Journal of Instrumentation

PAPER • OPEN ACCESS

A Bayesian tune of the Herwig Monte Carlo event generator

To cite this article: Salvatore La Cagnina *et al* 2023 *JINST* **18** P10033

Published by IOP Publishing for Sissa Medialab

Received: *September 4, 2023*
Accepted: *October 11, 2023*
Published: *October 26, 2023*

A Bayesian tune of the Herwig Monte Carlo event generator

Salvatore La Cagnina,[a,*] **Kevin Kröninger,**[a] **Stefan Kluth**[b] **and Andrii Verbytskyi**[b]

[a] *Fakultät Physik, Technische Universität Dortmund, Otto-Hahn-Str. 4, 44227 Dortmund, Germany*

[b] *Max-Planck-Institut für Physik, Föhringer Ring 6, 80805 Munich, Germany*

E-mail: salvatore.lacagnina@tu-dortmund.de

Abstract: The optimisation (tuning) of the free parameters of Monte Carlo event generators by comparing their predictions with data is important since the simulations are used to calculate experimental efficiency and acceptance corrections, or provide predictions for signatures of hypothetical new processes in experiments. We present a tuning procedure that is based on Bayesian reasoning and that allows for a proper statistical interpretation of the results. The parameter space is fully explored using Markov Chain Monte Carlo. We apply the tuning procedure to the Herwig7 event generator with both the cluster and the string hadronization models and a large set of measurements from hadronic Z-boson decays produced at LEP in e^+e^- collisions. Furthermore, we introduce a coherent propagation of uncertainties from the realm of parameters to the realm of observables and we show the effects of including experimental correlations of the measurements. To allow comparison with the approaches of other groups, we repeat the tuning considering weights for individual measurements.

Keywords: Analysis and statistical methods; Pattern recognition, cluster finding, calibration and fitting methods

ArXiv ePrint: 2302.01139

*Corresponding author.

https://doi.org/10.1088/1748-0221/18/10/P10033

2023 JINST 18 P10033

Contents

2023 JINST 18 P10033

1 Introduction

Monte Carlo event generators (MCEGs) are an indispensable tool in experimental and theoretical particle physics. They simulate final states in high-energy particle collisions according to the predictions of the Standard Model of Particle Physics (SM). While the interactions of partons, i.e., quarks and gluons, are calculated from first principles, the transition from partons to hadrons is modeled using a semi-empirical approach, see, e.g. ref. [1] for a review. In the analysis of

experimental data, the predictions from MCEGs are combined with a detailed simulation of the experimental setup, which is crucial for understanding of the data. Examples include the precise determinations of the top-quark mass and the W-boson mass [2, 3]. The use of MCEGs to calculate predictions of particle physics theory to investigate a data likelihood is an application of simulation-based inference [4].

Recent advances in the development of MCEGs include an improved treatment of higher-order perturbative corrections for the simulation of hard scattering processes and an improved theoretical precision of the simulation of the parton showers. To tune the simulation of parton showers, MCEGs typically have a small number of free parameters, such as the value of the strong coupling constant α_S, the lower limit of the evolution parameter in the parton shower, and the parameters of the hadronization models. It is these parameters that have to be optimized in order to achieve the best possible description of the experimental data.

A first complete workflow for the determination of optimal MCEG parameters was developed by the DELPHI collaboration [5]. Today, the standard for such a workflow is a set of data-to-prediction comparison codes (analysis modules), which are included in the `Rivet` package [6]. The tuning workflow in the `Professor` [7] system creates multiple event samples with different parameter settings, approximates the predictions as functions of the tuning parameters for all data points, and finally finds the optimal values of the parameters by comparing the approximation functions to the data.

The original application for tuning workflows were data from LEP including event shape observables, jet production rates, charged and identified particle multiplicities, and charged particle momentum spectra [5, 7]. In previous studies, some datasets were weighted during the final optimization step to improve the description of the data after fitting. Detailed studies of the choice of these weights were performed in refs. [8, 9]. While this procedure was introduced to ensure a good description of the phenomenologically important observables, a proper statistical interpretation of the parameter estimation, in particular the uncertainty estimation, was not possible.

The application of the `Professor` workflow on parameter subspaces was explored in ref. [10]. Other approaches to tuning MCEGs are Bayesian optimisation [11] and machine learning-based reweighting using the complete simulated final states [12, 13]. Even though many approaches for complete tuning workflows have been proposed, manual procedures are still followed [14].

In this study, a tuning procedure based on Bayesian reasoning that allows for a proper statistical interpretation of the results was developed. Using Markov chain Monte Carlo algorithms, the full posterior distributions for the MCEG parameters were determined. Those posterior distributions provide a better understanding of the parameter space and may also reveal the limitations of the underlying physical models. In addition, the full posterior distribution allows uncertainties to be coherently propagated from the parameter space to the space of observables. As a result, the tuning procedure provides not only an estimate of the optimal MCEG parameters (with uncertainties and correlations), but also uncertainty estimates for the predictions. As a study case, the developed tuning procedure was applied to a well-defined set of measurements and the parameters of the `Herwig7` MCEG with two different hadronization models were optimized. In addition, the effects of including experimental correlations of the measurements were studied. To allow comparison with the approaches of other groups, the tuning was repeated considering the weights for individual measurements.

2023 JINST 18 P10033

2 Data selection

Similar to previous tuning studies [7, 15, 16], the data measured in e^+e^- collisions were used. The main motivation for this choice was to avoid ambiguities in the choice of parton density functions, which are unavoidable when using the data from proton-proton or proton-lepton colliders. In addition, the process $e^+e^- \to (Z/\gamma)^* \to$ partons is very well understood in perturbative QCD (pQCD), and in modern MCEGs it is used together with parton showers for accurate and stable simulations of the process $e^+e^- \to (Z/\gamma)^* \to$ hadrons.

The focus of the study is on the tuning of hadronization models used in the MCEGs. Therefore, a set of measurements that can be used to assess the quality of the modeling of hadronic final states [5, 6, 17–20] was selected. The corresponding `Rivet` analysis modules are ALEPH_1996_S3486095 [17], ALEPH_2001_S4656318 [18], DELPHI_1996_S3430090 [5], JADE_OPAL_2000_S4300807 [19] and PDG_HADRON_MULTIPLICITIES [20]. The `Rivet` analysis modules ALEPH_1996_S3486095 and DELPHI_1996_S3430090 provide calculations for the event shape observables such as sphericity, thrust, aplanarity etc. These observables and the differential jet rates provided by the JADE_OPAL_2000_S4300807 analysis module are sensitive to α_S. The ALEPH_1996_S3486095 analysis module also provides identified particle spectra sensitive to fragmentation parameters. The b-quark fragmentation function observables were calculated with the ALEPH_2001_S4656318 analysis module. The PDG_HADRON_MULTIPLICITIES module was used to study the simulation of particle multiplicities.

In total, 100 observables from these five `Rivet` analysis modules were used. The full list of the used observables is given in tables 8, 10 and 9 in appendix B.

3 Monte Carlo event generators

To perform the calculations for the $e^+e^- \to (Z/\gamma^*) \to 2, 3, 4, 5$ partons processes the `Herwig7` MCEG version 7.2.2 [21] was used with the MENLOPS method [22] using the `MadGraph5` [23] matrix element generator and the `OpenLoops` [24] one loop library. The two-parton final states were predicted with full NLO accuracy in perturbative QCD in this scheme. The QCD matrix elements were calculated with mass effects taken into account for massive b-quarks. Two models were used for the modeling of the hadronization process and the details on these models are given below. The simulated events produced by the MCEG were put into the `HepMC` format [25] and passed to the `Rivet` package, where they were processed by the `Rivet` analysis modules for the corresponding data sets. The comparison between the MCEG predictions and the data is explained in section 4.

3.1 `Herwig7` with the cluster hadronization model

The default hadronization model of the `Herwig7` MCEG is the cluster hadronisation model [26]. This model is referred to as `Herwig7-H7`. For this model, the list of parameters, as well as the ranges in which the parameters were varied, are listed in table 1.

The selection of the parameter ranges was driven by two factors: the desire for a wide coverage of the physically meaningful parameter space and the constraints imposed on the parameters by the MC generator code and models. As a result, most parameter ranges were chosen to be in a ±50% window around their default settings with some exceptions that will be discussed in the following.

2023 JINST 18 P10033

Table 1. Parameters for the `Herwig7-H7` tune, their ranges and default values. The quantities without units are dimensionless.

Parameter	Range	Default
AlphaQCD	$[0.1000, 0.1417]$	0.1181
IRcutoff (GeV)	$[0.5004, 1.5012]$	1.0080
m_g (GeV)	$[0.7445, 1.1400]$	0.9500
m_s (GeV)	$[0.4734, 0.5000]m_g$	0.4500
ClMax (GeV)	$[1.9334, 5.8000]$	3.8667
ClPow	$[0.8295, 2.4885]$	1.6590
ClSmr	$[0.1719, 0.8593]$	0.3437
PSplit	$[0.3450, 1.0348]$	0.6899

The ranges for the `AlphaQCD` parameter were chosen taking into account the hardcoded minimal and maximal values in the generator code. Similarly, the gluon and strange quark constituent masses are limited by the boundaries set by the hadronization model [27]. The implementation of the model in `Herwig7` also requires the gluon and the strange quark to fulfill the condition $m_g > \frac{m_s}{2}$ for a successful run. Hereby, a lower bound can be set with the condition $m_s > m_{u,d} = 0.35$. Given the constraints of the constituent masses, m_s was chosen to be varied as a fraction of m_g with the lower fraction equal to the default ratio of $m_s/m_g = 0.47$ which gives a lower bound of $\min(m_g) = 0.74$.

3.2 `Herwig7` with the Lund string hadronization model

The second hadronization model available for `Herwig7` is the Lund string hadronization model as implemented in `Pythia8` [28]. The `Pythia8` hadronization code is interfaced to `Herwig7` using the `ThePEG` interface [10, 29]. This model is referred to as `Herwig7-P8` and used `Pythia 8.306` to generate the `Herwig7-P8` samples. The selection of the parameter ranges for the `Herwig7-P8` model was done in the same way as for the `Herwig7-H7` model. The ranges for the tune were chosen within the allowed scope of the `Pythia8` framework. The parameters `aExtraDiQuark` and `aExtraSQuark` have been varied and fitted as well. However, due to a lack of sensitivity of the observables to these parameters, they were fixed when running the tune. Their fixed value was derived from the Monash Tune [14] with 0.0 for `aExtraSQuark` and 0.97 for `aExtraDiQuark`. The parameter ranges for `Herwig7-P8` are shown in table 2.

Table 2. Parameters for the `Herwig7-P8` tune, their ranges and default values. The parameters marked as fixed are set to their default values for the tuning process. The quantities without units are dimensionless.

Parameter	Range	Fixed	Default
AlphaQCD	$[0.1000, 0.1417]$	×	0.1181
IRcutoff (GeV)	$[0.2002, 1.8014]$	×	1.0080
SigmaPT (GeV)	$[0.000, 1.000]$	×	0.335
aLund	$[0.20, 2.00]$	×	0.68
bLund (GeV^{-2})	$[0.00, 2.00]$	×	0.98
aExtraDiQuark	$[0.00, 2.00]$	✓	0.97
aExtraSQuark	$[0.00, 2.00]$	✓	0.00

2023 JINST 18 P10033

4 Tuning procedure and statistical model

The tuning procedure was implemented in the `JULIA` package `BAT.jl` [30] which is a software tool for Bayesian analysis containing algorithms for parameter estimation, hypothesis testing, model comparison, and goodness-of-fit tests. `BAT.jl` provides interfaces to define arbitrary data likelihood functions and prior distributions for statistical models defined by the user. The data likelihood and the prior were multiplied to obtain the (unnormalized) posterior distribution which was then explored with dedicated algorithms, in particular Markov Chain Monte Carlo techniques and derivatives thereof.

