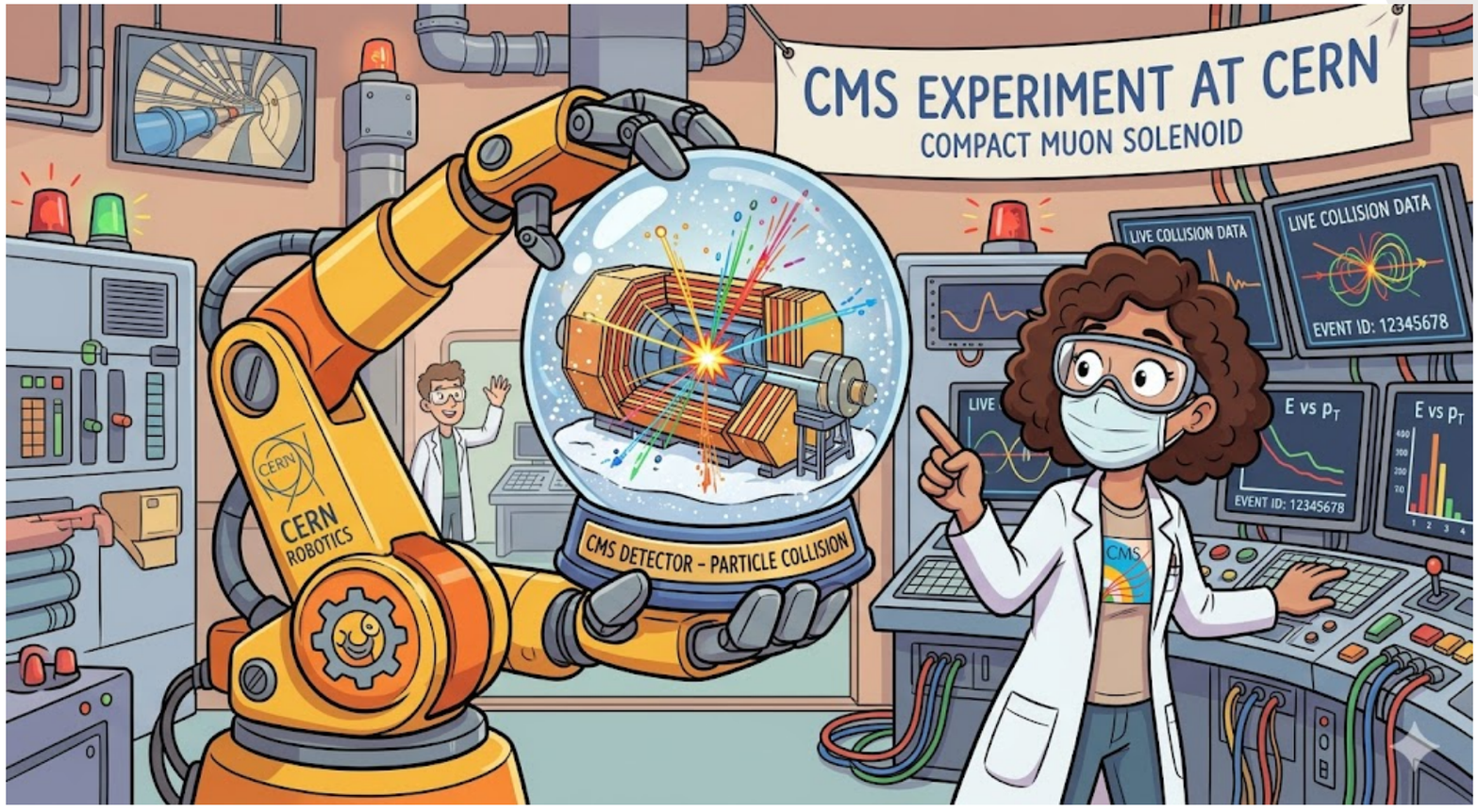
Accelerated AI Algorithms for Data-Driven Discovery



Automating Particle (and Astro) Physics research w/AI

Phil Harris

Phil Harris Lab@MIT

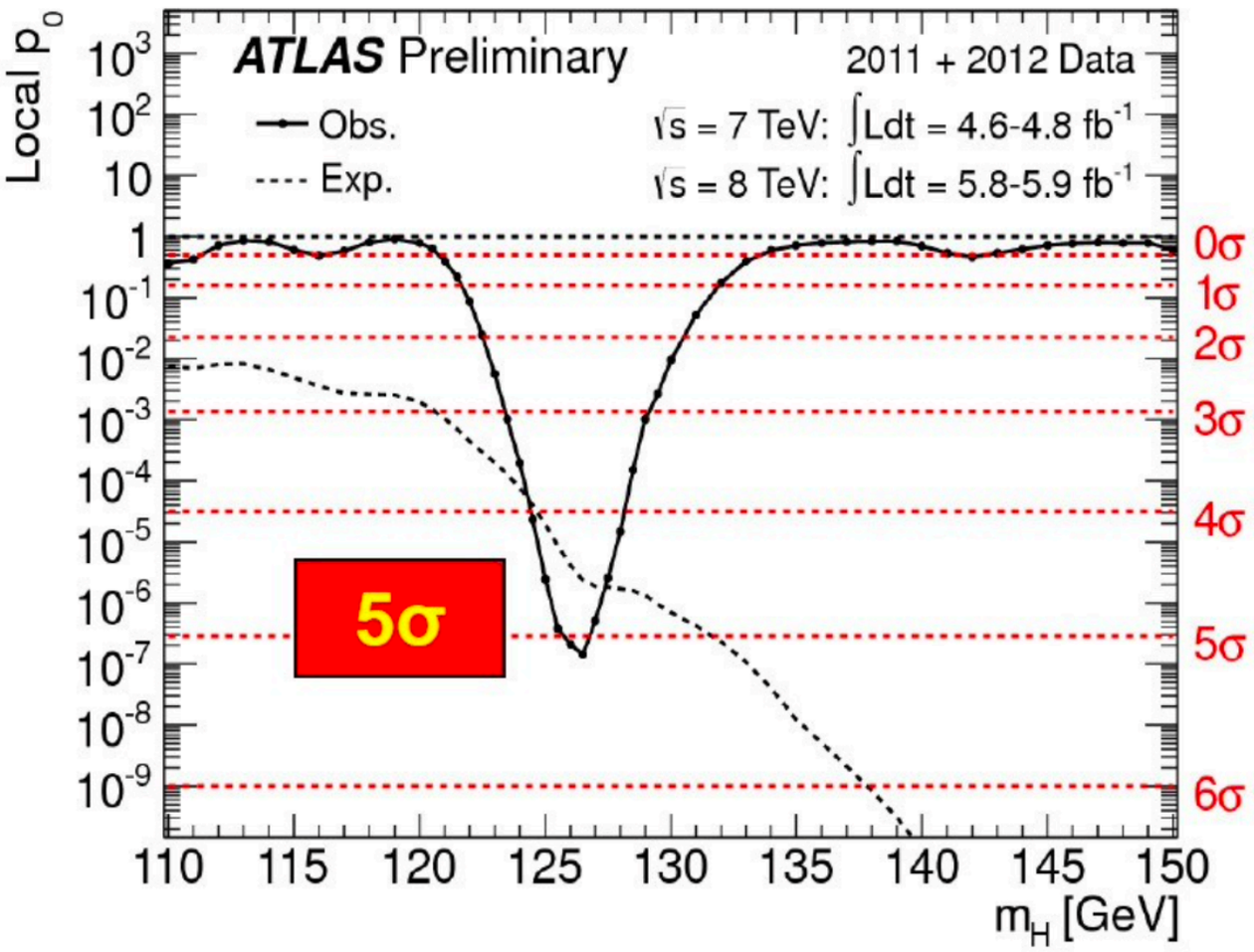
What you may not know?

Or Remember

At the Higgs discovery

ATLAS

γγ, 4l updated with ~6 fb^{-1} of 8 TeV data

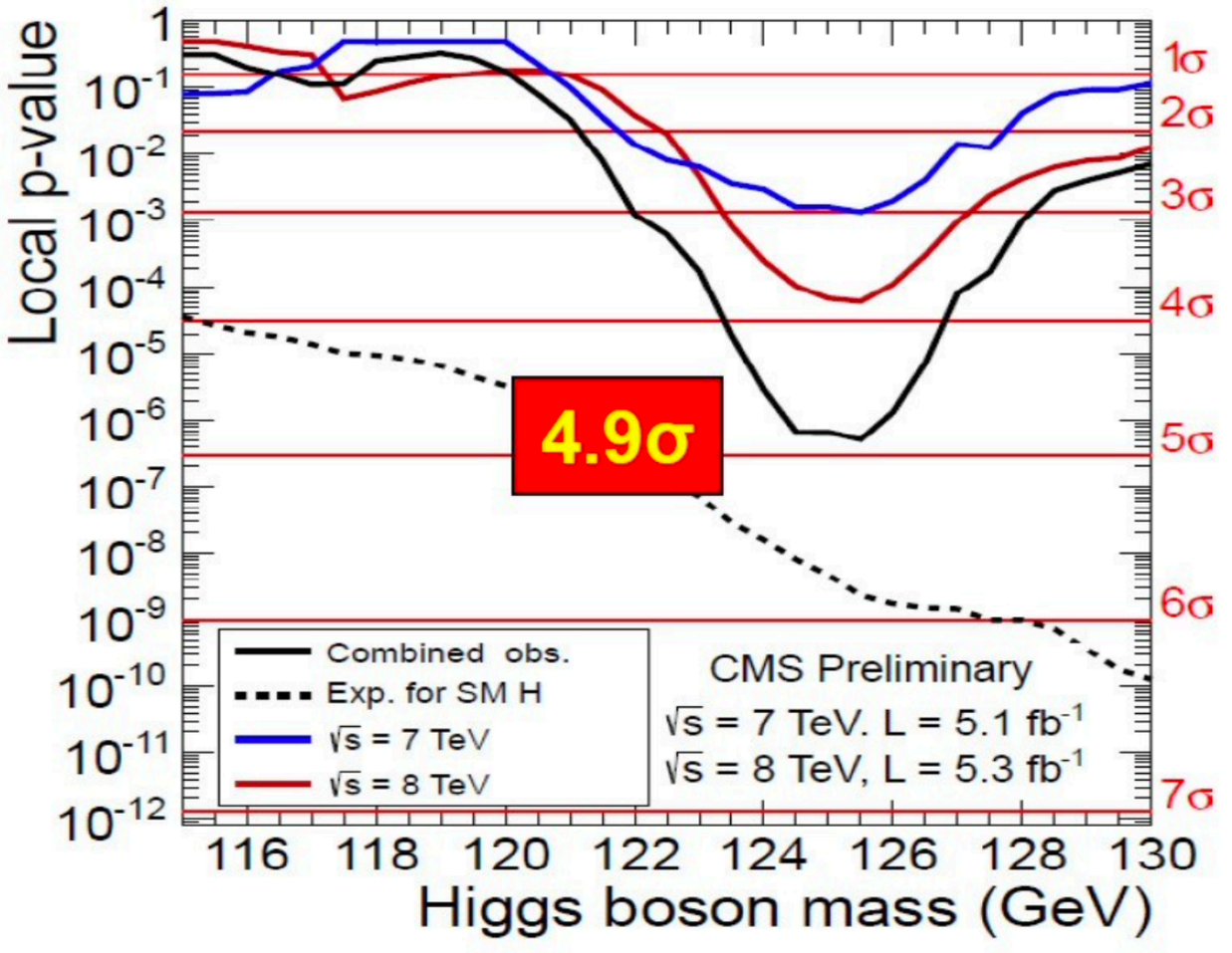


Largest local excess:

5σ at m_H= 126.5 GeV

CMS

All channels updated with ~5 fb^{-1} of 8 TeV data



Largest local excess:

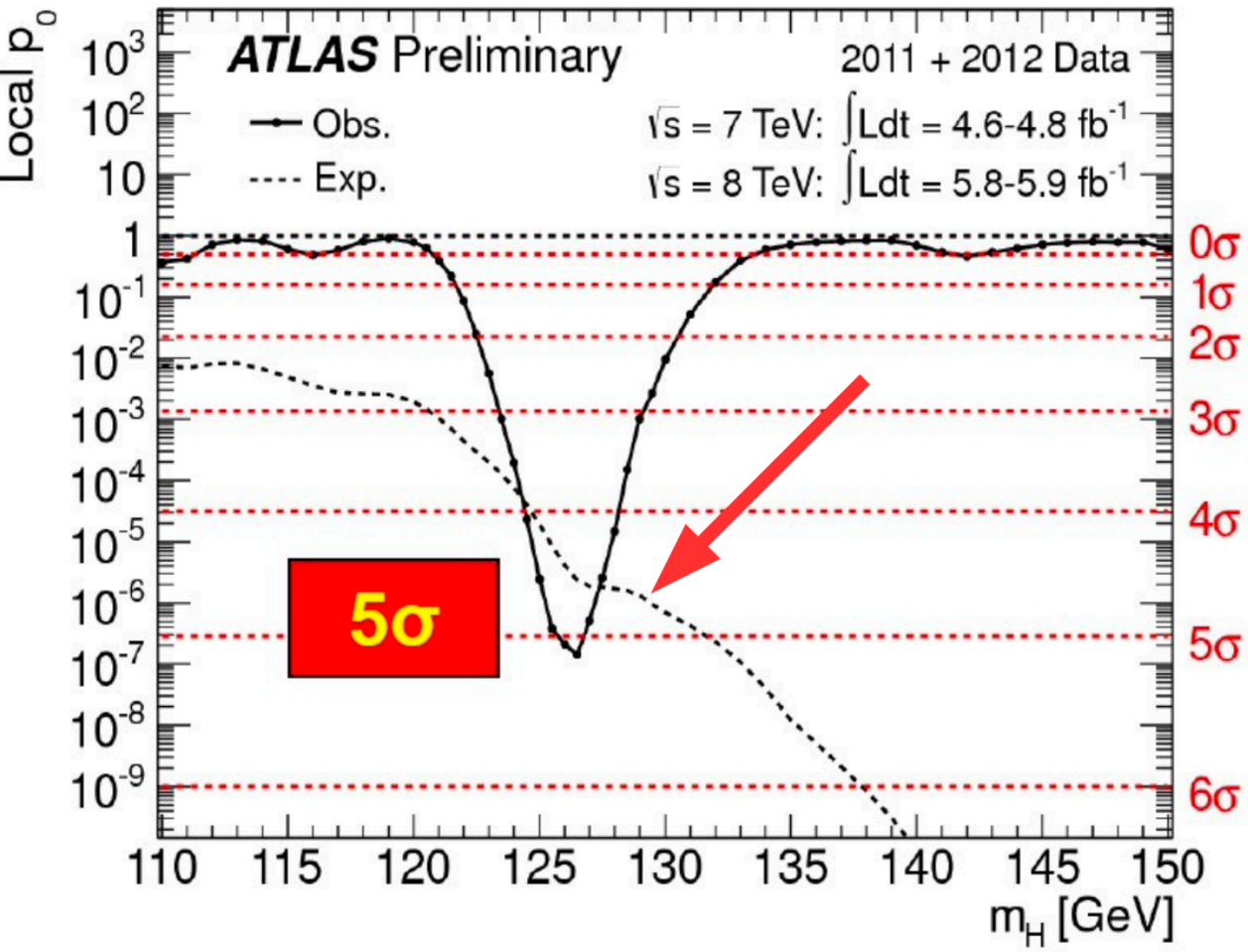
4.9σ around m_H= 125 GeV

(using H→γγ and H→4l: 5.0σ)

A big difference was present

ATLAS

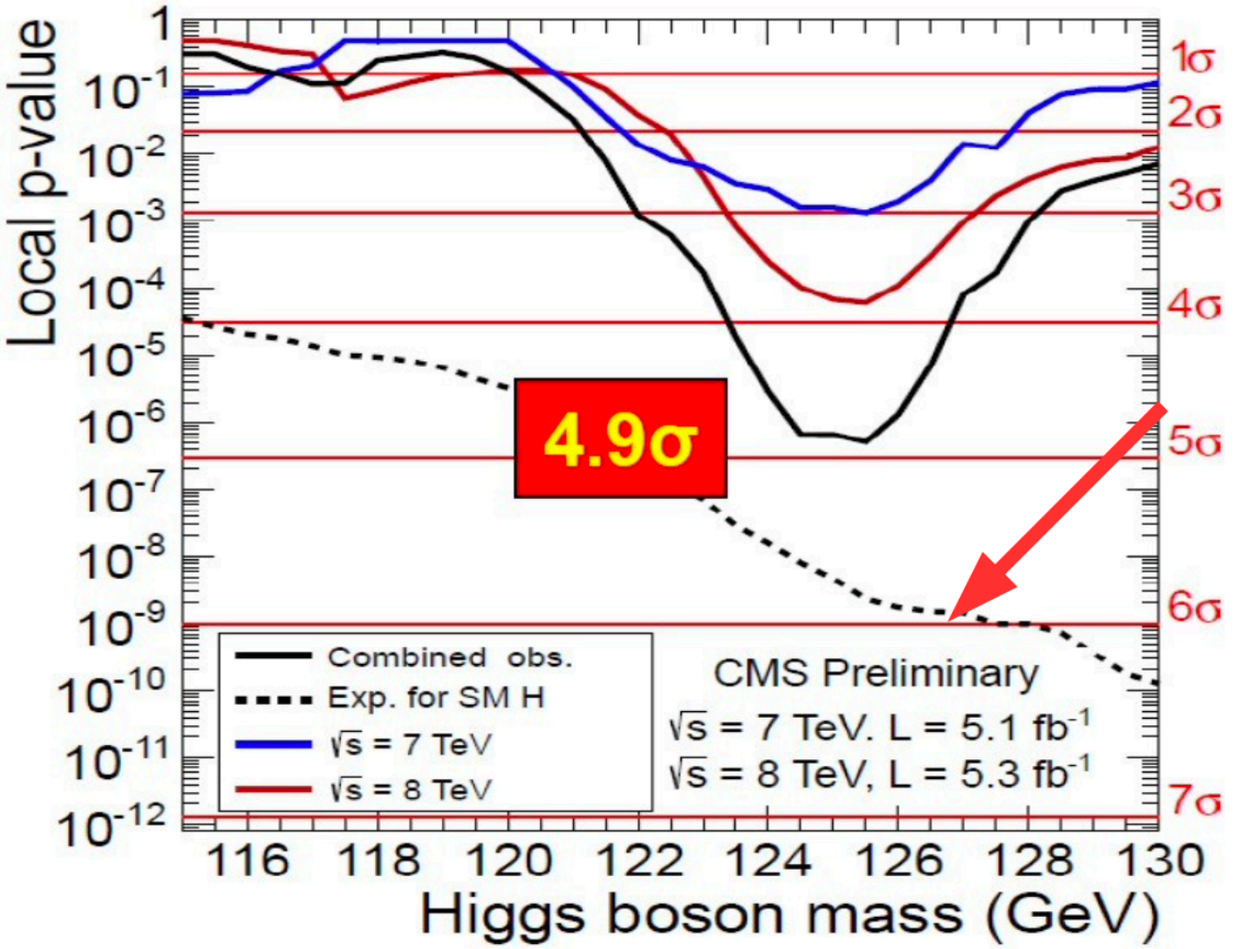
γγ, 4l updated with ~6 fb^{-1} of 8 TeV data



Largest local excess:
5σ at m_H= 126.5 GeV

CMS

All channels updated with ~5 fb^{-1} of 8 TeV data

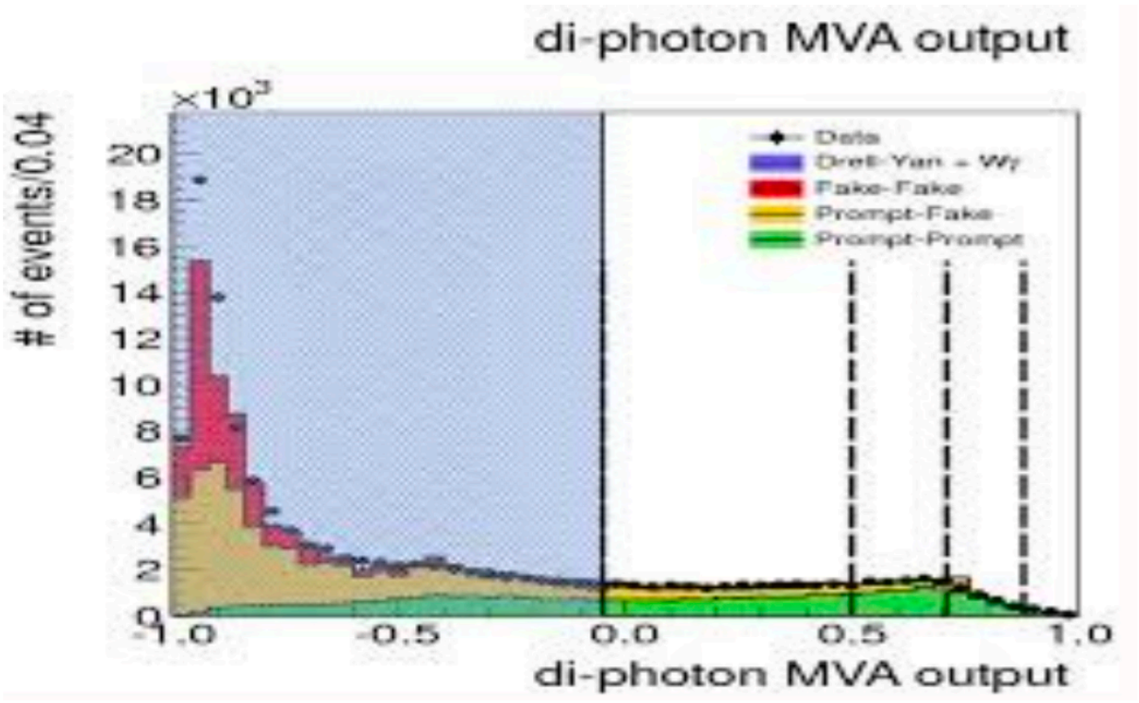


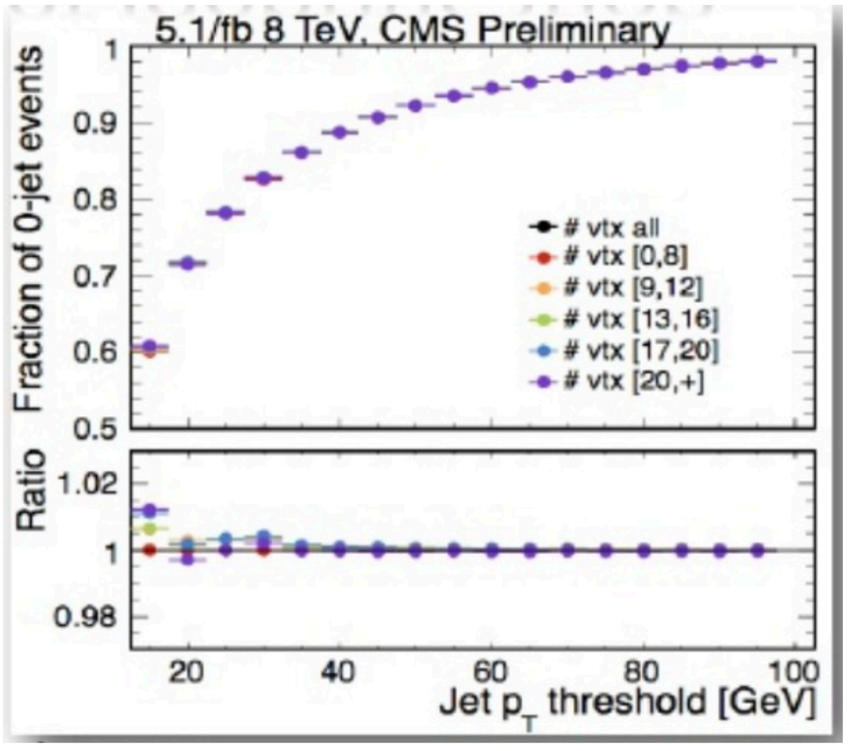
Largest local excess:
4.9σ around m_H= 125 GeV
(using H→γγ and H→4l: 5.0σ)

