Uncovering Hidden New Physics Patterns at High-Energy Colliders

Cambridge University
Cavendish-DAMPT seminar, Feb 25th 2021

Darius A. Faroughy

University of Zurich UZH

Overview

- Motivation
- Build step by step a probabilistic model for event data
- BSM jet physics application

Based on:

1904.04200
2005.12319

- Jernej F. Kamenik
- Barry Dillon
- Manuel Swezc

Introduction

- Since 2012, the SM has been experimentally verified.
- Strong motivations for physics beyond the SM:

Insert here favorite motivations for BSM_

- Many BSM theories address some of these problems:

Insert here favorite BSM theories_

- High-energy hadron colliders like the LHC play a fundamental role in BSM searching.

So far null results! Why?

https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ExoticsPublicResults

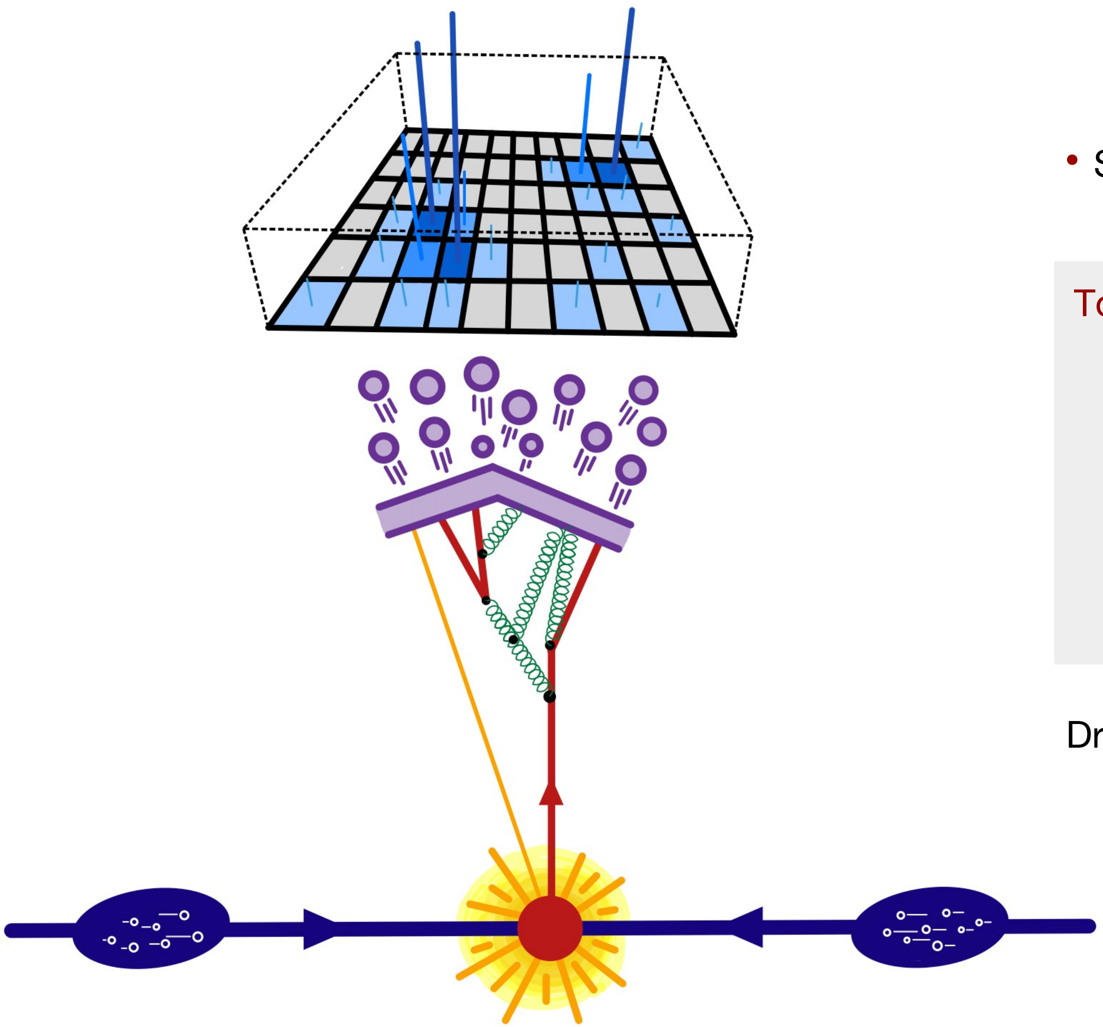
Exotics Physics Searches

Contact: ATLAS Exotics Working Group Conveners

ATLAS EXPERIMENT

This page contains public results from the ATLAS Exotics Working Group, which is searching for physics beyond the Standard Model with a signature-based program. Our aim is to cover all experimentally viable signatures focusing on non-supersymmetric models from Extra Dimensions and mini Black Holes to Dark Matter, extended Higgs models, and Compositeness to name a few.

Signature-Based approach



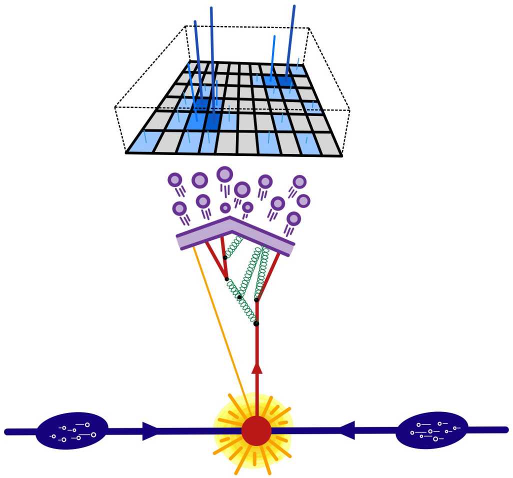
- Signature-based approach:

Top-down

- Propose Lagrangian
- Build dedicated physically motived observables
- propose signature
- implement search

Driven completely by our theoretical biases...

Complementary approach



- Data-based approach:

Bottom-up

- Select a data representation
- Model the data
- Train on data
- Build event classifiers
- Extract signature
- Characterize BSM signal

- Relies completely on our ability to model collider data — Collider data is very complex!
- Advances in Unsupervised Machine Learning (ML) offer an opportunity to pursue this approach

- Auto-encoders (AE) — Farina et al (2018), Roy et al (2019)
- Variational AEs — Cerri et al (2018)
- CWOLA — Metodiev et al (2018), Collins et al (2019), Amram et al (2020)
- Demix — Metodiev, Thaler (2018), Komiske et al (2019), Alvarez et al (2019)
- JUNIPR — Andreassen et al (2018, 2019)
- ...

Unsupervised ML
Semi-supervised ML

Take-away messages of this talk:

- It is possible to write down <u>simple</u> statistical models for generic collider events, useful for unsupervised event classification tasks.

Latent Dirichlet Allocation (Bayesian Probabilistic Generative Model)

- Use these models to discover resonances in jet substructure! $t\bar{t}$ W'

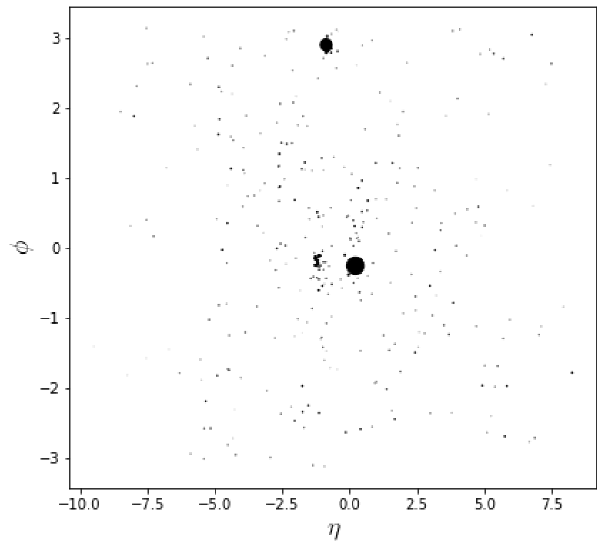
Data representation for events

- At the lowest level a collider event is a:

  A collection of reconstructed four-momenta of the visible
  f nal states from the scattering process.

$$e = \{p_1, p_2, \cdots, p_n\} \qquad \begin{cases} p_1 = (\eta, \phi, p_T)_1 \\ p_2 = (\eta, \phi, p_T)_2 \\ \vdots \\ p_n = (\eta, \phi, p_T)_n \end{cases}$$

$n \sim \mathcal{O}(10^2 - 10^3)$ High-dimensional phase space

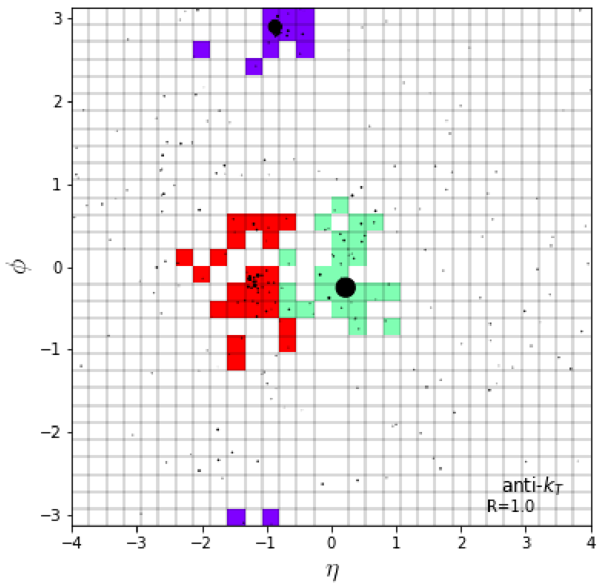


- High-level representations:
  - clustering
  - applying cuts
  - build physically motivated observables
  - ...

e.g.

$$\begin{cases} j_1 = (\eta, \phi, p_T)_1 \\ j_2 = (\eta, \phi, p_T)_2 \\ j_3 = (\eta, \phi, p_T)_3 \end{cases} \implies e = \{m_{12}^2\}$$

Low-dimensional phase space



Event data as random point patterns

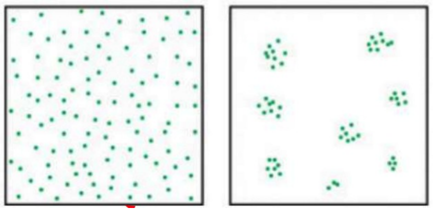
- Event: sequence of 'measurements' living in some vector space of observables.

$$e = \{o_1, o_2, \cdots, o_n\} \qquad o_i \in \mathcal{O} \subset \mathbb{R}^k$$

- Distribution of points: $e(o) = \sum_{i=1}^{n} \delta^{(k)}(o - o_i)$ $\quad n$ is a random variable

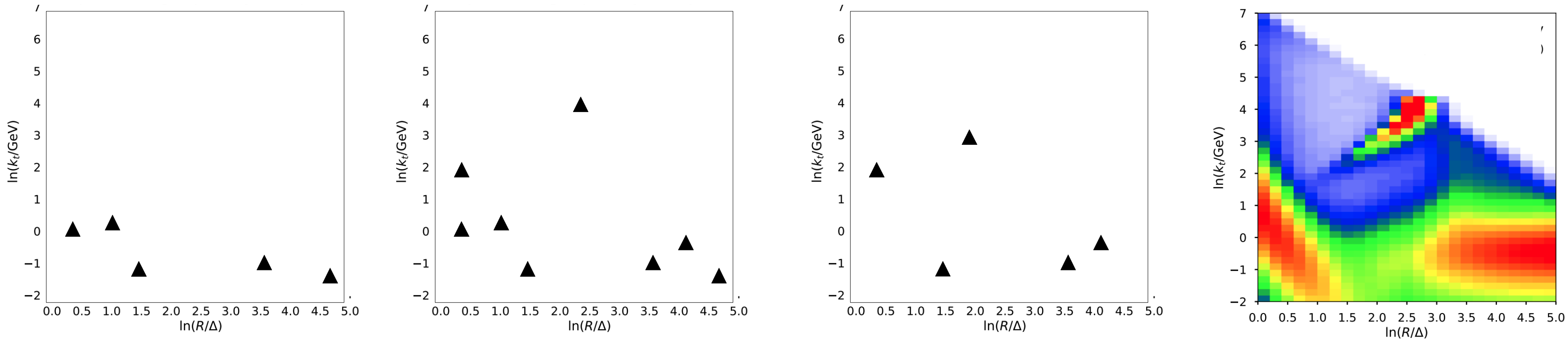
- Suggests that individual events are realizations of a stochastic point process in $\mathcal{O}$

Poisson process



- Events are typically sparse and irregular point patterns:

example: Lund jet plane



Probabilistic models for events

- What is the joint probability of an ensemble of collider events?

$$\mathcal{D} = \{e_1, \cdots, e_N\} \qquad \mathcal{P}(\mathcal{D}|\alpha) = \prod_{j=1}^{N} p(e_j|\alpha)$$

- What is the joint probability of a single collider event?

$$\mathcal{P}(e|\alpha) = \mathcal{P}(\{o_1, \cdots, o_n\}|\alpha)$$

How can we model this probability in a simple, yet, useful way?

Goal: event classification (not event generation!)

- We impose three model-buiding assumptions for the event probability:

(1) Exchangeability of measurements.

(2) Discretization of the observable space.

(3) Multiple *latent* categories contribute to the event-generating process.

Assumptions are *data-independent*

1) Exchangeability

- Exchangeability of event measurements (i.e. Permutation symmetry)

$$\mathcal{P}(e) = \mathcal{P}(\{o_1, o_2, o_3, \cdots\}) = P(\{o_{\pi(1)}, o_{\pi(2)}, o_{\pi(3)}, \cdots\})$$

$\pi \in \mathcal{S}$
(permutation group)

De Finnetti's representation theorem (1931):

A sequence of measurements is exchangeable if and only if there exists a *latent variable* ω and two distributions *p* and *P* such that

$$\mathcal{P}(e) = \int_\Omega \mathrm{d}\omega \, P(\omega) \prod_{i=1}^{n} p(o_i|\omega)$$

Latent space "Prior" "Likelihood"

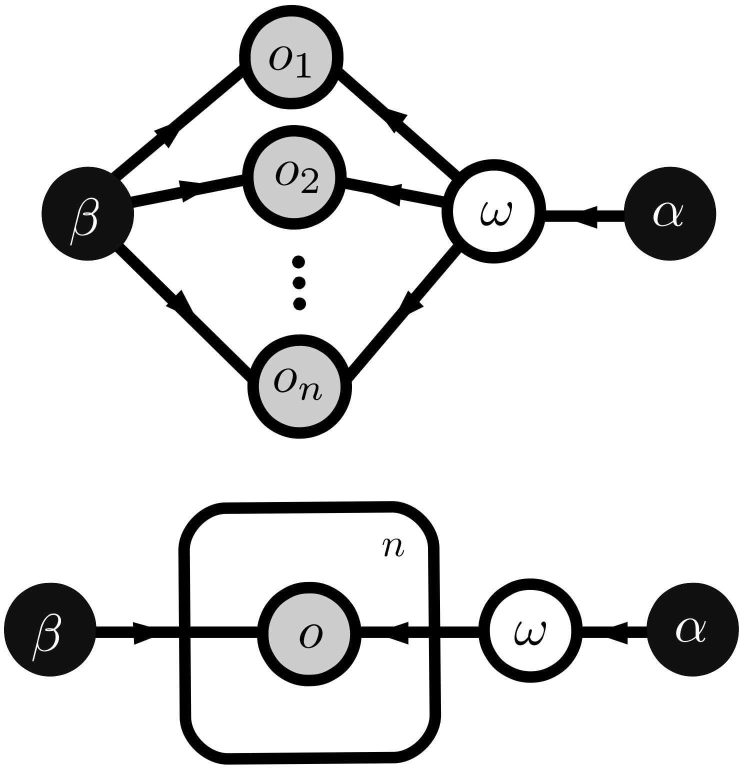
Justifies Bayesian methods!

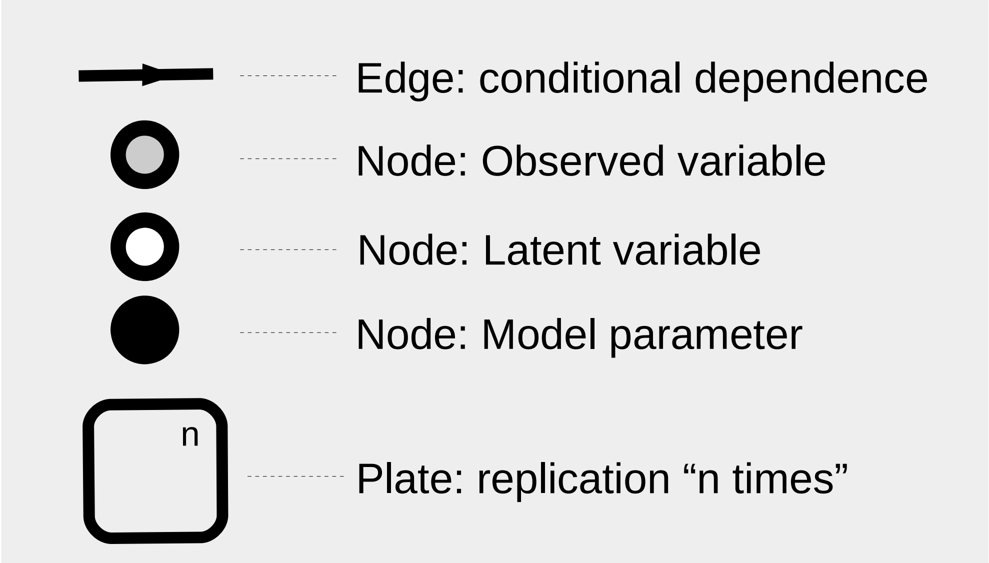
- Measurements are considered conditionally independent given a latent variable $\omega \in \Omega$
- Exchangeable not to be confused with independent and identically distributed (iid) !! $\mathcal{P}(e) = \prod_{i=1}^{n} \mathcal{P}(o_i)$
- We will need extra model-building assumptions to fix *p*, *P*, *omega*

$$\mathcal{P}(e|\alpha,\beta) = \int_\Omega \mathrm{d}\omega \, P(\omega|\alpha) \prod_{i=1}^{n} p(o_i|\omega,\beta)$$

Free parameters

- Graph models:





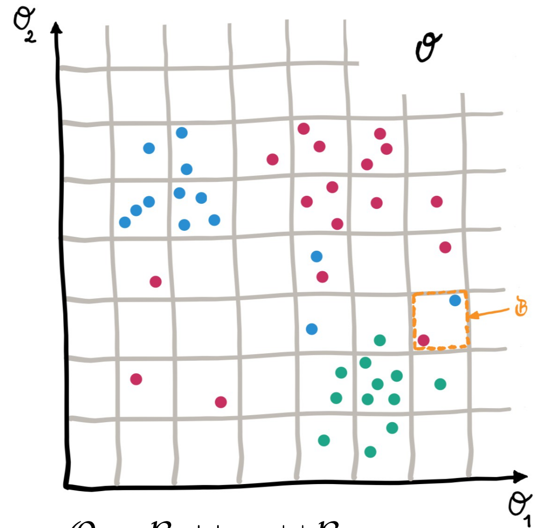
2) Event data discretization

- $\mathcal{P}(e|\alpha,\beta) = \int_\Omega \mathrm{d}\omega\, P(\omega|\alpha) \prod_{i=1}^{n} \boxed{p(o_i|\omega,\beta)}$

What to take for $p(o|\omega,\beta)$?

- *Binned* measurements:

$$o \sim \mathrm{Multinomial}(\beta) \quad \begin{cases} \beta = (\beta_1,\ldots,\beta_M) & M \text{ bins} \\ \sum_{m=1}^{M} \beta_m = 1 & \\ 0 \leq \beta_m \leq 1 & \end{cases}$$



$\mathcal{O} = \mathcal{B}_1 \cup \ldots \cup \mathcal{B}_M$

- Multinomial from Poisson process in $\mathcal{O}$:

$$\begin{cases} \text{Counts per-bin:} \quad N(\mathcal{B}) \equiv \#\{o \in \mathcal{B}\} \quad \Longleftarrow \quad N(\mathcal{B}) \sim \mathrm{Poisson}(\lambda_\mathcal{B})\,, \quad \lambda_\mathcal{B} = \int_\mathcal{B} \prod^{k} \mathrm{d}\mathcal{O}\; \mu(\mathcal{O}_1,\ldots,\mathcal{O}_k) \\ \text{Total Count:} \quad N = \sum_\mathcal{B} N(\mathcal{B}) \quad \Longleftarrow \quad N \sim \mathrm{Poisson}(\lambda)\,, \quad \lambda = \sum_\mathcal{B} \lambda_\mathcal{B} \end{cases}$$

Non-homogenous intensity function

$$\boxed{P\left(N(\mathcal{B}_1),\cdots,N(\mathcal{B}_M)\,|N\right) = \prod_\mathcal{B} \frac{\mathrm{Poisson}(\lambda_\mathcal{B})}{\mathrm{Poisson}(\lambda)} = \frac{N!}{N(\mathcal{B}_1)!\cdots N(\mathcal{B}_M)!} \prod_{m=1}^{M} \left(\frac{\lambda_m}{\lambda}\right)^{N(\mathcal{B}_m)}}$$

$\beta_m \equiv \lambda_m/\lambda$

Multinomial Distribution!

3) Multiple Latent Categories

- $\mathcal{P}(e|\alpha,\beta) = \int_\Omega \mathrm{d}\omega\, P(\boxed{\omega}|\alpha) \prod_{i=1}^{n} p(o_i|\boxed{\omega,}\beta)$

What to take for ther latent variable?

- Event measuremnts are generated from **multiple** latent Multinomial distributions over $\mathcal{O}$

$$p(o|\beta_t) \quad t = 1, \ldots, T$$

Mixture of multinomials: $$p(o|\omega,\beta) = \sum_{t=1}^{T} p(t|\omega)\, p(o|\beta_t)$$

“Themes” or “Topics”

Theme mixing parameter $\omega = (\omega_1, \ldots, \omega_t)$

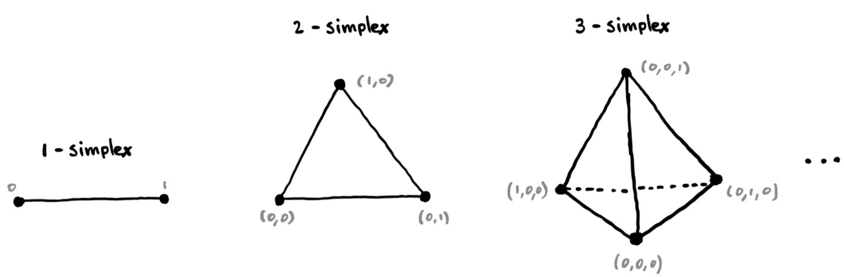
$$p(t|\omega) = \omega_t \qquad 0 \le \omega_t \le 1, \ \sum \omega_t = 1$$

- **Themes***: distributions encoding different physical contributions to a single event.

