A Reproducible Python Data-Analysis Pipeline for Z Boson Invariant-Mass Fitting Using ATLAS Open Data

Eldor Dustmurodov[1, 2], Temurbek Fayziev[1, 2], Gulnora Kasimova[1], Shakhlo Allakulieva[1] and Lyazzat Zhaidakbaeva[3]
[1]*Tashkent State Transport University, Temiryolchilar Str. 1, 100167 Tashkent, Uzbekistan*
[2]*Institute of Nuclear Physics, Academy of Sciences of the Republic of Uzbekistan, Khuroson Str. 1, 100214 Tashkent, Uzbekistan*
[3]*Mukhtar Auezov South Kazakhstan Research University, Tauke Khan Str. 5, 160012 Shymkent, Kazakhstan*
eldordustmurodov@mail.ru, temurfayziyev@outlook.com, gulnoratash2019@list.ru, shahloallquliyeva@gmail.com, luizca18@mail.ru

Keywords: Scientific Python, Reproducible Data Analysis, ATLAS Open Data, Jupyter Notebooks, Uproot, Zfit, Likelihood Fitting, Computational Workflow.

Abstract: This study develops a reproducible Python-based data-analysis pipeline for processing open collider data and performing likelihood-based invariant-mass fitting. Z-boson reconstruction using ATLAS Open Data is used as a scientific-computing case study. The workflow integrates Jupyter Notebooks, Uproot, NumPy, pandas, Matplotlib, and the zfit fitting framework in a modular sequence comprising ROOT-file ingestion, event-variable access, lepton-quality filtering, vectorized event selection, invariant-mass construction, visualization, and statistical model fitting. Events containing electron or muon pairs were filtered using trigger, isolation, impact-parameter, pseudorapidity, transverse-momentum, charge, and same-flavor requirements. The selected dilepton mass distributions were fitted with a Double Crystal Ball model using zfit. The fitted mass parameter was 90.459 GeV for the electron channel, 90.823 GeV for the muon channel, and 90.678 GeV for the combined dilepton sample. The principal contribution is the organization of high-energy-physics data processing as a transparent and reusable scientific Python workflow rather than a new particle-physics measurement. The case study demonstrates how open ROOT-formatted datasets can be processed and statistically analyzed without building the complete analysis around a monolithic ROOT/C++ environment. The proposed pipeline supports notebook-level reproducibility, modular replacement of processing stages, and adaptation to other open scientific datasets with event-based structures.

1 INTRODUCTION

Modern experimental science increasingly depends on reproducible computational workflows for accessing, filtering, transforming, and statistically interpreting large event-based datasets. In high-energy physics, scientific data are commonly stored in ROOT-based structures and processed through specialized software environments. The availability of open experimental datasets creates an opportunity to investigate not only physical observables but also the architecture of transparent and reusable scientific-computing pipelines.

Python-based analysis environments provide a modular alternative for selected open-data workflows. Jupyter Notebooks support executable analysis documents, while likelihood-based statistical modeling can be separated into an independent fitting stage implemented with zfit [1]. The ATLAS Open Data ecosystem provides event collections and analysis-oriented datasets for notebook-based computational studies [2]. Uproot enables direct access to ROOT-formatted files from Python [3], array and tabular processing can be performed using NumPy and pandas [4], [5], and Matplotlib can be used for visualization [6], as schematically demonstrated in Figure 1.

The transition from earlier datasets to richer event representations increases both the number of accessible variables and the diversity of simulated processes. As illustrated in Figure 2, the analyzed data environment contains an expanded variable structure, increasing the importance of controlled variable selection and modular processing.

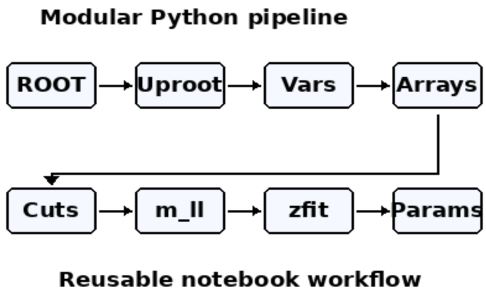

Figure 1: Modular Python workflow for open collider-data processing and invariant-mass fitting.

