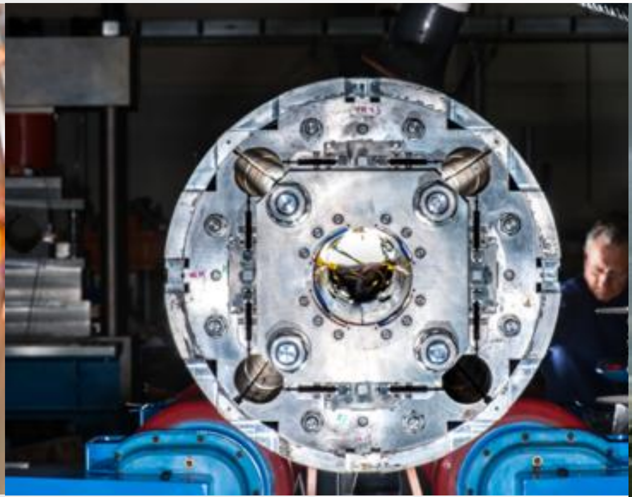
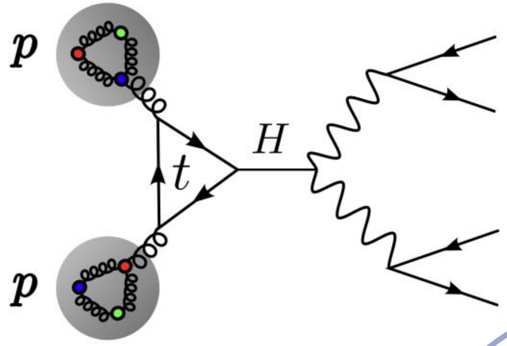
EXPLORING EFT WITH ML AT THE LHC

R. Schöfbeck (HEPHY Vienna), Dec. 2nd, 2022

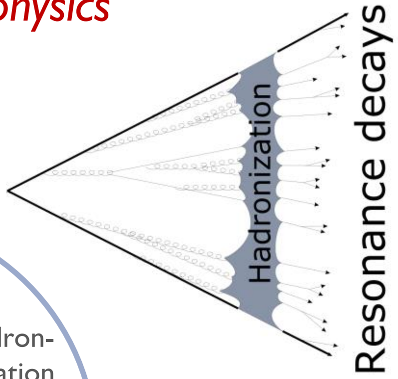
HEP MODELLING FROM *ML* POINT OF VIEW

Short-distance physics

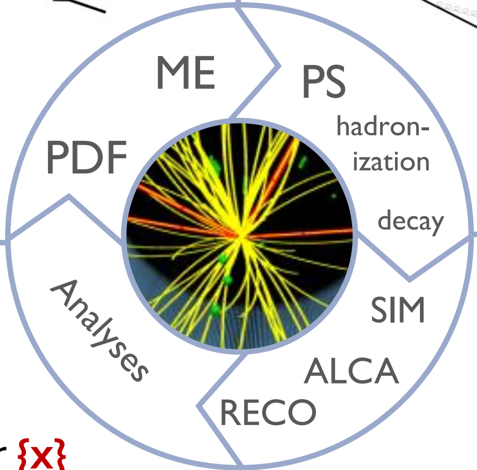


- *generative models* for MEs from perturbative QFT, sample parton level, and provide likelihood ratios via $\frac{|\mathcal{M}(z_p|\boldsymbol{\theta}\ \ )|^2}{|\mathcal{M}(z_p|\mathrm{SM})|^2}$

Long-distance physics



- parton shower
- factorization / hadronization models
- decay branchings, calculated & measured
- large *latent space*

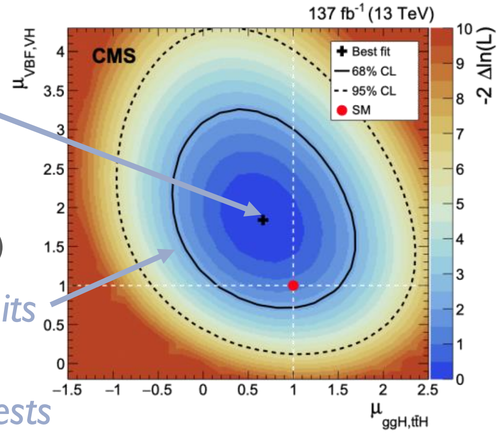


Event reconstruction & data analysis

Maximum-likelihood estimate

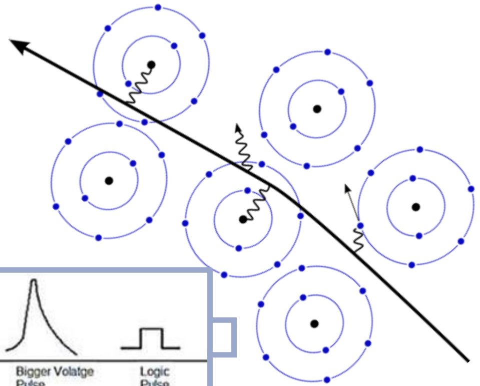
$\boldsymbol{\theta}_{\mathrm{MLE}} = \mathrm{argmax}_{\boldsymbol{\theta}}\ L(\mathcal{D},\boldsymbol{\theta})$

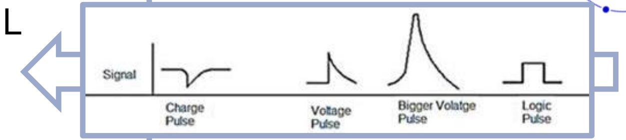
Confidence limits based on likelihood-ratio tests

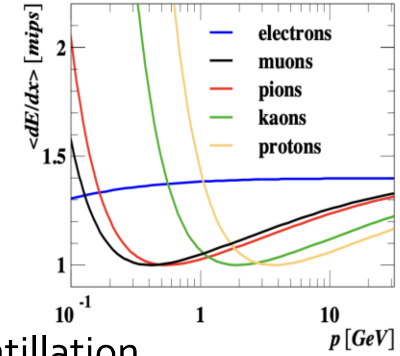


- observed feature vector **{x}**
  - used in MLE
  - hypo.-tests & CL
- Simulation: sampling of **{x,z}** for ML training & test design

Detector interactions



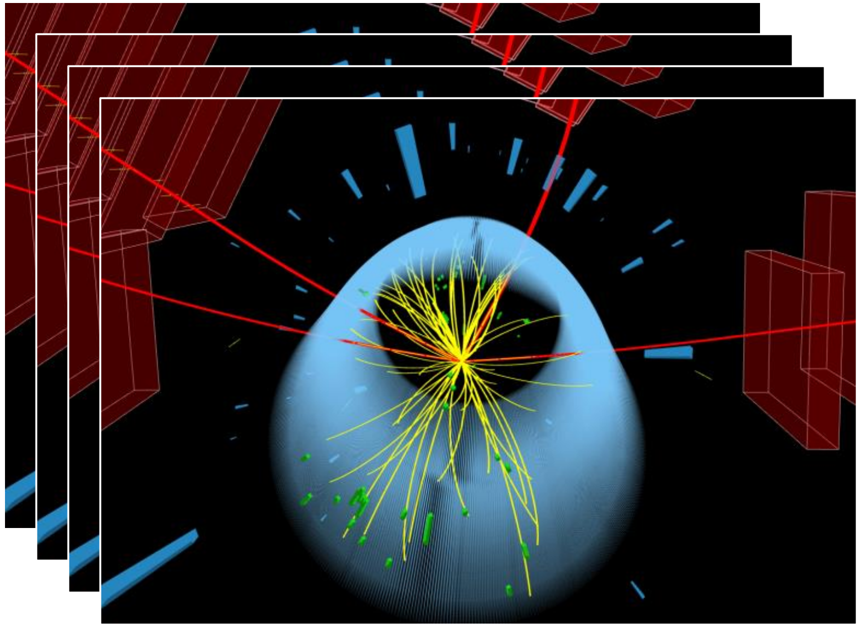




- ionization (Bethe-Bloch), szintillation, brem., transition-radiation … more latent variables

NEYMAN-PEARSON & LIKELHOOD RATIO "TRICK"

arxiv:1503.0x7622



Neyman-Pearson Lemma: The *likelihood* ***ratio*** *test statistic is optimal*

data-set with feature vectors **x**

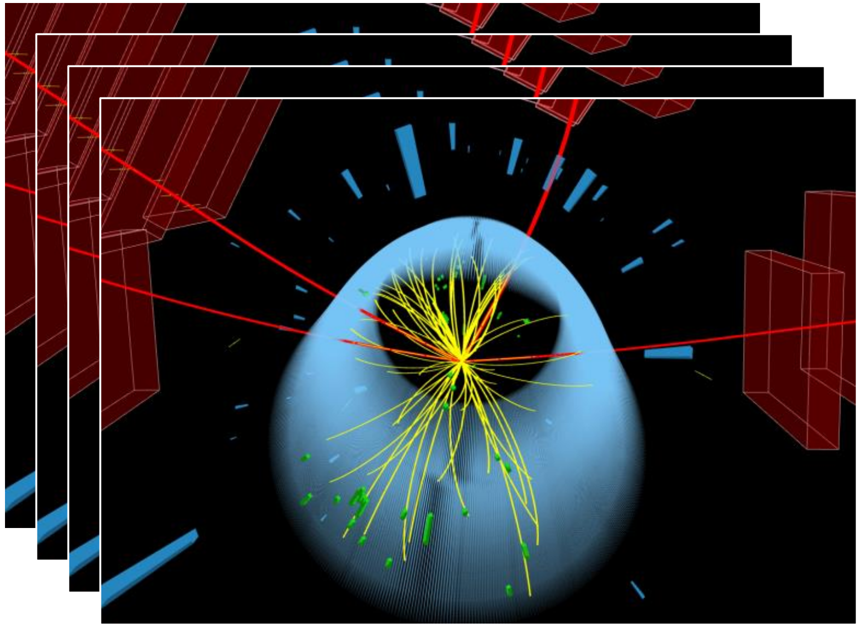
diff xsec ratio

$$q_{\boldsymbol{\theta}}(\mathcal{D}) \sim -\sum_{\boldsymbol{x}_i \in \mathcal{D}} \log \frac{\sigma(\boldsymbol{\theta})}{\sigma(\mathrm{SM})} \frac{p(\boldsymbol{x}_i|\boldsymbol{\theta}\quad)}{p(\boldsymbol{x}_i|\mathrm{SM})}$$

theory parameters

NEYMAN-PEARSON & LIKELHOOD RATIO "TRICK"

arxiv:1503.0x7622



Neyman-Pearson Lemma: The *likelihood* ***ratio*** *test statistic is optimal*

data-set with feature vectors **x**

diff xsec ratio

$$q_{\boldsymbol{\theta}}(\mathcal{D}) \sim -\sum_{\boldsymbol{x}_i \in \mathcal{D}} \log \frac{\sigma(\boldsymbol{\theta})}{\sigma(\mathrm{SM})} \frac{p(\boldsymbol{x}_i|\boldsymbol{\theta})}{p(\boldsymbol{x}_i|\mathrm{SM})}$$

theory parameters

Likelihood ratio "trick": label two values: **θ**, SM

$$L = \int \mathrm{d}\boldsymbol{x} \sum_{z \in \{0,1\}} p(\boldsymbol{x}, z) \left(z - \hat{f}(\boldsymbol{x})\right)^2$$

training samples

truth (supervised)

classifier

$$f^*(\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \mathrm{SM})}{p(\boldsymbol{x}, \mathrm{SM}) + p(\boldsymbol{x}, \boldsymbol{\theta})} = \frac{1}{1 + \frac{\sigma(\boldsymbol{\theta})}{\sigma(\mathrm{SM})} r(\boldsymbol{x})}$$

supervised learning provides (close-to) optimal test statisticsc

Can we avoid retrain for each hypothesis separately?

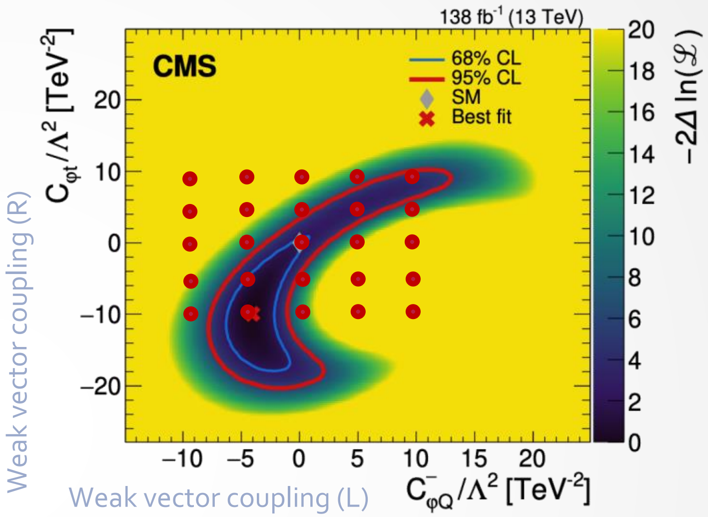
SENDING MIXED SIGNALS TO THE LOSS FUNCTION

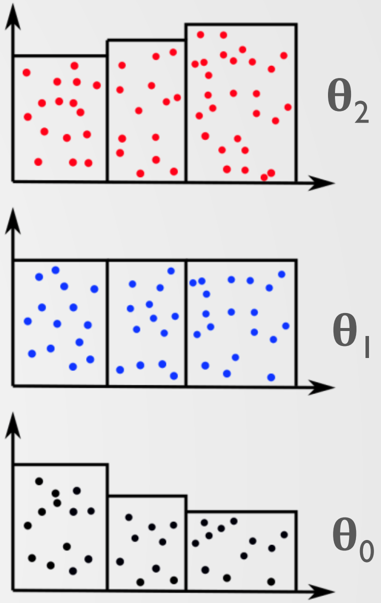
$$L = \sum_{\boldsymbol{\theta}\in\mathcal{B}} \int \mathrm{d}\boldsymbol{x} \left(p(\boldsymbol{x}|\boldsymbol{\theta})\hat{f}(\boldsymbol{x})^2 + p(\boldsymbol{x}|\mathrm{SM})(1-\hat{f}(\boldsymbol{x}))^2 \right)$$

$\boldsymbol{\theta}$ - ignorant

mixing signals &
case dependent mixes

$$f^*(\boldsymbol{x}) = \frac{1}{1+r_{\mathcal{B}}(\boldsymbol{x})} \qquad r_{\mathcal{B}}(\boldsymbol{x}) = \frac{\frac{1}{|\mathcal{B}|}\sum_{\boldsymbol{\theta}\in\mathcal{B}} p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x}|\mathrm{SM})}$$





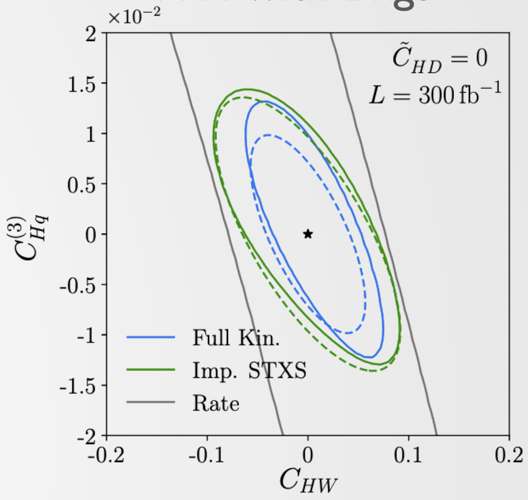
simulating different samples at each parameter point "SUSY style"

- Sending 'mixed signals' to the loss function
  - Averages the training data set - suboptimal when linear effects dominate
  - Classifier does not reflect knowledge on the $\boldsymbol{\theta}$-dependence
  - Separate trainings per $\boldsymbol{\theta}$ not feasible for high parameter dimensions → parametrize classifiers
  - For simulation: We can do better in SMEFT than the "simplified model" - style simulation with 1 sample per $\boldsymbol{\theta}$
- Definition: *SMEFT-specific ML exploits the analytic structure of the SMEFT predictions*
  - The challenge of global SM-EFT searches will require a high degree of "semi" automatization

REFERENCES (SELECTION)

- Madminer: Neural networks based likelihood-free inference & related techniques
  - K. Cranmer , J. Pavez , and G. Louppe [1506.02169]
    J. Brehmer, K. Cranmer, G. Louppe, J. Pavez [1805.00013] [1805.00020] [1805.12244]
    J. Brehmer, F. Kling, I. Espejo, K. Cranmer [1907.10621]
  - J. Brehmer, S. Dawson, S. Homiller, F. Kling, T. Plehn [1908.06980]
  - A. Butter, T. Plehn, N. Soybelman, J. Brehmer [2109.10414]
  - established many of the *main ideas* & *statistical interpretation* in various *NN applications*
- Weight derivative regression (A.Valassi) [2003.12853]
- Parametrized classifiers for SM-EFT: NN with quadratic structure
  - S. Chen, A. Glioti, G. Panico, A. Wulzer [*JHEP 05 (2021) 247*]
- Boosted Information Trees: Tree algorithms & boosting
  - S. Chatterjee, S. Rohshap, N. Frohner, R.S., D. Schwarz [2107.10859], [2205.12976]
- ML4EFT R. Ambrosio, J. Hoeve, M. Madigan, J. Rojo, V. Sanz [2211.02058]
- All approaches are “SMEFT-specific ML” with differences mostly on the practical side

WH with Bkgs



my practical experience

→ talk later today

LIKELIHOOD-FREE INFERENCE FOR SM-EFT

[Madminer 1805.00020]

1. Sampling of $p(\boldsymbol{x}, \boldsymbol{z}_\mathrm{d}, \boldsymbol{z}_\mathrm{s}, \boldsymbol{z}_\mathrm{p} | \boldsymbol{\theta})$, true L. intractable: $p(\boldsymbol{x}|\boldsymbol{\theta}) = \int \mathrm{d}\boldsymbol{z}_\mathrm{d}\, \mathrm{d}\boldsymbol{z}_\mathrm{s}\, \mathrm{d}\boldsymbol{z}_\mathrm{p}\; p(\boldsymbol{x}|\boldsymbol{z}_\mathrm{d})\; p(\boldsymbol{z}_\mathrm{d}|\boldsymbol{z}_\mathrm{s})\; p(\boldsymbol{z}_\mathrm{s}|\boldsymbol{z}_\mathrm{p})\; p(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta})$

LIKELIHOOD-FREE INFERENCE FOR SM-EFT

[Madminer 1805.00020]

1. Sampling of $p(\boldsymbol{x}, \boldsymbol{z}_\mathrm{d}, \boldsymbol{z}_\mathrm{s}, \boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta})$, true L. intractable: $p(\boldsymbol{x}|\boldsymbol{\theta}) = \int \mathrm{d}\boldsymbol{z}_\mathrm{d}\, \mathrm{d}\boldsymbol{z}_\mathrm{s}\, \mathrm{d}\boldsymbol{z}_\mathrm{p}\; p(\boldsymbol{x}|\boldsymbol{z}_\mathrm{d})\; p(\boldsymbol{z}_\mathrm{d}|\boldsymbol{z}_\mathrm{s})\; p(\boldsymbol{z}_\mathrm{s}|\boldsymbol{z}_\mathrm{p})\; p(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta})$

