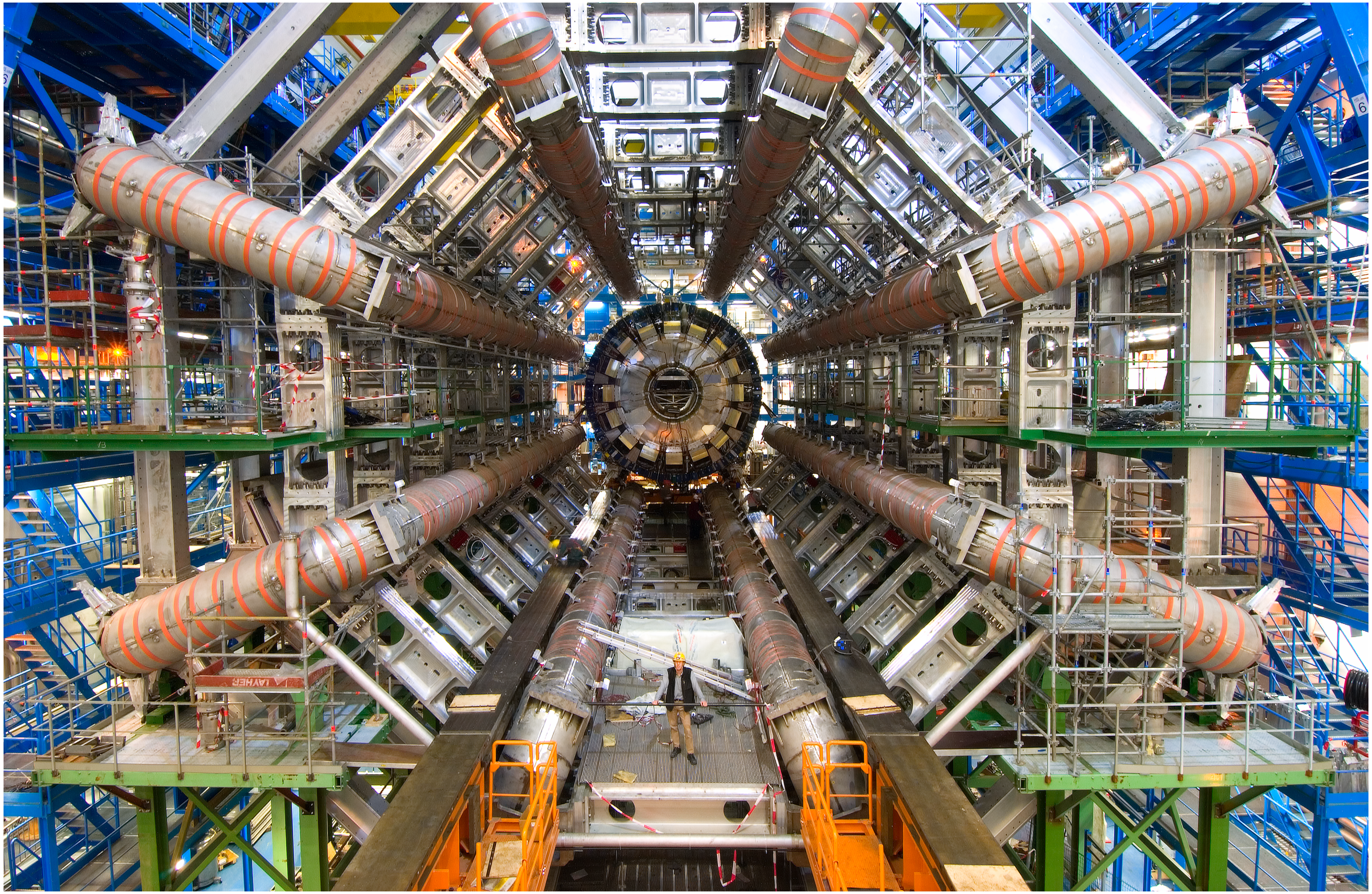
recent ML highlights in the DESY ATLAS group

