Non-Parametric Data-Driven Background Modelling using Conditional Probabilities

Júlia Silva

15th November 2022

UNIVERSITY OF BIRMINGHAM

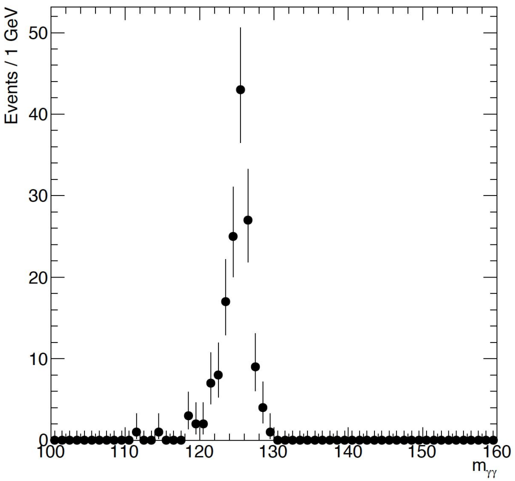
erc

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)

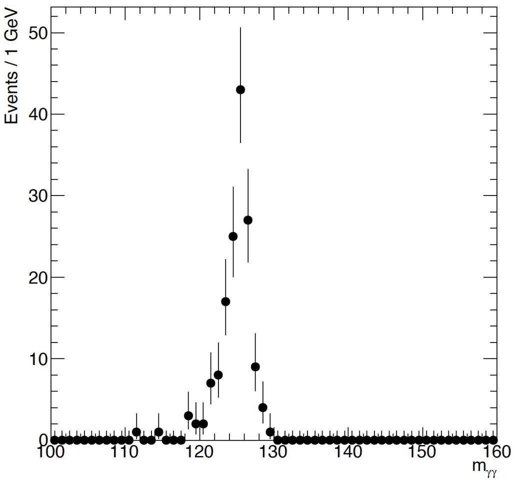
Searching for a new physics process

when searching for a **signal**

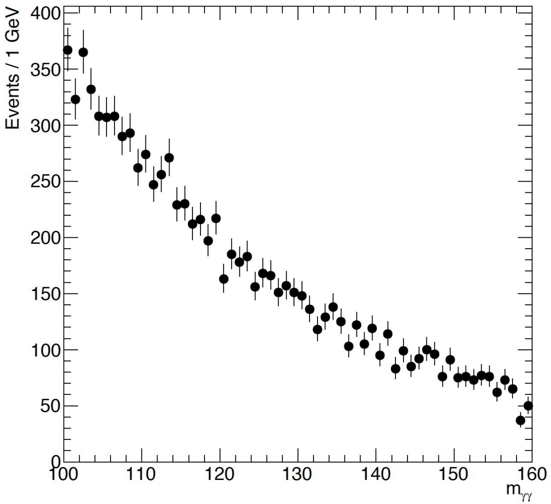


Searching for a new physics process

when searching for a **signal**

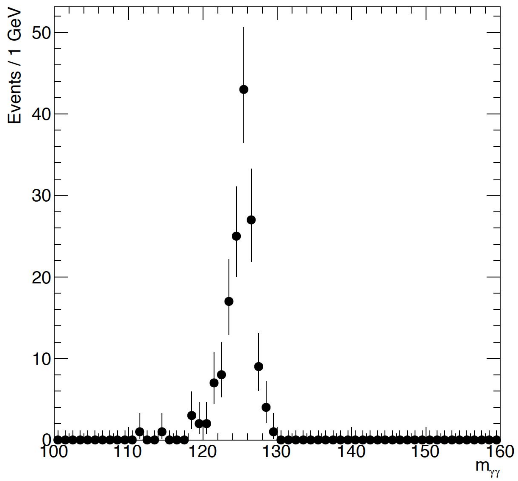


one needs to understand the **background**

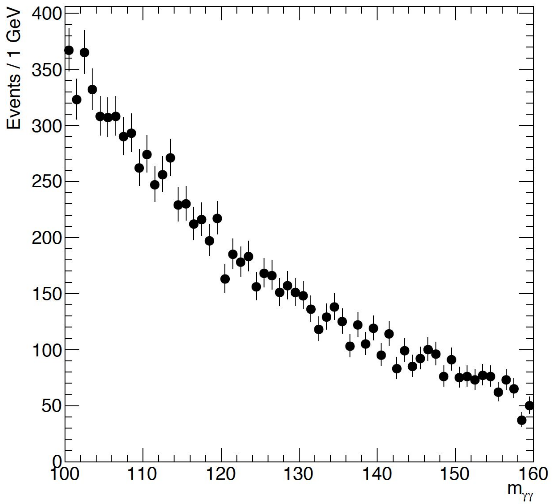


Searching for a new physics process

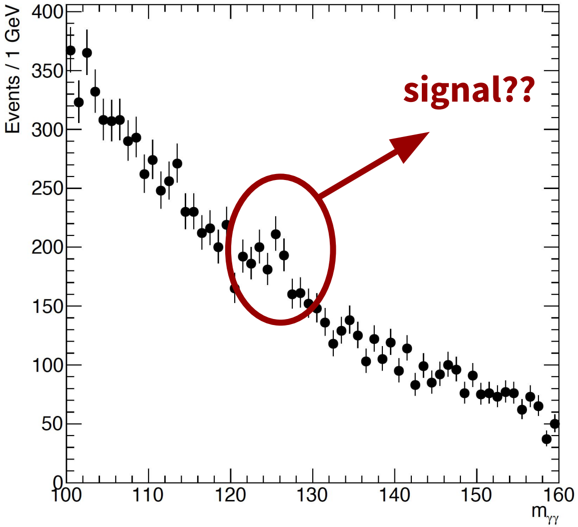
when searching for a **signal** one needs to understand the **background**



+

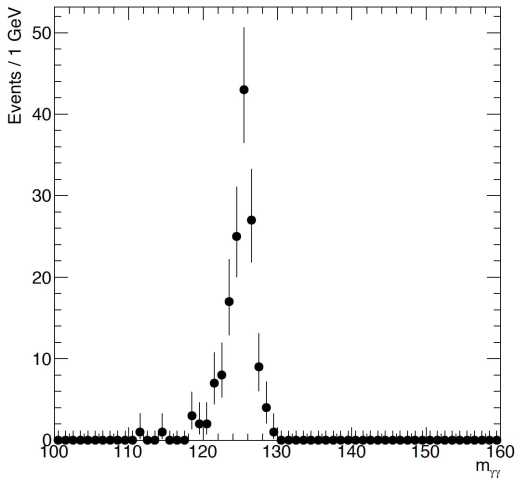


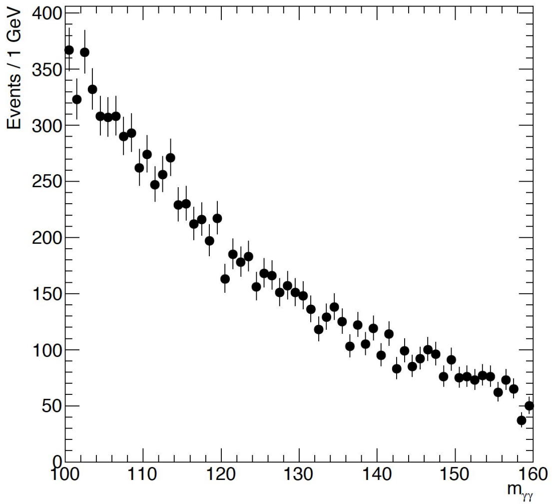
=

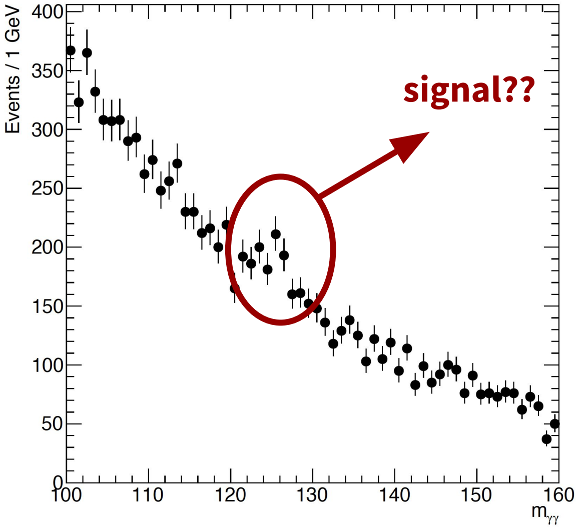


Searching for a new physics process

when searching for a **signal** one needs to understand the **background**





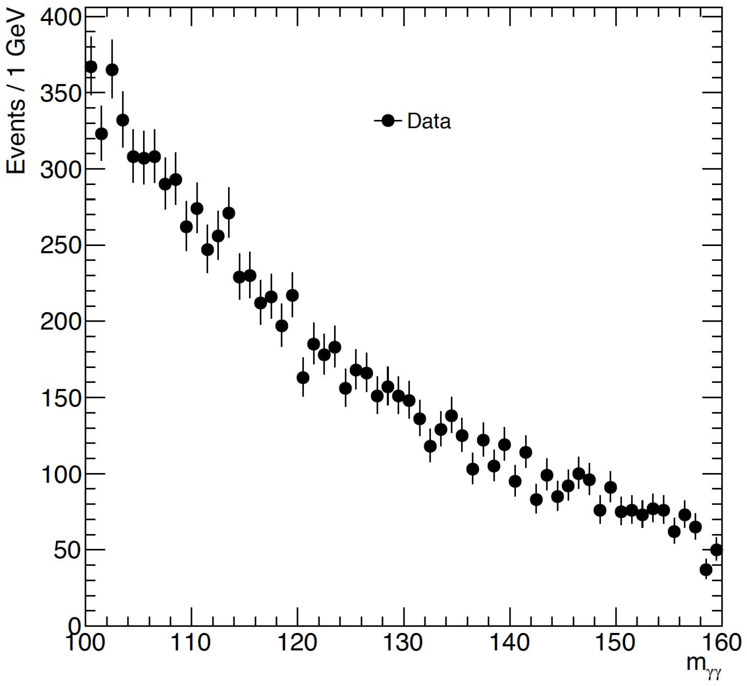


How can we model our background?

Modelling the background

What about **parametric models**?

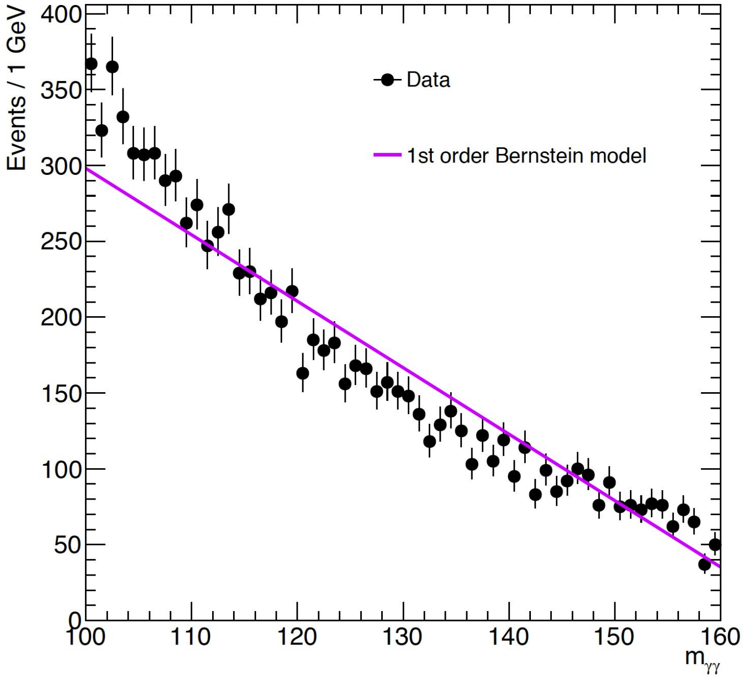
- Which functional form?
- How many parameters?



Modelling the background

What about **parametric models**?

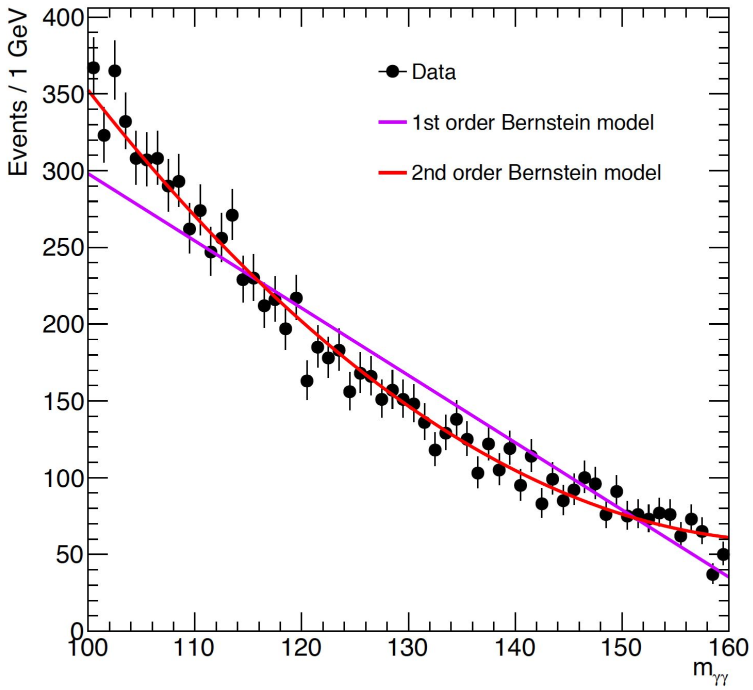
- Which functional form?
- How many parameters?



Modelling the background

What about **parametric models**?

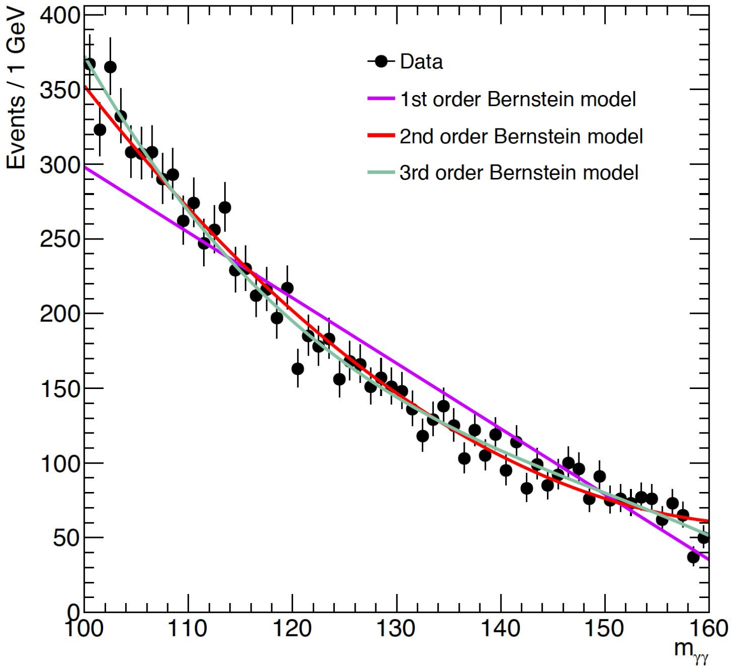
- Which functional form?
- How many parameters?



