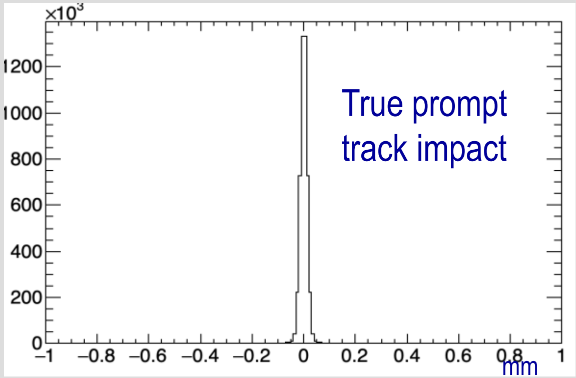
Metrics for vertex finding in particle physics

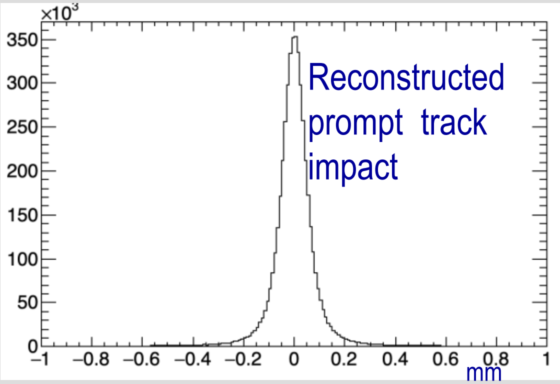
V. Kostyukhin, D.Biswas, M.Cristinziani
Siegen university

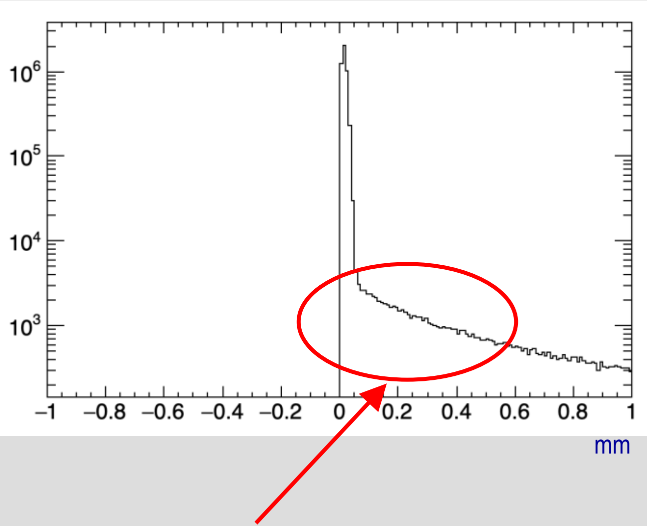
Primary+Secondary vertex challenge

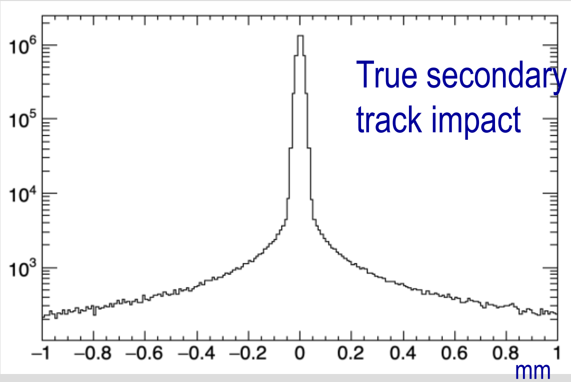
Terminology:

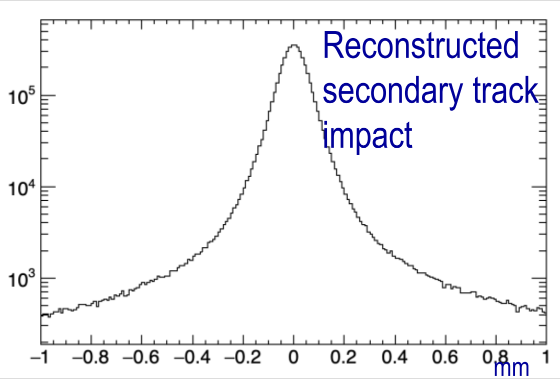
- Prompt – tracks produced in primary proton-proton interactions. Production vertices are distributed along a (beam)line
- Secondary – tracks are produced in subsequent interaction/decay vertices scattered in 3D space, sometimes very far from the beamline











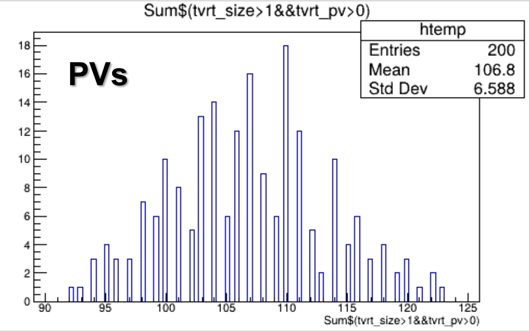
Problematic region for PV+SV reconstruction due to resolution (~equivalent to image blurring?). Hoped to resolve using ML.