A big difference was present

What caused the difference

- Time of the Higgs discovery there was new tech
  - It was at its infancy
- In CMS we were developing new strategies
  - ML Photon/electron identification and reconstruction
  - Substructure + ML For Jet identification
  - Many more approaches came after





Deep Learning Progression

Deep Learning started to impact research significantly in the mid 2010s

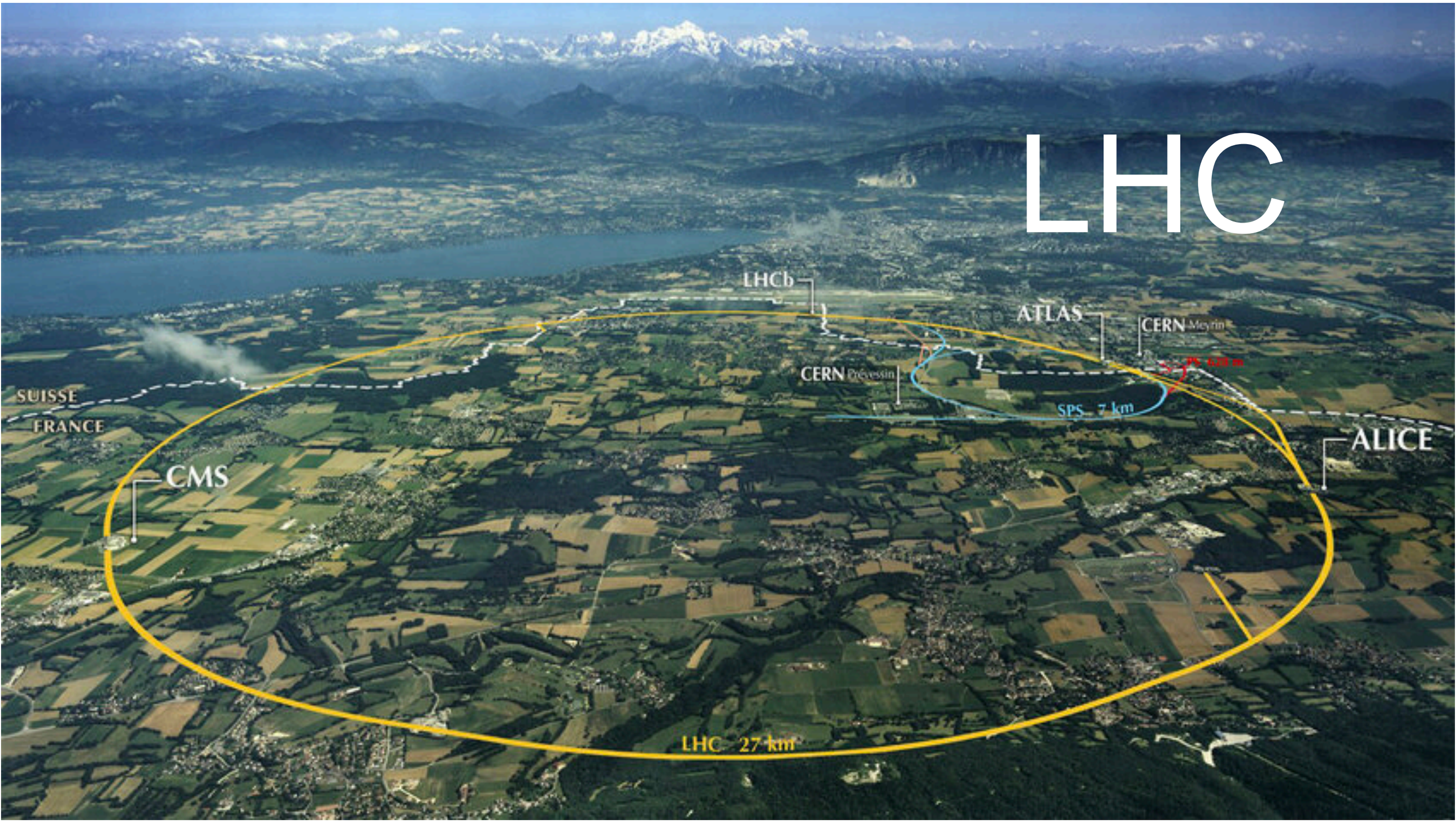
Software frameworks and GPUs rapidly accelerated its use

Progressively moving towards use of more info

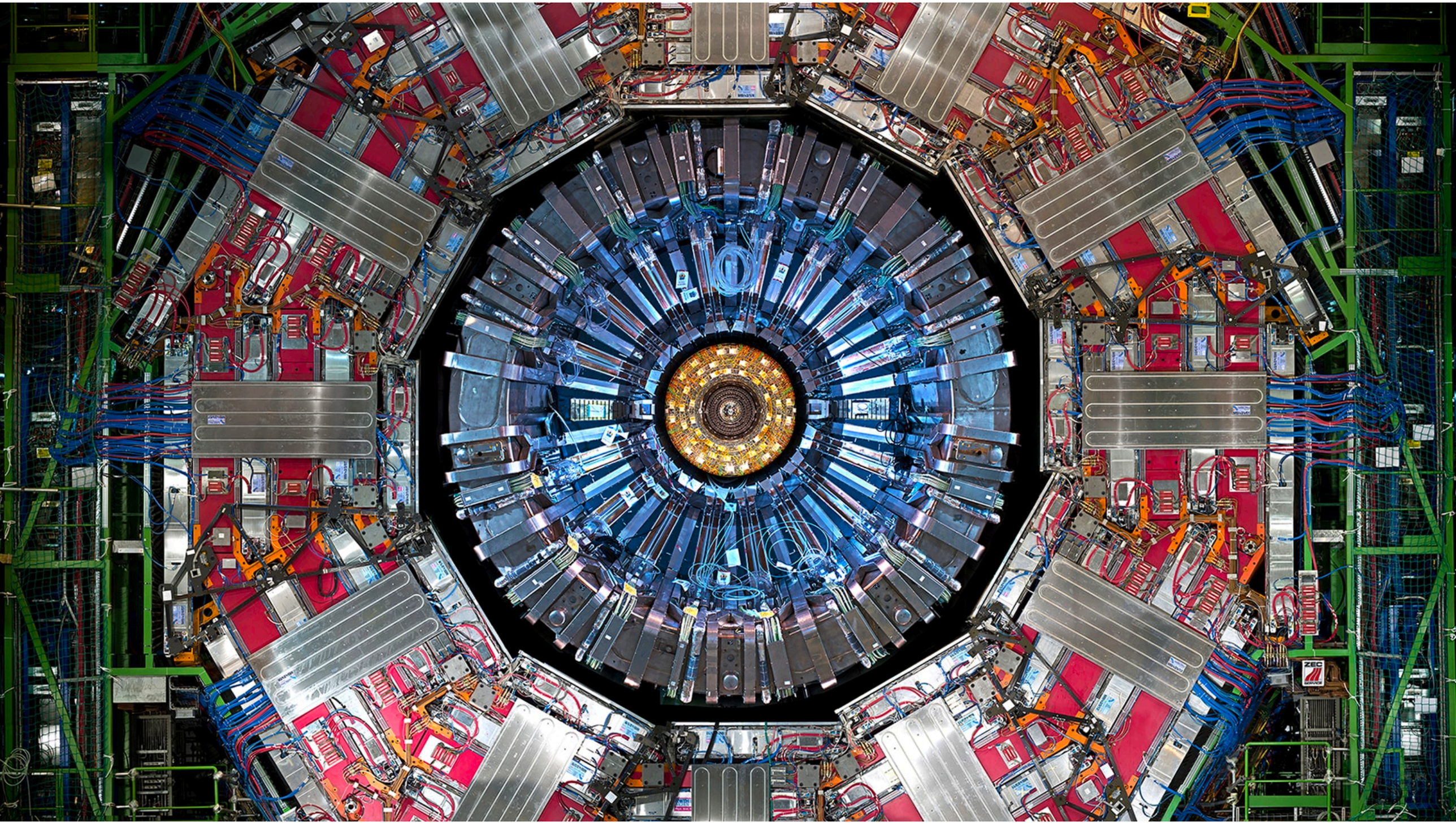
This talk

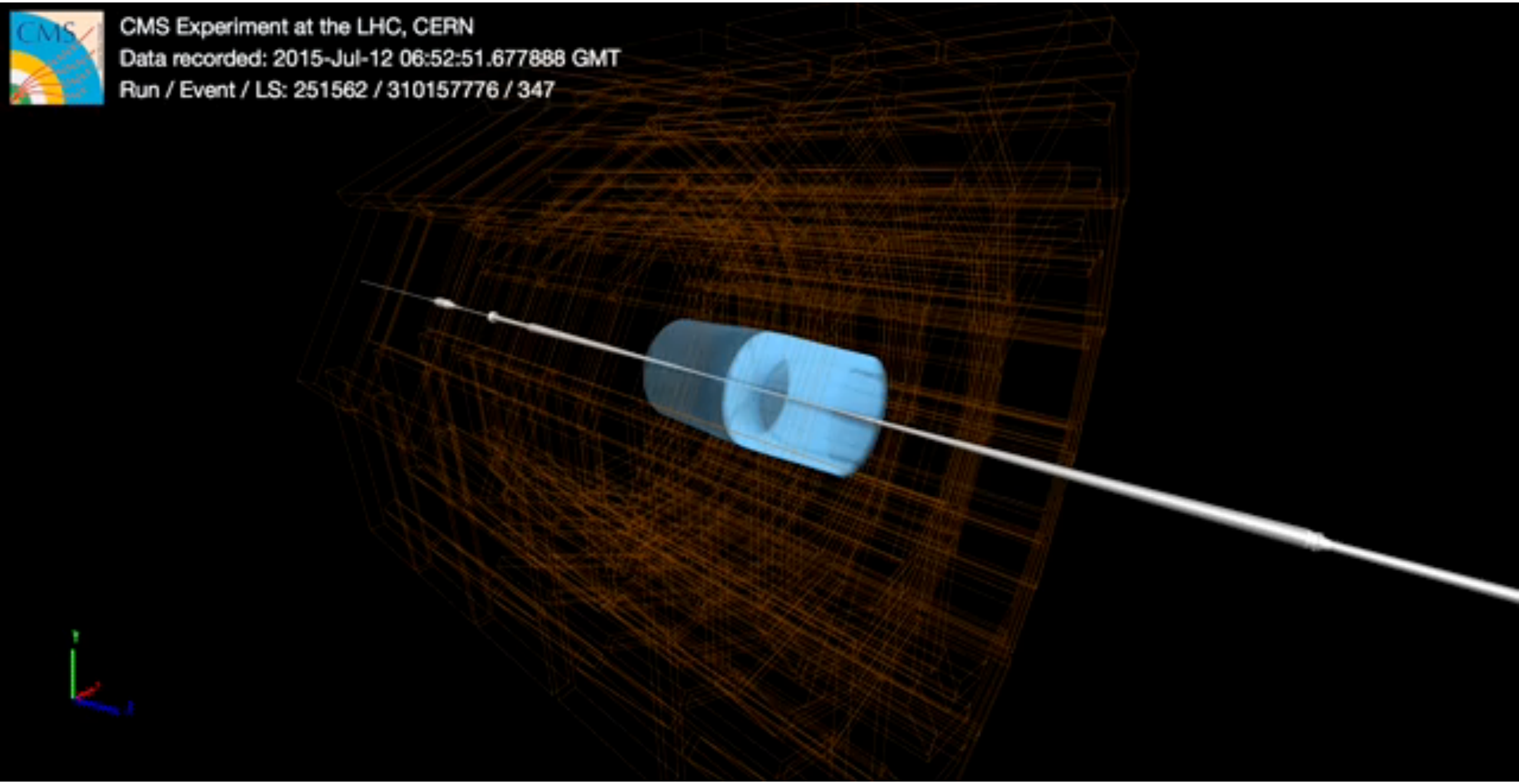
- Here, we ask the question of how AI changes HEP
  - We are already using it throughout HEP
    - I want to touch on new developments
    - Get you excited about the extended capabilities
- I am presenting a biased view touching on many projects
  - We start with a vision
  - We show a path to this vision
  - Its a long journey until we ge there
- You do not need to be an expert: AI is accessible to everyone