Modelling the background

What about **parametric models**?

- Which functional form?
- How many parameters?

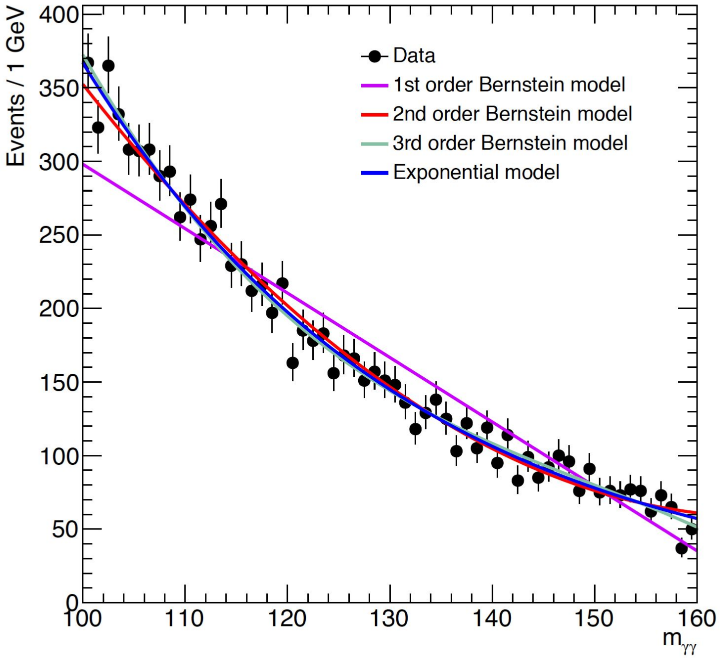


Modelling the background

What about **parametric models**?

- Which functional form?
- How many parameters?

There is no guarantee the true background shape is part of the family of curves parameterized by the chosen function



Modelling the background

- **MC simulation** is a commonly used technique

Modelling the background

- **MC simulation** is a commonly used technique
  - not always possible to model the background with sufficient accuracy → significant **theoretical uncertainties**

ATLAS ttH(bb)

Uncertainty source	$\Delta\mu$	
Process modelling		
$t\bar{t}H$ modelling	+0.13	−0.05
$t\bar{t}+\geq 1b$ modelling		
$t\bar{t}+\geq 1b$ NLO matching	+0.21	−0.20
$t\bar{t}+\geq 1b$ fractions	+0.12	−0.12
$t\bar{t}+\geq 1b$ FSR	+0.10	−0.11
$t\bar{t}+\geq 1b$ PS & hadronisation	+0.09	−0.08
$t\bar{t}+\geq 1b$ p_{T}^{bb} shape	+0.04	−0.04
$t\bar{t}+\geq 1b$ ISR	+0.04	−0.04
$t\bar{t}+\geq 1c$ modelling	+0.03	−0.04
$t\bar{t}+$ light modelling	+0.03	−0.03
tW modelling	+0.08	−0.07
Background-model statistical uncertainty	+0.04	−0.05
b-tagging efficiency and mis-tag rates		
b-tagging efficiency	+0.03	−0.02
c-mis-tag rates	+0.03	−0.03
l-mis-tag rates	+0.02	−0.02
Jet energy scale and resolution		
b-jet energy scale	+0.00	−0.01
Jet energy scale (flavour)	+0.01	−0.01
Jet energy scale (pile-up)	+0.00	−0.01
Jet energy scale (remaining)	+0.01	−0.01
Jet energy resolution	+0.02	−0.02
Luminosity	+0.01	−0.00
Other sources	+0.03	−0.03
Total systematic uncertainty	+0.30	−0.28
$t\bar{t}+\geq 1b$ normalisation	+0.04	−0.07
Total statistical uncertainty	+0.20	−0.20
Total uncertainty	+0.36	−0.34

arXiv:2111.06712

Modelling the background

- **MC simulation** is a commonly used technique
  - not always possible to model the background with sufficient accuracy → significant **theoretical uncertainties**

ATLAS ttH(bb)

Uncertainty source	$\Delta\mu$	
Process modelling		
$t\bar{t}H$ modelling	+0.13	−0.05
$t\bar{t}+\geq 1b$ modelling		
$t\bar{t}+\geq 1b$ NLO matching	+0.21	−0.20
$t\bar{t}+\geq 1b$ fractions	+0.12	−0.12
$t\bar{t}+\geq 1b$ FSR	+0.10	−0.11
$t\bar{t}+\geq 1b$ PS & hadronisation	+0.09	−0.08
$t\bar{t}+\geq 1b$ p_{T}^{bb} shape	+0.04	−0.04
$t\bar{t}+\geq 1b$ ISR	+0.04	−0.04
$t\bar{t}+\geq 1c$ modelling	+0.03	−0.04
$t\bar{t}+$ light modelling	+0.03	−0.03
tW modelling	+0.08	−0.07
Background-model statistical uncertainty	+0.04	−0.05
b-tagging efficiency and mis-tag rates		
b-tagging efficiency	+0.03	−0.02
c-mis-tag rates	+0.03	−0.03
l-mis-tag rates	+0.02	−0.02
Jet energy scale and resolution		
b-jet energy scale	+0.00	−0.01
Jet energy scale (flavour)	+0.01	−0.01
Jet energy scale (pile-up)	+0.00	−0.01
Jet energy scale (remaining)	+0.01	−0.01
Jet energy resolution	+0.02	−0.02
Luminosity	+0.01	−0.00
Other sources	+0.03	−0.03
Total systematic uncertainty	+0.30	−0.28
$t\bar{t}+\geq 1b$ normalisation	+0.04	−0.07
Total statistical uncertainty	+0.20	−0.20
Total uncertainty	+0.36	−0.34

arXiv:2111.06712

ATLAS VH(cc)

Source of uncertainty		$\mu_{VH(c\bar{c})}$
Total		15.3
Statistical		10.0
Systematic		11.5
Statistical uncertainties		
Signal normalisation		7.8
Other normalisations		5.1
Theoretical and modelling uncertainties		
$VH(\rightarrow c\bar{c})$		2.1
Z + jets		7.0
Top quark		3.9
W + jets		3.0
Diboson		1.0
$VH(\rightarrow b\bar{b})$		0.8
Multi-jet		1.0
Simulation samples size		4.2
Experimental uncertainties		
Jets		2.8
Leptons		0.5
$E_{\mathrm{T}}^{\mathrm{miss}}$		0.2
Pile-up and luminosity		0.3
Flavour tagging	c-jets	1.6
	b-jets	1.1
	light-jets	0.4
	τ-jets	0.3
Truth-flavour tagging	ΔR correction	3.3
	Residual non-closure	1.7

arXiv:2201.11428

Modelling the background

- **MC simulation** is a commonly used technique
  - not always possible to model the background with sufficient accuracy → significant **theoretical uncertainties**
  - Often computationally costly to produce large samples → significant **statistical uncertainties**

ATLAS ttH(bb)

Uncertainty source	$\Delta\mu$	
Process modelling		
$t\bar{t}H$ modelling	$+0.13$	-0.05
$t\bar{t}+\geq 1b$ modelling		
$t\bar{t}+\geq 1b$ NLO matching	$+0.21$	-0.20
$t\bar{t}+\geq 1b$ fractions	$+0.12$	-0.12
$t\bar{t}+\geq 1b$ FSR	$+0.10$	-0.11
$t\bar{t}+\geq 1b$ PS & hadronisation	$+0.09$	-0.08
$t\bar{t}+\geq 1b$ p_T^{bb} shape	$+0.04$	-0.04
$t\bar{t}+\geq 1b$ ISR	$+0.04$	-0.04
$t\bar{t}+\geq 1c$ modelling	$+0.03$	-0.04
$t\bar{t}+$ light modelling	$+0.03$	-0.03
tW modelling	$+0.08$	-0.07
Background-model statistical uncertainty	$+0.04$	-0.05
b-tagging efficiency and mis-tag rates		
b-tagging efficiency	$+0.03$	-0.02
c-mis-tag rates	$+0.03$	-0.03
l-mis-tag rates	$+0.02$	-0.02
Jet energy scale and resolution		
b-jet energy scale	$+0.00$	-0.01
Jet energy scale (flavour)	$+0.01$	-0.01
Jet energy scale (pile-up)	$+0.00$	-0.01
Jet energy scale (remaining)	$+0.01$	-0.01
Jet energy resolution	$+0.02$	-0.02
Luminosity	$+0.01$	-0.00
Other sources	$+0.03$	-0.03
Total systematic uncertainty	$+0.30$	-0.28
$t\bar{t}+\geq 1b$ normalisation	$+0.04$	-0.07
Total statistical uncertainty	$+0.20$	-0.20
Total uncertainty	$+0.36$	-0.34

arXiv:2111.06712

ATLAS VH(cc)

Source of uncertainty		$\mu_{VH(c\bar{c})}$
Total		15.3
Statistical		10.0
Systematic		11.5
Statistical uncertainties		
Signal normalisation		7.8
Other normalisations		5.1
Theoretical and modelling uncertainties		
$VH(\rightarrow c\bar{c})$		2.1
$Z+$jets		7.0
Top quark		3.9
$W+$jets		3.0
Diboson		1.0
$VH(\rightarrow b\bar{b})$		0.8
Multi-jet		1.0
Simulation samples size		4.2
Experimental uncertainties		
Jets		2.8
Leptons		0.5
E_T^{miss}		0.2
Pile-up and luminosity		0.3
Flavour tagging	c-jets	1.6
	b-jets	1.1
	light-jets	0.4
	τ-jets	0.3
Truth-flavour tagging	ΔR correction	3.3
	Residual non-closure	1.7

arXiv:2201.11428

Modelling the background

- **MC simulation** is a commonly used technique
  - not always possible to model the background with sufficient accuracy → significant **theoretical uncertainties**
  - Often computationally costly to produce large samples → significant **statistical uncertainties**

These uncertainties become more and more relevant as larger datasets become available

ATLAS ttH(bb)

Uncertainty source	$\Delta\mu$	
Process modelling		
$t\bar{t}H$ modelling	+0.13	−0.05
$t\bar{t}+\geq 1b$ modelling		
$\boxed{t\bar{t}+\geq 1b \text{ NLO matching}}$	+0.21	−0.20
$t\bar{t}+\geq 1b$ fractions	+0.12	−0.12
$t\bar{t}+\geq 1b$ FSR	+0.10	−0.11
$t\bar{t}+\geq 1b$ PS & hadronisation	+0.09	−0.08
$t\bar{t}+\geq 1b$ p_{T}^{bb} shape	+0.04	−0.04
$t\bar{t}+\geq 1b$ ISR	+0.04	−0.04
$t\bar{t}+\geq 1c$ modelling	+0.03	−0.04
$t\bar{t}+$ light modelling	+0.03	−0.03
tW modelling	+0.08	−0.07
Background-model statistical uncertainty	+0.04	−0.05
b-tagging efficiency and mis-tag rates		
b-tagging efficiency	+0.03	−0.02
c-mis-tag rates	+0.03	−0.03
l-mis-tag rates	+0.02	−0.02
Jet energy scale and resolution		
b-jet energy scale	+0.00	−0.01
Jet energy scale (flavour)	+0.01	−0.01
Jet energy scale (pile-up)	+0.00	−0.01
Jet energy scale (remaining)	+0.01	−0.01
Jet energy resolution	+0.02	−0.02
Luminosity	+0.01	−0.00
Other sources	+0.03	−0.03
Total systematic uncertainty	+0.30	−0.28
$t\bar{t}+\geq 1b$ normalisation	+0.04	−0.07
Total statistical uncertainty	+0.20	−0.20
Total uncertainty	+0.36	−0.34

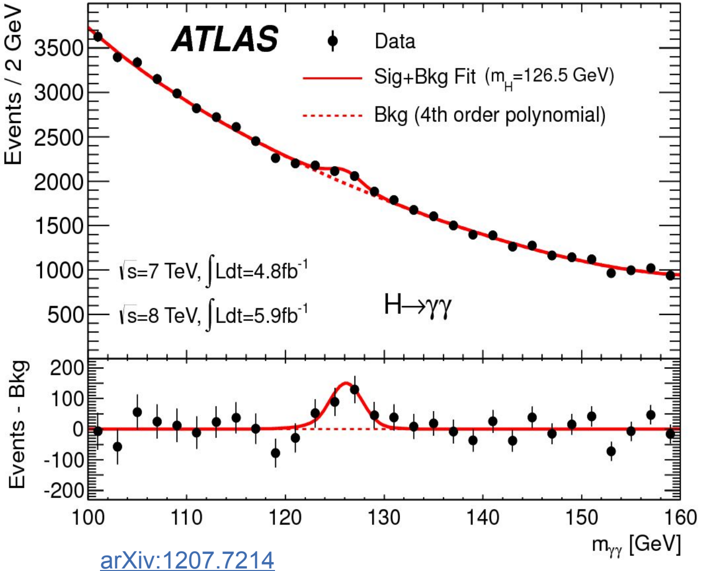
arXiv:2111.06712

ATLAS VH(cc)

