Disentangling Multiple Light-Flavor Jets at Colliders

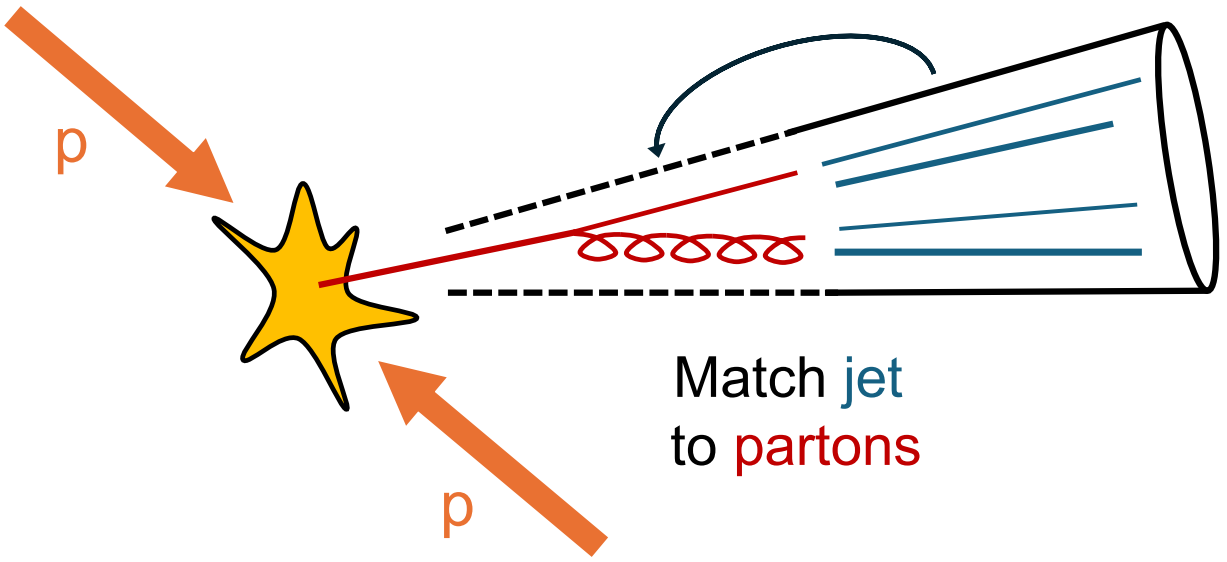
PHENO 2026

Gregorio de la Fuente

05/12/2026

Joint work with Jesse Thaler

Naively, jet flavor is assigned using parton labels from a Monte Carlo generator



This is unphysical. Can we come up with a hadron-level definition of jet flavor instead?

[Gras, Höche, Kar, Larkoski, Lönnblad, Plätzer, Siódmok, Skands, Soyez, Thaler, 1704.03878]

Toward a hadron-level definition of jet flavor

2015: Jet flavor as an enriched region of phase space.

- Well-defined, but impractical.

[Andersen et al., 1605.04692]

2018: An operational definition of quark and gluon jets.

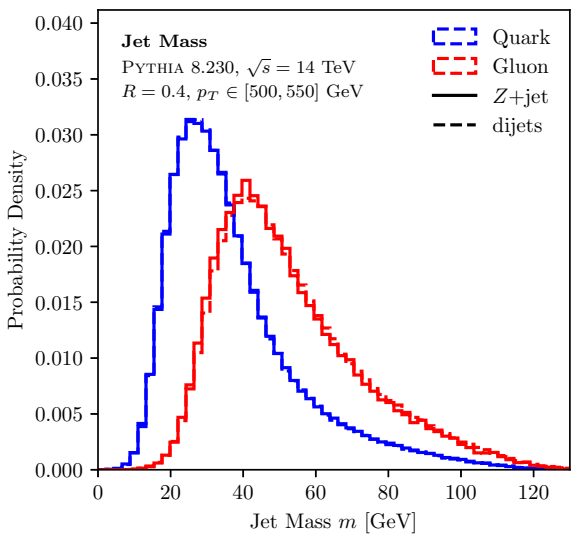
- Practical, but limited to two flavors.

[Komiske, Metodiev, Thaler, 1809.01140]

Today: **A practical definition of multiple jet flavors.**

Quark jet: a phase space region (as defined by an unambiguous hadronic fiducial cross section measurement) that yields an enriched sample of quarks (as interpreted by some suitable, though fundamentally ambiguous, criterion).

