Reducing the number of negative weights in fixed-order $pp \to t\bar{t} + X$ samples

M.V. Garzelli [1], S.-O. Moch[1], O. Zenaiev[1],

with input from J. Andersen, A. Maier + MATRIX experts....

[1] II Institut für Theoretische Physik, Universität Hamburg

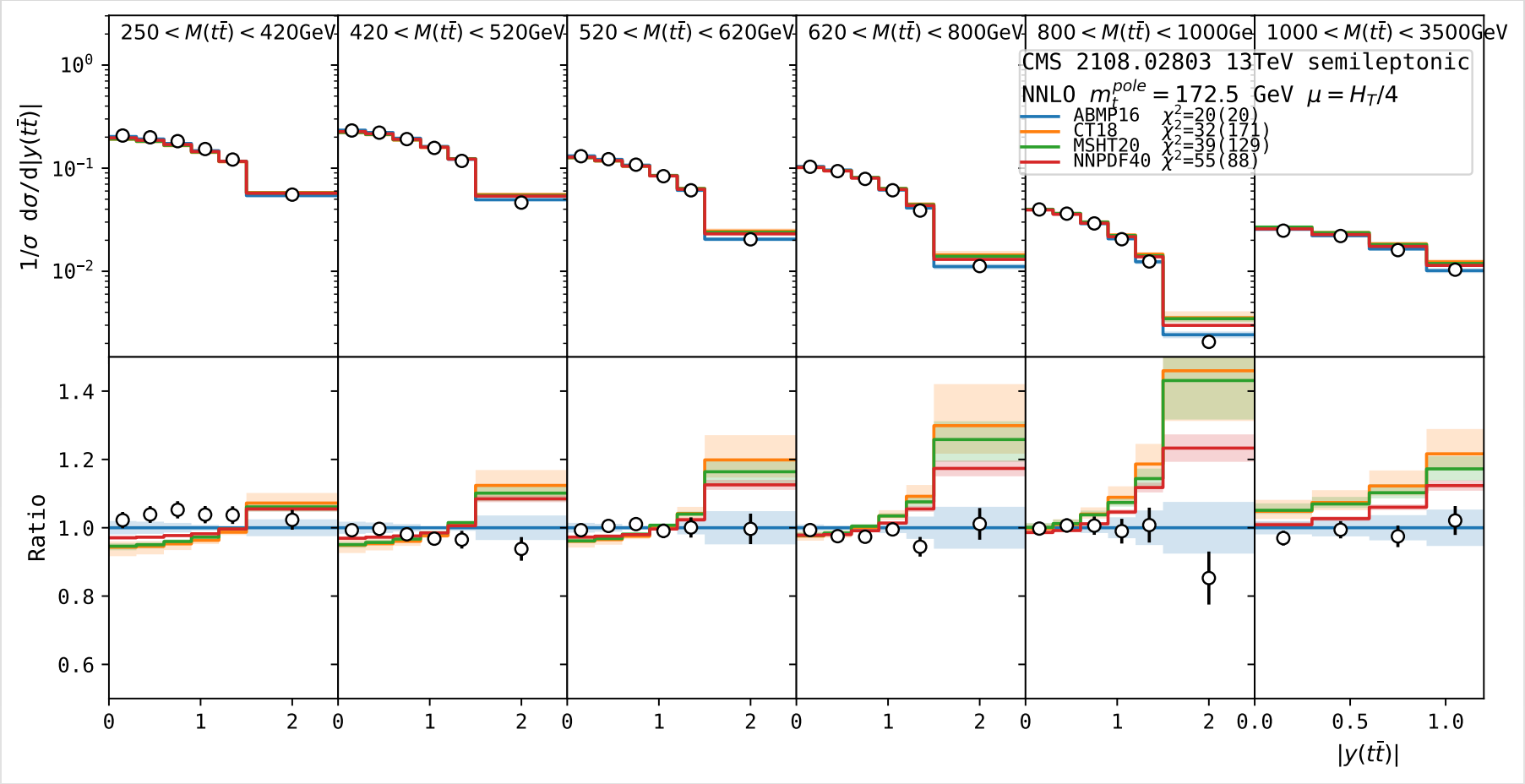
on the basis of work in progress
(see also [arXiv:2311.05509[hep-ph]], [arXiv:2407.00545[hep-ph]]
for some motivations underlying this work).

Physics at TeV Colliders, Les Houches, June 16 - 25, 2025

$pp \to t\bar{t} + X$ production

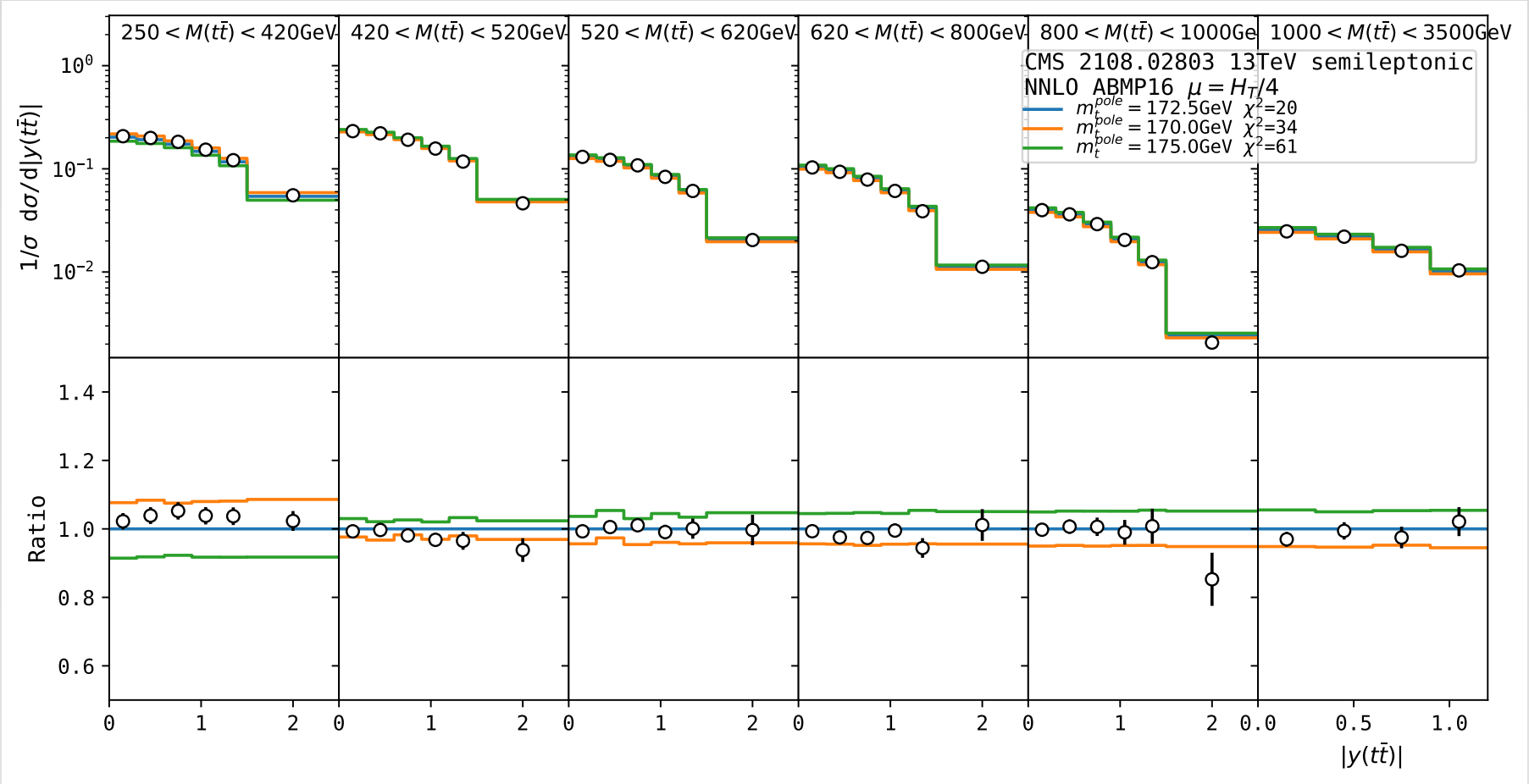
- $pp \to t\bar{t} + X$ production relevant for testing pQCD, PDF fits, top-quark mass fits, background for various SM and BSM processes at the LHC, BSM searches, etc...
- copiously produced at the LHC: good statistics available requires precise theory predictions.
- fixed-order predictions at NLO known since long.
- fixed-order predictions at NNLO QCD computed by two different IR-divergence subtraction methods, founding consistency.
- NLO+PS and NNLO+PS also available, as well as predictions including threshold resummation, top-quark decays, etc. (not considered in the following of this presentation, where we limit ourselves to predictions with stable top quarks at fixed order).

