Signals of New Physics as an Anomaly

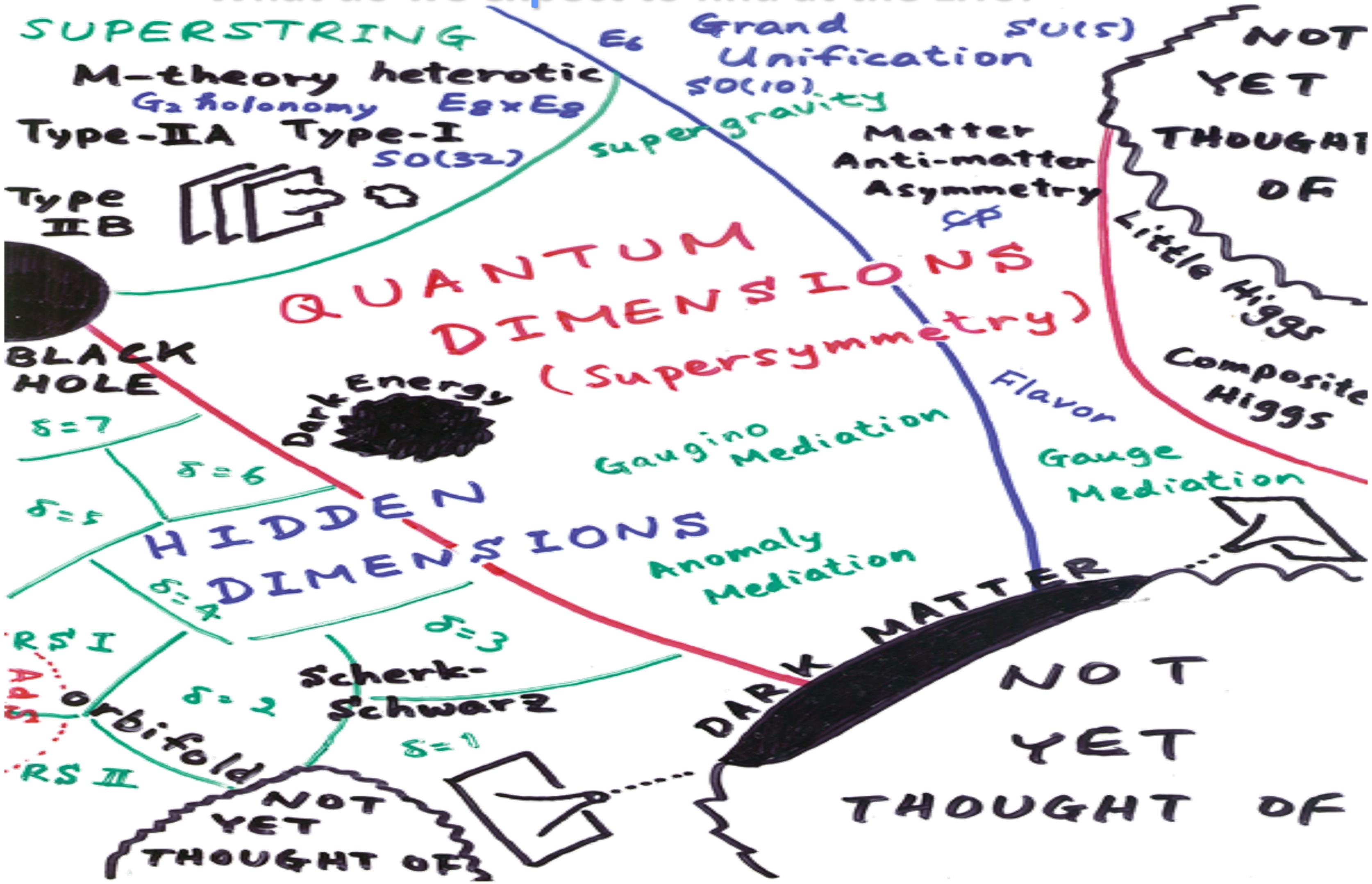
Supervised, unsupervised and data-derived signal regions
Phystat Anomaly 2022

Sascha Caron
(Radboud University and Nikhef)

Data visualisation art by Ouchhh Studio

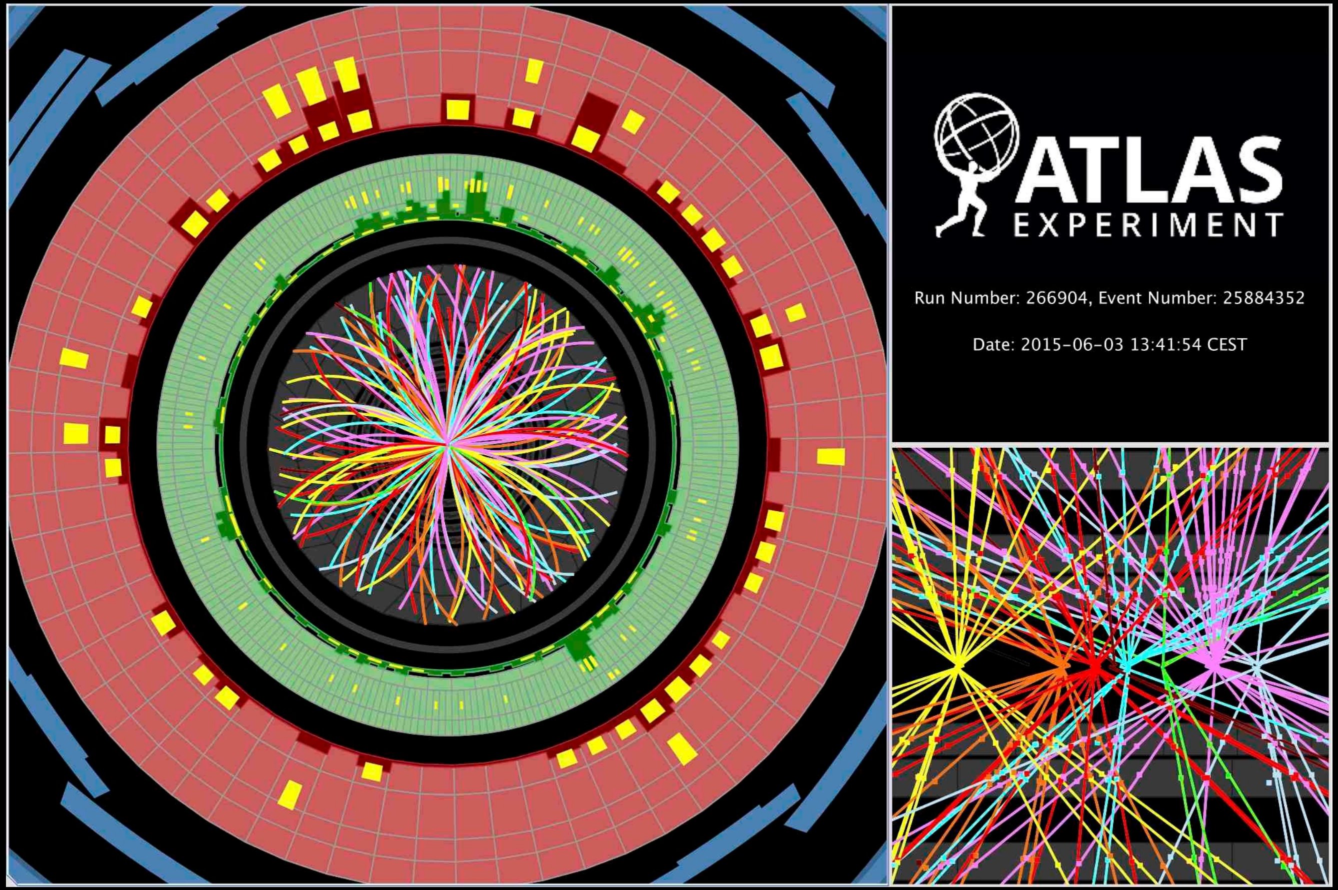
The situation in 2006

What do we expect to find at the LHC?



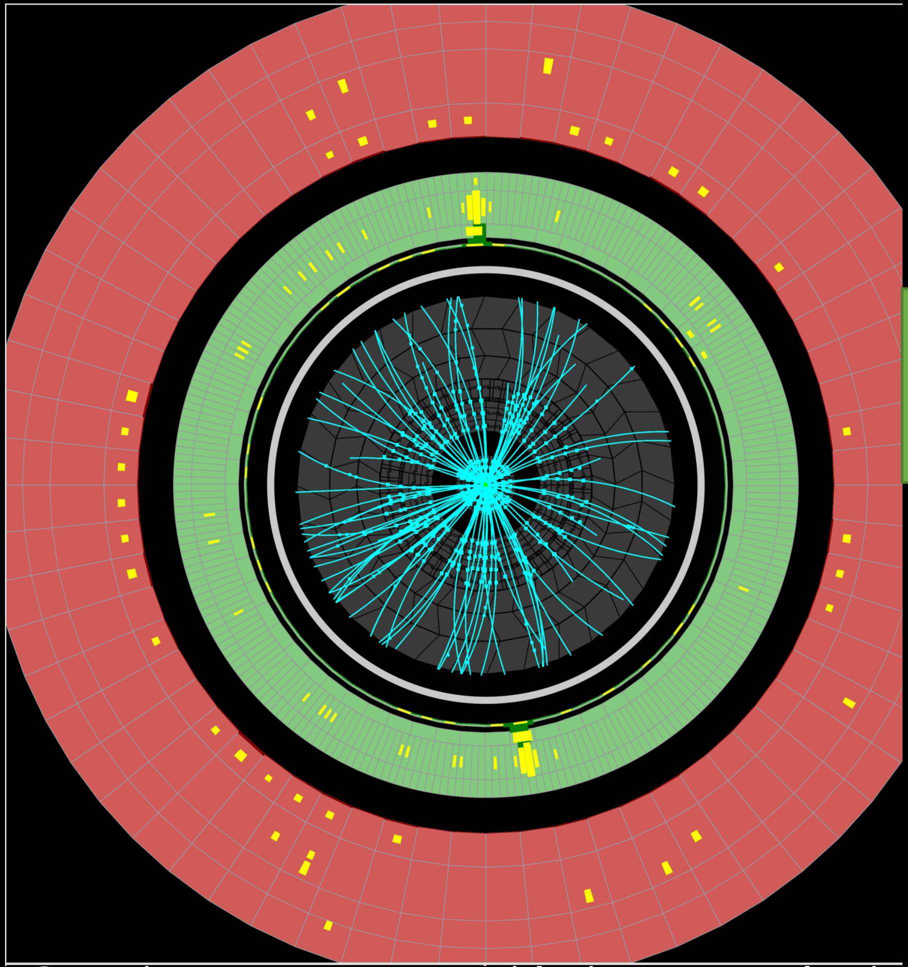
One physicist's schematic view of particle physics in the 21st century
(Courtesy of Hitoshi Murayama)

Most events look like this…



Event from LHC run-2

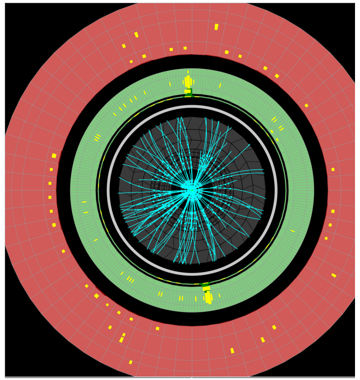
1 in >1000 billion events looks like this



