



Optimizing the dE/dx Estimator to enhance searches for HSCPs

Mentee: Michael John Noamesi (Gettysburg College)

Mentor: Dr. Emery Nibigira (University of Tennessee)





Overview

Motivation – Why HSCPs and why precise dE/dx matters for PID and in context of long-lived particle (LLP) searches.

Problem – Landau tails & outliers blur mass reconstruction.

New Tools – Prototype estimator + caveats + rescue plan.

Results – Core width, tail suppression, impact on mass.

Future Work



Why Heavy Stable Charged Particles?

- Predicted in SUSY & Hidden Valley
- Sub-relativistic speeds ($\beta = \frac{v}{c} < 1$) $\Rightarrow$ deposit large dE/dx
- Charged + Large mass $\Rightarrow$ Large ionization $\Rightarrow$ deposit large dE/dx
 - proton mass: 0.938 GeV vs HSCP mass: 200 GeV – 2600 GeV
 - **Mass is our handle!**
- Traverse full CMS tracker or decay late $\Rightarrow$ part of **Long-Lived Particles (LLPs)**

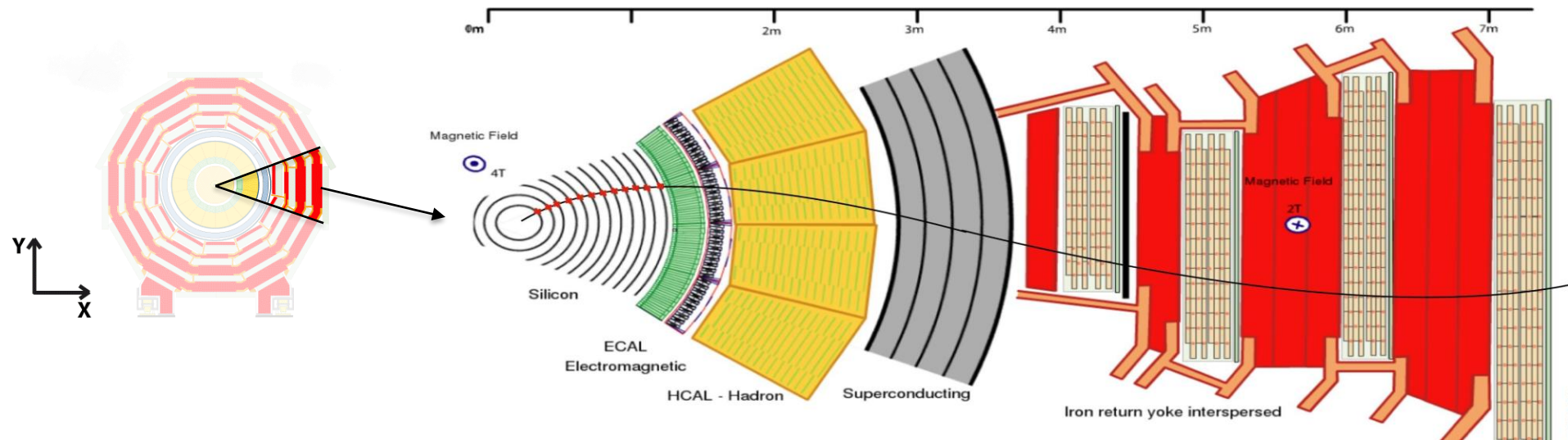