The `BAT.jl`-based package `EFTFitter.jl` [31] provides a data likelihood function for combining several quantities including uncertainties and correlated data. It has been developed for the interpretation of data in the context of effective field theories, which is a mathematically similar problem. The data likelihood $L(\vec{D}|\vec{\lambda})$ is a multivariate Gaussian, i.e.

$$\ln L(\vec{D}|\vec{\lambda}) = -\frac{1}{2}[\vec{D} - \vec{f}(\vec{\lambda})]^T \cdot M^{-1} \cdot [\vec{D} - \vec{f}(\vec{\lambda})] \ . \tag{4.1}$$

The data are represented by $\vec{D}$, where each component of the vector corresponds to a measured value of an observable (either a single measurement or a bin of a differential distribution). The covariance matrix of the data is denoted by M. The components $f_{b,O}(\vec{\lambda})$ of the vector $\vec{f}$ represent the MCEG predictions of bin b for observable O as a function of the parameters $\vec{\lambda}$. For the sake of simplicity and to facilitate comparison with previous results, the prior distributions of the MCEG parameters are chosen to be uniform over the ranges shown in tables 1 and 2. For the `Herwig7-H7` tune, an additional constraint is imposed on the data likelihood by demanding $m_g > m_s/2$ in the generated samples. The covariance matrix was chosen to be a diagonal matrix, the case of non-negligible off-diagonal elements is discussed in section 8.

The posterior distribution was sampled using the Metropolis-Hasting (MH) algorithm for which convergence is achieved within a few tuning cycles. The MH sampling was performed using six chains with 10^6 sampling steps. In order to check for convergence of the different chains, `BAT.jl` uses the Gelman-Rubin test [32], which has been generalized for the multivariate case by Brooks and Gelman [33]. The convergence parameter R, which is calculated by comparing the variance of the samples within a chain to the variance of samples between different chains, see ref. [33], was slightly increased to 1.3 from its default value of 1.1 to account for the larger number of chains which leads to higher distances, causing higher R values of the chains during tuning.

The global mode values of the posterior distribution were chosen as the optimal parameter set. While the full multidimensional posterior distribution can be used for further investigations, e.g. studies of global modes or the propagation of uncertainties, the uncertainties for the individual parameters were defined as the smallest 68% credibility interval of the marginalized distributions.

5 Parameterization of the MCEG predictions

The data likelihood contains a vector of functions $\vec{f}$ that represents the MCEG predictions as a function of the free parameters, e.g. the mean charged multiplicity as a function of α_s. Since the evaluation of these predictions using MC simulated events is very CPU-time intensive, they were replaced by approximative parametrizations. The resulting analytic expressions for the functions $f_{b,O}$ are much faster to calculate during the sampling of the parameter space during the optimization process.

2023 JINST 18 P10033

To parameterize the MCEG response the multidimensional cubic polynomials were used, i.e. functions of the form

$$f_{b,O}(\vec{\lambda}) \approx f_{b,O}^{\text{cubic}}(\vec{\lambda}) = c_0 + \sum_i c_i \lambda_i + \sum_i \sum_{j \leq i} c_{ij} \lambda_i \lambda_j + \sum_i \sum_{j \leq i} \sum_{k \leq j} c_{ijk} \lambda_i \lambda_j \lambda_k \tag{5.1}$$

for each bin b and observable O. The parameters $\vec{\lambda}$ for the used hadronization models are shown in tables 1 and 2. The linear, quadratic and cubic polynomial coefficients are denoted c_i, c_{ij} and c_{ijk}, respectively. The total number of coefficients needed for the approximation with N parameters was $1 + N + N(N+1)/2 + N(N+1)(N+2)/6$.

As the set of reference points for the fit, 500 (700) randomly chosen parameter sets were used for the parametrization of the `Herwig7-H7` (`Herwig7-P8`) models. For each set 10^6 MC events were generated. Below, those samples are referred to as *analysis samples*. For each resulting sample of events, the predictions for each observable were calculated using the `Rivet` framework. Finally, the polynomial model was fitted to the MCEG predictions using the `LsqFit.jl` [34] package, which implements the Levenberg-Marquardt algorithm [35, 36] for non-linear fitting procedures.

A reasonable agreement between the MCEG predictions and the corresponding parametrization functions was observed. Studies of the goodness-of-fit can be found in appendix D. While the focus of the current study is on the tuning process itself, it is worth noting that for a high-precision tune that is used, e.g., by experimental collaborations, the parametrization of the predictions should be addressed in more detail and alternative techniques, e.g. template morphing, should be considered.

To give a visual impression of the agreement between the MCEG predictions and the parametrization functions, additional *test samples* were produced for which only one parameter at a time was varied. All parameters were set to their default values according to tables 1 and 2, and one parameter was varied in the range given there in eleven equidistant steps. The first bin content of a sphericity distribution as a function of four parameters is shown in figure 1 as an example. The markers indicate the MC calculation and the red line represents the fitted parameterization function. The red area represents the uncertainties from the fitting procedure. To guide the eye, the blue area represents the range of predicted values from the parametrization model when changing the default values by ±5%. In general, the parameterization functions describe the MCEG predictions reasonably well with deviations of the order of a few percent.

6 Propagation of uncertainties

One of the main advantages of using a Bayesian approach in combination with a fast parametrization of the predictions is the opportunity to propagate the uncertainty from the parameter space to the space of observables. The resulting uncertainty is then related to the tuning procedure and limited knowledge about the MCEG parameters. The uncertainty propagation is done by re-sampling the posterior distribution to generate 10^6 parameter points $\vec{p}$. For each bin of each observable, these points were evaluated according to the parametrization function with the corresponding coefficients $\vec{y}_{b,O} = f_{b,O}(\vec{p})$ using eq. (5.1). The resulting values $\vec{y}_{b,O}$ represent the statistical distribution of the bin content and the width of this distribution reflects the statistical uncertainty from the tuning process.

As an example, figure 2 shows these distributions for the first bin of the sphericity observable (top) and for the multiplicity of B_u^+ mesons (bottom). One often uses the standard deviation as a measure of

2023 JINST 18 P10033

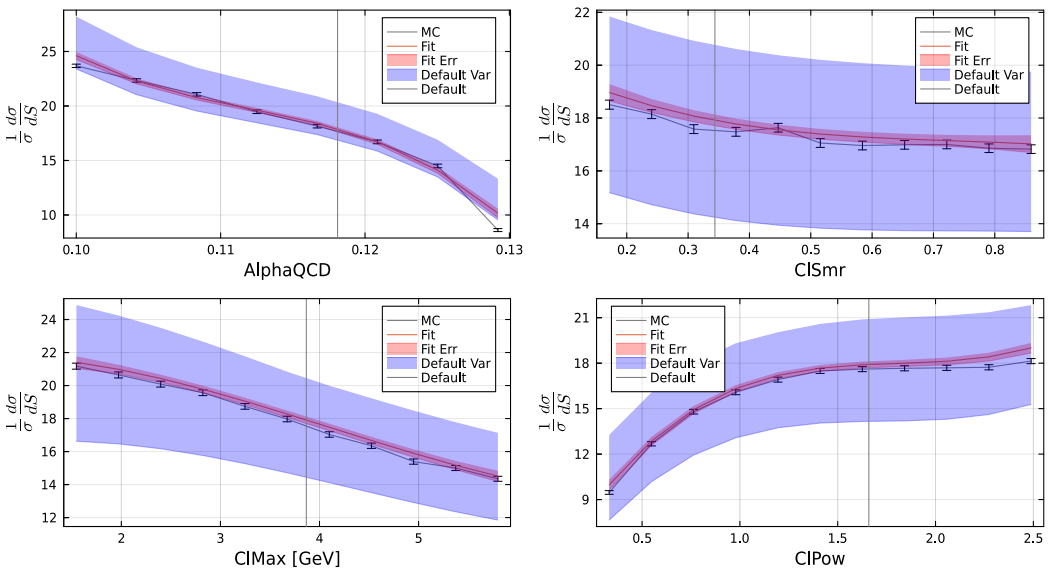

Figure 1. The content of the first bin of the sphericity distribution from the `Rivet` module DELPHI_1996_S3430090 as a function of `AlphaQCD`, `ClSmr`, `ClMax` and `ClPow` with parameter sets from the test samples. The approximation model is shown in red in comparison to the MCEG test sample predictions. The red band represents the uncertainty obtained by propagating the uncertainty of the fitted coefficients. The blue band represents variations caused by shifting the default values for the evaluation of the approximation model by ±5%.

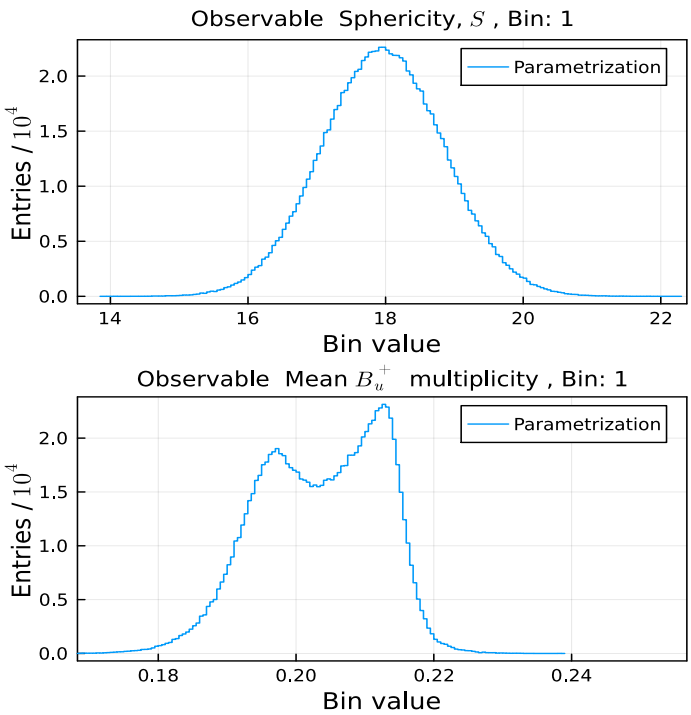

Figure 2. Distribution of the bin content for two bins of two observables using the parametrization with the posterior samples as input. The figure on top shows the first bin of the sphericity observable while the bottom figure shows the multiplicity of B_u^+ mesons.

2023 JINST 18 P10033

uncertainty, silently assuming a Gaussian distribution. In the case of the sphericity observable, this assumption is roughly valid as the distribution is uni-modal and symmetric. However, in the case of the B_u^+-meson multiplicity, the distribution has two separate modes and is not symmetric. In such cases, a full propagation of the uncertainty is required even if it is difficult to display graphically. For simplicity, the standard deviation was used as a measure of uncertainty for the tuning process in the following.

