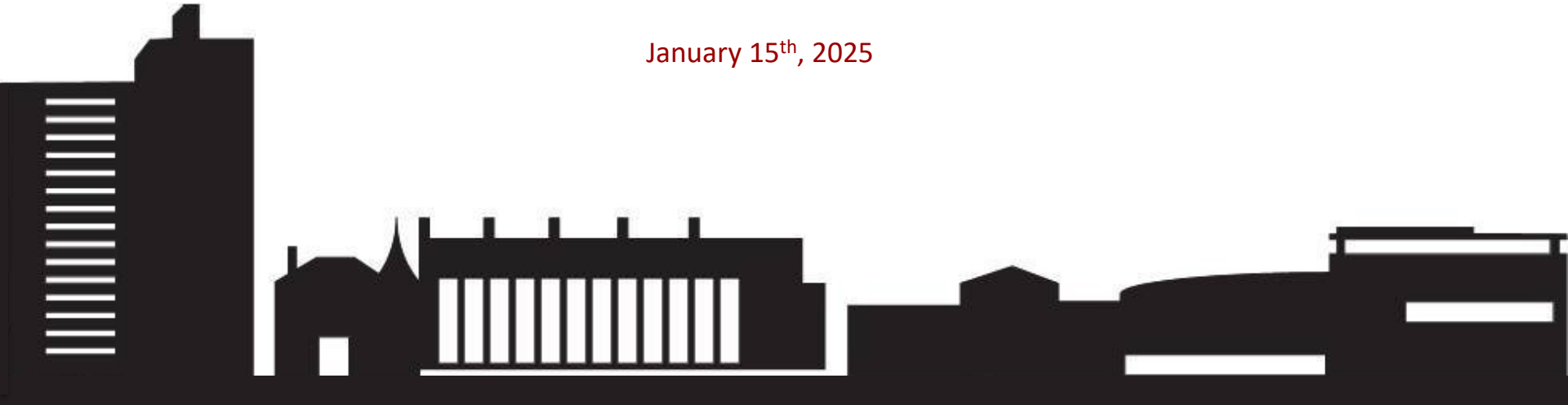
University of Massachusetts Amherst

Distributed analysis using GPU

Rafael Coelho Lopes de Sa (UMass Amherst) and Jay Sandesara (U. Wisconsin)

January 15th, 2025



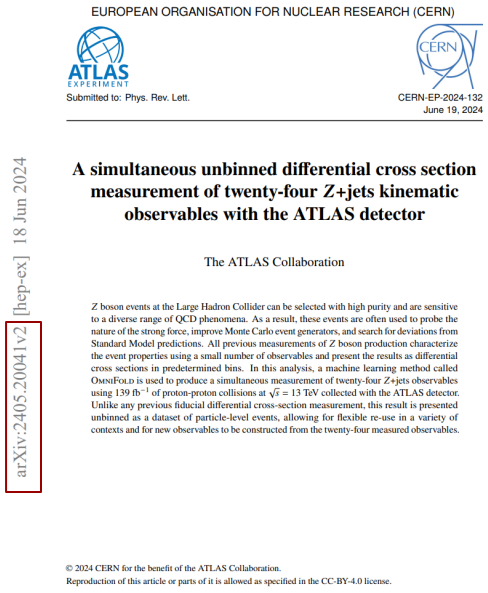
Introduction

- Two sessions in this TIM: “looking back” and “looking ahead”.
- We were planning to discuss several interesting projects at UMass in the “looking back”
  - Automating provisioning, virtualization, and site administration with OKD
  - Using shared tape systems for ATLAS ADC
  - Rucio/SENSE testing at NET2

- Unfortunately, due to personal reasons, we were not able to find the time to prepare the talk and these interesting R&D topics will be discussed somewhere else.
- So here I will discuss the “look forward”.

New analyses

- In order to look forward, let's first look (a bit) back.
- Last year, ATLAS published three papers that (I think) can change how we do data analysis.



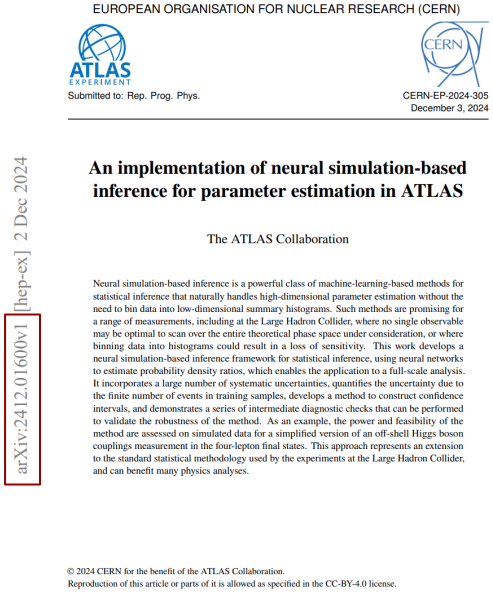
EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

Submitted to: Phys. Rev. Lett.
CERN-EP-2024-132
June 19, 2024

A simultaneous unbinned differential cross section measurement of twenty-four Z+jets kinematic observables with the ATLAS detector