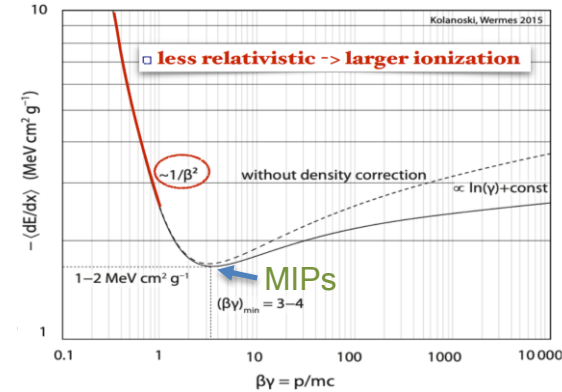
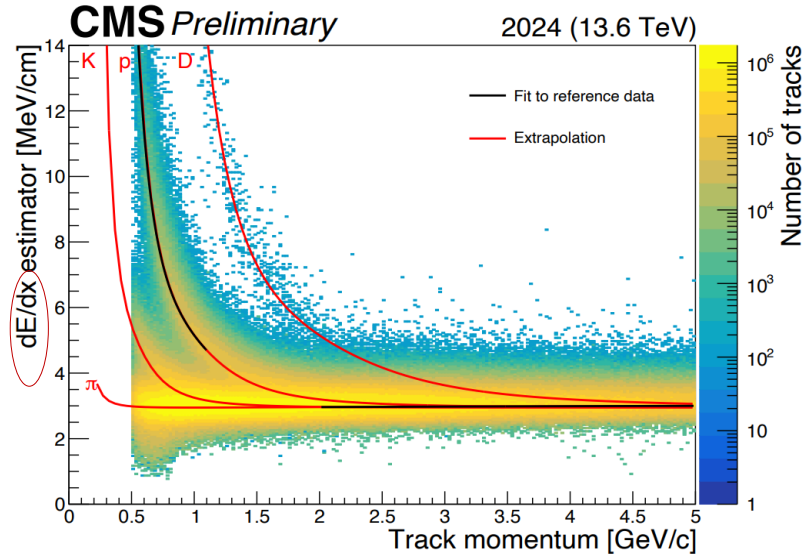
Figure 1. Transverse cut-out of CMS with an HSCP track [—]



How is mass reconstructed?

$dE/dx + \text{Momentum} \Rightarrow \text{Mass}$

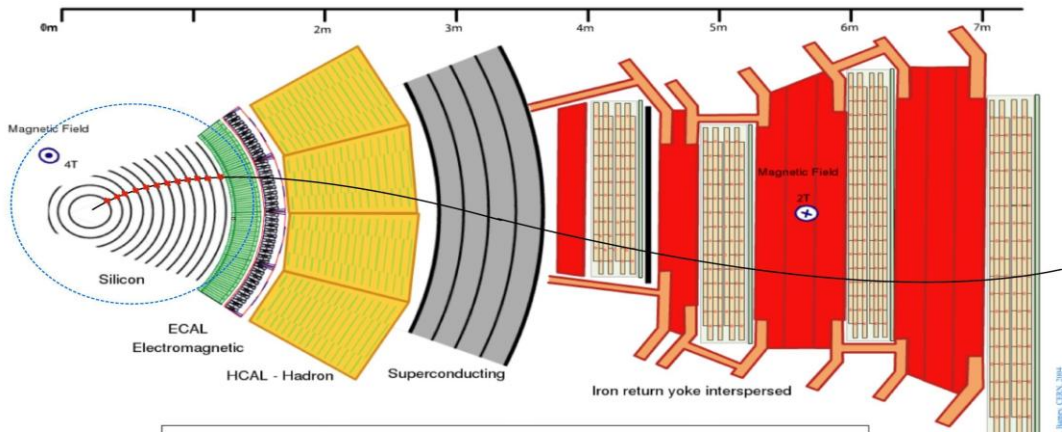
- Bethe-Bloch: $dE/dx \leftrightarrow \beta\gamma$
- $$-\left\langle \frac{dE}{dx} \right\rangle = \frac{4\pi}{m_e c^2} \cdot \frac{n z^2}{\beta^2} \cdot \left(\frac{e^2}{4\pi\epsilon_0} \right)^2 \cdot \left[\ln \left(\frac{2m_e c^2 \beta^2}{I \cdot (1 - \beta^2)} \right) - \beta^2 \right]$$



- Linearizing the first order Bethe-Bloch parameterization
- $\frac{dE}{dx} \approx f\left(\frac{1}{\beta}\right) \approx f\left(\frac{m^2}{p^2}\right) \Rightarrow \frac{dE}{dx} = K \frac{m^2}{p^2} + C$
- Bethe-Bloch fit to a sample of low-momentum particles composed of protons, kaons, and pions reference band in the distribution to extract **K** and **C**.
- The fit is extrapolated for other particles and K and C determined consequently.

