Vertex preselection

for data and Monte Carlo

Sarah Gaiser
Stanford/SLAC
HPS Collaboration Meeting June 2025

U.S. DEPARTMENT OF ENERGY | Stanford University
SLAC NATIONAL ACCELERATOR LABORATORY

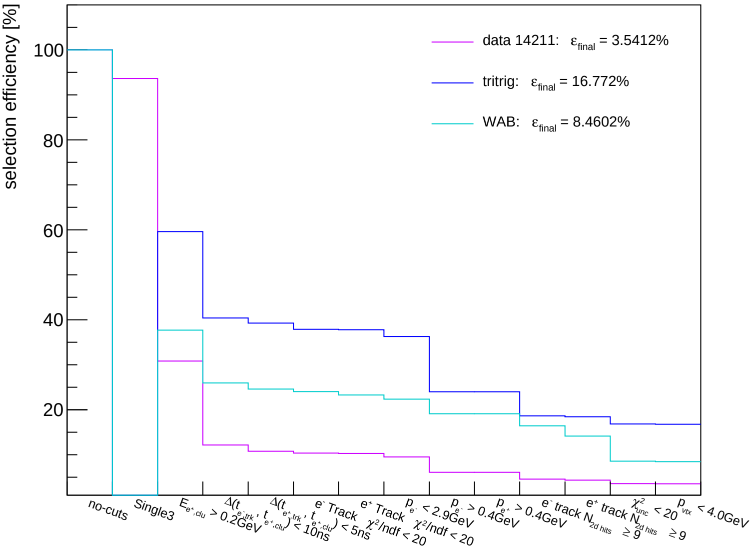
Outline

1. Introduction
2. Side-band subtraction
3. Effect on invariant mass and A' signal
4. Track quality in side-bands and signal region
5. Dependence on tracking volume
6. Requiring an electron ECAL cluster
7. Radiative acceptance and fraction

Data samples and preselection cuts

- Using MC from ‘new’ pass v7
  - tritrig+pulser, WAB+pulser
- Using data from run 14211 from pass v7
  - Data luminosity of full run: $\mathcal{L} = 0.940\,424\,\mathrm{pb}^{-1}$
- All cuts will refer to the following list of cuts:
  - Track cuts
    - $0.4\,\mathrm{GeV} < p_{e^-} < 2.9\,\mathrm{GeV}$
    - $0.4\,\mathrm{GeV} < p_{e^+} < 4.0\,\mathrm{GeV}$
    - $\Delta(t_{\mathrm{trk},e^-}, t_{\mathrm{clu},e^+}) < 10\,\mathrm{ns}$
    - $\Delta(t_{\mathrm{trk},e^+}, t_{\mathrm{clu},e^+}) < 5\,\mathrm{ns}$
    - $\chi^2_{\mathrm{trk}} < 20$
    - $N_{\mathrm{2d\ hits}} \geq 9$
  - Vertex cuts
    - $p_{\mathrm{vtx}} < 4.0\,\mathrm{GeV}$
    - $\chi^2_{\mathrm{vtx}} < 20$
  - Cluster cuts
    - $0.2\,\mathrm{GeV} < E_{\mathrm{clu},e^+} < 4.0\,\mathrm{GeV}$
  - Requiring single3 trigger for the event

Preslection cutflows



Data samples – MC luminosity scaling

- Calculate effective cross section for MC samples

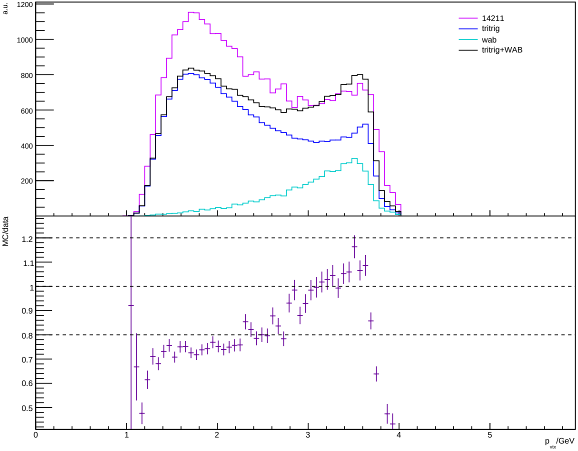
$$\sigma_{\text{eff}} = \sigma_{\text{gen}} \times \frac{N_{\text{selected}}}{N_{\text{gen}}}$$

- Use data luminosity (for 4 files: $\mathcal{L}_{\text{subset}} = 0.001\,492\,\text{pb}^{-1}$) and cross section to scale MC histograms with scaling factor f_{scale}

$$f_{\text{scale}} = \frac{N_{\text{MC, scaled}}}{N_{\text{selected}}} = \frac{\sigma_{\text{eff}} \times \mathcal{L}_{\text{subset}}}{N_{\text{selected}}} = \frac{\sigma_{\text{gen}} \times \mathcal{L}_{\text{subset}}}{N_{\text{gen}}}$$

sample	$\sigma_{\text{gen}}/\text{pb}$	N_{gen}	f_{scale}
tritrig	4.025×10^{9}	$17\,792 \times 10^{4}$	0.0337
WAB	8.249×10^{10}	$38\,700 \times 10^{4}$	0.318
14211	–	–	1.00

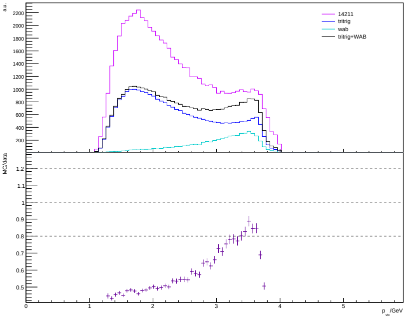
MC – data comparison: p_{sum} with all cuts



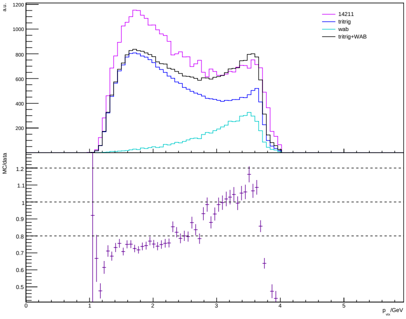
- Seems like we are underestimating the number of events in the low p_{sum} region in MC.

MC – data comparison: effect of $\Delta(t_{\mathrm{trk},e^-}, t_{\mathrm{clu},e^+})$ cut

no $\Delta(t_{\mathrm{trk},e^-}, t_{\mathrm{clu},e^+})$ cut



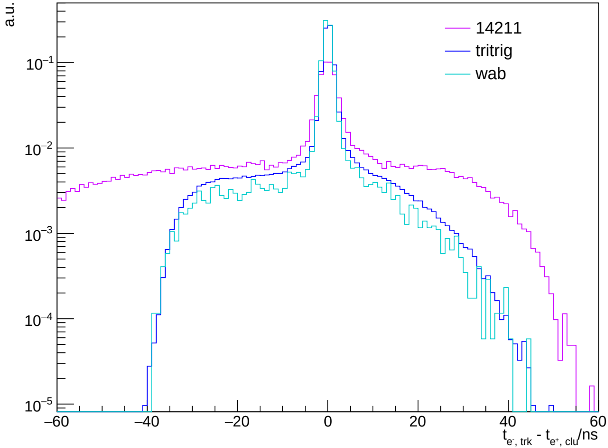
full preselection applied



- For reference: p_{sum} distributions with all cuts applied except for $\Delta(t_{\mathrm{trk},e^-}, t_{\mathrm{clu},e^+})$ cut compared to fully preselected distribution
- Electron time cut removes low p_{sum} events
  - Effect seems to be more pronounced in data

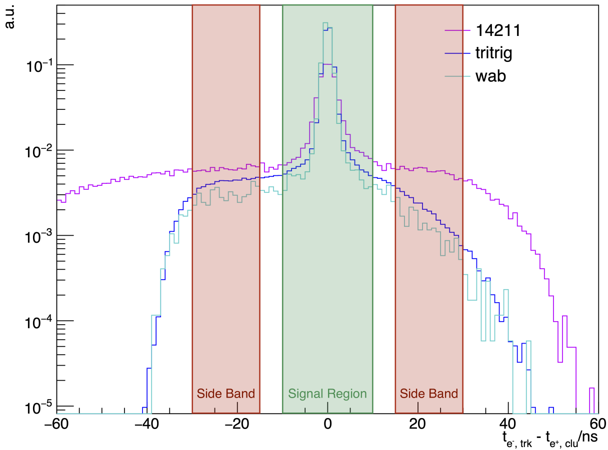
MC – data comparison: $(t_{\text{trk},e^-} - t_{\text{clu},e^+})$ distribution