Mass of the Higgs is reconstructed with photon energies

Higgs to 2 photon candidate with mass of 125 GeV

Traditional approach Model driven



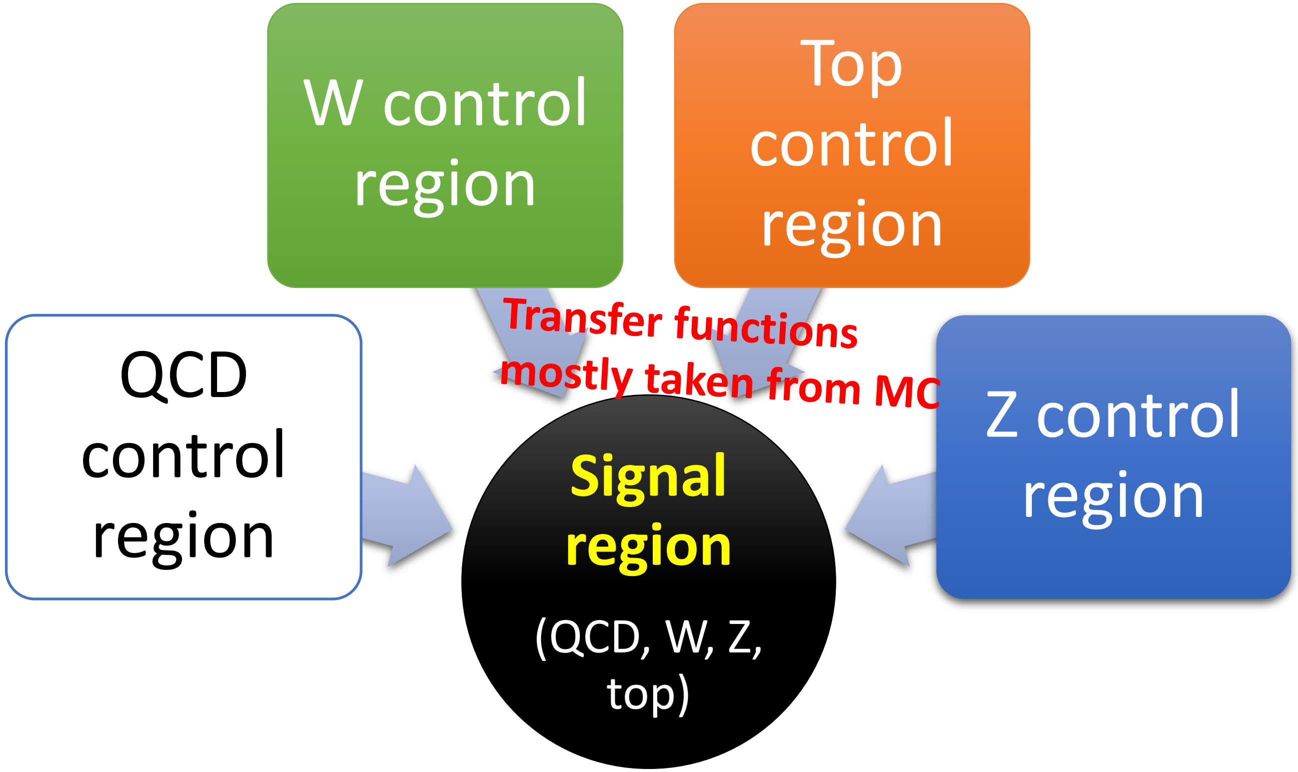
1. Pick a model of new physics
2. Simplify
3. Pick a likely (?) set of parameters
4. Make a prediction ➔ p_BSM(x)
5. Train classifier (p_BSM(x) vs p_SM(x)) to test the prediction
6. Hypothesis test with data|old model vs data|new model on classifier output
7. Exclude the model parameter point ?
8. Go to 3 or 1

- Best approach if the model + parameter set is true

→ Predicts the “right signal”

- Bad approach if the model + parameter set is wrong. How bad ?

SUSY Analysis model - control regions

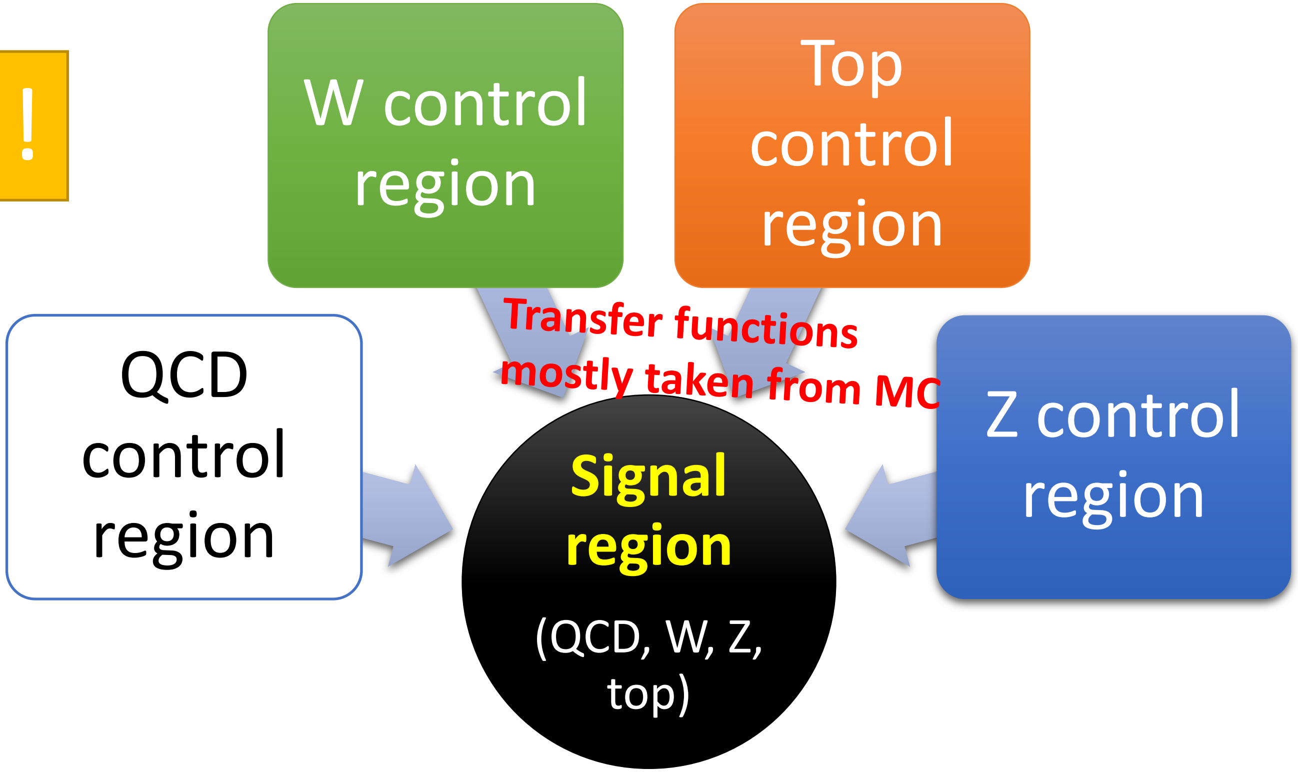


- Measure number of events in control selections
- Predict number of events in signal region via a fit to control regions
- Important : Test model and transfer functions
  (e.g. by alternative control regions or methods)

SUSY Analysis model - control regions

No bumps !!

SUSY
and Dark Matter
have no bumps.



- Measure number of events in control selections
- Predict number of events in signal region via a fit to control regions
- Important : Test model and transfer functions
  (e.g. by alternative control regions or methods)

SUSY Analysis model - control regions

Uncertainty estimated as in every other search !!

W control region

Top control

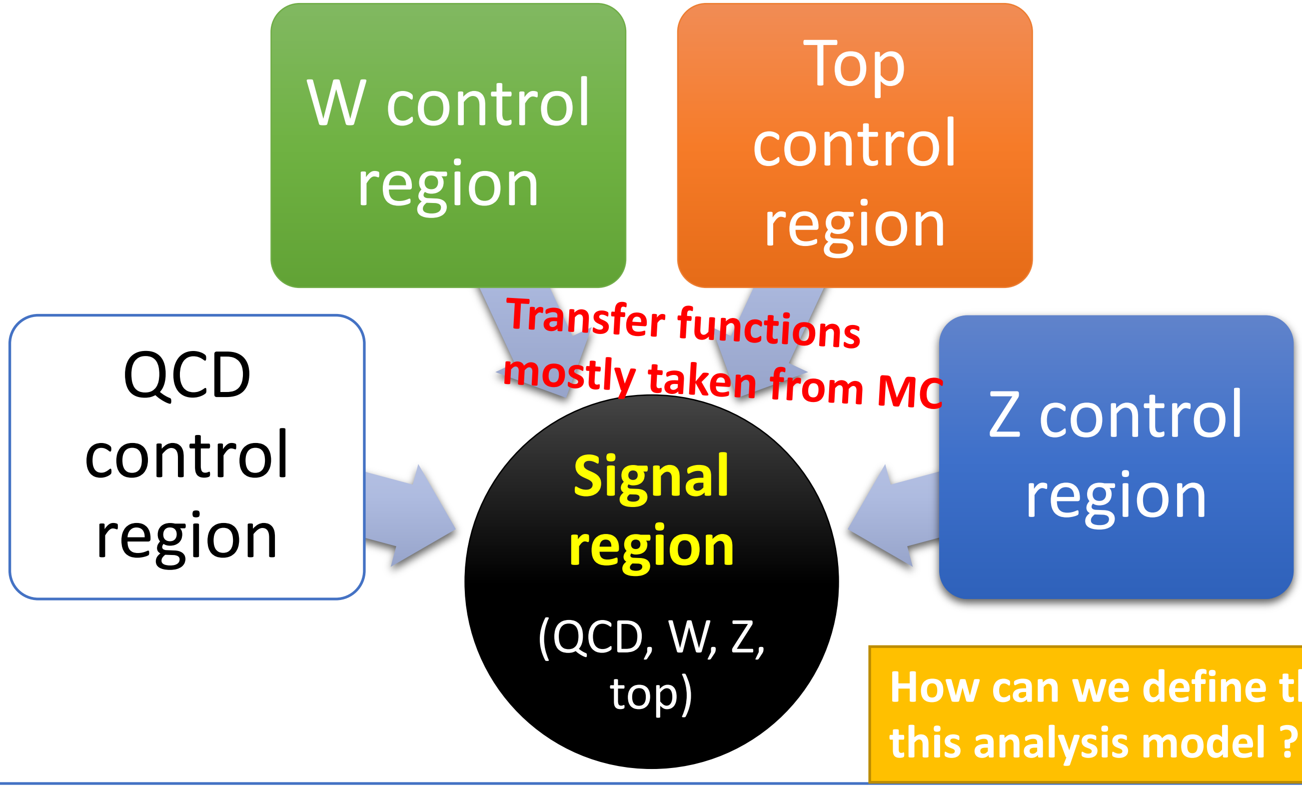
top)

…mental uncertainties:
- … efficiency
- …rgy scale and resolution
- …n energy scale and efficiency
- E_T^{miss} soft component
- b-tagging
- Luminosity
- pileup modelling

Theory uncertainties:
- Generator modelling (μ_F, μ_R, ME/PS matching, α_s scale choice when possible - otherwise compare generators)
- PS uncertainties (typically compare Pythia and Herwig)
- PDF choice

- Measure number of events in control selections
- Predict number of events in signal region via a fit to control regions
- Important : Test model and transfer functions
  (e.g. by alternative control regions or methods)

SUSY Analysis model - control regions



How can we define the signal regions for this analysis model ?

- Measure number of events in control selections
- Predict number of events in signal region via a fit to control regions
- Important : Test model and transfer functions
  (e.g. by alternative control regions or methods)

Idea: **Extend model-by-model supervised search for new physics**

What can we change / improve ?

Found 3 more directions (are there more?):

→ Look systematically in all data for new physics (brute force)

→ Hyper-class augmentation: Train a ML classifier on many models of new physics

→ Anomaly detection: Train ML classifier only on known physics

Brute force

- The brute force algorithm tries out all (many) possibilities till a significant signal is found.

