Multi-differential Studies of Strangeness Production with the CBM at FAIR using Machine Learning Techniques

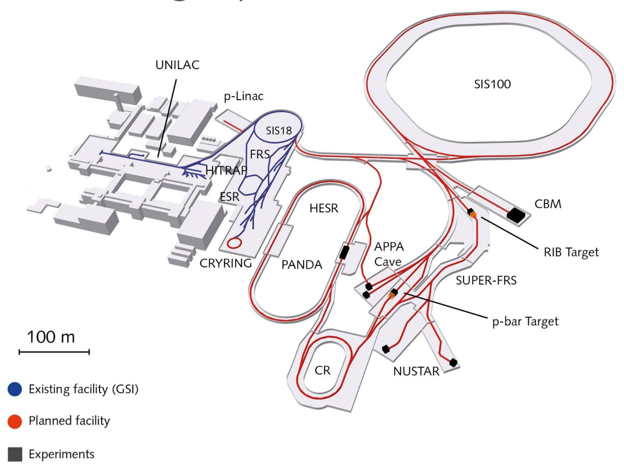
Axel Puntke[1], Andrea Dubla[2], Felix Fidorra[1], Shahid Khan[3], Lisa Katrin Kümmerer[4], Oleksii Lubynets[2,5], Ilya Selyuzhenkov[2] *for the CBM Collaboration*

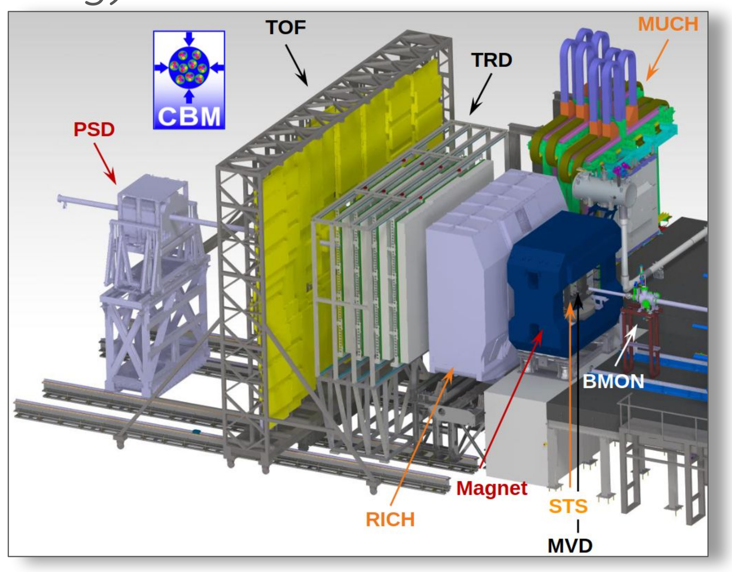
[1] WWU Münster, Germany, [2] GSI Darmstadt, Germany, [3] Eberhard Karls University of Tübingen, Germany, [4] Heidelberg University, Germany, [5] Goethe University, Frankfurt am Main, Germany

Quark Matter 2023, Houston

1. The CBM Experiment

- Future fixed target heavy-ion experiment at FAIR
- Setup at FAIR with proton and ion beams
  - Energy range $\sqrt{s_{NN}}$ = 2.9-4.9 GeV
- Explores the QCD phase diagram at high net-baryon densities (μ_B > 500 MeV)
  - Search for critical point and QGP-hadrons phase transition
- High interaction rates of up to 10^7 Hz
  - Precise Multi-differential (p_T, y, centrality) measurements of rare multi-strange particles

