Information Geometry of Latent Representations for Interpretable Jet Tagging

Part II: What does geometry tell us about jets?

ML4Jets Vienna 2026 [2603.02310]

Benedikt Schosser | Institute for Theoretical Physics, Heidelberg University

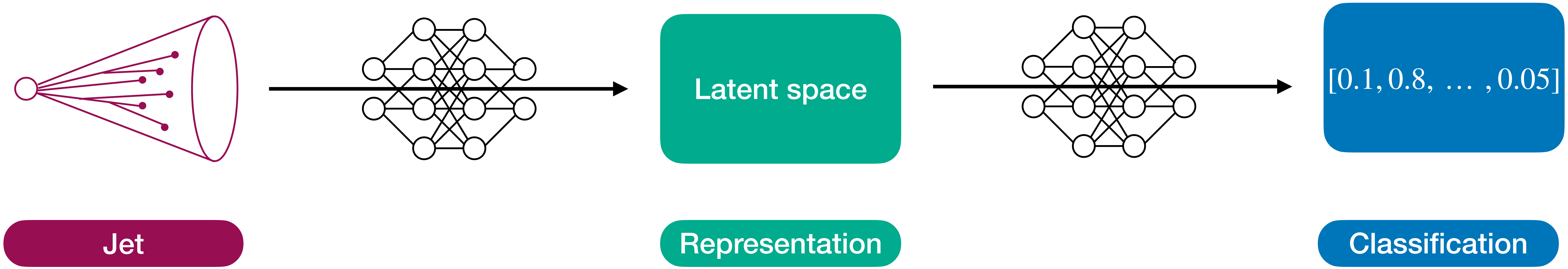
Rebecca Maria Kuntz, Tilman Plehn, Björn Malte Schäfer, Sophia Vent

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Why understand a jet tagger?

Modern jet tagging:

Powerful classification from a
learned high-dimensional representation



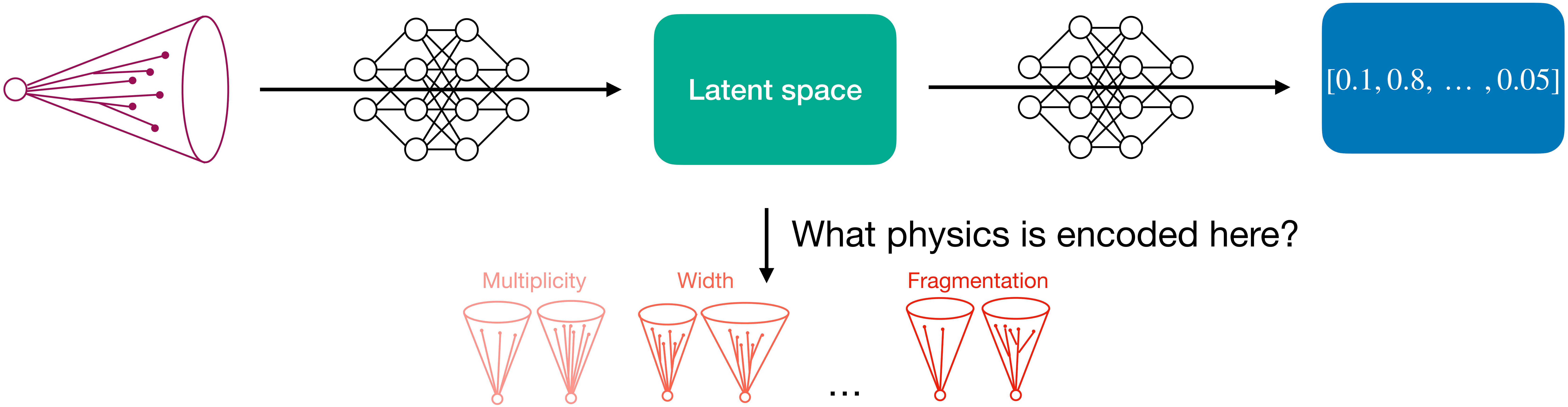
Why understand a jet tagger?

Modern jet tagging:

Powerful classification from a learned high-dimensional representation

Physical questions:

- Which physical features does the classifier use?
- Where in the latent space do they matter?
- How does the classifier move between physical jet configurations?



Quark-gluon tagging

Quark-gluon tagging

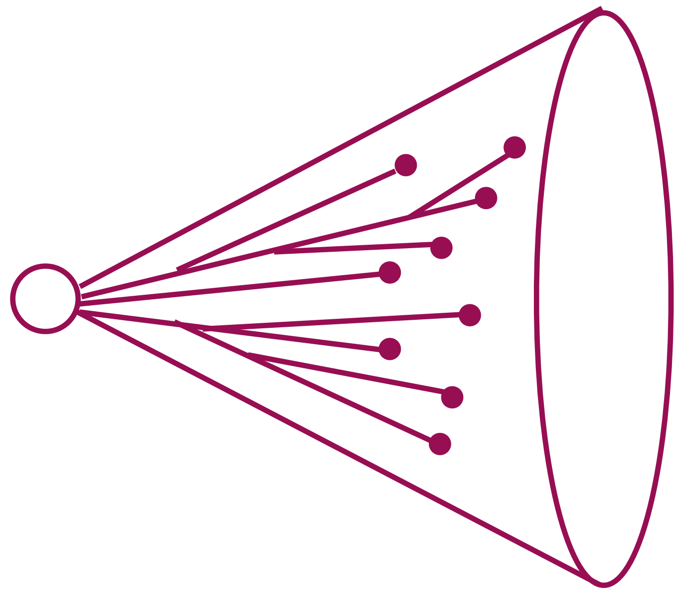
Physics expectation:

Gluon jet

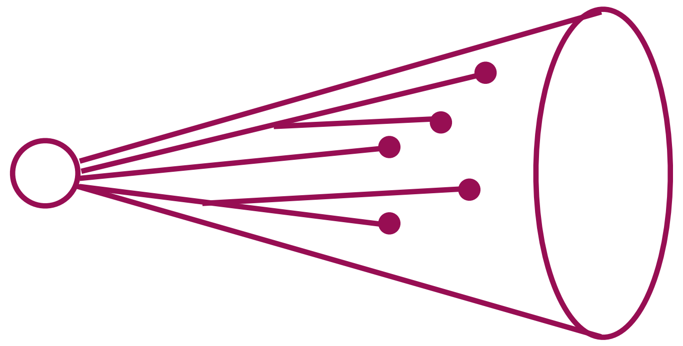
- More particles
- Wider
- Softer fragmentation

Quark jet

- Fewer particles
- Narrower
- Harder fragmentation



$$\boxed{C_A > C_F}$$



Quark-gluon tagging

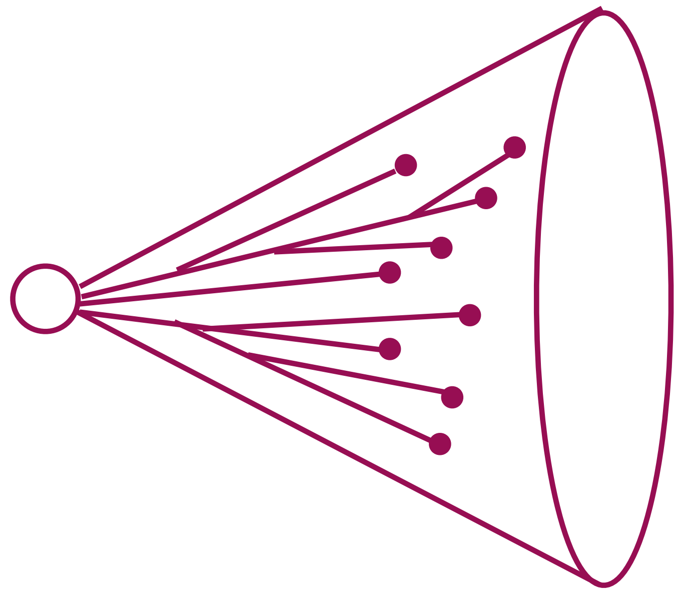
Physics expectation:

Gluon jet

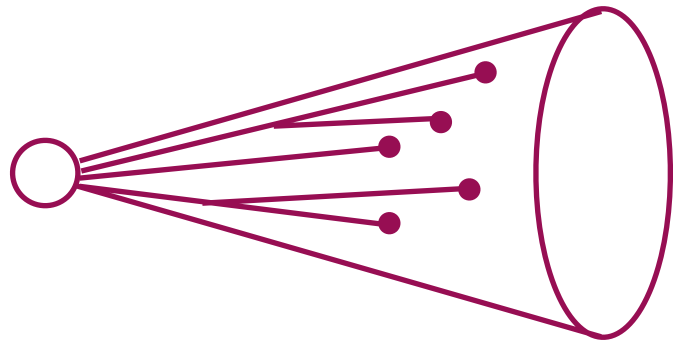
- More particles
- Wider
- Softer fragmentation

Quark jet

- Fewer particles
- Narrower
- Harder fragmentation



$\boxed{C_A > C_F}$

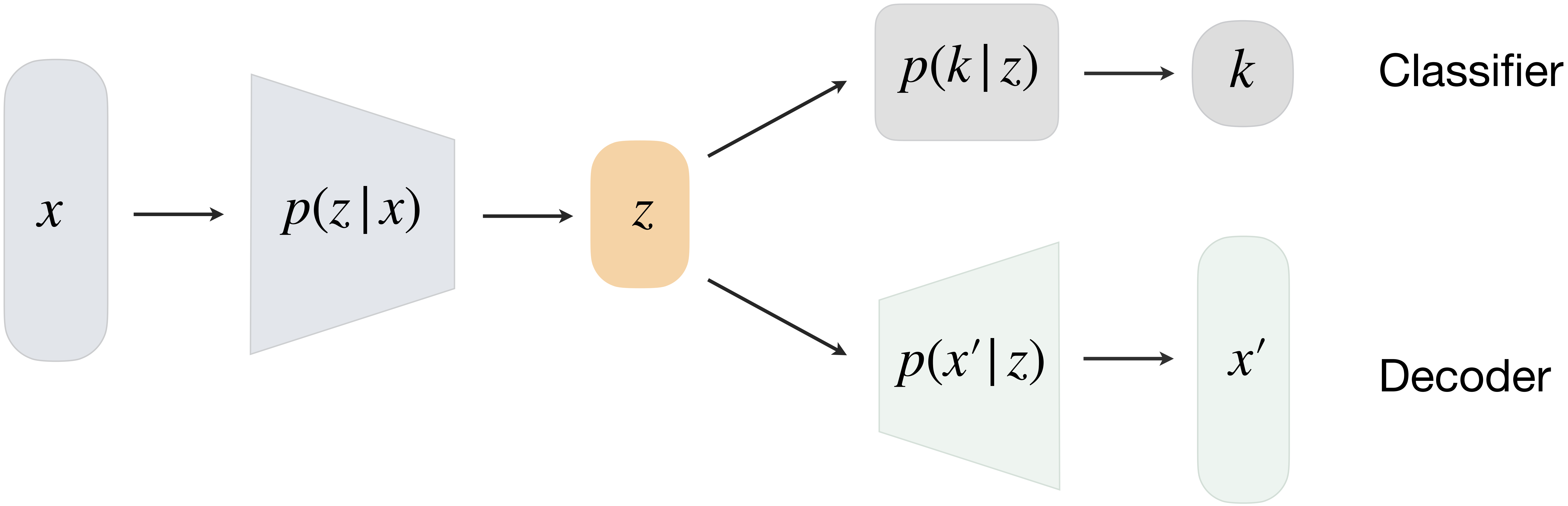


However …

- q/g labels are not unambiguous beyond leading order
- Discrimination is sensitive to showering and hadronization

Does the learned representation organize itself according to this expected QCD structure?