Chris Pollard

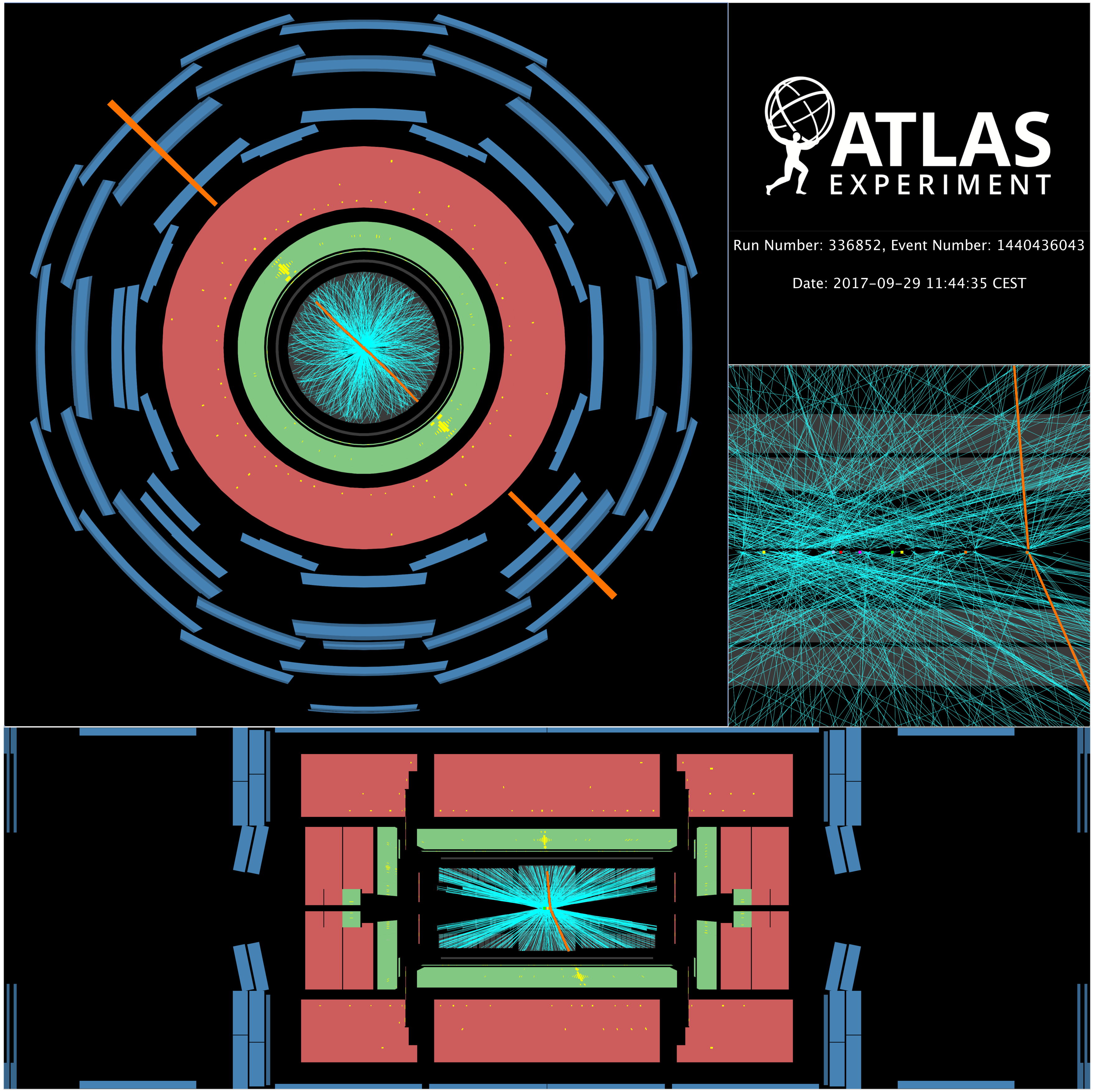


the LHC collides protons at extremely high energies

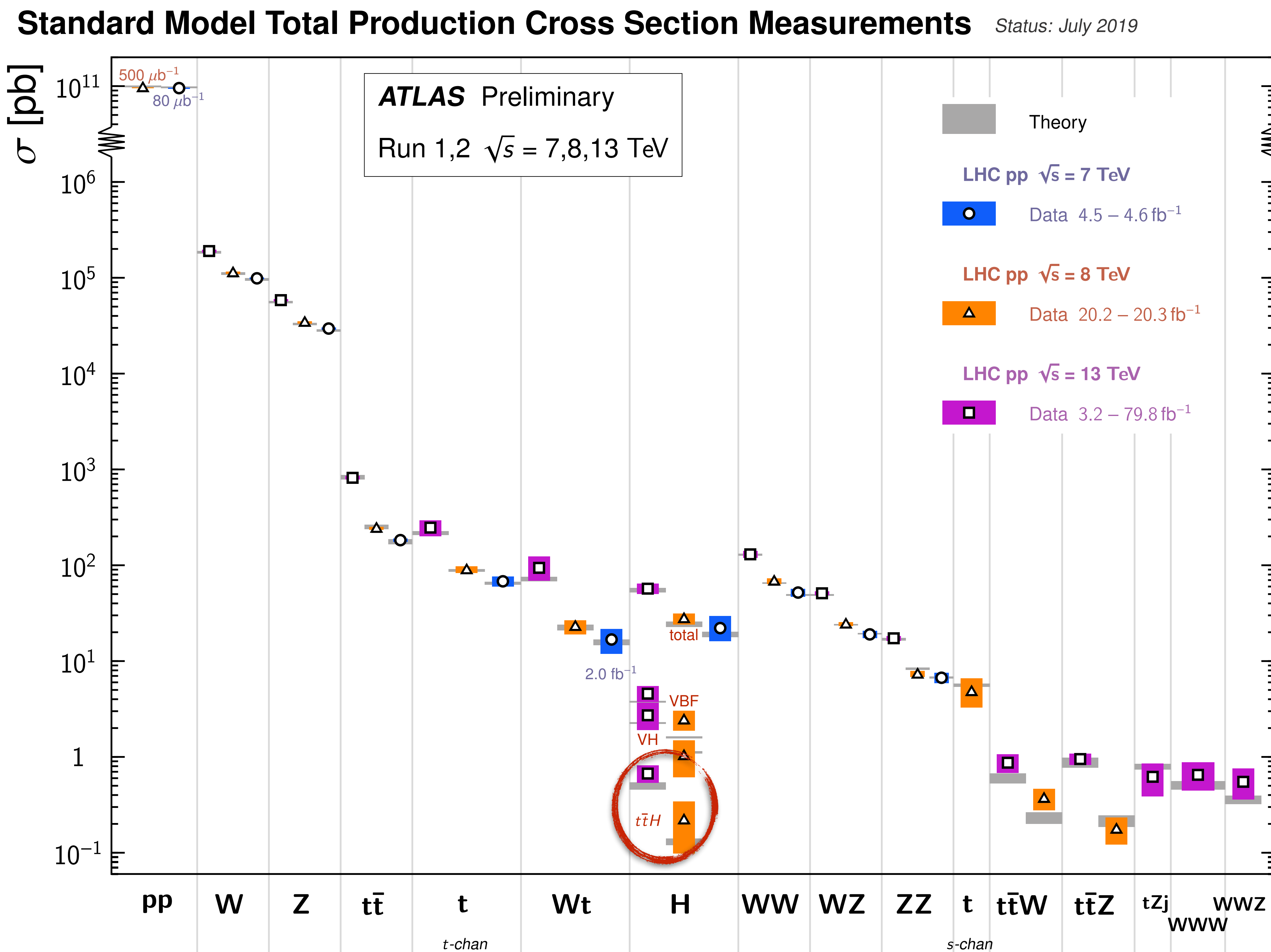
exotic particles are produced and studied by enormous, complex detectors like ATLAS and CMS

