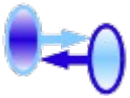
Argonne NATIONAL LABORATORY



HepSim

a repository with theoretical predictions for HEP community

S. Chekanov (ANL)

HEP Software Foundation Workshop,

20-22 January 2015

SLAC

Public Monte Carlo event samples

- **Studies of future colliders are community driven (i.e. Snowmass, FCC, etc.)**
  - theorists & experimentalists from different experiments, etc.
- **Common simulated samples are often required by current experiments**
- **We need:**
  - public access to Monte Carlo simulation samples (CPU intensive to generate!)
  - to ensure long-term availability & preservation of predictions
  - access to common data-analysis tools for data access & analysis
- **Similar databases exist in other areas (star&galaxy catalogs, NOMAD, Tycho..)**
  - “data catalogs” are more common in astronomy (diverse experiments!)
  - HEP has numerous Monte Carlo simulations (NLO,NNLO,NLO+PS)
- **Storing predictions in ``ntuples'' makes sense if:**

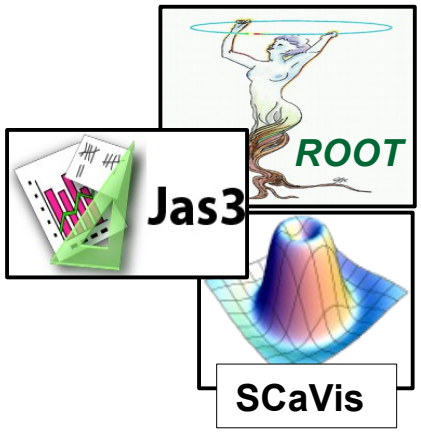
$$\frac{\text{time to download \& analyse on commodity computer}}{\text{CPU*h needed to create the prediction}} \equiv \varepsilon << 1$$

$\varepsilon \sim 0.01\text{-}1$ for LO MC
$\varepsilon << 0.01$ for NLO etc.

Technology choices

- **Public http//**
  - no authentication

- **Highly-compressed data format based on variable-byte encoding**
  - based on **ProMC** & **Google's Protocol Buffers**
    arXiv:1311.1229, Com. Phys. Com 185 (2014), 2629
  - data streaming to the clients (similar to video streaming)



- **Support for major programming languages & OS**
  - C++, CPython + **ROOT/PyROOT**
  - Java, Jython, Groovy, JRuby, BeanShell + **Jas, SCaVis**

- **Java as the primary language for the toolkit & online programs**
  - # 1 OO language (tiobe.com)
  - Low maintenance
  - Multi-platform

HepSim project

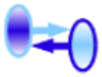
HepSim: http://atlaswww.hep.anl.gov/hepsim/

History:
- 2013: Designed for community studies at Snowmass 2013 (Top/Higgs group)
- 2014: New front-end, additional back-ends; software toolkit, samples for FCC studies

- **Current status**:
  - A suite of Monte Carlo simulations: LO+PS, NLO, NLO+PS using unified file format
  - Several storage back-ends (ANL, UChicago Atlas connect + XRootD)
  - Large fraction of NLO simulations are done on BlueGene/Q (Mira)
  - High compression using variable-byte encoding & multiplatform
  - Data “streaming” over the network (similar to video streaming)
  - Analysis tools do not require complex installation
  - Analysis support for Java, Jython, Groovy, (J)Ruby, C++/ROOT, CPython

- **HepSim can be used for fast & full detector simulation, calculations of LO/NLO cross sections, kinematic distributions etc.**
- **Currently used for FCC community studies + several ATLAS analyses**

