Machine Learning in the Search for New Fundamental Physics

Gregor Kasieczka
gregor.kasieczka@uni-hamburg.de
Twitter: @GregorKasieczka

ACAT 2022, 25.10.2022

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

PUNCH 4NFDI

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

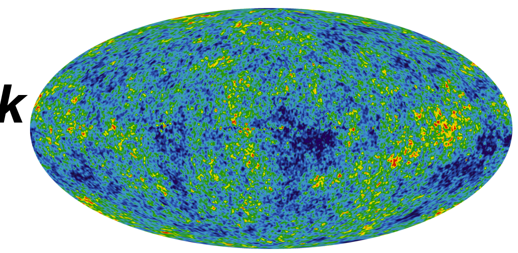
Bundesministerium für Bildung und Forschung

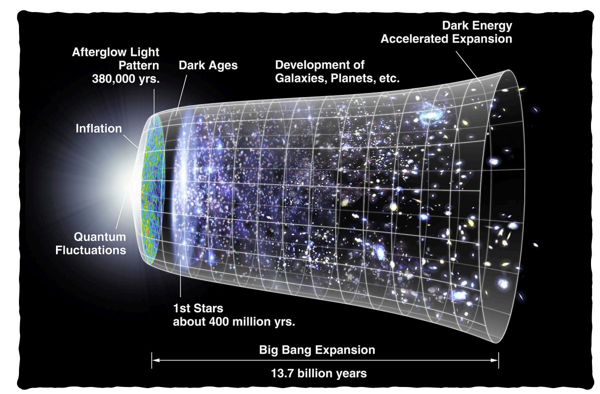
Motivation

- Theoretical and experimental reasons to expect new physics beyond the Standard Model

Why are neutrinos massive?

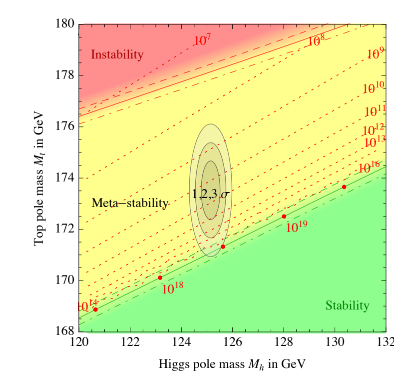
What is the nature of dark matter & dark energy?





What are the origins of the LHCb flavour anomaly?

Why is there more matter than anti-matter?



Why is there more matter than anti-matter?

Is the electroweak vacuum stable?

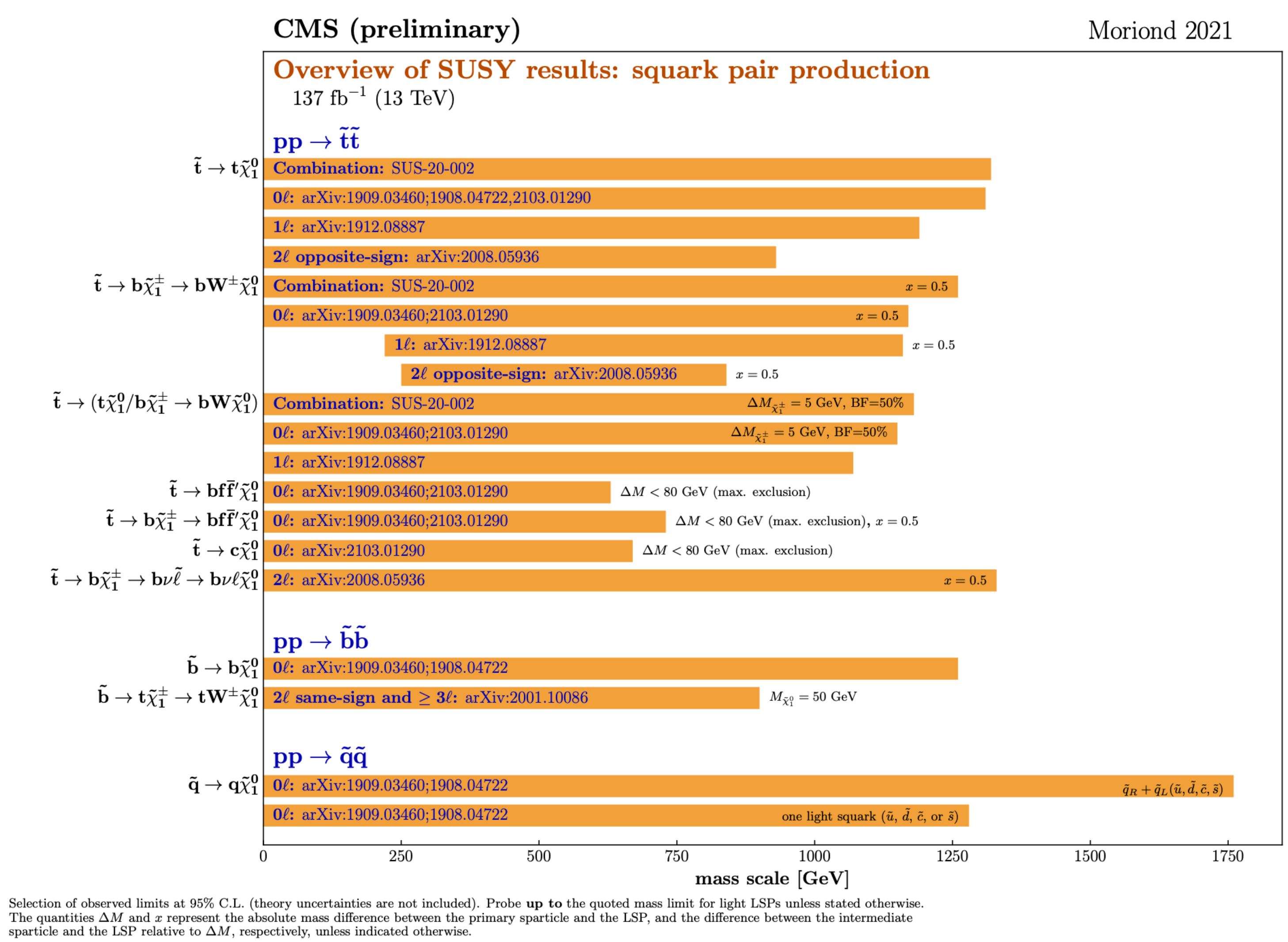
How can the Higgs boson be light when the mass receives large quantum corrections?

What are the details of cosmic inflation?

Motivation

- Theoretical and experimental reasons to expect new physics beyond the Standard Model

- However, so far only negative results in searches

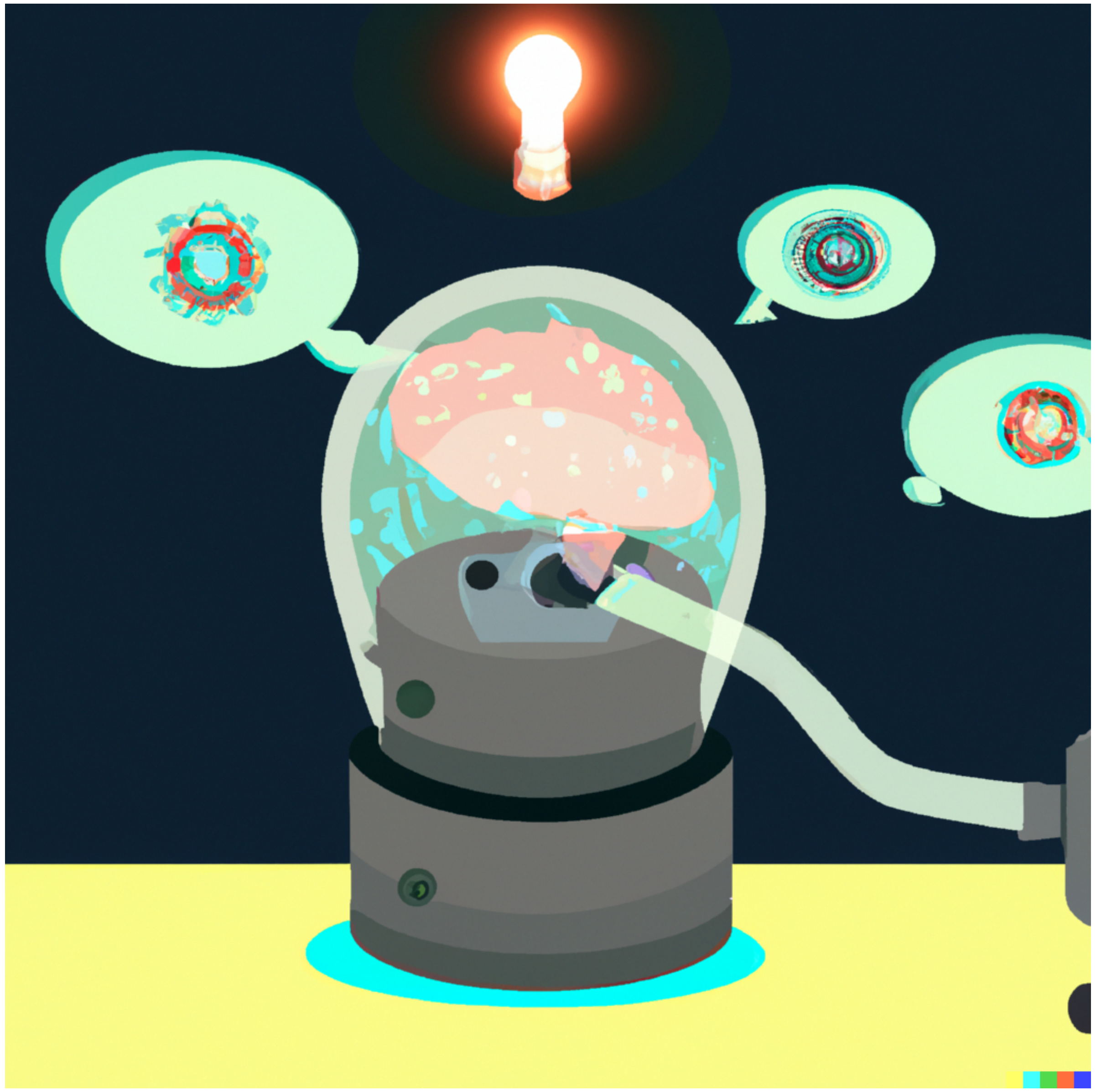


Motivation

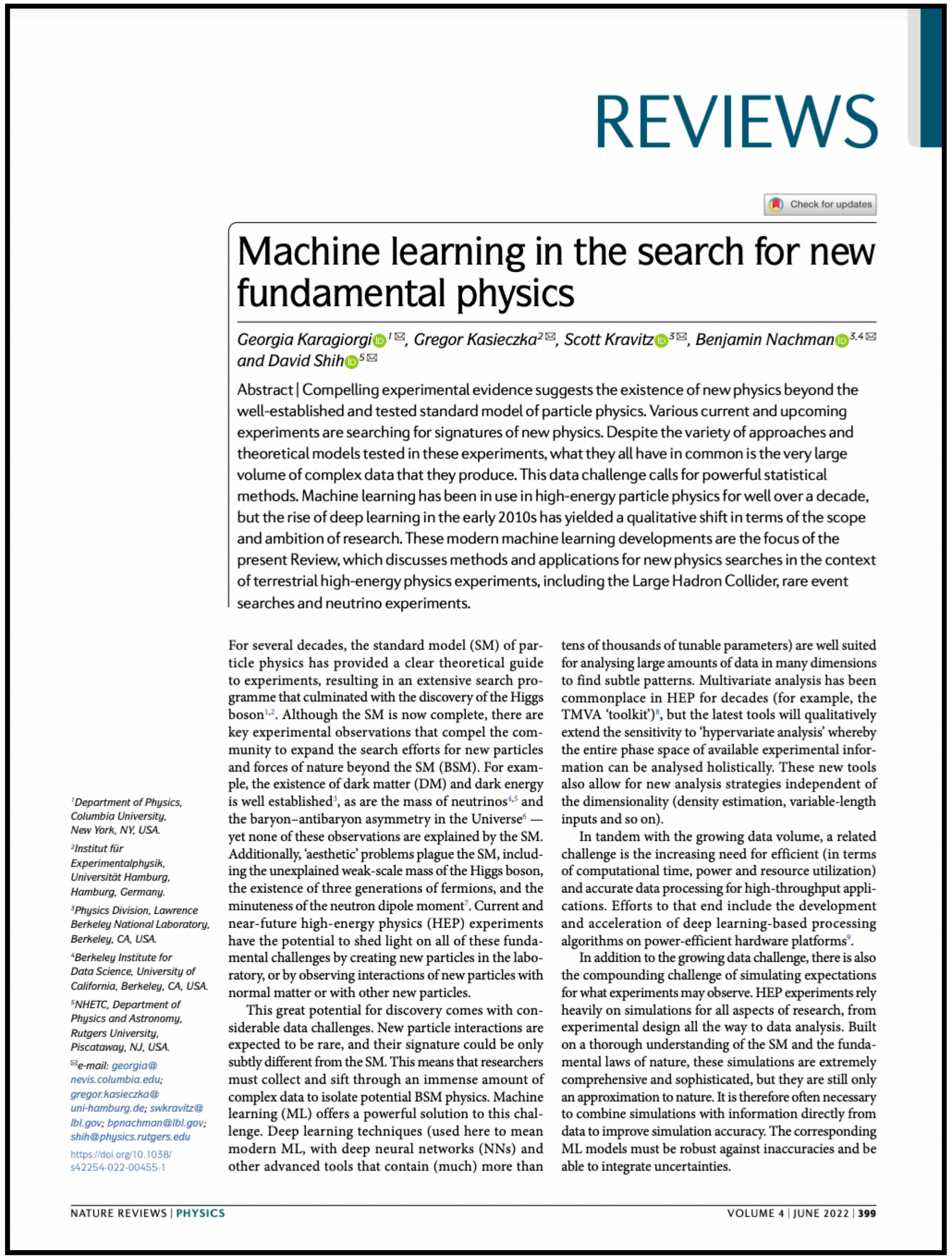
- Theoretical and experimental reasons to expect new physics beyond the Standard Model

- However, so far only negative results in searches

- Make sure that we do not miss potential discoveries at the LHC: Use machine learning to **improve existing approaches** and to **inspire new ideas**



Outline



REVIEWS

Check for updates

Machine learning in the search for new fundamental physics

Georgia Karagiorgi[1], Gregor Kasieczka[2], Scott Kravitz[3], Benjamin Nachman[3,4] and David Shih[5]

Abstract | Compelling experimental evidence suggests the existence of new physics beyond the well-established and tested standard model of particle physics. Various current and upcoming experiments are searching for signatures of new physics. Despite the variety of approaches and theoretical models tested in these experiments, what they all have in common is the very large volume of complex data that they produce. This data challenge calls for powerful statistical methods. Machine learning has been in use in high-energy particle physics for well over a decade, but the rise of deep learning in the early 2010s has yielded a qualitative shift in terms of the scope and ambition of research. These modern machine learning developments are the focus of the present Review, which discusses methods and applications for new physics searches in the context of terrestrial high-energy physics experiments, including the Large Hadron Collider, rare event searches and neutrino experiments.

For several decades, the standard model (SM) of particle physics has provided a clear theoretical guide to experiments, resulting in an extensive search programme that culminated with the discovery of the Higgs boson[1,2]. Although the SM is now complete, there are key experimental observations that compel the community to expand the search efforts for new particles and forces of nature beyond the SM (BSM). For example, the existence of dark matter (DM) and dark energy is well established[3], as are the mass of neutrinos[4,5] and the baryon–antibaryon asymmetry in the Universe[6] — yet none of these observations are explained by the SM. Additionally, 'aesthetic' problems plague the SM, including the unexplained weak-scale mass of the Higgs boson, the existence of three generations of fermions, and the minuteness of the neutron dipole moment[7]. Current and near-future high-energy physics (HEP) experiments have the potential to shed light on all of these fundamental challenges by creating new particles in the laboratory, or by observing interactions of new particles with normal matter or with other new particles.

This great potential for discovery comes with considerable data challenges. New particle interactions are expected to be rare, and their signature could be only subtly different from the SM. This means that researchers must collect and sift through an immense amount of complex data to isolate potential BSM physics. Machine learning (ML) offers a powerful solution to this challenge. Deep learning techniques (used here to mean modern ML, with deep neural networks (NNs) and other advanced tools that contain (much) more than tens of thousands of tunable parameters) are well suited for analysing large amounts of data in many dimensions to find subtle patterns. Multivariate analysis has been commonplace in HEP for decades (for example, the TMVA 'toolkit')[8], but the latest tools will qualitatively extend the sensitivity to 'hypervariate analysis' whereby the entire phase space of available experimental information can be analysed holistically. These new tools also allow for new analysis strategies independent of the dimensionality (density estimation, variable-length inputs and so on).

In tandem with the growing data volume, a related challenge is the increasing need for efficient (in terms of computational time, power and resource utilization) and accurate data processing for high-throughput applications. Efforts to that end include the development and acceleration of deep learning-based processing algorithms on power-efficient hardware platforms[9].

In addition to the growing data challenge, there is also the compounding challenge of simulating expectations for what experiments may observe. HEP experiments rely heavily on simulations for all aspects of research, from experimental design all the way to data analysis. Built on a thorough understanding of the SM and the fundamental laws of nature, these simulations are extremely comprehensive and sophisticated, but they are still only an approximation to nature. It is therefore often necessary to combine simulations with information directly from data to improve simulation accuracy. The corresponding ML models must be robust against inaccuracies and be able to integrate uncertainties.

[1]Department of Physics, Columbia University, New York, NY, USA.
[2]Institut für Experimentalphysik, Universität Hamburg, Hamburg, Germany.
[3]Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
[4]Berkeley Institute for Data Science, University of California, Berkeley, CA, USA.
[5]NHETC, Department of Physics and Astronomy, Rutgers University, Piscataway, NJ, USA.
e-mail: georgia@nevis.columbia.edu; gregor.kasieczka@uni-hamburg.de; swkravitz@lbl.gov; bpnachman@lbl.gov; shih@physics.rutgers.edu
https://doi.org/10.1038/s42254-022-00455-1

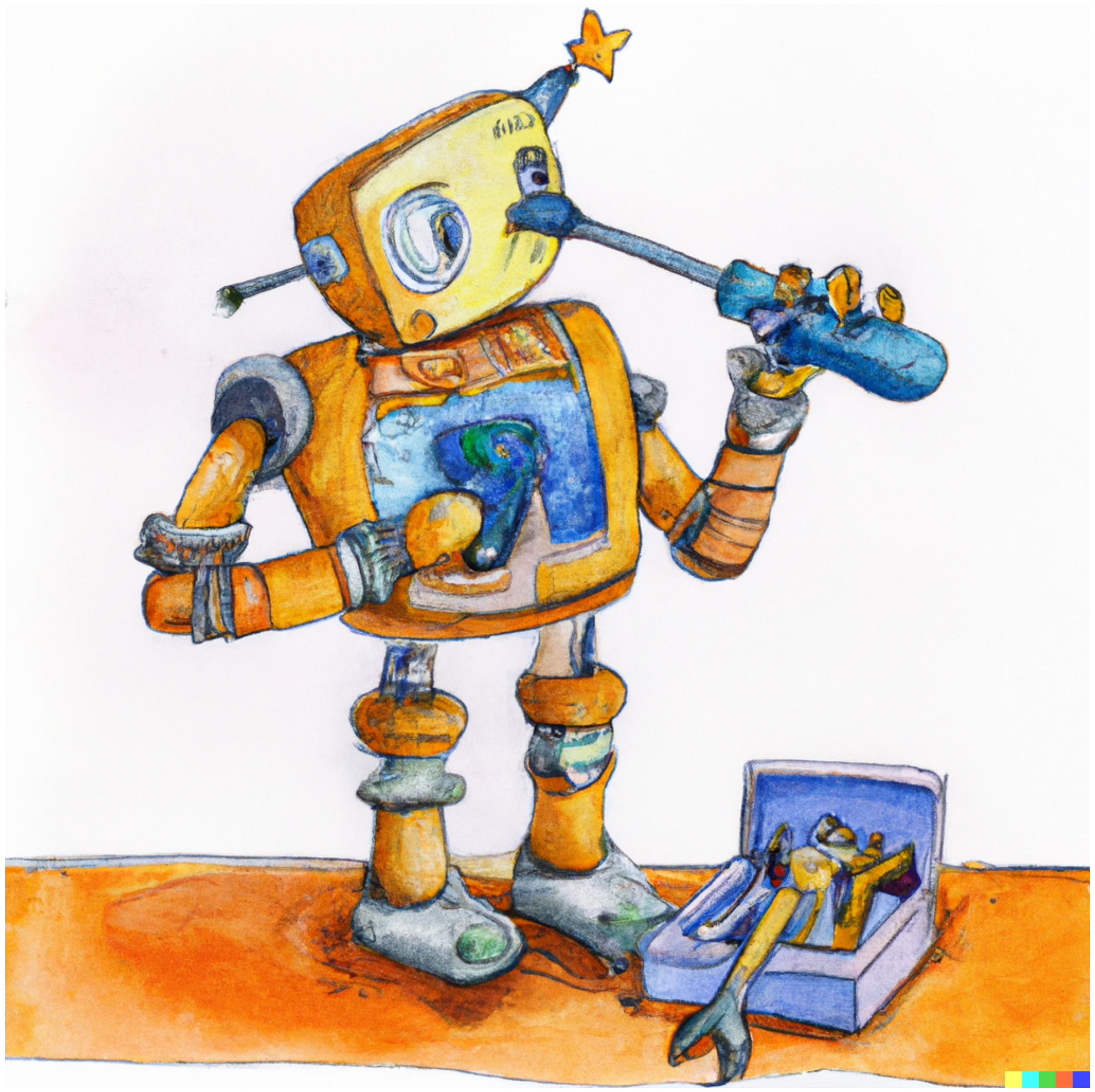
NATURE REVIEWS | PHYSICS VOLUME 4 | JUNE 2022 | 399

Despite the same title, this talk will be more focused on LHC physics and on recent work than our review (2112.03769)

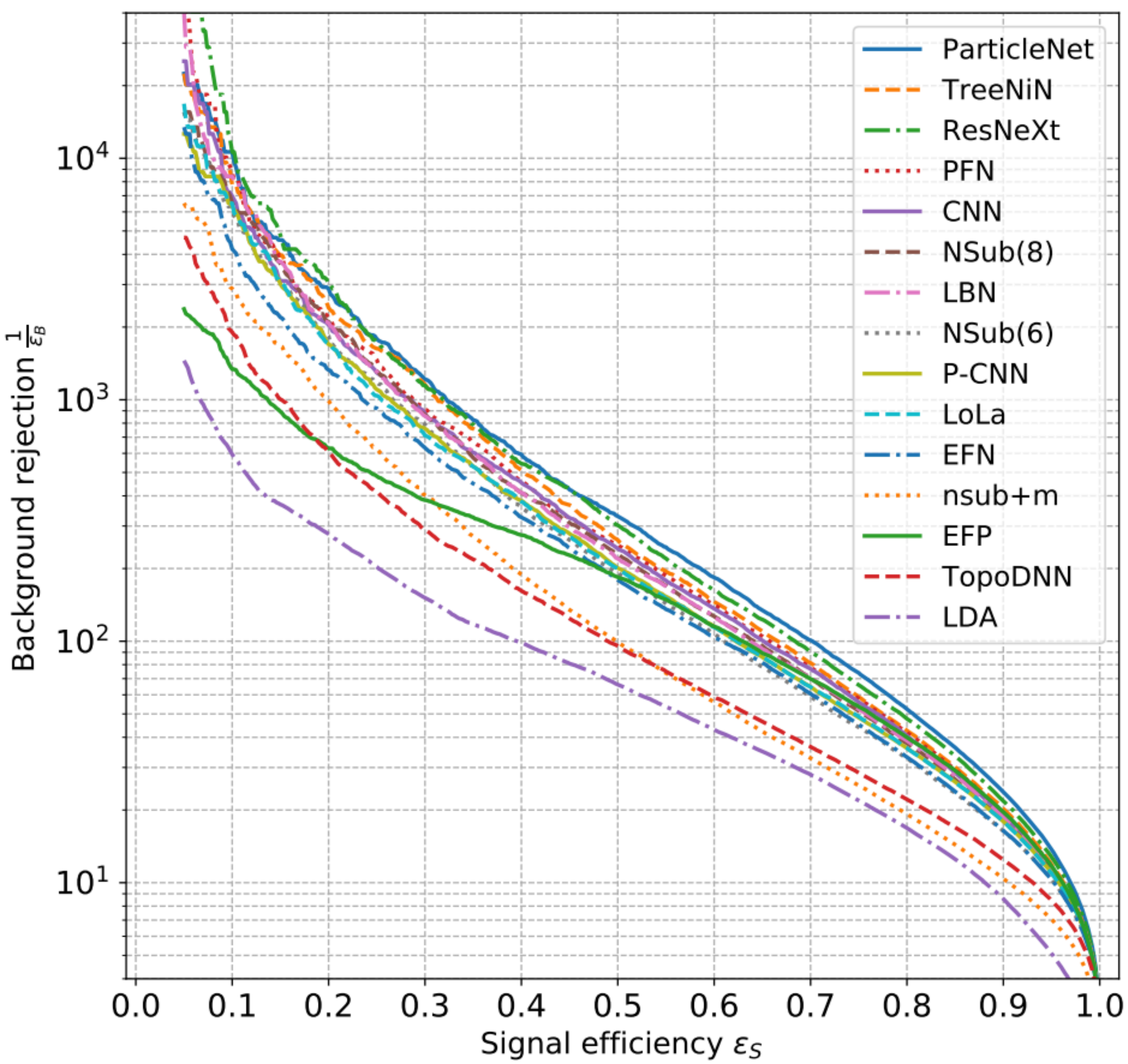
- Improve existing approaches:
  - Better **performance** for new physics taggers
  - Increase **stability**
- New ideas:
  - Build **model independent** searches

See Danilo's excellent talk from Monday for a big picture view of computing challenges in the future

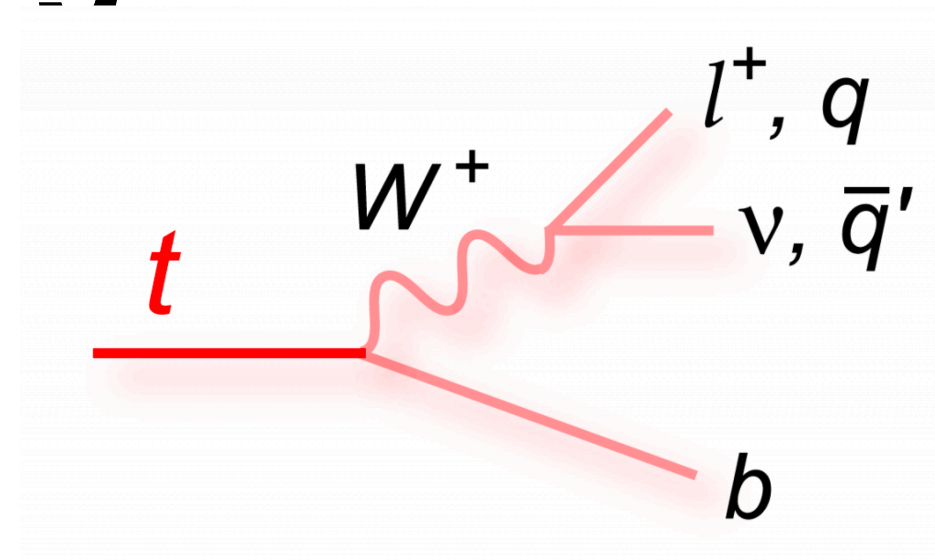
Improving Performance



Enter: Deep Learning



Improvement of factor 2-3 over shallow ML (and more over non-ML methods) in benchmark classification task



- Example application: **hadronic top quark decays**
- 1.2M simulated top quark and background events
- Either four-momenta of individual particles or high-level features
- Great test-bed to compare different data representations
    - (and, of course, useful for new physics searches)

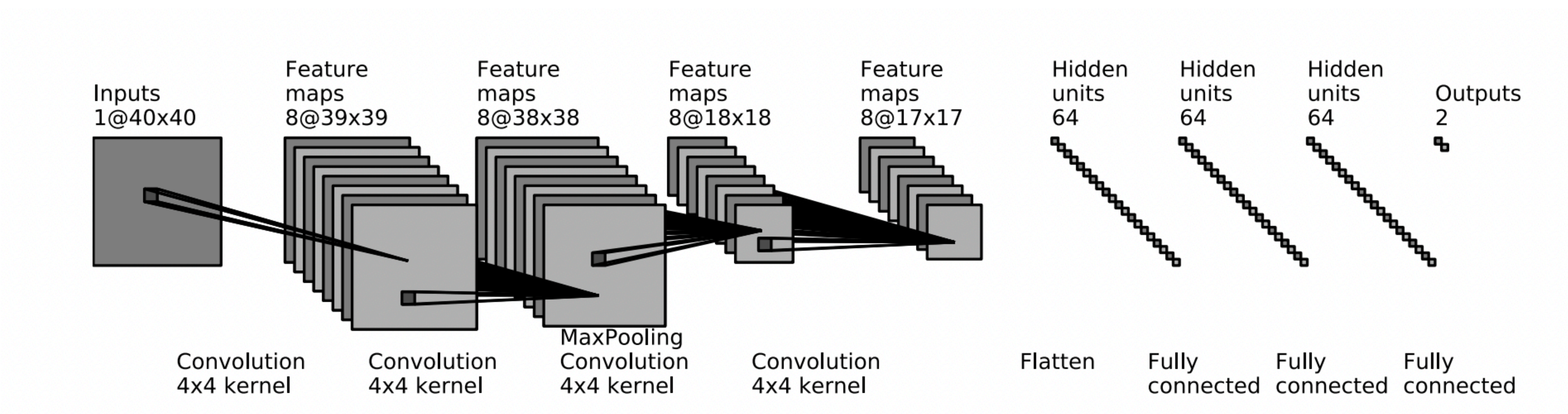
GK, Plehn (eds), et al, 1902.09914

Architectures

- Basic motivation: Use physicists' knowledge about data as an implicit (or explicit bias) to help networks train faster / achieve better performance

Architectures

- Basic motivation: Use physicists' knowledge about data as an implicit (or explicit bias) to help networks train faster / achieve better performance
- Either by phrasing physics problems so that outside-solutions can be used …

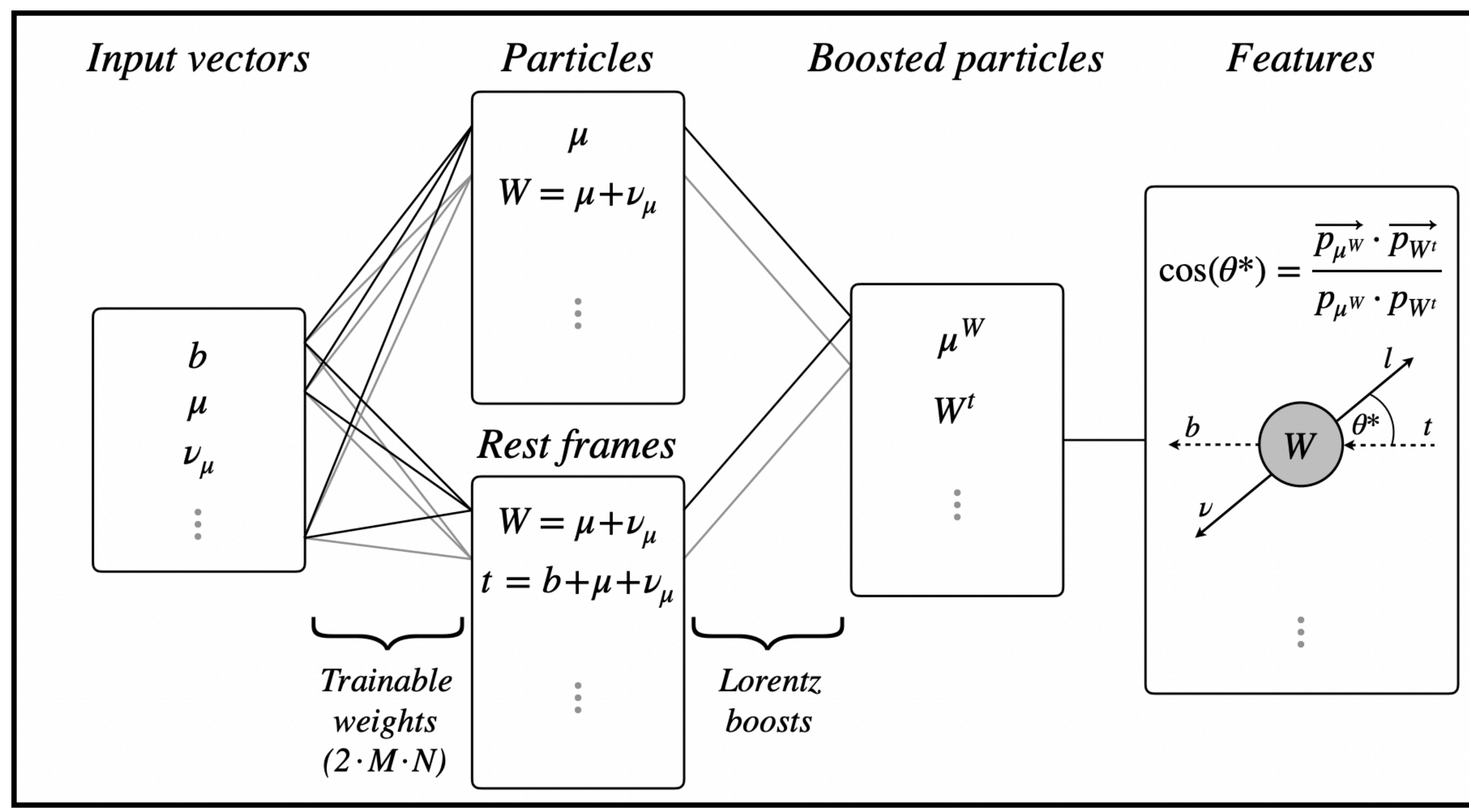


Very simple convolutional architecture, using locality and translation invariance.