each collision produces a large number of particles

millions of readout channels required to reconstruct them adequately



y-axis range = factor of ~ a trillion

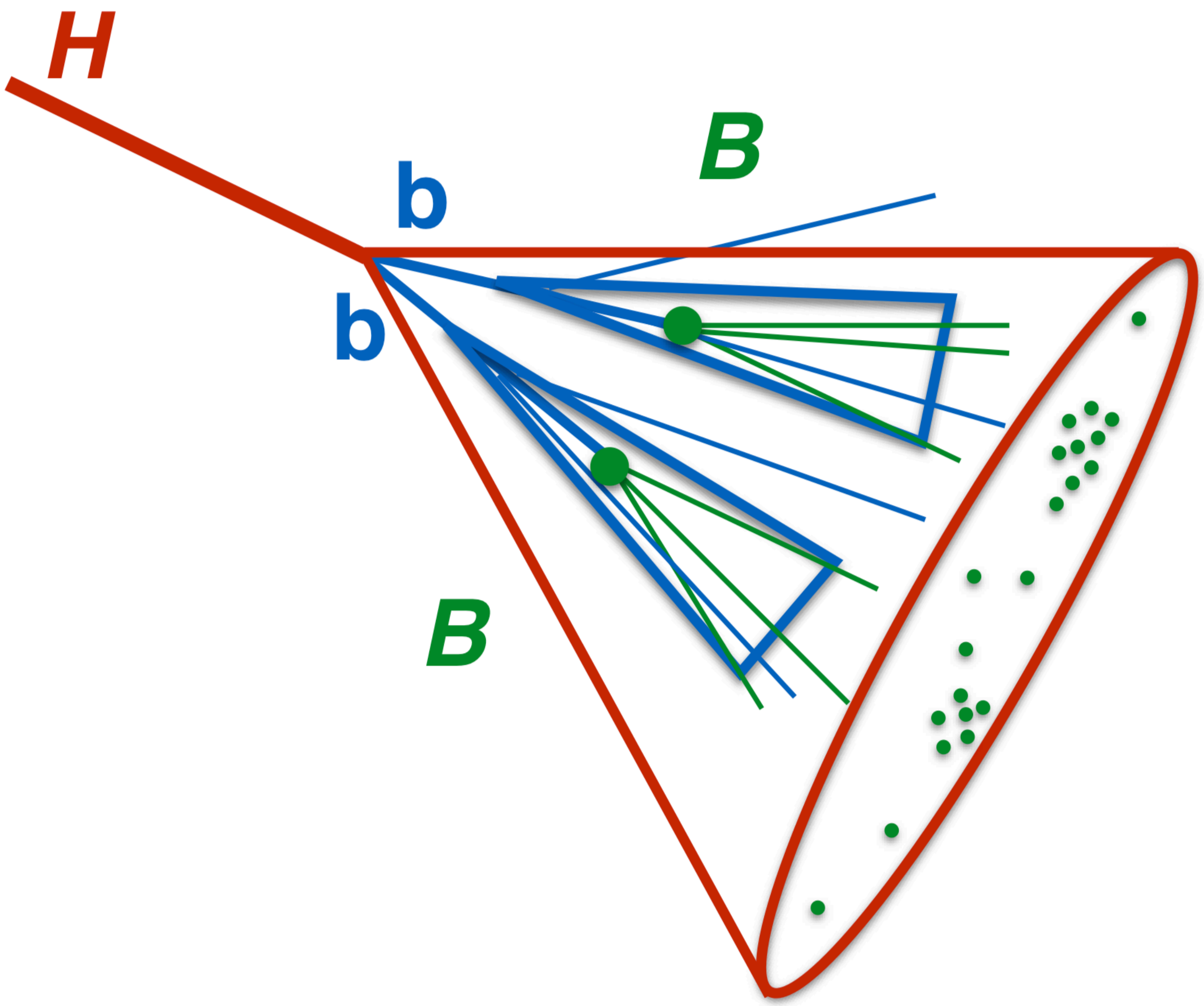


"interesting" processes to observe occur very rarely

produce ~ one billion collisions per second

- **we have a large number of observations per event**
- **we need to separate "interesting" and "uninteresting events with high accuracy**
- **we need to be able to do this very fast (in some cases *extremely* fast)**

- decades of specialist knowledge is built up in our field.
- in the last ~decade machine learning has had a huge impact.
- I'll focus on some examples that are worked-on actively in the DESY ATLAS group.



about 60% of the time Higgs bosons decay to pairs of bottom quarks

jets resulting from bottom quarks ("*b*-jets") contain particles that are not stable

i.e. these particles fly ~millimeters in the detector before decaying

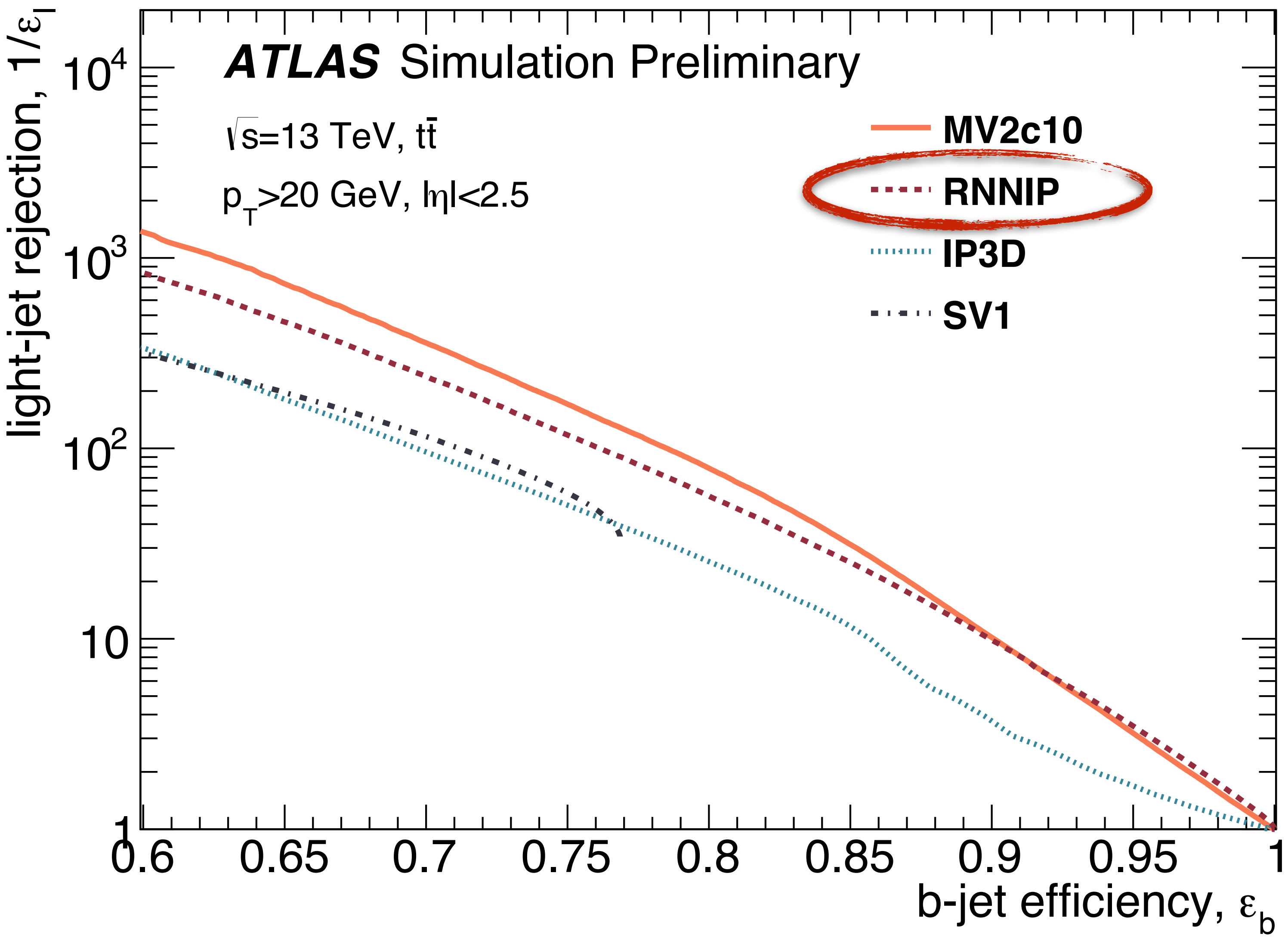
background jets ("light-jets") mostly do not contain unstable particles

