Suplemento de la Revista Mexicana de Física **3** 040902 (2022) 1–6

A tale of two jets

Christine Nattrass, Antonio Da Silva, Patrick Steffanic, Charles Hughes *University of Tennessee, Knoxville, 1408 Circle Drive, Knoxville TN 37996-1200.*

Received 3 July 2022; accepted 15 September 2022

We used model studies to investigate approaches to distin-guish signal and combinatorial jets.

Keywords: Jets

1 Introduction

The standard paradigm To date, most measurements of jets in heavy ion collisions impose a standard paradigm for separation of the signal from the background, shown schematically in fig. 1. There is a signal from particles produced in a single hard scattering and a background from all other processes. That background possibly includes particles from other hard scatterings. This background leads to two types of background which must be suppressed and/or subtracted

1. **Combinatorial jets** - jets whose constituents were not created by the same, or even correlated, processes, but which are simply spatially correlated.

2. **Background particles in a signal jet** - particles grouped into a signal jet which were not produced by the primary hard scattering.

These sources of background are the natural consequence of jet finders which group all particles in an event into a jet.

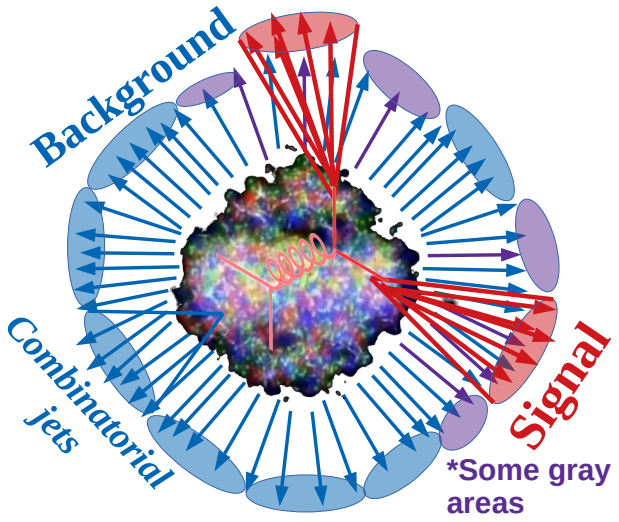

FIGURE 1. Schematic diagram showing standard paradigm for jet background subtraction. Particles and jets are assumed to be clearly classified as signal or background. In reality, some ambiguity exists.

This approach usually elides ambiguities in the classification of particles or jets as signal or combinatorial jets, but such ambiguities exist both experimentally and theoretically. If a hard parton scatters off a medium parton, that medium parton can carry some of the hard parton's original momentum and become spatially correlated with it. The resulting hadrons would be indistinguishable experimentally from particles formed exclusively from fragmentation and hadronization of the original hard parton. Experimentally, heuristic procedures are developed to suppress and subtract background contributions, but possible biases imposed by these approaches as well as how robust they are to partonic energy loss are frequently uninterrogated.

The resolution to such ambiguities in particle physics was the Snowmass Accord [1], which avoided heuristic approaches designed to get the expected result, leading to a definition of a jet as, essentially, what a jet finder finds. A rigorous approach following the Snowmass Accord is to use the same jet finding algorithm with the same parameters on signal and background in order to ensure that the two are comparable. In heavy ion collisions, this would mean applying the background suppression and subtraction techniques applied to data to model calculations. This would guarantee that the comparison is valid, but this approach has not yet been widely implemented or accepted in the heavy ion community.

Approaches to background Combinatorial jets lead to contamination of the measured signal by processes unrelated to hard scatterings. Background particles in signal jets lead the measured energy to be systematically higher than true energy; while the average contribution from these particles can be subtracted, fluctuations in the number of background particles and their energies smears the reconstructed energy. Contributions from combinatorial jets are suppressed and corrected for by

1. focusing on high momentum jets, where such contributions lead to a negligible contribution,

2. suppressing combinatorial jets by requiring a high momentum hadron [2] or otherwise requiring that they contain high momentum constituents [3], and/or

3. subtracting any residual contribution, e.g. through unfolding [4] or estimating the contribution using a technique such as mixed events [5].