SLAC



- Idea: excess might come from additional ‘random’ vertices under $(t_{\text{trk},e^-} - t_{\text{clu},e^+})$ time peak in data

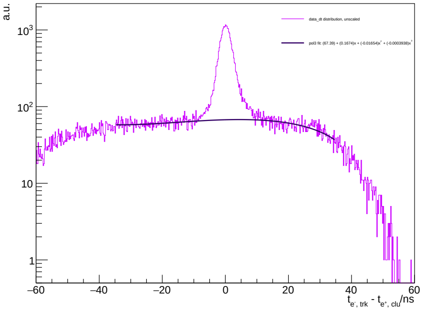
$(t_{\mathbf{trk},e^-} - t_{\mathbf{clu},e^+})$ – signal region and side-bands



- Defining signal region (inside $\Delta(t_{\mathrm{trk},e^-}, t_{\mathrm{clu},e^+}) < 10\,\mathrm{ns}$ cut window) and side-bands $15\,\mathrm{ns} < \Delta(t_{\mathrm{trk},e^-}, t_{\mathrm{clu},e^+}) < 30\,\mathrm{ns}$

Fitting $(t_{\mathbf{trk},e^-} - t_{\mathbf{clu},e^+})$ background

Example distribution – data



- Fit function
  $f_{\text{bkg}} = a + bx + cx^2 + dx^3$
- Excluding the signal region $(\Delta(t_{\text{trk},e^-}, t_{\text{clu},e^+}) < 10\,\text{ns})$ region from fit
- All fits in backup

Determining scaling factors

- Following this explanation
- Find number of events in side-bands $N_{\text{bkg}}^{\text{SB}}$ and signal region $N_{\text{bkg}}^{\text{SR}}$

$$N_{\text{bkg}}^{\text{SB}} = \int_{-30}^{-15} f_{\text{bkg}} + \int_{15}^{30} f_{\text{bkg}} \quad , \quad N_{\text{bkg}}^{\text{SR}} = \int_{-10}^{10} f_{\text{bkg}}$$

- Determine scaling factor ξ_{bkg} between side-bands and signal region

$$\xi_{\text{bkg}} = N_{\text{bkg}}^{\text{SR}} / N_{\text{bkg}}^{\text{SB}}$$

- Rescale $\left.\frac{\mathrm{d}N_{\text{bkg}}}{\mathrm{d}p_{\text{sum}}}\right|_{SB}$ to get estimate in of random background in signal region

$$\left.\frac{\mathrm{d}N_{\text{bkg}}}{\mathrm{d}p_{\text{sum}}}\right|_{SR} = \xi_{\text{bkg}} \cdot \left.\frac{\mathrm{d}N_{\text{bkg}}}{\mathrm{d}p_{\text{sum}}}\right|_{SB}$$

Determining scaling factors

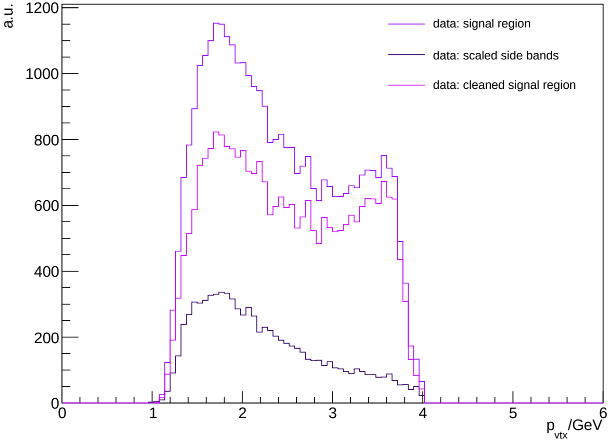
	fit parameters				
sample	a	b	c	d	χ^2/ndf
tritrig	828.00	-12.40	-0.59	0.0081	6.55
WAB	11.06	-0.19	-0.01	0.0001	1.19
14211	67.39	0.17	-0.02	-0.0004	1.22

- Fit for tritrig not ideal, but good enough for now!
  - Tried different functions: pol3 is the best for all three samples
- Determining integrals and scaling factor:

sample	$N_{\text{bkg}}^{\text{SB}}$	$N_{\text{bkg}}^{\text{SR}}$	ξ_{bkg}
tritrig	16 165	15 507	1.042
WAB	216	202	1.07
14211	1336	1761	0.759

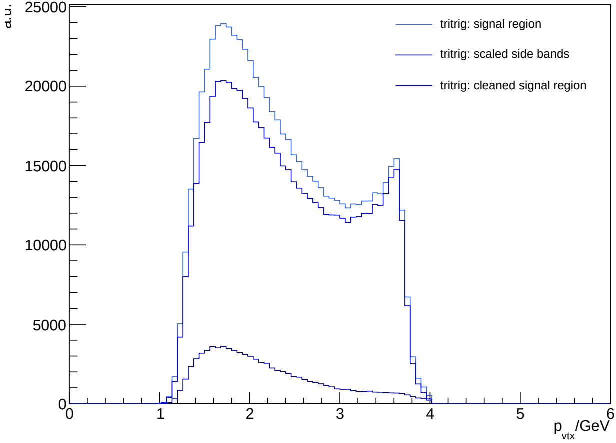
p_{sum} distributions – data

SLAC



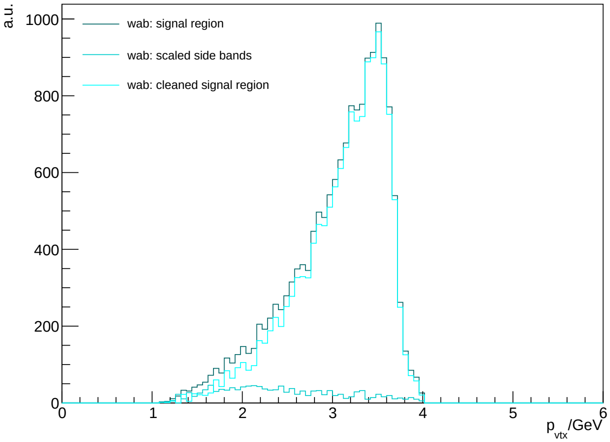
- p_{sum} distribution for data in signal region and side-bands, as well as side-band subtracted distribution in signal region

p_{sum} distributions – tritrig



- p_{sum} distribution for tritrig in signal region and side-bands, as well as side-band subtracted distribution in signal region

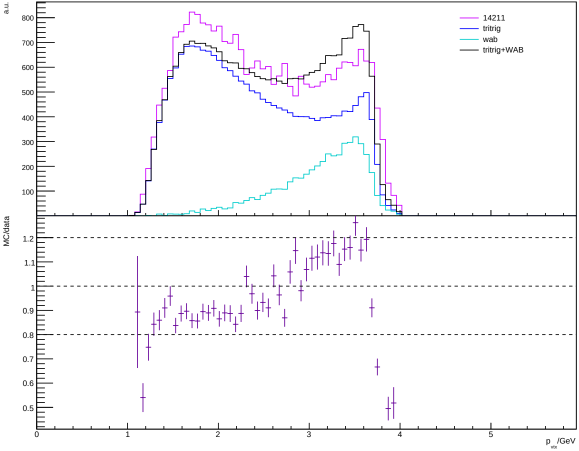
p_{sum} distributions – WAB



- p_{sum} distribution for WAB in signal region and side-bands, as well as side-band subtracted distribution in signal region

MC – data comparison: p_{sum} side-band subtracted

SLAC

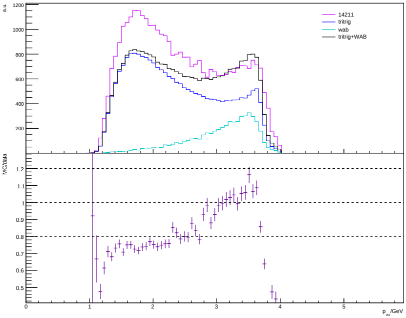


- Side-band subtracted and luminosity scaled MC compared to data
- Achieve higher low p_{sum} count in MC compared to data than before

Summary I

- Simply applying the preselection to data and MC leads to underestimating the number of low-psum events in Monte Carlo.

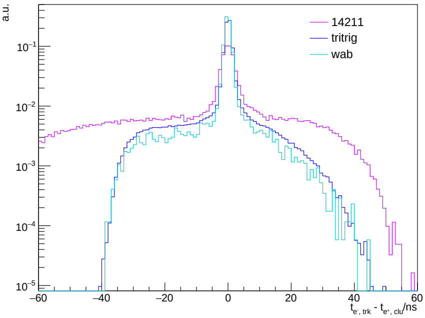
No side-band subtraction, luminosity scaled



Summary I

- Simply applying the preselection to data and MC leads to underestimating the number of low-psum events in Monte Carlo.
- The $(t_{\text{track},e^-} - t_{\text{clu},e^+})$ peak in data sits on a higher plateau of background events.

