AI SSAI
Anomaly Detection Workshop

WEIZMANN INSTITUTE OF SCIENCE

Anomaly Detection & Particle Physics

Traditional & ML-based efforts

Shikma Bressler I AISSAI, March 6, 2024

- The Challenge at hand
- Traditional (non-ML) efforts
- ML-based efforts
  - Supervised
  - Weakly supervised
  - Unsupervised
  - Others

The challenge at hand

- Huge number of potential topologies
  - In MC - 3527
- Huge number of combinations
  - 888022 simple combinatorics
  - 19497 when imposing boundary conditions
- Much fewer studies were carried out so far
- This is even before
  - Considering different distributions within each selection
  - Considering kinematic cut optimization

[2311.09012, Chekanov]

$$(N_{\not{E}_T},\ Nj,\ Nb,\ Ne,\ N\mu,\ N\tau,\ N\gamma)$$

$$N_{\not{E}_T} < 2, \quad Nj < 18, \quad Nb < 9, \quad Nj + Nb < 19,$$

$$Ne < 5, \quad N\mu < 5, \quad N\tau < 5, \quad N\gamma < 5,$$

$$Nj + N\ell < 19, \quad Nj + N\gamma < 19,$$

$$Nb + N\ell < 9, \quad Nb + N\gamma < 9,$$

$$N\ell < 6, \quad N\ell + N\gamma < 6,$$

$$Nj + Nb + N\ell + N\gamma < 21,$$

The challenge at hand

- Hundreds of searches have yielded no significant deviation from the SM prediction
  - We didn’t search in the right place for the right signature
  - Out of all models, we don’t know what is the right place to search in
  - Moreover, the one true model may haven’t been written yet
- Lack of resources to search for each and every individual signature
  - Impossible to cover all possible signatures with dedicated analyses

So…

- The potential of the data is far from being fully exhausted
- Clear need for complementary approaches → Anomaly Detection

Type of anomalies

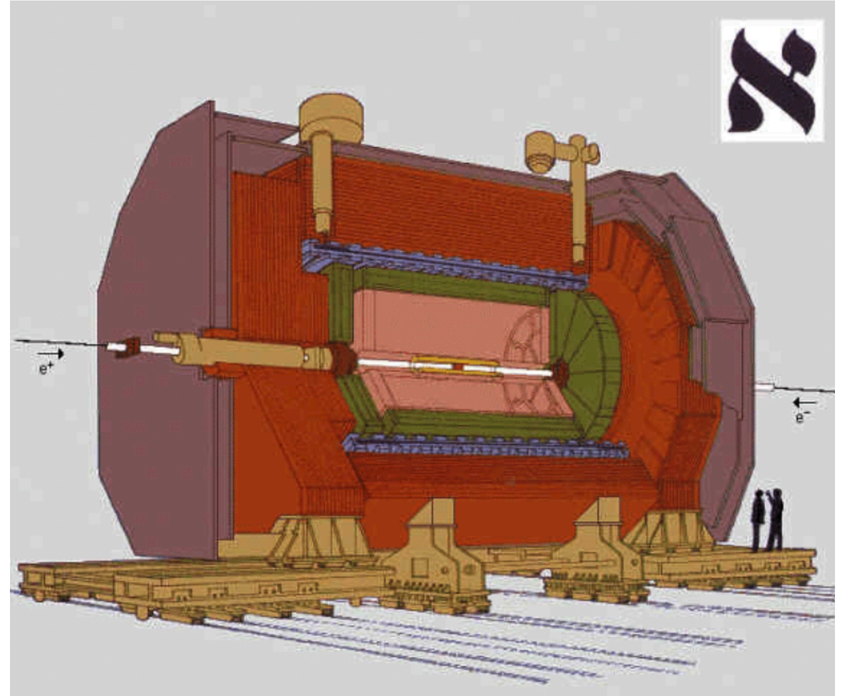
- Commonly discussed - Out-of-distribution [2312.14190, Belis et.al.]
  - Outlier detection - events that “should not be there”
  - Finding over-densities - e.g., bump hunting for new particles accessible by the LHC
- Unexpected differential cross section - e.g., new physics at above LHC energy scale (Shape of distribution)

Traditional (non ML) efforts

- The Challenge at hand
- **Traditional (non-ML) efforts**
- ML-based efforts
  - Supervised
  - Weakly supervised
  - Unsupervised
  - Others

MC/Data comparison

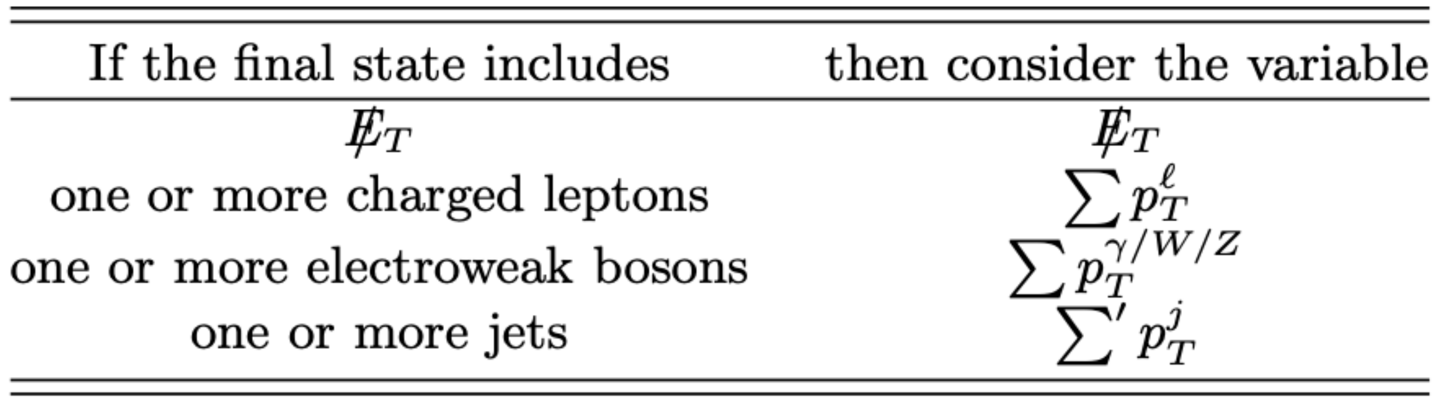
- Main idea - look at data/MC differences in large number of final states
  - Certain list of objects and object multiplicity
  - Certain list of parameters to look at, e.g., m_{inv}
  - Search algorithm
  - Treatment for look-elsewhere-effect
- Main limitations
  - MC mis-modeling
    - Systematic uncertainties
  - limited MC statistics



MC/Data comparison @ e.g., DØ

- $e\mu X$ final states [hep-ex/0006011]

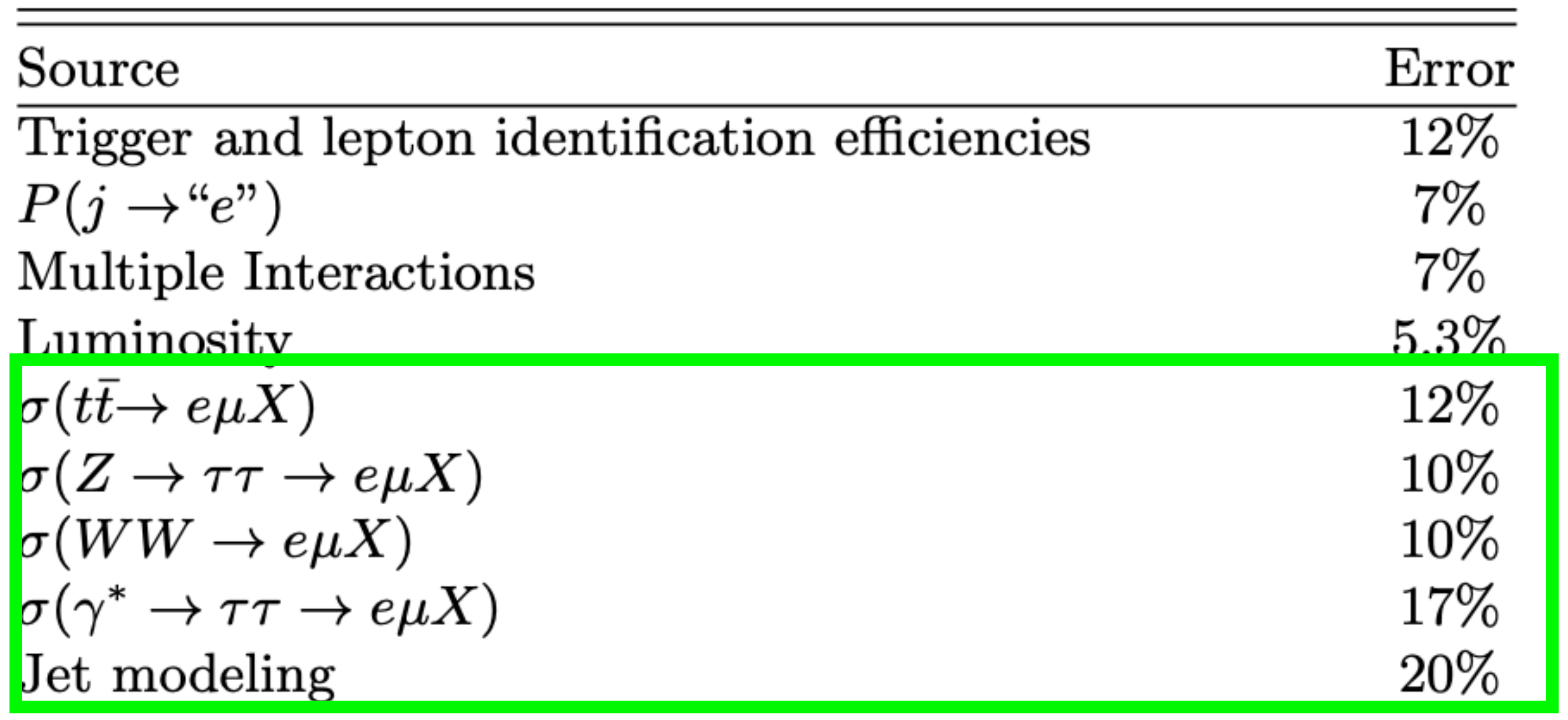
  $X \in \{\ell's, jets, \gamma's, \not\!\!E_T, W, Z\}$
- Variables of interest



If the final state includes	then consider the variable
$\not\!\!E_T$	$\not\!\!E_T$
one or more charged leptons	$\sum p_T^\ell$
one or more electroweak bosons	$\sum p_T^{\gamma/W/Z}$
one or more jets	$\sum' p_T^j$

TABLE I. A quasi-model-independently motivated list of interesting variables for any final state. The set of variables to consider for any particular final state is the union of the variables in the second column for each row that pertains to that final state. Here ℓ denotes e, μ, or τ. The notation $\sum' p_T^j$ is shorthand for $p_T^{j_1}$ if the final state contains only one jet, $\sum_{i=2}^n p_T^{j_i}$ if the final state contains $n \geq 2$ jets, and $\sum_{i=3}^n p_T^{j_i}$ if the final state contains n jets and nothing else, with $n \geq 3$. Leptons and missing transverse energy that are reconstructed as decay products of W or Z bosons are not considered separately in the left-hand column.

- Search employing the SLEUTH algorithm
  - Based on definition of regions within the parameter space
- Uncertainties



Source	Error
Trigger and lepton identification efficiencies	12%
$P(j \rightarrow$"e"$)$	7%
Multiple Interactions	7%
Luminosity	5.3%
$\sigma(t\bar{t} \rightarrow e\mu X)$	12%
$\sigma(Z \rightarrow \tau\tau \rightarrow e\mu X)$	10%
$\sigma(WW \rightarrow e\mu X)$	10%
$\sigma(\gamma^* \rightarrow \tau\tau \rightarrow e\mu X)$	17%
Jet modeling	20%

TABLE V. Sources of systematic uncertainty on the number of expected background events in the final states $e\mu\not\!\!E_T$, $e\mu\not\!\!E_T j$, $e\mu\not\!\!E_T jj$, and $e\mu\not\!\!E_T jjj$. $P(j \rightarrow$"e"$)$ denotes the probability that a jet will be reconstructed as an electron. "Jet modeling" includes systematic uncertainties in jet production in PYTHIA and HERWIG in addition to jet identification and energy scale uncertainties.

MC/Data comparison @ e.g., CMS MUSIC

- Objects considered [2010.02984, CMS]

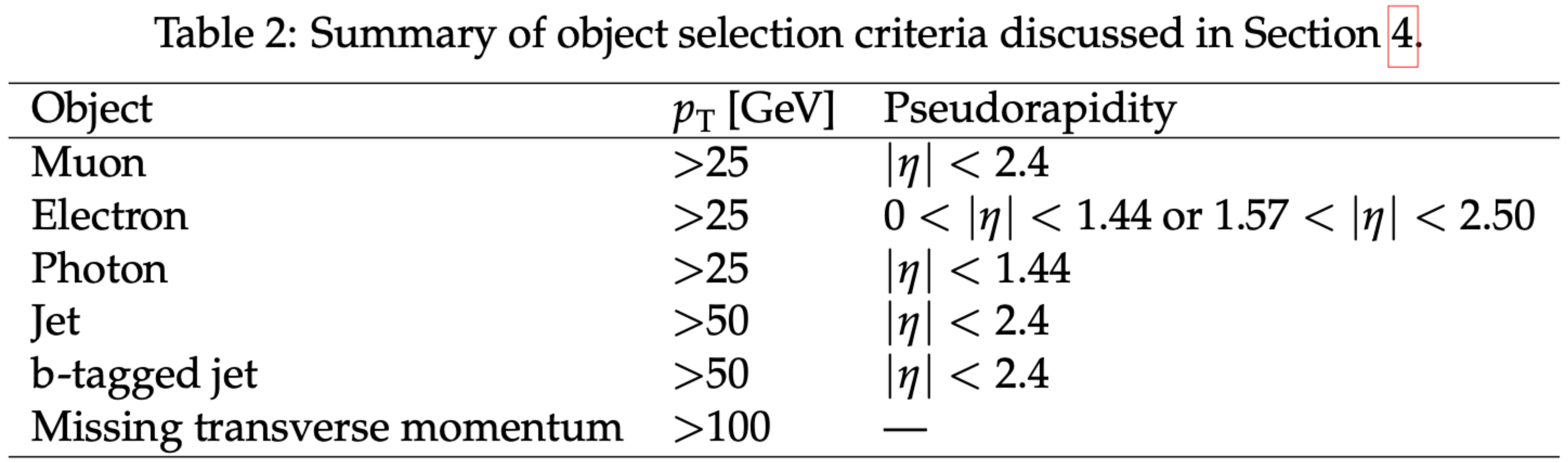
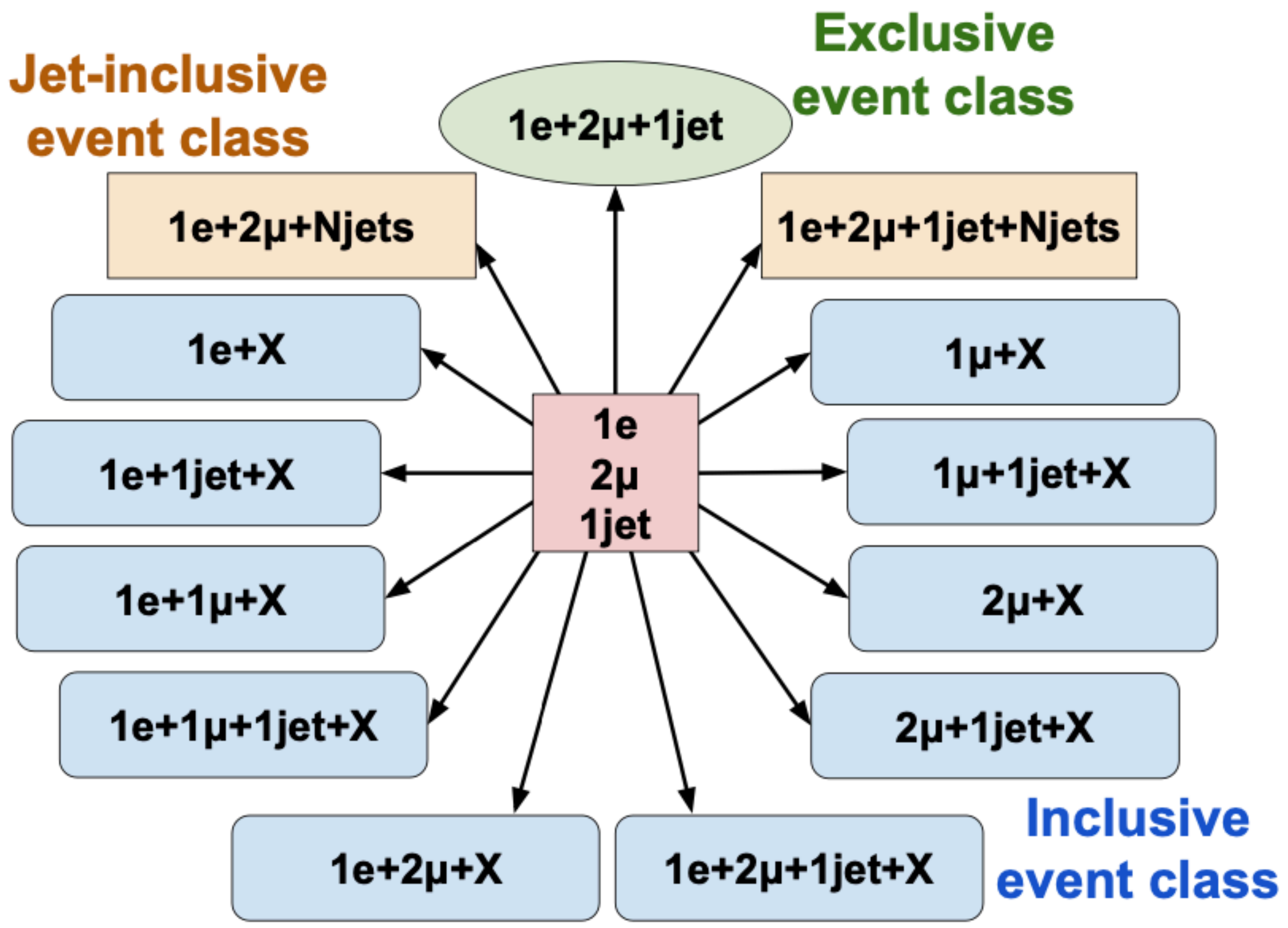


Table 2: Summary of object selection criteria discussed in Section 4.

