Accelerating the search for mass bumps using the Data-Directed Paradigm

Jean-François Arguin
Georges Azuelos
Émile Baril
Fannie Bilodeau
Ali El Moussaouy
Bruna Pascual
Muhammad Usman

Maryna Borysova
Shikma Bressler
Michael Chu
Etienne Dreyer
Nilotpal Kakati
Amit Shkuri

Samuel Calvet
Julien Donini
Eva Mayer

Université de Montréal

מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

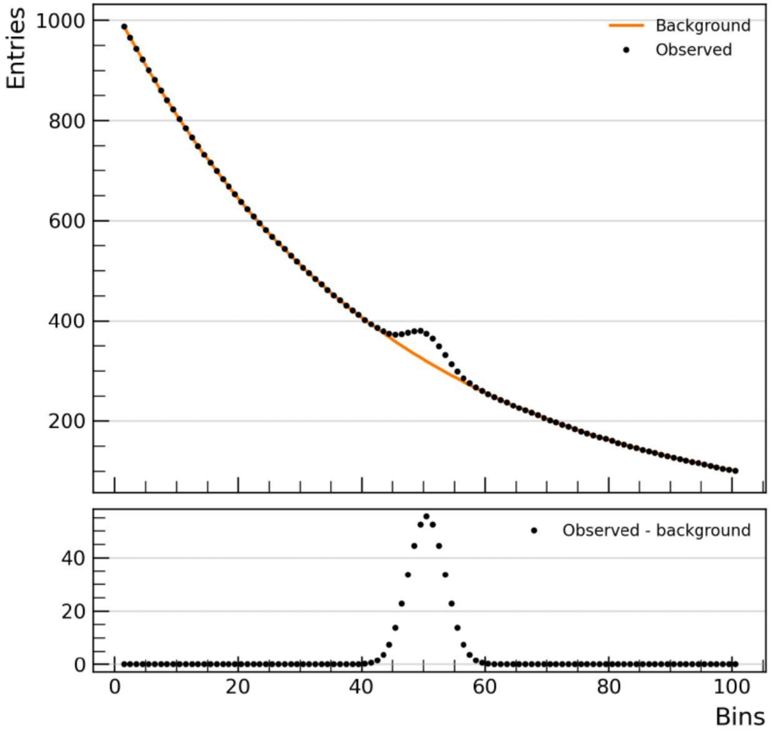
UNIVERSITÉ Clermont Auvergne
LPC Laboratoire de Physique de Clermont

Motivation

★ So far, none of the efforts to find new physics at LHC have succeeded
★ Traditional approach: confirm/exclude theory predictions in **small observable space**
★ Change of approach: scan **all data** for anomalies to motivate specific searches

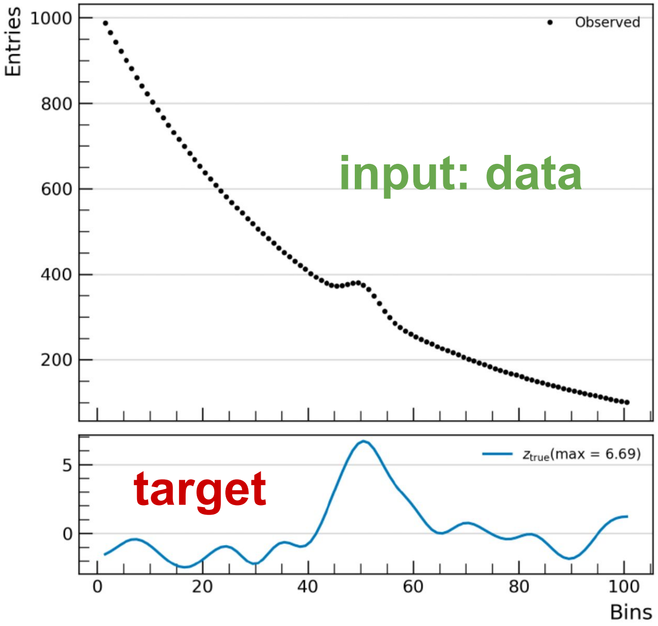
Data-Directed Paradigm

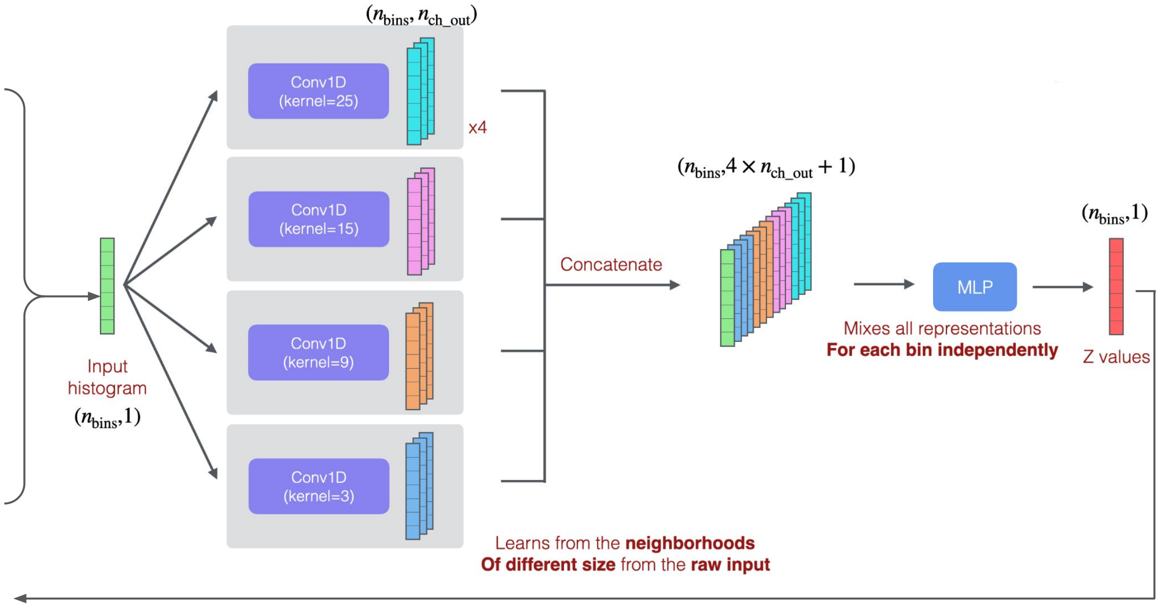
- ★ Smoothly falling mass distributions
- ★ New physics could cluster in data as **bumps** in mass histograms
- ★ Training of a network to find bumps
    - ○ Predicts the **z value**: statistical significance of an excess at each mass value



