No classification without representation

Gregor Kasieczka
Email: gregor.kasieczka@uni-hamburg.de
Twitter: @GregorKasieczka
Mastodon: @gregor_k@sciencemastodon.com

AI and Quantum in Fundamental Physics

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

PUNCH4NFDI

DASHH

Emmy Noether-Programm
Deutsche Forschungsgemeinschaft DFG

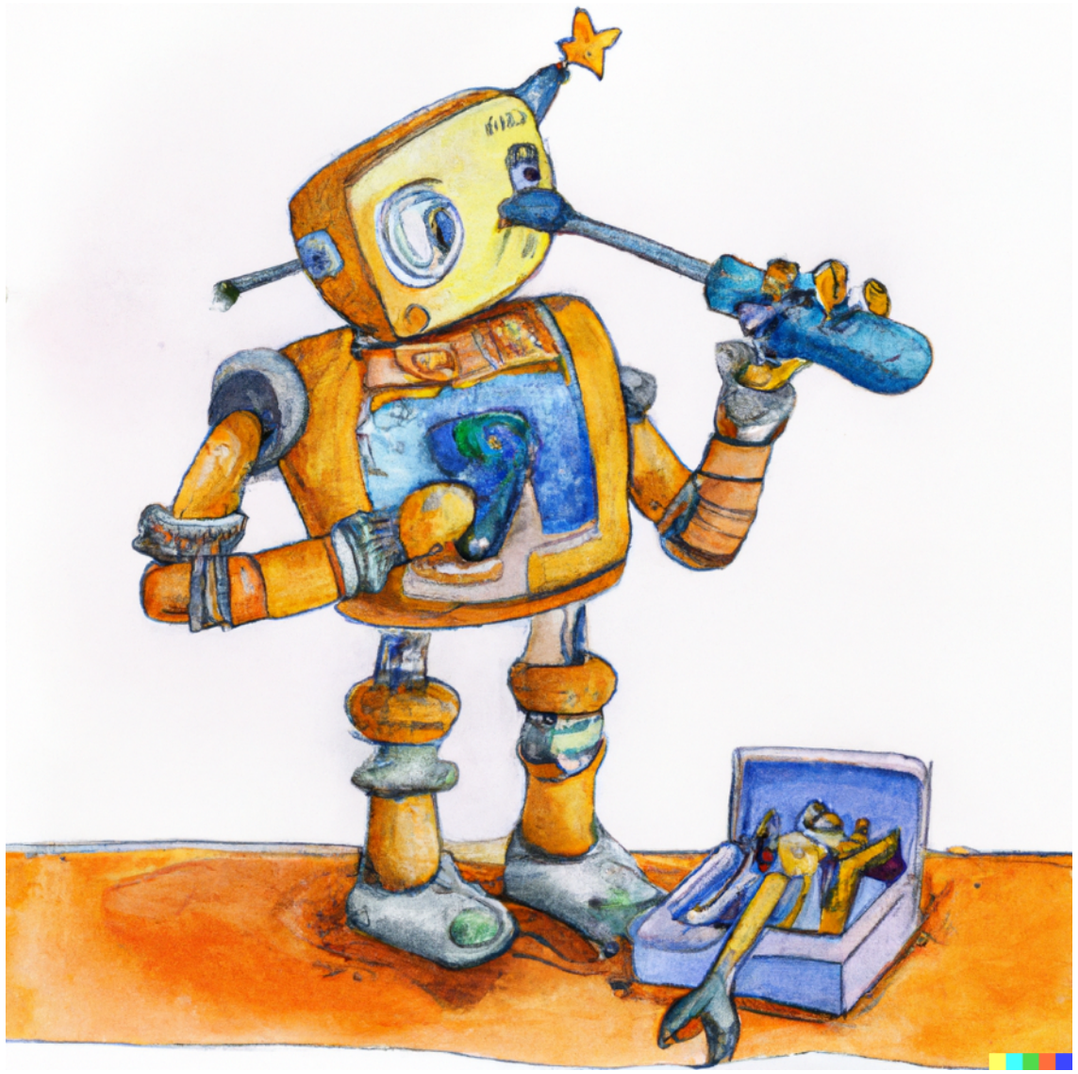
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CDCS
CENTER FOR DATA AND COMPUTING IN NATURAL SCIENCES

PIER
Partnership of Universität Hamburg and DESY

Bundesministerium für Bildung und Forschung

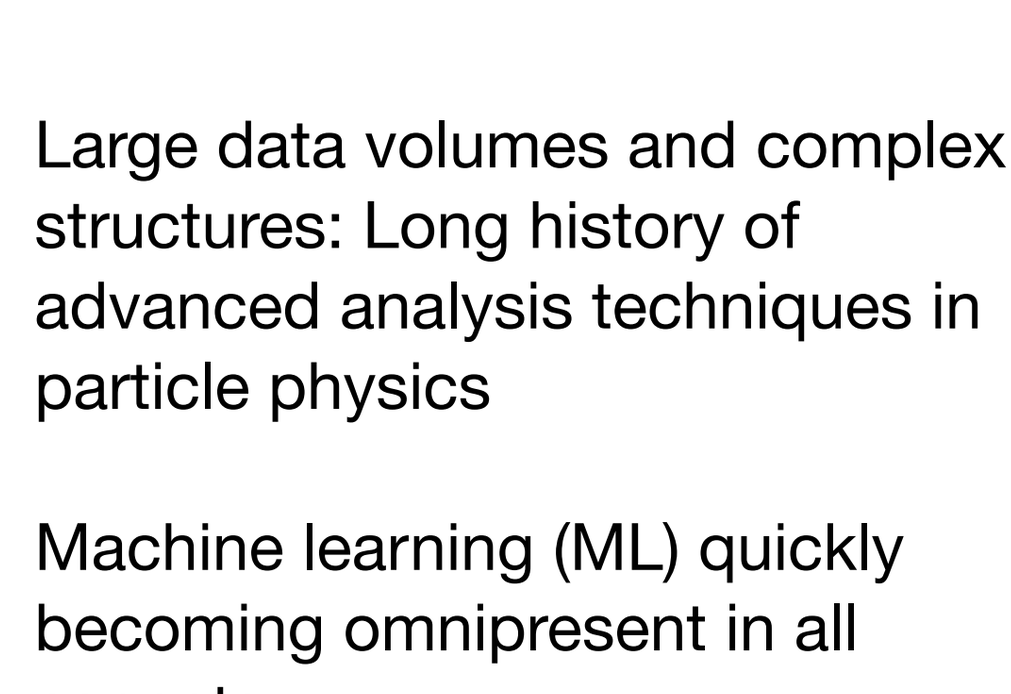
Introductions



State of the field

- Large data volumes and complex structures: Long history of advanced analysis techniques in particle physics

- Machine learning (ML) quickly becoming omnipresent in all aspects

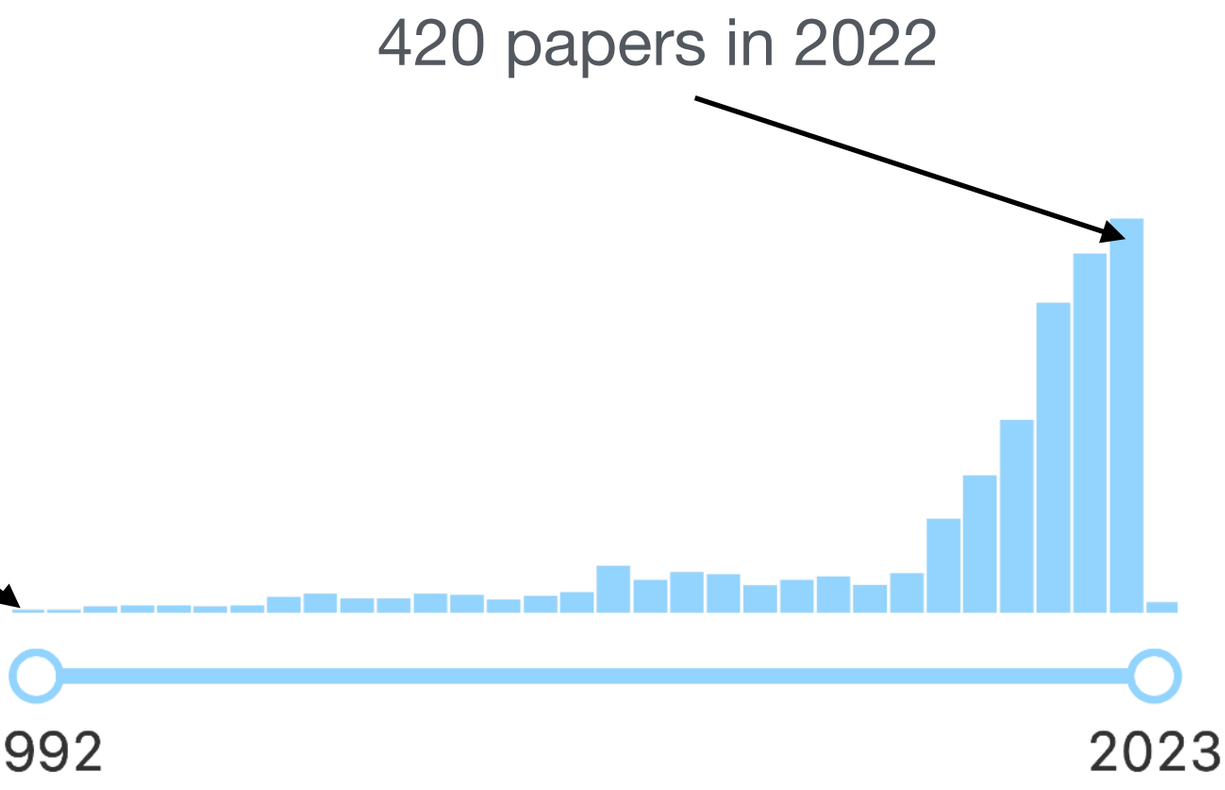


THE USE OF NEURAL NETWORKS
IN HIGH ENERGY PHYSICS*

BRUCE DENBY
Fermi National Accelerator Laboratory
M.S. 318
Batavia, Illinois 60510 U.S.A.
denby@fnal.bitnet

ABSTRACT

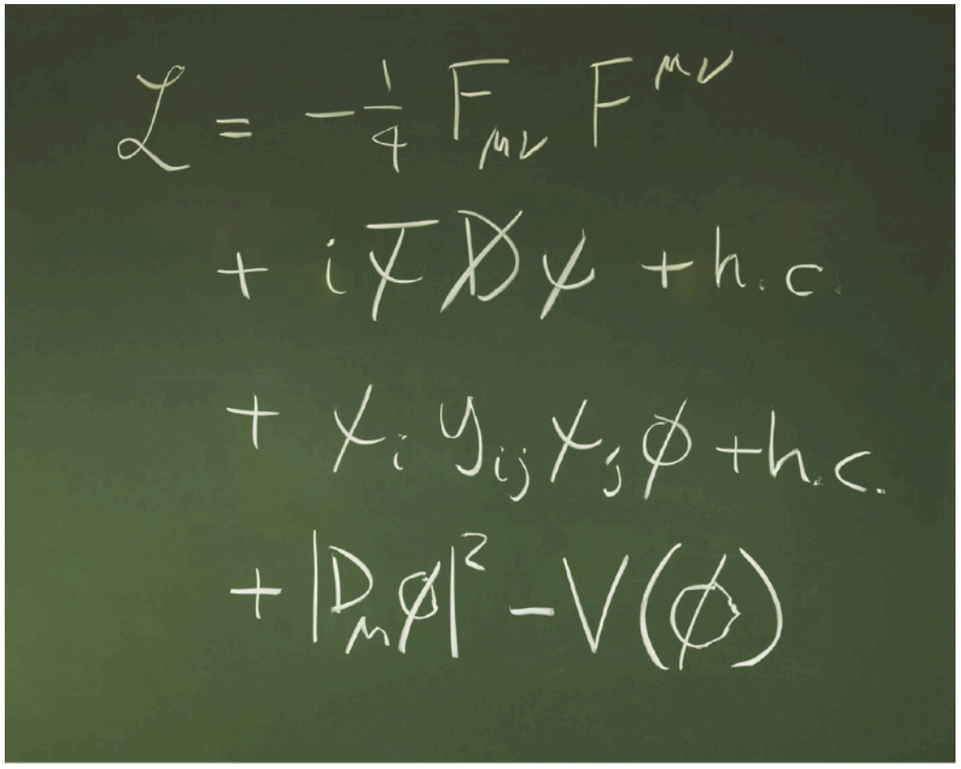
In the past few years a wide variety of applications of neural networks to pattern recognition in experimental high energy physics has appeared. The neural network solutions are in general of high quality, and, in a number of cases, are superior to those obtained using 'traditional' methods. But neural networks are of particular interest in high energy physics for another reason as well: much of the pattern recognition must be performed online, i.e., in a few microseconds or less. The inherent parallelism of neural network algorithms, and the ability to implement them as very fast hardware devices, may make them an ideal technology for this application.

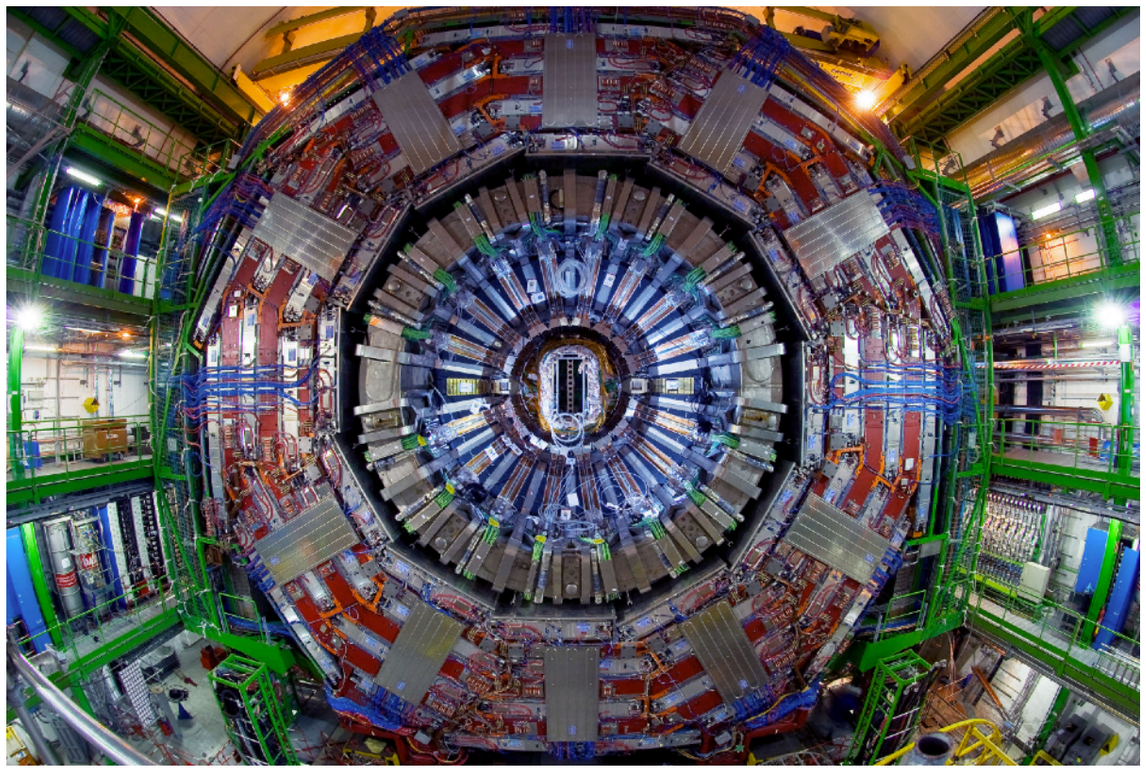


Inspire: ("machine learning" or "deep learning" or neural) and (hep-ex or hep-ph or hep-th)

Micro-Intro: Particle Physics

- Particle physics: study smallest constituents of matter
- Standard Model: incredible scientific achievement, describes 3/4 fundamental forces
- Mathematical, quantum theoretical understanding of matter at the smallest scales





- Experimental evidence (e.g. dark matter) & theoretical considerations: Standard Model is not sufficient, need *new physics*
- Comprehensive program at Large Hadron Collider (LHC)
- Experimental data is complemented by large volumes of high quality simulations (synthetic data)

Micro-Intro: Data

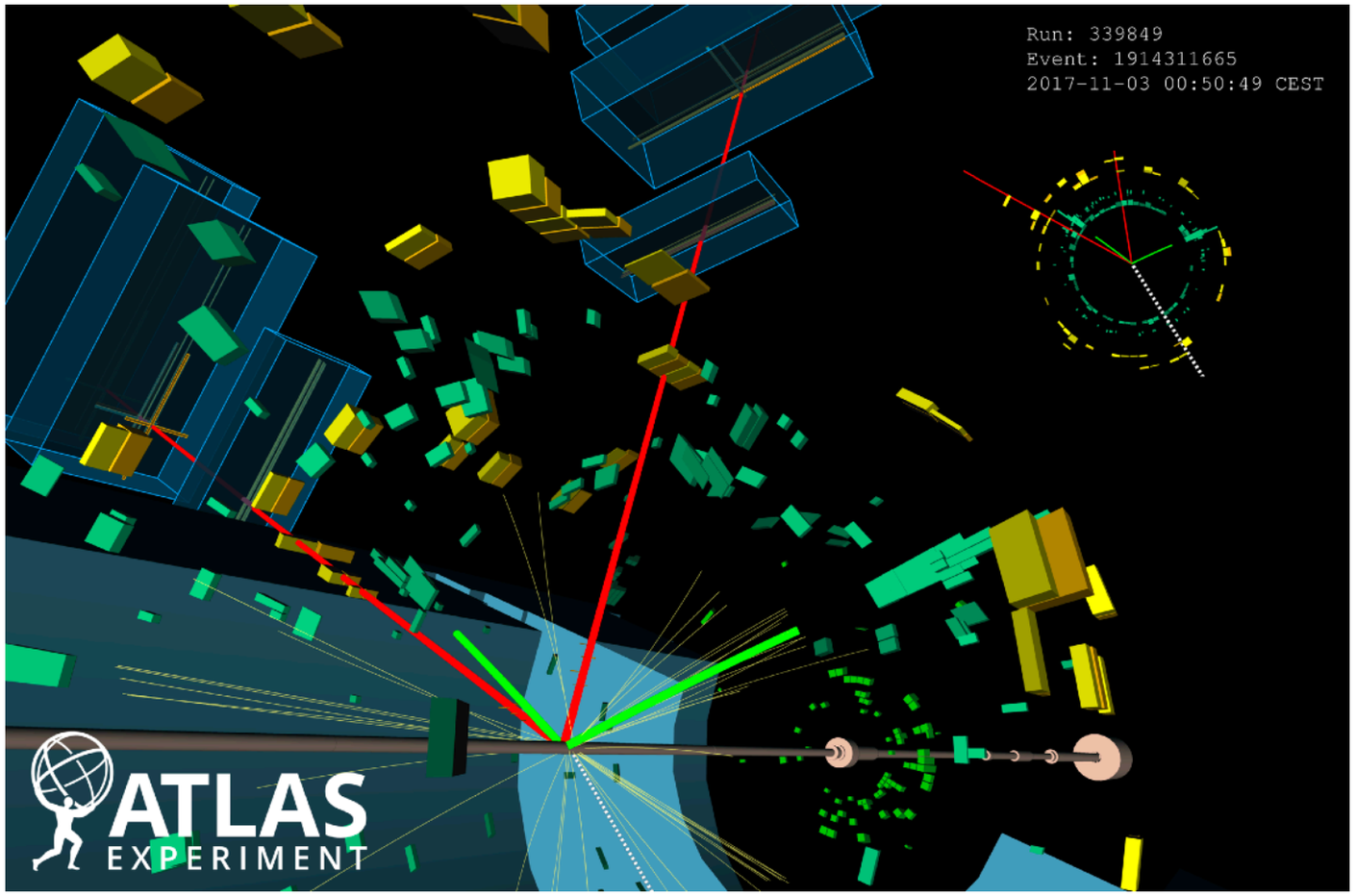
- Particle collisions with ~1 MB/ event happen at a rate of 40 MHz

- Distill to ~1 kHz via lossy, irreversible filtering algorithms (Trigger)

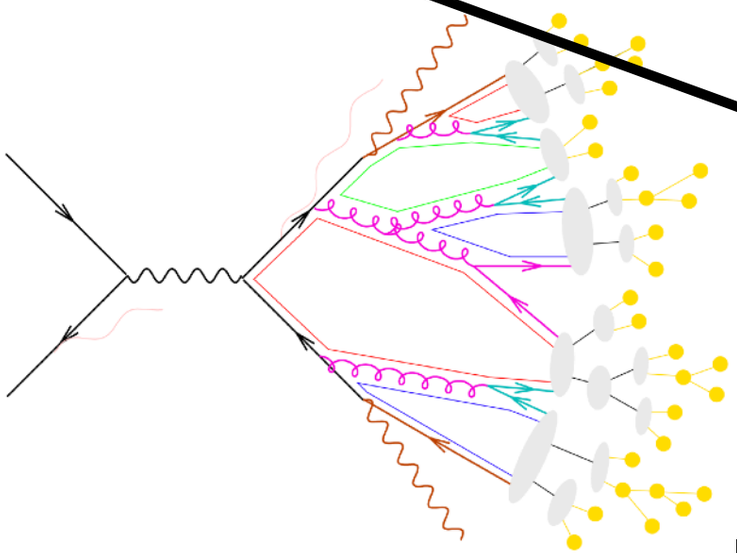
- Very heterogenous data: low-level readouts in ~100M channels; can condense to O(10) high-level features

- One collision/event = “one image”

  sample i.i.d. from underlying physics distribution (e.g. the Standard Model + potential new physics)

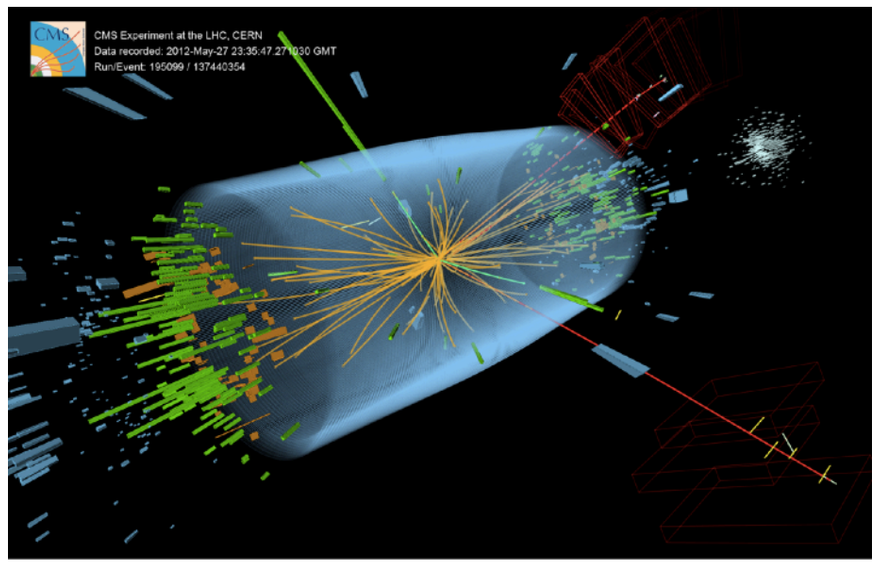


Experimental particle physics



This happens in the experiment

This is what we want to know



Connect observational data
with underlying theory:
Statistics & simulation

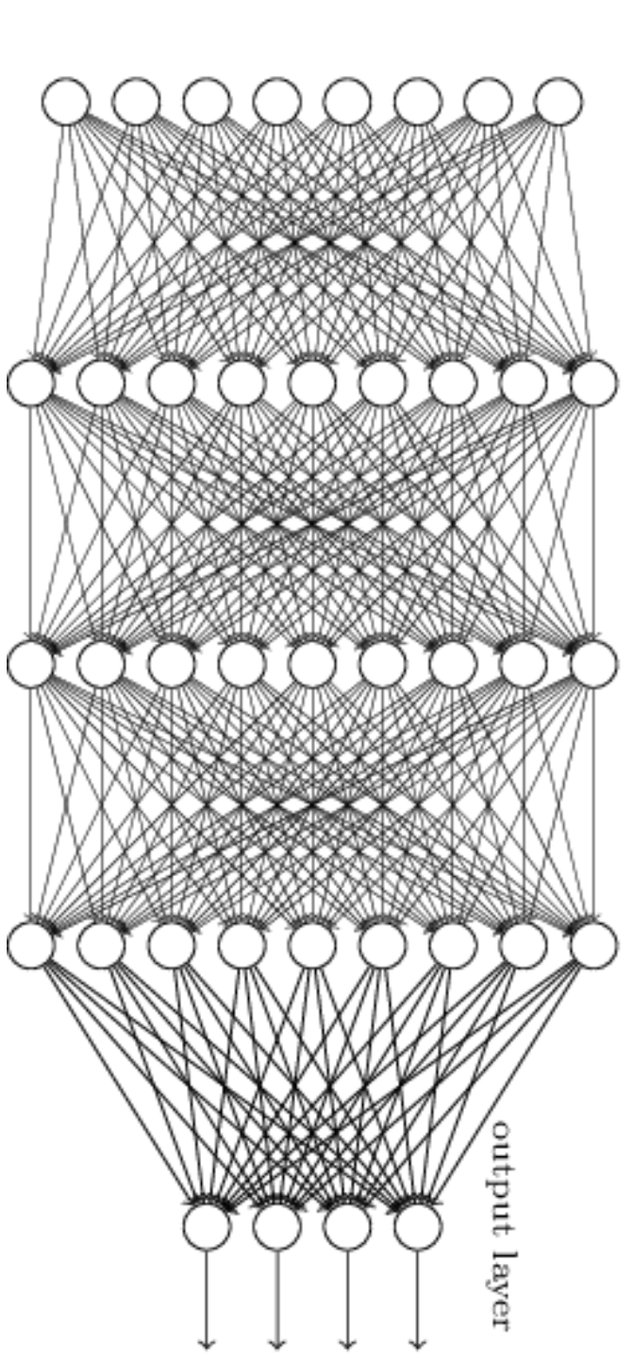
Micro-Intro: Machine Learning

- Rephrase task as a minimisation problem..
- ..and “simply” solve:

$$\theta^* = \text{argmin}_\theta \, \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})} \left[\mathcal{L}(f_\theta(\mathbf{x}), \mathbf{x}) \right]$$

- Modern ML: *function f is a deep neural network & minimisation carried out via gradient descent*

- Devil in the details:
  - How to map physics objective to loss function *L*
  - How to structure f to make maximum use of physics knowledge
  - How learn in a robust way from minimum amount of data
  - …

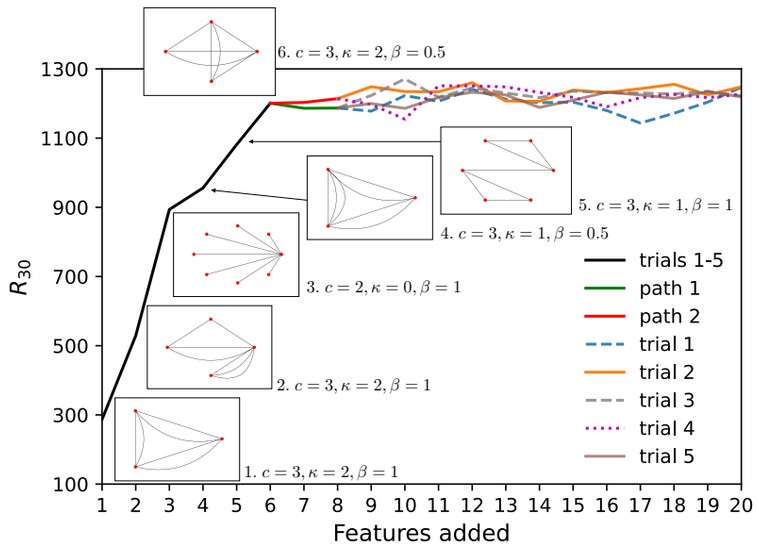


Overview

- Detailed domain knowledge — mathematical structures, useful observables, symmetries — in particle physics
- How to include in neural networks to improve performance and/or data and resource efficiency?
- Will cover three different areas & approaches

Overview

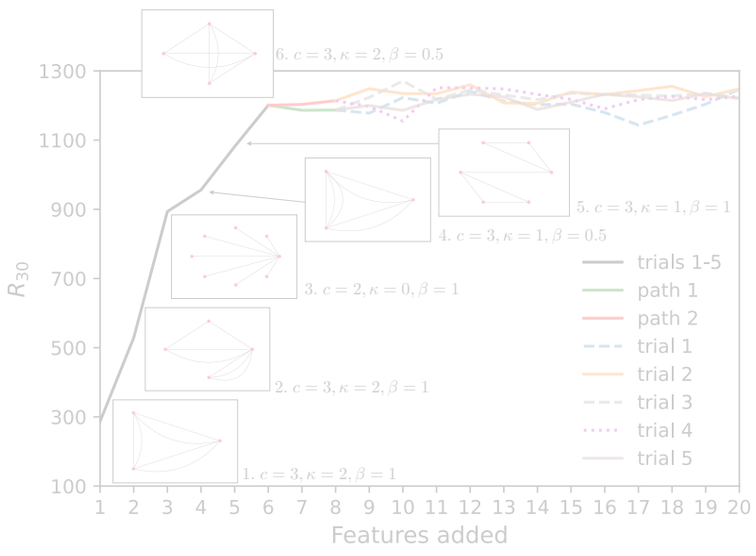
- Detailed domain knowledge — mathematical structures, useful observables, symmetries — in particle physics
- How to include in neural networks to improve performance and/or data and resource efficiency?
- Will cover three different areas & approaches