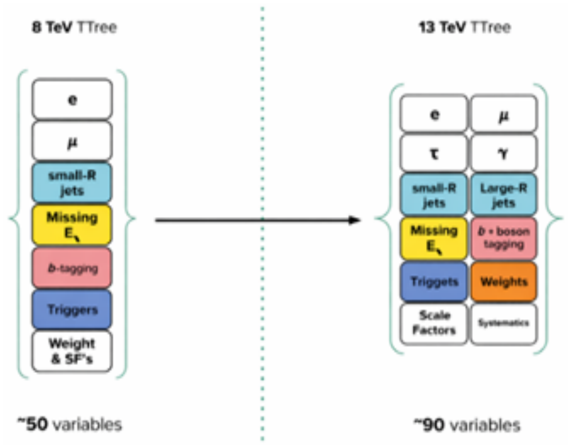

Figure 2: Expansion of the event-variable structure in the analyzed ATLAS Open Data environment.

Comparable applied engineering studies also show how computational workflows can transform measured variables into quantitative indicators and decision-relevant outputs. Computational modeling has been used to convert operational observations into quantitative engineering evidence [7], while data-driven workflows can convert heterogeneous infrastructure observations into structured decision-support priorities [8]. These examples support the broader methodological role of reproducible computational pipelines beyond a single scientific domain.

This study treats Z-boson invariant-mass reconstruction as a scientific-computing case study. The objective is to develop and describe a reproducible Python data-analysis pipeline that integrates event-data ingestion, quality filtering, physics-based selection logic, invariant-mass calculation, visualization, and likelihood fitting within one modular workflow. The contribution is therefore centered on computational organization and reproducibility rather than on claiming a new precision measurement of Standard Model parameters.

The case study focuses on dilepton final states because electron and muon channels provide a comparatively clean reconstruction path for Z-boson candidates. Table 1 summarizes the principal decay modes and motivates the selection of events containing at least two reconstructed leptons [9], [10].

Table 1: Basic decays of the Z boson (PDG, 2024).

Z decay modes	Fraction (Γ_i/Γ)	Scale factor/ Confidence level	p (MeV/c)
$e^+ e^-$	(3.3632 ± 0.0042) %	[h]	45594
$\mu^+ \mu^-$	(3.3662 ± 0.0066) %	[h]	45594
$\tau^+ \tau^-$	(3.3696 ± 0.0083) %	[h]	45559
$l^+ l^-$	(3.3658 ± 0.0023) %	[h]	45594
$l^+ l^-$- $l^+ l^-$	(3.3658 ± 0.0023) %	[b, h]	-
invisible	(4.63 ± 0.21) %	[i]	45559
hadrons	(20.000 ± 0.055) × 10^{-6}	[h]	45594
hadrons	(69.911 ± 0.056) %	[–]	-

Most Z-boson decays produce hadrons, while invisible channels contain neutrinos that are not directly reconstructed in the detector. The dilepton channels therefore form a practical open-data case study for testing a transparent Python processing and fitting workflow.

2 COMPUTATIONAL PIPELINE AND DATA ANALYSIS

2.1 Event Selection as a Rule-Based Processing Stage

Event selection was implemented as a deterministic rule-based processing stage within the Python workflow. After ROOT-file access and event-variable extraction, Boolean masks were applied sequentially to the lepton collections. The event-level filtering sequence comprised the following reconstruction-quality requirements:

- Electronic or muon trigger triggered.
- Leptons must have good quality measurements using data from the muon spectrometer and internal detector.

- Leptons must be isolated, i.e. in a cone around the lepton, the transverse energy and momentum of the remaining particles should not exceed 10% of the transverse energy and momentum of the lepton.
- Leptons are associated with the primary vertex: the significance of the impact parameter in the transverse plane $d_0/\sigma(d_0)$ is less than 3 for muons and 5 for electrons, in the longitudinal direction $|z_0 \sin(\theta)| < 0.5$ mm, where d_0 and z_0 are the transverse and longitudinal impact parameters.
- Leptons must be in the detector sensitivity region in terms of pseudorapidity $|\eta| < 2.47$. Electrons that fall into the gap between the sections of the calorimeter are also discarded ($1.37 < |\eta| < 1.52$).