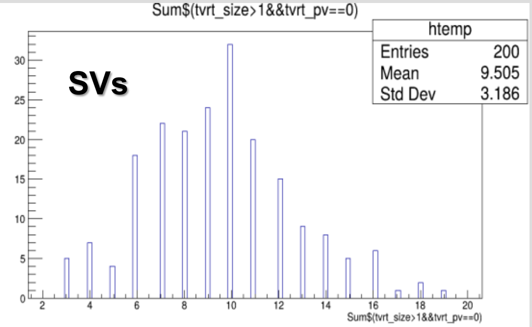
Hope – efficient PV+SV vertex finding must efficiently resolve close vertices in the region around beamline with big track density and bad resolution

Input data model (LLP, p_T>0.7GeV)

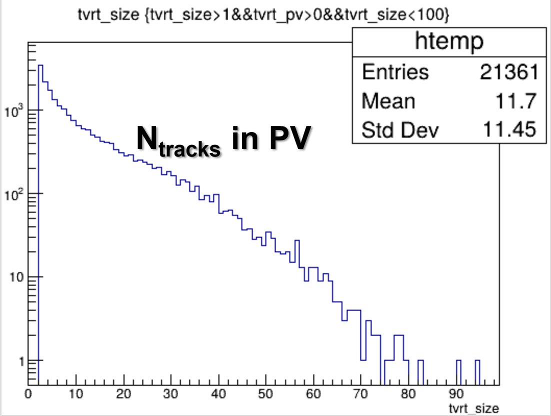
MC data features

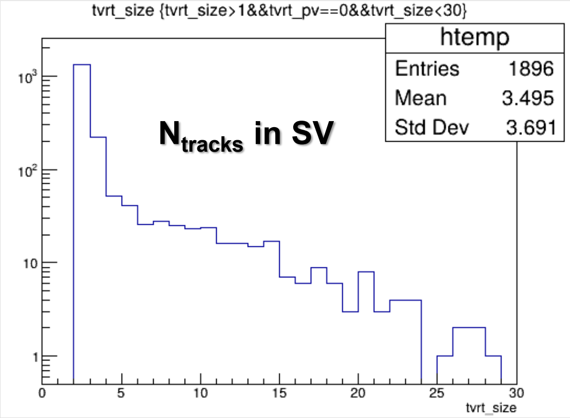
- ~1380 tracks/event (LLP sample, reco track_p_T>0.7GeV, pileup=200)
- ~7.3% secondary track (produced in SV)
- ~98 vertices/event have only one reconstructed track
  - 68% of them are SVs (interaction in material)
- ~116 vertices/events have >=2 reconstructed tracks
  - 9.5 SVs/event
  - 107 PVs/event

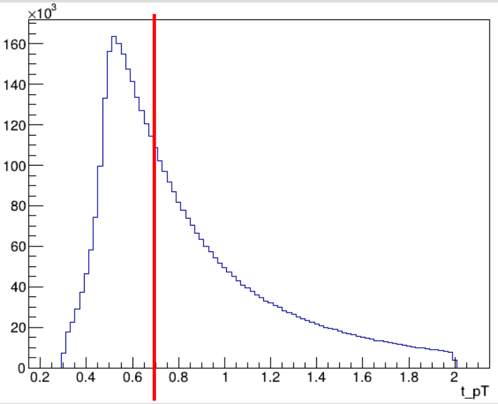




SVs and PVs categories are imbalanced (factor ~10)







Track p_T spectrum is exponentially falling:
1380(p_T>0.7GeV), 760(p_T>1.GeV), 340(p_T>1.5GeV)

Problem complexity can be easily tuned/scanned by changing the track p_T cut.
Good for the algorithm development. Physics – higher cut less physics info.