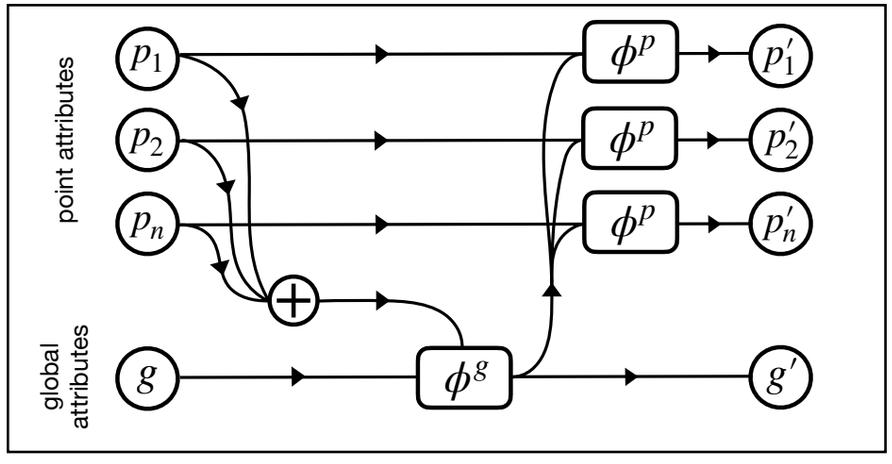
Select physically relevant features for supervised classification

Overview

- Detailed domain knowledge — mathematical structures, useful observables, symmetries — in particle physics
- How to include in neural networks to improve performance and/or data and resource efficiency?
- Will cover three different areas & approaches



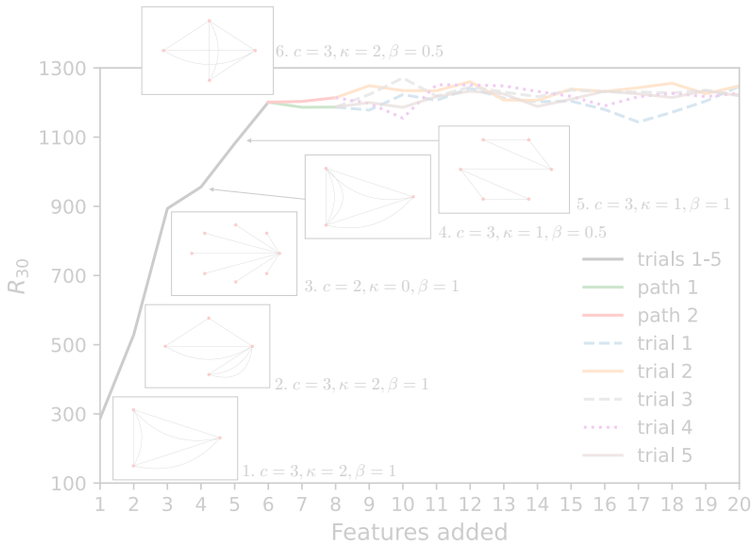
Select physically relevant features for supervised classification



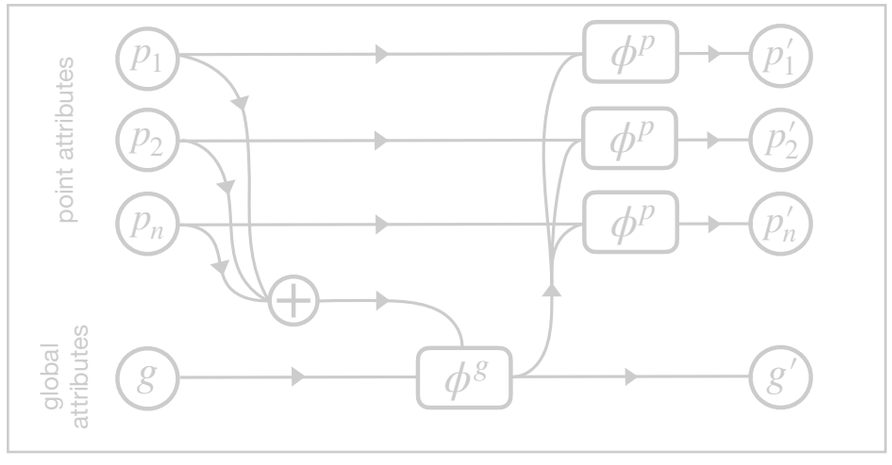
Add permutation symmetry to generative networks

Overview

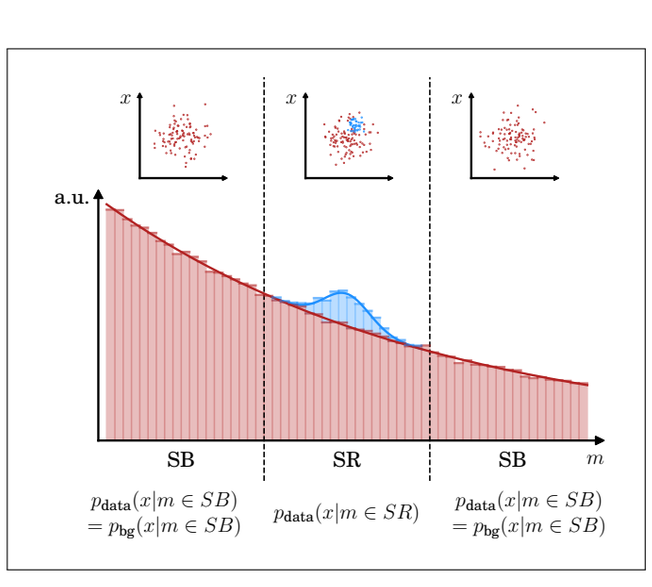
- Detailed domain knowledge — mathematical structures, useful observables, symmetries — in particle physics
- How to include in neural networks to improve performance and/or data and resource efficiency?
- Will cover three different areas & approaches



Select physically relevant features for supervised classification

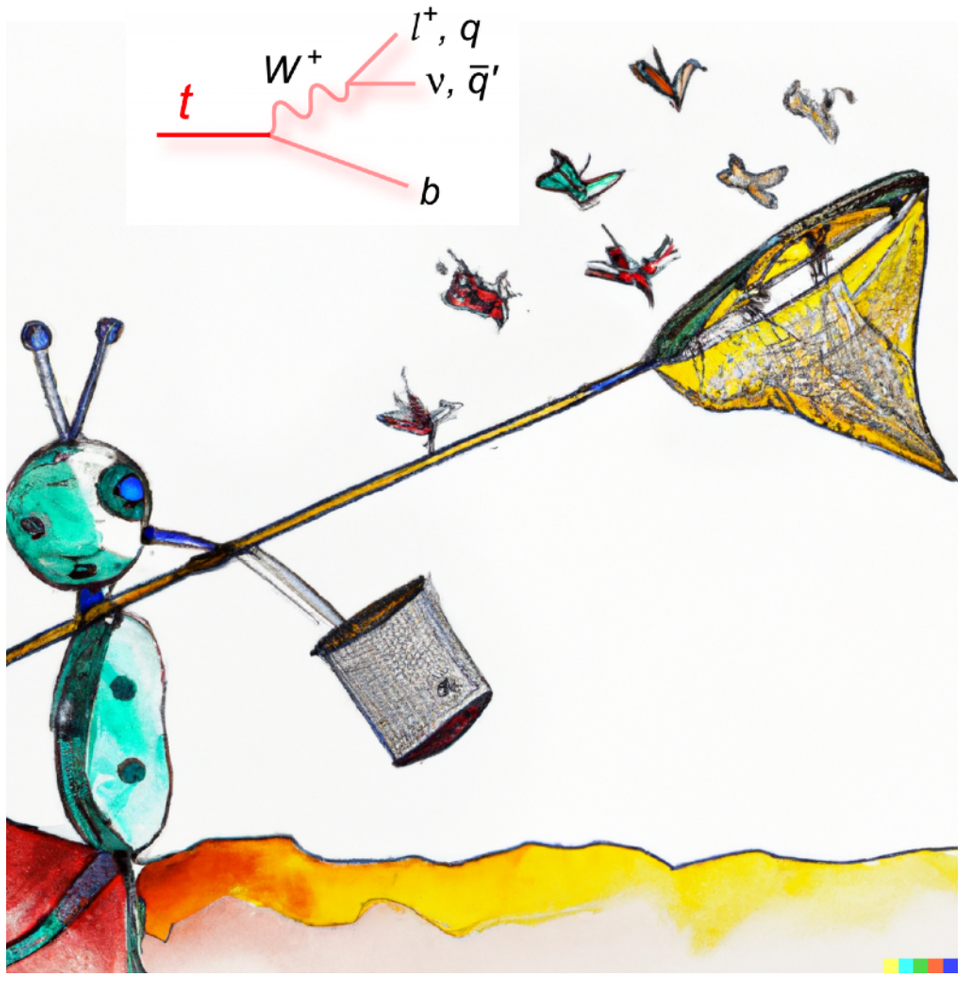


Add permutation symmetry to generative networks



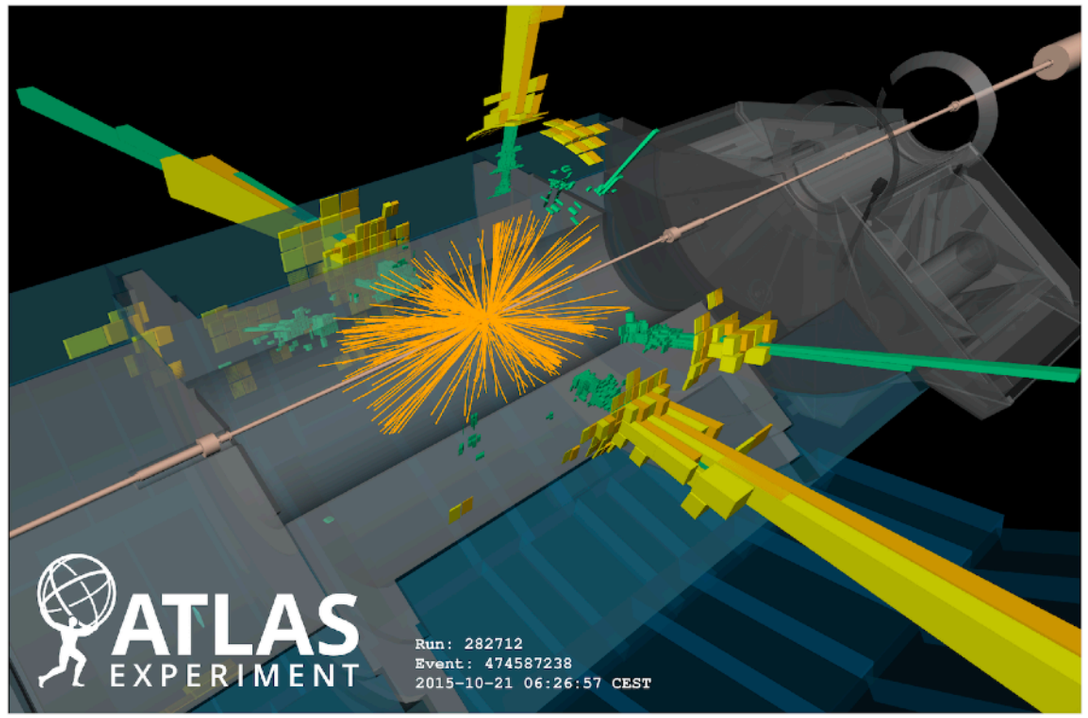
See how these issues affect the current frontier of anomaly detection

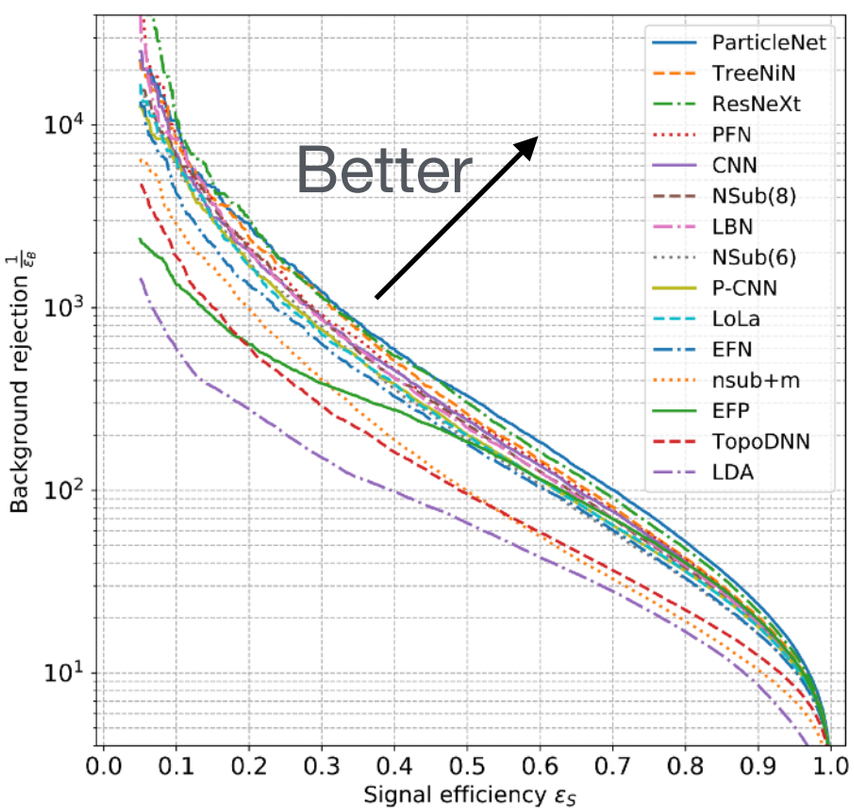
Particle taggers



Concrete Task

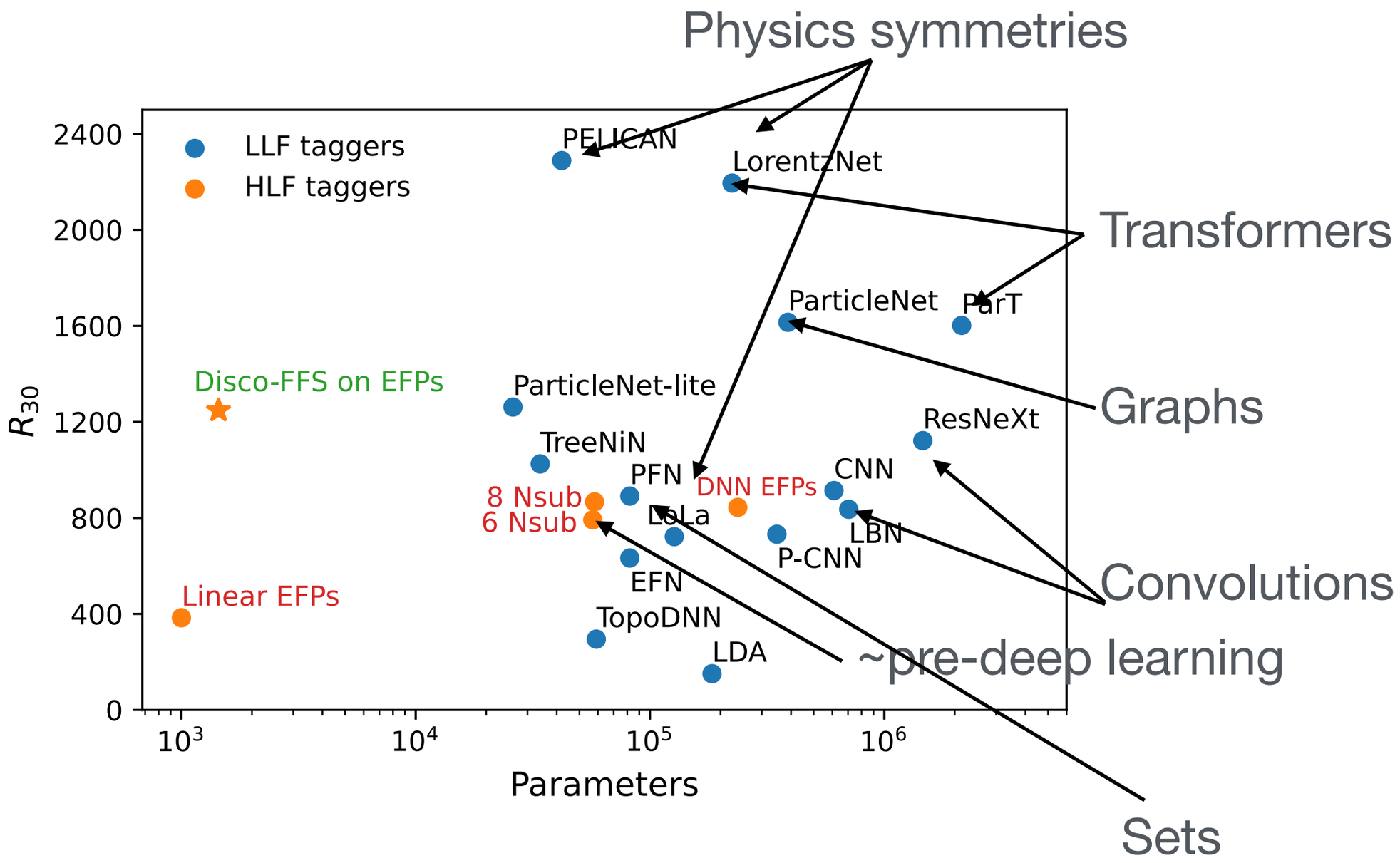
- Distinguish jets initiated by a top quarks from jets from other particles
  - Binary classification task
- Use simulation as synthetic training data: perfect class labels available
  - (Leads to domain shift when applied to collider data)





GK, Plehn, et al 1902.09914

- 1.2M training examples (*jets*), 400k each for testing and validation
- Each example: Up to 200 particles with 3 features/particle (2D position on detector surface+ energy)
- Metrics: AUC: area under curve and R_{30}:1/FPR @ TPR=0.3



- Many ways to encode symmetries in network architecture to improve tagging performance: Dedicated talk later today
- Instead, a way to automatically find best features to describe a jet

Das, GK, Shih 2212.00046

Motivation

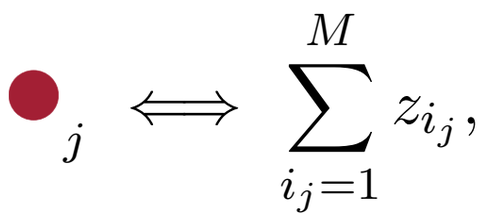
- Advantage of few high-level features:
  -easy to understand and calibrate
  -cheap to evaluate

- Advantage of complex architecture and low-level features: performance

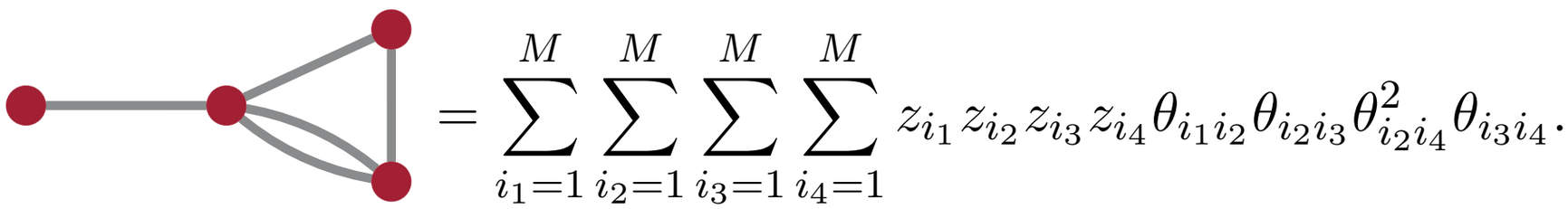
- Can we combine both?

We need a basis

- Energy Flow Polynomials (EFPs) form a basis of jet substructure
  - Nodes: energy fractions
  - Edges: angular distances
- Depending on order considered, too many (e.g 7k) to efficiently train NN (many features work if there is structure, not so much for EFPs)



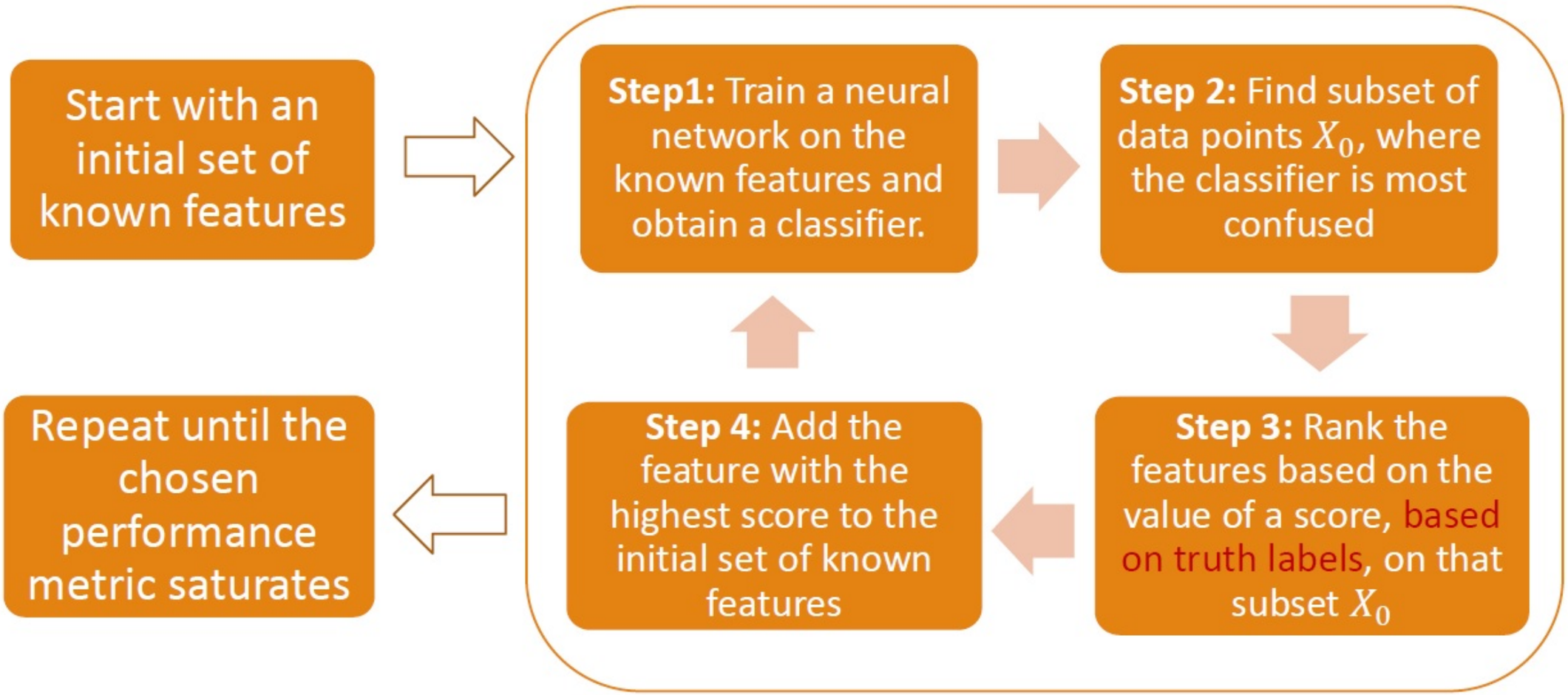
e.g.



Komiske, Metodiev, Thaler 1712.07124

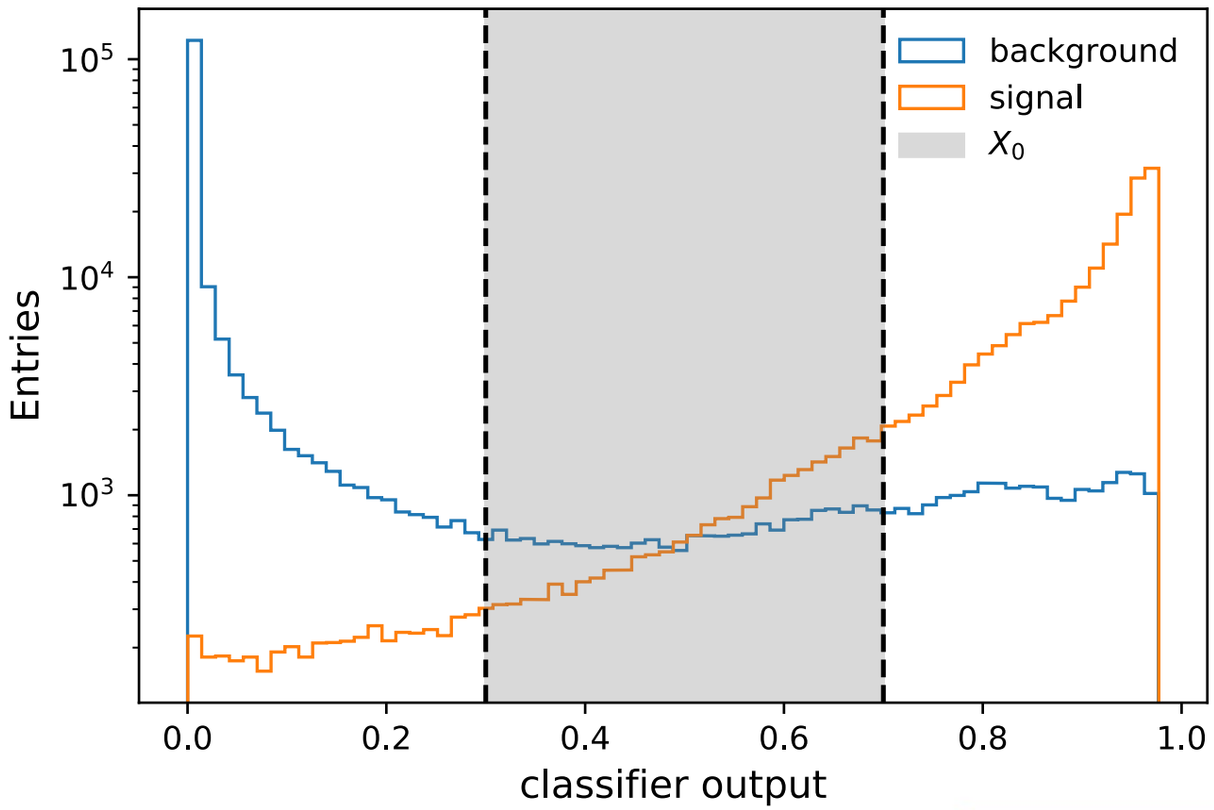
Look for optimal feature set

Solution: Iterative feature selection



Das, GK, Shih 2212.00046; Faucett, Thaler, Whiteson, 2010.11998

Algorithm



Focus on maximally confused region

Start with an initial set of known features

Step1: Train a neural network on the known features and obtain a classifier.