The ATLAS Collaboration

Z boson events at the Large Hadron Collider can be selected with high purity and are sensitive to a diverse range of QCD phenomena. As a result, these events are often used to probe the nature of the strong force, improve Monte Carlo event generators, and search for deviations from Standard Model predictions. All previous measurements of Z boson production characterize the event properties using a small number of observables and present the results as differential cross sections in predetermined bins. In this analysis, a machine learning method called OmniFold is used to produce a simultaneous measurement of twenty-four Z+jets observables using 139 fb^{-1} of proton-proton collisions at $\sqrt{s}$ = 13 TeV collected with the ATLAS detector. Unlike any previous fiducial differential cross-section measurement, this result is presented unbinned as a dataset of particle-level events, allowing for flexible re-use in a variety of contexts and for new observables to be constructed from the twenty-four measured observables.



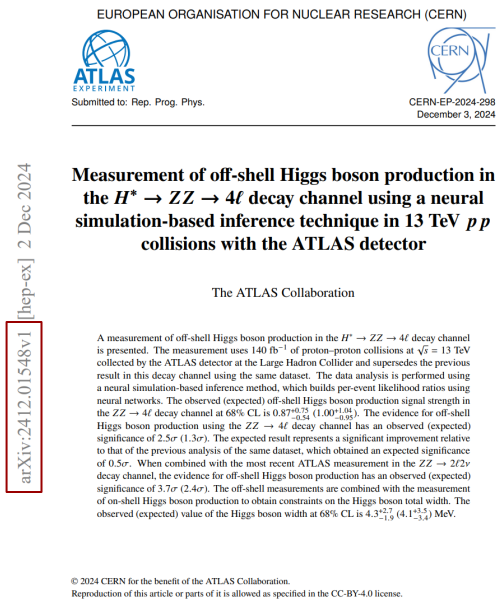
EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

Submitted to: Rep. Prog. Phys.
CERN-EP-2024-305
December 3, 2024

An implementation of neural simulation-based inference for parameter estimation in ATLAS

The ATLAS Collaboration

Neural simulation-based inference is a powerful class of machine-learning-based methods for statistical inference that naturally handles high-dimensional parameter estimation without the need to bin data into low-dimensional summary histograms. Such methods are promising for a range of measurements, including at the Large Hadron Collider, where no single observable may be optimal to scan over the entire theoretical phase space under consideration, or where binning data into histograms could result in a loss of sensitivity. This work develops a neural simulation-based inference framework for statistical inference, using neural networks to estimate probability density ratios, which enables the application to a full-scale analysis. It incorporates a large number of systematic uncertainties, quantifies the uncertainty due to the finite number of events in training samples, develops a method to construct confidence intervals, and demonstrates a series of intermediate diagnostic checks that can be performed to validate the robustness of the method. As an example, the power and feasibility of the method are assessed on simulated data for a simplified version of an off-shell Higgs boson couplings measurement in the four-lepton final states. This approach represents an extension to the standard statistical methodology used by the experiments at the Large Hadron Collider, and can benefit many physics analyses.



EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

Submitted to: Rep. Prog. Phys.
CERN-EP-2024-298
December 3, 2024

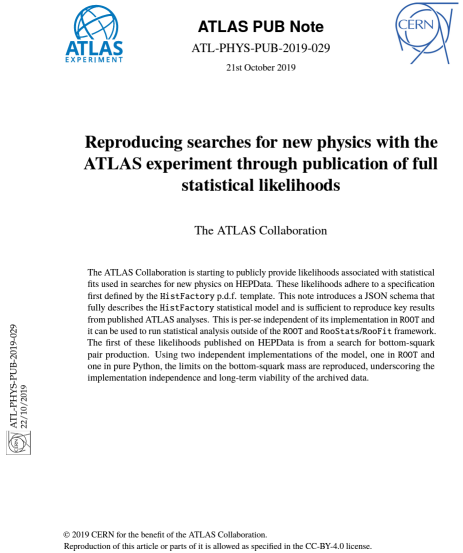
Measurement of off-shell Higgs boson production in the $H^* \to ZZ \to 4\ell$ decay channel using a neural simulation-based inference technique in 13 TeV pp collisions with the ATLAS detector

The ATLAS Collaboration

A measurement of off-shell Higgs boson production in the $H^* \to ZZ \to 4\ell$ decay channel is presented. The measurement uses 140 fb^{-1} of proton–proton collisions at $\sqrt{s}$ = 13 TeV collected by the ATLAS detector at the Large Hadron Collider and supersedes the previous result in this decay channel using the same dataset. The data analysis is performed using a neural simulation-based inference method, which builds per-event likelihood ratios using neural networks. The observed (expected) off-shell Higgs boson production signal strength in the $ZZ \to 4\ell$ decay channel at 68% CL is $0.87^{+0.75}_{-0.54}$ ($1.00^{+1.04}_{-0.95}$). The evidence for off-shell Higgs boson production using the $ZZ \to 4\ell$ decay channel has an observed (expected) significance of 2.5σ (1.3σ). The expected result represents a significant improvement relative to that of the previous analysis of the same dataset, which obtained an expected significance of 0.5σ. When combined with the most recent ATLAS measurement in the $ZZ \to 2\ell2\nu$ decay channel, the evidence for off-shell Higgs boson production has an observed (expected) significance of 3.7σ (2.4σ). The off-shell measurements are combined with the measurement of on-shell Higgs boson production to obtain constraints on the Higgs boson total width. The observed (expected) value of the Higgs boson width at 68% CL is $4.3^{+2.7}_{-1.9}$ ($4.1^{+3.5}_{-3.4}$) MeV.

