(Re)interpretation of the LHC
results for new physics:

Publication and reuse of ML models

7th workshop of the
LHC Reinterpretation Forum

12 Dec 2022

CERN and online

Simulation-based reinterpretation (“recasting”)

- Aims at reproducing experimental analyses in Monte Carlo simulation
- Nowadays well established for traditional cut-based analyses. Information needed: cf. arXiv:2003.07868

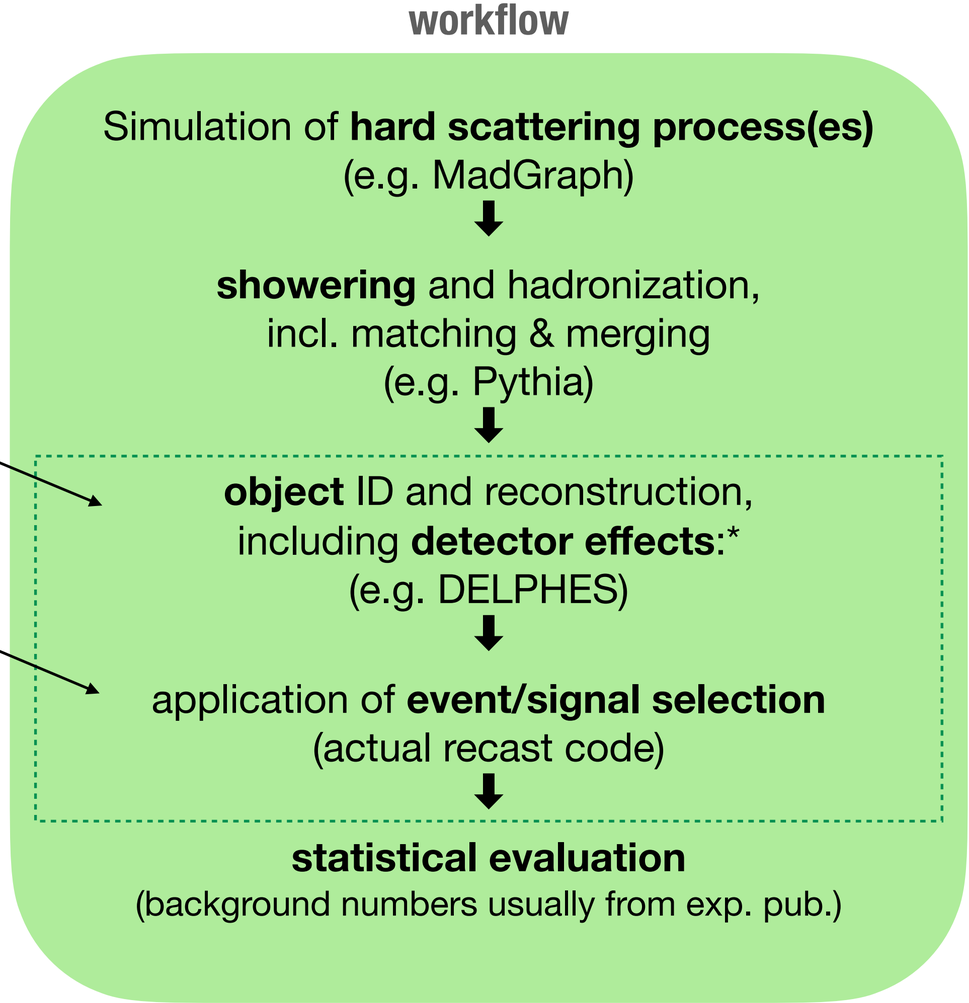
object definitions; identification, tagging, reconstruction efficiencies

detailed preselection and signal (+control) region cuts

- However, more and more analyses exploit ML techniques to gain in sensitivity

e.g. ML-based taggers, signal/bkg discrimination with ML classifiers

Pb: how can we reuse those?



* except for detector-unfolded results (Rivet/Contur)

ML as a bottleneck for reinterpretation?

- More and more analyses exploit ML techniques to gain in sensitivity.
- Serious difficulty for analysis preservation and reuse unless
  - resulting id/reco efficiencies can be (and are!) parametrised in terms of quantities accessible in a simulation, e.g., p_t, η, …
  - the actual ML model is published in appropriate form.

