IRN Terascale:
A surrogate model for a LLP search
(and how to build your own!)

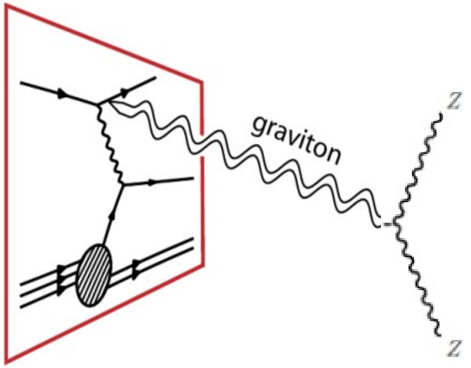
Louie Dartmoor Corpe
&
Abdelhamid Haddad

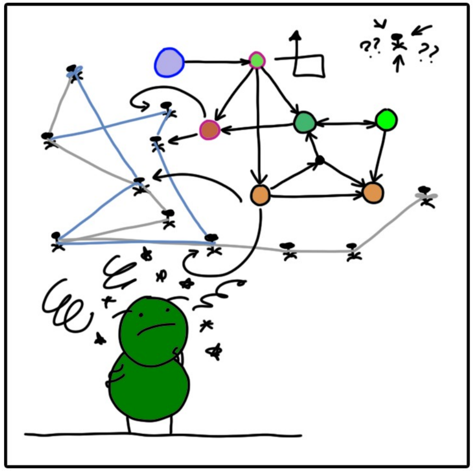
14th of November 2024

General context

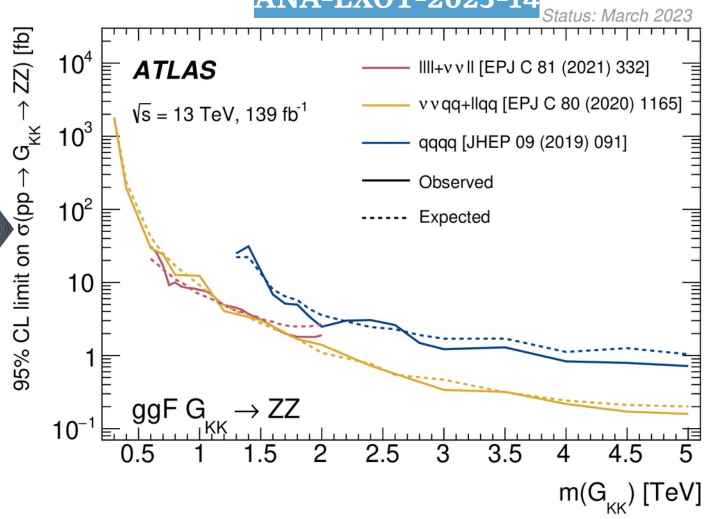
Theory Motivation → Model → … Analysis … → Results

Time Axis





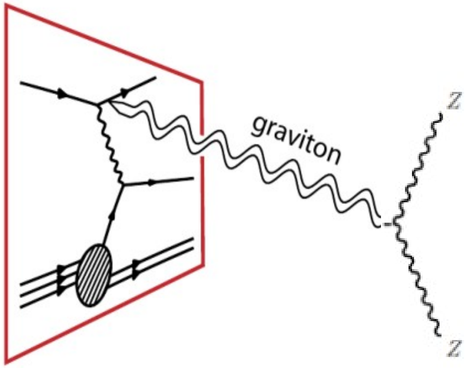
ANA-EXOT-2023-14

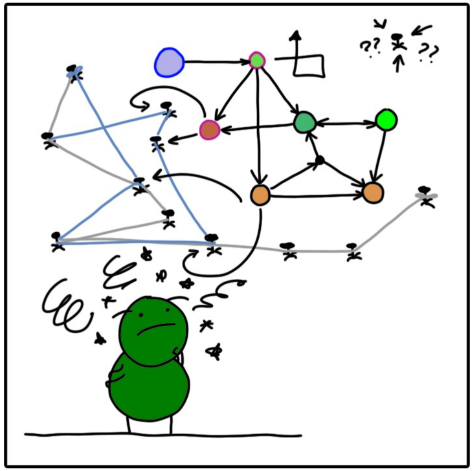


General context

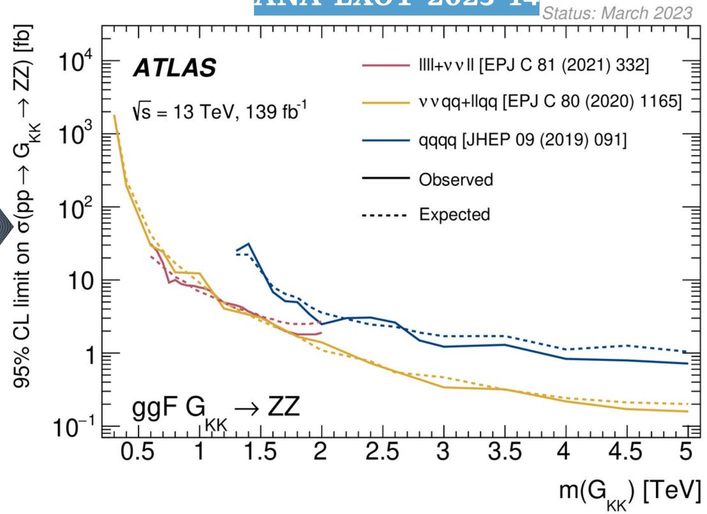
Theory Motivation → Model → ... Analysis ... → Results

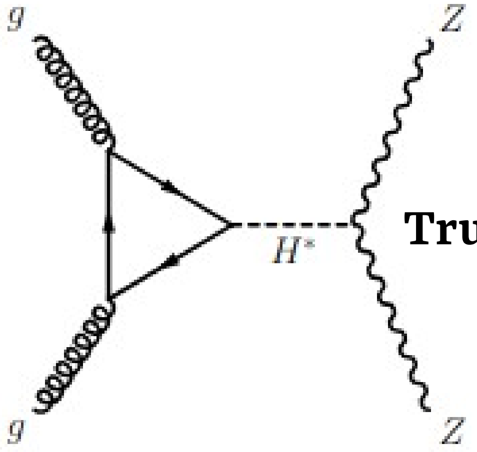
Time Axis





ANA-EXOT-2023-14





At Truth-Level

TURBO TURTLE

THIS WORK!

This is not meant to replace a full analysis, but rather to approximate the previous analysis!

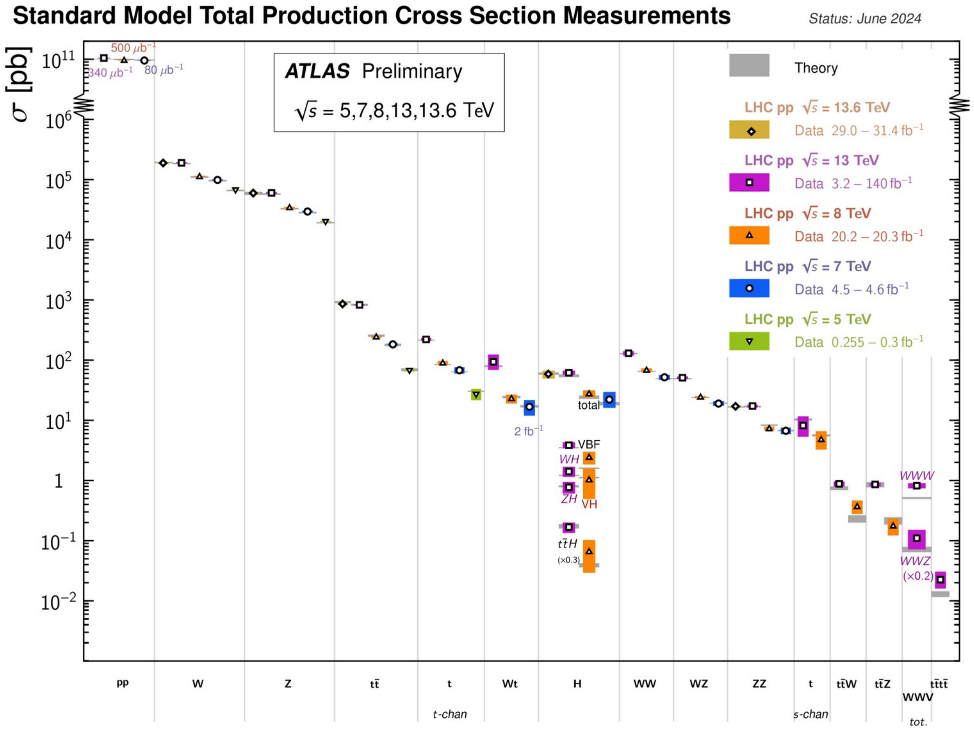
What this work about ?

