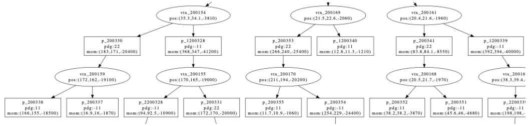
http://lcgapp.cern.ch/project/simu/HepMC



Open Event Generation

Reinterpretation Forum Workshop

26 February 2025, CERN

Rakhi Mahbubani (Rudjer Boskovic Institute), Zach Marshall (LBNL)
With thanks to Giovanni Guerrieri (CERN)

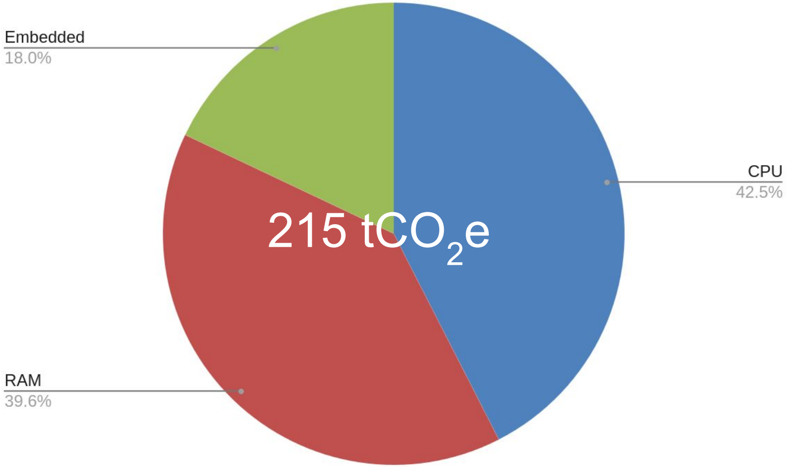
BERKELEY LAB

Why Open Event Generation?

- Theorists and phenomenologists generate their own MC simulation when needed
- The LHC experiments put enormous effort and CPU into running event generation
- Why should we experimentalists help the theorists out?
  - It’s great for scientific **transparency** (what *we* see is what you get)
  - It would support the community — this would be a great service
  - It would avoid **lots** of duplicate and wasted CPU — good for the environment
  - It would avoid some duplicated effort (some folks wouldn’t have to learn how to properly configure various event generators on their own)
  - It would encourage phenomenologists and theorists to look at our MC simulation, which could have a number of ancillary benefits: experts coming to help with configurations, identifying issues in our configurations, improved documentation, maybe validation help?
  - (Of course, this would also be a means of sharing our event generation between e.g. ATLAS and CMS, but that is an interesting-to-explore side-effect here)

Evtgen emissions (preliminary)

ATLAS annual EvtGen emissions ~ 60M CPU-h



Using global average carbon intensity for electricity. Assumptions: Data storage neglected; PUE=1; CPU usage factor=1; Dell server 2x32 core, 512 GB RAM.

Total EvtGen footprint (ATLAS+CMS+pheno)

~220 return transatlantic flights

Assumptions: CMS evtgen emissions same order as ATLAS; 150 pheno papers annually, each using 10k CPU hrs on 8-core MacBook Pro; CPU usage factor = 1.

Total will scale with lumi and need for increasing precision

What Would We Release?

- We would release our event generation via the CERN Open Data Portal
- Experiments release data on the portal already, just need to agree on technicalities
  - Need **robust metadata** for the samples (xsec, filter efficiency, k-factor, generators…)
- Rough estimate based on existing ATLAS MC for Run 2 and Run 3: 4800 samples
  - Could include some BSM signals if we want, but this is for SM backgrounds primarily
  - Could try to cut this down, but most of the disk space is the big baseline samples
- Could try to push reasonably regular updates (if effort and resources allow)
  - Depending on scenario and space, could keep old samples or delete them
  - Deleting them would make many people sad, but we have to make hard choices sometimes

