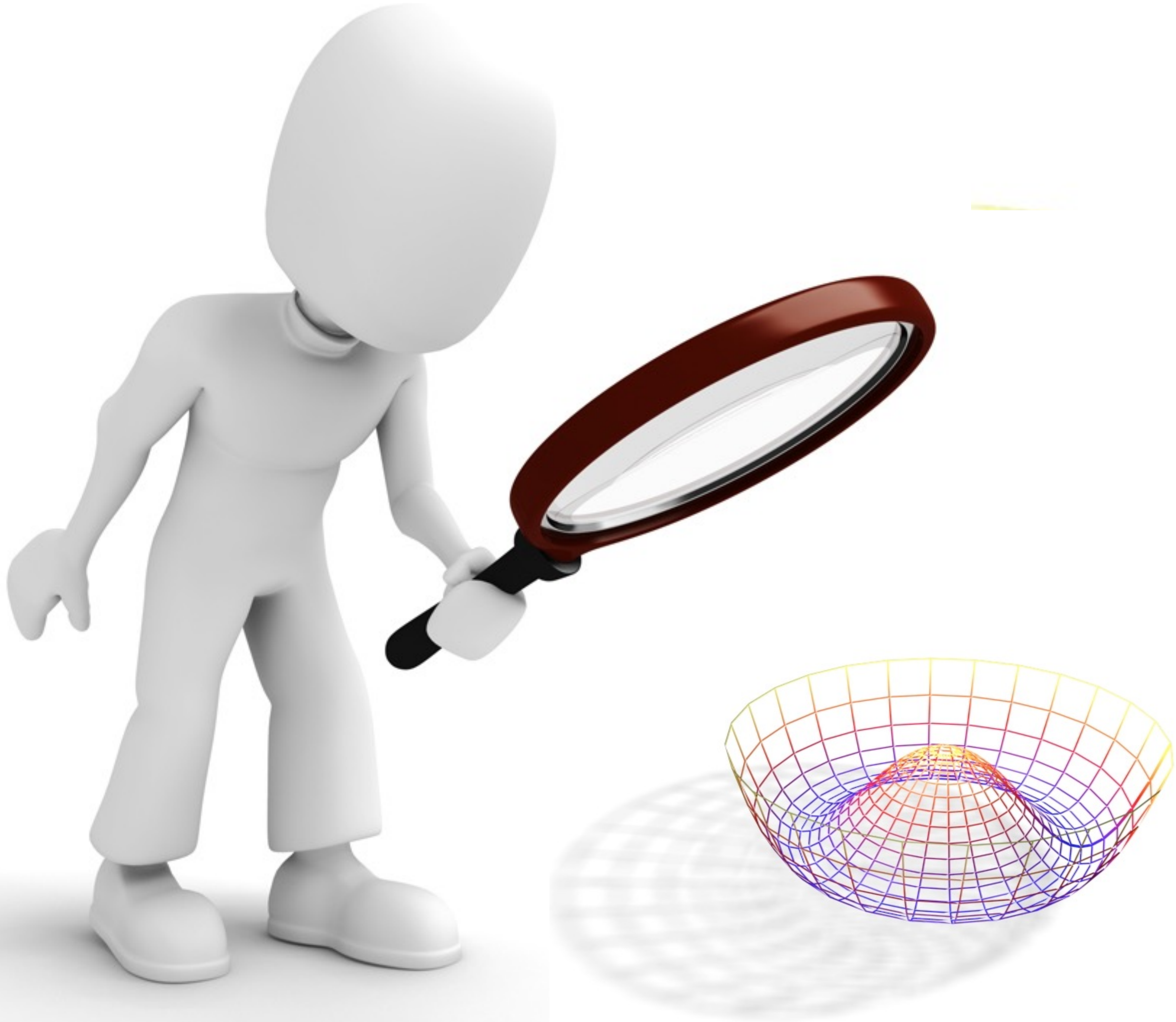
NYU Center for
Data Science

Center for
Cosmology and
Particle Physics

# NEW ANALYSIS TECHNIQUES

+ TOOLS & INFRASTRUCTURE



**@KyleCranmer**
New York University
Department of Physics
Center for Data Science
CILVR Lab

# Acknowledgements


Johann Brehmer


Gilles Louppe


Juan Pavez


Markus Stoye

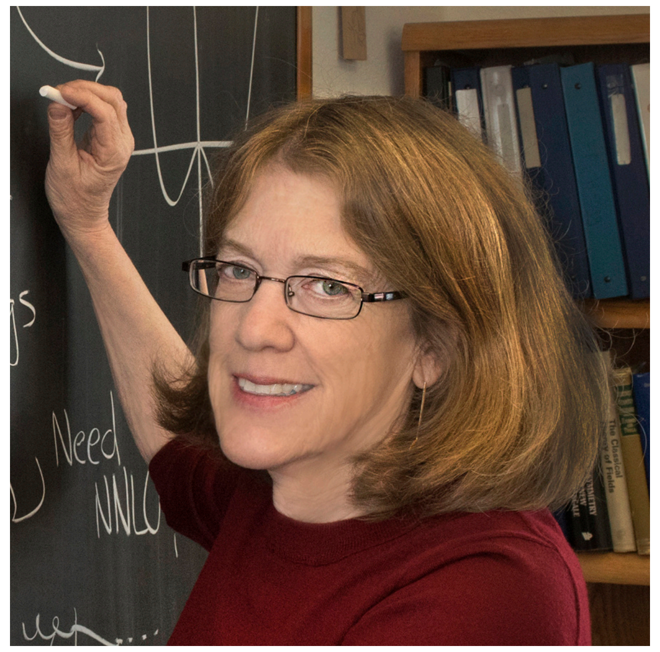

Felix Kling


Irina Espejo


Sinclert Perez

Special thanks to Johann for slides I borrowed


Tilman Plehn


Sally Dawson


Sam Homiller


Giordon Stark


Matthew Feickert


Lukas Heinrich

NYU

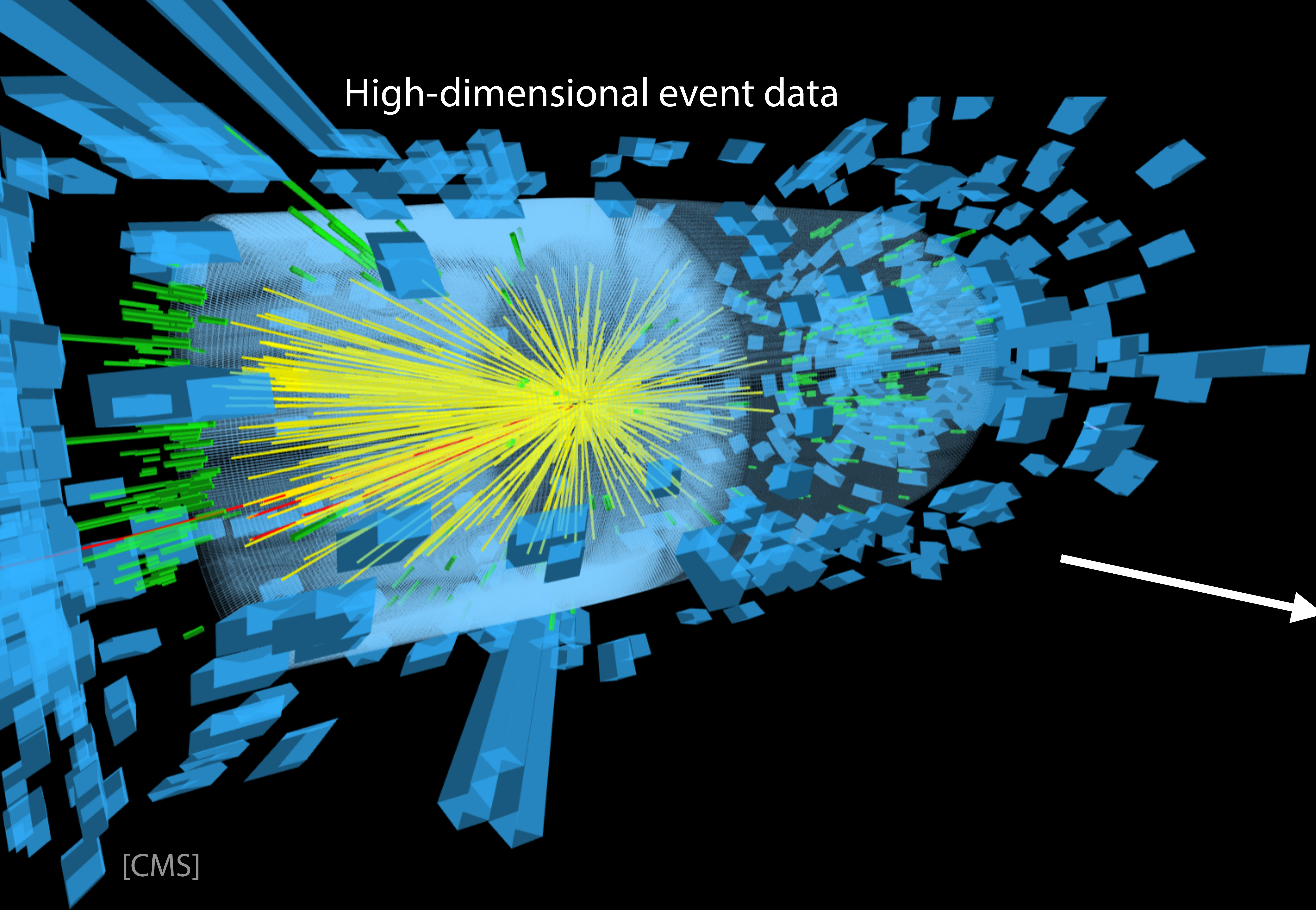
MSDSE
Moore - Sloan
Data Science Environments

dianahep

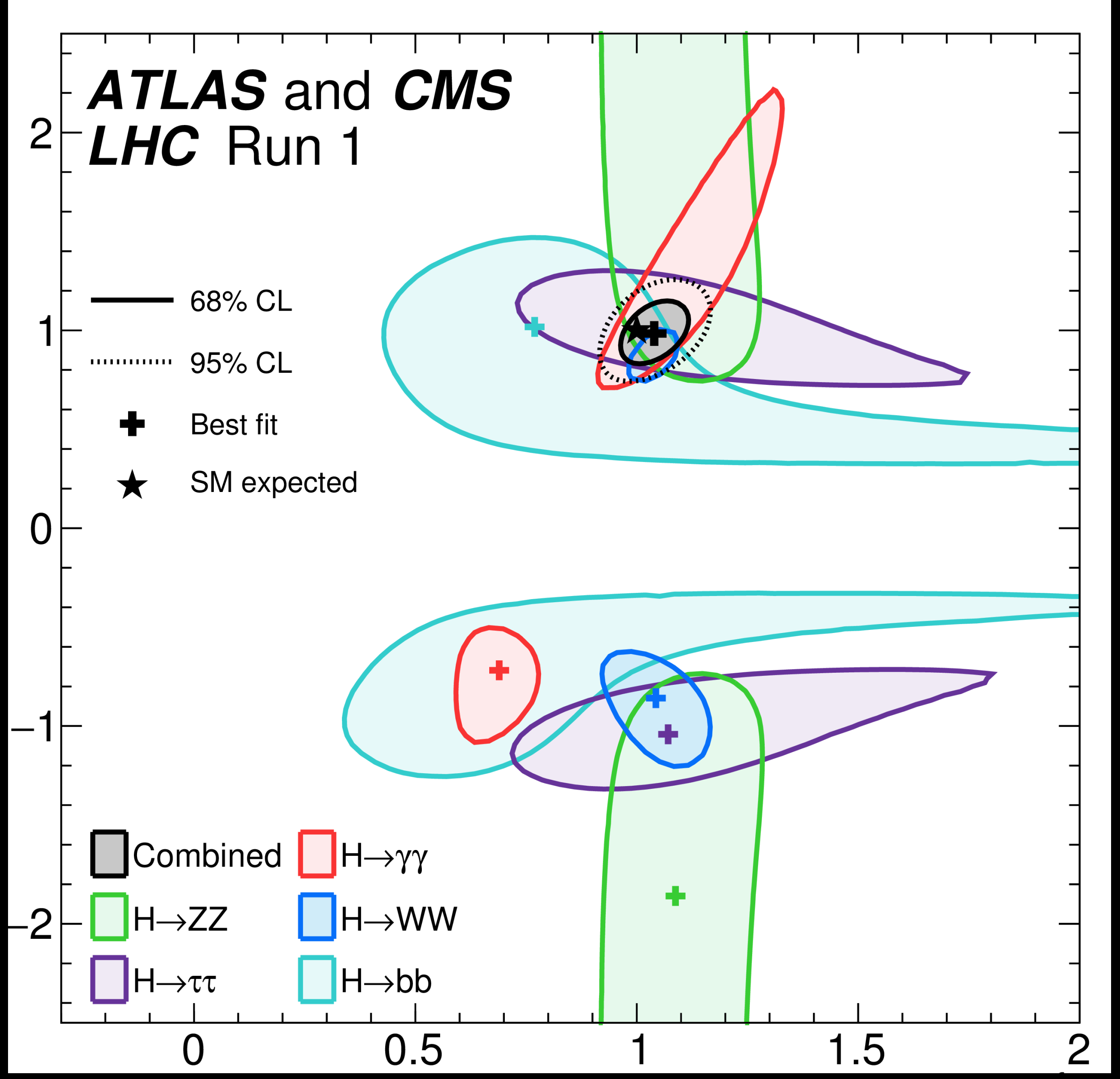
iris hep
Institute for Research and Innovation
in Software for High Energy Physics

**The SCAILFIN Project**
scailfin.github.io

High-dimensional event data



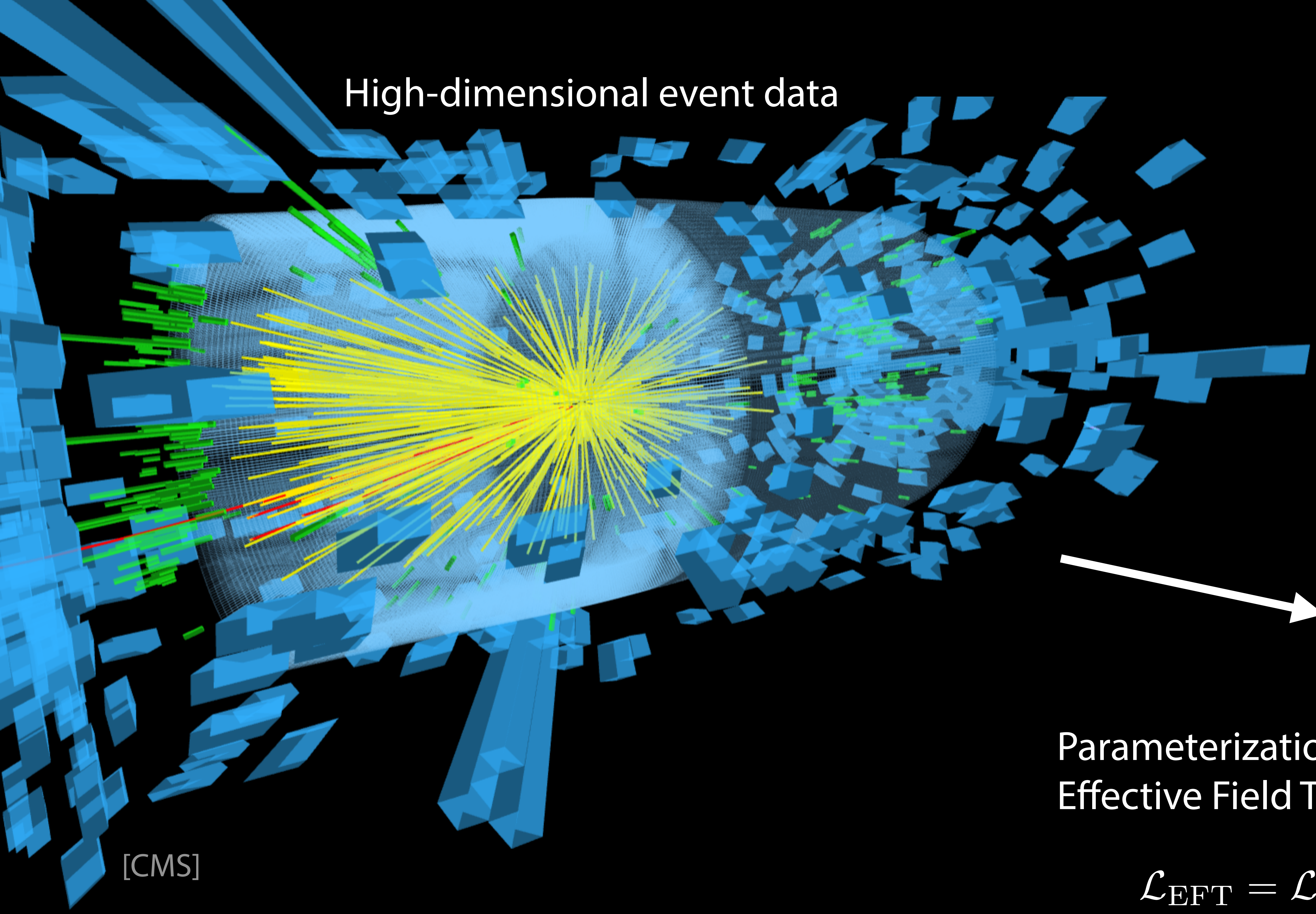
[CMS]



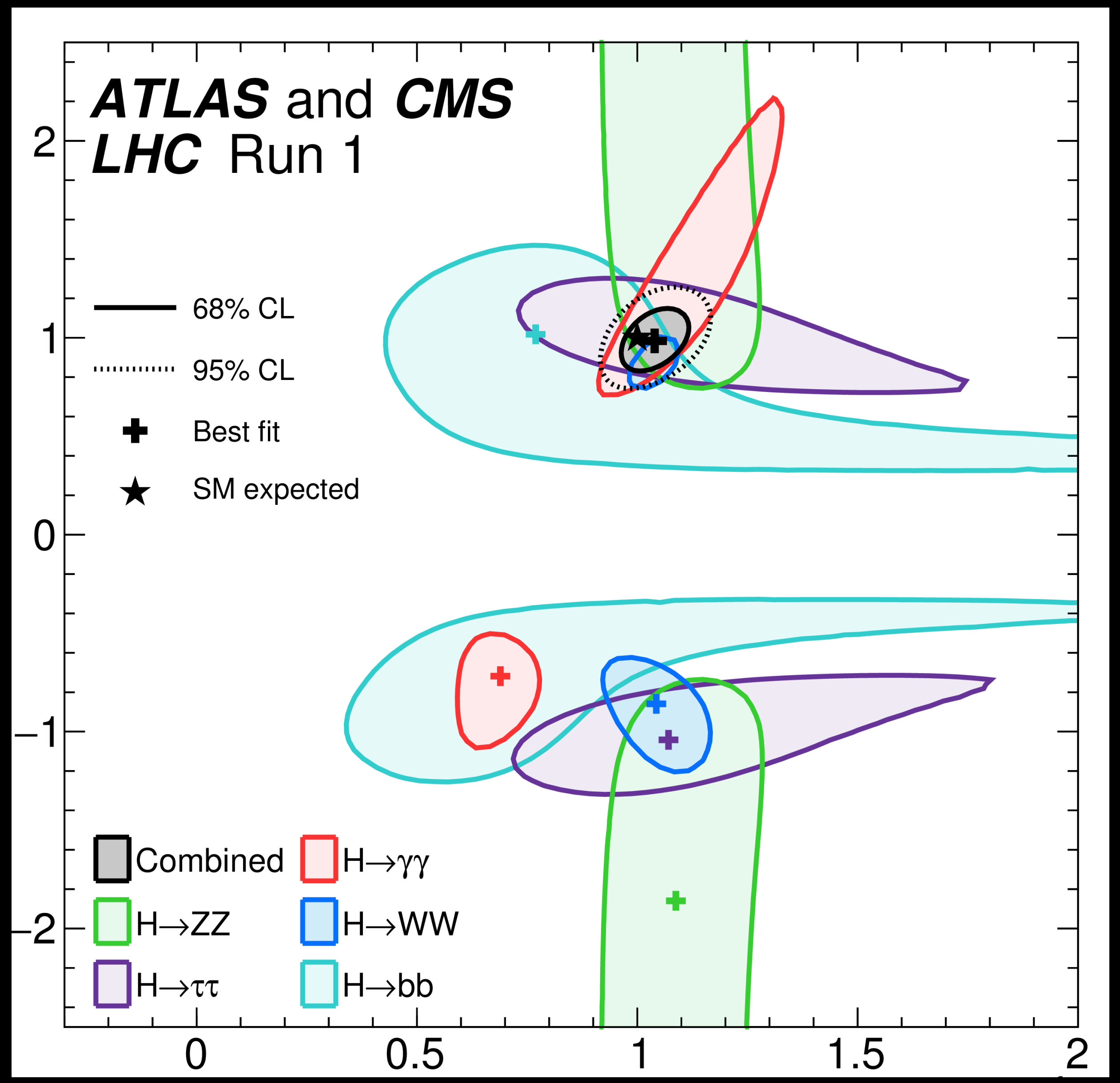

[ATLAS, CMS 1606.02266]

Precision constraints on new physics

High-dimensional event data



[CMS]




[ATLAS, CMS 1606.02266]

Precision constraints on new physics

Parameterization e.g. in Effective Field Theory:

$$\mathcal{L}_{\mathrm{EFT}} = \mathcal{L}_{\mathrm{SM}} + \sum_i \frac{f_i}{\Lambda^2} \mathcal{O}_i + \dots$$

systematic expansion of new physics around Standard Model

10s to 100s “universal” parameters to measure

# Reinterpretation
# The theory / experiment interface

# From yesterday

## Different analysis strategies

- Highly optimised analyses targeting specific properties / operators
  - → "best possible" sensitivity
  - → very model specific
- Fiducial and differential cross section measurements
  - → minimise model dependence
  - → relatively restricted sensitivity (hard to combine different channels)
  - → re-interpretable outside experiment
- Differential measurements in experimentally sensitive observables per production mode (STXS)
  - → model dependence from production mode definition
  - → easy combination of different Higgs decay channels → sensitivity to large number of EFT operators
  - → re-interpretable outside experiment

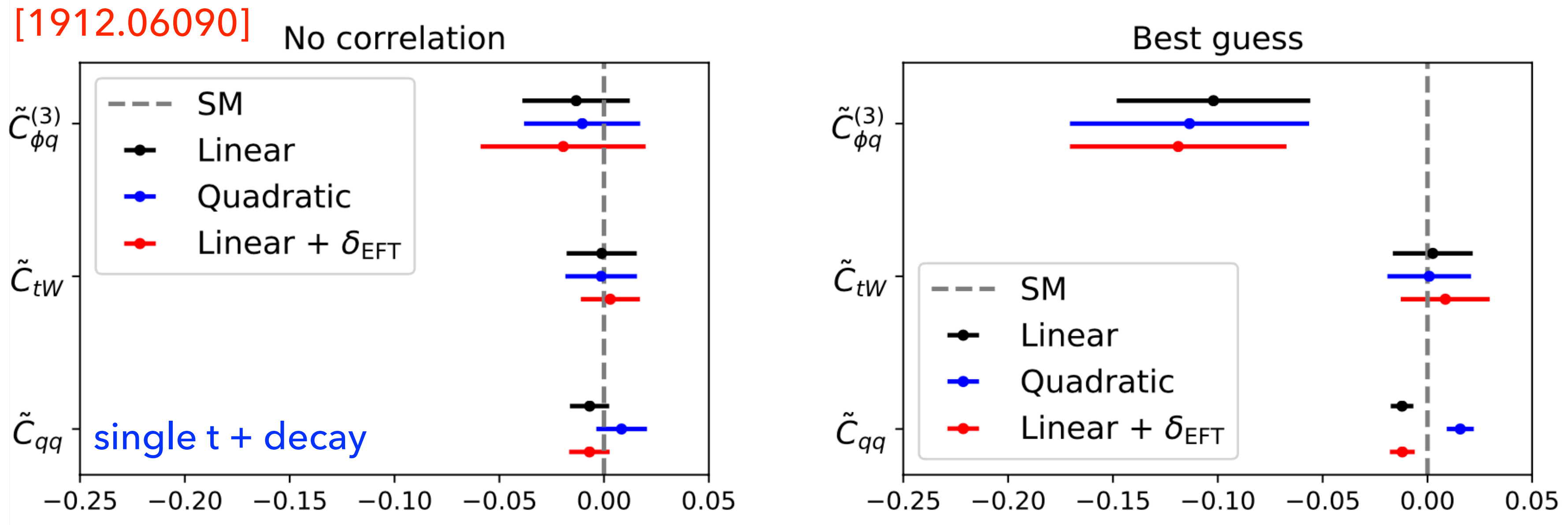
Saskia Falke | Inputs to EFT fits | 29/10/2020 | 3 /

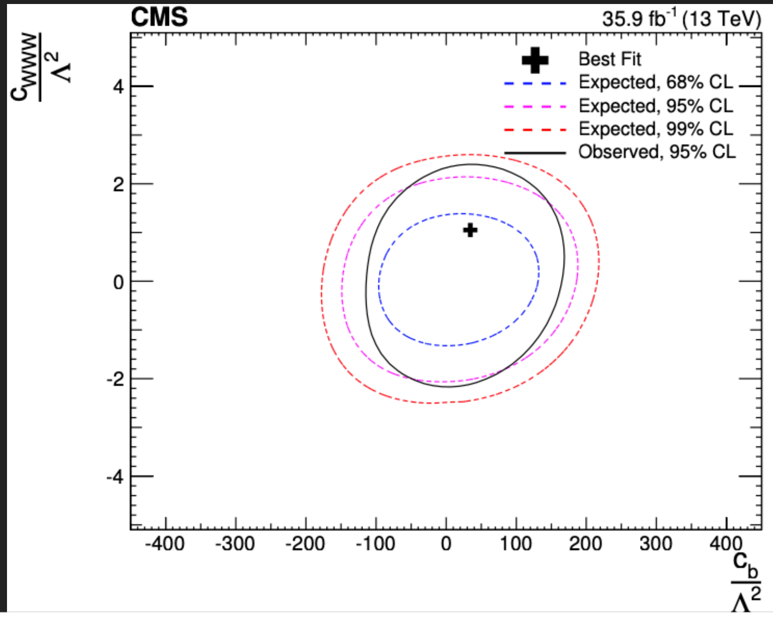
← Note the pros and cons

- This list of what is needed looks a lot like the discussions in reinterpretation of BSM searches ↓

WHAT IS NEEDED — OSCAR ÉBOLI — 12

- Information on experimental cut flows, efficiencies
- Information on backgrounds
- Information on results and corresponding correlations (becoming standard)
- Information on the likelihood
- Desirable to have results at particle level, and distributions (STXS or fiducial distr.)




- Makes me think of this ↓ from 2012

**Searches for New Physics: Les Houches Recommendations for the Presentation of LHC Results**

S. Kraml[1], B.C. Allanach[2], M. Mangano[3], H.B. Prosper[4], S. Sekmen[3,4] (editors), C. Balazs[5], A. Barr[6], P. Bechtle[7], G. Belanger[8], A. Belyaev[9,10], K. Benslama[11], M. Campanelli[12], K. Cranmer[13], A. De Roeck[3], M.J. Dolan[14], T. Eifert[15], J.R. Ellis[16,3], M. Felcini[17], B. Fuks[18], D. Guadagnoli[8,19], J.F. Gunion[20], S. Heinemeyer[17], J. Hewett[15], A. Ismail[15], M. Kadastik[21], M. Krämer[22], J. Lykken[23] F. Mahmoudi[3,24], S.P. Martin[25,26,27], T. Rizzo[15], T. Robens[28], M. Tytgat[29], A. Weiler[30]

# Comments on Unfolding

- Fiducial and differential cross section measurements
  - → minimise model dependence
  - → relatively restricted sensitivity (hard to combine different channels)
  - → re-interpretable outside experiment

- Desirable to have results at particle level, and distributions (STXS or fiducial distr.)

## Unfolding is deceptively attractive

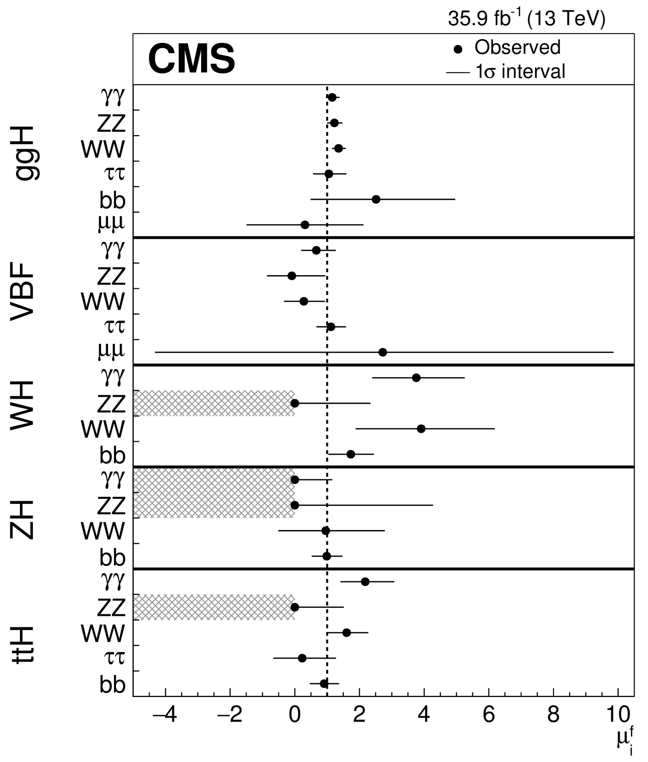
- It seems very convenient and to address what we want to know, but
- unfolding is a can of worms statistically and pushes many problems down stream
  - Combinations, correlated systematics, artifacts and bias introduced by the unfolding procedure. These will all turn into systematics in the final results.
- It is good for fast approximate answers, but
- I do not recommend it as a platform for the final "gold standard" results.

# Simplified Likelihoods

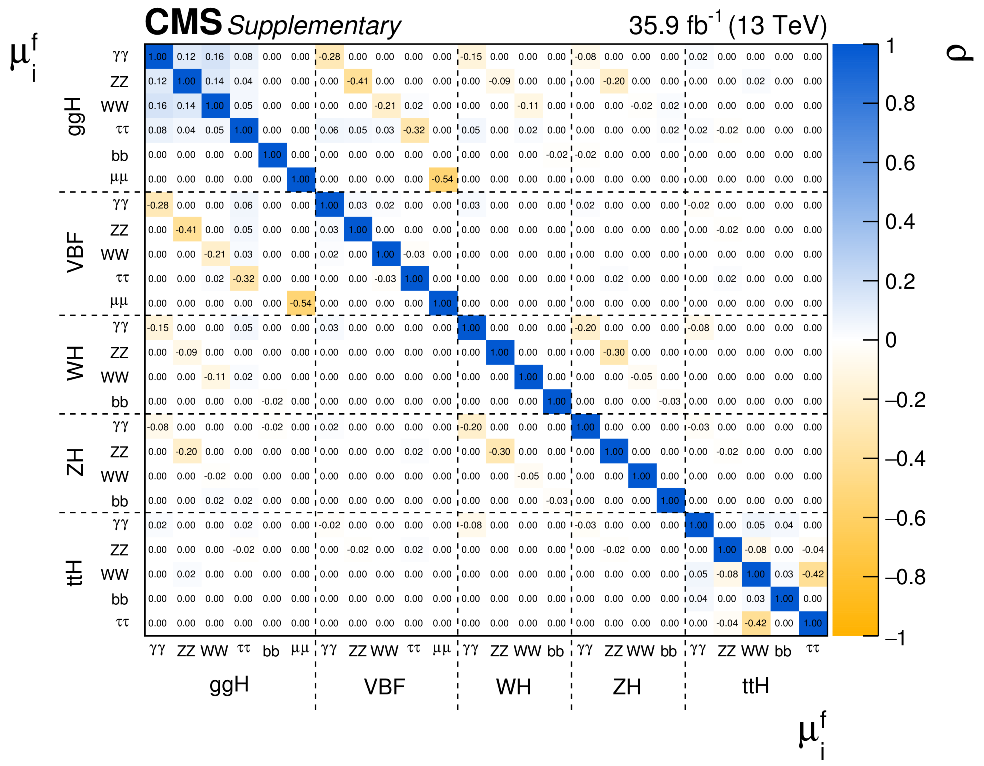
- Differential measurements in experimentally sensitive observables per production mode (STXS)
  - → model dependence from production mode definition
  - → easy combination of different Higgs decay channels → sensitivity to large number of EFT operators
  - → re-interpretable outside experiment

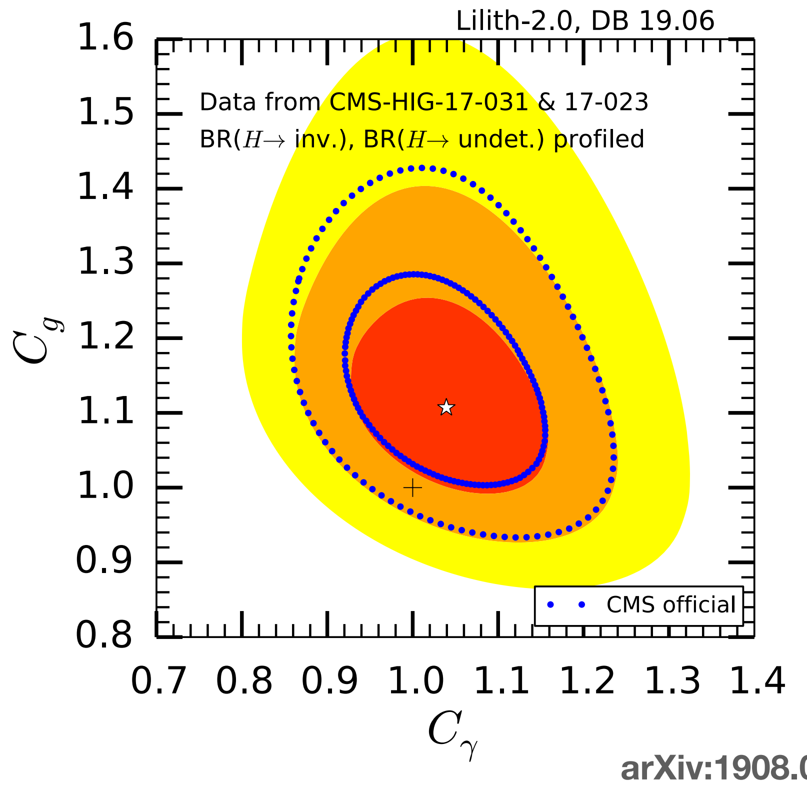
## So far: O±δO plus correlations (so

- Simplified likelihood, Gaussian approximation
  - e.g., Higgs measurements, channel-by-channel correlation matrix




CMS-HIG-17-031




arXiv:1908.03952

*«Correlation data […] has proven excellent for stabilising and ensuring better statistical definition in global fits as well as avoiding either overly conservative or over-enthusiastic interpretations.»*

Reinterpretation Forum Report, 2003.07868



This approach also doesn't expose sources of uncertainties needed to combine with other external measurements with shared systematics

- Simplified likelihood, Gaussian approximation
  - Extremely useful, but only an approximation, lots of relevant information is lost.
  - One problem is that possible non-Gaussian tails are ignored. Not an issue if the uncertainties are small. However, if the uncertainties are large the likelihood should be modelled in more detail.

https://indico.cern.ch/event/957797/contributions/4026032/

# Gold Standard: Full Likelihoods

## Different analysis strategies

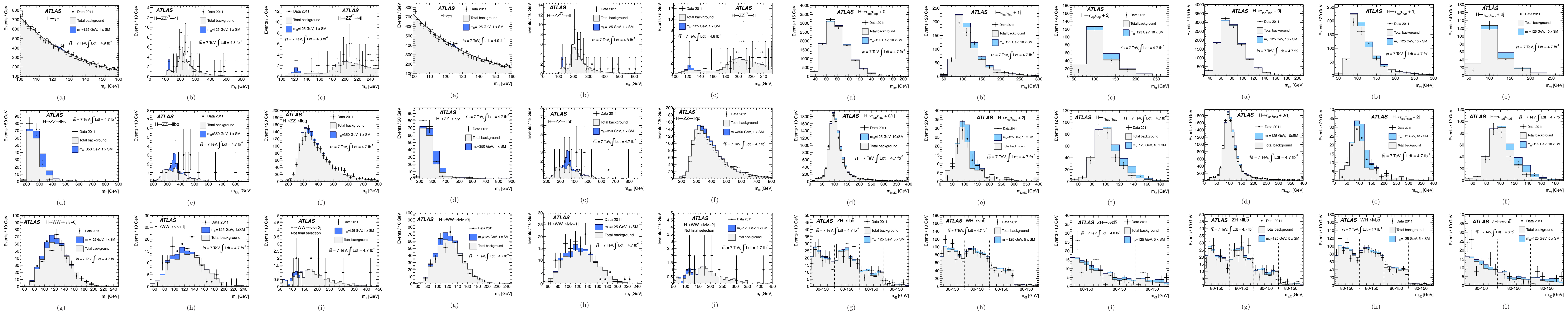
- Highly optimised analyses targeting specific properties / operators
  - → "best possible" sensitivity
  - → very model specific
- Fiducial and differential cross section measurements
  - → minimise model dependence
  - → relatively restricted sensitivity (hard to combine different channels)
  - → re-interpretable outside experiment
- Differential measurements in experimentally sensitive observables per production mode (STXS)
  - → model dependence from production mode definition
  - → easy combination of different Higgs decay channels → sensitivity to large number of EFT operators
  - → re-interpretable outside experiment

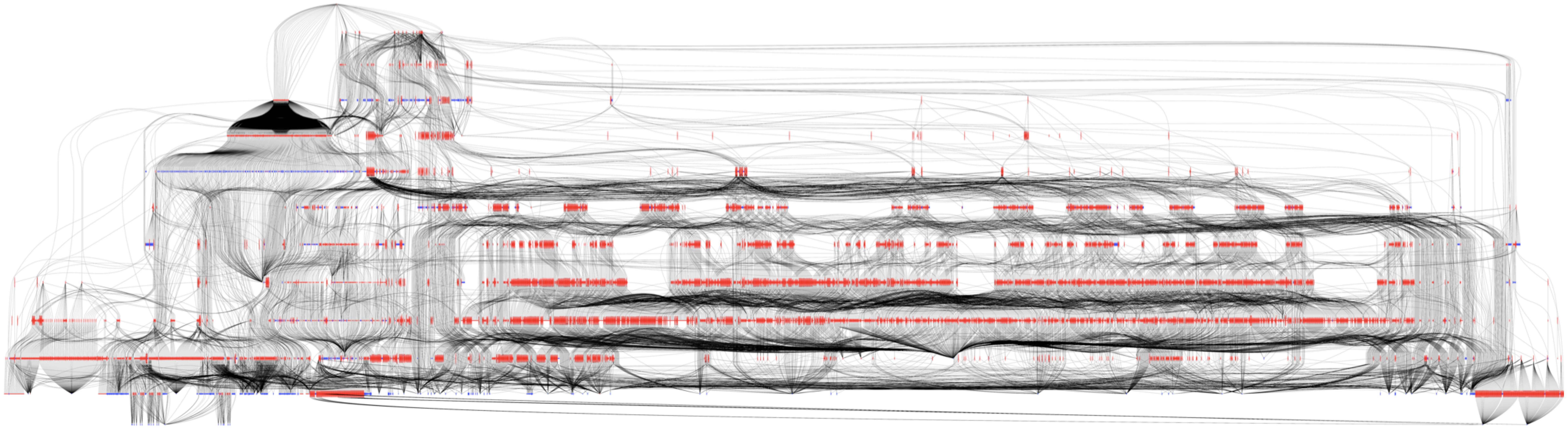


For EFT measurements, I think combined fits based on the full likelihood models in the "folded" data space are the most principled approach

- Likelihood can be based on STXS or dedicated analysis
- This should be the gold standard for the flagship EFT results

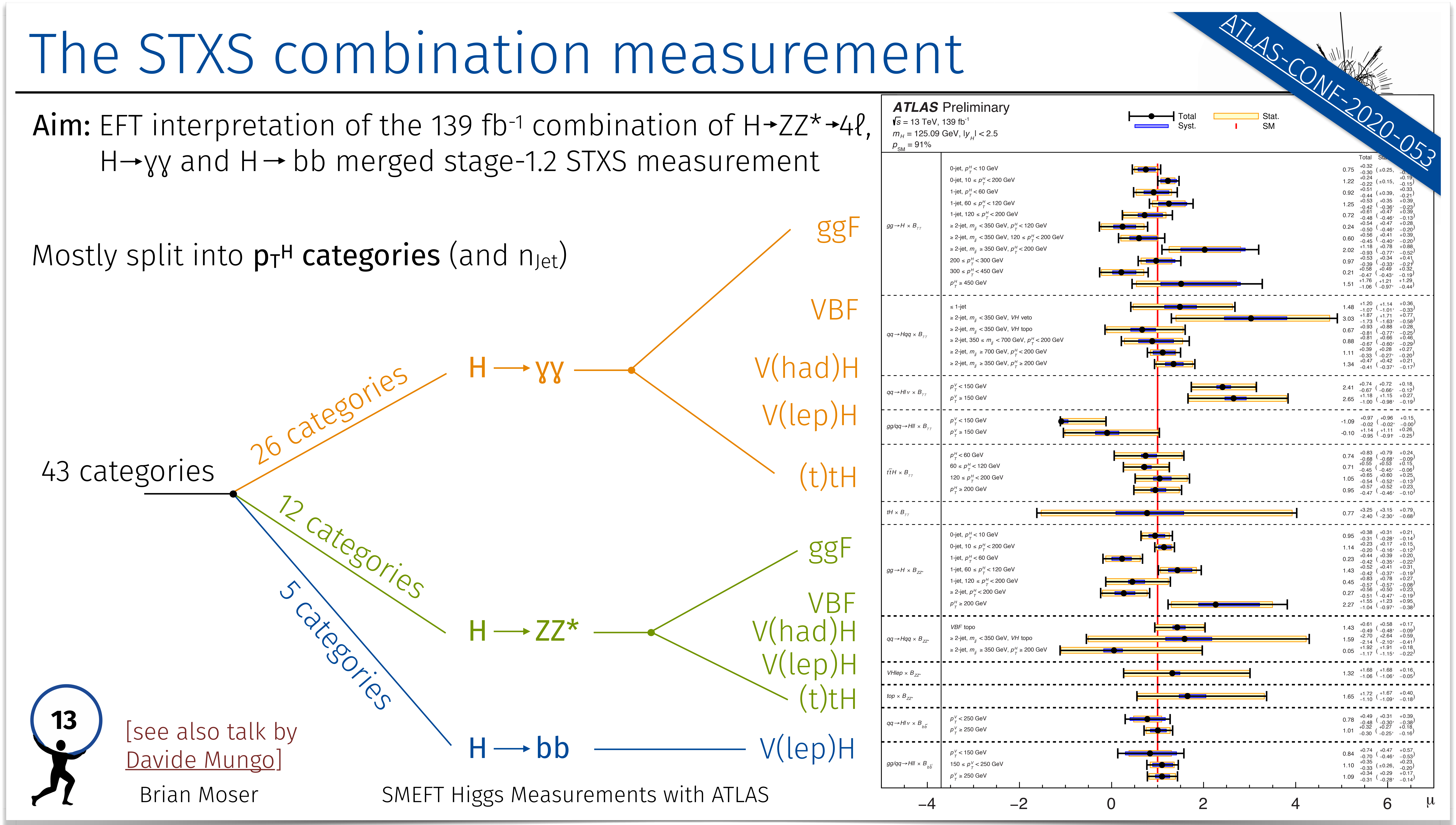
# Combined Fits for the Higgs discovery





$$\mathbf{f}_{\mathrm{tot}}(\mathcal{D}_{\mathrm{sim}}, \mathcal{G}|\boldsymbol{\alpha}) = \prod_{c \in \mathrm{channels}} \left[ \mathrm{Pois}(n_c|\nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce}|\boldsymbol{\alpha}) \right] \cdot \prod_{p \in \mathbb{S}} f_p(a_p|\alpha_p)$$

# Combined fits for EFTs



# 2012 recommendations

## Why public likelihoods

- The statistical model of an experimental analysis provides the complete mathematical description of that analysis

  *p(o|α)* relating the observed quantities *o* to the parameters *α*

- Given the likelihood, all the standard statistical approaches are available for extracting information from it

- Essential information for any detailed interpretation of experimental results

  = determining the compatibility of the observations with theoretical predictions

**Searches for New Physics: Les Houches Recommendations for the Presentation of LHC Results**

S. Kraml[1], B.C. Allanach[2], M. Mangano[3], H.B. Prosper[4], S. Sekmen[3,4] (editors), C. Balazs[5], A. Barr[6], P. Bechtle[7], G. Belanger[8], A. Belyaev[9,10], K. Benslama[11], M. Campanelli[12], K. Cranmer[13], A. De Roeck[3], M.J. Dolan[14], T. Eifert[15], J.R. Ellis[16,3], M. Felcini[17], B. Fuks[18], D. Guadagnoli[8,19], J.F. Gunion[20], S. Heinemeyer[17], J. Hewett[15], A. Ismail[15], M. Kadastik[21], M. Krämer[22], J. Lykken[23] F. Mahmoudi[3,24], S.P. Martin[25,26,27], T. Rizzo[15], T. Robens[28], M. Tytgat[29], A. Weiler[30]

**Les Houches Recommandations (2012)**

**3b:** When feasible, provide a mathematical description of the final likelihood function in which experimental data and parameters are clearly distinguished, either in the publication or the auxiliary information. Limits of validity should always be clearly specified.

**3c:** Additionally provide a digitized implementation of the likelihood that is consistent with the mathematical description.

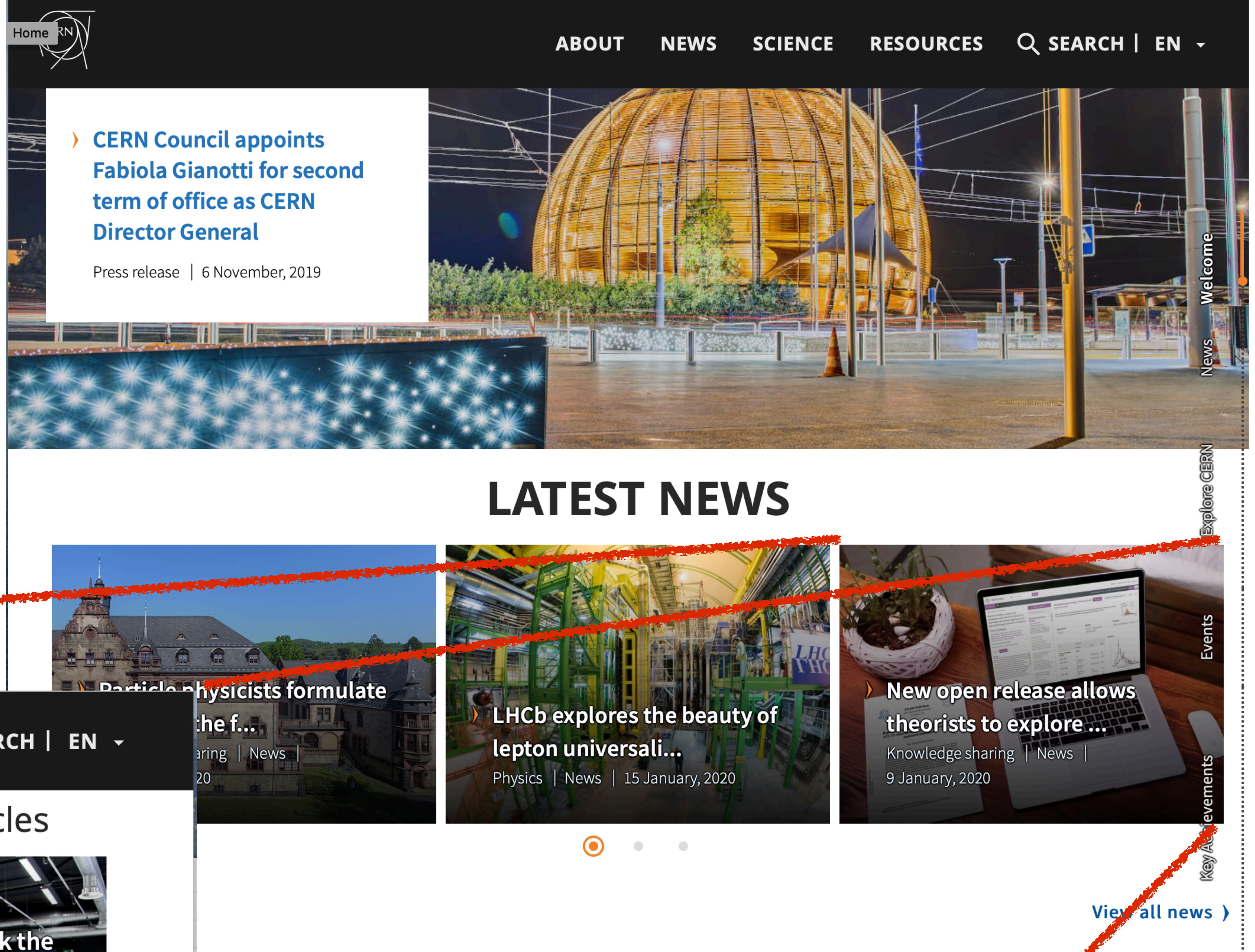
arXiv:1203.2489

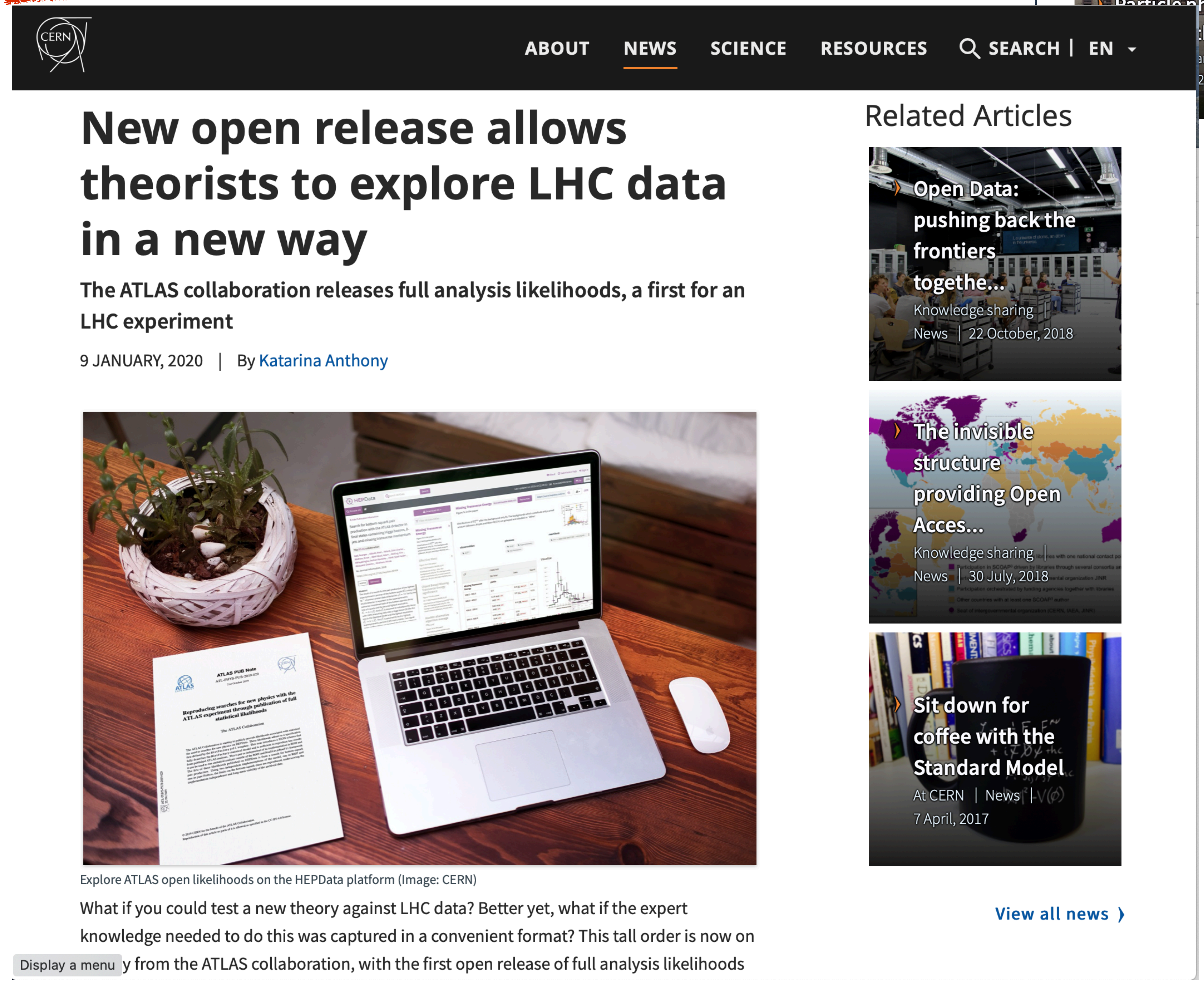
S. Kraml - Feedback on use of public likelihoods - 24 Sep 2020

# Recent progress

Recently ATLAS has started publishing full likelihoods to HEPData for SUSY and exotics searches

- Perfect for STXS-type measurements



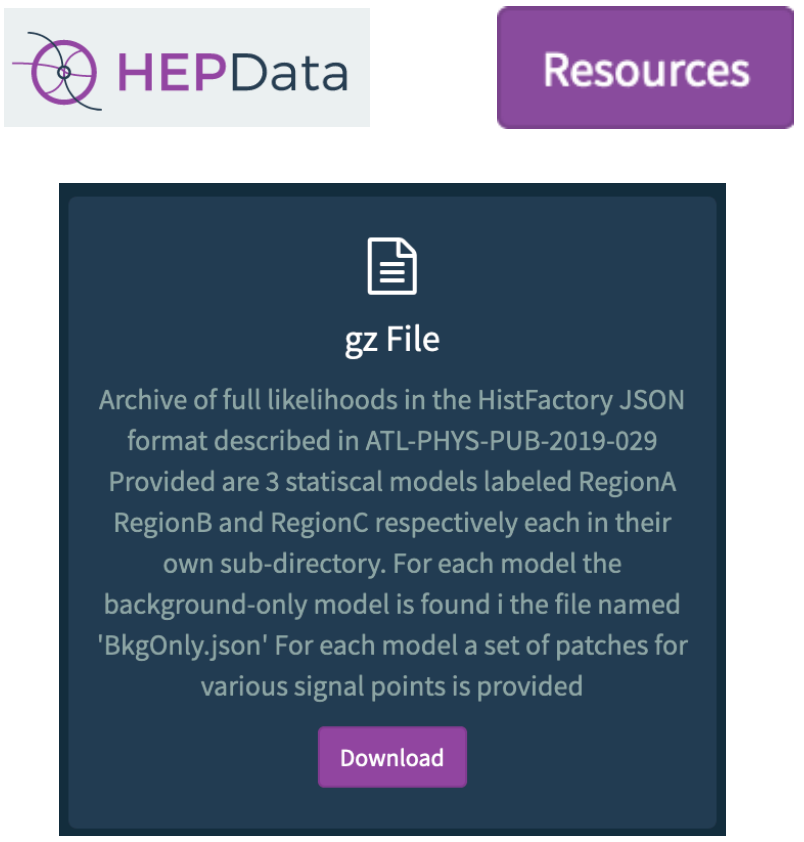

https://scikit-hep.org/pyhf/

# Theorist rejoice

## Now: full likelihoods !!

ATL-PHYS-PUB-2019-029 (05 Aug 2019)

- Plain-text serialisation of HistFactory workspaces, JSON format
  - Provides background estimates, changes under systematic variations, and observed data counts at the same fidelity as used in the experiment.




| | Description | Modification | Constraint Term $c_\chi$ | Input |
|---|---|---|---|---|
| constrained | Uncorrelated Shape | $\kappa_{scb}(\gamma_b) = \gamma_b$ | $\prod_b \text{Pois}\left(r_b = \sigma_b^{-2} \middle| \rho_b = \sigma_b^{-2}\gamma_b\right)$ | $\sigma_b$ |
| | Correlated Shape | $\Delta_{scb}(\alpha) = f_p\left(\alpha \middle| \Delta_{scb,\alpha=-1}, \Delta_{scb,\alpha=1}\right)$ | $\text{Gaus}\,(a = 0 \vert\, \alpha, \sigma = 1)$ | $\Delta_{scb,\alpha=\pm1}$ |
| | Normalisation Unc. | $\kappa_{scb}(\alpha) = g_p\left(\alpha \middle| \kappa_{scb,\alpha=-1}, \kappa_{scb,\alpha=1}\right)$ | $\text{Gaus}\,(a = 0 \vert\, \alpha, \sigma = 1)$ | $\kappa_{scb,\alpha=\pm1}$ |
| | MC Stat. Uncertainty | $\kappa_{scb}(\gamma_b) = \gamma_b$ | $\prod_b \text{Gaus}\left(a_{\gamma_b} = 1 \middle| \gamma_b, \delta_b\right)$ | $\delta_b^2 = \sum_s \delta_{sb}^2$ |
| | Luminosity | $\kappa_{scb}(\lambda) = \lambda$ | $\text{Gaus}\,(l = \lambda_0 \vert\, \lambda, \sigma_\lambda)$ | $\lambda_0, \sigma_\lambda$ |
| free | Normalisation | $\kappa_{scb}(\mu_b) = \mu_b$ | | |
| | Data-driven Shape | $\kappa_{scb}(\gamma_b) = \gamma_b$ | | |

Rate modifications defined in HistFactory for bin *b*, sample *s*, channel *c*.

