Lipschitz Networks and Energy Movers Distance

Niklas Nolte

15. December 2022

Overview

1. 15-30 seconds of ads
2. The LHCb trigger
3. Lipschitz Networks - Robustness and Monotonicity
4. Energy Movers Distance
5. Differentiable EMD

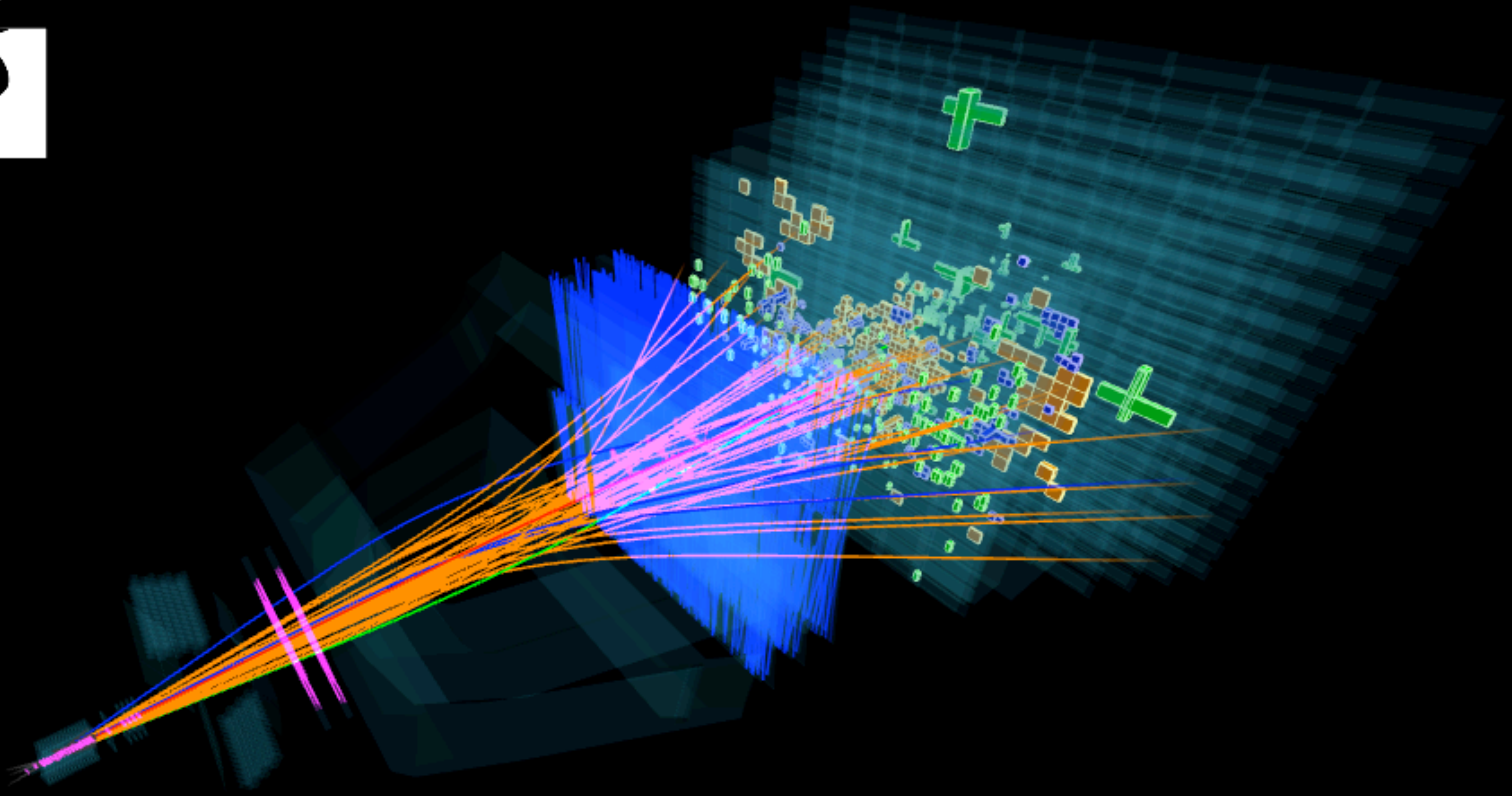

IAIFI.org

LHCb

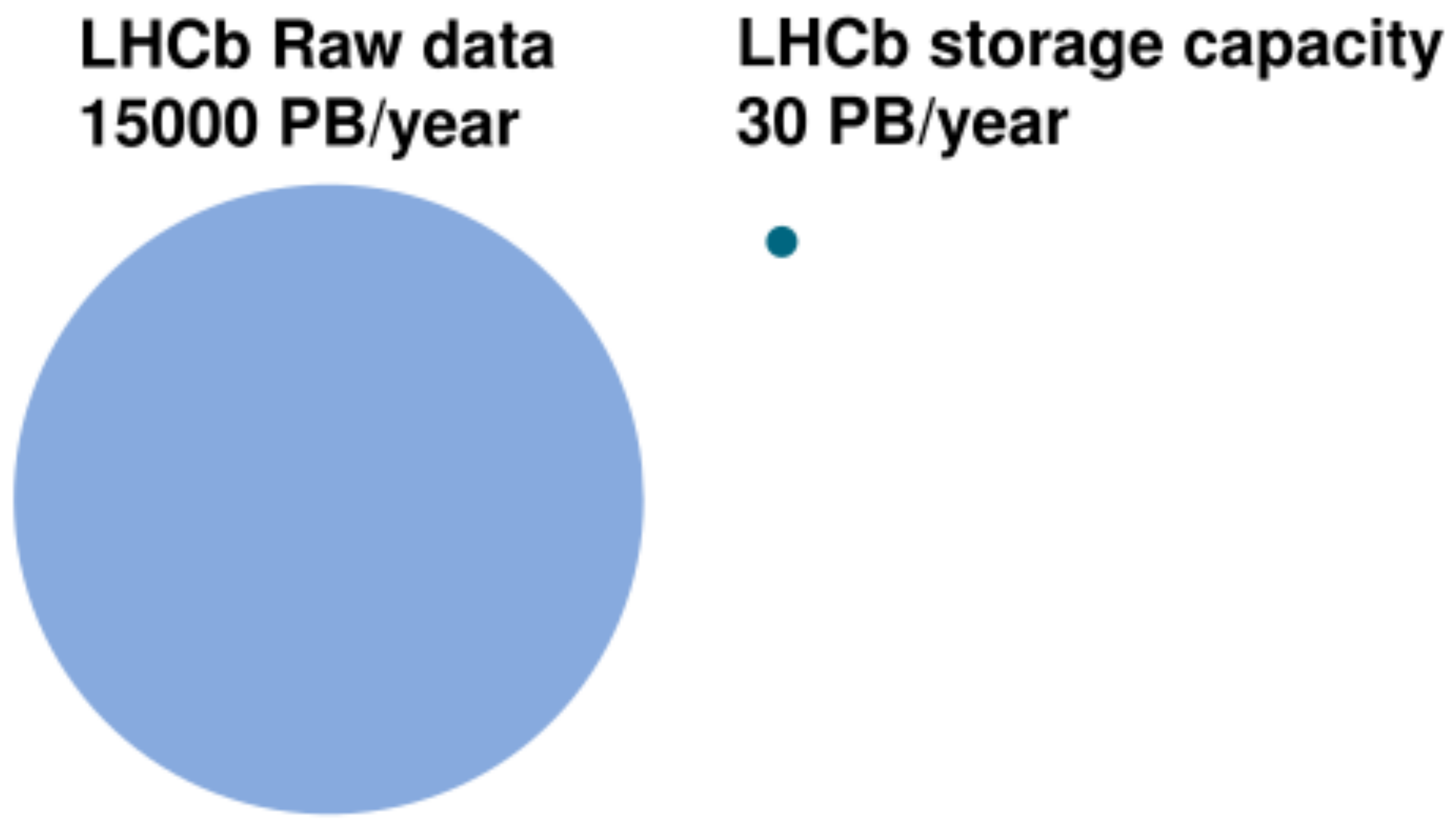
Event 11177639

Run 169235

Mon, 07 Dec 2015 10:38:25

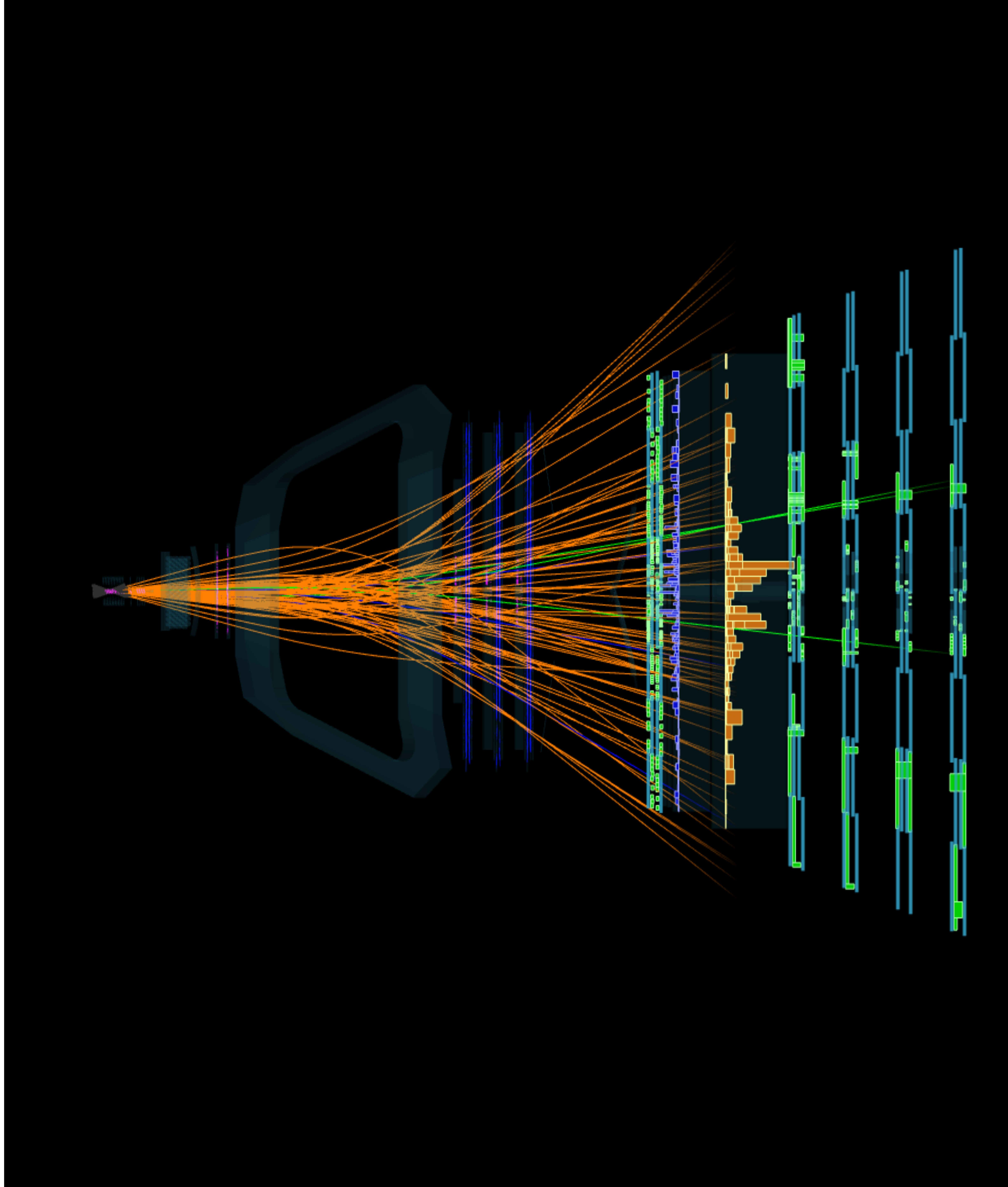


The Trigger at LHCb



Select **only interesting** events:
Hybrid approach of expert systems and ML

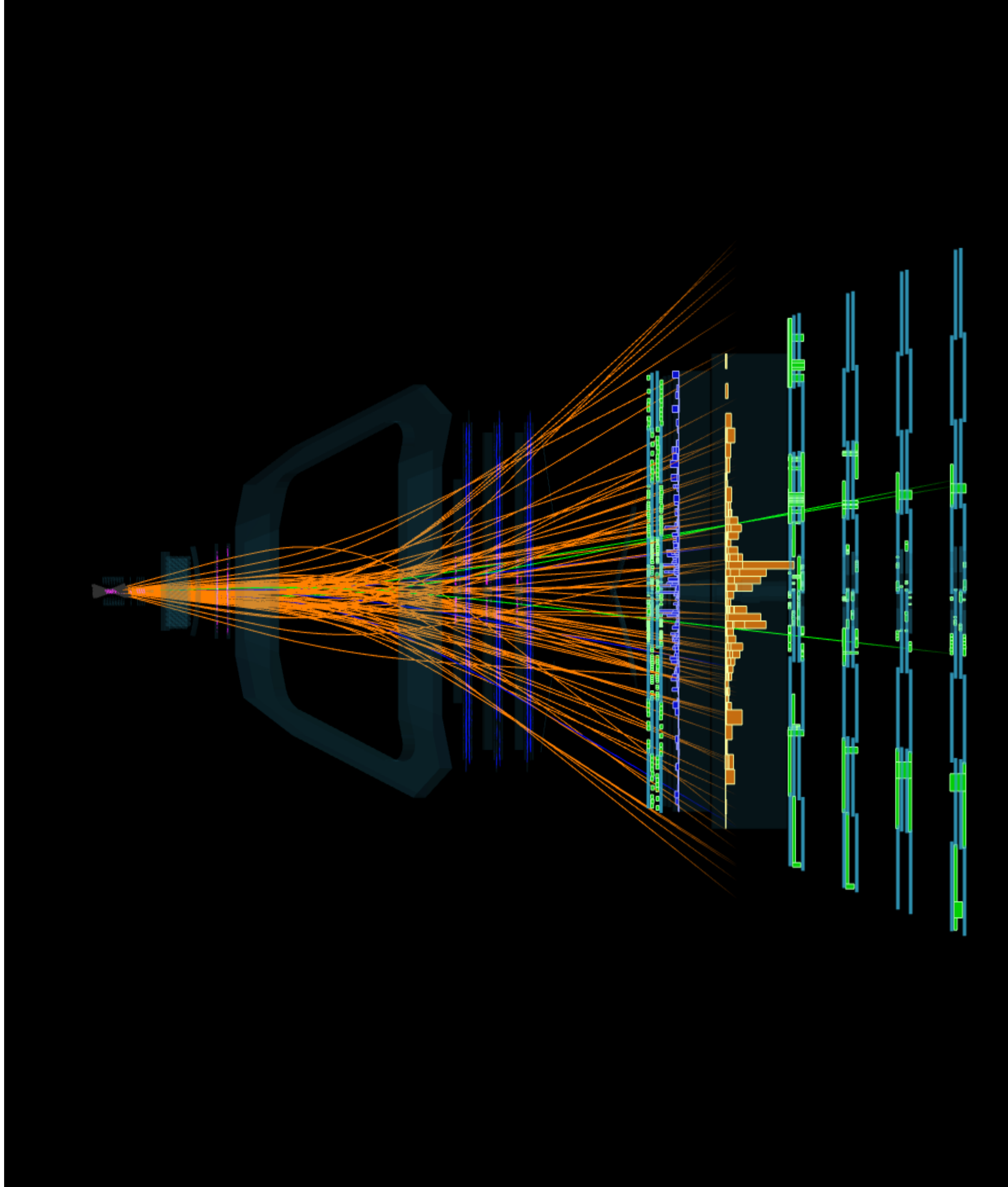
Real time data reduction: 5 TB/s → 10GB/s