The physics-based Z-candidate filtering sequence comprised the following requirements:

- Only two good leptons with $pT > 25$ GeV .
- Leptons have opposite charges .
- Leptons have the same flavor.
- The proximity of the invariant mass of leptons to the mass of the Z boson $|mll - mZ| < 20$ GeV.

The filtering logic was implemented through sequential array-level masks. Each mask reduced the candidate event set while preserving the event variables required by subsequent stages. This separation between data access, quality filtering, dilepton selection, and fitting reduces coupling between analysis components and allows individual selection conditions to be modified without rewriting the complete workflow. Similar selection and simulation practices are used in collider analyses and Monte Carlo-based comparisons [11] - [15].

2.2 Likelihood-Based Fitting with zfit

The final processing stage constructed the dilepton invariant-mass distribution from the selected events and passed the resulting numerical array to the statistical fitting component. Figure 3 shows the invariant-mass distributions on linear and logarithmic scales. A pronounced peak is visible in the Z-boson mass region, while the logarithmic representation also reveals the contribution of background components.

For statistical modeling, the selected invariant-mass array was processed using the zfit framework [1]. The fitting stage was kept separate from data ingestion and event filtering so that the probability model could be modified independently of the preceding processing logic. A Double Crystal Ball function was used to describe the central peak and asymmetric tails. The model contains a central location parameter, a width parameter, and left- and right-tail parameters. Parameter estimation was performed through likelihood-based minimization.

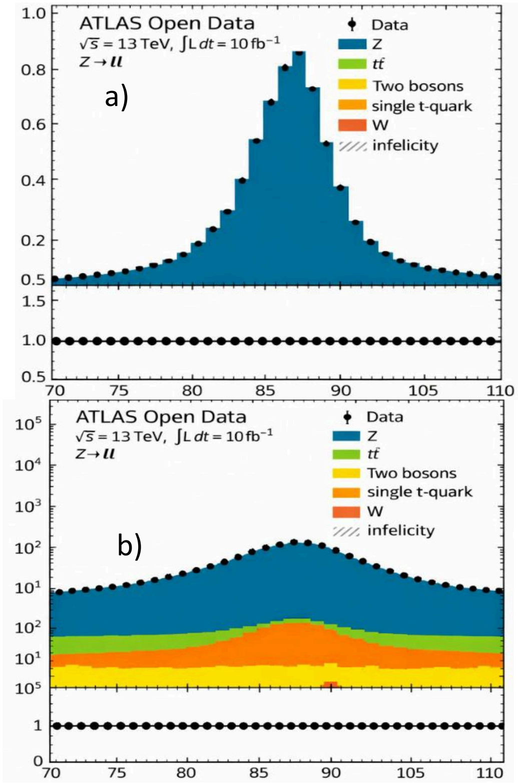

Figure 3: Distribution of the invariant mass of two leptons on a linear (a) and logarithmic (b) scale.

Table 2: Parameters for approximation of the Z boson mass.

Z→ee		Z→μμ		Z→ll	
M_Z	90.4589±0.0029	M_Z	90.8230±0.0020	M_Z	90.6779±0.0017
σ_Z	2.685±0.004	σ_Z	2.5180±0.0029	σ_Z	2.5870±0.0022
Γ_Z	6.323±0.009	Γ_Z	5.929±0.007	Γ_Z	6.091±0.006
α_L	1.035±0.004	α_L	1.272±0.003	α_L	1.1390±0.0023
n_L	3.86±0.03	n_L	2.035±0.011	n_L	2.727±0.013
α_R	1.560±0.005	α_R	1.417±0.004	α_R	1.4770±0.0027
n_R	1.956±0.015	n_R	2.456±0.017	n_R	2.229±0.013

The resulting fits are shown in Figure 4. The points represent selected open-data events, and the solid curves represent the fitted probability model applied consistently across the electron, muon, and combined dilepton channels.

The model parameters after minimization are summarized in Table 2. From a computational perspective, the same fitting component was applied to three data views: the electron channel, the muon channel, and the combined dilepton sample. Only the input array changed; the fitting interface and model structure remained unchanged. This illustrates the reusable character of the pipeline and separates channel-specific data selection from the common statistical-modeling stage.