** Terminology from Natural Language processing (NLP)*

- $\mathcal{P}(e|\alpha,\beta) = \int_{\boxed{\Omega}} \mathrm{d}\omega\, P(\omega|\alpha) \prod_{i=1}^{n} p(o_i|\omega,\beta)$

Latent space: **(T-1)-simplex** (space of mixings)



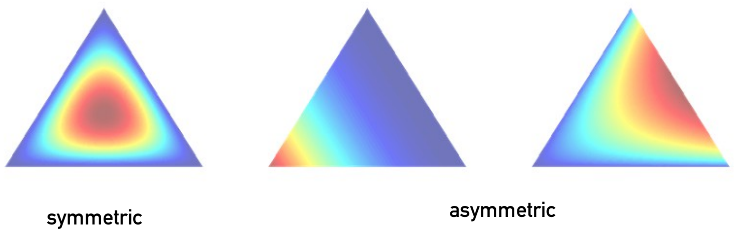
- $\mathcal{P}(e|\alpha,\beta) = \int_\Omega \mathrm{d}\omega\, \boxed{P(}\omega|\alpha) \prod_{i=1}^{n} p(o_i|\omega,\beta)$

What to take for the prior distribiton $P(\omega)$?

- Dirichlet distributions:

$$D(\omega|\alpha) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_T)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_T)} \prod_{t=1}^{T} \omega_t^{\alpha_t - 1}$$

$\alpha = (\alpha_1, \cdots, \alpha_T)$ shape parameter

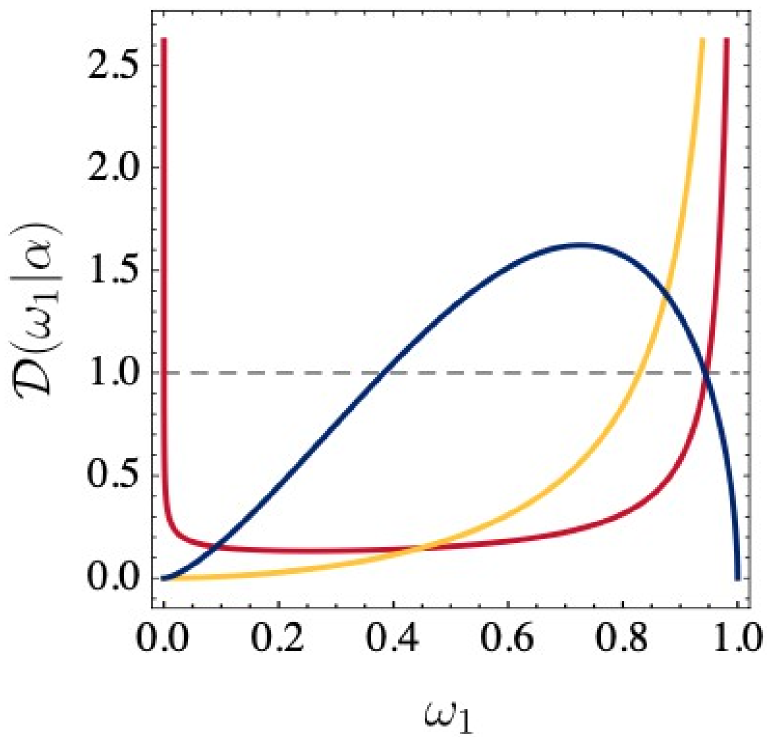


- Belongs to the exponential familiy and is **conjugate** to the multinomial.

- Two-theme model (T = 2) :

$D(\omega|\alpha_1, \alpha_2)$ is the Beta distribution over [0,1]

flat: $\alpha_1 = \alpha_2 = 1$
uni-modal bell-shape: $\alpha_1, \alpha_2 > 1$
uni-modal J-shape: $\alpha_1 > 1, \alpha_2 < 1$
bi-modal U-shape: $\alpha_1, \alpha_2 < 1$



Latent Dirichlet Allocation (LDA)

$$\mathcal{D} = \{e_1, \cdots, e_N\}$$

$$\mathcal{P}(\mathcal{D}|\alpha,\beta) = \prod_{j=1}^{N}\left[\int_{\Omega_{T-1}} \mathrm{d}\omega_j \, D(\omega_j|\alpha) \prod_{i=1}^{n_j}\left(\sum_{t=1}^{T} p(t|\omega_j)\; p(o_{ij}|\beta_t)\right)\right]$$

Event index

Simplex

Dirichlet Prior

Theme assignment variable

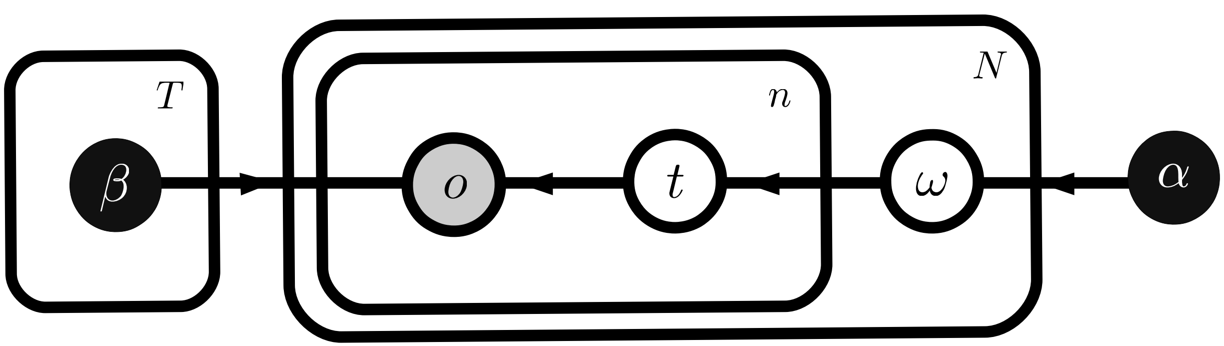
Themes (Multinomials)

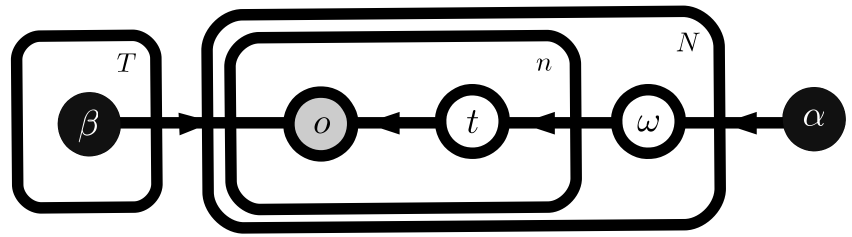
Theme mixing

- LDA is a mixed-membership model.

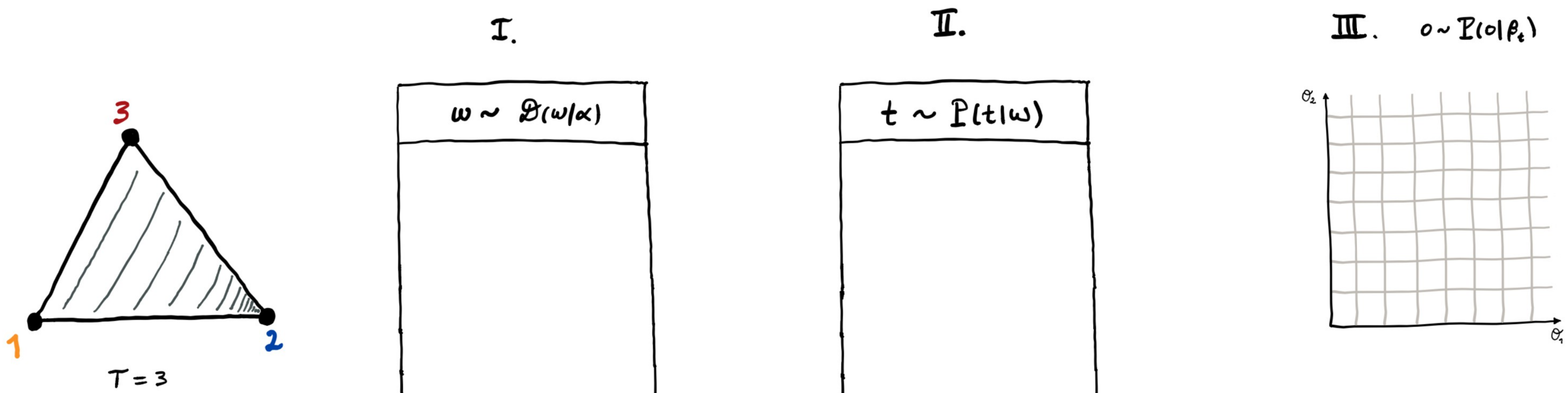
Individual events are described by mixture of multiple themes:

- LDA graphical model:

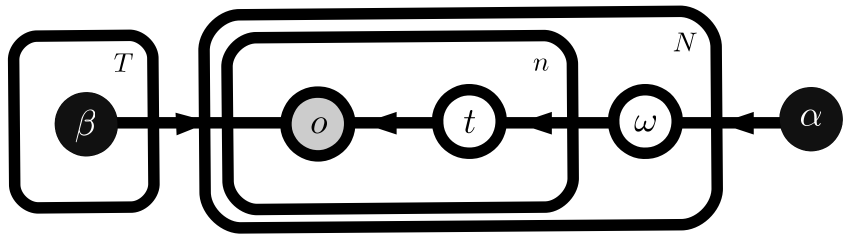




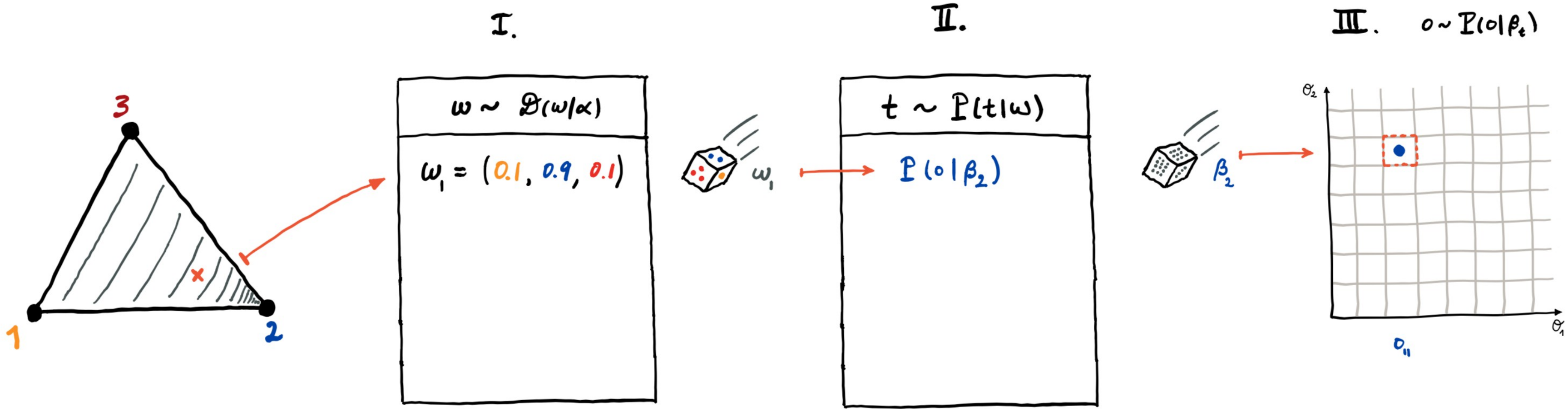
- Generative process for a 3-theme LDA model:

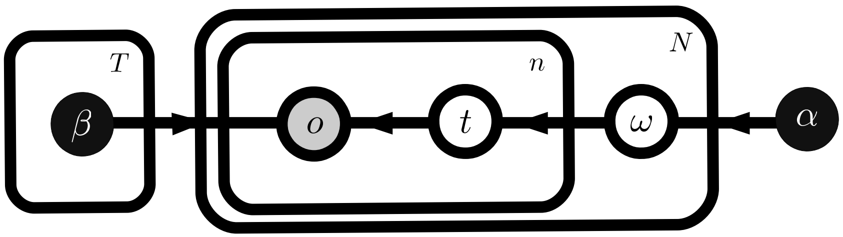


$D(\omega|\alpha_0, \alpha_1, \alpha_2)$

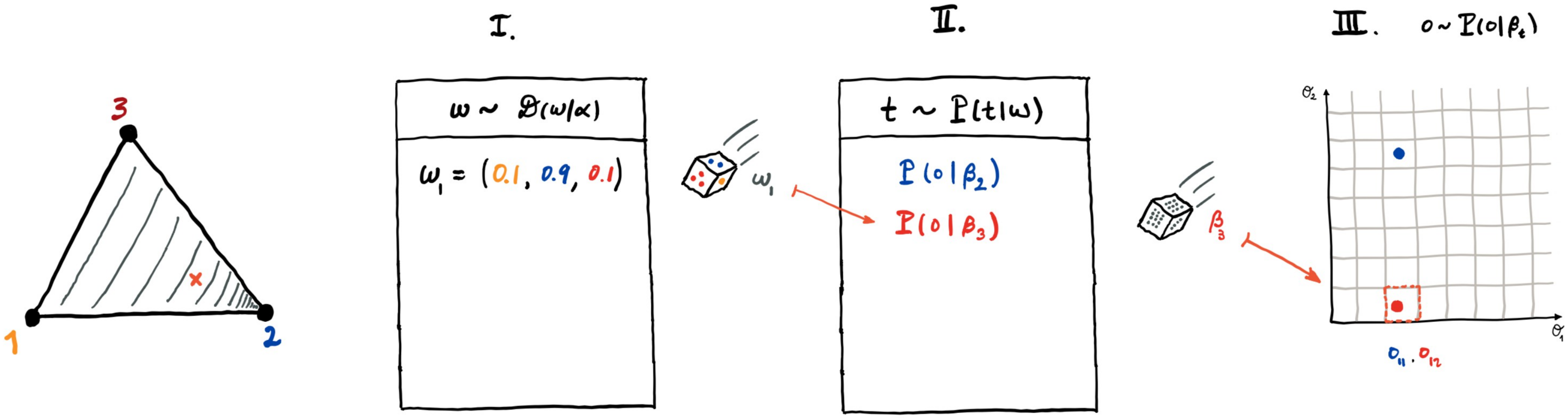


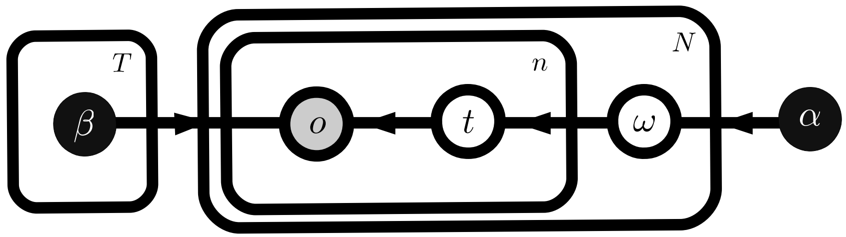
- Generative process for a 3-theme LDA model:



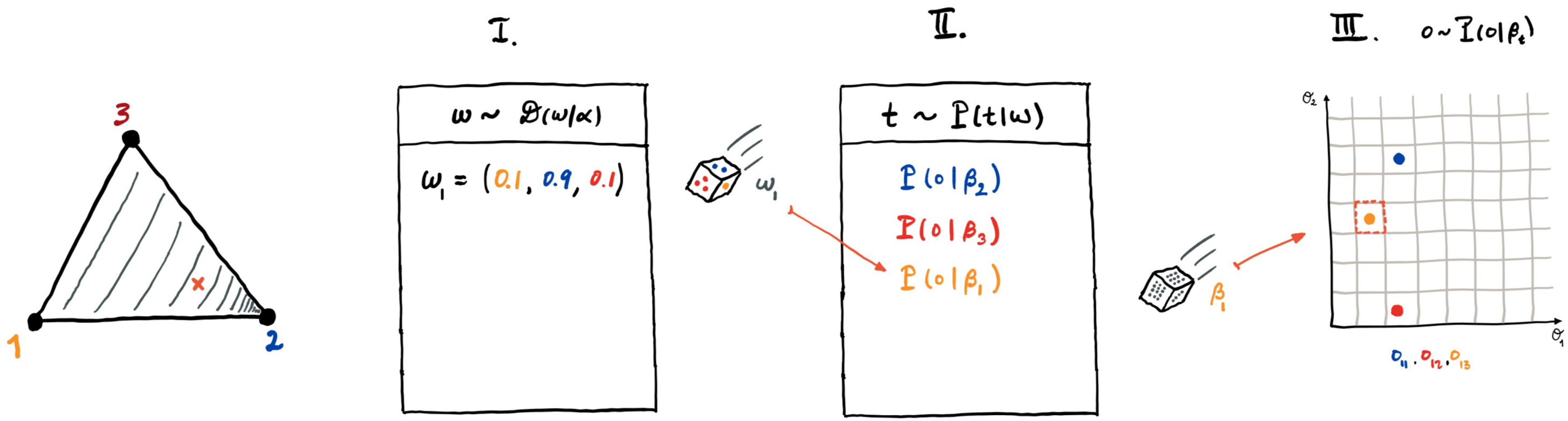


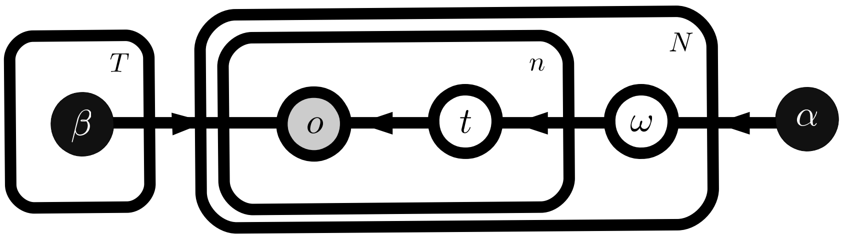
- Generative process for a 3-theme LDA model:



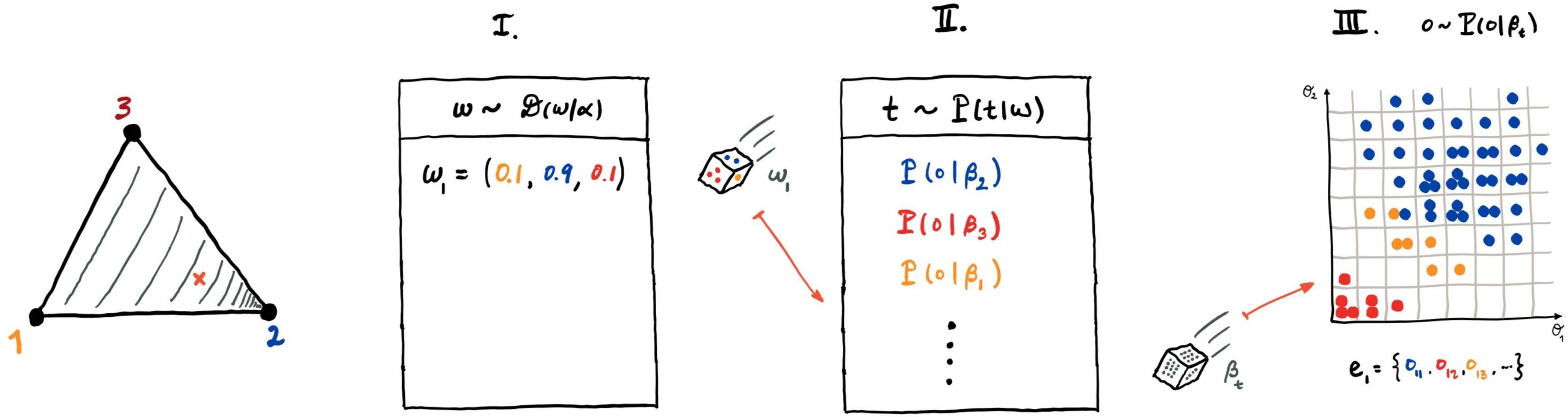


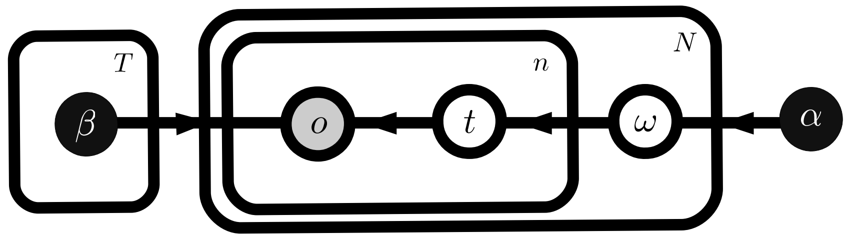
- Generative process for a 3-theme LDA model:



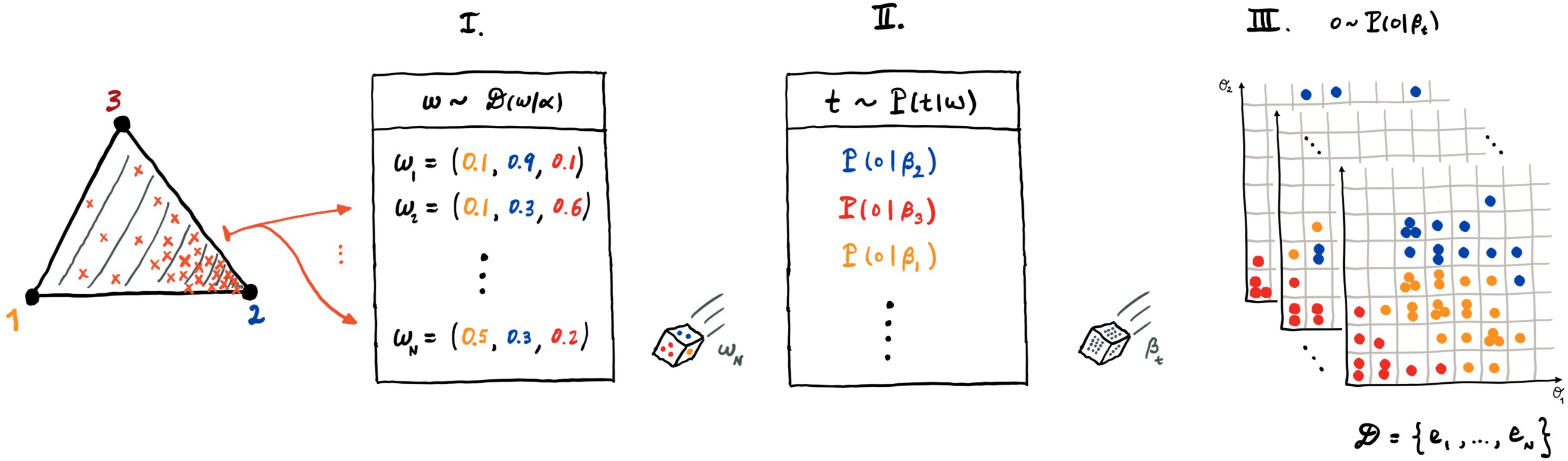


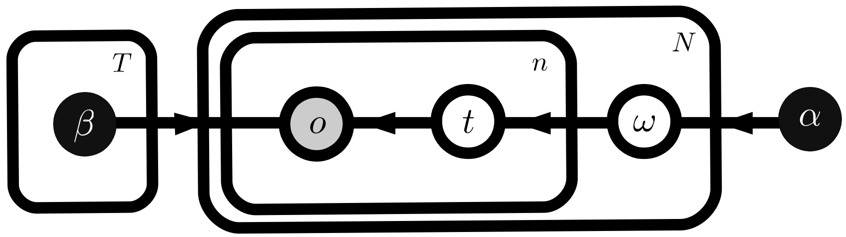
- Generative process for a 3-theme LDA model:



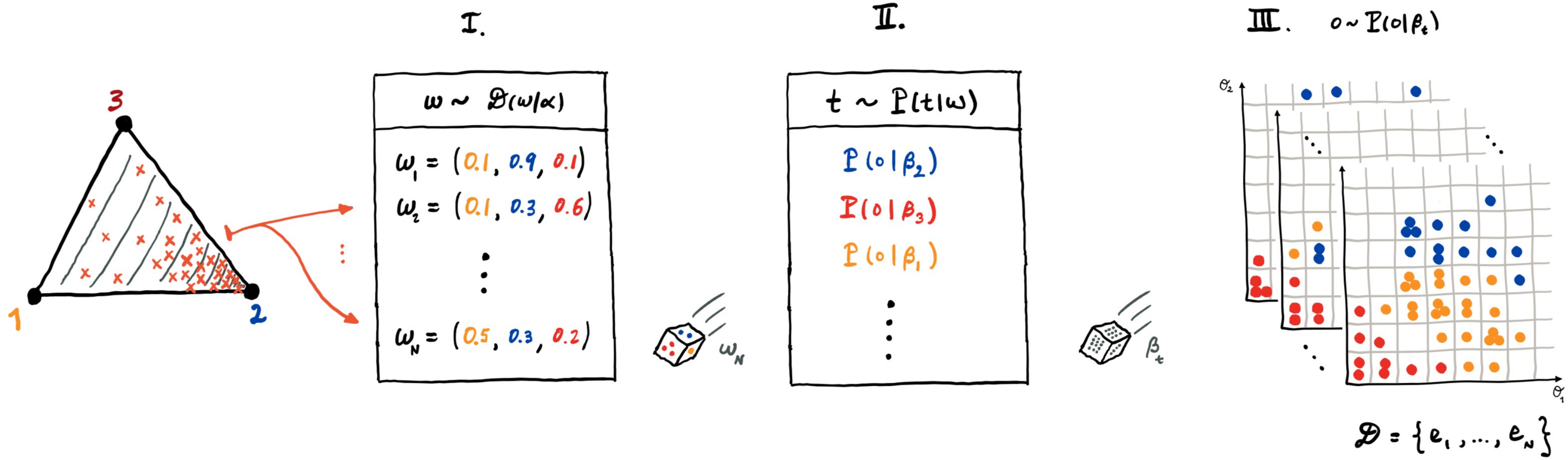


- Generative process for a 3-theme LDA model:





- Generative process for a 3-theme LDA model:



- Mixed-Membership Models not to be confused with Mixture models!

Allows for shared features between events!

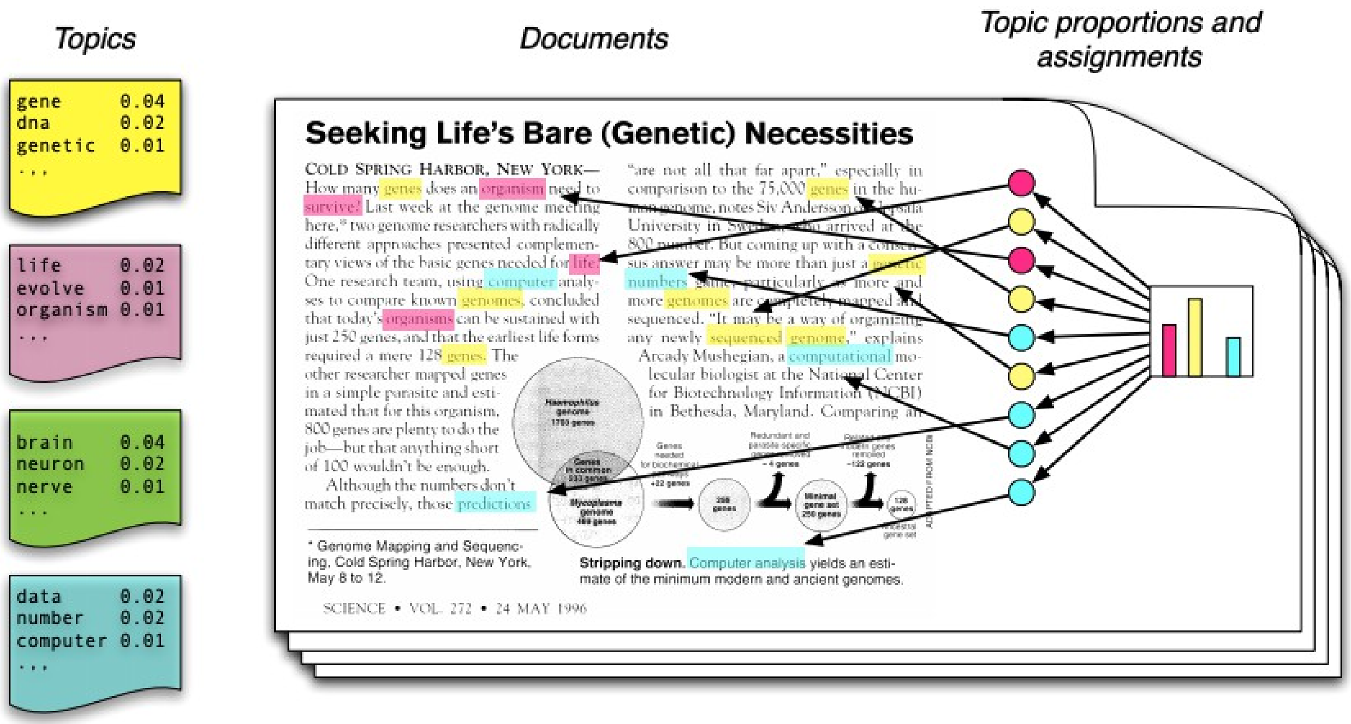
All measurments in an event come from only <u>one</u> theme...

Topic Models for texts

- LDA conceived for Natural Language Processing

Blei, Ng, Jordan,
Journal of Machine Learning
Research, 3 (2003) 993-1022.

over 30K citations!



Fully Unsupervised ML

- LDA uncovers the hidden topics in a collection of documents

- Documents: unstructured collection of words

  Bag-of-words (Bow)

Topics: distributions over vocabulary

- Text / Collider Physics correspondance:

  corpus -------------- event samples
  document ---------- event
  vocabulary --------- space of observables
  word ----------------- bin
  topic ----------------- histogram

Learning the latent variables

- The posterior for an event:

$$p(\omega, t, \beta | e, \alpha, \eta) = \frac{p(\omega, t, \beta, e | \alpha, \eta)}{p(e|\alpha,\eta)}$$

$$D(\omega|\alpha)\left(\prod_{t=1}^{T} D(\beta_t|\eta)\right)\left(\prod_{j=1}^{N}\prod_{i=1}^{n} p(t_i|\omega)\, p(o_i|t_i,\beta)\right)$$

$$\sum_t \int \mathrm{d}\omega \mathrm{d}\beta\, p(\omega, t, \beta, e|\alpha,\eta)$$ "evidence"

Intractable integral!

- Variational inference: inference problem $\longrightarrow$ optimization problem

Propose a simple family of distributions $\mathcal{Q}$

$$q^* = \underset{q\,\in\,\mathcal{Q}}{\mathrm{argmin}}\, d_{\mathrm{KL}}[q, p]$$

posterior

Kullback-Liebler divergence $d_{\mathrm{KL}}[q,p] = \langle \log q \rangle - \langle \log p \rangle + \underbrace{\log p(e)}$

Log-evidence..... still intractable

Instead we maximize evidence lower-bound (ELBO):

$$q^* = \underset{q\,\in\,\mathcal{Q}}{\mathrm{argmax}}\, \mathcal{L}[q]$$

$$\mathcal{L}[q] := \langle \log p \rangle - \langle \log q \rangle$$

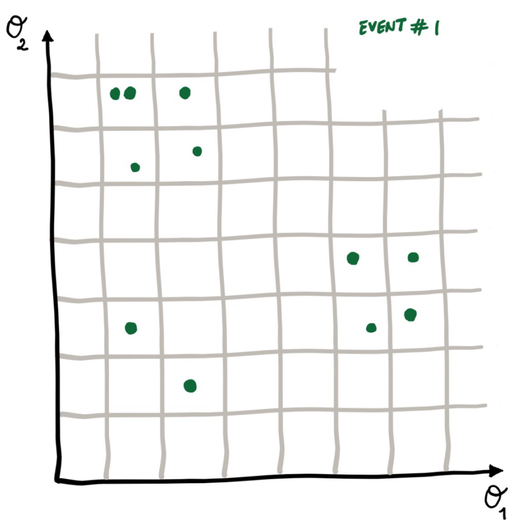
$$\log p(e) = d_{\mathrm{KL}}[q,p] + \mathcal{L}[q] \quad \Longrightarrow \quad \log p(e) \geq \mathcal{L}[q]$$

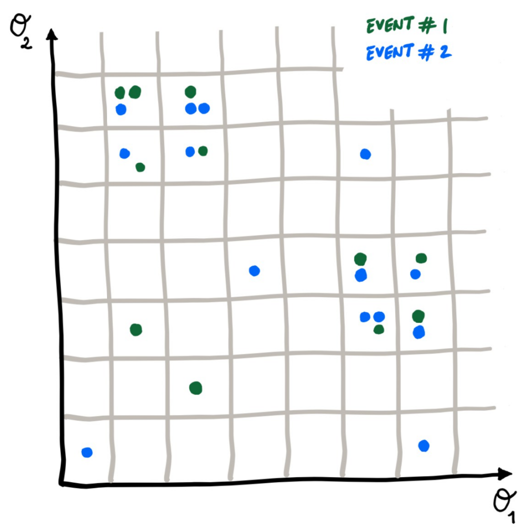
Co-ocurrences

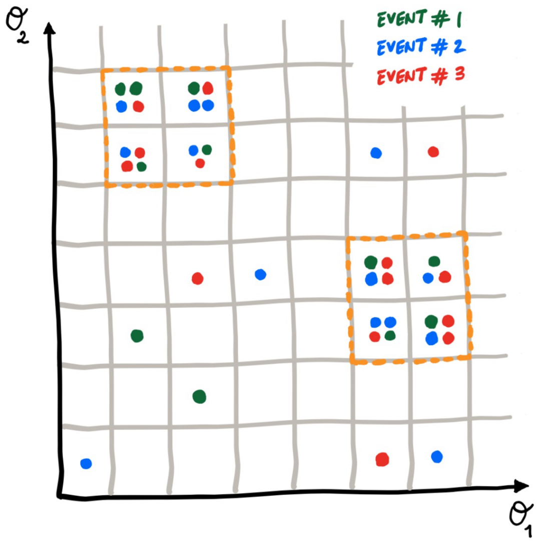
- What does LDA learn?

- LDA learns by identifying recurring measurement patterns

Captures the statistical dependencies between event measurements in the event ensemble







Finds Co-ocurrences between event measurement throughout the event sample.

(LDA clusters in the same themes measurments that tend to co-occur together)

Two-theme LDA classifiers

- For most applications we wish to classify events into two categories

  We focus on Two-theme LDA models $T = 2$

- This gives rise to two possiblel binary classif iers:

1) Likelihood-ratio of themes:

sliding threshold

$$L(e|\alpha) := \prod_{o \in e} \frac{p(o\,|\beta_2)}{p(o\,|\beta_1)}$$

$$\begin{cases} L(e|\alpha) > c & \Rightarrow \quad e \in \mathcal{C}_1 \\ L(e|\alpha) \leq c & \Rightarrow \quad e \in \mathcal{C}_2 \end{cases}$$

2) 'Cluster' assignment:

$$w(e|\alpha) := \omega(\alpha)|_e$$

Probability of event belonging to f irst cluster (theme)

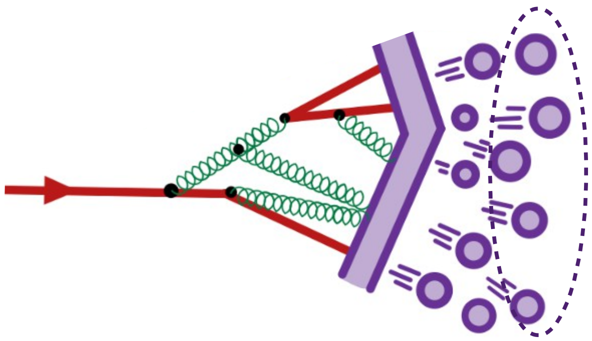
$$\begin{cases} w(e|\alpha) > c & \Rightarrow \quad e \in \mathcal{C}_1 \\ w(e|\alpha) \leq c & \Rightarrow \quad e \in \mathcal{C}_2 \end{cases}$$

LDA can be interpreted as a fuzy clustering algorithm

Both classif iers yield similar performances

Application to jet substructure

Jets



Jets are collimated spray of hadrons

- Jet clustering: sequential recombination schemes

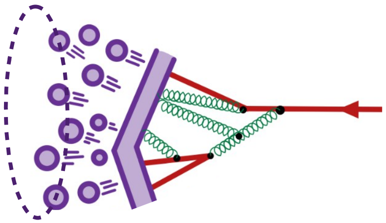
$$d_{ij} = \min\left\{{p_T}_i^{2\alpha}, {p_T}_j^{2\alpha}\right\}\left(\delta_{ij} + \frac{\Delta R_{ij}^2}{R^2}\right)$$

$$\begin{cases} \alpha = -1 & \text{anti-kT} \\ \alpha = 0 & \text{Cambridge/Aachen (CA)} \\ \alpha = +1 & \text{kT} \\ R = \mathcal{O}(1) & \text{jet cone radius} \end{cases}$$

jet merging criteria: $d_{ii} \geq d_{ij} \quad \Longrightarrow \quad i \cup j \to k \quad p_k^\mu = p_i^\mu + p_j^\mu$

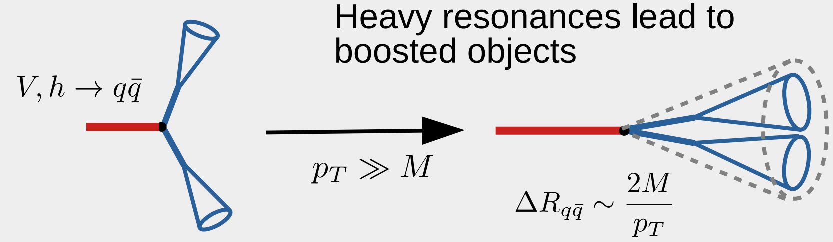
The jet classification problem

Jet



??? How to infer the underlying particle(s) responsible for the jet?

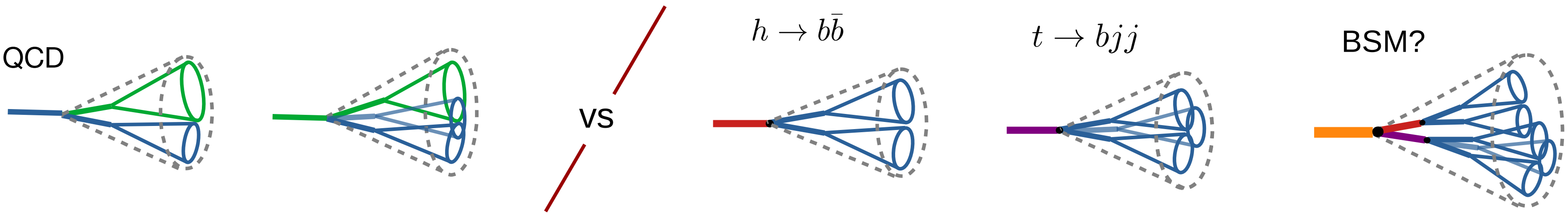
1-prong jets



2 resolved jets

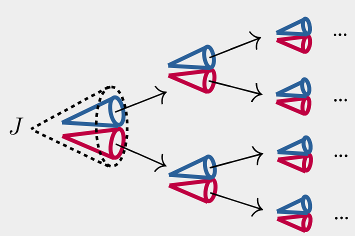
One 'fat jet' (2-prongs)

- Can LDA identify boosted decays?



- Jet substructure observables that resolve the inner structure of (fat) jets:

Jet declustering



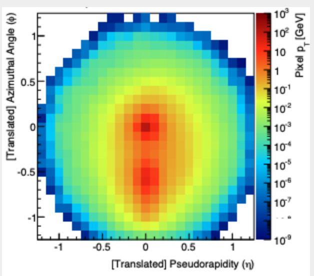
Jet shapes

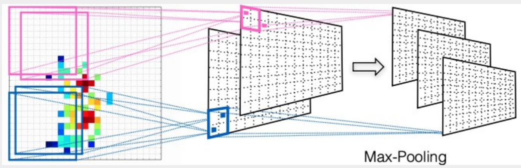
Angularities

N-subjettiness

Energy correlation functions...

Jet Images (Deep learning)

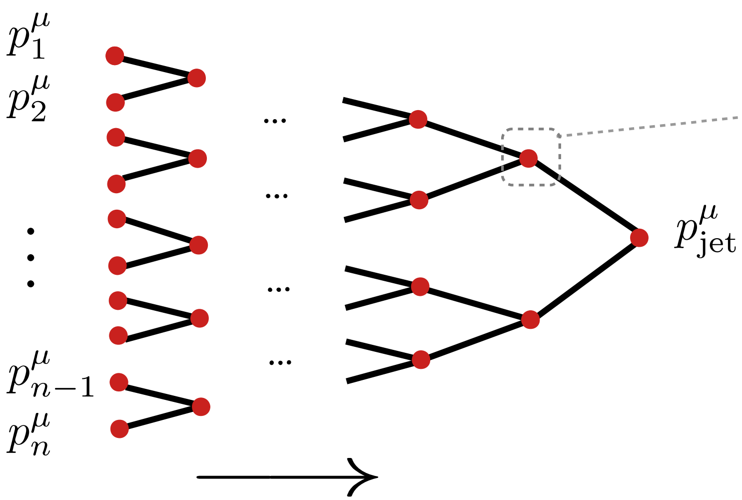




Cogan, Kagan, Strauss, Schwartzman 2015

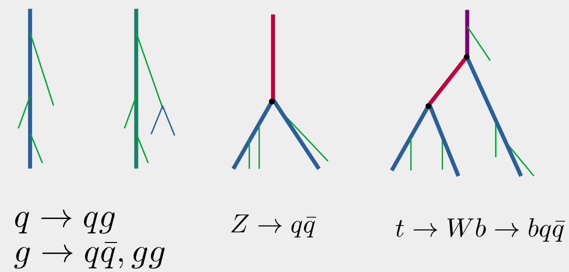
Jet clustering history

- Jet clustering hierarchy is sensitive to the underlying physics.
- Jet binary tree: proxy for the radiation pattern during jet formation.

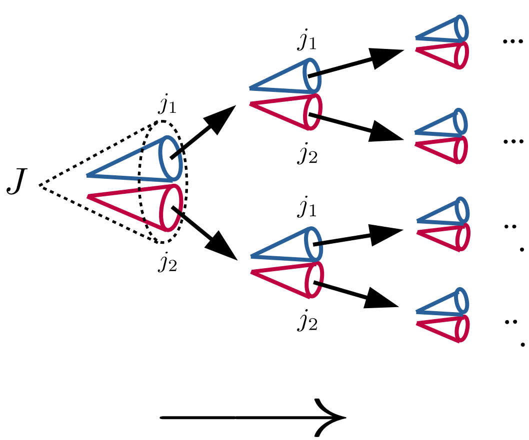


clustering (C/A) angular-ordered clustering tree.

Proxy for $1 \to 2$ splitting or massive decay



- Jet declustering:



declustering

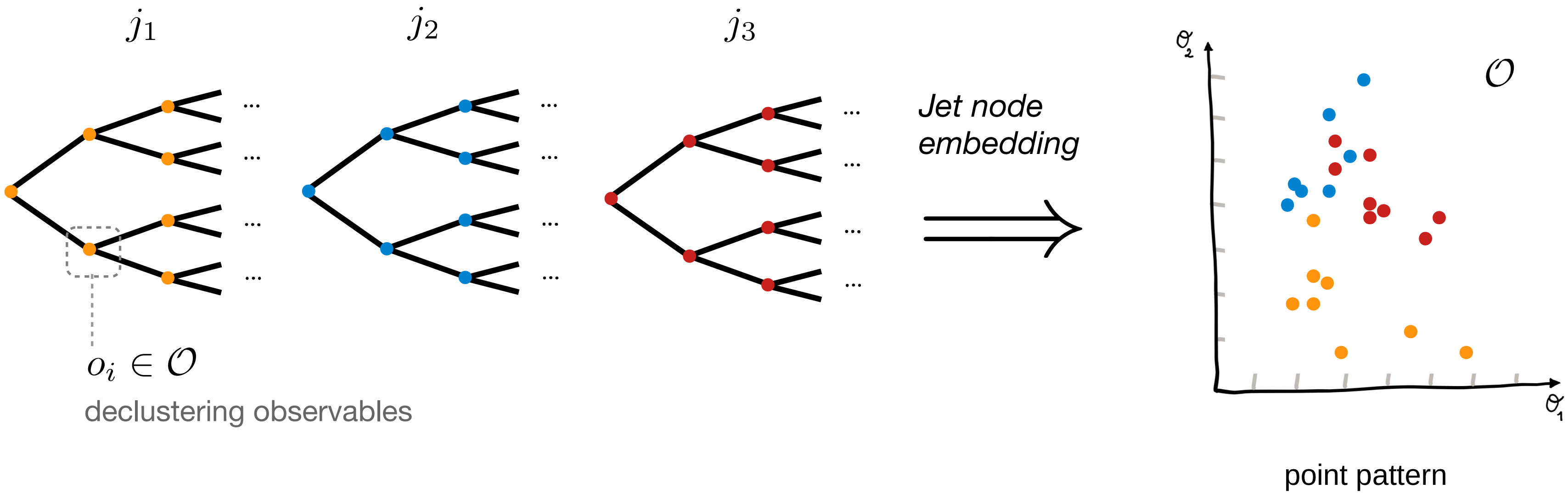
- Jet tagging/grooming:

Decluster jet iteratively following hardest branch until some “hard/collinear” branching condition is identified...

