A Machine Learning Approach to Track Identification in Emulsion Cloud Chambers

Benda Xu <benda.xu@ipmu.jp>
Kavli IPMU, The University of Tokyo, Japan

Emulsion Cloud Chamber and τ Physics

- τ is the heaviest and least-understood lepton generation.
- Emulsion cloud chamber (ECC): the key technology to ν_τ physics,
  - DONUT for ν_τ discovery.
  - OPERA for ν_τ oscillation appearance.
- Proposed SHiP (Search for Hidden Particles) is designed with an ECC neutrino detector having 100× more ν_τ events than OPERA.

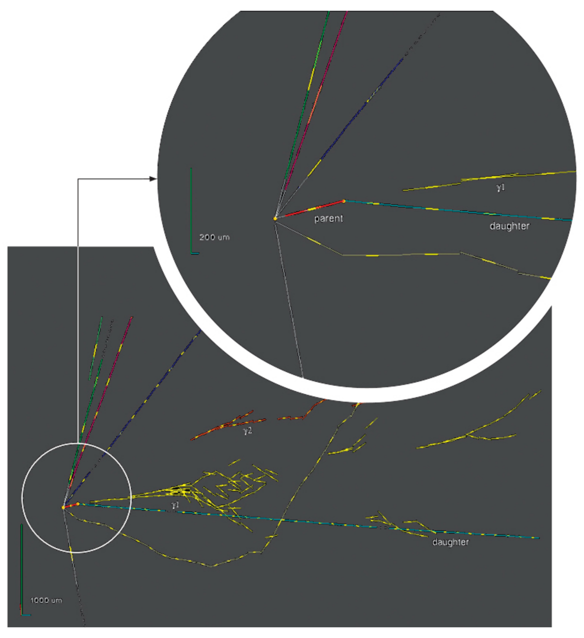


Figure: $\nu_\mu \to \nu_\tau$ appearance observed by OPERA with excellent resolution of event topology and particle identification.

Scope of This Work: An OPERA-Like ECC Brick

- Focusing on emulsion cloud chamber (ECC), ignoring other parts.
  - Signal: electromagnetic (EM) shower.
  - Background: random cosic ray.

Base tracks are already reconstructed in each emulsion layer and parameterized as $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{TX}, \boldsymbol{TY}, \chi^2$.