Clustering results classification

1) One-track clusters (C1t) - not vertices, garbage collector
2) Primary clusters (PVc) – made of only prompt tracks
3) Secondary clusters (SVc) – made of only secondary tracks
4) Mixed n-track primary clusters (mixPVc) – fraction of prompt tracks >50%
5) Mixed n-track secondary clusters (mixSVc) – fraction of secondary tracks >50%
6) Mixed 2-track clusters (mixC2t) – prompt+secondary tracks

Clustering metrics: statistical

Idea

After clustering select only prompt/secondary tracks and check statistical metrics for them only as these categories have different properties.

1) Variation of information (comparison with truth, VI==0 if identical.
   Smaller VI -> better clustering) [Journal of Multivariate Analysis 98 (2007) 873–895]
2) Adjusted Rand Index (comparison with truth, ARI==1 – perfect agreement,
   ARI==0 – random clustering)
3) Possibly Silhouette for PVs only. Cluster compactness vs inter-cluster distance.
   [Journal of Computational and Applied Mathematics, v20,1987, pp.53-65]

Clustering metrics: physics

Evident requirements

1) C1t/N_{tracks} → truth value
2) PVc+SVc+mixPVc+mixSVc+mixC2t -> truth number of N-track vertices
3) mixPVc →0 , mixSVc →0, mixC2t →0

Due to SV/PV imbalance, the PV reconstruction is not affected much by secondary track presence → stat metrics only for the moment (metric from PV paper comes later…)

Physics metrics

1) SV efficiency: $(SVc+mixSVc)/nSV(nTrack>1)_{true}$
2) SV purity: SVc/(SVc+mixSVc)
3) Number of mixC2t vertices: N_mixC2t
4) Secondary track fraction in mixPVc: nTsec_in_mixPVc/$N_{secondary\_tracks}$

(3)+(4) estimate how many secondary tracks are "lost" due to prompt track presence

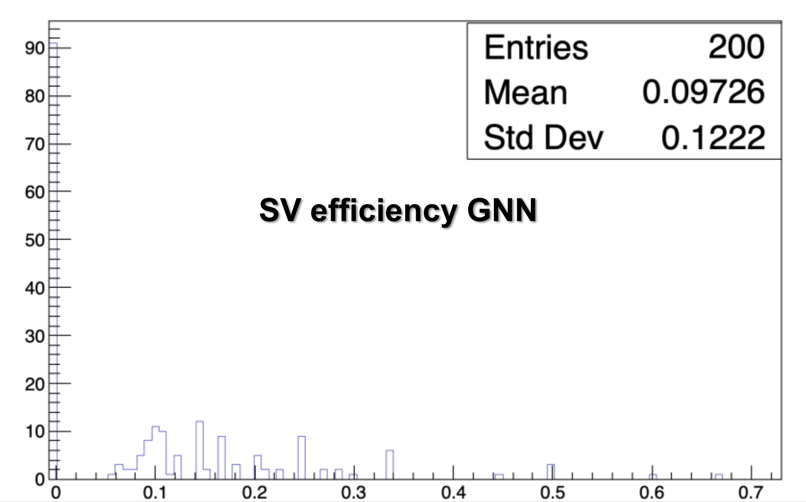
Results (LLP, p_T>0.7GeV, pu=200)

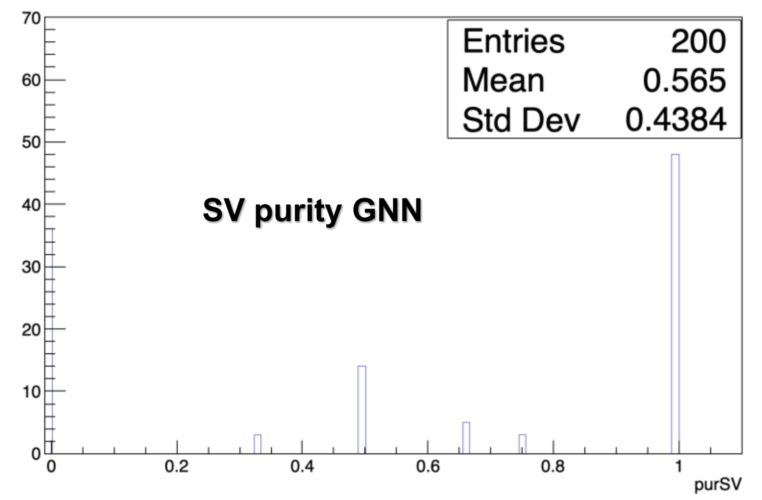
Compare LMC clustering in 2 cases

1) Edge weights are obtained in GNN processing
2) Edge weights are based on simple 2-track vertex fit χ^2