GK, Plehn, et al, 1701.08784; Macaluso, Shih, *1803.00107; …*

Architectures

- Basic motivation: Use physicists' knowledge about data as an implicit (or explicit bias) to help networks train faster / achieve better performance
- Either by phrasing physics problems so that outside-solutions can be used …
- …or by constructing networks layers based on physical symmetries

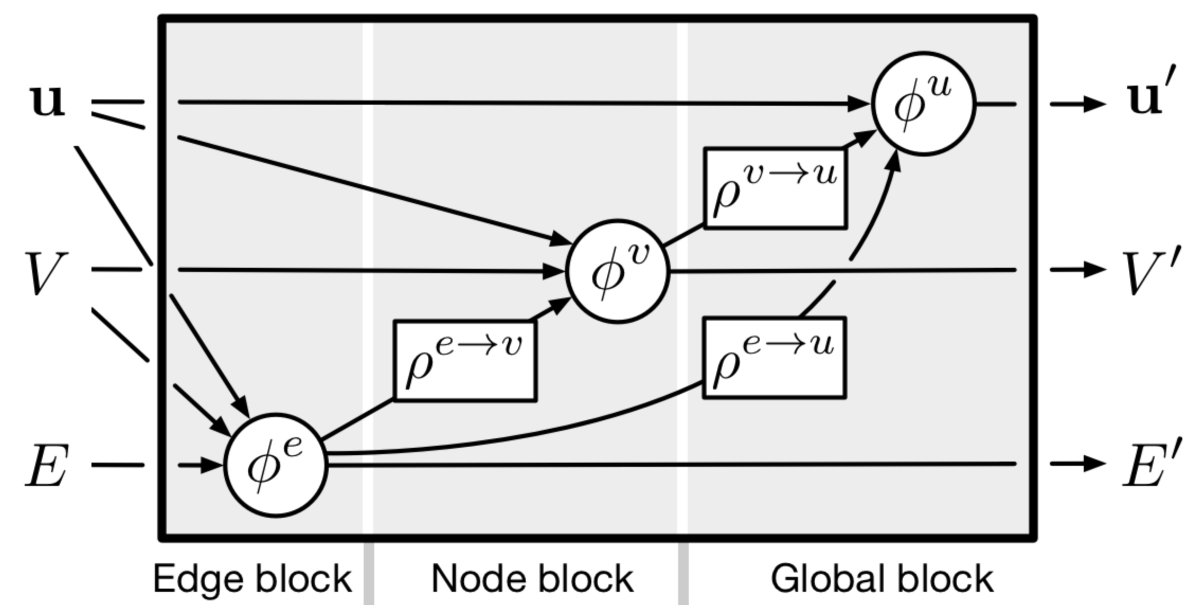


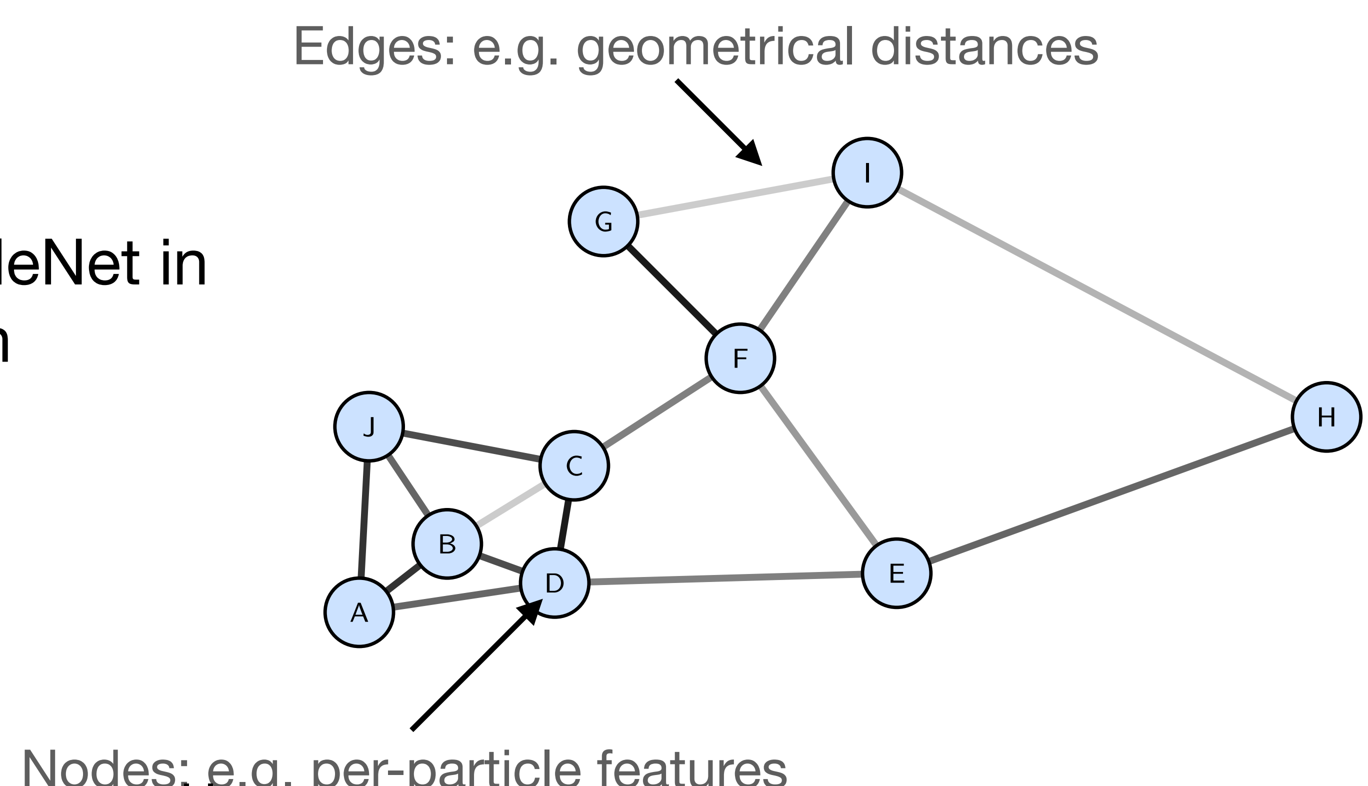
Learn combinations of particles and suitable rest frames

GK, Plehn, et al, 1707.08966; Erdmann et al 1812.09722; Bogatskiy et al 2006.04780; …

Architectures

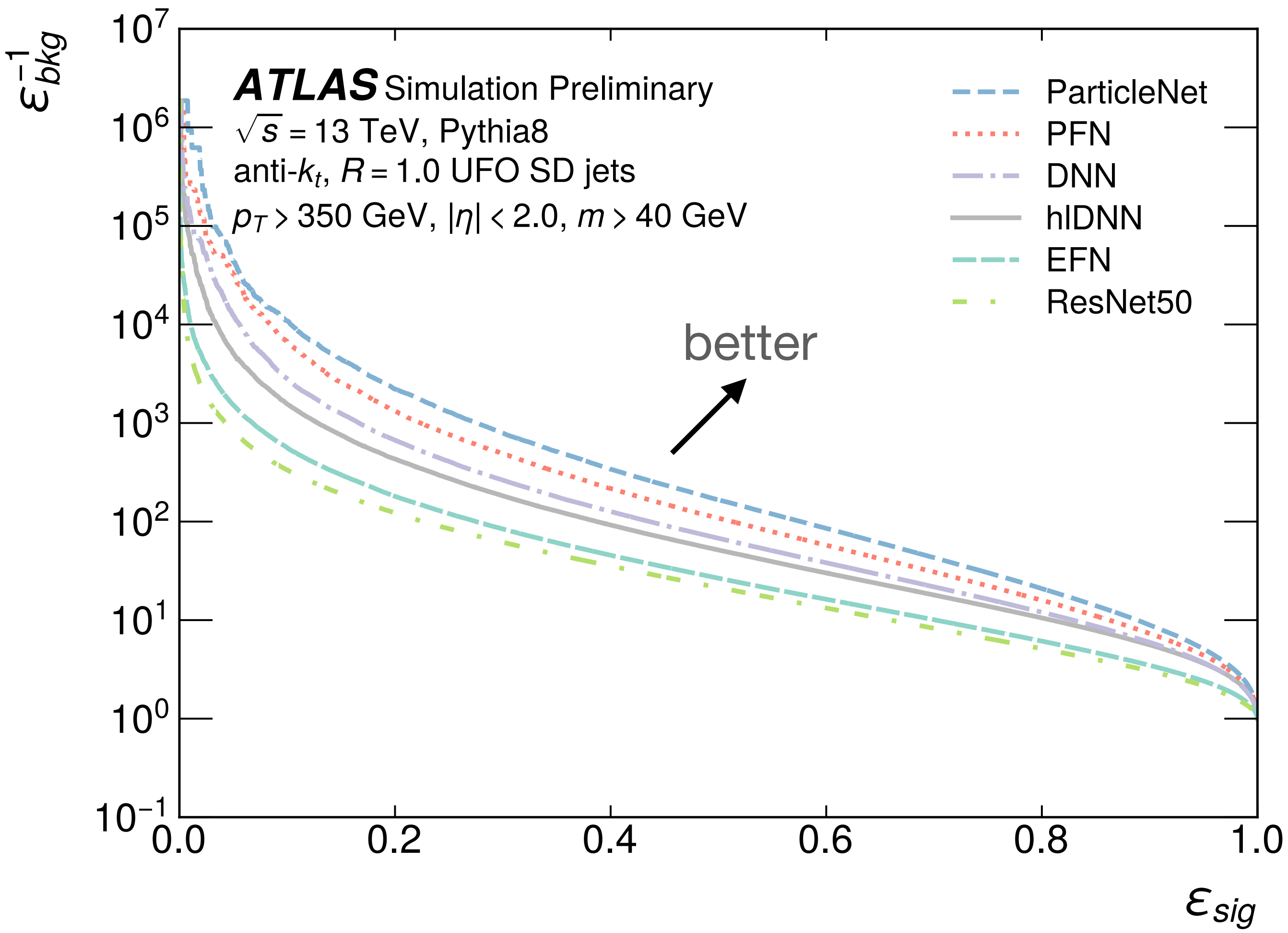
- Basic motivation: Use physicists' knowledge about data as an implicit (or explicit bias) to help networks train faster / achieve better performance
- Either by phrasing physics problems so that outside-solutions can be used …
- …or by constructing networks layers based on physical symmetries
- **Graphs** are a general + powerful framework that captures relevant properties for particle tagging
  - e.g. best performance of ParticleNet in original top tagging comparison
  - versatile and well suited

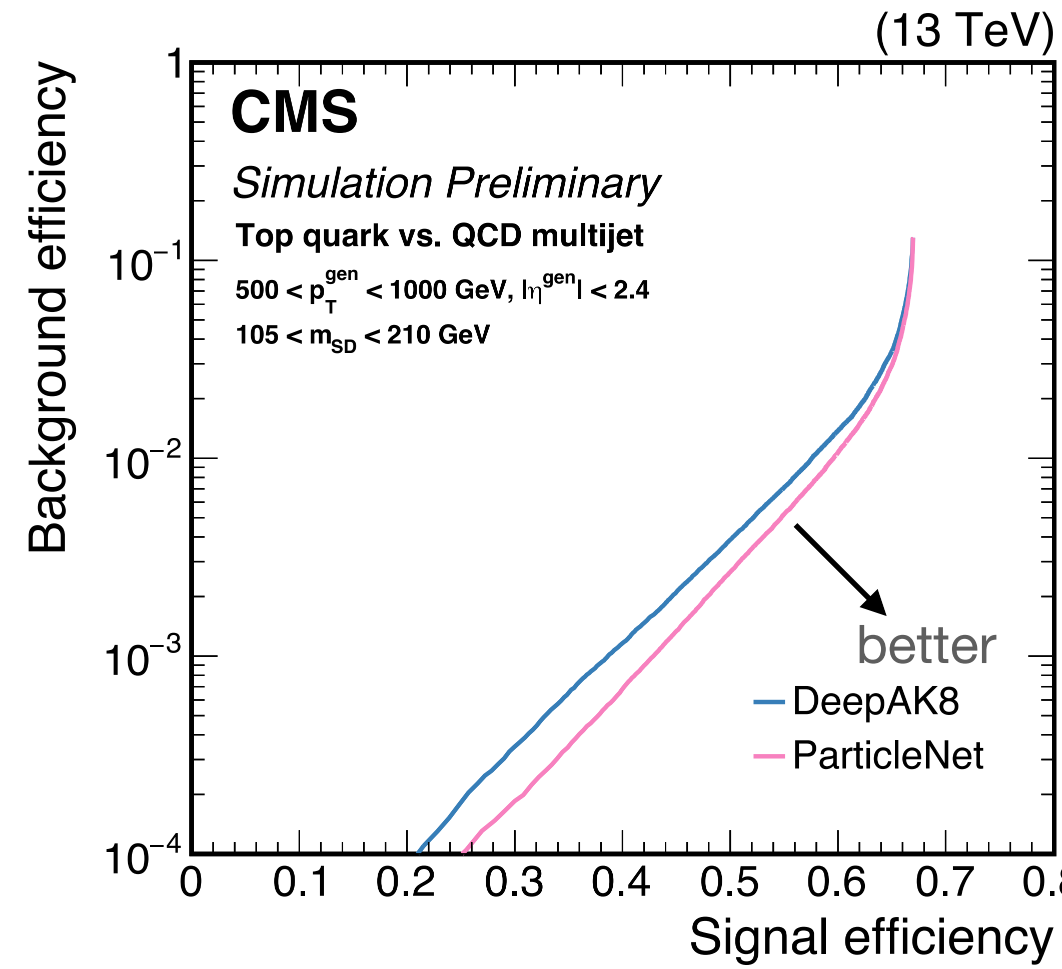




Henrion et al ML4PS 2017; Qu, Gouskos 1902.08570; Shalom, Battaglia, Valiant 2007.13681 (review)

In practice



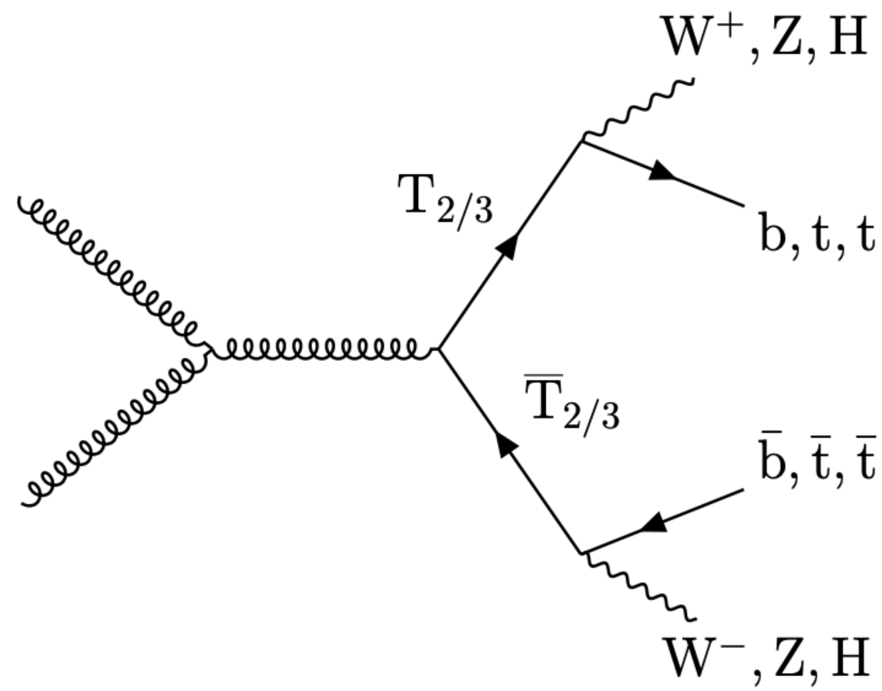


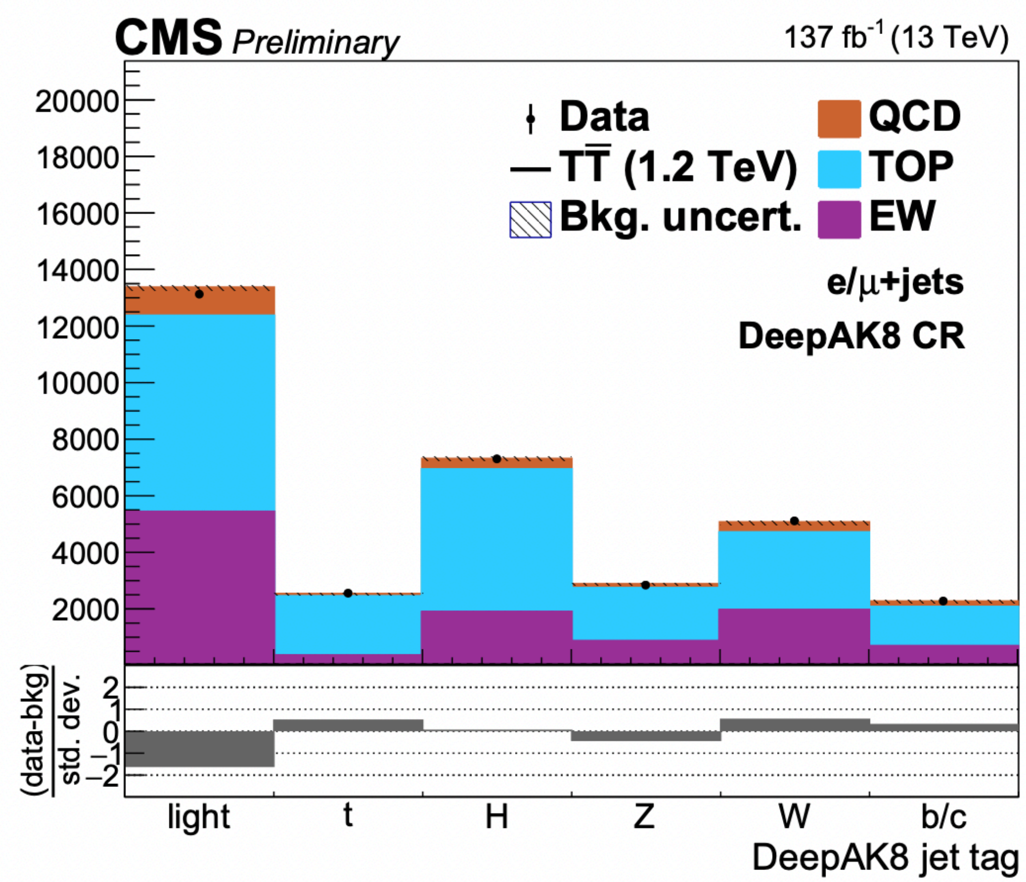
Both ATLAS and CMS confirm performance of graph-based (ParticleNet) approach on realistic simulations

CMS 2004.08262; CMS DP -2020/002; ATL-PHYS-PUB-2022-039

Use of object tagging

- Top tagging (+other heavy resonances, flavour, tau,…): standard model particles
    - Still important for BSM searches

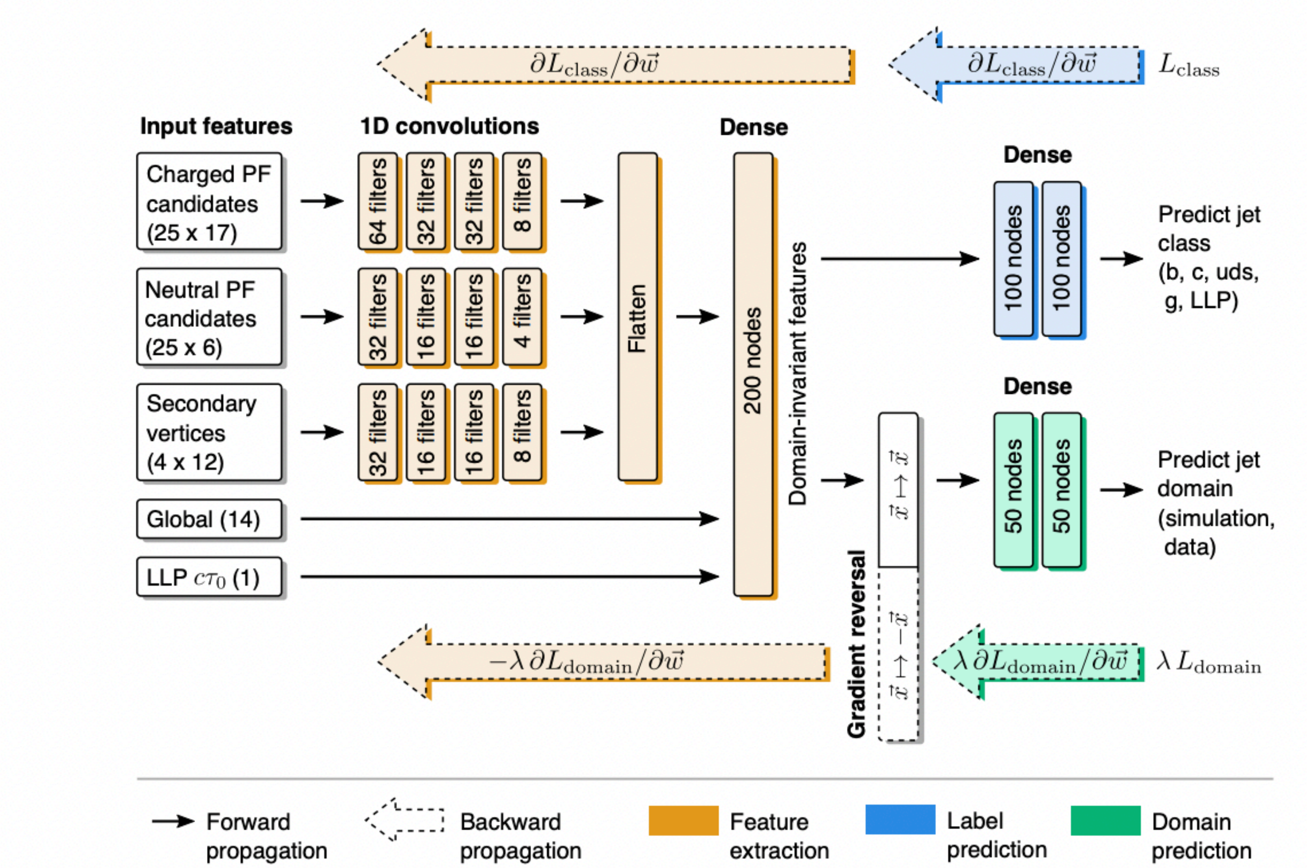




CMS B2G-20-011 (just an example; majority of searches uses flavour/resonance tagging as ingredient)

Tagging other particles

- Top tagging (+other heavy resonances, flavour, tau,…): standard model particles
    - Still important for BSM searches
    - Relatively easy calibration (signal & background samples in data exist)
      For new physics: background calibration possible, larger uncertainty on signal

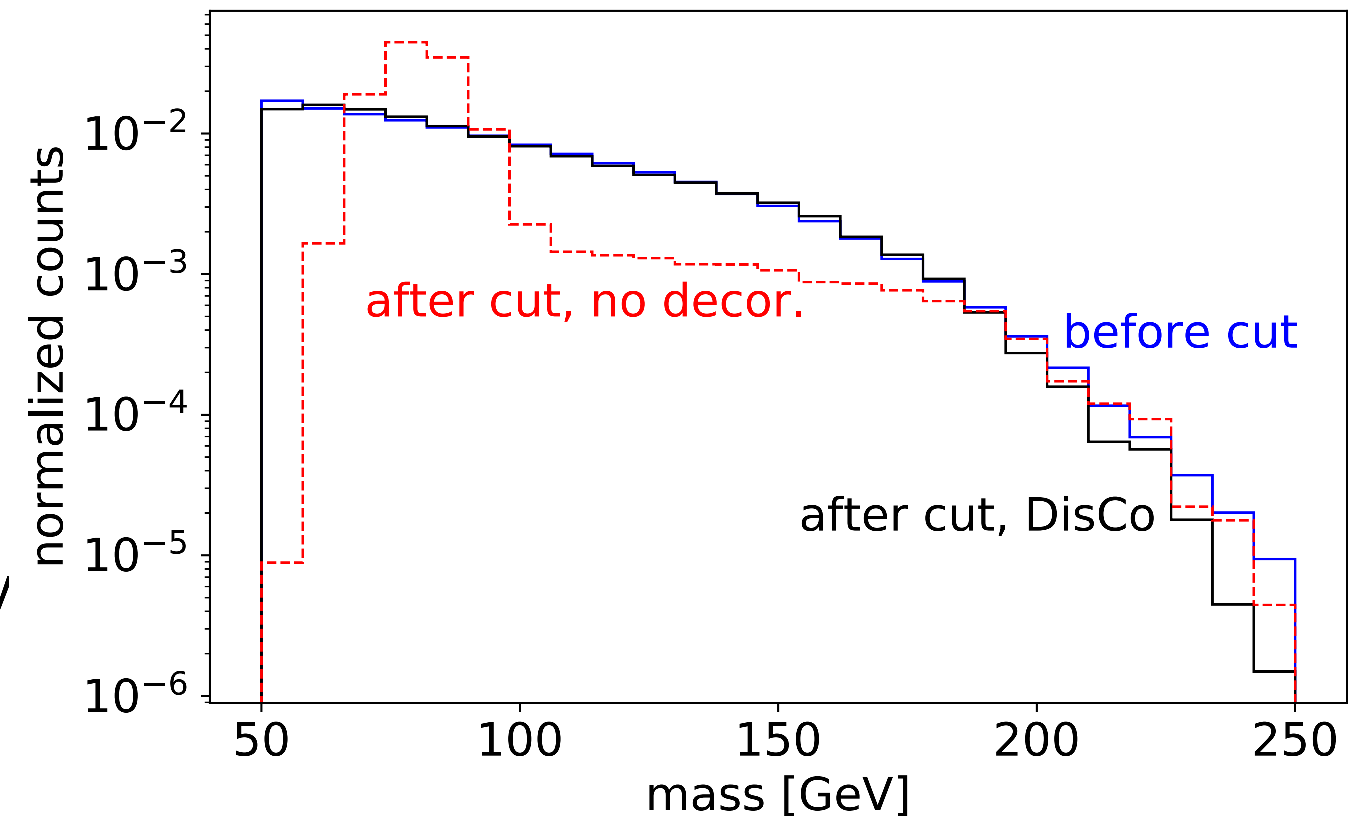


Example for domain adaptation

CMS 1912.12238; …

Tagging other particles

- Top tagging (+other heavy resonances, flavour, tau,…): standard model particles
  - Still important for BSM searches
  - Relatively easy calibration (signal & background samples in data exist)
    For new physics: background calibration possible, larger uncertainty on signal
  - Can assume all properties (mass,..)
    For new physics: Parametrised networks or decorrelation



$$L = L_{classifier}(\vec{y}, \vec{y}_{true}) + \lambda \, \mathrm{dCorr}^2_{y_{true}=0}(\vec{m}, \vec{y})$$

