Design by intelligent committee:

machine learning as a scientific advisor

IML Workshop, 19th-23rd October 2020

Stephen Menary, Darren Price

University of Manchester

Research supported by The Alan Turing Institute, London

Introduction

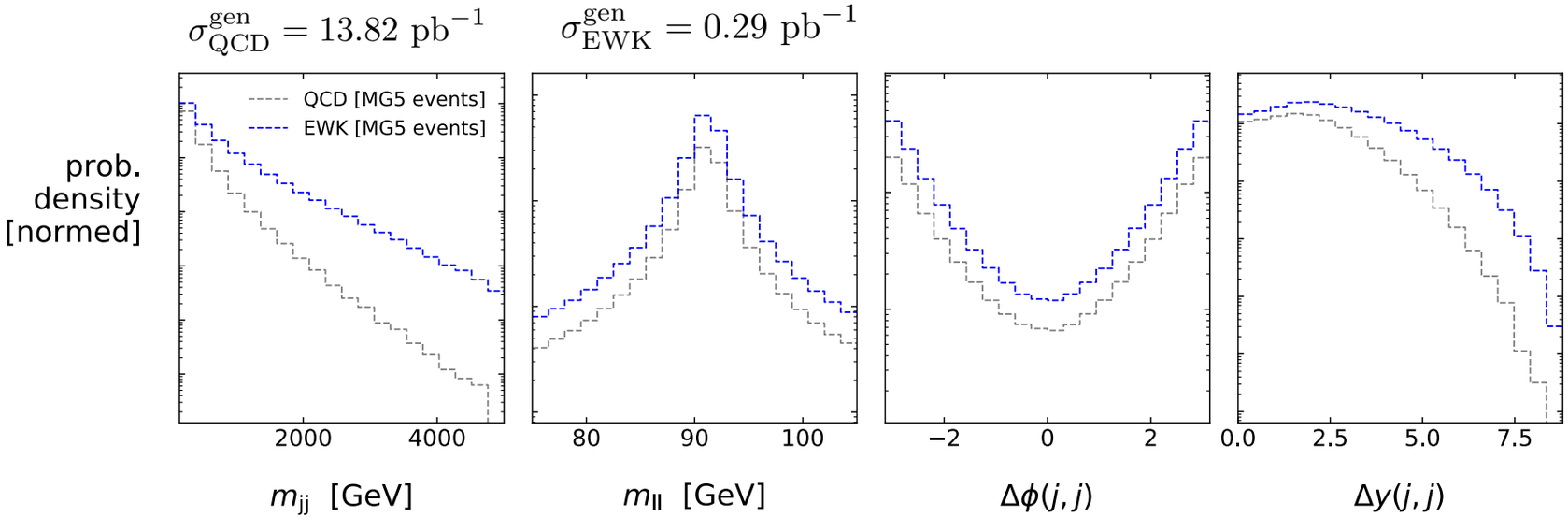
- ML good at understanding ***high dimensional dependencies***
  - e.g. in parameter space / space of observables

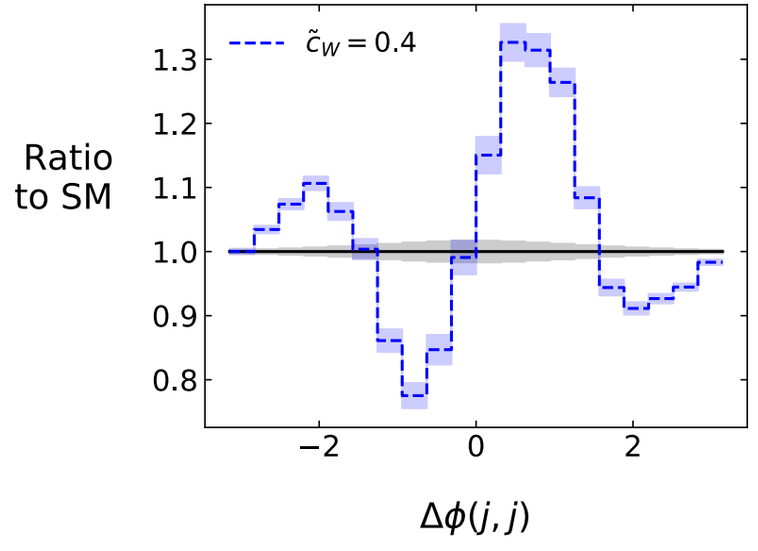
- Humans often bad at this, and make ad-hoc choices when designing analyses
  - e.g. which fiducial cuts and/or differential observables

- Promote idea of ***ML as insight extractor***, which can report on physical dependencies
  - analysers better understand salient features of models and data
  - —> design more sensitive measurements
  - —> better constrain models in multi-dimensional spaces (instead of 1D or 2D projections)
  - actual measurements don't inherit any bias in ML tool

- Real-world example

Real-world example: EWK Z + 2j production





signal appears at high m_{jj}

cannot separate sig/bkg

Δy(j,j) correlated with m_{jj}

signal ΔΦ(j,j) deformed by $\tilde{c}_W$

Key features of the data

Aim: ML workflow to design analysis, suitable for automation —> **apply when we don't know answer**

- This is just a well-understood example! Method contains nothing specific to the physics, so useful for all analyses.
- See arXiv:2006.15458 for ATLAS experimental analysis

Proposed workflow

1. Construct probability model for the data

↓

2. Approximate sensitivity bound

↓

3. Extract insights

↓

4. Optimal fiducial volumes to constrain models around sensitivity bound

Auto-regressive density models

Model overview

- High-dimensional joint density = product of 1D conditionals

easy to add dimensions

$$p\left(x_1, x_2, x_3|\theta\right) \;=\; p\left(x_1|x_2, x_3, \theta\right) \cdot p\left(x_2|x_3, \theta\right) \cdot p\left(x_3|\theta\right)$$

- Each 1D conditional modelled with **Gaussian mixture model**, with mean/variance/amplitude controlled by **neural net**

$$p\left(x_j|x_{<j}, \theta\right) \;=\; \sum_{i=1}^{N_G} f_i\left(x_{<j}, \theta\right) \cdot \mathcal{N}\left(\mu\left(x_{<j}, \theta\right), \log\sigma\left(x_{<j}, \theta\right)\right)$$

- Can ***evaluate PDF*** of datapoint *and* ***sample*** new datapoints ***from the same probability model***

(possible alternative: normalising flows)

Auto-regressive density models

Model overview

- High-dimensional joint density = product of 1D conditionals

easy to add dimensions

$$p\left(x_1, x_2, x_3|\theta\right) \;=\; p\left(x_1|x_2, x_3, \theta\right) \cdot p\left(x_2|x_3, \theta\right) \cdot p\left(x_3|\theta\right)$$

- Each 1D conditional modelled with **Gaussian mixture model**, with mean/variance/amplitude controlled by **neural net**

$$p\left(x_j|x_{<j}, \theta\right) \;=\; \sum_{i=1}^{N_G} f_i\left(x_{<j}, \theta\right) \cdot \mathcal{N}\left(\mu\left(x_{<j}, \theta\right), \log\sigma\left(x_{<j}, \theta\right)\right)$$

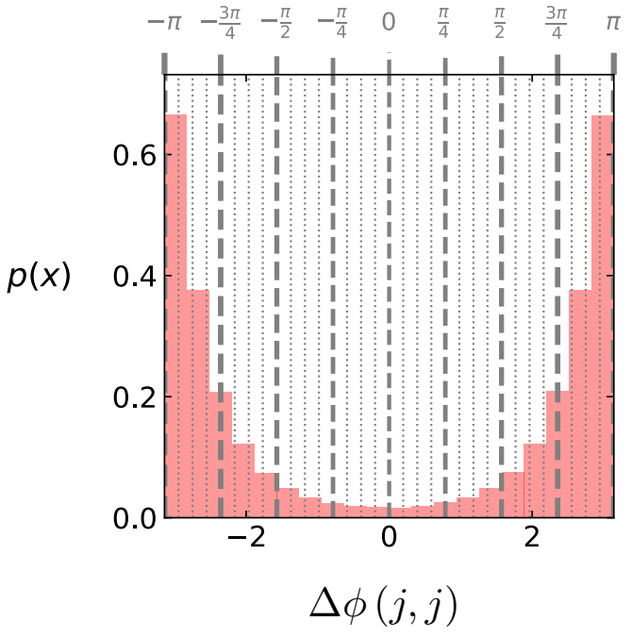
- Can ***evaluate PDF*** of datapoint *and* ***sample*** new datapoints ***from the same probability model***

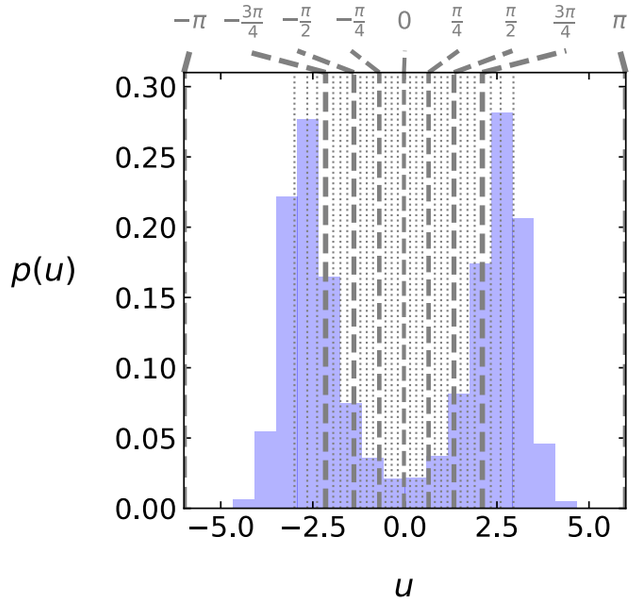
(possible alternative: normalising flows)

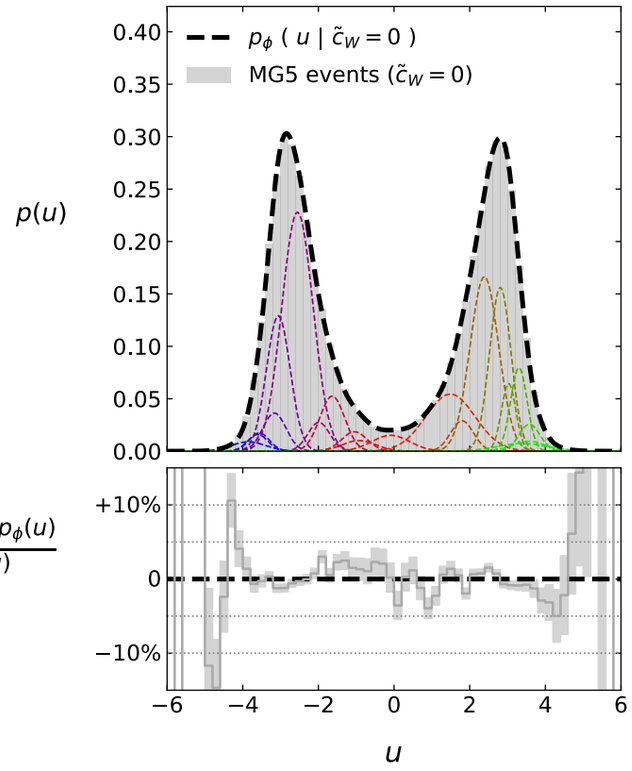
Footnote: features of density models

- Use maximum likelihood estimation to train on MC samples generated w/ MadGraph5
- PDF evaluation fast as each 1D conditional computed in parallel, sampling in sequence so speed linear with data dimensionality
- *N.B. can reverse this behaviour using IAFs if sampling speed is important*
- Each conditional trained independently - can parallelise for speed, and diagnose/improve independently
- Model is dependent on observable ordering (can create ensemble of many orderings if worried)
- Can act as both an inference network **and** a stochastic generator which does not require variational approximations and has a constant objective function during training (e.g. sometimes more stable than GANs).
- Unfortunately no access to marginal distributions

Novel method







Project onto latent space

- Removes hard physical boundaries
- Encourages distribution well described by Gaussian mixture