Object	p_T [GeV]	Pseudorapidity
Muon	>25	$\lvert\eta\rvert < 2.4$
Electron	>25	$0 < \lvert\eta\rvert < 1.44$ or $1.57 < \lvert\eta\rvert < 2.50$
Photon	>25	$\lvert\eta\rvert < 1.44$
Jet	>50	$\lvert\eta\rvert < 2.4$
b-tagged jet	>50	$\lvert\eta\rvert < 2.4$
Missing transverse momentum	>100	—

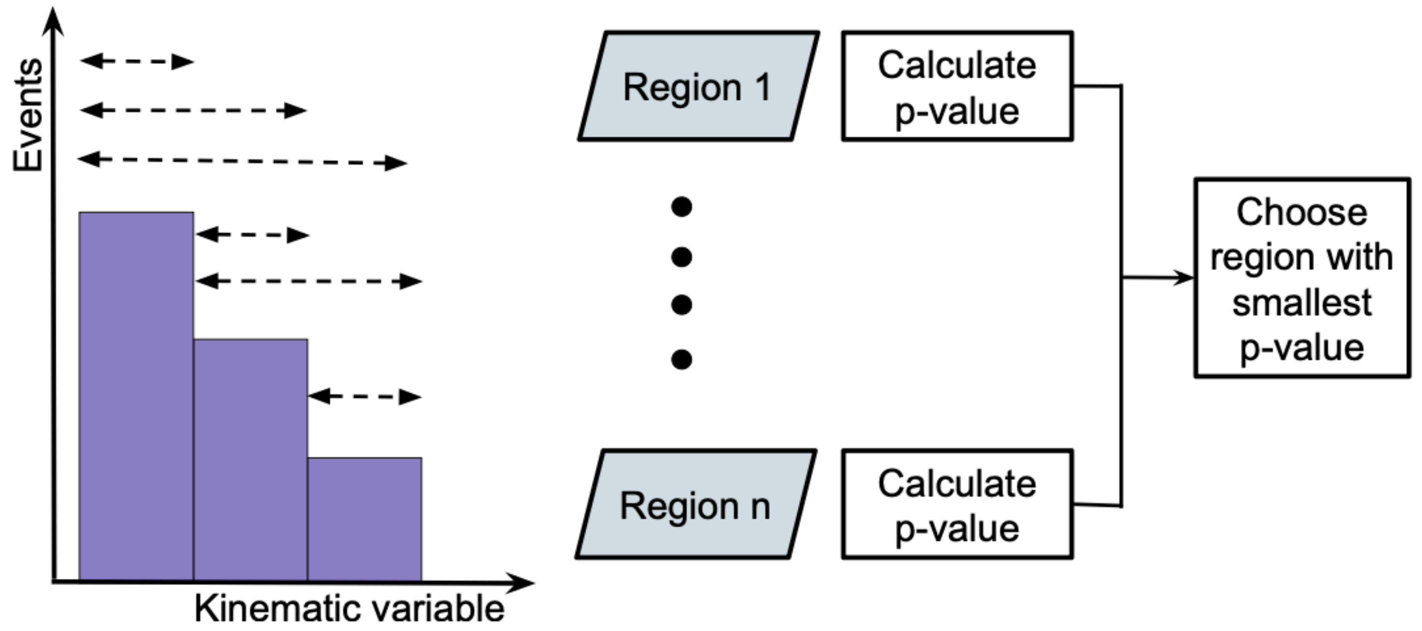
- Final state categorization
  - Exclusive, inclusive, jet-inclusive



- Parameters of interest

$$S_T = \sum |\vec{p}_{T,i}| \qquad M \text{ or } M_T \qquad p_T^{\text{miss}};$$

- Search algorithm



- Results

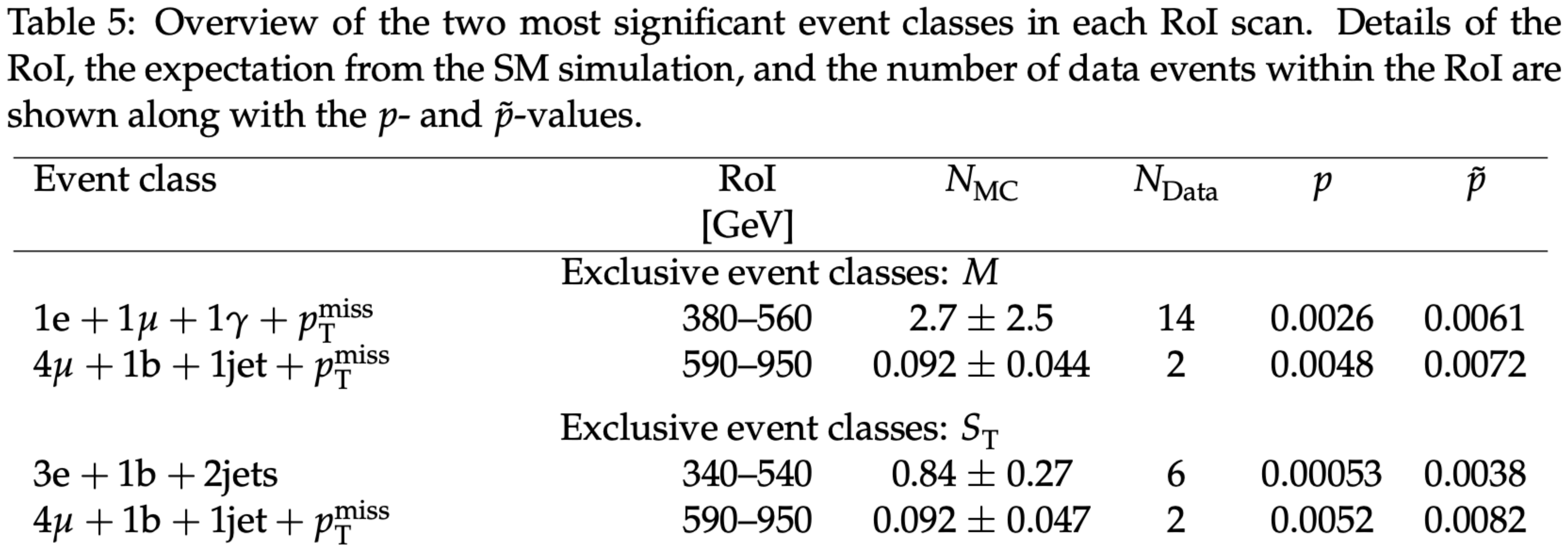


Table 5: Overview of the two most significant event classes in each RoI scan. Details of the RoI, the expectation from the SM simulation, and the number of data events within the RoI are shown along with the p- and $\tilde{p}$-values.

Event class	RoI [GeV]	N_{MC}	N_{Data}	p	$\tilde{p}$
Exclusive event classes: M					
$1e + 1\mu + 1\gamma + p_T^{\text{miss}}$	380–560	2.7 ± 2.5	14	0.0026	0.0061
$4\mu + 1b + 1\text{jet} + p_T^{\text{miss}}$	590–950	0.092 ± 0.044	2	0.0048	0.0072
Exclusive event classes: S_T					
$3e + 1b + 2\text{jets}$	340–540	0.84 ± 0.27	6	0.00053	0.0038
$4\mu + 1b + 1\text{jet} + p_T^{\text{miss}}$	590–950	0.092 ± 0.047	2	0.0052	0.0082

BumpHunter

[1101.0390, Choudalakis]

- Finding the largest deviating region in the data from a predefined background distribution
  - No assumption is made on the signal shape
  - Background shape is known

    → Test statistic pdf under $\mathscr{H}_0$ is known
  - Taking into account the look-elsewhere-effect

BumpHunter

- Recent advances implemented in pyBumpHunter [2208.14760, Vaslin et.al.]
  - Using Poisson statistics
  - Look-Elsewhere-Effect evaluated with pseudo experiments
  - Solution for 2D distributions is given

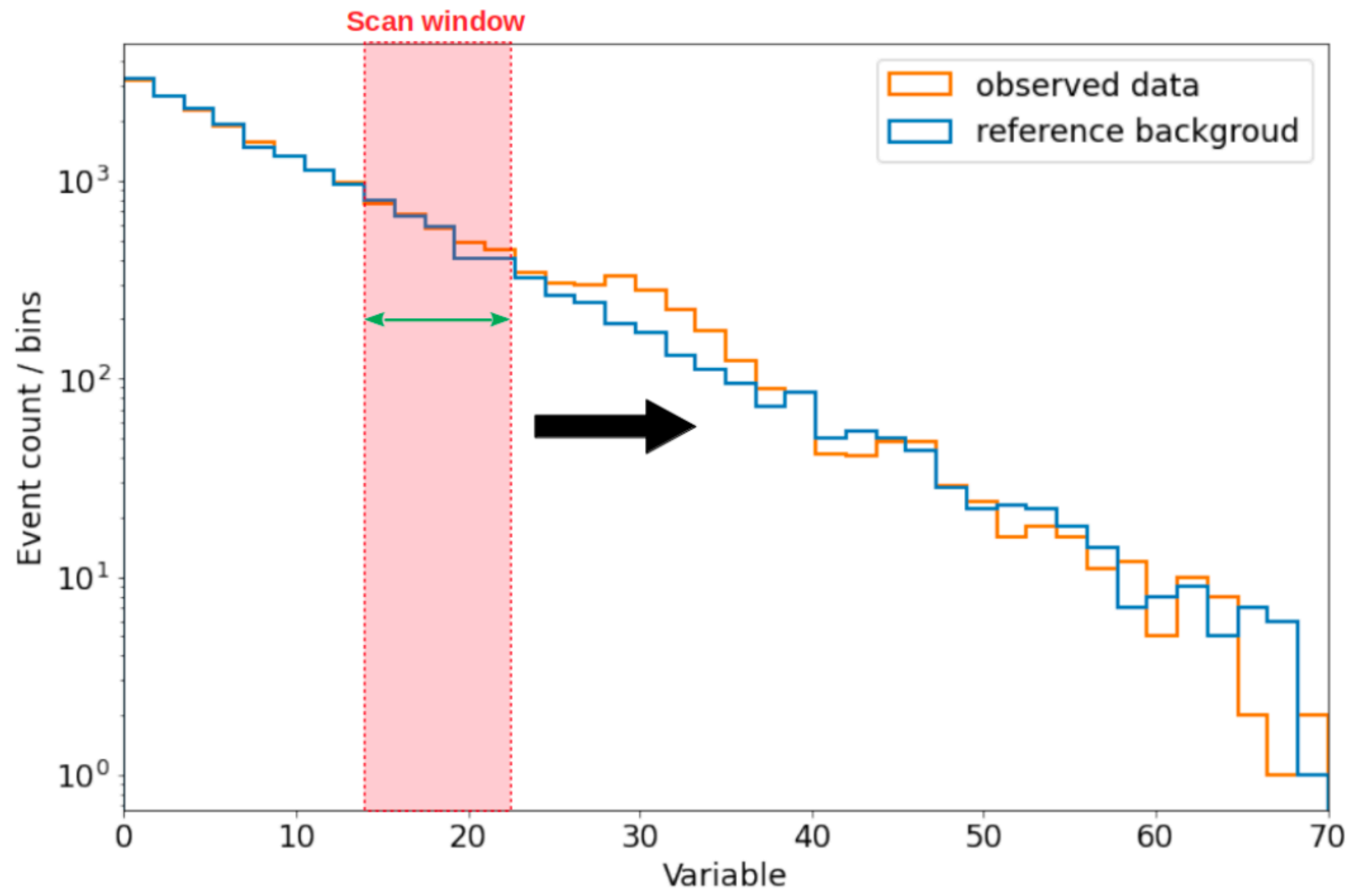


Figure 1: Scanning procedure performed by the BumpHunter algorithm. The red rectangle shows the interval that is currently being analyzed and the black arrow represents the motion of the scan window over the histogram range. In this example, the scan width is 5 bins and the scanned distributions have 40 bins.

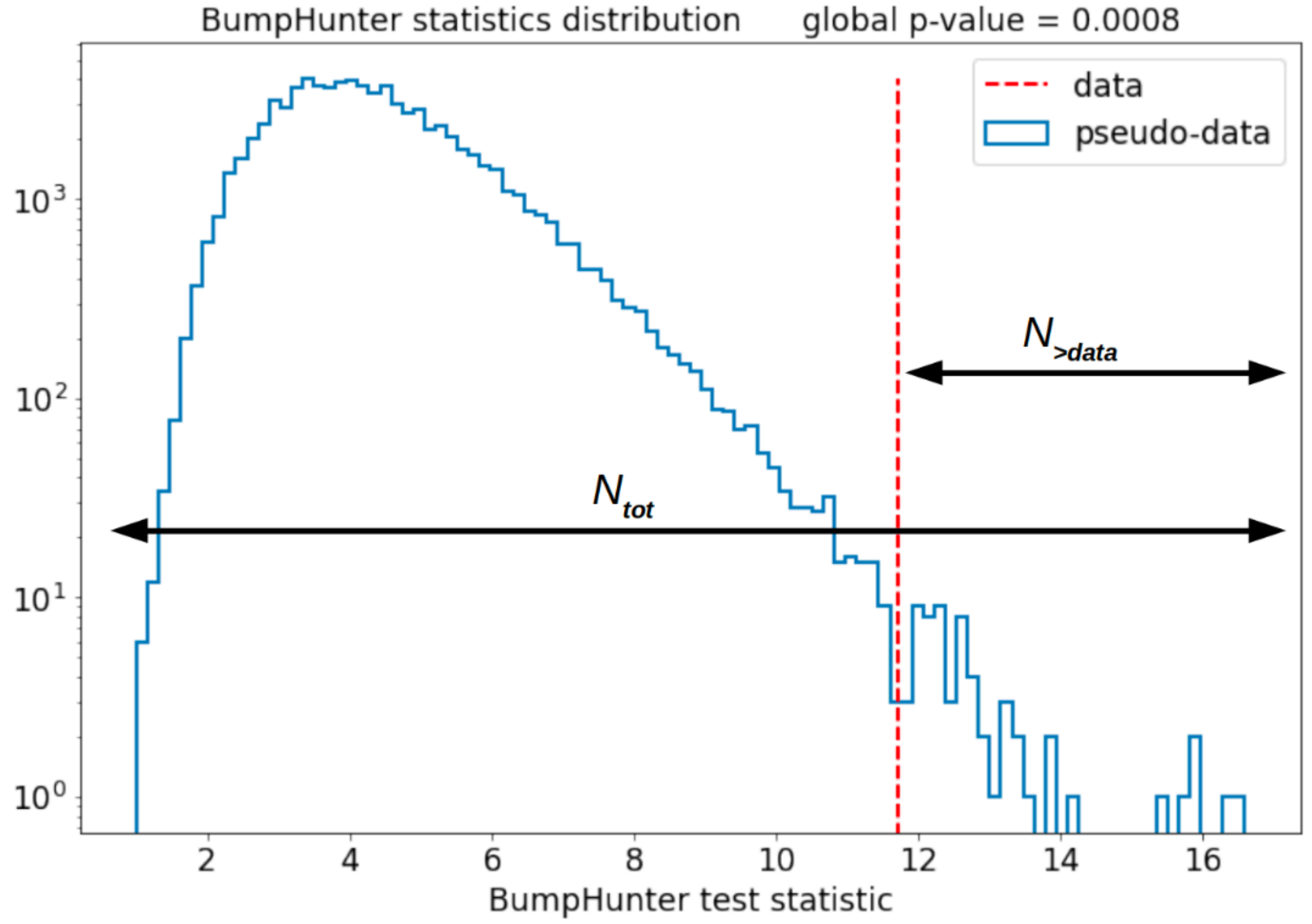


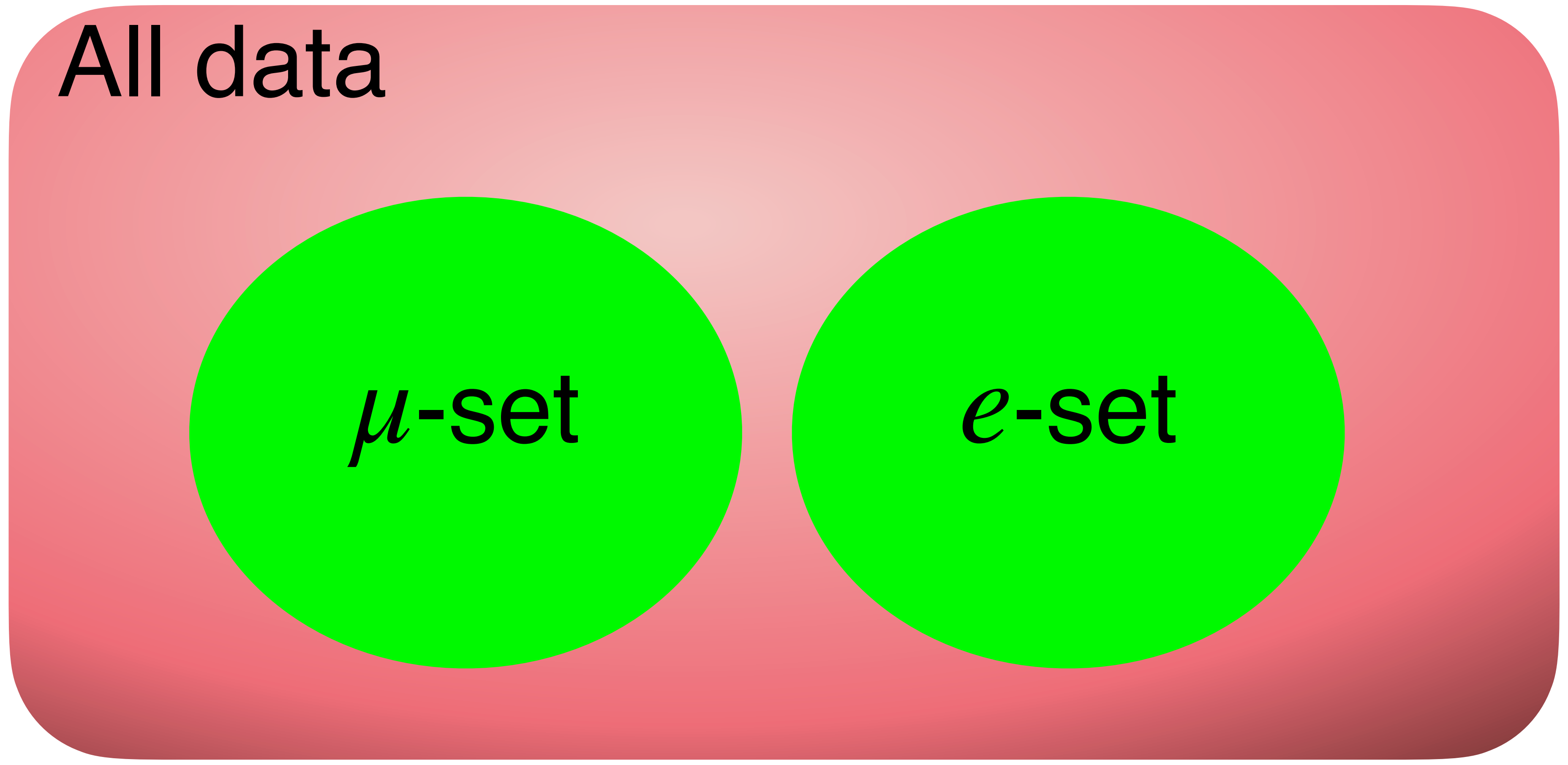
Figure 3: Distribution of BumpHunter test statistic value obtained for the pseudo-data (blue). The dashed red line correspond to the value of test statistic associated to observed data. The arrows illustrate how the global p-value is computed in equation 5. The global p-value is 0.0008, corresponding to 3.16σ.