Add term to loss function to decrease correlation with specific observable (e.g. invariant mass)

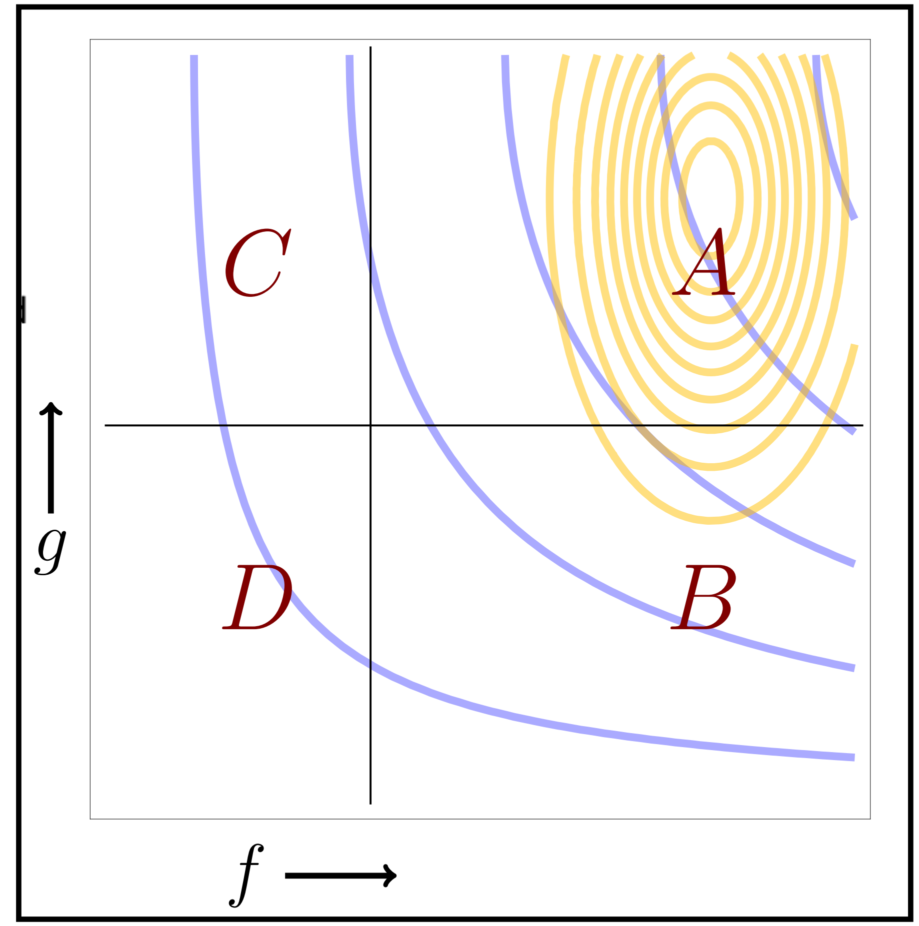
$$\begin{aligned}\mathrm{dCov}^2(X,Y) = &\langle |X - X'||Y - Y'| \rangle \\ &+ \langle |X - X'| \rangle \langle |Y - Y'| \rangle \\ &- 2\langle |X - X'||Y - Y''| \rangle\end{aligned}$$

$$\mathrm{dCorr}^2(X,Y) = \frac{\mathrm{dCov}^2(X,Y)}{\mathrm{dCov}(X,X)\mathrm{dCov}(Y,Y)}$$

GK, Shih 2001.05310; Louppe, Kagan, Cranmer 1611.01046; Shimmin et al 1703.03507; Bradshaw et al 1908.08959; Dolen et al 1603.00027; Kitouni et al 2010.09745, …

Estimating Backgrounds

- Can take decorrelation further.
- For new physics searches, need to
  - Find two variables that:
  - Isolate a possible signal &
  - Are independent
    (and can be used for
    ABCD background estimation)
- Can phrase this directly as training task,
  again using DisCo

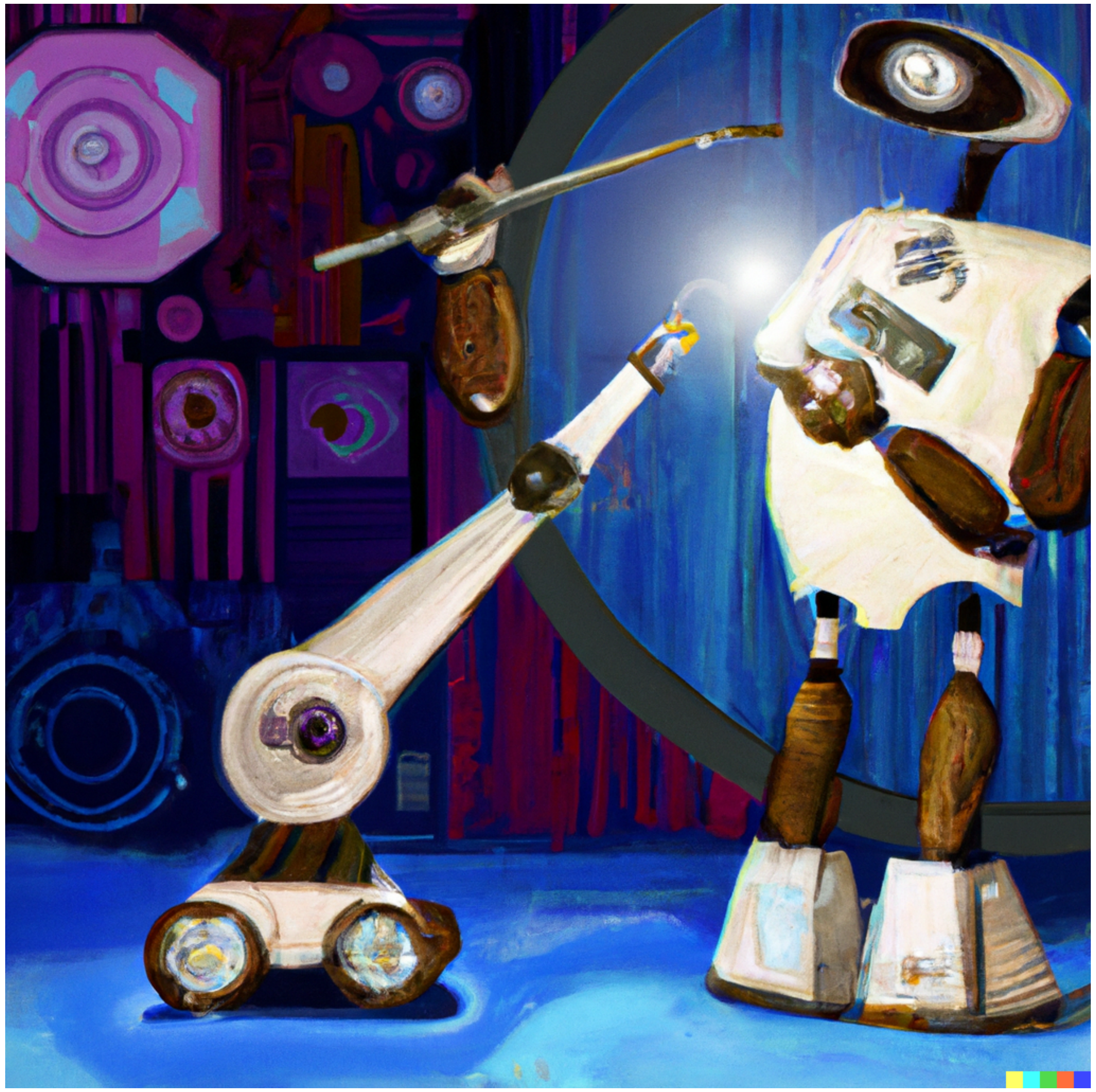


$$N_A = \frac{N_B N_C}{N_D}$$

$$\mathcal{L}[f,g] = \mathcal{L}_{\text{classifier}}[f(X),y] + \mathcal{L}_{\text{classifier}}[g(X),y] + \lambda\,\text{dCorr}^2_{y=0}[f(X),g(X)]$$

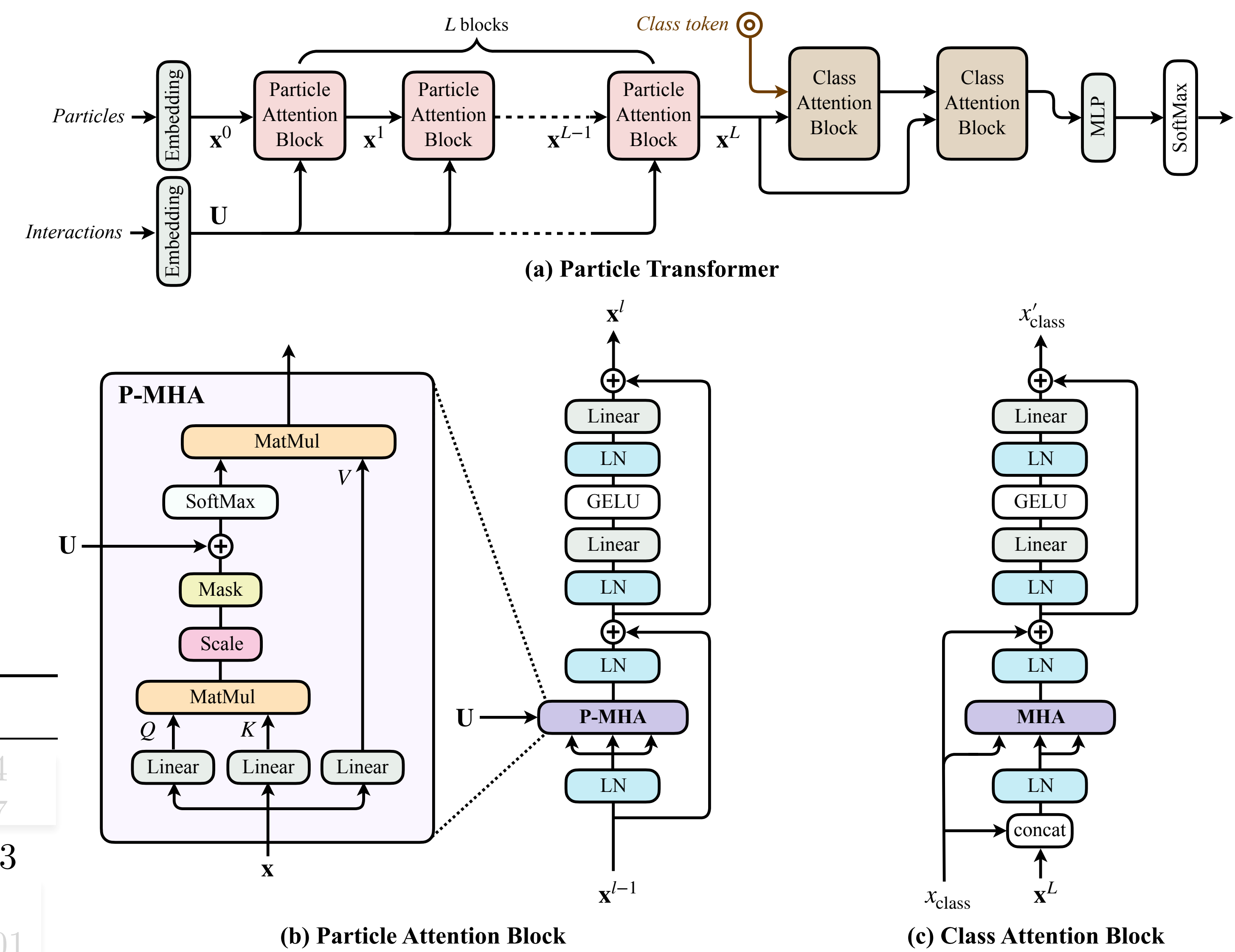
GK et al 2007.14400; See also Mikuni, Nachman, Shih for decorrelating autoencoders: 2111.06417

Back to Performance



Attention is all you need

- In ParticleNet, data-space geometry defines neighbourhood in graph; aggregation over all neighbours
- Attention allows the network to learn which parts of the input are truly relevant
- Attention is data-hungry, transfer-learning helps!



ParT architecture diagram

	Accuracy	AUC	$\mathrm{Rej}_{50\%}$	$\mathrm{Rej}_{30\%}$
P-CNN	0.930	0.9803	201 ± 4	759 ± 24
PFN	—	0.9819	247 ± 3	888 ± 17
ParticleNet	0.940	0.9858	397 ± 7	1615 ± 93
JEDI-net (w/ $\sum O$)	0.930	0.9807	—	774.6
PCT	0.940	0.9855	392 ± 7	1533 ± 101
LGN	0.929	0.964	—	435 ± 95
rPCN	—	0.9845	364 ± 9	1642 ± 93
LorentzNet	0.942	0.9868	498 ± 18	2195 ± 173
ParT	0.940	0.9858	413 ± 16	1602 ± 81
ParticleNet-f.t.	0.942	0.9866	487 ± 9	1771 ± 80
ParT-f.t.	**0.944**	**0.9877**	$\mathbf{691 \pm 15}$	$\mathbf{2766 \pm 130}$

Performance comparison on landscape dataset

Vaswani et al 1706.03762; Qu, Li, Qian 2202.03772; Mikuni, Canelli 2001.05311; ..

Attention is all you need

- In ParticleNet, data-space geometry defines neighbourhood in graph; aggregation over all neighbours
- Attention allows the network to learn which parts of the input are truly relevant
- Attention is data-hungry, transfer-learning helps! (Motivation for foundation models?)
- So far, observed trend: Higher physics performance comes at the cost of higher algorithm complexity & compute cost

	Accuracy	# params	FLOPs
PFN	0.772	86.1 k	4.62 M
P-CNN	0.809	354 k	15.5 M
ParticleNet	0.844	370 k	540 M
ParT	**0.861**	2.14 M	340 M
ParT (plain)	0.849	2.13 M	260 M

(plain: standard multi-head-attention vs particle-multi-head-attention)

Vaswani et al 1706.03762; Qu, Li, Qian 2202.03772; Mikuni, Canelli 2001.05311; Gong et al 2201.08187 for a combination of transformers and attention;

Attention is all you need

- In ParticleNet, data-space geometry defines neighbourhood in graph; aggregation over all neighbours
- Attention allows the network to learn which parts of the input are truly relevant
- Attention is data-hungry, transfer-learning helps! (Motivation for foundation models?)
- So far, observed trend: Higher physics performance comes at the cost of higher algorithm complexity & compute cost
- **Is this the only way?**

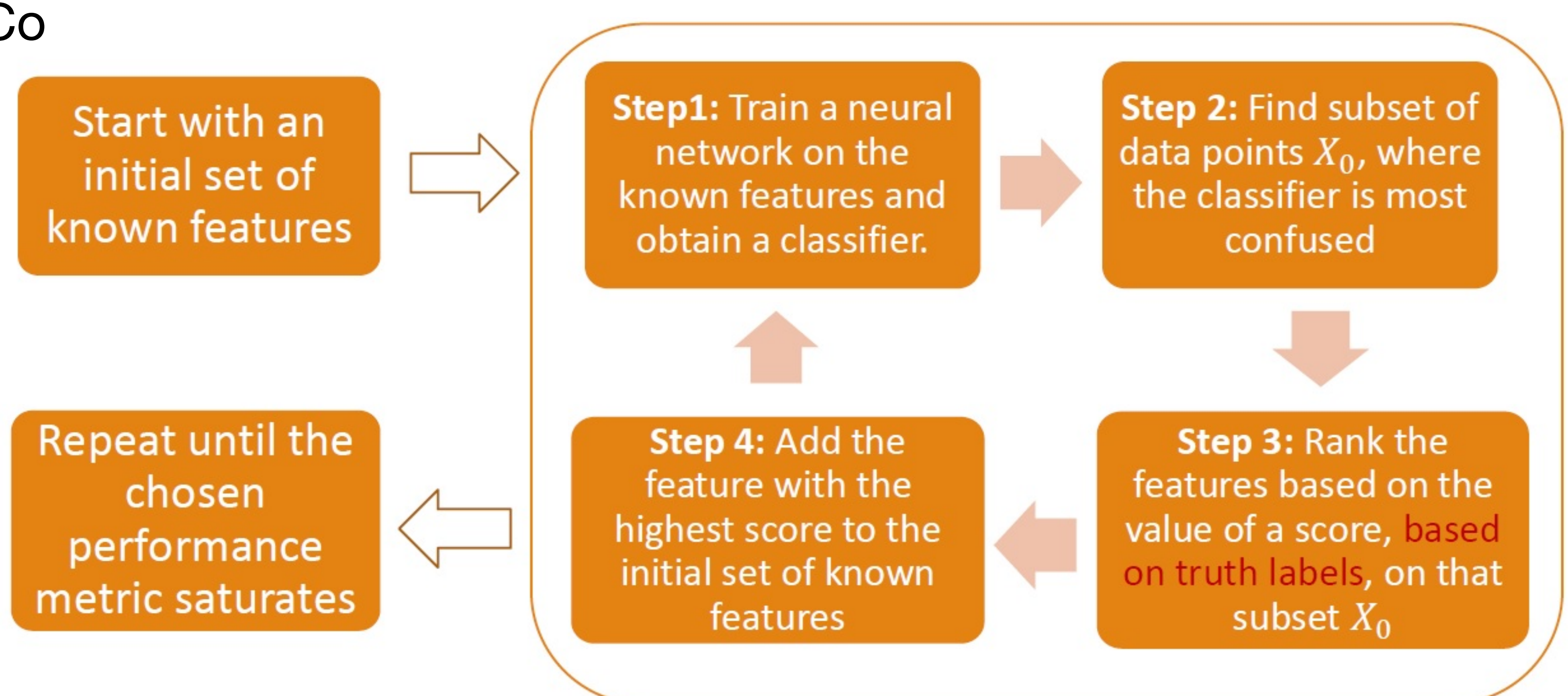
	Accuracy	# params	FLOPs
PFN	0.772	86.1 k	4.62 M
P-CNN	0.809	354 k	15.5 M
ParticleNet	0.844	370 k	540 M
ParT	**0.861**	2.14 M	340 M
ParT (plain)	0.849	2.13 M	260 M

(plain: standard multi-head-attention vs particle-multi-head-attention)

Vaswani et al 1706.03762; Qu, Li, Qian 2202.03772; Mikuni, Canelli 2001.05311; Gong et al 2201.08187 for a combination of transformers and attention

Looking for optimal feature set

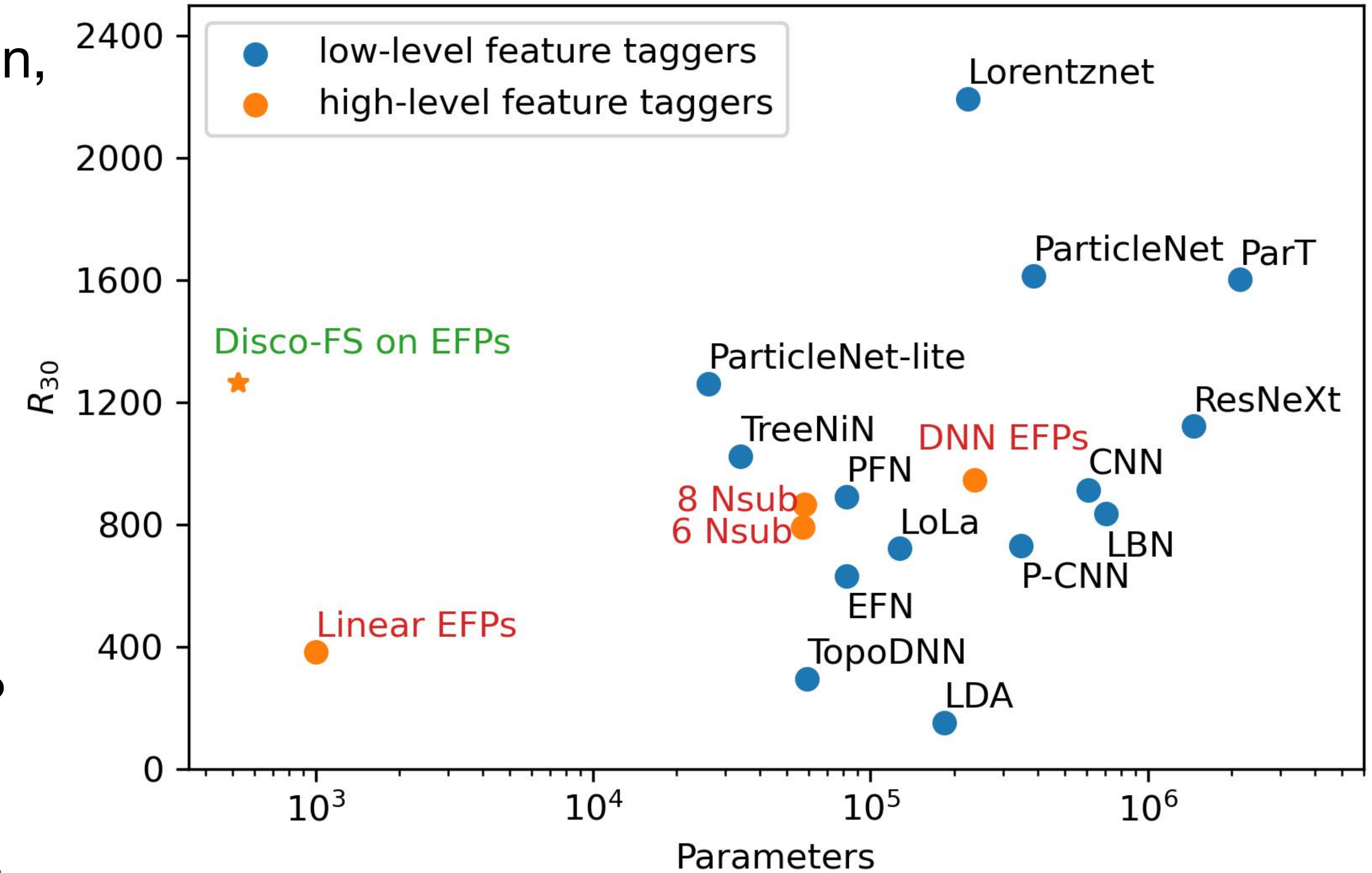
- Energy Flow Polyonomials (EFPs) form a basis of jet substructure
- Depending on order considered, too many (e.g 7k) to efficiently train NN (many features work if there is structure, not so much for EFPs)
- Solution: Iterative feature selection, again based on DisCo



Das, GK, Shih, to be published; Faucett, Thaler, Whiteson, 2010.11998

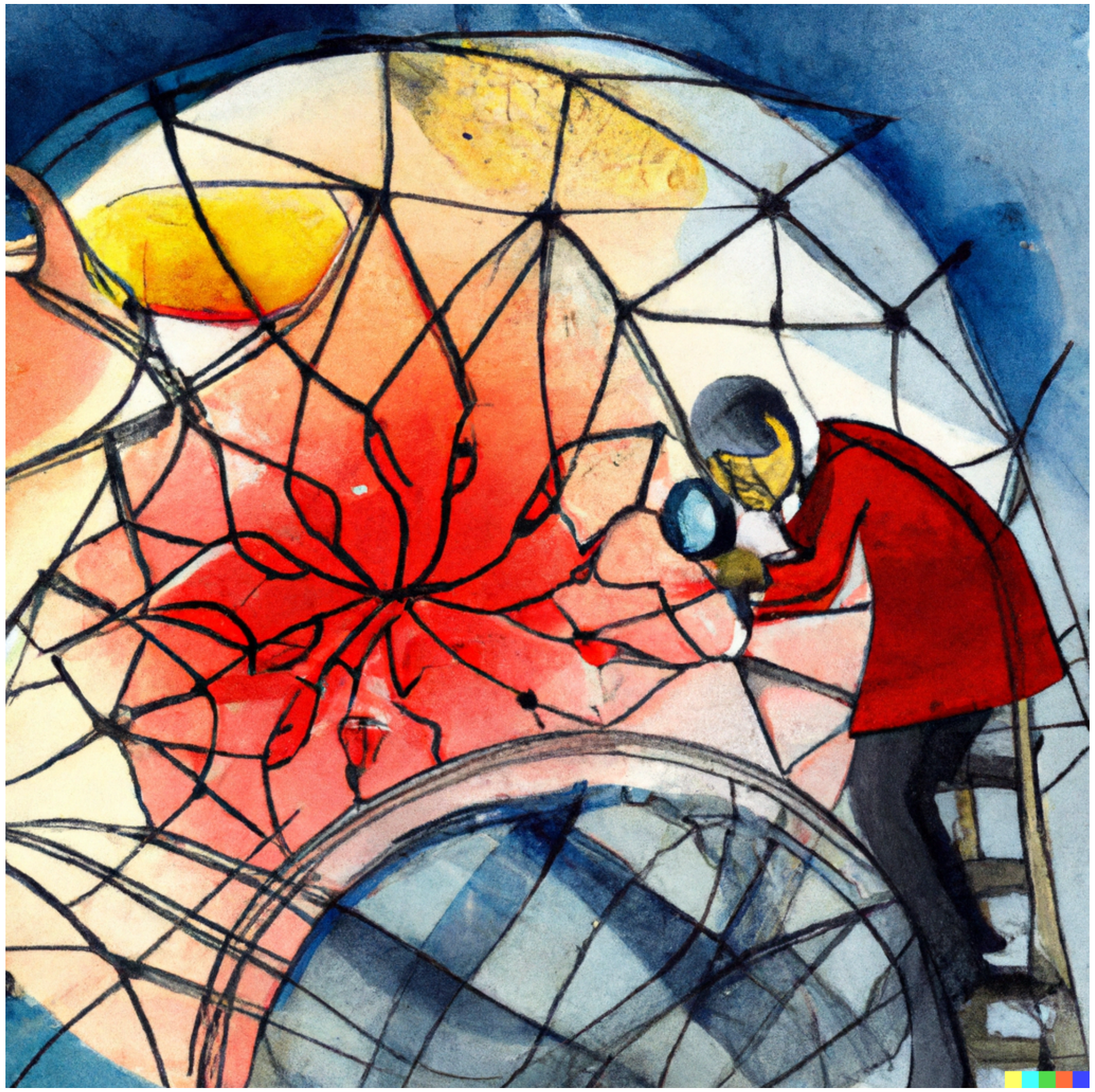
Looking for optimal feature set

- Energy Flow Polyonomials (EFPs) form a basis of jet substructure
- 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)
- Solution: Iterative feature selection, again based on DisCo
- Same top tag performance as simple graph network but only O(10) inputs; factor 50 less parameters
- Also helps interpretability, calibration
- Useful for new physics searches?



