Evaluating Analysis Description Language Concept as a First Introduction to Analysis in Particle Physics

A. Adıgüzel¹, O. Çakır², Ü. Kaya³, V. E. Özcan³,⁴, S. Öztürk⁵, S. Sekmen⁶, İ. Türk Çakır⁷, G. Ünel⁸

¹ İstanbul University, Physics Dept., İstanbul, Turkey
² Ankara University, Physics Dept., Ankara, Turkey
³ Boğaziçi University, Physics Dept., İstanbul, Turkey
⁴ Boğaziçi University, Feza Gürsey Center for Physics and Mathematics, İstanbul, Turkey
⁵ Gaziosmanpaşa University, Physics Dept., Tokat, Turkey
⁶ Kyungpook National University, Physics Dept., Korea
⁷ Giresun University, Dept. of Energy Systems Engineering, Giresun, Turkey
⁸ University of California, Irvine, Physics and Astronomy Dept., USA

August 28, 2020

Abstract

The fifth edition of the “Computing Applications in Particle Physics” school was held on 3-7 February 2020, at Istanbul University, Turkey. This particular edition focused on the processing of the (pseudo) data from the Large Hadron Collider collisions using an Analysis Description Language, namely CutLang. 24 undergraduate and 6 graduate students were initiated to collider data analysis during the school. After 3 days of lectures and exercises, the students were grouped into teams of 3 or 4 and each team was assigned a public analysis note from ATLAS or CMS experiments. After 1.5 days of independent study, each team was able to reproduce the assigned analysis using CutLang.

1 Introduction

The utilization of computational tools for doing research in particle and accelerator physics is an essential requirement and it is often omitted in the curricula of Turkish universities. Based on the experiences of the academics working in these two fields, the “Computing Applications in Accelerator and Particle Physics” school in Turkey (Turkish: HPFBU) was conceived in 2008 as a one-intensive-week event for graduate students. It aimed to educate new generations of scientists through a very hands-on attitude, providing them with skills that are readily applicable, training them by compelling them to actually perform research tasks.

The school has evolved over the years, targeting different sections of the Turkish student pool and updating itself to keep in touch with the latest developments in the field. Its fifth iteration, focused on data analysis of the Large Hadron Collider (LHC) data, was held between 3-7 February 2020, and targeted mostly undergraduate students, along with a few embedded graduate students. To ease their task, especially for the pre-literate ones in terms of computer programming languages, and to separate the physics analysis algorithm from the computational part, it was decided to introduce the Analysis Description Language, ADL, concept.

The ADL concept uses a plain text file containing blocks with a keyword-value structure. The blocks make clear the separation of analysis components such as object definitions, variable definitions, and event selections while the keywords specify analysis concepts and operations. The syntax includes mathematical and logical operations, comparison and optimization operators, reducers, four-vector algebra and common HEP-specific functions (e.g. $\Delta \phi$, $\Delta R$, etc.). ADL files can refer to self-contained functions encapsulating variables with complex algorithms (e.g. $M_{T_2}$, aplanarity, etc.) or non-analytic variables (e.g. efficiency tables, machine learning discriminators, etc.).

Not having the necessity to write or compile code, combined with the simple, human-readable nature of ADL syntax makes it a practical construct for quickly performing phenomenological and educational analyses as the ones in the school.
2 School Concept

2.1 Student Selection

Although the school had so far accepted graduate students only, this year, to test a new idea, participants were selected amongst the undergraduate students from physics and engineering departments of various universities. About 132 applications were received and 24 undergraduate, 3 MSc and 3 PhD students were accepted after consideration of their GPAs, resumes, reference letters and their scores on a simple web-based quiz focusing on the very basics of particle physics. About one third of the students were affiliated with some engineering department. No prior computer, programming or operating system knowledge requirements were imposed during the selection process.