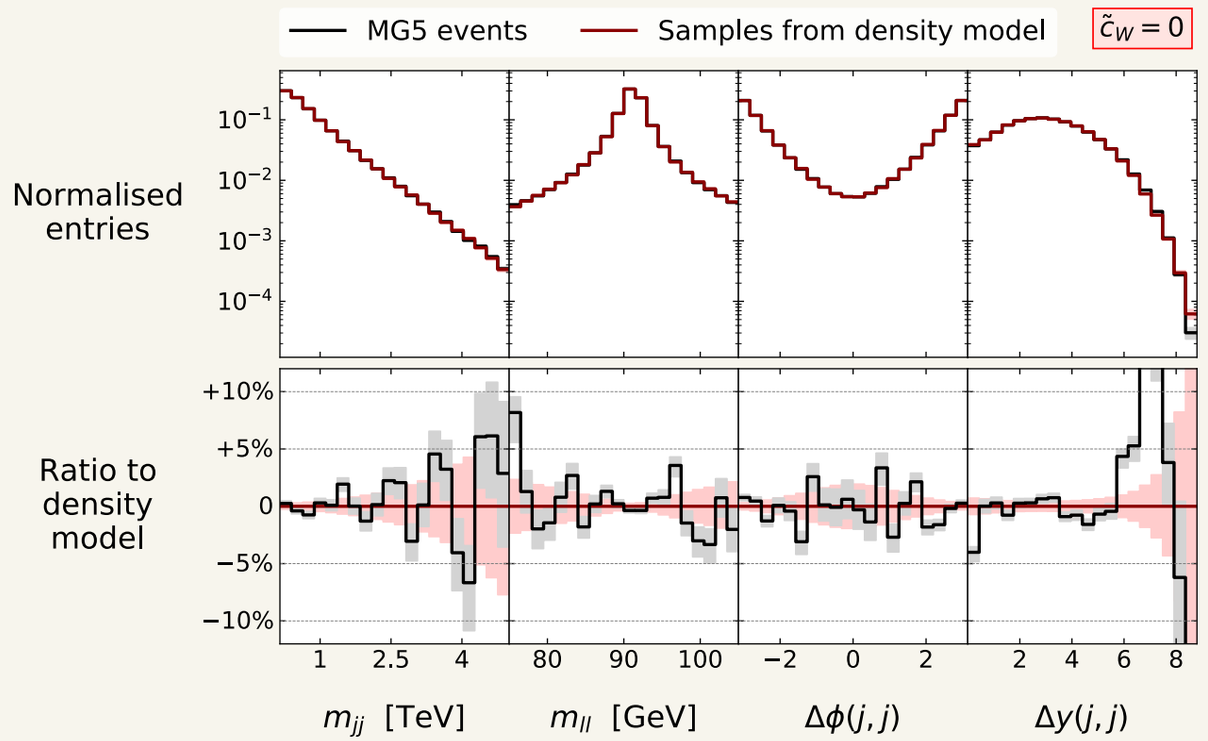
Density model on latent space

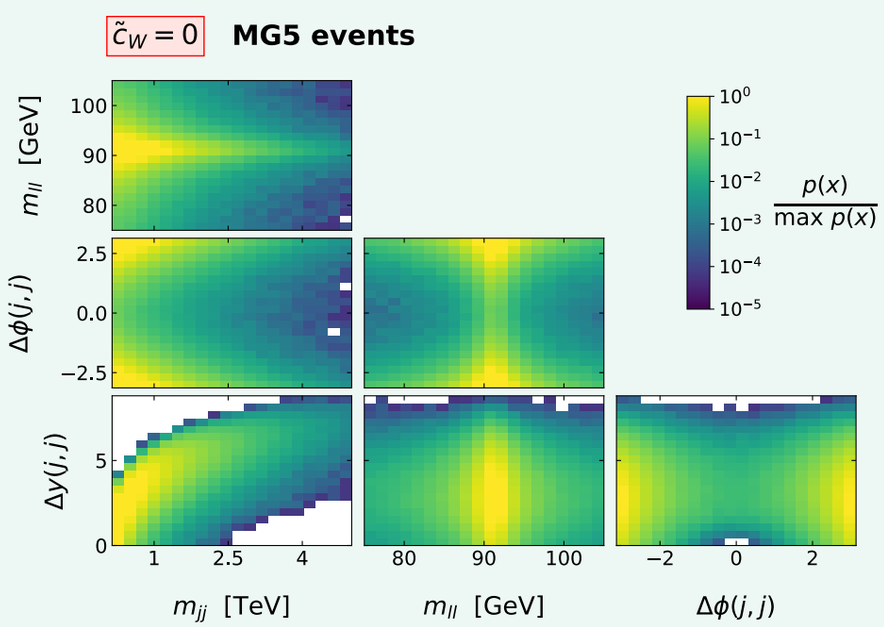
- Deformations of spectra expressed by modifying local Gaussians
- Model captures parameter dependence

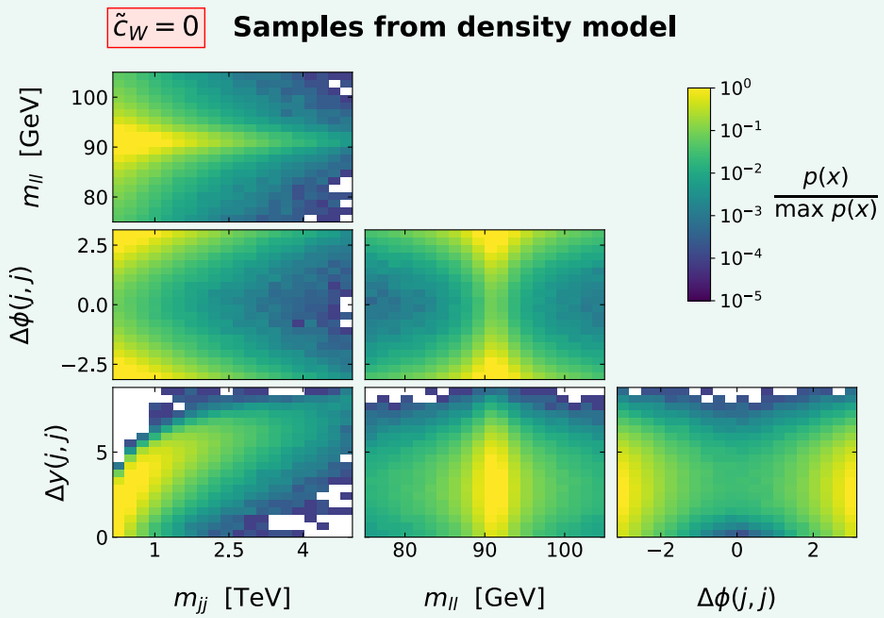
EWK density model at SM

1D projections within 5%

(except in sparse tail of Δy(j,j))

