A complete linear basis for (machine) learning jet substructure

Machine Learning for Jet Physics Workshop, 2017

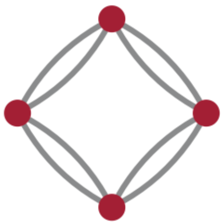
Eric M. Metodiev

Center for Theoretical Physics

Massachusetts Institute of Technology

Based on work with Patrick T. Komiske and Jesse Thaler

December 12, 2017



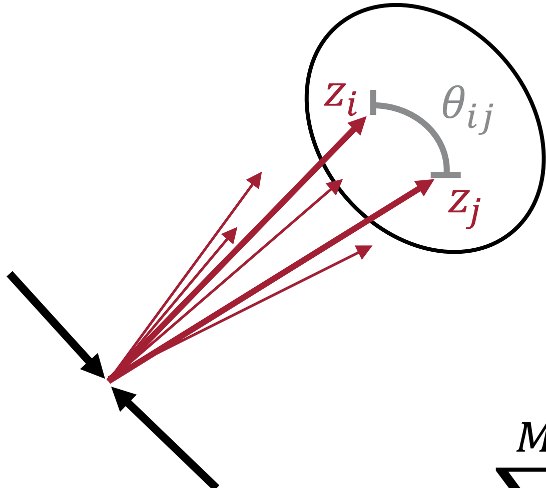
Energy Flow Polynomials (EFPs)

The Energy Flow Basis from IRC safety

Taming the (IRC-safe) Substructure Zoo

Spanning Substructure with Linear Regression

Anatomy of an Energy Flow Polynomial:



Energy Fraction | Pairwise Angular Distance

$$e^+e^-: \; z_i = \frac{E_j}{\sum_k E_k}, \quad \theta_{ij} = \left(\frac{2p_i^\mu p_{j\mu}}{E_i E_j}\right)^{\frac{\beta}{2}}$$

$$\text{Hadronic:} \; z_i = \frac{p_{Tj}}{\sum_k p_{Tk}}, \quad \theta_{ij} = \left(\Delta y_{ij}^2 + \Delta\phi_{ij}^2\right)^{\frac{\beta}{2}}$$

In equations:

$$\mathrm{EFP}_G = \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} \cdots \sum_{i_N=1}^{M} z_{i_1} z_{i_2} \cdots z_{i_N} \prod_{(k,l)\in G} \theta_{i_k i_l}$$

multigraph

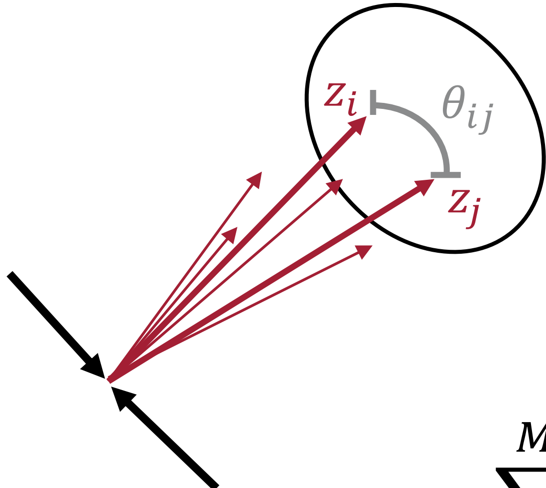
In words:

Correlator — Sum over all *N*-tuples of particle in the event

of **Energies** — Product of the *N* energy fractions

and **Angles** — One $\theta_{i_k i_l}$ for each edge in $(k,l) \in G$

Anatomy of an Energy Flow Polynomial:



Energy Fraction | Pairwise Angular Distance

$$e^+e^-: \; z_i = \frac{E_j}{\sum_k E_k}, \quad \theta_{ij} = \left(\frac{2p_i^\mu p_{j\mu}}{E_i E_j}\right)^{\frac{\beta}{2}}$$

$$\text{Hadronic: } \; z_i = \frac{p_{Tj}}{\sum_k p_{Tk}}, \quad \theta_{ij} = \left(\Delta y_{ij}^2 + \Delta\phi_{ij}^2\right)^{\frac{\beta}{2}}$$