2.2 Logistics

The event took place in the main campus of Istanbul University, Astronomy and Spaces Sciences Department, Turkey. Istanbul University’s participation to CERN related activities started in 1977. Currently, the university takes part in the ATLAS and CMS experiments through the members of the physics department.

The student dormitories of the university were allocated to participating students (including those who normally reside nearby) to allow the discussions and peer-to-peer education to continue up to as late as possible hours. The lectures were both recorded and broadcast live on the internet [3]. The computing exercises were conducted in the allocated computer room with 15 desktop machines. Students who wanted to use their own laptop computers were requested to have downloaded a virtual computer image and installed the Virtual Box host application [4] before attending the exercise sessions.

2.3 Lecture Contents

The lectures start with the basics of elementary particles and introduce the necessary physics and mathematics background for understanding the data analysis in particle physics. The relevant computer software used in the field are presented by some hands-on training on their utilization. Recently, utilization of dedicated analysis languages instead of general purpose languages (e.g. C++ or Python) is becoming popular as it allows physicists to become faster and more efficient in designing analyses. Using such a language, CutLang, object selection, object reconstruction and event selection concepts are shown through examples. The main goal is presented as being able to analyse the Monte Carlo pseudodata and real data from the LHC experiments (obtained through the open data initiative). Towards that goal, the statistical techniques used in particle physics are also presented.

2.3.1 Introduction to Analysis

This lecture starts from the basics of elementary particles and interactions. The most relevant concepts for experimental research are presented, such as the process cross sections and branching fractions. Particle detectors are then introduced as a means of observing known particles and searching for new particles. The concept of a particle jet is discussed at this stage. For understanding the particle collision concept, the kinematics variables and hard scattering are discussed in detail. At this stage, the invariant mass idea is presented with examples from Z and W bosons, which motivates the teaching of concepts such as transverse momentum, pseudorapidity and missing energy.
2.3.2 Analysis Structure

This lecture is envisaged as a practical approach to HEP analysis concepts. It starts with an overview of quantum field theory and Feynman diagrams to later cover quantum electrodynamics, quantum chromodynamics and weak interactions. Once a collision event is defined, the main steps of data analysis are presented: 1) to define a signal and to design a search channel, 2) to define a trigger for signal events (the trigger concept is introduced at this stage.), 3) to reconstruct new objects and eventually signal events, 4) to estimate the contribution from background events (the background concept is introduced at this stage.) 5) statistical analysis. The systematic uncertainties are also introduced to properly interpret the experimental results. Details of these concepts are left to the following lectures.

2.3.3 Introduction to ROOT

This lecture starts with an introduction to Linux and terminal commands. After all, a minimum command line literacy is needed to edit a file, to look at a histogram etc. The rest of the lecture introduces ROOT, its command line interface, simple calculations using the command line, graphics focusing on histograms and functions. These concepts are enforced through exercises executed with simple macros. Second part of the lecture brings more advanced concepts like error representation, histogram operations (normalization, scaling etc) and finally reading from and writing into the root files.

2.3.4 Introduction to CutLang: summary

This lecture covers a particular analysis description language (ADL) and its runtime interpreter. During the school, CutLang [1, 5], a runtime interpreter written in C++ and based on ROOT [6] classes for Lorentz vector operations and histogramming, was used. CutLang is able to operate directly on events without the need for compilation. It uses automatically generated dictionaries and grammar rules based on Unix tools Lex and Yacc [7]. The typical output of an analysis in CutLang is a file containing surviving events and histograms which can be used for statistical analysis. The lecture covers the basic steps of data analysis using examples and exercises executed using CutLang.

2.3.5 Introduction to Probabilities and Statistics

Probabilities and Poisson statistics were introduced via real life examples such as proton decay, German rocket hits to London during WWII, etc. Exercises were built on rudimentary use of previous knowledge (like the basic use of ROOT macros). Bayesian and frequentist statistical methods were discussed to introduce hypothesis testing. False positives and negatives were discussed together with $\chi^2$ distribution for decision making within the context of particle physics analysis. Concepts like particle identification and limit setting were presented as relevant examples. Finally a simple limit setting software, TLimit from ROOT, was shown with a working example.