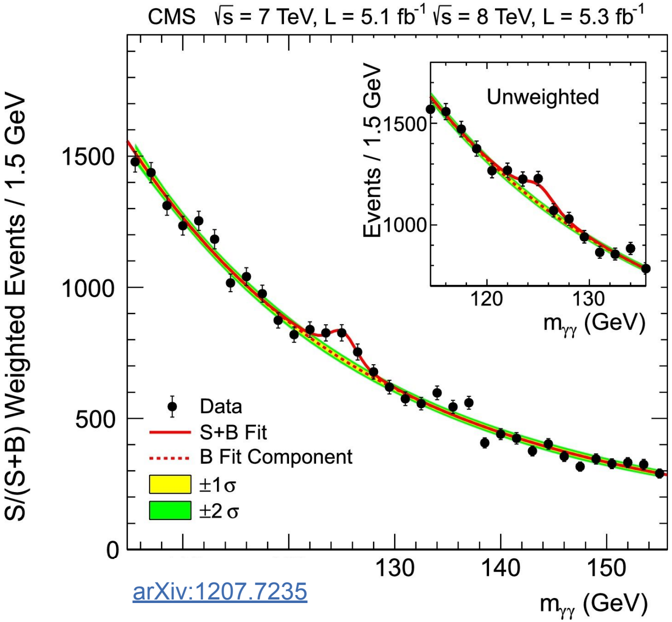
Source of uncertainty		$\mu_{VH(c\bar{c})}$
Total		15.3
Statistical		10.0
Systematic		11.5
Statistical uncertainties		
Signal normalisation		7.8
Other normalisations		5.1
Theoretical and modelling uncertainties		
$VH(\rightarrow c\bar{c})$		2.1
Z + jets		7.0
Top quark		3.9
W + jets		3.0
Diboson		1.0
$VH(\rightarrow b\bar{b})$		0.8
Multi-jet		1.0
Simulation samples size		4.2
Experimental uncertainties		
Jets		2.8
Leptons		0.5
$E_{\mathrm{T}}^{\mathrm{miss}}$		0.2
Pile-up and luminosity		0.3
Flavour tagging	c-jets	1.6
	b-jets	1.1
	light-jets	0.4
	τ-jets	0.3
Truth-flavour tagging	ΔR correction	3.3
	Residual non-closure	1.7

arXiv:2201.11428

Searching for H→γγ



arXiv:1207.7214



arXiv:1207.7235

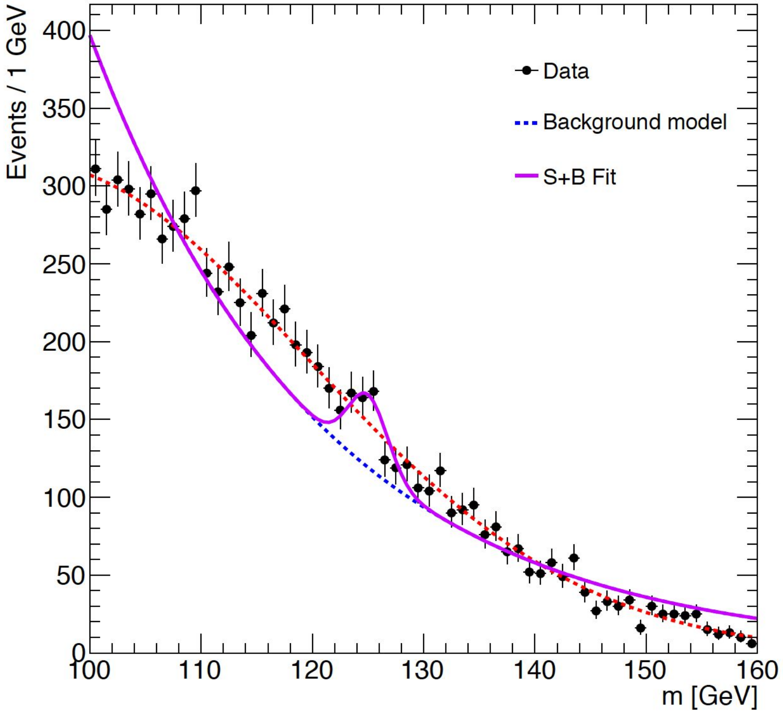
- Backgrounds arising from **di-jet**, **jet+photon** and di-photon processes
- Both experiments use **parametric models**

Spurious Signal

- ATLAS performs "**spurious-signal**" calculations
  - Test for **bias** in signal extraction arising from choice of functional form

Spurious Signal

- ATLAS performs "**spurious-signal**" calculations
  - Test for **bias** in signal extraction arising from choice of functional form
  - Fit **background-only Monte Carlo** samples with S+B model for different background models:
    - No background in sample, so take signal yield (N_{SP}) from fit as estimate for bias for specific model being tested
    - Ultimately pick model with **lowest N_{SP}**



Spurious Signal

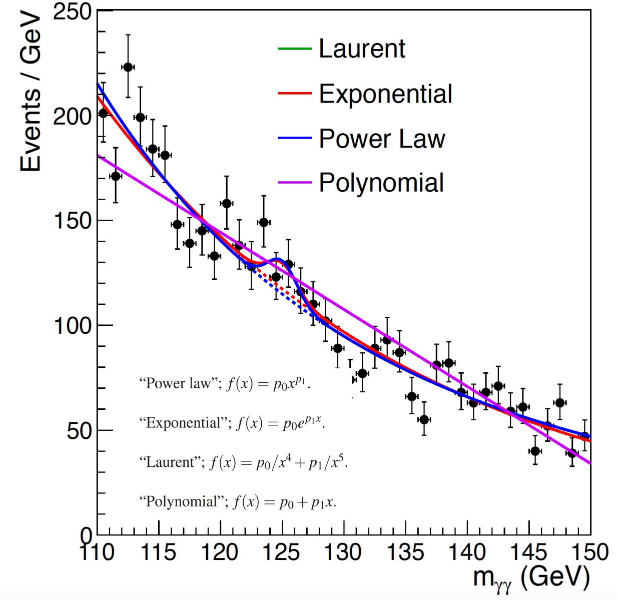
- ATLAS performs “**spurious-signal**” calculations
  - Test for **bias** in signal extraction arising from choice of functional form
  - Fit **background-only Monte Carlo** samples with S+B model for different background models:
    - No background in sample, so take signal yield (N_{SP}) from fit as estimate for bias for specific model being tested
    - Ultimately pick model with **lowest N_{SP}**
    - N_{SP} taken as **systematic uncertainty**

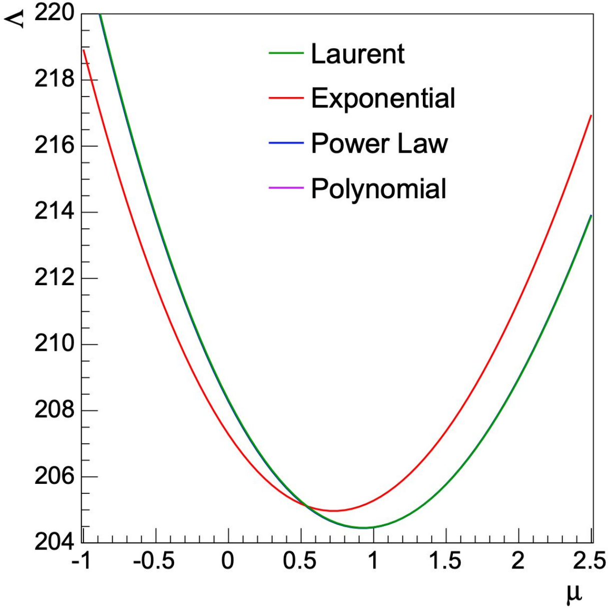
Source	Uncertainty (%)
Fit (stat.)	10
Fit (syst.)	8.3
Photon energy scale & resolution	4.0
Background modeling (spurious signal)	7.3
Correction factor	5.2
Photon isolation efficiency	4.6
Pileup	1.9
Photon ID efficiency	1.3
Trigger efficiency	0.7
Dalitz Decays	0.4
Theoretical modeling	$^{+0.3}_{-0.4}$
Diphoton vertex selection	0.1
Photon energy scale & resolution	0.1
Luminosity	2.0
Total	14

ATLAS-CONF-2018-028

Discrete profiling method

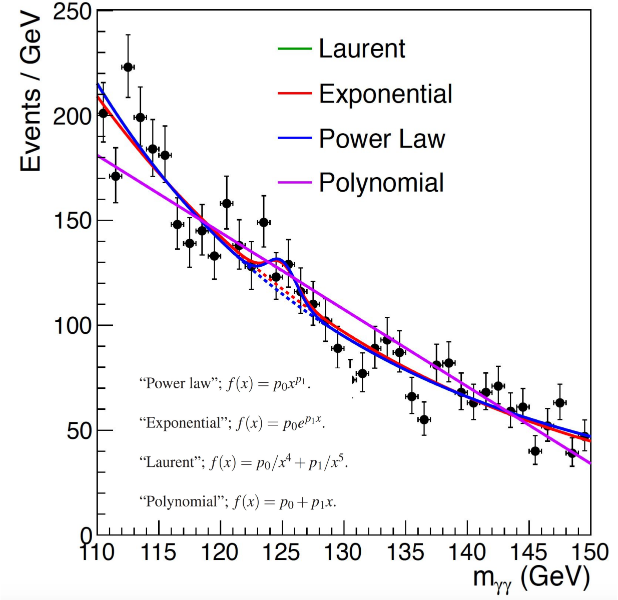
- CMS uses the discrete profiling of ensemble of parametric forms [arXiv:1408.6865]
  - Choice of functional form treated as a discrete nuisance parameter

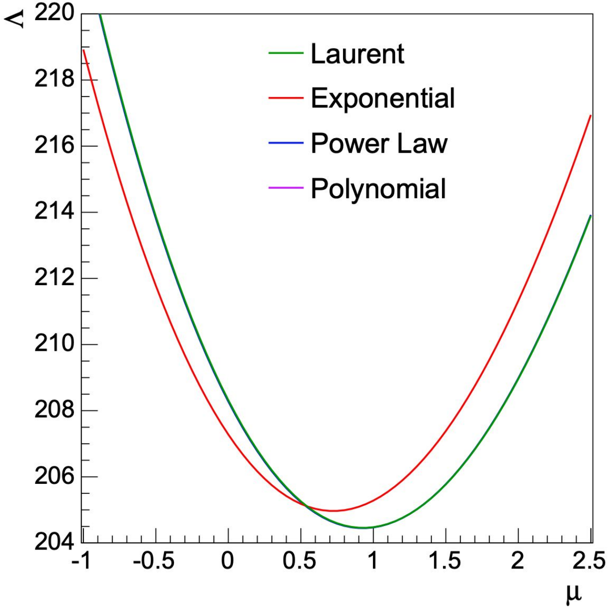


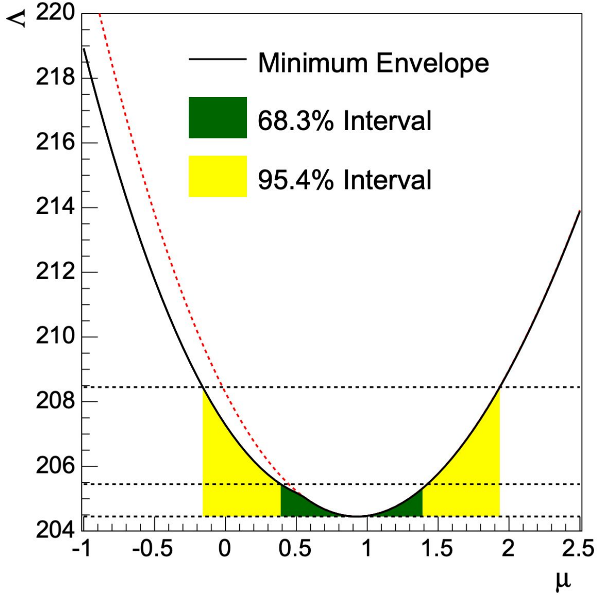


Discrete profiling method

- CMS uses the discrete profiling of ensemble of parametric forms [[arXiv:1408.6865](https://arxiv.org/abs/1408.6865)]
  - Choice of functional form treated as a discrete nuisance parameter
  - Minimum envelope of individual likelihood scans gives overall likelihood profile

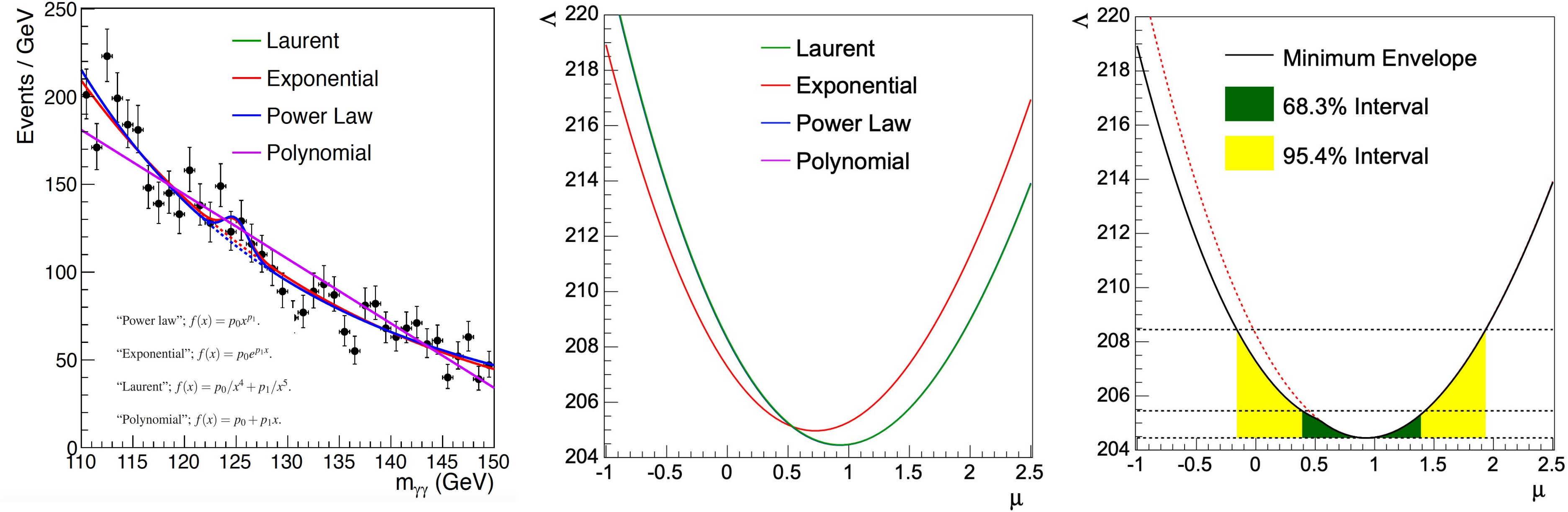






Discrete profiling method

- CMS uses the discrete profiling of ensemble of parametric forms [arXiv:1408.6865]
  - Choice of functional form treated as a discrete nuisance parameter
  - Minimum envelope of individual likelihood scans gives overall likelihood profile
  - Correction to penalise models with more parameters



Beyond parametric methods

- Some conceptual and practical complications:
  - **Spurious signal calculations**
    - Use samples that were considered not reliable to model the background
    - Need high statistics samples, which are not always available
  - **Discrete profiling method**
    - Dealing with common systematic effects across categories
      - All possible combinations of each function in each category must be fitted
      - Approximations have to be taken

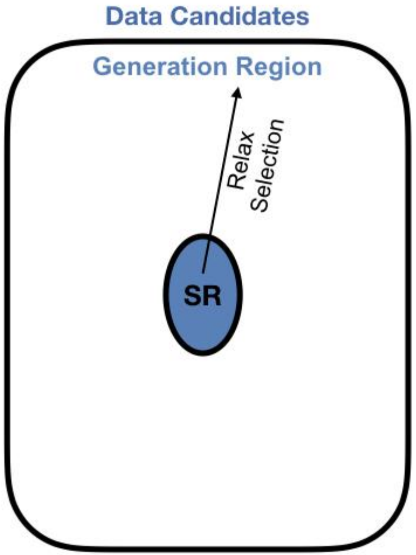
- Today will present a novel **non-parametric data-driven background modelling** technique
  - 2 different implementations through :
    - **ancestral sampling** (exemplified with $H \rightarrow \phi\gamma$ case study)
    - **generative adversarial networks** (exemplified with $H \rightarrow Za$ case study)

arXiv:2112.00650

Non-parametric data-driven background modelling

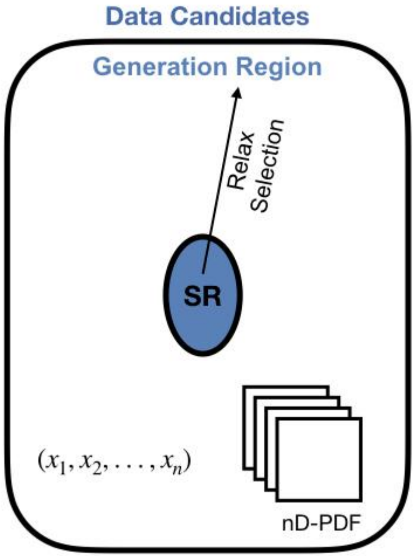
Non-parametric data-driven background modelling