  - Usage: RooFit, **pyhf**
  - Target: long-term data/analysis preservation, reinterpretation purposes

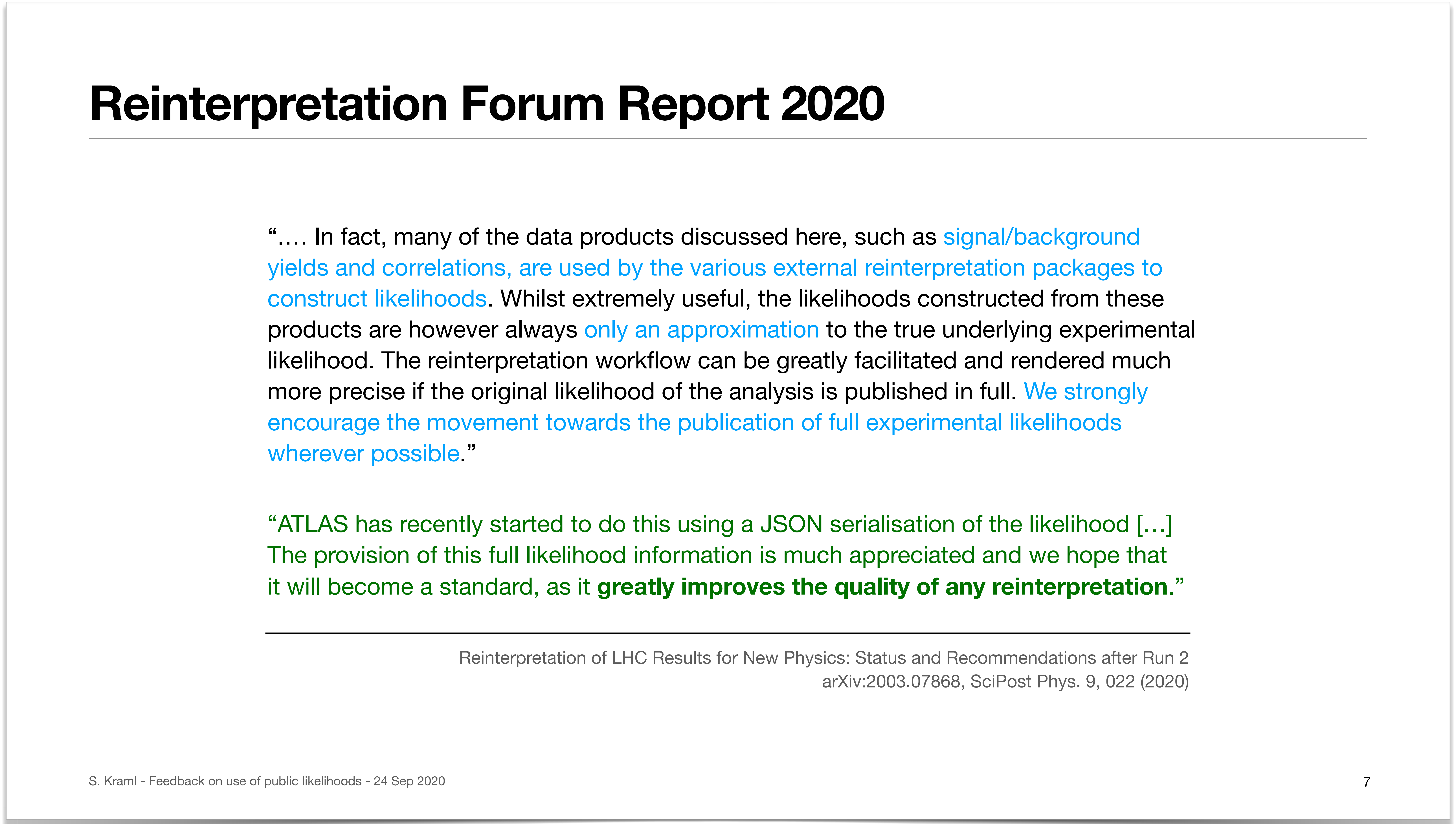
So far available for 4/12 SUSY analyses with 139 $fb^{-1}$

| | |
|---|---|
| SUSY-2018-31 (1908.03122) | multi-b sbottom: 2b+2H(bb) |
| SUSY-2018-04 (1911.06660 ) | stau search, 2 hadr. taus |
| SUSY-2019-08 (1909.09226) | 1 lept. + H(bb), EW-ino |
| SUSY-2018-06 (1912.08479 ) | 3 lept. EW-ino |



https://indico.cern.ch/event/957797/contributions/4026032/

# Theorist rejoice




https://indico.cern.ch/event/957797/contributions/4026032/

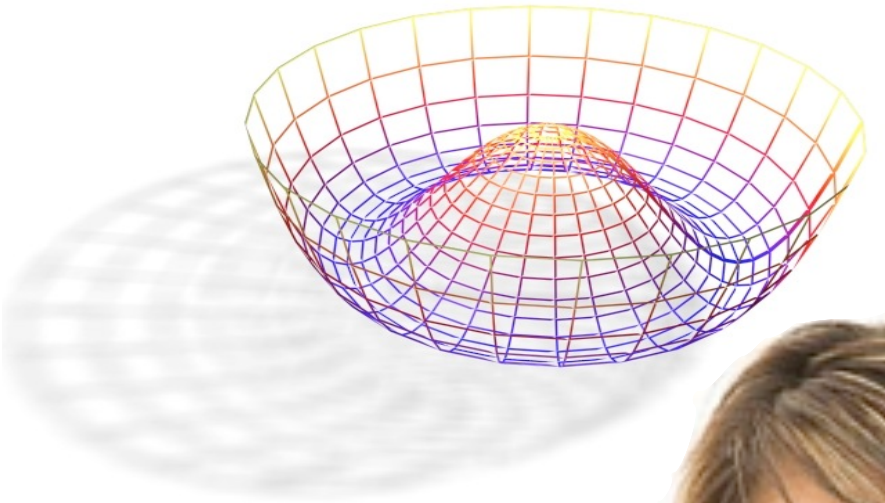
Likelihoods don’t address reinterpretation for signals with different final states or kinematics (eg. Exotic Higgs)

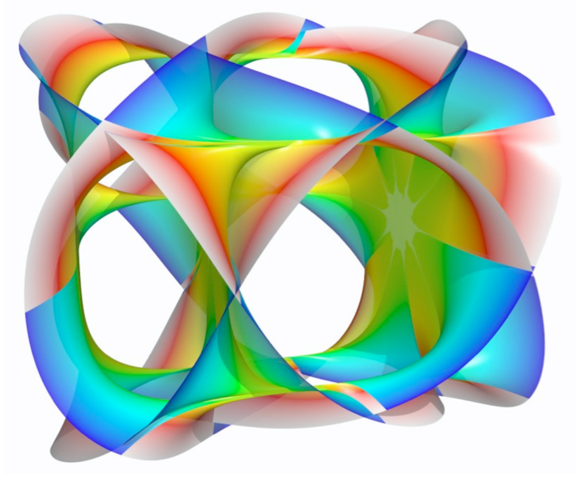
# THEORY recast SERVICE

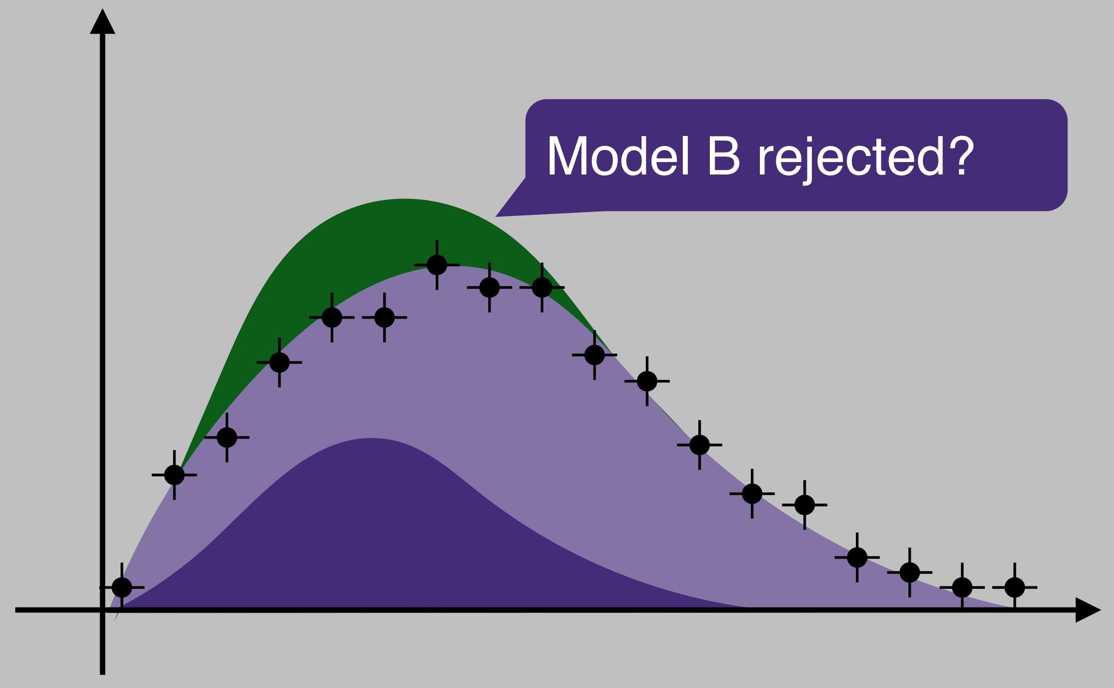
$$
\begin{aligned}
\mathcal{L}_{SM} = \; & \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
& + \underbrace{\bar{L}\gamma^\mu(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)L + \bar{R}\gamma^\mu(i\partial_\mu - \frac{1}{2}g'YB_\mu)R}_{\text{kinetic energies and electroweak interactions of fermions}} \\
& + \underbrace{\frac{1}{2}\left|(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)\,\phi\right|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{and Higgs masses and couplings}} \\
& + \underbrace{g''(\bar{q}\gamma^\mu T_a q)\,G^a_\mu}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1\bar{L}\phi R + G_2\bar{L}\phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}}
\end{aligned}
$$

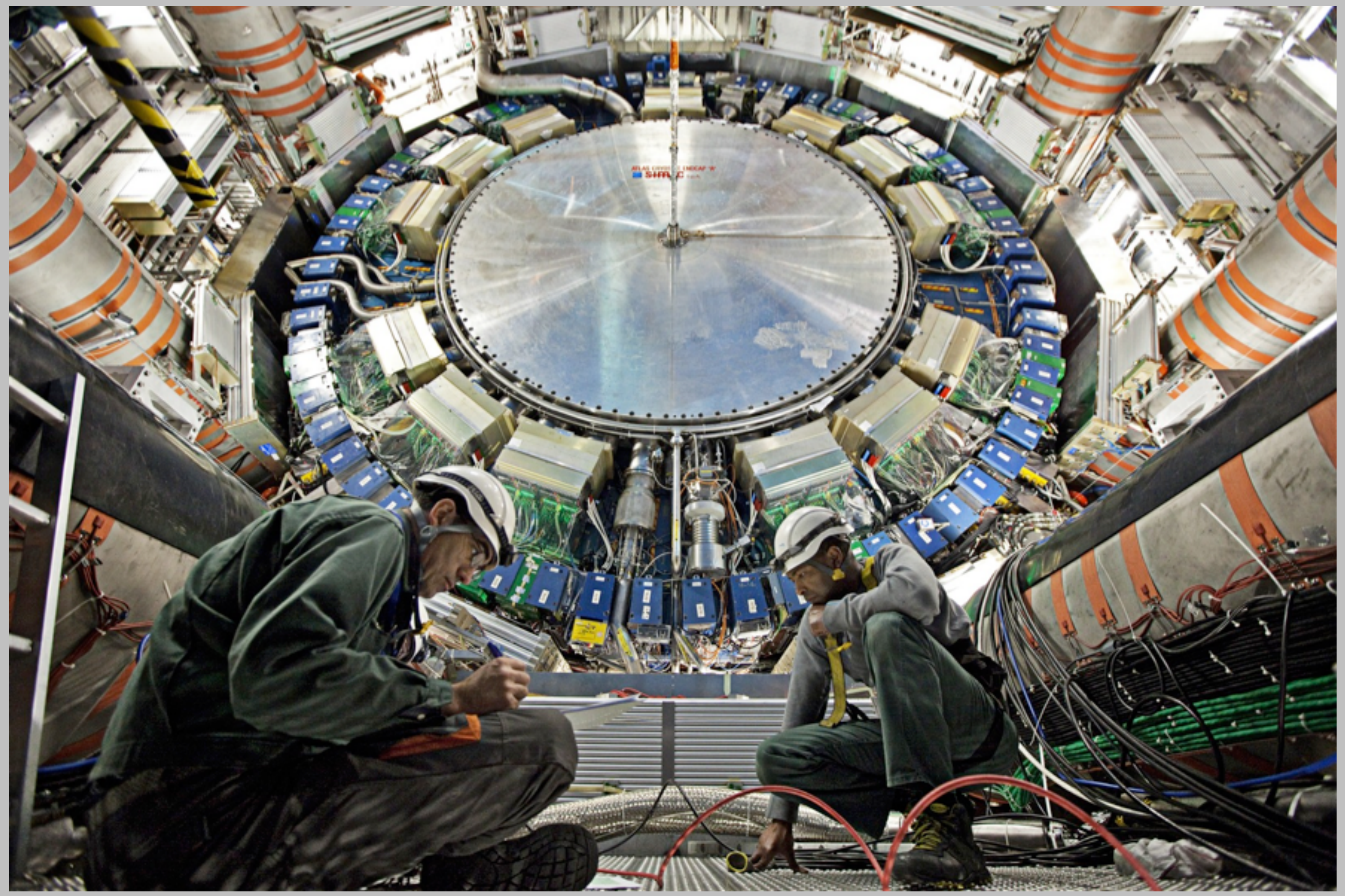
Q

A







# Analysis Preservation

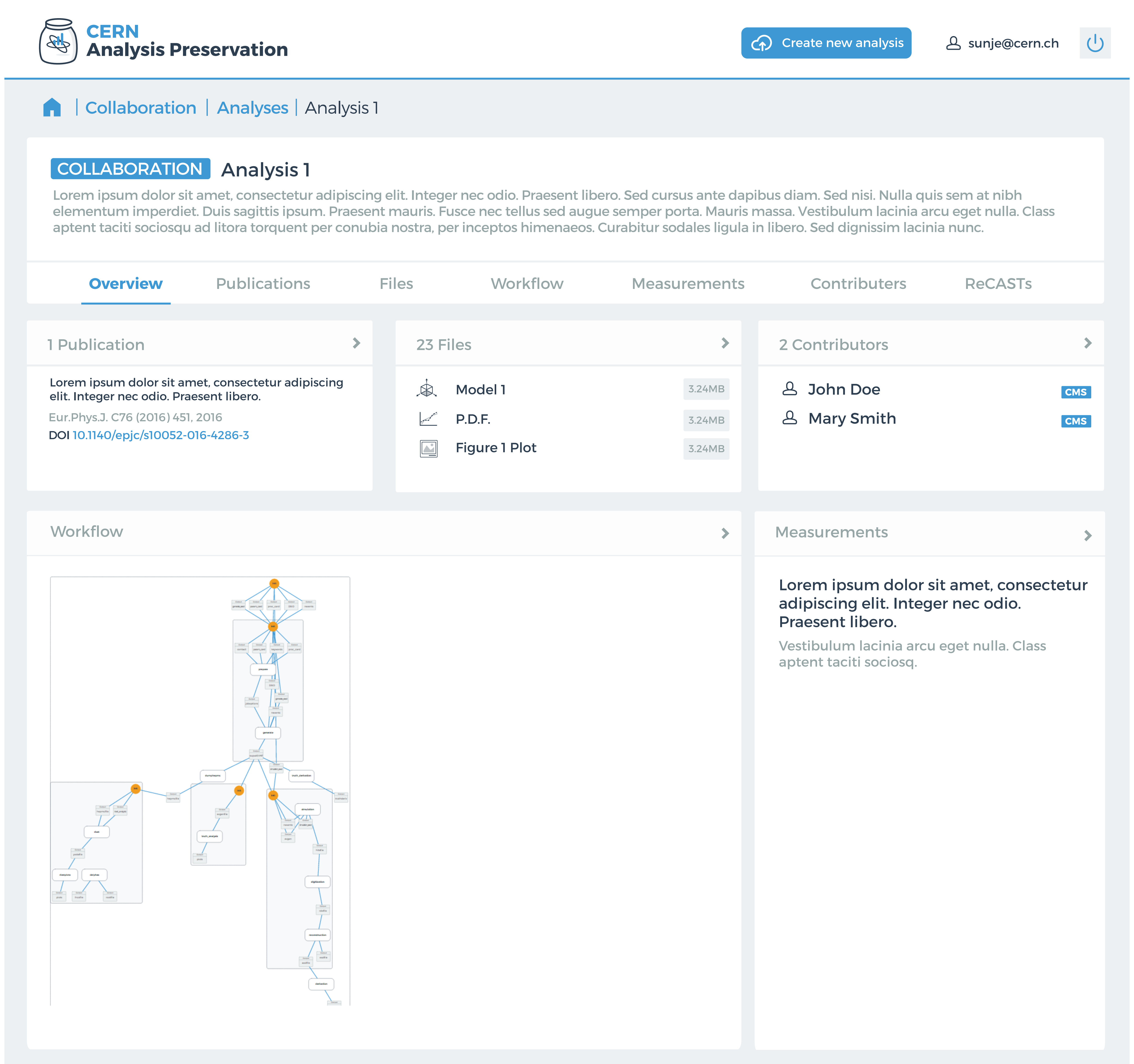
Recently we've made enormous progress in preserving the full analysis chain.

- Makes it possible to run a new new signal through the full analysis chain

reana

Reproducible research data analysis platform

| Flexible | Scalable | Reusable | Free |
| --- | --- | --- | --- |
| Run many computational workflow engines. | Support for remote compute clouds. | Containerise once, reuse elsewhere. Cloud-native. | Free Software. MIT licence. Made with ❤ at CERN. |
| COMMON WORKFLOW LANGUAGE | kubernetes | | |




**CERN Analysis Preservation**

https://doi.org/10.1038/s41567-018-0342-2

# RECAST in action

## ATLAS has started using RECAST to reinterpret SUSY and exotics searches

- Also relevant for exotic BSM Higgs scenarios

**ATLAS PUB Note**

ATL-PHYS-PUB-2019-032

11th August 2019

**RECAST framework reinterpretation of an ATLAS Dark Matter Search constraining a model of a dark Higgs boson decaying to two $b$-quarks**

The ATLAS Collaboration

The reinterpretation of a search for dark matter produced in association with a Higgs boson decaying to $b$-quarks performed with RECAST, a software framework designed to facilitate the reinterpretation of existing searches for new physics, is presented. Reinterpretation using RECAST is enabled through the sustainable preservation of the original data analysis as re-executable declarative workflows using modern cloud technologies and integrated with the wider CERN Analysis Preservation efforts. The reinterpretation targets a model predicting dark matter production in association with a hypothetical dark Higgs boson decaying into $b$-quarks where the mass of the dark Higgs boson $m_s$ is a free parameter, necessitating a faithful reinterpretation of the analysis. The dataset has an integrated luminosity of 79.8 $fb^{-1}$ and was recorded with the ATLAS detector at the Large Hadron Collider at a centre-of-mass energy of $\sqrt{s}$ = 13 TeV. Constraints on the parameter space of the dark Higgs model for a fixed choice of dark matter mass $m_\chi$ = 200 GeV exclude model configurations with a mediator mass up to 3.2 TeV.

ATL-PHYS-PUB-2019-032
12 August 2019



**ATLAS PUB Note**

ATL-PHYS-PUB-2020-007

27th March 2020

**Reinterpretation of the ATLAS Search for Displaced Hadronic Jets with the RECAST Framework**

The ATLAS Collaboration

A recent ATLAS search for displaced jets in the hadronic calorimeter is preserved in RECAST and thereafter used to constrain three new physics models not studied in the original work. A Stealth SUSY model and a Higgs-portal baryogenesis model, both predicting long-lived particles and therefore displaced decays, are probed for proper decay lengths between a few cm and 500 m. A dark sector model predicting Higgs and heavy boson decays to collimated hadrons via long-lived dark photons is also probed. The cross-section times branching ratio for the Higgs channel is constrained between a few millimetres and a few metres, while for a heavier 800 GeV boson the constraints extend from tenths of a millimetre to a few tens of metres. The original data analysis workflow was completely captured using virtualisation techniques, allowing for an accurate and efficient reinterpretation of the published result in terms of new signal models following the RECAST protocol.

ATL-PHYS-PUB-2020-007
28/03/2020

# RECAST + STXS overcomes model dependence

## Different analysis strategies

- Highly optimised analyses targeting specific properties / operators
  - → "best possible" sensitivity
  - → very model specific
- Fiducial and differential cross section measurements
  - → minimise model dependence
  - → relatively restricted sensitivity (hard to combine different channels)
  - → re-interpretable outside experiment
- Differential measurements in experimentally sensitive observables per production mode (STXS)
  - → model dependence from production mode definition
  - → easy combination of different Higgs decay channels → sensitivity to large number of EFT operators
  - → re-interpretable outside experiment

Saskia Falke | Inputs to EFT fits | 29/10/2020 | 3 / 22

The model dependence in STXS mainly connected to how results are conveyed.

- The phase space regions are just phase space regions, they don't assume any model
- Paired with RECAST one could reinterpret any model using the STXS phase space regions

***Likelihood Publishing + RECAST*** =

# Going fully differential

# STXS vs. dedicated analysis

## Different analysis strategies

- Highly optimised analyses targeting specific properties / operators
  - → "best possible" sensitivity
  - → very model specific
- Fiducial and differential cross section measurements
  - → minimise model dependence
  - → relatively restricted sensitivity (hard to combine different channels)
  - → re-interpretable outside experiment
- Differential measurements in experimentally sensitive observables per production mode (STXS)
  - → model dependence from production mode definition
  - → easy combination of different Higgs decay channels → sensitivity to large number of EFT operators
  - → re-interpretable outside experiment



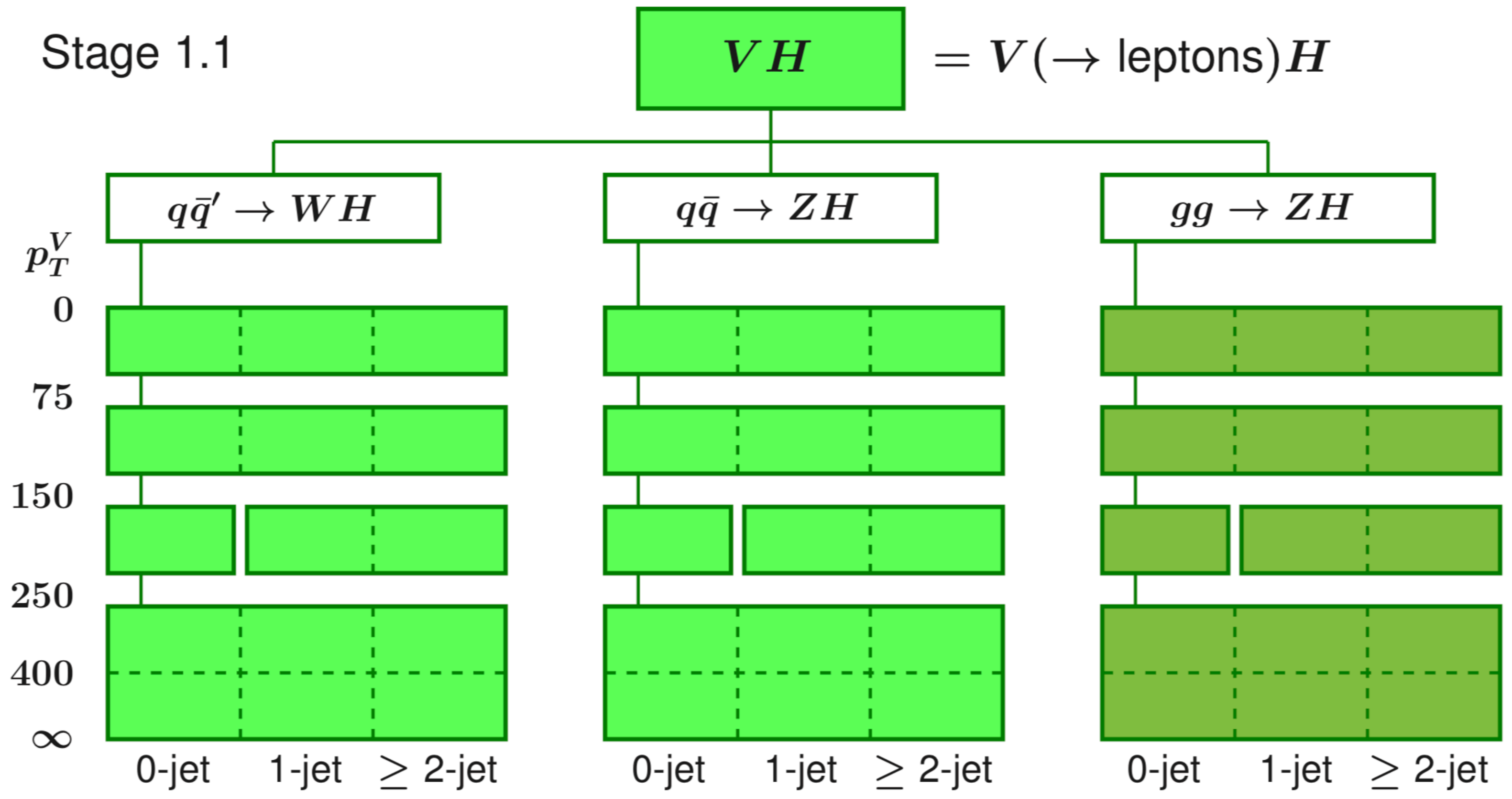
The bigger issue with the STXS approach is that it is not as sensitive as it could be

- Is it exploiting enough differential information?
- If this approach only leads to a small loss in sensitivity maybe it is worth it for convenience
- But is it?

# Benchmarking STXS in fully differential in WH

[JB, S. Dawson, S. Homiller, F. Kling, T. Plehn 1908.06980]

- Simplified Template Cross-Sections (STXS) define observable bins that are supposed to capture as much information on NP as possible
  [N. Berger et al. 1906.02754; HXSWG YR4]




- Let's check! How much information on

$$\tilde{\mathcal{O}}_{HD} = \mathcal{O}_{H\Box} - \frac{\mathcal{O}_{HD}}{4} = (\phi^\dagger\phi)\Box(\phi^\dagger\phi) - \frac{1}{4}(\phi^\dagger D^\mu\phi)^*(\phi^\dagger D_\mu\phi)$$

$$\mathcal{O}_{HW} = \phi^\dagger\phi W^a_{\mu\nu}W^{\mu\nu a}$$

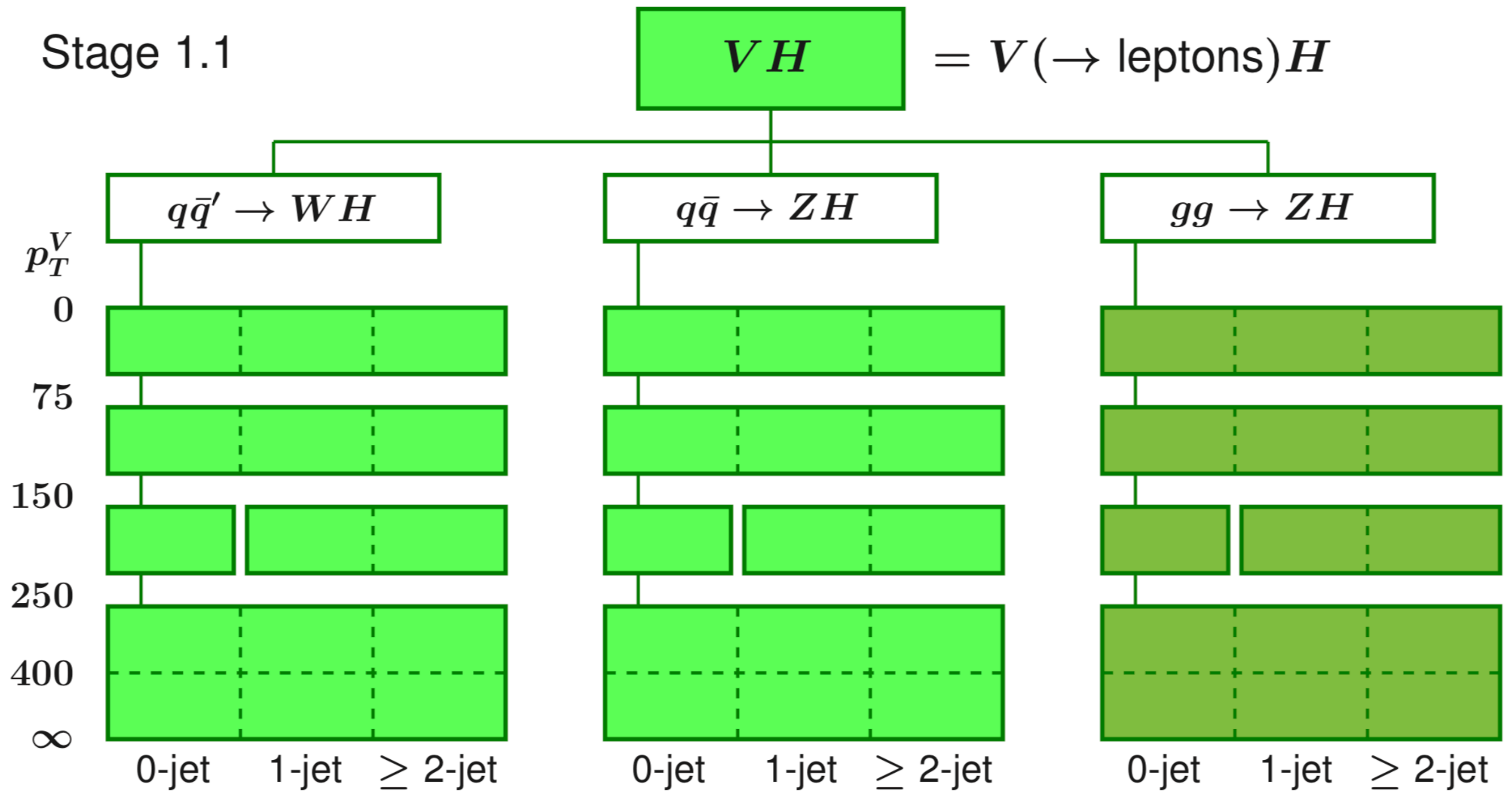
$$\mathcal{O}^{(3)}_{Hq} = (\phi^\dagger i\overleftrightarrow{D}^a_\mu\phi)(\overline{Q}_L\sigma^a\gamma^\mu Q_L)\,,$$

can we extract from $pp \to WH \to \ell\nu\, b\bar{b}$ ?

# Benchmarking STXS in fully differential in WH

[JB, S. Dawson, S. Homiller, F. Kling, T. Plehn 1908.06980]

- Simplified Template Cross-Sections (STXS) define observable bins that are supposed to capture as much information on NP as possible
  [N. Berger et al. 1906.02754; HXSWG YR4]




- Let's check! How much information on

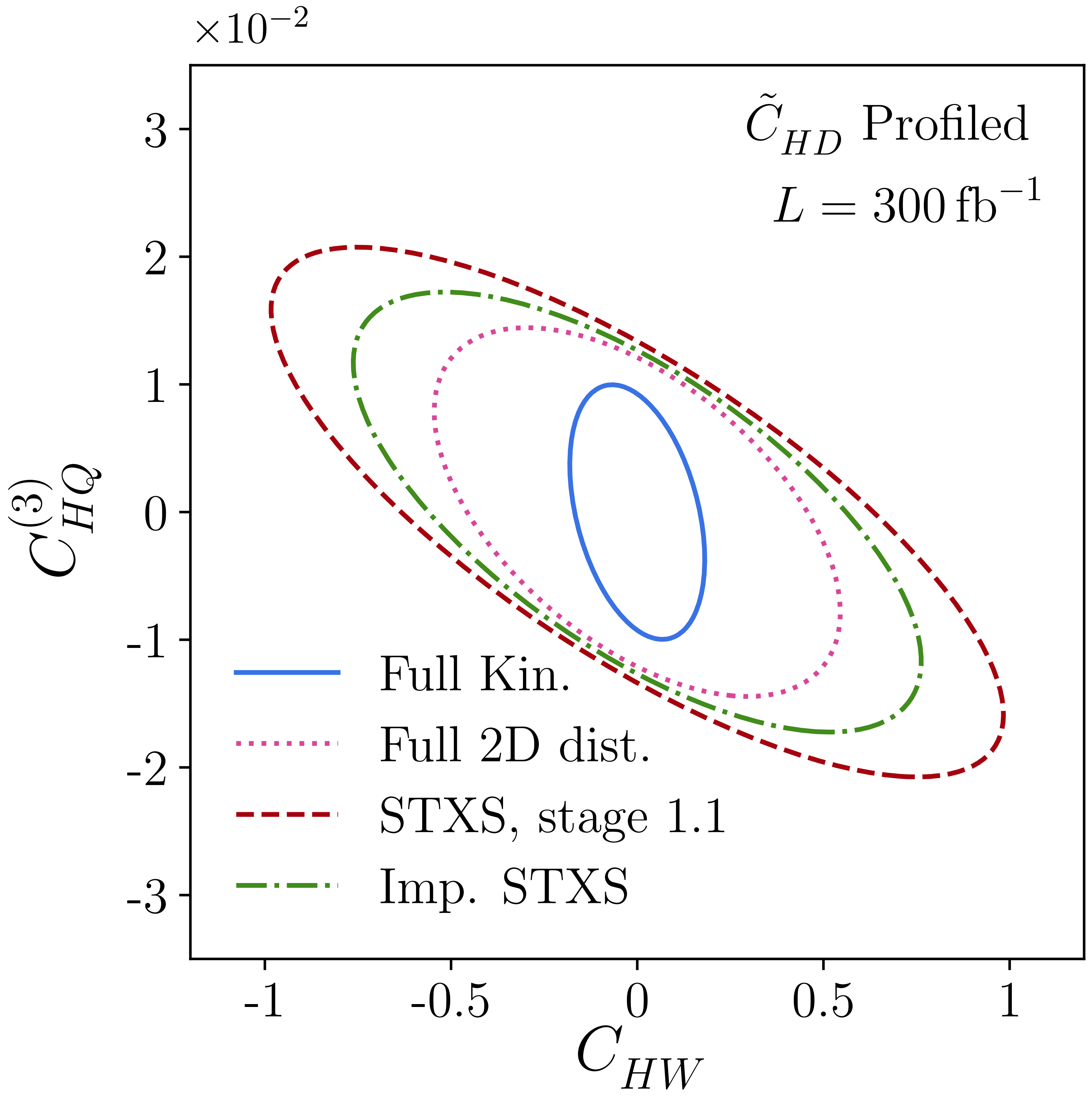
$$\tilde{\mathcal{O}}_{HD} = \mathcal{O}_{H\Box} - \frac{\mathcal{O}_{HD}}{4} = (\phi^\dagger\phi)\Box(\phi^\dagger\phi) - \frac{1}{4}(\phi^\dagger D^\mu\phi)^*(\phi^\dagger D_\mu\phi)$$

$$\mathcal{O}_{HW} = \phi^\dagger\phi W^a_{\mu\nu}W^{\mu\nu a}$$

$$\mathcal{O}^{(3)}_{Hq} = (\phi^\dagger i \overleftrightarrow{D}^a_\mu \phi)(\overline{Q}_L\sigma^a\gamma^\mu Q_L)\ ,$$

can we extract from $pp \to WH \to \ell\nu\, b\bar{b}$ ?

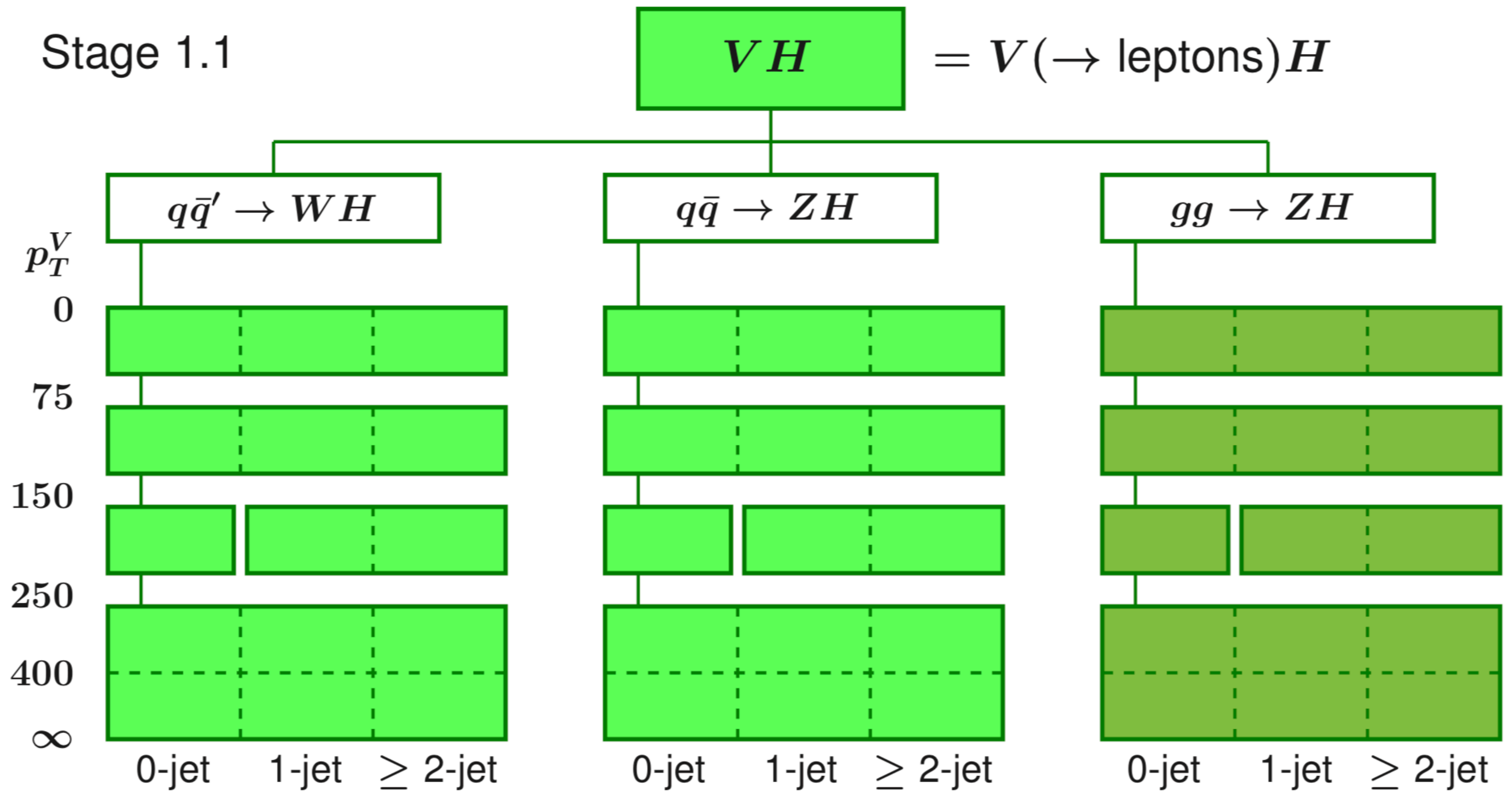
- Results: STXS are indeed sensitive to operators, adding a few more bins improve them, but a multivariate analysis is *much* stronger



# Benchmarking STXS in fully differential in WH

[JB, S. Dawson, S. Homiller, F. Kling, T. Plehn 1908.06980]

- Simplified Template Cross-Sections (STXS) define observable bins that are supposed to capture as much information on NP as possible
  [N. Berger et al. 1906.02754; HXSWG YR4]




- Let's check! How much information on

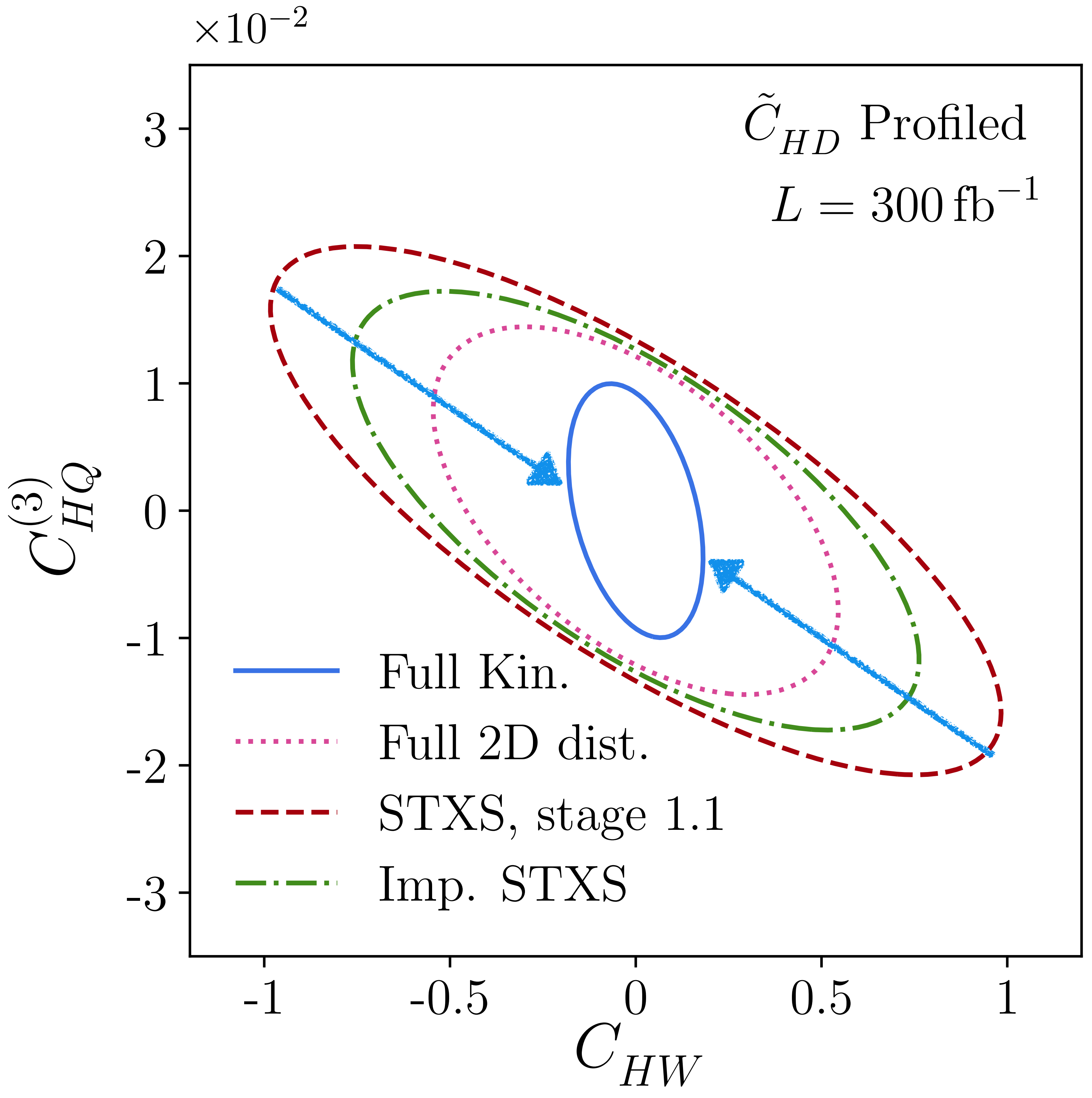
$$\tilde{\mathcal{O}}_{HD} = \mathcal{O}_{H\Box} - \frac{\mathcal{O}_{HD}}{4} = (\phi^\dagger\phi)\Box(\phi^\dagger\phi) - \frac{1}{4}(\phi^\dagger D^\mu\phi)^*(\phi^\dagger D_\mu\phi)$$

$$\mathcal{O}_{HW} = \phi^\dagger\phi W^a_{\mu\nu}W^{\mu\nu a}$$

$$\mathcal{O}^{(3)}_{Hq} = (\phi^\dagger i\overleftrightarrow{D}^a_\mu\phi)(\overline{Q}_L\sigma^a\gamma^\mu Q_L)\ ,$$

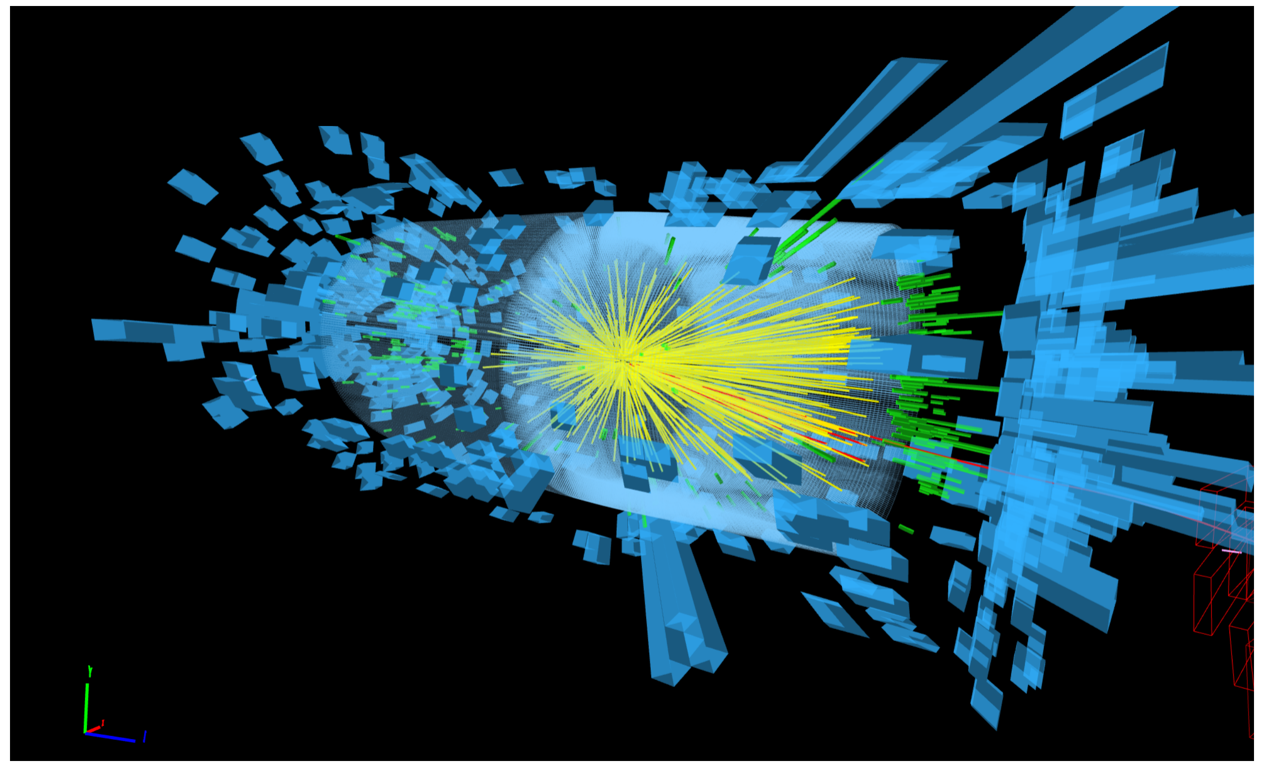
can we extract from $pp \to WH \to \ell\nu\, b\bar{b}$ ?

