Understand and Improving Higgs Measurements using Information Geometry

Felix Kling

SLAC NATIONAL ACCELERATOR LABORATORY

19 June 2020
Joint INFN-UNIMI-UNIMIB Pheno Seminars

Motivation

Status of the Field:
- Higgs discovery: Standard Model complete
- no discovery of new physics yet

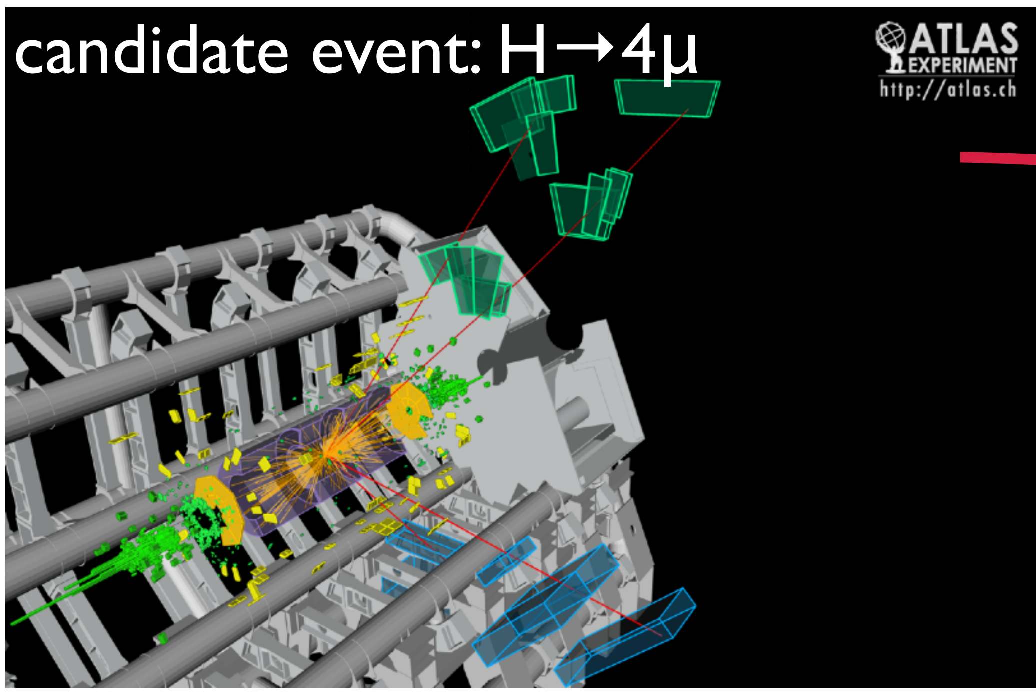
- maybe BSM physics will be obvious (di-X resonance at XXX GeV)
- maybe BSM physics is more subtle (Higgs couplings in dim6 SMEFT)

Status of the Field:

- Higgs discovery: Standard Model complete
- no discovery of new physics yet

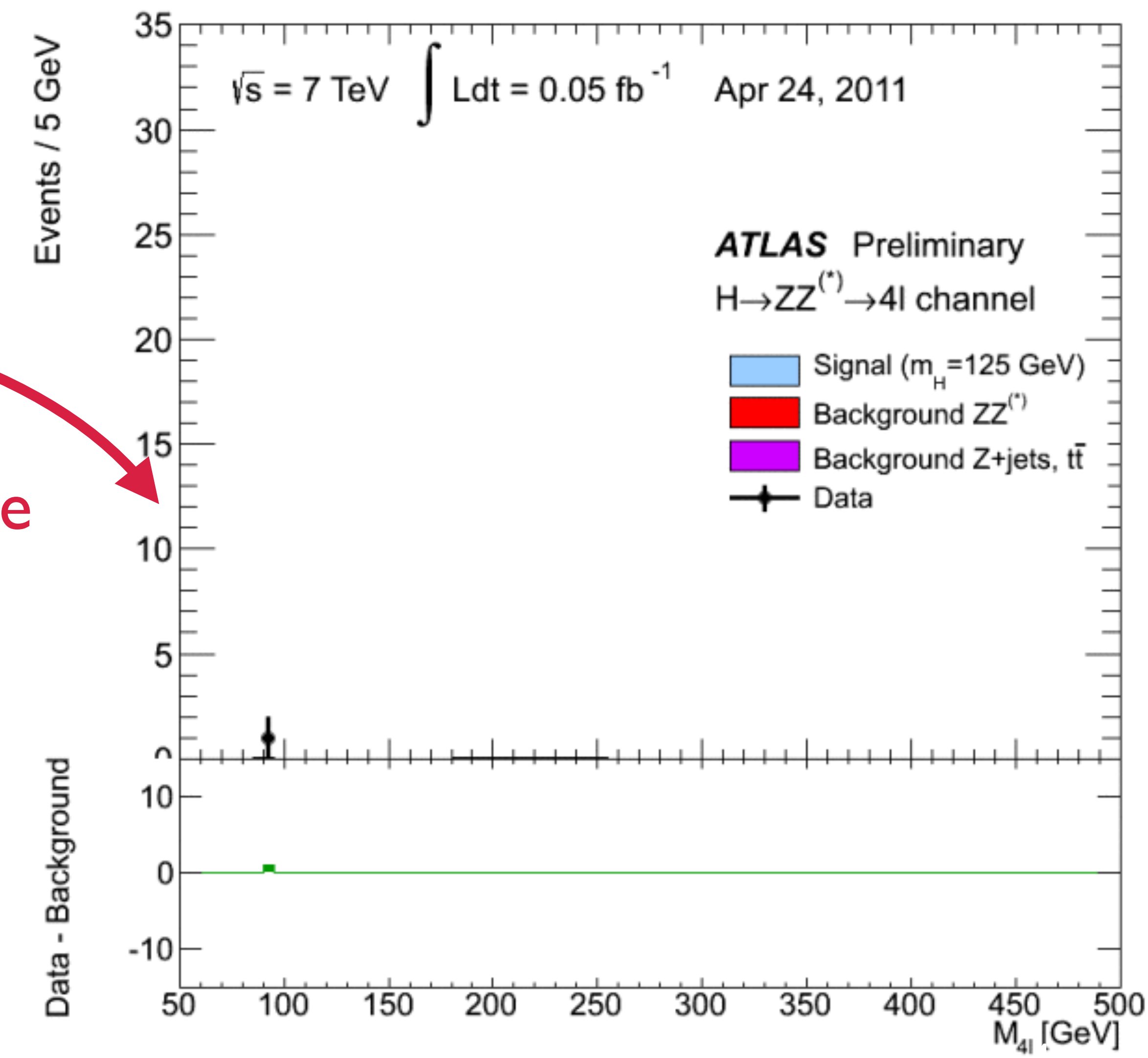
- maybe BSM physics will be obvious (di-X resonance at XXX GeV)
- maybe BSM physics is more subtle (Higgs couplings in dim6 SMEFT)

2012: discovery of the Higgs boson



collect many events

plot carefully chosen variable



invariant mass: m(4μ)

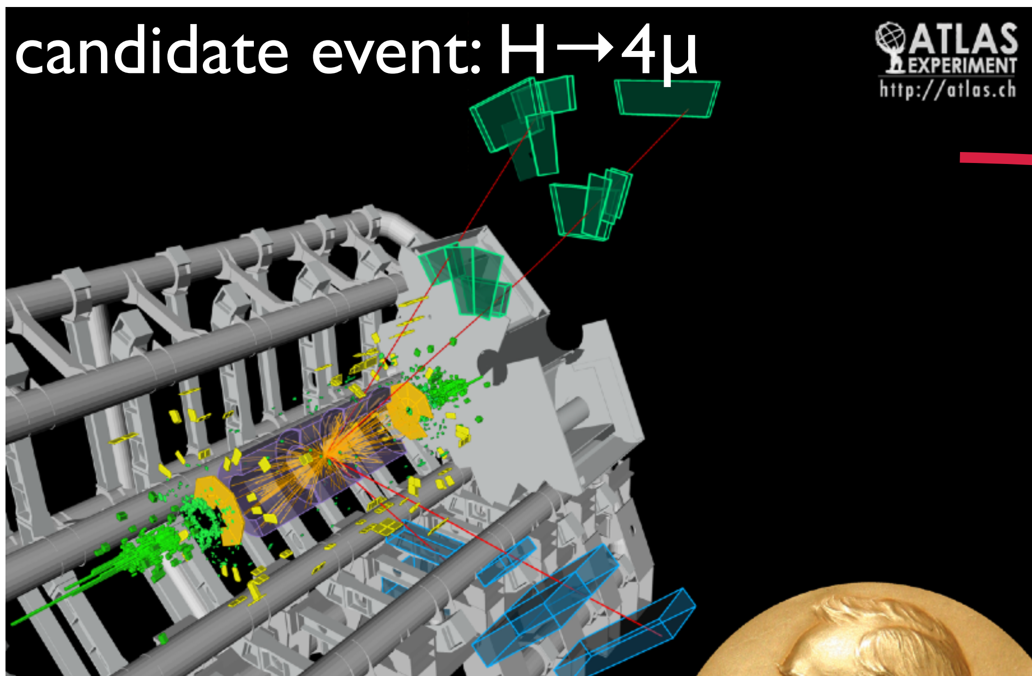
Motivation

Status of the Field:

- Higgs discovery: Standard Model complete
- no discovery of new physics yet

- maybe BSM physics will be obvious (di-X resonance at XXX GeV)
- maybe BSM physics is more subtle (Higgs couplings in dim6 SMEFT)