Adapted from [Andersen et al., 1605.04692]

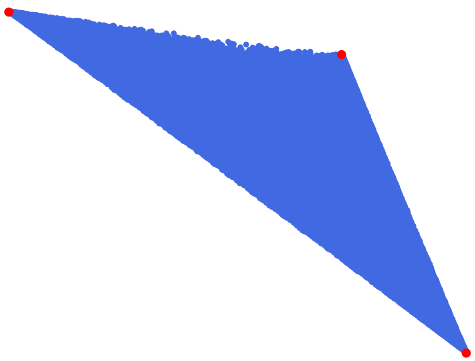


New perspective: jet flavors are the vertices of a simplex

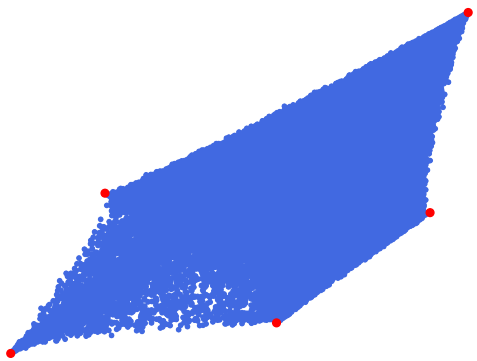
The simplex is the bounding shape of the output of a ML classifier trained on jet samples.

Each jet gets mapped to a point within the simplex, and the flavors are the vertices of the simplex. We extract...

u, d, and g flavors from 3 artificial samples.



five flavors from 14 quasi-realistic dijet samples.



Our method works for an arbitrary number of samples M and flavors $T \leq M$, with enough information.

2018: The operational definition of **two** flavors

Classification without labels (CWoLa). Train a classifier on p_1, p_2 using the cross-entropy loss:

$$L[c_1, c_2] = -\int d^D x \, [p_1 \ln c_1 + p_2 \ln c_2] \,,$$

subject to $c_1, c_2 \geq 0$, $c_1 + c_2 = 1$ (probability 1-simplex).

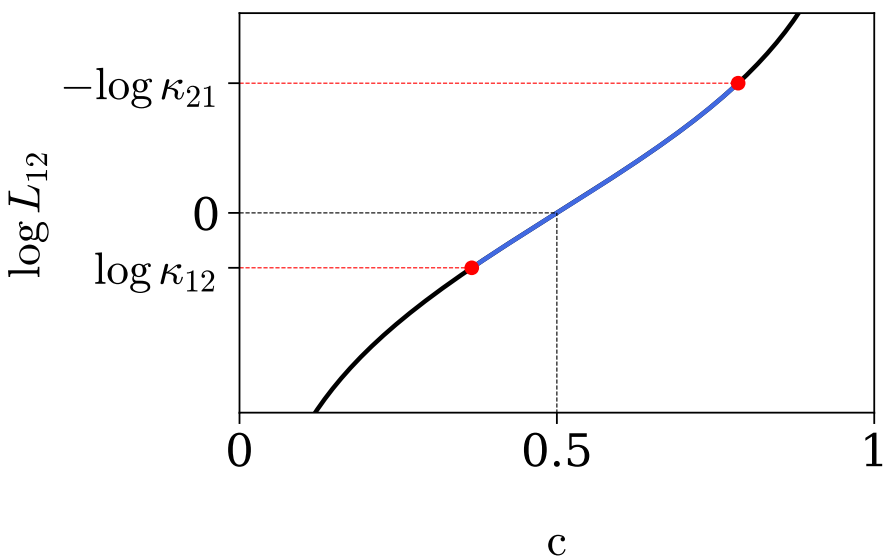
At $\delta L = 0$, the output $c \equiv c_1$ is monotonic to the likelihood ratio $L_{12} \equiv p_1/p_2$.

The jets lie between

$$\kappa_{12} \equiv \min_x L_{12},$$

$$\kappa_{21} \equiv \min_x \frac{1}{L_{12}}.$$

[Metodiev, Nachman, Thaler, 1708.02949]



2018: The operational definition of **two** flavors

Jet topics. Assume a mixture model of separable distributions:

$$p_1 = f_1 p_q + (1 - f_1) p_g,$$
$$p_2 = f_2 p_q + (1 - f_2) p_g.$$

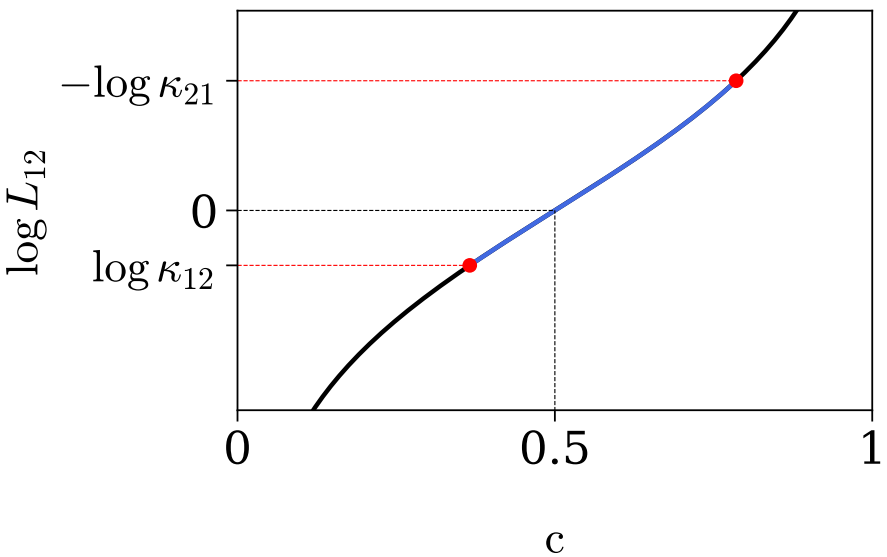
[Metodiev, Thaler, 1802.00008] [Dillon, Faroughy, Kamenik, 1904.04200] [Alvarez, Lamagna, Szewc, 1911.09699]

If $f_1 \neq f_2$, the endpoints give:

$$f_1 = \frac{1 - \kappa_{12}}{1 - \kappa_{12}\kappa_{21}},$$
$$f_2 = \frac{\kappa_{21}(1 - \kappa_{12})}{1 - \kappa_{12}\kappa_{21}}.$$

[Komiske, Metodiev, Thaler, 1809.01140]

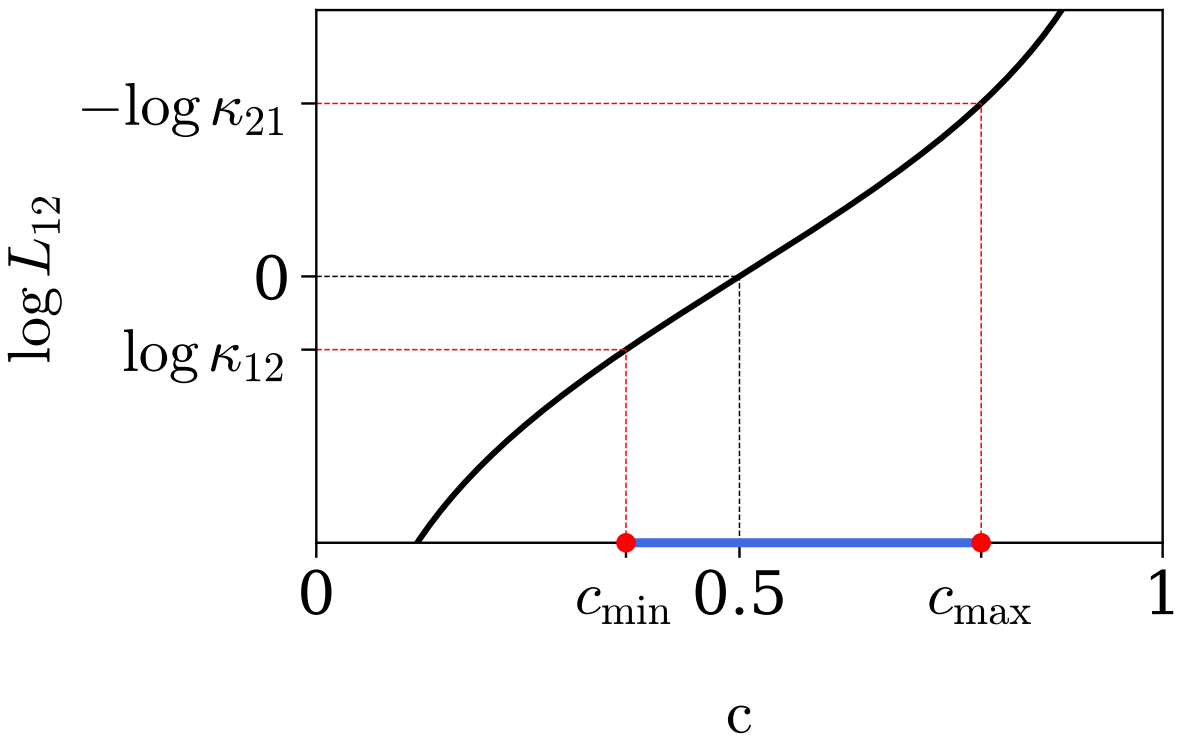
[Katz-Samuels, Blanchard, Scott, 1710.01167]



The classifier distribution is a **geometric object**

Every jet gets mapped between c_{min} and c_{max}: c_{min} is quark-like, while c_{max} is gluon-like (or vice versa).