light-jets are produced at a *much* higher rate than *b*-jets

lots of technology already exists for b-jet identification

for instance, algorithms that attempt to reconstruct "secondary vertices" from the decays of long-lived particles

most recently an RNN, taking reconstructed particles as inputs, was introduced resulting in significant performance gains

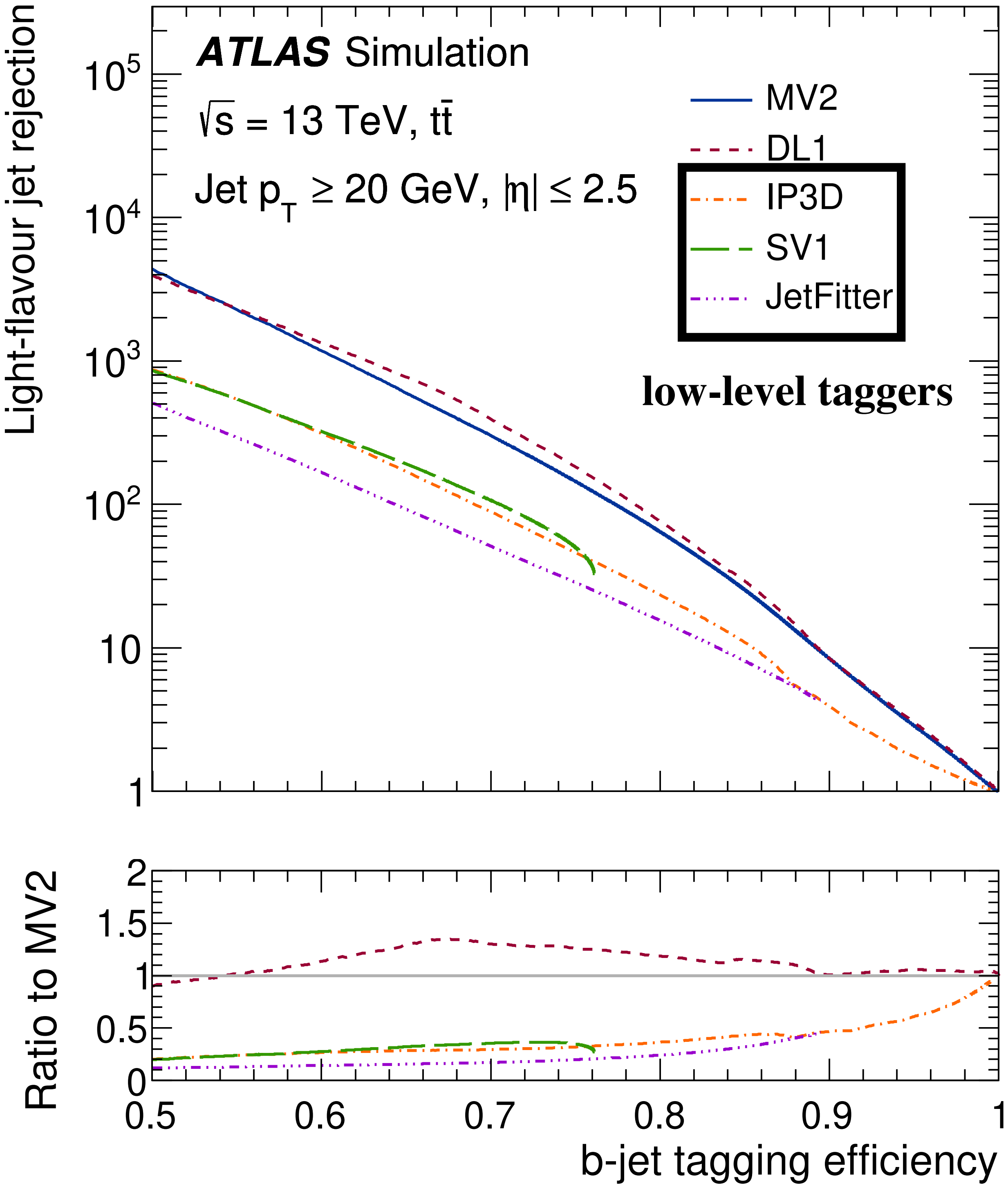


technical challenges -- time-consuming to train even on GPU farms

physics challenges -- introduces ordering which is not necessarily motivated in ATLAS we use both BDTs ("MV2") and DNNs ("DL1") to consolidate

