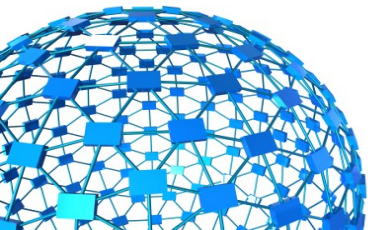


Monte Carlo samples for a 100 TeV collider

S. Chekanov (ANL)

FCC software meeting

(July 10, 2014)



Event repository with MC samples

- A number of new MC event samples for a 100 TeV collider were created in 2014 (after Snowmass in 2013)
 - Main focus: Higgs, top and some background SM processes
 - Used to study physics reach at a 100 collider, identify interesting processes
- Keep only truth-level events at LO+PS and NLO: No pileup, No reco samples.
- Many NLO+PS samples were generated using ANL BlueGene/Q supercomputer
 - MIRA <https://www.alcf.anl.gov/user-guides/mira-cetus-vesta>
 - MCFM and JETPHOX etc. were interfaced with MPI

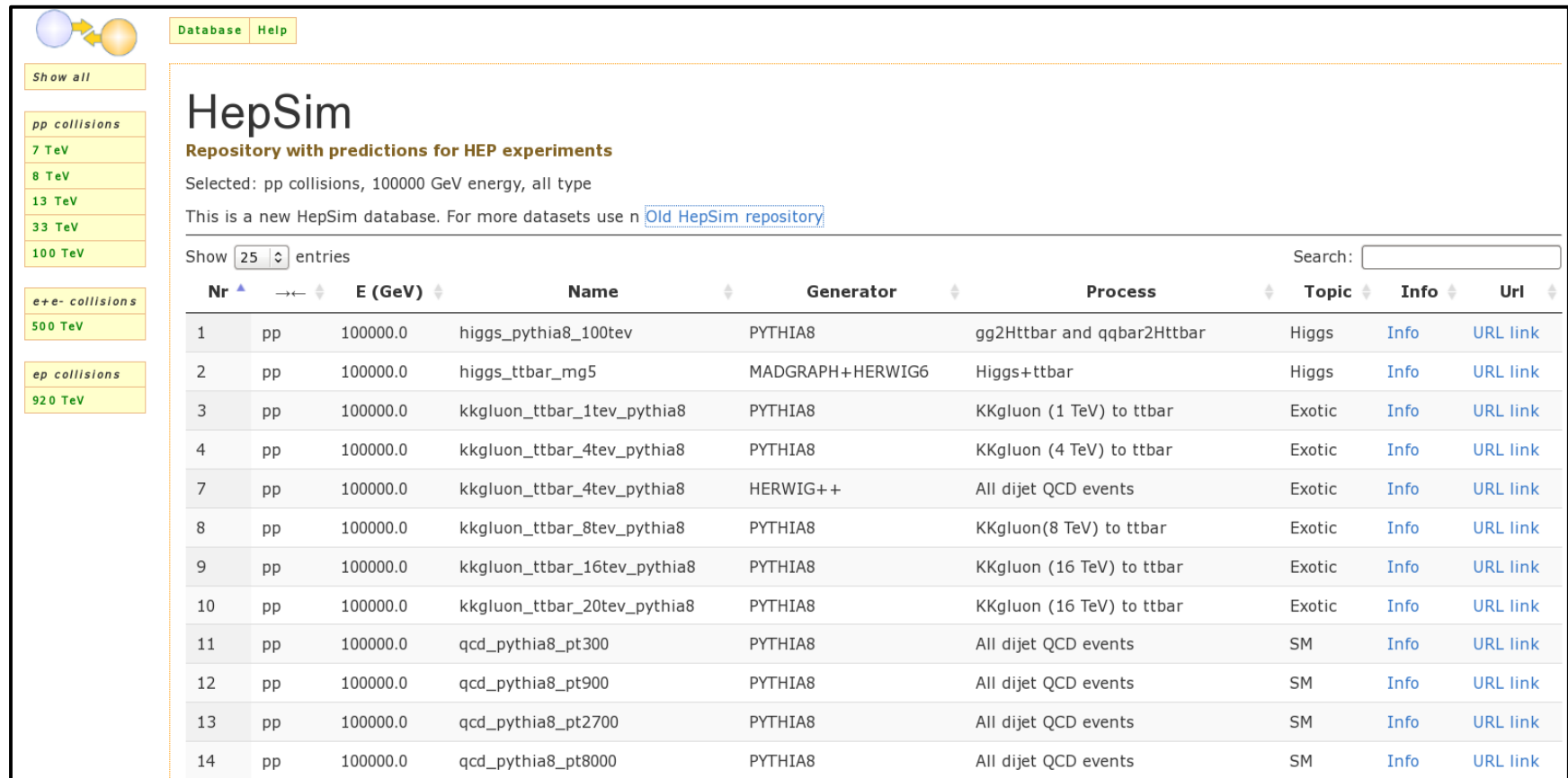


- ANL hosts ~40 “signal” samples (20k files, ~150M events) for top, Higgs, etc. processes

