Uncovering latent jet substructure

Phys.Rev. D100 (2019) no.5, 056002 [arxiv:1904.04200]

Barry M. Dillon

in collaboration with:
D. Faroughy, J. Kamenik, & M. Swezc

Jozef Stefan Institute, Ljubljana

Institut
"Jožef Stefan"
Ljubljana, Slovenija

Brda 2019

Outline of the talk

- Machine-learning in particle physics
- Unsupervised ML: A new approach for new physics
- Uncovering latent jet substructure

  Phys.Rev. D100 (2019) no.5, 056002
  BMD, D. A. Faroughy, J. F. Kamenik (& M. Szewc)

Machine-learning in particle physics

Supervised ML

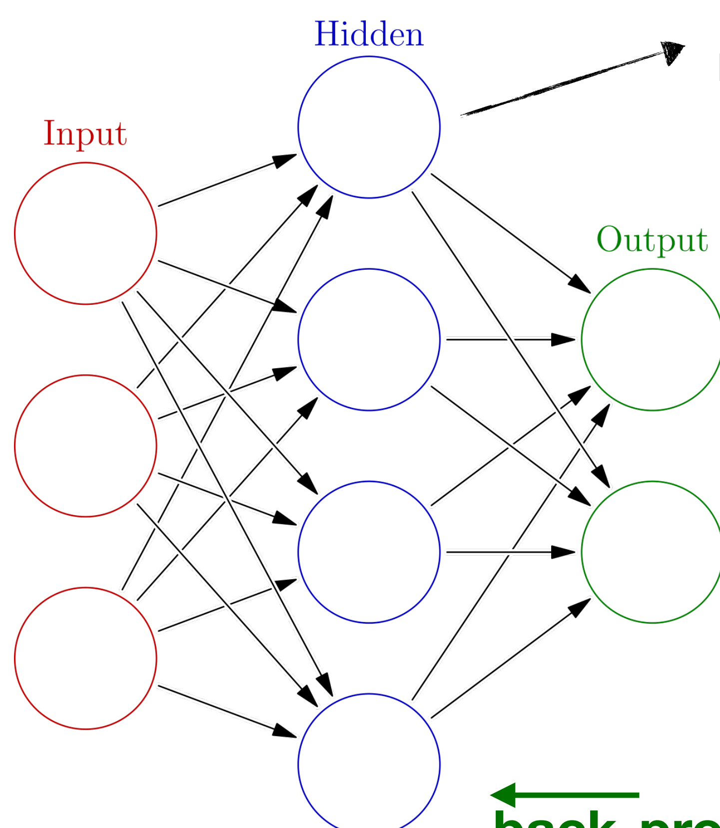
Wikipedia-style overview

Machine-learning: algorithms used to perform specific tasks without explicit instructions, relying on inference instead.

Supervised: learning a complex non-linear function that maps a high-dimensional input to an output.
User provides input and output.

Hidden layer learns the mapping

Input data where 'truth-labels' are known



Fix truth labels on output layer

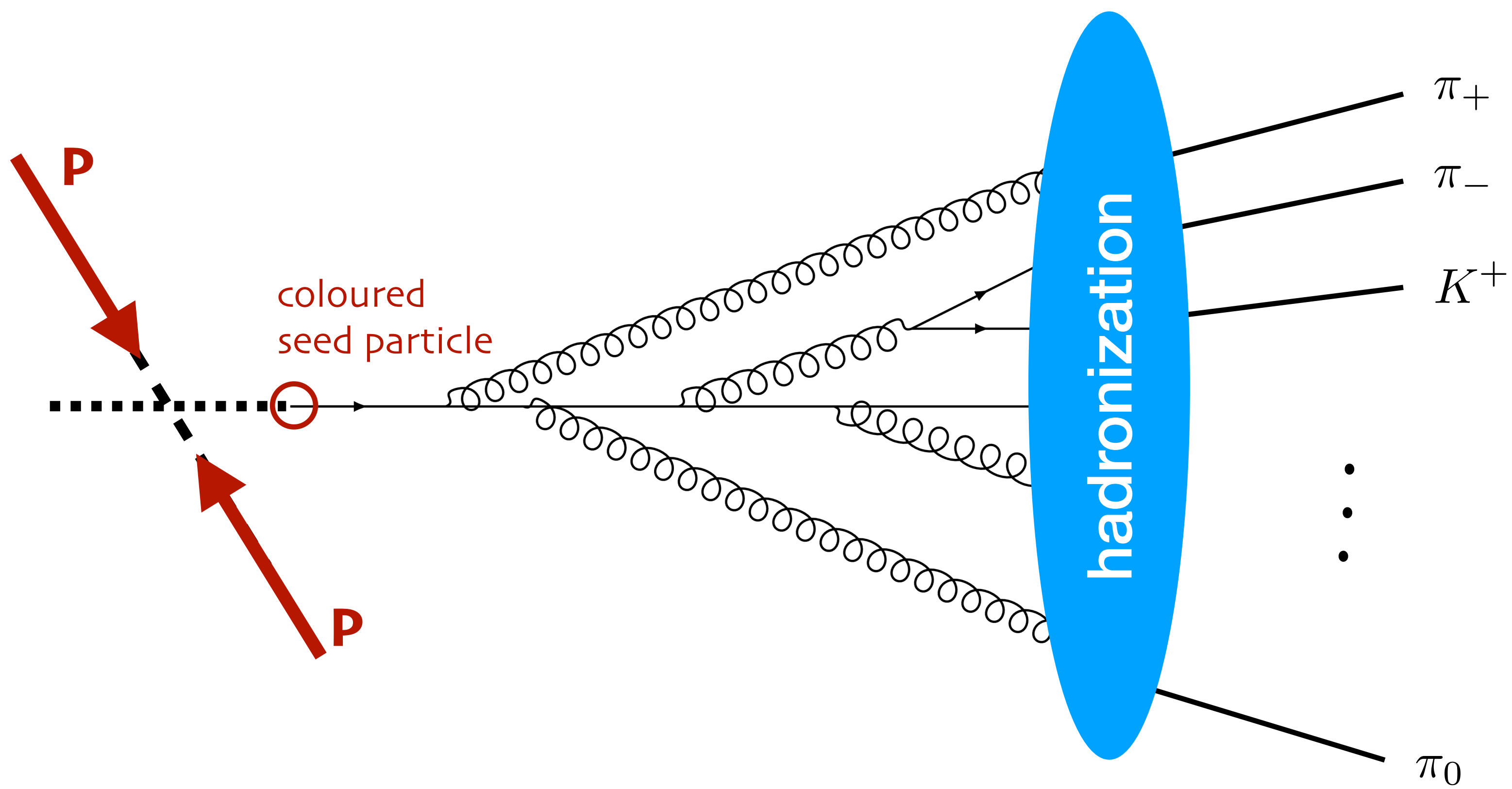
back-propagation

Supervised top-tagging

Most popular testing ground for ML tools in high-energy physics.

The task: to classify QCD jets from top-jets.

QCD jet:

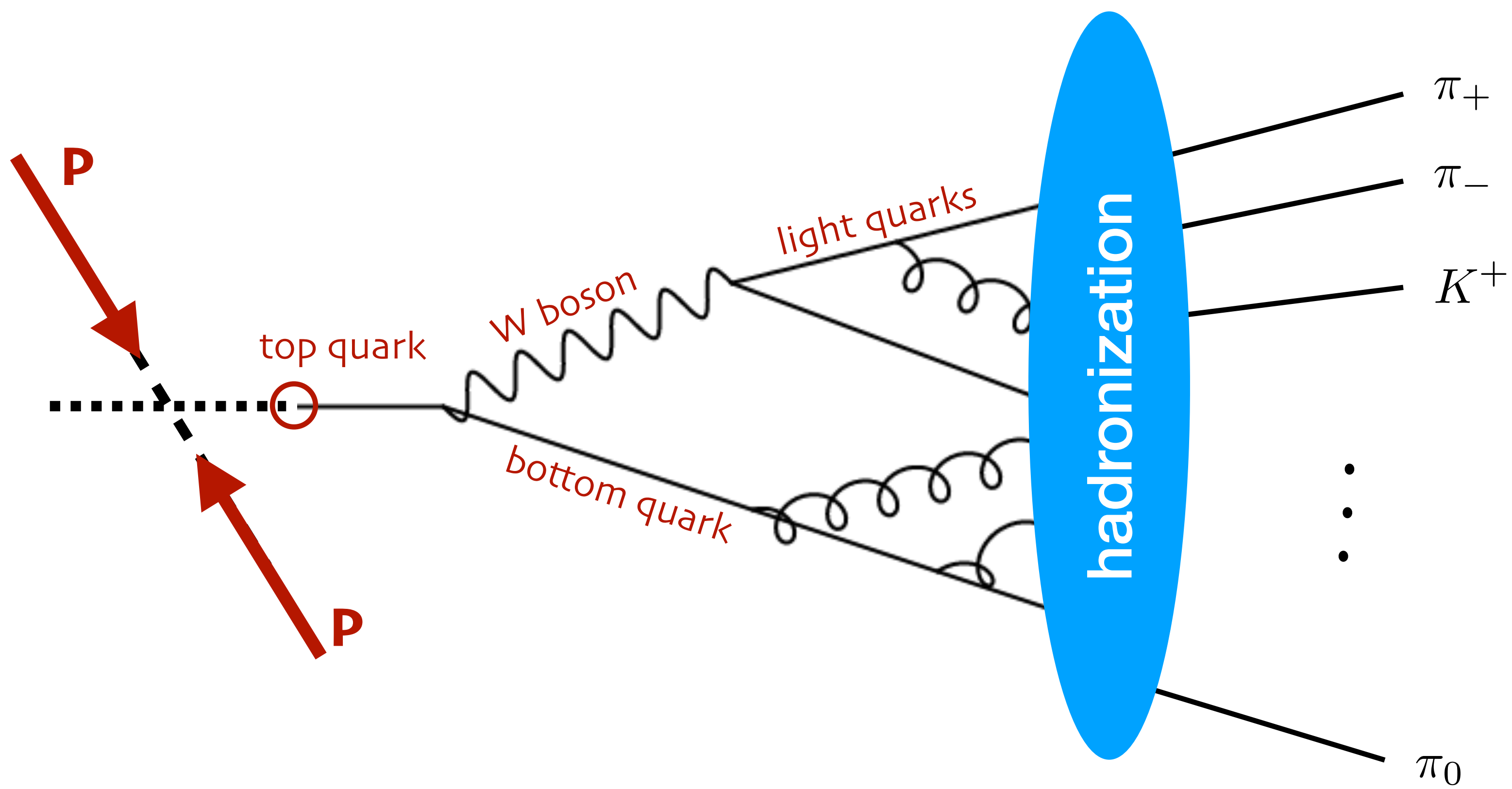


Supervised top-tagging

Most popular testing ground for ML tools in high-energy physics.

