LUMIN

A DATA SCIENCE AND DEEP LEARNING ECOSYSTEM FOR HIGH-ENERGY PHYSICS

Giles Strong

INFN-ML KnowledgeBase Use Cases, Online - 27/07/2020

giles.strong@outlook.com

twitter.com/Giles_C_Strong

Amva4newphysics.wordpress.com

github.com/GilesStrong

OVERVIEW

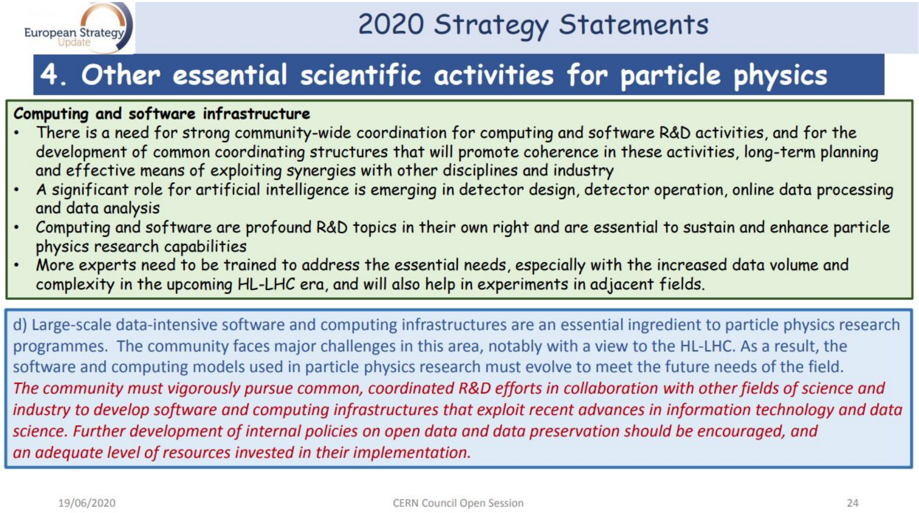
1. Motivation
2. LUMIN overview
3. Project examples
4. Summary

MOTIVATION

Machine learning in high-energy physics

MACHINE LEARNING IN HEP

- Many analyses and experiment software now aim to benefit from using machine learning approaches; often necessary in order to achieve competitive performance
- ML is now an integral part of HEP, and well recognised as such:
  - Establishment of dedicated forums & groups (IML, ATLAS & CMS ML groups)
  - Identified in 2020 update of the European Strategy for Particle Physics as essential R&D
- But! Hardware and timing for model training can be a limitation for analysis-level researchers



2020 Strategy Statements

4. Other essential scientific activities for particle physics

Computing and software infrastructure

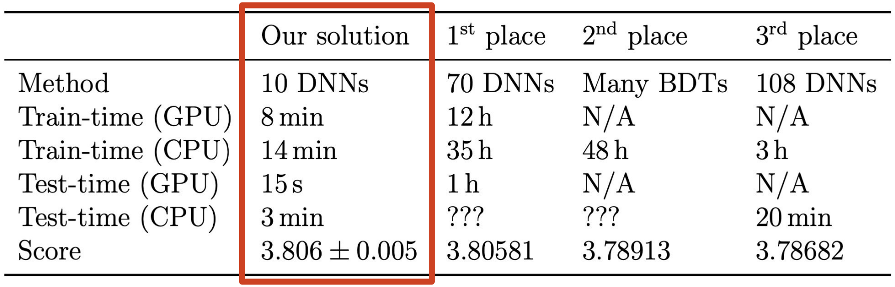
- There is a need for strong community-wide coordination for computing and software R&D activities, and for the development of common coordinating structures that will promote coherence in these activities, long-term planning and effective means of exploiting synergies with other disciplines and industry
- A significant role for artificial intelligence is emerging in detector design, detector operation, online data processing and data analysis
- Computing and software are profound R&D topics in their own right and are essential to sustain and enhance particle physics research capabilities
- More experts need to be trained to address the essential needs, especially with the increased data volume and complexity in the upcoming HL-LHC era, and will also help in experiments in adjacent fields.

d) Large-scale data-intensive software and computing infrastructures are an essential ingredient to particle physics research programmes. The community faces major challenges in this area, notably with a view to the HL-LHC. As a result, the software and computing models used in particle physics research must evolve to meet the future needs of the field. *The community must vigorously pursue common, coordinated R&D efforts in collaboration with other fields of science and industry to develop software and computing infrastructures that exploit recent advances in information technology and data science. Further development of internal policies on open data and data preservation should be encouraged, and an adequate level of resources invested in their implementation.*

19/06/2020 CERN Council Open Session 24

MODERN DEEP-LEARNING TECHNIQUES

- Strong, 2020 studied the impact of new DNNs techniques on performance and timing using benchmark HEP dataset (HiggsML)
  - HEP-specific data augmentation
  - 1cycle learning-rate scheduling
  - New architecture, activation function, etc.
  - Full details in paper
- Solution matched top performance, but trained in 14 minutes on a laptop CPU
  - 86% effective speedup over 1st-place GPU (accounting for hardware improvements)



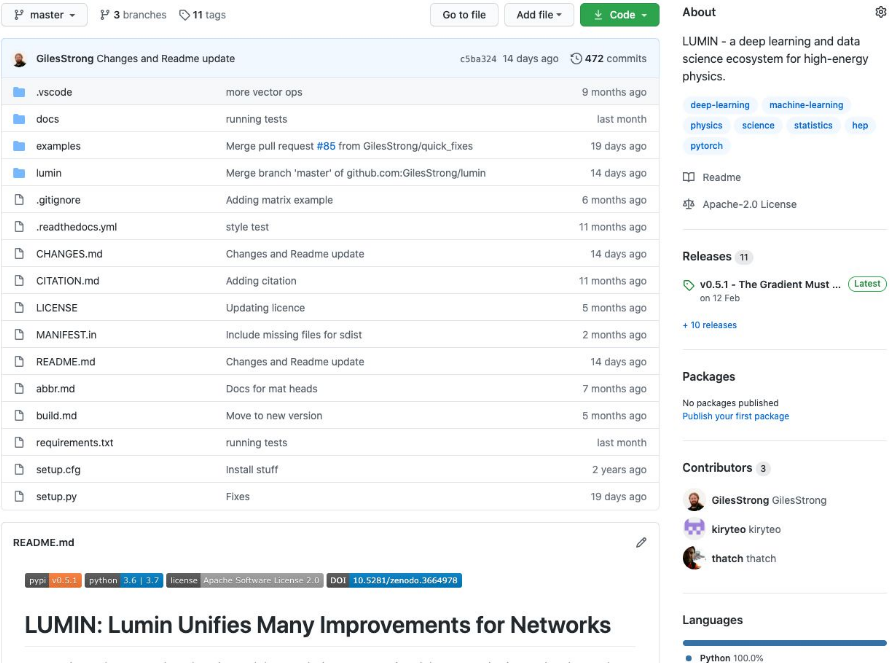
	Our solution	1st place	2nd place	3rd place
Method	10 DNNs	70 DNNs	Many BDTs	108 DNNs
Train-time (GPU)	8 min	12 h	N/A	N/A
Train-time (CPU)	14 min	35 h	48 h	3 h
Test-time (GPU)	15 s	1 h	N/A	N/A
Test-time (CPU)	3 min	???	???	20 min
Score	3.806 ± 0.005	3.80581	3.78913	3.78682

LUMIN

Lumin Unifies Many Improvements for Networks

LUMIN

- LUMIN is a PyTorch wrapper library that provides implementations for these methods
- Also includes other useful methods & classes for working with HEP data and columnar data in general, and more
  - E.g. recent update adds RNNs, CNNs, and a few graph-nets
- Links:
  - Docs
  - Github
  - Colab examples
  - Issues - contributions welcome!