CMS TOP-20-001 vs NNLO predictions using different PDFs



- Fixed $m_t^{\text{pole}} = 172.5$ GeV, $\mu_r = \mu_f = H_T/4$
- Reported χ^2 values with (and without) PDF uncertainties
- All PDF sets describe data reasonably well, with best description by ABMP16
  - CT18, MSHT20 and NNPDF40 show clear trend w.r.t data at high $y(t\bar{t})$ (large x)
- This is most precise currently available dataset with finest bins

CMS TOP-20-001 vs NNLO predictions with ABMP16 and different m_t^{pole}



- Using ABMP16, $\mu_r = \mu_f = H_T/4$
- Reported χ^2 values with PDF uncertainties
- Large sensitivity to m_t^{pole} in the first $M(t\bar{t})$ bin (and even in other $M(t\bar{t})$ bins, thanks to cross-section normalisation)
- Fluctuations of theory predictions are $\lesssim 1\%$

How to speed-up the production of predictions for this process ?

NNLO computations with MATRIX are CPU intensive....

For reaching an accuracy $\Delta\sigma_{t\bar{t}} = 0.02\%$ on total cross sections:

- $\approx$ 350000 CPU hours/run ($\sim$30 years on a single CPU)
- for differential distributions, this corresponds to a statistical uncertainty in bins $\lesssim 0.5\%$

To save CPU consumption:

- generation of PineAPPL (or FastNLO, etc...) interpolation grids, with the possibility of changing a-posteriori PDFs, α_s, multiples of the μ_r, μ_f scales without re-running from scratch, and publications of these grids (e.g. via Ploughshare).
- MATRIX as an "event + counterevents"/ntuples generator
- Reduction of the number of negative weights in the ntuple samples

$t\bar{t}$ @ NNLO in MATRIX

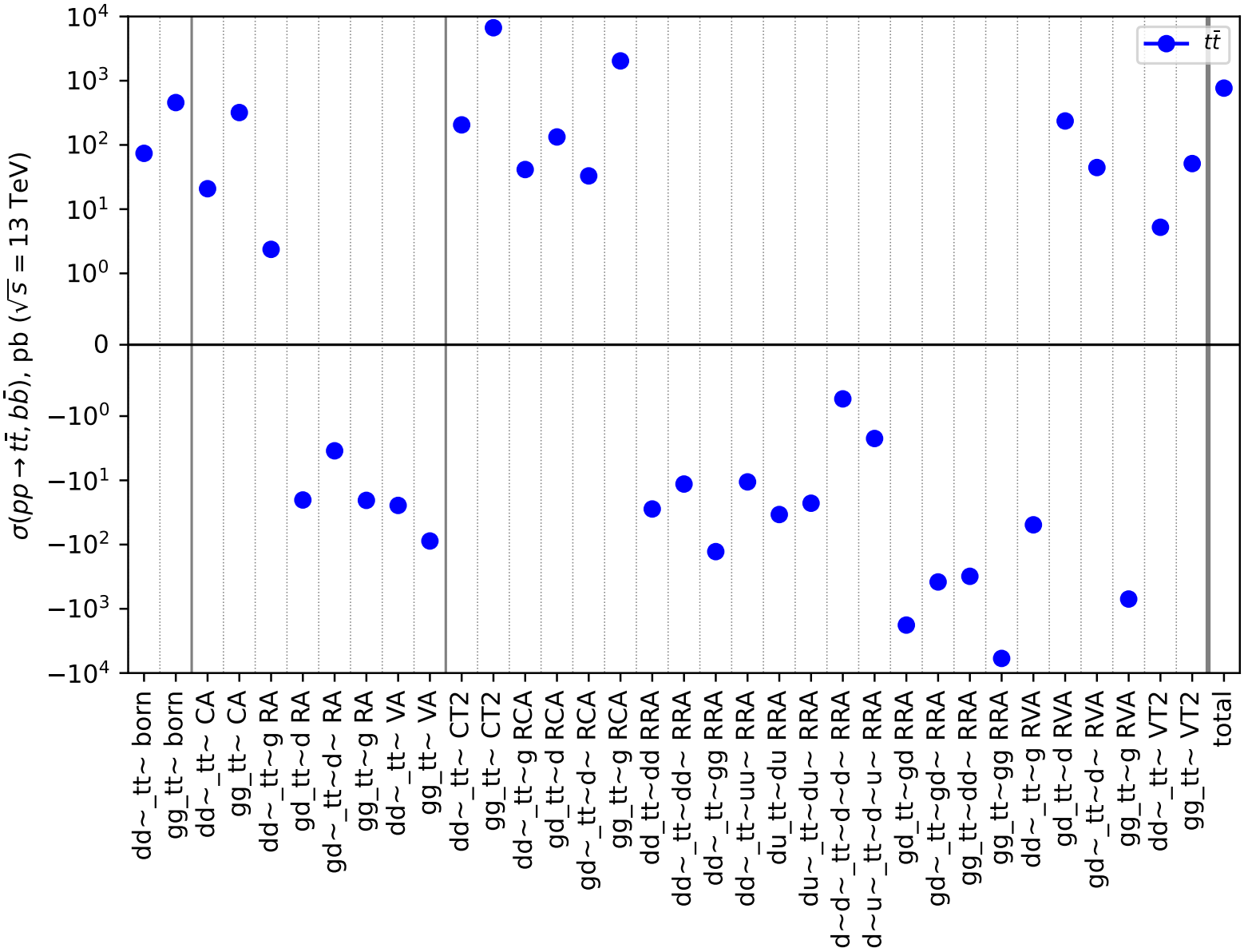
- q_T-subtraction:

$$d\sigma^{t\bar{t}}_{(N)NLO} = \mathcal{H}^{t\bar{t}}_{(N)NLO} \otimes d\sigma^{t\bar{t}}_{LO} + \left[d\sigma^{t\bar{t}+jet}_{(N)LO} - d\sigma^{t\bar{t},CT}_{(N)NLO} \right]$$