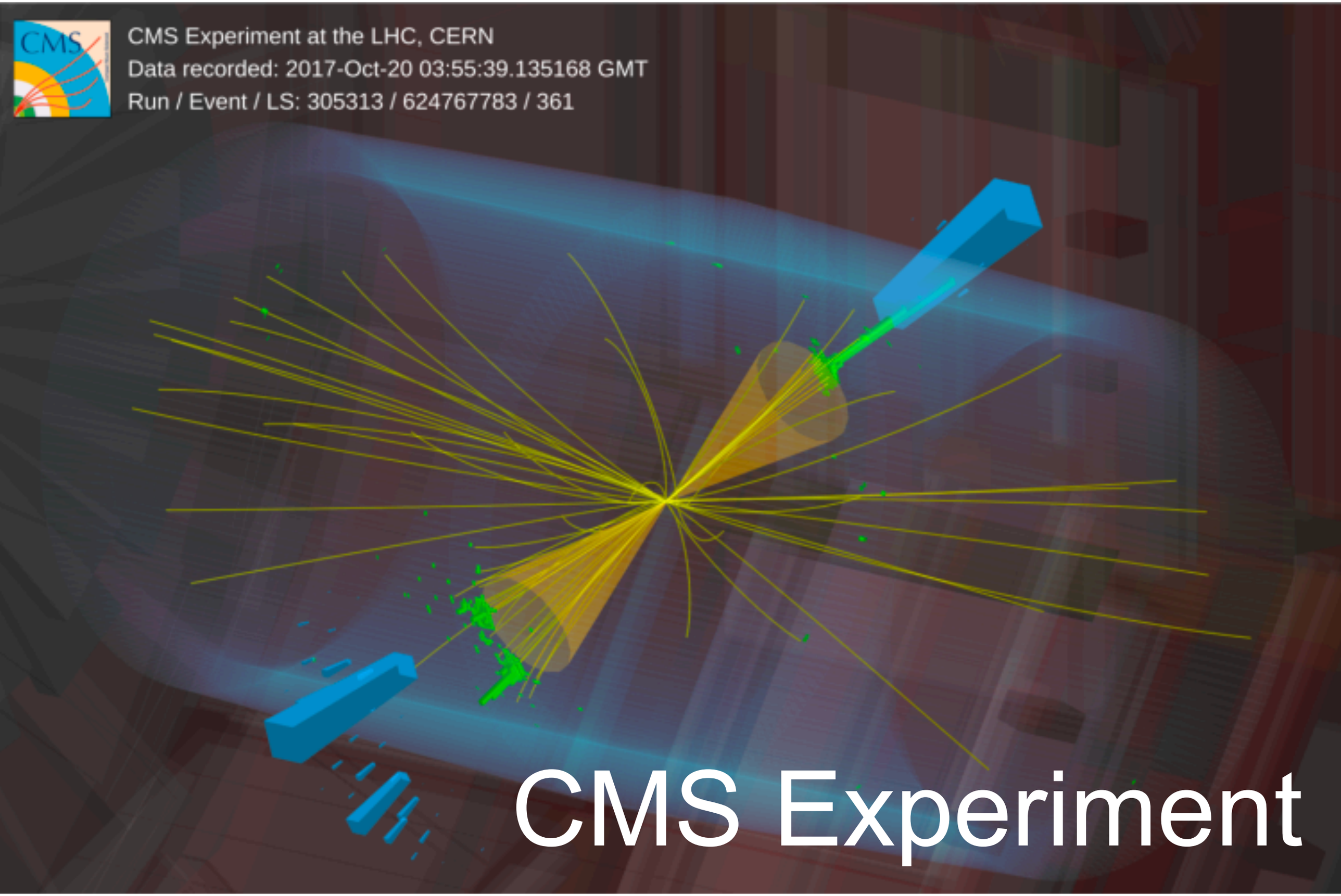
Geneva


LHC
LHCb
ATLAS
CERN Meyrin
CERN Prévessin
SPS 7 km
SUISSE
FRANCE
CMS
ALICE
LHC 27 km

With CMS Experiment

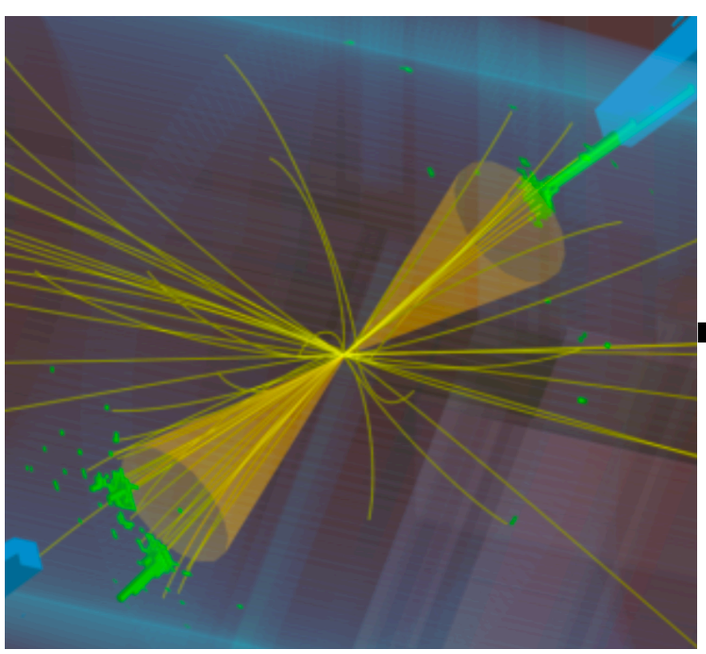



CMS Experiment at the LHC, CERN
Data recorded: 2015-Jul-12 06:52:51.677888 GMT
Run / Event / LS: 251562 / 310157776 / 347

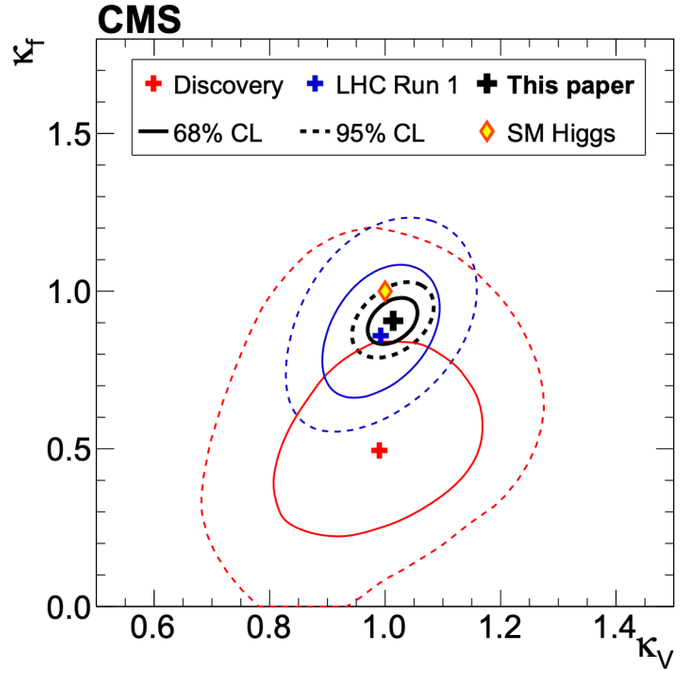

CMS Experiment at the LHC, CERN
Data recorded: 2017-Oct-20 03:55:39.135168 GMT
Run / Event / LS: 305313 / 624767783 / 361
CMS Experiment

The Dream

- Want to immediately translate event analysis to physics
  - Can we map a collision to a probabilistic hypothesis
  - Given this hypothesis, can we map this to physics



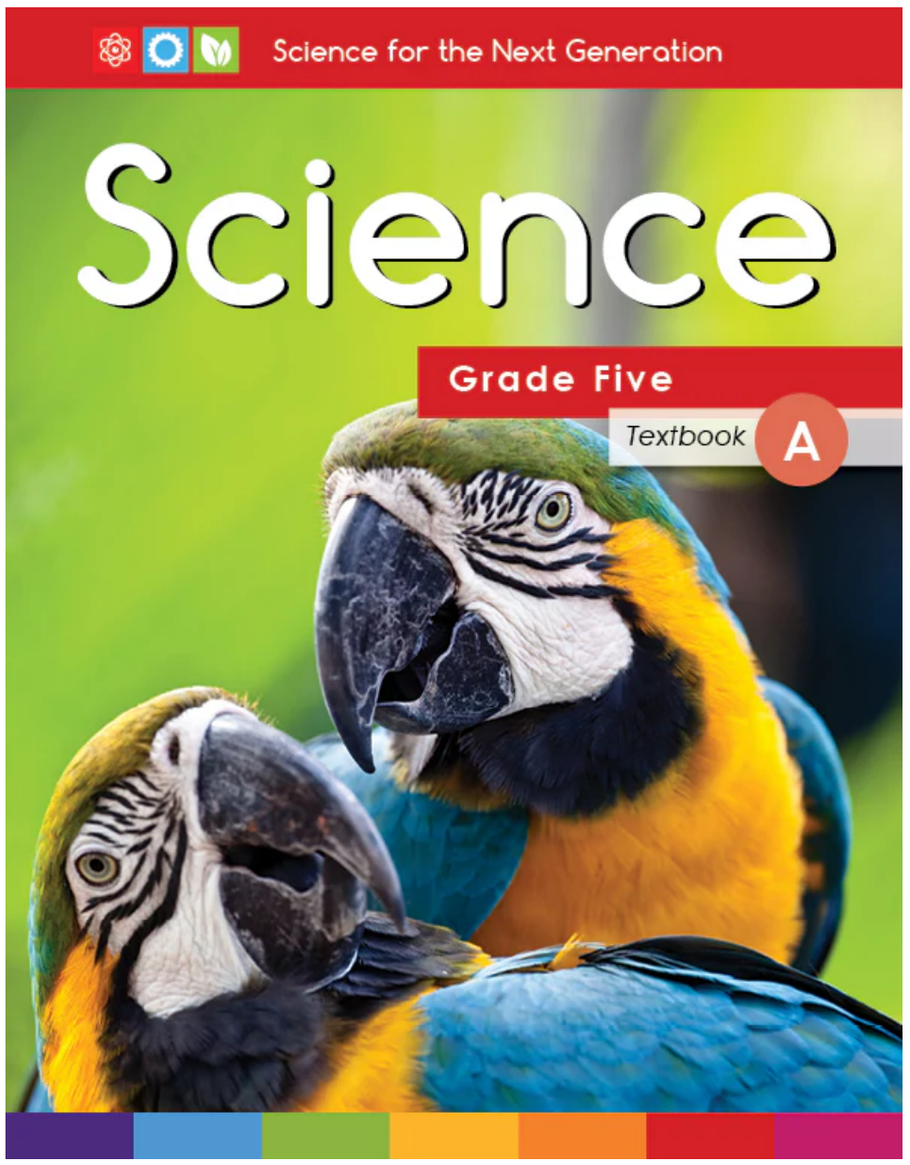
$$\begin{bmatrix} 0.6\ \text{Higgs} \to \bar{b}b \\ 0.2\ \text{dijet} \\ 0.1\ Z \to bb \\ 0.05\ ??? \\ \cdots \end{bmatrix}$$