Das, GK, Shih, to be published; Faucett, Thaler, Whiteson, 2010.11998

Model Independent Searches



Anomaly Searches

- 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

Anomaly Searches

- 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

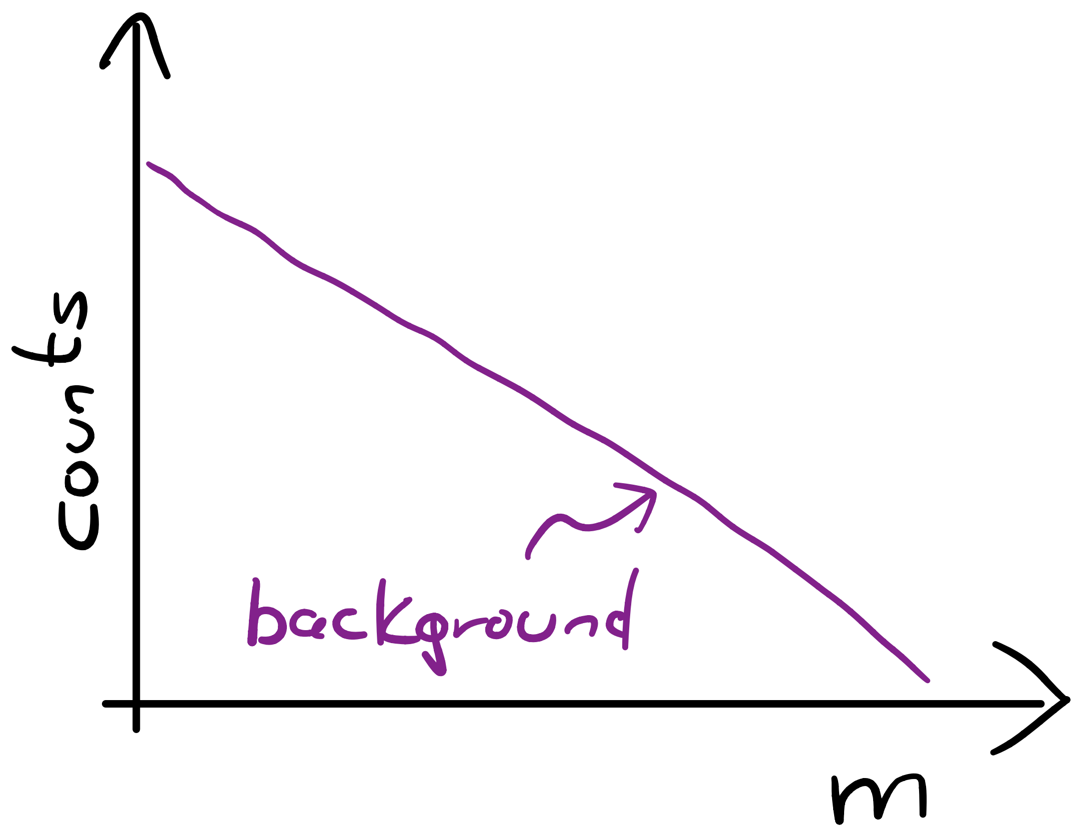
Resonant anomaly detection / Enhanced bump hunts

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

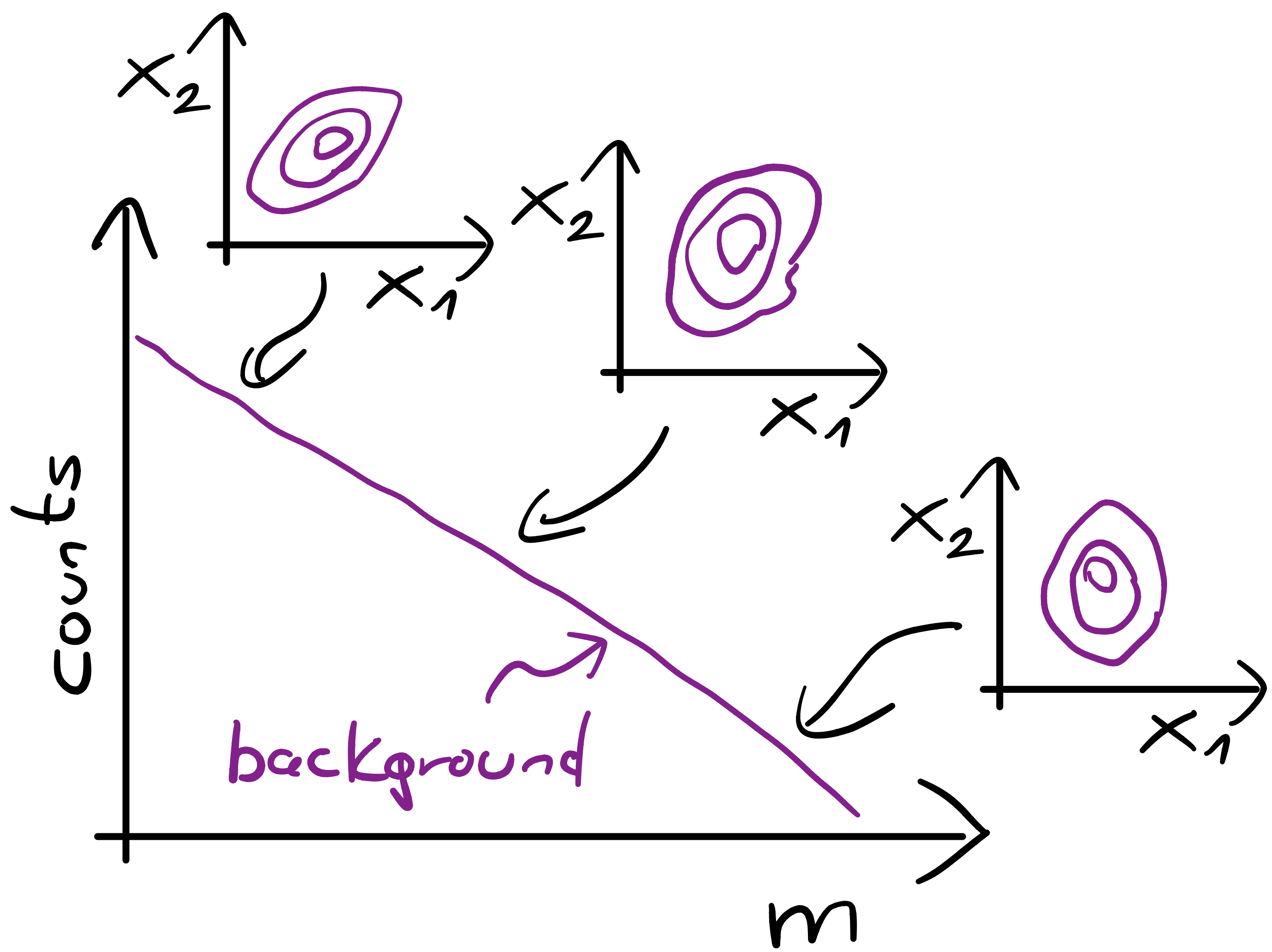
- Will focus on these for rest of the talk
- See Andrea tomorrow for an alternative view

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

Resonant Anomaly Detection

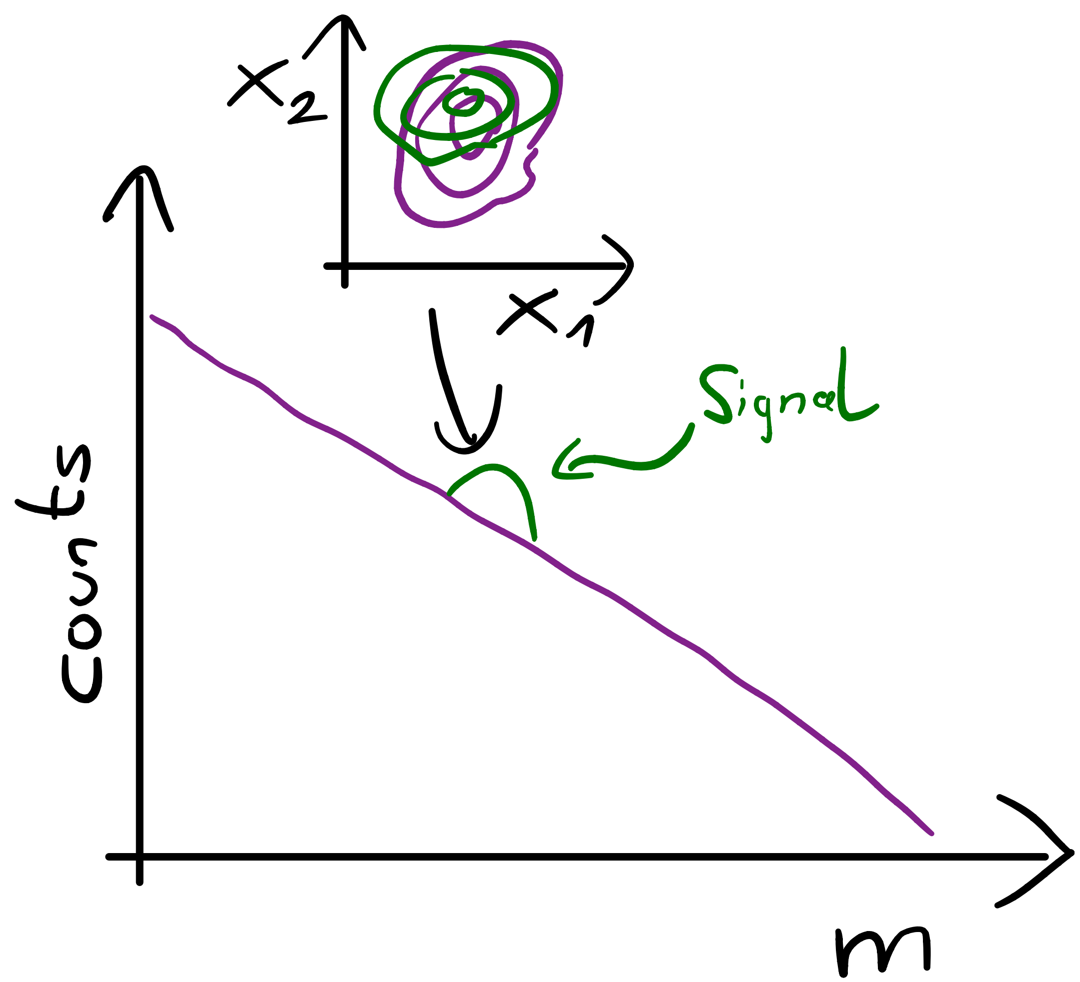


Resonant Anomaly Detection



+ additional dimensions
(in general correlated with m)

Resonant Anomaly Detection



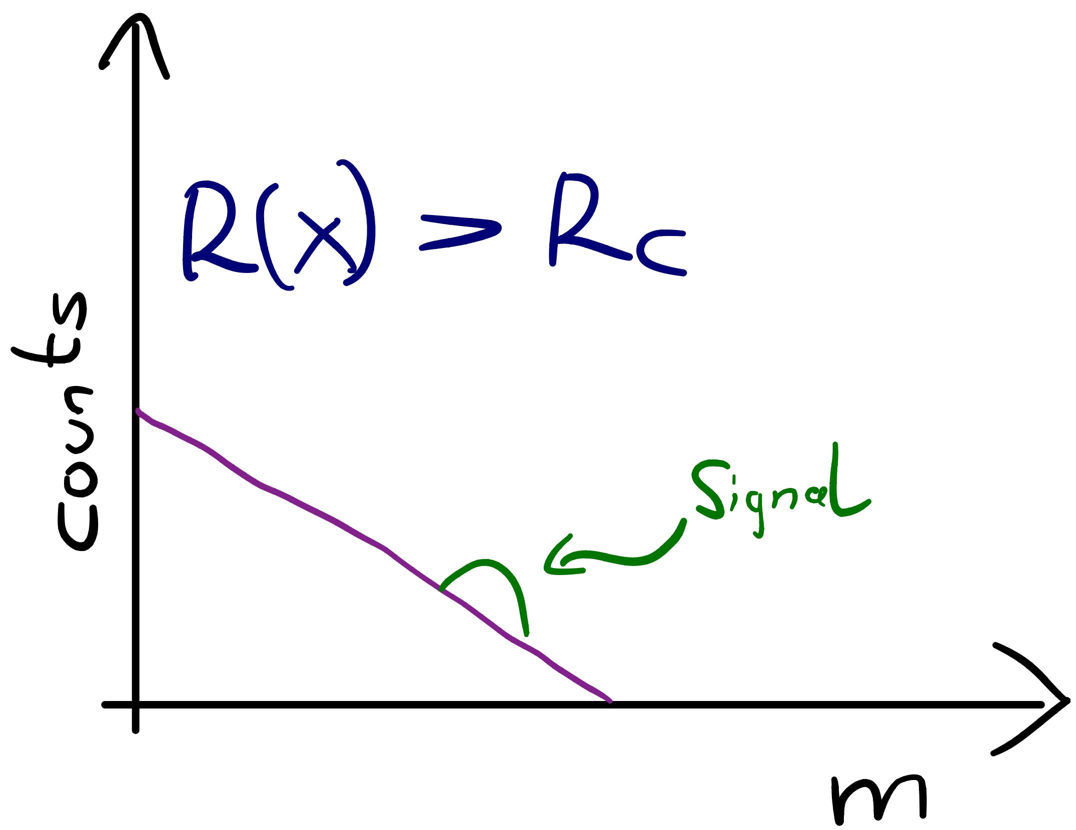
Look for a small signal, localised in m, and different shape in other features

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

No assumptions in other features.

Further generalisation as open issue.

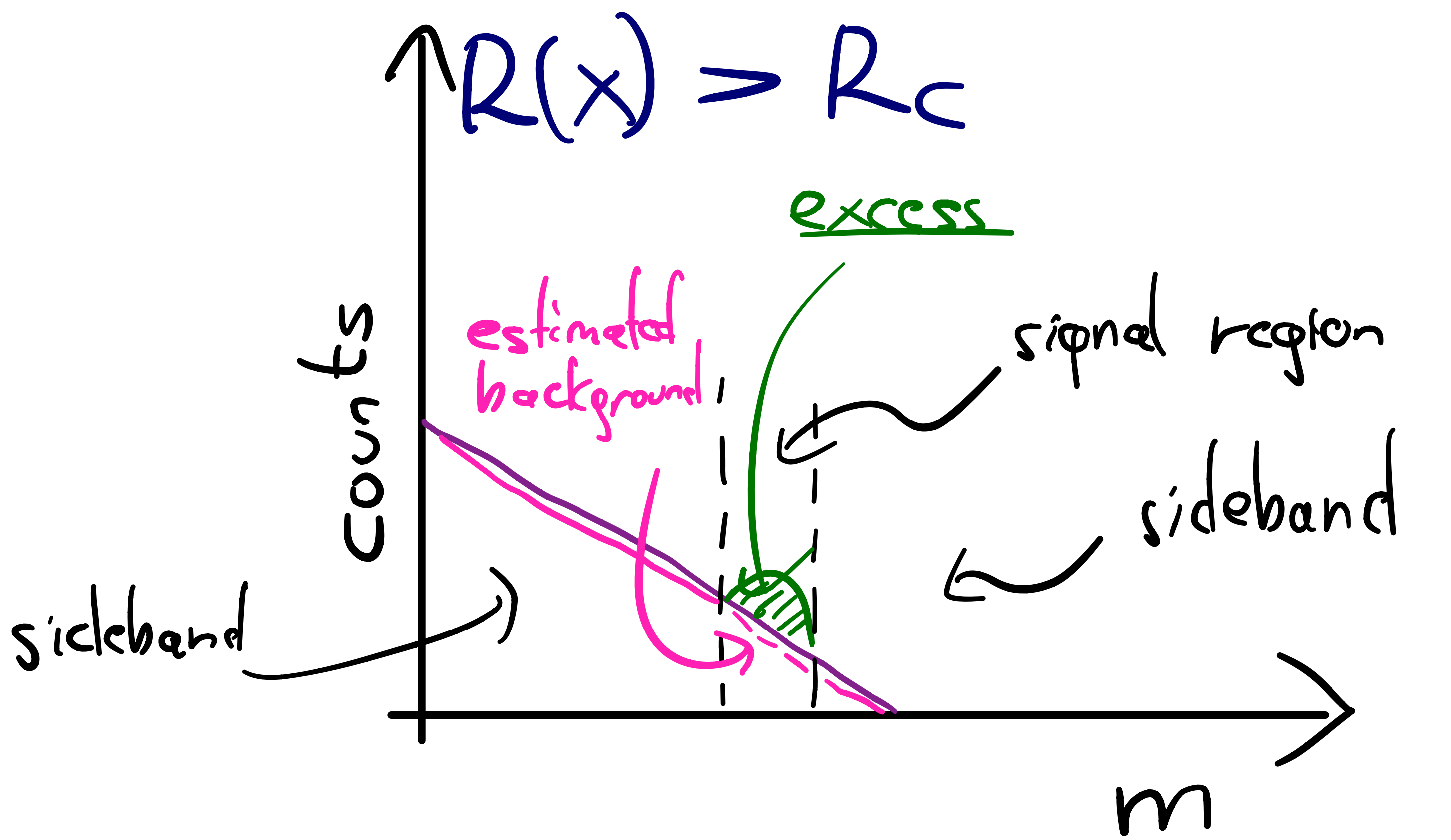
Resonant Anomaly Detection



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

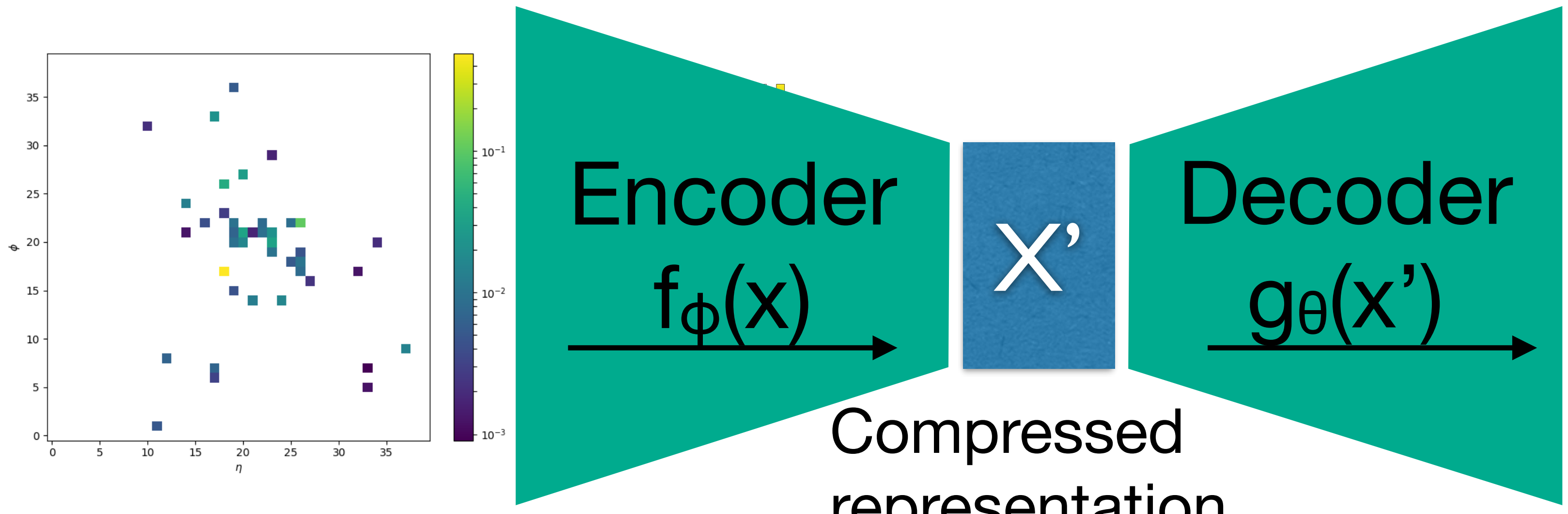
Resonant Anomaly Detection



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

…so HOW to construct the anomaly score?

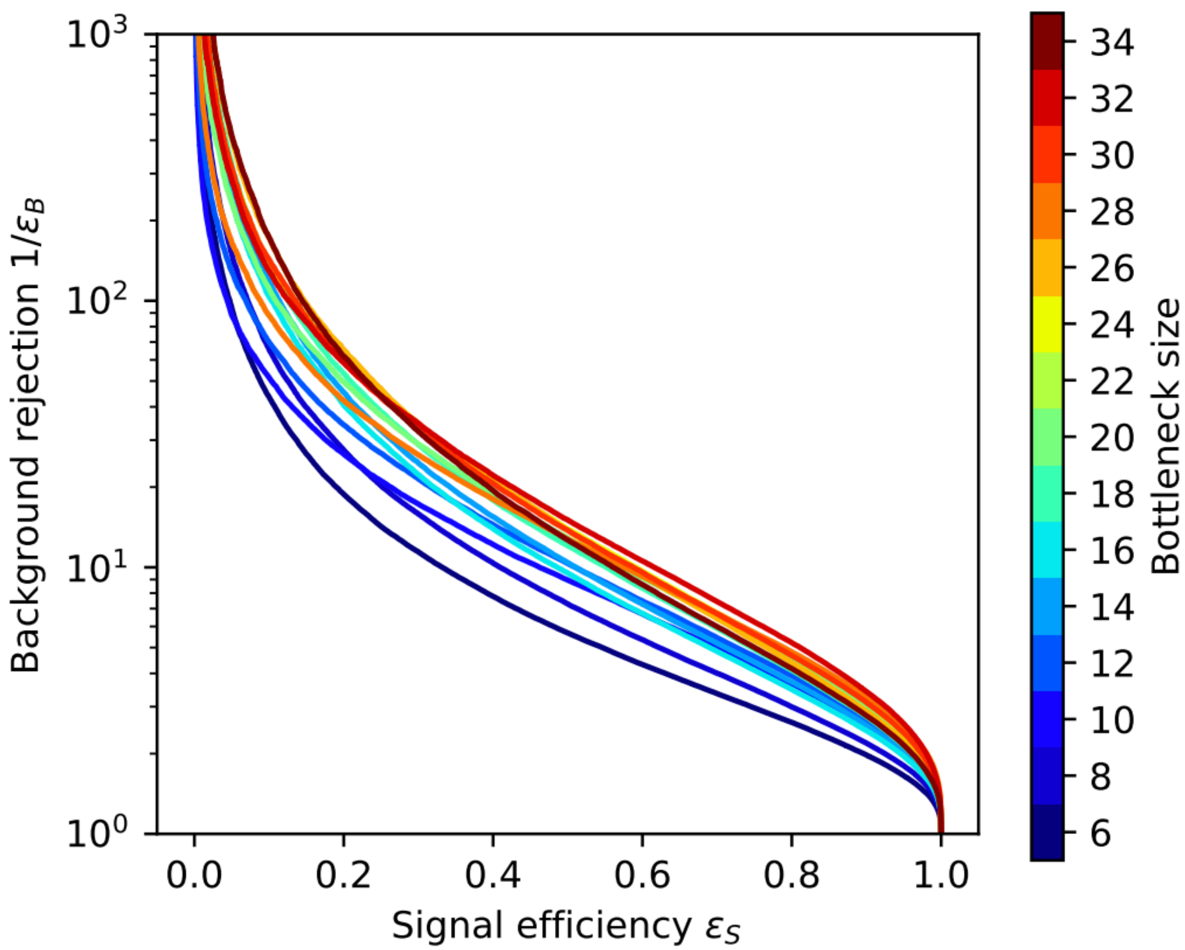
Autoencoders



Compressed
representation
Latent space

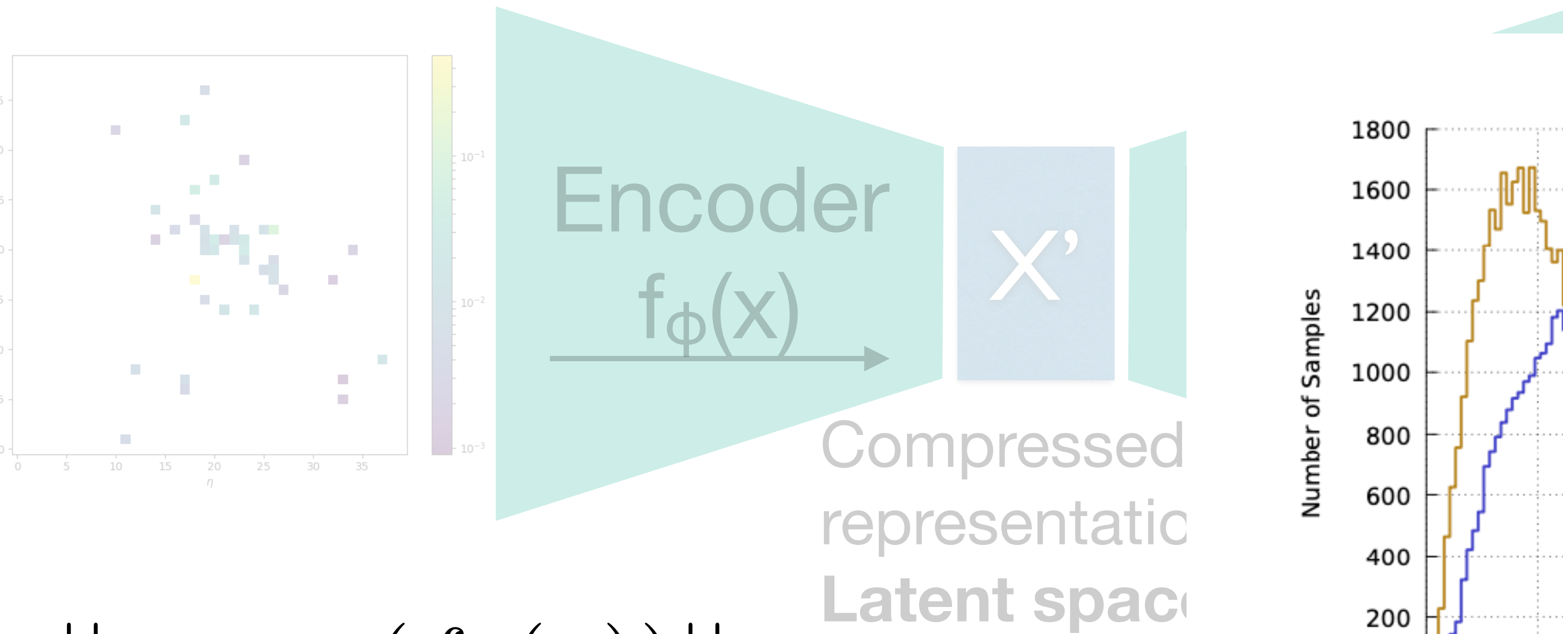
$$\mathcal{L}(x) = ||x - g_\theta(f_\phi(x))||_2$$
$$a(x) = \mathcal{L}(x)$$

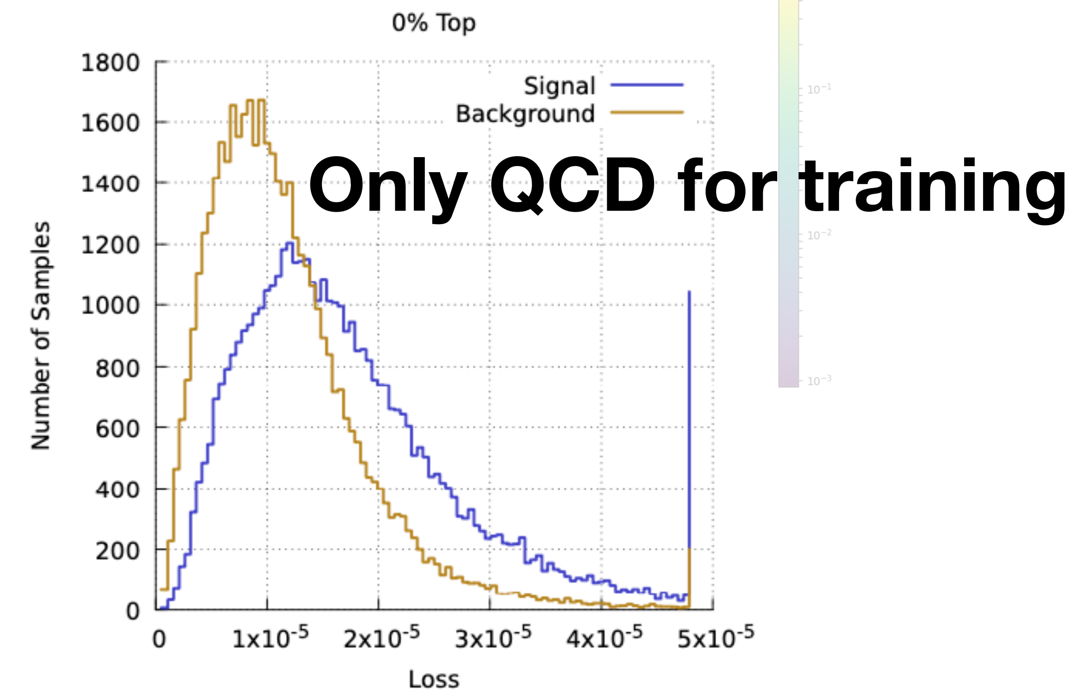
- Use that autoencoder approximates background density
- Loss = anomaly score
- Proof of concept on top tagging dataset

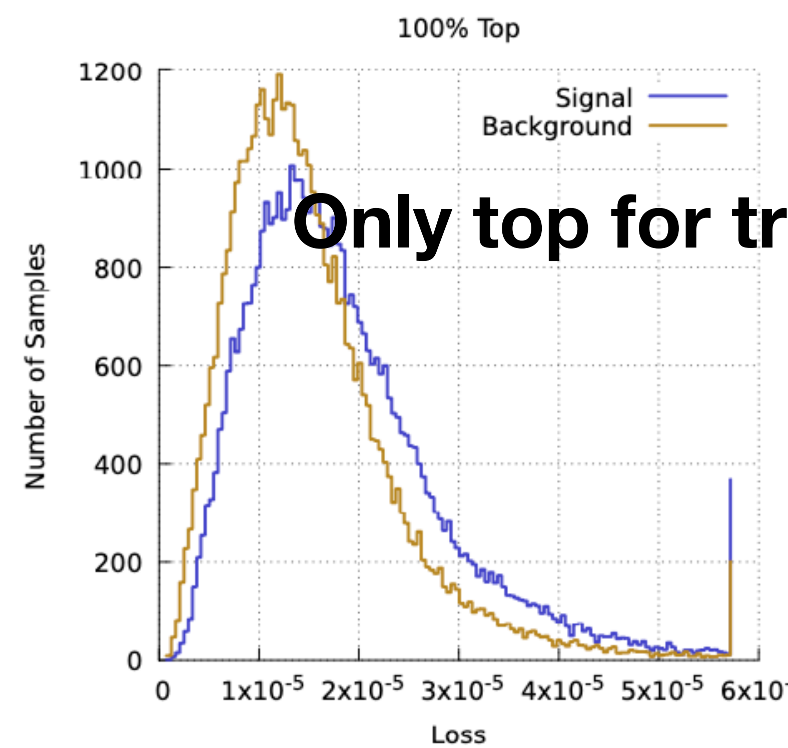


Heimel, GK, Plehn, Thompson, 1808.08979; Farina, Nakai, Shih, *1808.08992; ATLAS application:* ATLAS-CONF-2022-045; *…*

Autoencoders





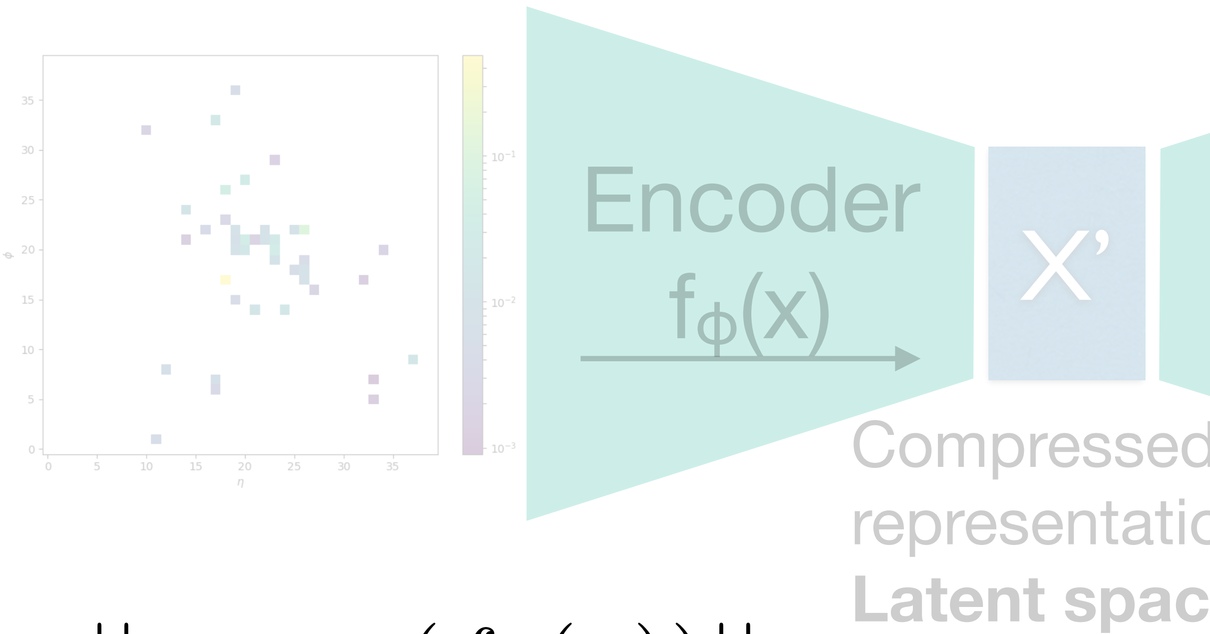


$$\mathcal{L}(x) = ||x - g_\theta(f_\phi(x))||_2$$
$$a(x) = \mathcal{L}(x)$$

- Upside: Powerful, conceptually simple, useful for trigger?
- Downside: Complexity bias

Weber MSc Thesis Hamburg 2019, Finke et al 2104.09051,..

