CERN-PH-TH/2013-105
DESY 13-104
LU-TP 13-23

MCPLOTS: a particle physics resource based on volunteer computing

A. Karneyeu[1], L. Mijovic[2,3], S. Prestel[2,4], P. Z. Skands[5]

[1]: Institute for Nuclear Research, Moscow, Russia
[2]: Deutsches Elektronen-Synchrotron, DESY, 22603 Hamburg, Germany
[3]: Irfu/SPP, CEA-Saclay, bat 141, F-91191 Gif-sur-Yvette Cedex, France
[4]: Dept. of Astronomy and Theoretical Physics, Lund University, Sweden
[5]: European Organization for Nuclear Research, CERN, CH-1211, Geneva 23, Switzerland

Abstract

The **mcplots.cern.ch** web site (MCPLOTS) provides a simple online repository of plots made with high-energy-physics event generators, comparing them to a wide variety of experimental data. The repository is based on the HEPDATA online database of experimental results and on the RIVET Monte Carlo analysis tool. The repository is continually updated and relies on computing power donated by volunteers, via the LHC@HOME 2.0 platform.

1 Introduction

Computer simulations of high-energy interactions are used to provide an explicit theoretical reference for a wide range of particle-physics measurements. In particular, Monte Carlo (MC) event generators [1–3] enable a comparison between theory and experimental data down to the level of individual particles. An exact calculation taking all relevant dynamics into account would require a solution to infinite-order perturbation theory coupled to non-perturbative QCD — a long-standing and unsolved problem. In the absence of such a solution, MC generators apply a divide-and-conquer strategy, factorizing the problem into many simpler pieces, and treating each one at a level of approximation dictated by our understanding of the corresponding parts of the underlying fundamental theory.

A central question, when a disagreement is found between simulation and data, is thus whether the discrepancy is within the intrinsic uncertainty allowed by the inaccuracy of the calculation, or not. This accuracy depends both on the sophistication of the simulation itself, driven by the development and implementation of new theoretical ideas, but it also depends crucially on the available constraints on the free parameters of the model. Using existing data to constrain these is referred to as “tuning”. Useful discussions of tuning can be found, e.g., in [1, 2, 4–10].

Typically, experimental studies include comparisons of specific models and tunes to the data in their publications. Such comparisons are useful both as immediate tests of commonly used models, and to illustrate the current amount of theoretical uncertainty surrounding a particular distribution. They also provide a set of well-defined theoretical reference curves that can be useful as benchmarks for future studies. However, many physics distributions, in particular those that are infrared (IR) sensitive[1] often represent a complicated cocktail of physics effects. The conclusions that can be drawn from comparisons on individual distributions are therefore limited. They also gradually become outdated, as new models and tunes supersede the old ones. In the long term, the real aim is not to study one distribution in detail, for which a simple fit would in principle suffice, but to study the degree of simultaneous agreement or disagreement over many, mutually complementary, distributions. This is also a prerequisite to extend the concept of tuning to more rigorous consistency tests of the underlying physics model, for instance as proposed in [8].

The effort involved in making simultaneous comparisons to large numbers of experimental distributions, and to keep those up to date, previously meant that this kind of exercise was restricted to a small set of people, mostly Monte Carlo authors and dedicated experimental tuning groups. The aim with the **mcplots.cern.ch** (MCPLOTS) web site is to provide a simple browsable repository of such comparisons so that anyone can quickly get an idea of how well a particular model describes various data sets[2]. Simultaneously, we also aim to make all generated data, parameter cards, source codes, experimental references, etc, freely and directly available in as useful forms as possible, for anyone who wishes to reproduce, re-plot, or otherwise make use of the results and tools that we have developed.

The MCPLOTS web site is now at a mature and stable stage. It has been online since Dec 2010 and is nearing a trillion generated events in total (900 billion as of June 2013). This paper is intended to give an overview of what is available on the site, and how to use it. In particular, section 2 contains a brief "user guide" for the site, explaining its features and contents in simple terms, how to navigate through the site, and how to extract plots and information about how they were made from it. As a reference for further additions and updates, and for the benefit of future developers, sections 3 – 6 then describe the more concrete details of the technical structure and implementation of the site, which the ordinary user would not need to be familiar with.