- Results: STXS are indeed sensitive to operators, adding a few more bins improve them, but a multivariate analysis is *much* stronger



# The likelihood is a key object

Let $\theta$ denote the coefficients of higher dimensional operators in the Lagrangian, $x$ be high-dimensional data associated to an event, and $p(x \mid \theta) = \frac{1}{\sigma(\theta)} \frac{d\sigma}{d\theta}$ be the distribution for the data

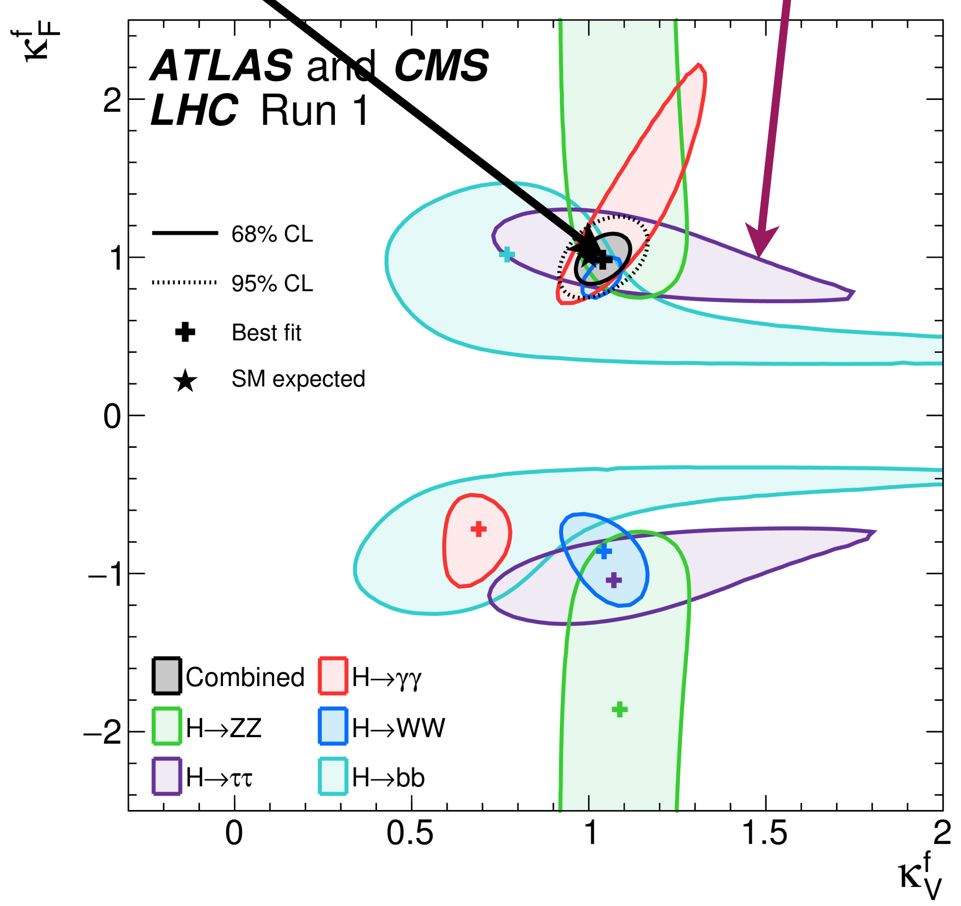


High-dimensional event data $x$

Likelihood function
$p(x|\theta)$

Maximum-likelihood estimator

Confidence limits based on likelihood ratio tests




Constraints on parameters $\theta$

Now for some bad news….

Particle physics processes do not have a tractable likelihood function.

# Modeling particle physics processes

Theory
parameters
$\theta$

Evolution

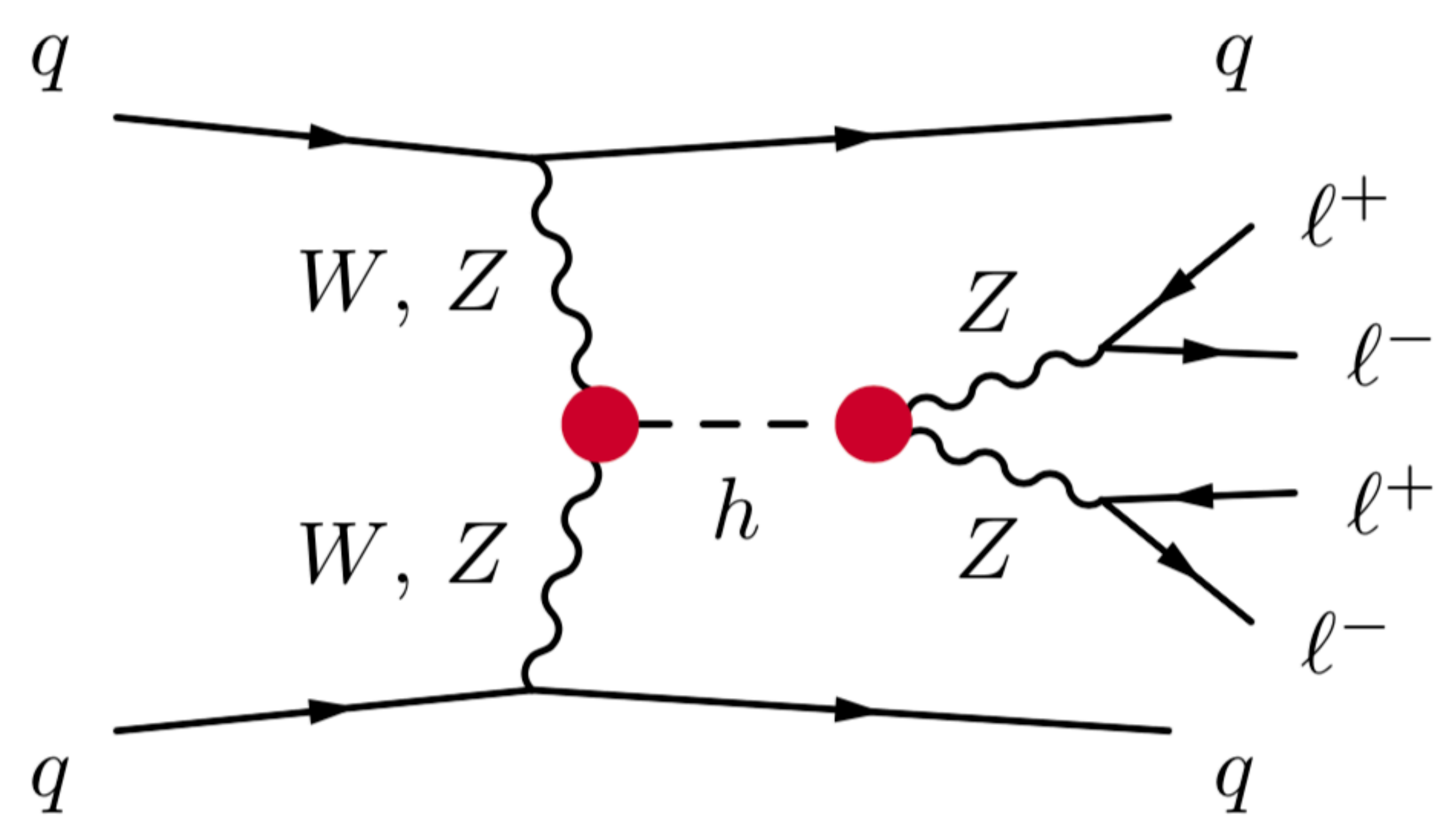
# Modeling particle physics processes

Latent variables

Parton-level momenta

Theory parameters

$z_p \longleftarrow \theta$

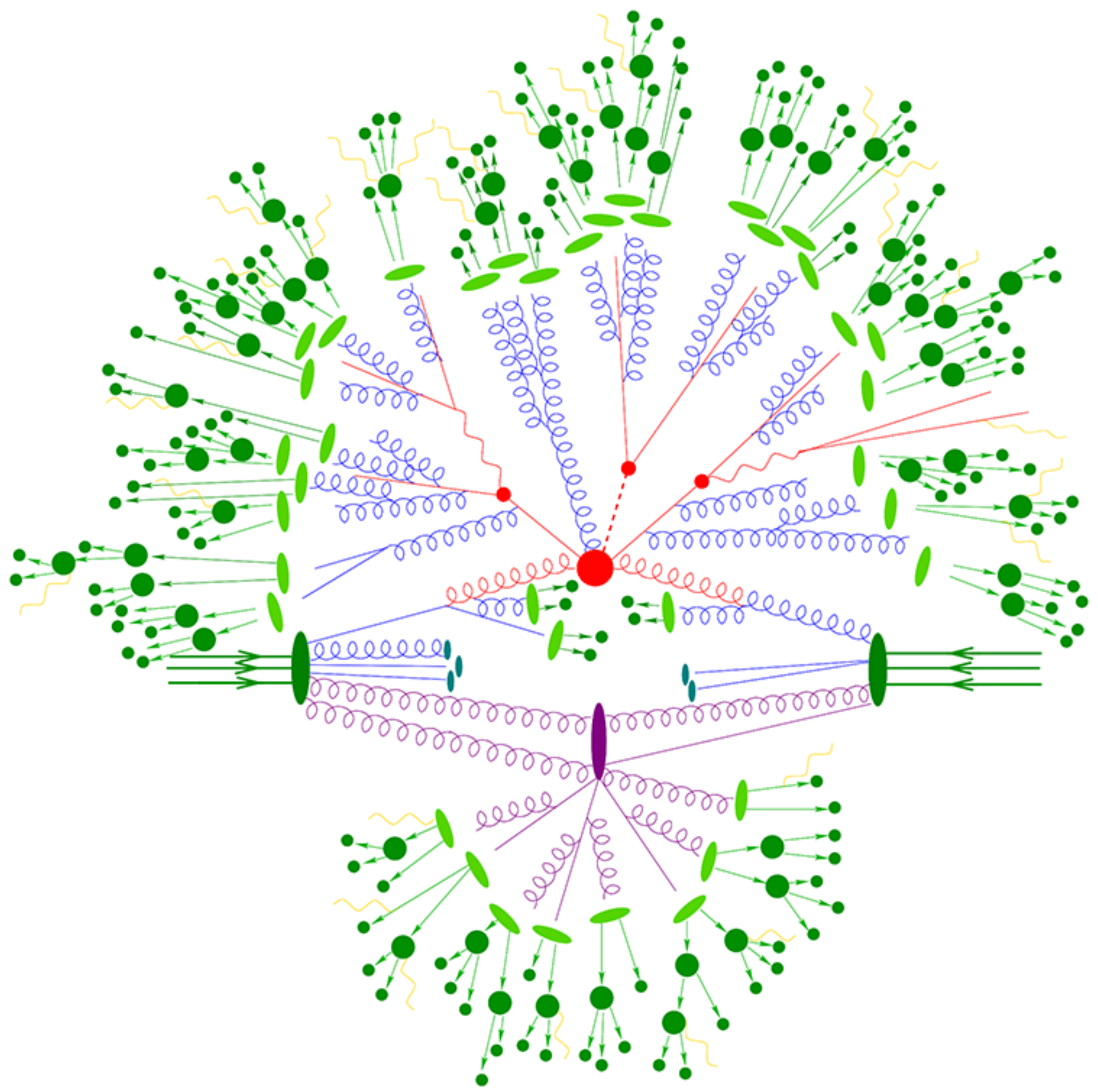



Evolution

# Modeling particle physics processes

Latent variables

| Shower splittings | Parton-level momenta | Theory parameters |
|---|---|---|
| $z_s$ ← | $z_p$ ← | $\theta$ |



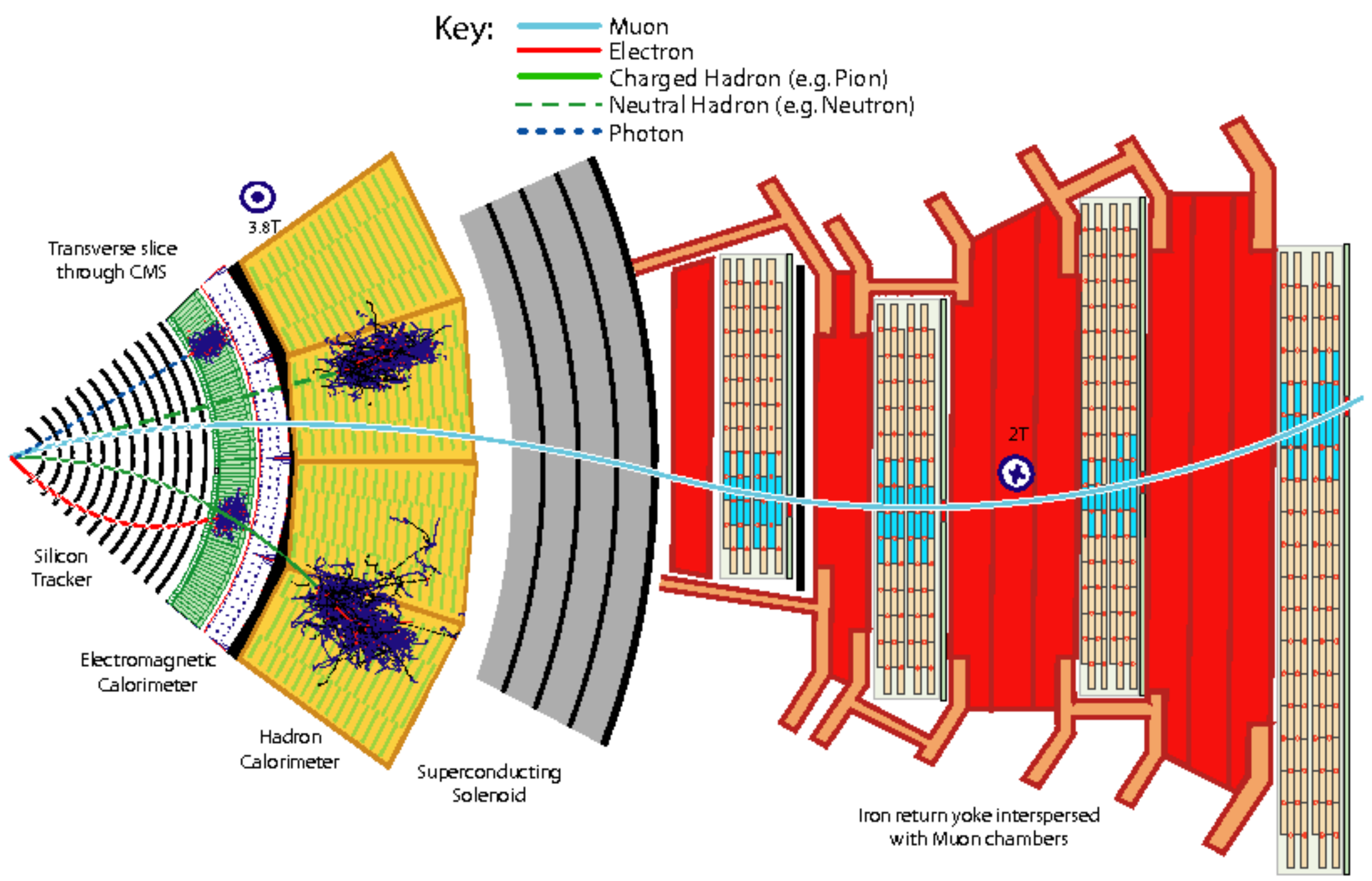
[F. Krauss]

← Evolution

# Modeling particle physics processes

Latent variables

| Detector interactions | | Shower splittings | | Parton-level momenta | | Theory parameters |
|---|---|---|---|---|---|---|
| $z_d$ | $\leftarrow$ | $z_s$ | $\leftarrow$ | $z_p$ | $\leftarrow$ | $\theta$ |



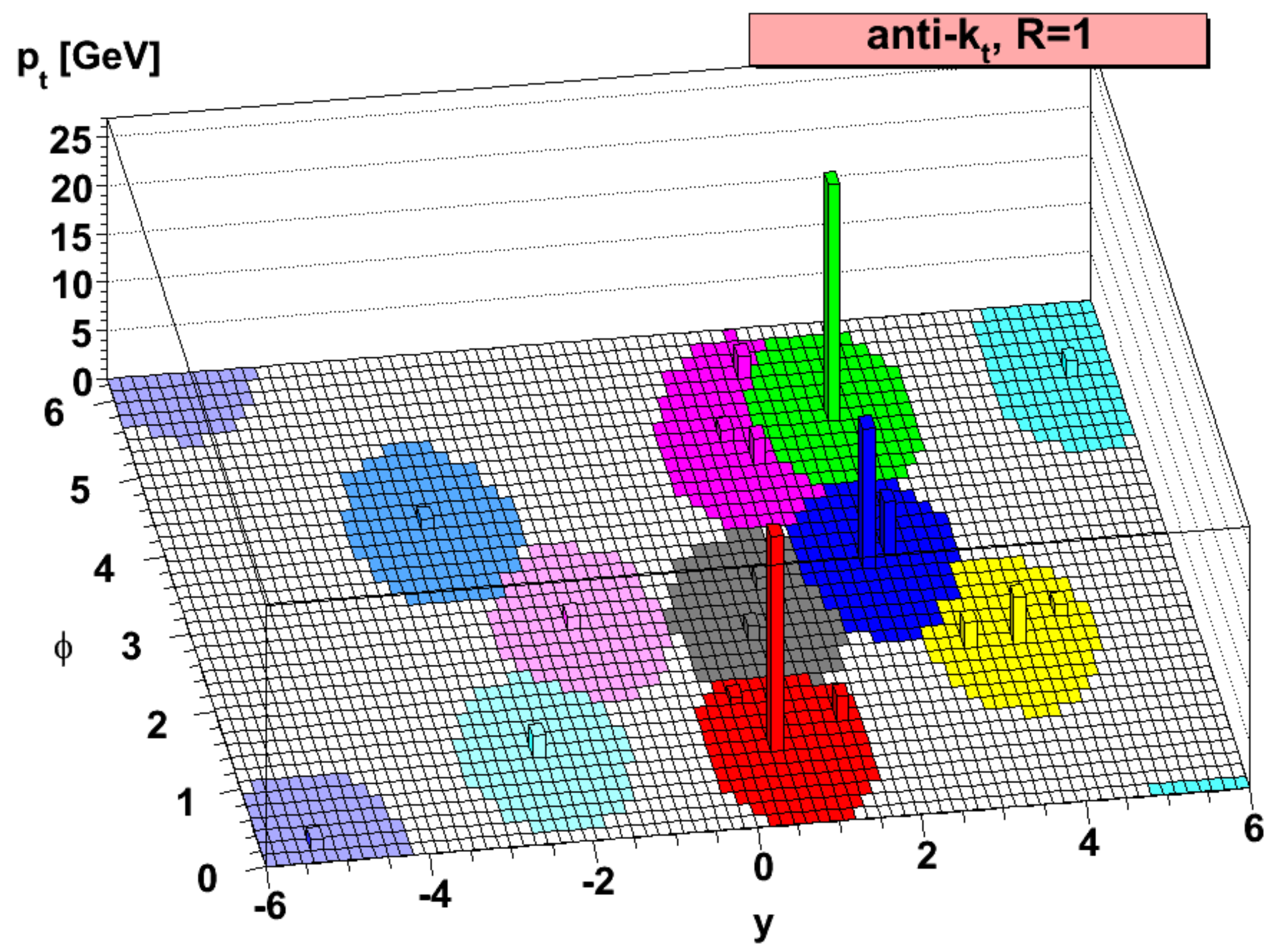

[CMS]

$\longleftarrow$ Evolution

# Modeling particle physics processes

Latent variables

| Observables | | Detector interactions | Shower splittings | Parton-level momenta | Theory parameters |
|---|---|---|---|---|---|
| $x$ | $\longleftarrow$ | $z_d \longleftarrow$ | $z_s \longleftarrow$ | $z_p \longleftarrow$ | $\theta$ |




[M. Cacciari, G. Salam, G. Soyez 0802.1189]

Evolution

# Modeling particle physics processes

Latent variables

| Observables | | Detector interactions | Shower splittings | Parton-level momenta | Theory parameters |
|---|---|---|---|---|---|
| $x$ | ← | $z_d$ ← | $z_s$ ← | $z_p$ ← | $\theta$ |

Sample from $p(x|z_d)$ $p(z_d|z_s)$ $p(z_s|z_p)$ $p(z_p|\theta)$

GEANT4 A SIMULATION TOOLKIT

MADGRAPH5_aMC@NLO

Prediction (simulation)

# Modeling particle physics processes

Latent variables

| Observables | | Detector interactions | | Shower splittings | | Parton-level momenta | | Theory parameters |
|---|---|---|---|---|---|---|---|---|
| $x$ | $\longleftarrow$ | $z_d$ | $\longleftarrow$ | $z_s$ | $\longleftarrow$ | $z_p$ | $\longleftarrow$ | $\theta$ |

$$p(x|\theta) = \int \mathrm{d}z_d \int \mathrm{d}z_s \int \mathrm{d}z_p \quad p(x|z_d) \quad p(z_d|z_s) \quad p(z_s|z_p) \quad p(z_p|\theta)$$

Inference $\longrightarrow$

# Modeling particle physics processes

Latent variables

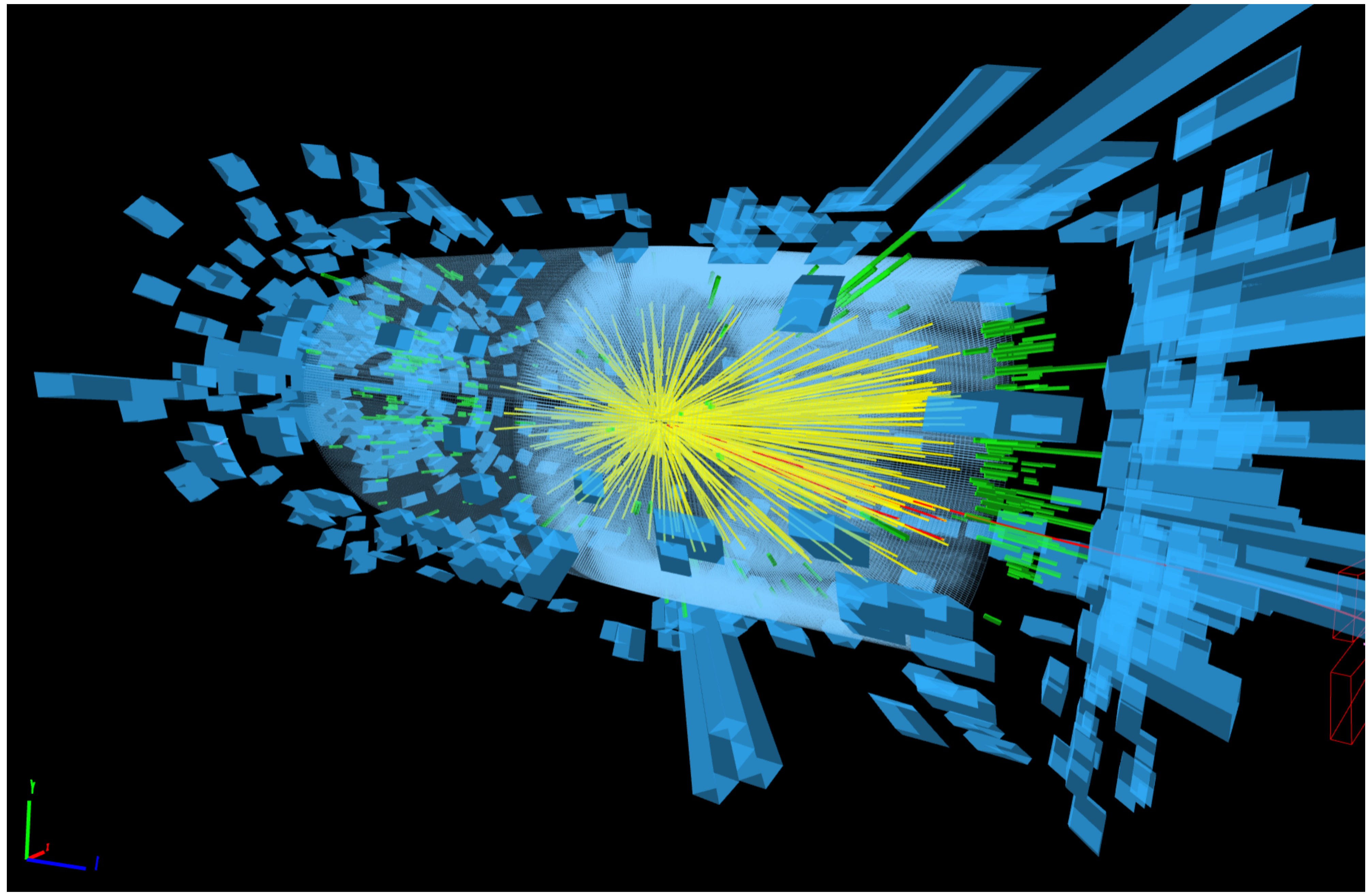
| Observables | | Detector interactions | | Shower splittings | | Parton-level momenta | | Theory parameters |
|---|---|---|---|---|---|---|---|---|
| $x$ | $\leftarrow$ | $z_d$ | $\leftarrow$ | $z_s$ | $\leftarrow$ | $z_p$ | $\leftarrow$ | $\theta$ |

$$p(x|\theta) = \boxed{\int \mathrm{d}z_d \int \mathrm{d}z_s \int \mathrm{d}z_p} \; p(x|z_d) \quad p(z_d|z_s) \quad p(z_s|z_p) \quad p(z_p|\theta)$$

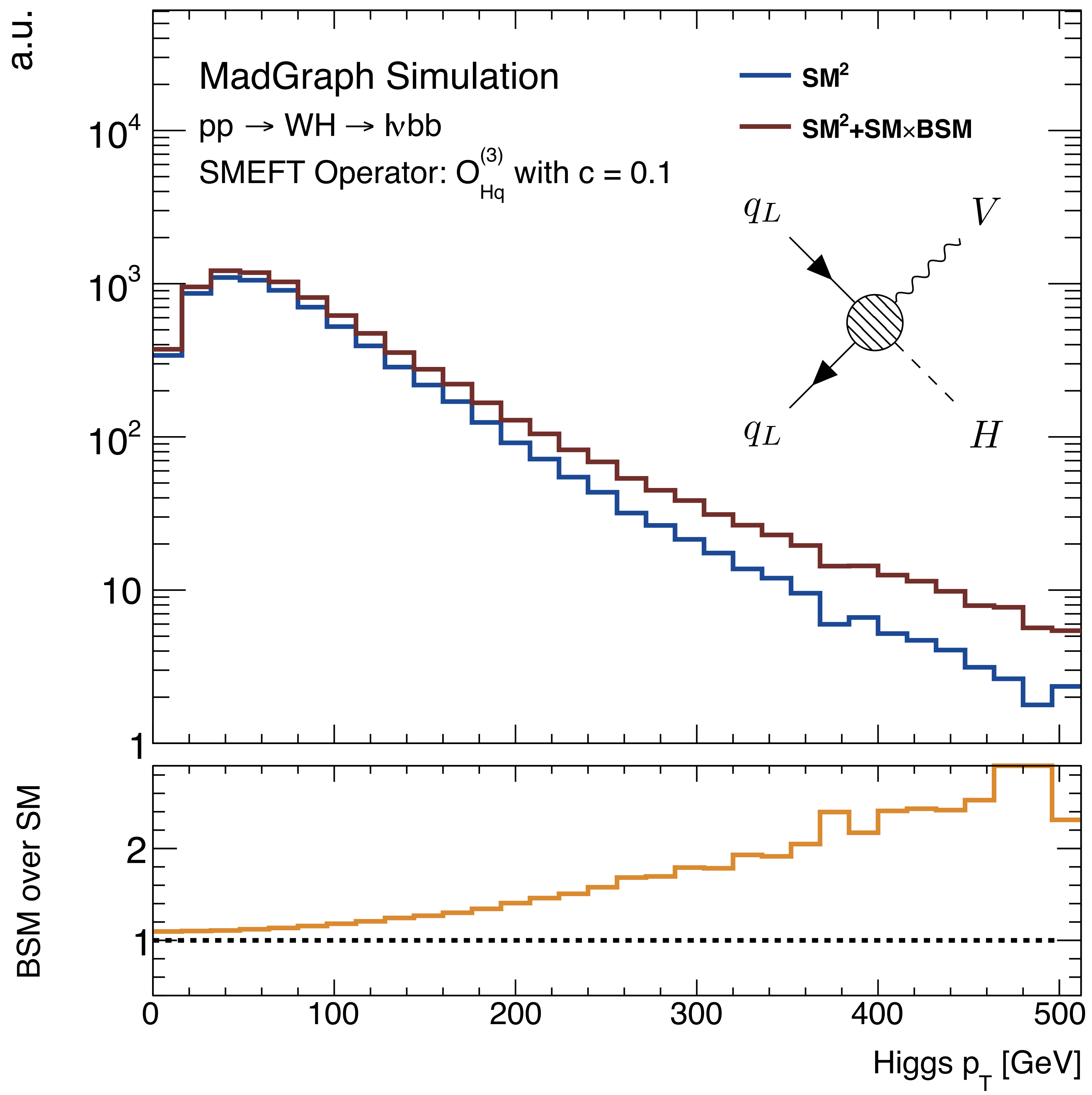
It's infeasible to calculate the integral over this enormous space!

$\longrightarrow$ Inference

# What we usually do: number counting or singly differential



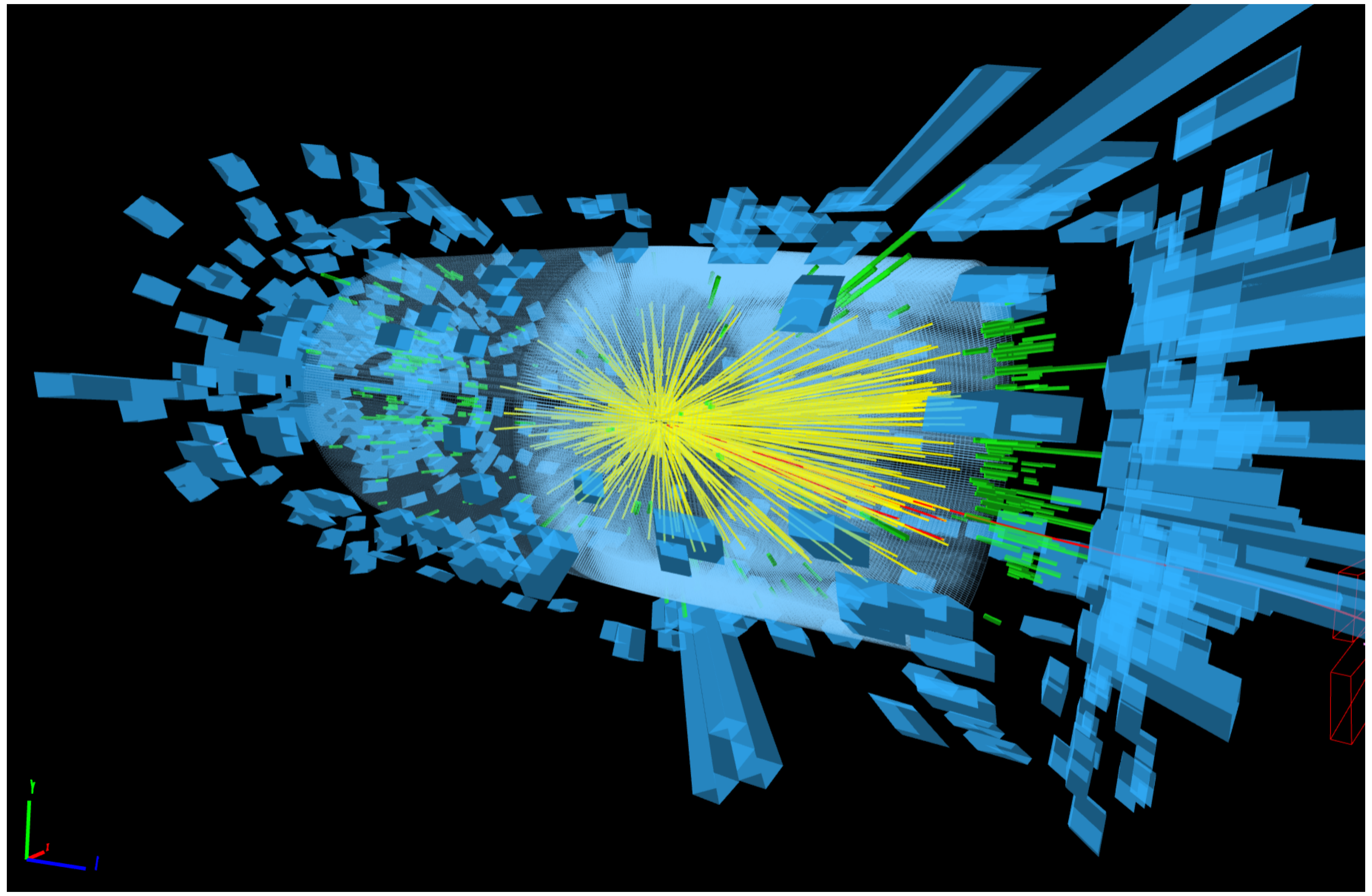
SMEFT: $O_{Hq}^{(3)}$ = 0.1



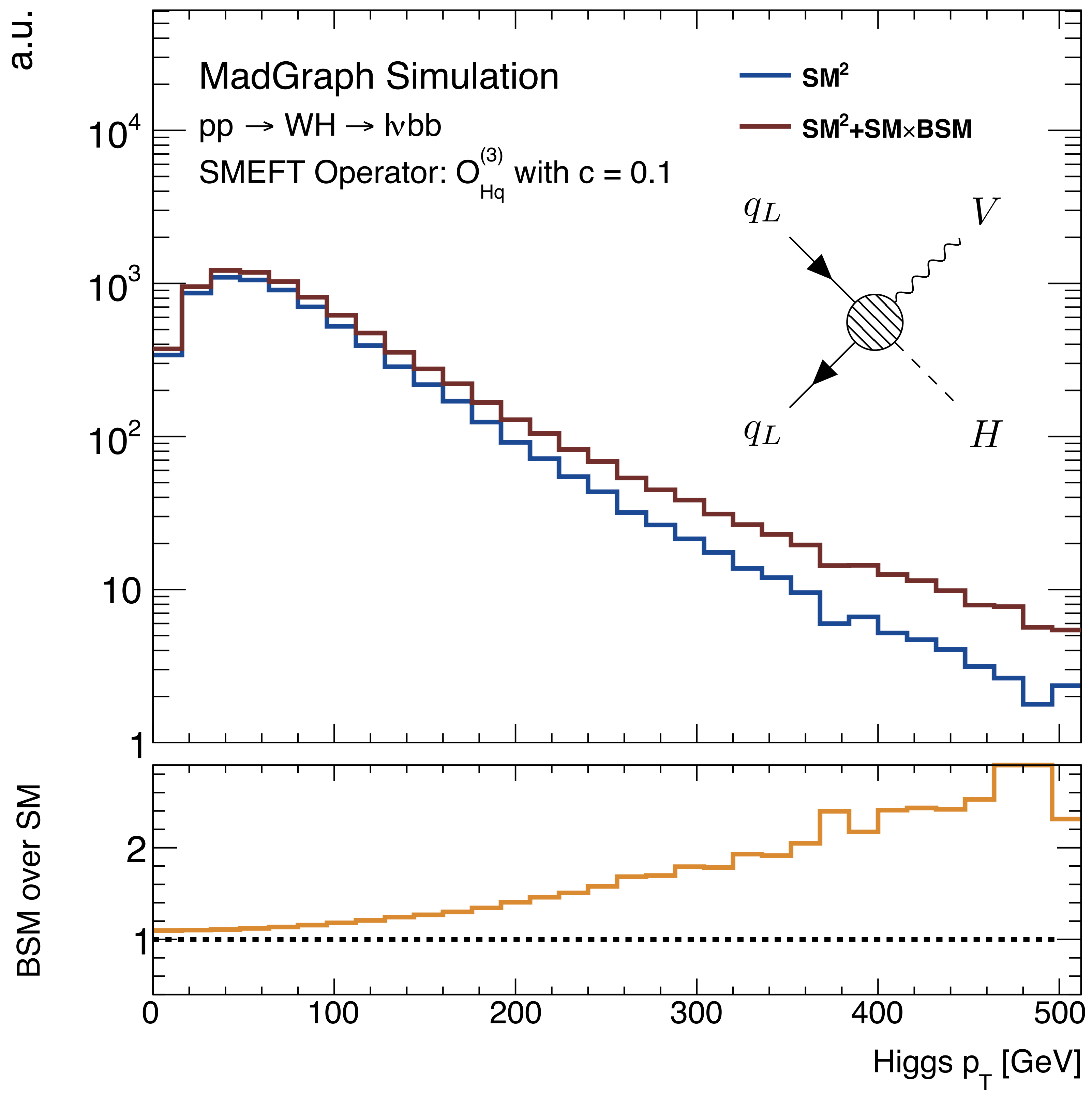

High-dimensional event data $x$

$p(x|\theta)$ cannot be calculated

# What we usually do: number counting or singly differential



SMEFT: $O_{Hq}^{(3)} = 0.1$




High-dimensional event data $x$

$p(x|\theta)$ cannot be calculated

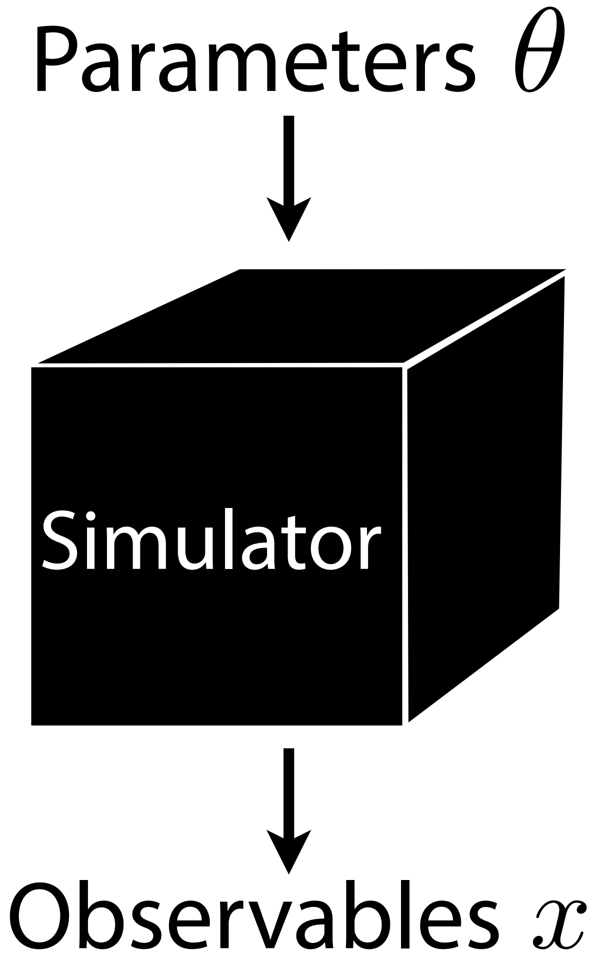
One or two summary statistics $x'$

$p(x'|\theta)$ can be estimated with histograms

n.b. "summary statistic" = a sensitive observable

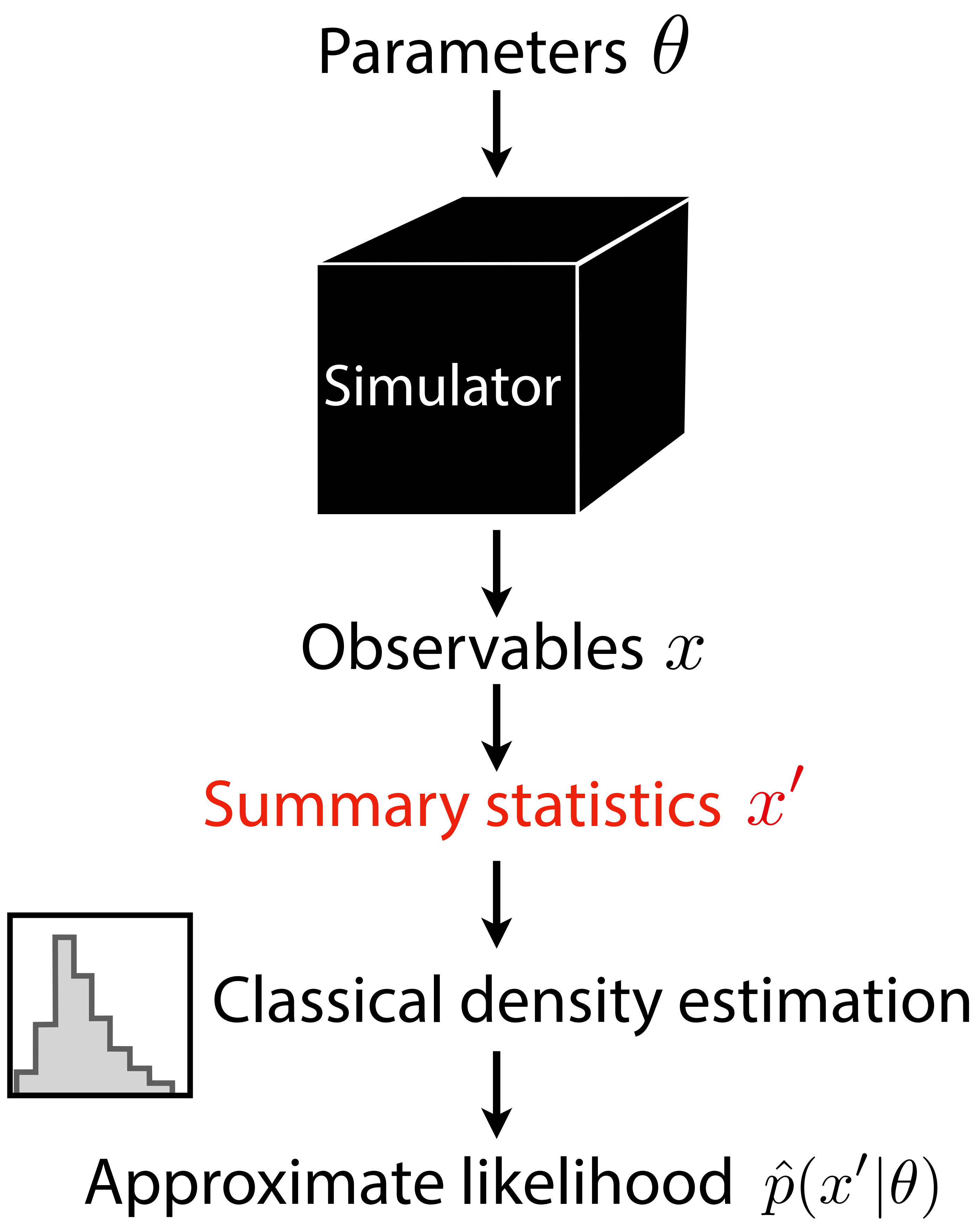
# Inference by estimating the likelihood

[e.g. P. Diggle, R. Gratton 1984]



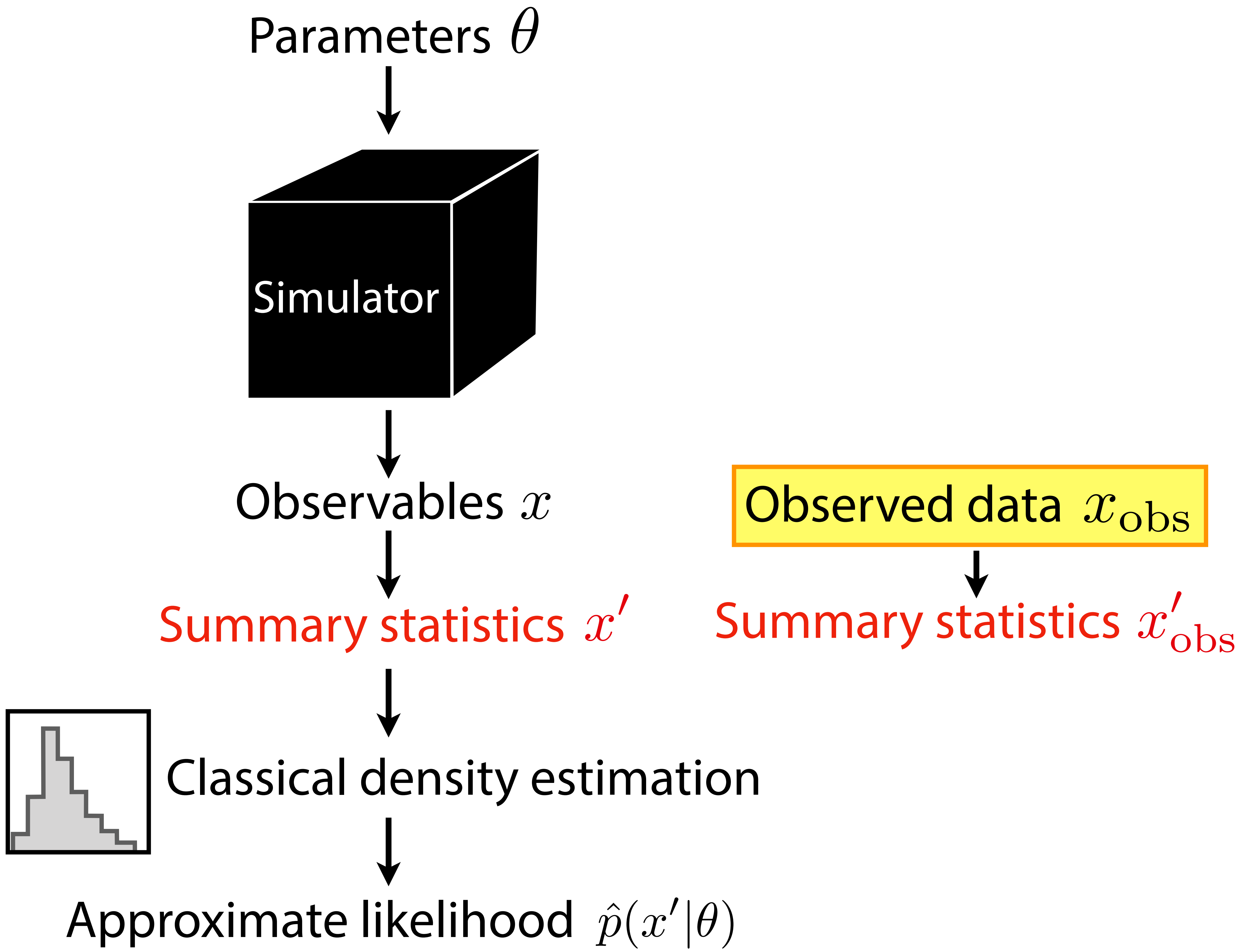
# Inference by estimating the likelihood

[e.g. P. Diggle, R. Gratton 1984]



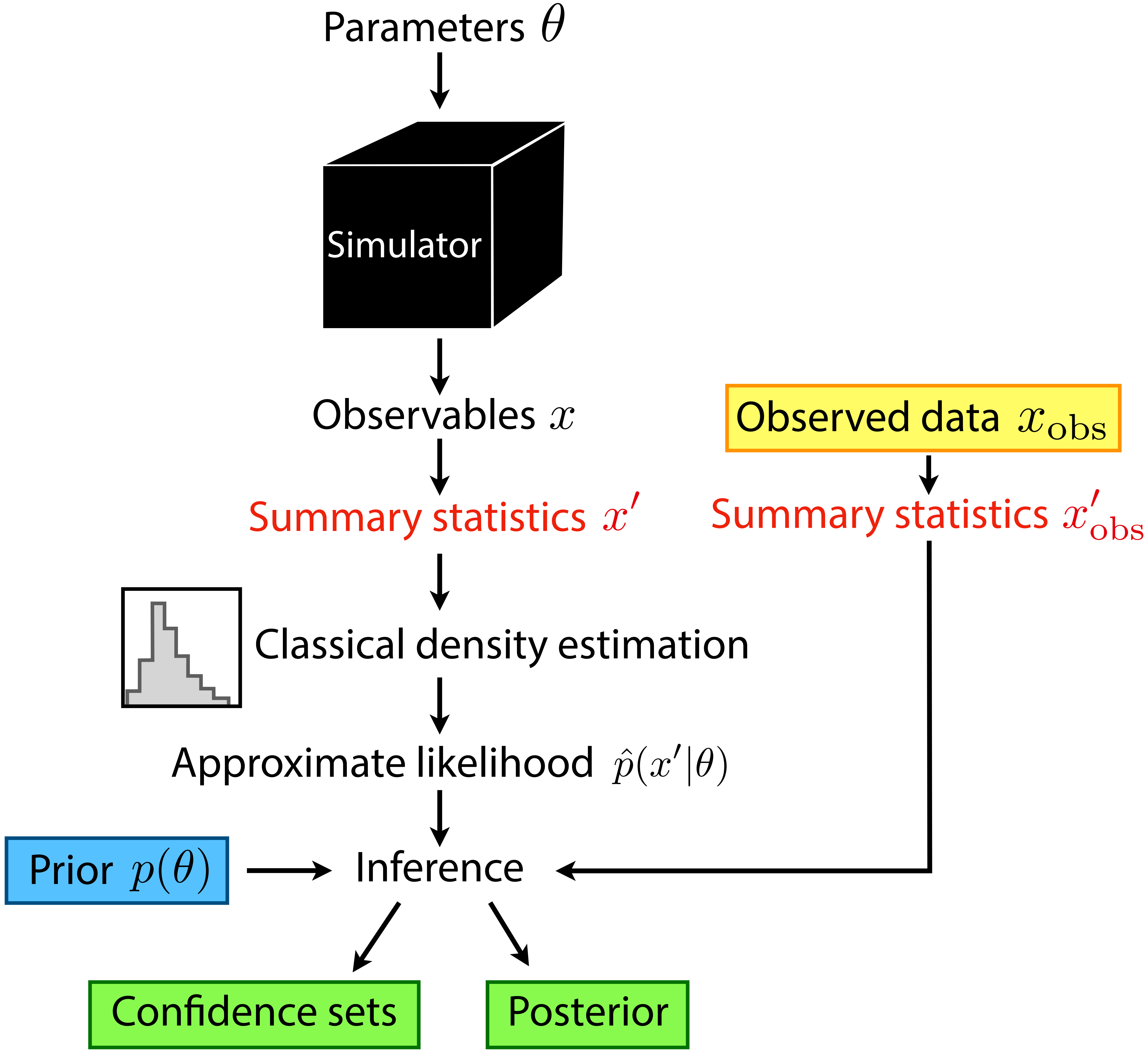
# Inference by estimating the likelihood

[e.g. P. Diggle, R. Gratton 1984]



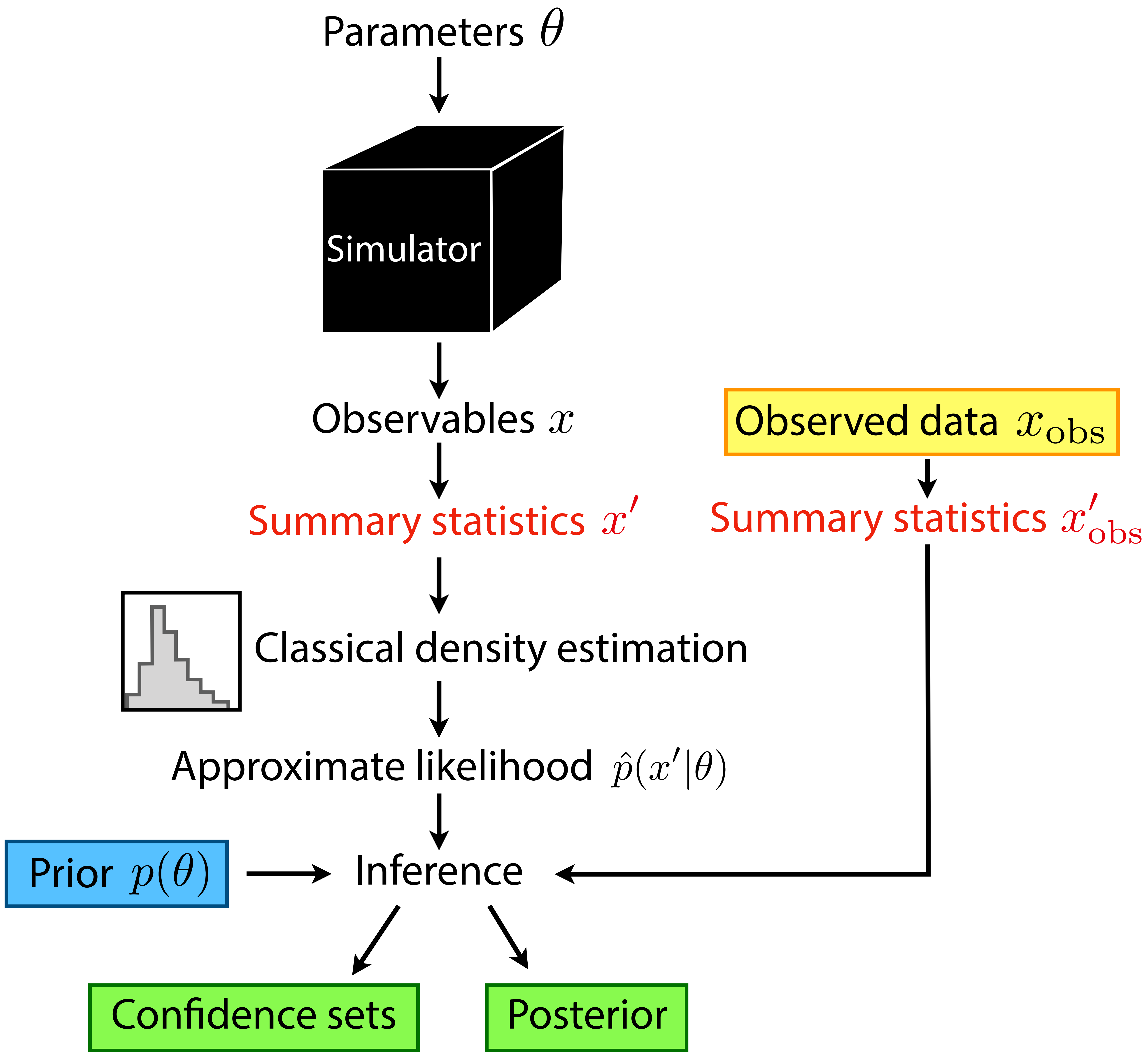
# Inference by estimating the likelihood

[e.g. P. Diggle, R. Gratton 1984]



# Inference by estimating the likelihood

[e.g. P. Diggle, R. Gratton 1984]




- Compression to summary statistics loses information & reduces quality of inference
- Curse of dimensionality: does not scale to more than a few summary statistics
- Related alternative: Approximate Bayesian Computation (ABC) [D. Rubin 1984]

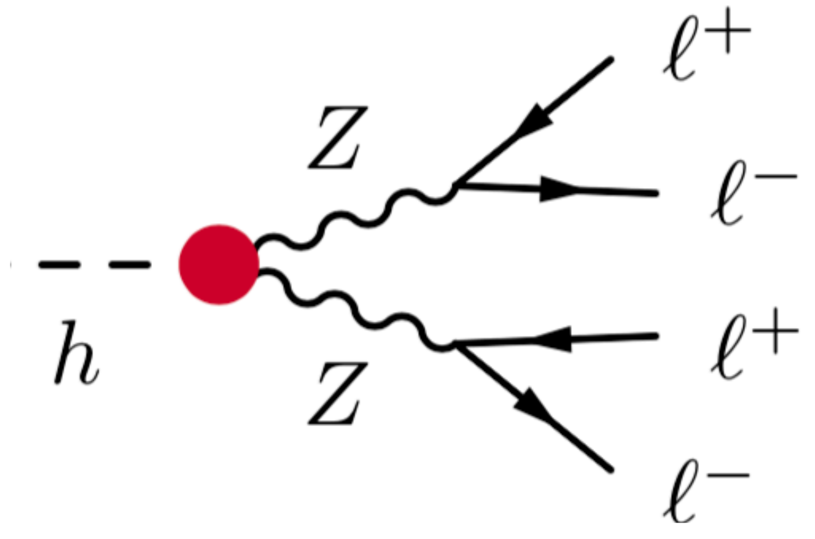
# Summary statistics for LHC measurements?