The fluctuations in the jet energy due to background particles are suppressed and corrected for by

1. restricting jet constituents to high momentum particles,
2. correcting for the residual energy resolution using unfolding, and/or
3. providing a response matrix so that it is possible to smear a theoretical calculation rather than correct an experimental measurement for smearing.

Possible biases imposed by these approaches often are corrected for using unfolding in combination with simulations, leaving the results potentially sensitive to model-dependent corrections.

2 Models

We have done a number of model studies to evaluate the impact of background suppression and subtraction techniques and evaluate how robust the assumptions behind them are. To do this we use two primary models, TennGen and PYTHIA.

We use TennGen [6, 7] for a realistic background Pb+Pb event at $\sqrt{s_{\rm NN}}$ = 2.76 TeV with correlations due to flow but no other physics correlations. TennGen uses fits of ALICE single particle spectra to a Boltzmann Gibbs Blast Wave function [8, 9] and the measured single particle azimuthal anisotropies [10] to throw a background roughly matching the data. By construction, this background has no contribution from hard scatterings.

We embed a PYTHIA 6 p+p event produced with the Perugia 2011c [11] tune at $\sqrt{s}$ = 2.76 TeV in the TennGen event to generate a jet signal. We cluster the combined event with the anti-k_T jet finder, producing a population of jet candidates with particles from both PYTHIA and TennGen.

3 Distinguishing signal and combinatorial jets

We can classify jet candidates as combinatorial or signal by how much momentum is from TennGen and PYTHIA particles. Jets which contain only particles from TennGen are classified as combinatorial. Jets which include 80% of the $p_T^{hard,min}$ used in PYTHIA are considered signal jets. All other jets are classified as "squishy," as they are difficult to unambiguously classify as signal or combinatoria. The majority of squishy jets are most likely more combinatorial, but some contain significant contributions from hard processes.

We then use four variables to characterize what these jets look like

1. **Area (A)** - determined by adding low momentum "ghost" particles to the event. The area is proportional to the number of ghost particles and is a measure of how large the jet is. High momentum anti-k_T jets will generally have an area near $A = \pi R^2$ where R is the resolution parameter of the jet, but lower momentum jets in particular are less likely to be conical and may consist of only a few particles. Since gluon-like jets tend to fragment into softer particles, any selection on the area could bias the sample of jets.
2. **Angularity [12] (λ_1^1)** - given by
$$\lambda_1^1 = \sum_{i=1}^{N} z_i \cdot \frac{\Delta R_{i,jet}}{R} \quad (1)$$
where N is the number of constituents in the jet, z_i is the momentum fraction carried by constituent i, and $\Delta R_{i,jet}$ is the distance in η-ϕ space between constituent i and the jet axis. This is a measure of how far constituents are from the jet axis on average. Generally signal jets should have a lower λ_1^1, but gluon-like jets are broader on average than quark-like jets, so any selection on λ_1^1 is likely to impose a bias.
3. **Mean p_T ($\langle p_T \rangle$)** - is the average momentum of all constituents, including both TennGen and PYTHIA particles.
4. **Leading hadron p_T (p_T^1)** - is the highest momentum hadron in the jet. Since quark-like jets tend to fragment harder, selecting on this this may lead to a bias in the surviving jets.

This collection of jet properties aims to both describe the complete properties of a jet and to be realistically applicable to jets reconstructed in data. The distribution of these properties for real, combinatorial, and squishy jets is shown in fig. 2 for anti-k_T jets with $R = 0.4$ for a range of $p_T^{hard,min}$. There is a clear region at low areas where there are primarily combinatorial jets, although for jets with $p_T^{hard,min} < 20$ GeV/c there is a significant contribution from signal jets. There is some separation between signal and combinatorial jets for low angularity and high p_T^1, while there is little separation for $\langle p_T \rangle$. We therefore restrict jets to $A < 0.6\pi R^2$, which matches the procedure used by ALICE [2] for further studies. At lower momenta, this selection may induce a bias at lower momenta, as the signal jets removed likely have different properties from those which are retained.