GilesStrong Changes and Readme update c5ba324 14 days ago 472 commits

.vscode	more vector ops	9 months ago
docs	running tests	last month
examples	Merge pull request #85 from GilesStrong/quick_fixes	19 days ago
lumin	Merge branch 'master' of github.com:GilesStrong/lumin	14 days ago
.gitignore	Adding matrix example	6 months ago
.readthedocs.yml	style test	11 months ago
CHANGES.md	Changes and Readme update	14 days ago
CITATION.md	Adding citation	11 months ago
LICENSE	Updating licence	5 months ago
MANIFEST.in	Include missing files for sdist	2 months ago
README.md	Changes and Readme update	14 days ago
abbr.md	Docs for mat heads	7 months ago
build.md	Move to new version	5 months ago
requirements.txt	running tests	last month
setup.cfg	Install stuff	2 years ago
setup.py	Fixes	19 days ago

README.md

pypi v0.5.1 python 3.6 | 3.7 license Apache Software License 2.0 DOI 10.5281/zenodo.3664978

LUMIN: Lumin Unifies Many Improvements for Networks

About

LUMIN - a deep learning and data science ecosystem for high-energy physics.

Releases 11

v0.5.1 - The Gradient Must ... Latest
on 12 Feb

+ 10 releases

Packages

No packages published
Publish your first package

Contributors 3

GilesStrong GilesStrong
kiryteo kiryteo
thatch thatch

Languages

Python 100.0%

USAGE

- LUMIN can be used to train neural networks for supervised classification and regression tasks using:
  - Columnar data (features in columns - events in rows)
  - And/or matrix data with arbitrary dimensions (i.e. 1D of 4-vectors, 2D & 3D grids of data, et cetera)
- Data must be coerced into a specific format: HDF5 with an expected layout
  - Methods provided to help with this
- Trained models can be exported to ONNX and TensorFlow
  - Can run in CMSSW via Tensorflow interface, see e.g cms_hh_tf_inference

A FEW DISTINGUISHING CHARACTERISTICS

- Ensembling - Training and applying with 10 models should be as easy as with 1 model
  - User defines **how** models should be built and training function creates and trains models
  - A opposed to the user building and training single models
- Modularity - Classes, methods, and workflow should be flexible and adaptable without heavy hacking
  - Expected workflow provided, but user free to cherry-pick specific aspects of the framework
  - User can inherit from existing classes to adjust to their own needs

A FEW DISTINGUISHING CHARACTERISTICS

- Automatic feature selection - Large menus of potential inputs can be filtered safely to only most useful set
- Modern techniques - Users should be able to easily apply the latest, useful, techniques
- Weight handling - All data-handling should expect sample weights
- Interpretation - Users should know what their models learnt and used during training

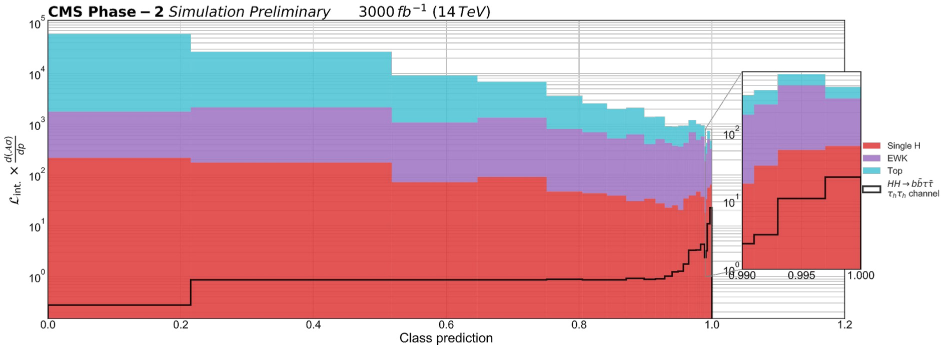
PROJECT EXAMPLES

Past and current usage of LUMIN (or its core techniques)

DI-HIGGS @ HL-LHC

M.Bengala, M.Gallinaro, R.Santo, & G.Strong, 2018-19

- HL-LHC projection studies for $hh \rightarrow bb\tau\tau$
- Completed prior to LUMIN, but used similar techniques as the Higgs ML study
- 20 DNNs trained as binary classifiers for signal|background
- Usage of advanced methods showed 30% improvement in sensitivity



CMS AN-2018/205, CMS PAS FTR-18-019, CERN-LPCC-2018-04

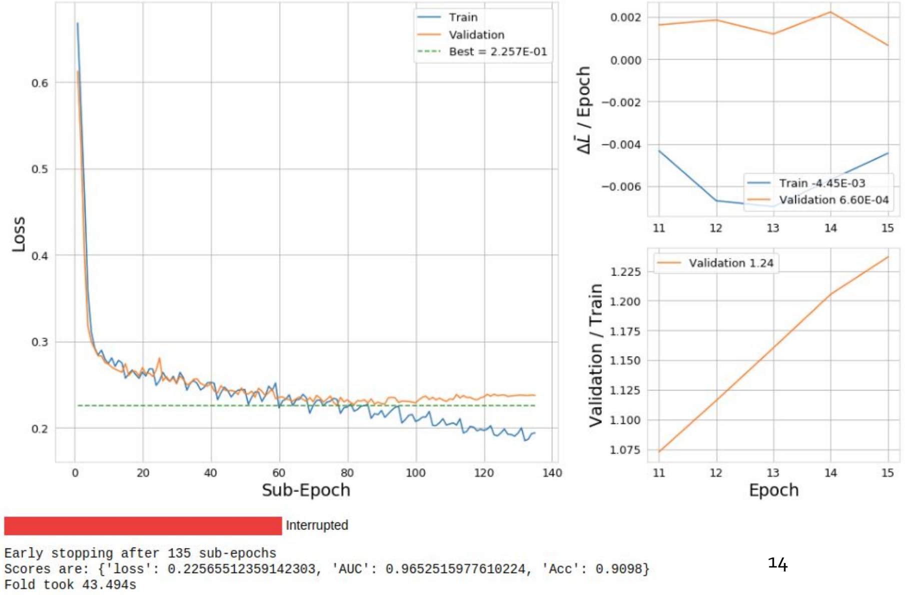
PREDICTIVE ANALYTICS

L.Cazon, R.Conceicao, A.Kocak, R.Lima, F.Riehn, C.Silva, & G.Strong, 2019

- Industry partnership between LIP and Nielsen (global data-measurement company)
- Aim was to develop a predictive model to help proactively retain employees
- LUMIN used to:
  - Automatically select relevant features from menu of several hundred
  - Highlight differences between datasets
  - Tune hyperparameters of model
- Unfortunately, most details are behind a NDA...