Step 2: Find subset of data points X_0, where the classifier is most confused

Step 3: Rank the features based on the value of a score, based on truth labels, on that subset X_0

Step 4: Add the feature with the highest score to the initial set of known features

Repeat until the chosen performance metric saturates

Das, GK, Shih 2212.00046

Algorithm

Already selected features

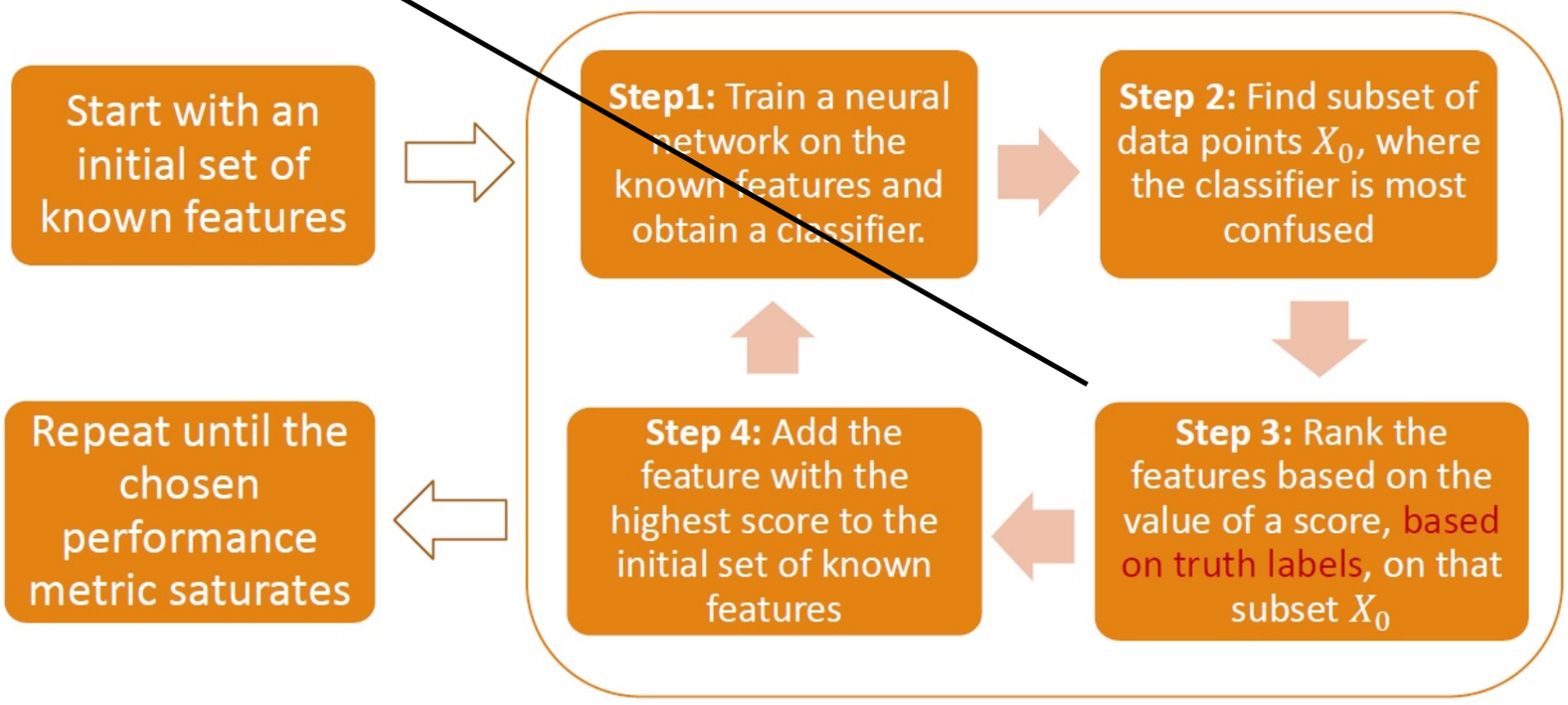
New feature under test

$$\mathcal{X} = \left\{ \left(f_{i_1}(\vec{x}), \ldots, f_{i_n}(\vec{x}), f_i(\vec{x}) \right) \middle| \vec{x} \in X_0 \right\}$$

$$\mathcal{Y} = \{ y_{\mathrm{ref}}(\vec{x}) | \vec{x} \in X_0 \}$$

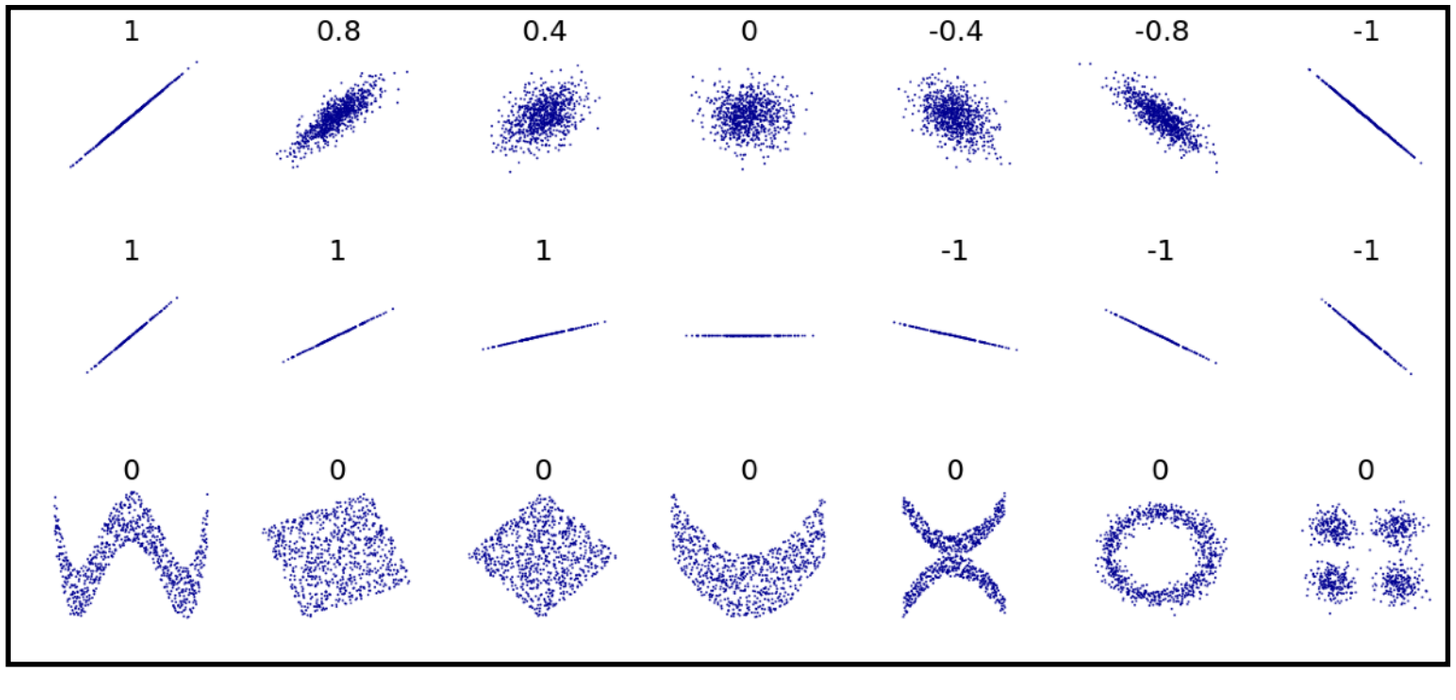
$$s_{f_i} = \text{Affine-DisCo}(\mathcal{X}, \mathcal{Y}).$$

Truth label or reference classifier



Das, GK, Shih 2212.00046;

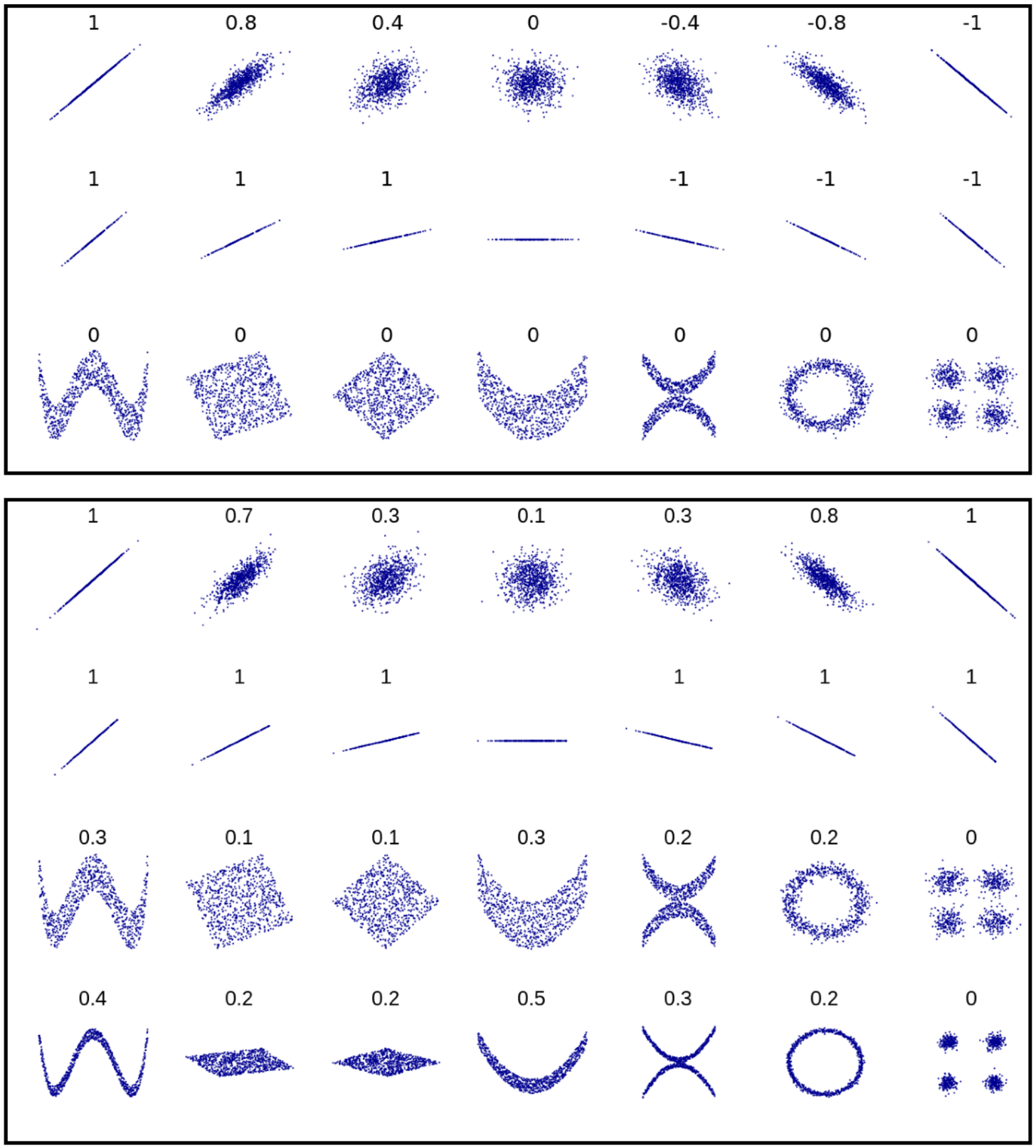
Example



Pearson Correlation

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Example

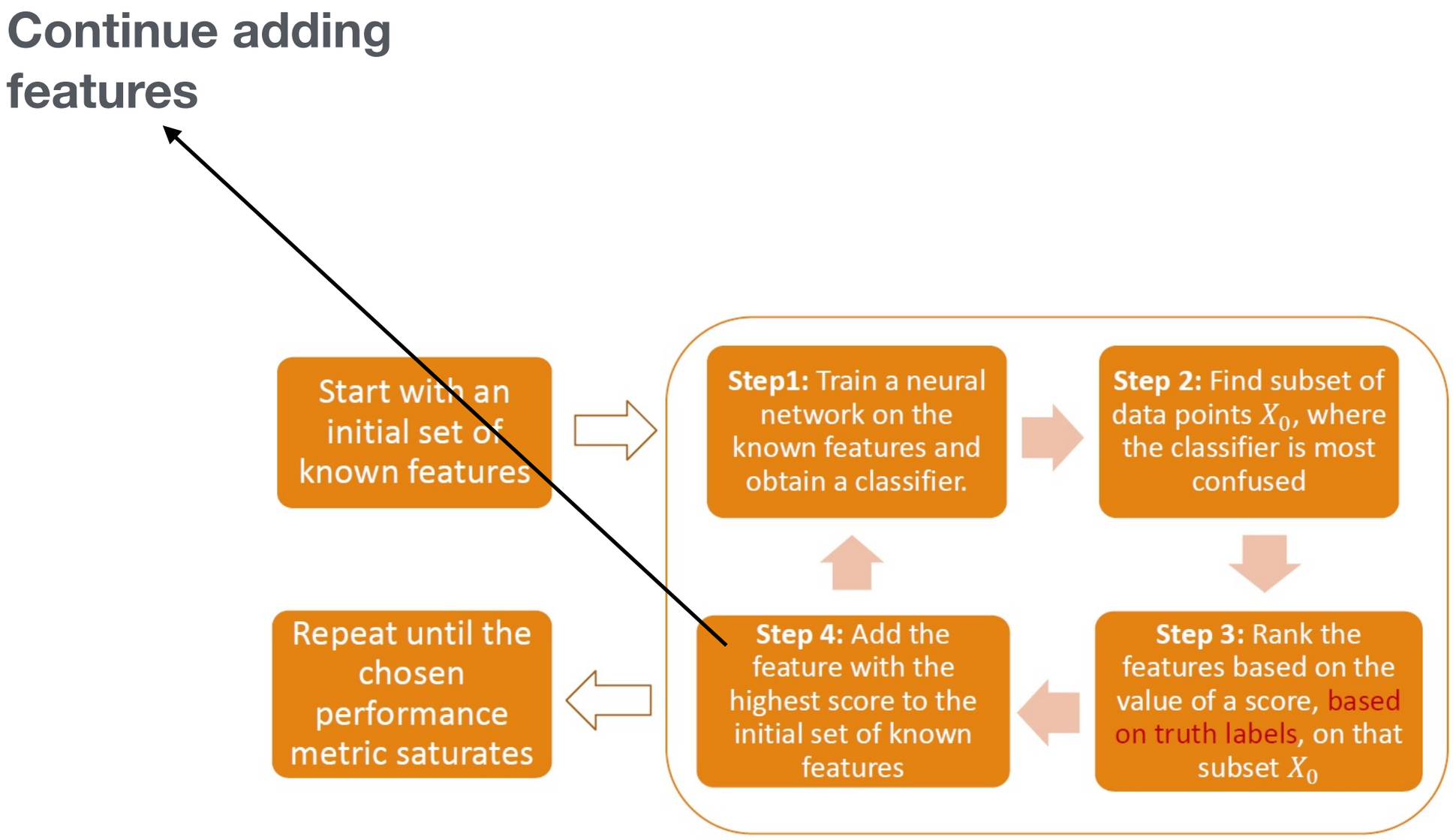


Pearson Correlation

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Distance Correlation

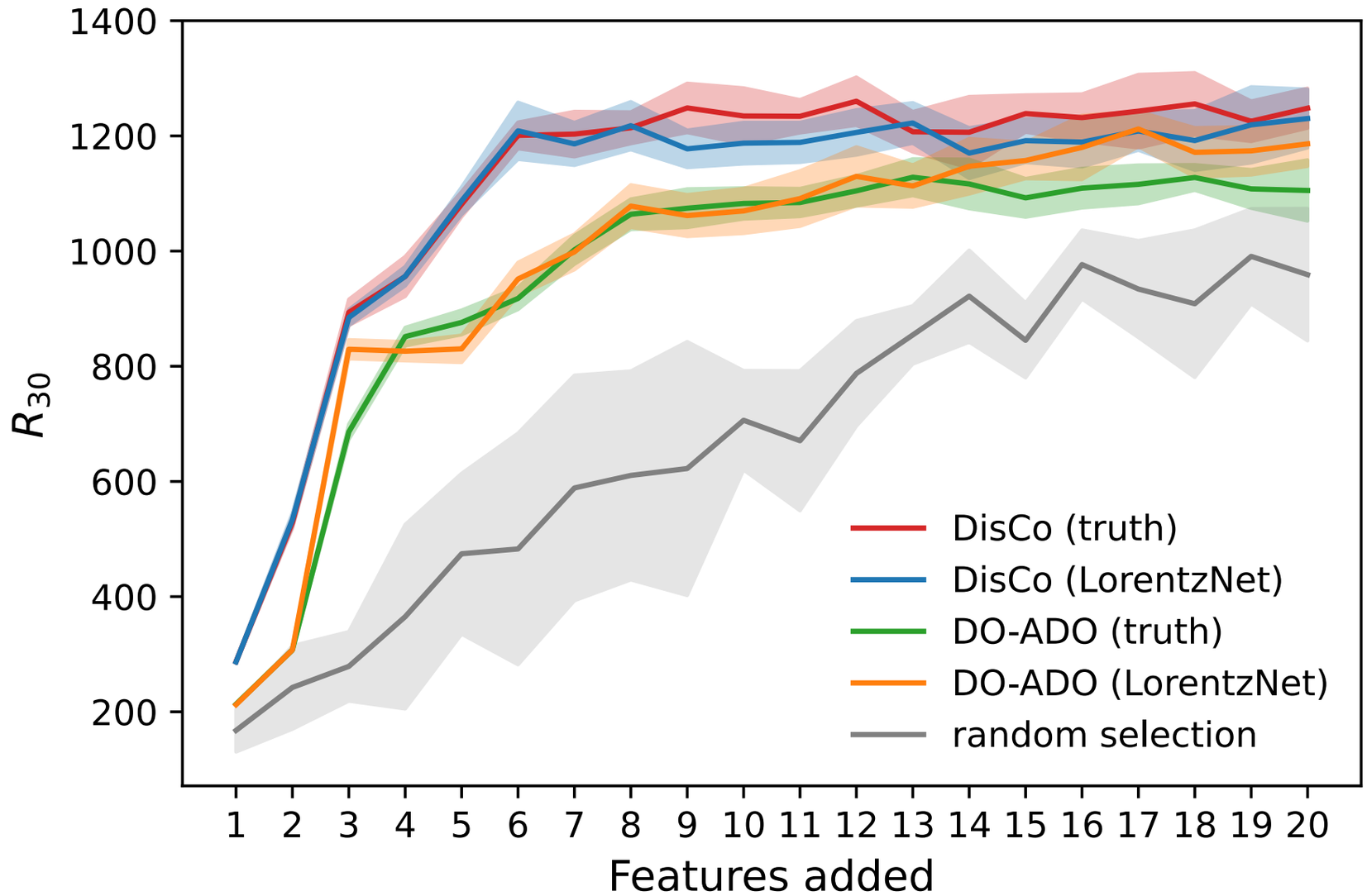
Algorithm



Das, GK, Shih 2212.00046;

Results

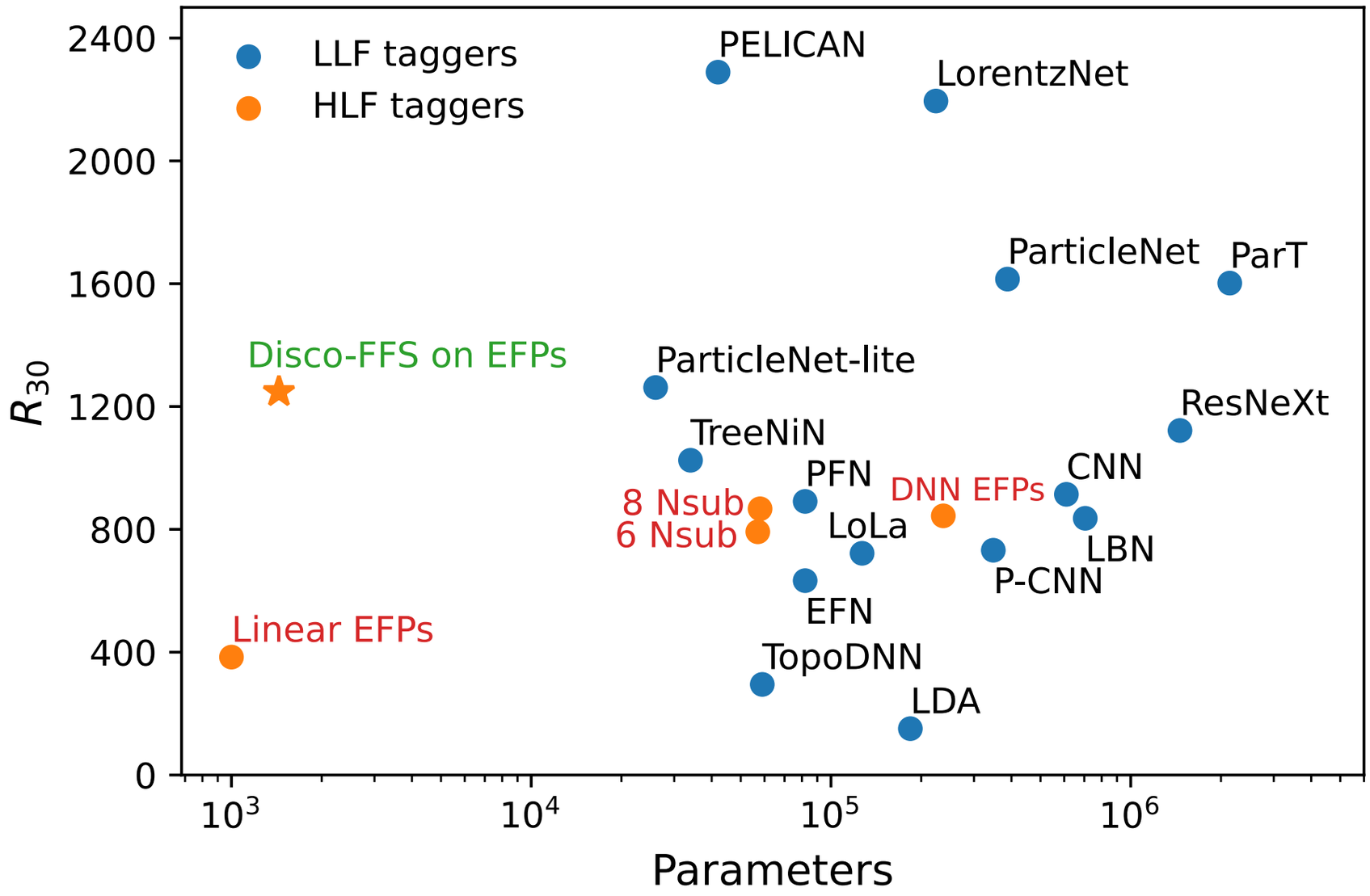
- DiscoFFS find relevant features quicker than alternative feature selection methods



Das, GK, Shih 2212.00046;

Results

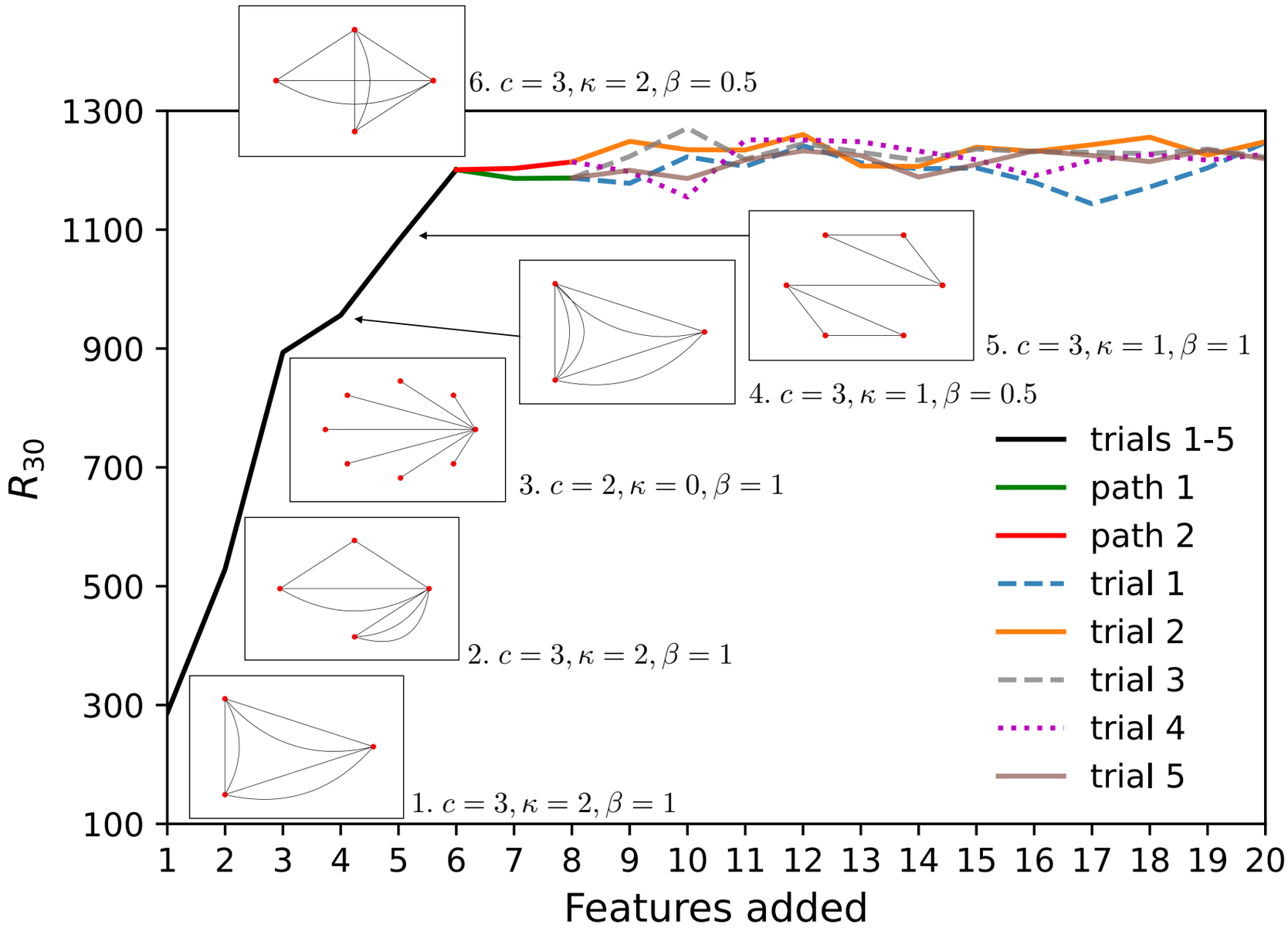
- DiscoFFS find relevant features quicker than alternative feature selection methods
- New best trade-off between performance and complexity



Das, GK, Shih 2212.00046;

Results

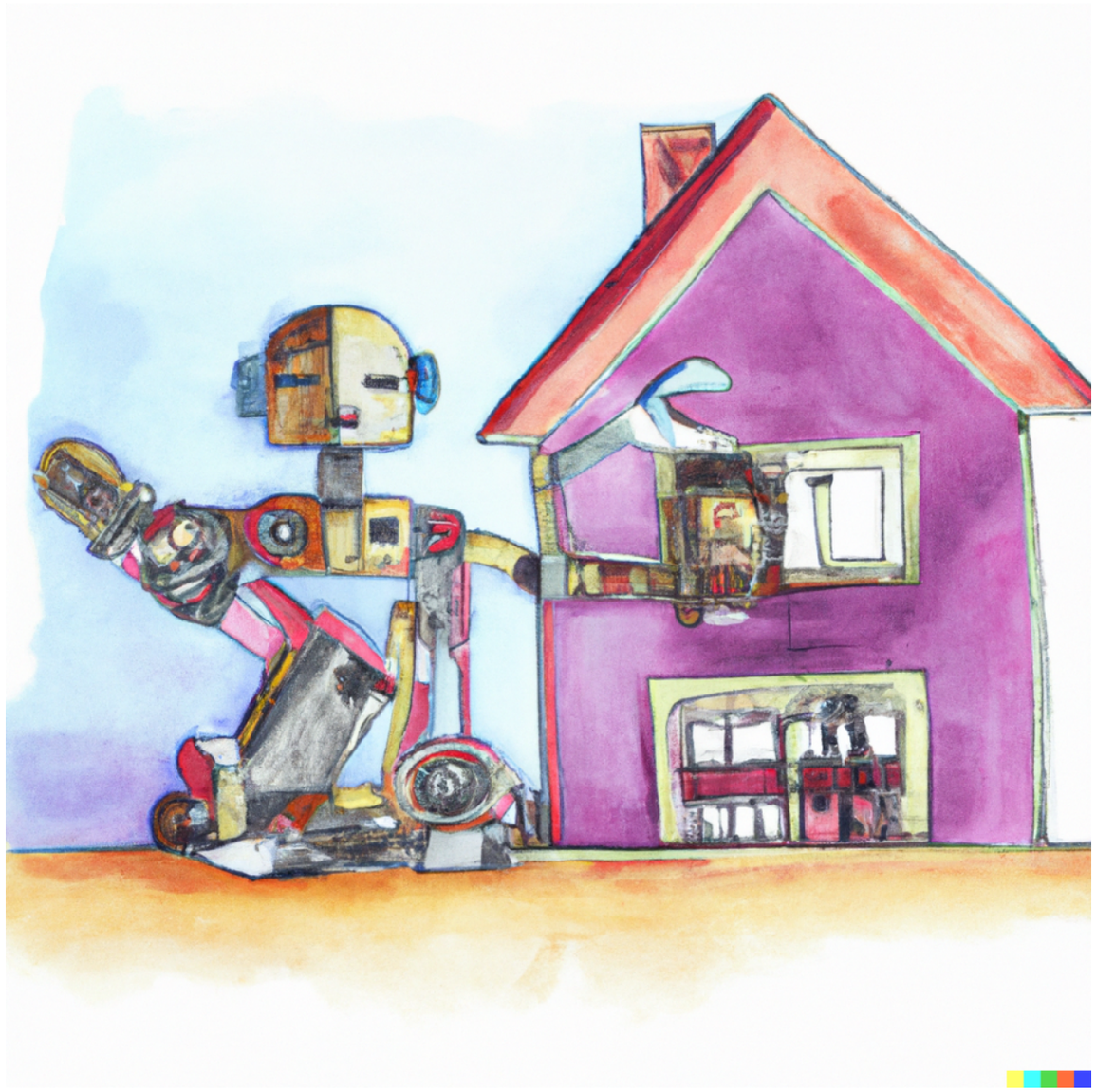
- DiscoFFS find relevant features quicker than alternative feature selection methods
- New best trade-off between performance and complexity
- Features stable under re-training



$$\mathrm{EFP}_G^{(\kappa,\beta)} = \sum_{i_1 \in J} \cdots \sum_{i_N \in J} z_{i_1}^{(\kappa)} \cdots z_{i_N}^{(\kappa)} \prod_{(m,\ell) \in G} \theta_{i_m i_\ell}^{(\beta)}. \quad \text{; chromatic number c}$$

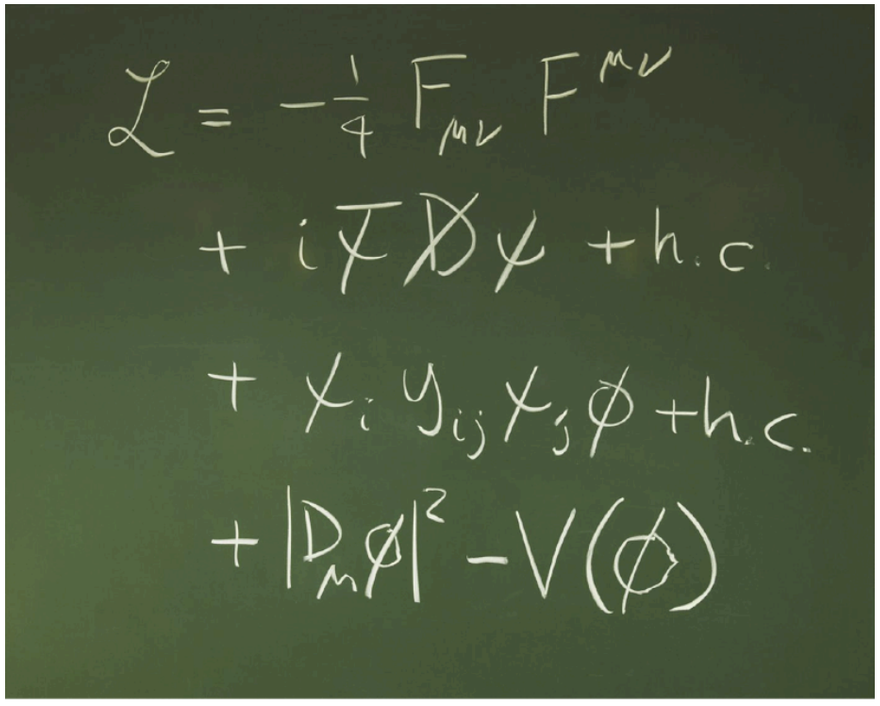
Das, GK, Shih 2212.00046;