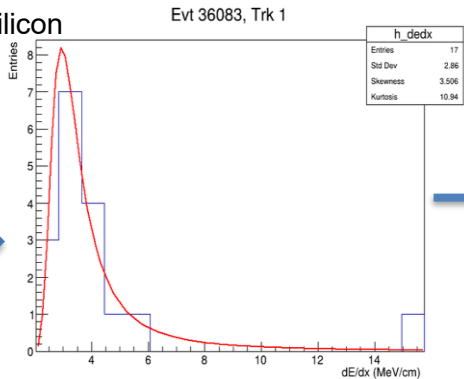
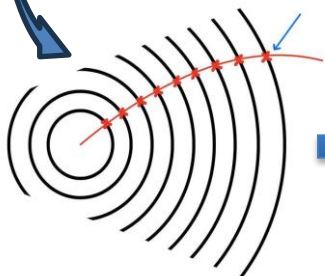
- becomes algebra: $mass_{rec} = p \sqrt{\frac{\frac{dE}{dx} - C}{K}}$

The Landau + outlier Challenge

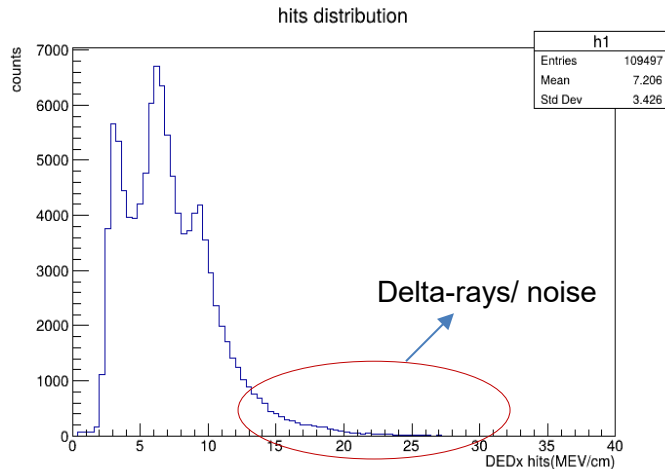


- Hit-level dE/dx is Landau-skewed with a long right tail
- Delta-rays & sensor artefacts inject extreme outliers
- Naïve means are biased $\rightarrow$ mass smear
- Need robust, outlier-tolerant estimators

Per hit dE/dx loss for each silicon detector layer traversed



Extracted $MPV_{dE/dx}$ to reconstruct per track mass



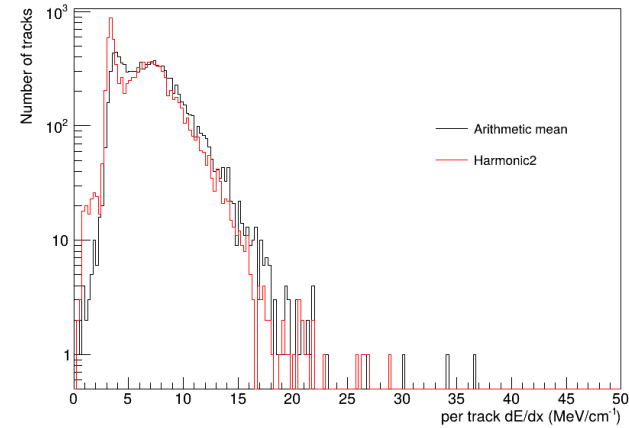


CMS Baseline Estimator – Harmonic-2 (I_h)

$$I_h = \left[\frac{1}{N} \sum_{i=1}^N \left(\frac{dE}{dx} \right)_i^{-2} \right]^{-\frac{1}{2}}$$

- Inverse-square weights tame tails: (**Harmonic-2 estimator (I_h)**, which weights each hit by $\frac{1}{\left(\frac{dE}{dx}\right)^2}$ and thus down weights extreme values)
- σ/μ of MPV distribution $\approx 15\%$ tighter than mean
- Tail persists $> 20 \text{ MeV cm}^{-1}$: possibly values heavily biased by outliers
- Feeds mass fit directly

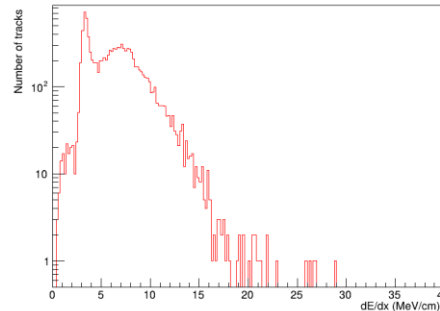
Arithmetic vs Harmonic2



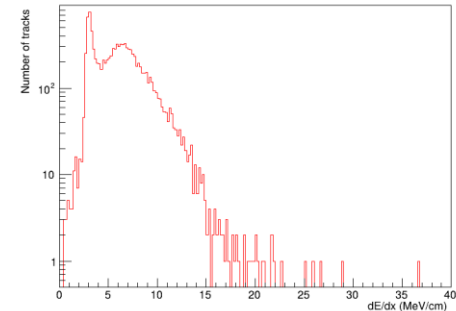
Other baseline estimator options:

- Truncated mean:
 - purges a fixed percentage of the per-track dataset, and calculates the arithmetic mean of surviving hits

Harmonic-2 Mean



40% Truncated Mean



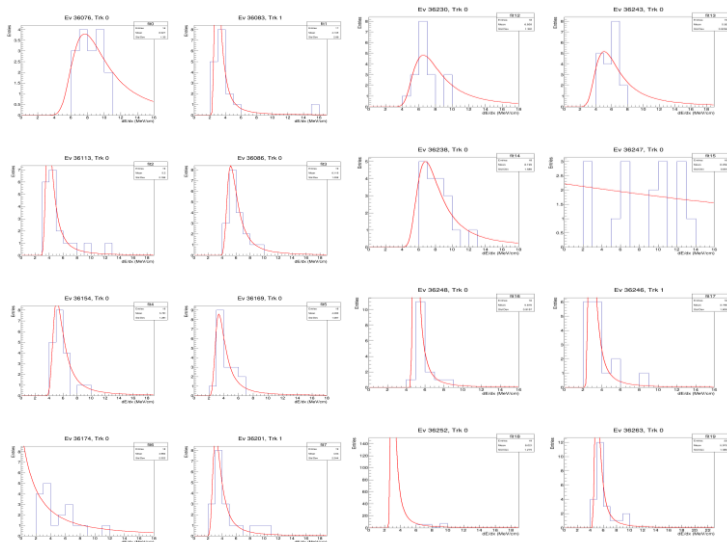


Alternative Prototype (Landau-MPV fit)

Philosophy..... Per track dE/dx is Landau- fit it!

Dataset.

2018 HSCP AOD1800 GeV gluino signal simulation sample (muon trigger and Isolated track filtered to $\approx 11k$ tracks) –no background.



Challenges.

- Tracks vary in hit count (10 – 30).
- Static binning distorts χ^2 wildly.
- 10 M tracks $\rightarrow$ runtime matters.
- Need > 90 % fit convergence.

Adaptive fit recipe

- Minimize the Integrated Mean Squared Error between the histogram and its PDF by setting per track bin width to:

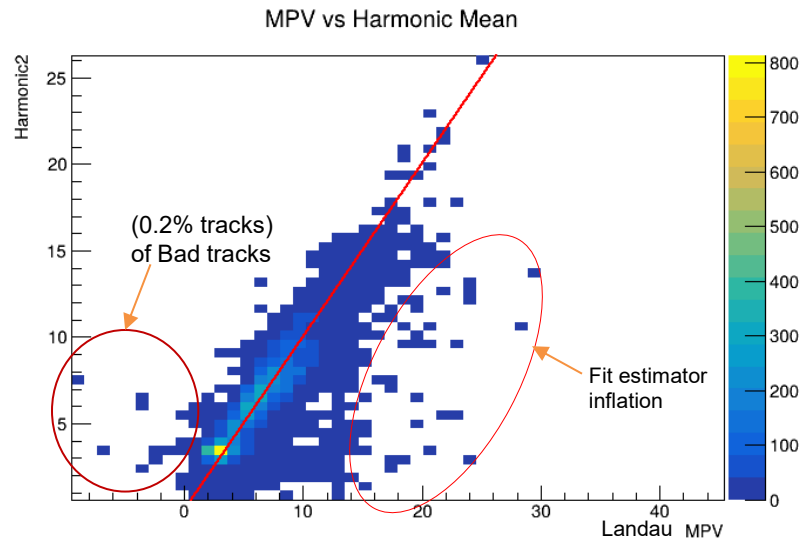
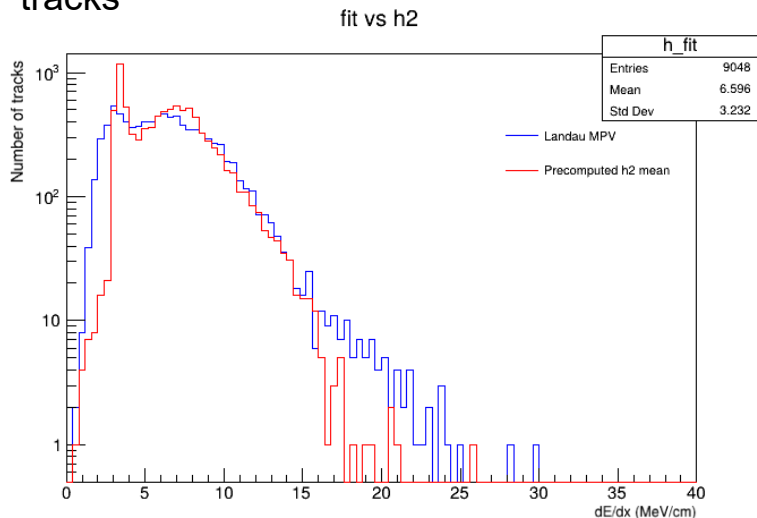
$$\text{bin width} = \frac{2(q_3 - q_1)}{\sqrt[3]{n}} \quad (\text{Freedman-Diaconis: robust on skewed data})$$