2. Exploit simplicity of the joint space: Intractable factors cancel in the joint likelihood ratio

$$r(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}, \mathrm{SM}) \equiv \frac{p(\boldsymbol{x}, \boldsymbol{z}_\mathrm{d}, \boldsymbol{z}_\mathrm{s}, \boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta})}{p(\boldsymbol{x}, \boldsymbol{z}_\mathrm{d}, \boldsymbol{z}_\mathrm{s}, \boldsymbol{z}_\mathrm{p}|\mathrm{SM})} = \frac{p(\boldsymbol{x}|\boldsymbol{z}_\mathrm{d})}{p(\boldsymbol{x}|\boldsymbol{z}_\mathrm{d})} \frac{p(\boldsymbol{z}_\mathrm{d}|\boldsymbol{z}_\mathrm{s})}{p(\boldsymbol{z}_\mathrm{d}|\boldsymbol{z}_\mathrm{s})} \frac{p(\boldsymbol{z}_\mathrm{s}|\boldsymbol{z}_\mathrm{p})}{p(\boldsymbol{z}_\mathrm{s}|\boldsymbol{z}_\mathrm{p})} \frac{p(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta})}{p(\boldsymbol{z}_\mathrm{p}|\mathrm{SM})} \propto \frac{|\mathcal{M}(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta})|^2}{|\mathcal{M}(\boldsymbol{z}_\mathrm{p}|\mathrm{SM})|^2} = \frac{w(\boldsymbol{\theta})}{w(\mathrm{SM})}$$

Change in likelihood of simulated observation x with latent "history" z going from "SM" to θ

staged simulation in forward mode: Intractable factors cancel

re-calcuable theory prediction

weighted simulation

LIKELIHOOD-FREE INFERENCE FOR SM-EFT

[Madminer 1805.00020]

1. Sampling of $p(\boldsymbol{x}, \boldsymbol{z}_\mathrm{d}, \boldsymbol{z}_\mathrm{s}, \boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta})$, true L. intractable: $p(\boldsymbol{x}|\boldsymbol{\theta}) = \int \mathrm{d}\boldsymbol{z}_\mathrm{d}\, \mathrm{d}\boldsymbol{z}_\mathrm{s}\, \mathrm{d}\boldsymbol{z}_\mathrm{p}\, p(\boldsymbol{x}|\boldsymbol{z}_\mathrm{d})\, p(\boldsymbol{z}_\mathrm{d}|\boldsymbol{z}_\mathrm{s})\, p(\boldsymbol{z}_\mathrm{s}|\boldsymbol{z}_\mathrm{p})\, p(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta})$

2. Exploit simplicity of the joint space: Intractable factors cancel in the joint likelihood ratio

$$r(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}, \mathrm{SM}) \equiv \frac{p(\boldsymbol{x}, \boldsymbol{z}_\mathrm{d}, \boldsymbol{z}_\mathrm{s}, \boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta}\ \ )}{p(\boldsymbol{x}, \boldsymbol{z}_\mathrm{d}, \boldsymbol{z}_\mathrm{s}, \boldsymbol{z}_\mathrm{p}|\mathrm{SM})} = \frac{p(\boldsymbol{x}|\boldsymbol{z}_\mathrm{d})}{p(\boldsymbol{x}|\boldsymbol{z}_\mathrm{d})}\frac{p(\boldsymbol{z}_\mathrm{d}|\boldsymbol{z}_\mathrm{s})}{p(\boldsymbol{z}_\mathrm{d}|\boldsymbol{z}_\mathrm{s})}\frac{p(\boldsymbol{z}_\mathrm{s}|\boldsymbol{z}_\mathrm{p})}{p(\boldsymbol{z}_\mathrm{s}|\boldsymbol{z}_\mathrm{p})}\frac{p(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta}\ \ )}{p(\boldsymbol{z}_\mathrm{p}|\mathrm{SM})} \propto \frac{|\mathcal{M}(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta}\ \ )|^2}{|\mathcal{M}(\boldsymbol{z}_\mathrm{p}|\mathrm{SM})|^2} = \frac{w(\boldsymbol{\theta})}{w(\mathrm{SM})}$$

Change in likelihood of simulated observation x with latent "history" z going from "SM" to θ

staged simulation in forward mode: Intractable factors cancel

re-calcuable theory prediction

weighted simulation

3. Regress (e.g.) in the joint likelihood ratio, ignoring the latent space. Available empirically.

$$L = \int \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{z}\, p(\boldsymbol{x}, \boldsymbol{z}|\mathrm{SM}) \left(r(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}, \mathrm{SM}) - \hat{f}_{\boldsymbol{\theta}}(\boldsymbol{x})\right)^2 \longrightarrow \min$$

4. Obtain change of likelihood for a specific observation, suitably integrating latent histories. NP optimal!

$$f^*_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\sigma(\theta)}{\sigma(\theta_0)} \frac{\int \mathrm{d}\boldsymbol{z}\, p(\boldsymbol{x}, \boldsymbol{z}) r(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\theta}_0)}{\int \mathrm{d}\boldsymbol{z}\, p(\boldsymbol{x}, \boldsymbol{z})} = \frac{\sigma(\theta)}{\sigma(\theta_0)} \frac{\int \mathrm{d}\boldsymbol{z}\, p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})}{\int \mathrm{d}\boldsymbol{z}\, p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}_0)} = \frac{\sigma(\theta)}{\sigma(\theta_0)} \frac{p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x}|\boldsymbol{\theta}_0)} = r(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\theta}_0)$$

Available from simulation

Latent space is integrated

what we actually want: change in likelihood of a specific observation

PARAMETRIZED CLASSIFIERS: NETS & TREES

Wulzer et.al. [JHEP 05 (2021) 247]

RS et. al., [2107.10859], [2205.12976]

$$L = \sum_{\boldsymbol{\theta}\in\mathcal{B}} \int \mathrm{d}\boldsymbol{x} \left(p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta}) \hat{f}(\boldsymbol{x};\boldsymbol{\theta})^2 + p(\boldsymbol{x},\boldsymbol{z}|\mathrm{SM})(1-\hat{f}(\boldsymbol{x};\boldsymbol{\theta}))^2 \right)$$

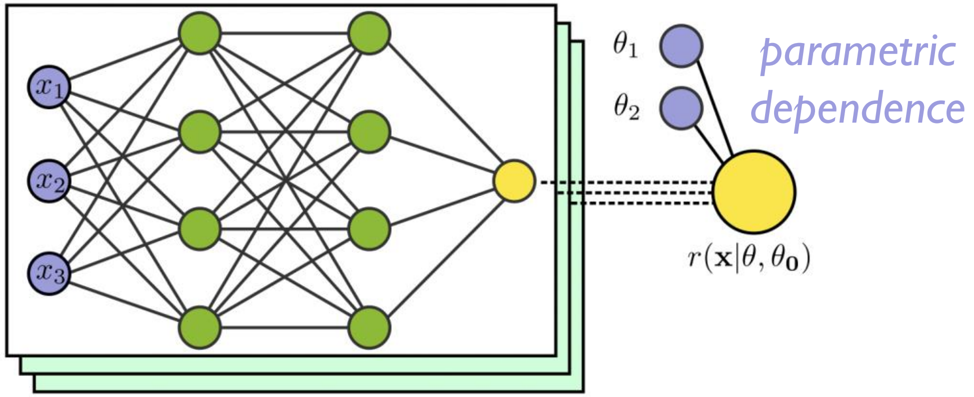
Make loss function aware of analytic SMEFT structure

Invert likelihood trick
with positive polynomial of NN -outputs

$$\hat{f}(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{1+\hat{r}(\boldsymbol{x};\boldsymbol{\theta})}$$

$$\hat{r}(\boldsymbol{x};\boldsymbol{\theta}) = \left(1 + \sum_a \theta_a \hat{n}_a(\boldsymbol{x})\right)^2 + \sum_a \left(\sum_{b\geq a} \theta_b \hat{n}_{ab}(\boldsymbol{x})\right)^2$$

Fit NNs simultaneously



$$L = \sum_{\boldsymbol{\theta}\in\mathcal{B}} \int \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{z}\, p(\boldsymbol{x},\boldsymbol{z}|\mathrm{SM}) \left(r(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta},\mathrm{SM}) - \hat{F}(\boldsymbol{x},\boldsymbol{\theta}) \right)^2$$

Tree ansatz with polynomial SMEFT dependence

$$\hat{F}(\boldsymbol{x},\boldsymbol{\theta}) = \sum_{j\in\mathcal{J}} \mathbb{1}_j(\boldsymbol{x}) F_j(\boldsymbol{\theta})$$

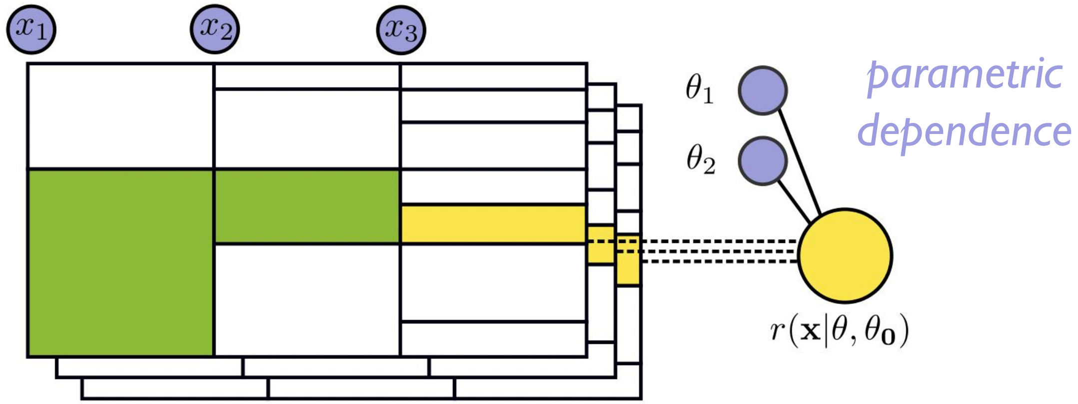
Can solve for degrees of freedom of the predictor
→ *Large training speedup*

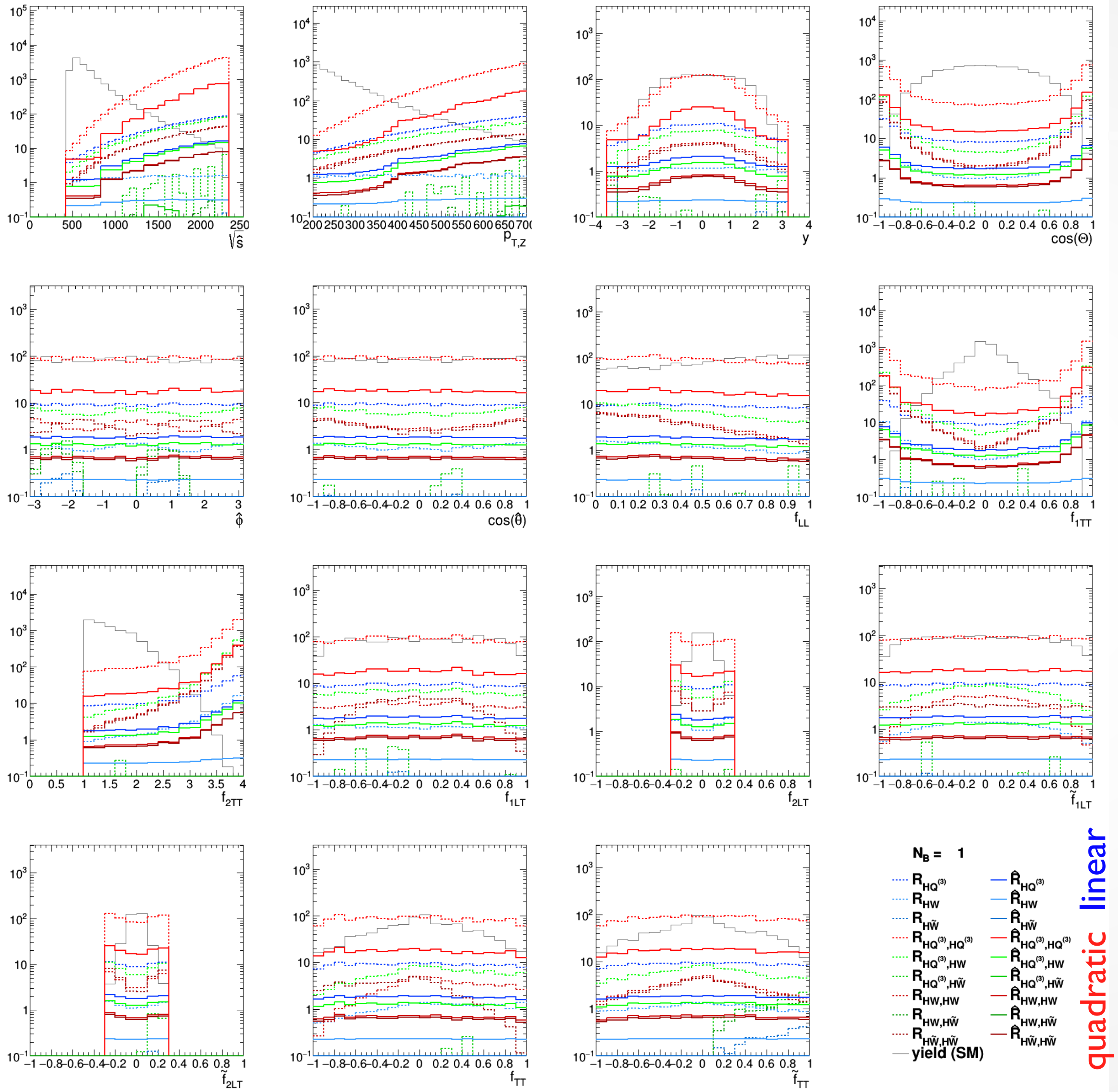
$$F_j(\boldsymbol{\theta}) = \frac{\sum_{i\in j} w_i(\boldsymbol{\theta})}{\sum_{i\in j} w_i(\boldsymbol{\theta}_0)} \equiv \frac{w_j(\boldsymbol{\theta})}{w_j(\boldsymbol{\theta}_0)}$$

Obtain loss function for optimal partitioning, solved by e.g. CART algorithm → Boost

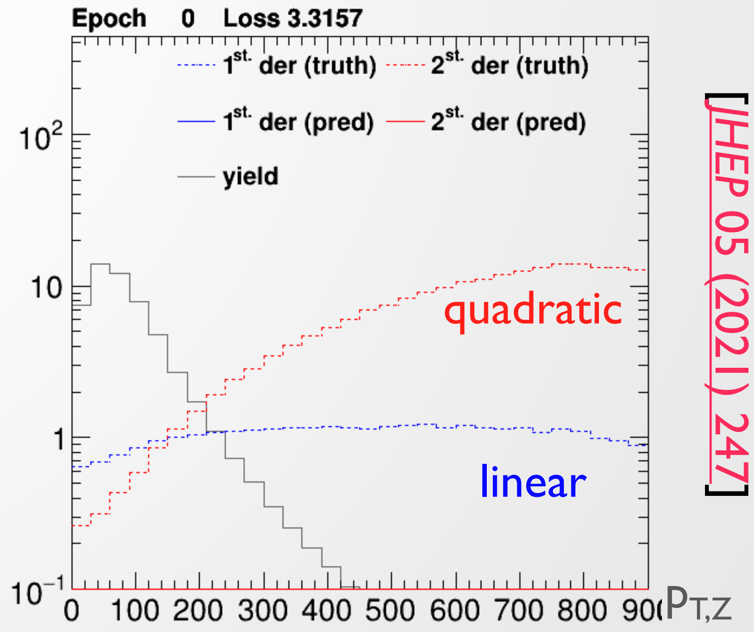
$$L = -\sum_{\boldsymbol{\theta}\in\mathcal{B}} \sum_{j\in\mathcal{J}} \frac{w_j^2(\boldsymbol{\theta})}{w_j(\boldsymbol{\theta}_0)}$$

linear truncation: optimize Fisher information





- Test-case: models of ZH and WZ
  - per-event weighting strategy
- Left: “Boosted Information Tree (BIT)”
  - 3 WC, 9 DOF, 500k events, ZH
  - 200 trees, D=5, 9 minutes of training
  - also more realistic study, including backgrounds [2107.10859], [2205.12976]
- Bottom: (weighted) Parametrized Classifiers
  - 1 WC, 2 DOF, 500k events, ZH, 2xNN

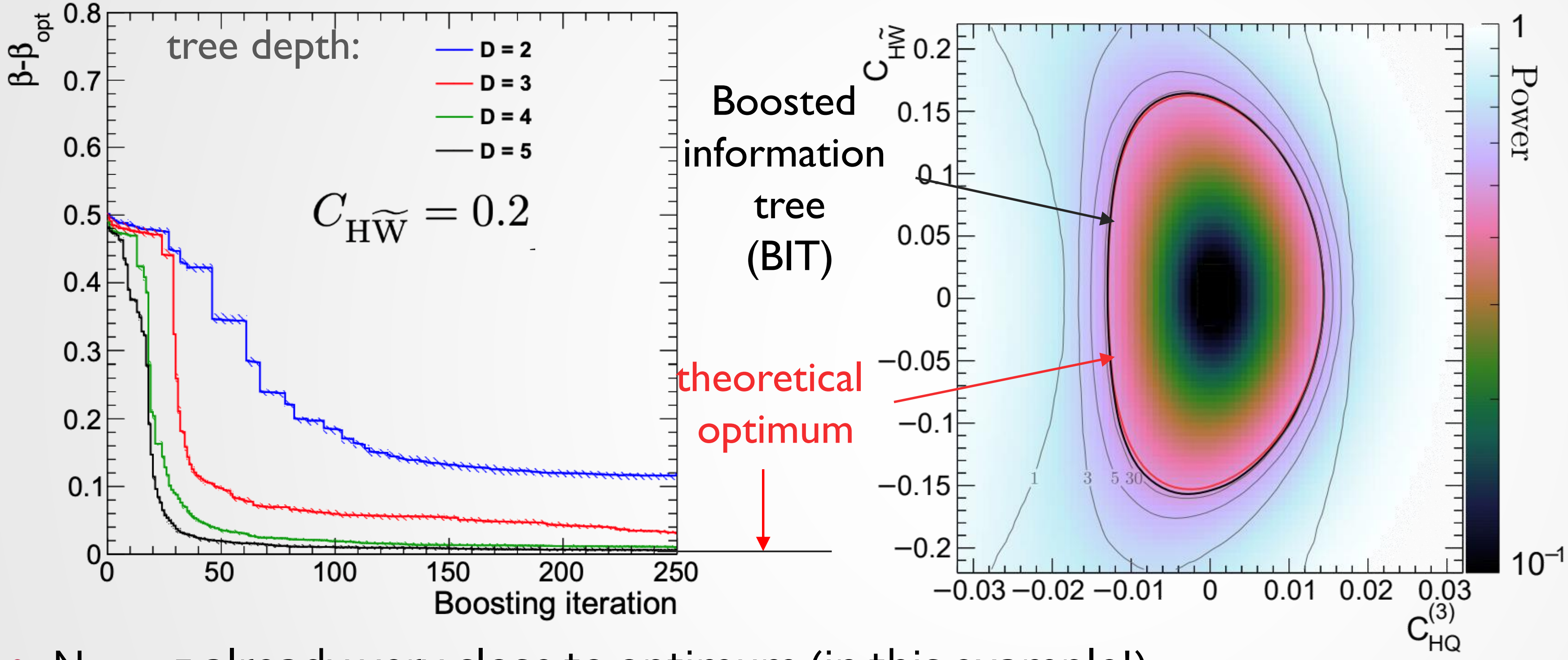


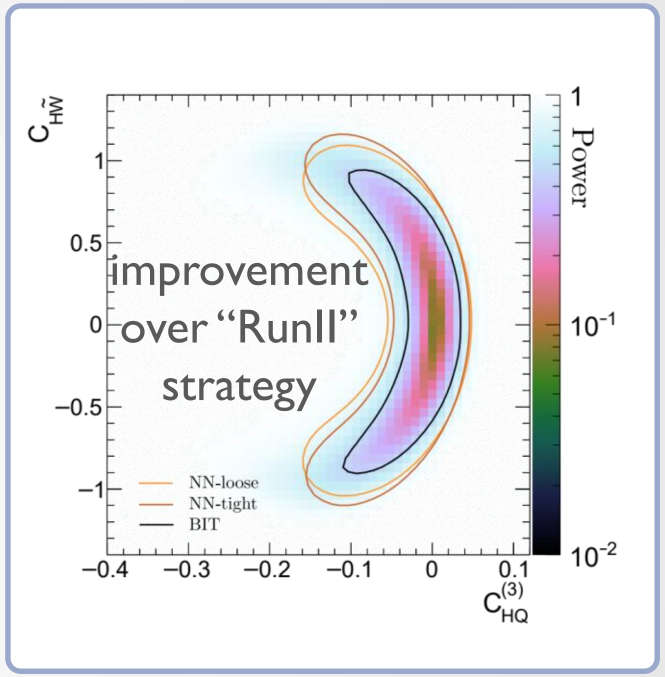
weighted training based on [JHEP 05 (2021) 247]

BINNING VS. OPTIMALITY

[arXiv:2107.10859, arXiv:2205:12976]

- Optimality holds for unbinned LL ratio tests. How important is the binning?