Brute force: Many hypotheses …

Searching for new physics with ‚minimal/less‘ assumptions on the signal

Consequences:

Less signal assumptions → more hypothesis tests (multiple testing)
→ more/all channels and data selections

Implementations:

- Search with an "algorithm": **automatizing data selections and testing**
- **Automatize/Generalize** the construction of **the background model**

A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment

Goal:

Strategy paper. Generalize previous attempts.
Define a “meta-algorithm“ for
automated / generic / unsupervised LHC searches
Show with 2015 data that this is - in principle – possible at the LHC

https://arxiv.org/pdf/1807.07447.pdf

Also approach by CMS called Music: https://arxiv.org/abs/2010.02984
Previous approaches in H1, DO, CDF

A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment

Define a 2-step approach:

First put available resources on generality

Then use available resources to test most interesting deviations…

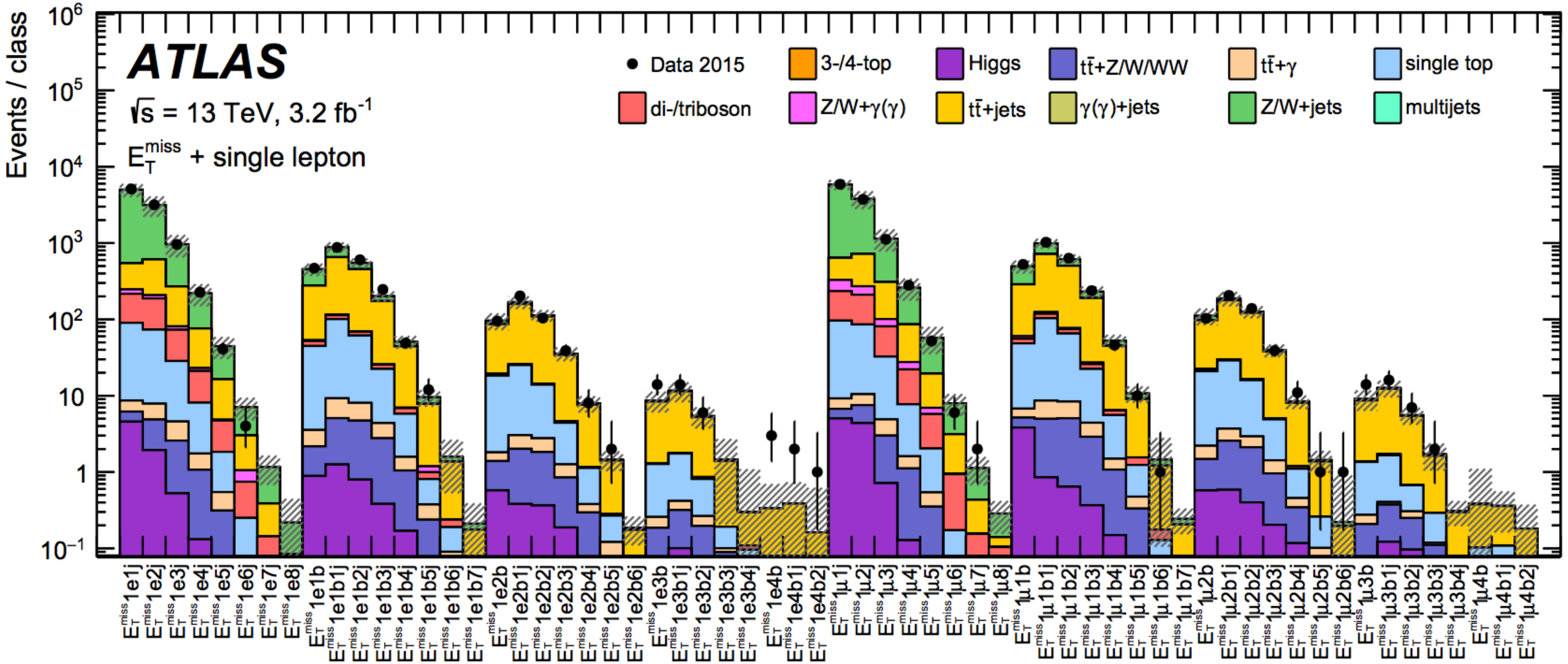
A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment

Define a 2-step approach:
First put **available resources on generality**
Then **use available resources to test most interesting deviations...**

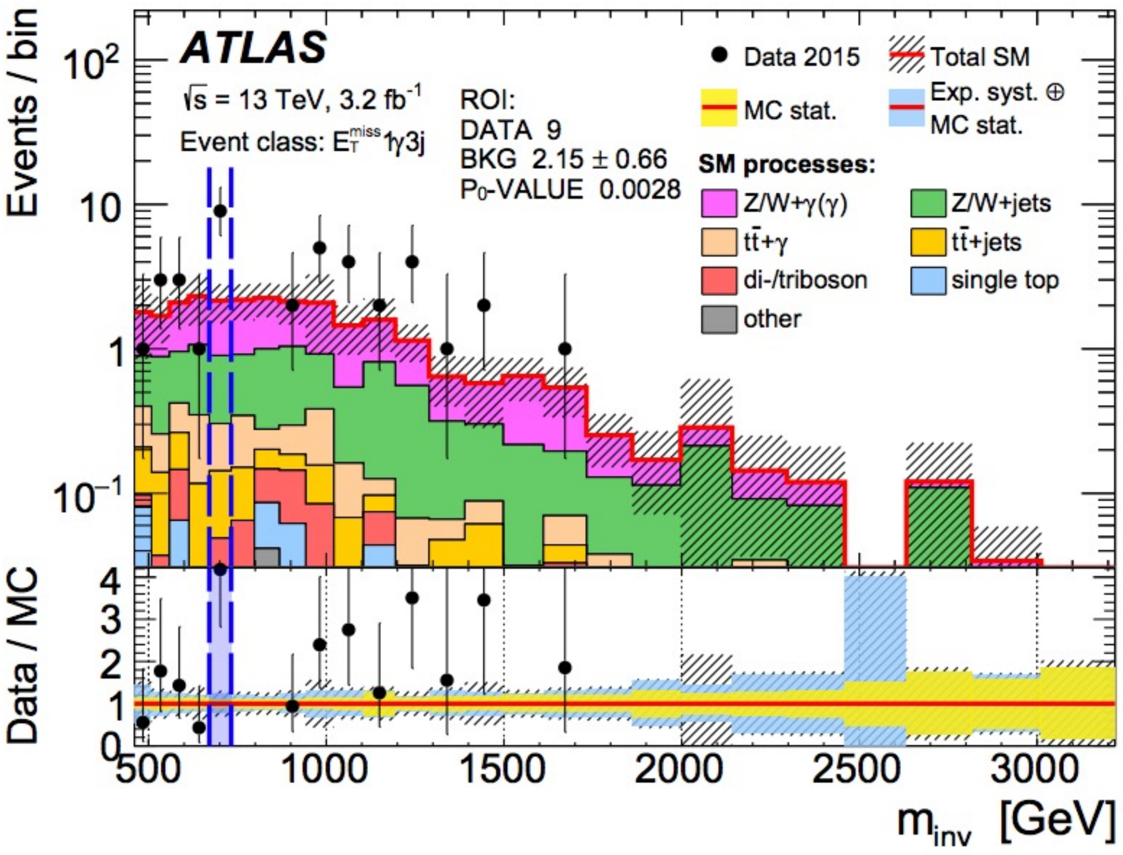
1. General Search: Automatically testing a large set of signal regions
Observation of one or more significant deviations in some phase-space region(s)

→ Trigger to perform dedicated and model-dependent analyses
where these **'data-derived' phase-space region(s) can be used as signal regions**

In ATLAS > 800 channels !
about 10^5 (correlated) signal regions/hypothesis tests !



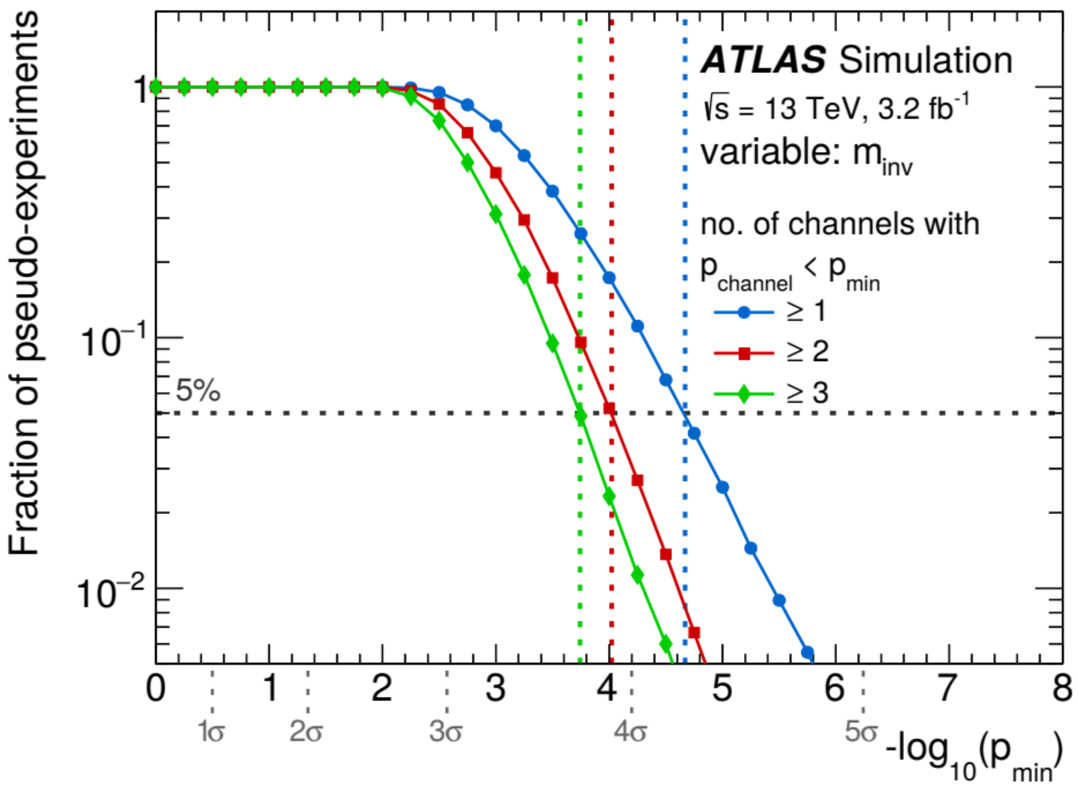
> 800 channels …. (plot shows a small selection)



> 30000 regions (hypothesis tests)

Determine p-value thresholds by asking how many toy datasets would give such a deviation

*→A regions is **interesting** if you find channels with p-values more significant than in 95% of the toys*



(yes, this is 5% and so high because of the trial-factor, note that we do not claim a discovery here, we just use this approach to select "signal regions" from data)

Outcome

0 signal region above threshold !

A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment

Define a 2-step approach:

First put available resources on generality

Then use available resources to test most interesting deviations…

1. General Search: ➔ Data derived-signal regions

2. Dedicated Search

- "Wave function collapsed" to test most interesting deviations with available resources on 2^{nd} dataset (➔ Statistically independent, unbiased p-value !!)

Advantage: ➔ Can make "traditional" control region analysis with 1^{st} and 2^{nd} dataset
1^{st} dataset corrected with trial factor, 2^{nd} dataset no need for correction

Why ? Lower resources, lower systematic uncertainties

Questions

When is the approach of dividing the data set into 2 optimal?

Minimizing available resources… (no time to check >2, would take more work), also mutual approach possible, resources (systematic uncertainties)

If there are n 'interesting deviations' in the first half, presumably the LEE factor is n.