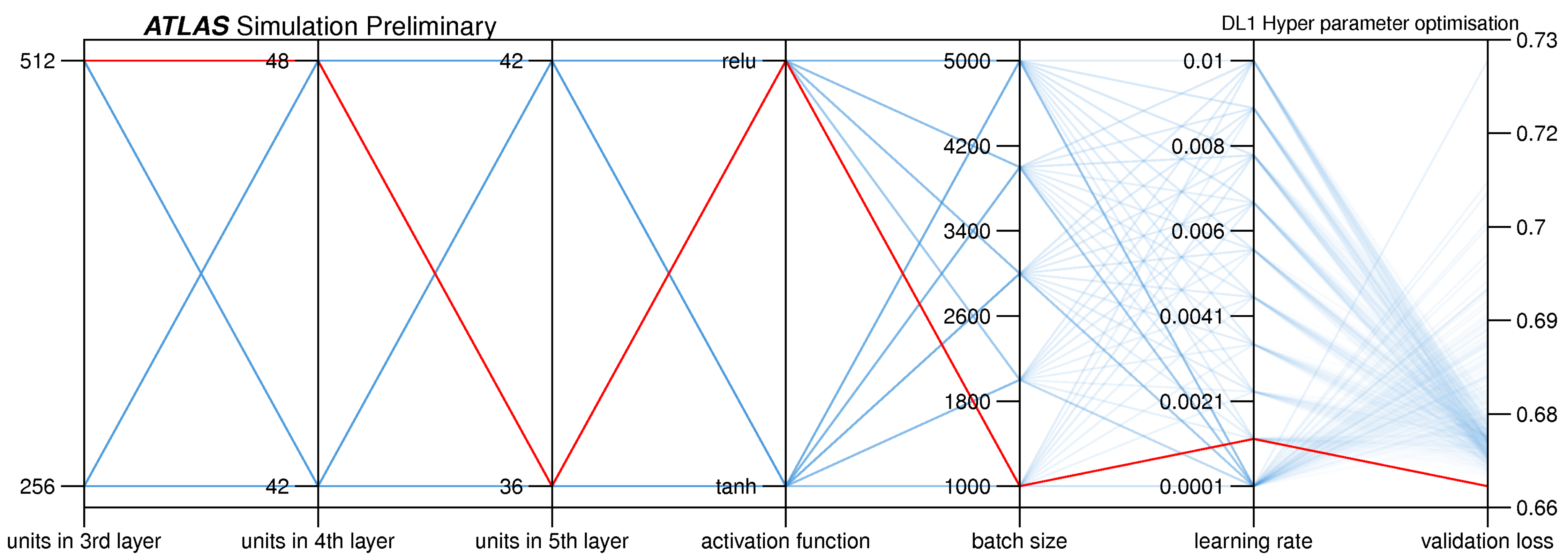
- the results of our best "specialist knowledge" algorithms
- outputs the RNN (mentioned previously)

these do a great job learning
- relative rates of b-hadron species production
- relative rates of b-hadron decays
- inefficiencies and resolutions of the detector
- correlations between various historical algorithms

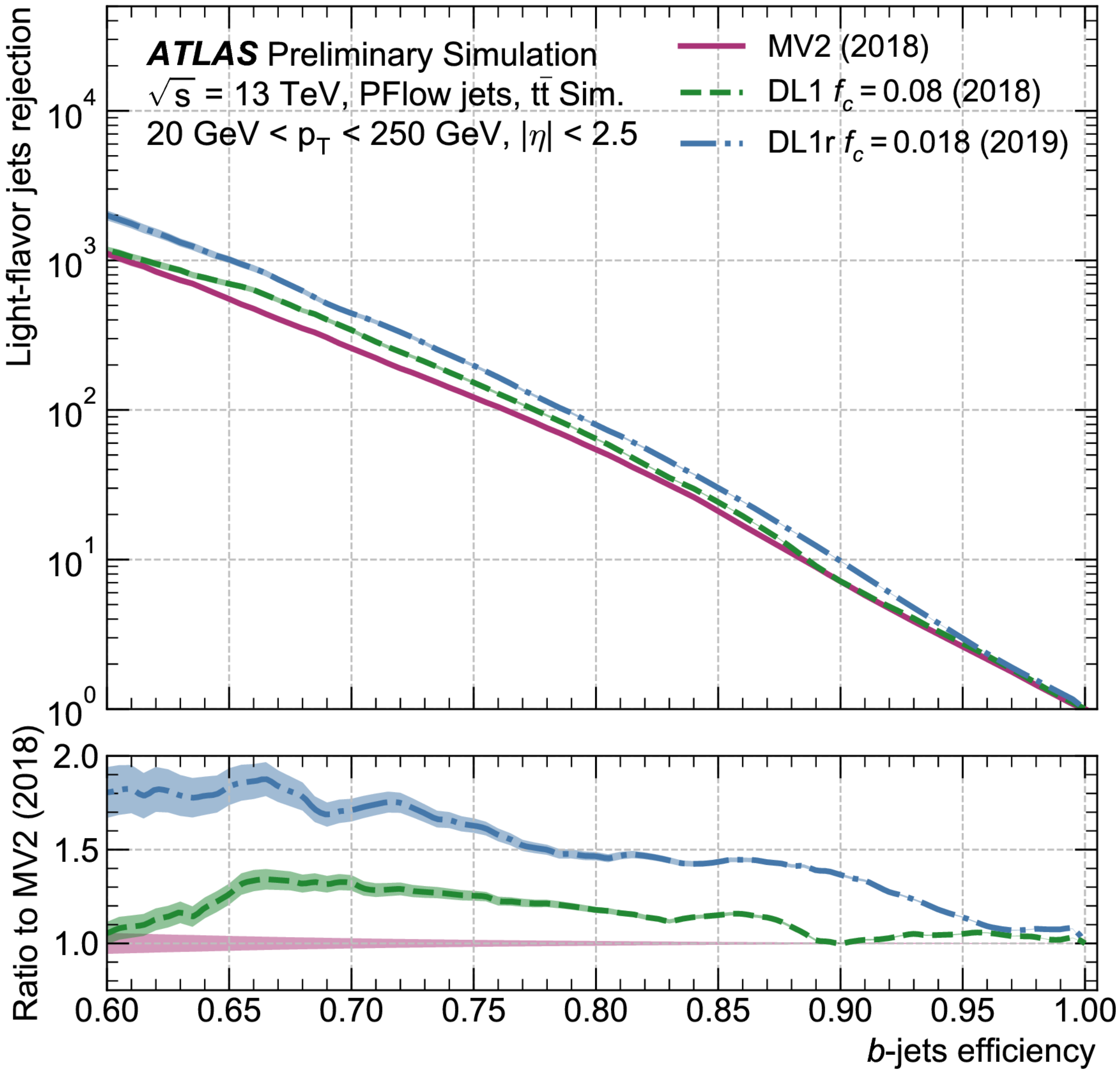


recently introduced more rigorous training of the DNN

- **first trainings performed locally**
- **hyperparameter scan performed on the LHC computing grid with GPUs**
- **containers crucial to making this possible**



more rigorous training and inclusion of the RNN pays off!



e.g. some measurements of the higgs self-coupling require 4 *b*-jets to be identified

this is a *huge* gain for such analyses of the LHC data

not so fast!