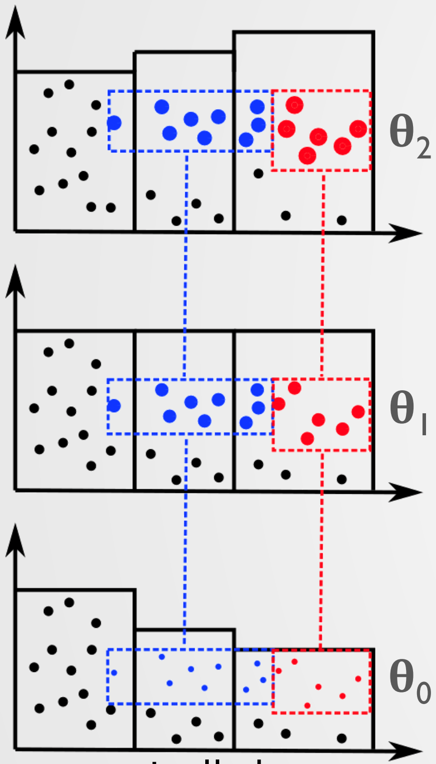




- N_{bin} = 5 already very close to optimum (in this example!)
  - How to chose? Smooth interpolation: Form N_{bin} evenly sized quantiles of p($r(x|\boldsymbol{\theta},SM)$ | SM)
- No free lunch – Analysis dependent choices are needed
  - A case-by-case compromise if background estimation is CPU intensive
  - Systematics treatment for unbinned analyses (beyond $M_{4\ell}$) less far developed

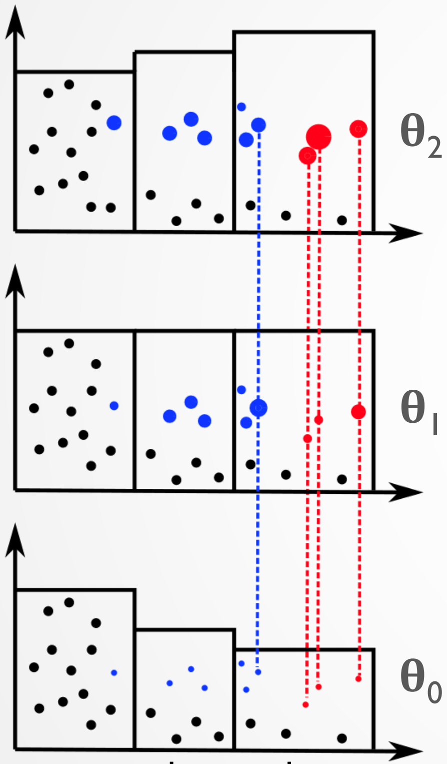
PRACTICALITIES OF SMEFT WEIGHTING

simulating interference & (dim-6)^2 ***samples***



- generically better sampling of EFT-relevant phase space
- need one sample per term in **θ** expansion

per-event weighting



- no independent stochastics in EFT-term
- better overtraining if weight distribution is uniform

- single-operator insertions: quadratic x-sec

$$d\sigma(\boldsymbol{\theta}) \propto |\mathcal{M}_{\mathrm{SM}}(\boldsymbol{z}) + \boldsymbol{\theta}_a \mathcal{M}^a_{\mathrm{BSM}}(\boldsymbol{z})|^2 d\boldsymbol{z}$$

  - compute probabilistic mass as polynomial event weights $w_i(\boldsymbol{\theta})$

1. ***sample-based***: Expand MEs and simulate independent samples
2. ***sampling at a fixed*** $\boldsymbol{\theta}_0$ evaluate $d\sigma(\boldsymbol{\theta})/dz$ for at base-points $\boldsymbol{\theta}$

$$w_i(\boldsymbol{\theta}) = w_{i,0} + \sum_a w_{i,a}\,\theta_a + \frac{1}{2}\sum_{a,b} w_{ab}\,\theta_a\theta_b = \frac{\sigma(\boldsymbol{\theta})}{\sigma(\boldsymbol{\theta}_0)} \cdot r(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta}, \boldsymbol{\theta}_0)$$

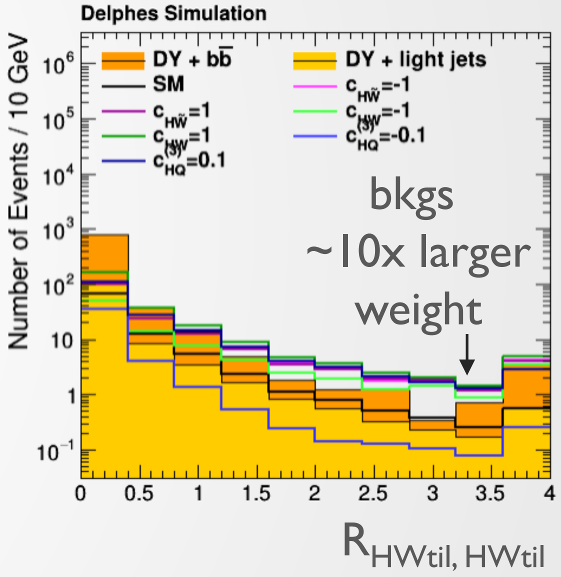
SM   interference   pure SMEFT   *interpret as "joint" LR*

- more robust EFT phase space coverage in **1.**
- higher ML sample efficiency in ***2.***
  - Comparative study [Cranmer et.al. 1808.00973]
  - Reduces risk of overtraining ML-training
- differences don't matter (much) for yield predictions

PRACTICALITIES OF EVENT WEIGHTING & ML

1. Training weight distribution can be uneven
   - In particular in the presence of backgrounds, e.g., Drell-Yan tails → factor 10^2 from event to event at SM
   - Can (/need) to regularize the regressors → much experience for NNs and trees

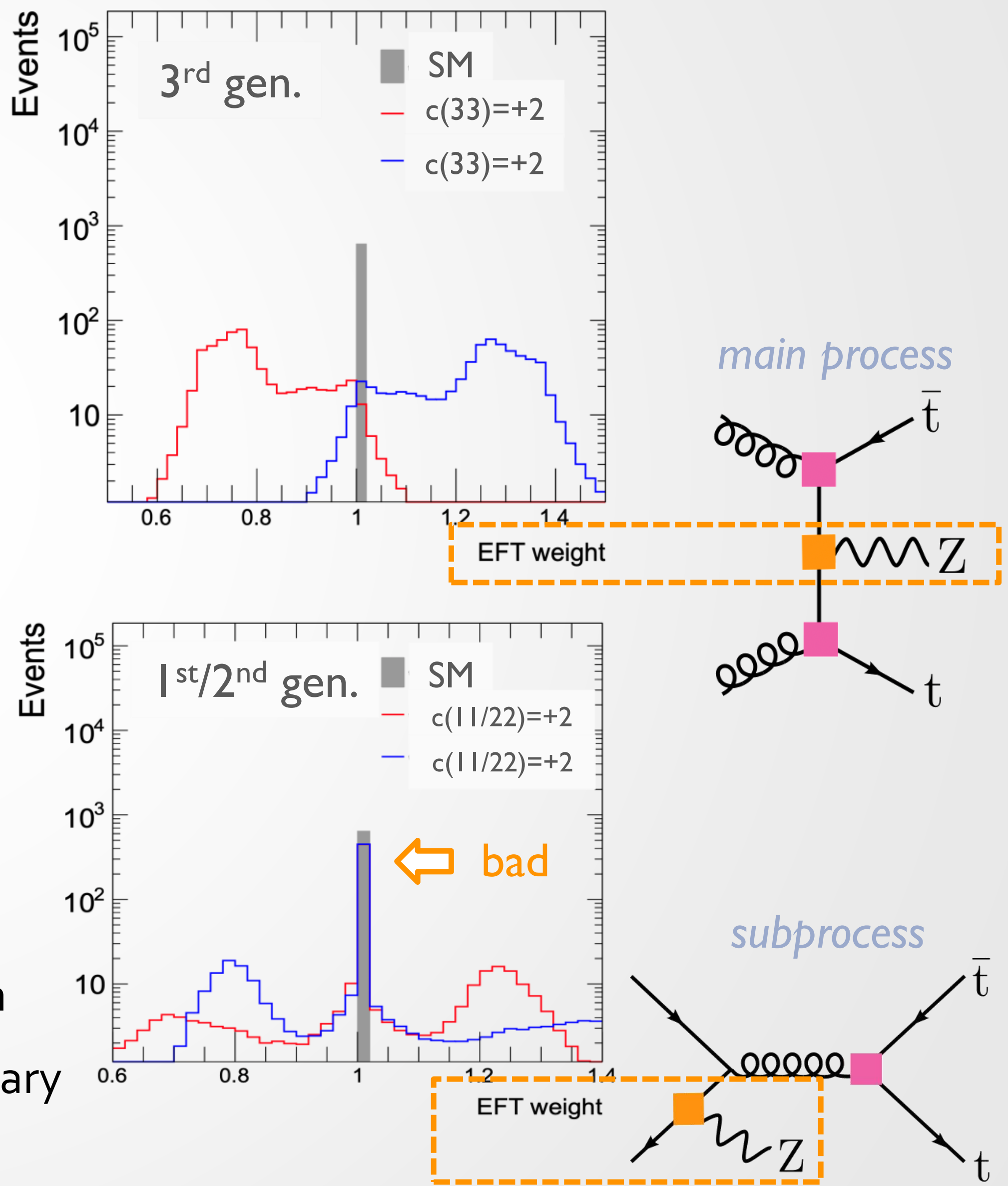
2. Negative weights in NLO samples
   - Spoil statistical interpretation of the *empirical* "joint pdf": $r(x_i,z_i|\boldsymbol{\theta},\text{SM})$
   - Empirical loss function not locally positive definite
   - What was "Overtraining" in LO samples may be a loss of convergence at NLO
     - (My experience:) only pathological toy cases
     - Tighten regulator by ≈ (1+f)/(1-f) with $f = n^+/n^-$
     - All yields are positive in large sample limit
   - Algorithms support positivity constraint (slows convergence)
     - In practice: Not needed, so far; although stricter regularization needed
   - If problematic: Maybe possible to develop new ideas based on, e.g., NLO re-sampling [B. Nachman, J. Thaler 2007.11586]



MORE ON THE SIMULATED SMEFT PREDICTION

3. Worse stochastics when rare subprocess is affected
   - Example: Consider the q-Z vector coupling in ttZ
   - Compare 1st/2nd with 3rd generation SMEFT effects:
   - 1st/2nd generation operators enter only in events where the Z couples to the initial state quark
     - w($\boldsymbol{\theta}^{(33)}$=±2) ≠ 1 for a most events → OK
     - w($\boldsymbol{\theta}^{(11,22)}$=±2) = 1 for a large fraction → reduces stats
   - A feature of the process, not the weighting strategy
   - Several ML tools to estimate variance of the estimator, not (to the best of my knowledge) used for SMEFT ML

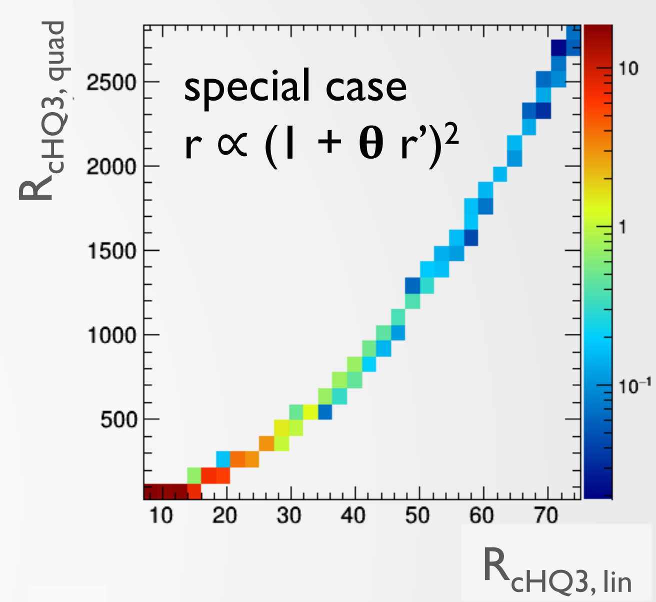
4. Linear and quadratic terms may be (perfectly) correlated.
   - For a point-like interaction r = (1 + $\boldsymbol{\theta}$ r′)², e.g. c_{HQ3} in VH at LO
   - backgrounds distort the relation but not break the correlation
   - Lesson: Case-by-case understanding of test statistic is necessary

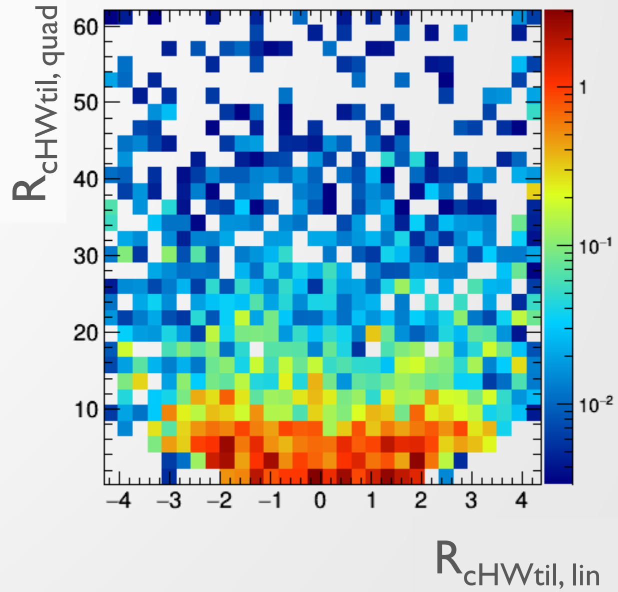


MORE ON THE SIMULATED SMEFT PREDICTION

3. Worse stochastics when rare subprocess is affected
   - Example: Consider the q-Z vector coupling in ttZ
   - Compare 1st/2nd with 3rd generation SMEFT effects:
   - 1st/2nd generation operators enter only in events where the Z couples to the initial state quark
     - $w(\boldsymbol{\theta}^{(33)}=\pm 2) \neq 1$ for a most events → OK
     - $w(\boldsymbol{\theta}^{(11,22)}=\pm 2) = 1$ for a large fraction → reduces stats
   - A feature of the process, not the weighting strategy
   - Several ML tools to estimate variance of the estimator, not (to the best of my knowledge) used for SMEFT ML

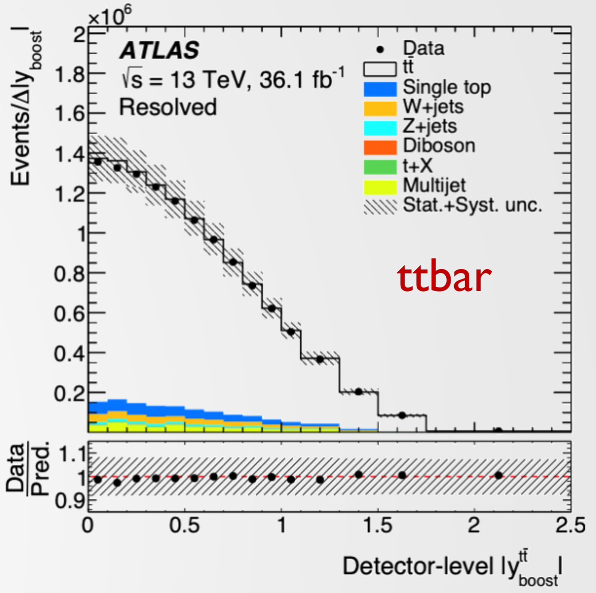
4. Linear and quadratic terms may be (perfectly) correlated.
   - For a point-like interaction $r = (1 + \boldsymbol{\theta}\, r')^2$, e.g. c_{HQ3} in VH at LO
   - backgrounds distort the relation but not break the correlation
   - Lesson: Case-by-case understanding of test statistic is necessary

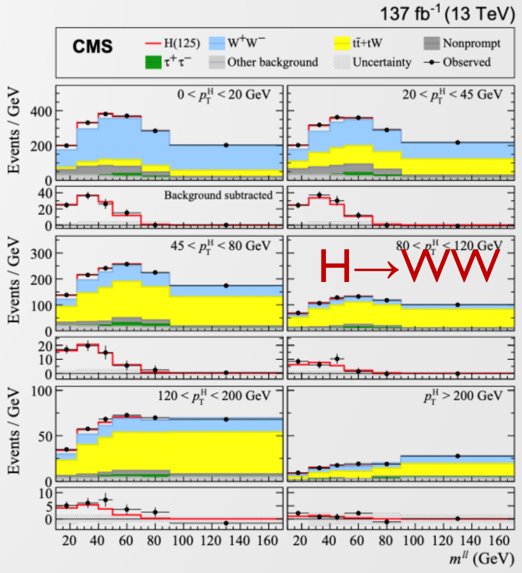




HELICITY-SUMMED REWEIGHTING

5. Operators with non-SM helicity configurations can't be weighted from the SM point *per helicity.*
   - extreme case of difference in phase space coverage
   - Approach #1: weight-based simulation for c≠0,
     to ensure sampling of all helicity configurations (used, e.g., in CMS)
   - Approach #2: Helicity-summed reweighting [arxiv:1607.00763]
     - Option in MG, not widely used but available in reweighting tools
     - Can reweight different models, provided LHE information is accessible
   - Helicity-summed reweighting *preserves* the possibility of in-experiment reinterpretation
     - No longer bound to initial choice of model?
     - need to keep LHE information from the events in SR
   - Persistency is important
     - If non-zero signal: Need to solve background correlations as triangular matrix
     - We will need Multi-differential high-dimensional SM-EFT analysis of candles





POSSIBLE ACTION ITEMS / CONCLUSION

A. Suggest comparative study of all approaches (& aware of STXS parametrisation)

	Sample based	**Event based (per hel.)**	**Event based (summed hel.)**
Persistency	1 sample per term	1 number per event & term	1 number per event & term
Madgraph reweighting	None	$W_{new} = \frac{\lvert M^h_{new}\rvert^2}{\lvert M^h_{orig}\rvert^2} W_{orig}$	$W_{new} = \frac{\lvert M_{new}\rvert^2}{\lvert M_{orig}\rvert^2} W_{orig}$
Phase space mismatches?	No problem	May require c≠0	May require c≠0 (fewer cases)
Can be staged?	Yes, including hel.	No new hel.	Yes, including hel.
ML sample efficient?	Less so	Yes	Yes

B. Analysis persistency for later reinterpretation – tools, practises, shortcomings

- Are we ready for an excess?