Primary URL link: <https://atlaswww.hep.anl.gov/hepsim/>

Old: <http://mc.hep.anl.gov/asc/hepsim/events/>

HepSim front-end: <https://atlaswww.hep.anl.gov/hepsim/>



The screenshot displays the HepSim web interface. At the top, there are navigation links for 'Database' and 'Help'. A sidebar on the left contains categories for 'pp collisions' (7 TeV, 8 TeV, 13 TeV, 33 TeV, 100 TeV), 'e+e- collisions' (500 TeV), and 'ep collisions' (920 TeV). The main content area features the title 'HepSim' and the subtitle 'Repository with predictions for HEP experiments'. Below this, it states 'Selected: pp collisions, 100000 GeV energy, all type' and provides a link to the 'Old HepSim repository'. A search bar and a 'Show 25 entries' dropdown are present. The central part of the interface is a table listing 14 Monte Carlo samples with columns for 'Nr', 'E (GeV)', 'Name', 'Generator', 'Process', 'Topic', 'Info', and 'Url'.

Nr		E (GeV)	Name	Generator	Process	Topic	Info	Url
1	pp	100000.0	higgs_pythia8_100tev	PYTHIA8	gg2Httbar and qqbar2Httbar	Higgs	Info	URL link
2	pp	100000.0	higgs_ttbar_mg5	MADGRAPH+HERWIG6	Higgs+ttbar	Higgs	Info	URL link
3	pp	100000.0	kkgluon_ttbar_1tev_pythia8	PYTHIA8	KKgluon (1 TeV) to ttbar	Exotic	Info	URL link
4	pp	100000.0	kkgluon_ttbar_4tev_pythia8	PYTHIA8	KKgluon (4 TeV) to ttbar	Exotic	Info	URL link
7	pp	100000.0	kkgluon_ttbar_4tev_pythia8	HERWIG++	All dijet QCD events	Exotic	Info	URL link
8	pp	100000.0	kkgluon_ttbar_8tev_pythia8	PYTHIA8	KKgluon(8 TeV) to ttbar	Exotic	Info	URL link
9	pp	100000.0	kkgluon_ttbar_16tev_pythia8	PYTHIA8	KKgluon (16 TeV) to ttbar	Exotic	Info	URL link
10	pp	100000.0	kkgluon_ttbar_20tev_pythia8	PYTHIA8	KKgluon (16 TeV) to ttbar	Exotic	Info	URL link
11	pp	100000.0	qcd_pythia8_pt300	PYTHIA8	All dijet QCD events	SM	Info	URL link
12	pp	100000.0	qcd_pythia8_pt900	PYTHIA8	All dijet QCD events	SM	Info	URL link
13	pp	100000.0	qcd_pythia8_pt2700	PYTHIA8	All dijet QCD events	SM	Info	URL link
14	pp	100000.0	qcd_pythia8_pt8000	PYTHIA8	All dijet QCD events	SM	Info	URL link

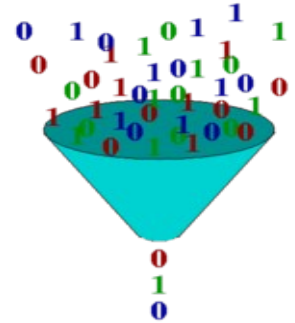
A SQL database stores metadata of each sample. Registered users can add MC samples

- HEPMC front-end can point to any URL location
- Currently, all links point to the ANL file storage with Apache web access
- Designed to support any file format ROOT, HepML, StdHEP, HEPMC, ProMC etc.
- Currently, all samples are in the ProMC file format (see later)