But why?

- Each physics paper published by ATLAS is a (highly optimized) statistical data analysis **with a single purpose** (measure a parameter, verify if a new particle exists, …). Each analysis usually takes 2-5 years to be published
- An immense amount of effort (brains and computing) have been devoted to try to understand if the results of a paper can be re-use/re-interpreted/re-analyzed since it takes so much effort to publish every single result.
- These efforts have only been partially successful. The reality is that most analysis are so optimized (blame ML) that efforts to re-interpret the result are almost impossible. Of course, we still do generic analysis and those are the cases where re-interpretation has been successful.



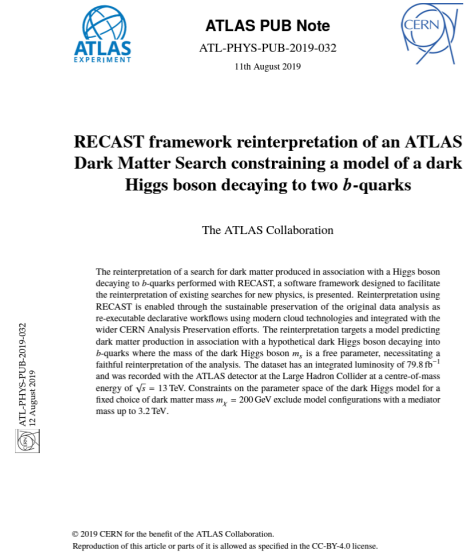
ATLAS PUB Note
ATL-PHYS-PUB-2019-029
21st October 2019

Reproducing searches for new physics with the ATLAS experiment through publication of full statistical likelihoods

The ATLAS Collaboration

The ATLAS Collaboration is starting to publicly provide likelihoods associated with statistical fits used in searches for new physics on HEPData. These likelihoods adhere to a specification first defined by the `HistFactory` p.d.f. template. This note introduces a JSON schema that fully describes the `HistFactory` statistical model and is sufficient to reproduce key results from published ATLAS analyses. This is per-se independent of its implementation in ROOT and it can be used to run statistical analysis outside of the ROOT and `RooStats`/`RooFit` framework. The first of these likelihoods published on HEPData is from a search for bottom-squark pair production. Using two independent implementations of the model, one in ROOT and one in pure Python, the limits on the bottom-squark mass are reproduced, underscoring the implementation independence and long-term viability of the archived data.

ATL-PHYS-PUB-2019-029
22/10/2019



ATLAS PUB Note
ATL-PHYS-PUB-2019-032
11th August 2019

RECAST framework reinterpretation of an ATLAS Dark Matter Search constraining a model of a dark Higgs boson decaying to two b-quarks

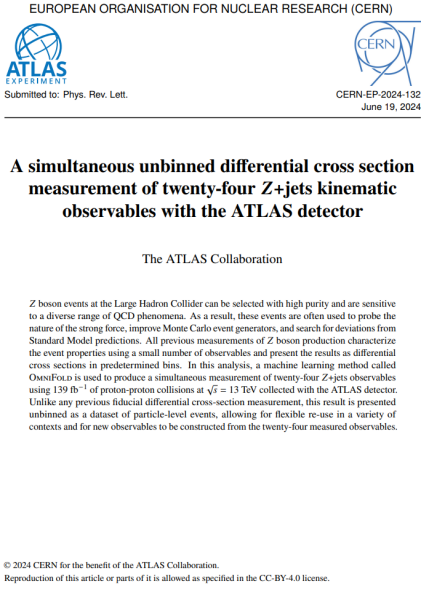
The ATLAS Collaboration

The reinterpretation of a search for dark matter produced in association with a Higgs boson decaying to b-quarks performed with RECAST, a software framework designed to facilitate the reinterpretation of existing searches for new physics, is presented. Reinterpretation using RECAST is enabled through the sustainable preservation of the original data analysis as re-executable declarative workflows using modern cloud technologies and integrated with the wider CERN Analysis Preservation efforts. The reinterpretation targets a model predicting dark matter production in association with a hypothetical dark Higgs boson decaying into b-quarks where the mass of the dark Higgs boson m_s is a free parameter, necessitating a faithful reinterpretation of the analysis. The dataset has an integrated luminosity of 79.8 fb^{-1} and was recorded with the ATLAS detector at the Large Hadron Collider at a centre-of-mass energy of $\sqrt{s} = 13$ TeV. Constraints on the parameter space of the dark Higgs model for a fixed choice of dark matter mass $m_\chi \sim 200$ GeV exclude model configurations with a mediator mass up to 3.2 TeV.