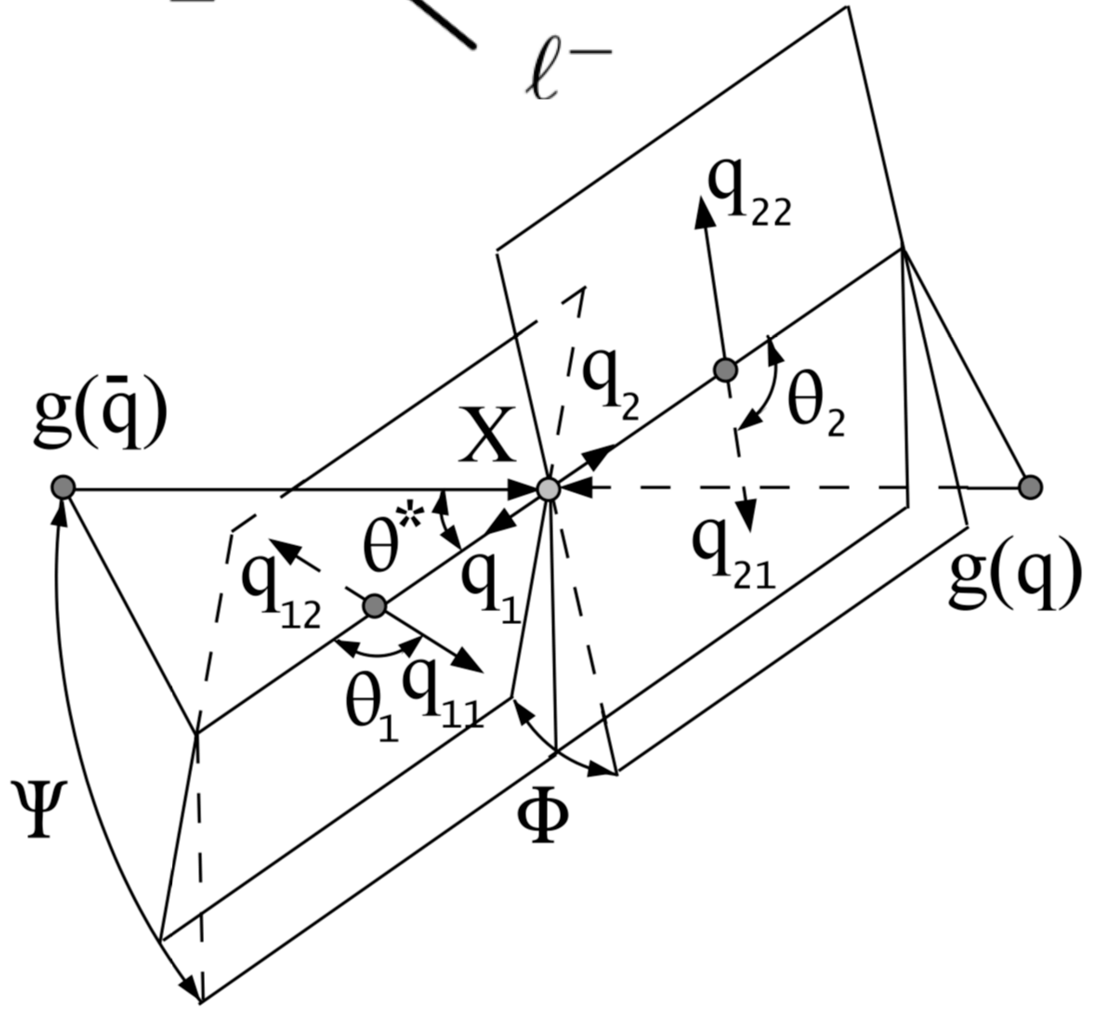
- In many LHC problems (eg. EFTs) there is no single good summary statistic: compressing to any $x'$ loses information!
  [JB, K. Cranmer, F. Kling, T. Plehn 1612.05261; JB, F. Kling, T. Plehn, T. Tait 1712.02350]

- Ideally: analyze all trustworthy high-level features (reconstructed four-momenta…), or some form of low-level features, including correlations
  ("fully differential cross section")

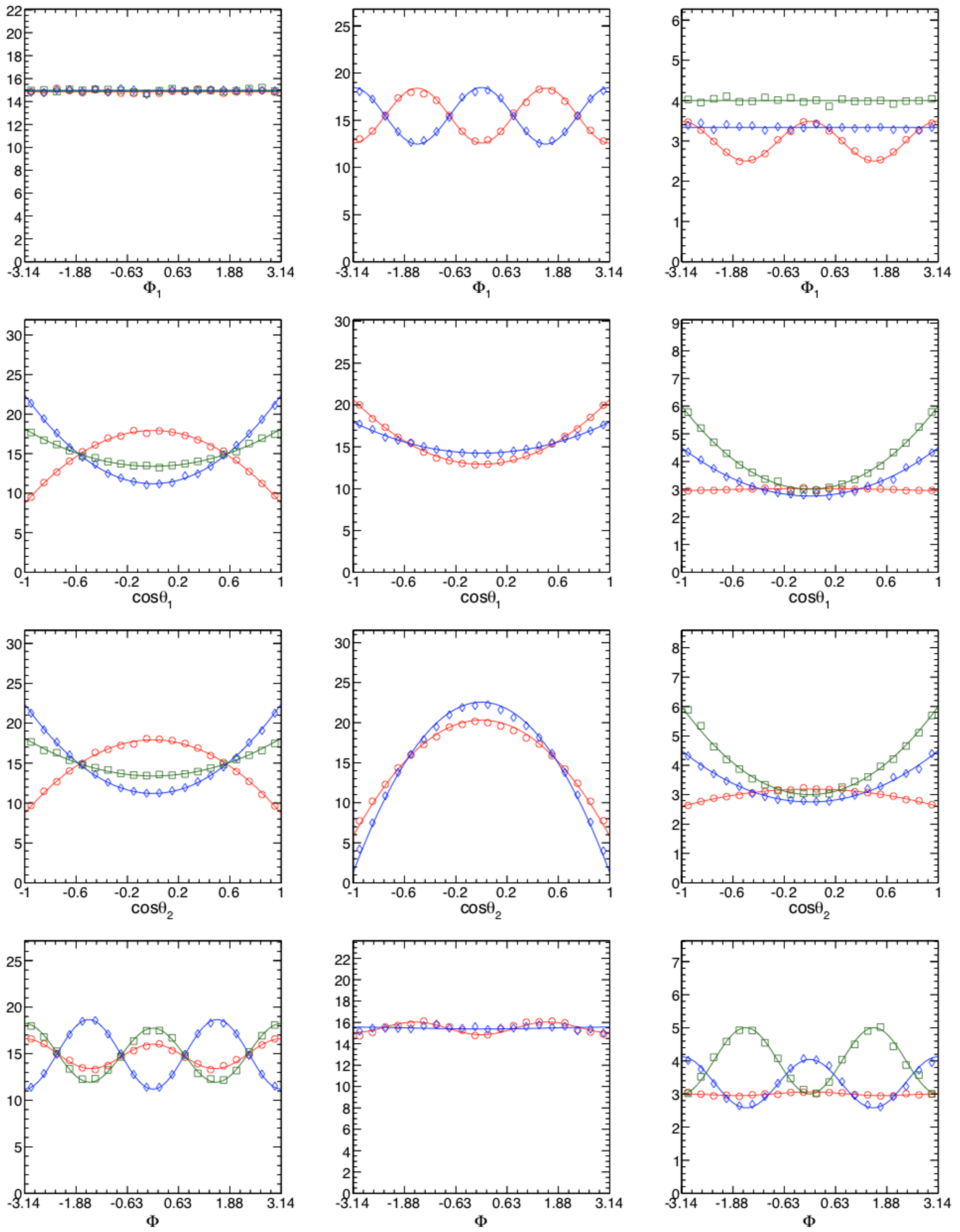
# Summary statistics for LHC measurements?

- In many LHC problems (eg. EFTs) there is no single good summary statistic: compressing to any $x'$ loses information!
  [JB, K. Cranmer, F. Kling, T. Plehn 1612.05261; JB, F. Kling, T. Plehn, T. Tait 1712.02350]




- Ideally: analyze all trustworthy high-level features (reconstructed four-momenta…), or some form of low-level features, including correlations
  ("fully differential cross section")

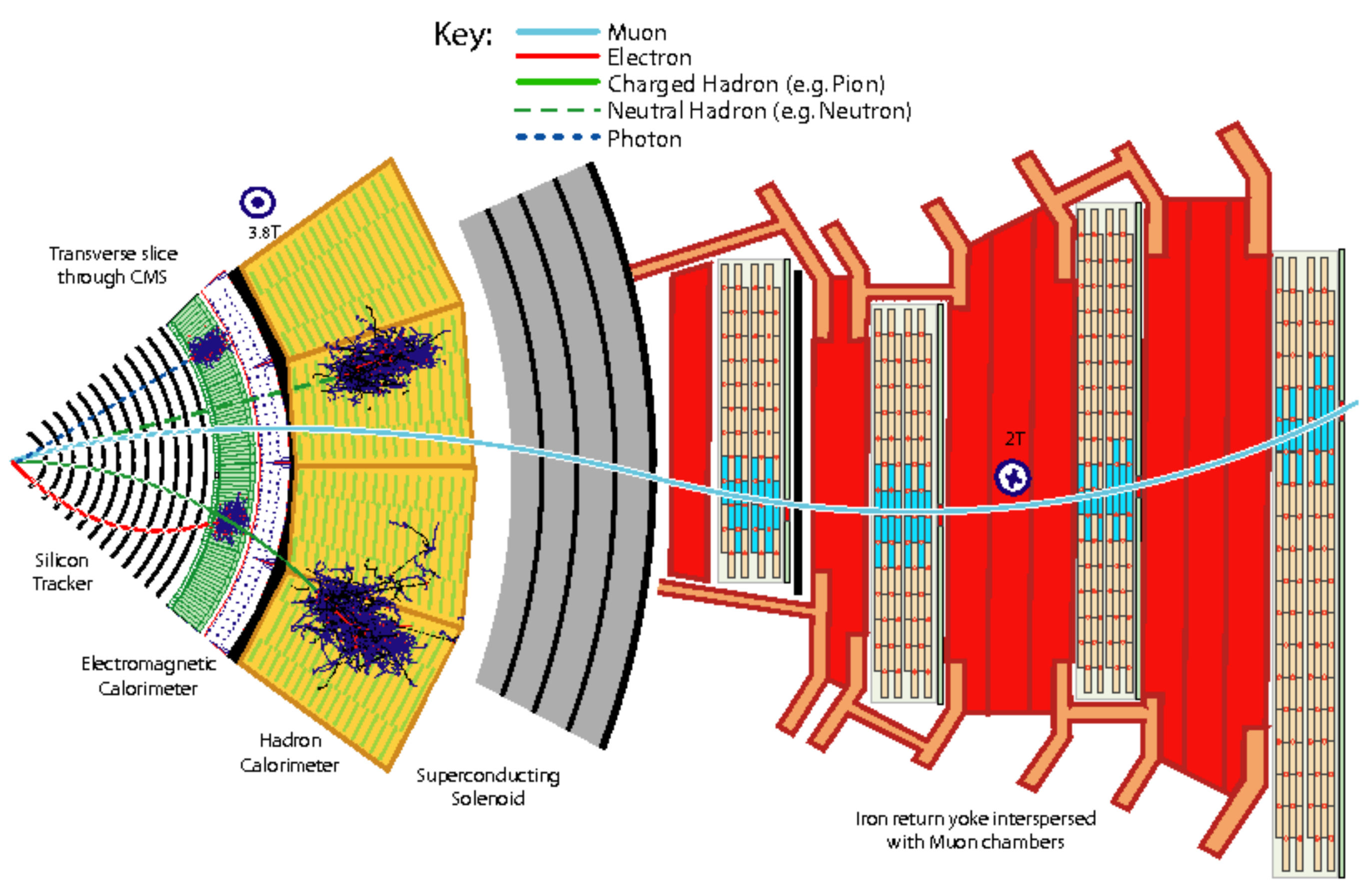



[Bolognesi et al. 1208.4018]

# Solve it by approximating the integral

- Problem: high-dimensional integral over shower / detector trajectories

$$p(x|\theta) = \int \mathrm{d}z_d \int \mathrm{d}z_s \int \mathrm{d}z_p \; p(x|z_d)\, p(z_d|z_s)\, p(z_s|z_p)\, p(z_p|\theta)$$

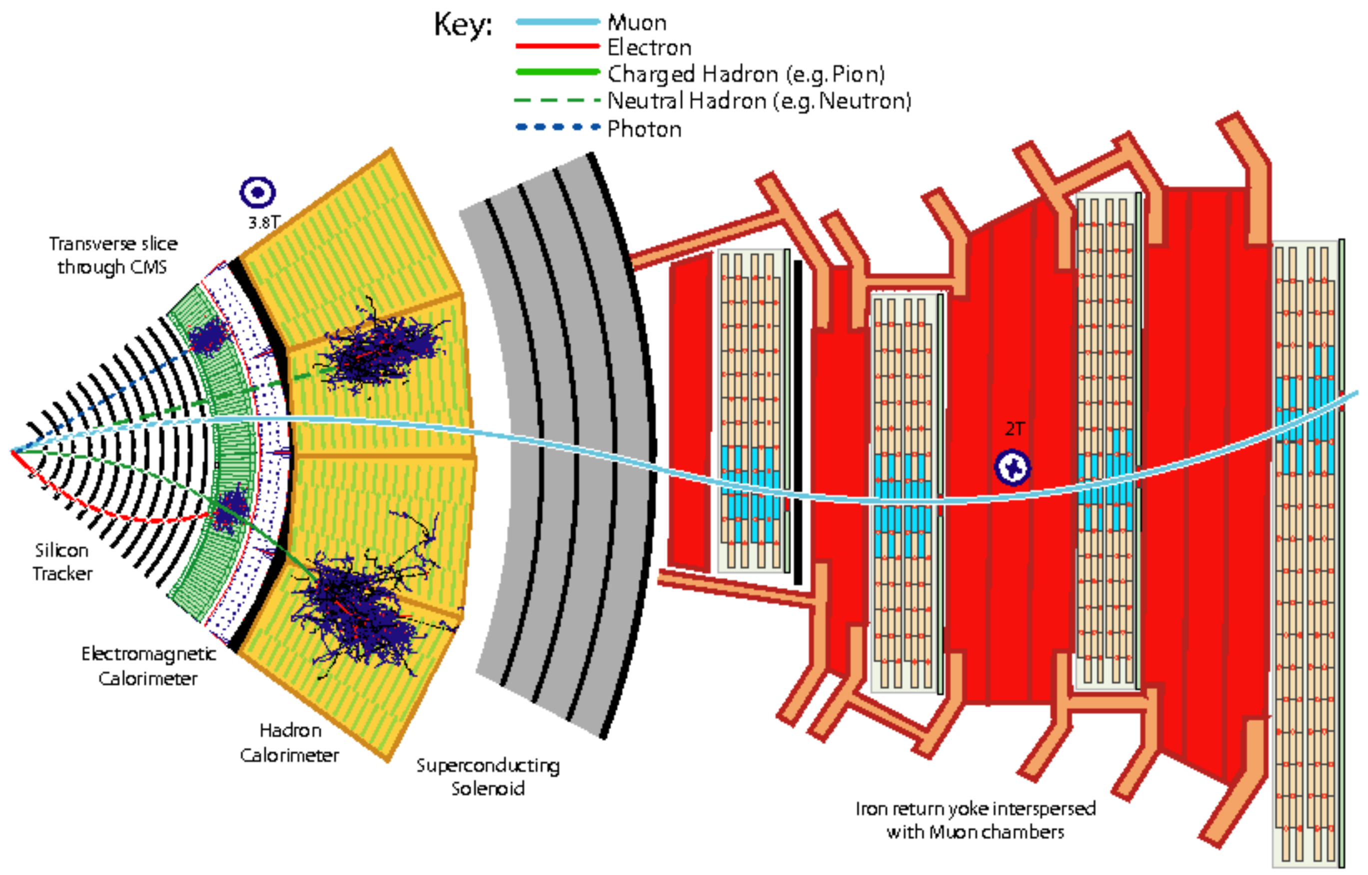


[CMS]

# Solve it by approximating the integral

- Problem: high-dimensional integral over shower / detector trajectories

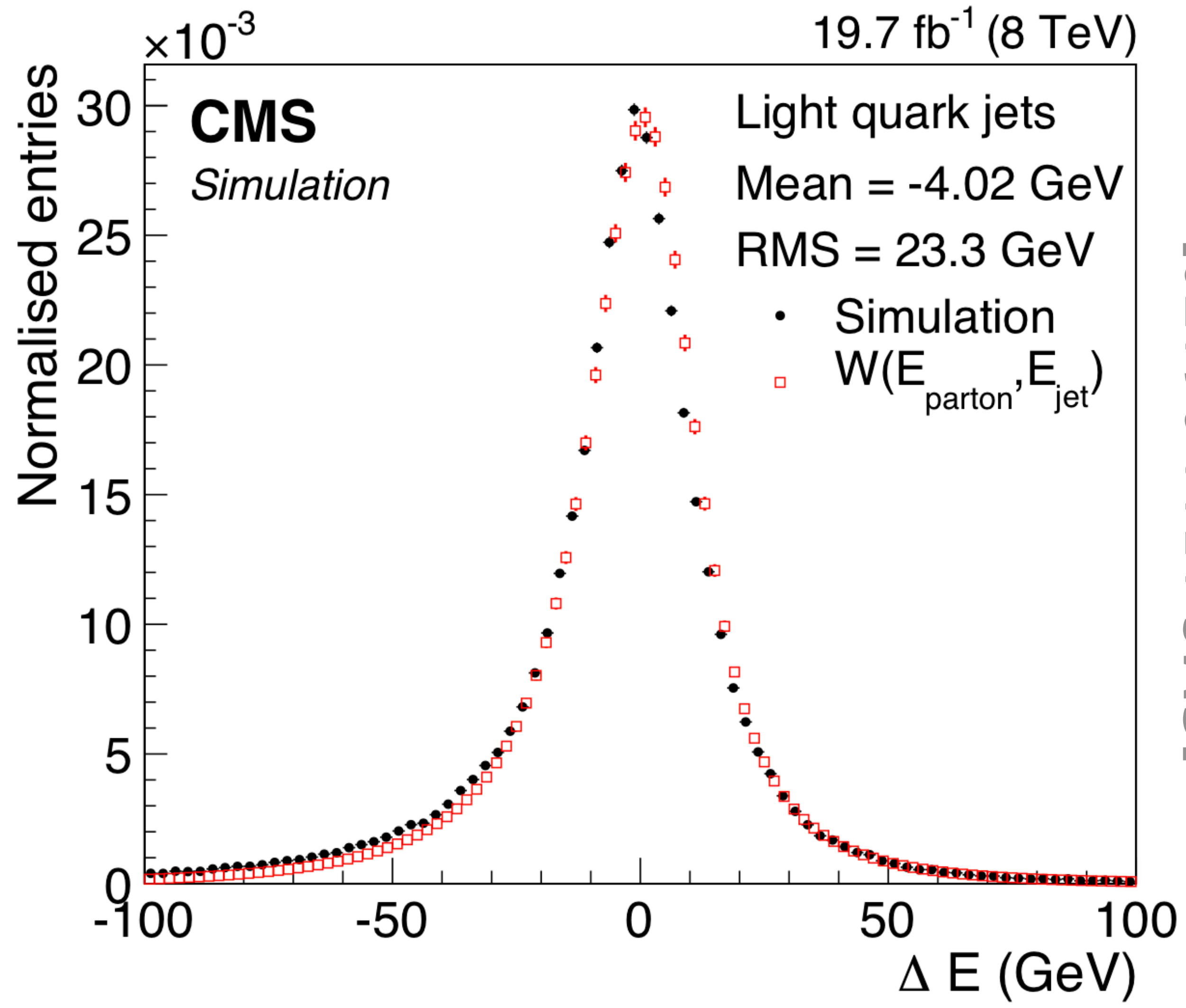
$$p(x|\theta) = \int \mathrm{d}z_d \int \mathrm{d}z_s \int \mathrm{d}z_p \; p(x|z_d)\, p(z_d|z_s)\, p(z_s|z_p)\, p(z_p|\theta)$$




[CMS]

- Matrix Element Method (and similarly Optimal Observables): [K. Kondo 1988]
  - approximate shower + detector effects into transfer function $\hat{p}(x|z_p)$
  - explicitly calculate remaining integral

$$\hat{p}(x|\theta) = \int \mathrm{d}z_p \; \hat{p}(x|z_p)\, p(z_p|\theta)$$

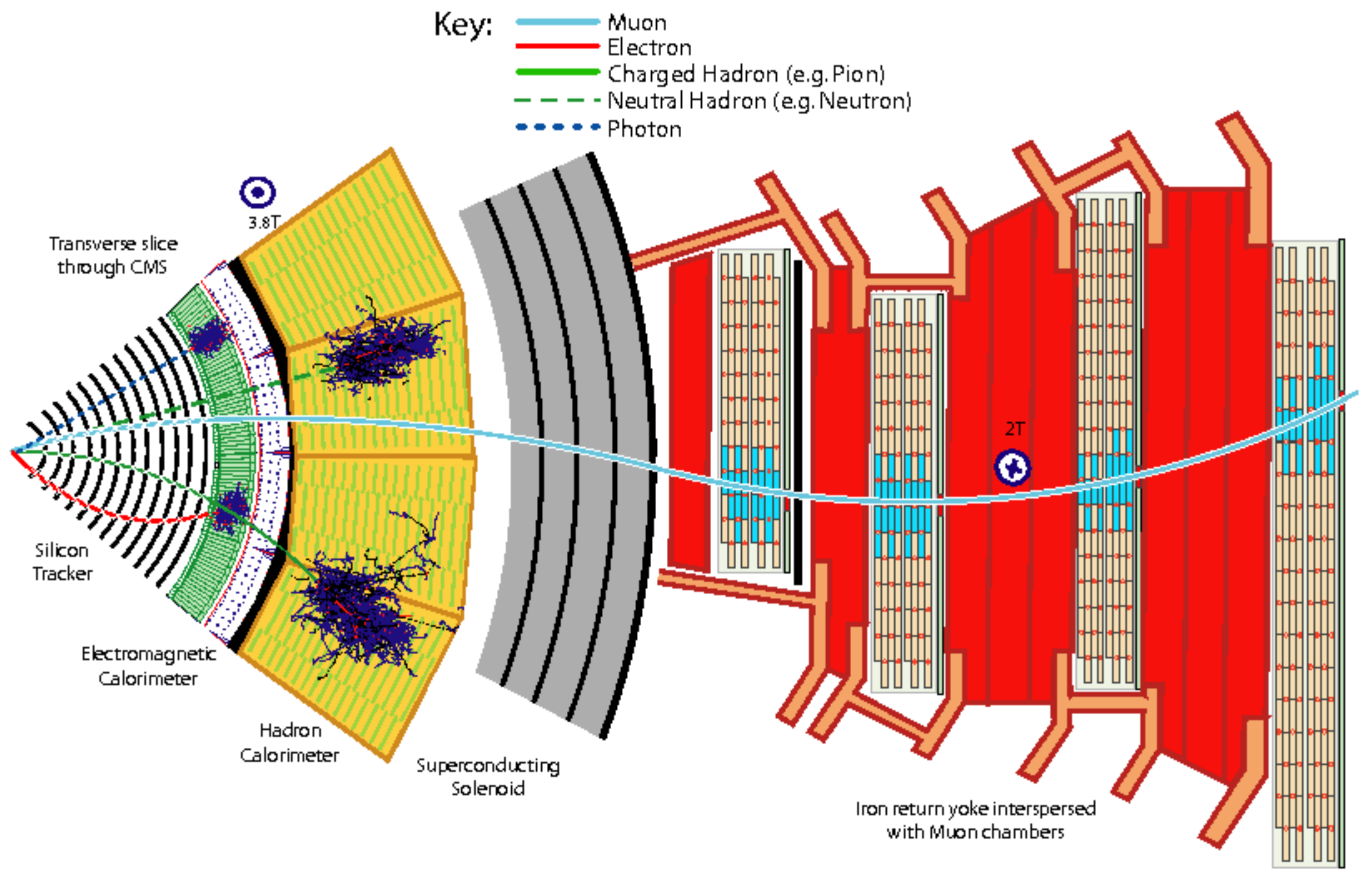



[CMS 1511.06170]

# Solve it by approximating the integral

- Problem: high-dimensional integral over shower / detector trajectories

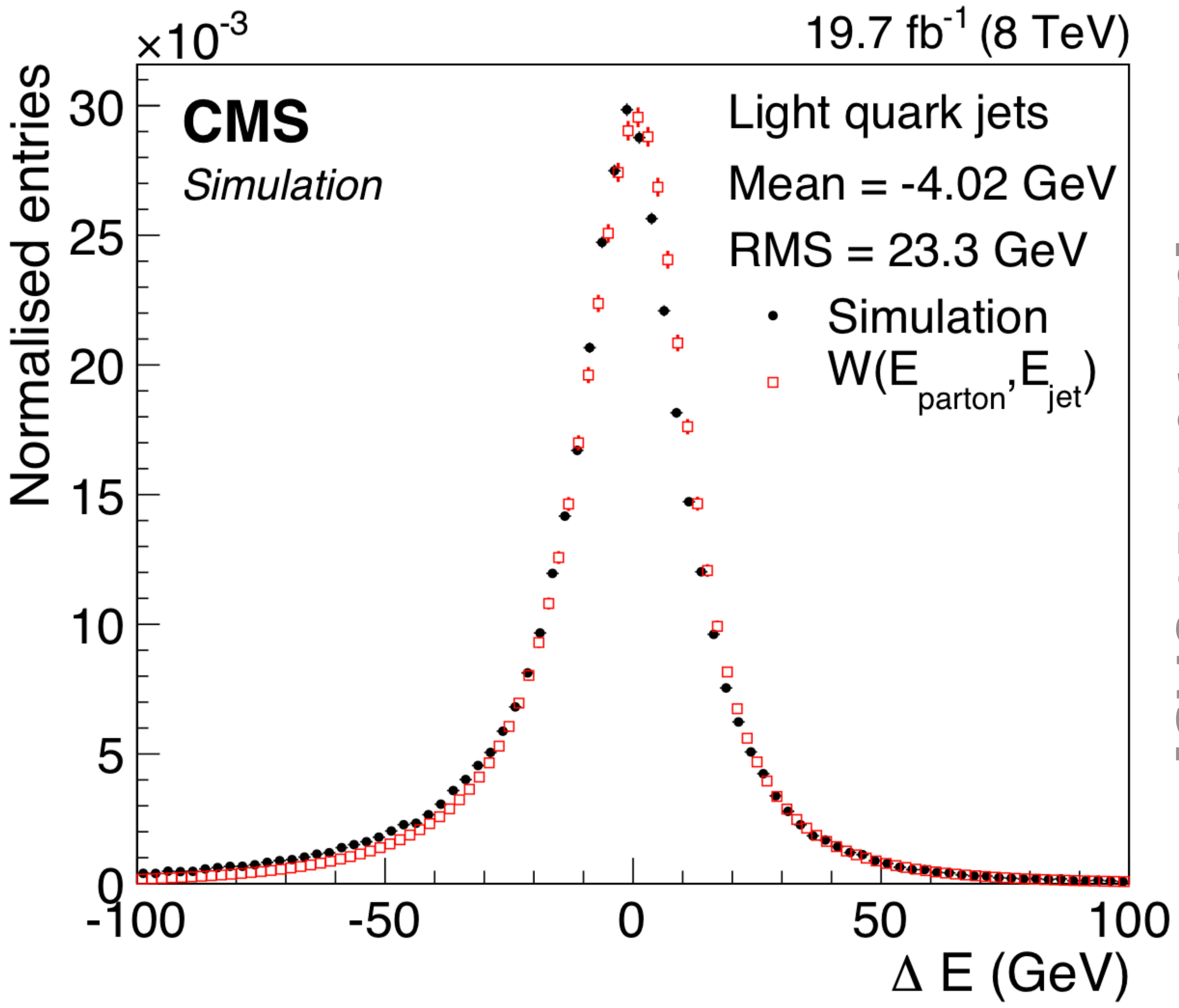
$$p(x|\theta) = \int \mathrm{d}z_d \int \mathrm{d}z_s \int \mathrm{d}z_p \; p(x|z_d)\, p(z_d|z_s)\, p(z_s|z_p)\, p(z_p|\theta)$$




[CMS]

- Matrix Element Method (and similarly Optimal Observables): [K. Kondo 1988]
  - approximate shower + detector effects into transfer function $\hat{p}(x|z_p)$
  - explicitly calculate remaining integral

$$\hat{p}(x|\theta) = \int \mathrm{d}z_p \; \hat{p}(x|z_p)\, p(z_p|\theta)$$




[CMS 1511.06170]

⇒ Uses matrix-element information, no summary statistics necessary, but:

- ad-hoc transfer functions (what about extra radiation?)
- evaluation still requires calculating an expensive integral

# What if we could estimate the likelihood...

- for high-dimensional observables, including correlations?

  like MEM: no need to pick summary statistics

- including state-of-the-art shower and detector models?

  allowing for extra radiation, no need for transfer functions

- in microseconds?

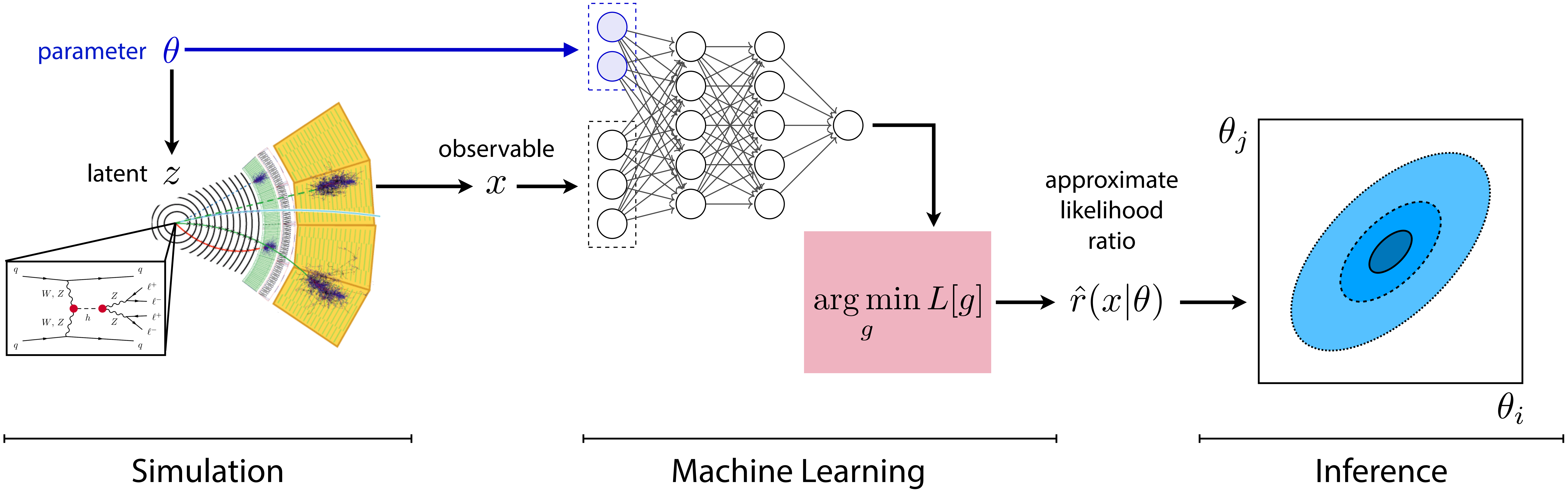
  amortized inference: train once, then always evaluate fast

- requiring less training examples than established machine learning methods?

  using matrix element information: "ML version of MEM"

# Learning with Simulated Data

arXiv:1805.12244
PRL, arXiv:1805.00013
PRD, arXiv:1805.00020
physics.aps.org/articles/v11/90



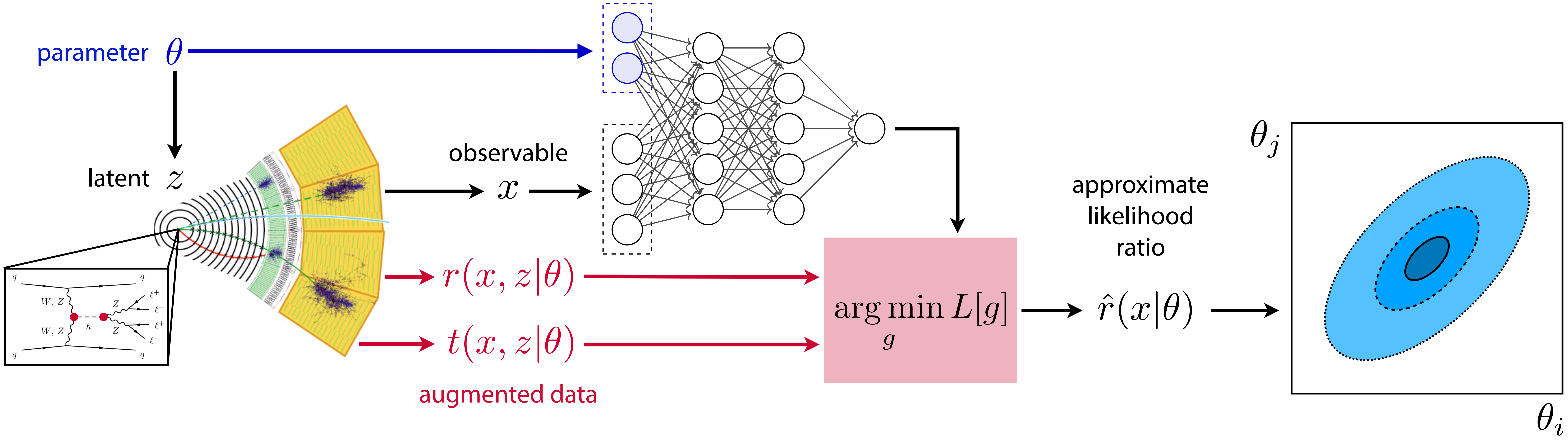

"Mining gold": Extract additional information from simulator

Use this information to train estimator for likelihood ratio

Limit setting with standard hypothesis tests

# Learning with Augmented Data

arXiv:1805.12244
PRL, arXiv:1805.00013
PRD, arXiv:1805.00020
physics.aps.org/articles/v11/90




**Simulation**

"Mining gold": Extract additional information from simulator

**Machine Learning**

Use this information to train estimator for likelihood ratio

**Inference**

Limit setting with standard hypothesis tests

# Mining gold from the simulator

Latent variables

| Observables | | Detector interactions | | Shower splittings | | Parton-level momenta | | Theory parameters |
|---|---|---|---|---|---|---|---|---|
| $x$ | ← | $z_d$ | ← | $z_s$ | ← | $z_p$ | ← | $\theta$ |

$$p(x|\theta) = \boxed{\int \mathrm{d}z_d \int \mathrm{d}z_s \int \mathrm{d}z_p} \; p(x|z_d) \quad p(z_d|z_s) \quad p(z_s|z_p) \quad p(z_p|\theta)$$

Intractable integrals

# Mining gold from the simulator

Latent variables

| Observables | | Detector interactions | Shower splittings | Parton-level momenta | Theory parameters |
|---|---|---|---|---|---|
| $x$ | $\longleftarrow$ | $z_d$ $\longleftarrow$ | $z_s$ $\longleftarrow$ | $z_p$ $\longleftarrow$ | $\theta$ |

$$p(x|\theta) = \int \mathrm{d}z_d \int \mathrm{d}z_s \int \mathrm{d}z_p \quad p(x|z_d) \quad p(z_d|z_s) \quad p(z_s|z_p) \quad \boxed{p(z_p|\theta)}$$

Parton-level likelihood is given by matrix element and can be evaluated!

⇒ For each simulated event, we can calculate the joint likelihood ratio which depends on the specific evolution of the simulation:

$$r(x,z|\theta_0,\theta_1) \equiv \frac{p(x,z_d,z_s,z_p|\theta_0)}{p(x,z_d,z_s,z_p|\theta_1)} = \frac{p(x|z_d)}{p(x|z_d)} \frac{p(z_d|z_s)}{p(z_d|z_s)} \frac{p(z_s|z_p)}{p(z_s|z_p)} \boxed{\frac{p(z_p|\theta_0)}{p(z_p|\theta_1)} \sim \frac{|\mathcal{M}(z_p|\theta_0)|^2}{|\mathcal{M}(z_p|\theta_1)|^2}}$$

# The value of gold

We can calculate the joint likelihood ratio

$$r(x, z|\theta_0, \theta_1) \equiv \frac{p(x, z_d, z_s, z_p|\theta_0)}{p(x, z_d, z_s, z_p|\theta_1)}$$

("How much more likely is this simulated event, including all intermediate states, for $\theta_0$ compared to $\theta_1$?")

↓

We want the likelihood ratio function

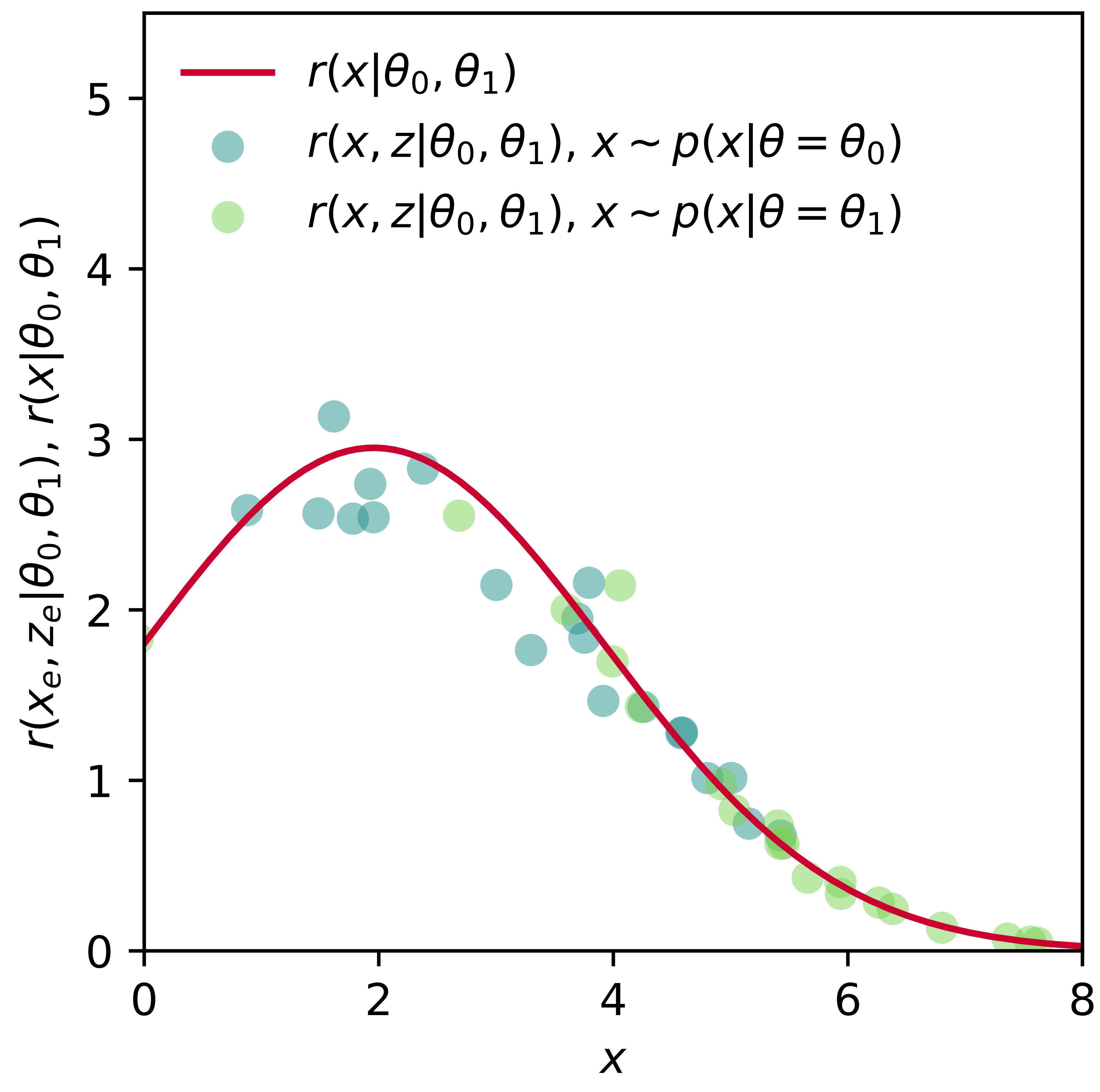
$$r(x|\theta_0, \theta_1) \equiv \frac{p(x|\theta_0)}{p(x|\theta_1)}$$

("How much more likely is the observation $x$ for $\theta_0$ compared to $\theta_1$?")

We can calculate the joint likelihood ratio

$$r(x, z|\theta_0, \theta_1) \equiv \frac{p(x, z_d, z_s, z_p|\theta_0)}{p(x, z_d, z_s, z_p|\theta_1)}$$

$r(x, z|\theta_0, \theta_1)$ are scattered around $r(x|\theta_0, \theta_1)$




We want the likelihood ratio function

$$r(x|\theta_0, \theta_1) \equiv \frac{p(x|\theta_0)}{p(x|\theta_1)}$$

# The value of gold

We can calculate the joint likelihood ratio

$$r(x, z|\theta_0, \theta_1) \equiv \frac{p(x, z_d, z_s, z_p|\theta_0)}{p(x, z_d, z_s, z_p|\theta_1)}$$

↓

We want the likelihood ratio function

$$r(x|\theta_0, \theta_1) \equiv \frac{p(x|\theta_0)}{p(x|\theta_1)}$$

With $r(x, z|\theta_0, \theta_1)$, we define a functional like

$$L_r[\hat{r}(x|\theta_0, \theta_1)] = \int \mathrm{d}x \int \mathrm{d}z \; p(x, z|\theta_1) \left[ (\hat{r}(x|\theta_0, \theta_1) - r(x, z|\theta_0, \theta_1))^2 \right]$$

It is minimized by

$$\mathbb{E}_{z \sim p(z|x, \theta_1)} [r(x, z|\theta_0, \theta_1)] = \underset{\hat{r}(x|\theta_0, \theta_1)}{\arg\min} \; L_r[\hat{r}(x|\theta_0, \theta_1)]!$$

(And we can sample from $p(x, z|\theta)$ by running the simulator.)

# The value of gold

We can calculate the joint likelihood ratio

$$r(x, z|\theta_0, \theta_1) \equiv \frac{p(x, z_d, z_s, z_p|\theta_0)}{p(x, z_d, z_s, z_p|\theta_1)}$$

↓

We want the likelihood ratio function

$$r(x|\theta_0, \theta_1) \equiv \frac{p(x|\theta_0)}{p(x|\theta_1)}$$

With $r(x, z|\theta_0, \theta_1)$, we define a functional like

$$L_r[\hat{r}(x|\theta_0, \theta_1)] = \int \mathrm{d}x \int \mathrm{d}z \; p(x, z|\theta_1) \left[ (\hat{r}(x|\theta_0, \theta_1) - r(x, z|\theta_0, \theta_1))^2 \right]$$

It is minimized by

$$\mathbb{E}_{z \sim p(z|x,\theta_1)} [r(x, z|\theta_0, \theta_1)] = \underset{\hat{r}(x|\theta_0,\theta_1)}{\arg\min} \; L_r[\hat{r}(x|\theta_0, \theta_1)]!$$

(And we can sample from $p(x, z|\theta)$ by running the simulator.)

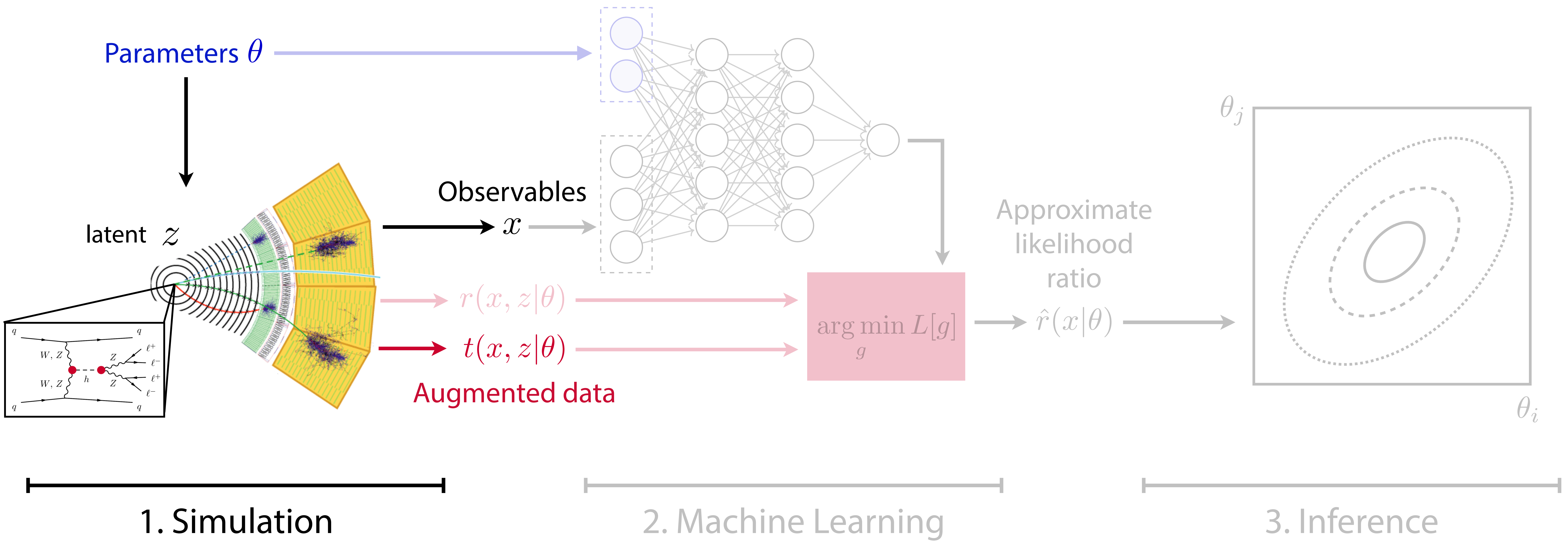
**…. and then magic …**

$$\mathbb{E}_{z \sim p(z|x,\theta_1)} [r(x, z|\theta_0, \theta_1)] = \int \mathrm{d}z \; p(z|x, \theta_1) \frac{p(x, z|\theta_0)}{p(x, z|\theta_1)}$$

$$= \int \mathrm{d}z \; \frac{p(x, z|\theta_1)}{p(x|\theta_1)} \frac{p(x, z|\theta_0)}{p(x, z|\theta_1)}$$

$$= r(x|\theta_0, \theta_1) \; !$$

# Learning with Augmented Data



# Learning the score (related to optimal observables)

Similar to the joint likelihood ratio, from the simulator we can extract the joint score

$$t(x, z|\theta_0) \equiv \nabla_\theta \log p(x, z_d, z_s, z_p|\theta)\Big|_{\theta_0}$$

↓

We want the score

$$t(x|\theta_0) \equiv \nabla_\theta \log p(x|\theta)\Big|_{\theta_0}$$

# Learning the score (related to optimal observables)

Similar to the joint likelihood ratio, from the simulator we can extract the joint score

$$t(x, z|\theta_0) \equiv \nabla_\theta \log p(x, z_d, z_s, z_p|\theta)\Big|_{\theta_0}$$

We want the score

$$t(x|\theta_0) \equiv \nabla_\theta \log p(x|\theta)\Big|_{\theta_0}$$

Given $t(x, z|\theta_0)$,
we define the functional

$$L_t[\hat{t}(x|\theta_0)] = \int \mathrm{d}x \int \mathrm{d}z \;\; p(x, z|\theta_0) \; \left[\left(\hat{t}(x|\theta_0) - t(x, z|\theta_0)\right)^2\right].$$

One can show it is minimized by

$$t(x|\theta_0) = \underset{\hat{t}(x|\theta_0)}{\arg\min}\, L_t[\hat{t}(x|\theta_0)].$$

Again, we implement this minimization through machine learning.

MadMiner automates all of these methods.

