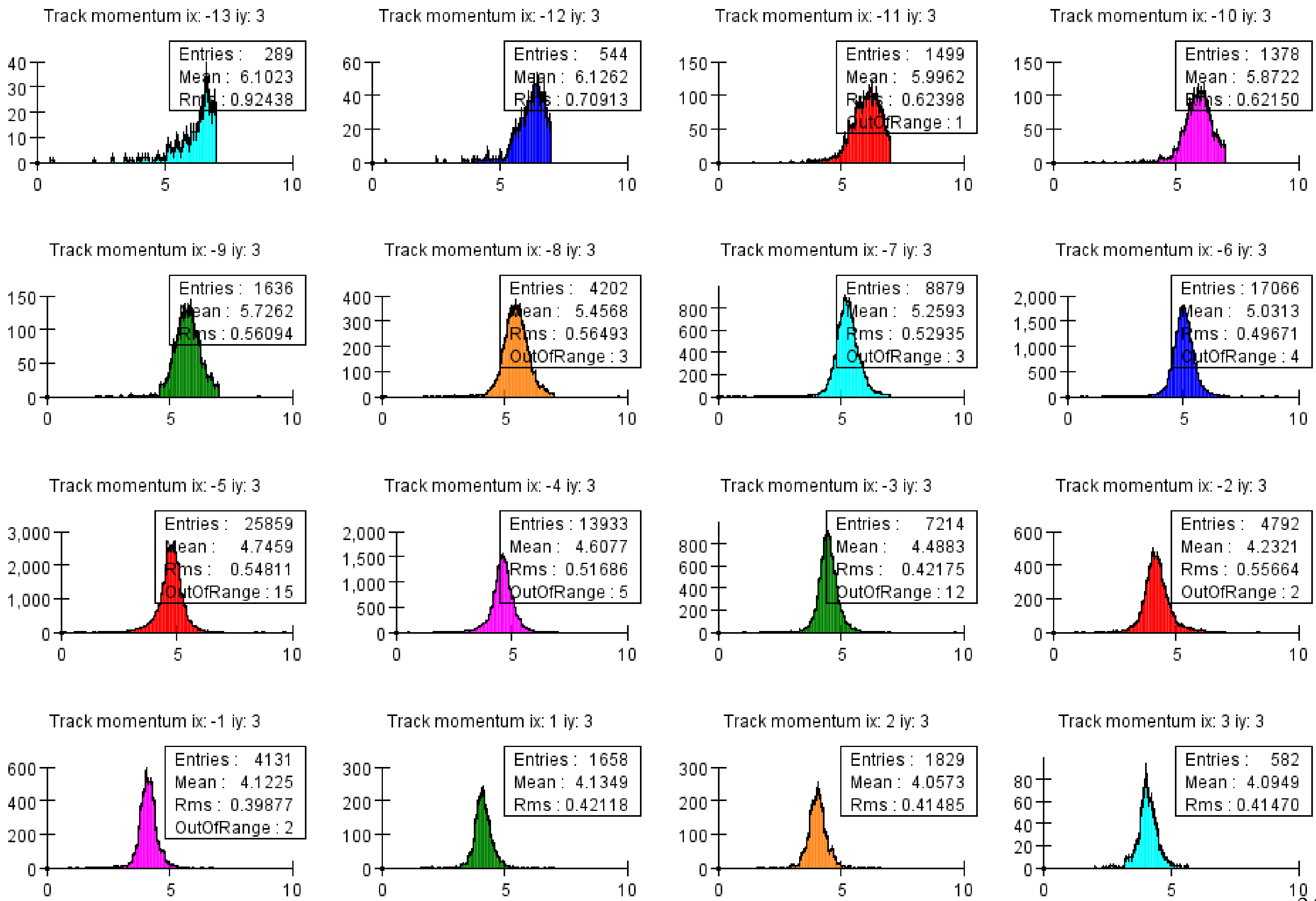


Top Track Momentum vs Cluster x

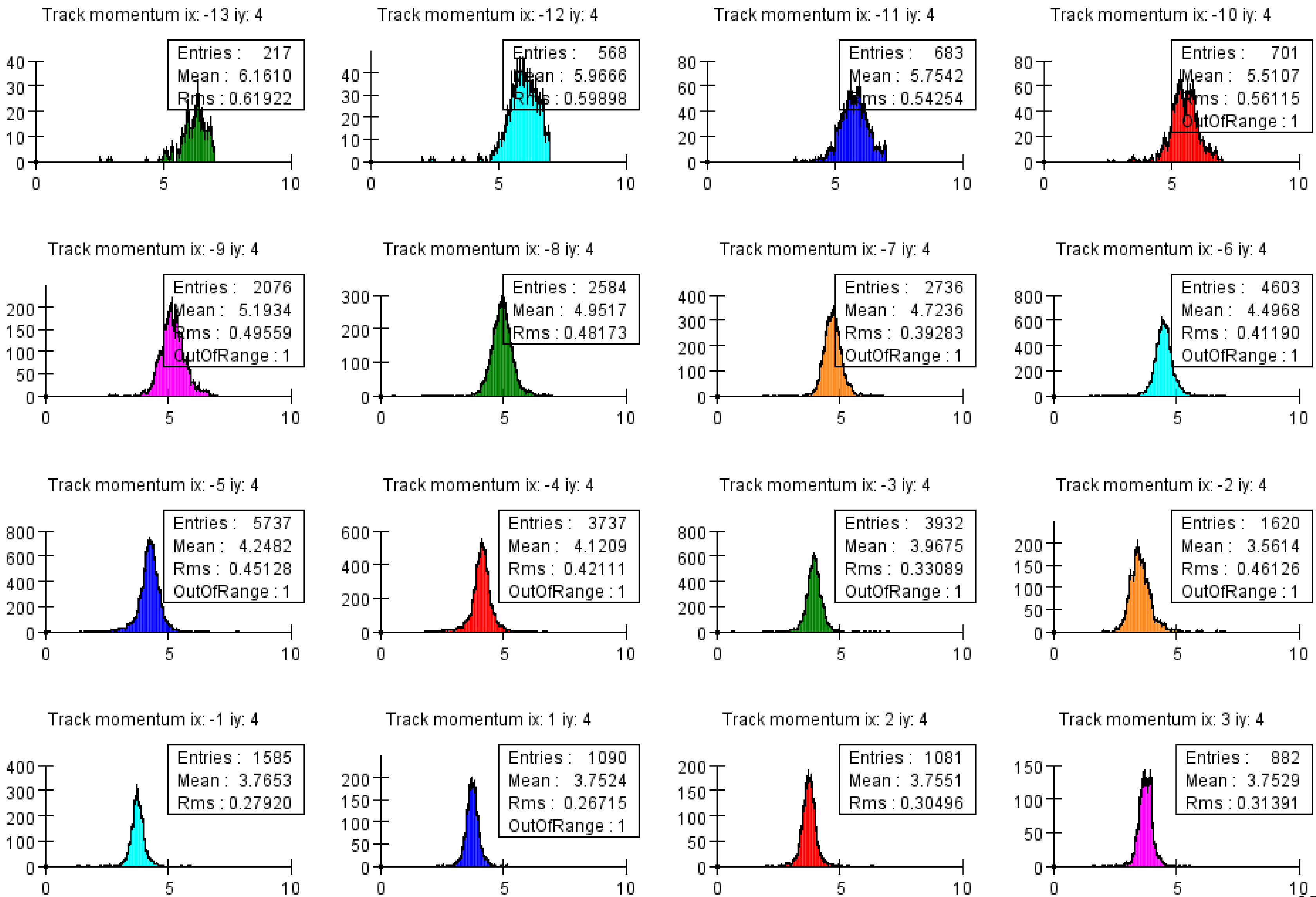




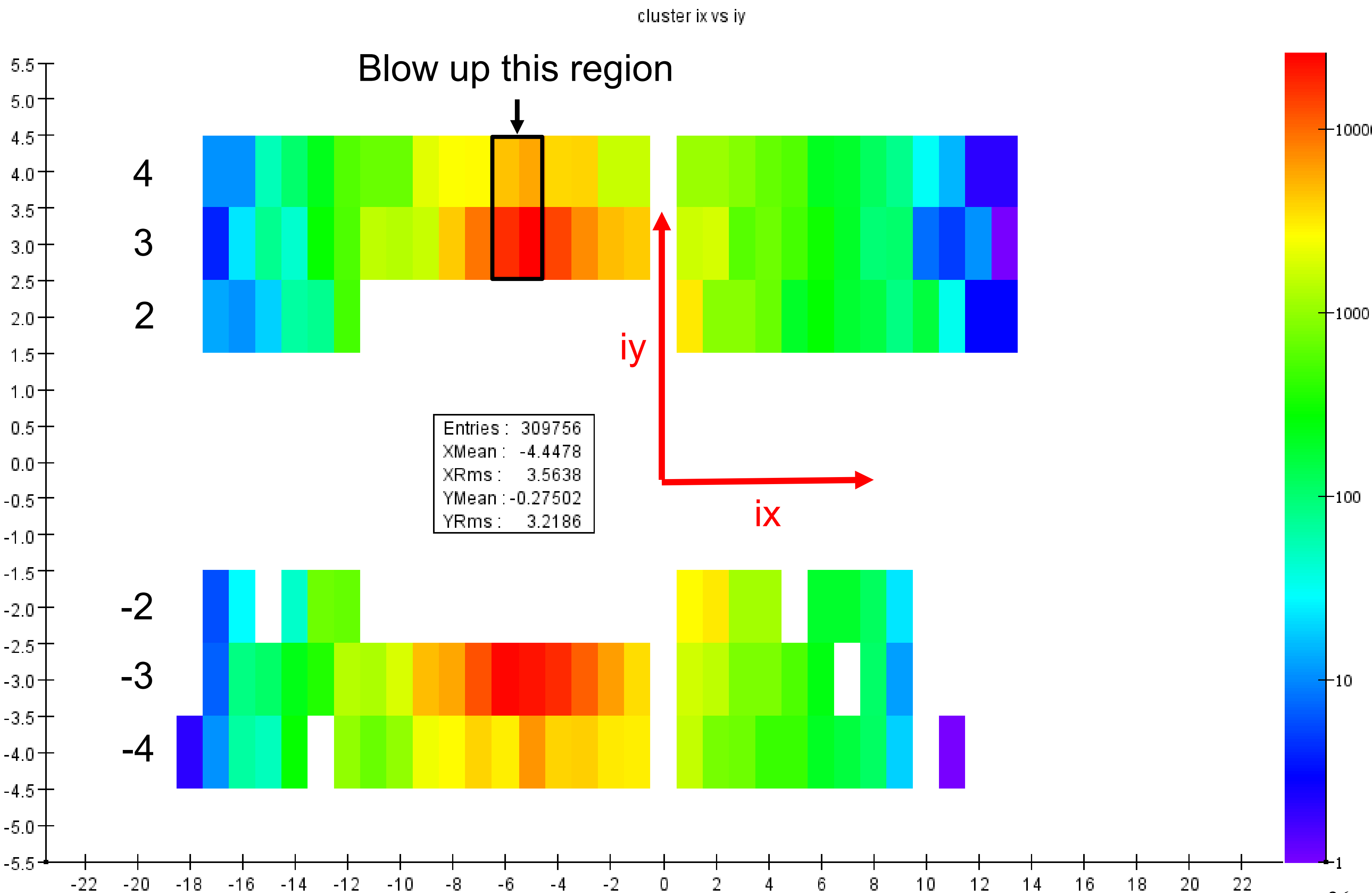
Top Track Momentum by Crystal iy=3



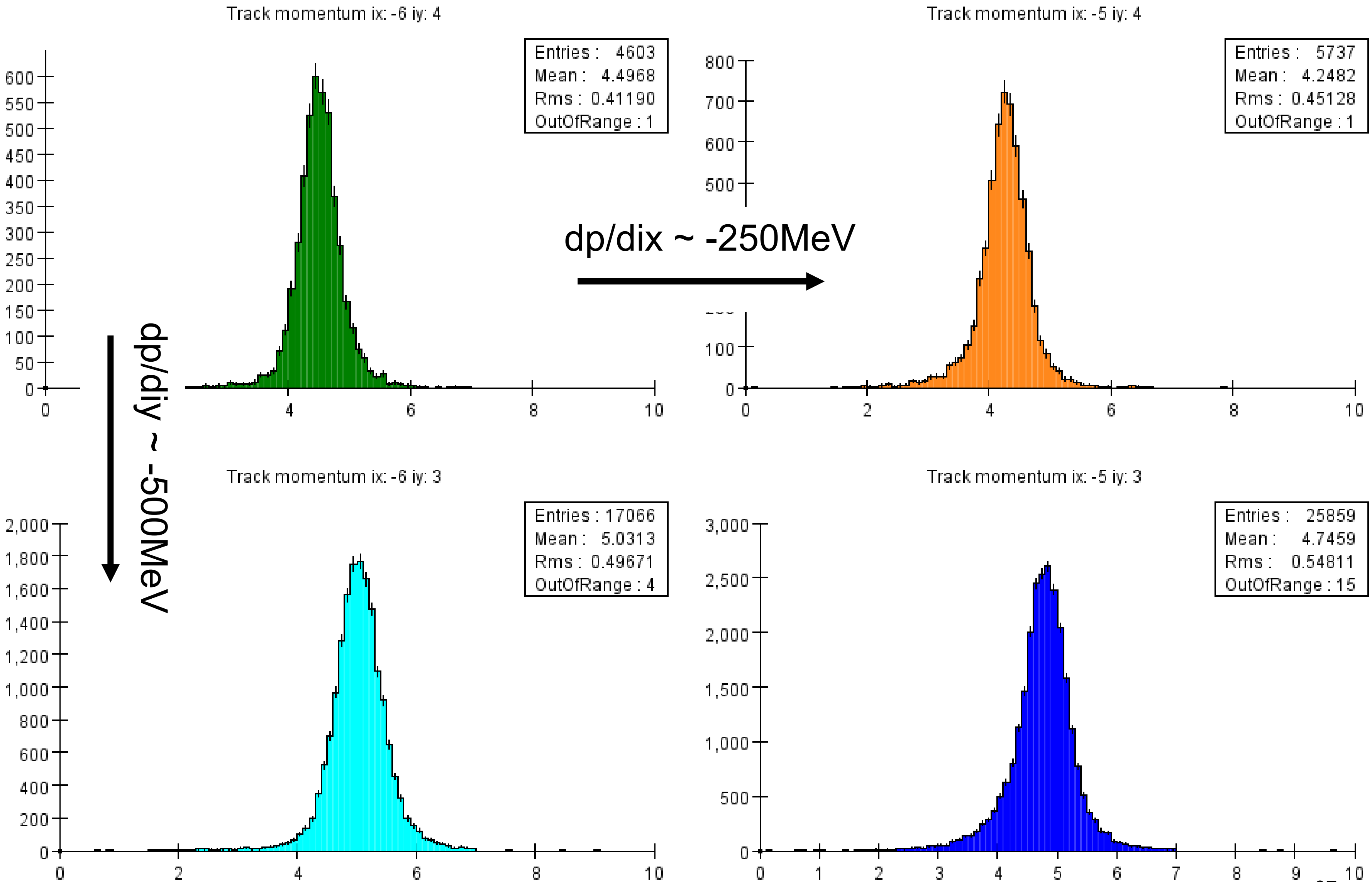
Top Track Momentum by Crystal iy=4



Clusters With Track Seed Crystal ID



Top Track Momentum vs ix & iy



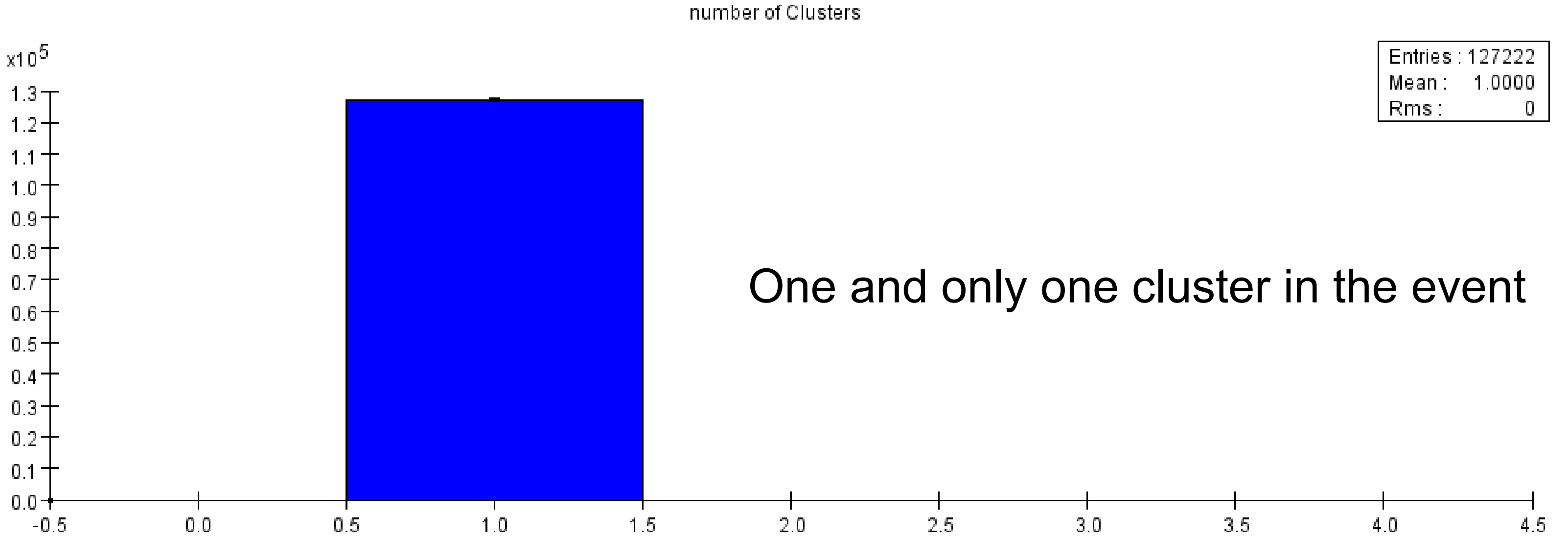
SVT Top Track Systematics

- Top SVT tracks appear to be afflicted with a number of rather severe systematic effects
  - “slot” appears disconnected from “hole”
  - momentum shifts as a function of x in Ecal
    - dp/dix ~ -250MeV
  - momentum shifts as a function of y in Ecal
    - dp/diy ~ -500MeV
- Is there some common geometrical misalignment which can be causing this?