ATL-PHYS-PUB-2019-032
12 August 2019

What is so new?

- We are changing what we consider the product of our analyses:



EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

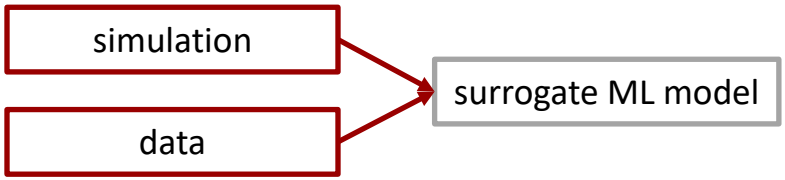
Submitted to: Phys. Rev. Lett.

CERN-EP-2024-132
June 19, 2024

A simultaneous unbinned differential cross section measurement of twenty-four Z+jets kinematic observables with the ATLAS detector

The ATLAS Collaboration

Z boson events at the Large Hadron Collider can be selected with high purity and are sensitive to a diverse range of QCD phenomena. As a result, these events are often used to probe the nature of the strong force, improve Monte Carlo event generators, and search for deviations from Standard Model predictions. All previous measurements of Z boson production characterize the event properties using a small number of observables and present the results as differential cross sections in predetermined bins. In this analysis, a machine learning method called OmniFold is used to produce a simultaneous measurement of twenty-four Z+jets observables using 139 fb^{-1} of proton-proton collisions at $\sqrt{s}$ = 13 TeV collected with the ATLAS detector. Unlike any previous fiducial differential cross-section measurement, this result is presented unbinned as a dataset of particle-level events, allowing for flexible re-use in a variety of contexts and for new observables to be constructed from the twenty-four measured observables.



The ML model is our result.

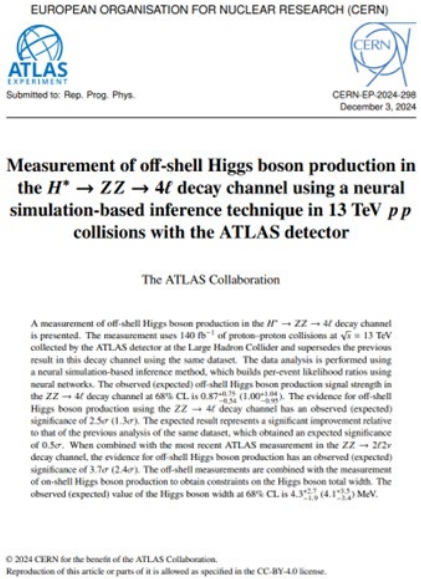
How it is used:

The ML model can be used to make unfolded comparisons using **any observable**.

The choice of observable can be made even after the paper has been published.

What is so new?

- These papers present changes in what we consider the product of an analysis.



EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

ATLAS EXPERIMENT

CERN

Submitted to: Rep. Prog. Phys.

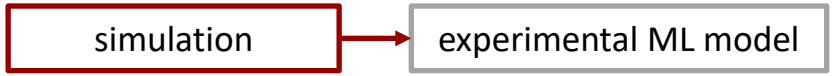
CERN-EP-2024-298
December 3, 2024

Measurement of off-shell Higgs boson production in the $H^* \to ZZ \to 4\ell$ decay channel using a neural simulation-based inference technique in 13 TeV pp collisions with the ATLAS detector

The ATLAS Collaboration

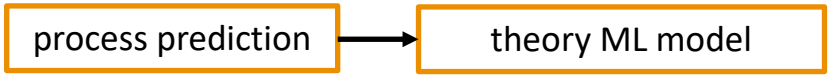
A measurement of off-shell Higgs boson production in the $H^* \to ZZ \to 4\ell$ decay channel is presented. The measurement uses 140 fb^{-1} of proton–proton collisions at $\sqrt{s} = 13$ TeV collected by the ATLAS detector at the Large Hadron Collider and supersedes the previous result in this decay channel using the same dataset. The data analysis is performed using a neural simulation-based inference method, which builds per-event likelihood ratios using neural networks. The observed (expected) off-shell Higgs boson production signal strength in the $ZZ \to 4\ell$ decay channel at 68% CL is $0.87^{+0.35}_{-0.34}$ ($1.00^{+0.04}_{-0.03}$). The evidence for off-shell Higgs boson production using the $ZZ \to 4\ell$ decay channel has an observed (expected) significance of 2.5σ (1.3σ). The expected result represents a significant improvement relative to that of the previous analysis of the same dataset, which obtained an expected significance of 0.5σ. When combined with the most recent ATLAS measurement in the $ZZ \to 2\ell2\nu$ decay channel, the evidence for off-shell Higgs boson production has an observed (expected) significance of 3.7σ (2.4σ). The off-shell measurements are combined with the measurement of on-shell Higgs boson production to obtain constraints on the Higgs boson total width. The observed (expected) value of the Higgs boson width at 68% CL is $4.3^{+2.7}_{-1.9}$ ($4.1^{+2.5}_{-2.4}$) MeV.