2.3.6 Particle Reconstruction

The first part of the lecture starts with the details of ATLAS and CMS detectors to focus on particle isolation, identification and reconstruction. Methods such as Particle Flow are discussed, as well as particle tracking and vertex identification. The performance of various LHC tracking detectors are compared and discussed. Accumulating on previous knowledge, muon and b-jet identification methods using track fitting and $\chi^2$ minimisation are presented. Finally electromagnetic calorimeters are introduced within the context of electron, photon and tau reconstruction and identification. The second part of the lecture is solely dedicated to jet reconstruction. It aims to familiarize the student with the strong interactions, hadronic collisions and jet structure. To that end, it also reviews parton distribution functions and hard scattering. Details of jet reconstruction algorithms like cone, kT and antikT are presented. Finally sub-jets, jet-charge and jet energy calibration concepts are discussed.

2.3.7 Introduction to Systematic Uncertainties

After introducing the concept of uncertainties, accuracy and precision, the main uncertainty sources are discussed: detector simulation and reconstruction, event recording and trigger and HEP theory. Details of the experimental uncertainties are elaborated: detector calibration, efficiency, and resolution. Their impact is discussed from both physics and statistics point of views. The concept of nuisance parameter is introduced for simple application in frequentist and bayesian statistics. PDFs are presented as an example to theory uncertainties. Simple exercises are performed using ROOT’s TLimit function to enforce learning.
2.3.8 BSM review

After a brief review of the shortcomings of the Standard Model, its four pillars are presented: the particle content, the forces, the symmetry group and the dimensions of space-time. Possible modifications for each of these four items are discussed as well as the possible solutions to the Standard Model’s shortcomings. For each case, searches from the LHC experiments and the limits for various models together with important model parameters are shown. Finally future accelerators and research plans are summarized to motivate the students.

2.3.9 Presentation Tips

Some tips on how to present results, how to display data, and how to prepare clear slides are discussed. Examples from ugly and cumbersome slides are given to show what not to do. Emphasis is given to make the presentation short and not go overtime and to give one single message that the listeners should take home.

2.4 Short Exercises

This is a set of short but complete exercises to get used to ROOT file manipulation, ADL file writing, CutLang utilization and histogram interpretation.

The first example is the Drell-Yan process of \( pp \rightarrow Z \rightarrow e^+e^- \). The final state is two opposite-sign “electrons”. First, events are required to have at least two electron candidates, and then requirement for opposite-sign electric charge is applied. After each selection, a \( Z \) boson candidate is reconstructed from the two highest \( p_T \) electrons, and its mass is plotted. The ADL file for this example is:

```adl
region test
  select ALL
  select Size(ELE) >= 2
  histon h1mReco, "Z candidate mass (GeV)", 100, 0, 200, {ELE_0 ELE_1}m
  select {ELE[0] ELE[1]} q==0
  histon h2mReco, "Z candidate mass (GeV)", 100, 0, 200, {ELE_0 ELE_1}m
```

As the second example, the Drell-Yan process of \( pp \rightarrow Z \rightarrow \mu^+\mu^- \) has been considered. The events are required to have at least two muons with the opposite-sign electric charge. After each selection, the \( Z \) candidate mass is plotted. The ADL file for this process is similar to the second example.