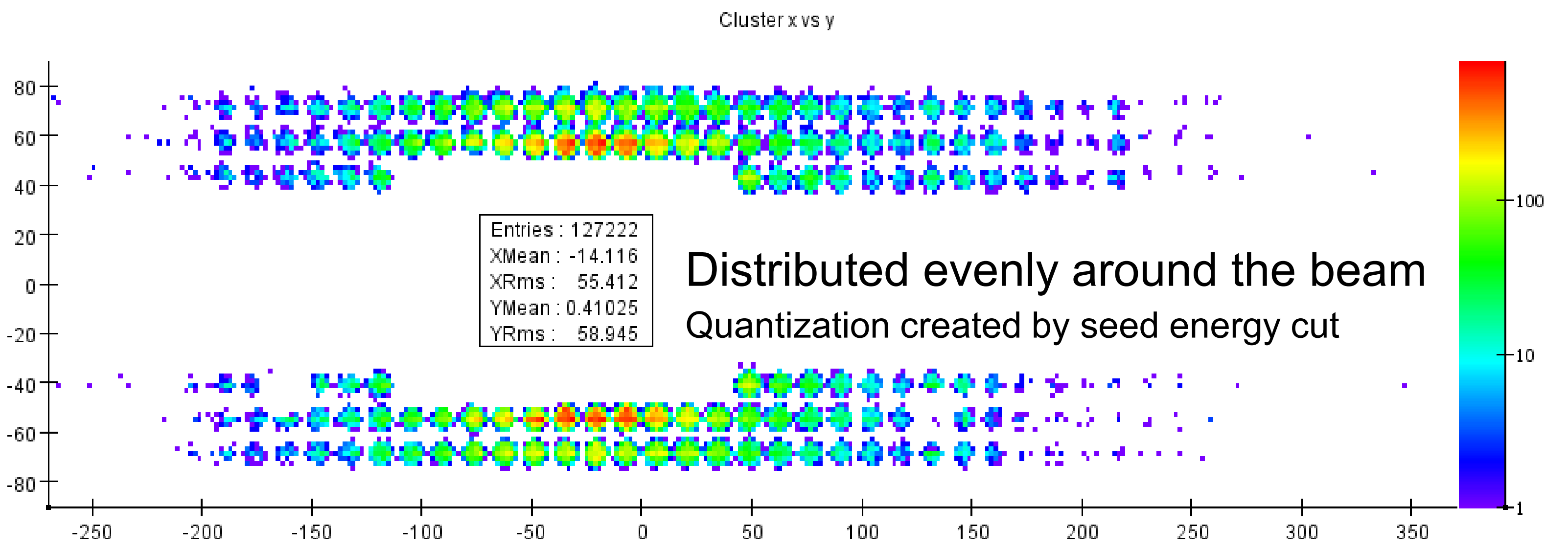
Tracking Efficiency with FEEs

- Skim events containing a single high-energy cluster in the fiducial region of the ECal with seed energy > 3.0GeV
- Provides 127222 clean FEE candidates
- With only a single cluster in the event, any ReconstructedParticle identified as a photon points to a failure either to reconstruct a track or to associate that track to the FEE cluster
- "FinalStateParticles“ contains tracks found using the SeedTracker strategies and fit using the Global Broken Lines fitter, labeled as GBL
- "FinalStateParticles_KF“ contains tracks found and fit using the Kalman Filter strategies, labeled as KF