The task: to classify QCD jets from top-jets.

Top jet:

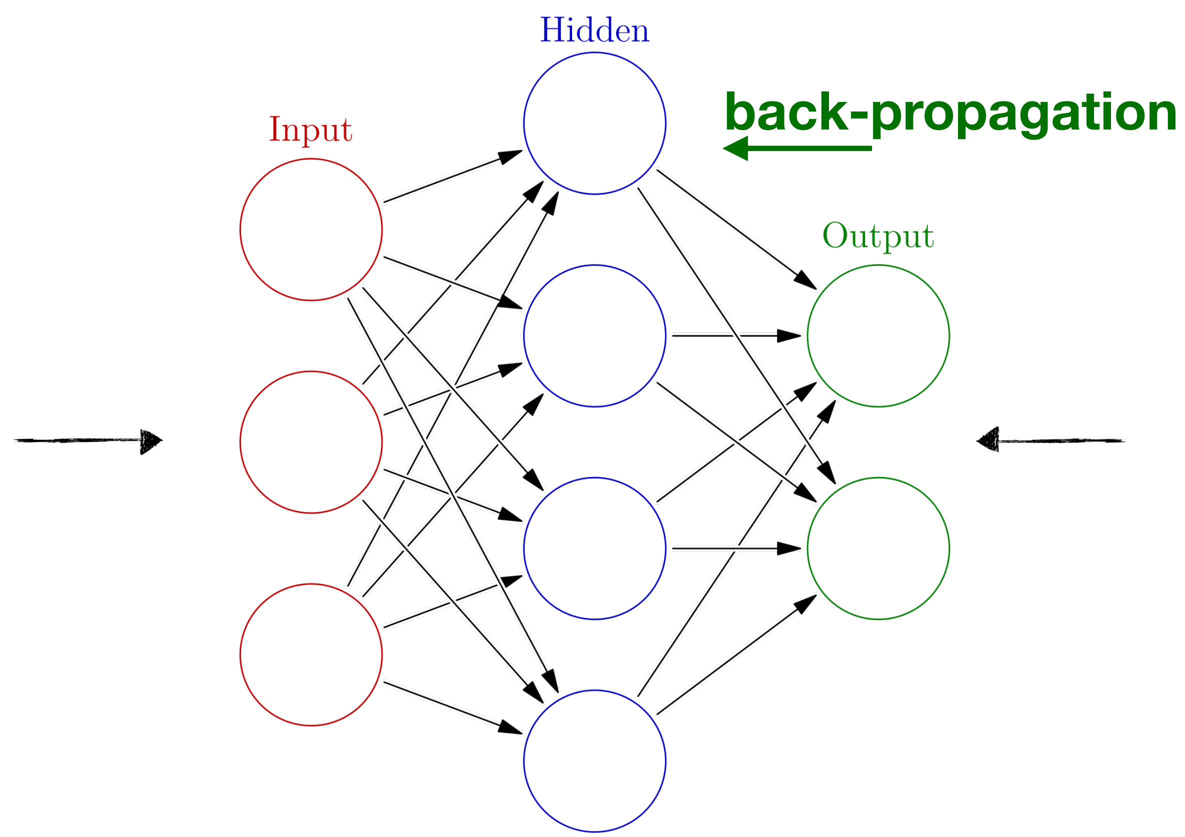


Traditional approach: study substructure in kinematics of final state particles

Supervised top-tagging

Machine-learning approach: input low-level kinematical data into a neural-network

Data obtained from simulation

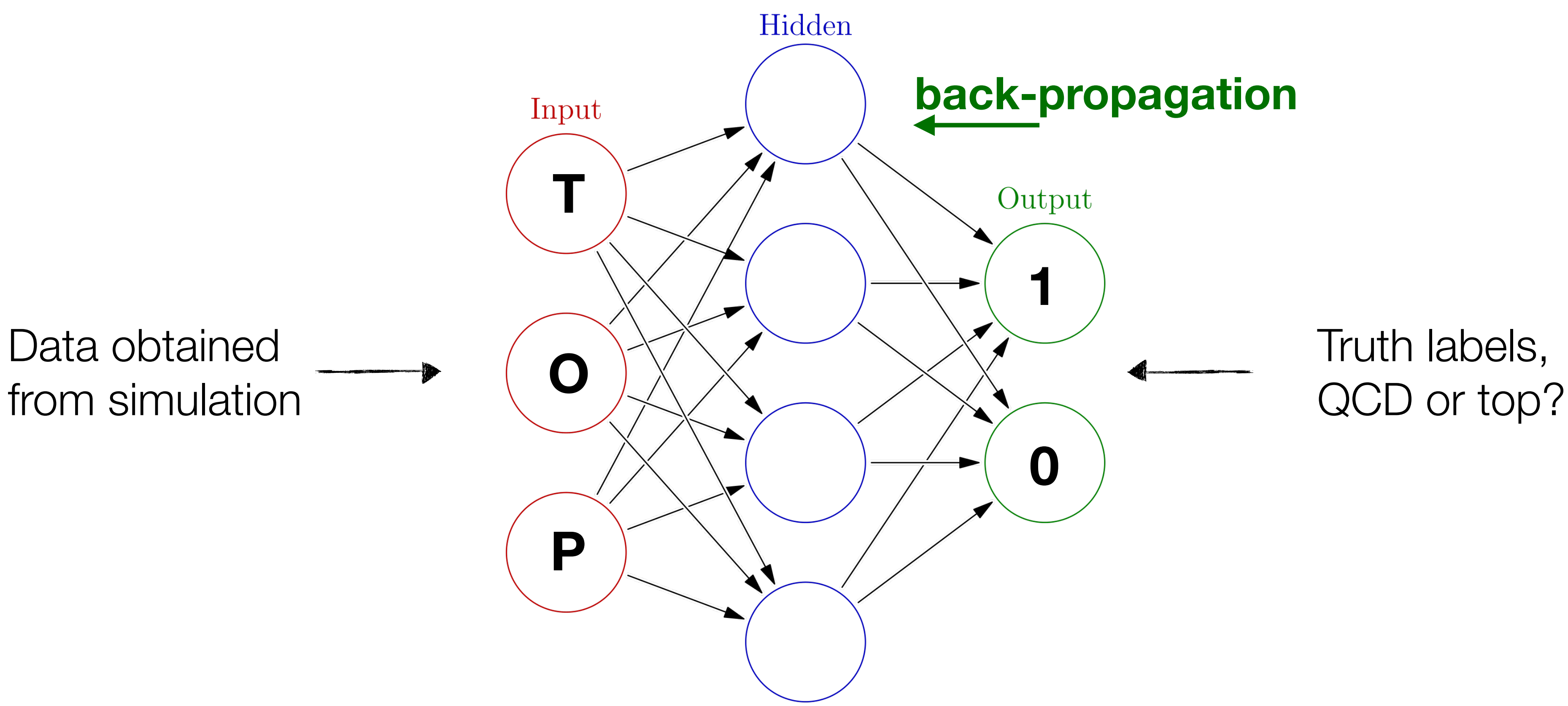


Truth labels, QCD or top?

Training: The hidden layer then learns the mapping, and given new unseen data can predict whether it came from a QCD and top jet.

Supervised top-tagging

Machine-learning approach: input low-level kinematical data into a neural-network

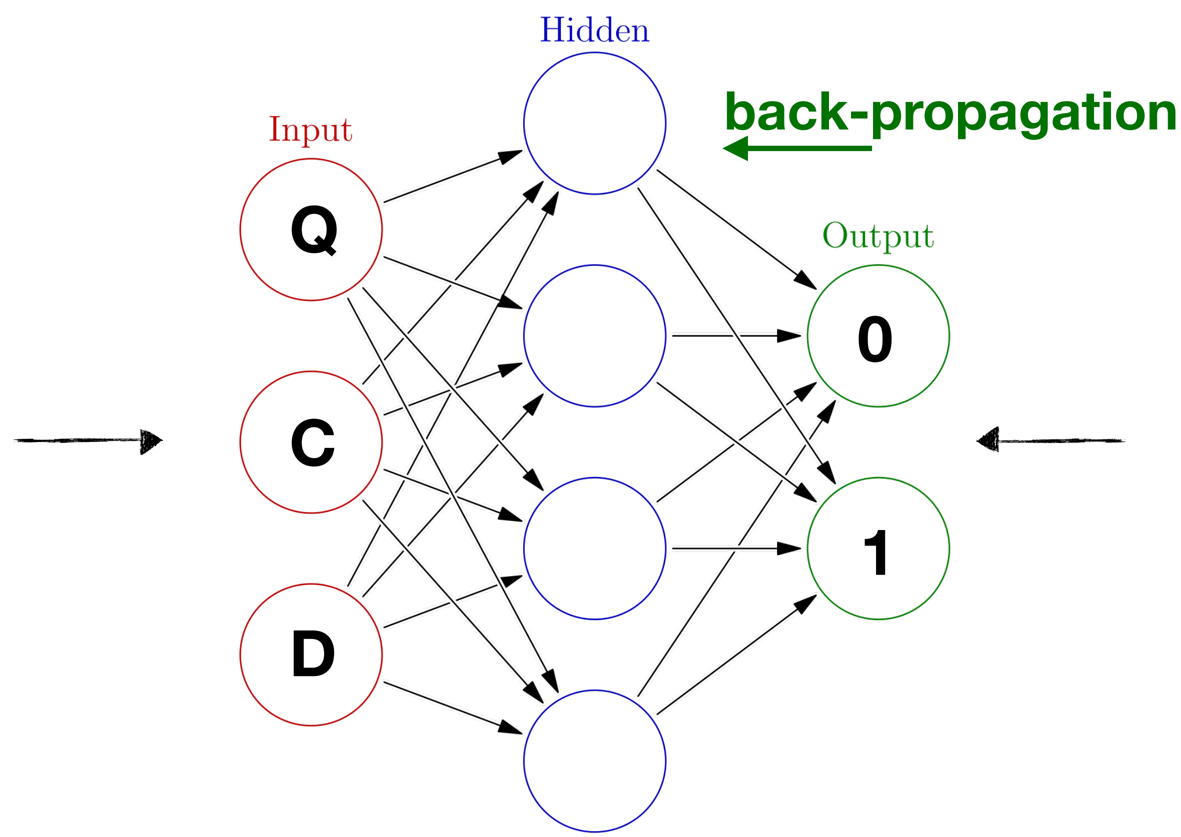


Training: The hidden layer then learns the mapping, and given new unseen data can predict whether it came from a QCD and top jet.