- Intuitive seeds: from $\sim 73\% \rightarrow 89.6\%$ convergence @ 0.44ms per track fit(single core).
 - MPV $\approx$ centre of modal bin
 - $\sigma \approx 0.3 \times \text{RMS}$
 - Amp $\approx$ modal-bin height



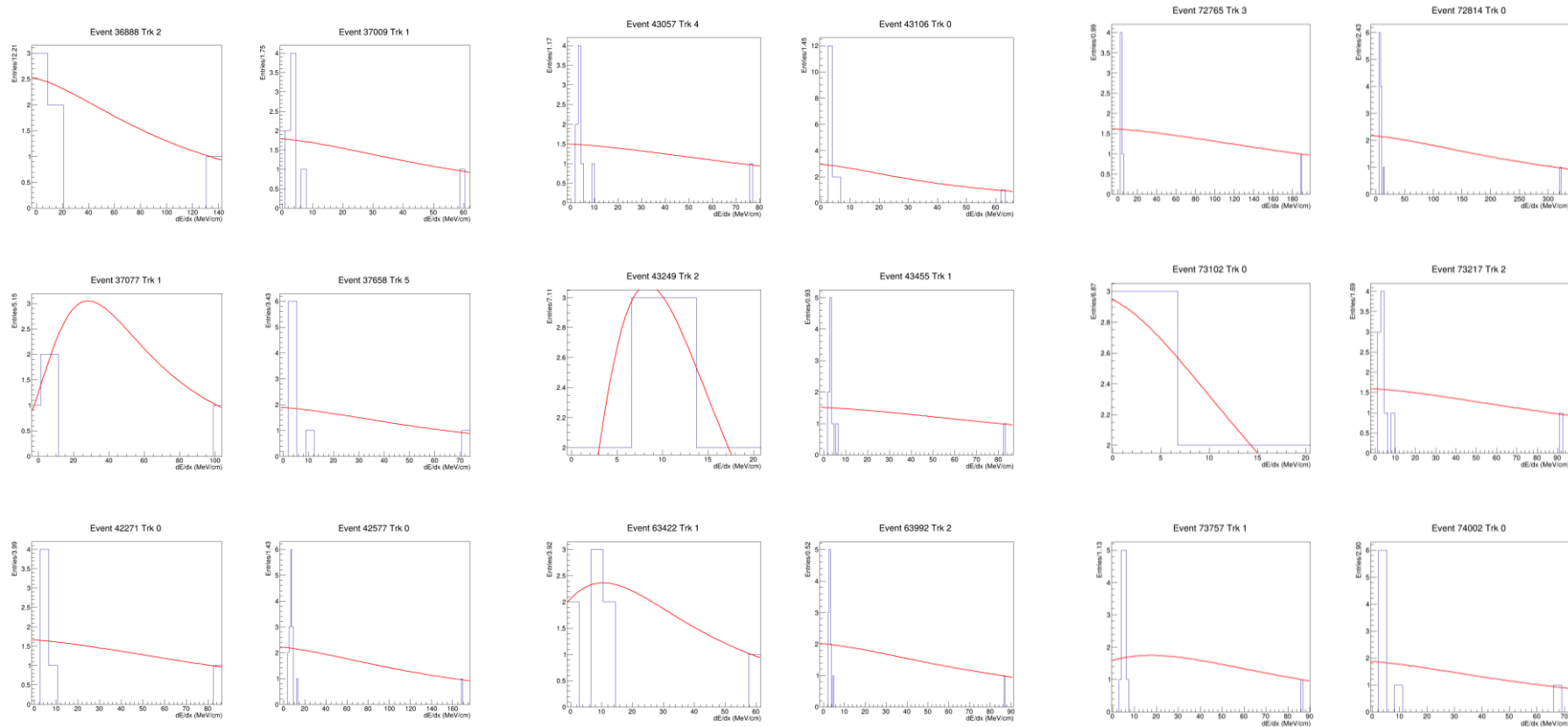
Landau-MPV fit

- **$\approx 5\%$ narrower than the arithmetic mean**
- It tracks Harmonic-2 closely in the MIP-to-15 MeV cm⁻¹ band (3.8–15 MeVcm⁻¹). Above that, the fit estimator starts to inflate above 15 MeVcm⁻¹.
- **Harmonic-2 still seems to suppresses the extreme tail more effectively.**
- Also the fit estimator fails badly(negative mpvs) for 0.2% of the tracks