The main idea is to build a Boosted Decision Tree (BDT) trained on truth-level kinematic information and analysis selection efficiencies → Surrogate Model !

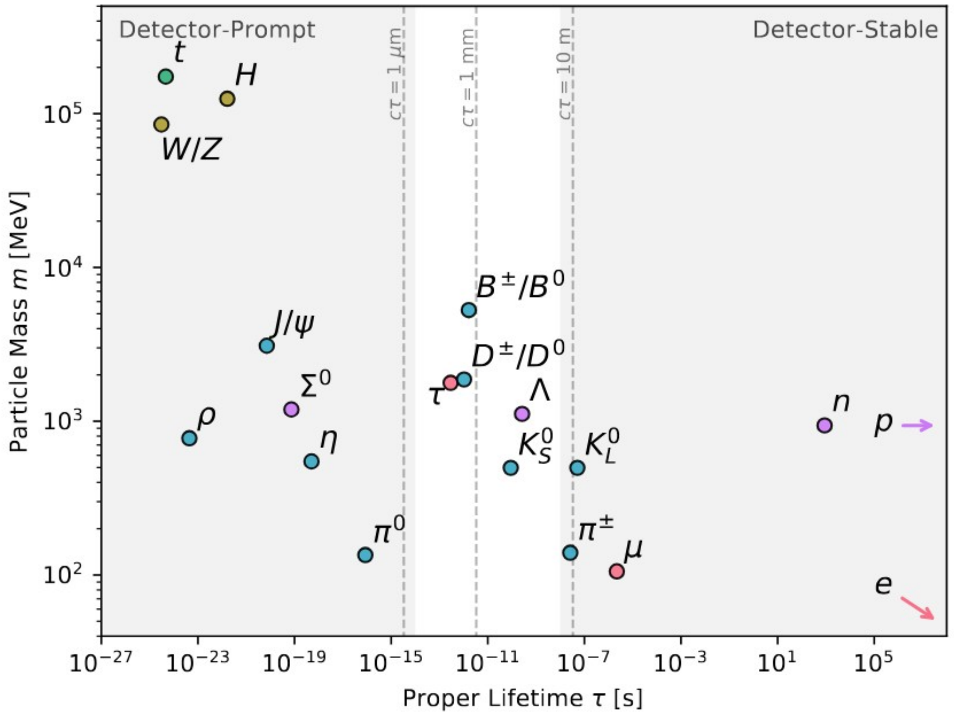
It can predict the probability that new-model events would be selected by an analysis, based solely on truth-level information.

This work was applied to the newly published CalRatio analysis, which searches for long-lived particles within the ATLAS calorimeter.

➢ The **Standard Model** works fine, but **remains incomplete:** Neutrino masses, DM, etc.



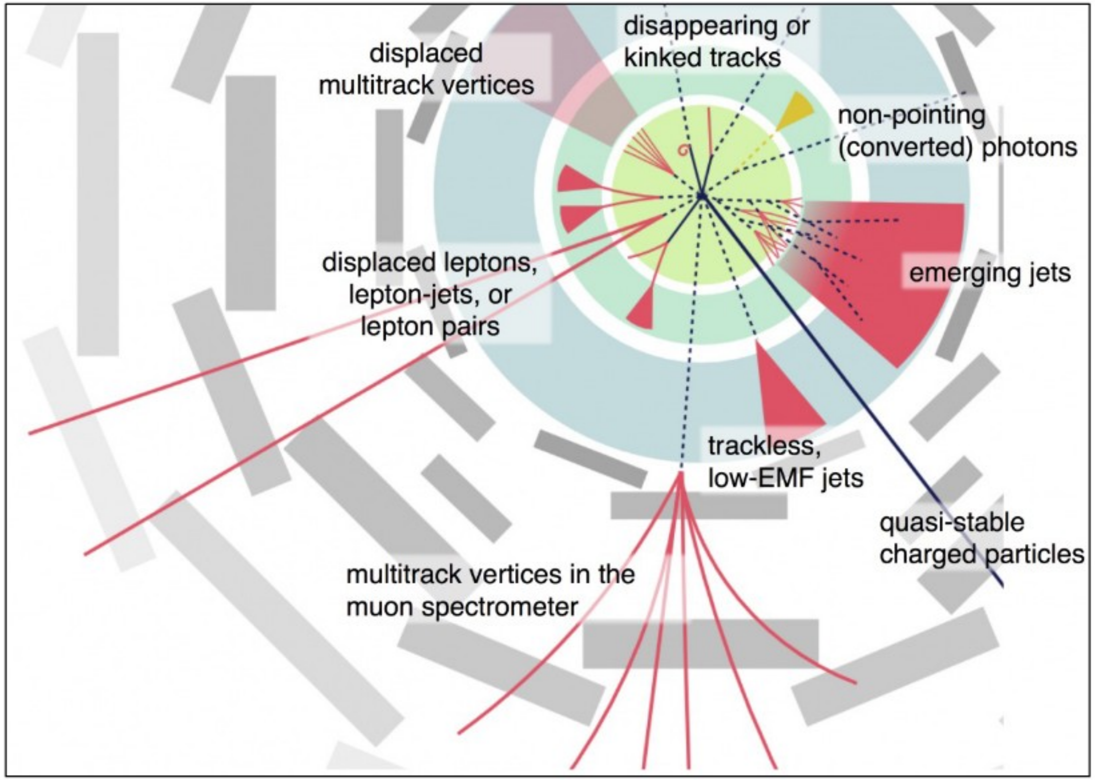
From arxiv:1810.12602



➢ **Some theoretical models** that attempt to address these issues **propose long-lived particles (LLPs),** which can **evade traditional prompt searches**!

LLPs distinctive signatures

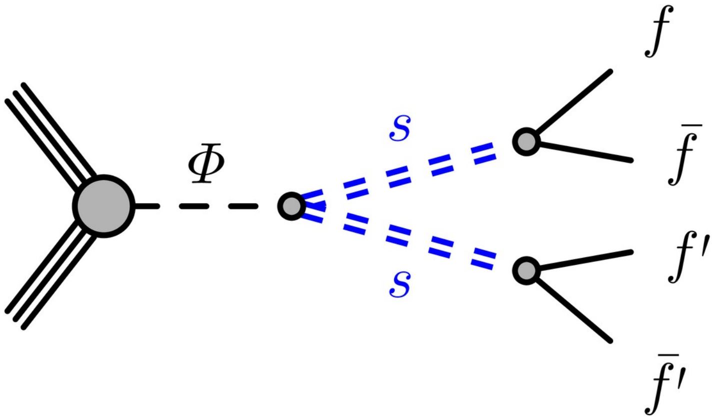
- **Inner Tracker**, displaced Vertex, disappearing or kinked tracks,..
- **Calorimeters**, displaced hadronic jets, trackless hadronic deposits with a low electromagnetic: A CalRatio signature.
- **Muon spectrometer,** displaced vertex, high track multiplicity.