Mass-drop tagger & mMDT	Butterworth, Davison, Rubin, Salam 2008 Dasgupta, Fregoso, Marzani, Salam 2018
HEP & JH Top taggers	Kaplan, Rehermann, Schwartz & Tweedie 2008
Soft-drop tagger/groomer	Larkoski, Marzani, Soyez, Thaler 2014 Dreyer, Necib, Soyez, Thaler 2018

Simpler data representation for jets

- Ordering in jet declustering procedure is ignored!



- For full events, can include jet kinematical “labels” based on some jet ordering.

- De Finnetti represenatation of jet:

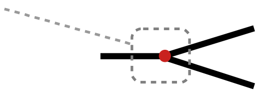
$$\mathcal{P}(\quad) \simeq \int_\Omega \mathrm{d}\omega\, \mathcal{P}(\omega) \prod_{\bullet \in j} \mathcal{P}(\bullet|\omega)$$

Justification: The 1->2 splitting pattern is dominated by QCD soft/collinear emissions, only a handful of splittings are relevant for identifying the underlying hard physics for jet/event classification

Text analogy: syntaxic structure of the document is removed when extracting the topics (bag-of words)

Jet declustering observables

$o_i \in \mathcal{O}$



$j_0 \to j_1\, j_2$

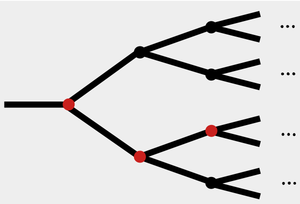
- Train LDA on full events with 2 types of substructure observables:

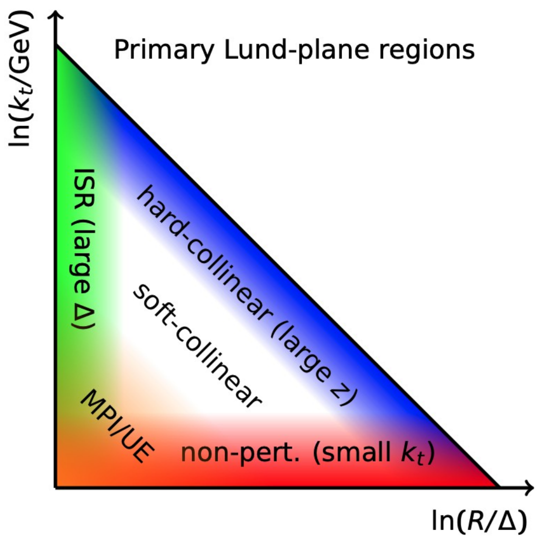
Mass observables: $\mathcal{O}_{\mathrm{Mass}} = \left\{ \ell, m_{j_0}, \dfrac{m_{j_1}}{m_{j_0}} \right\}$ $\quad m_{j_0} > 30\ \mathrm{GeV}$

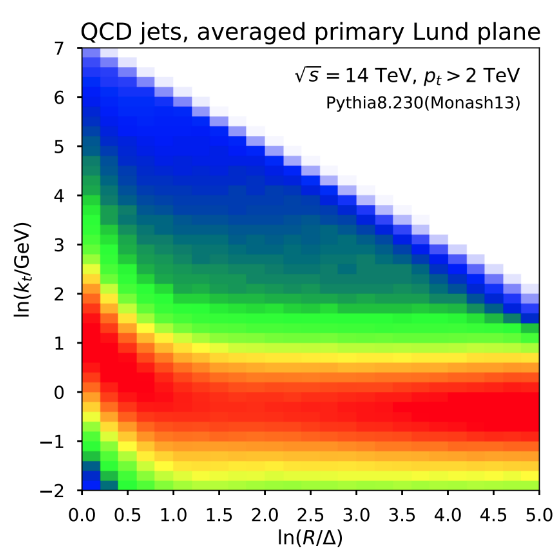
ℓ Label indicating to which jet the measurement belongs too, with jets ordered by mass.

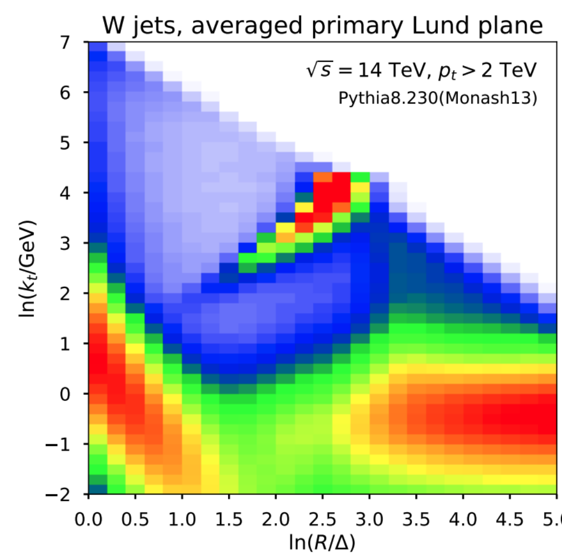
Lund observables: $\mathcal{O}_{\mathrm{Lund}} = \left\{ \ell, \log(k_t), \log\left(\dfrac{R}{\Delta R}\right) \right\}$

Primary Lund plane
Dreyer et al (2018)



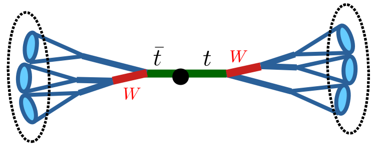


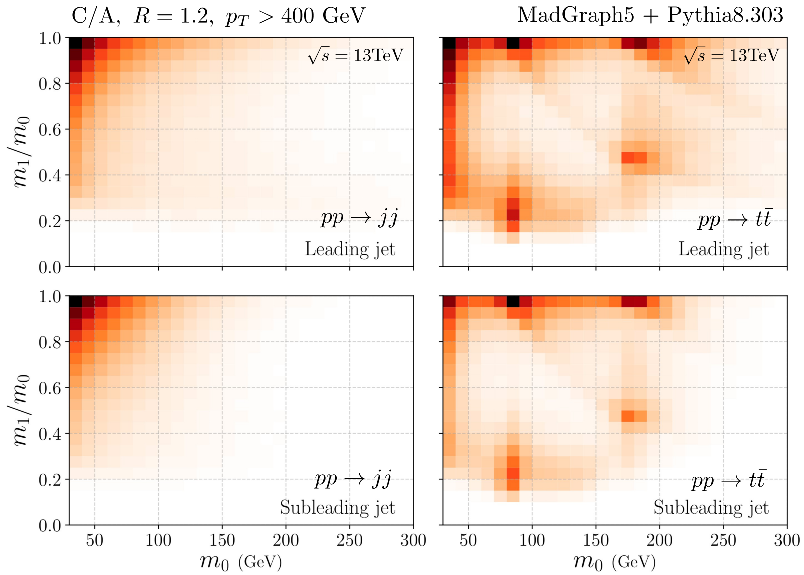




• Top-quarks vs QCD

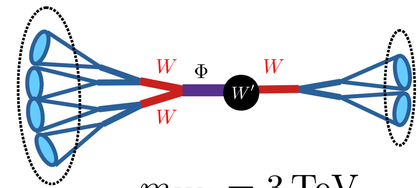
$$pp \to t\bar{t}$$



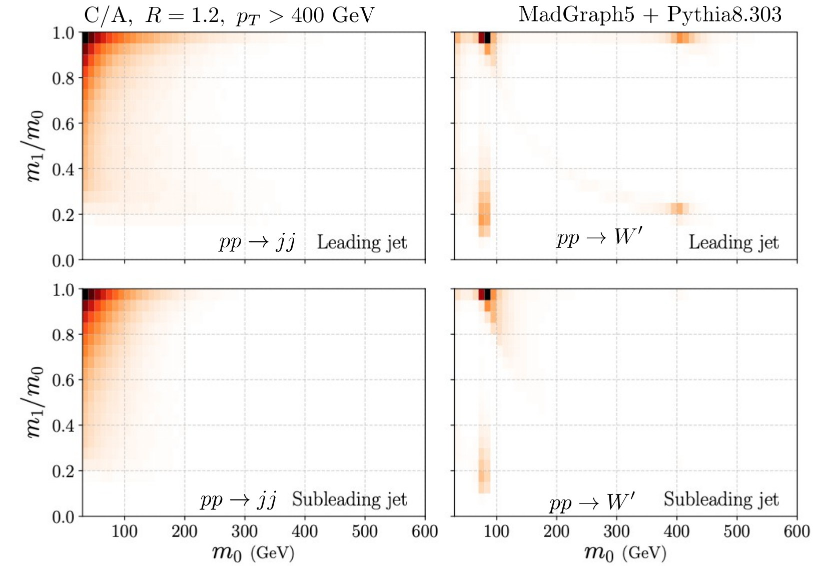


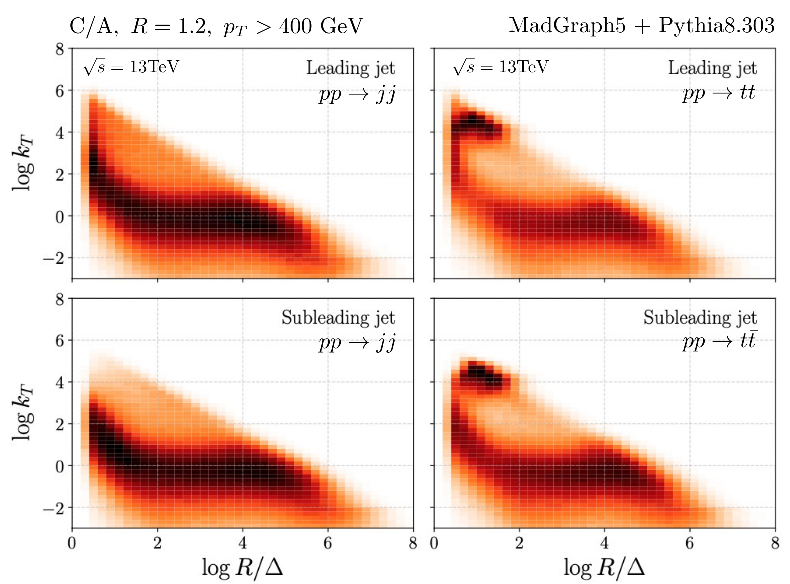
• BSM model:

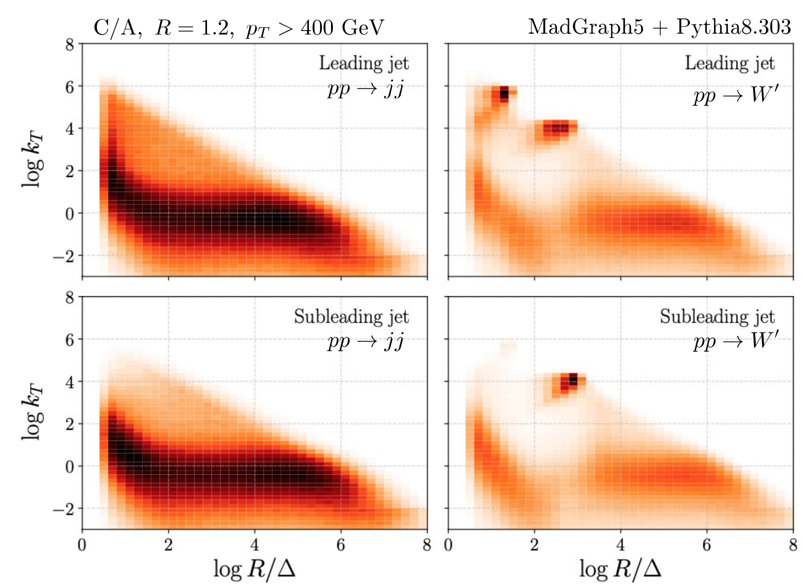
$$pp \to W' \to \Phi W^{\pm}$$
$$\Phi \to W^{\pm} W^{\mp}$$



$$m_{W'} = 3\,\mathrm{TeV}$$
$$m_{\Phi} = 400\,\mathrm{GeV}$$

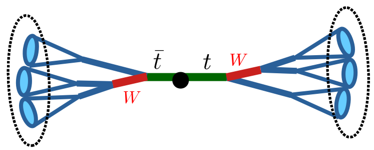






- Top-quarks

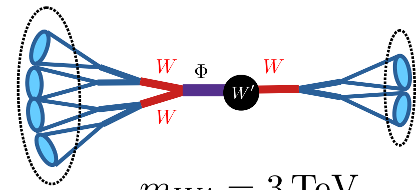
$pp \to t\bar{t}$



- BSM model:

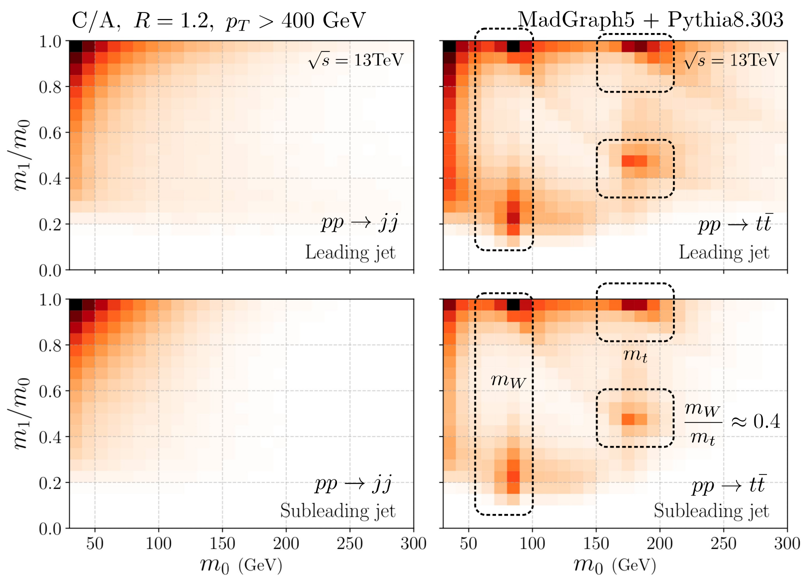
$pp \to W' \to \Phi W^{\pm}$

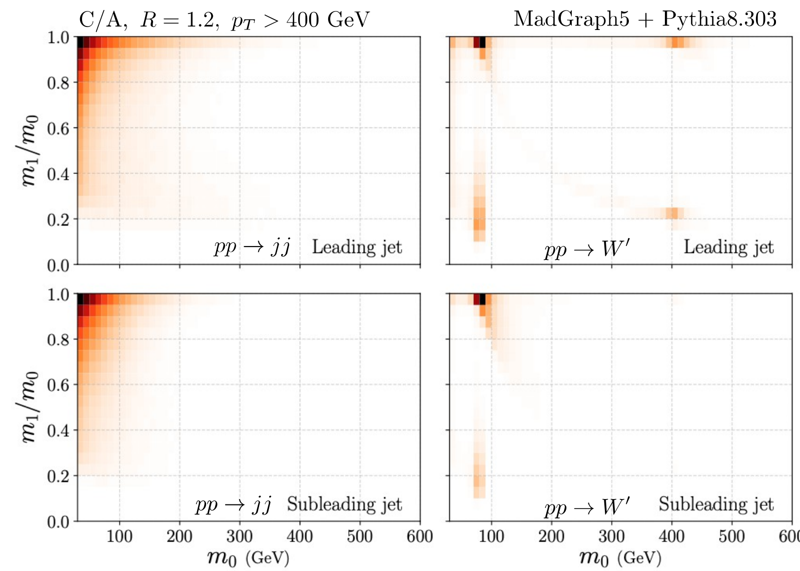
$\Phi \to W^{\pm} W^{\mp}$

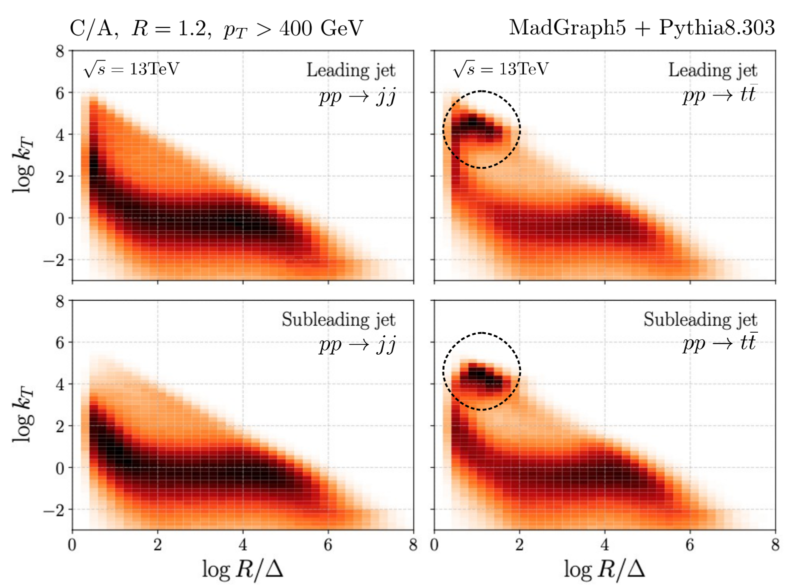


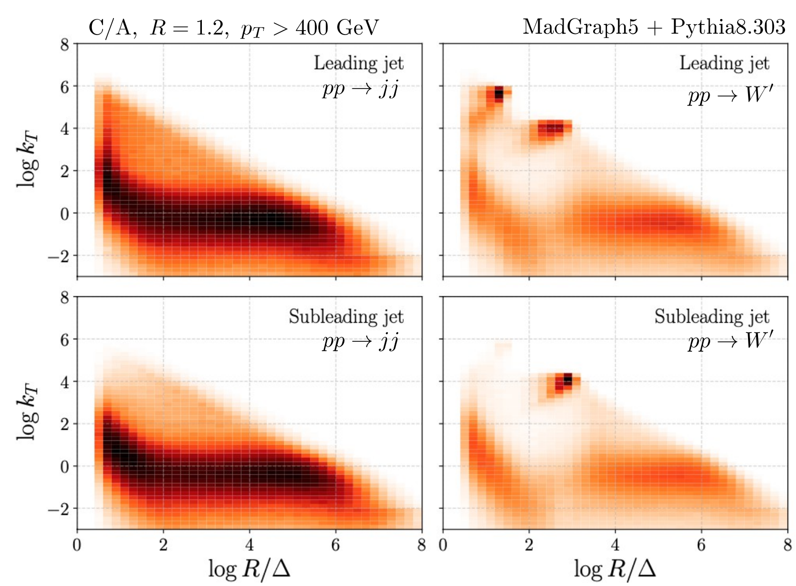
$m_{W'} = 3\,\text{TeV}$

$m_{\Phi} = 400\,\text{GeV}$



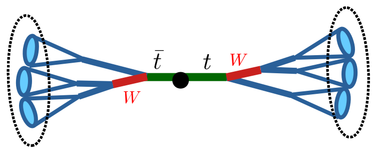


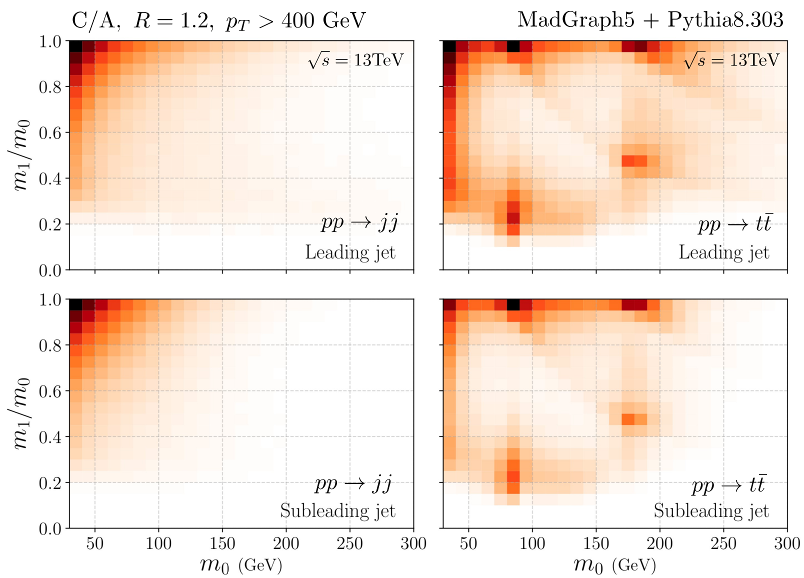


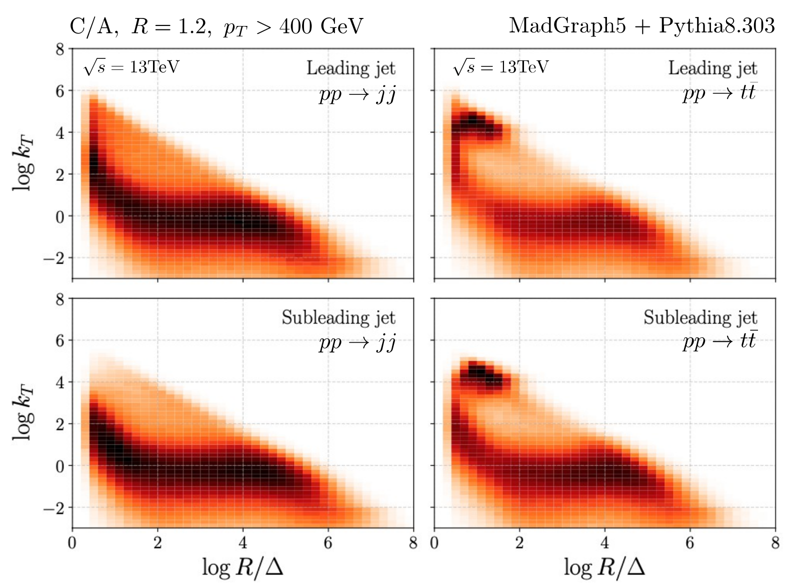


• Top-quarks

$pp \to t\bar{t}$

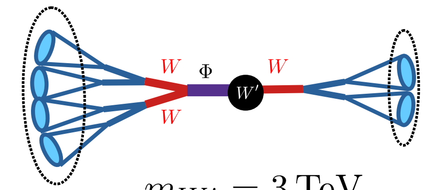




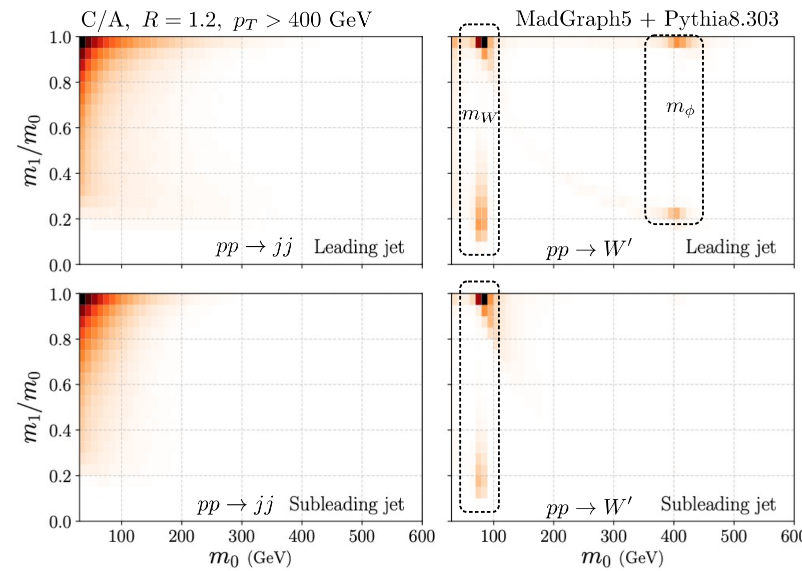


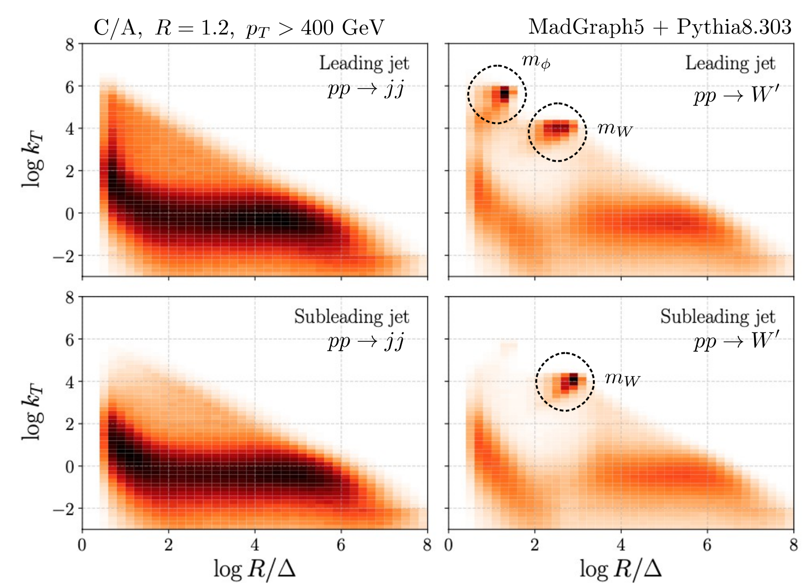
• BSM model:

$$pp \to W' \to \Phi W^{\pm}$$
$$\Phi \to W^{\pm} W^{\mp}$$



$$m_{W'} = 3\,\text{TeV}$$
$$m_{\Phi} = 400\,\text{GeV}$$





Rare signals with LDA

- Two-theme LDA:

If LDA works well: $\begin{cases} p_1 := p(o|\beta_1) & \text{1st theme: should include most QCD features} \\ p_2 := p(o|\beta_2) & \text{2nd theme: should include most signal features (e.g. BSM)} \end{cases}$

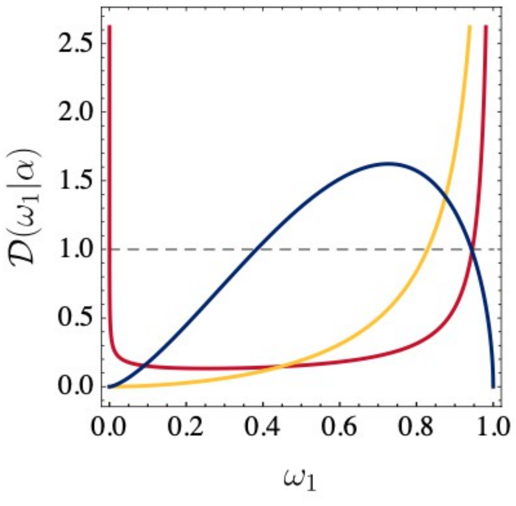
- Which Dirichlet prior for the theme mixture? $\omega \sim D(\omega|\alpha_1, \alpha_2)$

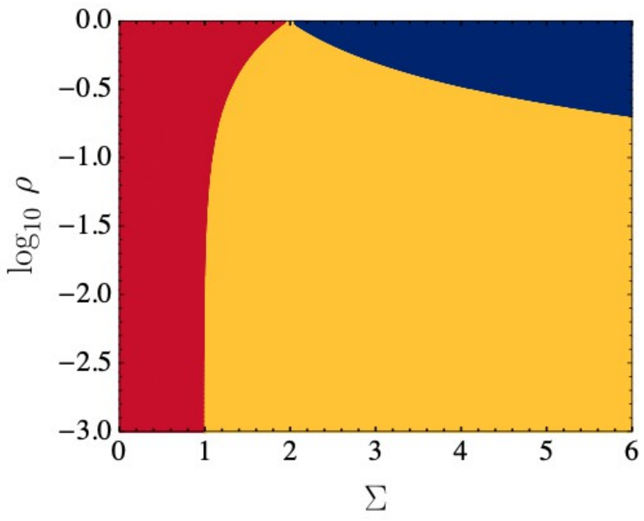
Prior with asymmetric shape $\Longleftarrow$ We want to discover *rare* signal in a data sample dominated by QCD $b \gg s$

Reparametrization: $(\alpha_1, \alpha_2) \to (\rho, \Sigma)$

$$\begin{cases} \Sigma = \alpha_1 + \alpha_2 \\ \rho = \dfrac{\alpha_2}{\alpha_1} \end{cases} \qquad \mathbb{E}\left[\omega p_1 + (1-\omega)p_2\right] = \frac{p_1 + \rho\, p_2}{1+\rho} \xrightarrow[\rho \ll 1]{} p_1$$

Controls prior shape asymmetry





$\rho \leq 0.1$ usually works...

Landscape of 2-theme LDA models!

Back to 1995: 're-discovering' Top-quarks

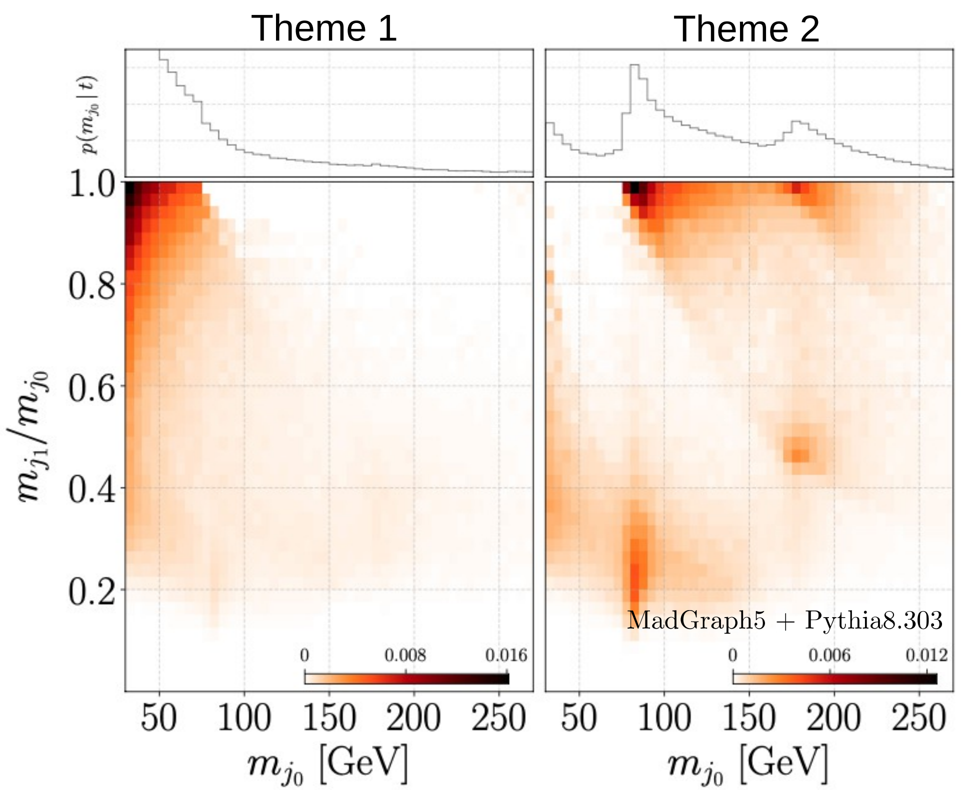
- Train two-theme LDA on mixed (unlabelled) QCD + tops sample **~ 50k events**
- Training performed with `Gensim` (python package)
- Unsupervised classifier results:

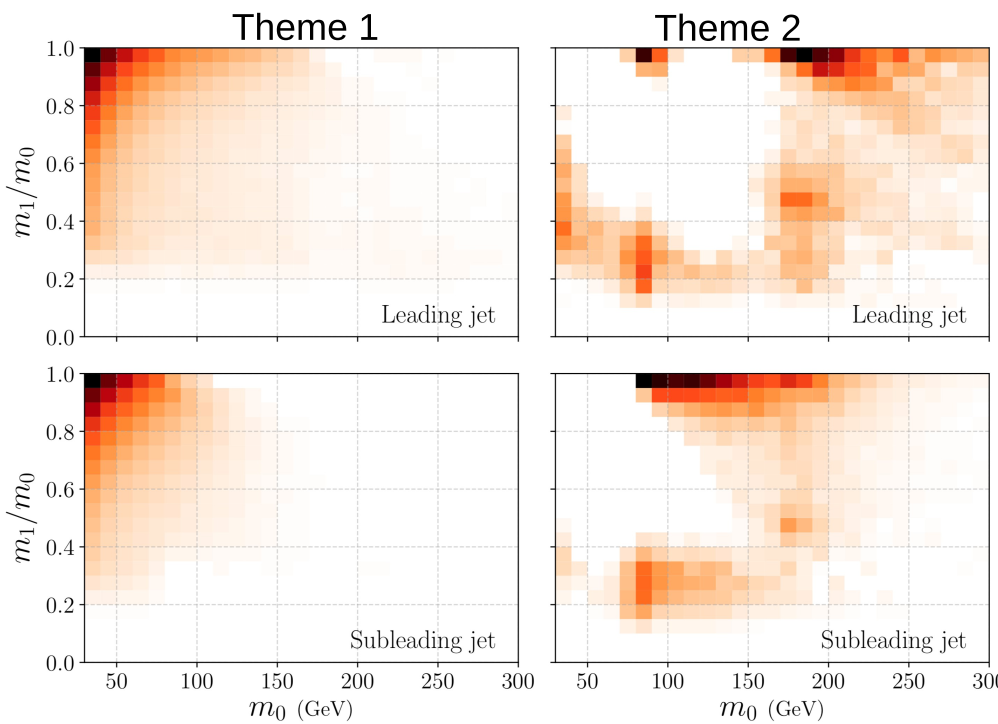
Proof of concept: $s/b = 1$

$$\mathcal{O}_{\text{Mass}} = \left\{ m_{j_0}, \frac{m_{j_1}}{m_{j_0}} \right\} \qquad (\rho, \Sigma) = (1, 1)$$

Small signal: $s/b = 0.05$

$$\mathcal{O}_{\text{Mass}} = \left\{ \ell, m_{j_0}, \frac{m_{j_1}}{m_{j_0}} \right\} \qquad (\rho, \Sigma) = (0.1, 1.5)$$

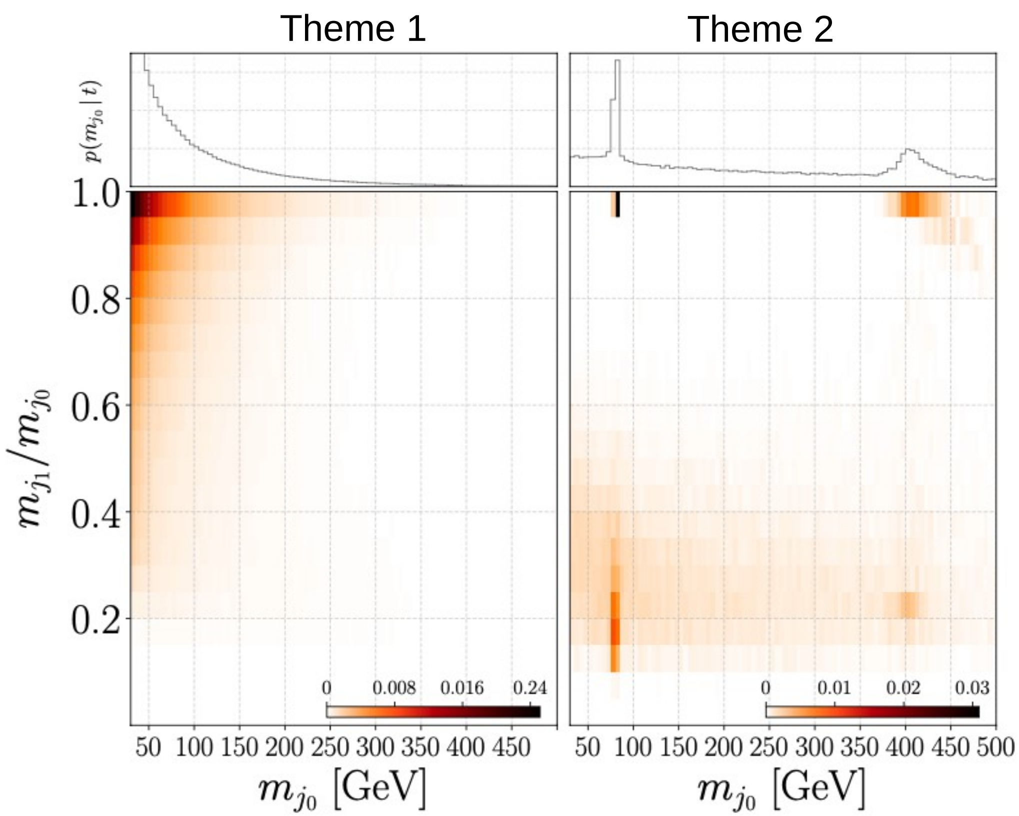


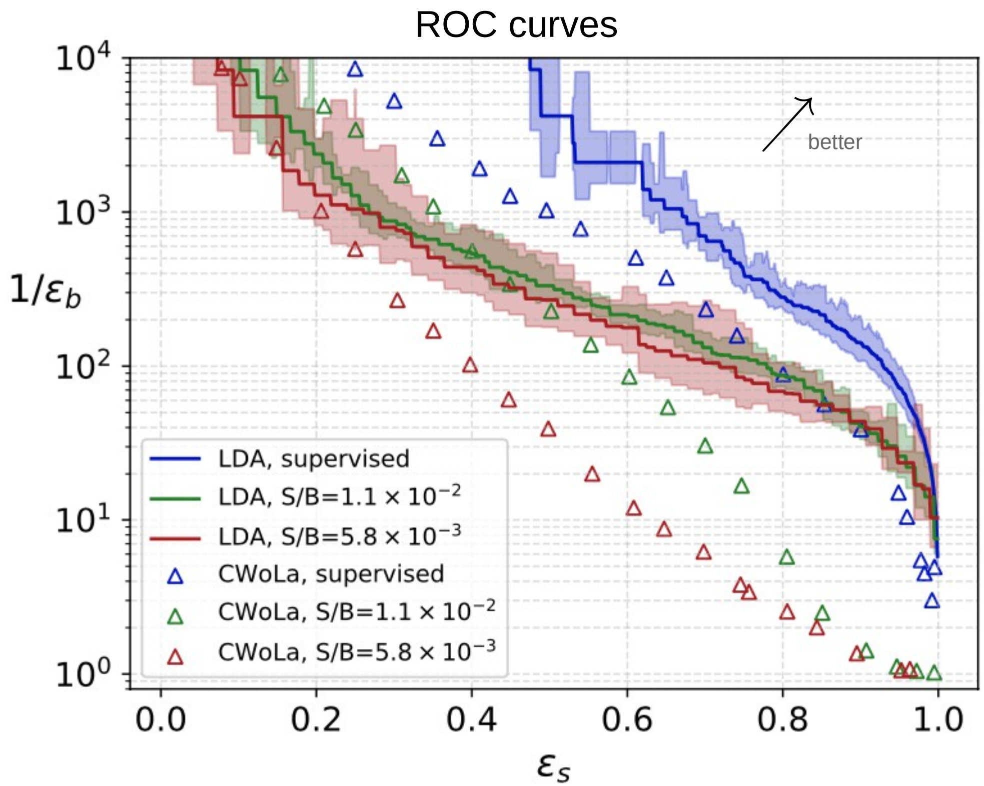


Uncovering BSM physics

$$pp \to W' \to \Phi W^{\pm},\ \ \Phi \to W^{\pm}W^{\mp} \qquad 2.7 \leq m_{JJ} \leq 3.2\ \text{TeV}$$

~ 100k events $\quad s/b = 0.01 \quad \mathcal{O}_{\text{Mass}} = \left\{ m_{j_0}, \frac{m_{j_1}}{m_{j_0}} \right\} \quad (\rho, \Sigma) = (0.1, 1)$



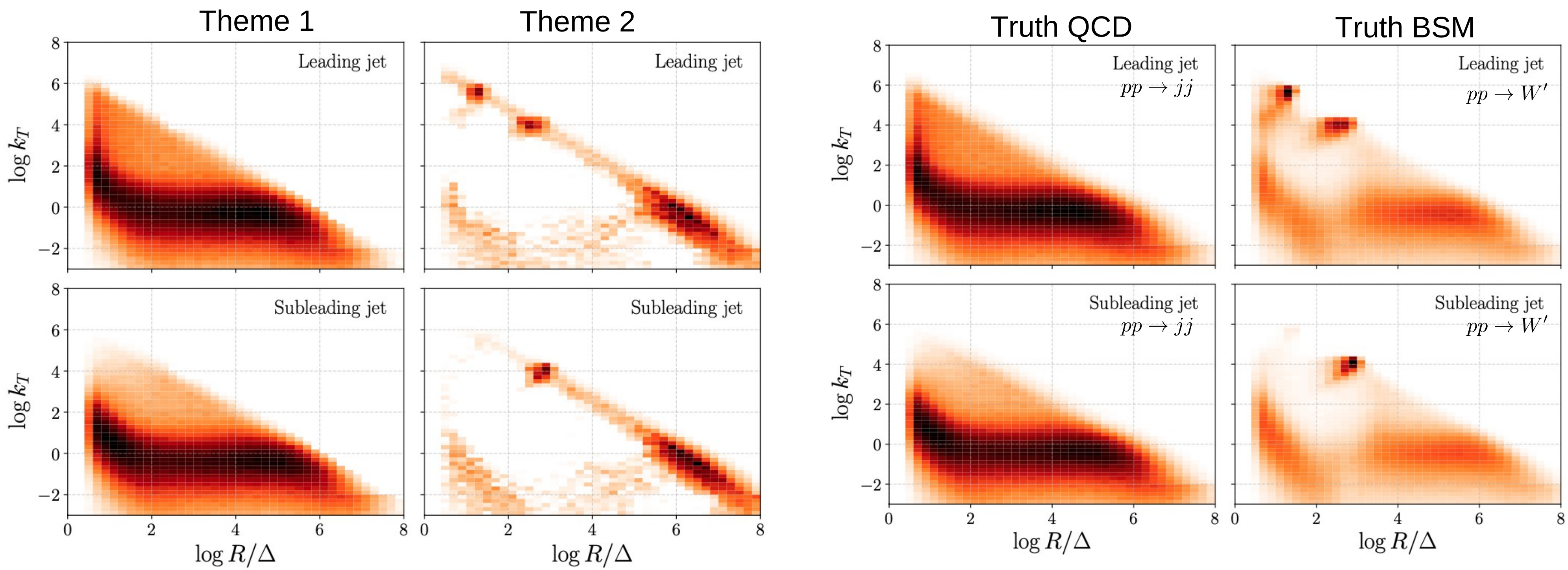


Excellent performance at small signal-to-background ratios!

Uncovering BSM physics from the Lund plane

$$pp \to W' \to \Phi W^{\pm},\ \Phi \to W^{\pm} W^{\mp}$$

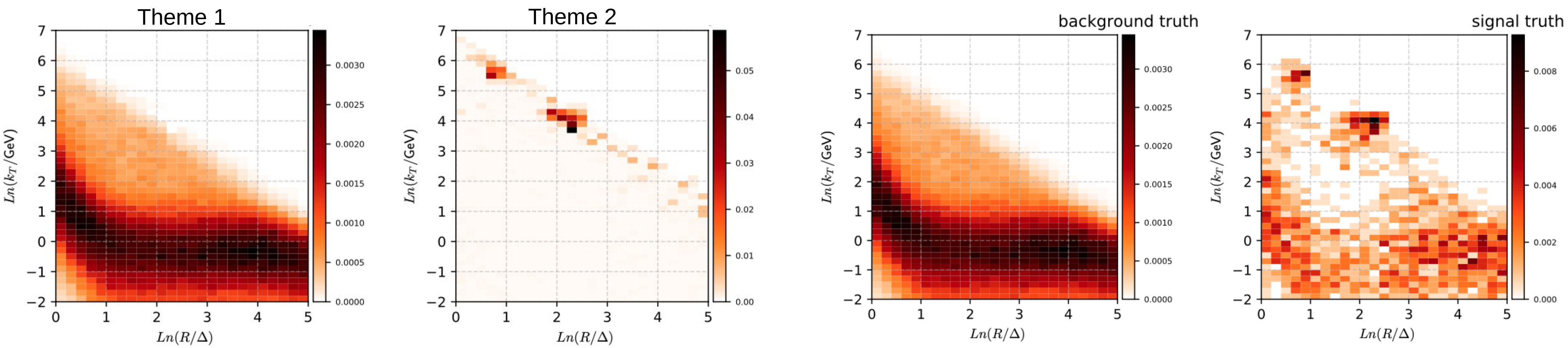
~ 100k events $s/b = 0.01$ $\mathcal{O}_{\rm Lund} = \left\{\ell, \log(k_t), \log\left(\frac{R}{\Delta R}\right)\right\}$ $(\rho, \Sigma) = (0.1, 1)$



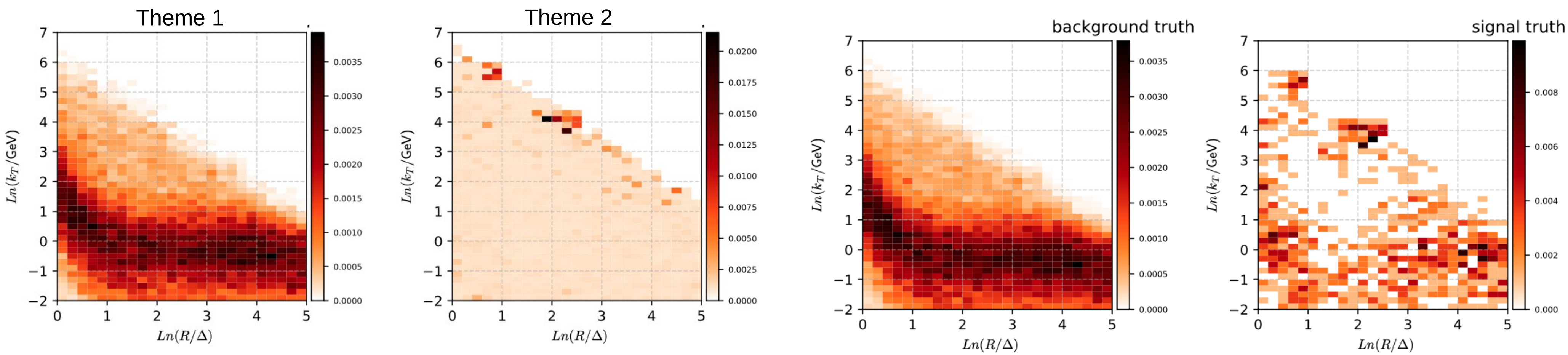
LDA discovers the hard/colinear splittings of the massive resonance decays in the Primary lund plane.

