DEUTSCHES ELEKTRONEN-SYNCHROTRON
Ein Forschungszentrum der Helmholtz-Gemeinschaft

DESY 22-077
MIT-CTP 5434
Nikhef 22-005
arXiv:2205.06818
May 2022

Power Counting Energy Flow Polynomials

P. Cal

Deutsches Elektronen-Synchrotron DESY, Hamburg

and

Institute for Theoretical Physics Amsterdam and Delta Institute for Theorectical Physics, University of Amsterdam, The Netherlands

and

Nikhef, Amsterdam, The Netherlands

J. Thaler

Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, USA

and

The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, Cambridge, USA

W. J. Waalewijn

Institute for Theoretical Physics Amsterdam and Delta Institute for Theorectical Physics, University of Amsterdam, The Netherlands

and

Nikhef, Amsterdam, The Netherlands

ISSN 0418–9833

NOTKESTRASSE 85 – 22607 HAMBURG

Herausgeber und Vertrieb:

Verlag Deutsches Elektronen-Synchrotron DESY

DESY Bibliothek
Notkestr. 85
22607 Hamburg
Germany

DESY-22-077
MIT-CTP 5434
Nikhef 22-005

Prepared for submission to JHEP

Power Counting Energy Flow Polynomials

Pedro Cal,[a,b,c] **Jesse Thaler,**[d,e] **Wouter J. Waalewijn**[b,c]

[a] *Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany*
[b] *Institute for Theoretical Physics Amsterdam and Delta Institute for Theoretical Physics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands*
[c] *Nikhef, Theory Group, Science Park 105, 1098 XG, Amsterdam, The Netherlands*
[d] *Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[e] *The NSF AI Institute for Artificial Intelligence and Fundamental Interactions*

E-mail: pedro.cal@desy.de, jthaler@mit.edu, w.j.waalewijn@uva.nl

Abstract: Power counting is a systematic strategy for organizing collider observables and their associated theoretical calculations. In this paper, we use power counting to characterize a class of jet substructure observables called energy flow polynomials (EFPs). EFPs provide an overcomplete linear basis for infrared-and-collinear safe jet observables, but it is known that in practice, a small subset of EFPs is often sufficient for specific jet analysis tasks. By applying power counting arguments, we obtain linear relationships between EFPs that hold for quark and gluon jets to a specific order in the power counting. We test these relations in the parton shower generator Pythia, finding excellent agreement. Power counting allows us to truncate the basis of EFPs without affecting performance, which we corroborate through a study of quark-gluon tagging and regression.

Contents

1 Introduction

Most collisions at the Large Hadron Collider (LHC) involve jets. The patterns exhibited by these collimated sprays of hadrons—known as a jet's *substructure*—reveal important information about the partons initiating jet formation and about the underlying hard scattering process. In the early days of jet substructure, tailored observables were developed to probe jets in specific theoretically-motivated ways [1–9]. In recent years, the focus has shifted towards using machine learning strategies, with the goal of exploiting all information available inside of jets [10–19]. This in turn raises the question about how jet information should be represented to ensure robust jet analyses, e.g. whether the inputs to a neural network should be a jet image, a sequence/set of particles, a clustering tree, a graph, or a collection of observables.

Degree		0	1	2	3	4	5	6
All EFPs (table 2)	by degree	1	1	3	8	23	66	212
	cumulative	1	2	5	13	36	102	314
Strongly-ordered basis (table 3)	by degree	1	1	2	4	7	12	22
	cumulative	1	2	4	8	15	27	49
2-collinear basis (table 5)	by degree	1	1	2	4	8	17	37
	cumulative	1	2	4	8	16	33	70
z^2-truncated basis (table 8)	by degree	1	1	3	5	9	13	20
	cumulative	1	2	5	10	19	32	52
Color-reduced basis (table 9)	by degree	1	1	2	4	8	16	34
	cumulative	1	2	4	8	16	32	66

Table 1: The number of all EFPs of/up to degree d, as well as the number of basis elements needed in the strongly ordered, 2-collinear, z^2-truncated, and color-reduced (1-collinear) expansion. For our studies, we restrict our attention to $d > 0$ EFPs.

In this paper, we use the technique of power counting to systematically organize jet observables from first principles. Power counting has already seen successes in identifying specific jet observables [20–22], and it forms the basis of precision calculations in soft-collinear effective theory [23–27]. Here, we apply power counting to study energy flow polynomials (EFPs) [28], which are an overcomplete linear basis for infrared-and-collinear safe jet substructure. EFPs have been used in a variety of machine learning tasks [29–36], so it is a natural context to study how best to represent jet information. By exploiting power counting, we show how to simplify the EFP basis for analysis tasks involving quark and gluon jets.

Each EFP is an N-point correlator on jets with angular degree d, which can be represented by a graph with N nodes and d edges. The set of all multigraphs corresponds to the set of all EFPs. While a complete basis of jet substructure observables is conceptually important, any practical application has to limit the basis in some way. For the related case of N-subjettiness [9, 37, 38], the convergence as a function of N was explored in refs. [39, 40] in the context of machine learning. For EFPs, a natural way to truncate the basis is to restrict the angular degree d of the EFPs, which corresponds to the limiting the total number of graph edges.

As we will show, even when restricting to fixed degree d, there is substantial redundancy between the EFPs when working to a certain level of approximation. To obtain these redundancy relations, we employ power counting, inspired by the pioneering work in ref. [20]. We consider several different schemes for performing the power counting, as summarized in table 1. These range from strongly-ordered emissions (typical of calculations at leading-logarithmic order) to an expansion in the number of collinear (energetic) emissions in which all angular correlations are kept. The EFP relations we obtain are valid for single prong jets, i.e. those initiated by a light quark or gluon. We test these EFP relations using the parton shower generator Pythia [41], finding reasonable to very good

agreement, depending on the choice of power counting scheme.

For any application of power counting, one has to decide the meaning of "leading power." This in turn corresponds to making an assumption about what form the optimal observable should take for a given task. We explore two different power counting assumptions in the body of the text:

- **Strongly-ordered expansion**: The emissions in the jet are assumed to be strongly ordered in both energy and angle.
- **Collinear expansion**: The jet is assumed to consist of collinear and collinear-soft emissions, and we expand in the number of collinear emissions, keeping all angular information.

We consider both the 1-collinear and 2-collinear expansion. The strongly-ordered expansion is a further expansion of the 1-collinear case. The expansions we consider are not directly related to the logarithmic accuracy of a calculation (which depends on the details of the observable). In general, leading-logarithmic (LL) accuracy lies between the strongly-ordered and 1-collinear expansion, while the 2-collinear expansion holds at next-to-leading logarithmic (NLL) order.[1]

In both the collinear and strongly-ordered expansions, we assume that the optimal observable for a given jet substructure task is well approximated by a single EFP, or by a sum of EFPs with no fine-tuned cancellations between terms. In practice, this means that we start from the full set of EFPs, using the redundancy relations to reduce the basis. Like with any linearly redundant system, the choice of EFP basis elements is not unique, with differences appearing beyond the chosen level of accuracy. We also explore an alternate scheme in appendix A:

- **Energy truncation**: The optimal observable for a given jet substructure task is assumed to have an expansion as a series in the momentum fractions z.

The energy expansion yields reasonable performance, but not much conceptual insight.

Power counting not only allows us to conceptually simplify the EFP basis, but in the 1-collinear case, it also reduces their computational cost. Naively, the complexity to compute an N-point EFP on M particles scales like $\mathcal{O}(M^N)$. Using variable elimination, this can be reduced to $\mathcal{O}(M^t)$, where t is the tree-width of the graph representing the EFP [28].[2] In the limit of just one collinear emission, power counting allows us to further "cut open" the highest tree-width graphs and express them in terms of lower tree-width graphs. As discussed in appendix B, this yields computational gains with a modest decrease in machine learning performance.

To demonstrate that our reduced bases of EFPs perform as well as using all EFPs on single-prong jets, we carry out quark/gluon tagging and regression studies. While the

[1] Here we count logarithms in the cross section, $\int^{\mathcal{O}} \mathrm{d}\mathcal{O}' \, \mathrm{d}\sigma/\mathrm{d}\mathcal{O}' = \sum_{n,k} c_{n,k} \alpha_s^n L^k + \ldots$, where L are logarithms of the observables $\mathcal{O}$. In this counting, $k = 2n$ corresponds to LL, and $k = 2n - 1$ corresponds to NLL.

[2] For the special case where the angular distance can be expressed through an inner product, this can be further reduced to $\mathcal{O}(v^3 M)$ [42], where v is the maximum number of lines connecting to a single node.

collinear and strongly-ordered expansions differ substantially in the numerical accuracy of the relations derived by power counting, interestingly, their performance in the regression study is nearly identical. This suggests that alternative approaches to obtaining the relations between EFPs should be possible to get the correct coefficients at LL accuracy. Indeed, we demonstrate that much better expressions for EFPs in terms of the strongly-ordered basis can be obtained via linear regression.

The rest of this paper is organized as follows. In section 2, we discuss the power counting for EFPs and obtain relations between them that hold in the strongly-ordered expansion, as well as for 1 or 2 collinear particles. Using these relationships, we identify a reduced basis of EFPs in section 3 using the strongly-ordered, 1- and 2-collinear expansions. The energy truncation and corresponding basis is discussed in appendix A, with complete results archived at ref. [43]. An alternative, computationally advantageous, basis for the 1-collinear expansion is given in appendix B. We test the accuracy of these relations between EFPs in section 4 and explore the machine learning performance of the reduced EFP bases in section 5. We conclude in section 6 with a summary and outlook.

2 Power counting of energy flow polynomials

2.1 Review of energy flow polynomials

EFPs are N-point correlators on jets [28], and they are represented by a graph with N nodes. For each node, the energy fractions z_i of all particles i in a jet are summed over. The terms in these nested sums are weighted by the angles between each of the particles whose momentum fractions appear, where the exponent of θ_{ij} is equal to the number of lines between nodes contributing z_i and z_j:

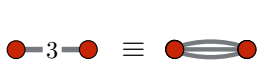
(2.1)

where M is the number of particles. The precise definition of z_i and θ_{jk} depends on the application of interest. For hadron colliders, it is typical to use:

$$z_i = \frac{p_{Ti}}{\sum_{j=1}^{M} p_{Tj}}, \qquad \theta_{jk} = \left(\Delta R_{jk}\right)^{\beta}, \tag{2.2}$$

where p_{Ti} is the transverse momentum of particle i, ΔR_{jk} is the angular distance between particles j and k on the rapidity-azimuth cylinder, and $\beta > 0$ is an angular weighting exponent. For the numerical studies in this paper, we use $\beta = 2$, though the power counting arguments are independent of β. Unlike for the angles, the energy fractions must have an exponent equal to one to ensure collinear safety.

As an example, the two-point correlator with three lines between the two nodes is given by:



(2.3)

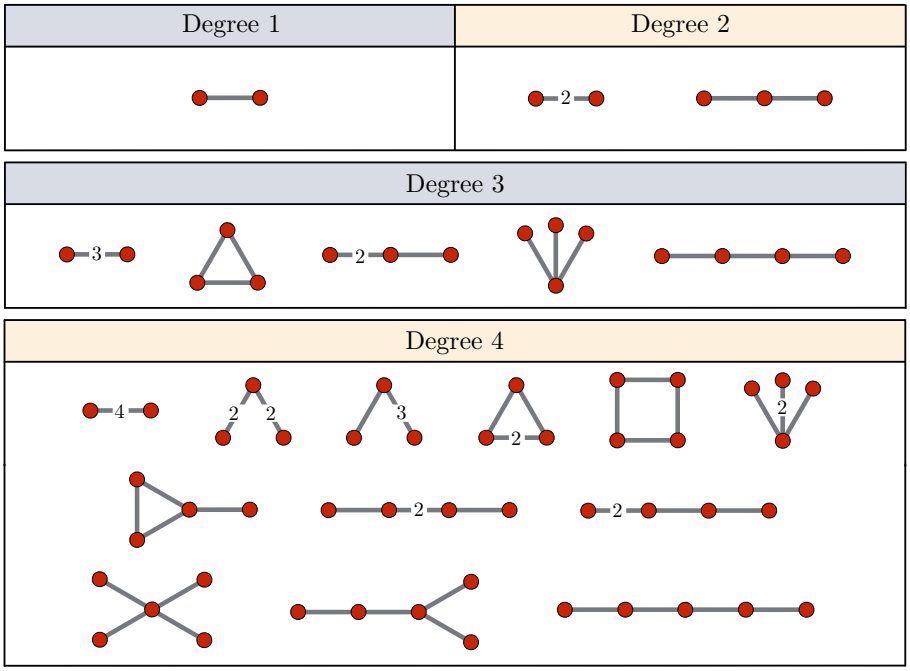


Table 2: All prime (connected) EFPs up to degree 4. For legibility, the numbers indicate the multiplicity of lines connecting the corresponding nodes. To form composite EFPs, one multiplies together these prime elements.

To simplify the graphs, we write the number of lines between nodes as a number, since drawing multiple lines between nodes becomes less legible for complicated EFPs. Two more non-trivial examples are:

$$\text{2}\!-\!\bullet\!-\!\bullet = \sum_{i_1=1}^{M}\sum_{i_2=1}^{M}\sum_{i_3=1}^{M}\sum_{i_4=1}^{M} z_{i_1} z_{i_2} z_{i_3} z_{i_4} \theta_{i_1 i_2} \theta_{i_2 i_3}^2 \theta_{i_3 i_4}\,,$$

$$\begin{array}{c}\bullet \\ \text{2}\!\diagdown \quad \diagup\!\text{3} \\ \bullet\end{array} = \sum_{i_1=1}^{M}\sum_{i_2=1}^{M}\sum_{i_3=1}^{M} z_{i_1} z_{i_2} z_{i_3} \theta_{i_1 i_2} \theta_{i_2 i_3}^3 \theta_{i_3 i_1}^2\,. \tag{2.4}$$

The prime (i.e. connected) EFPs up to degree 4 are shown in table 2. Composite EFPs can be written as the product of prime EFPs. The chromatic number of a graph is the number of colors needed to paint the nodes such that no two connected nodes have the same color. A single dot corresponds to the constant 1, which we omit in our EFP regression studies.

