ICHEP 2022 BOLOGNA

ICHEP 2022
XLI International Conference on High Energy Physics
Bologna (Italy)

Non-Parametric Data-Driven Background Modelling using Conditional Probabilities

A. Chisholm, T. Neep, K. Nikolopoulos, R. Owen, E. Reynolds, **J. Silva**

9th July 2022

UNIVERSITY OF BIRMINGHAM

erc

Exclusive Higgs

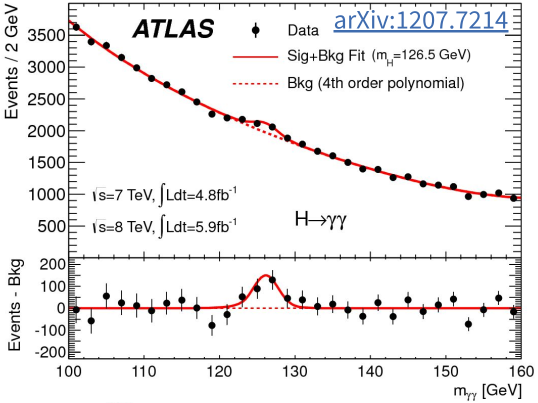
This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement no 714893 (ExclusiveHiggs)

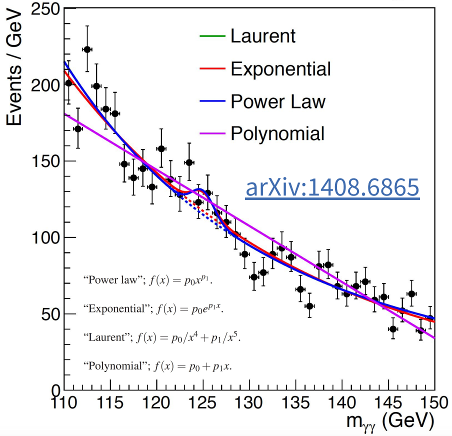
Motivation

➔ **Background modelling** is one of the main challenges in particle physics analyses

➔ Common techniques of background modelling:

- **MC simulation:**
  - Not always possible to model background with sufficient accuracy
  - Often computationally costly to produce large samples → significant statistical uncertainties
- **Parametric Models:**
  - Does the true shape belong to the family of curves parametrised by the chosen function?
    - Taken into account as **"spurious signal"** systematic uncertainty [1, 2]
    - Discrete **profiling of an ensemble of parametric forms** [3, 4, 5]





[arXiv:1207.7214 (1), arXiv:2007.07830 (2), arXiv:1408.6865 (3), arXiv:2002.06398 (4), arXiv:2009.04363 (5)]

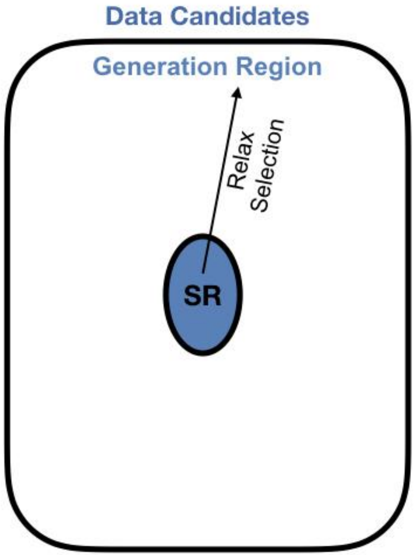
Background Modelling

➔ Non-parametric data-driven background modelling:

Background Modelling

➔ Non-parametric data-driven background modelling:

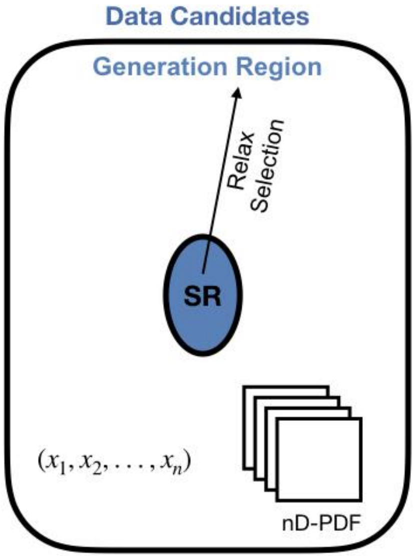
1. Obtain sample of data events enriched in background by relaxing event selection requirements (**Generation Region**)



Background Modelling

➔ Non-parametric data-driven background modelling:

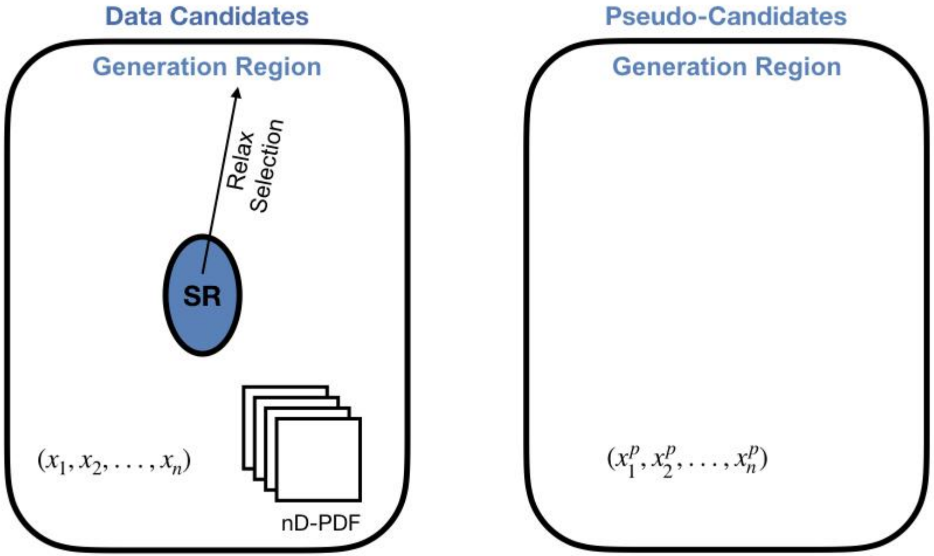
1. Obtain sample of data events enriched in background by relaxing event selection requirements (**Generation Region**)
2. Obtain conditional PDF of relevant variables ($x_1, x_2, \ldots, x_n$)



Background Modelling

➔ Non-parametric data-driven background modelling:

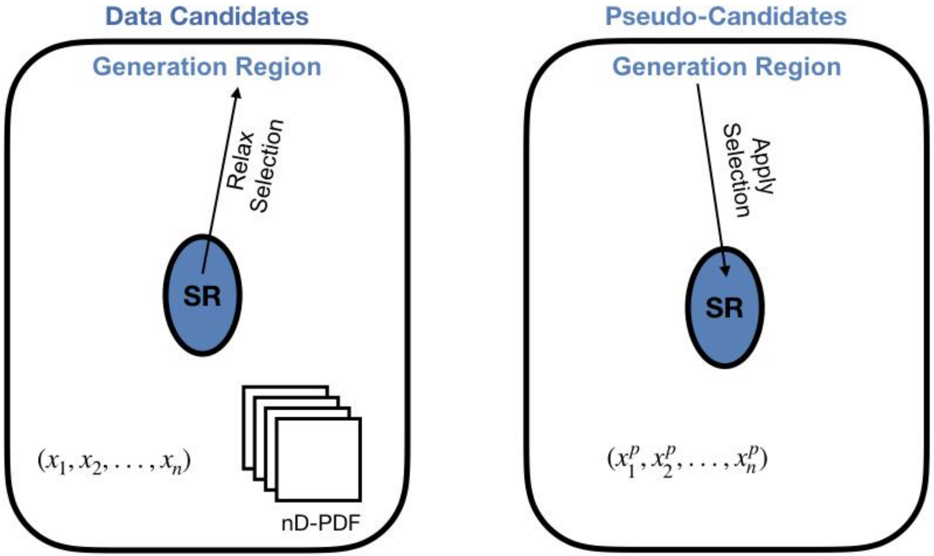
1. Obtain sample of data events enriched in background by relaxing event selection requirements (**Generation Region**)
2. Obtain conditional PDF of relevant variables (x_1, x_2,…, x_n)
3. Generate sample of pseudo-candidates



Background Modelling

➔ Non-parametric data-driven background modelling:

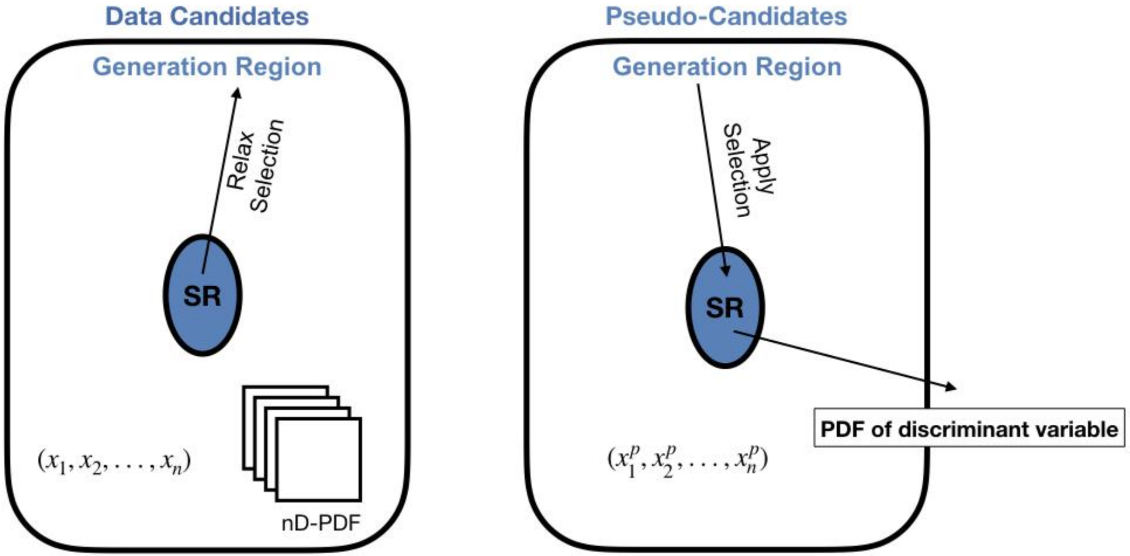
1. Obtain sample of data events enriched in background by relaxing event selection requirements (**Generation Region**)
2. Obtain conditional PDF of relevant variables ($x_1, x_2, \ldots, x_n$)
3. Generate sample of pseudo-candidates
4. Apply **Signal Region** requirements to pseudo-candidates sample



Background Modelling

➔ Non-parametric data-driven background modelling:

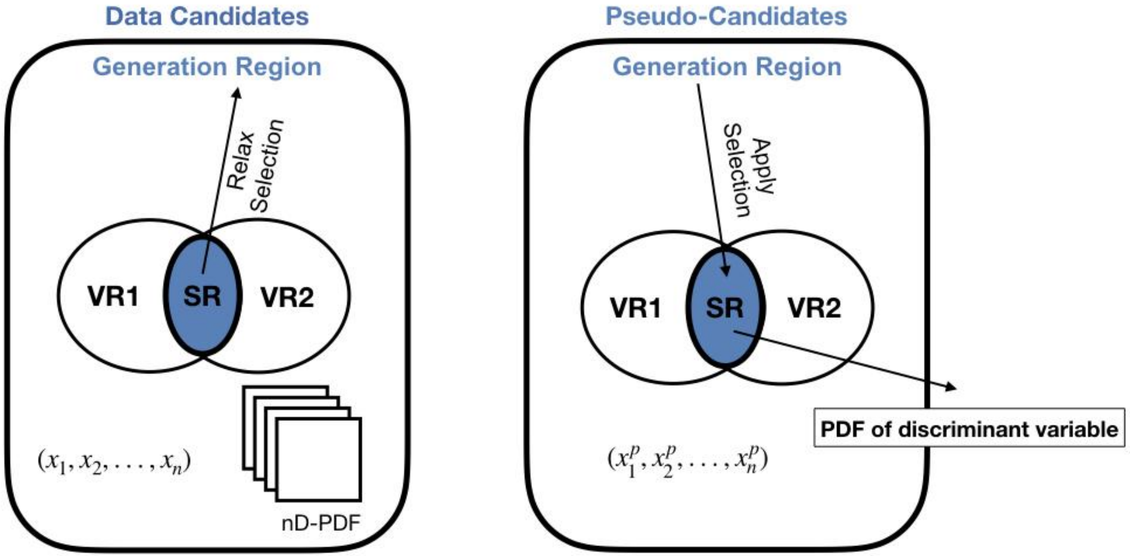
1. Obtain sample of data events enriched in background by relaxing event selection requirements (**Generation Region**)
2. Obtain conditional PDF of relevant variables (x_1, x_2,…, x_n)
3. Generate sample of pseudo-candidates
4. Apply **Signal Region** requirements to pseudo-candidates sample - obtain PDF of **discriminant variable** for statistical analysis



Background Modelling

➔ Non-parametric data-driven background modelling:

1. Obtain sample of data events enriched in background by relaxing event selection requirements (**Generation Region**)
2. Obtain conditional PDF of relevant variables (x_1, x_2,…, x_n)
3. Generate sample of pseudo-candidates
4. Apply **Signal Region** requirements to pseudo-candidates sample - obtain PDF of **discriminant variable** for statistical analysis
   - Intermediate **Validation Regions** to check method

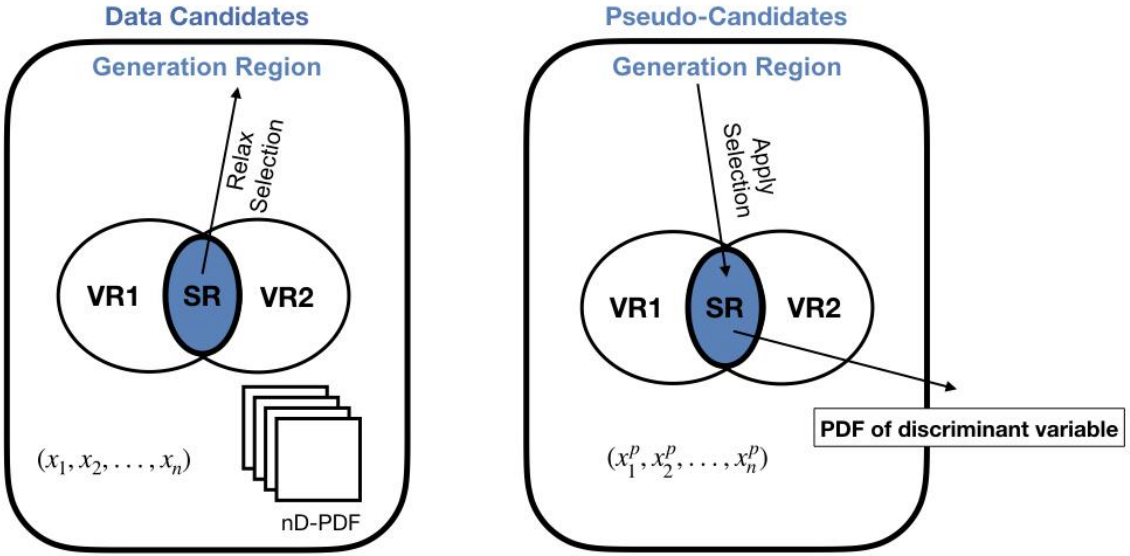


Background Modelling

➔ Non-parametric data-driven background modelling:

1. Obtain sample of data events enriched in background by relaxing event selection requirements (**Generation Region**)
2. Obtain conditional PDF of relevant variables ($x_1, x_2, \ldots, x_n$)
3. Generate sample of pseudo-candidates
4. Apply **Signal Region** requirements to pseudo-candidates sample - obtain PDF of **discriminant variable** for statistical analysis
   - Intermediate **Validation Regions** to check method

Today will present 2 implementations of the method

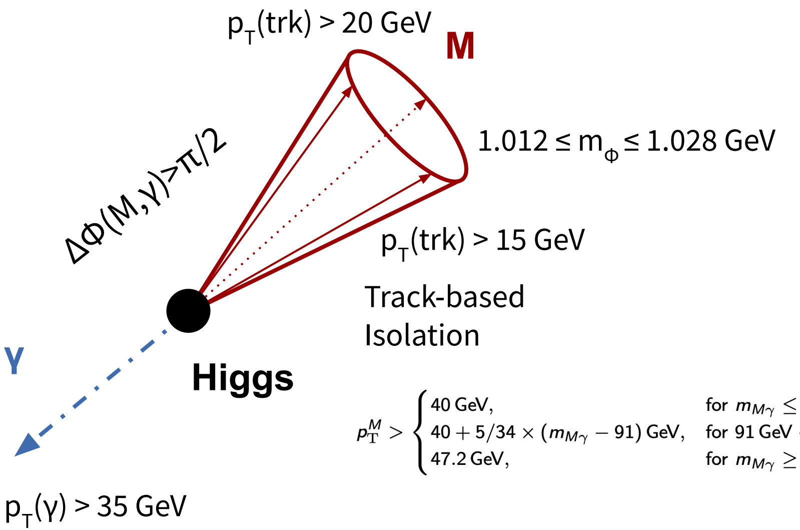


Ancestral Sampling

Case Study: H→Φγ

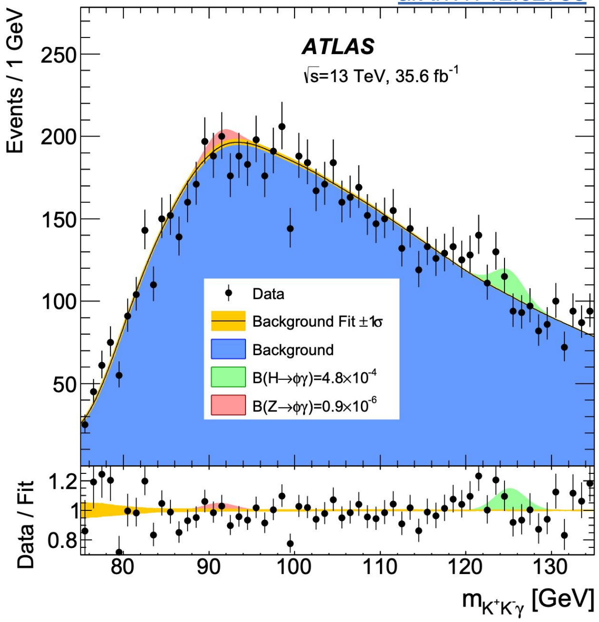
➔ H→ϕ(K⁺K⁻)γ has potential to probe **Higgs coupling to strange quark**

