11/01/2022

Rutgers University

# Topological Data Analysis for Collider Events

ML4Jets2022 Workshop

Session: Equivariance and New Architectures

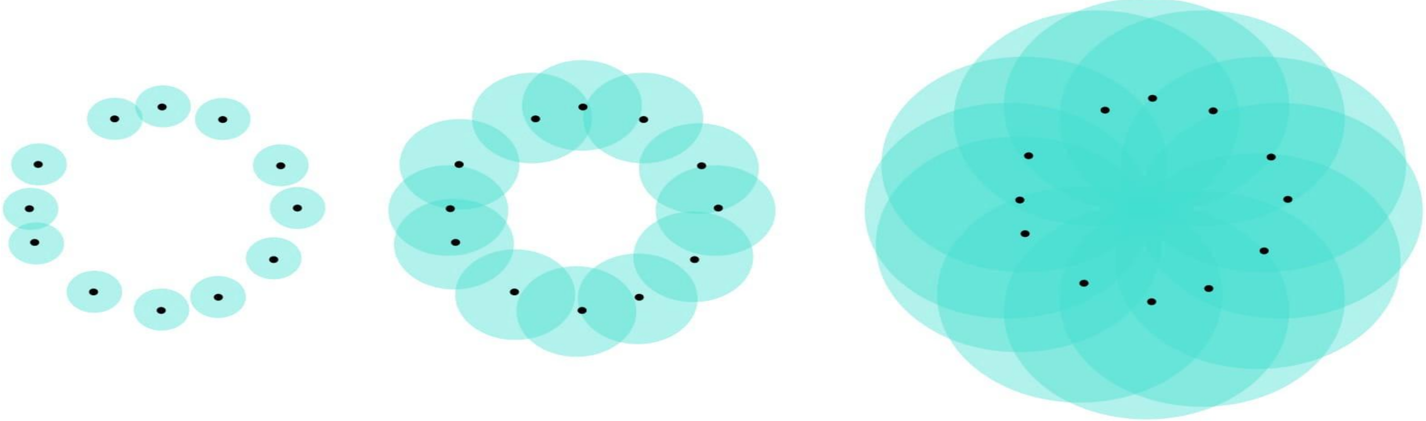
**Speaker: Tianji Cai (UCSB)**

Collaborators: Junyi Cheng (Harvard), Ian Dyckes (LBNL), Ben Nachman (LBNL).

*Work in progress. Coming Soon!*

Department of Physics

UC SANTA BARBARA



# Contents

Figure from this blog.

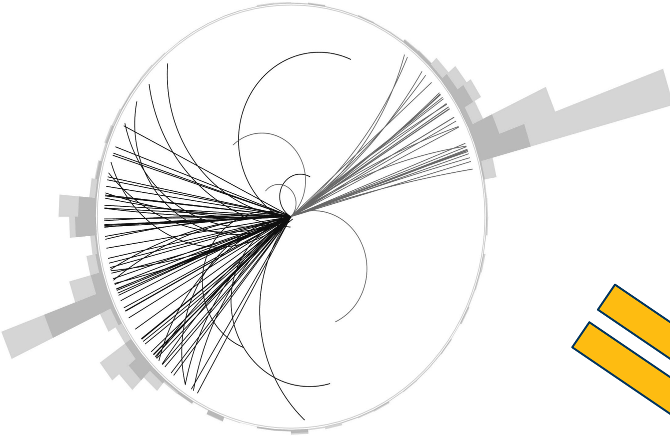
- **Introduction** — *Why, What, & How?*
- **Persistence Homology in a Nutshell** — *Not too scary math…*

  Filtration; Persistence Diagram (PD); PD Representation; Metric on the PD Space.

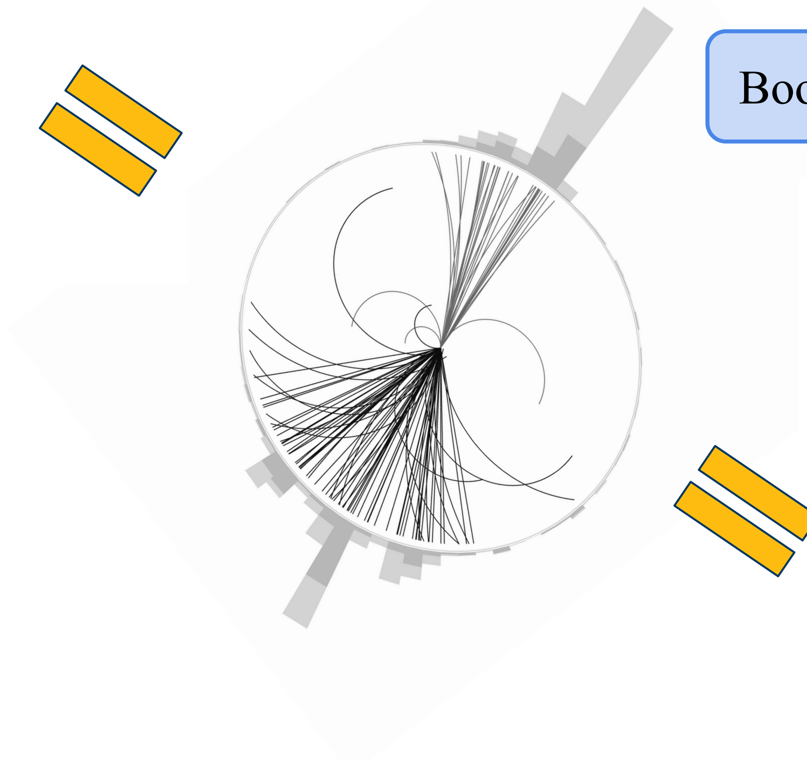
- **TDA for Jet Tagging** — *Data pre-processing not too much a trouble.*
- **TDA for Event Classification** — *Data pre-processing now a real issue.*
- **Summary & Outlook** — *So what's next?*

# 1. Introduction: *Why?*

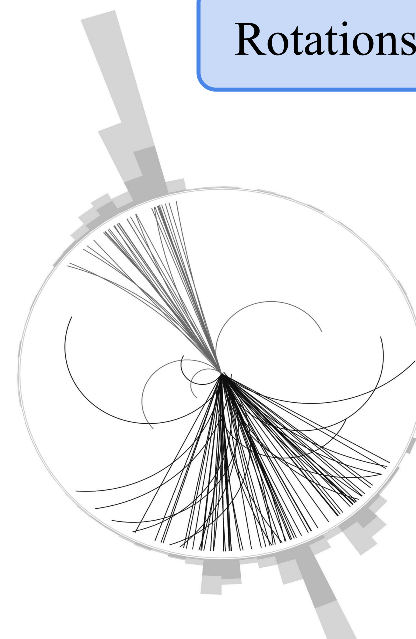
## *A Practical Nuisance…*



**Collider events and jets are inherently invariant under certain transformations.**



Boosts along the beam axis

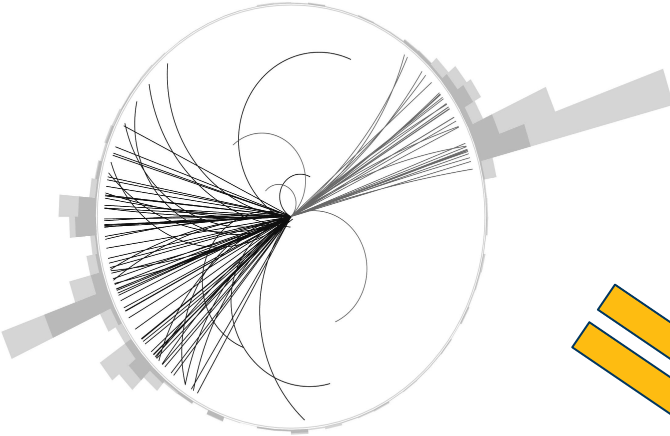


Rotations

Source: [CMS website](#).

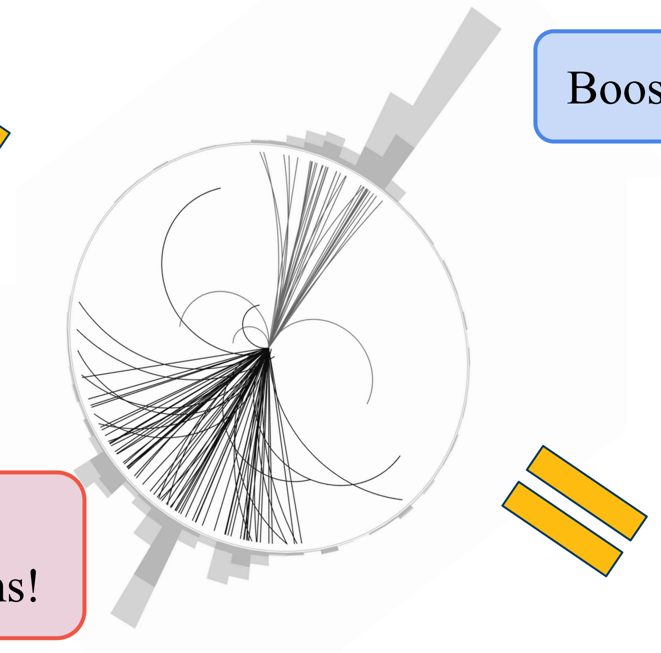
# 1. Introduction: *Why?*

## *A Practical Nuisance…*

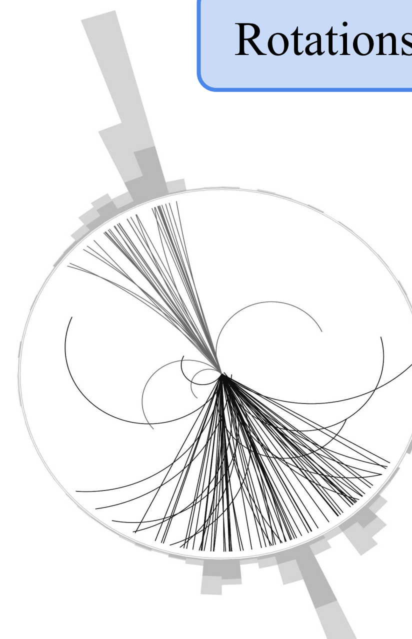


**Collider events and jets are inherently invariant under certain transformations.**

Boosts along the beam axis



Rotations



**Usual Solution:** Pre-process the events/jets to get rid of any artificial difference.

**Problem:** Such pre-processing is *ad-hoc* and only based on conventions!

**Proposal:** Design analysis frameworks that are invariant under these transformations, e.g., based on topology.

**Bonus of Topology-based Framework:** Can see the tagging power of topology alone when compared to the results of other geometry-based frameworks such as the optimal transport approach.

Source: [CMS website](CMS website).

# 1. Introduction: *What?*

**Topological Data Analysis** (TDA) aims at studying the complex topological structure of the underlying data.

# 1. Introduction: *What?*

**Topological Data Analysis** (TDA) aims at studying the complex topological structure of the underlying data.

One great TDA tool is **Persistence Homology** (PH). PH builds continuous shapes for a point cloud at different *scales* and analyzes the *evolution* of these shapes.