Available samples: NLO, NLO+PS, LO+PS

- MG5 (NLO+PS+hadr): TTbar
- MG5 (NLO+PS+hadr):: Higgs+jj
- MG5 (NLO+PS+hadr):: Higgs+TTbar
- PYHIA8, HERWIG++ for dijet QCD (~ 100 fb)
- MCFM (NLO):: Higgs $\rightarrow \gamma\gamma$
- MCFM (NLO): Inclusive gamma
- MCFM (NLO): TTbar
- PYTHIA8 (LO) for Z' and g(KK) with masses from 6 to 20 TeV
- PYTHIA8 (LO) for W'
- PYTHIA8 (LO) W/Z+jets
- NLOjet++ (NLO) for inclusive jets (bins in pT)
- JETPHOX (NLO) for inclusive photons (bins in pT)

Data size reduction & data storage issues



- **Disk problem at Snowmass:**
 - Traditional “text/xml” data formats (like HEPMC) were too large to keep
 - Example: 100 ttbar events with 140 pileup use ~1.2 GB (300 MB after compression)
 - “algorithmic” compression (gzip/zip) typically makes difficult to read / write
 - HEPMC / LHE samples with truth information were removed to save space

- **Removing truth-level files is OK when we are certain about what detector geometry is used and no further processing is need**
 - this is not the case for future colliders. Detector geometry is not settled!

- **We need to archive truth-level samples!**
 - Generation of truth-level becomes CPU intensive
 - large CM energies, NLO, NLO+PS matching etc
 - Need to ensure a long-term preservation of theoretical predictions
 - Need for a public access to the samples



Future can be “exotic”

- We do not know what computer technologies / OS/ programming language will be available in 20-30 years from now
- To ensure long-term availability of predictions, storing data in TEXT/XML format is safest, but technically is too challenging for “large data”
- During Snowmass13, a new data format has been developed:
 - highly compressed binary format
 - language- neutral (C++, Java, CPython, Jython, PHP, Ruby, Groovy, Scala) etc.
 - platform-neutral (Windows, Linux, Mac OS, IBM BG/Q, Android etc.)
 - XML-like, but ~10 faster
 - supported in industry (I.e. Google)
 - uses “varints” (the number of allocated bytes depends on numeric value)
 - very effective compression for pile-up events → particles with small momenta use small disk space
 - ROOT uses a fixed-bytes, i.e. 0 uses the same number of bytes as 2^{31} value



Data format

<http://atlaswww.hep.anl.gov/asc/promc/>

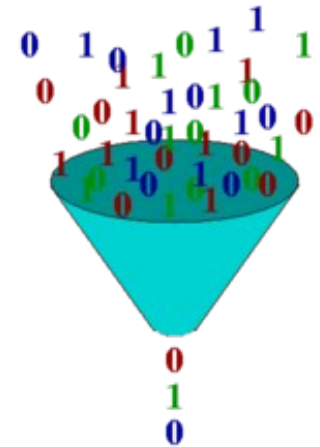
S.C., E.May, K. Strand, P. Van Gemmeren,
<http://arxiv.org/abs/1311.1229> (CPC in press)

■ ProMC is “archive” format for structured data:

- files are smaller compared to any known binary format (uses variable-byte encoding)
- Supported on known platform & any programming language
- Uses ProtocolBuffers to encode data. Used by Google to encode personal info.
- Random access to any event. Includes “meta” information (MC logfiles & data layouts)
- Included to Delphes3 & Pythia8

Benchmarks for 10,000 tbar MC events

File format	File size (in MB)	Read speed (in seconds)		
		C++	JAVA VM	JYTHON
ProMC	307	15.8	11.7,12.1*	33.3, 34.6*
ROOT	423	20.4	–	–
LHEF	2472	84.7	9.0, 9.6*	–
HEPMC	2740	175.1	–	–
LHEF (gzip)	712	–	–	–
HEPMC(gzip)	1021	–	–	–
LHEF (bzip2)	552	–	–	–
HEPMC(bzip2)	837	–	–	–
LHEF (lzma)	513	–	–	–
HEPMC(lzma)	802	–	–	–



ProMC files are ~30-40% smaller than ROOT and are faster to process in C++
Java and C++ show surprisingly similar performance



MC outputs

- Different MC generators use different methods to write ProMC files
 - **Pythia8** writes ProMC naively (see *main46.cc example*)
 - **Pythia6** uses FortranProMC package for output
 - (<https://www.hepforge.org/downloads/promc> K.Strand, E.May)
 - **HERWIG++** uses the conversion *hepmc2promc*
 - **MG5** uses the conversion *stdhep2promc*
- NLO programs (**MCFM**, **Jetphox**, **NLOjet++**) are interfaced with ProMC to write events in one step
- Many MC generators and ProMC package are deployed on Mira (BlueGene/Q supercomputer at ANL)
- See **BG@HepLib** library:
 - <https://atlaswww.hep.anl.gov/asc/wikidoc/doku.php?id=hpc:bghep:start>



ProMC tools <http://atlaswww.hep.anl.gov/asc/promc/>

- A special event viewer has been developed
- The file browser can read arbitrary event in >4 GB files (using random access) and show events in human-readable form (with particle names!)