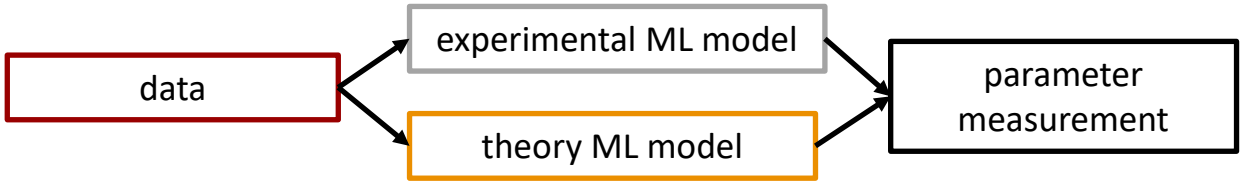
arXiv:2412.01548v1 [hep-ex] 2 Dec 2024



The ML model is our result.

How it is used:





The ML model can be used to measure **any parameter**.

The choice of parameter can be made even after the paper has been published.

Change in perspective

- These papers presented full analyses (unfolded observables and parameter estimations) and represent a first step towards the idea presented in the previous slide.
- In theory, those same analyses can be used to **optimally** perform any unfolded comparison (not only the 24 distributions presented) and any parameter estimation (not only the signal strength presented).
- These general ideas have not been fully implemeted yet. There are still practical and open questions.
- For instance:
  - How exactly do we publish the ML model?
  - How do we instruct the readers to use it correctly?
  - (in the case of parameter estimation), how do we publish our data? (**add open data discussion here**)

- But the point of this presentation is really what allowed these analyses to exist. It's more a "looking back to look forward".

What were the challenges?

- I will focus on the parameter estimation paper, since I know that one more closely.
- The ML models we built are quite large and quite complex.
- We were very fortunate to partner with the ATLAS-GCP (Google Cloud Platform) project, which provided infrastructure to perform distributed GPU analysis.
- **In many ways, it was the existence of this infrastructure that allowed us attempt to do novel type of analyses and find novel ways to explore our data with ML**

ATL-SOFT-PROC-2023-023
15/01/2024

ATLAS Data Analysis using a Parallel Workflow on Distributed Cloud-based Services with GPUs

Jay Sandesara[1,*], *Rafael* Coelho Lopes de Sa[1], *Verena* Martinez Outschoorn[1], *Fernando* Barreiro Megino[2], *Johannes* Elmsheuser[3], and *Alexei* Klimentov[3] on behalf of the ATLAS Computing Activity

[1]University of Massachusetts Amherst, Amherst, MA, USA
[2]University of Texas at Arlington, Arlington, TX, USA
[3]Brookhaven National Laboratory, Upton, NY, USA

Abstract. A new type of parallel workflow is developed for the ATLAS experiment at the Large Hadron Collider, that makes use of distributed computing combined with a cloud-based infrastructure. This has been developed for a specific type of analysis using ATLAS data, one popularly referred to as Simulation-Based Inference (SBI). The JAX library is used for the parts of the workflow to compute gradients as well as accelerate program execution using just-in-time compilation, which becomes essential in a full SBI analysis and can also offer significant speed-ups in more traditional types of analysis.

1 Introduction to Simulation-Based Inference

In high energy physics experiments like ATLAS [1] at the Large Hadron Collider, the advancement in Monte Carlo simulators over the years has enabled the accurate modelling of highly dense and complex physical interactions taking place inside particle detectors. Using such simulators, it is possible to sample an aribtrarily large number of events corresponding to a physics hypothesis.

In order to do statistical inference given an observed set of data, it is not enough to sample events using the Monte Carlo simulations but also to be able to quantify their agreement with a hypothesis of interest. The way to do this is to estimate the so-called likelihood function $L(\theta|x)$ associated with a set of observations x, where θ is the parameter corresponding to the hypothesis under study. We can think of the simulator as a computer program that takes as input a vector of parameters θ and produces a data vector $x \sim p(x|\theta)$ as output, sampled from some probability density $p(x|\theta)$ of the given hypothesis model, as shown in Figure 1. This Monte Carlo sampling is done over a series of thousands of internal states or latent variables that describe the complex interactions of particles inside the detector. This integration over thousands of variables is impractical to perform analytically making it challenging to reverse the direction from the simulated or observed set of events x to calculate $p(x|\theta)$, which is needed to build the likelihood $L(\theta|x)$ for statistical inference.

*e-mail: jay.ajitbhai.sandesara@cern.ch

ATL-SOFT-PROC-2023-020
04 September 2023

Accelerating science: the usage of commercial clouds in ATLAS Distributed Computing

Fernando Barreiro Megino[1,*], *Mikhail* Borodin[2], *Kaushik* De[1], *Johannes* Elmsheuser[3], *Alessandro* Di Girolamo[4], *Nikolai* Hartmann[5], *Lukas* Heinrich[6], *Alexei* Klimentov[3], *Mario* Lassnig[4], *FaHui* Lin[1], *Tadashi* Maeno[3], *Zachary* Marshall[7], *Gonzalo* Merino[8], *Paul* Nilsson[3], *Jay* Sandesara[9], *Cedric* Serfon[3], *David* South[10], *Harinder* Singh[11] on behalf of the ATLAS Computing Activity