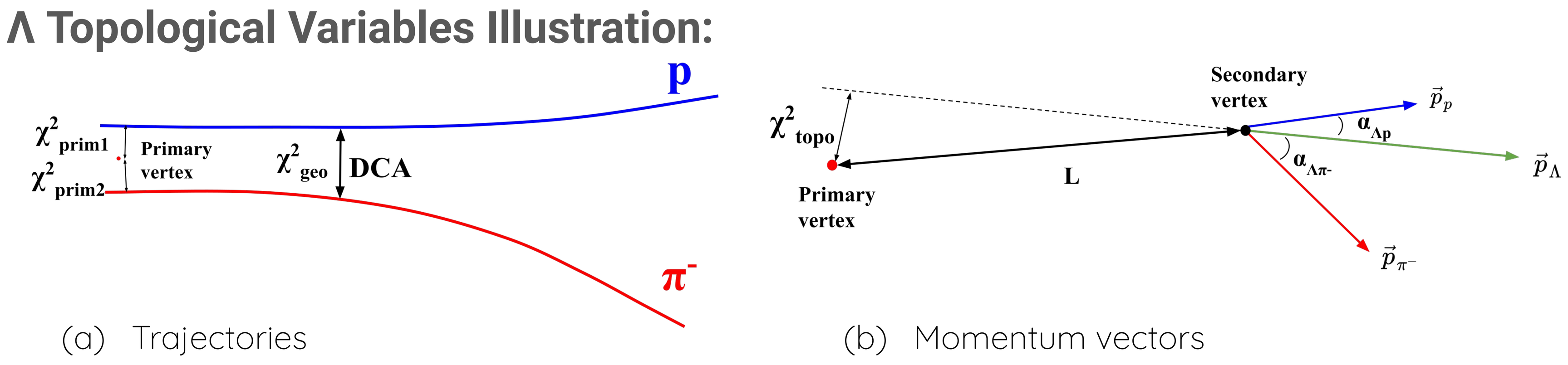




2. Reconstruction of short-lived strange Particles

- Performed with **PFSimple**, a KFParticleFinder-based C++ package optimized for reconstruction of short-lived particles
- Combines CbmROOT tracks ($e^\pm$, $\mu^\pm$, $\pi^\pm$, $p^\pm$, $K^\pm$, Ions) to mother particle candidate according to their PID hypothesis and corresponding decay channels
- Computes various topological variables for decay candidates
- Challenge: Reject combinatorial background candidates

Λ Topological Variables Illustration:



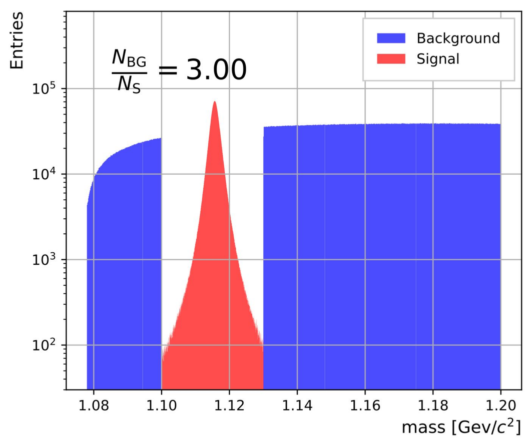
(a) Trajectories (b) Momentum vectors

3. Λ Selection using XGBoost

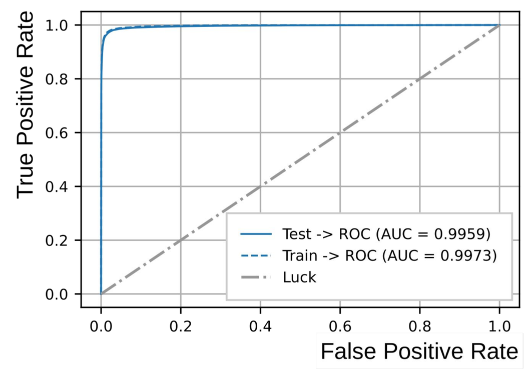
- XGBoost: Boosted Decision Tree (BDT) algorithm with some additional features
  - XGBoost shows better performance compared to other ML methods
  - Boosted decision tree model maps feature input vector of a candidate to the probability of being a signal candidate (BDT score)

Model Training:

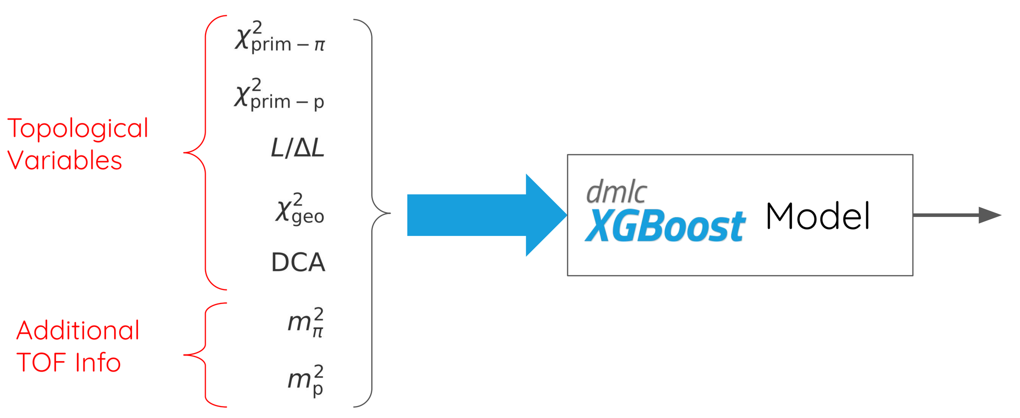
- Au-Au collisions @ $\sqrt{s_{NN}}$ = 4.93 GeV mbias
- Signal: DCM-QGSM-SMM in 5σ region around Λ peak
- Background: UrQMD (final experiment: real data)