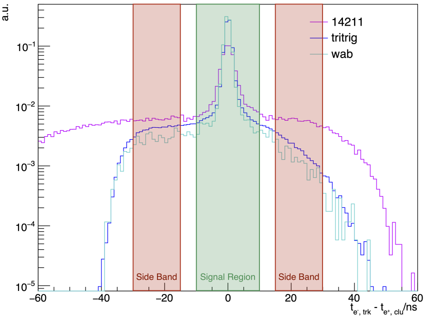
$(t_{\text{track},e^-} - t_{\text{clu},e^+})$ for data, tritrig, and WAB



Summary I

- Simply applying the preselection to data and MC leads to underestimating the number of low-psum events in Monte Carlo.
- The $(t_{\mathrm{track},e^-} - t_{\mathrm{clu},e^+})$ peak in data sits on a higher plateau of background events.
  - Using side-bands to estimate the background under the peak enables us to remove these random events.

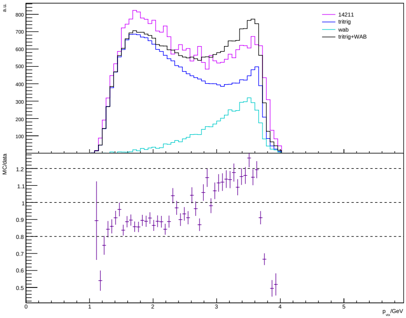
$(t_{\mathrm{track},e^-} - t_{\mathrm{clu},e^+})$ for data, tritrig, and WAB



Summary I

- Simply applying the preselection to data and MC leads to underestimating the number of low-psum events in Monte Carlo.
- The $(t_{\text{track},e^-} - t_{\text{clu},e^+})$ peak in data sits on a higher plateau of background events.
  - Using side-bands to estimate the background under the peak enables us to remove these random events.
- The side-band subtraction equalizes the number of events at low p_{sum} in data and MC.

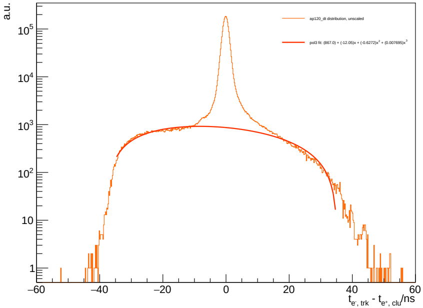
Side-band subtracted, luminosity scaled



Effect on invariant mass and A' signal

- How does the side-band subtraction affect the signal samples?
  - Similar to tritrig and WAB samples, the A' $(t_{\mathrm{trk},e^-} - t_{\mathrm{clu},e^+})$ time peak sits atop random background.

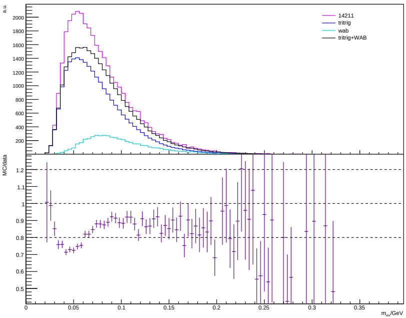
120 MeV A' – all cuts but $(t_{\mathrm{trk},e-} - t_{\mathrm{clu},e+})$ applied



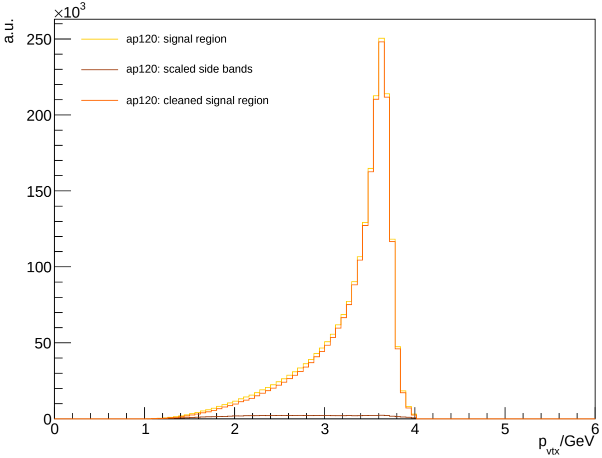
Effect on invariant mass and A' signal

- How does the side-band subtraction affect the signal samples?
  - Similar to tritrig and WAB samples, the A' $(t_{\mathrm{trk},e^-} - t_{\mathrm{clu},e^+})$ time peak sits atop random background.
- What is the effect of the side-band subtraction on the invariant mass distributions?
  - Taking a look at data, luminosity scaled tritrig and WAB samples, as well as 120 MeV A' signal

No side-band subtraction, luminosity scaled

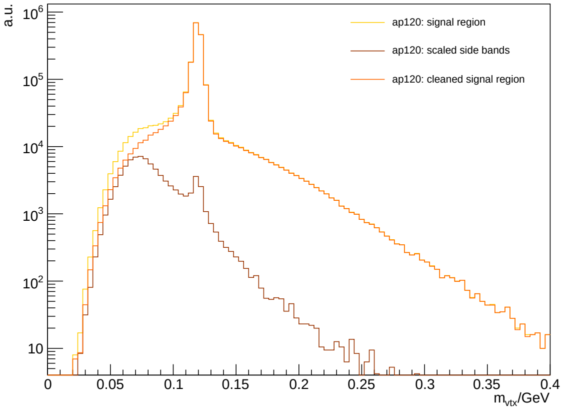


p_{sum} distributions – 120 MeV A'



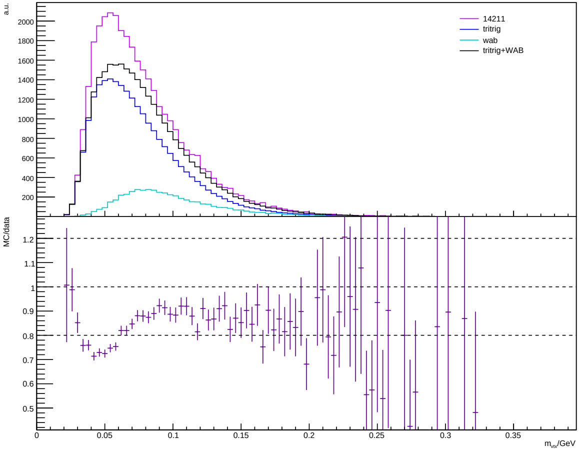
- p_{sum} distribution for A' in signal region and side-bands, as well as side-band subtracted distribution in signal region

m_{vtx} distributions – 120 MeV A'



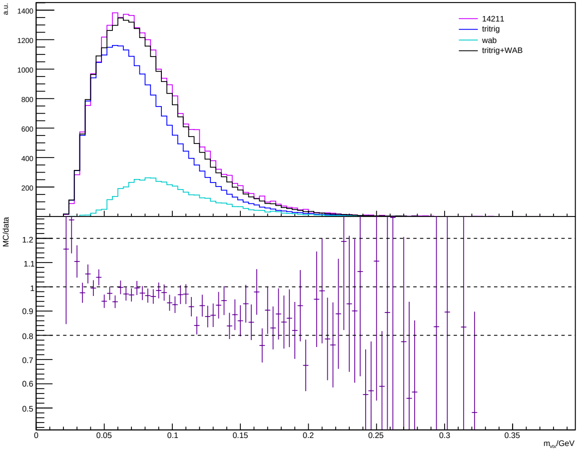
- m_{vtx} distribution for A' in signal region and side-bands, as well as side-band subtracted distribution in signal region
  - Side-band subtraction removes only very few events from invariant mass peak.

MC – data comparison: m_{vtx} with all cuts



- Without side-band subtraction MC underestimates the number of events in the peak of the invariant mass distribution.

MC – data comparison: m_{vtx} side-band subtracted



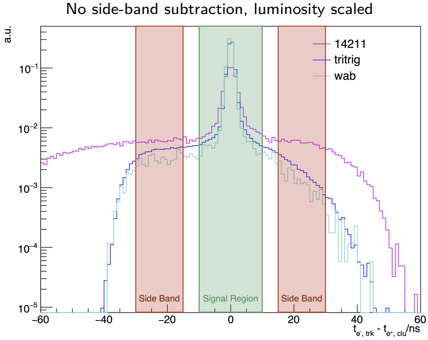
- While shape of p_{sum} does not agree fully after side-band subtraction, the invariant mass distribution looks pretty good.