Predicting the statistical significance using a neural network

ZNet3: supervised learning: both z_{true} and signal width are given in training





Training data

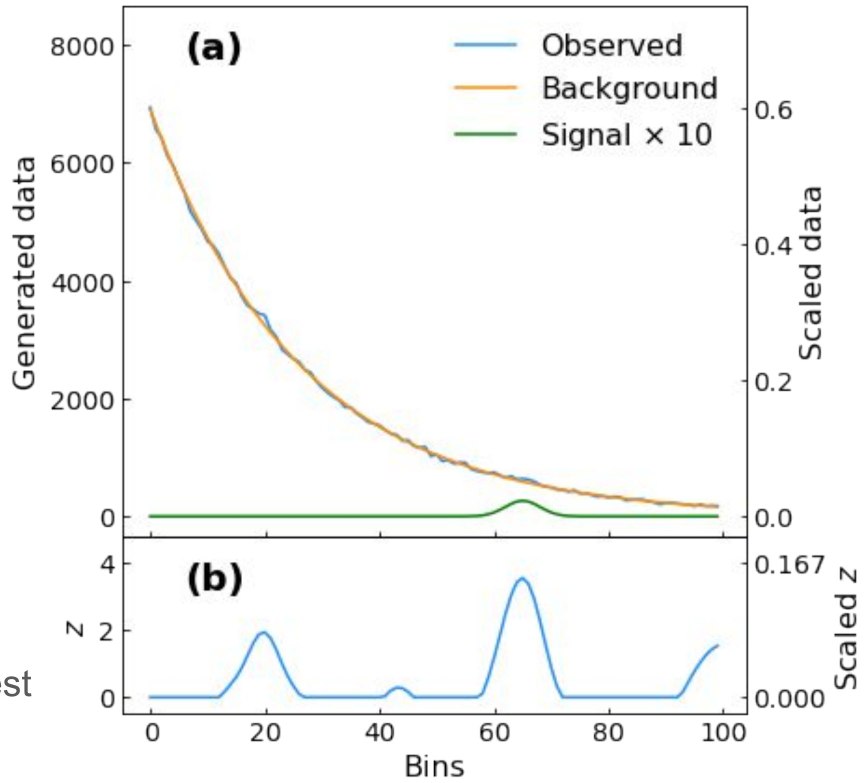
Background distributions:

★ analytical functions and/or simulations

$$be^{-ax}, \quad ax+b, \quad \frac{1}{ax}+b, \quad \frac{1}{ax^2}+b, \quad \frac{1}{ax^3}+b,$$
$$\frac{1}{ax^4}+b, \quad a\left(x-x_2\right)^2+y_2, \quad -a\cdot\ln\left(x\right)+b,$$
$$\left(y_1-y_2\right)\cos\left(a\left(x-b\right)\right)+y_2, \quad \cosh\left(a\left(x-x_2\right)\right)+b$$

Adding the signal:

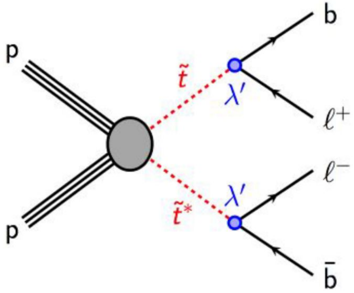
★ select **background**
★ add **gaussian signal**
★ add random fluctuations
★ calculate **true significance** with likelihood-ratio test

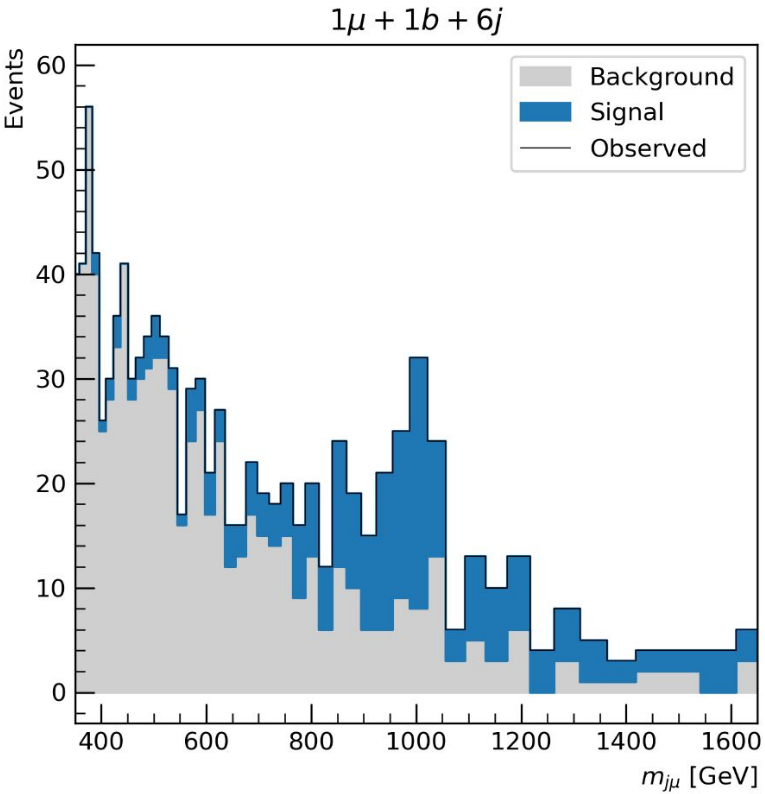


Dark Machines samples

- ★ Simulations including the **highest cross section processes at LHC**
  - ○ Designed for anomaly studies
  - ○ Produced using Madgraph, Pythia, and DELPHES fast detector simulations
- ★ Various simulated **new-physics signals** added to the background