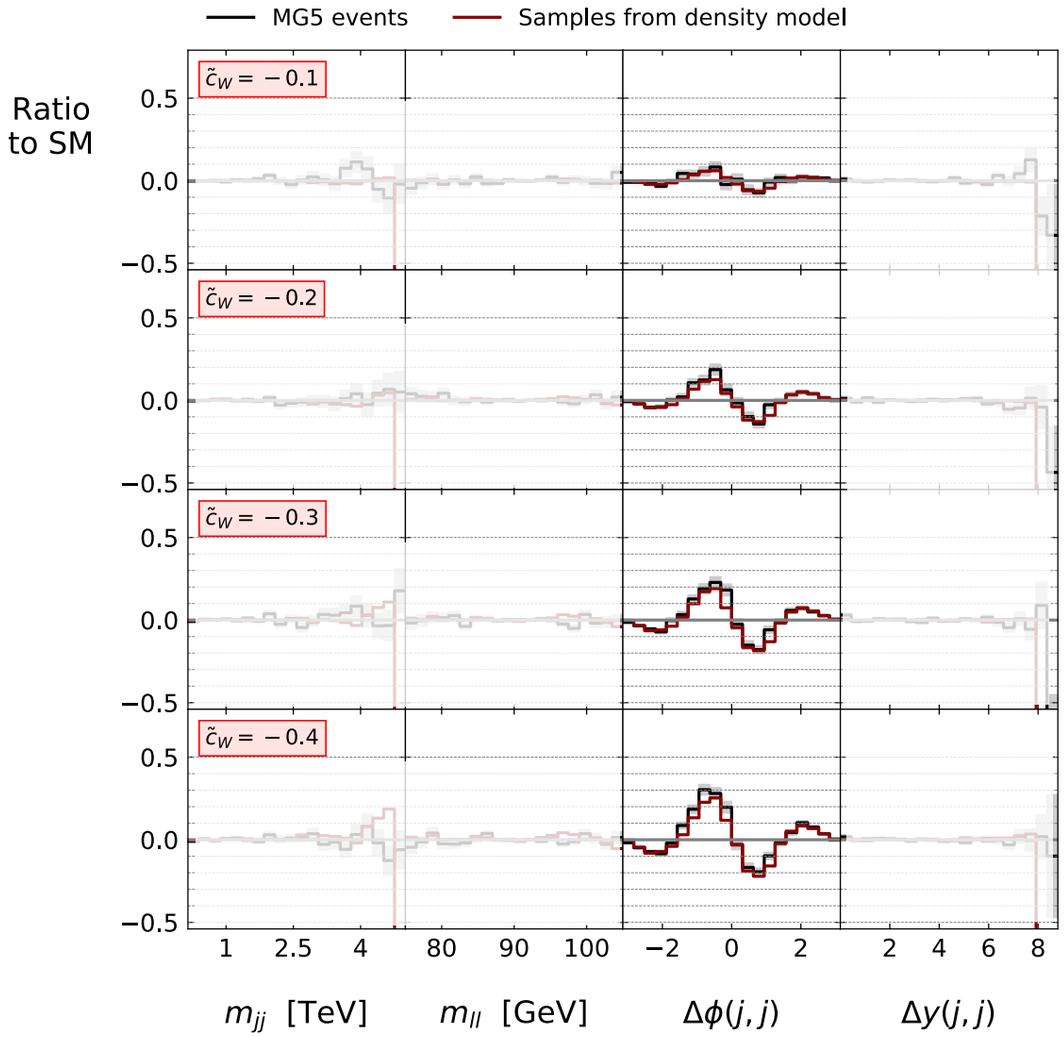


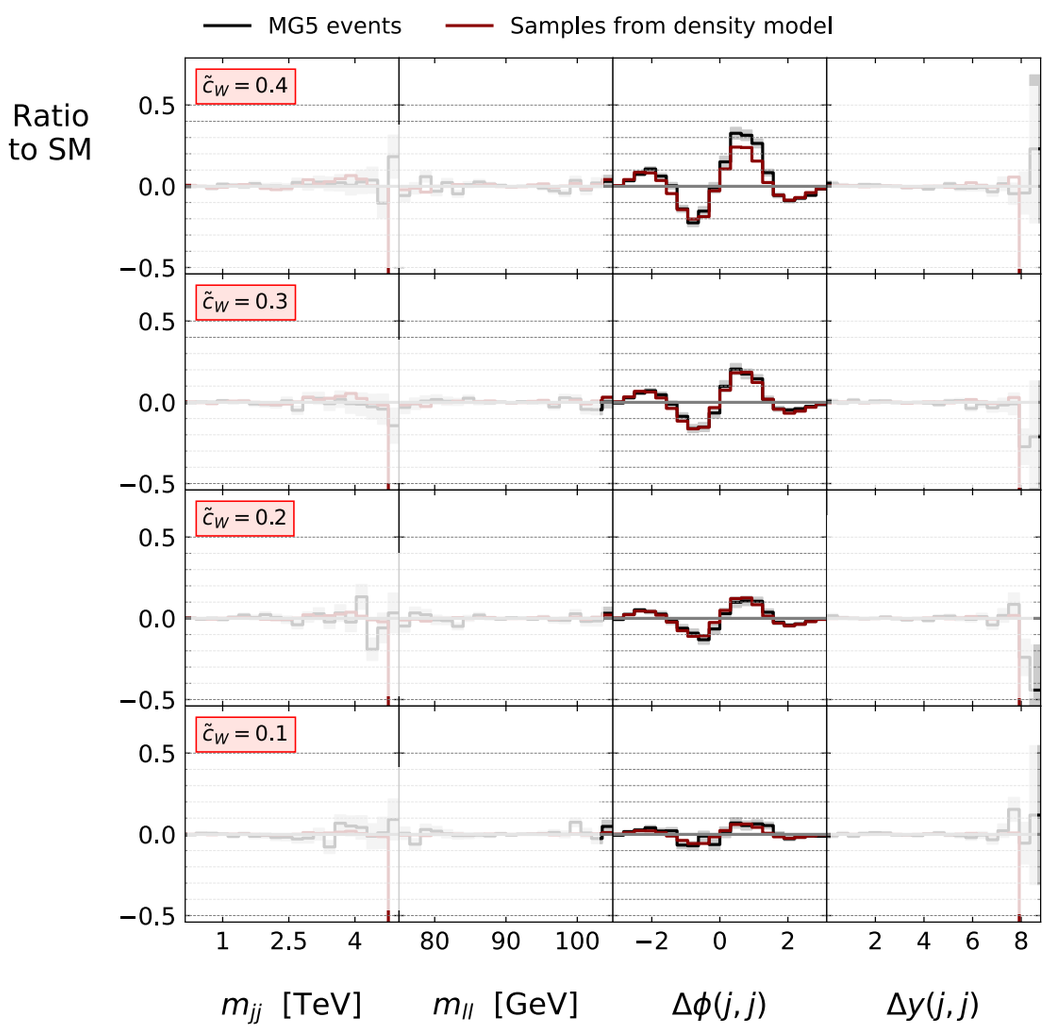




2D projections have captured observable dependencies

EWK density model parameter dependence

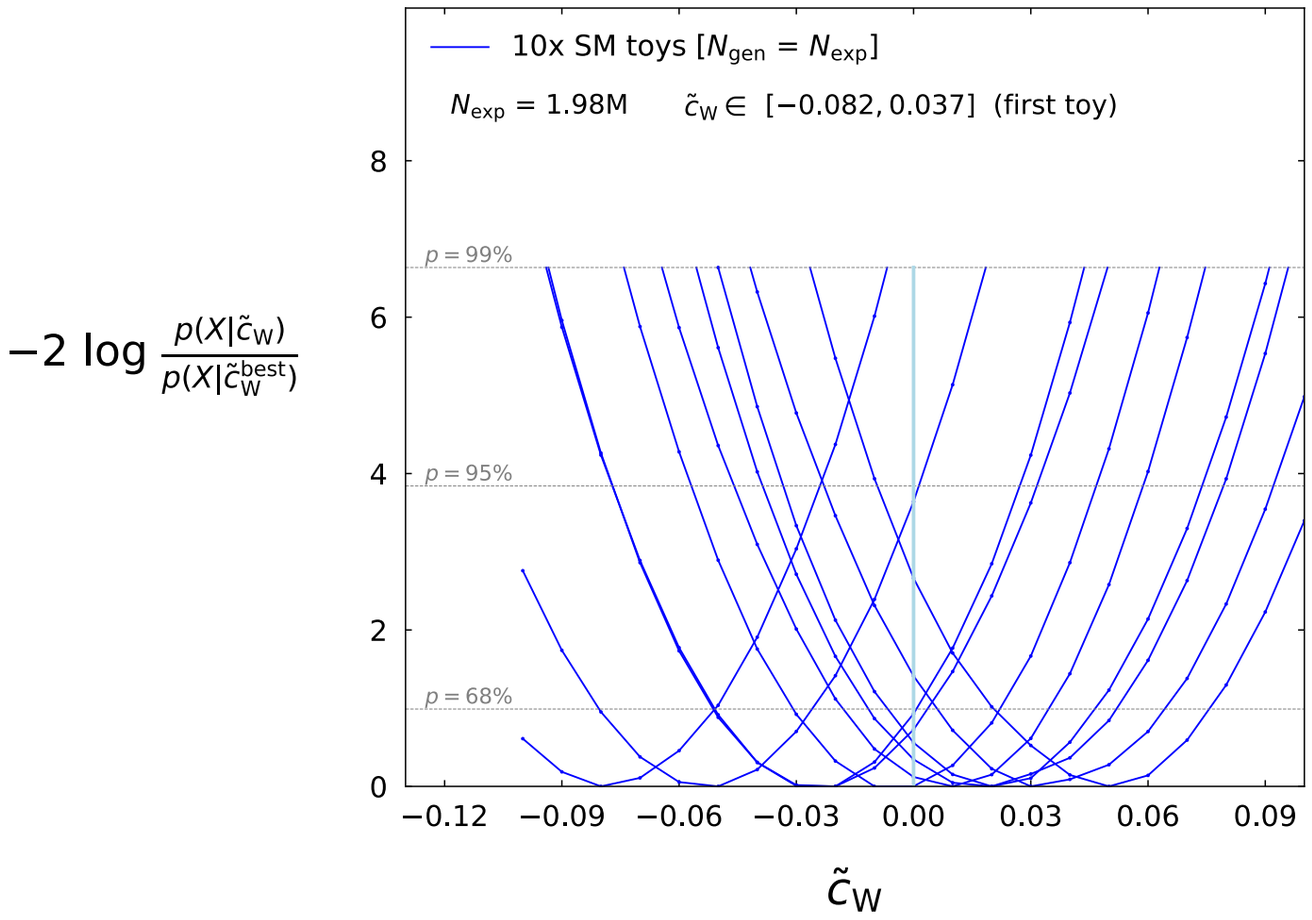




We capture the parameter dependence!

N.B. sample events at SM and reweight —> correlated stat fluctuations in ratio so only 5k events needed! - power of density model!

Reach of given dataset



Estimate sensitivity of unbinned analysis:
sample toy & profile likelihood

Target luminosity: 140/fb, no detector effects

However…

- Sampling fluctuations —> unstable estimates of experimental reach (especially in >1D)
- When $N_{gen} = N_{exp}$ —> sampling variance ≈ estimator variance
- When $N_{gen} \gg N_{exp}$ —> prohibitively large number of events required

Solution: importance sampling

How to improve sampling efficiency:

$$q_{\text{asimov}} = \mathbb{E}_{x\sim p(x|\theta_1)}\left[q\left(x|\theta_1,\theta_2\right)\right]$$

$$= -2\int_{-\infty}^{+\infty} p\left(x|\theta_1\right)\cdot\log\frac{p\left(x|\theta_2\right)}{p\left(x|\theta_1\right)}\,\mathrm{d}x$$

Usual toy method - inefficient as any events in bkg regions contribute q=0

Method 1

$$\approx -2\cdot\frac{N_{\text{exp}}}{N_{\text{gen}}}\cdot\sum_{x\sim p(x|\theta_1)}\log\frac{p\left(x|\theta_2\right)}{p\left(x|\theta_1\right)}$$

Sampling efficient when

$$Q\left(x\right) \propto \left|p\left(x|\theta_1\right)\cdot\log\frac{p\left(x|\theta_2\right)}{p\left(x|\theta_1\right)}\right|$$

Method 2

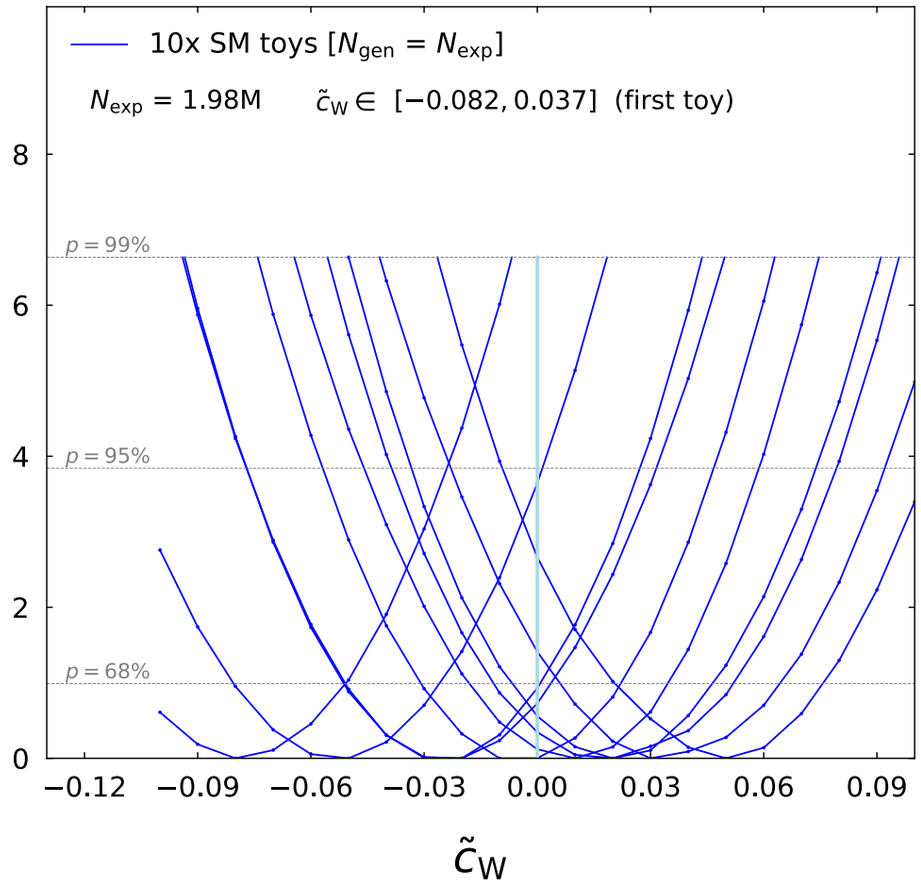
$$\approx -2\cdot\frac{N_{\text{exp}}}{N_{\text{gen}}}\cdot\sum_{x\sim Q(x)}\frac{p\left(x|\theta_1\right)}{Q\left(x\right)}\cdot\log\frac{p\left(x|\theta_2\right)}{p\left(x|\theta_1\right)}$$

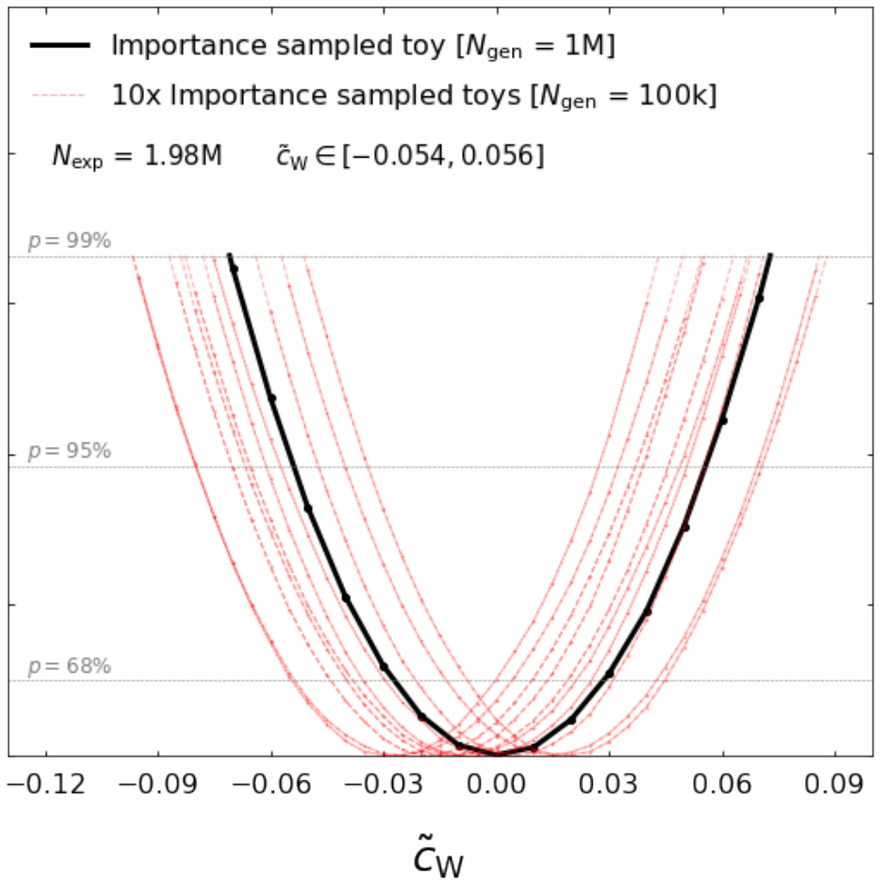
Proposed solution…

- Choose a reference value of cWtil
- Sample datapoints at SM, and calculate values of absolute log likelihood ratio wrt cWtil
- Train **auxiliary density model** to describe this distribution
- Model can be sampled and evaluated, allowing ~ efficient importance sampling as we scan cWtil

Importance sampling: efficiency gains

$-2 \log \frac{p(X|\tilde{c}_W)}{p(X|\tilde{c}_W^{best})}$





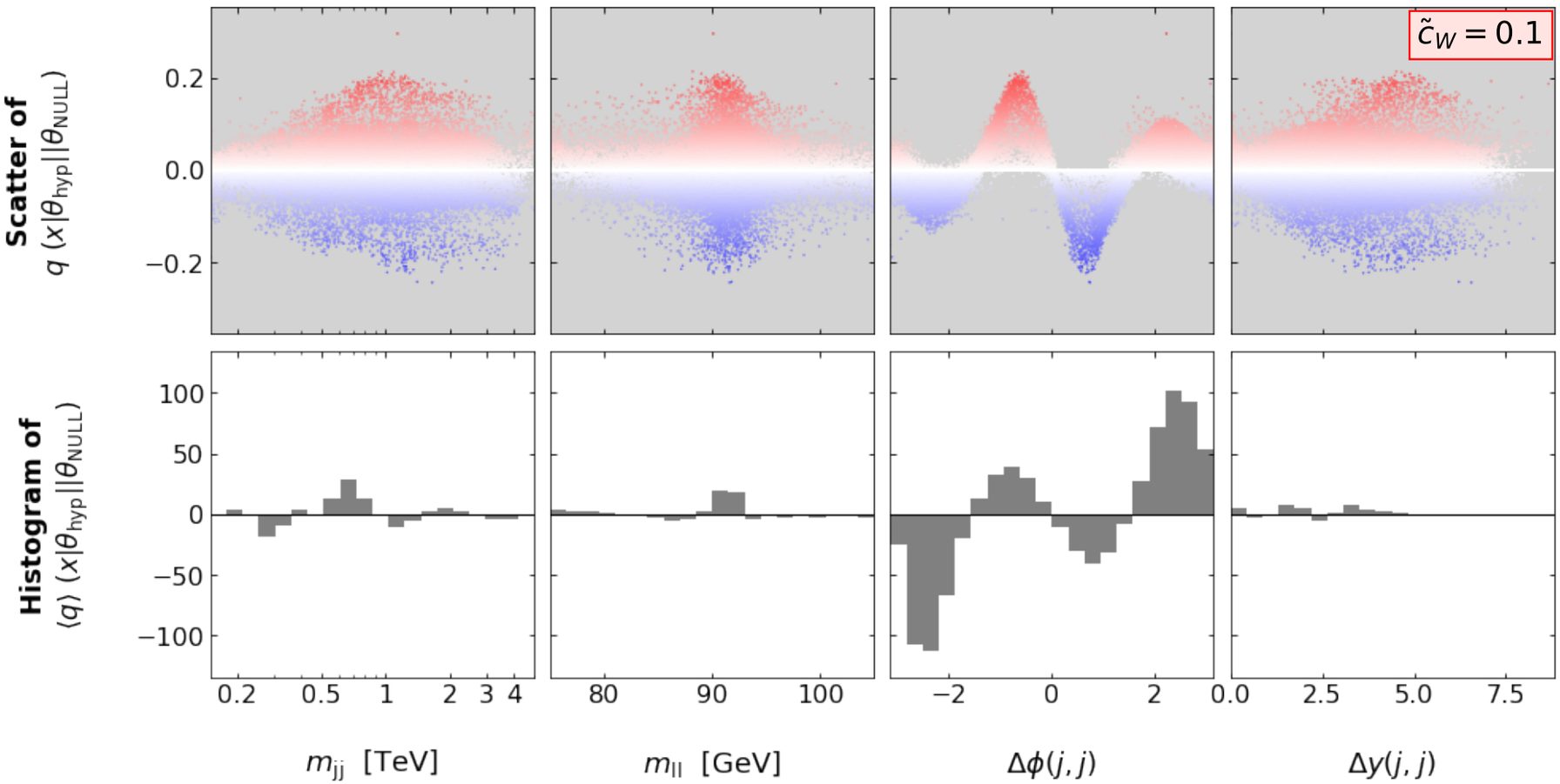
Toys from before, 2M events

Importance sampling, 100k - 1M events

Result

- Achieved smaller sampling variance with many fewer events :)
- N.B. unbinned analysis likely more sensitive than binned, see e.g. J. Brehmer et al [arXiv:1805.00013]