Summary II

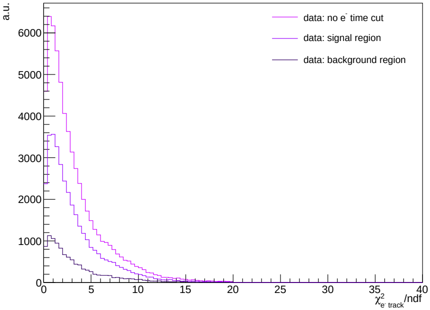
- Side-band subtraction preserves A' p_{sum} and invariant mass distribution
  - Only very few events are subtracted from invariant mass peak at $m_{A'} = 120\,\text{MeV}$
- After side-band subtraction, the tritrig+WAB MC invariant mass distribution follows the data distribution more closely.
- Overall, side-band subtraction seems to remove discrepancies between data and MC without introducing obvious new problems.

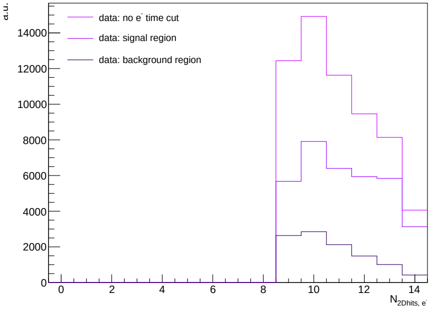
Track quality in side-bands and signal region

- Why do we have more low p_{sum} events in data than in MC?
- 'Excess' low p_{sum} events in data seem to be part of the random time background under the $(t_{\text{track},e^-} - t_{\text{clu},e^+})$ peak.
- Do these background events have any other distinct characteristics?



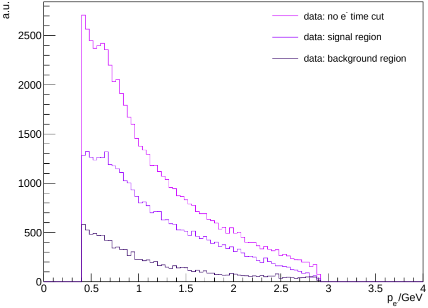
$\chi^2_{\mathbf{trk},e^-}/$ndf and $N_{\mathbf{2d\ hits},e^-}$ distributions – data

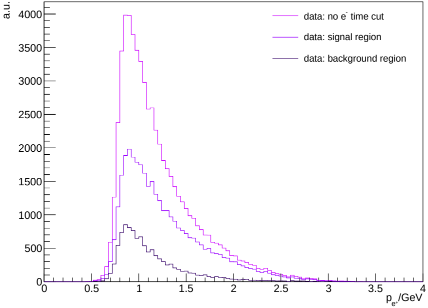




- $\chi^2/$ndf looks similar in signal region and side-bands
  - Perhabs slightly higher $\chi^2/$ndf in SB?
- Side-bands have more low $N_{2\text{d hits},e^-}$ tracks compared to signal region

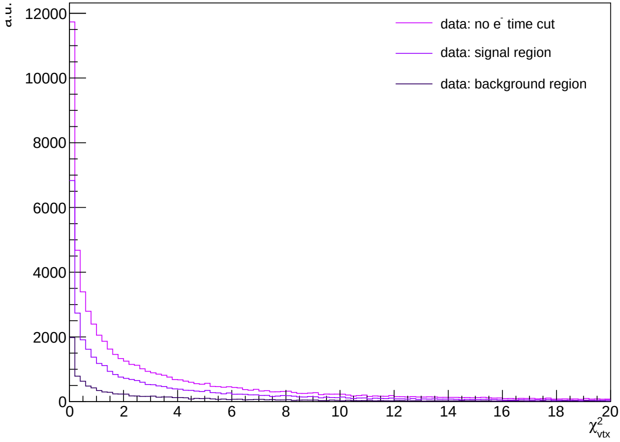
p_{e^-} and p_{e^+} distributions – data





- The slight double-peak structure in the uncut p_{e^-} distribution comes from a plateau at low p_{e^-} in the signal region and a rising contribution from the side-bands.
- The positron momentum looks similar across the sample regions

$\chi^2_{\mathbf{vtx}}$ distribution – data



- The vertex fits in the side-bands are not obviously worse than in the signal region.

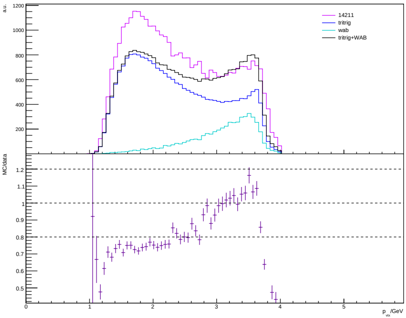
Summary III

- The higher contribution of random in-time background in data might come from more low-momentum and low $N_{\text{2d hits},e^-}$ in data compared to MC.
  - Side-bands have tracks with lower $N_{\text{2d hits},e^-}$ than in the signal region.
  - The electron momentum distribution seems to peak more strongly towards low values in the side-bands.
- Does this come from reconstruction inefficiencies in data?
- Or is our MC not including enough low-momentum tracks?

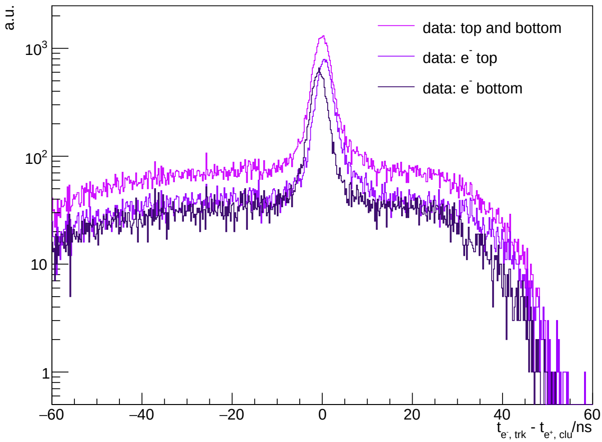
Dependence on tracking volume

- At the moment: alignment of bottom volume of SVT is better than top
- If tracking inefficiencies in data are causing an excess of events at low p_{sum}, we should see a difference in the distributions between top and bottom volume.

No side-band subtraction, luminosity scaled



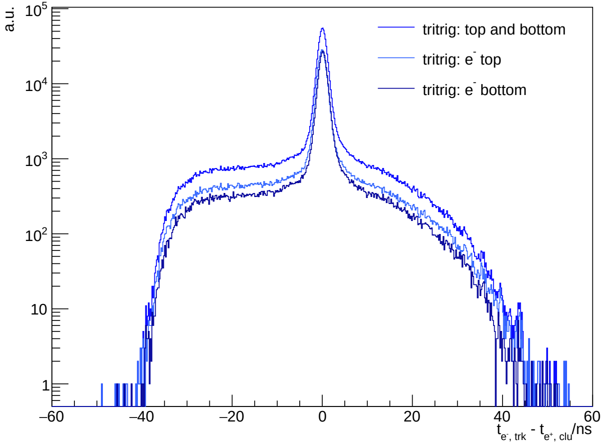
Tracking volume – data $(t_{\mathbf{trk},e^-} - t_{\mathbf{clu},e^+})$



- Applying all cuts except for $\Delta(t_{\mathrm{trk},e^-}, t_{\mathrm{clu},e^+})$ cut
- Time peaks are slightly shifted for electron in top or bottom

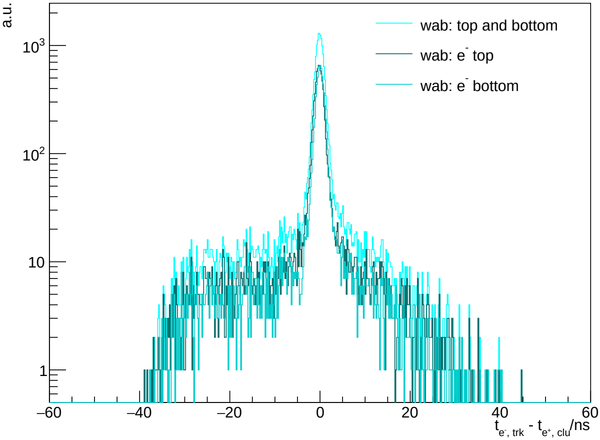
Tracking volume – tritrig ($t_{\mathrm{trk},e^-} - t_{\mathrm{clu},e^+}$)

SLAC



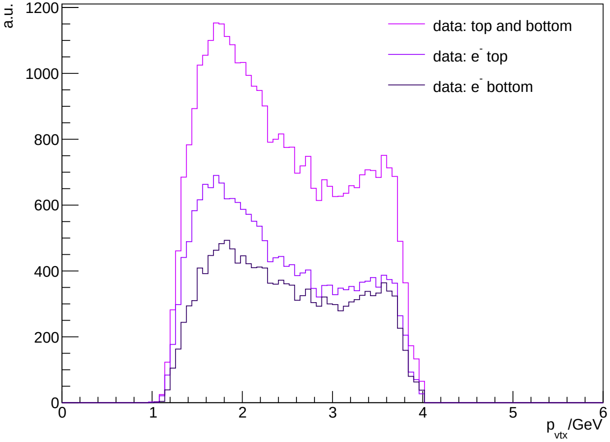
- Applying all cuts except for $\Delta(t_{\mathrm{trk},e^-}, t_{\mathrm{clu},e^+})$ cut
- Top and bottom time peaks mostly overlap for MC