In equations:

$$\text{EFP}_G = \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} \cdots \sum_{i_N=1}^{M} z_{i_1} z_{i_2} \cdots z_{i_N} \prod_{(k,l)\in G} \theta_{i_k i_l}$$

multigraph

In words:

Correlator of Energies and Angles

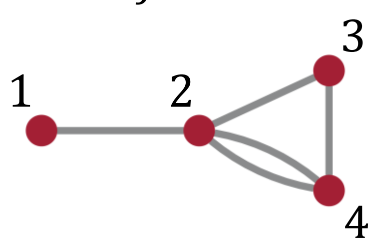
- **Correlator**: Sum over all N-tuples of particle in the event
- **Energies**: Product of the N energy fractions
- **Angles**: One $\theta_{i_k i_l}$ for each edge in $(k,l) \in G$

In pictures:

$\longleftrightarrow \; z_{i_j}$

$\longleftrightarrow \; \theta_{i_k i_l}$

(e.g.)

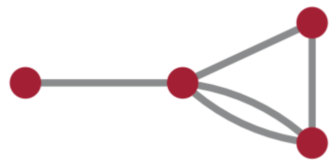


$$= \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} \sum_{i_3=1}^{M} \sum_{i_4=1}^{M} z_{i_1} z_{i_2} z_{i_3} z_{i_4} \, \theta_{i_1 i_2} \theta_{i_2 i_3} \theta_{i_3 i_4} \theta_{i_2 i_4}^2$$

(any index labelling works)

Multigraph/EFP Correspondence

Multigraph ⟷ EFP



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

j ⟷ z_{i_j}

k — l ⟷ $\theta_{i_k i_l}$

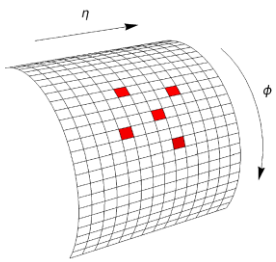
N	Number of vertices	⟷	*N*-particle correlator
d	Number of edges	⟷	Degree of angular monomial
χ	Treewidth + 1	⟷	Optimal VE Complexity

e.g. Tree graph EFPs are $O(M^2)$! Surprisingly efficient to compute. Stay tuned… See [P. Komiske's talk].

Connected	⟷	Prime
Disconnected	⟷	Composite

⋮

Energy Flow Polynomials (EFPs)



The Energy Flow Basis from IRC safety

Taming the (IRC-safe) Substructure Zoo

Spanning Substructure with Linear Regression

EFPs linearly span IRC-safe observables

IRC-safe Observable

EFPs linearly span IRC-safe observables

IRC-safe Observable

Energy Expansion: Expand/approximate the observable in polynomials of the particle energies

IR safety: Observable unchanged by addition of infinitesimally soft particle

C safety: Observable unchanged by the collinear splitting of a particle

Relabeling Symmetry: All ways of indexing particles are equivalent

New, direct argument from IRC safety
See also: F. Tkachov, hep-ph/9601308
N. Sveshnikov and F. Tkachov, hep-ph/9512370

Energy correlators linearly span IRC-safe observables

EFPs linearly span IRC-safe observables

IRC-safe Observable

Energy Expansion: Expand/approximate the observable in polynomials of the particle energies

IR safety: Observable unchanged by addition of infinitesimally soft particle

C safety: Observable unchanged by the collinear splitting of a particle

Relabeling Symmetry: All ways of indexing particles are equivalent

New, direct argument from IRC safety
See also: F. Tkachov, hep-ph/9601308
N. Sveshnikov and F. Tkachov, hep-ph/9512370

Energy correlators linearly span IRC-safe observables

Angular Expansion: Expansion/approximation of angular part of correlators in pairwise angular distances
Analyze: Identify the unique analytic structures that emerge as non-isomorphic multigraphs/EFPs

Similar expansions & emergent multigraphs in:
M. Hogervorst *et al.* arXiv:1409.1581
B. Henning *et al.* arXiv:1706.08520

EFPs linearly span/approximate IRC-safe observables!