Search for Asymmetries

- The SM has known and well tested symmetries [2203.07529, Birman et.al.]
  - Lepton universality, Lepton flavour, CP, etc.
  - Corrections are known and can be accounted for
- Splitting the data into 2 supposedly symmetric datasets
- Search in a model agnostic way for breaking of these symmetries

Search for Asymmetries

- The SM has known and well tested symmetries [2203.07529, Birman et.al.]
  - Lepton universality, Lepton flavour, CP, etc.
  - Corrections are known and can be accounted for
- Splitting the data into 2 supposedly symmetric datasets
- Search in a model agnostic way for breaking of these symmetries



- Flavour symmetric → μ-set and e-set originate from the same pdf
- Flavour asymmetric → μ-set and e-set originate from different pdfs

Search for Asymmetries

- The SM has known and well tested symmetries
  - Lepton universality, Lepton flavour, CP, etc.
  - Corrections are known and can be accounted for
- Splitting the data into 2 supposedly symmetric datasets
- Search in a model agnostic way for breaking of these symmetries
- Several possible implementations
  - N_σ test between 2 matrices $N_\sigma(B,A) = \frac{1}{\sqrt{M}}\sum_{i=1}^{M}\frac{B_i - A_i}{\sqrt{\sigma_{Ai}^2 + \sigma_{Bi}^2}}$.
  - Compare performance with traditional profile likelihood test statistics
  - Can also be treated with BumpHunter

[2203.07529, Birman et.al.]

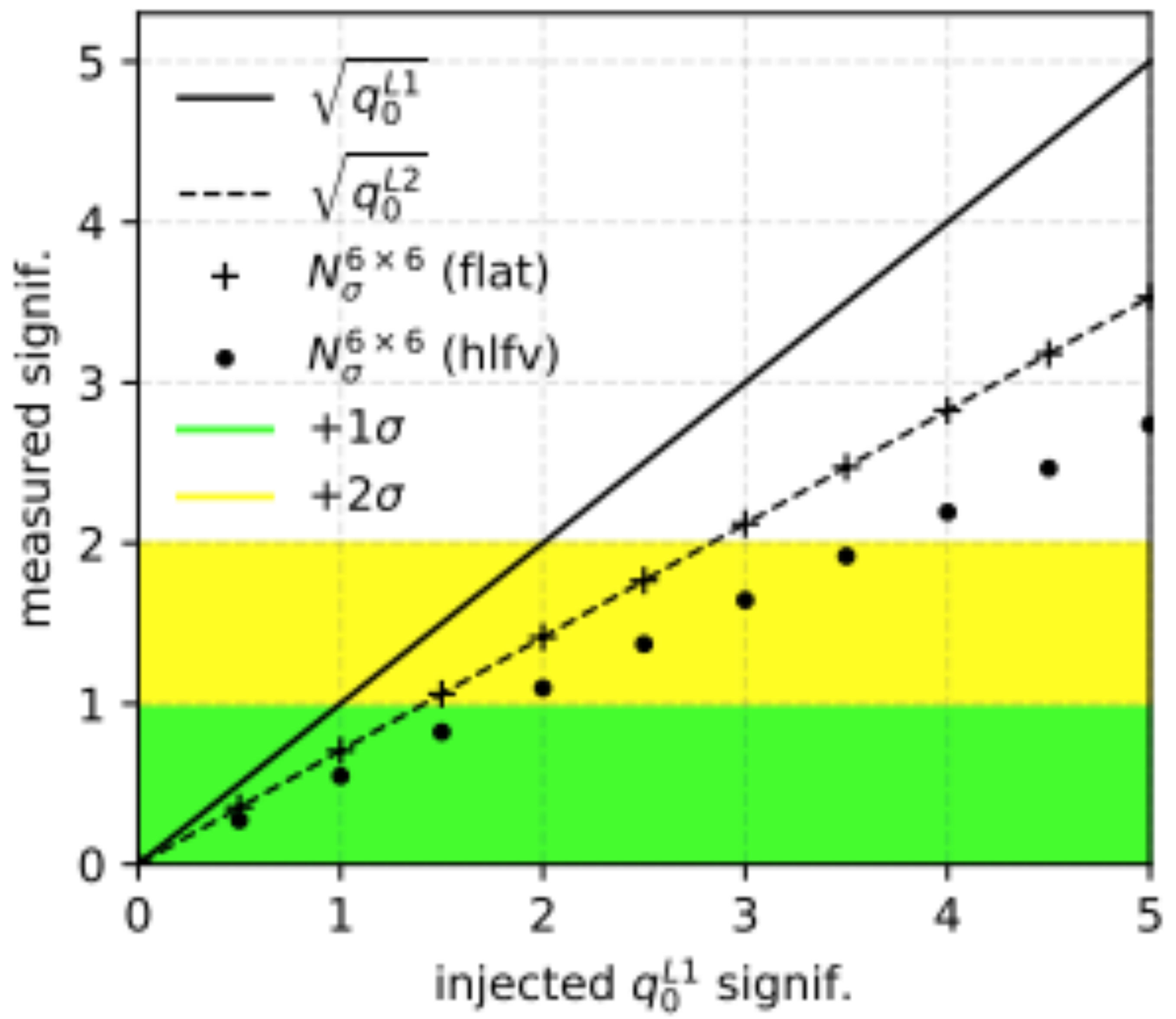


Fig. 5: Significance measured from the Asimov data for increasing injected signal, comparing results of the N_σ, q_0^{L1} and q_0^{L2} tests. Results for the Higgs LFV example and the ideal (flat) scenario are shown. The green and yellow bands correspond to the 1σ and 2σ deviations from the symmetry (no signal) assumption, respectively.

ML based efforts

- The Challenge at hand
- Traditional (non-ML) efforts
- **ML-based efforts**
    - Supervised
    - Weakly supervised
    - Unsupervised
    - Others

ML based efforts

- Based on the following reviews
  - “Machine learning for anomaly detection in particle physics”, Belis et. al.
  - “Machine Learning in the Search for New Fundamental Physics”, Karagiorgi et. al.
  - “The LHC Olympics 2020”, Kasieczka et. al.
  - “The Dark Machines Anomaly Score Challenge” et. al.
  - …and many more. Some can be found in https://iml-wg.github.io/HEPML-LivingReview/

ML based efforts

- ML schematic

Input

ML architecture

Target

- Analysis schematic

Signal enhancement

Background modeling

Statistical inference

ML based efforts

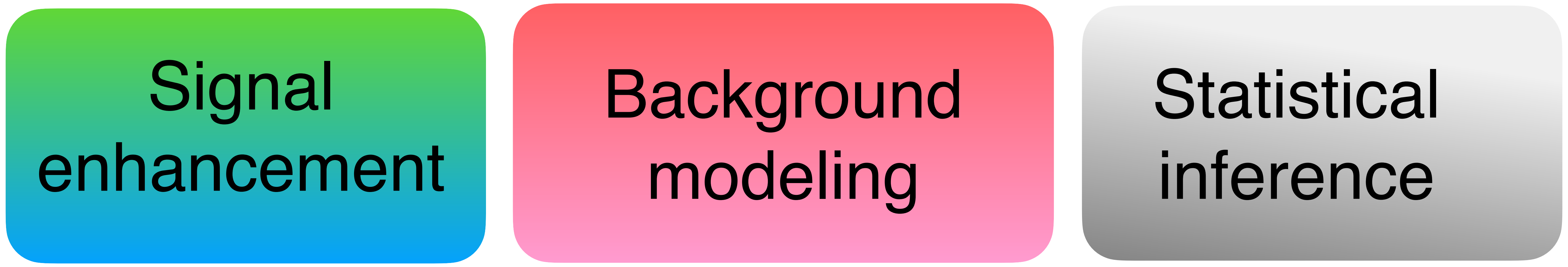
- ML schematic

Input

ML architecture

Target

- Analysis schematic



Signal enhancement

Background modeling

Statistical inference

ML techniques can be used in each stage of the analysis

ML based efforts

- ML schematic

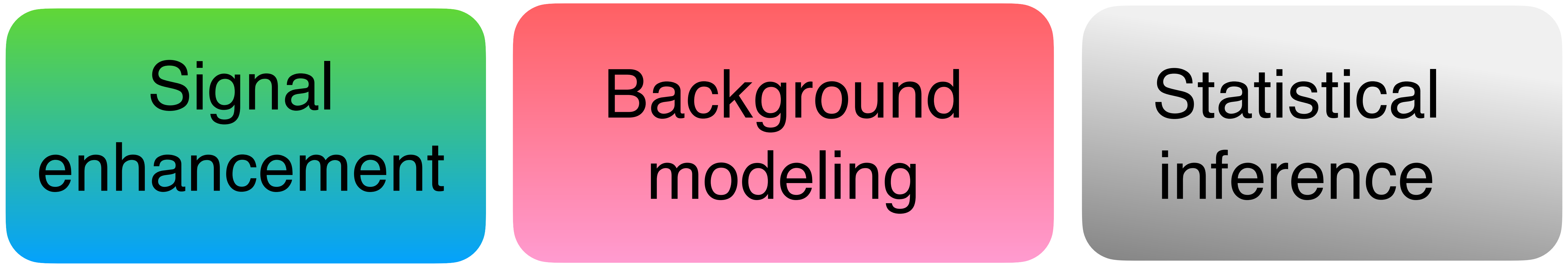
Input

ML architecture

Target

Novel techniques can be employed in each part of the ML

- Analysis schematic



Signal enhancement

Background modeling

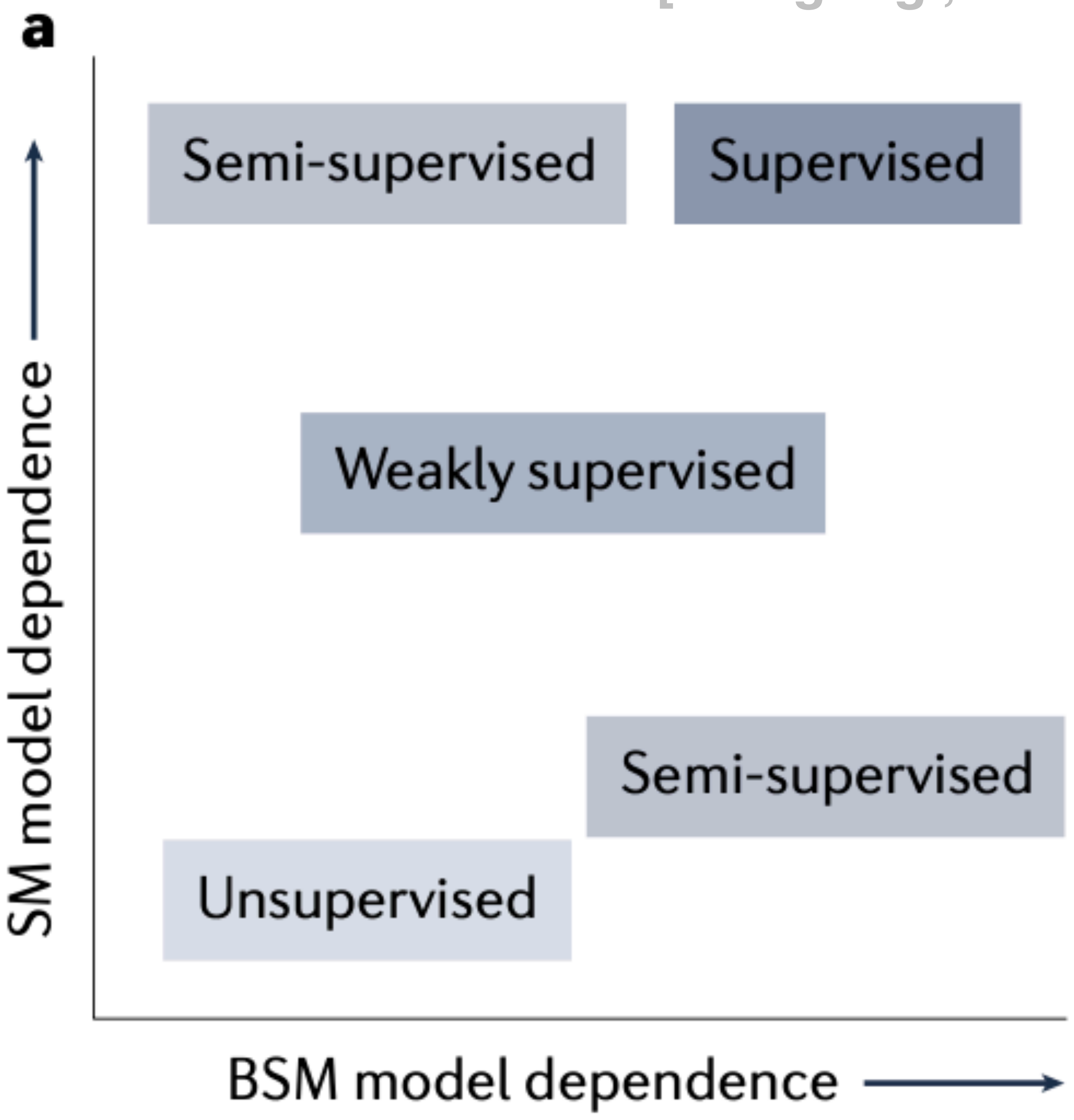
Statistical inference

Targets

- *Supervised* - use simulation for the signal and the SM background → MC labelled by construction
- *Semi-supervised* - use data for either the background or the signal-sensitive sample → data unlabelled by construction
- *Weakly supervised* - have labels for every example, but the labels are noisy
- *Unsupervised* - do not use any label information

- The choice of approach depends on the level of prior knowledge one wishes to assume
  - Greater knowledge better sensitivity
  - Greater knowledge more model dependency

[Karagiorgi, et.al.]



Type of anomalies

- Commonly discussed - Out-of-distribution [2312.14190, Belis et.al.]
  - Outlier detection - events that should no be there
  - Finding over-densities - e.g., bump hunting

- Unexpected differential cross section (Shape of distribution)

Supervised efforts

- The Challenge at hand
- Traditional (non-ML) efforts
- ML-based efforts
  - **Supervised**
  - Weakly supervised
  - Unsupervised
  - Others

Supervised efforts

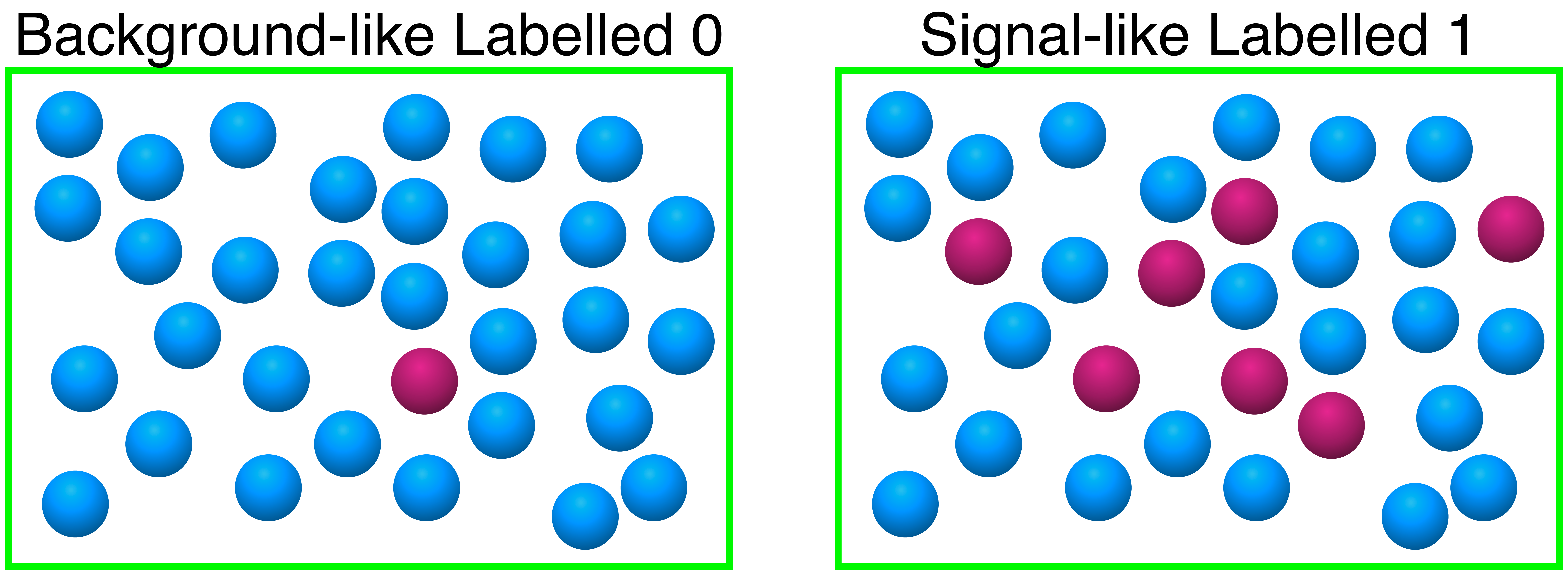
- To the best of my knowledge, AD with supervised ML is not addressed in literature but..
- One can think of training, e.g., a classifier to distinguish between
  - Background events - simulated or other
  - Signal events - simulated from a branch of models or a single model but over a brand range of the parameter space
- This is not a true-full Anomaly Detection process, but it has the potential of covering many possible signatures and topologies and reasonable sensitivity