Extracting information from 1D marginals



test statistic @ individual datapoints:

$$q\,(x;\theta_2||\theta_1) \;=\; -\,2\cdot\log\frac{p\,(x|\theta_2)}{p\,(x|\theta_1)}$$

test statistic integrated over bin:

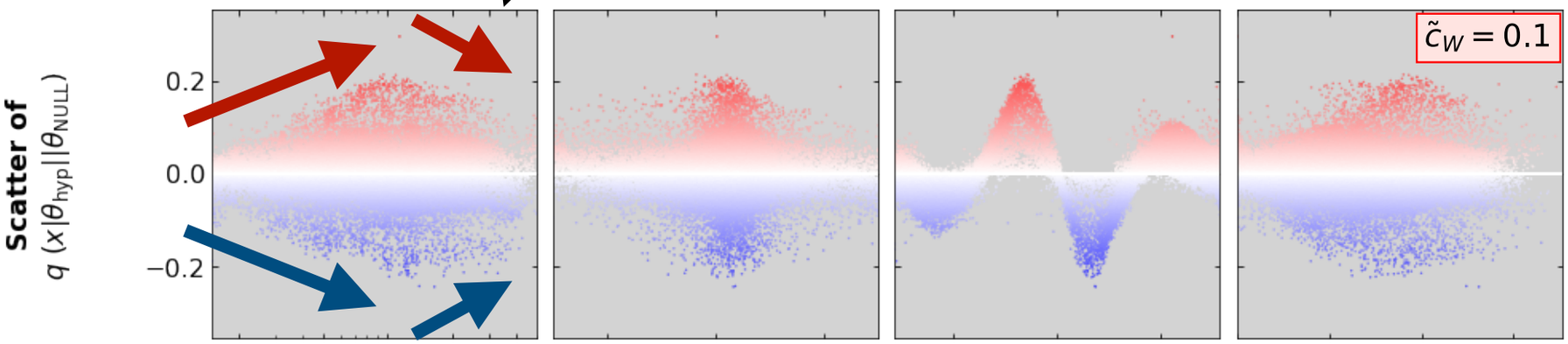
$$\langle q\,(x;\theta_2||\theta_1)\rangle_{\theta_1} \;=\; p\,(x|\theta_1)\cdot q\,(\theta_2||\theta_1)$$

Extracting information from 1D marginals

low m_{jj}, QCD dominated, less separability

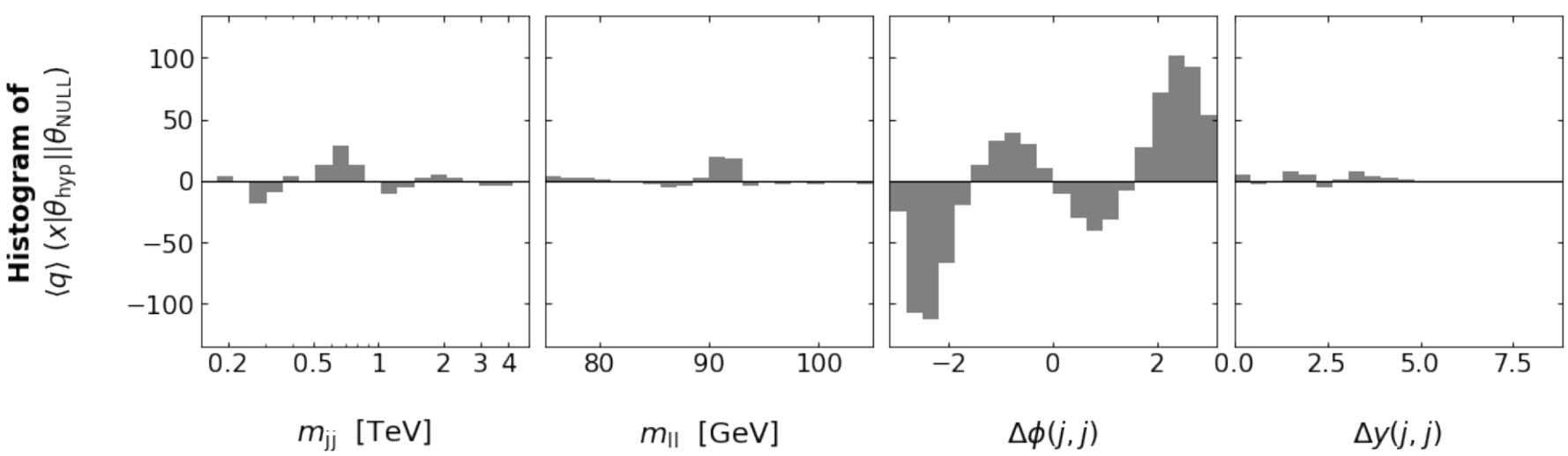
mid m_{jj}, many events & good separability

high m_{jj}, few events but good separability



test statistic @ individual datapoints:

$$q\,(x;\theta_2||\theta_1) \;=\; -\,2\cdot\log\frac{p\,(x|\theta_2)}{p\,(x|\theta_1)}$$



test statistic integrated over bin:

$$\langle q\,(x;\theta_2||\theta_1)\rangle_{\theta_1} \;=\; p\,(x|\theta_1)\cdot q\,(\theta_2||\theta_1)$$

Just as many datapoint provide +ve and -ve contributions, so binning in m_{jj} gives no sensitivity

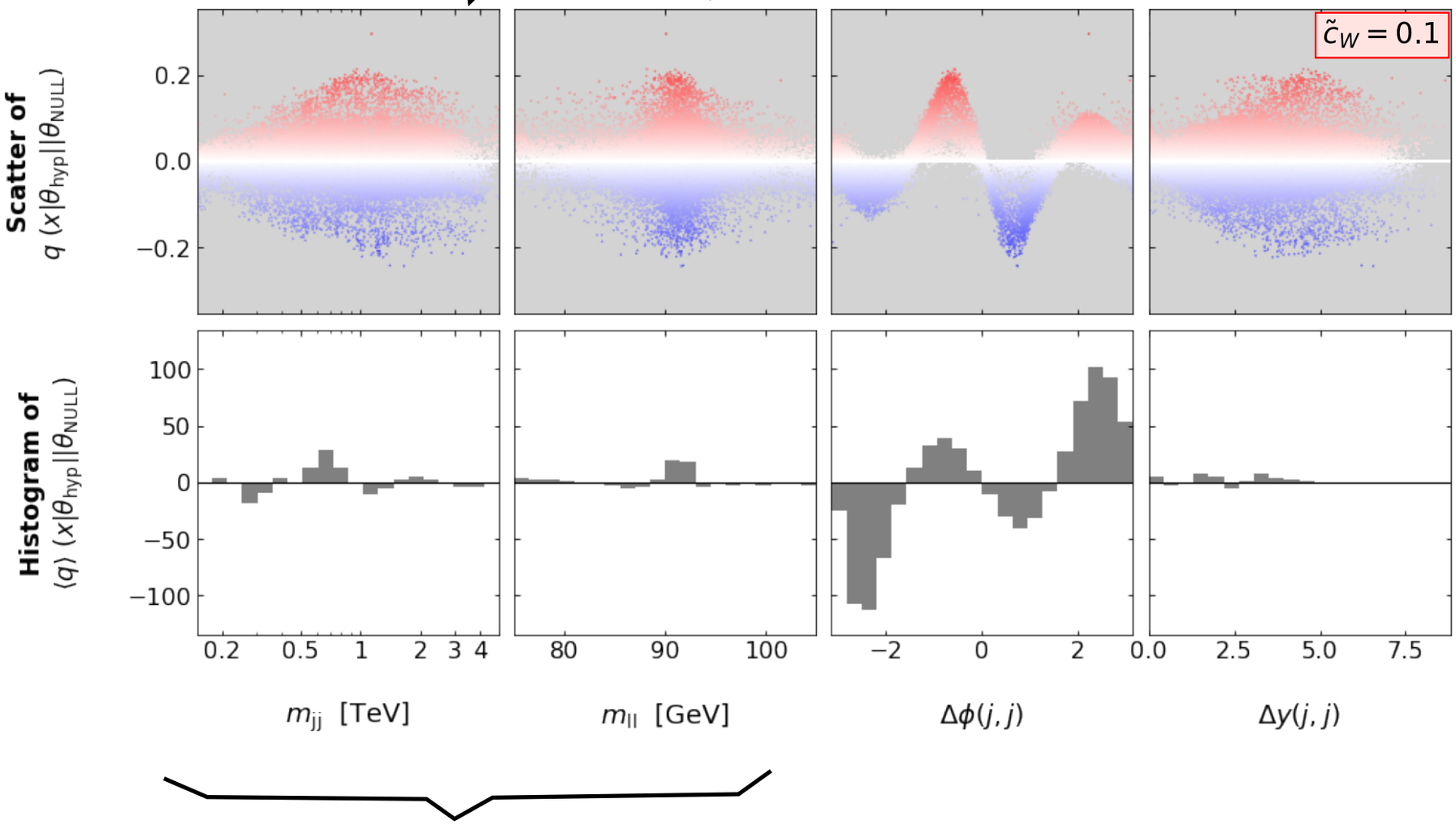
Extracting information from 1D marginals

low m_{jj}, QCD dominated, less separability

mid m_{jj}, many events & good separability

high m_{jj}, few events but good separability

events concentrated at Z peak (obviously)



test statistic @ individual datapoints:

$$q\,(x;\theta_2||\theta_1) \;=\; -\,2\cdot\log\frac{p\,(x|\theta_2)}{p\,(x|\theta_1)}$$

test statistic integrated over bin:

$$\langle q\,(x;\theta_2||\theta_1)\rangle_{\theta_1} \;=\; p\,(x|\theta_1)\cdot q\,(\theta_2||\theta_1)$$

Just as many datapoint provide +ve and -ve contributions, so binning in m_{jj} gives no sensitivity

Extracting information from 1D marginals

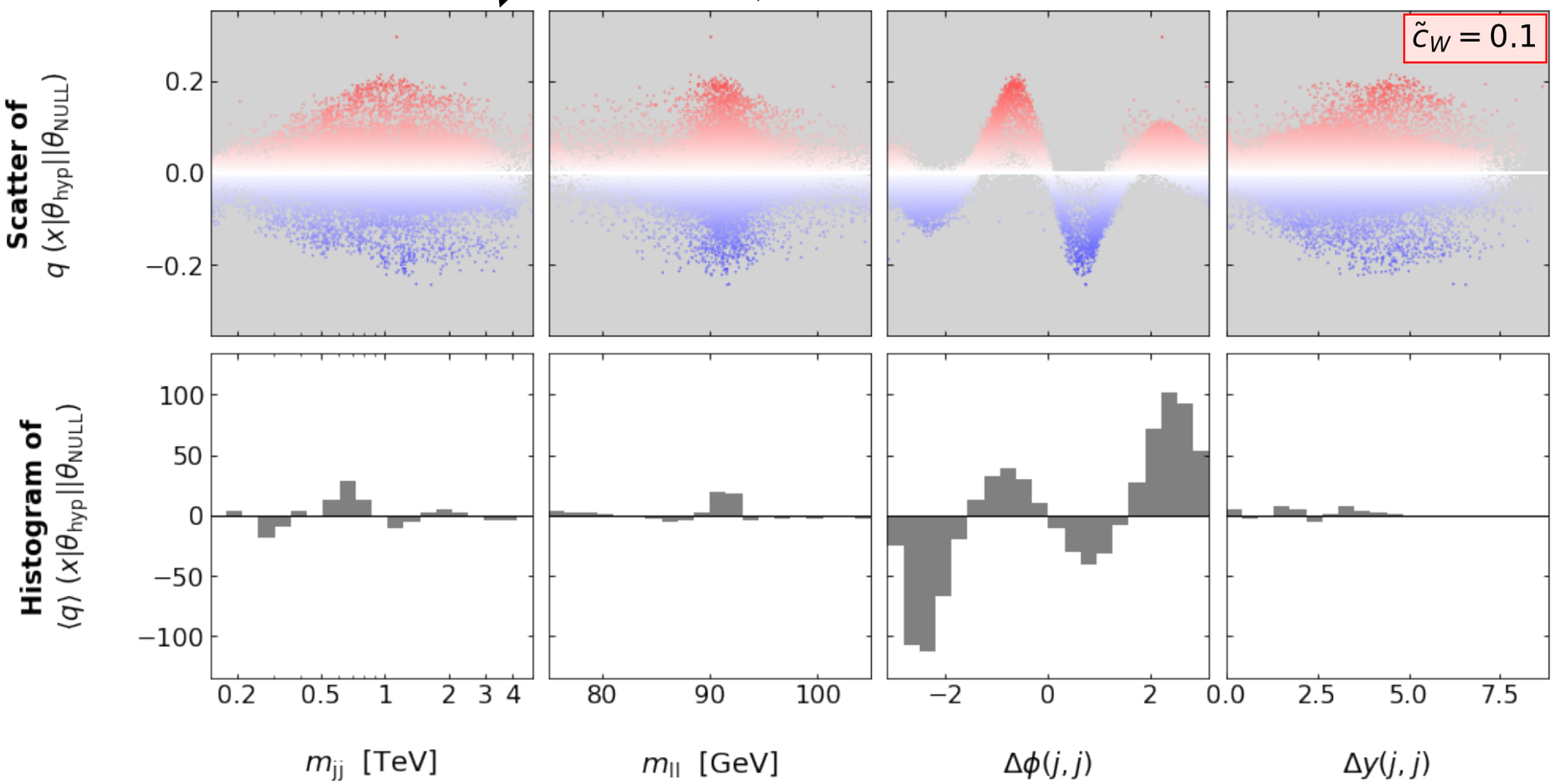
low m_{jj}, QCD dominated, less separability

mid m_{jj}, many events & good separability

high m_{jj}, few events but good separability

events concentrated at Z peak (obviously)