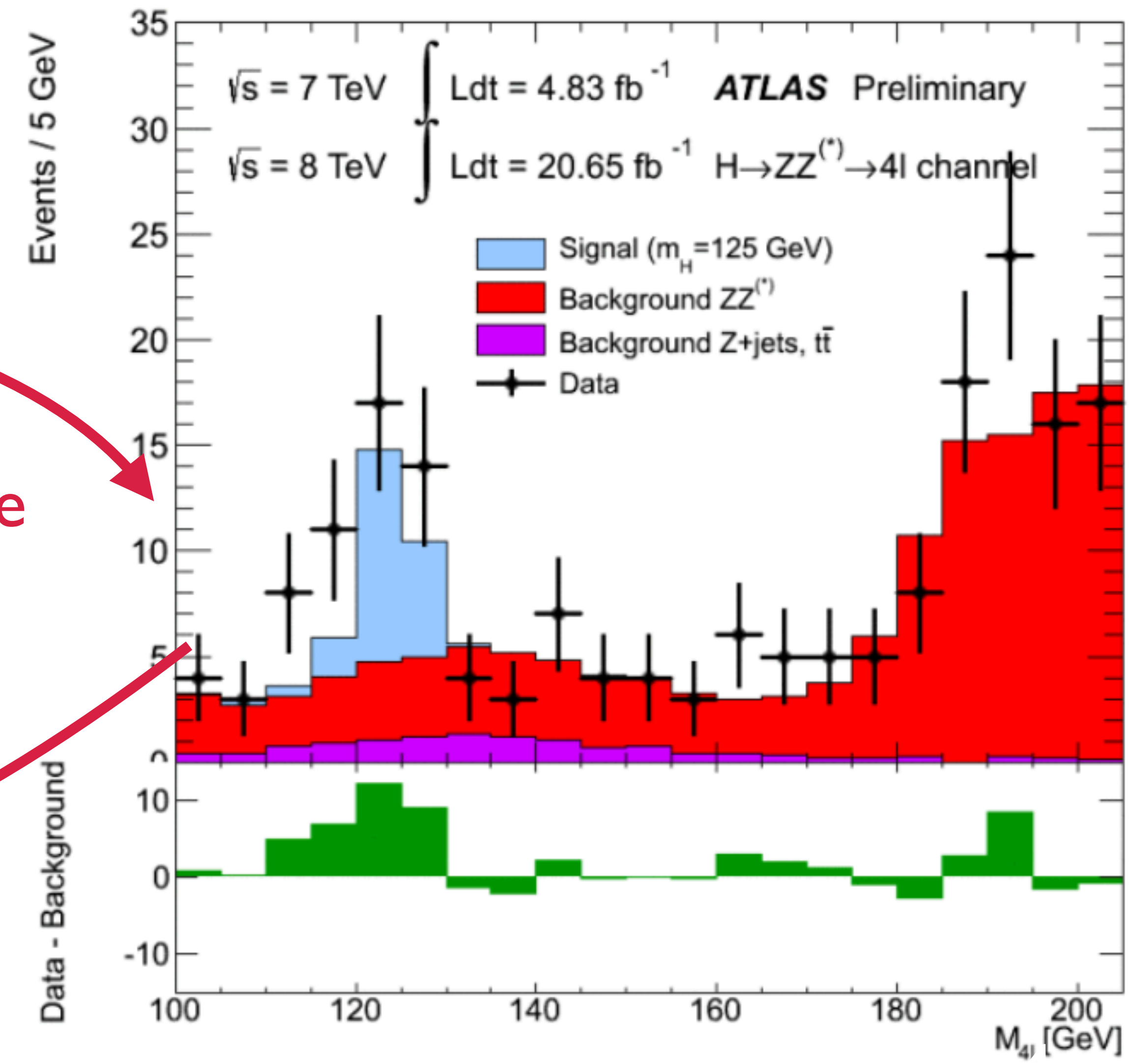
2012: discovery of the Higgs boson



collect many events

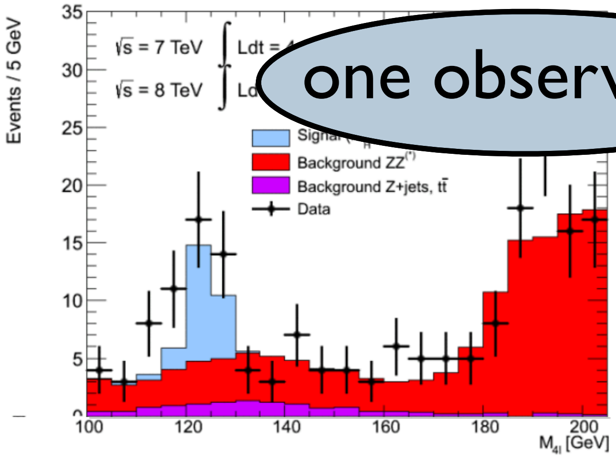
plot carefully chosen variable

discovery



invariant mass: $m(4\mu)$

Theory in an Era of Data



one observable

Year

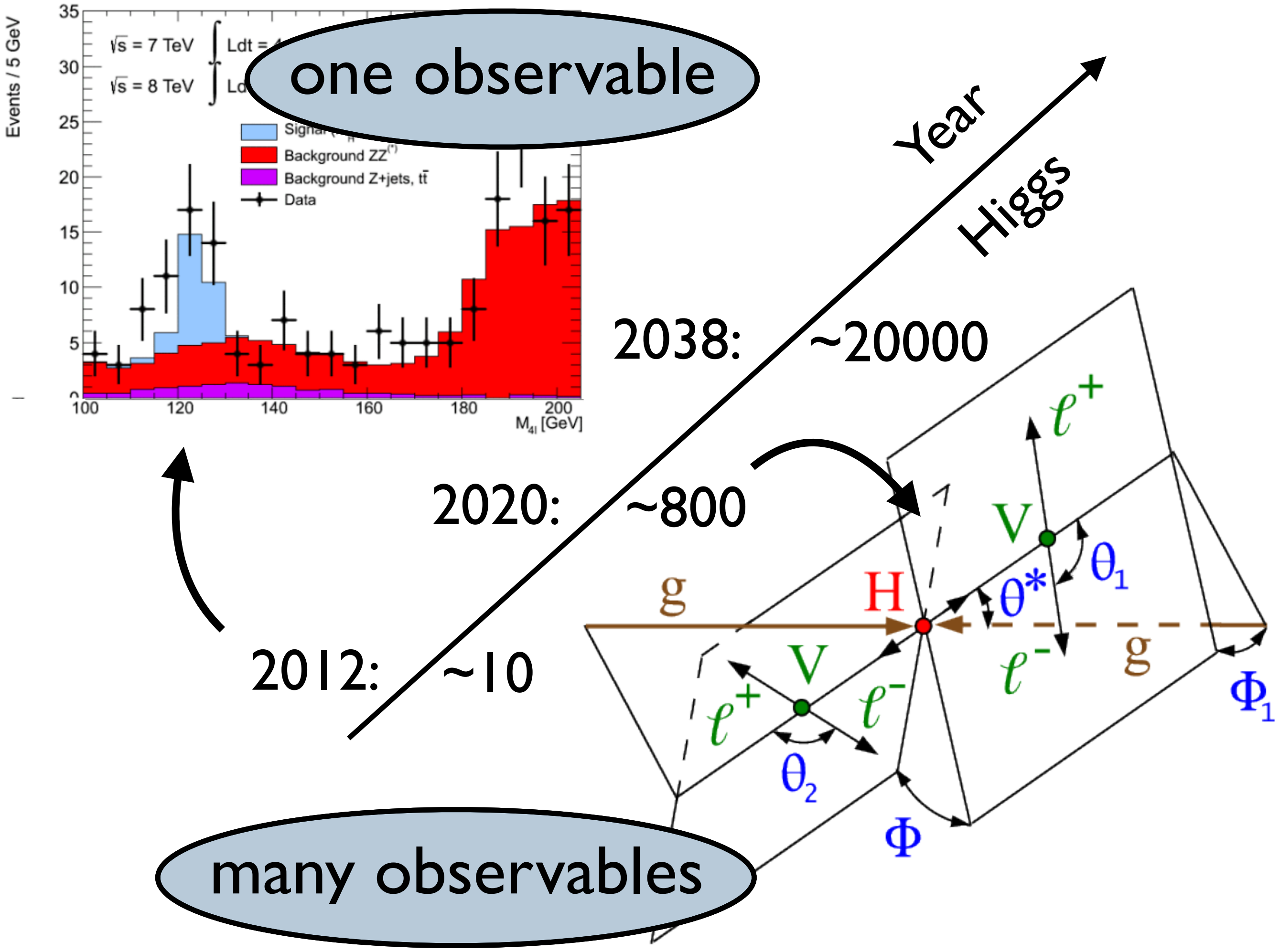
Higgs

2038: ~20000

2020: ~800

2012: ~10

Theory in an Era of Data

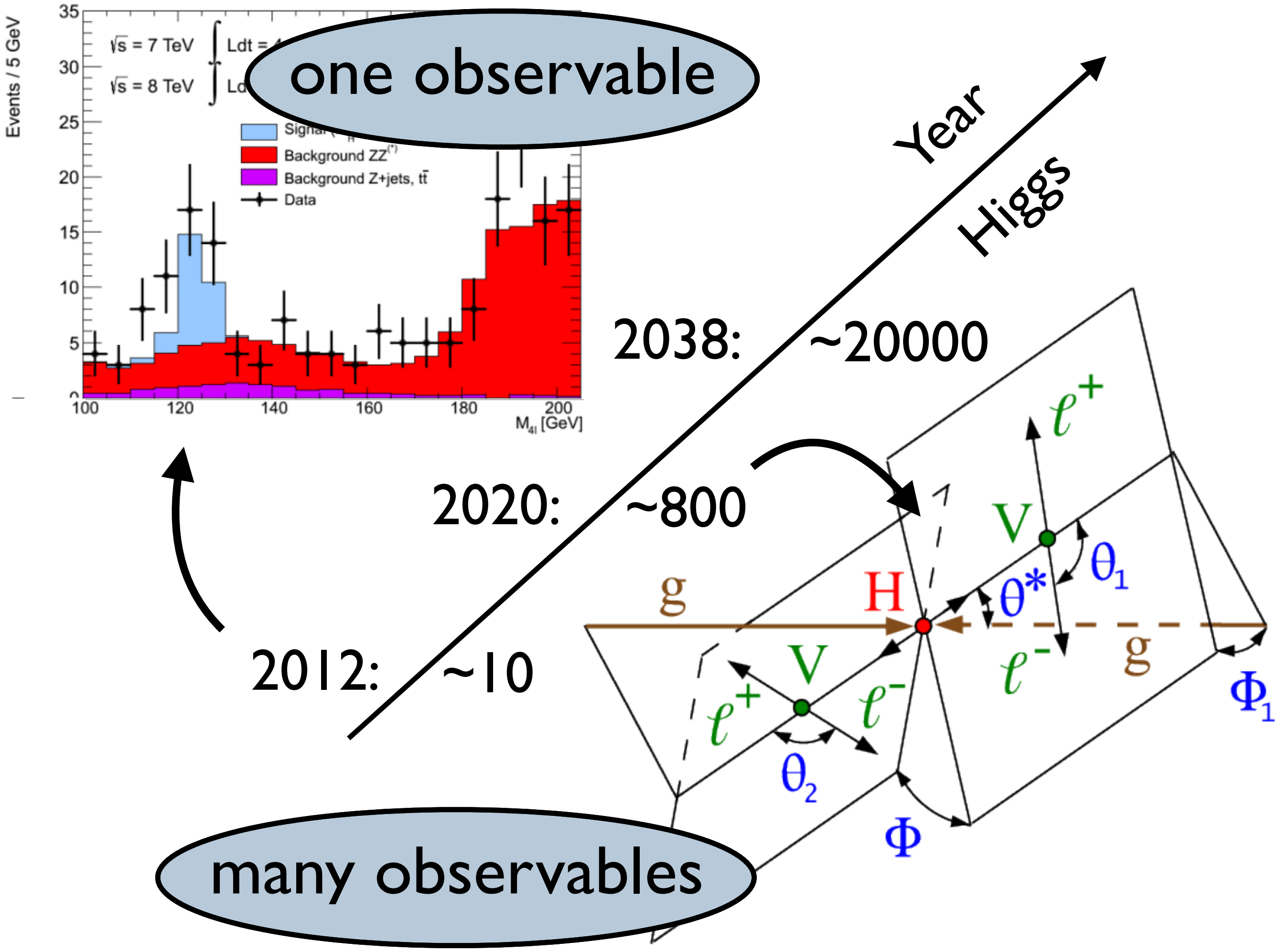


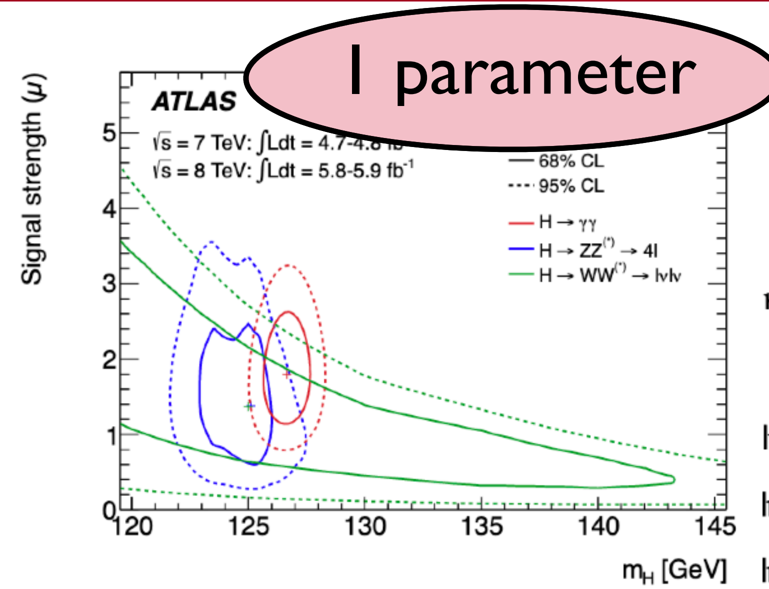
Era of Data

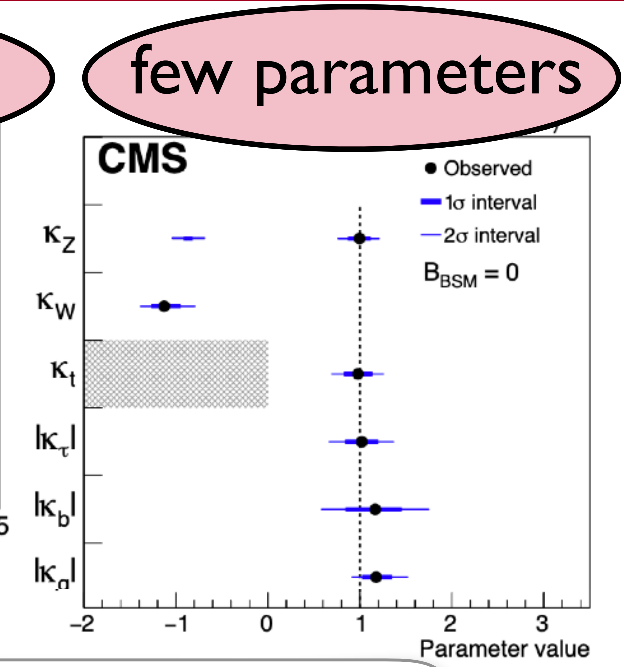
large statistics

many observables

Theory in an Era of Data







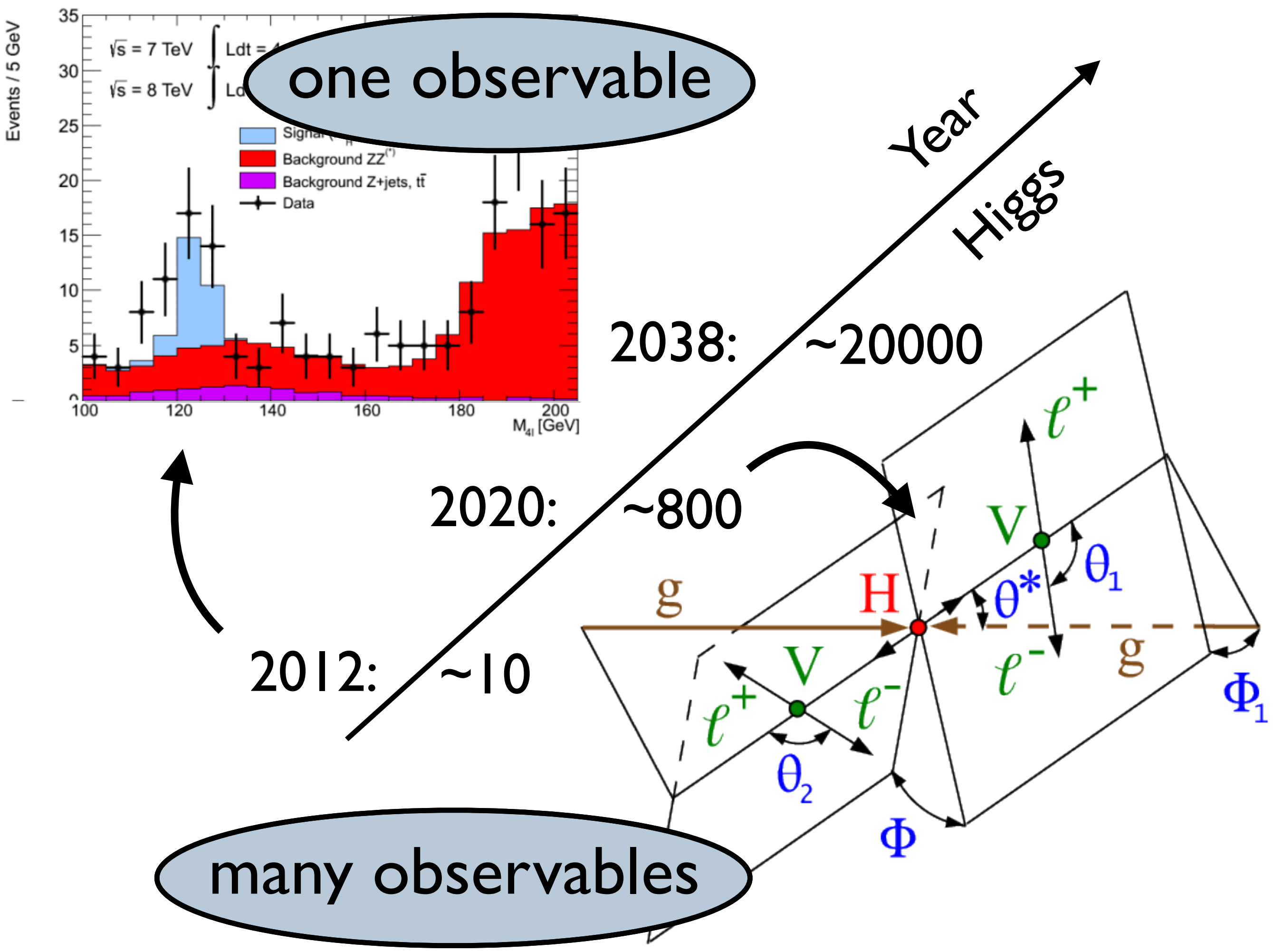
$$\Delta\mathcal{L}_{SILH} = \frac{\bar{c}_H}{2v^2}\partial^\mu\left(H^\dagger H\right)\partial_\mu\left(H^\dagger H\right) + \frac{\bar{c}_T}{2v^2}\left(H^\dagger \overleftrightarrow{D^\mu} H\right)\left(H^\dagger \overleftrightarrow{D}_\mu H\right) - \frac{\bar{c}_6\lambda}{v^2}\left(H^\dagger H\right)^3$$
$$+\left(\left(\frac{\bar{c}_u}{v^2}y_u H^\dagger H \bar{q}_L H^c u_R + \frac{\bar{c}_d}{v^2}y_d H^\dagger H \bar{q}_L H d_R + \frac{\bar{c}_l}{v^2}y_l H^\dagger H \bar{L}_L H l_R\right) + \text{h.c.}\right)$$
$$+\frac{i\bar{c}_W g}{2m_W^2}\left(H^\dagger \sigma^i \overleftrightarrow{D^\mu} H\right)\left(D^\nu W_{\mu\nu}\right)^i + \frac{i\bar{c}_B g'}{2m_W^2}\left(H^\dagger \overleftrightarrow{D^\mu} H\right)\left(\partial^\nu B_{\mu\nu}\right)$$
$$+\frac{i\bar{c}_{HW} g}{m_W^2}\left(D^\mu H\right)^\dagger \sigma^i \left(D^\nu H\right) W^i_{\mu\nu} + \frac{i\bar{c}_{HB} g'}{m_W^2}\left(D^\mu H\right)^\dagger\left(D^\nu H\right) B_{\mu\nu}$$
$$+\frac{\bar{c}_\gamma g'^2}{m_W^2} H^\dagger H B_{\mu\nu}B^{\mu\nu} + \frac{\bar{c}_g g_S^2}{m_W^2} H^\dagger H G^a_{\mu\nu}G^{a\mu\nu}$$

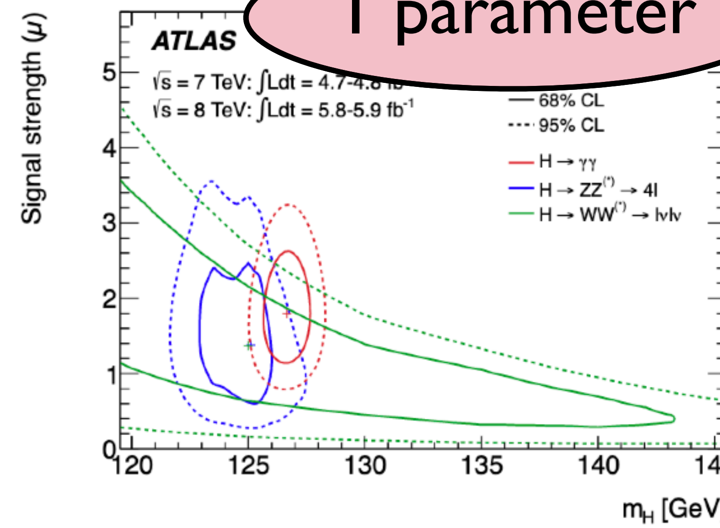
many parameters

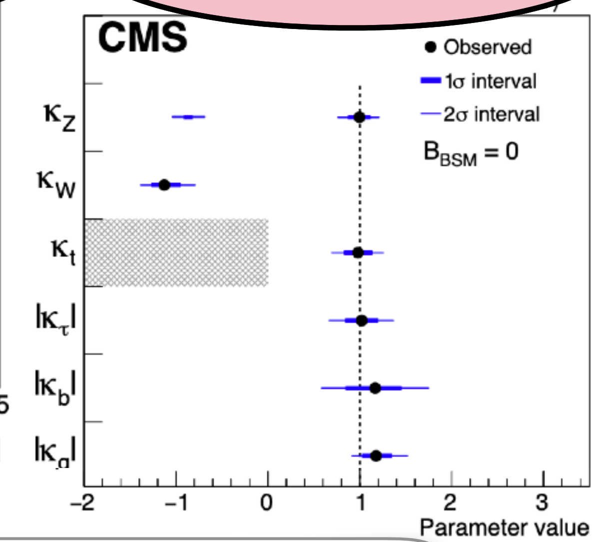
Era of Data
large statistics
many observables

Complex Theories
e.g. SM effective field theory
predict subtle kinematic features

Theory in an Era of Data







$$
\begin{aligned}
\Delta\mathcal{L}_{SILH} =\; & \frac{\bar{c}_H}{2v^2}\partial^\mu\left(H^\dagger H\right)\partial_\mu\left(H^\dagger H\right) + \frac{\bar{c}_T}{2v^2}\left(H^\dagger \overleftrightarrow{D^\mu} H\right)\left(H^\dagger \overleftrightarrow{D}_\mu H\right) - \frac{\bar{c}_6\lambda}{v^2}\left(H^\dagger H\right)^3 \\
& + \left(\left(\frac{\bar{c}_u}{v^2} y_u H^\dagger H \bar{q}_L H^c u_R + \frac{\bar{c}_d}{v^2} y_d H^\dagger H \bar{q}_L H d_R + \frac{\bar{c}_l}{v^2} y_l H^\dagger H \bar{L}_L H l_R\right) + \text{h.c.}\right) \\
& + \frac{i\bar{c}_W g}{2m_W^2}\left(H^\dagger \sigma^i \overleftrightarrow{D^\mu} H\right)\left(D^\nu W_{\mu\nu}\right)^i + \frac{i\bar{c}_B g'}{2m_W^2}\left(H^\dagger \overleftrightarrow{D^\mu} H\right)\left(\partial^\nu B_{\mu\nu}\right) \\
& + \frac{i\bar{c}_{HW} g}{m_W^2}\left(D^\mu H\right)^\dagger \sigma^i \left(D^\nu H\right) W^i_{\mu\nu} + \frac{i\bar{c}_{HB} g'}{m_W^2}\left(D^\mu H\right)^\dagger\left(D^\nu H\right) B_{\mu\nu} \\
& + \frac{\bar{c}_\gamma g'^2}{m_W^2} H^\dagger H B_{\mu\nu} B^{\mu\nu} + \frac{\bar{c}_g g_S^2}{m_W^2} H^\dagger H G^a_{\mu\nu} G^{a\mu\nu}
\end{aligned}
$$

many parameters

Era of Data

large statistics
many observables

Complex Theories

e.g. SM effective field theory
predict subtle kinematic features

How can we extract all the information from the data?

Outline

Some Statistic Concepts
Likelihoods, Optimal Observables and Fisher Information

Information Geometry in Theory
Parton Level and Matrix Elements

Information Geometry in Reality
Detector Effects and Machine Learning

Summary and Conclusion

Outline

Some Statistic Concepts

Likelihoods, Optimal Observables and Fisher Information

Information Geometry in Theory

Parton Level and Matrix Elements

Information Geometry in Reality

Detector Effects and Machine Learning

Summary and Conclusion

Likelihood Function

Simulation and Measurements

$$\theta \xrightarrow[\text{simulation}]{p(x|\theta)} x \xrightarrow[\text{inference}]{p(x|\theta)} \hat{\theta} \pm \Delta\theta$$

parameters — simulation — observables — inference — estimator/uncertainty

Likelihood Function p(x|θ)

likelihood of an observation x as a function of the theory parameter θ

$$p_{\text{full}}(\{x\}|\theta) = \text{Pois}(n|L\sigma(\theta)) \times \prod_{i=1}^{n} p(x_i|\theta)$$

Likelihood Ratio $r(x|\theta_{ref},\theta)=p(x|\theta)/p(x|\theta_{ref})$

"how much more likely is data x described by theory θ than θ_{ref}"

Neyman-Pearson Lemma:

The log-likelihood ratio log $r(x|\theta_{ref},\theta)$ is the most powerful test statistic to discriminate between two hypotheses θ_{ref} and θ.

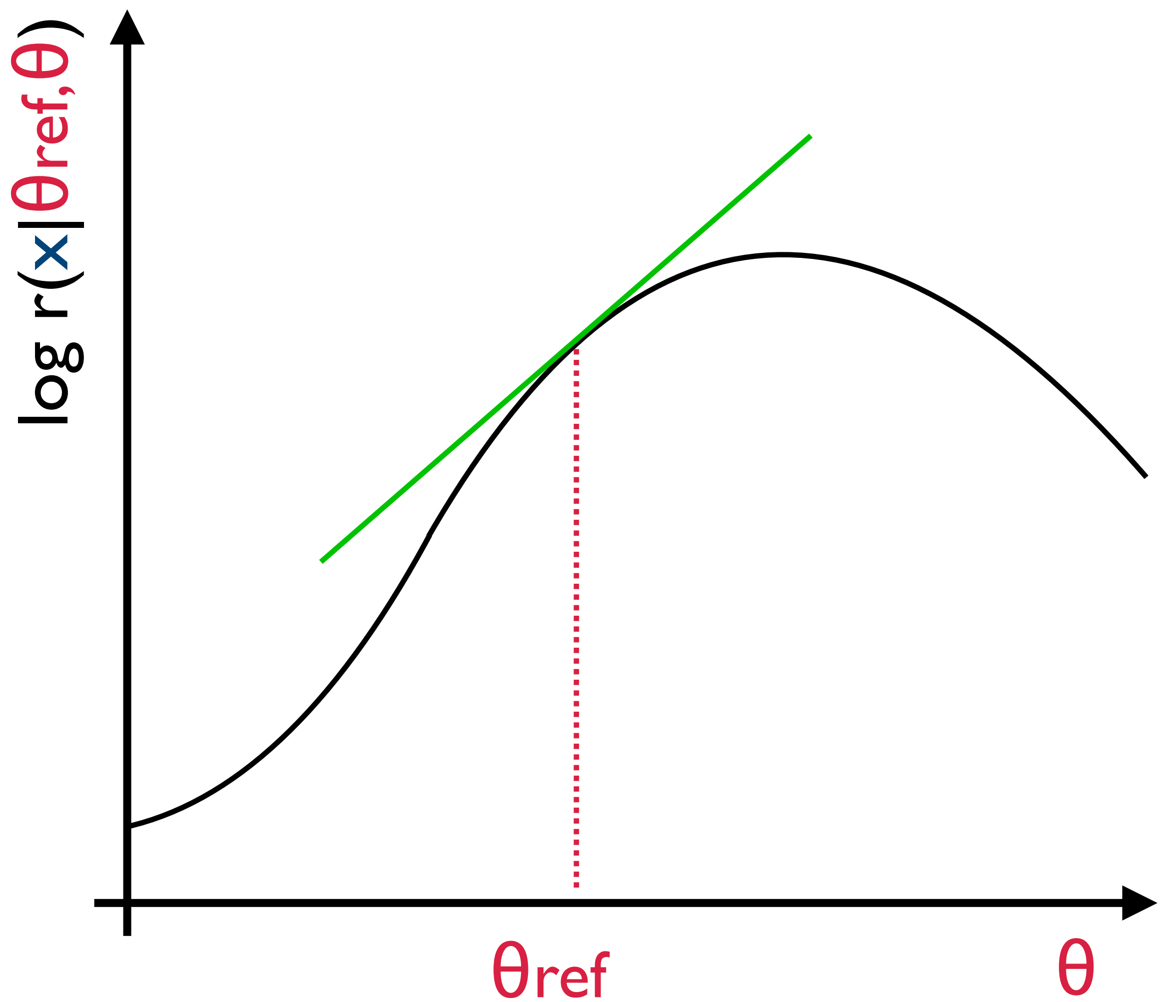
Optimal Observables

Score: $t(x|\theta)=d\log p(x|\theta)/d\theta$

How much does the data x change, when I change my theory θ

Scores as Optimal Observables

- expand LLR around θref: $\log r(x|\theta) = \log r(x|\theta_{\text{ref}}) + t(x)|_{\theta_{\text{ref}}} \cdot (\theta - \theta_{\text{ref}}) + \ldots$



- close to θref
  * score is sufficient statistics
  * knowing $t(x)|_{\theta_{\text{ref}}}$ is as powerful as knowing $r(x|\theta)$
  * $t(x)|_{\theta_{\text{ref}}}$ are optimal observables

- in SMEFT: $t(x)|_{SM}$ is sensitive to interference

[F. Edgeworth 1908; R. Fisher 1925; ...]

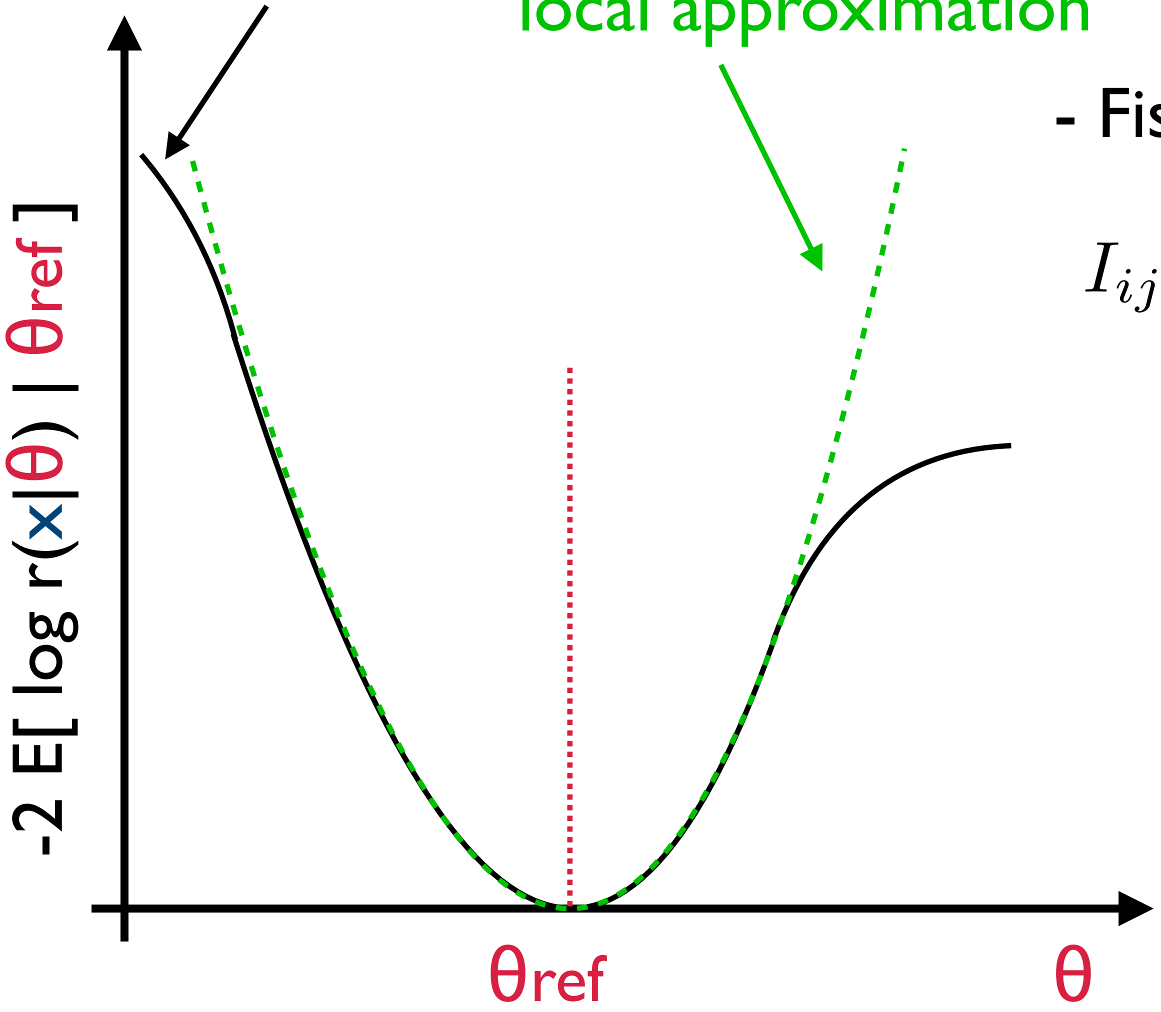
Fisher Information

- expand expected LLR around θ_{ref}:

$$\mathbb{E}[-2\log r_{\text{full}}(x|\theta)|\theta_{\text{ref}}] = I_{ij}(\theta_{\text{ref}}) \times (\theta-\theta_{\text{ref}})_i(\theta-\theta_{\text{ref}})_j + \cdots$$

full results

local approximation



- Fisher Information

$$I_{ij}(\theta_{\text{ref}}) = \mathbb{E}\left[\frac{\partial \log\, p_{\text{full}}(x|\theta)}{\partial\theta_i}\frac{\partial \log\, p_{\text{full}}(x|\theta)}{\partial\theta_j}\middle|\theta_{\text{ref}}\right]$$

$$= \mathcal{L}\frac{\partial_i\sigma(\theta)\partial_j\sigma(\theta)}{\sigma(\theta)} + \frac{\mathcal{L}\,\sigma(\theta)}{n}\sum_{x\sim p(x|\theta_{\text{ref}})} t_i(x)t_j(x)$$

scores

Fisher Information

[C. R. Rao 1945; H. Cramér 1946]

Cramer Rao Bound

The possible precision of a measurement is bounded from below by the inverse of the Fisher Information

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

The Fisher information encodes the maximum sensitivity of observables to theory parameters for a given experiment

Other useful properties of the Fisher Information

* simple: n×n matrix (for n theory parameters)
* independent of parameterizations of x
* covariant under $\theta \to \theta'$
* additive between experiments / phase-space
* easy to include systematics
* defines metric on theory parameter space

See [J. Brehmer, K. Cranmer, FK, T. Plehn 1612.05261]

Outline

Some Statistic Concepts
Likelihoods, Optimal Observables and Fisher Information

Information Geometry in Theory
Parton Level and Matrix Elements

Information Geometry in Reality
Detector Effects and Machine Learning

Summary and Conclusion

Fisher Information at Parton Level

Likelihood Function

- at parton level, we can simply calculate the likelihood function using matrix elements
  * $p(x|\theta) \sim |M(x|\theta)|^2 \times$ phase space factor
  * $r(x|\theta_{ref},\theta) \sim |M(x|\theta)|^2 / |M(x|\theta_{ref})|^2$
- MC generators generate a set of events with momenta x with weights $\Delta\sigma(x|\theta)$
  * $p(x|\theta) = \Delta\sigma(x|\theta) / \sigma(\theta)$
  * $r(x|\theta_{ref},\theta) = \Delta\sigma(x|\theta) / \Delta\sigma(x|\theta_{ref})$

EFTs and Morphing

- EFT Lagrangian: $L = L_{\rm SM} + \sum (f_i/\Lambda^2) \times \mathcal{O}_i$
- the matrix element is a quadratic function of the Wilson coefficients:
  * $\Delta\sigma(x|\theta) = \Delta\sigma_{SM}(x) + \theta\, \Delta\sigma_{dim6}(x) + \theta^2\, \Delta\sigma_{dim8}(x)$
- morphing: evaluate $\Delta\sigma(x|\theta)$ for 3 different θ to obtain quadratic function
  * use reweighing: $\Delta\sigma(x|\theta) = \Delta\sigma(x|\theta_{ref}) * |M(x|\theta)|^2 / |M(x|\theta_{ref})|^2$

Fisher Information

- can be easily calculated using MC:

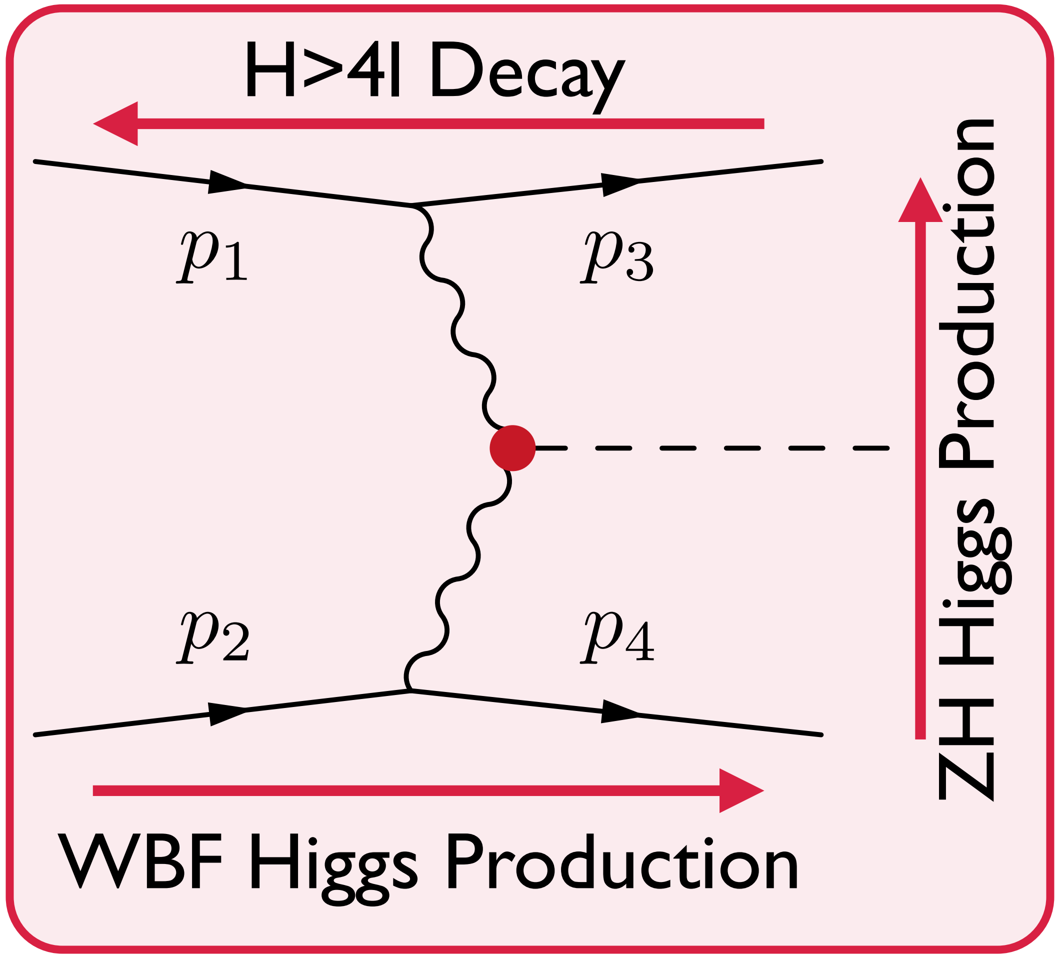
$$I_{ij}(\theta) = \sum_{\rm events} I_{ij}(x|\theta) = \mathcal{L} \sum_{\rm events} \frac{1}{\Delta\sigma(x|\theta)} \frac{\partial \Delta\sigma(x|\theta)}{\partial\theta_i} \frac{\partial \Delta\sigma(x|\theta)}{\partial\theta_j}$$

- possible to include efficiencies, backgrounds with same final state, mass smearing

See [J. Brehmer, K. Cranmer, FK, T. Plehn 1612.05261]

Physics Example

Which observables are sensitive Higgs CP?



