

August 30, 2024

Summer Student Report

Particle Identification capabilities of ALICE Inner Tracking System

Author:

Giorgio Alberto Lucia
giorgio.alberto.lucia@cern.ch

Supervisors:

Francesco Mazzaschi
Stefano Politanò

Abstract

The Inner Tracking System is a silicon pixel detector designed with very a high spatial resolution and a low material budget. To handle the high interaction rates of LHC Run 3, the detector employs a digital readout system. This work focuses on studying the cluster topology of signals left by charged particles passing through the detector. By combining this information with the momentum of the particles, it is possible to infer the particle species. The performance of particle identification is further improved through the use of machine learning algorithms.

1 Introduction

ALICE (A Large Ion Collider Experiment) is one of the largest experiments at the Large Hadron Collider (LHC) at CERN. The primary goal of ALICE is to study the physics of strongly interacting matter at extreme energy densities, where a state of matter known as quark-gluon plasma is formed. The experiment primarily focuses on Pb-Pb collisions, which are characterized by very high multiplicity events. The Inner Tracking System (ITS) plays a key role in reconstructing these events, tracking charged particles and providing a highly precise reconstruction of the primary and secondary vertices. The ITS is a silicon pixel detector consisting of seven layers. The detector utilizes Monolithic Active Pixel Sensors, achieving very high spatial resolution (with a pixel size of $29.24 \mu\text{m} \times 26.88 \mu\text{m}$) and a very low material budget ($0.35\% X_0$) [1].

Each pixel in the ITS can be described as a diode made of two layers of silicon with different doping types. When a charged particle passes through a silicon sensor, it deposits energy that generates electron-hole pairs in the material. The electrons drift through silicon and are accelerated by an electric field in a region surrounding the anode. These electrons produce a signal at the anode, which detects the passage of the charged particle. The tracker is designed with a digital readout to handle the interaction rates of 50 kHz and 500 kHz at LHC Run 3 for Pb-Pb and pp collisions, respectively. The pixels have a binary output, which is positive when the signal at the anode exceeds a set threshold. The number of electrons produced is proportional to the specific energy loss of the charged particle. Electrons produced by ionization in the silicon layer are typically collected by multiple neighboring pixels. The number of pixels activated by a single charged particle passing through a layer is called cluster size. This work focuses on characterizing the cluster size for different particle species and combining this variable with momentum to achieve particle identification (PID) with the ITS.

2 Dataset composition

This work is completely data driven, using data collected in 2022 and 2023 during LHC Run 3, and focuses on the characterization of the cluster size of π^\pm , K^\pm , (anti-)p, Ξ^\pm , Ω^\pm , (anti-)d and (anti-) ^3He . To fully understand the cluster size distribution for different species, we need to select them with very high purity. Different selection criteria are applied for the different species. π and p are selected as daughters of a Λ particle. The Λ is a neutral particle that decays in two charged particle, $\Lambda \rightarrow \pi + p$. Since the mother is not charged, only the tracks of the daughters are reconstructed by the ITS and by the TPC. Particles that decay with this particular topology are called V0 and are selected with very high purity by reconstructing the kinematics of the decay. Ω^\pm and Ξ^\pm decay into a charged particle and a V0, the latter of which in turn decays into two charged particles. The decays are $\Omega \rightarrow K + \Lambda \rightarrow K + \pi + p$ and $\Xi \rightarrow K + K_S^0 \rightarrow K + \pi + \pi$. This decay topology, called cascade, is identified in ALICE using the strangeness tracking technique [2]. The K^\pm are selected as daughters of the Ω^\pm . d and ^3He are selected through the Time-Of-Flight (TOF) detector and the Time Projection Chamber (TPC), two of the main detectors in ALICE that can achieve particle identification (PID) by measuring momentum and time-of-flight, and momentum and specific energy loss, respectively. To characterize the behavior of the cluster size for different particle species, we focus on the average ITS cluster size, $\langle \text{ITS Cluster Size} \rangle$. This variable is defined as the average cluster size measured across all ITS layers that registered a hit for a given track. We multiply $\langle \text{ITS Cluster Size} \rangle$ by $\cos \lambda$ to compensate for the fact that the cluster size also depends on the angle between the track and the ITS layer, where $\cos \lambda = p_T/p$.