HepSim front-end: http://atlaswww.hep.anl.gov/hepsim/



Requesting events | Help | Login

Show all

p →← p
- 7 TeV
- 8 TeV
- 13 TeV
- 14 TeV
- 33 TeV
- 100 TeV

e →← e
- 500 GeV

e →← p
- 920 GeV

HepSim

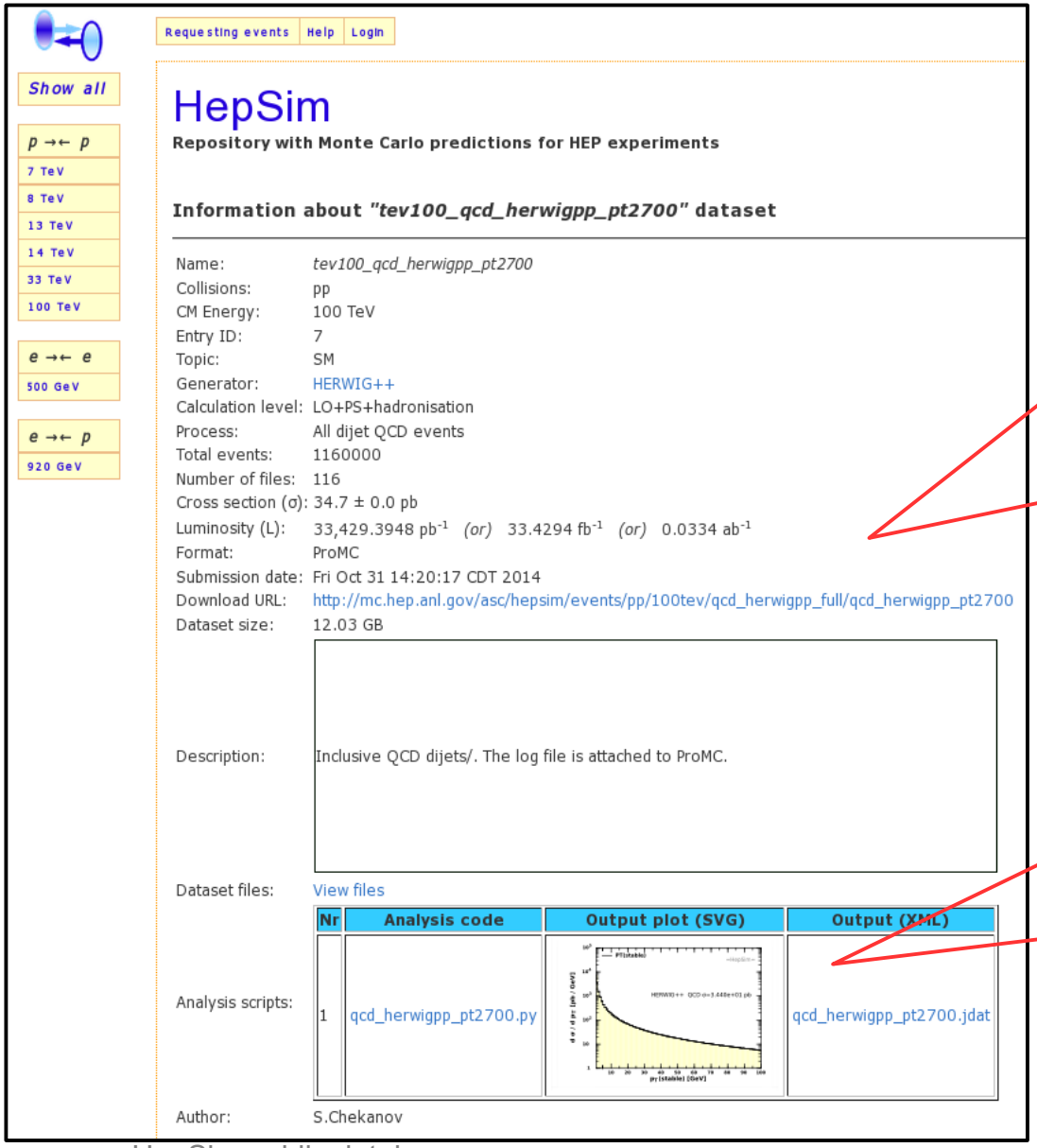
Repository with Monte Carlo predictions for HEP experiments

Show 25 entries Previous 1 2 3 4 Next Search:

Id	→←	E [TeV]	Name	Generator	Process	Topic	Info	L [fb^{-1}]	Link
1	pp	100	tev100_higgs_pythia8	PYTHIA8	gg2Httbar and qqbar2Httbar	Higgs	Info	1.77E+01	URL
2	pp	100	tev100_higgs_ttbar_mg5	MADGRAPH/HW6	Higgs+ttbar (NLO+PS)	Higgs	Info	3.13E+00	URL
5	pp	8	tev8_ww_excl_fpmc	FPMC	Exclusive Higgs	Higgs	Info	1.14E+05	URL
6	pp	8	tev8_gamma_herwigpp	HERWIG++	Direct photons	SM	Info	1.21E+03	URL
7	pp	100	tev100_qcd_herwigpp_pt2700	HERWIG++	All dijet QCD events	SM	Info	3.34E+01	URL
10	pp	100	tev100_kkgluon_ttbar_pythia8	PYTHIA8	KKgluon to ttbar M=1-20 TeV	Exotic	Info	-	URL
11	pp	100	tev100_qcd_pythia8_pt300	PYTHIA8	All dijet QCD events	SM	Info	3.01E-04	URL
12	pp	100	tev100_qcd_pythia8_pt900	PYTHIA8	All dijet QCD events	SM	Info	3.12E-02	URL
13	pp	100	tev100_qcd_pythia8_pt2700	PYTHIA8	All dijet QCD events	SM	Info	1.20E+04	URL
14	pp	100	tev100_qcd_pythia8_pt8000	PYTHIA8	All dijet QCD events	SM	Info	3.37E+03	URL
15	pp	100	tev100_ttbar_mg5	MADGRAPH/HW6	p p > t t~ [QCD] (ttbar at NLO)	Top	Info	3.39E-03	URL
16	pp	100	tev100_ttbar_pt2500_mg5_lo	MADGRAPH/HW6	p p > t t~ (ttbar at LO)	Top	Info	5.00E+02	URL

HepSim public database

Entry metadata: http://atlaswww.hep.anl.gov/hepsim/



HepSim public database

URL for download or data streaming

ProMC format with "variable-byte" encoding based on Google's protocol buffers

40% smaller than fixed-bytes in ROOT

Shows a typical validation distribution created using Jython script.