Higgs-Gauge Coupling
- WBF and ZH production, H>4l decay
- same hard process
- different final state (charge measurement)

Theory Language:
- dim-6-operators of SMEFT:
- operators such as $L = L_{\text{SM}} + \sum (f_i/\Lambda^2) \times \mathcal{O}_i$

  CP-even: $\mathcal{O}_{WW} \sim (\phi^\dagger \phi)\, W_{\mu\nu}\, W^{\mu\nu}$

  CP-odd: $\mathcal{O}_{W\widetilde{W}} \sim (\phi^\dagger \phi)\, W_{\mu\nu}\, \widetilde{W}^{\mu\nu}$
- goal: measure Wilson coefficients: $\theta = f_i$

Observables: 4 independent 4-momenta

4 C-even and P-even scalar products

2 C-odd and P-even scalar products: p_i

1 C-even and P-odd $\epsilon_{\alpha\beta\gamma\delta}\, p_1^\alpha p_2^\beta p_3^\gamma p_4^\delta$

up to 3 CP sensitive observables

	WBF	ZH	H>4l
		$\Delta p_{T,ll}, \Delta E_{ll}$	$\theta_{1,2}$
	$\Delta\phi^s_{jj}$	$\Delta\phi^s_{ll}$	Φ

[WBF: Hankele, Klamke, Zeppenfeld hep-ph/0609075,
ZH: Christensen, Han, Li 1005.5393,
H>4l: Bolognesi et al. 1208.4018]

i) Total Information

What is the maximum precision to measure theory parameters?

- encoded in Fisher Information $I = \sum_{all\ events} I_{event}$

Example: WBF Higgs Production with H→ττ in SMEFT

$$I_{ij}(\mathbf{0}) = \begin{pmatrix} 715 & -191 & 1 & 0 \\ -191 & 321 & -1 & 0 \\ 1 & -1 & 359 & -81 \\ 0 & 1 & -81 & 23 \end{pmatrix}$$

(columns: f_W, f_{WW}, $f_{W\tilde{W}}$, $\mathrm{Im} f_{WW}$; rows: f_W, f_{WW}, $f_{W\tilde{W}}$, $\mathrm{Im} f_{WW}$)

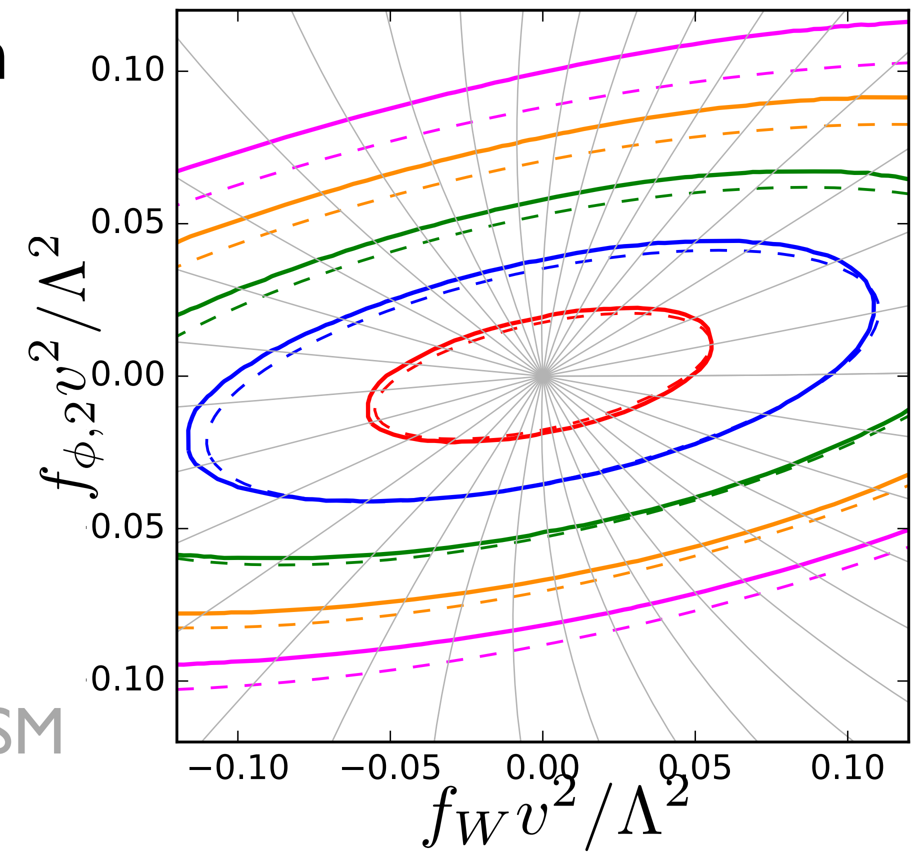
- sensitivity to CP-violating operator
- large mixing between CP-conserving operators
- no mixing between CP-conserving and CP-violating operators
- re-scattering can mimic CP-violation
- best case uncertainties: $\Delta\theta > 1\sqrt{I}$

→ calculate the maximum sensitivity of any LHC process

we assume13TeV LHC, L=100 fb^{-1}, take into account ggF and Z+jets BG,
for more analysis details see 1612.05261, 1712.02350

Geometric Interpretation of Fisher Information

- Distance Measure ~ unlikeliness to measure θ if θref is true 'in sigmas'
- local distance: $d^2 = I_{ij}(\theta_0)(\theta^i - \theta^i_0)(\theta^j - \theta^j_0)$ (dashed)
- global distance: $d = \min_{\theta(s)} \int_{s_a}^{s_b} ds \sqrt{I_{ij} \frac{d\theta_i}{ds} \frac{d\theta_j}{ds}}$ (solid)



Contours of distance d=1,2,3,4,5 from SM

- I_{ij}(θref) only sensitive to linear effects: Δσ(x|θ) ~ θ $\Delta\sigma_{dim6}$(x)
- Information geometry for dim-6 operators $\theta_i = f_i^{d=6} v^2/\Lambda^2$

$I_{ij}(\mathbf{0})$, local distances at SM

always missing

$$\Delta\sigma = \Delta\sigma_{SM} + \sum_i \frac{f_i^{d=6}}{\Lambda^2}\Delta\sigma_i + \sum_i \frac{f_i^{d=6} f_j^{d=6}}{\Lambda^4}\Delta\sigma_{ij} + \sum_i \frac{f_k^{d=8}}{\Lambda^4}\Delta\sigma_k + \mathcal{O}(\Lambda^{-6})$$

$I_{ij}(\theta \neq 0)$, global distances

Difference between local/global distance ⟷ size of $\mathcal{O}(\Lambda^{-4})$ effects

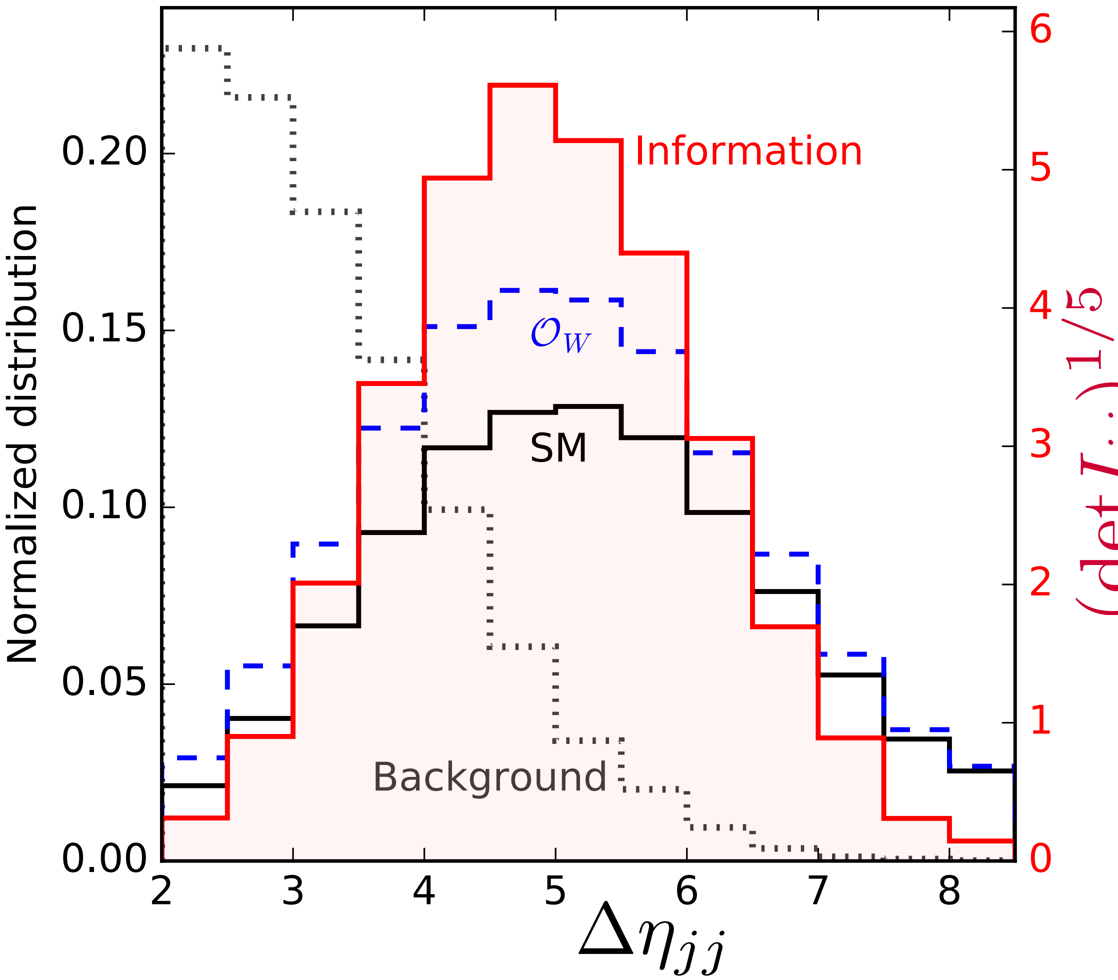
iii) Differential Information

Where in phase space is the information?
- binned kinematic distribution of information

$$I_{bin} = \sum_{events \in bin} I_{event}$$

Example: Jet Rapidity Difference in WBF
- smaller background at large $\Delta\eta_{jj}$
- momentum dependent operator
  * largest effect at medium $\Delta\eta_{jj}$
- strong WBF cuts ($\Delta\eta_{jj}$ > 4.2):
  * lose information of dim-6 operators

→ identify relevant phase-space regions



What are the most powerful observables?

- information of binned kinematic distribution

$$I = \sum_{bins} I_{bin}$$

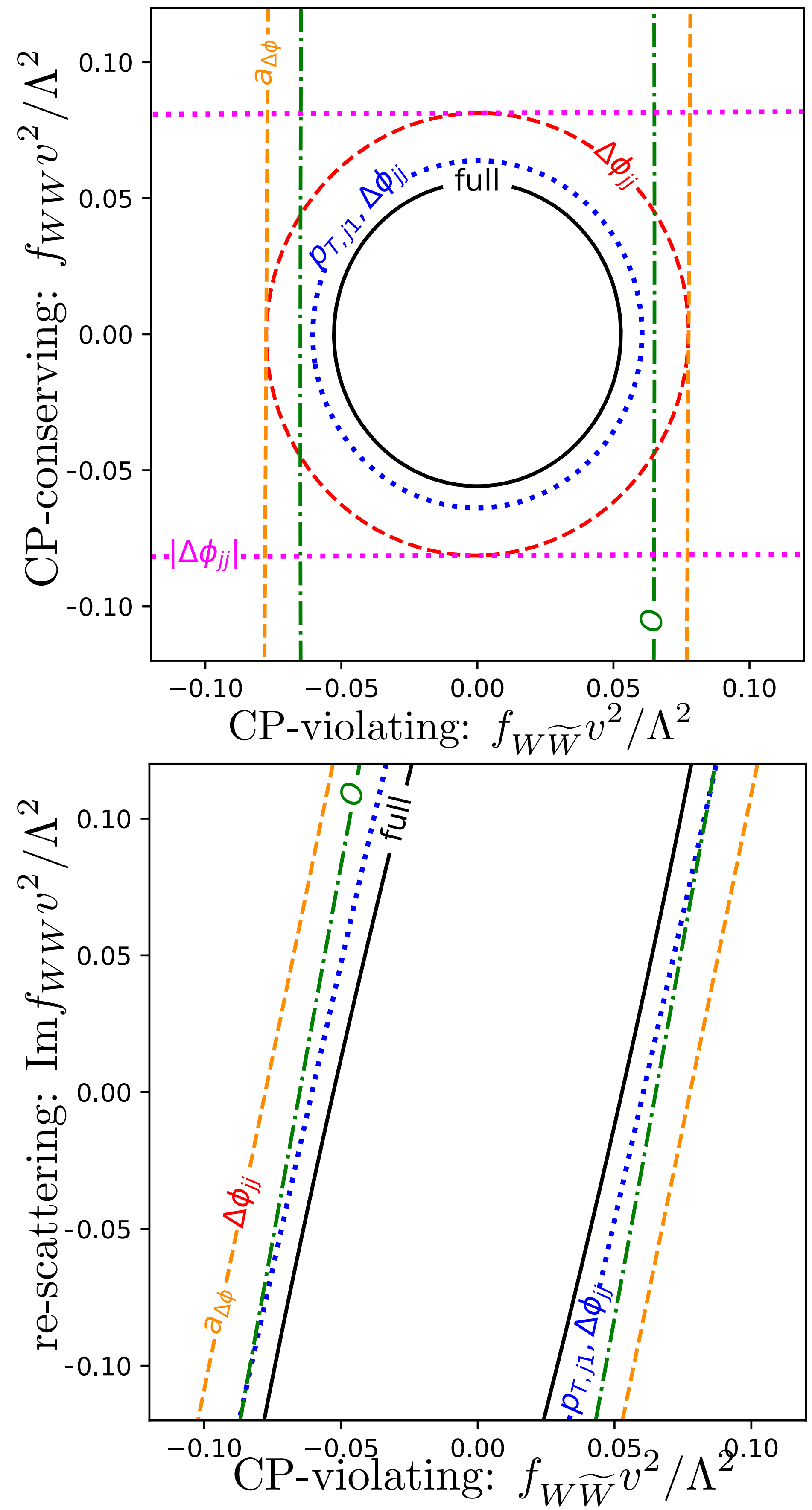
- minimum measurement error $\Delta f \geq 1/\sqrt{I}$

Example: Higgs coupling measurement in WBF

- $|\Delta\phi_{jj}|$ sensitive to CP-conserving physics only
- asymmetry sensitive to CP-violating physics only
- signed $\Delta\phi_{jj}$ probes both
- 2D histogram better, but still not close to **full** information
- re-scattering effects can mimic CP-violation

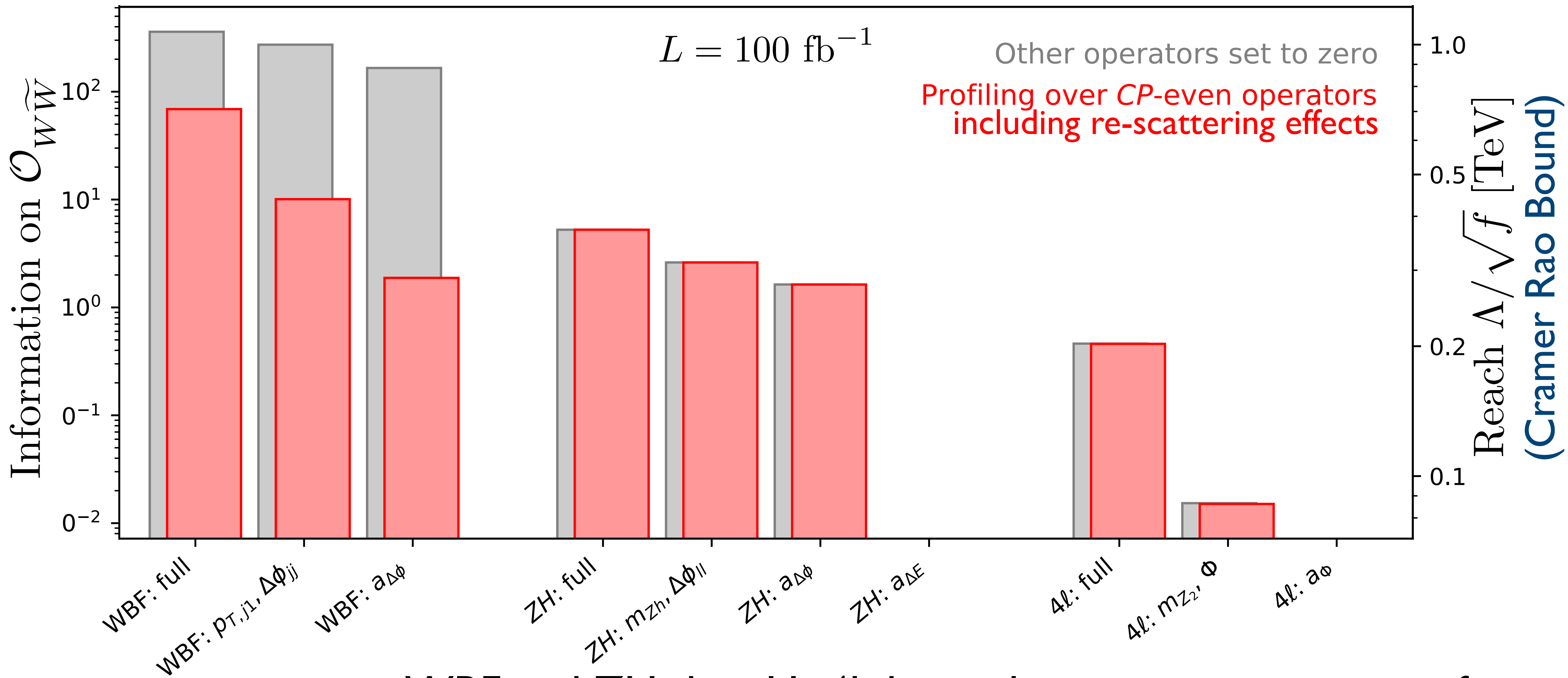
- asymmetry in $\Delta\phi_{jj}$ implies CP violation (in the absence of re-scattering)
- re-scattering small in SM

→ identify most powerful observables



How do histogram-based and multivariate analyses compare?

Example: Information on CP-violating Higgs couplings



- more sensitivity in WBF and ZH than H>4l due to larger momentum transfer
- WBF requires additional theory assumption on re-scattering
- CP-information mostly captured in asymmetry of $\epsilon_{\alpha\beta\gamma\delta}\ p_1^\alpha p_2^\beta p_3^\gamma p_4^\delta \sim \Delta\phi$
- adding momentum transfer measures/multivariate analysis increase sensitivity

→ quantitatively compare histogram-based vs. multivariate analyses

Outline

Some Statistic Concepts
Likelihoods, Optimal Observables and Fisher Information

Information Geometry in Theory
Parton Level and Matrix Elements

Information Geometry in Reality
Detector Effects and Machine Learning

Summary and Conclusion

Simulation in Real Life

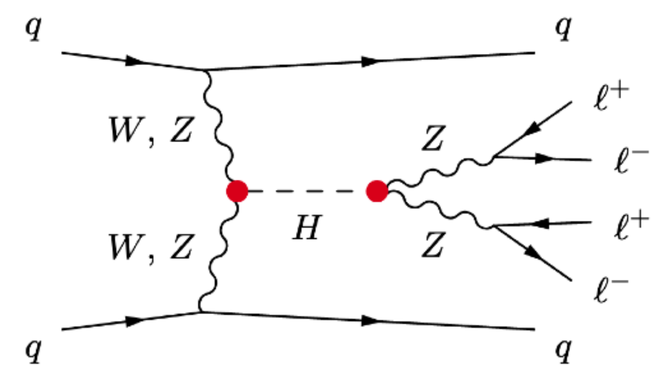
How can we compute the likelihood function?

sophisticated tools model parts of process

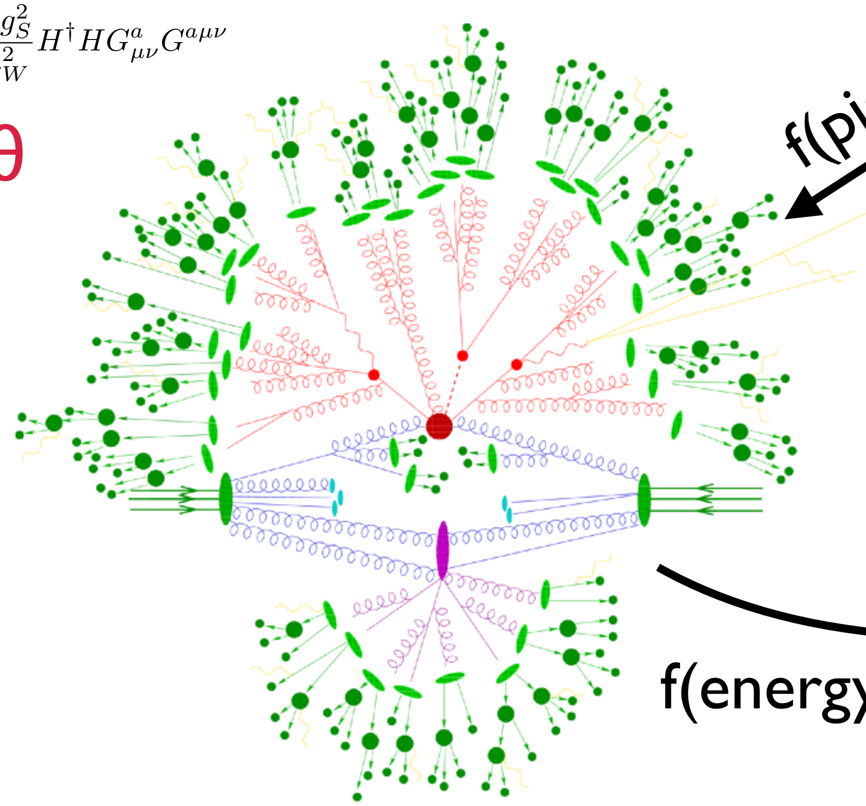
$$
\begin{aligned}
\Delta\mathcal{L}_{SILH} =& \frac{\bar{c}_H}{2v^2}\partial^\mu\left(H^\dagger H\right)\partial_\mu\left(H^\dagger H\right) + \frac{\bar{c}_T}{2v^2}\left(H^\dagger \overleftrightarrow{D^\mu} H\right)\left(H^\dagger \overleftrightarrow{D}_\mu H\right) - \frac{\bar{c}_6\lambda}{v^2}\left(H^\dagger H\right)^3 \\
&+ \left(\left(\frac{\bar{c}_u}{v^2} y_u H^\dagger H \bar{q}_L H^c u_R + \frac{\bar{c}_d}{v^2} y_d H^\dagger H \bar{q}_L H d_R + \frac{\bar{c}_l}{v^2} y_l H^\dagger H \bar{L}_L H l_R\right) + \text{h.c.}\right) \\
&+ \frac{i\bar{c}_W g}{2m_W^2}\left(H^\dagger \sigma^i \overleftrightarrow{D^\mu} H\right)\left(D^\nu W_{\mu\nu}\right)^i + \frac{i\bar{c}_B g'}{2m_W^2}\left(H^\dagger \overleftrightarrow{D^\mu} H\right)\left(\partial^\nu B_{\mu\nu}\right) \\
&+ \frac{i\bar{c}_{HW} g}{m_W^2}\left(D^\mu H\right)^\dagger \sigma^i \left(D^\nu H\right) W^i_{\mu\nu} + \frac{i\bar{c}_{HB} g'}{m_W^2}\left(D^\mu H\right)^\dagger \left(D^\nu H\right) B_{\mu\nu} \\
&+ \frac{\bar{c}_\gamma g'^2}{m_W^2} H^\dagger H B_{\mu\nu} B^{\mu\nu} + \frac{\bar{c}_g g_S^2}{m_W^2} H^\dagger H G^a_{\mu\nu} G^{a\mu\nu}
\end{aligned}
$$

model of nature θ

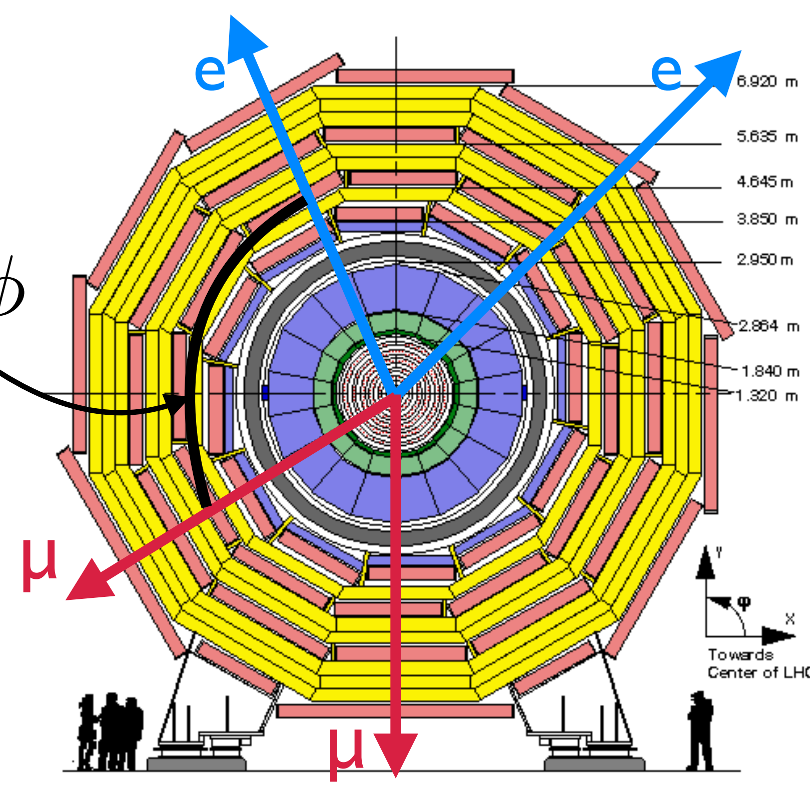
f(Higgs|model)



f(pion|quark)



f(energy deposit|pion)



interesting observables x

Simulation in Real Life

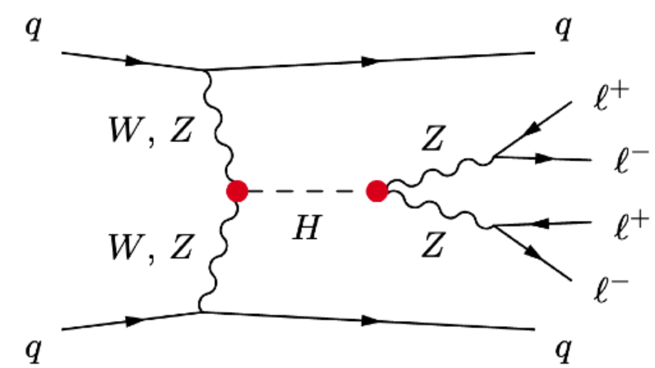
How can we compute the likelihood function?