From the $\langle \text{ITS Cluster Size} \rangle \times \cos \lambda$ distribution plotted as a function of momentum (Fig. 1), two features of the cluster size are promising: the average cluster size scales with momentum and in a given momentum slice, the cluster size distributions for different species are distinct. These features are characteristic of the specific energy loss.

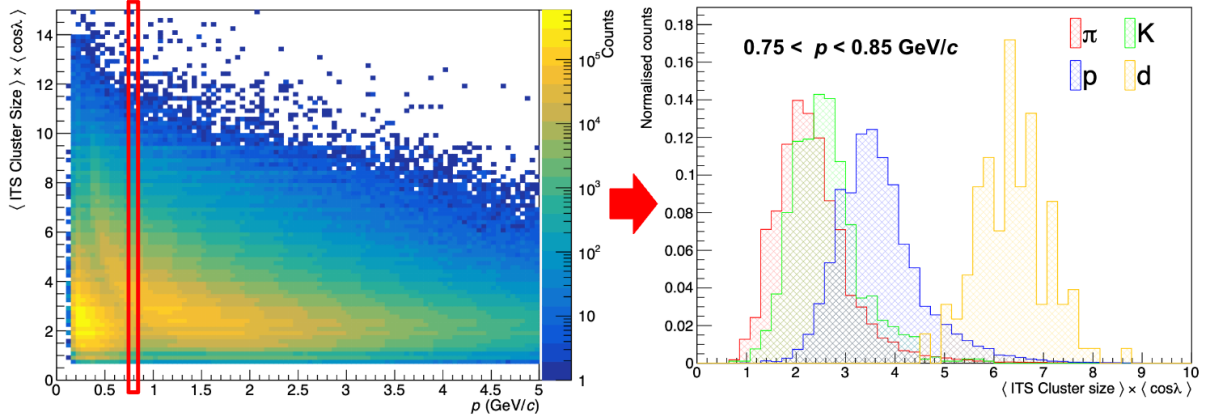


Figure 1: (a) Distribution of $\langle \text{ITS Cluster Size} \rangle \times \cos \lambda$ with respect to p for Ppim, K^\pm , (anti-)p, (anti-)d. (b) Distribution of $\langle \text{ITS Cluster Size} \rangle \times \cos \lambda$ for different particle species in an interval of momentum of [750, 850] MeV/c.

3 Fake matches

The cluster size information is proven to be very useful in identifying the fake match effect. This term refers to cases where a proton track reconstructed in the ITS is incorrectly associated with a track of a different particle reconstructed in the TPC. The global track (ITS + TPC) is often classified as a proton track, but has poorer quality due to incorrect association. In the plot in Fig. 2, with $\langle \text{ITS Cluster Size} \rangle \times \cos \lambda$ on the y-axis and the relativistic $\beta\gamma$ on the x-axis, these tracks are visible as they form an independent distribution, with an approximately constant value of $\langle \text{ITS Cluster Size} \rangle \times \cos \lambda$ (~ 2). The two distributions are well distinguished at low momentum, and it is possible to fit them in different $\beta\gamma$ bins with a double Gaussian. The fake match probability is defined as the ratio between the integral of the fake match Gaussian and the sum of the integrals of both distributions. This effect is not negligible. The cluster size information allows us to quantify the fake matches and even eliminate them at low values of $\beta\gamma$.