FEE Data Sample



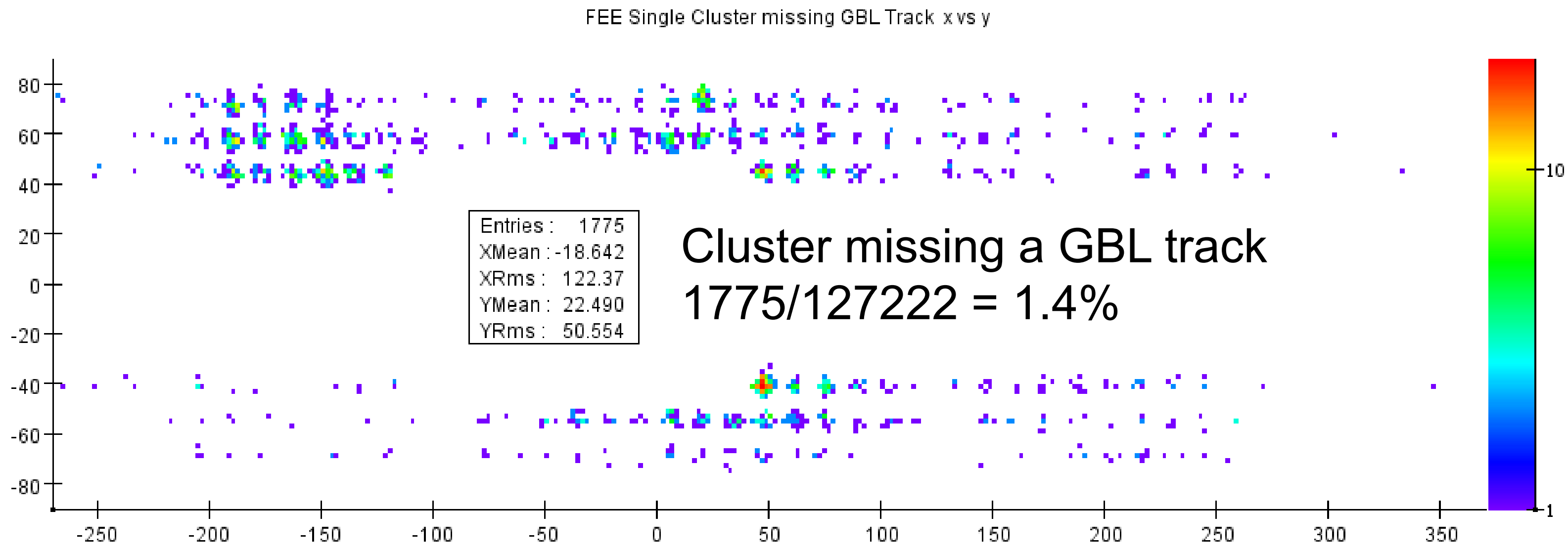
One and only one cluster in the event

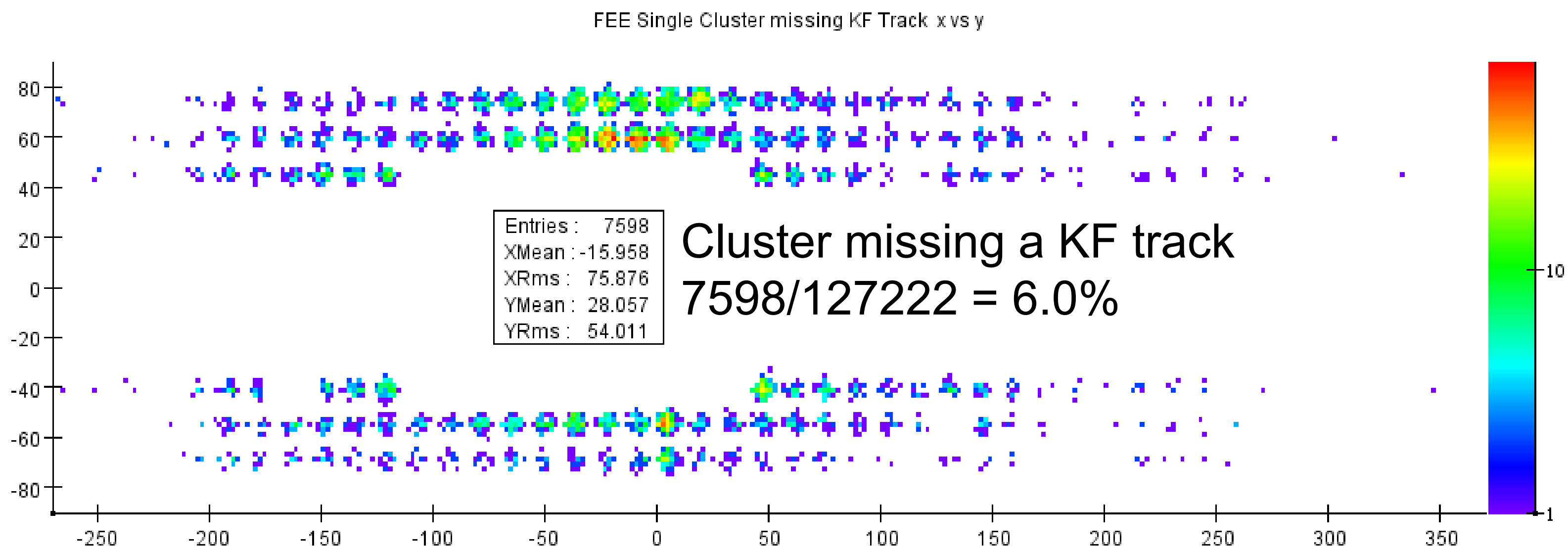


Distributed evenly around the beam

Quantization created by seed energy cut

FEE Track-Finding Inefficiency





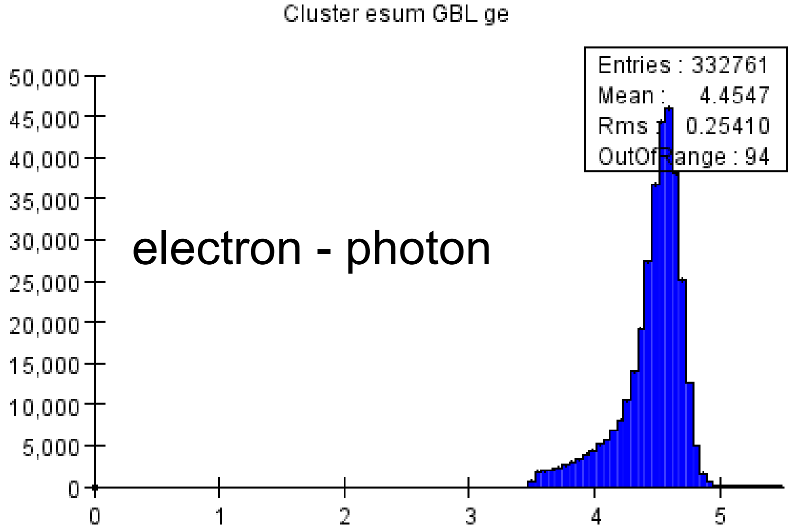
Track-Finding Efficiency with WABs

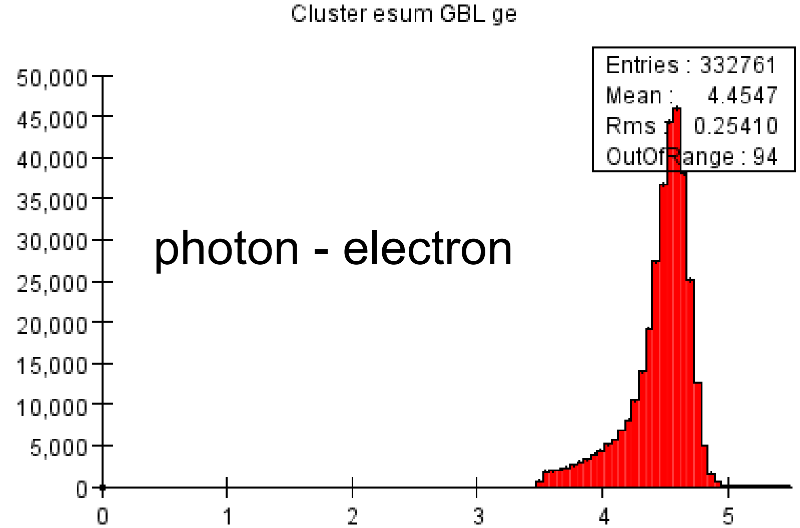
- Data Samples
  - Data run 31
- Reconstruction Version
  - hps-java 5.1 snapshot
- Detector
  - HPS_PhysicsRun2019-v2-FEE-Pass0
- Skim events containing two and only two clusters in the fiducial region of the calorimeter
- Clusters in diagonally opposite quadrants
- Cluster times within 2ns
- Cluster Esum > 3.5GeV
- Provides 708542 WAB candidates

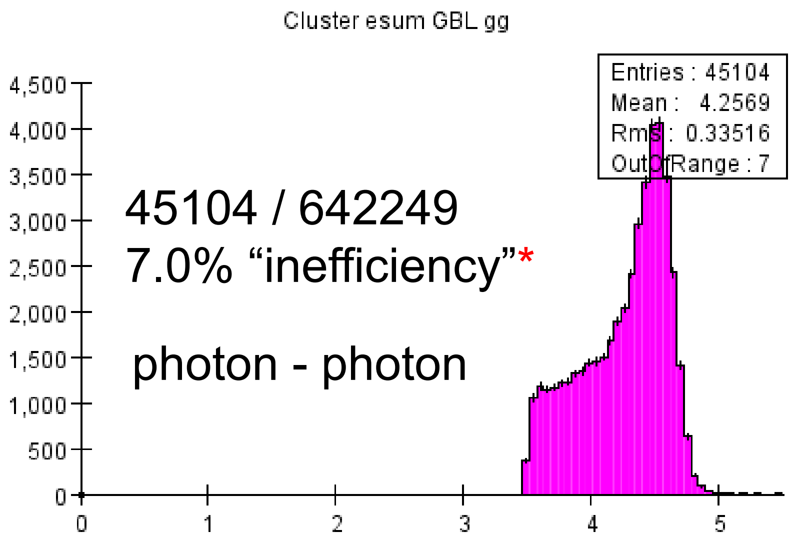
Event Classification

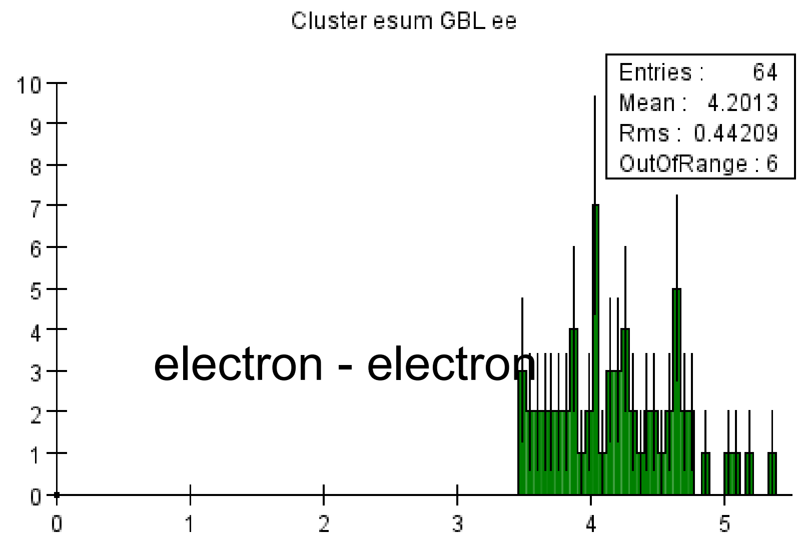
- Hypothesis is that these events are wide-angle bremsstrahlung (WAB) candidates where we have detected both the inelastically-scattered electron and the radiated photon, $e^-\gamma$.
- Esum should equal beam energy
- One, or the other, of the clusters should have an associated track, the other should not.
- Discard events with a reconstructed positron, as these may be real trident events.
  - $\rightarrow$ 642249 events
- Events reconstructed with two photons is a measure of the track inefficiency.