7 Results

We perform the tune of the `Herwig7-H7` and the `Herwig7-P8` models according to the procedure defined above. The global modes are quoted as the final results of the tune, while the modes and smallest 68% intervals of the marginalized distributions are quoted as results for the individual parameters and their uncertainties. We also display and discuss the one- and two-dimensional marginalized distributions. In order to evaluate the quality of the tune, we generate a new MC sample with the parameters set to the global mode and compare each data and MCEG distribution by calculating the χ^2 and the corresponding p-value.

7.1 `Herwig7-H7` tune

The global mode as well as the marginal and the smallest 68% credibility intervals of the marginalized posterior probability are summarized in table 3. The global and marginalized modes are close. The marginalized mode values are in agreement with the default settings shown in table 1 within the uncertainties for each parameter except for `ClSmr`. For that parameter, the default setting of 0.3437 lies within the 95 percentile.

Table 3. Results of the tune of the `Herwig7-H7` model. The values of the global and marginalized mode of the posterior samples as well as the smallest intervals containing 68% of the probability are listed.

Parameter	Global mode	Marginal mode	Smallest 68% interval
AlphaQCD	0.115	0.115	[0.112, 0.118]
IRCutoff (GeV)	0.879	0.755	[0.580, 1.020]
m_g(GeV)	0.709	0.738	[0.700, 0.955]
m_s(GeV)	0.353	0.375	[0.346, 0.470]
ClMax (GeV)	2.591	4.025	[3.200, 4.750]
ClPow	0.823	0.910	[0.740, 1.260], [1.540, 2.260]
ClSmr	0.675	0.725	[0.480, 0.885]
PSplit	0.868	0.728	[0.615, 0.865]

The one- and two-dimensional marginalized distributions are shown in figure 3. None of the parameters are described by a normal distribution. The distributions for the parameters `AlphaQCD`, `IRCutoff` and `PSplit` have only a slight asymmetry in their distributions, while `ClSmr` and `ClMax` show a more pronounced asymmetry. The distributions representing the gluon and strange quark constituent masses, (m_g and m_s), are almost one-sided with a sharp increase and slow drop-off towards higher values. The marginalized distribution for `ClPow` is multimodal. The global mode is located in the first peak, while the default value is closer to the second peak. Most two-dimensional distributions

2023 JINST 18 P10033

show a mild linear correlation. The most prominent exceptions are a very strong correlation between $m(g)$ and $m(s)$ and the highly non-Gaussian shape of the distributions of `ClPow` and `ClMax`.

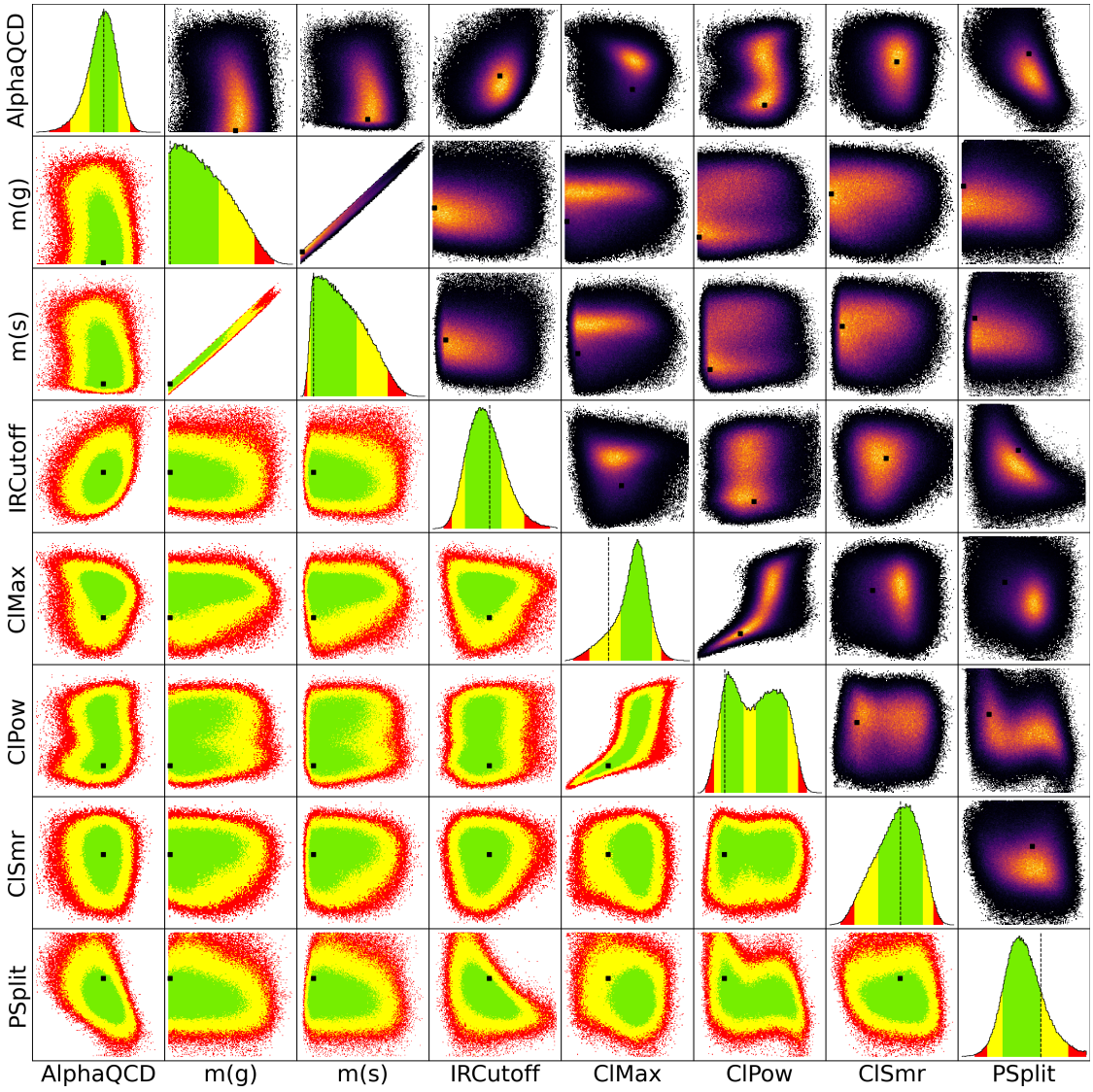

Figure 3. One and two-dimensional marginalized posterior distributions of the parameters for the tune of the `Herwig7-H7` model. The green, yellow and red areas contain the smallest 68, 95 and 99% intervals of the marginalized probability distributions, respectively. The dots and the lines are projections of the global mode representing the point with the highest probability.

The distribution of the p-values for each observable using the `Herwig7-H7` model with tuned and nominal parameter sets are shown in figure 4. The mean p-value increases from 0.095 for the default parameter values to about 0.133 for the tuned values. Since a majority of observables still tend to have low p-values we compare the two parameter sets on a logarithmic scale.

To demonstrate the impact of the tuning, figure 5 shows two observables as an example, namely the sphericity (left) and the B_u^+ multiplicity (right), calculated from the tuned and the nominal MCEG sample as well as from the data. The uncertainty bands for the tuned sample contain the MC statistical

2023 JINST 18 P10033

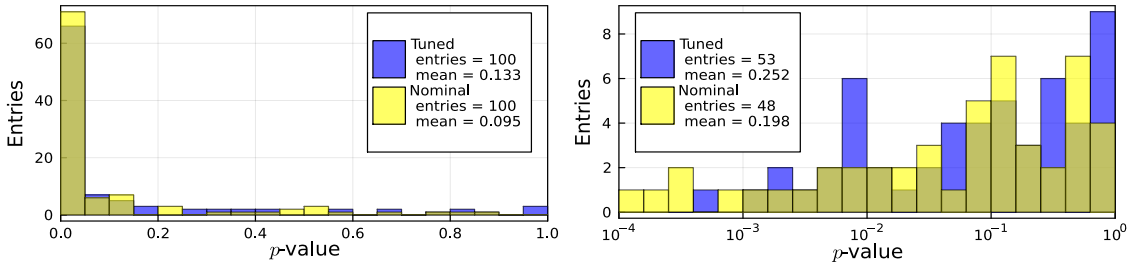

Figure 4. Distributions of p-values for the tuned and nominal MC samples for the `Herwig7-H7` tune. Each observable contributes one p-value toward the histogram. The bottom figure shows a subrange of the distribution $p > 10^{-4}$ in a logarithmic scale.

uncertainty and the propagated uncertainty from the tuning process as discussed in section 6. The uncertainties associated with the tuned sample are thus larger than those associated with the nominal sample. The tuned MCEG model results either in the same or in a better agreement with the data compared to the results obtained with the nominal sample. We find values of χ^2/ndf after (before) tuning of 7.60 (9.82) for the sphericity distribution and of 3.86 (25.6) for the mean B_u^+ multiplicity. This trend is seen in most observables and expected from the average p-value distribution.

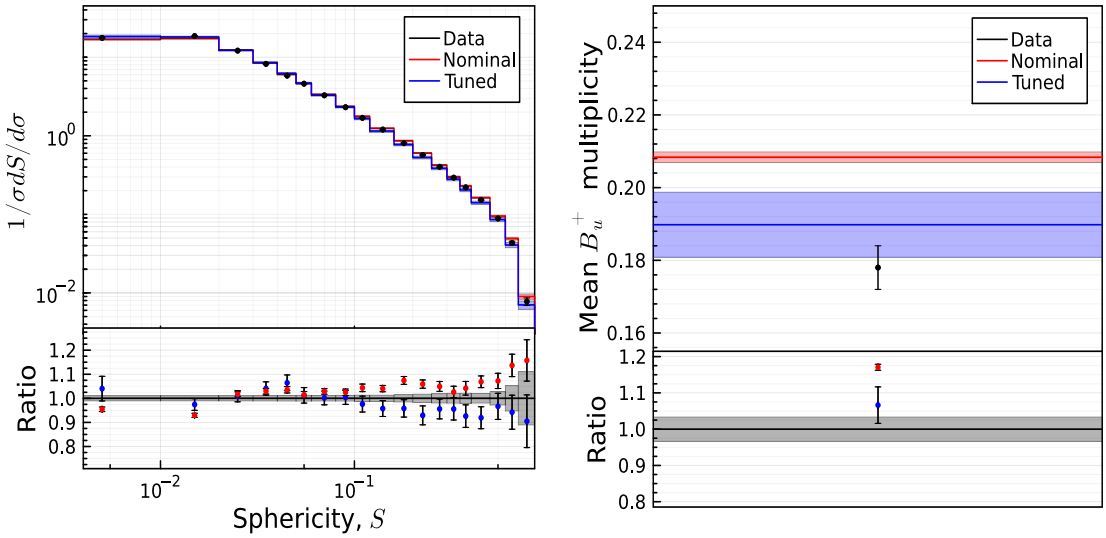

Figure 5. Distribution of the sphericity observable from DELPHI [5] (left) and the mean B_u^+ multiplicity [20] (right) for the data and the tuned and nominal MCEG samples for the `Herwig7-H7` tune. The calculation of the uncertainties for the nominal and tuned results is explained in the text. The bottom sections of the figures show the ratio to data.

7.2 `Herwig7-P8` tune

The marginal mode and smallest 68% intervals for the `Herwig7-P8` tune can be found in table 4. For `AlphaQCD`, `SigmaPT`, `IRCutoff` and `aLund` the default values, listed in table 2, are within the smallest 68% intervals while the `bLund` default values are significantly smaller.