EXAMPLE: TOP-TAGGING FROM JET CONSTITUENTS

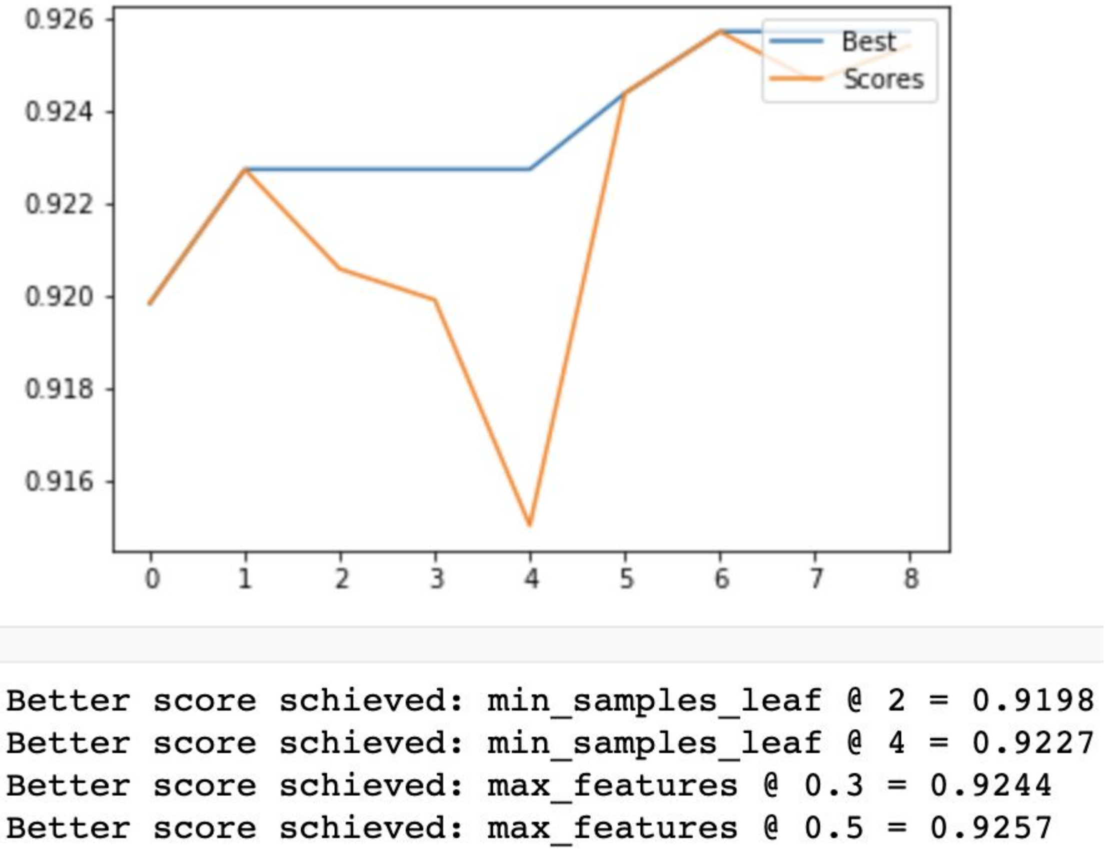
- HEP benchmark dataset for top tagging
- Data format: flat, 200 4-vectors, 1.2M jets
- LUMIN example #9 demonstrates:
    - Recursive networks
    - Convolutional networks (inc. ResNet, ResNeXT blocks)
    - Graph nets: Interaction net [1,2], Lorentz Boost Networks (LBN only in bleeding edge version)
- Only uses ~8% of total data and only 15 hardest constituents (to reduce runtime):
    - But, achieves ROC AUC of 0.965 in under 1 minute (c.f. SOTA 0.984)



```
Early stopping after 135 sub-epochs
Scores are: {'loss': 0.22565512359142303, 'AUC': 0.9652515977610224, 'Acc': 0.9098}
Fold took 43.494s
```