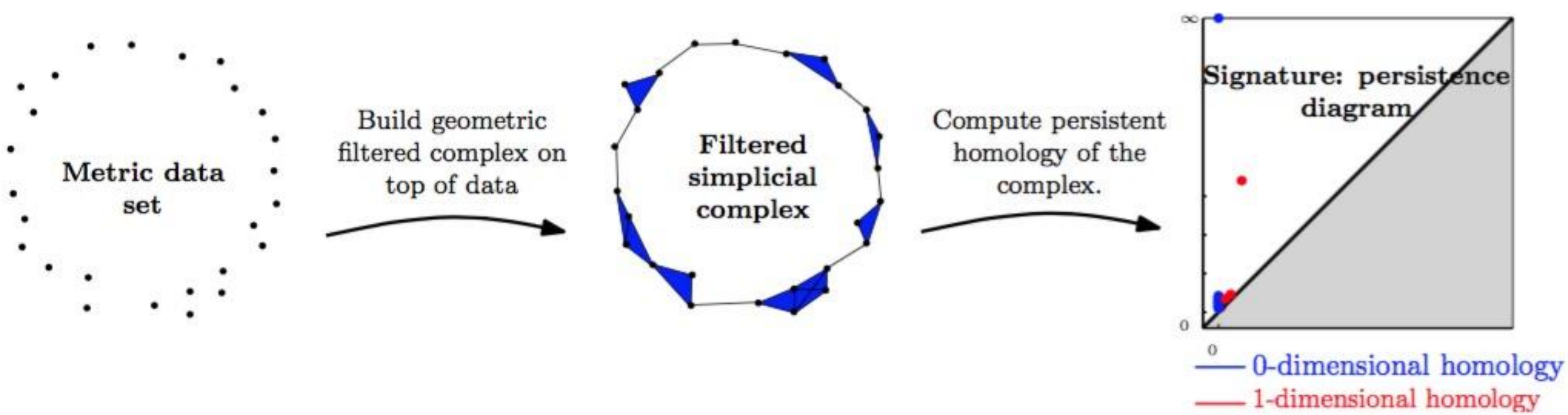
# 1. Introduction: *What?*

**Topological Data Analysis** (TDA) aims at studying the complex topological structure of the underlying data.

One great TDA tool is **Persistence Homology** (PH). PH builds continuous shapes for a point cloud at different *scales* and analyzes the *evolution* of these shapes.

PH is great because (i) well-understood theoretical framework based on algebraic geometry; (ii) efficient to compute; (iii) robust against small perturbations in input data.

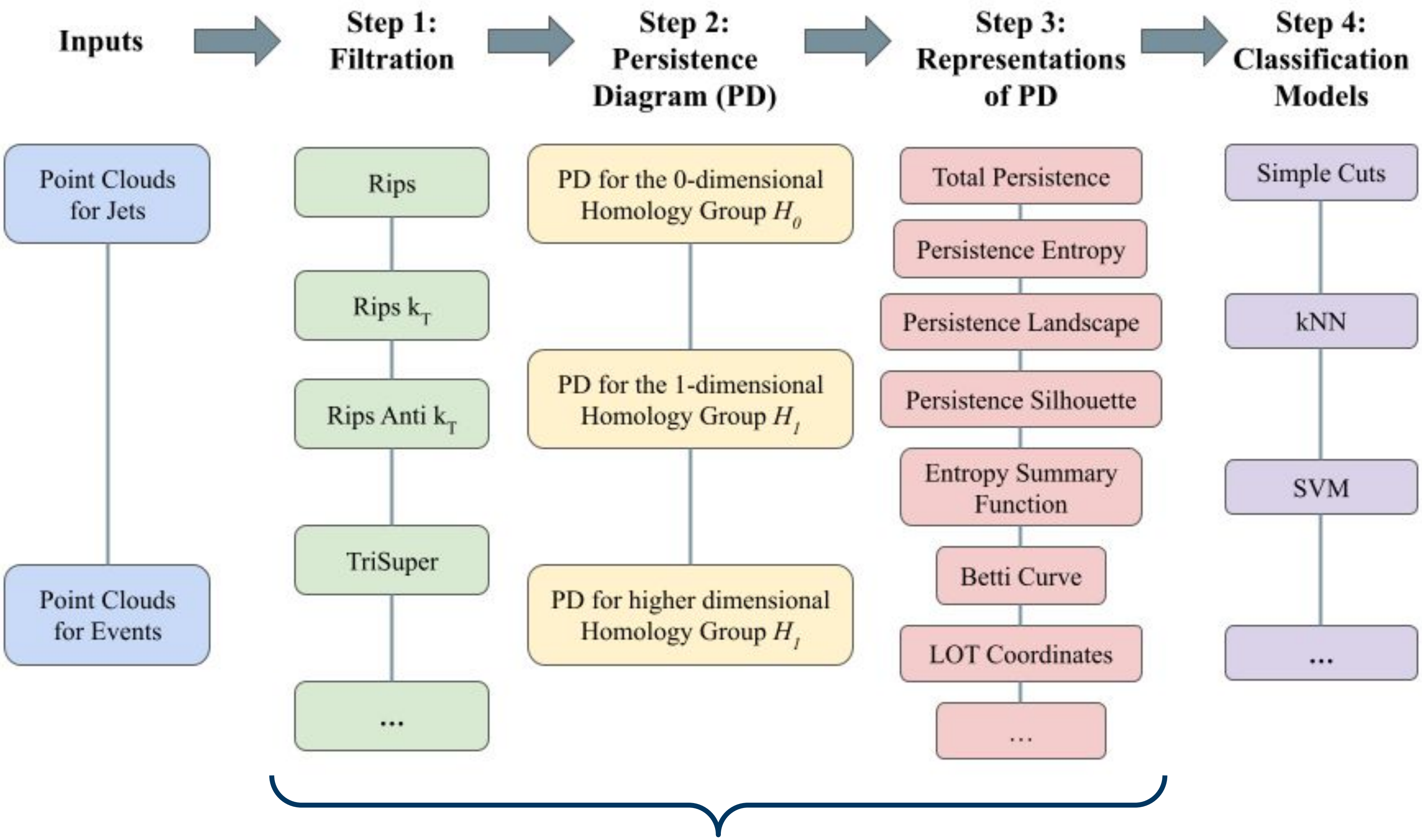
Suitable data types for PH: **finite metric spaces (point clouds)**, digital images, networks.



# 1. Introduction: *How?*

Related works: 2006.12446.

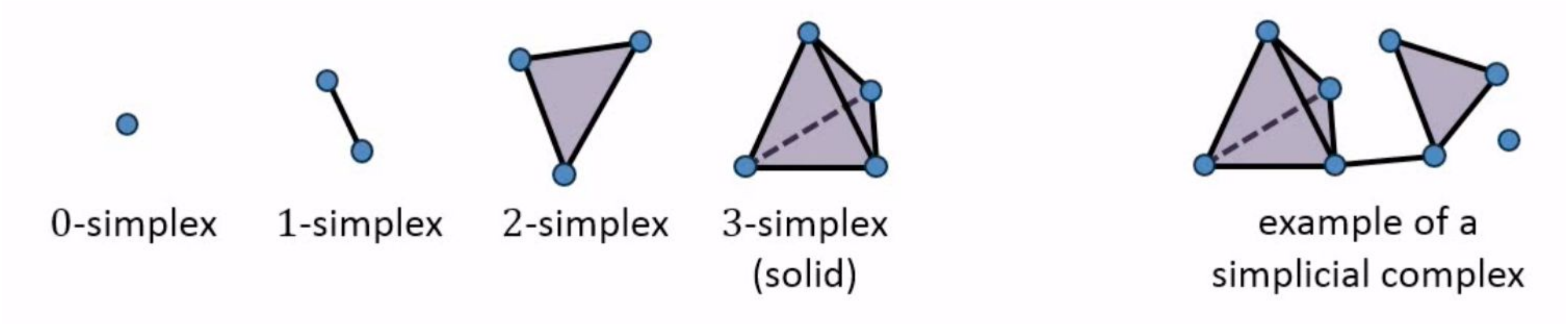
## TDA Workflow




**Topics to be discussed in §2.**
**Still under active research on the math side.**

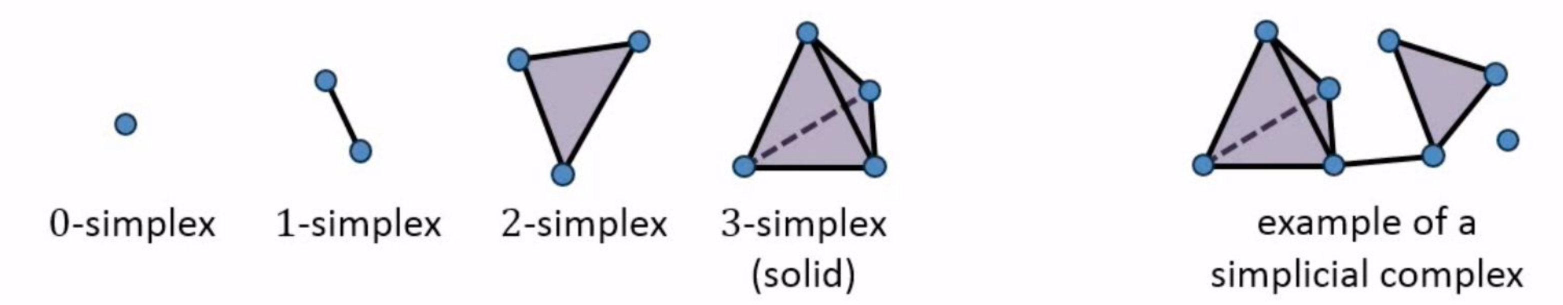
# 2. Persistence Homology in a Nutshell: *Filtration*

Build a "**simplicial complex**" from a point cloud.

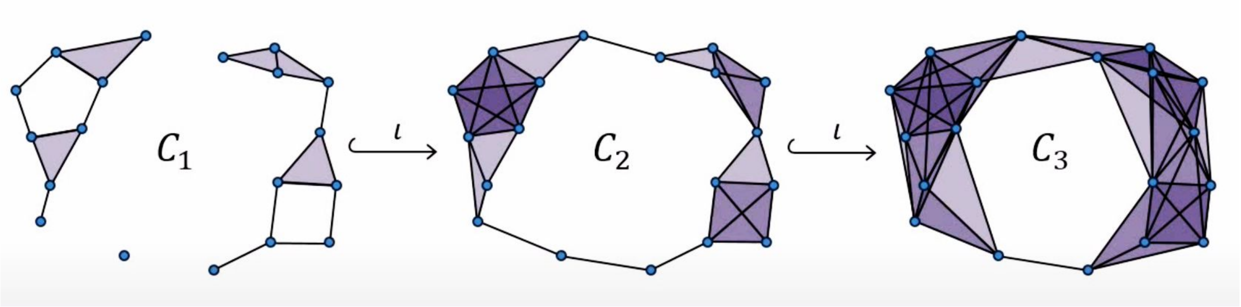


# 2. Persistence Homology in a Nutshell: *Filtration*

Build a "**simplicial complex**" from a point cloud.



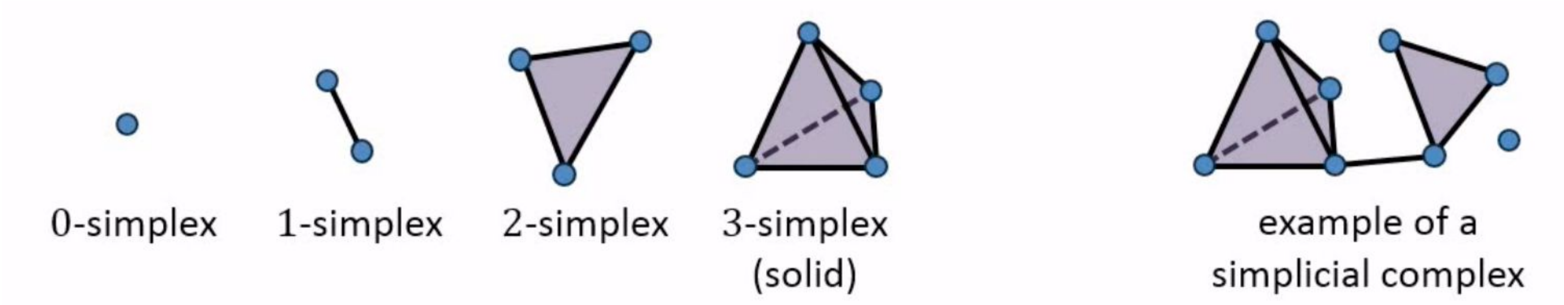

Construct a nested family of simplicial complexes called a **filtration**.