Section 3 describes the architecture of the site, and the thinking behind it. It currently relies on the following basic prerequisites,

- The HEPDATA database [13] of experimental results.
- The RIVET Monte Carlo analysis tool [14] which contains a large code library for comparing the output of MC generators to distributions from HEPDATA. RIVET in turn relies on the HEPMC universal event-record format [15], on the FASTJET package for jet clustering [16, 17], and on the LHAPDF library for parton densities [18, 19].
- Monte Carlo event generators. Currently implemented generators include ALPGEN [20], EPOS [21] HERWIG++ [22], PHOJET [23], PYTHIA 6 [24], PYTHIA 8 [25], SHERPA [26], and VINCIA [27]. Some of these in turn use the Les Houches Event File (LHEF) format [28, 29] to pass parton-level information back and forth.
- The LHC@HOME 2.0 framework for volunteer cloud computing [30–33]. LHC@HOME 2.0 in turn relies on the CERNVM computing environment [34, 35] (a Virtual-Machine environment based on Scientific Linux), and on the COPILOT job submission system [35, 36].

The main benefit of using CERNVM is that this has enabled MCPLOTS to draw on significant computing resources made available by volunteers, via LHC@HOME 2.0. Through the intermediary of

[1] IR sensitive observables change value when adding an infinitely soft particle or when splitting an existing particle into two collinear ones. Such variables have larger sensitivity to non-perturbative physics than IR safe ones, see, e.g., [2, 11]. Note also that we here use the word "IR" as a catchall for both the soft and collinear limits.

[2] Note: this idea was first raised in the now defunct JETWEB project [12]. The CERN Generator Services (GENSER) project also maintains a set of web pages with basic generator validation plots/calculations.

CERNVM, generator authors can concentrate on developing code that is compatible with the Scientific Linux operating system, a fairly standard environment in our field. This code can then be run on essentially any user platform by encapsulating it within CERNVM. The resulting "TEST4THEORY" project [33, 37] was the first application developed for the LHC@HOME 2.0 framework, see [30, 33], and it represents the world's first virtualization-based volunteer cloud.

The basic procedure to include a new measurement on MCPLOTS is, first, to provide the relevant experimental data points to HEPDATA, second to provide a RIVET routine for the corresponding analysis, and lastly to provide a very small additional amount of information to MCPLOTS, essentially specifying the placement of the observable in the MCPLOTS menus and summarizing the cuts applied in a LaTeX string, as e.g. exemplified by the already existing analyses on the site. This is described in more detail in section 4.

To update MCPLOTS with a new version of an existing generator, the first step is to check whether it is already available in the standard CERN Generator Services (GENSER) repository [38], and if not announce the new version to the GENSER team. The MCPLOTS steering scripts should then be updated to run jobs for the new version, as described in section 5.

To add a new generator to MCPLOTS, the first step is to check that it can run within CERNVM. CERNVM provides a standardized Scientific-Linux environment that should be appropriate for most high-energy physics (HEP) applications, including several commonly used auxiliary packages such as the GNU Scientific Library (GSL), the C++ BOOST libraries, and many others. A standalone version of CERNVM can be downloaded from [34] for testing purposes. To the extent that dependencies require additional packages to be installed, these should be communicated to the MCPLOTS and CERNVM development teams. The code should then be provided to the GENSER team for inclusion in the standard CERN generator repository. The complete procedure is described in more detail in section 6.

2 User Guide

In this section, we describe the graphical interface on the MCPLOTS web site and how to navigate through it. Care has been taken to design it so as to make all content accessible through a few clicks, in a hopefully intuitive manner.