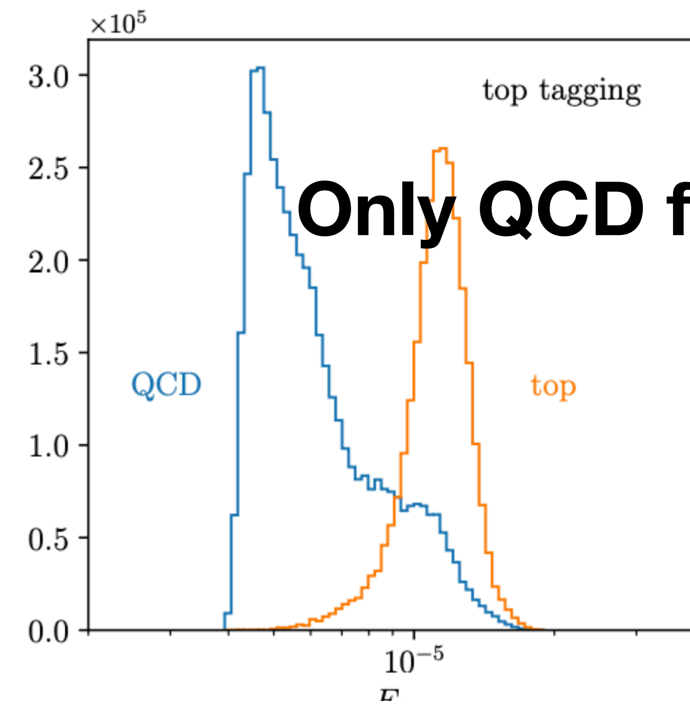
Autoencoders

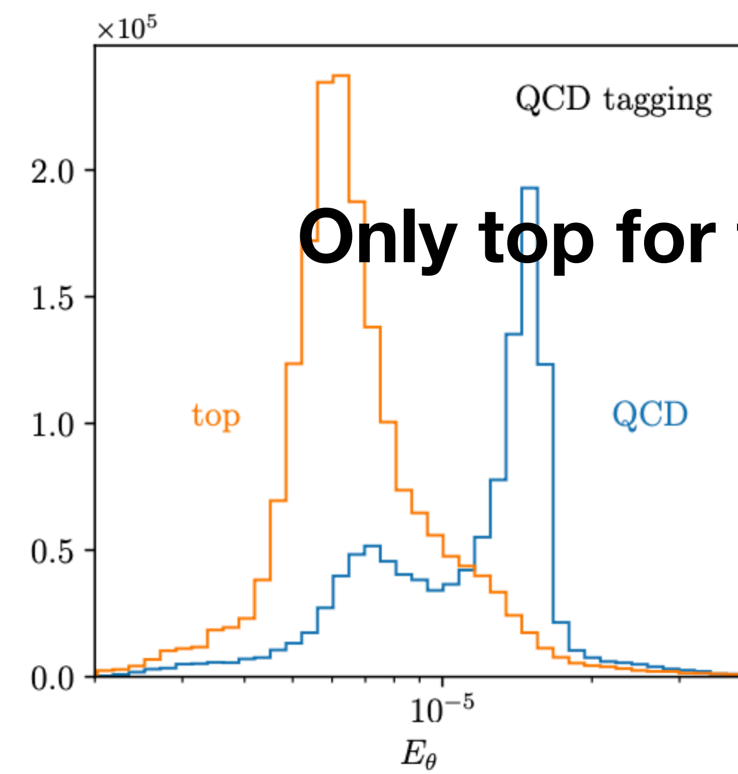


$$\mathcal{L}(x) = ||x - g_\theta(f_\phi(x))||_2$$

$$a(x) = \mathcal{L}(x)$$

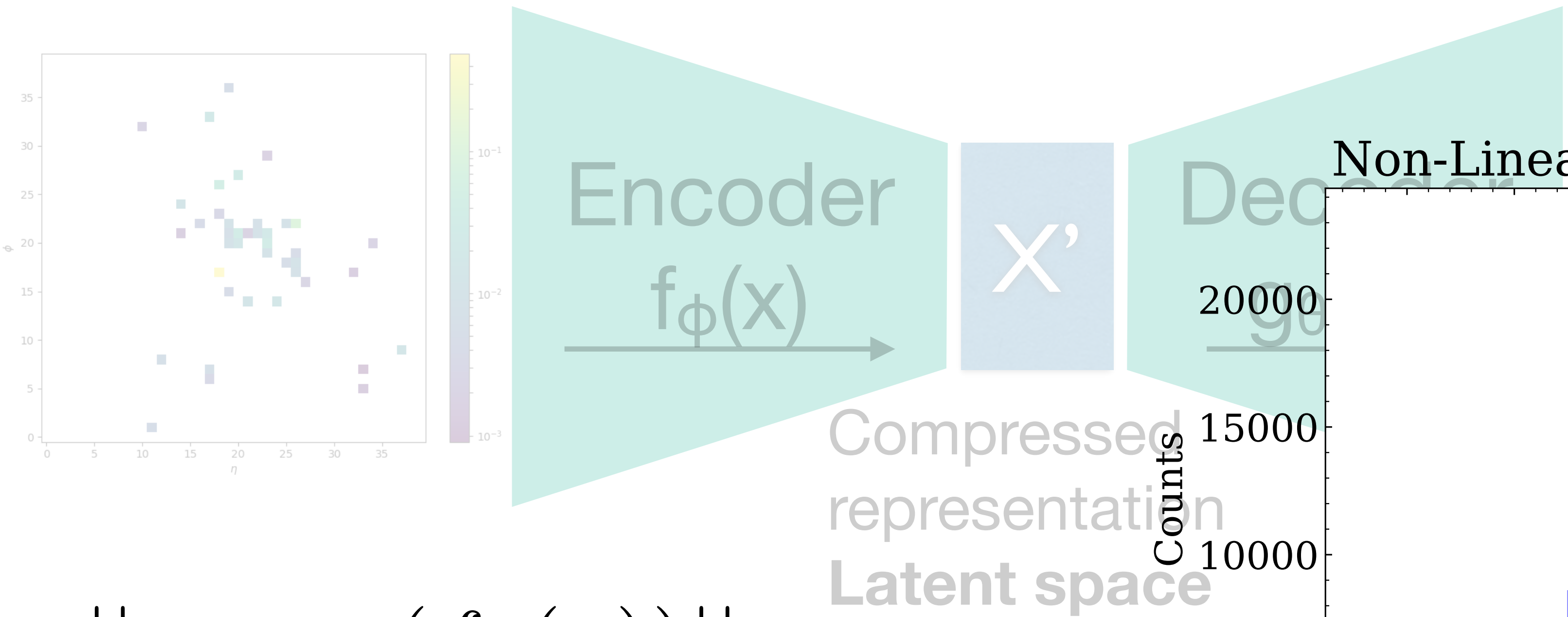
- Upside: Powerful, conceptually simple, useful for trigger?
- Downside: Complexity bias (Overcome e.g. by normalised auto encoders)

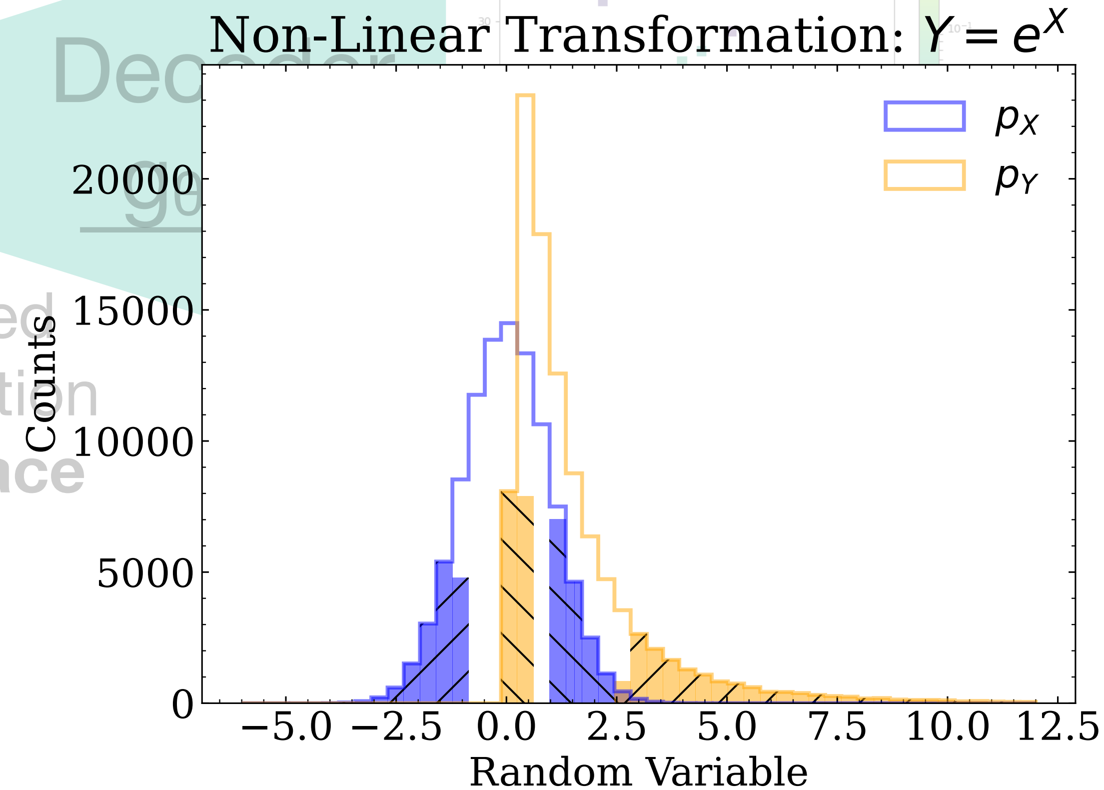




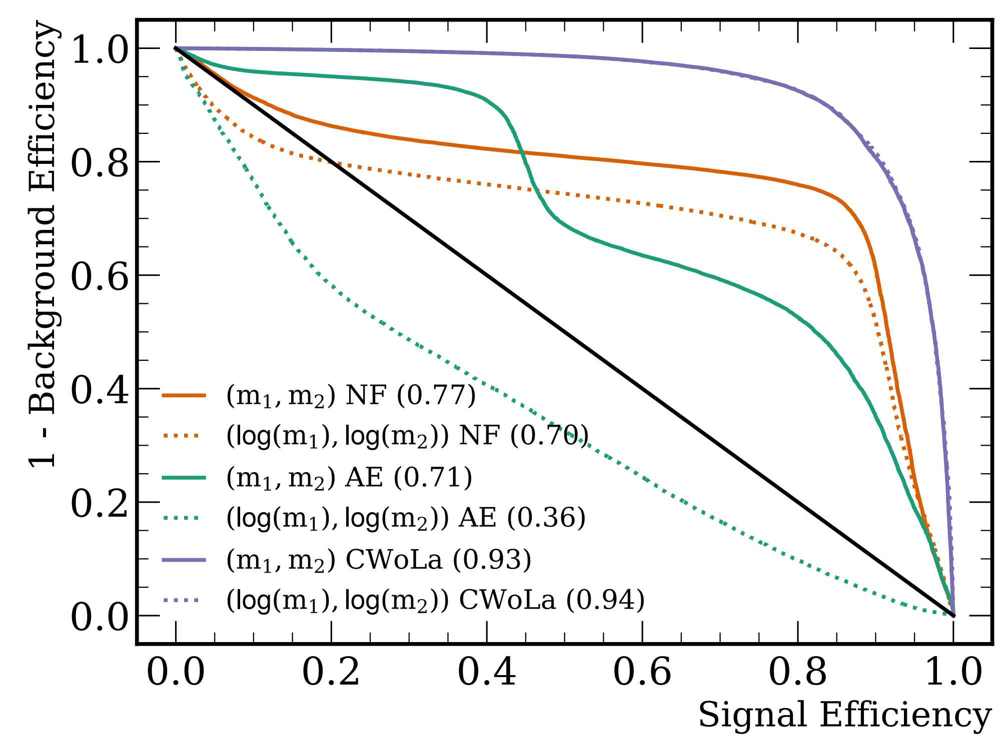
Yoon, Noh, Park 2105.05735; Dillon et al 2206.14225

Autoencoders





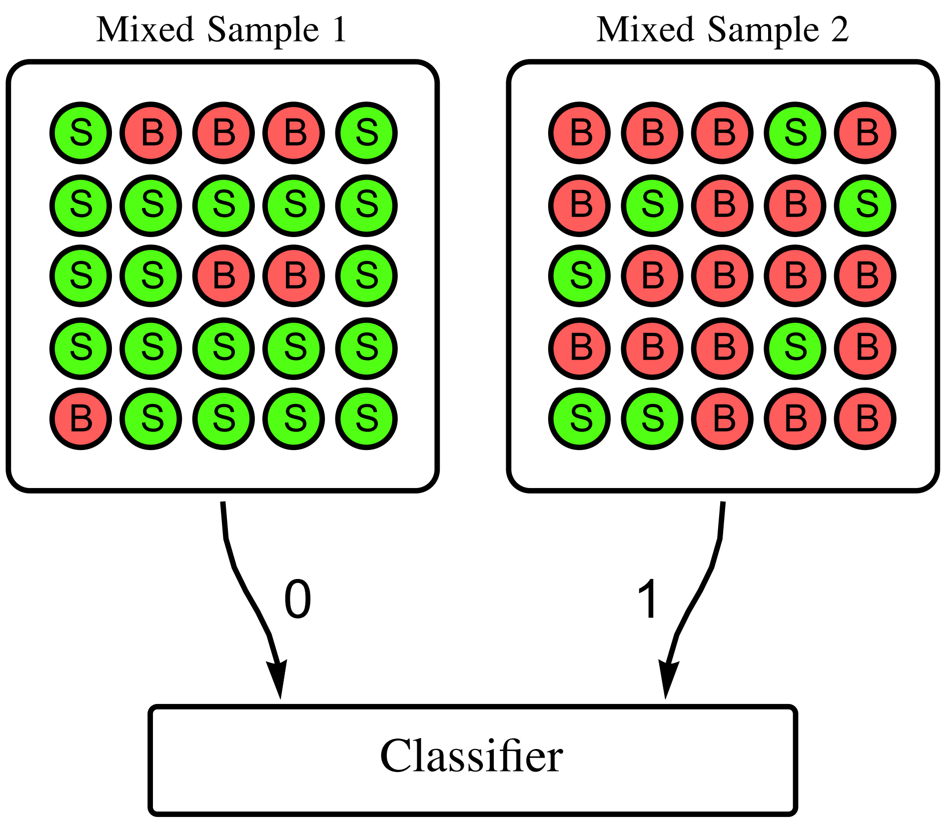
$$\mathcal{L}(x) = ||x - g_\theta(f_\phi(x))||_2$$
$$a(x) = \mathcal{L}(x)$$



- Upside: Powerful, conceptually simple, useful for trigger?
- Downside: Complexity bias (Overcome e.g. by normalised auto encoders)
  Ill-defined density under coordinate change

Lan, Dinh, 2012.03808; GK et al, 2209.06225

CWoLa



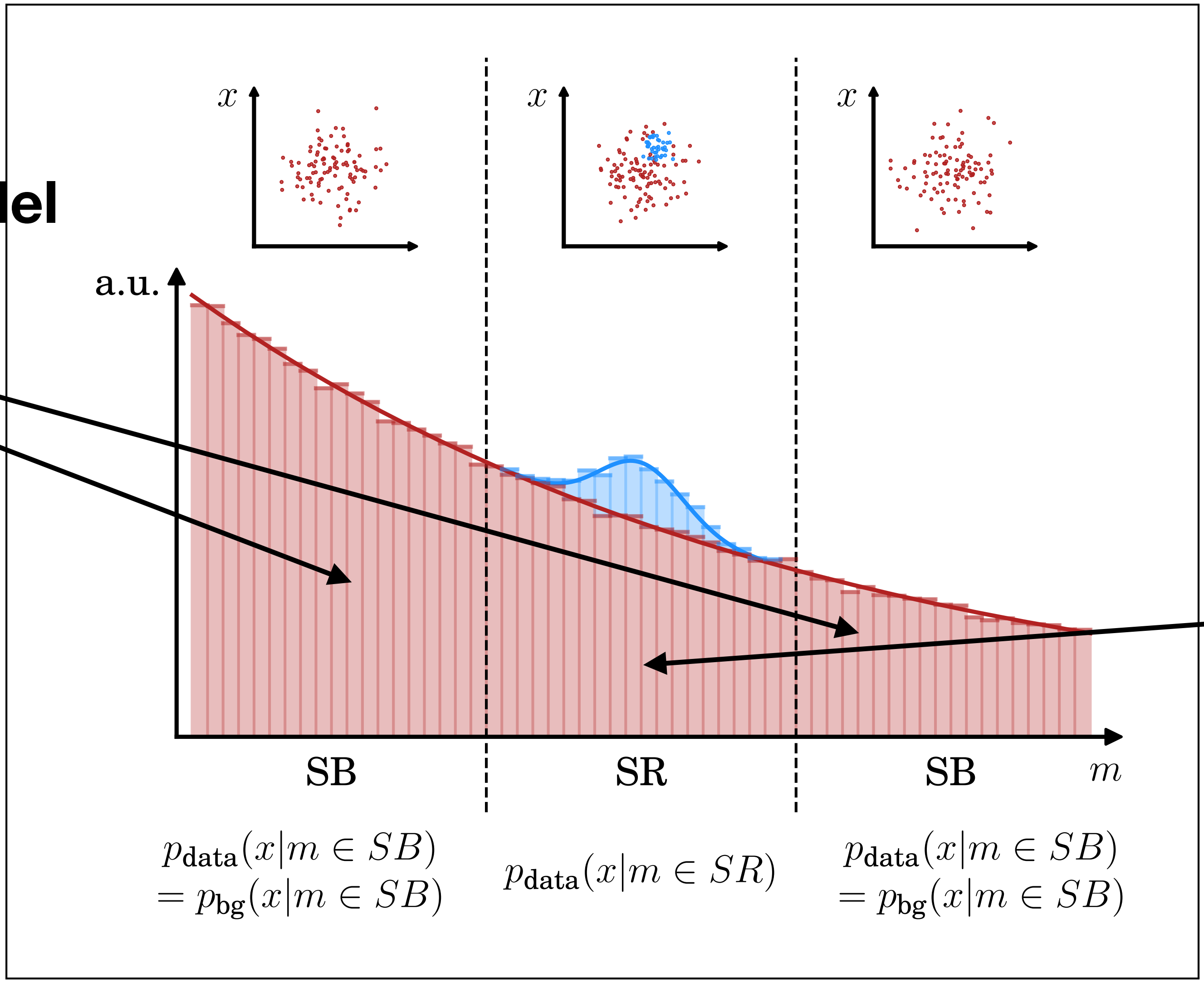
- A classifier (i.e. a neural network) trained to distinguish two mixed samples learns to distinguish the components
- But needs S/B from same underlying distribution (e.g uncorrelated) between mixed samples 1 and 2 (does not hold in general)

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

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

CATHODE

Train generative model here

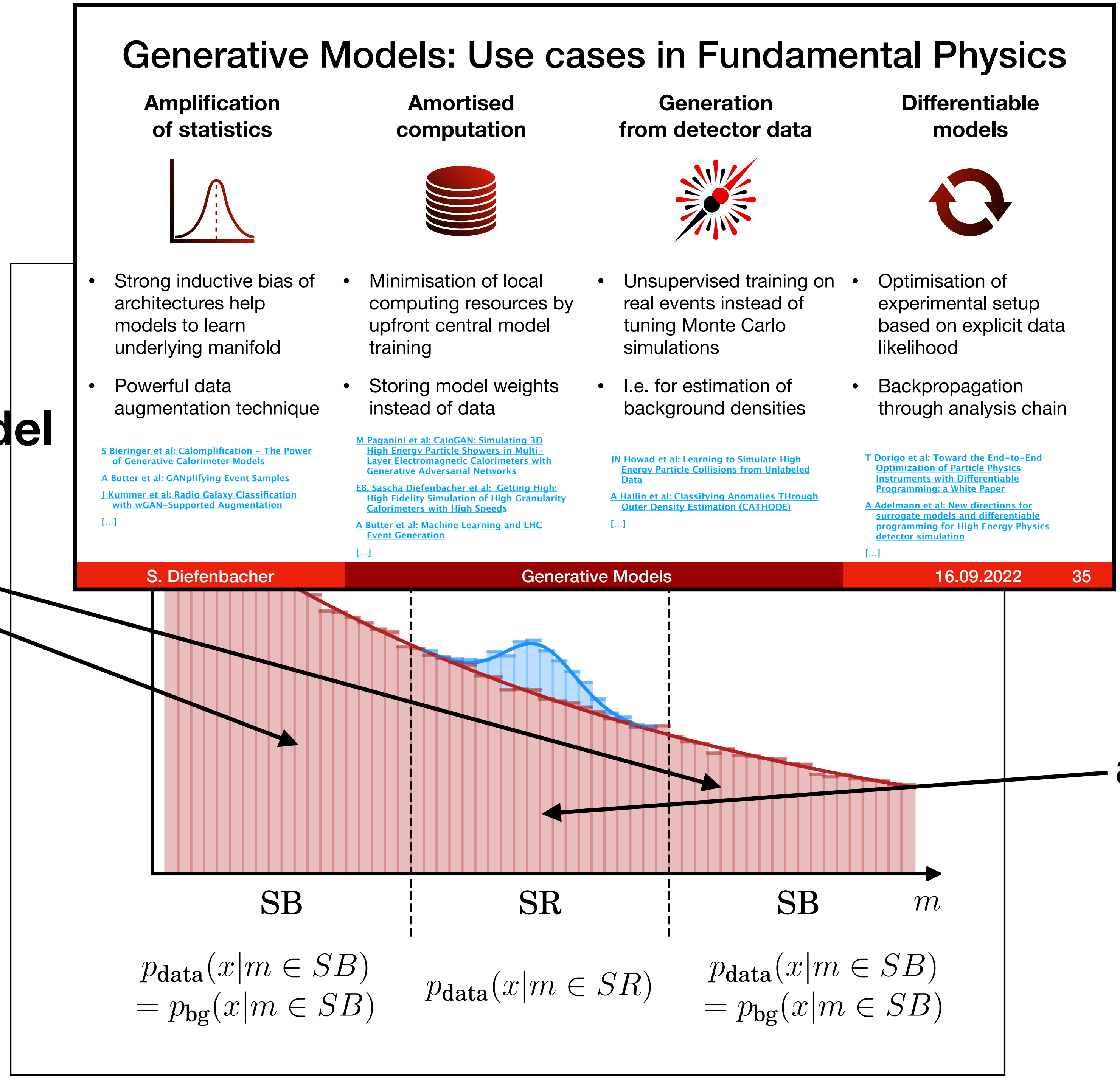


and sample here

- 1) Train a generative model p(x|m) on auxiliary features in SB
- 2) Sample from p(x|m) in SR. Designate as $p_{\mathrm{bg,est}}$.

Hallin, Isaacson, GK, et al 2109.00546; Nachman, Shih 2001.04990; Raine et al 2203.09470

Train generative model here



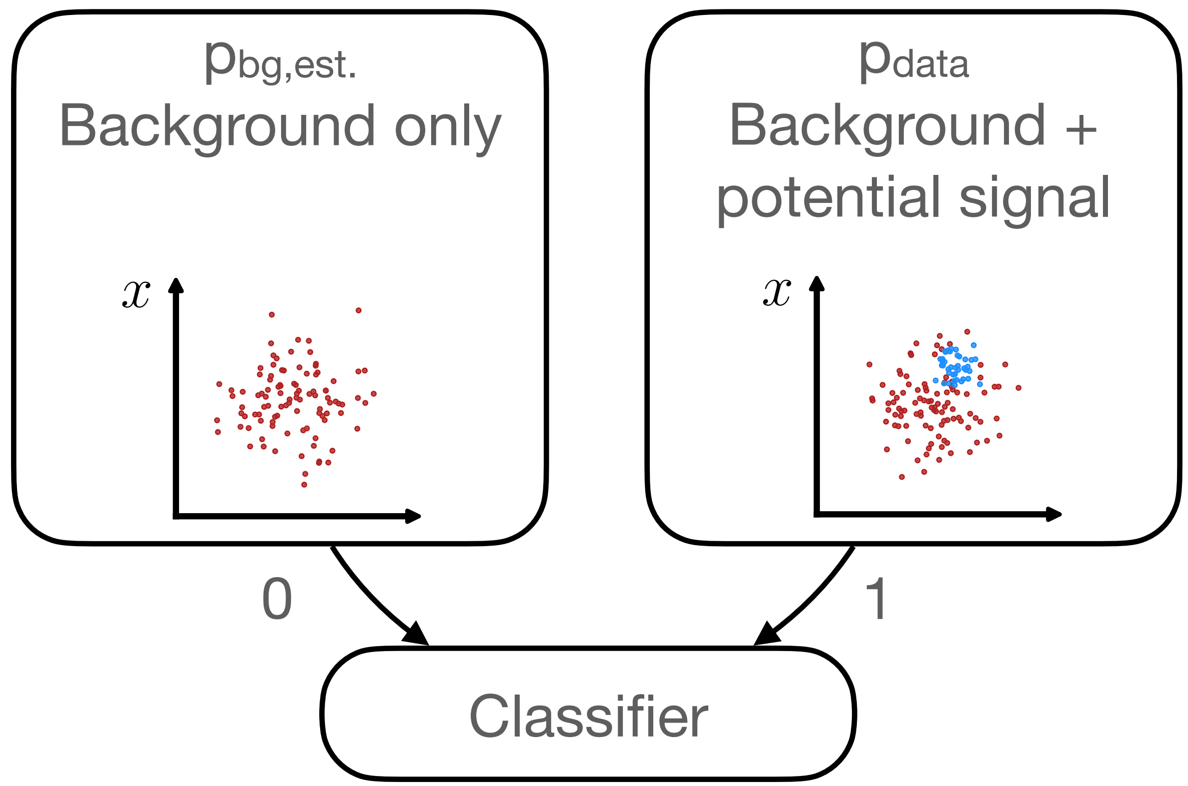
Remember Sascha's talk from yesterday

and sample here

- 1) Train a generative model p(x|m) on auxiliary features in SB
- 2) Sample from p(x|m) in SR. Designate as $p_{\text{bg,est}}$.