A line segment = 1-simplex



Can get f_1 and f_2 from c directly!

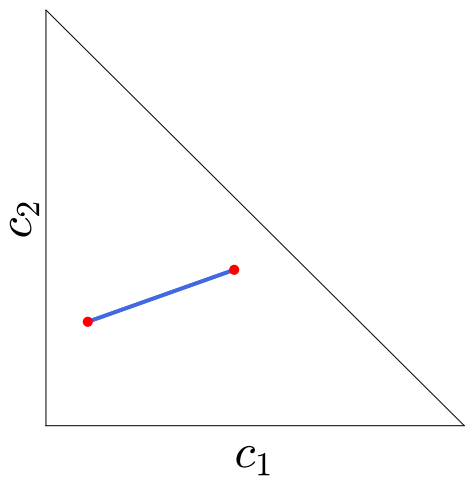
2026: For M samples, the distribution is a $(T-1)$-simplex

As an example, if we train a classifier on $M = 3$ jet samples p_1, p_2, p_3:

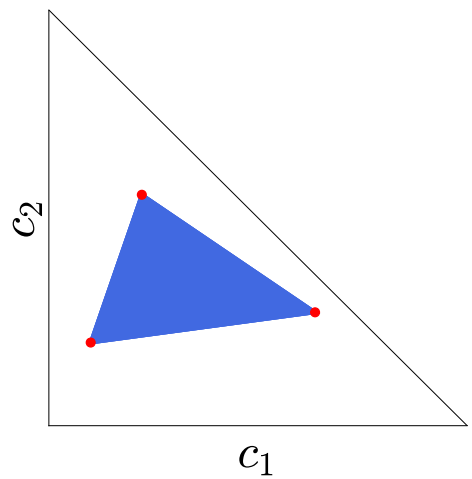
$$L[c_1, c_2, c_3] = -\int d^D x \, [p_1 \ln c_1 + p_2 \ln c_2 + p_3 \ln c_3] + L_{\text{auxiliary}},$$

subject to $c_1, c_2, c_3 \geq 0$, $c_1 + c_2 + c_3 = 1$ (probability 2-simplex).

T = 2 flavors, **1-simplex**:



T = 3 flavors, **2-simplex**:

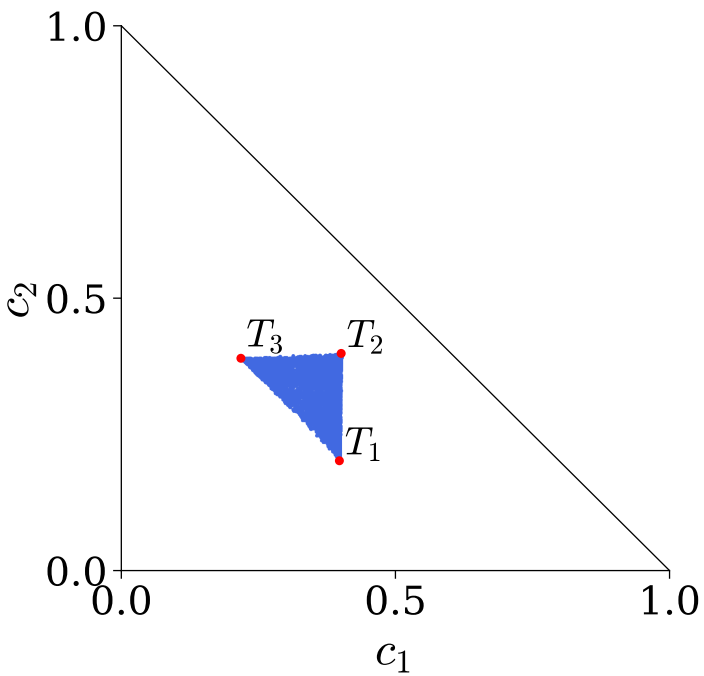


Can get fractions from vertices! [de la Fuente, Thaler, in prep]