Tracking volume – WAB $(t_{\mathbf{trk},e^-} - t_{\mathbf{clu},e^+})$



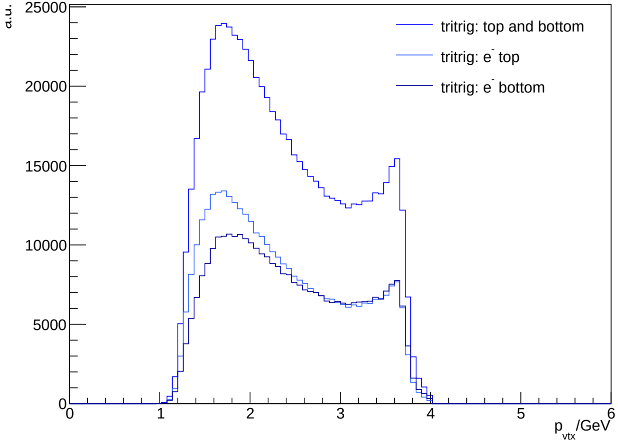
- Applying all cuts except for $\Delta(t_{\mathrm{trk},e^-}, t_{\mathrm{clu},e^+})$ cut
- Top and bottom time peaks mostly overlap for MC

Tracking volume – signal region



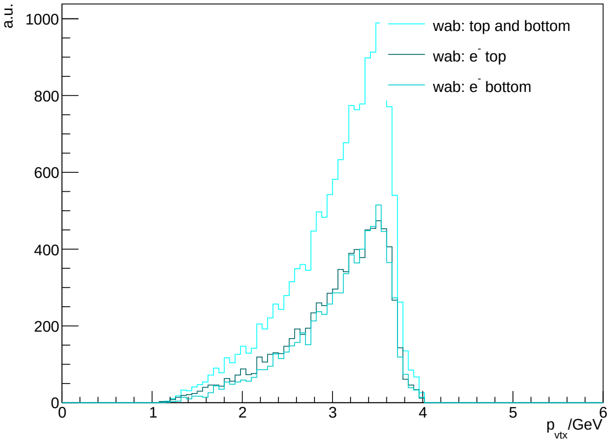
- Applying all cuts and splitting sample into top $\tan\lambda_{e^-} > 0$ and bottom $\tan\lambda_{e^-} < 0$

Tracking volume – signal region



- Applying all cuts and splitting sample into top $\tan\lambda_{e^-} > 0$ and bottom $\tan\lambda_{e^-} < 0$

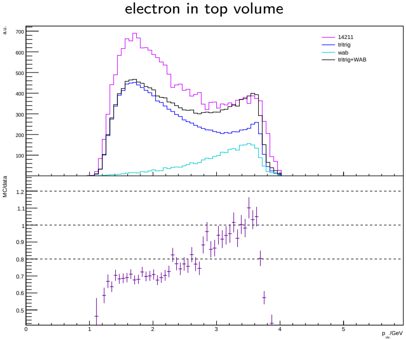
Tracking volume – signal region

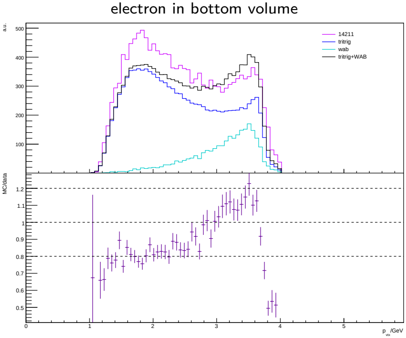


- Applying all cuts and splitting sample into top $\tan\lambda_{e^-} > 0$ and bottom $\tan\lambda_{e^-} < 0$

MC – data comparison: no side-band subtraction

SLAC

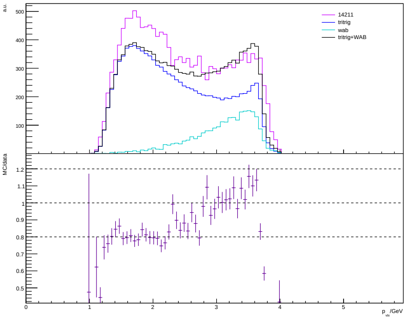




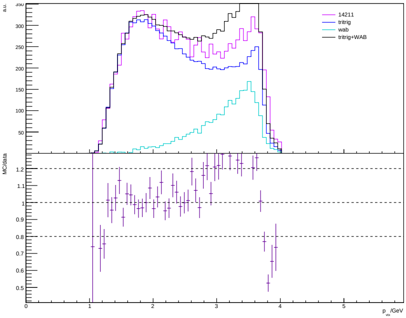
- Applying all preselection cuts, no side-band subtraction

MC – data comparison: with side-band subtraction

electron in top volume



electron in bottom volume



- Applying all preselection cuts, with side-band subtraction

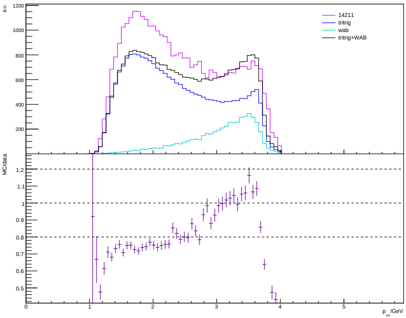
MC – data comparison IV

- There seem to be more background low p_{sum} events in the top than in the bottom.
  - This could point to alignment effects that are present in the top but not in the bottom of the SVT.
- For data, the $t_{\mathrm{trk},e^-} - t_{\mathrm{clu},e^+}$ distribution in the top has a small shoulder on the right side of the time peak.
- Generally, we need to make an effort to equalize the track time distributions between top and bottom, e.g. by introducing separate track and/or cluster time shifts for each half.

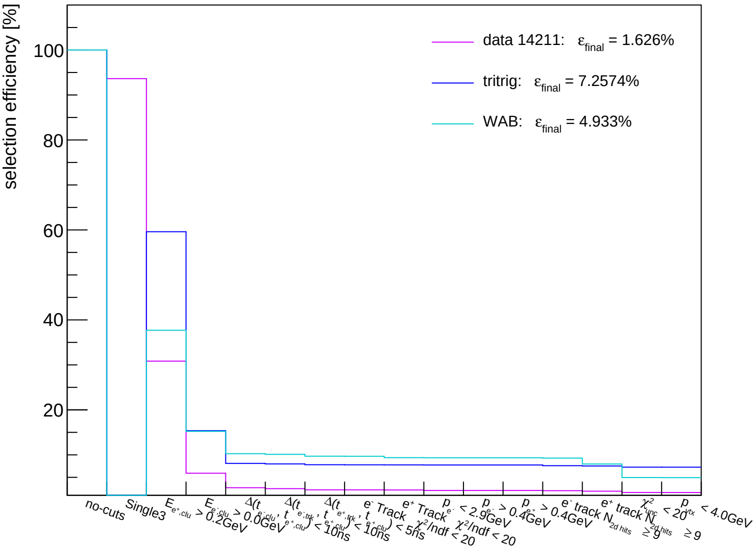
Requiring an electron ECAL cluster

- Lastly, I wanted to check if mimicking a pairs trigger would improve the data – MC ratio.
- Adding e^- energy requirement and $\Delta(t_{\mathrm{clu},e^-}, t_{\mathrm{clu},e^+}) < 10\,\mathrm{ns}$ cut
- Study effect on p_{sum} and $\Delta(t_{\mathrm{trk},e^-}, t_{\mathrm{clu},e^+})$ distribution

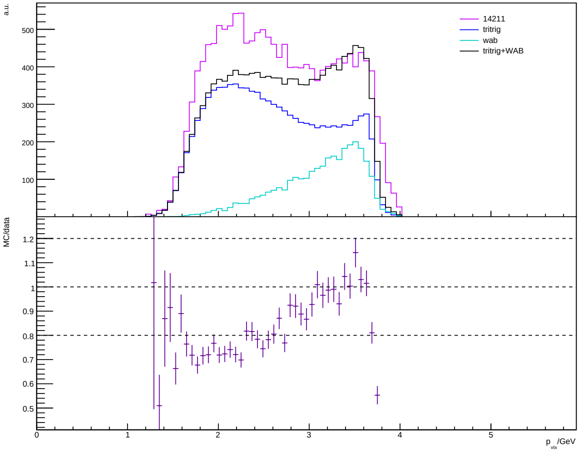
No side-band subtraction, luminosity scaled



Preslection cutflows – including e^- cluster cuts

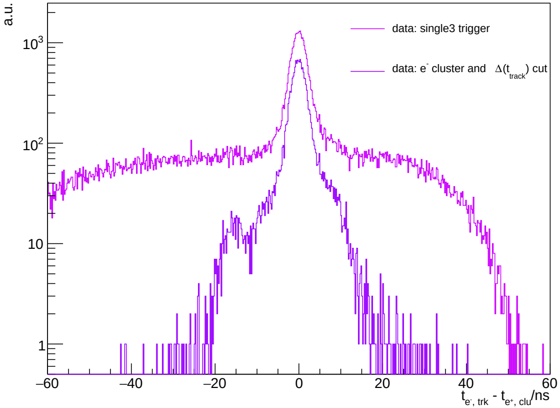


MC – data comparison: p_{sum} with e^- cuts



- Electron cluster cuts don't remove the difference at low p_{sum}
- N_{selected} is reduced drastically compared to standard preselection.