Supervised top-tagging

Machine-learning approach: input low-level kinematical data into a neural-network

Data obtained from simulation



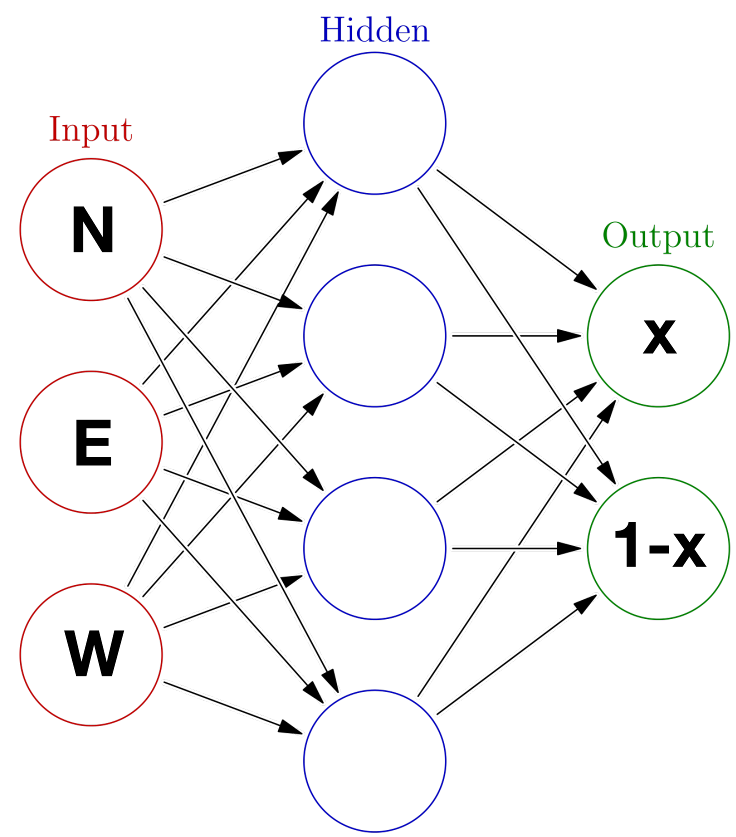
Truth labels, QCD or top?

Training: The hidden layer then learns the mapping, and given new unseen data can predict whether it came from a QCD and top jet.

Supervised top-tagging

Machine-learning approach: input low-level kinematical data into a neural-network

Data obtained from experiment

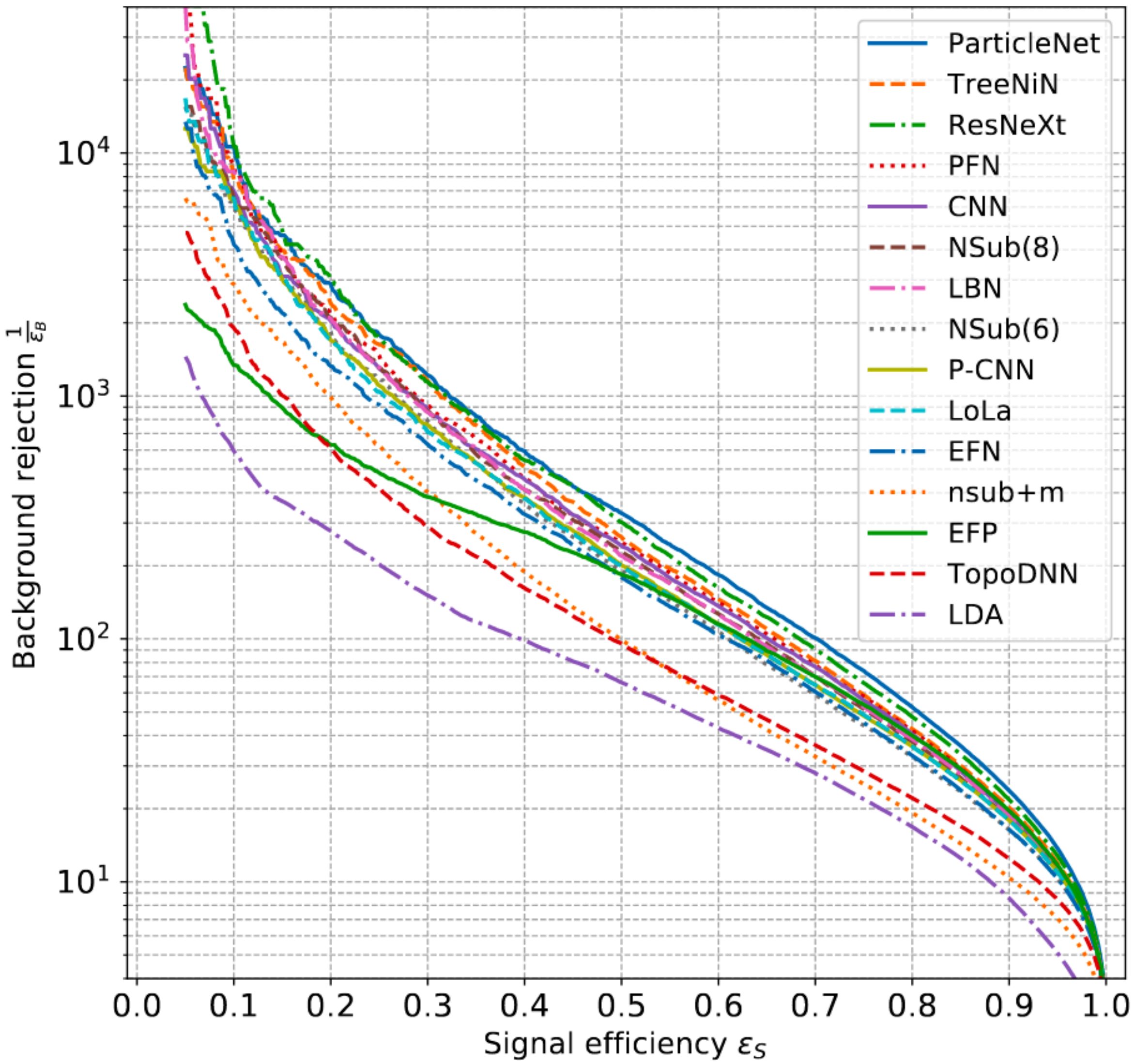


Truth labels, prediction

Training: The hidden layer then learns the mapping, and given new unseen data can predict whether it came from a QCD and top jet.

Supervised top-tagging

Scanning over ‘x’, and measuring the percentage of top-jets tagged, and the percentage of QCD-jets mis-tagged, gives us the Receiver Operating Characteristic (ROC) curve.



`Machine-learning landscape of top-taggers’
Kasieczka, Plehn et al
SciPost Phys. 7, 014 (2019)

Other applications

There are many other applications studied as well:

Quark/gluon tagging	Kasieczka, Kiefer, Plehn, Thompson: SciPost Phys. 6, 069
Recursive NNs for jets	Louppe, Cho, Becot, Cranmer: arXiv:1702.00748
Pile-up mitigation	Komiske, Metodiev, Nachman, Schwartz: JHEP 12 (2017) 051
Constraining EFTs with ML	Brehmer, Cranmer, Louppe, Pavez: Phys. Rev. D 98, 052004

Searching for new physics

ML taggers could be very important for NP searches.
Eg: could significantly improve searches for NP decaying to boosted top-jets.

However supervised algorithms suffer some serious drawbacks:

- they rely on accurate modelling of the event in simulations
- it is very difficult to know ‘what the machine has learned’
- they require a-priori knowledge of the what the signal is
- for every signal a new algorithm needs to be designed

All of these can be addressed using an **unsupervised** machine learning approach.

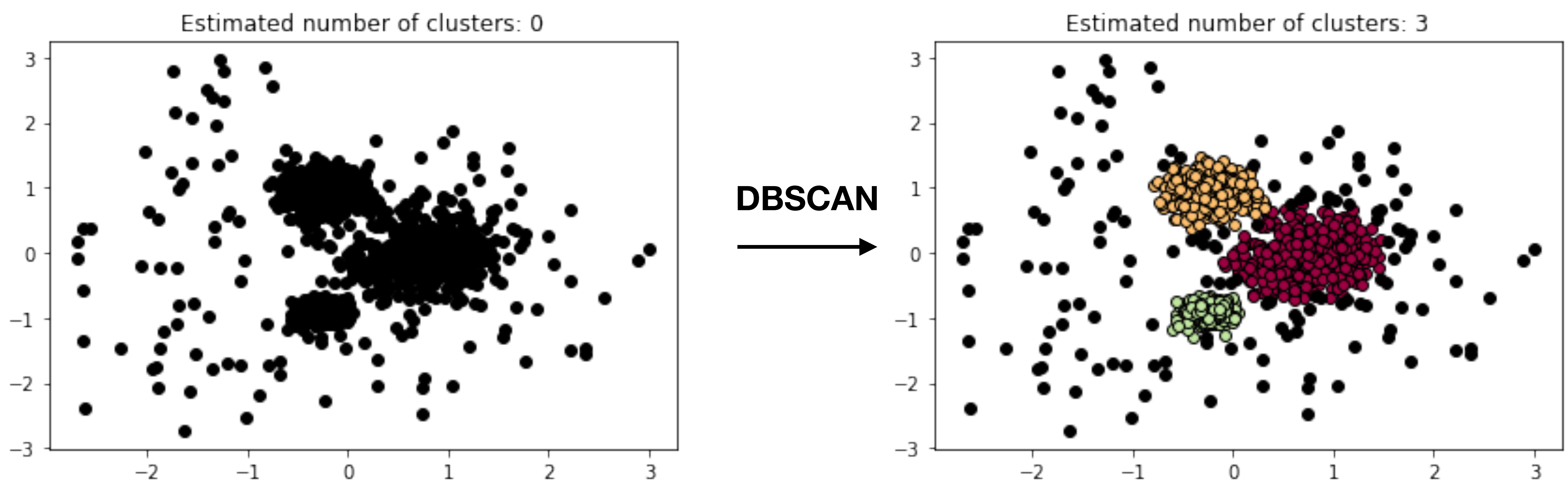
A new approach for new physics

Unsupervised ML

Wikipedia-style overview

Unsupervised learning: an algorithm that helps find previously unknown patterns in a data set without pre-existing labels.