Fast simulation

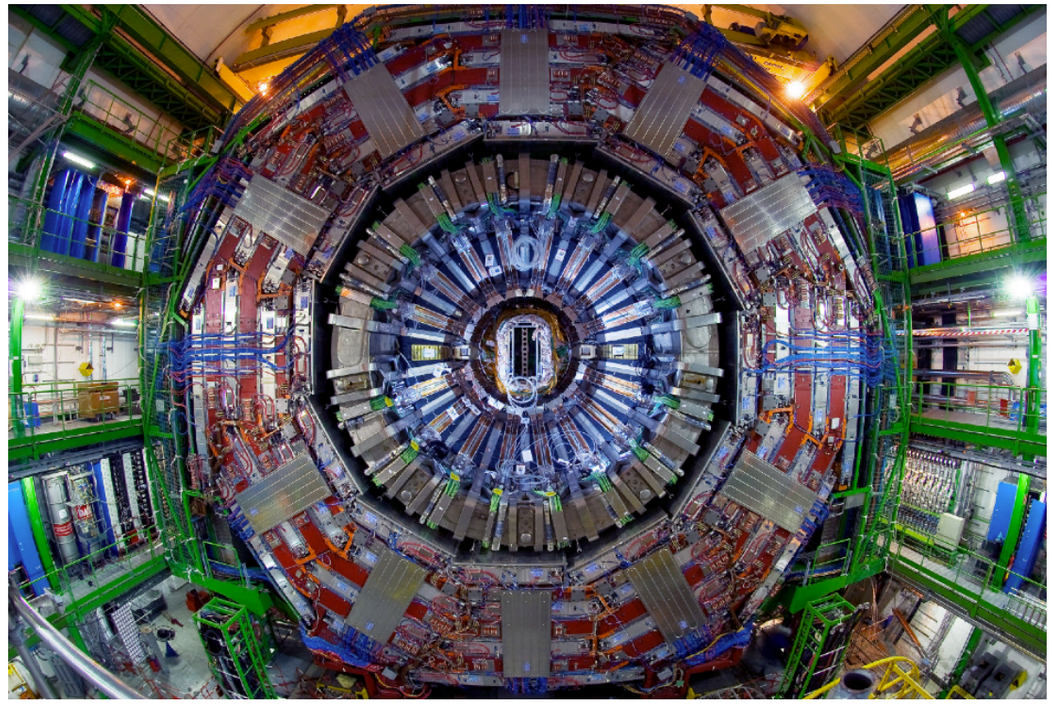


Generative Models

Simulation is crucial to connect experimental data with theory predictions



Complex chain of simulations →

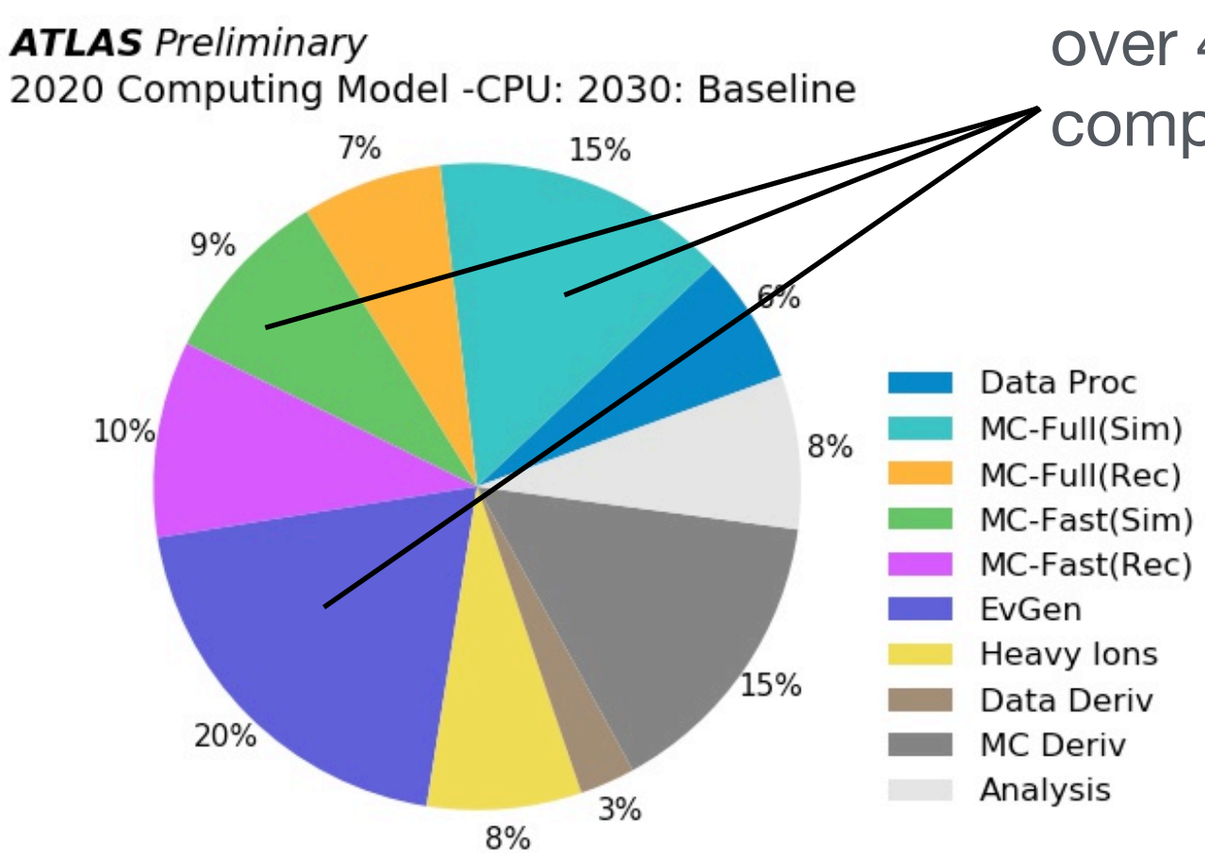


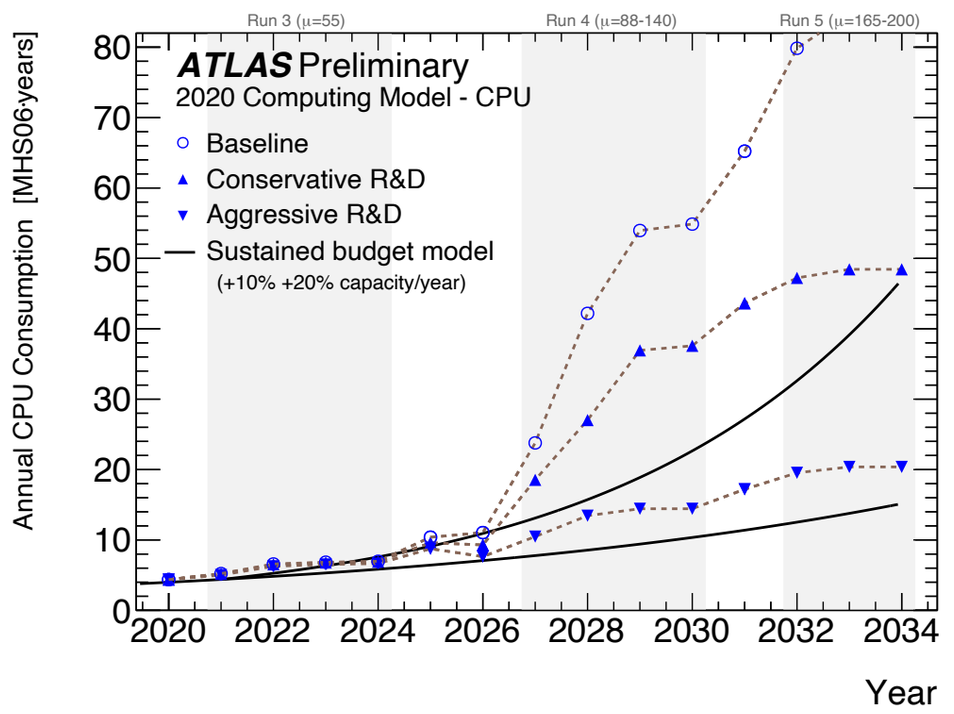
Generative Models

Simulation is crucial to connect experimental data with theory predictions
but computationally very expensive

Simulation and Generation steps over 40% of ATLAS compute effort..

..and projected to exceed available resources.

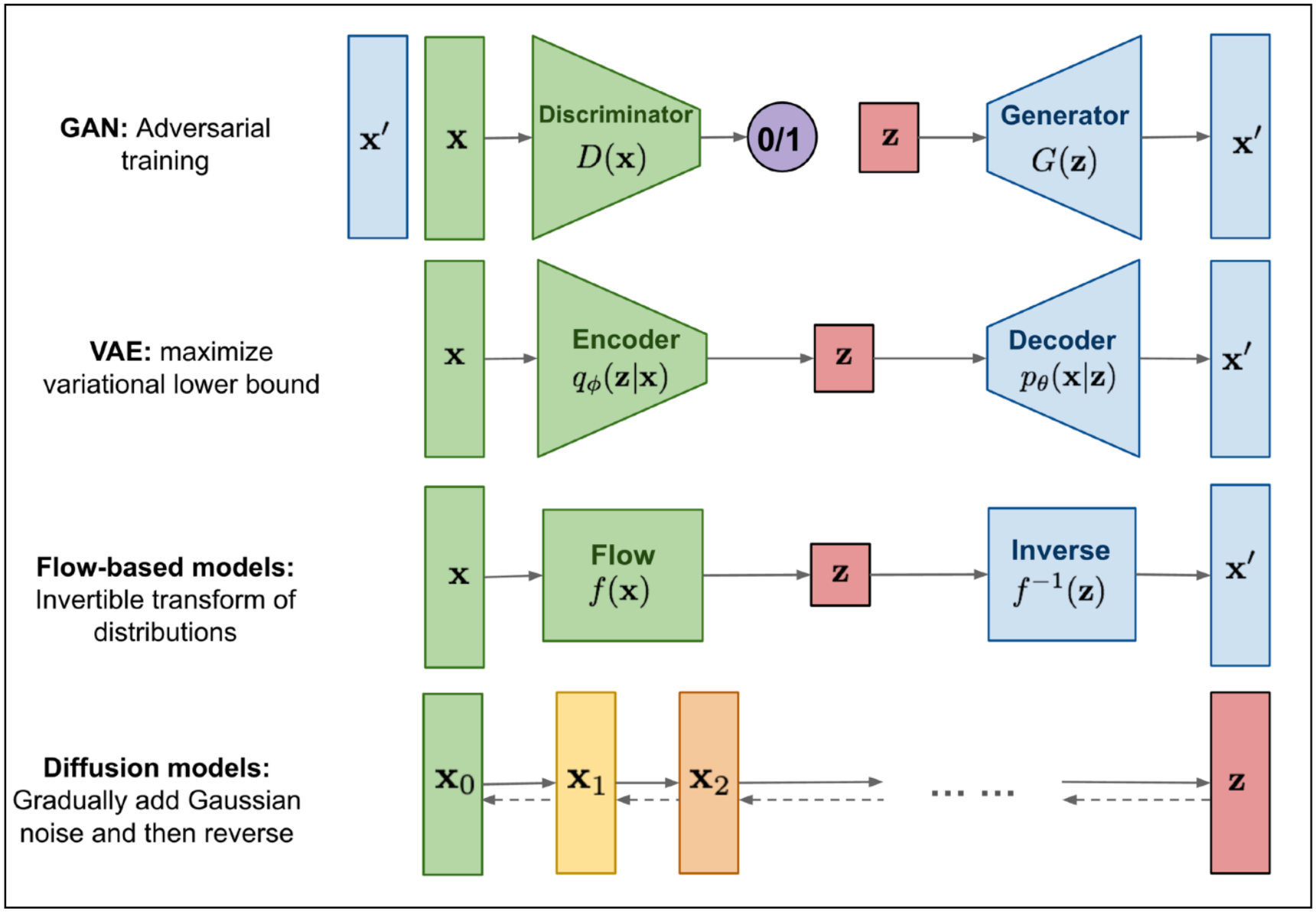




Generative Models

Simulation is crucial to connect experimental data with theory predictions
but computationally very expensive

Use generative models trained on initial data to augment statistics



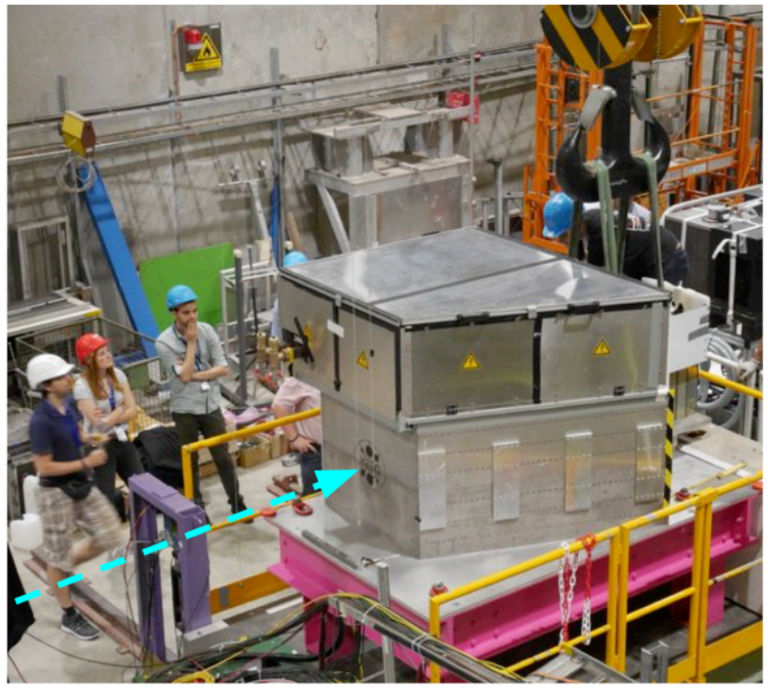
Overview of generative architectures

Generative Models

Simulation is crucial to connect experimental data with theory predictions
but computationally very expensive

Use generative models trained on initial data to augment statistics

Want to achieve particle showers in calorimeters



CALICE AHCal testbeam

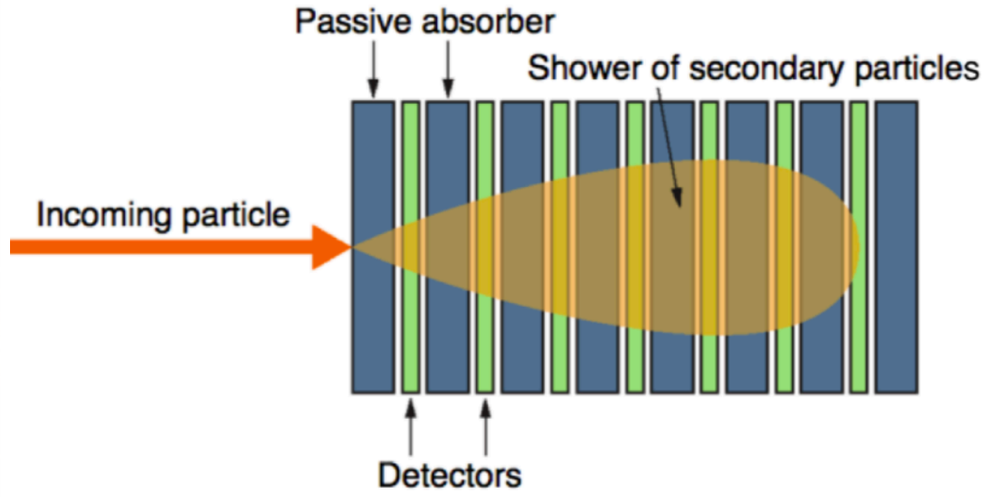
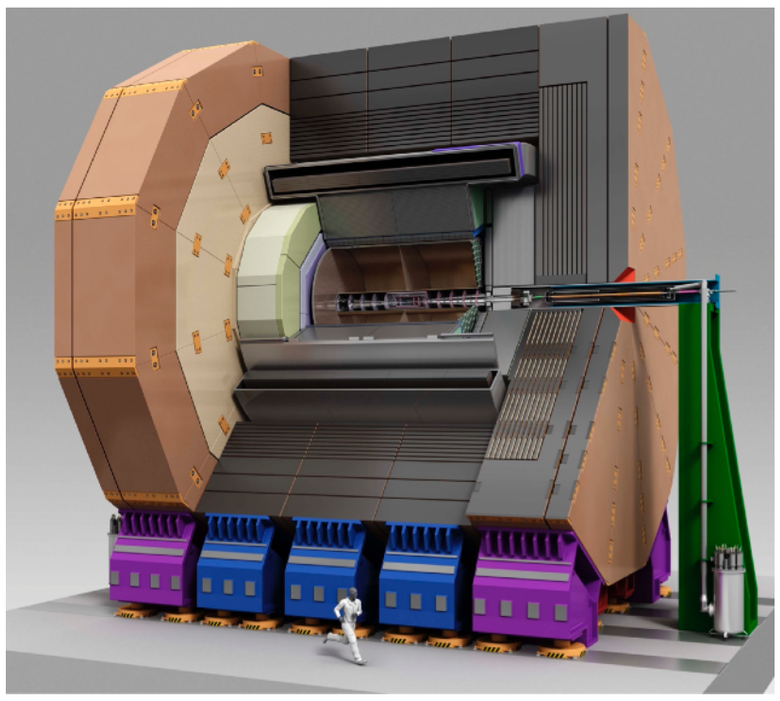
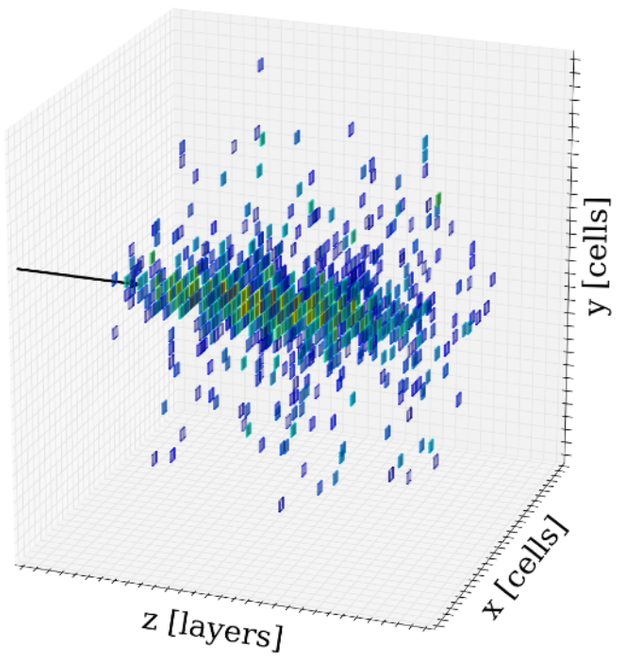


Illustration of particle shower in a sampling calorimeter



ILD Detector



One data example

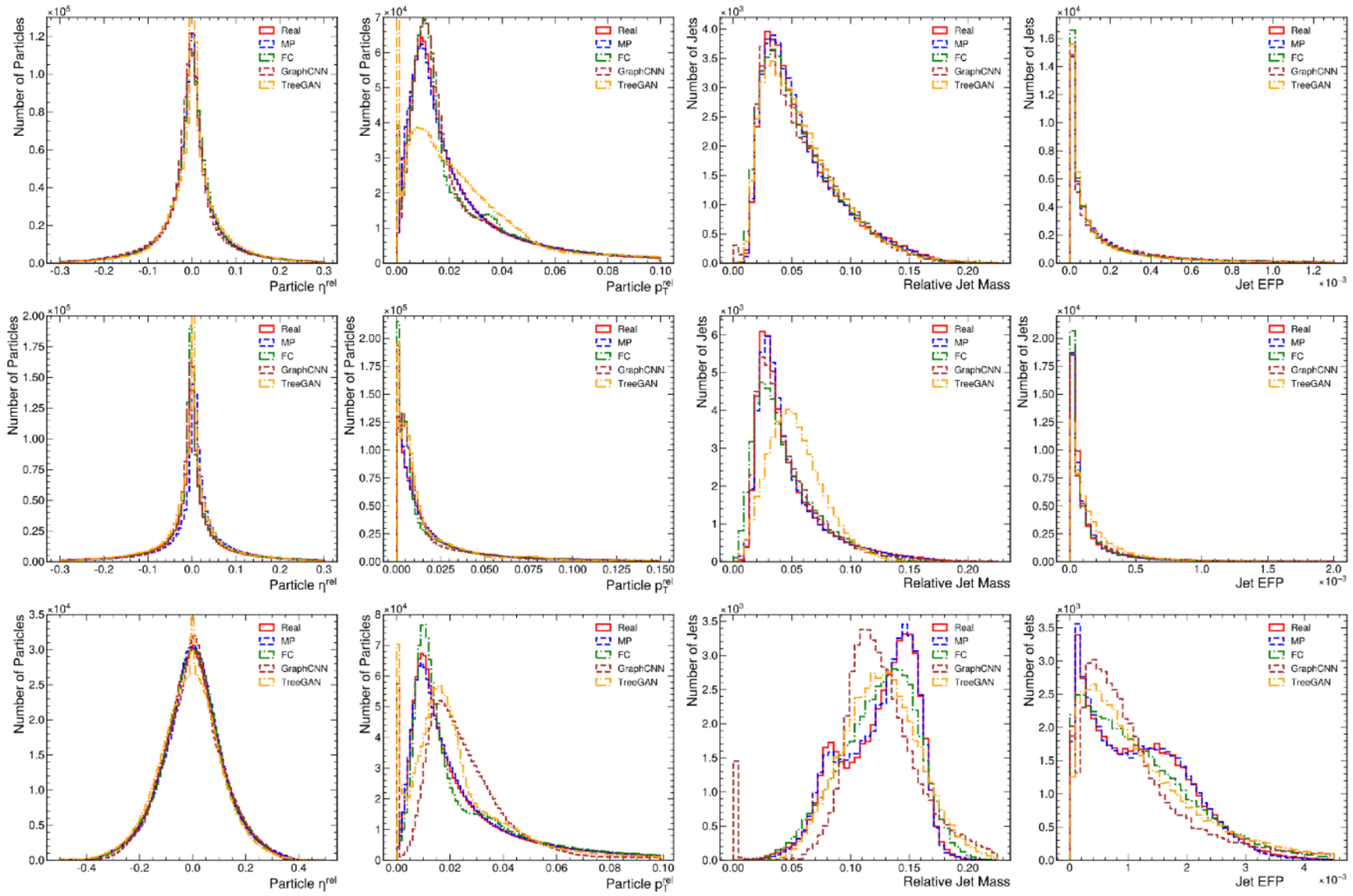
Generative Models

Simulation is crucial to connect experimental data with theory predictions
but computationally very expensive

Use generative models trained on initial data to augment statistics

Want to achieve particle showers in calorimeters

But start with a simpler task (Claudius will talk about Calorimeters)

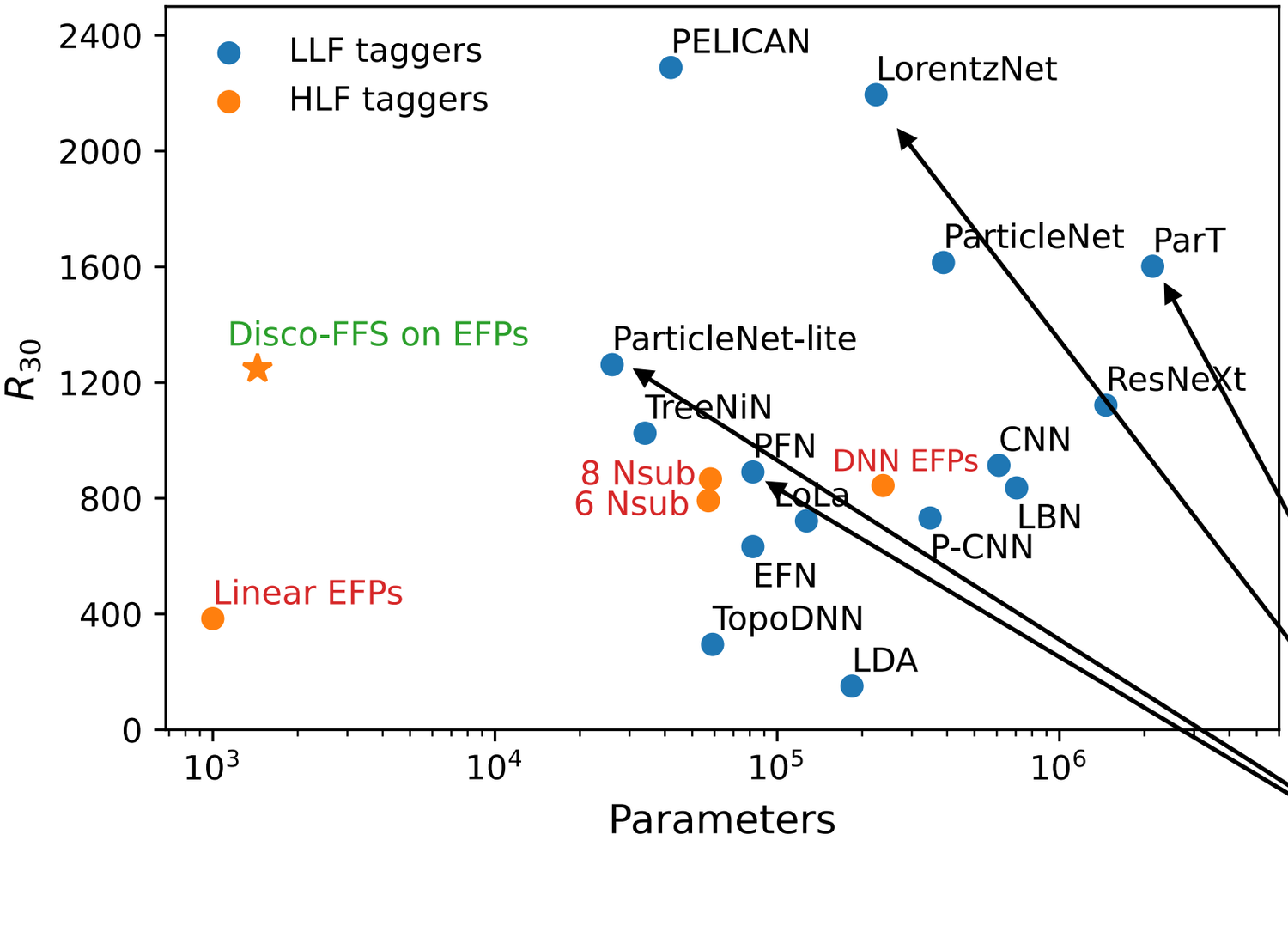


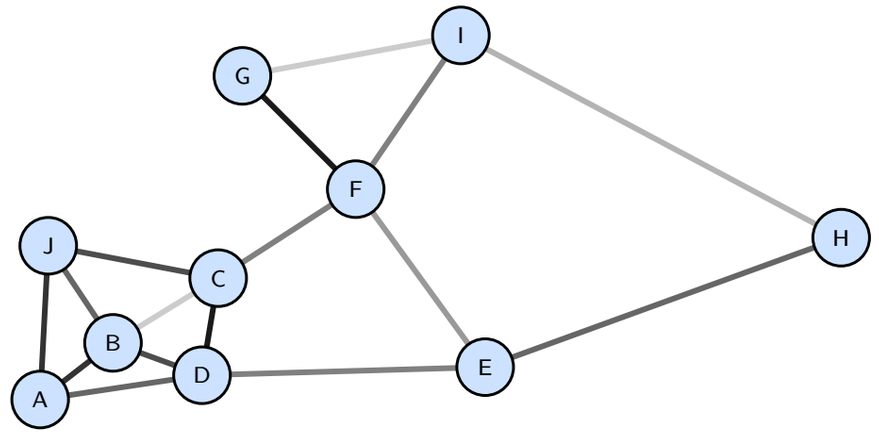
JetNet dateset: Pythia final-state particles from a jet. Either up to 30 or 150 particles/jet

Kansal et al 2106.11535;

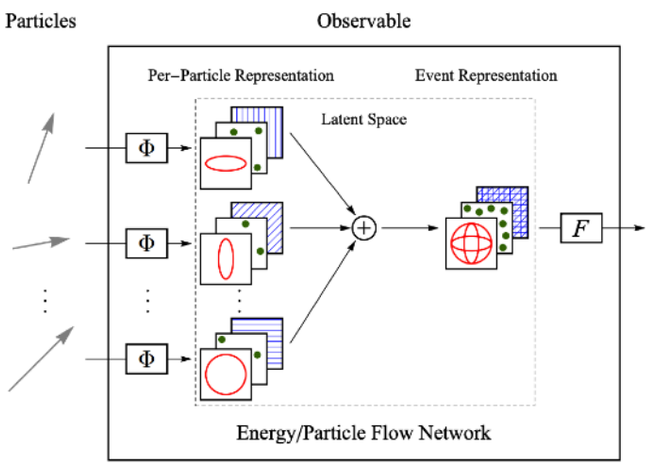
Point cloud generation

- In both cases (calorimeter showers, JetNet):
  - Each jet/shower is a point cloud with 30 to O(100k) points (particles)
  - Each point has a 2D/3D position in space and other associated quantities
- This poses a limit to using fixed structures (e.g. convolutions)
  - How to simulate point clouds directly?
  - Specifically: how to add permutation symmetry?

