Concept of measuring (Multi-)Strange Hadron Yields in the CBM Experiment using Machine Learning Techniques

Axel Puntke[1], Shantanu Bhalerao[2], Andrea Dubla[3], Felix Fidorra[1], Shahid Khan[2], Oleksii Lubynets[3,4], Philipp Munkes[1], Ilya Selyuzhenkov[3]
for the CBM Collaboration

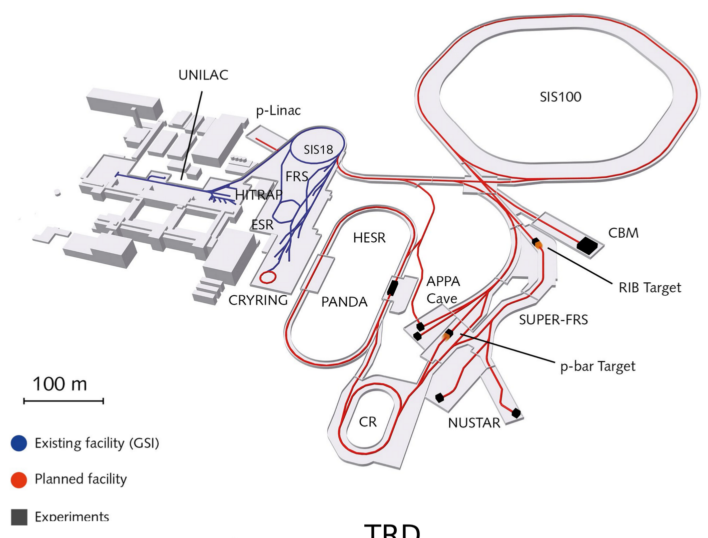
[1] WWU Münster, Germany, [2] Eberhard Karls University of Tübingen, Germany, [3] GSI Darmstadt, Germany, [4] Goethe University, Frankfurt am Main, Germany

Compressed Baryonic Matter experiment at FAIR

Experimental Background

CBM Experiment

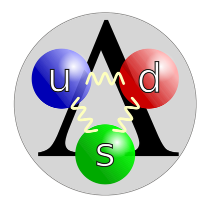
- Future fixed target heavy-ion experiment at FAIR
- Setup at FAIR with proton and ion beams
- Explores the QDC phase diagram at high net-baryon densities (μ_B > 500 MeV)
  → Search for critical point and possible QGP phases
- High interaction rates of up to 10^7 Hz at foil target
  → Multi-differential (p_T, y, centrality) measurements of rare multi-strange particles
- Increased production of strangeness in QGP phases compared to hadron phase



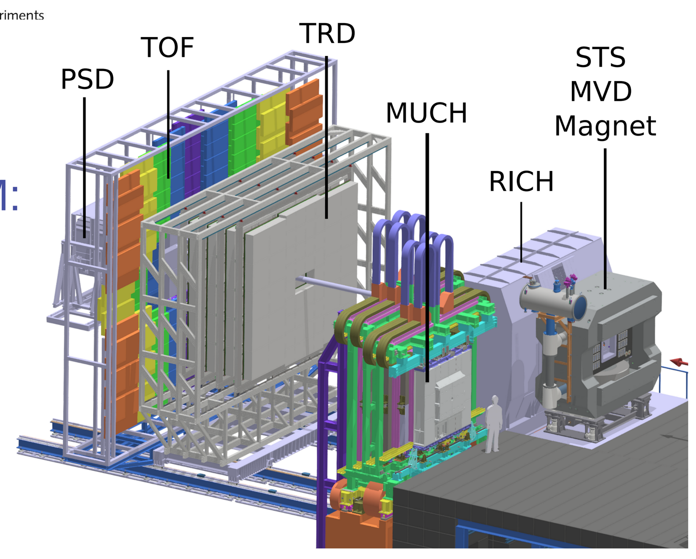
Simulation and Reconstruction

- Event generators: UrQMD, DCM-QGSM-SMM
- Simulation of interaction with CBM detector with GEANT4
- Detector response, digitization, identification & tracking with CbmRoot
- Investigate most abundantly produced hyperon: $\Lambda \rightarrow p + \pi^-$
  → Simulated: $2 \cdot 10^6$ AuAu-collisions @ 12 A GeV/c beam momentum, minimum bias

Directly reconstructed by CBM:
$e^\pm$, $\mu^\pm$, $\pi^\pm$, $p^\pm$, $K^\pm$, Ions