Simplest example: clustering algorithms.

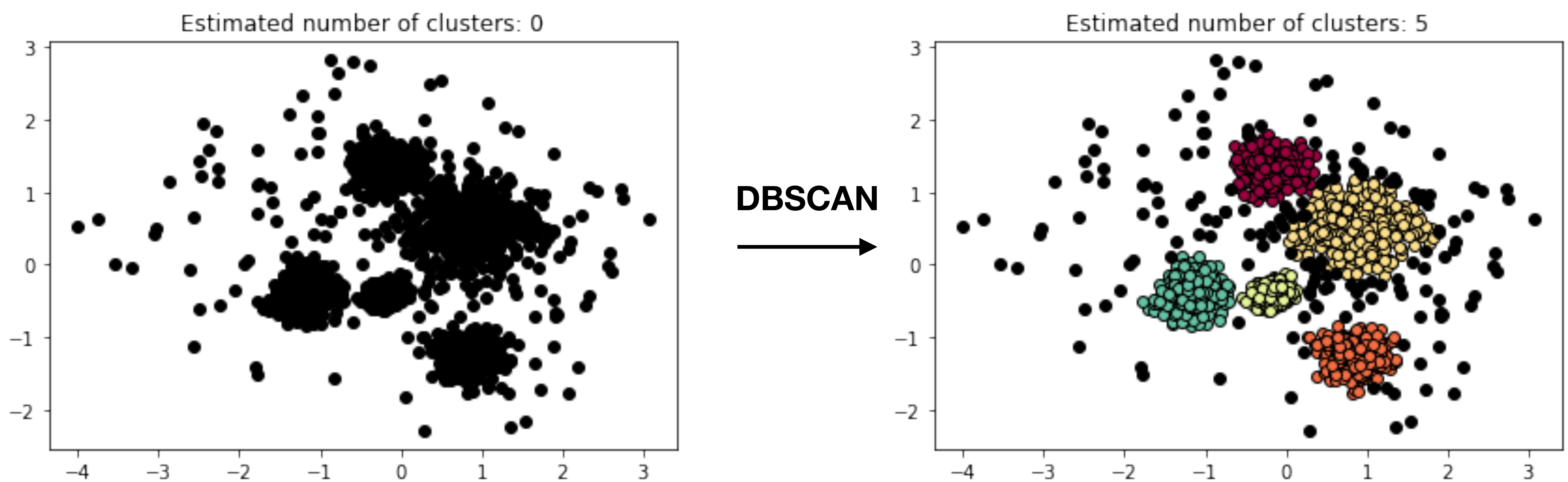


Unsupervised ML

Wikipedia-style overview

Unsupervised learning: an algorithm that helps find previously unknown patterns in a data set without pre-existing labels.

Simplest example: clustering algorithms.



The outline

The goals:

- identify events as signal or background without any prior knowledge on the how the events look

- do so in samples with small S/B.

The steps:

1. Construct a statistical model to parameterise physical processes in the events

2. Use inference algorithms to infer the parameters of the model from the data

3. Use the results of the inference to classify events

Building a model

Consider an event, represented by a list of measurements made on the event:

$$e_j = \{f_1, f_2, \dots, f_n\}$$

Suppose events can be generated either by signal or background processes, the model can be written as:

$$P(e_j|\omega, t) = \sum_{z=b,s} \omega_z \prod_{i=1}^{n_f} P(f_i|t_z)$$

It can help to think of the probability as a 'generative' model

Latent parameters

Parameterise the S/B

Latent probability distributions

Parameterise processes within signal and background events

Building a model

Consider an event, represented by a list of measurements made on the event:

$$e_j = \{f_1, f_2, \ldots, f_n\}$$

Now suppose events can be generated by a mixture of signal and background processes; this is the Latent Dirichlet Allocation (LDA) model:

$$P(e_j|\alpha, t) = \int d\omega \; \pi(\omega_z|\vec{\alpha}) \prod_{i=1}^{n_f} \sum_{z=b,s} \omega_z P(f_i|t_z)$$

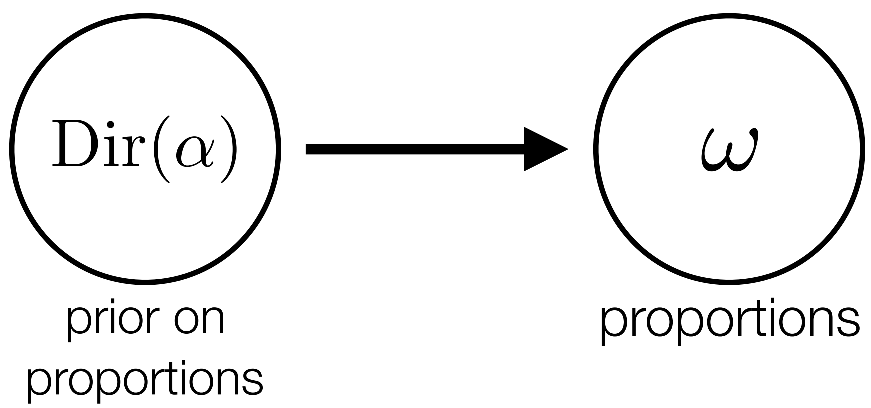
Prior on the relevance of signal and background processes within the event.