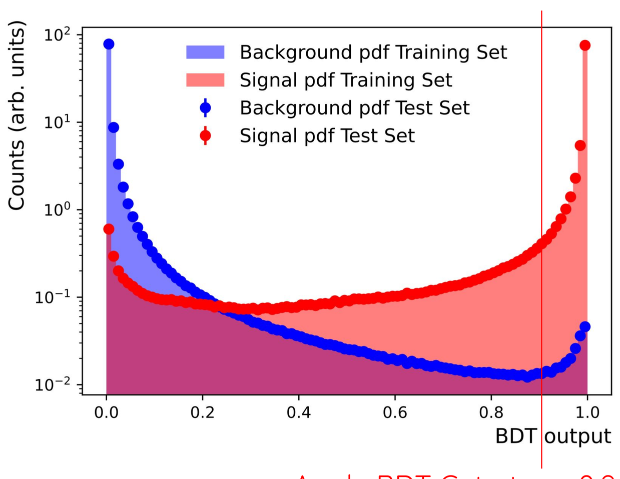


- Hyperparameter optimization using Optuna package
  - Maximization of accuracy using bayesian optimization and 5-fold cross validation
- Overfitting control: ROC curves and BDT probability plots have to match for train and test sets



Final Model:

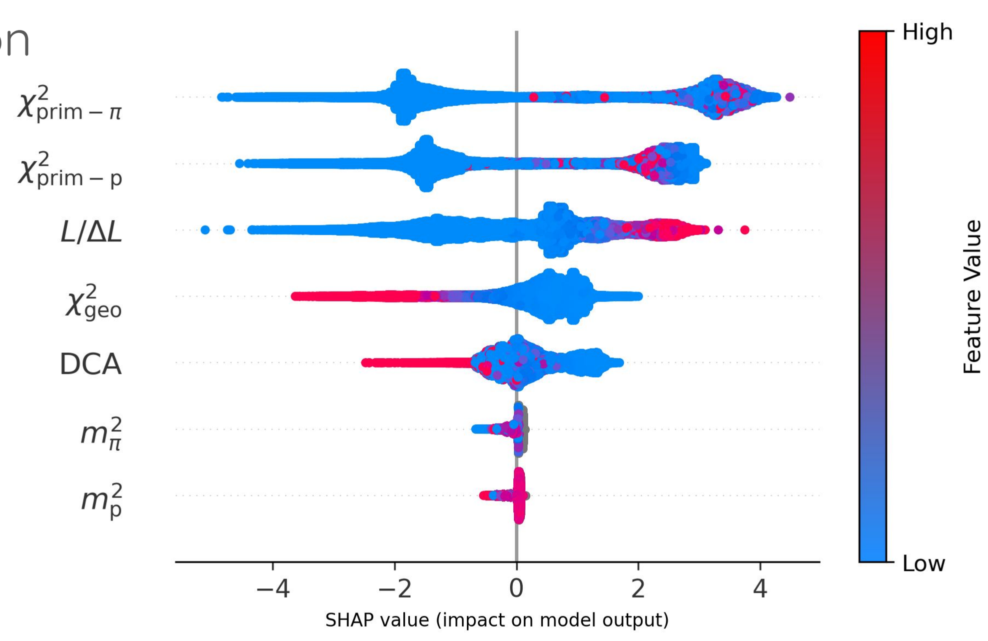




Apply BDT Cut at e.g. 0.9 to classify as signal or background

Feature Importance:

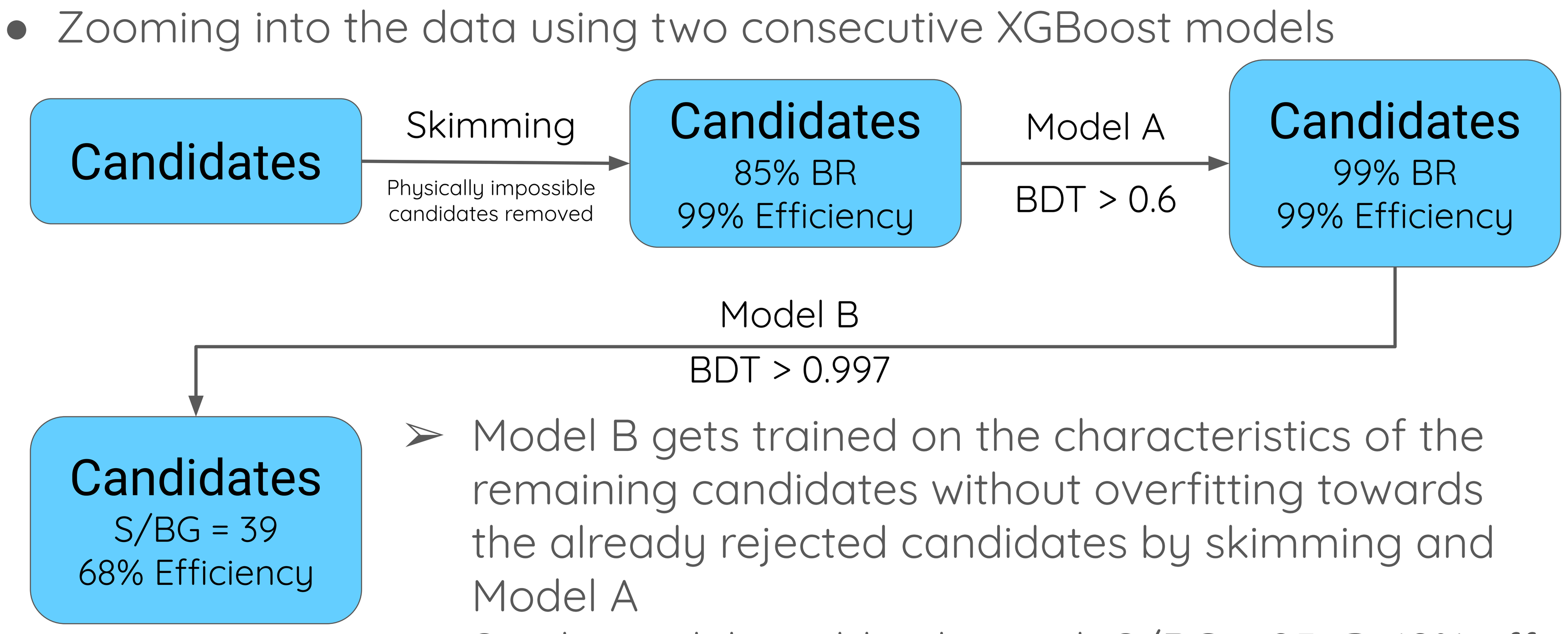
- Insights on how the specific feature variables contribute to the models decision are gained by computing SHAP values
  - SHAP value ~ contribution to being signal
- E.g. low values of $\chi^2_{prim-\pi}$ often lead to classification as background