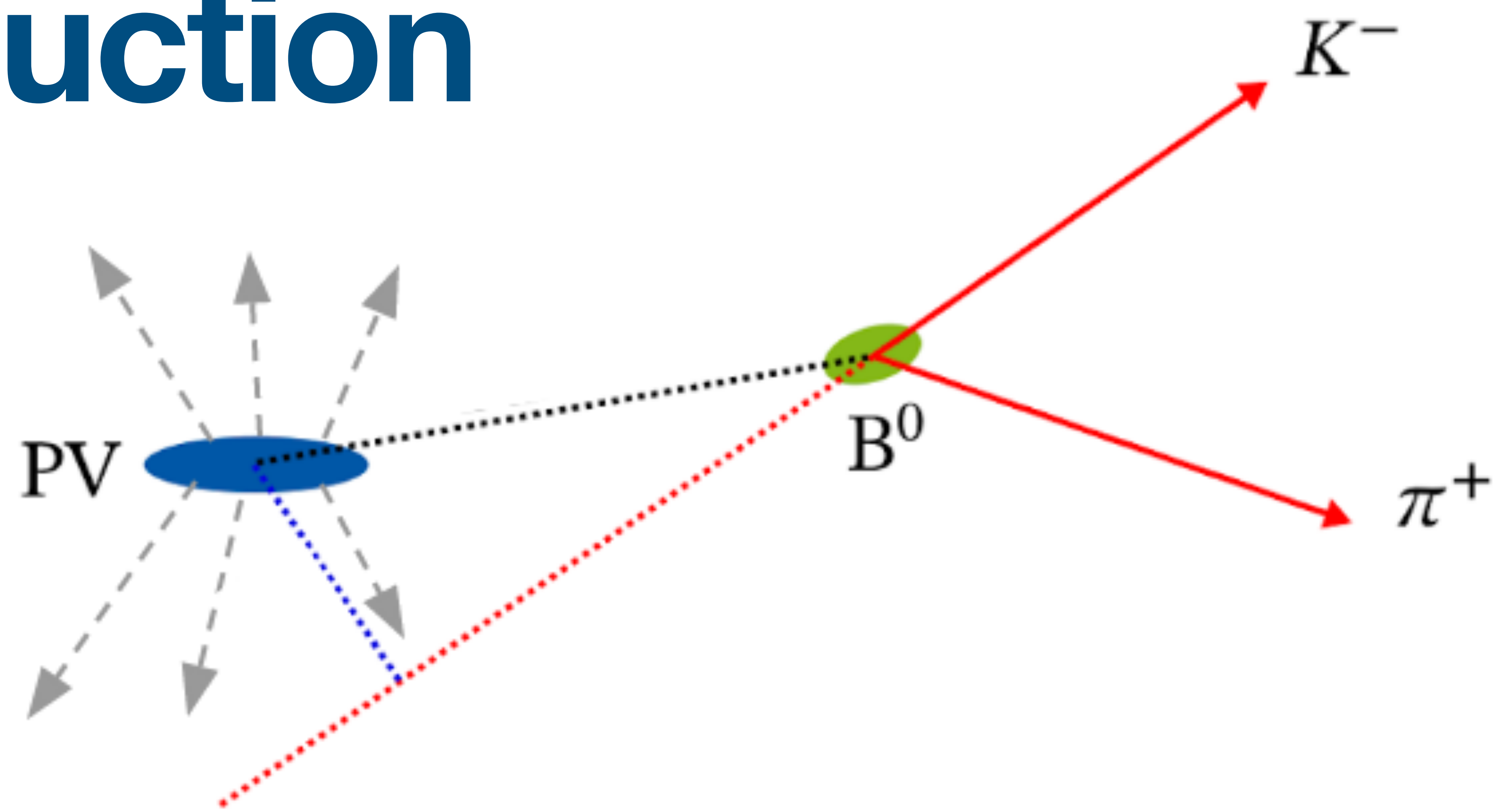
The Trigger at LHCb

Real time expert systems + ML to process 5 TB/s
→ High performance software on GPU and CPU

500x data reduction
→ high purity selection and good event compression



Event Selection -- data reduction



Trigger: mostly an expert system

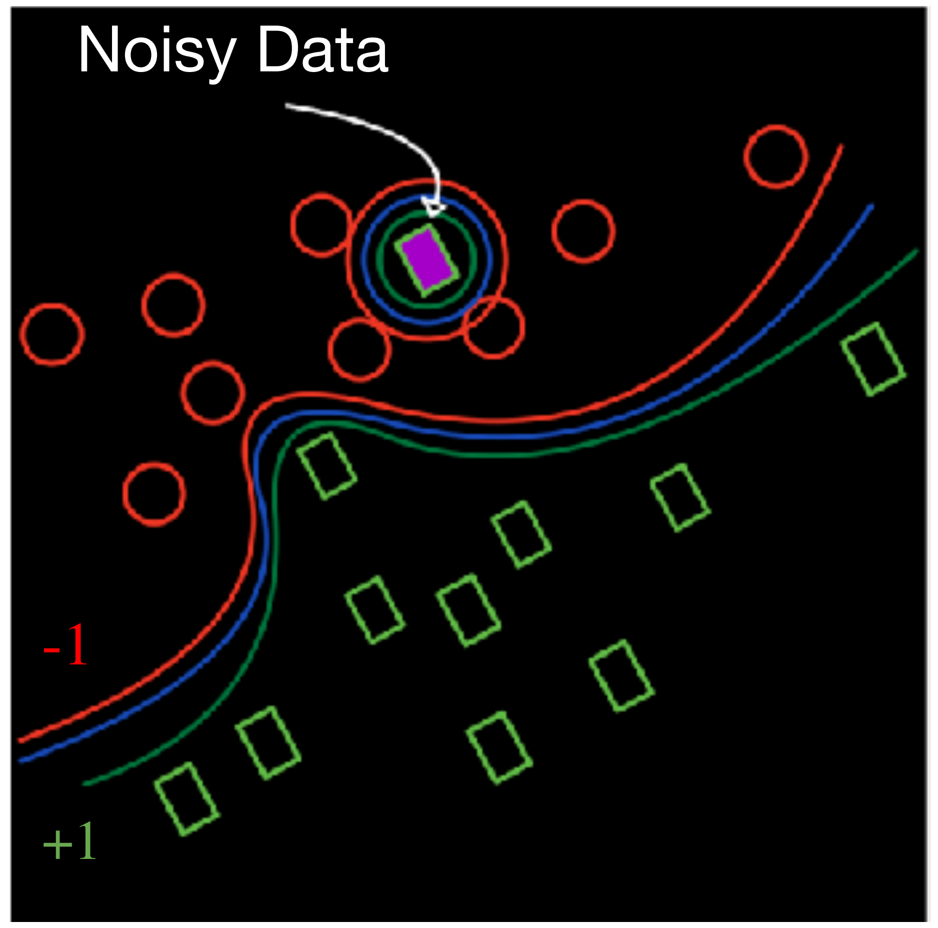
Many subsystems look for particular decays
→ Strong reduction and purity ✅

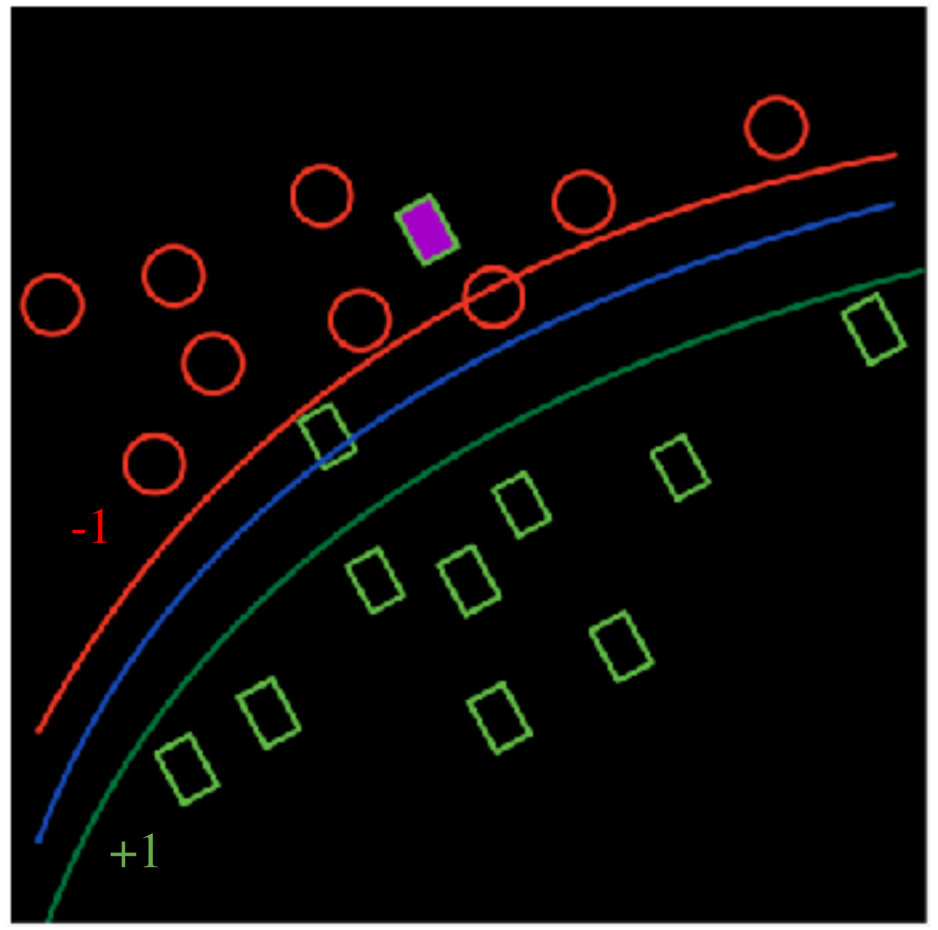
Some look for general signatures, weaker selection
→ Achieve good purity with ML classifiers
Need guarantees to employ these! **No room for error**

Guarantees needed:
1. Robustness w.r.t small changes
2. Monotonicity in certain features for OOD guarantees