Can we directly translate a collision to measurements of parameters
+ Searches for new physics

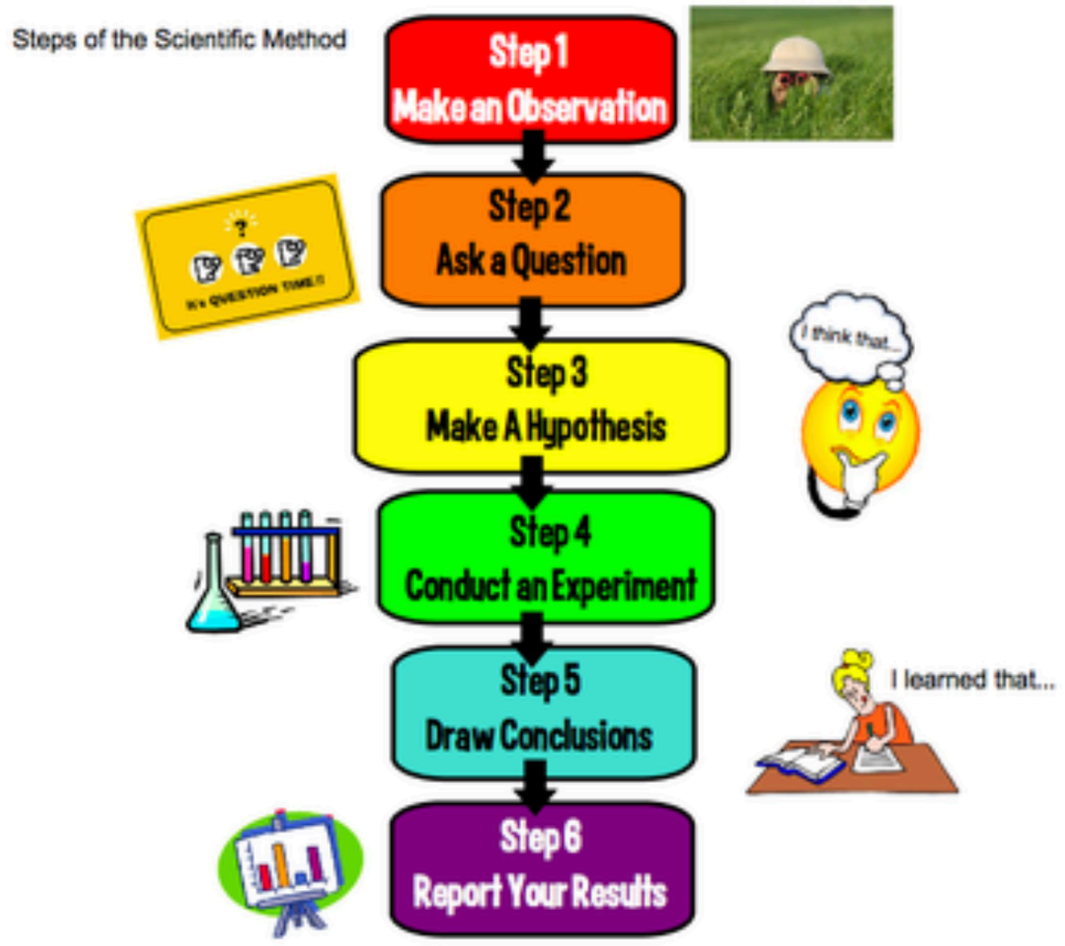
Arxiv:2510.21935
Neruips 2025

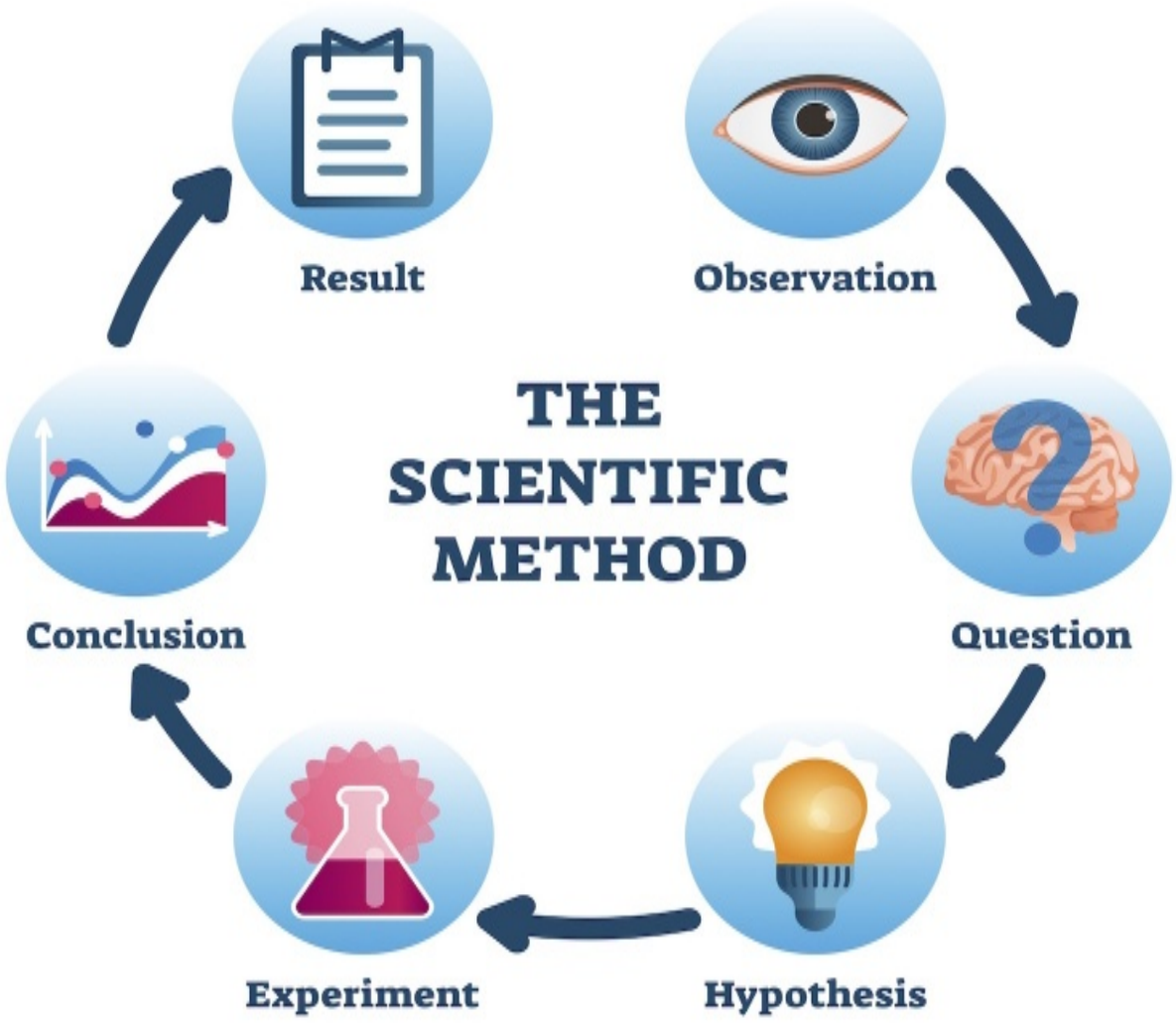
Back to the basics



Arxiv:2510.21935
Neruips 2025

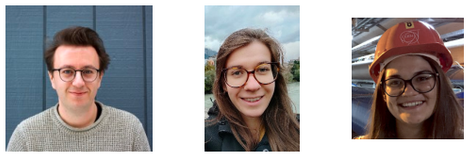
Back to the basics



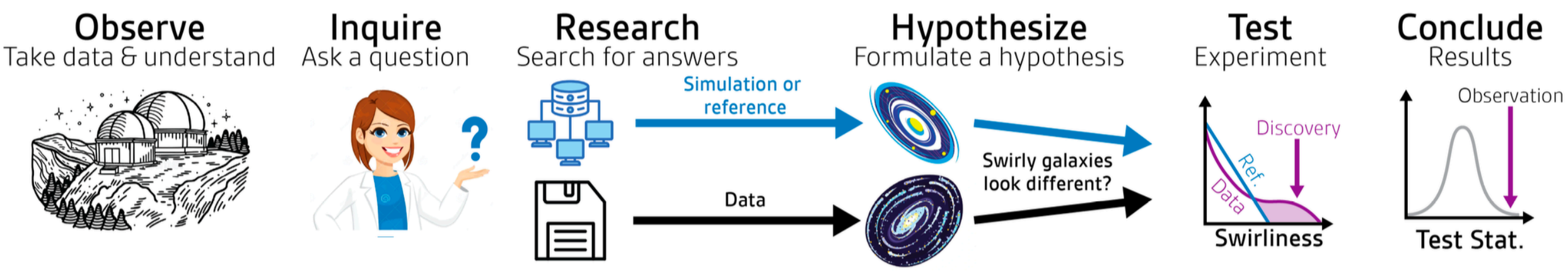


The Scientific Method

Arxiv:2510.21935
Neruips 2025