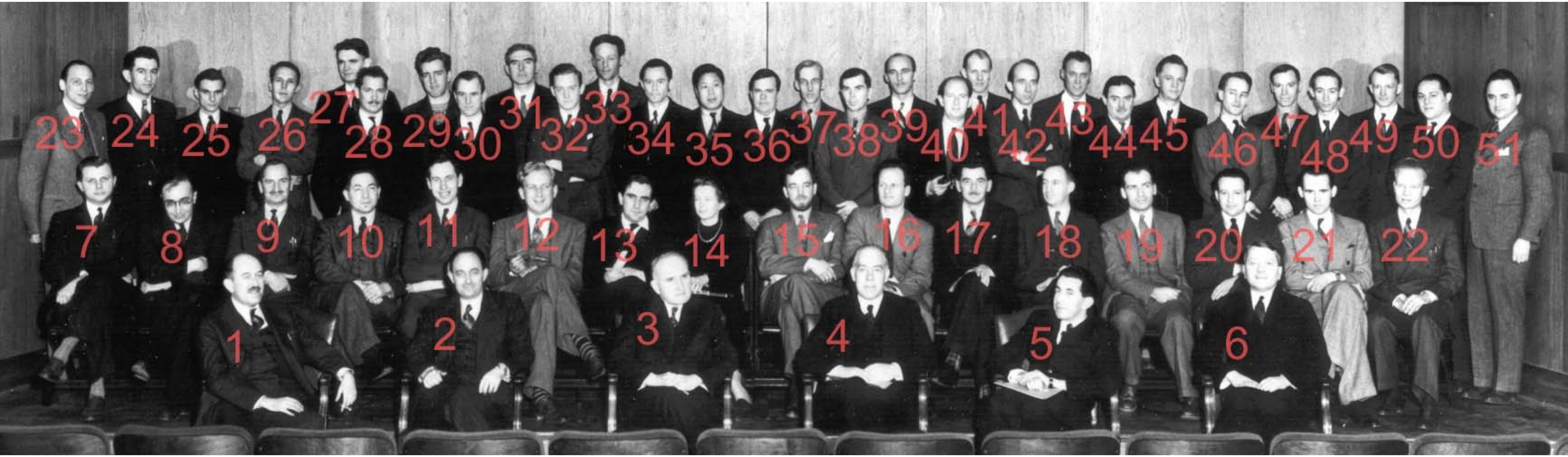
Yes, then we would define n "data-derived" signal regions and have a trial factor of n in the 2^{nd} half (Bonferoni)

New approach: Hyperclass of models

Search via “Hyperclass: Mixture of theories”

Assume the model/parameter set is not the correct one, but includes some knowledge about the new phenomenon we expect in the data..

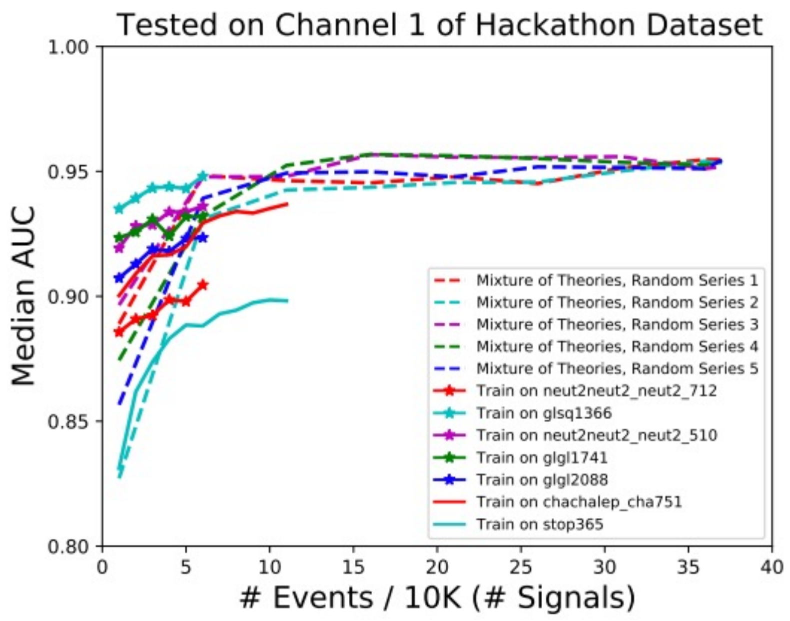
Maybe we should **mix the knowledge of the theory community**.

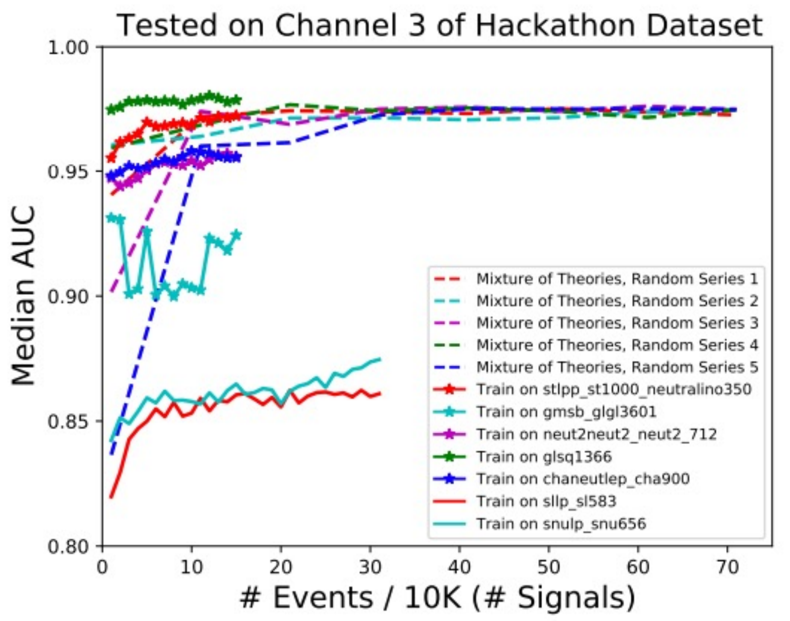


https://home.gwu.edu/~kargaltsev/HEA/washington-conferences.html

Our approach Model driven

1. Pick many “model of new physics”
2. Pick many likely (?) sets of parameters!
3. Make many predictions
4. Mix them
5. Train a classifier (NN, BDT) on $\sum_i^N w_i p_{S,i}(x)$ vs p_SM(x)
6. Hypothesis test in signal region data|SM





Mixture theories outperforms
“on average”
compared to single theory training

→ See later for comparison with
other approaches

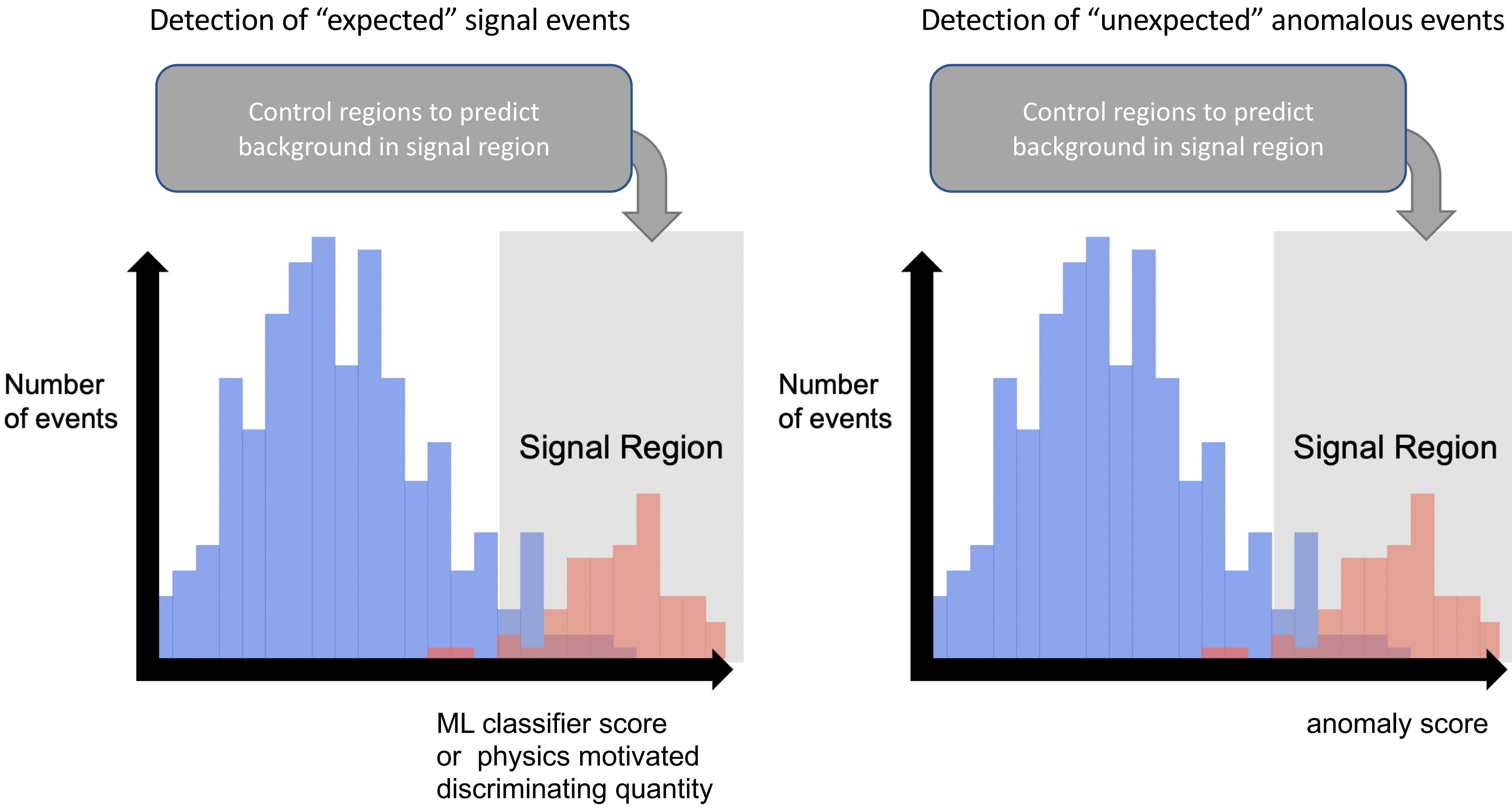
With Zhongyi Zhang, Roberto di Austri

Anomaly detection: Out of distribution

Anomaly detection

1. Pick **no** “new physics model"
2. Learn the background model
3. Train ML classifier to test the prediction (is event background or not?)
4. Hypothesis test with data|background model on classifier output
5. Exclude the background model?

In which variable should you search?
Need a variable to "flag" an outlier


Detection of “expected” signal events
Control regions to predict background in signal region
Number of events
Signal Region
ML classifier score or physics motivated discriminating quantity
Detection of “unexpected” anomalous events
Control regions to predict background in signal region
Number of events
Signal Region
anomaly score

Advantages

Minimal changes to old approach

You could just "add" a new signal region to your analysis

Background prediction via transfer functions, control regions etc.

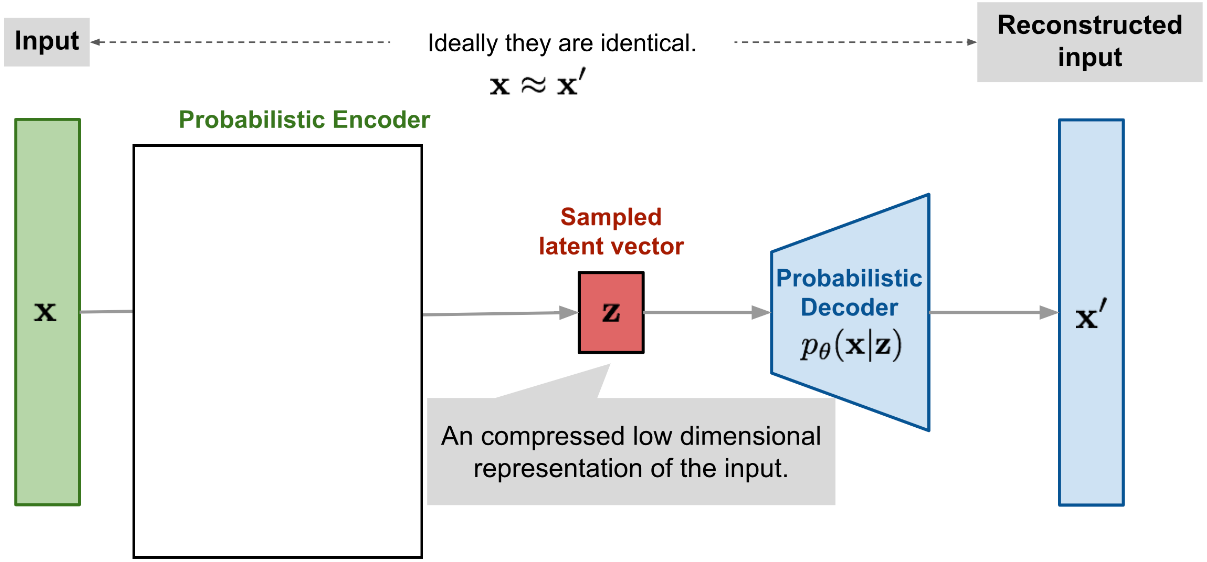
No *extra* Look-elsewhere effect (Why ? → only 1 more statistical test in the new "anomaly SR")

No training of NN data vs SM prediction needed

How to define anomalies ? ML approaches

2018: The new standard approach

Various papers on arxiv now proposing this → Autoencoder



Idea:
Signal region is region outside the SM, i.e. simulation

Then determine a distance between x and x' , e.g. MSE = (x-x')^2

But various other possibilities… needs comparison etc.