(Adversarial) Robustness

many SOTA ML models are proven to be highly unstable

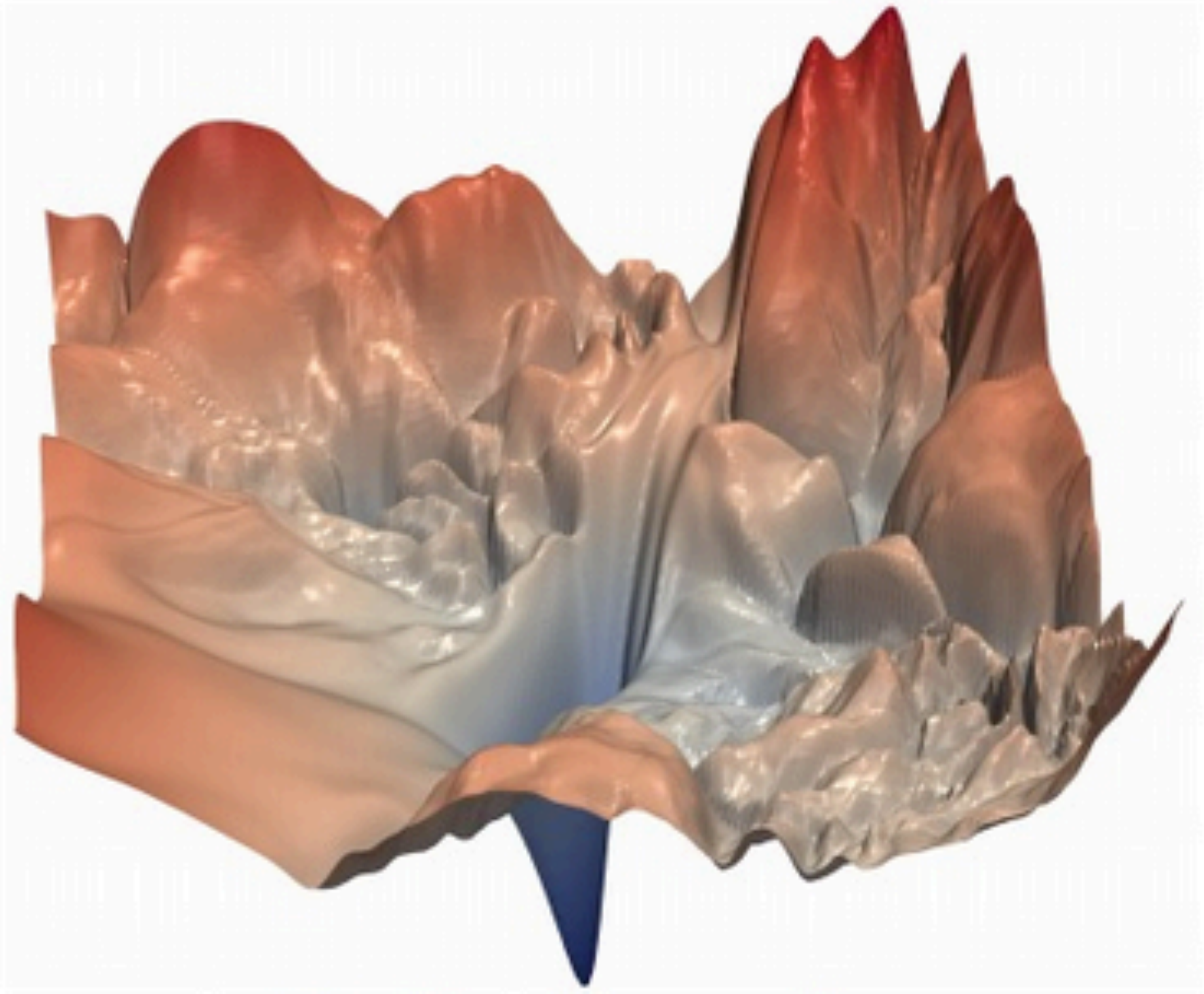


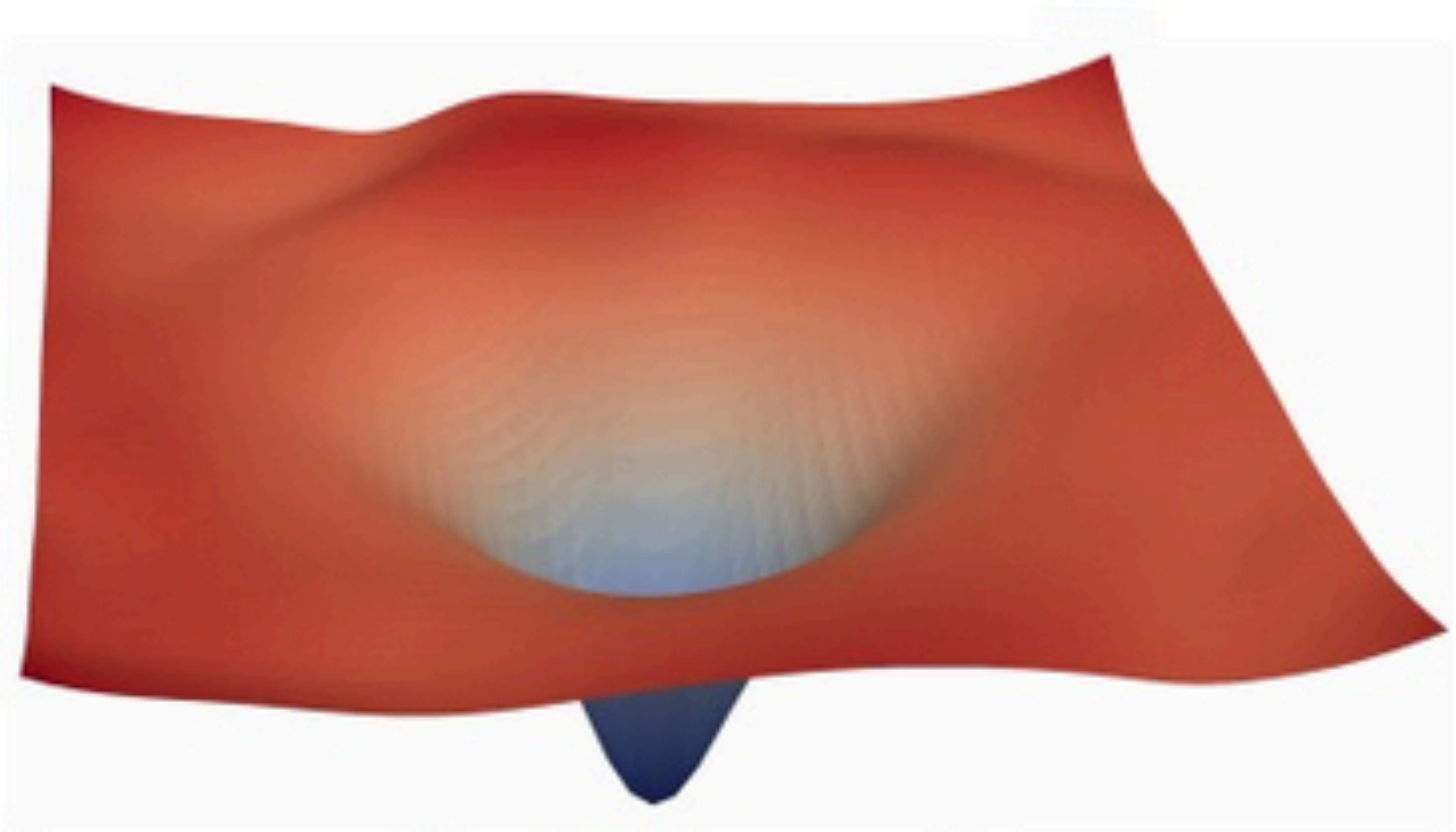


Robustness $:= \quad F(x+\epsilon) = F(x) + O(\epsilon)$

I am looking for deterministic robustness, i.e. provably robust networks!

Two Decision Landscapes





Deterministic Robustness - How?

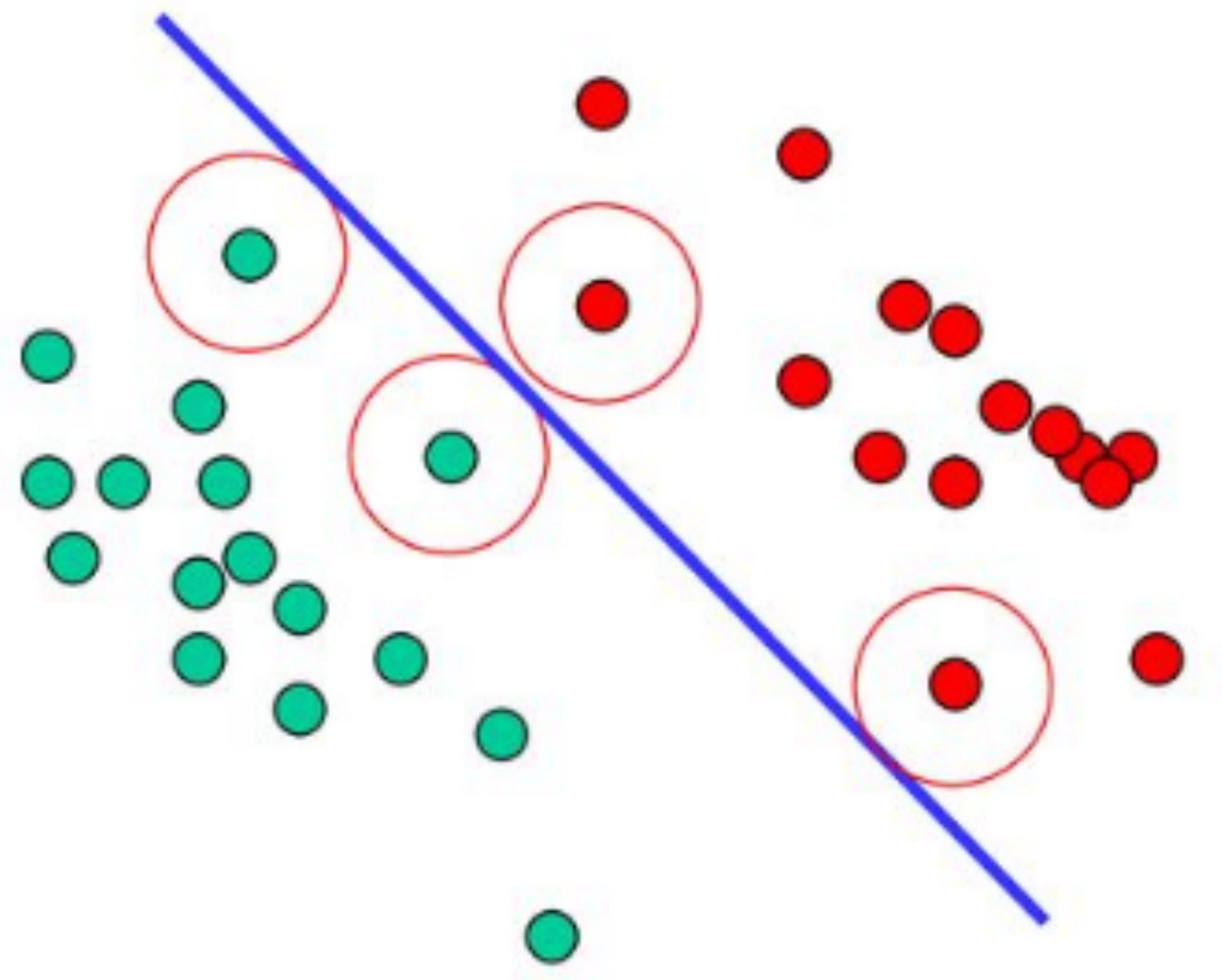
WLOG: Binary classifier: $F : \mathbb{R}^n \to \mathbb{R}$

Constrain gradient wrt inputs, i.e. make it Lipschitz-L

$||\nabla F|| \leq L$

A perturbation ϵ to an input x needs certain magnitude to flip the sign:

$$\operatorname{sign} F(x+\epsilon) = -\operatorname{sign} F(x) \;\Rightarrow\; ||\epsilon|| > \frac{|F(x)|}{L}$$



Lipschitz Networks

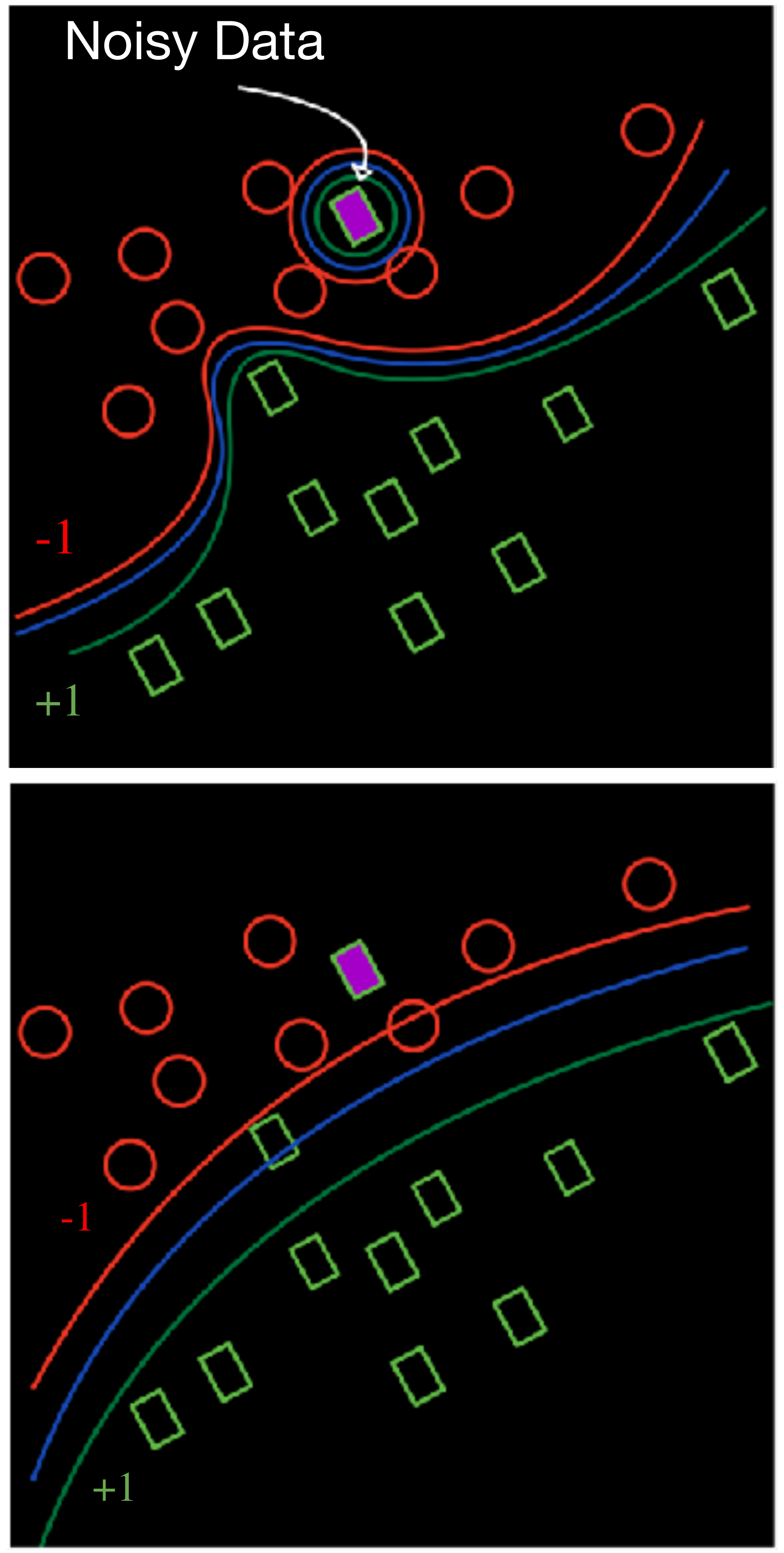
WLOG: $F(x) = W^{(2)} \cdot \sigma(W^{(1)} \cdot x + b^{(1)}) + b^{(2)}$

$||\nabla F|| \leq L$ can be enforced by constraining weights

In an MLP with Lipschitz-1 activations: $L \leq \prod ||W^{(i)}||$
(Toeplitz matrix for CNNs)

Maintain a maximum operator norm $||W^{(i)}||$ in every layer

Lipschitz-L guaranteed!