- **Distinct experimental signature**: pair of collimated high-p_T isolated tracks recoiling against isolated photon
- Main background : **photon + jet** and **dijet**
  - difficult to model accurately using MC - ideal use case for method
  - **photon + jet MC sample** used to exemplify model application



$$p_T^M > \begin{cases} 40\ \text{GeV}, & \text{for } m_{M\gamma} \le 91\ \text{GeV} \\ 40 + 5/34 \times (m_{M\gamma} - 91)\ \text{GeV}, & \text{for } 91\ \text{GeV} < m_{M\gamma} < 140\ \text{GeV} \\ 47.2\ \text{GeV}, & \text{for } m_{M\gamma} \ge 140\ \text{GeV} \end{cases}$$

arXiv:1712.02758



Building the model for H→Φγ

Relax Selection

Obtain Conditional PDFs

Generate pseudo candidates

Apply Selection

1. Relax $p_T(M)$ and Iso(M) requirements

Region	**$p_T(M)$ cut**	**Iso(M) cut**
GR	x	x
VR1	✓	x
VR2	x	✓
SR	✓	✓

Building the model for H→Φγ

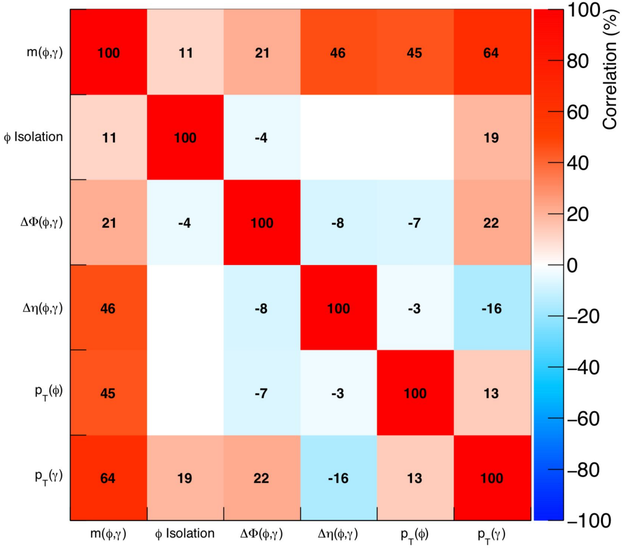
Relax Selection

Obtain Conditional PDFs

Generate pseudo candidates

Apply Selection

2. Build PDFs of relevant **kinematic** and **isolation** variables in generation region - we need the ϕ and γ 4-momentum vectors to ultimately obtain **m(ϕγ)** + Iso(ϕ)
   - 1D, 2D and 3D **histograms** to be sampled from in generation step
   - only most important correlations are explicitly described



$p_T(\phi)$	$p_T(\gamma)$

$p_T(\gamma)$	ϕ Iso

$p_T(\phi)$	$p_T(\gamma)$	$\Delta\Phi(\phi,\gamma)$

$\Delta\Phi(\phi,\gamma)$	$\Delta\eta(\phi,\gamma)$

Building the model for H→Φγ

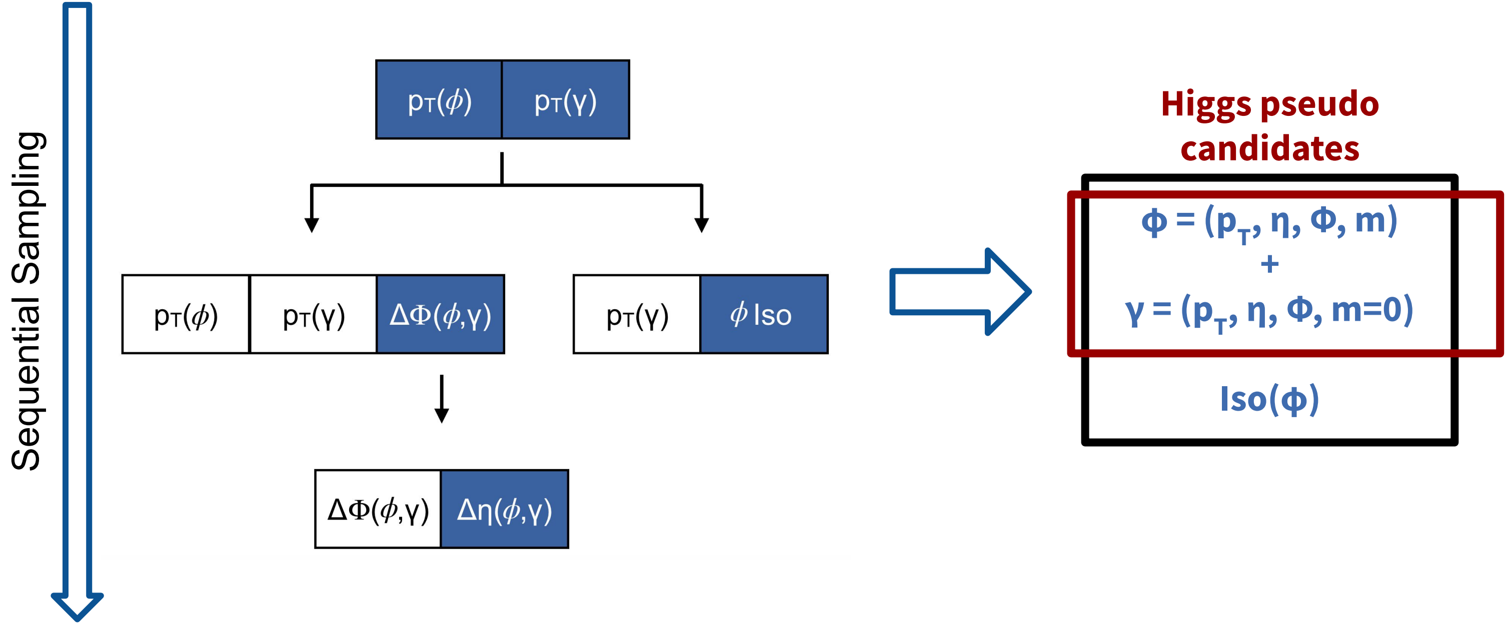
Relax Selection

Obtain Conditional PDFs

Generate pseudo candidates

Apply Selection

3. **Sample** from PDFs and construct pseudo-candidates
   - each pseudo-candidate is defined by the ϕ and γ 4-momentum vectors, and an associated Φ isolation variable