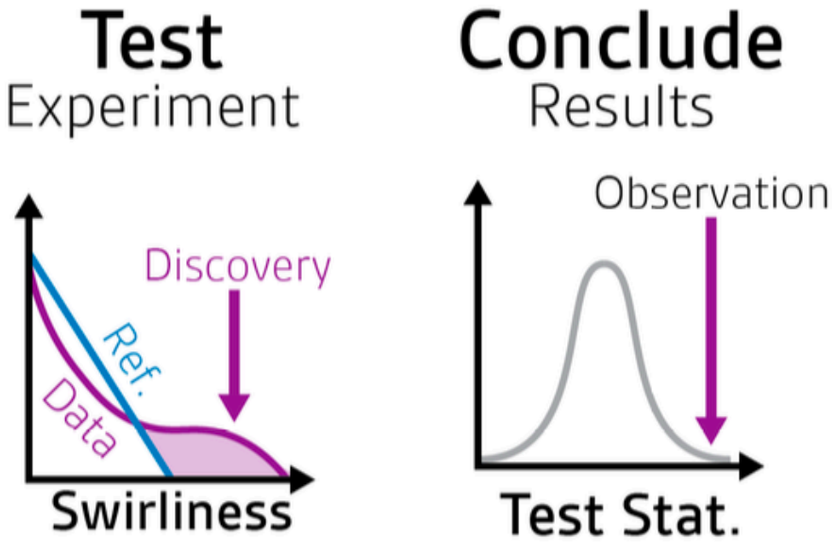
Back to the basics



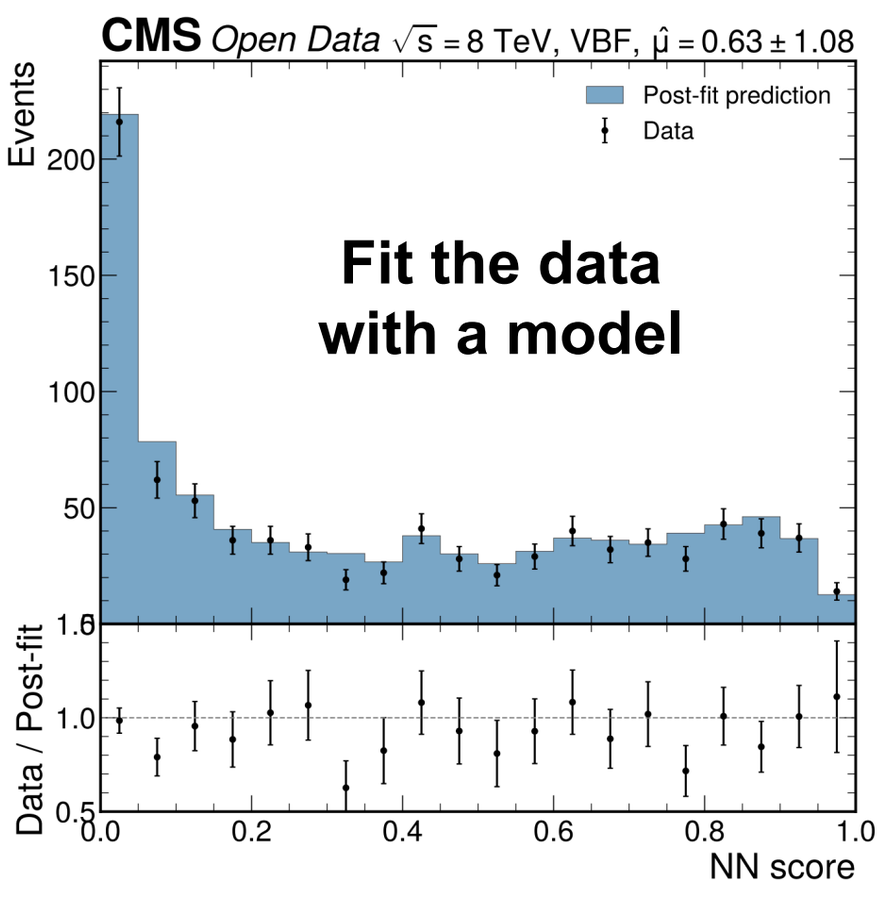
The Scientific Method

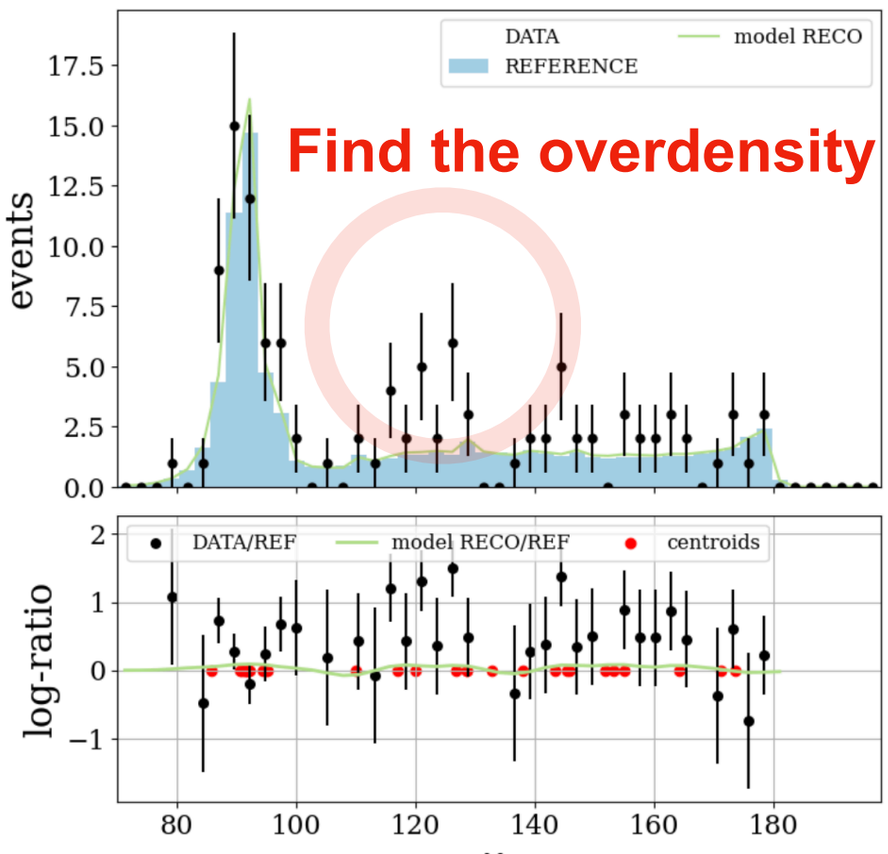
Flow of this talk

Testing Procedure



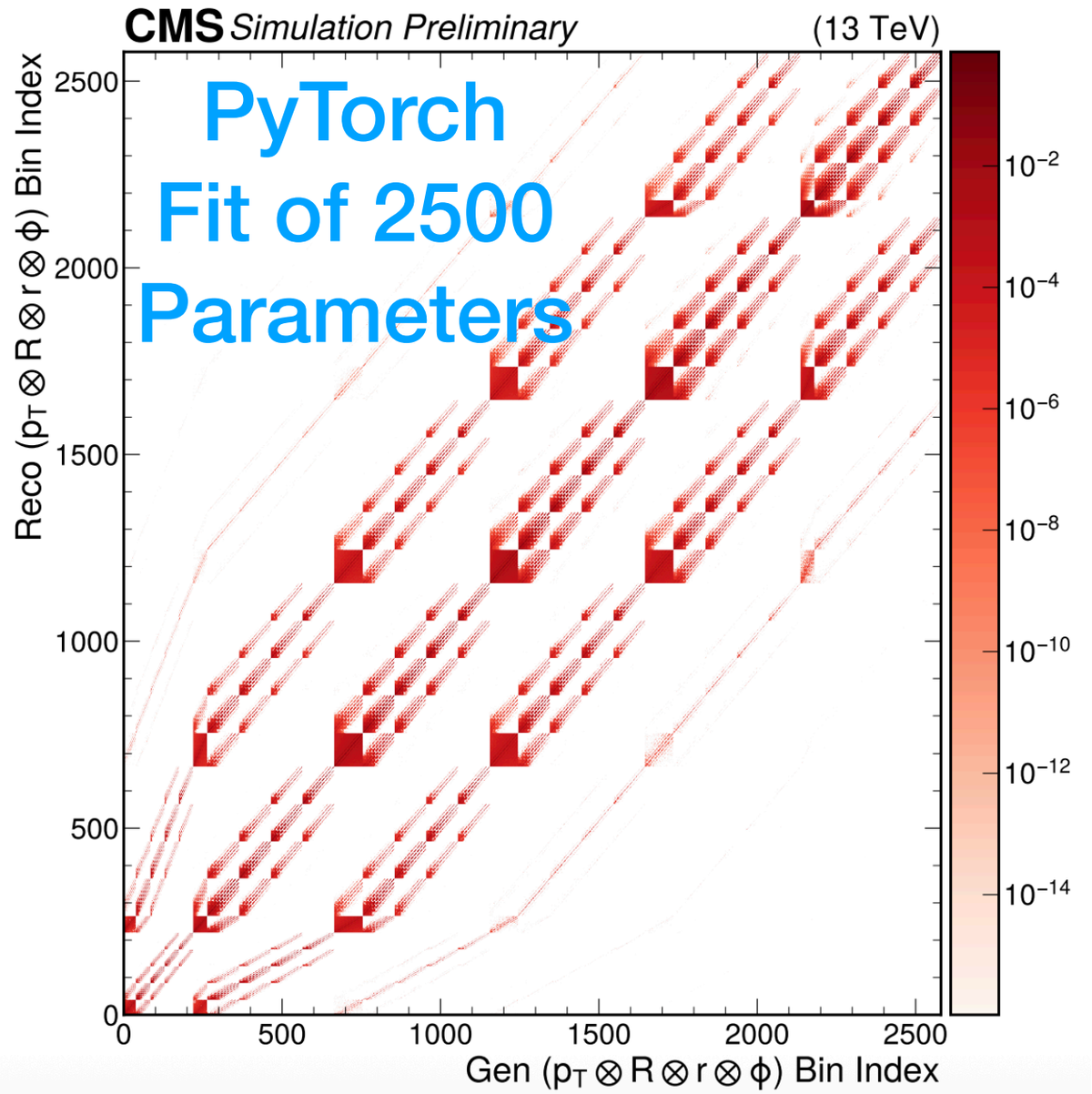
- Two types of tests
  - We know the model we want to test
    - Supervised Learning (Classic Hypoth)
  - We don't know the model (exactly)
    - Anomaly detection (Modified Hypoth)