Setup — variational autoencoder

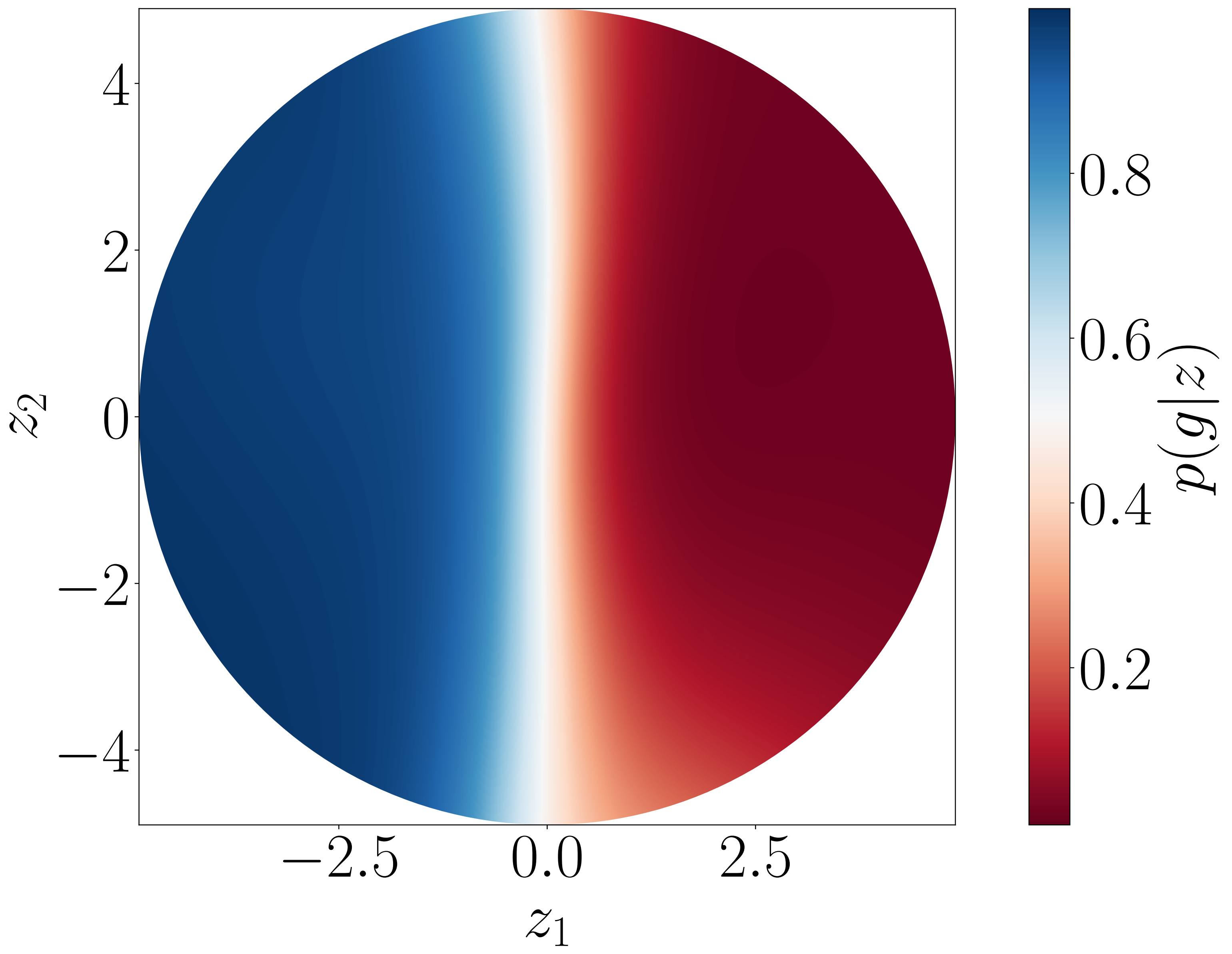


Latent space $z \sim \mathcal{N}(\mu, \sigma)$

Reconstructed features: $\{n_{\mathrm{PF}}, w_{\mathrm{PF}}, p_T D, C_{0.2}, \epsilon, m_{\mathrm{Jet}}, x_{\max}, S_{\mathrm{frag}}\}$