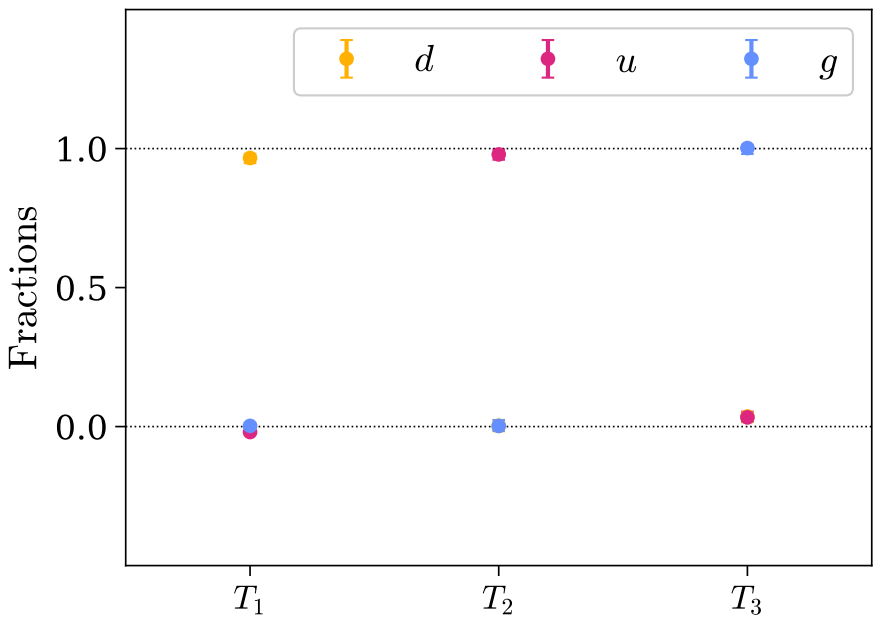
Case study 1: u, d, and g flavors from $M = 3$ samples

We use artificial mixed samples of PYTHIA-generated u-quark, d-quark, and gluon jets: p_1, p_2, and p_3 [Bierlich et al., 2203.11601], and a Particle Flow Network (PFN-ID) architecture. [Komiske, Metodiev, Thaler, 1810.05165].

The classifier learns a 2-simplex with $T = 3$ flavors:



PYTHIA flavors p_f^* in terms of operational flavors p_t, $p_f = F_{ft} p_t$:

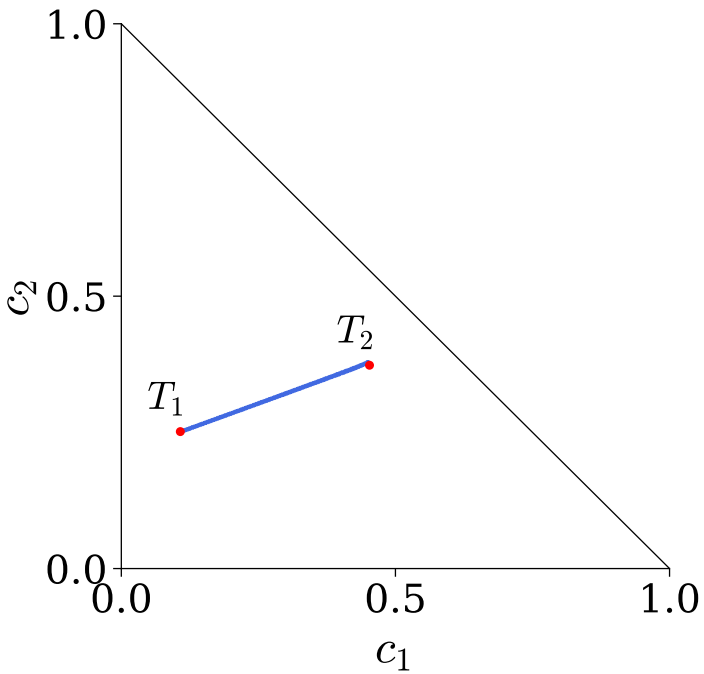


*Up to small non-separability.

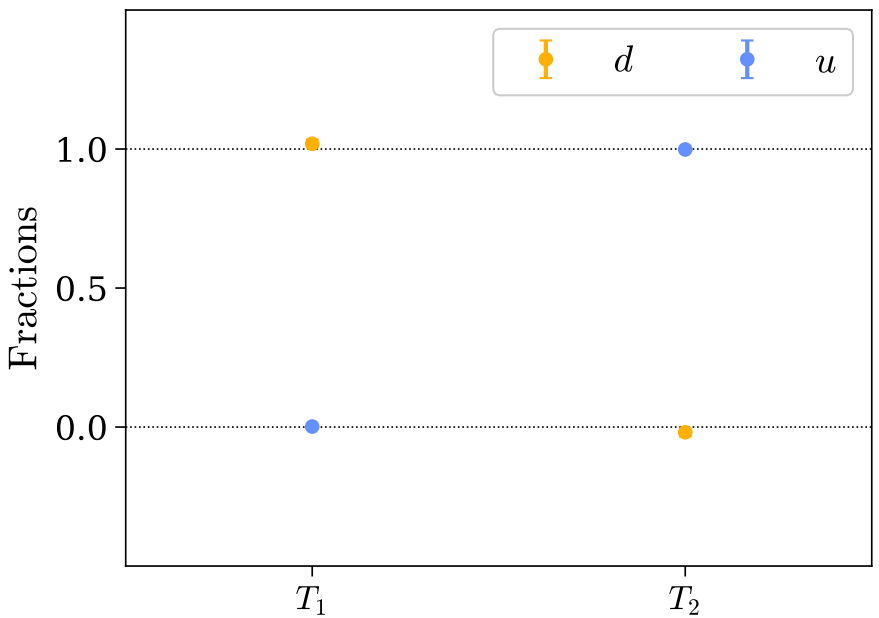
Edge case: what if the gluon flavor is missing?