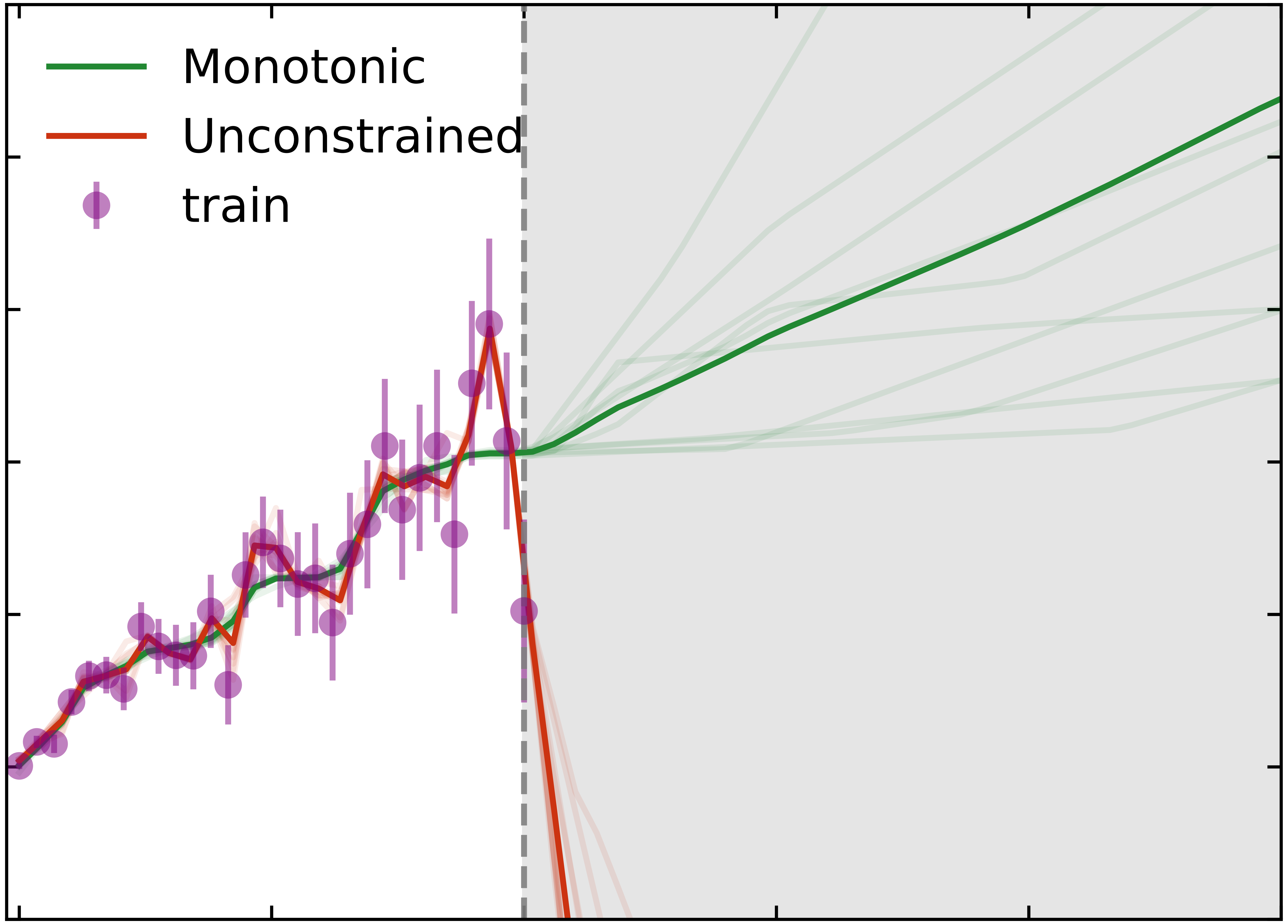
Monotonic Networks

We care about tails!

Guarantees about OOD with monotonicity

Expressive monotonic networks are not obvious

Existing algorithms involve monotonic regularization or non-expressive architectures



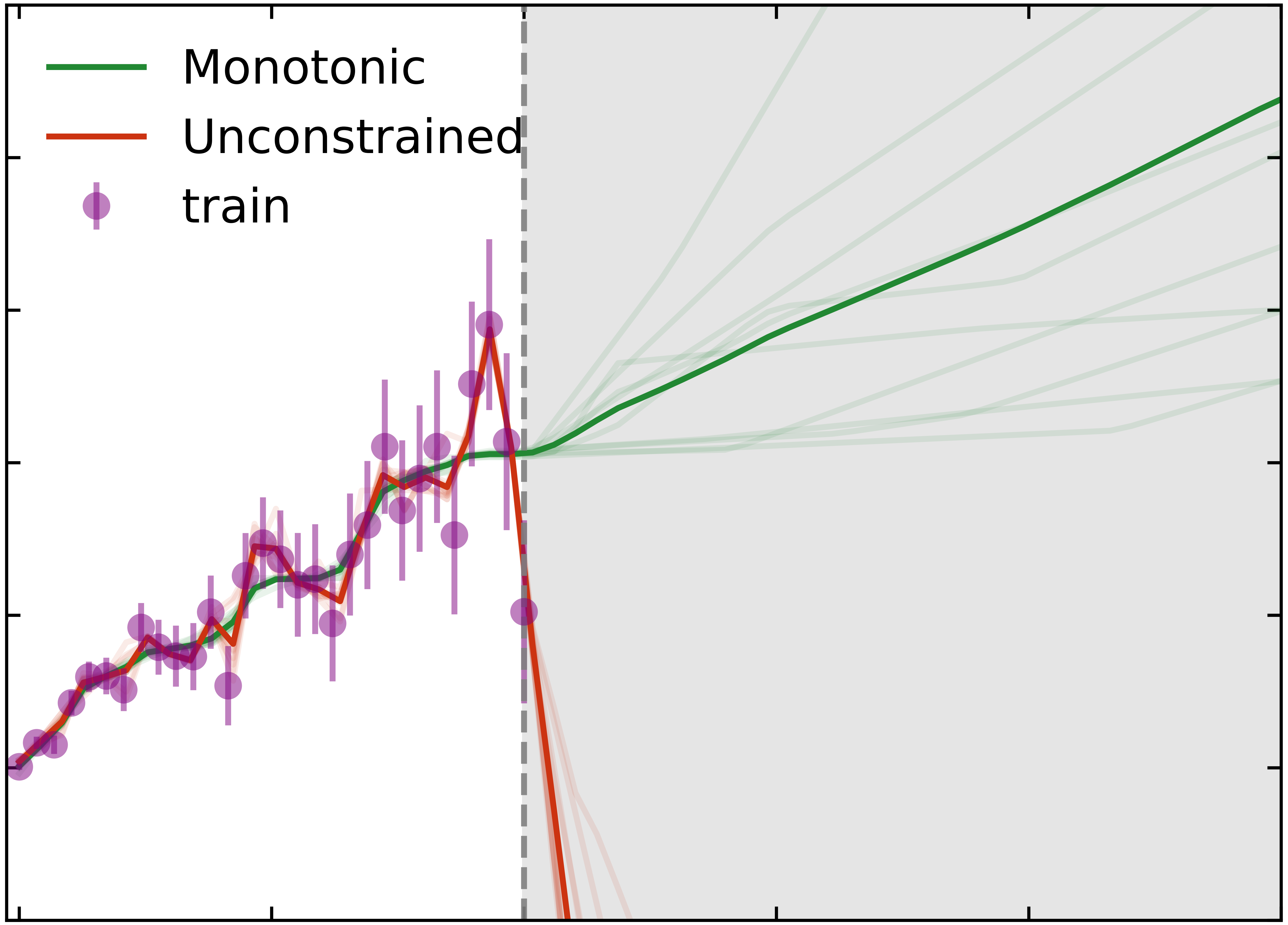
Monotonic Lipschitz Networks

Combine Lipschitz networks with monotonicity!

$$M(x) = F(x) + L \sum_i x_i$$

$$\frac{\partial M}{\partial x_i} = \frac{\partial F}{\partial x_i} + L \geq 0$$

Lipschitz-L : $\|\nabla F\| \leq L$



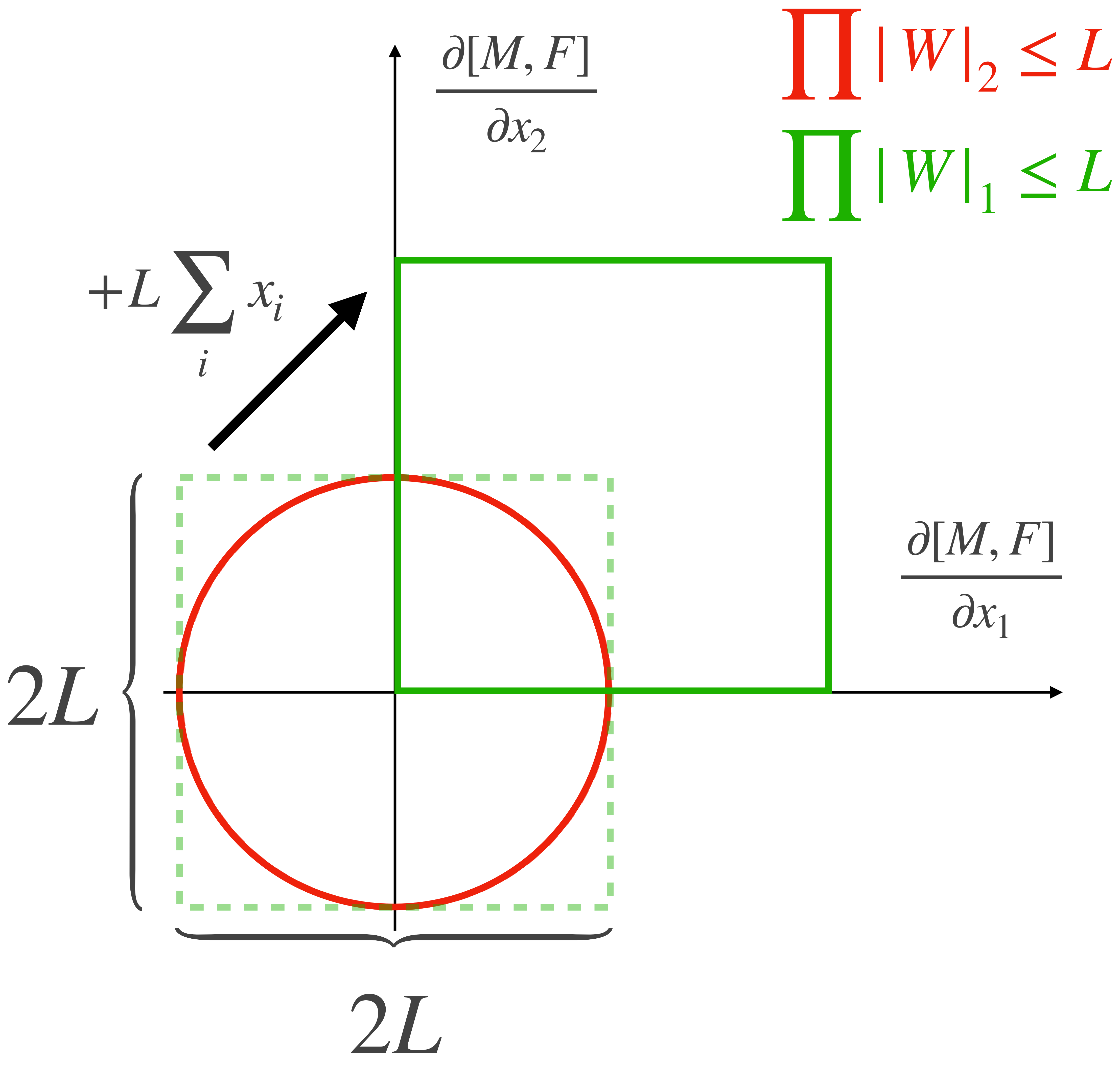
Monotonic Lipschitz Networks

$$M(x) = F(x) + L\sum_i x_i$$

$$\frac{\partial M}{\partial x_i} = \frac{\partial F}{\partial x_i} + L$$

$+L$ contribution in every direction x_i
$\|\nabla F\| \leq L$ is not good enough

We want $\|\nabla F\|_\infty \leq L$!



Activations

Universal Lipschitz Approximation

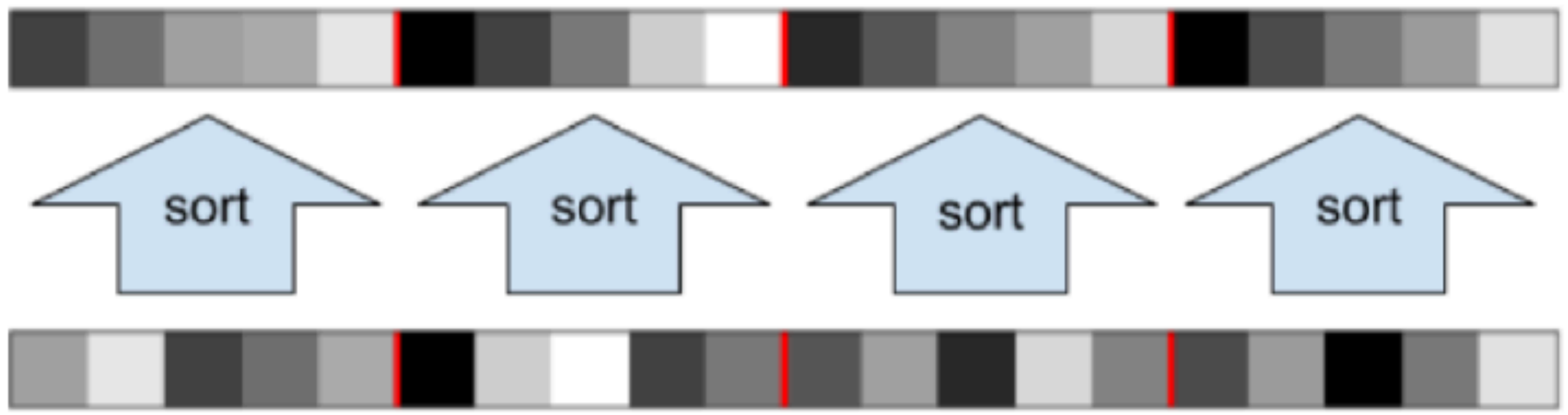
$$F(x) = W^{(2)} \cdot \sigma(W^{(1)} \cdot x + b^{(1)}) + b^{(2)}$$