Event Cluster Types GBL

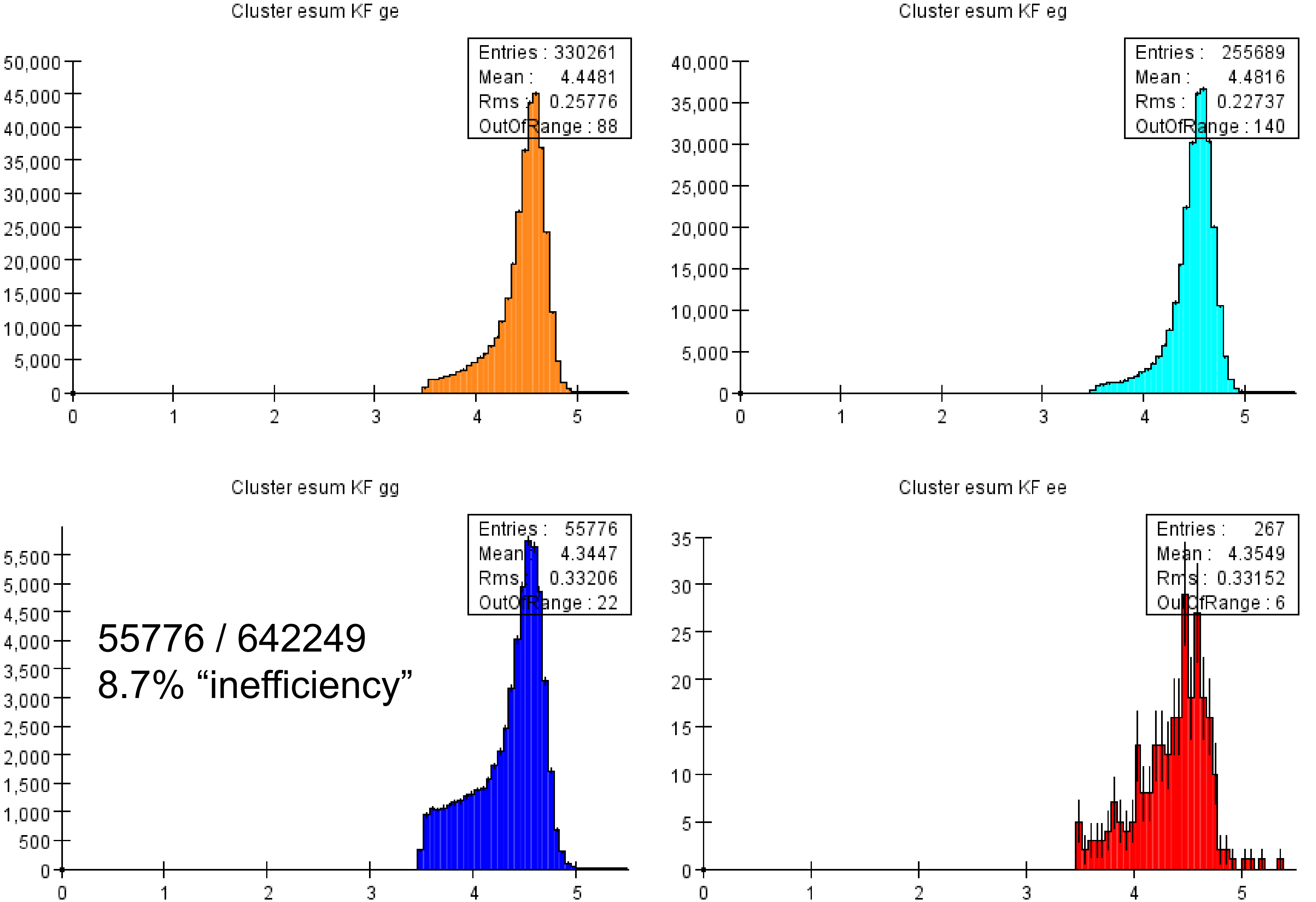






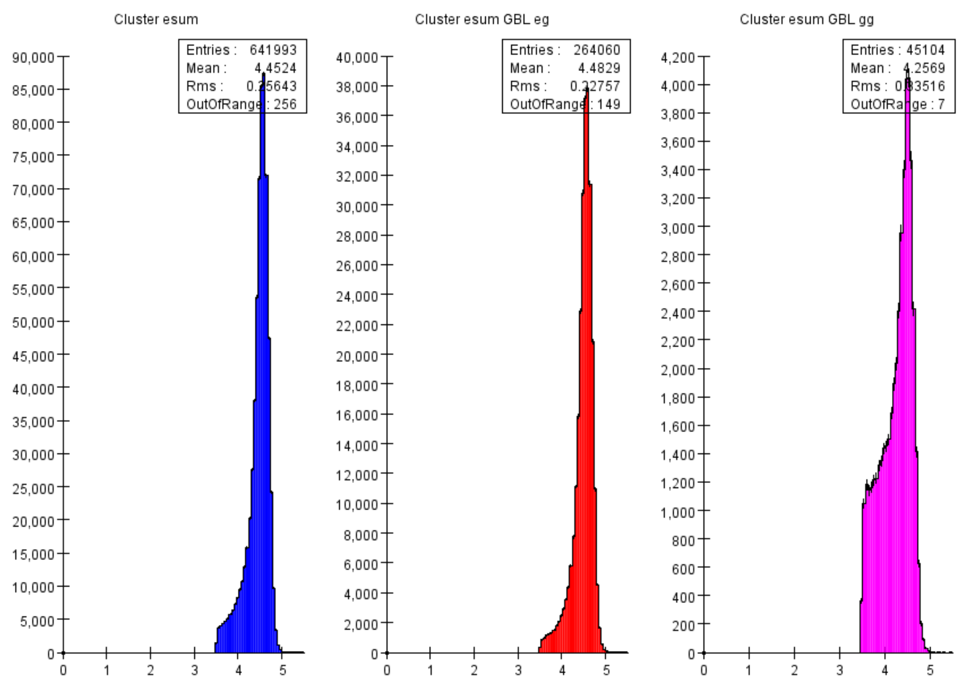


Event Cluster Types KF



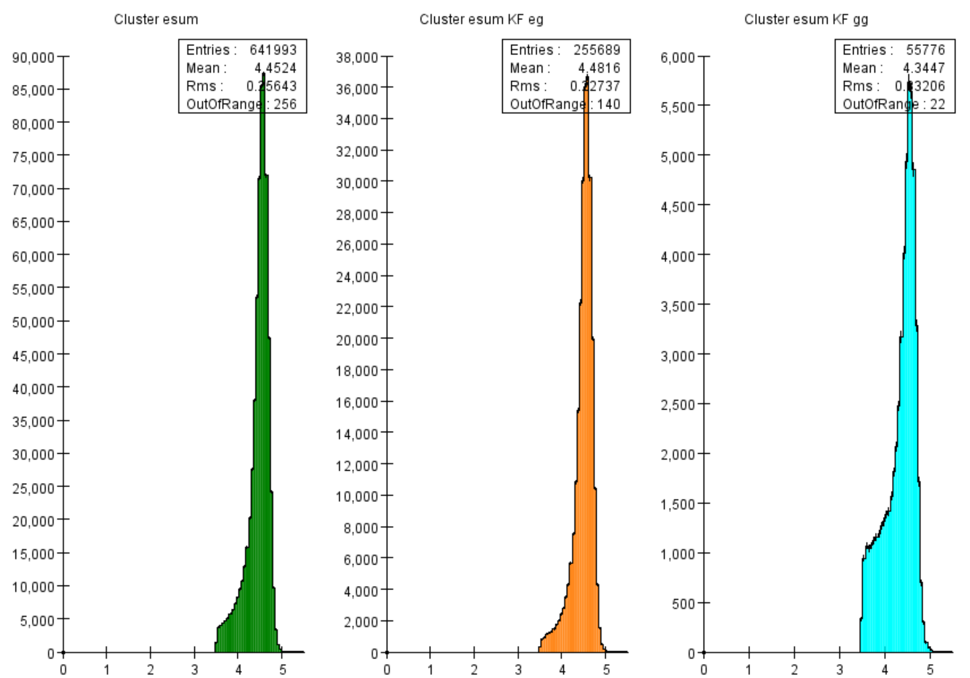
"Inefficiency" GBL

- Measured "inefficiency" is affected by the purity of the parent sample.
- Note that the gg Esum distribution has more of a "porch" at low Esum than the eg (or ge) sample



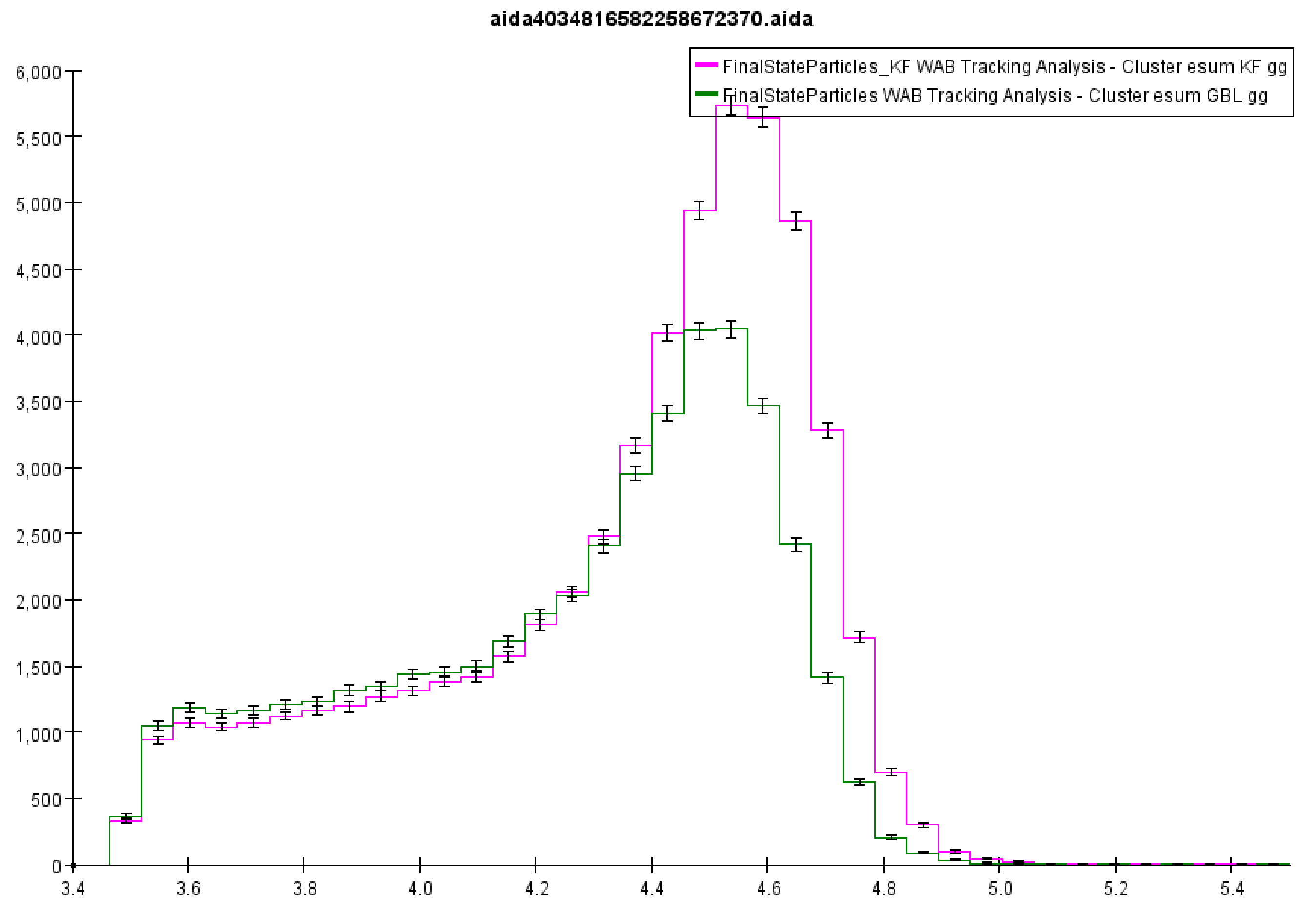
“Inefficiency” KF

- Measured “inefficiency” is affected by the purity of the parent sample.
- Note that the gg Esum distribution has more of a “porch” at low Esum than the eg (or ge) sample