Use three artificial mixed samples of PYTHIA-generated u-quark and d-quark jets (no gluon jets).

The classifier learns a 1-simplex with $T = 2$ latent flavors.

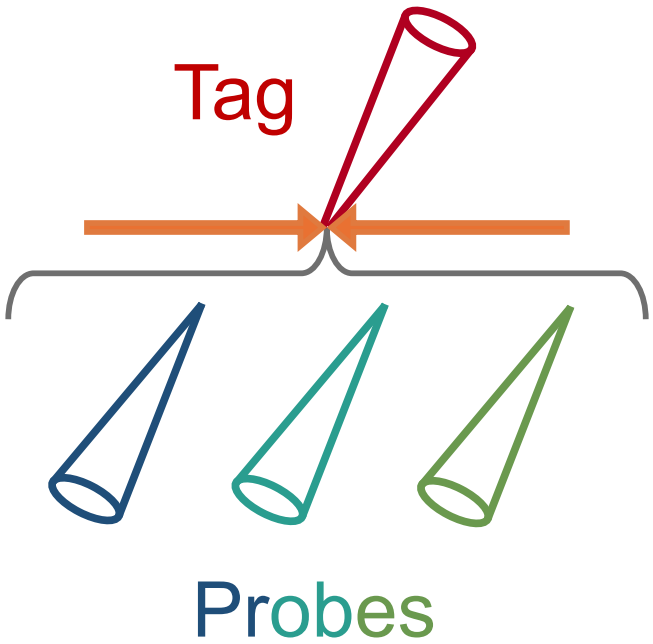


PYTHIA flavors p_f in terms of operational flavors p_t, $p_f = F_{ft} p_t$:



Case study 2: multiple flavors from $M = 14$ samples

Tag-and-probe: each PYTHIA dijet event is split into a "tag" jet and a "probe" jet by randomizing over ϕ.



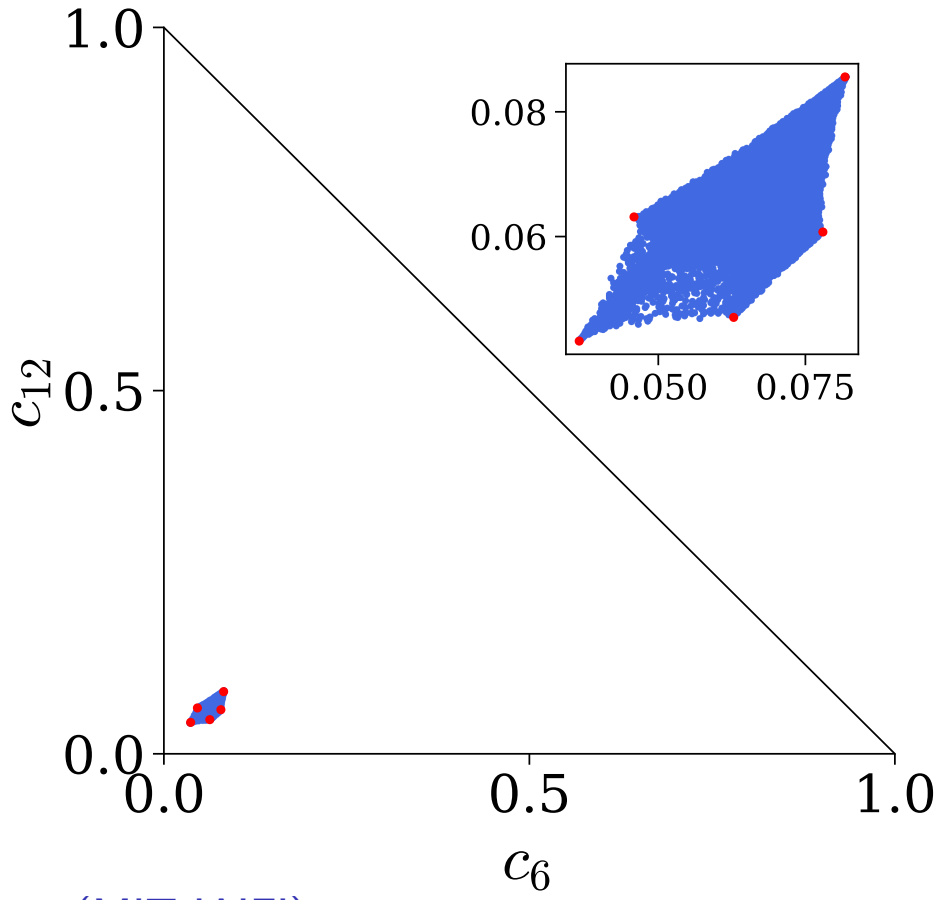
Each sample is defined by collecting all probe jets with the same tag-jet flavor and in the same probe-jet η bin.