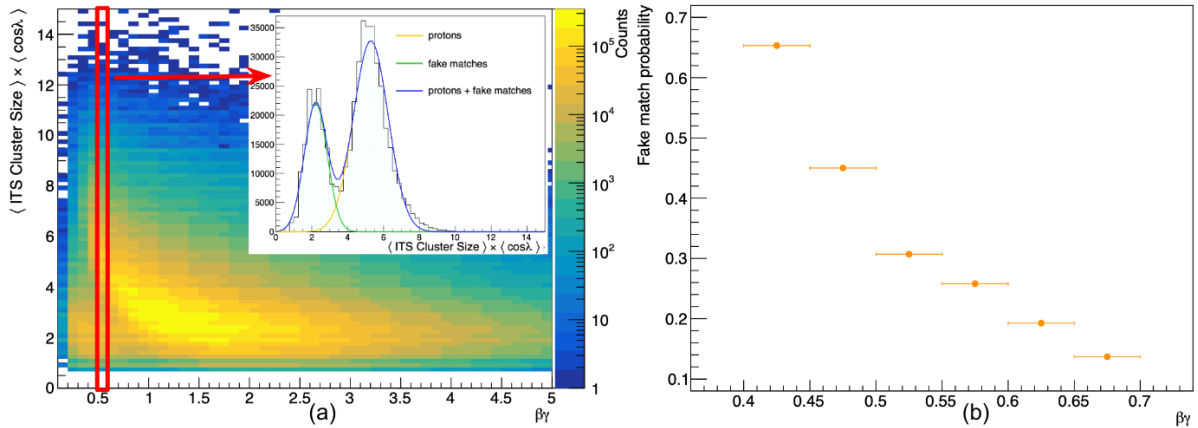


Figure 2: (a) Distribution of $\langle \text{ITS Cluster Size} \rangle \times \cos \lambda$ with respect to $\beta\gamma$ for protons. Fake matches are visible in the bottom left corner. An example of the double Gaussian fit is shown in the subplot. (b) Fake match probability for protons in different bins of $\beta\gamma$.

4 Bethe-Bloch parametrization

The number of pixels that a single charged particle activates when passing through a layer of the ITS is expected to be proportional to the specific energy loss, $-\langle dE/dx \rangle$, of the particle. This energy loss

depends on the particle species (specifically its charge z), its velocity β , and the properties of the material (such as electron density n , mean excitation energy I , and the atomic number of its constituents Z). This relationship is expressed in the Bethe-Bloch formula:

$$-\left\langle \frac{dE}{dx} \right\rangle = \frac{4\pi}{m_e c^2} \left(\frac{e^2}{4\pi\epsilon_0} \right)^2 \frac{nz^2}{\beta^2} \left[\ln \left(\frac{2m_e \beta^2 c^2}{I} \right) - \ln(1 - \beta^2) - \beta^2 \right]$$

If the cluster size is proportional to the specific energy loss, it is expected that the cluster size will be compatible between different species with $z = 1$ for particles with the same $\beta\gamma$. This is verified by fitting the $\langle \text{ITS Cluster Size} \rangle \times \cos\lambda$ distribution of individual species in multiple intervals of $\beta\gamma$. The distribution is fitted with a exponentially modified Gaussian function:

$$f(x; N, \mu, \sigma, \tau) = \begin{cases} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), & \text{if } x \leq \mu + \tau\sigma \\ \exp\left(-\left(\frac{x-\mu}{\sigma} + \frac{\tau}{2}\right)\tau\right), & \text{otherwise} \end{cases}$$

The procedure is repeated as a function of the momentum, and the mean of the distribution obtained from the fit is shown in the plots in Fig 3.

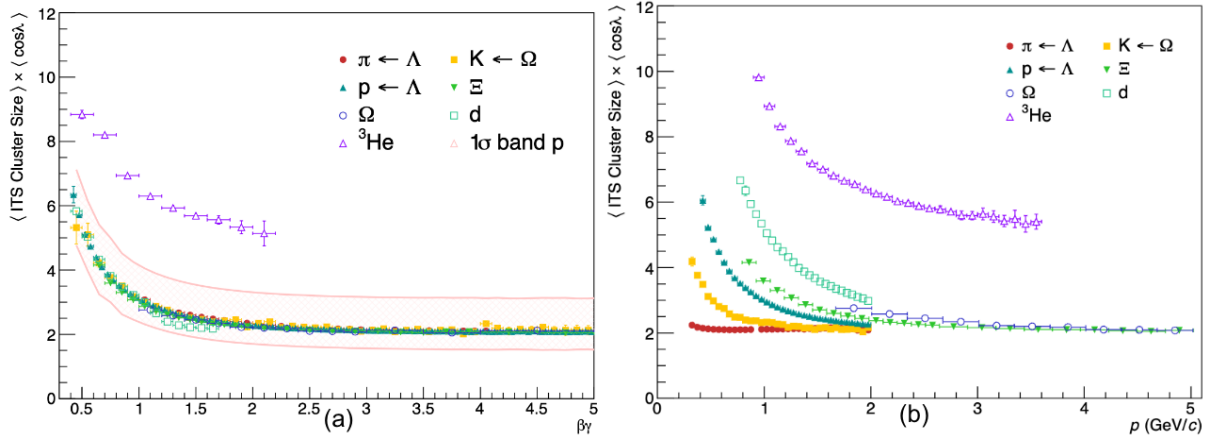


Figure 3: (a) Mean value of the $\langle \text{ITS Cluster Size} \rangle \times \cos\lambda$ from the fits for individual species as a function of $\beta\gamma$. (b) Mean value of the $\langle \text{ITS Cluster Size} \rangle \times \cos\lambda$ from the fits for individual species as a function of momentum