Hallin, Isaacson, GK, et al 2109.00546; Nachman, Shih 2001.04990; Raine et al 2203.09470

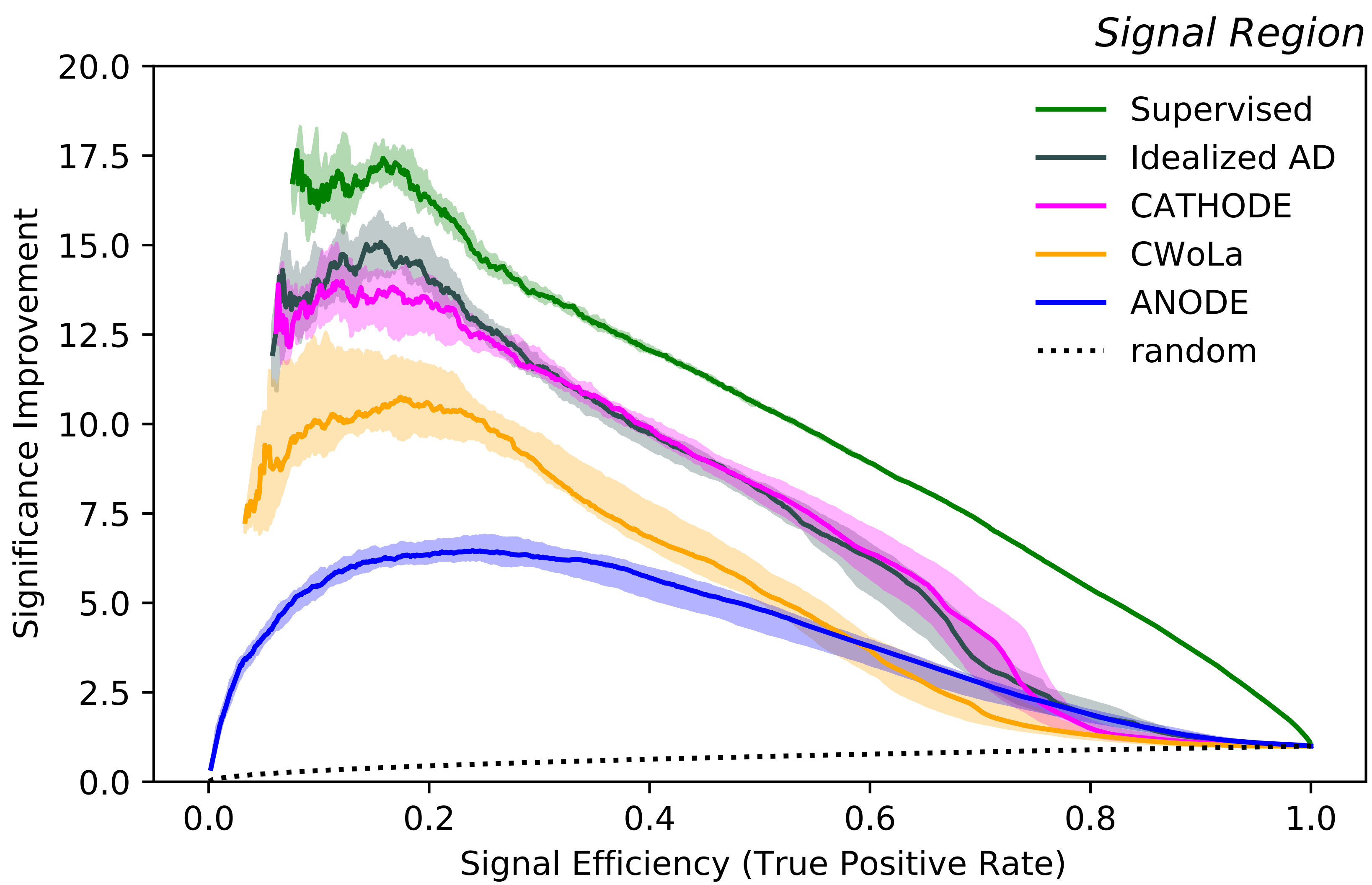
CATHODE



- 1) Train a generative model p(x|m) on auxiliary features in SB
- 2) Sample from p(x|m) in SR. Designate as $p_{bg,est.}$
- 3) Train binary classifer between p_{data} and $p_{bg,est.}$

Hallin, Isaacson, GK, et al 2109.00546; Nachman, Shih 2001.04990; Raine et al 2203.09470

CATHODE

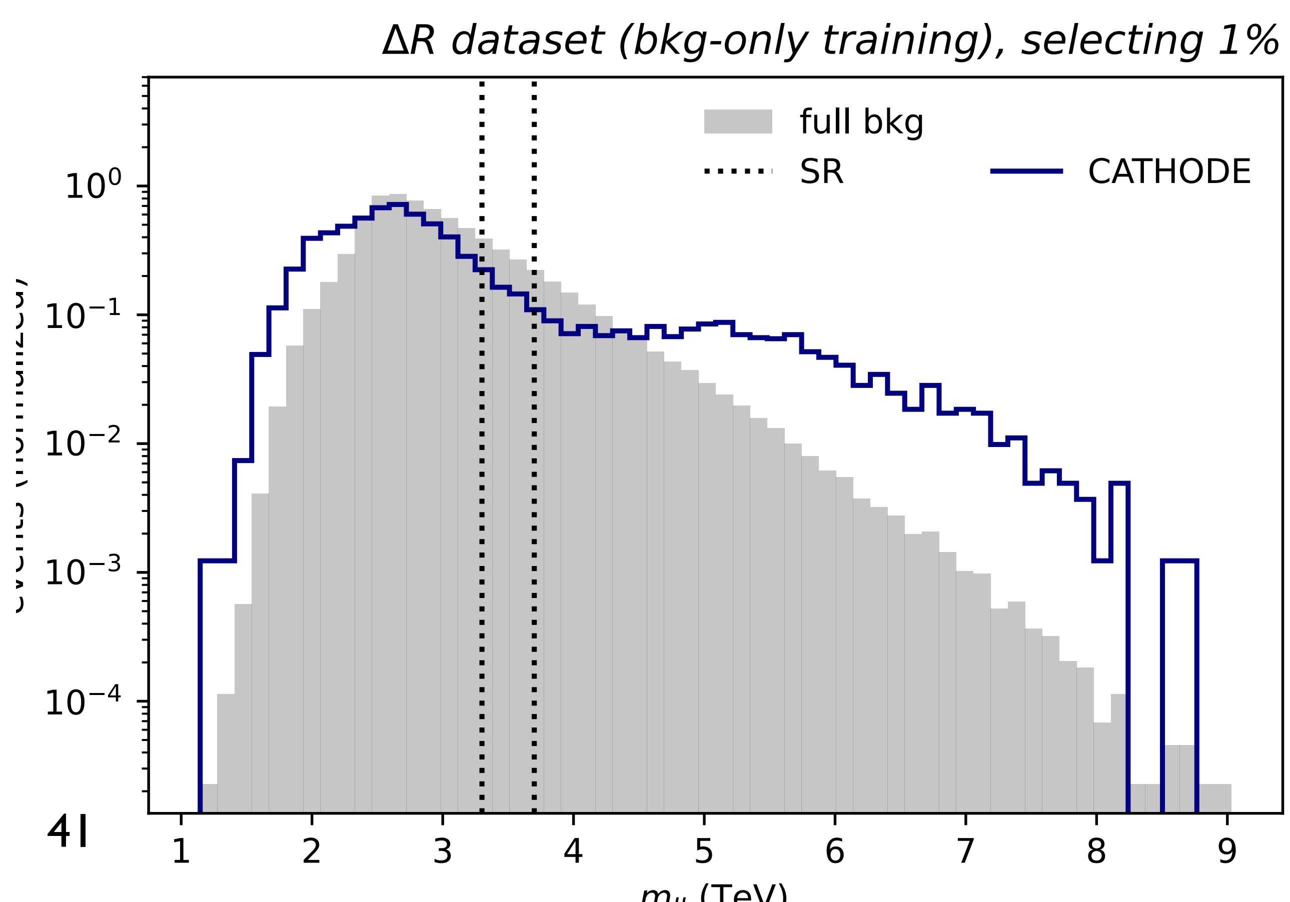


- 1) Train a generative model p(x|m) on auxiliary features in SB
- 2) Sample from p(x|m) in SR. Designate as $p_{bg,est.}$
- 3) Train binary classifer between p_{data} and $p_{bg,est.}$
- 4) Evaluate score, use in enhanced bump hunt

Hallin, Isaacson, GK, et al 2109.00546; Nachman, Shih 2001.04990; Raine et al 2203.09470

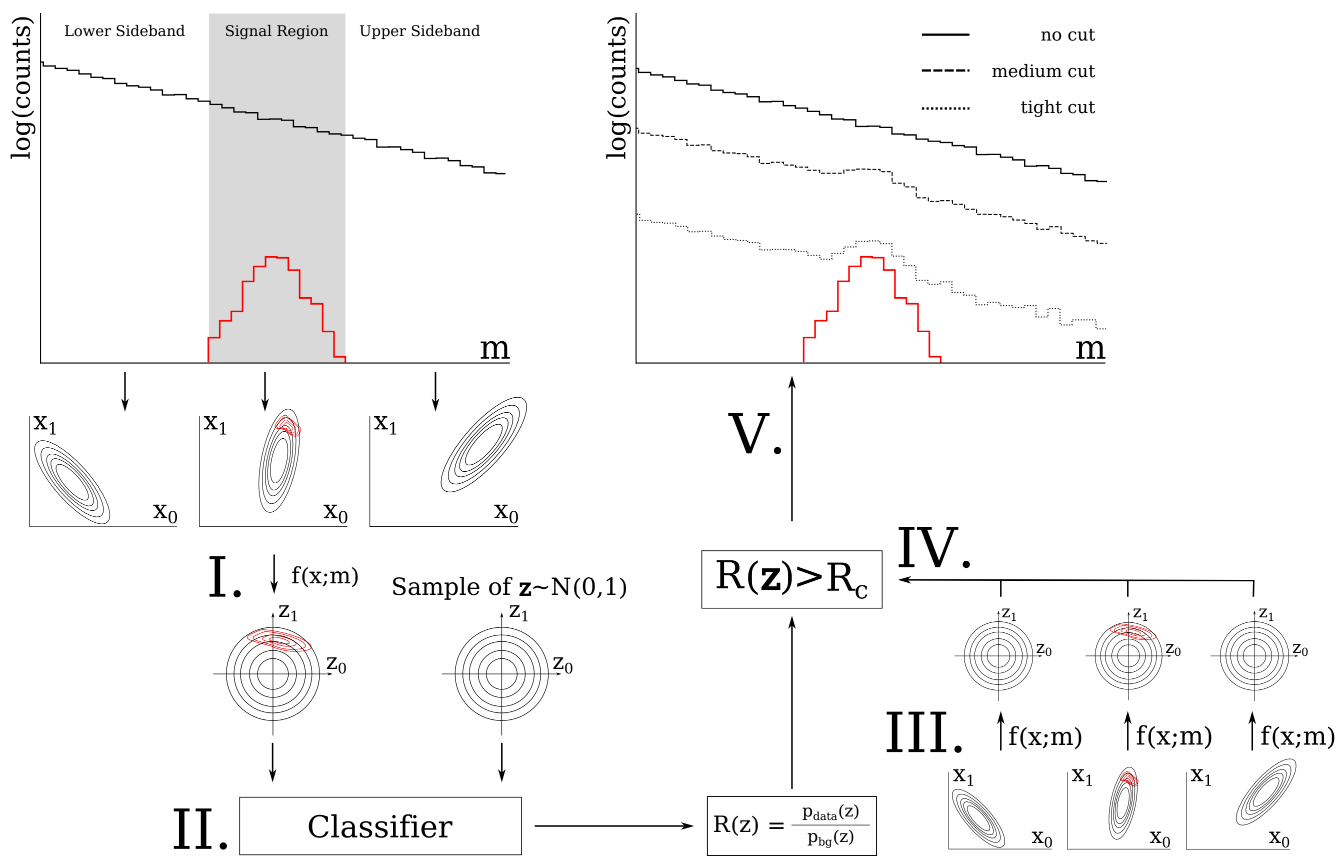
La(tent) CATHODE

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

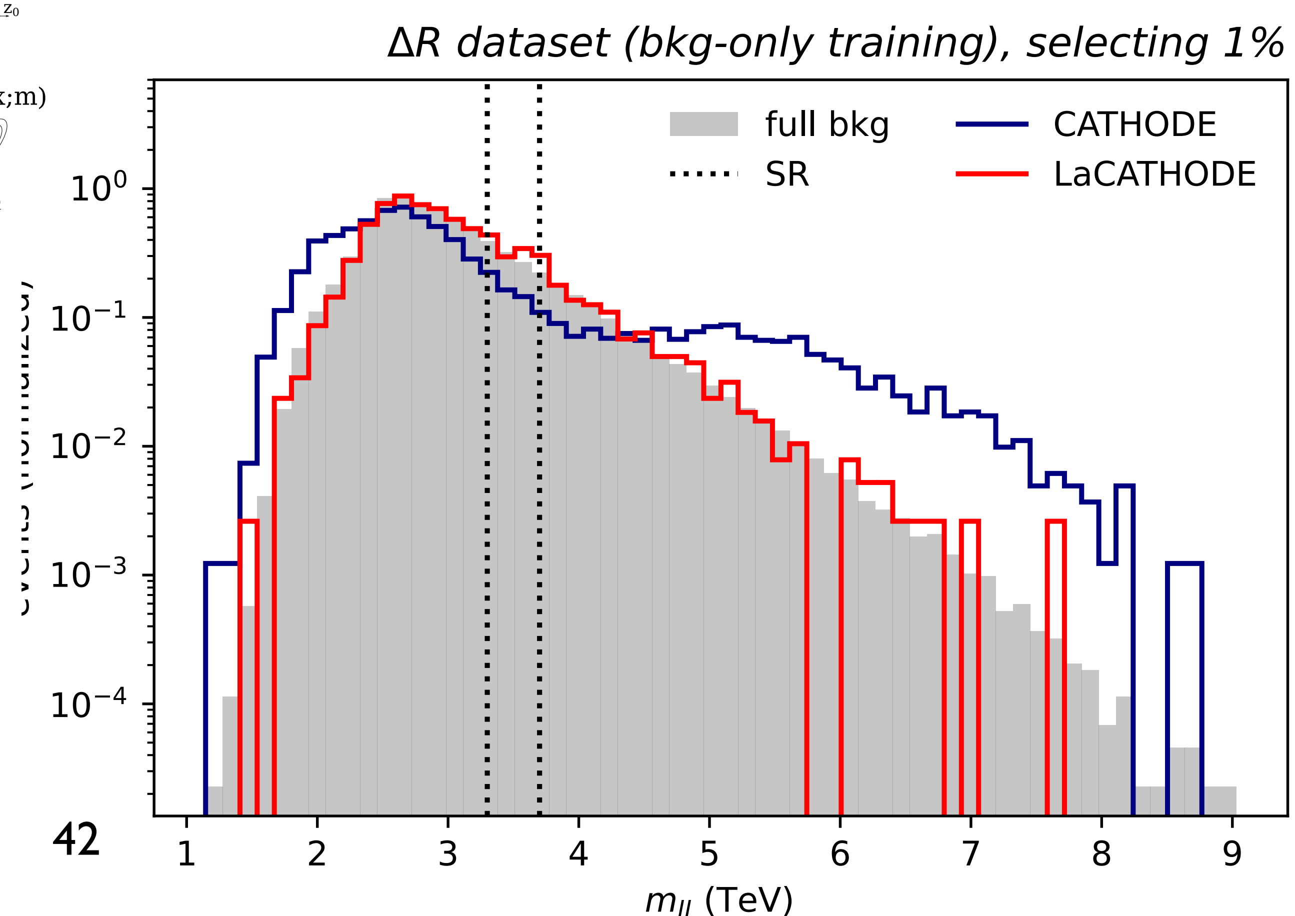


Hallin, GK, et al, coming soon

L-aCAT-HODE

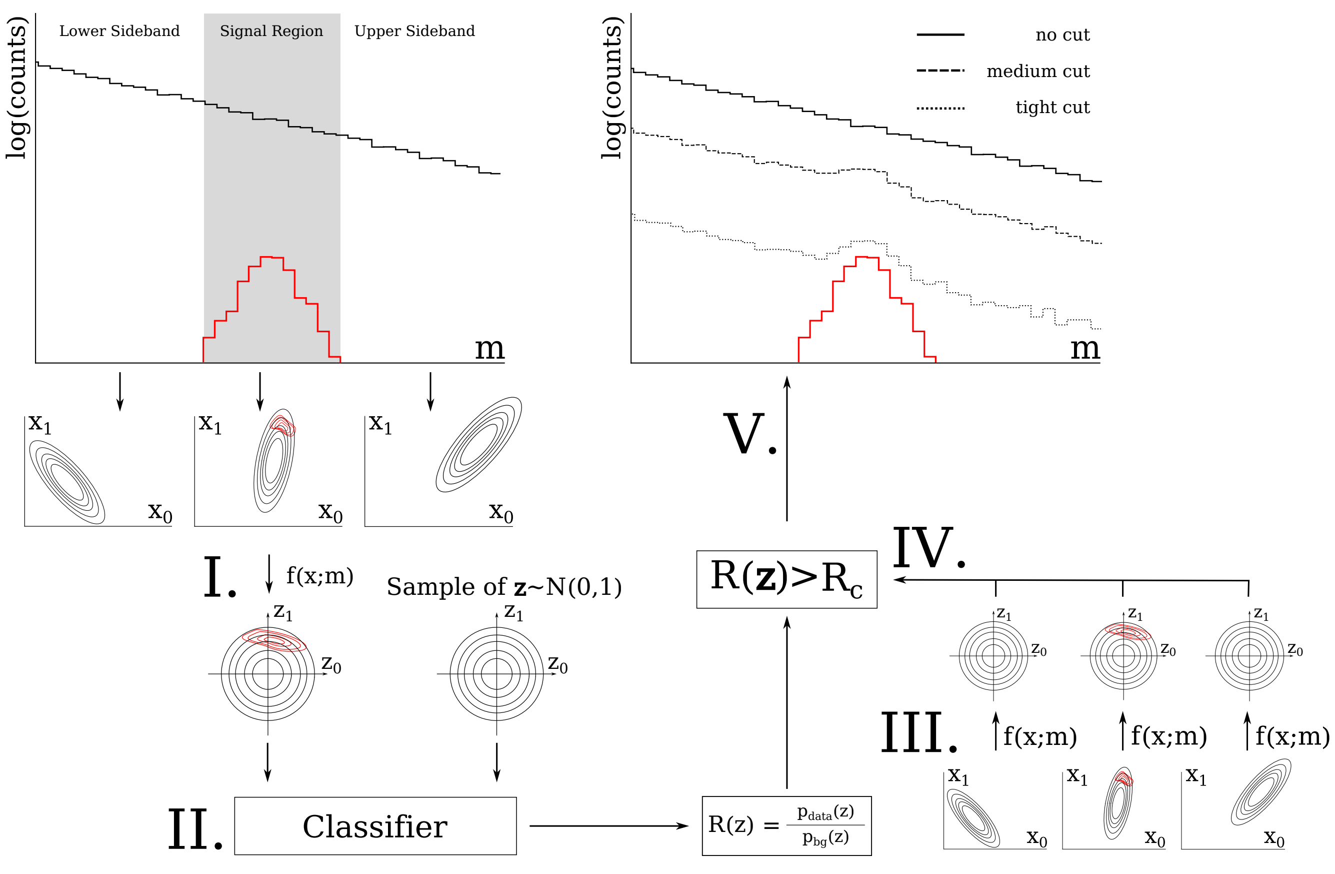


- If R(x) is only calculated in signal region, it's extrapolation behaviour is not well-defined
- Potential problem for bump-hunt if it shapes distributions
- Can overcome by training the classifier in latent space instead

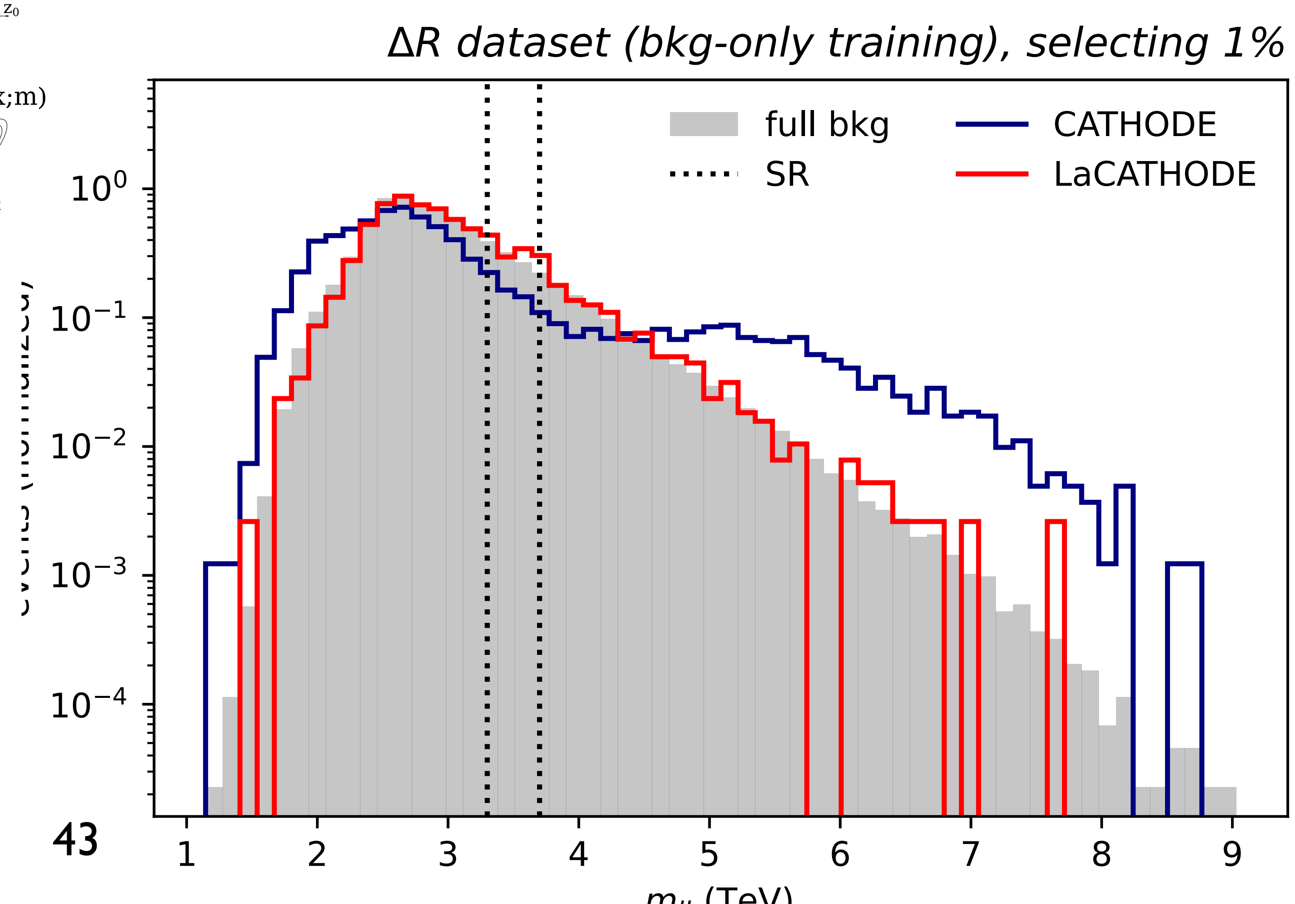


Hallin, GK, et al, coming soon

LaCATHODE



- If R(x) is only calculated in signal region, it's extrapolation behaviour is not well-defined
- Potential problem for bump-hunt if it shapes distributions
- Can overcome by training the classifier in latent space instead



Hallin, GK, et al, coming soon

Closing

Closing

- Machine learning aids new physics searches by improving existing approaches and opening up new techniques
    - Convergence of architectures - share foundation models?
    - Understanding and dealing with correlations
    - Generative models as in-situ background estimators
- Rapid pace of innovation, still no end in sight

Thank you!

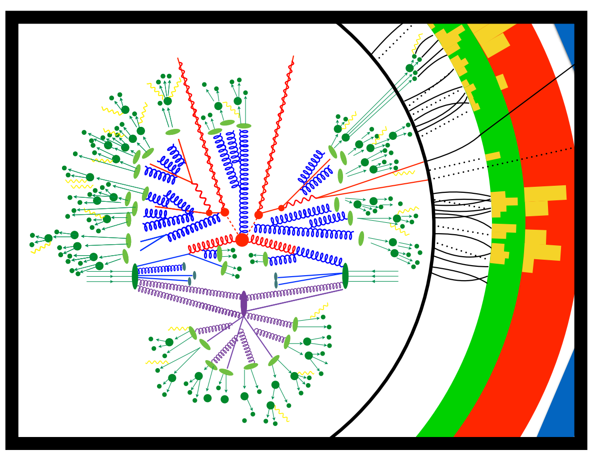
Bonus Material

Online Learning

Emphemeral Learning

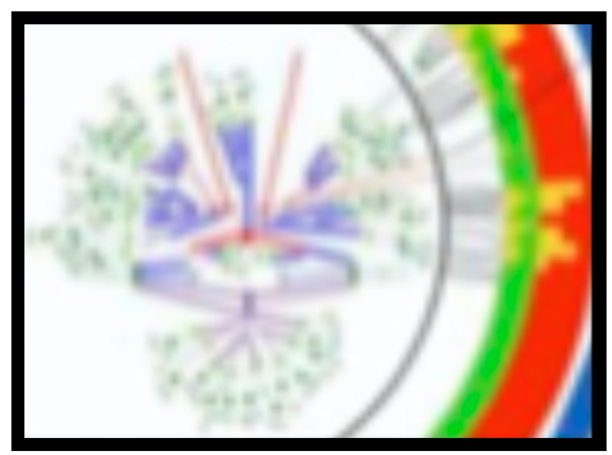
- CMS/ATLAS triggers:
  - Only able to store a subset (<1 in 10.000) of events
- Possible (wild) alternative:
  - Train a generative model online during data taking

No compression



Many numbers per **event**

Compress per event



Small set of numbers per **event**

Compress entire dataset

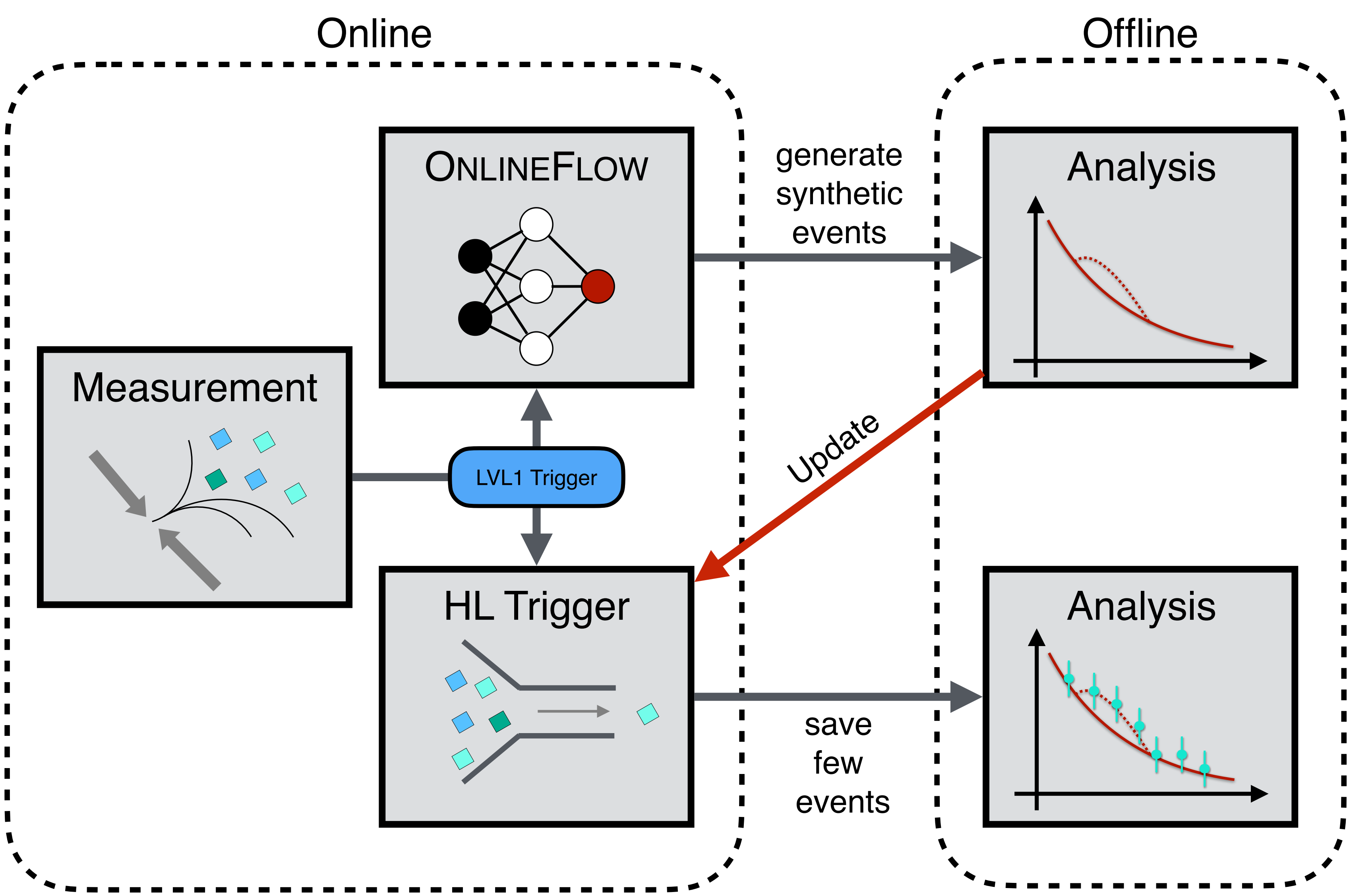
$$p(x\,|\,\theta)$$

Small set of numbers per **dataset**

- Fixed size, independent of training data amount
- Radically different format from usual way of storing data, but might open up new approaches

2202.0937

OnlineFlows



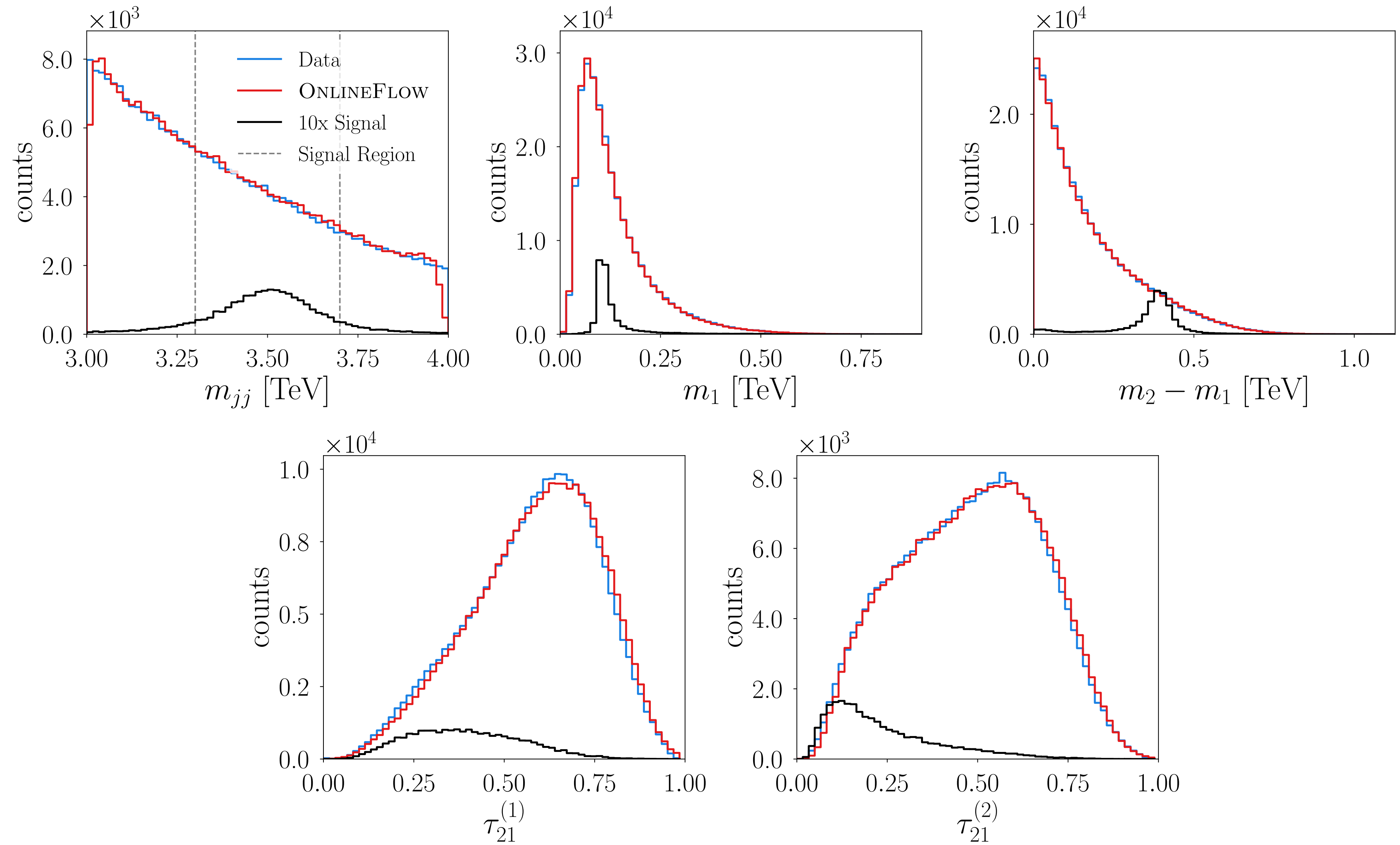
Schematic of proposed approach.