2.1 The Main Menu

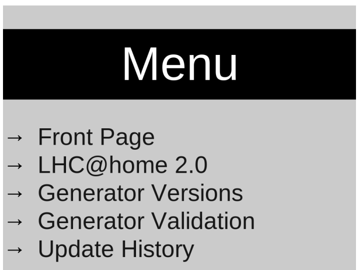


Figure 1: The main MCPLOTS menu.

The main menu, shown in figure 1, is always located at the top left-hand corner of the page. The *Front Page* link is just a "home" button that takes you back to the starting page for MCPLOTS, and the *LHC@home 2.0* one takes you to the external LHC@HOME 2.0 web pages, where you can connect your computer to the volunteer cloud that generates events for MCPLOTS.

The *Generator Versions* link opens a configuration page that allows you to select which generator versions you want to see results for on the site. The default is simply to use the most recently implemented ones, but if your analysis, for instance, uses an older version of a particular generator, you can select among any of the previously implemented versions on the site by choosing that specific version on the *Generator Versions* page. All displayed content on the site will thereafter reflect your choice, as you can verify by checking the explicit version numbers written at the bottom of each plot. You can return to the *Generator Versions* page at any time to modify your choice. After making your choice, click on the *Front Page* button to exit the *Generator Versions* view.

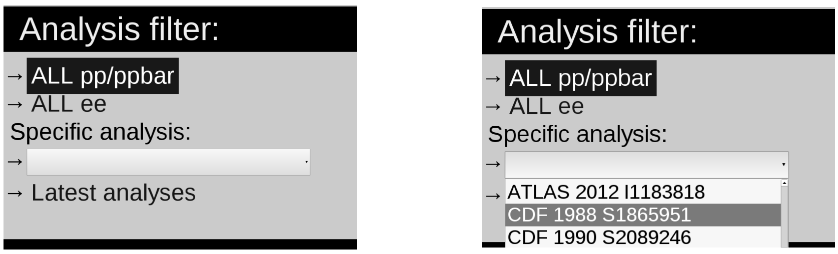
The *Generator Validation* link changes the page layout and content from the *Front Page* one, to one in which different generator versions can be compared both globally, via χ^2 values, and individually on each distribution. This view will be discussed in more detail in section 2.4.

The *Update History* link simply takes you to a page on which you can see what the most recent changes and additions to the site were, and its previous history.

As an experimental social feature, we have added a “like” button to the bottom of the front page, which you can use to express if you are happy with the MCPLOTS site.

2.2 The Analysis Filter

Immediately below the main menu, we have collected a few options to control and organize which analyses you want to see displayed on the site, under the subheading *Analysis Filter*, illustrated in figure 2.



(a) Analysis Filter: Normal View

(b) Analysis Filter: Dropdown Menu

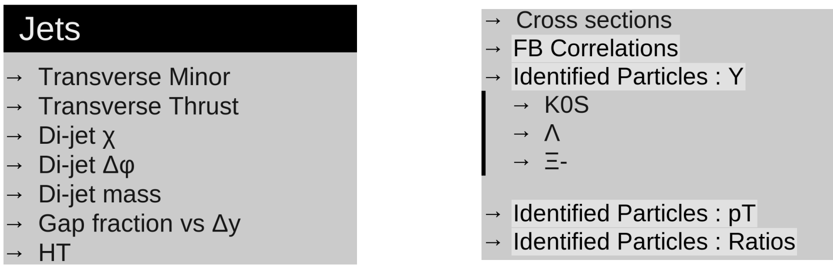
Figure 2: The analysis filter submenu; normal view (left) and after clicking on the *Specific Analysis* dropdown menu (right).

At the time of writing, the main choice you have here is between *ALL pp/ppbar* (for hadron collisions) and *ALL ee* (for fragmentation in electron-positron collisions). The default is *ALL pp/ppbar*, so if you are interested in seeing all hadron-collider analyses, you will not have to make any changes here. Using the *Specific Analysis* dropdown menu, you can also select to see only the plots from one particular RIVET analysis. The latter currently requires that you know the RIVET ID of the analysis you are interested in. The ID is typically formed from the experiment name, the year, and the inSPIRE ID (or SPIRES ID, for older analyses) of the paper containing the original analysis, as illustrated in figure 2b (the numbers beginning with “I” are inSPIRE codes, while ones beginning with “S” are SPIRES ones). You can also find this information in the RIVET user manual [14] and/or on the RIVET web pages.