Latent parameters

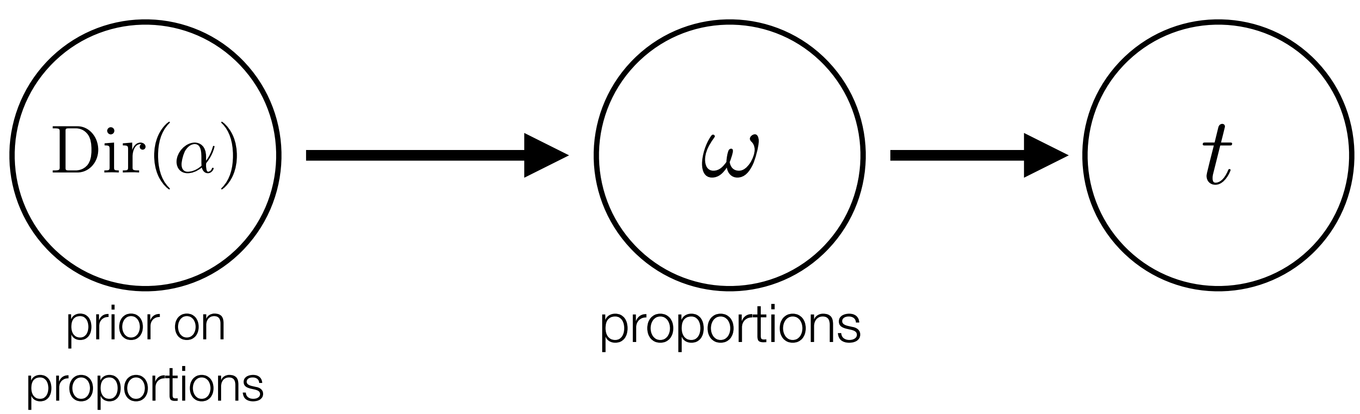
Latent probability distributions

Parameterise processes within signal and background events

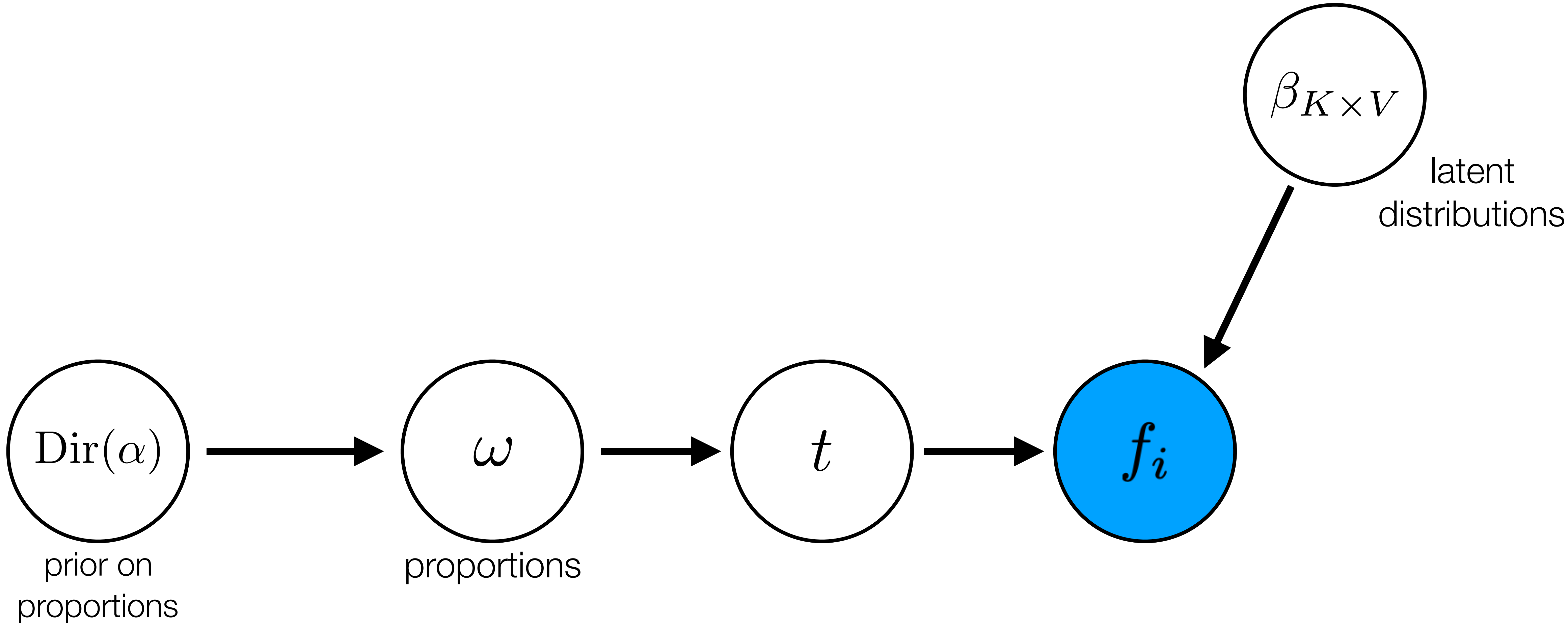
LDA as a generative model



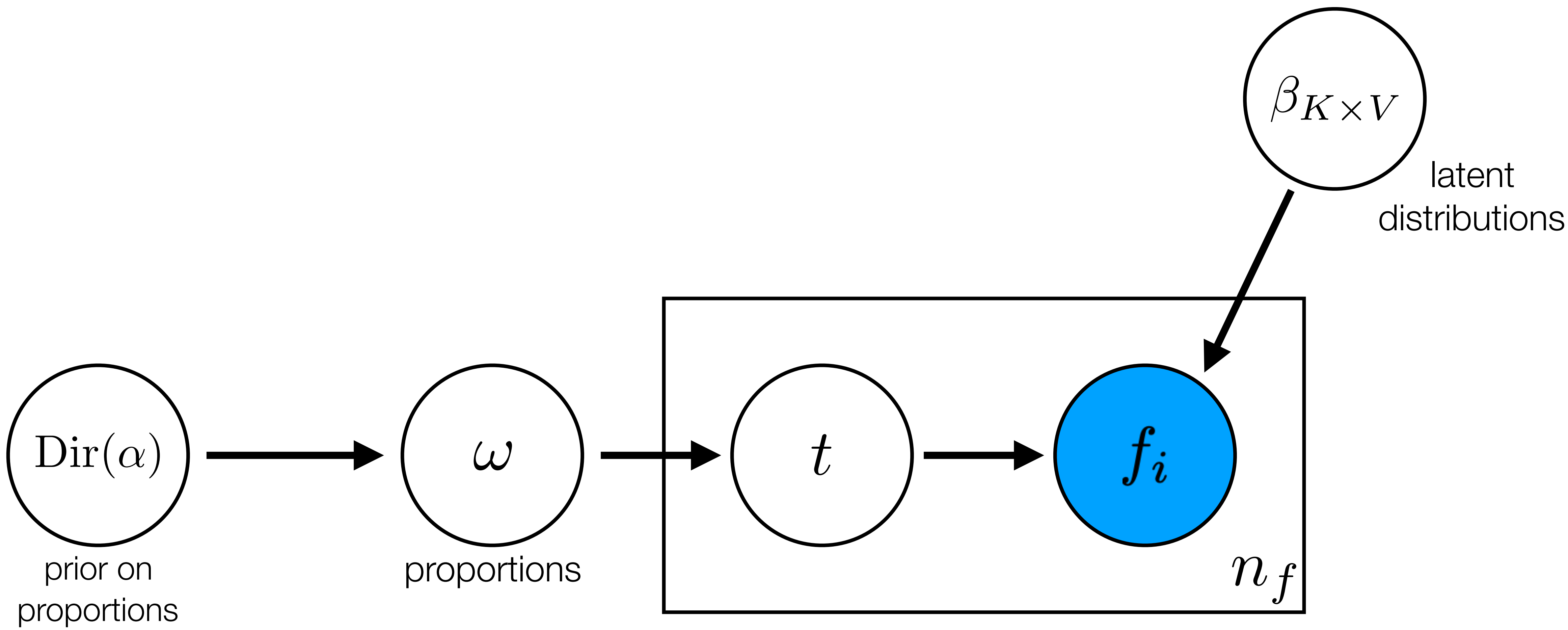
LDA as a generative model



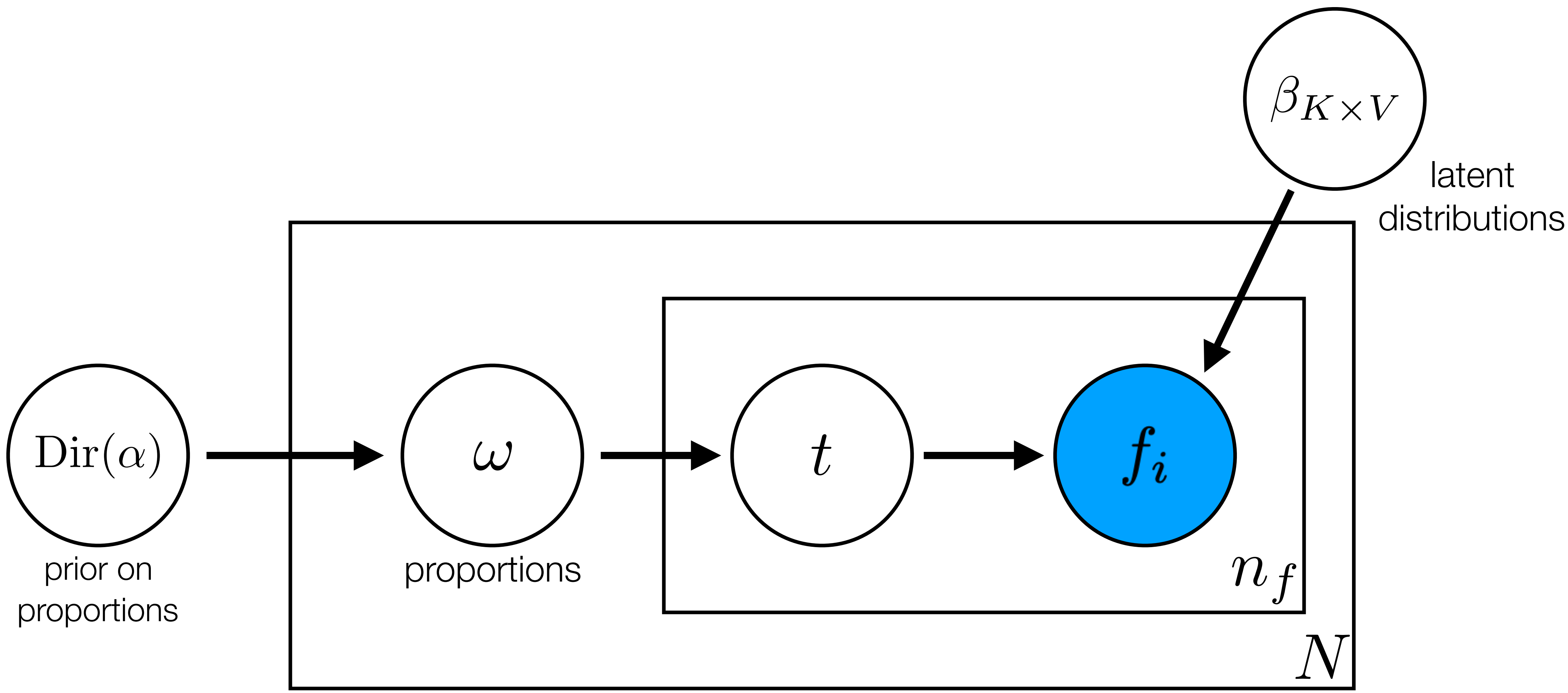
LDA as a generative model



LDA as a generative model



LDA as a generative model



Inference

Given the model, and the data: $\mathcal{D} = \{e_1, e_2, \ldots, e_{n_e}\}$

The latent distributions need to be extracted.
This is done through variational inference, a technique used to estimate the latent distributions that maximise the log-likelihood:

$$\log \prod_{j=1}^{n_e} P(e_j) = \sum_{j=1}^{n_e} \log P(e_j)$$

The success relies on **co-occurrences** of observables within the jet.
Observables which co-occur often, will have larger weights in the same latent distributions.

The prior on the proportions of signal and background processes is incredibly important for focusing the inference algorithm towards the extraction of rare processes. (more information in additional slides)

Classification

Once we have extracted $P(f_i|t_b)$ & $P(f_i|t_s)$ we need to use these for classification. There are two methods:

1. Inference using the model

$$\hat{\omega}(e_j) = \underset{\omega}{\operatorname{argmax}} \left(P(e_j|\omega, t) \right)$$

   i.e. using the proportions of processes inferred in the event.