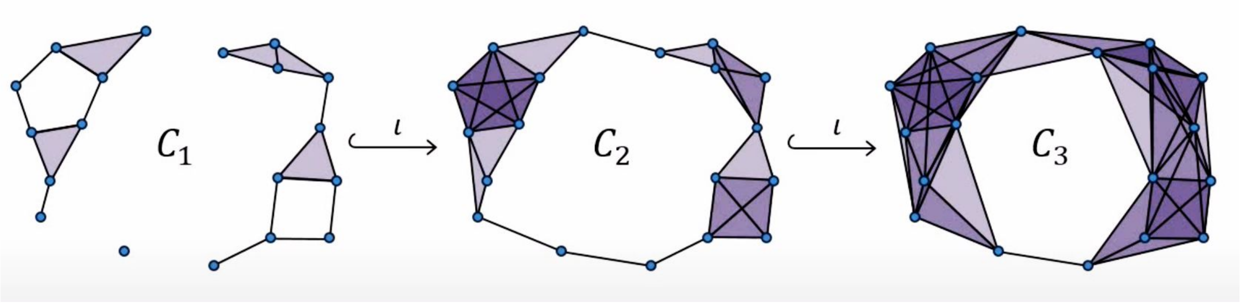

Source: Images from youtube.

# 2. Persistence Homology in a Nutshell: *Filtration*

Build a "**simplicial complex**" from a point cloud.

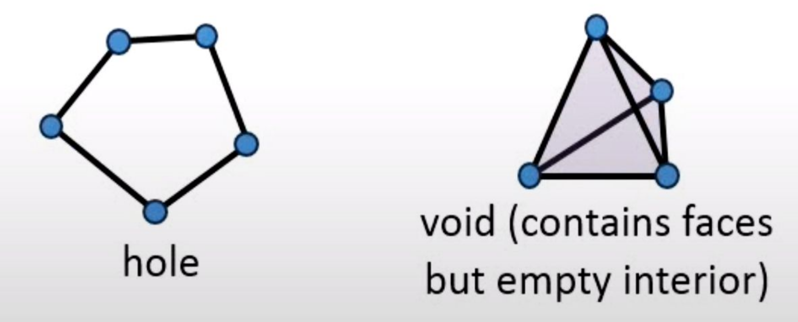



Construct a nested family of simplicial complexes called a **filtration**.



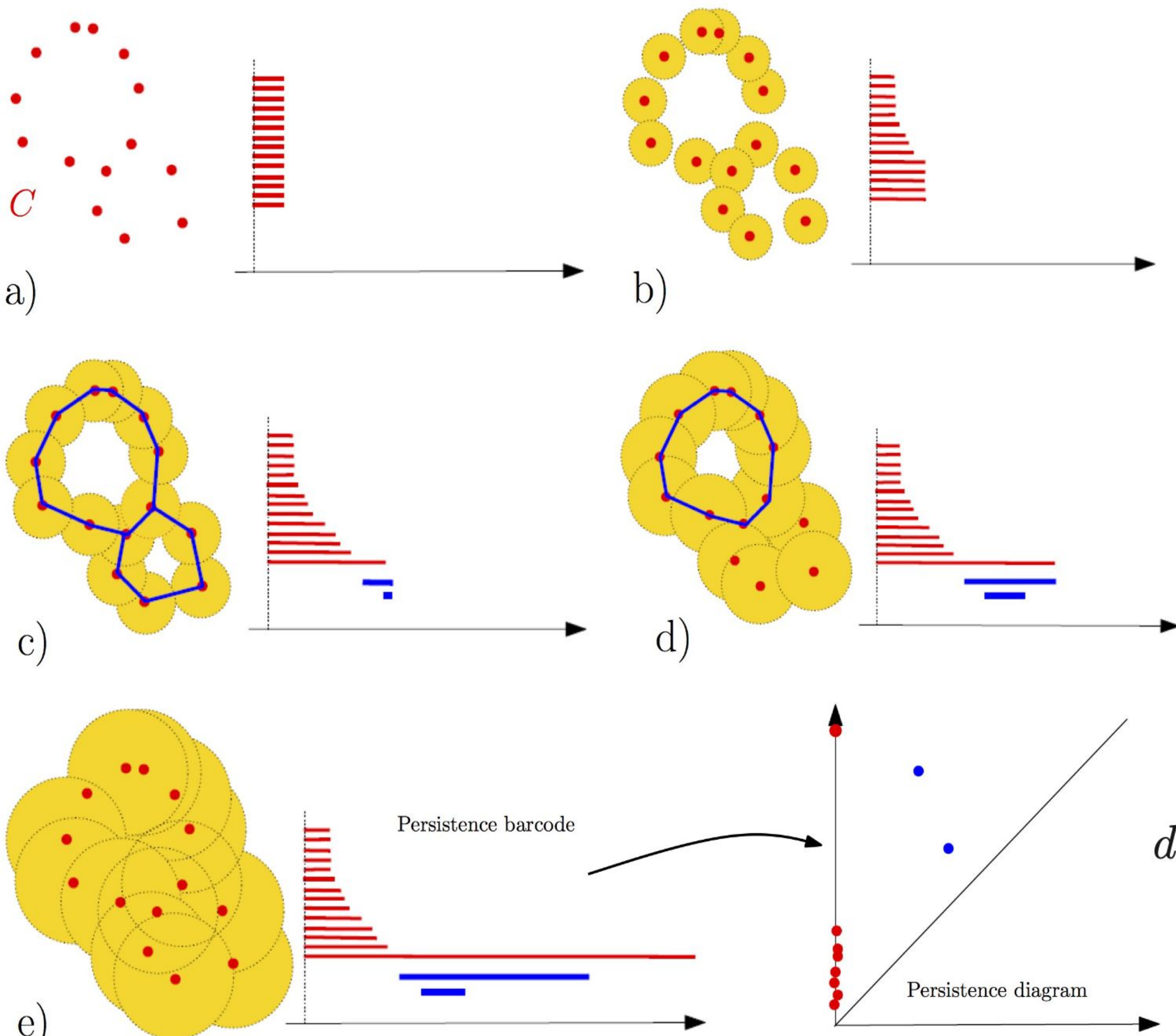

Source: Images from [youtube](youtube).

**Homology** counts connected components (0th dim), holes (1st dim), voids (2nd dim)...
*Homology of simplicial complexes are easily computable via linear algebra.*



# 2. Persistence Homology in a Nutshell: *Rips Filtration & PD*

Include simplex if the pairwise distances between all its vertices satisfy $d_{ij} < \alpha$.
Can be computed efficiently.




***We consider 3 physical distance functions:***

- ***Rips with C/A distance***

$$\Delta R^2 = \Delta y^2 + \Delta\phi^2$$

$$d_{ij} = \Delta R_{ij}^2 / R^2$$

- ***Rips with $k_T$ distance***

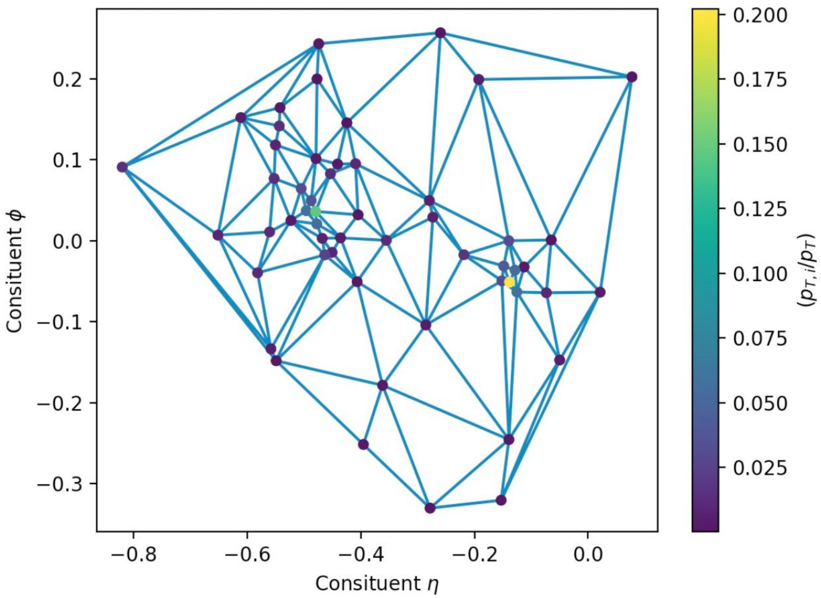
$$d_{ij} = \min(p_{ti}^2, p_{tj}^2)\, \Delta R_{ij}^2 / R^2$$

- ***Rips Anti $k_T$ distance***

$$d_{ij} = \min(1/p_{ti}^2, 1/p_{tj}^2)\, \Delta R_{ij}^2 / R^2$$

Source: Fig 10 in paper.

# 2. Persistence Homology in a Nutshell: *TriSuper Filtration & PD*



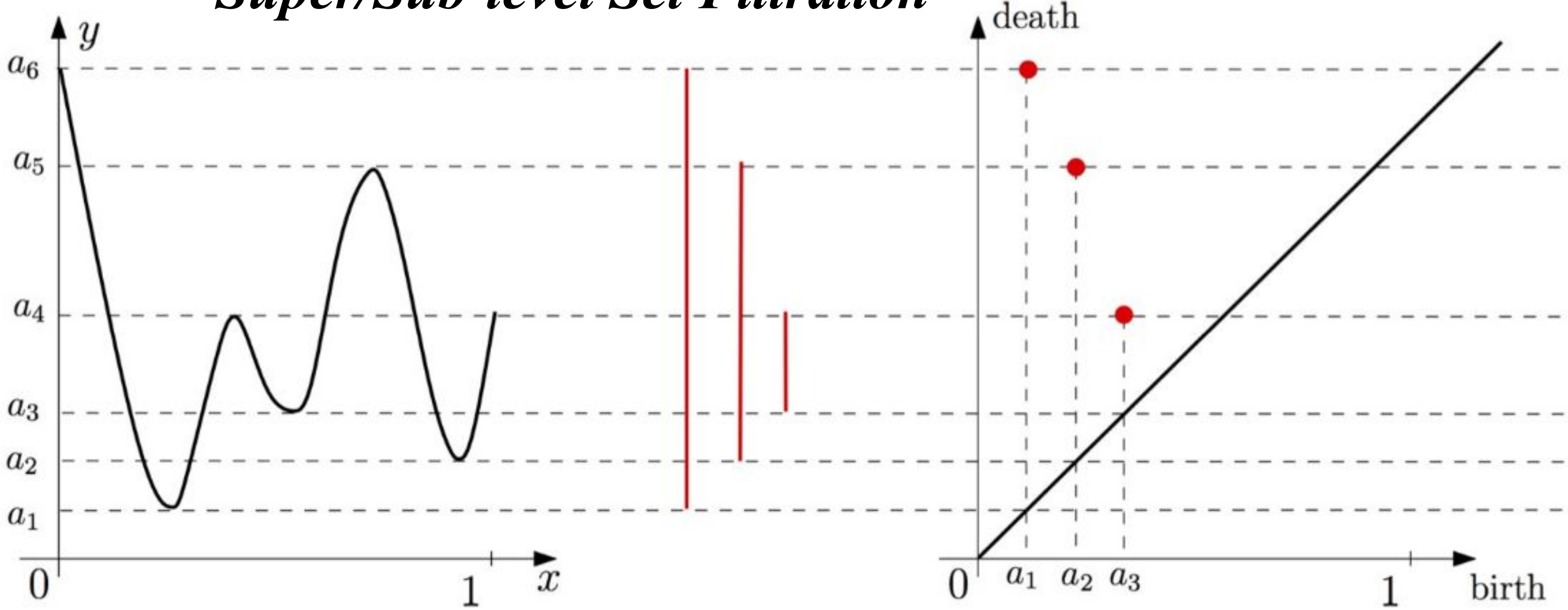

## ***Delaunay Triangulation (DT)***

The Dual graph of Voronoi Diagram:

$$V_s = \left\{x \in \mathbb{R}^d \mid \mathrm{d}(x,s) \leq \mathrm{d}\left(x,s'\right) \text{ for all } s' \in S\right\}.$$