Is the data in the simulation ?

- Autoencoder:

data → Simulation^-1 → code → Simulation → data’

→Is data = data’ or distance in latent space from target

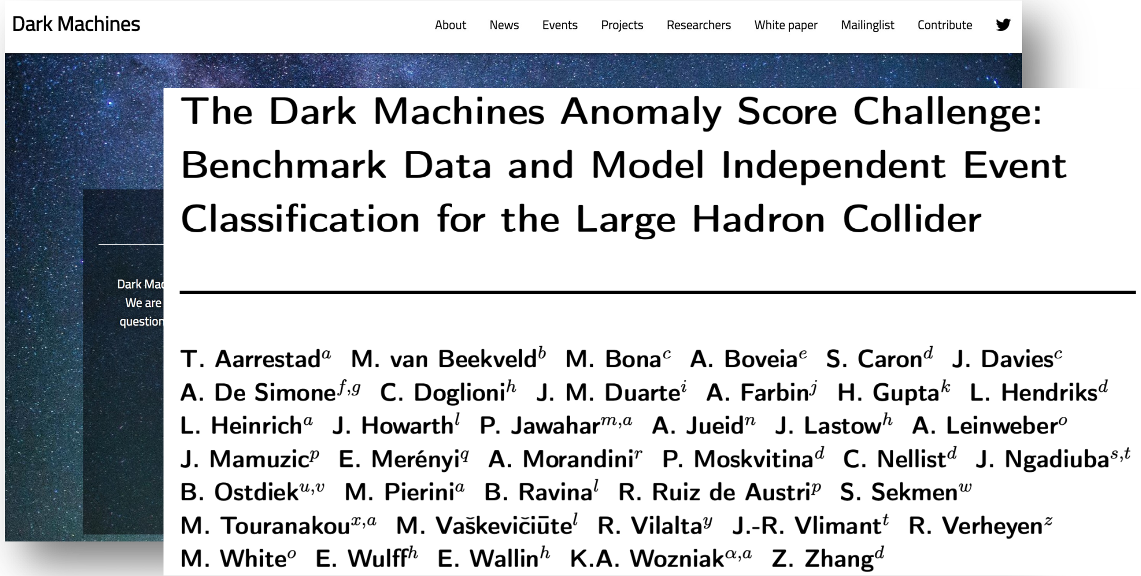
→Is this a good question ?

→Is this the best approach ?

→Comparison

Comparisons of approaches

Darkmachines (www.darkmachines.org) anomaly score challenge:
Objective → compare different approaches to define an "event- by-event" anomaly score



Event data:
4-vectors, jets, leptons, charge, photons

Different to
LHC Olympics (full signal and bump hunting / density comparisons with a few signals + background expectation)
→ *Talk by Georg*

Contact persons: Comparisons: B. Ostdiek (bostdiek@g.harvard.edu)
Datasets: M. van Beekveld (melissa.vanbeekveld@physics.ox.ac.uk)

Results (on arxiv https://arxiv.org/abs/2105.14027)

Compared performance of **>20 methods** to define anomalies
With > 1000 hyperparameter settings (i.e. algorithms to define anomalies)
Using
>20 signals
Using
> **1 Billion LHC events**
Using
A secret dataset (labels are still blind, *only Melissa van Beekveld (Oxford) knows*)

Task: Classify 100000s of events as SM or not by assigning **a score between 0 and 1…**

Figure of merit: **By how much can we improve the significance for that signal i.e. Significance Improvement SI per signal**

Organizers:
C. Doglioni, M. Pierini, S.C

$$\sigma'_S = \frac{S'}{\sqrt{B'}} = \frac{\epsilon_S S}{\sqrt{\epsilon_B B}} = \frac{\epsilon_S}{\sqrt{\epsilon_B}}\sigma_S \quad \Rightarrow \quad \mathrm{SI} \equiv \frac{\epsilon_S}{\sqrt{\epsilon_B}},$$

Many signals
many algorithms
many channels

Charginos

Gluinos

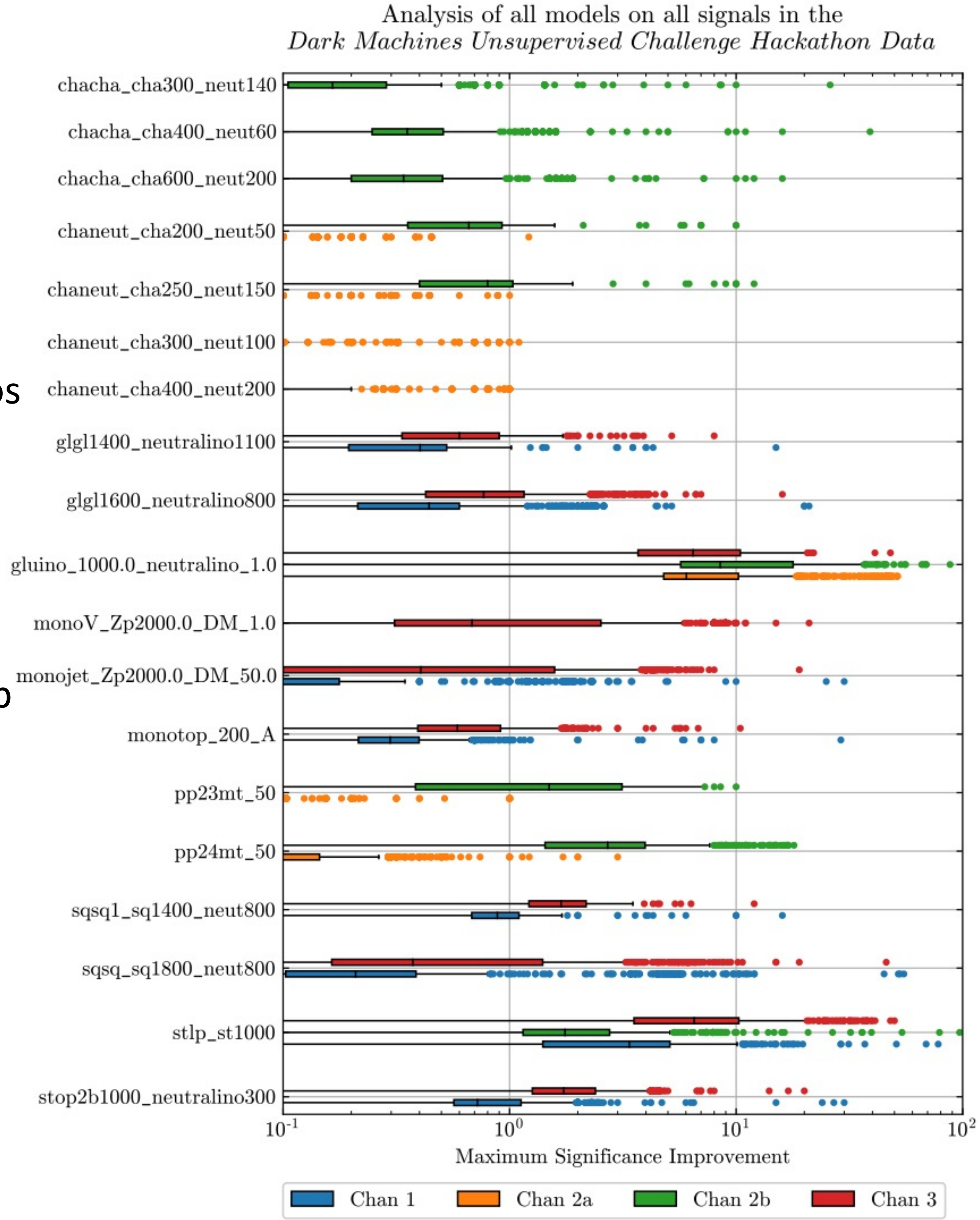
Zprime

Monotop

Squarks

Stops

....

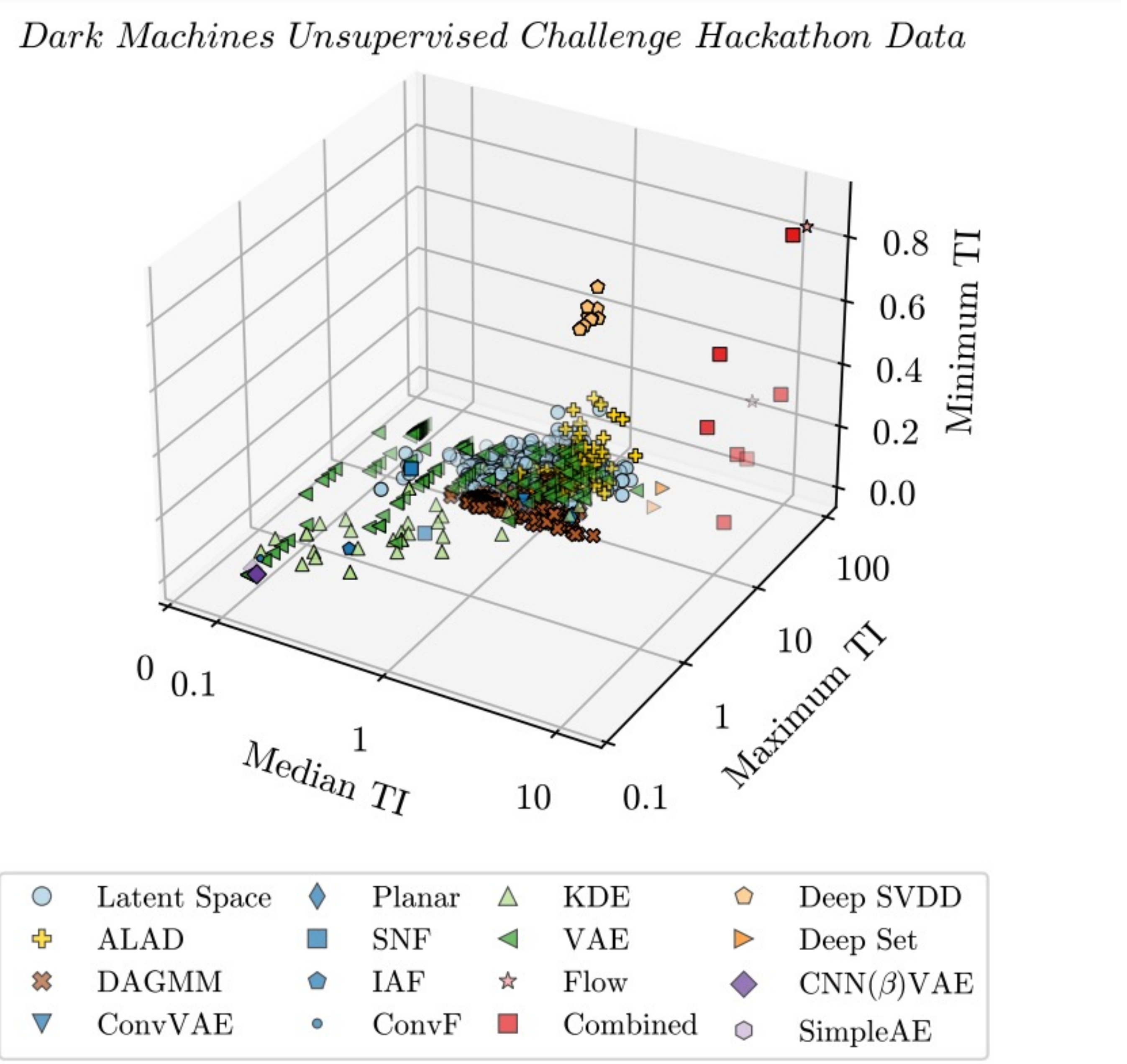


Summary plot

TI =
Total Improvement. (over many signals)