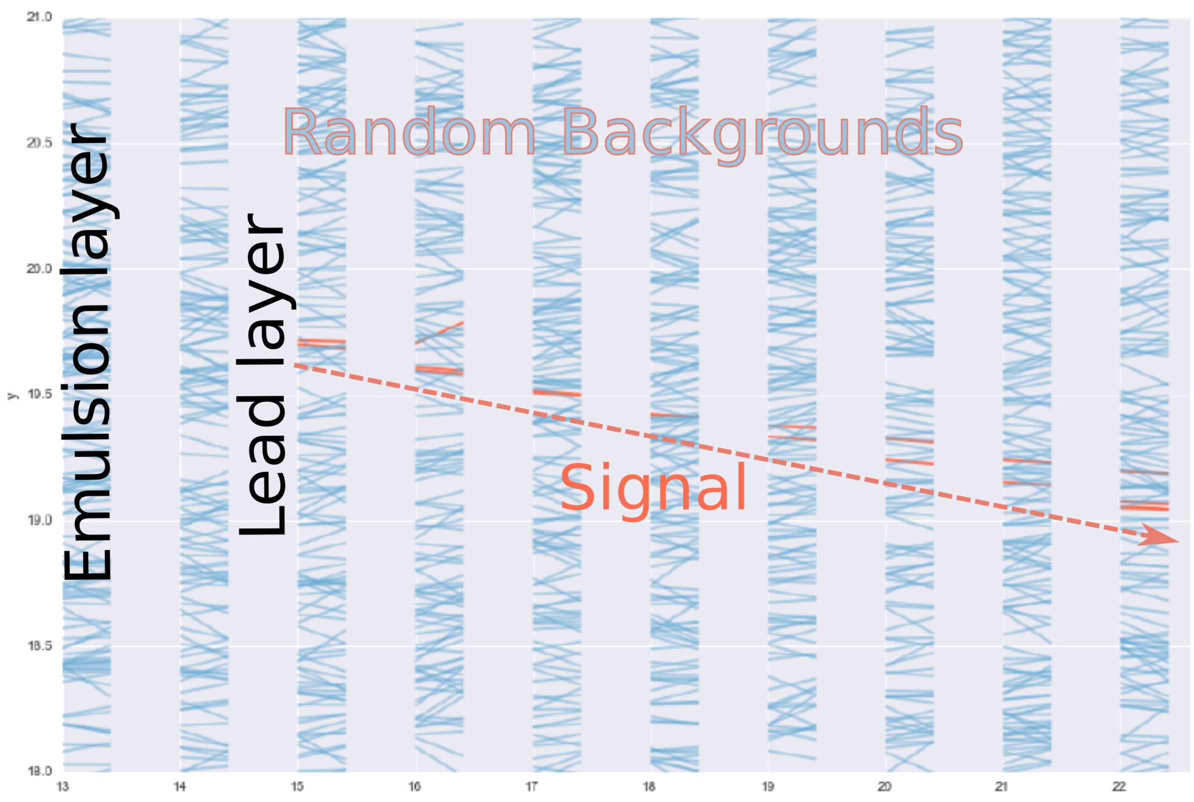
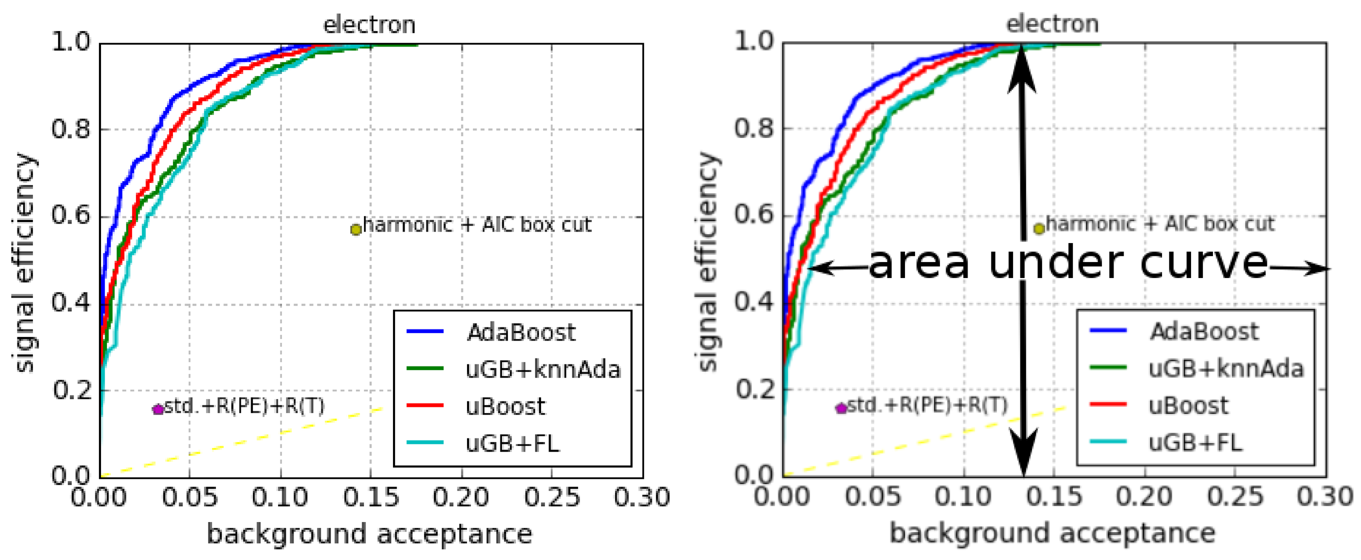


Figure: Adapted from `https://www.kaggle.com/c/dark-matter-signal-search-episode-1`