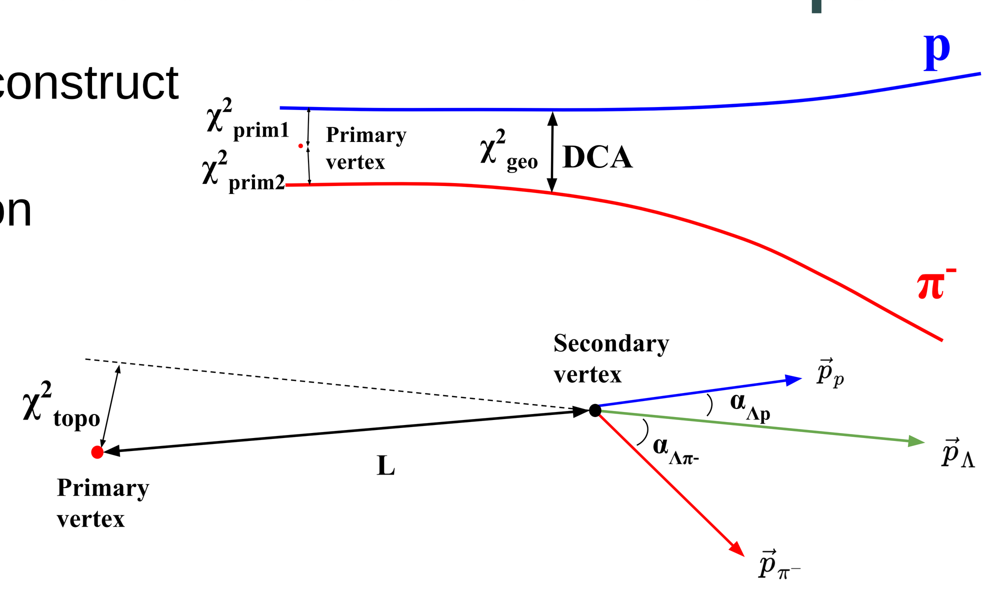
m_Λ = 1116 MeV/c²



Machine Learning Approach

Data Preparation: Reconstruction of Λ Candidates with PFSimple

- Kalman filter mathematics based C++ package to reconstruct particles via their weak decay topology
- **Input:** All reconstructed tracks from the simulation (event-wise)
- **Output:** Decay candidates for a given particle hypothesis
- Selection criteria are applied to remove numerical artifacts and reduce combinatorics



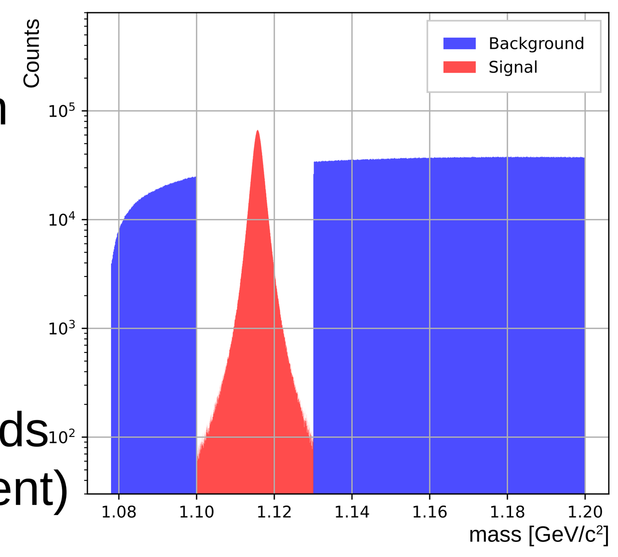
PFSimple Main Topological Variables

Variable	Description
$\chi^2_{prim1/2}$	Squared distance between daughter track and primary vertex divided by covariance matrix (CV)
χ^2_{geo}	Squared distance between daughter tracks divided by CV
χ^2_{topo}	Squared distance between mother particle trajectory and primary vertex
DCA	Distance of closest appoach between daughter tracks
L/ΔL	Distance between primary and secondary vertex divided by its error
$\alpha_{\Lambda p}$	Angle between proton and lambda track momenta

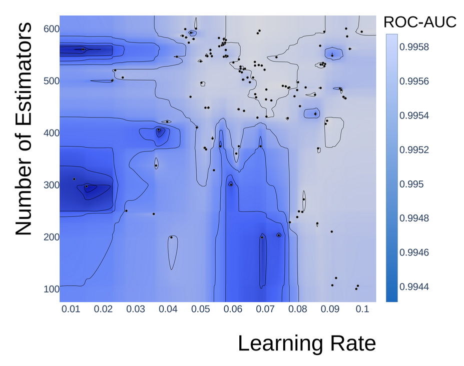
→ Additionally, TOFs m² information of the two daughter tracks is available

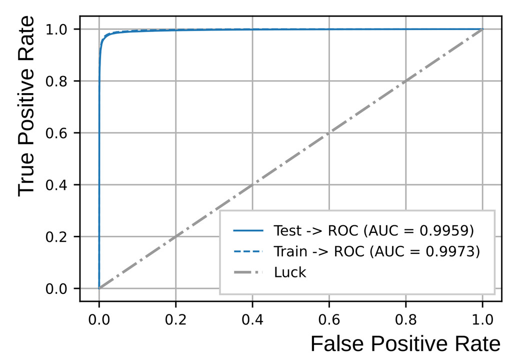
XGBoost Model Training with Hipe4ML

- XGBoost classifier is trained to distinguish signal candidates from combinatorial background using Monte Carlo information (best peforming ML model for this task)
- Data set (50/50 split test/train):
  → **Signal:** Signal-only DCM candidates in the 5 σ region around the Λ peak
  → **Background:** UrQMD candidates for Λ from side bands (proxy for real data in running experiment)