Also supports Java, Groovy, (J)Ruby, CPython and C++

The manual explains how to download or stream events using client-side Java analysis tool

Available simulations

- 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 -> γγ
- 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)

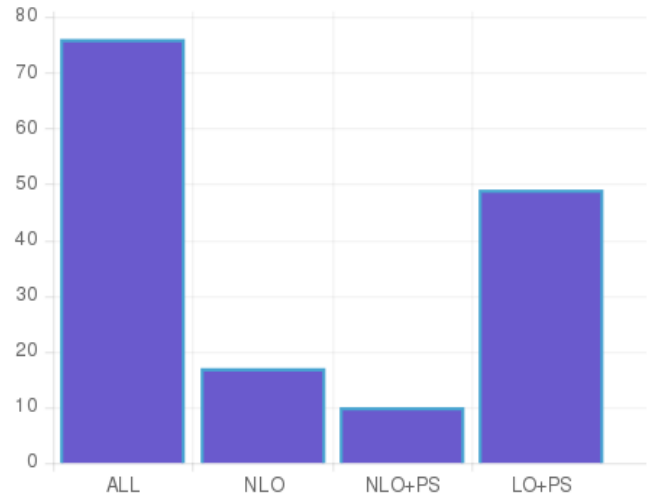
~40% samples generated on BlueGene/Q (Mira) (Jetphox, MCFM)
~50% HEP-ANL (mainly Madgraph)
~10% USATLAS "Connect" during deployment testing (large Pythia samples)

HepSim statistics

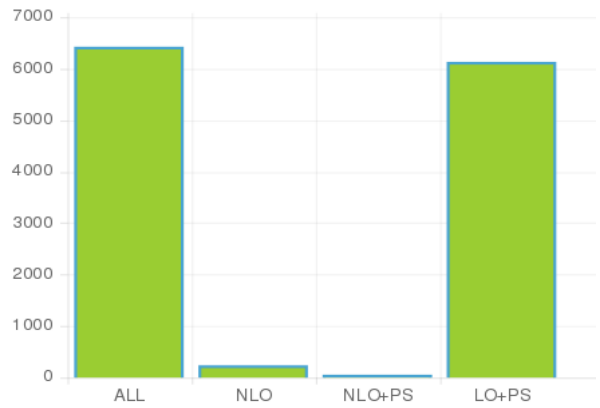
Statistics of HepSim

Total event samples	**76**
NLO samples	17
NLO+PS samples	10
LO (+PS) samples	49
Total number of events	**1203200000**
NLO events	583000000
NLO+PS events	10655000
LO (+PS) events	609545000
Total size (GB)	**6426.186**
NLO size (GB)	238.06
NLO+PS size (GB)	55.95
LO (+PS) size (GB)	6132.176
Total number of files	**281514**

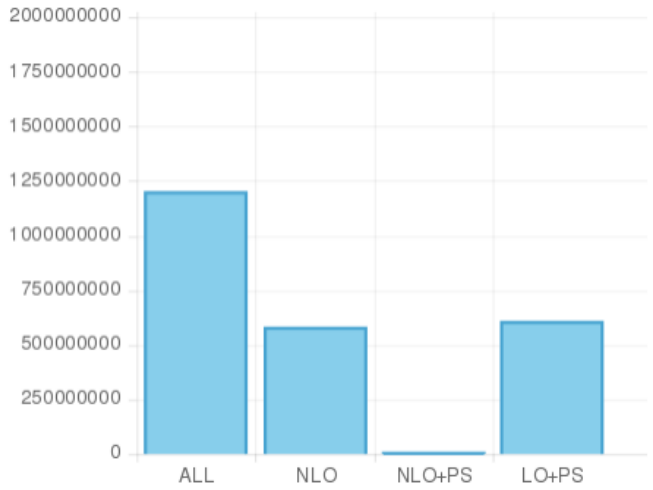
Nr of data samples



Size of HepSim datasets (GB)