A 1σ band is displayed for protons in the plot with $\beta\gamma$ on the x-axis, where the σ parameter from the fits is used as the width of the band. As expected, the ITS cluster size of particles with $z = 1$ as a function of $\beta\gamma$ is consistent across different species. The cluster size of ^3He , the only particle with $z = 2$ in the dataset, is well separated from the other species. A parametrization of the $\langle \text{ITS Cluster Size} \rangle \times \cos\lambda$ as a function of $\beta\gamma$ for particles with $z = 1$ is achieved by fitting the proton data points in the plot in Figure 3 (b) using the following Bethe-Bloch parametrization:

$$\text{ExpectedITSClusterSize} = \left\{ [kp2] - \left(\frac{\beta\gamma}{\sqrt{1 + (\beta\gamma)^2}} \right)^{[kp4]} - \ln \left((\beta\gamma)^{-[kp5]} + [kp3] \right) \right\} \cdot [kp1] \cdot \left(\frac{\sqrt{1 + (\beta\gamma)^2}}{\beta\gamma} \right)^{[kp4]}$$

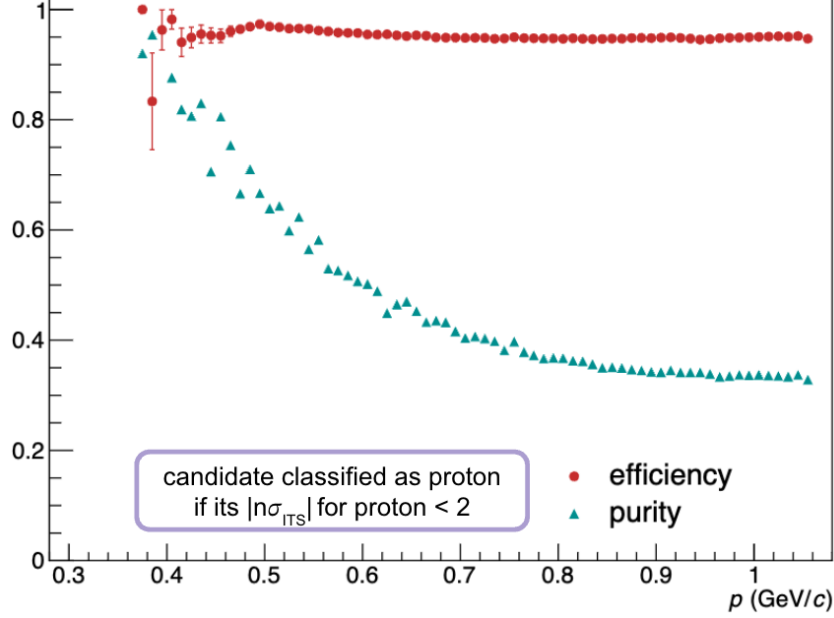
The parameters obtained from the fit are listed in Table 1.

Figure 3 (b) shows that the average value of the $\langle \text{ITS Cluster Size} \rangle \times \cos\lambda$ for different species is separated at low momentum. We define the $n\sigma_{ITS}^X$, $n\sigma_{ITS}$ variable for the species X, as follows:

$$n\sigma_{ITS}^X = \frac{(\langle \text{ITS Cluster Size} \rangle \times \cos\lambda)_{\text{expected}}^X - (\langle \text{ITS Cluster Size} \rangle \times \cos\lambda)_{\text{observed}}}{\sigma_{ITS}^X},$$

Table 1: Fit parameters

$kp1$	-0.031712	$kp2$	-45.0275
$kp3$	-0.997645	$kp4$	1.68228
$kp5$	0.010848		

Figure 4: Efficiency and purity of protons selecting requiring that the candidate has a $|(n\sigma_{ITS}^p)| \leq 2$

where $(\langle \text{ITS Cluster Size} \rangle \times \cos \lambda)^X_{\text{expected}}$ is the value obtained from the parametrization in Table 1 for a particle with $\beta\gamma = p / m_X$ and σ_{ITS}^X is the value from the exponentially modified Gaussian fits for the particle X at momentum p and p is the momentum of the candidate. It is possible to achieve PID by assigning the species X to a candidate if its $n\sigma_{ITS}^X$ is contained in an interval $[th_a, th_b]$. The performance of the PID with the $n\sigma_{ITS}$ cut technique is evaluated by measuring the efficiency and purity of the selection. These two variables are defined as follows:

$$\text{Efficiency}_X = \int_{th_a}^{th_b} dn\sigma_{ITS}^X f_X(n\sigma_{ITS}^X)$$

$$\text{Purity}_X = \frac{\int_{th_a}^{th_b} dn\sigma_{ITS}^X f^X(n\sigma_{ITS}^X)}{\int_{th_a}^{th_b} dn\sigma_{ITS}^X (f^X(n\sigma_{ITS}^X) + \sum_{i \neq X} f^i(n\sigma_{ITS}^X))},$$

where $f^i(n\sigma_{ITS}^X)$ is the normalised $(n\sigma_{ITS}^X)$ distribution for the species i . Figure 4 shows the efficiency and purity values for the proton as a function of the momentum labelling as protons all the candidates with $|(n\sigma_{ITS}^p)| \leq 2$. For this study, only π , K and p are considered in the dataset.

This result demonstrates that the ITS cluster size information can be reliably used to achieve PID using ITS information alone, with good performance.

5 Machine Learning classifier

The result presented in Figure 4 can be further improved using machine learning algorithms. Unlike traditional algorithms that follow explicitly programmed instructions, machine learning algorithms learn directly from data. The goal of the algorithm is to differentiate between classes, starting from a set of examples known as the training set. Each element of the training set has a label containing its class,

which is assigned based on the selections described in paragraph 2. The training process adjusts the internal parameter of the algorithm to maximize the separation power between the different classes.

For this study, a Boosted Decision Tree (BDT) classifier was used, specifically the XGBoost implementation [3]. A BDT is a machine learning algorithm that constructs multiple smaller models, called decision trees, and combines their predictions. A decision tree organizes data into a tree-like model of decisions based on the features of the data. Each internal node represents a decision rule, while each leaf node represents an outcome. During training, a BDT creates a new decision tree at each iteration, refining the rules to improve the prediction accuracy.

The input features for the BDT classifier included the cluster size ($\langle \text{ITS Cluster Size} \rangle \times \cos \lambda$), the average ITS cluster size ($\langle \text{ITS Cluster Size} \rangle$), the cosine of the lambda angle ($\cos \lambda$), the azimuthal angle in the transverse plane (ϕ), the pseudorapidity (η), and the ITS cluster size on the sixth and seventh layers of the detector¹. The model outputs probability scores for each candidate, representing the likelihood of the candidate belonging to each of the species considered during training. The sum of these probability scores for each candidate is 1.

In this study, only π , K, and p are considered. The BDT is sensitive to class imbalance in the training set, and its performance is expected to vary with the particle's momentum because the separation between species based on the $\langle \text{ITS Cluster Size} \rangle \times \cos \lambda$ distribution decreases at higher momentum. To address the class imbalance, a bootstrap method was applied to the training dataset to create a uniformly distributed dataset of candidates across species and momentum. To account for momentum dependence, seven independent classifiers were trained in non-overlapping momentum intervals ranging from 350 to 1050 MeV/c, each with a width of 100 MeV/c.

A candidate passed to the trained classifier is associated with three probability scores. To assign a species to the candidate, its probability score is compared to a threshold. If the score exceeds the threshold, the corresponding label is assigned to the particle.

The performance of each classifier was evaluated by computing the efficiency and purity of the selection for different threshold values. Defining as $f^i(\text{score}_X)$ the normalised distribution of the probability score X for the candidates with true label i, efficiency and purity can be computed as follows:

$$\text{Efficiency}_X = \int_{\text{th}}^1 d(\text{score}^X) f^X(\text{score}^X)$$

$$\text{Purity}_X = \frac{\int_{\text{th}}^1 d(\text{score}^X) f^X(\text{score}^X)}{\int_{\text{th}}^1 d(\text{score}^X) (f^X(\text{score}^X) + \sum_{i \neq X} f^i(\text{score}^X))}.$$

Figure 5 shows a plot of efficiency versus purity for proton selection in the momentum interval of [350, 450] MeV/c. The points corresponding to the efficiency and purity obtained with a selection requiring $|n\sigma_{ITS}^p| \leq \text{th}$ for thresholds (th) of 1, 2, and 3 are also shown.

The classifier outperforms the $n\sigma$ cut technique, particularly when high purity is required. Additionally, this method is very flexible, allowing users to choose a working point on the graph by setting the threshold, thereby tuning the selection to their desired efficiency and purity.