Weakly supervised efforts

- The Challenge at hand
- Traditional (non-ML) efforts
- ML-based efforts
  - Supervised
  - **Weakly supervised**
  - Unsupervised
  - Others

Weakly supervised efforts

Background event ●
Signal event ●

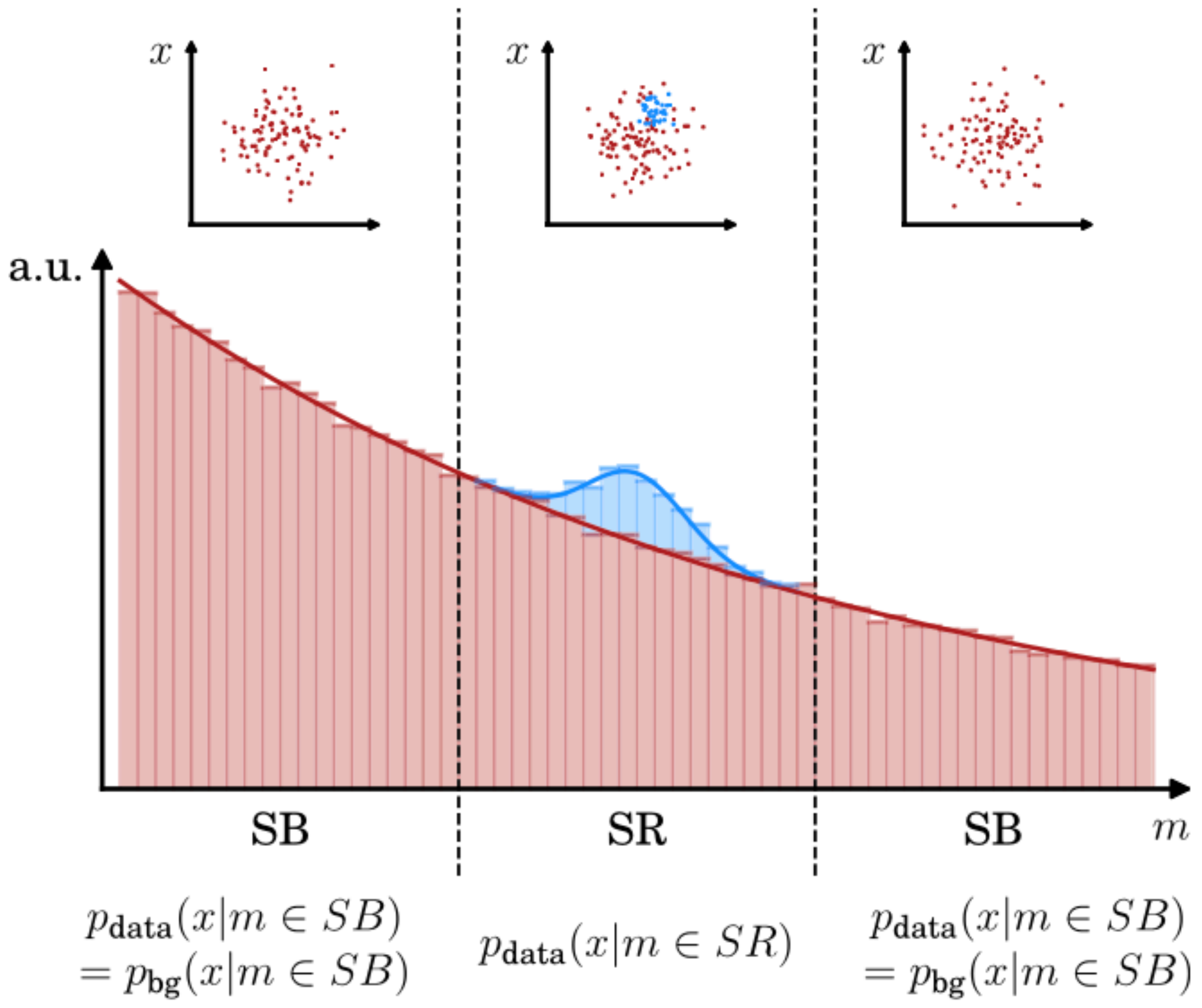


Background-like Labelled 0

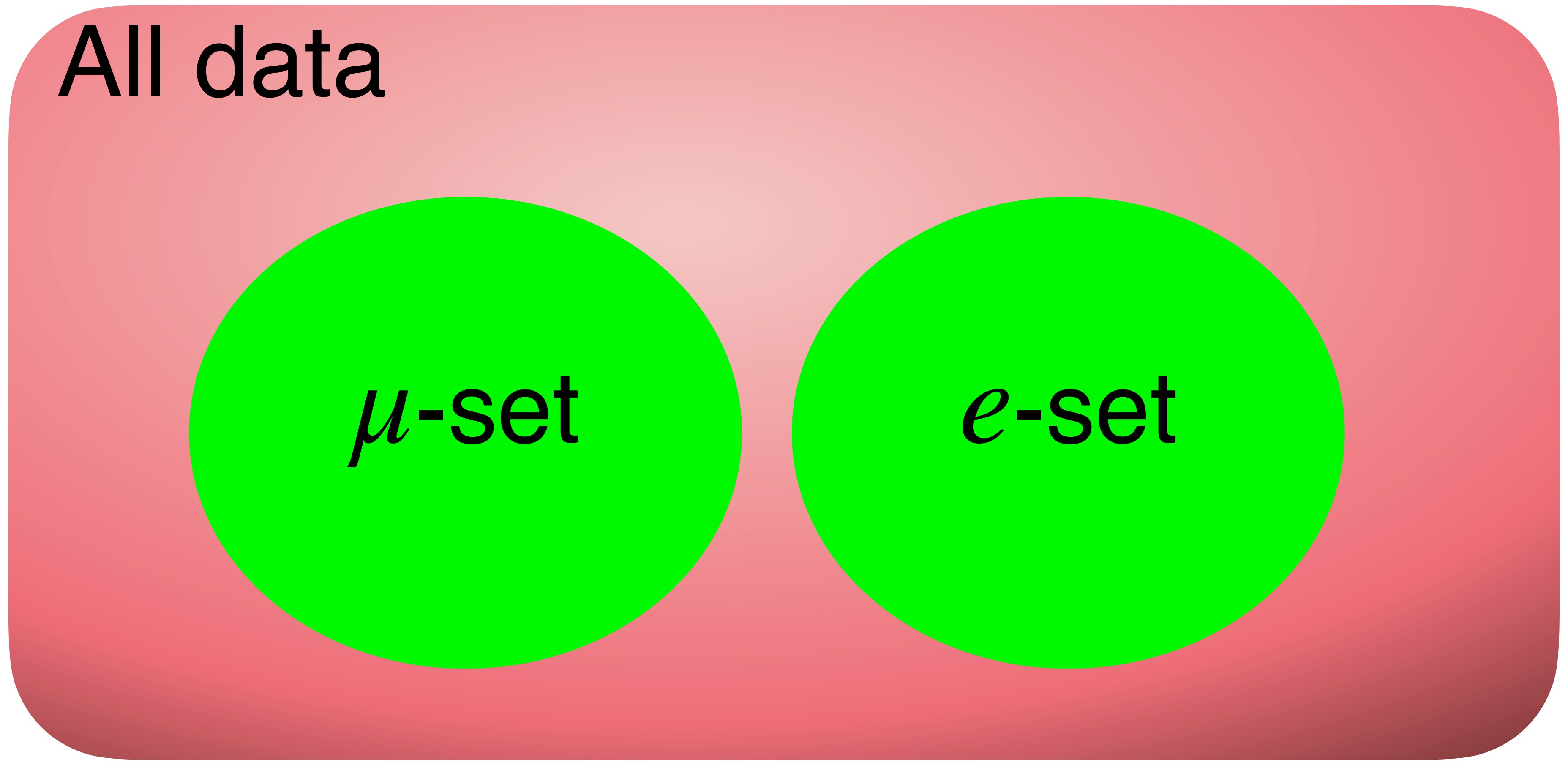
Signal-like Labelled 1

Noisy datasets

Sidebands



Symmetries



All data

μ-set

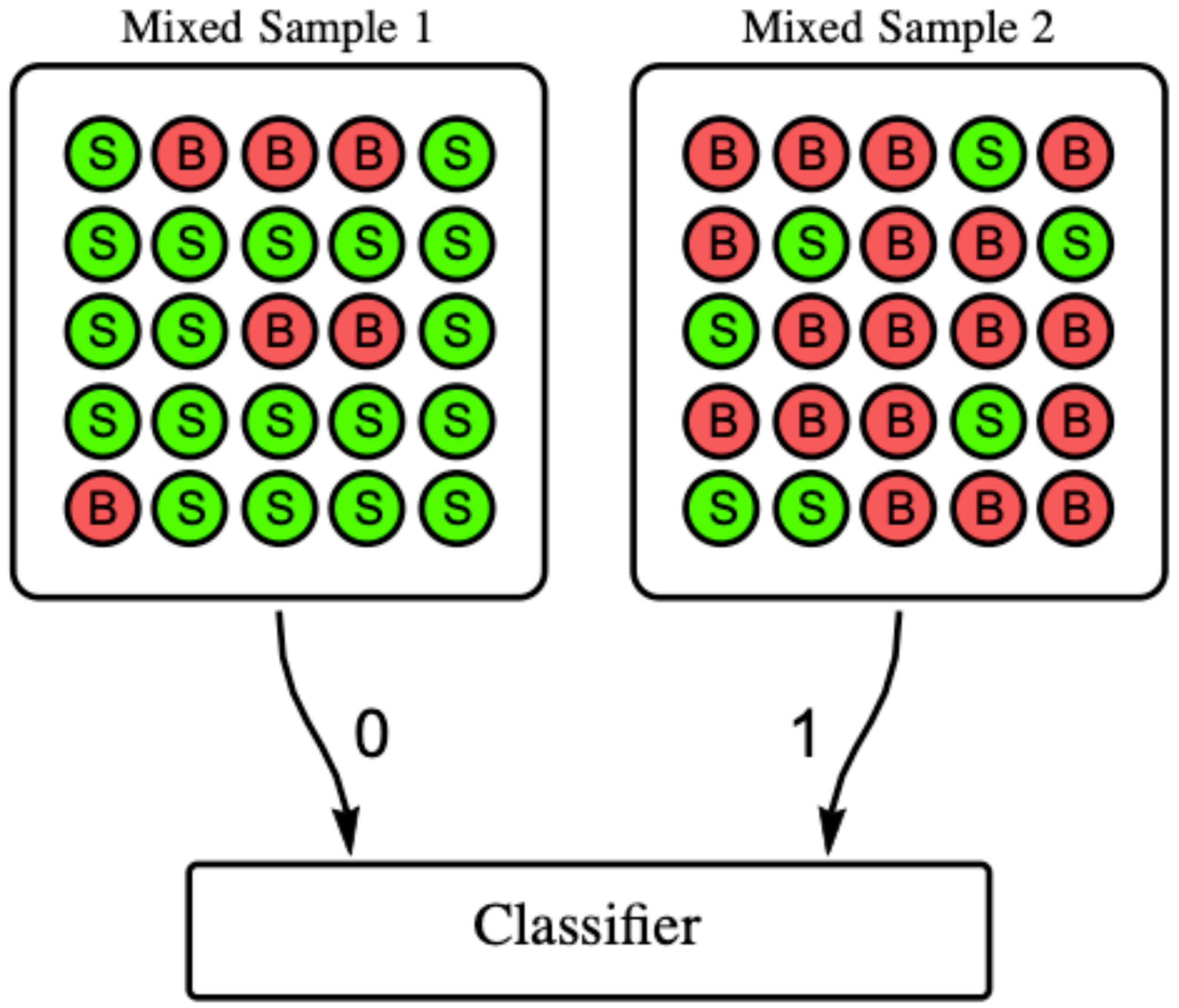
e-set

CWola - Classification Without Labels

ML used to enhance S/B

[1708.02949, Metodiev et.al.]

- An optimal classifier (likelihood ratio) trained to distinguish two mixed samples M_1 and M_2 is also optimal for distinguishing S from B
  - For large enough datasets
  - As long as the relative S and B propositions in M_1 and M_2, f_1 and f_2, are different
- For $f_1 > f_2$ event classified as M_1 is also classified as S event



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

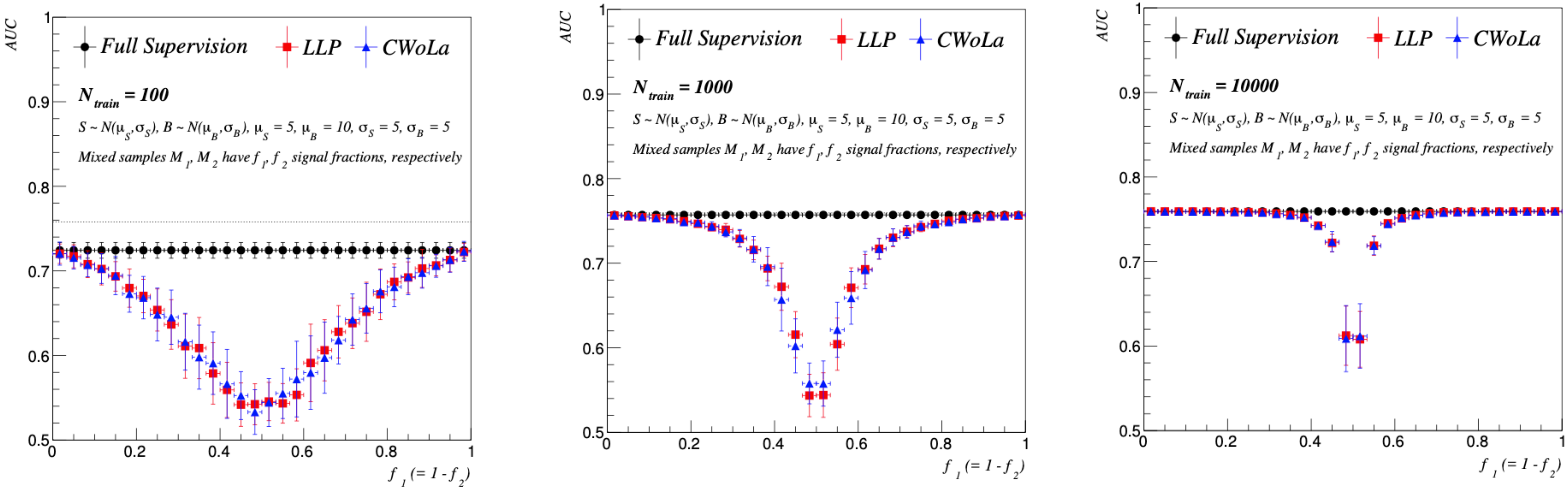
CWola - Classification Without Labels

[1708.02949, Metodiev et.al.]

- Separation with 2 Gaussian examples
  - Approaches Likelihood ratio with large enough dataset

$$h_{\text{optimal}}(x) = \frac{p_S(x)}{p_B(x)}.$$

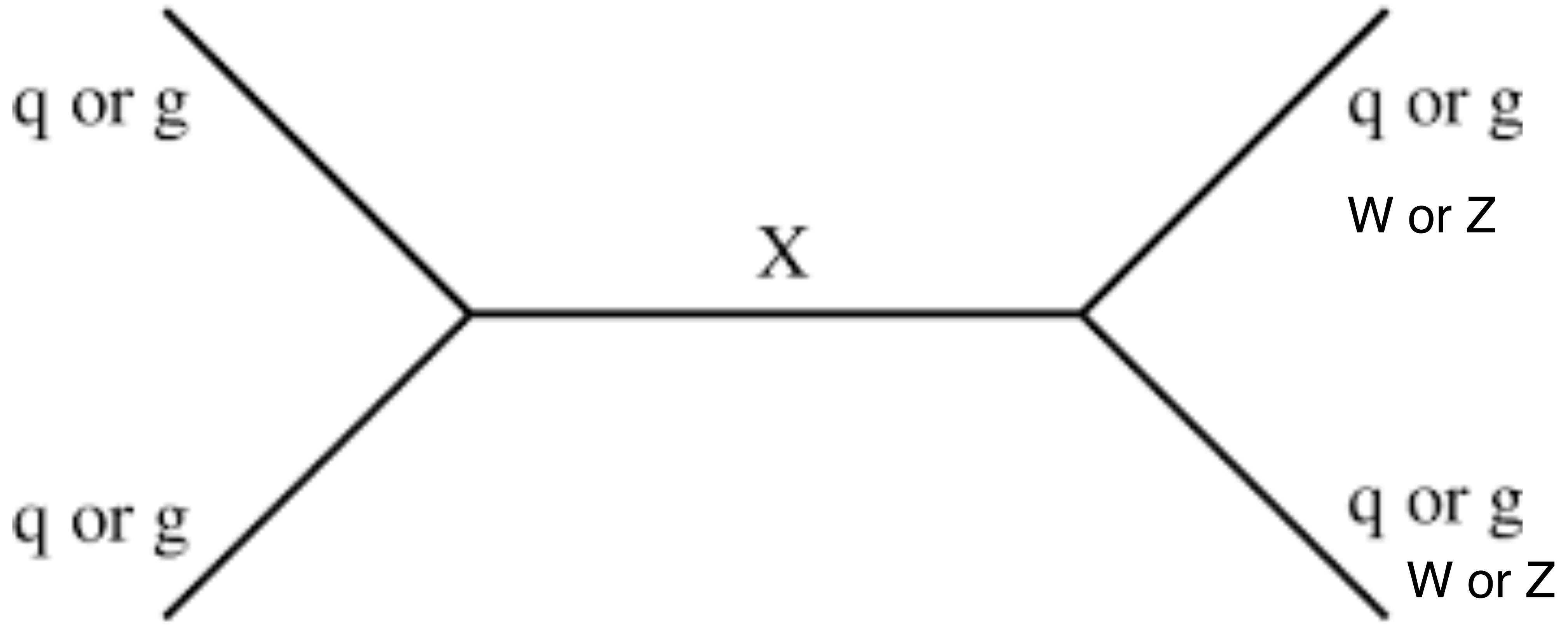
$$h_{\text{full}}(x) = \frac{\sum_{y\in\mathcal{S}} \mathbb{I}[y=x]}{\sum_{y\in\mathcal{B}} \mathbb{I}[y=x]}. \qquad h_{\text{CWoLa}}(x) = \frac{\sum_{y\in\mathcal{M}_1} \mathbb{I}[y=x]}{\sum_{y\in\mathcal{M}_2} \mathbb{I}[y=x]}$$