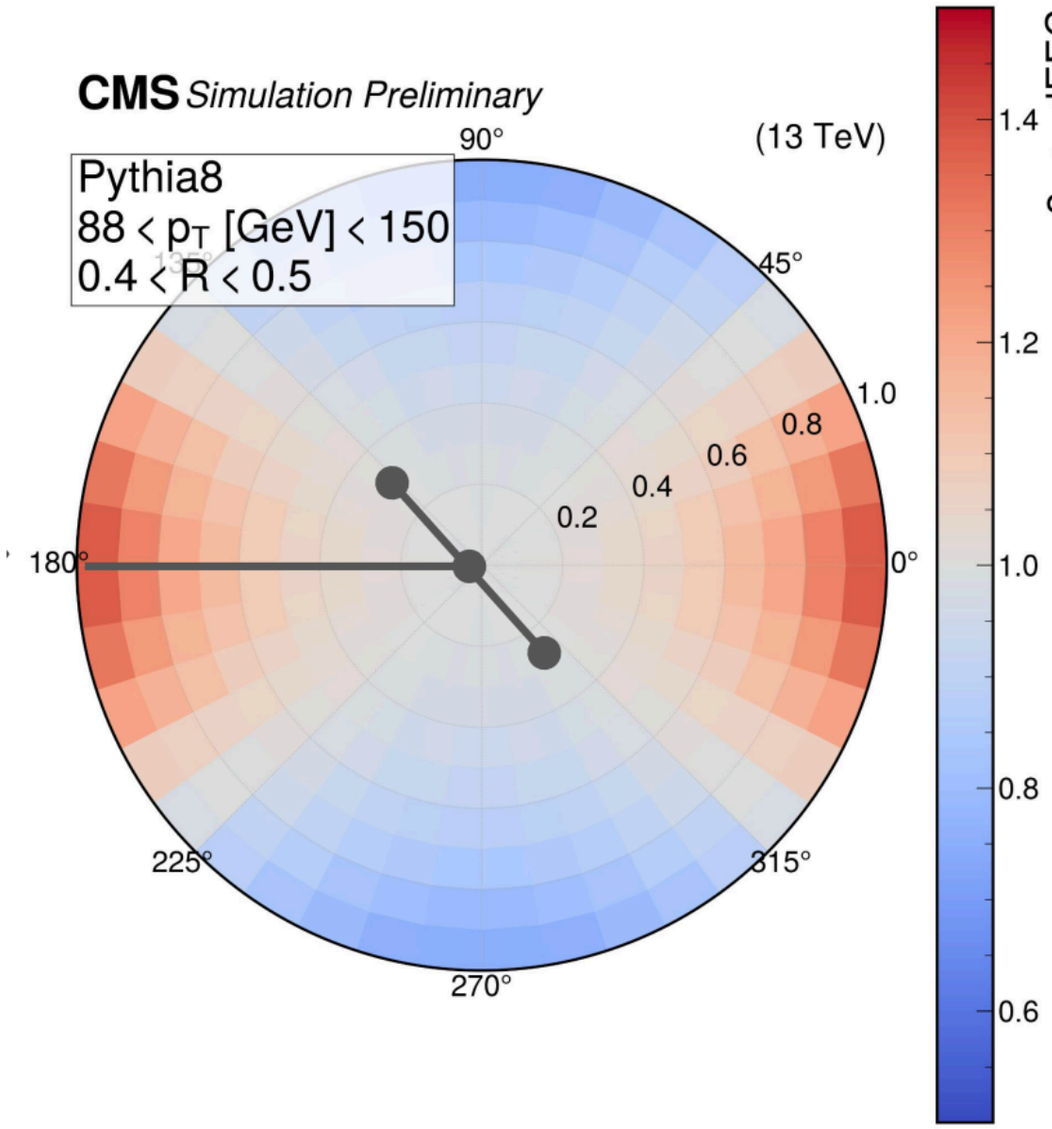




CMS-NOTE-2025-011

Fits with many parameters





A fully differentiable 2500 parameter fit runs in a few seconds

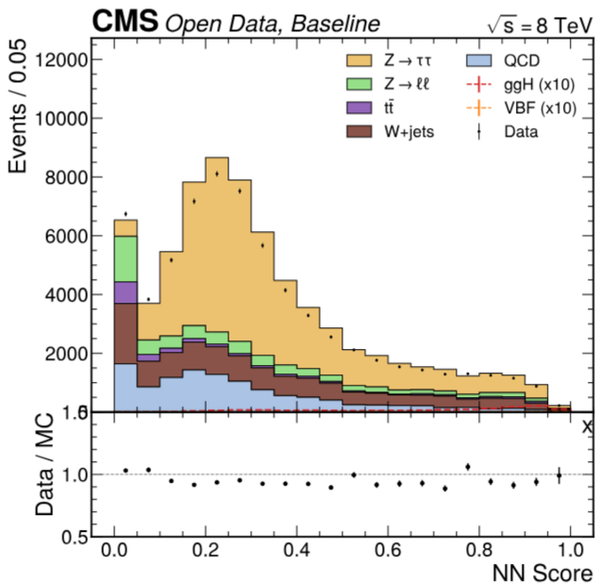
Using PyTorch/Jax under the hood allows for full fitting

Arxiv:Soon
Neruips 2025

Differentiable Fits

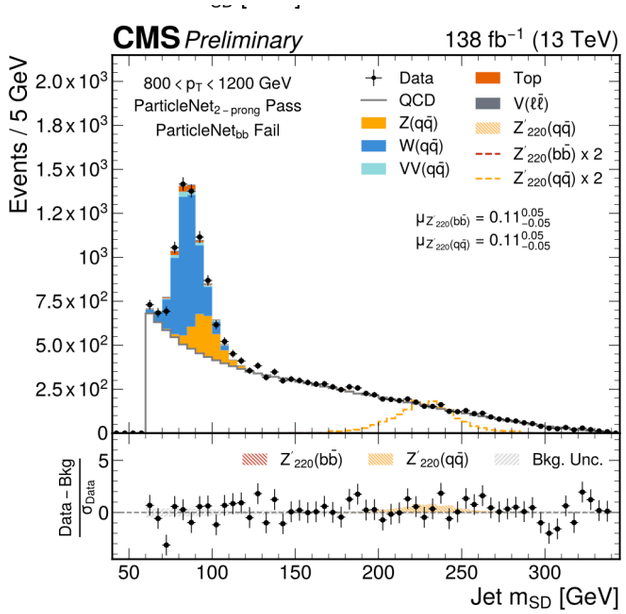
What would you like?

Option 1:
Fit a discriminator



Optimal Separation

Option 2:
Fit a Mass

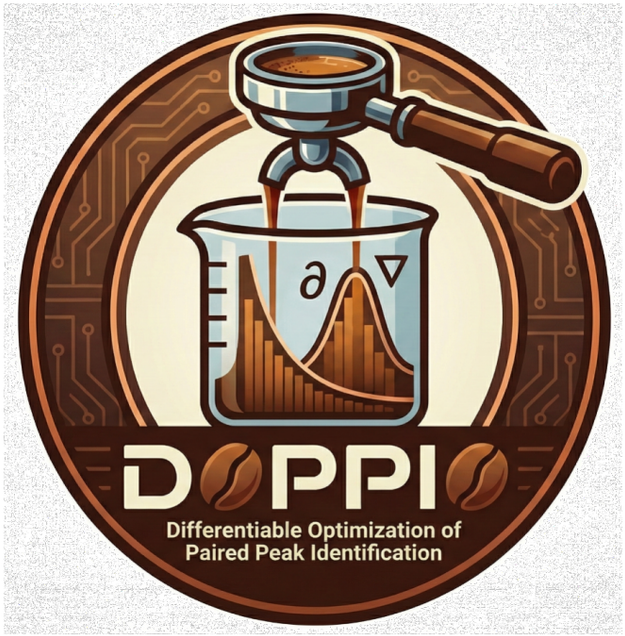


Control of backgrounds

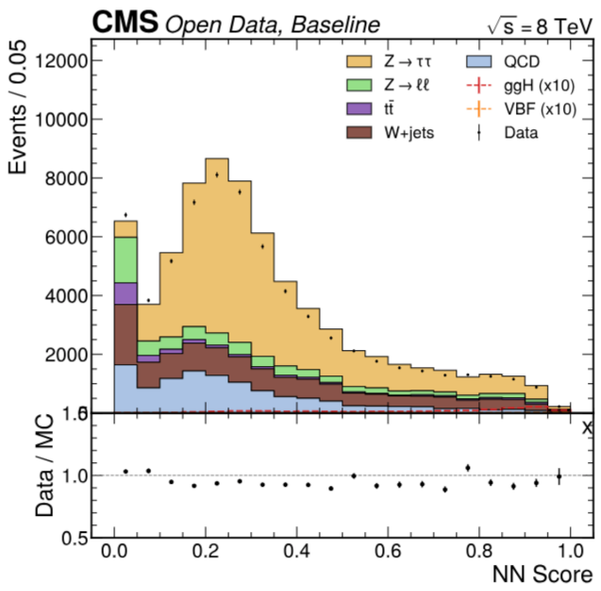
Differentiable Fits

What would you like?