C. Best practices for publication of ML results

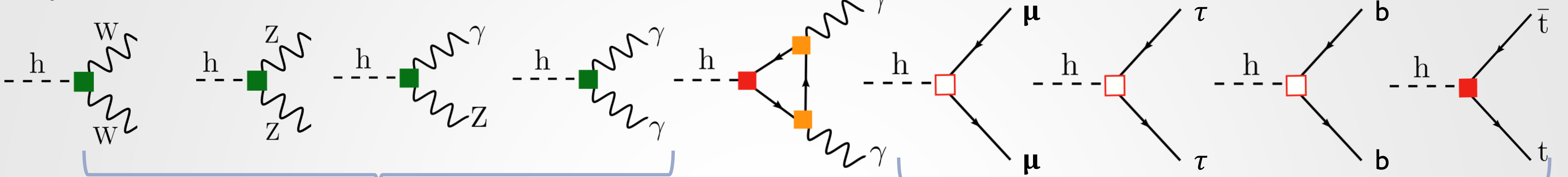
- proposal [Publishing statistical models: Getting the most out of particle physics experiments, 2109.0491]

PER-SAMPLE VS REWEIGHTING – LOSS FUNCTIONS

$$
\begin{aligned}
L &= \int \mathrm{d}\sigma_{\boldsymbol{\theta}} \hat{f}(\boldsymbol{x})^2 + \int \mathrm{d}\sigma_{\mathrm{SM}} (1 - \hat{f}(\boldsymbol{x}))^2 \\
&= \int \mathrm{d}\sigma_{\mathrm{SM}} \left(\frac{\mathrm{d}\sigma_{\boldsymbol{\theta}}}{\mathrm{d}\sigma_{\mathrm{SM}}} \hat{f}(\boldsymbol{x})^2 + (1 - \hat{f}(\boldsymbol{x}))^2 \right) \\
&= \int \mathrm{d}\boldsymbol{x}\, p(\boldsymbol{x}, \boldsymbol{z} | \mathrm{SM}) \left(r(\boldsymbol{x}, \boldsymbol{z} | \boldsymbol{\theta}, \mathrm{SM}) \hat{f}(\boldsymbol{x})^2 + (1 - \hat{f}(\boldsymbol{x}))^2 \right)
\end{aligned}
$$

DISENTANGLING SM-EFT IN THE HIGGS-SECTOR

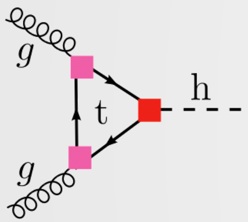
decay: 21% 2.6% 0.15% 0% 0.23% 0.02% 6.3% 58%



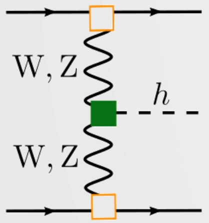
Higgs – Boson couplings

Higgs – Boson loop induced

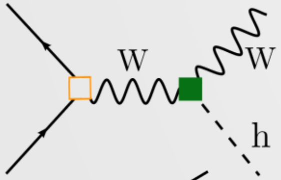
Higgs – Fermion couplings



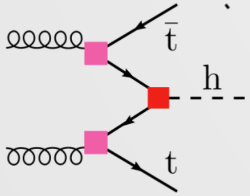
45 pb gluon fusion



3.5 pb vector-boson fusion



1.2 pb associate W



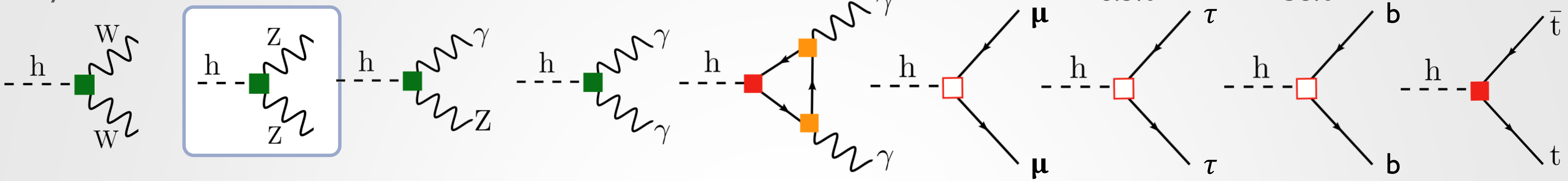
0.5 pb ttH

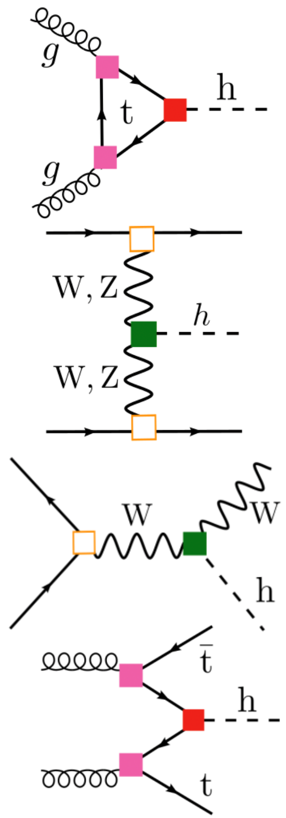
Higgs production modes with their SM-EFT couplings

THE HIGGS IN THE GOLDEN CHANNEL

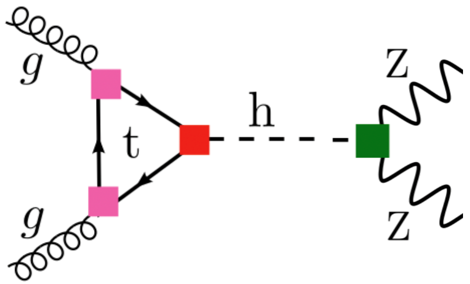
EPJC 80 (2020) 957

decay: 21% 2.6% 0.15% 0% 0.23% 0.02% 6.3% 58%





$H \to ZZ^* \to 4l$

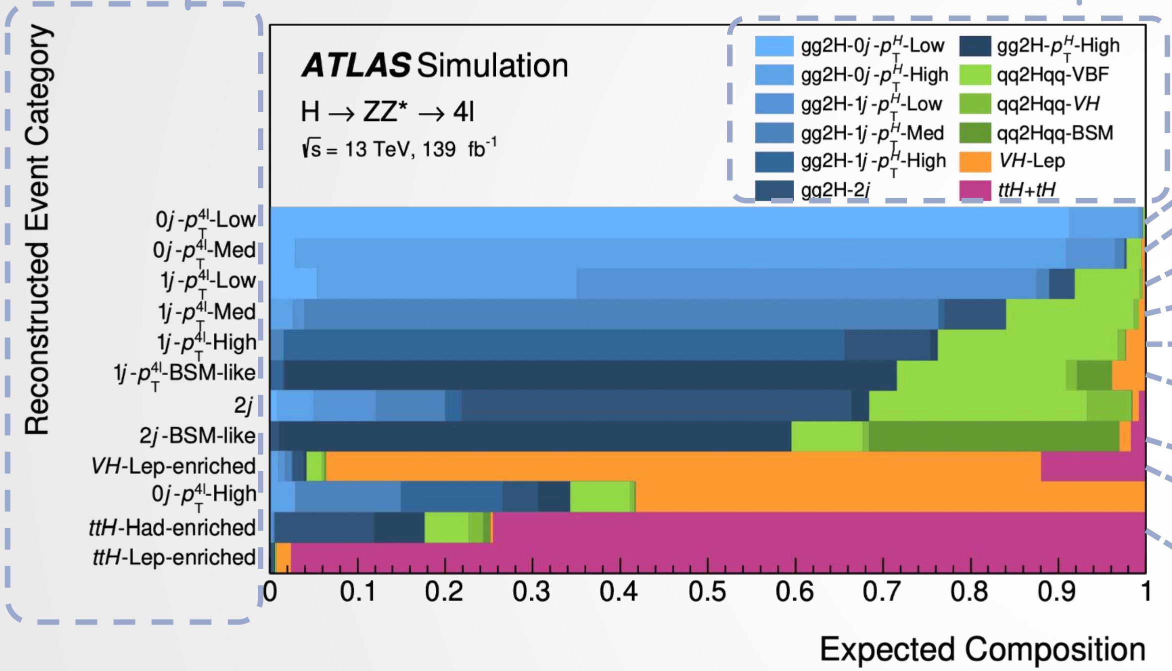


- example #1: ZZ^* decay channel in all production modes
- experimentally clean ("golden channel")
- 10 = 5 (+5 CP odd) operators: c_{HW}, c_{HB}, c_{HW}, c_{uH}, c_{HWB}
- attempt to optimally disentangle production modes

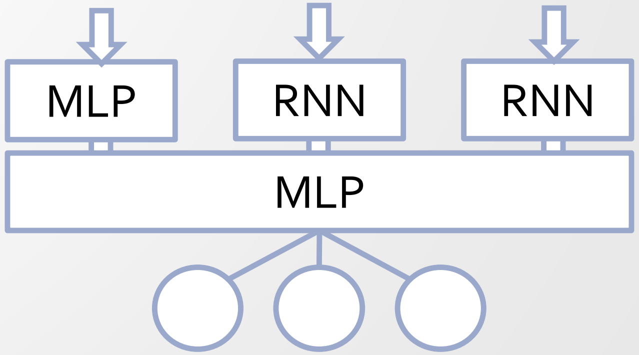
THE HIGGS IN THE GOLDEN CHANNEL

EPJC 80 (2020) 957

- Reconstructed bins contain a mixture of production channels and backgrounds (mostly ZZ*)



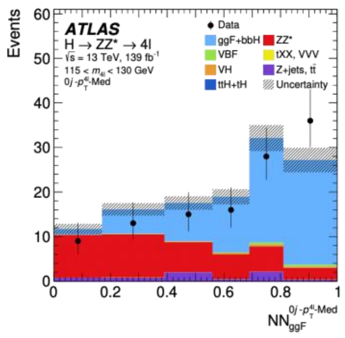
Processes	MLP	Lepton RNN	Jet RNN
ggF, ZZ^*	$p_T^{4\ell}, D_{ZZ^*}, m_{12}, m_{34}$, $\lvert\cos\theta^*\rvert, \cos\theta_1, \phi_{ZZ}$	p_T^ℓ, η_ℓ	-
ggF, VBF, ZZ^*	$p_T^{4\ell}, p_T^j, \eta_j$, $\Delta R_{4\ell j}, D_{ZZ^*}$	p_T^ℓ, η_ℓ	-
ggF, VBF, ZZ^*	$p_T^{4\ell}, p_T^j, \eta_j, E_T^{miss}$, $\Delta R_{4\ell j}, D_{ZZ^*}, \eta_{4\ell}$	p_T^ℓ, η_ℓ	-
ggF, VBF	$p_T^{4\ell}, p_T^j, \eta_j$, $E_T^{miss}, \Delta R_{4\ell j}, \eta_{4\ell}$	p_T^ℓ, η_ℓ	-
ggF, VBF, VH	$m_{jj}, p_T^{4\ell jj}$	p_T^ℓ, η_ℓ	p_T^j, η_j
ggF, VBF	$\eta_{ZZ}^{Zepp}, p_T^{4\ell jj}$	p_T^ℓ, η_ℓ	p_T^j, η_j
VH, ttH	$N_{jets}, N_{b\text{-jets},70\%}$, E_T^{miss}, H_T	p_T^ℓ	-
ggF, ttH, tXX	$p_T^{4\ell}, m_{jj}$, $\Delta R_{4\ell j}, N_{b\text{-jets},70\%}$,	p_T^ℓ, η_ℓ	p_T^j, η_j

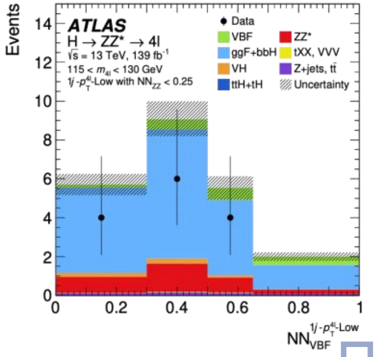


- ML is used to separate production modes in each category
- per reco-channel: NNs trained with 2-7 observables
  - combine with RNNs (LSTMs) using variable-length jets and leptons
  - common network layer for multiclassification in e.g., ggF, VBF, ZZ*

THE HIGGS IN THE GOLDEN CHANNEL

EPJC 80 (2020) 957





Likelihood = prod. of Poissonians

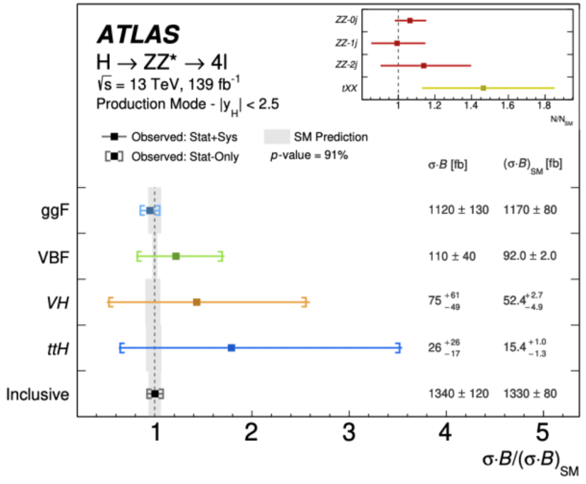
auxiliary measurements

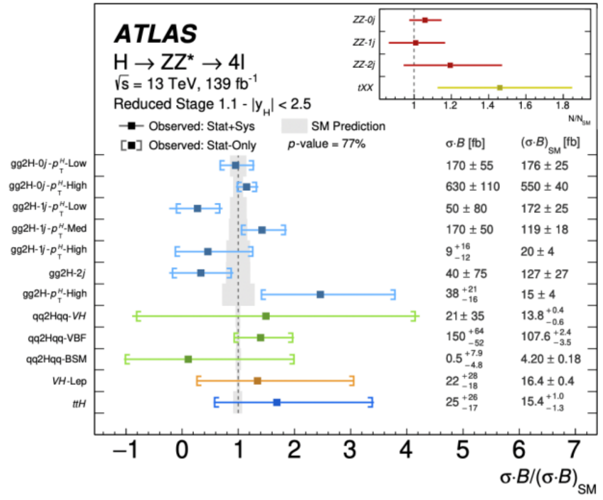
$$L(\mathcal{D}|\boldsymbol{\sigma},\boldsymbol{\nu}) = \prod_{j=1}^{N_{\text{cat.}}} \prod_{i=1}^{N_{\text{bin}}^{(j)}} \text{Pois}\left(N_{i,j} \middle| \mathcal{L} \sum_{p=1}^{N_{\text{prod}}} \sigma_{\text{SM}}^{(p)} \mathcal{B}_{\text{SM}}^{4\ell} A_{i,j}^{(p)}(\boldsymbol{\nu}) + B_{i,j}(\boldsymbol{\nu}) \right) \times \prod_{m=1}^{N_{nuis.}} C_m(\boldsymbol{\nu})$$

Log-likelihood ratio test statistic

$$q(\boldsymbol{\sigma}) = -2\log \frac{L(\mathcal{D}|\boldsymbol{\sigma}, \hat{\hat{\boldsymbol{\nu}}}_{\boldsymbol{\sigma}})}{L(\mathcal{D}|\hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\nu}})}$$

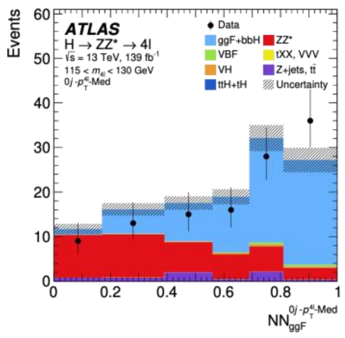
(profiled, to deal with nuisances)

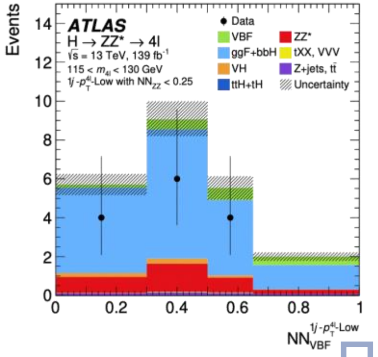




THE HIGGS IN THE GOLDEN CHANNEL

EPJC 80 (2020) 957





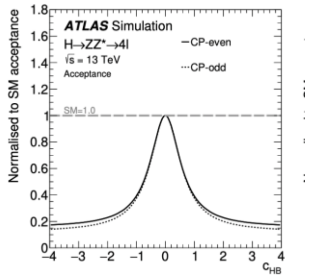
signal-strength modifiers

defined at the fiducial level

production | decay | acceptance

$$\mu^{(p)}(\boldsymbol{\theta}) = \frac{\sigma^{(p)}(\boldsymbol{\theta})}{\sigma^{(p)}_{\mathrm{SM}}} \times \frac{\mathcal{B}^{4\ell}(\boldsymbol{\theta})}{\mathcal{B}^{4\ell}_{\mathrm{SM}}} \times \frac{A(\boldsymbol{\theta})}{A_{\mathrm{SM}}}$$

universal!