Finally, if you click on *Latest Analyses*, only those analyses that were added in the last update of the site will be shown. This can be useful to get a quick overview of what is new on the site, for instance to check for new distributions that are relevant to you and that you may not have been able to see on your last visit to the site. More options may of course be added in the future, in particular as the number of observables added to the pp/ppbar set grows.

2.3 Selecting Observables

Below the *Analysis Filter*, the list of processes and observables for the selected set of analyses is shown. This is illustrated in figure 3.



(a) The Process and Observables List

(b) An expanded dropdown menu

Figure 3: Illustrations of the process and observables list; normal view (left) and after clicking on a shaded dropdown menu (right), in this case *Identified Particles: Y*.

Clicking on any blue link below one of the process headers (e.g. below the "Jets" header), figure 3a, or any blue link in the shaded drop-down menus, figure 3b, will open the plot page for that observable in the right-hand part of the page.

At the top of the plot page, figure 4, you can select which generators and tune combinations you want to see on the page. By default, you are shown the results obtained with default settings of the available generators, but the links for each generator give you access to see results for different tune and model variations. Use the *Custom* link to specify your own set of generators and tunes. The available plots for the chosen settings are shown starting with the highest CM energies at the top of the page, and, for each CM energy, cascading from left to right alphabetically.

For many observables, measurements have been made using a variety of different cuts and triggers. These are indicated both above the plots and on the plots themselves, so as to minimize the potential for misinterpretation. In the example of charged-particle multiplicity distributions shown in figure 4, the two first plots that appear are thus ALICE (left), for their INEL>0 cuts [39], and ATLAS (right), using their $N_{\mathrm{ch}} \geq 1$ and $p_T > 2.5\,\mathrm{GeV}$ cuts [40]. We explain how to find the correct references and run cards for each plot and generator below. Note that, for the Monte Carlo runs, the number of events in the smallest sample is shown along the right-hand edge of each plot. I.e., if two generators were used, and the statistics were N_1 and N_2 events, respectively, the value printed is $\min(N_1, N_2)$.

Underneath each plot is shown a ratio pane, showing the same results normalized to the data (or to the first MC curve if there are no data points on the plot). This is illustrated in figure 5. The vertical range of the ratio plot is fixed to $[0.5, 1.5]$, thus larger deviations than this will exceed the boundaries of the ratio plot. The central shaded (yellow) band corresponds to the experimental uncertainty, at the 1σ level, as reported by RIVET. For MC-to-MC comparisons, it represents the statistical uncertainty of the reference MC.

Immediately below the ratio pane are links to download the plot (main plot + ratio pane) in higher resolution and/or in vector graphics formats. Currently, *pdf*, *eps*, and *png* formats are available. For publications or presentations, we strongly recommend using these high-resolution versions rather than the web-optimized ones displayed on the page, to avoid undesirable pixelation effects.

