MEM from the ttH(bb) trenches

Joosep Pata (ETH Zurich)

Higgs Toppings

Benasque 2018

Problem statement

- In a nutshell: given observed data **y** ∈ {jets, leptons, ...}, exclude alternate hypotheses

- Search: signal (tt+H, $\mathbf{H_1}$) and bkg (tt+jets, $\mathbf{H_0}$)

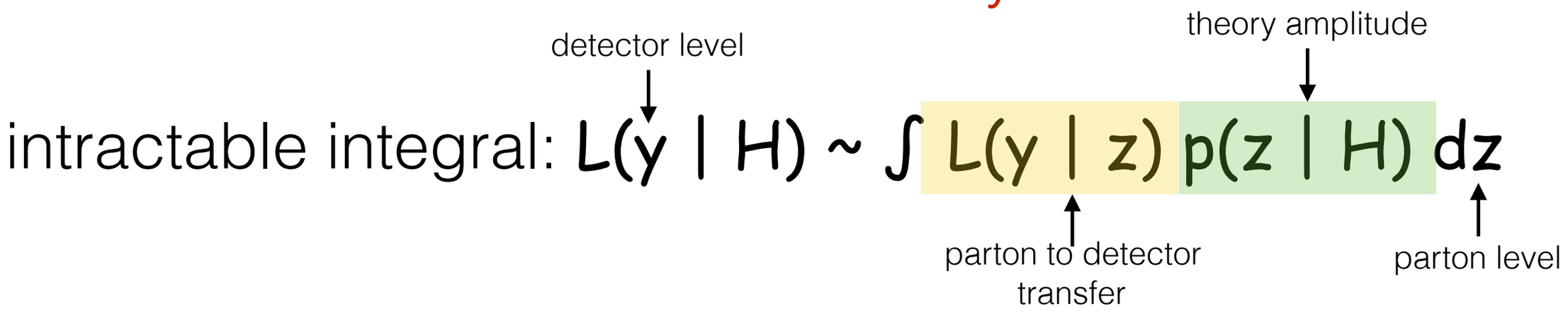
Neyman-Pearson lemma: given observable event **y** and two hypotheses, the ratio

$$\lambda(\boldsymbol{y}) = \frac{L(\boldsymbol{y}|\mathrm{H}_1)}{L(\boldsymbol{y}|\mathrm{H}_0)}$$

is the **most powerful test statistic**, with $L(y \mid H)$ being the likelihood of data given observation

Methods

We cannot directly compute the likelihood **L(y | H)** given an observed event y!



We have sets of MC simulation **events with different H**:

$$\{y_{1,0}\} \sim L(y \mid H_{1,0})$$

(I) Choose **clever observable d=f(y)**, estimate **L(d(y) | H)** from simulation events: templates

(II) Compute **approximate L(y | H)** directly given theory ideas

Clever observables

- Most analyses: **d(y)** is a parametric function (BDT, DNN), **numerically optimized** to discriminate signal from background based on MC samples

- Amount of MC statistics needed? Half a billion full-MC per experimental conditions is routine...
- Which features exploited? Study shaping of inv. masses, interpretability...
- What if MC does not accurately represent data? Cover with systematic uncertainties, but any bias?

"Modelling studies can be seen which tend to overestimate certainty, pretending to produce crisp numbers precise to the third decimal digits even in situation of pervasive uncertainty or ignorance."