MC – data comparison: p_{sum} with e^- cuts



- Requiring an e^- cluster and applying a $\Delta(t_{\mathrm{clu},e^-}, t_{\mathrm{clu},e^+})$ removes many out of time background events.
- However, we now see a ‘strange’ peak structure with secondary peaks/shoulders.

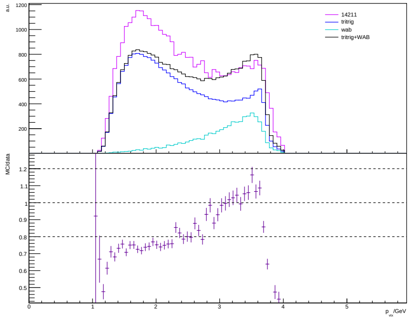
MC – data comparison: summary

- The number of background events at low p_{sum} is higher in data than in MC.
  - This is probably the reason why we are getting fewer than expected MC events using luminosity scaling.
  - Side-band subtraction using the $(t_{\text{trk},e^-} - t_{\text{clu},e^+})$ is an effective tool to remove random background events under the in-time peak.
  - We should either
    - Improve our data reconstruction and selection,
    - Create more accurate conditions in MC to reflect the real experimental conditions, or both
- There are some notable differences between the top and bottom of the detector
  - Timing distributions peak at slightly different values.
  - Requiring $\tan\lambda_{e^-} > 0$ gives more events at low p_{sum} than $\tan\lambda_{e^-} < 0$
- Switching to a pairs-like trigger is not an easy fix to the differences that we see between data and MC

Radiative acceptance and fraction

- For now: trust luminosity scaled MC

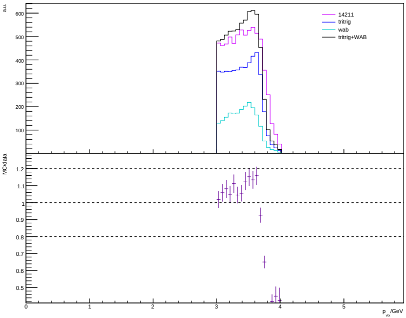
No side-band subtraction, luminosity scaled



Radiative acceptance and fraction

- For now: trust luminosity scaled MC
- Apply additional requirements:
  - Tracks have hits in L2
  - Selecting radiative peak ($p_{\text{sum}} > 3\,\text{GeV}$)

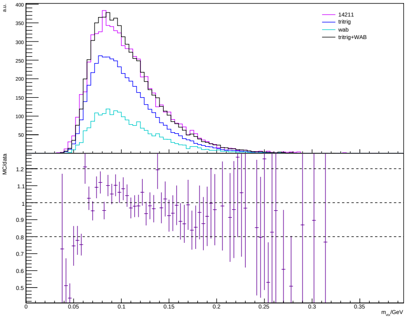
No side-band subtraction, luminosity scaled, radiative selection



Radiative acceptance and fraction

- For now: trust luminosity scaled MC
- Apply additional requirements:
  - Tracks have hits in L2
  - Selecting radiative peak ($p_{\text{sum}} > 3\,\text{GeV}$)

No side-band subtraction, luminosity scaled, radiative selection



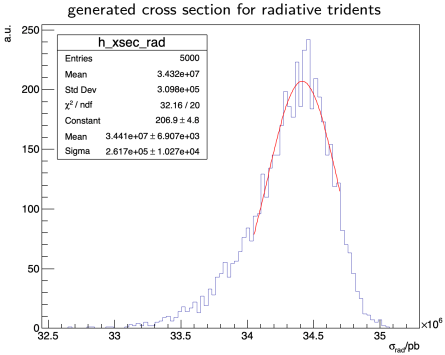
Radiative acceptance and fraction

- For now: trust luminosity scaled MC
- Apply additional requirements:
  - Tracks have hits in L2
  - Selecting radiative peak ($p_{\mathrm{sum}} > 3\,\mathrm{GeV}$)
- To calculate the radiative fraction, we need the correct scaling.
  - Follow same process as for other MC samples

$$f_{\mathrm{scale}} = \frac{\sigma_{\mathrm{gen}}^{\mathrm{rad}} \times \mathcal{L}_{\mathrm{subset}}}{N_{\mathrm{gen}}^{\mathrm{rad}}}$$

Radiative acceptance and fraction

- For now: trust luminosity scaled MC
- Apply additional requirements:
  - Tracks have hits in L2
  - Selecting radiative peak ($p_{\text{sum}} > 3\,\text{GeV}$)
- To calculate the radiative fraction, we need the correct scaling.
  - Follow same process as for other MC samples



Radiative acceptance and fraction

- For now: trust luminosity scaled MC
- Apply additional requirements:
  - Tracks have hits in L2
  - Selecting radiative peak ($p_{\mathrm{sum}} > 3\,\mathrm{GeV}$)
- To calculate the radiative fraction, we need the correct scaling.
  - Follow same process as for other MC samples

sample	$\sigma_{\mathrm{gen}}/\mathrm{pb}$	N_{gen}	f_{scale}
tritrig	4.025×10^9	$17\,792 \times 10^4$	0.0337
WAB	8.249×10^{10}	$38\,700 \times 10^4$	0.318
rad	3.441×10^7	9595×10^4	0.0005
14211	–	–	1.00

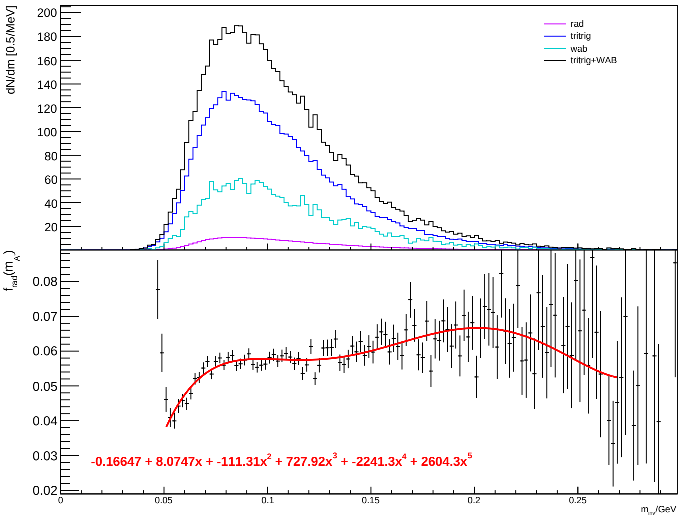
Radiative fraction

SLAC

$$f_{\text{rad}} = \frac{\frac{d\sigma_{\text{rad}}}{dm_{\text{truth}}}}{\frac{d\sigma_{\text{tritrig}}}{dm_{\text{reco}}} + \frac{d\sigma_{\text{WAB}}}{dm_{\text{reco}}}} = \frac{\frac{dN_{\text{rad}}}{dm_{\text{truth}}}}{\frac{dN_{\text{tritrig}}}{dm_{\text{reco}}} + \frac{dN_{\text{WAB}}}{dm_{\text{reco}}}}$$

- N_{rad}, N_{tritrig}, and N_{WAB} refer to luminosity scaled numbers
- Get true invariant mass m_{truth} of e^+e^- vertex from MC truth of radiative trident event.
  - I am pretty confident that I am selecting the correct truth vertex but I need to double-check some things.
- m_{reco} is the reconstructed invariant mass.
- All samples include pulser overlay.