2.2 Power counting in the strongly-ordered expansion

In the strongly-ordered (SO) expansion, we consider emissions that are both collinear and soft, as depicted in figure 1. The jet thus consists of a hard parton ($i = 1$) and $M-1$

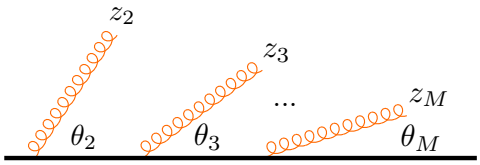


Figure 1: At strongly-ordered (SO) accuracy, all emissions (orange) in a jet are assumed to be collinear *and* soft, as well as strongly ordered. The parton initiating the jet (black) has momentum fraction $z_1 = 1 - \sum_{i=2}^{M} z_i \approx 1$ after the emissions.

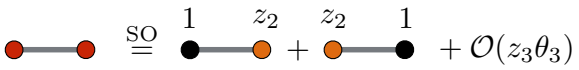
collinear-soft gluons ($i = 2, \ldots, M$), which are strongly ordered in energy and angle:

$$\textbf{Strongly-ordered expansion:} \quad z_{i+1} \ll z_i\,, \qquad \theta_{1,i+1} \ll \theta_{1,i} \text{ for } i > 1\,. \tag{2.5}$$

In this case, the squared matrix element, describing the probability of producing a certain jet, can be factorized into a product of matrix elements for $1 \to 2$ processes.

The measurement can also be simplified in the SO expansion. For example, for EFPs with bipartite graphs (i.e. with chromatic number 2), the observable will be dominated by the leading collinear-soft emission, so one can simplify the EFPs up to power corrections. This is closely related to LL accuracy in calculations, though there one only has strong ordering in a single variable; e.g. for the LL resummation of jet mass it is convenient to order emissions in their contribution to the mass. In the SO limit, we are effectively assuming simultaneous strong ordering for multiple variables.

We start by applying the power counting in eq. (2.5) to the simplest dumbbell EFP from eq. (2.3):



$$\overset{\text{SO}}{=} 2z_2\theta_2 + \mathcal{O}(z_3\theta_3). \tag{2.6}$$

Since the momentum fraction of the hard parton is larger than that of the collinear-soft, we want to focus on the corresponding terms in the sums arising from the nodes. However, the term in which both sums involve the hard parton vanishes, because it is weighted by a power of $\theta_{11} = 0$. The leading contribution thus arises from the hard parton (black) contribution in the sum of one of the nodes and a collinear-soft (orange) contribution from the other. There are two permutations, resulting in an overall factor of 2.

By using the expansion in eq. (2.5), more complicated EFPs can be written in terms of simpler building blocks. As an example, consider the four-dot EFP:

$$\overset{\text{SO}}{=} 2z_2^2\theta_2^3 + \mathcal{O}(z_2 z_3 \theta_2^2 \theta_3). \tag{2.7}$$

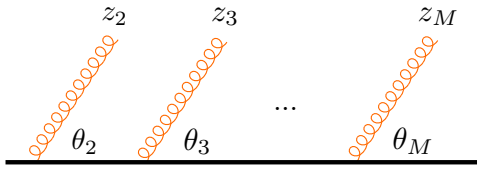
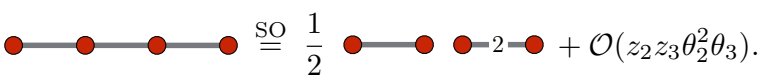


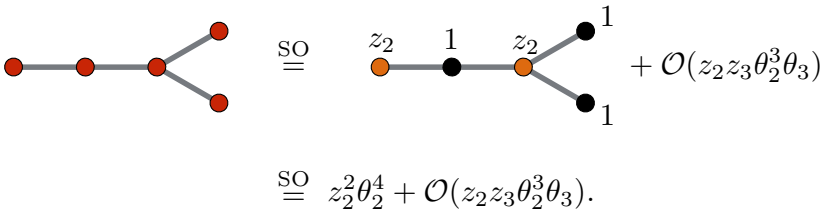
Figure 2: At 1-collinear (1c) accuracy, all emissions (orange) are assumed to be collinear and soft. Unlike the SO expansion, though, here the emissions are *not* strongly ordered, i.e. no assumptions are made on their relative energy and angular scalings.

By comparing to eq. (2.6), we see that the 4-dots EFP can be identified as a product of dumbbells. Therefore, we conclude that in the SO expansion:



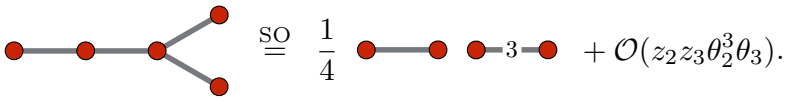
(2.8)

As another example, consider the crocodile EFP:



(2.9)

This leads to the relationship:



(2.10)

Using these results, we will derive the strongly-ordered EFP basis in section 3.1.

2.3 Power counting in the 1-collinear expansion

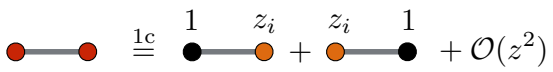
The relationships between EFPs obtained when assuming strong ordering in eq. (2.5) are not numerically accurate, as will be shown in section 4.1. We therefore explore a different power counting scheme that yields much more robust relationships.

In the n-collinear approximation, a jet consists of $n \geq 1$ collinear parton(s) and $M - n$ collinear-soft gluons, as shown in figure 2. In this case, we do not assume a hierarchy in angles, and we treat the energies of all collinear-soft gluons as being parametrically of the same size. For the 1-collinear approximation, we thus assume:

$$\textbf{1-collinear expansion:} \quad z_1 = 1 + \mathcal{O}(z)\,, \qquad z_i \sim z \ll 1 \text{ for } i > 1\,, \qquad \theta_{ij} \sim \theta \ll 1\,, \tag{2.11}$$

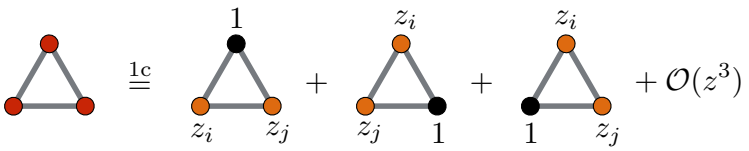
The strongly-ordered expansion in eq. (2.5) is therefore a further expansion of the 1-collinear case. Whereas the strongly-ordered expansion is worse than LL accurate, the 1-collinear expansion is better than LL accurate, since it includes no strong ordering of the collinear-soft emissions.

We start again by examining the dumbbell EFP, but this time applying the power counting in eq. (2.11):



$$\overset{1c}{=} 2\sum_{i=2}^{M} z_i\theta_i + \mathcal{O}(z^2). \tag{2.12}$$

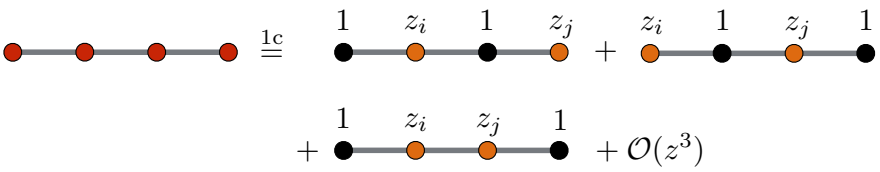
In contrast to eq. (2.6), the sum over all collinear-soft emissions is kept. Similarly, for the triangle EFP:



$$\overset{1c}{=} 3\sum_{i,j=2}^{M} z_i z_j \theta_i\theta_j\theta_{ij} + \mathcal{O}(z^3). \tag{2.13}$$

Since the graph is fully connected, the hard parton can only contribute in the sum of one of the nodes.

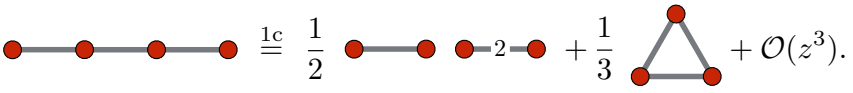
By using the expansion in eq. (2.11), more complicated EFPs can be written in terms of simpler building blocks. As an example, consider the four-dot EFP:



$$\overset{1c}{=} \sum_{i,j=2}^{M} z_i z_j (2\theta_i^2\theta_j + \theta_i\theta_j\theta_{ij}) + \mathcal{O}(z^3). \tag{2.14}$$

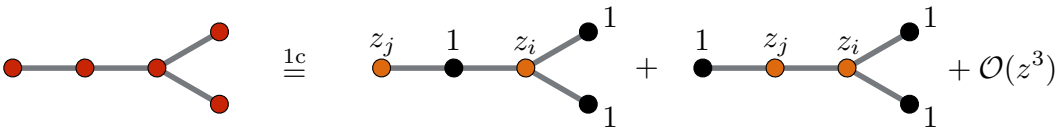
The sums in the first term of the expansion factorize and can be represented as a product of two EFPs. By comparing to eq. (2.12), these can be identified as dumbbells. The remaining term can be identified as the triangle EFP in eq. (2.13). Therefore we conclude

that at 1-collinear accuracy:



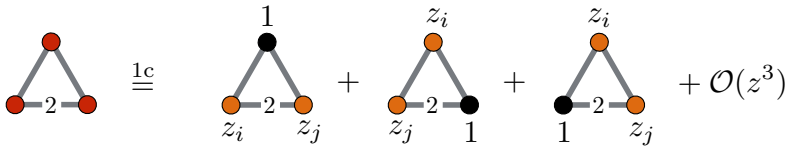
(2.15)

As another example, consider the crocodile EFP:



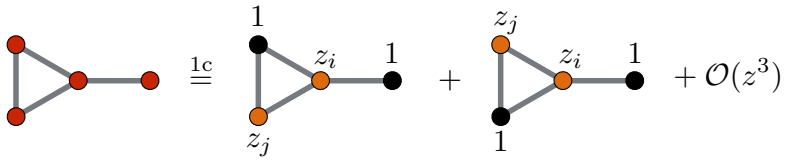
$$\overset{1c}{=} \sum_{i,j=2}^{M} z_i z_j (\theta_i^3 \theta_j + \theta_i^2 \theta_j \theta_{ij}) + \mathcal{O}(z^3). \tag{2.16}$$

The first term can directly be identified as a product of dumbbells. The second term looks like the triangle in eq. (2.13) except that $\theta_i \to \theta_i^2$. This suggests considering a triangle with two lines on one side:



$$\overset{1c}{=} \sum_{i,j=2}^{M} z_i z_j (\theta_i \theta_j \theta_{ij}^2 + 2\theta_i^2 \theta_j \theta_{ij}) + \mathcal{O}(z^3). \tag{2.17}$$

This, however, also produces an unwanted $\theta_i \theta_j \theta_{ij}^2$ term. It turns out that the desired angular structure is instead produced by the martini glass EFP:



$$\overset{1c}{=} 2 \sum_{i,j=2}^{M} z_i z_j \theta_i^2 \theta_j \theta_{ij} + \mathcal{O}(z^3). \tag{2.18}$$

This allows us to write the crocodile EFP in eq. (2.16) as:

(2.19)

In our basis of EFPs in the 1-collinear expansion, described in section 3.2, we will also need the $\theta_i \theta_j \theta_{ij}^2$ structure, which we obtain by taking the difference of eqs. (2.17) and (2.18).

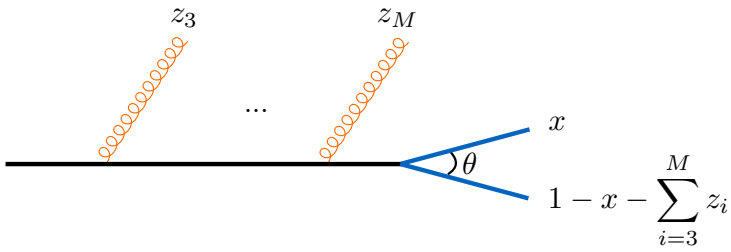


Figure 3: In the 2-collinear (2c) approximation, one can have one collinear emission with momentum fraction x, in addition to collinear-soft emissions, such that the momentum fraction of the initial parton is $1 - x - \sum_i z_i \approx 1 - x$. The angle between the collinear partons is denoted by $\theta \equiv \theta_{12}$.

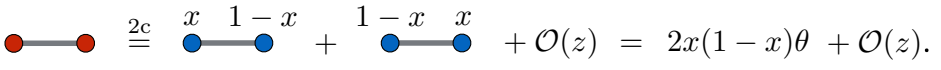
2.4 Power counting in the 2-collinear expansion

In the 2-collinear approximation, we have two collinear partons in the final state, in addition to collinear-soft emissions, as pictured in figure 3. This corresponds to:

$$\textbf{2-collinear expansion:} \quad z_1 = x\,, \quad z_2 = 1 - x\,, \quad z_i \sim z \ll 1 \text{ for } i > 2\,, \quad \theta_{ij} \sim \theta \ll 1\,, \tag{2.20}$$

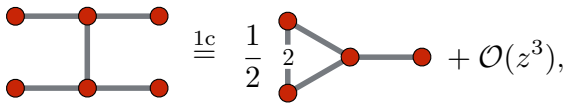
where z_2 has $\mathcal{O}(z)$ corrections, and x is $\mathcal{O}(1)$. This expansion goes beyond NLL accuracy, for which one of the emissions from the initial parton is soft or collinear, because the collinear-soft emissions are not strongly ordered.[3]

We now show some examples of the 2-collinear approximation. Let us start by expanding the dumbbell EFP:



(2.21)

As another example, we look at the "H" EFP. In the 1-collinear approximation it reads:

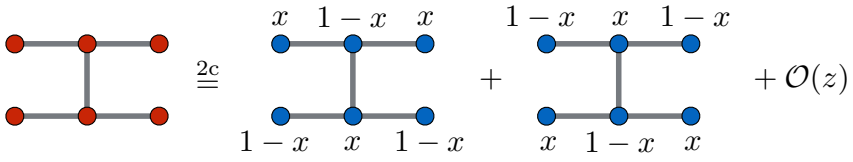


(2.22)

which also holds when one of the emissions is soft. In the 2-collinear approximation, the

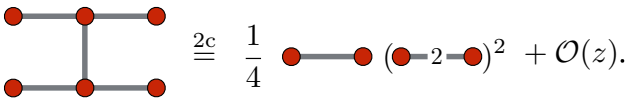
[3]This statement is true when we count logarithms in the cross section, see footnote 1. When counting logarithms in the exponent of the cross section (i.e. in $\ln \mathrm{d}\sigma$), NLL requires the exponentiation of the one-loop non-cusp anomalous dimension, and it is not obvious that this would be reproduced in the 2-collinear expansion.

