ANALYSIS FOR HEP DATA/DATASETS DATA COMPRESSION

EHIZOJIE ALIGBEH

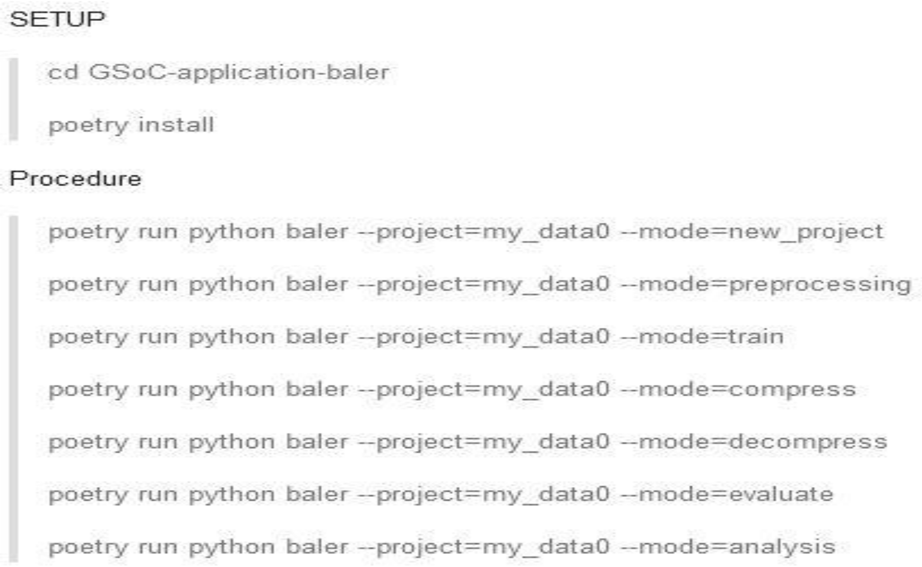
GSoC-application-baler

INTRODUCTION

In this presentation, I will share my report on applying Baler to perform a simple analysis before and after compression on a selected dataset (589F075F-48D6-E511-B46F-001E67DBE79B.root).

The goal is to present the effectiveness and reproducibility in the utilization of the baler tool across a wide variety of data

The Dataset here is a root data (root://eospublic.cern.ch//eos/opendata/cms/mc/RunIIFall15MiniAODv2/ZprimeToTT_M-3000_W-30_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/MINIAODSIM/PU25nsData2015v1_76X_mcRun2_asymptotic_v12-v1/10000/589F075F-48D6-E511-B46F-001E67DBE79B.root)