The third example is the Drell-Yan process of \( pp \rightarrow Z \rightarrow \mu^+\mu^-/e^+e^- \). Since the final state is two opposite sign same flavour muons or electrons, events are categorised into the \( ee \) and \( \mu\mu \) channels. At least two leptons are required to exist in the final state before the OS-SF criteria. The ADL file for this study can be written as:

```adl
region chele
  select Size(ELE) >= 2
  select {ELE[0] ELE[1]} q == 0
  histon h1mReco, "Z candidate mass (GeV)", 100, 0, 200, {ELE_0 ELE_1}m
region chmuo
  select Size(MUO) >= 2
  select {MUO[0] MUO[1]} q == 0
  histon h2mReco, "Z candidate mass (GeV)", 100, 0, 200, {MUO_0 MUO_1}m
```

In the fourth example, the same Drell-Yan process of the previous example has been elaborated to specify “good” leptons. The leptons are defined as the union of electrons and muons. The good leptons are defined based on the cuts on their transverse momenta (\( p_T > 25 \text{ GeV} \)) and pseudorapidity (\( |\eta| < 2.5 \)). The students are expected to use CutLang syntax to find the best lepton combination to reconstruct a \( Z \)-boson candidate and they plot its invariant mass to comment on the decay width and eventually production cross section.

As the fifth example, the \( pp \rightarrow W \rightarrow \mu\nu \) process has been studied. The neutrino in the final state is assumed to have been measured as the missing transverse energy (MET), which, combined with the charged lepton, is used to reconstruct the \( W \)-boson candidate. In this case, the Lorentz vector from MET, METLV, is defined as \( (\text{MET}, \text{MET}_x, \text{MET}_y, 0) \). Afterwards, the \( p_T \) of the muon and the \( W \) candidate mass are plotted. The ADL file for this example is:

```adl
region testW
  select ALL
  select Size(MUO) >= 1
  define Wreco : MUO_0 METLV_0
  histon h1muPt, "pT of muon (GeV)", 100, 0, 200, Pt{MUO_0}
  histon h2mReco, "W candidate mass (GeV)", 100, 0, 200, {Wreco}m
```
The sixth example is the process \( pp \rightarrow WW \rightarrow \mu\nujj \). The final state consisted of a muon, two jets and MET. Initially both leptonically and hadronically decaying W bosons are defined. The event selection starts with a requirement of at least one muon and sufficiently large MET (\( > 20 \text{ GeV} \)). The events are also required to have at least two jets. Finally, the W boson candidate masses are obtained from the highest momentum physics objects and are plotted for both leptonic and hadronic cases separately. The ADL file for this exercise is:

```adl
define WLreco : MUO,0 METLV,0
define WHreco : JET,0 JET,1
region test
  select ALL
  select Size(MUO) >= 1
  select MET > 20
  select Size(JET) >= 2
hists lmnWL, "W_L candidate mass (GeV)" , 100, 0, 200, {WLreco}m
hists l2mnWH, "W_H candidate mass (GeV)" , 100, 0, 200, {WHreco}m
```

The last short example is the process \( pp \rightarrow t\bar{t} \rightarrow WWbb \rightarrow jjjjjj \). The final state contains at least six jets. Events are required to have two hadronically reconstructed W bosons and two jets that can be identified as b-jets. Here, first the Ws are defined from jets and then the optimal reconstruction is determined using a \( \chi^2 \) algorithm based on the known mass of the W boson. Next, the top quarks are defined by pairing the reconstructed W bosons and jets. The best pairing is again found via a \( \chi^2 \) algorithm without using the b-tagging. After all definitions, at least six jets are required. The MET in the event is expected to be small (\(< 100 \text{ GeV}\)). Finally, the total \( \chi^2 \) is minimized (\( \chi^2 = 0 \)) to plot the masses of the reconstructed objects. The ADL file for this example is given below:

```adl
define WH1 : JET[-1] JET[-1]
define WH2 : JET[-3] JET[-3]
define Wchi2 : ((({WH1}m - 80.4)/2.1)^2 + ((({WH2}m - 80.4)/2.1)^2
define Top1 : WH1 JET[-2]
define Top2 : WH2 JET[-4]
define mTop1 : m(Top1)
define mTop2 : m(Top2)
define topchi2 : ((mTop1 - mTop2)/4.2)^2
algo best
  select ALL
  select Size(JET) >= 6
  select MET < 100
  select Wchi2 + topchi2 = 0
hists lmnWH1, "Hadronic W reco (GeV)" , 50, 50, 150, m(WH1)
hists lmnWH2, "Hadronic W reco (GeV)" , 50, 50, 150, m(WH2)
hists lmnTop1, "Hadronic Top reco (GeV)" , 70, 0, 700, m(Top1)
hists lmnTop2, "Hadronic Top reco (GeV)" , 70, 0, 700, m(Top2)
```