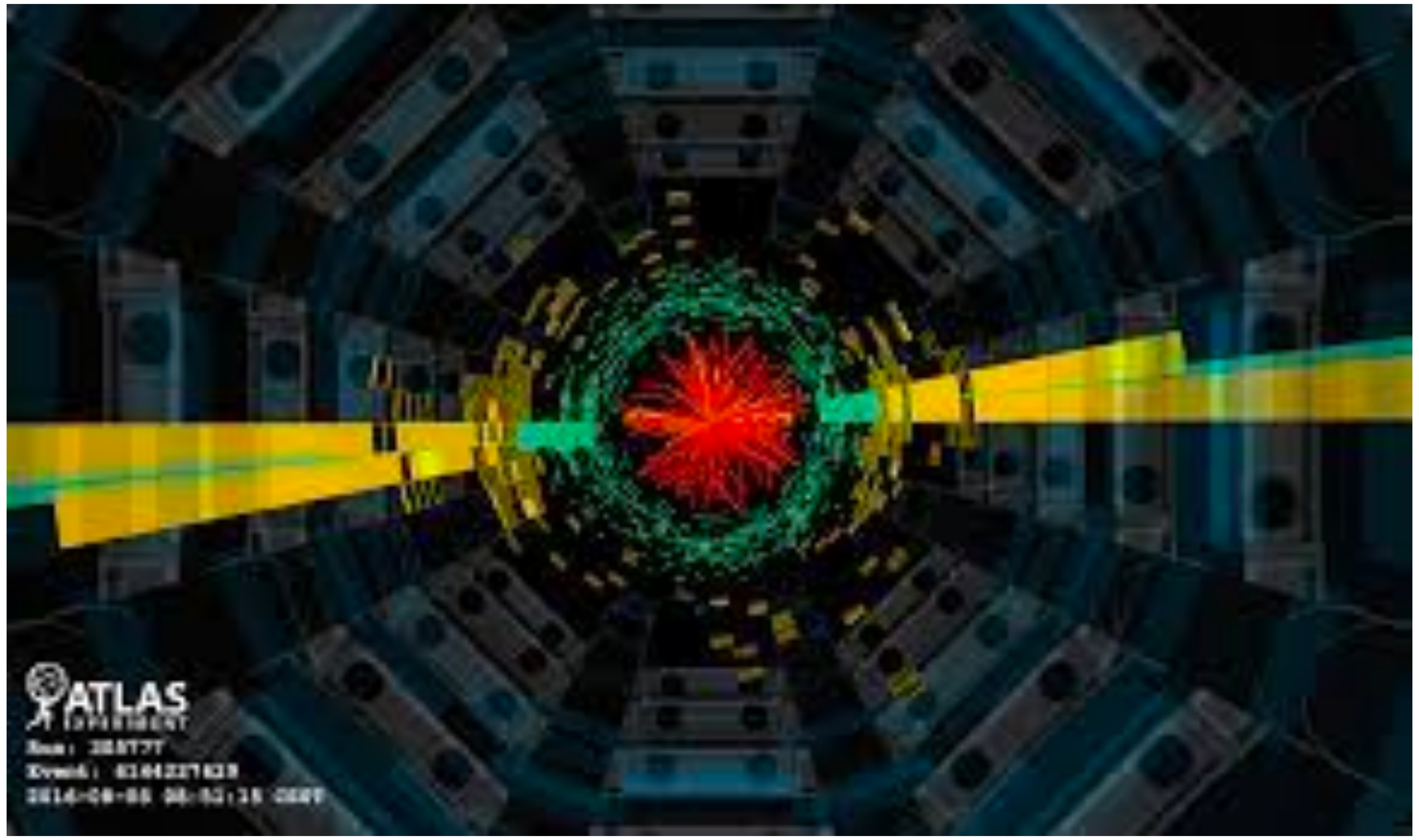


(a) (b) (c)

CWola - Classification Without Labels

- ATLAS search for di-jet resonance [2005.02983, ATLAS]
  - Topology - m_{jj}
  - Discriminating features - m_{j_1} and m_{j_2}





CWola - Classification Without Labels

- ATLAS search for di-jet resonance [2005.02983, ATLAS]
  - Topology - m_{jj}
  - Discriminating features - m_{j_1} and m_{j_2}
- The discriminating variables can't be correlated with the topology variables
- Threshold set for different signal efficiency benchmarks
- Background modeled in standard sideband fit

$$dn/dx = p_1(1-x)^{p_2-\xi_1 p_3} x^{-p_3+(p_4-\xi_2 p_3-\xi_3 p_2)\log(x)}$$

- Search outperforms generic inclusive searches
- Dedicated searches have better sensitivity for searched fore models

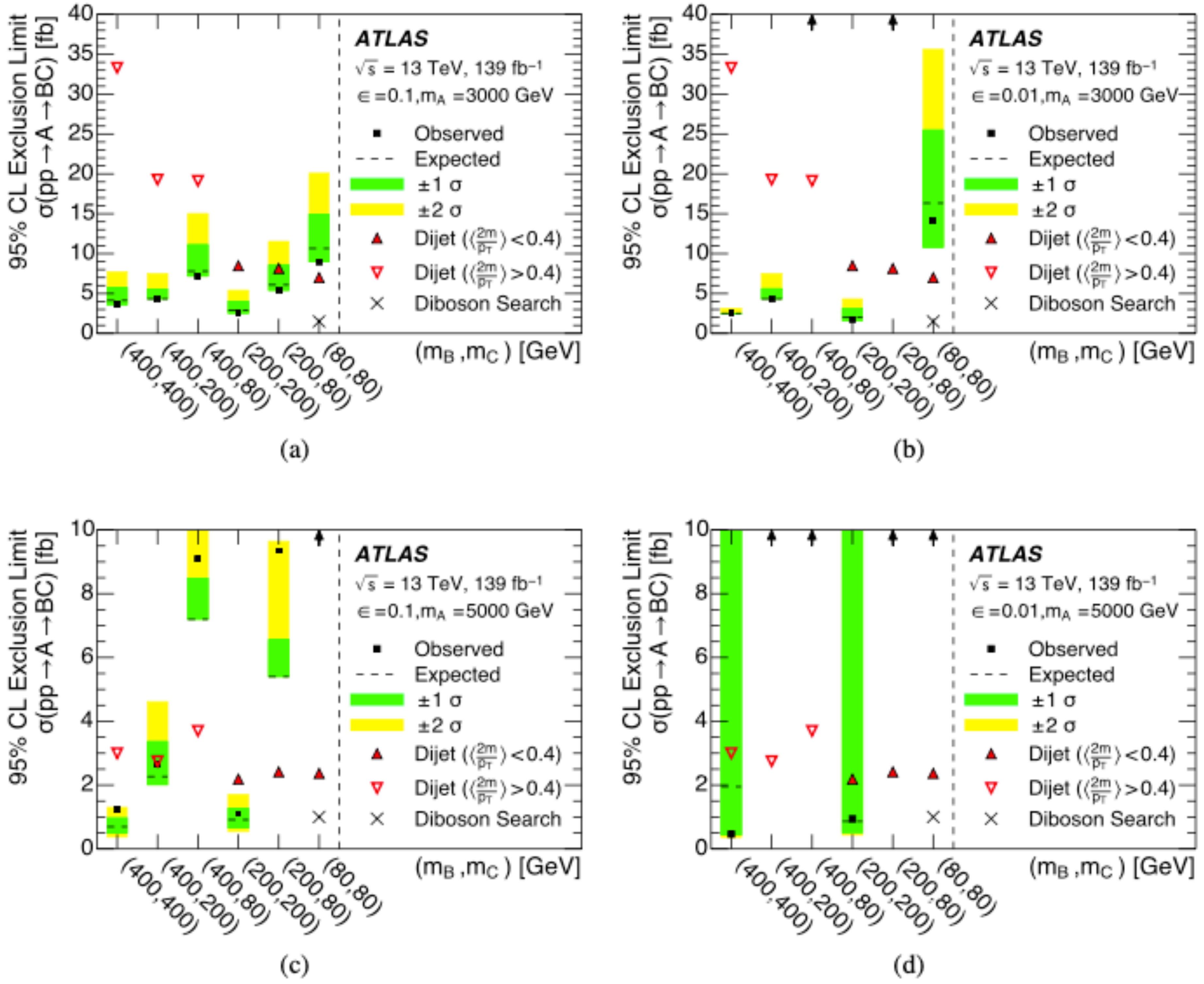
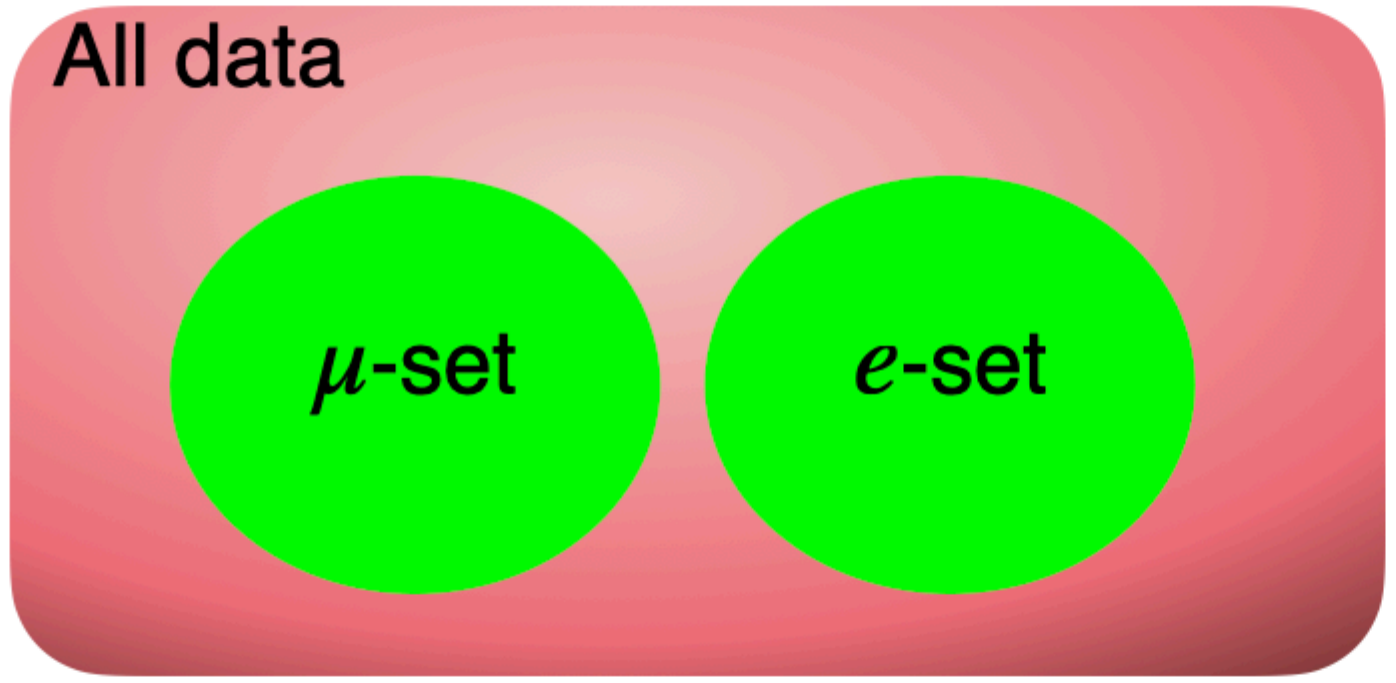


Figure 3: 95% confidence level upper limits on the cross section for a variety of signal models, labeled by (m_B, m_C), in GeV. The limits are shown for signal models with (a,b) m_A = 3000 GeV and NN trained on signal region 2; and (c,d) m_A = 5000 GeV and NN trained on signal region 5. The limits are broken down between the analyses with (a,c) ϵ = 0.1 and (b,d) ϵ = 0.01. Also shown are the limits from the ATLAS dijet search [101] and the ATLAS all-hadronic diboson search [112]. The inclusive dijet limits are calculated using the W' signals from this paper and the full analysis pipeline of Ref. [101]; the diboson search limits are computed using the Heavy Vector Triplet [113] W' signal from Ref. [112]. The acceptance for the W' in this paper, compared to the W' acceptance in Ref. [112], is 86% and 54% for $m_{W'}$ = 3 and 5 TeV, respectively. Missing observed markers are higher than the plotted range. Poor limits occur when the NN fails to tag the signal.

Search for Asymmetries

ML used to infer statistics

[2203.07529, Birman et.al.]



- Flavour symmetric → μ-set and e-set originate from the same pdf
- Flavour asymmetric → μ-set and e-set originate from the different pdf

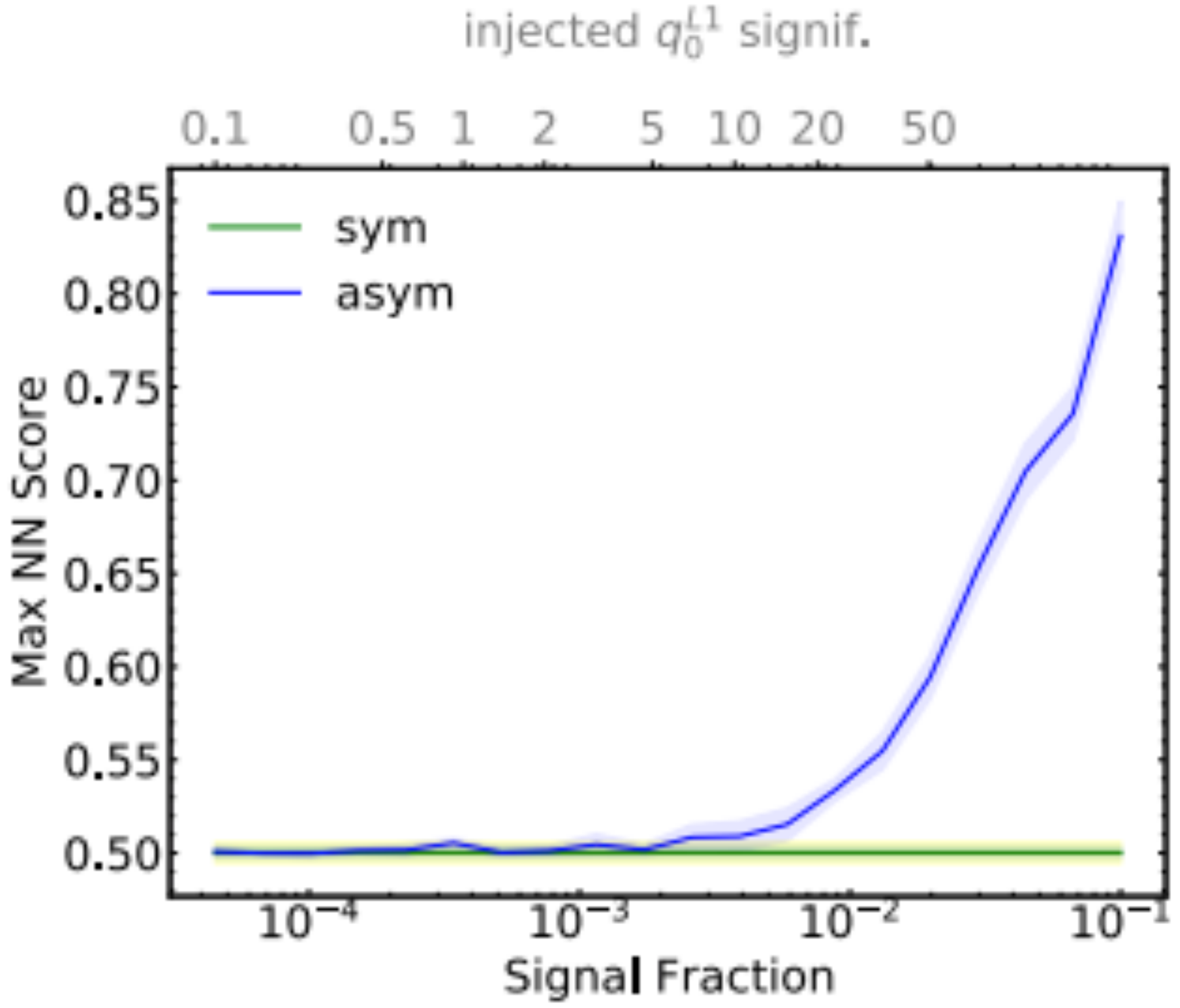


Fig. 6: The maximum neural network score from training a classifier to distinguish the $e\mu$ from μe samples with (asym) and without (sym) a BSM contribution. The green (yellow) and blue bands represent (twice) the standard deviation over 10 bootstrap samples.The separation power is shown as a function of the injected signal fraction (bottom scale) and the corresponding significance calculated with the ideal q_0^{L1} test. Note that these results are not directly comparable to the binned DDP because it is not possible to ignore signal statistical uncertainties.