2023 JINST 18 P10033

Table 4. Results of the tune of the `Herwig7-P8` model. The values of the global and marginalized mode of the posterior samples as well as the smallest intervals containing 68% of the probability are listed. Fixed parameters are set to their mode value. The quantities without units are dimensionless.

Parameter	Global mode	Marginal mode	Smallest 68% interval	Fixed
AlphaQCD	0.120	0.120	[0.117, 0.122]	×
IRCutOff (GeV)	1.079	1.079	[0.730, 1.390]	×
SigmaPT (GeV)	0.303	0.311	[0.284, 0.336]	×
aLund	1.287	1.435	[0.950, 1.760]	×
bLund (GeV^{-2})	1.302	1.325	[0.940, 1.720]	×
aExtraDiquark	0.970	0.970	-	✓
aExtraSQuark	0.0	0.0	-	✓

Figure 6 shows the posterior distributions for the `Herwig7-P8` tune. The `AlphaQCD` and `SigmaPT` parameters are well constrained with both marginalized distributions being almost symmetric and showing a smaller width compared to the full prior size. The `IRCutoff` variable shows similar behavior. `aLund` and `bLund` show a correlation as visible in their two-dimensional marginalized distribution. Contrary to the other variables, their constraint is rather weak as the one-dimensional marginal distribution has a larger width which results in them being cut-off by the prior edges towards higher values. Compared to the posterior of the `Herwig7-H7` tune in figure 3, the posterior has a single mode indicating a less ambiguous solution tune.

The distributions of the p-values for the `Herwig7-P8` tune compared to the default parameter set are shown in figure 7. Similarly to the `Herwig7-H7` tune, the mean of the p-values increases from 0.096 to 0.143 for the tuned MCEG with fewer observables showing a p-value below $p < 10^{-4}$. As examples, the sphericity and B_u^+ multiplicity observables for the tuned and nominal samples are shown in figure 8. In this case, the sphericity distribution shows an improved agreement to data, while the multiplicity distribution shows no improvement. We find values of χ^2/ndf after (before) tuning of 6.42 (16.6) for the sphericity distribution and of 13.5 (13.3) for the mean B_u^+ multiplicity. However, the overall agreement to the data improves as indicated by the p-values.

7.3 `Herwig7-H7` and `Herwig7-P8` comparison

In addition to the comparison of the tuned and nominal samples, both hadronization models can be compared to each other. While it is not possible to compare the parameters one to one, the distributions of the observables and their agreement with the data can be compared. The distributions of the p-values of the observables suggest that the `Herwig7-P8` model has a better overall agreement to data, as can be seen in figure 9, with an increase in p-value of about 8% when compared to `Herwig7-H7`. A comparison of the sphericity and B_u^+ multiplicity observables is shown in figure 10 together with their tuning uncertainty as discussed in section 6. It is noticeable that this uncertainty is systematically larger for the `Herwig7-H7` tune. This behavior is, at least in part, expected as the larger number of parameters leads to larger uncertainties, even if those are individually constrained to the same degree. The predictions from the models are in agreement within their uncertainties for the sphericity observable, while the central values of the `Herwig7-P8` model better reproduce the

2023 JINST 18 P10033

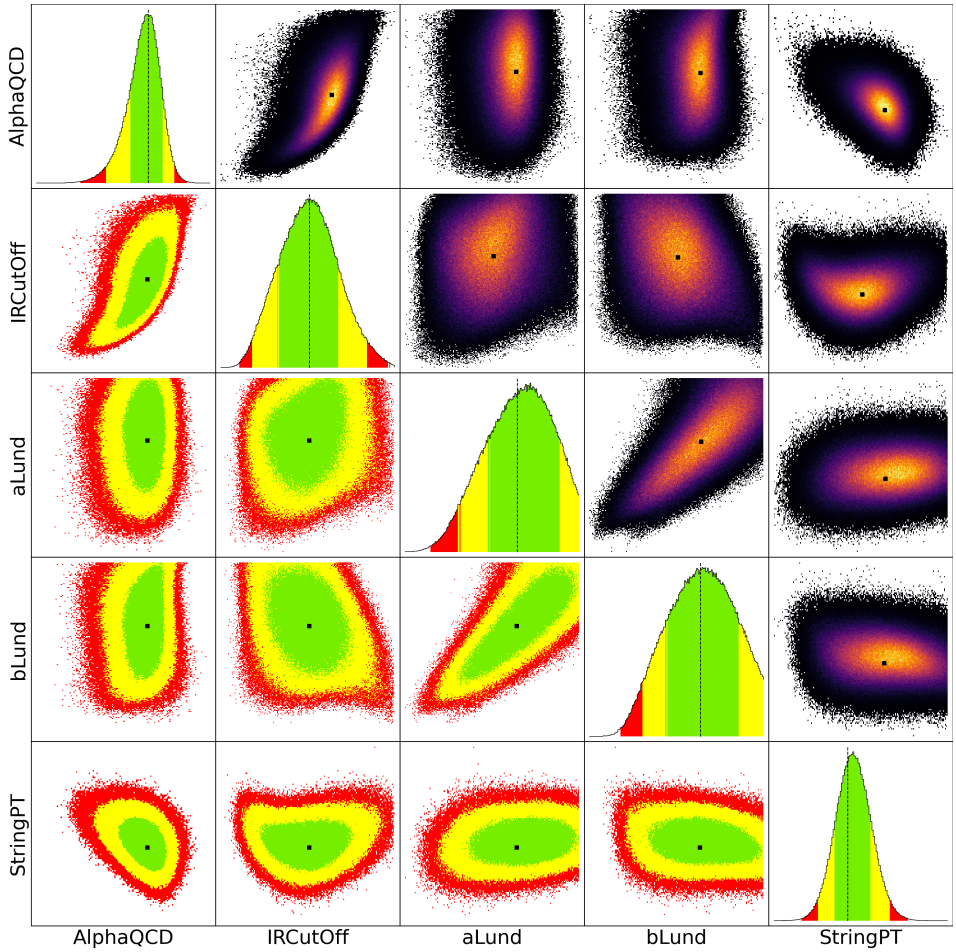

Figure 6. Results for the tune of the `Herwig7-P8` hadronization model. See the caption on figure 3 for a detailed explanation.

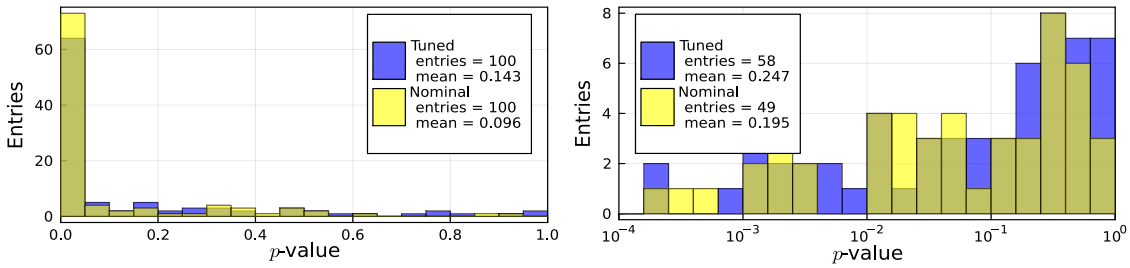

Figure 7. Distributions of p-values for the tuned and nominal MC samples for the `Herwig7-P8` tune. Each observable contributes one p-value toward the histogram. The right figure shows a subrange $p > 10^{-4}$ of the distribution in a logarithmic scale.

data. However, the multiplicity, shown in figure 10, is described better by the `Herwig7-H7` model. In conclusion, both models have similar performance with some observables being better described by either of them. Overall, the `Herwig7-P8` model has a marginally better agreement with the data in this exemplary set of observables.

2023 JINST 18 P10033

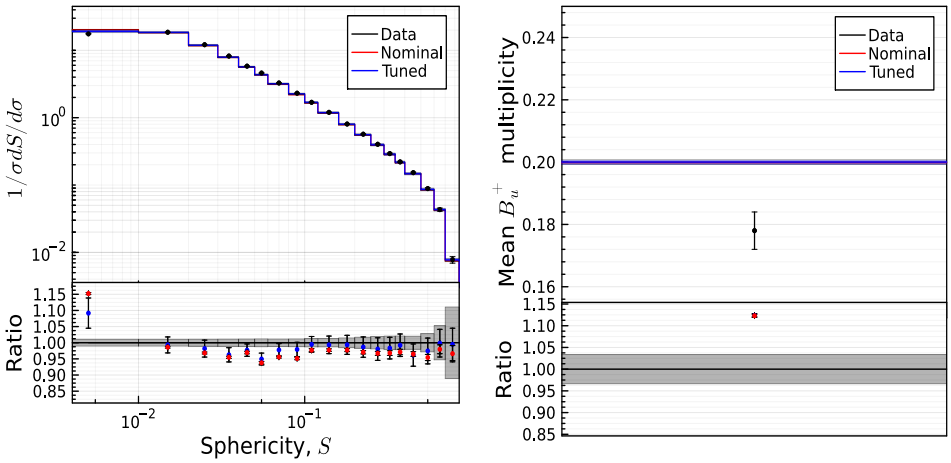

Figure 8. Distribution of the sphericity observable from DELPHI [5] on the left and the mean B_u^+ multiplicity [20] on the right for the data and the tuned and nominal MCEG samples for the `Herwig7-P8` tune. See the caption on figure 5 for further details.

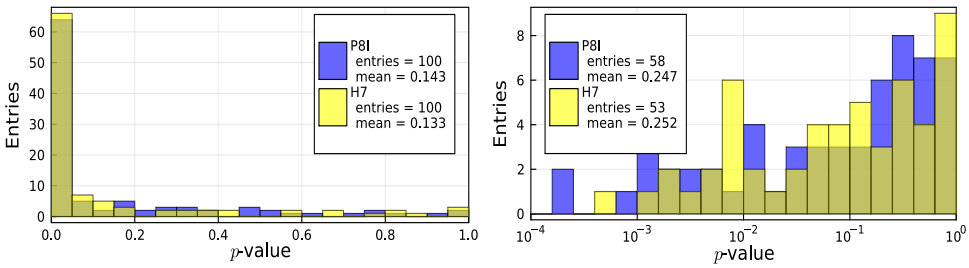

Figure 9. Distributions of p-values for the tuned `Herwig7-H7` and `Herwig7-P8` hadronization models. Each observable contributes one p-value toward the histogram. The bottom figure shows a subrange $p > 10^{-4}$ of the distribution in a logarithmic scale.

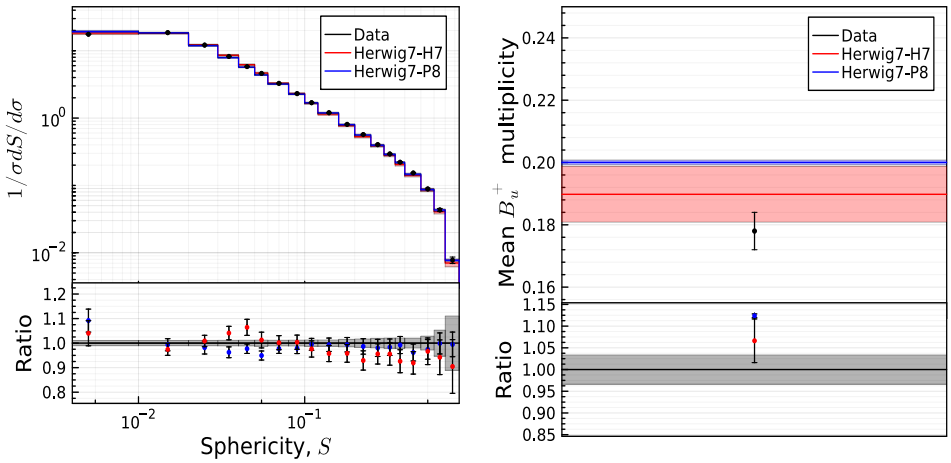

Figure 10. Distribution of the sphericity observable from DELPHI [5] on the left and the mean B_u^+ multiplicity [20] on the right for the data and the tuned MCEG samples for the `Herwig7-H7` and `Herwig7-P8` hadronization models. See the caption on figure 5 for further details.