2. Likelihood-ratio

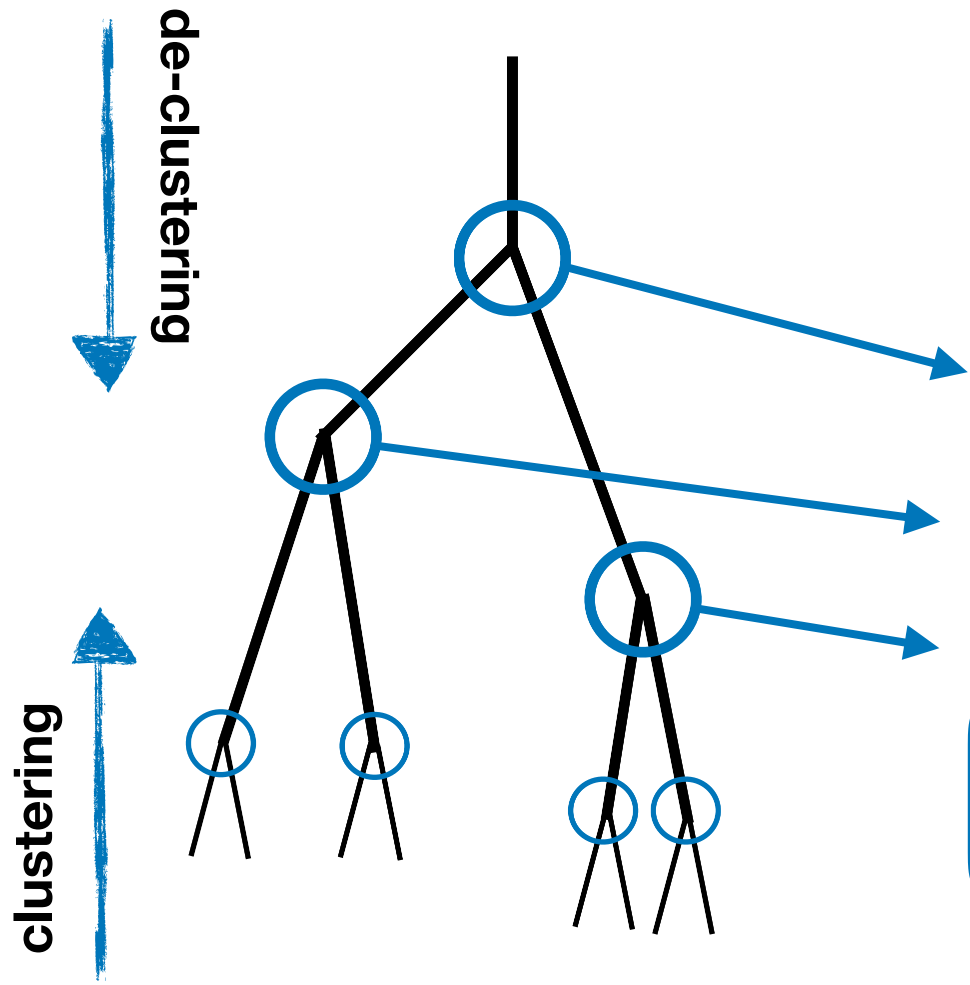
$$L(e_j) = L(f_1, \ldots, f_{n_f}) = \frac{\prod_{i=1}^{n_f} P(f_i|t_s)}{\prod_{i=1}^{n_f} P(f_i|t_b)}$$

We can classify and construct ROC curves using these test-statistics.

Uncovering latent jet substructure

Modelling jets with LDA

The only thing to decide upon is the representation of the observables.



$$\text{feature1} = \left[m_{\text{child}}, \frac{m_{\text{parent1}}}{m_{\text{child}}}, \frac{m_{\text{parent2}}}{m_{\text{parent1}}}, \Delta R_{\text{parents}} \right]$$

$$\text{feature2}$$

$$\ldots$$

$$\boxed{\text{jet} = \{\text{feature1}, \text{feature2}, \text{feature3}, \ldots\}}$$

The jets and latent distributions are defined over this space.

Unsupervised top-tagging

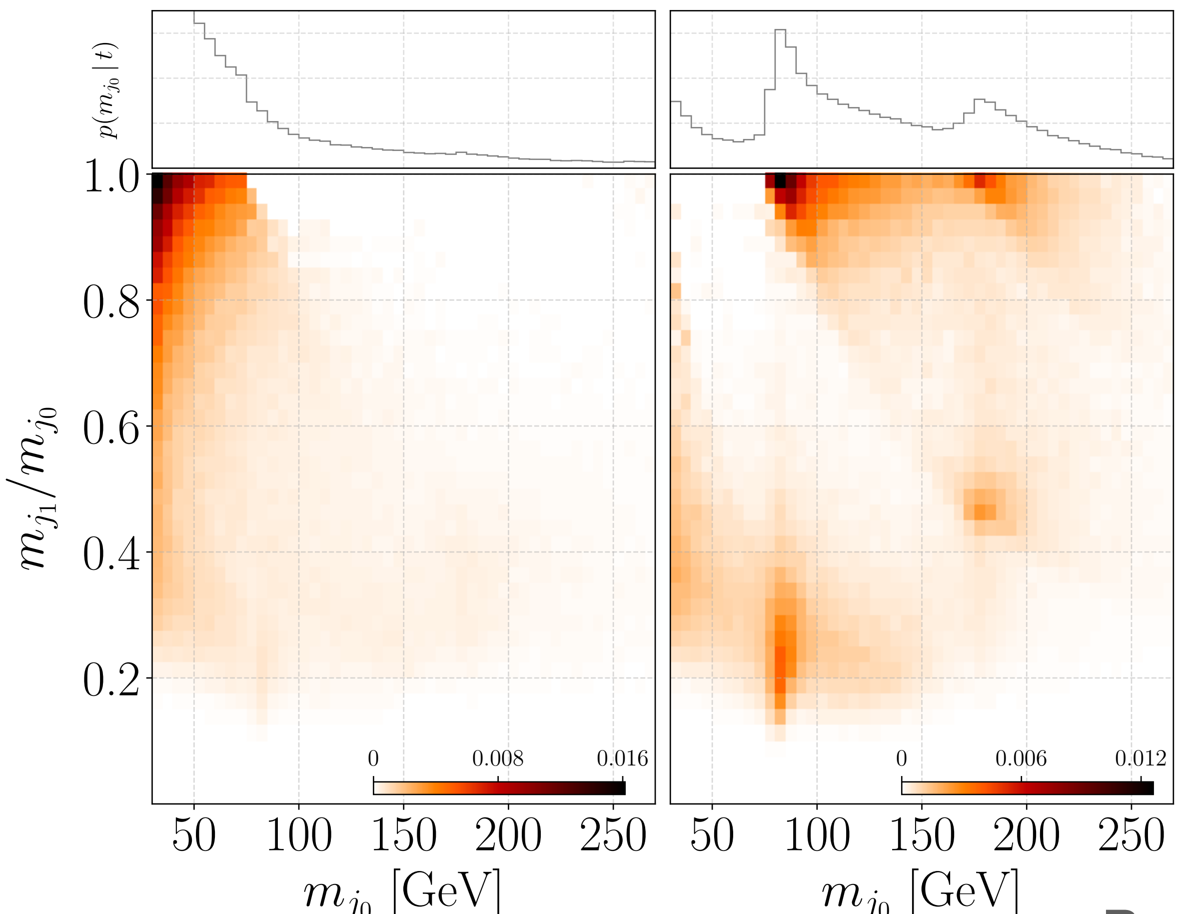
Proof-of-principle test: unsupervised classification of ttbar events.

The challenge: given a mixed, unlabelled sample of QCD and ttbar di-jet events, extract the signal and background latent distributions without any prior knowledge of what the signal is.

The latent distributions:

Subjet masses and mass drops in exactly the right places for the top jet signal!

Some sculpting in the distributions…



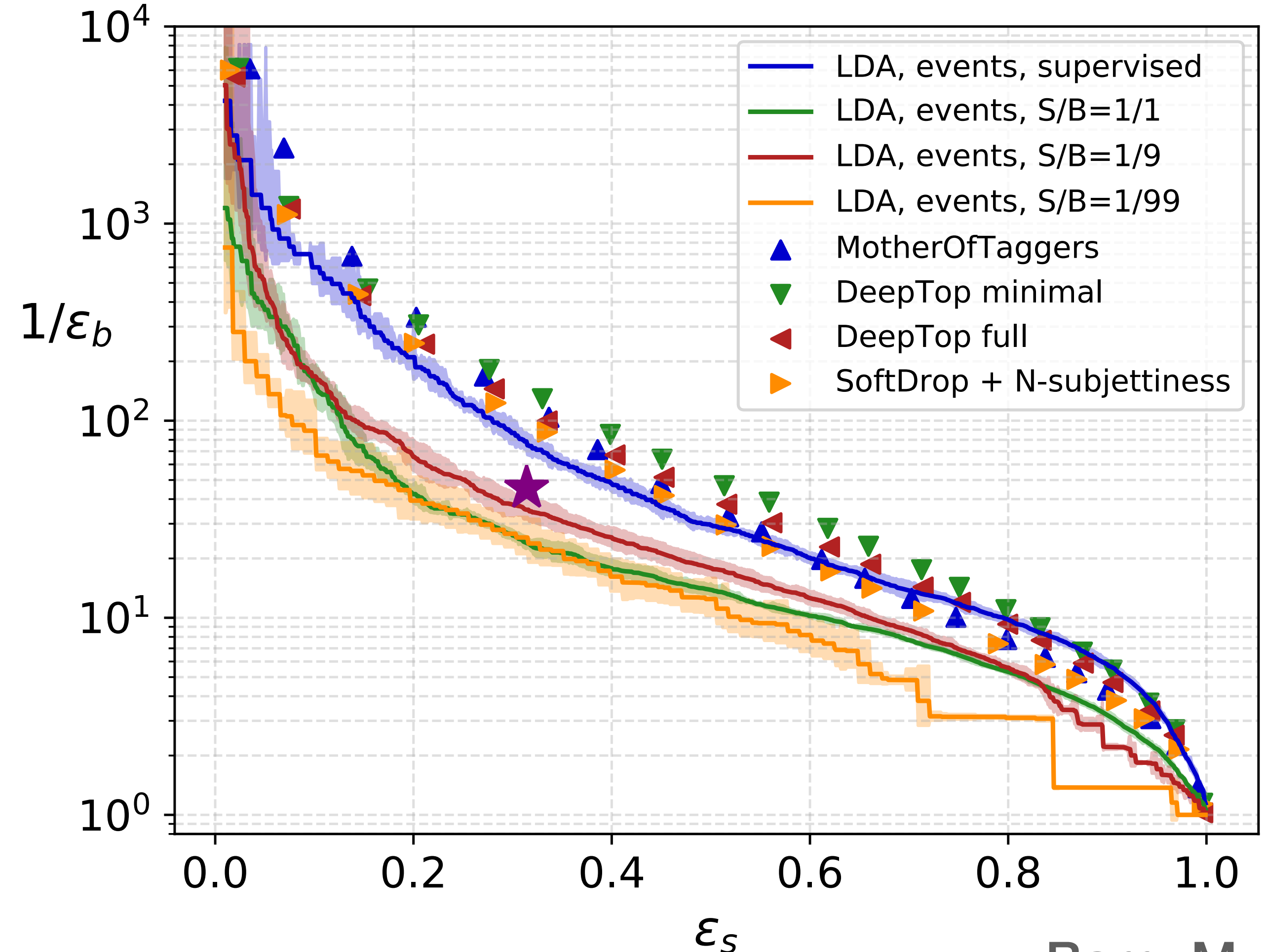
Unsupervised top-tagging

Proof-of-principle test: unsupervised classification of ttbar events.