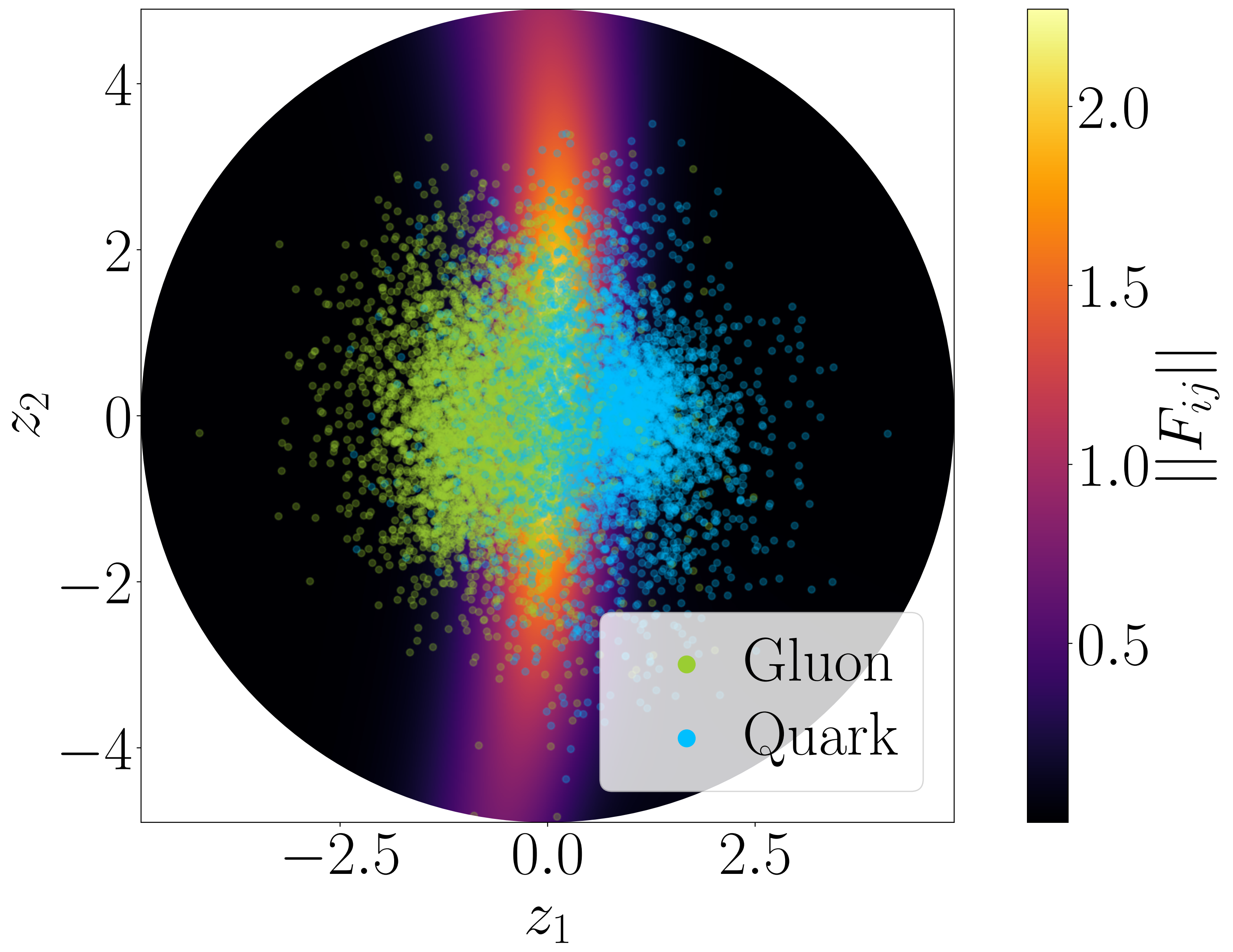
Classifier metric

Decision boundary



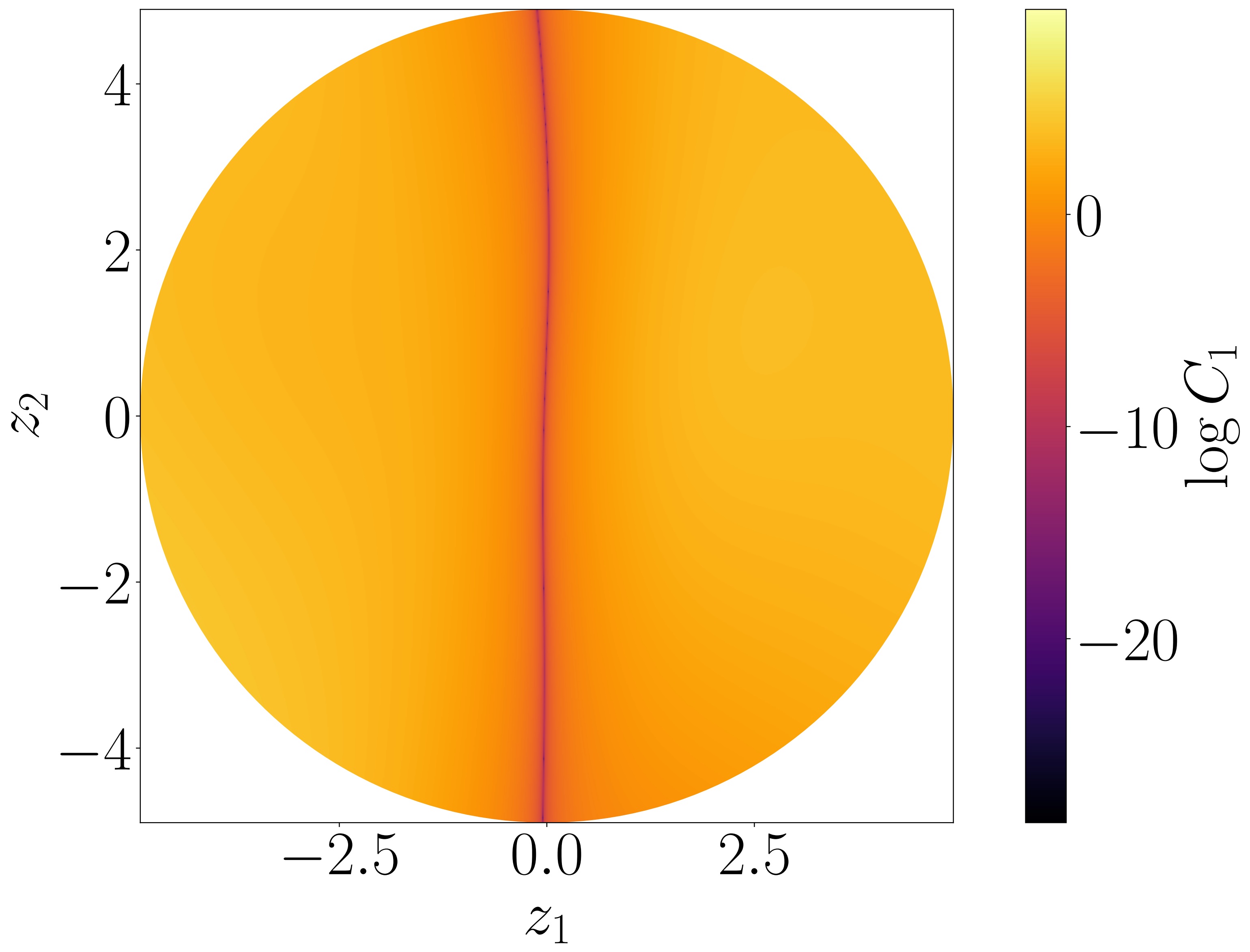
$$p = \frac{1}{2}$$

Fisher metric



Network sensitivity

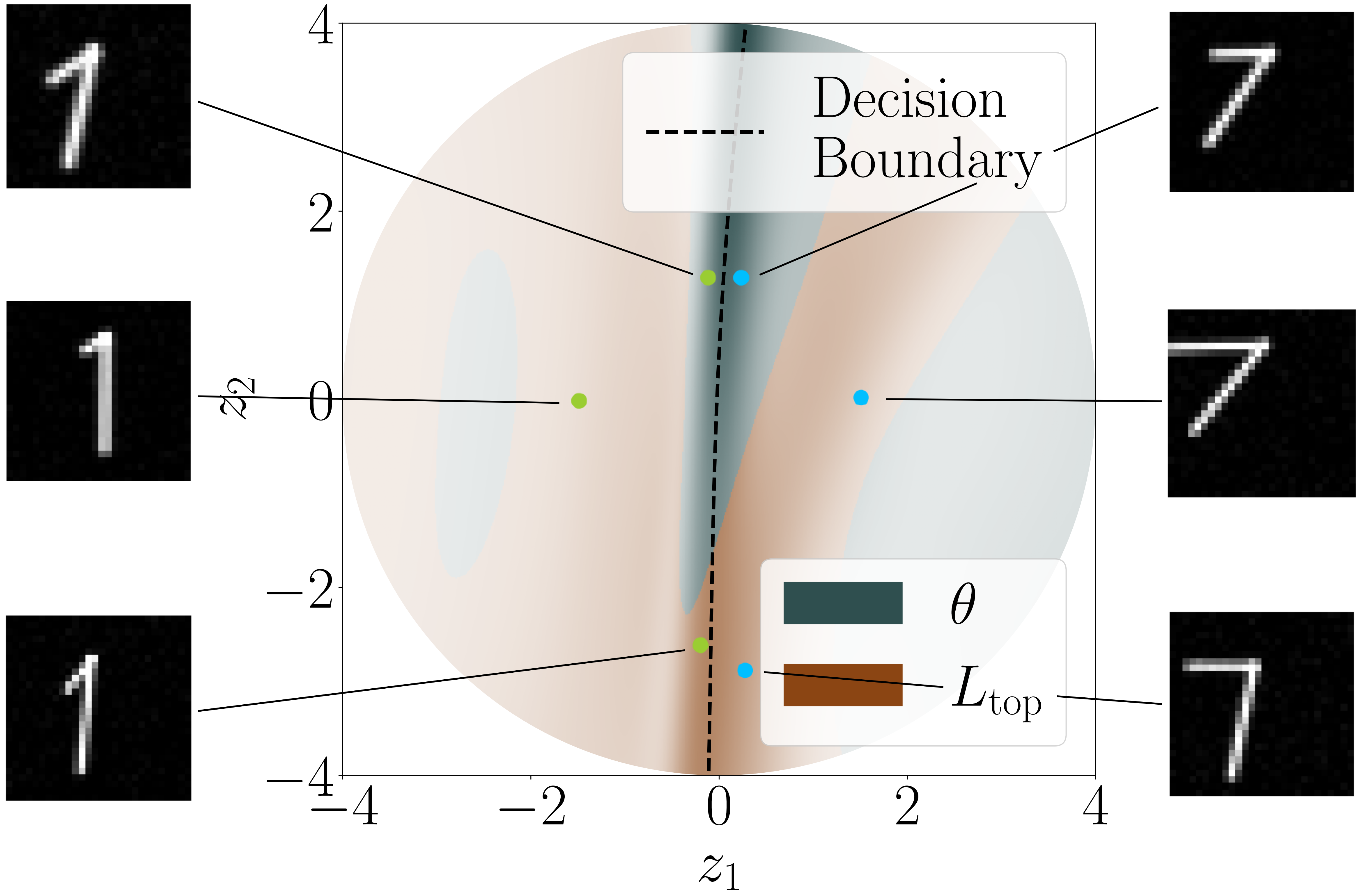
Fully contracted ACT



Likelihood skewness

Which features change with the decision?

Reminder MNIST

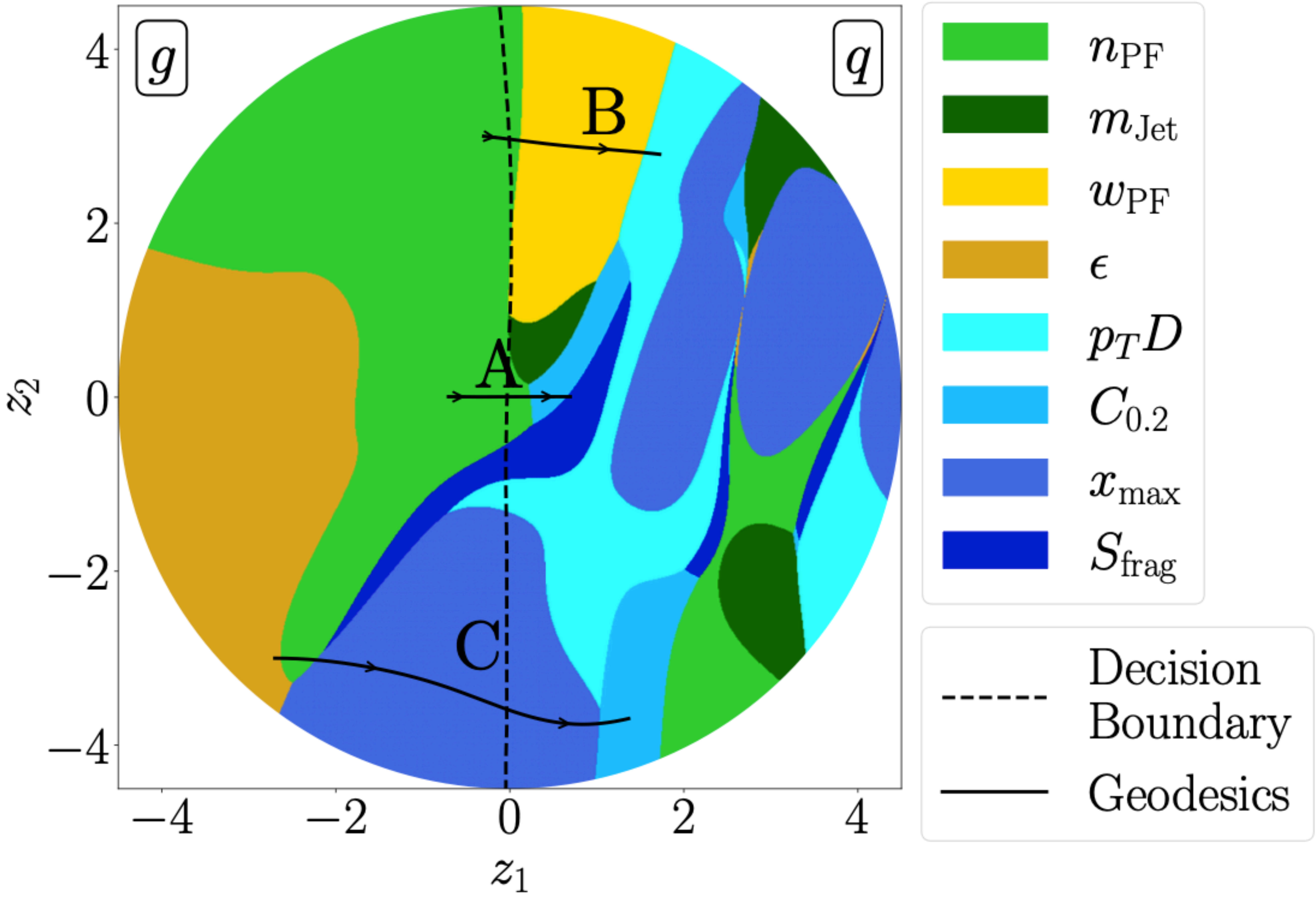


Fisher directional derivative:

How do the reconstructed features change in the most sensitive direction of the classifier?