EFP is dominated by the terms involving the two collinear particles (blue nodes):



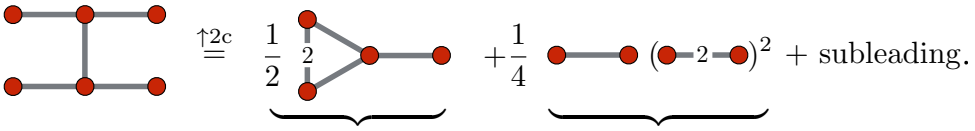
$$\overset{2c}{=} \; 2x^3(1-x)^3\theta^5 + \mathcal{O}(z). \tag{2.23}$$

This means that in the 2-collinear approximation we can use eq. (2.21) to write the H EFP as



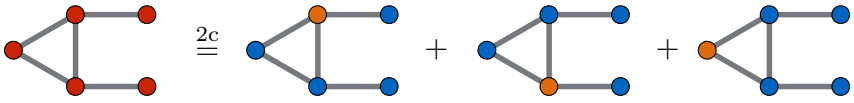
$$+ \mathcal{O}(z). \tag{2.24}$$

Interestingly, the general solution (denoted by ↑2c for "up to 2-collinear") is the sum of eq. (2.22) and eq. (2.24), because the martini glass is suppressed in the 2-collinear approximation and the product of dumbbells is suppressed in the 1-collinear approximation:



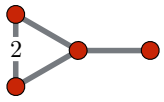
$1c : \mathcal{O}(z^2)$, $2c : \mathcal{O}(z)$ $\qquad$ $1c : \mathcal{O}(z^3)$, $2c : \mathcal{O}(1)$ $\qquad$ (2.25)

As a less trivial example, consider the "A" EFP. Because this is a 3-color graph and there are only two collinear particles, one needs to include collinear-soft emissions. The leading contributions are shown below, where the assignment of the nodes is indicated by blue nodes for collinear and orange nodes for collinear-soft emissions:



$$\overset{2c}{=} \; 2\sum_{i=3}^{M} z_i\, x(1-x)\theta^2\theta_{1i}\theta_{2i}[x\theta_{1i} + (1-x)\theta_{2i}]$$
$$+ 2\sum_{i=3}^{M} z_i\, x^2(1-x)^2\theta^3\theta_{1i}\theta_{2i} \; + \mathcal{O}(z^2), \tag{2.26}$$

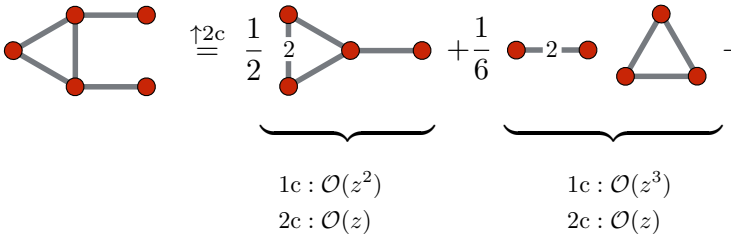
where we simplified the first term using $x(1-x)^2 + x^2(1-x) = x(1-x)$. The momentum fractions of the blue nodes can be x or $1-x$, must differ for connected nodes, and all possibilities are summed over. In the 2-collinear approximation, we also obtain:



$$\overset{2c}{=} 4\sum_{i=3}^{M} z_i\, x(1-x)\theta^2\theta_{1i}\theta_{2i}[x\theta_{1i} + (1-x)\theta_{2i}] + \mathcal{O}(z^2),$$

$$\overset{2c}{=} 6\sum_{i=3}^{M} z_i\, x(1-x)\theta\,\theta_{1i}\theta_{2i} + \mathcal{O}(z^2). \tag{2.27}$$

From this we obtain the 2-collinear solution, which in this case also holds in the 1-collinear approximation:



(2.28)

Specifically, the first term in the above equation is the solution in the 1-collinear approximation and the second term is suppressed in this case.

The above example shows that, in general, the up-to-2-collinear solution is not obtained by simply adding the individual 1- and 2-collinear solutions. The key to systematically obtaining the full solution lies in considering the degeneracy of the individual solutions, i.e. including all EFPs terms that are higher order in z as part of the solution, with unspecified coefficients. The up-to-2-collinear solution is then obtained by taking the intersection of the individual 1- and 2-collinear solutions, which is by definition a solution to both approximations. We pursue this strategy in section 3.3.

3 Constructing reduced EFP bases

In the previous section, we discussed specific relations between EFPs in the SO, 1- and 2-collinear approximations. We can now put these to use systematically to reduce the full basis of EFPs. We consider two approaches:

- **Strongly-ordered basis:** Starting from all EFPs, we use relationships obtained in the SO expansion to eliminate as many elements as possible.
- **n-collinear basis:** Similar to how we obtain the SO basis, except that we only use relationships obtained in the up-to-n-collinear expansion.

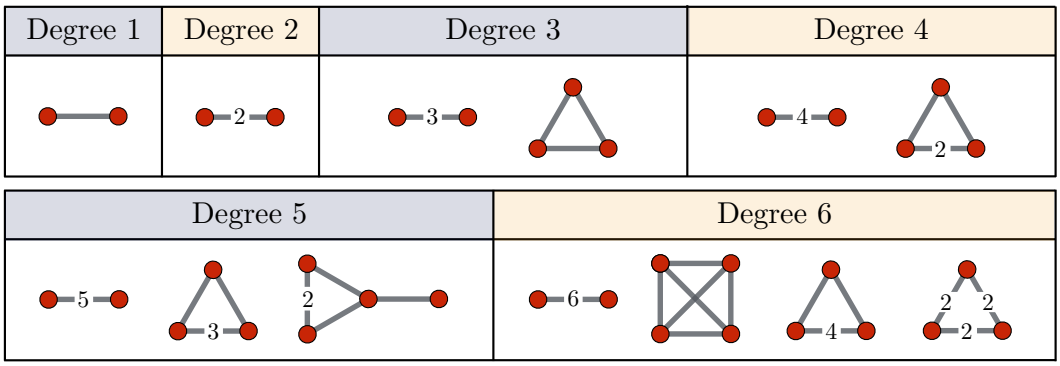


Table 3: Basis of prime EFPs up to degree 6 in the strongly-ordered (SO) expansion.

A third approach based on a z expansion is presented in appendix A. The linear relations obtained in these limits are presented in ref. [43] up to degree 6. As shown in section 5, quark/gluon discrimination starts to saturate at around degree 6, which motivates truncating our analysis at this degree; expanding to higher degrees is a straightforward but tedious exercise.

There is of course a certain arbitrariness when using relationships to reduce a basis. Though the difference is formally beyond the order that we are working at, the choice may matter. In the 1-collinear approximation, in particular, it is possible to use a color-reduced basis in which only graphs with chromatic number up to $c-1$ are needed as basis elements for graphs of chromatic number c, as discussed in appendix B. This color-reduced basis is not as useful when considering the 2-collinear approximation, however, as it would require a substantial extension. In the body of this paper, we choose a 1-collinear basis that requires a minimal extension to lift to the 2-collinear case.

3.1 The strongly-ordered basis

In the SO approximation from eq. (2.5), any EFP is reduced to a polynomial of terms of the form:

$$\prod_{i=2}^{c} z_i^{n_i} \theta_i^{m_i}, \tag{3.1}$$

with $n_2 \geq ... \geq n_c$ and $m_2 > ... > m_c$, where c is the chromatic number of the EFP in question. In fact, up to degree 6 all except for one EFP consist of a single such term (with the exception being the martini glass with a double line on the rim at degree 5). The construction of the SO basis is then performed by considering all terms of this form that can appear at a given degree:

$$\begin{aligned}
&\text{Degree 1:} && z_2\theta_2; \\
&\text{Degree 2:} && z_2\theta_2^2; \\
&\text{Degree 3:} && z_2\theta_2^3, \quad z_2 z_3 \theta_2^2 \theta_3; \\
&\text{Degree 4:} && z_2\theta_2^4, \quad z_2 z_3 \theta_2^3 \theta_3;
\end{aligned}$$

$$\begin{aligned}
&\text{Degree 5:} \quad z_2\theta_2^5, \;\; z_2z_3\theta_2^4\theta_3, \;\; z_2z_3\theta_2^3\theta_3^2; \\
&\text{Degree 6:} \quad z_2\theta_2^6, \;\; z_2z_3\theta_2^5\theta_3, \;\; z_2z_3\theta_2^4\theta_3^2, \;\; z_2z_3z_4\theta_2^4\theta_3\theta_4.
\end{aligned} \tag{3.2}$$

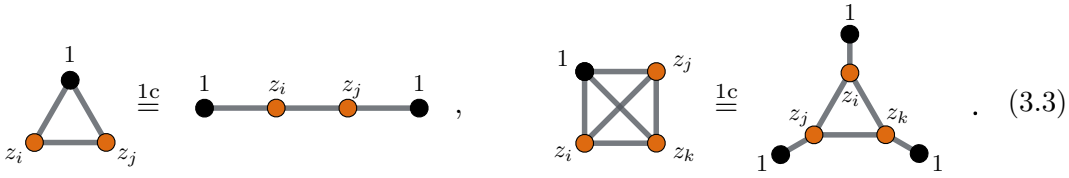
Once we select representative EFPs that correspond to each of these terms in the SO approximation, all others can be written in terms of the chosen elements. This yields the SO basis, which can be seen in table 3.

Note that we could not have chosen a basis for which all of the graphs are fully connected.[4] This would require replacing the martini glass at degree 5 by a triangle. However, the triangles with sides 2-2-1 and 3-1-1 are equivalent, since there are two large angles $\theta_{12} \sim \theta_{23}$ and one small angle θ_{23}. Thus, a partially connected graph is needed to capture the angular scaling of the martini glass.

3.2 The 1-collinear basis

To systematically construct the 1-collinear basis, we consider all possible monomial structures in the sum over momentum fractions and angles that can appear in the 1-collinear approximation from eq. (2.11). These structures and the corresponding EFPs are summarized in table 4 up to degree 5. Generic EFPs can be obtained as polynomials of these ingredients. The full basis of prime EFPs up to degree 6 is given in table 5. We have validated this basis by expressing all EFPs with $d \leq 6$ as polynomials of these basis elements.

In the 1-collinear approximation, we could have alternatively used a color-reduced basis, discussed further in appendix B. In the 1-collinear approximation, there is a single collinear parton with momentum fraction 1. This collinear parton that has be inserted somewhere on the EFP, allowing us to "cut" the EFP there. Letting a black dot indicating the collinear parton, we can cut open fully connected graphs via:



(3.3)

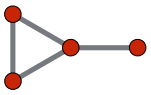
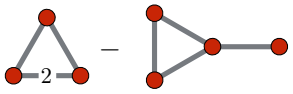
This leads to a computational gain, since cutting open a graph generally decreases the computational cost by a factor of $\mathcal{O}(M)$. That said, the above relationships do not hold in 2-collinear approximation, and one cannot, for example, write the 3-color triangle as some combination of 2-color graphs.

3.3 The 2-collinear basis

While there is considerable freedom in choosing basis elements in the 1-collinear case, the specific elements in table 5 were selected to minimize the difference between the 1-collinear and 2-collinear bases. In particular, table 5 was constructed to ensure that essentially the same basis elements could be used in the up-to-2-collinear approximation.

[4] In ref. [44] by one of the authors, an erroneous statement was made implying that fully connected graphs form a complete basis for infrared-and-collinear-safe observables. This statement was corrected in ref. [28]. Here, we see that this statement is not even true in the SO limit. Mea culpa.

	Degree 1	Degree 2	Degree 3	
Expression	$\sum_{i=2}^{M} z_i\theta_i$	$\sum_{i=2}^{M} z_i\theta_i^2$	$\sum_{i=2}^{M} z_i\theta_i^3$	$\sum_{i,j=2}^{M} z_iz_j\theta_i\theta_j\theta_{ij}$
EFP term	$\frac{1}{2}$ 	$\frac{1}{2}$ 	$\frac{1}{2}$ 	$\frac{1}{3}$ 

	Degree 4		
Expression	$\sum_{i=2}^{M} z_i\theta_i^4$	$\sum_{i,j=2}^{M} z_iz_j\theta_i^2\theta_j\theta_{ij}$	$\sum_{i,j=2}^{M} z_iz_j\theta_i\theta_j\theta_{ij}^2$
EFP term	$\frac{1}{2}$	$\frac{1}{2}$ 	

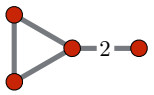
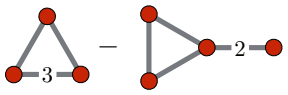
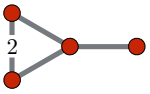
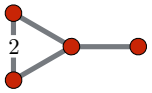
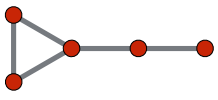
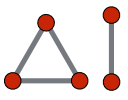
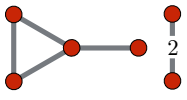
	Degree 5		
Expression	$\sum_{i=2}^{M} z_i\theta_i^5$	$\sum_{i,j=2}^{M} z_iz_j\theta_i^3\theta_j\theta_{ij}$	$\sum_{i,j=2}^{M} z_iz_j\theta_i\theta_j\theta_{ij}^3$
EFP term	$\frac{1}{2}$	$\frac{1}{2}$ 	
Expression	$\sum_{i,j=2}^{M} z_iz_j\theta_i^2\theta_j^2\theta_{ij}$	$\sum_{i,j=2}^{M} z_iz_j\theta_i^2\theta_j\theta_{ij}^2$	
EFP term	$\frac{1}{2}$ 	$\frac{1}{2}$  $-\frac{1}{4}$ 	
Expression	$\sum_{i,j,k=2}^{M} z_iz_jz_k\theta_i\theta_j\theta_{ij}\theta_{jk}$		
EFP term	$\frac{1}{2}$  $-\frac{1}{12}$  $-\frac{1}{2}$ 		

Table 4: Possible monomial structures that can appear in an EFP in the 1-collinear approximation up to degree 5, and the corresponding EFPs in terms of which they can be expressed.