The challenge: given a mixed, unlabelled sample of QCD and ttbar di-jet events, extract the signal and background latent distributions without any prior knowledge of what the signal is.

The classification power:

Performance equalling that of the HEP top-tagger, shown in the purple star.

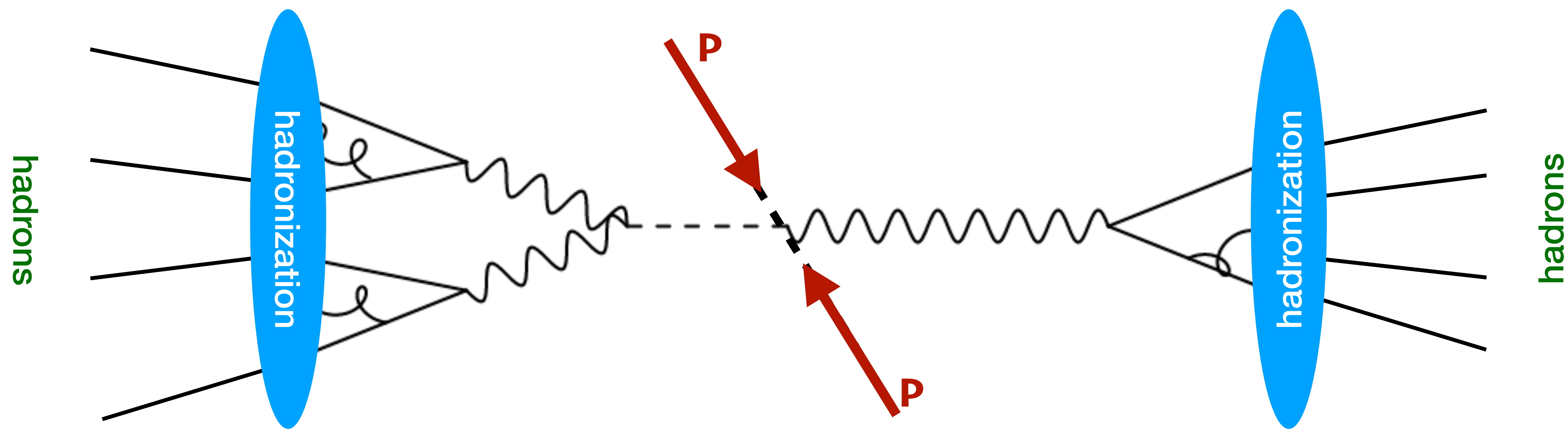


New physics extraction

There are well-known new physics signatures that aren't covered by traditional searches, such as the jets with a di-boson substructure.

For example: $W' \to W\phi \to WWW \to \text{jets}$

$$m_{W'} = 3\text{TeV},\ m_\phi = 400\text{GeV},\ m_W = 80\text{GeV}$$



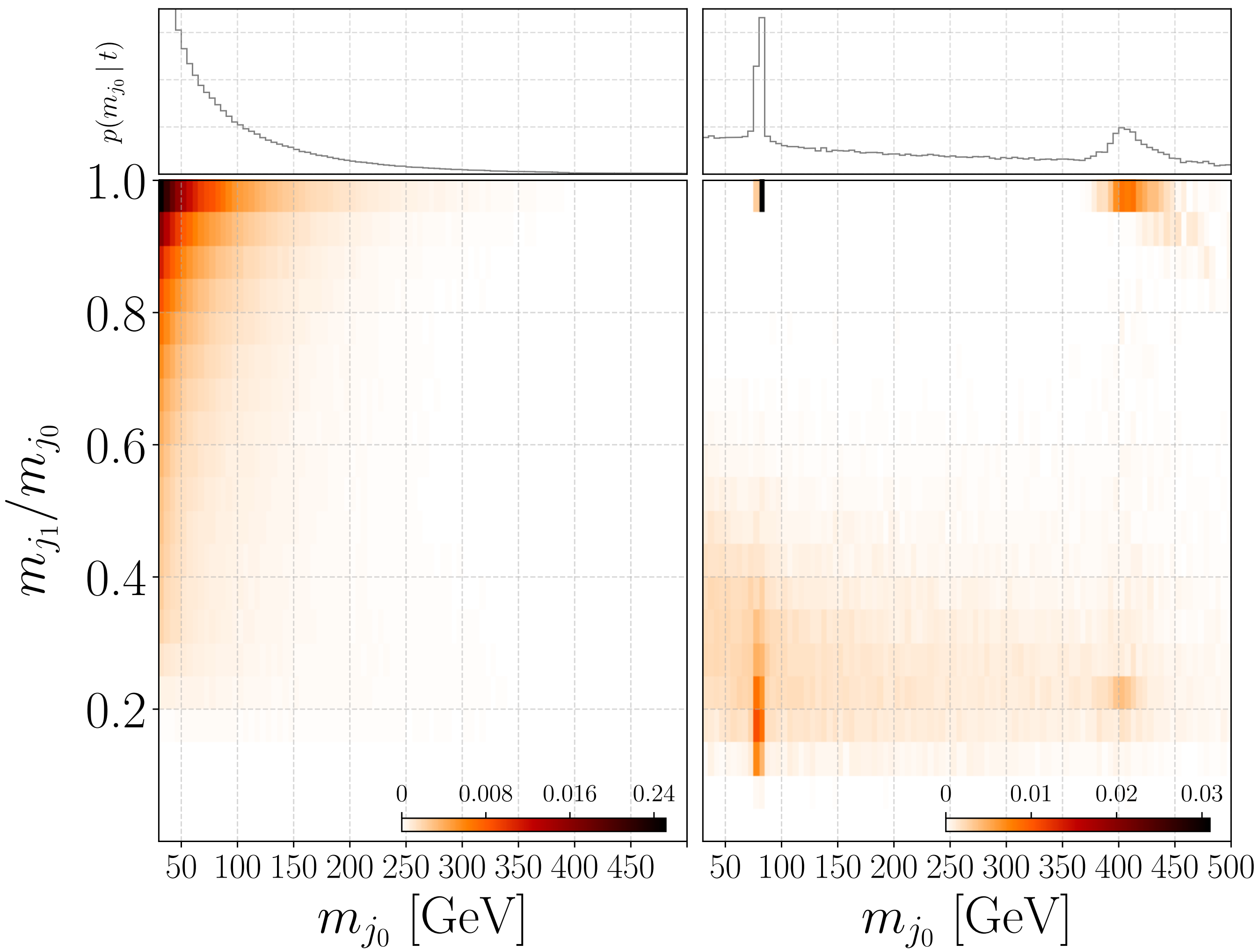
New physics extraction

The challenge: these signals are rare, so we must be able to extract the signal from samples with very small S/B

The set-up: we take a sample with 50,000 di-jet events, and S/B = 0.01 and 0.0058.

The latent distributions:

subjet masses and mass drops in exactly the right places for the W' signal!



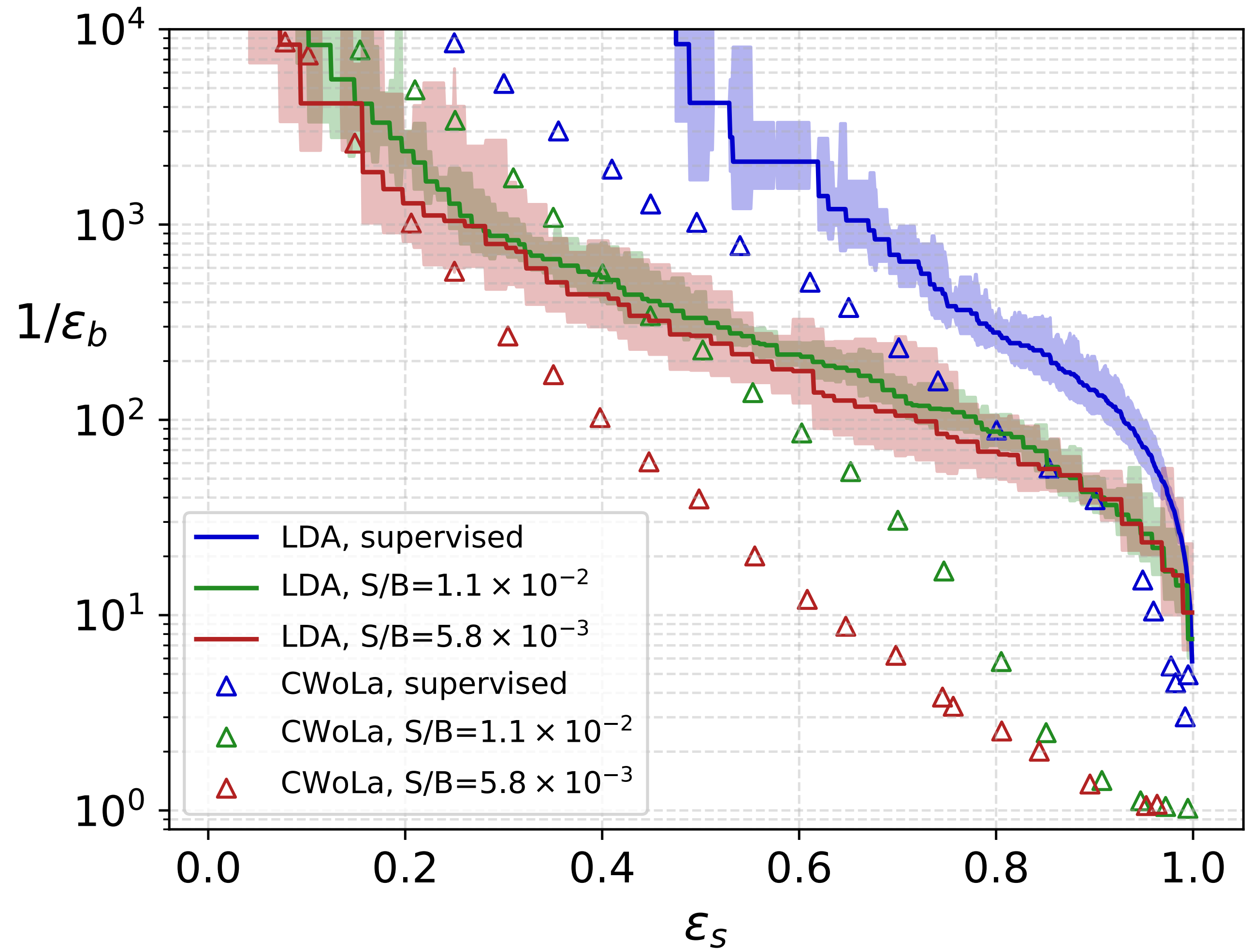
New physics extraction

The challenge: these signals are rare, so we must be able to extract the signal from samples with very small S/B

The set-up: we take a sample with 50,000 di-jet events, and S/B = 0.01 and 0.0058.