Radiative fraction

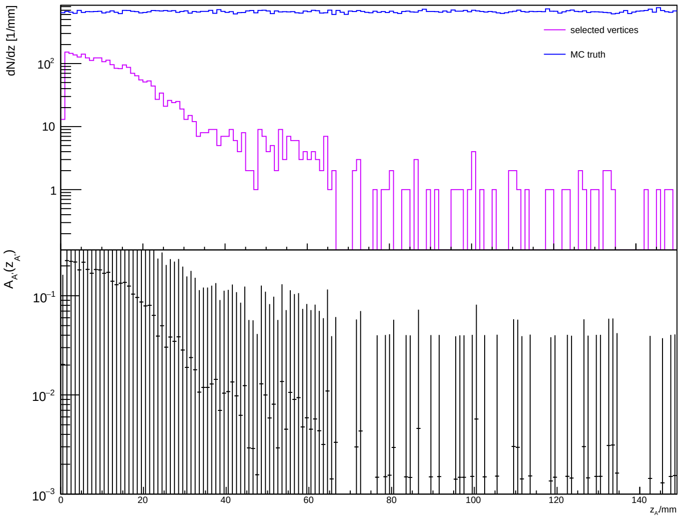


A' acceptance

$$A_{A'} = \frac{dN_{A'}^{\text{selected}}}{dz_{\text{truth}}} \Bigg/ \frac{dN_{A'}^{\text{gen}}}{dz_{\text{gen}}}$$

- Applying preselection and additional requirements like before
  - Tracks have hits in L2
  - Selecting radiative peak ($p_{\text{sum}} > 3\,\text{GeV}$)
- This is by no means a real analysis yet, but I wanted to share my progress with you!

A' acceptance



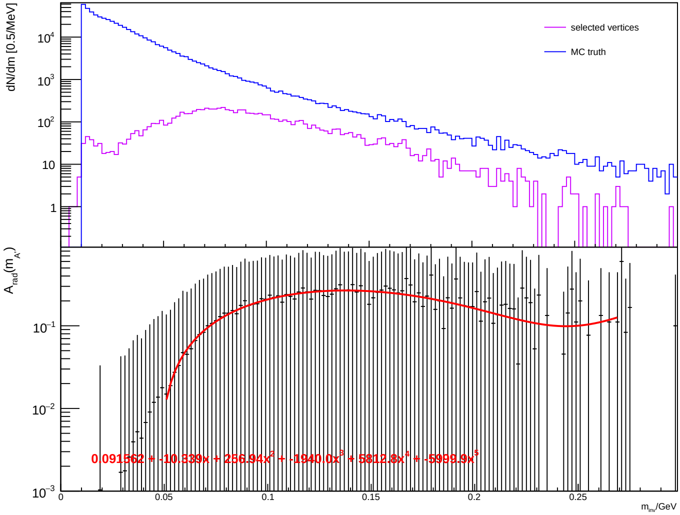
Radiative acceptance

- And because I basically get this for free, I also calculated the radiative acceptance:

$$A_{\text{rad}} = \frac{dN_{\text{rad}}^{\text{selected}}}{dm_{\text{truth}}} \Bigg/ \frac{dN_{\text{rad}}^{\text{gen}}}{dm_{\text{gen}}}$$

- Applying preselection and additional requirements like before
  - Tracks have hits in L2
  - Selecting radiative peak ($p_{\text{sum}} > 3\,\text{GeV}$)

Radiative acceptance



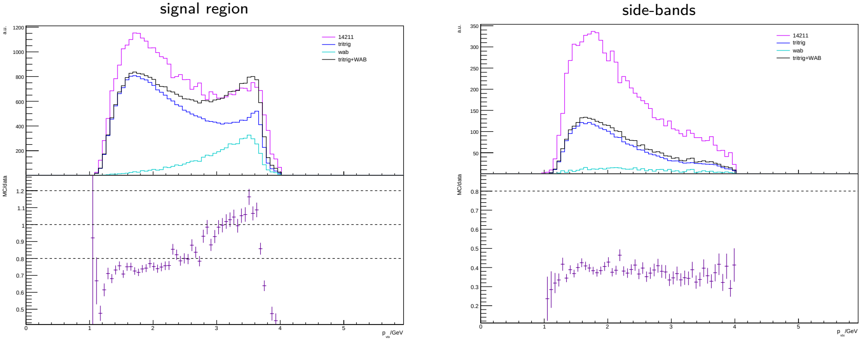
Summary

- With the preselection mostly in place, it is possible to start working on the next steps of the analysis.
- Currently, MC does not include hit killing and smearing effects which will change the distributions I have just shown.
- We will need to monitor the preselection and the comparison of data and MC as we improve the alignment and include inefficiencies in MC.

Backup

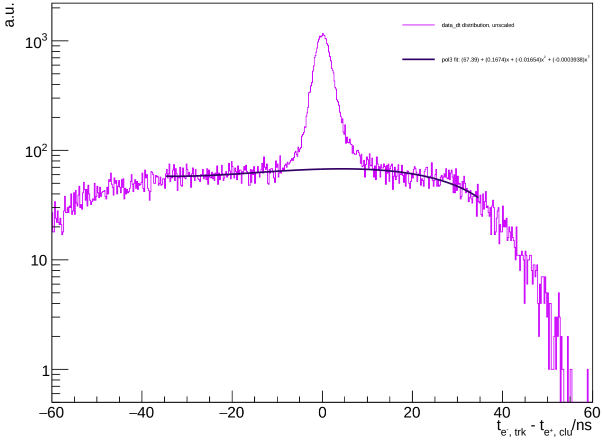
BACKUP BEGIN

MC – data comparison: p_{sum} in SR and SB



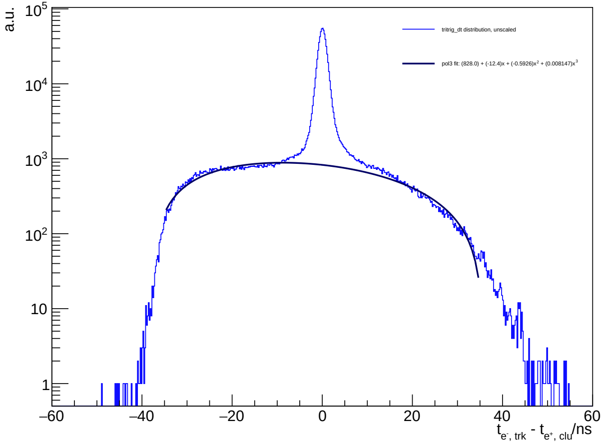
- p_{sum} distributions in signal region and side-bands
  - Signal region is the same as all preselection cuts applied
  - Side-bands see peak at low p_{sum} for data and tritrig, WAB seems to be mostly constant

Fitting $(t_{\mathbf{trk},e^-} - t_{\mathbf{clu},e^+})$ background – data



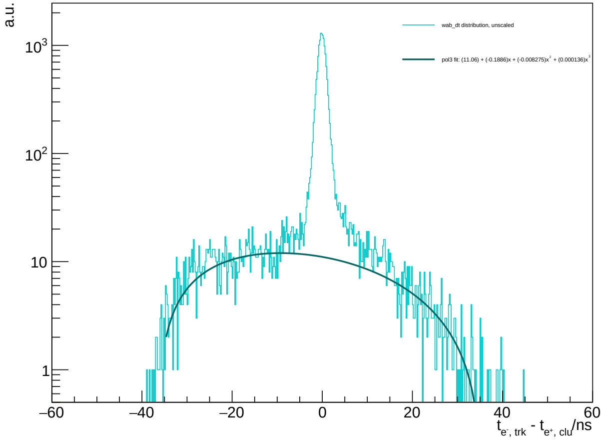
- Fit function $f_{\text{bkg}}^{\text{data}} = a_d + b_d x + c_d x^2 + d_d x^3$

Fitting $(t_{\mathbf{trk},e^-} - t_{\mathbf{clu},e^+})$ background – tritrig



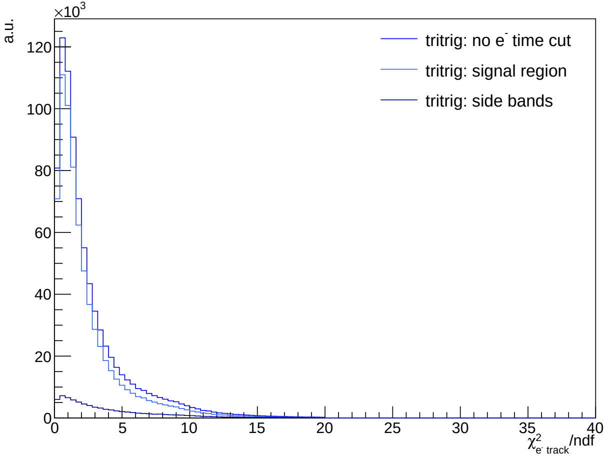
- Fit function $f_{\mathrm{bkg}}^{\mathrm{tritrig}} = a_t + b_t x + c_t x^2 + d_t x^3$
  - Fit not ideal but good enough for now