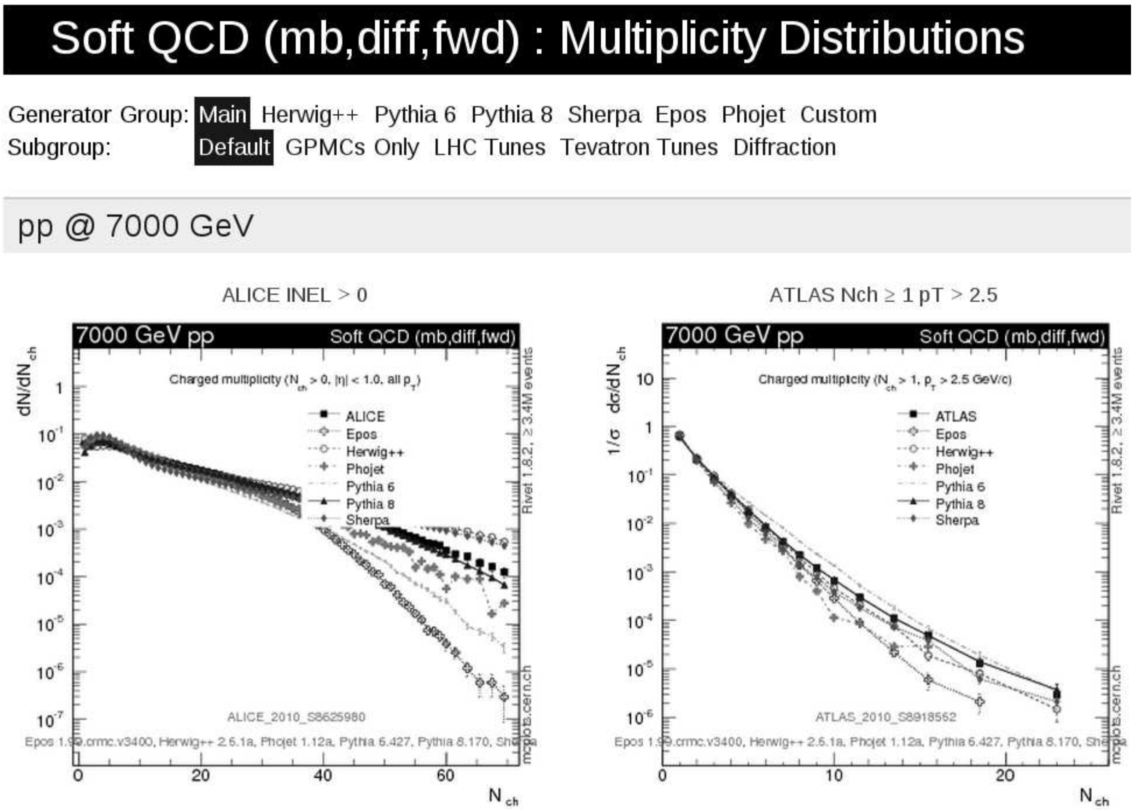
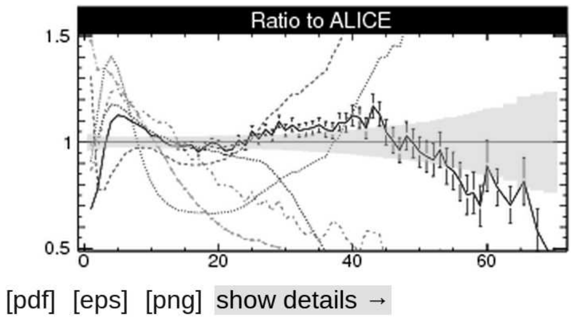
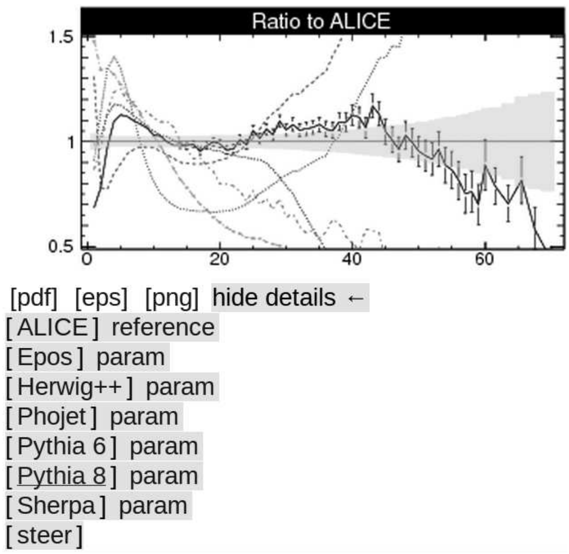


Figure 4: The plot page. The generator and tune group selections are at the top, followed by the available plots for the chosen observable, ordered by CM energy and subordered alphabetically.