(7 light flavors) $\times$ (2 η bins) = 14 samples

We "discover" multiple light flavors in the dijet samples

Our procedure identifies 5 robust flavors in the 14 mixtures.

The flavors emerge as 5 vertices in a probability 13-simplex given by $c_1 + \ldots + c_{14} = 1$, one of whose 2D projections is shown below:



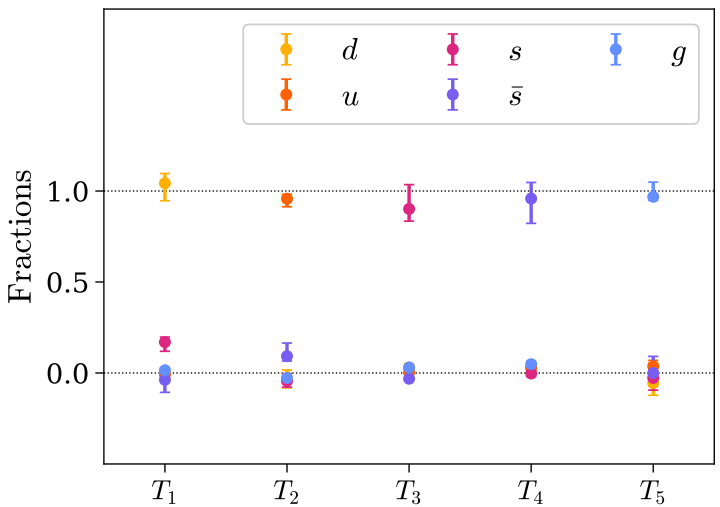
The d, u, s, $\bar{s}$, and g jets are identifiable

Like before, we write the unphysical PYTHIA flavors p_f as a combination of operational flavors p_t,

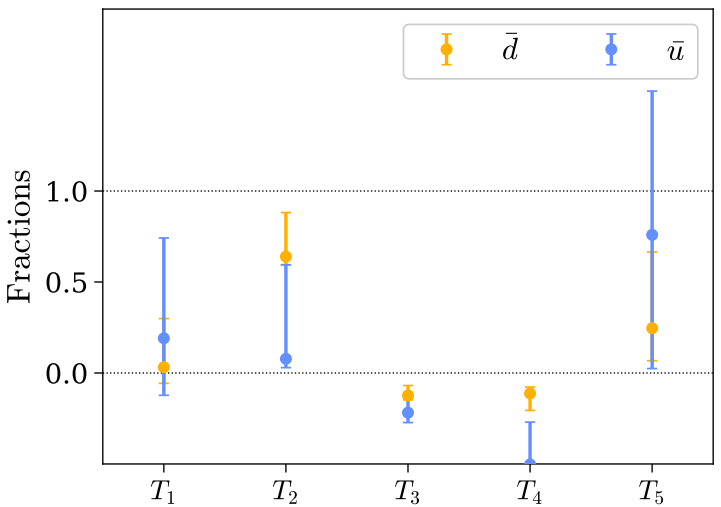
$$p_f = F_{ft}\, p_t.$$

We bootstrap the training data to get error bars on the F_{ft}.