- MATRIX run structure:
  - ∗ the number of events in different channels is chosen to optimise the total uncertainty: balance of cross section magnitude and runtime
  - ∗ in many channels, 'event+counterevents'

order	contr.	channel	σ [pb]	time [s]	events	ms/event	counterevents
LO	born	dd~_tt~	73.9 ± 0.2	19	234748	0.08	1
		gg_tt~	456.5 ± 0.4	125	1483164	0.08	1
NLO	CA	dd~_tt~	20.8 ± 0.1	13	89374	0.14	6
		gg_tt~	319.3 ± 0.4	275	1834780	0.15	6
	RA	dd~_tt~g	2.4 ± 0.0	20	49998	0.40	13
		gd_tt~d	−20.2 ± 0.1	39	122419	0.32	7
		gd~_tt~d~	−3.5 ± 0.0	16	49999	0.32	7
		gg_tt~g	−20.5 ± 0.2	66	131387	0.50	13
	VA	dd~_tt~	−24.6 ± 0.1	70	50000	1.40	1
		gg_tt~	−88.0 ± 0.4	130	57892	2.24	1
NNLO	CT2	dd~_tt~	205.3 ± 0.6	500	225198	2.22	13
		gg_tt~	6675.6 ± 4.3	19150	4316824	4.44	9
	RCA	dd~_tt~g	41.4 ± 0.2	74	367542	0.20	6
		gd_tt~d	133.1 ± 0.3	189	857314	0.22	6
		gd~_tt~d~	32.9 ± 0.1	45	203916	0.22	6
		gg_tt~g	2032.7 ± 1.3	3275	14423630	0.23	6
	RRA	dd_tt~dd	−27.9 ± 0.3	118	131068	0.90	1–17
		dd~_tt~dd~	−11.4 ± 0.2	56	68013	0.82	1–13
		dd~_tt~gg	−128.8 ± 0.7	1158	815758	1.42	1–37
		dd~_tt~uu~	−10.6 ± 0.1	24	66735	0.36	1–5
		du_tt~du	−34.1 ± 0.2	126	217950	0.58	1–9
		du~_tt~du~	−22.7 ± 0.2	78	129277	0.60	1–9
		d~d~_tt~d~d~	−0.8 ± 0.0	45	49996	0.90	1–17
		d~u~_tt~d~u~	−2.2 ± 0.0	29	49997	0.58	1–9
		gd_tt~gd	−1793.9 ± 2.0	10374	6924587	1.50	1–29
		gd~_tt~gd~	−382.1 ± 1.0	1848	1248710	1.48	1–29
		gg_tt~dd~	−312.2 ± 0.8	1195	933749	1.28	1–21
		gg_tt~gg	−5926.6 ± 5.4	54688	18892527	2.89	1–37
	RVA	dd~_tt~g	−49.3 ± 0.4	420	81682	5.14	1
		gd_tt~d	236.1 ± 1.0	2183	419884	5.20	1
		gd~_tt~d~	44.5 ± 0.4	400	76878	5.20	1
		gg_tt~g	−705.0 ± 2.9	39974	1264466	31.61	1
	VT2	dd~_tt~	5.2 ± 0.2	672	50000	13.44	13
		gg_tt~	51.2 ± 1.1	2664	104476	25.50	9
total			**766.5 ± 8.2**	**140056**	**56023938**		**981854334**

$t\bar{t}$ @ NNLO in MATRIX: summary



- $\sigma(t\bar{t}) = 767$ pb, while RRA and CT2 contributions $\sim \pm 6000$ pb: large unc. on the sum

Applying cell resampling on top of $t\bar{t}$ events + counterevents

- Cell resampling redistributes the weights within a cell centered around a negative weight event and containing events and counterevents with both positive and negative weights, in such a way to decrease/minimize(?) the number of negative weights, preserving the total weight in the cell. See [arXiv:2109.07851] and [arXiv:2303.15246] for more info.
- A cell includes a number of neighbor events: key aspect is the definition of distance among events and its calculation.
- Various parameters affect the result, here we run with anti-kt jet-algorithm, with R=0.5 and pt(jet)=50 GeV, and we consider in particular the effect of the variation of the maximum cell size (mcs) parameter and the minimum weight (minw) parameter, which allows to put to 0 the weight of a number of events, reducing the size of the event sample
(minw = weight after weight redistribution, below which the corresponding event is discarded with probability=(1-|w|)/minweight)

  $\Rightarrow$ We study the resulting maximal and average distorsion of various NLO and NNLO distributions, the CPU-time consumption, the memory consumption, and the reduction of file size.

  reduction of file size = file size after cell resampling / file size before it

  distorsion definition = $|\sigma_{\textit{after cell resampling}} - \sigma_{\texttt{MATRIX}}|$ / numerical uncertainty$_{\texttt{MATRIX}}$

  distorsion evaluated for the CMS TOP-20-001 differential distributions

Effects of variations of maximum cell size and minimun weight parameters on NLO samples

mcs,minw	0.001	0.01	0.1	1.0	10.0
0.1	97s/1GB/0.860/0.00,0.07	94s/1GB/0.744/0.01,0.11	91s/1GB/0.581/0.06,1.02	93s/1GB/0.349/0.57,7.90	78s/1GB/0.116/3.21,69.58
1.0	97s/1GB/0.849/0.11,0.88	93s/1GB/0.744/0.12,0.88	88s/1GB/0.570/0.17,1.16	83s/1GB/0.337/0.57,6.53	77s/1GB/0.105/2.78,52.76
10.0	115s/1GB/0.837/0.39,2.00	115s/1GB/0.721/0.41,2.00	110s/1GB/0.547/0.42,1.97	104s/1GB/0.314/0.70,8.85	96s/1GB/0.081/2.29,20.06
25.0	215s/1GB/0.837/0.47,1.84	212s/1GB/0.733/0.48,1.81	203s/1GB/0.558/0.54,1.84	196s/1GB/0.291/0.86,8.21	193s/1GB/0.070/2.18,14.64
100.0	792s/1GB/0.826/1.90,44.90	644s/1GB/0.686/1.56,42.98	706s/1GB/0.430/1.73,42.68	634s/1GB/0.151/1.69,41.69	751s/1GB/0.035/3.18,44.10
-1	1827s/1GB/0.802/3.55,43.80	1704s/1GB/0.628/3.20,43.52	1831s/1GB/0.372/3.78,45.34	1913s/1GB/0.140/3.49,46.48	1691s/1GB/0.023/3.58,43.00

mcs,minw	0.001	0.01	0.1	1.0	10.0
0.1	558s/6GB/0.845/0.07,1.20	550s/6GB/0.720/0.09,1.20	550s/6GB/0.440/0.31,4.18	450s/6GB/0.147/1.03,15.32	424s/6GB/0.034/4.97,61.47
1.0	548s/6GB/0.843/0.25,1.75	527s/6GB/0.715/0.23,1.09	489s/6GB/0.437/0.43,3.92	444s/6GB/0.143/1.14,14.12	429s/6GB/0.032/4.49,55.74
10.0	665s/6GB/0.835/0.53,2.27	646s/6GB/0.708/0.55,2.22	590s/6GB/0.428/0.64,2.30	531s/6GB/0.133/1.16,9.99	515s/6GB/0.025/2.87,27.00
25.0	1503s/6GB/0.803/0.67,3.29	1417s/6GB/0.658/0.68,3.29	1365s/6GB/0.361/0.90,4.31	1244s/6GB/0.098/1.20,11.99	1331s/6GB/0.015/3.56,39.26
100.0	7397s/6GB/0.654/2.44,49.97	6967s/6GB/0.425/2.33,49.48	7007s/6GB/0.184/2.42,49.67	6319s/6GB/0.047/2.92,47.20	7103s/6GB/0.007/4.18,50.04
-1	28473s/6GB/0.580/4.21,55.12	26784s/6GB/0.359/4.20,53.69	28895s/6GB/0.162/4.06,53.69	25254s/6GB/0.042/3.45,51.49	30102s/6GB/0.007/5.26,54.97

mcs,minw	0.001	0.01	0.1	1.0	10.0
0.1	893s/23GB/0.787/0.15,1.01	846s/23GB/0.574/0.17,1.07	973s/23GB/0.263/0.47,2.98	793s/23GB/0.059/1.68,32.42	775s/23GB/0.014/7.15,92.25
1.0	1060s/23GB/0.785/0.38,1.68	864s/23GB/0.572/0.41,1.68	812s/23GB/0.261/0.74,8.51	810s/23GB/0.057/2.04,29.38	803s/23GB/0.012/7.46,131.51
10.0	1306s/23GB/0.777/0.58,1.99	1299s/23GB/0.563/0.60,1.99	1265s/23GB/0.245/0.76,3.01	1502s/23GB/0.048/1.70,19.36	1233s/23GB/0.007/5.17,46.80
25.0	5797s/23GB/0.678/0.72,3.16	4912s/23GB/0.425/0.74,3.26	4903s/23GB/0.141/0.91,8.52	4965s/23GB/0.026/1.91,11.84	4958s/23GB/0.003/6.65,105.9
100.0	14338s/23GB/0.480/2.13,47.21	14353s/23GB/0.238/2.11,47.29	15175s/23GB/0.077/2.18,46.73	13960s/23GB/0.015/2.72,46.52	14153s/23GB/0.002/6.21,53.7
-1	37447s/23GB/0.449/3.10,47.37	34686s/23GB/0.228/3.13,47.50	34899s/23GB/0.075/3.19,47.49	34971s/23GB/0.014/3.54,47.96	34761s/23GB/0.002/6.93,58.6

Different tables correspond to NLO cross-section accuracy of $\sim$ 0.5%, 0.1%, 0.06%, respectively.