2023 JINST 18 P10033

8 Studies of correlations

For the tuning presented in section 7 the uncertainties were assumed to be uncorrelated. In general however, it is expected that uncertainties, especially systematic uncertainties, are in fact correlated to some degree. In order to evaluate the impact of such correlations, additional tunes were performed. The main difference compared to the tunes in section 7 is that the covariance matrix M in eq. (4.1) is no longer purely diagonal. The covariance matrix is now constructed to be blockwise diagonal with a block for each observable distribution. For the off-diagonal entries in each block, the product $r \cdot \sigma_i \sigma_j$ was inserted, where $\sigma_{i,j}$ are the systematic uncertainties of data points i and j, and r quantifies the amount of correlation between both in terms of a linear correlation coefficient. Due to the lack of information regarding the correlation of systematic uncertainties from the analyses, a pragmatic approach was chosen by scanning through the values of the correlation factor $r = [0.0, 0.4, 0.6, 0.8, 0.9]$.

While these values are merely suggestive, they represent scenarios with mild, medium and strong correlations. As an example, figure 11 shows the two-dimensional posterior probability distribution for the two parameters `ClMax` and `ClPow` for different assumptions about r. The smallest areas containing 68% of the marginalized posterior shrink with increasing correlation factor r. For large correlation coefficients, such as $r = 0.9$, this area splits into two parts indicating the presence of a second mode within the posterior distribution.

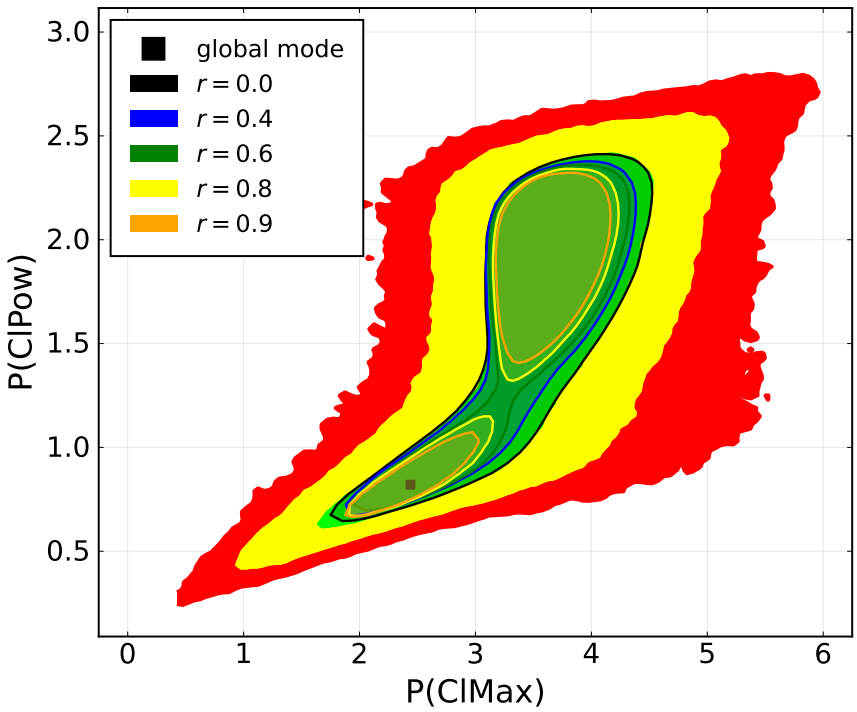

Figure 11. Marginalized two-dimensional distribution of the posterior probability for the parameters `ClMax` and `ClPow` for the `Herwig7-H7` tune. The underlying distribution represents the contours of the posterior for uncorrelated uncertainties with the green, yellow and red contours representing the smallest intervals containing 68%, 95% and 99%, respectively. The overlayed colored contours represent the smallest interval containing 68% of the posterior for different configurations of correlation coefficients.

2023 JINST 18 P10033

Introducing and increasing the correlation of uncertainties predominantly results in a narrowing of the posterior phase space, i.e. reduced uncertainties of the global modes of the MCEG parameters. The global mode for the MCEG parameters remains stable with only small deviations which are within the parameter uncertainties. This behavior of the global mode can be seen in table 5 where the global mode and the standard deviation of the parameters are listed for $r = 0.0$ and $r = 0.9$. Hence, the results of the tune can be regarded as stable, while the estimation of uncertainties of the tuned parameter depends on the choice of correlation factor.

Table 5. The mode and standard deviation of the tuned parameters for the tune of `Herwig7-H7`. Results are shown without correlation and for a correlation of $r = 0.9$. The quantities without units are dimensionless.

	Correlation			
	$r = 0.0$		$r = 0.9$	
Parameter	Mode	σ	Mode	σ
AlphaQCD	0.114	0.0032	0.115	0.0018
IRCutoff (GeV)	0.810	0.221	0.730	0.126
m_g (GeV)	0.716	0.129	0.780	0.106
m_s (GeV)	0.357	0.062	0.390	0.050
ClMax (GeV)	2.543	0.872	2.081	0.742
ClPow	0.800	0.560	0.667	0.545
ClSmr	0.662	0.194	0.461	0.148
PSplit	0.916	0.128	1.050	0.101

9 Studies on the effects of weighting observables

Traditionally, MC tuning relies on the use of weights to stabilize the tuning process and/or increase the importance of certain observables [10]. This procedure, however, compromises the statistical interpretation of the resulting uncertainties and potentially biases the obtained results of the tune. To test the effect, non-unity weights were introduced in the study and the results were compared to the tune discussed above with all weights equal to unity. The weights were introduced by adding a vector of coefficients $\vec{w}_{i=1\ldots N_{\text{bins}}}$ into the eq. (4.1) and dividing the likelihood by the sum of those weights. All weights for the bins from the same distributions were set to be equal.

In order to investigate the impact of weighting on the tune, the tuning process was repeated for two different weighting schemes w_1 and w_2 with the weight values for these schemes given in tables 8, 10 and 9 in appendix B. The weighting scheme w_1 applies higher weights for multiplicities while leaving the event shape variables mostly unchanged. The weighting scheme w_2 sets the weights of the multiplicities to zero and in contrast, increases the weights of the event shape variables and the weights of the mean charged multiplicities observable.

While changing from the unweighted tune to the weighting schemes w_1 and w_2, most of the posteriors retain their shape with only very minor changes. Most notably the positions of the global

2023 JINST 18 P10033

modes were shifted by the weighting procedure, as seen in tables 6 and 7. The mode values of these weighted tunes were also used for the MCEG to compare to data similarly to section 7. The p-values were calculated for the `Herwig7-H7` and `Herwig7-P8` tunes for both weighting schemes.

Table 6. The mode and standard deviation of the tuned parameters for the tune of `Herwig7-H7` using different weighting schemes. The quantities without units are dimensionless.

	Weighting scheme					
	none		w_1		w_2	
Parameter	Mode	σ	Mode	σ	Mode	σ
AlphaQCD	0.115	0.003	0.113	0.004	0.115	0.003
IRCutoff (GeV)	0.879	0.223	0.859	0.245	0.837	0.214
m_g (GeV)	0.709	0.128	0.706	0.136	0.708	0.13
m_s (GeV)	0.353	0.062	0.352	0.066	0.346	0.063
ClMax (GeV)	2.591	0.871	3.187	0.911	3.761	0.832
ClPow	0.823	0.561	0.847	0.567	2.147	0.541
ClSmr	0.675	0.193	0.501	0.228	0.806	0.213
PSplit	0.868	0.130	0.867	0.141	0.776	0.137

Table 7. The mode and standard deviation of the tuned parameters for the tune of `Herwig7-P8` using different weighting schemes. The quantities without units are dimensionless.

	Weighting scheme					
	none		w_1		w_2	
Parameter	Mode	σ	Mode	σ	Mode	σ
AlphaQCD	0.120	0.003	0.120	0.004	0.120	0.003
IRCutOff (GeV)	1.079	0.313	1.135	0.363	1.115	0.339
aLund	1.287	0.376	1.380	0.416	1.381	0.396
bLund (GeV^{-2})	1.302	0.359	1.369	0.379	1.393	0.363
aExtraDiquark	0.97	-	0.97	-	0.97	-
aExtraSQuark	0.0	-	0.0	-	0.0	-
SigmaPT (GeV)	0.303	0.026	0.304	0.031	0.302	0.027

The `Herwig7-H7` tune seems to benefit from the first weighting scheme as the mean of p-values increases from 0.133 to 0.151 while the second scheme decreases the overall agreement to data. In contrast, the `Herwig7-P8` benefits from both weighting schemes, although, only yielding a small additional improvement from 0.143 to 0.147 when compared to the unweighted tunes from section 7. The posterior distributions for the `Herwig7-H7` and `Herwig7-P8` tune are shown in figure 14 and figure 15 respectively in appendix C.

2023 JINST 18 P10033

10 Conclusions

A MCEG tuning procedure based on proper statistical grounds using a Bayesian approach is presented. As an application example, the procedure was used to tune the `Herwig7` MCEG with two different hadronization models. The data used for those studies were collected by the LEP experiments and include event-shape and jet rate distributions, charged hadron momentum spectra and multiplicities from the process $e^+e^- \to (Z/\gamma)^* \to$ hadrons. The `Rivet` framework was used to generate the analysis code. The global mode values of the posterior distribution were chosen as the optimal parameter set. In addition, appropriate uncertainty measures for the individual parameters are provided, e.g. from the smallest 68% credibility intervals, as well as uncertainty propagation of these parameters. The impact of different correlation assumptions and data-weighting schemes on the final results was investigated.

Several conclusions can be drawn from these observations. First it was found that the Bayesian approach to tuning MCEG works successfully. In particular, sets of optimized parameters including appropriate uncertainty estimates were obtained. These estimates can then be used to propagate the uncertainties on the optimal parameter set to the space of observables. As a result, the tuning of MCEG improves the agreement between the data and the MCEG predictions, and it provides an estimate for the uncertainties of prediction from the tuning process itself.

Second, it was shown that correlations between measurements can have an impact on optimization results, especially when uncertainties are estimated. Therefore it is recommended to the experimental collaborations if possible to derive and publish those correlations, and suggest that such correlations be carefully considered in further tuning studies. Third, it was found that for the particular data sets, the MCEG models examined, and the tuning parameters considered, the `Herwig7` MCEG using the Lund string hadronization model describes the data slightly better than the standard cluster hadronization model.

Acknowledgments

This work was supported by the German Science Foundation DFG through the Collaborative Research Center SFB1491 and project KR 4060/7-1

The authors are grateful to the authors and contributors of the software packages which made this analysis possible. The authors are grateful to Andrzej Siodmok for valuable discussions of this manuscript and the presented analysis.