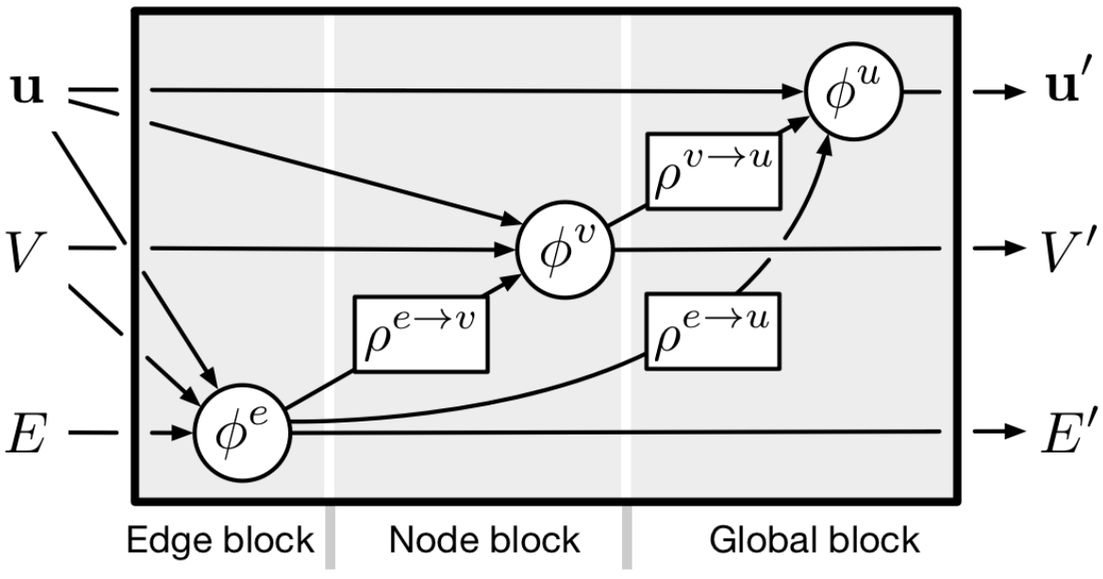




Nodes: Individual particles,
Edges: Geometrical distance



Set, graph and transformer based techniques

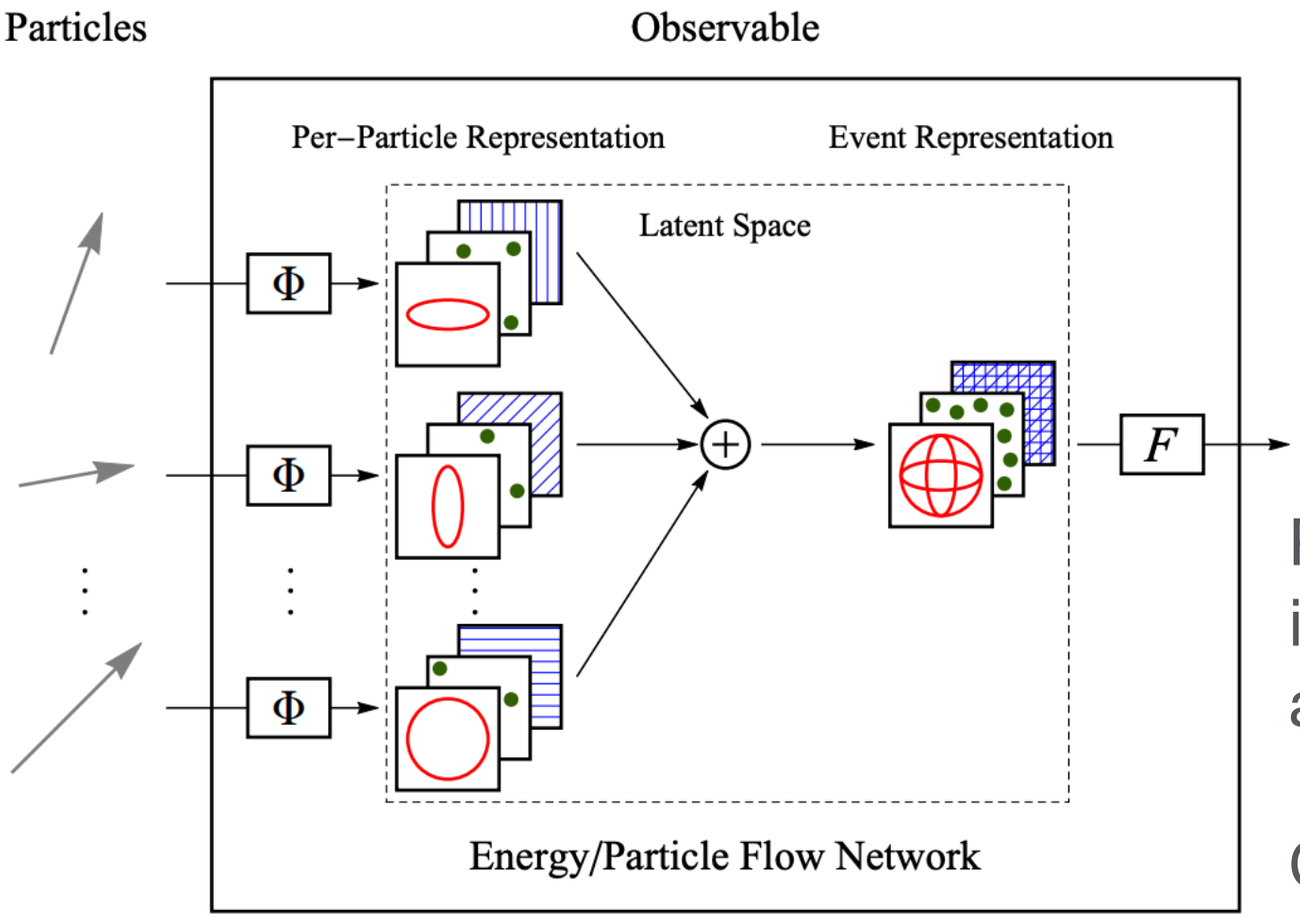


General message-passing graph block

Das, GK, Shih 2212.00046; Qu, Gouskos 1902.08570; Qu et al 2202.03772; Gong et al 2201.08187; Shlomi et al 2007.13681;

Deep Sets

Theorem 2 *A function $f(X)$ operating on a set X having elements from a countable universe, is a valid set function,* i.e., ***invariant*** *to the permutation of instances in X, iff it can be decomposed in the form $\rho\left(\sum_{x\in X}\phi(x)\right)$, for suitable transformations ϕ and ρ.*

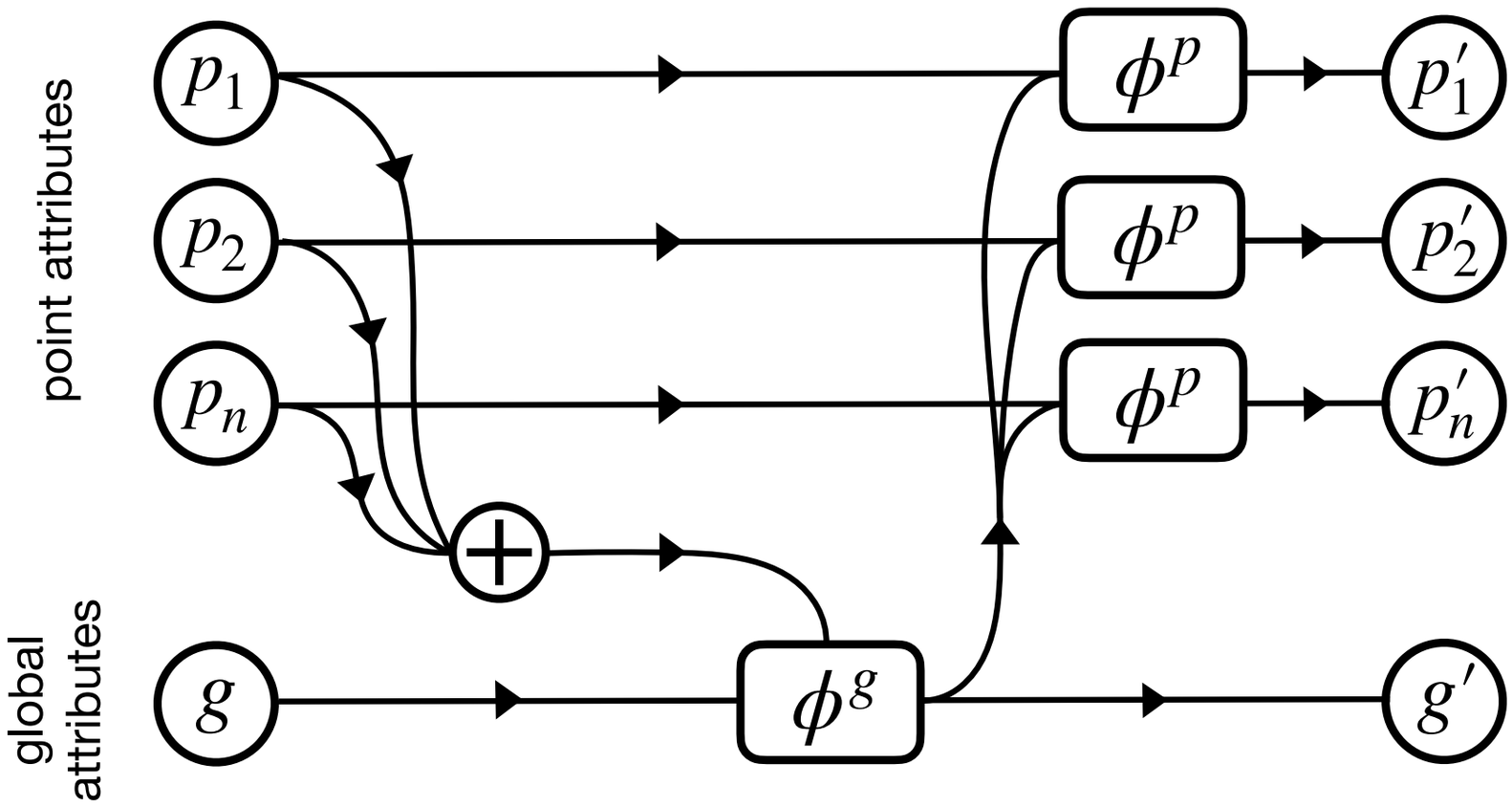


Permutation invariance for a tagger is ‘easy’: Use a permutation invariant aggregation function

Can extend to generation?

Zaheer et al 1703.06114; Komiske, Metodiev, Thaler 1810.05165

EPiC

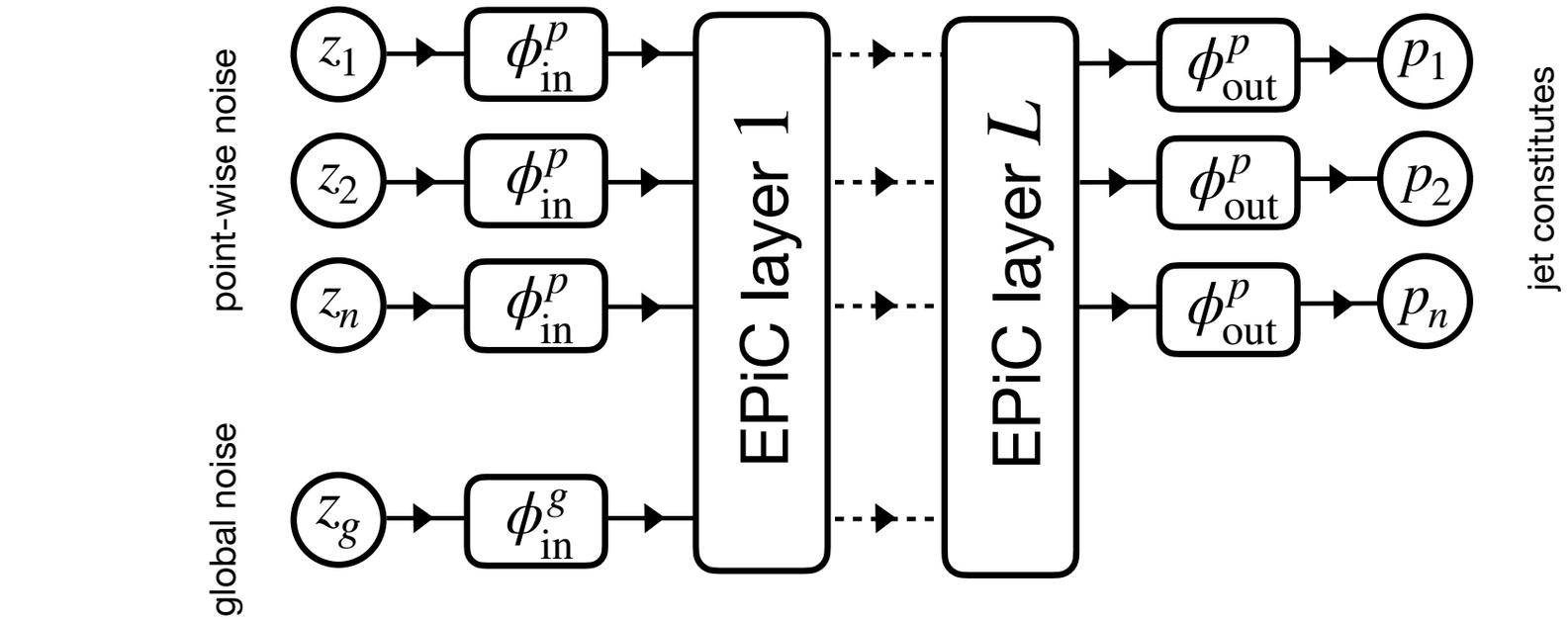


Equivariant Point Cloud interaction (EPiC) block:
Similar to deep sets, but with additional global information exchange.

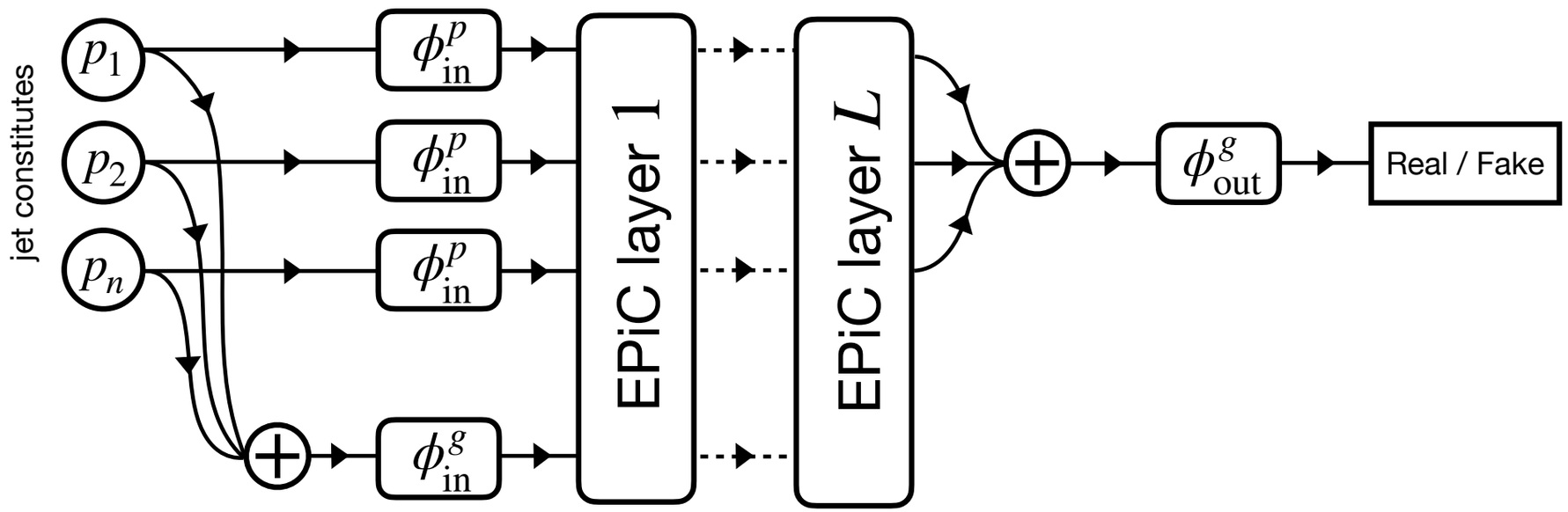
Still permutation equivariant

Buhmann, GK, Thaler 2301.08128;

EPiC GAN



(a) Generator

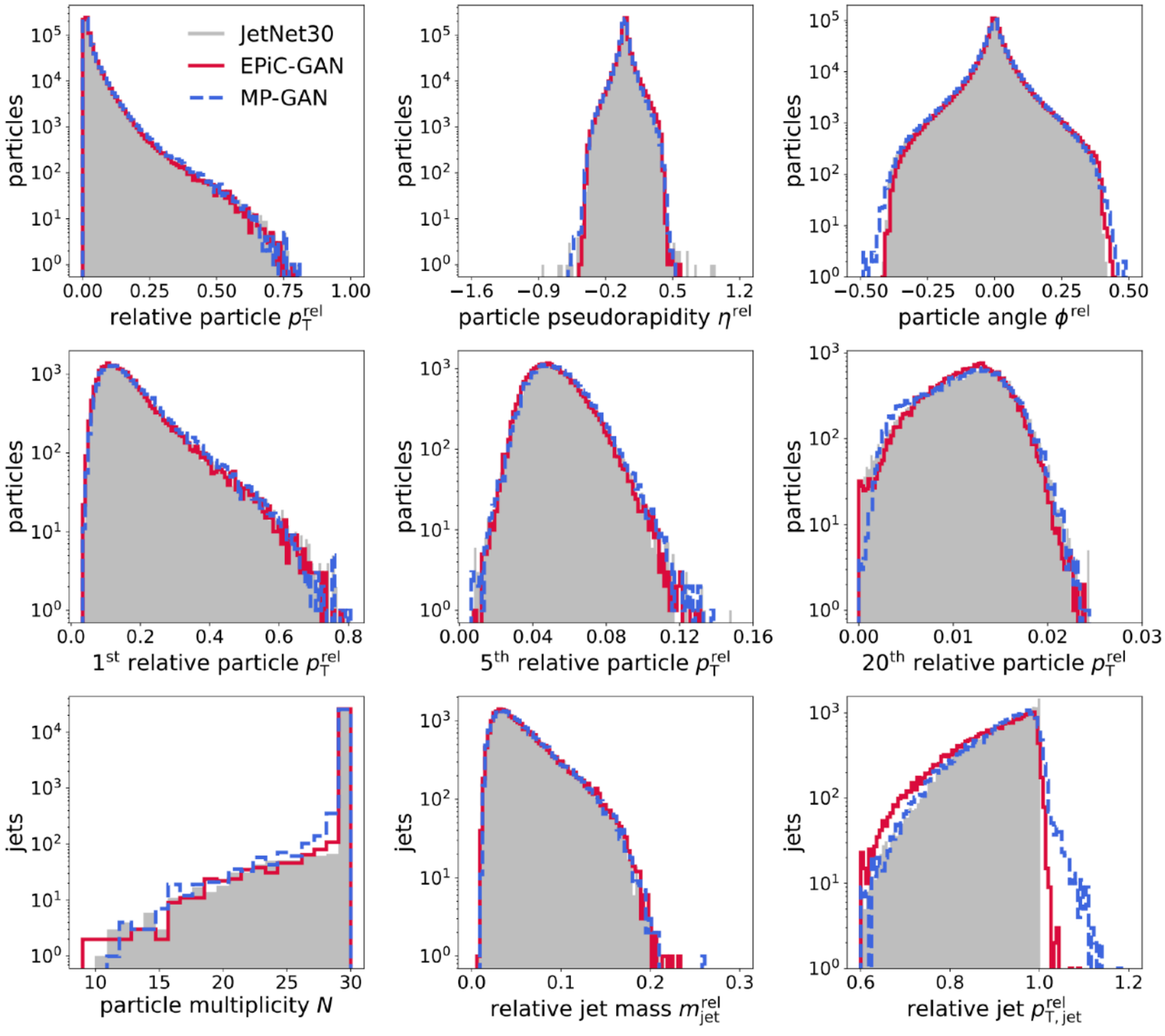


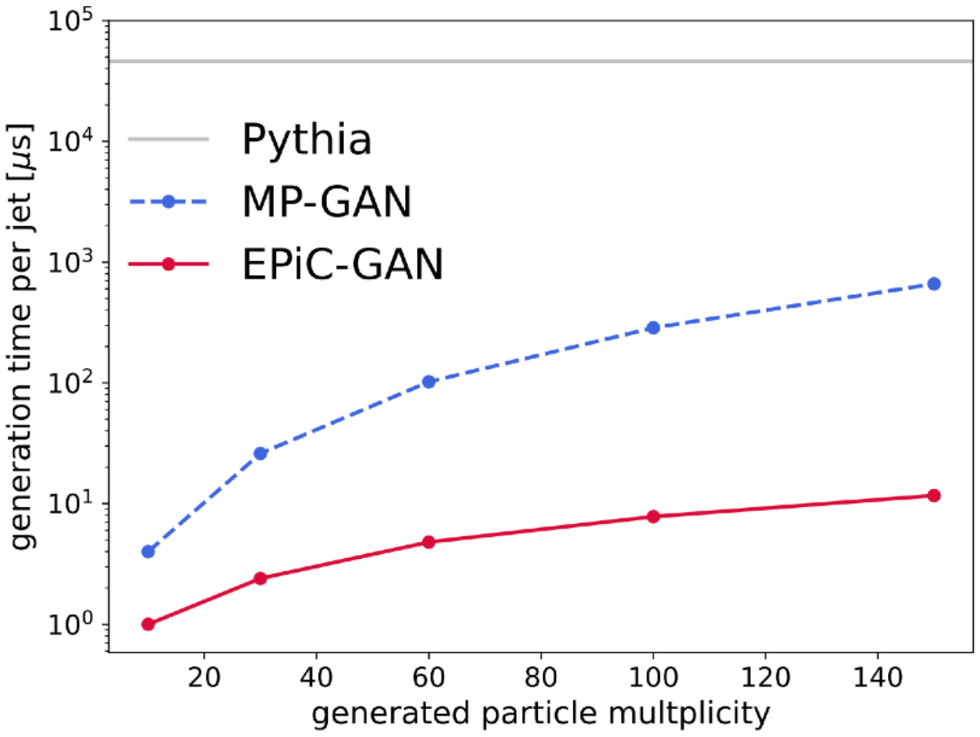
(b) Discriminator

Can use to build generator and discriminator blocks for classical GAN architecture, but fully respecting permutation symmetry.

Buhmann, GK, Thaler 2301.08128;

Results





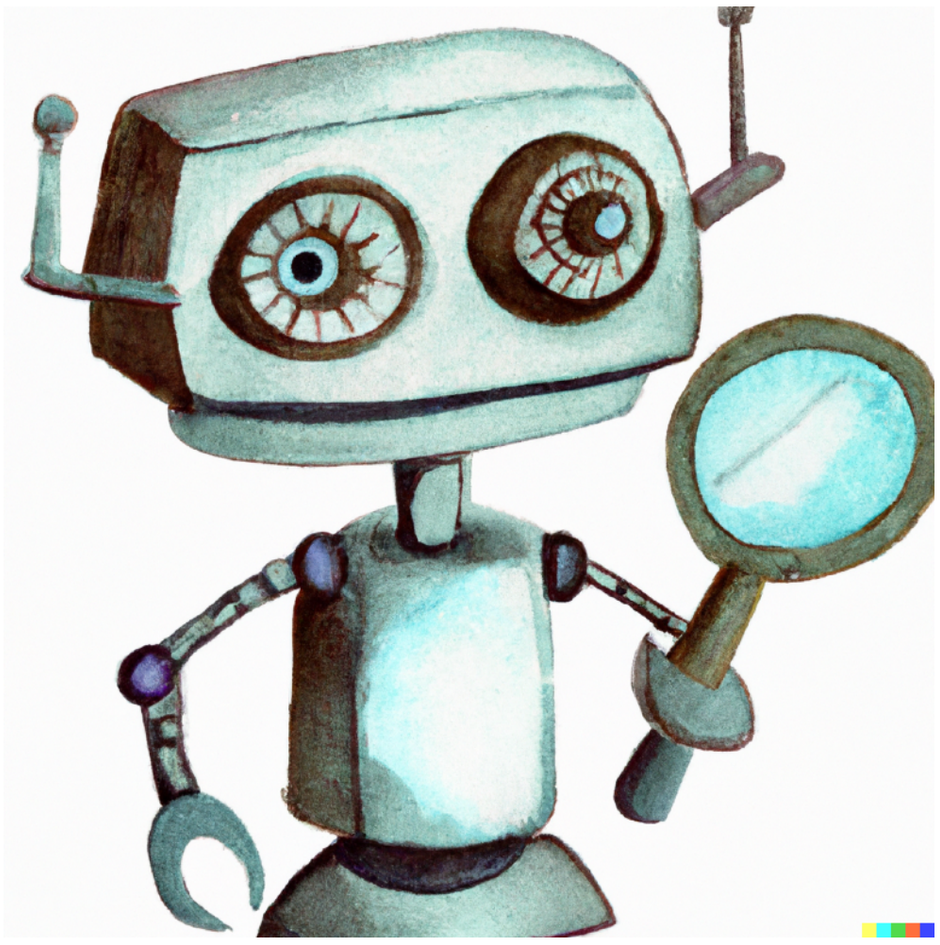
Good fidelity of distributions, very fast and linear scaling with number of particles

Buhmann, GK, Thaler 2301.08128;

Where to go from here?

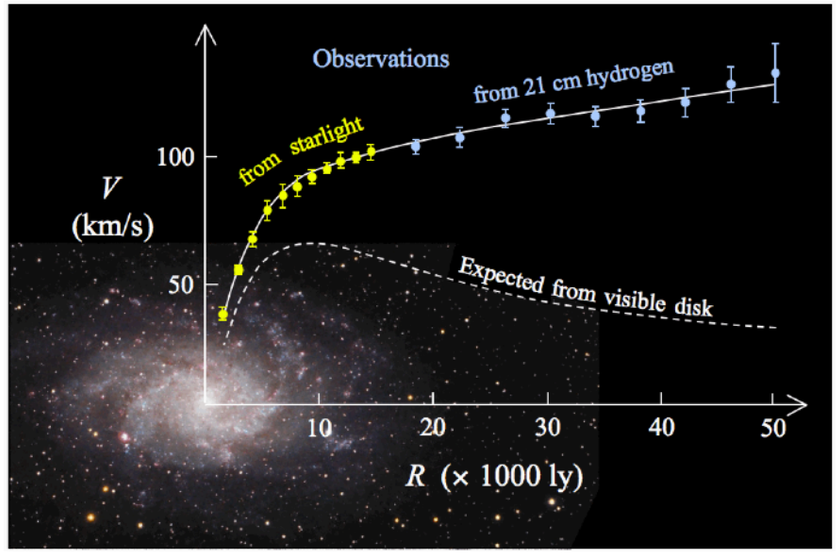
- Extend to larger point clouds
- Increase fidelity
- GANs work, what about other architectures (VAEs, flows, diffusion)?

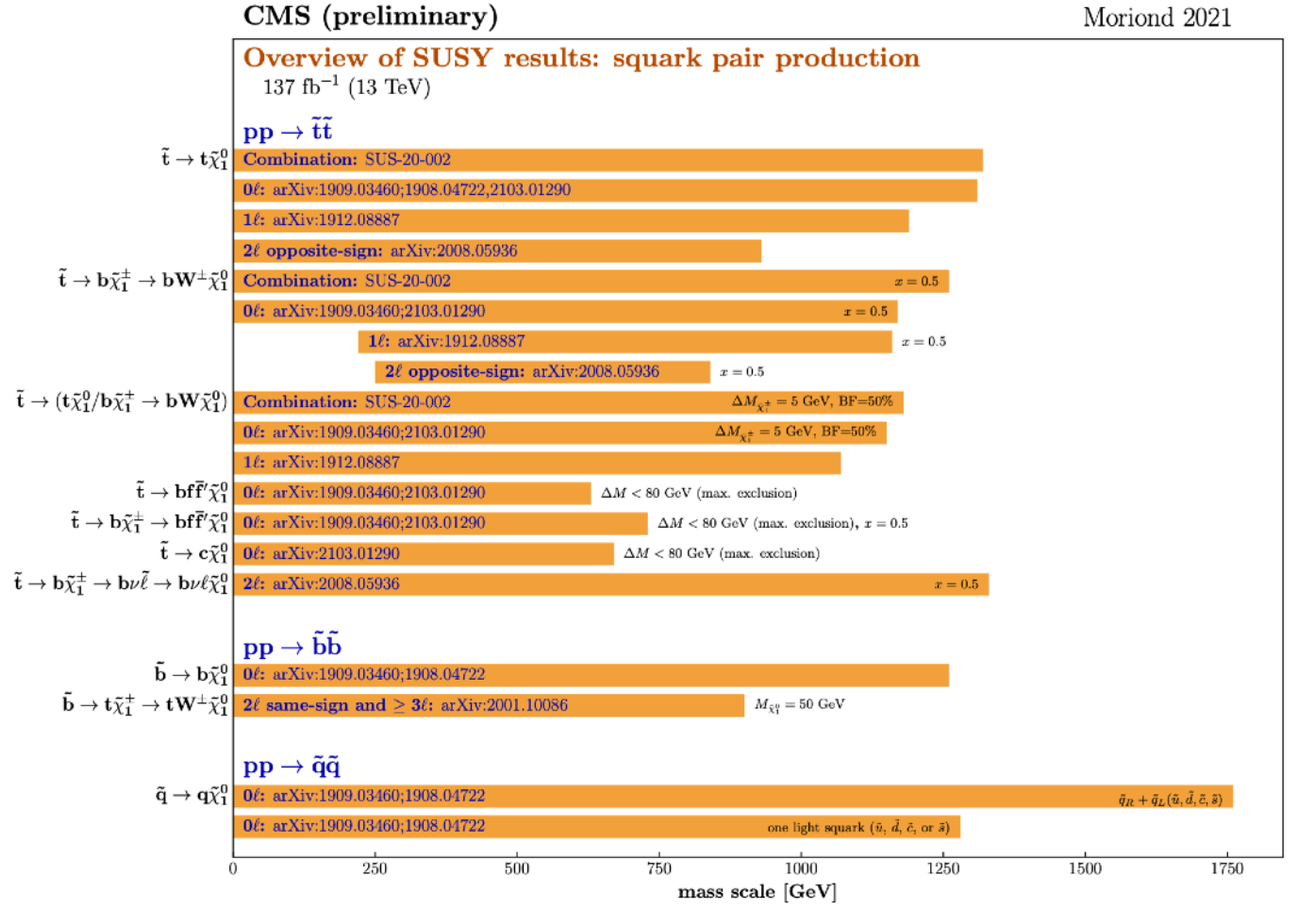
Finding new physics



Motivation

- Theoretical and experimental reasons to expect new physics beyond the Standard Model
- However, so far only negative results in direct (model driven) searches
- Two discovery strategies:
  - Model-specific
  - Model independent
- Machine learning plays a key role in both — focus on anomaly detection now





Selection of observed limits at 95% C.L. (theory uncertainties are not included). Probe **up to** the quoted mass limit for light LSPs unless stated otherwise. The quantities ΔM and x represent the absolute mass difference between the primary sparticle and the LSP, and the difference between the intermediate sparticle and the LSP relative to ΔM, respectively, unless indicated otherwise.