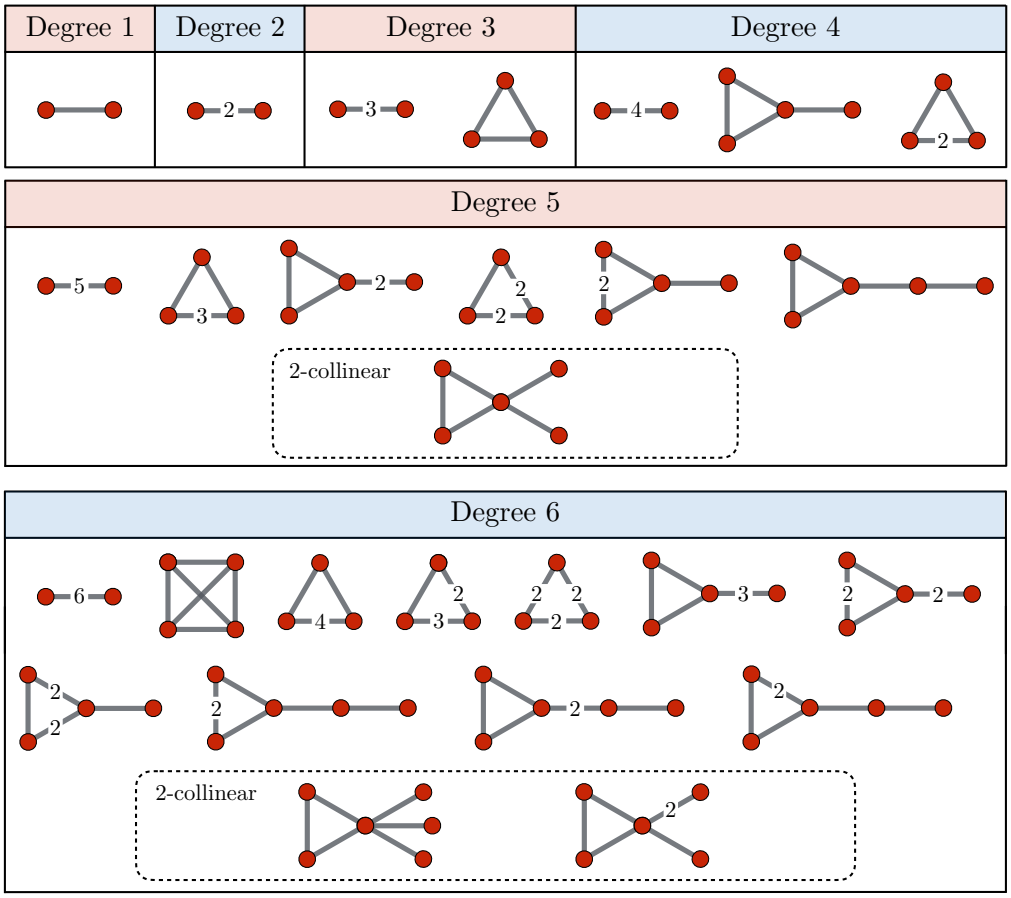
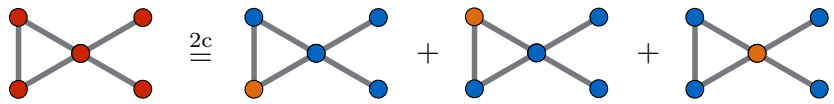


Table 5: Basis of prime EFPs up to degree 6 in the 1-collinear expansion. A generic EFP up to degree 6 can be expressed as a polynomial in terms of these bases elements. To extend this basis to the 2-collinear approximation requires one new basis element at degree 5 and two at degree 6.

To verify this, we expressed all EFPs up to degree 6 in terms of our 1-collinear basis, using the relationships that hold in the up-to 2-collinear approximation from eqs. (2.11) and (2.20). We found that this was possible for all but one EFP at degree 5 and three at degree 6, requiring the new basis elements shown in the dashed boxes of table 5. The structures of these new basis elements in the 2-collinear approximation do not match that

of any of the other EFPs. For example,



$$\overset{2c}{=} 2x(1-x)\theta \sum_{i=3}^{M} z_i \theta_{1i}\theta_{2i}\Big\{(1-2x+2x^2)\theta^2+\big[x\theta_{1i}+(1-x)\theta_{2i}\big]^2\Big\} + \mathcal{O}(z^2). \tag{3.4}$$

The last contribution on the first line is the challenging one, as it involves four angles θ_{1i}, θ_{2i} with a complicated x dependence. Since there is only a single momentum fraction z_i, it cannot be written as a product of two EFPs. Similarly, removing one of the "antennae" to change one of the lines to a double line does not capture the x dependence. We therefore conclude that this is the minimal extension needed to go from the 1-collinear to the 2-collinear basis.

4 Testing linear relations

We now test the linear relations that express EFPs in terms of basis elements in various approximations. The dataset used for this study was retrieved from refs. [19, 45], in which gluon jets are obtained from $Z(\to \nu\bar{\nu}) + g$ and quark jets from $Z(\to \nu\bar{\nu}) + (u, d, s)$. These events are generated in Pythia 8.226 [41] with $\sqrt{s} = 14$ TeV, hadronization and underlying event turned on, using the `WeakBosonAndParton:qqbar2gmZg` and `WeakBosonAndParton:qg2gmZq` processes. The jets are identified using the anti-k_T jet algorithm [46, 47] with radius parameter $R = 0.4$, and we select for jets with transverse momentum $p_T^{\text{jet}} \in [500, 550]$ GeV and rapidity $|y^{\text{jet}}| < 1.7$.

4.1 Results with the strongly-ordered basis

We start with some of the examples we discussed explicitly before: the four dots in eq. (2.8) and the crocodile in eq. (2.10). In the left panels of figure 4, we compute the left-hand and right-hand side of the equation on each jet, and plot the values as the horizontal and vertical coordinates. In the right panels of figure 4, we show the ratio between these values. While there is a clear correlation, the average ratio μ differs from 1 by a factor of 0.6, and the spread σ in the correlation is at the 10% level. This motivated us to consider the more accurate n-collinear expansions, which we discuss in section 4.2.

Moving beyond individual cases, we tested the expression of all EFPs up to degree 6 in terms of the SO basis. In the blue histograms of figures 5a and 5b, we summarize the average μ and spread σ of the ratio for prime EFPs (i.e. EFPs represented by connected graphs) that have a non-trivial power counting relation. As expected, μ is peaked around 1 and σ is peaked around 0. While there are some relationships that are very accurate, there is quite a spread, in line with figure 4. In figure 5c, we show the same results as a scatter plot with $|\mu - 1|$ and σ on the axes, allowing one to see the correlation between them.

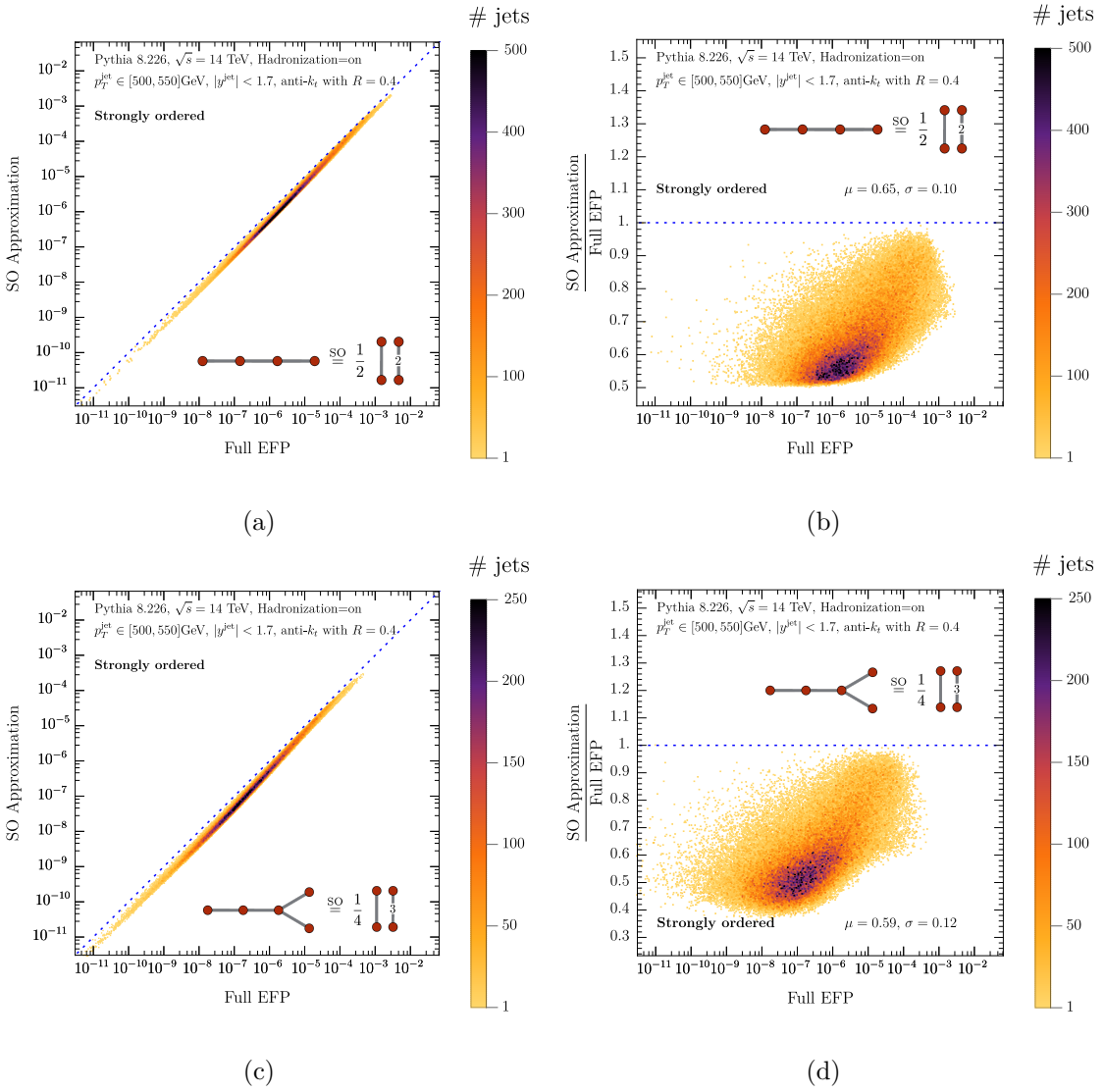


Figure 4: Testing the strongly-ordered relationship for the 4-dot EFP in eq. (2.8) (top row) and the crocodile EFP in eq. (2.10) (bottom row). Results are shown as a correlation plot (left column) and ratio (right column). The average μ and standard deviation σ of the ratio is also shown.

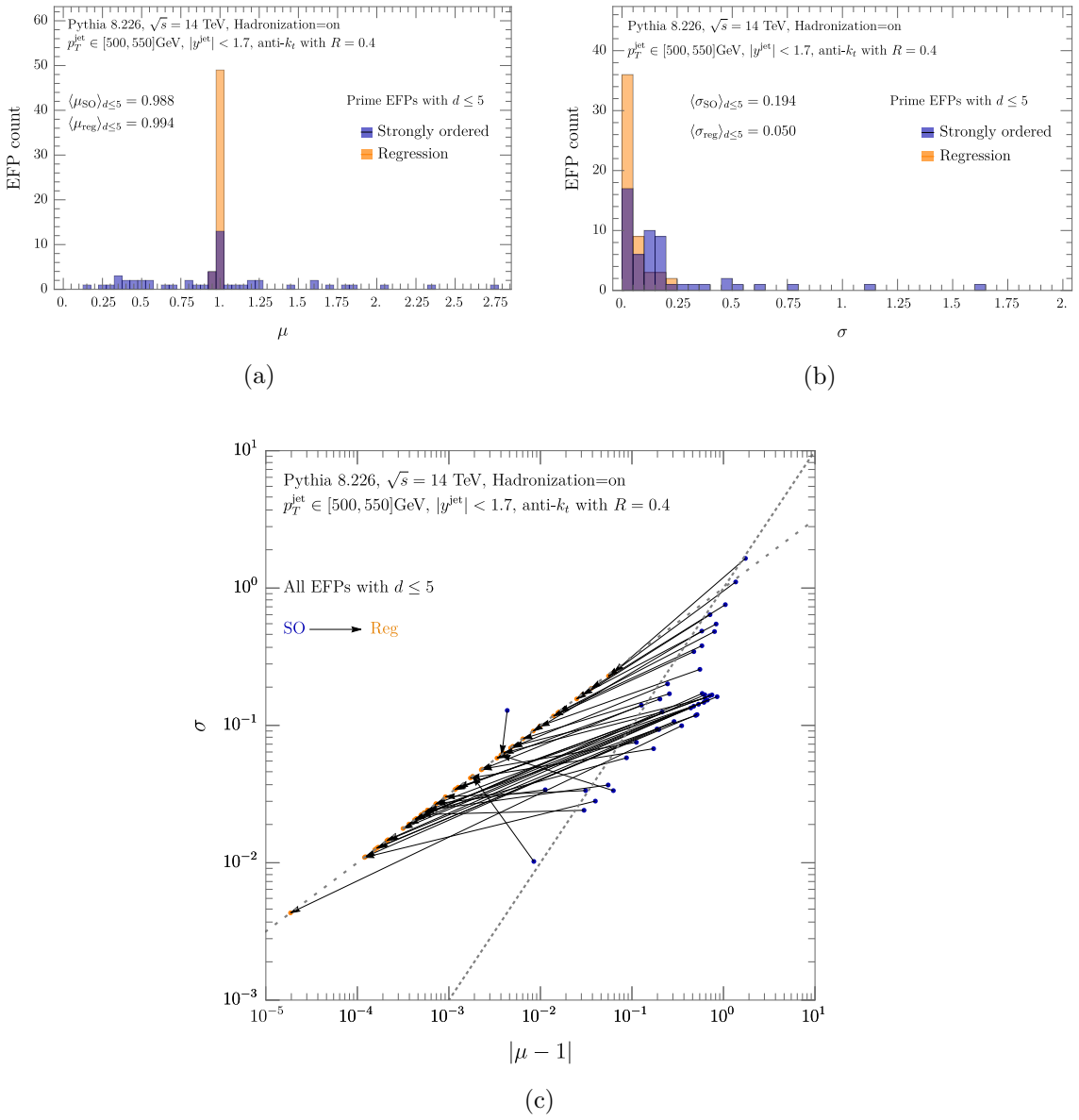


Figure 5: Histogram for the average μ (a) and standard deviation σ (b) of the ratio between prime EFPs up to degree 5 and their expression in terms of SO basis elements. Here, the relationship was found using power counting (blue) and regression (orange). The improvement from using regression is visualized with arrows in the scatter plot (c) for $|\mu - 1|$ and σ. The dashed line corresponds to $\sigma = |\mu - 1|$, and the dotted line corresponds to $\sigma^2 = |\mu - 1|$.