sophisticated tools model parts of process

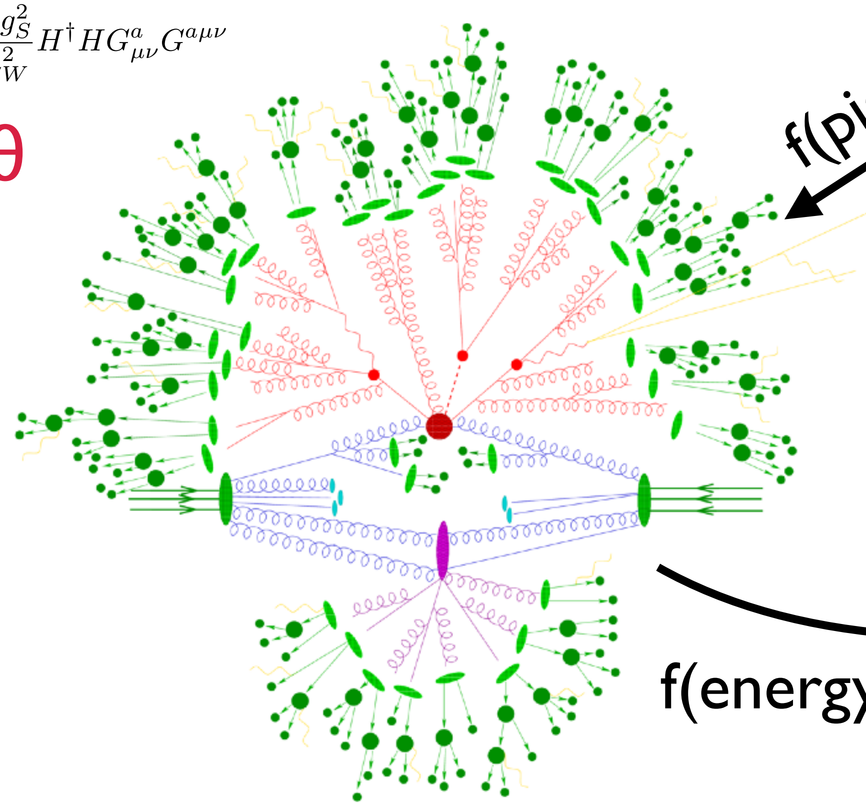
$$
\begin{aligned}
\Delta\mathcal{L}_{SILH} =& \frac{\bar{c}_H}{2v^2}\partial^\mu\left(H^\dagger H\right)\partial_\mu\left(H^\dagger H\right) + \frac{\bar{c}_T}{2v^2}\left(H^\dagger \overleftrightarrow{D^\mu} H\right)\left(H^\dagger \overleftrightarrow{D}_\mu H\right) - \frac{\bar{c}_6\lambda}{v^2}\left(H^\dagger H\right)^3 \\
&+ \left(\left(\frac{\bar{c}_u}{v^2}y_u H^\dagger H \bar{q}_L H^c u_R + \frac{\bar{c}_d}{v^2}y_d H^\dagger H \bar{q}_L H d_R + \frac{\bar{c}_l}{v^2}y_l H^\dagger H \bar{L}_L H l_R\right) + \text{h.c.}\right) \\
&+ \frac{i\bar{c}_W g}{2m_W^2}\left(H^\dagger \sigma^i \overleftrightarrow{D^\mu} H\right)\left(D^\nu W_{\mu\nu}\right)^i + \frac{i\bar{c}_B g'}{2m_W^2}\left(H^\dagger \overleftrightarrow{D^\mu} H\right)\left(\partial^\nu B_{\mu\nu}\right) \\
&+ \frac{i\bar{c}_{HW} g}{m_W^2}\left(D^\mu H\right)^\dagger \sigma^i \left(D^\nu H\right) W^i_{\mu\nu} + \frac{i\bar{c}_{HB} g'}{m_W^2}\left(D^\mu H\right)^\dagger \left(D^\nu H\right) B_{\mu\nu} \\
&+ \frac{\bar{c}_\gamma g'^2}{m_W^2} H^\dagger H B_{\mu\nu} B^{\mu\nu} + \frac{\bar{c}_g g_S^2}{m_W^2} H^\dagger H G^a_{\mu\nu} G^{a\mu\nu}
\end{aligned}
$$

model of nature θ

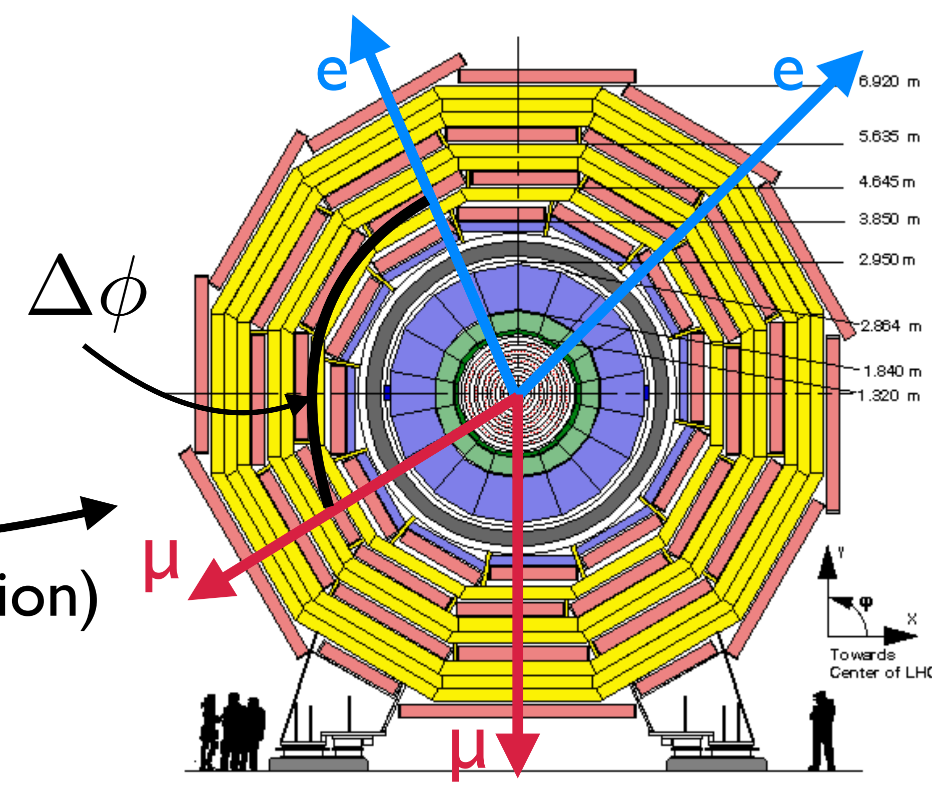
f(Higgs|model)



f(pion|quark)



f(energy deposit|pion)



interesting observables x

we obtain events distributed according to f(x|θ)

we cannot calculate f(x|θ)

Simulation in Real Life

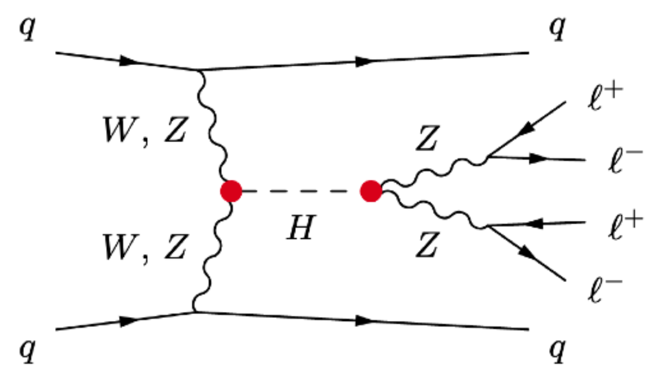
How can we compute the likelihood function?

sophisticated tools model parts of process

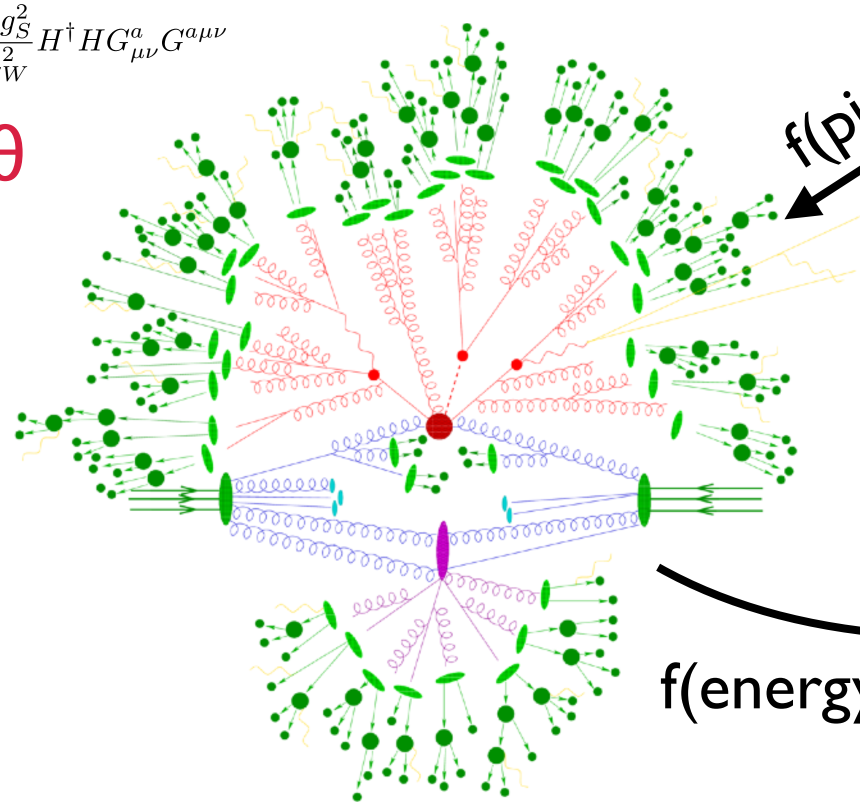
$$\begin{aligned}\Delta\mathcal{L}_{SILH} = &\frac{\bar{c}_H}{2v^2}\partial^\mu\left(H^\dagger H\right)\partial_\mu\left(H^\dagger H\right) + \frac{\bar{c}_T}{2v^2}\left(H^\dagger \overleftrightarrow{D^\mu} H\right)\left(H^\dagger \overleftrightarrow{D}_\mu H\right) - \frac{\bar{c}_6\lambda}{v^2}\left(H^\dagger H\right)^3 \\ &+ \left(\left(\frac{\bar{c}_u}{v^2} y_u H^\dagger H \bar{q}_L H^c u_R + \frac{\bar{c}_d}{v^2} y_d H^\dagger H \bar{q}_L H d_R + \frac{\bar{c}_l}{v^2} y_l H^\dagger H \bar{L}_L H l_R\right) + \text{h.c.}\right) \\ &+ \frac{i\bar{c}_W g}{2m_W^2}\left(H^\dagger \sigma^i \overleftrightarrow{D^\mu} H\right)\left(D^\nu W_{\mu\nu}\right)^i + \frac{i\bar{c}_B g'}{2m_W^2}\left(H^\dagger \overleftrightarrow{D^\mu} H\right)\left(\partial^\nu B_{\mu\nu}\right) \\ &+ \frac{i\bar{c}_{HW} g}{m_W^2}\left(D^\mu H\right)^\dagger \sigma^i \left(D^\nu H\right) W^i_{\mu\nu} + \frac{i\bar{c}_{HB} g'}{m_W^2}\left(D^\mu H\right)^\dagger \left(D^\nu H\right) B_{\mu\nu} \\ &+ \frac{\bar{c}_\gamma g'^2}{m_W^2} H^\dagger H B_{\mu\nu} B^{\mu\nu} + \frac{\bar{c}_g g_S^2}{m_W^2} H^\dagger H G^a_{\mu\nu} G^{a\mu\nu}\end{aligned}$$

model of nature θ

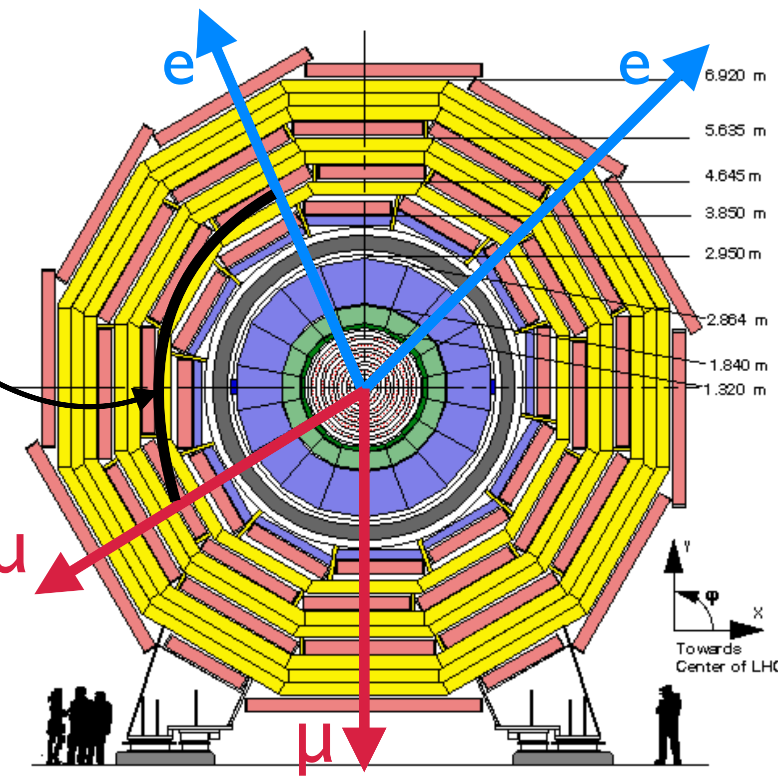
f(Higgs|model)



f(pion|quark)



f(energy deposit|pion)



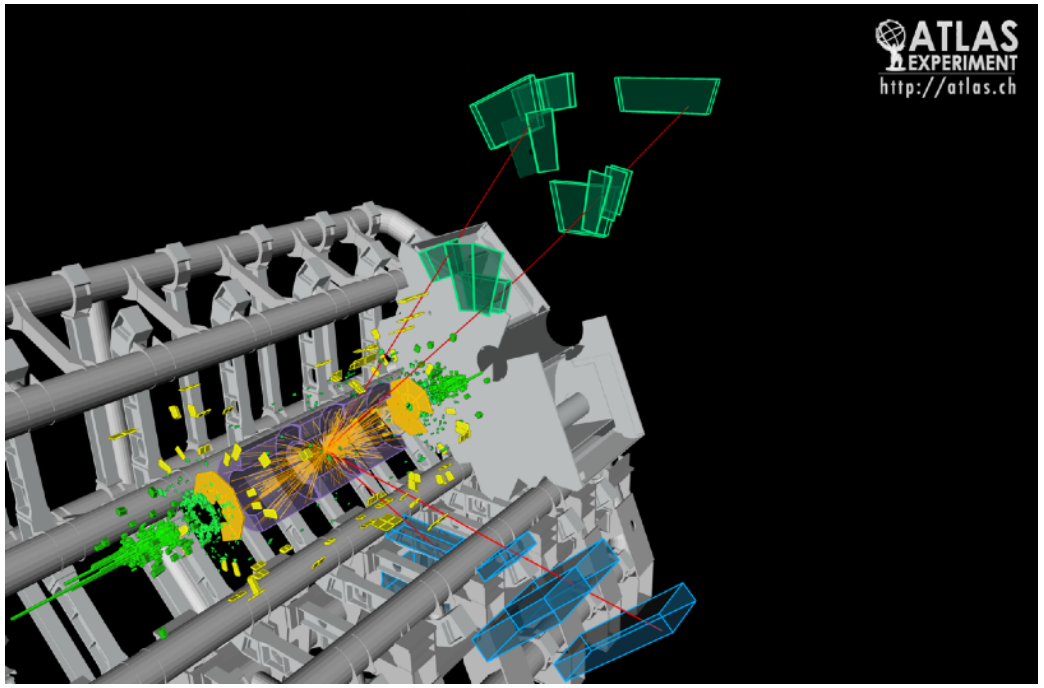
we obtain events distributed according to f(x|θ)

we cannot calculate f(x|θ)

interesting observables x

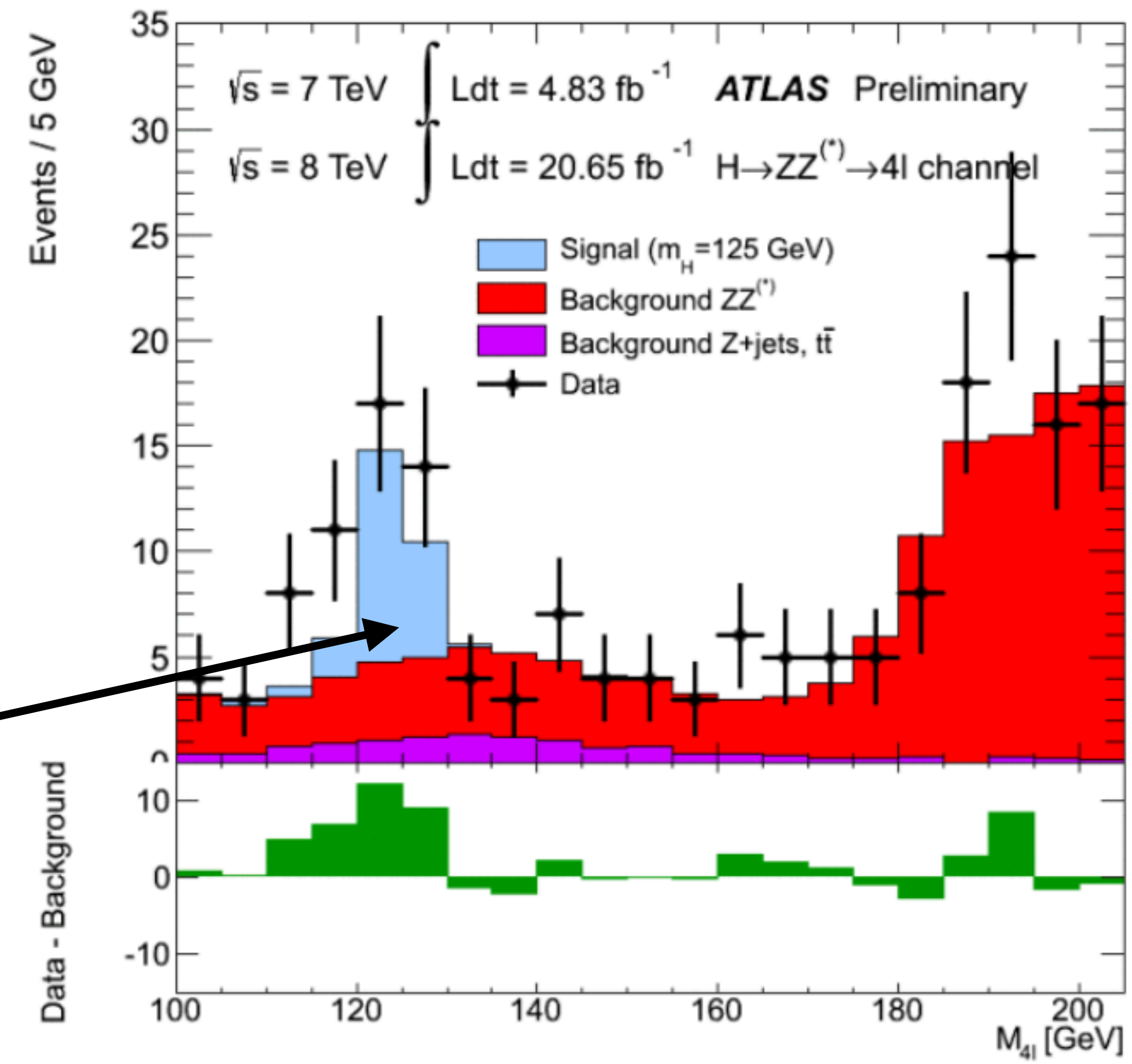
learn likelihood from simulated data - likelihood free inference

Traditional Method: Histograms



complex data: 10^8 sensors

analyze one carefully chosen summary statistics



invariant mass: m(4μ)

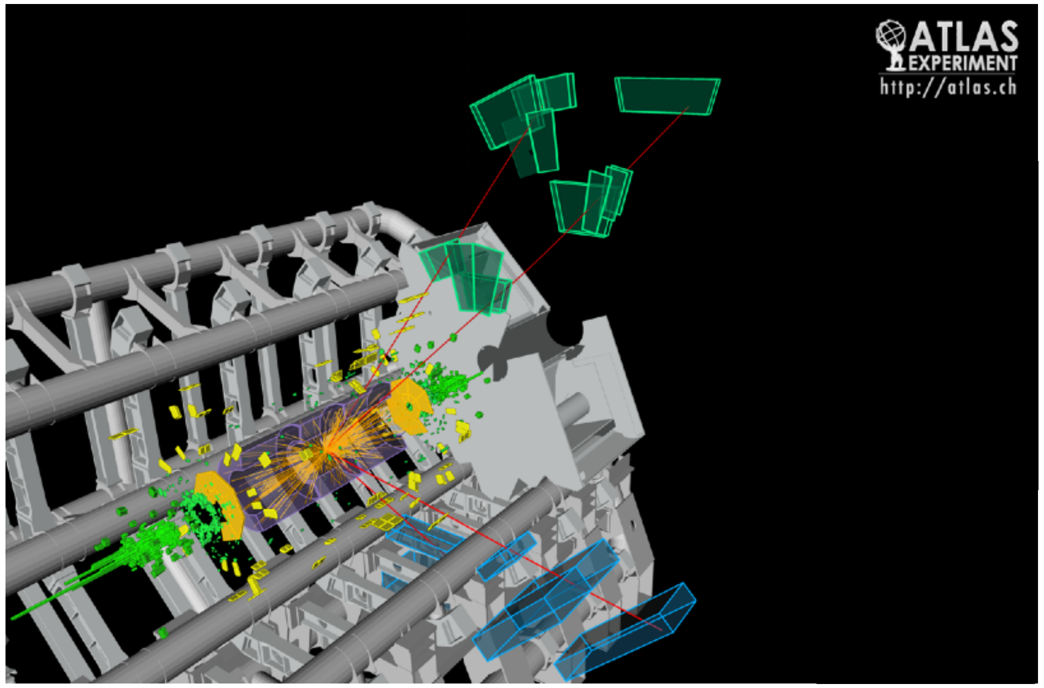
+ very easy

- ignore all the information in rest of the data

histogram ↔ $p(x|\theta)$

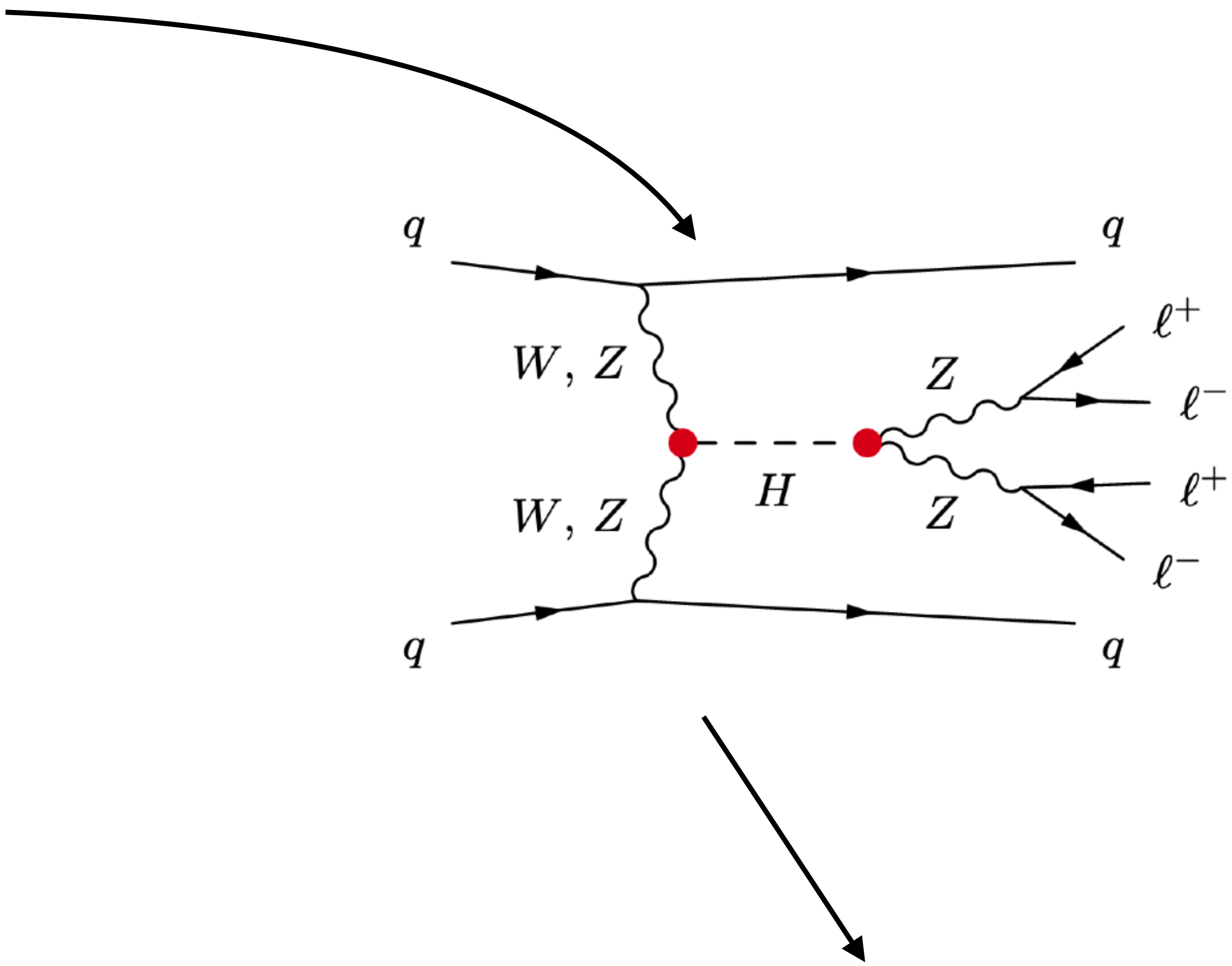
Inference in Real Life

Multivariate Method: Matrix Elements



complex data: 10^8 sensors

match to parton level event



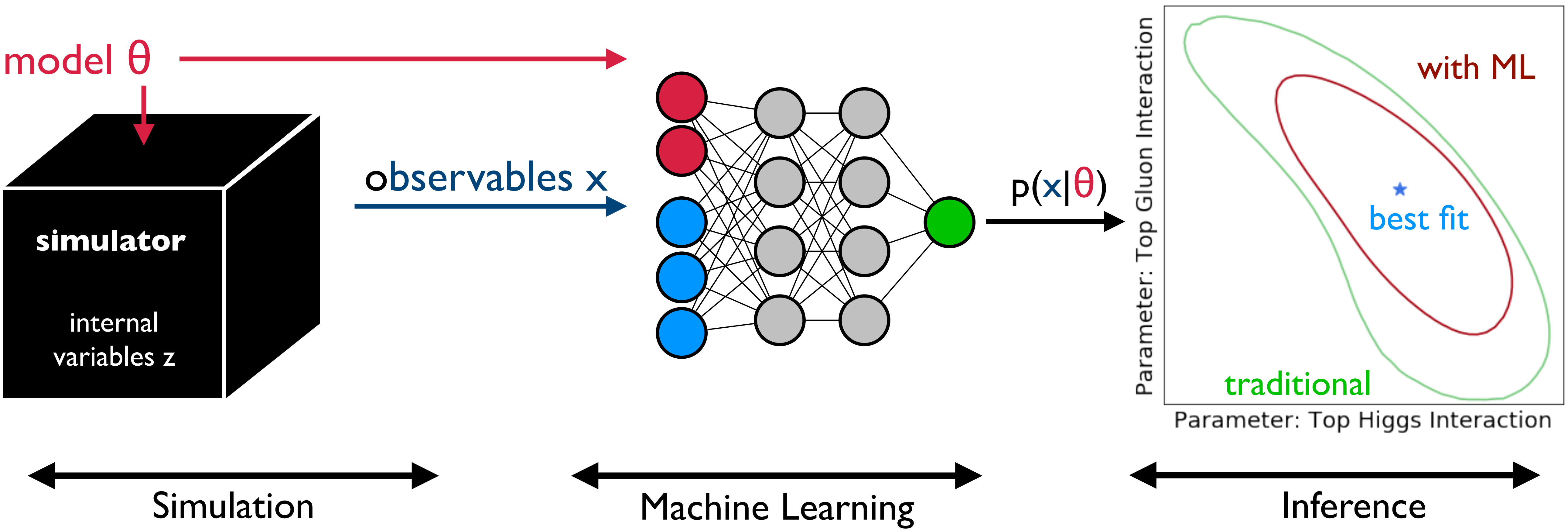
\+ works great at parton level:
  * S' vs S is easy