Building the model for H→Φγ

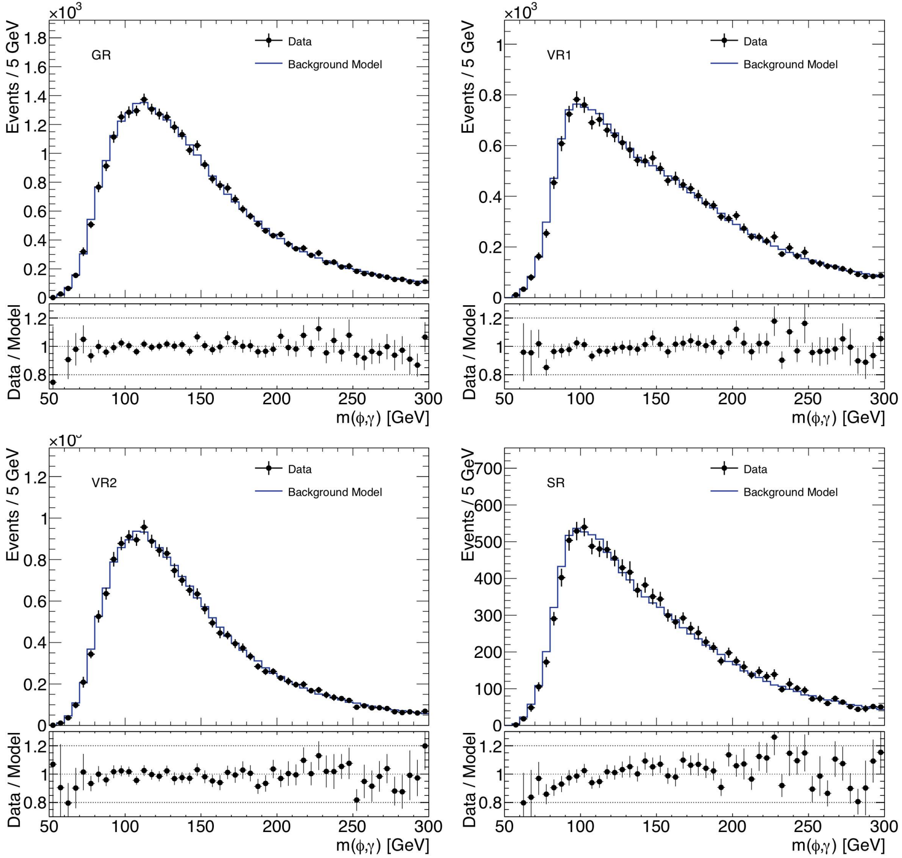
Relax Selection

Obtain Conditional PDFs

Generate pseudo candidates

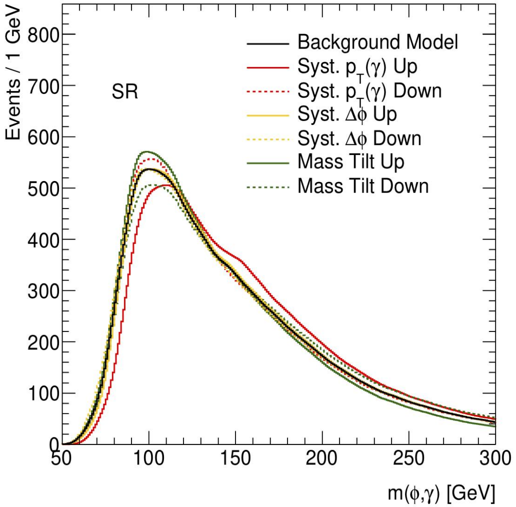
Apply Selection

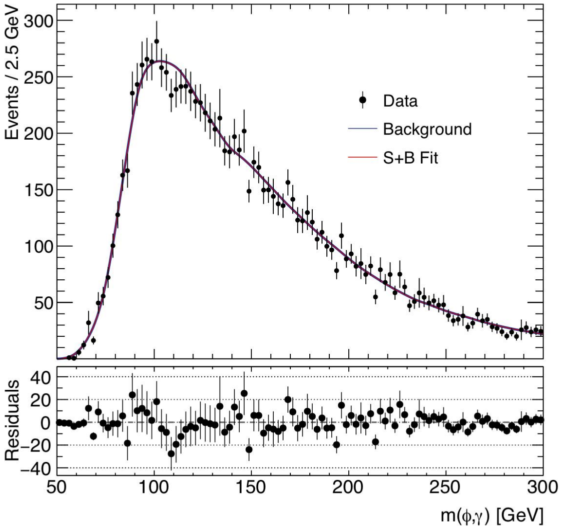
4. Apply $p_T(M)$ and Iso(M) requirements to sample of pseudo-candidates
   - obtain PDF of **m(ϕγ)** for statistical analysis in Signal and Validation Regions



Implementation in Statistical Analysis

- ➔ **Systematic uncertainties** are provided through variations of the nominal PDFs
  - ◆ selected to capture different modes of potential deformations of the background shape
- ➔ Maximum likelihood fit to invariant mass
  - ◆ each variation controlled by a nuisance parameter - directly constrained by data in fit





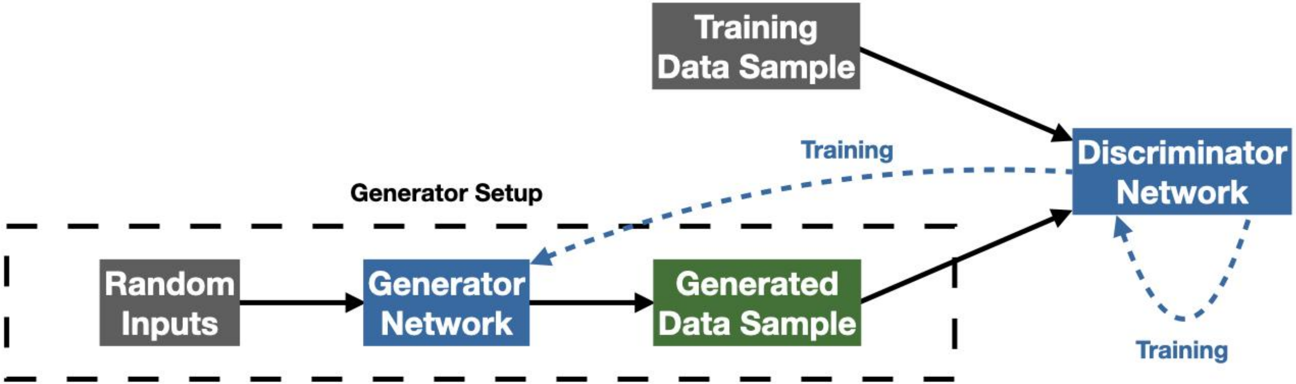
Parameter	Value	Uncertainty $(\pm 1\sigma)$
μ_{signal}	-0.03	± 0.55
μ_{bkgd}	1.01	± 0.01
Shape: $p_T(\gamma)$ shift	0.27	± 0.15
Shape: $\Delta\Phi(\phi,\gamma)$ tilt	0.26	± 0.43
Shape: $m(\phi,\gamma)$ tilt	0.11	± 0.24

Conditional Generative Adversarial Networks

Generative Adversarial Networks

➔ Challenge for ancestral sampling:
  - application in multivariate analyses
  - signal region blinding