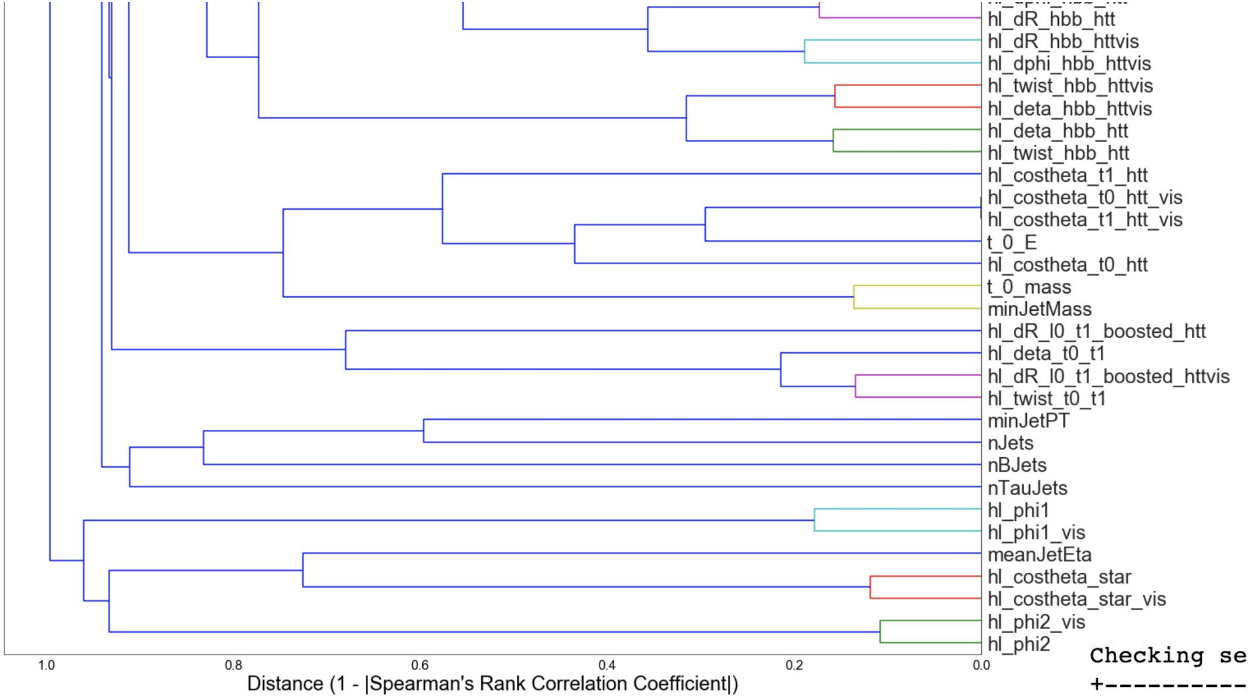
SCREENSHOTS

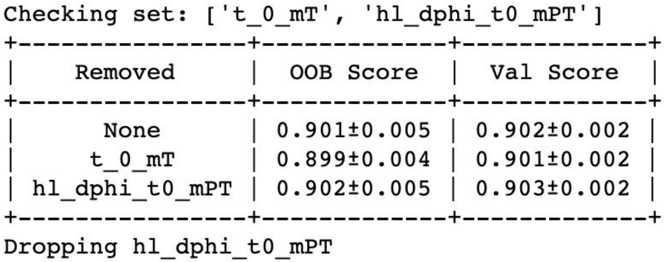
HYPER-PARAMETER OPTIMISATION: RANDOM FOREST



```
Better score schieved: min_samples_leaf @ 2 = 0.9198
Better score schieved: min_samples_leaf @ 4 = 0.9227
Better score schieved: max_features @ 0.3 = 0.9244
Better score schieved: max_features @ 0.5 = 0.9257
```