Likelihood = prod. of Poissonians

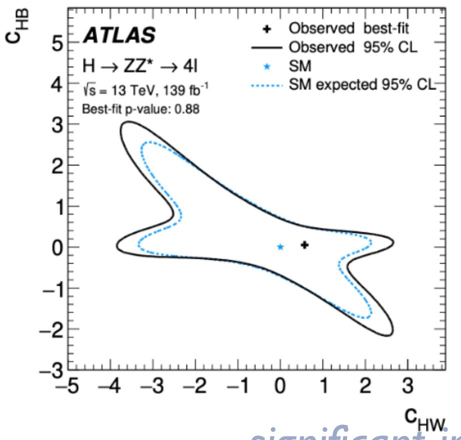
auxiliary measurements

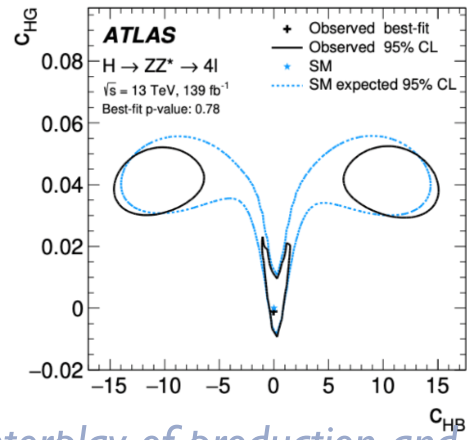
$$L(\mathcal{D}|\boldsymbol{\theta},\boldsymbol{\nu}) = \prod_{j=1}^{N_{\mathrm{cat.}}} \prod_{i=1}^{N^{(j)}_{\mathrm{bin}}} \mathrm{Pois}\left(N_{i,j} \middle| \mathcal{L} \sum_{p=1}^{N_{\mathrm{prod}}} \mu^{(p)}(\boldsymbol{\theta}) \sigma^{(p)}_{\mathrm{SM}} \mathcal{B}^{4\ell}_{\mathrm{SM}} A^{(p)}_{i,j}(\boldsymbol{\nu}) + B_{i,j}(\boldsymbol{\nu}) \right) \times \prod_{m=1}^{N_{nuis.}} C_m(\boldsymbol{\nu})$$

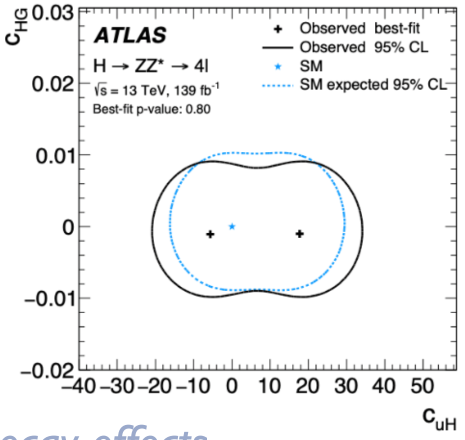
Log-likelihood ratio test statistic

$$q(\boldsymbol{\theta}) = -2 \log \frac{L(\mathcal{D}|\boldsymbol{\theta}, \hat{\hat{\boldsymbol{\nu}}}_{\boldsymbol{\theta}})}{L(\mathcal{D}|\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\nu}})}$$

(profiled, to deal with nuisances)





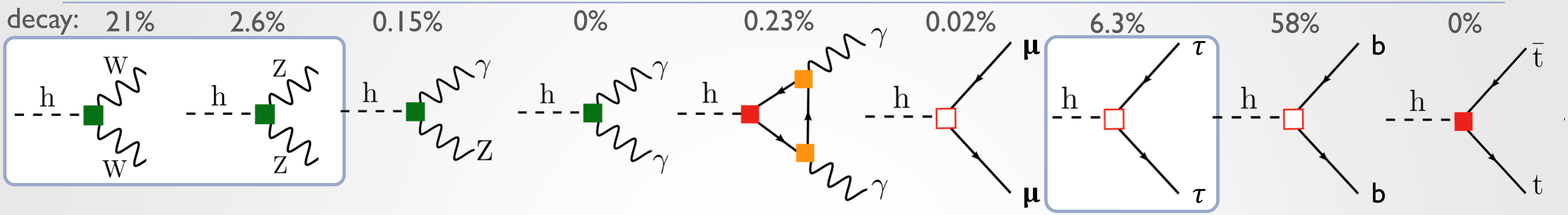


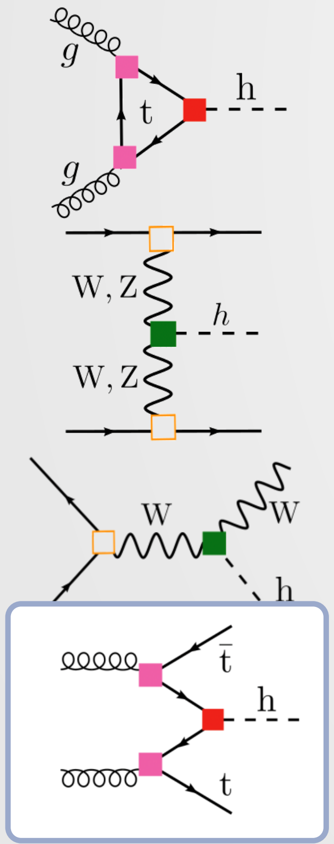
significant interplay of production and decay effects

learn "only" the likelihood ratio of different SM production modes

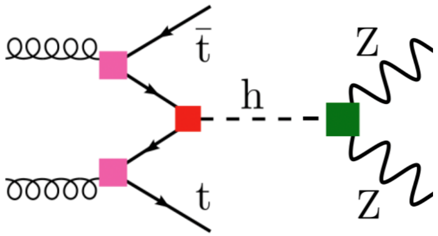
TTH IN THE MULTILEPTON CHANNEL

JHEP (submitted)



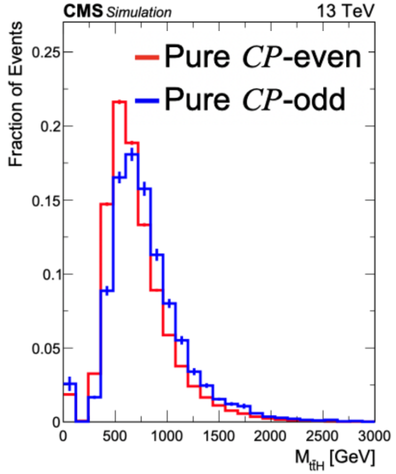


ttH multilepton



- example #2: t(t)H multilepton in 2ℓSS+0τ, 2ℓSS+1τ, 3ℓ final states
- 3 DNNs for signal/background multi-classification
- targets t-t-H Yukawa coupling (■) in κ-framework
- in SM-EFT: "CP" structure (complex phase) of $HH^\dagger \bar{q}_p u_r \tilde{H}$

- use ML for separating CP-even vs. odd effects
  - gradient-BDT XGBoost
  - 38 input features (kinematic properties)

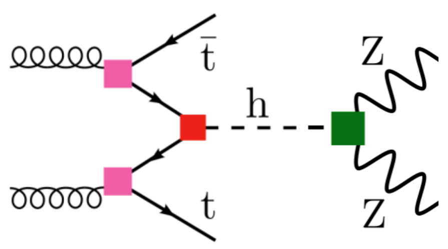


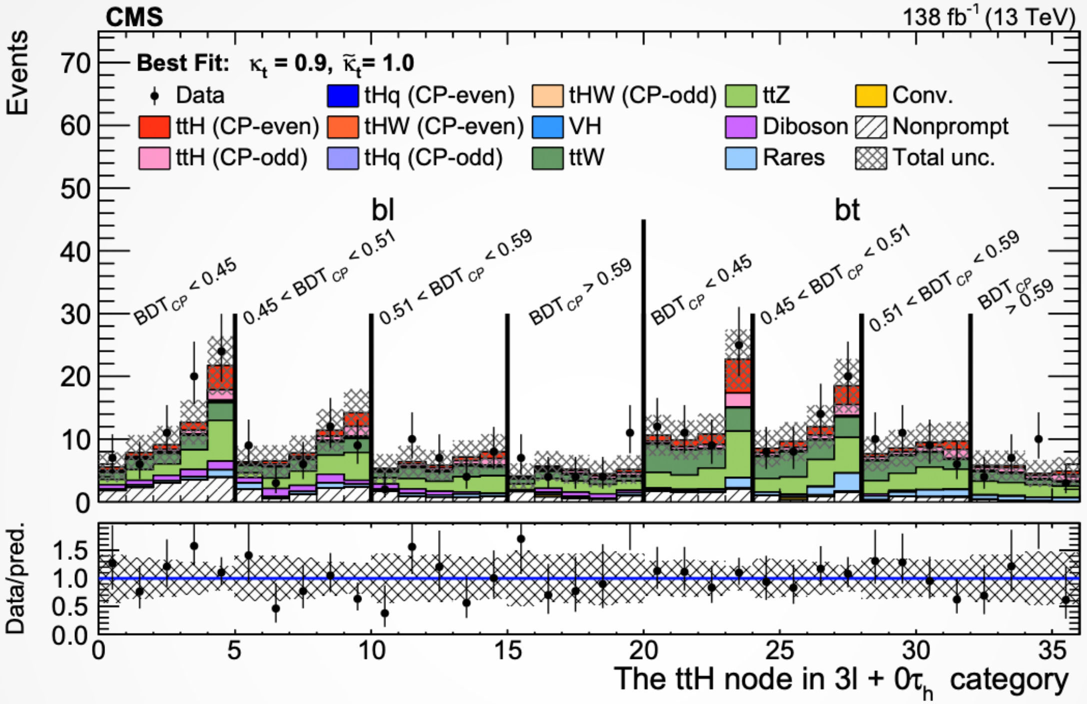
+ 37 other observables

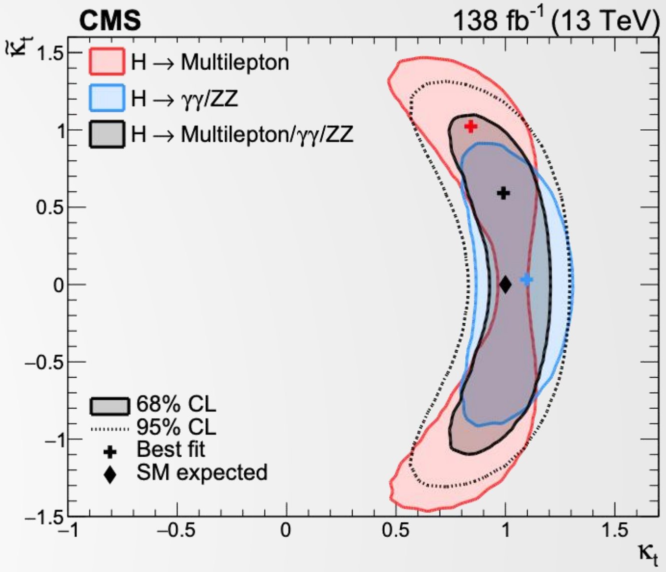
TTH IN THE MULTILEPTON CHANNEL

[JHEP (submitted)]

ttH multilepton







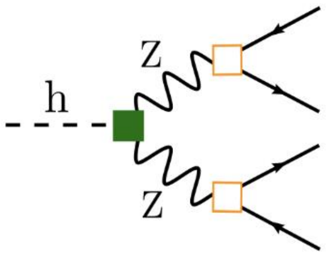
- BDT exploits the likelihood trick to obtain CP even/odd fraction from the data
- limits on deviations of the t-t-H interaction (κ_t , $\tilde{\kappa}_t$) including combinations with other final states
- example of learning "of" SM-EFT effects
- issue: large top backgrounds from ttZ and ttW in all measurement regions → combine sectors!
- τ lepton ID performance has significant impact

RECENT SM-EFT RESULTS (SELECTION!)

4l JHEP 07 (2021) 005

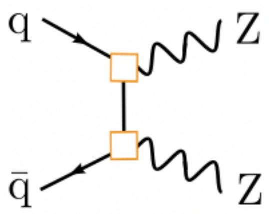
ATL-PHYS-PUB-2021-010

$H \to Z^*Z$



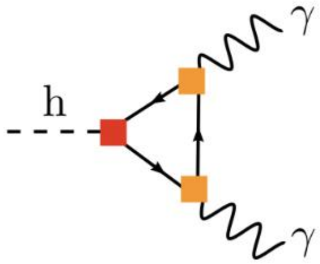
$H \to 4l$ EPJC 80 (2020) 957
$H \to 4l$ PRD 104, 052004 (2021)

ZZ



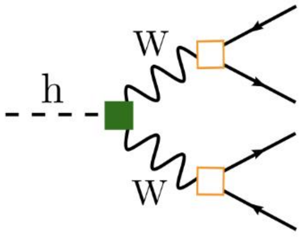
PRD 97 (2018) 032005
EPJC 81 (2021) 200

$H \to \gamma\gamma$



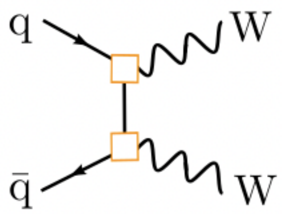
arxiv:2202.00487

$H \to W^*W$



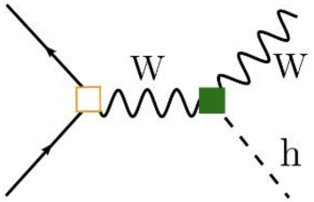
$H \to WW$, eμ
EPJC 82 (2022) 622

$W^\pm W^\mp$



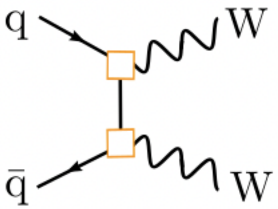
$W^\pm W^\mp$
PRD 102, 092001 (2020)

W/Z+H ($H \to bb$)



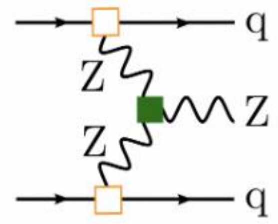
resolved EPJC 81 (2021) 178
boosted PLB 816 (2021) 136204

$W^\pm W^\mp$ (+ ≥ 1 jet)



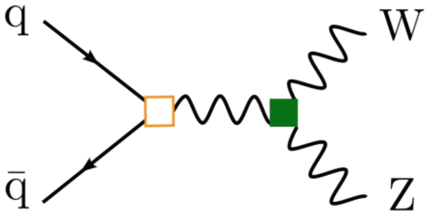
JHEP 06 (2021) 003
PRD102, 092001 (2020)

VBF Z + jj



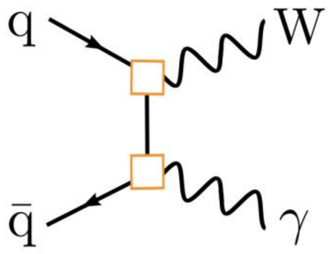
EPJC 81(2021)163

WZ



CMS-SMP-20-0014

Wγ



PRD sub.
CMS-SMP-20-005

RECENT SM-EFT RESULTS (SELECTION!)

4l JHEP 07 (2021) 005 | ATL-PHYS-PUB-2021-010

H→Z*Z

O_{uH} $HH^\dagger \bar{q}_p u_r \tilde{H}$
O_{HG} $HH^\dagger G^A_{\mu\nu} G^{\mu\nu A}$
O_{HW} $HH^\dagger W^l_{\mu\nu} W^{\mu\nu l}$
O_{HB} $HH^\dagger B_{\mu\nu} B^{\mu\nu}$
O_{HWB} $HH^\dagger \tau^l W^l_{\mu\nu} B^{\mu\nu}$
+CP odd

H→4l EPJC 80 (2020) 957
H→4l PRD 104, 052004 (2021)

ZZ

aTGC

PRD 97 (2018) 032005
EPJC 81 (2021) 200

H→γγ

O_{uH} $HH^\dagger \bar{q}_p u_r \tilde{H}$
O_{HG} $HH^\dagger G^A_{\mu\nu} G^{\mu\nu A}$
O_{HW} $HH^\dagger W^l_{\mu\nu} W^{\mu\nu l}$
O_{HB} $HH^\dagger B_{\mu\nu} B^{\mu\nu}$
O_{HWB} $HH^\dagger \tau^l W^l_{\mu\nu} B^{\mu\nu}$

arxiv:2202.00487

H→W*W

HC framework
CP even/odd
(O_{HW}, O_{HB} + CP odd)

H→WW, eμ
EPJC 82 (2022) 622

W±W∓

$$\mathcal{O}_{WWW} = \frac{c_{WWW}}{\Lambda^2} W_{\mu\nu} W^{\nu\rho} W_\rho{}^\mu$$
$$\mathcal{O}_W = \frac{c_W}{\Lambda^2} (D^\mu \Phi)^\dagger W_{\mu\nu} (D^\nu \Phi)$$
$$\mathcal{O}_B = \frac{c_B}{\Lambda^2} (D^\mu \Phi)^\dagger B_{\mu\nu} (D^\nu \Phi)$$

+CP odd

W±W∓
PRD 102, 092001 (2020)

W/Z+H (H→bb)

c_{HWB}	$O_{HWB} = H^\dagger \tau^I H W^I_{\mu\nu} B^{\mu\nu}$
c_{HW}	$O_{HW} = H^\dagger H W^I_{\mu\nu} W^{\mu\nu}_I$
c_{Hq3}	$O^{(3)}_{Hq} = (H^\dagger i \overleftrightarrow{D}^I_\mu H)(\bar{q}_p \tau^I \gamma^\mu q_r)$
c_{Hq1}	$O^{(1)}_{Hq} = (H^\dagger i \overleftrightarrow{D}_\mu H)(\bar{q}_p \gamma^\mu q_r)$
c_{Hu}	$O_{Hu} = (H^\dagger i \overleftrightarrow{D}_\mu H)(\bar{u}_p \gamma^\mu u_r)$
c_{Hd}	$O_{Hd} = (H^\dagger i \overleftrightarrow{D}_\mu H)(\bar{d}_p \gamma^\mu d_r)$
c_{dH}	$O_{dH} = (H^\dagger H)(\bar{q} d H)$

resolved EPJC 81 (2021) 178
boosted PLB 816 (2021) 136204

W±W∓ (+ ≥ 1 jet)

$$\mathcal{O}_{WWW} = \frac{c_{WWW}}{\Lambda^2} W_{\mu\nu} W^{\nu\rho} W_\rho{}^\mu$$
$$\mathcal{O}_W = \frac{c_W}{\Lambda^2} (D^\mu \Phi)^\dagger W_{\mu\nu} (D^\nu \Phi)$$
$$\mathcal{O}_B = \frac{c_B}{\Lambda^2} (D^\mu \Phi)^\dagger B_{\mu\nu} (D^\nu \Phi)$$

JHEP 06 (2021) 003
PRD102, 092001 (2020)

VBF Z + jj

$$\mathcal{O}_{WWW} = \frac{c_{WWW}}{\Lambda^2} W_{\mu\nu} W^{\nu\rho} W_\rho{}^\mu$$
$$\mathcal{O}_W = \frac{c_W}{\Lambda^2} (D^\mu \Phi)^\dagger W_{\mu\nu} (D^\nu \Phi)$$