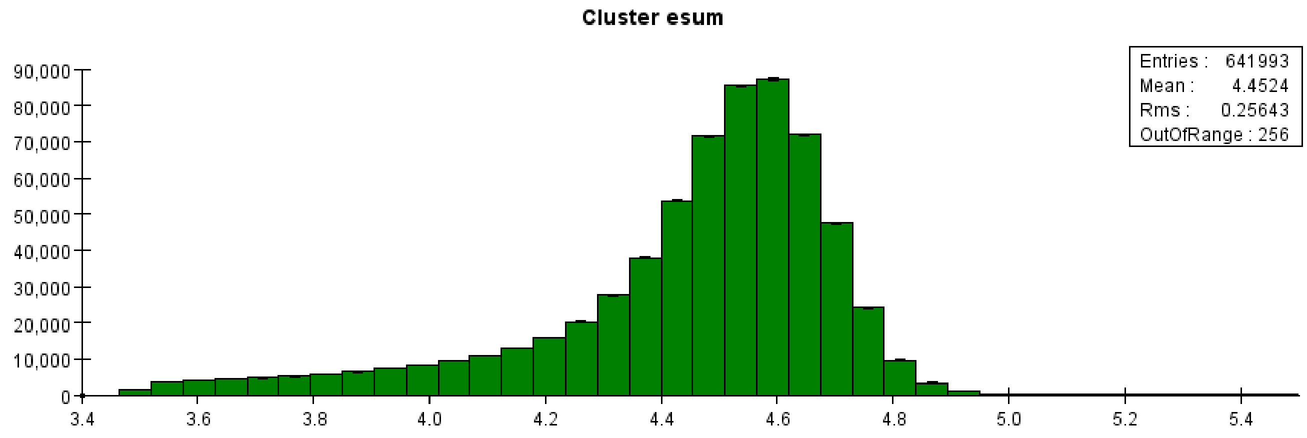
gg Esum

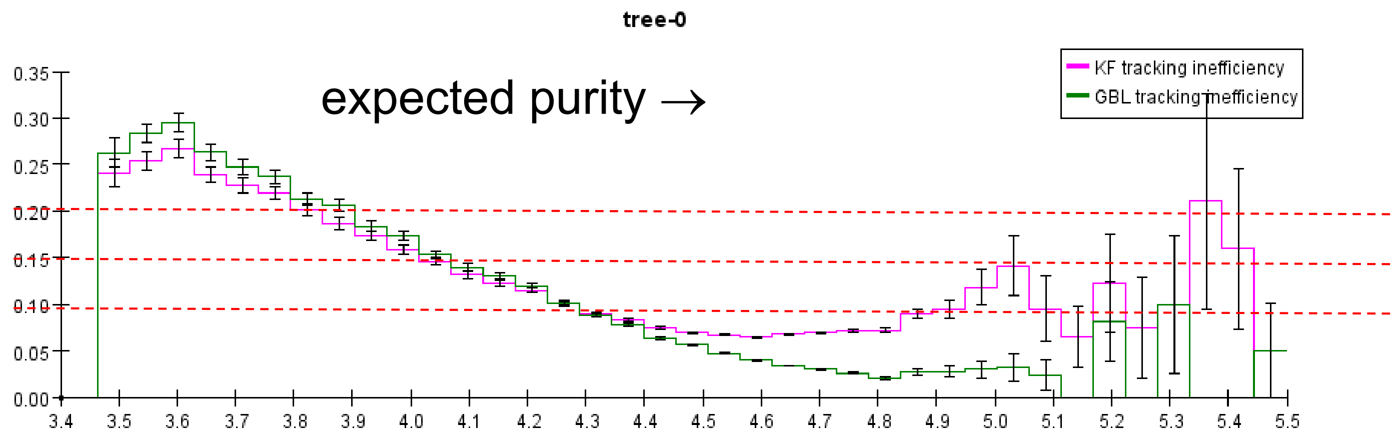
- Esum in events with no track matched to either cluster, “gg”



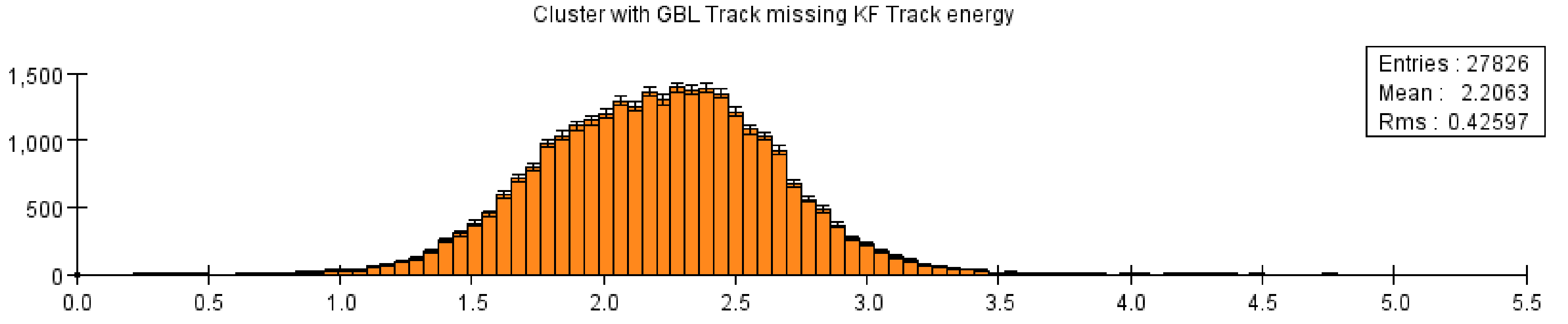
"Inefficiency" vs Esum

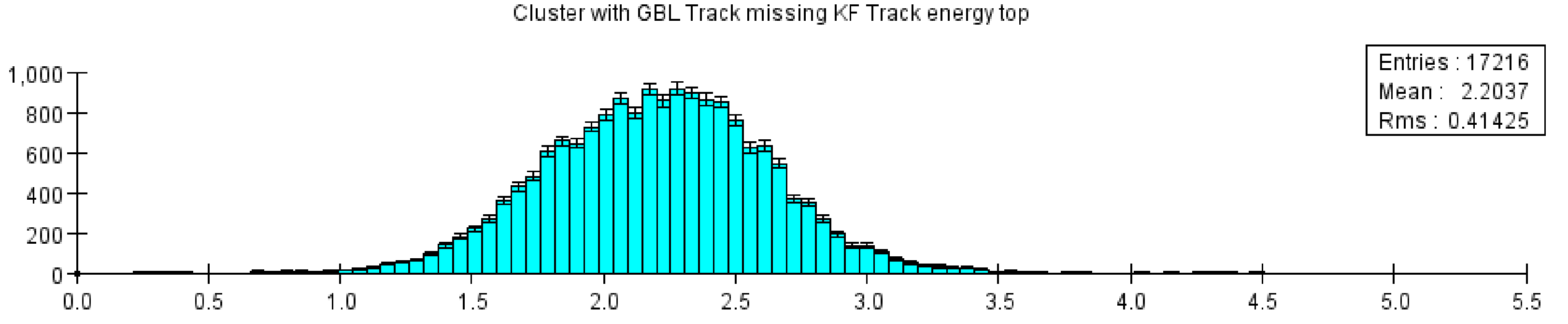
- Expect purity of the parent sample to increase as Esum nears beam energy.