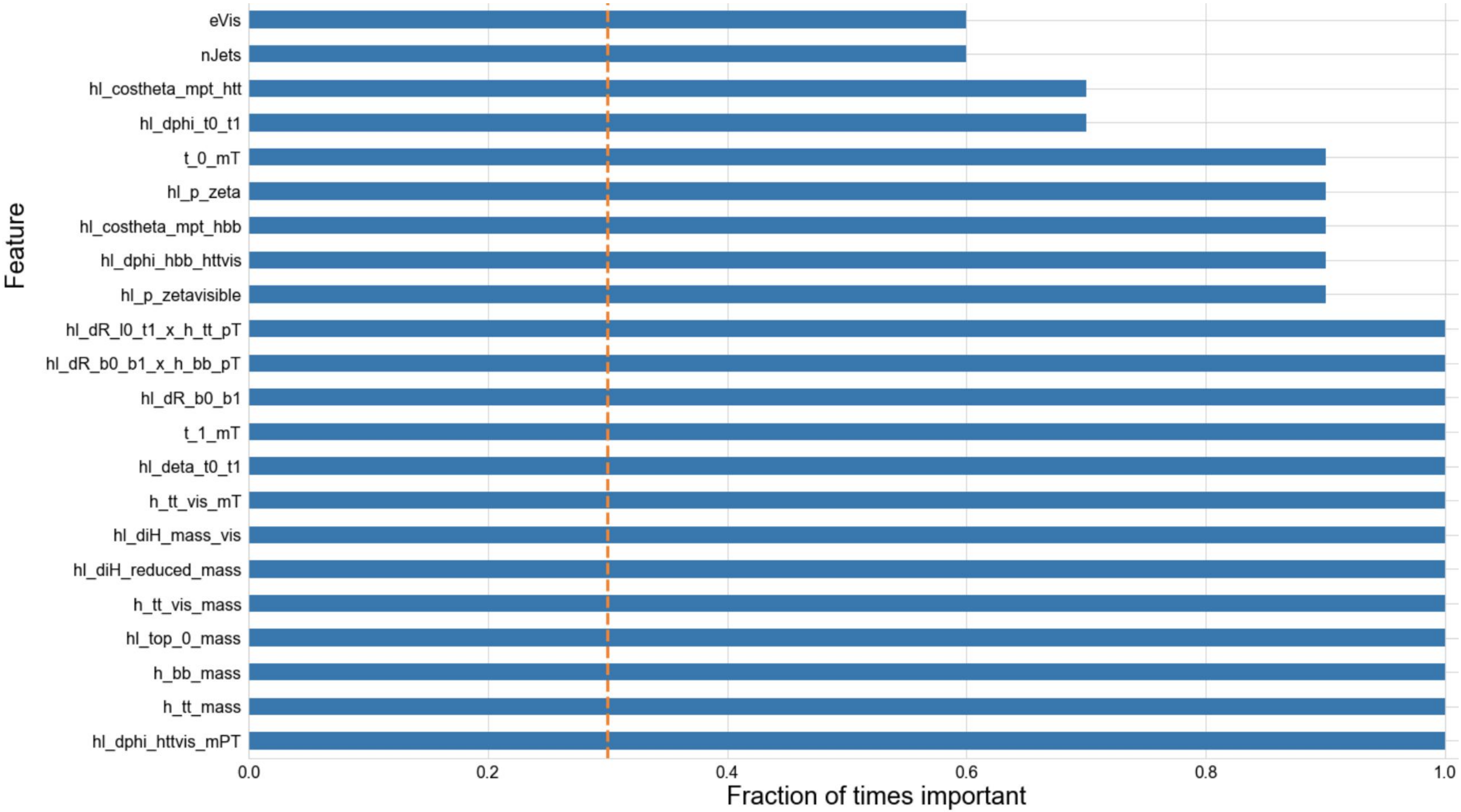
FEATURE SELECTION: CLUSTERING & REMOVAL OF CORRELATED FEATURES



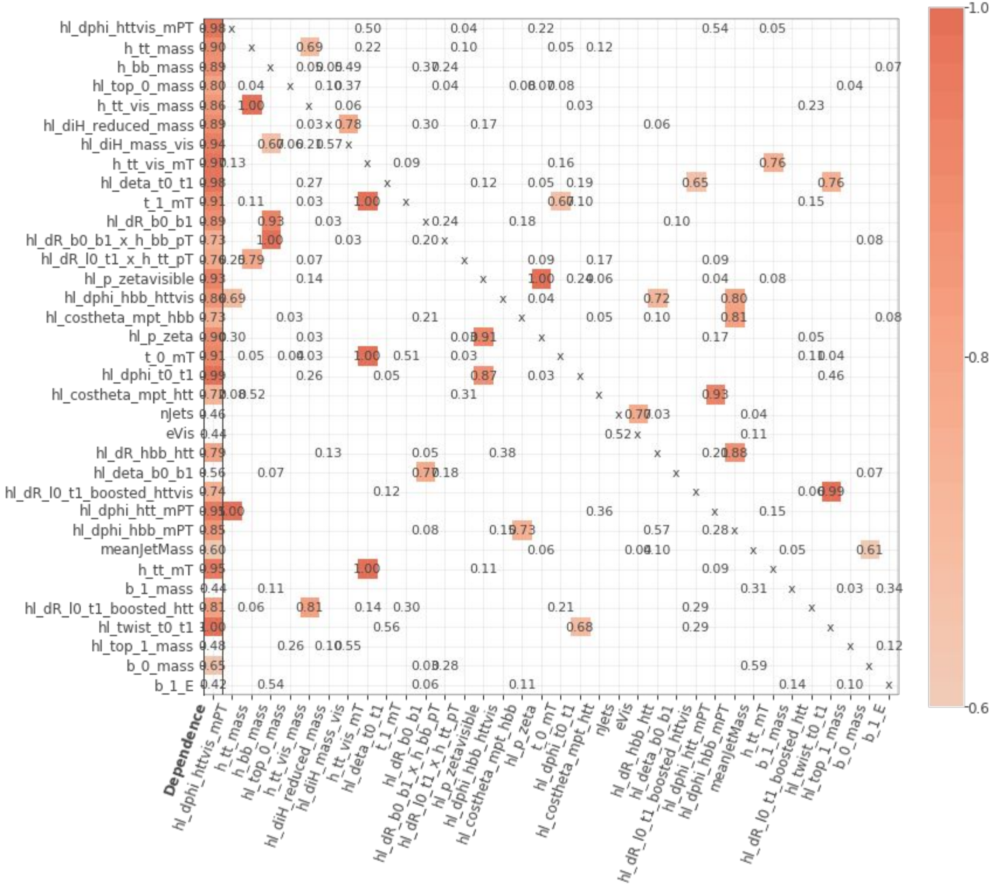


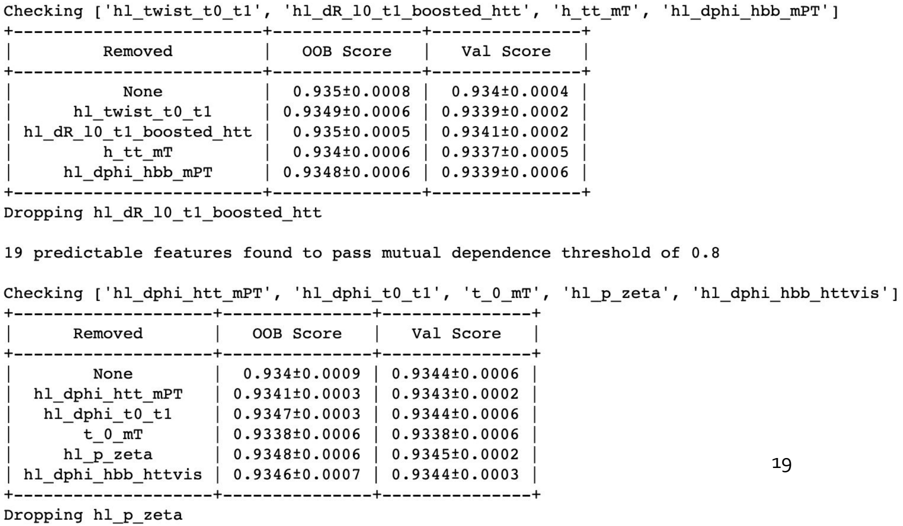
```
Checking set: ['t_0_mT', 'hl_dphi_t0_mPT']
+----------------+-------------+-------------+
|    Removed     |  OOB Score  |  Val Score  |
+----------------+-------------+-------------+
|      None      | 0.901±0.005 | 0.902±0.002 |
|     t_0_mT     | 0.899±0.004 | 0.901±0.002 |
| hl_dphi_t0_mPT | 0.902±0.005 | 0.903±0.002 |
+----------------+-------------+-------------+
Dropping hl_dphi_t0_mPT
```