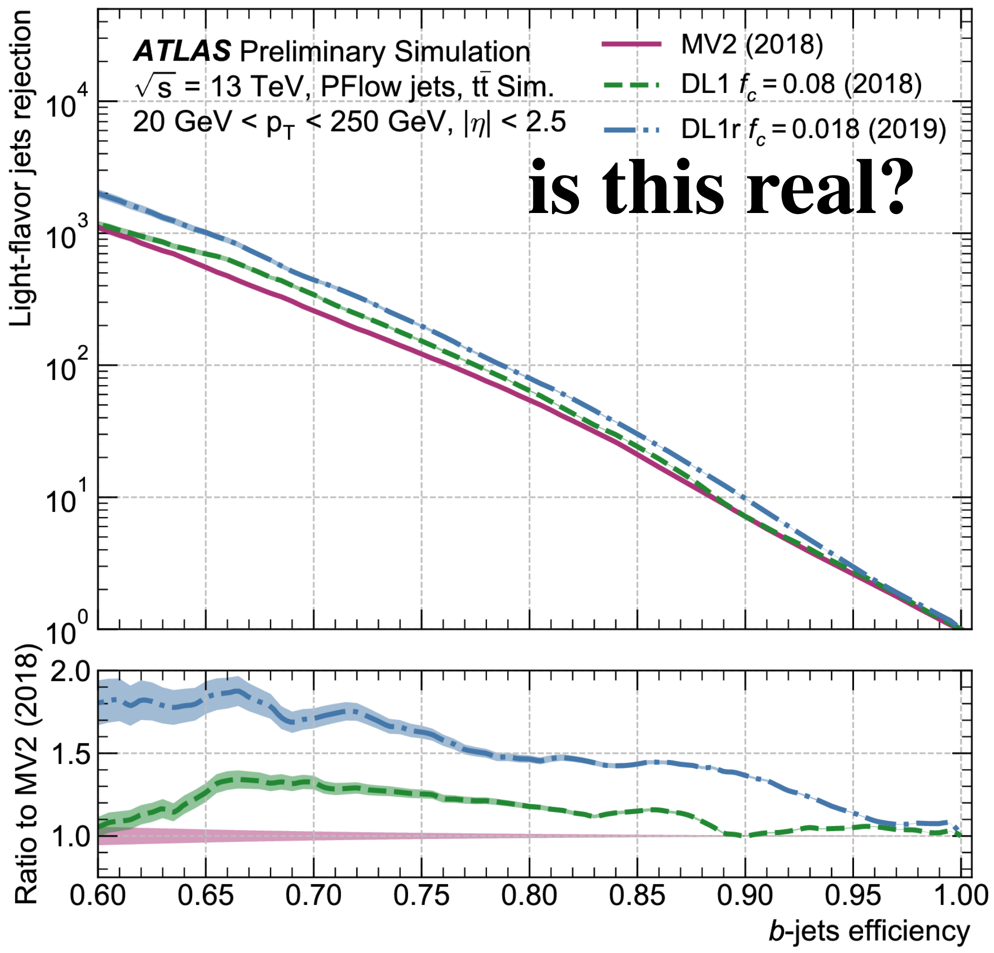
in practice our understanding of collision events and the detector itself is far from perfect.

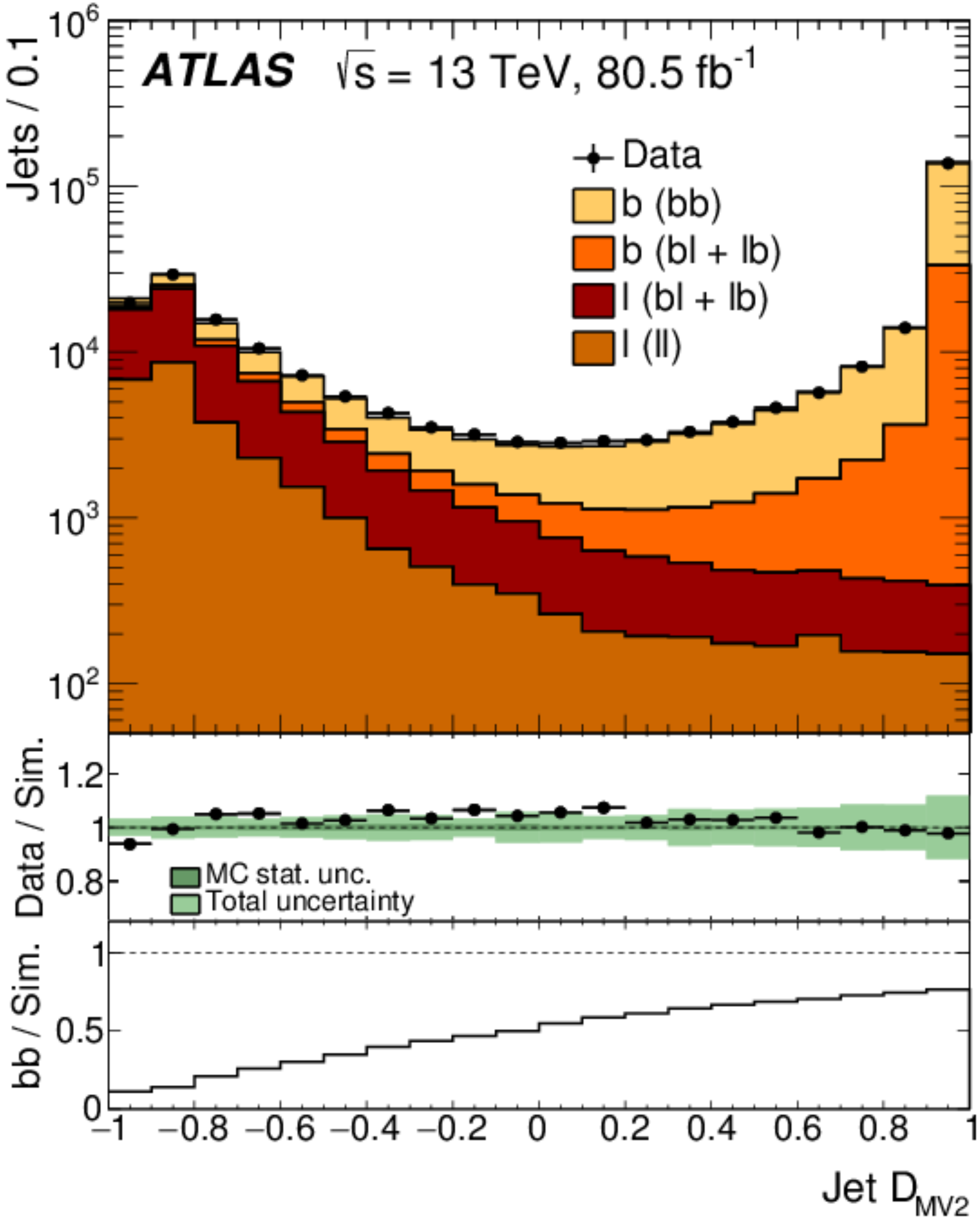
machine learning-based algorithms benefit from correlations between observables

these correlations are very difficult to predict correctly.

there is an enormous effort at DESY to quantify the performance of *b*-tagging in real collision data

and to correct our simulation accordingly





for this to be possible we carefully prune through collisions...