After the area selection, the surviving population of combinatorial jets look nearly indistinguishable from signal jets for angularity, $\langle p_T \rangle$, and p_T^1. There is still some separation between signal and combinatorial jets, but there is significant overlap, particularly for jets with $p_T^{hard,min} < 20$ GeV/c. For p_T^1, there is a clear separation between signal and combinatorial jets, but there are significant populations of signal jets for all p_T^1 so any restriction which removes combinatorial jets would also remove signal jets. Furthermore, this would have a physics bias, as quark jets fragment harder than gluon jets [13, 14].

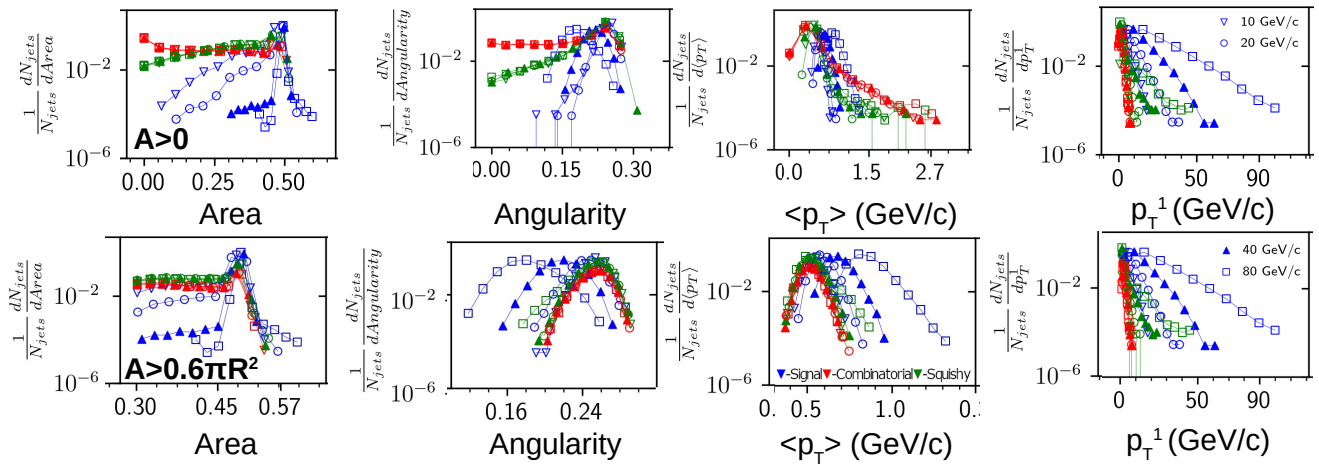

FIGURE 2. Distribution of jet area A, angularity, $\langle p_T \rangle$, and p_T^1 for real, combinatorial, and squishy jets before (upper) and after (lower) a selection of $A > 0.6\pi R^2$ for anti-k_T jets with $R = 0.4$.

4 Silhouette values

In order to investigate whether there is a separation between these populations of jets, we investigate the distributions of silhouette values. The silhouette value for the jet with index i is given by

$$S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{2}$$

The $a(i)$ are the mean in-class distance,

$$a(i) = \frac{1}{N_I - 1} \sum_{j \neq i}^{N_I} d_{i,j} \tag{3}$$

where the sum runs over jets of the same class as jet i, N_I is the number of jets in the same class as jet i, d^{max} is the largest distance between any two jets, and d^{min} is the smallest distance between any two jets. The $b(i)$ are the mean out-of-class distance, similar to eq. 3. This is typically the Euclidean distance computed in feature space, but instead we use

$$d(i,j) = \sqrt{\sum_{k=1}^{4} \left(\frac{x_j - x_i}{x_{max} - x_{min}} \right)^2} \tag{4}$$

where the x represent the four observables used (A, λ_1^1, $\langle p_T \rangle$, and p_T^1) and the x_{max} and x_{min} represent the maximum and minimum, respectively, of each of these observables. The normalized distances are used because the units and scales of each observable are different.