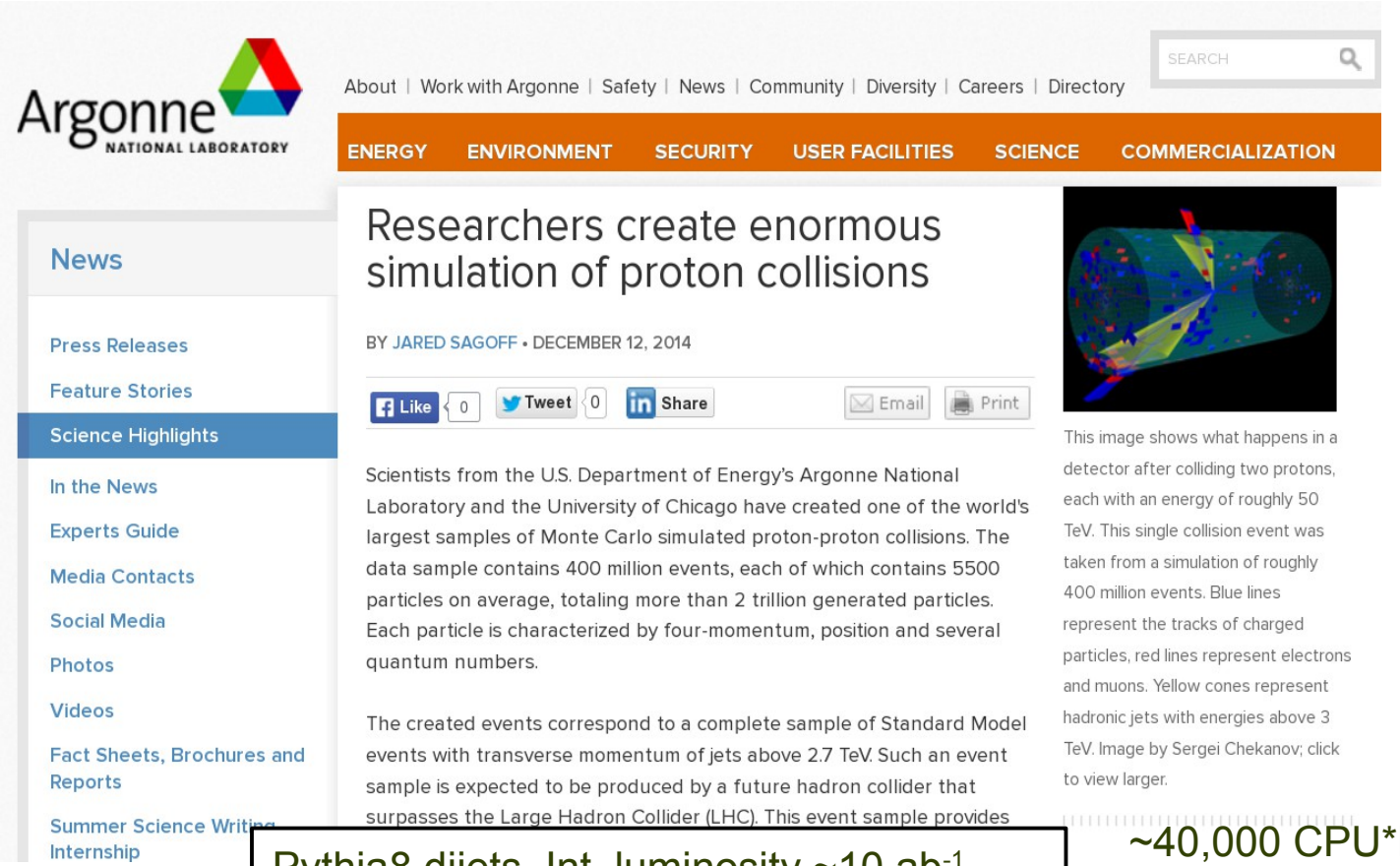
Nr of simulated events



HepSim public database

World's largest public MC sample hosted by HepSim

http://www.anl.gov/articles/researchers-create-enormous-simulation-proton-collisions



Pythia8 dijets. Int. luminosity ~10 ab^{-1}
0.4 billion pp events at 100 TeV

~40,000 CPU*h to create
~ 2 days for download &
analysis on 16 cores $\varepsilon \sim 0.01$

HepSim public database

Backup

HepSim structure

- **(1) Web front-end based on a SQL database with user login**
  - Describes samples
  - Provide macro to illustrate analysis / data access

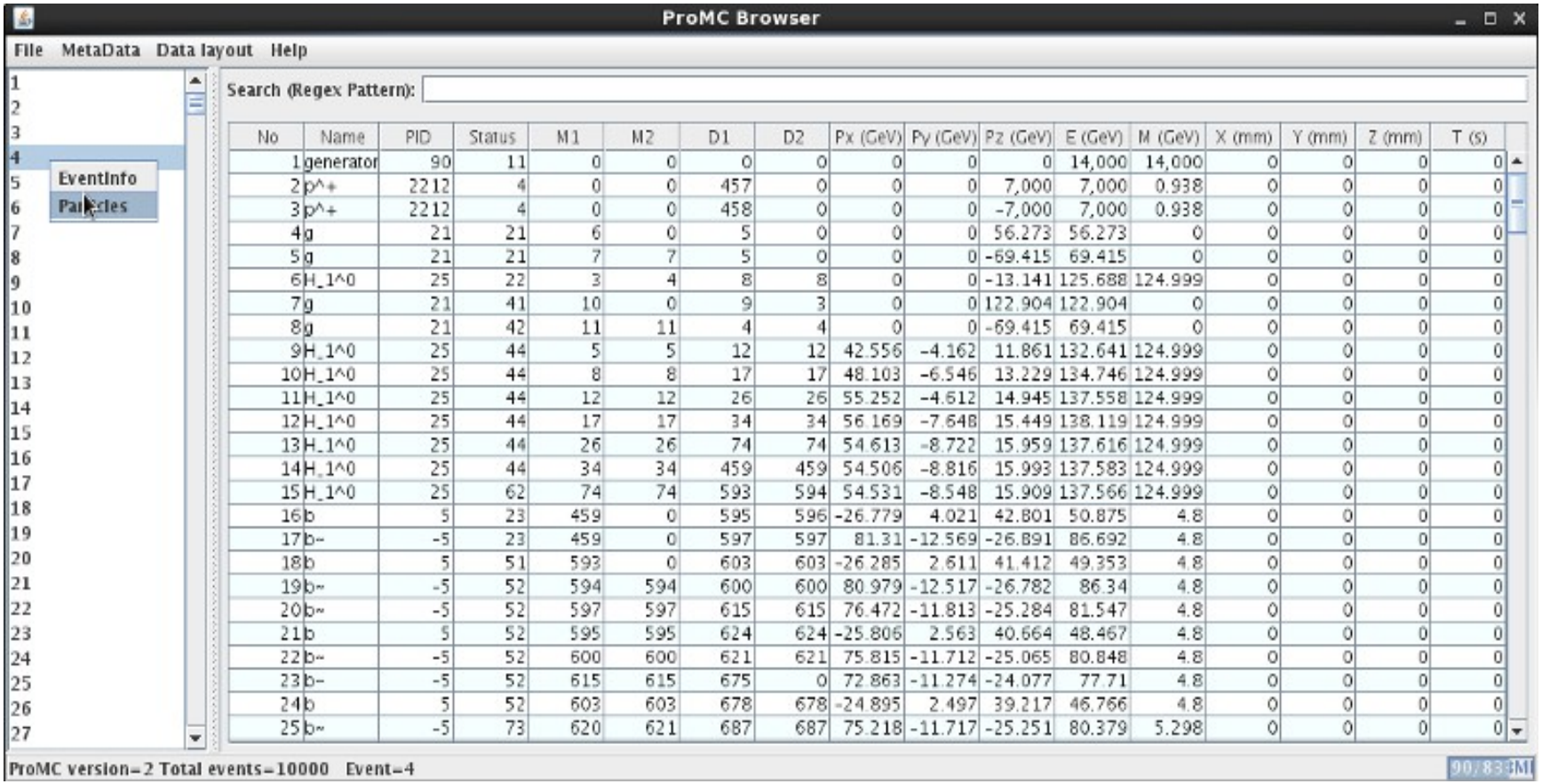
- **(2) Data storage back-end.**
  - Can be distributed on multiple servers & clouds
  - Currently: ANL & UChicago storages