\- requires approximations in reality
  * S vs BG can be hard

calculate matrix elements
$p(x|\theta) \sim |M(x|\theta)|^2$

Inference in Real Life

Multivariate Method: Machine Learning

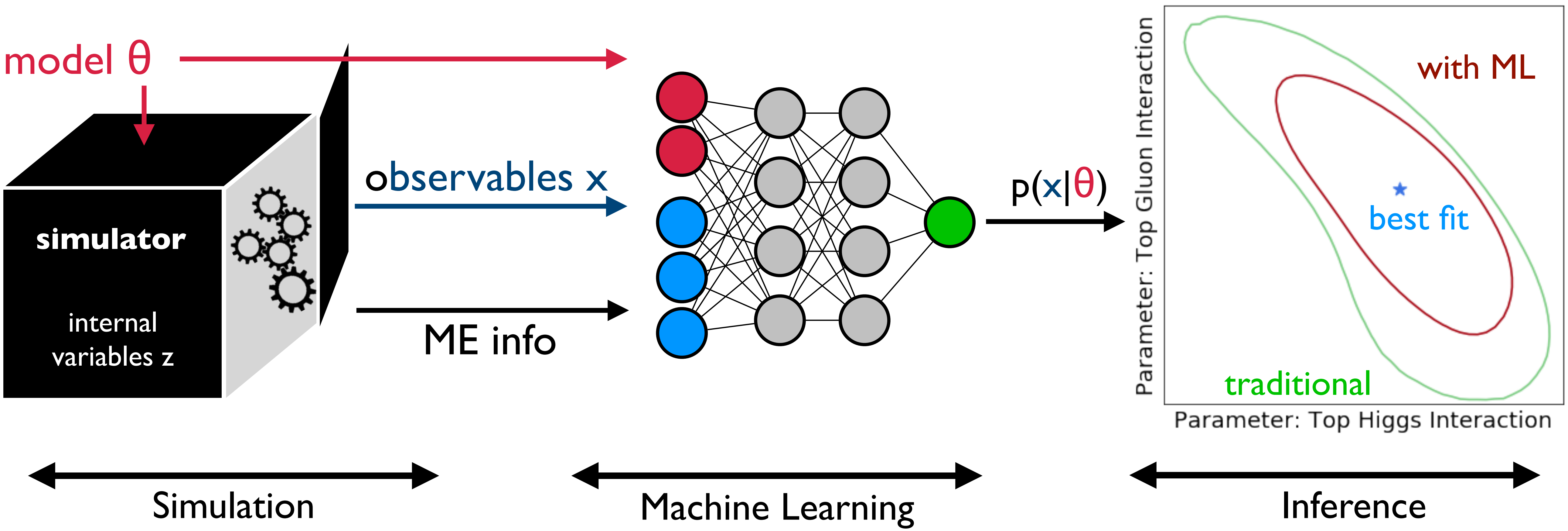


+ works great for S vs BG

- struggles with S' vs S
  * large number of S'
  * very similar S', S

Inference in Real Life

Multivariate Method: The MadMiner Approach



power of machine learning

physics insight of matrix element information

MadMiner

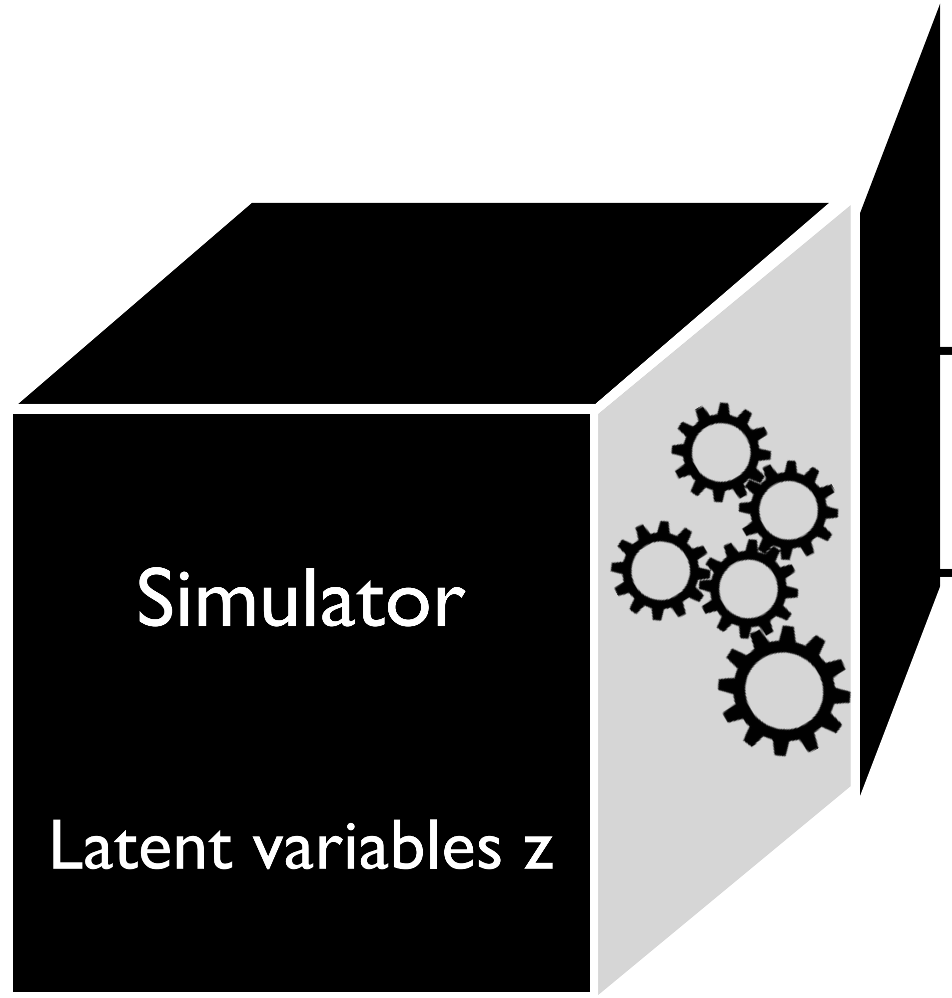
[J. Brehmer, K. Cranmer, G. Louppe, J. Pavez 1805.00013, 1805.00020,1805.12244]

[J. Brehmer, FK, I. Espejo, K. Cranmer 1907.10621]

The MadMiner Approach

Theory Parameters

θ



Set of Events:

Observables

x

matrix elements

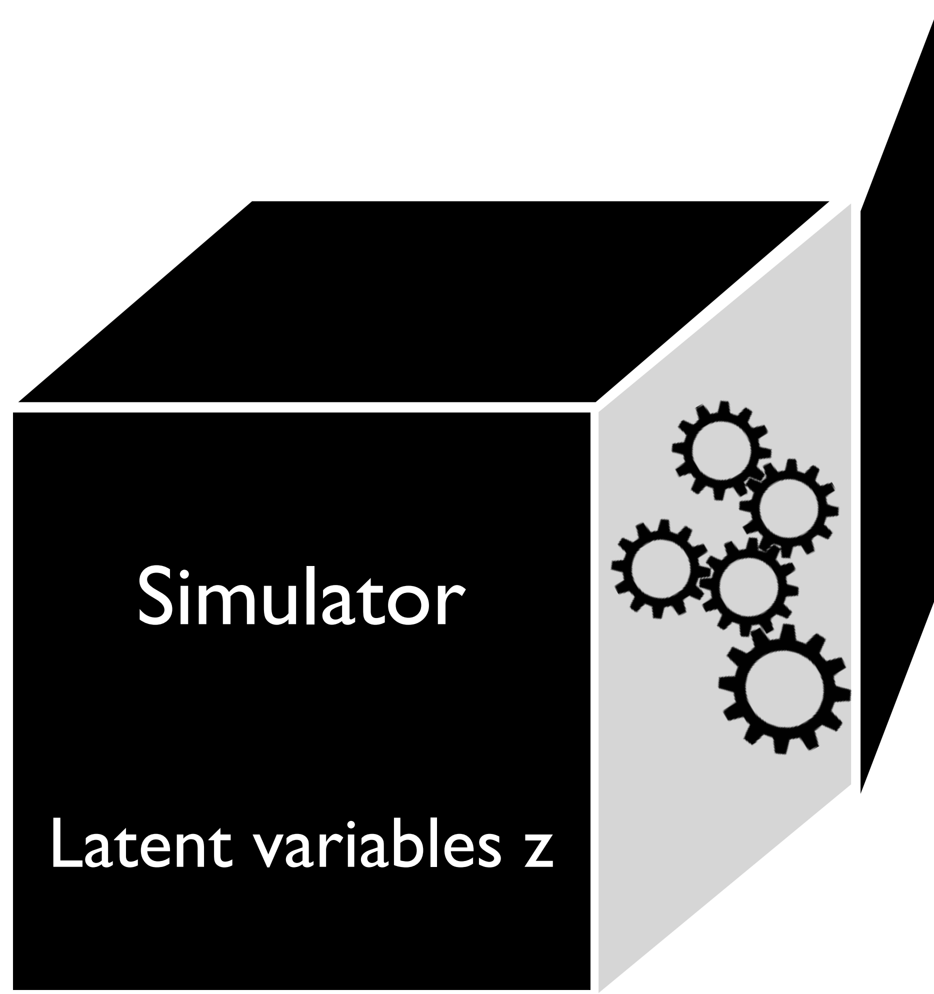
$|\mathcal{M}(z_p|\theta)|^2$

joint likelihood ratio

$$r(x, z|\theta) = \frac{|\mathcal{M}(z_p|\theta)|^2}{|\mathcal{M}(z_p|\theta_{\text{ref}})|^2}$$

Theory Parameters

θ



Set of Events:

Observables

x

matrix elements

$|\mathcal{M}(z_p|\theta)|^2$

joint likelihood ratio

$$r(x, z|\theta) = \frac{|\mathcal{M}(z_p|\theta)|^2}{|\mathcal{M}(z_p|\theta_{\text{ref}})|^2}$$

Problem Setup:

We want: Likelihood-ratio $r(x|\theta)$

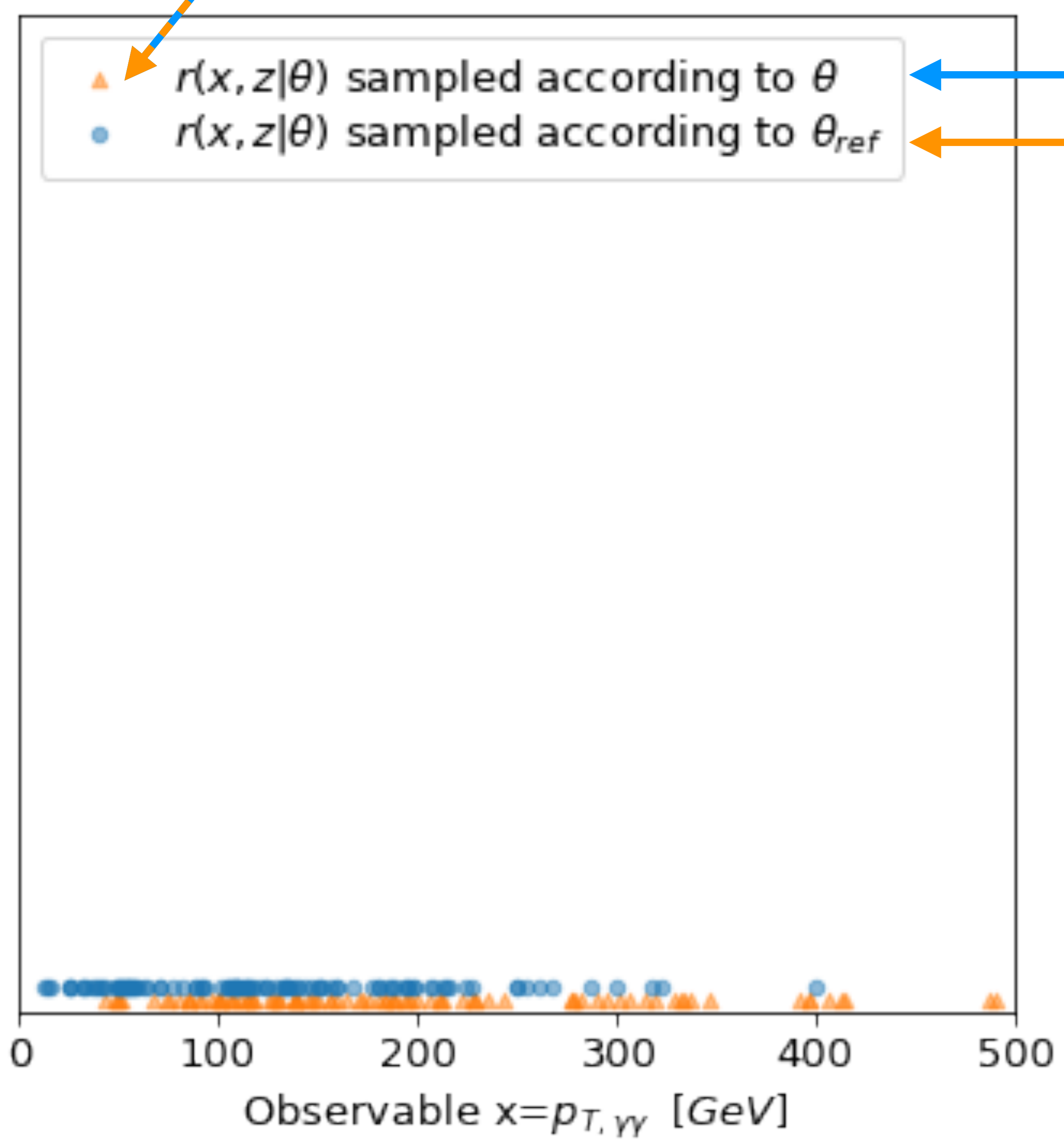
data ↗ ↖ theory

Generators give us: Joint Likelihood-ratio $r(x,z|\theta)$

The MadMiner Approach

How is Likelihood Estimated?

- consider two models BSM (θ=1) vs SM (θ_{ref}=0)
  * sampled events according to p(x|SM) , p(x|BSM)



sampled events according to p(x|SM)

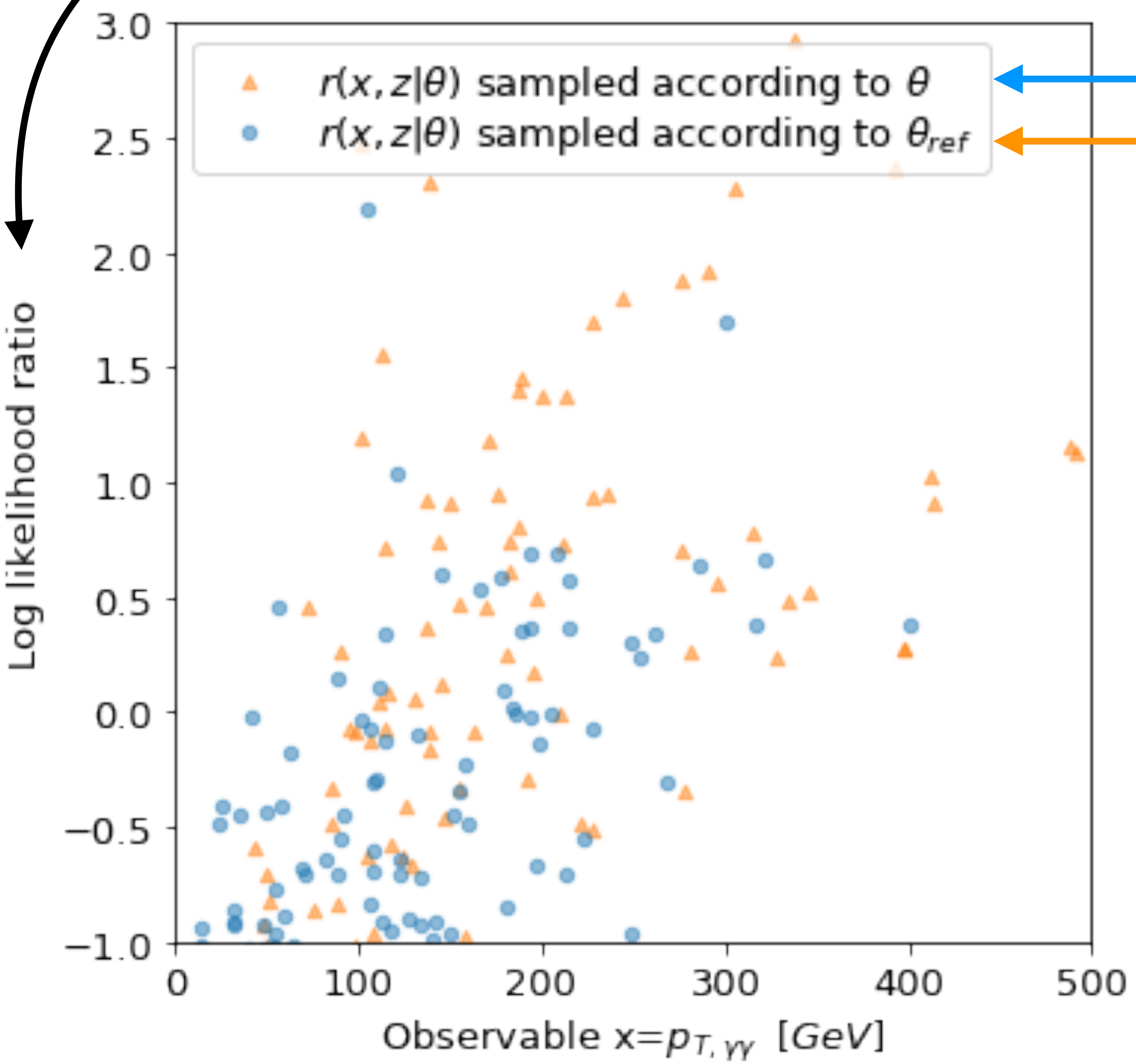
sampled events according to p(x|BSM)

- more BSM points at high pT

The MadMiner Approach

How is Likelihood Estimated?

- consider two models BSM (θ=1) vs SM (θ_{ref}=0)
  * sampled events according to p(x|SM) , p(x|BSM)
  * y-axis: joint likelihood ratio r(x,z|BSM,SM)



sampled events according to p(x|SM)

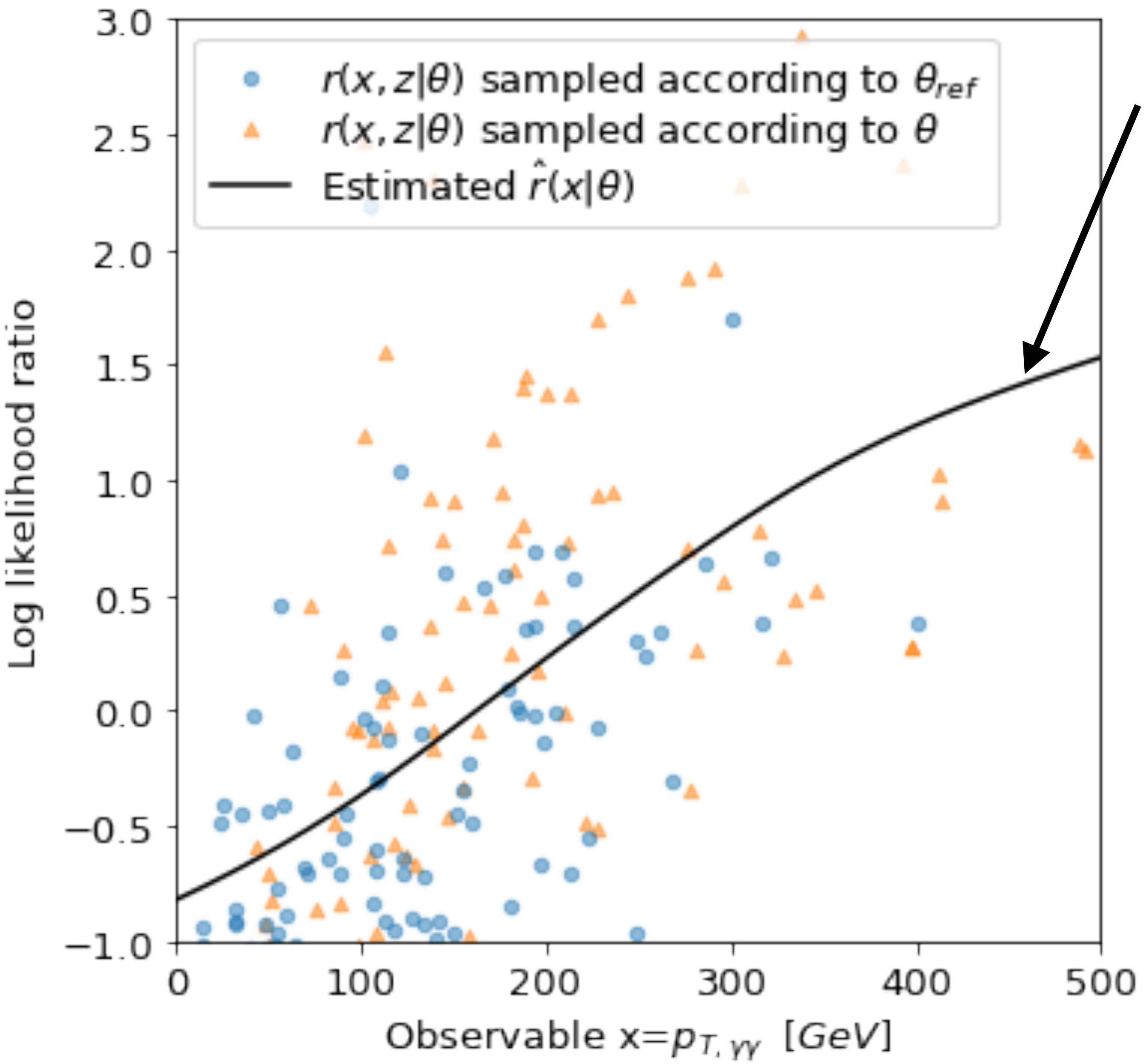
sampled events according to p(x|BSM)

- more BSM points at high pT
  → BSM more likely than SM at high pT

The MadMiner Approach

How is Likelihood Estimated?