(median, max and min Improvement of many toy signals)

→ **Good algorithms have large max, min and mean TI**



https://arxiv.org/abs/2105.14027

Summary plot

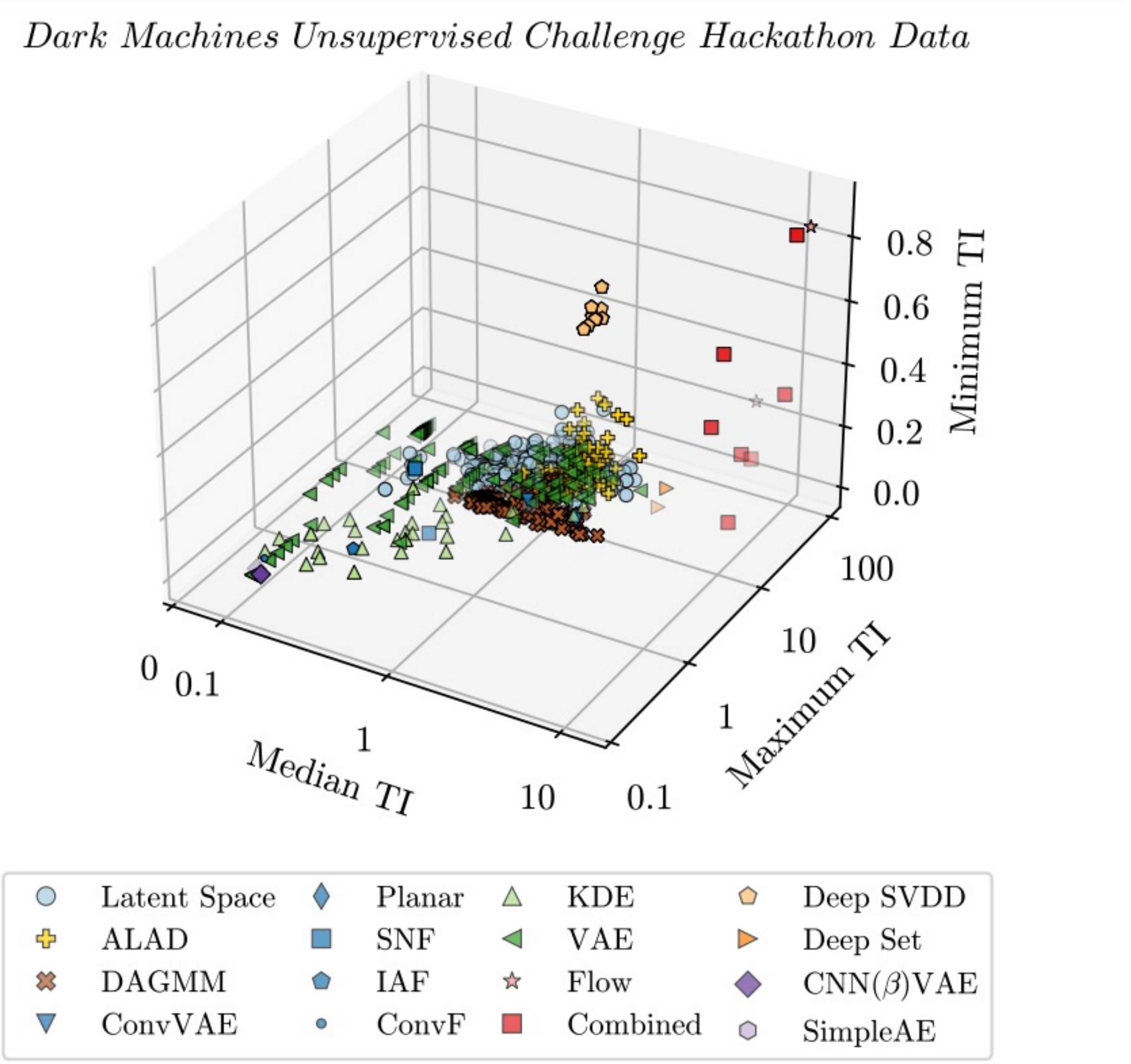
TI =
Total Improvement. (over many signals)

(median, max and min Improvement of many toy signals)

→ **Good algorithms have large max, min and mean TI**

→ ***DeepSVDD, Flow , Combined, DeepSets***
largely outperform traditional approaches (e.g. KDE), but also all autoencoder and VAEs !!

Why ? --> decoder seems not to be needed!



https://arxiv.org/abs/2105.14027

Rare and Different

Idea:

Anomalies can be either rare, meaning that these events are a minority in the normal dataset, or different, meaning they have values that are not inside the dataset.

We **quantify and combine these two properties/objectives**

https://arxiv.org/abs/2106.10164

Rare and Different

- A- **density wise**: events that have a low likelihood as determined by a (ML-)model that knows the **likelihood p(x)**

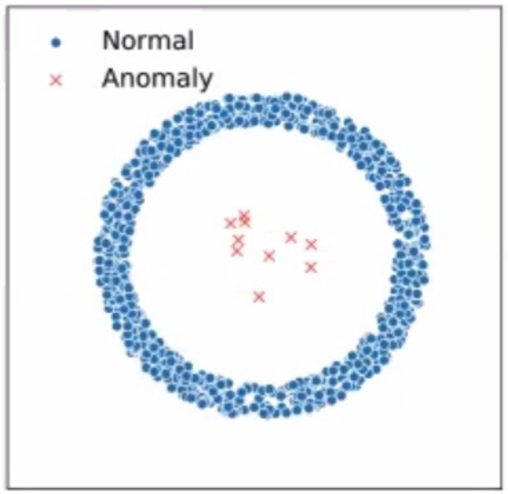
- B- **event wise/out of manifold:**

Is the event on the SM manifold (yes/no) ? **One class classification.**

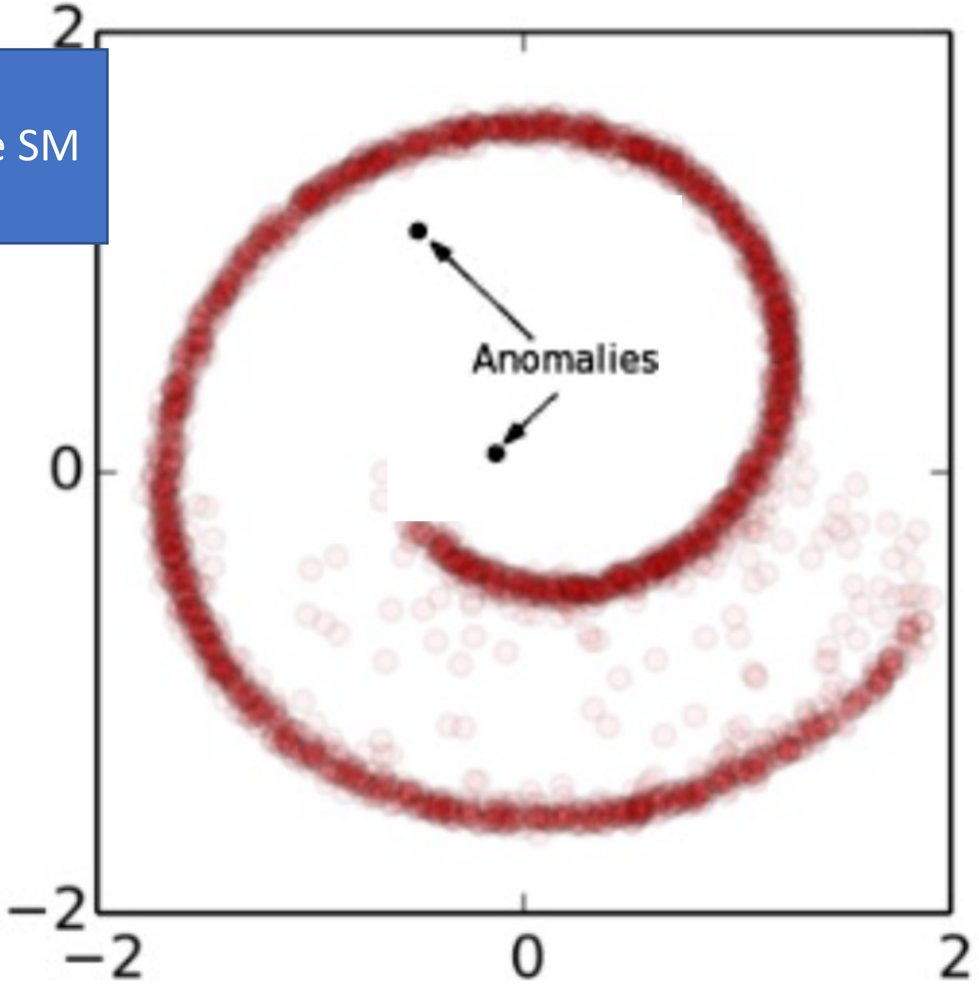
https://inspirehep.net/literature/1869277

With Luc Hendriks, Rob Verheyen

Rare → Density estimation



Idea:
Signal region is region outside the SM /simulation

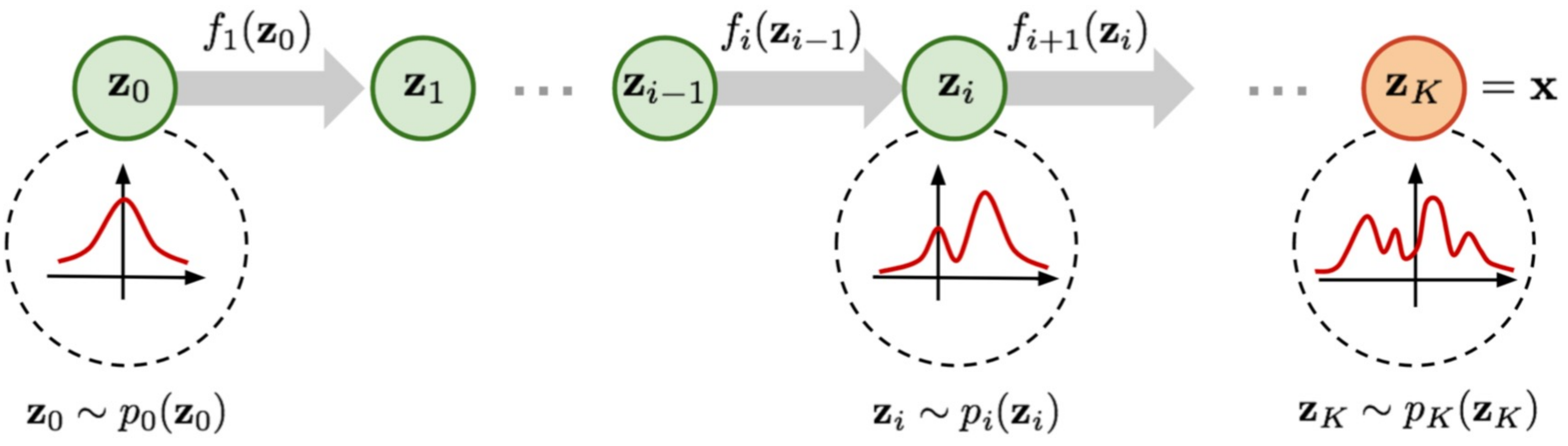


Rob Verheyen : Surjective normalizing flows work even better as anomaly detectors…
→ https://inspirehep.net/literature/2077178

Bob Stienen, Rob Verheyen

Rare ?