jet-to-parton associations

parton level momenta

observed event

$$P(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_{k=1}^{N_a} \int \frac{\mathrm{d}x_1 \mathrm{d}x_2}{2x_1 x_2 s} \int \prod_{i=1}^{n} \frac{\mathrm{d}^3 p_i}{(2\pi)^3 2E_i}$$

hypothesis

$$\times \delta^4(q_1 + q_2 - \sum_{i=1}^{n} p_i)$$

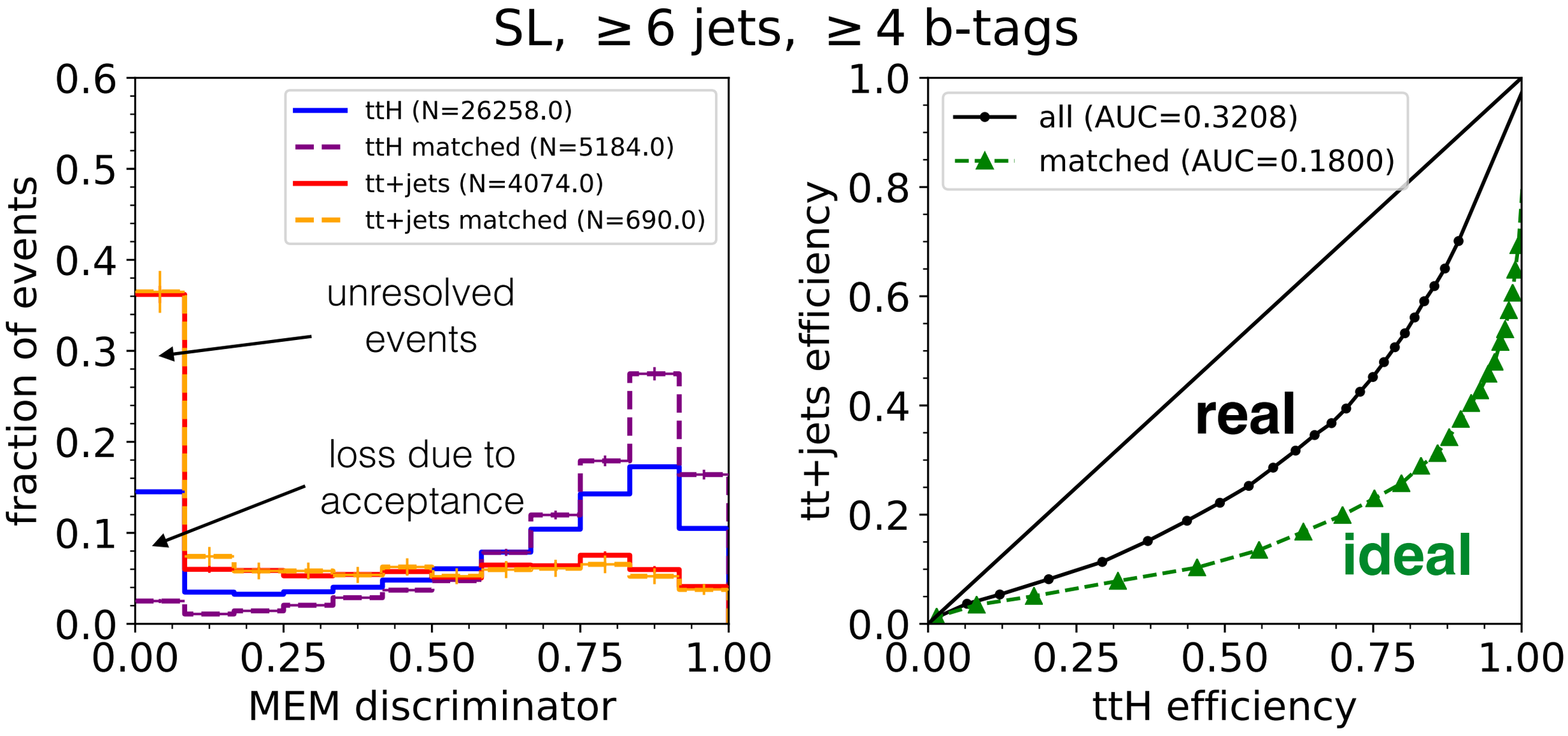
PDFs	$\times\, g(x_1) g(x_2)$
recoil	$\times\, \mathcal{R}(\tilde{\boldsymbol{\rho}}_T, \boldsymbol{\rho}_T)$
scattering amplitude	$\times\, \lvert\mathcal{M}_{\boldsymbol{\theta}}(q_1, q_2, p_1, \ldots, p_n)\rvert^2$
transfer function	$\times\, W(\boldsymbol{y}, \boldsymbol{p}).$

Integrate numerically on the LHC computing grid.