e.g.: RPV stop -> bl

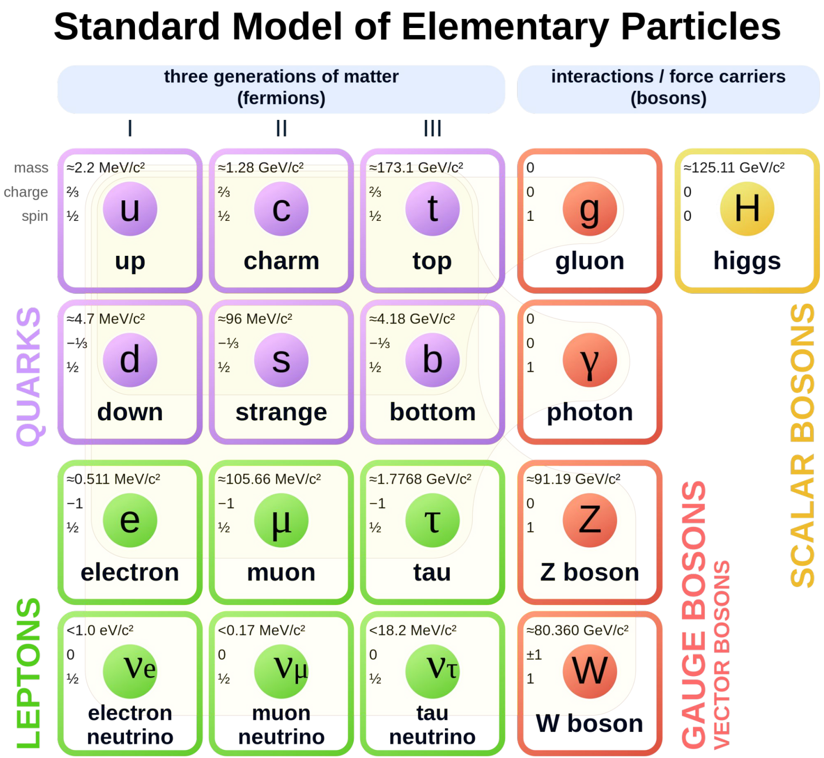


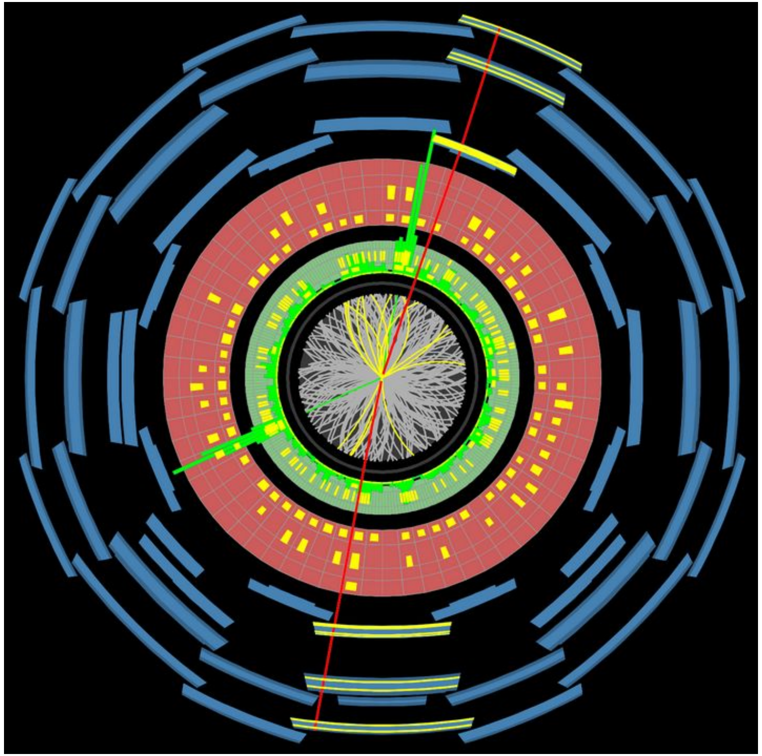


T. Aarrestad et al., SciPost Phys. 12, 043 (2022), arXiv:2105.14027 [hep-ph]

Signatures

Each event consists of a combination of reconstructed physics objects

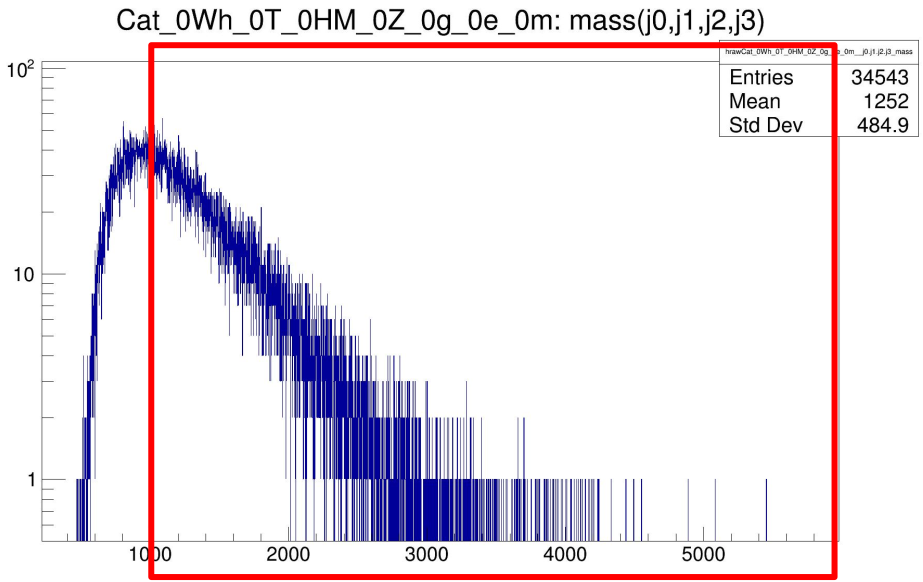




Higgs -> eeμμ

Histogram production

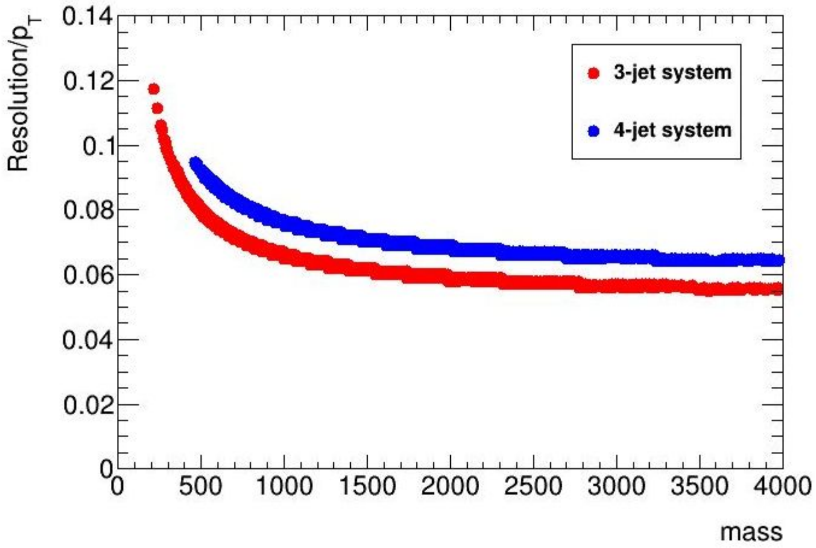
★ Consider all possible signatures:
  ○ e.g.: 1e + 1μ + 3j
★ Compute masses of all possible combinations of objects
  ○ e.g.: mass(e), mass(μ, j0), mass(e, μ,j1),...,mass(e,μ, j0, j1, j2)