Strategies

- Orthogonal strategy to model specific searches:
  - Discover new physics with making minimal assumptions
- Less sensitive to one specific model, broader coverage

ML-assisted global comparison

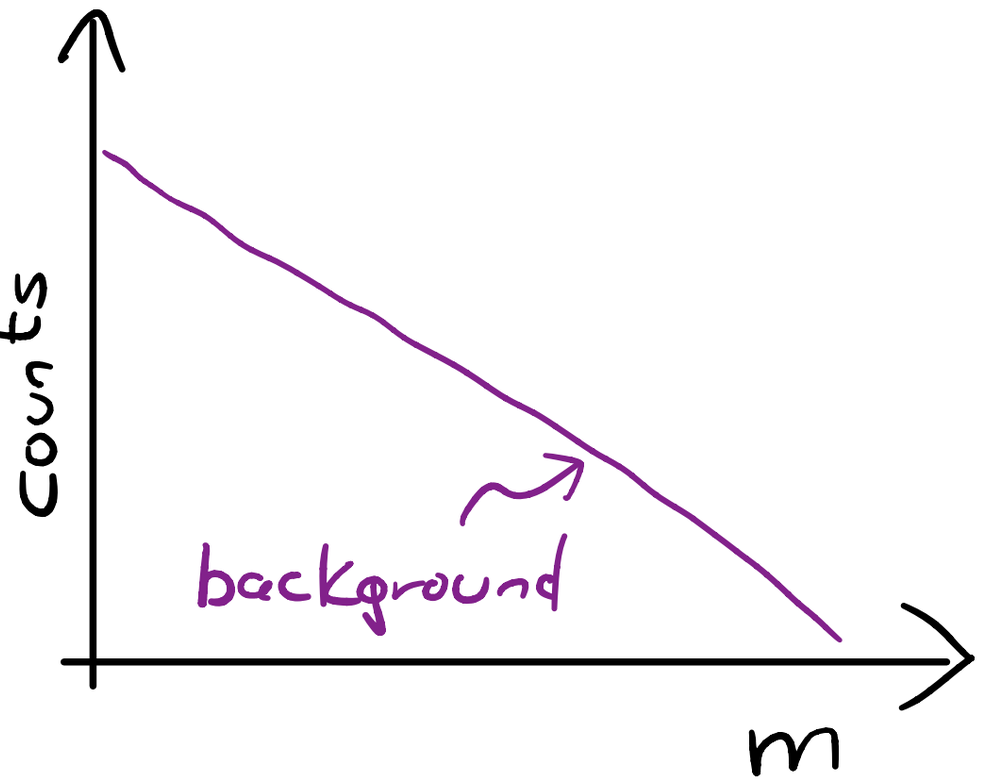
- Systematically compare simulation to recorded data, look for differences
- Con: Rely on imperfect simulation, maximally background model dependent
- Pro: Sensitive to all types of anomalies

Resonant anomaly detection / Enhanced bump hunts

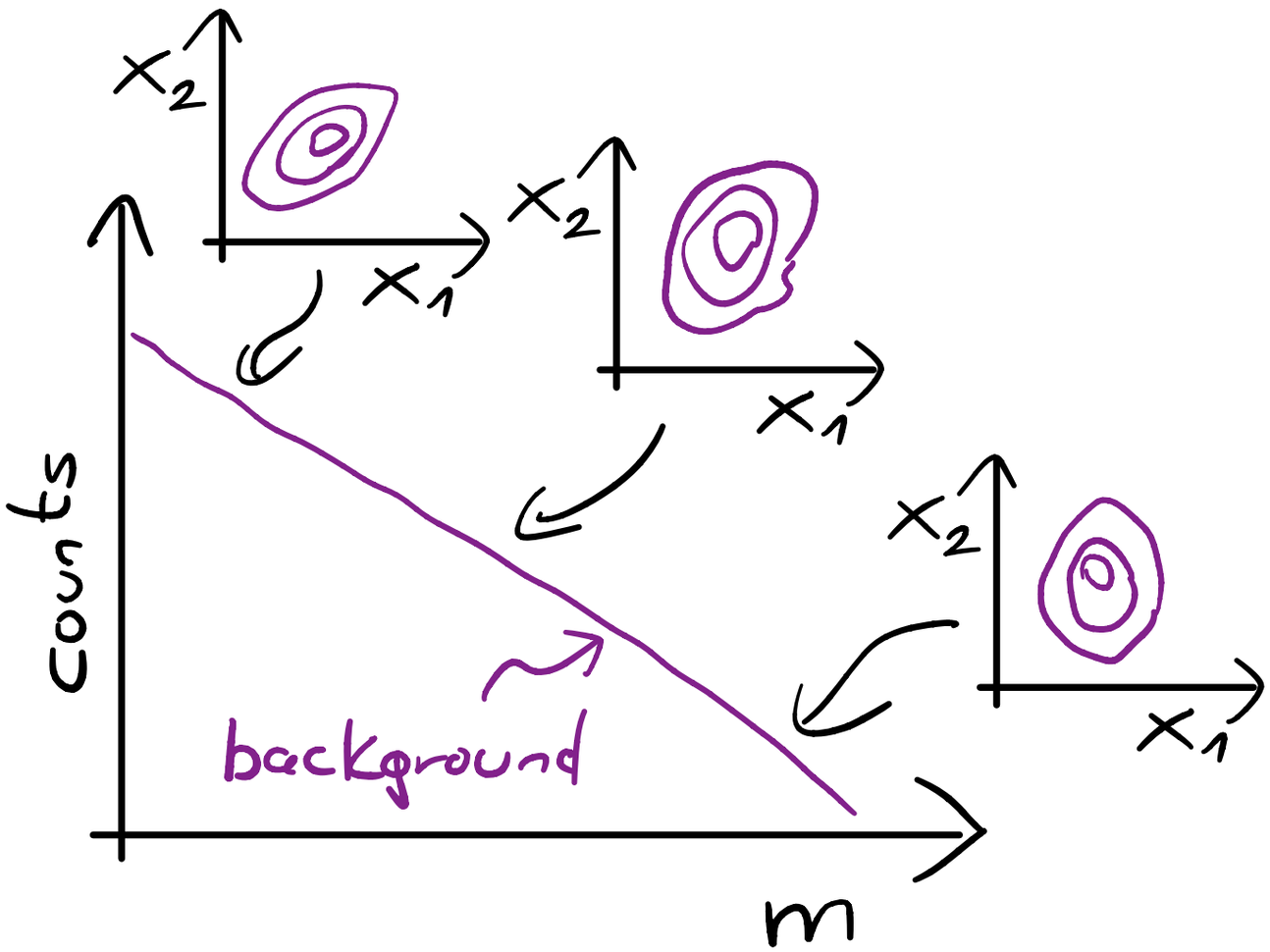
- Estimate background in-situ from data
- Con: Need to make assumptions about signal shape
- Pro: Data-driven on background model

GK et al 2101.08320; Arrested et al 2105.14027; CMS 2010.02984; ATLAS 1807.07447

Resonant Anomaly Detection

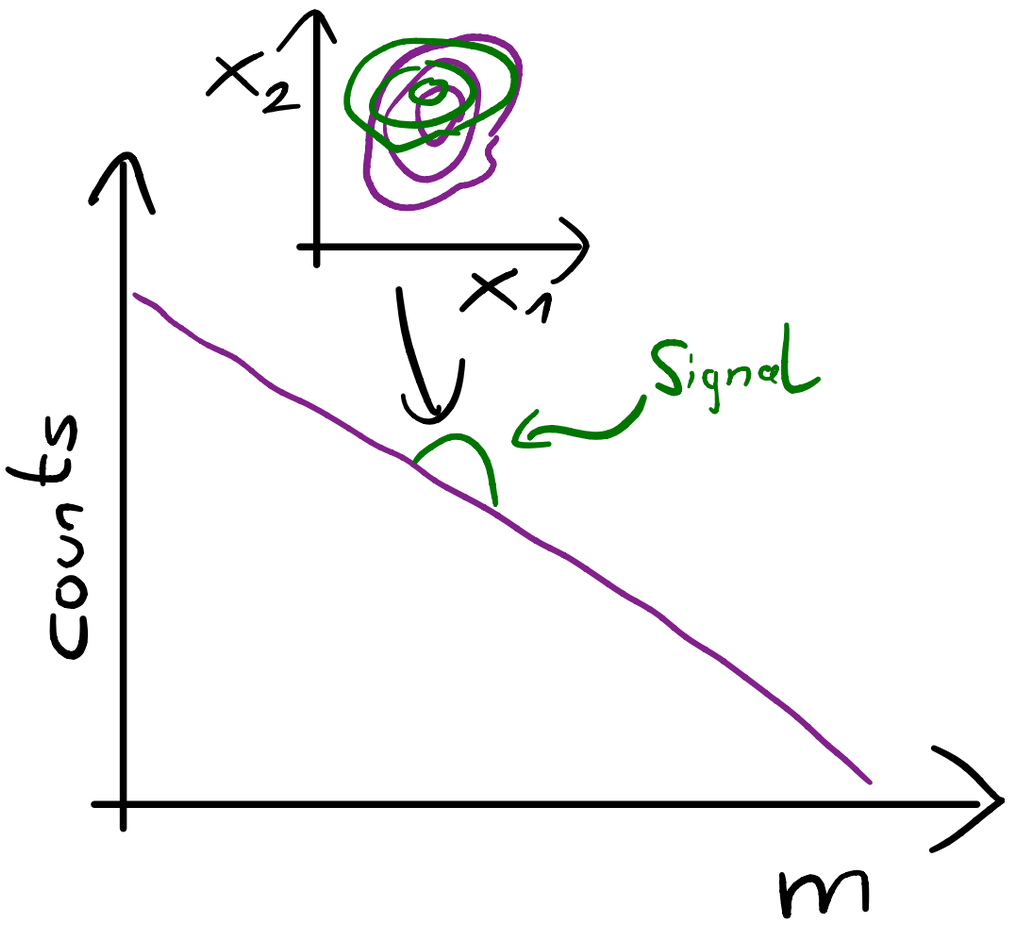


Resonant Anomaly Detection



+ additional dimensions
(in general correlated with m)

Resonant Anomaly Detection



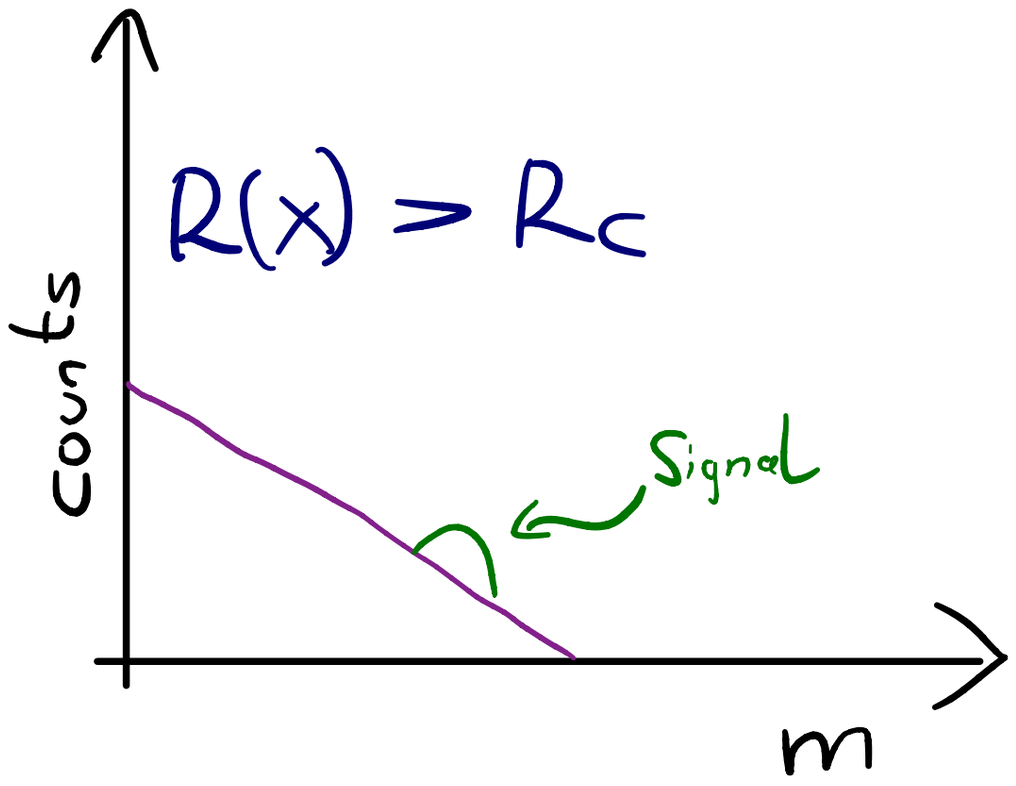
Look for a small signal, localised in m, and different shape in other features

Need to find a feature in which signal is resonant and background smooth.

No assumptions in other features.

Further generalisation as open issue.

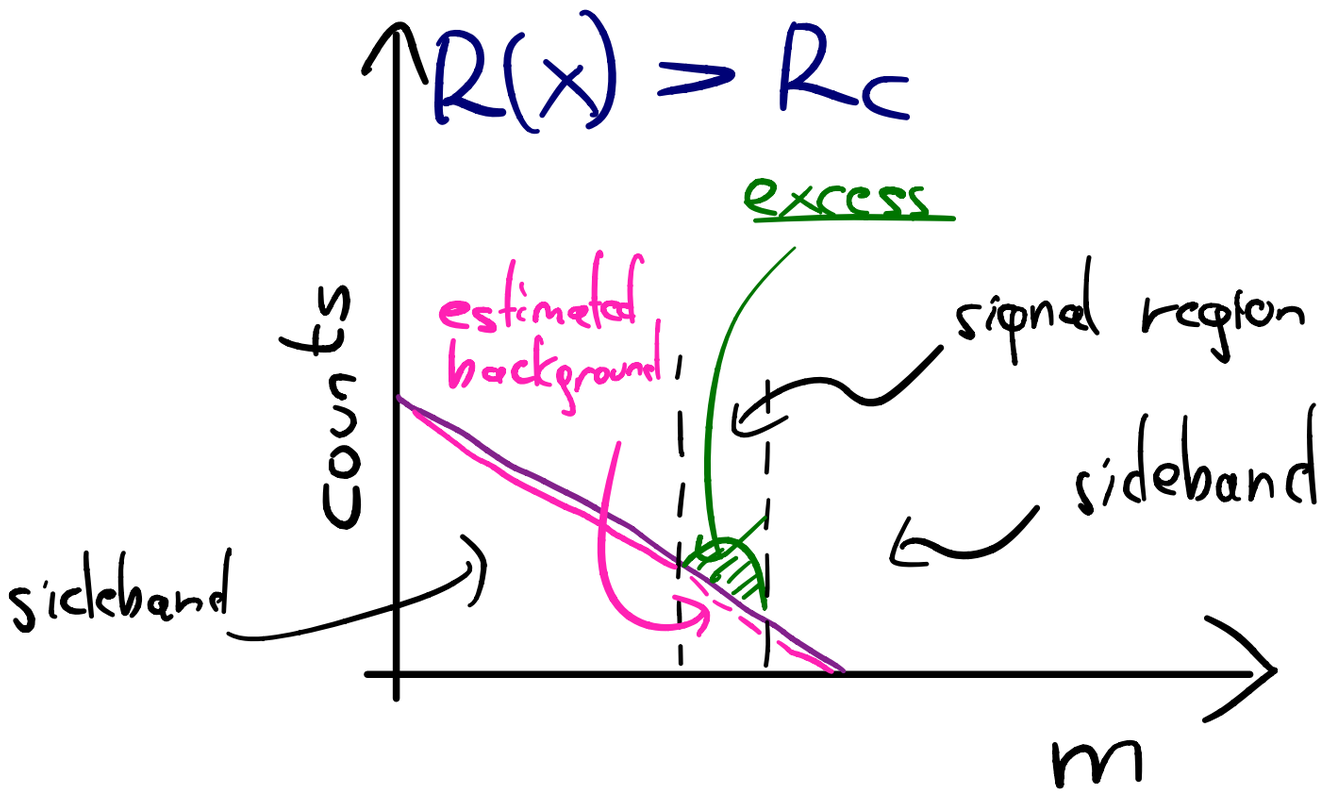
Resonant Anomaly Detection



Enhanced bump-hunt: Use ML to build classifier $R(x)$ so that selecting $R(x) > c$ enhances signal fraction

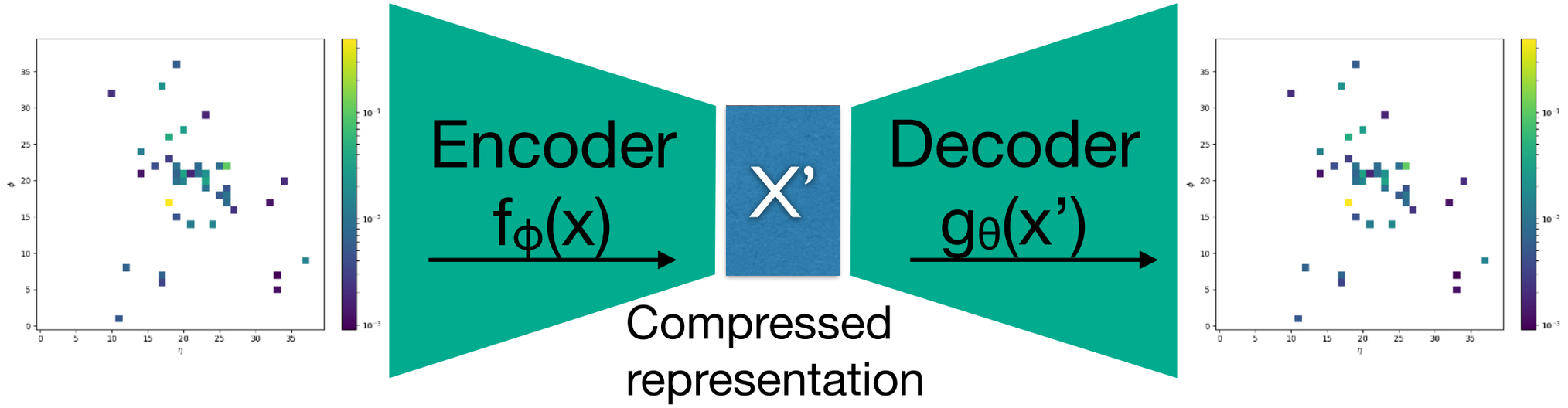
No worry, will come back to HOW
this is done is a moment

Resonant Anomaly Detection



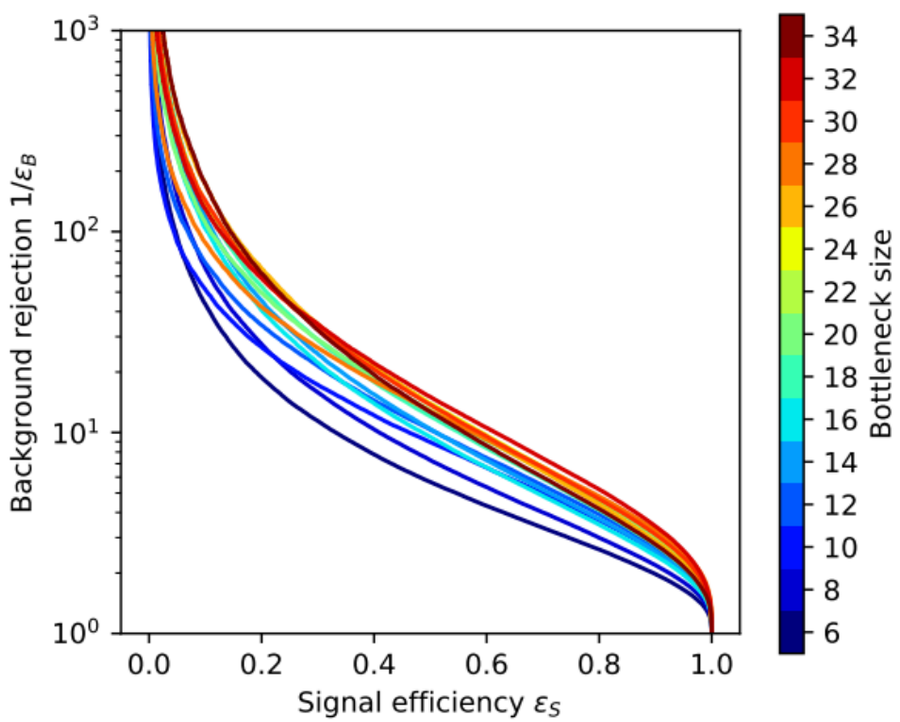
Enhanced bump-hunt: Then fit background from sidebands, compare to data in signal region

Autoencoders



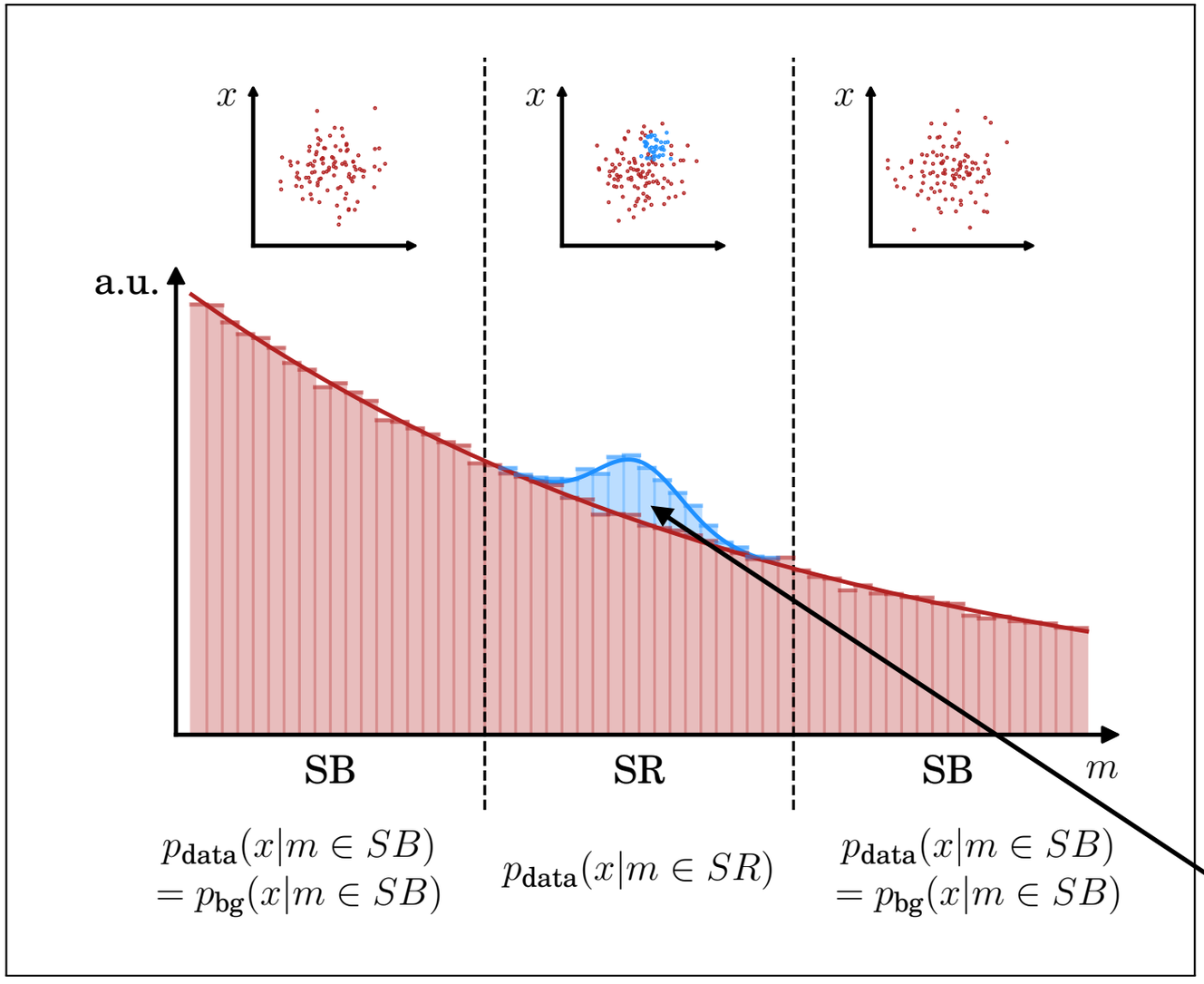
$$\mathcal{L}(x) = ||x - g_\theta(f_\phi(x))||_2$$

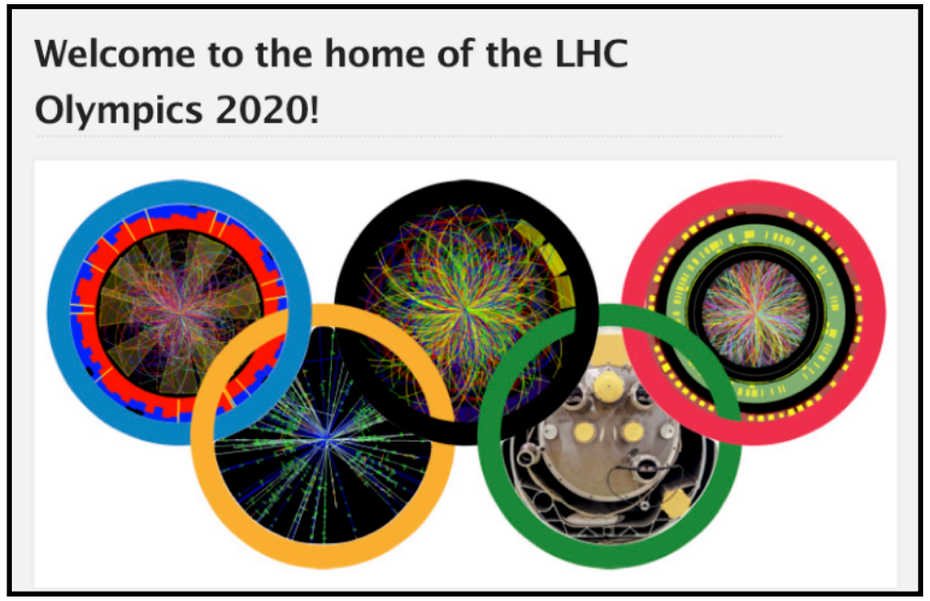
Autoencoders:
Learn-compression/ decompression on signal free sample and use loss as anomaly score



Heimel, GK, et al 1808.08979; Farina et al 1808.08992; …

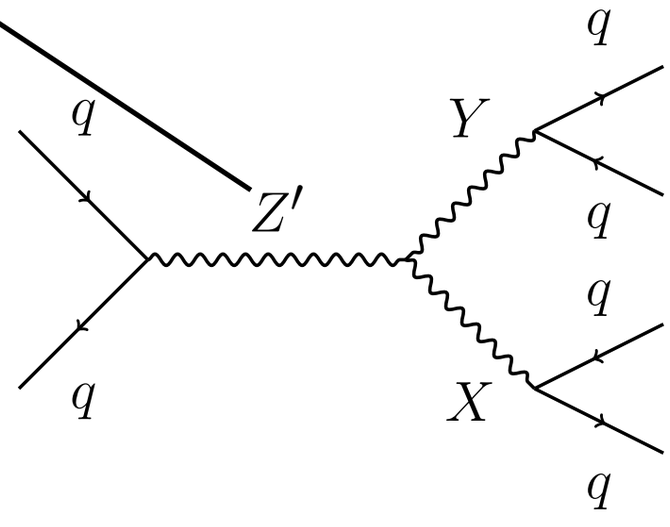
Setup





https://lhco2020.github.io/homepage/

Consider resonant anomalies: reduces generality — but allows fully data-driven construction of anomaly detection score



GK, Nachmann, Shih et al 2101.08320;
Hallin, .., GK et al 2109.00546;