+CP odd

EPJC 81(2021)163

WZ

$$\mathcal{O}_{WWW} = \frac{c_{WWW}}{\Lambda^2} W_{\mu\nu} W^{\nu\rho} W_\rho{}^\mu$$
$$\mathcal{O}_W = \frac{c_W}{\Lambda^2} (D^\mu \Phi)^\dagger W_{\mu\nu} (D^\nu \Phi)$$

+CP odd

CMS-SMP-20-0014

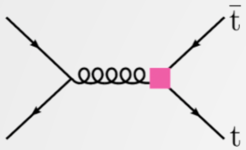
Wγ

$$\mathcal{O}_{WWW} = \frac{c_{WWW}}{\Lambda^2} W_{\mu\nu} W^{\nu\rho} W_\rho{}^\mu$$

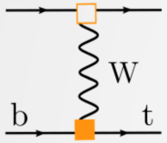
PRD sub.
CMS-SMP-20-005

TOP AND DIBOSON SECTORS

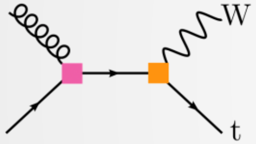
tt
811 pb



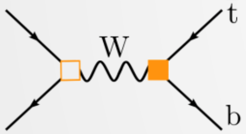
t (t-channel)
217 pb



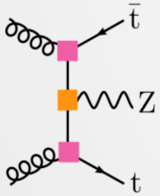
tW
72 pb



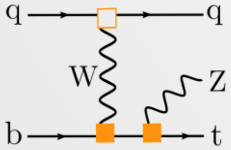
t (s-channel)
10 pb



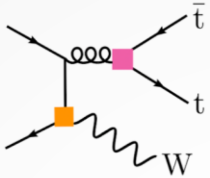
ttZ
1 pb



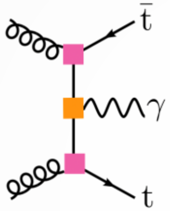
tZq
0.088 pb



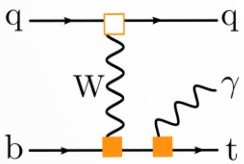
ttW
0.55 pb



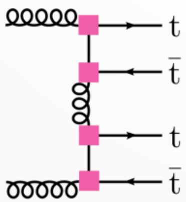
ttɣ
0.80 pb



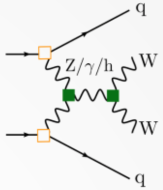
tɣq
0.4 pb

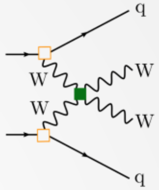


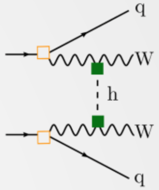
tttt
0.009 pb



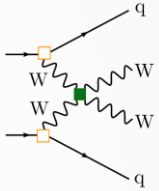
W±W∓
EWK 10 fb

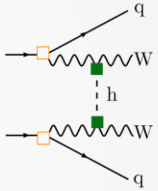




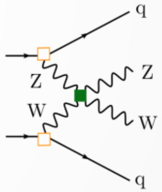


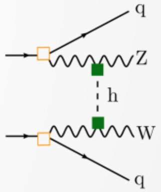
W±W±
EWK 4 fb



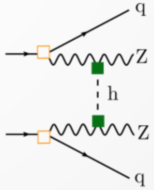


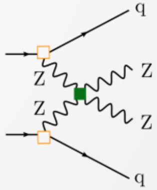
WZ
EWK 2.15 fb



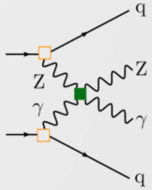


ZZ
EWK 0.82 fb

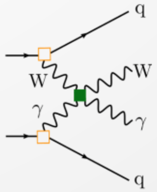




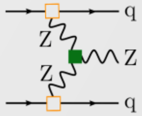
Zɣ
EWK 7.8 fb



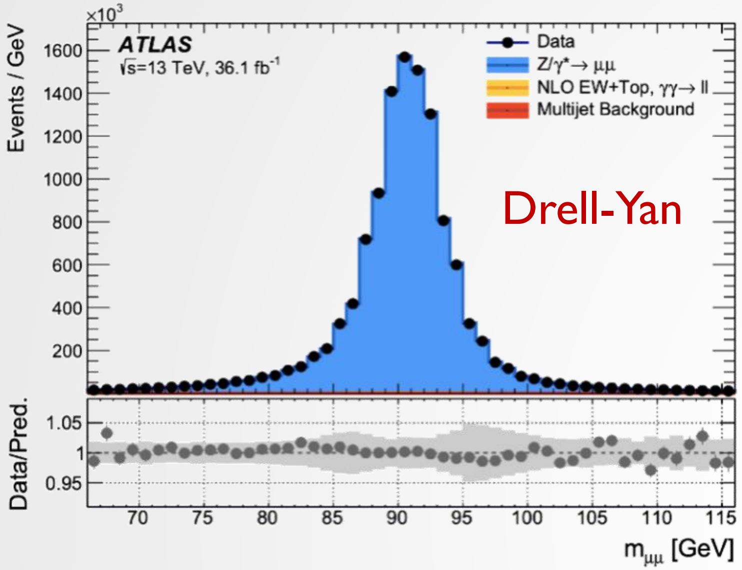
Wɣ
EWK 20.4 fb

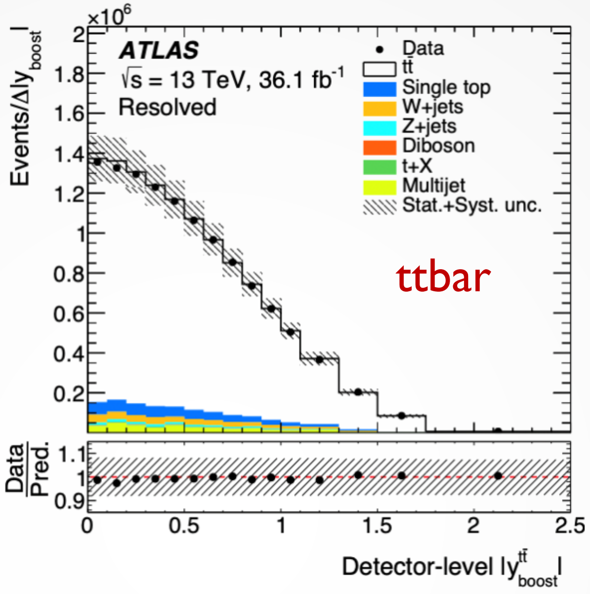


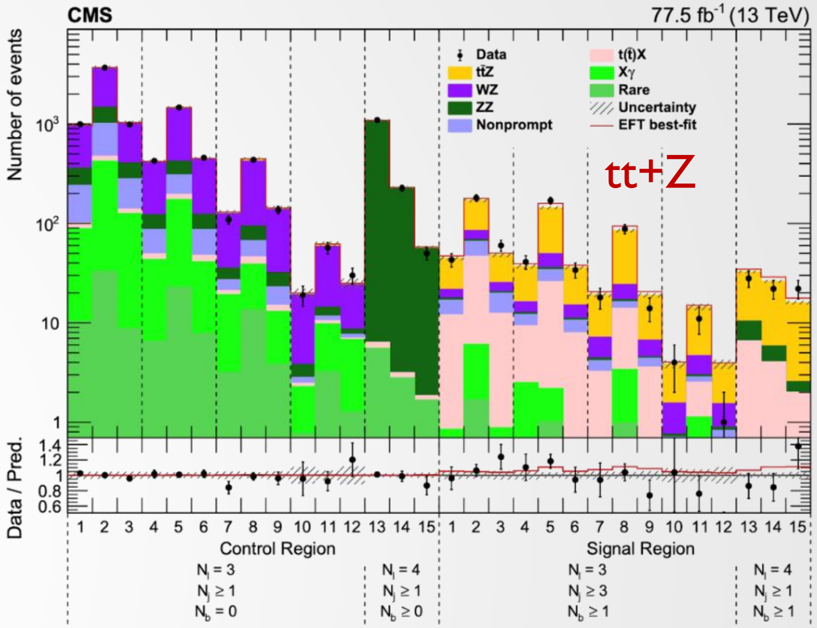
Z
EWK 37.4 fb

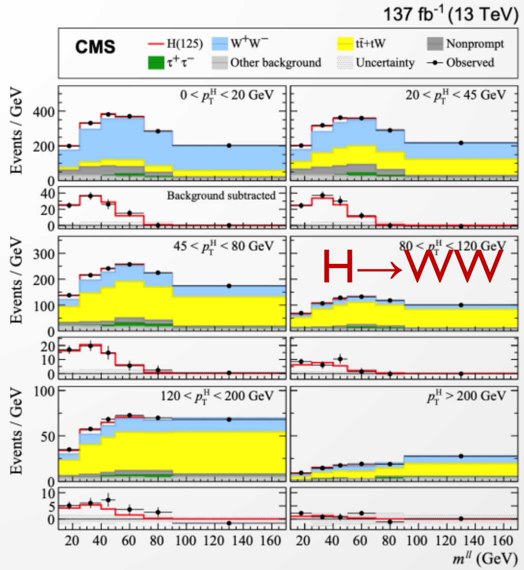


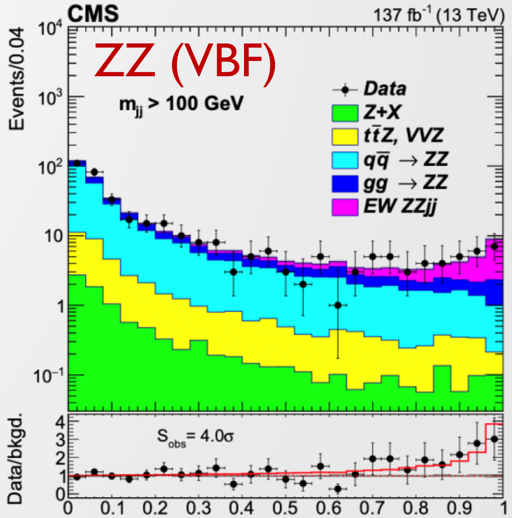
SM-EFT EFFECTS ARE EVERYWHERE











- Solve background correlations like a triangular matrix (i.e. staged):
  - Multi-differential high-dimensional SM-EFT analysis of candles:
    - Drell-Yan, W+Jets, ttbar, single-top (t), etc.
  - Then move to ZH (+ Drell-Yan), WH (+ttbar), H→WW (+WW and ttbar)
- Can go in parallel provided re-interpretation is feasible
  - Needs close-to complete likelihood → a whole separate discussion
- ML versatile tools to optimally extract SM-EFT effects without too much tuning need → parametrized classifiers are an example

TREES & BOOSTING

[arXiv:2107.10859, arXiv:2205:12976]

weak learner

index-function (non-linearity)

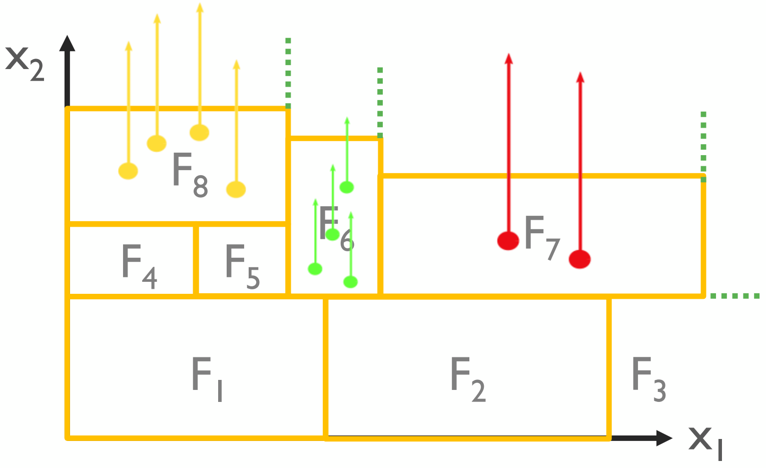
$$\hat{F}(\boldsymbol{x}) = \sum_{j\in\mathcal{J}} \mathbb{1}_j(\boldsymbol{x}) F_j$$

phase space partitioning J

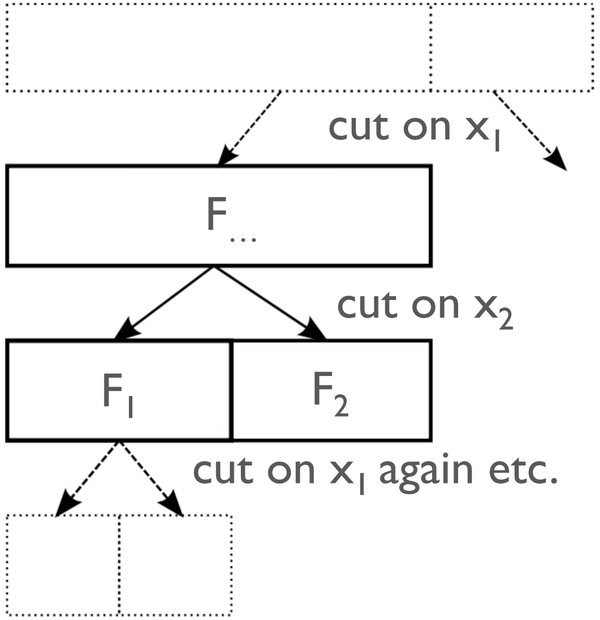
prediction F_j

need to solve for partitioning J and $\{F_j\}$

phase-space partitioning



training phase:
e.g. "CART" algo



- Let us make a tree-based ansatz for the differential cross-section ratio R
- The "weak learner" is a tree associating a sub-region (j) of a partitioning $\mathcal{J}$ with a predictive function F_j
- Fitting tree: Optimize "node split positions" on some loss. Trained (e.g. greedily) on the *ensemble*.
  - An axis-aligned tree is limited. Remove the limitation iteratively with "boosting".

LEARNING MORE WITH TREES

[arXiv:2107.10859, arXiv:2205:12976]

Regress in R, including its the polynomial $\boldsymbol{\theta}$ dependence

$$R(\boldsymbol{x}|\boldsymbol{\theta}, \mathrm{SM}) = \frac{\mathrm{d}\sigma(\boldsymbol{x}, \boldsymbol{\theta})/\mathrm{d}\boldsymbol{x}}{\mathrm{d}\sigma(\boldsymbol{x}, \mathrm{SM})/\mathrm{d}\boldsymbol{x}}$$

→ will allow to compute the optimal LLR test statistic $q(\mathcal{D})$

$$L = \sum_{\boldsymbol{\theta}\in\mathcal{B}} \int \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{z}\, p(\boldsymbol{x}, \boldsymbol{z}|\mathrm{SM}) \left(R(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}, \mathrm{SM}) - \hat{F}(\boldsymbol{x}, \boldsymbol{\theta})\right)^2$$

$$F^*(\boldsymbol{x}, \boldsymbol{\theta}) = R(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\theta}_0)$$

Tree ansatz for each a, ab:
$F_j(\boldsymbol{\theta})$ polynomial with const. coeff. (per node)

$$\hat{F}(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{j\in\mathcal{J}} \mathbb{1}_j(\boldsymbol{x}) F_j(\boldsymbol{\theta})$$

find optimal partitioning | find optimal predictor

Solve for the predictor on the empirical distribution (simulated sample)

$$F_j(\boldsymbol{\theta}) = \frac{\sum_{i\in j} w_i(\boldsymbol{\theta})}{\sum_{i\in j} w_i(\boldsymbol{\theta}_0)} \equiv \frac{w_j(\boldsymbol{\theta})}{w_j(\boldsymbol{\theta}_0)}$$

No trainable parameters in the predictor

Solve for optimal partitioning with greedy CART algorithm

$$L = -\sum_{\boldsymbol{\theta}\in\mathcal{B}} \sum_{j\in\mathcal{J}} \frac{w_j^2(\boldsymbol{\theta})}{w_j(\boldsymbol{\theta}_0)}$$

split only if
$w_j(\boldsymbol{\theta})$ is *positive* $\forall \boldsymbol{\theta}$

We'll find an optimized tree.
→ boost

CONCRETE SOLUTION: TREE BOOSTING

[arXiv:2107.10859, arXiv:2205:12976]

- Boosting: Fit linear model iteratively to pseudo-residuals of the preceding iteration with learning rate η
- Ansatz :

$$\hat{F}^{(b)}(\boldsymbol{x},\boldsymbol{\theta}) = \underbrace{\hat{f}(\boldsymbol{x},\boldsymbol{\theta})}_{\text{current iteration}} + \eta \underbrace{\hat{F}^{(b-1)}(\boldsymbol{x},\boldsymbol{\theta})}_{\text{previous iteration}}$$

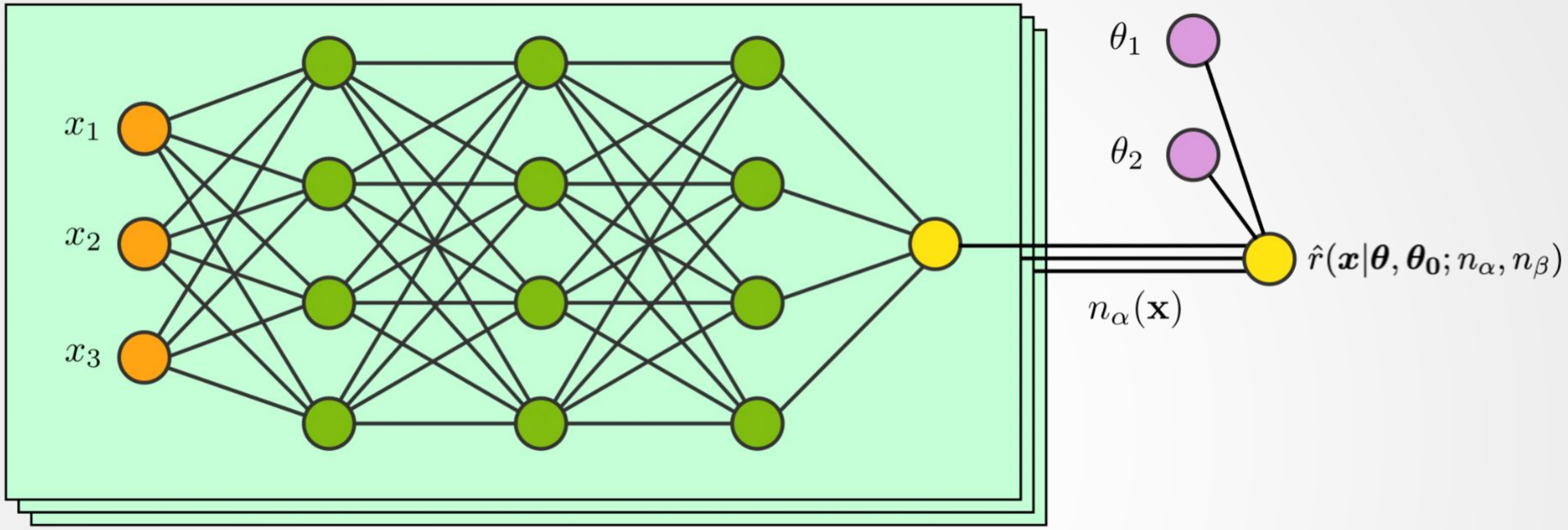
- Insert into the loss function:

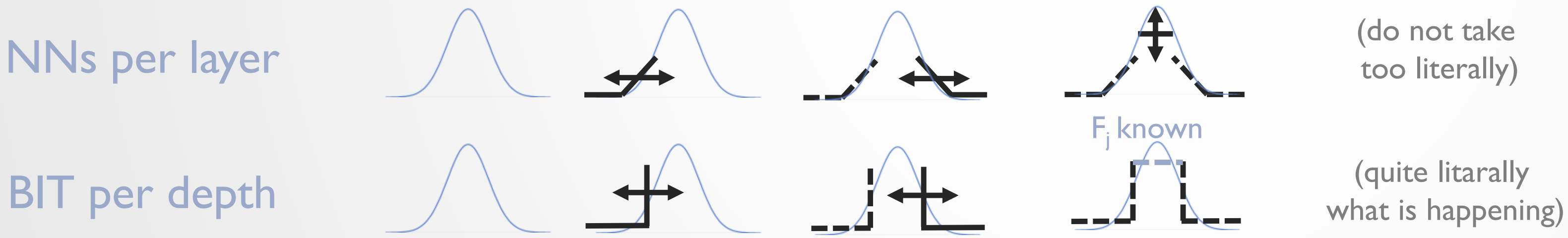
$$\underbrace{L[\hat{f}^{(b)}]}_{\text{current iteration}} = \sum_{\boldsymbol{\theta}\in\mathcal{B}} \int \mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{z}\, p(\boldsymbol{x},\boldsymbol{z}|\mathrm{SM}) \Big(\underbrace{R(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta},\mathrm{SM}) - \eta \underbrace{\hat{F}^{(b-1)}(\boldsymbol{x},\boldsymbol{\theta})}_{\text{previous iteration}}}_{\text{pseudo-residual, amounting to event-leve reweighting}} - \underbrace{\hat{f}^{(b)}(\boldsymbol{x},\boldsymbol{\theta})}_{\text{current iteration}} \Big)^2$$

$$w_i^{(b)}(\boldsymbol{\theta}) \quad \to \quad w_i^{(b-1)}(\boldsymbol{\theta}) - \eta\, w_i^{(b-1)}(\boldsymbol{\theta}_0) \hat{F}^{(b-1)}(\boldsymbol{x}_i,\boldsymbol{\theta})$$

.... perform this iteratively

NEURAL NETWORKS REGRESS; THE BIT DOESN'T





→ fewer DOF need regressing

- Each NN layer maps $L_{n+1} = \sigma(W_{ij}\, L_n + b_i)$. These DOF need to select & predict the regressed values.
- In the BIT, we only select. The prediction (F_j) is computed from the boxed events. This is possible, because a tree algorithm is (greedely) trained on the *ensemble*. The BITs' DOF are NOT updated event-by-event.

HOW TO PARAMETRIZE?

- Quantum field theory: Differential cross section predict polynomial SM-EFT dependence:

$$d\sigma(\boldsymbol{\theta}) \propto |\mathcal{M}_{\rm SM}(\boldsymbol{z}) + \boldsymbol{\theta}_a \mathcal{M}^a_{\rm BSM}(\boldsymbol{z})|^2 d\boldsymbol{z}$$

probability = wave function, squared

  - additivity of the matrix element → incur a simple (polynomial) dependence in $\boldsymbol{\theta}$ for fixed configuration z

$$\frac{d\sigma(\boldsymbol{x},\boldsymbol{\theta})}{d\boldsymbol{x}} = \frac{d\sigma_{\rm SM}(\boldsymbol{x})}{d\boldsymbol{x}} + \sum_a \theta_a \frac{d\sigma^a_{\rm int.}(\boldsymbol{x})}{d\boldsymbol{x}} + \frac{1}{2}\sum_{a,b}\theta_a\theta_b \frac{d\sigma^{ab}_{\rm BSM}(\boldsymbol{x})}{d\boldsymbol{x}}$$

- Neyman-Pearson: $q(\mathcal{D}) = \frac{L(\mathcal{D}|\boldsymbol{\theta}\ \ )}{L(\mathcal{D}|\mathrm{SM})}$ where $L(\mathcal{D}|\boldsymbol{\theta}) = \mathrm{P}_{\mathcal{L}\sigma(\boldsymbol{\theta})}(N) \times \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\theta})$

"normalization" "shape"

$$q_{\boldsymbol{\theta}}(\mathcal{D}) = \underbrace{\mathcal{L}(\sigma_{\boldsymbol{\theta}} - \sigma_{\rm SM})}_{\rm const.} - \sum_{\boldsymbol{x}_i\in\mathcal{D}} \log R(\boldsymbol{x}_i|\boldsymbol{\theta},\mathrm{SM})$$

Optimality can be achieved with cross-section ratio R or its **universal** coefficient functions R_a, R_{ab}

$$R(\boldsymbol{x}|\boldsymbol{\theta},\mathrm{SM}) = \frac{d\sigma(\boldsymbol{x},\boldsymbol{\theta})/d\boldsymbol{x}}{d\sigma(\boldsymbol{x},\mathrm{SM})/d\boldsymbol{x}} = 1 + \sum_a \theta_a R_a(\boldsymbol{x}) + \frac{1}{2}\sum_{a,b}\theta_a\theta_b R_{ab}(\boldsymbol{x})$$

$$\widehat{=} \left(1 + \sum_a \theta_a \hat{n}_a(\boldsymbol{x})\right)^2 + \sum_a \left(\sum_{b\geq a} \theta_b \hat{n}_{ab}(\boldsymbol{x})\right)^2$$

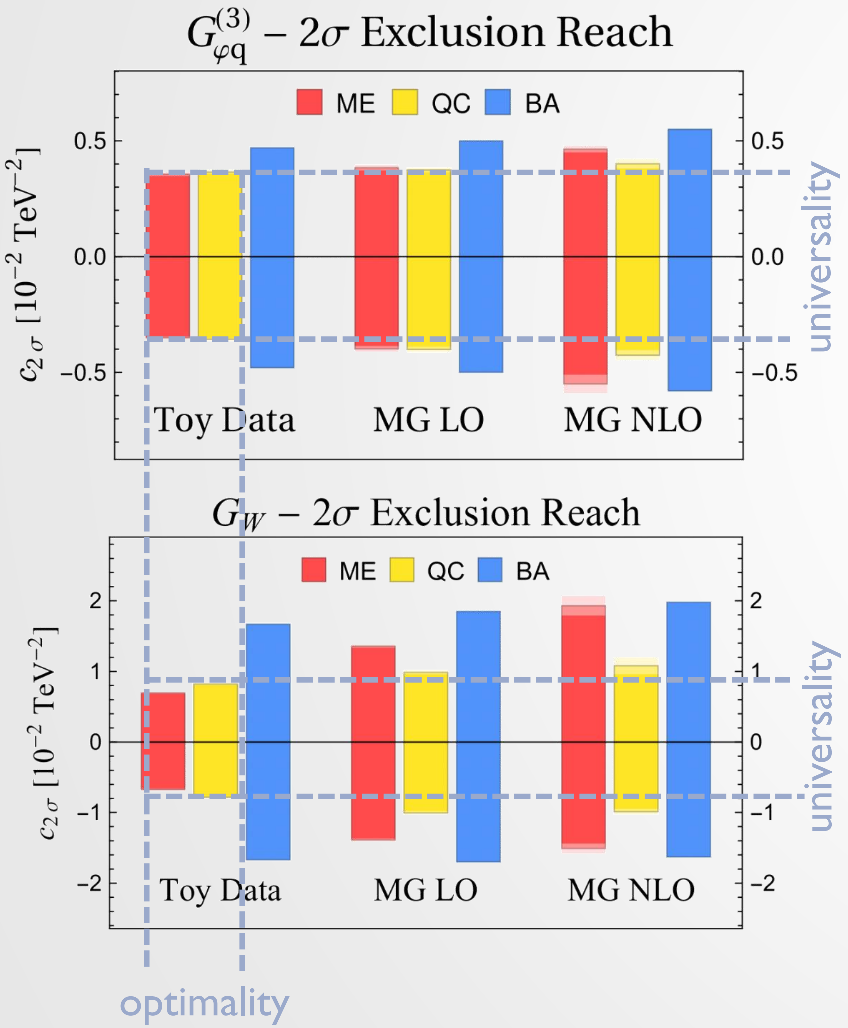
NB #1 Curse of dimensionality is lifted!!
15 operators → 136 coefficients

NB #2: R is positive: Fit universal dependence using the most general quadratic polynomial

OPTIMAL PARAMETRIZED CLASSIFIERS

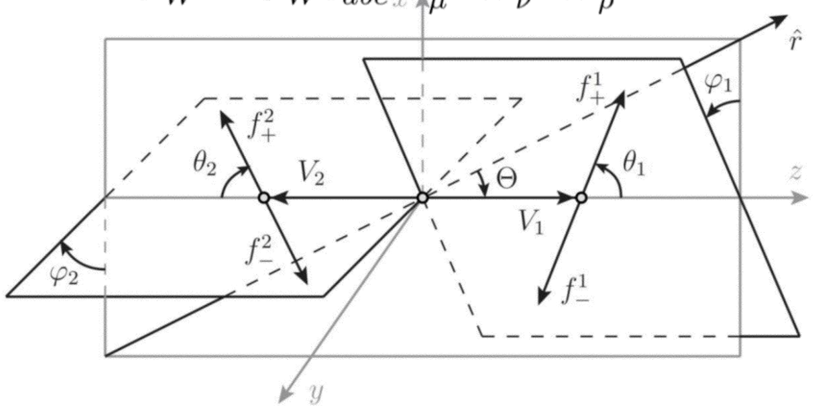
JHEP 05 (2021) 247
ATLAS-CONF-NOTE-2016-043,
JHEP 07 (2022) 032

- studied in the context of $p\,p \to W^{\pm}\,Z \to (l^{\pm}\,\nu)\,(l^{+}\,l^{-})$ for the most important SM-EFT operators

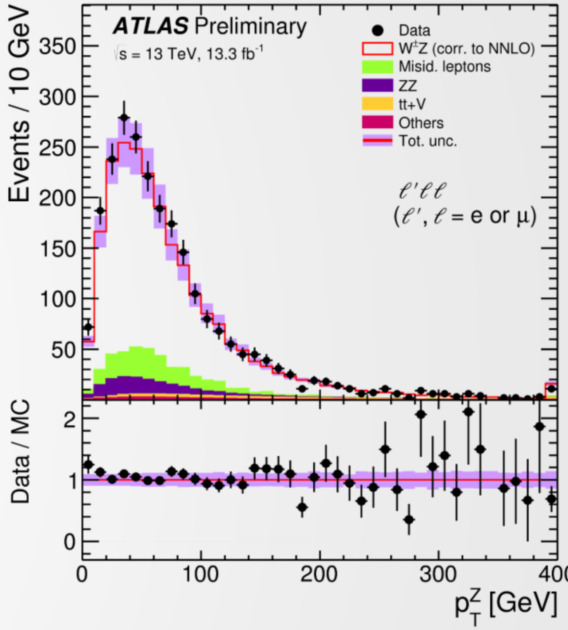


$$\mathcal{O}_{\varphi q}^{(3)} = G_{\varphi q}^{(3)}\,(\overline{Q}_L \sigma^a \gamma^\mu Q_L)\,(i H^\dagger \overleftrightarrow{D}_\mu^a H)$$

$$\mathcal{O}_W = G_W \varepsilon_{abc} W_\mu^{a\,\nu} W_\nu^{b\,\rho} W_\rho^{c\,\mu}$$



$$x = \{s,\,\Theta,\,\theta_W,\,\theta_Z,\,p_{T,\mathrm{ZW}},\,p_{T,\mathrm{Z}},\,\sin\varphi_W,\,\cos\varphi_W,\,\sin\varphi_Z,\,\cos\varphi_Z\}$$

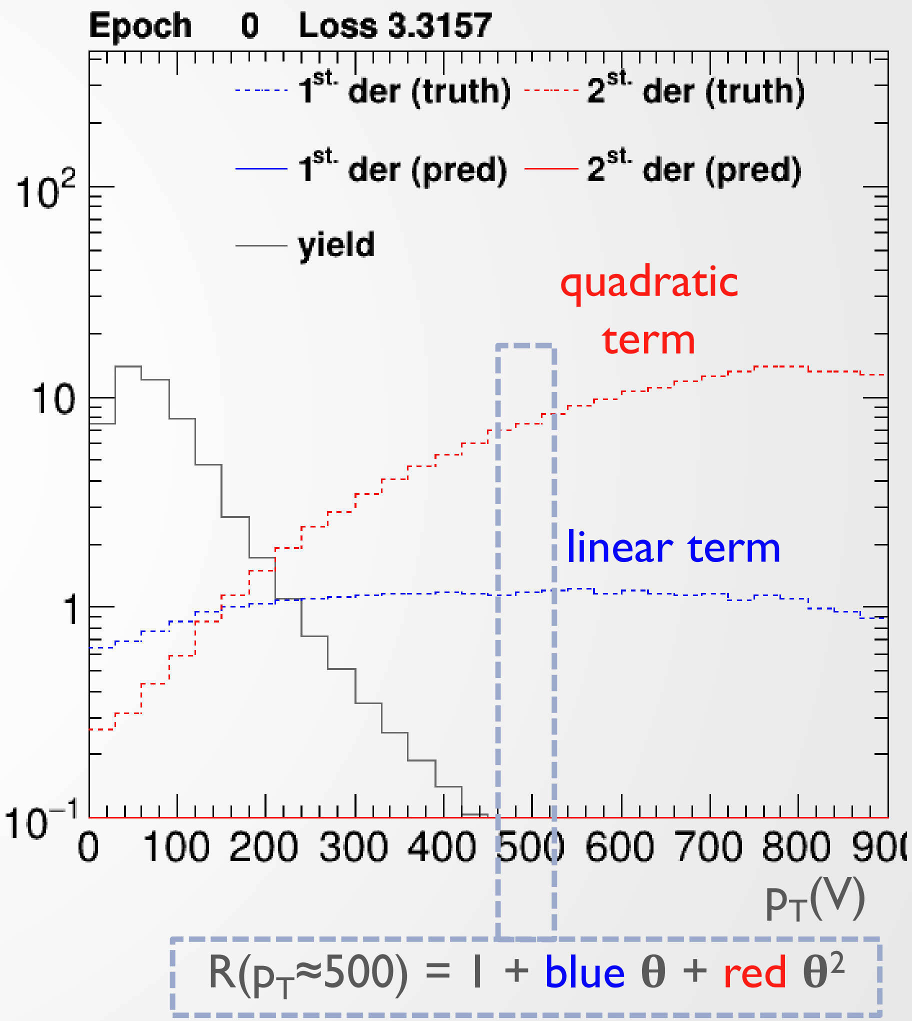


- high purity, ~85%-90% as seen by ATLAS and CMS (with SM-EFT)
- Adam optimizer, pytorch, 10^4 epochs, learning rate of 10^{-4}
  - 4 hidden layers á 32 nodes, 2 networks simultaneously trained
  - alternatives configurations studied
- establish optimality with analytic model (Toy), very similar at (N)LO

PYTORCH IMPLEMENTATION

S. Chen, A. Glioti, G. Panico, A. Wulzer
JHEP 05 (2021) 247

- ZH production, analytic model, 500k events
  - Single coefficient: c_{HW}
  - 4 hidden layers á 32 nodes, 2 networks simultaneously trained
  - 10^4 epochs, Adam optimizer, LR=10^{-4}
- The training is *simultaneous* and it must be!
  - Positivity is a property of the polynomial, not of an individual coefficient.
- several options to emphasise the tails
  - bias loss with function of A(x) or choosing base points
- just a proof of principle implementation



"PARTICLE PHYSICS STRUCTURE"

K. Cranmer , J. Pavez , and G. Louppe [1506.02169]
J. Brehmer, K. Cranmer, G. Louppe, J. Pavez [1805.00013] [1805.00020] [1805.12244]
J. Brehmer, F. Kling, I. Espejo, K. Cranmer [1907.10621]

- It's somewhat of a miracle that one can regress on the observable-level likelihood ratio

Observables	Integration over intractable factors	Detector & reconstruction	Parton shower	Parton-level momenta	Theory parameters

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \int \mathrm{d}\boldsymbol{z}_\mathrm{d}\, \mathrm{d}\boldsymbol{z}_\mathrm{s}\, \mathrm{d}\boldsymbol{z}_\mathrm{p} \quad p(\boldsymbol{x}|\boldsymbol{z}_\mathrm{d}) \quad p(\boldsymbol{z}_\mathrm{d}|\boldsymbol{z}_\mathrm{s}) \quad p(\boldsymbol{z}_\mathrm{s}|\boldsymbol{z}_\mathrm{p}) \quad p(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta})$$

$$q_{\boldsymbol{\theta}}(\mathcal{D}) = \text{const.} - \sum_{\boldsymbol{x}_i \in \mathcal{D}} \log \frac{\sigma(\boldsymbol{\theta})}{\sigma(\boldsymbol{\theta}_0)} \frac{p(\boldsymbol{x}_i|\boldsymbol{\theta}\ )}{p(\boldsymbol{x}_i|\boldsymbol{\theta}_0)}$$

superpowers

$$\mathrm{d}\sigma(\boldsymbol{\theta}) \propto$$

$$|\mathcal{M}_{\mathrm{SM}}(\boldsymbol{z}) + \boldsymbol{\theta}_a \mathcal{M}^a_{\mathrm{BSM}}(\boldsymbol{z})|^2 \mathrm{d}\boldsymbol{z}$$

calcuable & re-calcuable
(aka tractable) theory prediction

based on this talk: C. Kranmer, J. Brehmer

"JOINT" DISTRIBUTIONS ARE MUCH SIMPLER

- To understand the power of simulation, look at the simpler "joint" pdf

1. The intractable factors cancel in the joint LR

$$r(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{\theta}_0) \equiv \frac{p(\boldsymbol{x},\boldsymbol{z}_\mathrm{d},\boldsymbol{z}_\mathrm{s},\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta})}{p(\boldsymbol{x},\boldsymbol{z}_\mathrm{d},\boldsymbol{z}_\mathrm{s},\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta}_0)} = \frac{p(\boldsymbol{x}|\boldsymbol{z}_\mathrm{d})}{p(\boldsymbol{x}|\boldsymbol{z}_\mathrm{d})}\frac{p(\boldsymbol{z}_\mathrm{d}|\boldsymbol{z}_\mathrm{s})}{p(\boldsymbol{z}_\mathrm{d}|\boldsymbol{z}_\mathrm{s})}\frac{p(\boldsymbol{z}_\mathrm{s}|\boldsymbol{z}_\mathrm{p})}{p(\boldsymbol{z}_\mathrm{s}|\boldsymbol{z}_\mathrm{p})}\frac{p(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta}\ )}{p(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta}_0)} \propto \frac{|\mathcal{M}(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta}\ )|^2}{|\mathcal{M}(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta}_0)|^2}$$