¹The exclusion of the cluster size from the other layers of the detector is related to the composition of the dataset used. It has been observed that π and p candidates rarely have hits on the first layers, while this is more common for K. This occurs because π and p are selected as daughters of a Λ , which has a longer lifetime than the Ω , from which the K are selected. Since this model is intended to be applied to more general datasets, where this may not be the case, only the cluster size information from the external layers is considered for the training process.

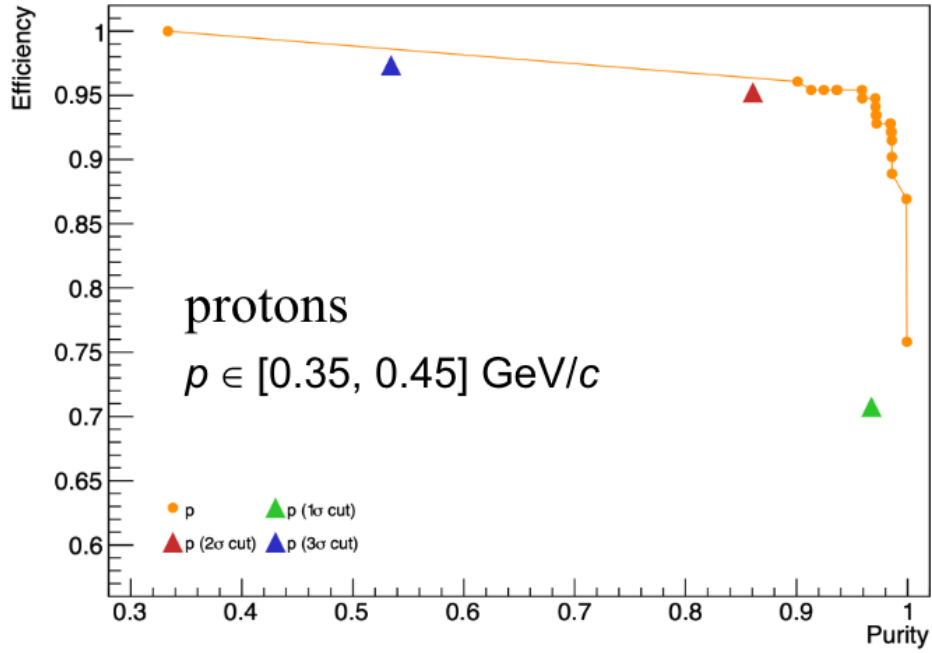


Figure 5: Efficiency and purity of protons selected based on the classifier output. A candidate is labeled as a proton if its proton output score exceeds a set threshold. The orange points represent selections for different threshold values. The green, red, and blue points show efficiency and purity for selections using the $n\sigma$ cut technique with symmetric cuts at 1, 2, and 3 σ , respectively.

6 Conclusions

This work has demonstrated the effectiveness of using cluster size information from the ITS for particle identification in the ALICE experiment. By examining the cluster size distributions for various particle species and applying a Bethe-Bloch parametrization, we showed that the ITS can distinguish between different particle types with high accuracy. Furthermore, we identified the fake match effect and quantified its impact, highlighting the importance of cluster size in improving track quality. Finally, the implementation of machine learning techniques, such as Boosted Decision Trees, further enhanced the precision of particle identification, surpassing traditional methods in both efficiency and purity.

Acknowledgements

I am deeply grateful to my supervisors for their constant guidance and availability. Their support, along with the vibrant environment fostered by the entire group from Turin, has been invaluable and filled with enriching discussions. I would also like to thank my fellow summer students for the laughter, conversations, and shared adventures—I feel fortunate to have met such wonderful people.

References

- [1] B. A. et al and T. A. Collaboration), “Technical design report for the upgrade of the alice inner tracking system”, *Journal of Physics G: Nuclear and Particle Physics* **41** no. 8, (Jul, 2014) .
<https://dx.doi.org/10.1088/0954-3899/41/8/087002>.
- [2] D. Dobrigkeit Chinellato, “Charm and multi-charm baryon measurements via strangeness tracking with the upgraded alice detector”, *EPJ Web of Conferences* **259** (2022) .
<http://dx.doi.org/10.1051/epjconf/202225909004>.
- [3] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system”, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16. ACM, Aug., 2016. <http://dx.doi.org/10.1145/2939672.2939785>.