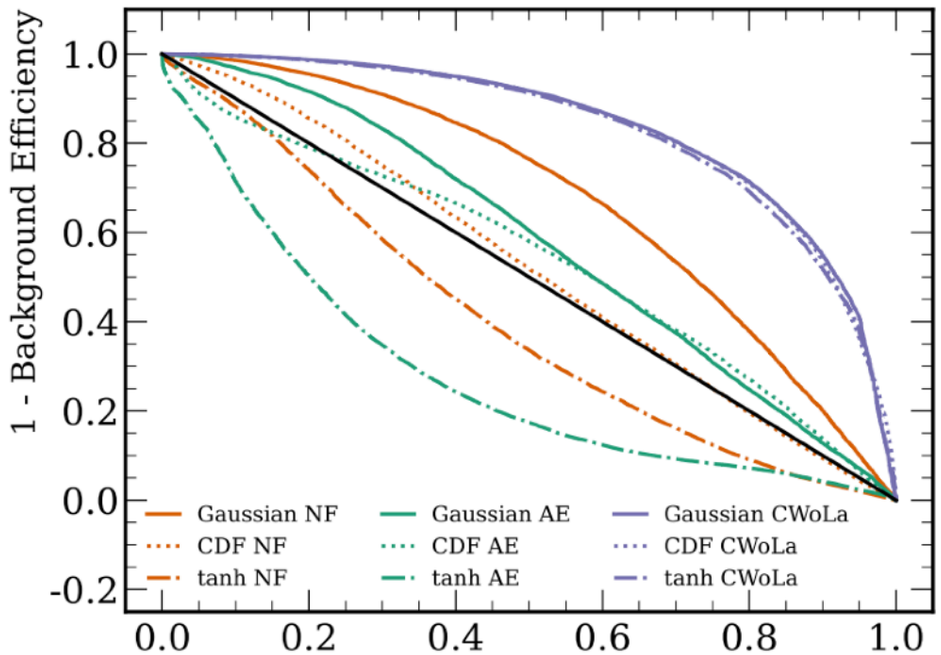
Autoencoder Challenges

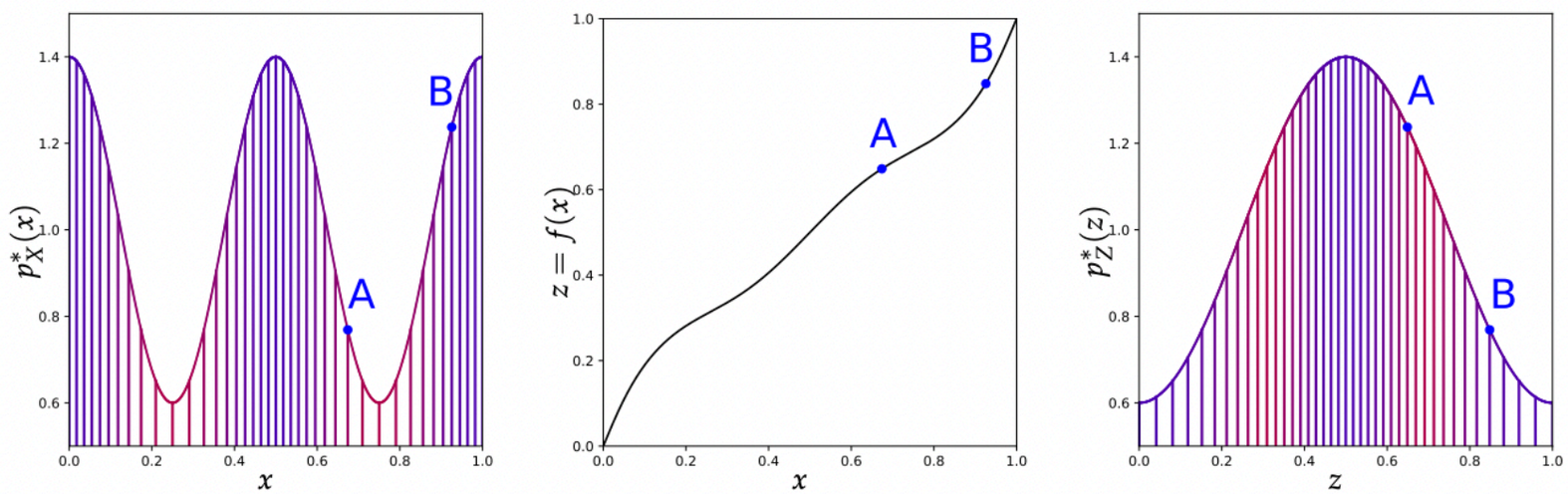
- Complexity bias

  If anomalies are much simpler than backgrounds: L will still be lower, despite never encountered in training. Overcome e.g. with normalising AE

- Change of variables

  Autoencoder approximates background density — not stable under change of variables

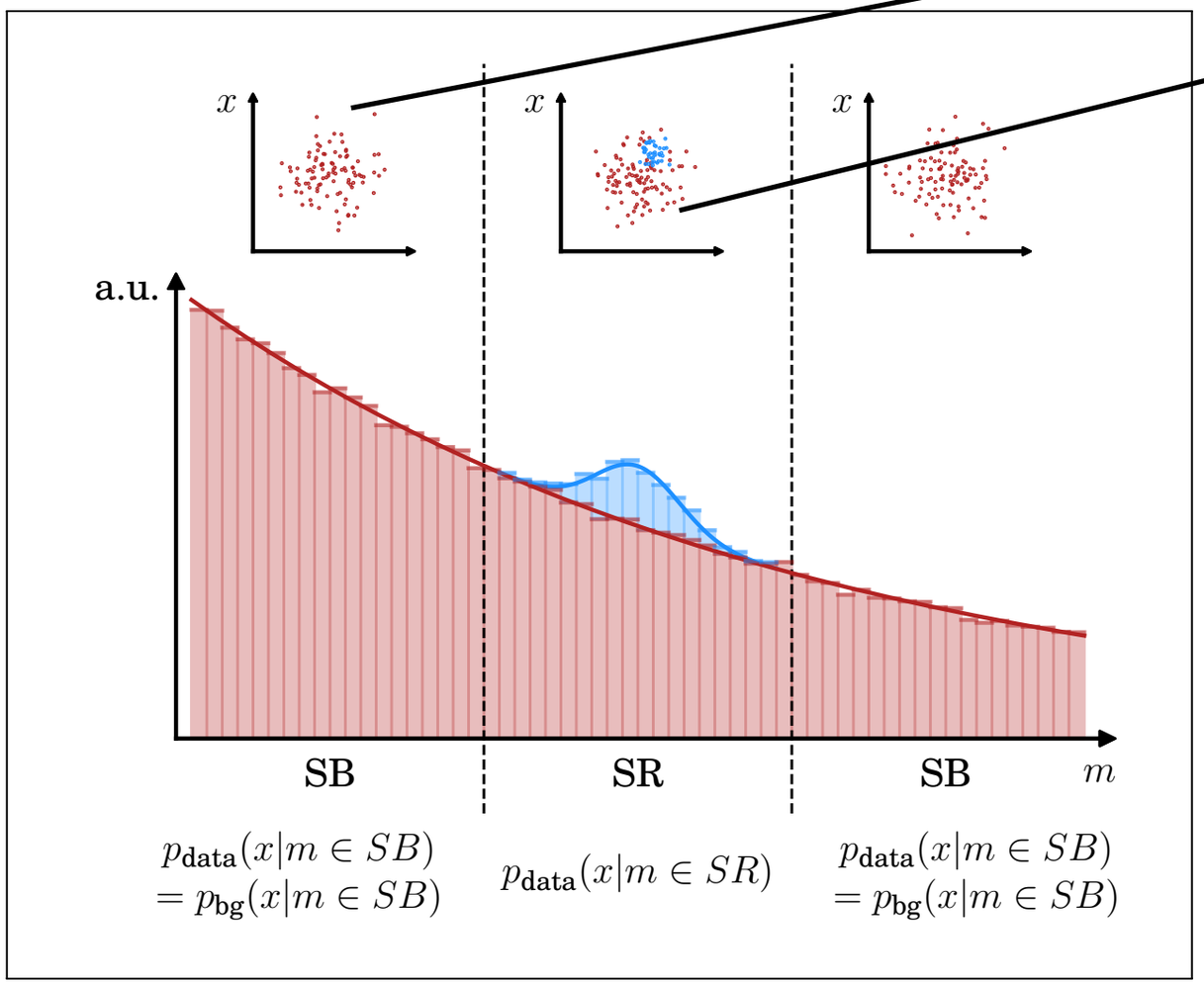


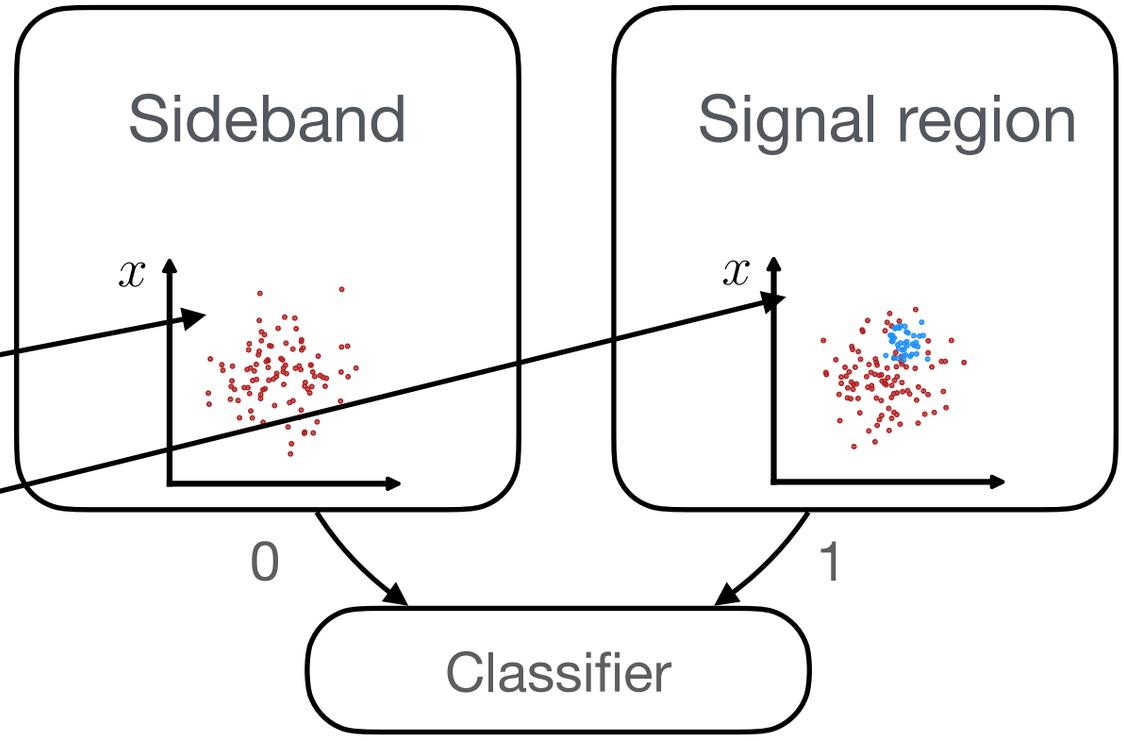


GK, …Shih, et al 2209.06225; Le Lan, Dinh 2012.03808; Weber MSc thesis; Finke et al 2104.09051; Dillon et al 2206.14225;

Weak supervision

$$L_{M_1/M_2} = \frac{p_{M_1}}{p_{M_2}} = \frac{f_1\, p_S + (1-f_1)\, p_B}{f_2\, p_S + (1-f_2)\, p_B} = \frac{f_1\, L_{S/B} + (1-f_1)}{f_2\, L_{S/B} + (1-f_2)}$$

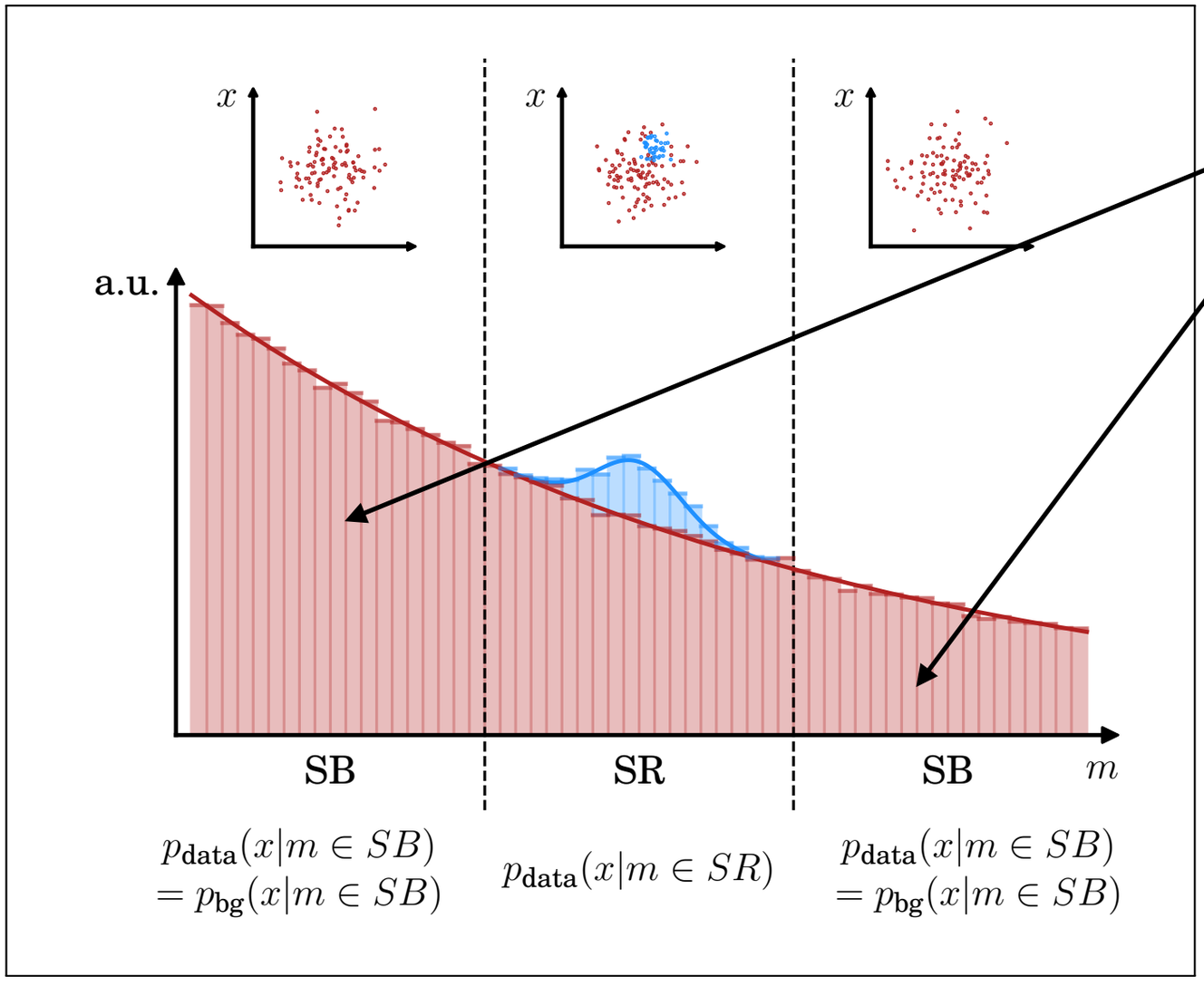




- Train a classifier between side-band and signal region
- Discrimination possible if different process
- Requires no-correlation for background

Metodiev, Nachman, Thaler, 1708.02949; Howe, Nachman 1805.02664

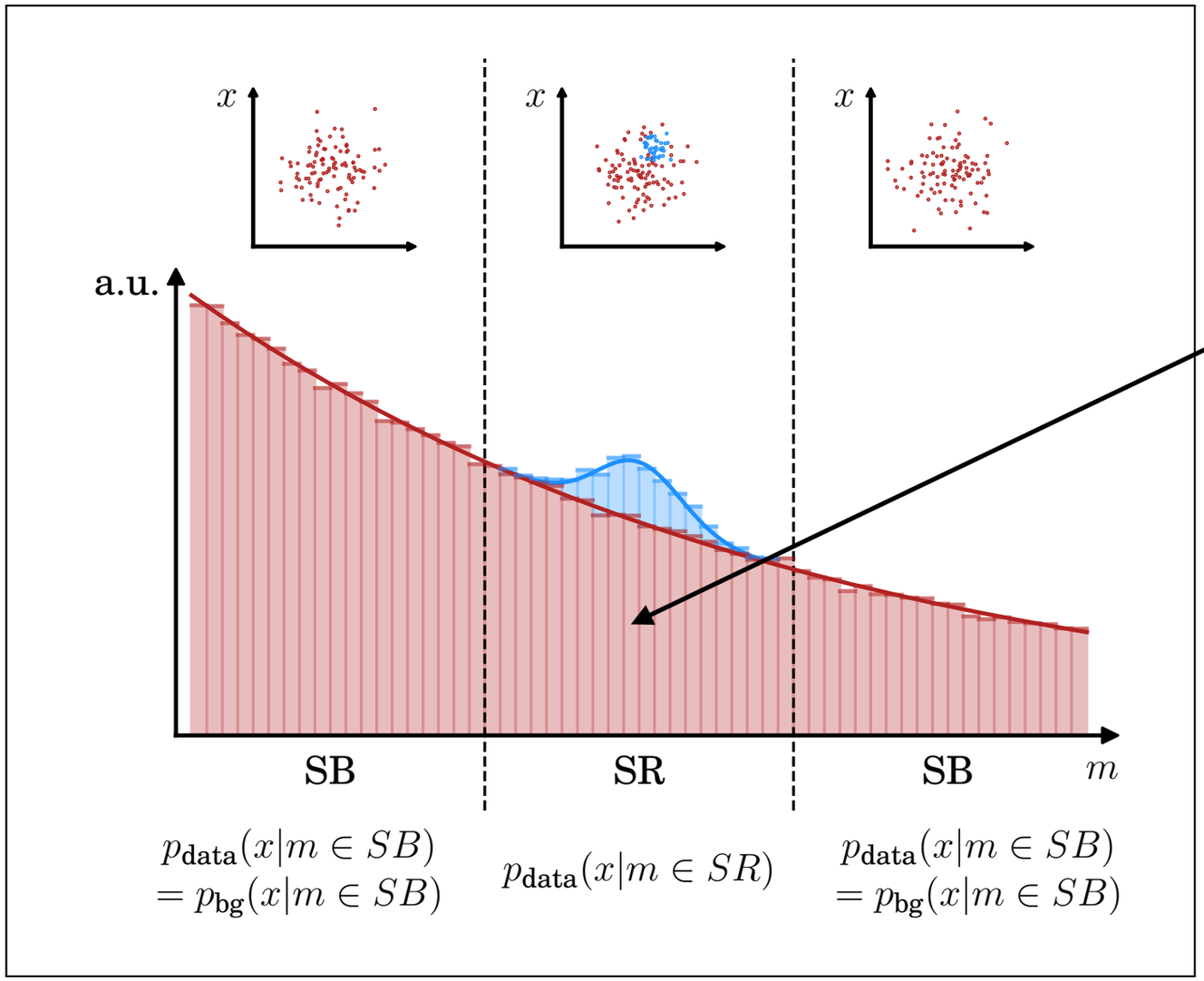
CATHODE & Friends



Train a generative model here (conditional normalising flow)

GK, Nachmann, Shih et al 2101.08320;
Hallin, .., GK et al 2109.00546;

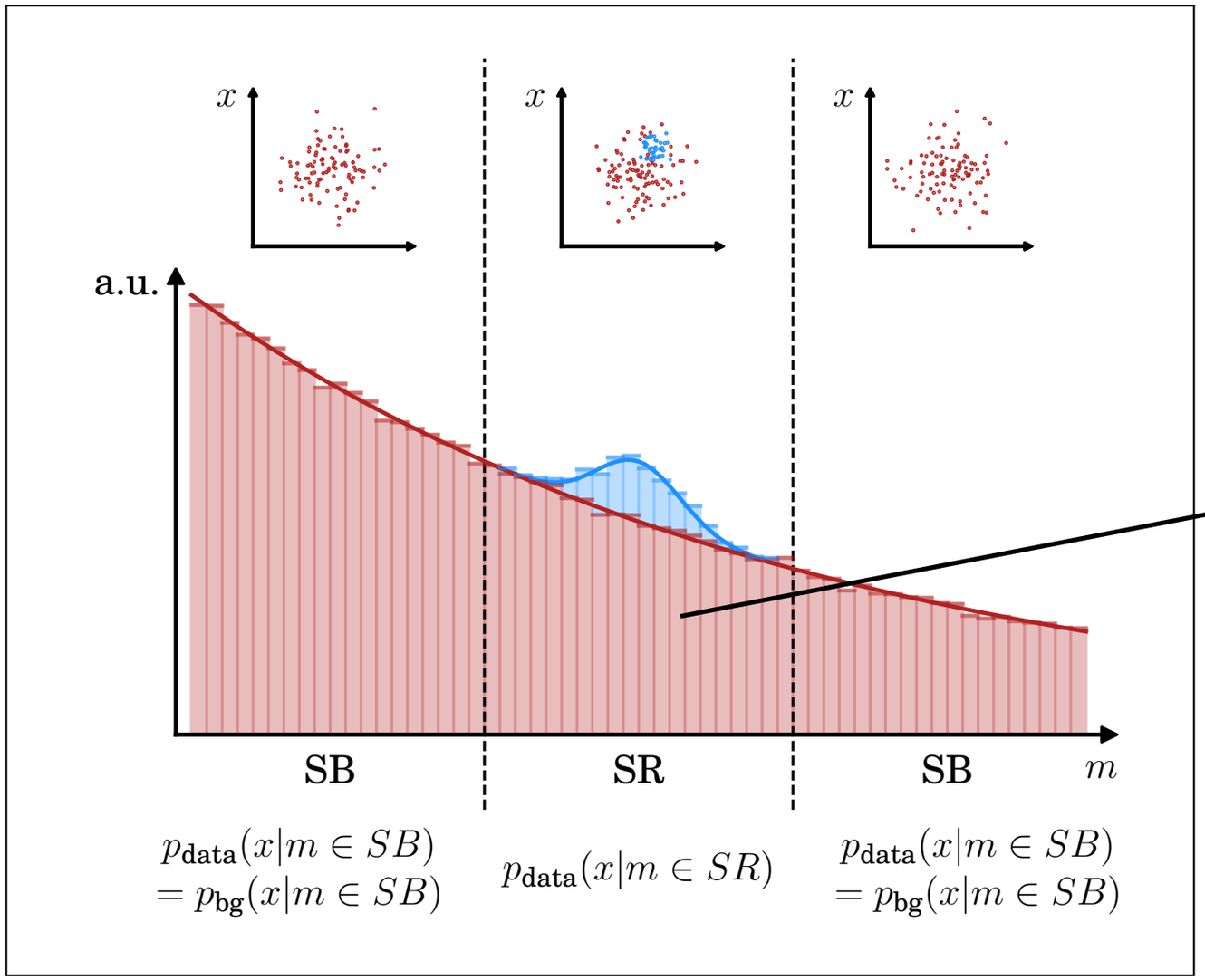
CATHODE & Friends



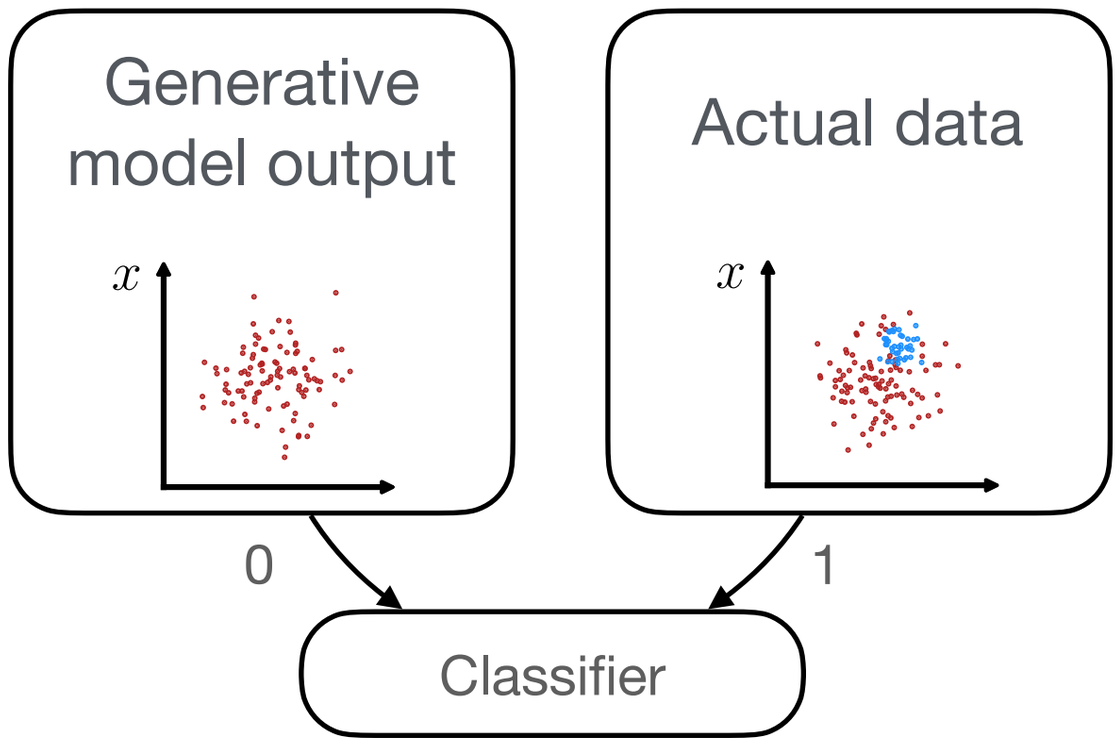
Interpolate &
and sample here

GK, Nachmann, Shih et al 2101.08320;
Hallin, .., GK et al 2109.00546;

CATHODE & Friends

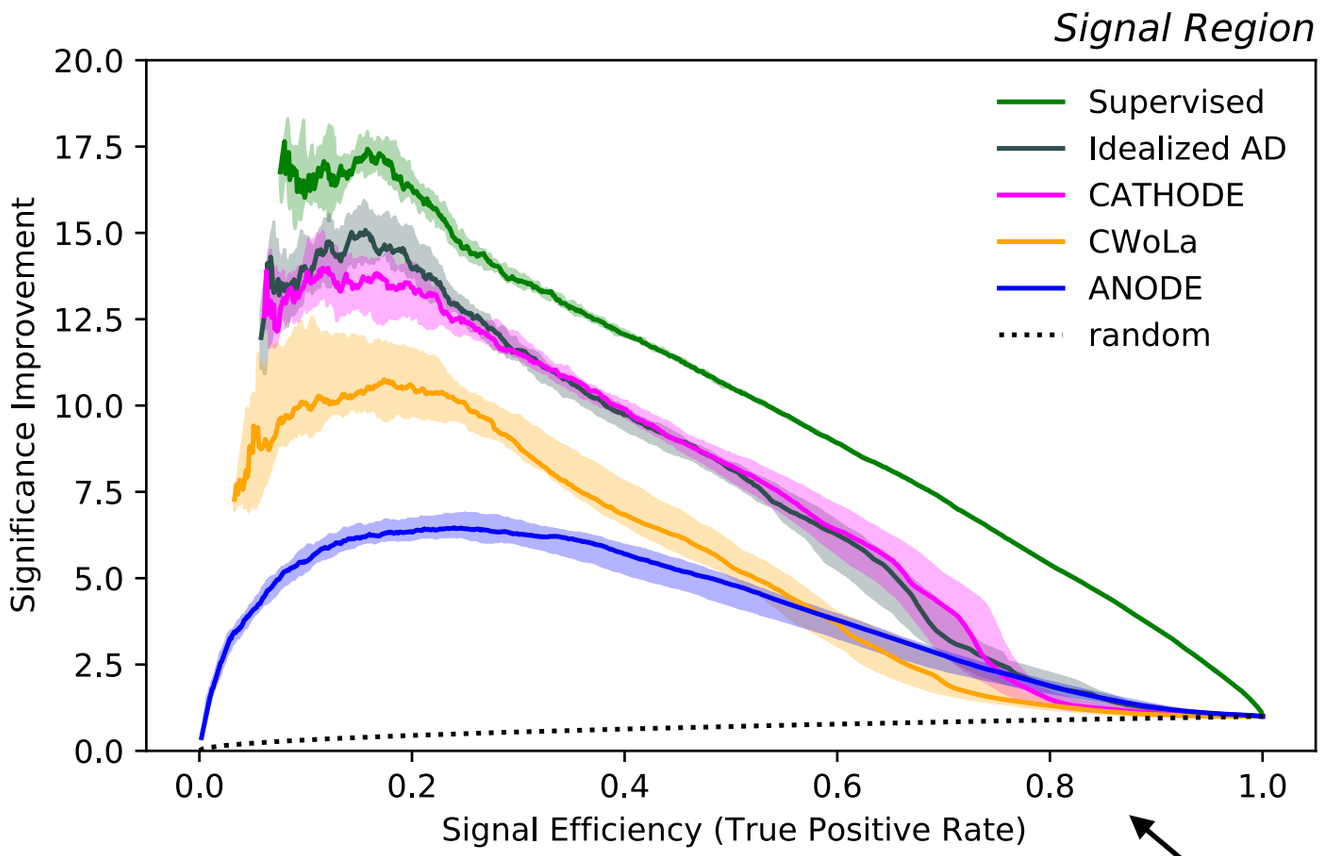


Train a binary classifier between prediction & data

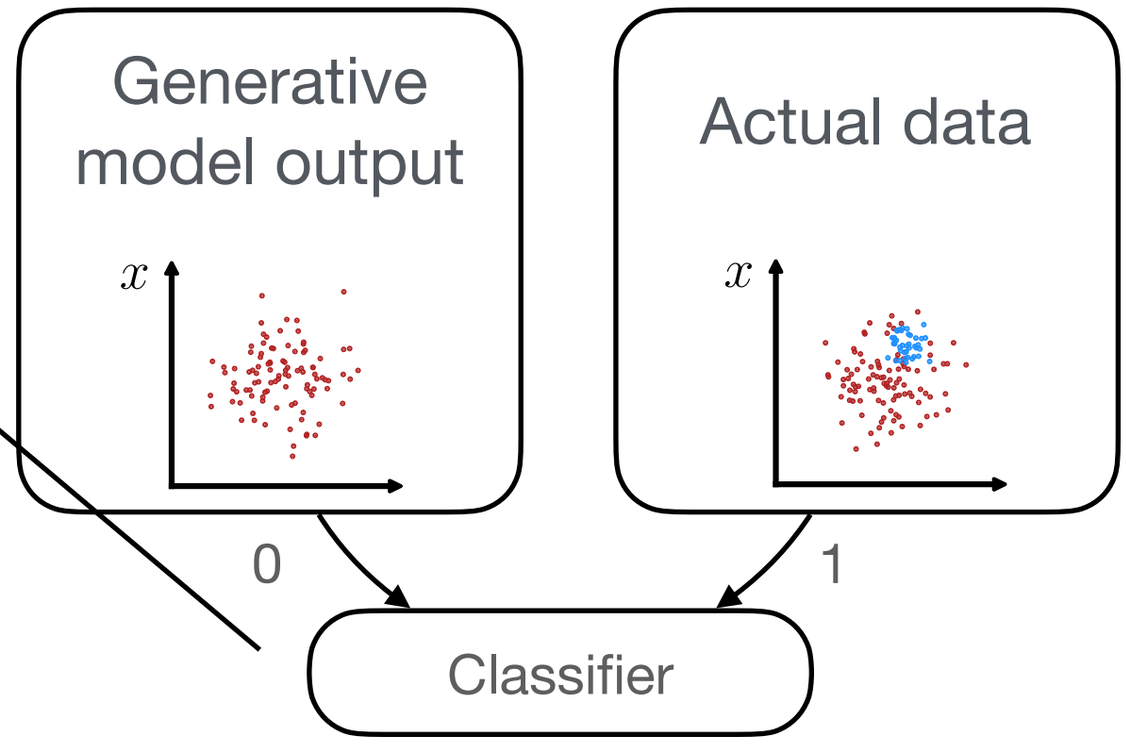


GK, Nachmann, Shih et al 2101.08320;
Hallin, .., GK et al 2109.00546;