- consider two models BSM (θ=1) vs SM (θ_{ref}=0)
  * sampled events according to p(x|SM) , p(x|BSM)
  * y-axis: joint likelihood ratio r(x,z|BSM,SM)



- likelihood ratio r(x|SM,BSM) is solution to minimization problem
- define functional (loss function)

$$L\left[\hat{r}(x|\theta)\right] \sim \sum \left|r(x|\theta) - r(x,z|\theta)\right|^2$$

and minimize it

$$r(x|\theta) = \underset{\hat{r}(x|\theta)}{\arg\min}\, L_r\left[\hat{r}(x|\theta)\right]$$

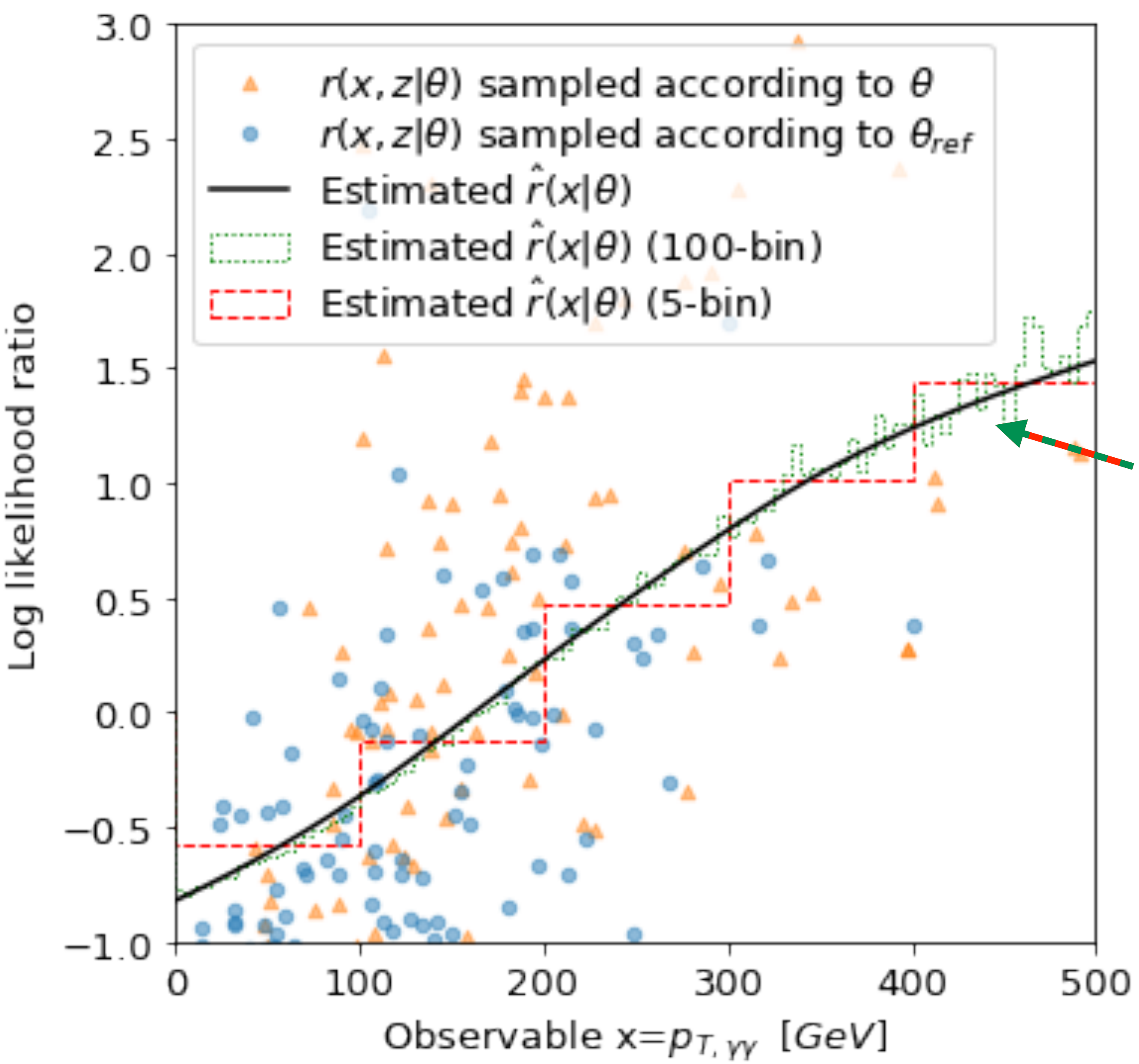
neural network — stochastic gradient descent — loss function

A sufficiently expressive network, efficiently trained in this way with enough data will learn the likelihood ratio function r(x|θ)!

[Proof: J. Brehmer, K. Cranmer, G. Louppe, J. Pavez 1805.00020]

How is Likelihood Estimated?

- consider two models BSM (θ=1) vs SM (θ_{ref}=0)
  - sampled events according to p(x|SM) , p(x|BSM)
  - y-axis: joint likelihood ratio r(x,z|BSM,SM)



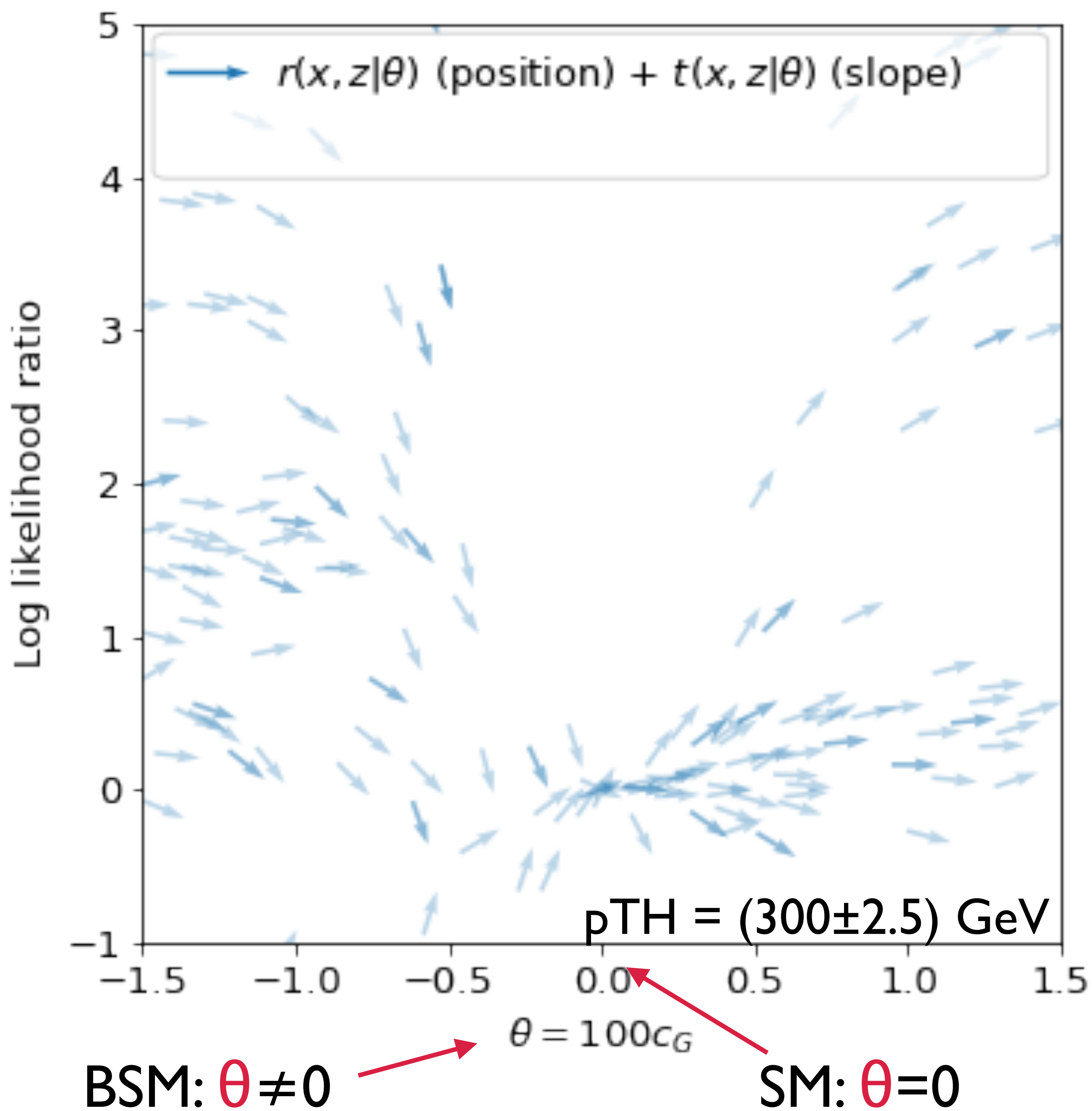
- estimated r(x|SM,BSM)
  → obtain "best fit"

- LLR obtained using histogram
  → agrees well :)
  → "continuum limit of for large # of bins"

The MadMiner Approach

Useful: the Score!

- knowing the derivative often helps: "*How does data x change, when theory θ is changed?*"
  → Score: $t(x|\theta)=d\log p(x|\theta)/d\theta$



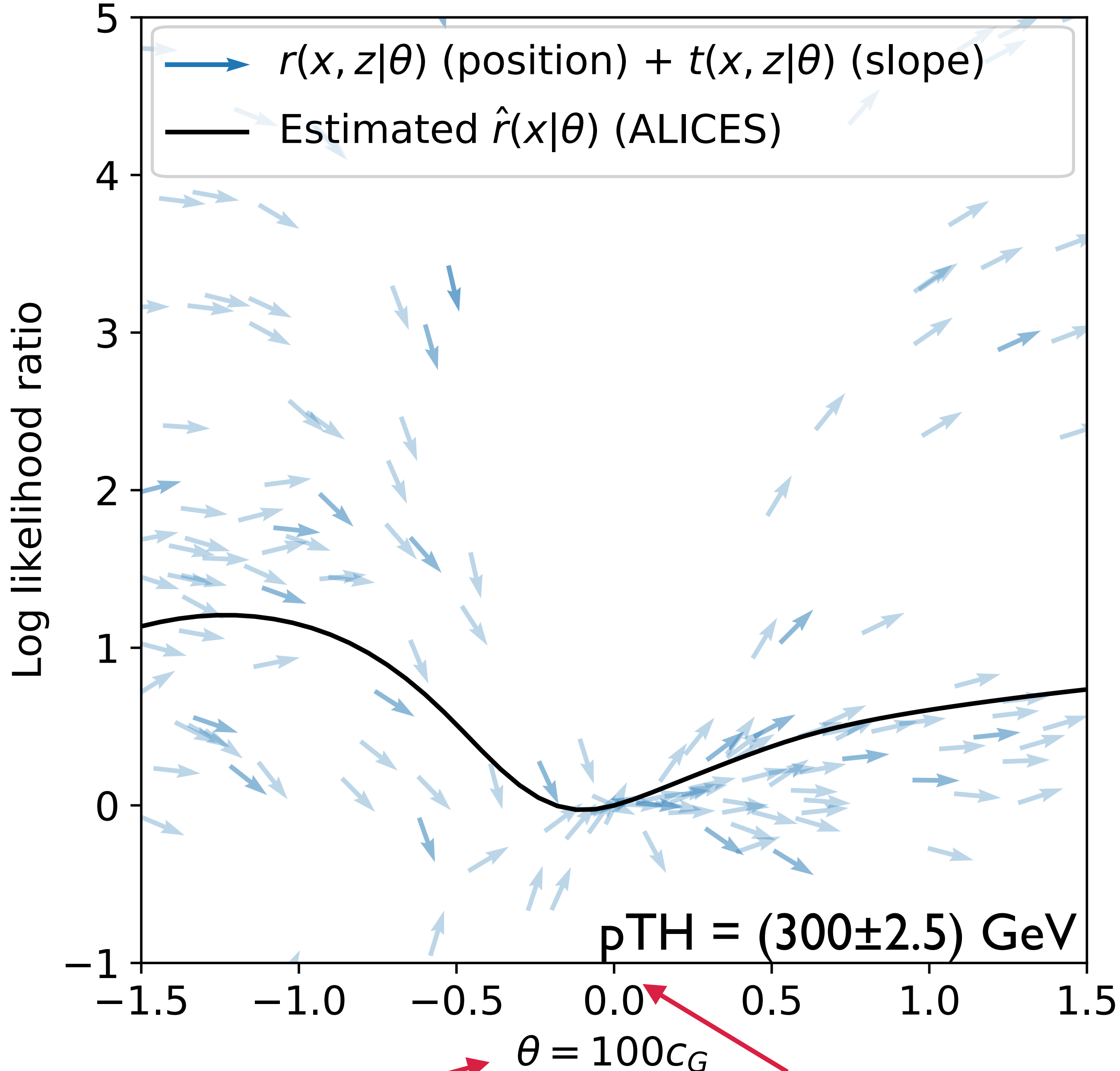
- position: joint likelihood ratio
  $\log r(x,z|\theta)$ as function of θ
- slope: joint score
  $t(x,z|\theta)$ as function of θ

The MadMiner Approach

Useful: the Score!

- knowing the derivative often helps: "*How does data x change, when theory θ is changed?*"

→ Score: $t(x|\theta) = d\log p(x|\theta)/d\theta$



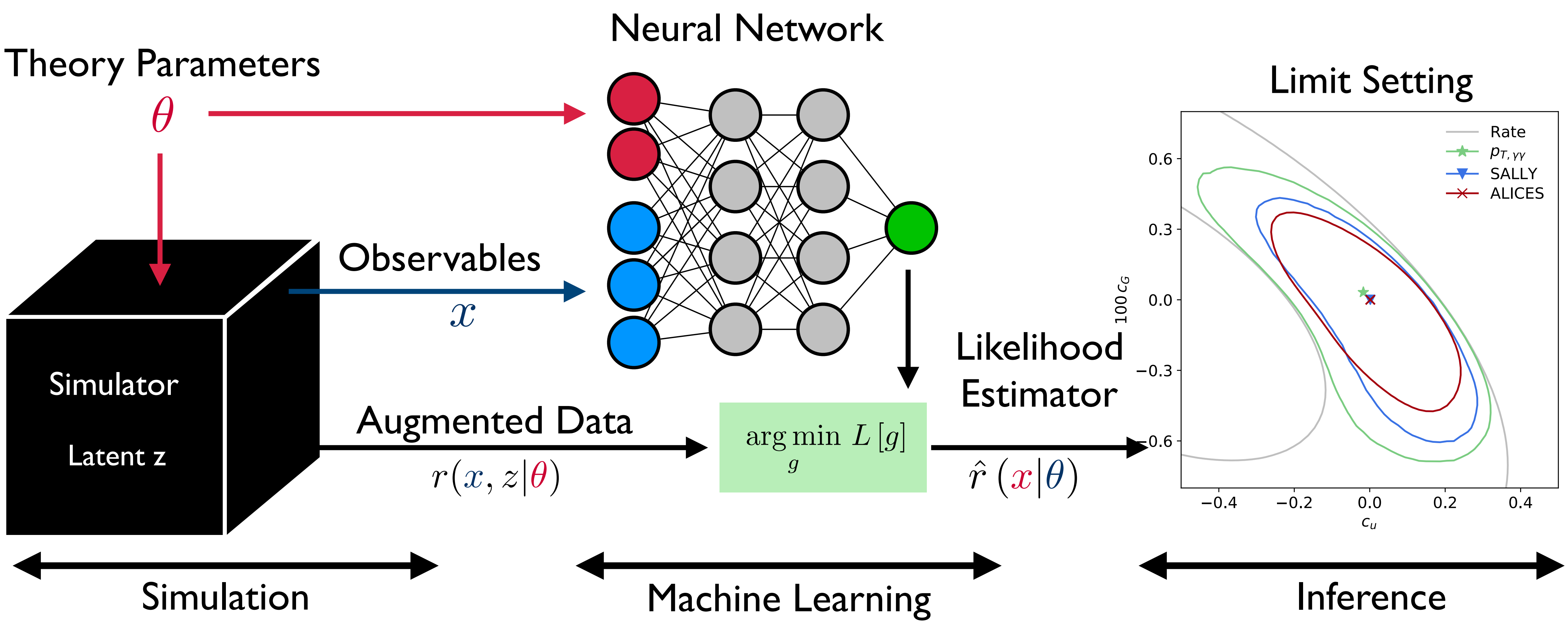
BSM: $\theta \neq 0$ SM: $\theta = 0$

- position: joint likelihood ratio
  $\log r(x,z|\theta)$ as function of θ
- slope: joint score
  $t(x,z|\theta)$ as function of θ

↓ obtain "best fit"

- estimate $r(x, \theta)$ and $t(x, \theta)$

A short summary



Likelihood Ratio Estimator (ALICES)

- learn LLR as function of x and θ
- use r(x,z|θ) and t(x,z|θ) as input

$$\mathrm{NN}:(x,\theta)\rightarrow\hat{r}(x|\theta)\approx p(x|\theta)/p(x|\theta_{\mathrm{ref}})$$

Score Estimator (SALLY)

- learn score as function of x at θ_{ref}
- t(x,z|θ_{ref}) as input

$$\mathrm{NN}:x\rightarrow\hat{t}(x)\approx\nabla_\theta\log(x|\theta)\big|_{\theta_{\mathrm{ref}}}$$

MadMiner: The Tool

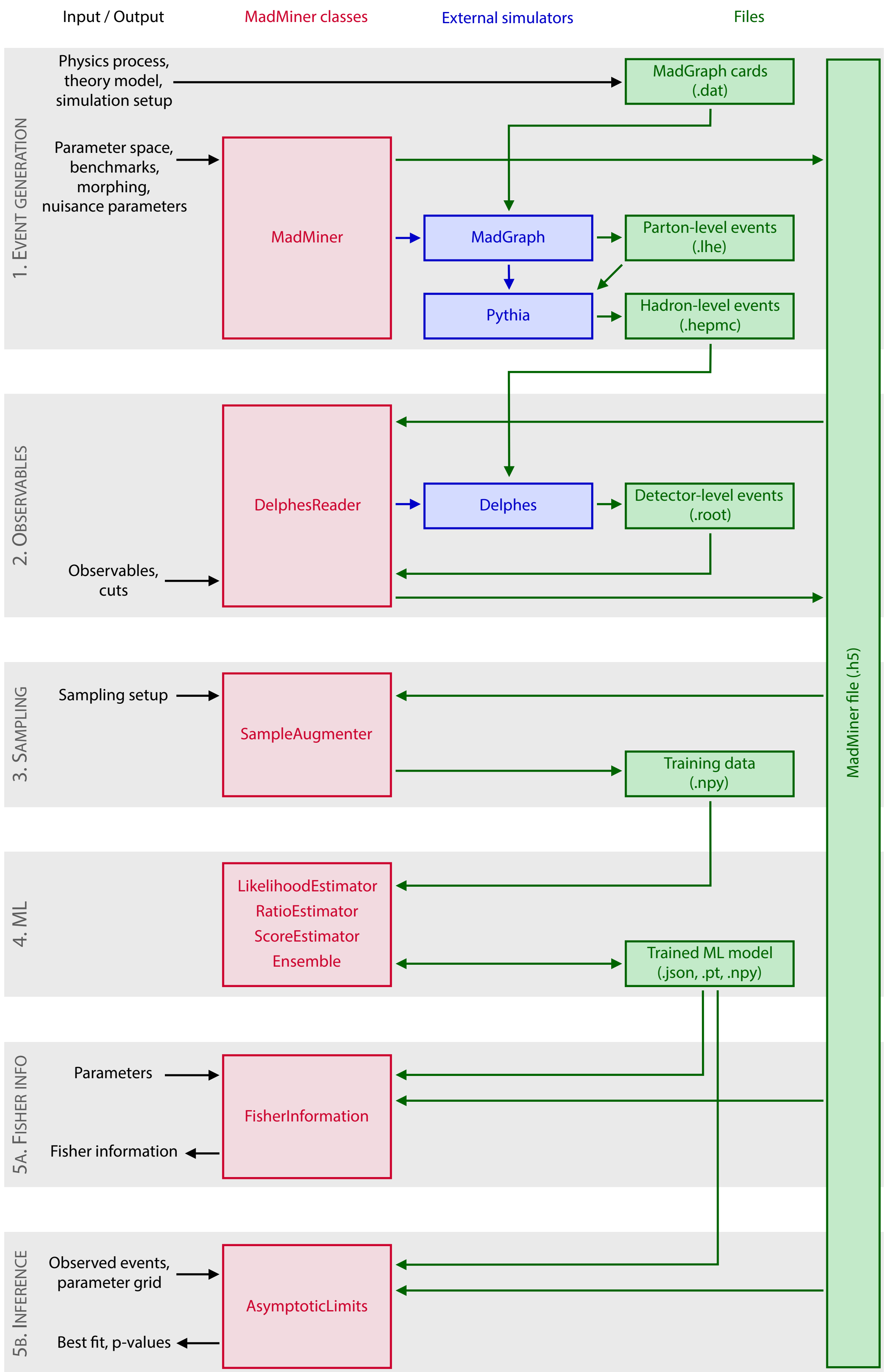
MadMiner [J. Brehmer, FK, I. Espejo, K. Cranmer 1907.10621]

- automizes these techniques
- straightforward to apply them to LHC problems

- out of the box: Pheno-level analysis
  * MadGraph, Pythia, Delphes
  * backgrounds
  * PDF/scale uncertainties
  * ML uncertainties
  * morphing
  * many inference techniques (SALLY, ALICES …)

- scalable to state-of-the-art experimental tools

- python package
  * modular interface
  * extensive documentation
  * on GitHub
    github.com/diana-hep/madminer
  * easy to install
    `pip install madminer`



MadMiner: The Tool

MadMiner Team

Johann Brehmer

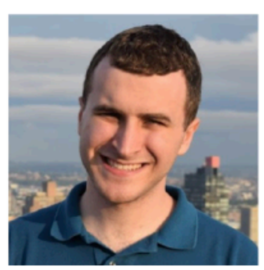
Kyle Cranmer

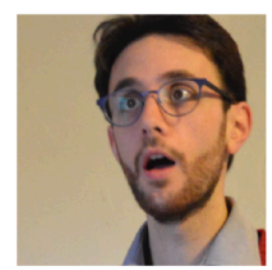
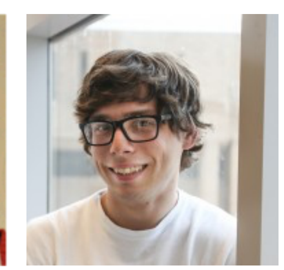
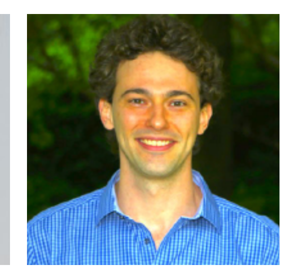
Irina Espejo

Felix Kling

Already used by many people






Thank you a lot for testing!

Example 1: Probing SMEFT in ttH

ttH production in SMEFT

[J. Brehmer, FK, I. Espejo, K. Cranmer 1907.10621]

- signal: fully leptonic tth $\quad pp \to t\bar{t}h \to \left(b\ell^+\right)\left(\bar{b}\ell^-\right)(\gamma\gamma)E_T^{\text{miss}}$
- 3 operators considered

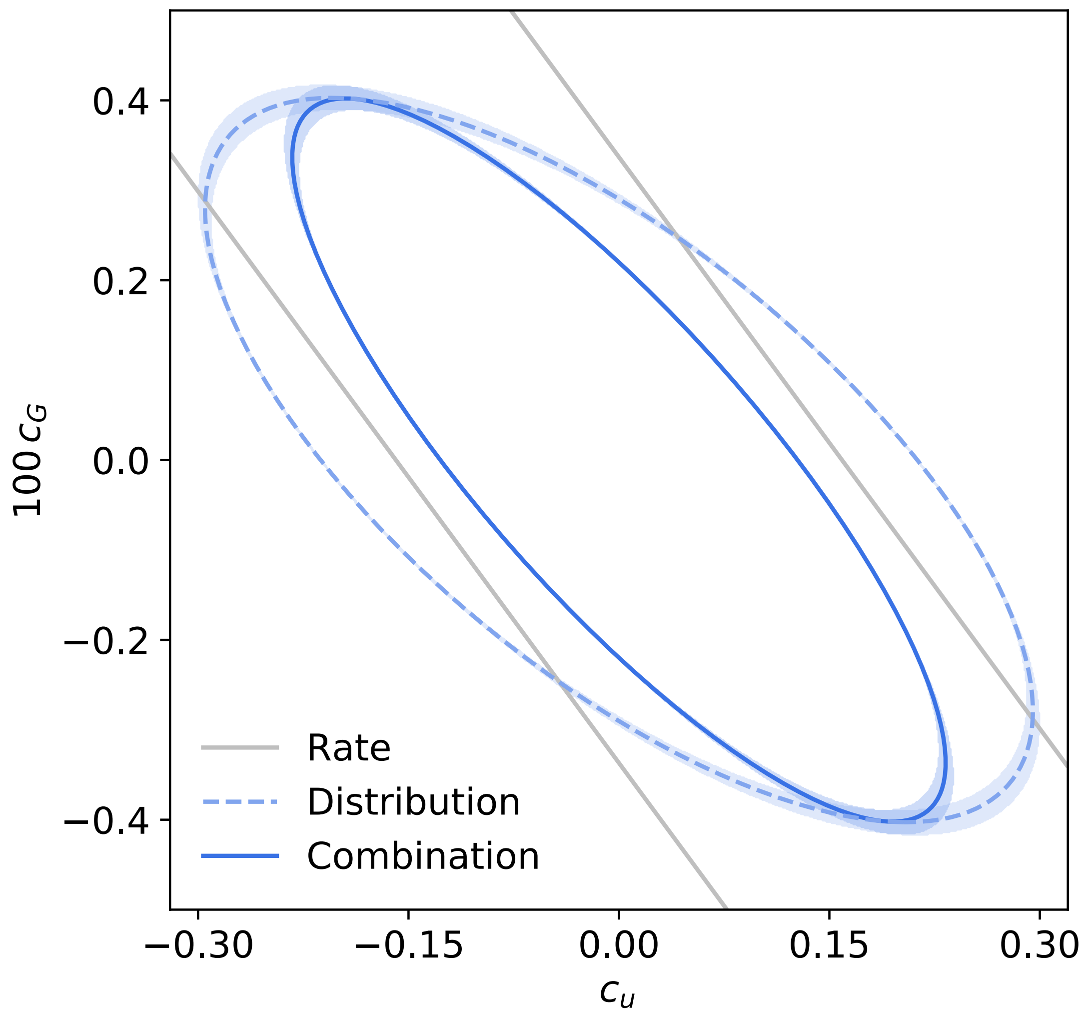
$$\mathcal{O}_u = -\frac{1}{v^2}\left(H^\dagger H\right)\left(H^\dagger \bar{Q}_L\right)u_R$$
- rescales top Yukawa: $y_t \to y_t\,(1+3/2\ c_u)$

$$\mathcal{O}_G = \frac{g_s^2}{m_W^2}\left(H^\dagger H\right)G^a_{\mu\nu}G_a^{\mu\nu}$$
- Higgs-gluon coupling: $g_{ggh} \to g_{ggh}(1 + 192\pi^2/g^2\ c_G)$

$$\mathcal{O}_{uG} = -\frac{4g_s}{m_W^2}y_u\left(H^\dagger \bar{Q}_L\right)\gamma^{\mu\nu}T_a u_R G^a_{\mu\nu}$$
- chromo-dipole moment: gtt, ggtt, gtth, ggtth