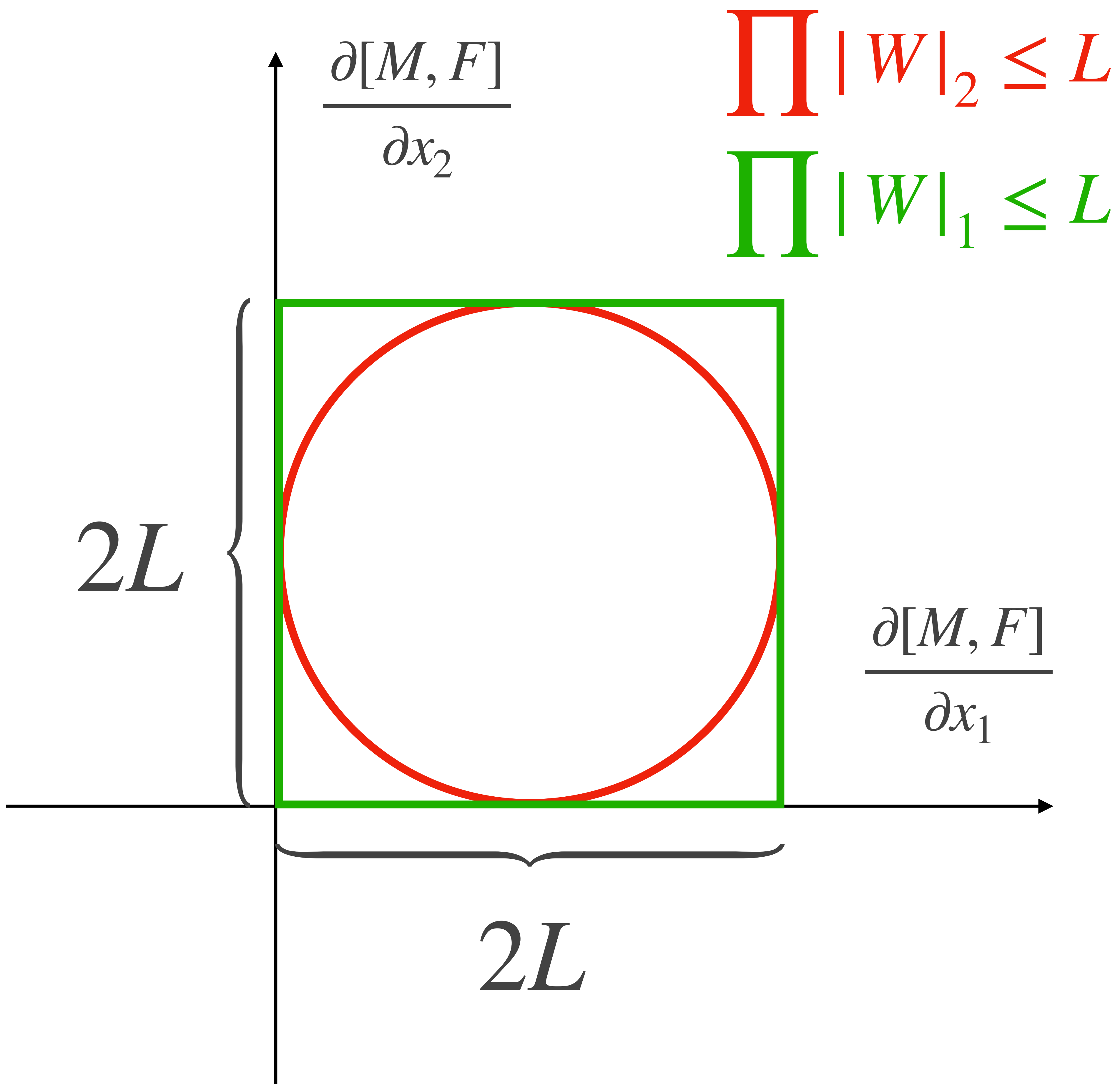
$$||W^{(i)}|| \leq 1 \leftarrow \text{over-constraining?}$$

Activations need $||\nabla \sigma|| = 1$

Pointwise activations are useless!

Solution:



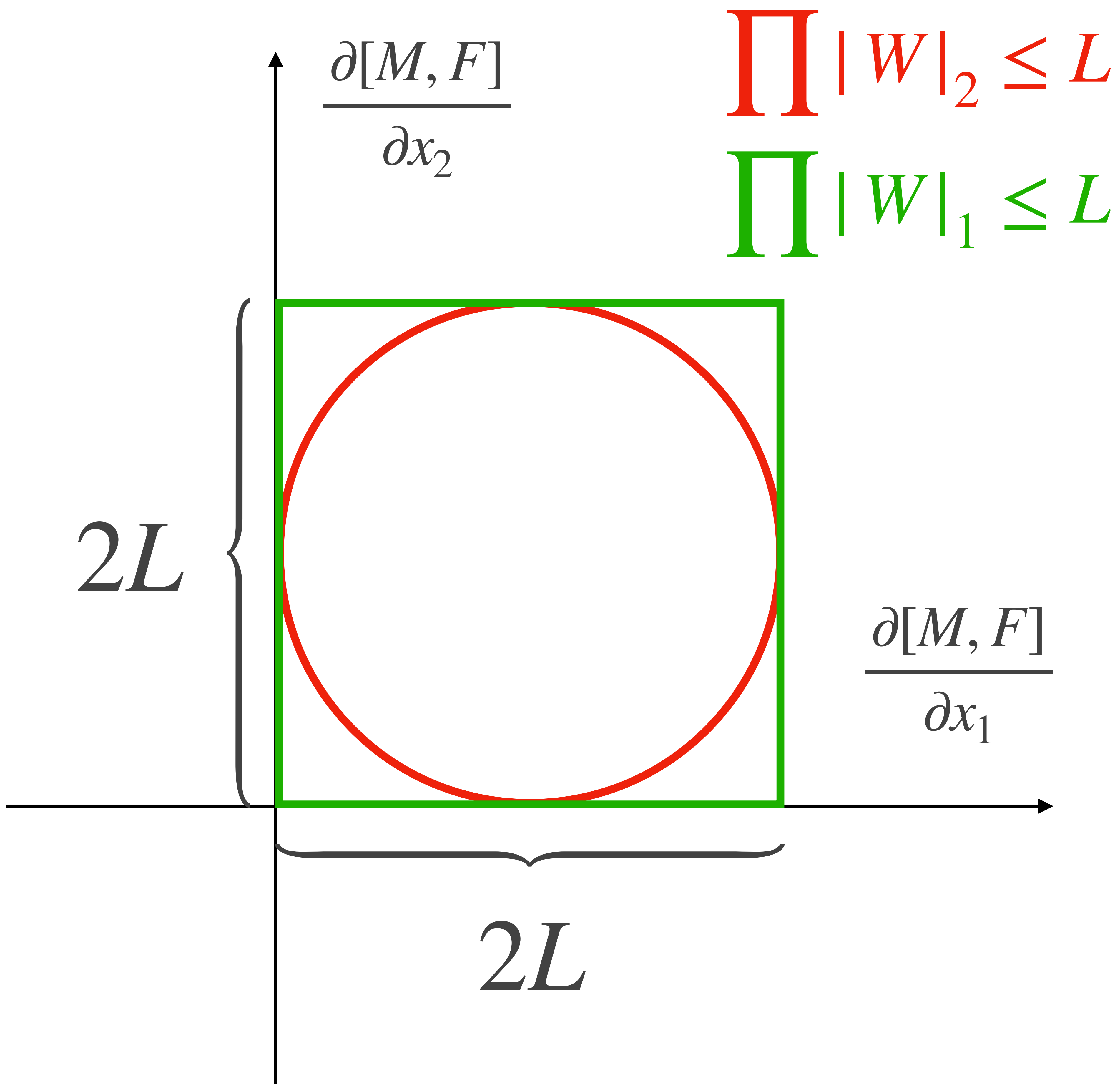


Summary - Monotonic Lipschitz Functions

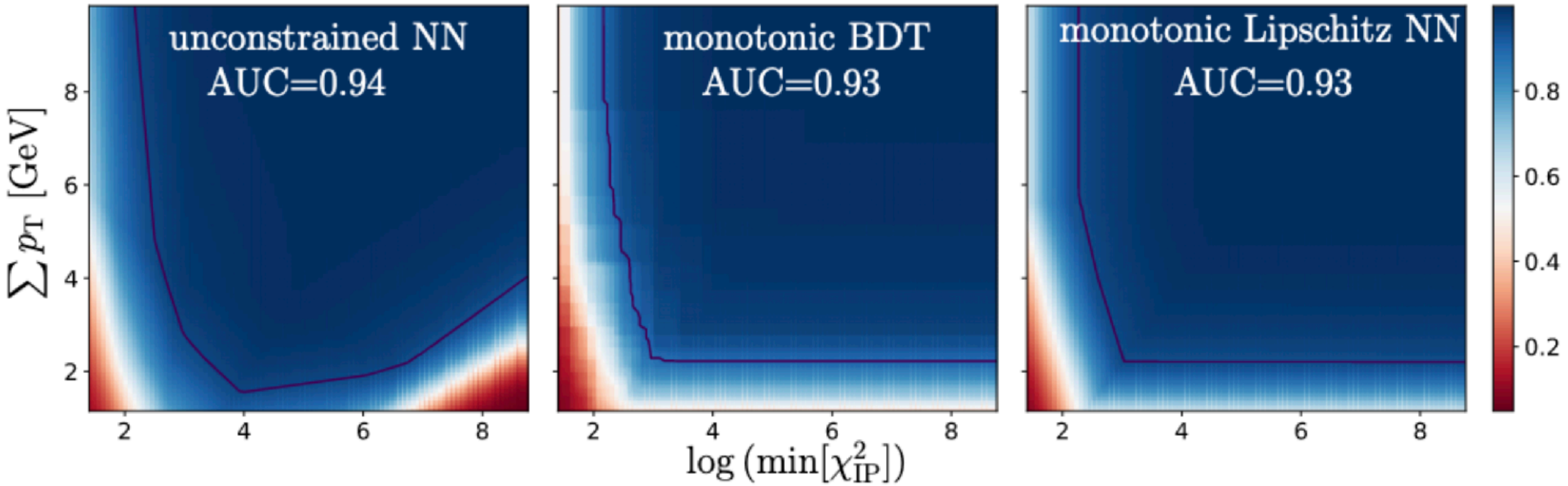
This architecture is

1. provably robust
2. provably monotonic
3. universally approximating
   the target function class
4. working well in practice

→ Implemented in the LHCb trigger for many major selections



HLT1 Inclusive b&c lines -- 2D subproblem



Energy Flow and Energy Movers Distance

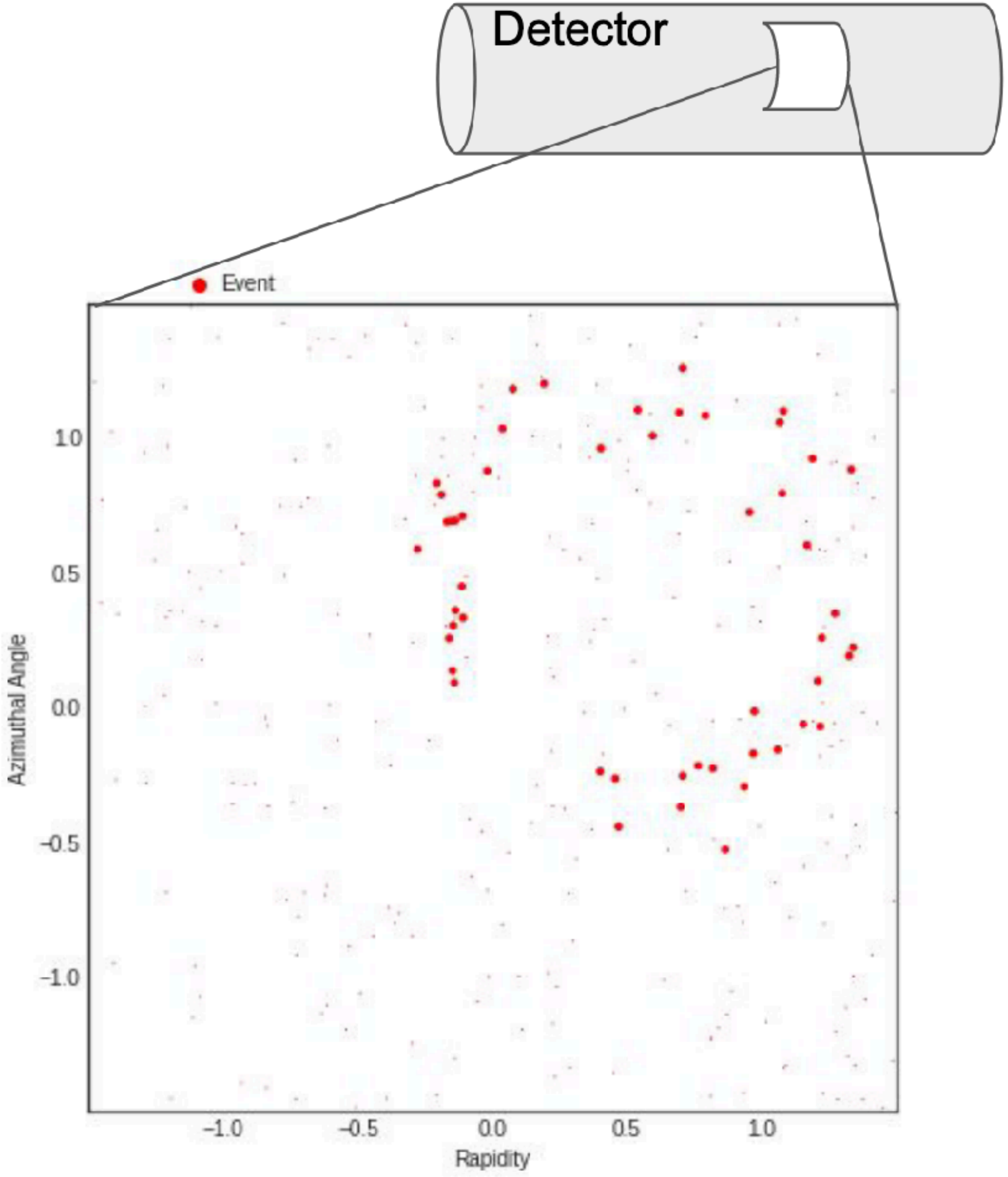
Energy Flow

$$\mathscr{E}(x) = \sum_i E_i \, \delta(x - x_i)$$

2D projection on (y, ϕ)

$$\left.\begin{array}{l}\text{Infrared robust}\\ \text{Collinear robust}\end{array}\right\}\ \text{IRC}$$