ProMC Browser

File Metadata Data layout Help

Search (Regex Pattern):

No	Name	PID	Status	M1	M2	D1	D2	Px (GeV)	Py (GeV)	Pz (GeV)	E (GeV)	M (GeV)	X (mm)	Y (mm)	Z (mm)	T (s)
1	generator	90	11	0	0	0	0	0	0	0	14,000	14,000	0	0	0	0
2	p^+	2212	4	0	0	457	0	0	0	7,000	7,000	0.938	0	0	0	0
3	p^+	2212	4	0	0	458	0	0	0	-7,000	7,000	0.938	0	0	0	0
4	g	21	21	6	0	5	0	0	0	56.273	56.273	0	0	0	0	0
5	g	21	21	7	7	5	0	0	0	-69.415	69.415	0	0	0	0	0
6	H_1^0	25	22	3	4	8	8	0	0	-13.141	125.688	124.999	0	0	0	0
7	g	21	41	10	0	9	3	0	0	122.904	122.904	0	0	0	0	0
8	g	21	42	11	11	4	4	0	0	-69.415	69.415	0	0	0	0	0
9	H_1^0	25	44	5	5	12	12	42.556	-4.162	11.861	132.641	124.999	0	0	0	0
10	H_1^0	25	44	8	8	17	17	48.103	-6.546	13.229	134.746	124.999	0	0	0	0
11	H_1^0	25	44	12	12	26	26	55.252	-4.612	14.945	137.558	124.999	0	0	0	0
12	H_1^0	25	44	17	17	34	34	56.169	-7.648	15.449	138.119	124.999	0	0	0	0
13	H_1^0	25	44	26	26	74	74	54.613	-8.722	15.959	137.616	124.999	0	0	0	0
14	H_1^0	25	44	34	34	459	459	54.506	-8.816	15.993	137.583	124.999	0	0	0	0
15	H_1^0	25	62	74	74	593	594	54.531	-8.548	15.909	137.566	124.999	0	0	0	0
16	b	5	23	459	0	595	596	-26.779	4.021	42.801	50.875	4.8	0	0	0	0
17	b~	-5	23	459	0	597	597	81.31	-12.569	-26.891	86.692	4.8	0	0	0	0
18	b	5	51	593	0	603	603	-26.285	2.611	41.412	49.353	4.8	0	0	0	0
19	b~	-5	52	594	594	600	600	80.979	-12.517	-26.782	86.34	4.8	0	0	0	0
20	b~	-5	52	597	597	615	615	76.472	-11.813	-25.284	81.547	4.8	0	0	0	0
21	b	5	52	595	595	624	624	-25.806	2.563	40.664	48.467	4.8	0	0	0	0
22	b~	-5	52	600	600	621	621	75.815	-11.712	-25.065	80.848	4.8	0	0	0	0
23	b~	-5	52	615	615	675	0	72.863	-11.274	-24.077	77.71	4.8	0	0	0	0
24	b	5	52	603	603	678	678	-24.895	2.497	39.217	46.766	4.8	0	0	0	0
25	b~	-5	73	620	621	687	687	75.218	-11.717	-25.251	80.379	5.298	0	0	0	0

ProMC version=2 Total events=10000 Event=4 90.833MI

The browser unpacks “varints” into the usual numbers and show particle names using a look-up table

Examples

- Show generator logfile (use browser). In command line:
 - `unzip -p file.promc logfile.txt`
- Converting to ROOT tree (ProMC package should be installed)
 - `promc2root file.promc output.root`
- Fast detector simulation (ProMC & Delphes should be installed):
 - `DelphesProMC delphes_card.tcl output.root file.promc`
- Looking at separate events:
 - `wget http://atlaswww.hep.anl.gov/asc/promc/download/browser_promc.jar`
 - `java -jar browser_promc.jar file.promc`

ProMC conversion tools	
<code>hepmc2promc</code>	Converts a HEPMC 2.03.11 file [4] to the ProMC file format.
<code>promc2hepmc</code>	Converts a ProMC file to a HEPMC 2.03.11 file [4].
<code>promc2root</code>	Converts stores a ProMC file in a ROOT tree [3].
<code>stdhep2promc</code>	converts a STDHEP file [5] to the ProMC file.