ΔΦ(j,j) almost optimal in separating +ve and -ve contributions



test statistic @ individual datapoints:

$$q\,(x;\theta_2||\theta_1) \;=\; -\,2\cdot\log\frac{p\,(x|\theta_2)}{p\,(x|\theta_1)}$$

test statistic integrated over bin:

$$\langle q\,(x;\theta_2||\theta_1)\rangle_{\theta_1} \;=\; p\,(x|\theta_1)\cdot q\,(\theta_2||\theta_1)$$

Just as many datapoint provide +ve and -ve contributions, so binning in m_{jj} gives no sensitivity

As we knew from origin events, differential ΔΦ(j,j) provides excellent separability

Extracting information from 1D marginals

low m_{jj}, QCD dominated, less separability

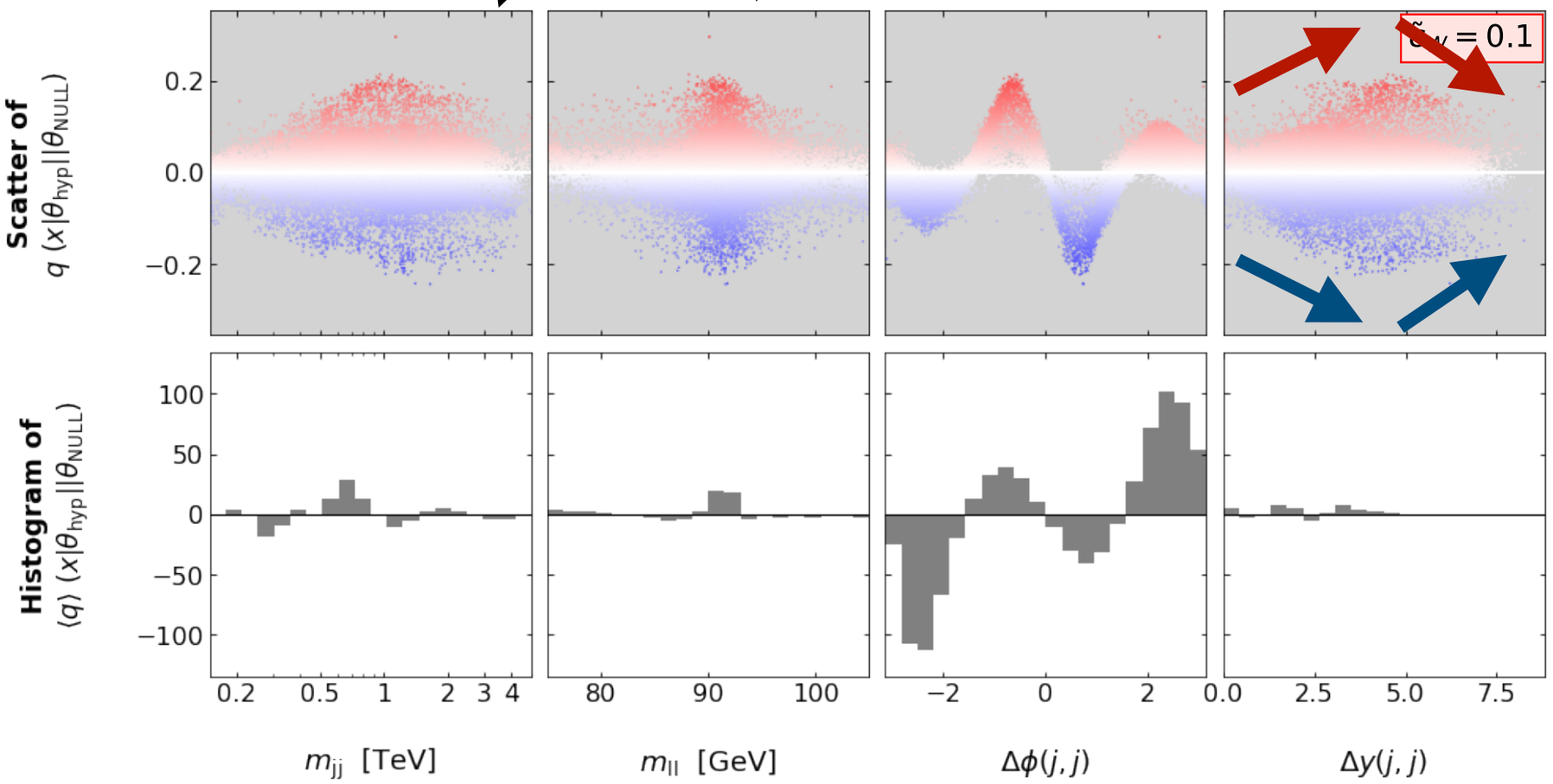
mid m_{jj}, many events & good separability

high m_{jj}, few events but good separability

events concentrated at Z peak (obviously)

ΔΦ(j,j) almost optimal in separating +ve and -ve contributions

Δy(j,j), similar to m_{jj}



test statistic @ individual datapoints:

$$q\,(x;\theta_2||\theta_1) \;=\; -\,2\cdot\log\frac{p\,(x|\theta_2)}{p\,(x|\theta_1)}$$

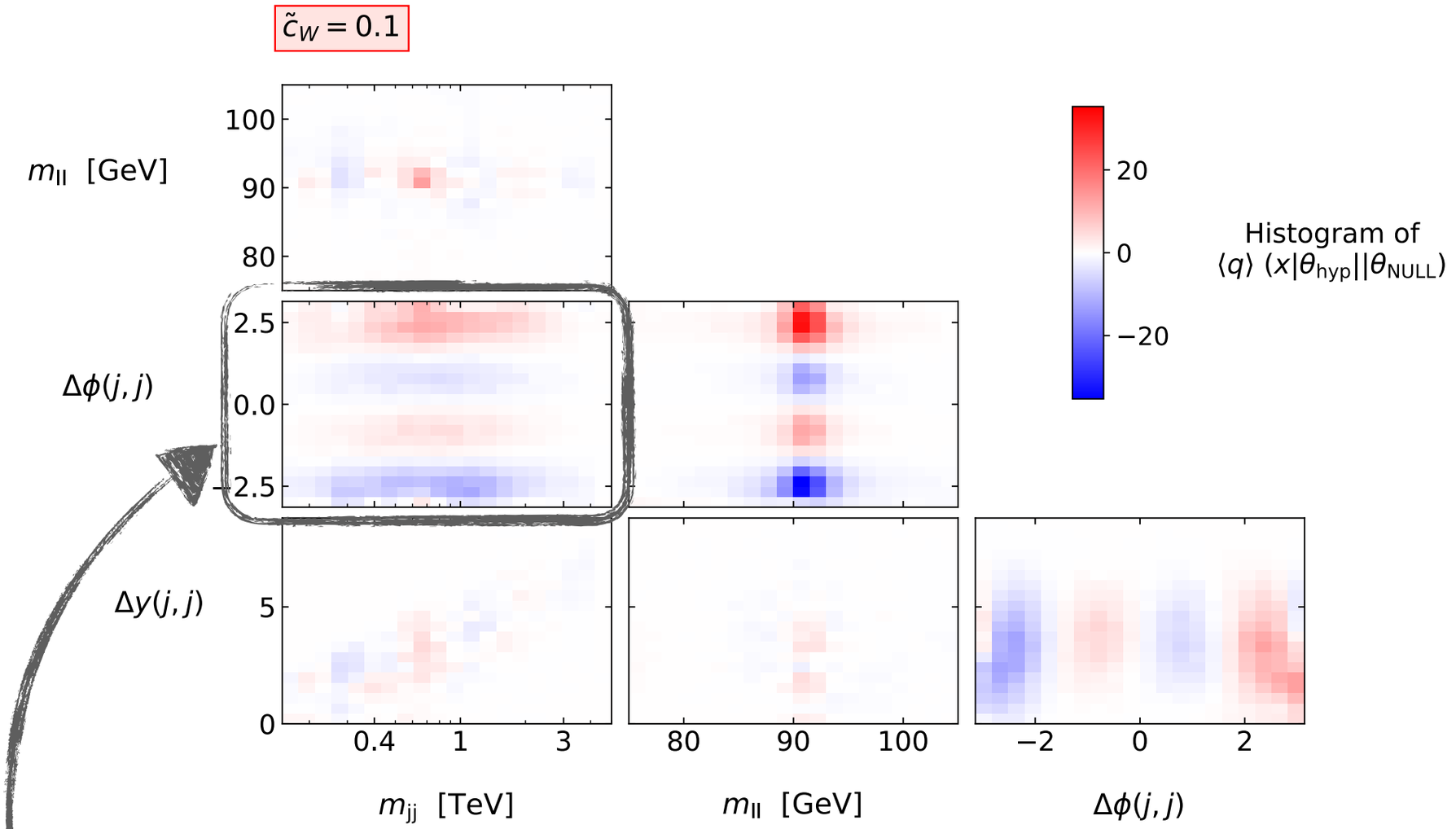
test statistic integrated over bin:

$$\langle q\,(x;\theta_2||\theta_1)\rangle_{\theta_1} \;=\; p\,(x|\theta_1)\cdot q\,(\theta_2||\theta_1)$$

Just as many datapoint provide +ve and -ve contributions, so binning in m_{jj} gives no sensitivity

As we knew from origin events, differential ΔΦ(j,j) provides excellent separability

Extracting information from 2D marginals



2D dependencies: ΔΦ(j,j) distribution has most sensitivity at mid m_{jj} / Δy(j,j)

Note related work

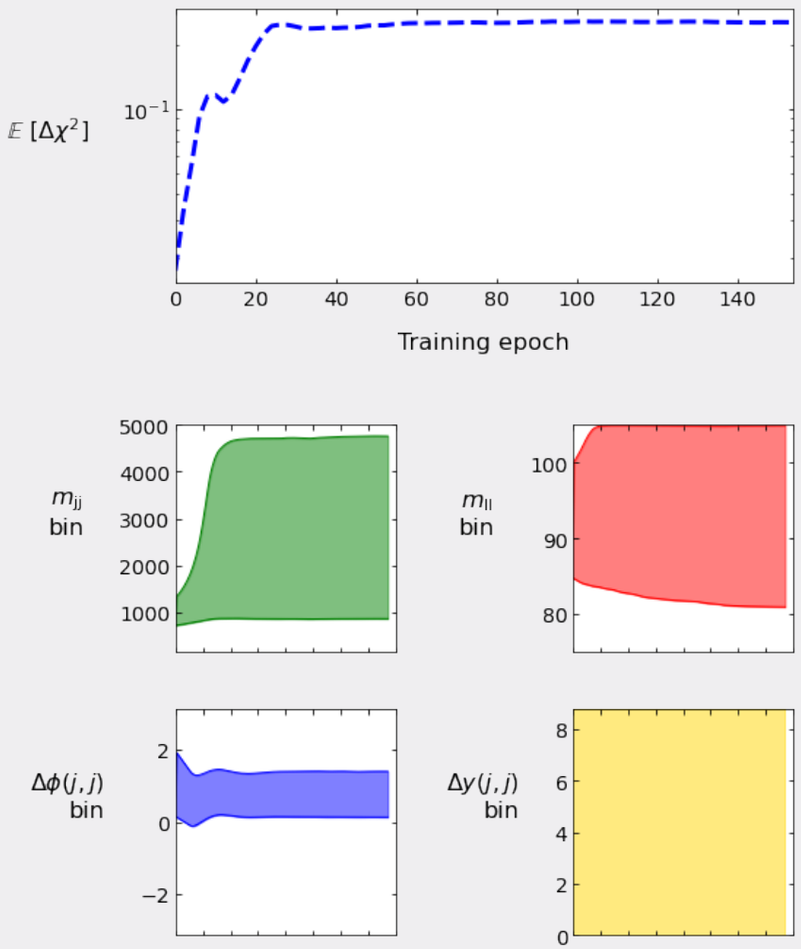
- J. Brehmer et al [arXiv:1612.05261, arXiv:1907.10621]
- Use of density/ratio/score estimators & Fisher information for similar purposes