Transfer functions

- Transition from parton-level event **p** (a few 4-momenta) to detector-level event **y** (~50-100 high-level quantities): showering, hadronization, detector effects, acceptance
- Assumptions, e.g. **Gaussian** quark-to-jet smearing
- Determined from **MC simulation by max likelihood fitting**
- **Need to hand-code** event-to-event transfer function, combination of possible jet-to-quark assignments
- In practice, integration becomes very expensive

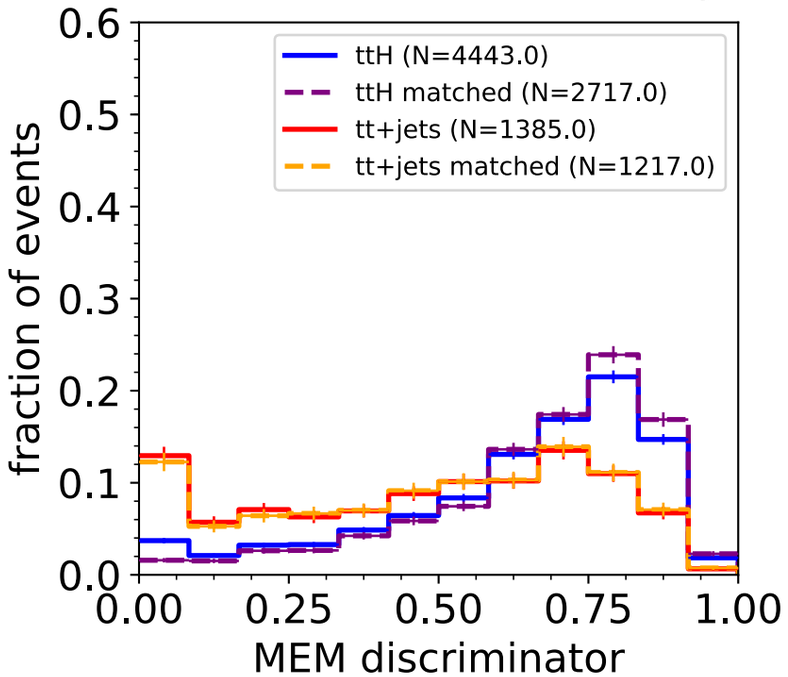
Reconstruction effects

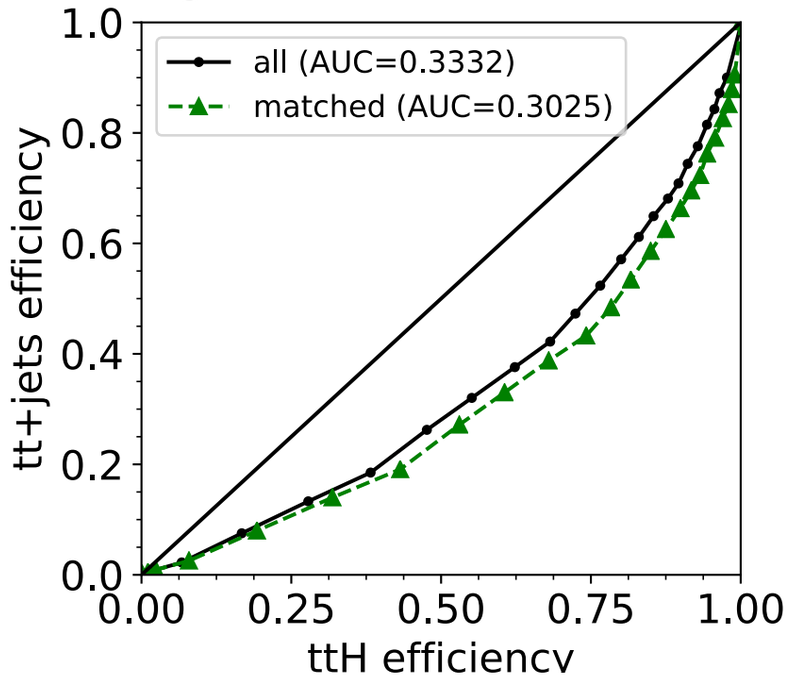


In ttH(bb), MEM performance is far from ideal due to detector effects & missed quarks!