Focus on HLT, more technical challenges for use in hardware Trigger.

Main problem:
How to make training work if each event is only available for short time?

Proof of concept

Use LHCO dataset, train on high-level features on a mixture of background (99%) and signal (1%).

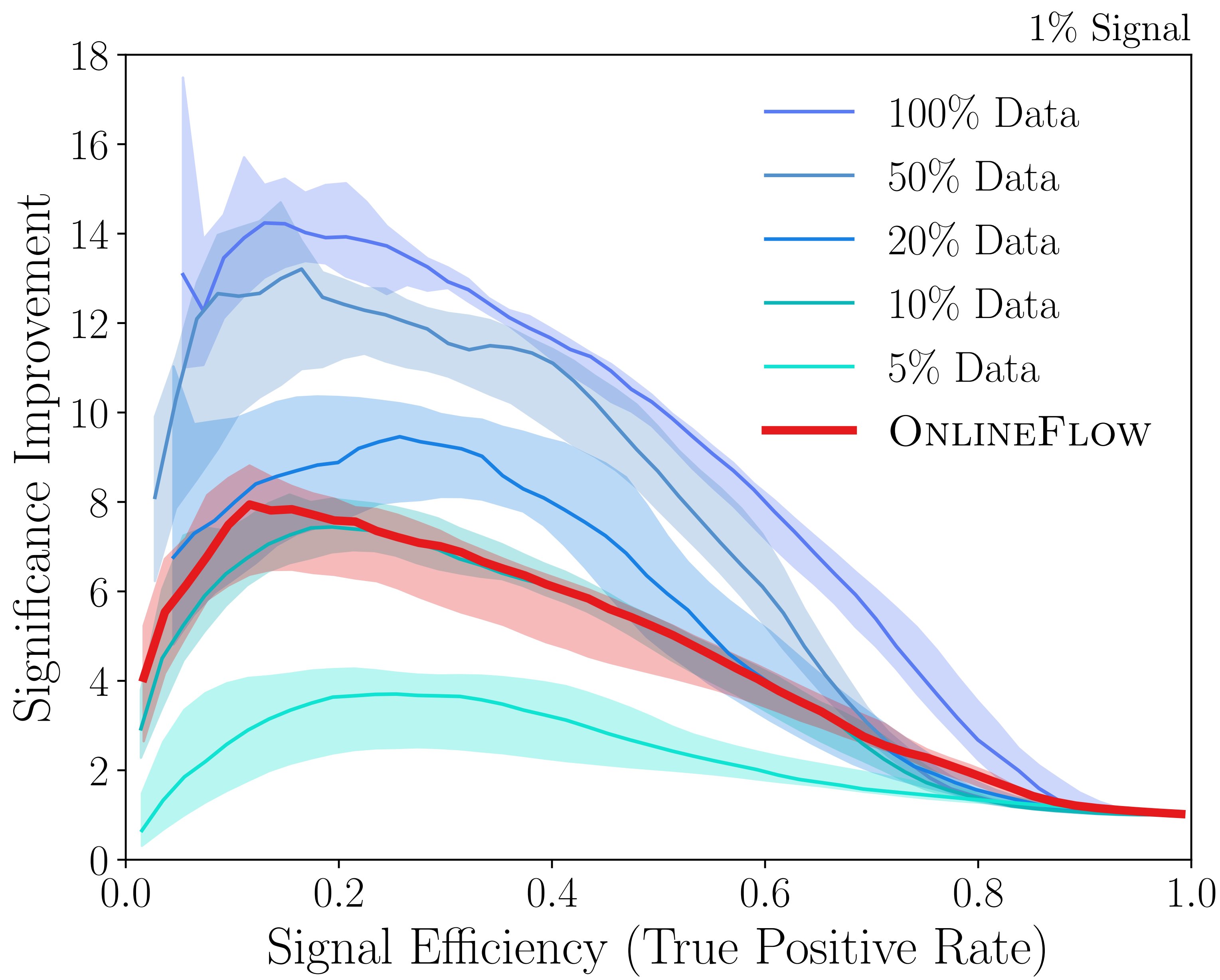


Proof of concept

Use LHCO dataset, train on high-level features on a mixture of background (99%) and signal (1%).

Train classifier to distinguish a signal region and sideband (CWoLA appaorach)

Compare procedure directly carried out on data with output of flow.



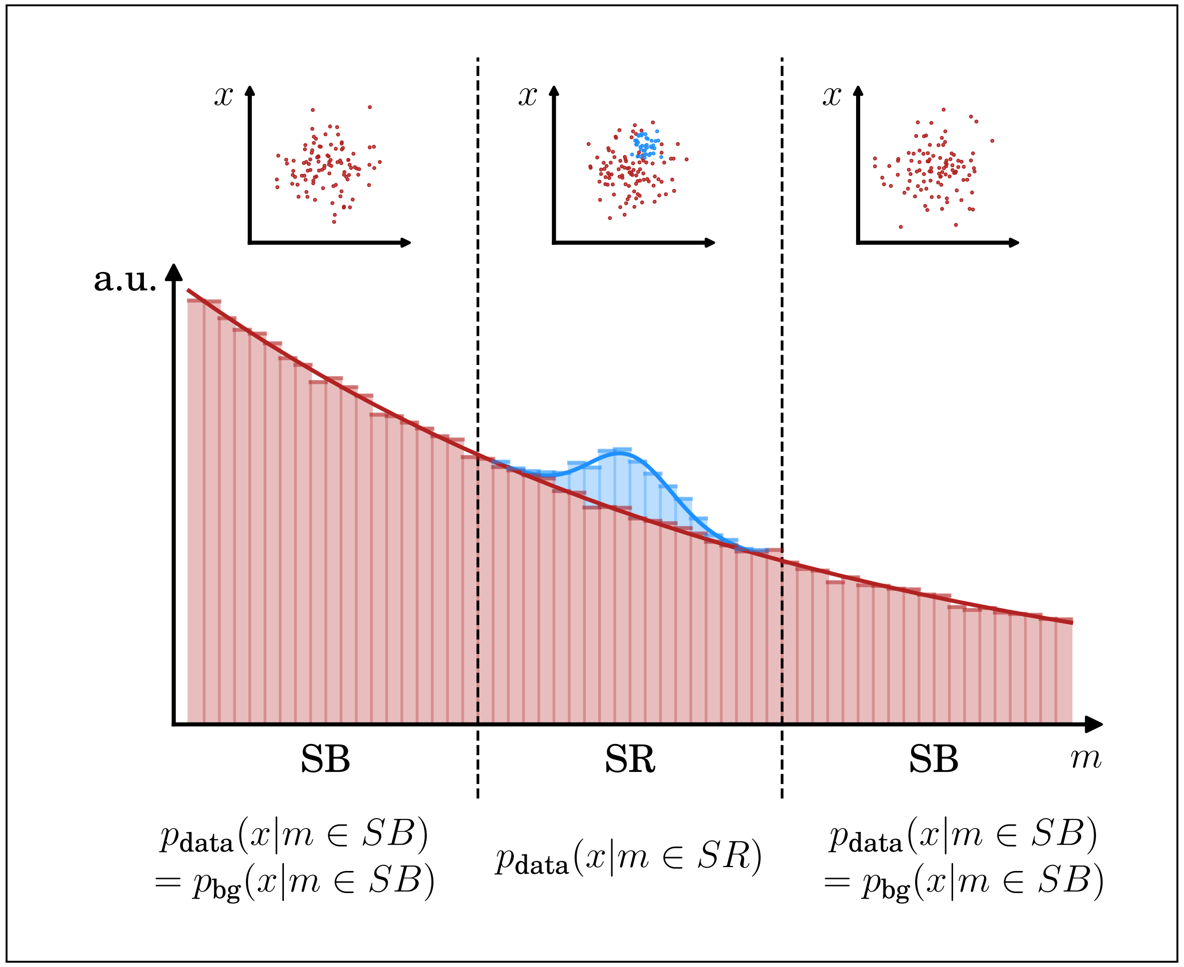
2202.0937

On Anomaly Detection

Types of anomalies

- **Outliers/Point anomalies**: Datapoints far away from regular distribution
- Examples:
  - Detector malfunctions
  - Background-free search

- **Group anomlies**: Individual examples not interesting,
  but signal is an overdensity with respect to background
- Examples:
  - Resonance searches
  - Transient signals in time series



Assumptions

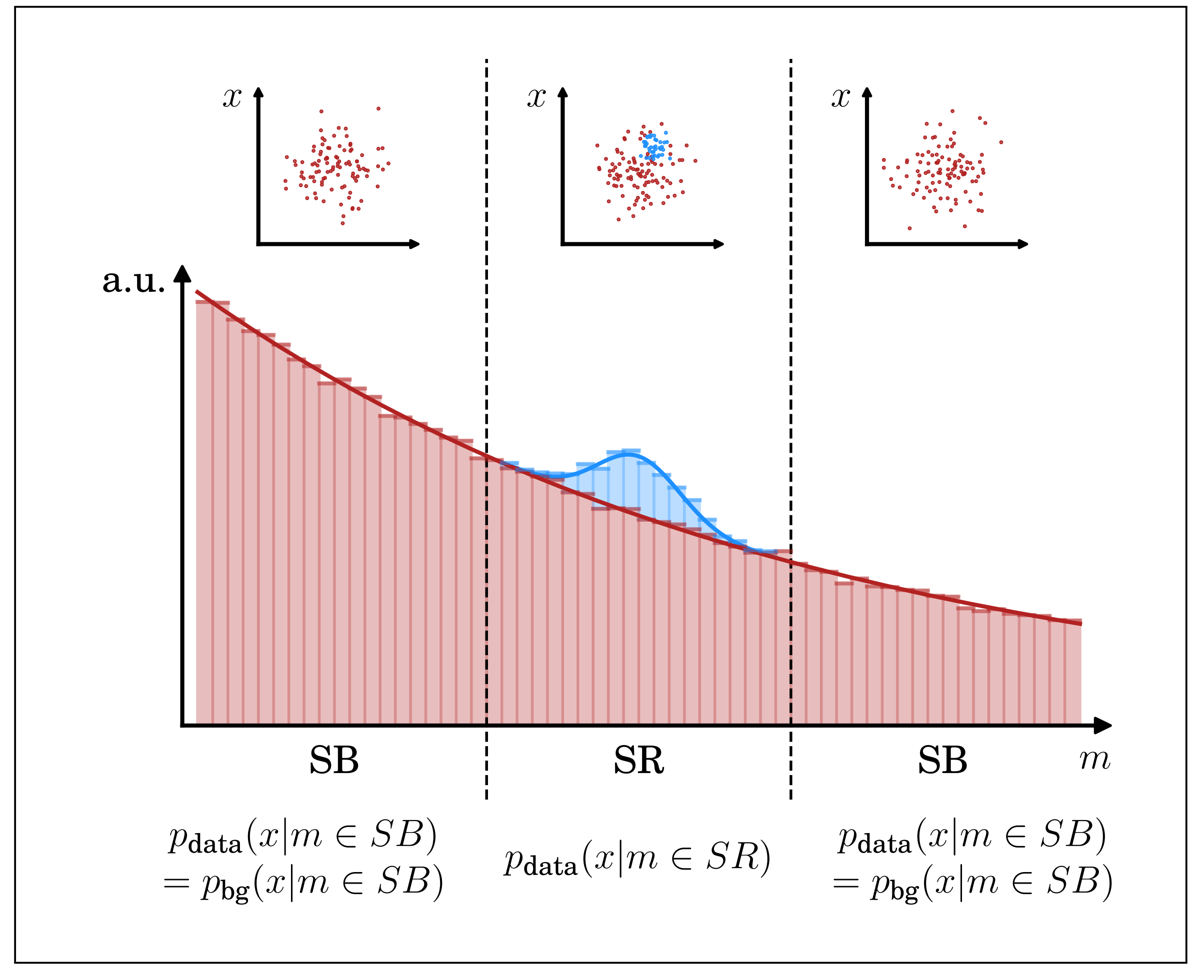
Rarity: Pr(anomaly) $\ll$ Pr(normal)

Overlap:

max $_x$ $p(x|\text{anomaly})/p(x|\text{normal}) < \infty$

Resonance: Pr($|m - m0| > \delta|$anomaly) $\approx$ 0 for some feature m (often a mass) and fixed $m0$, δ

Smoothness: $p(x|m$, normal) varies slowly with m so that one can use data with $|m - m0| > \delta$ to estimate $p(x|m$,normal) for $|m - m0| < \delta$



2101.08320

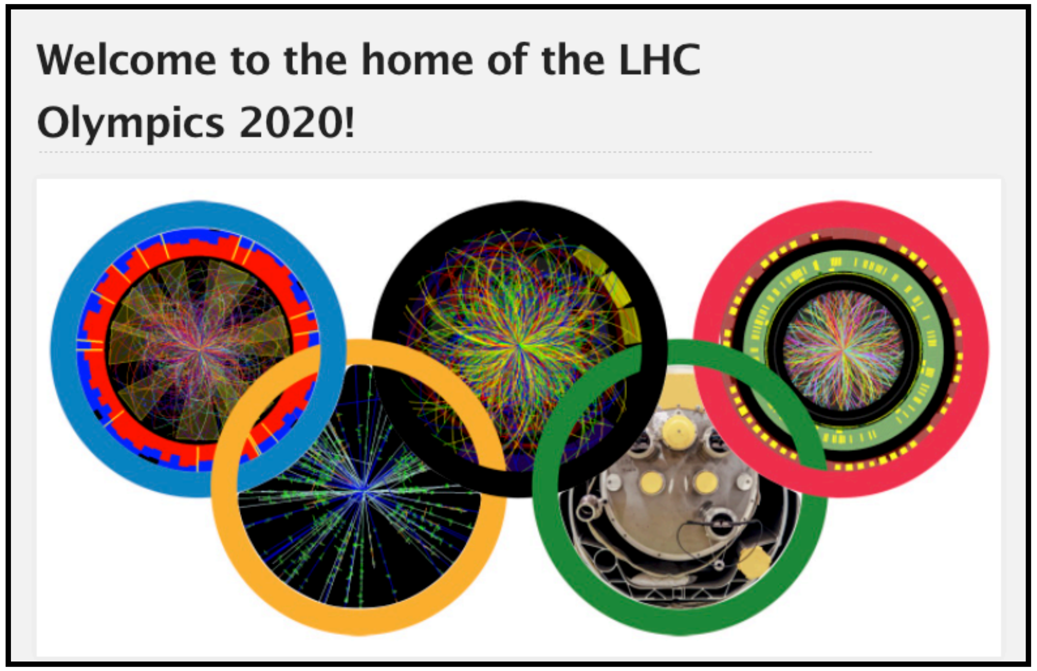
Introducing: LHC Olympics

- Encourage development and comparison of model-agnostic search strategies
    - Focus on group anomalies, data-driven searches
    - Use for a convenient overview of space of techniques
    - Complementary to 2105.14027

- Provide a complete package, balance details vs accessiblity

- Datasets:
    - One R&D dataset for algorithm development
    - Three black box datasets (BB1-BB3)
        - Unblinded over time

- Timeline:
    - Spring 2019: Release R&D dataset (link)
    - Autumn 2019: Release BB datasets (link)
    - January 2020: Winter Olympics as part of ML4Jets, unblinding of BB1 (link)
    - July 2020: (Virtual) Summer Olympics, unblinding of BB2 and BB3 (link)
    - LHC Olympics paper (https://arxiv.org/abs/2101.08320) public



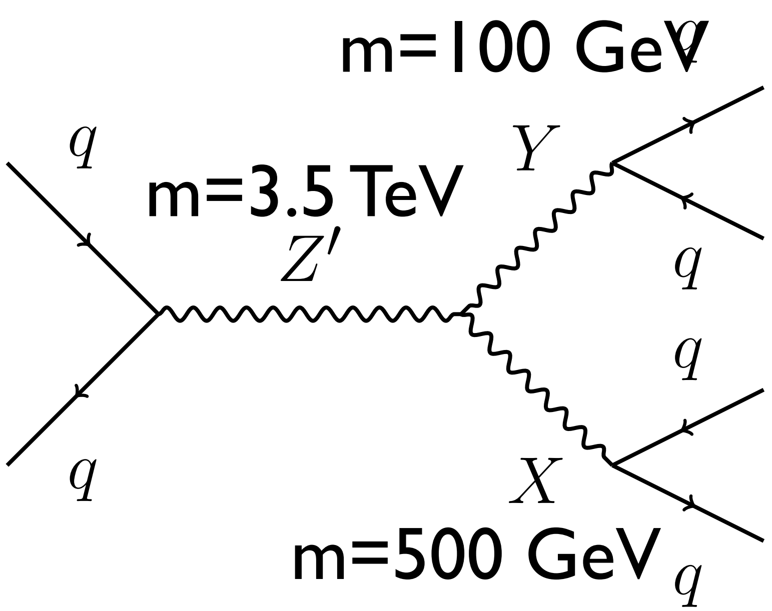
https://lhco2020.github.io/homepage/

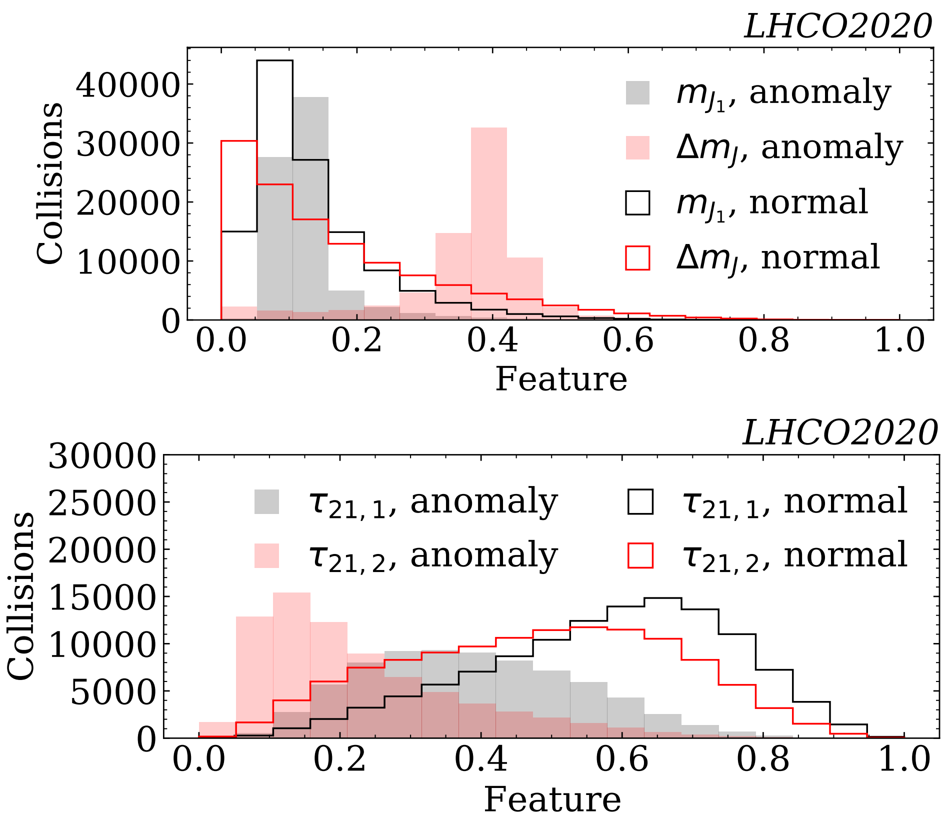
2101.08320

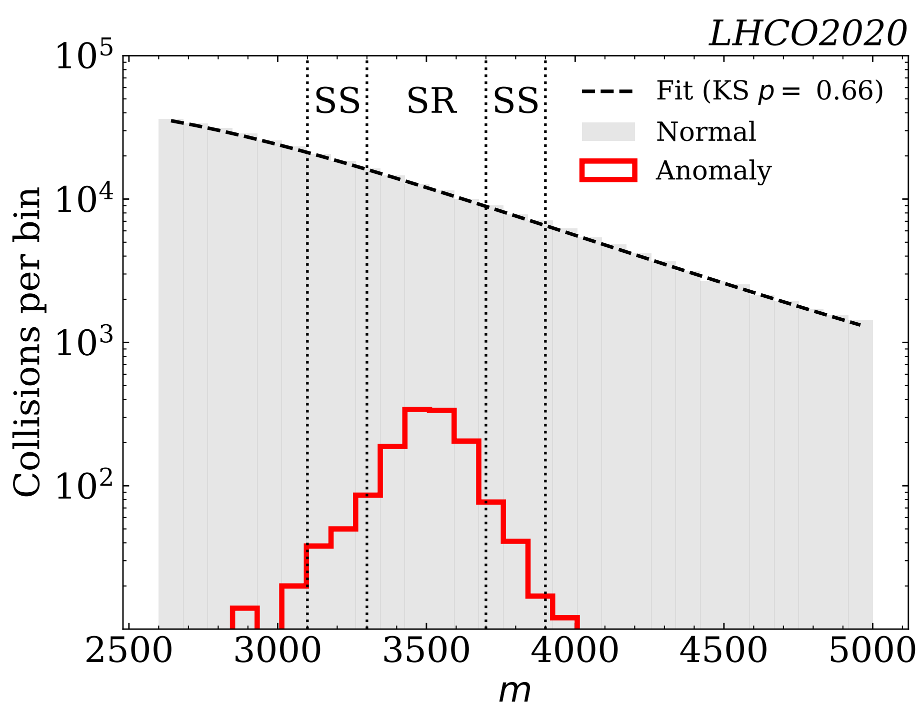
R&D dataset

- For building and testing methods
- 1M background examples (Standard Model), 100k signal examples (signal, see Feynman diagram on the right)
- Labels provided

- Relatively simple signal
    - Known to differ in previously mentioned features from background distribution
- Unrealistically high S/B