CATHODE & Friends

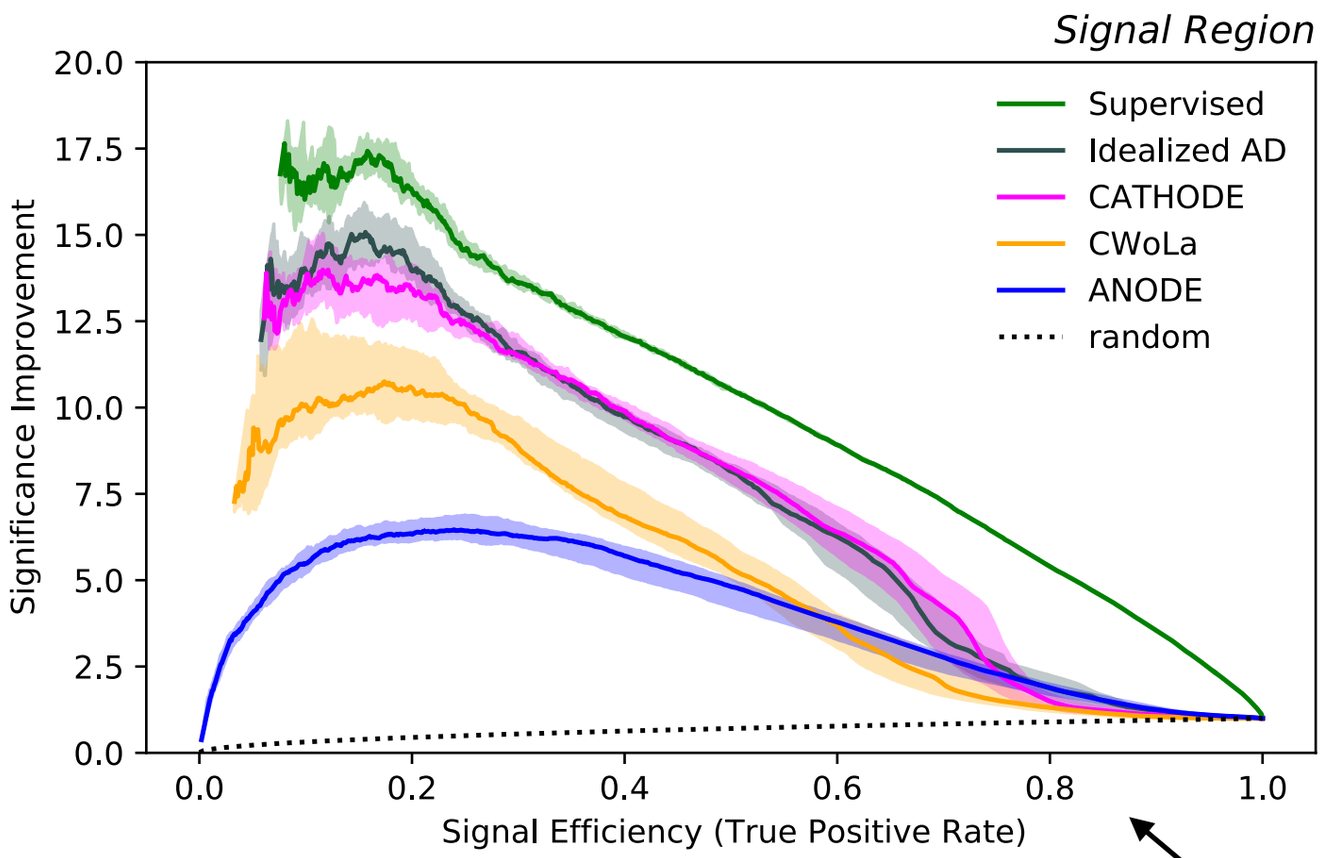


Cut on classifier output to identify possible anomalies

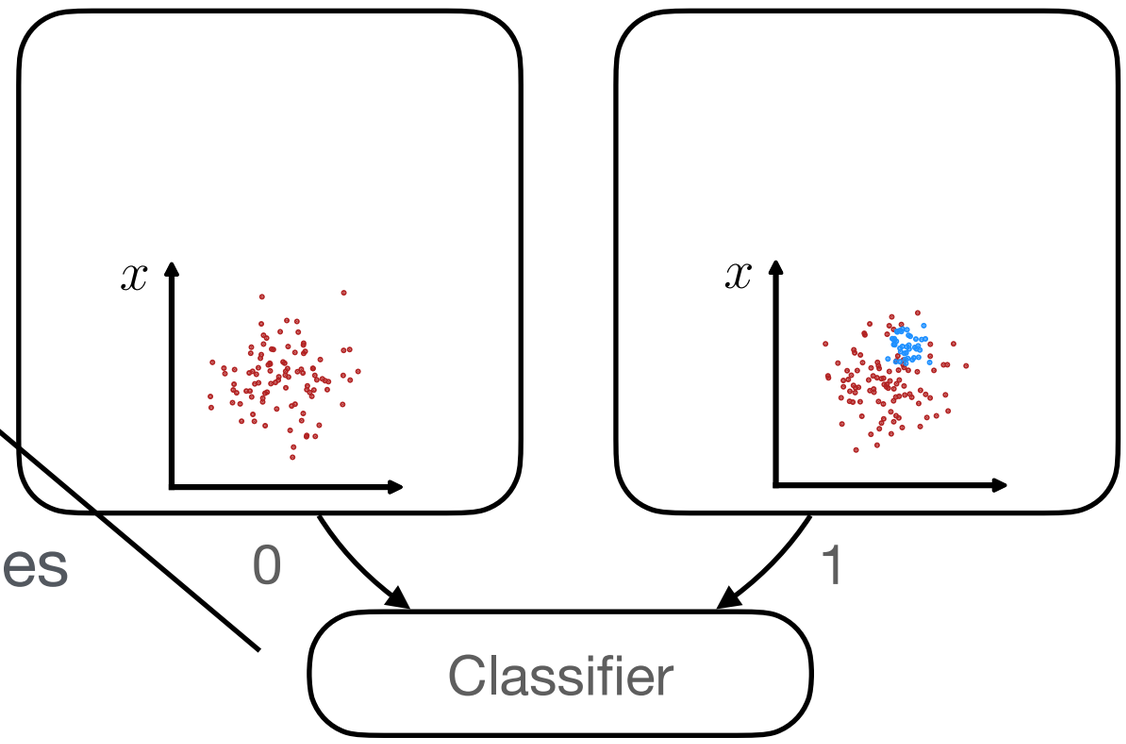


GK, Nachmann, Shih et al 2101.08320;
Hallin, .., GK et al 2109.00546;

CATHODE & Friends



- Most sensitive current resonant anomaly detection technique
- Stability improvement: coming next
- Currently preparing application on CMS data

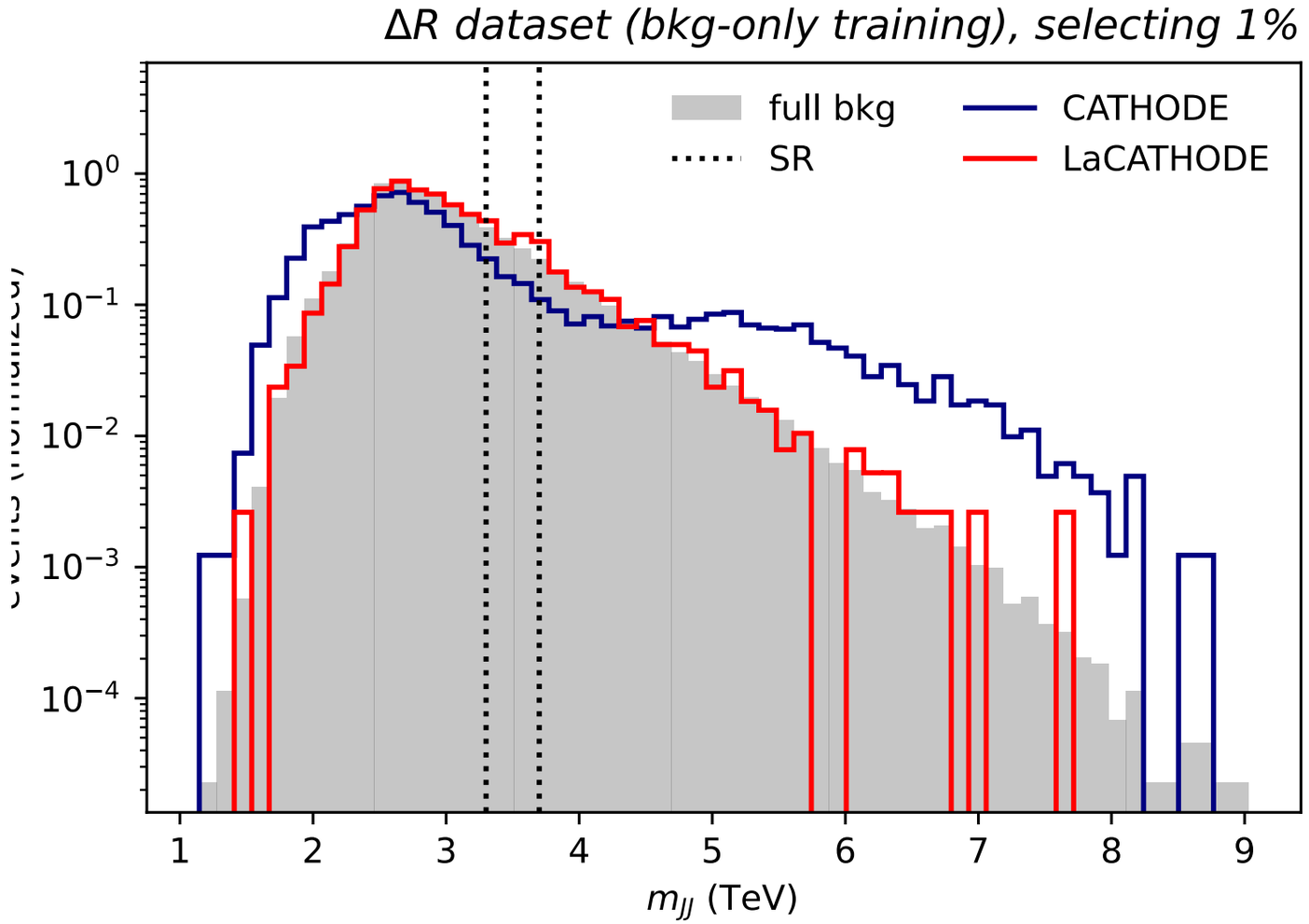


Cut on output to enhance possible anomalies

GK, Nachmann, Shih et al 2101.08320; Hallin, .., GK et al 2109.00546; Hallin, GK, et al 2210.14924

LaCATHODE

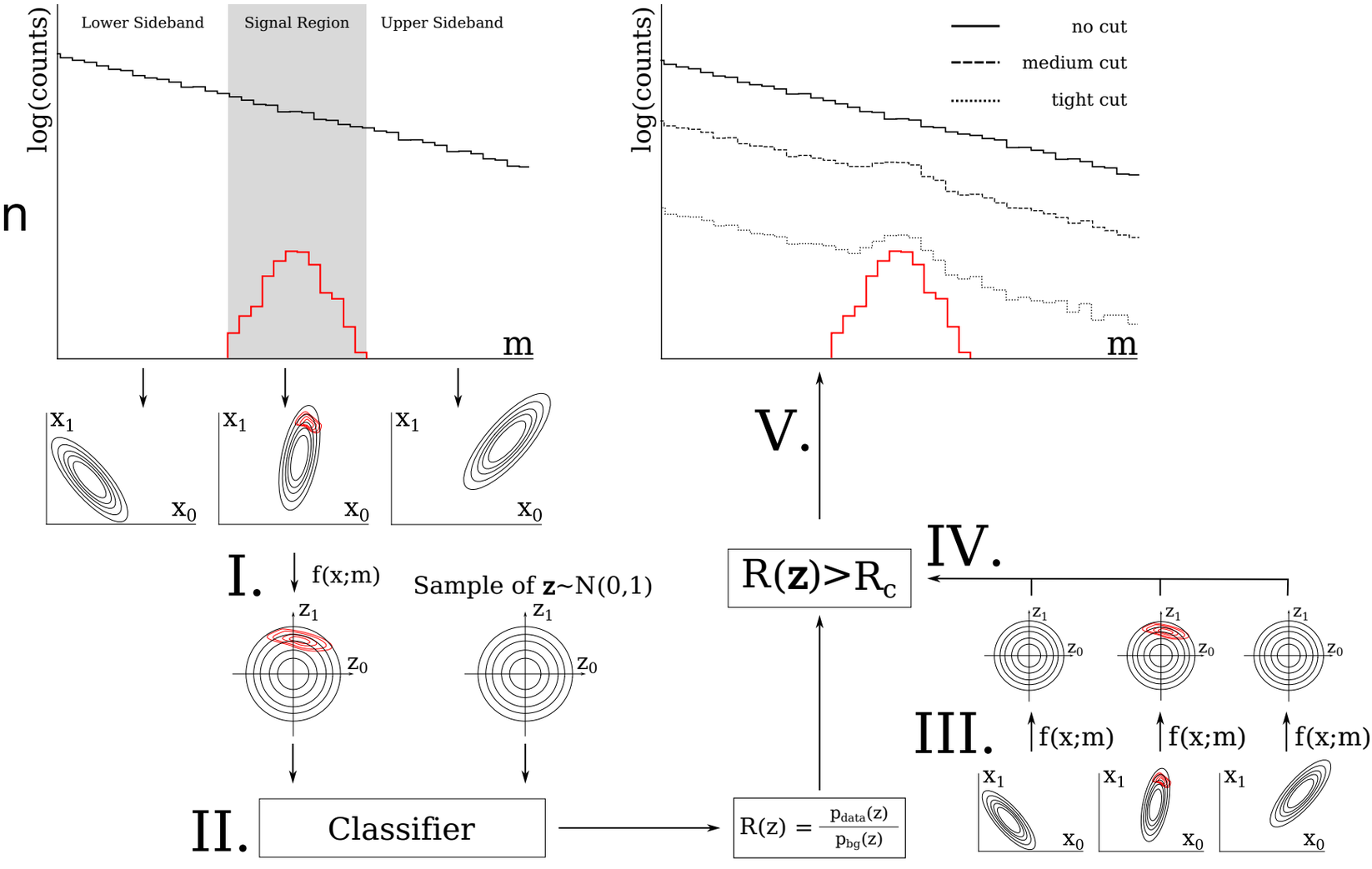
- If R(x) is only calculated in signal region, it's extrapolation is not well-defined
- Potential problem for bump-hunt if it shapes distributions



Hallin, GK, Shih, et al, 2210.14924

LaCATHODE

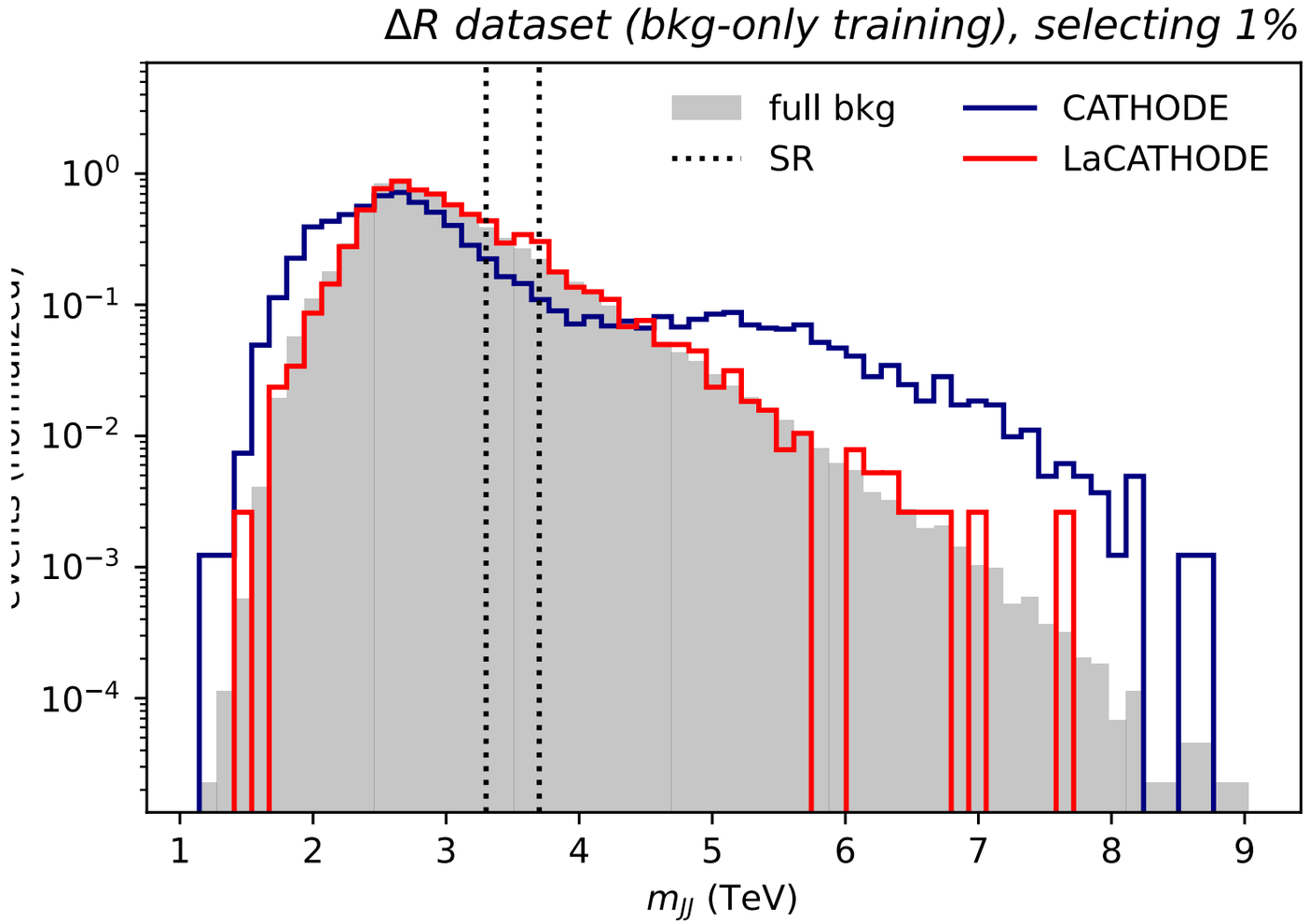
- If R(x) is only calculated in signal region, it's extrapolation is not well-defined
- Potential problem for bump-hunt if it shapes distributions
- Instead, train classifier in latent space



Hallin, GK, Shih, et al, 2210.14924

LaCATHODE

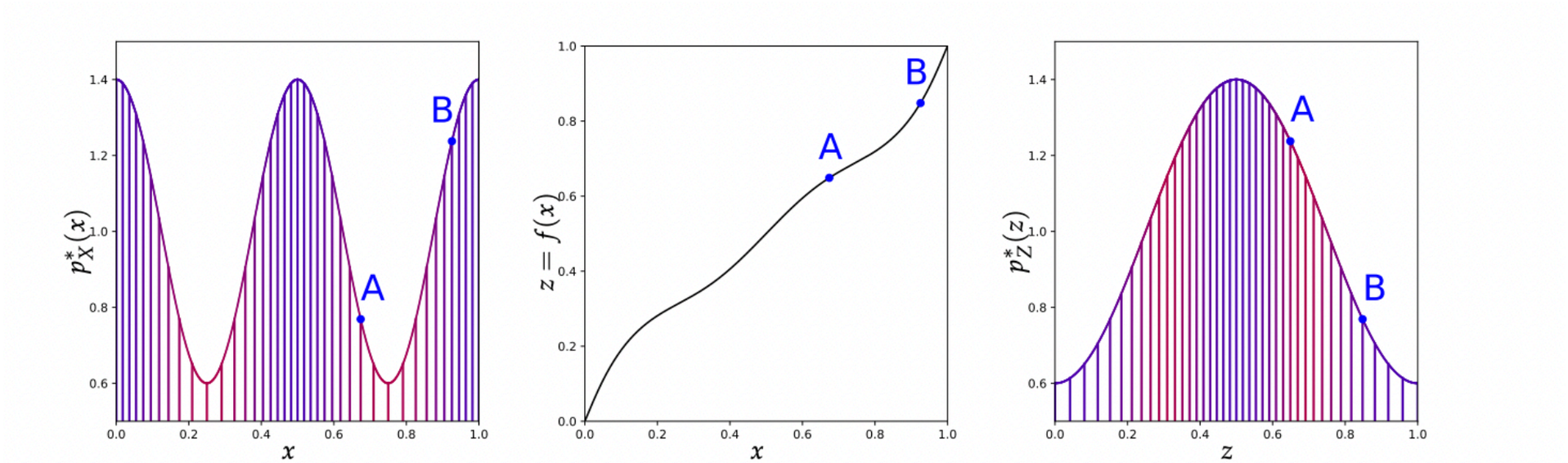
- If R(x) is only calculated in signal region, it's extrapolation is not well-defined
- Potential problem for bump-hunt if it shapes distributions
- Instead, train classifier in latent space to achieve flat distributions



Hallin, GK, Shih, et al, 2210.14924

Comments on anomaly detection

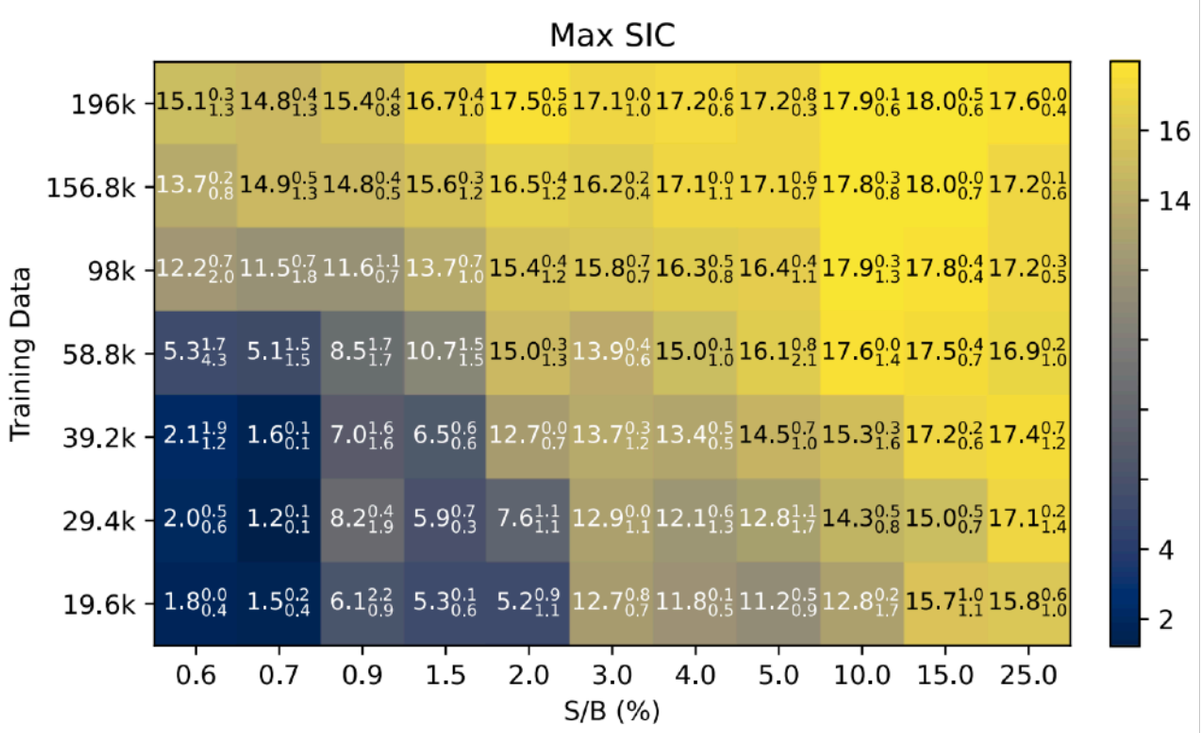
- As CATHODE approximates a likelihood ratio, it should be robust compared to methods that only use $p_{\text{Background}}$ (e.g. autoencoders)



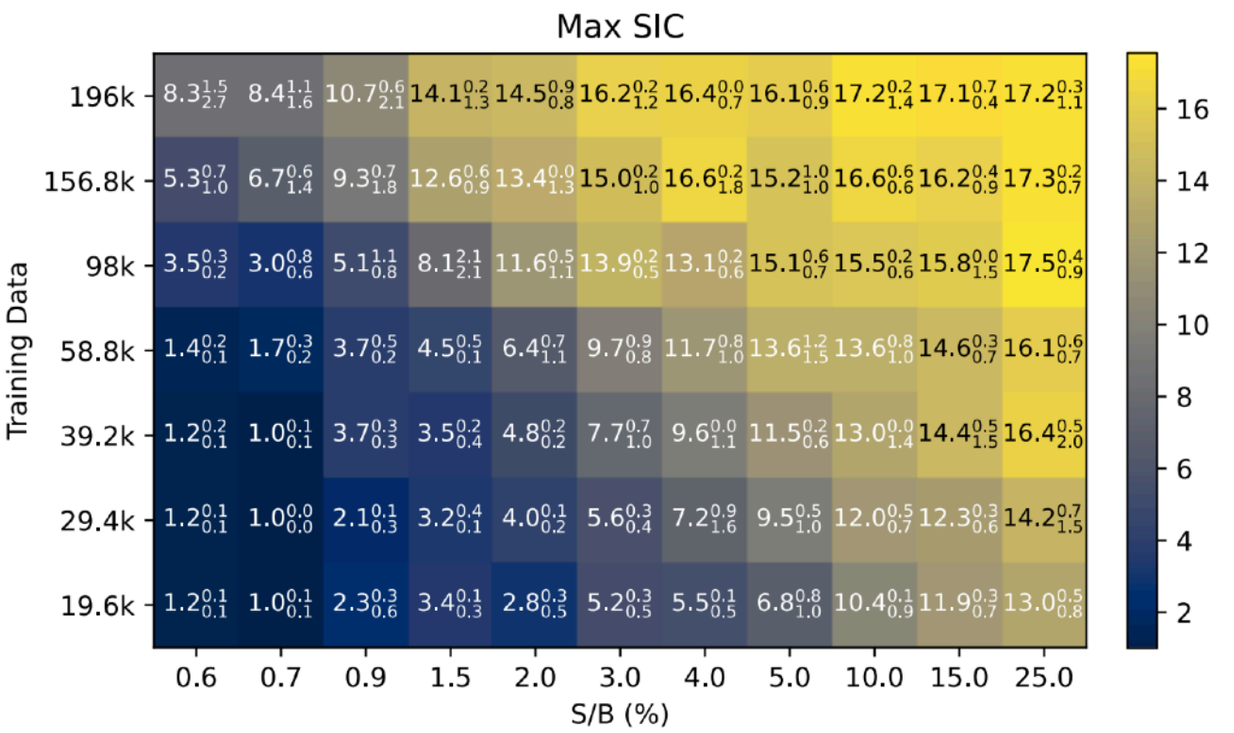
2012.03808 (Lan & Dinh)

Comments on anomaly detection

- As CATHODE approximates a likelihood ratio, it should be robust compared to methods that only use $p_{\text{Background}}$ (e.g. autoencoders)
- However, still can be sensitive to choice of input features
  - Here shown: idealised anomaly detector (perfect DE)



No noisy features



Two noisy features

Comments on anomaly detection

- As CATHODE approximates a likelihood ratio, it should be robust compared to methods that only use $p_{Background}$ (e.g. autoencoders)
- However, still can be sensitive to choice of input features

- Need also consider
  - Higher dimensional input data
  - Stable training
  - Holistic ‘end-to-end’ setups

Closing & future outlook

Conclusions

- Large potential and wide range of applications of machine learning in particle physics

- Key issue of including physics knowledge in ML training:
  - Either by starting from physics observables and constructing efficient selections
  - or by building architectures that are invariant/equivariant under relevant symmetries

- Recent progress in anomaly detection — but scaling stable techniques to more complex data representations still unsolved

Thank you!

Backup

Distance Correlation

$$\begin{aligned}\mathrm{dCov}^2(\vec{X},\vec{Y}) := {} & \mathrm{E}[\|\vec{X}-\vec{X}'\|\,\|\vec{Y}-\vec{Y}'\|]+ \\ & \mathrm{E}[\|\vec{X}-\vec{X}'\|]\;\mathrm{E}[\|\vec{Y}-\vec{Y}'\|] \\ & -2\,\mathrm{E}[\|\vec{X}-\vec{X}'\|\,\|\vec{Y}-\vec{Y}''\|].\end{aligned}$$

$$\mathrm{dCor}^2(\vec{X},\vec{Y}) = \frac{\mathrm{dCov}^2(\vec{X},\vec{Y})}{\sqrt{\mathrm{dCov}^2(\vec{X},\vec{X})\;\mathrm{dCov}^2(\vec{Y},\vec{Y})}},$$

$$\overline{\mathrm{dCor}}^2(\vec{X},\vec{Y}) = \mathrm{dCor}^2(\Sigma_X^{-1/2}\vec{X},\Sigma_Y^{-1/2}\vec{Y}).$$

Covariance matrixes

Das, GK, Shih 2212.00046;