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



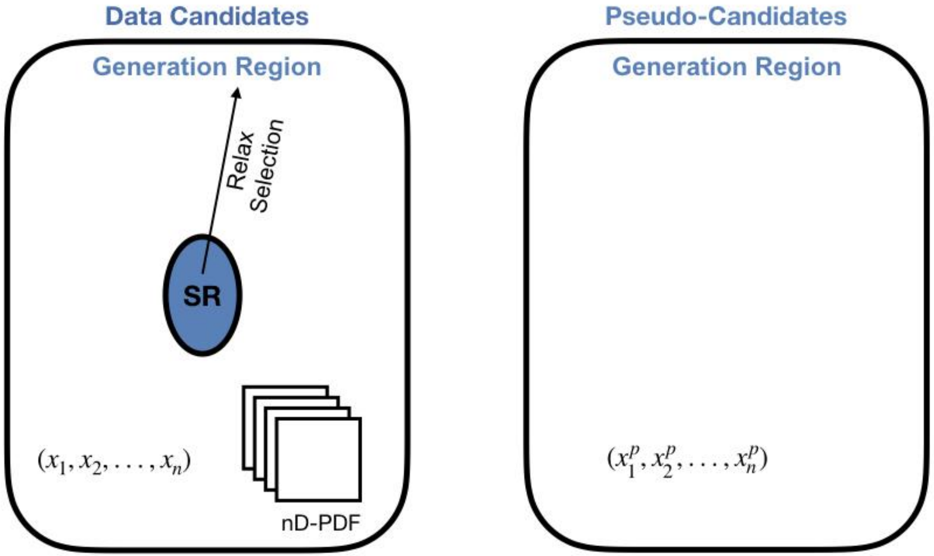
Non-parametric data-driven background modelling

1. Obtain sample of data events enriched in background by relaxing event selection requirements (**Generation Region**)
2. Obtain conditional PDF of relevant variables (x1, x2,…, xn)



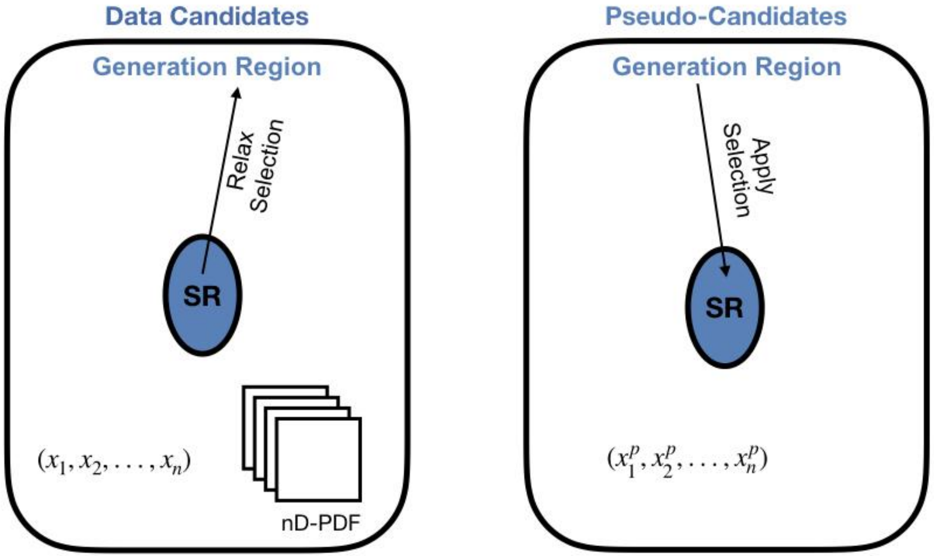
Non-parametric data-driven background modelling

1. Obtain sample of data events enriched in background by relaxing event selection requirements (**Generation Region**)
2. Obtain conditional PDF of relevant variables (x1, x2,…, xn)
3. Generate sample of pseudo-candidates



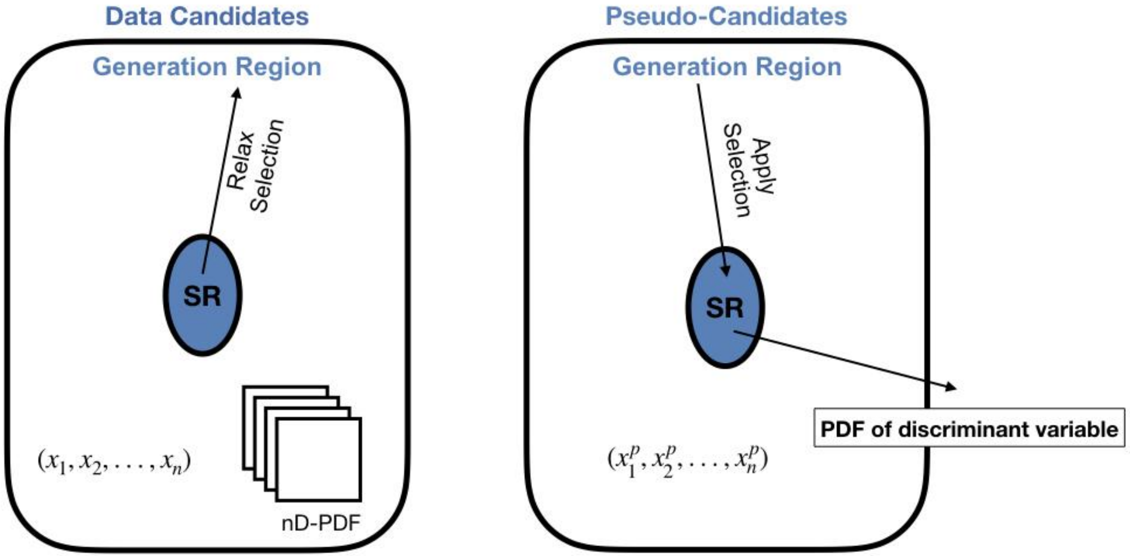
Non-parametric data-driven background modelling

1. Obtain sample of data events enriched in background by relaxing event selection requirements (**Generation Region**)
2. Obtain conditional PDF of relevant variables (x1, x2,…, xn)
3. Generate sample of pseudo-candidates
4. Apply Signal Region requirements to pseudo-candidates sample



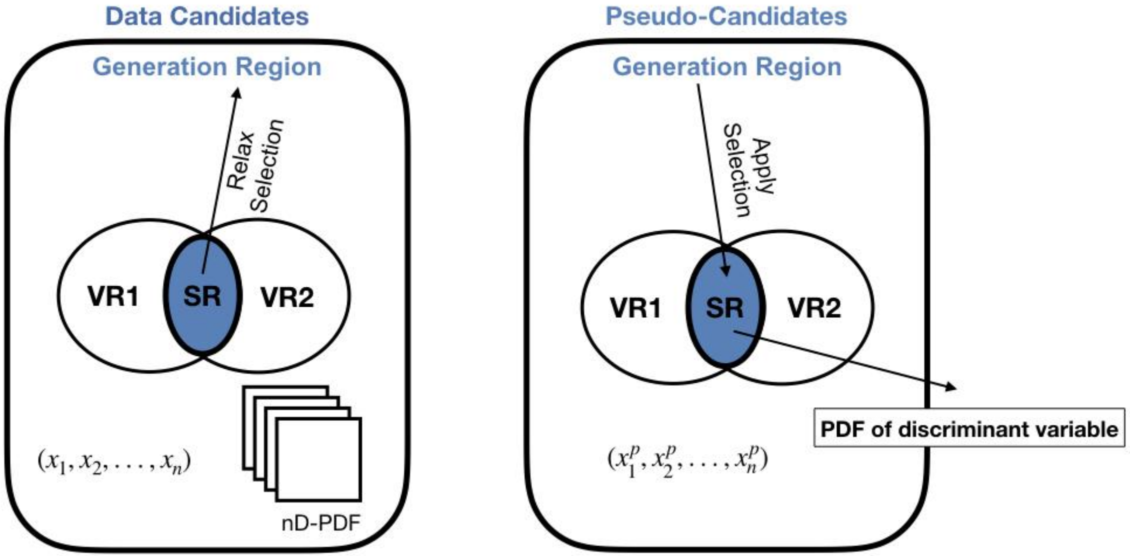
Non-parametric data-driven background modelling

1. Obtain sample of data events enriched in background by relaxing event selection requirements (**Generation Region**)
2. Obtain conditional PDF of relevant variables (x1, x2,…, xn)
3. Generate sample of pseudo-candidates
4. Apply **Signal Region** requirements to pseudo-candidates sample - obtain PDF of **discriminant variable** for statistical analysis



Non-parametric data-driven background modelling

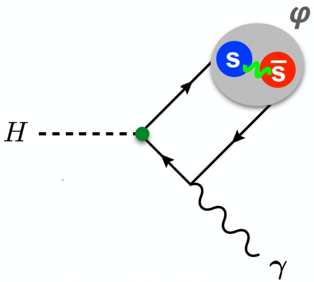
1. Obtain sample of data events enriched in background by relaxing event selection requirements (**Generation Region**)
2. Obtain conditional PDF of relevant variables (x1, x2,…, xn)
3. Generate sample of pseudo-candidates
4. Apply **Signal Region** requirements to pseudo-candidates sample - obtain PDF of **discriminant variable** for statistical analysis
   - Intermediate Validation Regions to check method



Ancestral Sampling

Case Study: H→Φγ

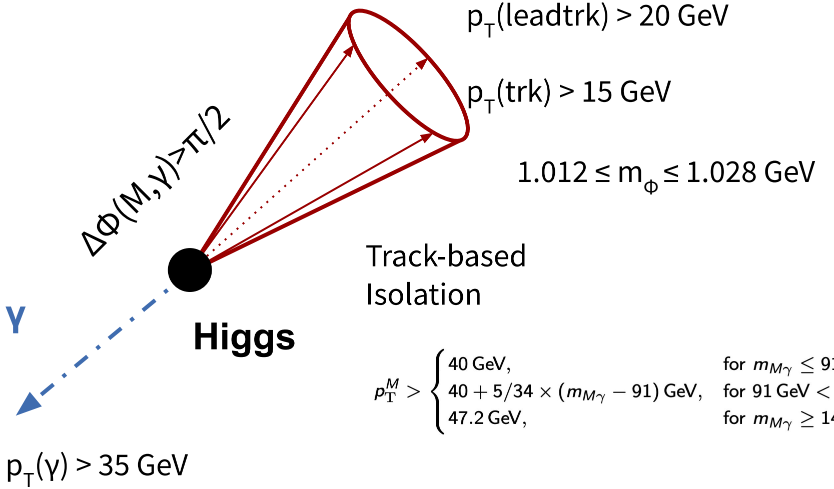
➔ H→ϕ(K^+K^-)γ suggested as probe of **Higgs coupling to strange quark** (arXiv:1406.1722)



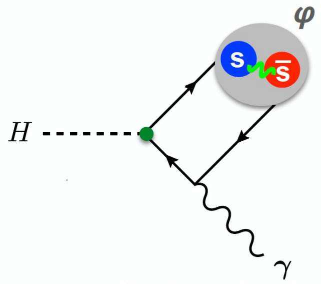
$$BR(h \to \phi\gamma) = (2.31 \pm 0.03_{f_\phi} \pm 0.11_{h\to\gamma\gamma}) \cdot 10^{-6}$$

Case Study: H→Φγ

- ➔ H→ϕ(K^+K^-)γ suggested as probe of **Higgs coupling to strange quark** (arXiv:1406.1722)
  - ◆ **Distinct experimental signature**: pair of collimated high-p_T isolated tracks recoiling against isolated photon



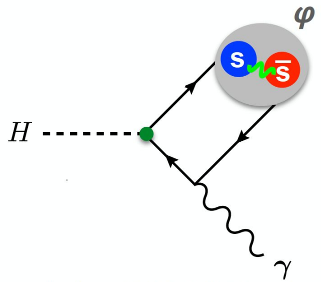
$$p_T^M > \begin{cases} 40\text{ GeV}, & \text{for } m_{M\gamma} \leq 91\text{ GeV} \\ 40 + 5/34 \times (m_{M\gamma} - 91)\text{ GeV}, & \text{for } 91\text{ GeV} < m_{M\gamma} < 140\text{ GeV} \\ 47.2\text{ GeV}, & \text{for } m_{M\gamma} \geq 140\text{ GeV} \end{cases}$$



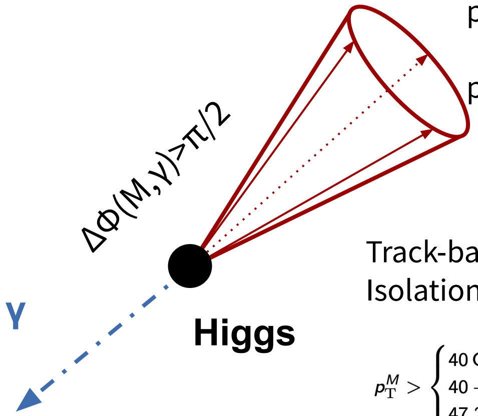
$$BR(h \to \phi\gamma) = (2.31 \pm 0.03_{f_\phi} \pm 0.11_{h\to\gamma\gamma}) \cdot 10^{-6}$$

Case Study: H→Φγ

- H→ϕ(K^+K^-)γ suggested as probe of **Higgs coupling to strange quark** (arXiv:1406.1722)
  - **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



$$BR(h \to \phi\gamma) = (2.31 \pm 0.03_{f_\phi} \pm 0.11_{h\to\gamma\gamma}) \cdot 10^{-6}$$



p_T(leadtrk) > 20 GeV

p_T(trk) > 15 GeV

$1.012 \leq m_\phi \leq 1.028$ GeV

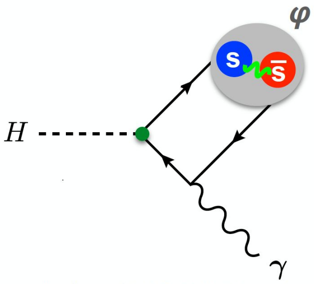
Track-based Isolation

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

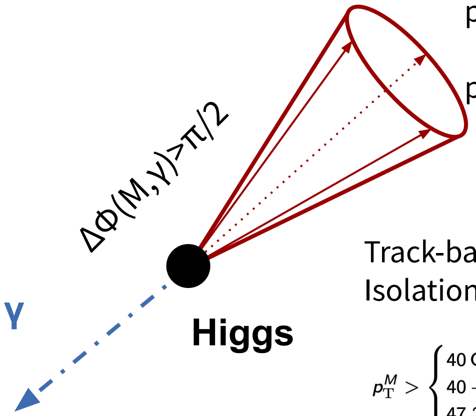
$p_T(\gamma)$ > 35 GeV

Case Study: H→Φγ

- ➔ H→ϕ(K^+K^-)γ suggested as probe of **Higgs coupling to strange quark** (arXiv:1406.1722)
  - ◆ **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



$$BR(h \to \phi\gamma) = (2.31 \pm 0.03_{f_\phi} \pm 0.11_{h\to\gamma\gamma}) \cdot 10^{-6}$$



p_T(leadtrk) > 20 GeV

p_T(trk) > 15 GeV

$1.012 \leq m_\phi \leq 1.028$ GeV

Track-based Isolation

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

$p_T(\gamma)$ > 35 GeV

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 variables following most important correlations
   - 1D, 2D and 3D histograms to be sampled from in generation step

Building the model for H→Φγ

Relax Selection

Obtain Conditional PDFs

Generate pseudo candidates

Apply Selection

2. Build PDFs of relevant variables following most important correlations
   - 1D, 2D and 3D histograms to be sampled from in generation step

Which variables do we need in H→ϕγ case?