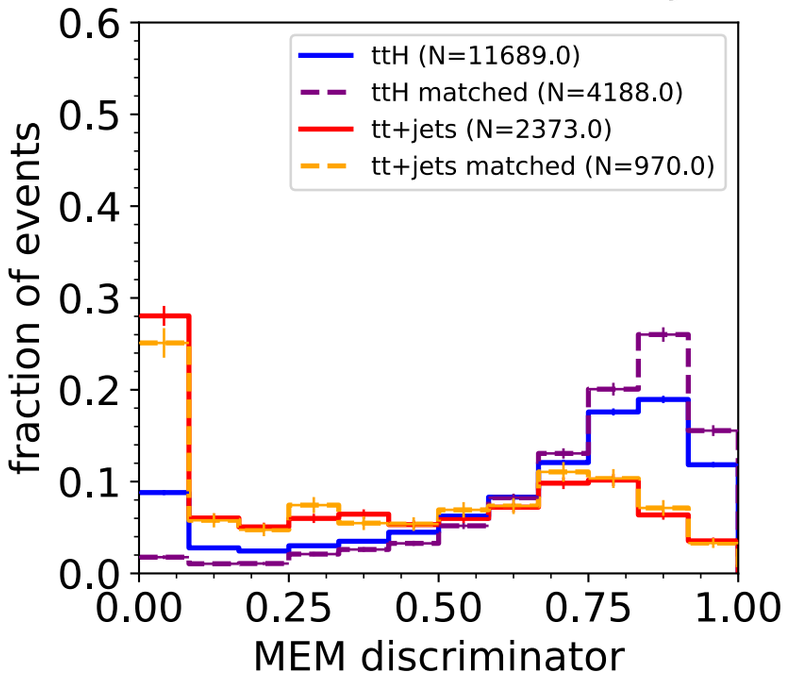
4j, 5j events

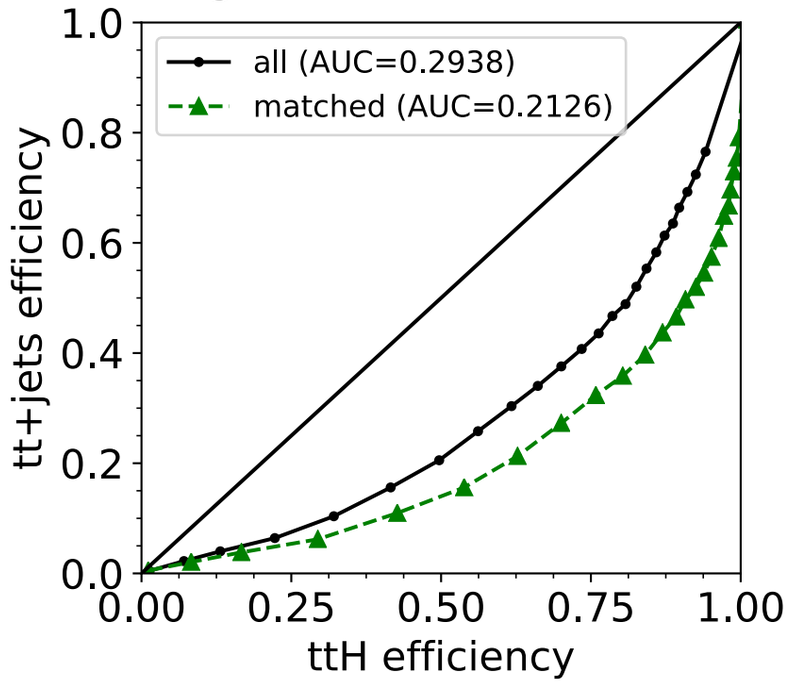
SL, 4 jets, 4 b-tags





SL, 5 jets, ≥ 4 b-tags





Questions for MEM

- **Which hypothesis?** ttH(bb) SL ≥6j, ≥4b fully matched ~30% of times, 30% miss quark from W

  - need to "degrade gracefully", but not implement MC by hand

- **Which assumptions?** Need to reduce integration space by assuming e.g. b-tagging, top tagging