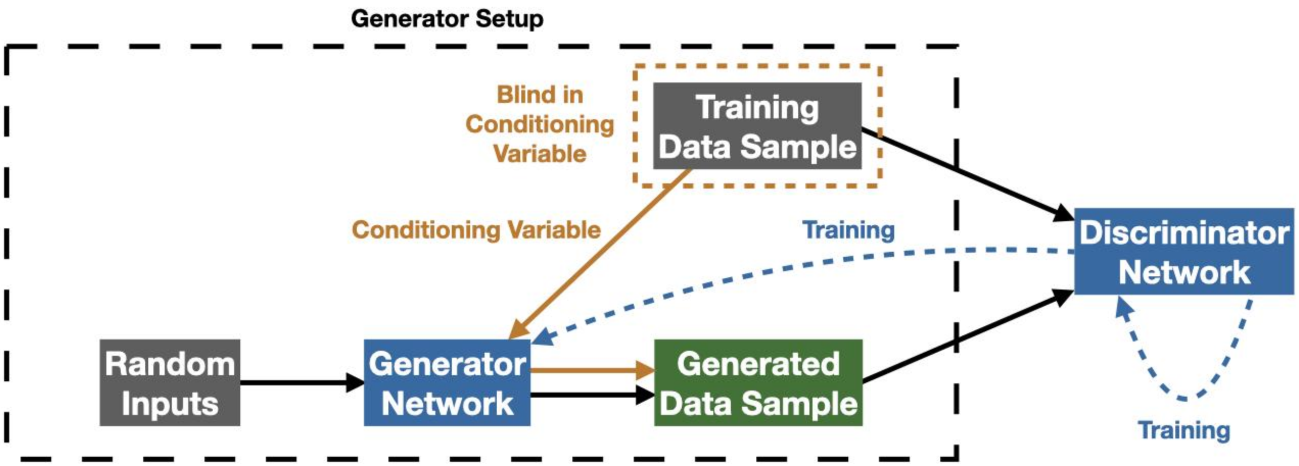
➔ Generalisation of method: use **GANs trained on data** to produce background model
  - **Generator** - learns generative model from data sample
  - **Discriminator** - simultaneously trained to discriminate the generator output from data



Conditional Generative Adversarial Networks

➔ Possible **signal contamination** in training data:

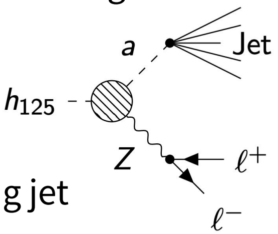
- **Condition** GAN (cGAN) on a blinding variable, allowing **SR to be blinded during training** - cGAN extrapolates prediction into SR



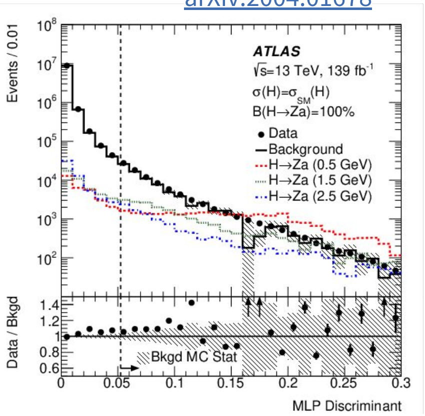
Case Study: H→Za

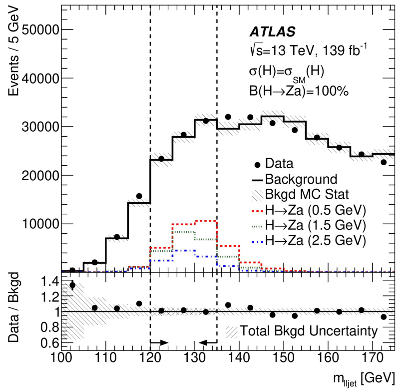
➔ Light pseudo-scalars produced in Higgs decays feature in BSM theories like two-Higgs-doublet model and the 2HDM with additional scalar singlet

➔ Search for **H→Z(ll)+a**, with a→hadrons:
- Main background: **Z + jets**
- background discrimination relies on **MVA** techniques, using jet substructure variables
- ideal case study for implementation of background modelling using cGANs
  - systematics arising by limited stats in MC simulation (arXiv:2004.01678)
  - use of MVA techniques makes it impractical to use ancestral sampling



arXiv:2004.01678





➔ **Z + jets MC sample** used to exemplify model application

Building the model for H→Za

Relax Selection

Obtain Conditional PDFs

Generate pseudo candidates

Apply Selection

1. Remove MLP-based selection
   - **& blind signal region** to avoid signal contamination

Use $m_{\mu\mu j}$ as blinding variable

⇩

123 GeV ≤ $m_{\mu\mu j}$ ≤ 135 GeV blinded

Building the model for H→Za

Relax Selection

Obtain Conditional PDFs

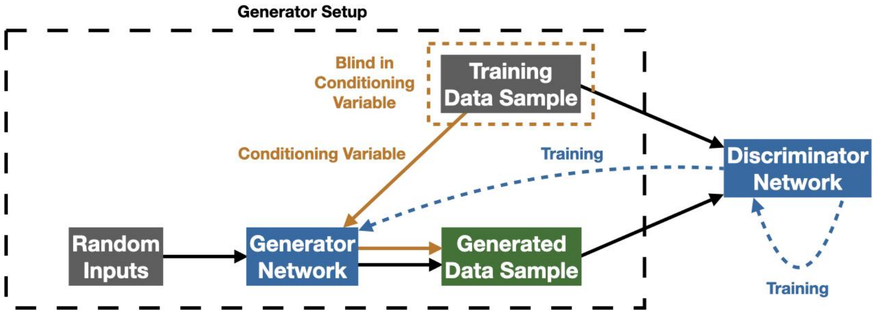
Generate pseudo candidates

Apply Selection

2. cGans trained using **blinded data**
   - learn generative model of the conditional probability distribution of the data, given value of blinding variable
   - Use **ensemble** of cGANs and take average:
     - 100 cGANs trained, 5 best based on χ^2 metric kept for analysis

Generator and discriminator:

- 5 layers x 256 hidden nodes with leaky ReLU activation function
- binary cross entropy loss function and L2 regularisation