- Trained NN with μ-set labeled 0 and e-set labeled 1
- Binary cross entropy loss as test statistics
- Compare performance for traditional profile likelihood test statistics

Unsupervised efforts

- The Challenge at hand
- Traditional (non-ML) efforts
- ML-based efforts
  - Supervised
  - Weakly supervised
  - **Unsupervised**
  - Others

Unsupervised efforts

- Several approaches
  - Mostly trained to learn the probability density of the data $p_{data}(x)$
  - Map a random variable z with know probability density to the data $f(z) \rightarrow X$

- Generative adversarial models (GAN)
  - A second network h trained to distinguish $f(z)$ from X
- (Variational) Autoencoders (VAE)
  - $f(z)$ is the decoder operating after the data is encoded into the latent space Z
- Normalizing Flows (NF)
  - A series of invertible functions f_i with tractable Jacobians in order to change Z into X

Finding-over-densities

Anode - Anomaly Detection with Density Estimation

ML used to enhance S/B and background modeling

[2001.04990, Nachman & Shih]

- Signal assumed to be localized somewhere in m
- Discriminating variable $x \rightarrow p_{data}(x|m)$ estimated in SR
- Estimating background pdf from sidebands
- Extrapolate into the signal region $\rightarrow p_{bg}(x|m)$
- Construct a likelihood ratio of data pdf over background pdf in the signal region $R = \dfrac{p_{data}(x|m)}{p_{bg}(x|m)}$
  - No signal $R(x|m) = 1$
  - Presence of signal → $R(x|m) > R_{cut} > 1$
- Various density estimation methods can be used

Anode with normalizing flow models [see e.g., 1908.09257]

[2001.04990, Nachman & Shih]

- Core idea - apply a change of variables from a random variable with a simple density (e.g. Gaussian or uniform) to one with a complex density that matches some training dataset
- Optimize for large $R(x/m)$ and define the threshold $R_c(x/m)$ for best discriminating S from B
- Predict the number of background events for $R_c(x/m)$

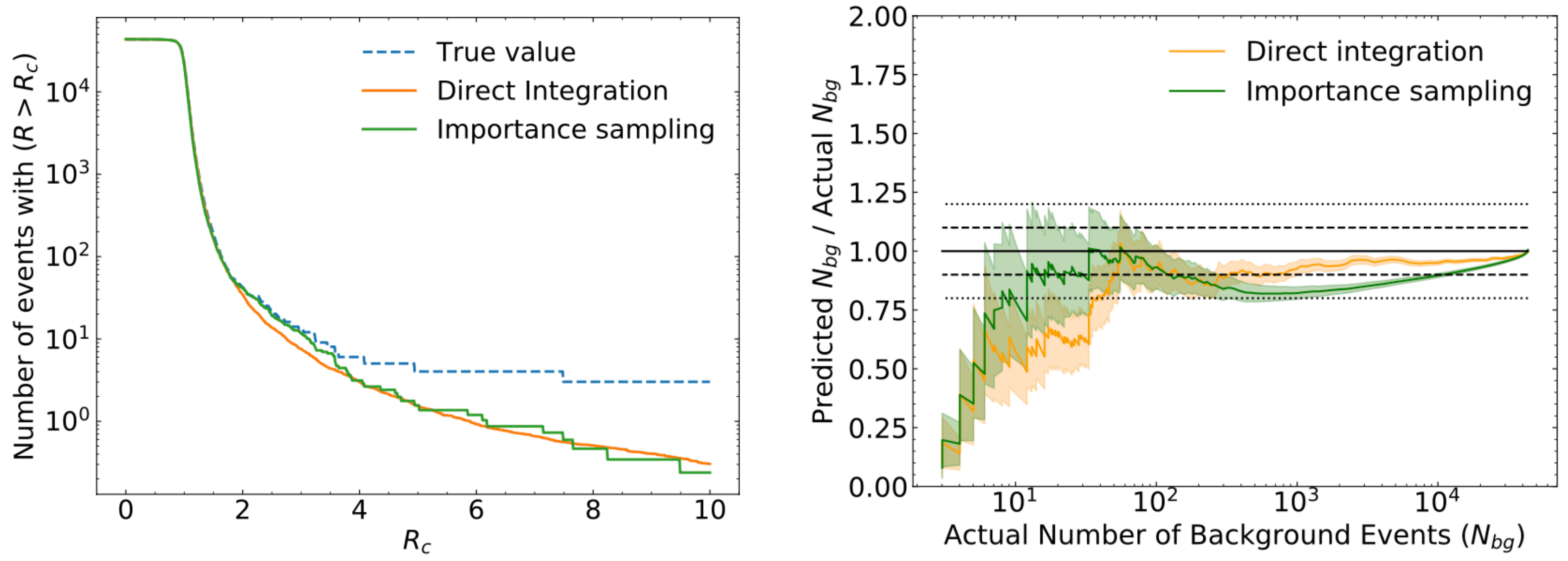


Figure 8. Left: The number of events after a threshold requirement $R > R_c$ using the two integration methods described in Sec. 3.2, as well as the true background yield. Right: The ratio of the predicted and true background yields from the left plot, as a function of the actual number of events that survive the threshold requirement. The shaded bands around the central predictions are the 1σ statistical (Poisson) uncertainty derived from the observed background counts. The black dashed and dotted lines are 10% and 20% around a ratio of 1.

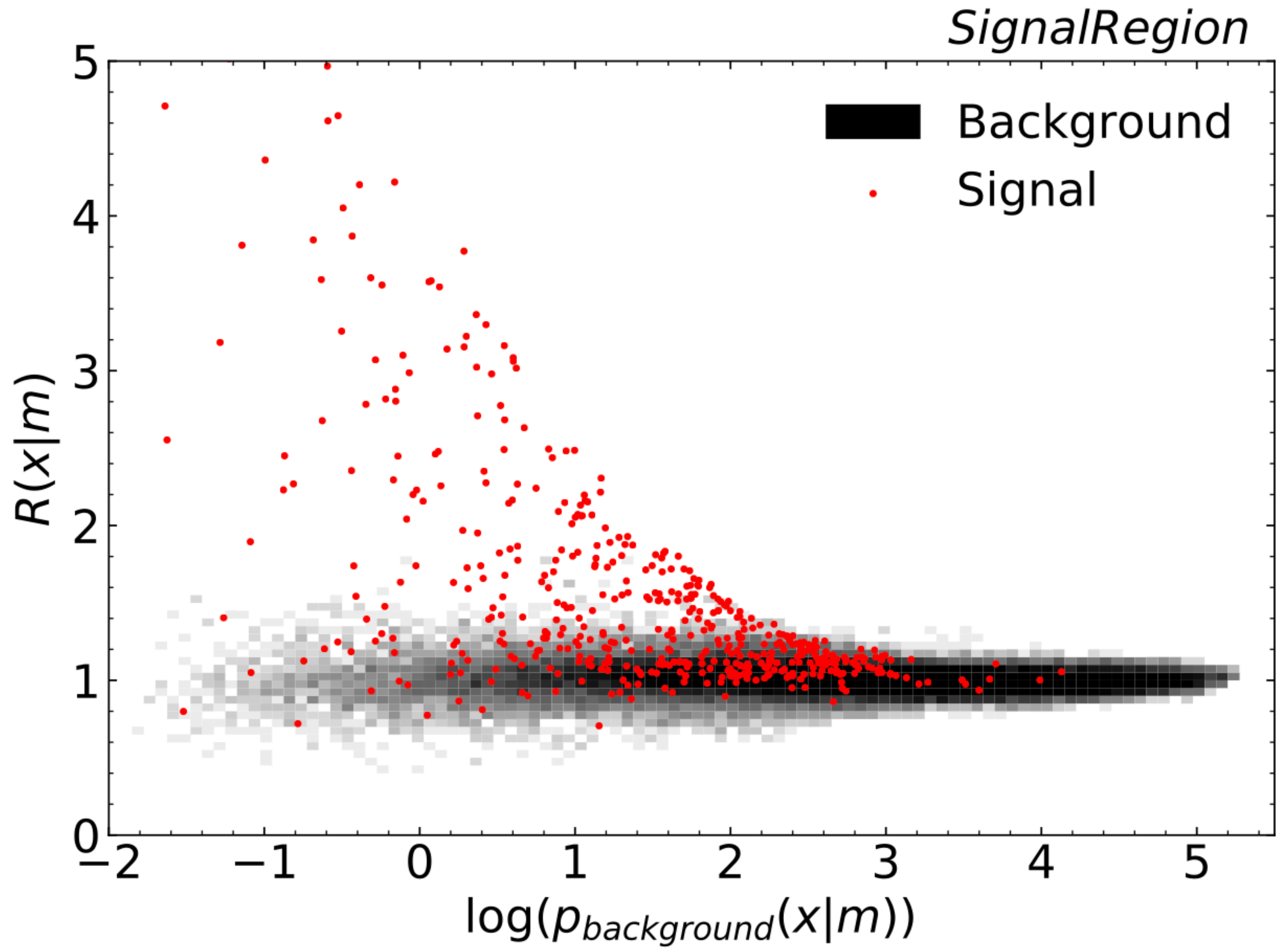


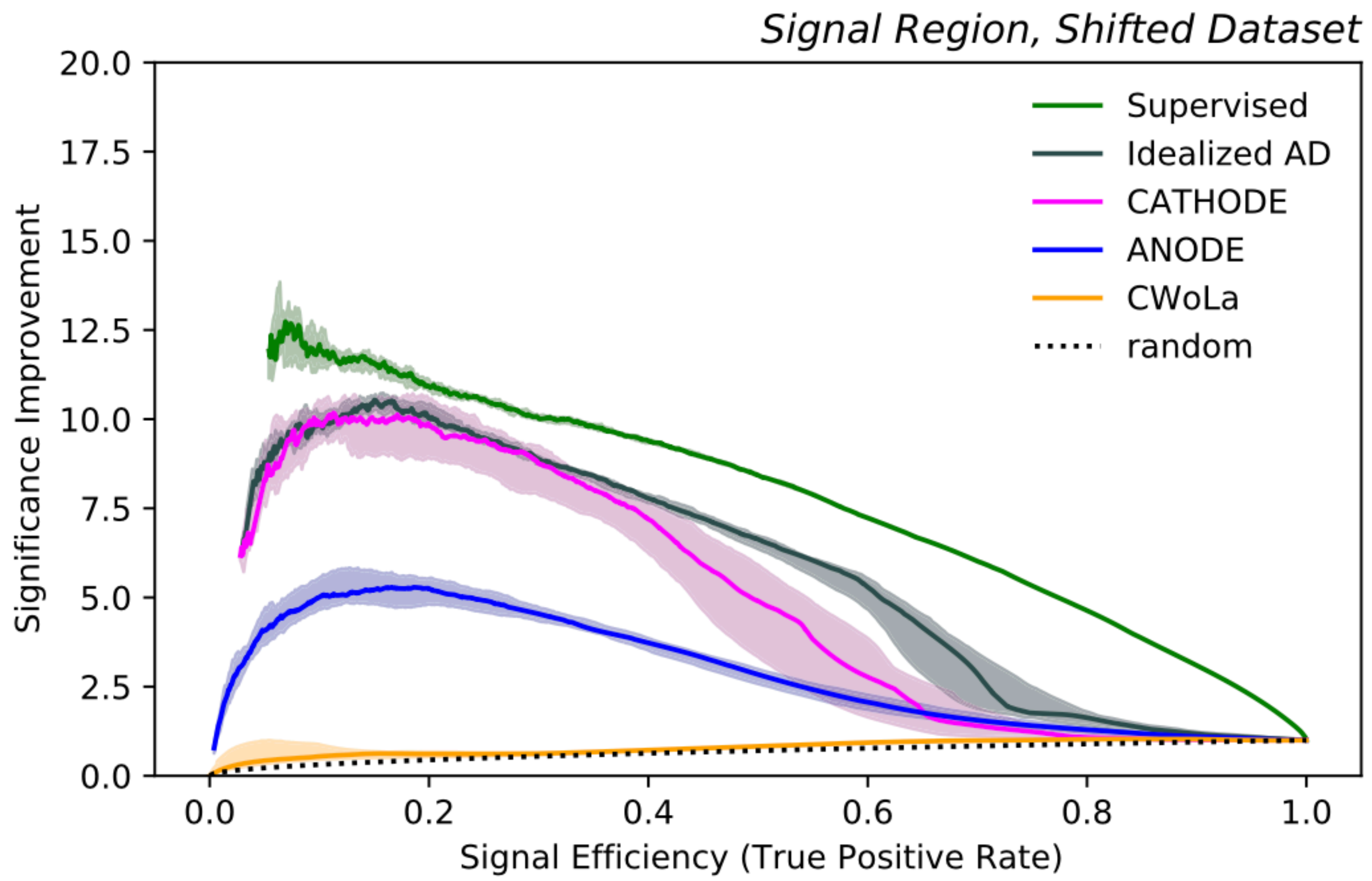
Figure 4. Scatter plot of $R(x|m)$ versus $\log p_{\text{background}}(x|m)$ across the test set in the SR. Background events are shown (as a two-dimensional histogram) in grayscale and individual signal events are shown in red.

Cathode* - Classifying anomalies through outer density estimation

ML used to enhance S/B and background modeling

[2109.00546 , Hallin, et.al.]

- Train a density estimator to learn the smooth background distribution in the sideband
- Interpolate into the → $p_{bg}(x\,|\,m)$
- Generate sample events from $p_{bg}(x\,|\,m)$
- Train a classifier to distinguish between $p_{bg}(x\,|\,m)$ and $p_{data}(x\,|\,m)$ and maximize $r(x\,|\,m)$
- See also yesterday's talk by Gregor Kasieczka



Idealized AD - trained on data vs. perfectly simulated background

* Cathode uses several NN, some unsupervised and some semi-supervised

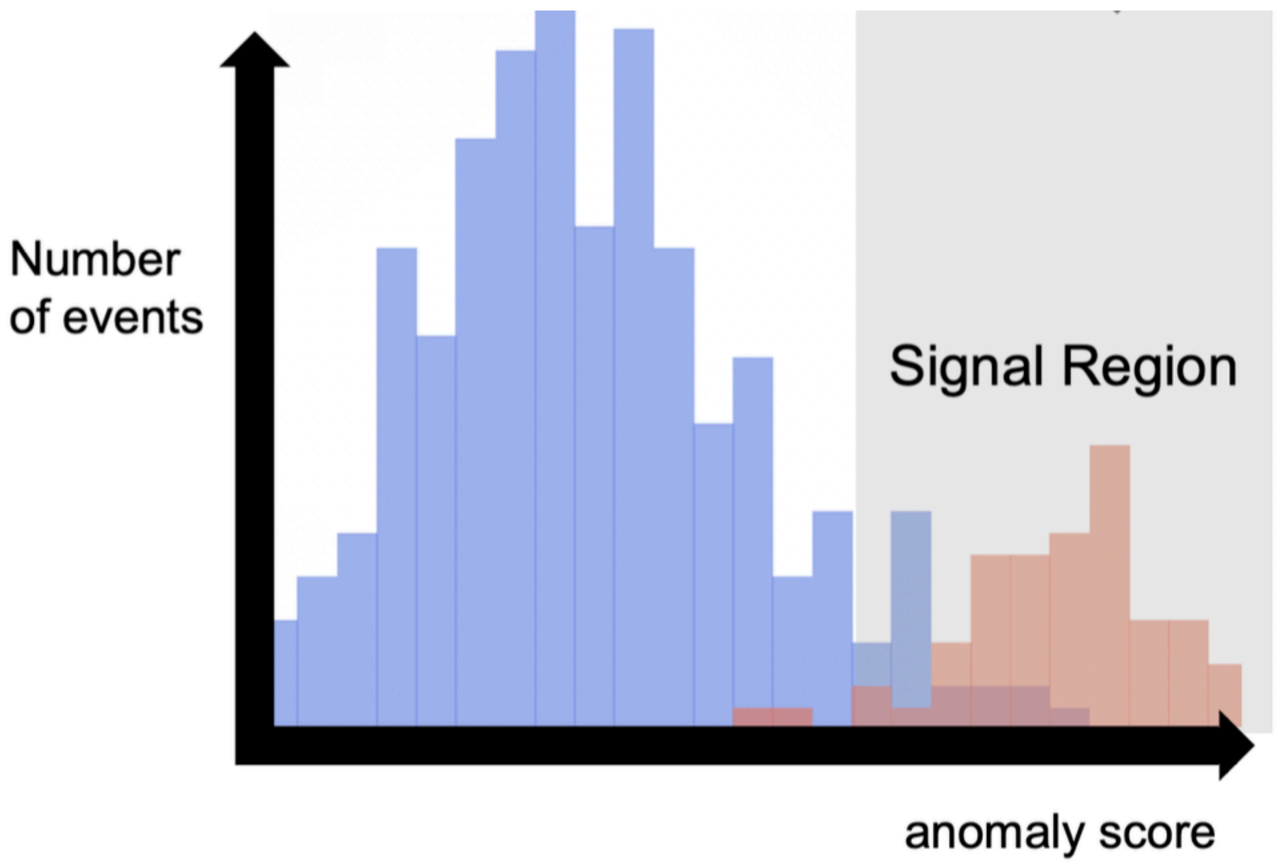
Outliers detection

Outliers detection

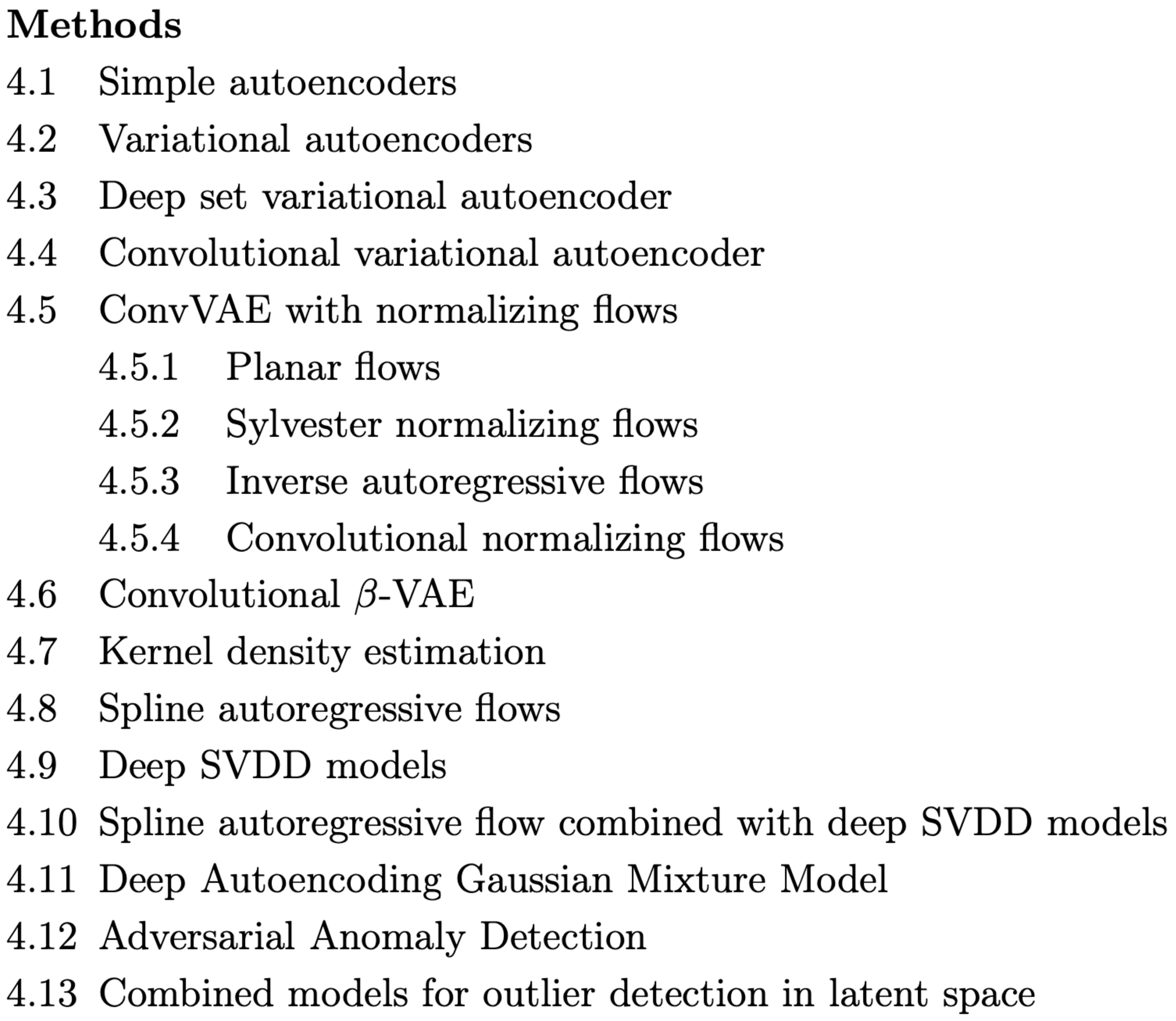
The Dark Machine challenge

[2105.14027, Arrestad, et.al.]

