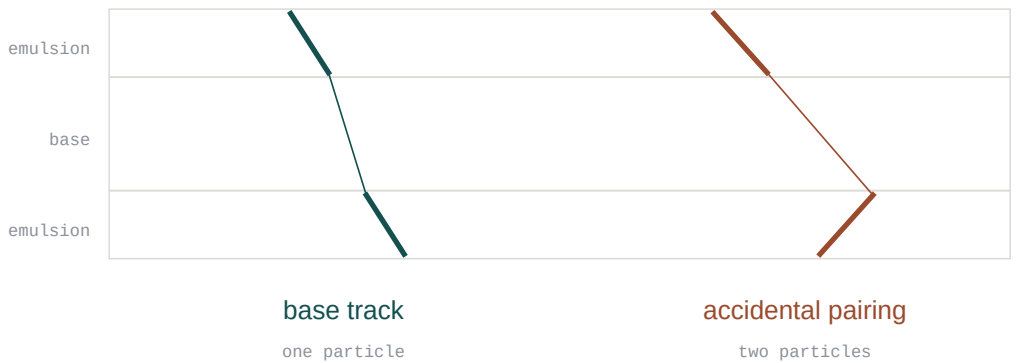
A public dataset and benchmark

The reconstruction problem, the tasks a challenge could ask, and what we would need from the simulation to make it worth doing.

01 . THE DATA

Emulsion records tracks one film at a time

Each film carries emulsion on both sides of a plastic base. A charged particle leaves a short segment in each layer, and the pair across the base gives a position and a direction. That pair, a base track, is the unit everything is built from. Pairing is geometric, so some pairs join segments from two different particles.



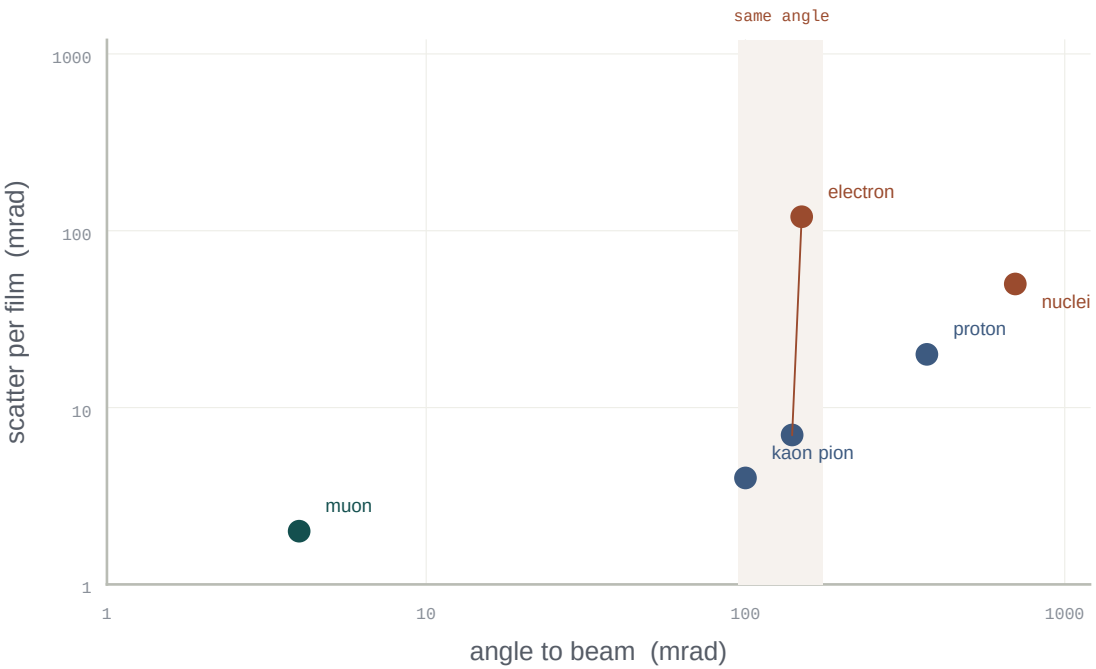
FILMS	BASE TRACKS	REAL PAIRINGS	DISTINCT PARTICLES	ANGULAR RESOLUTION
226	~178 M	~62 %	~10 M	~1 mrad

Two populations, and everything else under a percent

	BASE TRACKS	PARTICLES	ANGLE TO BEAM	FILMS SPANNED	SCATTER / FILM	BEHAVIOUR
muon	59 %	3 %	4 mrad	whole volume	2 mrad	crosses everything, straight to within resolution
electron	41 %	96 %	150 mrad	4	120 mrad	short, numerous, wide, the cascade products
pion	0.2 %	0.2 %	140 mrad	10	7 mrad	track-like, interacts or decays in the volume
kaon	0.02 %	0.01 %	100 mrad	15	4 mrad	longer and stiffer than a pion
proton	0.09 %	0.14 %	370 mrad	3	20 mrad	short and heavily scattered
nuclei	< 0.01 %	< 0.01 %	600 - 800	1 - 2	40 - 60 mrad	fragments, a segment or two at large angle

Most of the data and most of the objects are different populations: muons are most of the base tracks and almost none of the particles, electrons the reverse. Any single aggregate number is largely a readout of that split.