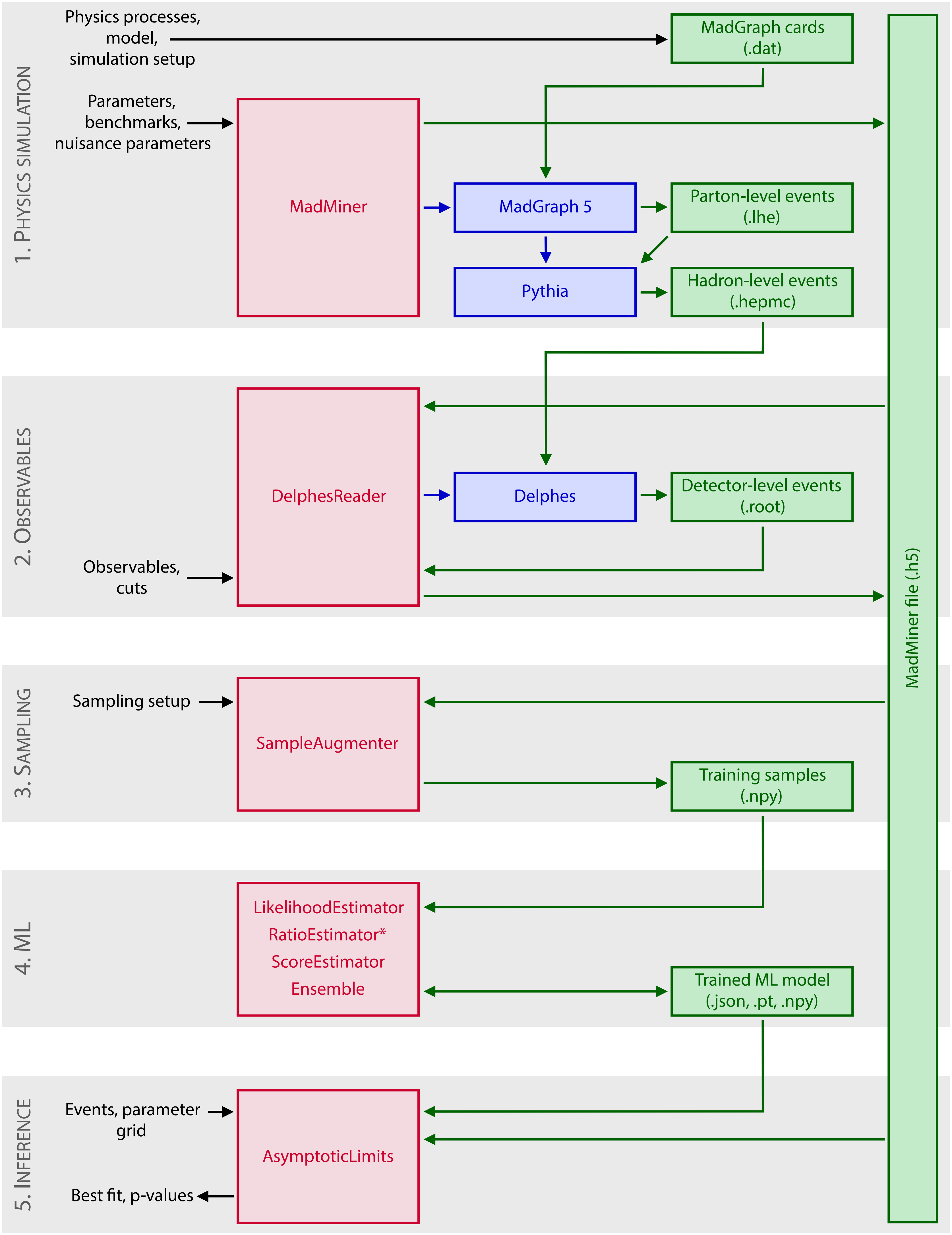
[JB, F. Kling, I. Espejo, K. Cranmer 1907.10621]

# MadMiner

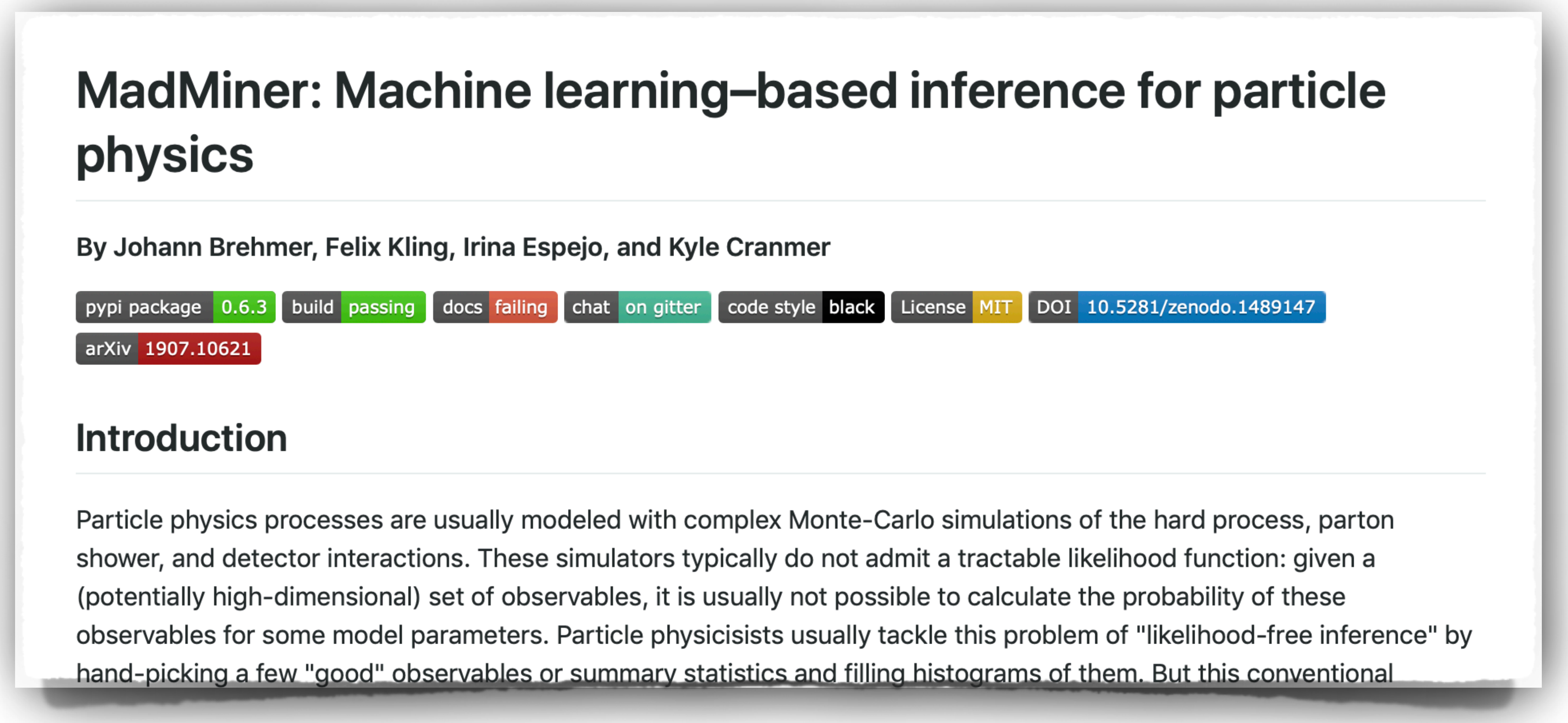
[JB, F. Kling, I. Espejo, K. Cranmer 1907.10621]

New Python package MadMiner makes it straightforward to apply the new techniques to LHC problems

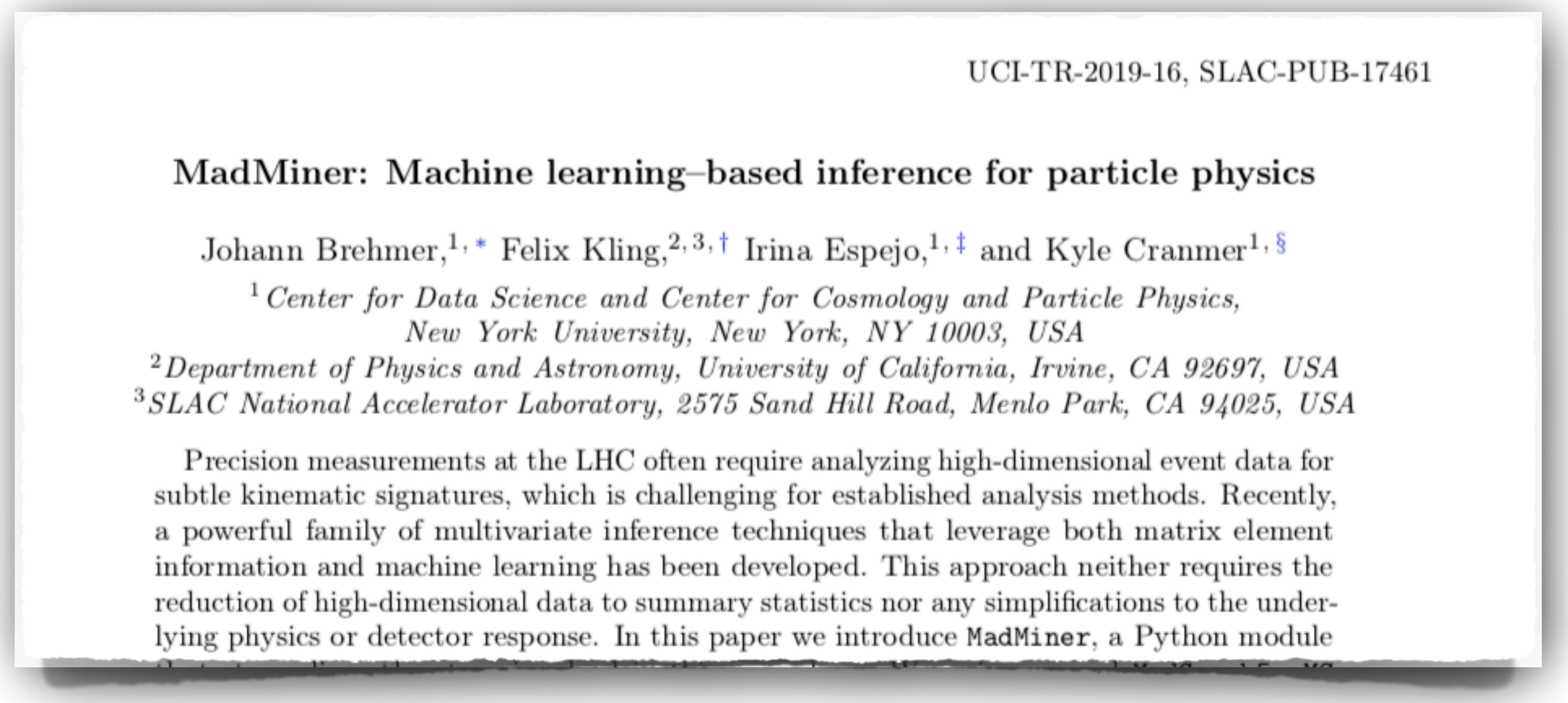
- Out of the box: Pheno-level analyses
  - MadGraph, Pythia, Delphes, (could be GEANT4)
  - Systematic uncertainties from PDF / scale variation
- Scalable to state-of-the-art experimental tools
  - Mostly requires bookkeeping of fully differential cross sections
- Modular interface
  - Extensive documentation
  - Embedded into Python / ML ecosystem



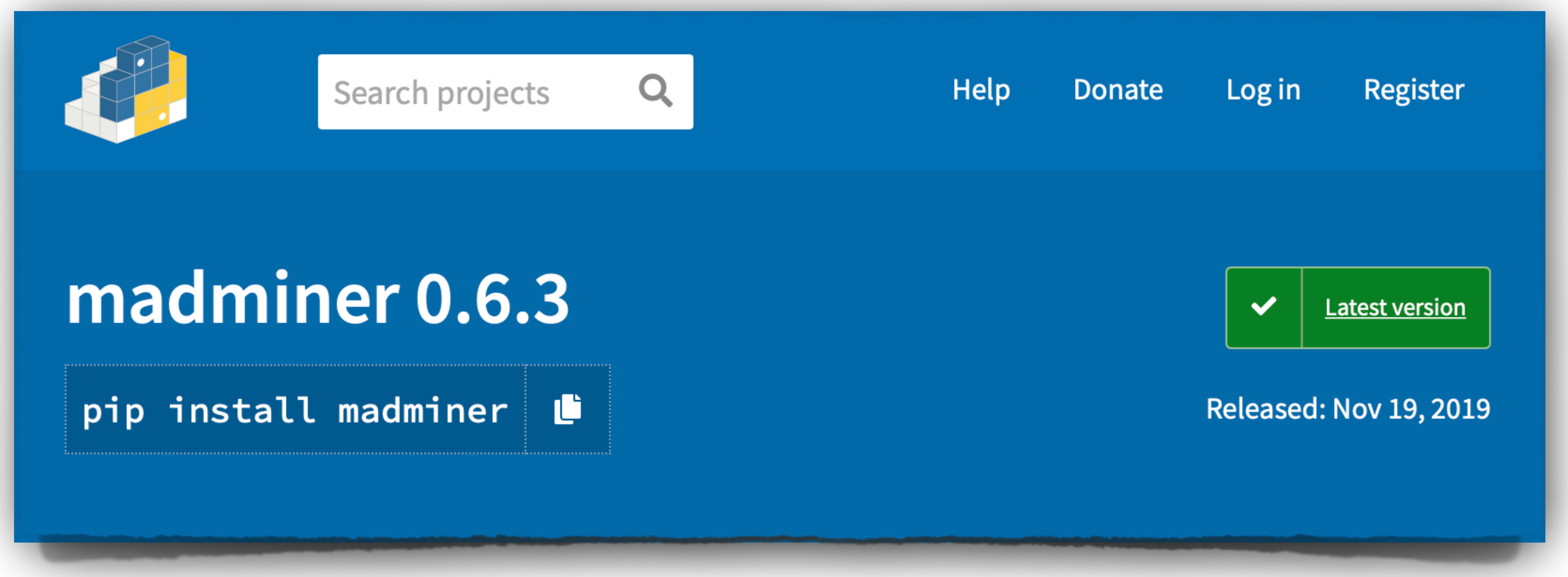
# MadMiner resources



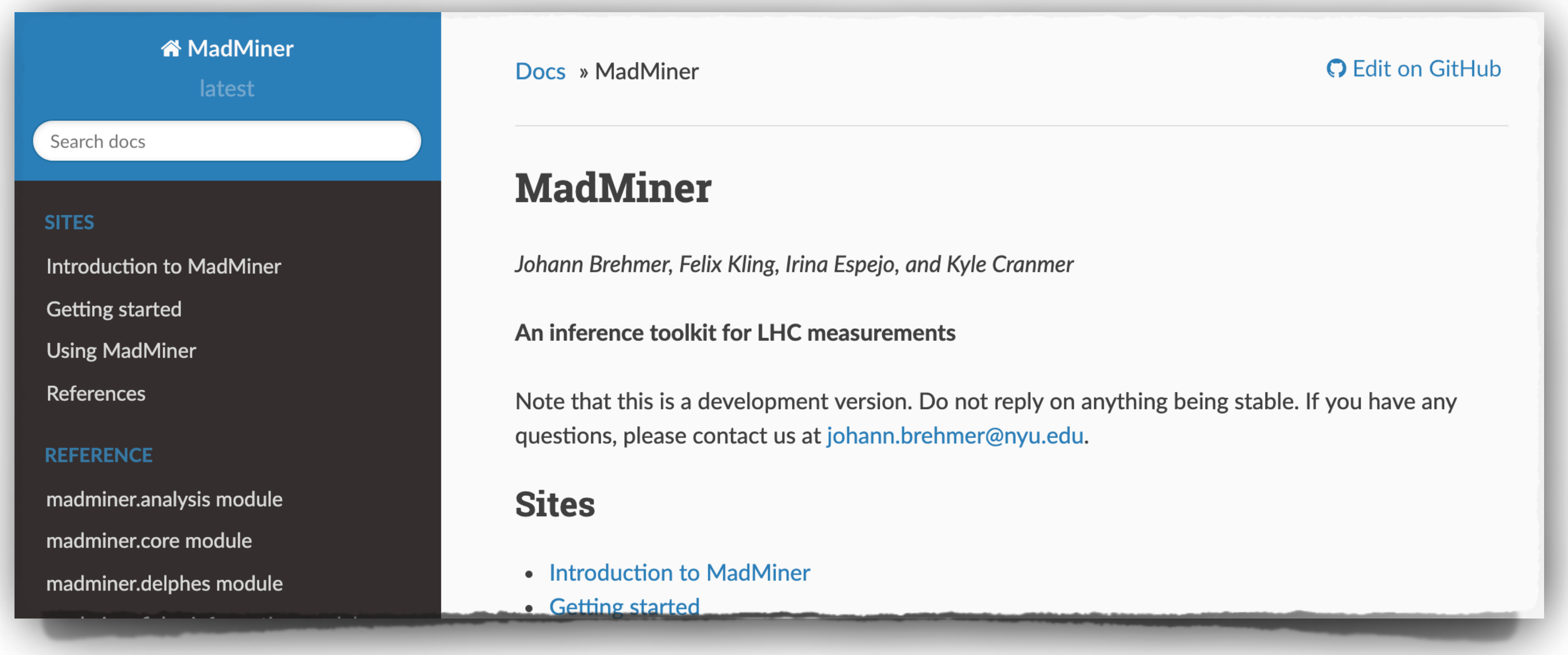

Repository and tutorials:
github.com/johannbrehmer/madminer




Paper with detailed explanations:
1907.10621




Installation:
`pip install madminer`



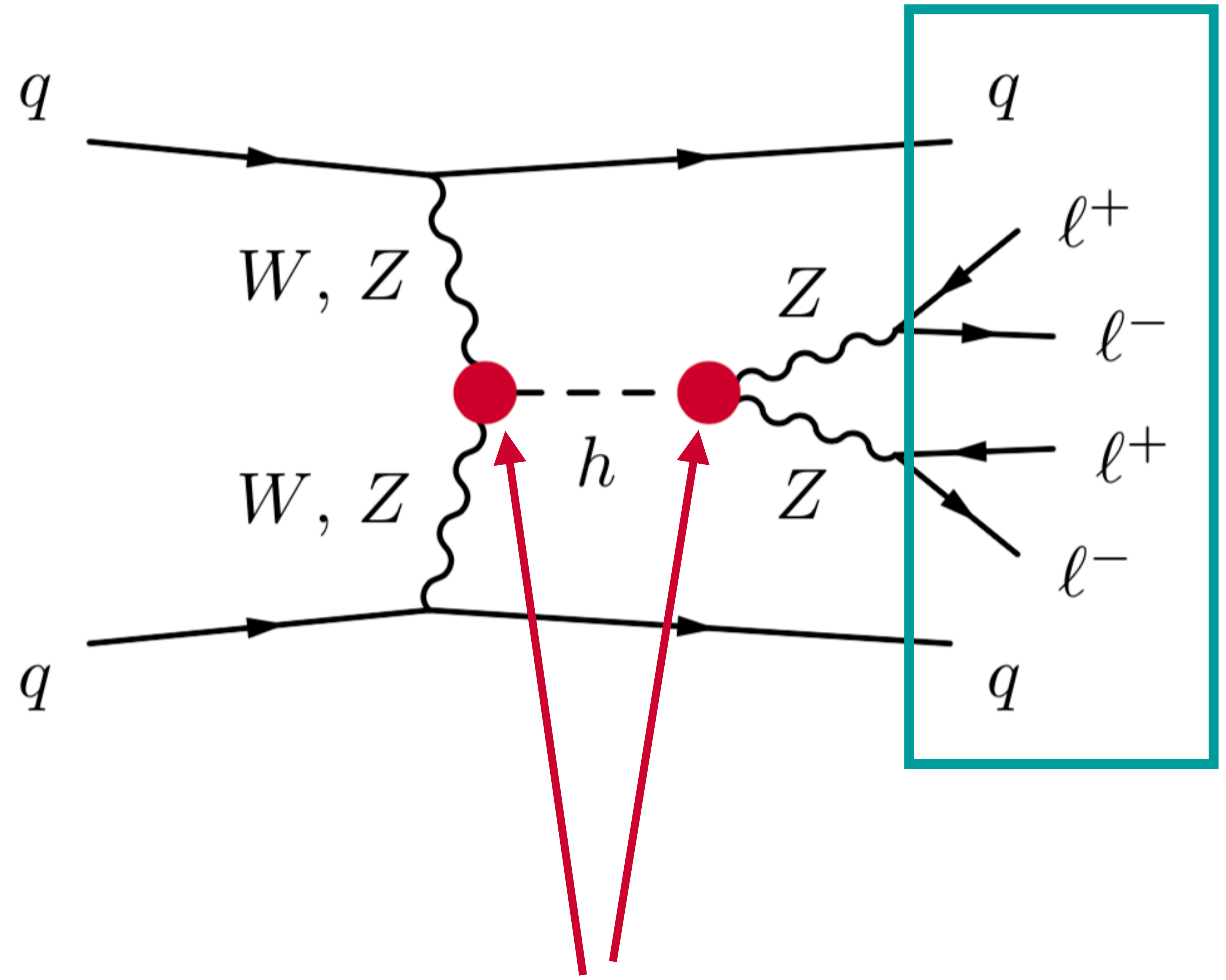

API documentation:
madminer.readthedocs.io

These techniques let us constrain effective theories more effectively.

# Proof of concept: Higgs production in weak boson fusion

[JB, K. Cranmer, G. Louppe, J. Pavez
1805.00013, 1805.00020]

at least 16-dimensional
observable space

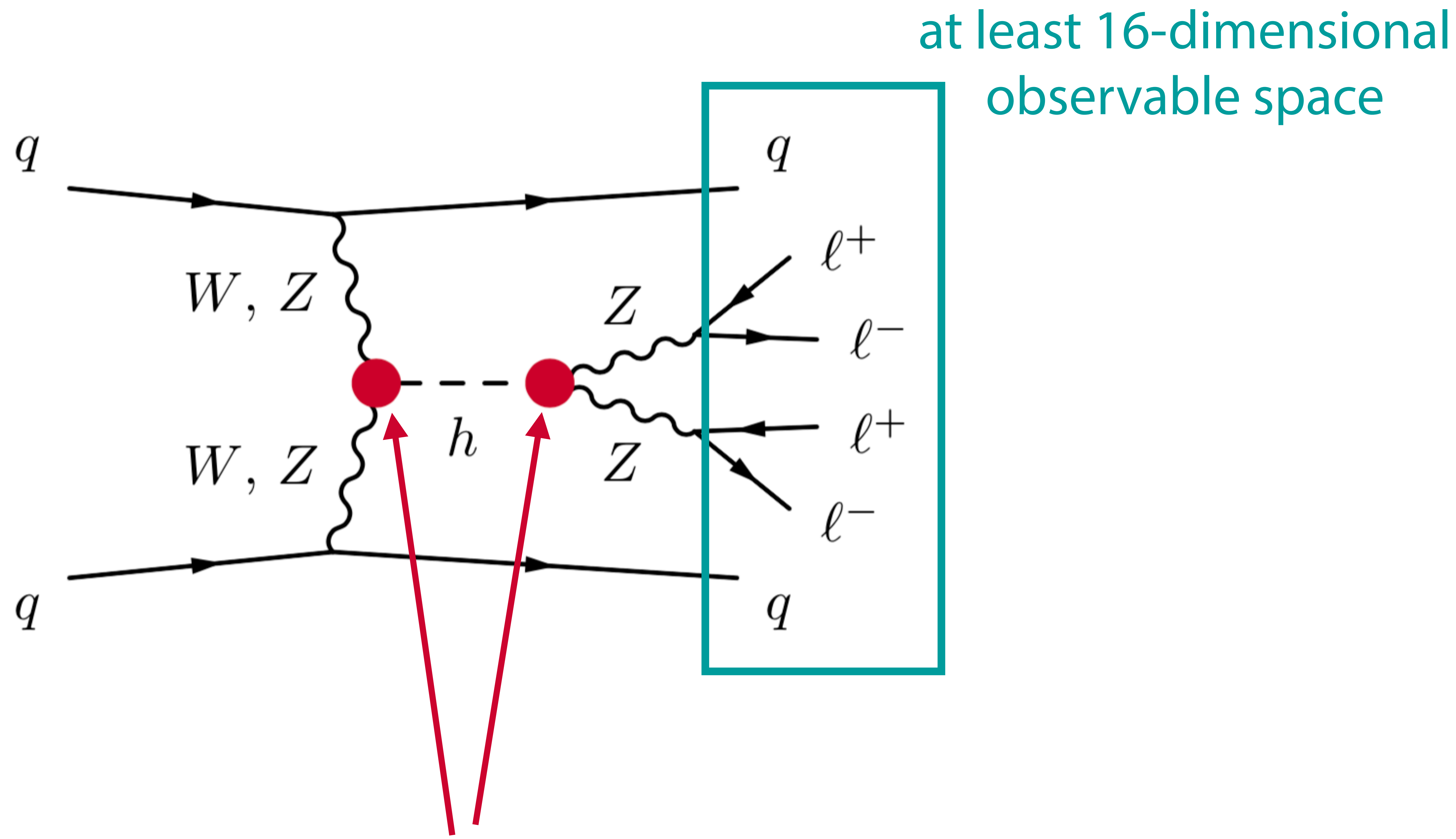



Exciting new physics might hide here!
We parameterize it with two EFT coefficients:

$$\mathcal{L} = \mathcal{L}_{SM} + \frac{f_W}{\Lambda^2} \underbrace{\frac{ig}{2} (D^\mu \phi)^\dagger \sigma^a D^\nu \phi \, W^a_{\mu\nu}}_{\mathcal{O}_W} - \frac{f_{WW}}{\Lambda^2} \underbrace{\frac{g^2}{4} (\phi^\dagger \phi) \, W^a_{\mu\nu} W^{\mu\nu\, a}}_{\mathcal{O}_{WW}}$$

# Proof of concept: Higgs production in weak boson fusion

[JB, K. Cranmer, G. Louppe, J. Pavez 1805.00013, 1805.00020]




Exciting new physics might hide here!
We parameterize it with two EFT coefficients:

$$\mathcal{L} = \mathcal{L}_{SM} + \frac{f_W}{\Lambda^2} \underbrace{\frac{ig}{2} (D^\mu\phi)^\dagger \sigma^a D^\nu\phi \, W^a_{\mu\nu}}_{\mathcal{O}_W} - \frac{f_{WW}}{\Lambda^2} \underbrace{\frac{g^2}{4} (\phi^\dagger\phi) \, W^a_{\mu\nu} W^{\mu\nu\,a}}_{\mathcal{O}_{WW}}$$

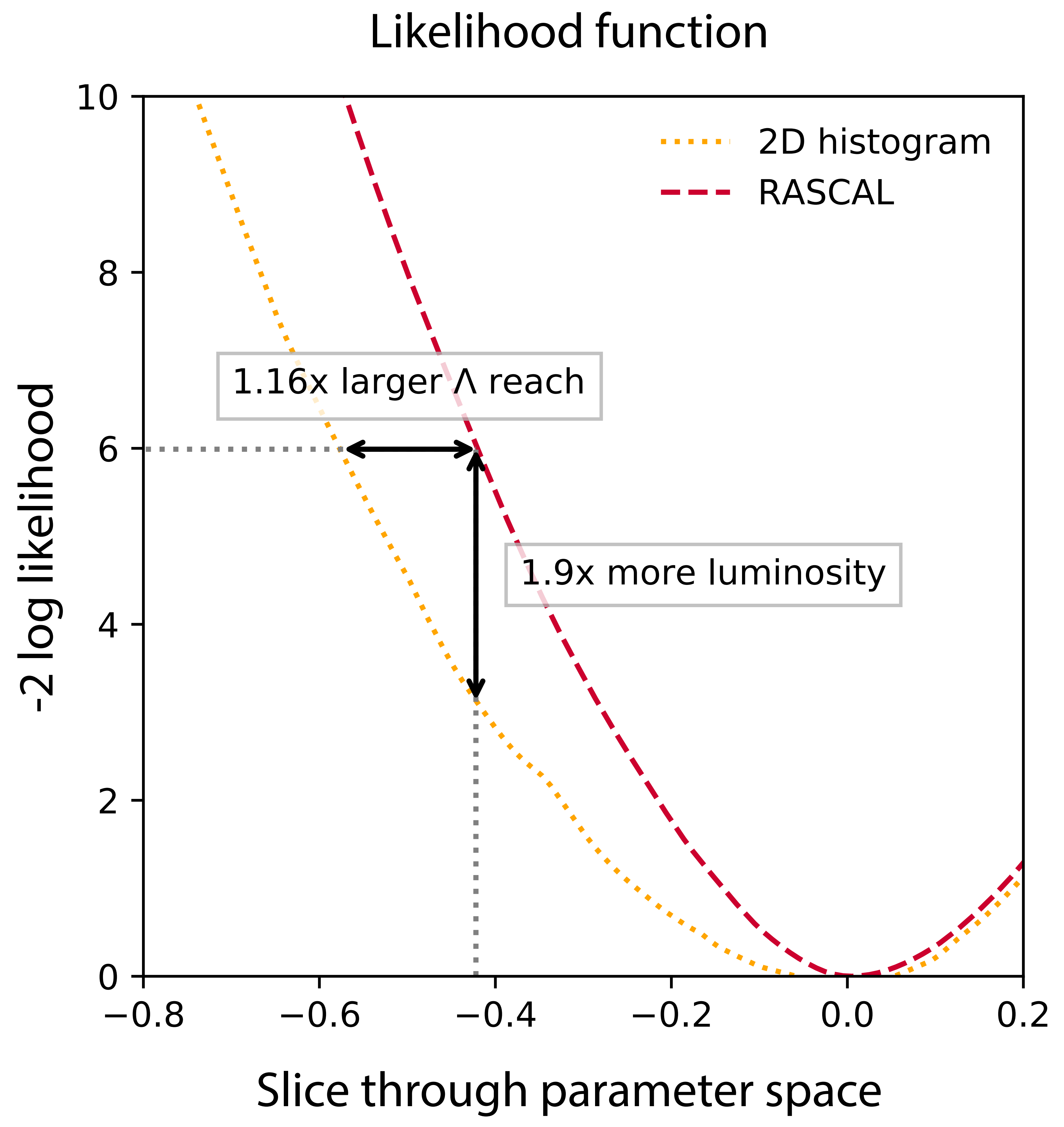
Goal: constrain the two EFT parameters

- new inference methods
- baseline: 2d histogram analysis of jet momenta & angular correlations

Two scenarios:

- Simplified setup in which we can compare to true likelihood
- "Realistic" simulation with approximate detector effects

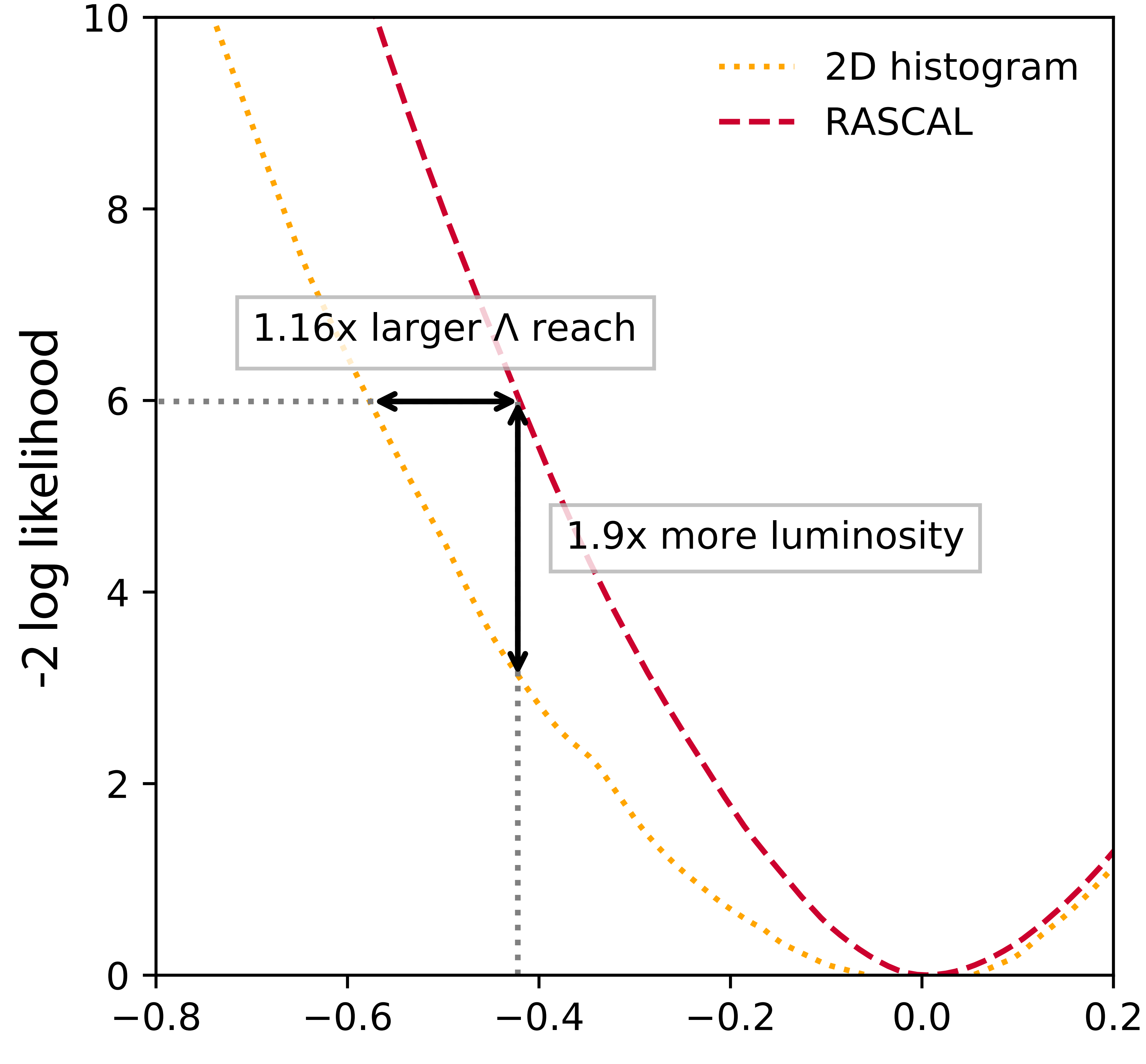
# Better sensitivity to new physics



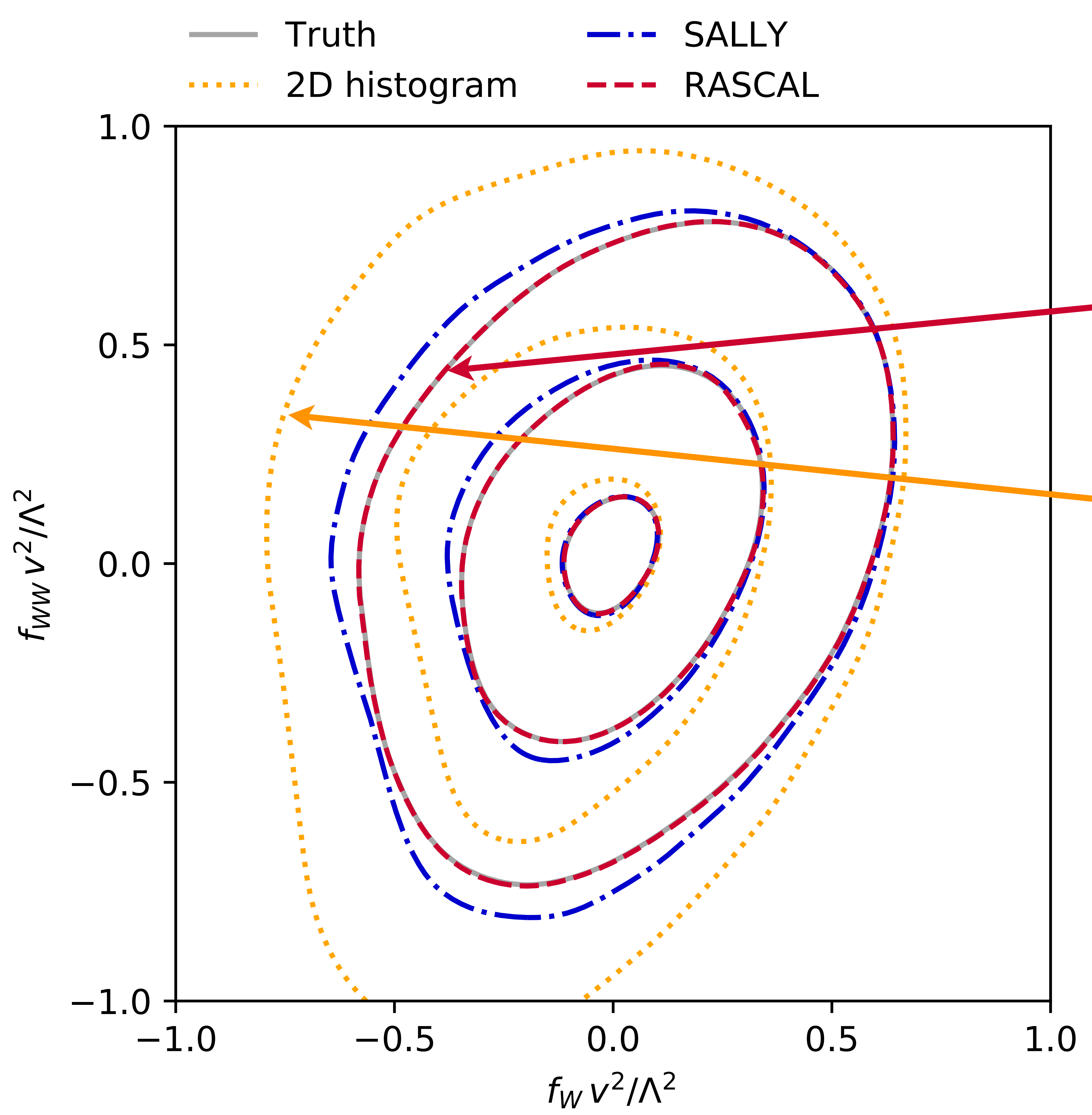

Results are based on 36 observed events, assuming SM

# Better sensitivity to new physics

Likelihood function



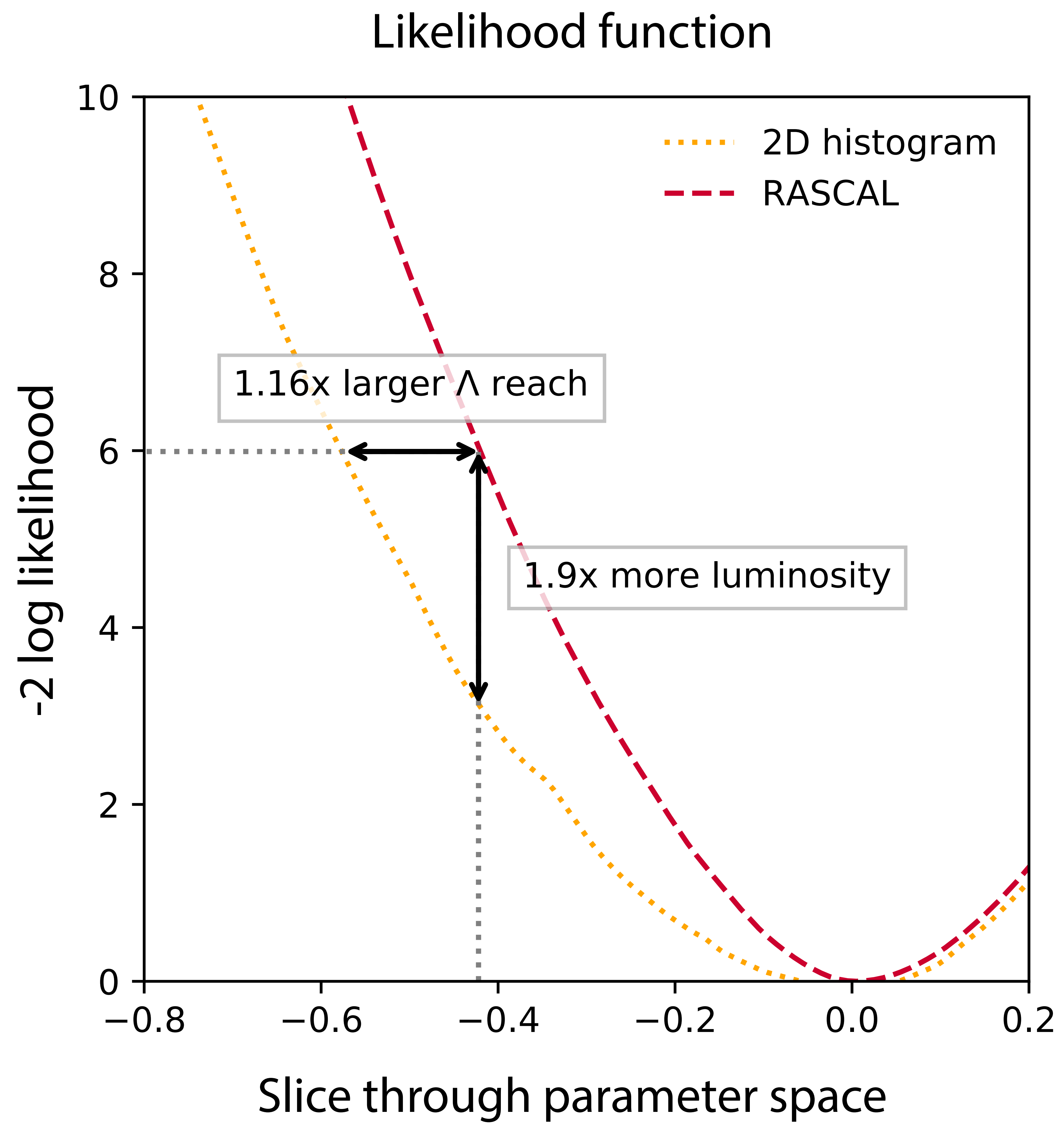

Expected exclusion limits at 68%, 95%, 99.7% CL

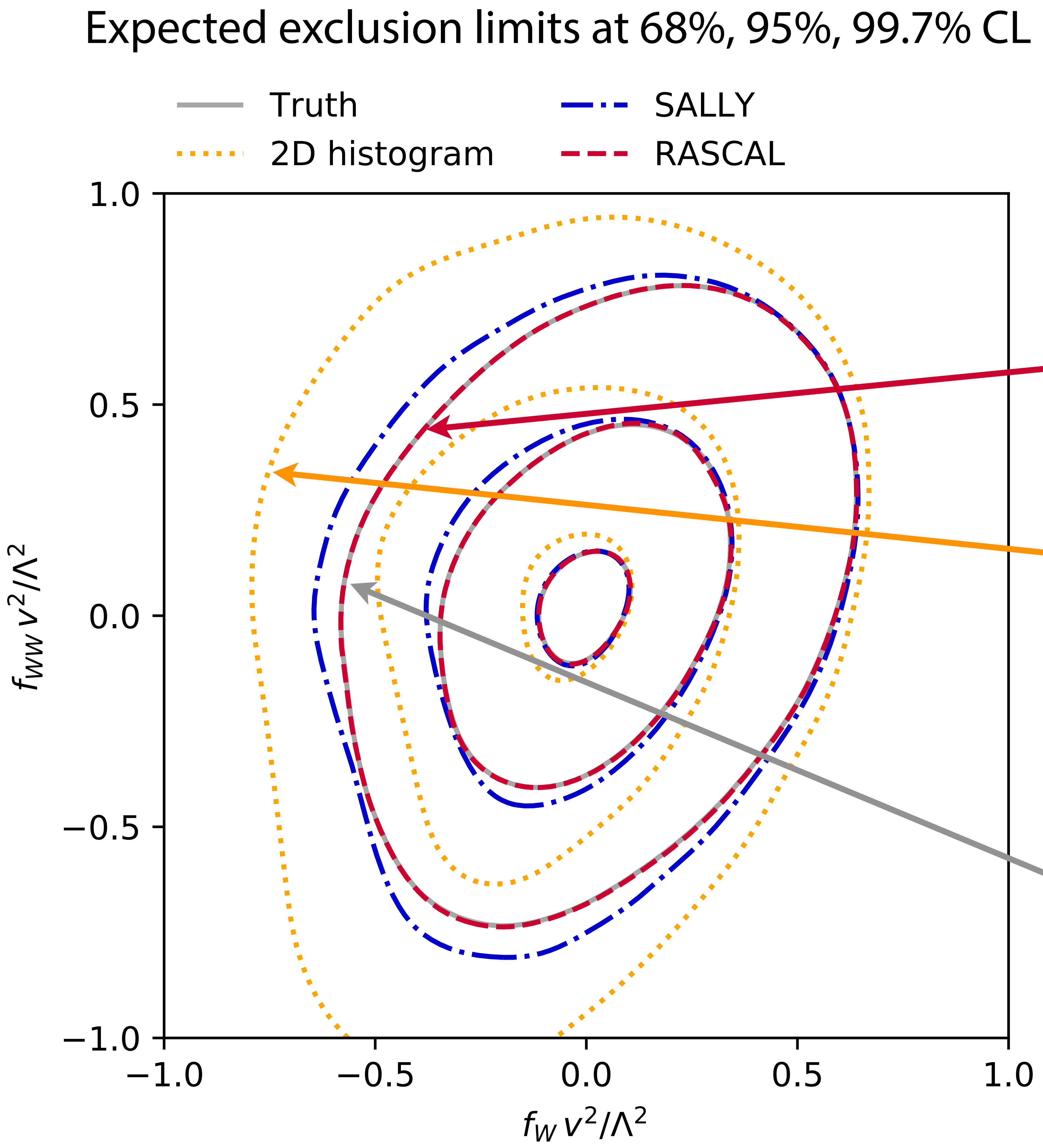



RASCAL and SALLY enables stronger limits than 2D histogram

Results are based on 36 observed events, assuming SM

# Better sensitivity to new physics




RASCAL and SALLY enables stronger limits than 2D histogram

Limits from RASCAL indistinguishable from true likelihood

(usually we don't have that)

Results are based on 36 observed events, assuming SM

# Constraining operators in ttH effectively

[JB, F. Kling, I. Espejo, K. Cranmer 1907.10621]

- Pheno-level analysis of

  $$pp \to t\bar{t}\, h \to (b\ell^+)\,(\bar{b}\ell^-)\,(\gamma\gamma)\,E_T^{\text{miss}}$$

  with MadGraph + Pythia + Delphes

- Inference on three EFT operators:

  $$\mathcal{O}_u = -\frac{1}{v^2}(H^\dagger H)(H^\dagger \bar{Q}_L)u_R\,,\ \mathcal{O}_G = \frac{g_s^2}{m_W^2}(H^\dagger H)G^a_{\mu\nu}G_a^{\mu\nu}\,,$$

  $$\mathcal{O}_{uG} = -\frac{4g_s}{m_W^2}y_u(H^\dagger \bar{Q}_L)\gamma^{\mu\nu}T_a u_R G^a_{\mu\nu}$$

# Constraining operators in ttH effectively

[JB, F. Kling, I. Espejo, K. Cranmer 1907.10621]

- Pheno-level analysis of

$$pp \to t\bar{t}\,h \to (b\ell^+)\,(\bar{b}\ell^-)\,(\gamma\gamma)\,E_T^{\text{miss}}$$

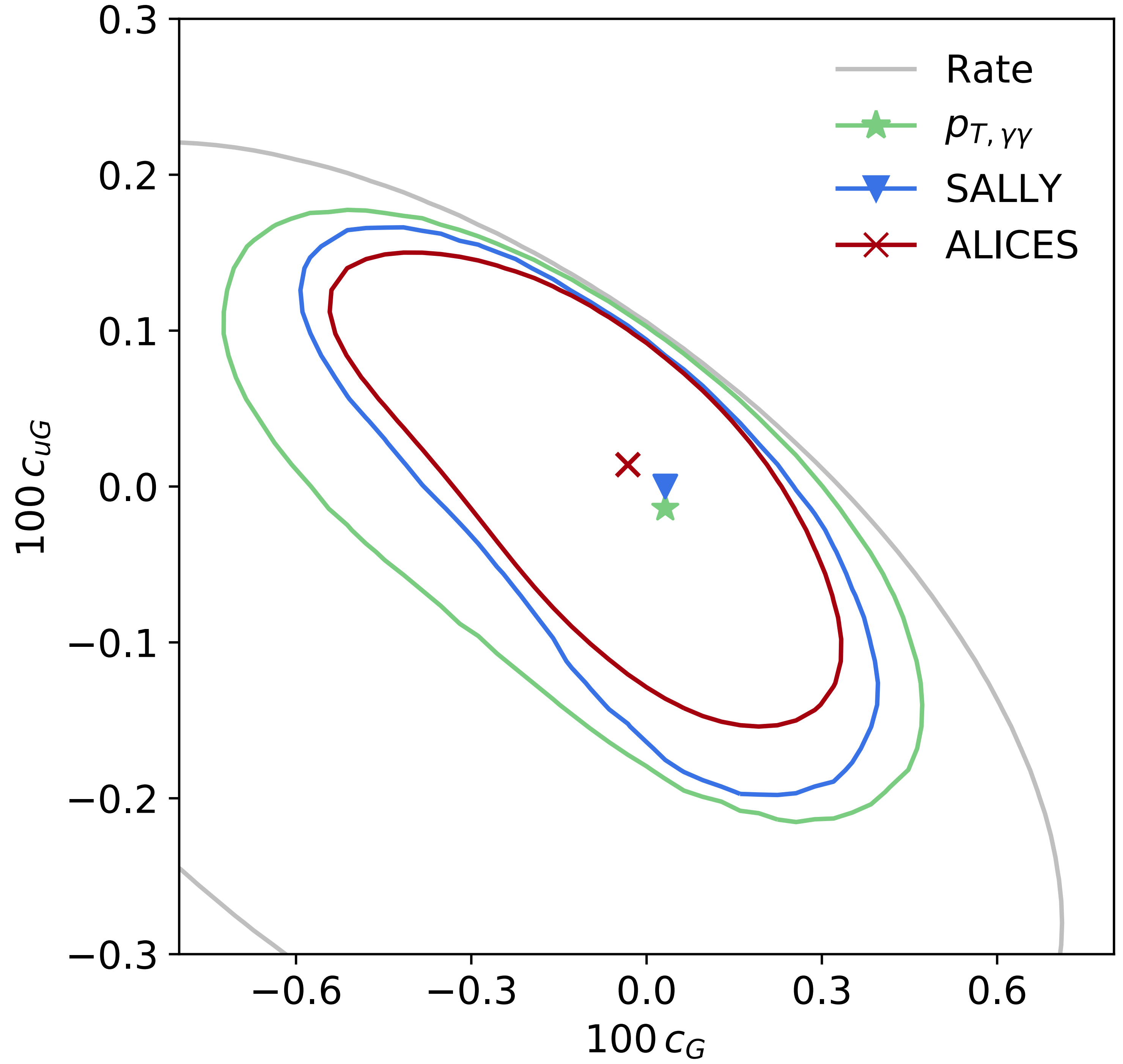
with MadGraph + Pythia + Delphes

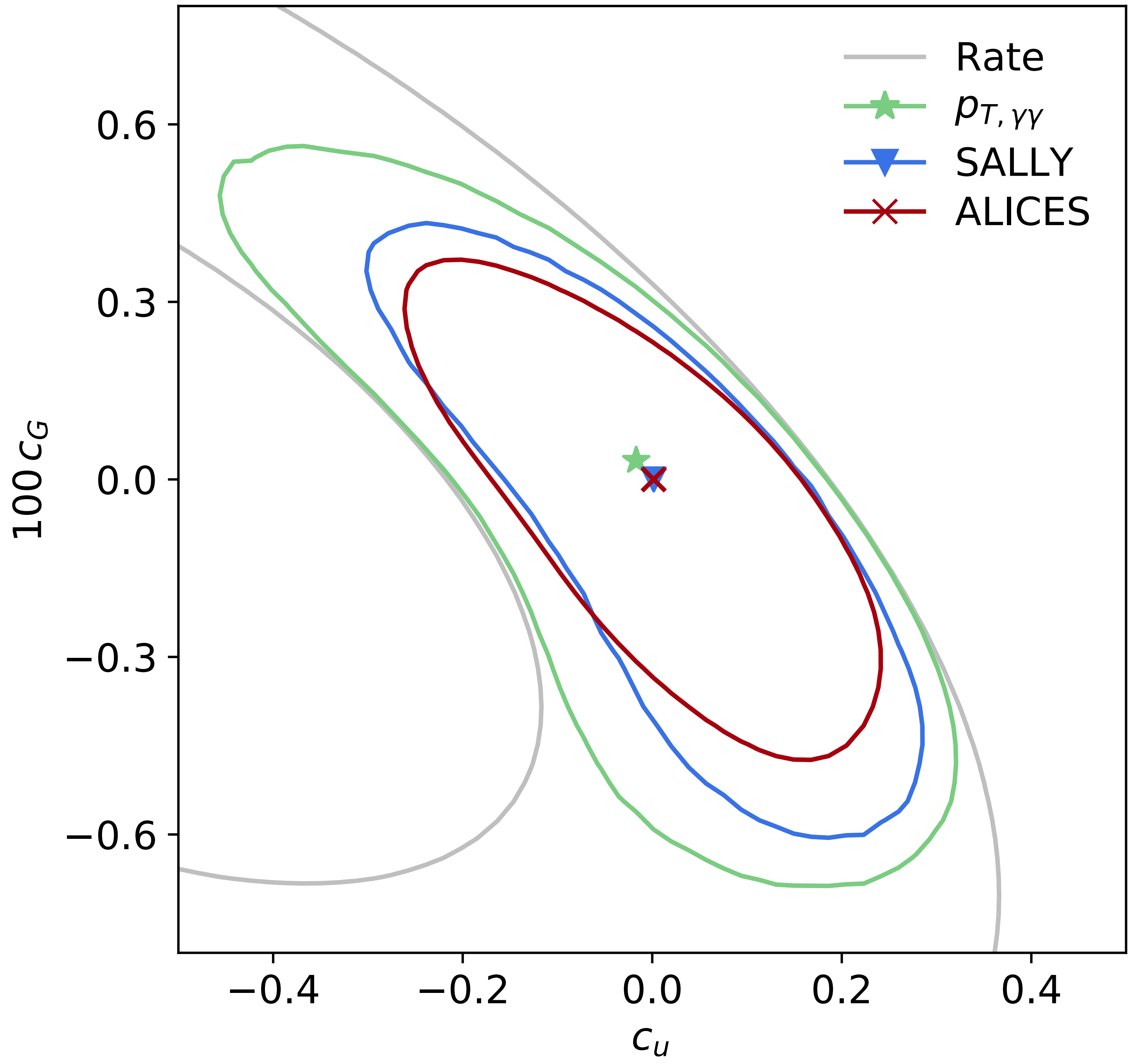
- Inference on three EFT operators:

$$\mathcal{O}_u = -\frac{1}{v^2}(H^\dagger H)(H^\dagger \bar{Q}_L)u_R\,,\ \mathcal{O}_G = \frac{g_s^2}{m_W^2}(H^\dagger H)G^a_{\mu\nu}G_a^{\mu\nu}\,,$$

$$\mathcal{O}_{uG} = -\frac{4g_s}{m_W^2}y_u(H^\dagger \bar{Q}_L)\gamma^{\mu\nu}T_a u_R G^a_{\mu\nu}$$

- New inference techniques improve expected HL-LHC limits compared to histogram baseline:

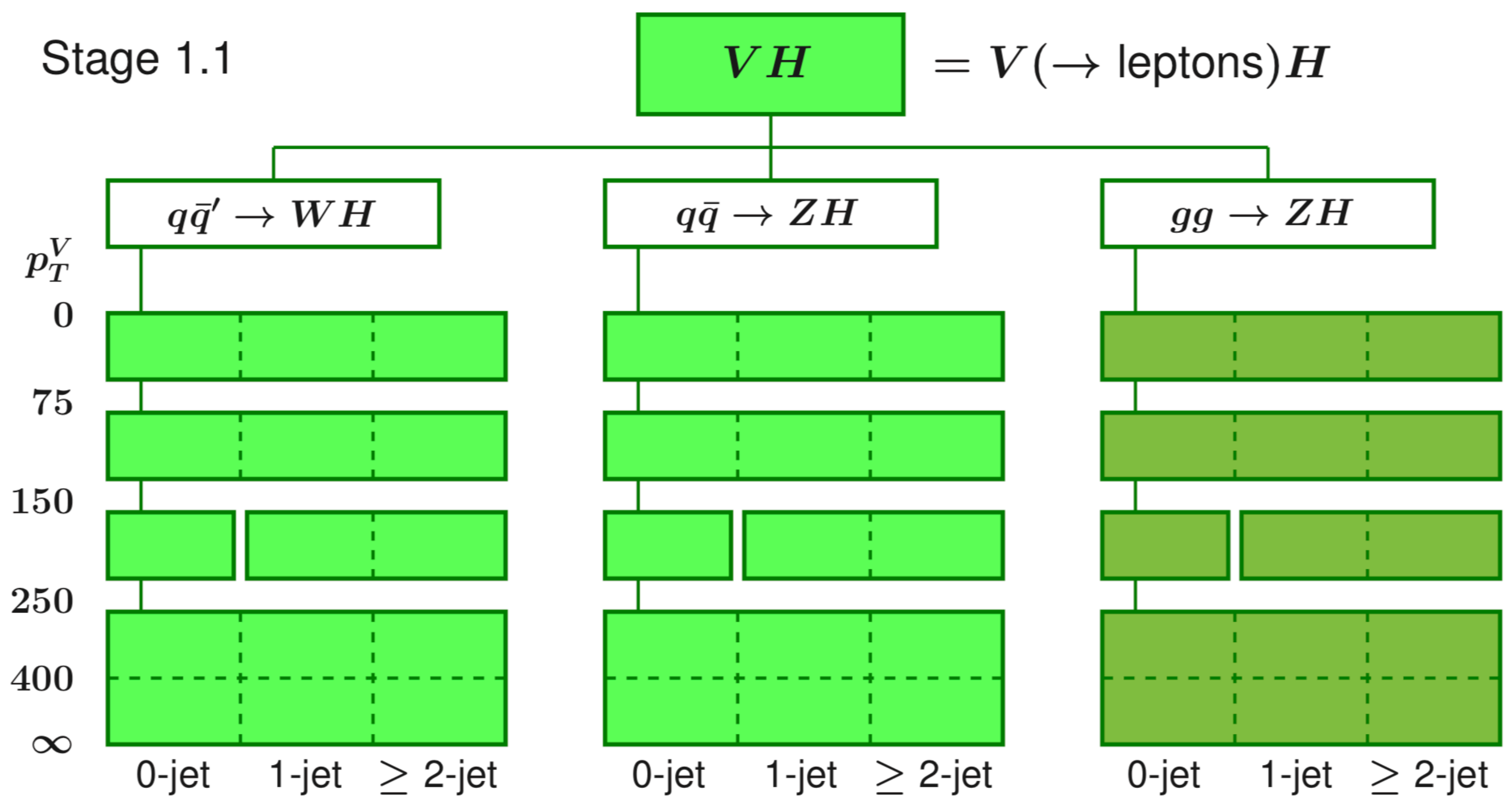


# Benchmarking STXS in WH

[JB, S. Dawson, S. Homiller, F. Kling, T. Plehn 1908.06980]

- Simplified Template Cross-Sections (STXS) define observable bins that are supposed to capture as much information on NP as possible
  [N. Berger et al. 1906.02754; HXSWG YR4]




- Let's check! How much information on

$$\tilde{\mathcal{O}}_{HD} = \mathcal{O}_{H\Box} - \frac{\mathcal{O}_{HD}}{4} = (\phi^\dagger\phi)\Box(\phi^\dagger\phi) - \frac{1}{4}(\phi^\dagger D^\mu\phi)^*(\phi^\dagger D_\mu\phi)$$

$$\mathcal{O}_{HW} = \phi^\dagger\phi W^a_{\mu\nu}W^{\mu\nu a}$$

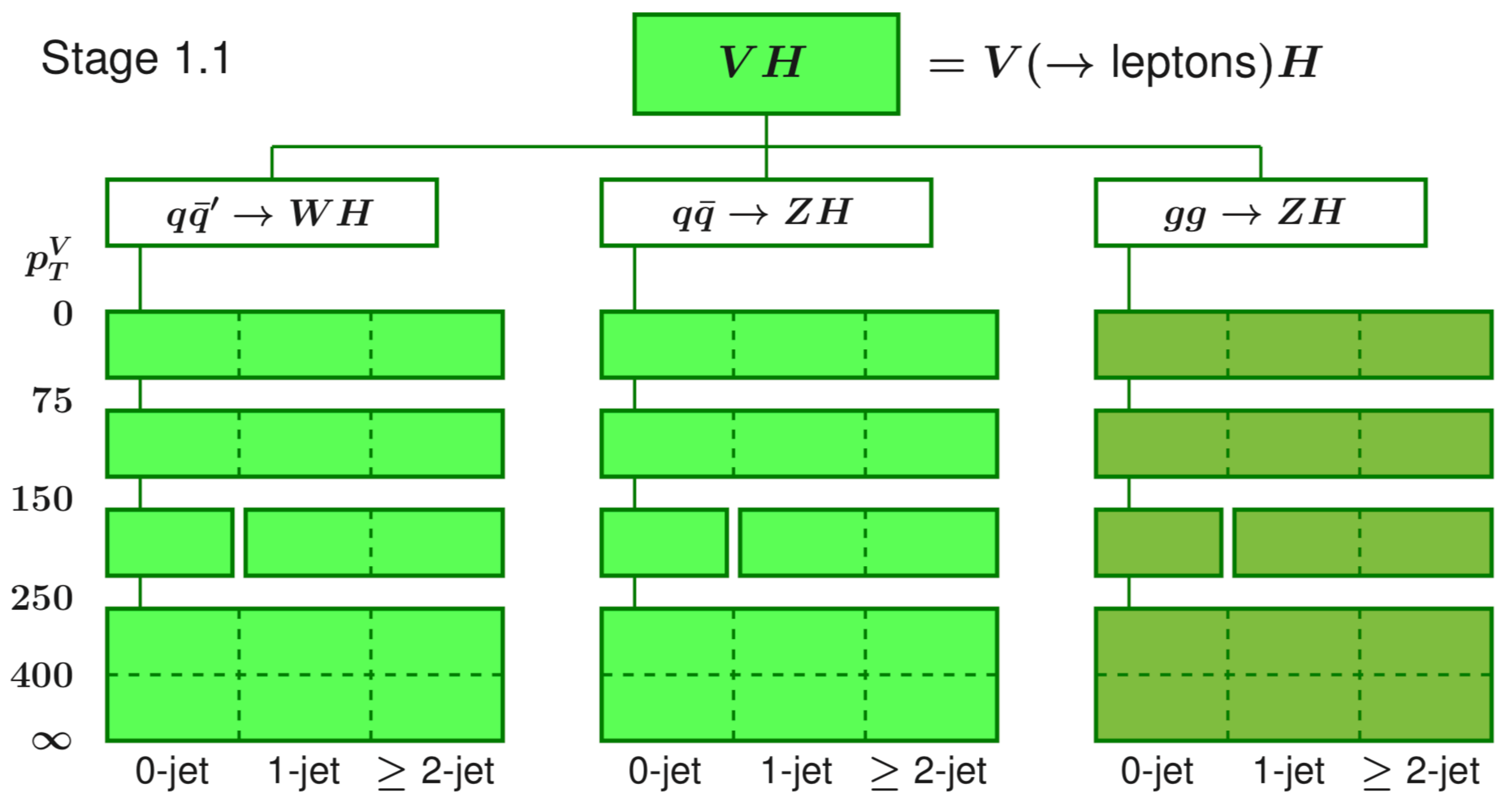
$$\mathcal{O}^{(3)}_{Hq} = (\phi^\dagger i\overleftrightarrow{D}^a_\mu\phi)(\overline{Q}_L\sigma^a\gamma^\mu Q_L)\,,$$

can we extract from $pp \to WH \to \ell\nu\, b\bar{b}$ ?

# Benchmarking STXS in WH

[JB, S. Dawson, S. Homiller, F. Kling, T. Plehn 1908.06980]

- Simplified Template Cross-Sections (STXS) define observable bins that are supposed to capture as much information on NP as possible
  [N. Berger et al. 1906.02754; HXSWG YR4]

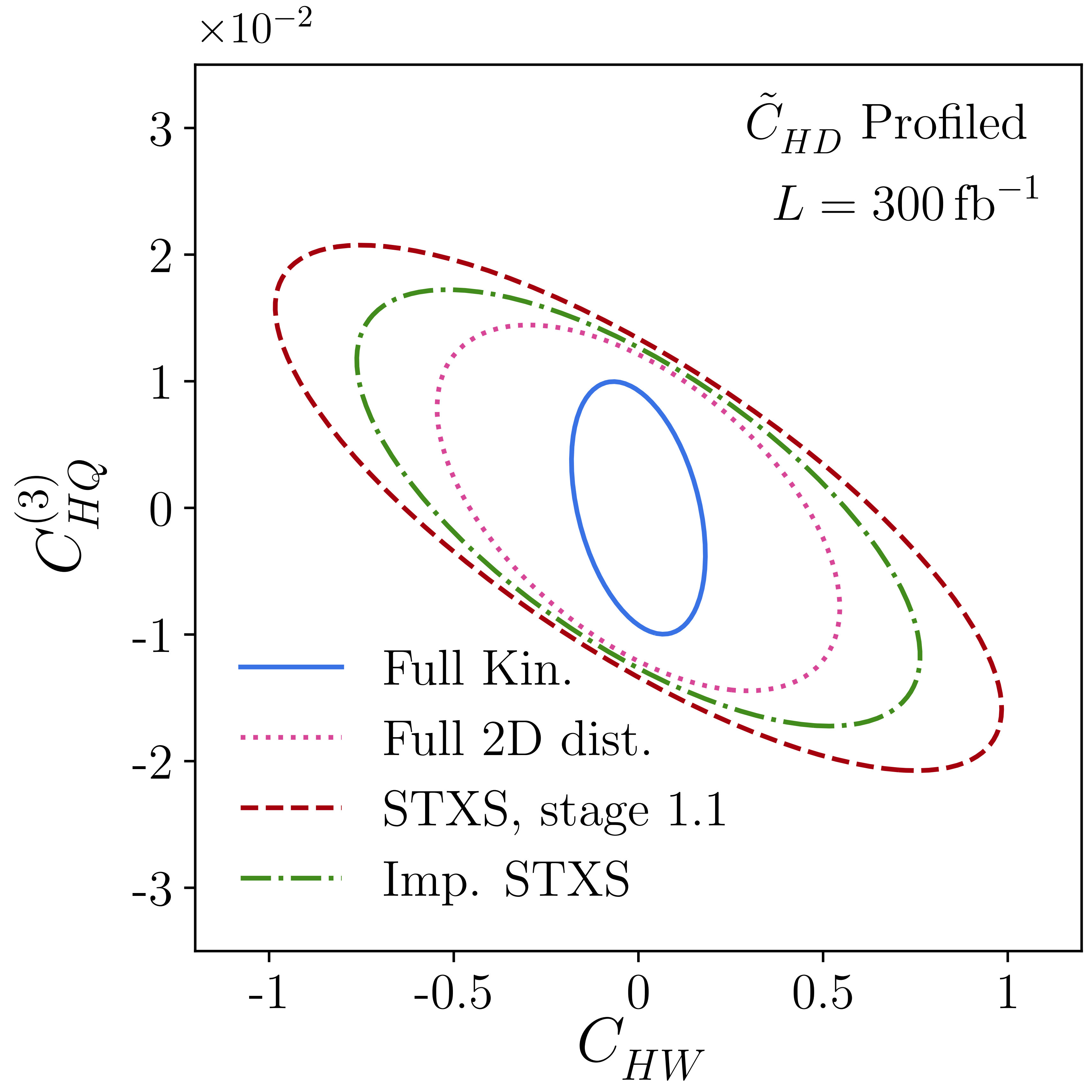



- Let's check! How much information on

$$\tilde{\mathcal{O}}_{HD} = \mathcal{O}_{H\Box} - \frac{\mathcal{O}_{HD}}{4} = (\phi^\dagger\phi)\Box(\phi^\dagger\phi) - \frac{1}{4}(\phi^\dagger D^\mu\phi)^*(\phi^\dagger D_\mu\phi)$$
$$\mathcal{O}_{HW} = \phi^\dagger\phi W^a_{\mu\nu}W^{\mu\nu a}$$
$$\mathcal{O}^{(3)}_{Hq} = (\phi^\dagger i \overleftrightarrow{D}^a_\mu \phi)(\overline{Q}_L\sigma^a\gamma^\mu Q_L) ,$$

can we extract from $pp \to WH \to \ell\nu\, b\bar{b}$ ?

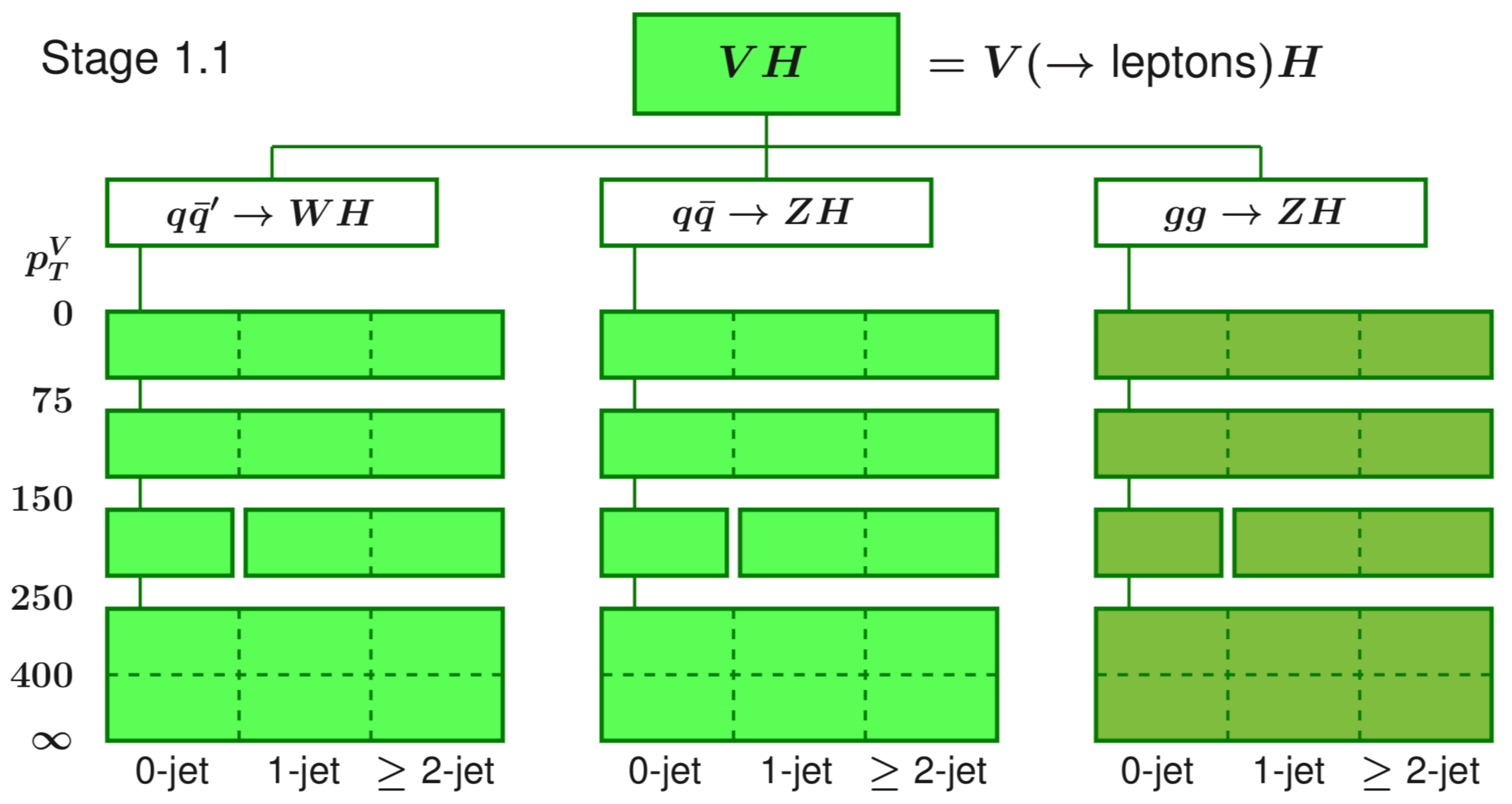
- Results: STXS are indeed sensitive to operators, adding a few more bins improve them, but a multivariate analysis is still stronger



# Benchmarking STXS in WH

[JB, S. Dawson, S. Homiller, F. Kling, T. Plehn 1908.06980]

- Simplified Template Cross-Sections (STXS) define observable bins that are supposed to capture as much information on NP as possible
  [N. Berger et al. 1906.02754; HXSWG YR4]

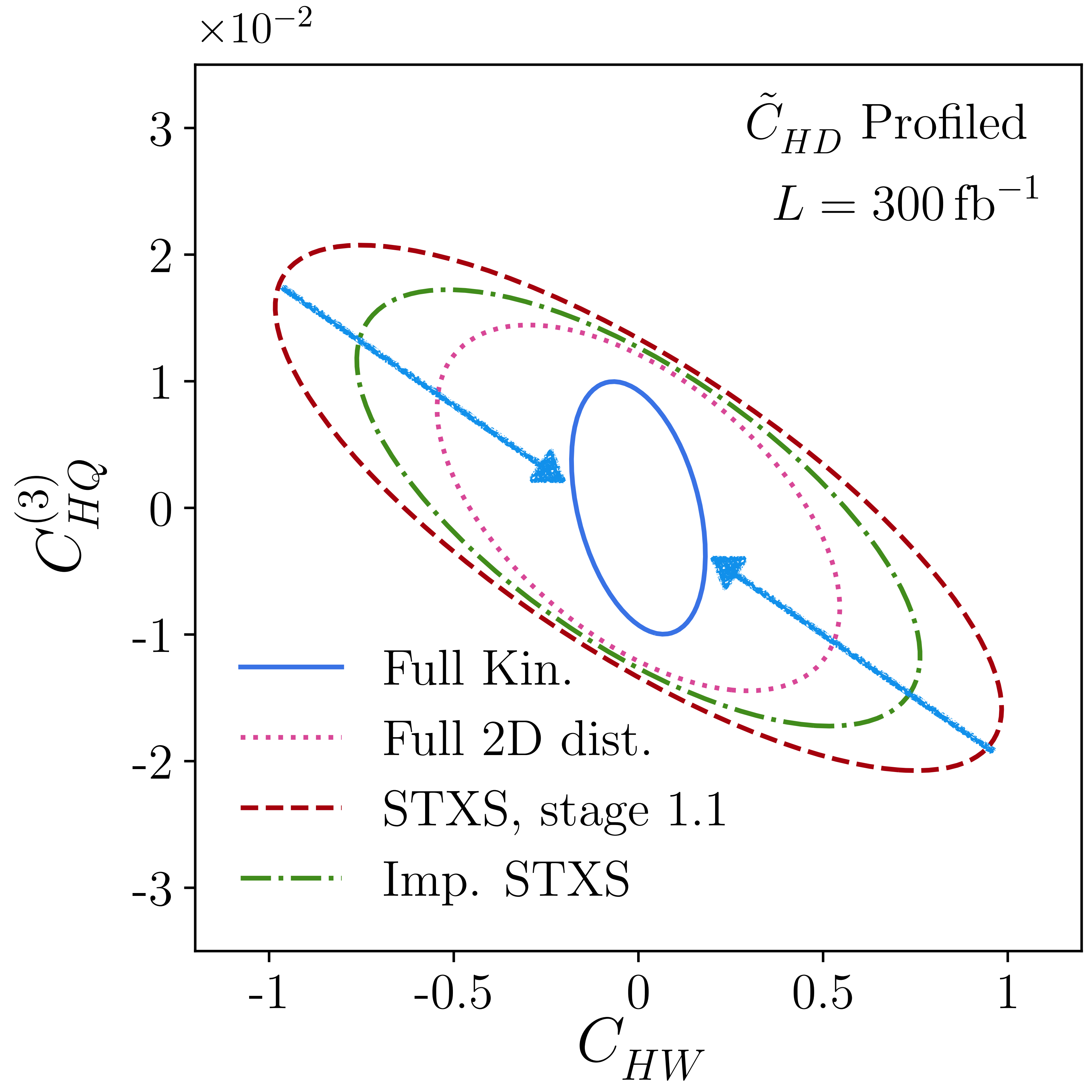



- Let's check! How much information on

$$\tilde{\mathcal{O}}_{HD} = \mathcal{O}_{H\Box} - \frac{\mathcal{O}_{HD}}{4} = (\phi^\dagger\phi)\Box(\phi^\dagger\phi) - \frac{1}{4}(\phi^\dagger D^\mu\phi)^*(\phi^\dagger D_\mu\phi)$$
$$\mathcal{O}_{HW} = \phi^\dagger\phi W^a_{\mu\nu}W^{\mu\nu a}$$
$$\mathcal{O}^{(3)}_{Hq} = (\phi^\dagger i\overleftrightarrow{D}^a_\mu\phi)(\overline{Q}_L\sigma^a\gamma^\mu Q_L)\,,$$

can we extract from $pp \to WH \to \ell\nu\, b\bar{b}$ ?

- Results: STXS are indeed sensitive to operators, adding a few more bins improve them, but a multivariate analysis is still stronger



# Diboson production

- In inclusive observables, the interference between SM and new physics amplitudes vanishes

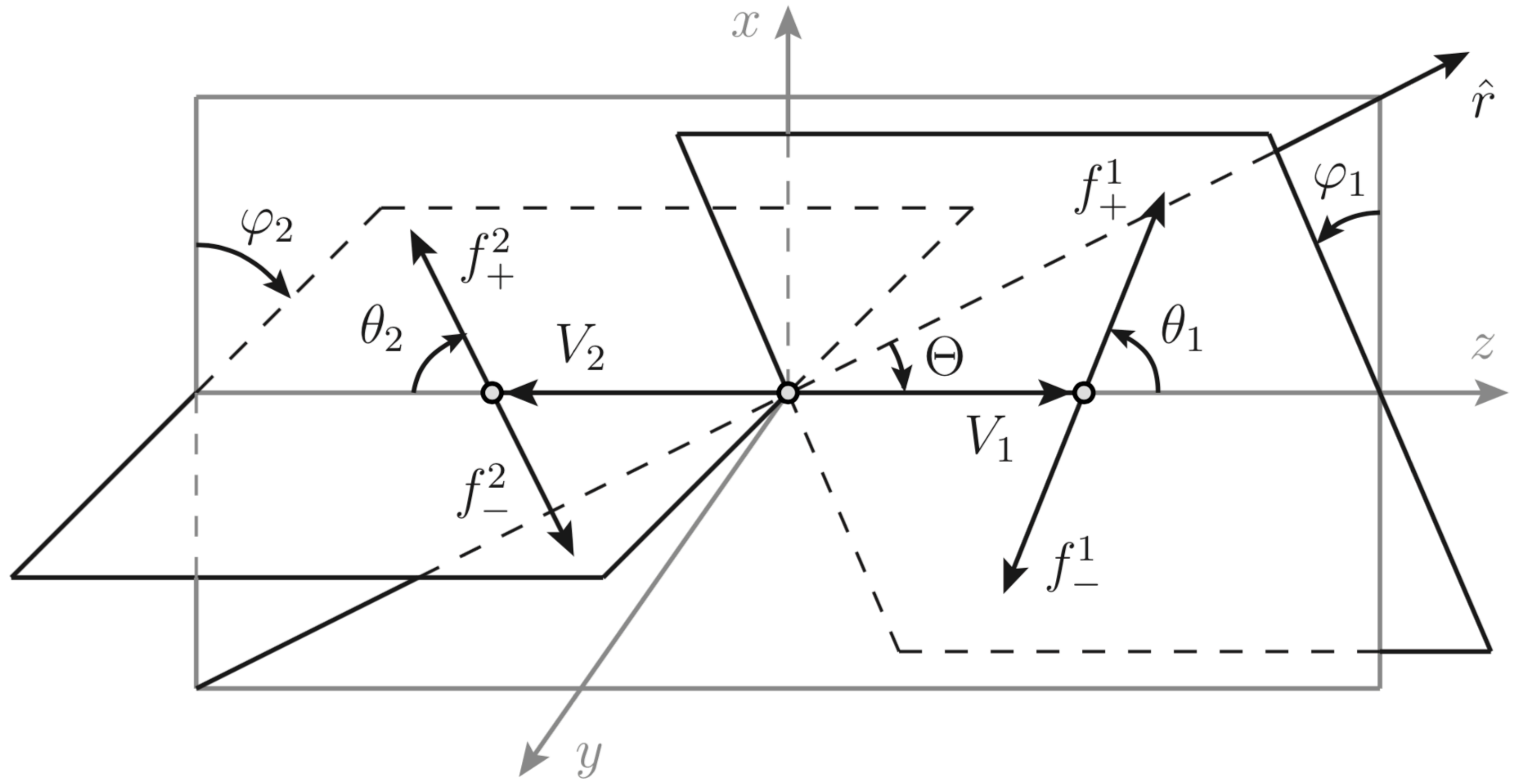
  ⇒ Reduced sensitivity to new physics

- "Diboson interference resurrection":
  an angular variable $\varphi$ can be constructed to be sensitive to this interference
  [G. Panico, F. Riva, A. Wulzer 1708.07823;
  A. Azatov, D. Barducci, E. Venturini 1901.04821]

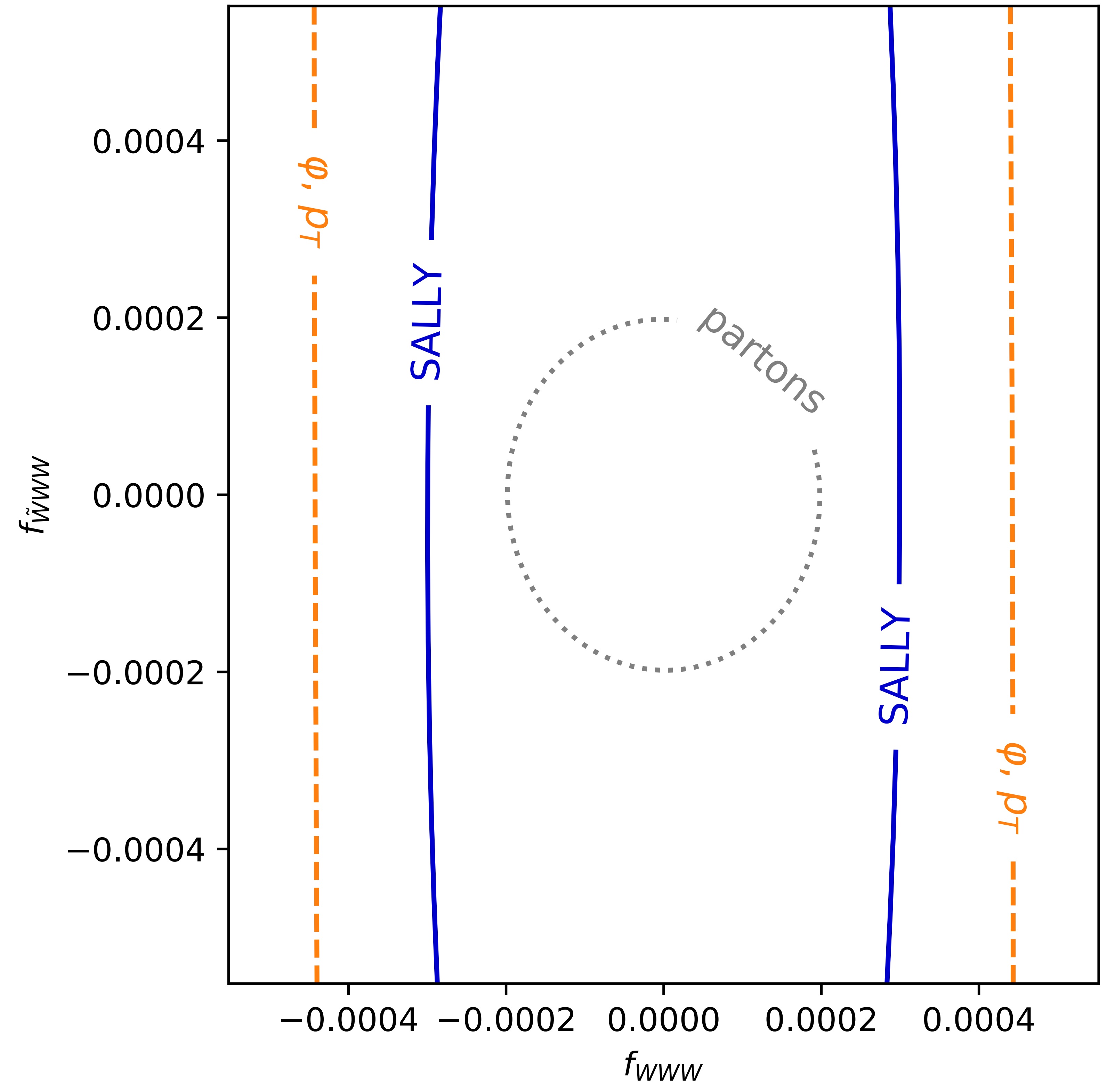
- We test the ML approach in EFT measurements in $W\gamma \to \ell\nu\ \gamma$
  [JB, K. Cranmer, M. Farina, F. Kling, D. Pappadopulo, J. Ruderman in progress]

  **New**: WZ→ ℓℓ ℓν by Chen, Glioti, Panico, Wulzer arXiv:2007.10356

- Preliminary results: we can extract more information when we analyze events with SALLY than with histograms of $\varphi$ and standard observables




Very preliminary results



# Conclusion

Likelihood fits in the data space are the gold standard for statistical inference

- RECAST and likelihood publishing are technical solutions that address model dependence and the theory-experiment interface
- STXS a good step, but more differential information can lead to large gain in sensitivity

Properties we want

- Ability to be fully differential
- Exploit highest fidelity simulation (QCD, detector simulation) without approximations that introduce additional systematic errors
- Clear statistical motivation and compatibility with traditional combined analyses
- Scalability in terms of channels and parameters

The approach I presented (implemented in MadMiner) achieves these goals

# References


**Opinionated review**
K. Cranmer, JB, G. Louppe:
“The frontier of simulation-based inference”
[1911.01429]

**Do It Yourself (for LHC physics)**
JB, F. Kling, I. Espejo, K. Cranmer:
“MadMiner: Machine learning—based inference for particle physics”
[CSBS, 1907.10621, https://github.com/diana-hep/madminer]

**LHC HXSWG YR4 STXS**
JB, S. Dawson, S. Homiller, F. Kling, T. Plehn:
“Benchmarking simplified template cross sections in WH production”
[JHEP, 1908.06980]

**Use in Astro: Strong lensing**
JB, S. Mishra-Sharma, J. Hermans, G. Louppe, K. Cranmer
“Mining for Dark Matter Substructure: Inferring subhalo population properties from strong lenses with machine learning”
[ApJ, 1909.02005]

**Original works**

JB, K. Cranmer, G. Louppe, J. Pavez:
“A guide to constraining Effective Field Theories with machine learning”
[PRD, 1805.00020]

JB, G. Louppe, J. Pavez, K. Cranmer:
“Mining gold from implicit models to improve likelihood-free inference”
[PNAS, 1805.12244]

**Follow-up with incremental improvements**
M. Stoye, JB, K. Cranmer, G. Louppe, J. Pavez:
“Likelihood-free inference with an improved cross-entropy estimator”
[NeurIPS workshop, 1808.00973]

# An incomplete wrap-up of simulation-based inference methods

| Method | Approximations | Upfront cost | Eval |
|---|---|---|---|
| Summary statistics: | | | |
| Likelihood for summary stats (standard histograms) | Reduction to summary stats | Fast | Fast |
| Approximate Bayesian Computation | Reduction to summary stats | Depends | Depends |
| Matrix elements: | | | |
| Matrix Element Method | Transfer fns | Fast | Slow |
| Optimal Observables | Transfer fns, optimal only locally | Fast | Slow |
| Neural networks: | | | |
| Neural likelihood | NN | Needs many samples | Fast |
| Neural posterior | NN | Needs many samples | Fast |
| Neural likelihood ratio | NN | Needs many samples | Fast |
| Neural networks + matrix elements: | | | |
| Neural likelihood (ratio) + gold mining (RASCAL etc) | NN | Needs less samples | Fast |
| Neural optimal observables (SALLY) | NN, optimal only locally | Needs less samples | Fast |

# Mining gold: A family of new inference techniques

| Method | Simulate | Extract $r(x,z)$ | Extract $t(x,z)$ | NN estimates | Asympt. exact | Generative |
|---|---|---|---|---|---|---|
| ROLR | $\theta_0 \sim \pi(\theta), \theta_1$ | ✓ | | $\hat{r}(x\|\theta_0, \theta_1)$ | ✓ | |
| CASCAL | $\theta_0 \sim \pi(\theta), \theta_1$ | | ✓ | $\hat{r}(x\|\theta_0, \theta_1)$ | ✓ | |
| ALICE | $\theta_0 \sim \pi(\theta), \theta_1$ | | ✓ | $\hat{r}(x\|\theta_0, \theta_1)$ | ✓ | |
| RASCAL | $\theta_0 \sim \pi(\theta), \theta_1$ | ✓ | ✓ | $\hat{r}(x\|\theta_0, \theta_1)$ | ✓ | |
| ALICES | $\theta_0 \sim \pi(\theta), \theta_1$ | ✓ | ✓ | $\hat{r}(x\|\theta_0, \theta_1)$ | ✓ | |
| SCANDAL | $\theta \sim \pi(\theta)$ | | ✓ | $\hat{p}(x\|\theta)$ | ✓ | ✓ |
| SALLY | $\theta_{\text{ref}}$ | | ✓ | $\hat{t}(x\|\theta_{\text{ref}})$ | in local approx. | |
| SALLINO | $\theta_{\text{ref}}$ | | ✓ | $\hat{t}(x\|\theta_{\text{ref}})$ | in local approx. | |

# Mining gold: A family of new inference techniques

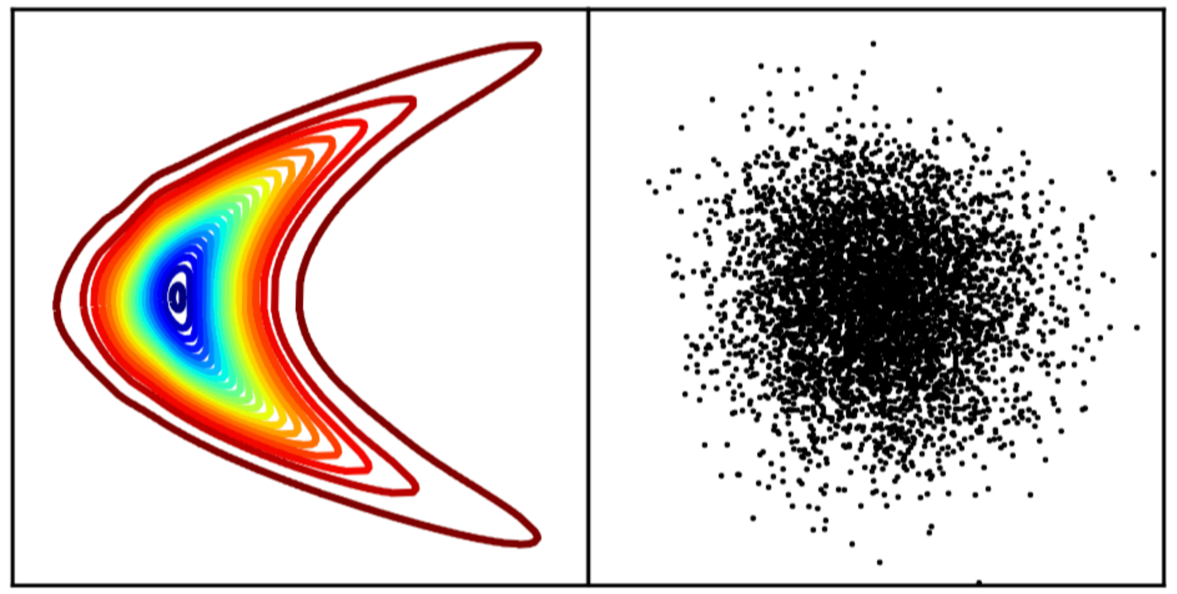
| Method | Simulate | Extract $r(x,z)$ | Extract $t(x,z)$ | NN estimates | Asympt. exact | Generative |
|---|---|---|---|---|---|---|
| ROLR | $\theta_0 \sim \pi(\theta), \theta_1$ | ✓ | | $\hat{r}(x\|\theta_0, \theta_1)$ | ✓ | |
| CASCAL | $\theta_0 \sim \pi(\theta), \theta_1$ | | ✓ | $\hat{r}(x\|\theta_0, \theta_1)$ | ✓ | |
| ALICE | $\theta_0 \sim \pi(\theta), \theta_1$ | | ✓ | $\hat{r}(x\|\theta_0, \theta_1)$ | ✓ | |
| RASCAL | $\theta_0 \sim \pi(\theta), \theta_1$ | ✓ | ✓ | $\hat{r}(x\|\theta_0, \theta_1)$ | ✓ | |
| ALICES | $\theta_0 \sim \pi(\theta), \theta_1$ | ✓ | ✓ | $\hat{r}(x\|\theta_0, \theta_1)$ | ✓ | |
| SCANDAL | $\theta \sim \pi(\theta)$ | | ✓ | $\hat{p}(x\|\theta)$ | ✓ | ✓ |
| SALLY | $\theta_{\text{ref}}$ | | ✓ | $\hat{t}(x\|\theta_{\text{ref}})$ | in local approx. | |
| SALLINO | $\theta_{\text{ref}}$ | | ✓ | $\hat{t}(x\|\theta_{\text{ref}})$ | in local approx. | |

Performance gains with cross-entropy-based loss

[M. Stoye, JB, K. Cranmer, G. Louppe, J. Pavez 1808.00973]

# Mining gold: A family of new inference techniques

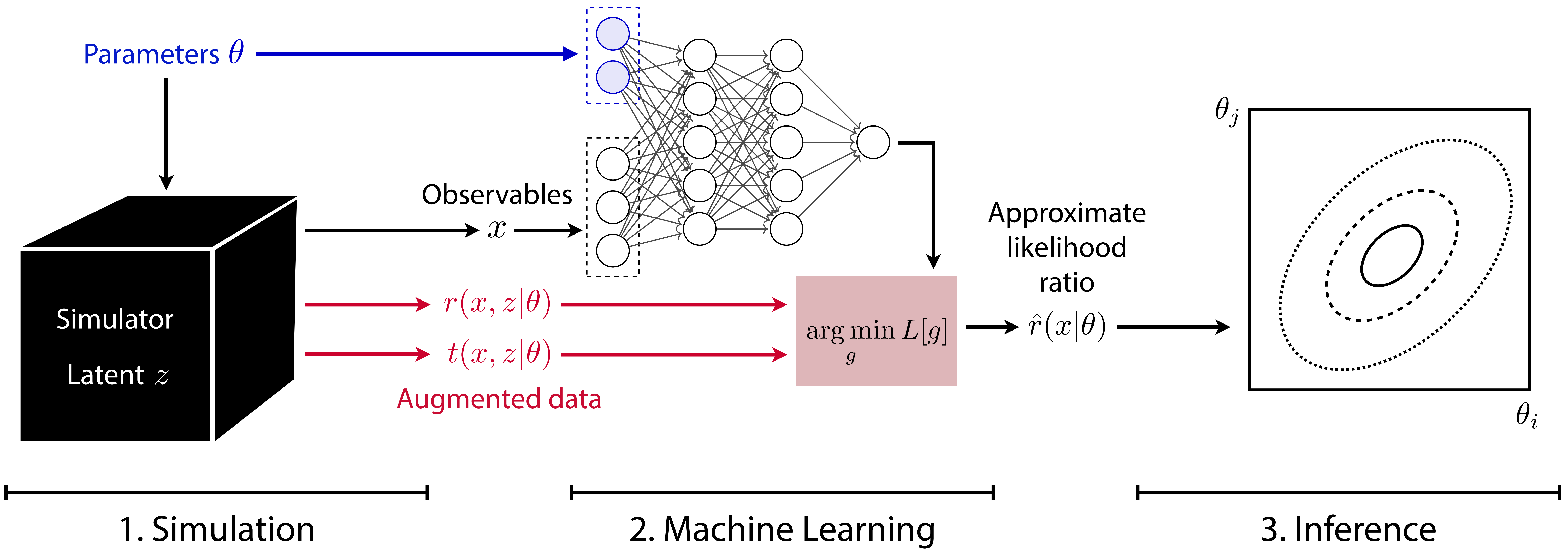
| Method | Simulate | Extract $r(x,z)$ | Extract $t(x,z)$ | NN estimates | Asympt. exact | Generative |
|---|---|---|---|---|---|---|
| ROLR | $\theta_0 \sim \pi(\theta), \theta_1$ | ✓ | | $\hat{r}(x\|\theta_0, \theta_1)$ | ✓ | |
| CASCAL | $\theta_0 \sim \pi(\theta), \theta_1$ | | ✓ | $\hat{r}(x\|\theta_0, \theta_1)$ | ✓ | |
| ALICE | $\theta_0 \sim \pi(\theta), \theta_1$ | | ✓ | $\hat{r}(x\|\theta_0, \theta_1)$ | ✓ | |
| RASCAL | $\theta_0 \sim \pi(\theta), \theta_1$ | ✓ | ✓ | $\hat{r}(x\|\theta_0, \theta_1)$ | ✓ | |
| ALICES | $\theta_0 \sim \pi(\theta), \theta_1$ | ✓ | ✓ | $\hat{r}(x\|\theta_0, \theta_1)$ | ✓ | |
| SCANDAL | $\theta \sim \pi(\theta)$ | | ✓ | $\hat{p}(x\|\theta)$ | ✓ | ✓ |
| SALLY | $\theta_{\text{ref}}$ | | ✓ | $\hat{t}(x\|\theta_{\text{ref}})$ | in local approx. | |
| SALLINO | $\theta_{\text{ref}}$ | | ✓ | $\hat{t}(x\|\theta_{\text{ref}})$ | in local approx. | |



Combination with state-of-the-art conditional neural density estimators, e.g. normalizing flows

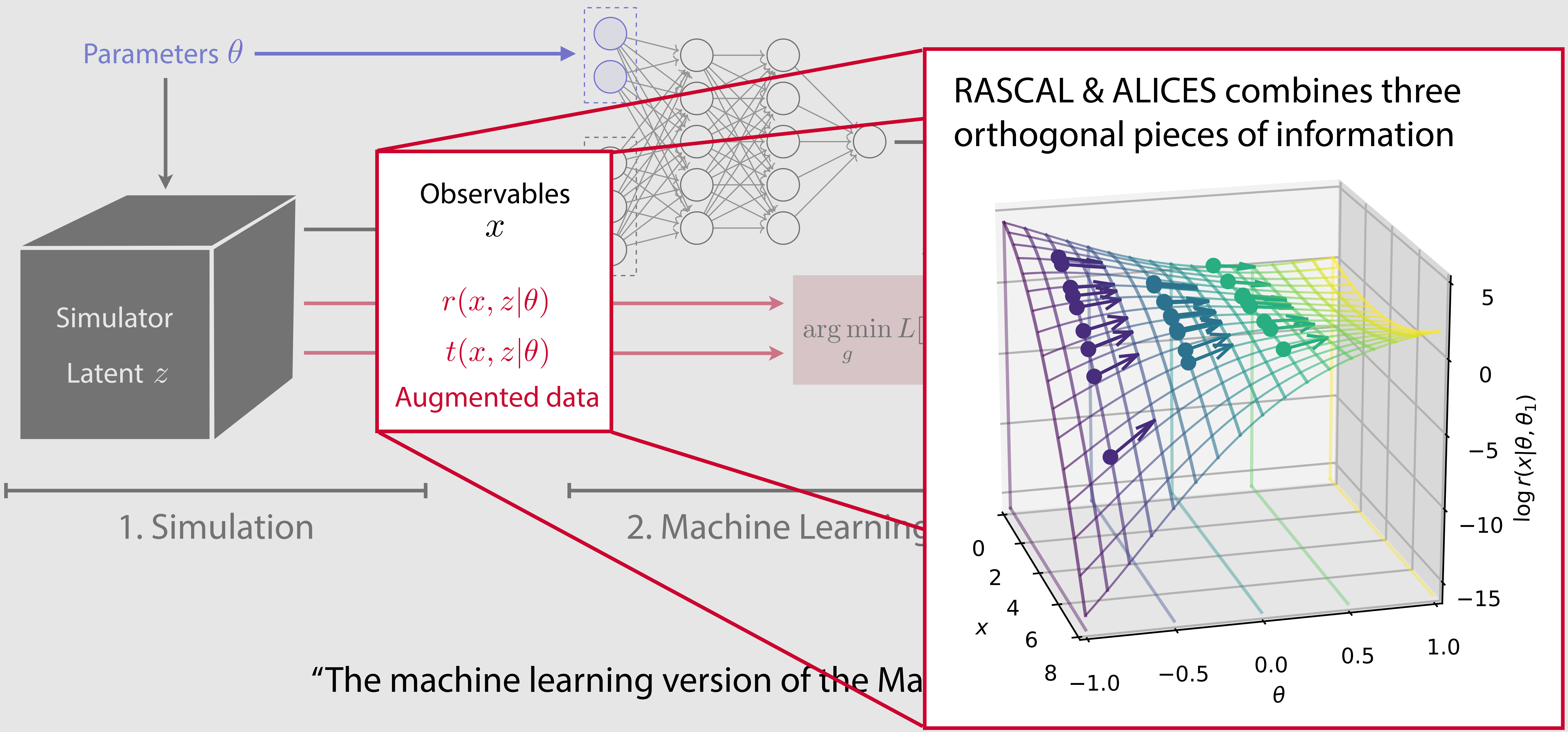
[everything by G. Papamakarios: G. Papamakarios, T. Pavlakou, I. Murray 1705.07057; G. Papamakarios, D. Sterratt, I. Murray 1805.07226; …]

# Putting the pieces together: RASCAL & ALICES




"The machine learning version of the Matrix Element Method"

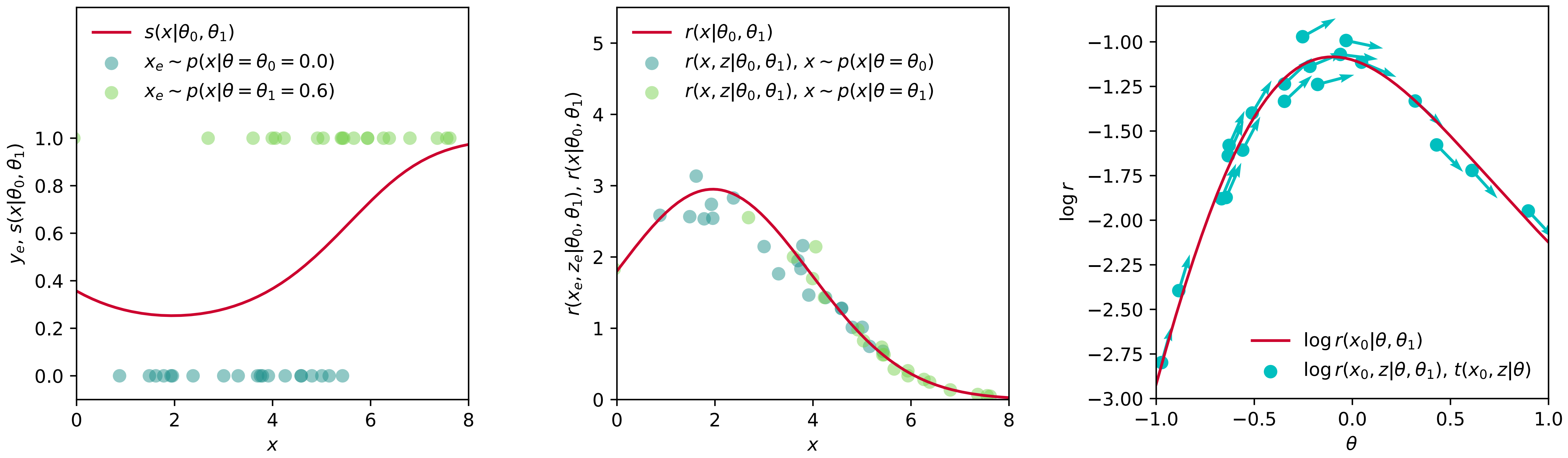
# Putting the pieces together: RASCAL & ALICES

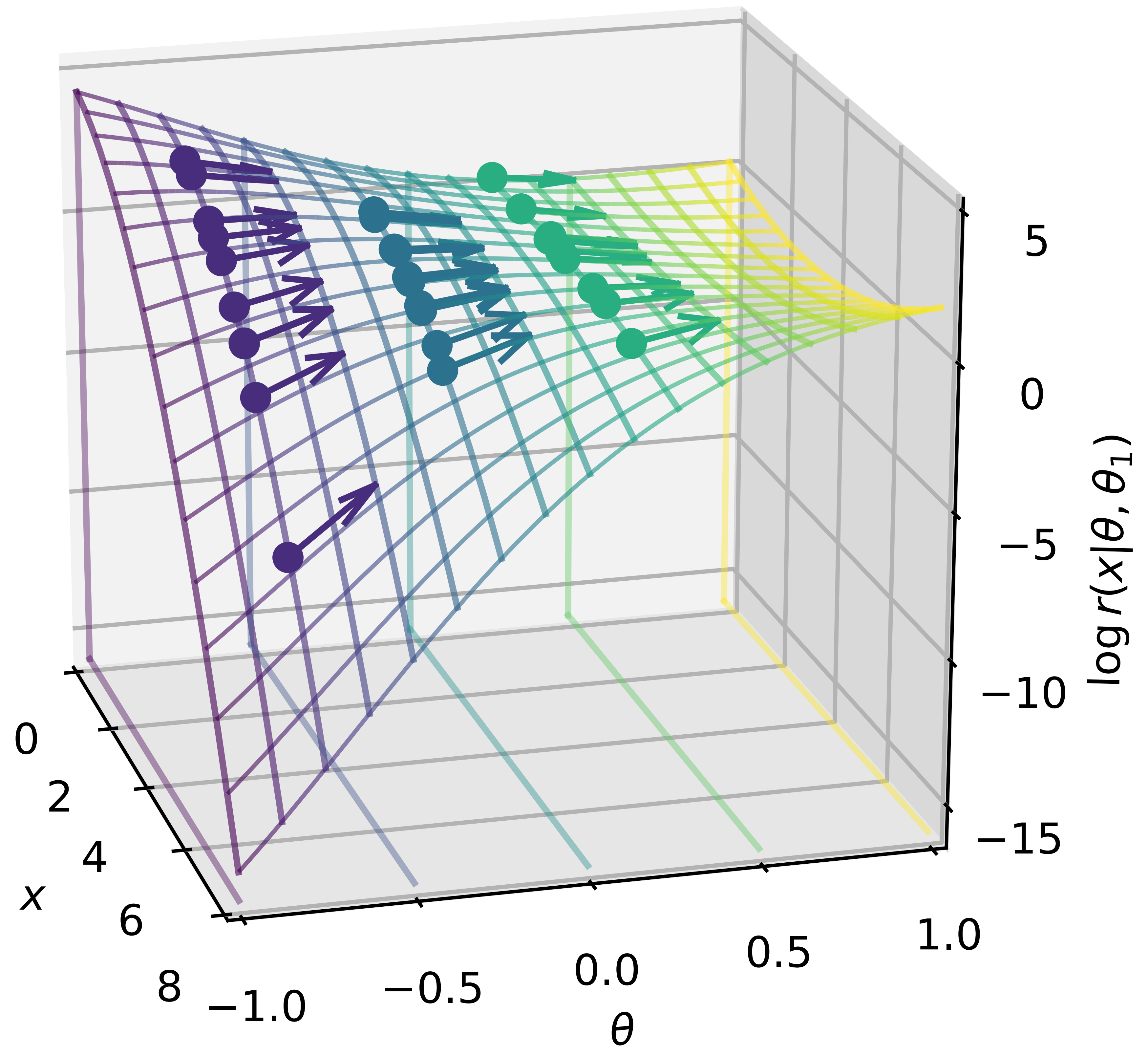


# Gold mining: augmenting the training data

The augmented training data converts supervised classification into supervised regression with lower variance