Fitted Bad Tracks

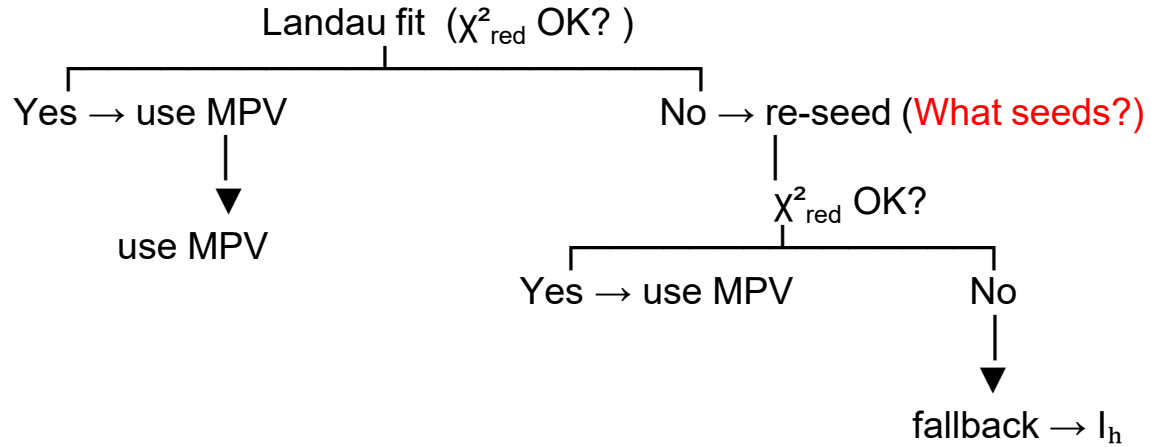




Pathologies and Prototype Optimization

What still goes wrong?

- **0.2 %** tracks $\rightarrow$ negative MPV
– single hit $\gg 25 \text{ MeV cm}^{-1}$
- **5 %** fits: $\sigma \rightarrow 0$ (≤ 4 hits)
- Both smear mass $\Rightarrow$ must flag



Interesting take on the 0.2% tracks.

Minuit fails badly even with tight parameter limits.

All those tracks have huge pathologies

Understanding the correlation between Minuit's MIGRAD's algorithm, this specific IMSE reducing binning and the bad hit inclusive tracks could probe a statistical method to discriminate bad hits.....