Building the model for H→Za

Relax Selection

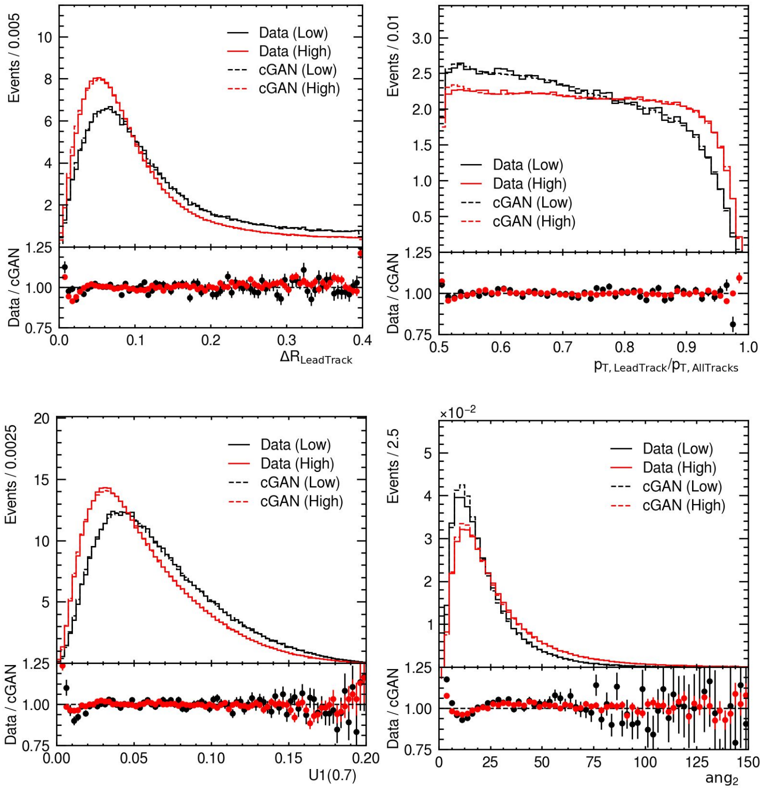
Obtain Conditional PDFs

Generate pseudo candidates

Apply Selection

3. Input inclusive distribution of the conditioning variable into cGAN:
   - cGAN **extrapolates** the conditional generative model into signal region
   - obtain prediction of **MLP input variables**

$m_{\mu\mu j}$ sidebands



Building the model for H→Za

Relax Selection

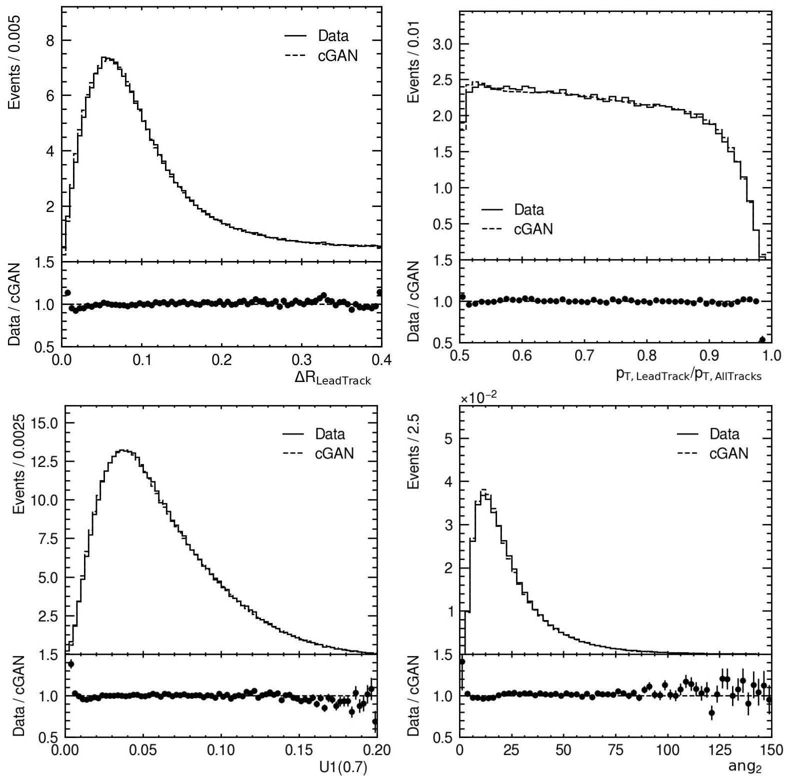
Obtain Conditional PDFs

Generate pseudo candidates

Apply Selection

3. Input inclusive distribution of the conditioning variable into cGAN:
   - cGAN **extrapolates** the conditional generative model into signal region
   - obtain prediction of **MLP input variables**

$m_{\mu\mu j}$ SR



Building the model for H→Za

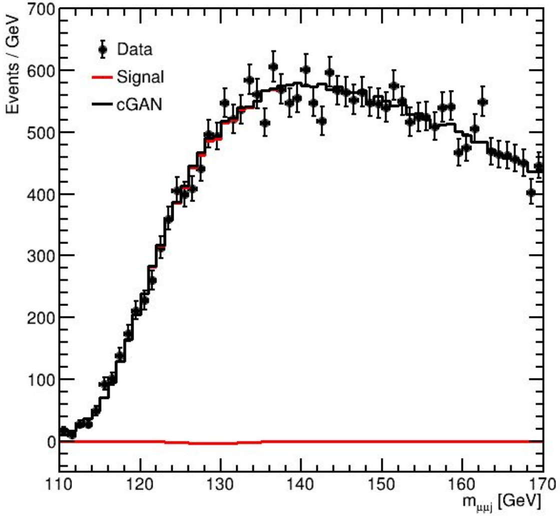
Relax Selection

Obtain Conditional PDFs

Generate pseudo candidates

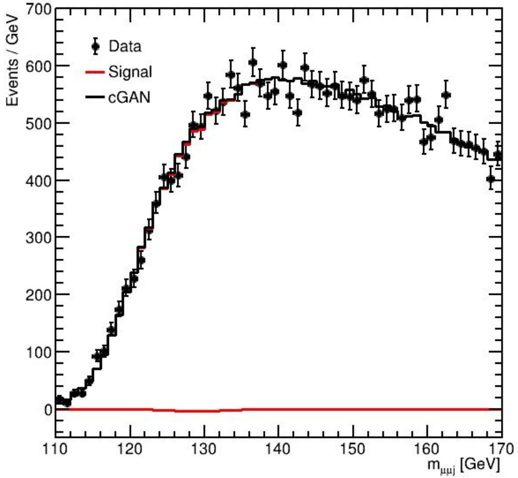
Apply Selection

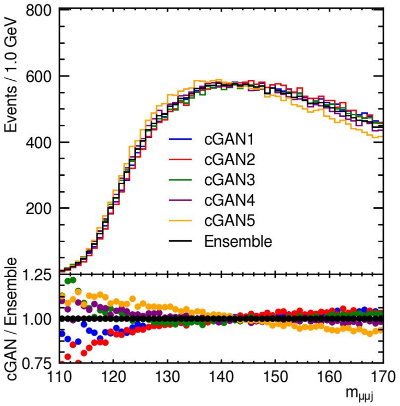
4. Apply MLP selection to pseudo-candidates sample
   - obtain PDF of $\mathbf{m}_{\mu\mu j}$ in SR for statistical analysis