FEATURE SELECTION: CONSISTENT IMPORTANCE



FEATURE SELECTION: REMOVAL OF MUTUALLY DEPENDENT FEATURES

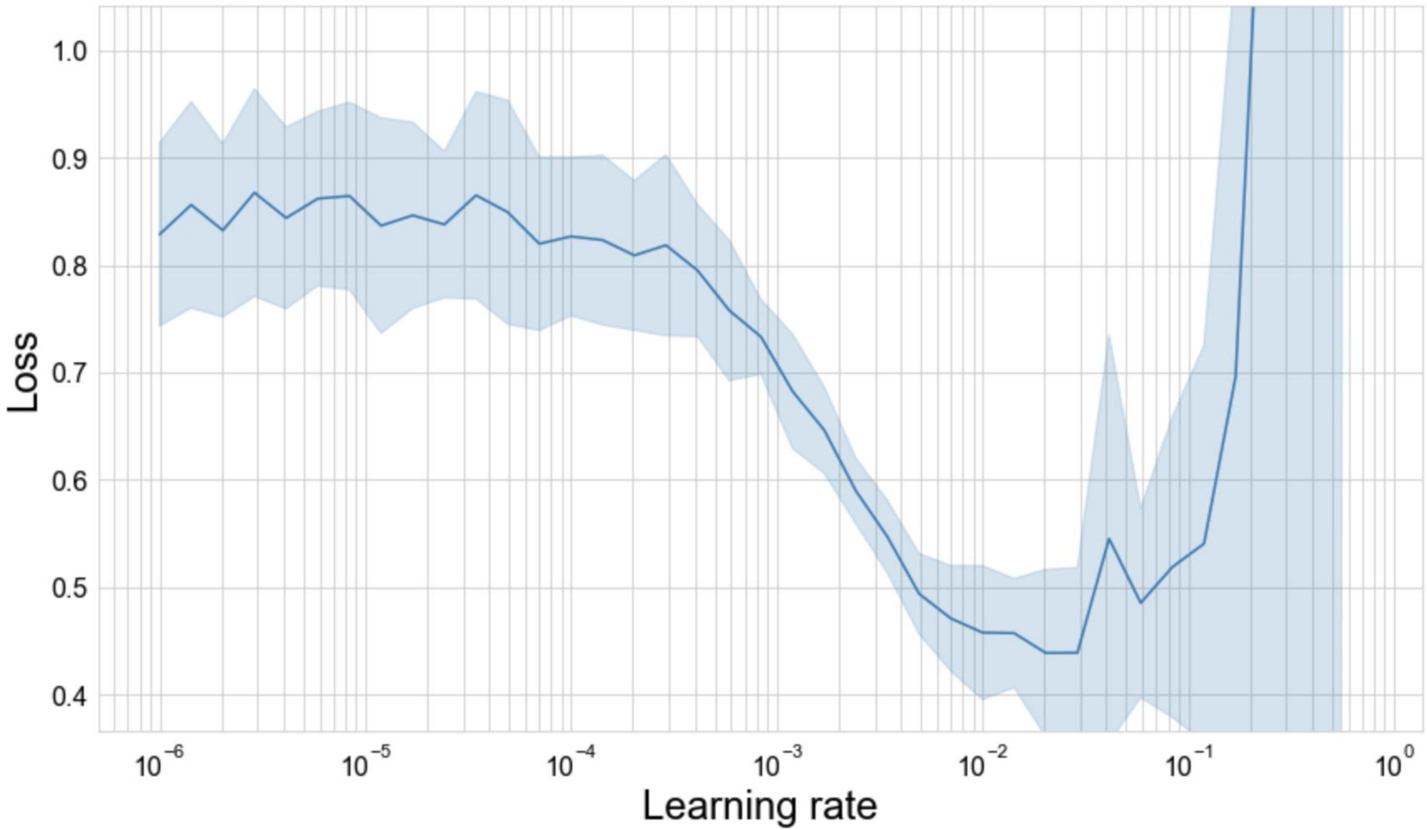




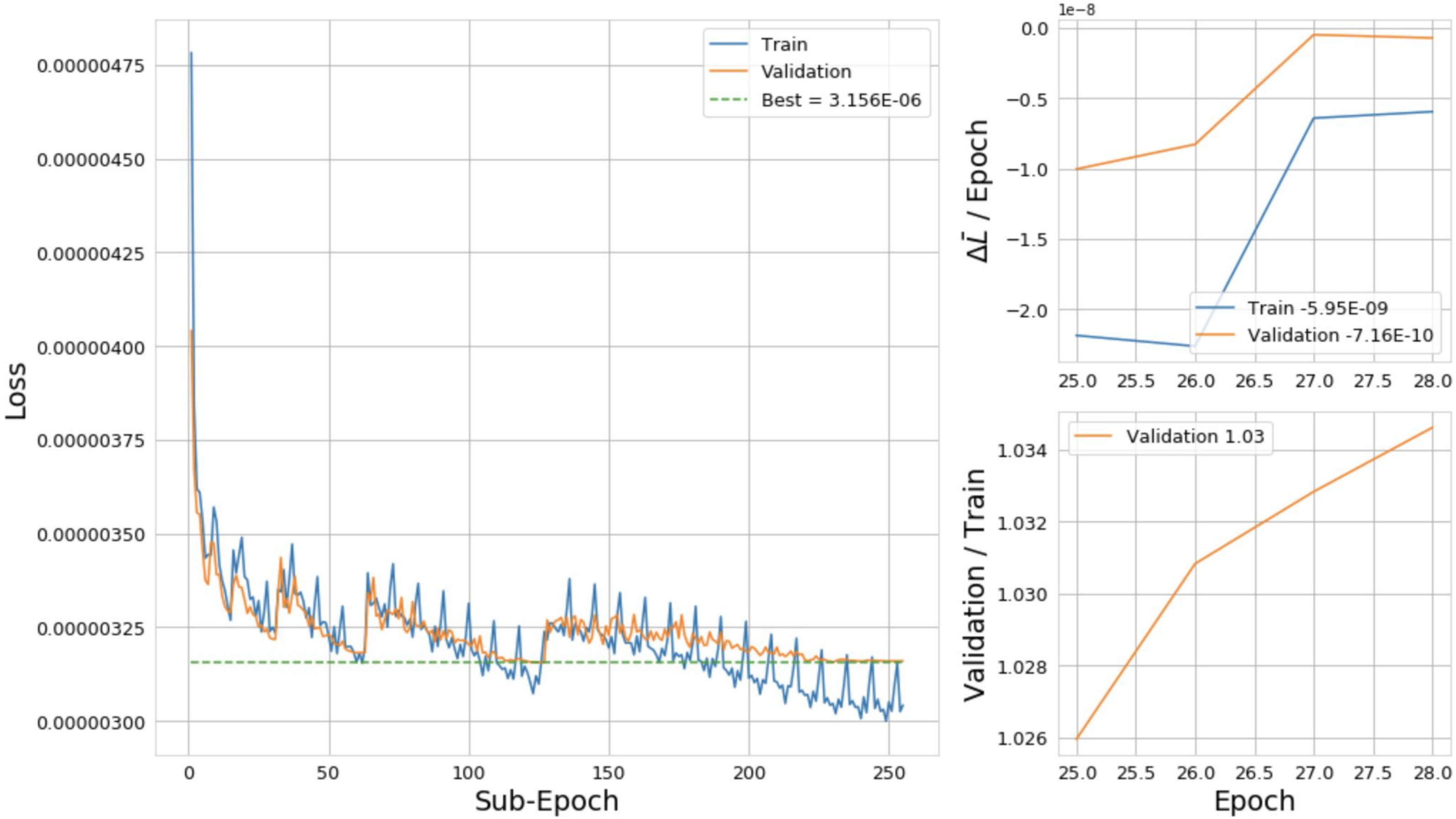
```
Checking ['hl_twist_t0_t1', 'hl_dR_l0_t1_boosted_htt', 'h_tt_mT', 'hl_dphi_hbb_mPT']
+-------------------------+---------------+---------------+
|         Removed         |   OOB Score   |   Val Score   |
+-------------------------+---------------+---------------+
|           None          |  0.935±0.0008 |  0.934±0.0004 |
|      hl_twist_t0_t1     | 0.9349±0.0006 | 0.9339±0.0002 |
| hl_dR_l0_t1_boosted_htt |  0.935±0.0005 | 0.9341±0.0002 |
|         h_tt_mT         |  0.934±0.0006 | 0.9337±0.0005 |
|     hl_dphi_hbb_mPT     | 0.9348±0.0006 | 0.9339±0.0006 |
+-------------------------+---------------+---------------+
Dropping hl_dR_l0_t1_boosted_htt

19 predictable features found to pass mutual dependence threshold of 0.8

Checking ['hl_dphi_htt_mPT', 'hl_dphi_t0_t1', 't_0_mT', 'hl_p_zeta', 'hl_dphi_hbb_httvis']
+--------------------+---------------+---------------+
|      Removed       |   OOB Score   |   Val Score   |
+--------------------+---------------+---------------+
|        None        |  0.934±0.0009 | 0.9344±0.0006 |
|  hl_dphi_htt_mPT   | 0.9341±0.0003 | 0.9343±0.0002 |
|   hl_dphi_t0_t1    | 0.9347±0.0003 | 0.9344±0.0006 |
|       t_0_mT       | 0.9338±0.0006 | 0.9338±0.0006 |
|     hl_p_zeta      | 0.9348±0.0006 | 0.9345±0.0002 |
| hl_dphi_hbb_httvis | 0.9346±0.0007 | 0.9344±0.0003 |
+--------------------+---------------+---------------+
Dropping hl_p_zeta
```