- **(3) Platform-neutral (Java) software toolkit that allows:**
  - list available files in a sample
  - search for a given sample
  - download files in multiple threads
  - validate and view files or separate events
  - GUI for event scan
  - analyzing data using downloaded files (or) streaming data from a server
  - file converters (→ ROOT, HEPMC, LHE, STDHEP)

What can be done with with HepSim?

- Download files in multiple threads
- View event metadata, look at separate events
- Run analysis code using Java, Jython, C++, Python, Groovy, Ruby
- Run analysis code without downloading files (“data streaming”)
  - similar to “video” streaming where user's frontend does calculations
- Data can be analyzed in a GUI and a “batch” mode.
- Data can be processed with fast simulations

Output of “hs-view” (based on http://atlaswww.hep.anl.gov/asc/promc/)



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

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

Processing events over a network using Jython/Java

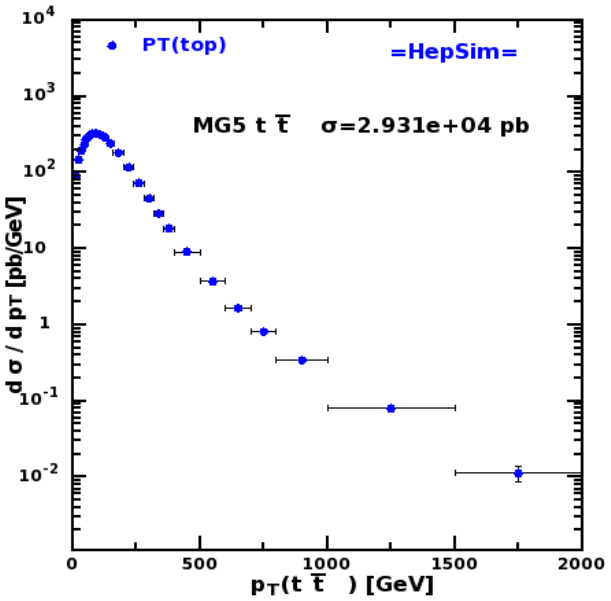
Calculate differential ttbar cross section using Madgraph5 (NLO) using the dataset:
http://atlaswww.hep.anl.gov/hepsim/info.php?item=15

```
wget http://mc.hep.anl.gov/asc/hepsim/events/pp/100tev/ttbar_mg5/macros/ttbar_mg5.py
wget -O scavis.zip http://sourceforge.net/projects/scavis/files/latest/download
unzip scavis.zip
./scavis/scavis_batch.sh ttbar_mg5.py \
http://mc.hep.anl.gov/asc/hepsim/events/pp/100tev/ttbar_mg5/ 100000
```

How does it work?