Our encoding of the likelihoods: Flow models



f_i are bijectors , have a known inverse

Jakobian can be calculated →
Try to use this to estimate likelihood and anomaly score:

$$s(x) = \frac{\log p(x) - \log p_{\min}}{\log p_{\max} - \log p_{\min}}$$

(we use the MADE network with rational quadratic splines as bijectors)

In particular, starting from a simple prior distribution $p_0(z_0)$, subsequent latent variables z_{i+1} are determined as

$$z_{i+1} = f_{i+1}(z_i, \theta_i), \tag{4}$$

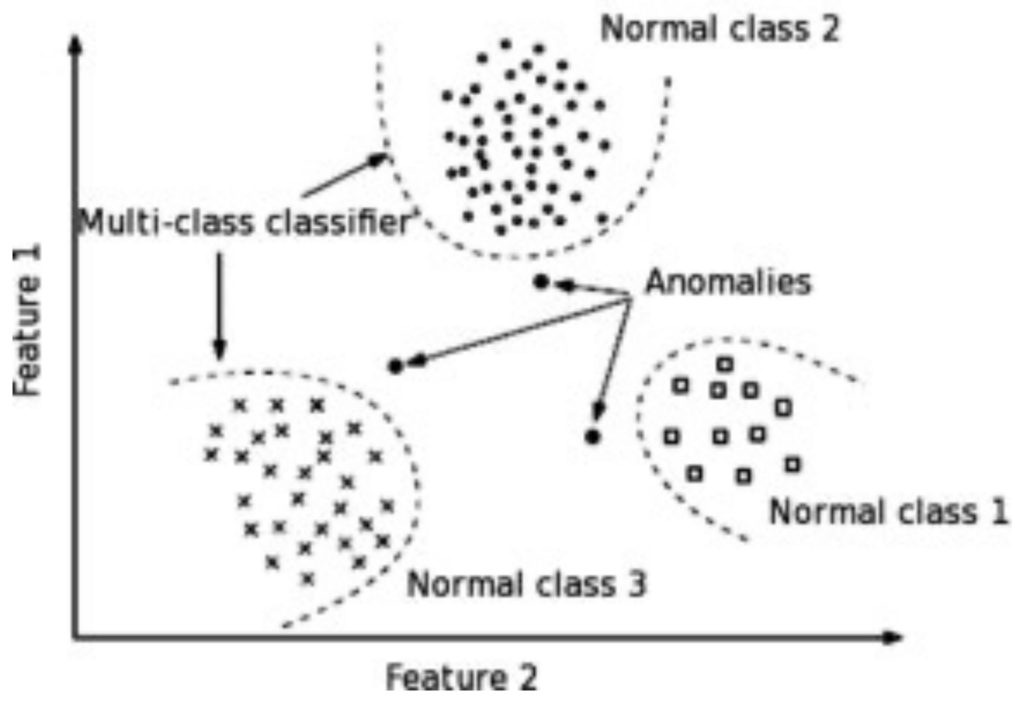
where θ_i are parameters inferred during training. The likelihood is transformed as

$$p_{i+1}(z_{i+1}) = p_i(z_i) \left| \det \frac{\partial z_{i+1}}{\partial z_i} \right|^{-1} \tag{5}$$

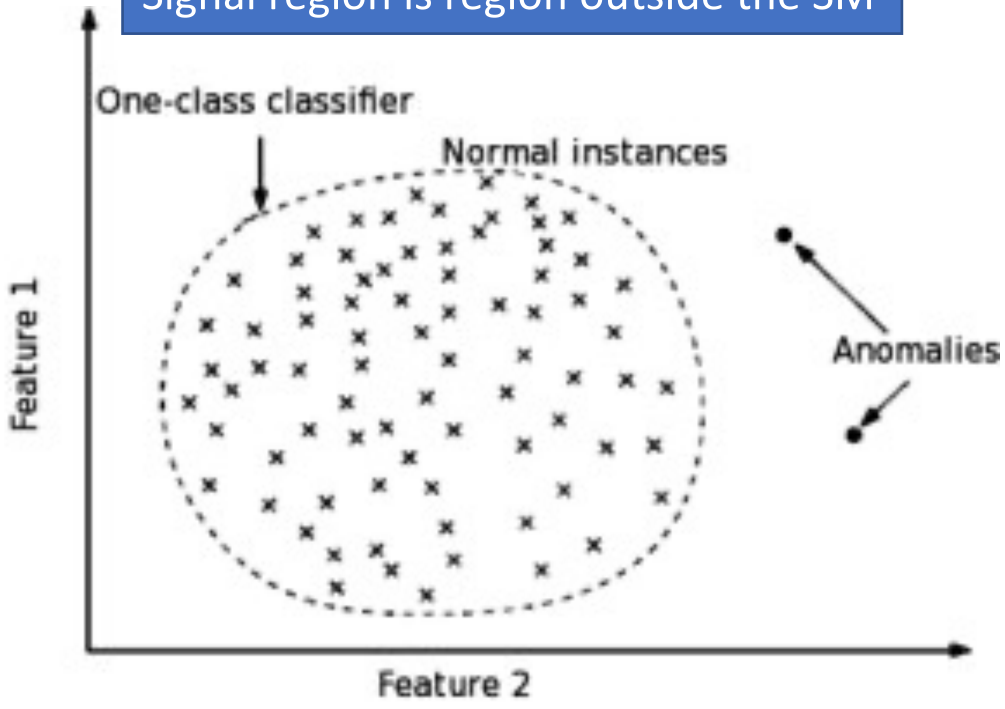
Identifying the last latent dimension z_n with the data x, the likelihood may be evaluated by propagating data backwards through the model, such that

$$\begin{aligned} \log p(x) &\equiv \log p_n(z_n) \\ &= \log p_0(z_0) + \sum_{i=0}^{n-1} \log \left| \det \frac{\partial z_{i+1}}{\partial z_i} \right|^{-1} \end{aligned} \tag{6}$$

Different ? One class classification



Idea:
Signal region is region outside the SM



Different? Deep SVDD

Alternatively one could try to pass the events through a trained “filter” that only allows events to pass if they belong to the training data

Here: Deep SVDD

X → Network → 42

Anomaly score:

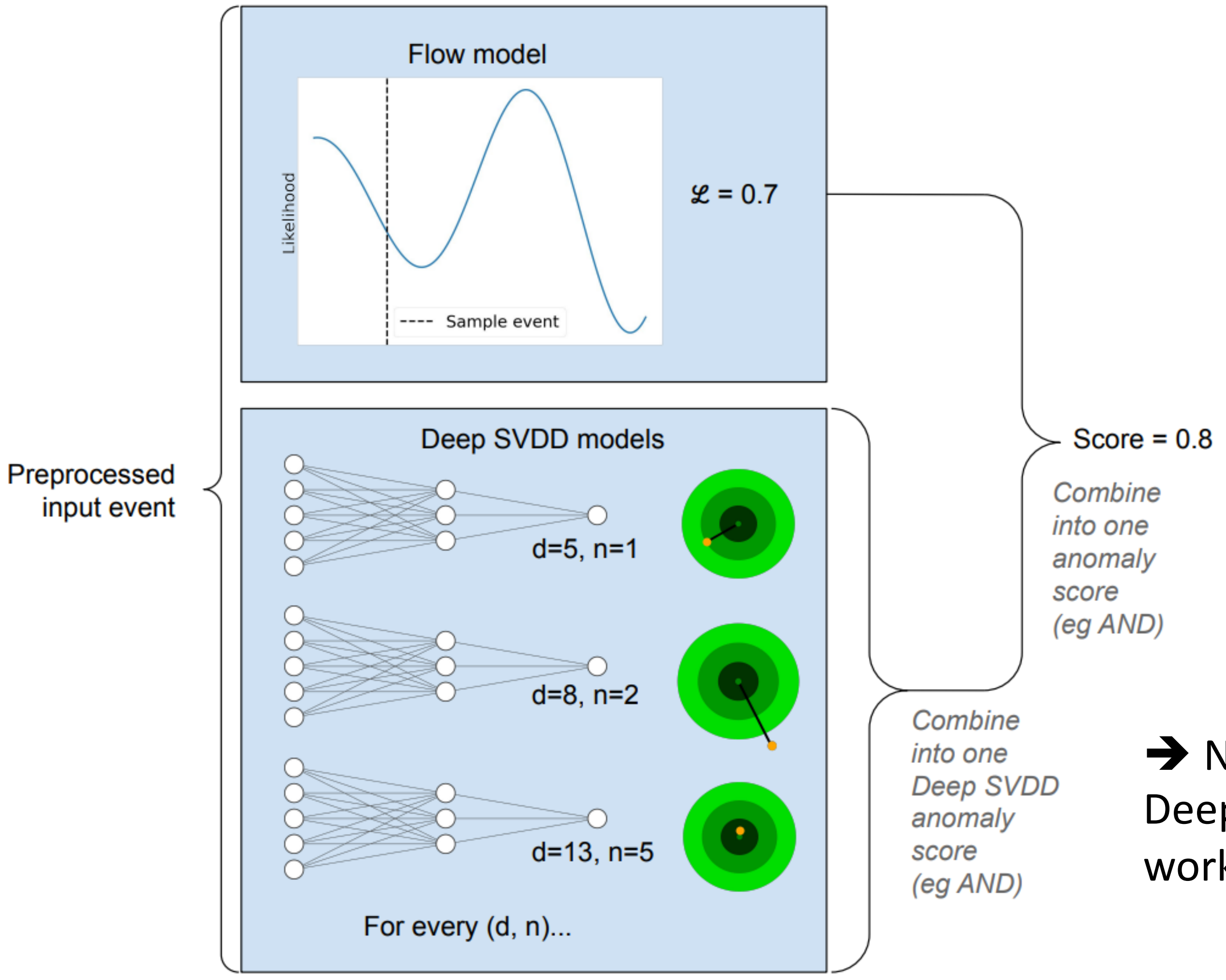
Difference from 42 !

The Deep SVDD network is similar to the encoder component of an autoencoder. The loss is defined as

$$s(x) = O_n^d - \mathrm{Model}(x), \tag{3}$$

where the model maps the input x to the same tensor shape as the manifold O. In our case, O is a vector of identical scalar values, with the subscript n defining the scalar value and superscript d the number of elements in the vector. For example, O_3^4 identifies the vector $(3, 3, 3, 3)$. The optimisation of the Deep SVDD model is fundamentally very simple: it is a NN that receives some input x and transforms it to some output O_n^d.