★ Zoom in so the histogram starts at the maximum -> smoothly falling

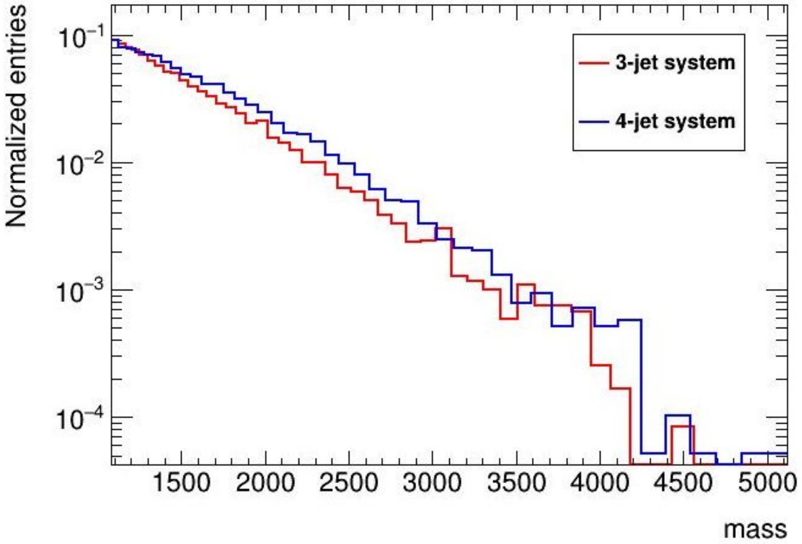


-> **Huge amount of data**: 63 000 signatures with many histograms each

Binning and detector resolution

- ★ Signal is expected to be narrow -> we train the network to find it in few bins
- ★ Adjust binning to ATLAS detector resolution
  - ○ Resolution depends on the objects involved, their $p_{T,}$ and η



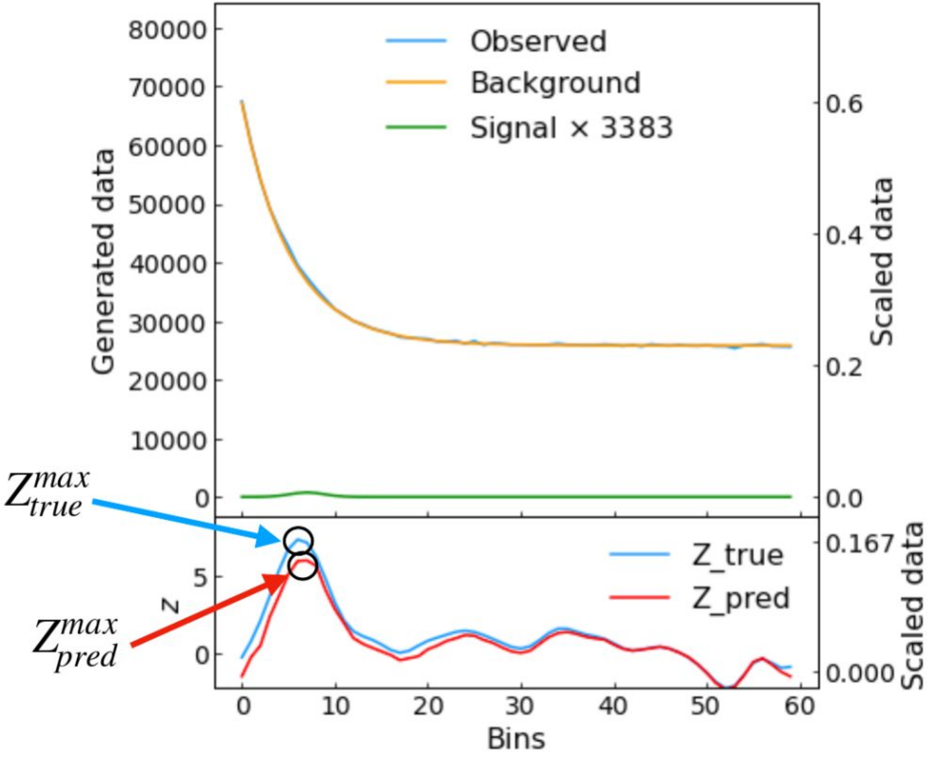


Performance

Performance evaluated in terms of:

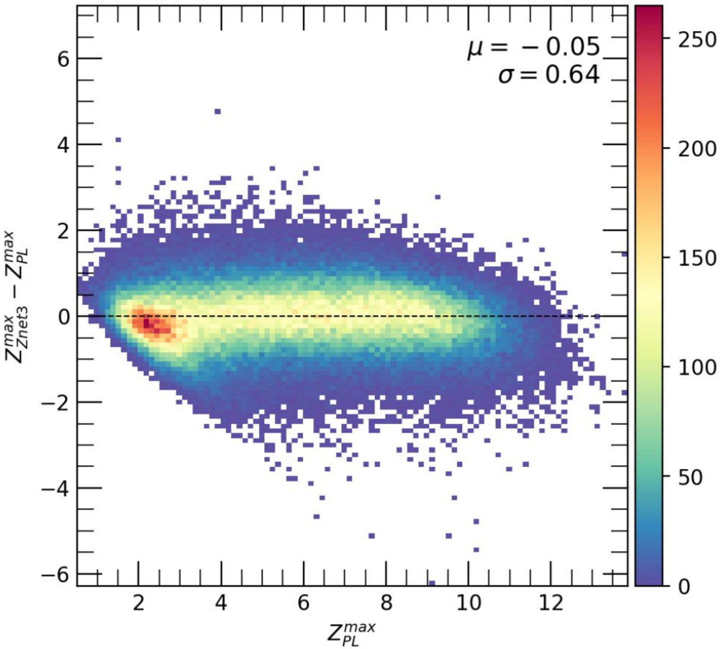
★ Difference between the true and predicted maximum significance: $z_{max,true} - z_{max,pred}$

★ False-positive rate

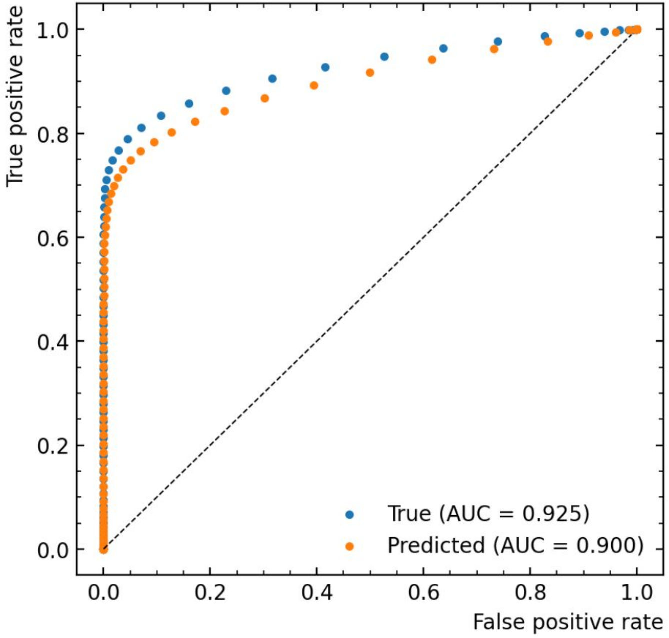


Performance

Prediction of z_{max} is **unbiased** with a variance of 0.64:

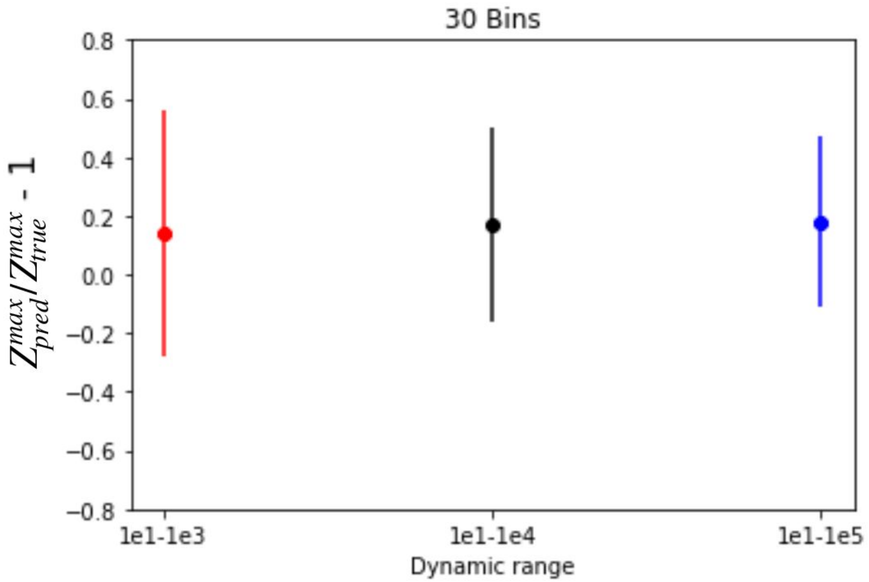


ROC curve: bump detection based on z_{pred} almost as good as based on z_{true}

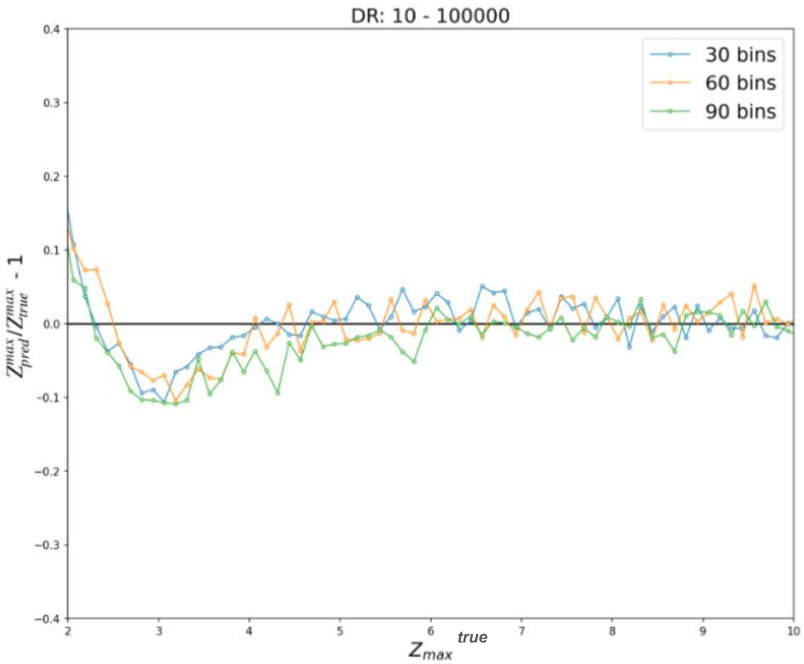


Performance stability

Dynamic range: number of entries per bin



Number of bins per histogramm

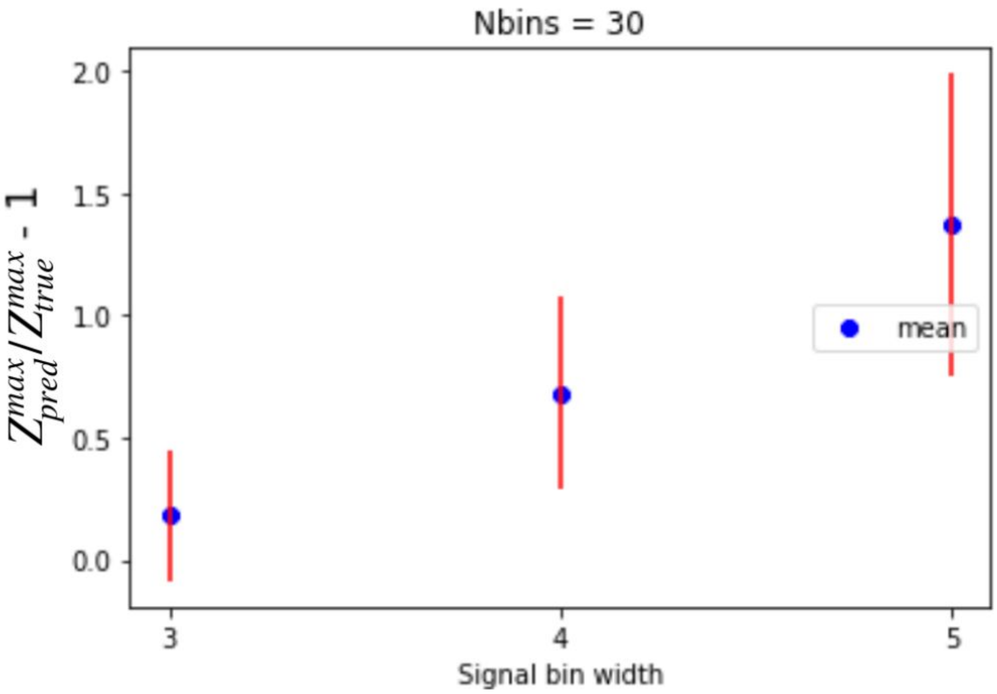


Performance stability

- ★ Training on fixed signal width: 1
- ★ Bias increases with signal width

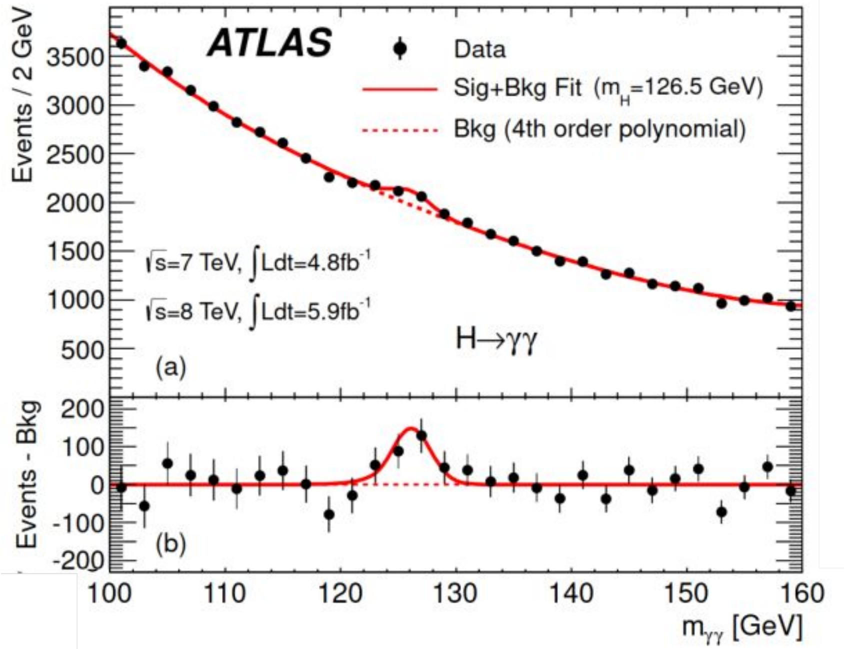
-> important to **adapt binning to experimental resolution**

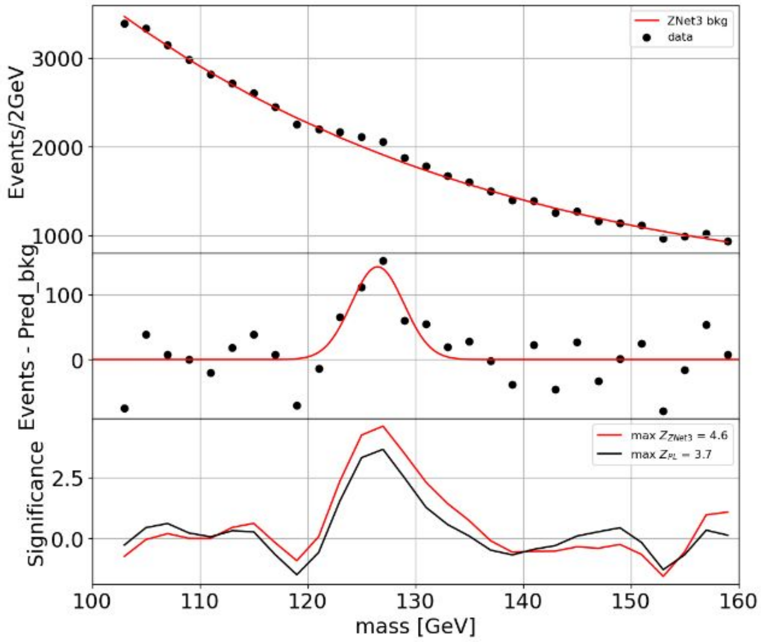
Mean & sigma as a function of signal width



Finding the Higgs bump

DDP was able to find the Higgs bump using the ATLAS plot:



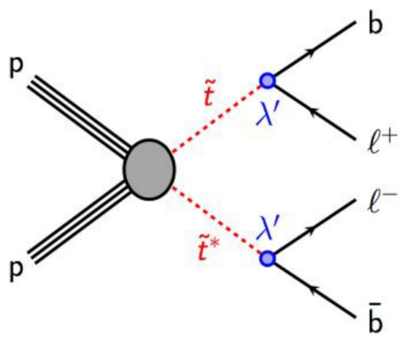


Bump was predicted at **correct mass** with $\mathbf{z_{pred}}$ **= 4.6**
(consistent with z_{true} of 3.7 within method precision)

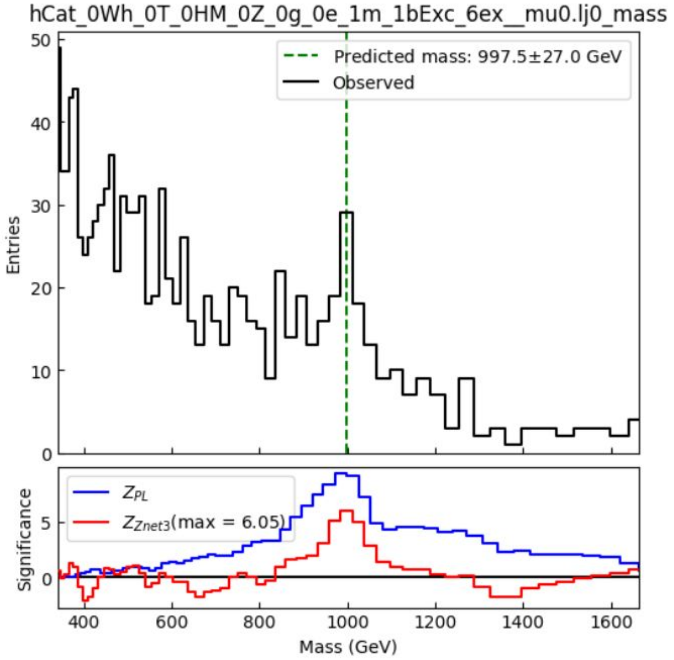
Finding new-physics signals

DDP was successful at finding the injected new-physics signals at correct mass:

RPV stop → bl

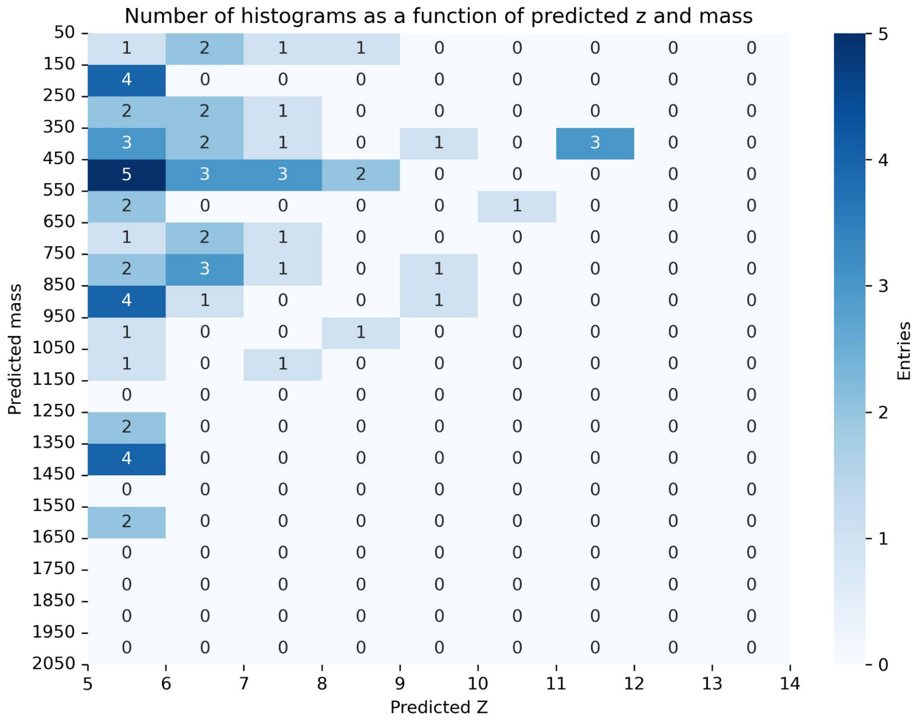


$m(j\mu)$ at 1μ+$1b$+$6j$



Finding new-physics signals

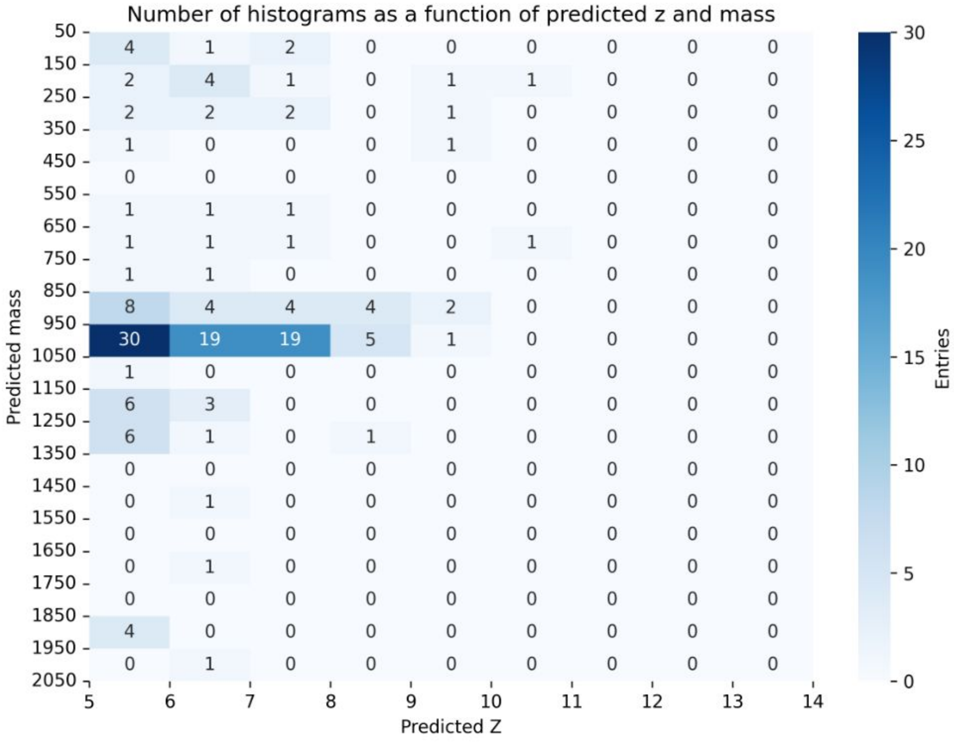
★ Testing over background-only samples results in **false-positive rate of 0.1%**