- get a Python analysis script
- get *scavis* from sourceforge (unzip it)
- run a batch job that streams 100000 online events

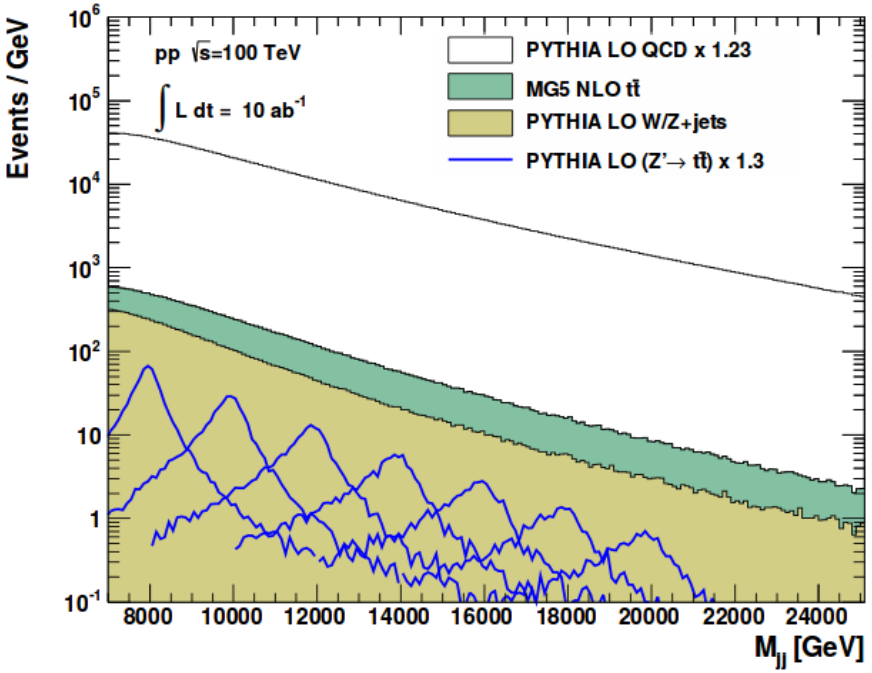
You may download data files first to make the program faster!



Should take ~3 min to create this plot →

High-pT dijets (10 ab^{-1})

Realistic example using C++/ROOT analysis program and HepSim input files (pp collider, 100 TeV)



Sensitivity to new high-mass states decaying to ttbar at a 100 TeV collider

(arXiv:1412.5951, PRD)

NLO calculations as “ntuples”

- Several NLO calculations are available (MCFM, JETPHOX, NLOjet++)
- Data structure is somewhat different compared to full PS Monte Carlo
- “Particle record”: Usually 4-momenta of 3-4 particles per events
- “Event record”:
  - Event weights (double)
  - Deviations from central weights for different PDF eigenvector sets for calculations of PDF uncertainties

$$w_n = \left[1000 \times \left(1 - \frac{PDF(n)}{PDF(0)}\right)\right]$$

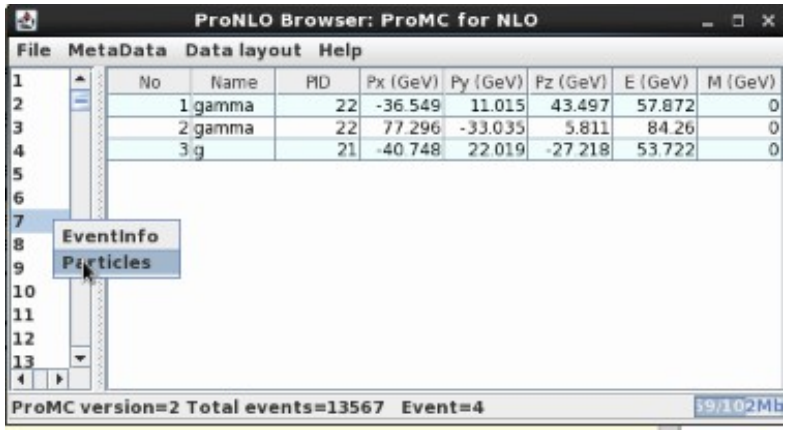
n=1...51 for CT10

Let's look at file structure of MCFM prediction for H(→ γγ)+jet

hs-view-nlo http://mc.hep.anl.gov/asc/hepsim/events/pp/100tev/higgsjet_gamgam_mcfm/hjetgamgam_0000000.promc

NLO calculations as “ntuples”

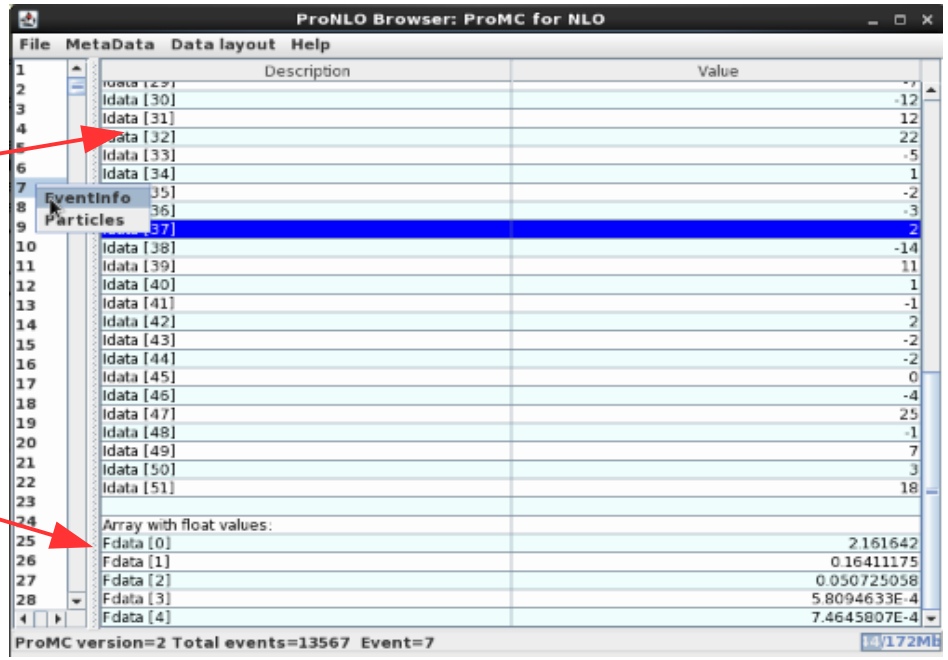
hs-view-nlo http://mc.hep.anl.gov/asc/hepsim/events/pp/100tev/higgsjet_gamgam_mcfm/hjetgamgam_0000000.promc