Building the model for H→Φγ

Relax Selection

Obtain Conditional PDFs

Generate pseudo candidates

Apply Selection

2. Build PDFs of relevant variables following most important correlations
   - 1D, 2D and 3D histograms to be sampled from in generation step

Which variables do we need in H→ϕγ case?

ϕ and γ 4-momentum vectors to ultimately obtain **m(ϕγ)**
+ Iso(ϕ)

pT(Φ), pT(γ), ΔΦ(ϕ,γ), Δη(ϕ,γ), Iso(Φ)

⇒ **PDF of m(ϕγ)**

Building the model for H→Φγ

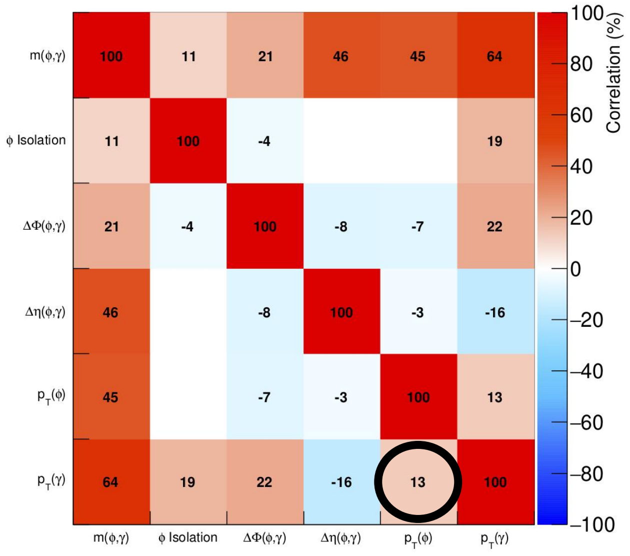
Relax Selection

Obtain Conditional PDFs

Generate pseudo candidates

Apply Selection

2. Build PDFs of kinematic and isolation variables following most important correlations
   - 1D, 2D and 3D histograms to be sampled from in generation step



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

Building the model for H→Φγ

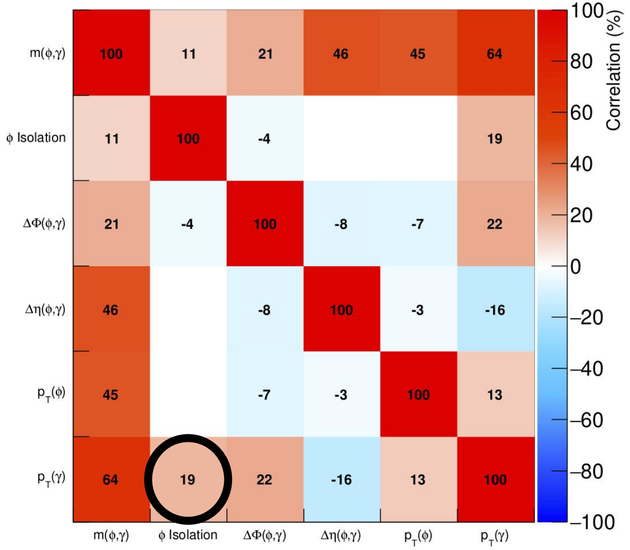
Relax Selection

Obtain Conditional PDFs

Generate pseudo candidates

Apply Selection

2. Build PDFs of kinematic and isolation variables following most important correlations
   - 1D, 2D and 3D histograms to be sampled from in generation step



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

$p_T(\gamma)$	ϕ Iso

Building the model for H→Φγ

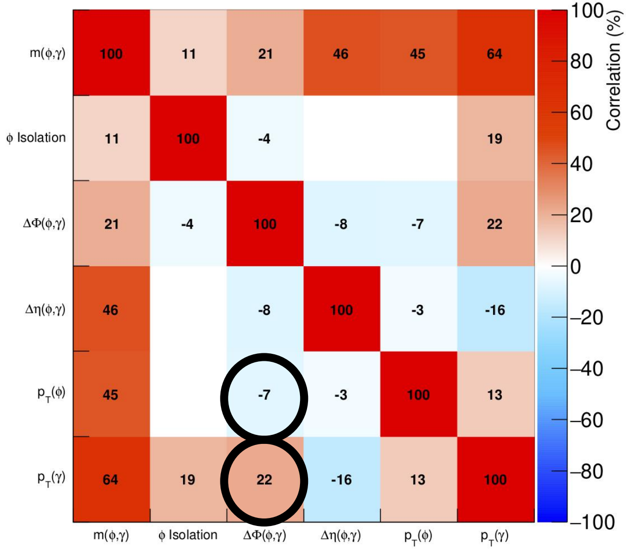
Relax Selection

Obtain Conditional PDFs

Generate pseudo candidates

Apply Selection

2. Build PDFs of kinematic and isolation variables following most important correlations
   - 1D, 2D and 3D histograms to be sampled from in generation step



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

$p_T(\gamma)$	ϕ Iso

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

Building the model for H→Φγ

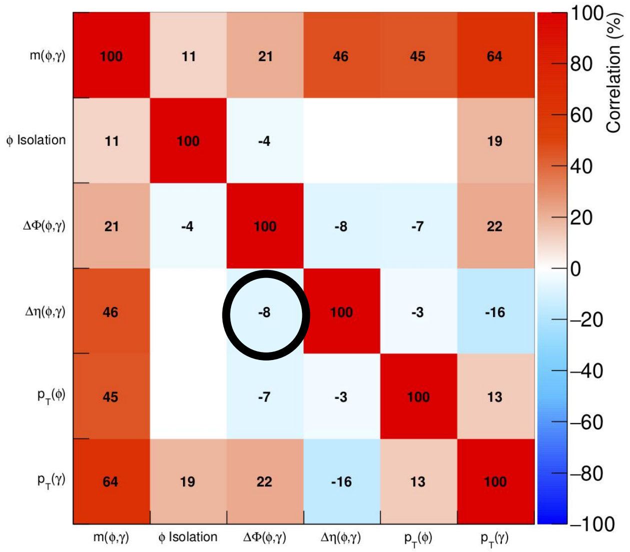
Relax Selection

Obtain Conditional PDFs

Generate pseudo candidates

Apply Selection

2. Build PDFs of kinematic and isolation variables following most important correlations
   - 2D and 3D histograms to be sampled from in generation step



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

$p_T(\gamma)$	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

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

$\phi = (\mathbf{p_T}, \eta, \Phi, m)$

$\gamma = (\mathbf{p_T}, \eta, \Phi, m)$

$\mathrm{Iso}(\phi)$

Building the model for H→Φγ

Relax Selection

Obtain Conditional PDFs

Generate pseudo candidates

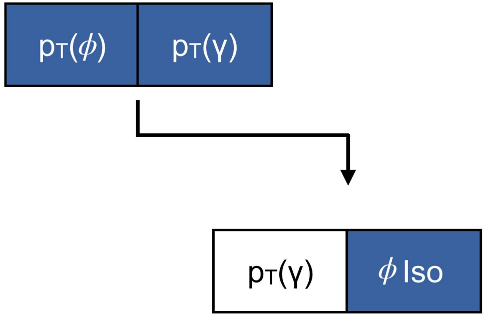
Apply Selection

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

$\phi = (\mathbf{p_T}, \eta, \Phi, m)$

$\gamma = (\mathbf{p_T}, \eta, \Phi, m)$

Iso(ϕ)