SHAP Beeswarm plot for 10k candidates, features ordered in descending importance

4. Iterative Ξ Selection

- Zooming into the data using two consecutive XGBoost models

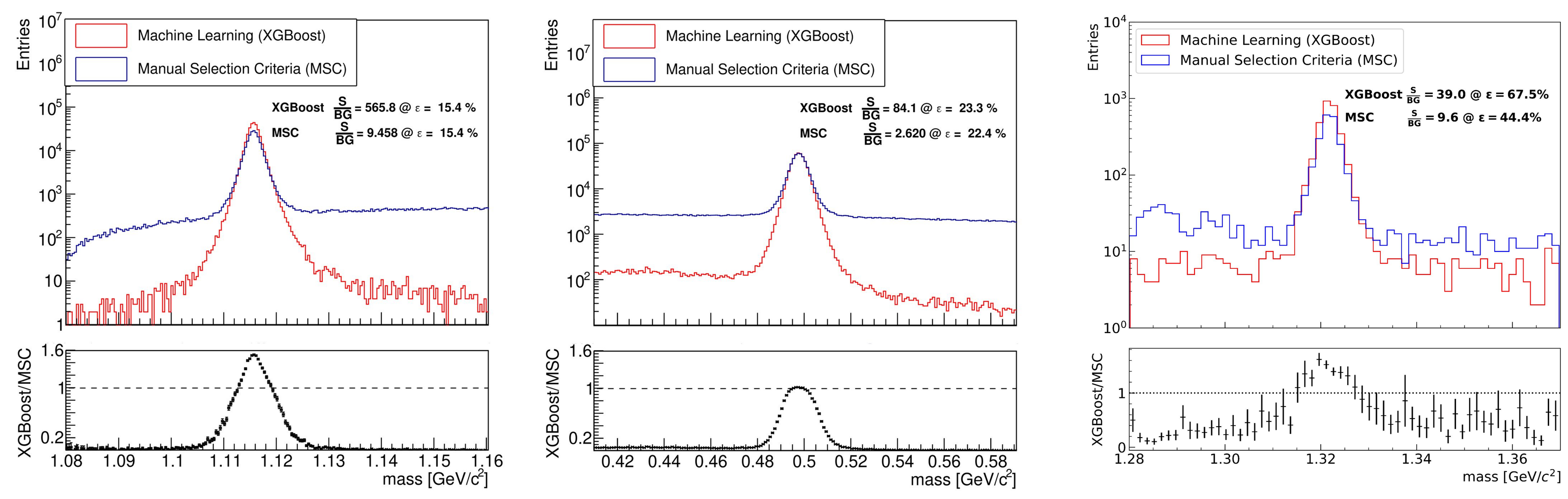


- ➢ Model B gets trained on the characteristics of the remaining candidates without overfitting towards the already rejected candidates by skimming and Model A
- ➢ Single model would only reach S/BG = 25 @ 49% eff.

5. XGBoost compared to classical Box Selection

- Classically candidates are selected by applying box selection criteria on their topological variables to maximize S/BG ratio
  - Linear selection only & laborious task to optimize them for every collision system and energy separately
- Machine learning allows non-linear multi-dimensional selection with automatic training on large data sets at much better S/BG ratio at same efficiency

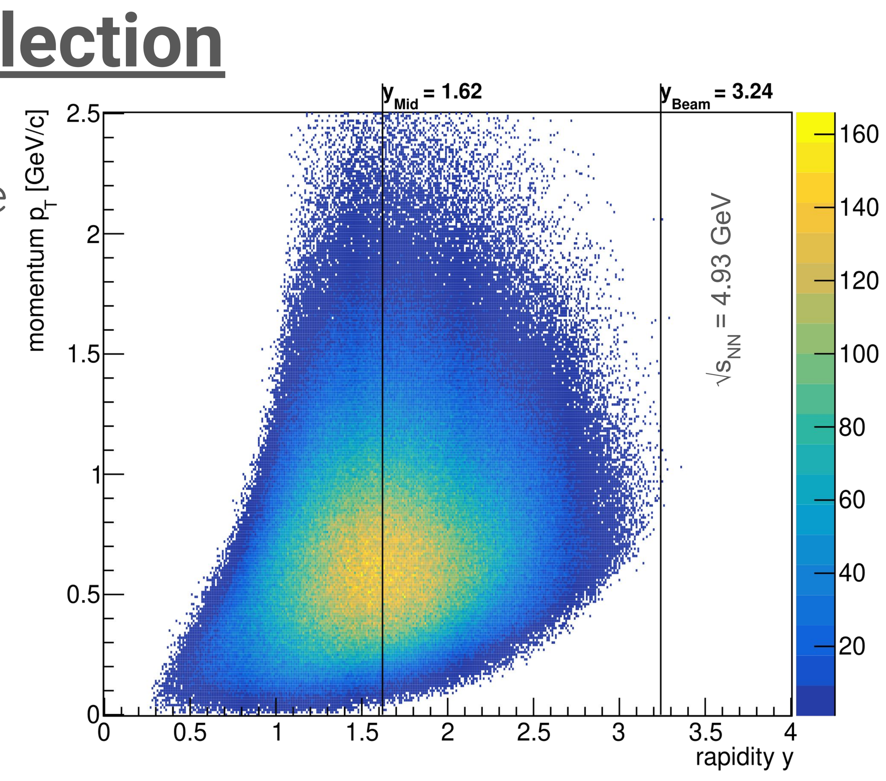
Resulting S/BG Ratios:



(a) Λ^0 (b) K^0_S (c) Ξ^-

6. Multi-Differential Λ Selection

- Model performance is evaluated over CBMs whole p_T-y phase space
- Training separate models for different p_T-y intervals further improves the model performance
  - 8 - 22 % increase in S/BG ratio per interval at same efficiency as p_T-y-integrated model achieved



Selected MC-true candidates by XGBoost model working at 90% ML efficiency and S/BG = 3.18