A Software used in the analysis

The list of the used software is given below. The statistical analysis was performed using the `BAT.jl` [30] and `EFTFitter.jl` [31] packages of the `JULIA` language. The fitting was performed using the `LsqFit.jl` package for the `JULIA` language. The generation of the MC event samples and their processing was done using `Herwig7.2.2` [21], `MadGraph5` [23], `OpenLoops` [24], `Pythia8` [28], `TheP8I` [29], `Rivet` [6], `ROOT 6.22` [37] and `HepMC3` packages packed into `singularity` containers based on Fedora Linux distribution.

2023 JINST 18 P10033

B Observables used in the analyses

Table 8. Lists of `Rivet` analyses and bins used in the tuning including their descriptions imported from `Rivet` as well as two different weighting schemes used for different tunes, part I [17–19].

`Rivet` analysis and bin code	Description	Weight scheme	
		w_1	w_2
ALEPH_1996_S3486095 [17]			
d01-x01-y01	Sphericity, S (charged)	1	5
d02-x01-y01	Aplanarity, A (charged)	2	10
d03-x01-y01	1-Thrust, $1 - T$ (charged)	1	5
d04-x01-y01	Thrust minor, m (charged)	2	10
d07-x01-y01	C parameter (charged)	1	5
d08-x01-y01	Oblateness, $M - m$ (charged)	1	5
d09-x01-y01	Scaled momentum, $x_p = \|p\|/\|p_{\text{beam}}\|$ (charged)	1	5
d11-x01-y01	In-plane p_T w.r.t. sphericity axes (charged)	1	5
d12-x01-y01	Out-of-plane p_T w.r.t. sphericity axes (charged)	1	5
d17-x01-y01	Log of scaled momentum, $\log(1/x_p)$ (charged)	1	5
d18-x01-y01	Charged multiplicity	2	10
d19-x01-y01	Mean charged multiplicity	150	750
d25-x01-y01	$\pi^{\pm}$ spectrum	1	1
d26-x01-y01	$K^{\pm}$ spectrum	1	1
d29-x01-y01	π^0 spectrum	1	1
d30-x01-y01	η spectrum	1	1
d31-x01-y01	η' spectrum	1	1
d32-x01-y01	K^0 spectrum	1	1
d33-x01-y01	Λ^0 spectrum	1	1
d34-x01-y01	Ξ^- spectrum	1	1
d35-x01-y01	$\Sigma^{\pm}(1385)$ spectrum	1	1
d36-x01-y01	$\Xi^0(1530)$ spectrum	1	1
d37-x01-y01	ρ spectrum	1	1
d38-x01-y01	$\omega(782)$ spectrum	1	1
d39-x01-y01	$K^{*0}(892)$ spectrum	1	1
d40-x01-y01	ϕ spectrum	1	1
d43-x01-y01	$K^{*\pm}(892)$ spectrum	1	1
ALEPH_2001_S4656318 [18]			
d01-x01-y01	b quark fragmentation function $f(x_B^{\text{weak}})$	7	35
d07-x01-y01	Mean of b quark fragmentation function $f(x_B^{\text{weak}})$	3	15
JADE_OPAL_2000_S4300807 [19]			
d26-x01-y01	2-jet Durham diff. rate	2	10
d26-x01-y02	3-jet Durham diff. rate	2	10
d26-x01-y03	4-jet Durham diff. rate	2	10
d26-x01-y04	5-jet Durham diff. rate	2	10

2023 JINST 18 P10033

Table 9. Lists of `Rivet` analyses and bins used in the tuning including their descriptions imported from `Rivet` as well as two different weighting schemes used for different tunes, part II [20].

`Rivet` analysis and bin code	Description	Weight scheme	
		w_1	w_2
PDG_HADRON_MULTIPLICITIES [20]			
d01-x01-y03	Multiplicity of π^+	10	0
d02-x01-y03	—"— π^0	10	0
d03-x01-y03	—"— K^+	10	0
d04-x01-y03	—"— K^0	10	0
d05-x01-y03	—"— η	10	0
d06-x01-y03	—"— $\eta'(958)$	10	0
d07-x01-y03	—"— D^+	10	0
d08-x01-y03	—"— D^0	10	0
d09-x01-y03	—"— D_s^+	10	0
d10-x01-y01	—"— B^+, B_d^0	10	0
d11-x01-y01	—"— B_u^+	10	0
d12-x01-y01	—"— B_s^0	10	0
d13-x01-y03	—"— $f_0(980)$	10	0
d14-x01-y01	—"— $a_0^+(980)$	10	0
d15-x01-y03	—"— $\rho^0(770)$	10	0
d16-x01-y01	—"— $\rho^+(770)$	10	0
d17-x01-y02	—"— $\omega(782)$	10	0
d18-x01-y03	—"— $K^{*+}(892)$	10	0
d19-x01-y03	—"— $K^{*0}(892)$	10	0
d20-x01-y03	—"— $\phi(1020)$	10	0
d21-x01-y03	—"— $D^{*+}(2010)$	10	0
d23-x01-y02	—"— $D_s^{*+}(2112)$	10	0
d24-x01-y01	—"— B^*	10	0
d25-x01-y02	—"— $J/\psi(1S)$	10	0
d26-x01-y01	—"— $\psi(2S)$	10	0
d27-x01-y01	—"— $\Upsilon(1S)$	10	0
d28-x01-y01	—"— $f_1(1285)$	10	0
d29-x01-y01	—"— $f_1(1420)$	10	0
d30-x01-y01	—"— $\chi_{c1}(3510)$	10	0
d31-x01-y03	—"— $f_2(1270)$	10	0
d32-x01-y01	—"— $f_2'(1525)$	10	0
d34-x01-y02	—"— $K_2^{*0}(1430)$	10	0
d35-x01-y01	—"— B^{**}	10	0
d36-x01-y01	—"— D_{s1}^+	10	0
d37-x01-y01	—"— D_{s2}^+	10	0
d38-x01-y03	—"— p	10	0
d39-x01-y03	—"— Λ	10	0

2023 JINST 18 P10033

Table 10. Lists of `Rivet` analyses and bins used in the tuning including their descriptions imported from `Rivet` as well as two different weighting schemes used for different tunes, part III [5].

`Rivet` analysis and bin code	Description	Weight scheme	
		w_1	w_2
PDG_HADRON_MULTIPLICITIES [20]			
d40-x01-y02	—"— Σ^0	10	0
d41-x01-y01	—"— Σ^-	10	0
d42-x01-y01	—"— Σ^+	10	0
d43-x01-y01	—"— $\Sigma^\pm$	10	0
d44-x01-y03	—"— Ξ^-	10	0
d45-x01-y02	—"— $\Delta^{++}(1232)$	10	0
d46-x01-y03	—"— $\Sigma^-(1385)$	10	0
d47-x01-y03	—"— $\Sigma^+(1385)$	10	0
d48-x01-y03	—"— $\Sigma^\pm(1385)$	10	0
d49-x01-y02	—"— $\Xi^0(1530)$	10	0
d50-x01-y03	—"— Ω^-	10	0
d51-x01-y03	—"— Λ_c^+	10	0
d52-x01-y01	—"— Λ_b^0	10	0
d54-x01-y02	—"— $\Lambda(1520)$	10	0
DELPHI_1996_S3430090 [5]			
d01-x01-y01	In-plane $p_\perp$ w.r.t. thrust axes	1	5
d02-x01-y01	Out-of-plane $p_\perp$ w.r.t. thrust axes	1	5
d03-x01-y01	In-plane $p_\perp$ w.r.t. sphericity axes	1	5
d04-x01-y01	Out-of-plane $p_\perp$ w.r.t. sphericity axes	1	5
d07-x01-y01	Scaled momentum, $x_p = \lvert p\rvert/\lvert p_{\text{beam}}\rvert$	1	5
d08-x01-y01	Log of scaled momentum, $\log(1/x_p)$	1	5
d09-x01-y01	Mean out-of-plane $p_\perp$ w.r.t. thrust axes vs. x_p	1	5
d10-x01-y01	Mean $p_\perp$ vs. x_p	1	5
d11-x01-y01	$1 -$ Thrust	1	5
d12-x01-y01	Thrust major, M	1	5
d13-x01-y01	Thrust minor, m	2	10
d14-x01-y01	Oblateness $= M - m$	1	5
d15-x01-y01	Sphericity, S	1	5
d16-x01-y01	Aplanarity, A	2	10
d17-x01-y01	Planarity, P	1	5
d18-x01-y01	C parameter	1	5
d19-x01-y01	D parameter	1	5
d33-x01-y01	Energy-energy correlation, EEC	1	5
d35-x01-y01	Mean charged multiplicity	150	750

2023 JINST 18 P10033

C Studies of the weights impact

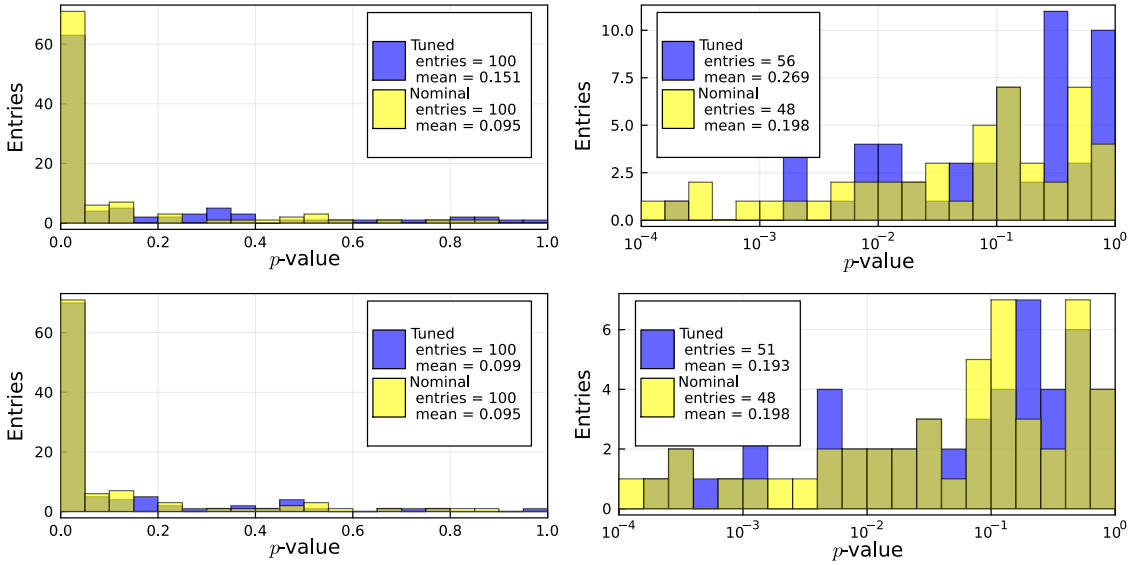

Figure 12. Distribution of p-values for tuned and nominal MC samples for the `Herwig7-H7` tune with weighting scheme one (two top distributions) and two (two bottom distributions). Each observable contributes one p-value toward the histogram. The bottom figures show a subrange $p > 10^{-4}$ of the distributions in a logarithmic scale.