By construction, $-1 < S(i) < 1$. Positive values indicate that the jet is more like others in its group; for instance, a given combinatorial jet is more like other combinatorial jets than signal jets. Negative values indicate that the jet is more like jets in the other group; for instance, a combinatorial jet is more like a signal jet than other combinatorial jets. A silhouette value of zero indicates that the jet is equally close to both its own group and the other group. The strength of silhouette values is that they are sensitive to multiple variables at once. Silhouette values are more sensitive to multivariable correlations which may enable the separation of signal and background jets than single variables.

The distribution of silhouette values for real and combinatorial jets is shown in fig. 3 for anti-k_T jets with $R = 0.4$ for a range of $p_T^{hard,min}$, both before and after the area selection. The silhouette values cannot be defined for squishy jets, as it is unclear which group they belong in. Since silhouette values are normalized to measure the difference between the two groups in the sample, the silhouette values in different panels in fig. 3 are not directly comparable to each other. For all samples, most signal jets have positive silhouette values, meaning that these look more like signal jets than combinatorial jets. There is always a significant population of combinatorial jets with negative silhouette values, indicating that these jets are more similar to signal jets than background jets. The fraction of combinatorial jets with negative silhouette values increases with decreasing jet momenta. This makes intuitive sense, as high momentum jets stand out more from the background.

At low momenta, the overlap between signal and combinatorial jets is significant. This means that combinatorial jets are indistinguishable from signal jets, even in our controlled simulations where we can clearly and unambiguously define these samples. The ALICE collaboration found that unfolding for the jet momentum resolution was not stable without an additional selection on p_T^1. Our studies here indicate that there is no hidden selection which overcomes the need for additional selection on variables such as p_T^1 which inherently bias the surviving jet population. Realistically, therefore, some kinematic selections which would impose a bias are necessary. This ought to be considered carefully in the interpretation of the results. Furthermore, in many cases, a residual combinatorial jet contribution is corrected for using unfolding; this would add an inherent but subtle model dependence in the way that these corrections are done.

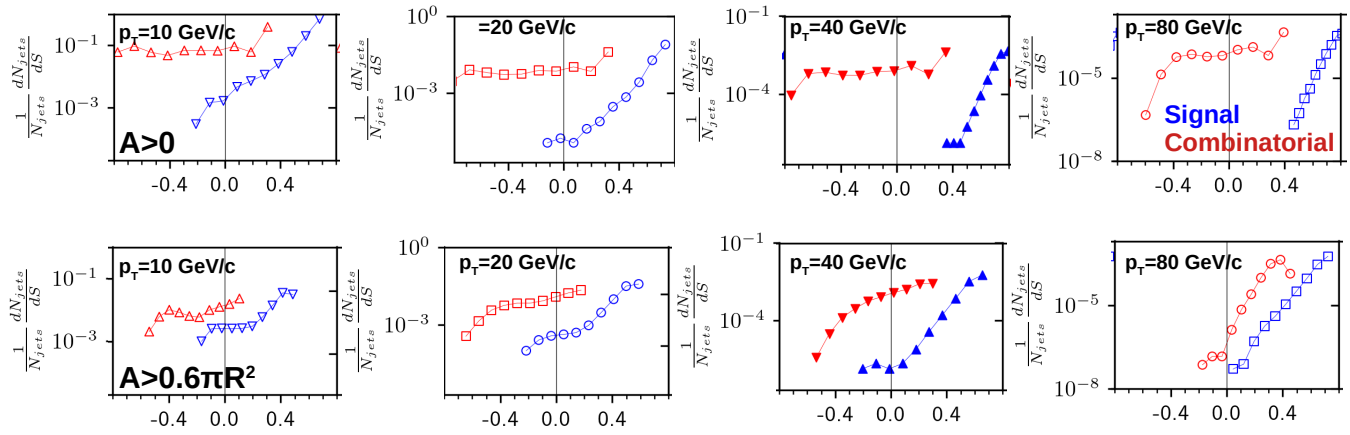

FIGURE 3. Distribution of silhouette values for real and combinatorial jets before (upper) and after (lower) a selection of $A > 0.6\pi R^2$ for anti-k_T jets with $R = 0.4$ for various $p_T^{hard,min}$.