4-momenta of particles

PDF variations for CT10 (51)

Event weights



Calculating Higgs differential cross section

We will use MCFM sample and Python script from:
https://atlaswww.hep.anl.gov/hepsim/info.php?item=52

It points to MCFM ntuples:
http://mc.hep.anl.gov/asc/hepsim/events/pp/100tev/higgsjet_gamgam_mcfm/

Copy and run these commands:

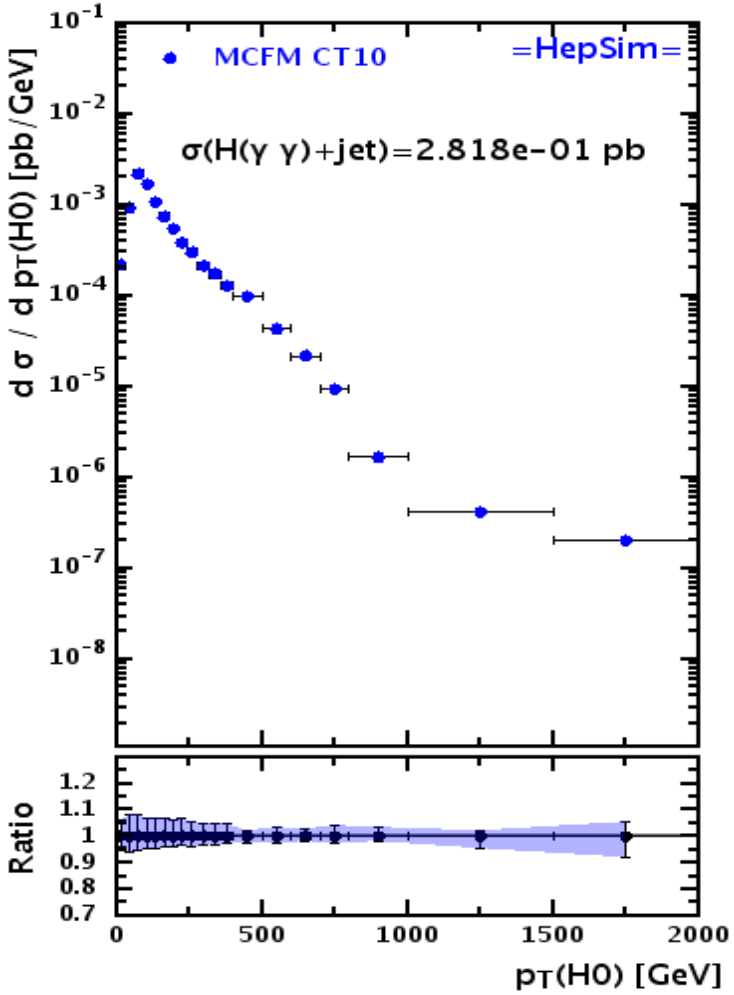
```
mkdir Higgs; cd Higgs;
wget http://mc.hep.anl.gov/asc/hepsim/events/pp/100tev/higgsjet_gamgam_mcfm/macros/higgsjet_gamgam_mcfm.py
wget -O scavis.zip http://sourceforge.net/projects/scavis/files/latest/download
unzip scavis.zip
wget http://atlaswww.hep.anl.gov/asc/promc/download/browser_promc.jar -O ./scavis/lib/physics/browser_promc.jar
./scavis/scavis_batch.sh higgsjet_gamgam_mcfm.py http://mc.hep.anl.gov/asc/hepsim/events/pp/100tev/higgsjet_gamgam_mcfm/
```

How does it work?
- copied analysis script "*higgs_gamgam_mcfm.py*"
- installed *scavis* to process the Python script
- updated ProMC library if it is too old
- run Python using online files and create cross section

Note:
it works faster if you download all files fist and then pass the directory with files, not http!