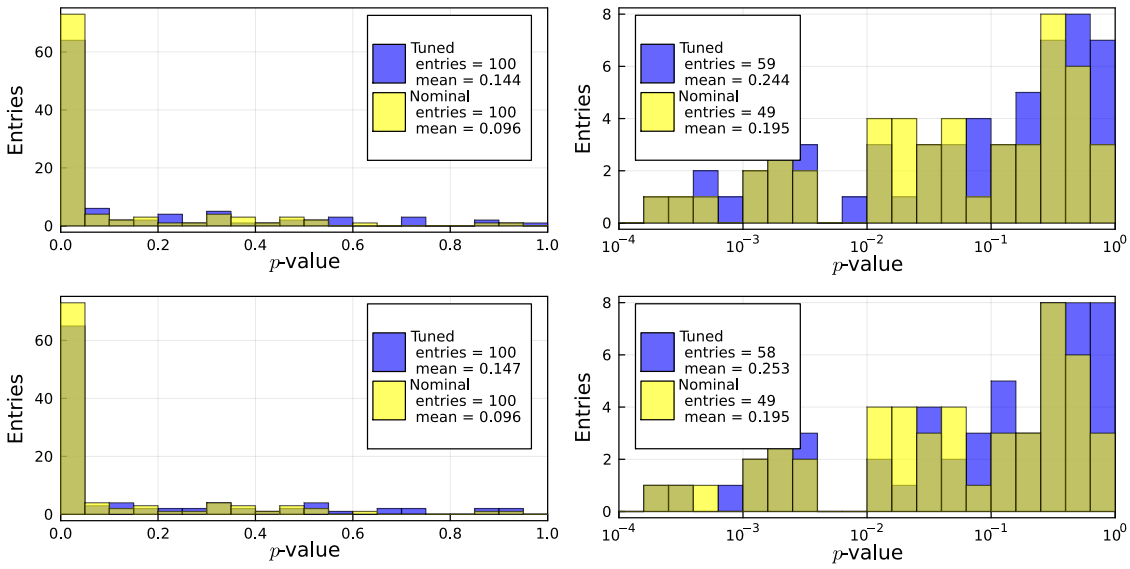

Figure 13. Distribution of p-values for tuned and nominal MC samples for the `Herwig7-P8` tune with weighting scheme one (two top distributions) and two (two bottom distributions). Each observable contributes one p-value toward the histogram. The bottom figures show a subrange $p > 10^{-4}$ of the distributions in a logarithmic scale.

2023 JINST 18 P10033

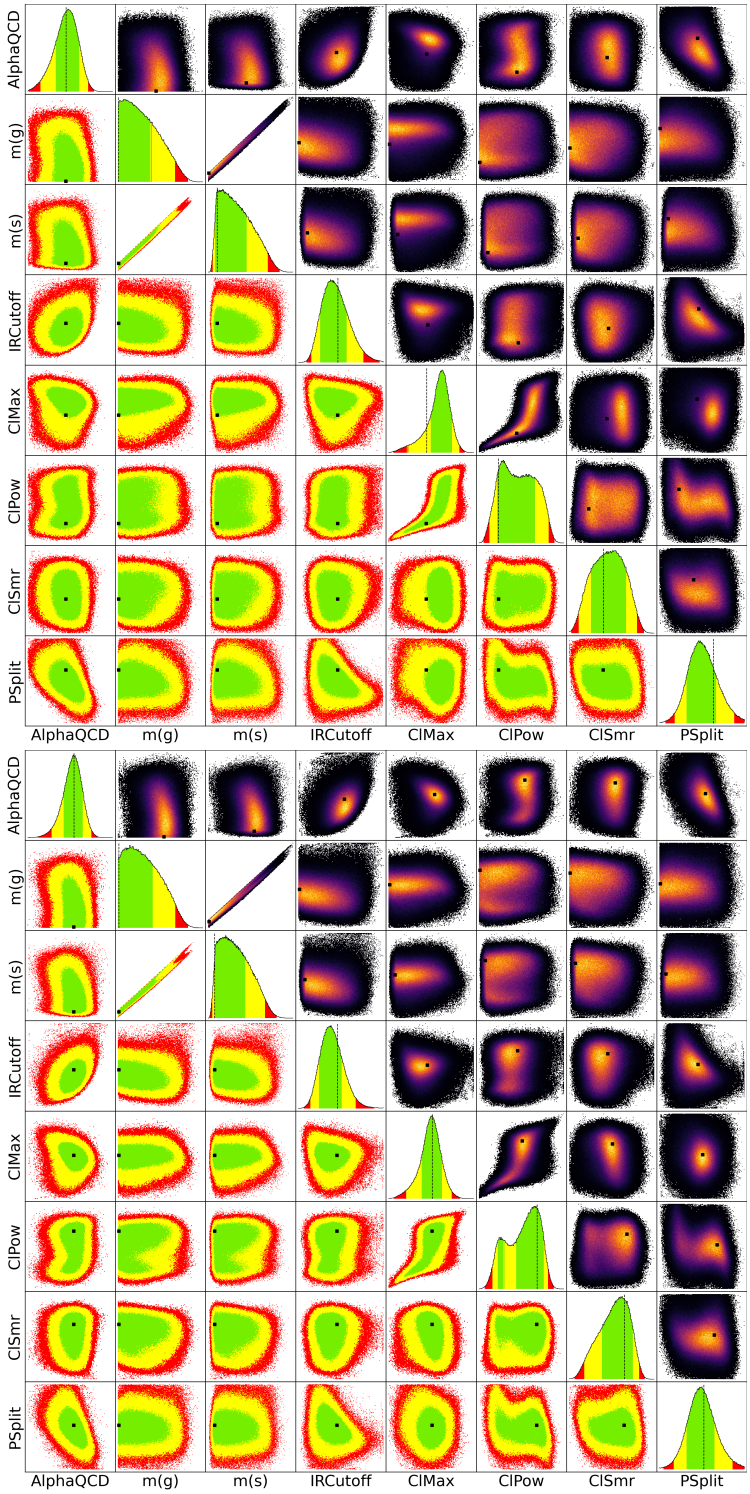

Figure 14. One and two dimensional marginalized posterior distributions for the `Herwig7-H7` tune using weighting scheme one (top) and two (bottom). See the caption on figure 3 for a detailed explanation.

2023 JINST 18 P10033

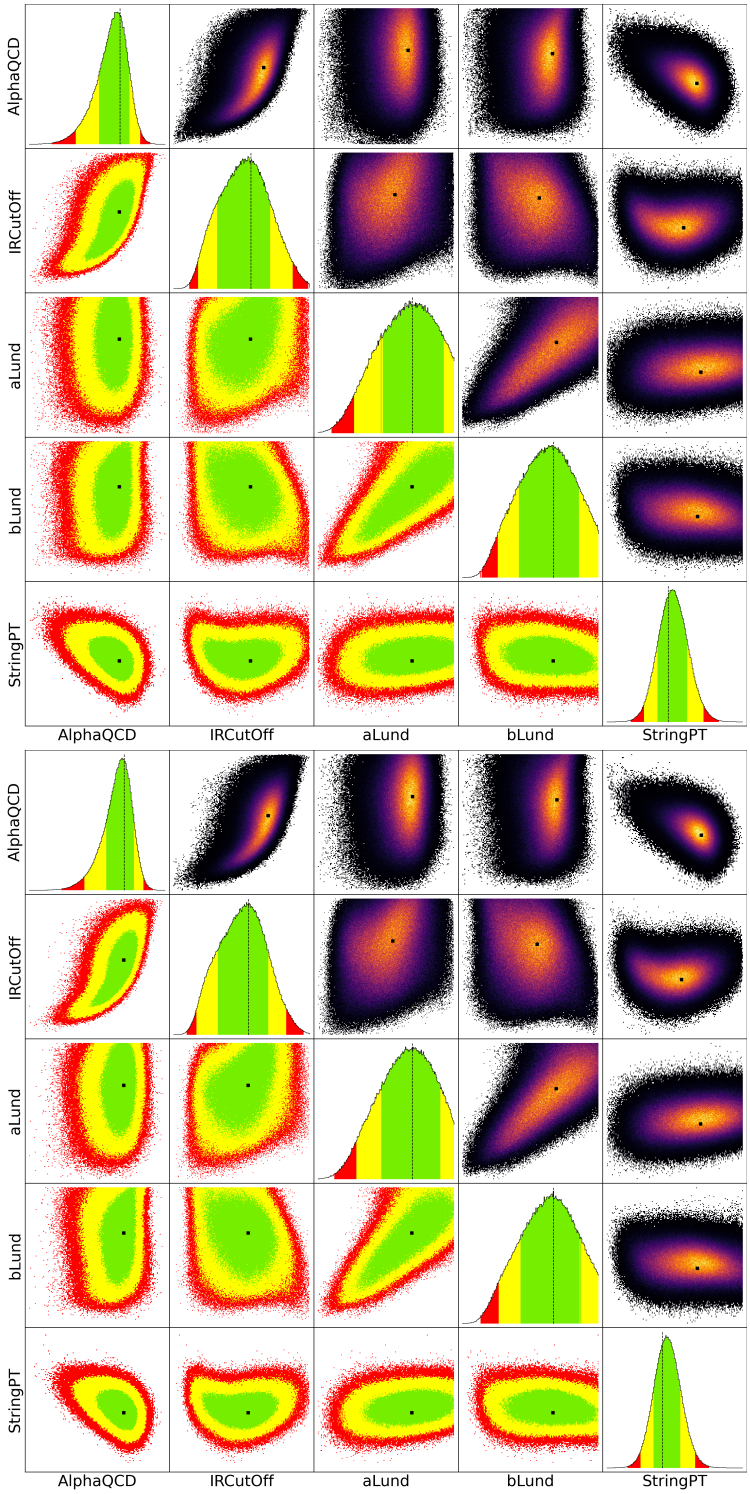

Figure 15. One and two dimensional marginalized posterior distributions for the `Herwig7-P8` tune using weighting scheme one (top) and two (bottom). See the caption on figure 3 for a detailed explanation.

D Studies on the goodness of fit

The test samples, described in section 5 were used to study the approximation model. The performance of the approximation model was evaluated using the residuals between the approximated values and the analysis and test samples. The corresponding pulls for those residuals in the analysis samples were calculated for each fitted bin as

$$p_{\text{analysis}} = \frac{f(\vec{\lambda}) - D_{MC}}{\sigma_r} = \frac{f(\vec{\lambda}) - D_{MC}}{\sqrt{\sigma^2_{D_{MC}} - \sigma^2_f}} \tag{D.1}$$

with the MCEG values D_{MC}, their uncertainties $\sigma_{D_{MC}}$, the value of the approximation $f(\vec{\lambda})$ and the uncertainty of the approximation σ^2_f which was calculated by propagating the fit uncertainties on the coefficients [38]. The calculations of pulls with the test samples take into account that the uncertainties of D_{MC} and f are uncorrelated, hereby

$$p_{\text{test}} = \frac{f(\vec{\lambda}) - D_{MC}}{\sigma_r} = \frac{f(\vec{\lambda}) - D_{MC}}{\sqrt{\sigma^2_{D_{MC}} + \sigma^2_f}} \tag{D.2}$$

In both cases, in eq. D.1 and eq. D.2 it is possible for the uncertainty of the fit to become larger than the uncertainty of the MCEG data points as the error propagation relies on a linear approximation of the cubic model. Hence, data points with $\sigma^2_f > \sigma^2_{D_{MC}}$ were omitted from the pull distributions and the behaviour of the approximation model for those points was studied on a case-by-case basis. The p_{test} and p_{analysis} follow normal distributions, although the p_{analysis} distribution shows larger values toward the tails. In addition, normal distributions were fitted to p_{analysis} and p_{test}, which resulted in mean values of -0.008 and -0.003, respectively. The fitted standard deviations are 1.66 and 1.18. The higher standard deviation and the larger tails of the p_{analysis} distribution indicate that the fitted approximation models do not always give a perfect description of the MCEG data, leading to larger average residuals than expected from statistical fluctuations alone. While it is beneficial for further tuning efforts to refine the description of the MCEG data, the approximation model suffices for the study's emphasis on presenting a new tuning method.