IRC robust information is contained in $\mathscr{E}$

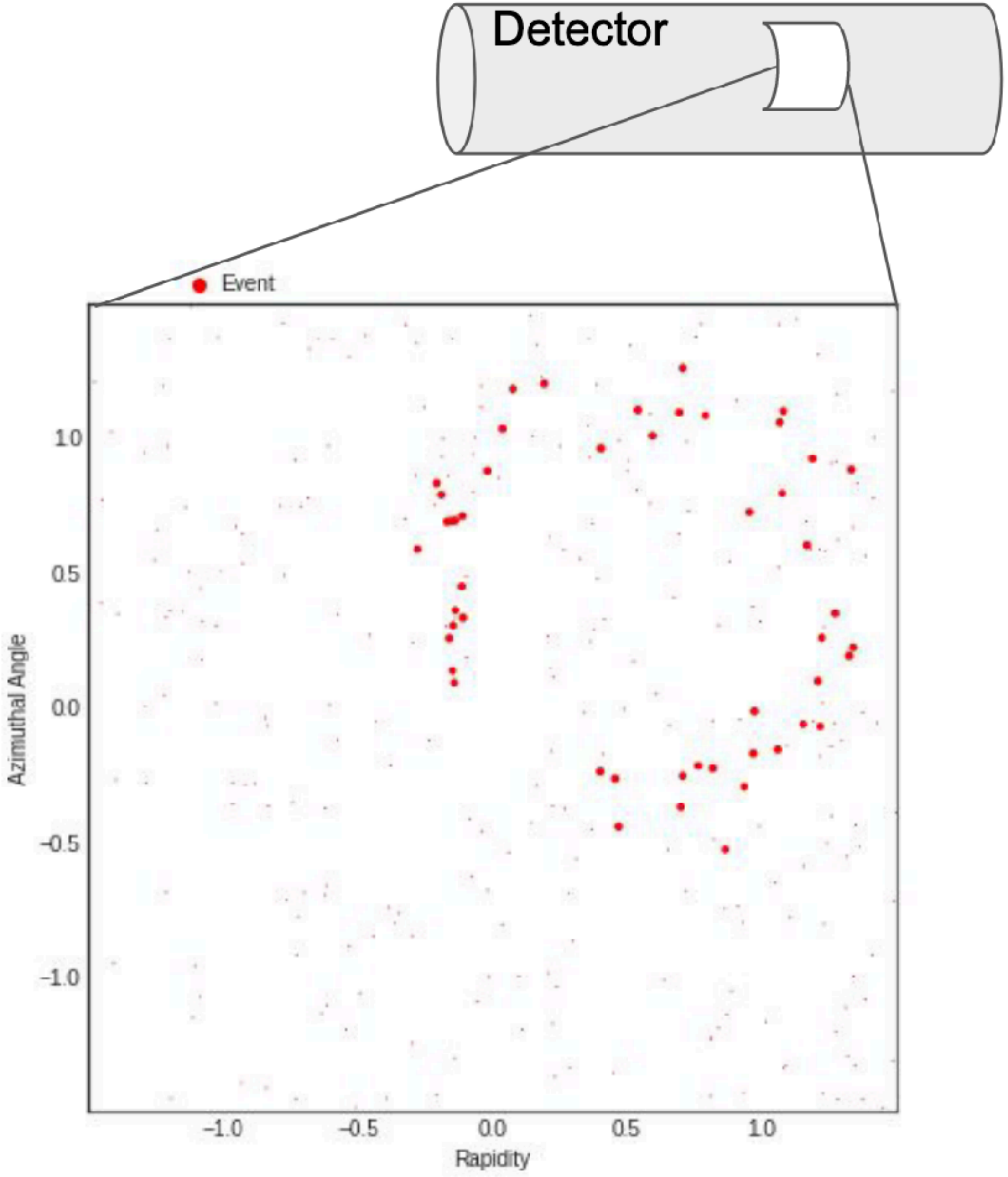


IRC invariance

Infrared:
Zero energy radiation

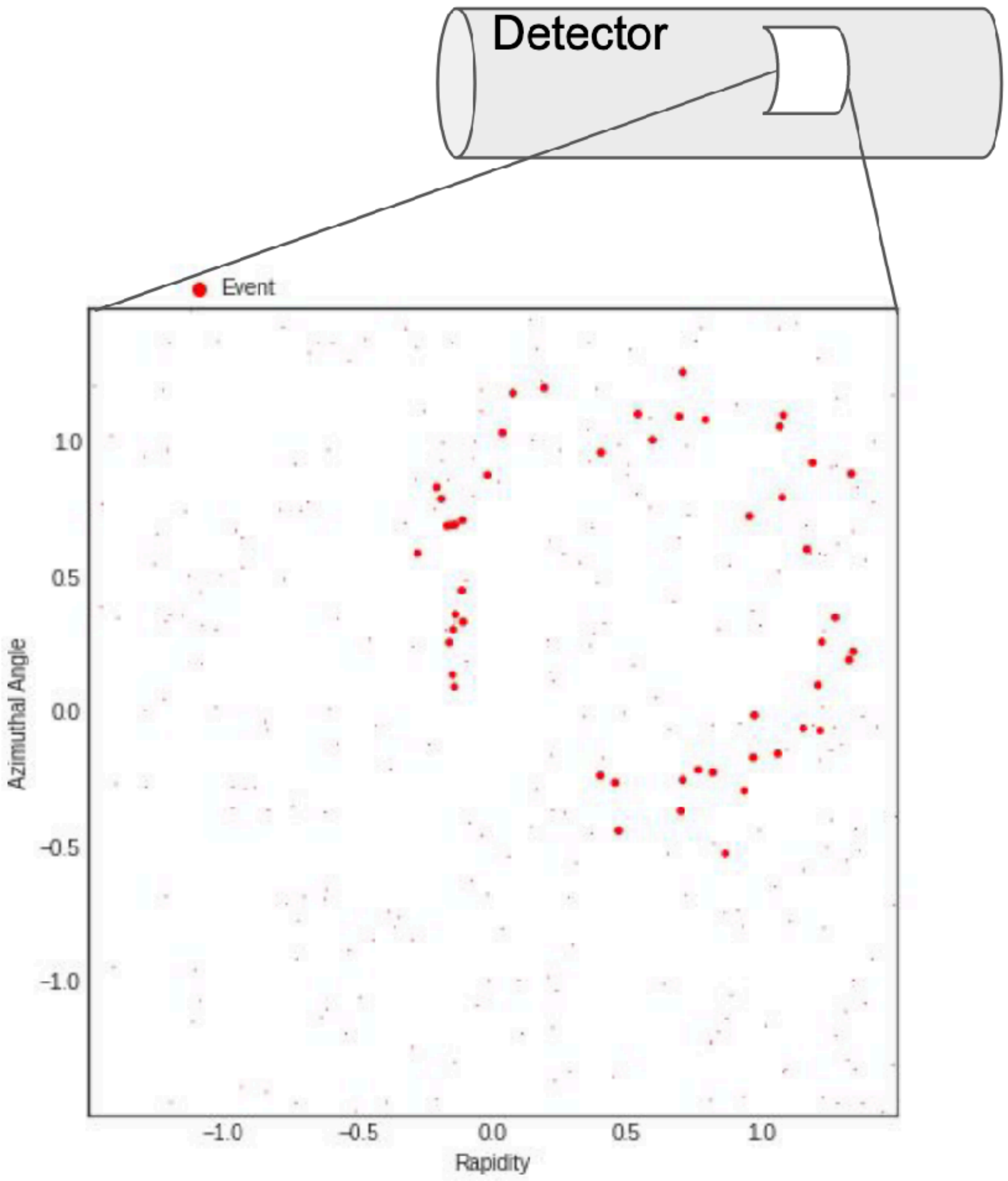
Collinear:
Split one particle into two identical ones with $\angle\ 0$

IRC invariant observable
→ perturbative description possible



Energy Flow

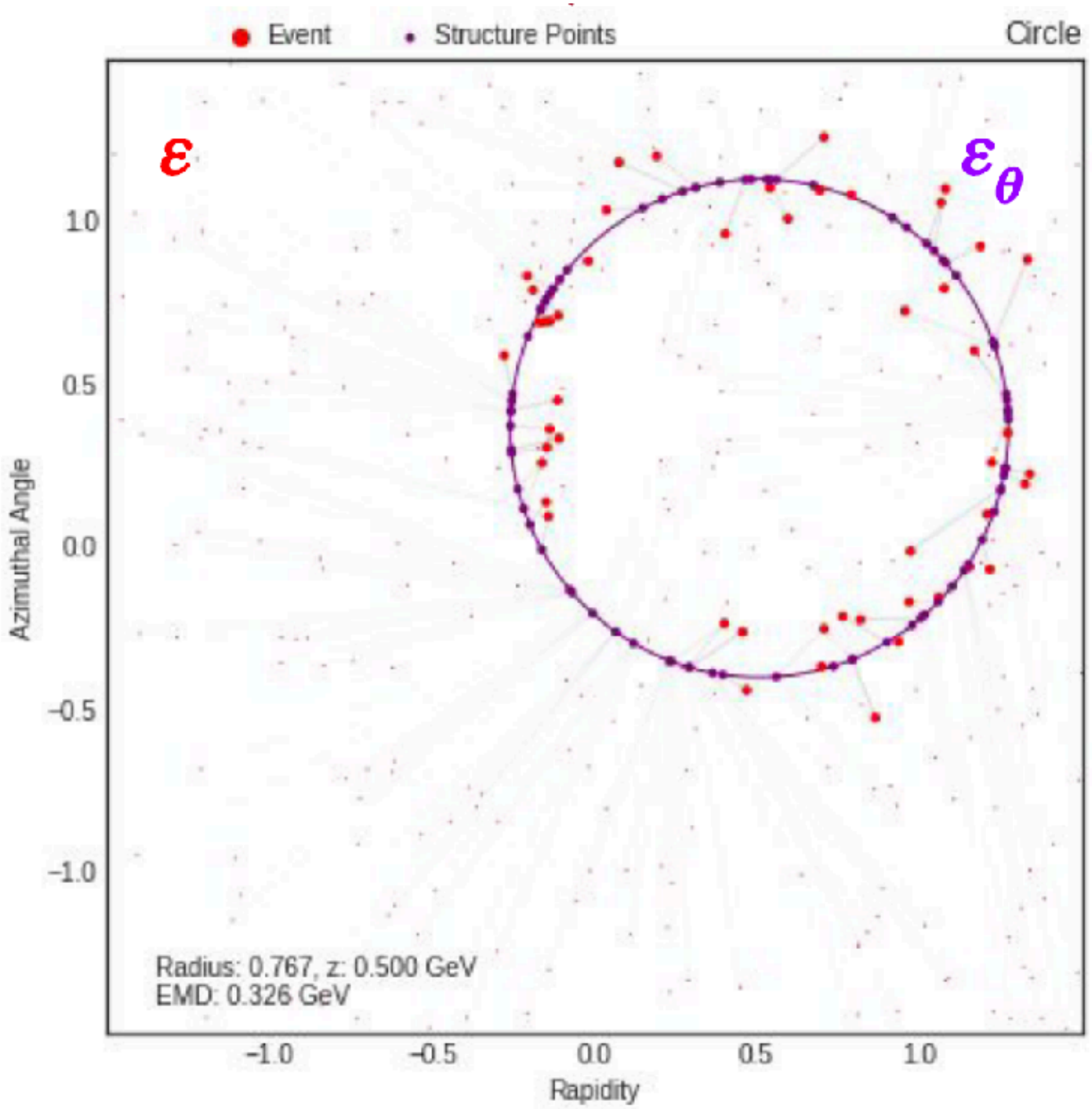
Question: What shape is that?



A circle!

Question: What shape is that?

Answer: It is a circle!



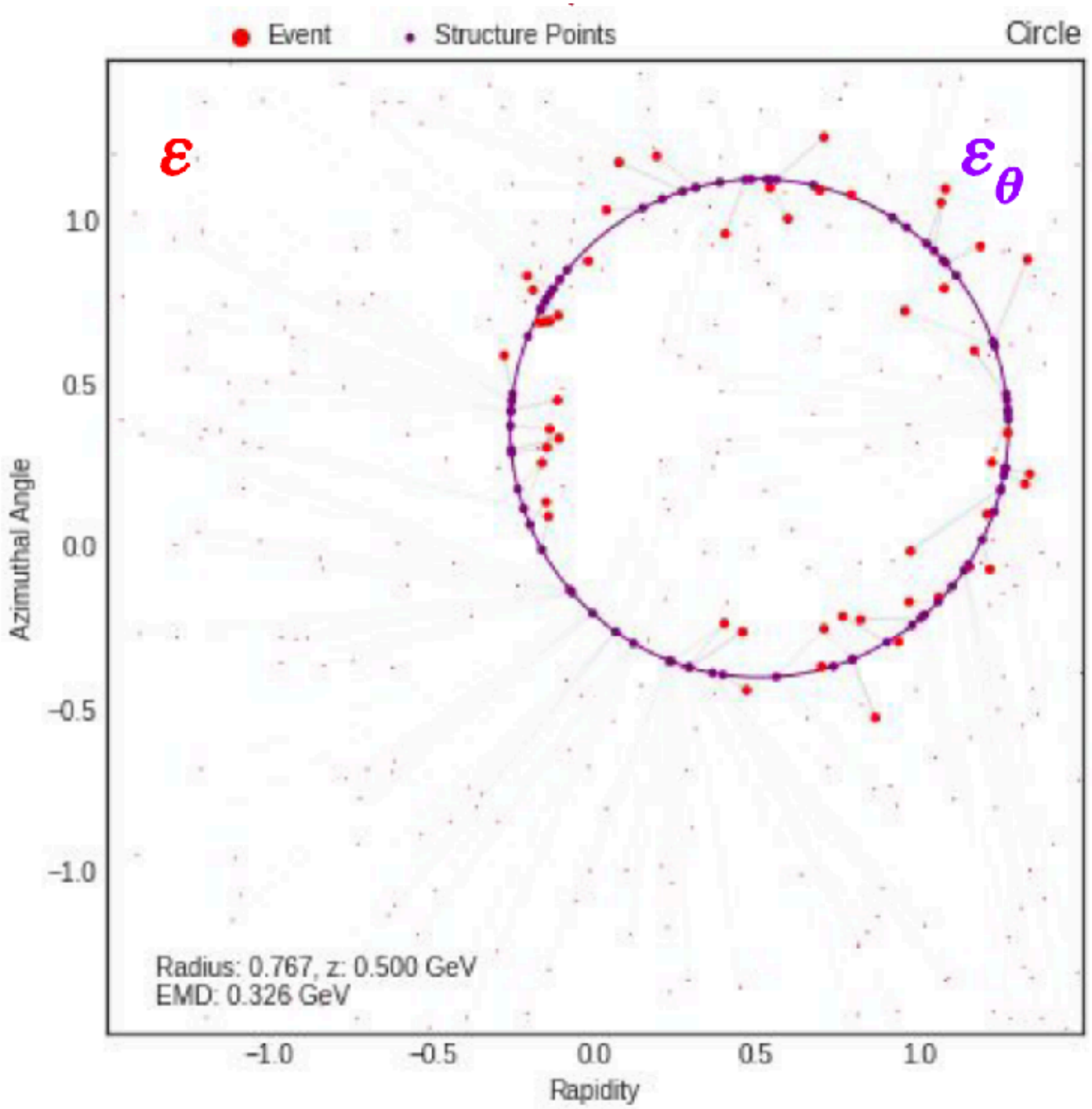
A circle!