Our biggest NLO sample includes 6.5 10^7 events + counterevents (1.5 GB).

CPU time increases with incresing mcs, memory consumption does not. Reduction of file size is more pronounced for large minw accompanied by large mcs. Distorsion increases with increasing mcs and increasing minw.

Effects of variations of maximum cell size and minimun weight parameters on NNLO samples

mcs,minw	0.001	0.01	0.1	1.0
0.1	1041s/17GB/0.682/0.02,0.14	1098s/17GB/0.560/0.03,0.14	1362s/17GB/0.408/0.07,0.99	1159s/17GB/0.230/0.31,6.11
1.0	1255s/17GB/0.662/0.07,0.42	1054s/17GB/0.535/0.07,0.44	1228s/17GB/0.377/0.10,1.22	1274s/17GB/0.197/0.28,2.93
10.0	2781s/17GB/0.642/0.20,1.41	2989s/17GB/0.513/0.21,1.45	2540s/17GB/0.354/0.23,1.55	2609s/17GB/0.177/0.39,4.13
25.0	114380s/17GB/0.678/0.70,10.61	97881s/17GB/0.577/0.71,10.91	93798s/17GB/0.374/0.69,10.61	104330s/17GB/0.145/0.86,10.69
100.0	0s/0GB/0.000/0.00,0.00	0s/0GB/0.000/0.00,0.00	0s/0GB/0.000/0.00,0.00	0s/0GB/0.000/0.00,0.00
-1	0s/0GB/0.000/0.00,0.00	0s/0GB/0.000/0.00,0.00	0s/0GB/0.000/0.00,0.00	0s/0GB/0.000/0.00,0.00

mcs,minw	0.001	0.01	0.1	1.0
0.1	24525s/341GB/0.507/0.08,0.98	24414s/341GB/0.352/0.14,1.68	23953s/341GB/0.177/0.64,15.67	23813s/341GB/0.050/1.71,14.39
1.0	27914s/341GB/0.477/0.16,1.07	27576s/341GB/0.315/0.19,1.07	26797s/341GB/0.143/0.41,5.16	26726s/341GB/0.034/1.26,8.58
10.0	495939s/341GB/0.487/0.52,4.50	495606s/341GB/0.303/0.53,4.43	624523s/341GB/0.108/0.61,4.43	501740s/341GB/0.020/1.15,7.20
25.0	0s/0GB/0.000/0.00,0.00	0s/0GB/0.000/0.00,0.00	0s/0GB/0.000/0.00,0.00	0s/0GB/0.000/0.00,0.00
100.0	0s/0GB/1.000/0.00,0.00	0s/0GB/1.000/0.00,0.00	0s/0GB/1.000/0.00,0.00	0s/0GB/1.000/0.00,0.00
-1	0s/0GB/1.000/0.00,0.00	0s/0GB/1.000/0.00,0.00	0s/0GB/1.000/0.00,0.00	0s/0GB/1.000/0.00,0.00

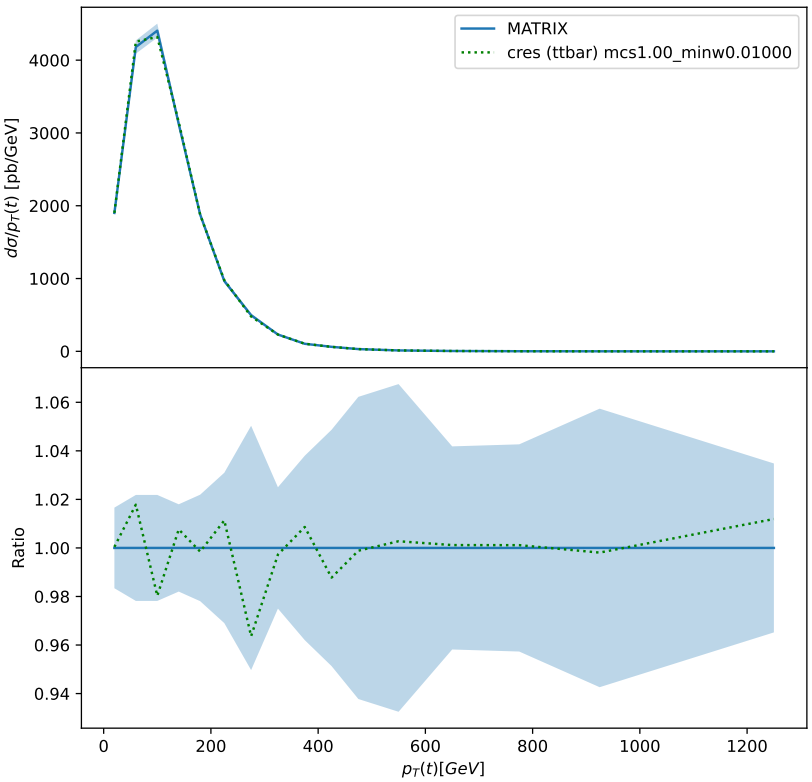
Different tables correspond to NNLO cross-section accuracy of 5% and 1%, respectively.

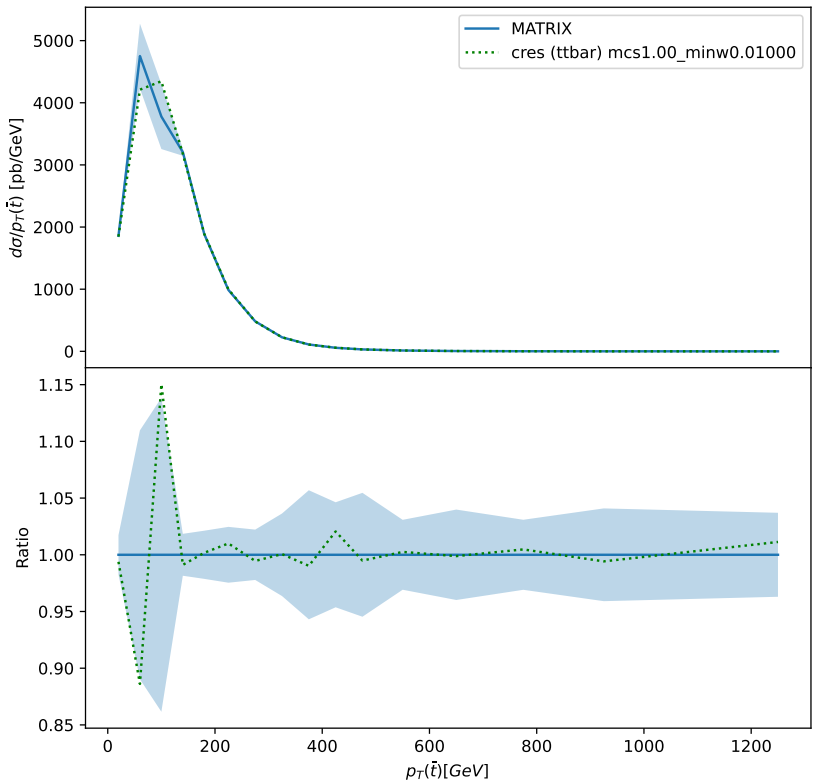
Our biggest NLO sample includes 10^9 events + counterevents (40 GB).