... to find events that are mostly *b*-jets

and we measure the discriminant distribution in the data

for up to 24 combinations of taggers and types of jets

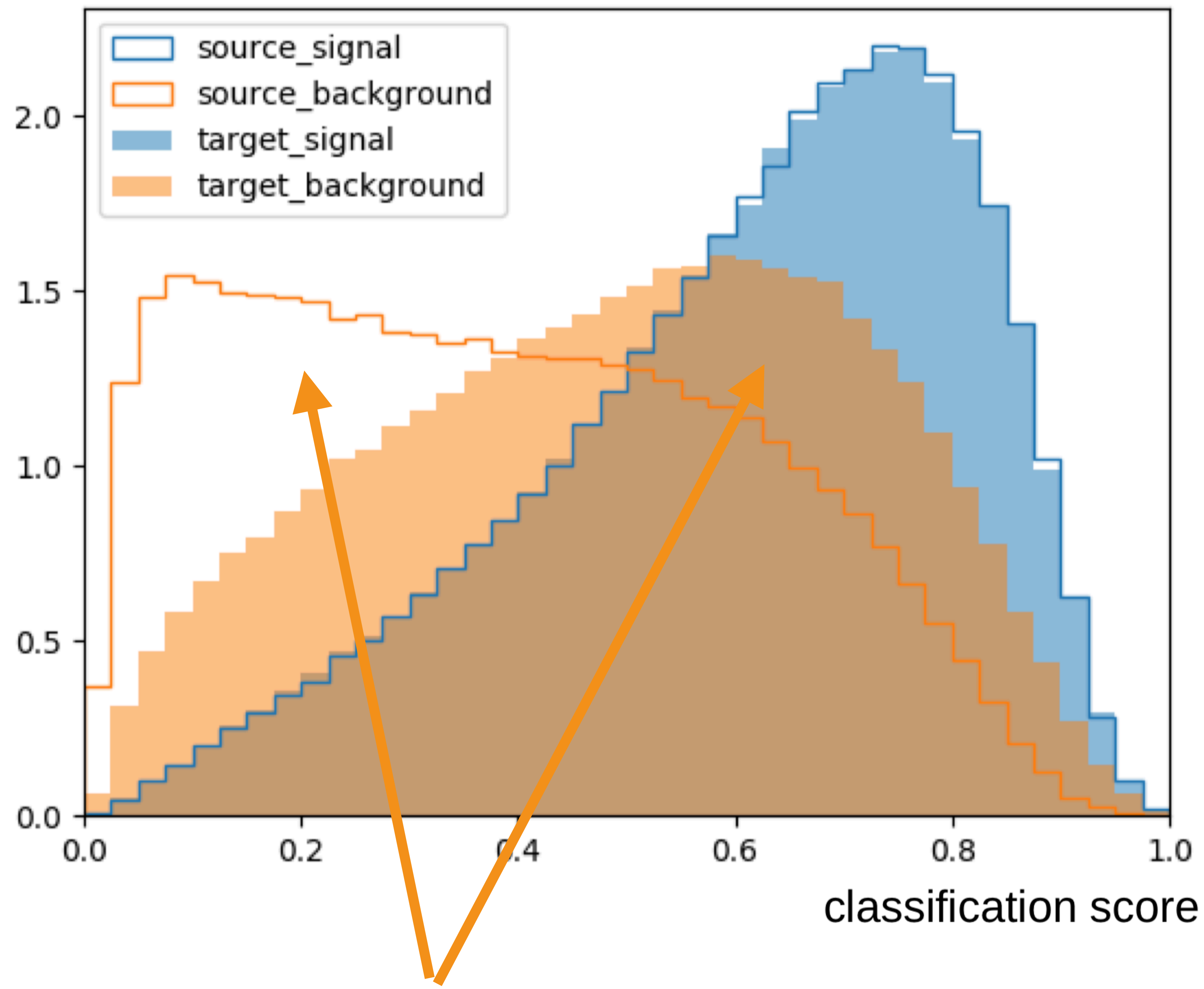
this uncertainty on the "real" performance is often the constraining factor in extracting parameters of interest from the data

Jose's talk at Terascale

this uncertainty on the "real" performance is often the constraining factor in extracting parameters of interest from the data