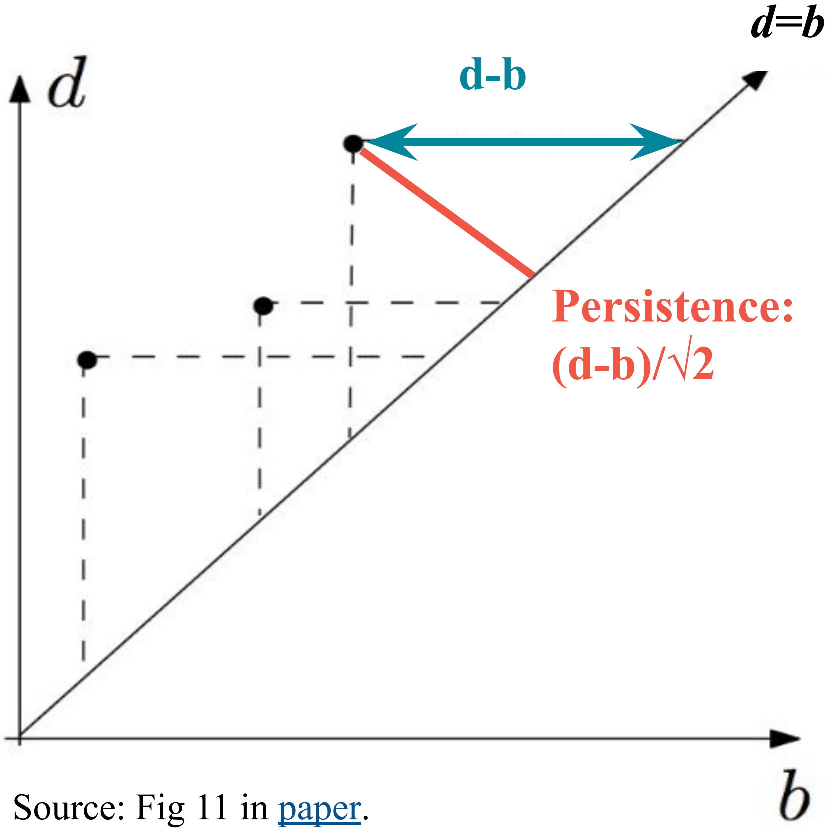
Original method in the paper *Jet Topology*.

## ***Super/Sub-level Set Filtration***




Source: Fig 9 in paper.

# 2. Persistence Homology in a Nutshell: *PD Representations*




Source: Fig 11 in [paper](paper).

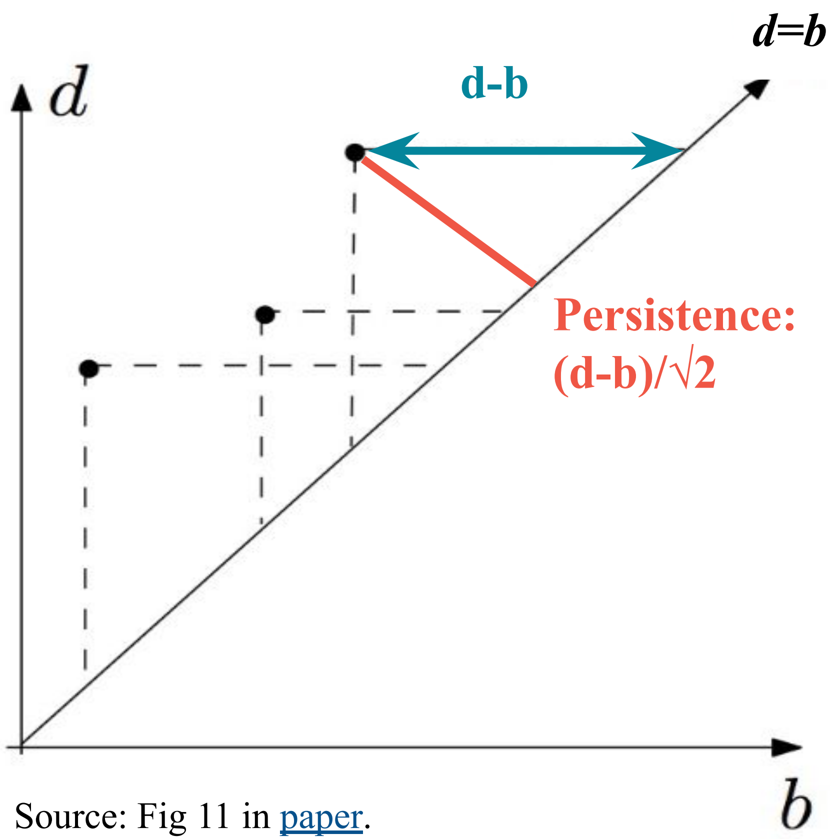
***A PD with $n_a$ off-diagonal points***

$$p_i = (b_i, d_i) \in \mathbb{R}^2 \quad i = 1, \ldots, n_a.$$

***Define the persistence of each point as***

$$m_i = \frac{l_i}{\sqrt{2}}, \qquad l_i = d_i - b_i$$

# 2. Persistence Homology in a Nutshell: *PD Representations*




Source: Fig 11 in paper.

***A PD with $n_a$ off-diagonal points***

$$p_i = (b_i, d_i) \in \mathbb{R}^2 \quad i = 1, \ldots, n_a.$$

***Define the persistence of each point as***

$$m_i = \frac{l_i}{\sqrt{2}}, \qquad l_i = d_i - b_i$$

**Total Persistence *T*: Scalar**

$$T[A] = \sum_{i=1}^{n_a} m_i = \frac{L}{\sqrt{2}}$$

**Betti Curve *β(t)*: Vector**

$$\beta_p[A](t) = \sum_{i=1}^{n_a} w_i(t)$$

$$w_i(t) = \begin{cases} 1 & b_i \le t \le d_i \\ 0 & \text{otherwise} \end{cases}$$

**Persistence Entropy *E*: Scalar**

$$E[A] = -\sum_{i=1}^{n_a} \frac{l_i}{L} \log\left(\frac{l_i}{L}\right) = -\sum_{i=1}^{n_a} \frac{m_i}{T} \log\left(\frac{m_i}{T}\right)$$

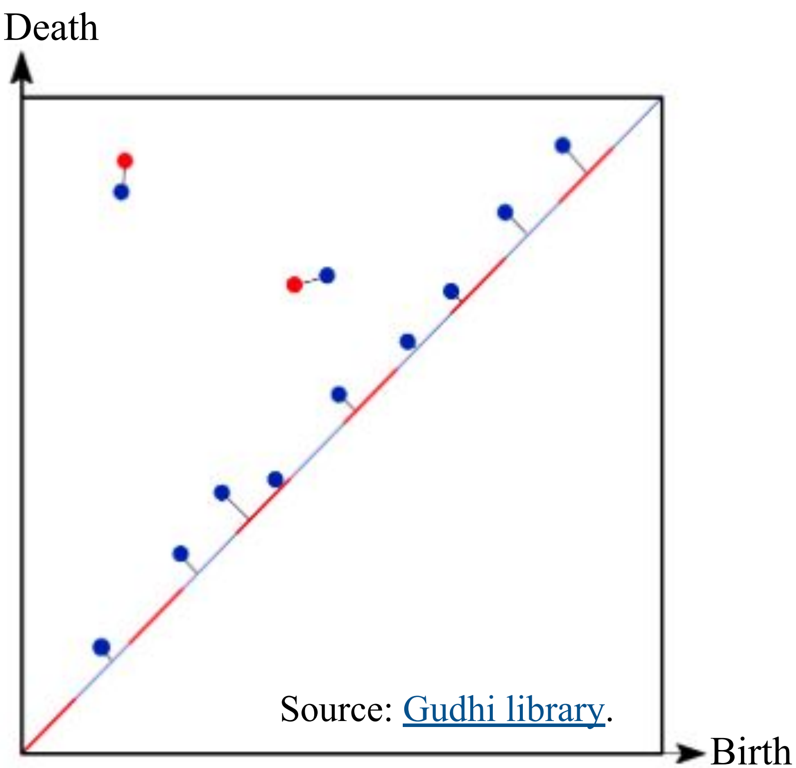
# 2. Persistence Homology in a Nutshell: *Metric on the PD Space*

**Under Active Research!**

***Existing Method:*** *The pth Wasserstein ($W_p$) distance*

$$W_p[\mathrm{d}](X,Y) := \inf_{\phi:X\to Y}\left[\sum_{x\in X}\mathrm{d}[x,\phi(x)]^p\right]^{1/p}$$

=> *The Bottleneck distance* $W_\infty[L_\infty]$ *with* $W_\infty[\mathrm{d}](X,Y) := \inf_{\phi:X\to Y}\sup_{x\in X}\mathrm{d}[x,\phi(x)]$




Source: Gudhi library.

# 2. Persistence Homology in a Nutshell: *Metric on the PD Space*

**Under Active Research!**

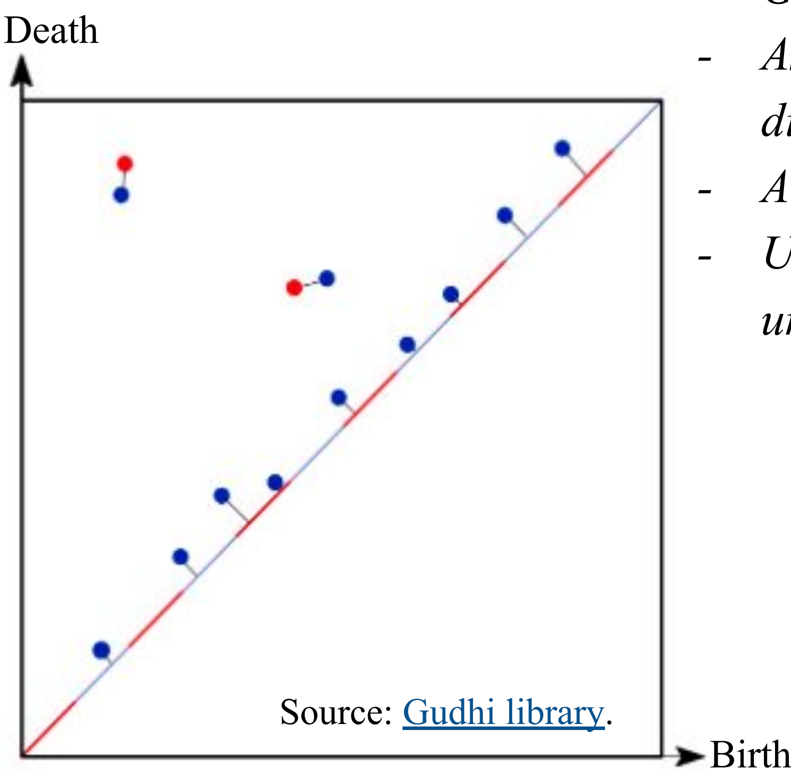
***Existing Method:*** *The pth Wasserstein ($W_p$) distance*

$$W_p[\mathrm{d}](X,Y) := \inf_{\phi:X\to Y}\left[\sum_{x\in X}\mathrm{d}[x,\phi(x)]^p\right]^{1/p}$$

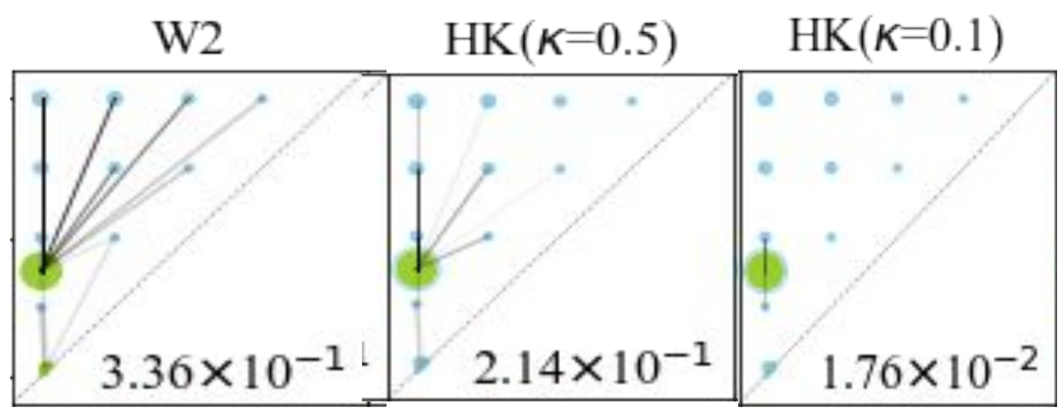
=> *The Bottleneck distance* $W_\infty[L_\infty]$ *with* $W_\infty[\mathrm{d}](X,Y) := \inf_{\phi:X\to Y}\sup_{x\in X}\mathrm{d}[x,\phi(x)]$

***Our Proposal:***

- *Get rid of the diagonal.*
- *Assign mass to points based on their distances to the diagonal.*
- *Allow mass creation & destruction.*
- *Use Hellinger-Kantorovich (HK) distance, an unbalanced OT generalization of the $W_2$ distance.*




Source: [Gudhi library](Gudhi library).