S.C., E.May, K. Strand, P. Van Gemmeren,
<http://arxiv.org/abs/1311.1229> (CPC in press)



Creating analysis code

- ProMC files are “self-describing”. Language-neutral data layouts are kept inside the files. One can generate C++, Java and Python analysis code for reading and writing data from existing ProMC file.
- Example: Create analysis code from a downloaded ProMC file:
 - **Step 1:** Download a file (“file.promc”) from <https://atlaswww.hep.anl.gov/hepsim>
 - **Step 2:** Look at info and generate analysis code:

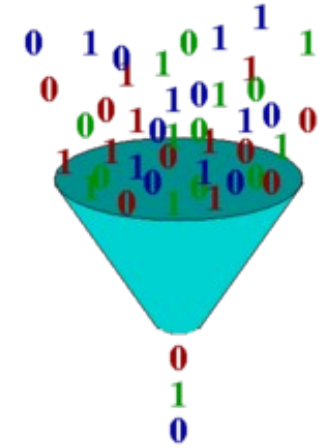
```
promc_info file.promc # check information about this file
promc_proto file.promc # extracts data layouts into the directory "proto"
promc_code # generate C++, Java and Python code in src/, java/, python/
make # compiles C++ code “reader.cc”
./reader file.promc # runs over all truth events
```

- **Step 3:** Use ROOT to fill histograms etc.
- **Step 4:** Try Python (slower!) and Java code (faster!) from python/ and java/ directories

Read more: <https://atlaswww.hep.anl.gov/asc/wikidoc/doku.php?id=asc:promc:examples>



Coming back to file size reduction



- HepSim data are slimmed to reduce disk space

(status=1 && pT>0.4 GeV) or # final states
(PID=5 || PID=6) or # b or top
(PID>22 && PID<38) or # exotics and Higgs
(PID>10 && PID<17) or # leptons/neutrinos

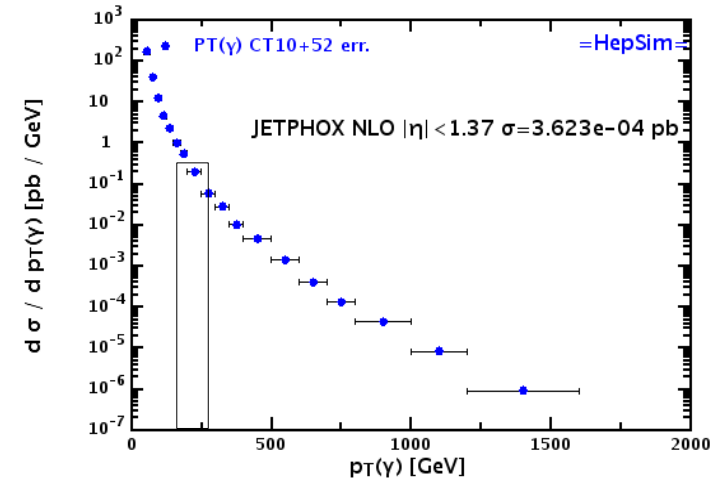
- For 90% cases this is OK, unless you need to find origin of some particles (like taus!)
- Typical file size ~ 20 MB for 5k events:
 - Example for 5000 events (4Kb/event):
 - wget http://atlaswww.hep.anl.gov/asc/promc/download/browser_promc.jar
 - java -jar browser_promc.jar
http://mc.hep.anl.gov/asc/hepsim/events/pp/100tev/ttbar_mg5/mg5_ttbar_100tev_001.promc



Data from NLO programs (MCFM, Jetphox, NLOjet+)

- Data from NLO/NNLO etc. are kept as ProMC “ntuples”
 - 4-momenta (~2-3 particles)
 - all weights for systematics (40-50 floats)
- Keep central weight as **double** and deviations from the central value in form of **varint** (int!)
 - $[(1-\text{PDF}(i)/\text{PDF}(0)) * 1000]$
 - effective varint compression: 50k events ~10 MB
- **Data creation time on BlueGene/Q is ~1-2K CPU/h**
- **Typical data output ~ a few GB**
- **Processing time on a desktop <1h**

Current challenges:



- One p_T bin \rightarrow 10h
- 20 bins \rightarrow 200h for all bins
- Few PDFs, scale variations \rightarrow 2000h