The sequence of 0's correspond to jobs that never finished...

Here reduction of file size and distorsion at fixed minw are not always increasing for larger mcs.

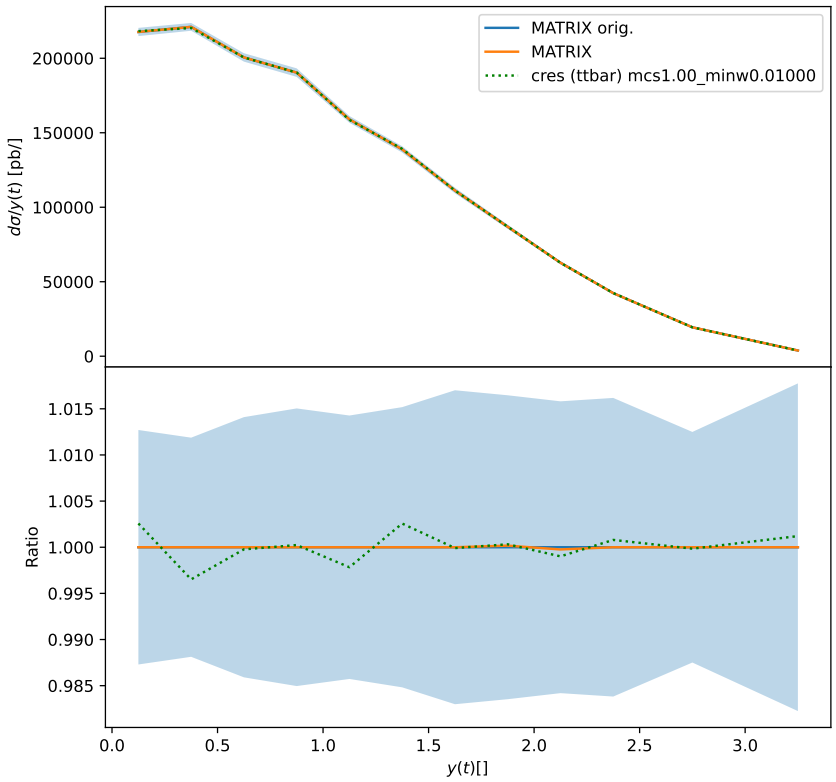
$p_{T,t}$ and $p_{T,\bar{t}}$ differential distributions at NLO

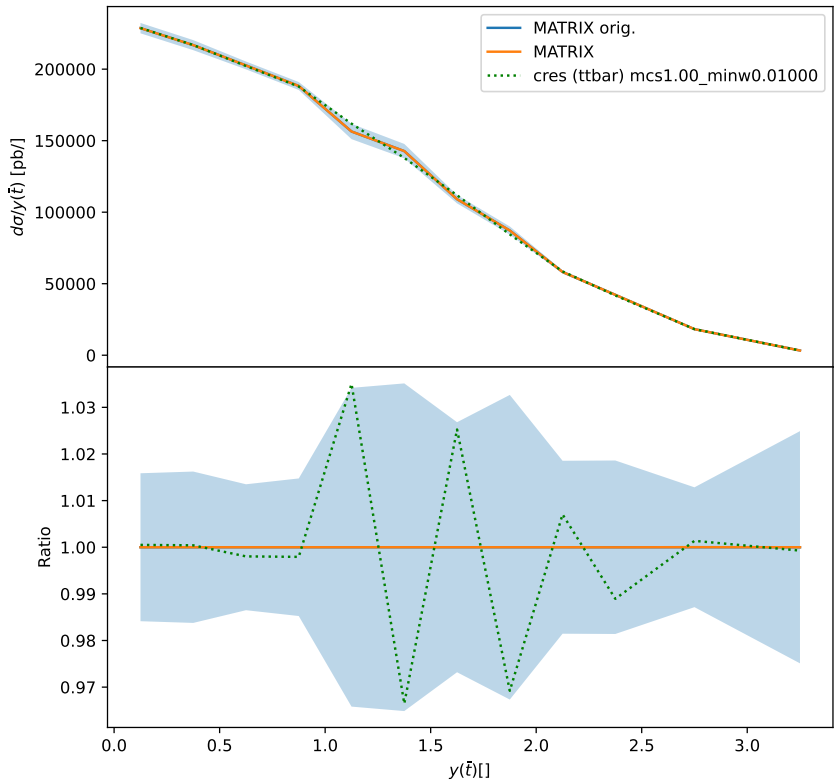




* Differences in the $p_{T,t}$ and $p_{T,\bar{t}}$ distributions seem to be reduced by cell resampling.
* Cell resampling with default metrics definition performs better at large p_T.

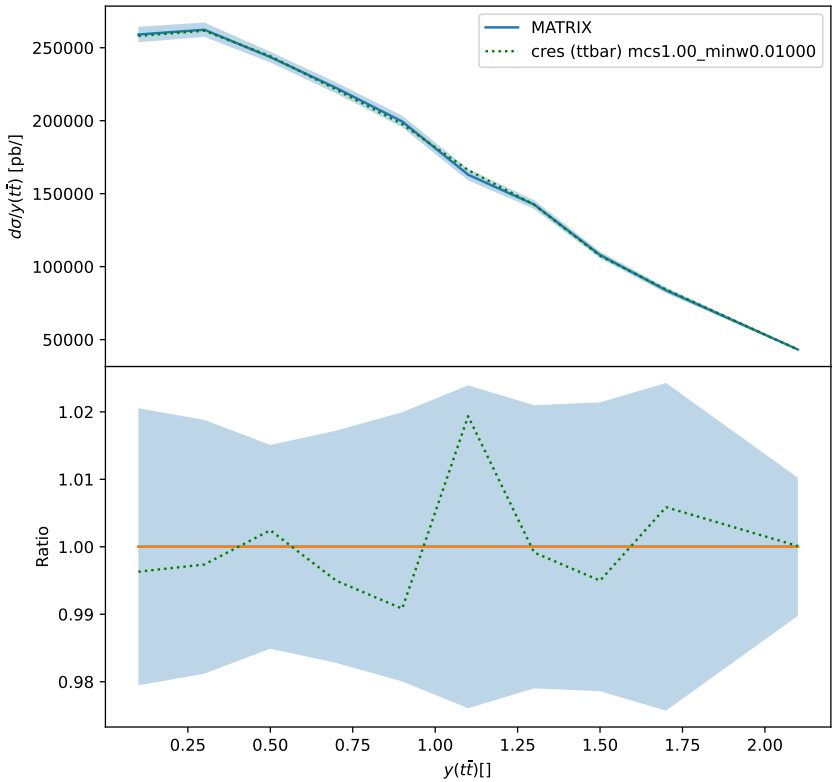
y_t and $y_{\bar{t}}$ differential distributions at NLO