Optimal fiducial volumes

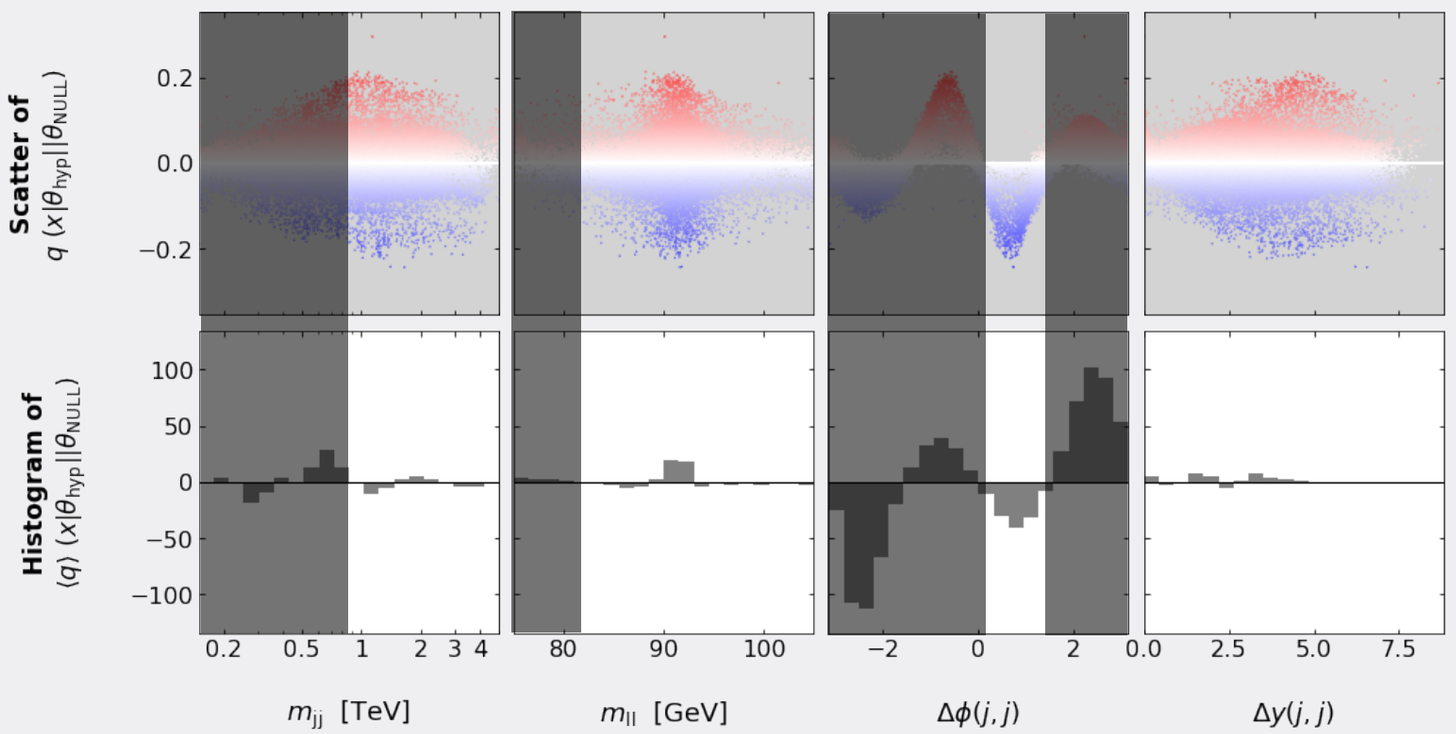
Work in progress
curated results

- Numerically estimate gradient of expected $\Delta\chi^2$ wrt selection cuts + gradient ascent
- Algorithm selected **high m_{jj} region** and a **single cluster of $\Delta\Phi(j,j)$** *(N.B. forced to be inclusive in $\Delta y(j,j)$)*

Training



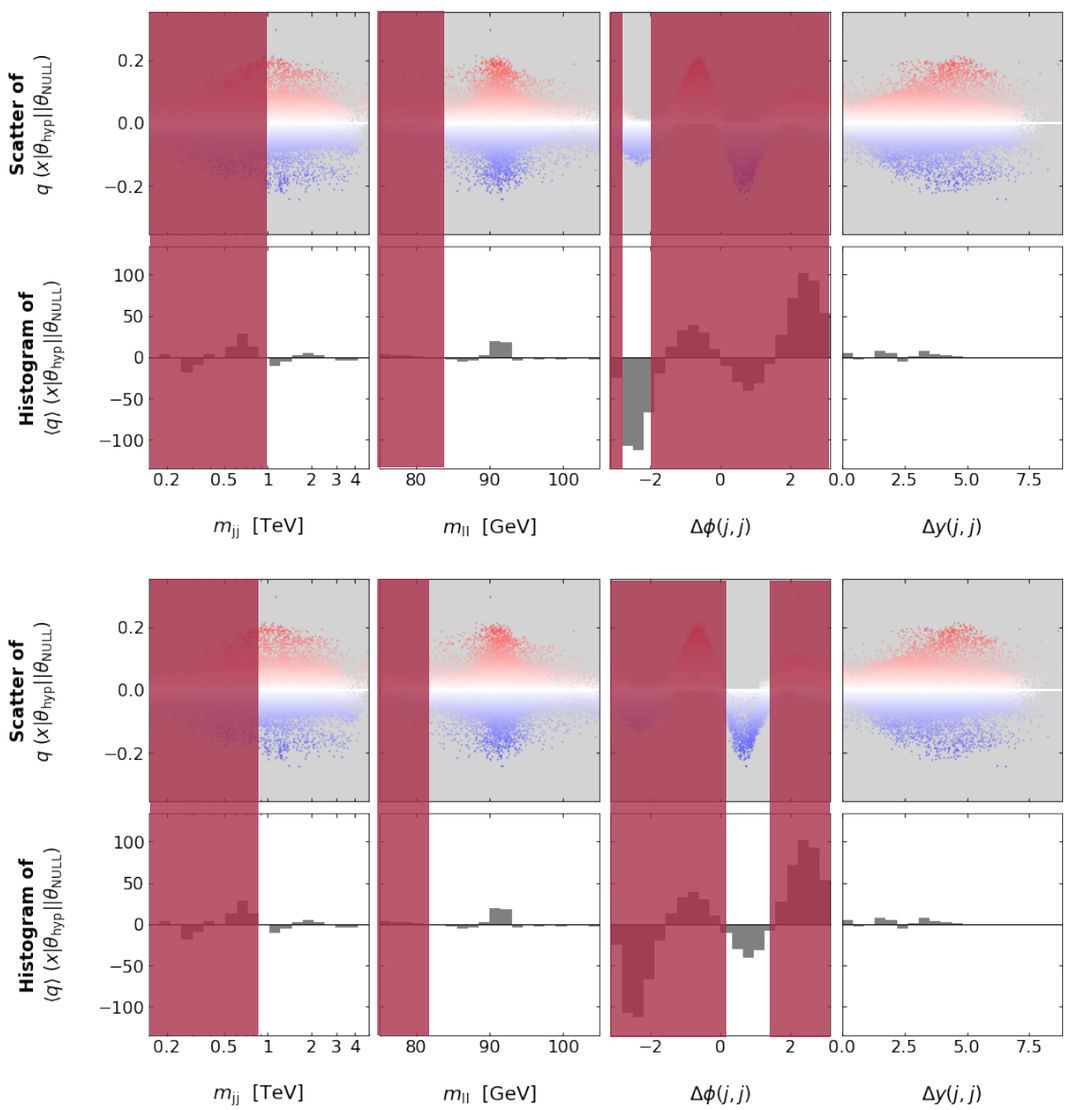
Suggested fiducial selection

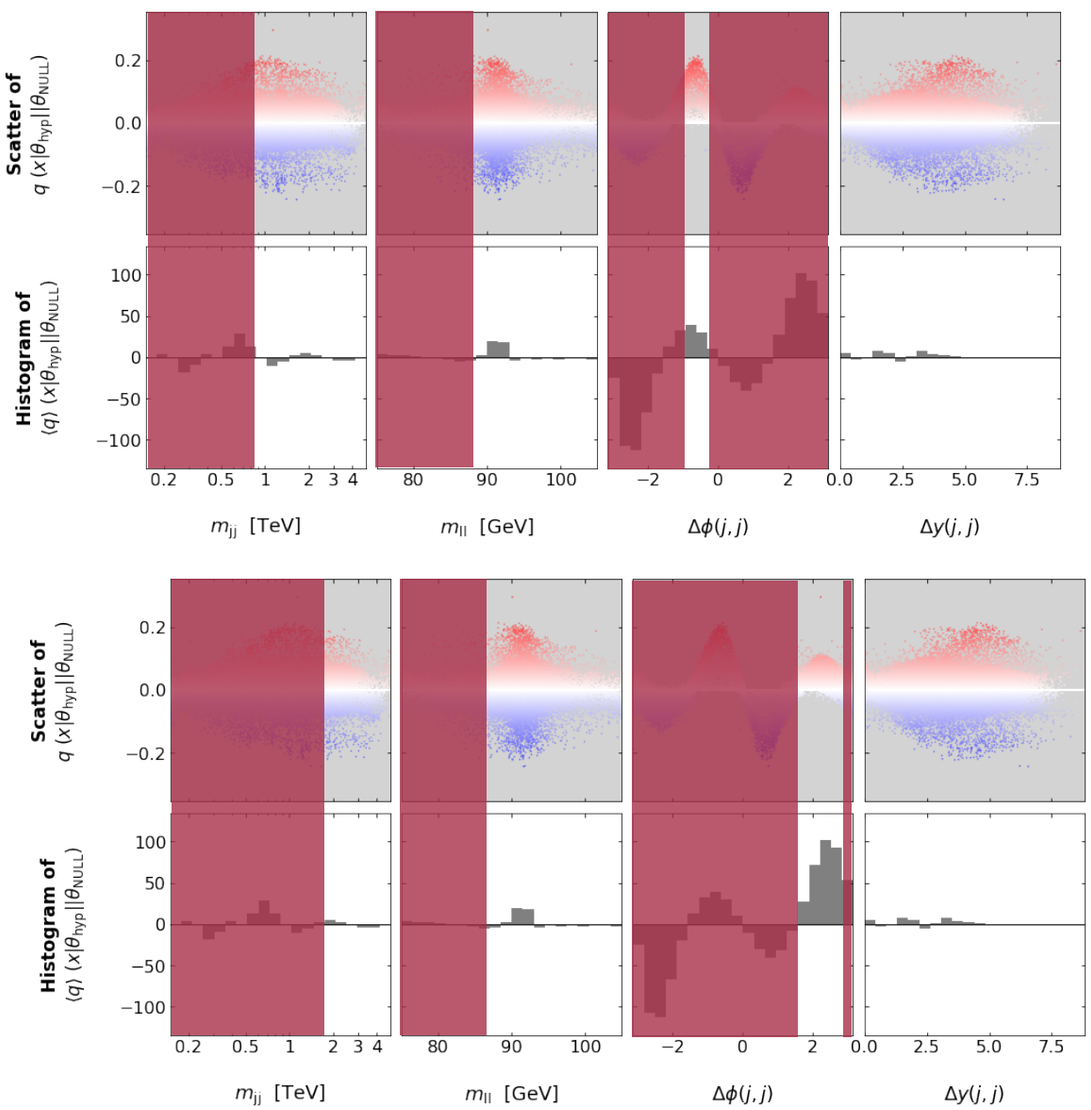


MANCHESTER 1824
The University of Manchester

Optimal fiducial volumes

Work in progress
curated results





- Ensemble of initial conditions —> find all useful fiducial volumes, i.e. peaks of ΔΦ(j,j)

Summary

- Promote idea of ***ML as an insight extractor***
  - understand sensitivity reach in parameter space
  - understand data features to design better analyses
  - measurements don't inherit ML bias *(but may be sub-optimal)*
  - **designed for automation —> create an AI scientific advisor**
  - **general tool for optimising data analyses for scientific discovery**

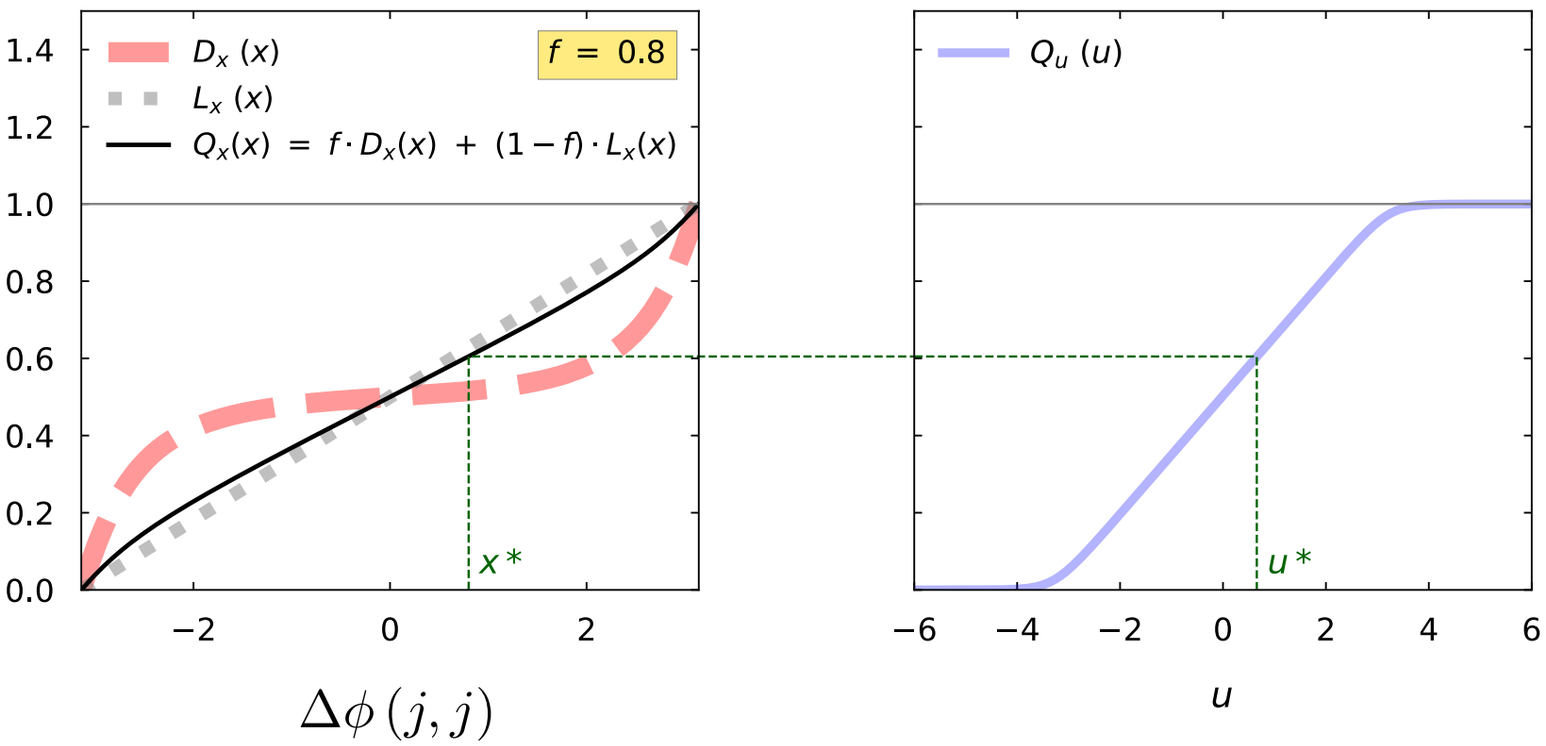
- Demonstrated workflow on well-understood VBFZ example
  - neural density models very powerful (allow sampling and tractable likelihood evaluation)
  - showed flexible 4D density model using novel latent space method + Gaussian mixture. Easily extends to higher dimensions because auto-regressive. Also powerful generator as alternative to GANs/VAEs.
  - showed importance sampling for efficient sensitivity estimation using multi-model data
  - ***showed insights can be extracted from model***, e.g. visualise sensitivity, suggest optimal fiducial selections