The classification power:

Results here compared to those from CWoLa



Concluding remarks

- The mixed-membership (LDA) model proves very successful in extracting rare signals from large datasets (at least for di-jet events).

- The signal needs to contain a substructure complex enough to provide the co-occurrences required for variational inference to work.

Next steps:

1. Construct statistical models to describe whole events:
   jets, isolated photons & leptons, missing energy, pile-up, …

2. Develop inference tools to extract latent parameters for these models.

3. Apply these methods on datasets from the CMS Open Data project.

Additional slides…

The Dirichlet distribution

- The prior on signal and background proportions is a Dirichlet distribution, and is conjugate to the binomial distribution

$$K=2 \quad \Rightarrow \quad \vec{\alpha}=[\alpha_0,\alpha_1] \qquad \& \qquad \mathrm{Dir}(\theta|\alpha_0,\alpha_1) = \frac{\Gamma(\alpha_0+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}\theta^{\alpha_0-1}(1-\theta)^{\alpha_1-1}$$

- The alpha parameters control the distribution of the signal and background features throughout the events

$\rho = \alpha_1/\alpha_0$ → controls the ratio of signal to background features in the whole sample

$$\int_0^1 d\theta\ \mathrm{Dir}\left(\theta|\alpha_0,\alpha_1\right)\left(\theta p_S(f_i)+(1-\theta)p_B(f_i)\right) = \frac{p_S(f_i)+\rho\ p_B(f_i)}{1+\rho}$$

$\Sigma = \alpha_0+\alpha_1$ → controls the distribution of signal and background features in each event

$\Sigma \ll 1$ → events mostly composed of a single process

$\Sigma \gg 1$ → events composed of a large mixture of processes

Finding the best model

- We need to find the 'best' model, without knowing what the signal or S/B is

  by 'best' I mean best choice of hyper-parameters

$$\text{perplexity} = e^{-\log \frac{P(\text{events}|\Sigma,\rho)}{N}}$$

- A model is good when the **perplexity is minimised**

- If the resulting model does not provide good classification on test samples

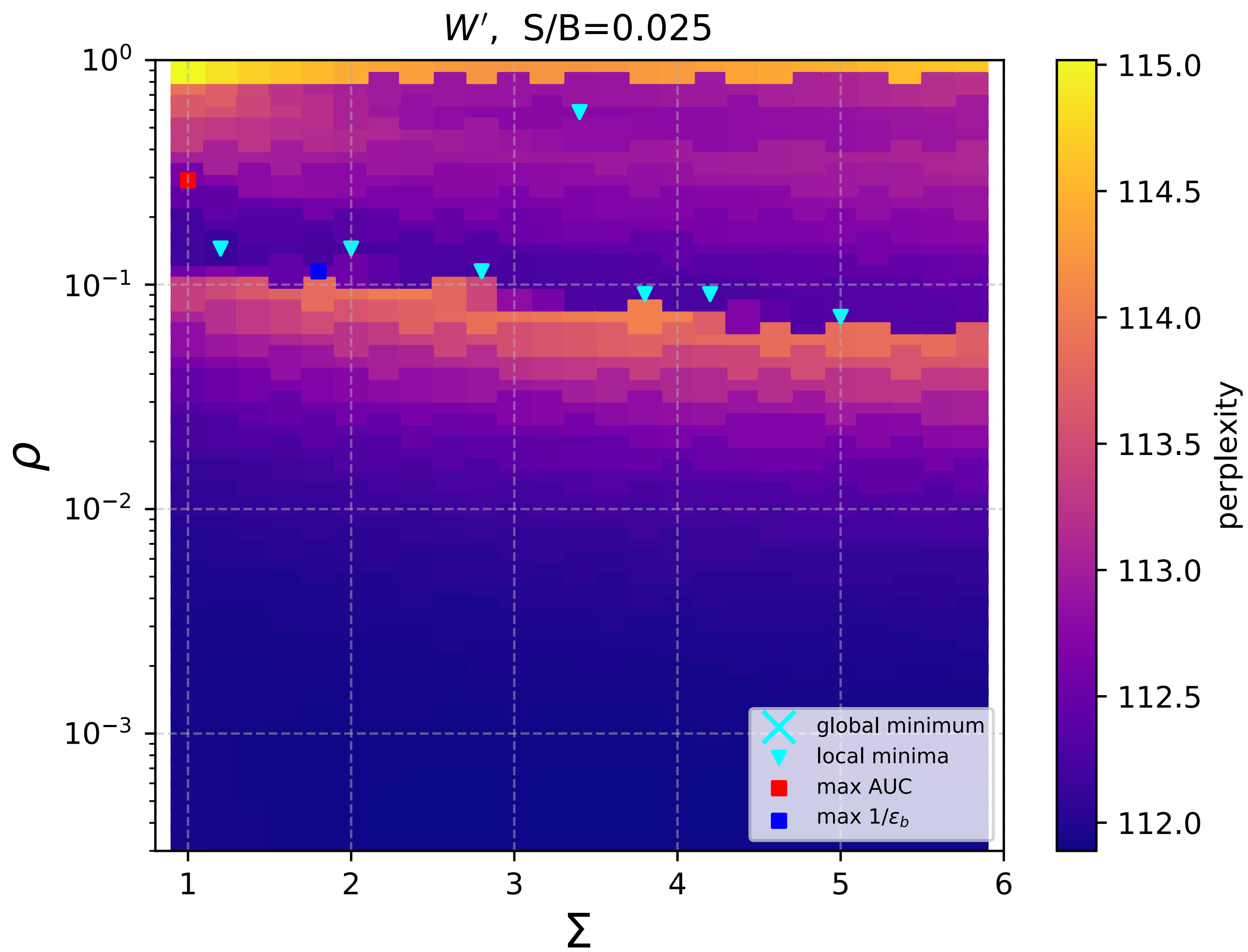
  - LDA does not work well for our physical scenario
  - our representation of the data is not optimal
  - the signal is just difficult to extract

Finding the best model

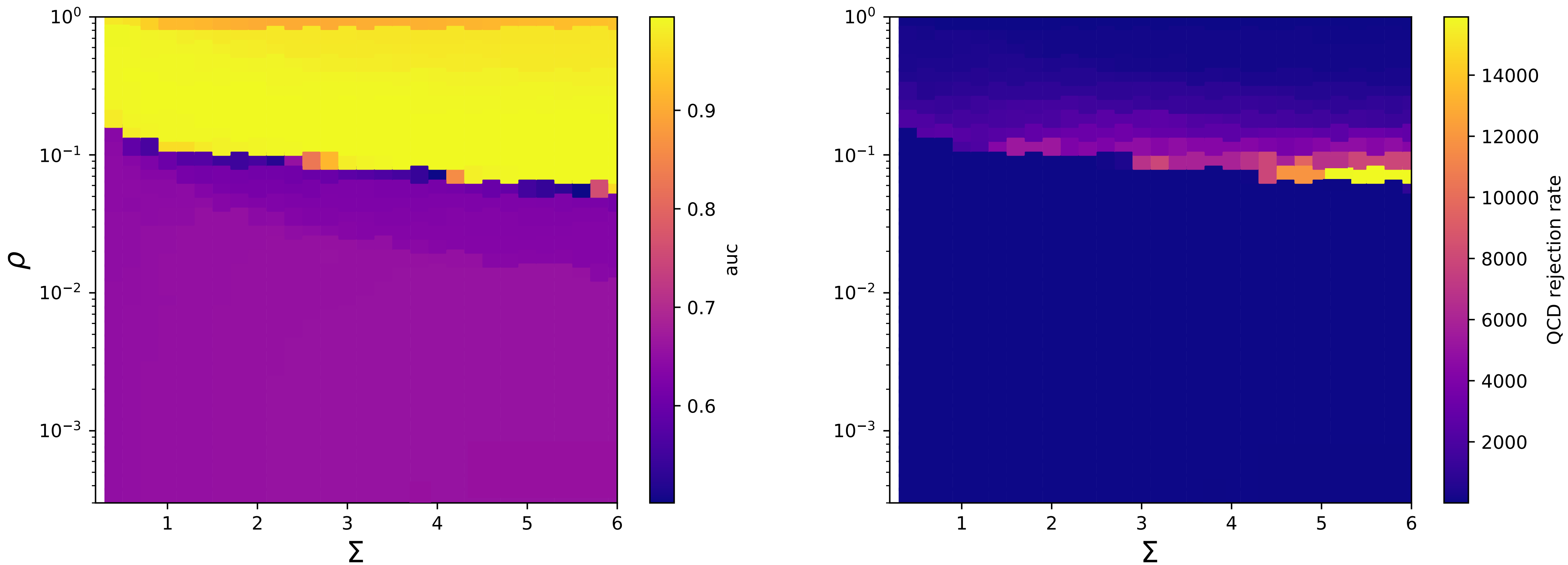
- We scan over the hyper-parameters:

lots of **local minima**, close to models with best AUC and best rejection rate at fixed mis-tag.

global minimum at vanishing rho, but this is a **trivial** solution.



Finding the best model



High-performance regions match those of (local) minimum perplexity.

Preliminary results