Performance Snapshot

```
mpv_h2 shape      : (11074,)
mpv_trunc40 shape: (11074,)
mpv_landau shape  : (9902,)
```

Core dE/dx band ($3.8\text{--}15\text{ MeV cm}^{-1}$):
MIPs in silicon will deposit energy at a rate of approximately **$1.664\text{ MeV g-1cm}^2 \rightarrow 3.876\text{ MeV cm}^{-1}$** in silicon.
Interested in $\gg$ MIP range

Tail cut: $> 20\text{ MeV cm}^{-1}$: Core σ/μ shows *precision*; the tail fraction ($> 20\text{ MeV cm}^{-1}$) shows the *robustness* to extreme outliers.

```
Harmonic-2  core  $\sigma/\mu$ = 0.312  full  $\sigma/\mu$ = 0.485  tail%= 0.15
Trunc 40 %   core  $\sigma/\mu$ = 0.296  full  $\sigma/\mu$ = 0.492  tail%= 0.14
Landau-MPV   core  $\sigma/\mu$ = 0.336  full  $\sigma/\mu$ = 0.492  tail%= 0.35
```

Estimator	core σ/μ	full σ/μ	tail % ($> 20\text{ MeV cm}^{-1}$)	sanity check
Harmonic-2	0.312	0.485	0.15 %	baseline
Trunc 40 %	0.296	0.492	0.14 %	tail-killer, slightly better core
Landau-MPV	0.336	0.492	0.35 %	core <i>wider</i> , tail fatter

Sharper MPV $\Rightarrow$ sharper mass, but these statistics alone can only say so much, full mass-resolution check is required.



Mass reconstruction

Event Selection

Events after full selection: 558

CUMULATIVE SELECTION EFFICIENCY

Selection	Events	Eff. (cum to All)

All events	7713	100.000 %
Trigger	914	11.850 %
$p_T > 55$ GeV	911	11.811 %
$ \eta < 1$	594	7.701 %
# valid pixel hits L2-L4 ≥ 2	574	7.442 %
Fraction of valid hits > 0.8	562	7.286 %
High-purity track	561	7.273 %
Track $\chi^2/\text{ndf} < 5$	561	7.273 %
$ dz < 0.1$ cm	559	7.248 %
$ dxy < 0.02$ cm	558	7.235 %

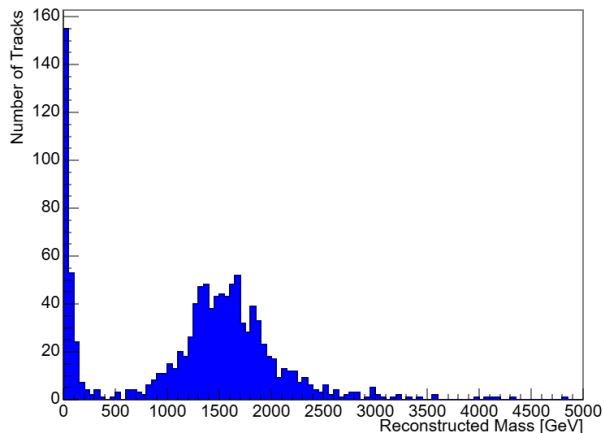
- **Small-N:** only 7 % of the sample NanoAOD makes it through; statistics are thin (558 events $\sim$ 1100 tracks)
- **Bethe-Bloch constants:** $K = 2.49$, $C = 3.18$ taken from the 2018 signal simulation data sample with the Harmonic-2 (I_h) estimator.
 - *Keeping K , C fixed lets us compare prototype estimators shape-to-shape without re-calibrating each one; scale tweaks come later.*

Parameter (MeV/cm)	Data		Simulation	
	2017	2018	2017	2018
K	2.54 ± 0.05	2.55 ± 0.05	2.50 ± 0.05	2.49 ± 0.05
C	3.14 ± 0.02	3.14 ± 0.02	3.18 ± 0.02	3.18 ± 0.02



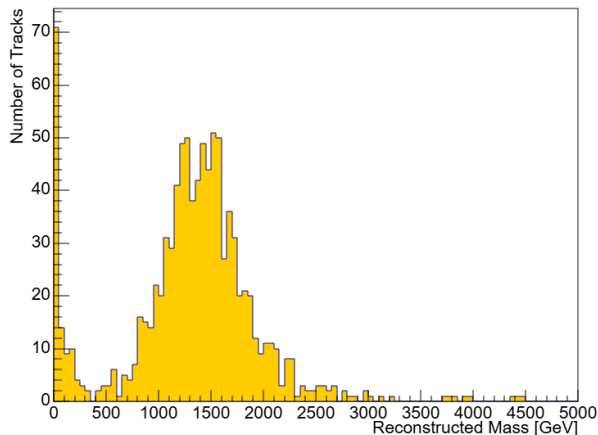
Mass reconstruction snapshots

Mass Distribution (Harmonic Mean Estimator - DeDx_IhStrip)