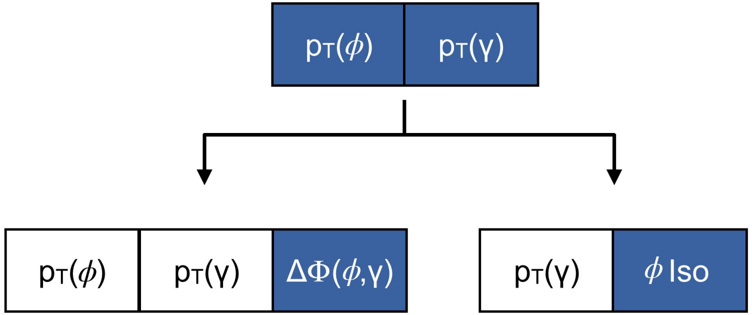
Building the model for H→Φγ

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

$$\phi = (\mathbf{p_T}, \eta, \mathbf{\Phi}, m)$$

$$\gamma = (\mathbf{p_T}, \eta, \mathbf{\Phi}, m)$$

$$\mathbf{Iso(\phi)}$$



Building the model for H→Φγ

Relax Selection

Obtain Conditional PDFs

Generate pseudo candidates

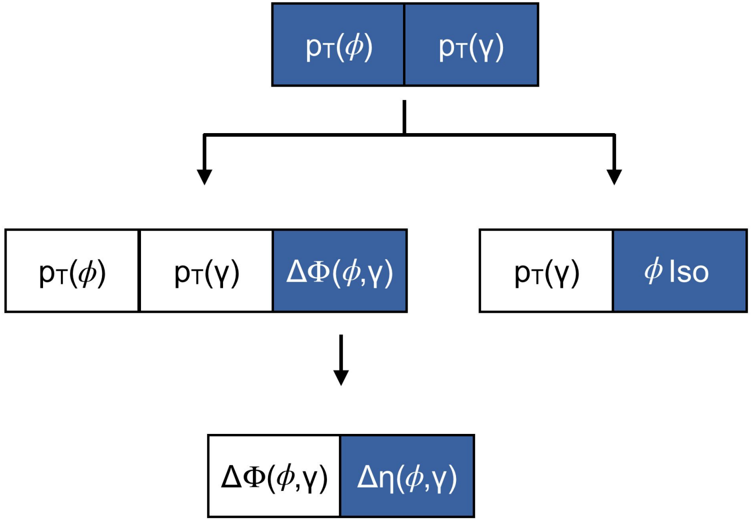
Apply Selection

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

$\phi = (\mathbf{p_T}, \boldsymbol{\eta}, \boldsymbol{\Phi}, m)$

$\gamma = (\mathbf{p_T}, \boldsymbol{\eta}, \boldsymbol{\Phi}, m)$

$\mathbf{Iso(\phi)}$



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

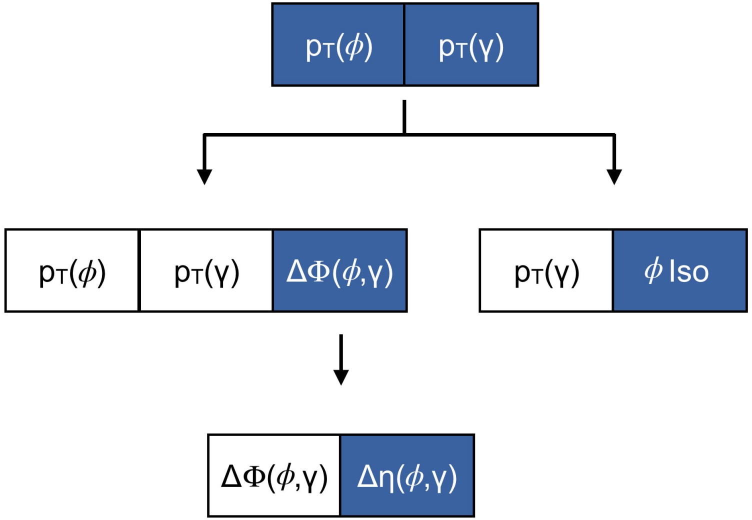
Higgs pseudo candidates

$\phi = (p_T, \eta, \Phi, m)$

$+$

$\gamma = (p_T, \eta, \Phi, m=0)$

$\mathrm{Iso}(\phi)$



Building the model for H→Φγ

Relax Selection

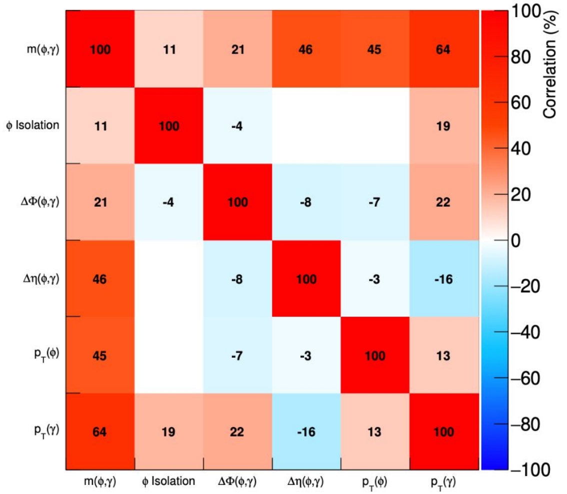
Obtain Conditional PDFs

Generate pseudo candidates

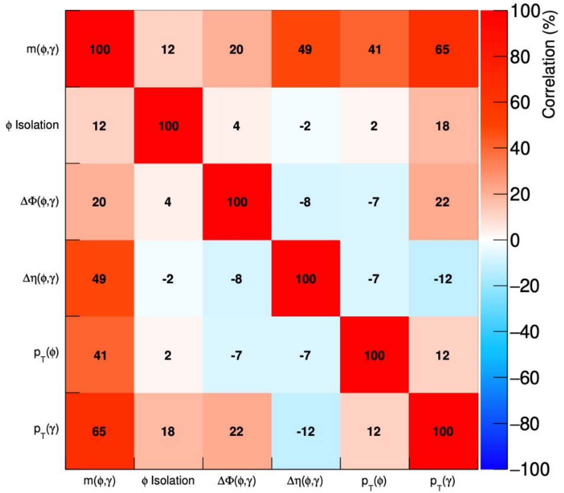
Apply Selection

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

γ+jet MC



Model



Building the model for H→Φγ

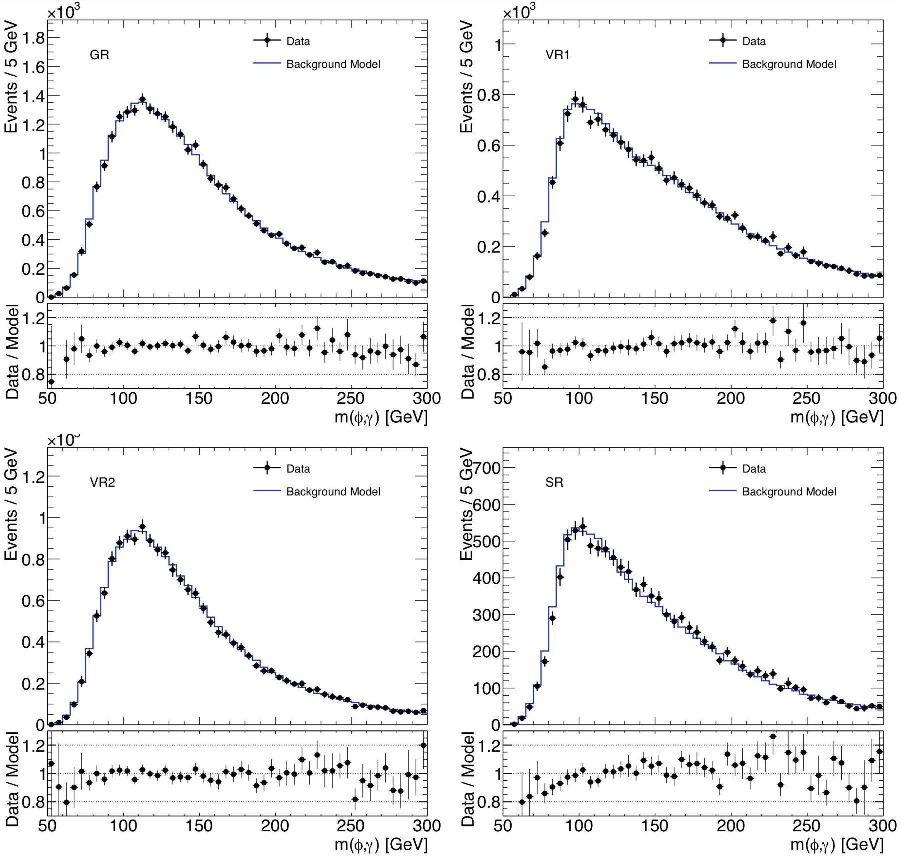
Relax Selection

Obtain Conditional PDFs

Generate pseudo candidates

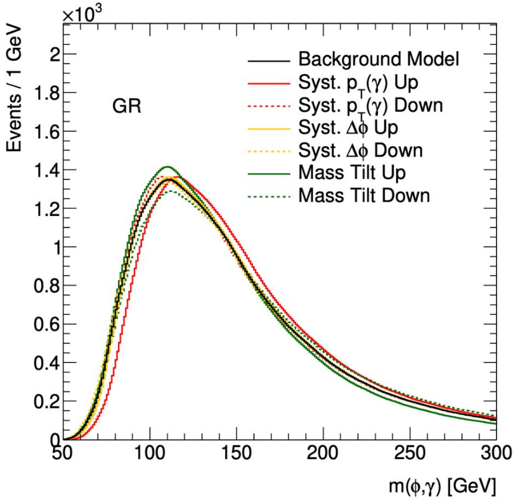
Apply Selection

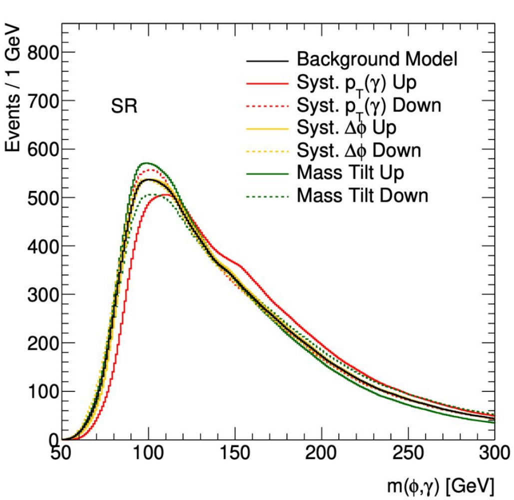
4. Apply $p_T(M)$ and Iso(M) requirements to sample of pseudo-candidates
   - ◆ obtain PDF of **m(ϕγ)** for statistical analysis in Signal and Validation Regions



Implementation in Statistical Analysis

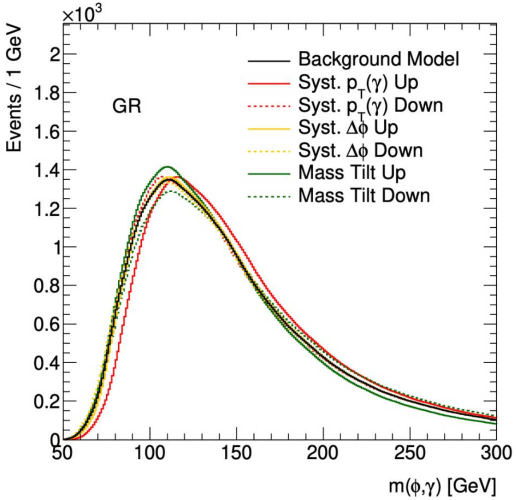
➔ **Systematic uncertainties** are provided through variations of the nominal PDFs
- selected to capture different modes of potential deformations of the background shape

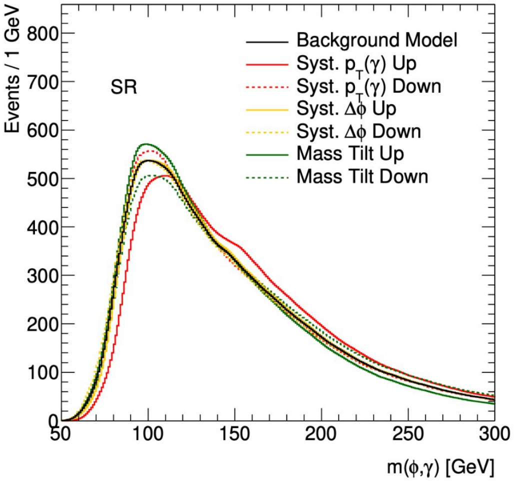


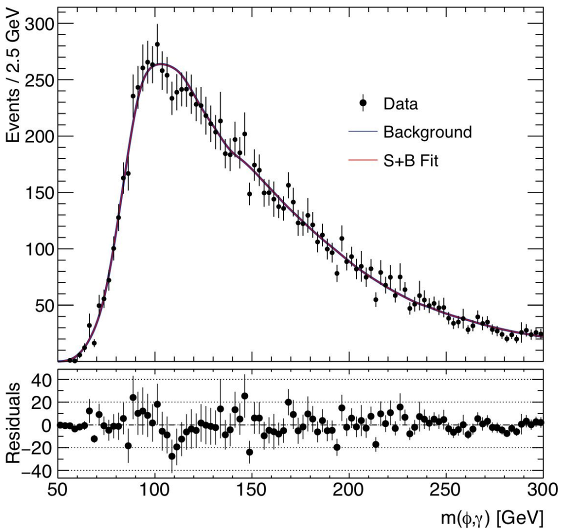


Implementation in Statistical Analysis

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



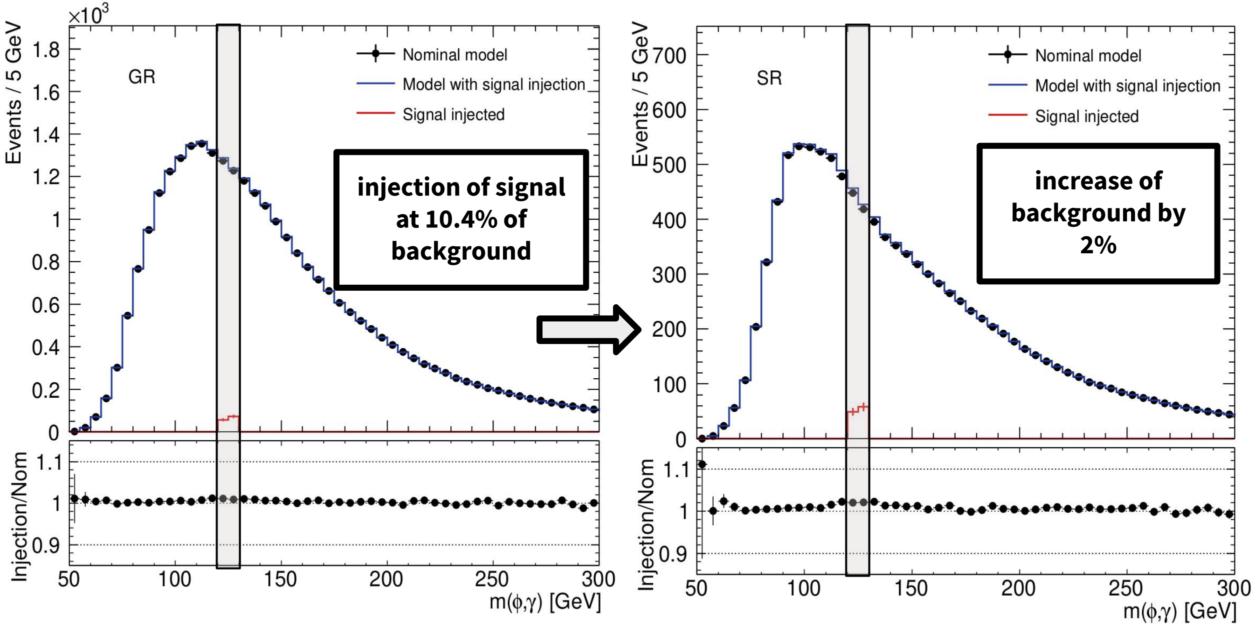




Parameter	Value	Uncertainty ($\pm 1\sigma$)
μ_{signal}	-0.07	± 0.54
μ_{bkgd}	1.01	± 0.01
Shape: $p_T(\gamma)$ shift	0.26	± 0.15
Shape: $\Delta\Phi(\phi, \gamma)$ tilt	0.30	± 0.43
Shape: $m(\phi, \gamma)$ tilt	0.10	± 0.24

Signal contamination test

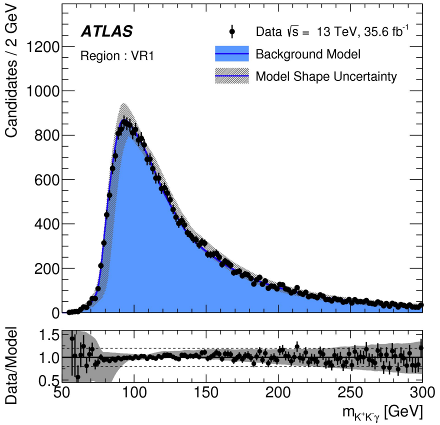
➔ **Robust** under signal contamination:
- Features of resonant contributions are diluted by process of factorising the background PDF
- Means that resonant backgrounds need to be modelled separately

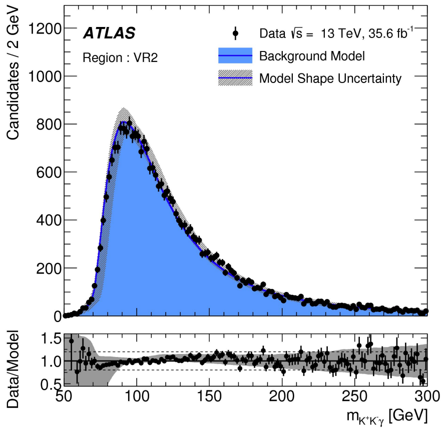


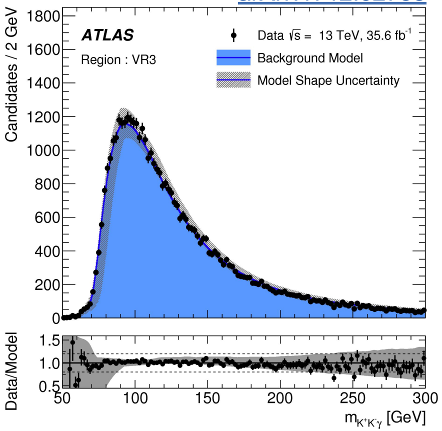
H/Z→Φγ Analysis

Model in Validation Regions

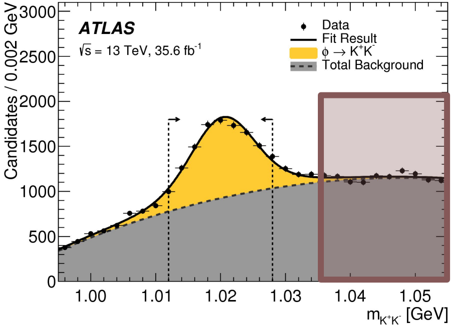
arXiv:1712.02758

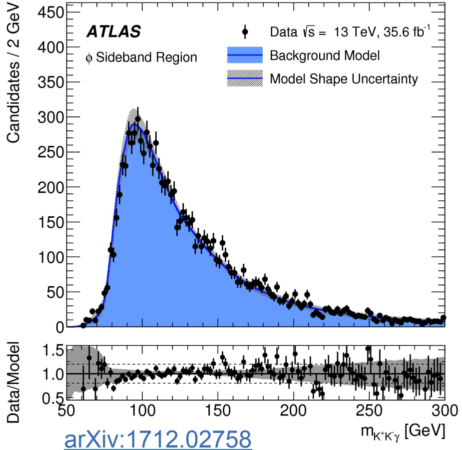






Validation with Φ sideband





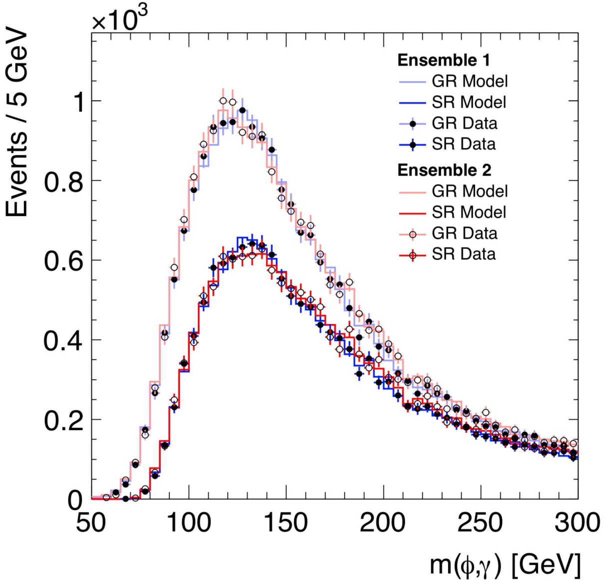
arXiv:1712.02758

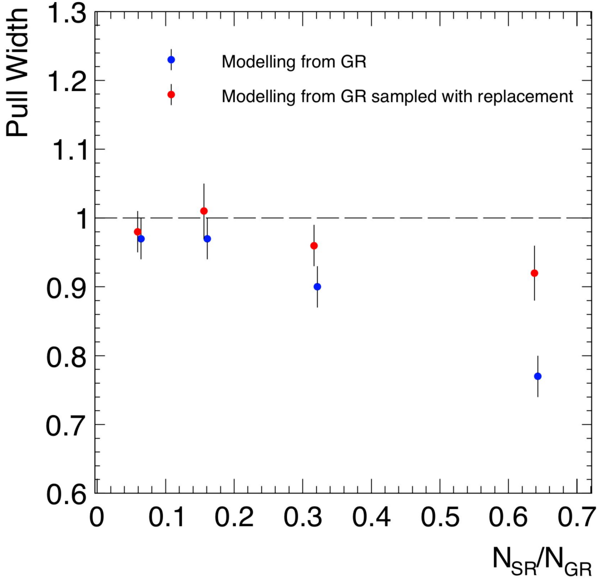
➔ **Model used in several other exclusive Higgs analyses already!** [Phys. Rev. Lett. 114 (2015) 121801, Phys. Rev. Lett. 117, 111802 (2016), JHEP 07 (2018) 127, Phys. Lett. B 786 (2018) 134]

Sampling from GR

- ➔ Events in SR are not independent from events in GR:
  - ◆ Adds information on fluctuations of each ensemble
  - ◆ Effect scales with the ratio of the number of events in SR over GR