# 3. TDA for Jet Tagging: *TriSuper Filtration & Its PD*

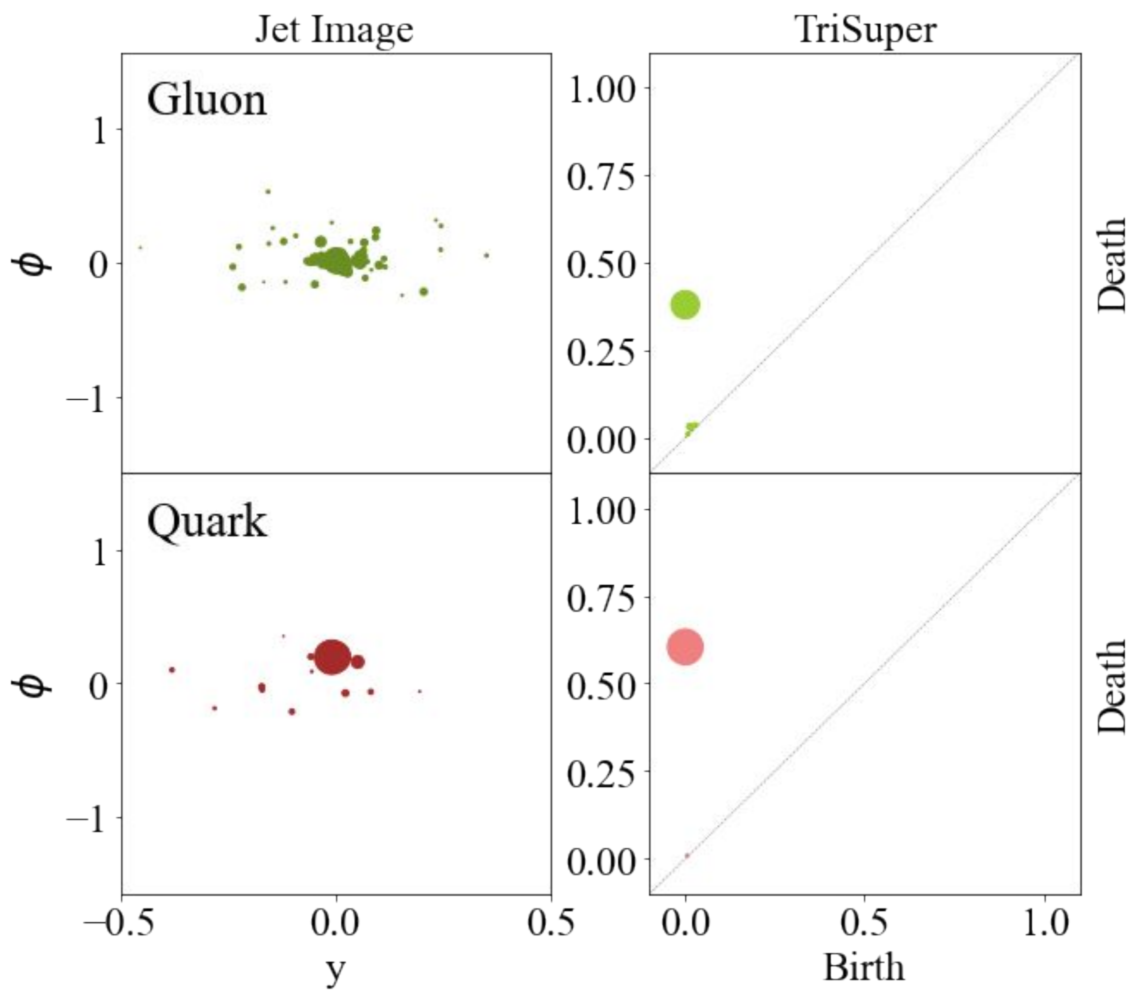
**Dataset:** 10k light QCD jets with total $p_T$ in [100, 350] GeV.

**Simulation:** pp collisions at $\sqrt{s}$=14 TeV; anti-kt jet clustering with R=0.6; jets selected with $|y| < 1.7$.

**Filtration:** Delaunay Triangulation+Superlevel Set Filtration (***TriSuper***).

**Dimension:** 0th homological dimension.

We also examined **Rips filtration** with C/A, $k_T$, and Anti $k_T$ distances for 0th-dim.



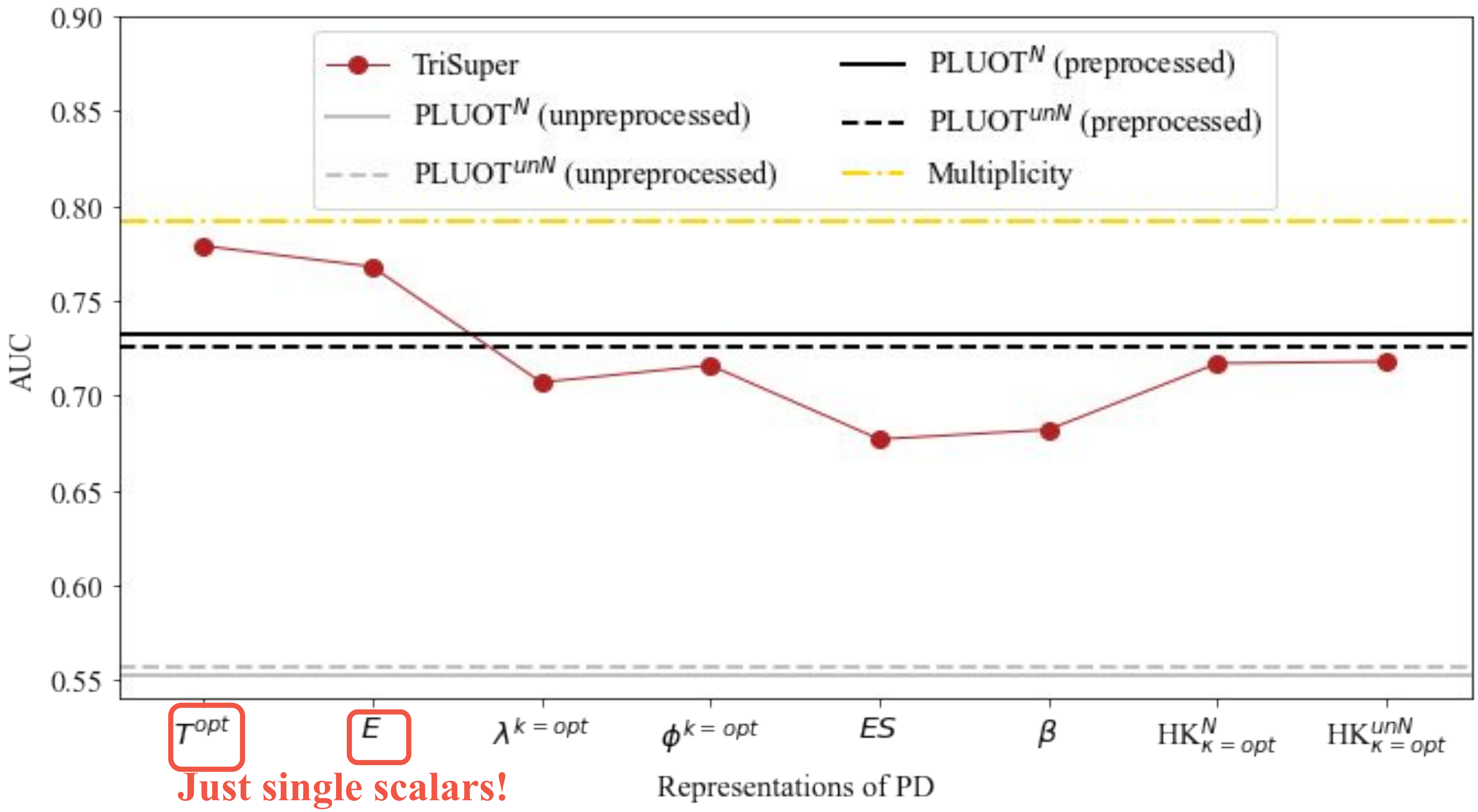
# 3. TDA for Jet Tagging: *Result for TriSuper*