- Define an anomaly score SR
- On an event-by-event basis
- Using unsupervised algorithm trained without defining a signal
  - On simulated SM events only
  - On data if signal is rare relative to background
- Background estimation and statistical inference done regularly



ML used to enhance S/B



Methods

4.1 Simple autoencoders

4.2 Variational autoencoders

4.3 Deep set variational autoencoder

4.4 Convolutional variational autoencoder

4.5 ConvVAE with normalizing flows

- 4.5.1 Planar flows
- 4.5.2 Sylvester normalizing flows
- 4.5.3 Inverse autoregressive flows
- 4.5.4 Convolutional normalizing flows

4.6 Convolutional β-VAE

4.7 Kernel density estimation

4.8 Spline autoregressive flows

4.9 Deep SVDD models

4.10 Spline autoregressive flow combined with deep SVDD models

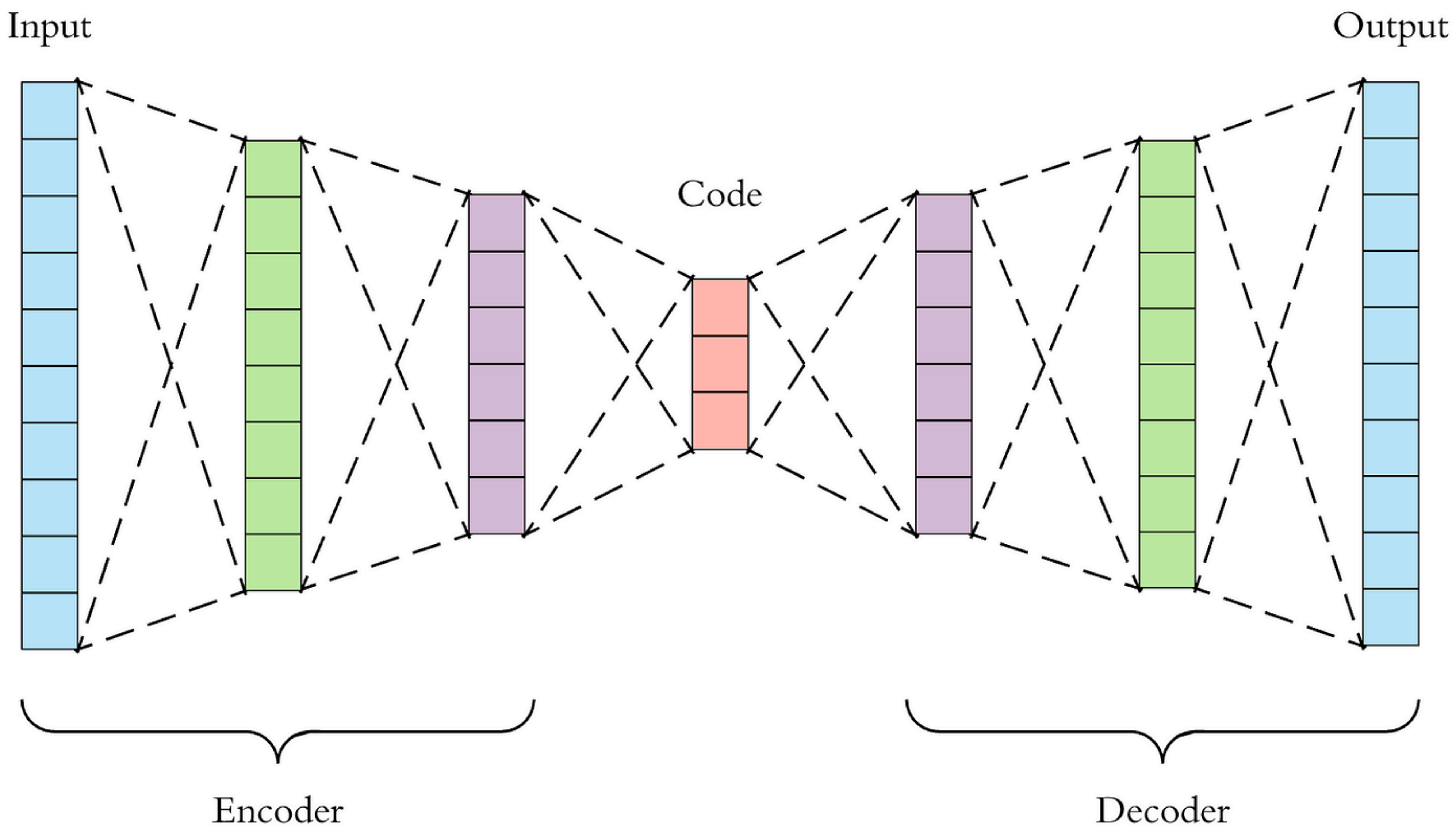
4.11 Deep Autoencoding Gaussian Mixture Model

4.12 Adversarial Anomaly Detection

4.13 Combined models for outlier detection in latent space

Autoencoders

- Unsupervised NN
- Learns to compress and encode data and to reconstruct the data back from the reduced encoded representation to a representation that is as close to the original input as possible
- Difference between the original and reconstructed data measured with a loss function
- By design, reduces data dimensions by learning how to ignore noise
- Here, for small S/B ratio supposed to learn mostly the background



Autoencoders

- ATLAS search for 2-body resonances
- 9 signal regions:

  $m_{jj}, m_{jb}, m_{bb}, m_{je}, m_{j\mu}, m_{be}, m_{b\mu}, m_{j\gamma}, m_{b\gamma}$
- Anomaly regions (cuts) assumes hypothetical cross-sections
- Statistical inference based on m_{inv} distribution using fit to $f(x) = p_1(1-x)^{p_2} x^{p_3+p_4 \ln x + p_5 \ln^2 x}$

[2307.01612, ATLAS]

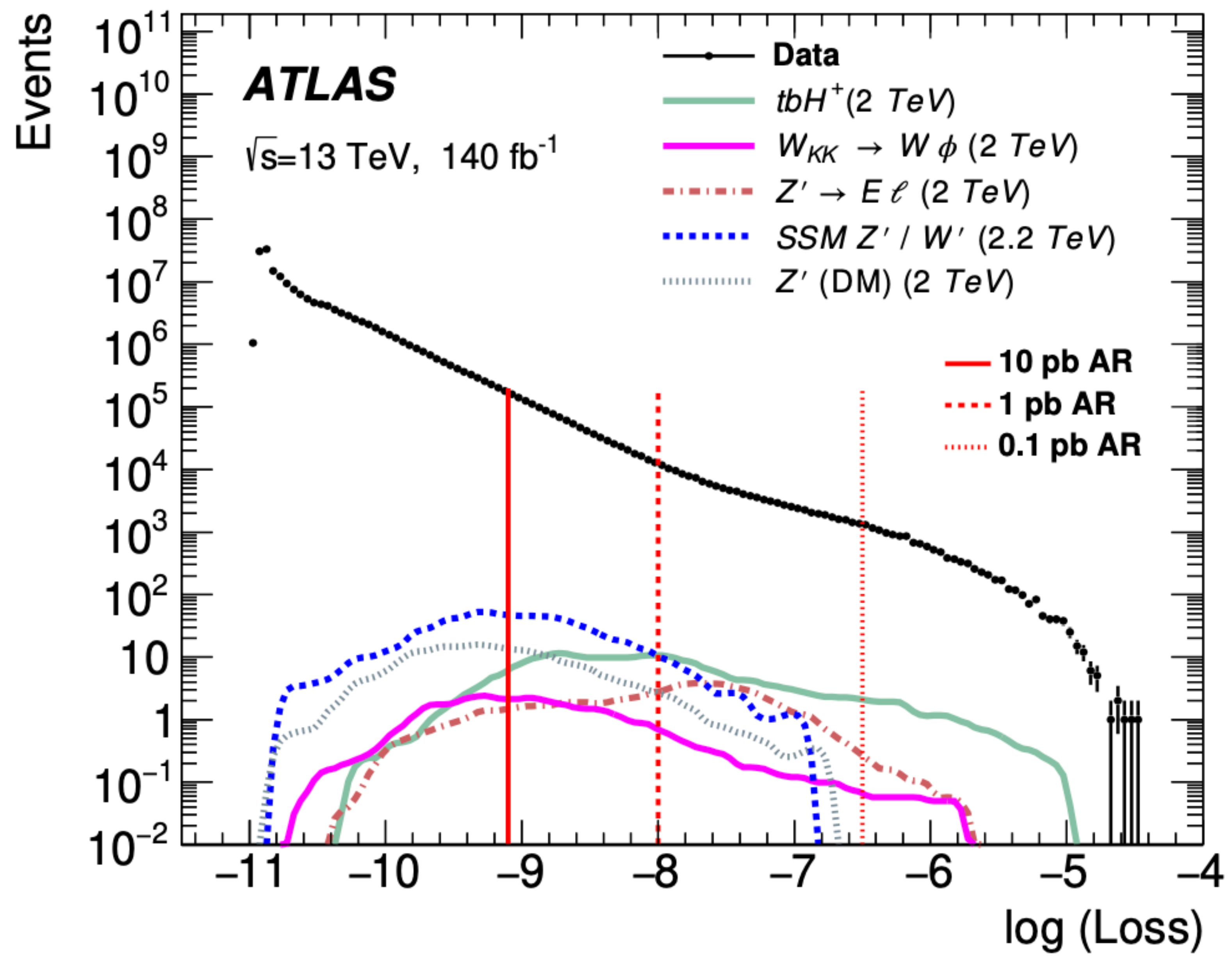
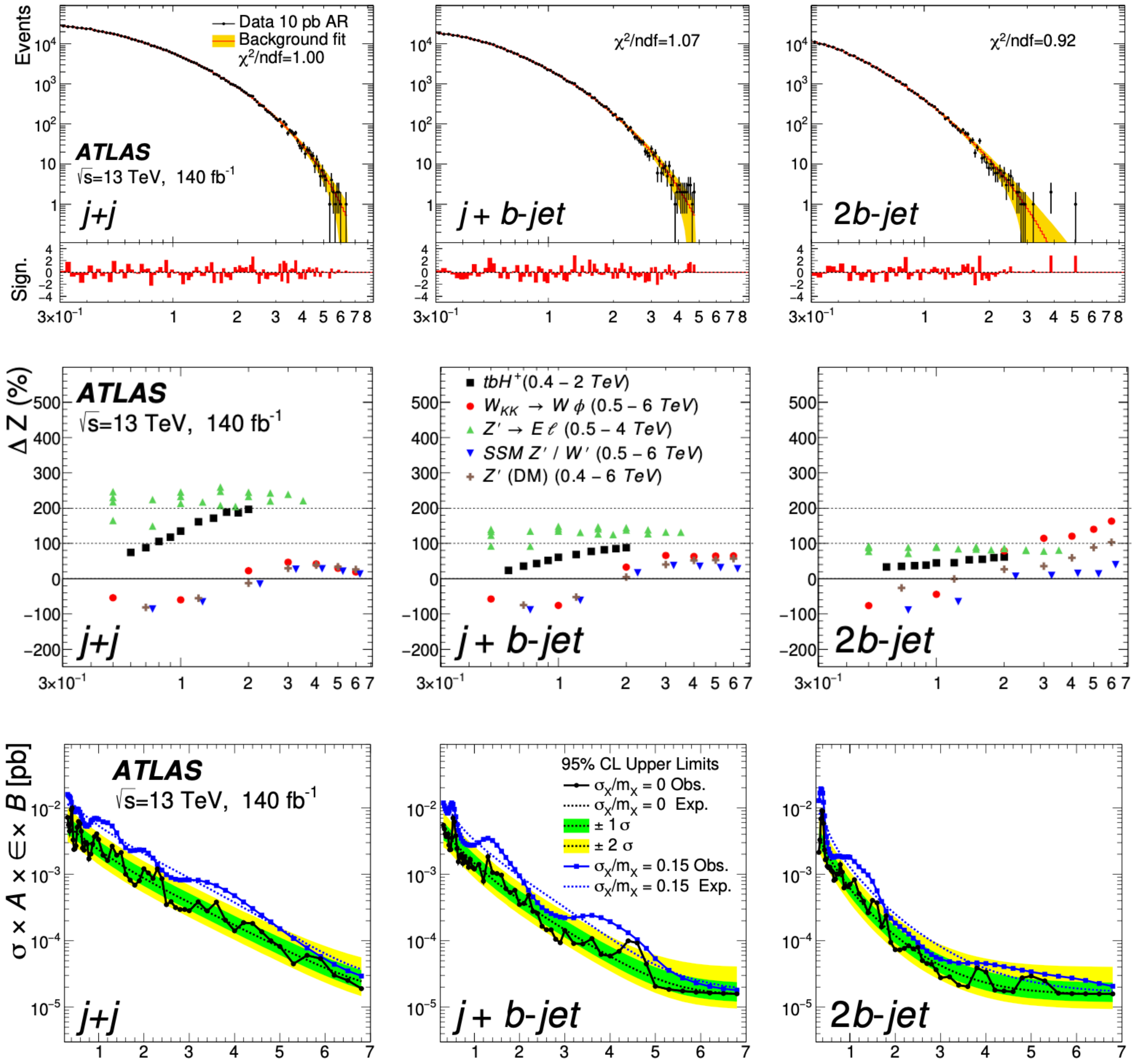


Figure 1: Distributions of the anomaly score from the AE for data and five benchmark BSM models. Their legends, from top to bottom, are: (1) charged Higgs boson production in association with a top quark, tbH^+ with $H^+ \to t\bar{b}$; (2) a Kaluza–Klein gauge boson, W_{KK}, with the SM W boson and a radion ϕ; (3) a Z' boson decaying to a composite lepton E and ℓ, with $E \to Z\ell$ with a mass of 0.5 TeV; (4) the SSM $W' \to WZ' \to \ell\nu q\bar{q}$; (5) a simplified dark-matter model with an axial-vector mediator $Z' \to q\bar{q}$, where one of the quarks radiates a W boson decaying to $\ell\nu$. The BSM predictions represent the expected number of events from 140 fb^{-1} of data for heavy particle (H^+, W_{KK}, Z', W' and Z', respectively) masses around 2 TeV. The distributions for the BSM models are smoothed to remove fluctuations due to low MC event counts. The vertical lines indicate the start of the three anomaly regions (ARs). The labels of the three ARs indicate the visible cross section for hypothetical processes yielding the same number of events as observed in the 140 fb^{-1} dataset. The AE is applied to preselected events without any requirements on invariant mass distributions.

Autoencoders

[2307.01612, ATLAS]



$$f(x) = p_1(1-x)^{p_2} x^{p_3+p_4 \ln x + p_5 \ln^2 x}$$

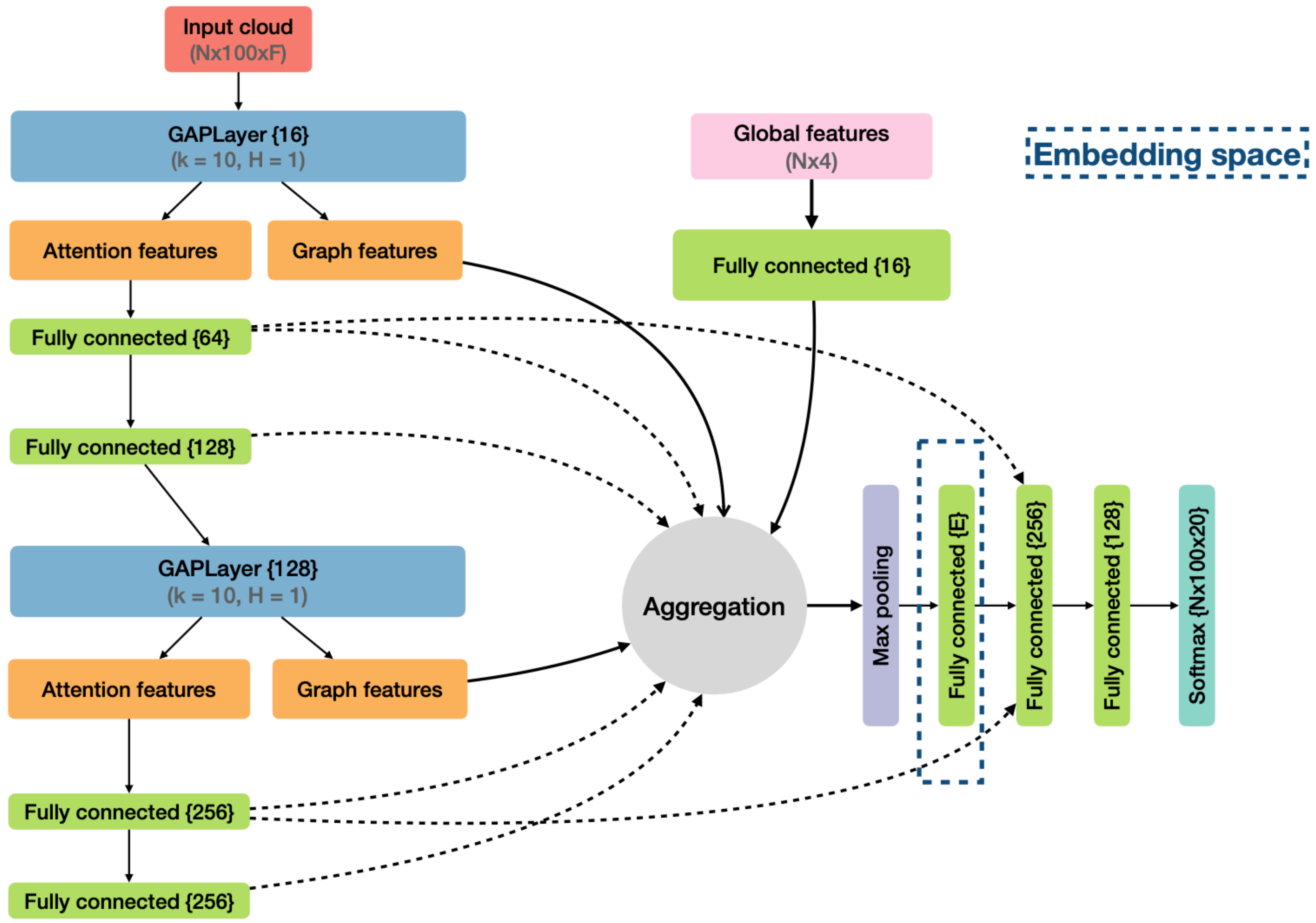
$$\Delta Z = ((Z_{AE}/Z) - 1) \times 100\%,$$

$$Z = \sqrt{2\left((s+b)\ln(1+s/b) - s\right)}$$

Unsupervised Clustering

- Procedure:
  - Reduce the data dimensionality to retain the main properties of the events
  - In the reduced representation add clustering objective to the training procedure → group together points in the reduced representation that share similar properties
    - Clustering is based on physical considerations e.g., constituent of jets
  - Loss function is defined to take into account to classification as well as the clustering properties