- background: ttγγ continuum
- simulation: MadGraph5 + Pythia8 + Delphes3
  * PDF4LHC: scale + PDF uncertainties
- 48 observable:
  * x={px, py, pz, E, pT, η, ΔΦ, MET, m_{ij}, …}
- Fisher Information:

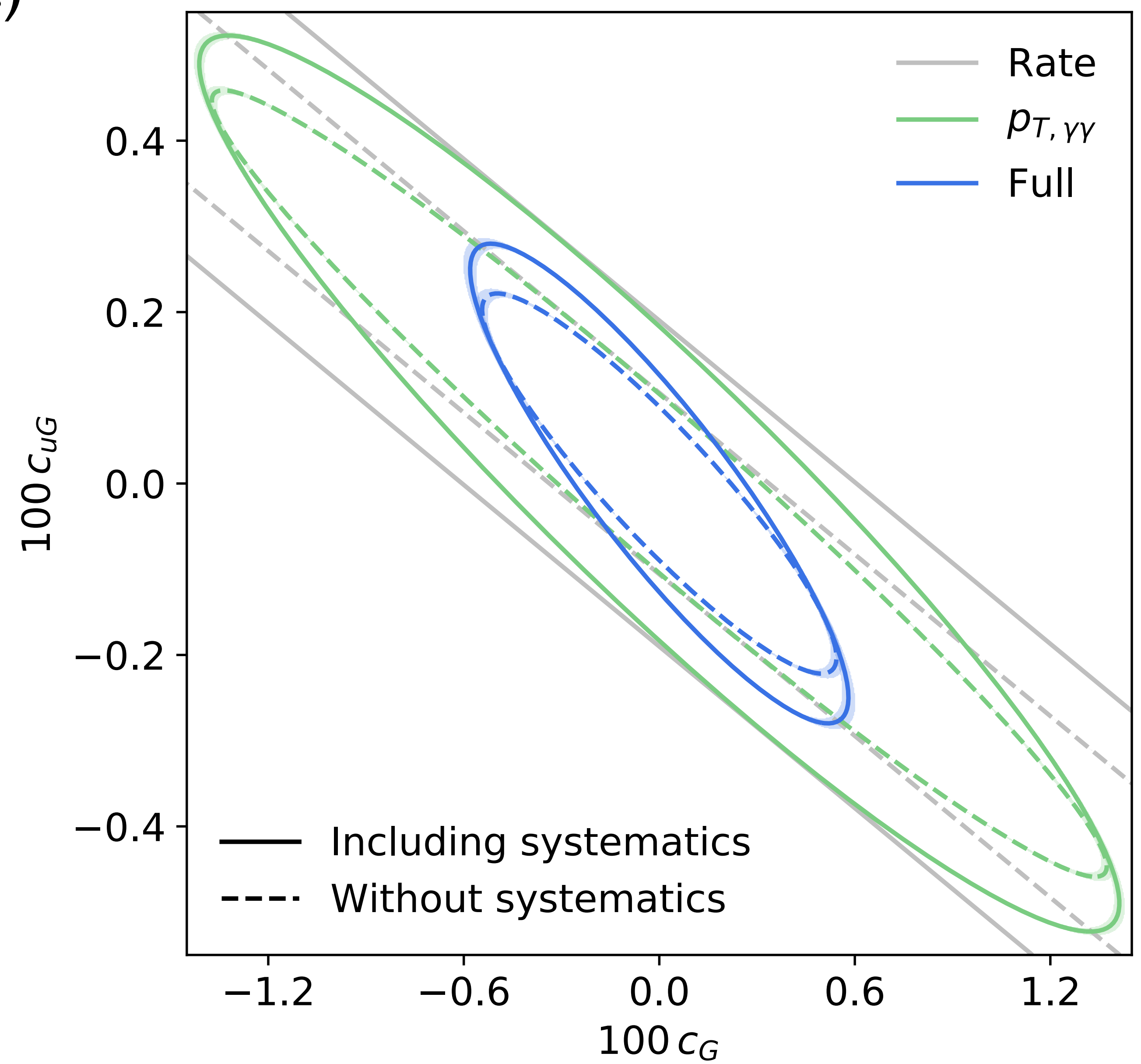
$$I_{ij} = \begin{array}{cc} & \begin{array}{ccc} c_u & 100c_G & 100c_{uG} \end{array} \\ & \left(\begin{array}{ccc} 140.5 & 68.1 & 170.6 \\ 68.1 & 47.1 & 105.7 \\ 170.6 & 105.7 & 283.3 \end{array}\right) \begin{array}{c} c_u \\ 100c_G \\ 100c_{uG} \end{array} \end{array}$$



How about Systematics?

- strategy
  * introduce nuisance parameter ν for scale (2) and PDF (30) uncertainties
  * replace r(x|θ)→r(x|θ,ν)
  * learn score, obtain Fisher Info (3+32 dim.)
  * profile over ν

- results:
  * mainly scale uncertainties
  * systematic reduce reach in rate-sensitive direction
  * multivariate analysis less affected



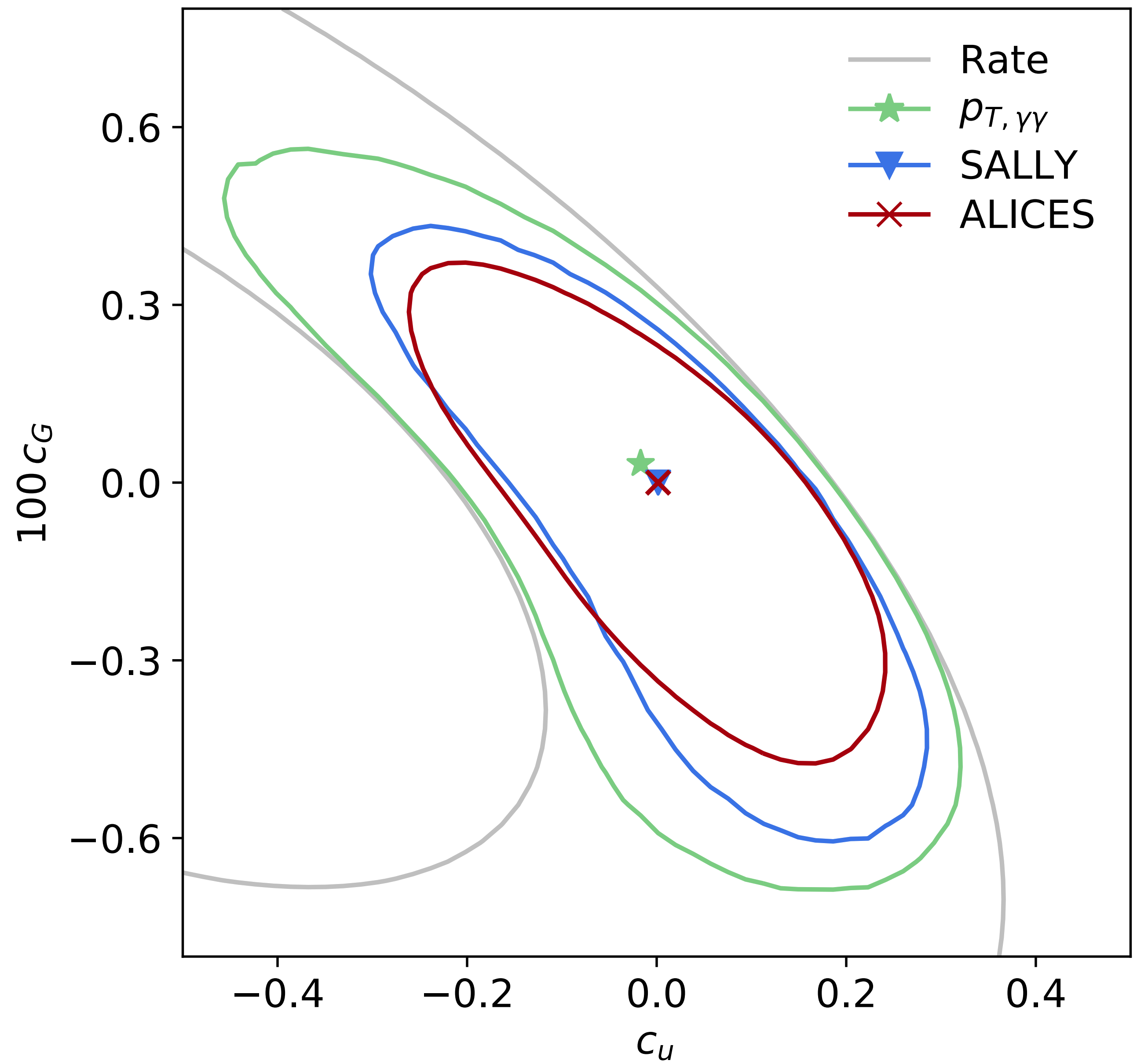
Full Results

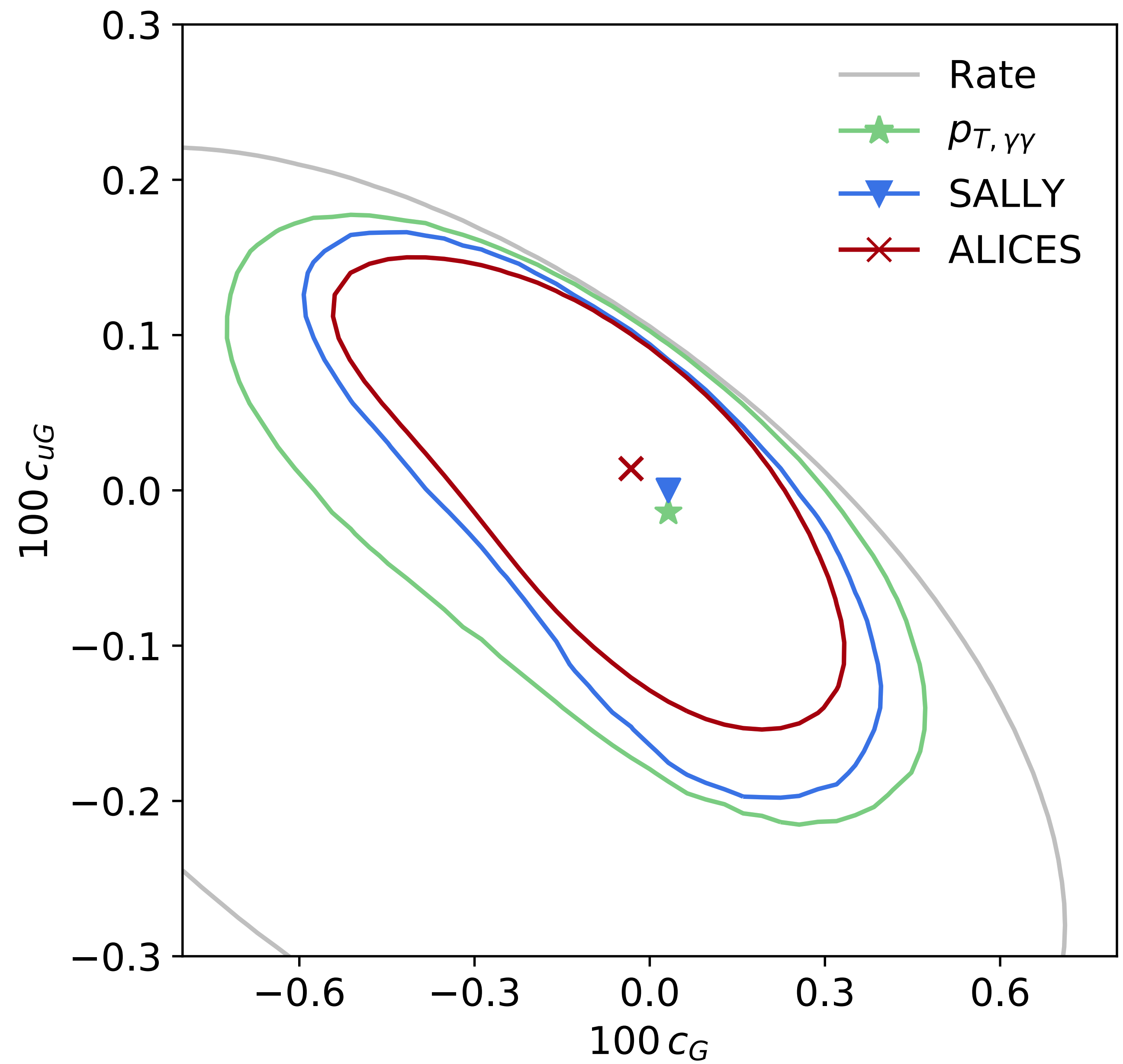
- inference methods

  total rate — SALLY: score $t(x)|_{SM}$ as locally optimal observables

  pTH histogram — ALICES: use full $r(x|\theta)$

- multivariate methods significantly improve reach





Example 2: Probing SMEFT in WH

[Brehmer, Dawson, Homiller, FK, Plehn 1908.06980]

WH production in SMEFT

- process WH>lvbb
- 3 operators contribute

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

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

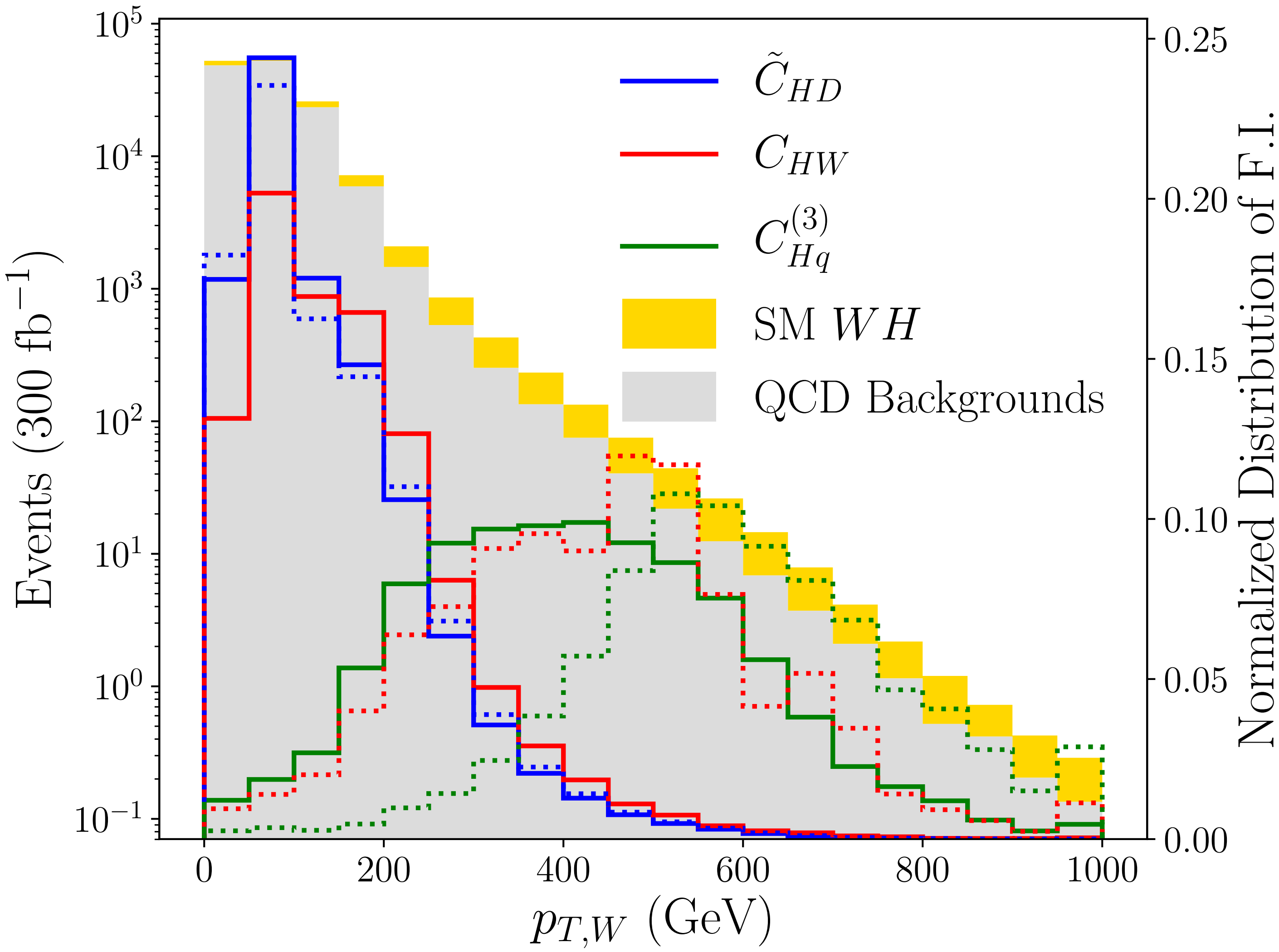
$$\mathcal{O}^{(3)}_{Hq} = \left(\phi^\dagger\, \overleftrightarrow{D}^a{}_\mu\, \phi\right)\left(\bar{Q}_L\sigma^a\gamma^\mu Q_L\right)$$

- 2 main observables: ptH and mT,tot
- where is the information
  * identify sensitive phase space regions

Distribution of Information:

— interference term

- - - - quadratic term

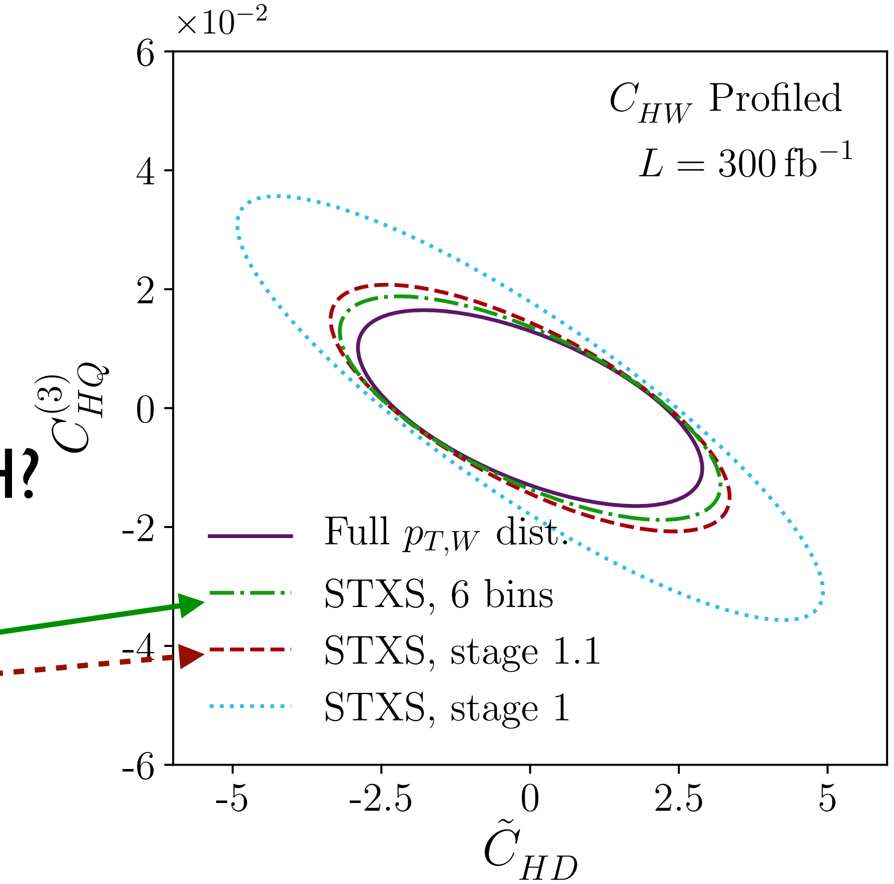


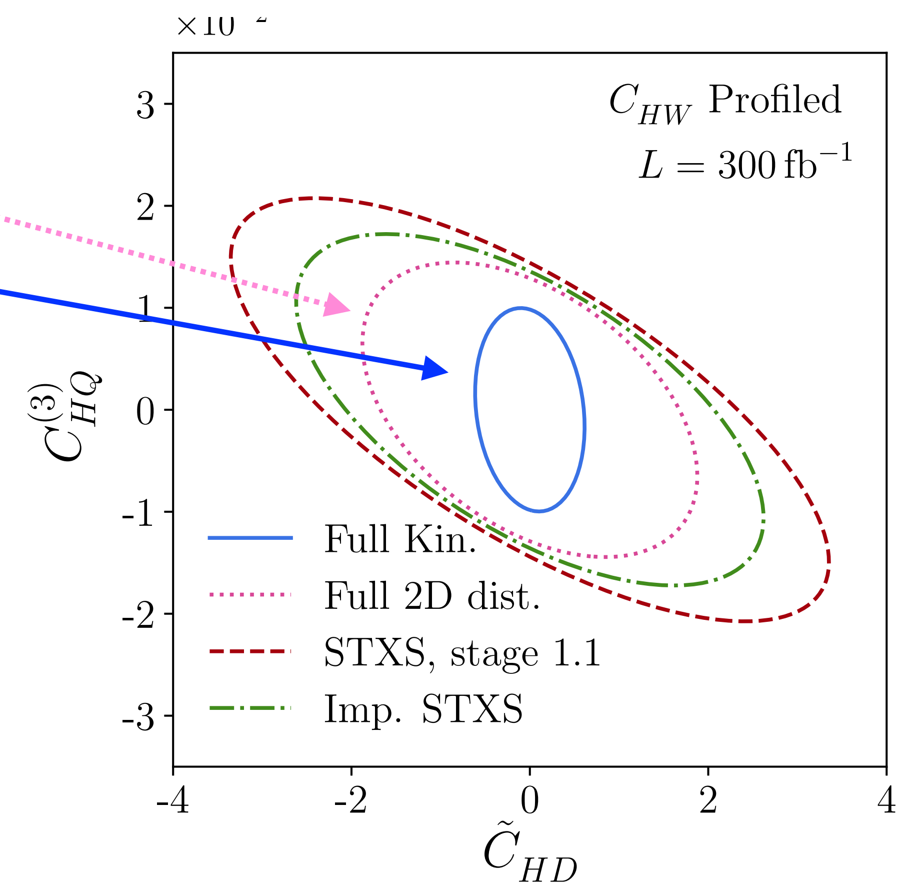
Example 2: Probing SMEFT in WH

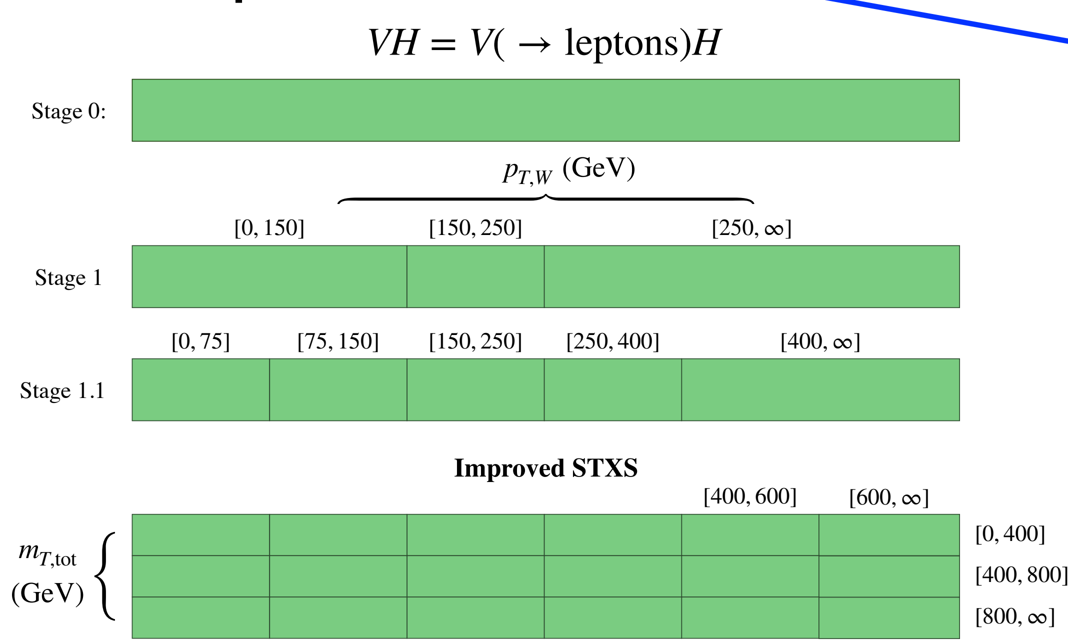
Another example: WH production in SMEFT

- process WH>lvbb
- 3 operators contribute
- 2 main observables: ptH and mT,tot

- how good can simplified template cross sections probe WH?
  * quantify performance using information geometry
  * compare to full information
  * additional high pT bin essential
  * include 2nd observable
  * multivariate analysis potentially much more powerful







Outline

Some Statistic Concepts
Likelihoods, Optimal Observables and Fisher Information

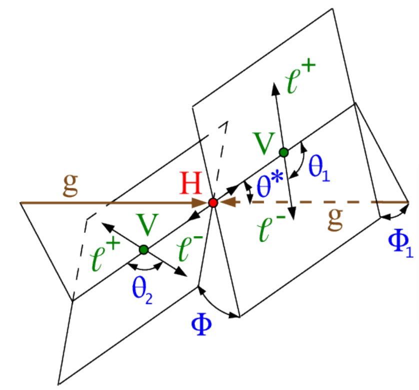
Information Geometry in Theory
Parton Level and Matrix Elements

Information Geometry in Reality
Detector Effects and Machine Learning

Summary and Conclusion

Summary and Conclusion

Theory in an Era of Data
- lots of data, kinematic features
- goal: constrain high-dimension theory space