- **More complex analyses?** Multi-hypothesis, multi-parameter EFT fits

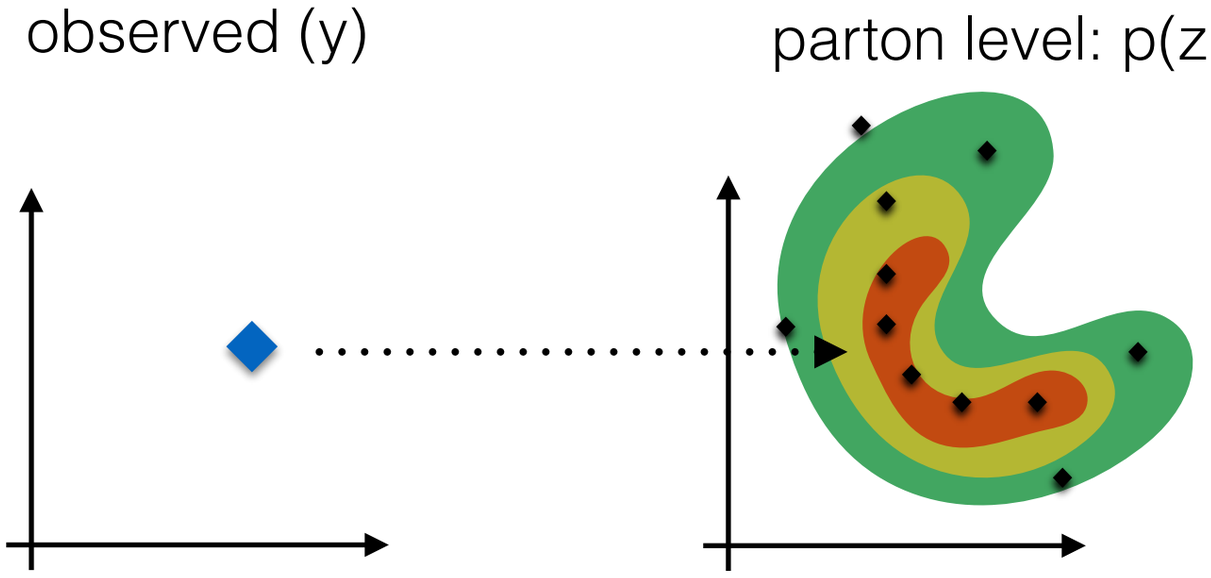
- **NLO:** given computational cost, how to make use in experiment besides samples?

MEM + ML combination

$$L(y \mid H) \sim \int L(y \mid z)\, p(z \mid H)\, dz$$

Use ML to approximate hadronization, showering, acceptance in $L(y \mid z)$.
Bonus: differentiable wrt. exp systematics, no reimplementation of MC.