Result: Calculating Higgs differential cross section

Data from: https://atlaswww.hep.anl.gov/hepsim/info.php?item=52



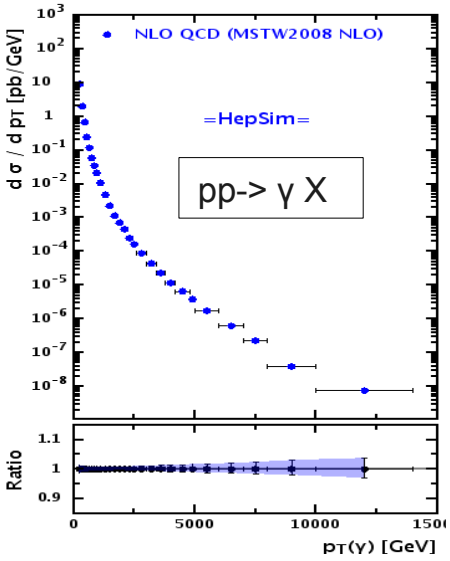
This calculation takes ~5-10 min

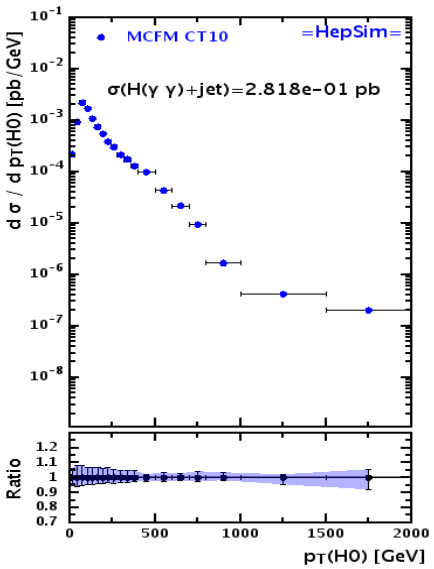
~1000 CPU*h to generate samples on MIRA (ANL)

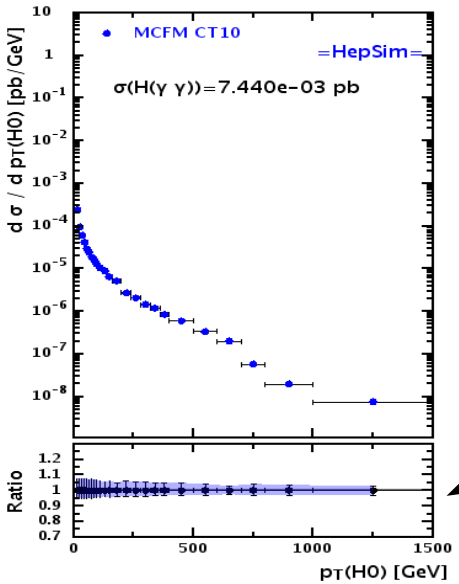
Using Root/C++ is also possible
(but takes more time to explain!)

PDF uncertainties

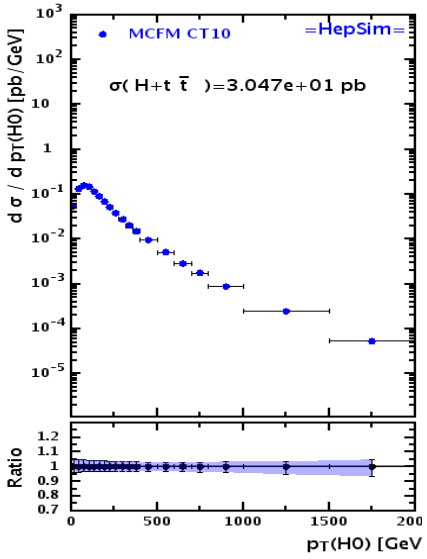
Examples of differential cross sections for 100 TeV

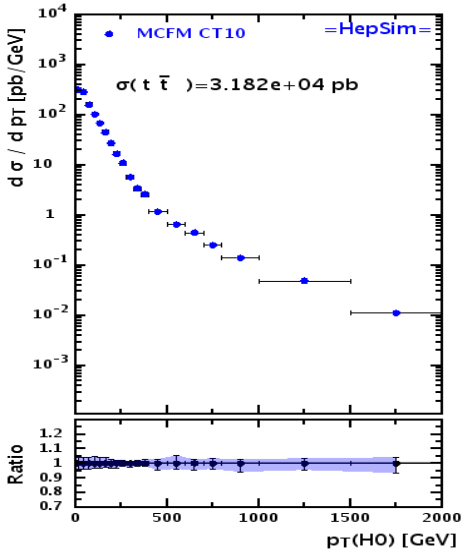


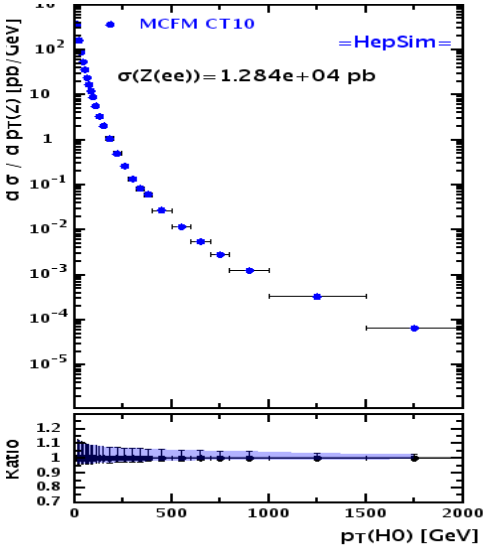




$$\frac{\sqrt{\sum_{i=1}^{N}(\sigma_i - \sigma_0)^2}}{\sigma_0}$$







PDF uncertainties are < 15% for all studied processes

HepSim public database