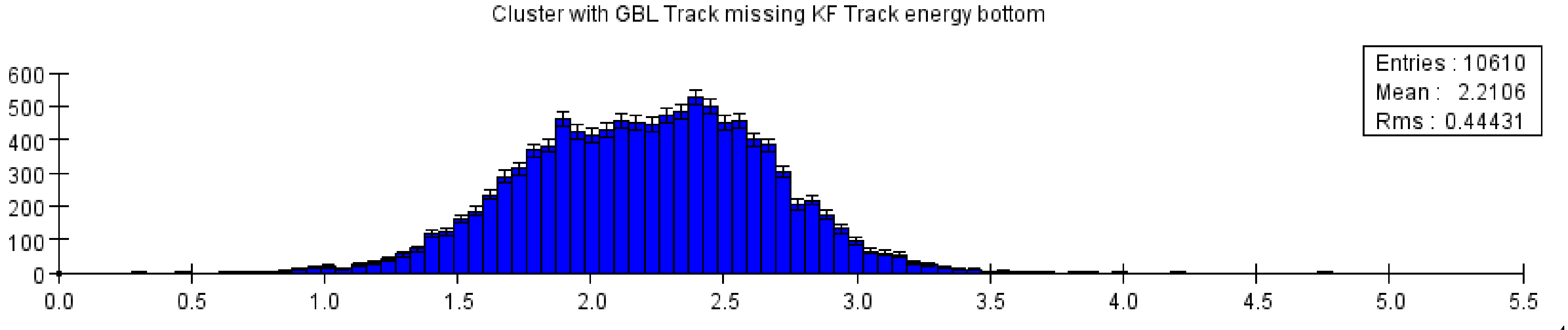




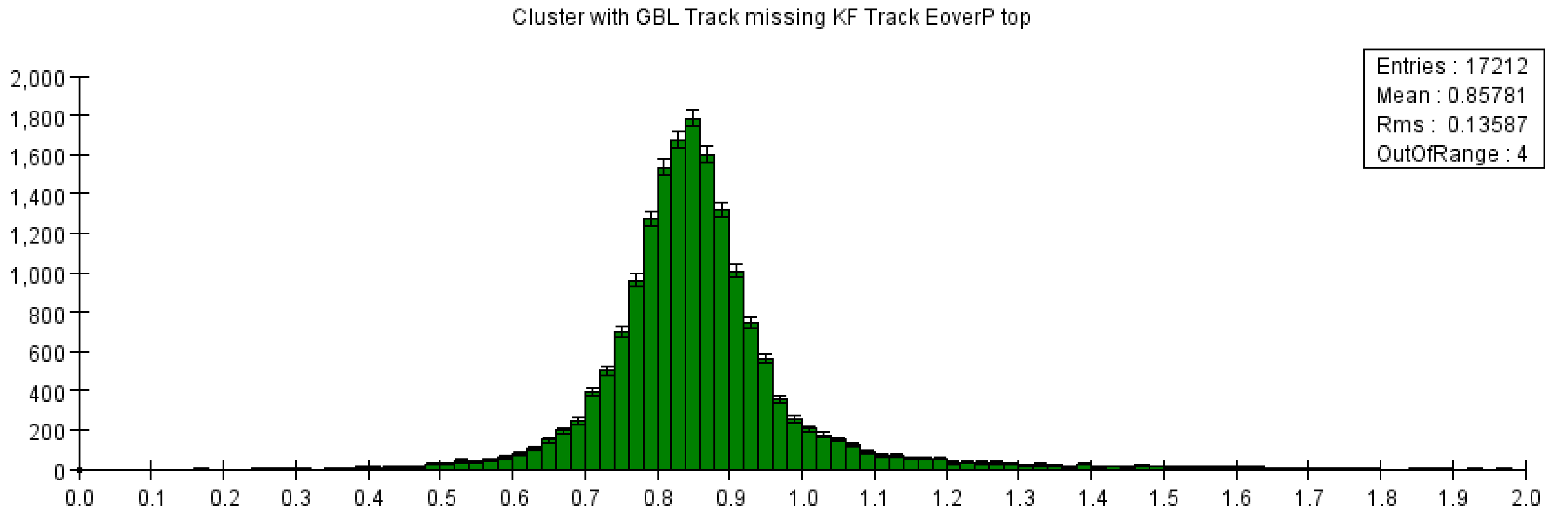
Events with GBL but no KF track

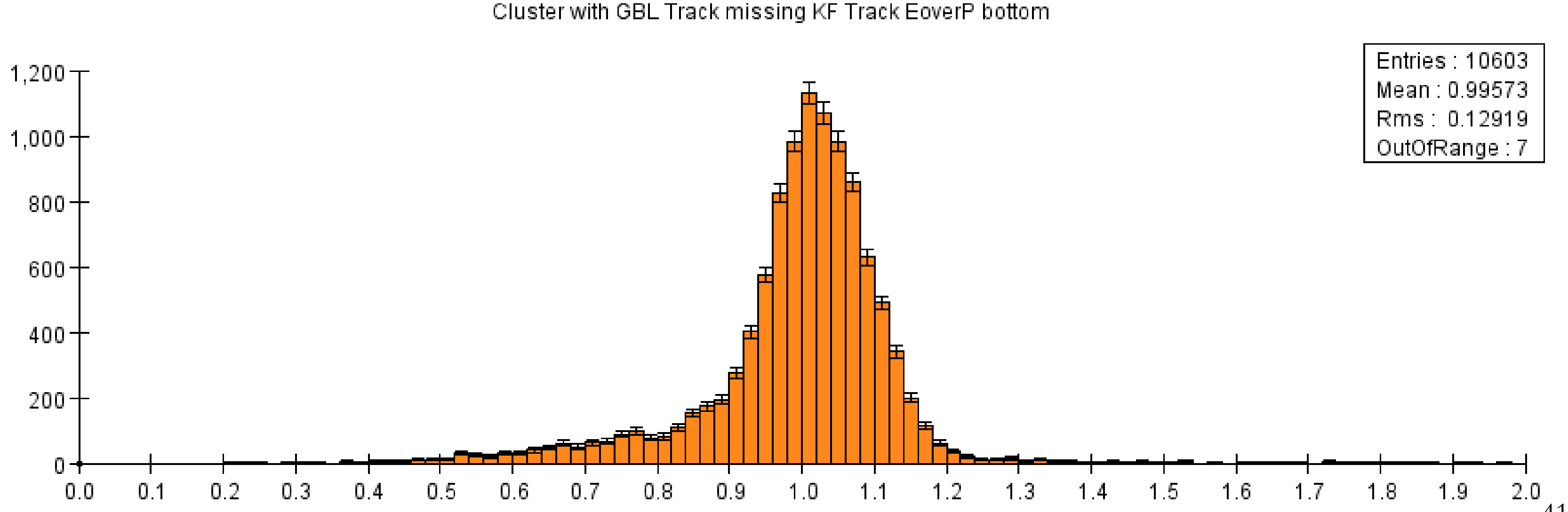




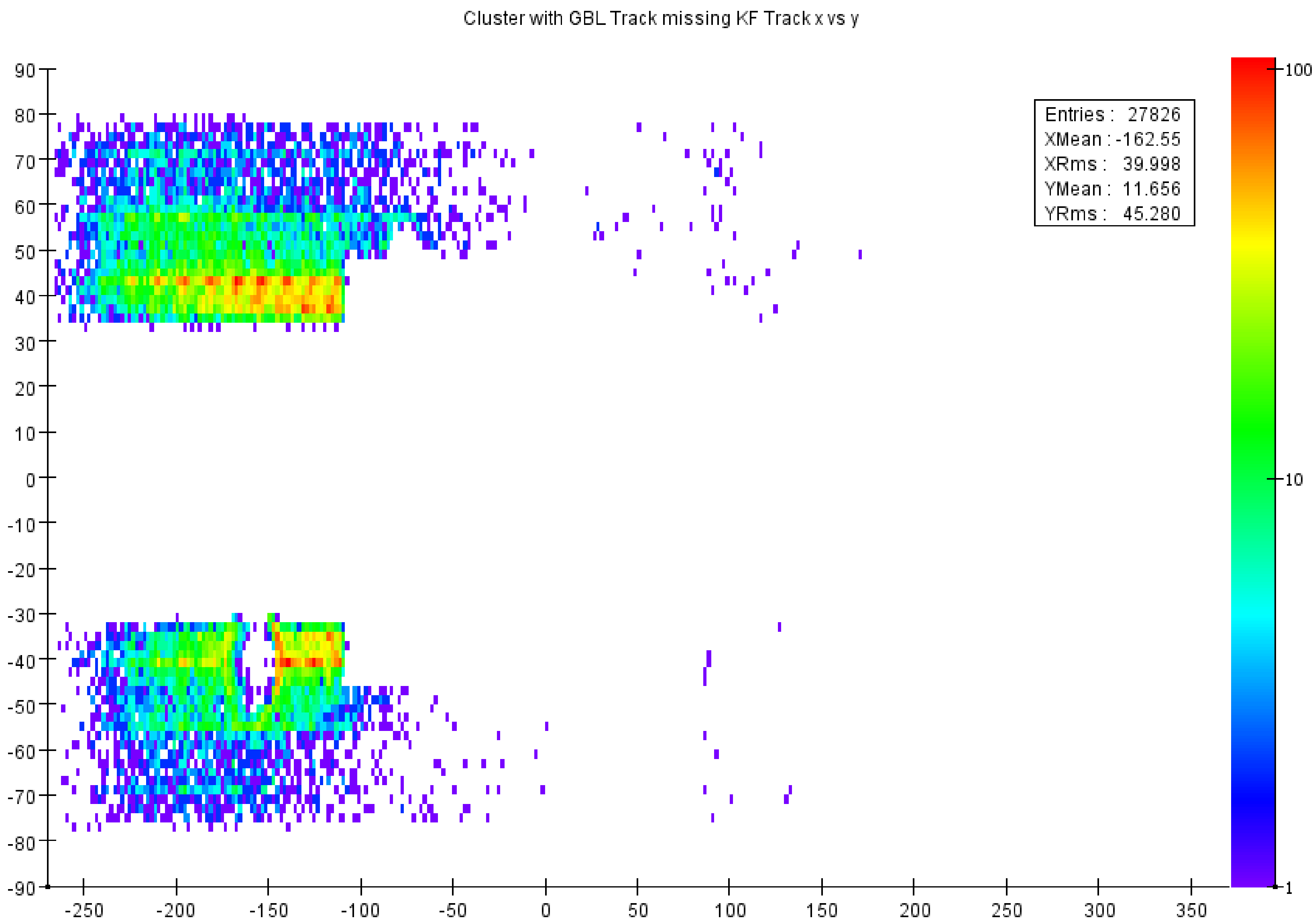


Events with GBL but no KF track





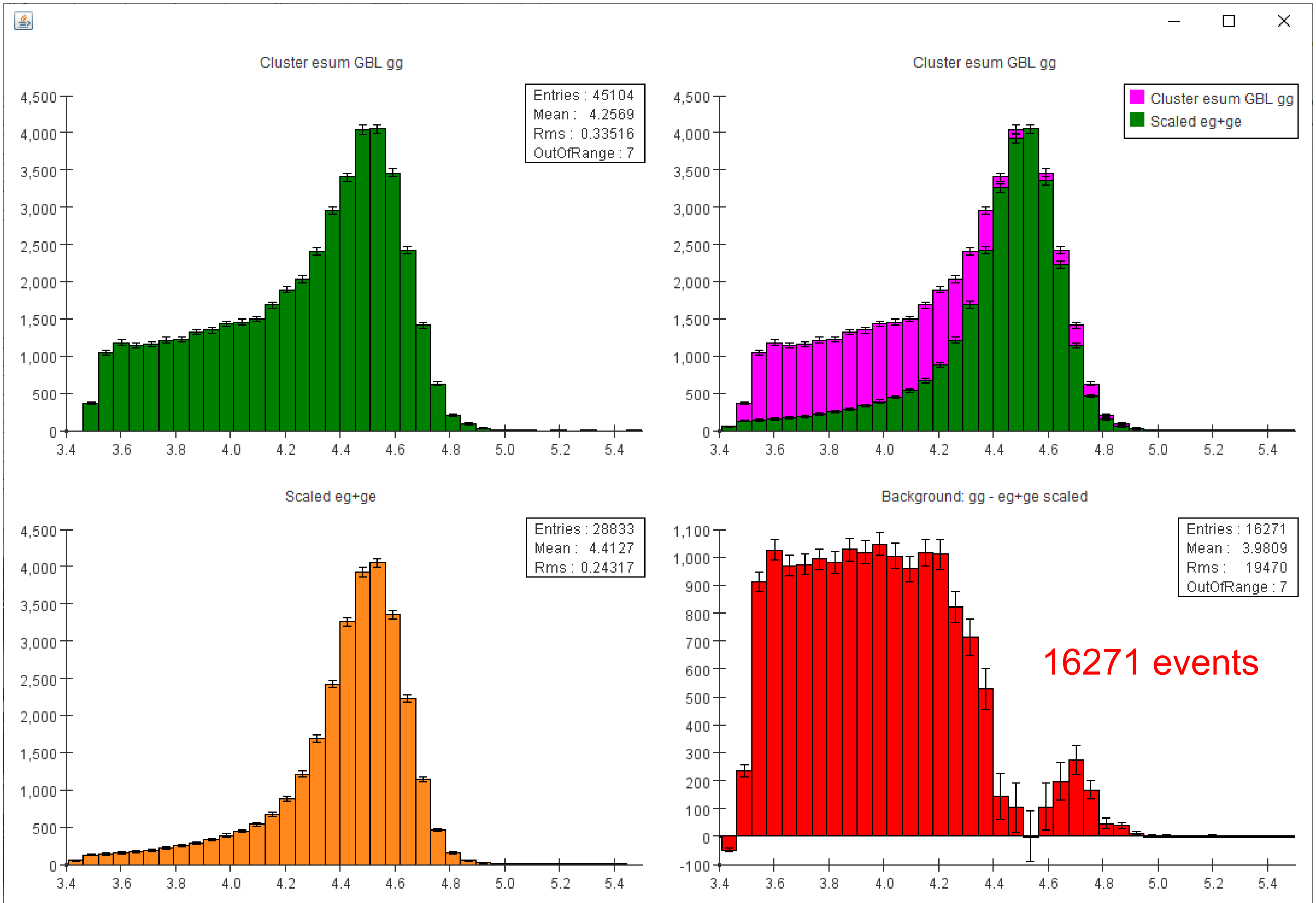
Events with GBL but no KF track



Track-Finding Efficiency Using Data

- Difference in shape of $e^-\gamma$ and $\gamma\gamma$ esum distribution points to evidence for non-WAB background.
- Sum $e^-\gamma$ and γe^- histograms.
- Subtract this WAB distribution from the $\gamma\gamma$ histogram after scaling to match the peak height.
- Provides estimate of non-WAB background contribution.

Track-Finding Efficiency Using Data



Track-Finding Efficiency Using Data

- Subtracting a scaled WAB esum distribution from the γγ esum distribution results in a flat background "porch" distribution.
- Can cut tighter on esum to get better purity, but want to study background as a function of esum
- Better estimate of the tracking efficiency for WABs in run 31 is now:

$$1-(45104-16271)/642249=95.5\%$$

Track-Finding Efficiency Using WABs

- Preliminary study of the track-finding efficiency using WABS indicates an efficiency of finding Seed Tracker / GBL tracks in the momentum range between 1 and 3.5GeV of ~95.5% for run 31.
- Seed Tracker appears to have a slightly higher efficiency of finding tracks than the Kalman Filter.
- 27826 events having one GBL track matched to an Ecal cluster but no KF track have been skimmed.
- These events fall into at least three categories:
  - Events with a nearby KF track, but the track was not matched to the cluster
    - may have to tweak some settings in track-cluster matcher
  - Events with KF tracks, but have picked up wrong hits in the earlier layers and have very low momentum
  - Events with no KF tracks at all
- Quite often there are essentially duplicate GBL tracks
  - MergeTrackCollections needs to be revisited.

Trident Momenta

- Similar studies of track-finding efficiency, E/p, etc. can be performed on the trident samples.
- Also provide a clean sample of positrons which can be used to study the trigger and the hodoscope performance.
- Can also use the tridents themselves as proxies for the beam to understand the alignment of HPS with respect to the beam direction.

Track Calibration

- The FEE and WAB samples, selected with only calorimeter cuts, can be used to add momentum constraints to the alignment procedures.
- Can also be used to characterize the performance of proposed detector alignments by comparing cluster energy to track momenta.
- These samples can also be used for a wide variety of other tracking studies, such as track-finding efficiencies.

Mass Calibration

- The search for A's at HPS ultimately relies on mass calculations, so it would be nice if we could identify any processes in the data which would allow a mass calibration or at least a check.
- $\pi^0 \rightarrow \gamma\gamma$ would have been a nice check of the Ecal energy calibration. I reported on the null results of my search quite some time ago.
- $\phi \rightarrow K^+K^-$ would have been a nice check of the SVT momentum calibration and alignment. I reported on the null results of my search quite some time ago.
- Unfortunately, direct calibration / check of our mass scale and resolution is not possible.

2019 Data Reconstruction

- The 2021 run is fast approaching.
- Do we intend to process the full 2019 data set before we accumulate the 2021 set?
- Where are we?
- Where do we need to be?
- When do we need to be there?

2019 “Good” Runs

- We have a preliminary list of 282 “good” runs broken into 278052 file partitions
- The “sample partitions” are 867 files ending in 041 and 042 which are intended as a faithful subset of the full run (~3‰).
  - More information can be found on confluence.
- Processed at JLab using:
  - Recent snapshot of hps-java
  - PhysicsRun2019FullRecon_pass0.lcsim
    - Runs both SeedTracker/GBL and Kalman Filter