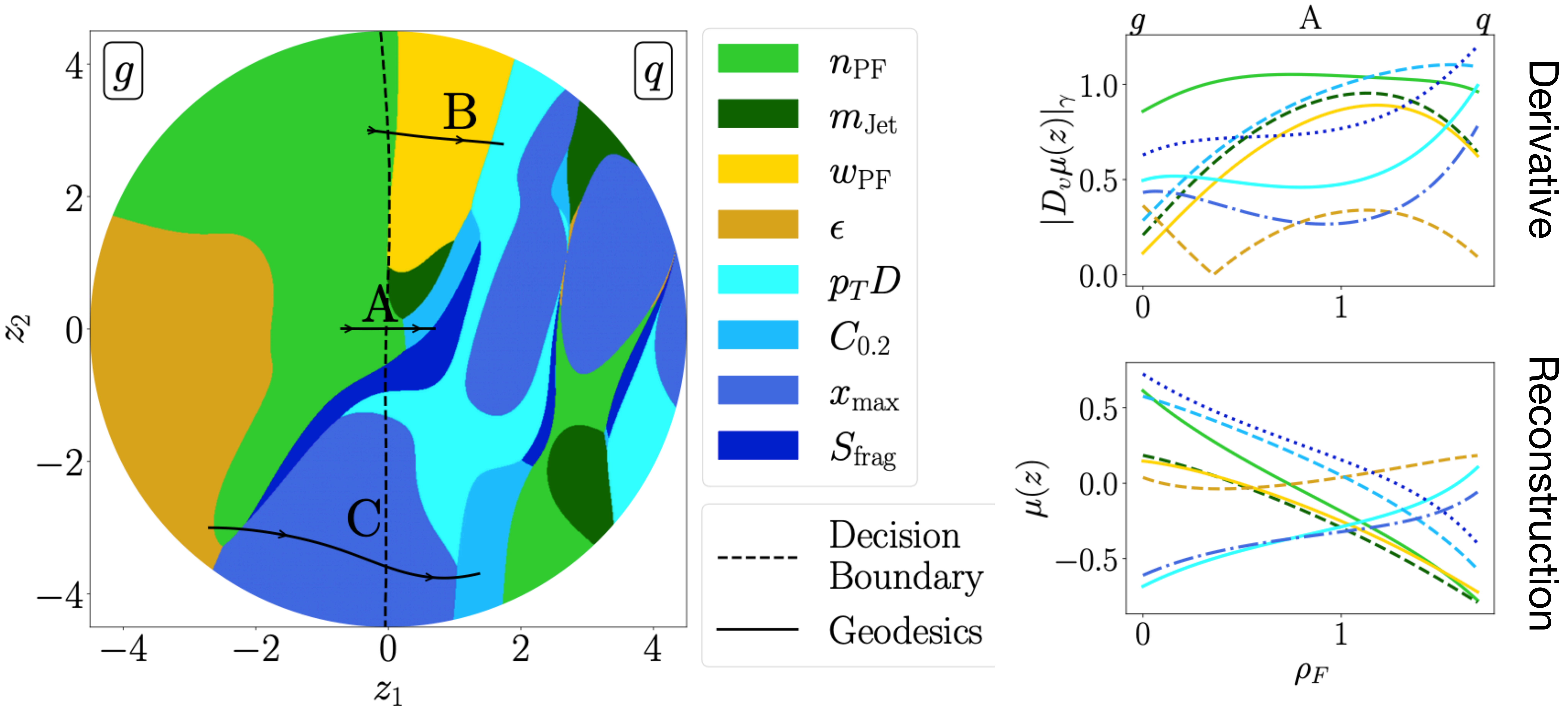
“Which feature is best aligned with the decision”

Which features change with the decision?



- Gluon-like region: dominated by multiplicity
- Quark-like region: more heterogeneous
- The latent space separates into regions governed by **multiplicity**, **fragmentation** & **radiation shape**

Which features change with the decision?