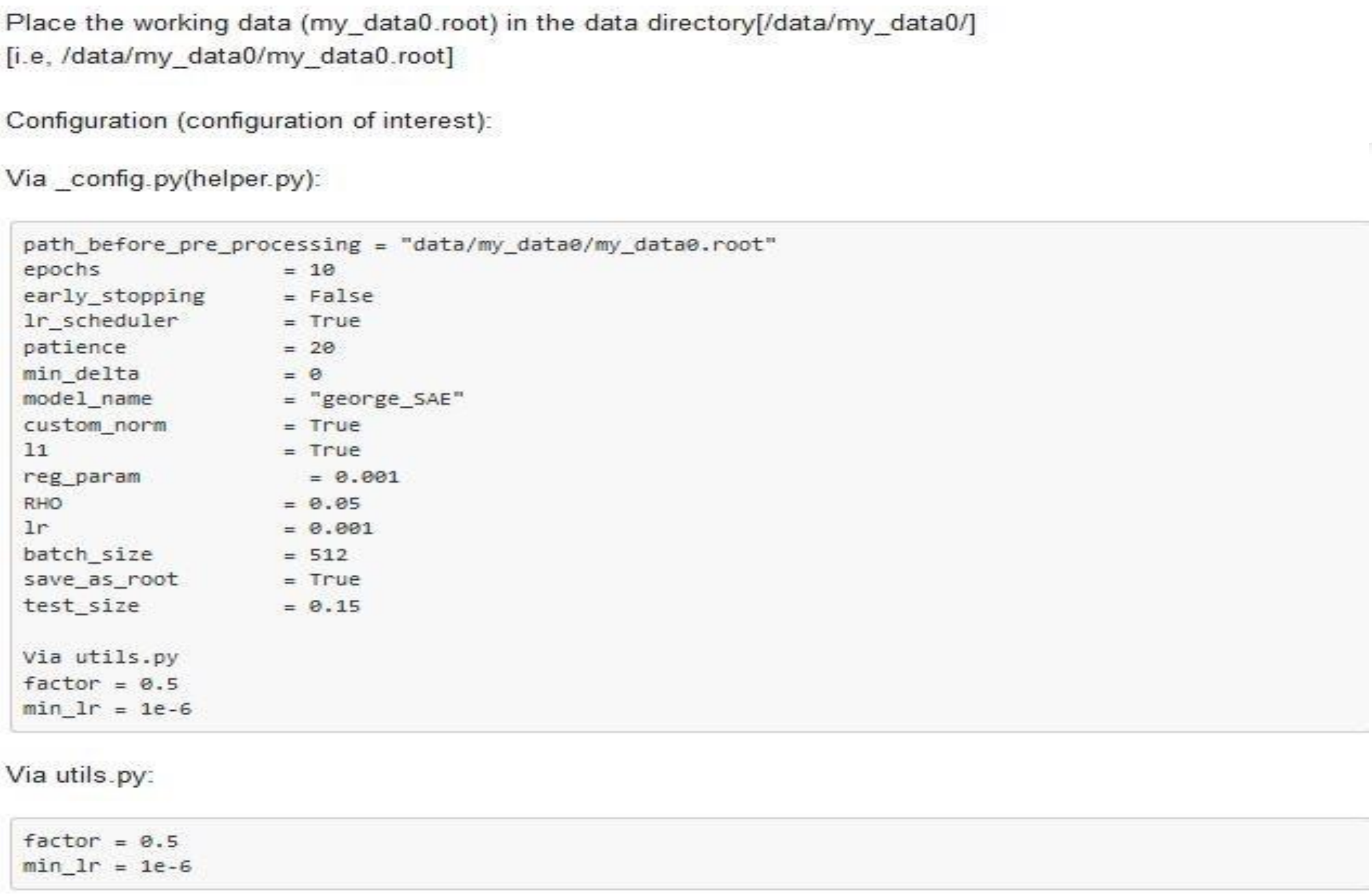


SETUP

```
cd GSoC-application-baler

poetry install
```

Procedure

```
poetry run python baler --project=my_data0 --mode=new_project

poetry run python baler --project=my_data0 --mode=preprocessing

poetry run python baler --project=my_data0 --mode=train

poetry run python baler --project=my_data0 --mode=compress

poetry run python baler --project=my_data0 --mode=decompress

poetry run python baler --project=my_data0 --mode=evaluate

poetry run python baler --project=my_data0 --mode=analysis
```



Place the working data (my_data0.root) in the data directory[/data/my_data0/]
[i.e, /data/my_data0/my_data0.root]

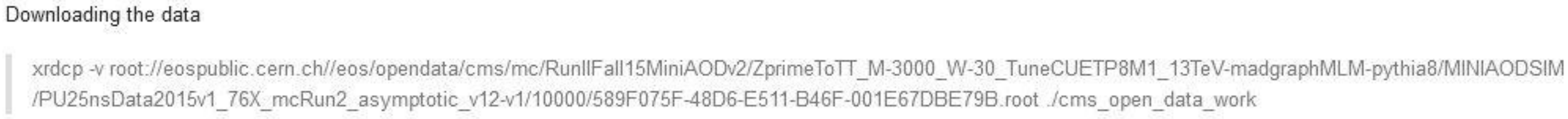
Configuration (configuration of interest):

Via _config.py(helper.py):

```
path_before_pre_processing = "data/my_data0/my_data0.root"
epochs                     = 10
early_stopping             = False
lr_scheduler               = True
patience                   = 20
min_delta                  = 0
model_name                 = "george_SAE"
custom_norm                = True
l1                         = True
reg_param                    = 0.001
RHO                        = 0.05
lr                         = 0.001
batch_size                 = 512
save_as_root               = True
test_size                  = 0.15

via utils.py
factor = 0.5
min_lr = 1e-6
```

Via utils.py:

```
factor = 0.5
min_lr = 1e-6
```



Downloading the data

```
xrdcp -v root://eospublic.cern.ch//eos/opendata/cms/mc/RunIIFall15MiniAODv2/ZprimeToTT_M-3000_W-30_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/MINIAODSIM
/PU25nsData2015v1_76X_mcRun2_asymptotic_v12-v1/10000/589F075F-48D6-E511-B46F-001E67DBE79B.root ./cms_open_data_work
```

DATASET

DATASET:

[File1] Simulated dataset ZprimeToTT_M-3000_W-30_TuneCUETP8M1_13TeV-madgraphMLM-pythia8 in MINIAODSIM format for 2015 collision data (`589F075F-48D6-E511-B46F-001E67DBE79B.root`)