**Topology-based observables:** TriSuper of 0th-dim with various PD representations.

**Geometry-based observables:** PLUOT framework.

**Traditional observable:** Multiplicity—*optimal observable*.

**Classifiers:** Simple cuts for scalars; kNN for vectors.

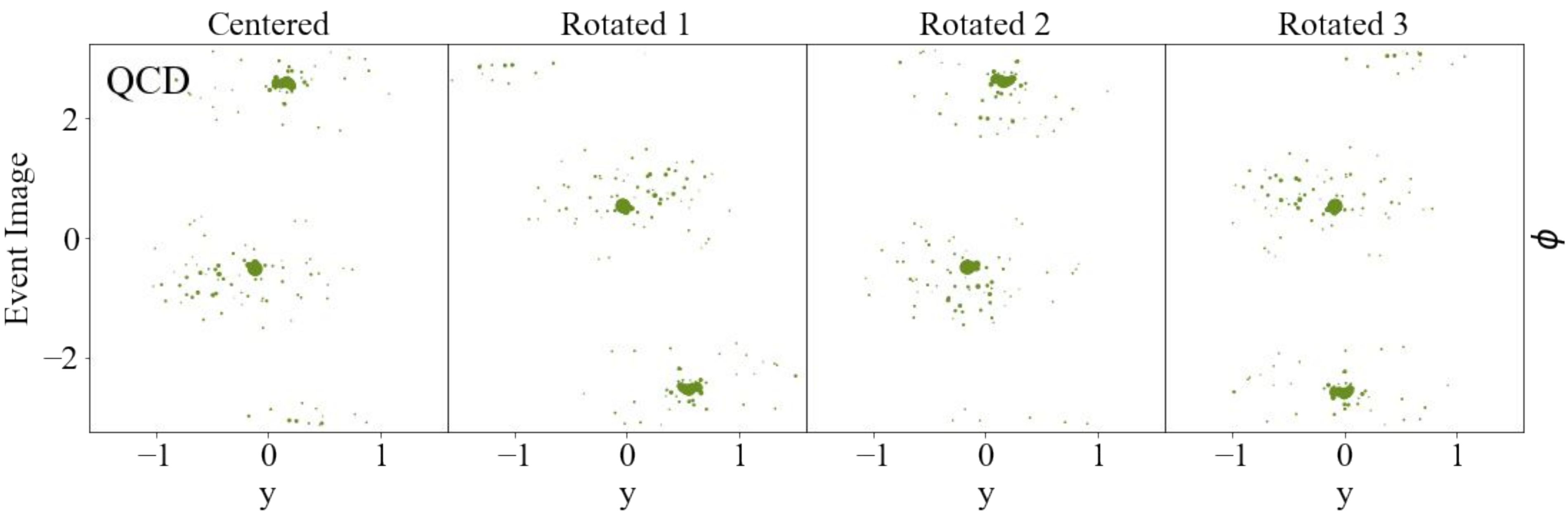


# 4. TDA for Event Classification: *Issue of Preprocessing*

**Dataset:** 10k dijet W boson and QCD events with √s=14 TeV.

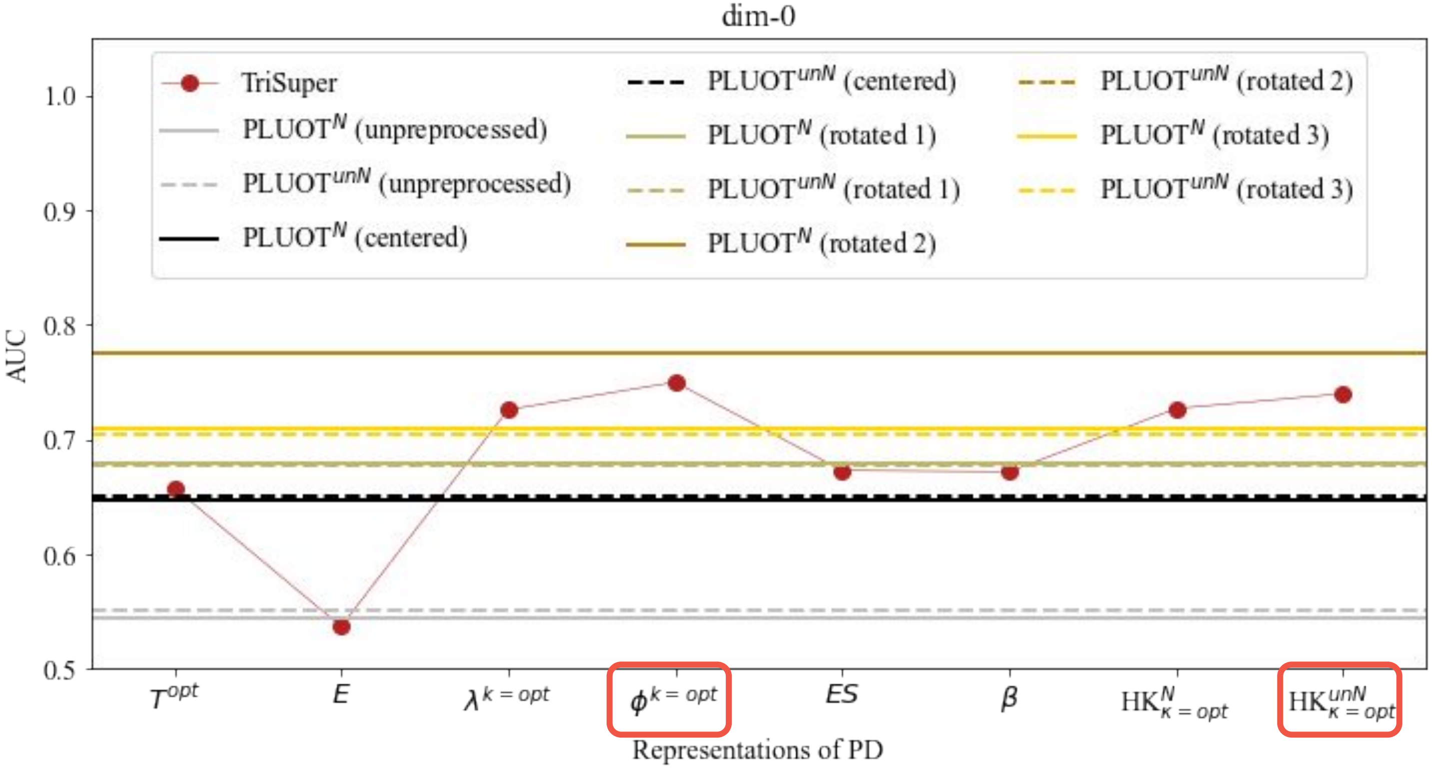
**Simulation Highlights:** Anti-kt algorithm for jet clustering with R=1; individual jet with a $p_T$ in [500, 550] GeV and |y| < 1.7 cut.

**Preprocessing:** (i) Boost the events to their CM frames (*centered*) ✅; (ii) Rotate the events according to three different schemes (*Rotated 1, 2, 3*).



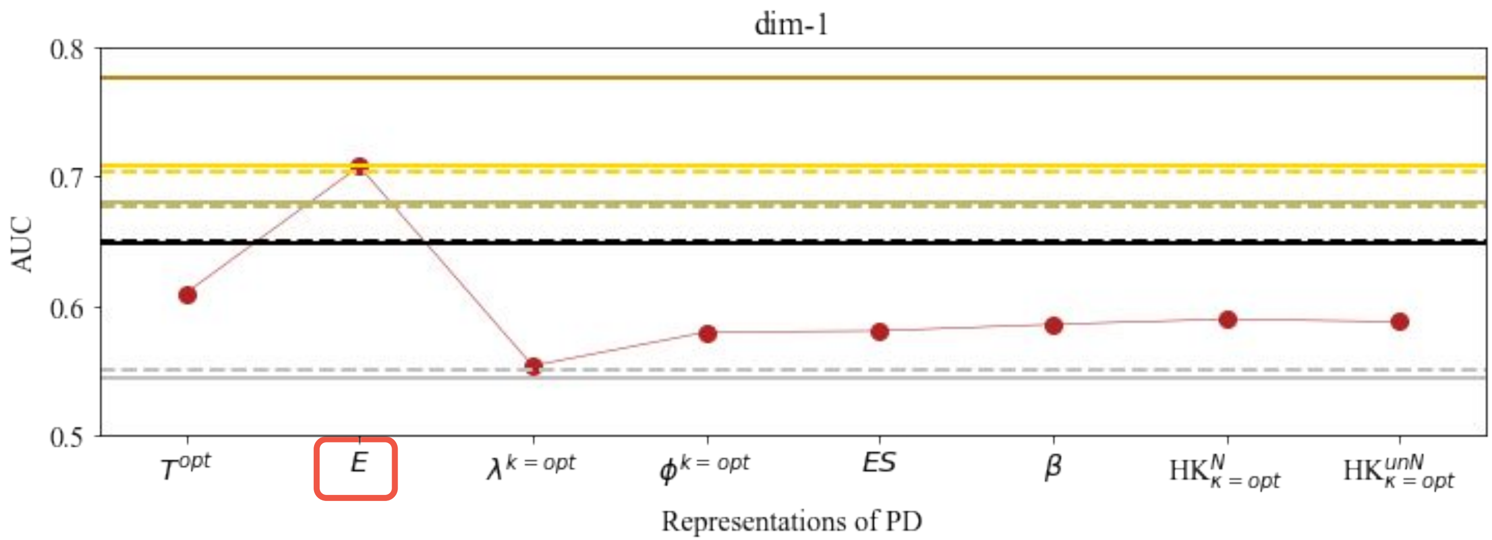

**Which is the best rotation scheme?**
**A priori unknown & all *ad-hoc*!**

# 4. TDA for Event Classification: *Result for TriSuper*




***0th Homological Dimension:*** *Connected components*

***1st Homological Dimension:*** *Holes*



# 5. Summary: *What have we done in this study?*

- **Physics side:**
  - We introduced the TDA framework to get rid of *ad-hoc* pre-processing of LHC events.
  - We examined several filtrations with the 0th and 1st homological dimensions
  - We compared tagging performance of various PD representations with standard observables and geometry-based optimal transport framework.
  - Even a single scalar representation of PD achieves close to optimal performance for jet tagging.
  - The topology of an event is more complex; therefore more sophisticated PD representations are preferred.
  - TDA-based taggers (0th-dim) perform consistently better than geometry-based frameworks without pre-processing. This is especially important for event analysis.
- **Math side:**
  - We proposed a new way to present homology in a persistence diagram, getting rid of the diagonal.
  - This enabled the full use of HK distances to define a metric on the space of PDs.
  - Linearized HK offers a novel way to represent a PD, potentially useful for topologically more challenging datasets.

# 5. Outlook: *What's next?*

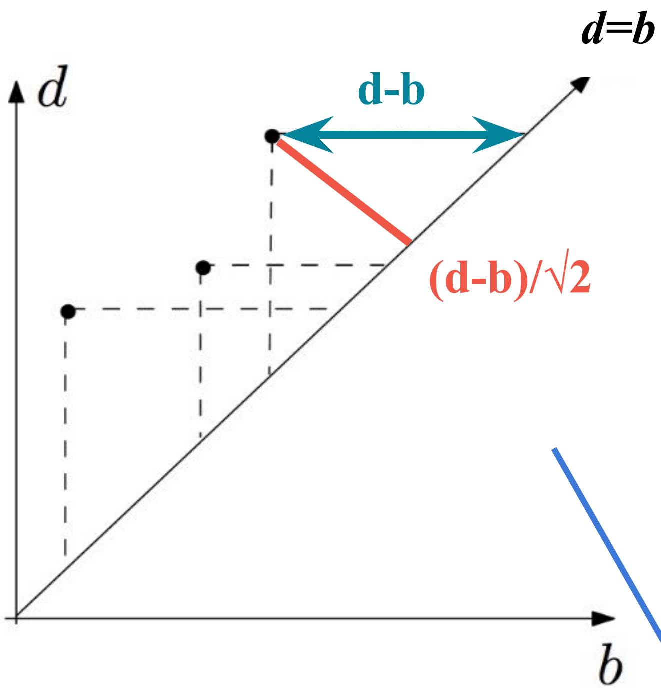
- **Math side:**
  - To rigorously prove that our proposed unbalanced HK distance is a good metric for the PD space (convergence, stability, etc).
  - To fully develop linearized HK embedding as a more sophisticated representation of a PD for statistical analysis.
  - To find more topologically challenging datasets that may showcase the power of our novel PD representation.
- **Physics side:**
  - To fully understand why certain combinations of filtrations and PD representations perform better for a given jet/event tagging task.
  - To find a non-trivial way to combine the tagging powers of the 0th and 1st homological dimensions.
  - To explore other filtrations based on physics considerations.
  - To apply PH on datasets with more complex topological structures, e.g., top events.
  - To explore further TDA tools for collider physics.

# THANKS!

Presented by Tianji Cai

# Backup Slides

# 2. Persistence Homology in a Nutshell: *PD Representations*




*A PD with n off-diagonal points*

$$p_i = (b_i, d_i) \in \mathbb{R}^2 \quad i = 1, \ldots, n_a.$$

*Define the persistence of each point as*

$$m_i = \frac{l_i}{\sqrt{2}}, \quad l_i = d_i - b_i$$

**Total Persistence $T$**

$$T[A] = \sum_{i=1}^{n_a} m_i = \frac{L}{\sqrt{2}}$$

**Persistence Entropy $E$**

$$E[A] = -\sum_{i=1}^{n_a} \frac{l_i}{L} \log\left(\frac{l_i}{L}\right) = -\sum_{i=1}^{n_a} \frac{m_i}{T} \log\left(\frac{m_i}{T}\right)$$

**Entropy Summary Function *ES(t)***

$$\mathrm{ES}[A](t) = -\sum_{i=1}^{n_a} w_i(t) \frac{l_i}{L} \log\left(\frac{l_i}{L}\right) = -\sum_{i=1}^{n_a} w_i(t) \frac{m_i}{T} \log\left(\frac{m_i}{T}\right)$$

$$w_i(t) = \begin{cases} 1 & b_i \leq t \leq d_i \\ 0 & \text{otherwise} \end{cases}$$

**Betti Curve $\boldsymbol{\beta(t)}$**

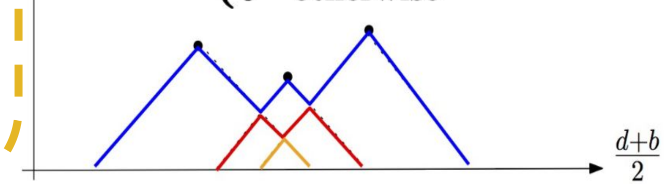
$$\beta_p[A](t) = \sum_{i=1}^{n_a} w_i(t)$$

**Persistence Landscape $\boldsymbol{\lambda^k(t)}$**

$$\lambda^k[A](t) = \mathrm{kmax}_{i=1,\ldots,n_a} \Lambda_i(t)$$

$$\frac{d-b}{2}$$

$$\Lambda_i(t) = \begin{cases} t - b_i & t \in \left[b_i, \frac{b_i + d_i}{2}\right] \\ d_i - t & t \in \left(\frac{b_i + d_i}{2}, d_i\right] \\ 0 & \text{otherwise} \end{cases}$$