### 2.5 Team analyses

During the last two days of the event, the students were grouped into teams of 3-4 students. Each group was assigned a particular ATLAS or CMS analysis and was expected to skim through the relevant publication and then write their own ADL files. The assigned analyses are listed below:

1. “Search for new resonances in mass distributions of jet pairs using 139 fb\(^{-1}\) of pp collisions at \( \sqrt{s} = 13 \text{ TeV} \) with the ATLAS detector”, [arXiv:1910.08447](https://arxiv.org/abs/1910.08447).
2. “Search for high-mass dilepton resonances using 139 fb\(^{-1}\) of pp collision data at \( \sqrt{s} = 13 \text{ TeV} \) with the ATLAS detector”, [arXiv:1903.06248v3](https://arxiv.org/abs/1903.06248v3).
3. “Search for a heavy charged boson in events with a charged lepton and missing transverse momentum from pp collisions at \( \sqrt{s} = 13 \text{ TeV} \) with the ATLAS detector”, [arXiv:1906.05609v2](https://arxiv.org/abs/1906.05609v2).
4. “Search for resonances decay into into top-quark pairs using fully hadronic decays in pp collisions with ATLAS at \( \sqrt{s} = 13 \text{ TeV} \)”, [arXiv:1211.2202v3](https://arxiv.org/abs/1211.2202v3).
5. “Measurement of the top quark pair production cross section at 13TeV with the CMS detector”, PoS (TOP2015) 026.
6. “Search for supersymmetry in multijet events with missing transverse momentum in proton-proton collisions at 13 TeV”, arXiv:1704.07781.

7. “Search for supersymmetry in events with one lepton and multiple jets exploiting the angular correlation between the lepton and the missing transverse momentum in proton-proton collisions at $\sqrt{s} = 13$ TeV”, arXiv:1709.09814.

8. “Search for new physics in events with two soft oppositely charged leptons and missing transverse momentum in proton-proton collisions at $\sqrt{s} = 13$ TeV”, arXiv:1801.01846.

2.5.1 Generation of the Pseudo-data

Monte Carlo event samples used for performing the analyses in this school were either produced by the instructors or taken from the CERN LHC Open Data portal [8]. Signal samples used in the SUSY exercises were generated using Pythia8. Signal models, consisting of several gluino, top squark or EW gaugino pair production benchmarks were defined by SUSY Les Houches Accord files. Signal and SM background samples in dijet and dilepton resonances analyzes were also produced using Pythia8. Detector response for all signal samples was modelled with the Delphes detector simulator using the default ATLAS or CMS detector cards. Background samples including $t\bar{t}$+jets, $W$+jets, etc were directly taken from the ATLAS open data.

2.6 Student Presentations

The students, after their independent analysis training, were expected to work as a team and prepare a short presentation. It is very important that each member of the team presents some aspect of the analysis, even with 2-3 slides. Public speaking and being able to defend the analysis algorithm or results in front of one’s peers is an important part of HEP carrier.