this is a recurring issue not only for *b*-jet identification

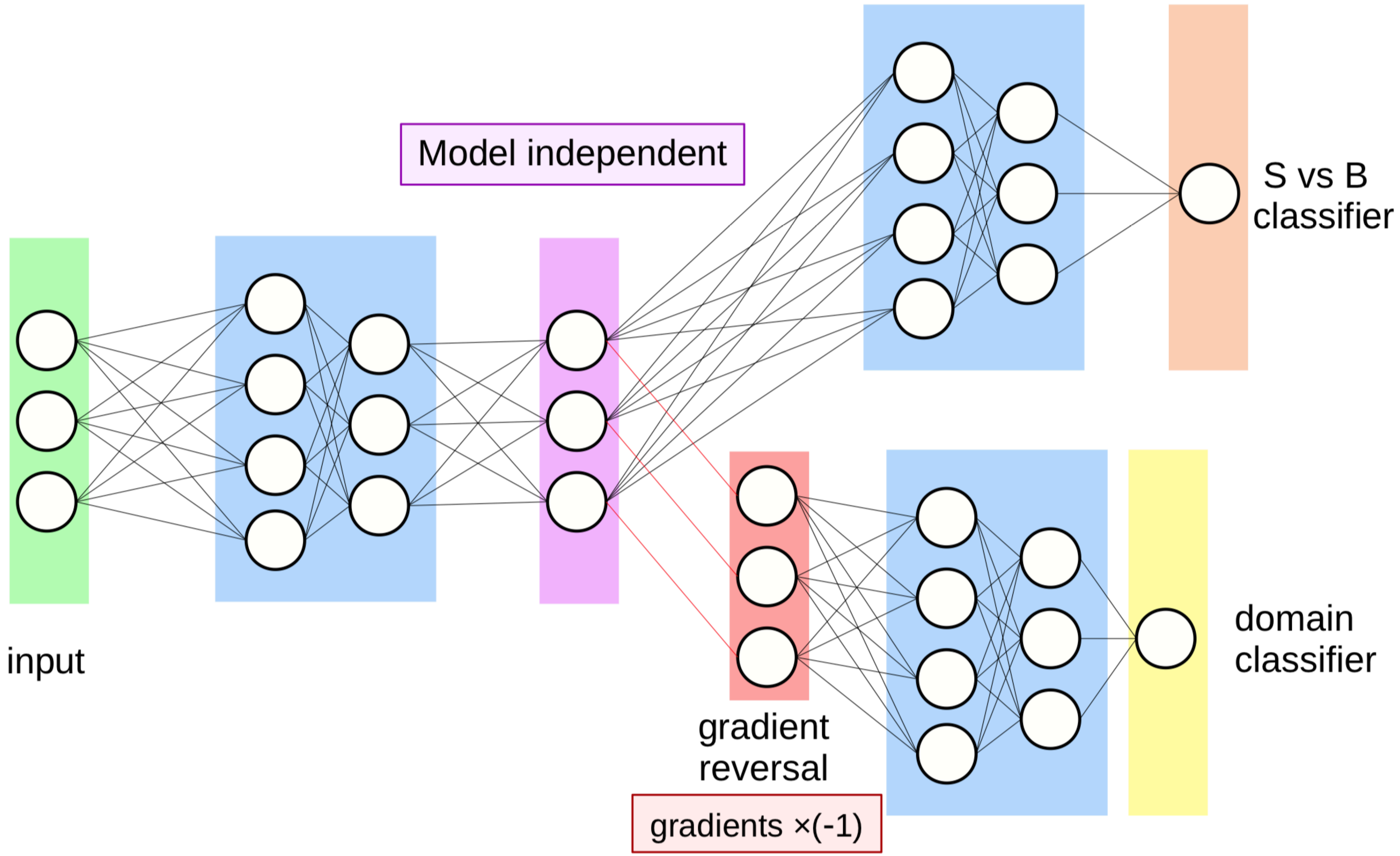
physics analyses using machine learning building up complex discriminants can incur large uncertainties on the algorithm response



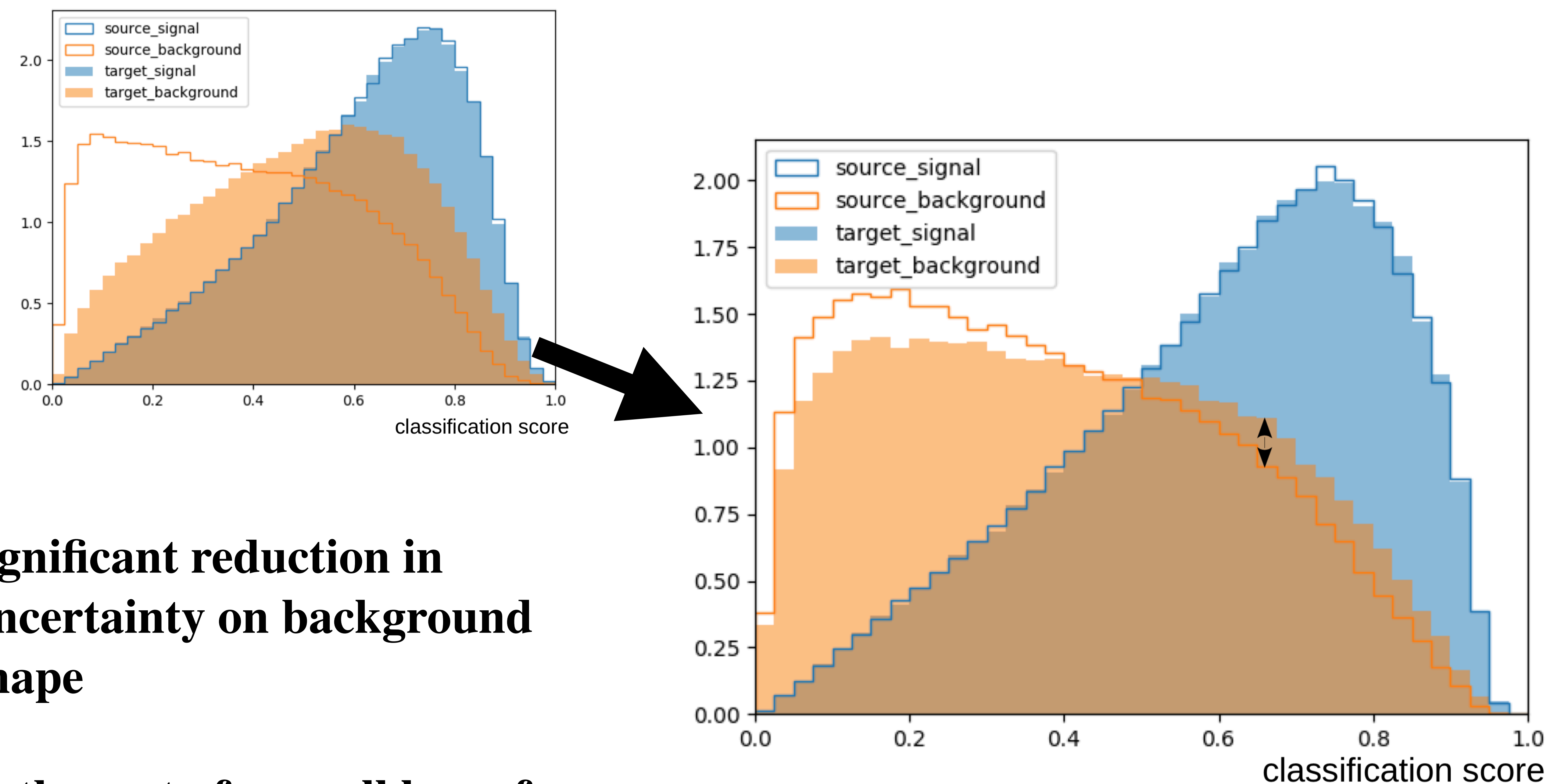
systematic variations on background

Jose's talk at Terascale

add a domain classifier that introduces a loss if it can *identify the systematic variation in question*



Jose's talk at Terascale



significant reduction in uncertainty on background shape

at the cost of a small loss of nominal performance

trained on GPUs on NAF nodes

looks very promising for systematically-limited analyses

- ML continues to undergo rapid proliferation in the LHC community for good reason!
- complex collision events with high performance requirements
- we are still learning how to make *the best* use of ML
- in many cases the constraining factor is actually "how well can we understand the performance in real collision data"?
- active work ongoing to reduce this issue!