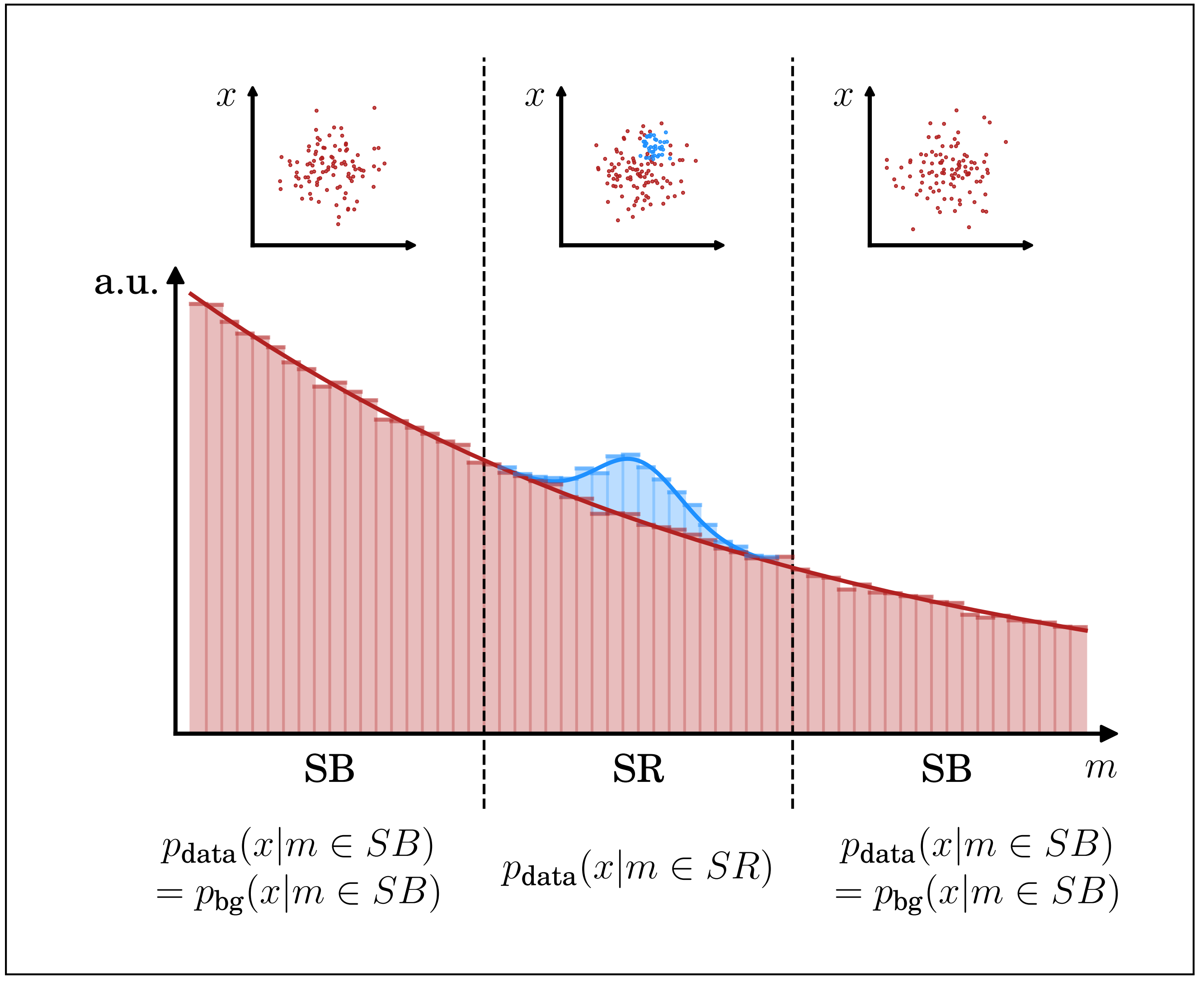






2107.02821

Generative models for anomaly detection

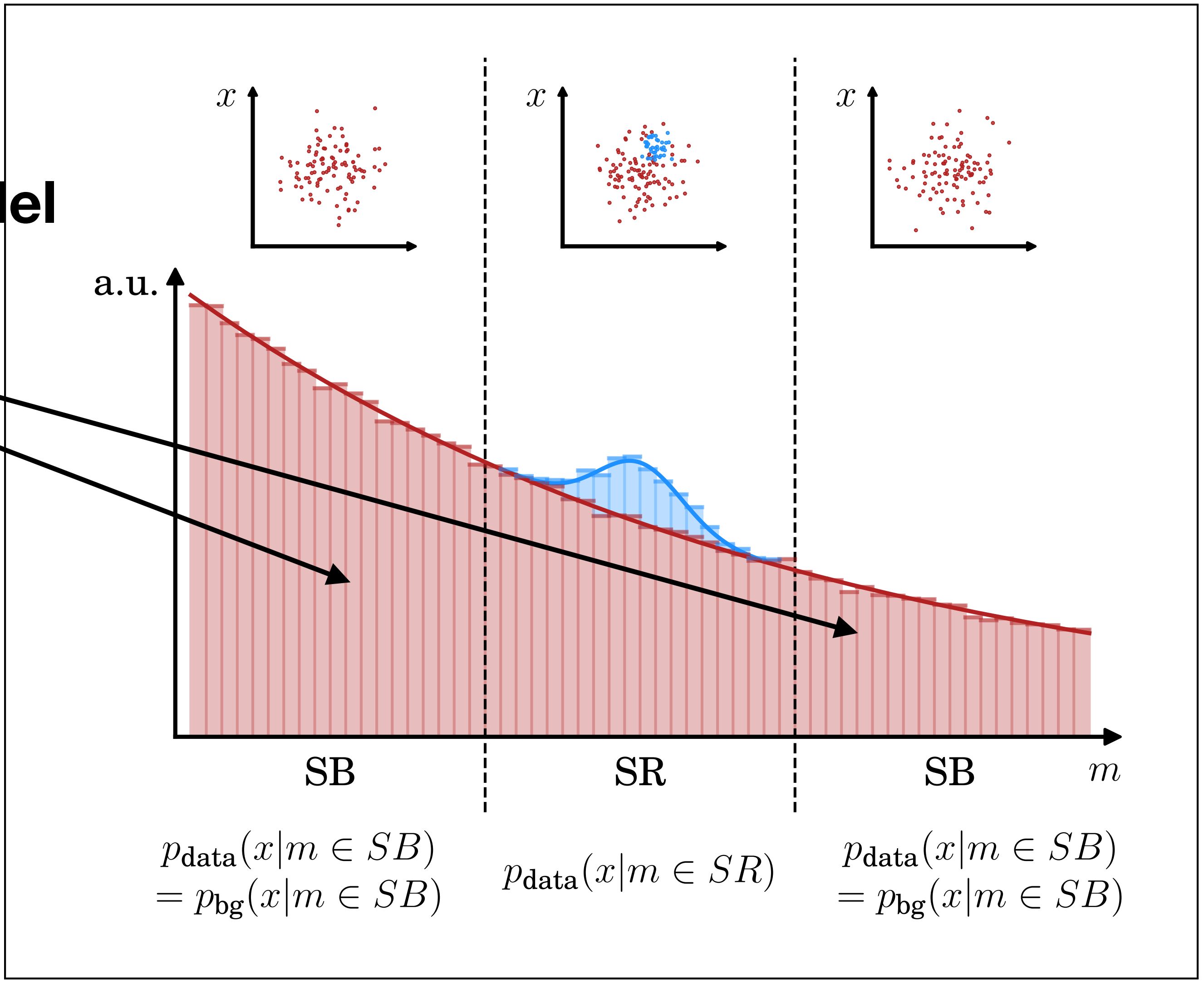


- 1): Choose one feature (m) in which to search for resonances

2109.00546

Generative models for anomaly detection

Train generative model here

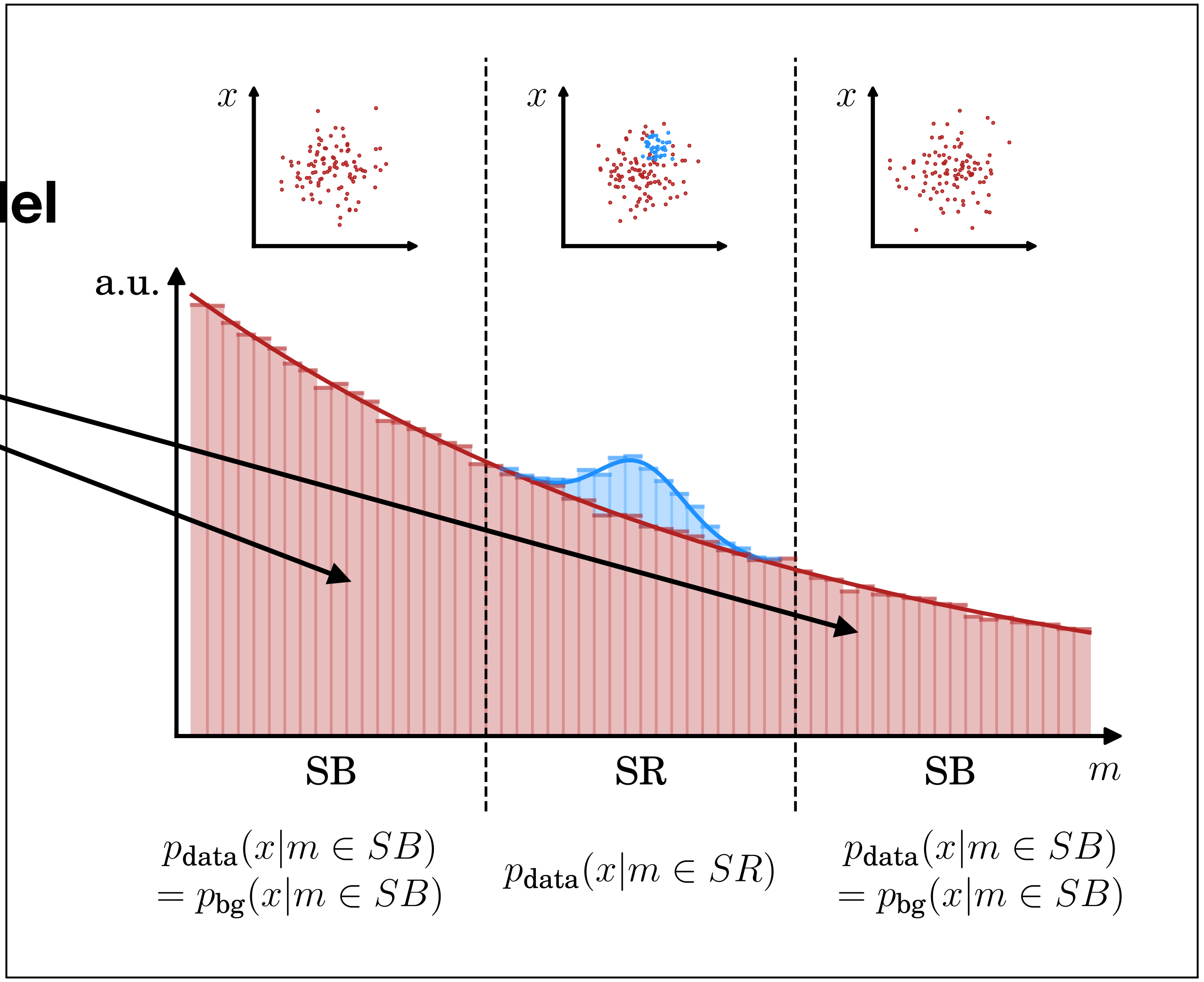


- 1): Choose one feature (m) in which to search for resonances
- 2): Use m divide spectrum into non-overlapping regions. Designate one as signal region (SR), others as sidebands (SB). Repeat the following for all choices of SR

2109.00546

Generative models for anomaly detection

Train generative model here

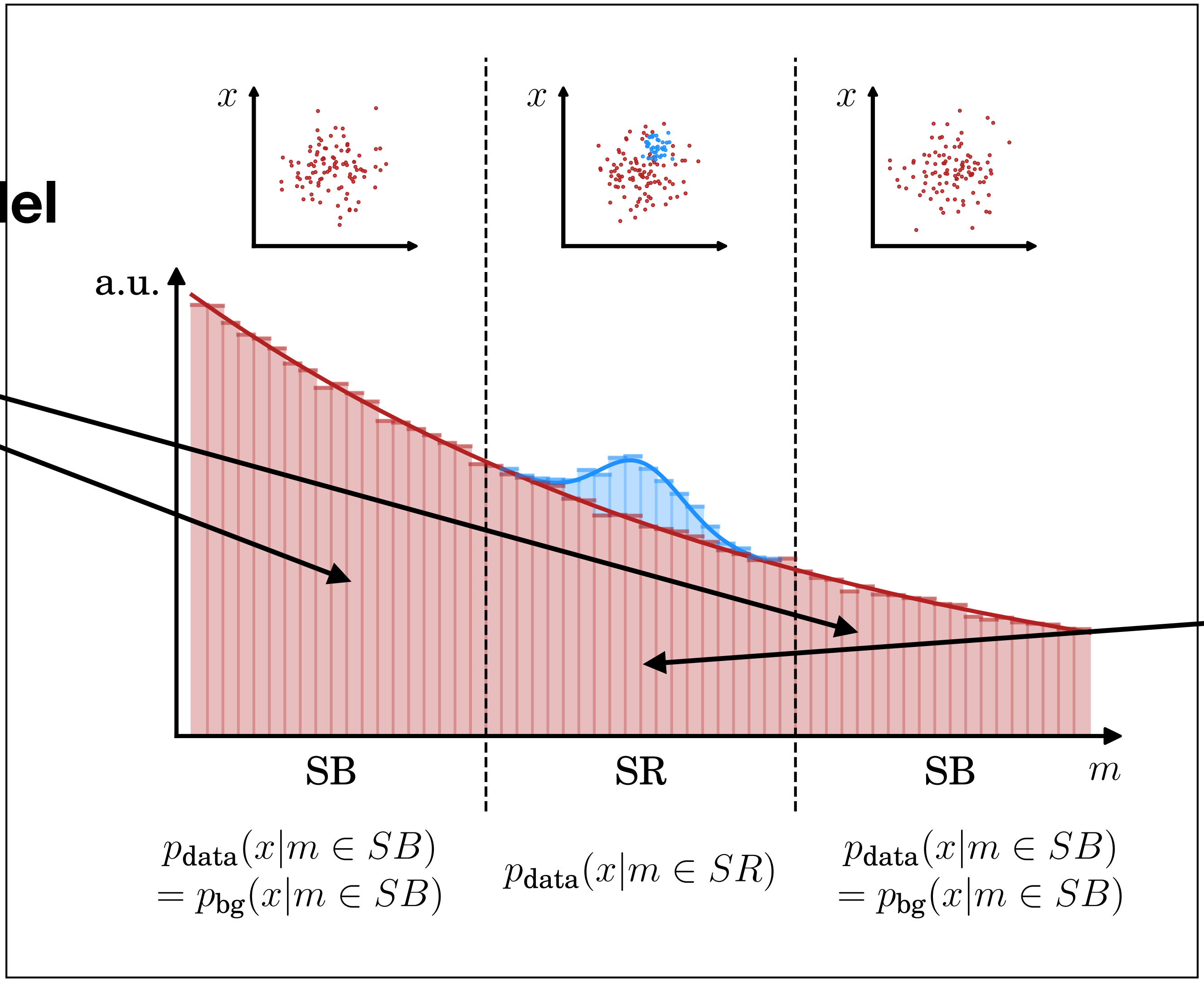


- 3) Train a generative model p(x|m) on auxiliary features in SB (used MAF, other choices including GAN/VAE possible as well)

2109.00546

Generative models for anomaly detection

Train generative model here

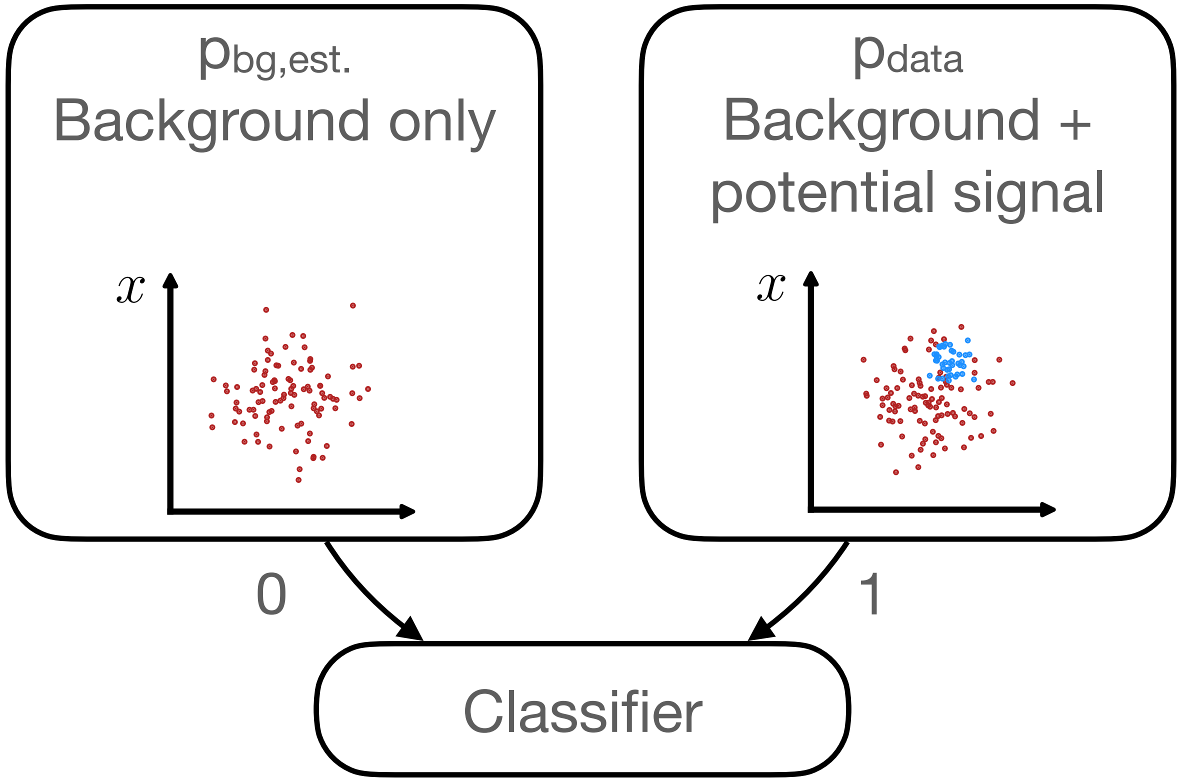


and sample here

- 3) Train a generative model p(x|m) on auxiliary features in SB
- 4) Sample from p(x|m) in SR. Designate as $p_{\text{bg,est}}$.

2109.00546

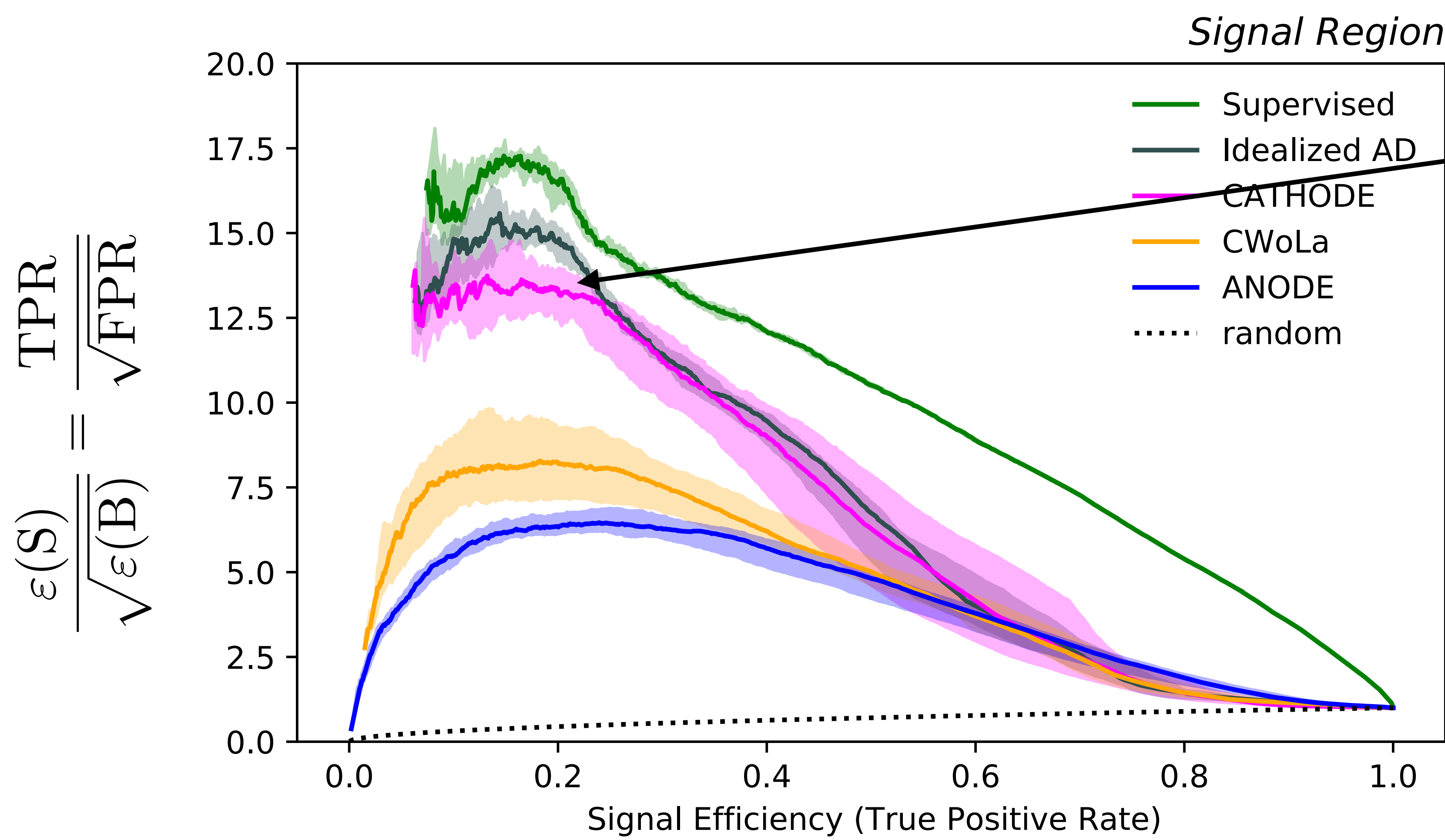
Generative models for anomaly detection



- 3) Train a generative model p(x|m) on auxiliary features in SB
- 4) Sample from p(x|m) in SR. Designate as $p_{bg,est}$
- 5) Train binary classifer between p_{data} and $p_{bg,est.}$

2109.00546

Generative models for anomaly detection



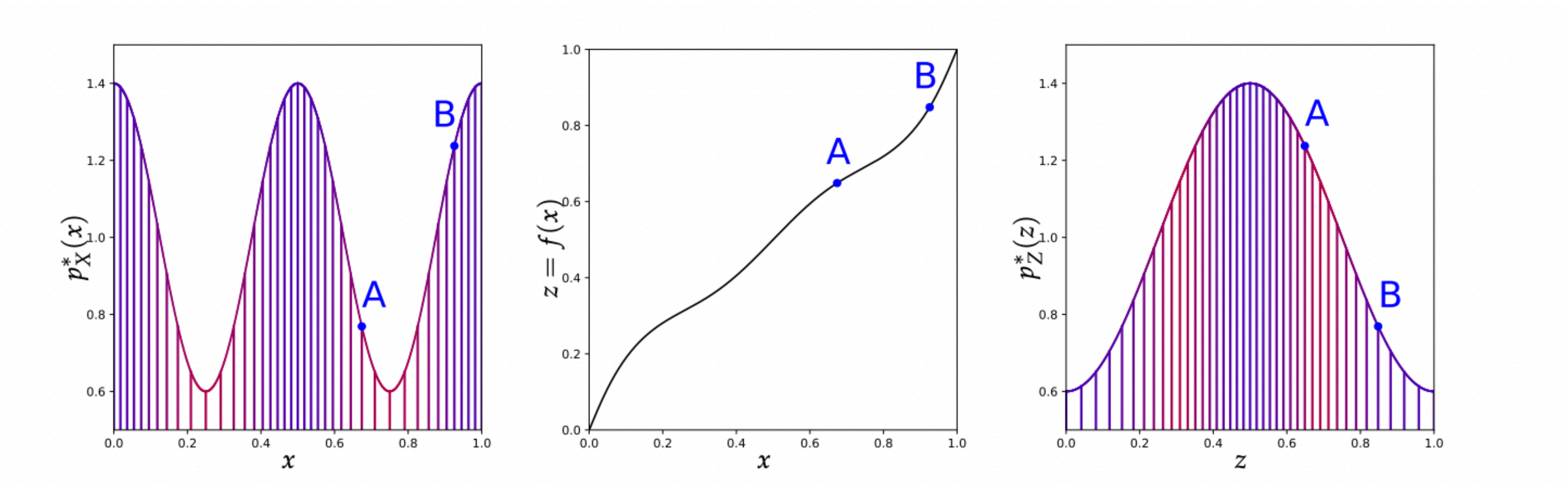
ours
CATHODE:
Classifying Anomalies THrough Outer Density Estimation

- 3) Train a generative model p(x|m) on auxiliary features in SB
- 4) Sample from p(x|m) in SR. Designate as $p_{bg,est.}$
- 5) Train binary classifer between p_{data} and $p_{bg,est.}$ (mixed sample classifer)
- 6) Cut on high classifier scores to enrich sample with anomalies
  (and perform statistical analysis)

2109.00546

Comments on anomaly detection

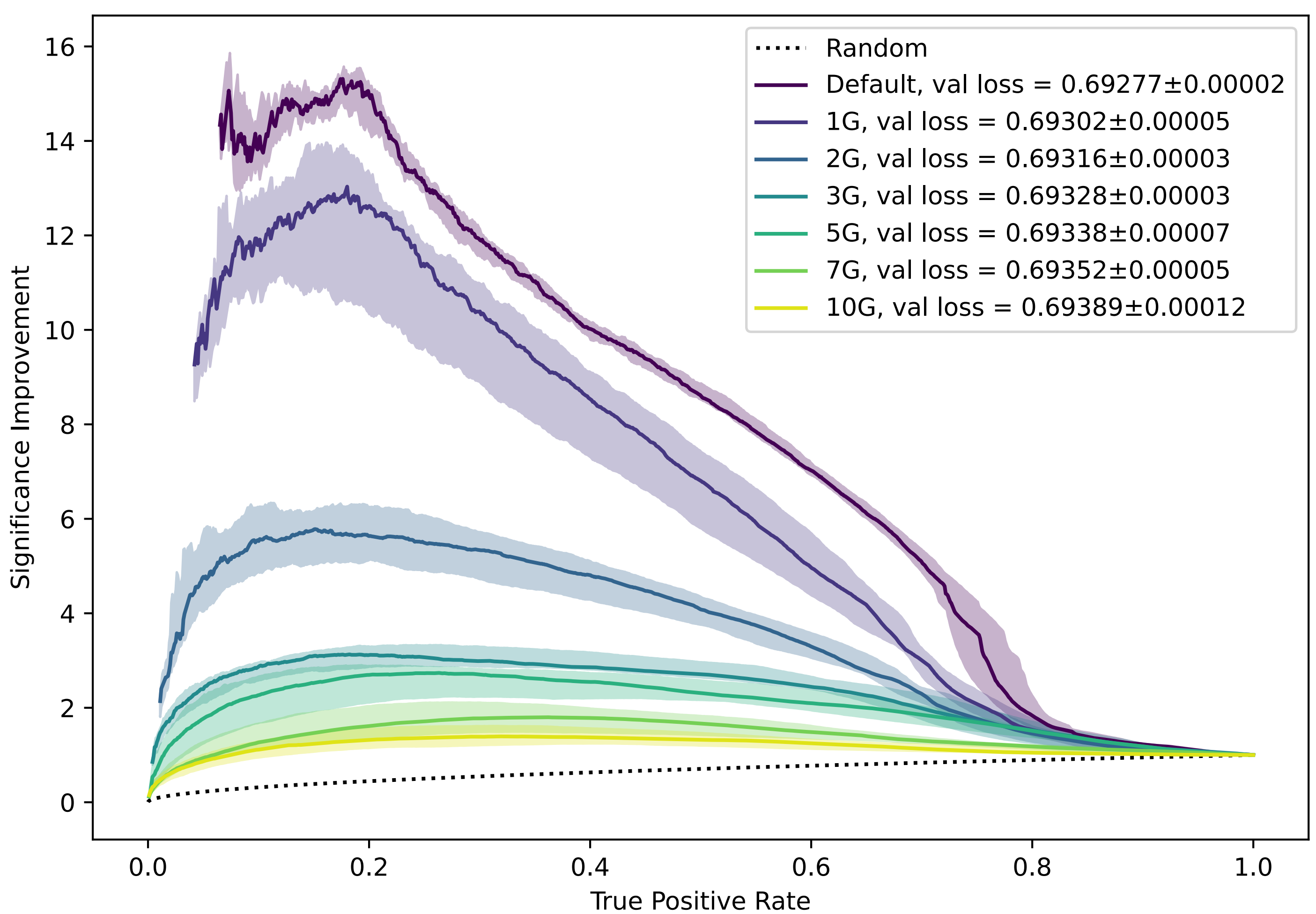
- As CATHODE approximates a likelihood ratio, it should be robust compared to methods that only use $p_{Background}$ (e.g. autoencoders)



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



MSc thesis work of P. Prangchaikul

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

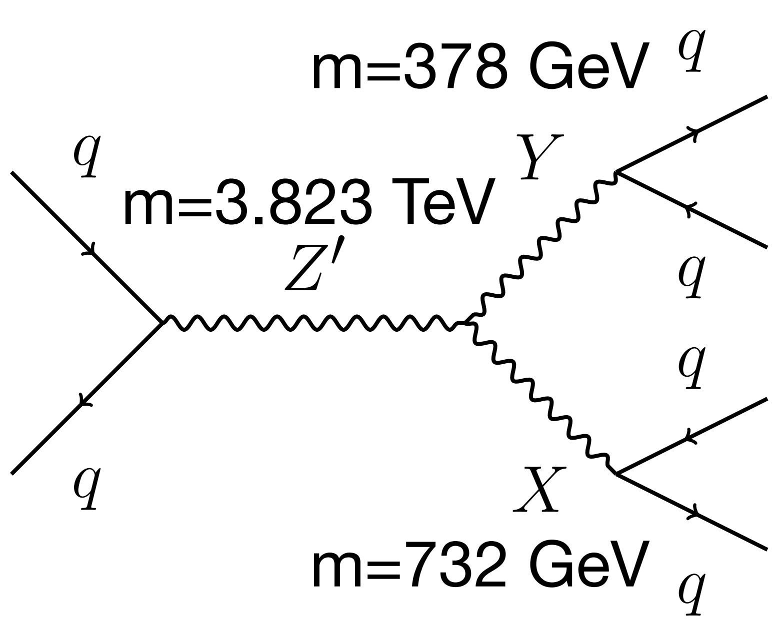
- Need also consider
  - Shaping of distributions under tigher anomaly detection cuts
  - Cost of signal-injection in training on data
  - How to efficiently estimate / compare / communicate sensitive regions of different anomaly detection algorithms
  - Make data-based anomaly detection more flexible

Challenge datasets

- All contain total of 1M examples; might contain signal; no labels provided during ‘content’ phase (labels available no)
- All used different simulation parameters for background (to avoid unrealistic exploits)

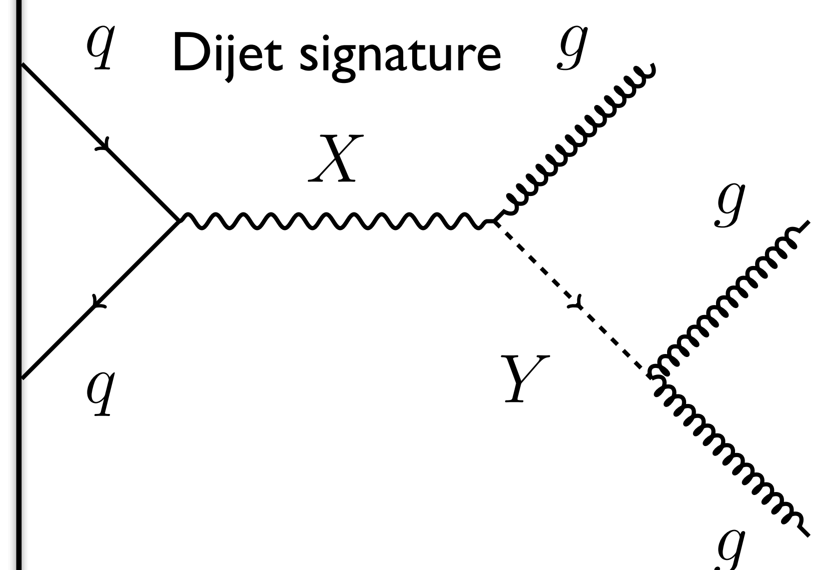
BB1: 834 signal examples
Same event topology as R&D dataset, different masses

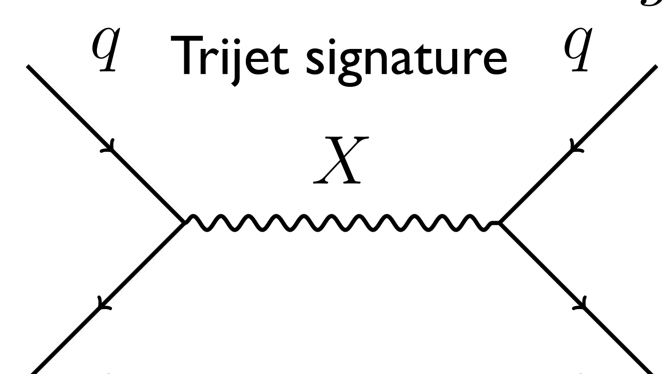
might be easy?



BB2: empty

BB3:





🐨 & Friends

- Situation seems better for density ratio based techniques (CWola, ANODE, CATHODE,..)

Per Neyman-Pearson: Likelihood-ratio is optimal test statistic
Unfortunatly, p(x|anomaly) is not available

$$L_{S/B} = \frac{p(x|\text{anomaly})}{p(x|\text{normal})}$$

Build data/background ratio:

$$L_{D/B} = \frac{p(x)}{p(x|\text{normal})}$$

Approximate background density using control measurement (e.g. sideband)

$$L_{D/B} \approx \frac{p(x)}{\tilde{p}(x|\text{normal})}$$

Expand $p(x) = f_{\text{normal}}\, p(x|\text{normal}) + f_{\text{anomaly}}\, p(x|\text{anomaly})$

And insert:

$$L_{D/B} \approx f_{\text{normal}} + f_{\text{anomaly}} \frac{p(x|\text{anomaly})}{\tilde{p}(x|\text{normal})}$$

- However…