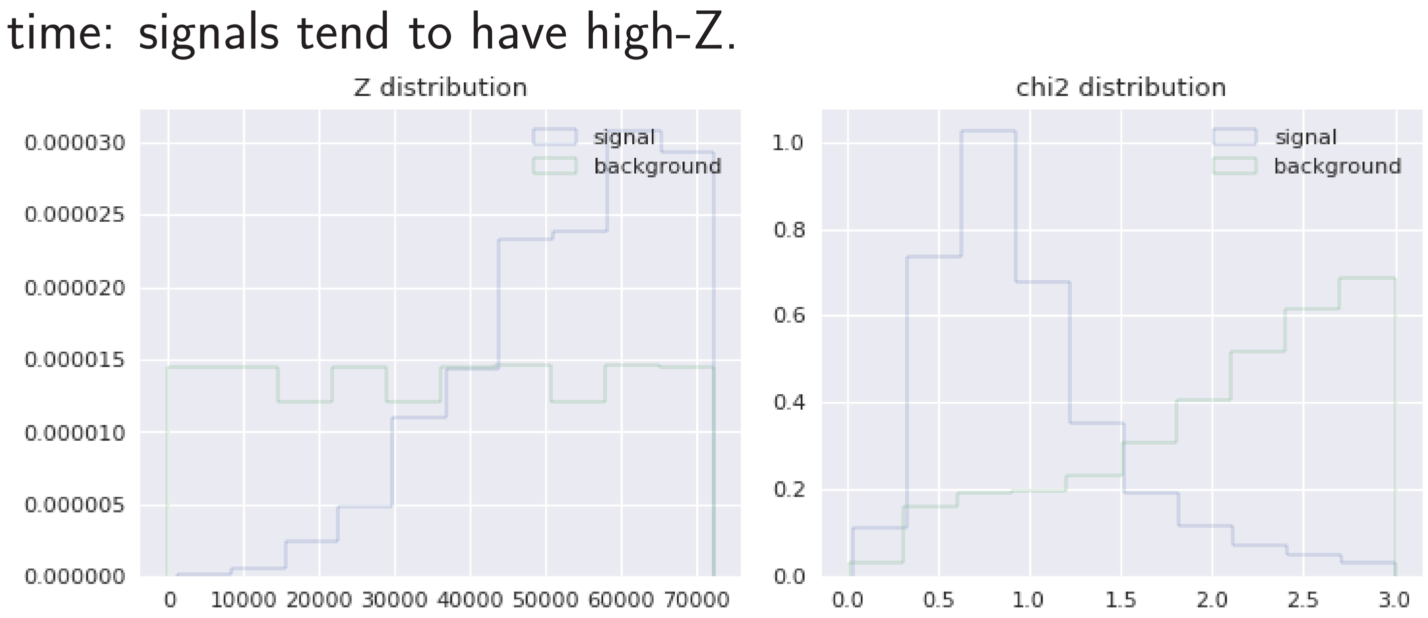
- ROC: receiver operating characteristic
  - In physics: signal efficiency vs. background acceptance.
  - Compare parameters (curves) and cuts (points).
  - AUC (area under ROC curve) is a performance measure for classifier: 0.5 means no classification and 1 means perfect classification.



- Energy resolution, etc., are also important, but we focus on AUC as a figure-of-merit. It is an adaptation to Kaggle and the machine learning community. `https://goo.gl/8N7BYG`

Baseline Performance – AUC 0.930

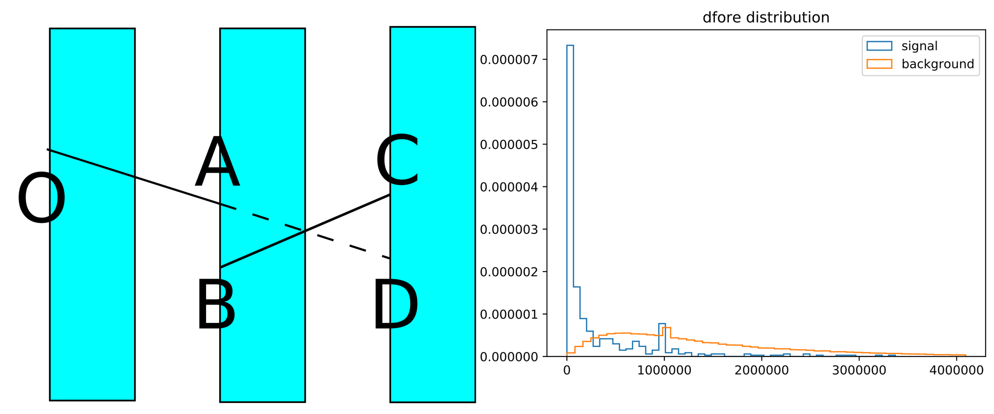
1. An eletromagnetic shower should be contained in the block to be useful: X, Y near the center, TX, TY near 0.
2. An eletromagnetic shower develops more base tracks when evolving in time: signals tend to have high-Z.