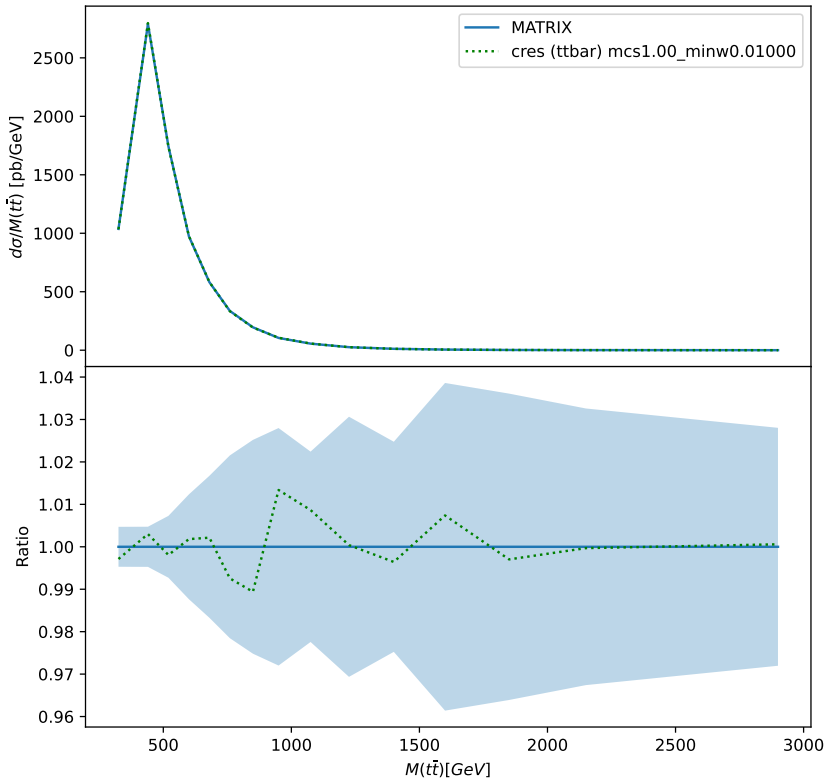




∗ Cell resampling does not seem to perform equally well on y_t and $y_{\bar{t}}$, but the distributions before it have already a different numerical accuracy.

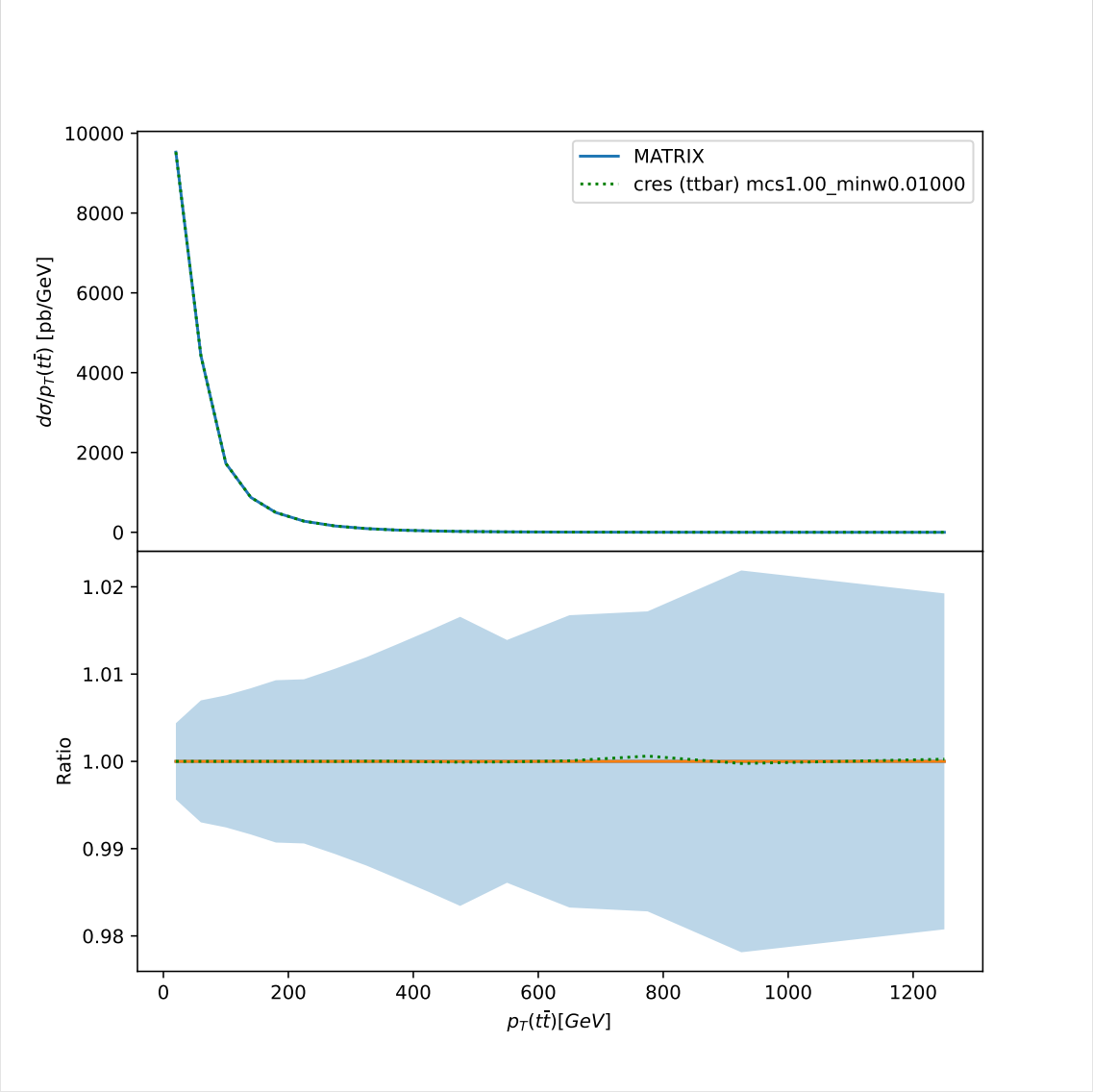
$y_{t\bar{t}}$ and $m_{t\bar{t}}$ differential distributions at NLO





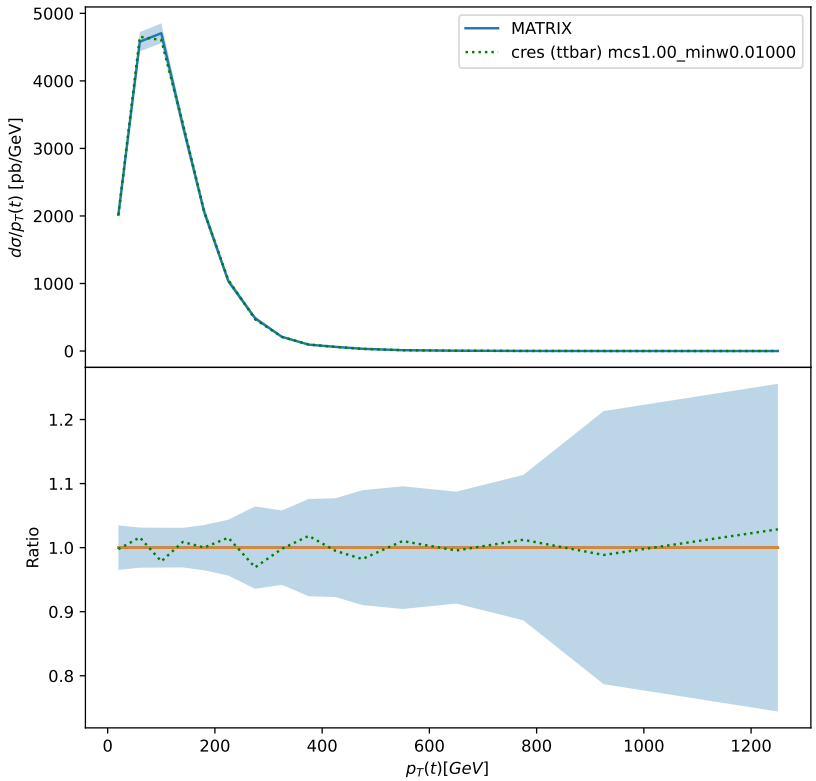
∗ Maximal distorsion of $\sim$ 1.5 - 2% on both distributions.