HYPER-PARAMETER OPTIMISATION: LEARNING RATE

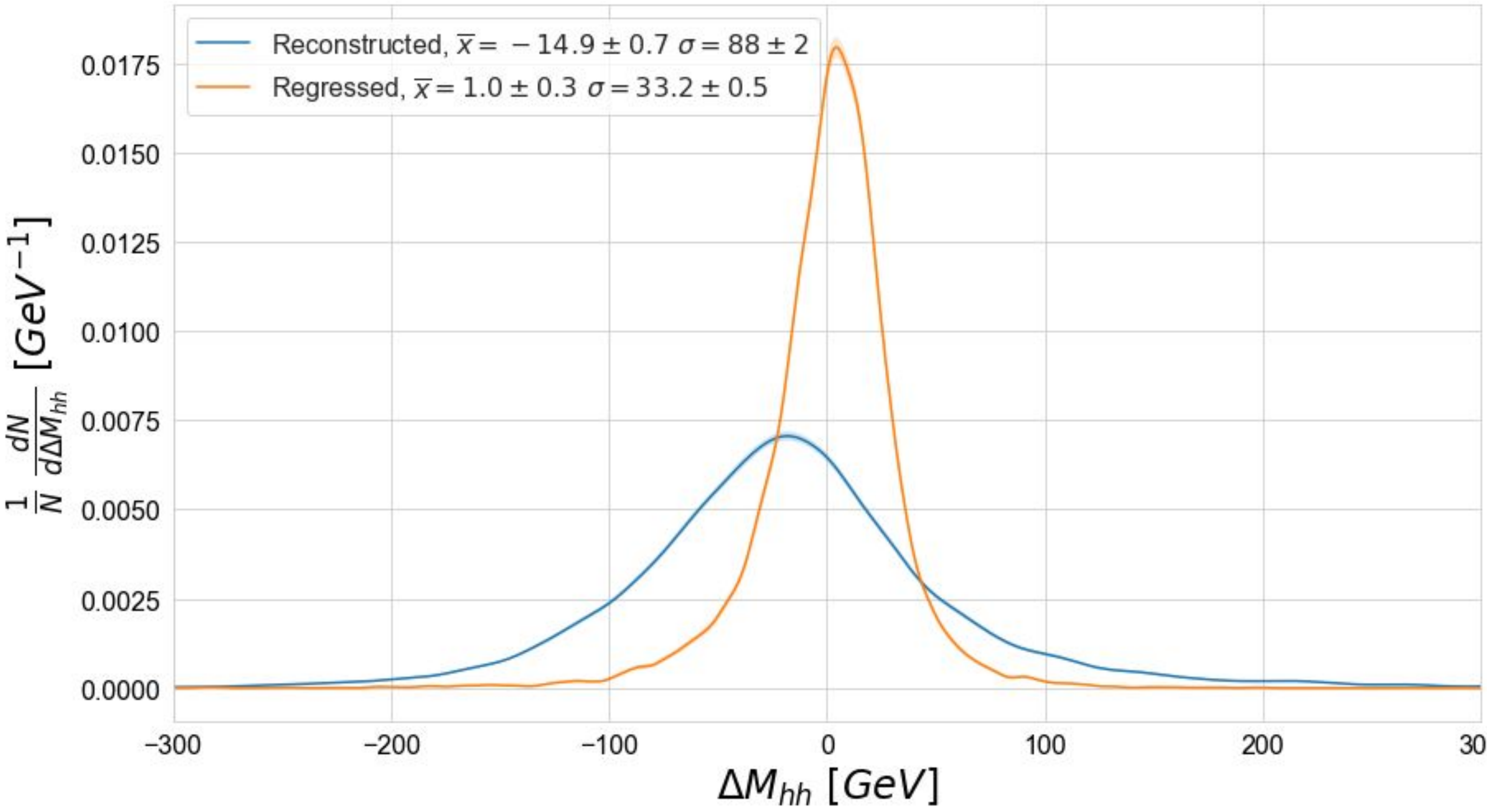
```
LR finder took 1.811s
```



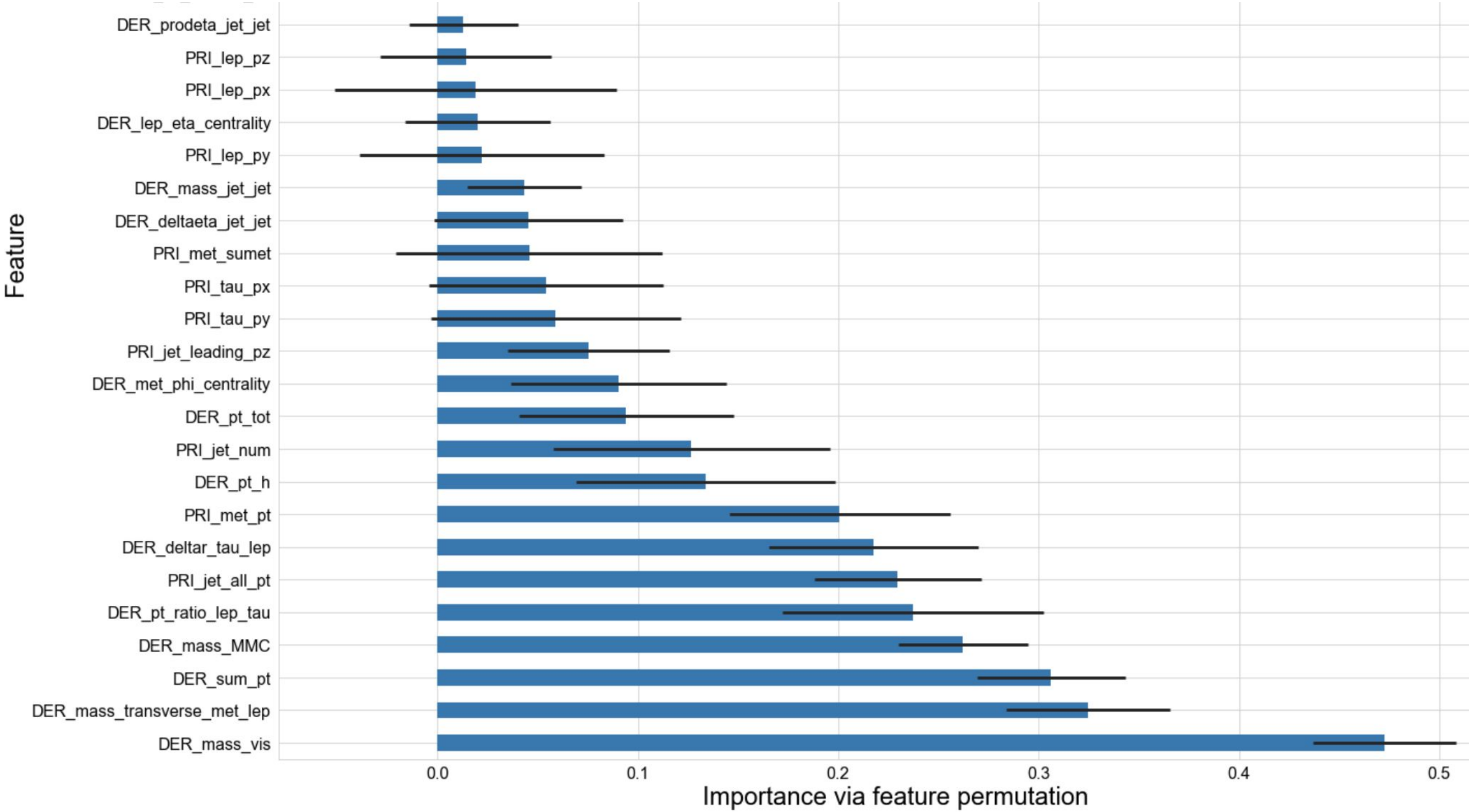
LIVE TRAINING-MONITORING



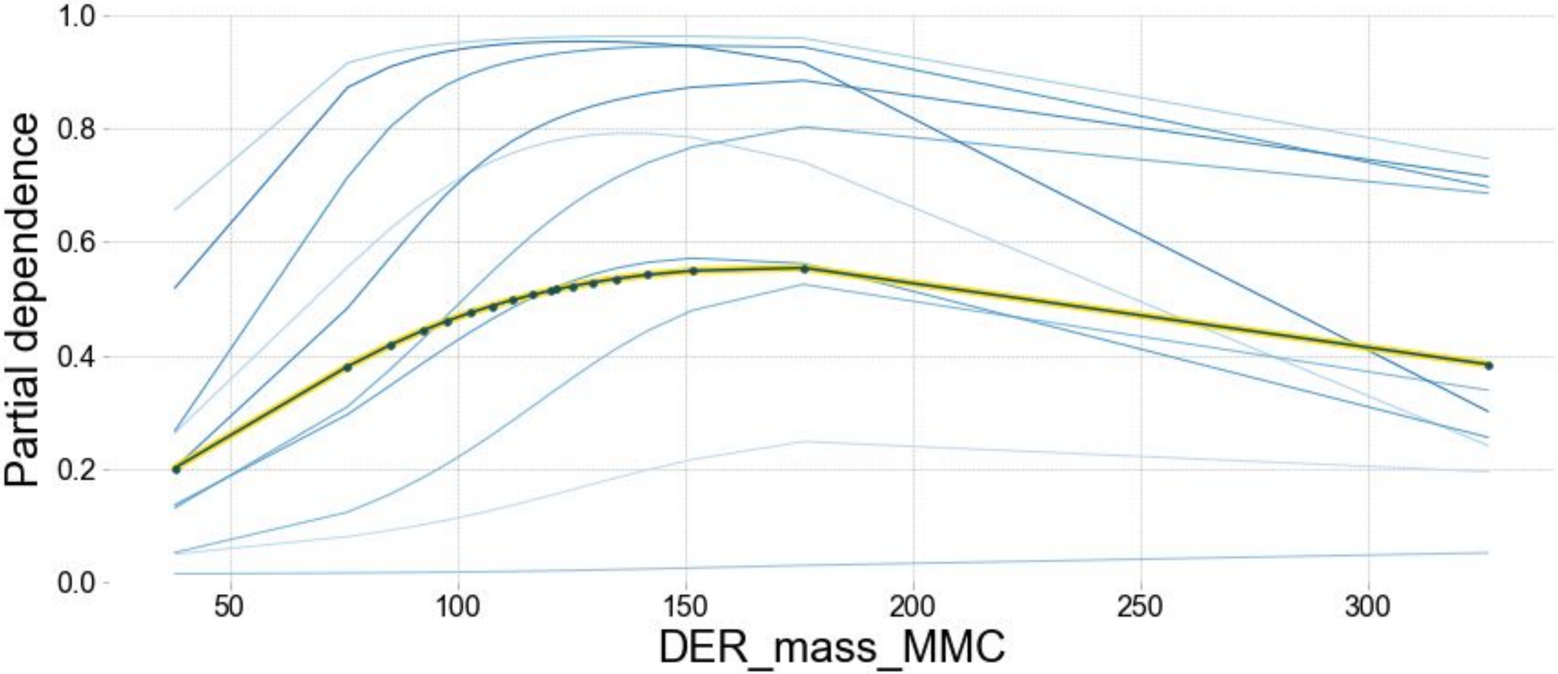
BOOTSTRAPPED KDE PLOTS



INTERPRETATION: FEATURE IMPORTANCE



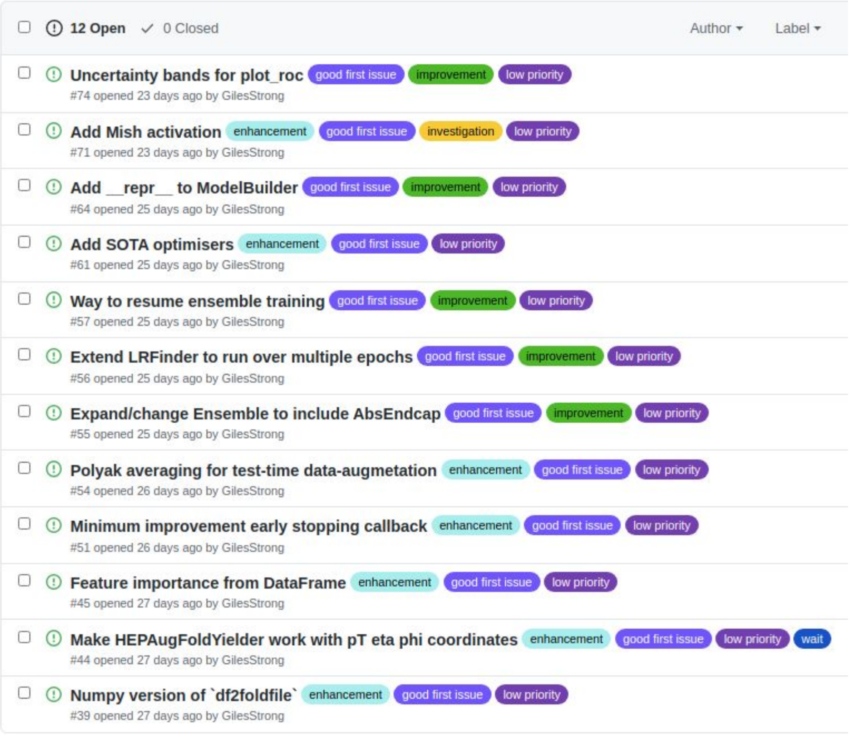
INTERPRETATION: PARTIAL DEPENDENCE



SUMMARY

CALL FOR CONTRIBUTIONS

- LUMIN has already been used in several diverse projects
  - But so far only used by me (to my knowledge)
- The package needs people trying it out, playing around, and giving feedback on:
  - Bugs
  - Design & layout choices
  - General suggestions
- Several examples available
- The issues include all my thoughts on possible improvements
  - Dedicated "good first issue" label for getting to know the code base



□ 12 Open ✓ 0 Closed Author ▾ Label ▾

□ Uncertainty bands for plot_roc good first issue improvement low priority
#74 opened 23 days ago by GilesStrong

□ Add Mish activation enhancement good first issue investigation low priority
#71 opened 23 days ago by GilesStrong

□ Add __repr__ to ModelBuilder good first issue improvement low priority
#64 opened 25 days ago by GilesStrong

□ Add SOTA optimisers enhancement good first issue low priority
#61 opened 25 days ago by GilesStrong

□ Way to resume ensemble training good first issue improvement low priority
#57 opened 25 days ago by GilesStrong

□ Extend LRFinder to run over multiple epochs good first issue improvement low priority
#56 opened 25 days ago by GilesStrong

□ Expand/change Ensemble to include AbsEndcap good first issue improvement low priority
#55 opened 25 days ago by GilesStrong

□ Polyak averaging for test-time data-augmetation enhancement good first issue low priority
#54 opened 26 days ago by GilesStrong

□ Minimum improvement early stopping callback enhancement good first issue low priority
#51 opened 26 days ago by GilesStrong

□ Feature importance from DataFrame enhancement good first issue low priority
#45 opened 27 days ago by GilesStrong

□ Make HEPAugFoldYielder work with pT eta phi coordinates enhancement good first issue low priority wait
#44 opened 27 days ago by GilesStrong

□ Numpy version of `df2foldfile` enhancement good first issue low priority
#39 opened 27 days ago by GilesStrong