- What if we train on much less events?

$\left\{\begin{array}{l}\text{10k QCD events}\\ \text{100 signal events}\end{array}\right.$ $s/b = 0.01$ $\mathcal{O}_{\text{Lund}} = \left\{\log(k_t), \log\left(\frac{R}{\Delta R}\right)\right\}$ $(\rho, \Sigma) = (0.0009, 5.2)$



$\left\{\begin{array}{l}\text{1600 QCD events}\\ \text{40 signal events}\end{array}\right.$ $s/b = 0.025$ $\mathcal{O}_{\text{Lund}} = \left\{\log(k_t), \log\left(\frac{R}{\Delta R}\right)\right\}$ $(\rho, \Sigma) = (0.09, 4.0)$



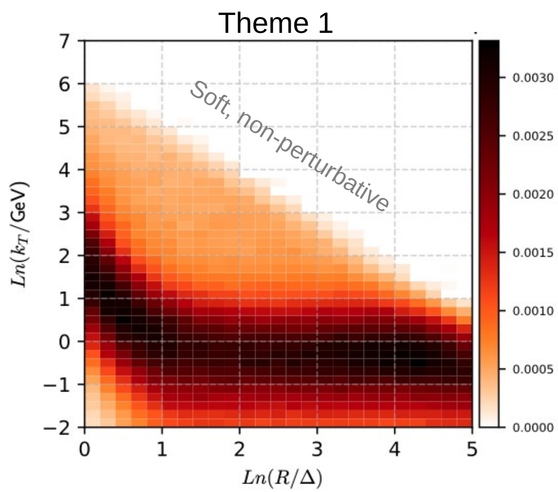
- LDA works well with small data samples!

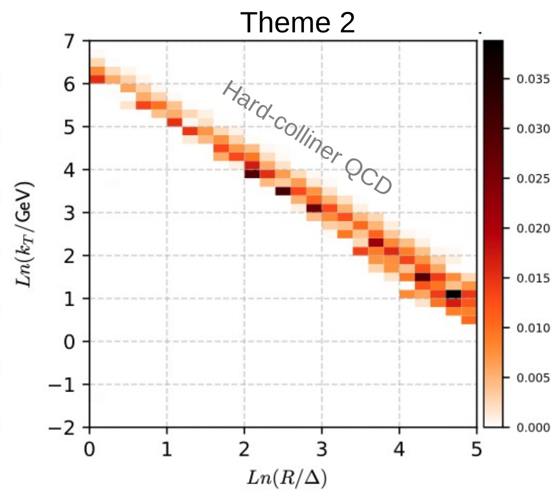
- What if there is NO signal?

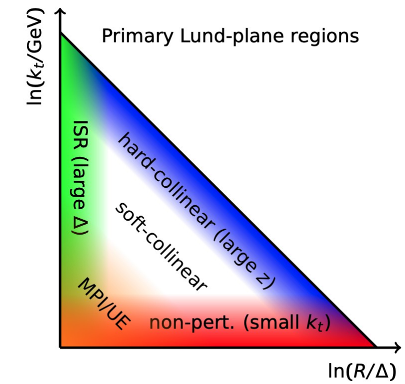
Train ~ 100k QCD events

Asymmetric Dirichlet prior

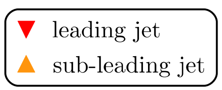
$$(\rho, \Sigma) = (0.1, 1)$$

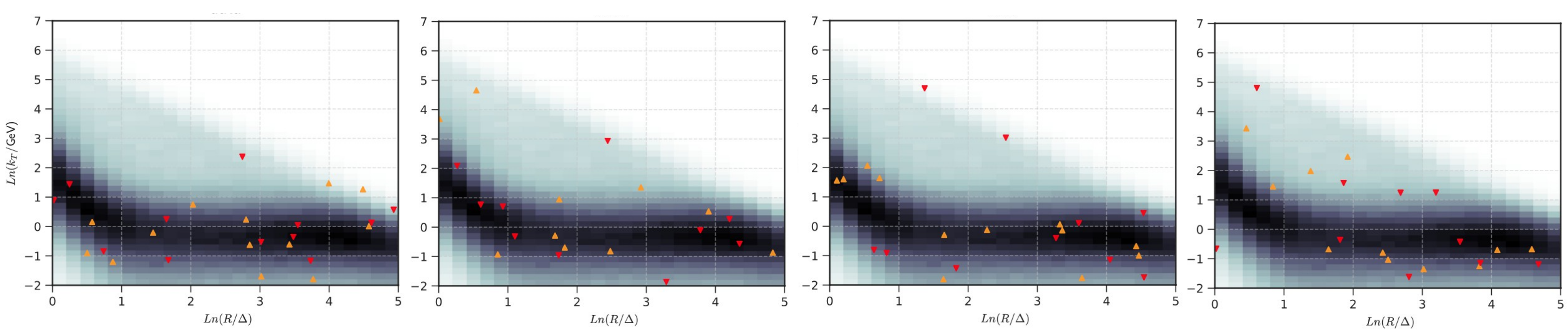






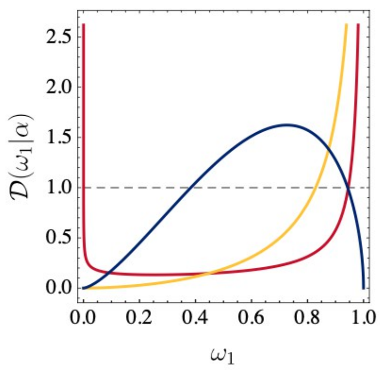
- 4 QCD events:

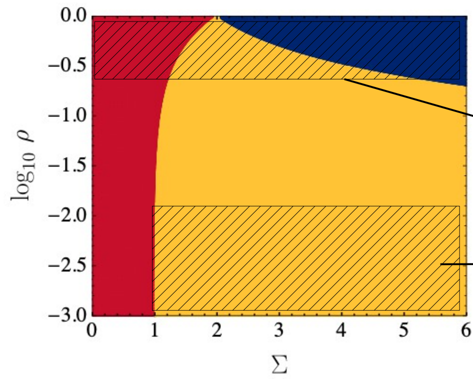




Landscape of 2-theme LDA models

- (ρ, Σ) - plane

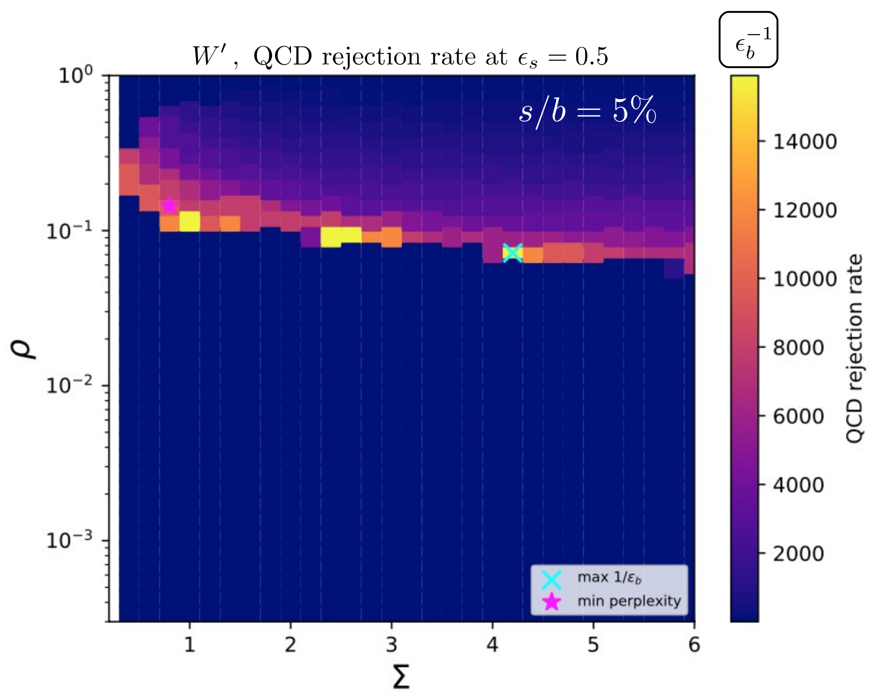


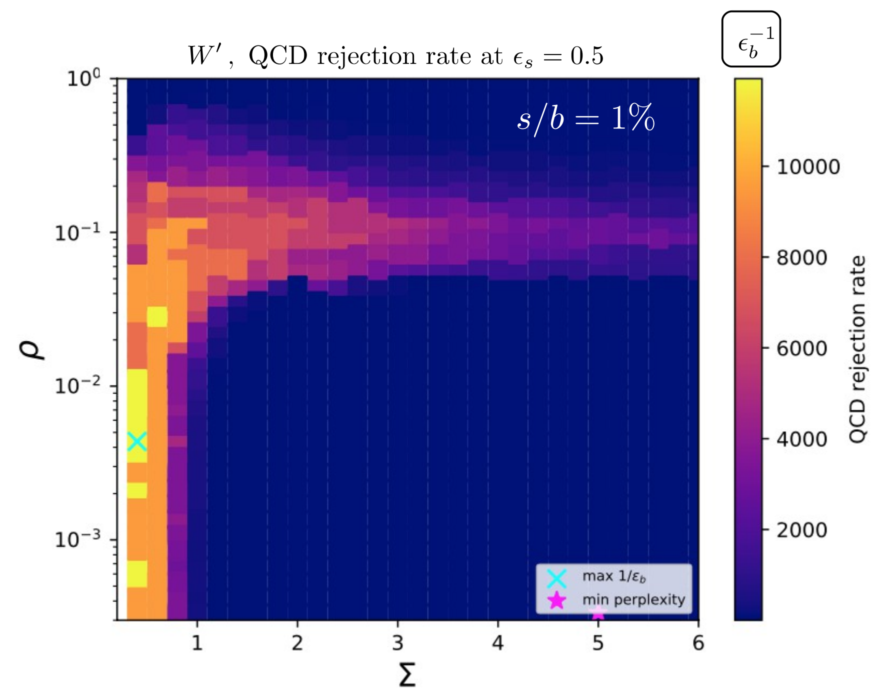


Symmetric shapes: not good for for our task...

Too asymmetric J-shape miss 'rare' themes

- Event classifcation performance over the LDA landscape:





Perplexity

- We need a criteria for selecting from all models in the Landscape the one with the “best” performance.

  We need a statistical goodness-of-fit test for the generative model.

- Perplexity:

  For an event sample $\mathcal{D} = \{e_1, \ldots, e_N\}$

$$\text{perplexity}(\mathcal{D}) := 2^{-b} \qquad b = \frac{1}{n_{\text{tot}}} \sum_{j=1}^{N} \log p(e_j) \approx \frac{1}{n_{\text{tot}}} \sum_{j=1}^{N} \mathcal{L}(e_j)$$

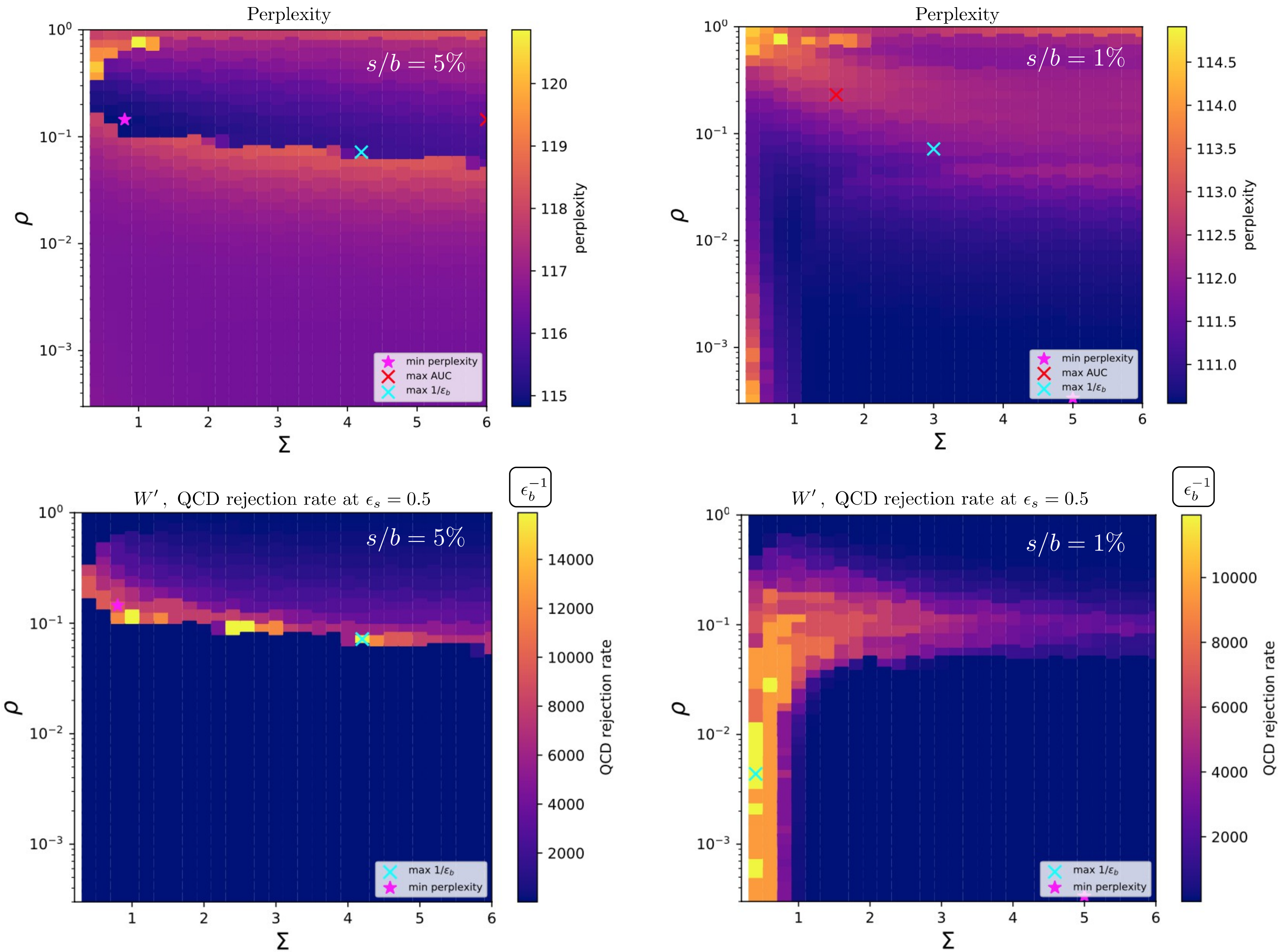
Total number of measurements

ELBO

  - Perplexity is the measure of how well a generative model fits the data sample.

Good models have a lower perplexity score, i.e. a greater probability it generated the observed data.

Trained ~1000 2-theme LDA models



Summary

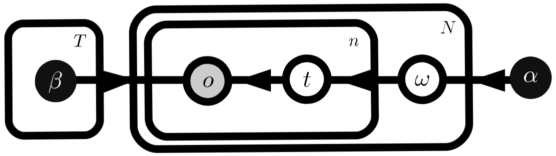
- We need new model indendent ways of searching for BSM physics at the LHC.

- We showed that simple generative probabilistic models can be used to describe generic data representations for collider events.

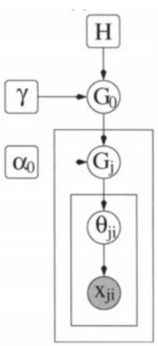
- Under broad assumptions we arrived to the Latent Dirichlet Allocation (LDA) model.

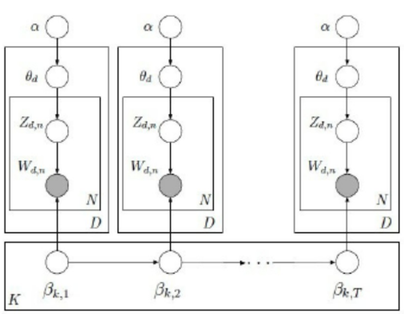
- We demonstrated that LDA can be used to uncover heavy resonances in dijet samples in a fully unsupervised manner.

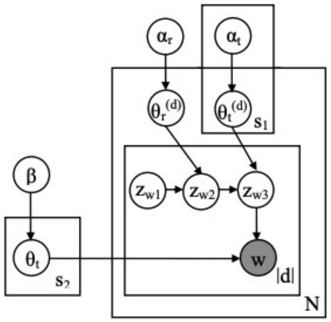
- LDA is just one possible probabilistic model...

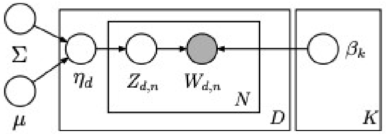


It can used as a building block for more complex probilistic models for collider events.

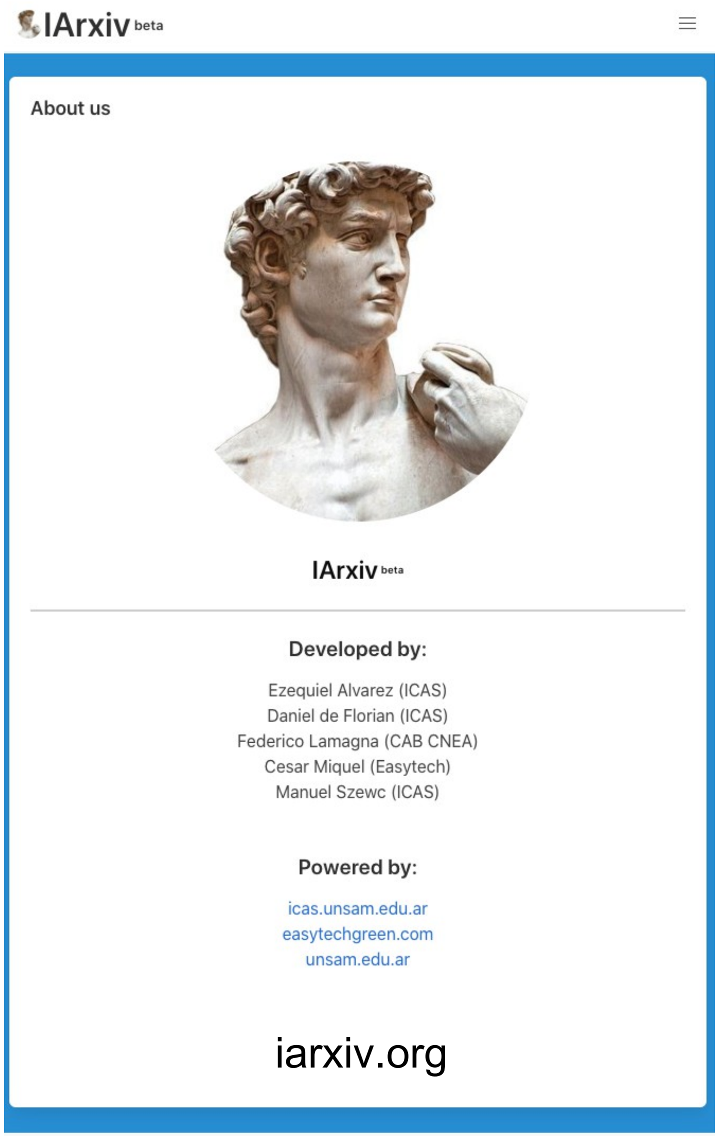






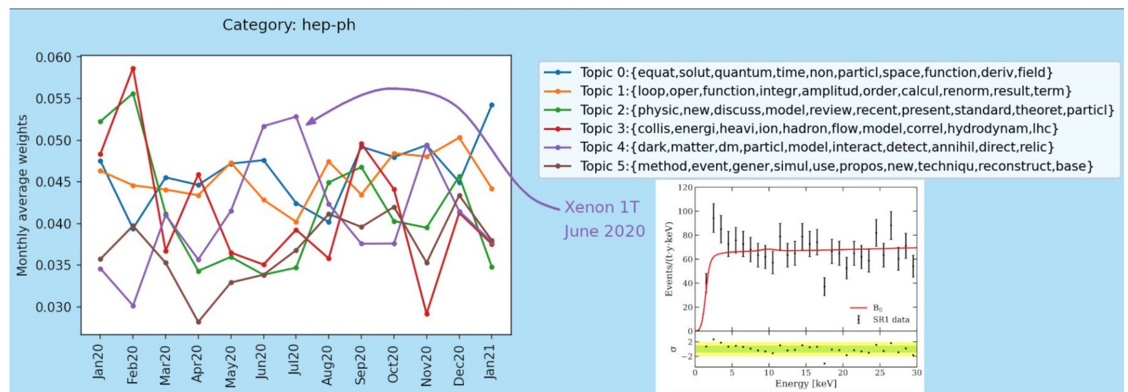


Thank You!



iarxiv.org

Uses LDA to sort daily papers by learning users topic preferences



LDA discovered the Xenon 1T anomaly

Backup

Latent Dirichlet Allocation (LDA)

$$\mathcal{D} = \{e_1, \cdots, e_N\}$$

$$\mathcal{P}(\mathcal{D}|\alpha, \beta) = \prod_{j=1}^{N} \left[\int_{\Omega_{T-1}} \mathrm{d}\omega_j \, D(\omega_j|\alpha) \prod_{i=1}^{n_j} \left(\sum_{t=1}^{T} p(t|\omega_j) \, p(o_{ij}|\beta_t) \right) \right]$$

Event index

Simplex

Dirichlet Prior

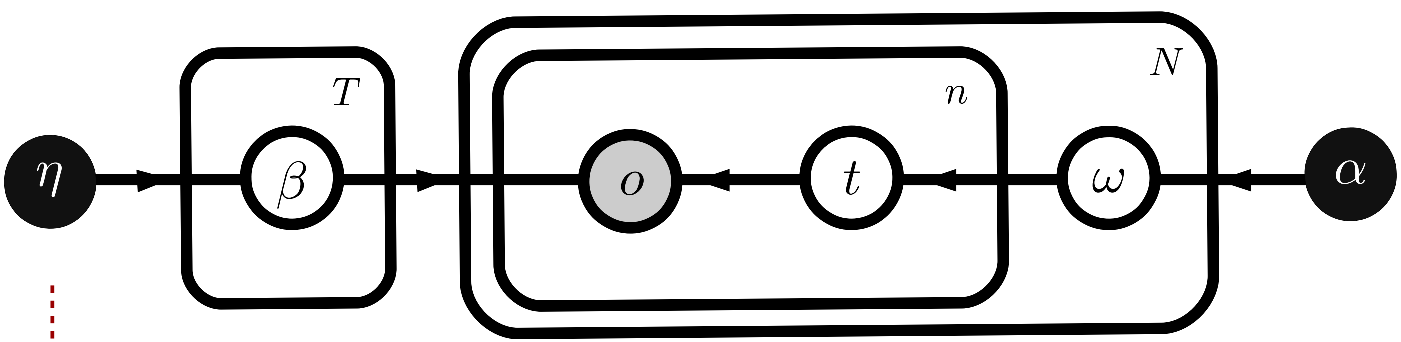
Theme assignment variable

Themes (Multinomials)

Theme mixing

- LDA is a mixed-membership model.
- Individual events are described by mixture of multiple themes:

- 'Smoothed' LDA graphical model:

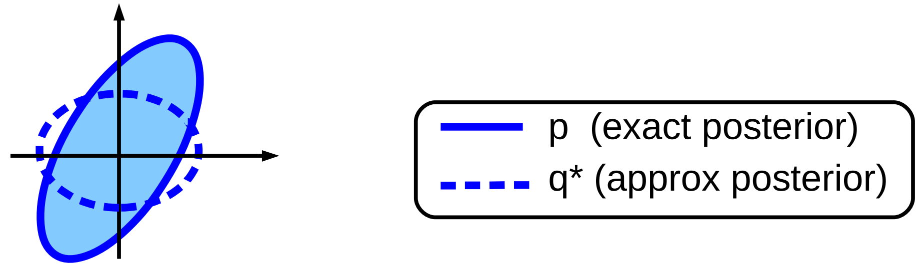


$D(\beta_t|\eta_t)$ Dirichlet priors for theme parameters (not to be confused with the Dirichlet for theme proportions...)