Empirically, we find that most relations satisfy $\sigma \simeq |\mu - 1|$, which one might anticipate because the average and spread are controlled by the same subleading power corrections.

Interestingly, even though the individual relations in the SO approximation are not very accurate, any EFP is still very well approximated by a linear combination of SO basis elements. This is shown in the orange histograms of figures 5a and 5b, where we perform a linear regression for all prime EFPs in terms of the elements of the SO basis. This helps explain the curious fact we will encounter in section 5, where the tagging performance of the SO basis matches that of the 2-collinear basis, even though the individual linear relations are considerably poorer. Note that this regression does not include a constant term (i.e. single dot EFP), so μ does not automatically equal 1, since $d > 0$ EFPs go to zero in the collinear and soft limits. Empirically in figure 5c, we see that the regression results satisfy $\sigma^2 \simeq |\mu - 1|$, which arises because with regression, the average can be fine-tuned to be smaller than the spread.

4.2 Results with the n-collinear basis

We now test linear relations in the collinear expansion. We start by considering the same examples as in figure 4, given for four dots in eq. (2.15) and the crocodile in eq. (2.19). While these relationships were derived in the 1-collinear approximation, they in fact also hold for the 2-collinear approximation. The correlation holds extremely well for four dots, while for the crocodile the mean of the ratio differs by about 10% from 1 and the spread is a few percent. In both cases, the improvement over the strongly ordered expansion is substantial.

Next, we consider two EFPs where the 2-collinear approximation differs from the 1-collinear approximation, namely the "H" and "A" EFPs from eqs. (2.25) and (2.28). In figure 7, we show the correlation plot for the 1-collinear approximation on the left and including the 2-collinear approximation on the right. While the ratio is not plotted, the average μ and standard deviation σ of the ratio are indicated in the figure. As is clear from their values, and from the correlation plots, there is a substantial improvement from including the 2-collinear approximation.

In figure 8, we test the expression of all prime EFPs up to degree 6 in terms of the basis elements, in both the 1- and 2-collinear approximation. The results up to degree 5 and 6 are shown separately, allowing one to see an increase in the number of outliers at higher degree. We have identified these outliers as star-like EFPs and will have more to say about them in section 4.3. As for the SO basis, μ is peaked around 1 and σ is peaked around 0. There is a noticeable improvement from including the 2-collinear approximation in both μ and σ.

Finally, in figure 9 we show the scatter plot with $|\mu - 1|$ and σ on the axes. In the left panel, results are shown for the 2-collinear approximation. In this plot, most EFPs are represented by a dot, though for some, the actual graph is drawn. As in the SO case, the EFPs fall along the line $|\mu - 1| = \sigma$. In the right panel, EFPs are shown for which the expression in terms of basis elements differs at the 1-collinear and 2-collinear approximation, with an arrow indicating the improvement.

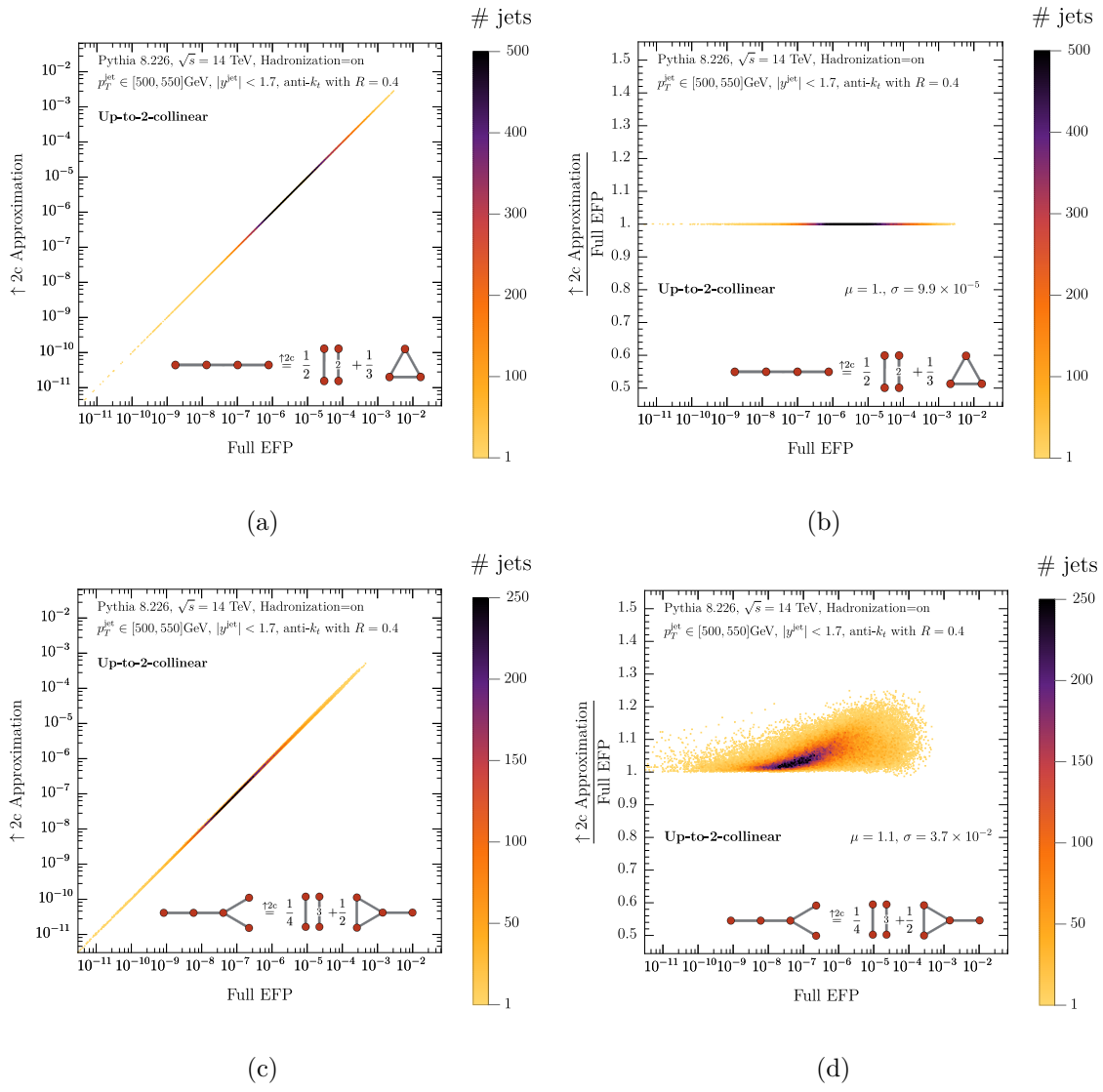


Figure 6: Same as figure 4, but using the relationships obtained from the up-to-2-collinear expansion, which are much more accurate.

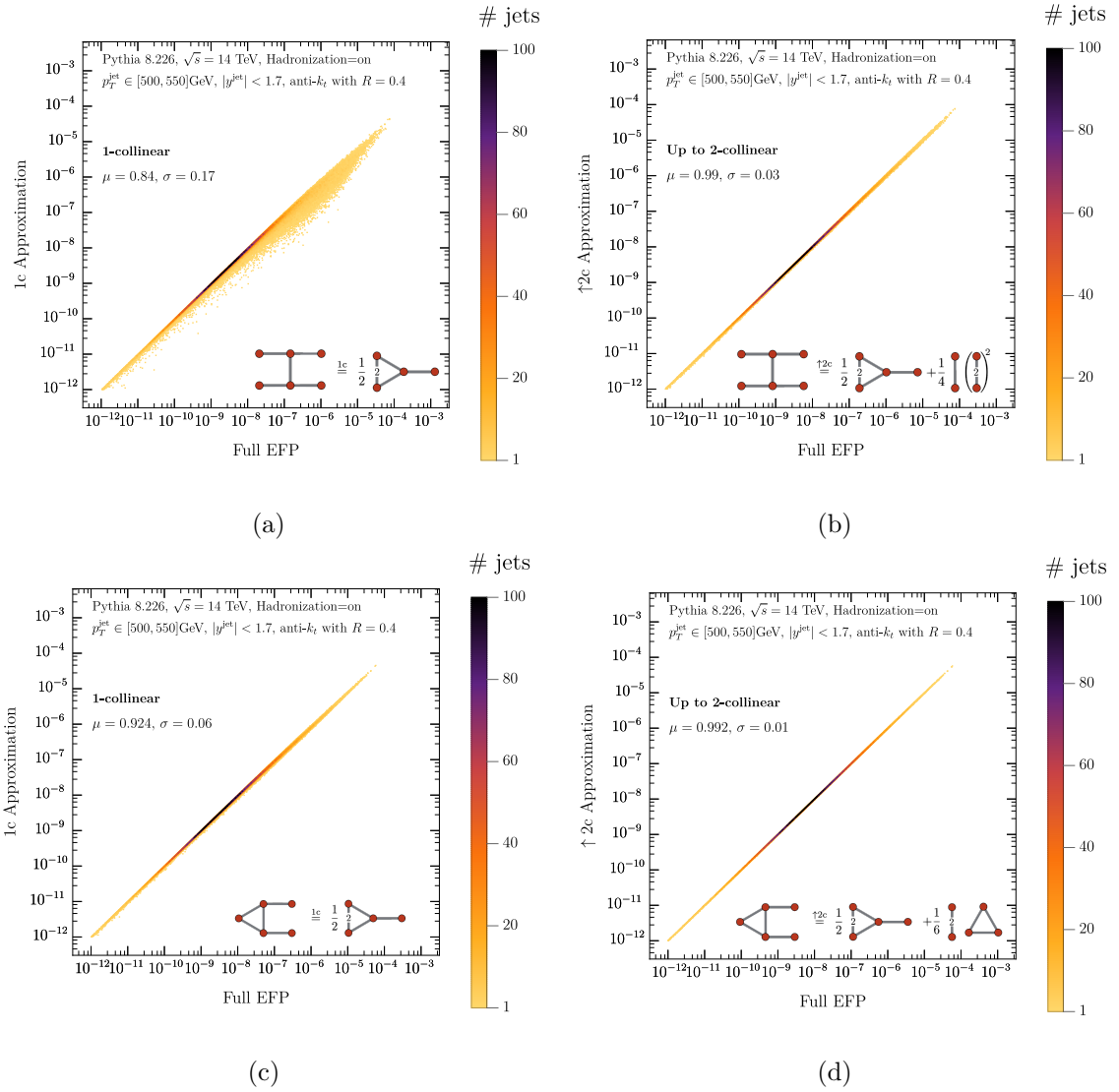


Figure 7: Testing the relationship for the "H" EFP in eq. (2.25) (top row) and "A" EFP in eq. (2.28) (bottom row) in the 1-collinear approximation (left column) and including the 2-collinear approximation (right column), which leads to a substantial improvement. The average μ and standard deviation σ of the ratio is shown.