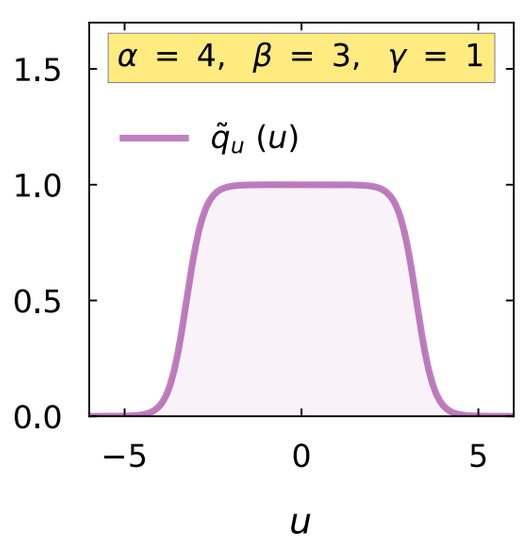
Backup

Using a latent space

Derive way to project data onto latent space

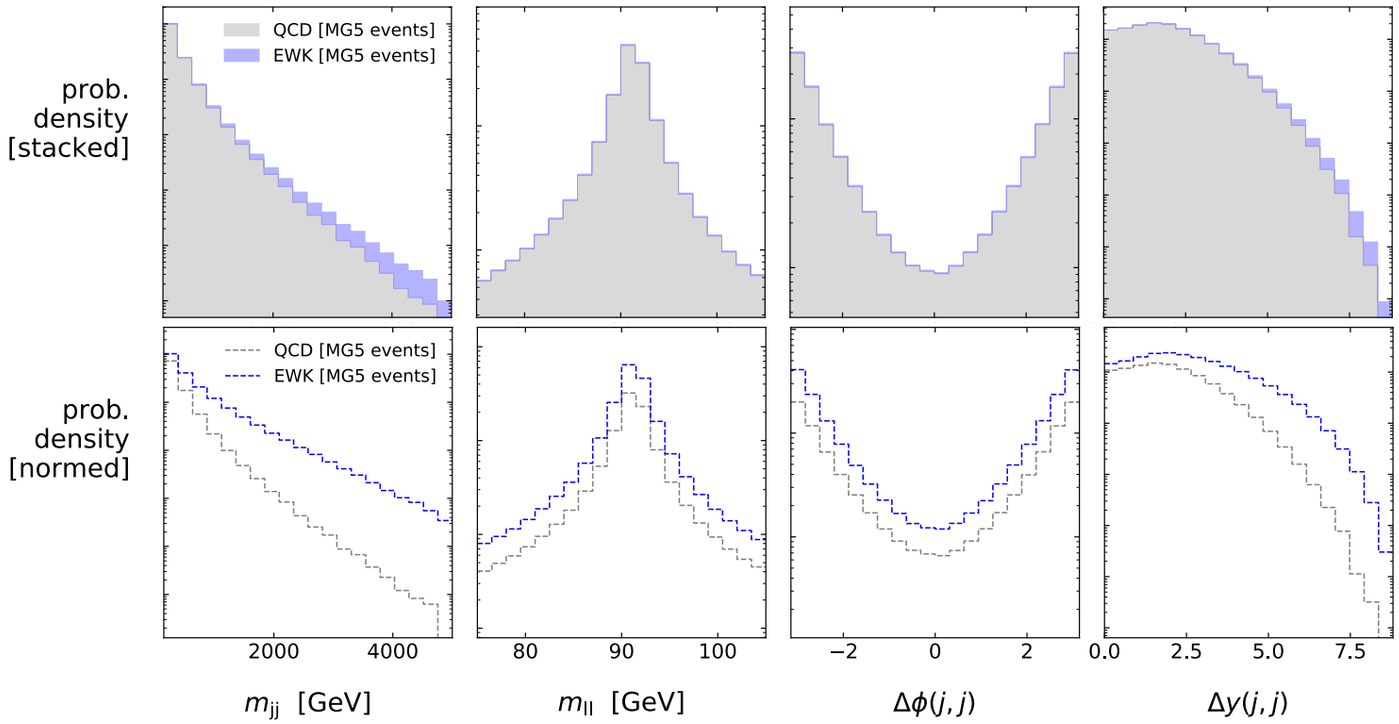
- Compute c.d.f. of data and c.d.f. of a reference distribution on latent space
- Mapping between c.d.f.s provides invertible non-linear projection function

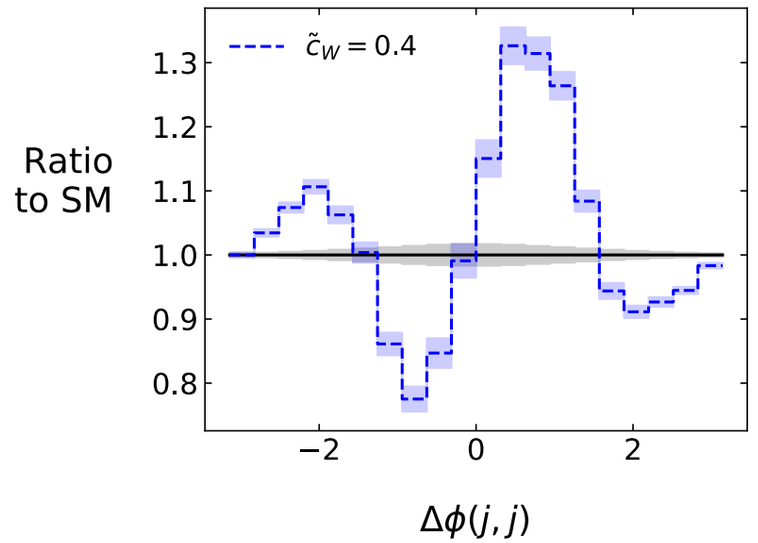




Real-world example: EWK Z + jets production

$\sigma^{\text{gen}}_{\text{QCD}} = 13.82\ \text{pb}^{-1}$ $\sigma^{\text{gen}}_{\text{EWK}} = 0.29\ \text{pb}^{-1}$





EWK signal appears at high m_{jj}

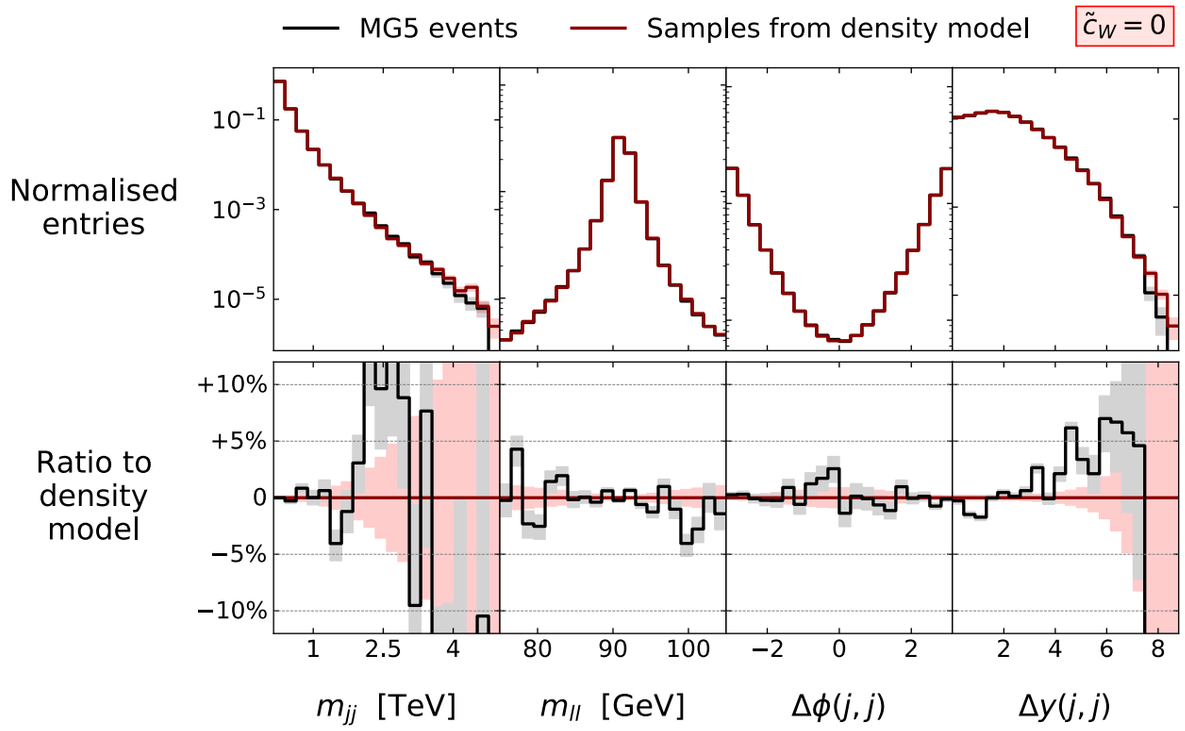
cannot separate sig/bkg (important to learn this too!)