**Persistence Silhouette $\boldsymbol{\phi^k(t)}$**

$$\phi^k[A](t) = \frac{\sum_{i=1}^{n_a} {l_i}^k \Lambda_i(t)}{\sum_{i=1}^{n_a} {l_i}^k}$$

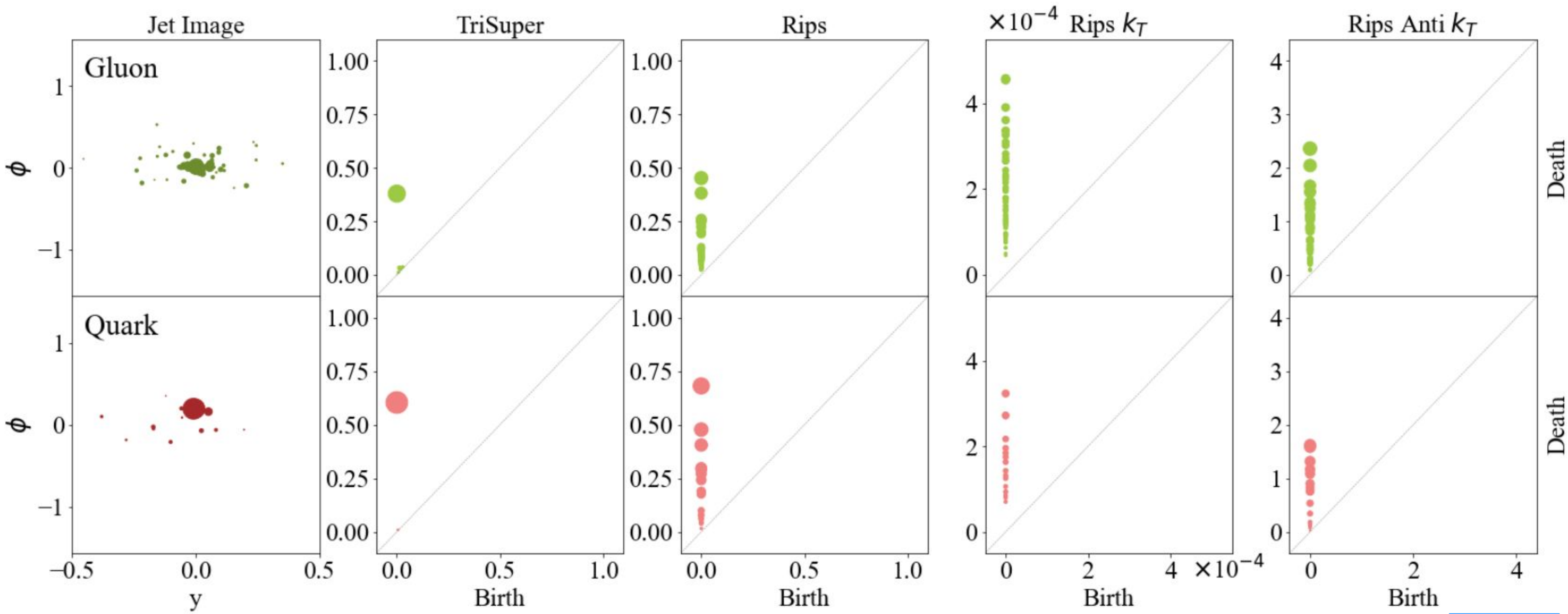
Source: Fig 11 in paper.

# 3. TDA for Jet Tagging: *Different Types of Filtrations & PDs*

**Dataset:** 10k light QCD jets with total $p_T$ in [100, 350] GeV.

**Simulation Highlights:** pp collisions at $\sqrt{s}$=14 TeV; anti-kt algorithm for jet clustering with R=0.6; jets selected with |y| < 1.7.

**Filtrations:** (i) Delaunay Triangulation+Superlevel Set Filtration (***TriSuper***); (ii) Rips Filtration with C/A distance (***Rips***); (iii) Rips with $k_T$ distance (***Rips*** $\boldsymbol{k_T}$); (iv) Rips with anti-$k_T$ distance (***Rips Anti*** $\boldsymbol{k_T}$).



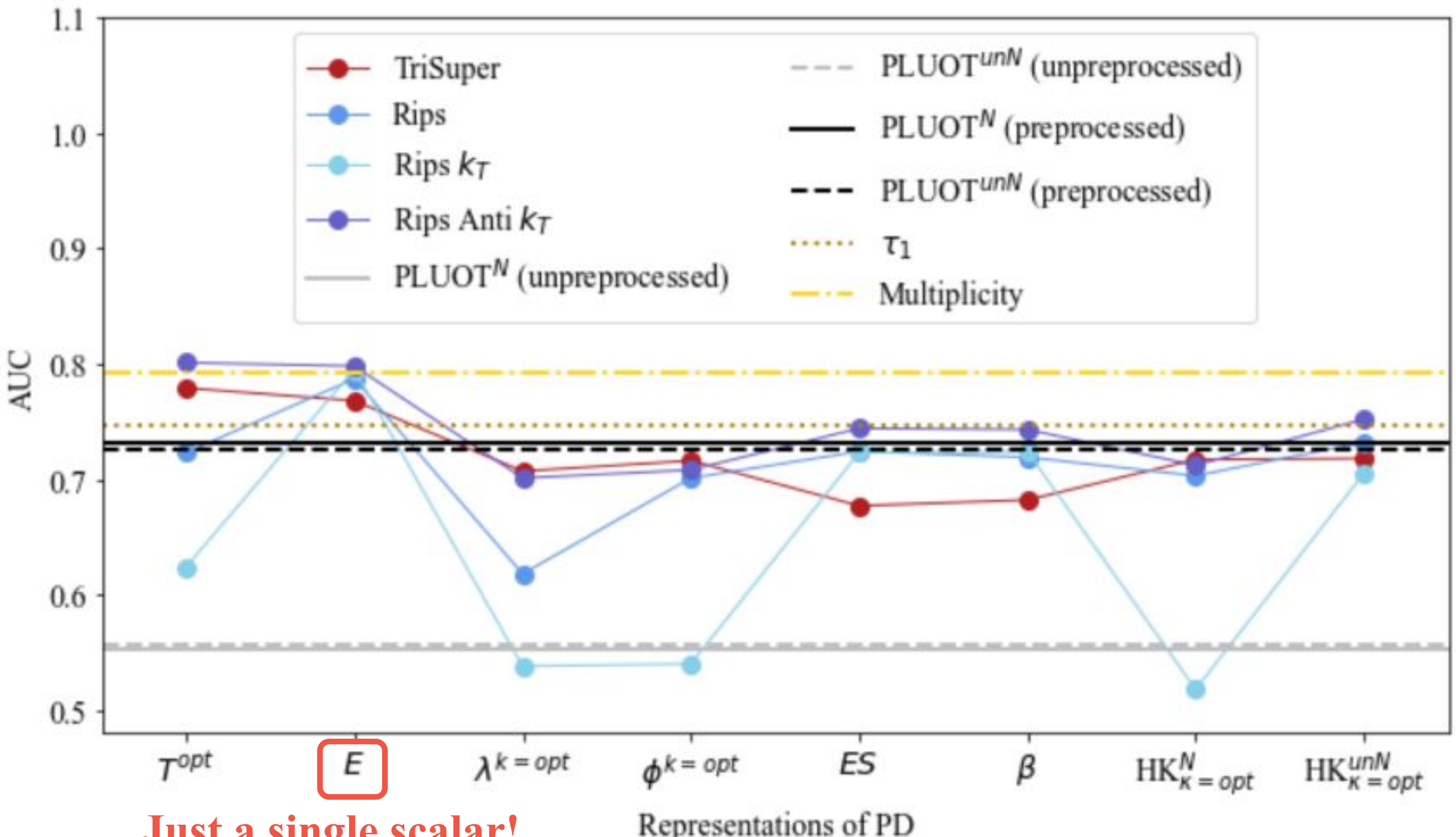
# 3. TDA for Jet Tagging: *Results*

**Topology-only observables:** Four filtrations with various PD representations (*colorful lines for filtrations; marks on x-axis for PD representations*).

**Geometry-based observables:** PLUOT framework (*gray & black*).

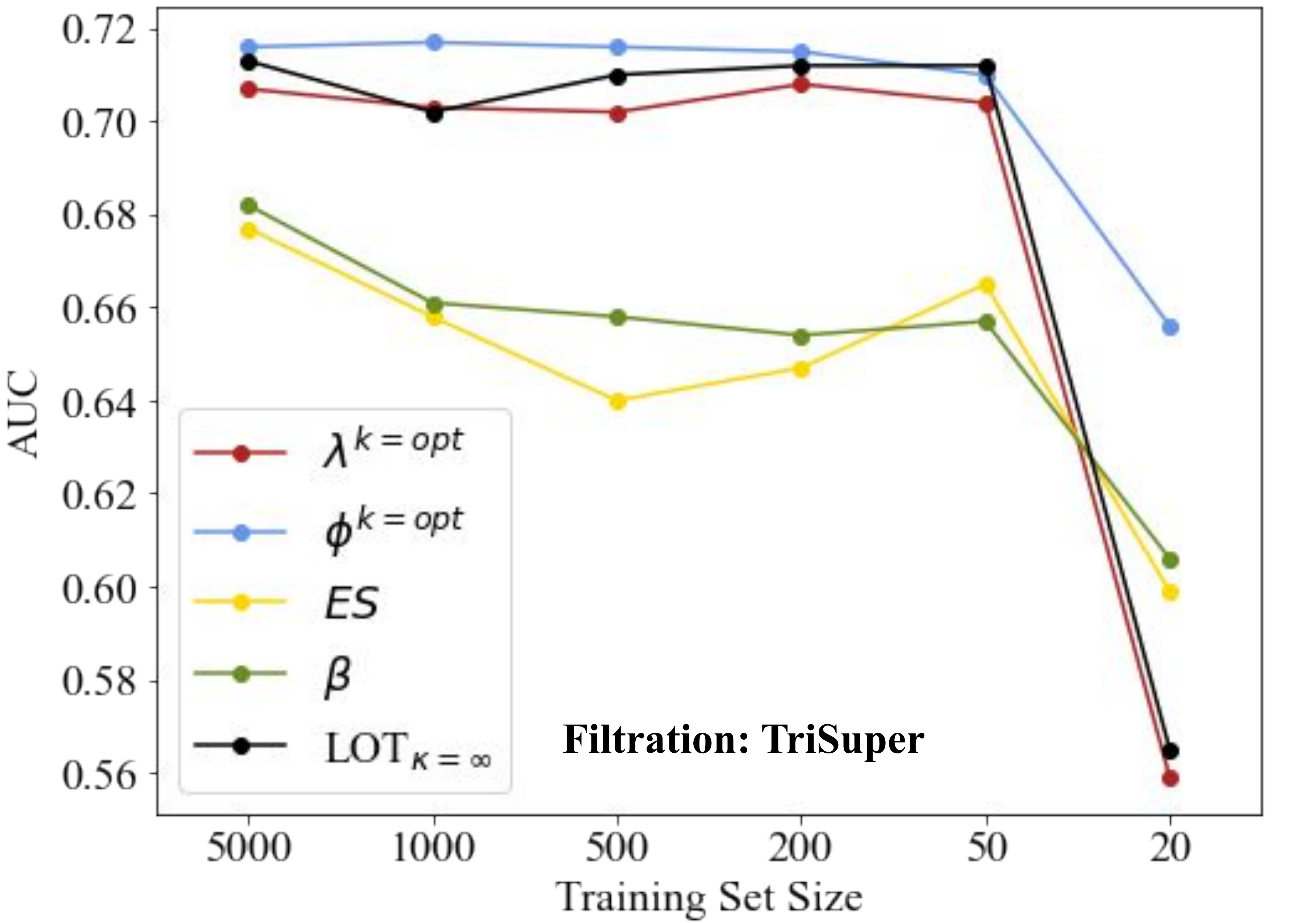
**Traditional observable:** *N*-subjettiness $\boldsymbol{\tau}_1$ (*brown*); Multiplicity (*yellow*).