Organization of the basis

EFPs are truncated by angular degree *d*, the order of the angular expansion.

Finite number at each order in *d*!
All prime EFPs up to *d=5* →

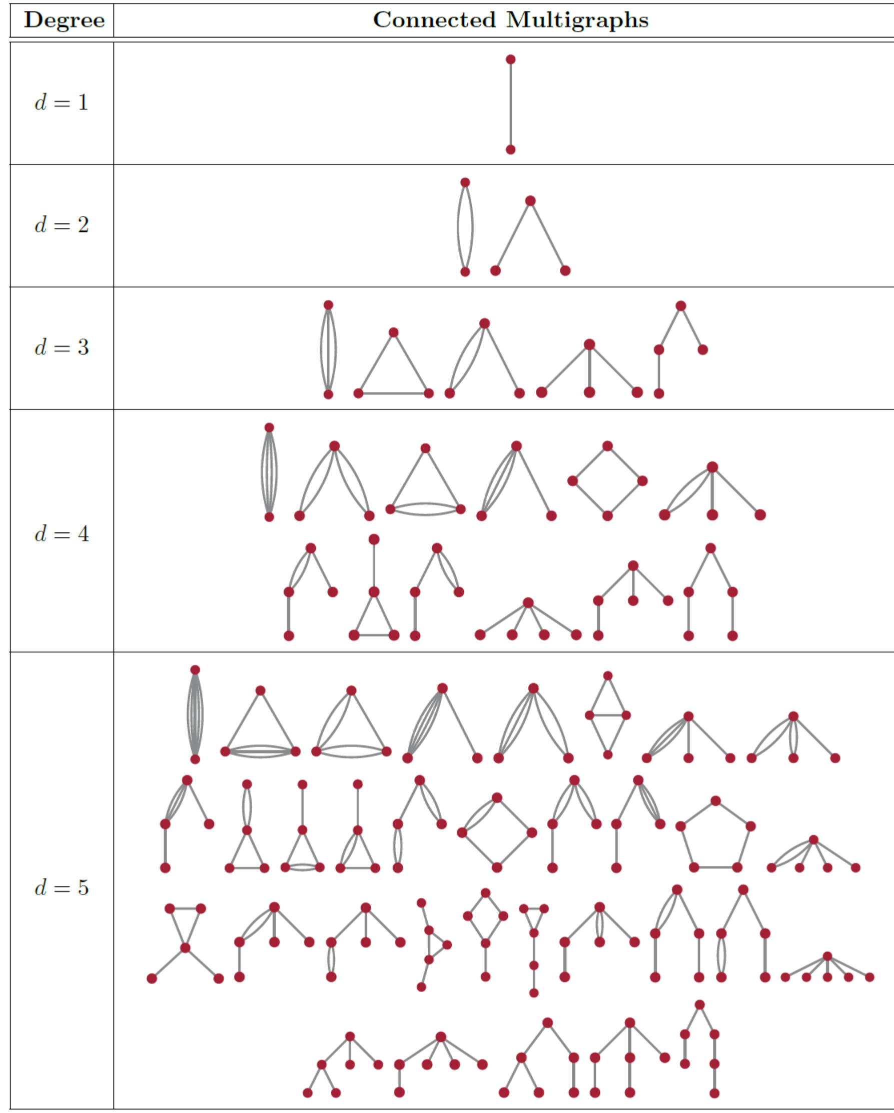


Image files for all of the prime EFP multigraphs up to d = 7 are available here.