2.6.1 An example analysis

Below is the final ADL file, for duplicating the analysis in the “Search for supersymmetry in events with one lepton and multiple jets” paper as written by group 7. It contains three sections: the first one defines new objects, the second defines new event variables, and finally the last section contains the selection algorithms for signal regions with and without $b$-jets.

```
object goodEle : ELE
    select abs({ELE}Eta) < 1.44
    select Pt(ELE) > 25

object goodMuo : MUO
    select Pt(MUO) > 25
    select abs({MUO}Eta) < 2.1

object goodLep : Union(goodEle, goodMuo)

object goodJet : JET
    select Pt(JET) > 30
    select abs({JET}Eta) < 2.4
    select dR(JET, goodLep) > 0.4

object vetoEle : ELE
    select Pt(ELE) > 10 AND Pt(ELE) < 25

object vetoMuo : MUO
    select Pt(MUO) > 10 AND Pt(MUO) < 25

object vetoLep : Union(vetoEle, vetoMuo)

object bjet : goodJet
    select bTag(goodJet) == 1

define Wlep : goodLep, METLV

define lepan : dPhi(goodLep, Wlep)

region preselection
    select ALL
    select Size(goodEle) >= 0
    select Size(goodMuo) >= 0
    select Size(goodLep) == 1
    select Size(vetoEle) >= 0
    select Size(vetoMuo) >= 0
```
select Size(vetoLep) < 1
select Size(JET) >= 5
select Size(goodJet) >= 5
select HT(goodJet) [] 500 1000
select Pt(goodJet_0) > 80
select Pt(goodJet_1) > 80
select Pt(goodLep_0) + MET > 250
select lepan > 0.5 #AND lepan < 1
histo h1jetpt, "Jet signals", 100, 0, 600, {goodJet_0}Pt
histo h1htjet, "Jet signals", 100, 300, 1000, HT(goodJet)
#histo h1ptele, "Ele signals", 20, 0, 100, Pt(goodEle)
#histo h1lepan, "Lepton Angle signals", 200, 0, 200, LT(lepan)
#histo goodJet, "Jet signals", 200, 0, 600, Lt(goodJet_0)

region nob
    preselection
    select Size(bjet) == 0

region multib
    preselection
    select Size(bjet) > 0
    select Size(goodJet) >= 6

3 Conclusions

On the whole, the school reached its goal of engaging the students in particle physics data analysis and they showed clear progress. The introductory level HEP data analysis, including the associated concepts of MC techniques and statistical methods is an appropriate topic to for junior and senior-level university students if the computer and programming technicalities can be surmounted. The utilization of the Analysis Description Language concept has helped greatly by bypassing the programming related difficulties especially for those that had never programmed before.

Finally we would like to encourage other particle physicists to engage undergraduate students with similar schools and eventually use the presented material and the ADL concept to tailor their own activities to spread the knowledge about HEP analyses, and to contact us in case there are any questions.

4 Funding and Acknowledgements

This school has been supported by Boğaziçi University Foundation (BÜVAK) Prof. Engin Ark Funds and Istanbul University rectorate. We would like to thank all the students for their enthusiasm and hard work.

References

[1] http://hpfbu.web.cern.ch/HPFU/Anasayfa.html
[2] Sekmen S. and Ünel G., 2018, “CutLang: A Particle Physics Analysis Description Language and Runtime Interpreter”, Comput. Phys. Commun., 233, 215 (2018), doi:10.1016/j.cpc.2018.06.023, arXiv:1909.10621[hep-ph].
[3] https://indico.cern.ch/event/877623/
[4] https://www.virtualbox.org/
[5] Ünel G., Sekmen S. and Toon A. M., 2019, “CutLang: a cut-based HEP analysis description language and runtime interpreter”, to appear in the refereed proceedings of ACAT 2019 to be published by IOP, arXiv:1909.10621[hep-ph] https://arxiv.org/abs/1909.10621
[6] https://root.cern/
[7] The Lex and Yacc Page, http://dinasour.compilertools.net
[8] http://opendata.cern.ch
Photos from the School

Figure 2: Snapshots from the lectures

Figure 3: A snapshot from an exercise session
Figure 4: Snapshots from long exercises

Figure 5: A snapshot from a group presentation

Figure 6: PFBU-5 Instructors
Figure 7: PFBU-5 Participants