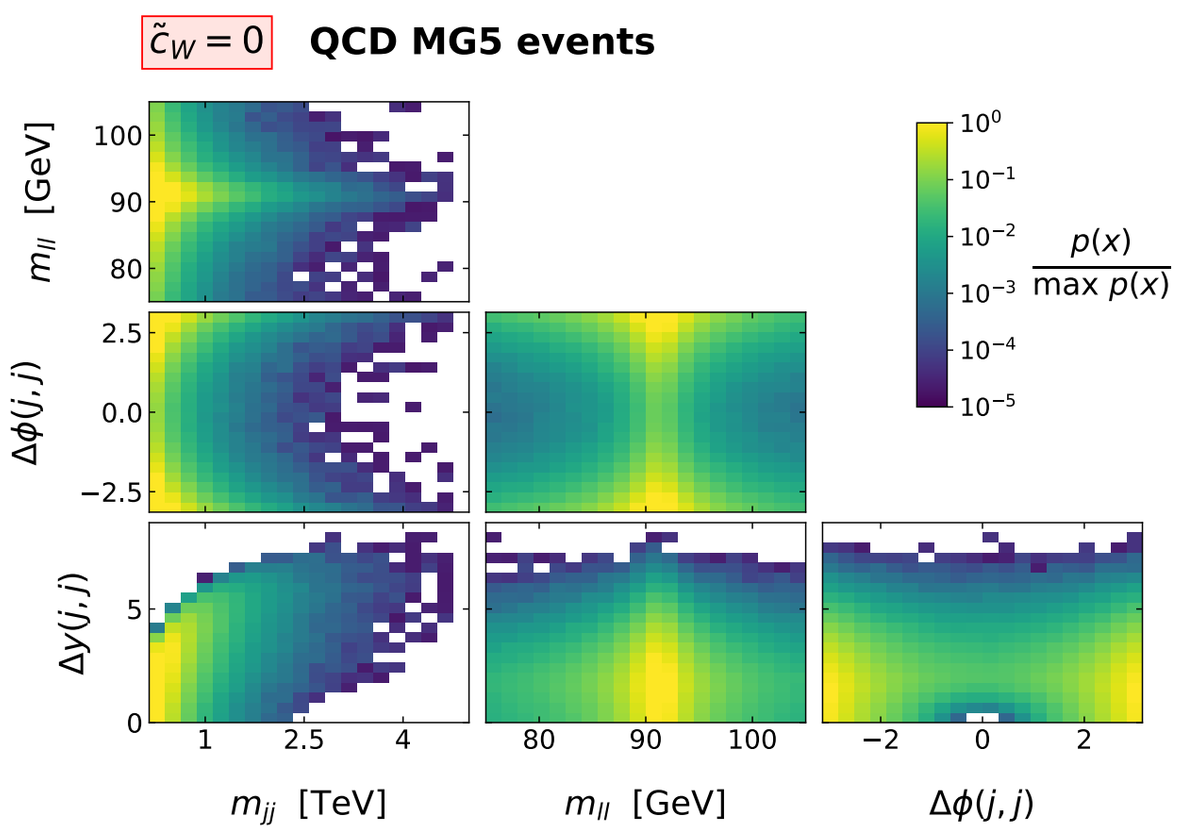
Δy(j,j) correlated with m_{jj}

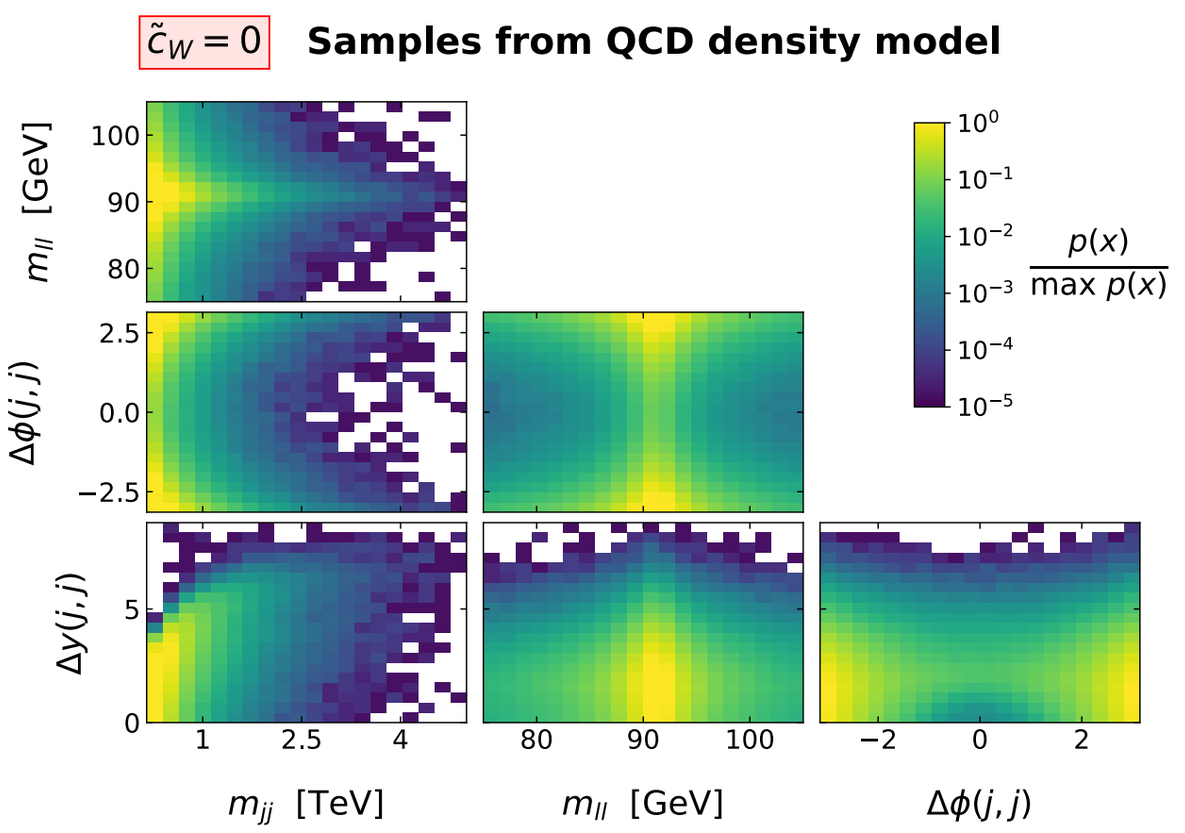
ΔΦ(j,j) deformed by variations of $\tilde{c}_W$

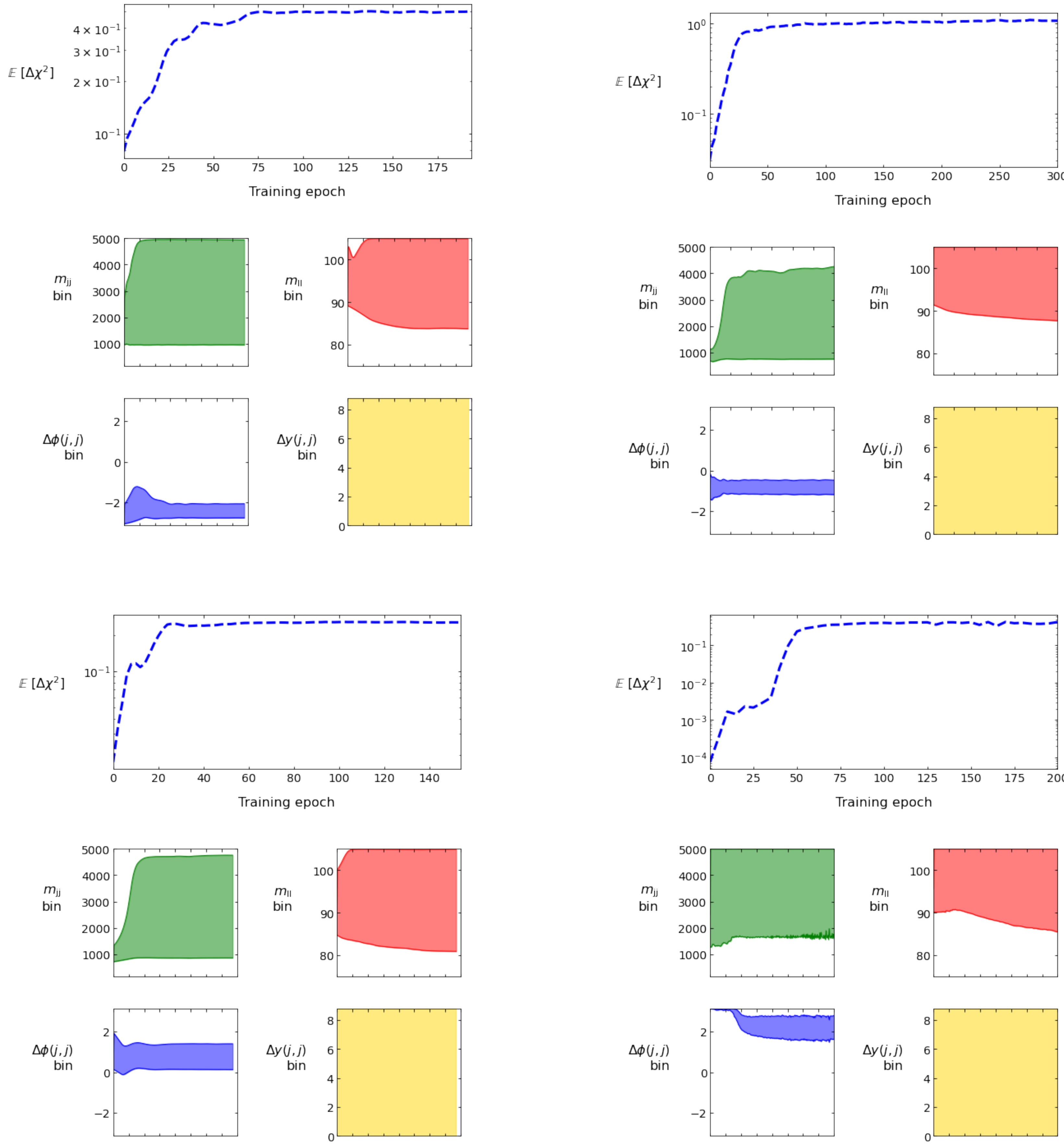
Key features of the data

- **Aim:** use ML workflow to design analysis, suitable for automation & **application when we don't know answer**
- This is just a well-understood example! Method contains nothing specific to the QCD / EWK / SMEFT, so useful for all analyses.
- See arXiv:2006.15458 for ATLAS experimental analysis

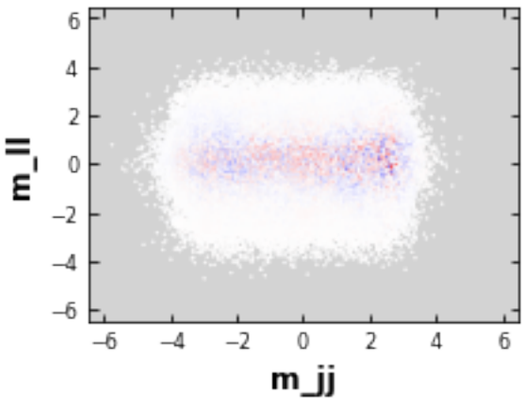

MG5 events
Samples from density model
$\tilde{c}_W = 0$
Normalised entries
10^{-1}
10^{-3}
10^{-5}
Ratio to density model
+10%
+5%
0
−5%
−10%
1
2.5
4
80
90
100
−2
0
2
2
4
6
8
m_{jj} [TeV]
m_{ll} [GeV]
$\Delta\phi(j,j)$
$\Delta y(j,j)$

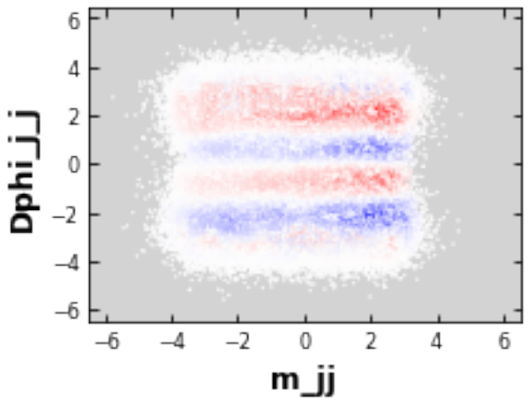

$\tilde{c}_W = 0$
QCD MG5 events
m_{ll} [GeV]
100
90
80
$\Delta\phi(j,j)$
2.5
0.0
−2.5
$\Delta y(j,j)$
5
0
1
2.5
4
80
90
100
−2
0
2
m_{jj} [TeV]
m_{ll} [GeV]
$\Delta\phi(j,j)$
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
$\frac{p(x)}{\max p(x)}$

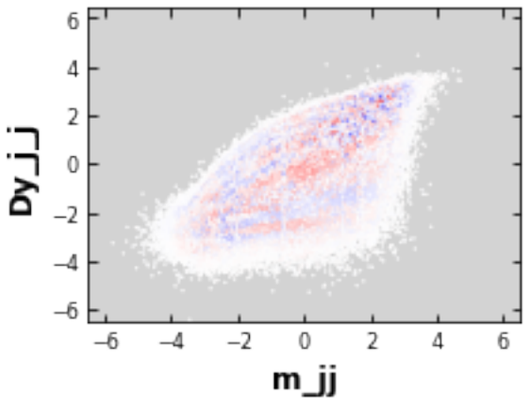

$\tilde{c}_W = 0$
Samples from QCD density model
m_{ll} [GeV]
100
90
80
$\Delta\phi(j,j)$
2.5
0.0
−2.5
$\Delta y(j,j)$
5
0
1
2.5
4
80
90
100
−2
0
2
m_{jj} [TeV]
m_{ll} [GeV]
$\Delta\phi(j,j)$
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
$\frac{p(x)}{\max p(x)}$

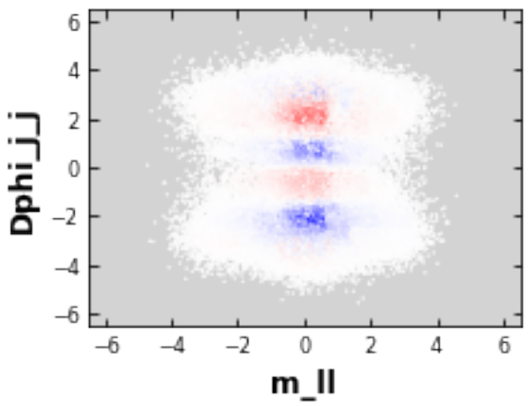


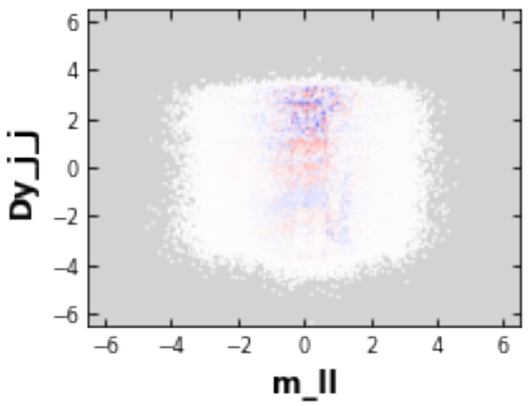
$\mathbb{E}\,[\Delta\chi^2]$
Training epoch
m_{jj} bin
m_{ll} bin
$\Delta\phi(j,j)$ bin
$\Delta y(j,j)$ bin
$\mathbb{E}\,[\Delta\chi^2]$
Training epoch
m_{jj} bin
m_{ll} bin
$\Delta\phi(j,j)$ bin
$\Delta y(j,j)$ bin
$\mathbb{E}\,[\Delta\chi^2]$
Training epoch
m_{jj} bin
m_{ll} bin
$\Delta\phi(j,j)$ bin
$\Delta y(j,j)$ bin
$\mathbb{E}\,[\Delta\chi^2]$
Training epoch
m_{jj} bin
m_{ll} bin
$\Delta\phi(j,j)$ bin
$\Delta y(j,j)$ bin

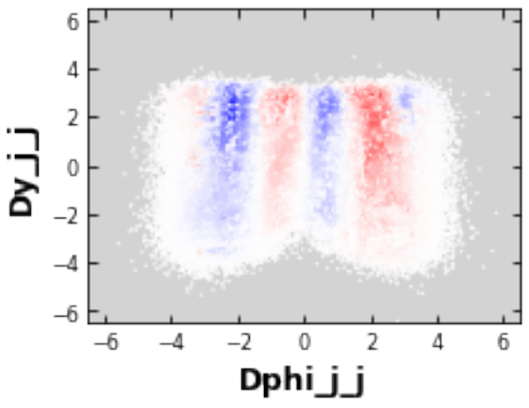

m_ll
m_jj


Dphi_j_j
m_jj


Dy_j_j
m_jj


Dphi_j_j
m_ll


Dy_j_j
m_ll


Dy_j_j
Dphi_j_j