Angle does not separate the species. Scattering does.



A pion sits at the same angle to the beam as an electron. No per-track angular cut separates them.

Their scattering differs by more than an order of magnitude, and the whole hadron family sits well below the electrons.

Identification therefore needs a chain rather than a single base track, which places it after grouping rather than before it.

Two tasks, with the interaction as the goal

SEGMENTATION
a label and an instance for every base track

task

RELATIONS
typed links between the objects

task

INTERACTION
where it happened and what it was

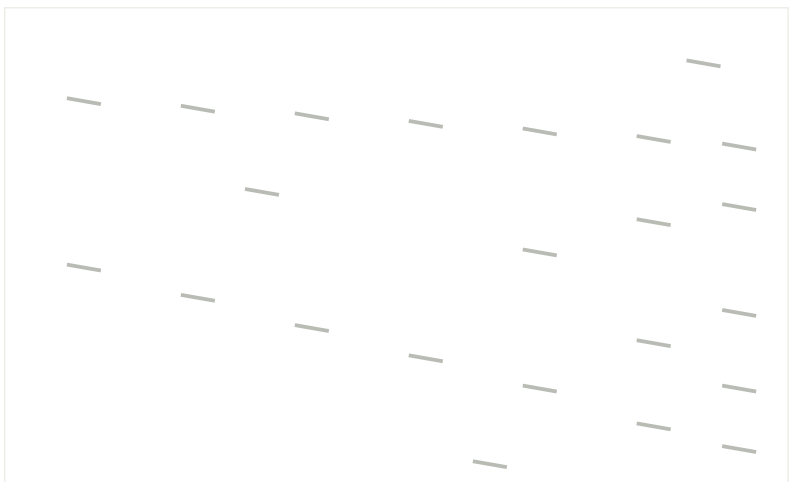
the goal

	GIVEN	PRODUCE
Segmentation	every base track in a volume, accidental pairings included, nothing pre-filtered	for each base track, what it is and which object it belongs to
Relations	the objects	for each object, what produced it, and which objects share an origin

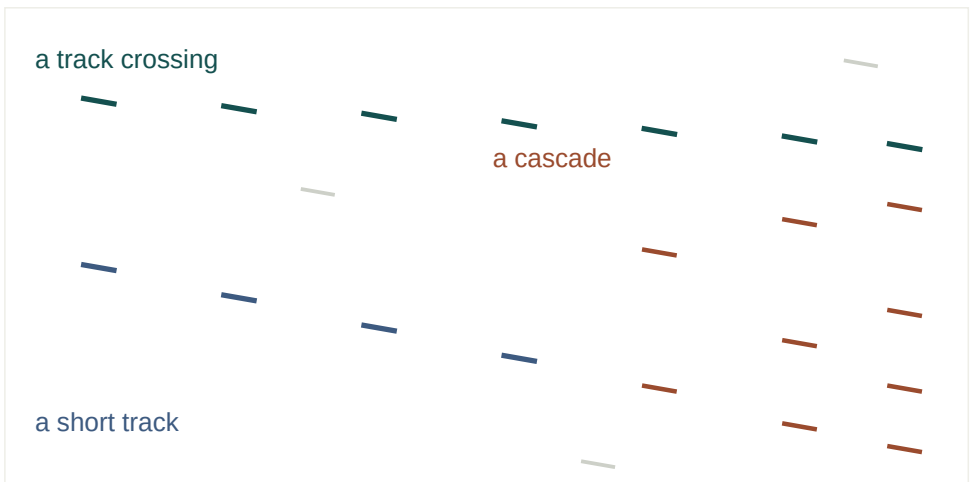
We are not fixing the label set, the granularity of an object, or whether an object may nest inside another. Those are real choices with consequences, and they are among the things we would like decided rather than assumed.