Background-only histograms with $Z_{max,pred} \geq 5$

Finding new-physics signals

- ★ Signal is usually detected in more than one histogram
- ★ Physics correlations help distinguish between anomaly and false-positives

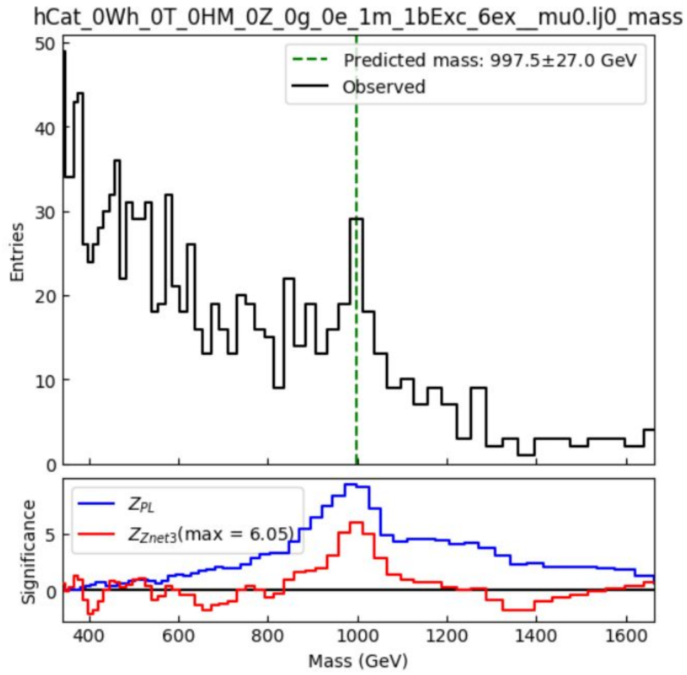


Histograms that find RPV stop -> bl with $Z_{max,pred} \geq 5$

Conclusion

★ DDP is successful at finding bumps
  ○ in analytical **functions**: 1st proof of concept
  ○ in **simulations**: 2nd proof of concept in progress
★ Method scans huge parameter space efficiently
★ Could be used to find interesting signal regions
★ Outlook: apply DDP to ATLAS data

$m(j\mu)$ at 1μ+1b+6j



Backup

Splitting the dataset by jet multiplicity

- ★ We split each signature by jet multiplicity (=number of jets)
- ★ Compare the position of $z_{max,pred}$ in neighboring jet multiplicities to reduce look-elsewhere effect

$z_{max,pred}$ of same signature with different jet multiplicity for a background-only:

6j	3.2
5j	1.7
4j	2.4
3j	1.9
2j	4.5
1j	2.2

mass at $z_{max,pred}$

Splitting the dataset by jet multiplicity

★ Signal should appear at same mass in several histograms with neighboring jet multiplicity

$Z_{max,pred}$ of same signature with different jet multiplicity for a signal:

6j	2.3
5j	1.7
4j	3.4
3j	5.3
2j	3.9
1j	2.2

mass at $z_{max,pred}$

Dark Machines samples

SM processes			
Physics process	Process ID	σ (pb)	N_{tot} ($N_{10\,\text{fb}^{-1}}$)
$pp \to jj(+2j)$	njets	$19718_{H_T>600\text{GeV}}$	415331302 (197179140)
$pp \to l^{\pm}\nu_l(+2j)$	w_jets	$10537_{H_T>100\text{GeV}}$	135692164 (105366237)
$pp \to \gamma j(+2j)$	gam_jets	$7927_{H_T>100\text{GeV}}$	123709226 (79268824)
$pp \to l^+l^-(+2j)$	z_jets	$3753_{H_T>100\text{GeV}}$	60076409 (37529592)
$pp \to t\bar{t}(+2j)$	ttbar	541	13590811 (5412187)
$pp \to t + \text{jets}(+2j)$	single_top	130	7223883 (1297142)
$pp \to \bar{t} + \text{jets}(+2j)$	single_topbar	112	7179922 (1116396)
$pp \to W^+W^-(+2j)$	ww	82.1	17740278 (821354)
$pp \to W^{\pm}t(+2j)$	wtop	57.8	5252172 (577541)
$pp \to W^{\pm}\bar{t}(+2j)$	wtopbar	57.8	4723206 (577541)
$pp \to \gamma\gamma(+2j)$	2gam	47.1	17464818 (470656)
$pp \to W^{\pm}\gamma(+2j)$	Wgam	45.1	18633683 (450672)
$pp \to ZW^{\pm}(+2j)$	zw	31.6	13847321 (315781)
$pp \to Z\gamma(+2j)$	Zgam	29.9	15909980 (299439)
$pp \to ZZ(+2j)$	zz	9.91	7118820 (99092)
$pp \to h(+2j)$	single_higgs	1.94	2596158 (19383)
$pp \to t\bar{t}\gamma(+2j)$	ttbarGam	1.55	95217 (15471)
$pp \to t\bar{t}Z$	ttbarZ	0.59	300000 (5874)
$pp \to t\bar{t}h(+1j)$	ttbarHiggs	0.46	200476 (4568)
$pp \to \gamma t(+2j)$	atop	0.39	2776166 (3947)
$pp \to t\bar{t}W^{\pm}$	ttbarW	0.35	279365 (3495)
$pp \to \gamma\bar{t}(+2j)$	atopbar	0.27	4770857 (2707)
$pp \to Zt(+2j)$	ztop	0.26	3213475 (2554)
$pp \to Z\bar{t}(+2j)$	ztopbar	0.15	2741276 (1524)
$pp \to t\bar{t}t\bar{t}$	4top	0.0097	399999 (96)
$pp \to t\bar{t}W^+W^-$	ttbarWW	0.0085	150000 (85)