Change in likelihood of observation x (with history z) going from θ_0 to θ

staged simulation: Intractable factors cancel

re-calcuable theory prediction

2. Now fit a general function on the join space with a regressor depending only on the observables:

$$L = \int \mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{z}\,p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta}_0)\left(f(\boldsymbol{x},\boldsymbol{z}) - \hat{f}(\boldsymbol{x})\right)^2 \longrightarrow \min \qquad f^*(\boldsymbol{x}) = \frac{\int \mathrm{d}\boldsymbol{z}\,p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta}_0)f(\boldsymbol{x},\boldsymbol{z})}{\int \mathrm{d}\boldsymbol{z}\,p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta}_0)}$$

Latent space is integrated

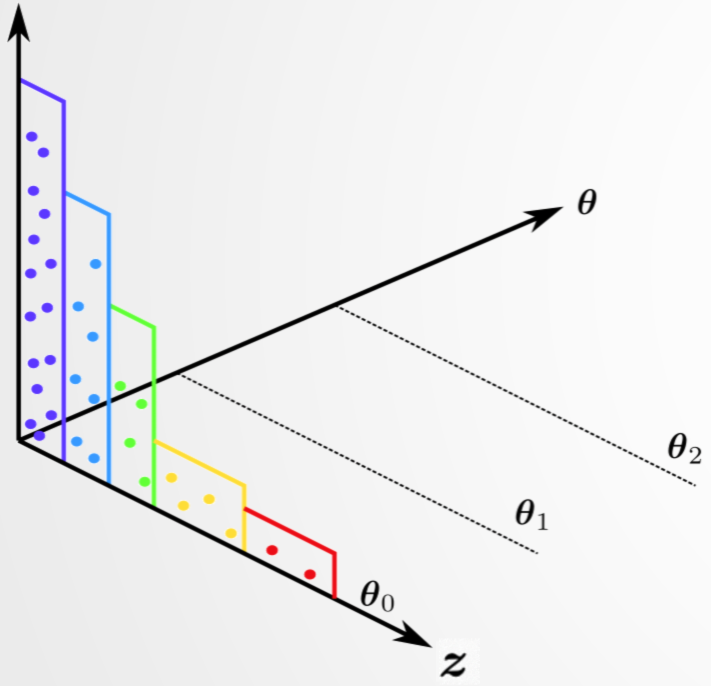
3. Now chose f(x,z) = r(x,z | $\boldsymbol{\theta}$, $\boldsymbol{\theta}_0$) which is available in simulation & fit with expressive function:

$$f^*(\boldsymbol{x}) = \frac{\int \mathrm{d}\boldsymbol{z}\,p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta}_0)f(\boldsymbol{x},\boldsymbol{z})}{\int \mathrm{d}\boldsymbol{z}\,p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta}_0)} = \frac{\int \mathrm{d}\boldsymbol{z}\,p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta}_0)r(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{\theta}_0)}{\int \mathrm{d}\boldsymbol{z}\,p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta}_0)} = \frac{\int \mathrm{d}\boldsymbol{z}\,p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta}_0)\frac{p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta})}{p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta}_0)}}{\int \mathrm{d}\boldsymbol{z}\,p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta}_0)} = \frac{\int \mathrm{d}\boldsymbol{z}\,p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta})}{\int \mathrm{d}\boldsymbol{z}\,p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta}_0)} = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x}|\boldsymbol{\theta}_0)}$$

Available from simulation ⟹ what we actually want: change in likelihood of a specific observation

... statistical framework of all the parametrized classifiers

EXPLOITING PARAMETRIZED *SIMULATION* WITH TREES



- Quantum field theory: Differential cross section have structure

$$\mathrm{d}\sigma(\boldsymbol{\theta}) \propto |\mathcal{M}_{\mathrm{SM}}(\boldsymbol{z}) + \boldsymbol{\theta}_a \mathcal{M}^a_{\mathrm{BSM}}(\boldsymbol{z})|^2 \mathrm{d}\boldsymbol{z}$$

probability = wave function, squared

  - sampling **z** at a fixed $\boldsymbol{\theta}_o$
  - evaluate dσ($\boldsymbol{\theta}$) for sufficient number of base-points $\boldsymbol{\theta}$
  - fix polynomial coefficients of event weights $w_i(\boldsymbol{\theta})$

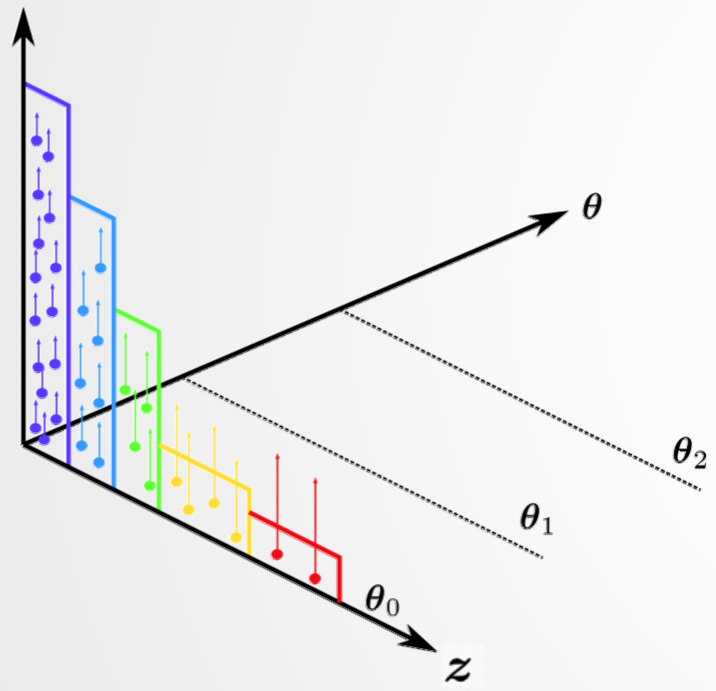
$$w_i(\boldsymbol{\theta}) = w_{i,0} + \sum_a w_{i,a}\,\theta_a + \frac{1}{2}\sum_{a,b} w_{ab}\,\theta_a\theta_b = \frac{\sigma(\boldsymbol{\theta}\ )}{\sigma(\boldsymbol{\theta}_0)} \cdot r(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta}, \boldsymbol{\theta}_0)$$

SM — interference — pure SM-EFT

interpretation valid at LO

  - obtain predictions for, e.g., yields for all x,z and $\boldsymbol{\theta}$

EXPLOITING PARAMETRIZED *SIMULATION* WITH TREES



- Quantum field theory: Differential cross section have structure

$$d\sigma(\boldsymbol{\theta}) \propto |\mathcal{M}_{\mathrm{SM}}(\boldsymbol{z}) + \boldsymbol{\theta}_a \mathcal{M}^a_{\mathrm{BSM}}(\boldsymbol{z})|^2 d\boldsymbol{z}$$

probability = wave function, squared

  - sampling **z** at a fixed $\boldsymbol{\theta}_0$
  - evaluate $d\sigma(\boldsymbol{\theta})$ for sufficient number of base-points $\boldsymbol{\theta}$
  - fix polynomial coefficients of event weights $w_i(\boldsymbol{\theta})$

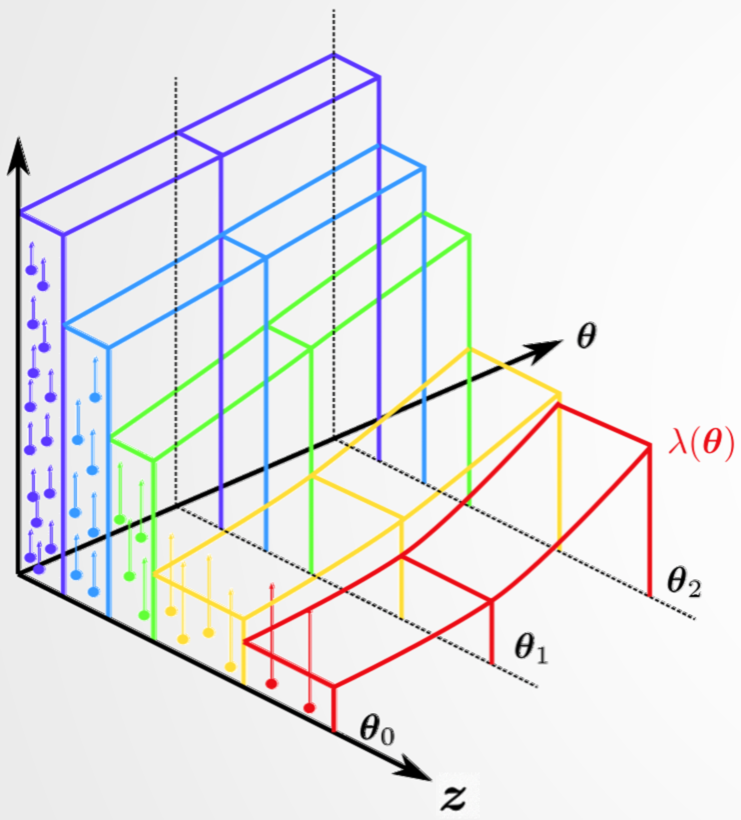
$$w_i(\boldsymbol{\theta}) = w_{i,0} + \sum_a w_{i,a}\,\theta_a + \frac{1}{2}\sum_{a,b} w_{ab}\,\theta_a\theta_b = \frac{\sigma(\boldsymbol{\theta}\ )}{\sigma(\boldsymbol{\theta}_0)} \cdot r(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta}, \boldsymbol{\theta}_0)$$

SM | interference | pure SM-EFT

interpretation valid at LO

  - obtain predictions for, e.g., yields for all x,z and $\boldsymbol{\theta}$

EXPLOITING PARAMETRIZED *SIMULATION* WITH TREES



- Quantum field theory: Differential cross section have structure

$$\mathrm{d}\sigma(\boldsymbol{\theta}) \propto |\mathcal{M}_{\mathrm{SM}}(\boldsymbol{z}) + \boldsymbol{\theta}_a \mathcal{M}^a_{\mathrm{BSM}}(\boldsymbol{z})|^2 \mathrm{d}\boldsymbol{z}$$

probability = wave function, squared

  - sampling **z** at a fixed $\boldsymbol{\theta}_0$
  - evaluate d$\sigma(\boldsymbol{\theta})$ for sufficient number of base-points $\boldsymbol{\theta}$
  - fix polynomial coefficients of event weights $w_i(\boldsymbol{\theta})$

$$w_i(\boldsymbol{\theta}) = w_{i,0} + \sum_a w_{i,a}\,\theta_a + \frac{1}{2}\sum_{a,b} w_{ab}\,\theta_a\theta_b = \frac{\sigma(\boldsymbol{\theta}\ )}{\sigma(\boldsymbol{\theta}_0)} \cdot r(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta}, \boldsymbol{\theta}_0)$$

SM | interference | pure SM-EFT | *interpretation valid at LO*

  - obtain predictions for, e.g., yields for all x,z and $\boldsymbol{\theta}$

TREES & BOOSTING

[arXiv:2107.10859, arXiv:2205:12976]

weak learner

index-function (non-linearity)

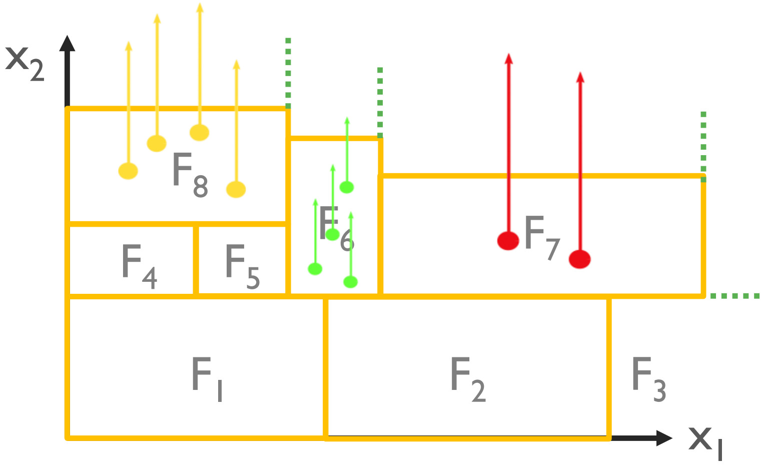
$$\hat{F}(\boldsymbol{x}) = \sum_{j \in \mathcal{J}} \mathbb{1}_j(\boldsymbol{x}) F_j$$

phase space partitioning J

prediction F_j

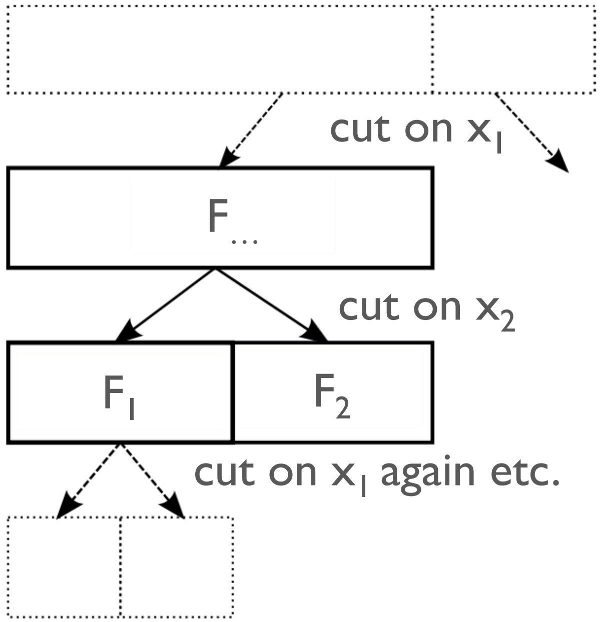
need to solve for partitioning J and $\{F_j\}$

phase-space partitioning



training phase:

e.g. "CART" algo



- Let us make a tree-based prediction for R or its coefficient function
- Weak learner: Tree ↔ Associates a predictive function F_j (flexible!) with a sub-region j of a partitioning
- Fitting tree: Optimize "node split positions" on some loss. Trained (e.g. greedily) on the *ensemble*.

CONCRETE SOLUTION: TREE BOOSTING

[arXiv:2107.10859, arXiv:2205:12976]

- Boosting: Fit linear model iteratively to pseudo-residuals of the preceding iteration
- Ansatz :

$$\hat{F}^{(b)}(\boldsymbol{x}) = \underbrace{\hat{f}^{(b)}(\boldsymbol{x})}_{\text{current iteration}} + \underbrace{\eta \hat{F}^{(b-1)}(\boldsymbol{x})}_{\text{previous iteration}}$$

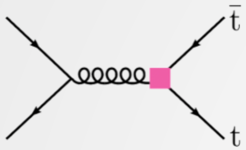
- Insert into the loss function:

$$\underbrace{\mathrm{MSE}[\hat{f}_a^{(b)}]}_{\text{current iteration}} = \sum_{(\boldsymbol{x},\boldsymbol{z},w)_i\in\mathcal{D}} w_{i,0} \left| \underbrace{\frac{w_{i,a}}{w_{i,0}} - \underbrace{\eta \hat{F}_a^{(b-1)}(\boldsymbol{x}_i)}_{\text{previous iteration}}}_{\text{pseudo-residual}} - \underbrace{\hat{f}_a^{(b)}(\boldsymbol{x}_i)}_{\text{current iteration}} \right|^2 = \underbrace{\sum_{(\boldsymbol{x},\boldsymbol{z},w)_i\in\mathcal{D}} w_{i,0} \left| \underbrace{\frac{w_{i,a} - \eta w_{i,0} \hat{F}_a^{(b-1)}(\boldsymbol{x}_i)}{w_{i,0}}}_{\text{reweighting}} - \hat{f}_a^{(b)}(\boldsymbol{x}_i) \right|^2}_{\text{MSE structure at iteration b}}$$

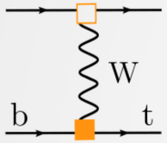
.... perform this iteratively

TOP AND DIBOSON SECTORS

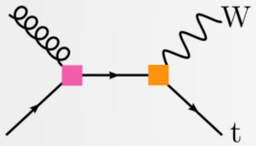
tt
811 pb



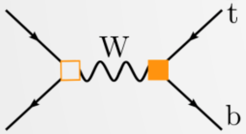
t (t-channel)
217 pb



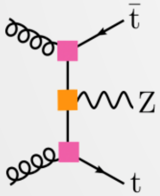
tW
72 pb



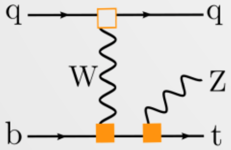
t (s-channel)
10 pb



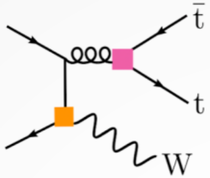
ttZ
1 pb



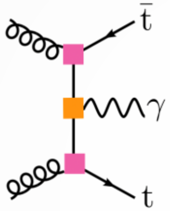
tZq
0.088 pb



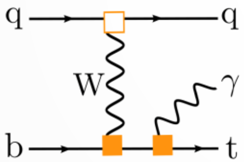
ttW
0.55 pb



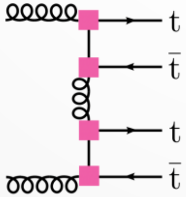
ttγ
0.80 pb



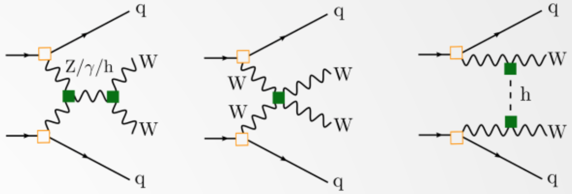
tγq
0.4 pb



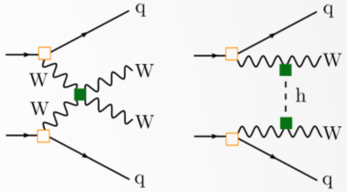
tttt
0.009 pb



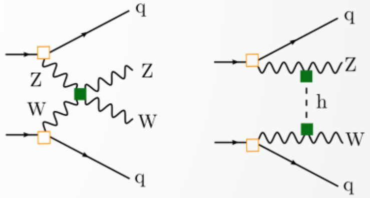
W±W∓
EWK 10 fb



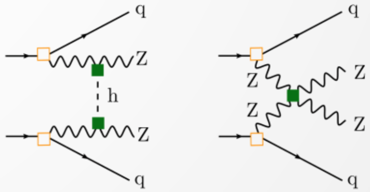
W±W±
EWK 4 fb



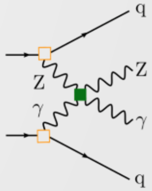
WZ
EWK 2.15 fb



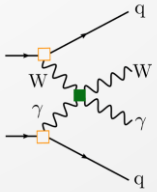
ZZ
EWK 0.82 fb



Zγ
EWK 7.8 fb



Wγ
EWK 20.4 fb



Z
EWK 37.4 fb