Fitting $(t_{\mathbf{trk},e^-} - t_{\mathbf{clu},e^+})$ background – WAB



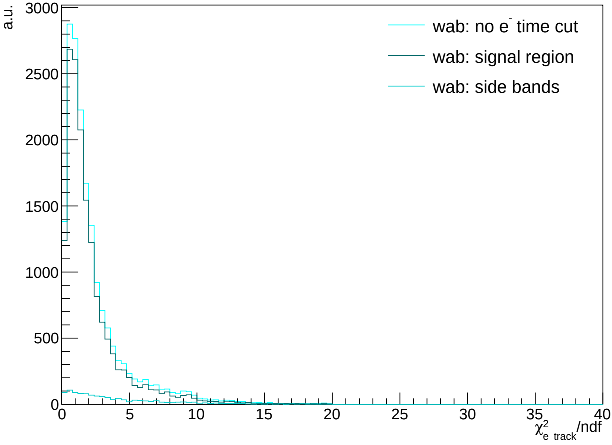
- Fit function $f_{\text{bkg}}^{\text{wab}} = a_w + b_w x + c_w x^2 + d_w x^3$

$\chi^2_{\mathbf{trk},e^-}$/ndf distribution – tritrig

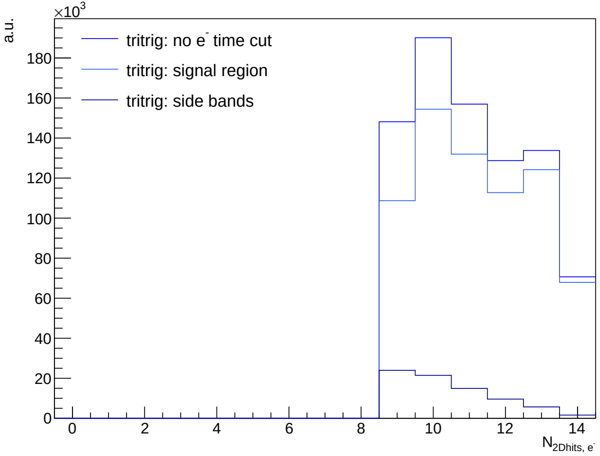


-

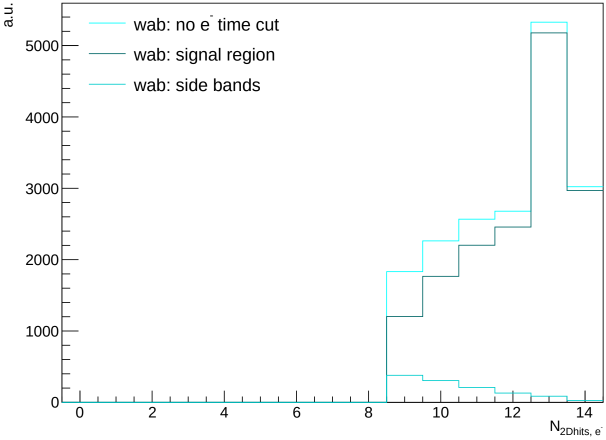
$\chi^2_{\mathbf{trk},e^-}$/ndf distribution – WAB



$N_{2\text{d hits},e^-}$ distribution – tritrig

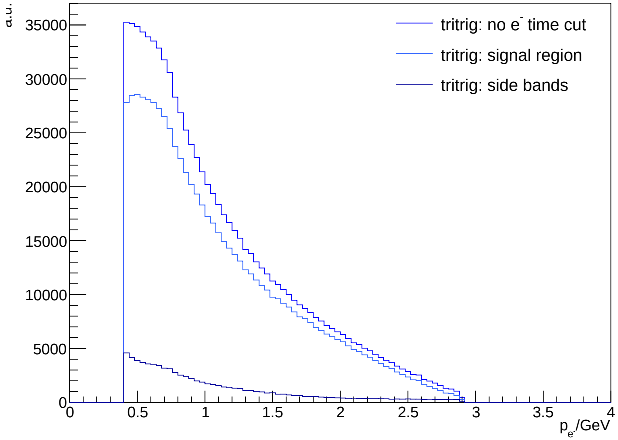


$N_{\text{2d hits},e^-}$ distribution – WAB



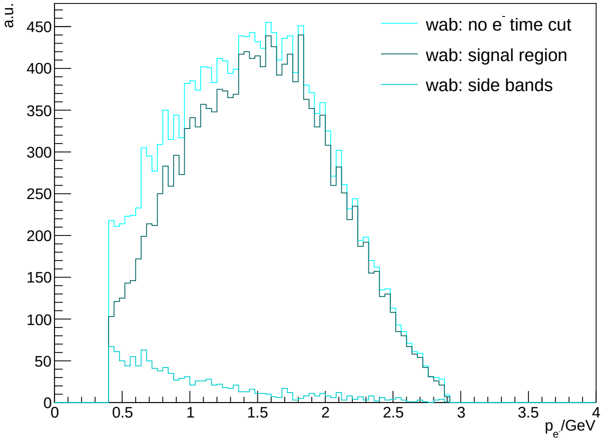
-

p_{e^-} distribution – tritrig



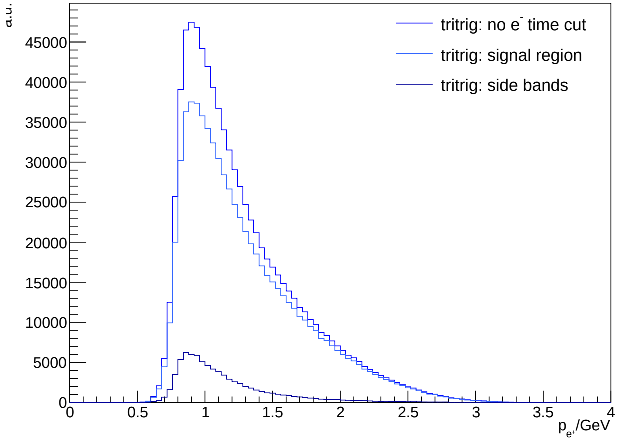
-

p_{e^-} distribution – WAB



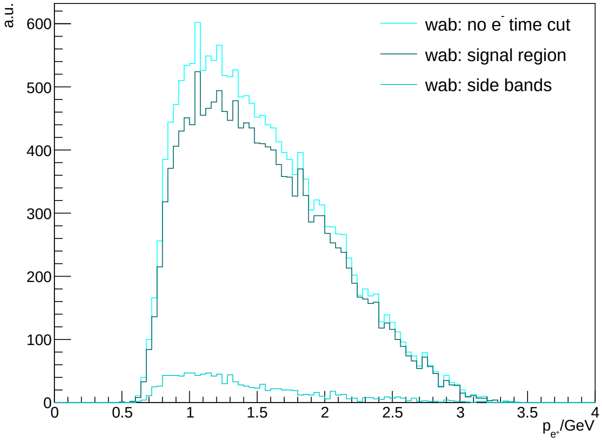
-

p_{e^+} distribution – tritrig

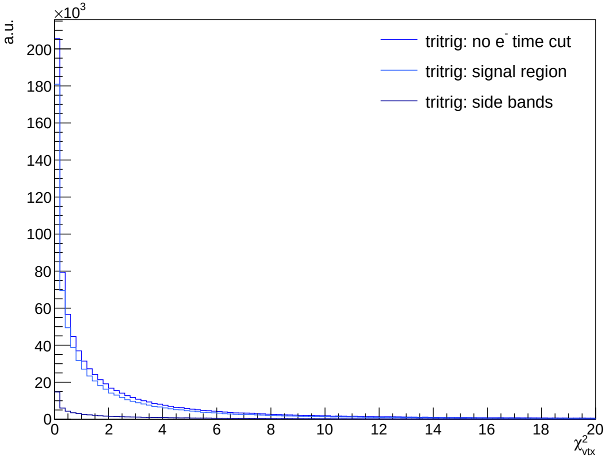


-

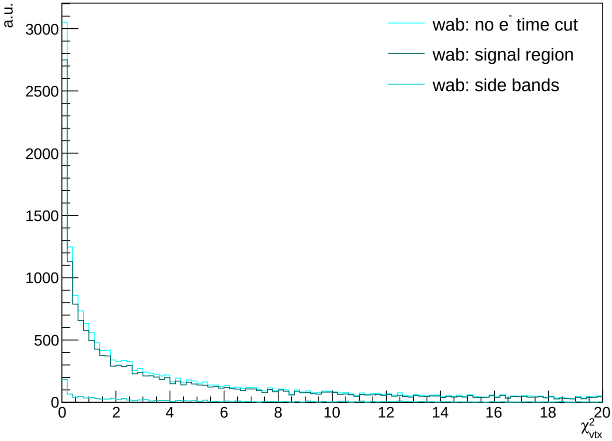
p_{e^+} distribution – WAB



$\chi^2_{\mathbf{vtx}}$ distribution – tritrig



χ^2_{vtx} distribution – WAB



-