Segmentation: the volume becomes objects

base tracks in a volume



objects, plus what belongs to nothing

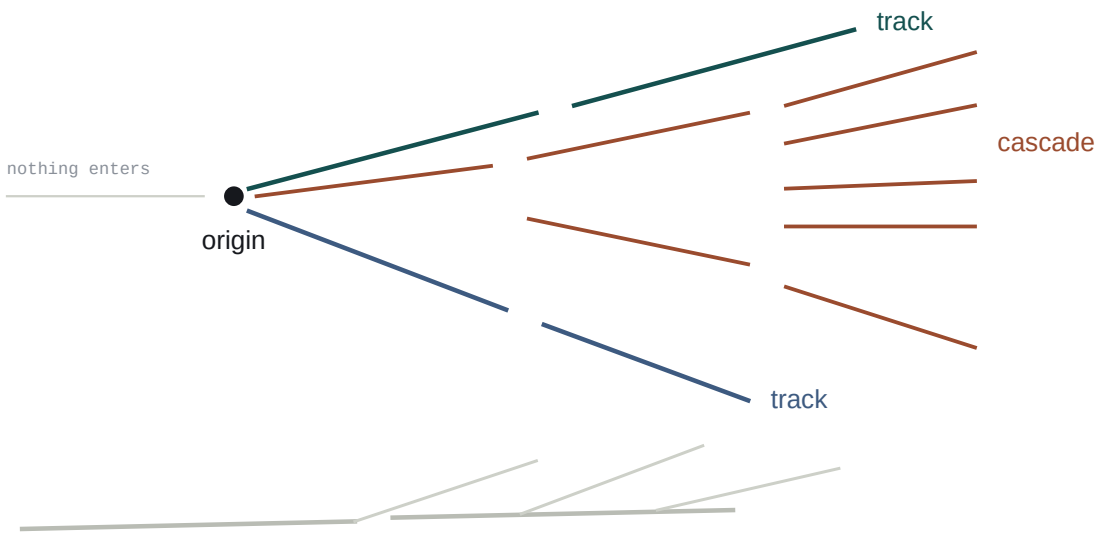


accidental pairings, belonging to no object

The label answers what a base track is, including whether its two segments came from the same particle at all. The instance answers which object it belongs to. Both are open as to how finely they divide.

Relations: two kinds of link, and one of them is hard

An object was produced by another, a cascade by the particle that radiated it. Several objects share an origin. An interaction is a set of objects sharing an origin, none of them produced by anything that entered from outside.



a cascade from a passing track, not an interaction

07 . THE OPEN QUESTION

What do we want a shower to be?

This is the question we would most like answered, because everything downstream depends on it and we do not think it is ours to settle alone.

A shower is not one track. It is a spray of short, wide, heavily scattered segments, and it carries the energy of whatever started it. What we call an object inside that spray is a choice rather than something the data dictates.

One object per shower, one per conversion, or one per branch: each is defensible, and each implies a different level at which two showers must be told apart.

The choice and the measurement are entangled. How finely we divide sets how crowded the objects are, and therefore what can be separated at all. Choosing finely enough makes the task impossible for reasons of definition rather than of instrument.

It also decides the scoring. Whether the easy majority is counted, and whether we count per base track or per object, moves any score enormously. We would rather agree the principle now and the metric together, once the level is settled.

What we would like from the discussion is the level the physics actually needs, since that is what should drive the definition rather than what is convenient to label.

08 . OUR BASELINE

A pipeline that finds muons and hadrons

CANDIDATE FINDER
every plausible continuation,
at every film gap at once

—

SCORER
how well each candidate
fits the track so far

—

ASSIGNMENT
one resolution of the pool,
free to leave a track unextended

The candidate finder propagates each track end along its own direction and collects every base track lying within a position and angle window of where it should be. Every film gap enters the same pool rather than being tried only when the previous one fails, so a track that is missing on the next film does not lose a competition it was never in.

The scorer rates each candidate on how well it fits: the residual left after propagation, the change in direction, how many films were skipped, and how much competition there is for the same continuation.

The assignment resolves the whole pool at once and carries a cost for leaving a track unextended, so it may decline rather than being forced to match. Without that, every track links to something.

What the tracks then are is decided by classic criteria: an angular cut against the beam separates the muons, and the scattering along a chain separates a hadron from electromagnetic tracks, since the two look alike at a single point and differ by an order of magnitude along their length. Chain length and how a track ends carry the rest.

What it contains, and how close is it to data?

Present	labels	the species of each base track, and the ancestor a base track descends from
	pairings	whether a base track joined two segments from the same particle
	momentum	at production, not along the track
Absent	parentage	no link from a particle to what produced it, inside a shower or anywhere
	ionisation	the grain weight carries no species information

The question we cannot answer ourselves is how closely this resembles real film. The only noise present is accidental pairings, with nothing instrumental; and the measurement error appears to sit on the base track rather than on each segment, which would make pairing quality look far cleaner than a scanner would deliver. Both matter for whether a method trained here works on data.

From the simulation

01 Parentage. Without it a shower cannot be labelled as an object at all, which is also why the question of what level to divide at cannot be answered from the data as it stands.

02 Neutrino interactions on this background, particularly the shower topology the muon background covers least well.

03 Ionisation, which would normally carry most of the identification.

04 Momentum along the track, not only at production.

05 Detector effects with their truth, so alignment can be scored rather than self-validated.