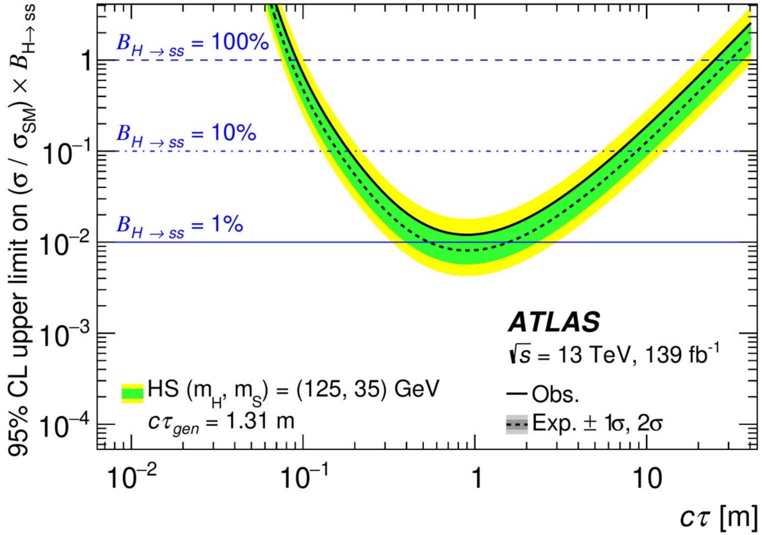


Russell (see, e.g., LLP White Paper 2019)

Analysis Public Page!

- **A Full-Run-2 CalRatio analysis:** (i.e. 2CalRatio)
  - Targeting a **Hidden Sector** (HS) benchmark model:





  - No significant excess was seen, and limits were set on a variety of HS models.

Analysis Public Page

➢ The new analysis is composed of **three channels**:

1

CalRatio + 2J: Same benchmark as the 2CalRatio analysis, but with one LLP reconstructed as two resolved jets (low-boost).

Image: from V.Sanchez

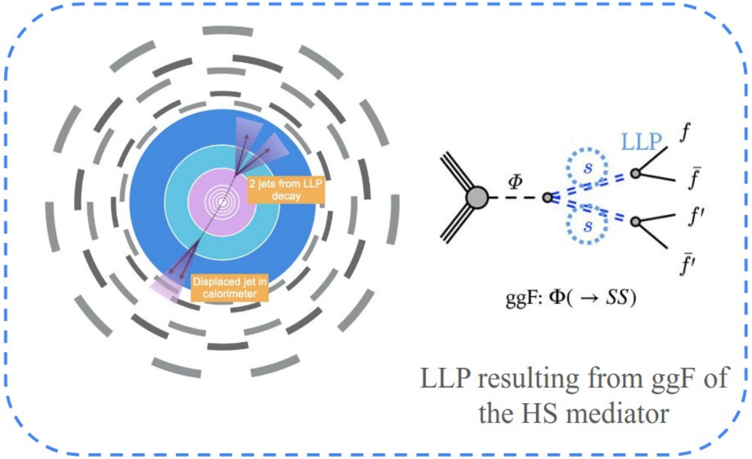
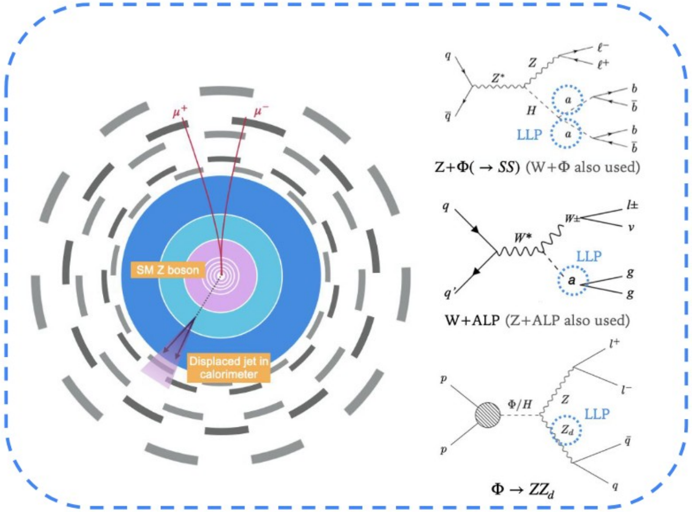


Image: from V.Sanchez



2 & 3

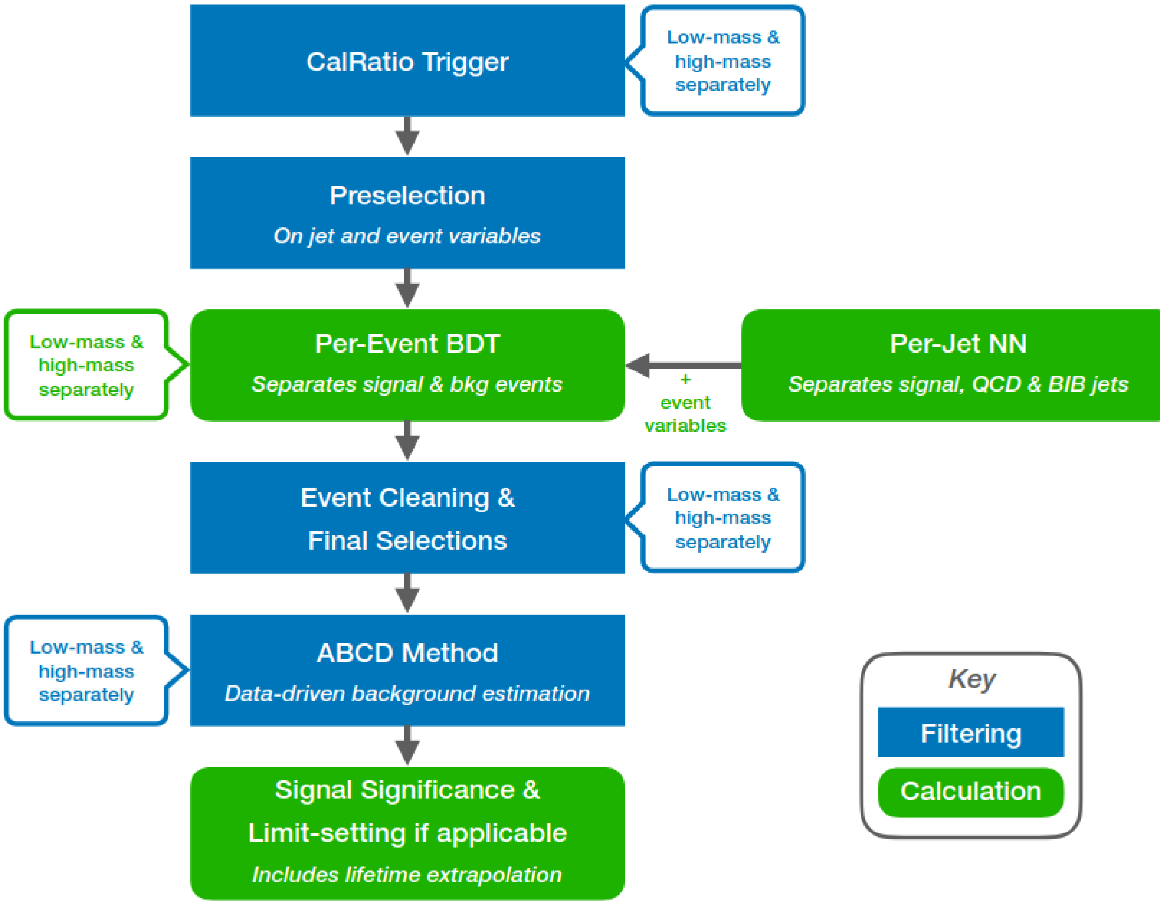
CalRatio + leptons: Improve sensitivity by requiring a single displaced object, access to single-production of LLPs:

CalRatio jet + prompt W/Z

Analysis Public Page

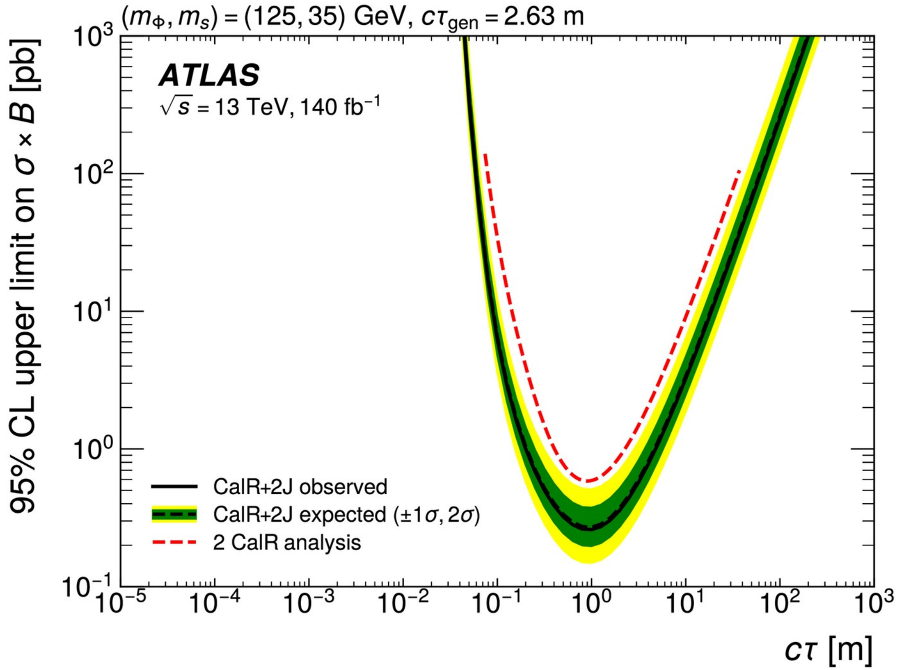
All channels use a per-jet NN followed by a per-event NN or BDT which render the reinterpretation trickier.

- Even though some guidance and advice were provided in Les Houches guide to reusable ML models in LHC analyses

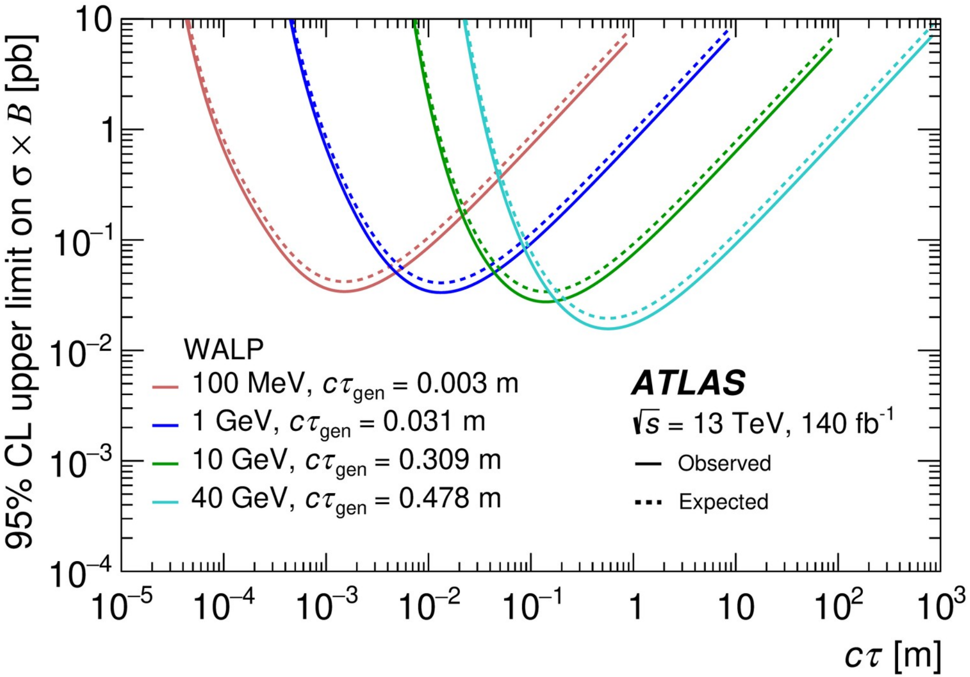


Skipping analysis details that you can find in the ArXiv Paper:
trigger, preselection, selection, event cleaning, and background estimation.

No significant excess is observed,
upper limits on LLP production cross-section times branching ratio have been set.



- Improvement by up to a factor of 3 over the previous 2CalRatio search for mediator masses below 200 GeV, with comparable results above.

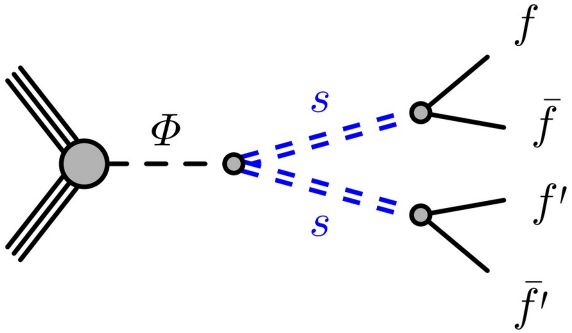


- First ATLAS limits on photo-phobic ALP models, excluding cross-sections above 0.1pb in the 0.1mm - 10m range.

REINTERPRETATION

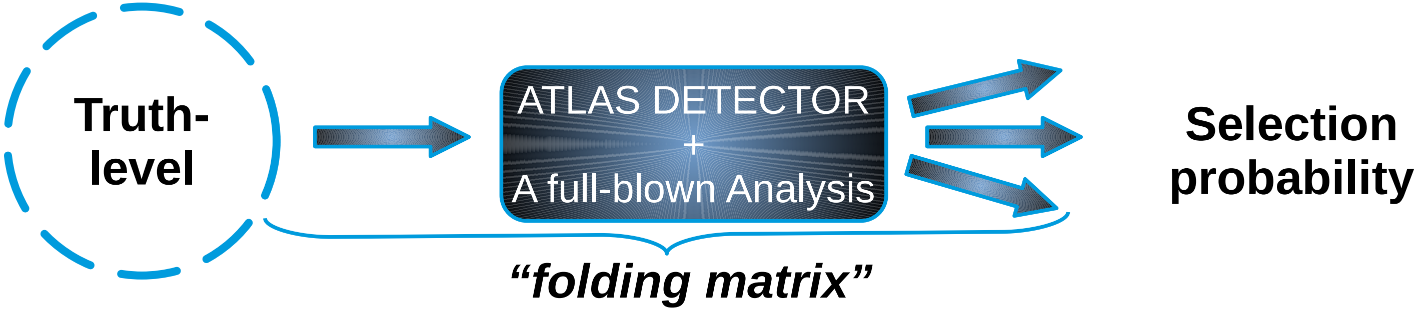
RAMP seminar by Louie Corpe

➔ Event selection probability should depend only on decay properties (position, flavor, pT), not the internal details of the model..



Tentative solution:

1. Map selection probability in Region A using LLP decay kinematics.
2. Calculate efficiency by summing selected event probabilities over total events.