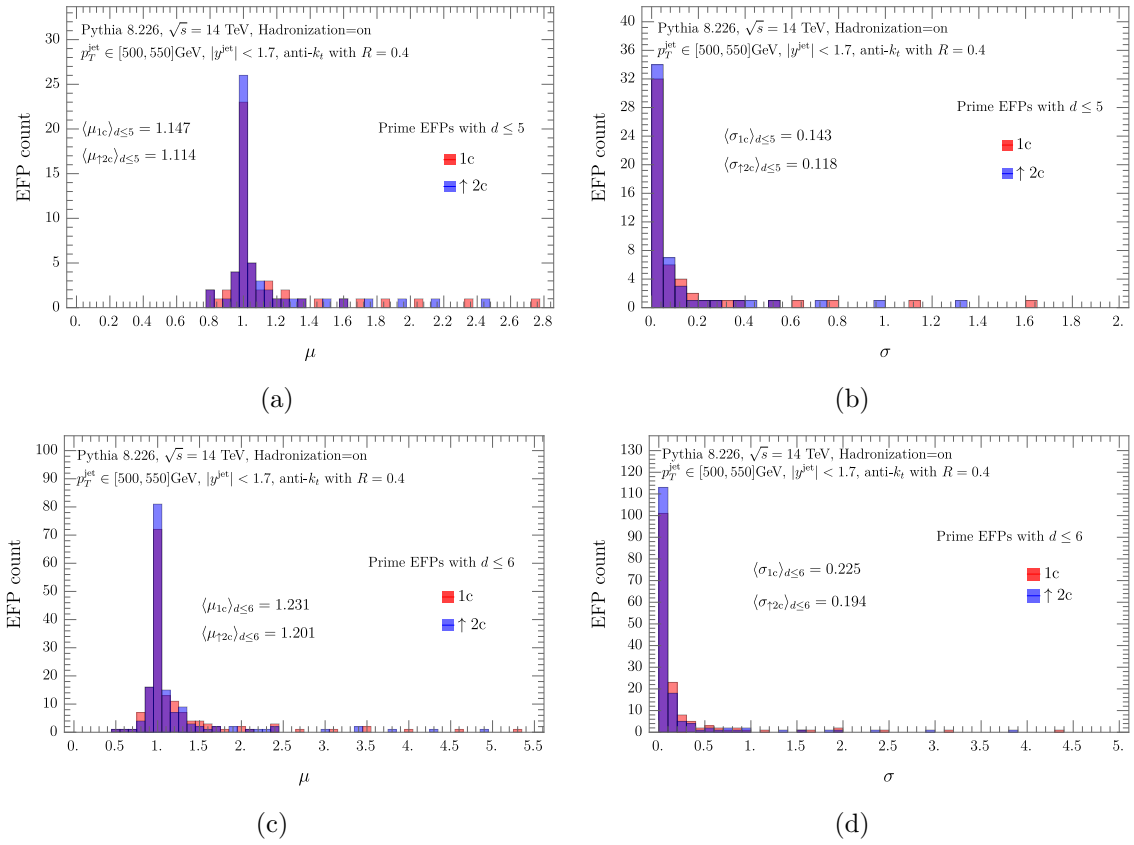
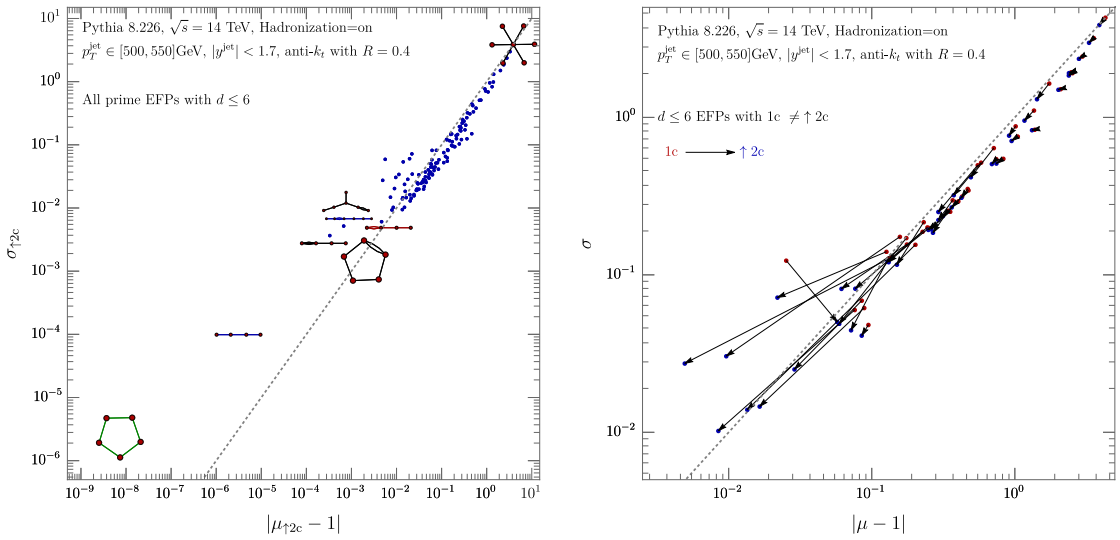


Figure 8: Histogram for the average μ (left) and standard deviation σ (right) of the ratio between prime EFPs and their expression in terms of basis elements in the 1- (red) and 2-collinear (blue) approximation. Going from EFPs up to degree 5 (top) to degree 6 (bottom), μ and σ become slightly worse because of a couple of outliers, which are star-like EFPs.



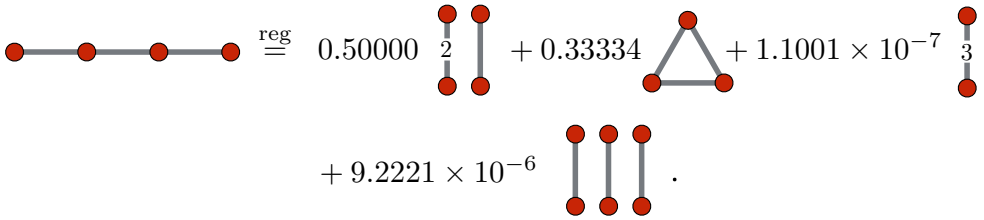
(a) (b)

Figure 9: Scatter plot for $|\mu - 1|$ and σ of the prime EFPs up to degree 6. Left: in the 2-collinear approximation. Right: the result in the 1- (red) and 2-collinear (blue) approximation, connected by an arrow. We only show whose EFPs where the expression for 2-collinear is different; note the difference in scale. The dashed line indicates $\sigma \simeq |\mu - 1|$.

4.3 Linear regression for star graphs

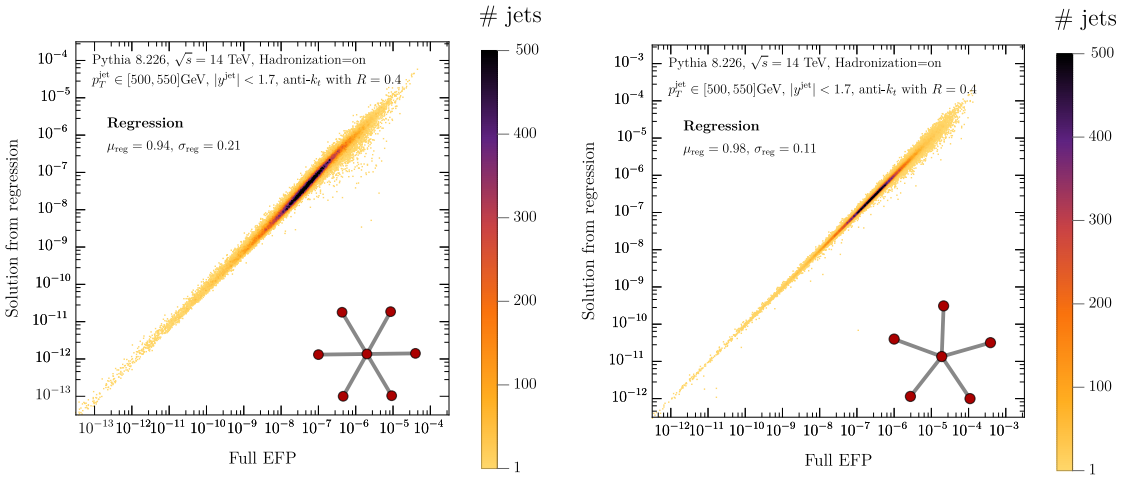
In figure 5, we saw that relationships obtained for the strongly-ordered basis performed poorly and could be improved by linear regression. We can perform a similar study for star-like EFPs, for which the n-collinear expansion does not perform very well.

Before proceeding, we first check our fit procedure. Performing regression on the 4-dot EFP, we reproduce the relationship in eq. (2.15) up to expected power corrections:



(4.1)

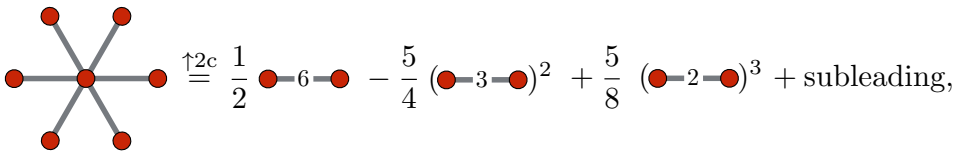
For the star-like EFPs, the n-collinear expansion breaks down. The EFP for which the collinear power counting relationship performs worst is a 6-pointed star, as indicated



(a) (b)

Figure 10: Testing the relationship between the 6-pointed (left) and 5-pointed (right) star EFP and a corresponding linear fit of EFPs belonging to the 2-collinear basis. The average and standard deviation of the ratio is greatly improved over that for the relationship obtained using 2-collinear power counting, which was $(\mu, \sigma)_6 = (4.89, 3.86)$ and $(\mu, \sigma)_5 = (2.42, 1.31)$, respectively.

in figure 9. In the 2-collinear approximation, the 6-pointed star can be expressed as:



(4.2)

which is simply a polynomial of dumbbells. At subleading power in z, though, one can assign a collinear-soft particle to the central node, which introduces strong dependence on the angle of this particle. Such angular complexity is not captured by dumbbells, explaining the poor power counting performance.

Nevertheless, we find that star-like EFPs can be well approximated by a linear combination of basis elements. Performing a linear regression (without a constant term) for the 6-pointed and 5-pointed stars in figure 10, we find μ much closer to 1 than from the 2-collinear power counting. Inspected the regression solution, we found no discernible pattern. In particular, we see no evidence that the regression solution looks like the power-counting relation plus corrections. It is an interesting open question whether this regression relationship could have been derived from first principles.

5 Logistic regression for quark/gluon jet tagging

For our last test of the power counting relations, we compare the performance of the reduced bases to that of all EFPs for the task of quark/gluon jet discrimination. Distinguishing quark jets from gluon jets has a long history [48–53], with a recent revival of interest [10, 54–57] including both analytic [58–61] and machine learning [14, 29, 62–65] approaches. Indeed, quark/gluon discrimination was one of the initial benchmark tests of EFPs [28].

For this study, we use the same dataset as in section 4, which has parton-level quark and gluon labels from Pythia. The classification is accomplished via logistic regression, which receives as input either all EFPs or the EFPs in one of four power-counting bases:

- strongly-ordered from section 3.1;
- 2-collinear from section 3.3;
- z^2-truncated from appendix A;
- color-reduced (1-collinear) from appendix B.

The classifier output for logistic regression takes the form:

$$c(\Phi) = \frac{1}{1 + e^{-\sum_G c_G \, \mathrm{EFP}_G(\Phi)}}, \tag{5.1}$$

where Φ represents the jet, c_G are regression coefficients, $\mathrm{EFP}_G(\Phi)$ are the EFPs, and the sum is over the relevant graphs. The classifier is trained to output 1 for quark jets and 0 for gluon jets.

To compare the performance of these classifiers, we examine their receiver operating characteristic (ROC) curves, which are obtained from the true-positive and false-positive rates as the decision threshold is varied. We consider quark jets as signal and gluon jets as background, such that the true-positive rate corresponds to the quark jet efficiency and the false-positive to the gluon jet mistag rate. To encapsulate the classifier performance into a single quantity, we take the area under the curve (AUC) of the ROC curve, with an AUC of 0.5 corresponding to a random classifier and an AUC of 1.0 to a perfect classifier.

The ROC curves for the different sets of inputs (all EFPs, strongly-ordered basis, 2-collinear basis, z^2-truncated, and color-reduced basis) are shown in figure 11, where the different panels correspond to the maximum degree of the EFPs. The corresponding AUCs are shown in table 6, and the gluon mistag rate for a quark efficiency of 0.5 is shown in table 7. At degree 1, there is only a single EFP and this is an element of all bases, which thus perform the same. At degrees 2 and 3, the z^2-truncated basis has the best AUC of the reduced bases (performing as well as using all EFPs), as it keeps some power-suppressed terms. It does not perform as well beyond degree 4, where the 2-collinear has the best AUC.

Perhaps surprisingly, the strongly-ordered basis has pretty much the same performance as the 2-collinear basis for degrees 4 and higher. The poor correlation plots for the strongly ordered expansion in figure 4 are not indicative of its tagging performance. This was

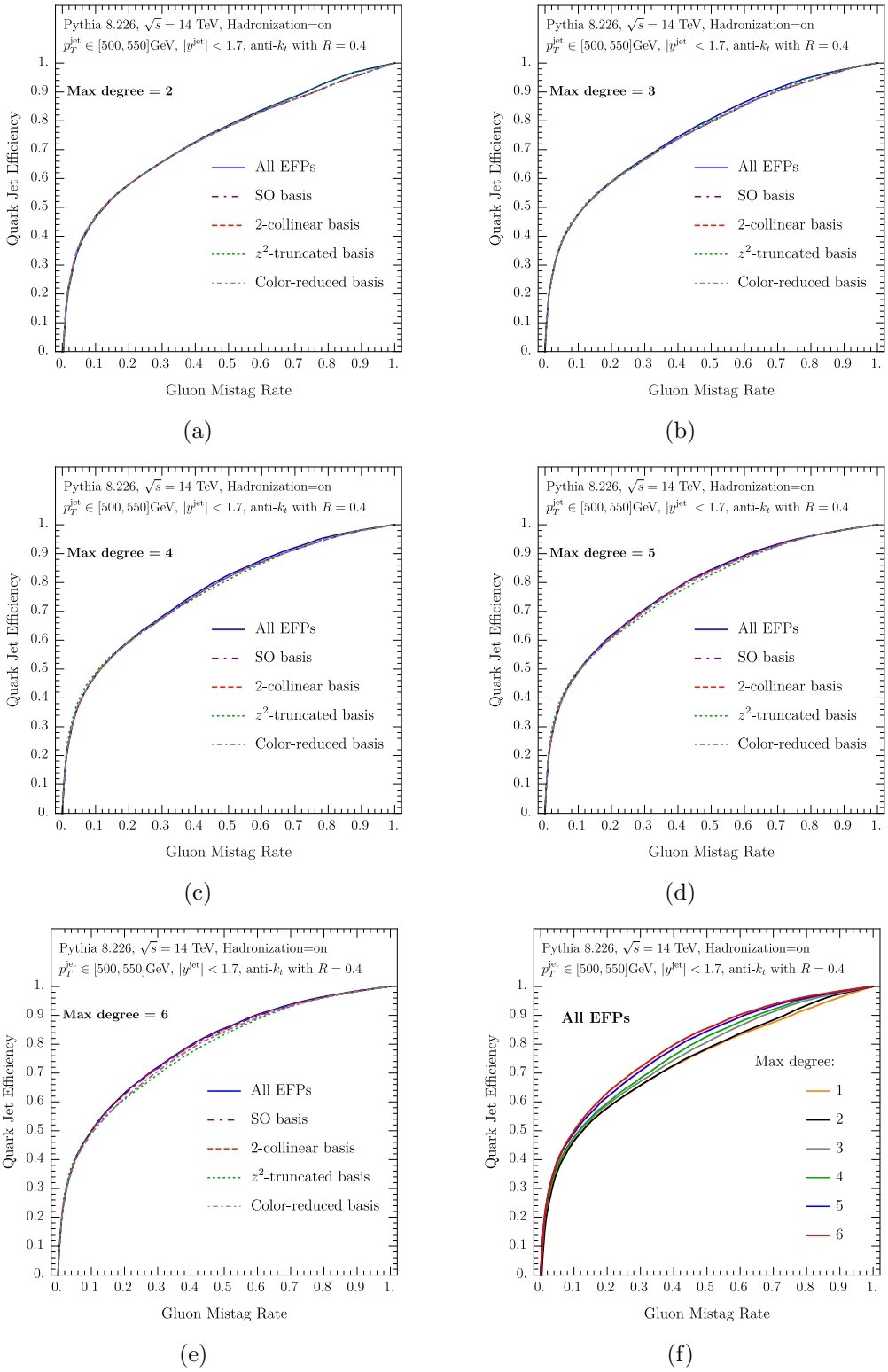


Figure 11: ROC curves for quark/gluon discrimination, organized by maximum degree for the different inputs used for classification: all EFPs, strongly-ordered basis, 2-collinear basis, z^2-truncated basis, and color-reduced (1-collinear) basis. To help interpret the comparison, in figure 11f we show the ROC curves for all EFPs by maximum degree.