Online Encyclopedia of Integer Sequences (OEIS)

A050535 # of multigraphs with *d* edges
# of EFPs of degree *d*

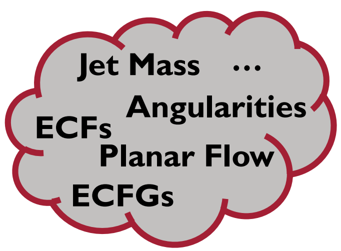
A076864 # of connected multigraphs with *d* edges
# of prime EFPs of degree *d*

↓

Exactly 1000 EFPs up to degree *d=7!*

Energy Flow Polynomials (EFPs)

The Energy Flow Basis from IRC safety



Taming the (IRC-safe) Substructure Zoo

Spanning Substructure with Linear Regression

Jet Observables with Energy Flow

Jet Mass

Dumbbell EFP

$$\frac{m_J^2}{p_{TJ}^2} = \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} z_{i_1} z_{i_2} (\cosh \Delta y_{i_1 i_2} - \cos \Delta \phi_{i_1 i_2}) = \frac{1}{2}$$

$$+ \cdots$$

Can include these using a fully general measure

Jet Observables with Energy Flow

Jet Mass
Dumbbell EFP

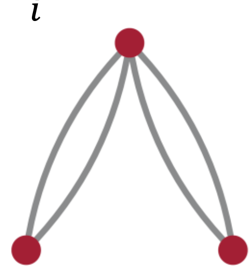
$$\frac{m_J^2}{p_{TJ}^2} = \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} z_{i_1} z_{i_2} (\cosh \Delta y_{i_1 i_2} - \cos \Delta \phi_{i_1 i_2}) = \frac{1}{2}$$  $$+ \cdots$$

Can include these using a fully general measure

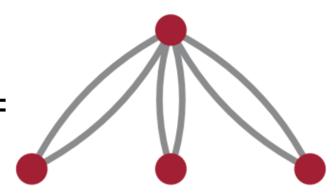
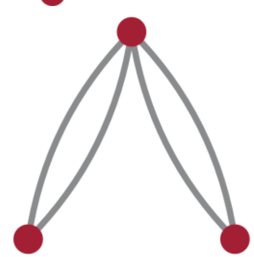
Angularities
Star Graph EFPs

$$\lambda^{(\alpha)} = \sum_{i}^{M} z_i \theta_i^{\alpha}$$

[C. Berger, T. Kucs, and G. Sterman, hep-ph/0303051]
[L. Almeida, *et al.*, arXiv:0807.0234]
[S. Ellis, *et al.*, arXiv:10010014]
[A. Larkoski, J. Thaler, and W. Waalewijn, arXiv:1408.3122]

$$\lambda^{(4)} =$$  $$-\frac{3}{4}$$ 

(and so on, for all even angularities)

$$\lambda^{(6)} =$$  $$-\frac{3}{2}$$   $$+\frac{5}{8}$$ 

using pT-centroid axis

Jet Observables with Energy Flow

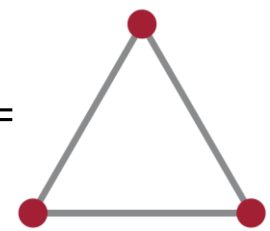
Energy Correlation Functions

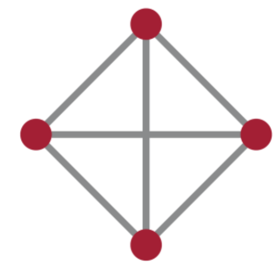
Complete Graph EFPs

$$e_N^{(\beta)} = \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} \cdots \sum_{i_N=1}^{M} z_{i_1} z_{i_2} \cdots z_{i_N} \prod_{k<l\in\{1,\cdots,N\}} \theta_{i_k i_l}^{\beta}$$

[A. Larkoski, G. Salam, and J. Thaler, arXiv:1305.0007]

$e_2^{(\beta)} =$ 

$e_3^{(\beta)} =$ 

$e_4^{(\beta)} =$  $\cdots$

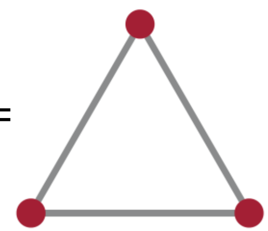
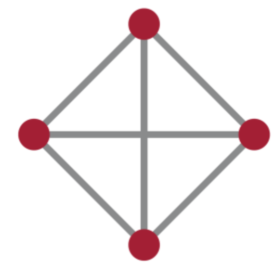
with measure choice of β