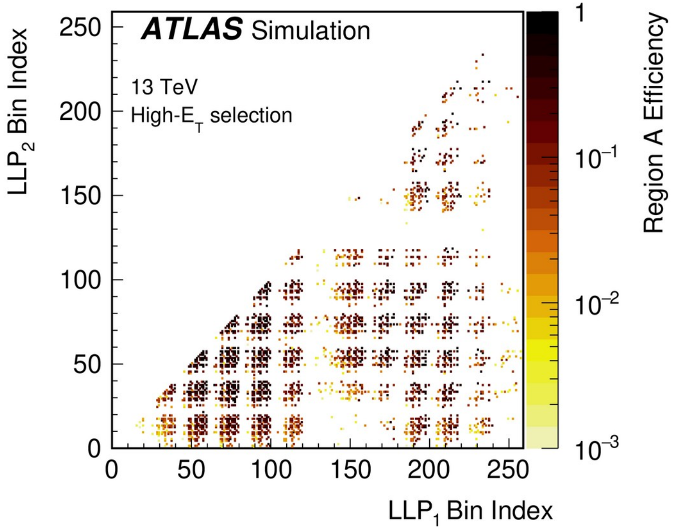
2CalRatio analysis reinterpretation

Decay flavours: 4 bins
c, b, t, τ

LLP pT: 5 bins
between 0, 50, 100, 200, 400, ∞

Decay positions: 13 bins
Lxy: 0, 1.5, 2, 2.5, 3, 3.5, 3.9, ∞
Lz: 0, 3.6, 4.2, 4.8, 5.5, 6, ∞

By folding the decay particles' properties into a "**Bin Index**," the **final efficiencies** for an event are **represented on an efficiency map.**



$$\begin{aligned}\text{Bin Index} = &\ \text{decay position bin index} \times (\text{number of pT bins} \\ &\times \text{number of decay type bins}) \\ &+ \text{pT bin index} \times (\text{number of decay type bin}) \\ &+ \text{decay type bin index}\end{aligned}$$

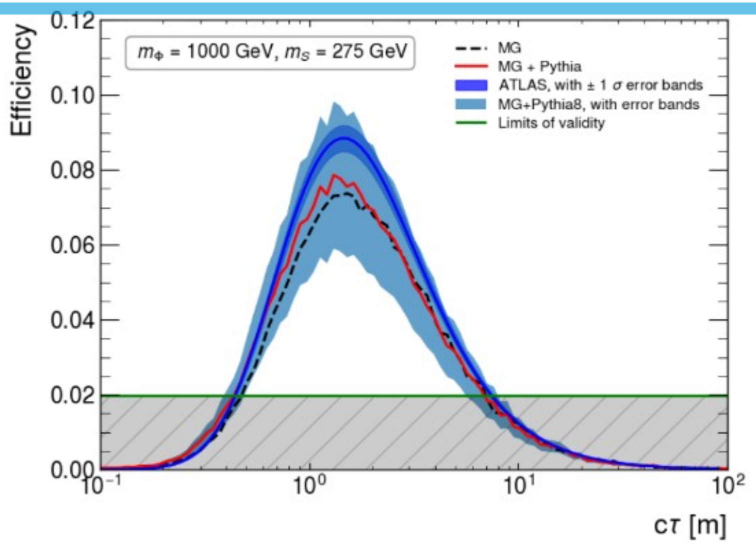
2CalRatio analysis reinterpretation

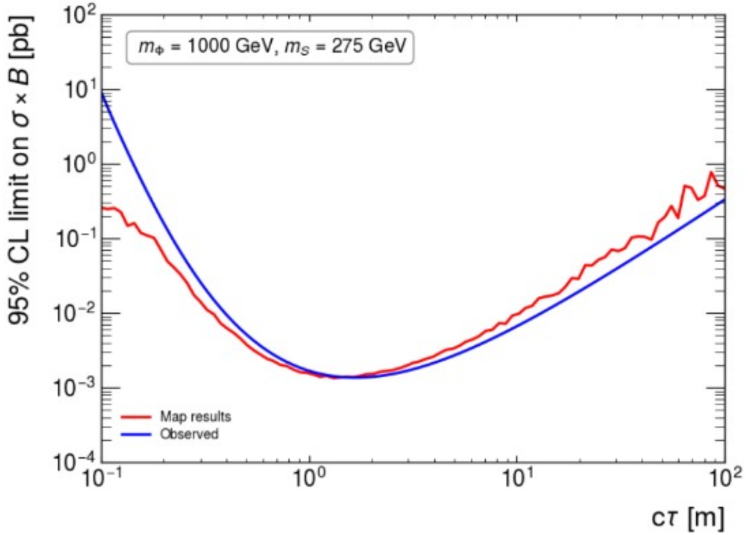
Results from a master project:
GitHub:ThomasChehab/recastingCodes

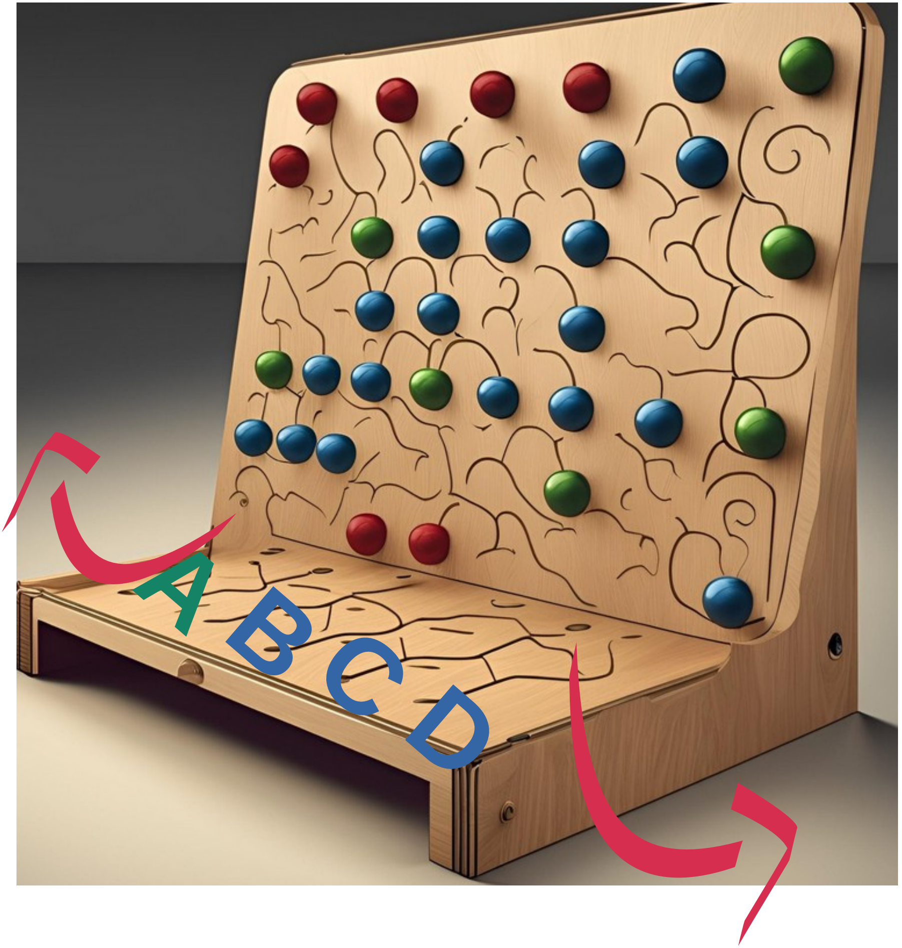
➢ The method showed quite **decent performances**, with a few exceptions (e.g. low masses).

➢ **Limitations:**

- Limited number of variables,
- Manual binning,
- Time-consuming process,
- Challenging to adapt to different analyses.
- Difficult to explain what the map represents.







- For the CalRatio+X analysis, **we considered potential improvements**.
- A **BDT is trained on truth-level information** to classify events within the ABCD plane or as unselected in the analysis.
- The trained BDT can then **estimate the likelihood that events** from a new model **would have been captured by the prior analysis**.

What Have We Gained from This Approach?

Previous efficiency map

- Limited number of variables,
- Manual binning,
- Time-consuming process,
- Challenging to adapt to different analyses.
- Difficult to explain what the map represents.