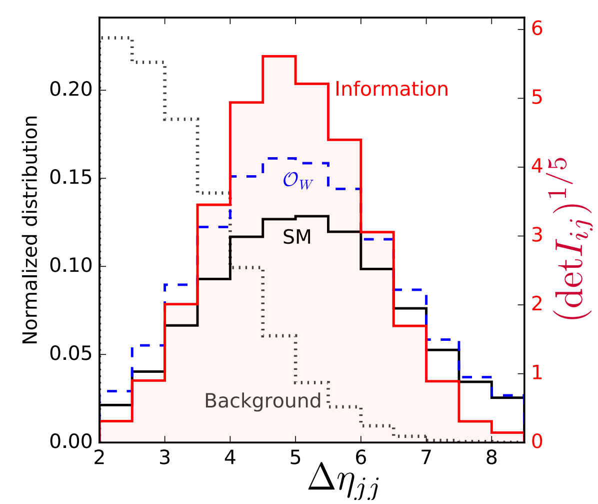


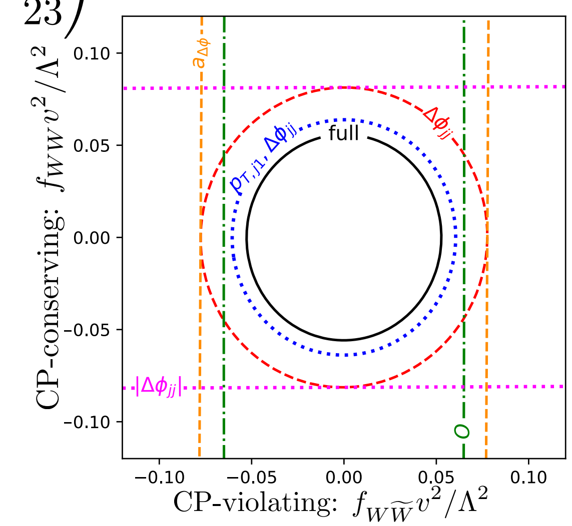
$$\begin{aligned}\Delta\mathcal{L}_{SILH} =& \frac{\bar{c}_H}{2v^2}\partial^\mu\left(H^\dagger H\right)\partial_\mu\left(H^\dagger H\right) + \frac{\bar{c}_T}{2v^2}\left(H^\dagger \overleftarrow{D'} H\right)\left(H^\dagger \overleftrightarrow{D}_\mu H\right) - \frac{\bar{c}_6\lambda}{v^2}\left(H^\dagger H\right)^3 \\ &+ \left(\left(\frac{\bar{c}_u}{v^2} y_u H^\dagger H \bar{q}_L H^c u_R + \frac{\bar{c}_d}{v^2} y_d H^\dagger H \bar{q}_L H d_R + \frac{\bar{c}_l}{v^2} y_l H^\dagger H \bar{L}_L H l_R\right) + \text{h.c.}\right) \\ &+ \frac{i\bar{c}_W g}{2m_W^2}\left(H^\dagger \sigma^i \overleftarrow{D^\mu} H\right)\left(D^\nu W_{\mu\nu}\right)^i + \frac{i\bar{c}_B g'}{2m_W^2}\left(H^\dagger \overleftarrow{D^\mu} H\right)\left(\partial^\nu B_{\mu\nu}\right) \\ &+ \frac{i\bar{c}_{HW} g}{m_W^2}\left(D^\mu H\right)^\dagger \sigma^i \left(D^\nu H\right) W^i_{\mu\nu} + \frac{i\bar{c}_{HB} g'}{m_W^2}\left(D^\mu H\right)^\dagger \left(D^\nu H\right) B_{\mu\nu} \\ &+ \frac{\bar{c}_\gamma g'^2}{m_W^2} H^\dagger H B_{\mu\nu} B^{\mu\nu} + \frac{\bar{c}_g g_S^2}{m_W^2} H^\dagger H G^a_{\mu\nu} G^{a\mu\nu}\end{aligned}$$

Information Geometry
- fisher information encodes the maximum sensitivity of observables to model parameters

$$I_{ij}(\mathbf{0}) = \begin{pmatrix} 715 & -191 & 1 & 0 \\ -191 & 321 & -1 & 0 \\ 1 & -1 & 359 & -81 \\ 0 & 1 & -81 & 23 \end{pmatrix}$$

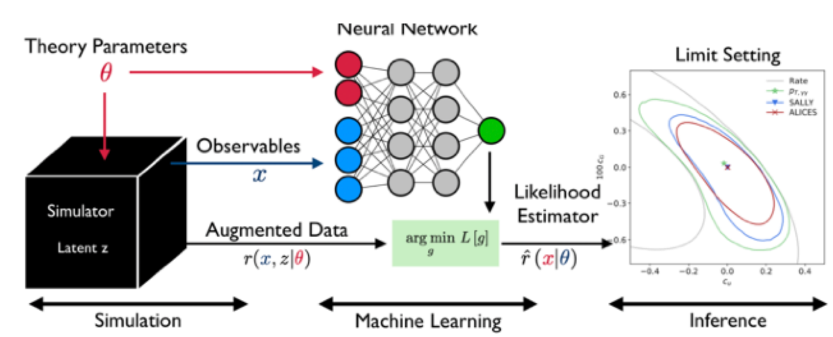
- calculate maximum sensitivity
- identify important phase space regions
- identify most powerful observables
- quantitatively compare analyses
- include systematics





- powerful and transparent analysis tool
- particularly easy to apply to EFT

MadMiner
- new inference techniques using ME and ML
- allows to apply information geometry at detector level
- MadGraph add-on / python package
- great for Pheno studies

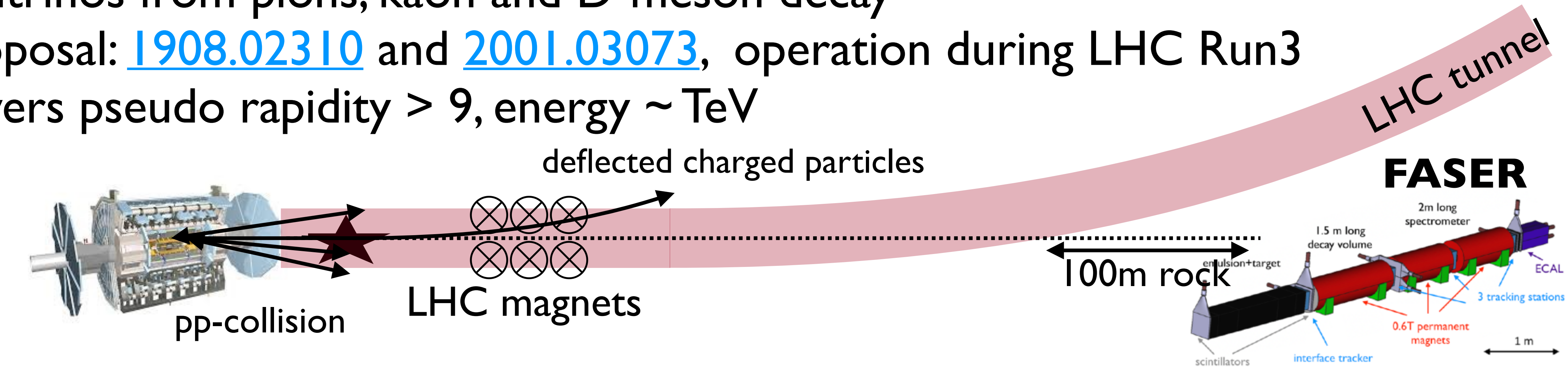


Backup

Cool stuff that didn't make it into the main part

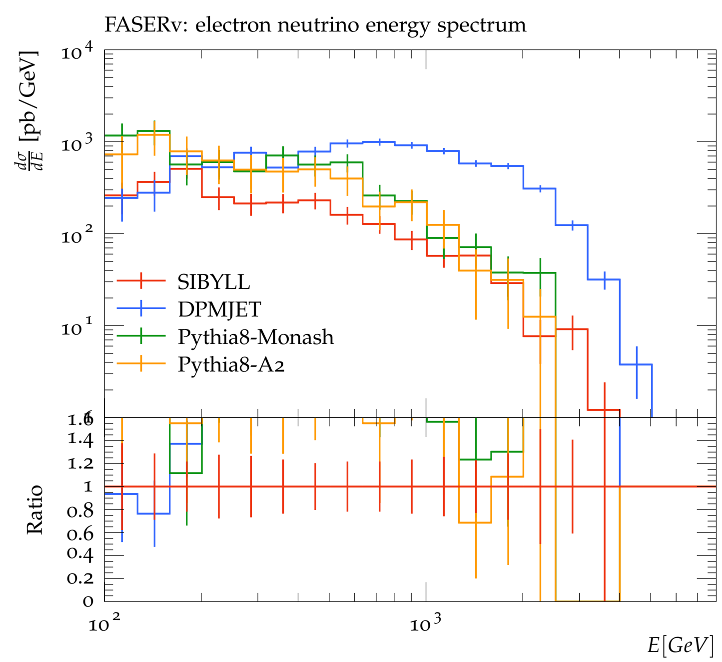
Completely Independent: FASERν

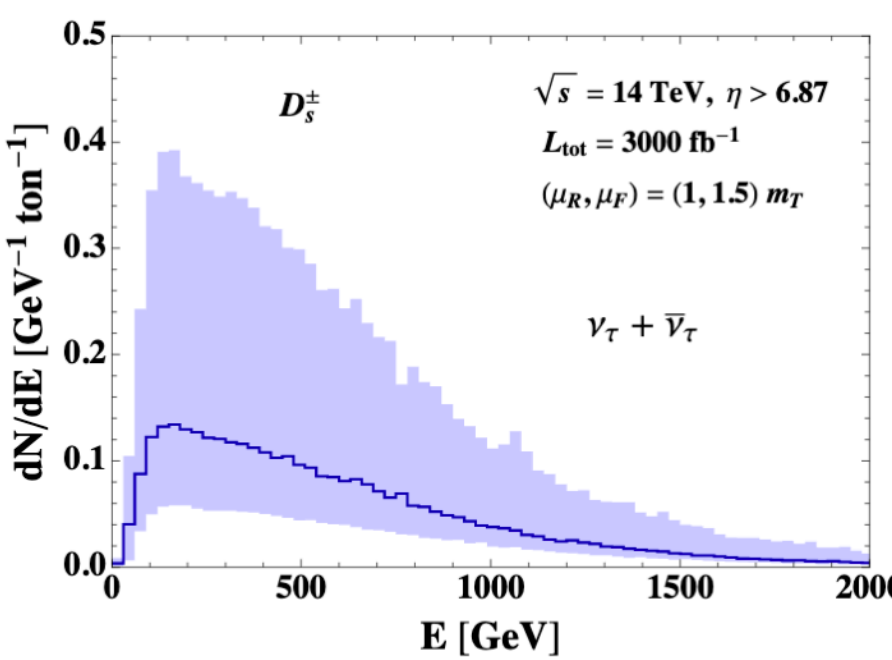
- the recently approved FASERν detector will detect for forward LHC neutrinos
  * neutrinos from pions, kaon and D-meson decay
  * proposal: 1908.02310 and 2001.03073, operation during LHC Run3
  * covers pseudo rapidity > 9, energy ~ TeV



- main physics goal: neutrino cross section at TeV energies
- currently large uncertainties on forward charm production
  * generators not yet tuned, differ by up to factor 10 → forward physics tune in progress
  * NLO QCD calculation has O(1) uncertainties as well: 2002.03012
- physics sensitivity ?
  * PDFs at low x,Q2 ?
  * intrinsic charm ?

Suggestions, feedback, help and ideas are welcome!

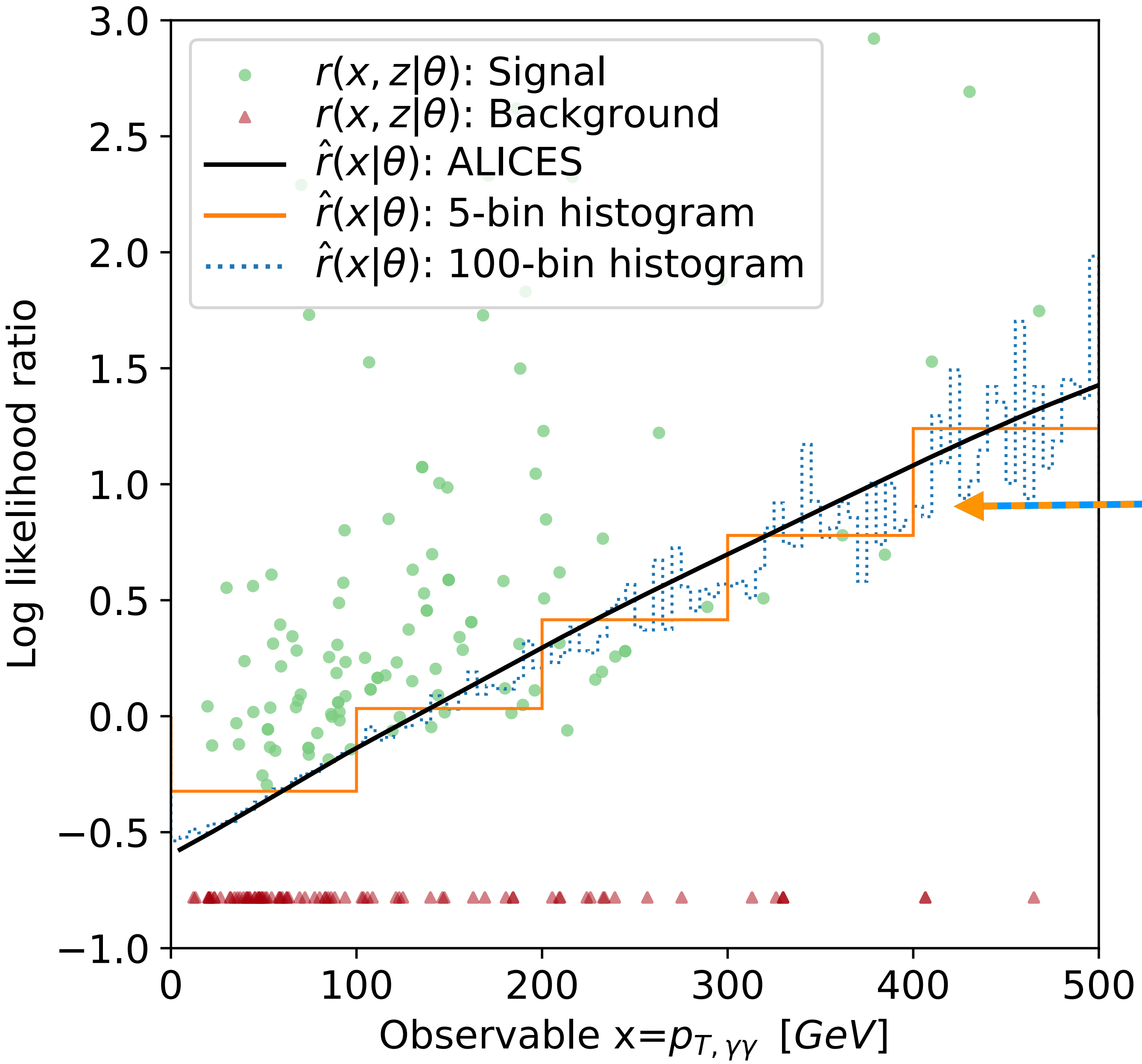




MadMiner with Backgrounds

How about Backgrounds?

- consider two models: BSM ($\theta=1$) vs SM ($\theta_{ref}=0$)
- include ttH signal and $tt\gamma\gamma$ background



- additional latent variable z: process label
- points: joint likelihood ratio

  estimate r(x, θ): "best fit"

Comparison with Histogram

- LLR obtained using histogram
- agrees well :)
- ML "continuum limit of for large # of bins"

- realistic problem: x and θ high dimensional

Probing SMEFT in tth

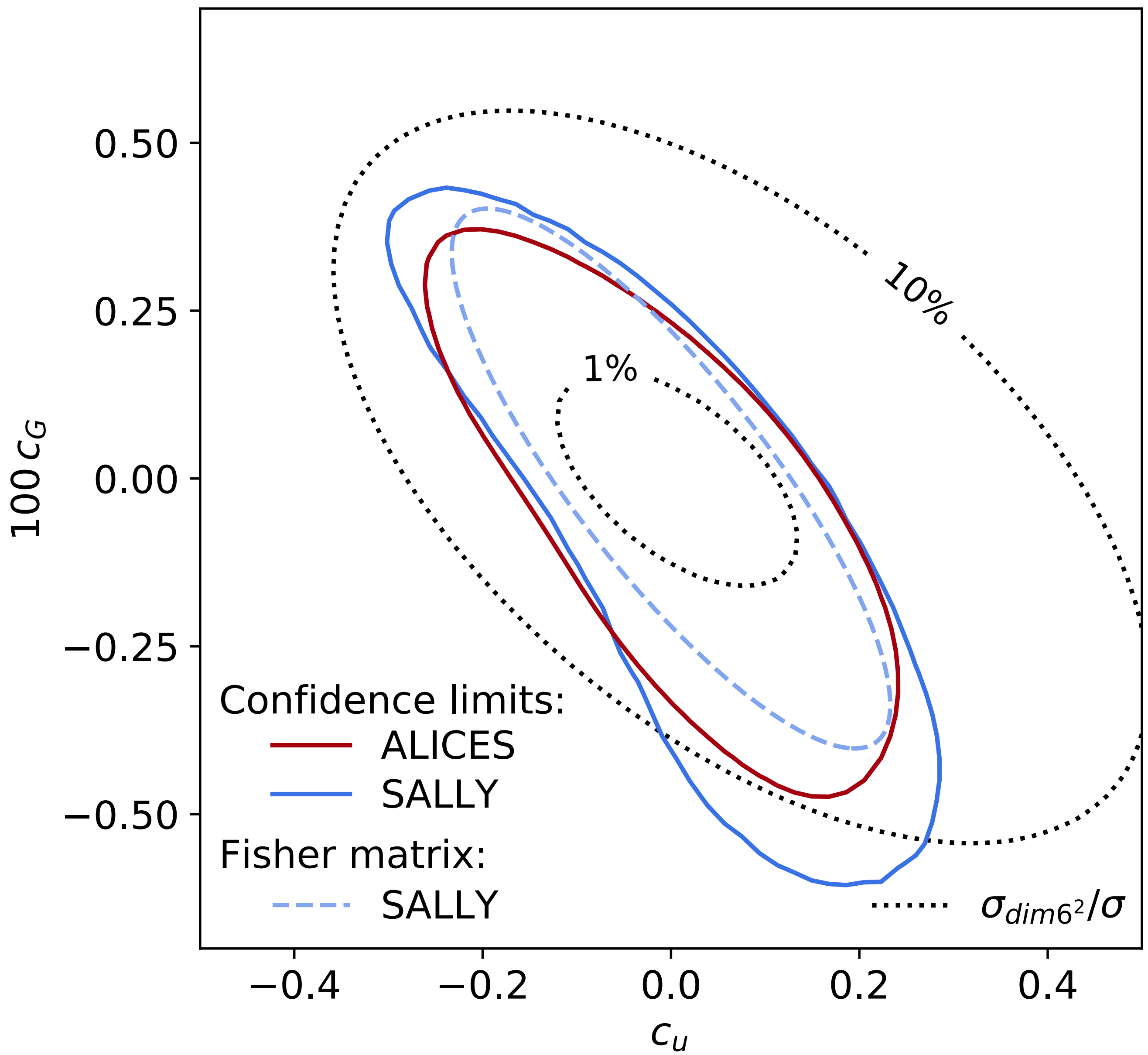
How good is local approximation?

- Fisher Information matrix
  * estimate
  * only sensitive to interference effects by construction
  * symmetric limits (ellipses)

- SALLY
  * estimates score t(x)|SM
  * use score as optimal observable (filled in histograms)
  * optimal only close to SM

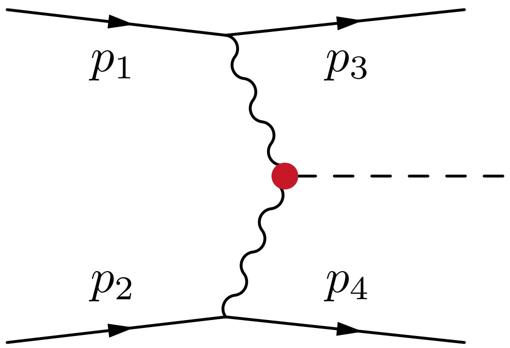
- ALICES:
  * estimate r(x|θ)
  * optimal limits in whole parameter space

- this example analysis:
  * few data, weak constraints
  * dim6 squared terms important in probed parameter space
  * local / full limits differ



Why is WBF Higgs production sensitive to CP?

- naive time reversal $\hat{T}: \; |\vec{p},\vec{s}\rangle \to |-\vec{p},-\vec{s}\rangle$
- $\hat{T}$-symmetric initial state at pp-collider
- $\hat{T}$-invariant squared matrix element in absence of CP-violation and re-scattering

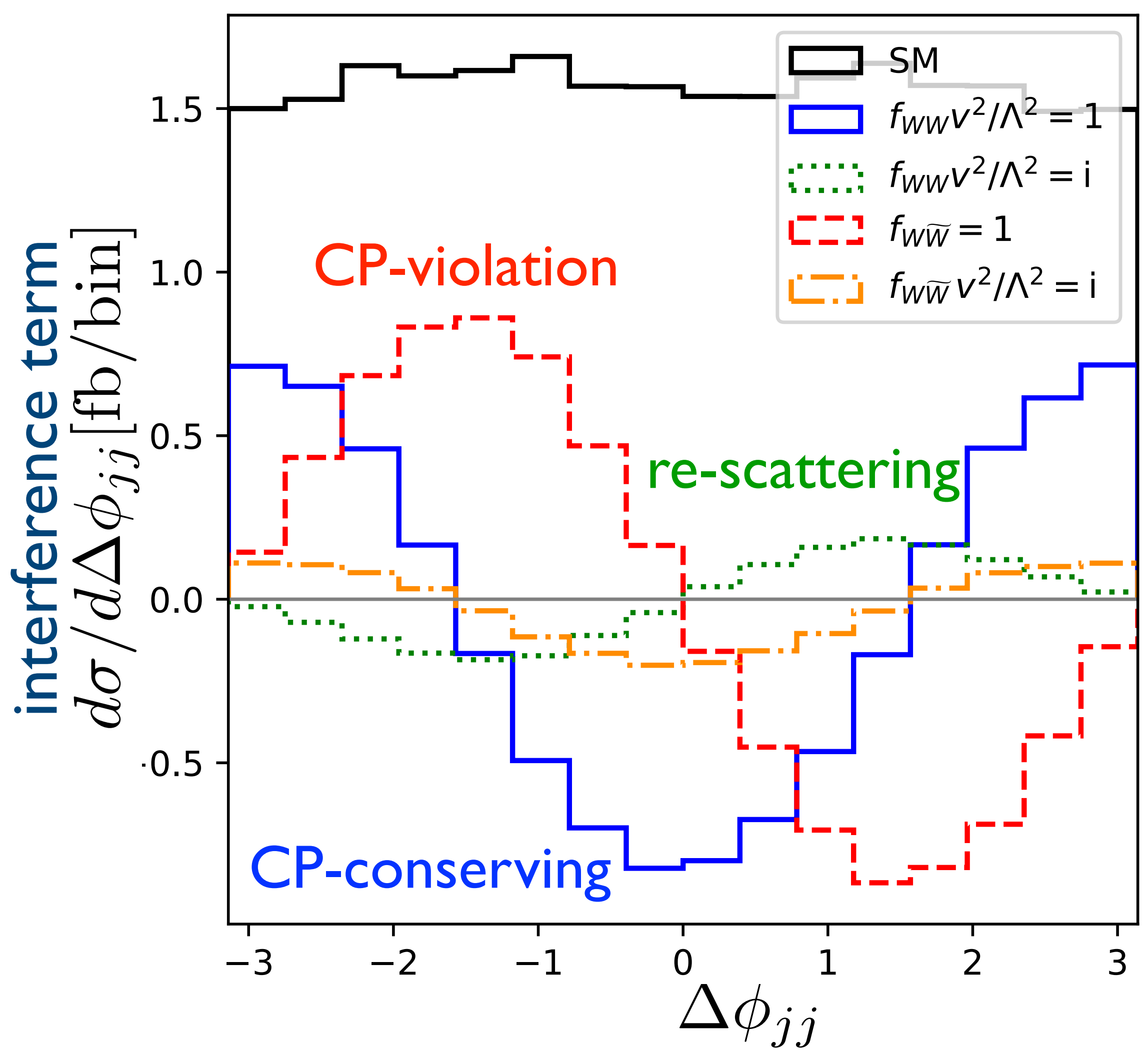


$$\langle f|\mathcal{T}|i\rangle \underset{\text{CPT-theorem}}{\overset{CP\text{-invariant}}{=}} \langle i_T|\mathcal{T}|f_T\rangle \underset{\text{optical theorem}}{\overset{\text{no re-scattering}}{=}} \langle f_T|\mathcal{T}|i_T\rangle^* \quad \Rightarrow \quad |\langle f|\mathcal{T}|i\rangle|^2 = |\langle f_T|\mathcal{T}|i_T\rangle|^2$$

- genuine $\hat{T}$-odd observable $\epsilon_{\alpha\beta\gamma\delta}\, p_1^\alpha p_2^\beta p_3^\gamma p_4^\delta$

→ signed angle $\Delta\phi_{jj}^s$

$\Delta\phi_{jj}^s$ is sensitive to CP-violation
if re-scattering effects are known to be small



Summary: Inference Techniques

Traditional

- use summary statistics
  * hand-picked observables x'
  * estimate $p(x'|\theta)$
- information loss
- problem dependent

Examples:

- rate only (cut and count)
- histograms
- Approximate Bayesian Computation
- STXS

Machine Learning

- multivariate analysis
- works great for S vs BG
- struggles with S' vs S
  * large number of S'
  * very similar S', S

Examples:

- Neural Density Estimator
- ML Classifier

Matrix Element Based

- multivariate analysis
- uses $p(x|\theta) \sim |M(x|\theta)|^2$
- works great at parton level
  * S' vs S is easy
- requires approximations in reality
  * S vs BG can be hard

Examples:

- Matrix Element Method
- Optimal Observables

power of machine learning → **MadMiner** ← physics insight of matrix element information

[J. Brehmer, K. Cranmer, G. Louppe, J. Pavez 1805.00013, 1805.00020, 1805.12244]

[J. Brehmer, FK, I. Espejo, K. Cranmer 1907.10621]