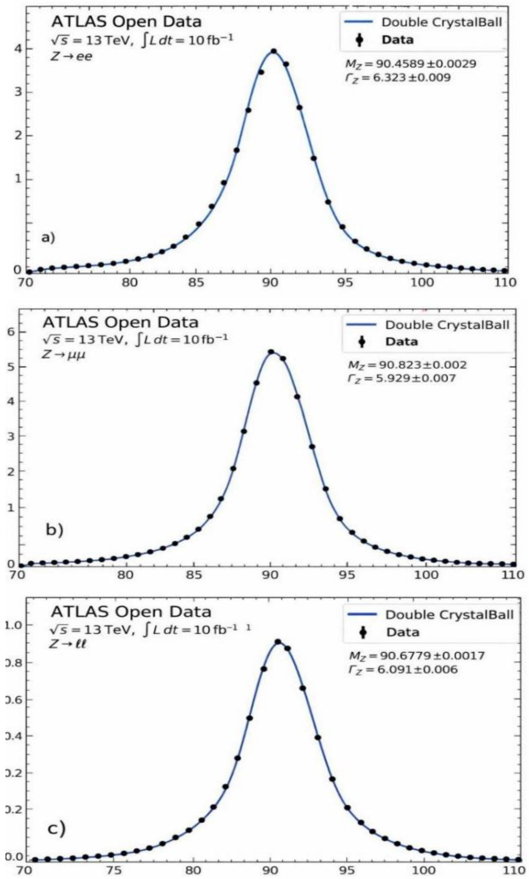

Figure 4: Zfit-based invariant-mass fits for the electron channel (a), muon channel (b), and combined dilepton sample (c).

The muon channel produced the fitted mass value closest to the reference Z-boson mass. This difference is consistent with the detector-level challenge of electron-energy reconstruction in open datasets, where not all calibration and correction variables used in full collaboration analyses are available. Consequently, the fitted values should be interpreted as outputs of the educational open-data workflow rather than as new precision measurements.

The combined dilepton fit returned a mass parameter of 90.678 GeV. All quoted errors in this study are statistical outputs of the fitting procedure, and systematic detector effects are not evaluated within the simplified open-data workflow.

For comparison, Figure 5 shows the corresponding invariant-mass distribution obtained from an earlier open-data example at lower collision energy.

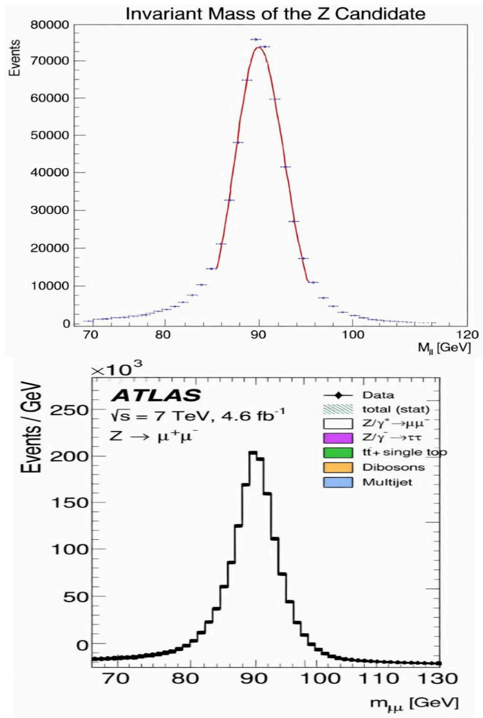

Figure 5: Invariant-mass distribution of two muons from open ATLAS data at lower collision energy.

3 CONCLUSIONS

This study developed a modular Python data-analysis pipeline for processing ROOT-formatted open collider data and performing likelihood-based invariant-mass fitting. Z-boson reconstruction from ATLAS Open Data was used as a case study to demonstrate the complete workflow from event-data ingestion to fitted statistical parameters.

The computational pipeline separated ROOT-file access, event-variable selection, quality filtering, dilepton candidate construction, invariant-mass calculation, visualization, and zfit-based statistical modeling into distinct processing stages. The fitted mass parameter was 90.459 GeV for the electron channel, 90.823 GeV for the muon channel, and 90.678 GeV for the combined sample. These values are reported as outputs of the implemented open-data workflow and are not presented as new precision measurements of Standard Model parameters.

The principal contribution is the reproducible organization of the analysis within the scientific Python ecosystem. The same fitting interface was reused for multiple lepton channels, while event-selection conditions remained explicit and independently modifiable. This modular structure improves transparency and facilitates adaptation of the workflow to other event-based open datasets.