Jet Observables with Energy Flow

Energy Correlation Functions
Complete Graph EFPs

$$e_N^{(\beta)} = \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} \cdots \sum_{i_N=1}^{M} z_{i_1} z_{i_2} \cdots z_{i_N} \prod_{k<l\in\{1,\cdots,N\}} \theta_{i_k i_l}^{\beta}$$

[A. Larkoski, G. Salam, and J. Thaler, arXiv:1305.0007]

$e_2^{(\beta)} =$  $e_3^{(\beta)} =$  $e_4^{(\beta)} =$  $\cdots$

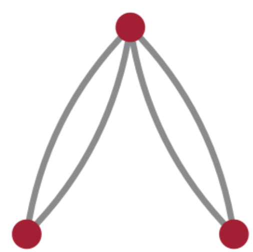
with measure choice of β

Geometric Moments
Higher dumbbell EFPs

$$\mathbf{C} = \sum_{i_1=1}^{M} z_i \begin{pmatrix} \Delta y_i^2 & \Delta y_i \Delta\phi_i \\ \Delta\phi_i \Delta y_i & \Delta\phi_i^2 \end{pmatrix}$$

$$(\text{e.g.})\ \text{Pf} = \frac{4 \det \mathbf{C}}{(\text{tr}\,\mathbf{C})^2}$$

[L. Almeida, *et al.*, arXiv:0807.0234]
[J. Thaler and L.-T. Wang, arXiv:0806.0023]
[J. Gallicchio and M. Schwartz, arXiv:1211.7038]

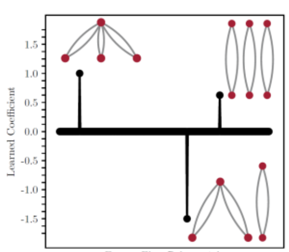
$\text{tr}\,\mathbf{C} = \frac{1}{2}$  $\det \mathbf{C} =$  $-\frac{1}{2}$ 

using pT-centroid axis

Energy Flow Polynomials (EFPs)

The Energy Flow Basis from IRC safety

Taming the (IRC-safe) Substructure Zoo



Spanning Substructure with Linear Regression

Linear Models and Energy Flow

$$S = \sum_G w_G \, \mathrm{EFP_G}$$

Machine learn these

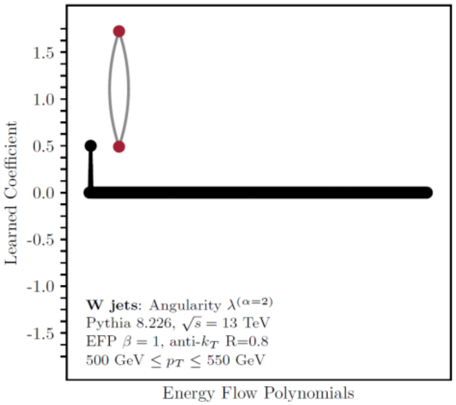
Linear methods:

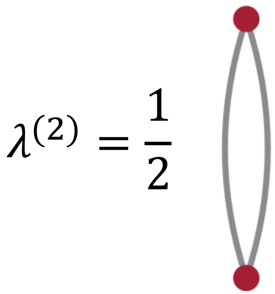
Utilize the linear completeness of the Energy Flow basis.
Convex and few/no hyperparameters to tune.
Achieve global optimum via closed-form solution or convergent iteration.
Simple models with the minimum number of parameters/input.