5 Unfolding

Experiments correct for distortions in the jet momentum due to the migration of jets from their correct momentum bin to another bin using a procedure called unfolding [15, 16]. Given the results in sec. 3 and sec. 4 indicating that some bias may, indeed, be a necessary consequence of kinematic selections which distinguish between signal and combinatorial jets, we investigate how robust unfolding is.

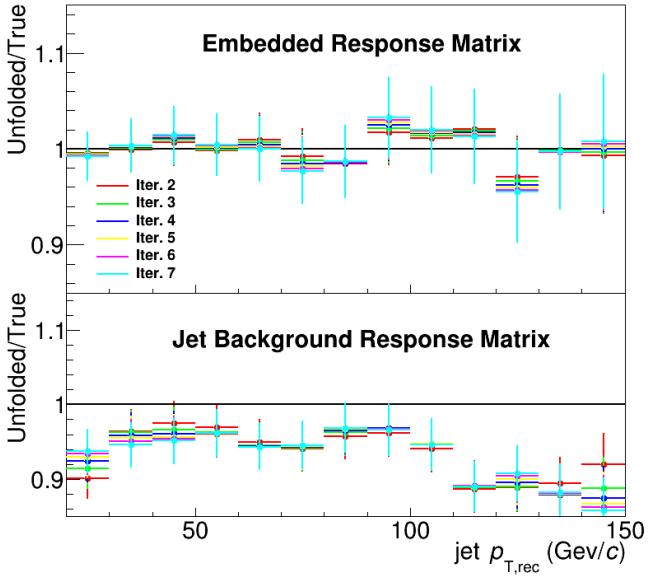

FIGURE 4. Comparison of the unfolded result to the true distribution for an embedding response matrix (upper) and a fluctuation-only response matrix (lower).

Momentum smearing arises due to both detector effects and jet background fluctuations. Usually, a response matrix is determined from a full simulation of the detector response to a jet, often with a simulated p+p collision embedded in either a real or a simulated Pb+Pb collision. An alternate approach is to factorize the response matrix into a detector response, constructed with a p+p collision, and smearing in the jet energy resolution from background fluctuations [2]. We use both approaches, using PYTHIA Angantyr for the Pb+Pb event.

Two sets of jets are reconstructed, the generated distribution, which contains only charged particles from the p+p event, and the smeared distribution, using all charged particles in either event. We do not include the impact of the finite single track reconstruction efficiency.

An unfolded distribution has to converge to the true distribution in order to demonstrate closure. We use the jets in the combined Pb+Pb plus p+p event which were matched to a jet in the p+p event and unfold this transverse momentum spectrum using Bayesian unfolding implemented in the RooUnfold [17] package. We compare this to the transverse momentum spectrum in PYTHIA p+p events. Results in fig. 4 demonstrate that closure is only achieved for the embedding response matrix. We note that this is consistent with previous studies indicating the two approaches were consistent, but those studies had limited statistics. This indicates that full embedding is required to achieve closure for Monte Carlo models which simulate the full event.