The current study does not provide a formal runtime or memory benchmark against ROOT/C++ and therefore does not claim superior computational performance. Future work should report dataset size, processed event count, wall-clock runtime, peak memory consumption, and input/output throughput for identical hardware configurations. A controlled comparison of Uproot-based Python processing with ROOT RDataFrame would provide a quantitative basis for evaluating scalability.

REFERENCES

[1] J. Eschle, A. Puig Navarro, R. S. Coutinho, and N. Serra, "zfit: scalable pythonic fitting," SoftwareX, vol. 11, Art. no. 100508, 2020, [Online]. Available: https://doi.org/10.1016/j.softx.2020.100508.

[2] ATLAS Collaboration, "ATLAS Open Data: 13 TeV documentation and tutorials," CERN Open Data Portal, 2024.

[3] J. Pivarski et al., "Uproot: reading and writing ROOT files in Python," Scikit-HEP project, 2024.

[4] C. R. Harris et al., "Array programming with NumPy," Nature, vol. 585, pp. 357-362, 2020, [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2.

[5] W. McKinney, "Data structures for statistical computing in Python," in Proc. 9th Python in Science Conference, pp. 56-61, 2010, [Online]. Available: https://doi.org/10.25080/Majora-92bf1922-00a.

[6] J. D. Hunter, "Matplotlib: A 2D graphics environment," Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007, [Online]. Available: https://doi.org/10.1109/MCSE.2007.55.

[7] M. M. Talipov, "Computational modeling and analysis of mechanical power consumption in train assemblers' work," Proceedings of International Conference on Applied Innovation in IT, vol. 13, no. 2, pp. 419-426, 2025, [Online]. Available: https://doi.org/10.25673/120513.

[8] R. Salmorbekova and M. Talipov, "Digital risk matrix and safety management workflow for airport infrastructure in developing countries: a data-driven prioritization approach," Vibroengineering Procedia, vol. 62, pp. 653-661, Jun. 2026, [Online]. Available: https://doi.org/10.21595/vp.2026.26121.

[9] CMS Collaboration, "Measurement of the rapidity and transverse momentum distributions of Z bosons in pp collisions at sqrt(s) = 7 TeV," Physical Review D, vol. 85, Art. no. 032002, 2012, [Online]. Available: https://doi.org/10.1103/PhysRevD.85.032002.

[10] LHCb Collaboration, "Measurement of the cross-section for Z -> e+e- production in pp collisions at sqrt(s) = 7 TeV," Journal of High Energy Physics, vol. 2013, Art. no. 106, 2013, [Online]. Available: https://doi.org/10.1007/JHEP02(2013)106.

[11] T. Sjostrand, S. Mrenna, and P. Z. Skands, "PYTHIA 6.4 physics and manual," Journal of High Energy Physics, vol. 2006, Art. no. 026, 2006, [Online]. Available: https://doi.org/10.1088/1126-6708/2006/05/026.

[12] CMS Collaboration, "Measurements of inclusive W and Z cross sections in pp collisions at sqrt(s) = 7 TeV," Journal of High Energy Physics, vol. 2011, Art. no. 080, 2011, [Online]. Available: https://doi.org/10.1007/JHEP01(2011)080.

[13] CMS Collaboration, "Angular coefficients of Z bosons produced in pp collisions at 8 TeV and decaying to mu+mu- as a function of transverse momentum and rapidity," Physics Letters B, vol. 750, pp. 154-175, 2015, [Online]. Available: https://doi.org/10.1016/j.physletb.2015.08.061.

[14] ATLAS Collaboration, "Measurement of the Z/gamma* boson transverse momentum distribution in pp collisions at sqrt(s) = 7 TeV with the ATLAS detector," Journal of High Energy Physics, vol. 2014, Art. no. 145, 2014, [Online]. Available: https://doi.org/10.1007/JHEP09(2014)145.

[15] ATLAS Collaboration, "Precision measurement and interpretation of inclusive W+, W- and Z/gamma* production cross sections with the ATLAS detector," European Physical Journal C, vol. 77, Art. no. 367, 2017, [Online]. Available: https://doi.org/10.1140/epjc/s10052-017-4911-9.