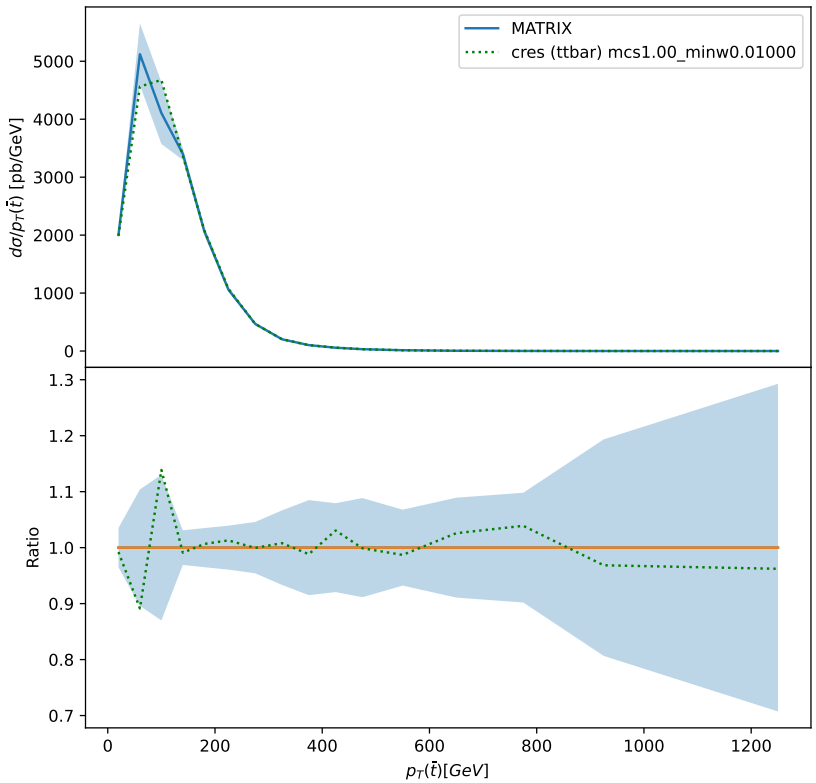
$p_{T,t\bar{t}}$ differential distribution at NLO



∗ It should not be modified by cell resampling: good cross-check.

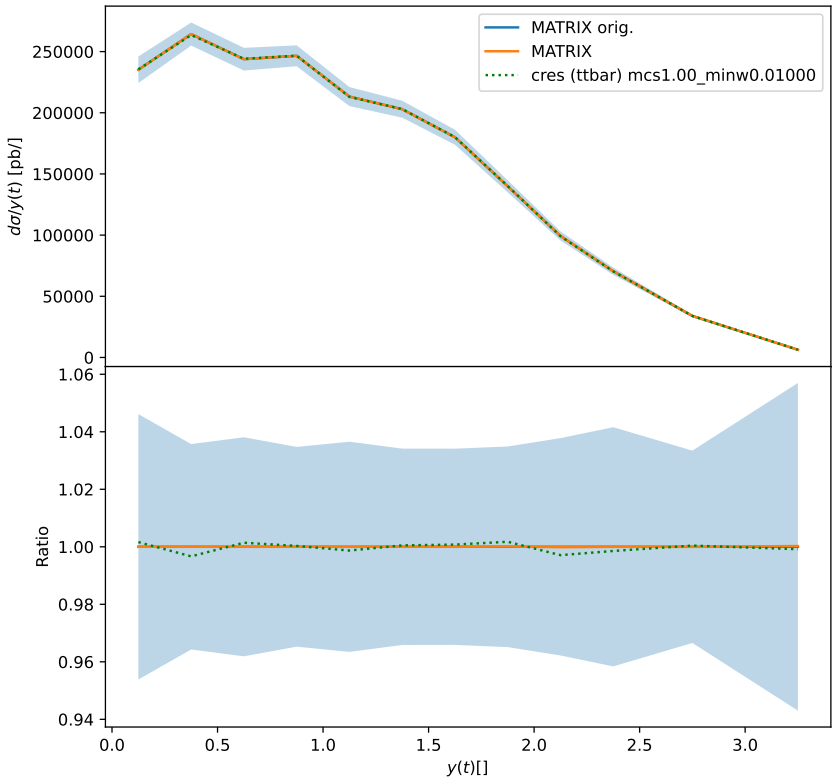
$p_{T,t}$ and $p_{T,\bar{t}}$ differential distributions at NNLO

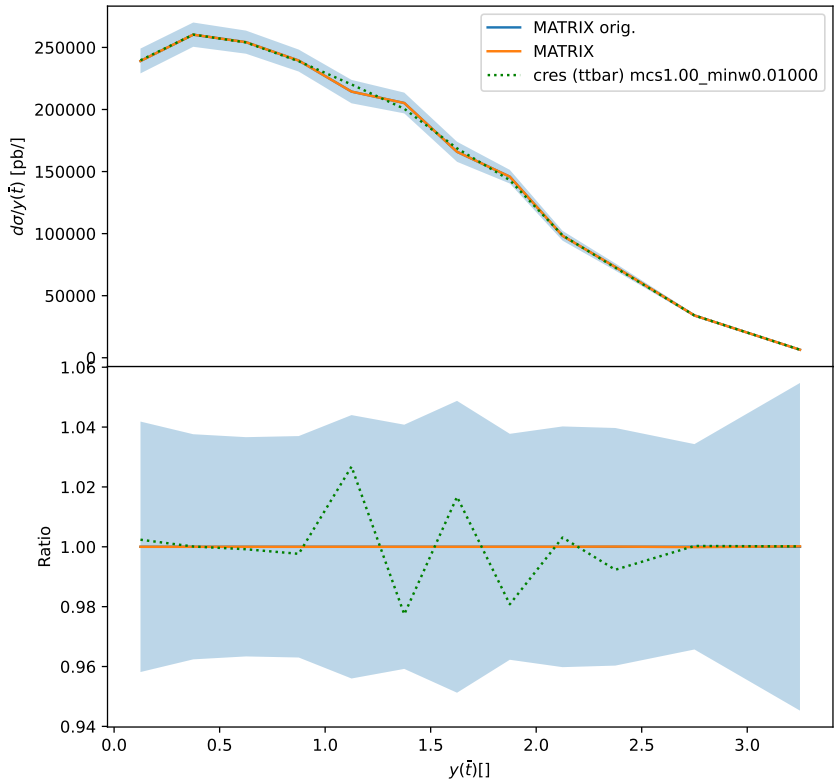




∗ Similar observations as for NLO.

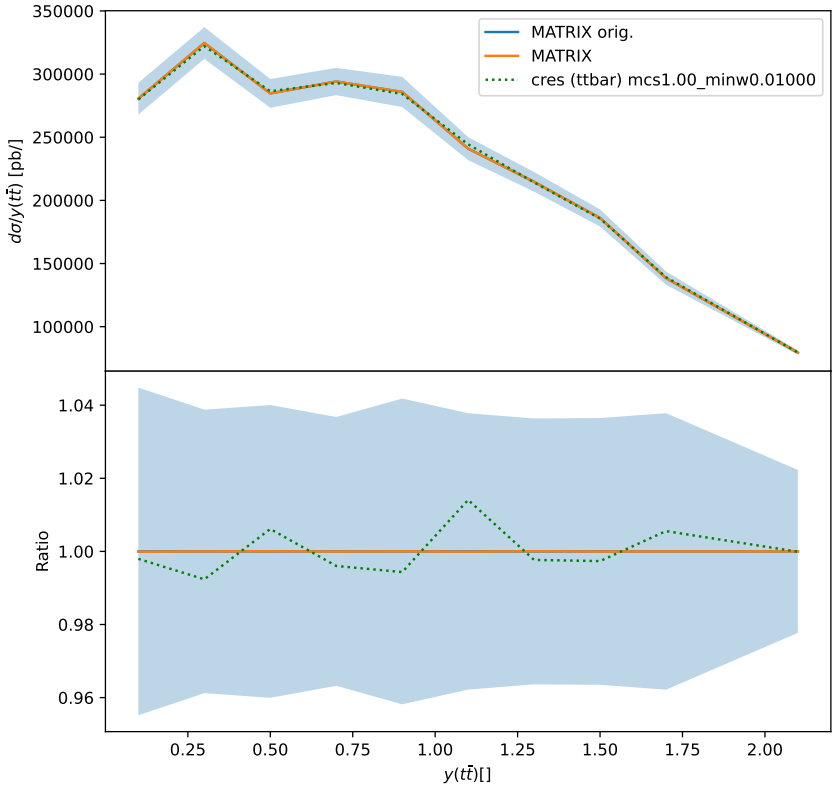
y_t and $y_{\bar{t}}$ differential distributions at NNLO