- improvement in training efficiency

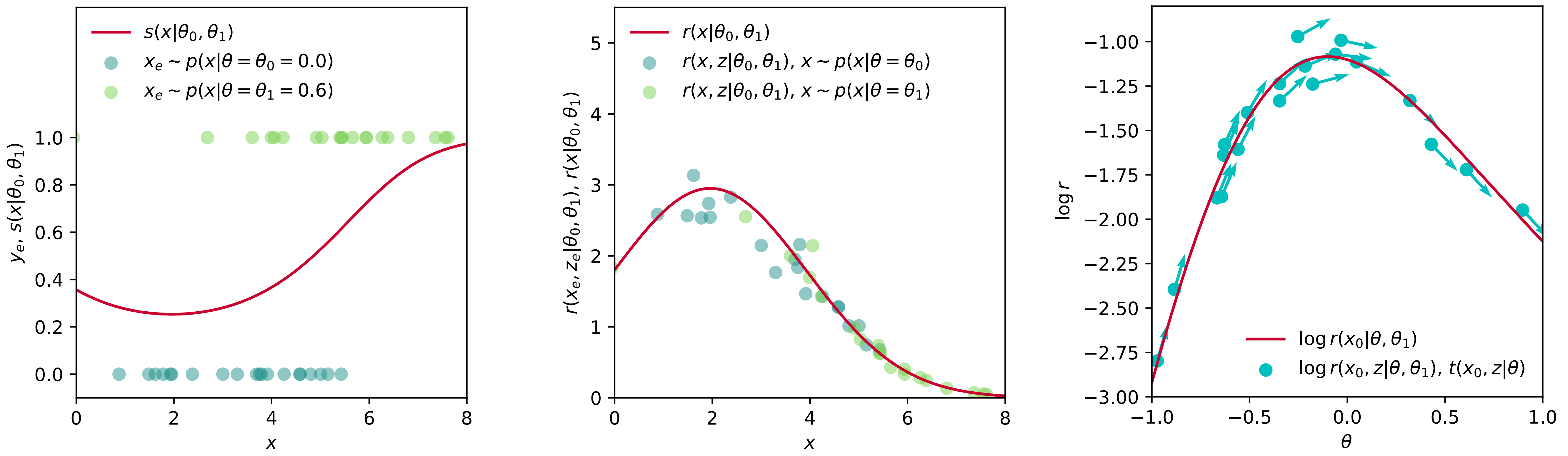


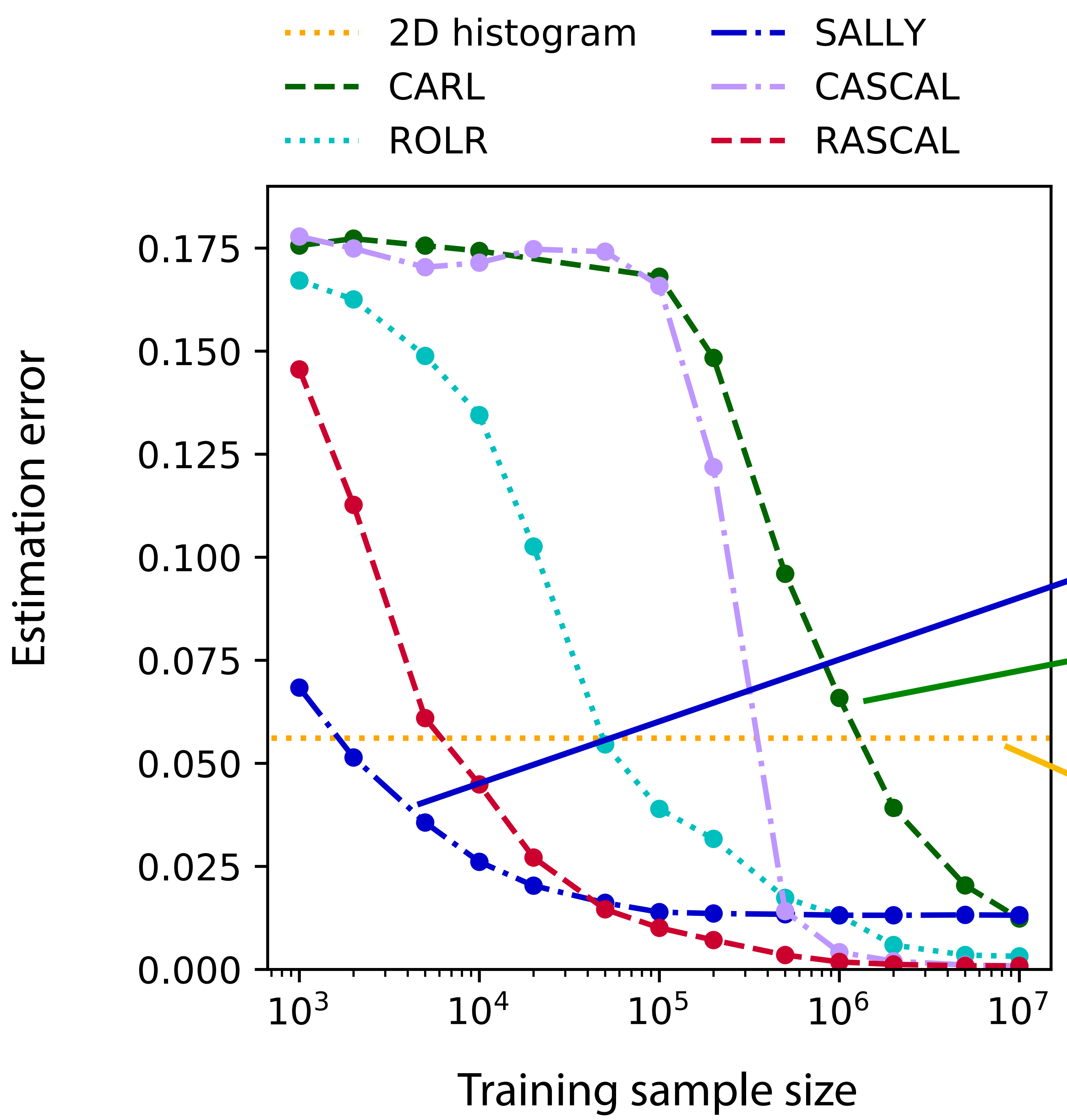



# Gold mining: augmenting the training data

The augmented training data converts supervised classification into supervised regression with lower variance

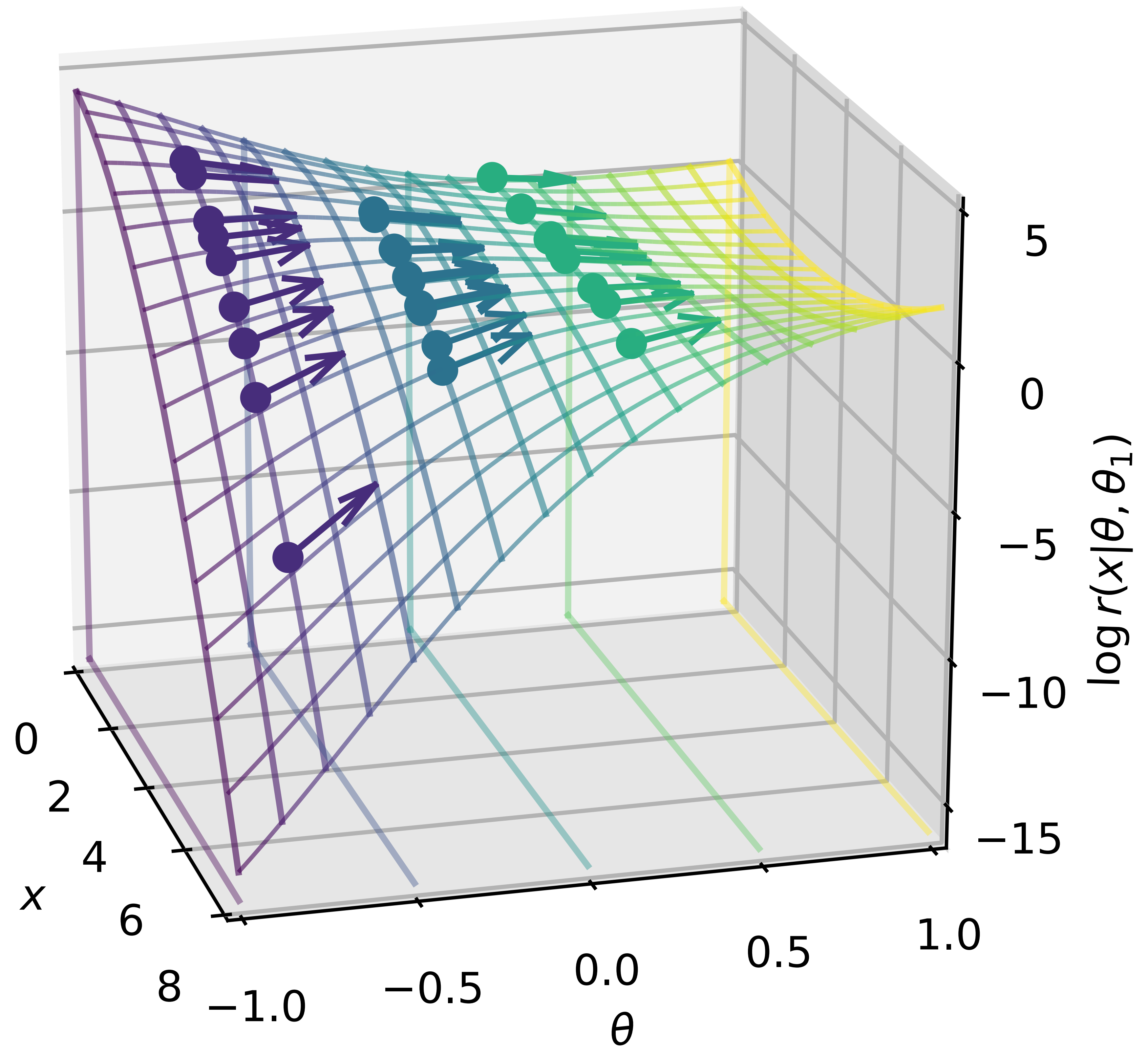
- improvement in training efficiency




New techniques require less data than without augmented data

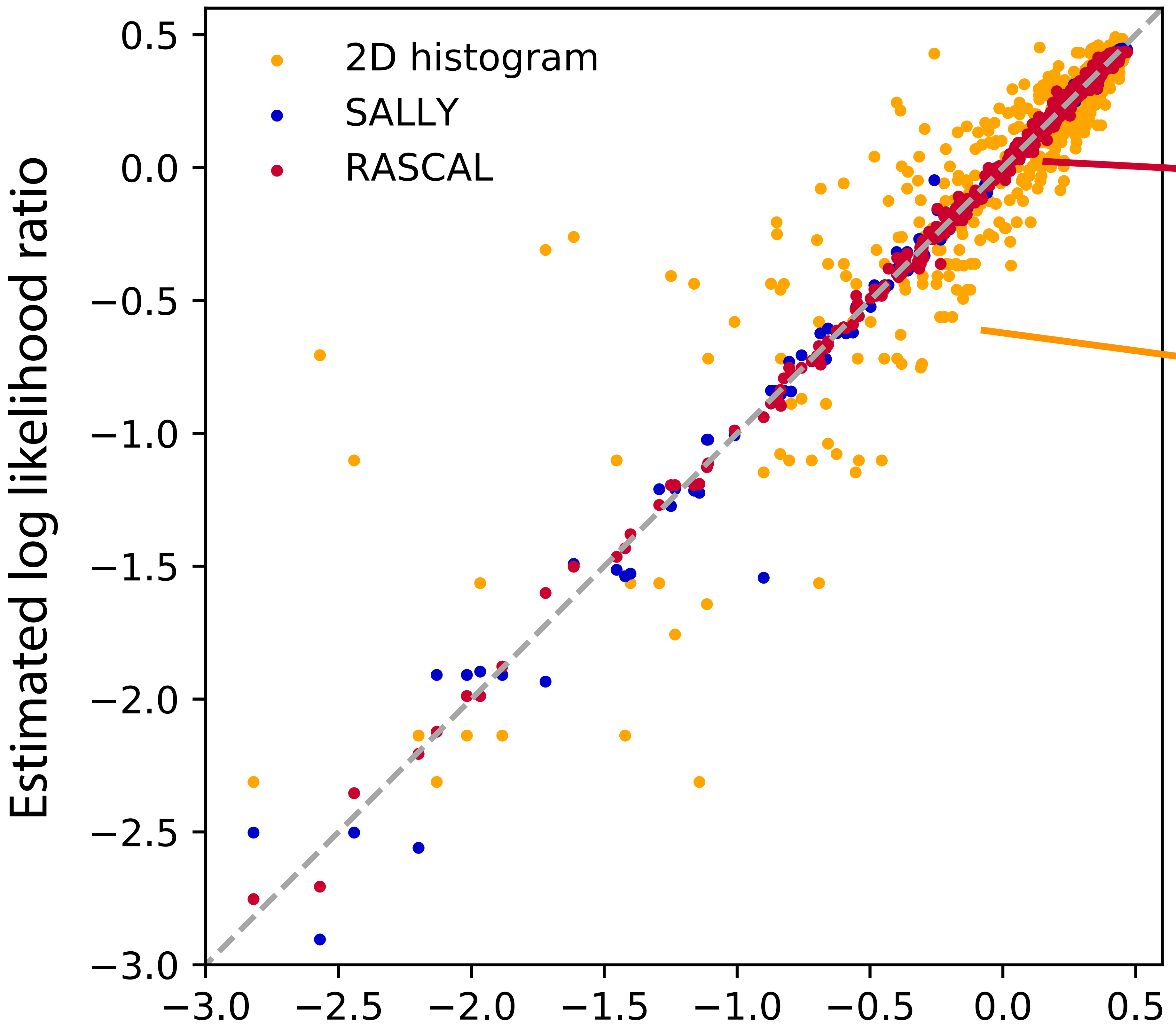
Traditional Approach no NN




Brehmer, Louppe, Pavez, KC, PNAS (2019), arXiv:1805.12244

# More precise likelihood ratio estimates with less training data

Each point = 1 event

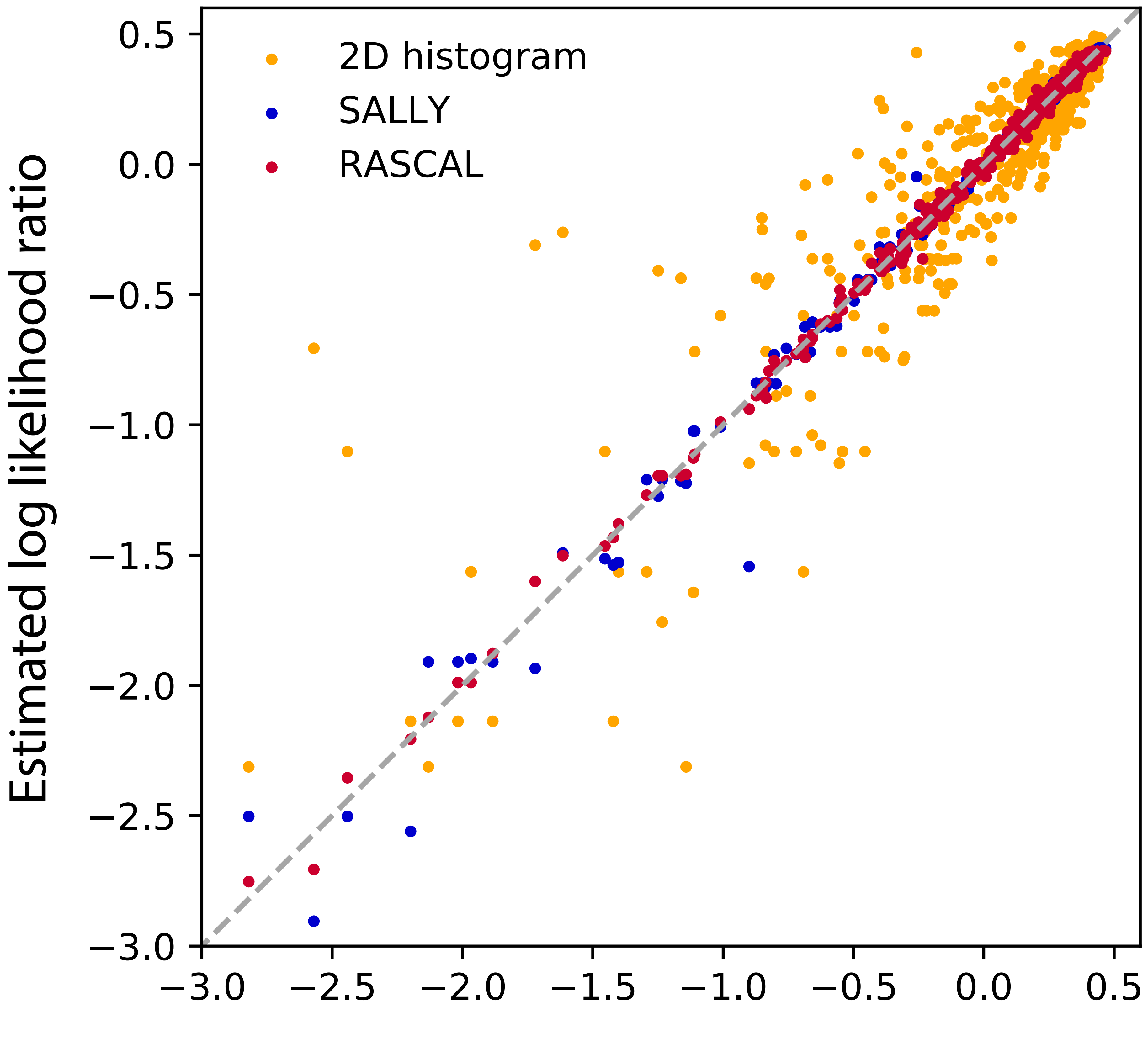



True log likelihood ratio
(usually don't have this)

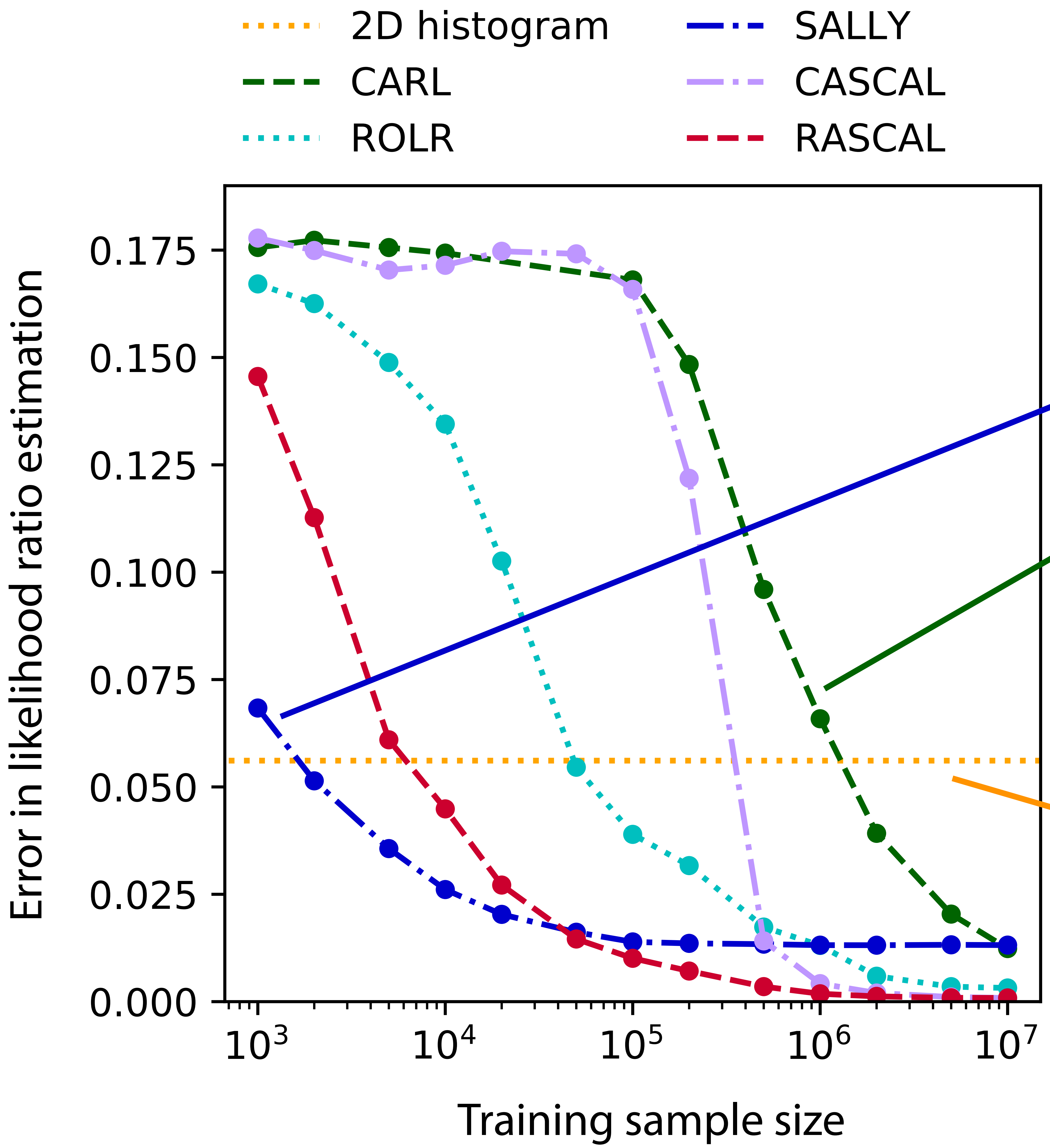
RASCAL gives more precise likelihood ratio estimates than the 2D histogram

# More precise likelihood ratio estimates with less training data

Each point = 1 event




(usually don't have this)



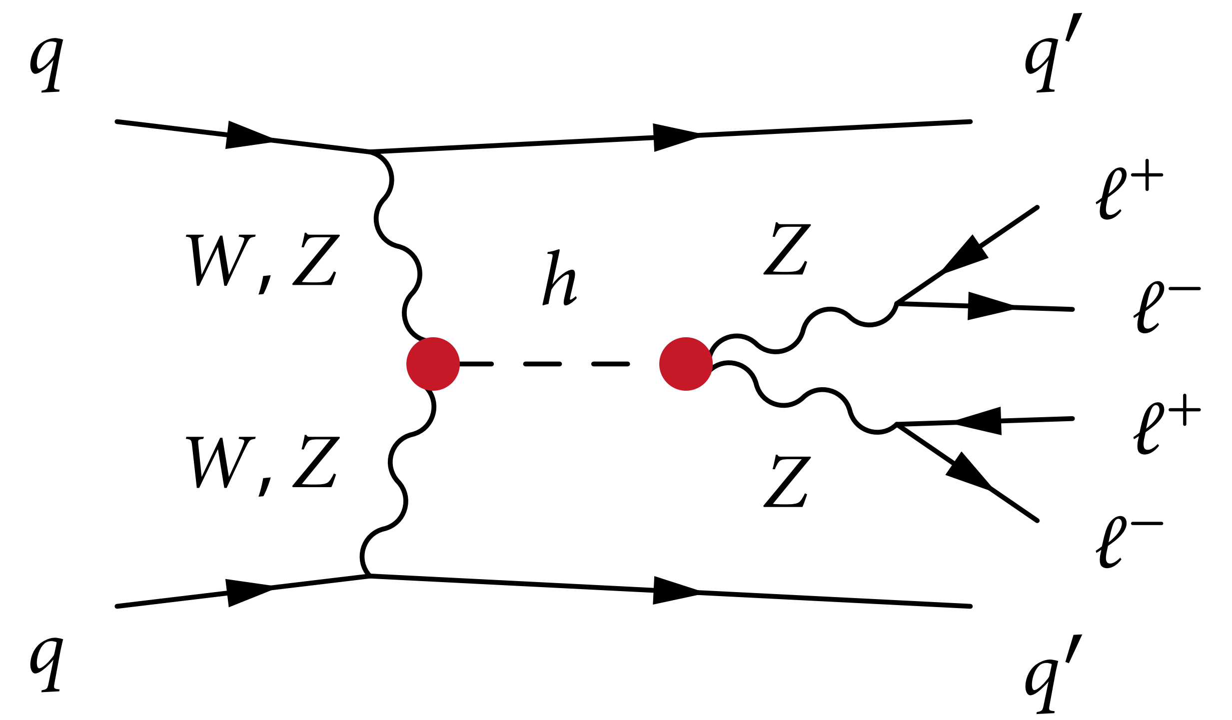

SALLY and RASCAL require less data than generic ML method

[CARL, K. Cranmer, G. Louppe, J. Pavez 1506.02169]

Baseline: 2D histogram (based on large training sample)

# Challenge for EFT

Let $\theta$ denote the coefficients of higher dimensional operators in the Lagrangian, $x$ be high-dimensional data associated to an event, and $p(x \mid \theta) = \frac{1}{\sigma(\theta)} \frac{d\sigma}{d\theta}$ be the distribution for the data



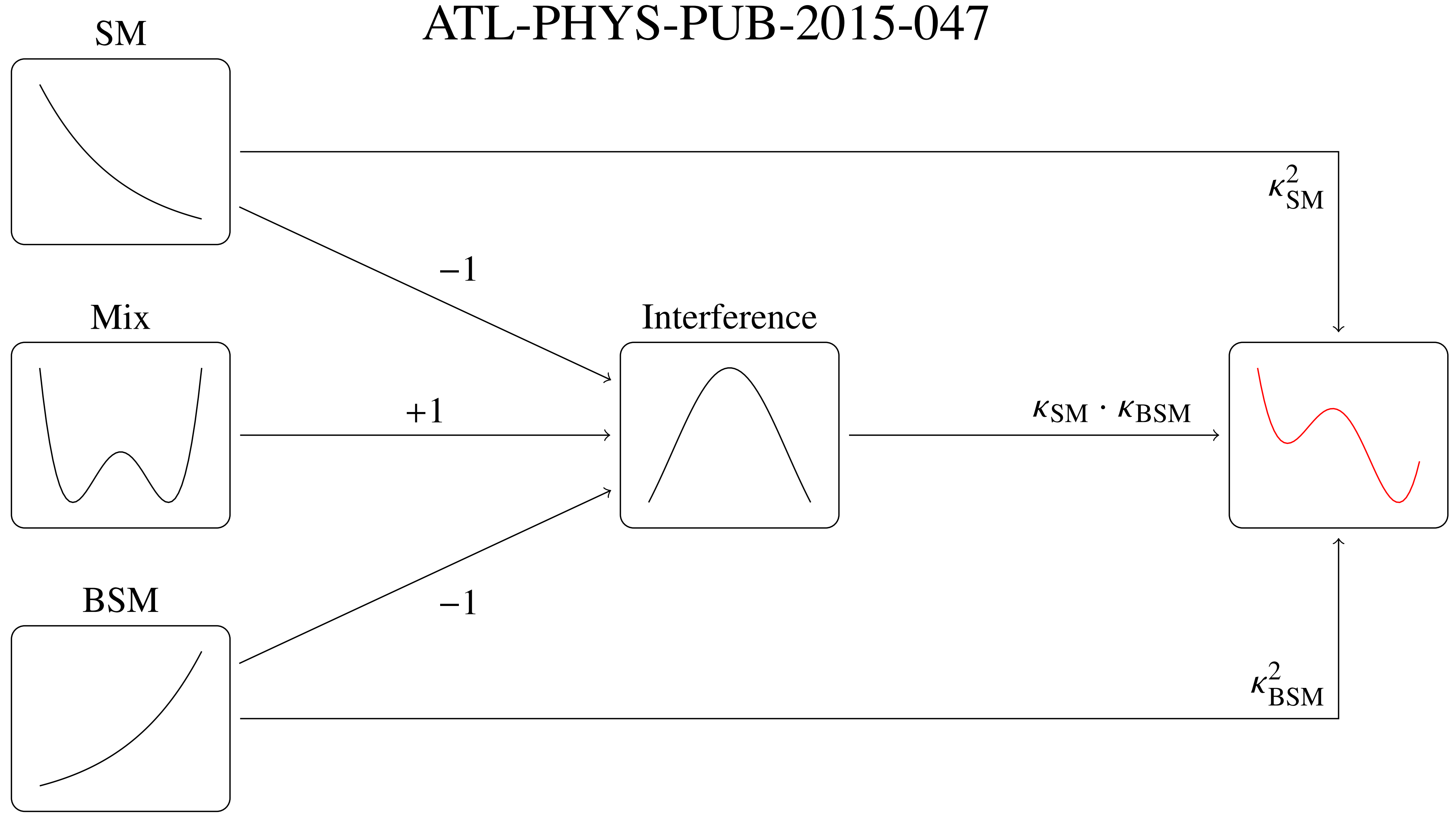

- we want to compare any two points in EFT parameter space
- evaluate the **likelihood ratio** $r(x|\theta_0, \theta_1) \equiv \frac{p(x|\theta_0)}{p(x|\theta_1)}$

Difficulty is that one changes the parameters of the EFT, the distributions p(x|θ) change due to interference.

- It would be very computationally expensive (infeasible) to generate samples for every value of θ and estimate $p(x \mid \theta)$ with histograms. Small changes mean we need a lot of MC events!
- Ideally we could directly estimate the **score** $t(x|\theta_0) \equiv \nabla_\theta \log p(x|\theta)\Big|_{\theta_0}$

# EFT Embedded in a vector space

ATL-PHYS-PUB-2015-047




Difficulty is that one changes the parameters of the EFT, the distributions p(x|θ) change due to interference.

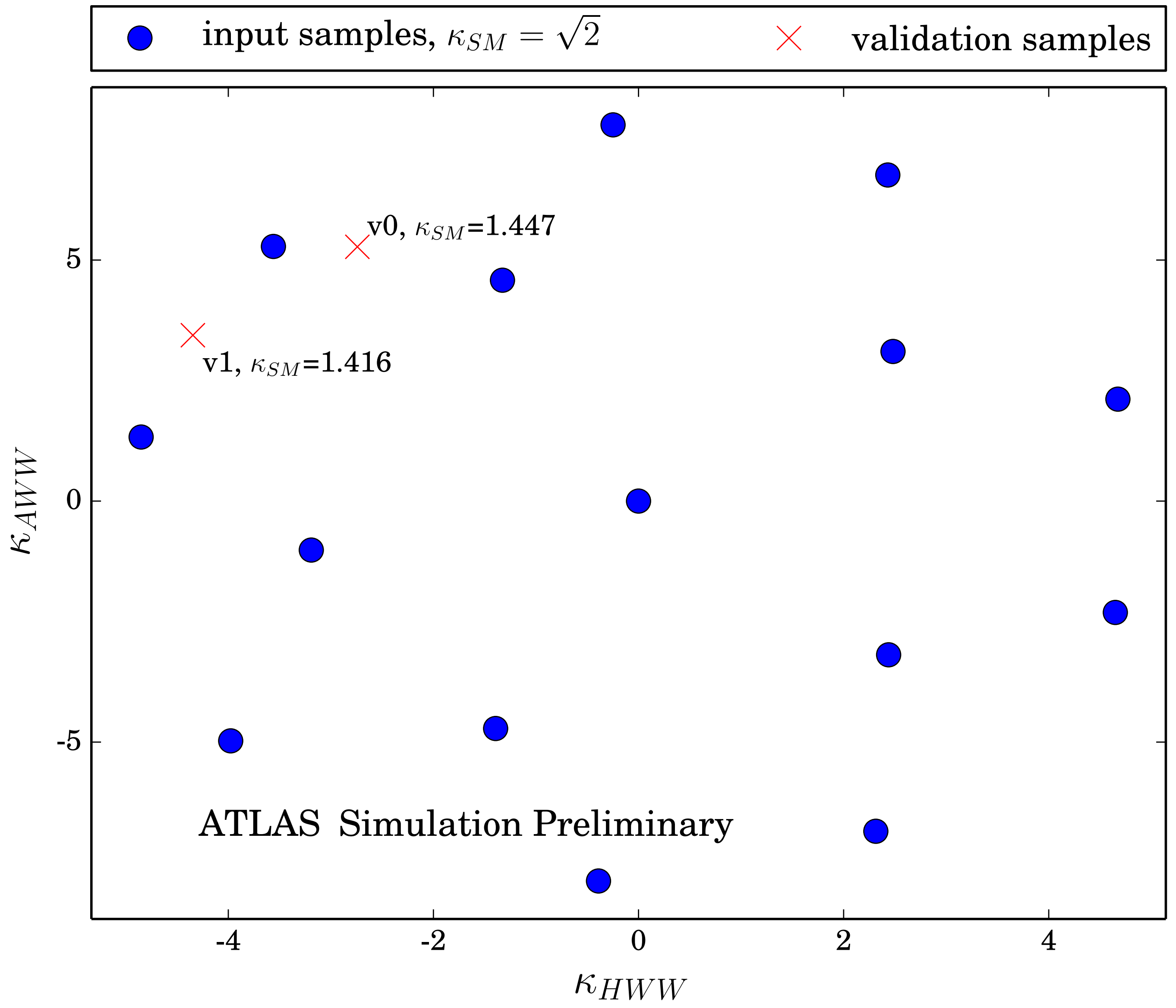
**But there is a trick:**

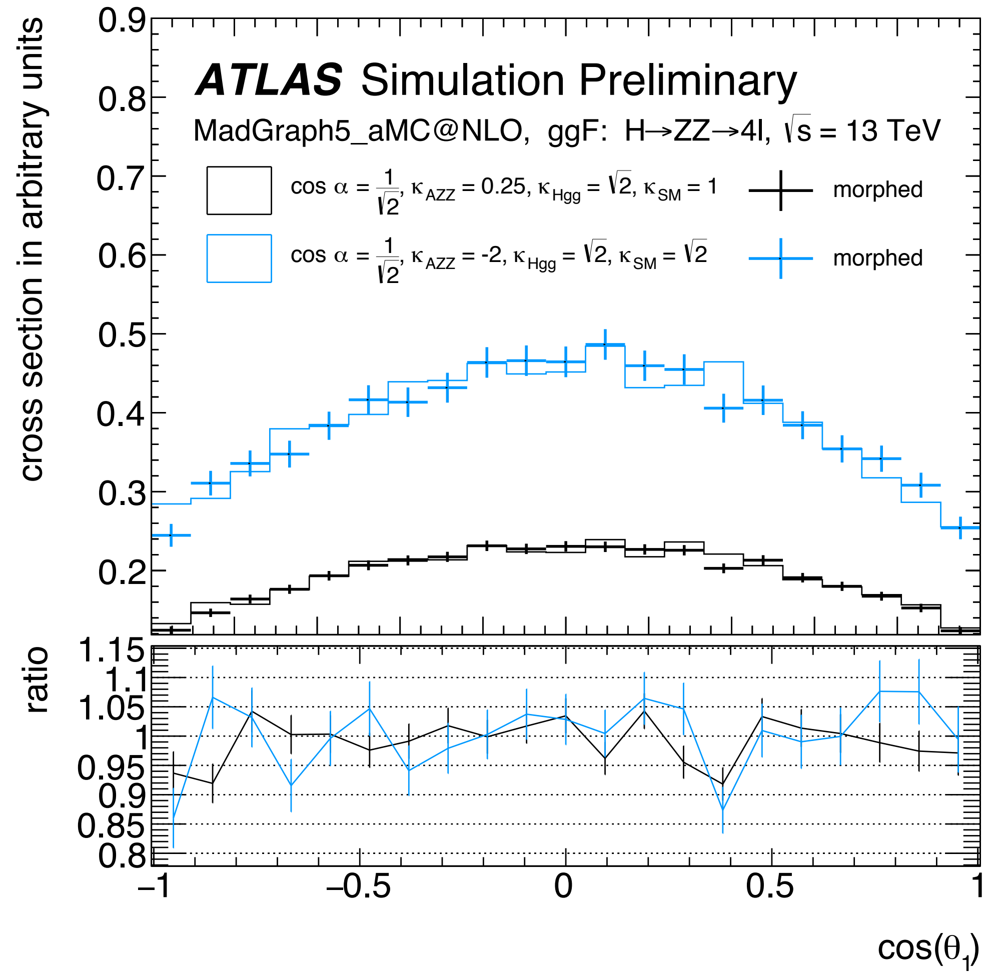
Simple example:

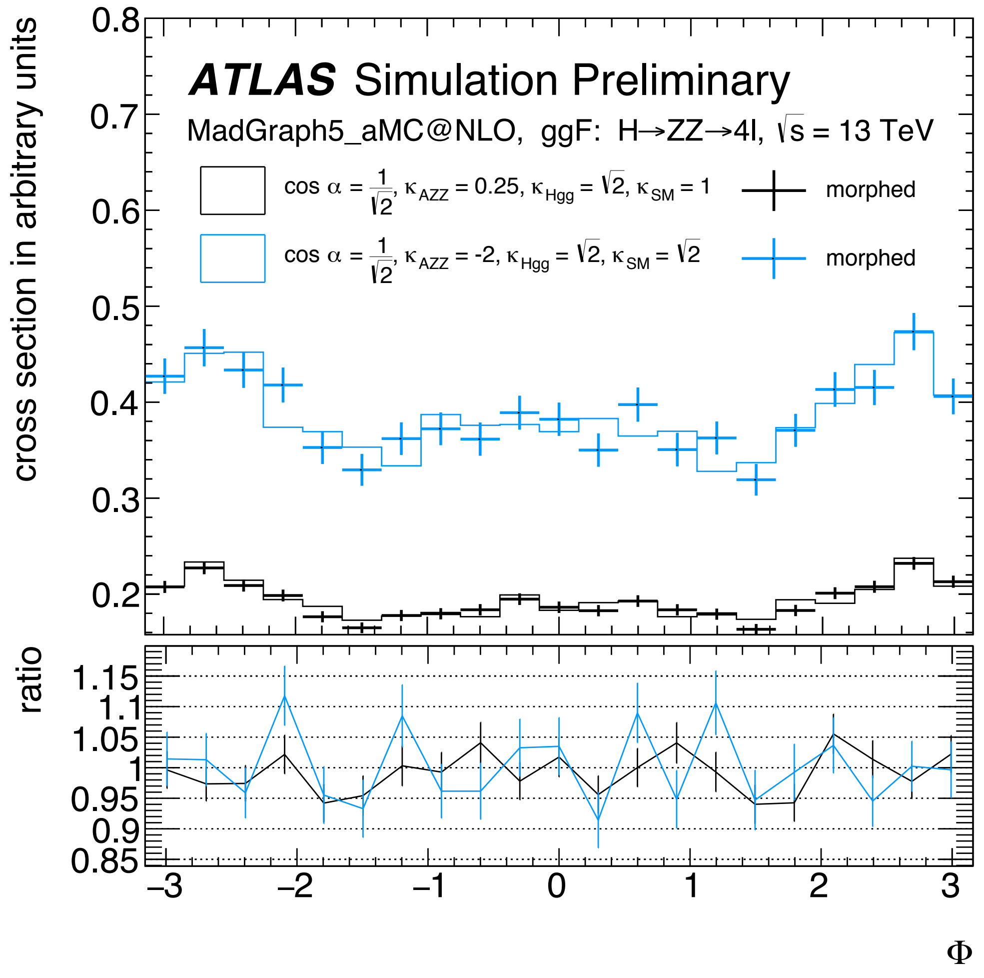
$$|g_1 M_{SM} + g_2 M_{BSM}|^2 = g_1^2 |M_{SM}|^2 + 2 g_1 g_2 Re\left[M^*_{SM} M_{BSM}\right] + g_2^2 |M_{BSM}|^2$$

3-d vector space, distribution for any point in this space is linear mixture of distribution for 3 basis samples!

(real examples need more basis samples)



# EFT Decomposition

$$d\sigma \propto \left| \underbrace{\left( \mathcal{M}^p_{\text{SM}} + \sum_i \frac{f_i}{\Lambda^2} \mathcal{M}^p_i \right)}_{\text{production}} \underbrace{\left( \mathcal{M}^d_{\text{SM}} + \sum_j \frac{f_j}{\Lambda^2} \mathcal{M}^d_j \right)}_{\text{decay}} \right|^2$$

Express EFT as a mixture:

$$p(x|\theta) = \sum_c w_c(\theta) p_c(x)$$

$w_c(\theta)$ are polynomials

$\nabla_\theta \log p(x|\theta)$ is now possible!

| Process | Number of components for $n$ operators | | | | | |
|---|---|---|---|---|---|---|
| | $\mathcal{O}(\Lambda^0)$ | $\mathcal{O}(\Lambda^{-2})$ | $\mathcal{O}(\Lambda^{-4})$ | $\mathcal{O}(\Lambda^{-6})$ | $\mathcal{O}(\Lambda^{-8})$ | $\sum$ |
| $hV$ / WBF production | 1 | $n$ | $\frac{n(n+1)}{2}$ | | | $\frac{(n+1)(n+2)}{2}$ |
| $h \to VV$ decay | 1 | $n$ | $\frac{n(n+1)}{2}$ | | | $\frac{(n+1)(n+2)}{2}$ |
| Production + decay | 1 | $n$ | $\frac{n(n+1)}{2}$ | $\binom{n+2}{3}$ | $\binom{n+3}{4}$ | $\binom{n+4}{4}$ |

Table 1: Number of components $c$ as given in Eq. (6) for different processes, sorted by their suppression by the EFT cutoff scale $\Lambda$.

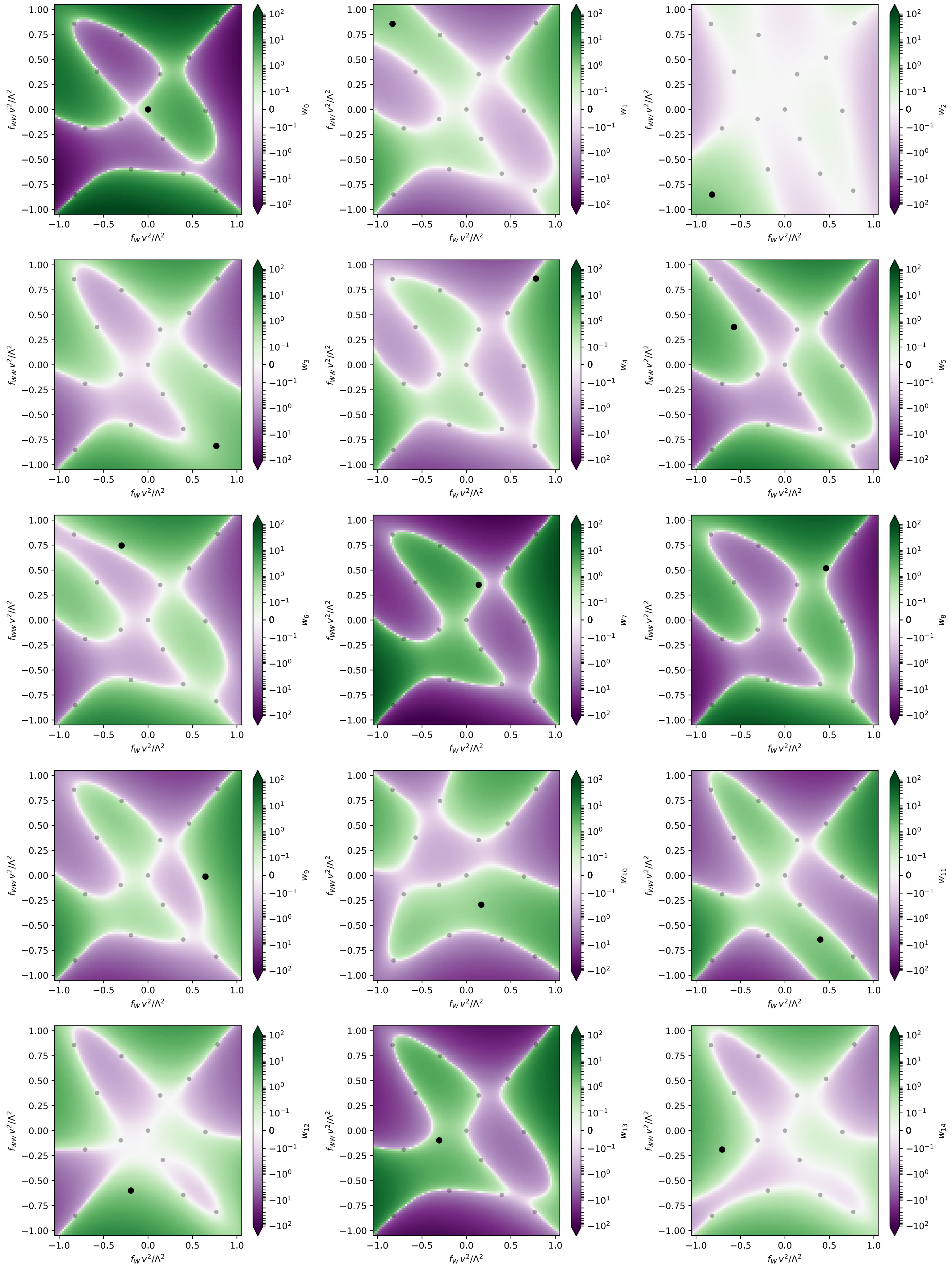
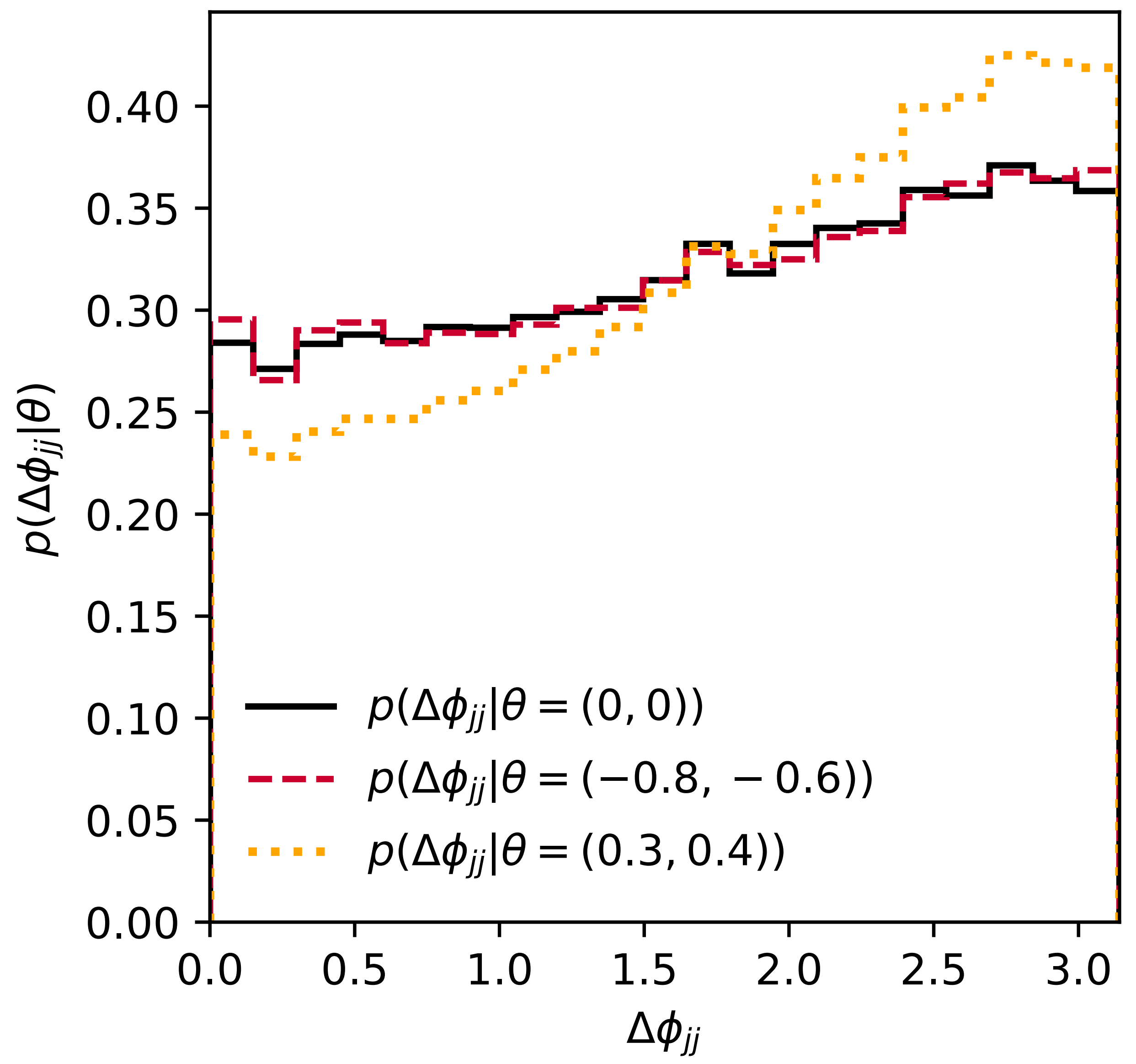


Figure 13: Morphing weights $w_i(\theta)$ for basis points distributed over the full relevant parameter space.