Apart from the checks described in section 5, additional studies were performed to ensure that the approximation procedure works as expected. The $\chi^2_{\text{red}} = \chi^2/n_{\text{dof}}$ values for each bin were used to evaluate the performance of the approximation procedure. The distribution of χ^2_{red} for all approximated observables with the `Herwig7` MCEG is given in figure 16. The distribution peaks at the $\chi^2_{\text{red}} \approx 1$, with most of the entries having $\chi^2_{\text{red}} < 15$. The tail of the distribution contains entries from the presumably poorly approximated bins, however, the number of those entries is quite small, which is a desirable property for the approximation procedure.

The p-values for the χ^2 tests for the test samples for the `Herwig7` MCEG are calculated and are shown in figure 17. Both the χ^2_{red} and p-values indicate potential issues with the approximation for certain bins, however, they do not provide a quantitative goodness of fit result. Hence, bins of observables with low p-values and large χ^2_{red} values were further inspected manually using the test samples, e.g. considering distributions similar to those shown in figure 1.

For most of those bins, the approximated values are compatible with the MC values if the variances introduced by the choice of the default parameters are taken into account. However, due

2023 JINST 18 P10033

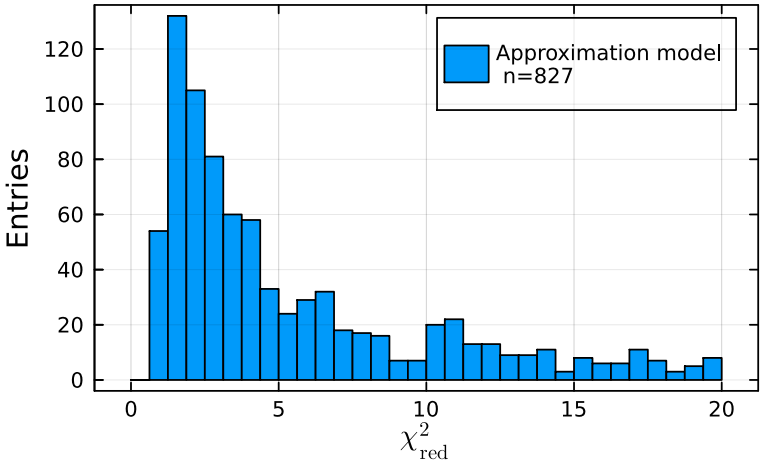

Figure 16. Distribution of χ^2_{red} for all the approximated observables and bins. Each bin of each observable contributes one value to the distribution.

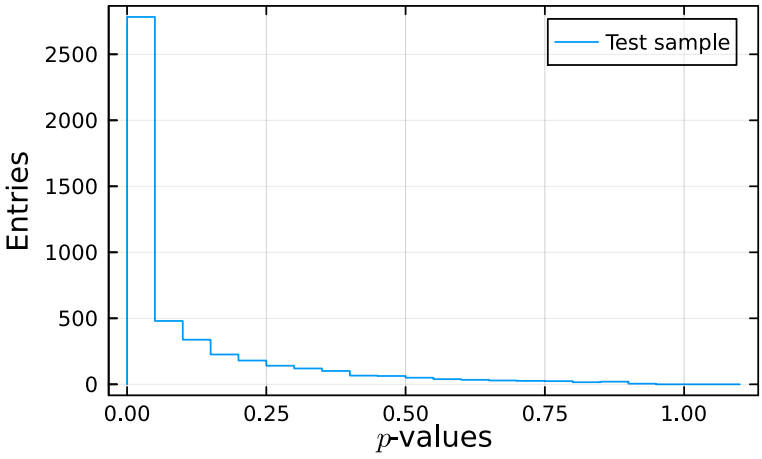

Figure 17. Calculated p-values for the χ^2 tests on the test samples evaluated for the `Herwig7` MCEG.

to the limited number of parameters used in the approximation it is not possible to reproduce the MC samples exactly.

References

[1] A. Buckley et al., *General-purpose event generators for LHC physics*, *Phys. Rept.* **504** (2011) 145 [arXiv:1101.2599].

[2] CDF collaboration, *High-precision measurement of the W boson mass with the CDF II detector*, *Science* **376** (2022) 170.

2023 JINST 18 P10033

[3] ATLAS collaboration, *Measurement of the top quark mass in the $t\bar{t} \to$ lepton+jets channel from $\sqrt{s} = 8$ TeV ATLAS data and combination with previous results*, Eur. Phys. J. C **79** (2019) 290 [arXiv:1810.01772].

[4] K. Cranmer, J. Brehmer and G. Louppe, *The frontier of simulation-based inference*, Proc. Nat. Acad. Sci. **117** (2020) 30055 [arXiv:1911.01429].

[5] DELPHI collaboration, *Tuning and test of fragmentation models based on identified particles and precision event shape data*, Z. Phys. C **73** (1996) 11.

[6] C. Bierlich et al., *Robust Independent Validation of Experiment and Theory: Rivet version 3*, SciPost Phys. **8** (2020) 026 [arXiv:1912.05451].

[7] A. Buckley et al., *Systematic event generator tuning for the LHC*, Eur. Phys. J. C **65** (2010) 331 [arXiv:0907.2973].

[8] W. Wang et al., *BROOD: Bilevel and Robust Optimization and Outlier Detection for Efficient Tuning of High-Energy Physics Event Generators*, SciPost Phys. Core **5** (2022) 001 [arXiv:2103.05751].

[9] M. Krishnamoorthy et al., *Apprentice for Event Generator Tuning*, EPJ Web Conf. **251** (2021) 03060 [arXiv:2103.05748].

[10] J. Bellm and L. Gellersen, *High dimensional parameter tuning for event generators*, Eur. Phys. J. C **80** (2020) 54 [arXiv:1908.10811].

[11] P. Ilten, M. Williams and Y. Yang, *Event generator tuning using Bayesian optimization*, 2017 JINST **12** P04028 [arXiv:1610.08328].

[12] A. Andreassen and B. Nachman, *Neural Networks for Full Phase-space Reweighting and Parameter Tuning*, Phys. Rev. D **101** (2020) 091901 [arXiv:1907.08209].

[13] M. Lazzarin, S. Alioli and S. Carrazza, *MCNNTUNES: Tuning Shower Monte Carlo generators with machine learning*, Comput. Phys. Commun. **263** (2021) 107908 [arXiv:2010.02213].

[14] P. Skands, S. Carrazza and J. Rojo, *Tuning PYTHIA 8.1: the Monash 2013 Tune*, Eur. Phys. J. C **74** (2014) 3024 [arXiv:1404.5630].

[15] N. Fischer, S. Gieseke, S. Plätzer and P. Skands, *Revisiting radiation patterns in e^+e^- collisions*, Eur. Phys. J. C **74** (2014) 2831 [arXiv:1402.3186].

[16] F. Klimpel, *Improved tuning methods for Monte Carlo generators*, M.Sc. thesis, Max Planck Institute, Munich, Germany (2018) [arXiv:1801.07187].

[17] ALEPH collaboration, *Studies of quantum chromodynamics with the ALEPH detector*, Phys. Rept. **294** (1998) 1.

[18] ALEPH collaboration, *Study of the fragmentation of b quarks into B mesons at the Z peak*, Phys. Lett. B **512** (2001) 30 [hep-ex/0106051].

[19] JADE and OPAL collaborations, *QCD analyses and determinations of α_S in e^+e^- annihilation at energies between 35 GeV and 189 GeV*, Eur. Phys. J. C **17** (2000) 19 [hep-ex/0001055].

[20] Particle Data Group collaboration, *Review of Particle Physics*, Phys. Lett. B **667** (2008) 1.

[21] J. Bellm et al., *Herwig 7.0/Herwig++ 3.0 release note*, Eur. Phys. J. C **76** (2016) 196 [arXiv:1512.01178].

[22] K. Hamilton and P. Nason, *Improving NLO-parton shower matched simulations with higher order matrix elements*, JHEP **06** (2010) 039 [arXiv:1004.1764].

[23] J. Alwall et al., *MadGraph 5: Going Beyond*, JHEP **06** (2011) 128 [arXiv:1106.0522].

2023 JINST 18 P10033

[24] F. Cascioli, P. Maierhofer and S. Pozzorini, *Scattering Amplitudes with Open Loops*, *Phys. Rev. Lett.* **108** (2012) 111601 [arXiv:1111.5206].

[25] A. Buckley et al., *The HepMC3 event record library for Monte Carlo event generators*, *Comput. Phys. Commun.* **260** (2021) 107310 [arXiv:1912.08005].

[26] B.R. Webber, *A QCD Model for Jet Fragmentation Including Soft Gluon Interference*, *Nucl. Phys. B* **238** (1984) 492.

[27] M. Bahr et al., *Herwig++ Physics and Manual*, *Eur. Phys. J. C* **58** (2008) 639 [arXiv:0803.0883].

[28] C. Bierlich et al., *A comprehensive guide to the physics and usage of PYTHIA 8.3*, arXiv:2203.11601 [DOI:10.21468/SciPostPhysCodeb.8].

[29] ThePSI developers, *ThePSI: an interface between the Pythia8 and the ThePEG toolkit*, https://gitlab.cern.ch/ThePSI/ThePSI.

[30] O. Schulz et al., *BAT.jl: A Julia-Based Tool for Bayesian Inference*, *SN Comput. Sci.* **2** (2021) 1 [arXiv:2008.03132].

[31] N. Castro et al., *EFTfitter — A tool for interpreting measurements in the context of effective field theories*, *Eur. Phys. J. C* **76** (2016) 432 [arXiv:1605.05585].

[32] A. Gelman and D.B. Rubin, *Inference from Iterative Simulation Using Multiple Sequences*, *Statist. Sci.* **7** (1992) 457.

[33] S.P. Brooks and A. Gelman, *General Methods for Monitoring Convergence of Iterative Simulations*, *J. Comput. Graphical Stat.* **7** (1998) 434.

[34] LsqFit.jl developers, *LsqFit.jl*, https://github.com/JuliaNLSolvers/LsqFit.jl.

[35] D.W. Marquardt, *An Algorithm for Least-Squares Estimation of Nonlinear Parameters*, *J. Soc. Ind. Appl. Math.* **11** (1963) 431.

[36] K. Levenberg, *A method for the solution of certain non-linear problems in least squares*, *Q. Appl. Math.* **2** (1944) 164.

[37] I. Antcheva et al., *ROOT: A C++ framework for petabyte data storage, statistical analysis and visualization*, *Comput. Phys. Commun.* **182** (2011) 1384.

[38] V. Blobel and E. Lohrmann, *Statistische und numerische Methoden der Datenanalyse*, Vieweg+Teubner Verlag (1998) [DOI:10.1007/978-3-663-05690-4].

2023 JINST 18 P10033