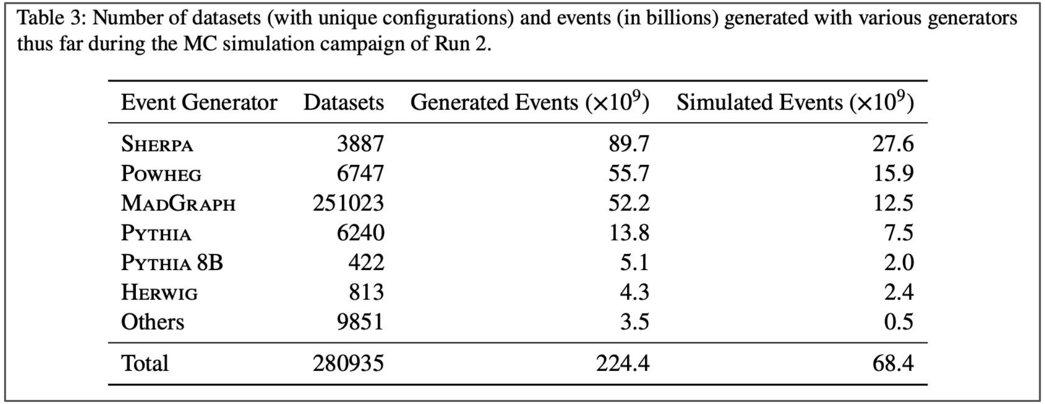


Table 3: Number of datasets (with unique configurations) and events (in billions) generated with various generators thus far during the MC simulation campaign of Run 2.

Event Generator	Datasets	Generated Events ($\times 10^9$)	Simulated Events ($\times 10^9$)
Sherpa	3887	89.7	27.6
Powheg	6747	55.7	15.9
MadGraph	251023	52.2	12.5
Pythia	6240	13.8	7.5
Pythia 8B	422	5.1	2.0
Herwig	813	4.3	2.4
Others	9851	3.5	0.5
Total	280935	224.4	68.4

Totals in Run 2-like configuration. Some samples are obsolete now (newer configurations exist)

When and how much should we release?

- **ATLAS has agreed in principle** to a **first release** of our event generator output
  - min(what we've generated, 2*luminosity) → 300 TB, 7.6B events
  - Our high-stats Baseline samples are **LARGE** per-event because of event weights
  - We are **not** committing *today* to regular updates — it depends on response, effort, etc
  - We will ask for citation and acknowledgement when you use these samples
  - This should come before summer
- CERN IT have indicated that they are willing to support this open data storage
  - For much more we should have a broader discussion to see whether other experiments want something similar and what the expectations for the space are
  - With regular updates, we could reach 1 PB in a few years. Needs to be watched.
- What sort of data volume would be of interest to the community? Which samples?
  - The Open Data Portal has helpers, so you would not need to swallow an entire sample
  - We have a *ton* of signal models… probably these are not a priority? Specific ones?

Format Technicalities

- Theorists/Phenomenologists seem to want compressed HEPMC (for now) — right?
  - Primarily because that interfaces well with existing tools
  - Could move to an alternative format if the tools support that format; we could push the community towards something ROOT-based(?) if desired
  - ATLAS and CMS both use ROOT-based representations of HEPMC (v3 for ATLAS)
- Quick size comparison test with an Run 3 (13.6 TeV C.o.M) ttbar file (10k events):
  - EVNT (ATLAS internal format): 58.2 kB/event (~2x variations depending on ROOT settings)
  - Compressed HEPMC: 54.5 kB/event (variations depending on compression settings)
  - Uncompressed HEPMC: 210 kB/event
  - TRUTH0 derivation (easy ROOT-readable EVNT): 35.9 kB/event
  - TRUTH1 derivation (TRUTH0+pre-built simplified collections like ‘jets’): 40.8 kB/event
- To do this we will convert our EVNT to HEPMC
  - O(4k) CPU-core-days (not much by modern standards)
  - Being prepared now, likely to stage the release so that folks can test / check before everything is converted

Long term thoughts / vision

- Could try to develop some notebook-like examples for running on HEPMC
  - We have these sorts of things for ATLAS Open Data already
  - HighTEA looks similar to some of our open data setups, at least in principle
- Could discuss whether ATLAS and CMS could share event generation
  - Raised last month in the Dark Showers workshop
  - Raised some years ago in the HSF Event Generation WG
  - Maybe now's the time for another round of that discussion… NB sharing does **not** mean that we don't allow private / internal / custom samples, etc or even force the same nominal
- If we develop some custom / common (e.g. ROOT-based) format, we should work to integrate it with existing tools
  - Delphes, PGS, Rivet, others? Need a (complete?) list if we embark on this path
  - Could consider a document recommending a simulation (configuration) and pointing out some of the known limitations (working document of improvements?)

Some discussion questions

- What would you want most?
  - And how much of that?
- Would you be willing to contribute to make it happen?
- Any other considerations before we move forward?

This is a also **test of the community**:

- Can we work together to support common samples?
- When issues are found, are they reported back? Do people help correct issues?
- If samples are insufficient for some reason, is that reported back? Do we make the samples better together?

Are we ready?