(a) The Ratio Pane



(b) After clicking on *show details*

Figure 5: The ratio pane, before (a) and after (b) clicking on the *show details* link. The example corresponds to the left-hand plot shown in figure 4.

Additional information about the plot, such as data tables, references, and run cards, can be accessed by clicking on the *show details* dropdown menu, illustrated in figure 5b. In this example, clicking on *[ALICE]* gives you a text file containing a table of the experimental data points, along with additional information (from RIVET) about the plot. Clicking on *reference* sends you to the inSPIRE page for the experimental paper in which the measurement was presented. Clicking on a generator name will give you a text file containing a table of the results obtained with that generator, together with additional technical information from MCPLOTS (including an additional table which is used by MCPLOTS to combine the results of several different runs). You can use or ignore the additional information in these files as you wish. Clicking on *param* gives you the exact generator run card that was used to make the plot, so that you can see precisely how the results were generated. These cards can also be useful as examples to start generating your own standalone results, or to check that you can reproduce ours. Finally the *steer* link contains the steering card used by the ROOT-based tool that makes the actual plots you see. Though this tool is not yet publicly accessible on the site, just contact us if you would like a copy, e.g. to use it to make your own standalone plots outside of MCPLOTS.

2.4 The Generator Validation View

Clicking on *Generator Validation* (see the main MCPLOTS menu in figure 1) opens the validation view. (You can click on *Front Page* to get back to the default view at any time.) In this view, instead of the *Analysis Filter* you will see a list of event generators beneath the main MCPLOTS menu. You can click on each generator to see a list of the available tunes and model variations for that generator. This is illustrated in figure 6.

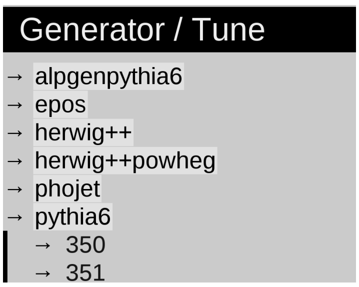
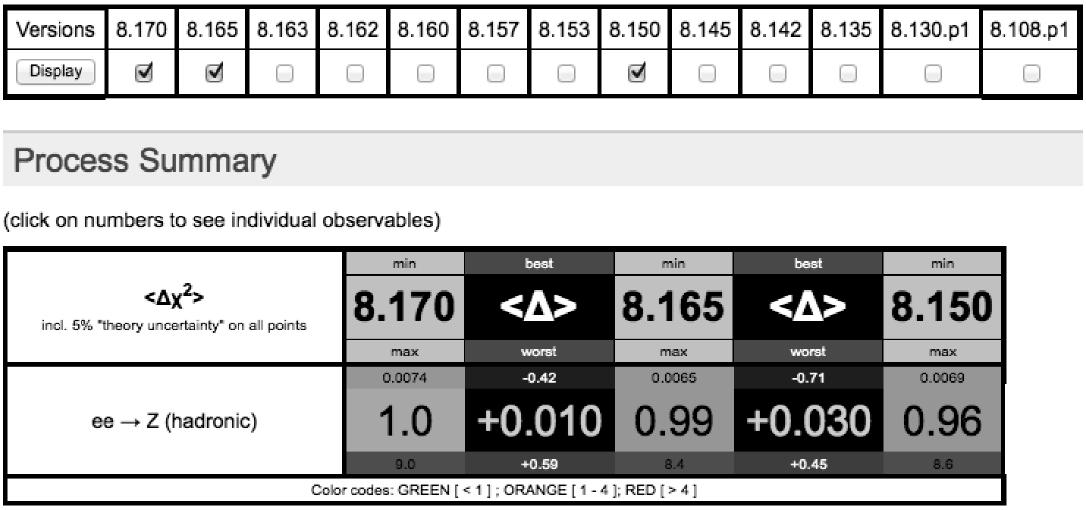


Figure 6: Generator Validation View: excerpt of the generator/tune selection menu.