HepSim for 100 TeV

<https://atlaswww.hep.anl.gov/asc/hepsim/>

=HepSim= reference HEP simulation samples

RefHepSim is a repository with reference Monte Carlo events (LO+PS, NLO, etc) for HEP experiments. Events are stored in the ProMC format. RefHepSim can be used to browser separate events, look at cross sections, reconstruct any distribution or use for fast detector simulations as described in the [HepSim manual](#). In order to download a folder with all files, right click on the directory below and select "Copy link location". Then use this command to copy all files: `wget -r -l1 -H -t1 -nd -N -np -A promc -E [URL]`, substituting [URL] with the correct directory name.

pp collisions	8 TeV, 13 TeV, 14 TeV, 100 TeV
e+e- collisions	500 GeV

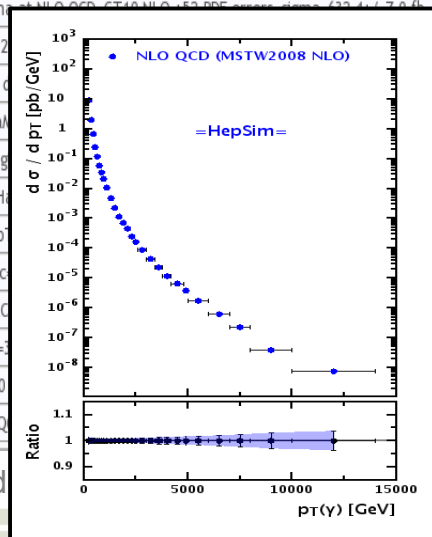
Send comments to: Sergei Chekanov (ANL) chekanov@anl.gov.

events/pp/100tev

16 directories, 0 files (51.44 GiB total)

Search files Case sensitive Current directory only

Nr	Directory/File name	Description		Size	Last Modified
1	gamma_jetphox/	JETPHOX 1.3.2 : 10M events, inclusive gamma at NLO QCD for pT>1 TeV. CT10 NLO +52 PDF errors	plot data script.py	1.24 GiB	2014-02-18 21:08
2	gamma_jetphox_ptbins/	JETPHOX 1.3.2 : 10M events, inclusive gamma at NLO QCD for pT>200 GeV. MSTW2008 NLO +41 sets. Binned in pT	plot data script.py	6.28 GiB	2014-03-01 18:57
3	gamma_mcfm/	MCFM 6.7 : 26M events, inclusive gamma at NLO QCD for pT>1 TeV. CT10 NLO +52 PDF errors	plot data script.py	6.19 GiB	2014-02-11 20:28
4	higgsjet_gamgam_mcfm/	MCFM 6.7 : 26M events, Higgs(->gamma+gamma)+jet at NLO QCD. CT10 NLO +52 PDF errors. sigma=672.3+/-7.0 fb	plot data script.py	1.47 GiB	2014-02-14 15:40
5	higgs_gamgam_mcfm/	MCFM 6.7 : 26M events, Higgs->gamma+gamma at NLO QCD. CT10 NLO +52 PDF errors. sigma=32.4+/-7.0 fb	plot data script.py	2.07 GiB	2014-02-14 15:41
6	higgs_pythia8/	PYTHIA8 : 10,000 events, gg2Httbar and qqbar2Httbar	plot data script.py	852.20 MiB	2014-03-01 07:26
7	higgs_ttbar_mcfm/	MCFM 6.7 : Higgs+ttbar. 20,000x512 events. No QCD	plot data script.py	2.34 GiB	2014-02-14 15:42
8	higgs_ttbar_mg5/	MadGraph5 : p p > h t t- [QCD], 100k events, aMC	plot script.py	477.77 MiB	2014-02-08 12:00
9	jets_nlojetpp/	NLOJET++ : Incl. antiKT4 jets at NLO QCD. pTg	plot data script.py	1.74 GiB	2014-02-25 13:16
10	qcd_pythia8/	PYTHIA8 : 100,000 events. All QCD processes. H	plot script.py	716.21 MiB	2014-02-13 16:04
11	qcd_pythia8_full/	PYTHIA8 : 400,000 events. All QCD processes. p		8.71 GiB	2014-03-10 09:17
12	ttbar_mcfm/	MCFM 6.7 : 26M events, ttbar at NLO QCD (proc	plot data script.py	5.57 GiB	2014-02-13 08:29
13	ttbar_mg5/	MadGraph5 : p p > t t- [QCD], 100k events, aMC	plot data script.py	365.49 MiB	2014-03-01 07:53
14	ttbar_pythia8_full/	PYTHIA8 : 400,000 events. ttbar processes. pT=3			09:19
15	wprime10000_pythia8/	PYTHIA8 : 50,000 events. Wprime to ttbar. M=10			20:32
16	zboson_ee_mcfm/	MCFM 6.7 : 20M events, Zboson->e+e at NLO Q			15:41