Description

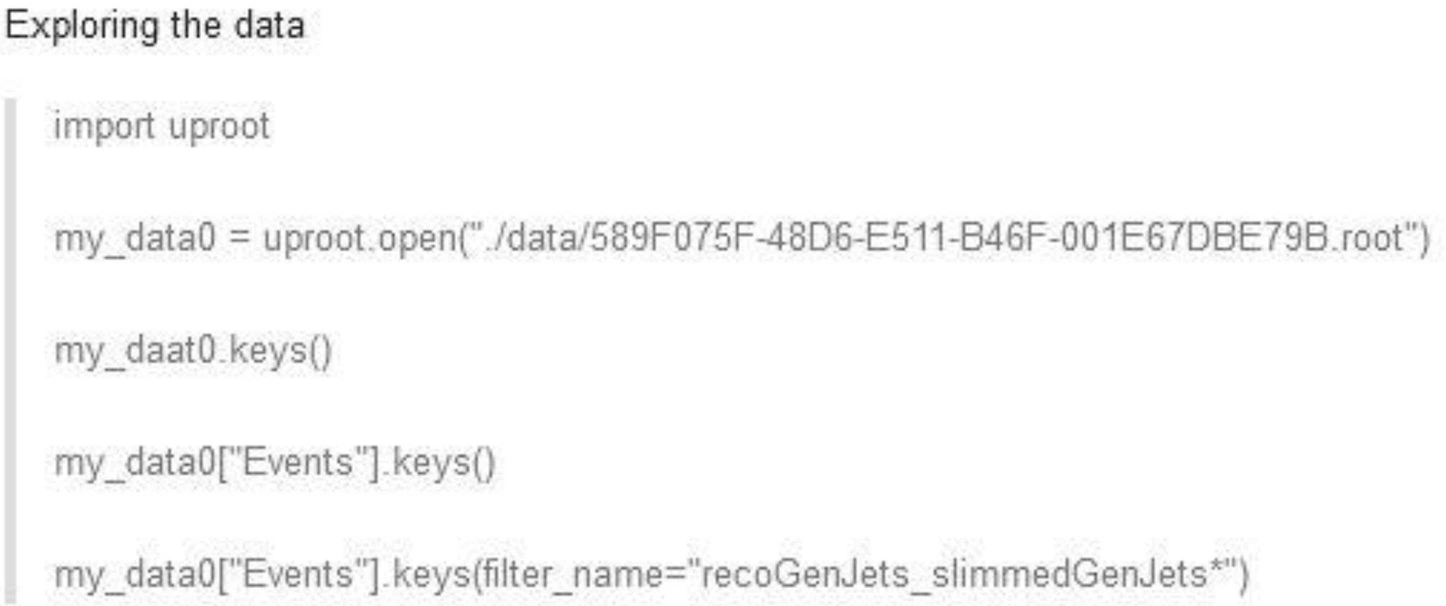
Simulated dataset ZprimeToTT_M-3000_W-30_TuneCUETP8M1_13TeV-madgraphMLM-pythia8 in MINIAODSIM format for 2015 collision data.

See the description of the simulated dataset names in: About CMS simulated dataset names.

These simulated datasets correspond to the collision data collected by the CMS experiment in 2015

Download the data

```
xrdcp -v root://eospublic.cern.ch//eos/opendata/cms/mc/RunIIFall15MiniAODv2/ZprimeToTT_M-3000_W-30_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/MINIAODSIM/PU25nsData2015v1_76X_mcRun2_asymptotic_v12-v1/10000/589F075F-48D6-E511-B46F-001E67DBE79B.root ./cms_open_data_work
```



Exploring the data

```
import uproot

my_data0 = uproot.open("./data/589F075F-48D6-E511-B46F-001E67DBE79B.root")

my_daat0.keys()

my_data0["Events"].keys()

my_data0["Events"].keys(filter_name="recoGenJets_slimmedGenJets*")
```

ANALYSIS

To analyse the data and performance, each variable of the distribution was plotted (besides few variables which were during the ore-processing)(view analysis [/my_data0_analysis.py])

After training the available data (from the 2015 collision data), response and correlation plots were created to analyse performance

These response plots are used to analyse the properties of the reconstructed data. These plots are of relative difference or residuals between input and output data.

This is well suited for analysing the effect of compression on the particles features(mass)