Why do we care?

Many classic observables are defined by "similarity to some shape"

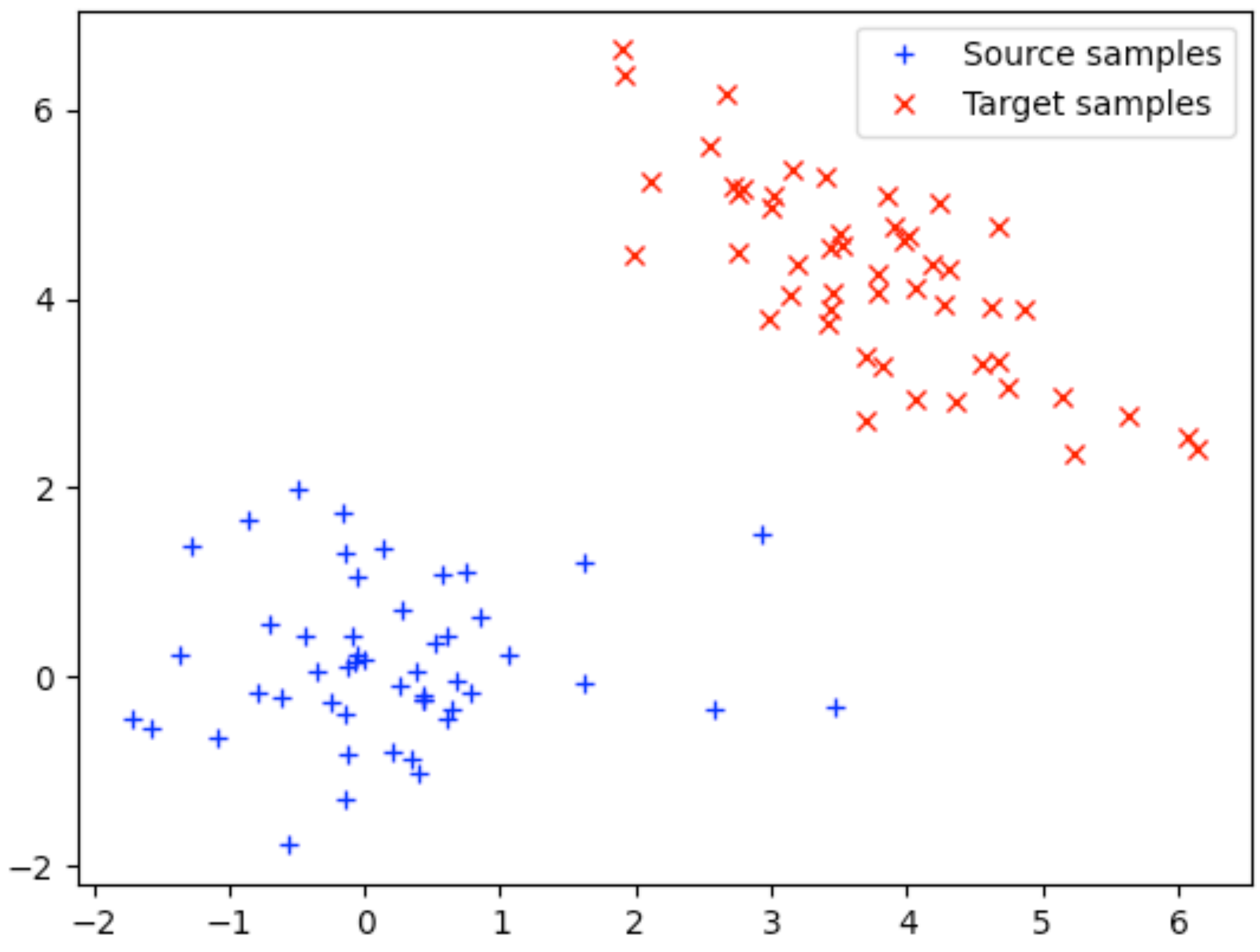
Many new ones will also be

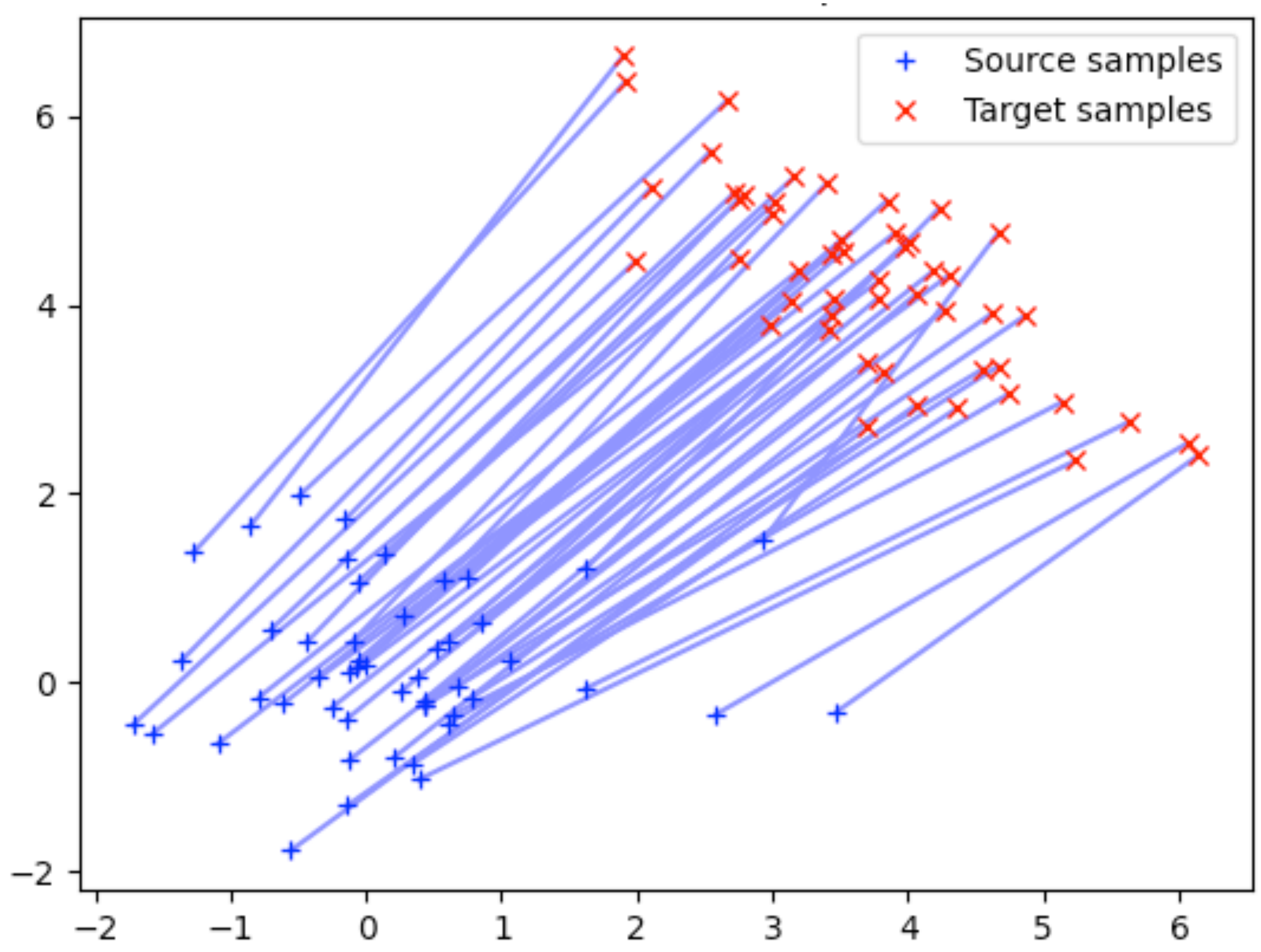
https://doi.org/10.1007/JHEP07(2020)006



Optimal Transport - Earth Movers Distance

Optimal Transport - Earth Movers Distance





Optimal Transport - Earth Movers Distance

$$\mathrm{EMD}(\mathbb{P}, \mathbb{Q}) = \inf_{\gamma \in \Pi(\mathbb{P}, \mathbb{Q})} \mathbb{E}_{(x,y)\sim\gamma}\left[\|x - y\|\right]$$

Distribution 1

Distribution 2

Joint Distribution

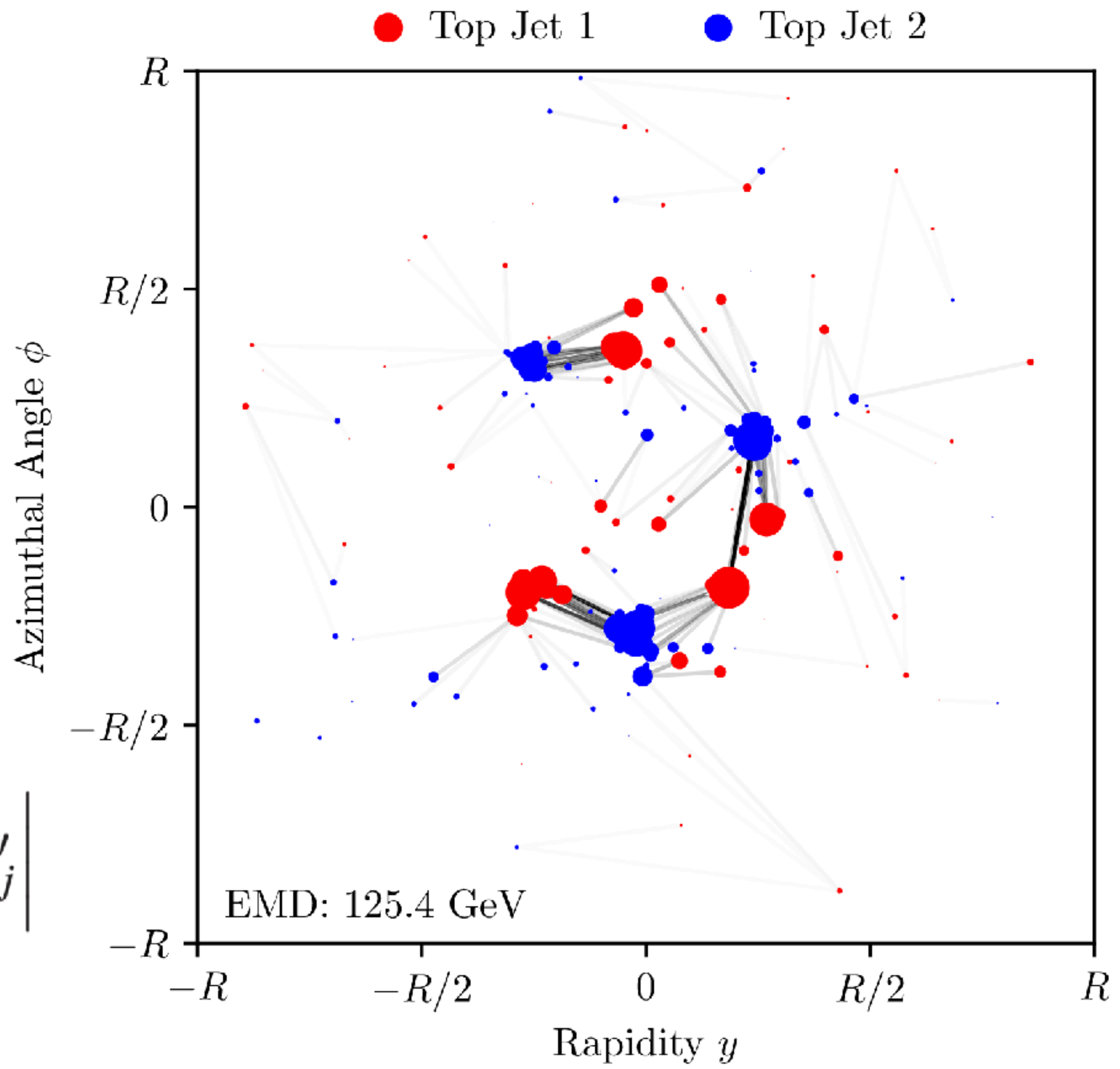
Coordinates

Energy Movers Distance

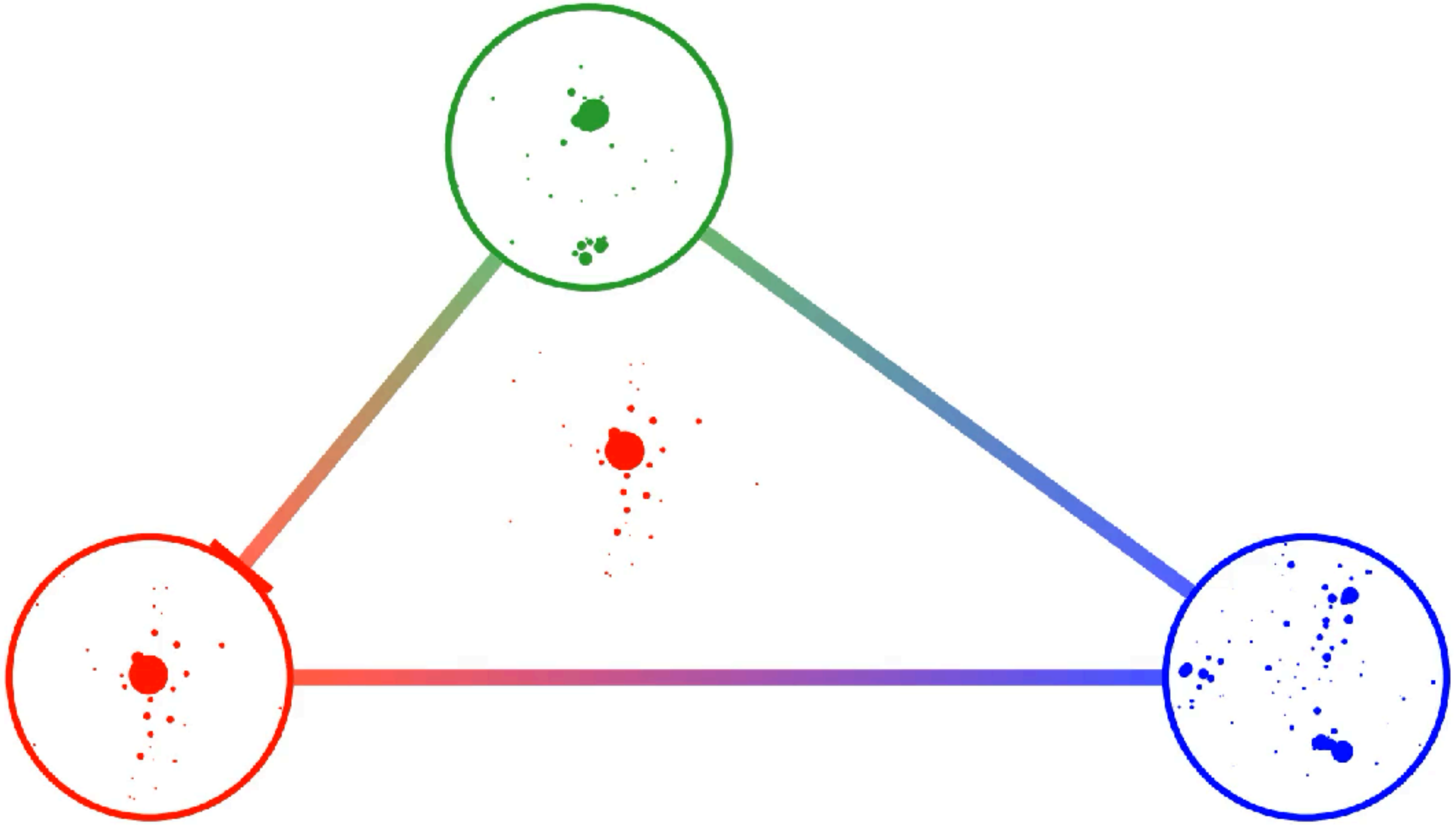
Energy Movers Distance

A proper metric!