If possible to generate samples cheaply, can integrate scattering ME efficiently.



$$\int p(z \mid y)\, p(z \mid H)\, dz$$

Or, replace per-event integral with regression over full MC:
[Brehmer, Cranmer, Louppe, Pavez 2018]

Bonus slides

MEM hypotheses

interpretation	bottom quarks	light quarks	description
SL $2_W 2_h 2_t$	4	2	fully-reconstructed semileptonic
SL $1_W 2_h 2_t$	4	1	$(l\nu'_l b)_t (\not{q} q' \bar{b})_{\bar{t}} (b\bar{b})_h$
SL $0_W 2_h 2_t$	4	0	$(l\nu'_l b)_t (\not{q}\not{q}' \bar{b})_{\bar{t}} (b\bar{b})_h$
SL $2_W 2_h 2_t + 1g$	4	3	fully-reconstructed, additional ISR gluon
DL $2_h 2_t$	4	0	fully-reconstructed dileptonic

Table 4.1: The detailed event interpretations for semileptonic and dileptonic $t\bar{t}H$ (signal) and $t\bar{t}+b\bar{b}$ (background) events. In the semileptonic channel, we consider cases where up to 2 light quarks may be lost. The minimum number of jets required for a hypothesis is the sum of the number of quarks. For the SL $1_W 2_h 2_t$ hypothesis the direction and energy of one of the light quarks is integrated out, denoted by $\not{q}$. For the fully-reconstructed semileptonic case with 7 jets, we also consider the ISR-modified interpretation.

MEM permutations

interpretation	8+ jets	7 jets	6 jets	5 jets	4 jets
SL $2_W 2_h 2_t$	72/36	36	12	-	-
SL $1_W 2_h 2_t$	-	-	-	12	-
SL $0_W 2_h 2_t$	-	-	-	-	12
SL $2_W 2_h 2_t + 1g$	72/36	36	-	-	-
DL $2_h 2_t$	12	12	12	12	12

Table 4.2: The number of MEM combinations in associating quarks to jet for various MEM categories. In SL events with 7 or more jets, we choose up to 4 candidates based on invariant mass among the light jets for the W-boson reconstruction, in order to prevent a combinatorial explosion for events with a very high jet multiplicity. In DL events, we always choose exactly 4 candidates for the b quarks.

CPU cost

method	time $t\bar{t}H$ (s)	time $t\bar{t} + b\bar{b}$ (s)	ROC AUC	ϵ_{bkg}	total (h) / 1k
SL, $\geq$ 7jet, $2_W 2_h 2_t$	45.8 ± 18.9	69.4 ± 26.1	0.315	0.232	32.00
SL, $\geq$ 7jet, $2_W 2_h 2_t 1_g$	71.7 ± 27.1	471.7 ± 50.6	0.317	0.233	150.94
SL, $\geq$ 6jet, $2_W 2_h 2_t$	30.2 ± 21.0	45.4 ± 30.9	0.321	0.233	21.00
SL, $\geq$ 6jet, $1_W 2_h 2_t$	64.8 ± 22.7	101.1 ± 33.0	0.307	0.210	46.07
SL, $\geq$ 6jet, $0_W 2_h 2_t$	83.9 ± 20.4	136.3 ± 28.9	0.294	0.218	61.16
SL, 5jet, $1_W 2_h 2_t$	25.4 ± 7.1	39.9 ± 9.6	0.293	0.198	18.13
SL, 5jet, $0_W 2_h 2_t$	84.7 ± 20.3	136.9 ± 28.9	0.291	0.217	61.56
SL, 4jet, $0_W 2_h 2_t$	84.3 ± 20.7	136.0 ± 29.2	0.333	0.275	61.21
DL, $\geq$ 4jet, $0_W 2_h 2_t$	55.7 ± 13.7	90.4 ± 19.3	0.223	0.124	40.58

Table 4.3: The CPU budget and separation power of the MEM in the SL channel using various event interpretations. We show the time required to evaluate the signal and background hypotheses, the receiver operating characteristic (ROC) area under curve (AUC), the efficiency of background at a signal efficiency of 50% (ϵ_{bkg}) and the total time required to compute the MEM discriminator for 1000 events.