→

New BDT method

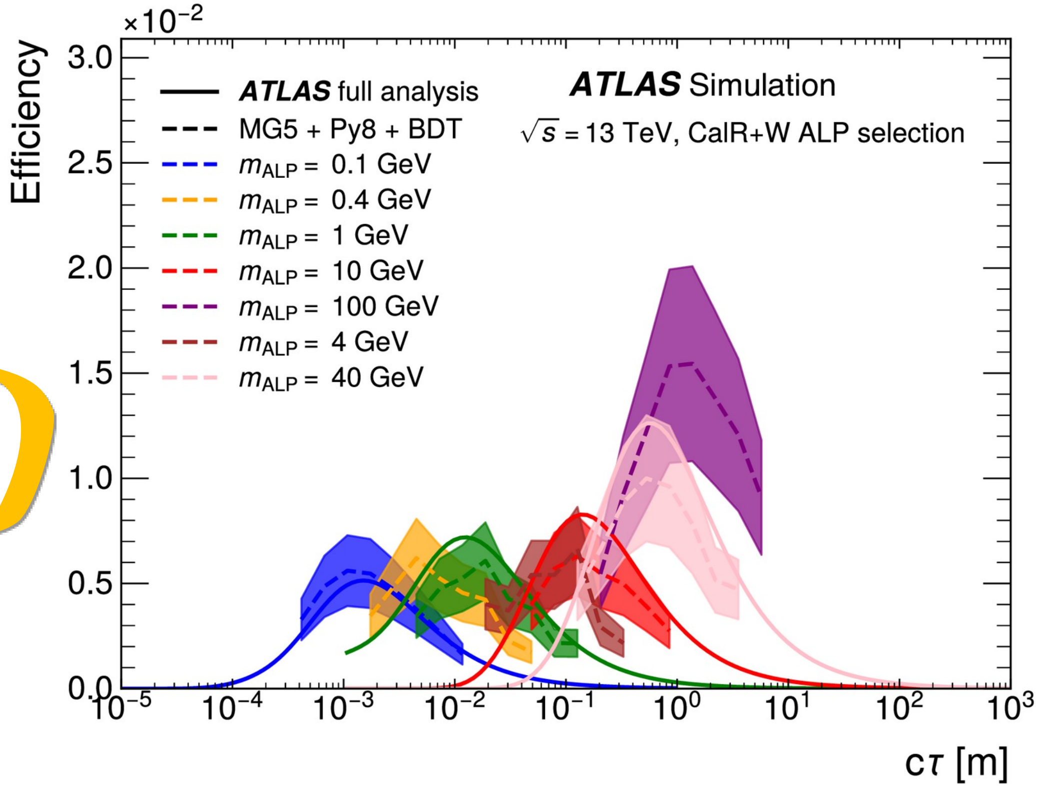
- ✔ As much as we want
- ✔ No need
- ✔ From 4h to 30min
- ✔ Easier: Input files and variables
- ✔ Largely used in the community

BDT reinterpretation: Results

- To properly assess the performance of this method, we utilized **standalone MadGraph** simulations of the same models as a proof of concept.

- As a result, we obtained the efficiencies of these new events (dashed), closely matching the real analysis results (solid) !

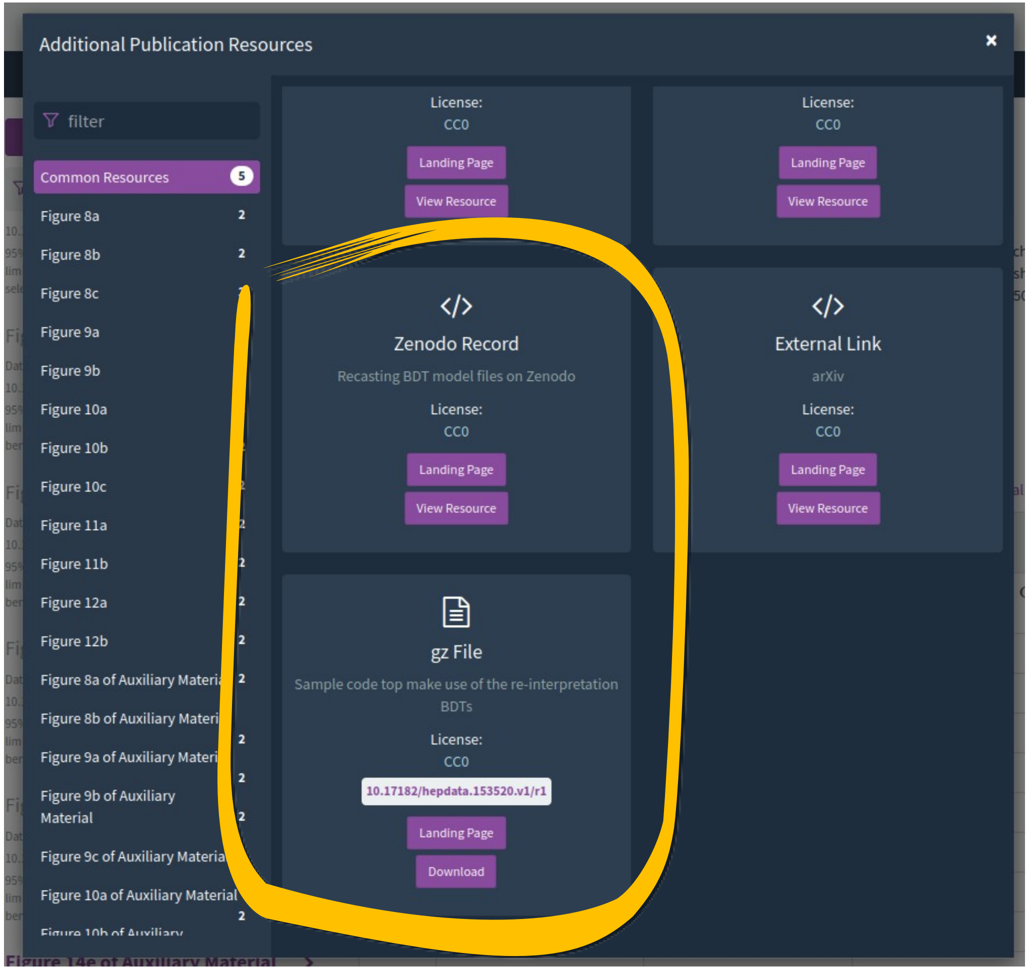
- Pleasingly, **we were also able to reasonably estimate the efficiencies for events with new masses.**



Note: The BDT-estimated efficiencies are assigned an uncertainty of 25%.

How YOU can benefit from this work!

HEPData page



➢ To test your own model events, you will need the **pickle files (BDT weights) and sample code** that we have **<u>made available for you on HEPData</u>**.

How YOU can benefit from this work!