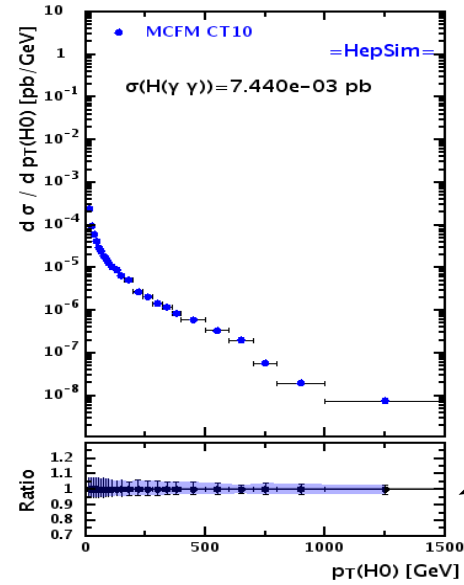
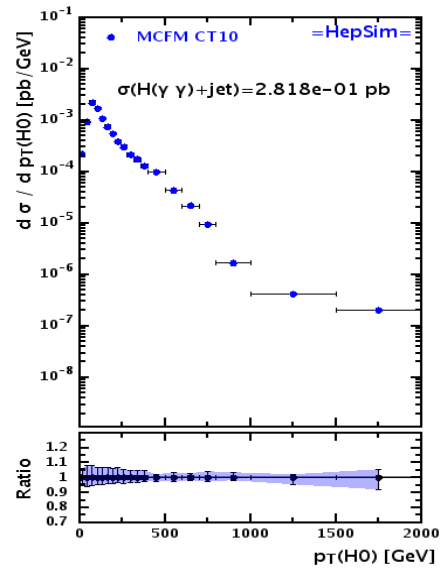
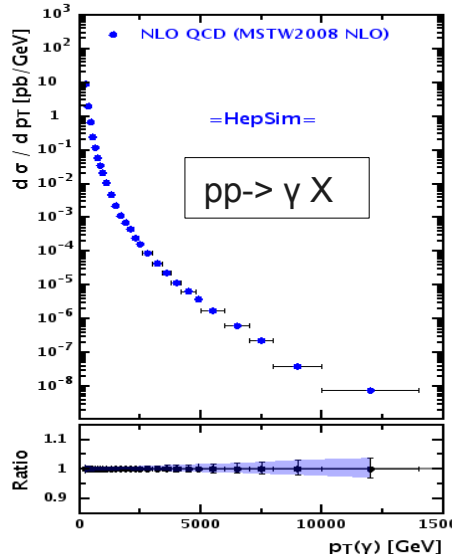


Truth levels
8, 13, 14, 100 TeV
MC LO, NLO+matched
showers, NLO (MCFM)

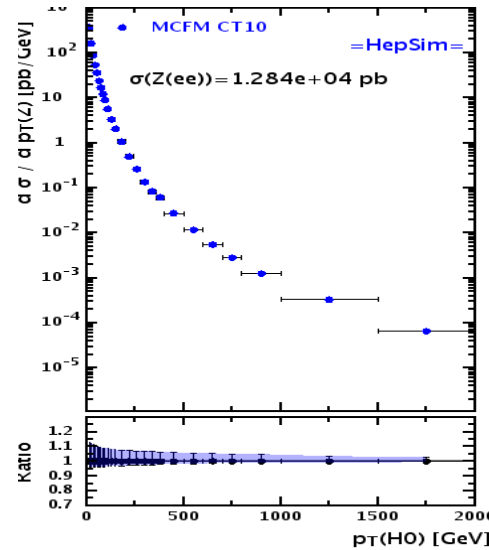
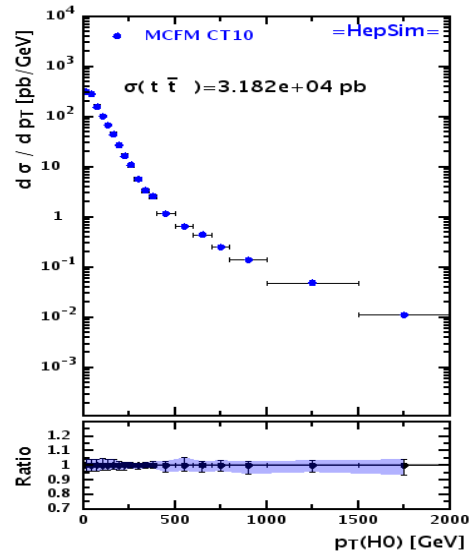
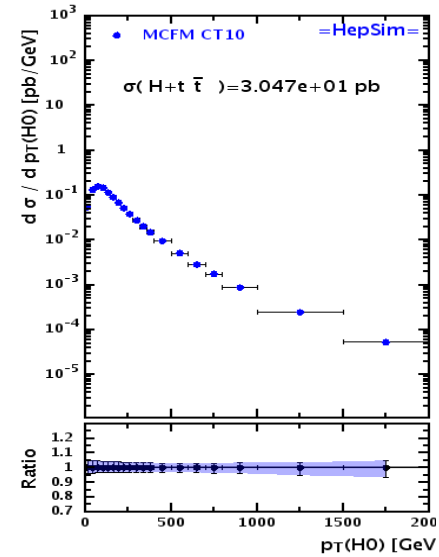
Monte Carlo for a 100 TeV pp collision