[1]University of Texas at Arlington, Arlington, TX, USA
[2]University of Iowa, Iowa City, IA, USA
[3]Brookhaven National Laboratory, Upton, NY, USA
[4]CERN, Geneva, Switzerland
[5]Ludwig-Maximilians-Universitaet Muenchen, Munich, Germany
[6]Max-Planck-Institut für Physik, Munich, Germany
[7]Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[8]Port d'Informació Científica, Barcelona, Spain
[9]University of Massachusetts at Amherst, Amherst, MA, USA
[10]DESY, Hamburg, Germany
[11]California State University at Fresno, Fresno, CA, USA

Abstract. The ATLAS experiment at CERN is one of the largest scientific machines built to date and will have ever growing computing needs as the Large Hadron Collider collects an increasingly larger volume of data over the next 20 years. ATLAS is conducting R&D projects on Amazon Web Services and Google Cloud as complementary resources for distributed computing, focusing on some of the key features of commercial clouds: lightweight operation, elasticity and availability of multiple chip architectures.

The proof of concept phases have concluded with the cloud-native, vendor-agnostic integration with the experiment's data and workload management frameworks. Google Cloud has been used to evaluate elastic batch computing, ramping up ephemeral clusters of up to O(100k) cores to process tasks requiring quick turnaround. Amazon Web Services has been exploited for the successful physics validation of the Athena simulation software on ARM processors.

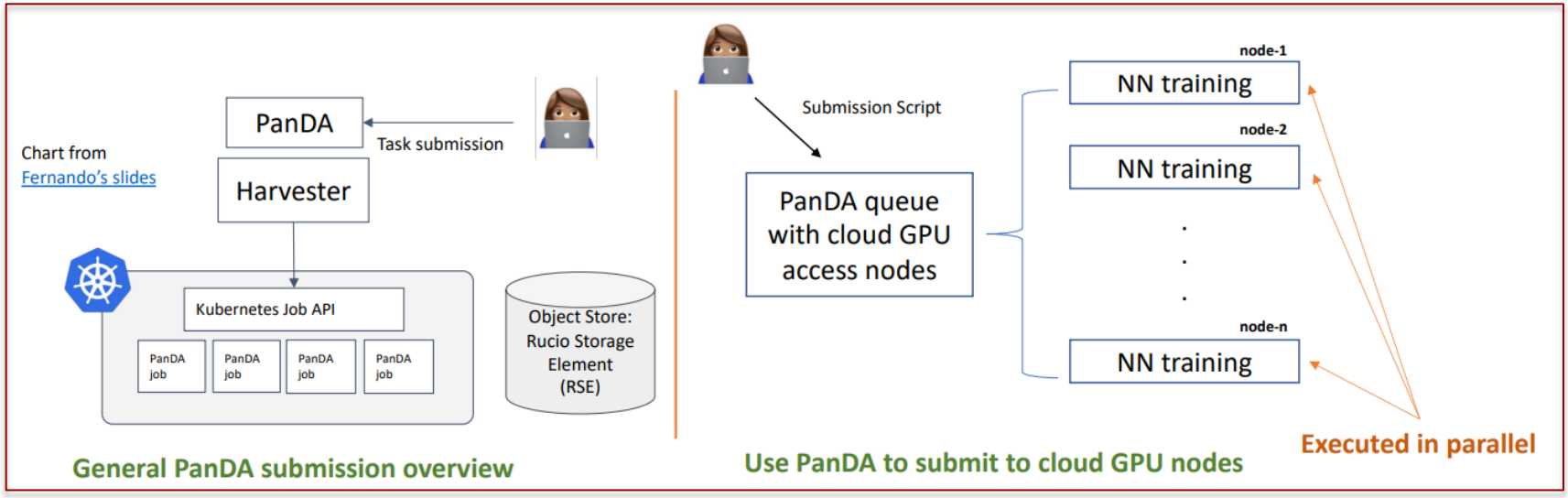
We have also set up an interactive facility for physics analysis allowing end-users to spin up private, on-demand clusters for parallel computing with up to 4 000 cores, or run GPU enabled notebooks and jobs for machine learning applications.

The success of the proof of concept phases has led to the extension of the Google Cloud project, where ATLAS will study the total cost of ownership of a production cloud site during 15 months with 10k cores on average, fully integrated with distributed grid computing resources and continue the R&D projects.