➢ A small **setup script** is also available along with those slides.

```
main_folder="ATLAS-EXOT-2022-04_CalRatio+X_Reinterpretation"
mkdir -p "$main_folder"

mkdir -p "$main_folder/BDTs"
mkdir -p "$main_folder/Example_Samples"

curl -OJLH "Accept: application/x-tar"
https://doi.org/10.17182/hepdata.153520.v1/r1

tar -xzvf reinterpretationMaterials.tar.gz

mv reinterpretationMaterials/*.csv "$main_folder/Example_Samples"

mv reinterpretationMaterials/*.py "$main_folder"

rm -f reinterpretationMaterials.tar.gz
rm -rf reinterpretationMaterials
```

```
ATLAS-EXOT-2022-04_CalRatio+X_Reinterpretation
  > BDTs
  > Example_Samples
  example.py
  recast_bdts.py
```

```
wget -O files-archive.zip https://zenodo.org/api/records/12957031/files-archive
unzip files-archive.zip -d files-archive

mv files-archive/* "$main_folder/BDTs"

rm files-archive.zip
rm -rf files-archive

rm "$main_folder/BDTs/reinterpretationMaterials.tar.gz"

for tarfile in "$main_folder/BDTs"/*.tar.gz; do
    foldername="${tarfile%.tar.gz}"  # Remove the .tar.gz extension
    mkdir "$foldername"

    echo -e "\e[36mExtracting \"$tarfile\" to \"$foldername\"...\e[0m"
    tar -xzvf "$tarfile" -C "$foldername"

    rm "$tarfile"
done
```

```
> CR+2J
> WALP
> WHS_highET
> WHS_lowET
> ZHS_highET
> ZHS_lowET
```

```
ggH60_S5_gg_py8_ct5.32.csv
ggH125_S55_ctau5p32_gg_py8.csv
ggH200_S50_gg_py8_ct1.25.csv
ggH400_S100_ctau1p6m_gg_py8.csv
ggH600_S275_gg_py8_ct4.288.csv
```

```
models/WALP_features.txt
models/WALP_model.pkl
models/WALP_scaler_mean.npy
models/WALP_scaler_std.npy
```

How YOU can benefit from this work!

- Adjust the paths for both the BDT pickle and model files as specified below:

recast_bdts.py

```
def load_model(sel, modelDir = "models"):
    scaler_mean =  np.load(f"./BDTs/{sel}/models/{sel}_scaler_mean.npy")
    scaler_std =  np.load(f"./BDTs/{sel}/models/{sel}_scaler_std.npy")
    with open(f"./BDTs/{sel}/models/{sel}_features.txt") as g: var=eval(g.read())
    f = open(f'./BDTs/{sel}/models/{sel}_model.pkl', 'rb')
    clf = pickle.load(f)
    return scaler_mean, scaler_std, var, clf
```

example.py

```
# The CSV was obtained from parsing HEPMC3 output for a ggH sample
for sample in ['./Example_Samples/ggH125_S55_ctau5p32_gg_py8.csv', './Example_Samples/ggH600_S275_gg_py8_ct4.288.csv']:
    print("===")
```

How **YOU** can benefit from this work!

- The provided "**example.py**" script is designed to give you insights into the efficiency of the method using the available standalone generated samples:

```
===
Sample ./Example_Samples/ggH125_S55_ctau5p32_gg_py8.csv has efficiency in the Region A of CR+2J selection of 0.09 %
===
Sample ./Example_Samples/ggH600_S275_gg_py8_ct4.288.csv has efficiency in the Region A of CR+2J selection of 2.46 %
===
```

	CalR+2J selection	Main dataset	BIB dataset	HS $m_\Phi = 125$ GeV $m_S = 55$ GeV	HS $m_\Phi = 600$ GeV $m_S = 275$ GeV
Preselection	CalRatio triggers			1.4%	11%
	≥ 3 clean jets			0.60%	7.7%
	$\sum \Delta R_{\text{min}} > 0.5$	5,738,136	446,794	0.59%	7.6%
Event cleaning	Trigger matching	2,068,592	154,986	0.53%	6.5%
	-3 ns $< t <$ 15 ns	2,609,223	99,398	0.51%	5.8%
	$\log_{10}(E_{\text{H}}/E_{\text{EM}}) > -1.5$	2,289,758	89,380	0.46%	5.1%
	$\lvert\eta\rvert \notin (1.45, 1.55)$	2,068,592	80,555	0.41%	4.5%
	$\text{NN}_{\text{CalR+2J}} \geq 3$	30,097	408	0.35%	4.4%
Region A	$\sum \Delta R_{\text{min}} \geq 0.71$, $\text{NN}_{\text{CalR+2J}} \geq 7.61$	92	2	0.10%	2.8%
Region B	$\sum \Delta R_{\text{min}} < 0.71$, $\text{NN}_{\text{CalR+2J}} \geq 7.61$	18	1	0.00%	0.01%
Region C	$\sum \Delta R_{\text{min}} \geq 0.71$, $\text{NN}_{\text{CalR+2J}} < 7.61$	25213	328	0.24%	1.6%
Region D	$\sum \Delta R_{\text{min}} < 0.71$, $\text{NN}_{\text{CalR+2J}} < 7.61$	4774	77	0.01%	0.04%

Aux. Mat.

What you need to do after having made the setup and testing it..

(1) **Pick your preferred theory model** predicting LLPs decaying within the ATLAS calorimeter.

(2) **Simulate it** and **save the kinematic variables** required as features for the BDTs (see the *_features.txt files for the required information), and save them **into a pandas dataframe.**

(3) **Pass the dataframe to the BDT handler** as in the example, and get your region A yield.

Remark: In some cases, **the selection works for cases with one or two LLPs**. In that case, the BDT has features for both LLPs, **but if your model predicts just one, you can set LLP2==LLP1**.

Theorist

Feedback are welcome to enhance this method's application and further refinement.

Experimentalist

Contact us for guidance on creating similar reinterpretation material, and to ensure your analysis remains impactful for future research.

Summary

- **A reinterpretation method based on Boosted Decision Trees has been developed** and tested within the full **CalRatio + X ATLAS analysis**, broadening its scope for testing new theoretical models.

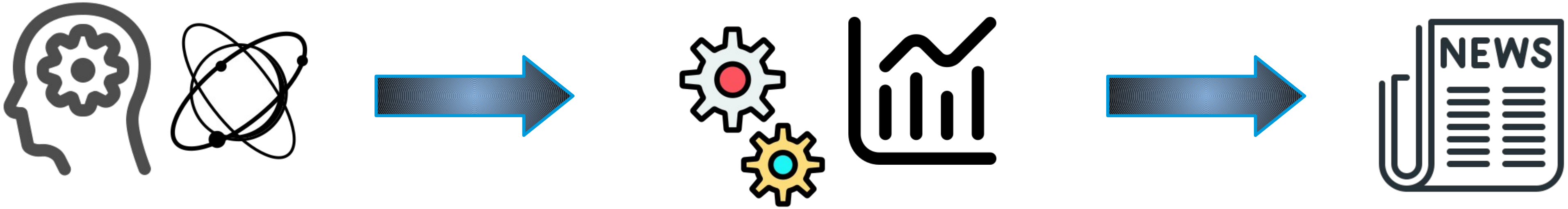
- **The results appear conclusive enough to be generalized, offering encouraging prospects for the future. Especially that it is easy to produce** on the experimental side, and to use on theory side.

BACKUP

1) **Theorists develop models** that extend or modify the Standard Model of particle physics.
   - ➢ Often predict **new particles and/or interactions** that haven't been observed yet.

2) **Experimentalists** create detailed plans to **test the predicted models**.
   - ➢ Simulation, Data Collection, Event Reconstruction, Background Estimation, Machine learning trainings & Statistical Analysis,...

3) **The analysis results are compared against the predictions of the BSM model**.
   - ➢ If the data matches the predictions, **the model is supported**.
   - ➢ If not, the model may need revision or even **rejected**.



1) **Theorists develop models** that exten y the Standard Model of particle physics.
- Often predict **new particles and/or s** that haven't been observed yet.

2) **Experimentalists** create detailed p **predicted models**.
- Simulation, Data Collection, Even ckground Estimation, Machine learning trainings & Statistical Ana sis,...

3) **The analysis results are compa SM model**.
- If the data mat
- If not, the m

This process may seem inefficient since we don't know where new physics is hiding, and we want to avoid testing models that aren't truly new but have been tested indirect in previous analysis, as it would be a waste of time and resources.