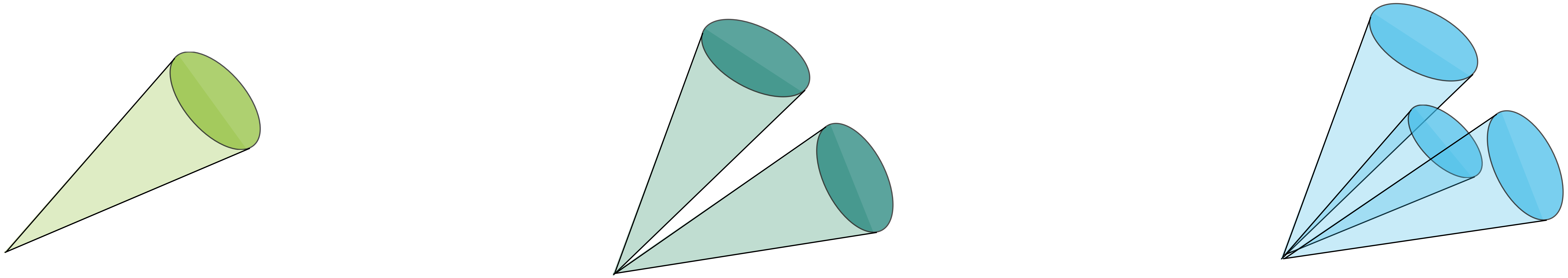


Along a classifier geodesic, we can track how the physics changes

Binary $\rightarrow$ multiclass

Three-fold jet tagging

q/g $\quad$ $Z \rightarrow q\bar{q}$ $\quad$ $t \rightarrow bq\bar{q}'$

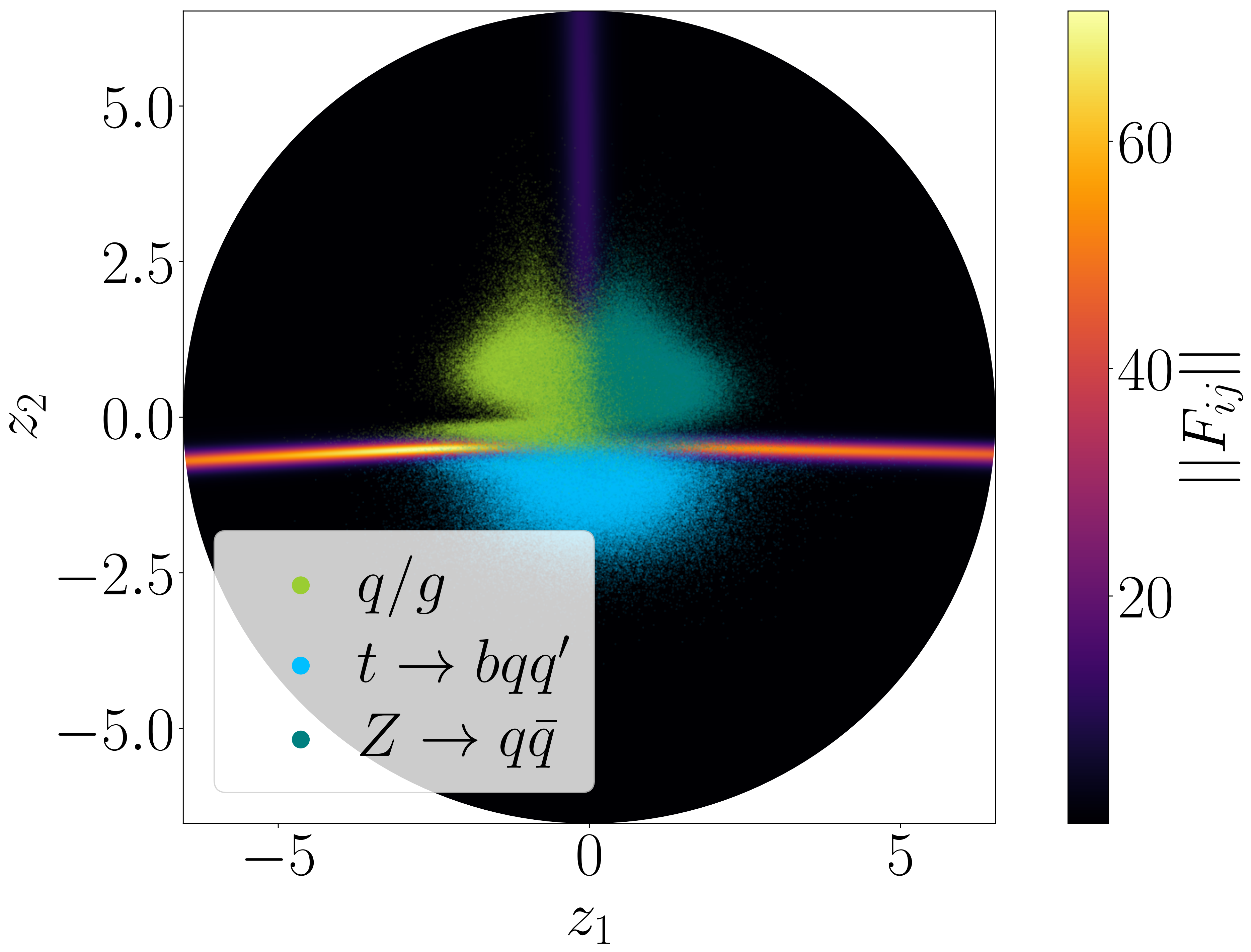


1-prong $\quad$ 2-prong $\quad$ 3-prong

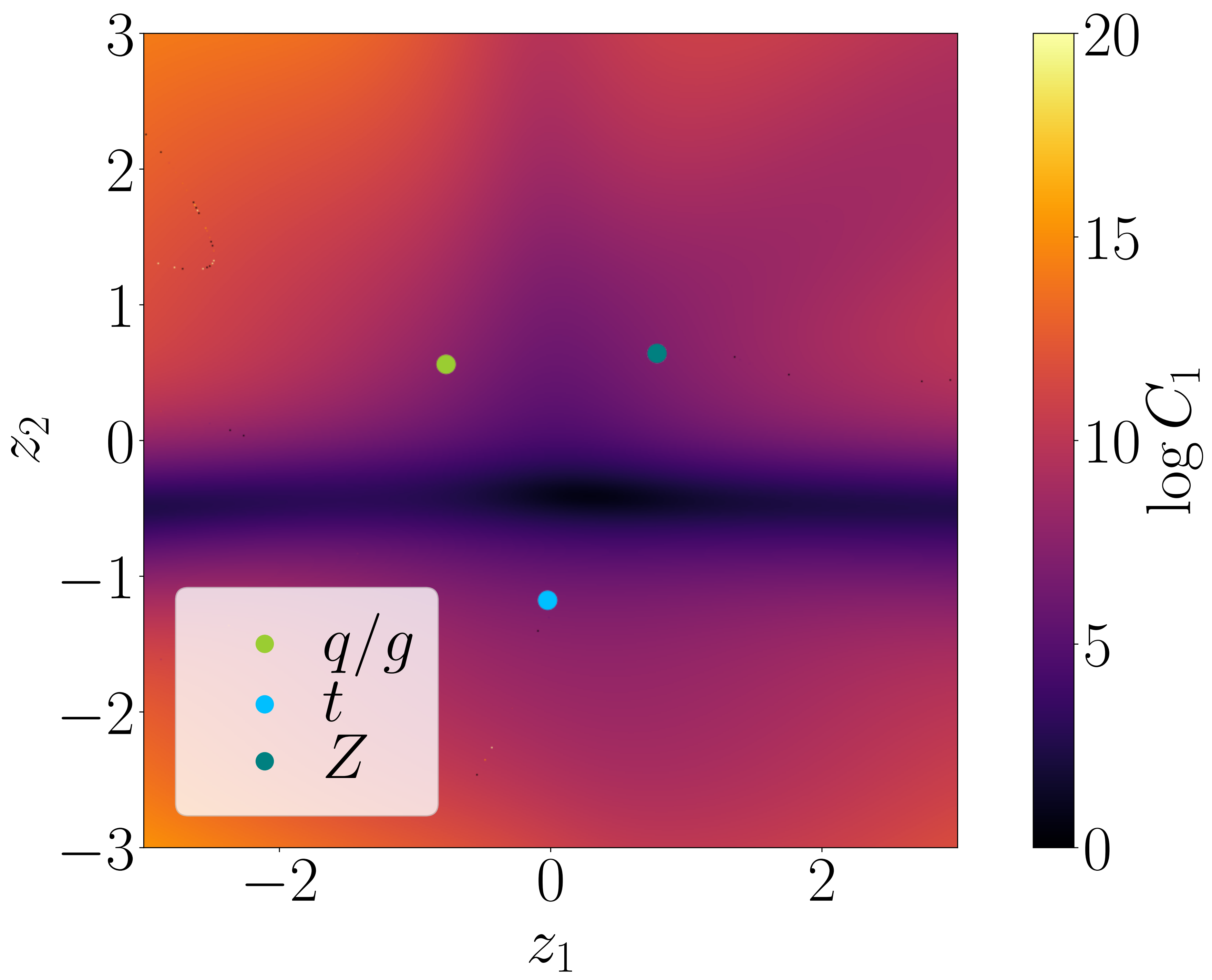
Features: m_{Jet} $\quad$ $\tau_{21} : 1 \leftrightarrow 2$ - prong $\quad$ $\tau_{32} : 3 \leftrightarrow 2$ - prong

Classifier metric

Fisher metric

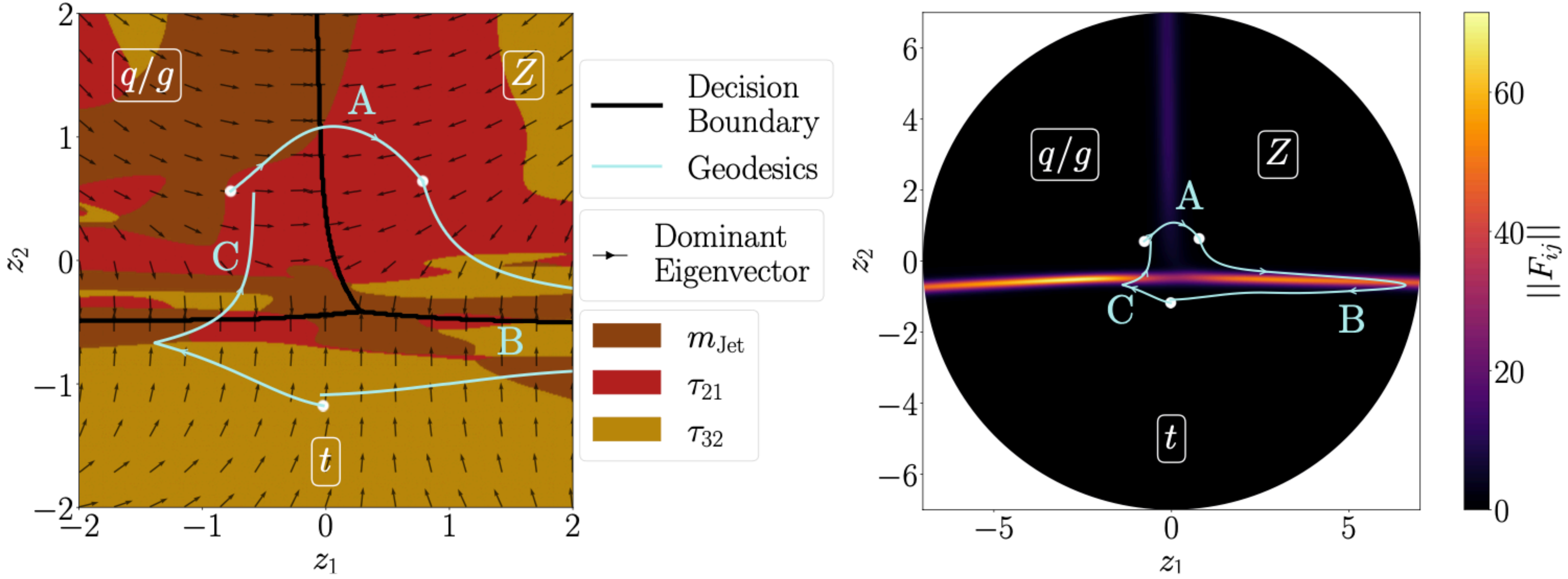


Fully contracted ACT



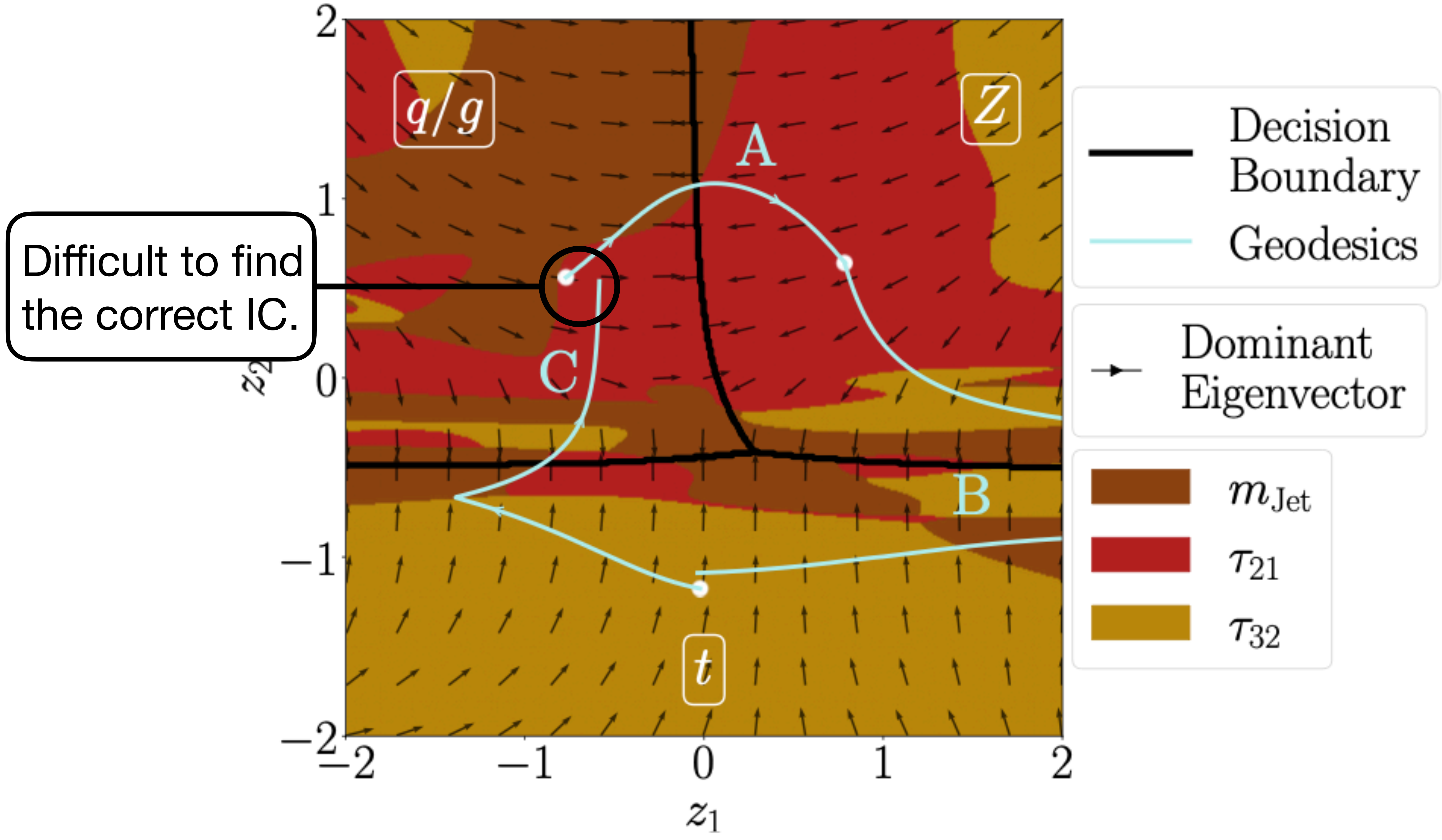
Top jets form a distinct region, while q/g and Z remain more closely connected