As an example, clicking on *pythia8* $\rightarrow$ *default* will show the validation view for default settings of the PYTHIA 8 generator, in the right-hand side of the page, illustrated in figure 7. In this view, no plots are shown immediately. Instead you are presented with a table of $\langle\chi^2\rangle$ values, averaged over all measurements within each process category. Note that we use a slightly modified definition of $\langle\chi^2\rangle$, discussed below.

At the top of the page, you can select which versions of the generator you want to include in the table. Click the *Display* button to refresh the table after making modifications.

Below the line labelled *Process Summary*, we show the main table of $\langle\chi^2\rangle$ values for the versions you have selected, as well as the relative changes between successive versions, thus allowing you to look for any significant changes that may have resulted from improvements in the modelling/tuning (reflected by decreasing χ^2 values) or mistunings/bugs (reflected by increasing χ^2 values). The largest and smallest individual $\langle\chi^2\rangle$ values (and changes) in the relevant data set are also shown, in smaller fonts, above and below the average values. To aid the eye, values smaller than 1 are shaded green



Versions	8.170	8.165	8.163	8.162	8.160	8.157	8.153	8.150	8.145	8.142	8.135	8.130.p1	8.108.p1
Display	☑	☑	☐	☐	☐	☐	☐	☑	☐	☐	☐	☐	☐

Process Summary

(click on numbers to see individual observables)

<Δχ²> incl. 5% "theory uncertainty" on all points	min 8.170 max	best <Δ> worst	min 8.165 max	best <Δ> worst	min 8.150 max
ee → Z (hadronic)	0.0074 1.0 9.0	-0.42 +0.010 +0.59	0.0065 0.99 8.4	-0.71 +0.030 +0.45	0.0069 0.96 8.6

Color codes: GREEN [< 1] ; ORANGE [1 - 4]; RED [> 4]

Figure 7: Generator Validation: example showing (an excerpt of) the validation view for default settings of the PYTHIA 8 generator.

(corresponding to less than 1σ average deviation from the data), values between 1 and 4 are shaded orange (corresponding to less than 2σ deviation), and values greater than 4 are shaded red, following the spirit of the Les Houches "tune killing" exercise reported on in [10]. In the example shown in figure 7, the changes are less than a few per cent, indicative of no significant overall change. Bear in mind that the statistical precision of the MC sample plays a role, hence small fluctuations in these numbers are to be expected, depending on the available numbers of generated events for each version.

Note: to make these numbers more physically meaningful, the MC predictions are assigned a flat 5% "theory uncertainty", as a baseline sanity limit for the achievable theoretical accuracy with present-day MC models. A few clear cases of GIGO[3] are excluded from the χ^2 calculation, but some problematic cases remain. Thus, e.g., if a calculation returns a too small cross section for a dimensionful quantity, the corresponding χ^2 value will be large, even though the shape of the distribution may be well described. It could be argued how this should be treated, how much uncertainty should be allowed for each observable, whether it is reasonable to include observables that a given model is not supposed to describe, etc. These are questions that we do not believe can be meaningfully (or reliably) addressed by a fully automated site containing tens of thousands of model/observable combinations. In the end, the interpretation of the information we display is up to you, the user. That is also why, at least for the time being, we do not display any direct comparisons between different MC generators.

To see the values for all of the individual distributions that enter a given process type, click on any of the $\langle\chi^2\rangle$ values in the table. Or, to see a comparison between two successive versions, click on any of the $\Delta\langle\chi^2\rangle$ values. The result of doing the latter is illustrated in figure 8. You can now scroll down the list of observables to get a more detailed view of which observables actually changed, and how. Also in this view, you can click on the numbers in the table. Doing so takes you to the final level of validation detail, which is a plot showing the two generator versions compared on the given observable. In the example of figure 8, clicking on the value *-0.29* in the third row will produce the plot shown in figure 9. As usual, you can use the links below the plot to download it in various formats or obtain all the numerical information that was used to make it.

[3] Garbage In, Garbage Out.