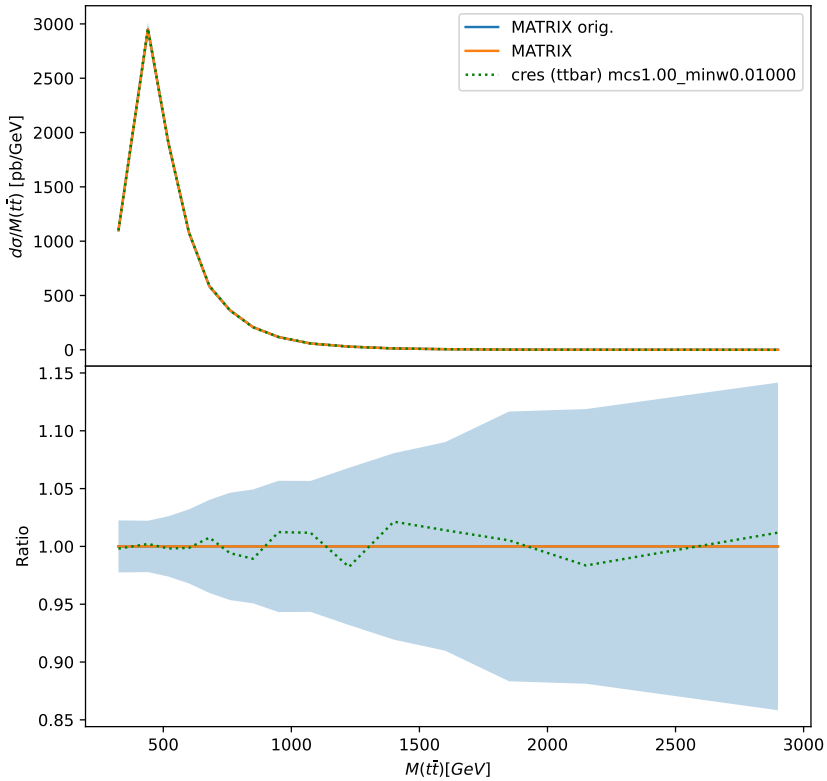




∗ Similar observations as for NLO.

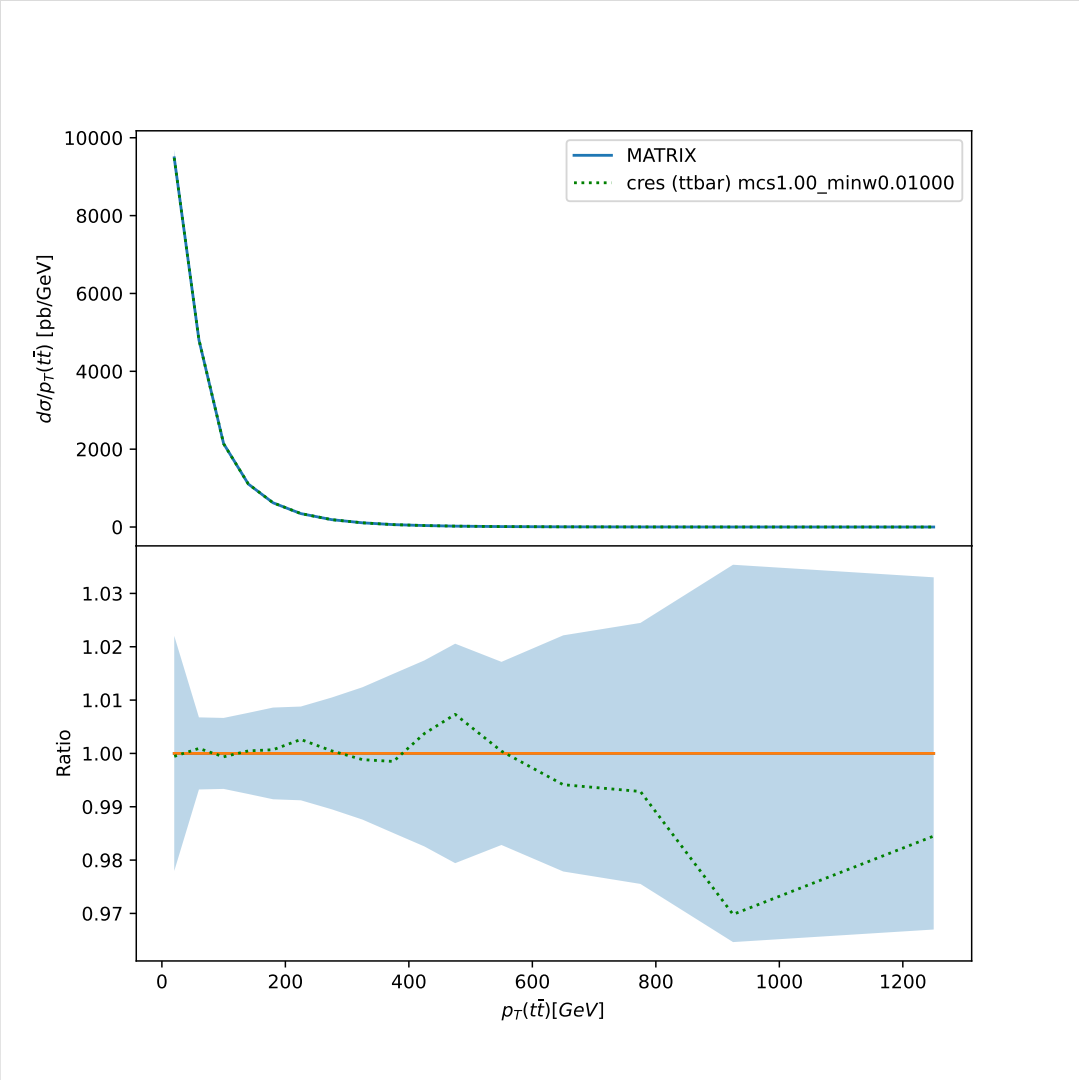
$y_{t\bar{t}}$ and $m_{t\bar{t}}$ differential distributions at NNLO





∗ Distorsions at the level of $\sim 1\% - 2\%$ for both distributions.

$p_{T,t\bar{t}}$ differential distribution at NNLO



∗ Here there is distorsion at large p_T.

Conclusions

- Cell resampling can be used to reduce the size of the ntuples to 30-60% without significant distortion of kinematic distributions.
  - However, assigning numerical uncertainties to distributions after cell resampling is still an open issue....
- So far, we could NOT reach a $< 1\%$ total uncertainty at NNLO due to the high memory consumption (> 1 TB). How can we apply scaling to huge samples ?
- Values of the parameters mcs and minw have to be adjusted depending on the total uncertainty and perturbative order, this might require time-consuming iterations
  - for 1% NNLO accuracy we find that mcs = 1 and minw = 0.01 are a good compromise and we used them for the plots of differential cross sections
  - some distributions are reproduced much better than others. This depends on the order, the metrics definition, the numerical uncertainty of the initial simulation.... How can we better fix the parameters for cell resampling and the metrics ? In principle we should be able to reproduce any distribution with a desired accuracy.
- Variation of other parameters/use of different distance definition in our case-study still to be investigated....

Thank you for your attention and your suggestions are welcome!