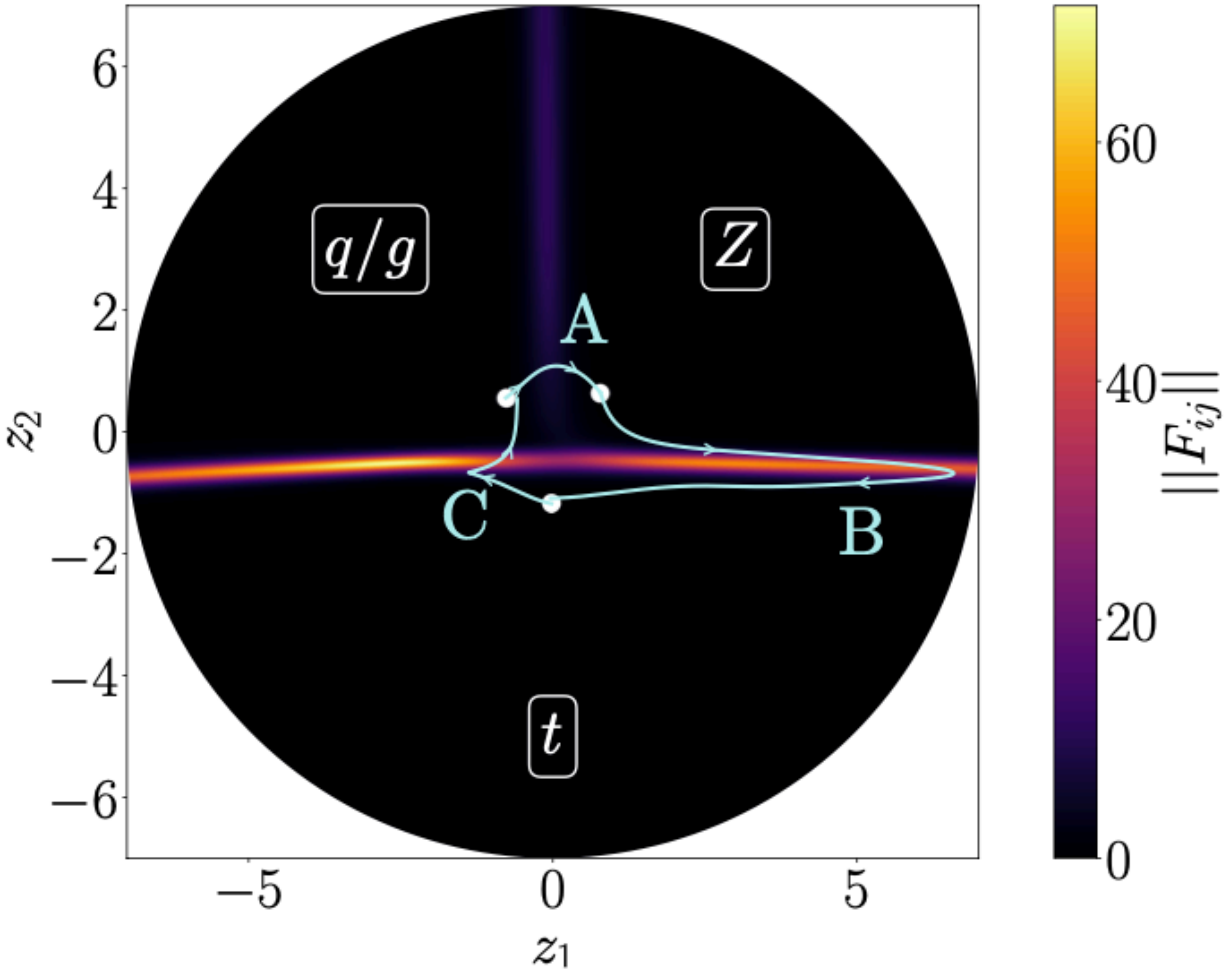
Dominant features & geodesics



- Substructure observables dominate most of the latent space
- Jet mass becomes important near class boundaries
- The natural $Z \to$ top path avoids the q/g region

Dominant features & geodesics

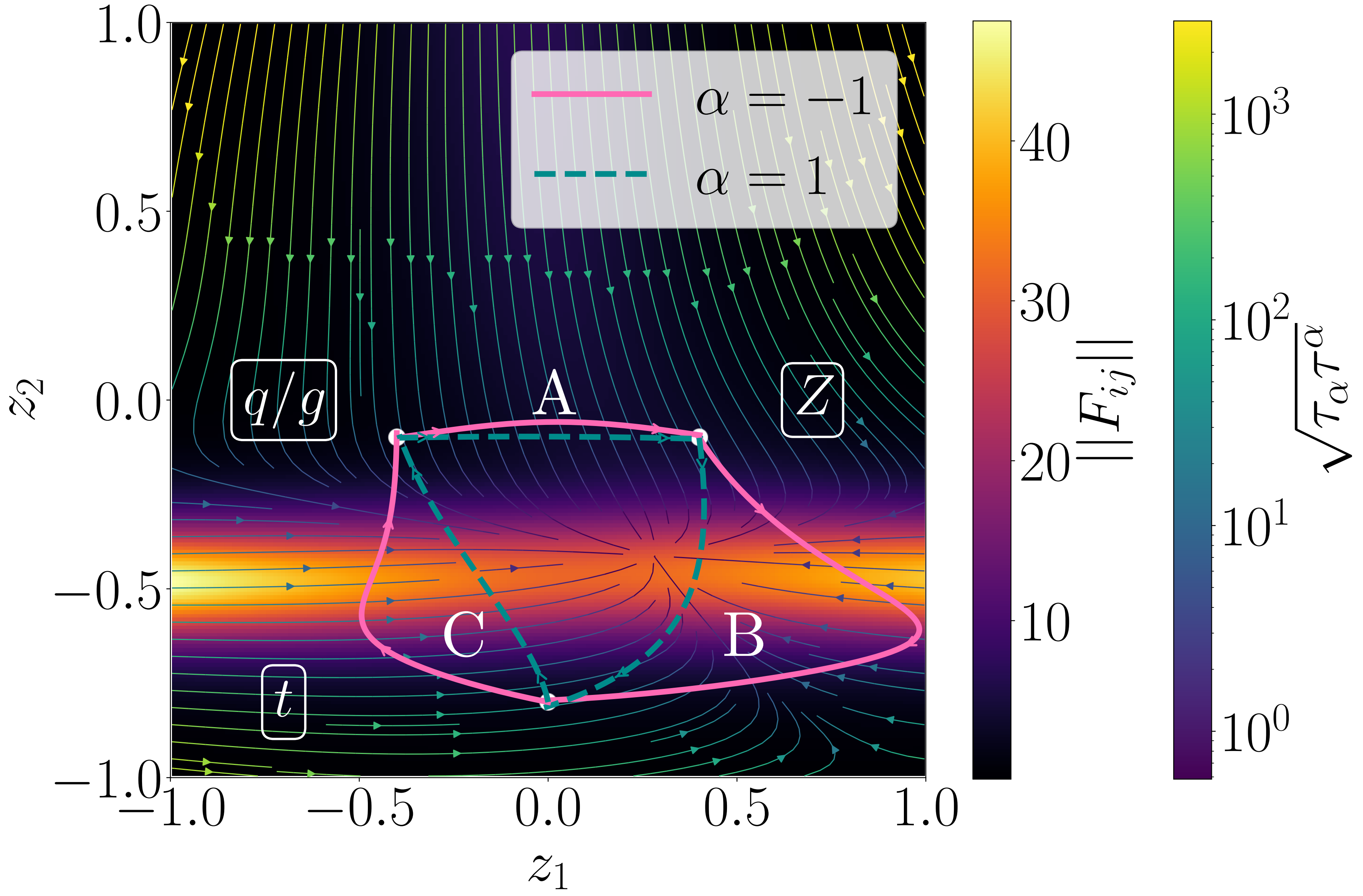




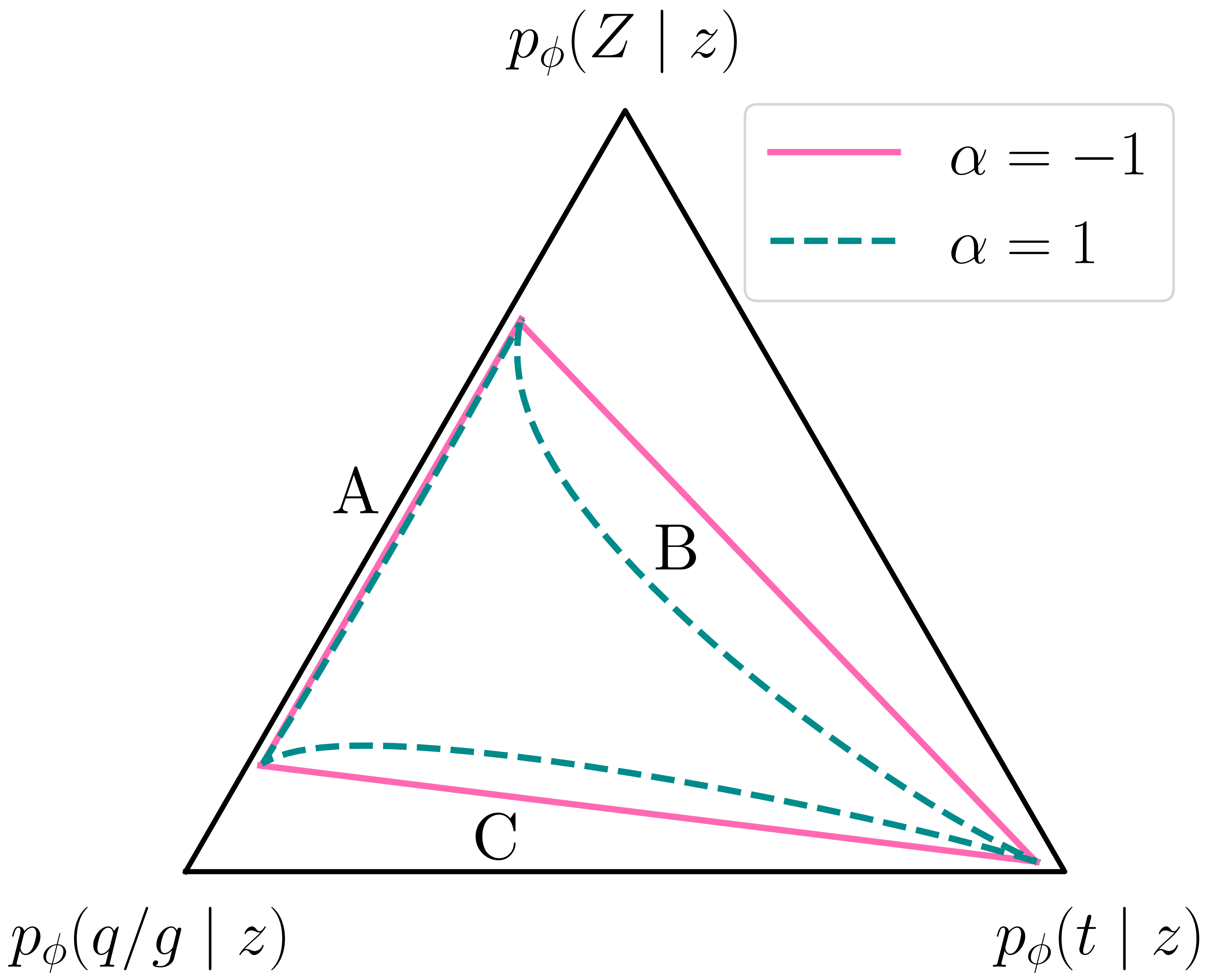
- Substructure observables dominate most of the latent space
- Jet mass becomes important near class boundaries
- The natural $Z \to$ top path avoids the q/g region