**Classifiers:** Simple cuts for scalars; kNN for vectors.

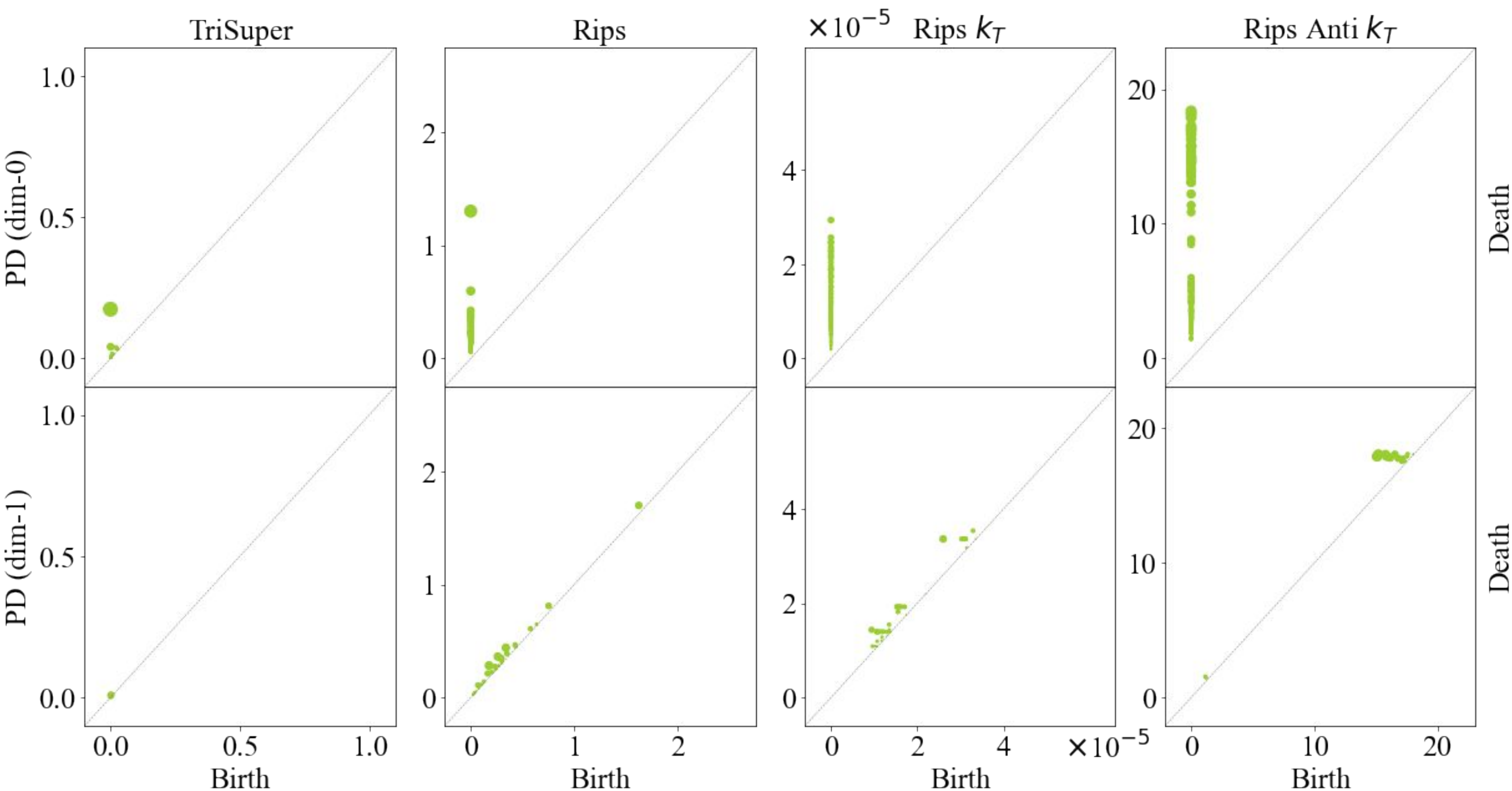



Just a single scalar!

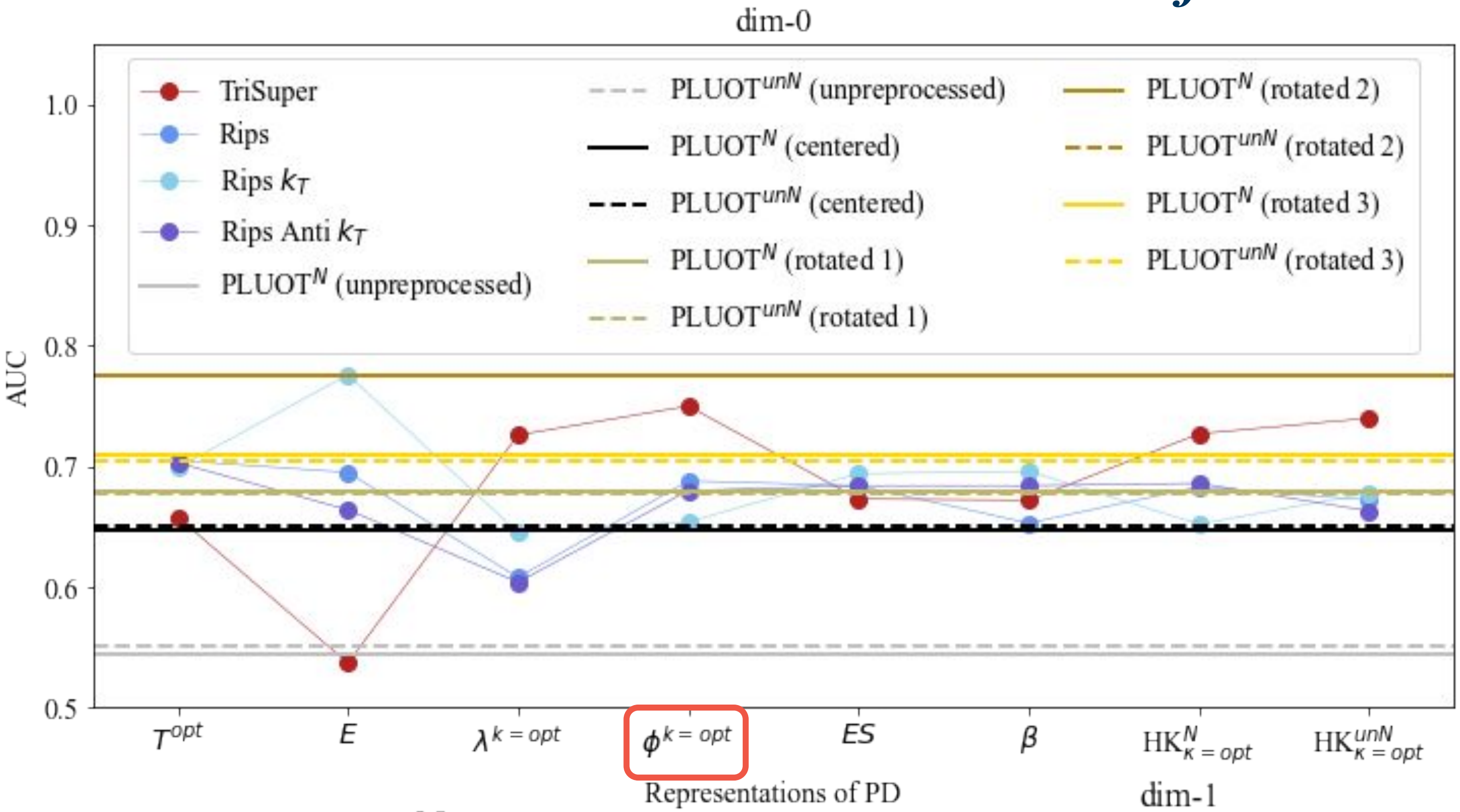
# 3. TDA for Jet Tagging: *Shrinking the Size of the Training Set*



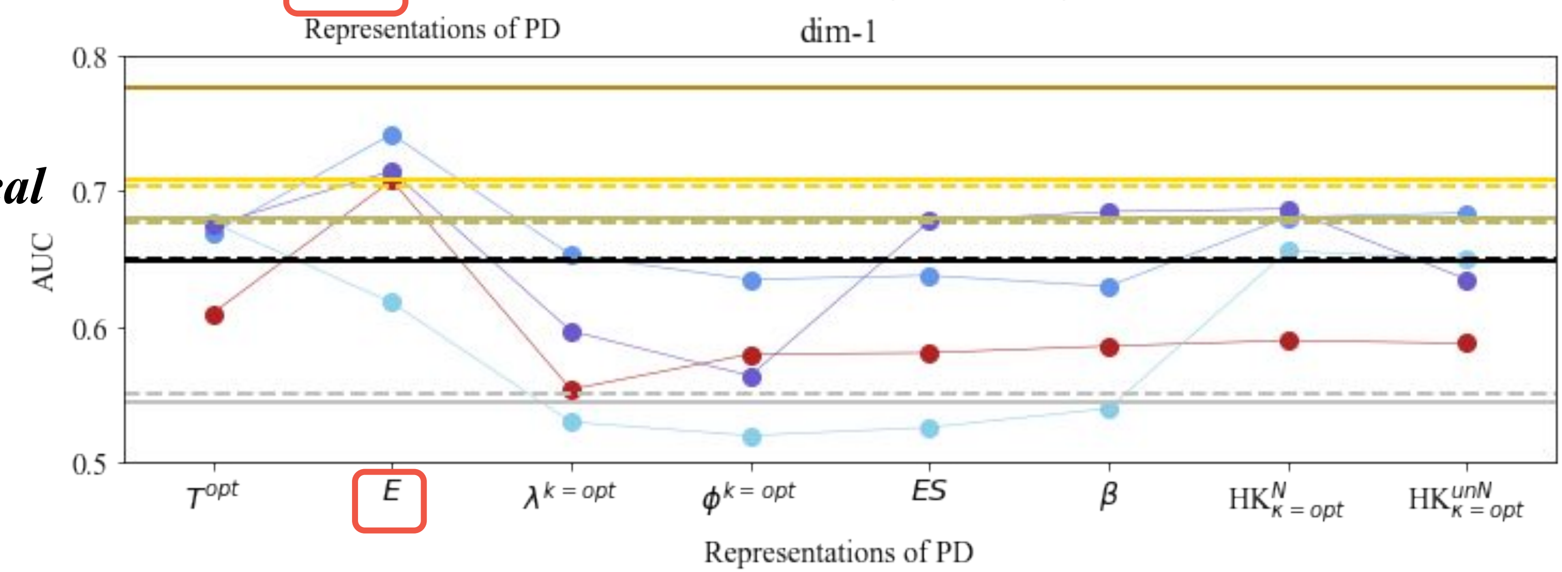
# 4. TDA for Event Classification:
## *Persistence Diagrams for 0th & 1st Homological Dimensions*



# 4. TDA for Event Classification: *Results for dim-0 & dim-1*




***0th Homological Dimension:*** *Connected components*




***1st Homological Dimension:*** *Holes*

# 5. Summary: *What have we done in this study?*

- **Motivation:** Get rid of the *ad-hoc* pre-processing for LHC jets and events.
- **Proposal:** Study the topology of jets/events (invariant under pre-processing) via the framework of Topological Data Analysis (TDA).
- **Tool:** Persistence homology to encode the evolution of topological features of certain filtration of a point cloud in persistence diagrams (PD).
    - **Topological Features:** Connected components (0th dim) for jets; Connected components (0th dim) & holes (1st dim) for events.
    - **Filtrations:** TriSuper; Rips; Rips $k_T$; Rips Anti $k_T$.
- **Statistical Analysis:** Simple cuts or kNN on various PD representations.
    - Studied six existing PD repres.
    - Proposed a new way to metricize the space of PDs via unbalanced optimal transport (OT) and introduced linearized OT as a novel PD representation.
- **Results:** The TDA framework achieve comparable or even better tagging performance than geometry-based approaches without the need of pre-processing.
    - A simple cut on a scalar PD repre for certain filtration performs surprisingly well.
    - Certain filtrations are more stable than others across different choices of PD repres.
    - The performance of the new HK repre is not so impressive.