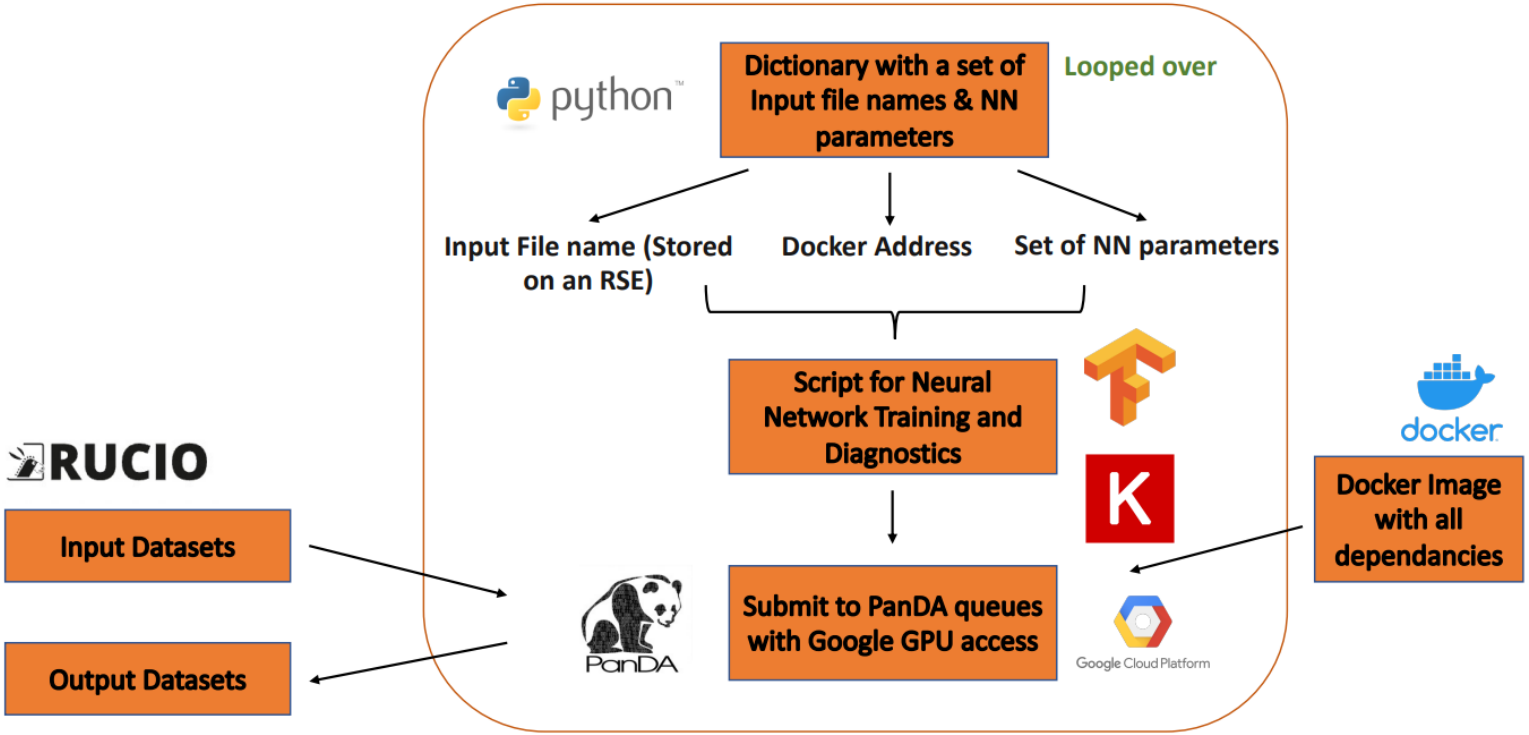
*e-mail: fernando.barreiro[at]uta.edu

Distributed analysis in GPU

- The R&D for the analysis was all done using PanDA based on the Kubernetes infrastructure presented below (charts copied from Fernando's slide during CHEP 2023)
- The infrastructure allowed for the training, optimization, and validation of very complex ML models



Workflow


python
Dictionary with a set of Input file names & NN parameters
Looped over
Input File name (Stored on an RSE)
Docker Address
Set of NN parameters
Script for Neural Network Training and Diagnostics
RUCIO
Input Datasets
Output Datasets
PanDA
Submit to PanDA queues with Google GPU access
Google Cloud Platform
docker
Docker Image with all dependancies