3. Signal base trackes have smaller χ^2. AUC: $\sim$ **0.845**
4. The baseline solution combines all of them by boosting. AUC: **0.930**

Parameter: Base Track Connection – AUC 0.995

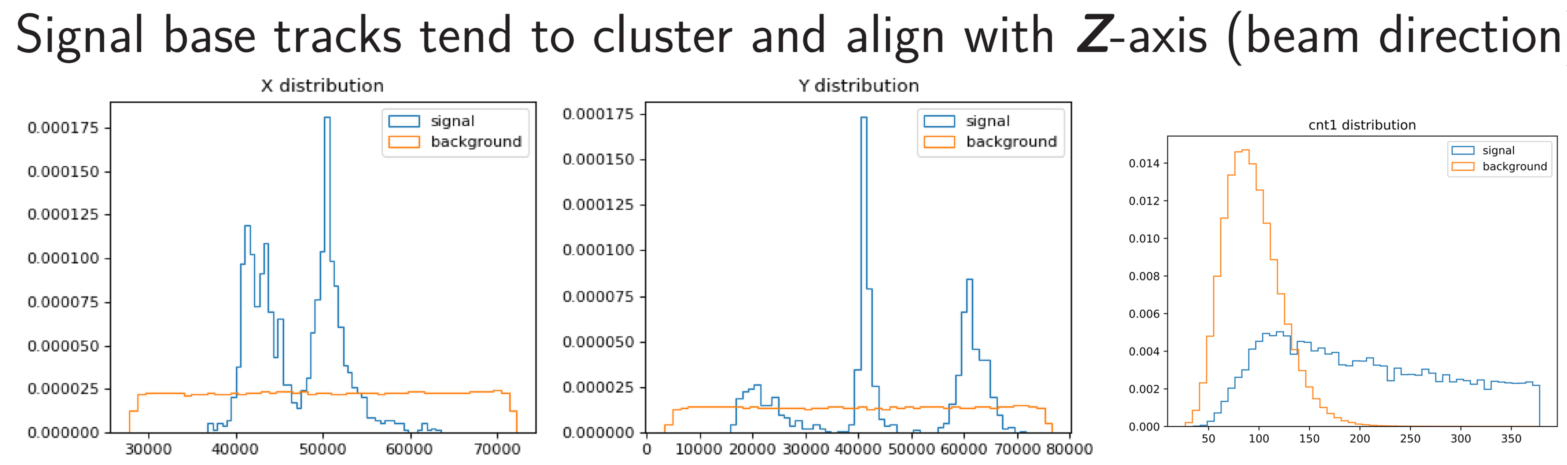
Most of the signal base tracks have neighbours.



- For each base track $\boldsymbol{OA}$.
  1. Select the closest downstream basetrack $\boldsymbol{BC}$ by $\boldsymbol{d}^2 = \|\boldsymbol{AB}\|^2 + \|\boldsymbol{CD}\|^2$. Use the $\boldsymbol{d}^2$ as a feature.
  2. Do the similar with upstream basetracks.
- Signals are concentrated towards $\boldsymbol{d}^2 = \mathbf{0}$ (dfore in the figure).
- Some layer might miss the signal base track:
  - Extension: jump 2 layers. AUC: **0.930** → **0.993**
  - ... and up to 6 layers. AUC: **0.993** → **0.995**

Parameter: Histogram in 3D – AUC 0.997

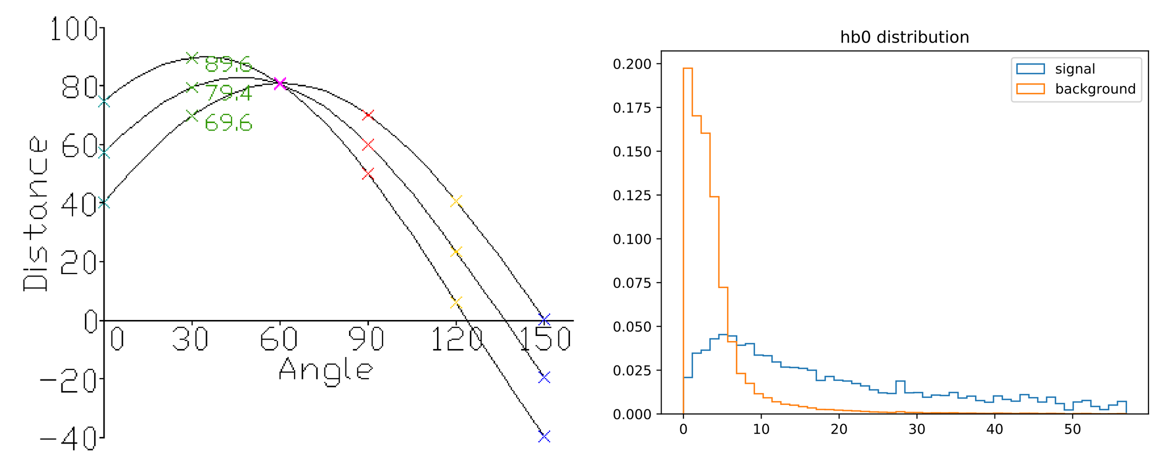
- Signal base tracks tend to cluster and align with $\boldsymbol{Z}$-axis (beam direction):