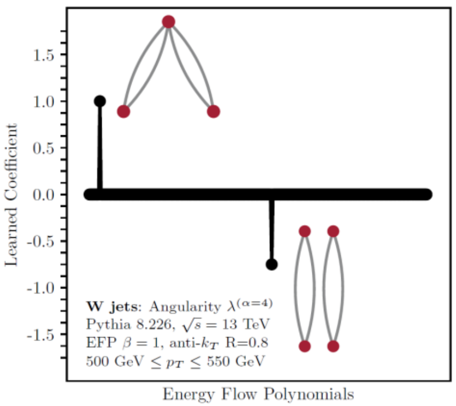
Rich in tools and applications:

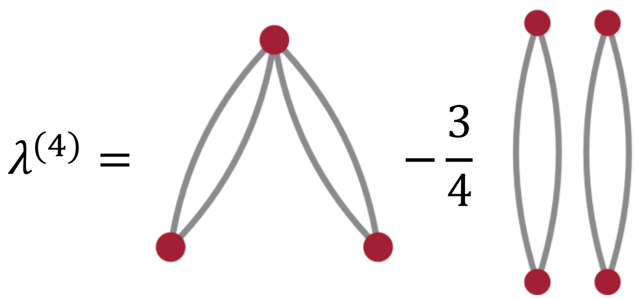
First few chapters of C. Bishop's *Pattern Recognition and Machine Learning*:

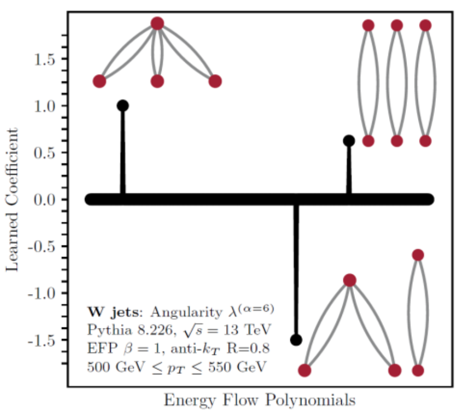
Confirming Analytic Relationships with Regression

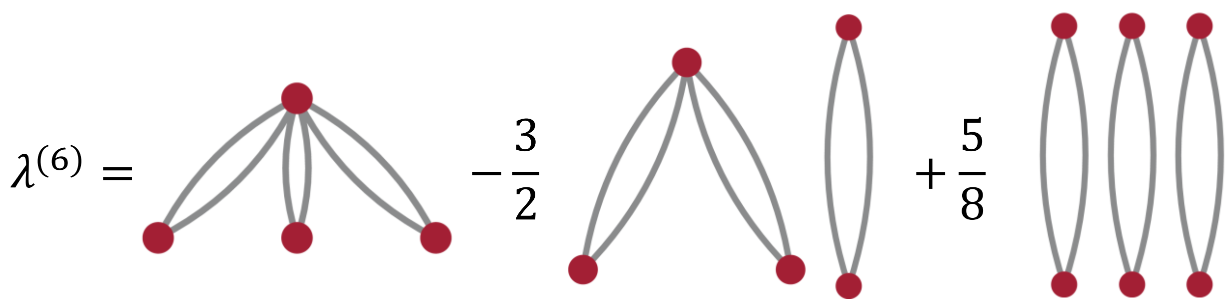












Linear Regression and IRC-safety

$\frac{m_J}{p_{TJ}}$: IRC safe. No Taylor expansion due to square root.

$\lambda^{(\alpha=1/2)}$: IRC safe. No simple analytic relationship.

τ_2: IRC safe. Algorithmically defined.

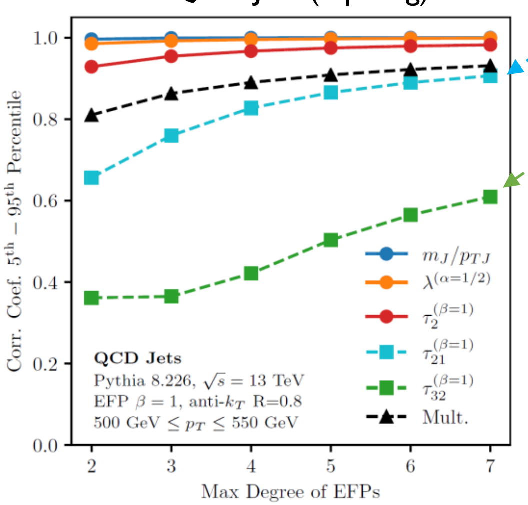
τ_{21}: Sudakov safe. Safe for 2-prong jets and higher.