Examples of differential cross sections for 100 TeV

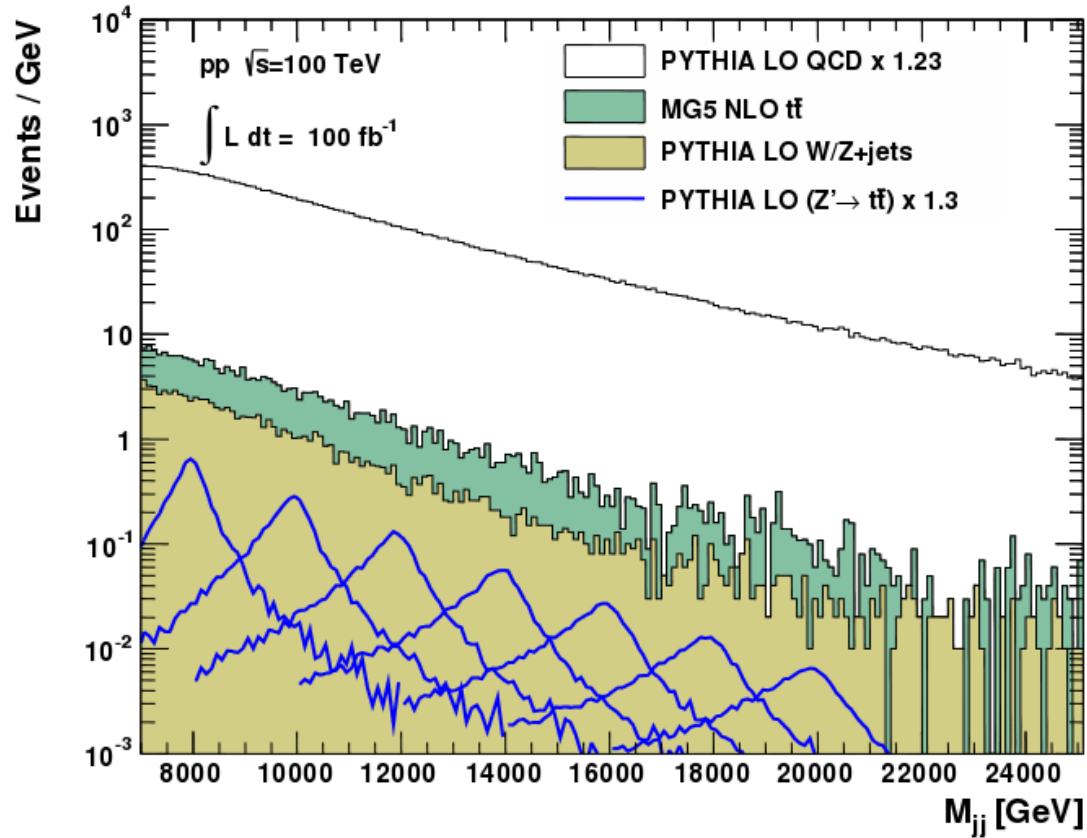


$$\frac{\sqrt{\sum_{i=1}^N (\sigma_i - \sigma_0)^2}}{\sigma_0}$$



PDF uncertainties are within 11% for all studied processes

Realistic plot for $Z' \rightarrow t\bar{t}$ using HepSim (gen-level)



Summary

- Try to use MC events from HepSim
<https://atlaswww.hep.anl.gov/hepsim/>
- 140 samples, 19757 files, 190000000 events (approx.)
- ~30 samples for pp collisions at 100 TeV
- Use DelphesProMC command to create Delphes outputs on-the-fly
 - Delphes processing time \ll download time!