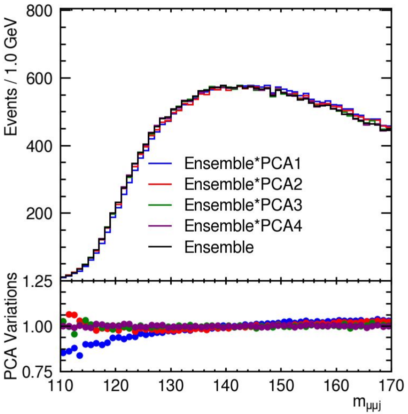


Implementation in Statistical Analysis

- ➔ **Systematic uncertainties** are provided through shape variations:
  - ◆ Differences between ensemble and individual cGANs
  - ◆ Principal component analysis performed to orthogonalise differences
- ➔ Maximum likelihood fit to Higgs invariant mass
  - ◆ each variation controlled by a nuisance parameter - directly constrained by data in fit







Parameter	Value	Uncertainty ($\pm 1\sigma$)
μ_{signal}	-0.003	± 0.010
μ_{bkgd}	1.001	± 0.008
Shape uncertainty 1	-0.36	± 0.27
Shape uncertainty 2	-0.31	± 0.52

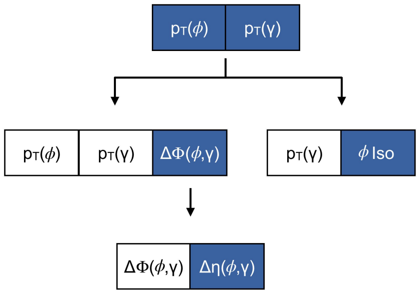
Summary

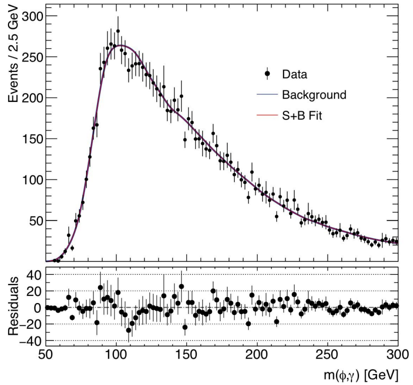
➔ A novel **non-parametric**, **data-driven** background modelling technique was presented
- Addresses typical shortcomings of often employed background modelling techniques
- Dataset from a **relaxed event selection** to create a model based on **conditional probabilities**
- Two distinct ways of building the conditional PDF:

arXiv:2112.00650

Ancestral sampling

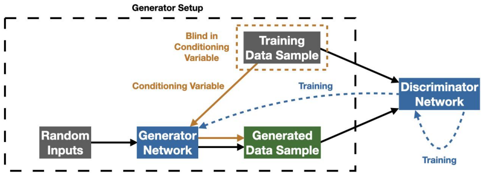
- Sample from histograms of relevant variables in data, built with respect to most important correlations
- Already used in multiple analysis! [Phys. Rev. Lett. 114 (2015) 121801, Phys. Rev. Lett. 117, 111802 (2016), JHEP 07 (2018) 127, Phys. Lett. B 786 (2018) 134]

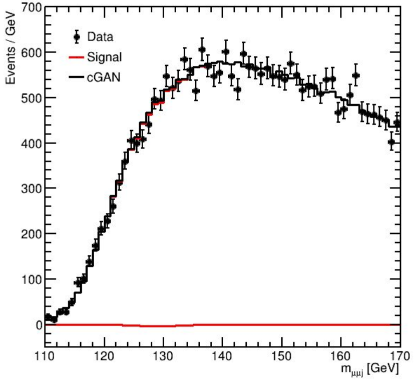




Conditional Generative Adversarial Networks

- Generalisation of ancestral sampling
- Use GANs trained on data to produce background model
- **Condition** GAN (cGAN) on a blinding variable, allowing **SR to be blinded during training**





BACK-UP

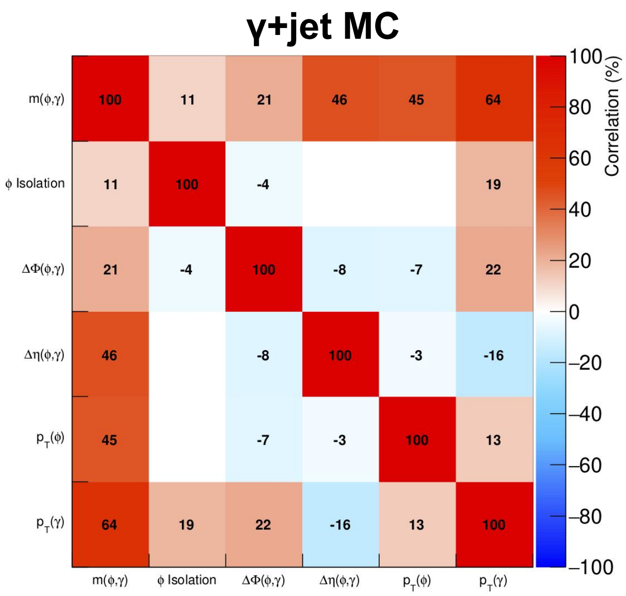
Building the model for H→Φγ

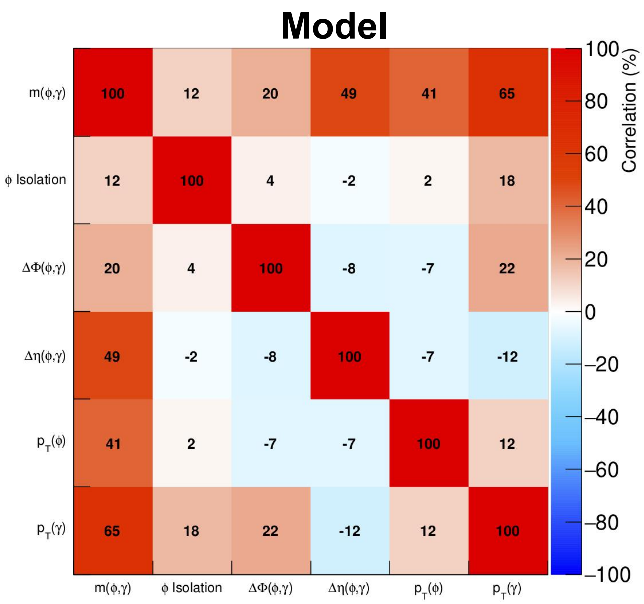
Relax Selection

Obtain Conditional PDFs

Generate pseudo candidates

Apply Selection





Ancestral Sampling

➔ Background model is **robust under signal contamination**
- ◆ Signal Injection tests to evaluate robustness