- Hyperparameters optimization: Optuna package
- Overfitting control: ROC curves and BDT probability plots for train-test sets



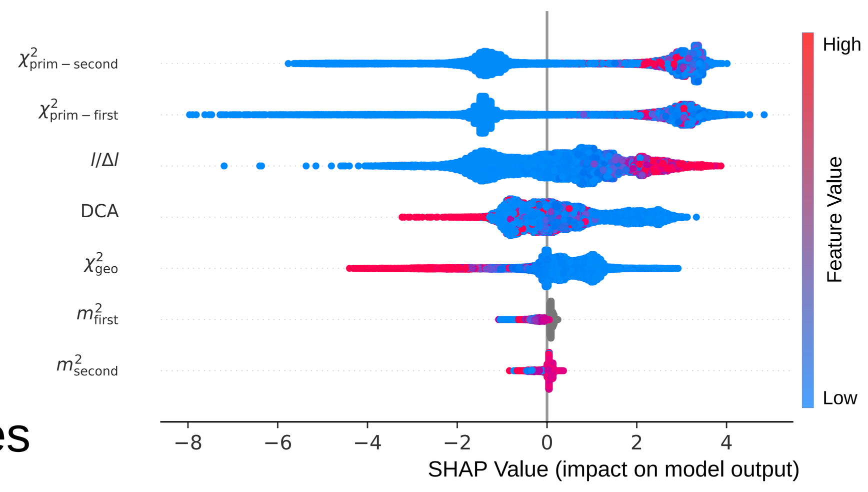


Feature Importance: SHAP Value Distributions

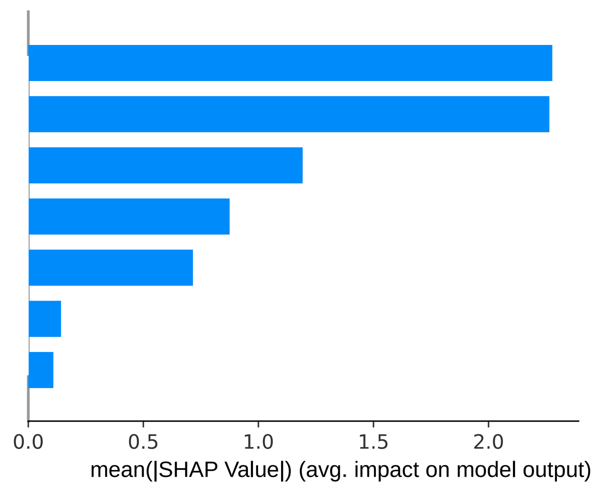
SHAP (SHapley Additive exPlanations):

- Calculate the contribution of a feature to a result
- For each candidate, an explanatory model (linear function) is fitted to the real model, which yields Shapley values when a specific fit weight is chosen
- Feature Importance: Average of these values over many candidates

SHAP Distribution

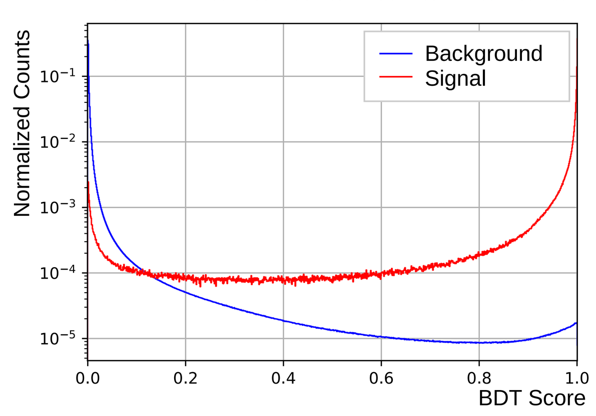


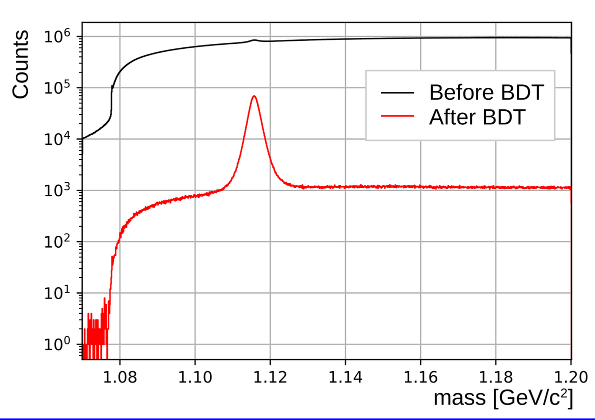
Input Variable Ranking



Resulting Λ Significance after XGBoost Selection

- Candidates with BDT score > 0.9 are selected as signal candidates
- Currently p_T integrated
- High signal to background ratio is achieved with 87.7 % efficiency





Outlook: Multi-Differential Yield Extraction

- Double-sided crystal ball function as signal distribution approximation
- Pol4 for background distribution approximation