For 2 BSM operators affecting VBF Higgs production and decay, we need a 15-D vector space

For 5 BSM operators we need 126-D vector space

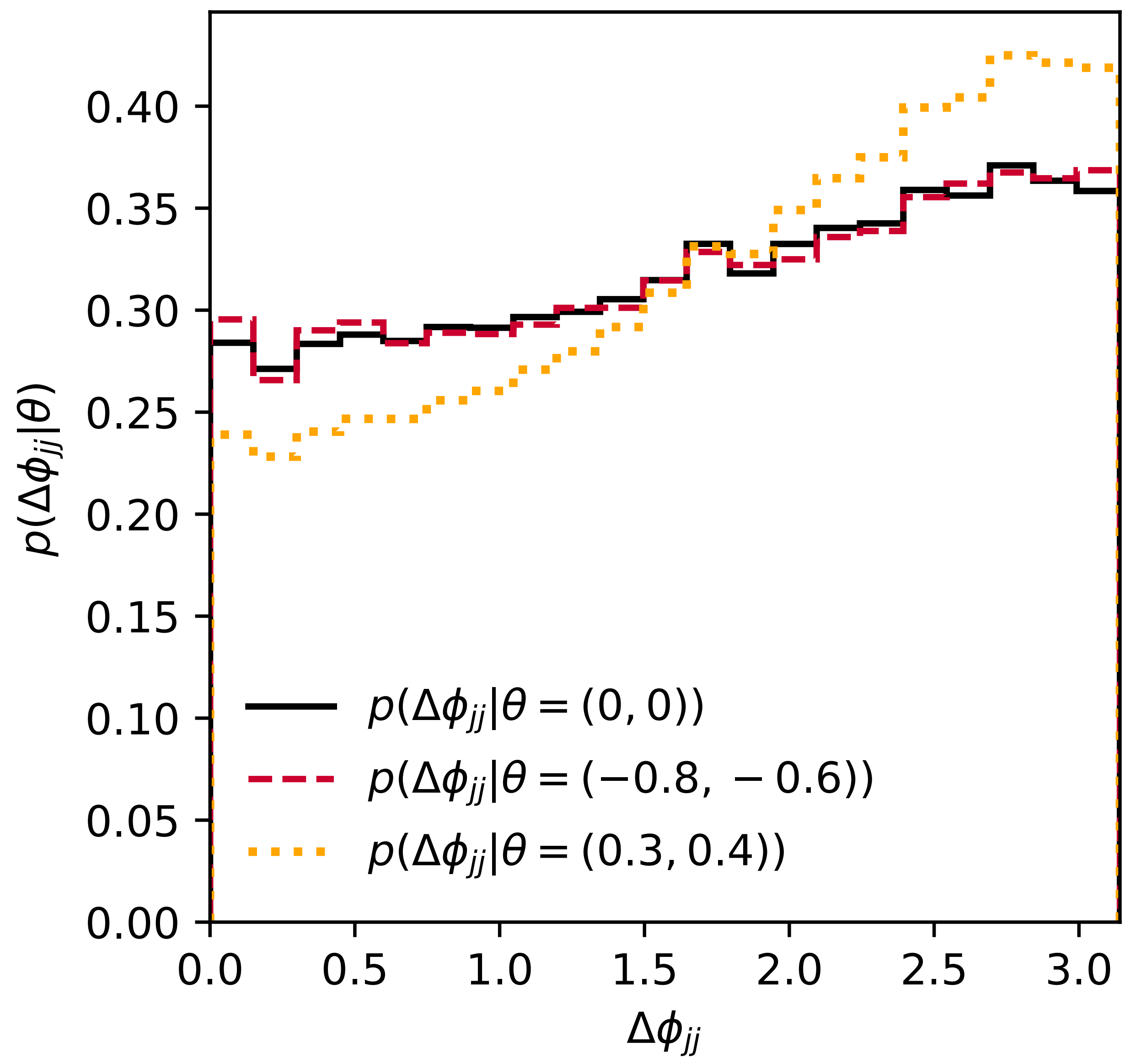
# Perfect match for EFT measurements

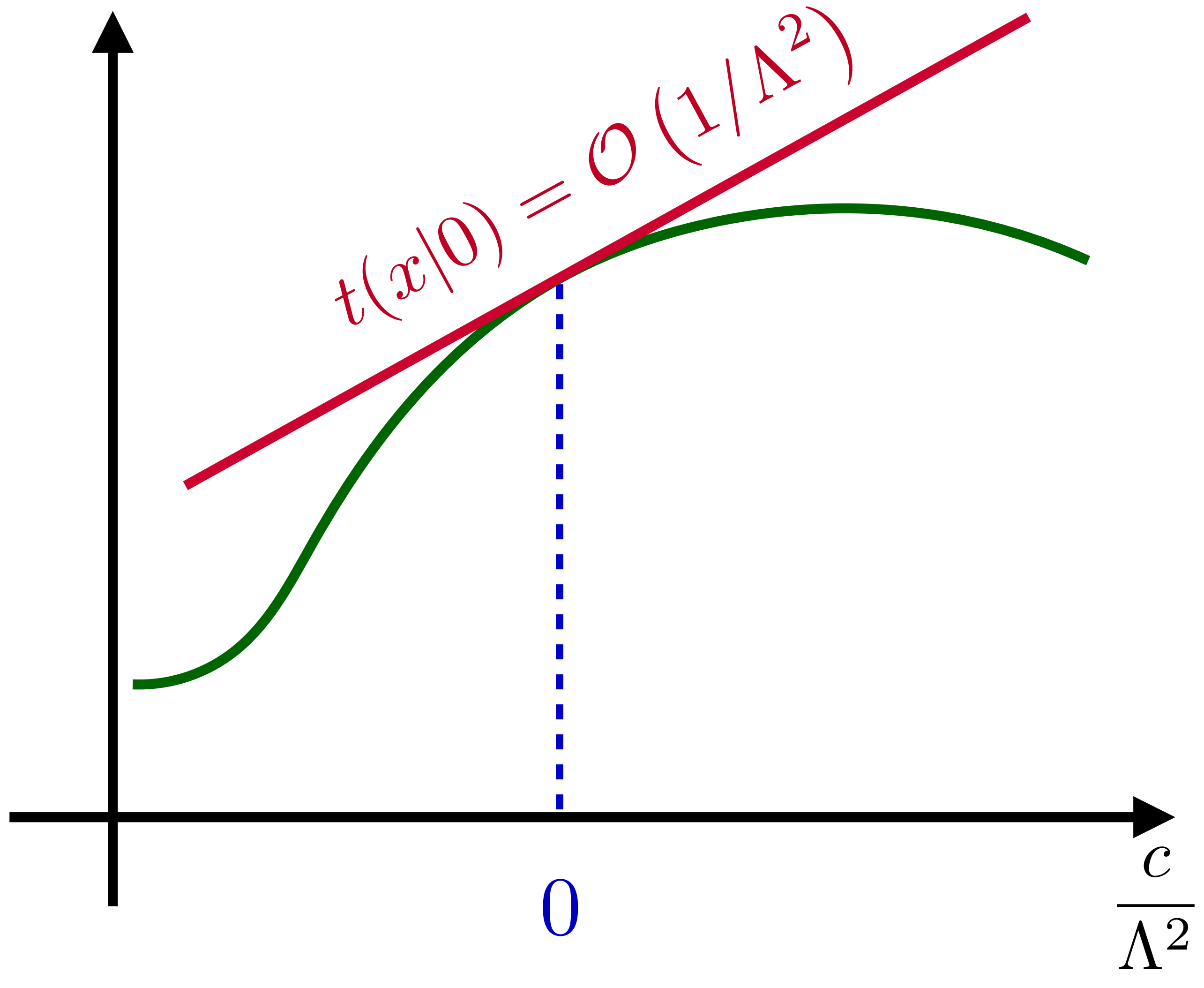



- Good for subtle kinematic effects

  (Subtle point: Large overlap of kinematic distributions reduces variance of joint likelihood ratio / joint score)

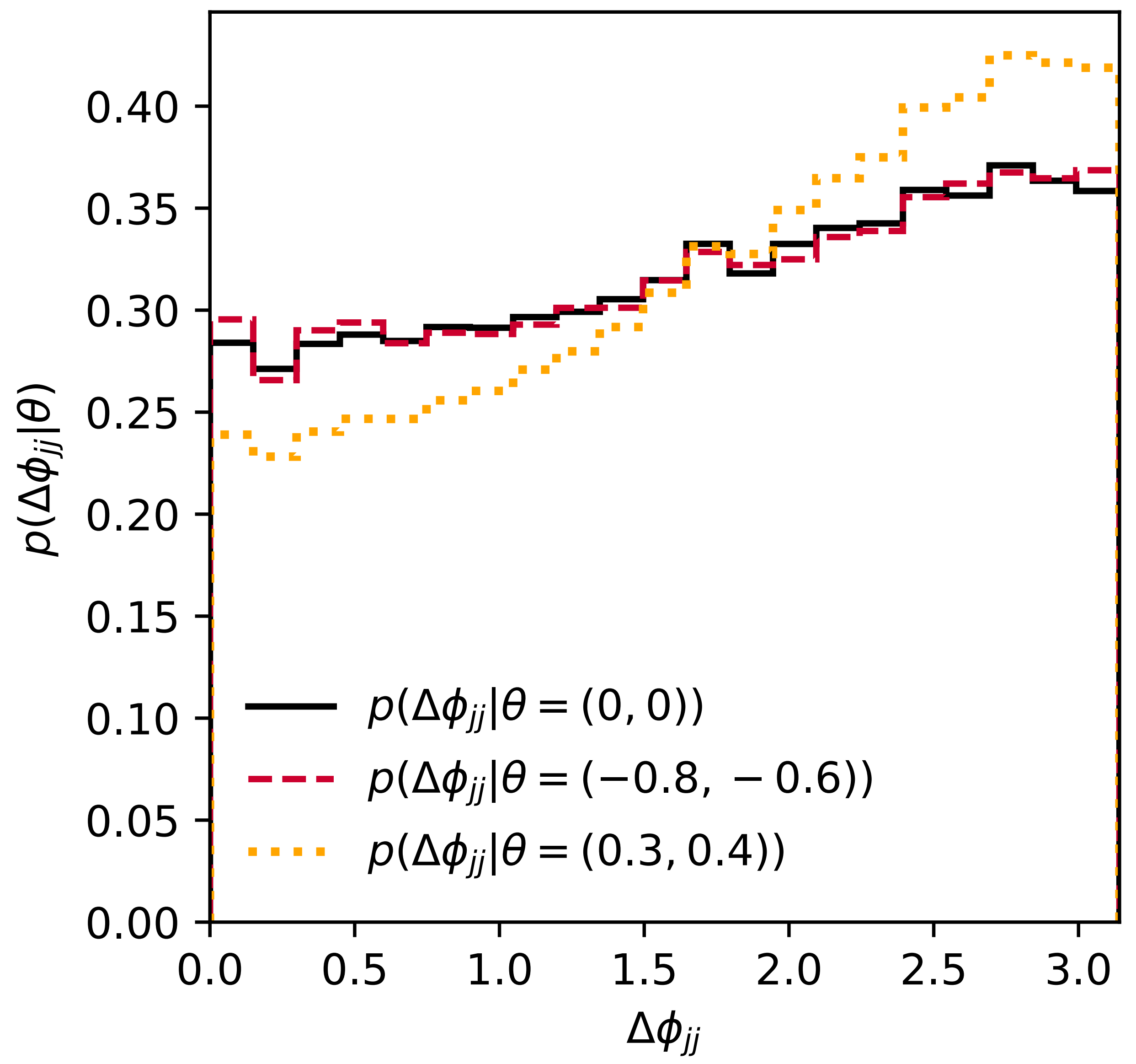
# Perfect match for EFT measurements

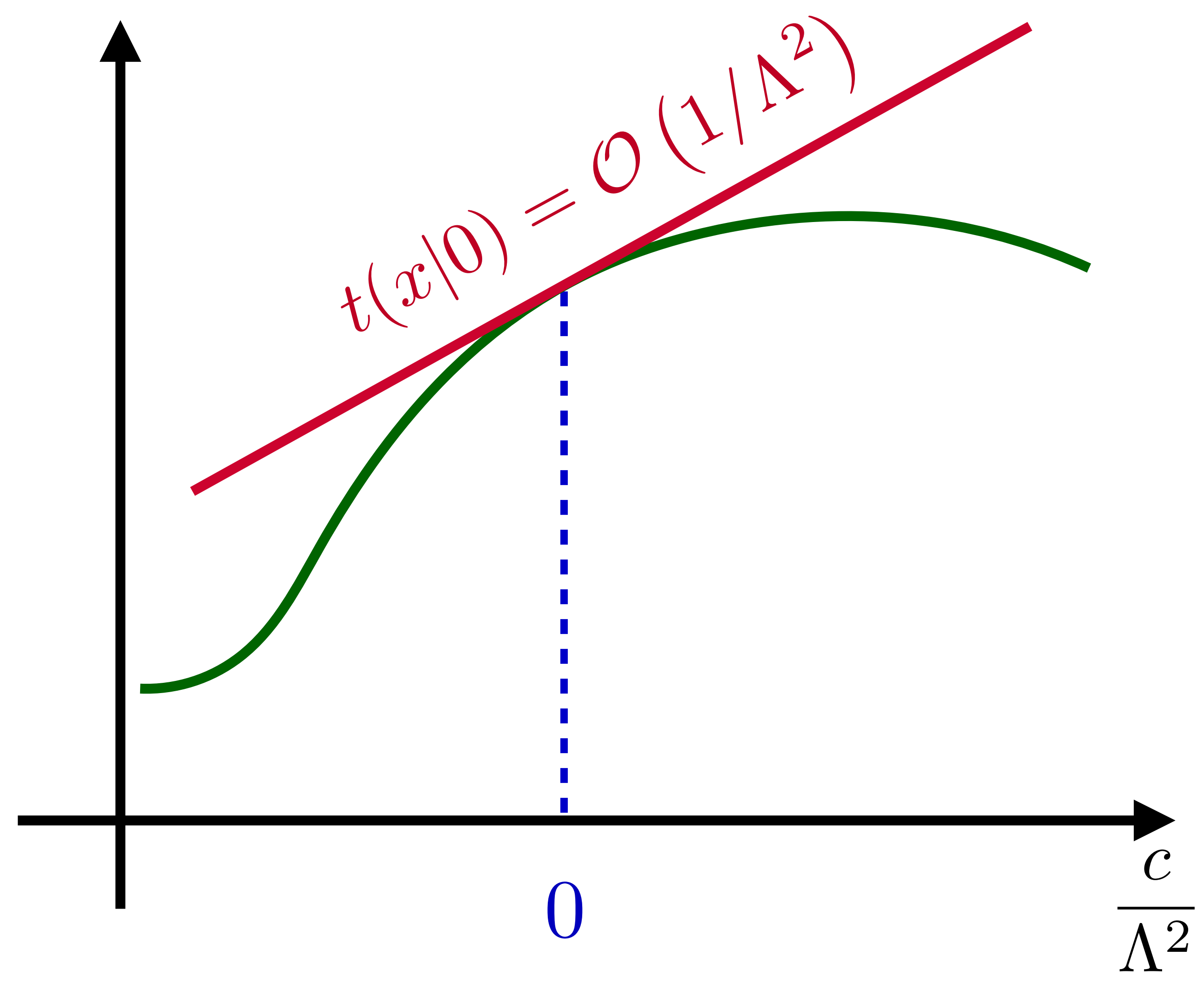


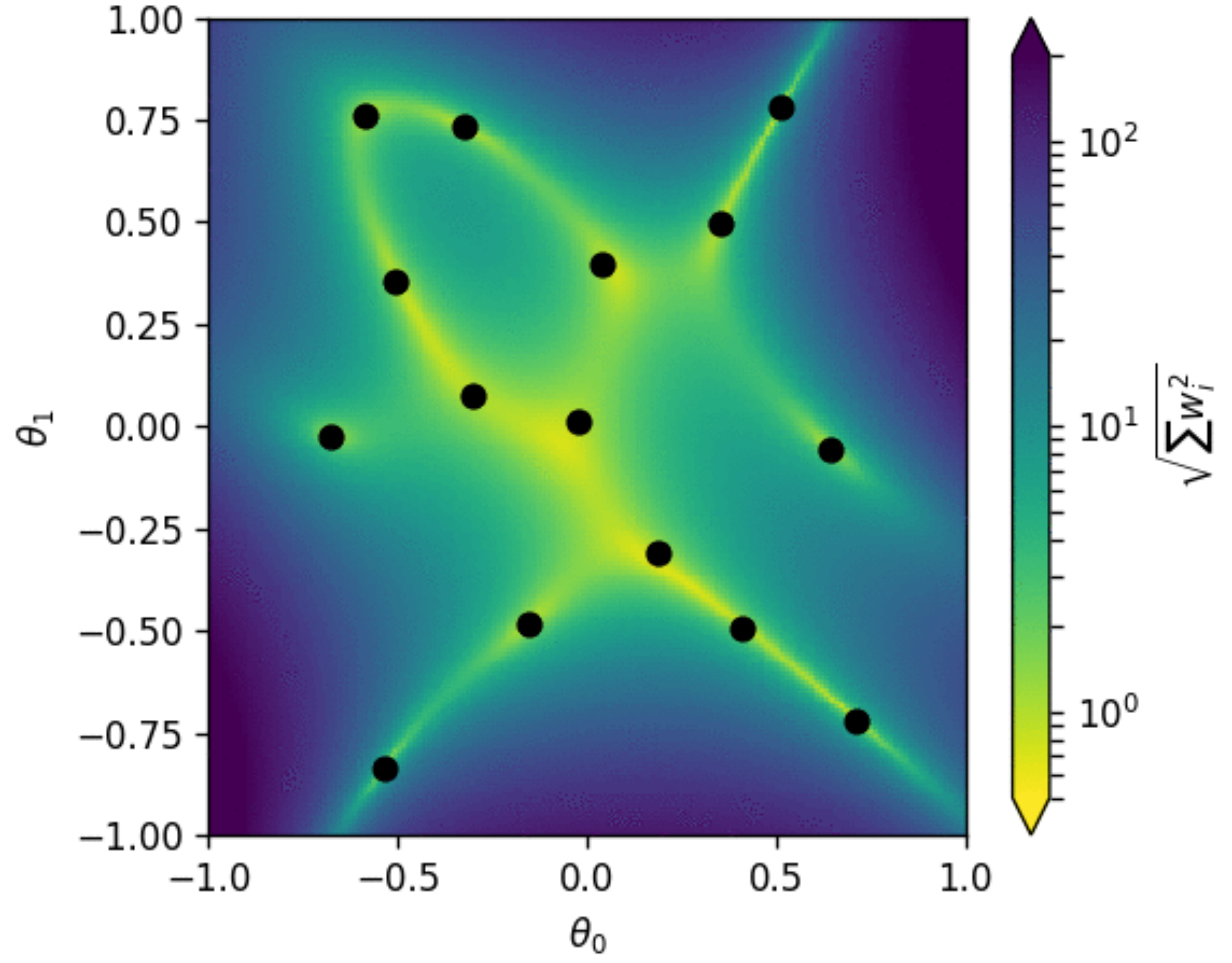




- Good for subtle kinematic effects
  (Subtle point: Large overlap of kinematic distributions reduces variance of joint likelihood ratio / joint score)
- Interference effects can be isolated using SALLY at the SM

# Perfect match for EFT measurements




- Good for subtle kinematic effects
  (Subtle point: Large overlap of kinematic distributions reduces variance of joint likelihood ratio / joint score)
- Interference effects can be isolated using SALLY at the SM
- Morphing techniques allow fast reweighting to any parameter points
  [e.g. ATL-PHYS-PUB-2015-047]

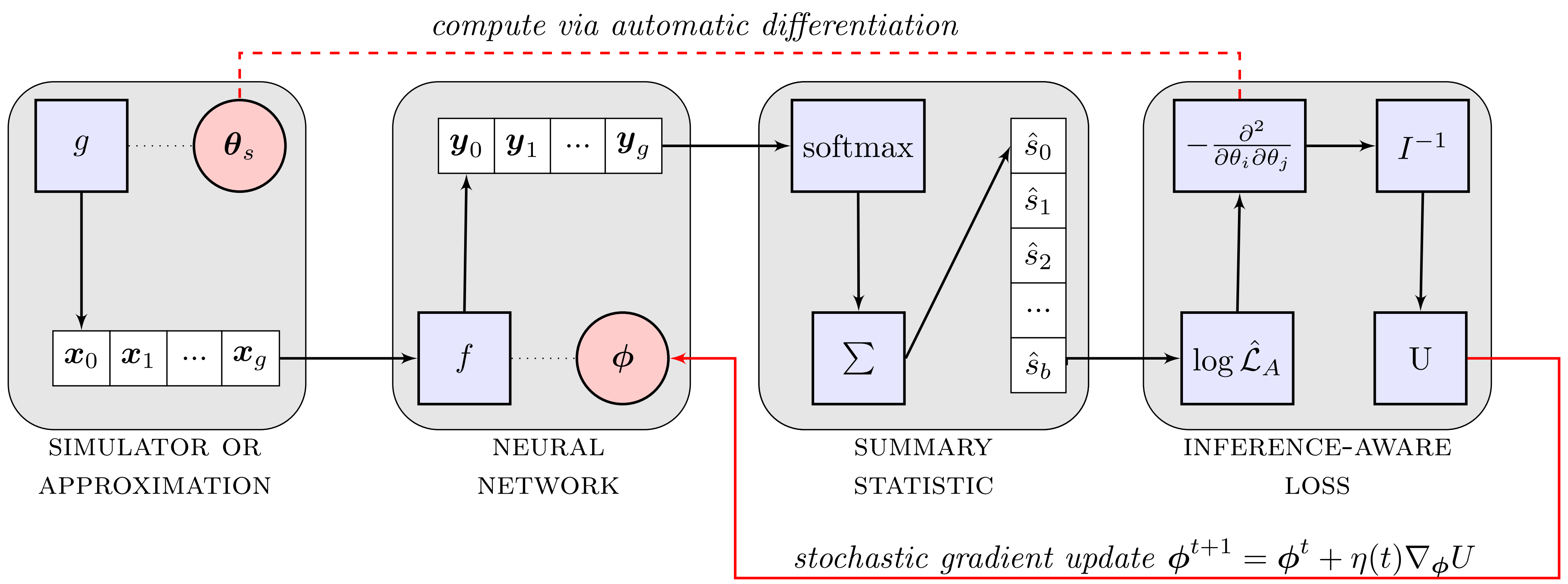
# End-to-end optimization with autodiff

- With tools like MadMiner the objective is to learn a likelihood ratio, which is known optimal properties for measurements etc.
- In INFERNO and Neo the inference objective is directly optimized

**INFERNO: Inference-Aware Neural Optimisation**


**Pablo de Castro**
INFN - Sezione di Padova
pablo.de.castro@cern.ch

**Tommaso Dorigo**
INFN - Sezione di Padova
tommaso.dorigo@cern.ch

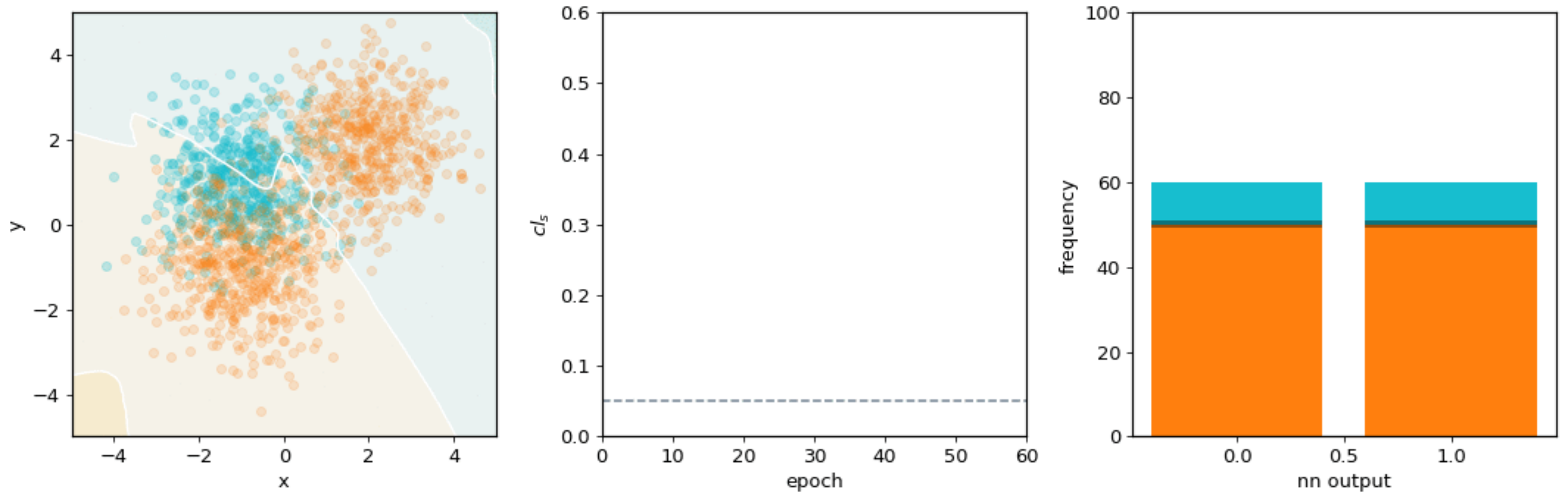

**Kyle Cranmer** @KyleCranmer · 19h
Take note! Here is a nice example of differentiable programming. It shows end-to-end optimization of a NN for event categorization wrt. final statistical analysis (using pyhf). Requires running gradients through results of maximum likelihood with fixed-point differentiation 👏

**Nathan Simpson @ CERN**
@phi_nate

I'm *very* excited to share with you what I've been working on recently in collaboration with @lukasheinrich_ !

We've developed a module that performs end-to-end learning with respect to statistical inference in particle physics.

try it yourself at github.com/pyhf/neos! :)




10:58 AM · Mar 5, 2020 · Twitter Web App

https://github.com/pyhf/neos

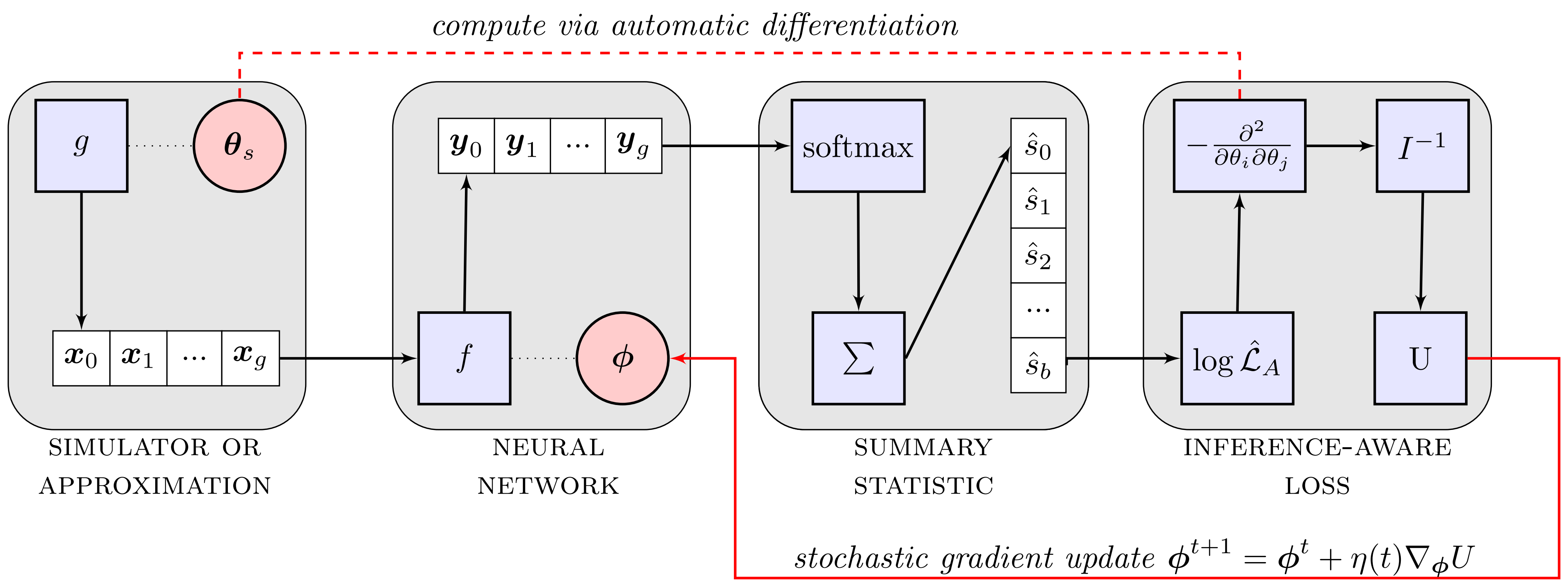
# End-to-end optimization with autodiff

- With tools like MadMiner the objective is to learn a likelihood ratio, which is known optimal properties for measurements etc.
- In INFERNO and Neo the inference objective is directly optimized

**INFERNO: Inference-Aware Neural Optimisation**


**Pablo de Castro**
INFN - Sezione di Padova
pablo.de.castro@cern.ch

**Tommaso Dorigo**
INFN - Sezione di Padova
tommaso.dorigo@cern.ch

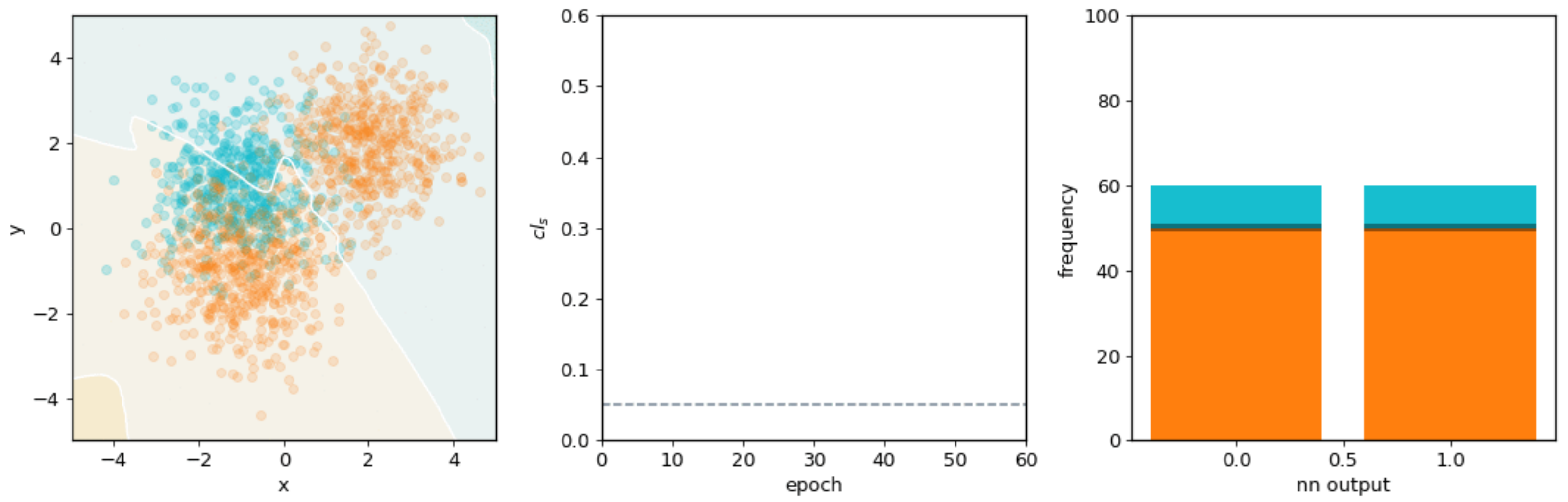

**Kyle Cranmer** @KyleCranmer · 19h
Take note! Here is a nice example of differentiable programming. It shows end-to-end optimization of a NN for event categorization wrt. final statistical analysis (using pyhf). Requires running gradients through results of maximum likelihood with fixed-point differentiation 👏

**Nathan Simpson @ CERN**
@phi_nate

I'm *very* excited to share with you what I've been working on recently in collaboration with @lukasheinrich_ !

We've developed a module that performs end-to-end learning with respect to statistical inference in particle physics.

try it yourself at github.com/pyhf/neos! :)




10:58 AM · Mar 5, 2020 · Twitter Web App

https://github.com/pyhf/neos

# Recap on Likelihood Ratios

$$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

For signal vs. background searches:

- **Neyman-Pearson Lemma**: optimal hypothesis test given by **likelihood ratio** (basis of Higgs search)
- Likelihood ratio $\frac{p(x|\theta_0)}{p(x|\theta_1)}$ also used for exclusion contours

For estimates of parameters $\hat{\theta}$

- **Cramér-Rao bound** states $\mathrm{cov}[\hat{\theta}|\theta_0]_{ij} \geq I^{-1}_{ij}(\theta_0)$ where $I_{ij}$ is the **Fisher-information matrix** (Hessian of log-likelihood)
- Motivates **Information Geometry** as a phenomenological tool
- Maximum-likelihood (asymptotically) saturates the bound

Note: $\nabla_\theta \log p(x|\theta)$ acts like a likelihood ratio locally

# Cramér-Rao Bound

The minimum variance bound on an unbiased estimator is given by the Cramér-Rao bound:

$$\mathrm{cov}[\hat{\theta}|\theta_0]_{ij} \geq I^{-1}_{ij}(\theta_0)$$

Expected error
of best-fit parameter

Inverse of
Fisher information

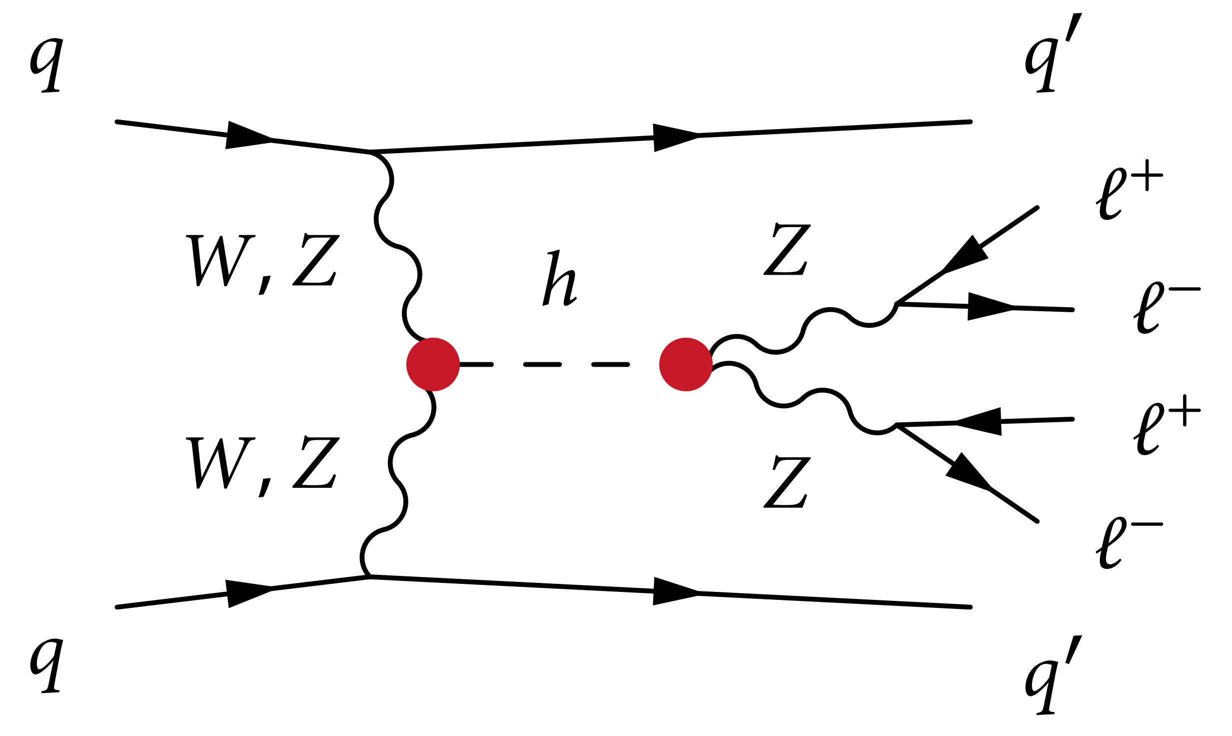
Fisher information matrix (is also a Riemannian metric! )

$$I_{ij}[\theta] = -\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial\theta_i\,\partial\theta_j}\middle|\theta\right]$$

Maximum Likelihood Estimators *asymptotically* reach this bound

# Challenge for EFT

Let $\theta$ denote the coefficients of higher dimensional operators in the Lagrangian, $x$ be high-dimensional data associated to an event, and $p(x \mid \theta) = \frac{1}{\sigma(\theta)} \frac{d\sigma}{d\theta}$ be the distribution for the data

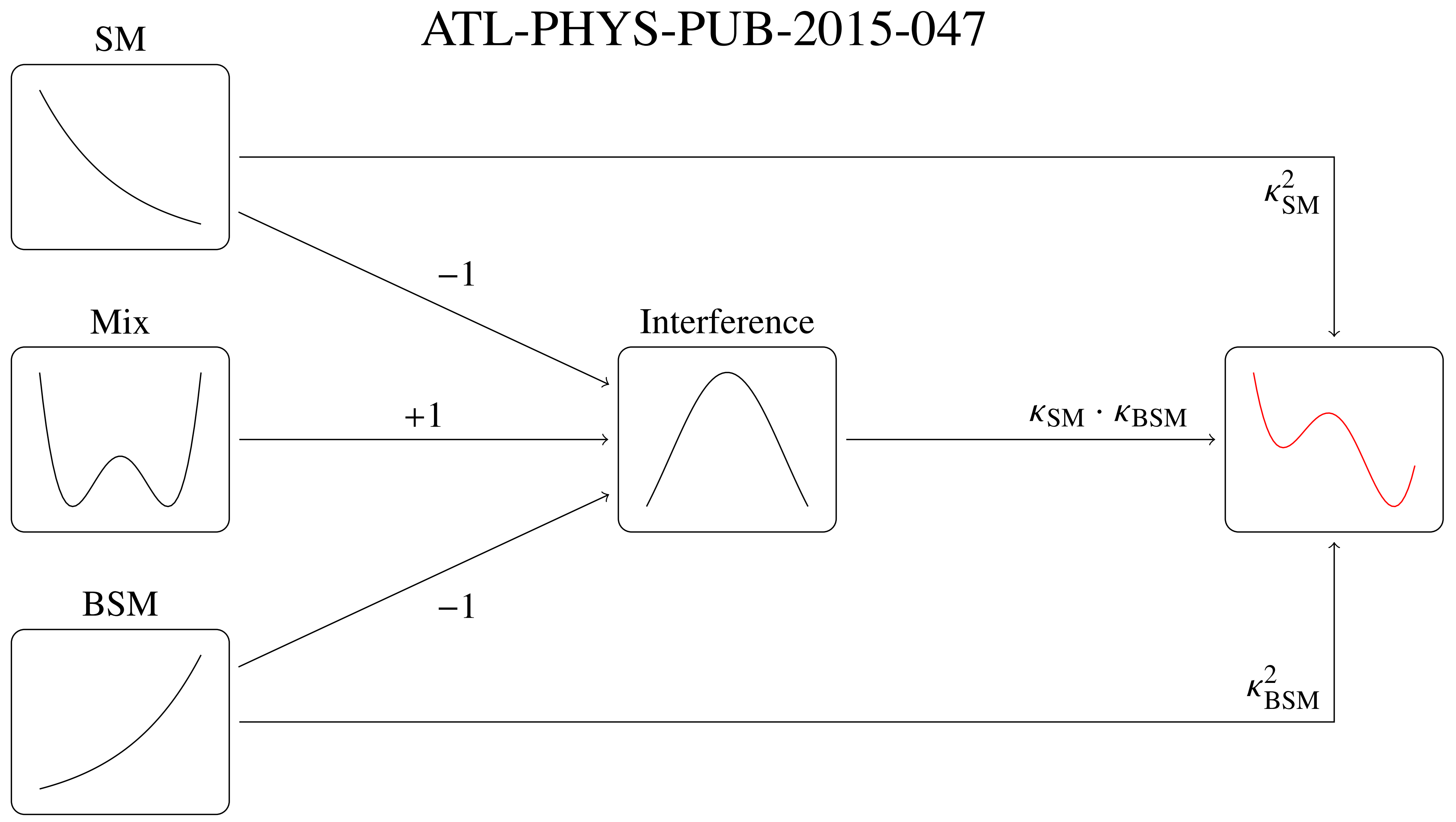



- we want to compare any two points in EFT parameter space
- evaluate the **likelihood ratio** $r(x|\theta_0, \theta_1) \equiv \frac{p(x|\theta_0)}{p(x|\theta_1)}$

Difficulty is that one changes the parameters of the EFT, the distributions p(x|θ) change due to interference.

- It would be very computationally expensive (infeasible) to generate samples for every value of θ and estimate $p(x \mid \theta)$ with histograms. Small changes mean we need a lot of MC events!
- Ideally we could directly estimate the **score** $t(x|\theta_0) \equiv \nabla_\theta \log p(x|\theta)\Big|_{\theta_0}$

# EFT Embedded in a vector space



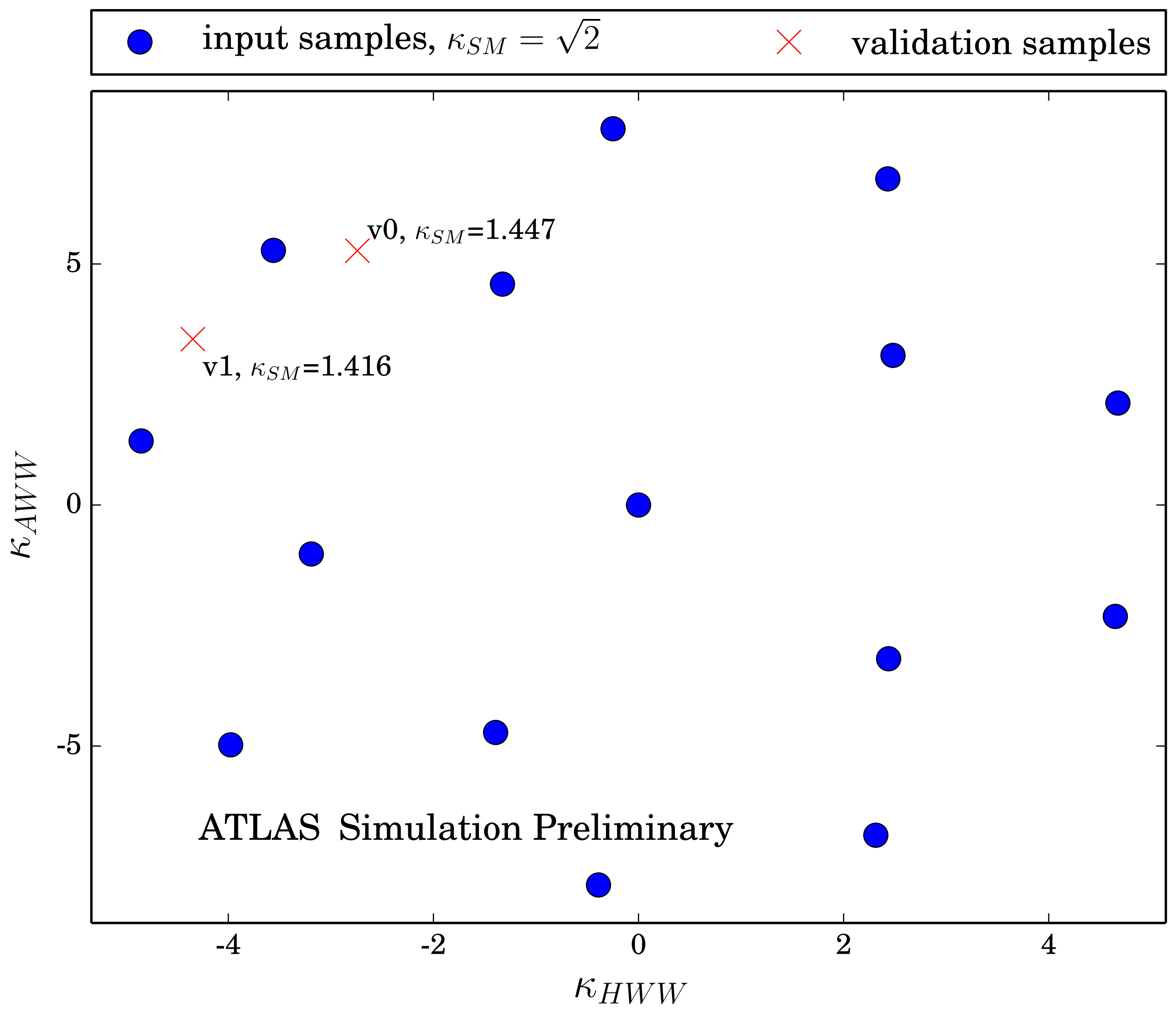

Difficulty is that one changes the parameters of the EFT, the distributions p(x|θ) change due to interference.

**But there is a trick:**

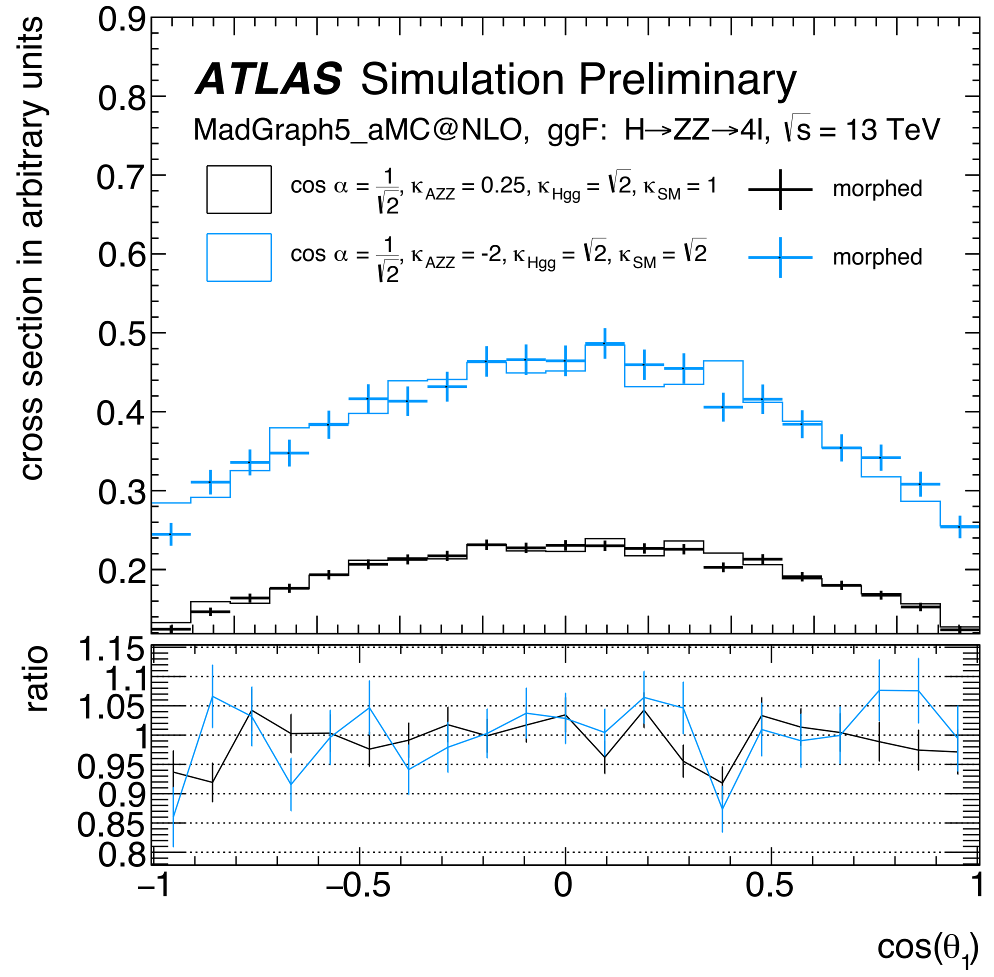
Simple example:

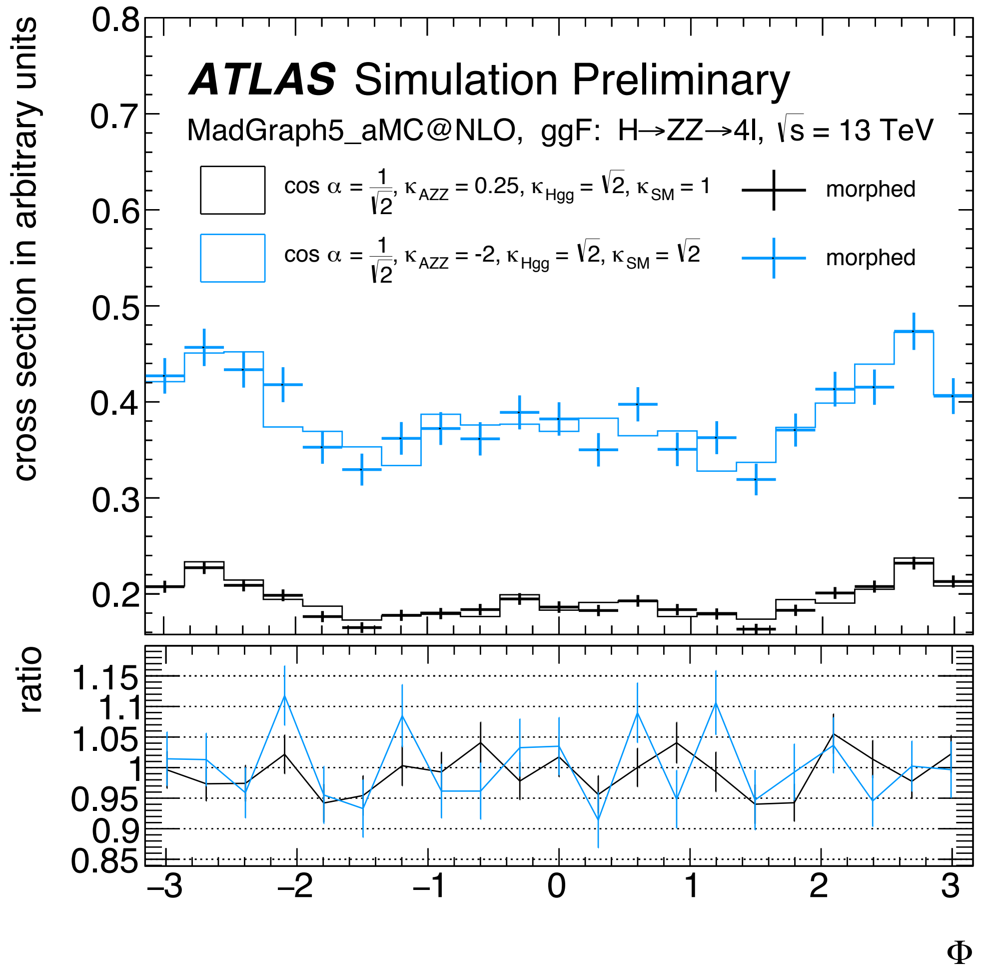
$$|g_1 M_{SM} + g_2 M_{BSM}|^2 = g_1^2 |M_{SM}|^2 + 2 g_1 g_2 Re\left[M^*_{SM} M_{BSM}\right] + g_2^2 |M_{BSM}|^2$$

3-d vector space, distribution for any point in this space is linear mixture of distribution for 3 basis samples!




(real examples need more basis samples)



# EFT Decomposition

production | decay

$$d\sigma \propto \left\| \left( \mathcal{M}^p_{\text{SM}} + \sum_i \frac{f_i}{\Lambda^2} \mathcal{M}^p_i \right) \left( \mathcal{M}^d_{\text{SM}} + \sum_j \frac{f_j}{\Lambda^2} \mathcal{M}^d_j \right) \right\|^2$$

Express EFT as a mixture:

$$p(x|\theta) = \sum_c w_c(\theta) p_c(x)$$

$w_c(\theta)$ are polynomials

$\nabla_\theta \log p(x|\theta)$ is now possible!

| Process | Number of components for $n$ operators | | | | | |
|---|---|---|---|---|---|---|
| | $\mathcal{O}(\Lambda^0)$ | $\mathcal{O}(\Lambda^{-2})$ | $\mathcal{O}(\Lambda^{-4})$ | $\mathcal{O}(\Lambda^{-6})$ | $\mathcal{O}(\Lambda^{-8})$ | $\sum$ |
| $hV$ / WBF production | 1 | $n$ | $\frac{n(n+1)}{2}$ | | | $\frac{(n+1)(n+2)}{2}$ |
| $h \to VV$ decay | 1 | $n$ | $\frac{n(n+1)}{2}$ | | | $\frac{(n+1)(n+2)}{2}$ |
| Production + decay | 1 | $n$ | $\frac{n(n+1)}{2}$ | $\binom{n+2}{3}$ | $\binom{n+3}{4}$ | $\binom{n+4}{4}$ |

Table 1: Number of components $c$ as given in Eq. (6) for different processes, sorted by their suppression by the EFT cutoff scale $\Lambda$.

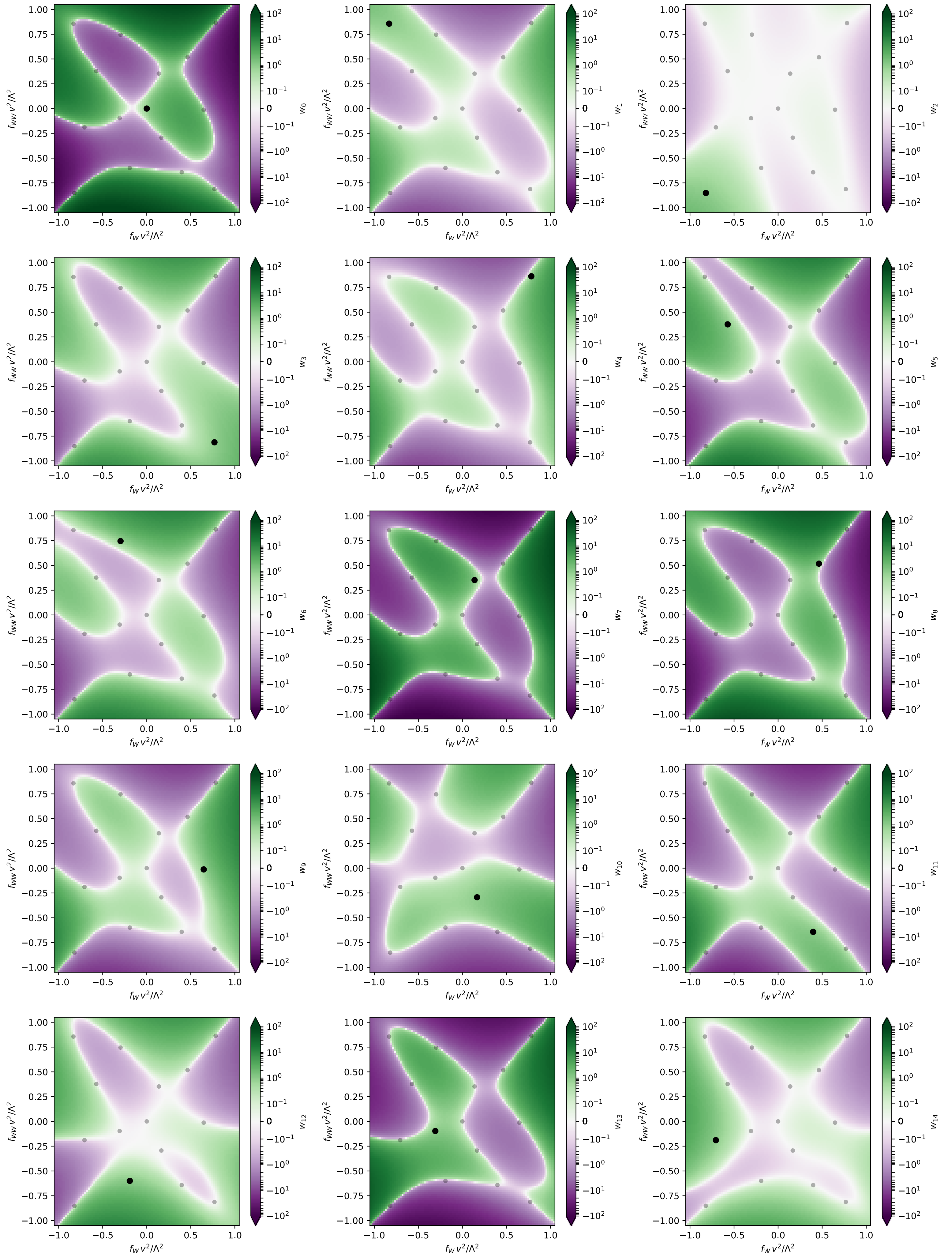


Figure 13: Morphing weights $w_i(\boldsymbol{\theta})$ for basis points distributed over the full relevant parameter space.

For 2 BSM operators affecting VBF Higgs production and decay, we need a 15-D vector space

For 5 BSM operators we need 126-D vector space