    Caveat: input variables need to be physics quantities that can be matched in a simulation

Two analyses where the latter has been attempted:

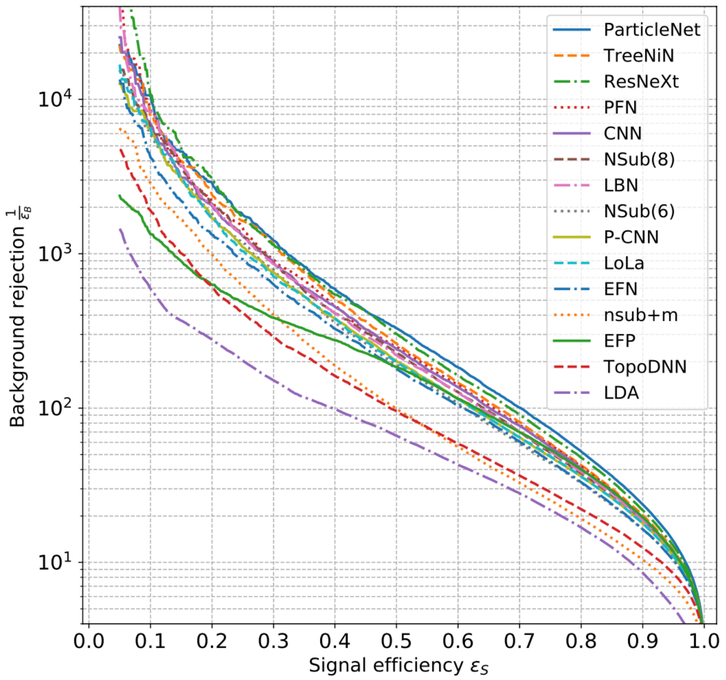
- **ATLAS-SUSY-2018-22** (0-lepton gluino/squark search) published **BDT weights** as XML file

  → RAMP seminar by Kenta Uno
- **ATLAS-SUSY-2019-04** (1-2 leptons + jets RPV search) published neural network as **ONNX** file

  → RAMP seminar by Javier Montejo Berlingen

RAMP: Reinterpretation Auxiliary Material Presentation

Example: top taggers



13 different algorithms: image-based (2), 4-vector-based (5), theory-inspired (6) taggers

"The Machine Learning Landscape of Top Taggers"
G. Kasieczka, T. Plehn et al., arXiv:1902.09914

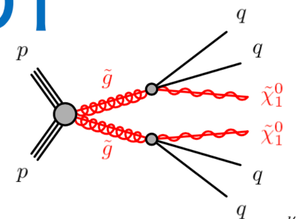
Kenta Uno on ATLAS-SUSY-2018-22

RAMP seminar #1
9 April 2021

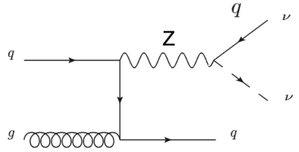
Analysis approach: BDT

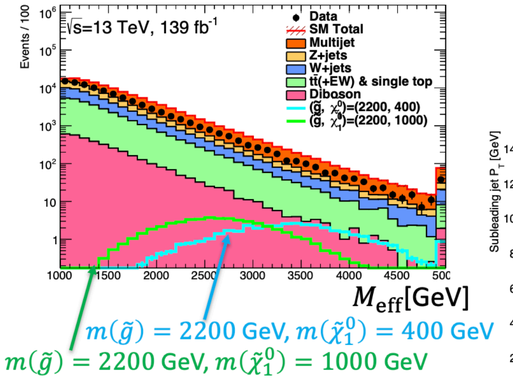
Key variable in the previous analysis: M_{eff}

$$M_{\text{eff}} = \sum_i |p_{\text{T}}(i)| + E_{\text{T}}^{\text{miss}}$$

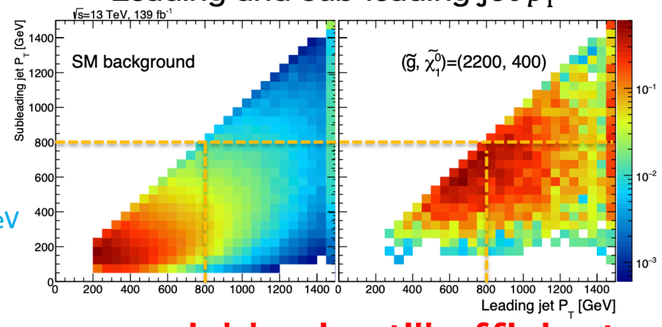


- However, it is not efficient as $\tilde{\chi}_1^0$ mass is large





Leading and sub-leading jet p_{T}



Using correlation between variables is still efficient.

→ BDT analysis is useful for SUSY search

Discussion point 1

- The auxiliary material provided for the BDT reinterpretation
  - We published HepData materials [Link]
    - Provided xml files of BDT classifiers
      → ZeroLepton2018-SRBDT-weight.tar.gz
    - Also provided analysis snipped code
      → ZeroLeptonBDT2018.cxx
      → By using this code, you can use same BDT classifiers.
    - Full likelihood is provided in the hepdata material

※ Signal events: fast sim (reco)
※ Background events: full reco

We provide data, SM total and signal yields in each BDT score bin and the acceptance and efficiencies of BDT SRs for simplified GG models.