$$\mathrm{EMD}(\mathcal{E},\mathcal{E}') = \min_{\{f_{ij}\geq 0\}} \sum_{ij} f_{ij}\frac{\theta_{ij}}{R} + \left|\sum_i E_i - \sum_j E'_j\right|$$



Energy Movers Distance



From https://energyflow.network/docs/emd/

Energy Movers Distance

$$\text{EMD}(\mathcal{E},\mathcal{E}') = \min_{\{f_{ij}\geq 0\}} \sum_{ij} f_{ij}\frac{\theta_{ij}}{R} + \left|\sum_i E_i - \sum_j E'_j\right|$$

$$\text{EMD}(\mathbb{P},\mathbb{Q}) = \inf_{\gamma\in\Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(x,y)\sim\gamma}\big[\|x-y\|\big]$$

Event 1

Event 2

Energies

Particle Coordinates

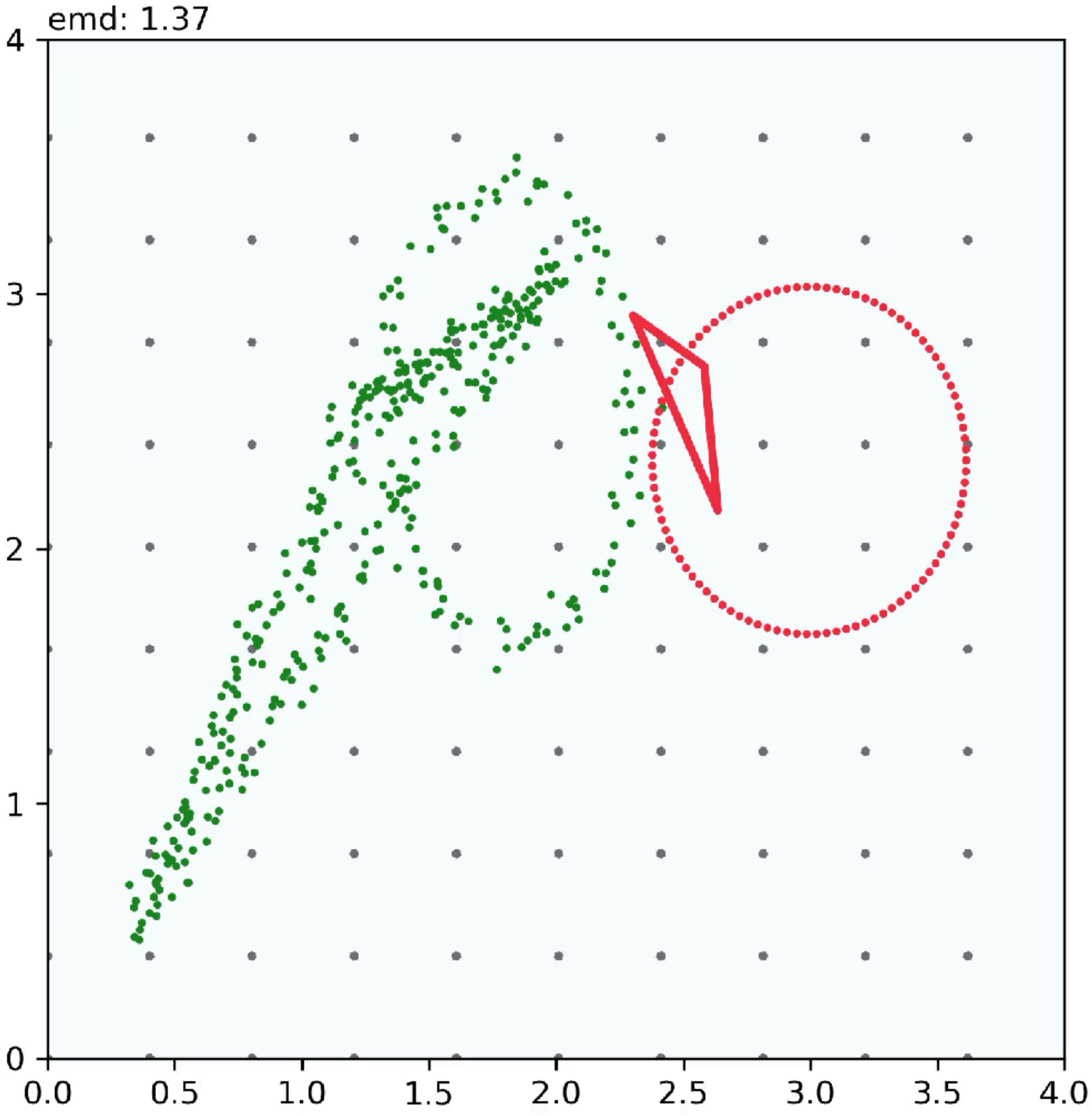
Kantorovich-Rubenstein Dual Formulation

$$\mathrm{EMD}(\mathbb{P}, \mathbb{Q}) = \sup_{\|\nabla F\| \leq 1} \mathbb{E}_{x \sim \mathbb{P}}\left[F(x)\right] - \mathbb{E}_{y \sim \mathbb{Q}}\left[F(y)\right]$$

Kantorovich potential

What shape is that?

Answer via geometric fitting!



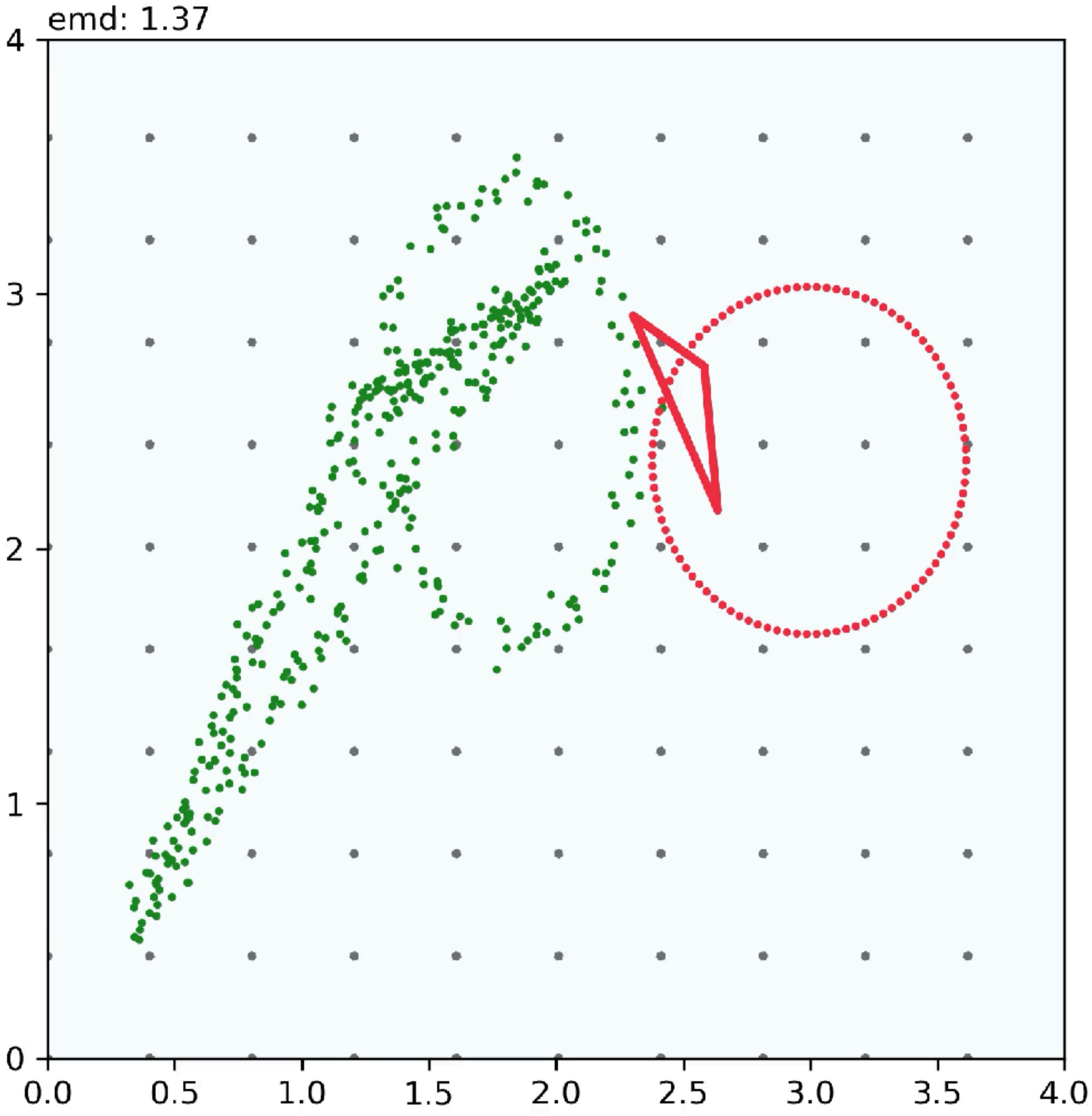
Joint EMD Optimization

Joint EMD Optimization

Use cases:

Unify many useful LHC observables

New observables at LHC?



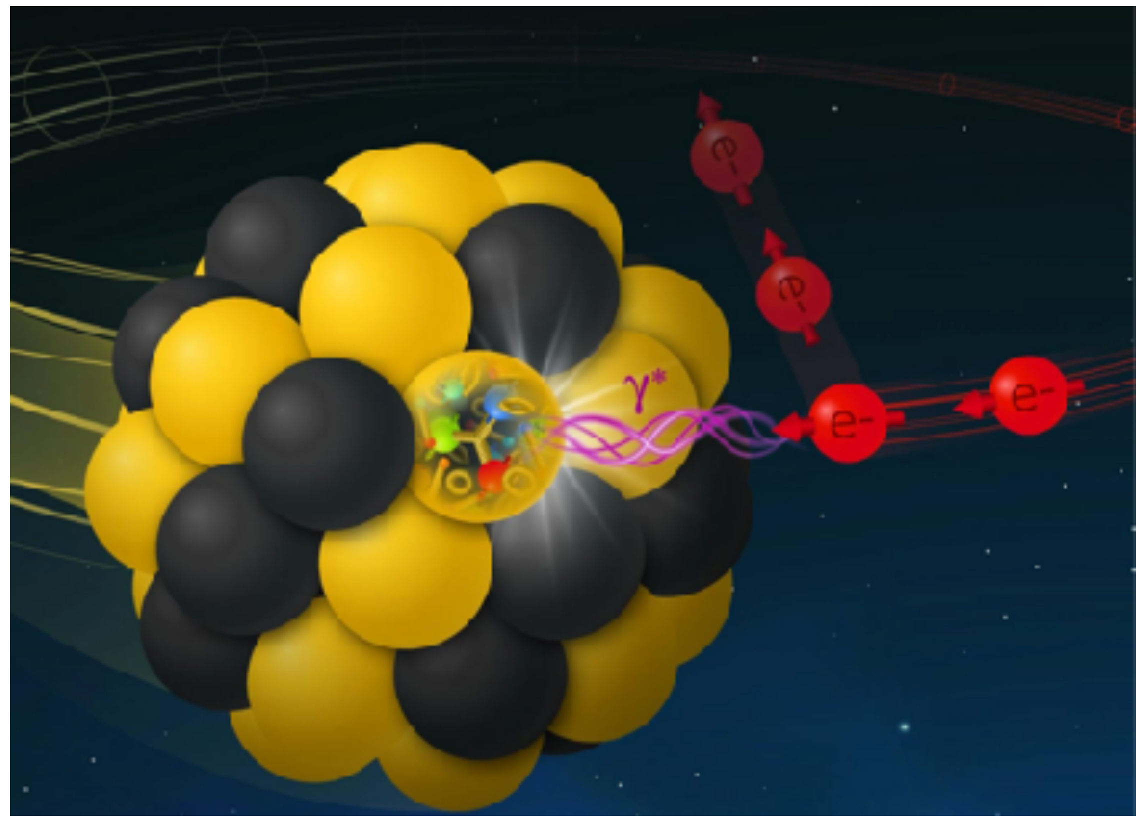
EMD observables

With geometric fitting we unify N-(sub)jettiness, Thrust, Event Isotropy etc.

These are handcrafted observables for the LHC, maybe find new ones?

A new playing field: the EIC

1. Electron Ion vs. Proton Proton
2. Lab Frame vs. Breit Frame



Find new observables that yield high information/discrimination!

N-Circliness, N-Ellipsiness or any shape of our choosing

Summary

1. 15-30 seconds of ads
2. The LHCb trigger
3. Lipschitz Networks - Robustness and Monotonicity
4. Energy Flow and Energy Movers Distance
5. Fitting with the EMD