- Histogram the tracks in $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$, and assign each track the bin count of the histogram.
- Signals have larger bin counts (cnt1 in the figure).
- Convolute with a smoothing kernel $(\mathbf{1}, \mathbf{3}, \mathbf{1})^{\mathbf{3}}$ 0–6 times. AUC: **0.995** → **0.997**

Parameter: Histogram in Hough Space – AUC 0.998

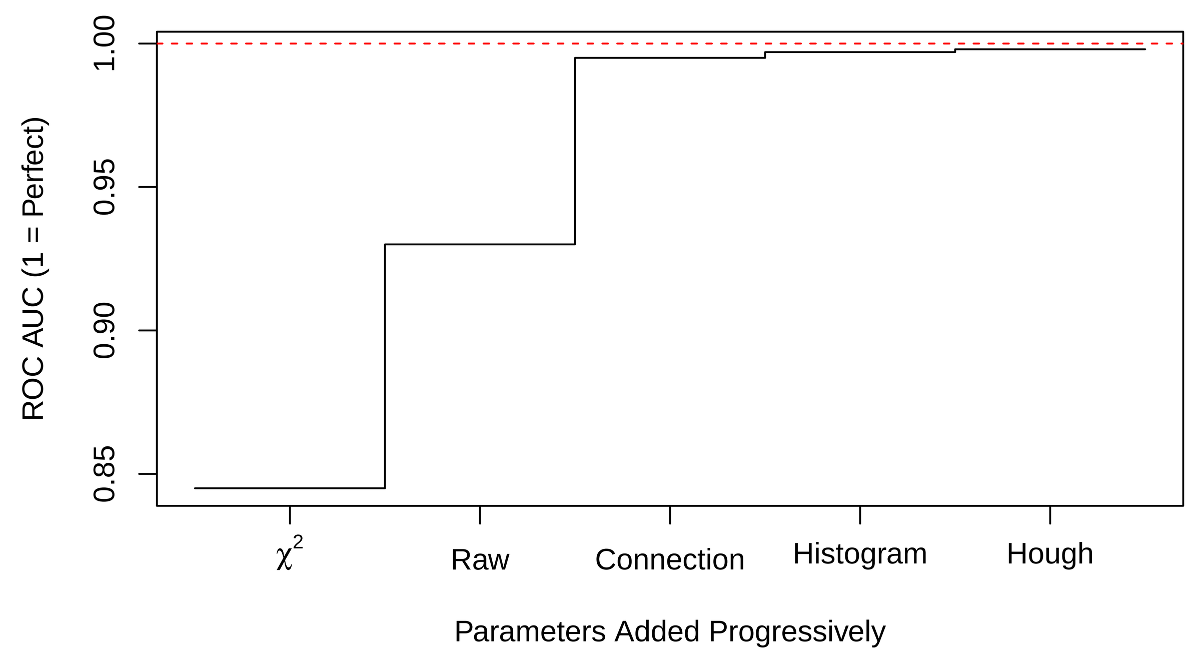
- Hough Transformation catches the line nature of signal tracks.



- A line in 3D space has degree-of-freedom 4.
- 3D line parameterization suitable for ECC is given by K.S. Roberts (1988), *A new representation for a line*.
- Signals have larger bin counts in the Hough-Roberts space (hb0 in the figure). AUC: **0.997** → **0.998**

The Way to Track Identification

- Tune the hyperparameters of extreme gradient boosting.
- Iteration: tracks close to 1-pass identified signals are more likely to be signals in the 2-pass. AUC (2 samples): **0.99835** → **0.99840, 0.99842** → **0.99845**



- The method will be applied to data from OPERA for evaluation of real world performance.
- Acknowledgements: Andrey Ustyuzhanin, Ulrik Egede, Miha Zgubic and all the friends at MLHEP2017; OPERA and SHiP collaborations.