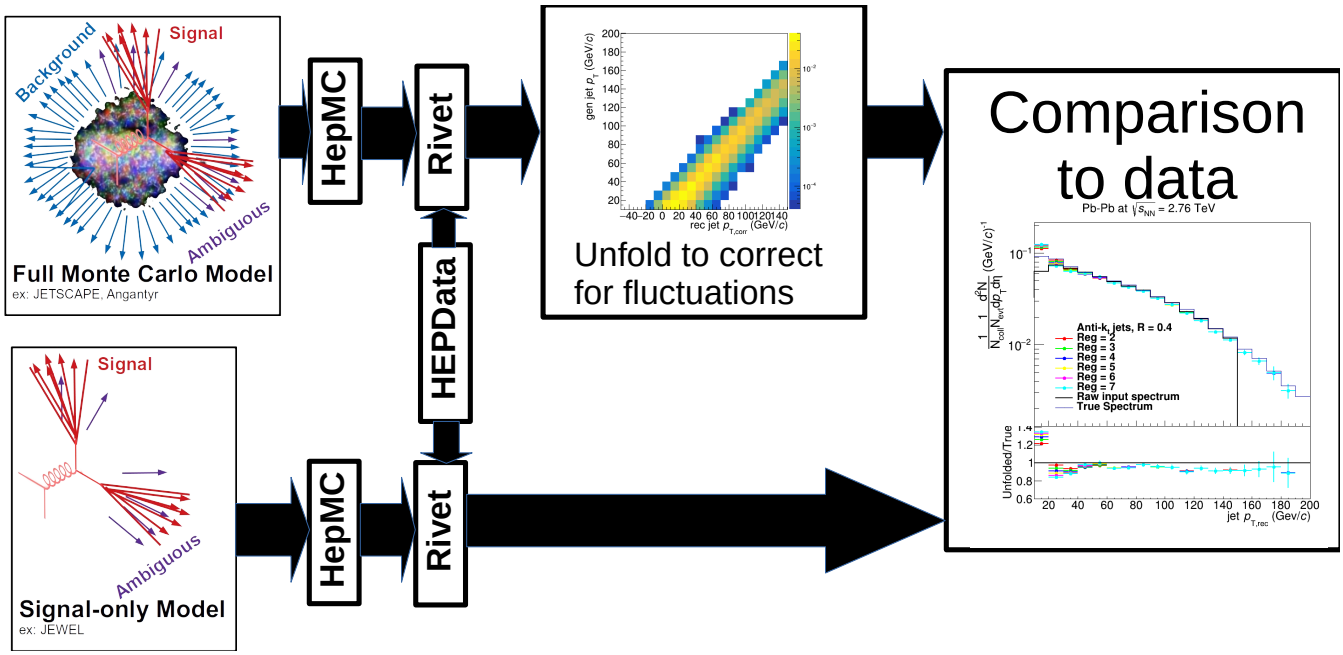

FIGURE 5. Distribution of silhouette values for real and combinatorial jets before (upper) and after (lower) a selection of $A > 0.6\pi R^2$ for anti-k_T jets with $R = 0.4$ for various $p_T^{hard,min}$.

6 Conclusions

We used model studies to investigate approaches to distinguish signal and combinatorial jets. The results in sec. 3 and sec. 4 demonstrate that some combinatorial jets are indistinguishable from signal jets, which indicates that the suppression of combinatorial jets at lower momenta requires kinematic selections which may bias the results. Of the four variables investigated to distinguish between signal and combiantorial jets, selected because they could actually be applied to data, only p_T^1 is relatively robust to a large background. Therefore, the ideal approach would involve applying these selections in a Monte Carlo which generates the full event so that model calculations faithfully reproduce any biases. Section 5 demonstrates that to achieve closure, a full Monte Carlo calculation requires unfolding using a response matrix constructed with embedding.

We therefore propose the scheme in fig. 5 to ensure that data and models are comparable. The RIVET [18] (Robust Independent Validation of Experiment and Theory) framework provides a generalized software framework for facilitating these comparisons in a systematic way. Analyses implemented in this framework can follow the procedure in the paper, including background subtraction. For models simulating the entire event, as opposed to models which only simulate the jet, an additional step constructing a response matrix and unfolding to correct for fluctuations in the background is necessary. This avoids many of the problems with separating signal and combinatorial background, as any biases or residual contamination

7 Acknowledgements

This work was supported in part by funding from the Division of Nuclear Physics of the U.S. Department of Energy under Grant No. DE-FG02-96ER40982 and from the National Science Foundation under Grant No. OAC-1550300. We also acknowledge support from the UTK and ORNL Joint Institute for Computational Sciences Advanced Computing Facility.