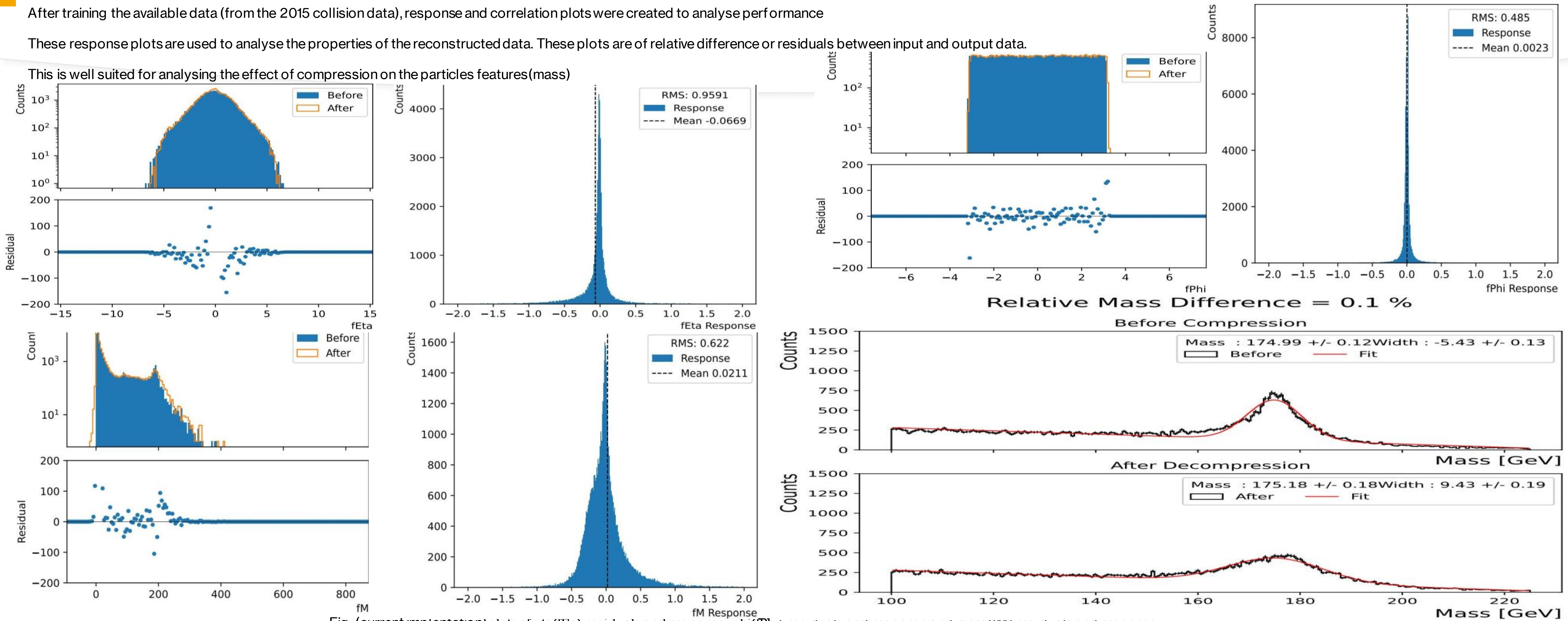


Fig: (current implentation)plots of eta(fEa) residuals and response, phi(fPhi) residuals and response and mass(fM) residuals and response

OBSERVATION

The response plots after compression are generated. It is observed(as could be seen in the evaluation output) that the results from the compression are the similar to the input data. This shows good performance across the features of the particles.

It is observed that the variations are small in pt(fPt), eta(fEta), phi and mass.

POSSIBLE IMPROVEMENT, HOW IT WORKS WITH BALER, IMPACT IT COULD HAVE ON SOCIETY

POSSIBLE IMPROVEMENT ON VARIETY OF DATASETS (USING BALER)

- Utilization of Adversarial Autoencoders and effective Variable Autoencoders.
- Extensive Normalization methods for compression techniques (Some distributions with large ranges and high distribution peaks, can benefit from undergoing a logarithm instead of just a division by the standard deviation).
- Compression and De-compression times could be further improved.

IMPLEMENTATION[OF ANALYSIS+DATA] WITH BALER

To work with this on Baler;

Exploring the data available[to be worked on] (using the Uproot library) sets the ground for proper implementation

(After performing modification to configurations for pre-processing, training and analysis)

Ensuring the Setup has been completed appropriately, I had the default_custom_analysis function allow for use of Baler on new datasets (besides the defualt example (example.root)

Same step was taken on the pre_processing and config scripts.

Dropping of fields of the input data to the Baler compressor (if necessary).

Refactoring of code for adequate analysis to be carried-out successfully

Updating Baler compressor utilization and functionalities according to new project and implementation

IMPACT ON SOCIETY

Data compression techniques can reduce the size of data drastically while giving a sufficiently similar representation of the uncompressed data. For high-energy physics, using new data-compression tools such Baler would allow for further storage savings without waiting for major technological advancements, allowing for new scientific discoveries to be made earlier by increasing the amount of data that can be recorded.

Finally, these reports and analyses show that reconstruction and compression techniques are useful for variety of applications such as TLAs which could be helpful to CERN LHC.

References:

https://github.com/baler-compressor/GSoC-application-baler, https://github.com/ehizojie1/GSoC-application-baler, [-e]