Dual structure

Learned latent space



Simplex

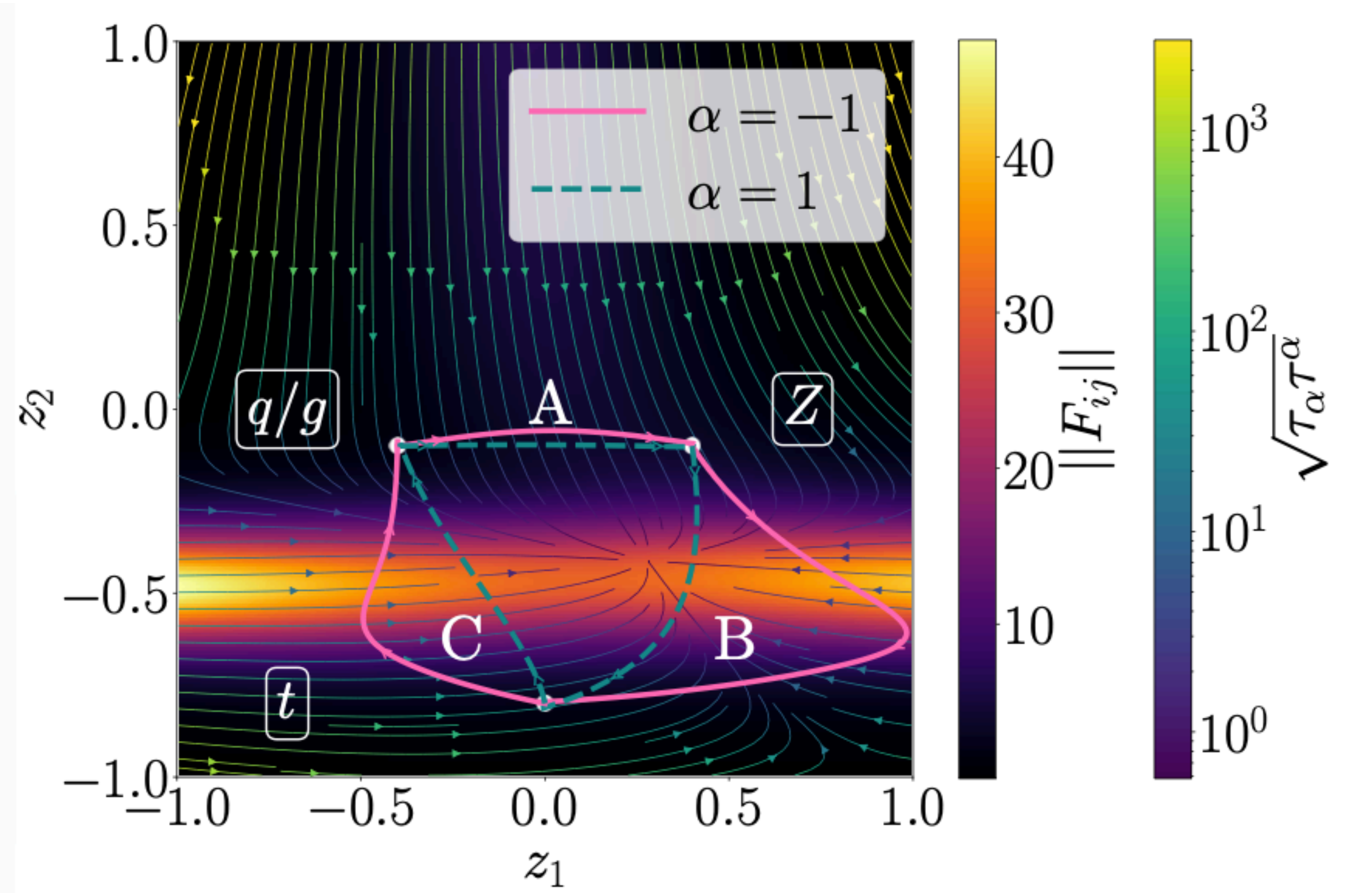


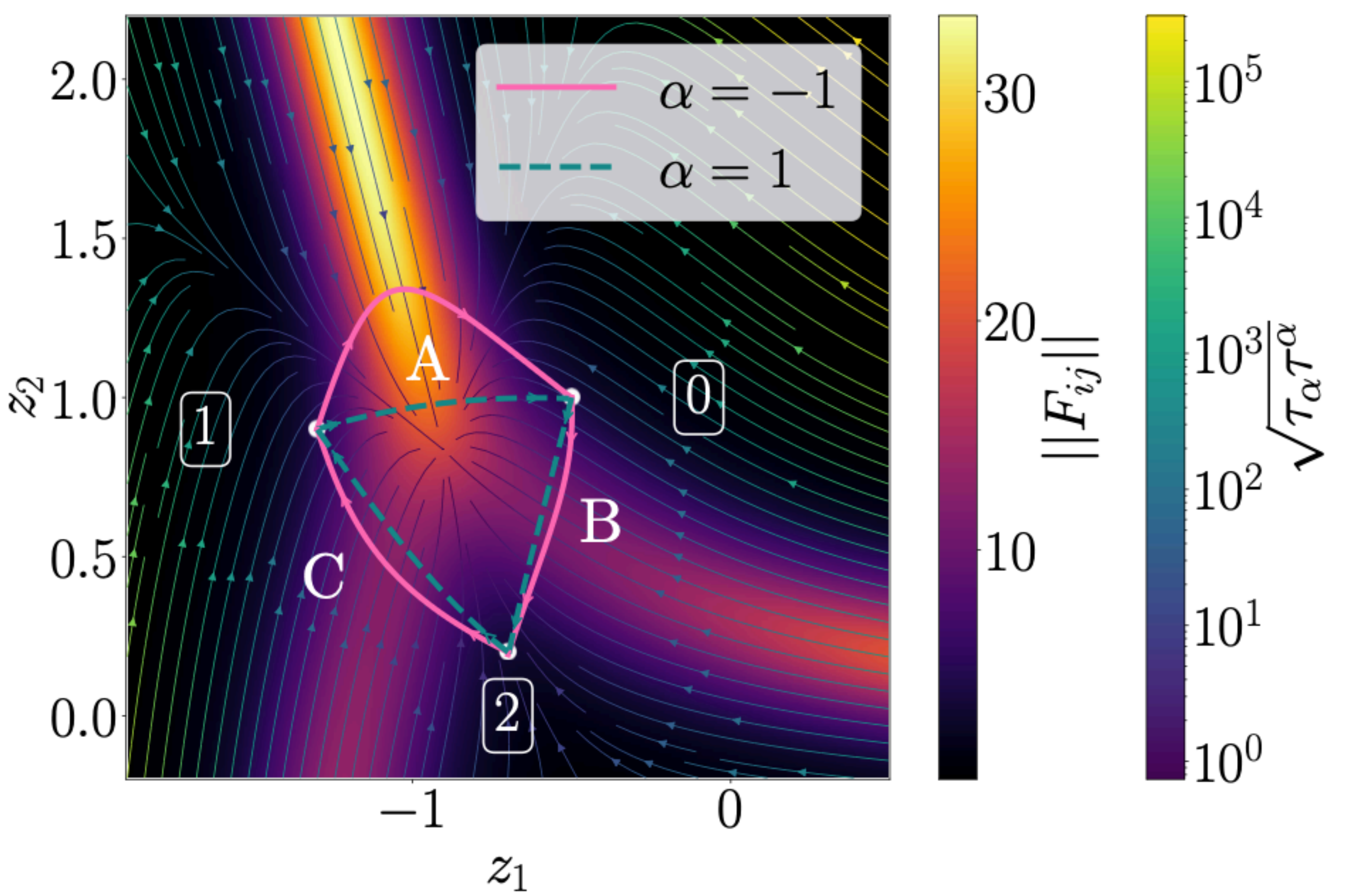
Connections Γ derived from

- Forward KL: $\alpha = -1$
- Backward KL: $\alpha = +1$

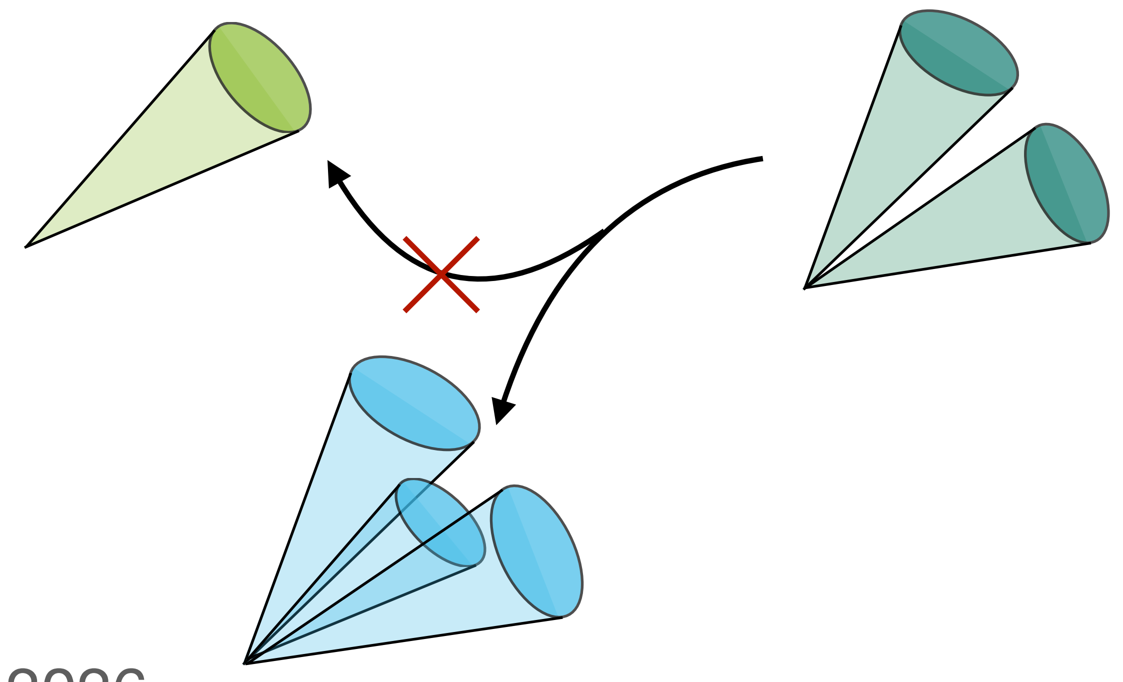
- Same endpoints, different natural paths
- The difference is largest when the third class becomes relevant