Choose $\mathcal{Q}$ flexible enough to approximate posterior... but simple enough for efficient optimization.

- "Mean-field" variational family:

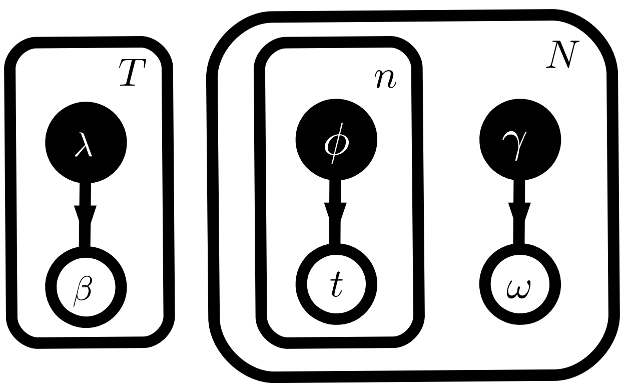
$$q(\theta|\mu) = \prod_i q(\theta_i|\mu_i)$$



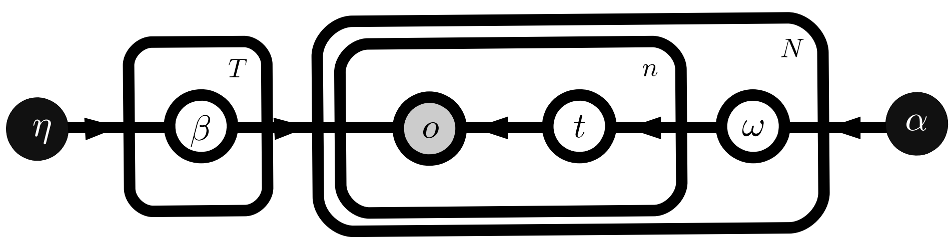
- LDA variational inference:

$$q(\omega, t, \beta|\lambda, \phi, \gamma) = q(\omega|\gamma)\, q(t|\phi)\, q(\beta|\lambda)$$

LDA mean-field approximation



LDA



$$q(\omega) = \mathrm{Dirichlet}(\omega|\gamma)$$
$$q(t) = \mathrm{Multinomial}(\phi)$$
$$q(\beta) = \mathrm{Dirichlet}(\beta|\lambda)$$

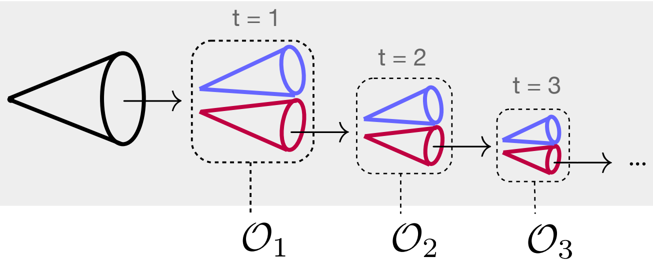
$$(\lambda^*, \phi^*, \gamma^*) = \underset{(\lambda,\phi,\gamma)}{\mathrm{argmax}}\, \mathcal{L}[q(\omega, t, \beta)|\lambda, \phi, \gamma)]$$

Mass-drop tagger or BDRS tagger

Butterworth, Davison, Rubin, Salam 2008

- Travel through ‘hardest’ branch of the declustering tree

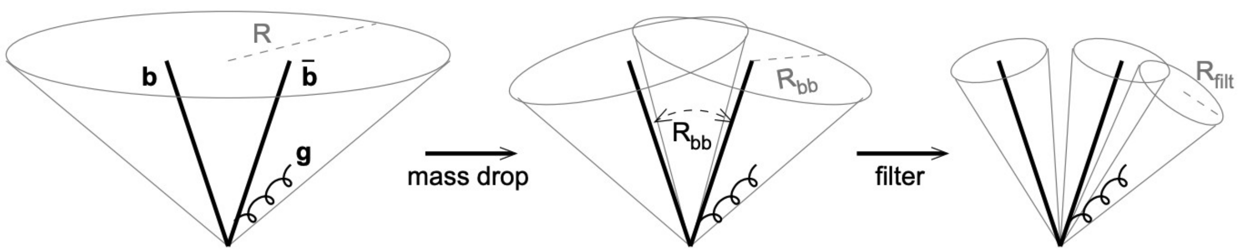
Cluster with C/A algorithm with R=1.2

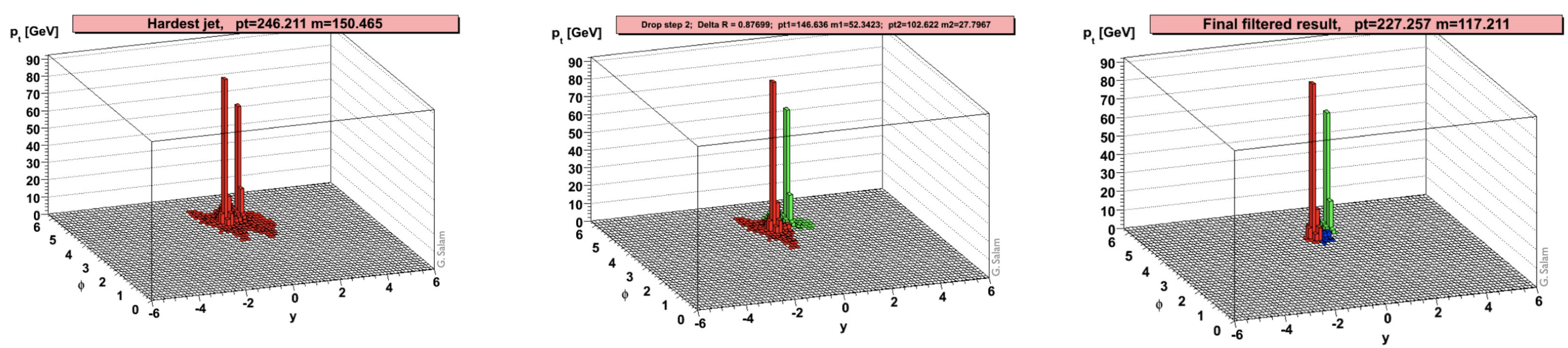


Compute: $\mathcal{O}_t = \left\{ \frac{m_{j_0}}{m_{j_1}}, \frac{\min({p_T}_1^2, {p_T}_2^2)}{m_{j_0}} \Delta R_{12}^2 \right\}$

mass-drop   asymmetry

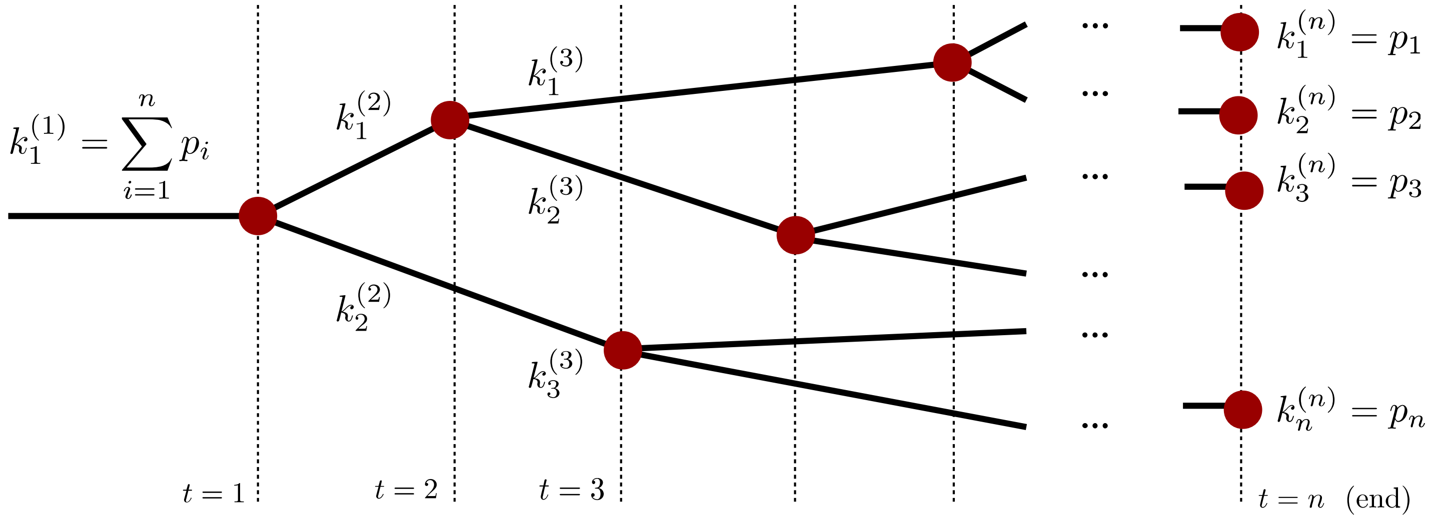
- $h \to b\bar{b}$ Higgs tagging condition: $\begin{cases} \text{massdrop} < 0.67 \\ \text{asymmetry} > 0.09 \end{cases}$





Probabilistic model for jet formation?

- If we are faithful to the jet declustering process:



- Probabilitic model for a jet:

$$P(p_1, p_2, ..., p_n) = \prod_{t=1}^{n-1} P\left(k_1^{(t+1)}, k_2^{(t+1)}, \ldots, k_{t+1}^{(t+1)} \middle| , k_1^{(t)}, \ldots, k_t^{(t)}, \theta_t\right) \times P\left(z \middle| k_1^{(n)}, \ldots, k_n^{(n)}\right)$$

Non-trivial conditional probabilities
(Markovian structure)

Latent variables

JUNIPR framework Andreassen ,Feige, Frye, Schwartz 2019

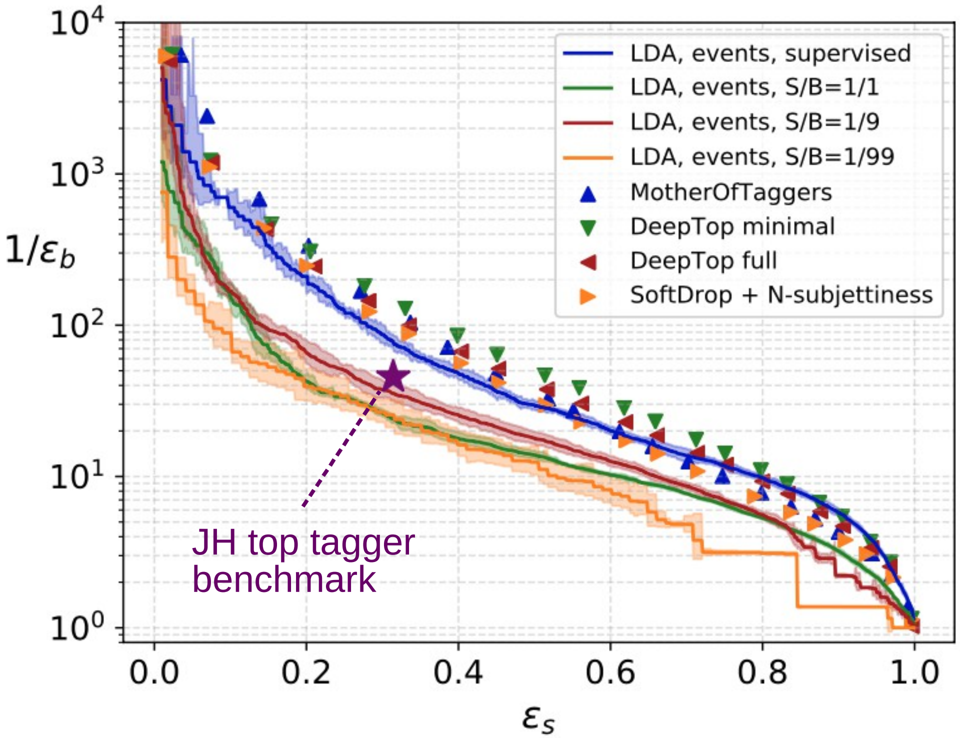
QCD-aware RNN Louppe, Cho, Becot, Cranmer 2017

Not clear how to generalize to
unsupervised jet/event classification tasks

Back to 1995: 're-discovering' Top-quarks

- Train two-theme LDA on mixed (unlabelled) QCD + tops sample ~ **50k events**
- Training performed with `Gensim` (python package)
- Unsupervised mass-drop classifier results:

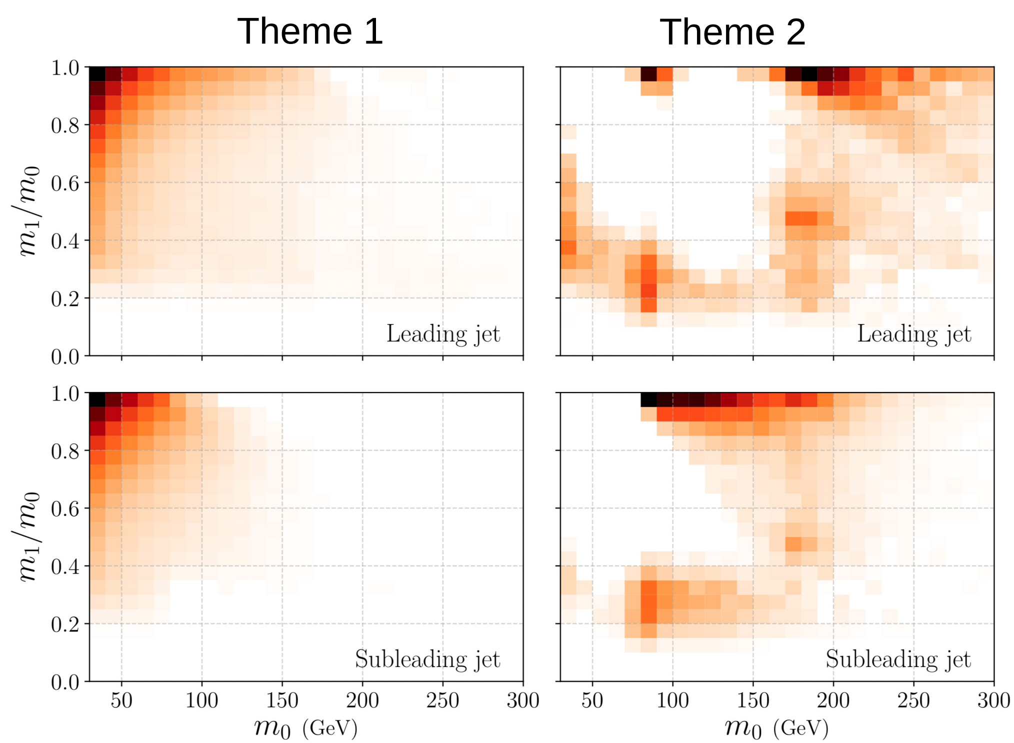
LDA classifier performance:



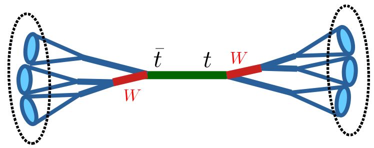
Moderate performance for unsupervised LDA classifiers

Small signal: $s/b = 0.05$

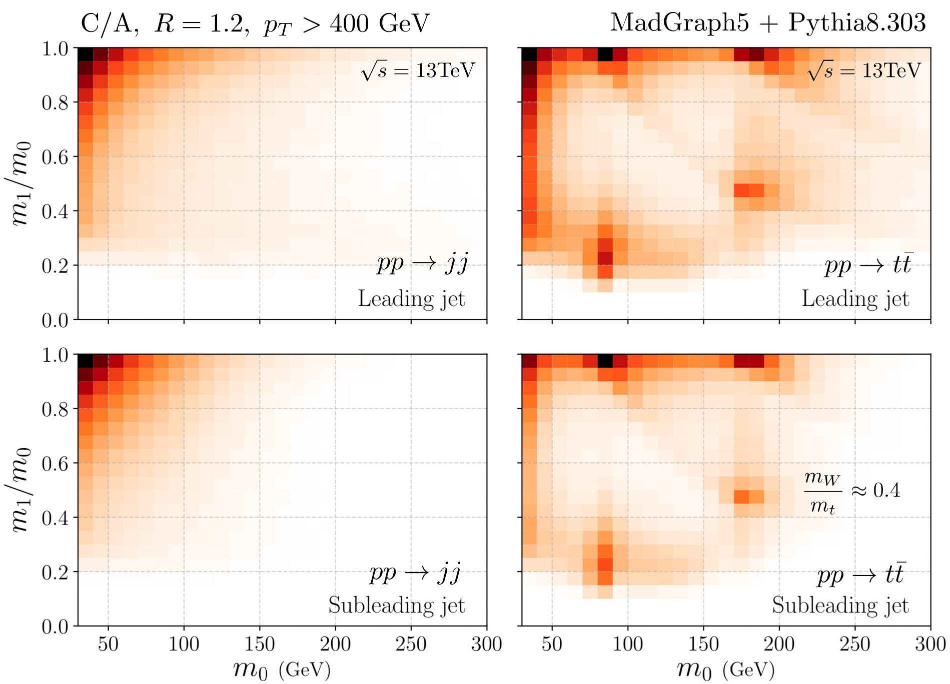
$$\mathcal{O}_{\text{Mass}} = \left\{ \ell, m_{j_0}, \frac{m_{j_1}}{m_{j_0}} \right\} \qquad (\rho, \Sigma) = (0.1, 1.5)$$

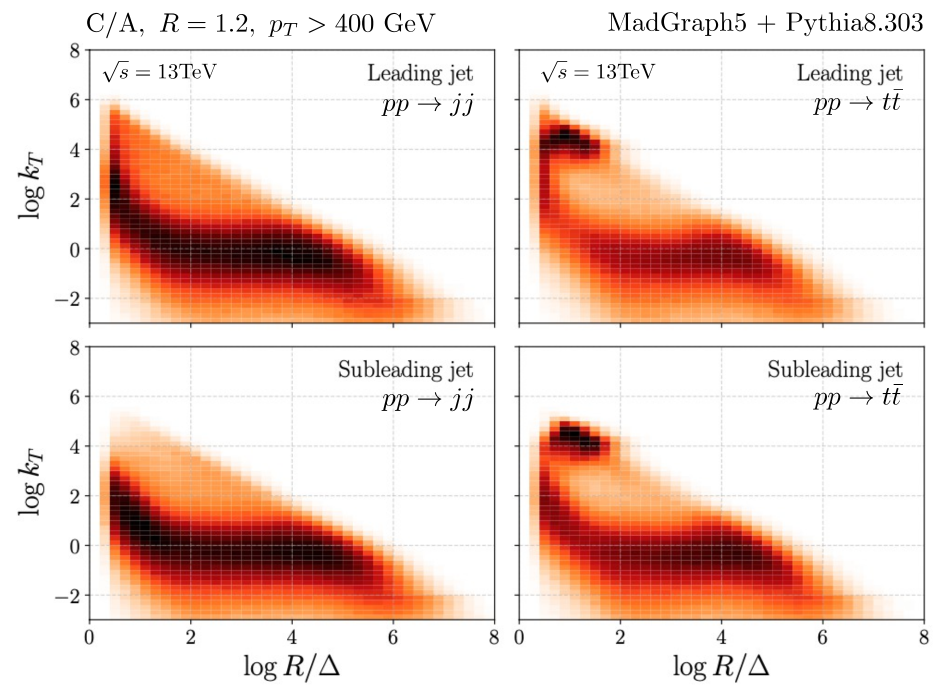


- Top pairs vs QCD:



100k events

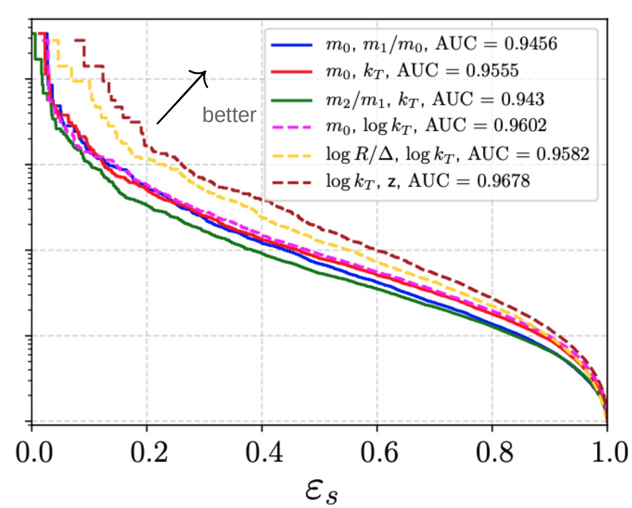




- substructure observable performance: AUC / ROC (Receiver Operator Characteristic) curves

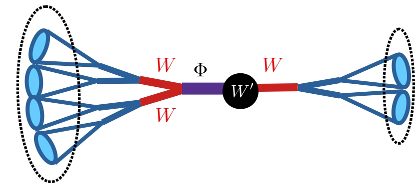
$$L_{NP}(e) = \prod_{o \in e} \frac{p_{\text{truth}}(o|s)}{p_{\text{truth}}(o|b)}$$

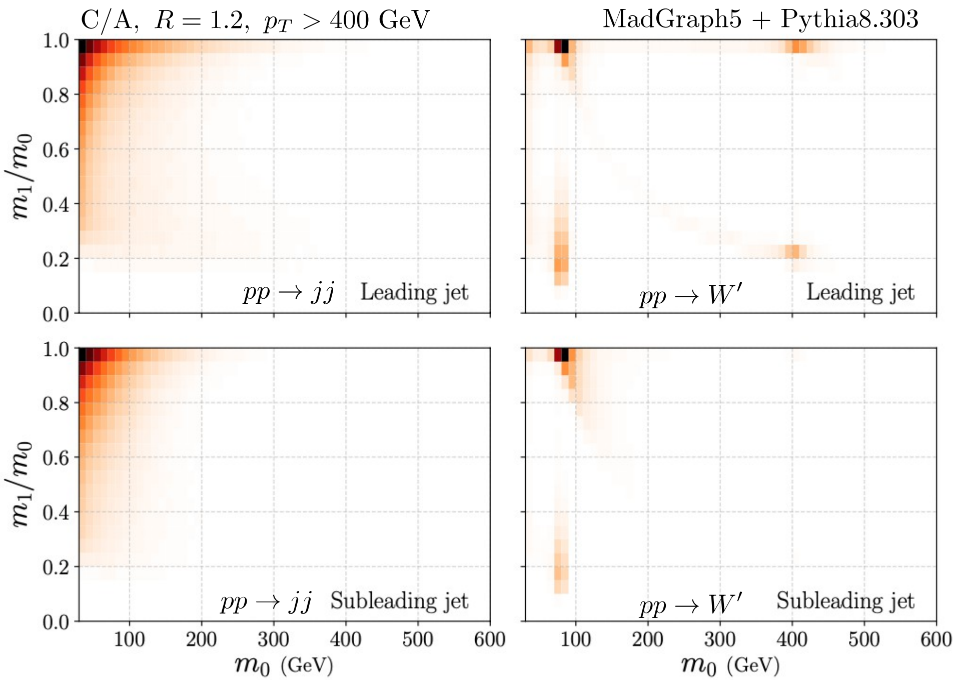
Neyman-Pearson classifier

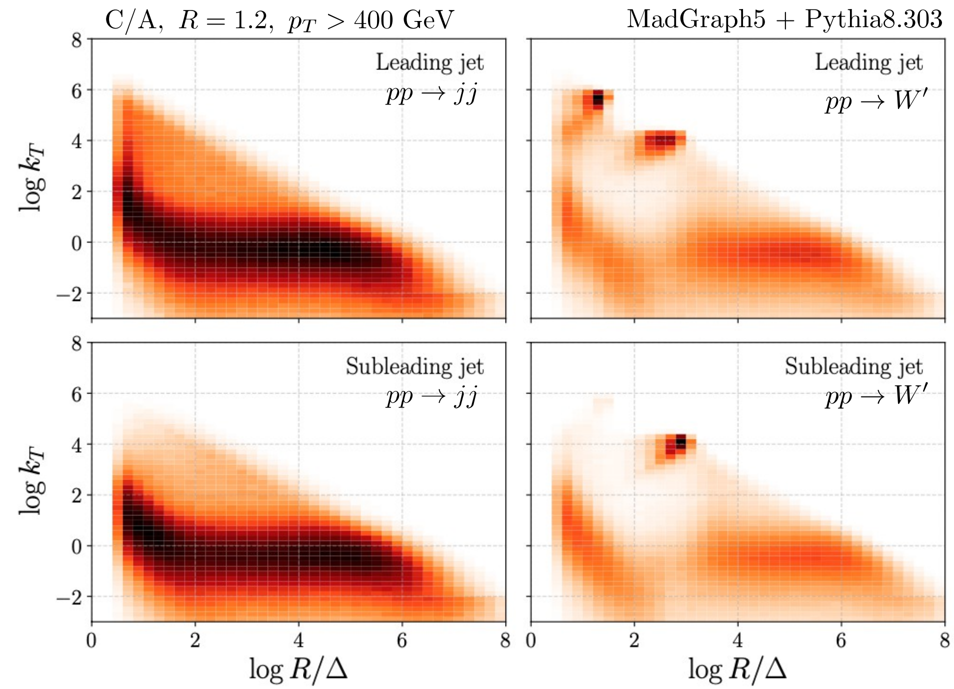


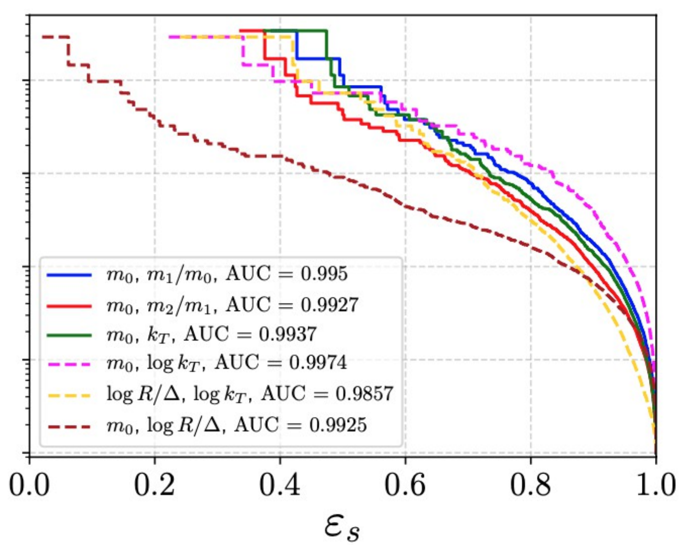
- BSM model: $pp \to W' \to \Phi W^{\pm},\ \Phi \to W^{\pm} W^{\mp}$

$$m_{W'} = 3\,\text{TeV}, \quad m_{\Phi} = 400\,\text{GeV}$$



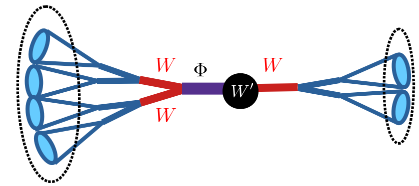


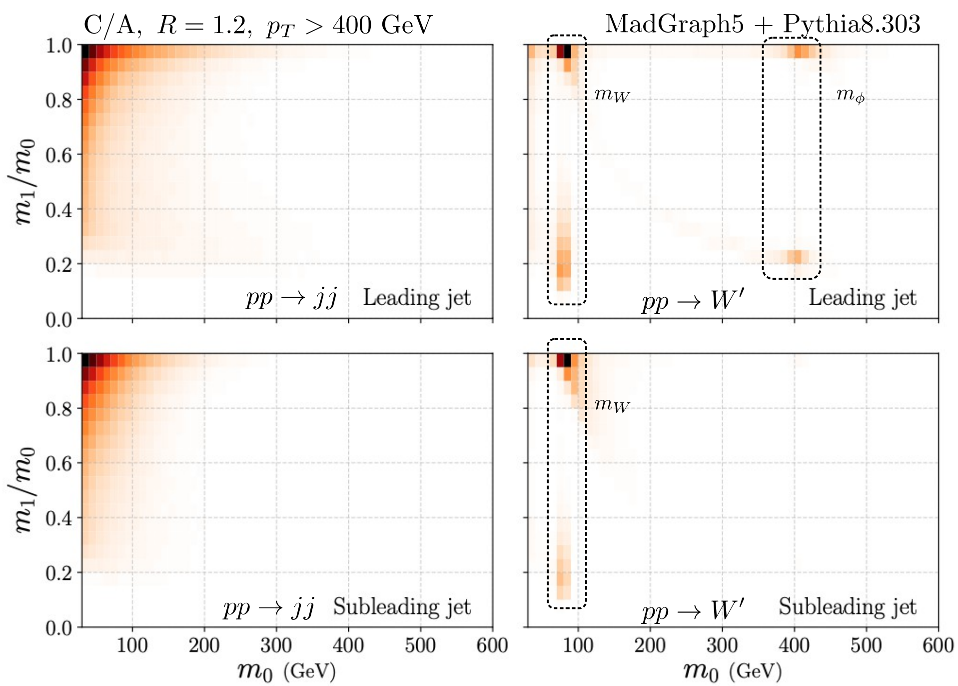


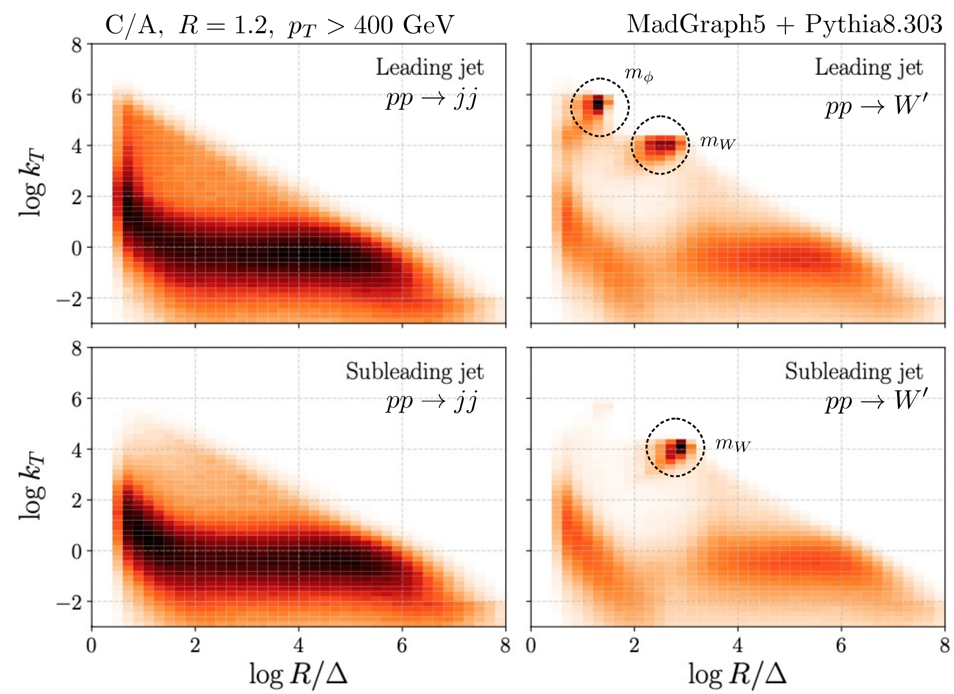


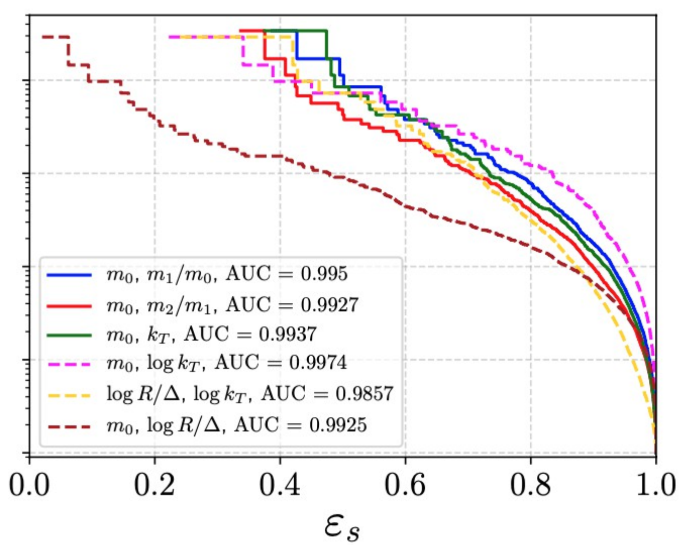
- **BSM model**: $pp \to W' \to \Phi W^{\pm},\ \Phi \to W^{\pm} W^{\mp}$

$$m_{W'} = 3\,\mathrm{TeV}, \quad m_{\Phi} = 400\,\mathrm{GeV}$$



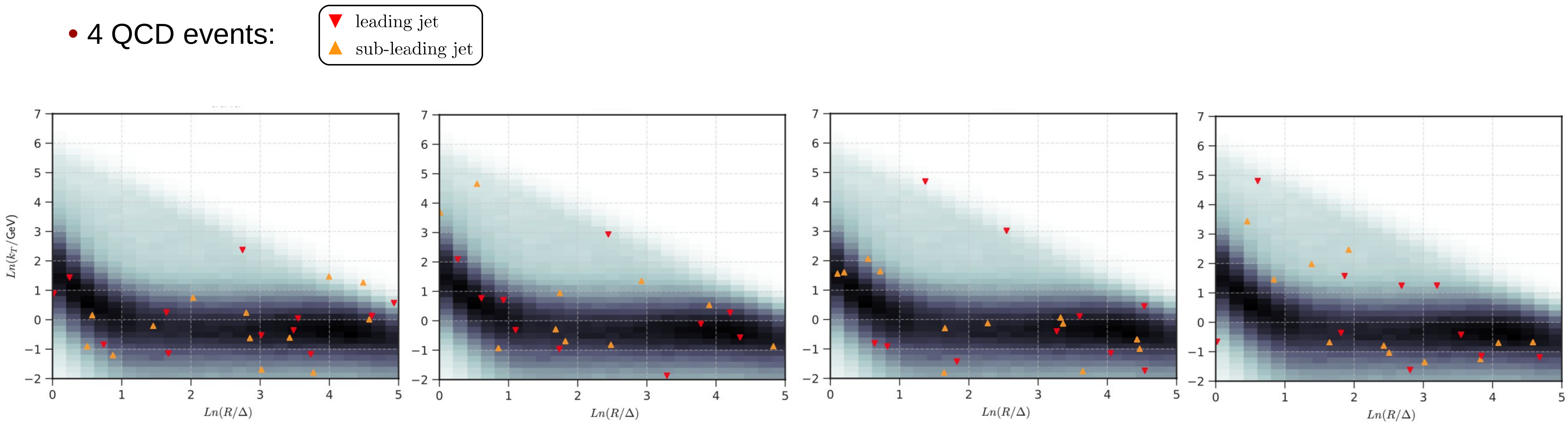




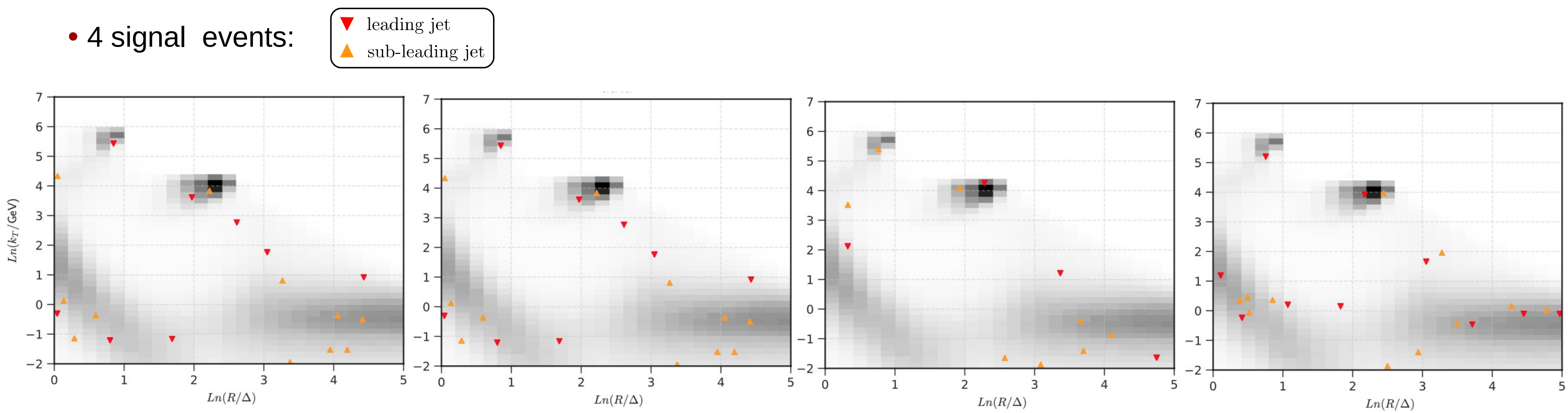


Point pattern co-ocurrences in the Lund plane

- 4 QCD events:



- 4 signal events:



- What if there is NO signal?

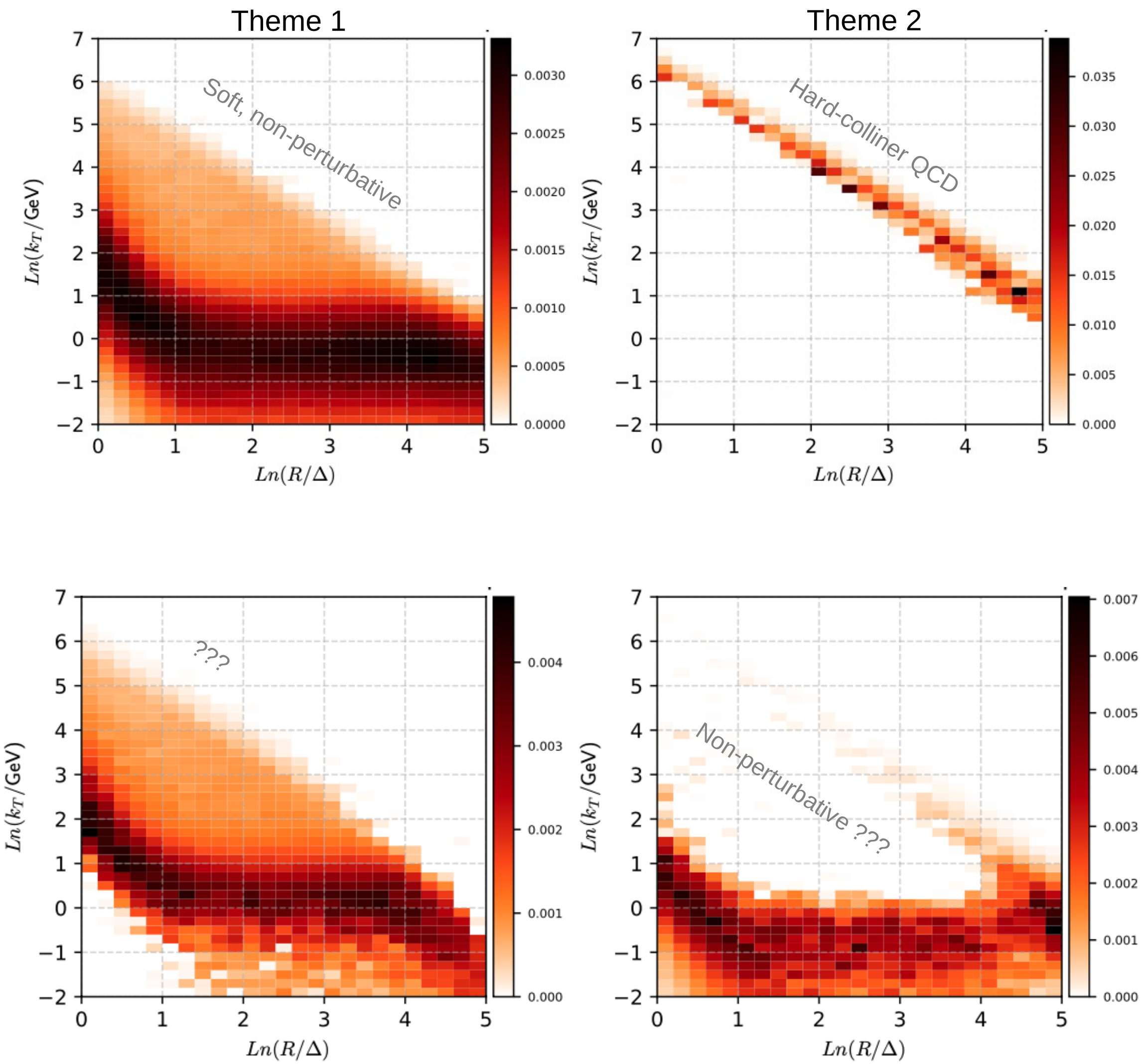
Train ~ 100k QCD events

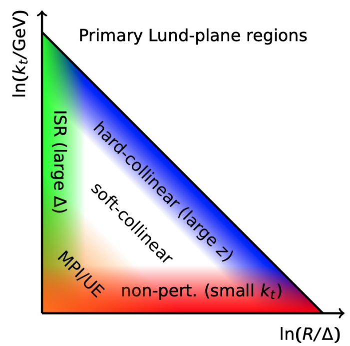
Asymmetric Dirichlet prior

$$(\rho, \Sigma) = (0.1, 1)$$

Symmetric Dirichlet prior

$$(\rho, \Sigma) = (0.75, 1.8)$$





Unphysical sculpting of data?

Perplexity

- We need a criteria for selecting from all models in the Landscape the one with the best performance without using truth data.

  We need a statistical goodness-of-fit for the generative model.

- Perplexity:

  For an event sample $\mathcal{D} = \{e_1, \dots, e_N\}$

$$\text{perplexity}(\mathcal{D}) := 2^{-b} \qquad b = \frac{1}{n_{\text{tot}}} \sum_{j=1}^{N} \log p(e_j) \approx \frac{1}{n_{\text{tot}}} \sum_{j=1}^{N} \mathcal{L}(e_j)$$

Total number of measurements

ELBO

- Perplexity is the measure of how well a generative model fits the data sample.

  Good models have a lower perplexity score, i.e. a greater probability it generated the observed data.

Trained ~1000 LDA models