- ➔ Leads to overestimation of signal strength statistical uncertainty for analyses in which low N_{SR}/N_{GR} can't be achieved:
  - ◆ This effect can be removed by building the model through sampling with replacement from GR
  - ◆ Statistical uncertainty can be corrected through toy MC studies



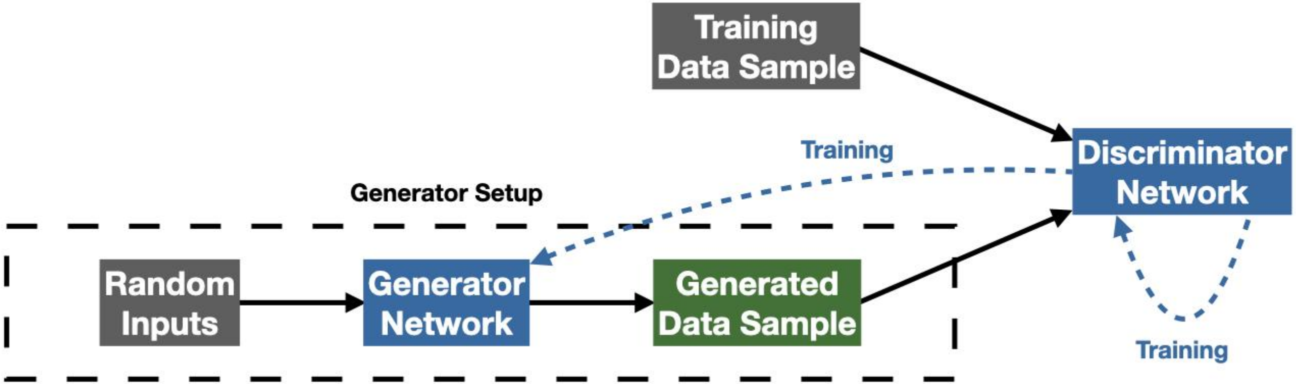


Conditional Generative Adversarial Networks

Generative Adversarial Networks

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

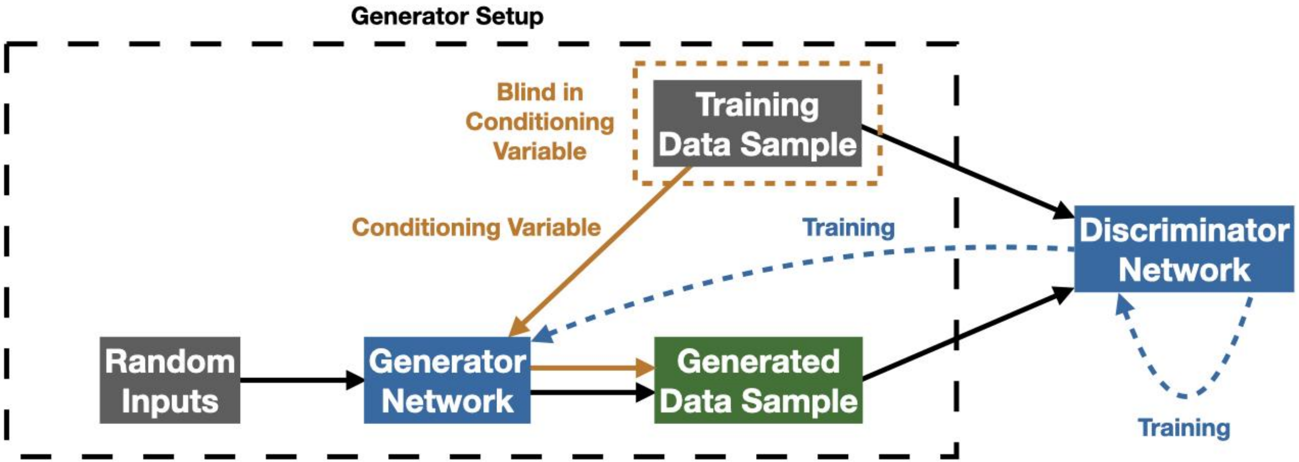
➔ Generalisation of method: use **GANs trained on data** to produce background model
  - **Generator** - learns generative model from data sample
  - **Discriminator** - simultaneously trained to discriminate the generator output from data



Conditional Generative Adversarial Networks

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

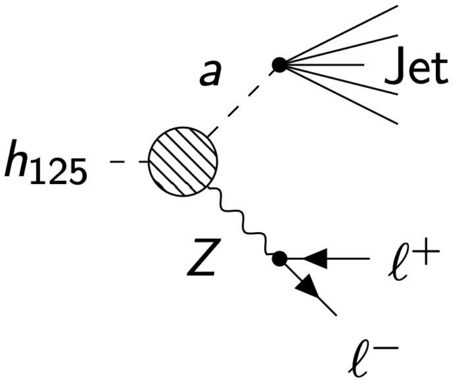
- **Condition** GAN (cGAN) on a blinding variable, allowing **SR to be blinded during training** - cGAN interpolates prediction into SR



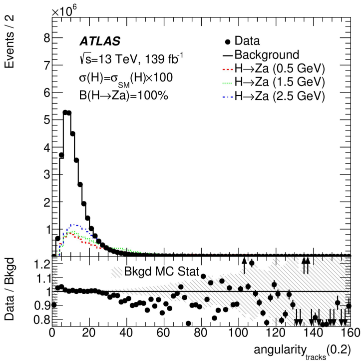
Case Study: H→Za

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

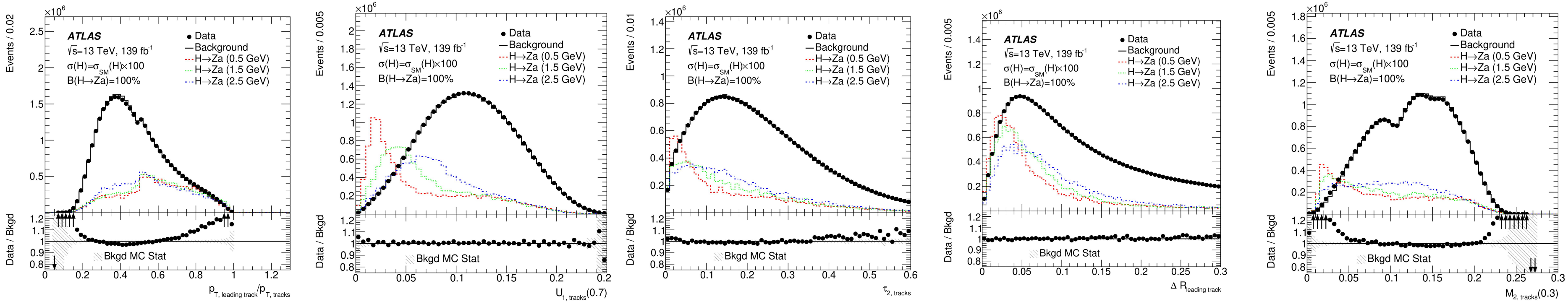
- Search for **H→Z(ll)+a**, with a→hadrons



Case Study: H→Za



- ➔ Light pseudo-scalars produced in Higgs decays feature in BSM theories 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

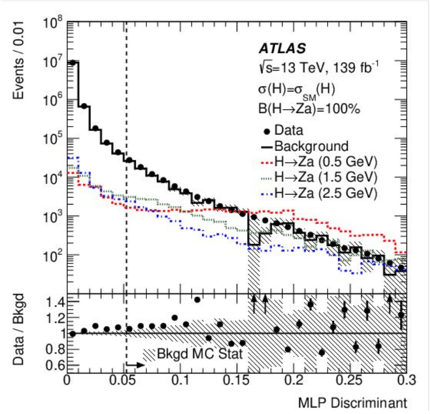


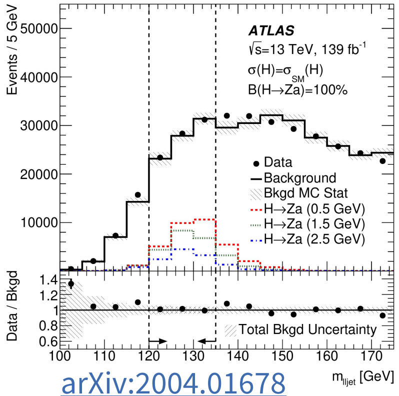
arXiv:2004.01678

Case Study: H→Za

- ➔ Light pseudo-scalars produced in Higgs decays feature in BSM theories 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
  - ◆ background estimation through modified ABDC method using mllj and MLP discriminant:
    - ● MC used to derive correction for correlation between variables





arXiv:2004.01678

Case Study: H→Za

- Light pseudo-scalars produced in Higgs decays feature in BSM theories 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
  - background estimation through modified ABDC method using mllj and MLP discriminant:
    - MC used to derive correction for correlation between variables
- ideal case study for implementation of background modelling using cGANs
  - background systematics arising by use of MC simulation (arXiv:2004.01678)
  - use of MVA techniques makes it impractical to use ancestral sampling

a mass	0.5 GeV	1.5 GeV	2.5 GeV
Total Uncertainty	8.3	10.7	20.3
Total Statistical Uncertainty	0.6	0.8	1.6
Total Systematic Uncertainty	8.2	10.7	20.2
Signal Systematic Uncertainties			
Jet Energy Scale	1.3	1.5	1.5
Parton Shower	1.4	1.4	1.4
Luminosity, Pileup, Trigger, Leptons, & JVT	0.2	0.3	0.5
MC Statistics	0.2	0.2	0.6
Renormalization Scale	0.1	< 0.1	0.2
Acceptance	0.1	< 0.1	0.2
Background Systematic Uncertainties			
MC Statistics	6.4	8.4	15.8
Parton Shower and ME	3.9	5.1	9.6
Renormalization Scale	3.4	4.4	8.3

arXiv:2004.01678

Use of GANs solves statistical limitations of background sample

Training on data avoids modelling limitations of MC

Case Study: H→Za

- ➔ Light pseudo-scalars produced in Higgs decays feature in BSM theories 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
  - ◆ background estimation through modified ABDC method using mllj and MLP discriminant:
    - ● MC used to derive correction for correlation between variables
- ➔ ideal case study for implementation of background modelling using cGANs
  - ◆ background systematics arising by use of MC simulation (arXiv:2004.01678)
  - ◆ use of MVA techniques makes it impractical to use ancestral sampling

a mass	0.5 GeV	1.5 GeV	2.5 GeV
Total Uncertainty	8.3	10.7	20.3
Total Statistical Uncertainty	0.6	0.8	1.6
Total Systematic Uncertainty	8.2	10.7	20.2
Signal Systematic Uncertainties			
Jet Energy Scale	1.3	1.5	1.5
Parton Shower	1.4	1.4	1.4
Luminosity, Pileup, Trigger, Leptons, & JVT	0.2	0.3	0.5
MC Statistics	0.2	0.2	0.6
Renormalization Scale	0.1	< 0.1	0.2
Acceptance	0.1	< 0.1	0.2
Background Systematic Uncertainties			
MC Statistics	6.4	8.4	15.8
Parton Shower and ME	3.9	5.1	9.6
Renormalization Scale	3.4	4.4	8.3

arXiv:2004.01678

Use of GANs solves statistical limitations of background sample
Training on data avoids modelling limitations of MC

Case Study: H→Za

- Light pseudo-scalars produced in Higgs decays feature in BSM theories 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
  - background estimation through modified ABDC method using mllj and MLP discriminant:
    - MC used to derive correction for correlation between variables
- ideal case study for implementation of background modelling using cGANs
  - background systematics arising by use of MC simulation (arXiv:2004.01678)
  - use of MVA techniques makes it impractical to use ancestral sampling
- **Z + jets MC sample** used to exemplify model application

a mass	0.5 GeV	1.5 GeV	2.5 GeV
Total Uncertainty	8.3	10.7	20.3
Total Statistical Uncertainty	0.6	0.8	1.6
Total Systematic Uncertainty	8.2	10.7	20.2
Signal Systematic Uncertainties			
Jet Energy Scale	1.3	1.5	1.5
Parton Shower	1.4	1.4	1.4
Luminosity, Pileup, Trigger, Leptons, & JVT	0.2	0.3	0.5
MC Statistics	0.2	0.2	0.6
Renormalization Scale	0.1	< 0.1	0.2
Acceptance	0.1	< 0.1	0.2
Background Systematic Uncertainties			
MC Statistics	6.4	8.4	15.8
Parton Shower and ME	3.9	5.1	9.6
Renormalization Scale	3.4	4.4	8.3

arXiv:2004.01678

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

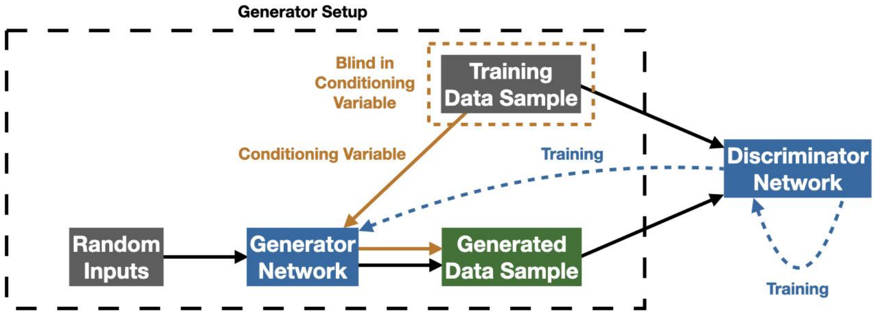
Obtain Conditional PDFs

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

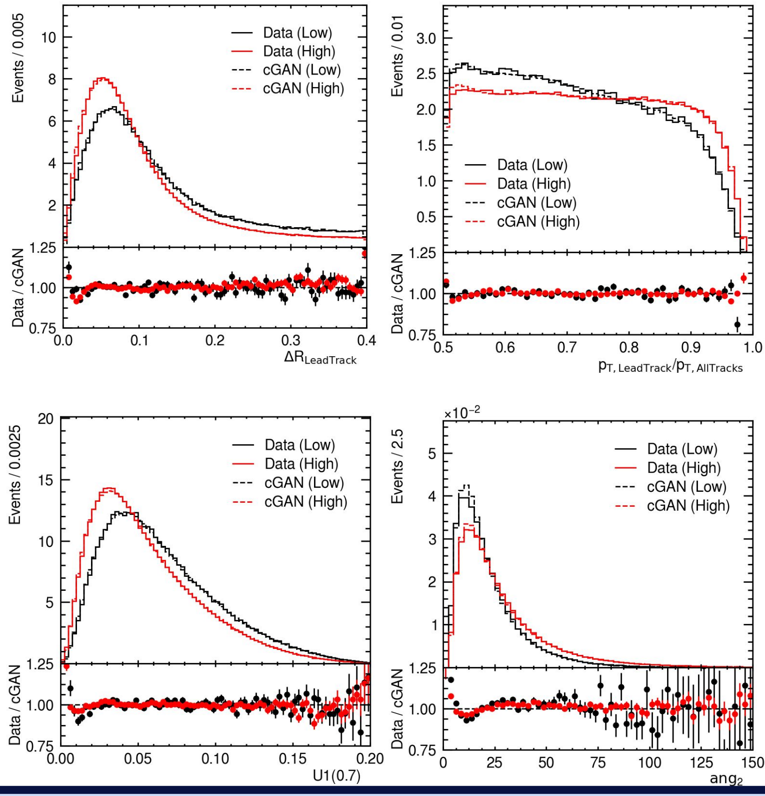
Obtain Conditional PDFs

Generate pseudo candidates

Apply Selection

3. Generate sample of pseudo-candidates:
   - input inclusive distribution of the conditioning variable into cGAN
   - cGAN **interpolates** 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