Max degree	All EFPs	SO	2-collinear	z^2-truncated	Color-reduced
1	0.741	**0.741**	**0.741**	**0.741**	**0.741**
2	0.745	0.741	0.741	**0.745**	0.741
3	0.761	0.755	0.755	**0.759**	0.755
4	0.770	0.765	**0.766**	**0.766**	0.765
5	0.784	0.781	**0.782**	0.776	0.779
6	0.792	**0.789**	**0.789**	0.780	0.782

Table 6: AUCs for the ROC curves displayed in figure 11. The uncertainty for all AUC values is ±0.003, obtained from the 95% confidence interval coming from 10-fold cross validation. Bold-faced entries show the best performing approximations in each row.

Max degree	All EFPs	SO	2-collinear	z^2-truncated	Color-reduced
1	0.125	**0.125**	**0.125**	**0.125**	**0.125**
2	0.123	**0.123**	**0.123**	0.127	**0.123**
3	0.117	0.117	0.117	**0.114**	0.117
4	0.115	0.115	0.115	**0.109**	0.115
5	0.105	0.107	0.106	**0.103**	0.107
6	0.100	**0.101**	**0.101**	0.102	0.106

Table 7: Gluon mistag rate at 0.5 quark efficiency. The uncertainty for all values is ±0.004, obtained from the 95% confidence interval coming from 10-fold cross validation. Bold-faced entries show the best performing approximations in each row.

anticipated in figure 5, as we found much better (albeit training-set dependent) relations with regression. The color-reduced basis also performs similarly to the 2-collinear case, except at degree 6.

Looking beyond AUCs at the structure of the ROC curves, we note that the strongly-ordered and 2-collinear bases have similar performance independent of the desired quark jet efficiency/gluon mistag rate. On the other hand, the ROC curve for the z^2-truncated basis has a different shape. In the high quark purity regime, it performs even better than using all EFPs, as also illustrated in table 7 (compared to table 6). As one goes towards higher quark efficiency, though, it performs worse. It would be interesting to understand what $O(z^3)$ physics explains this behavior.

6 Conclusions

Energy flow polynomials (EFPs) form an (over)complete linear basis for jet substructure. While one could simply use all EFPs as input for machine learning studies, computational cost limits how many EFPs can be used in practice. By employing power counting, we find relations between EFPs that hold to a certain level of accuracy, and we can use these to substantially reduce the basis of EFPs, providing a more efficient choice of inputs.

Such reductions are also beneficial to streamline the interpretation of machine learning algorithms [31].

We considered two power counting schemes in this body of this paper: strong ordering (SO) in the energy and angle of emissions; and an expansion involving n-collinear emissions and arbitrary collinear-soft emissions, with no further hierarchies. We found that it was possible to choose a basis for the 1-collinear case such that only three new basis elements were needed up to degree 6 to obtain the 2-collinear basis. Alternatively in appendix B, we find that it is possible to obtain a color-reduced basis for the 1-collinear case, such that EFPs with chromatic number c can be described using those with chormatic number $c-1$, reducing the computational overhead. In appendix A, we consider keeping terms up to a certain power in the energy fraction (z^n truncation). One might expect that this would perform better at low degree (where additional basis elements are kept) than at high degree (where more are dropped), which was borne out in our quark/gluon discrimination study.

While the linear relationships obtained between EFPs in the SO expansion perform substantially worse than in the up-to-2-collinear case, interestingly, the two expansions have similar performance for quark/gluon discrimination, particularly at higher degree. Thus, while the power counting relationships from the SO limit are not as clean, the same relevant information is apparently still present. This echoes the argument in ref. [39] that only a small number of observables are needed to map out N-body phase space. Here, though, there is a crucial difference that we only performed simple logistic regression and not a fully non-linear machine learning study. We leave a study of this curious result to future work.

We limited our study to the case of single-prong jets initiated by gluons or light quarks. It will be interesting to extend this analysis to the case of jets produced from the hadronic decay of heavy resonances such as the Higgs boson or top quark. Because we did not assume a relative hierarchy of angles in our n-collinear expansion, we expect that similar (and possibly the same) basis elements will appear for more complex decay topologies.

Power counting is sufficient for reducing the basis of EFPs, but further work is needed to make precise predictions for jet substructure. The natural next step would be the calculation of cross sections that are *simultaneously* differential in the basis EFPs. There has been some work on multi-differential cross sections for angularities [66–69], which correspond to dumbbells in the language of EFPs. Following on the pioneering study of ref. [70], it will be interesting to extend this analysis to higher-point EFPs. Multi-differential EFP studies could provide insights into jet properties that are complementary to the parton showers approach. We anticipate that power counting will continue to play an essential role in organizing and simplifying systematic studies of jet substructure.

Acknowledgments

We thank Patrick Komiske for collaboration in the initial stages of this research. P.C. thanks Miguel Jaques for machine learning related discussions. J.T. is supported by the U.S. Department of Energy (DOE) Office of High Energy Physics under Grant No. DE-SC0012567, and by the National Science Foundation under Cooperative Agreement PHY-2019786

(The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/). W.W. is supported by the D-ITP consortium, a program of NWO that is funded by the Dutch Ministry of Education, Culture and Science (OCW). P.C. is supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 101002090).

A z^2-truncated basis

In this appendix, we consider energy truncation as a strategy to organize the basis of EFPs. We assume the same configuration of one hard emission and $M-1$ collinear-soft emissions as in the 1-collinear approximation in eq. (2.11). Rather than keeping only leading terms and dropping all subleading terms, however, we use energy scaling to determine which elements to keep:

- **z^n-truncated basis:** We remove EFPs that scale as $\mathcal{O}(z^{n+1})$ and only use relationships between EFPs where the dropped terms are $\mathcal{O}(z^{n+1})$, where $z \ll 1$ is the collinear-soft momentum fraction.

The z^n-truncated basis seems natural when linearly combining EFPs, as it takes their relative z scaling into account. (Of course, this suppression could be compensated by the size of the coefficients.) In the z^n-truncated bases, certain EFPs can be directly eliminated due to their scaling, but there are fewer relationships that can be used to reduce the basis. Because the degree of the EFP limits the z scaling,[5] z^2-truncated bases will have more elements at low degree and fewer at high degree compared to the 2-collinear basis.

The z-truncated basis is rather simple: it consists of EFPs with two nodes connected by any number of lines, i.e. dumbbells. Moving on to a more interesting case, we consider the z^2-truncated basis, for which the basis elements up to degree 6 are shown in table 8. Here we show composite EFPs, such that all EFPs can be expressed as a *linear* combination of these ingredients. This is different from tables 3 and 5 which only showed prime EFPs, and the EFP power counting relationships involved polynomials of the basis elements.

The reason for the different presentation in table 8 is that for z^n-truncation, it is in general difficult to determine by inspection which products are kept and whether they are related to other EFPs. For z^2-truncation this is not too complicated yet: only products of two dumbbells are allowed. Because we can only use relationships between EFPs that hold up to corrections that are order z^3, new basis elements appear compared to the 2-collinear case in table 5, such as 3 dots in a row or star-like configurations. On the other hand, other basis elements from table 5 do not appear here because they themselves are order z^3, such as the fully-connected four dot graph or a martini glass with an elongated stem.

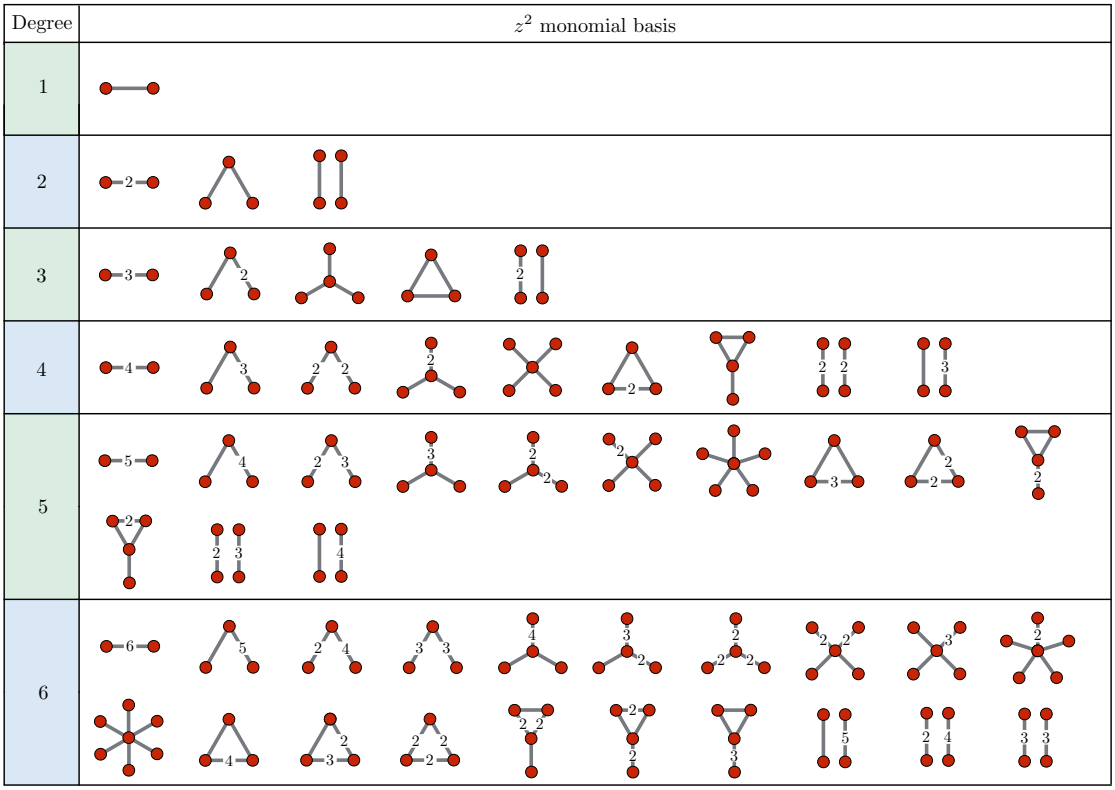


Table 8: The z^2-truncated basis of EFPs up to degree 6. A generic EFP up to degree 6 can be expressed as a *linear* combination in terms of these bases elements.

B 1-collinear color-reduced basis

As discussed in section 3.2, we have freedom to choose a different basis set for the 1-collinear expansion which is more computational efficient. The essential trick is shown in eq. (3.3), where any vertex with a collinear parton can be cut open at 1-collinear order. For a chromatic number c graph, this trick allows it to be expressed in terms of chromatic number $c-1$ basis elements.

In table 9, we present a color reduced basis that can be used in the 1-collinear expansion. This basis has a reduced number of loop-like graphs, which offers a computational speedup in evaluating the EFP. There are no loop-like graphs up to degree 5, and only one loop-like graph at degree 6. Unfortunately, this color reduction does not persist to higher orders, and a non-minimal extension of table 9 is needed to lift the basis to 2-collinear accuracy.

For tagging and regression, this color-reduced 1-collinear basis exhibits somewhat worse performance than the default 1-collinear basis in table 5. This is to be expected, since the default 1-collinear basis already contains the majority of the 2-collinear information. It

[5]EFPs with N nodes and degree d scale like $z^\kappa\theta^d$ with $\kappa \leq N-1 \leq d$, since always at least one of the nodes involves a collinear parton. Disconnected EFPs also satisfy $\kappa \leq d$, since they are simply products of connected EFPs.

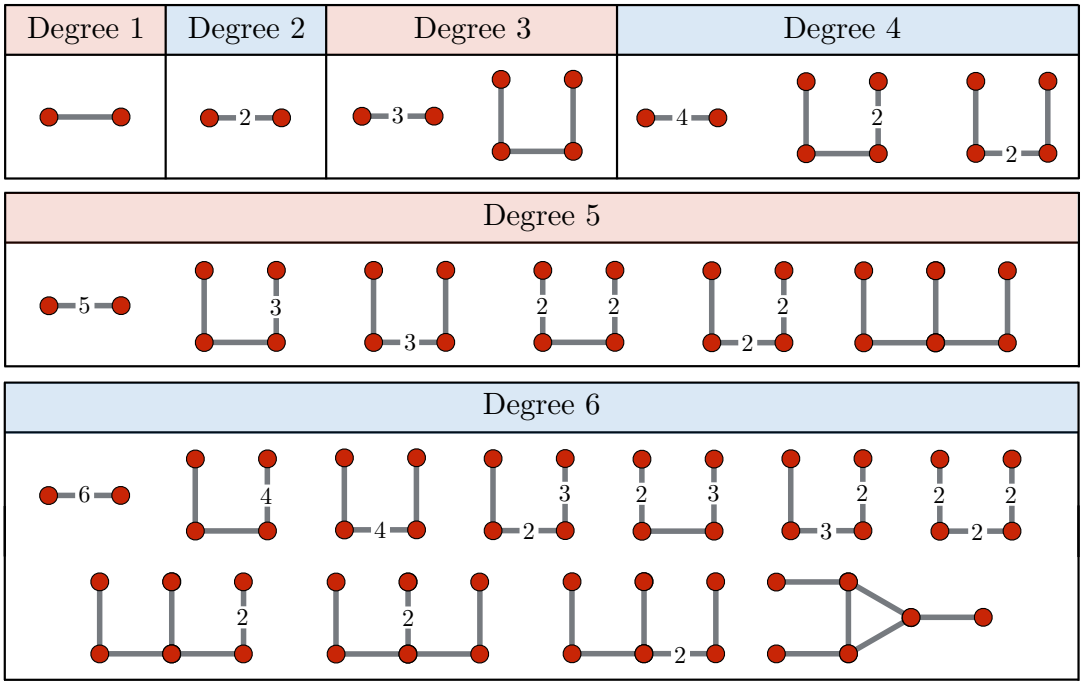


Table 9: Prime EFPs for the color-reduced basis of the 1-collinear expansion.

would be interesting to explore trade offs between performance and computational cost. At degree 6, the fully-connected 4-point graph is the most costly EFP to compute, so one could consider a hybrid basis where this chromatic number 4 graph is cut open but most chromatic number 3 graphs are kept.