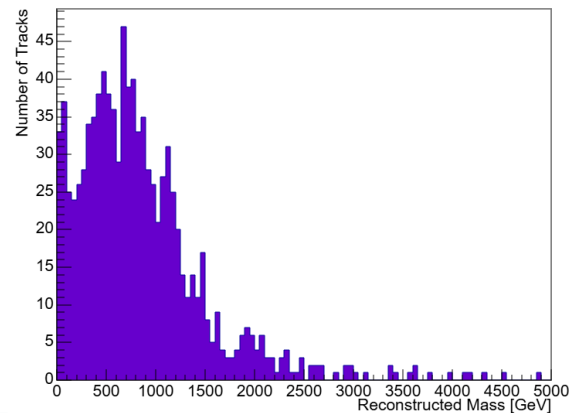
```
=== Mass Reconstruction for Harmonic2 Mean Estimator (DeDx_IhStrip) ===  
Data shape - p: (1187,), Ih: (1187,)   
pt range: 9.895 to 458320.549 GeV   
Ih range: 0.764 to 21.065 MeV/cm   
Tracks with Ih > C: 1113 / 1187   
Mass range: 0.524 - 479212.671 GeV
```

Mass Distribution (Default Truncated Mean Estimator - DeDx_ItStrip)



```
=== Mass Reconstruction for 40% Truncated Mean Estimator (DeDx_ItStrip) ===  
Data shape - p: (1187,), Ih: (1187,)   
pt range: 9.895 - 458320.549 GeV   
Ih range: 0.764 - 20.111   
Tracks with Ih > C: 973 / 1187   
Mass range: 3.305 - 437435.885 GeV
```

Mass Distribution (Fit Estimator - Landau MPV)



```
=== Mass Reconstruction with Fit Estimator ===  
Working with 1167 MPV-p pairs   
  
Mass reconstruction results:  
Tracks with MPV > C: 984 / 1167   
Mass range: 3.667 - 137744.627 GeV   
MPV range: -6.872 - 19.657   
Fit Estimator (Landau MPV): 984 tracks with valid mass reconstruction
```

Single K,C calibration.

-Accurate scale for **Harmonic-2**.

Fit prototype glimpse: small secondary peak at ≈ 1.9 TeV hints that, once we gate on χ^2 , reseed bad fits, and mask single-hit outliers, a dedicated K,C fit should land **right on 1.8 TeV** with a core possibly narrower than Harmonic-2.



Future Work

Algorithm clean-up

- χ^2 -gated reseed for Landau fits
- Isolation-Forest single-hit masking in a pre-pass
- Per-track dynamic f-truncation to shave residual tails

Calibration & validation

- Re-fit K,C separately for
 - 40 % truncator
 - Landau-MPV v2
- Re-run mass reconstruction on full signal data (better statistics).

Even more

- Prototype an adaptive truncation algorithm
- Understand why MIGRAD fails for 0.2% tracks with pathological hits



Conclusion

Take-aways

HSCP searches live on precise dE/dx .

Better MPV $\Rightarrow$ sharper reconstructed mass $\Rightarrow$ deeper reach for long-lived new physics.

Baseline vs prototype:

Harmonic-2 (CMS default) $\sigma/\mu \approx 0.31$

40 % truncator tightens core by ≈ 4 % and halves extreme-tail rate.

Landau-MPV fit shows a little 1.9 TeV bump; optimizing the algorithm better should yield different results.

Single K,C:

I_h determined K and C put all peaks within 10 % of the true 1.8 TeV mass—scale should lock once each estimator gets its own K,C.

What I learned?

- Writing efficient code:
 - Vectorised ROOT/VecOps & NumPy pipelines
 - Lazy loading, one-pass histogramming, functional helpers, caching
- Statistics
- Interesting new physics!!



Acknowledgements

This work was supported by the DOE grant RENEW-HEP: U.S. CMS SPRINT - "A Scholars Program for Research Internship" at Tougaloo College, MS (DE-SC0023681), Brown University (DE-SC0023651), RI, University of Puerto Rico-Mayaguez, PR (DE-SC0023680), and University of Wisconsin – Madison, WI (DE-SC0023643) and U.S. CMS Operations at the Large Hadron Collider (NSF-2121686).

THANK YOU!!