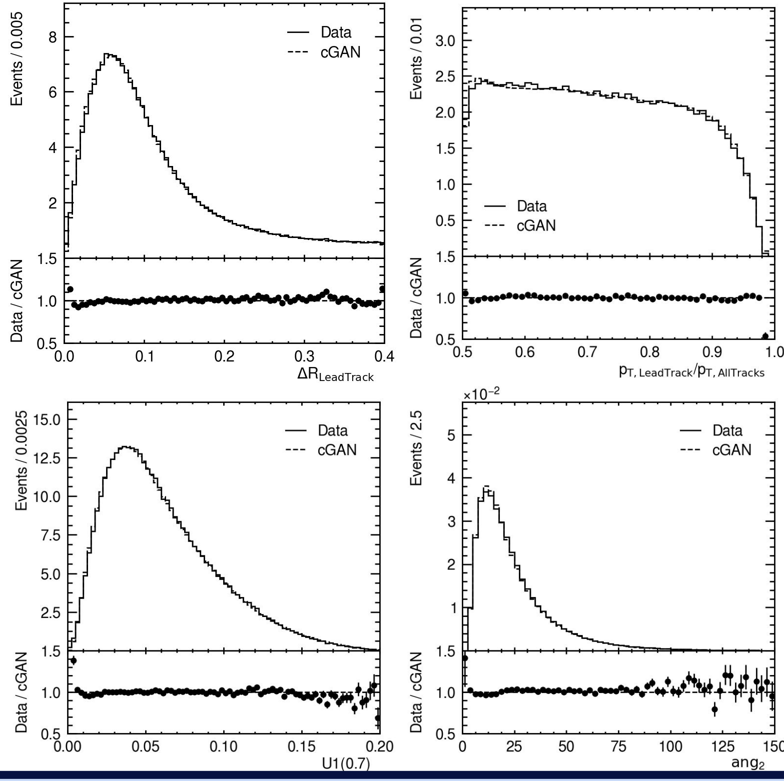
Obtain Conditional PDFs

Generate pseudo candidates

Apply Selection

3. Generate sample of pseudo-candidates:
   - input inclusive distribution of the conditioning variable into cGAN
   - cGAN **interpolates** the conditional generative model into signal region
   - obtain prediction of **MLP input variables**

$m_{\mu\mu j}$ SR



Building the model for H→Za

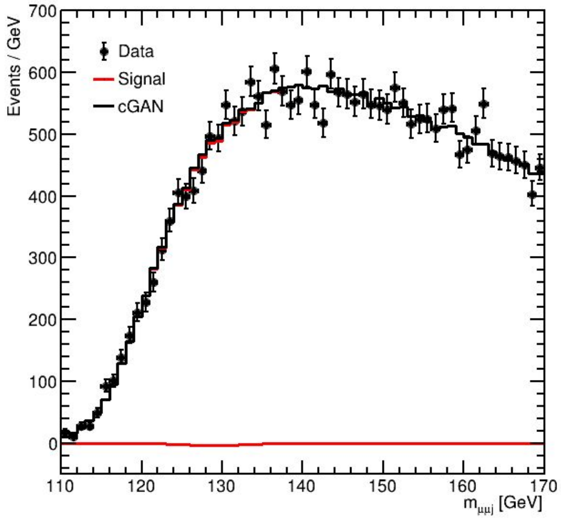
Relax Selection

Obtain Conditional PDFs

Generate pseudo candidates

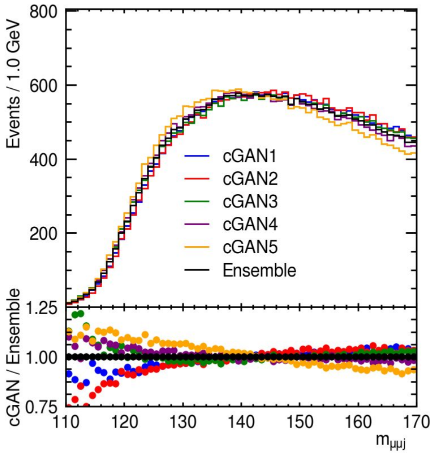
Apply Selection

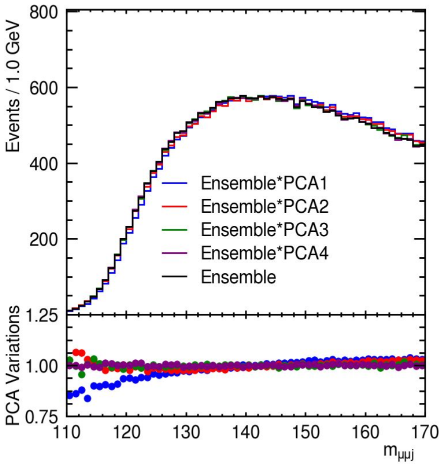
4. Apply MLP selection to pseudo-candidates sample
   - obtain PDF of $\mathbf{m}_{\boldsymbol{\mu\mu j}}$ in SR for statistical analysis



Implementation in Statistical Analysis

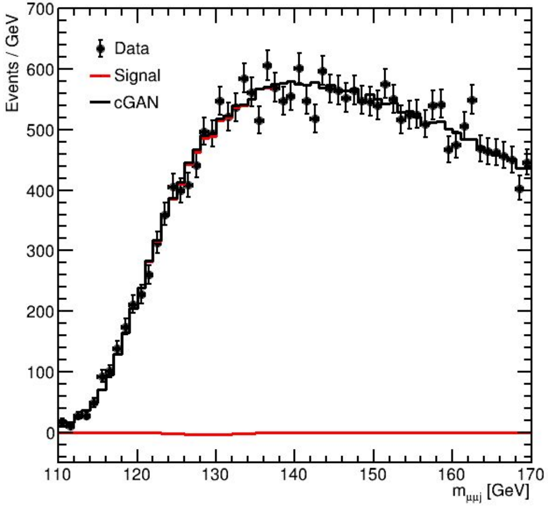
➔ **Systematic uncertainties** are provided through shape variations:
- Differences between ensemble and individual cGANs
- **Principal component analysis** performed to orthogonalise differences
- 2 biggest differences considered in statistical analysis

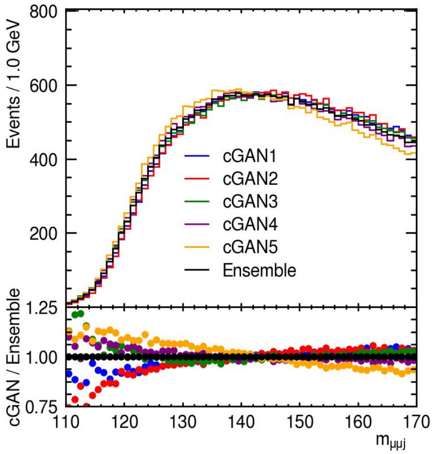


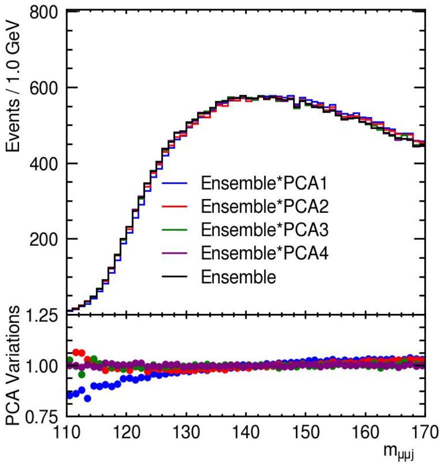


Implementation in Statistical Analysis

- **Systematic uncertainties** are provided through shape variations:
  - Differences between ensemble and individual cGANs
  - **Principal component analysis** performed to orthogonalise differences
  - 2 biggest differences considered in statistical analysis
- Binned 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

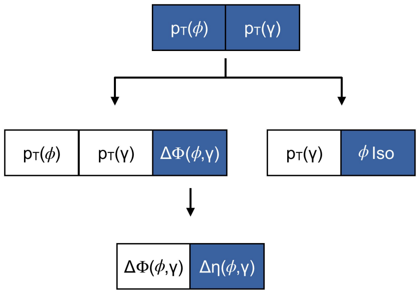
➔ A novel **non-parametric**, **data-driven** background modelling technique was presented

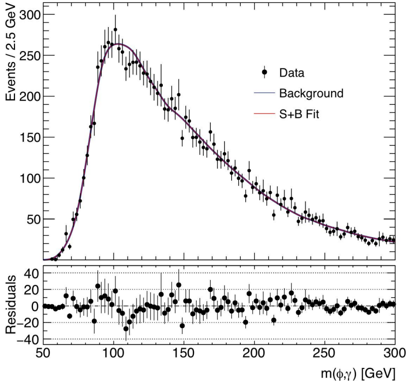
- Addresses typical shortcomings of often employed background modelling techniques
- Dataset from a **relaxed event selection** to create a model based on **conditional probabilities**
- Two distinct ways of building the conditional PDF:

arXiv:2112.00650

Ancestral sampling

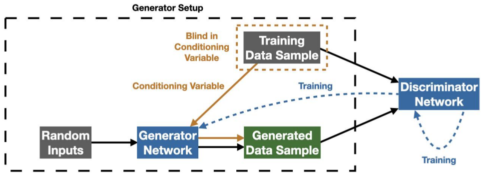
- Sample from histograms of relevant variables in data, built with respect to most important correlations
- Already used in multiple analysis! [Phys. Rev. Lett. 114 (2015) 121801, Phys. Rev. Lett. 117, 111802 (2016), JHEP 07 (2018) 127, Phys. Lett. B 786 (2018) 134]

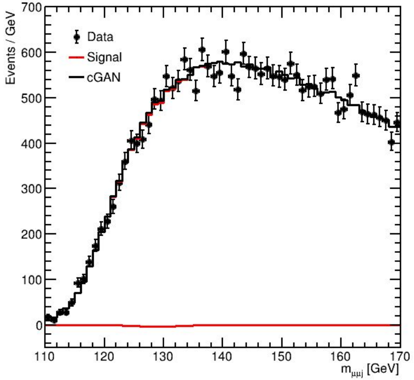




Conditional Generative Adversarial Networks

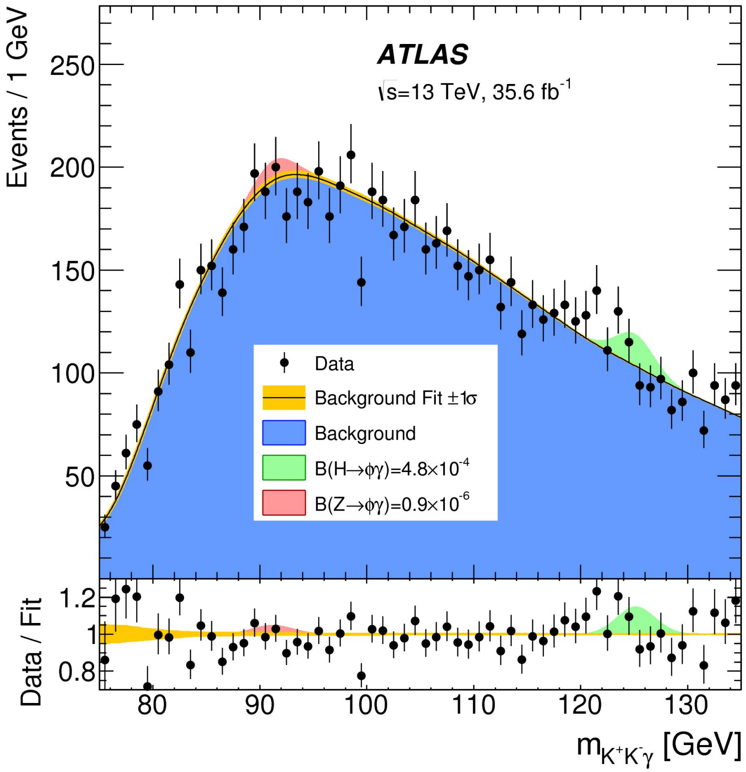
- Generalisation of ancestral sampling
- Use GANs trained on data to produce background model
- **Condition** GAN (cGAN) on a blinding variable, allowing **SR to be blinded during training**





BACK-UP

H/Z→Φγ Analysis

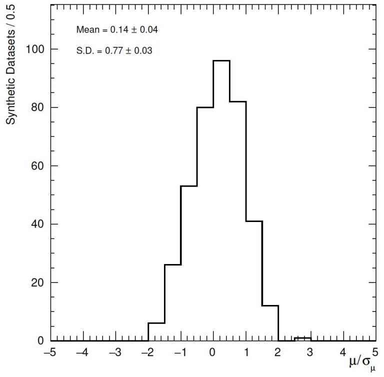


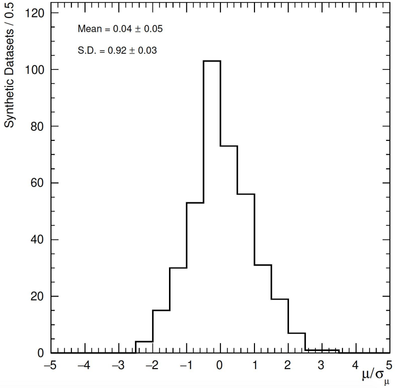
Branching Fraction Limit (95% CL)	Expected	Observed
$\mathcal{B}(H \to \phi\gamma)\,[\,10^{-4}\,]$	$4.2^{+1.8}_{-1.2}$	4.8
$\mathcal{B}(Z \to \phi\gamma)\,[\,10^{-6}\,]$	$1.3^{+0.6}_{-0.4}$	0.9

Ensemble Tests

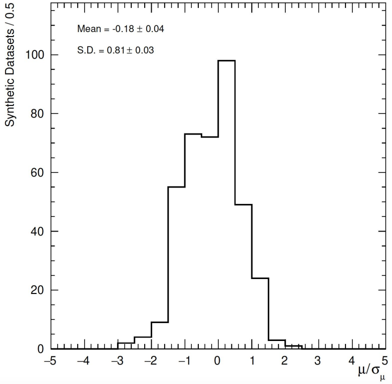
Sampling from GR | **Sampling with replacement**

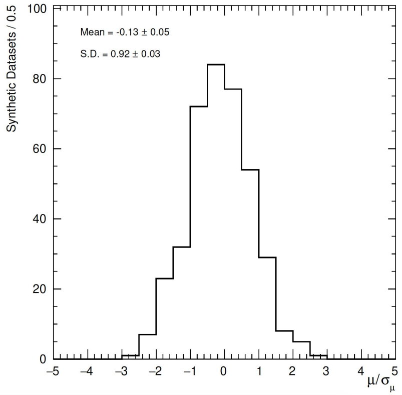
m = 125 GeV





m = 150 GeV





Ensemble Tests

- Due to the computational cost, number of training steps lowered, and training stopped before saturation