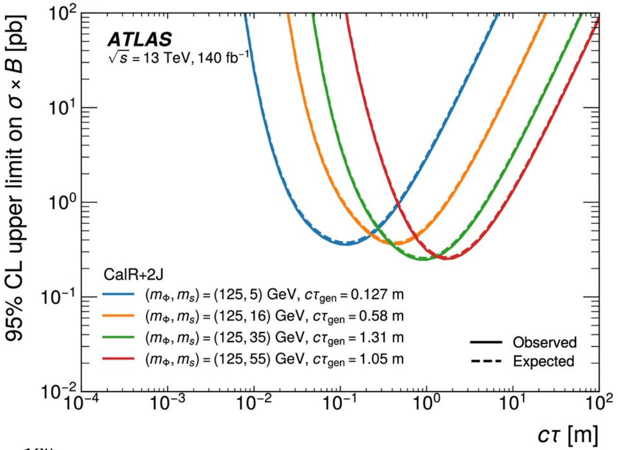
Reinterpretation in Experimental Particle Physics, A Pathway to New Discoveries

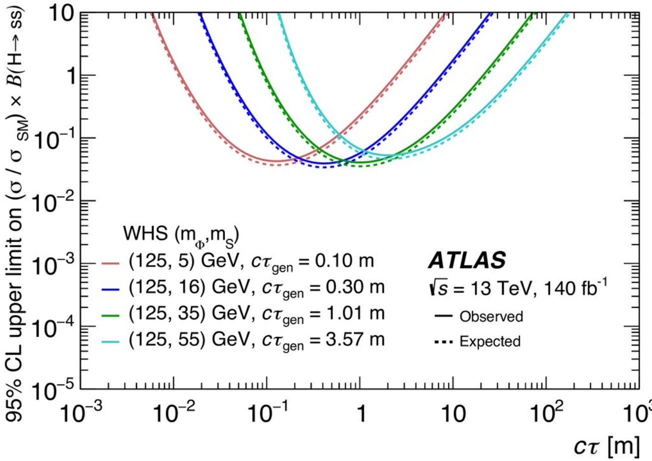
Reinterpretation involves using existing analysis to test new theoretical models.

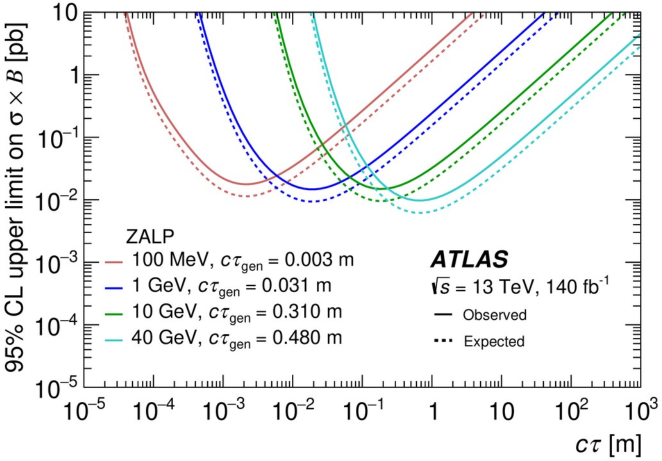
- **For Theorists**, it allows to test new models against existing analysis, accelerating the validation or refinement of their hypotheses without waiting for new analysis.
- **For Experimentalists**, it maximizes the utility of collected data and already tuned analysis.

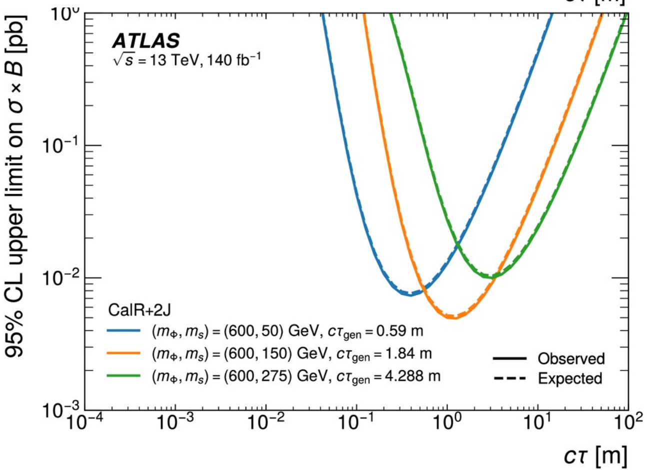
Skipping analysis details that you can find in the ArXiv Paper:
trigger, preselection, selection, event cleaning, and background estimation.

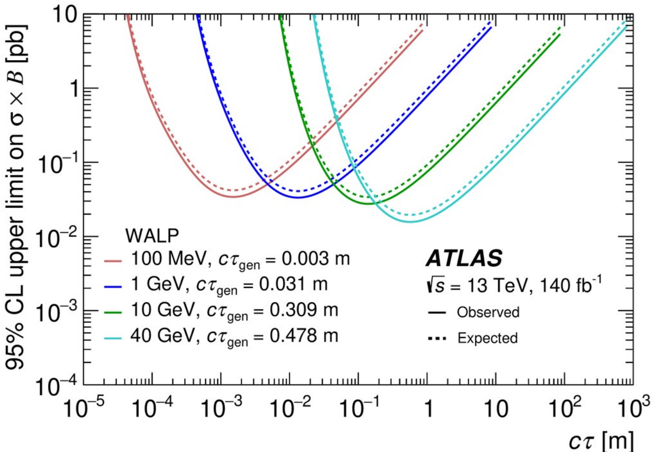
No significant excess is observed in any regions, upper limits on LLP production cross-section times branching ratio are extracted using the CLs method have been set.

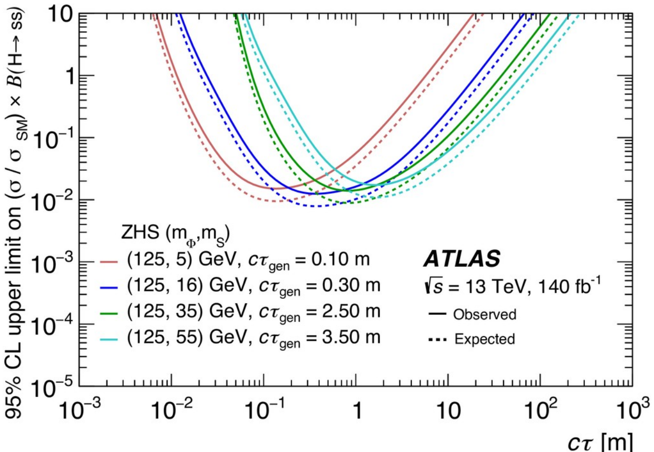












One could have a look at the efficiencies per sample!

	Efficiency in A	Efficiency in B	Efficiency in C	Efficiency in D
testing input file selection_mALP0p1GeV_W_selLLP_w.root				
Real	0.49%	0.15%	0.22%	0.06%
BDT	0.46%	0.13%	0.19%	0.06%
Relative diff	*-6.1%*	*-13.3%*	*-13.6%*	*0.0%*
testing input file selection_mALP10GeV_W_highctau_selLLP_w.root				
Real	0.39%	0.12%	0.15%	0.05%
BDT	0.38%	0.11%	0.14%	0.05%
Relative diff	*-2.6%*	*-8.3%*	*-6.7%*	*0.0%*
testing input file selection_mALP10GeV_W_selLLP_w.root				
Real	0.56%	0.18%	0.18%	0.06%
BDT	0.53%	0.16%	0.18%	0.06%
Relative diff	*-5.4%*	*-11.1%*	*0.0%*	*0.0%*
testing input file selection_mALP1GeV_W_selLLP_w.root				
Real	0.48%	0.15%	0.23%	0.07%
BDT	0.46%	0.15%	0.19%	0.06%
Relative diff	*-4.2%*	*0.0%*	*-17.4%*	*-14.3%*
testing input file selection_mALP40GeV_W_selLLP_w.root				
Real	1.09%	0.32%	0.17%	0.06%
BDT	1.01%	0.29%	0.17%	0.06%
Relative diff	*-7.3%*	*-9.4%*	*0.0%*	*0.0%*