Dual structure

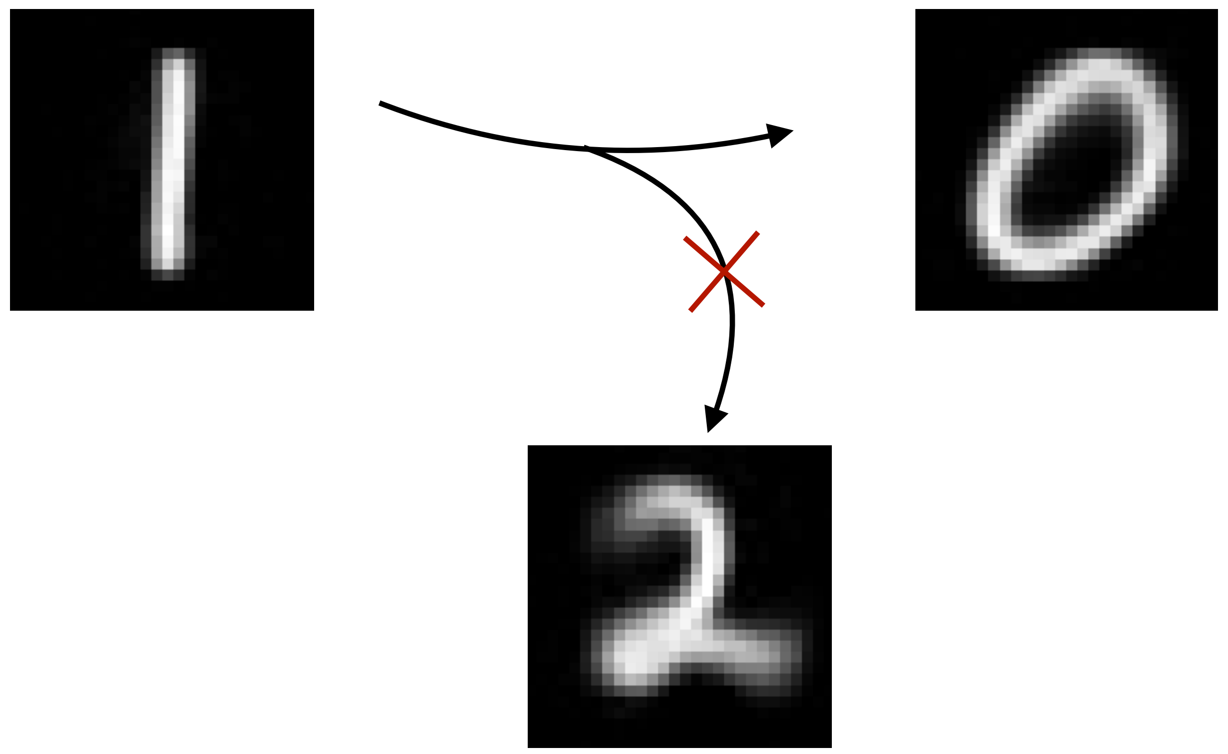




- The $Z \rightarrow$ top path is shaped by the q/g region



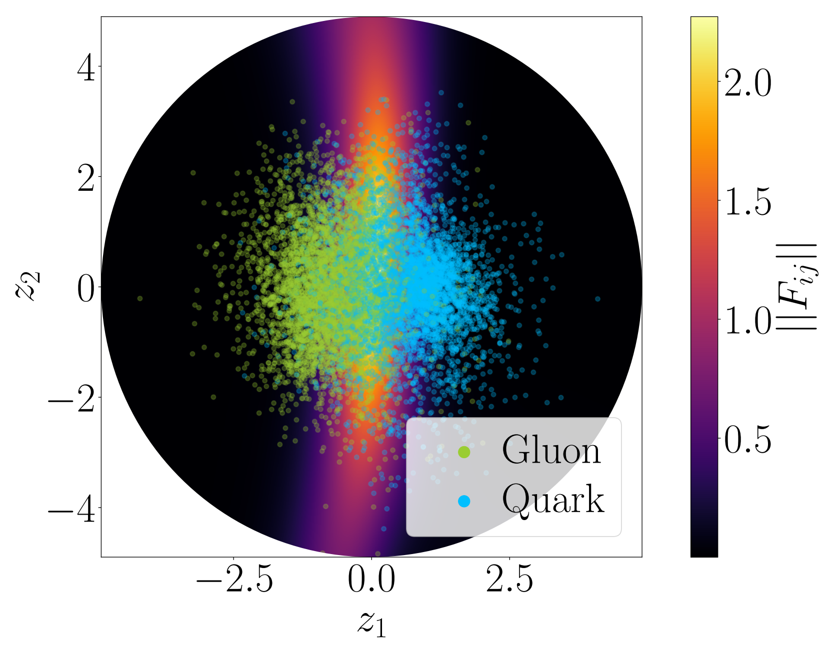
- The $1 \rightarrow 0$ path is shaped by the 2 region



Information gained through geometry

Where?

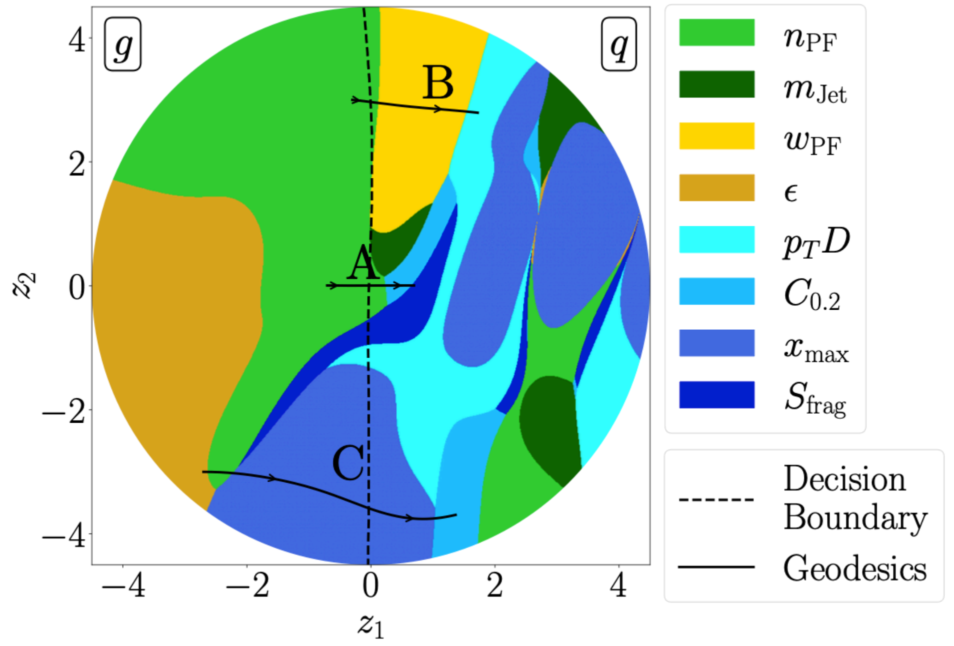
Classifier geometry



Identifies classifier-sensitive regions and directions

What?

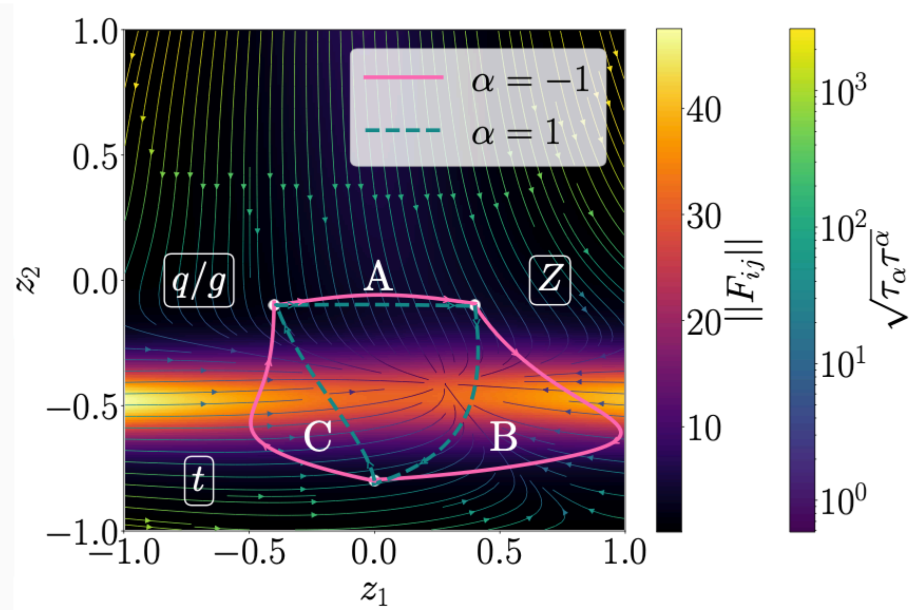
Feature-decision alignment



Connects classifier-relevant directions to physical jet features

How?

Geodesics



Reveals physically interpretable paths between classes