Rare and Different



➔ Need an ensemble of Deep SVDDs to make it work

https://arxiv.org/abs/2106.10164

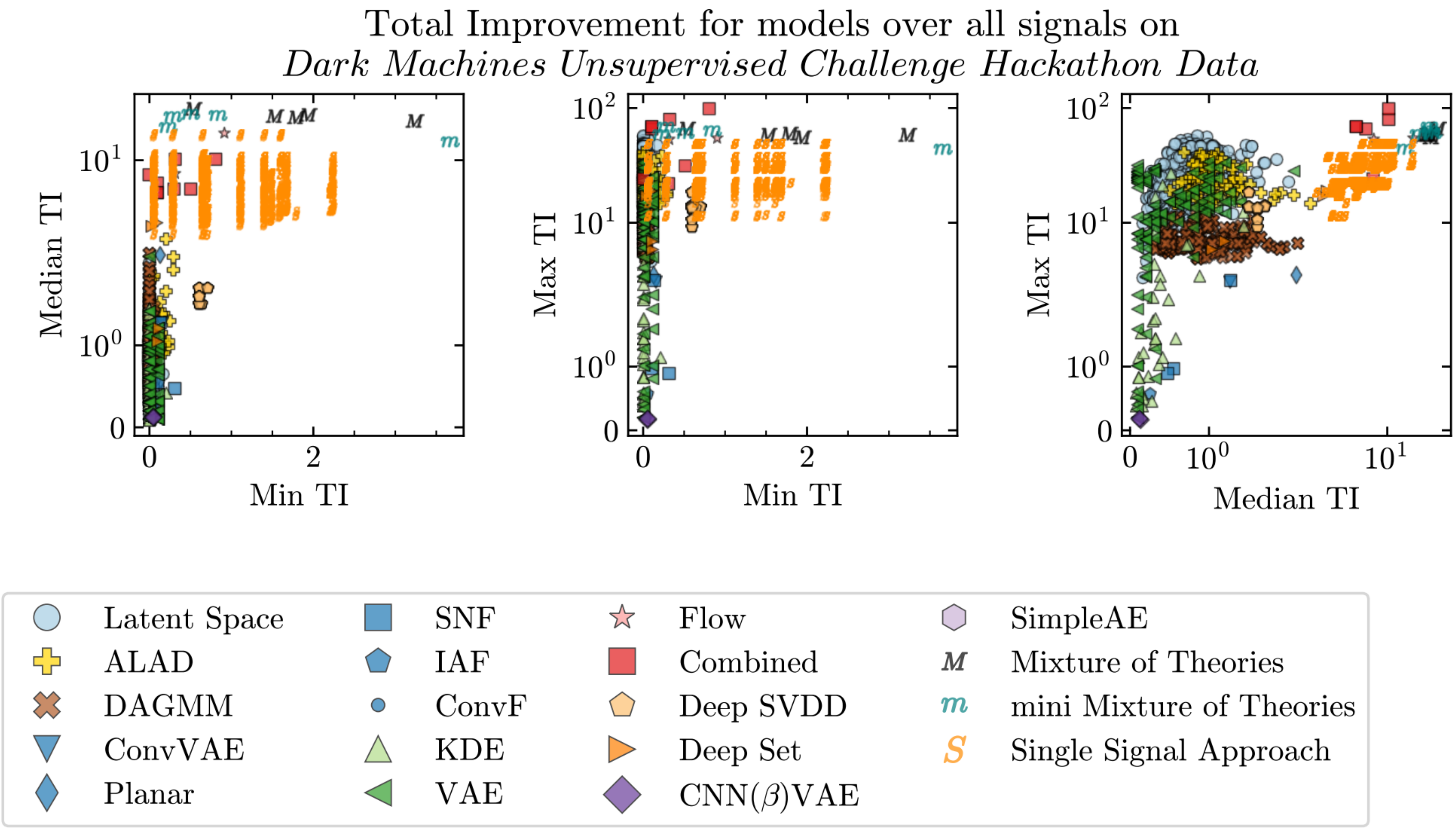
Compare them all (besides brute force)

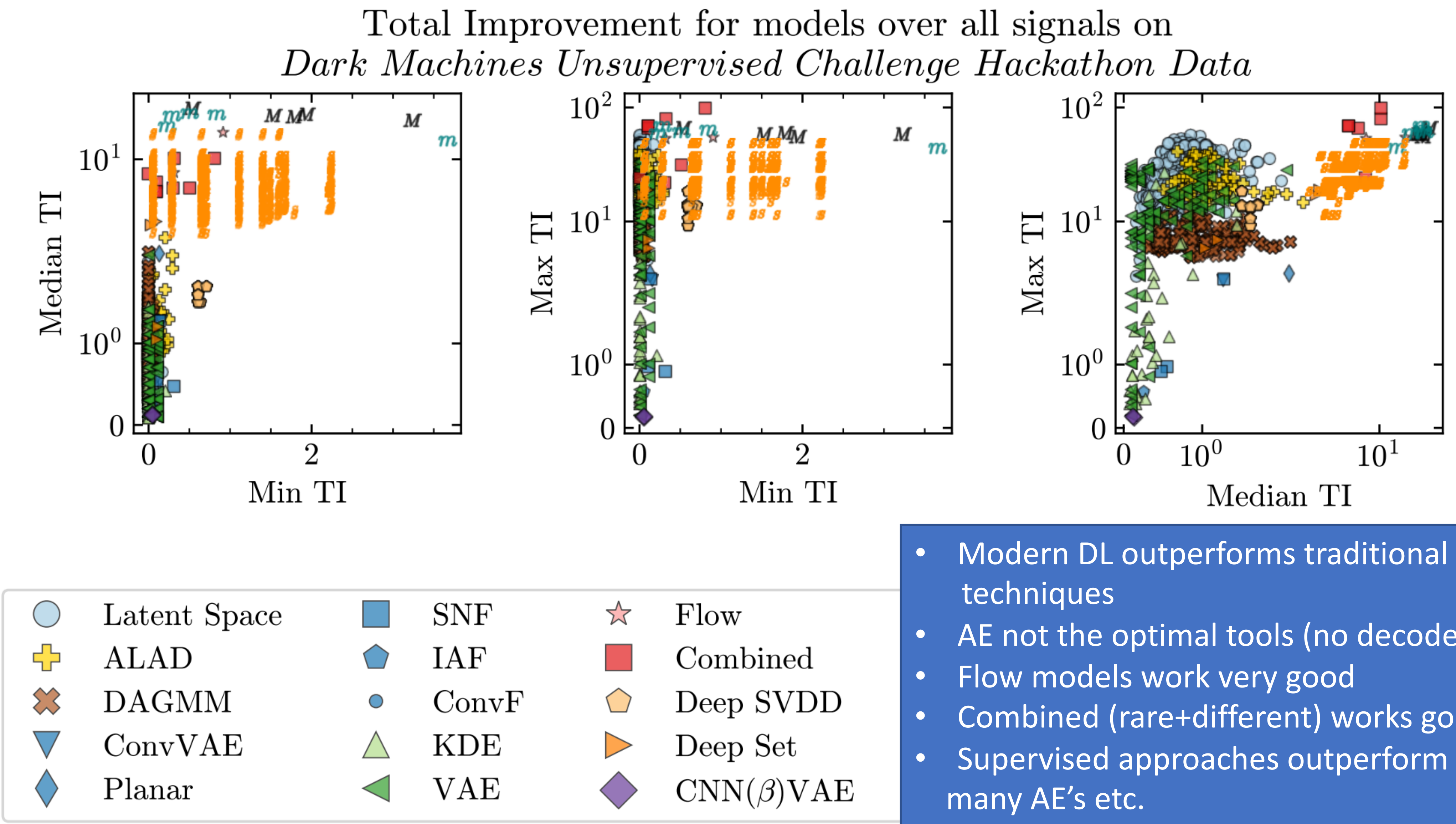
Compare them all

Compared:

- Supervise approaches (100s trained on different “single” signals)
- Mixture of Theory approach
- Unsupervised approaches

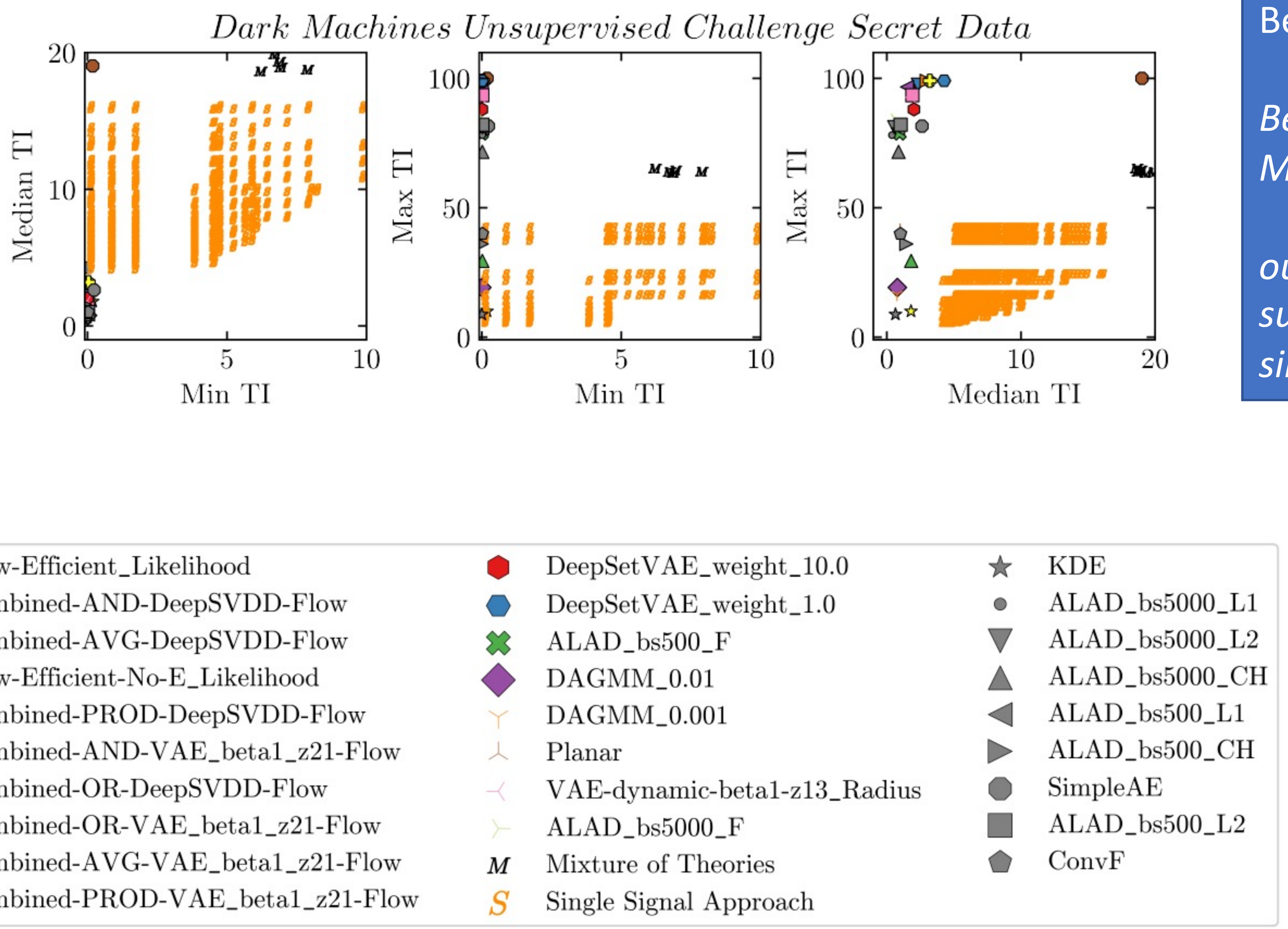
Who wins?


Total Improvement for models over all signals on
Dark Machines Unsupervised Challenge Hackathon Data
Median TI
Min TI
Max TI
Min TI
Max TI
Median TI
Latent Space
ALAD
DAGMM
ConvVAE
Planar
SNF
IAF
ConvF
KDE
VAE
Flow
Combined
Deep SVDD
Deep Set
CNN(β)VAE
SimpleAE
Mixture of Theories
mini Mixture of Theories
Single Signal Approach



- Modern DL outperforms traditional techniques
- AE not the optimal tools (no decoder needed)
- Flow models work very good
- Combined (rare+different) works good
- Supervised approaches outperform many AE's etc.
- Mixed signal approach outperform all supervised approaches

Secret dataset!!



Best:

Best unsupervised Mixed-model

outperform supervised and simple unsupervised

Summary

- **Searching for the unknown**

- **Exploring different methods to define signal regions**

- Brute force / General Search
https://www.nature.com/articles/d41586-018-05972-7
- Anomaly Scores → Darkmachines
https://cerncourier.com/a/whats-in-the-box/
- Hyper-data of theories → Upcoming !

Apply them all ?

Extra Slides