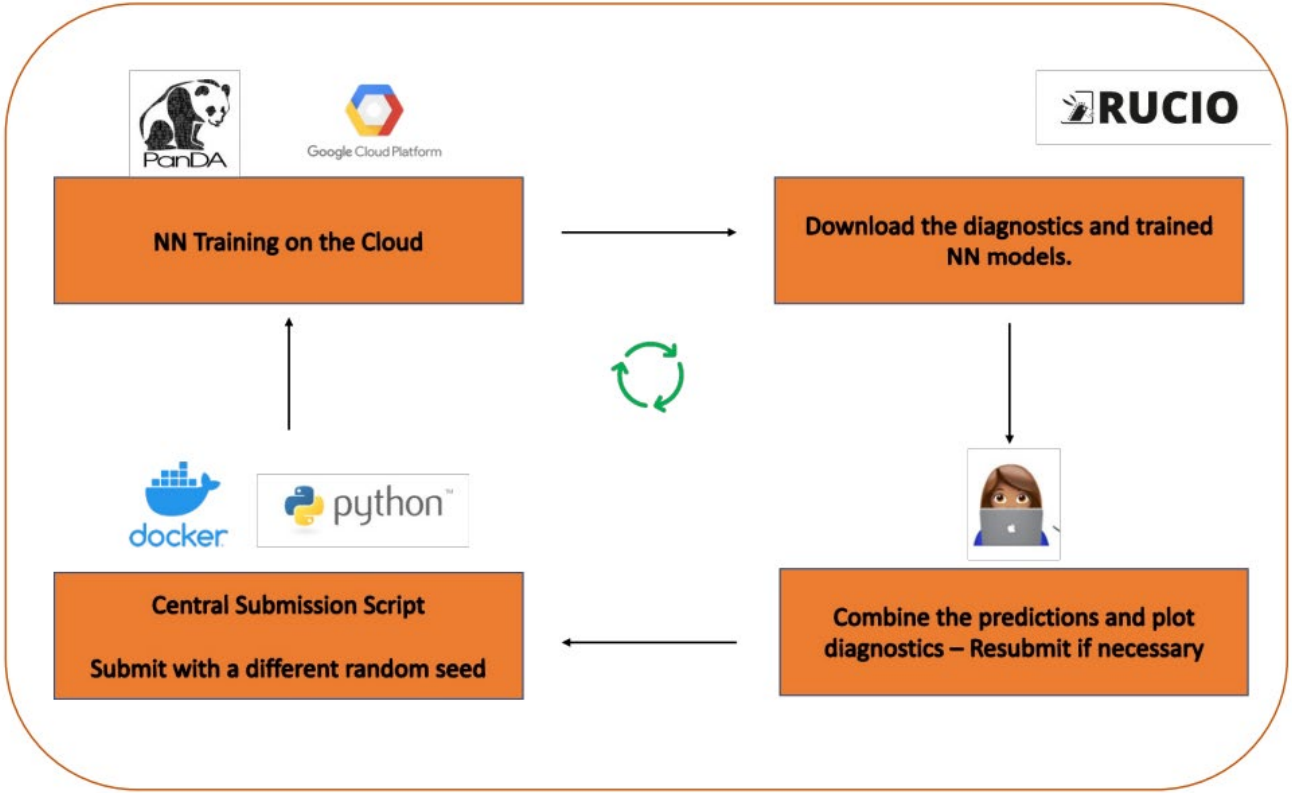
Real usage of GPU during R&D

Total of 200 nodes were made available for the analysis, each with four cores of CPU and an Nvidia T4 GPU

Number of CPU cores in use at given time

Summary of Cloud resources used for the SBI analysis R&D, in October 2022

ML model optimization/validation



Analysis facilities

Unfortunately, at some point, the ATLAS-GCP project ended. We were able to complete the R&D phase, but a lot was still needed for the analyses to be completed/reviewed/approved (still on going).

We were able to find similar infrastructure in non-ATLAS analysis facilities.

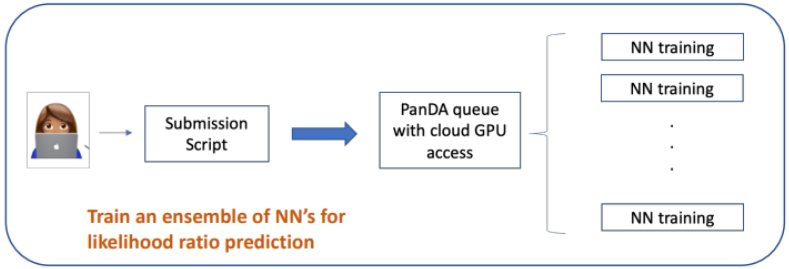
O'Donnell Data Science and Research Computing Institute

SMU

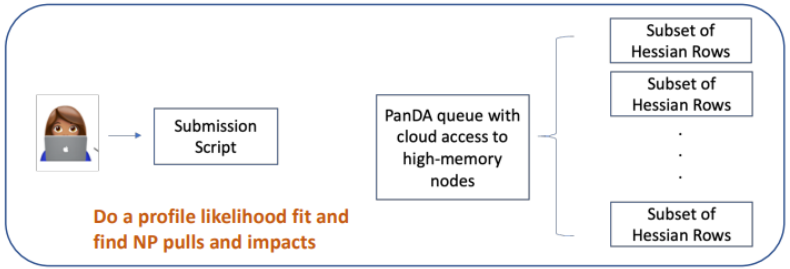
UNITY

What infrastructure exactly?

The ATLAS-GCP project and the analysis facilities in these universities provided distributed computing resources with unusual capabilities (O(500-1000) simultaneous jobs)



Distributed GPU nodes



Distributed high-memory (500+ GB) CPU nodes

Looking forward

- The take-home message is that new computing resources can enable new form of analysis.
- We haven't seen many innovation in analysis methods for a long time now and these two papers from last year show that it is possible to leverage ML in ways that we haven't yet.
- But, for that to happen ADC has to be ready to work with analyzers to provide custom computational resources for dedicated analyses.
- I think that the ATLAS GCP project was unique and very successful. We were able to communicate the resources needed for each development stage and the resources were readily available. Unfortunately, this does not exist anymore.
- The project did not have a duration compatible with a typical ATLAS analysis.
- **Can we develop long-term, flexible distributed computing infrastructure?**
- We would be able to do much more with our data.

Questions

- Let me start: “oh, but this is expensive”
  - True, it is expensive. The question here is if this is less expensive than producing hundreds of papers with short shelf-life and that will never be used or cited again.
  - I think not.
- Any other question?