LMP weight	Ncl 1-track	Ncl N-track	effic. SVc	purity SVc	mixed 2trk clst	Lost sec.track	VI PVc	ARI PVc	VI SVc	ARI SVc
GNN wgt	321	74	10%	54%	3.0	52%	3.28	0.285	1.45	0.107
Edge Chi2	197	81	6%	43%	1.6	69%	4.40	0.153	1.89	0.095
Truth	98	116	100%	100%	0	0	0	1	0	1

Prb=TMath::Prob(Chi2,nDoF=1)

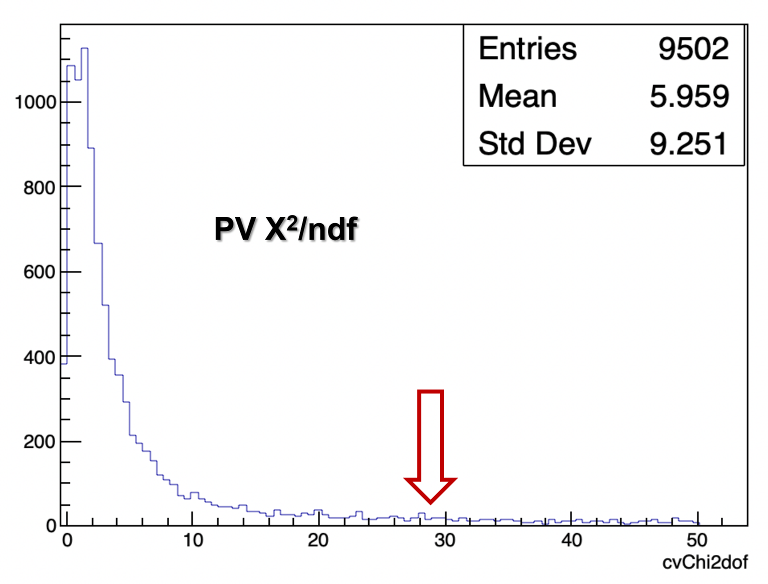


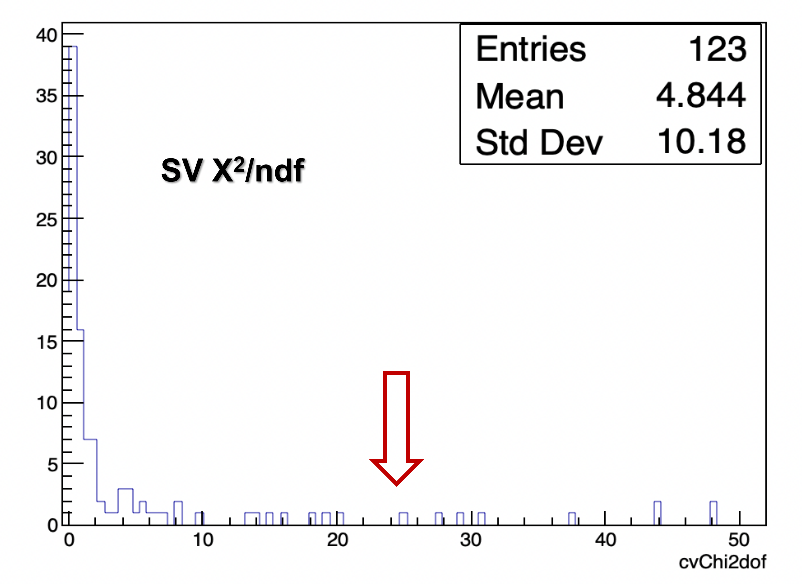


Clustering with GNN weights is significantly better

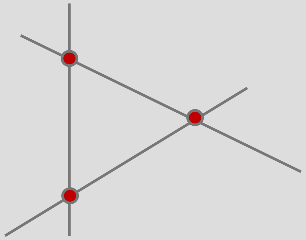
Lifted minimum cost multi-cut problem

Current LMC (no constraints, hyperedges, etc.) setup provides non-compact clusters. Can be seen in cluster vertex fit χ^2/nDoF.
They can't become physics vertices.





In SV case it might be a triangular(quartic, etc.) anomaly:



This is not a 3-track SV!

In PV case no such explanation (effective 1D space) – further study needed

Next steps

1) Metric to compare/tune simultaneous SV+PV finding is proposed
2) Graph processing does improve the vertex finding efficiency as compared to the simple vertex Chi2 based one.
3) Still far from expected:
   - LMC clustering doesn't look ideal for the moment
   - 1000ev GNN training only
   - Balancing Secondary/Prompt fractions
   - Try LMC constraints
   - Better physics information in GNN input
   - ....