[A. Larkoski, S. Marzani, and J. Thaler, arXiv:1502.01719]

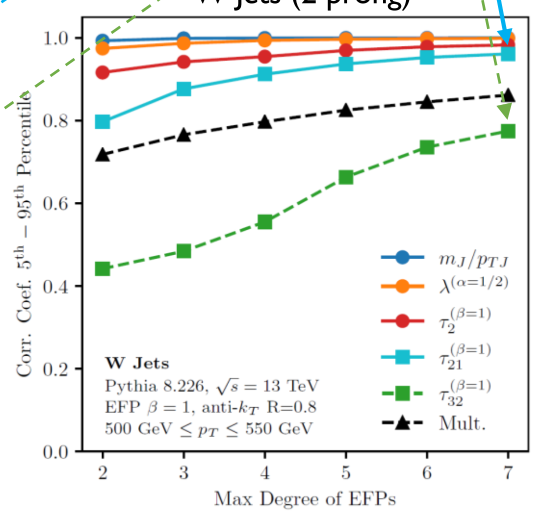
τ_{32}: Sudakov safe. Safe for 3-prong jets and higher.

Multiplicity: IRC unsafe.

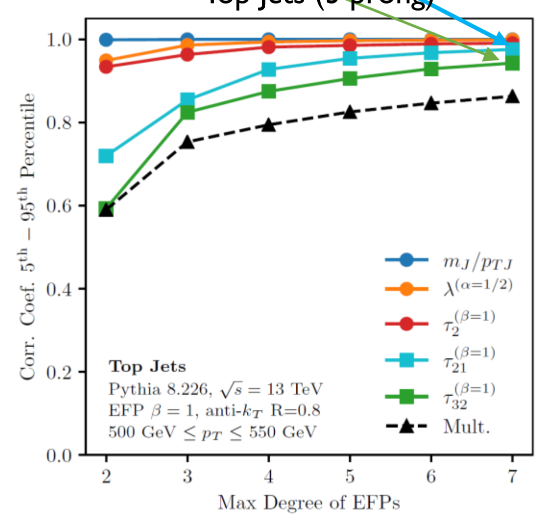
QCD Jets (1 prong)



W Jets (2 prong)



Top Jets (3 prong)



Expected to be IRC safe = Solid.
Expected to be IRC unsafe = Dashed.

Conclusions

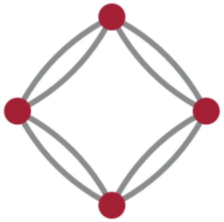
EFPs form a complete, linear representation of the jet

- EFPs energy correlators with monomial angular structure
- Encompass many existing classes of expert variables
- Opens the door to using linear methods for jet substructure
- IRC-unsafe information? Combine!
  - Use EFPs & linearity to reduce radiation pattern to a single optimal observable

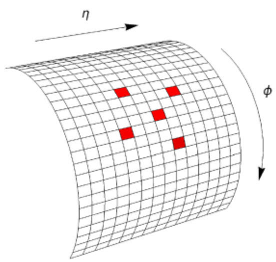
(Linear) Learning is easy

- Linear models are convex & even closed-form at times
- Few or no hyperparameters to tune at all
- Guaranteed global optima

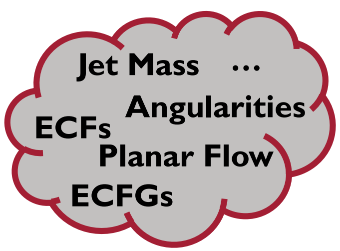
The End



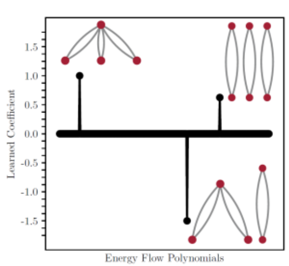
Energy Flow Polynomials (EFPs)



The Energy Flow Basis from IRC safety



Taming the (IRC-safe) Substructure Zoo



Spanning Substructure with Linear Regression