[LHC Olympics]



Other ideas

- The Challenge at hand
- Traditional (non-ML) efforts
- ML-based efforts
  - Supervised
  - Weakly supervised
  - Unsupervised
  - **Others**

NPLM - Learning New Physics from Data

ML used for S/B enhancement and statistical inference

[1806.02350, D'Agnolo & Wulzer]

- Goal: detect data departures from a given reference model
- Train a classifier with a loss function equivalent to the likelihood ratio
- For two datasets A and B define:
  - $N_b >> N_A$

$$\mathcal{H}_0: \qquad n_{\mathbf{A}}(x|\mathcal{H}_0) = \frac{N_{\mathbf{A}}}{N_{\mathbf{B}}} n_{\mathbf{B}}(x,\nu)$$

$$\mathcal{H}_1: \qquad n_{\mathbf{A}}(x|\mathcal{H}_1) = \frac{N_{\mathbf{A}}}{N_{\mathbf{B}}} e^{f(x,\mu)} n_{\mathbf{B}}(x,\nu)$$

$$t = 2\log\left(\frac{\max_{\nu,\mu}\left(\mathcal{L}\left(\mathcal{H}_1|\mathbf{A}\right)\right)}{\max_{\nu}\left(\mathcal{L}\left(\mathcal{H}_0|\mathbf{A}\right)\right)}\right) \qquad \mathcal{L}\left(\mathcal{H}|\mathbf{A}\right) = \frac{e^{-N_{\mathbf{A}}(\mathcal{H})}}{\tilde{N}_{\mathbf{A}}!}\prod_{x\in\mathbf{A}} n_{\mathbf{A}}(x|\mathcal{H})$$

- The test statistics

$$t = t_{\mathbf{B}}\left(\mathbf{A}\right) \equiv -2\left(\frac{\hat{N}_{\mathbf{A}}\left(\mathcal{H}_0\right)}{\tilde{N}_{\mathbf{B}}}\sum_{x\in\mathbf{B}}\left(e^{\hat{f}(x)}-1\right) - \sum_{x\in\mathbf{A}}\hat{f}\left(x\right)\right)$$

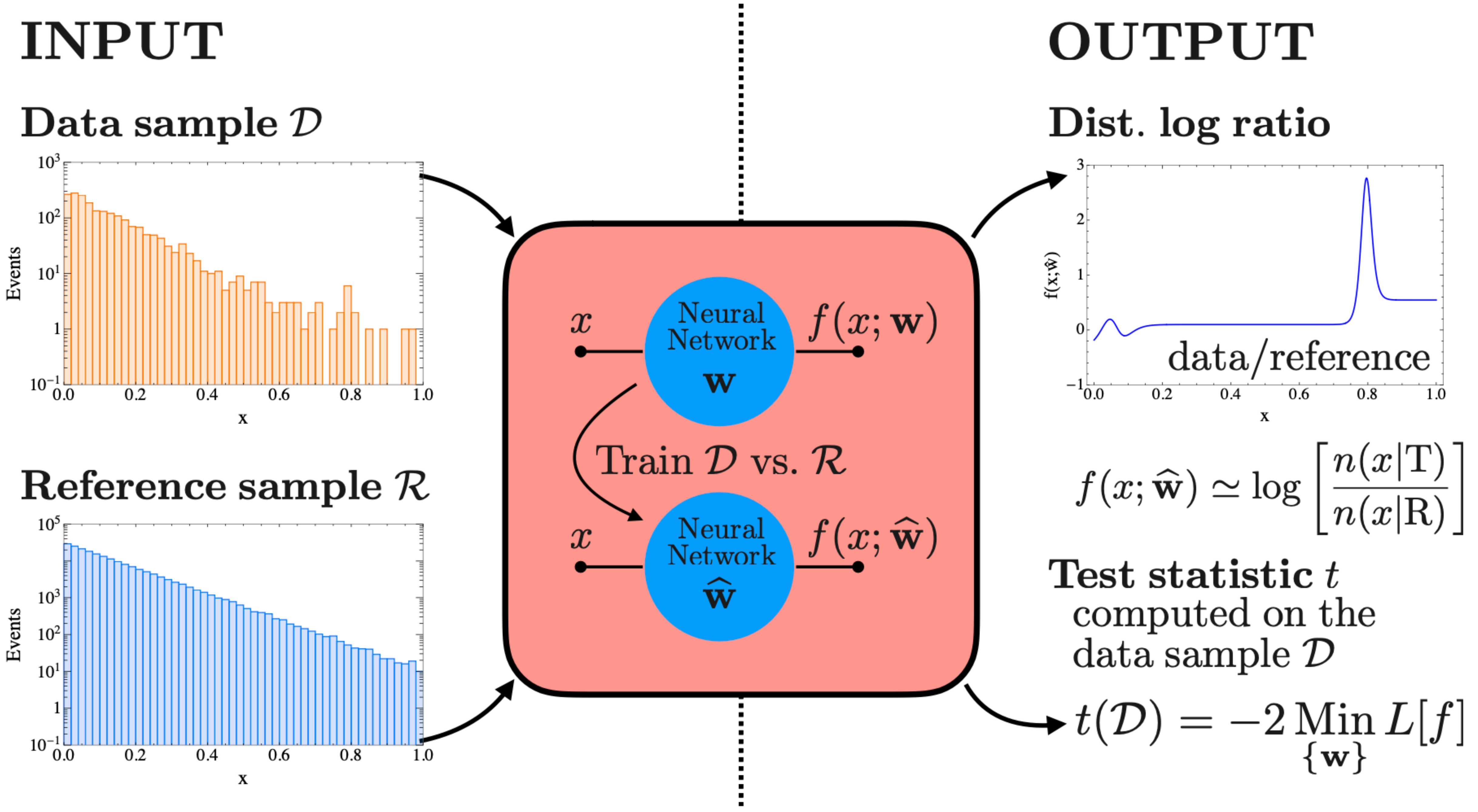
$$f\left(x\right) = b_{\text{out}} + \sum_{\alpha=1}^{N_{\text{neu}}} w_{\text{out}}^{\alpha}\sigma\left(w_{\alpha}x + b_{\alpha}\right)$$

- Use NN machinery to fit f that maximizes the log likelihood ratio

NPLM - Learning New Physics from Data

[1806.02350, D’Agnolo & Wulzer]

- Log likelihood ratio test statistics → background only distribution follows $\chi^2_{n_{dof}}$
- Fitting procedure → n_{dof} determined by the number of free parameters in the fit model
  - The w’s and b’s in $f\,(x) = b_{\text{out}} + \sum_{\alpha=1}^{N_{\text{neu}}} w^{\alpha}_{\text{out}}\sigma\,(w_\alpha x + b_\alpha)$
- The reference sample assumed to be much larger than the data
- Some fine tuning is needed in weight clippings



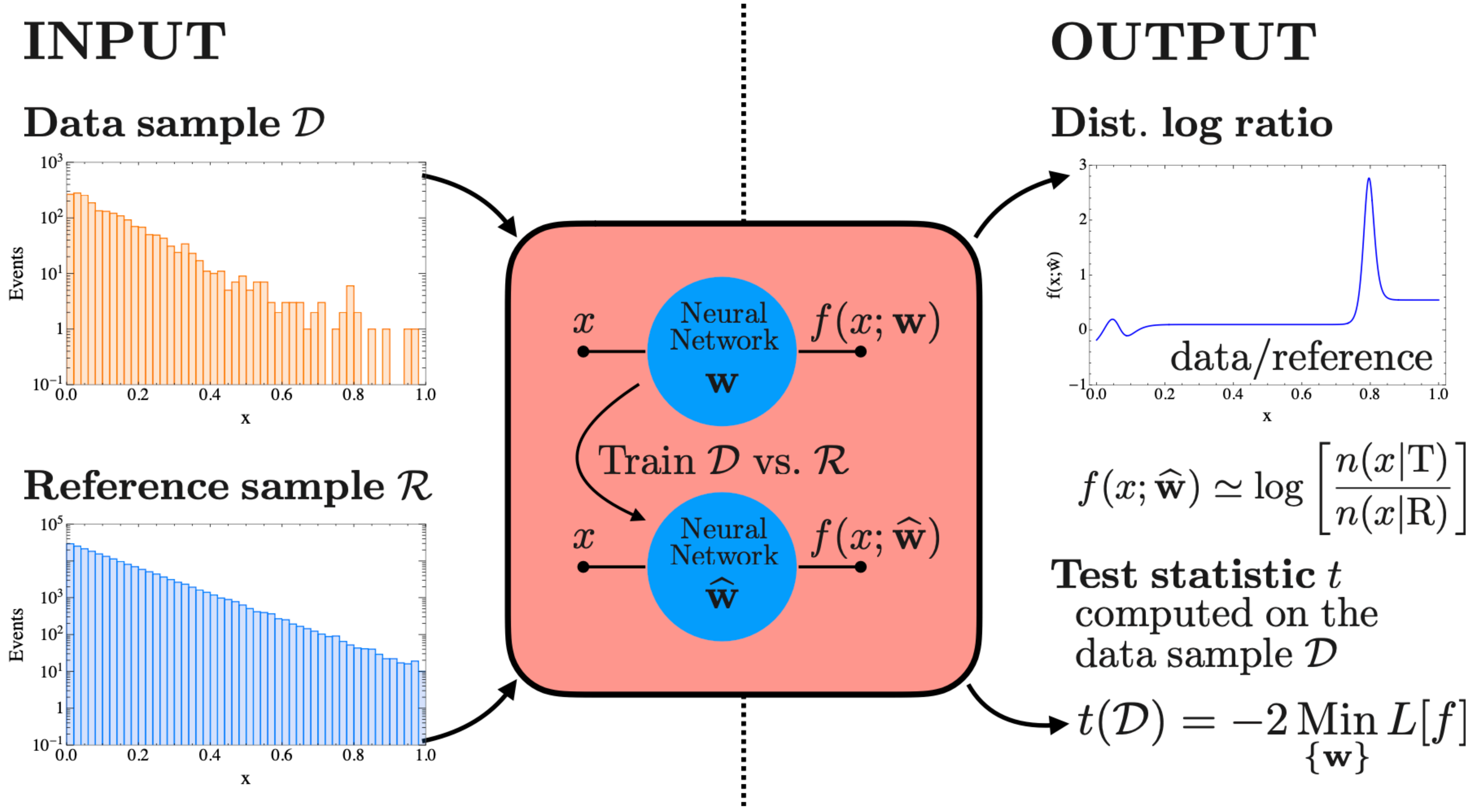
INPUT
Data sample $\mathcal{D}$
Events
x
Reference sample $\mathcal{R}$
Events
x
x
Neural Network
$\mathbf{w}$
$f\,(x;\mathbf{w})$
Train $\mathcal{D}$ vs. $\mathcal{R}$
x
Neural Network
$\widehat{\mathbf{w}}$
$f\,(x;\widehat{\mathbf{w}})$
OUTPUT
Dist. log ratio
f(x;ŵ)
data/reference
x
$f\,(x;\widehat{\mathbf{w}}) \simeq \log\left[\frac{n(x|\mathrm{T})}{n(x|\mathrm{R})}\right]$
Test statistic t
computed on the data sample $\mathcal{D}$
$t(\mathcal{D}) = -2\,\underset{\{\mathbf{w}\}}{\mathrm{Min}}\,L[f]$

Figure 1: A schematic representation of the implementation of our strategy.

NPLM - Learning New Physics from Data

- Log likelihood ratio test statistics → background only distribution follows $\chi^2_{n_{dof}}$
- Fitting procedure → n_{dof} determined by the number of free parameters in the fit model
  - The w's and b's in $f\left(x\right) = b_{\text{out}} + \sum_{\alpha=1}^{N_{\text{neu}}} w^{\alpha}_{\text{out}} \sigma\left(w_{\alpha} x + b_{\alpha}\right)$
- The reference sample assumed to be much larger than the data
- Some fine tuning is needed in weight clippings
- See also Mikael Kuusla's talk las Monday

[1806.02350, D'Agnolo & Wulzer]



INPUT
Data sample $\mathcal{D}$
Events
10^{-1}
1
10^1
10^2
10^3
0.0
0.2
0.4
0.6
0.8
1.0
x
Reference sample $\mathcal{R}$
Events
10^{-1}
1
10^1
10^2
10^3
10^5
0.0
0.2
0.4
0.6
0.8
1.0
x
x
Neural Network
$\mathbf{w}$
$f(x;\mathbf{w})$
Train $\mathcal{D}$ vs. $\mathcal{R}$
x
Neural Network
$\widehat{\mathbf{w}}$
$f(x;\widehat{\mathbf{w}})$
OUTPUT
Dist. log ratio
$f(x;\hat{w})$
-1
0
1
2
3
0.0
0.2
0.4
0.6
0.8
1.0
x
data/reference
$f(x;\widehat{\mathbf{w}}) \simeq \log\left[\frac{n(x|\mathrm{T})}{n(x|\mathrm{R})}\right]$
Test statistic t
computed on the data sample $\mathcal{D}$
$t(\mathcal{D}) = -2\,\underset{\{\mathbf{w}\}}{\mathrm{Min}}\,L[f]$

Figure 1: A schematic representation of the implementation of our strategy.

Data Directed Paradigm - DDP

ML used for rapid statistical inference

- Two key ingredients:
  - A property of the SM based on which an anomaly can be searched for
  - A tool allowing to infer statistically the significance of a deviation from this property
- Allow scanning rapidly many sub-selections of the data
- Can be combined with other optimization algorithms
- Two examples developed
  - The BumpHunt DDP → see talk by Evan Mayer on Thursday [2107.11573, Volkovich, et.al.]

→ See poster by Bruna Pascual

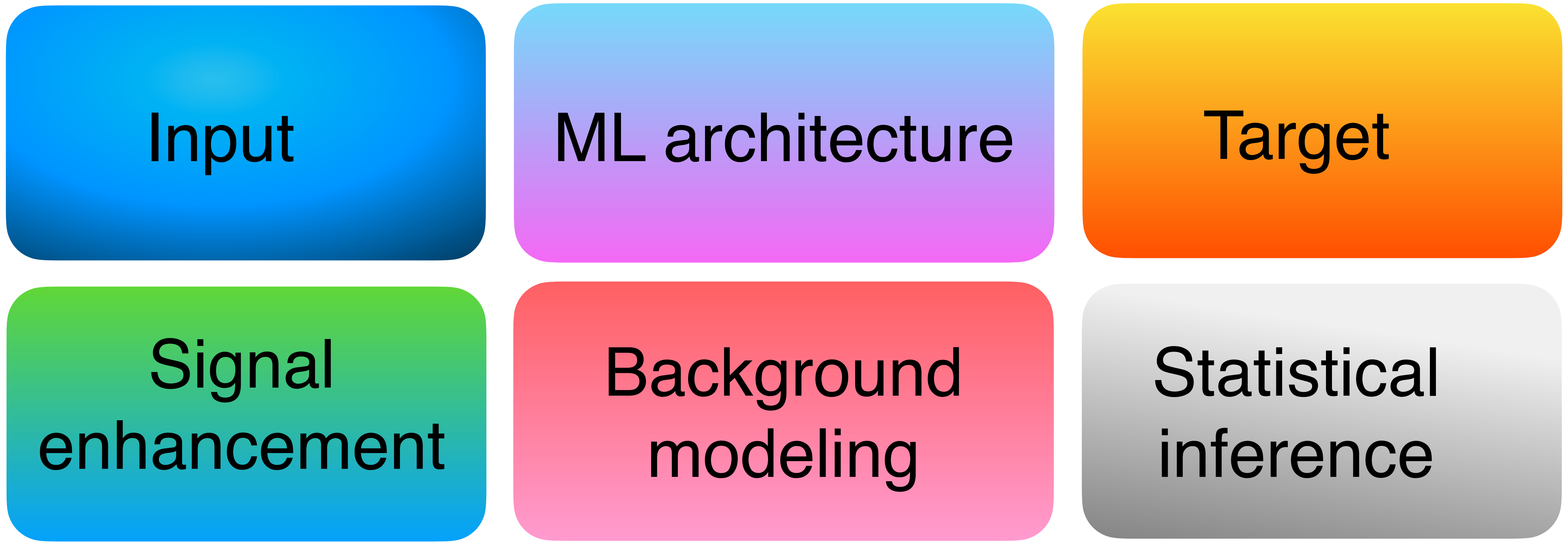
"Accelerating the search for mass bumps using the Data-Directed Paradigm"

- Symmetry DDP → see my talk on Thursday [2401.09530, SB, et.al.]
  "Exploiting the discovery potential of the LHC data using the Data Directed Paradigm"

Summary

Summary

- The LHC (and other accelerators) data is far from being fully exploited
- ML is in the process of revolutionizing anomaly detection also in particle physics
  - Yet huge difference between suggesting an idea and actually apply it to data and… get it approved by the collaborations



Input

ML architecture

Target

Signal enhancement

Background modeling

Statistical inference

- Nothing to summarize since it is clearly just the beginning