8 References

1. J. E. Huth et al., Toward a standardization of jet definitions, In 1990 DPF Summer Study on High-energy Physics: Research Directions for the Decade (Snowmass 90) Snowmass, Colorado, June 25-July 13, 1990 (1990) pp. 0134–136, URL http://lss.fnal.gov/cgi-bin/find_paper.pl?conf-90-249.
2. J. Adam et al., Measurement of jet suppression in central Pb-Pb collisions at $\sqrt{s_{\rm NN}}$ = 2.76 TeV, Phys. Lett. B 746 (2015) 1,

10.1016/j.physletb.2015.04.039

3. G. Aad et al., Measurements of the Nuclear Modification Factor for Jets in Pb+Pb Collisions at $\sqrt{s_{\rm NN}} = 2.76$ TeV with the ATLAS Detector, Phys.Rev.Lett. 114 (2015) 072302, 10.1103/PhysRevLett.114.072302

4. J. Adam et al., Azimuthal anisotropy of charged jet production in $\sqrt{s_{\rm NN}}$ = 2.76 TeV Pb-Pb collisions, Phys. Lett. B 753 (2016) 511, 10.1016/j.physletb.2015.12.047

5. L. Adamczyk et al., Measurements of jet quenching with semi-inclusive hadron+jet distributions in Au+Au collisions at $\sqrt{s_{NN}}$ = 200 GeV, Phys. Rev. C 96 (2017) 024905, 10.1103/PhysRevC.96.024905

6. C. Hughes, A. C. Oliveira Da Silva, and C. Nattrass, Model studies of fluctuations in the background for jets in heavy ion collisions (2020)

7. C. Hughes and C. Nattrass, TennGen Source Code and Documentation, https://github.com/chughes90/TennGen, Accessed: 2020-05-05.

8. O. Ristea, et al., Study of the freeze-out process in heavy ion collisions at relativistic energies, J. Phys. Conf. Ser. 420 (2013) 012041, 10.1088/1742-6596/420/1/012041

9. E. Schnedermann, J. Sollfrank, and U. W. Heinz, Thermal phenomenology of hadrons from 200-A/GeV S+S collisions, Phys. Rev. C 48 (1993) 2462, 10.1103/PhysRevC.48.2462

10. J. Adam et al., Higher harmonic flow coefficients of identified hadrons in Pb-Pb collisions at $\sqrt{s_{\rm NN}}$ = 2.76 TeV, JHEP 09 (2016) 164, 10.1007/JHEP09(2016)164

11. P. Z. Skands, Tuning Monte Carlo Generators: The Perugia Tunes, Phys. Rev. D 82 (2010) 074018, 10.1103/PhysRevD.82.074018

12. D. Reichelt, et al., Phenomenology of jet angularities at the LHC, JHEP 03 (2022) 131, 10.1007/JHEP03(2022)131

13. P. Abreu et al., Energy dependence of the differences between the quark and gluon jet fragmentation, Z. Phys. C70 (1996) 179, 10.1007/s002880050095

14. R. Akers et al., A Model independent measurement of quark and gluon jet properties and differences, Z. Phys. C68 (1995) 179

15. G. D'Agostini, Improved iterative Bayesian unfolding, In Alliance Workshop on Unfolding and Data Correction (2010) .

16. A. Hocker and V. Kartvelishvili, SVD approach to data unfolding, Nucl. Instrum. Meth. A 372 (1996) 469, 10.1016/0168-9002(95)01478-0

17. T. Adye, Unfolding algorithms and tests using RooUnfold, In PHYSTAT 2011 (CERN, Geneva, 2011) pp. 313–318, 10.5170/CERN-2011-006.313.

18. C. Bierlich et al., Confronting experimental data with heavy-ion models: RIVET for heavy ions, Eur. Phys. J. C 80 (2020) 485, 10.1140/epjc/s10052-020-8033-4