→ I think it might help everyone..

10 -12 input variables: jet $p_{\text{T}}, \eta, E_{\text{T}}^{\text{miss}}, m_{\text{eff}}$, Aplanarity

Javier Montejo Berlingen on ATLAS-SUSY-2019-04

Reinterpretation material

Additional resources

C++ code snippet with the implementation of the analysis selection at truth-level

- Can be used with SimpleAnalysis framework

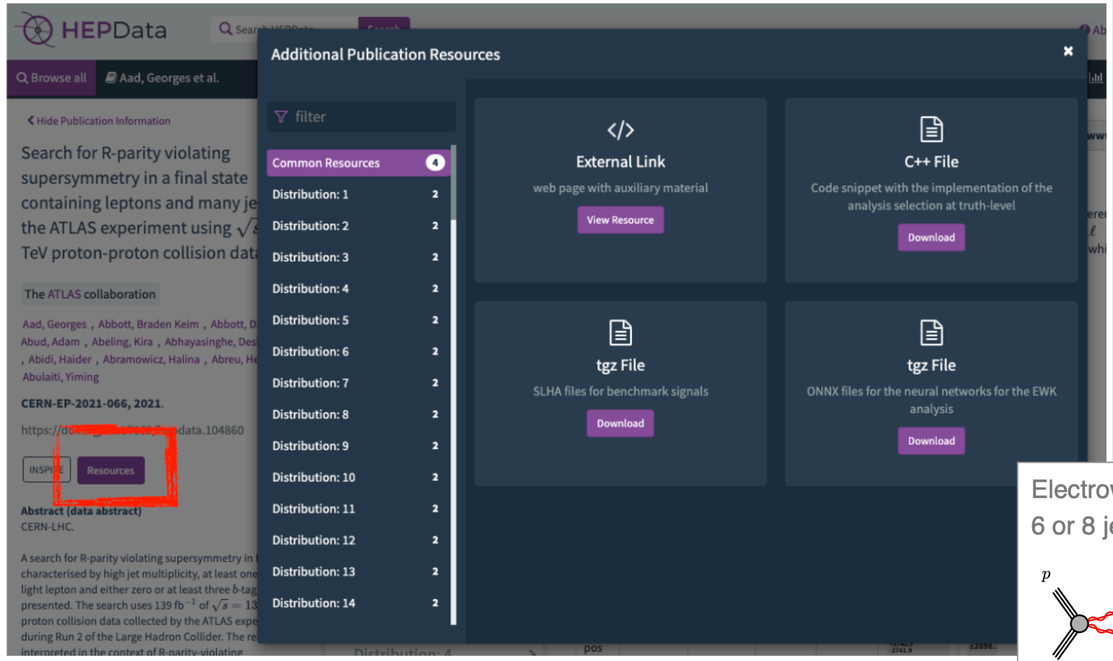
SLHA files for benchmark signals

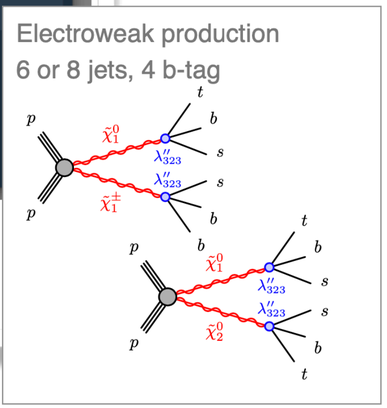
ONNX files for the neural networks for the EWK analysis

- Not possible to use lwtnn because the architecture of one of the layers is not supported

Upcoming: **ROOT workspaces**, containing data and the fitted background model in all SRs

- not using pyhf as it doesn't support our parameterised background model





Snowmass white paper on data and analysis preservation and reinterpretation

S. Bailey et al., arXiv:2203.10057

Reinterpretation and Recasting Recommendations

5.1: En
ar
M
te
nc
th

5.2: De
uc

5.3: Improve the coordination among the different public reinterpretation frameworks with the goal of a centralised database of recast codes, common input/output formats, and a unified statistical treatment.

5.4: Encourage the FAIR-ification of codes and data products from (theory) reinterpretation studies outside the experimental collaborations at the same level of sophistication as asked for experimental analyses and results. Suitable repositories are, e.g., GitHub and Zenodo; appropriate versioning is essential.

Encourage that **reinterpretability and reuse be kept in mind early on in the analysis design**. This concerns, for instance, the **choice of input parameters in ML models**, the full specification of the fiducial phase space of a measurement in terms of the final state, including any vetos applied, and generally the choice of non-overlapping regions and standard naming of shared nuisances to facilitate the combination of analyses.