References

[1] M.H. Seymour, *Tagging a heavy Higgs boson*, in *ECFA Large Hadron Collider (LHC) Workshop: Physics and Instrumentation*, pp. 557–569, 1, 1991.

[2] M.H. Seymour, *Searches for new particles using cone and cluster jet algorithms: A Comparative study*, *Z. Phys. C* **62** (1994) 127.

[3] J.M. Butterworth, B.E. Cox and J.R. Forshaw, *WW scattering at the CERN LHC*, *Phys. Rev. D* **65** (2002) 096014 [hep-ph/0201098].

[4] J.M. Butterworth, J.R. Ellis and A.R. Raklev, *Reconstructing sparticle mass spectra using hadronic decays*, *JHEP* **05** (2007) 033 [hep-ph/0702150].

[5] J.M. Butterworth, A.R. Davison, M. Rubin and G.P. Salam, *Jet substructure as a new Higgs search channel at the LHC*, *Phys. Rev. Lett.* **100** (2008) 242001 [0802.2470].

[6] D.E. Kaplan, K. Rehermann, M.D. Schwartz and B. Tweedie, *Top Tagging: A Method for Identifying Boosted Hadronically Decaying Top Quarks*, *Phys. Rev. Lett.* **101** (2008) 142001 [0806.0848].

[7] J. Thaler and L.-T. Wang, *Strategies to Identify Boosted Tops*, *JHEP* **07** (2008) 092 [0806.0023].

[8] L.G. Almeida, S.J. Lee, G. Perez, G.F. Sterman, I. Sung and J. Virzi, *Substructure of high-p_T Jets at the LHC*, *Phys. Rev. D* **79** (2009) 074017 [0807.0234].

[9] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness*, *JHEP* **03** (2011) 015 [1011.2268].

[10] J. Gallicchio and M.D. Schwartz, *Quark and Gluon Tagging at the LHC*, *Phys. Rev. Lett.* **107** (2011) 172001 [1106.3076].

[11] J. Cogan, M. Kagan, E. Strauss and A. Schwartzman, *Jet-Images: Computer Vision Inspired Techniques for Jet Tagging*, *JHEP* **02** (2015) 118 [1407.5675].

[12] L.G. Almeida, M. Backović, M. Cliche, S.J. Lee and M. Perelstein, *Playing Tag with ANN: Boosted Top Identification with Pattern Recognition*, *JHEP* **07** (2015) 086 [1501.05968].

[13] P. Baldi, K. Bauer, C. Eng, P. Sadowski and D. Whiteson, *Jet Substructure Classification in High-Energy Physics with Deep Neural Networks*, *Phys. Rev. D* **93** (2016) 094034 [1603.09349].

[14] P.T. Komiske, E.M. Metodiev and M.D. Schwartz, *Deep learning in color: towards automated quark/gluon jet discrimination*, *JHEP* **01** (2017) 110 [1612.01551].

[15] D. Guest, J. Collado, P. Baldi, S.-C. Hsu, G. Urban and D. Whiteson, *Jet Flavor Classification in High-Energy Physics with Deep Neural Networks*, *Phys. Rev. D* **94** (2016) 112002 [1607.08633].

[16] G. Louppe, K. Cho, C. Becot and K. Cranmer, *QCD-Aware Recursive Neural Networks for Jet Physics*, *JHEP* **01** (2019) 057 [1702.00748].

[17] F.A. Dreyer, G.P. Salam and G. Soyez, *The Lund Jet Plane*, *JHEP* **12** (2018) 064 [1807.04758].

[18] A. Andreassen, I. Feige, C. Frye and M.D. Schwartz, *JUNIPR: a Framework for Unsupervised Machine Learning in Particle Physics*, *Eur. Phys. J. C* **79** (2019) 102 [1804.09720].

[19] P.T. Komiske, E.M. Metodiev and J. Thaler, *Energy Flow Networks: Deep Sets for Particle Jets*, *JHEP* **01** (2019) 121 [1810.05165].

[20] A.J. Larkoski, I. Moult and D. Neill, *Power Counting to Better Jet Observables*, *JHEP* **12** (2014) 009 [1409.6298].

[21] A.J. Larkoski, I. Moult and D. Neill, *Building a Better Boosted Top Tagger*, *Phys. Rev. D* **91** (2015) 034035 [1411.0665].

[22] A.J. Larkoski and I. Moult, *The Singular Behavior of Jet Substructure Observables*, *Phys. Rev. D* **93** (2016) 014017 [1510.08459].

[23] C.W. Bauer, S. Fleming, D. Pirjol and I.W. Stewart, *An Effective field theory for collinear and soft gluons: Heavy to light decays*, *Phys. Rev. D* **63** (2001) 114020 [hep-ph/0011336].

[24] C.W. Bauer and I.W. Stewart, *Invariant operators in collinear effective theory*, *Phys. Lett. B* **516** (2001) 134 [hep-ph/0107001].

[25] C.W. Bauer, D. Pirjol and I.W. Stewart, *Soft collinear factorization in effective field theory*, *Phys. Rev. D* **65** (2002) 054022 [hep-ph/0109045].

[26] C.W. Bauer, S. Fleming, D. Pirjol, I.Z. Rothstein and I.W. Stewart, *Hard scattering factorization from effective field theory*, *Phys. Rev. D* **66** (2002) 014017 [hep-ph/0202088].

[27] M. Beneke, A.P. Chapovsky, M. Diehl and T. Feldmann, *Soft collinear effective theory and heavy to light currents beyond leading power*, *Nucl. Phys. B* **643** (2002) 431 [hep-ph/0206152].

[28] P.T. Komiske, E.M. Metodiev and J. Thaler, *Energy flow polynomials: A complete linear basis for jet substructure*, *JHEP* **04** (2018) 013 [1712.07124].

[29] P.T. Komiske, E.M. Metodiev and J. Thaler, *An operational definition of quark and gluon jets*, *JHEP* **11** (2018) 059 [1809.01140].

[30] A. Butter et al., *The Machine Learning landscape of top taggers*, *SciPost Phys.* **7** (2019) 014 [1902.09914].

[31] T. Faucett, J. Thaler and D. Whiteson, *Mapping Machine-Learned Physics into a Human-Readable Space*, *Phys. Rev. D* **103** (2021) 036020 [2010.11998].

[32] J. Collado, J.N. Howard, T. Faucett, T. Tong, P. Baldi and D. Whiteson, *Learning to identify electrons*, *Phys. Rev. D* **103** (2021) 116028 [2011.01984].

[33] J. Collado, K. Bauer, E. Witkowski, T. Faucett, D. Whiteson and P. Baldi, *Learning to isolate muons*, *JHEP* **21** (2020) 200 [2102.02278].

[34] B.M. Dillon, G. Kasieczka, H. Olischlager, T. Plehn, P. Sorrenson and L. Vogel, *Symmetries, Safety, and Self-Supervision*, 2108.04253.

[35] Y. Lu, A. Romero, M.J. Fenton, D. Whiteson and P. Baldi, *Resolving Extreme Jet Substructure*, 2202.00723.

[36] L. Bradshaw, S. Chang and B. Ostdiek, *Creating Simple, Interpretable Anomaly Detectors for New Physics in Jet Substructure*, 2203.01343.

[37] J. Thaler and K. Van Tilburg, *Maximizing Boosted Top Identification by Minimizing N-subjettiness*, *JHEP* **02** (2012) 093 [1108.2701].

[38] I.W. Stewart, F.J. Tackmann and W.J. Waalewijn, *N-Jettiness: An Inclusive Event Shape to Veto Jets*, *Phys. Rev. Lett.* **105** (2010) 092002 [1004.2489].

[39] K. Datta and A. Larkoski, *How Much Information is in a Jet?*, *JHEP* **06** (2017) 073 [1704.08249].

[40] L. Moore, K. Nordström, S. Varma and M. Fairbairn, *Reports of My Demise Are Greatly Exaggerated: N-subjettiness Taggers Take On Jet Images*, *SciPost Phys.* **7** (2019) 036 [1807.04769].

[41] T. Sjöstrand, S. Ask, J.R. Christiansen, R. Corke, N. Desai, P. Ilten et al., *An introduction to PYTHIA 8.2*, *Comput. Phys. Commun.* **191** (2015) 159 [1410.3012].

[42] P.T. Komiske, E.M. Metodiev and J. Thaler, *Cutting Multiparticle Correlators Down to Size*, *Phys. Rev. D* **101** (2020) 036019 [1911.04491].

[43] P. Cal, J. Thaler and W. Waalewijn, *Power counting relations for Energy Flow Polynomials*, *Zenodo* (2022) 6542205.

[44] I. Moult, L. Necib and J. Thaler, *New Angles on Energy Correlation Functions*, *JHEP* **12** (2016) 153 [1609.07483].

[45] P. Komiske, E. Metodiev and J. Thaler, *Pythia8 quark and gluon jets for energy flow*, *Zenodo* (2019) 3164691.

[46] M. Cacciari, G.P. Salam and G. Soyez, *The anti-k_t jet clustering algorithm*, *JHEP* **04** (2008) 063 [0802.1189].

[47] M. Cacciari, G.P. Salam and G. Soyez, *FastJet User Manual*, *Eur. Phys. J. C* **72** (2012) 1896 [1111.6097].

[48] H.P. Nilles and K.H. Streng, *Quark - Gluon Separation in Three Jet Events*, *Phys. Rev.* **D23** (1981) 1944.

[49] L.M. Jones, *Tests for Determining the Parton Ancestor of a Hadron Jet*, *Phys. Rev.* **D39** (1989) 2550.

[50] Z. Fodor, *How to See the Differences Between Quark and Gluon Jets*, *Phys. Rev.* **D41** (1990) 1726.

[51] L. Jones, *Towards a Systematic Jet Classification*, *Phys. Rev.* **D42** (1990) 811.

[52] L. Lönnblad, C. Peterson and T. Rognvaldsson, *Using neural networks to identify jets*, *Nucl. Phys.* **B349** (1991) 675.

[53] J. Pumplin, *How to tell quark jets from gluon jets*, *Phys. Rev.* **D44** (1991) 2025.

[54] J. Gallicchio and M.D. Schwartz, *Quark and Gluon Jet Substructure*, *JHEP* **04** (2013) 090 [1211.7038].

[55] B. Bhattacherjee, S. Mukhopadhyay, M.M. Nojiri, Y. Sakaki and B.R. Webber, *Associated jet and subjet rates in light-quark and gluon jet discrimination*, *JHEP* **04** (2015) 131 [1501.04794].

[56] D. Ferreira de Lima, P. Petrov, D. Soper and M. Spannowsky, *Quark-Gluon tagging with Shower Deconstruction: Unearthing dark matter and Higgs couplings*, *Phys. Rev. D* **95** (2017) 034001 [1607.06031].

[57] P. Gras, S. Höche, D. Kar, A. Larkoski, L. Lönnblad, S. Plätzer et al., *Systematics of quark/gluon tagging*, *JHEP* **07** (2017) 091 [1704.03878].

[58] A.J. Larkoski, J. Thaler and W.J. Waalewijn, *Gaining (Mutual) Information about Quark/Gluon Discrimination*, *JHEP* **11** (2014) 129 [1408.3122].

[59] C. Frye, A.J. Larkoski, J. Thaler and K. Zhou, *Casimir Meets Poisson: Improved Quark/Gluon Discrimination with Counting Observables*, *JHEP* **09** (2017) 083 [1704.06266].

[60] J. Mo, F.J. Tackmann and W.J. Waalewijn, *A case study of quark-gluon discrimination at NNLL' in comparison to parton showers*, *Eur. Phys. J. C* **77** (2017) 770 [1708.00867].

[61] A.J. Larkoski and E.M. Metodiev, *A Theory of Quark vs. Gluon Discrimination*, *JHEP* **10** (2019) 014 [1906.01639].

[62] E.M. Metodiev and J. Thaler, *Jet Topics: Disentangling Quarks and Gluons at Colliders*, *Phys. Rev. Lett.* **120** (2018) 241602 [1802.00008].

[63] T. Cheng, *Recursive Neural Networks in Quark/Gluon Tagging*, *Comput. Softw. Big Sci.* **2** (2018) 3 [1711.02633].

[64] H. Luo, M.-x. Luo, K. Wang, T. Xu and G. Zhu, *Quark jet versus gluon jet: fully-connected neural networks with high-level features*, *Sci. China Phys. Mech. Astron.* **62** (2019) 991011 [1712.03634].

[65] G. Kasieczka, N. Kiefer, T. Plehn and J.M. Thompson, *Quark-Gluon Tagging: Machine Learning vs Detector*, *SciPost Phys.* **6** (2019) 069 [1812.09223].

[66] A.J. Larkoski, I. Moult and D. Neill, *Toward Multi-Differential Cross Sections: Measuring Two Angularities on a Single Jet*, *JHEP* **09** (2014) 046 [1401.4458].

[67] M. Procura, W.J. Waalewijn and L. Zeune, *Resummation of Double-Differential Cross Sections and Fully-Unintegrated Parton Distribution Functions*, *JHEP* **02** (2015) 117 [1410.6483].

[68] M. Procura, W.J. Waalewijn and L. Zeune, *Joint resummation of two angularities at next-to-next-to-leading logarithmic order*, *JHEP* **10** (2018) 098 [1806.10622].

[69] G. Lustermans, A. Papaefstathiou and W.J. Waalewijn, *How much joint resummation do we need?*, *JHEP* **10** (2019) 130 [1908.07529].

[70] A.J. Larkoski, I. Moult and D. Neill, *Analytic Boosted Boson Discrimination*, *JHEP* **05** (2016) 117 [1507.03018].