    - Fits SVT data, but does not run tracking over events with greater than 200 SVT strip clusters (aka “monster” events)
  - HPS_PhysicsRun2019-v2-FEE-Pass0

Reconstruction Times

- For each run:
  - average the CPU times over the number of sample partitions for that run
  - multiply this average times the total number of partitions for that run
- Total time is then the sum over all the runs of the average time to reconstruct one partition times the number of partitions for that run.
- Total time: 1.56809e+06 CPU hours
- In good agreement with the estimate (1.3e+06) that Nathan had made.

- 1.6M CPU hr/ (2.6k CPU)/ (24 hr/day) = 26 days at HPS Hall B fairshare
- 1.6M CPU hr/ (5.2k CPU)/ (24 hr/day) = 13 days at full Hall B fairshare

Recon Timing @ JLab

- Quite a bit of run-to-run variability
  - factor of ~3 between runs 22 and 515
- Quite a bit of “options” variability
  - factor of ~1.5 between keeping/skipping “monster” events
- Still some work to be done to release code
- Still some work to be done to finalize the reconstruction steering files
- But we appear to be in the ballpark in terms of reconstruction time.

Output File Size

- Critical path now shifts to output file size.
- We start with 278052 file partitions at 2GB giving us 556.104 TB
- What information do we NEED?
- What information do we WANT?
- How much can we keep disk-resident?
  - All the output files?
  - Only skims?
  - If so, which skims? trigger? recon?
- Do we want to be able to re-run (some of) the reconstruction on Icio output?

Output File Size

- Start by dropping whole collections
- If needed, drop individual objects from remaining collections
- If needed add extra collections or extra information to existing collections (e.g. TrackData)
- Have investigated the following scenarios:
  - Drop all “raw” hit collections
  - If we don’t run ST/GBL, then we drop a number of hit and track collections automatically
  - Drop SVT fitted hits, keep only 1D strip clusters
  - Work our way up the chain…

Input Needed

- First crude pass to investigate what can be done easily.
- Instead of dropping all SVT strip clusters, could also drop individual strips not in the fiducial region of the track-finding (low amplitude, early/late times, physical regions)
- Could also only keep skims on disk.
  - trigger skims? recon skims?
- Input from analysis group and individuals doing analysis is clearly needed.
- Ideally we would like to have output size ~10%

Short-term Plans

- We have been repeatedly processing the "sample partitions" from each of the "good" runs.
- SVT group has requested a larger fraction of the data to be processed in order to determine baseline calibrations
  - 20 partitions vs few (2-6) per run
- Having a larger reconstruction sample available would also allow other variables to be accumulated, such as beam spot and beam tilt and track-finding efficiencies, etc., on a run-by-run basis
- Not quite a 10% subset, but sufficiently large to constitute the "unblinded" sample on which to develop our physics analyses.

Next Steps

- Get latest code development into master branch
  - At least two pull requests need to be approved ASAP
- Review steering file and detector to be used
- hps-java needs to be released
- Release any other software to be run as part of this process, e.g skims, tuple generation, etc.
- Develop list of run partitions to be processed
  - SVT, Analysis & Recon groups, individuals
- Develop scripts, identify resources and individuals to oversee processing.
- Analyze the data!
- Iterate if necessary
- Repeat for full data set.

Next Steps

- Most importantly we need your feedback!
- Help us align and calibrate the detector!
- Analyze the data that's already available.
  - Track-finding efficiencies & new strategies
  - Track-cluster matching
  - Energy & Momentum scales and resolutions
  - V0 efficiencies, purity and backgrounds
- Let us know what data you absolutely NEED in the reconstruction output.