Five flavors are identifiable:



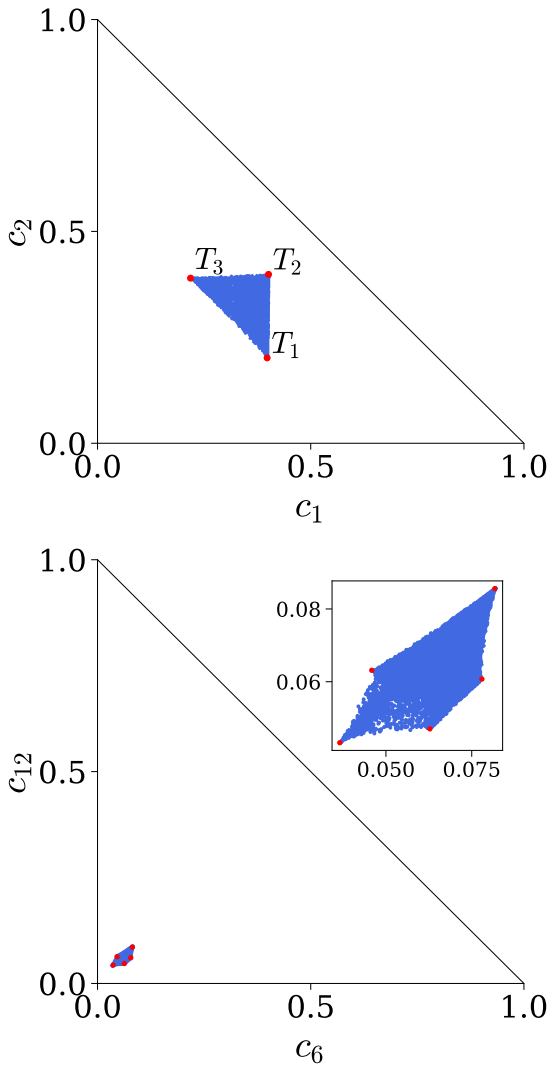
$\bar{d}$ and $\bar{u}$ quarks are not:



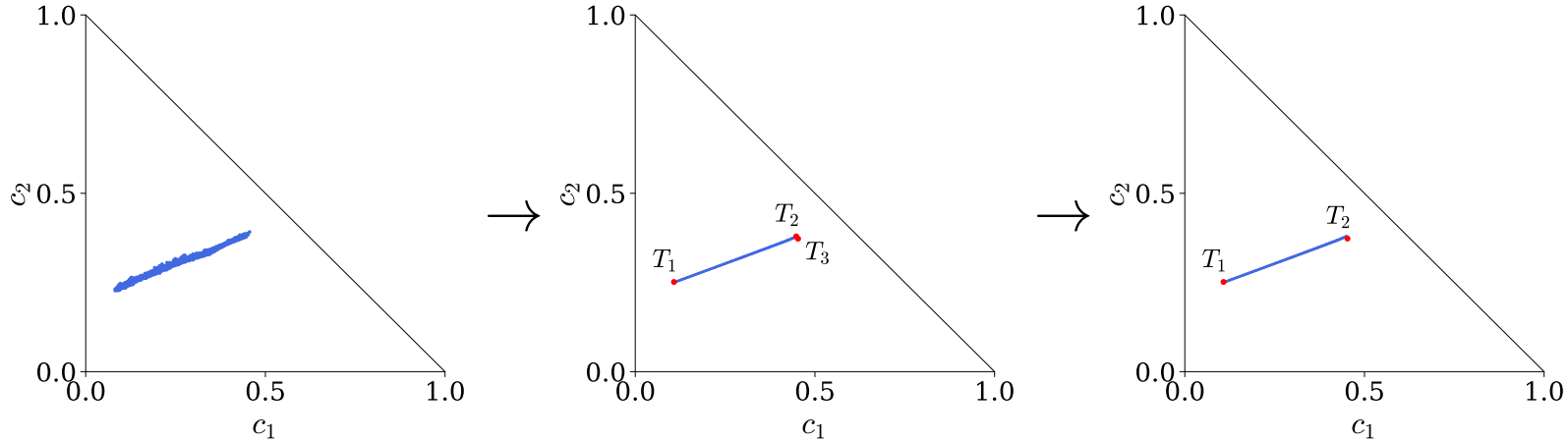
The $\bar{d}$ and $\bar{u}$ fractions are not so different across mixtures, making them less identifiable.

Summary. Email me at grego137@mit.edu!

1. We have provided a hadron-level, data-driven definition of multiple jet flavors.
2. Jet flavors are vertices of a classifier distribution in our framework.
3. In a toy example, we extracted down, up, and gluon flavors that agree with the PYTHIA parton-level flavors.
4. In a quasi-realistic example, we identified the down, up, strange, anti-strange, and gluon flavors in a data-driven way.
5. **Next steps:** make the procedure fully deployable on real data.



What is $L_{\text{auxiliary}}$?



Turn on a *perimeter* loss:

$$\alpha \sum_{t>t'} d\left(\vec{V}_t, \vec{V}_{t'}\right).$$

Turn on an L1 loss:

$$\beta \sum_{t} w_t.$$