Yes Please



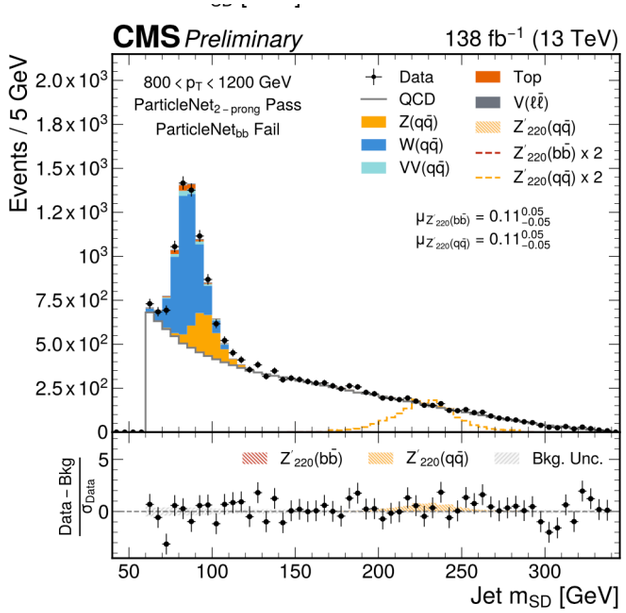
Option 1: Fit a discriminator



Optimal Separation

X

Option 2: Fit a Mass



Control of backgrounds

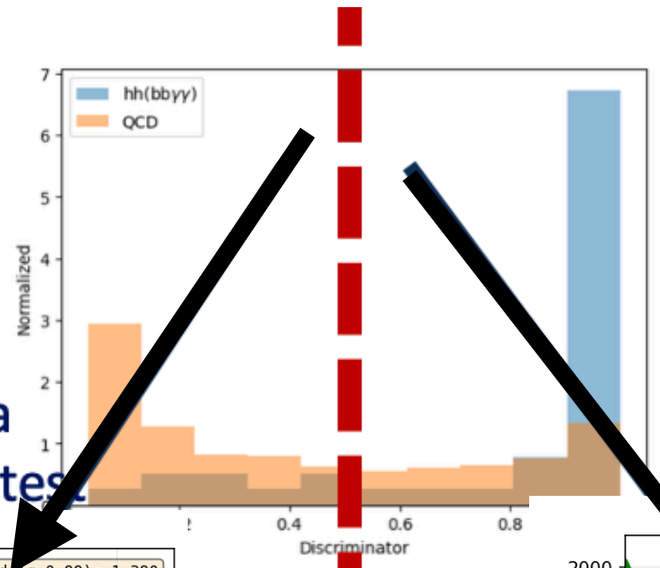
Doppio
Adjust training
Based on fit

In Data

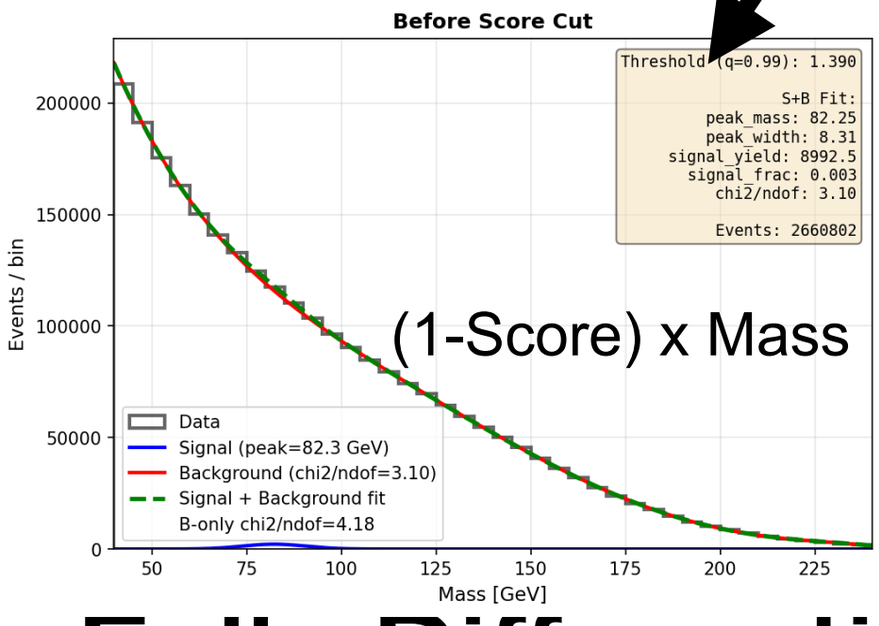
Doppio

1 .(Pre-)Train a Neural network:
2. Use the NN score to categorize

Split data
By NN score

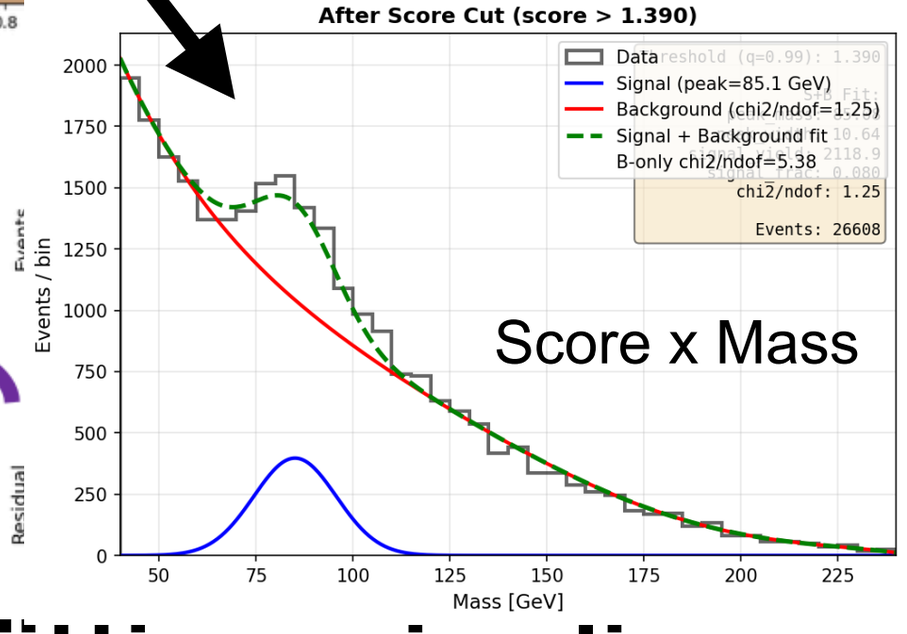


Fit functions to the data
by performing hypothesis test



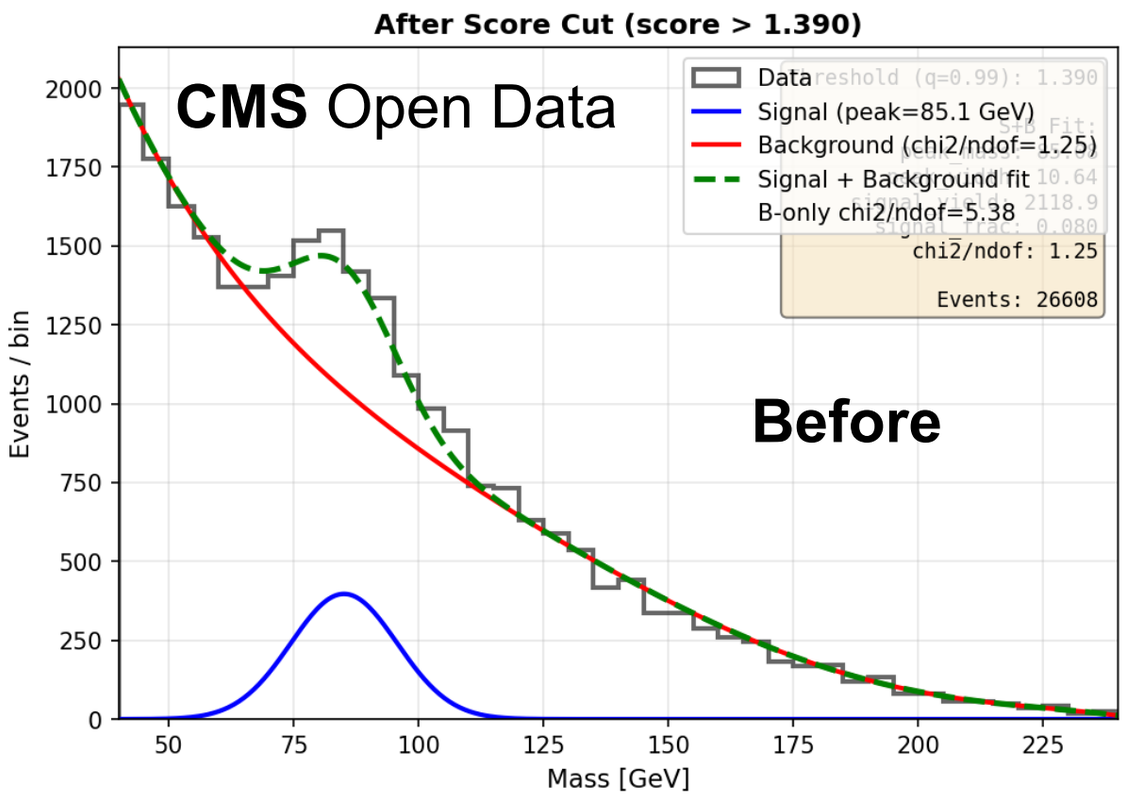
Backpropagate the best fit hypothesis test to the fitted shape

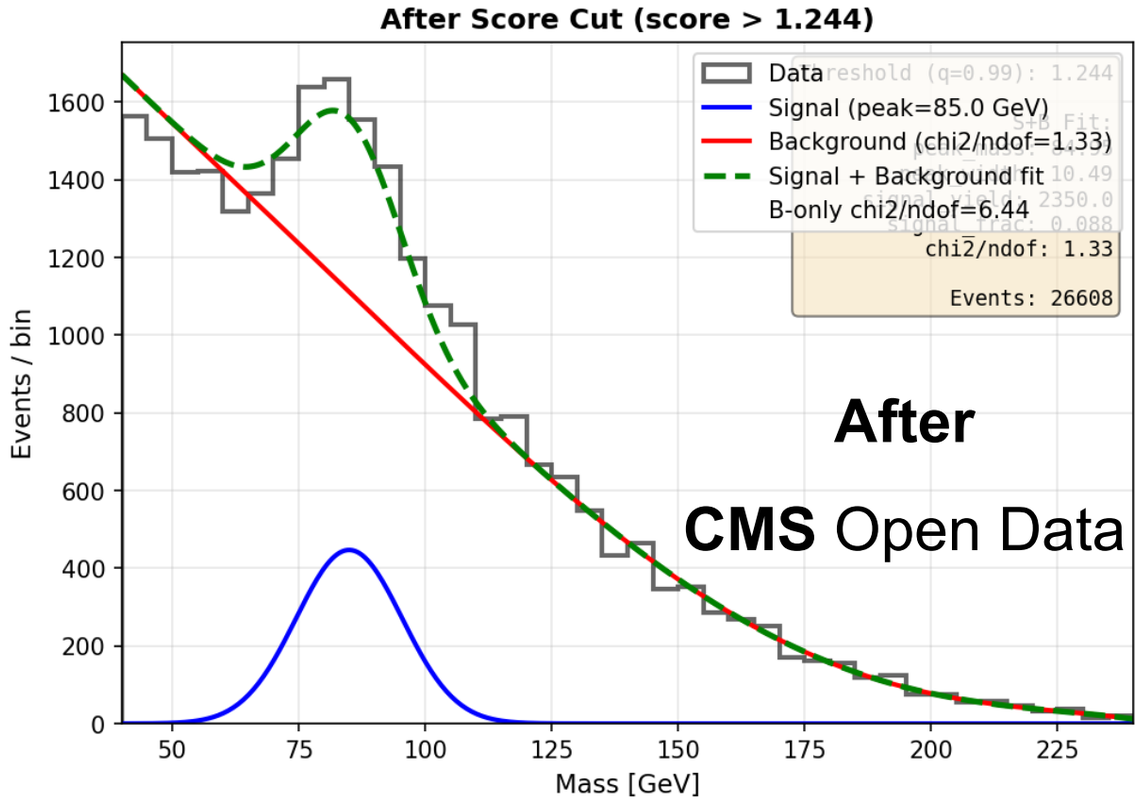
Fit functions to the data
by performing hypothesis test



Fully Differentiable Fitting pipeline

In CMS Open Data

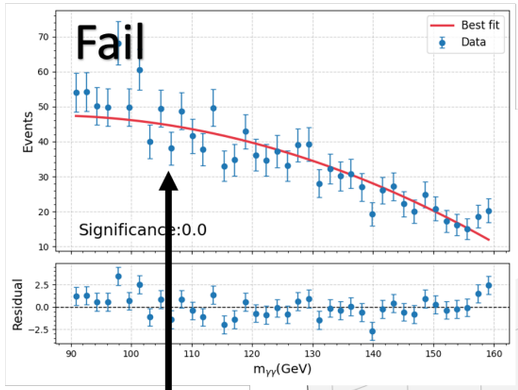


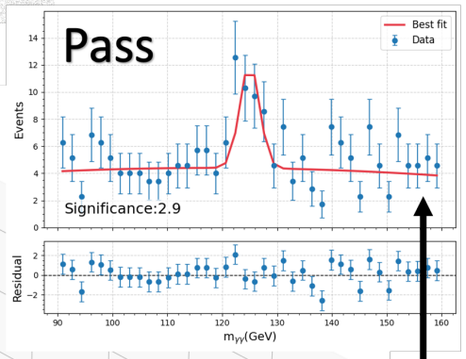


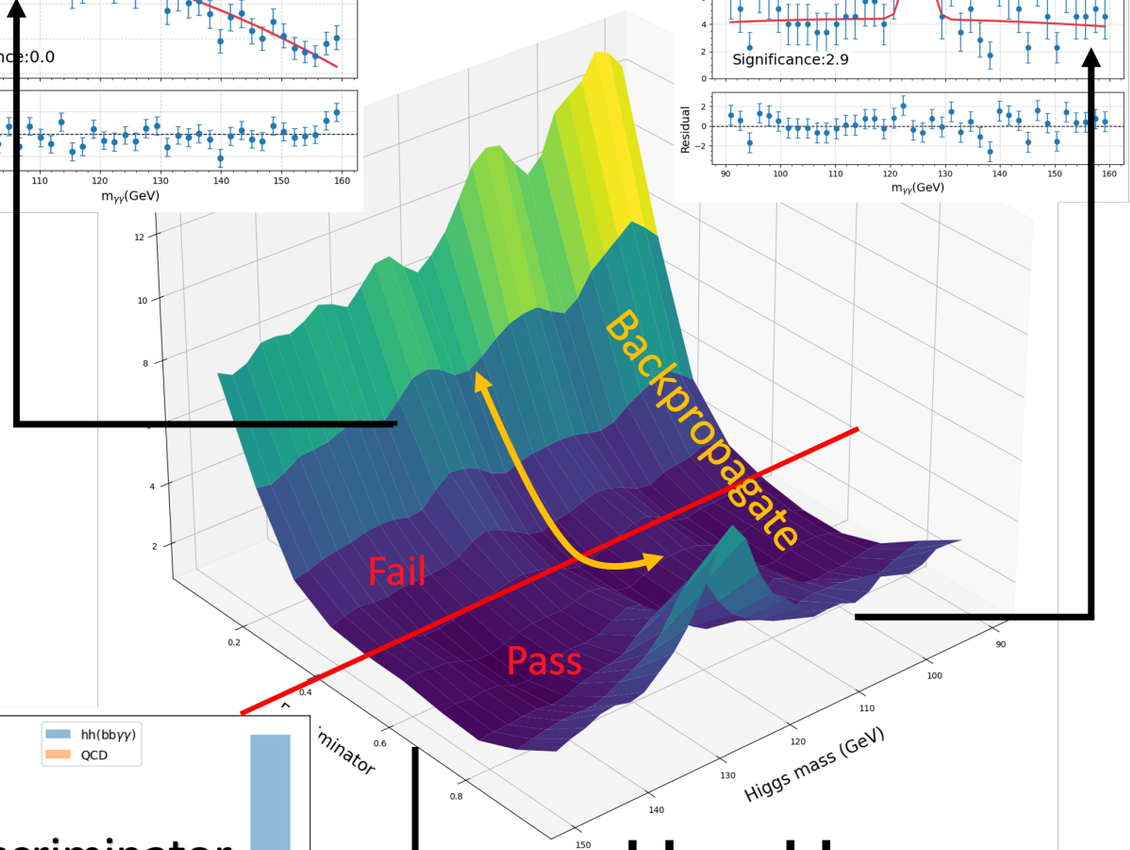
- This works in data: Here optimizing the W boson peak
  - Now we can validate trainings/optimizations on data
  - We can loosen prior to try to just maximize a bump

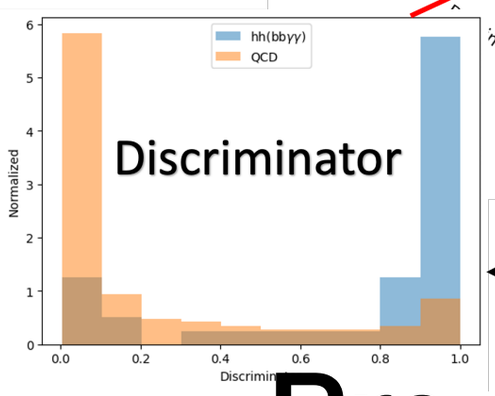
Arxiv:Soon
Neruips 2025

Optimally measure the Higgs Self Coupling

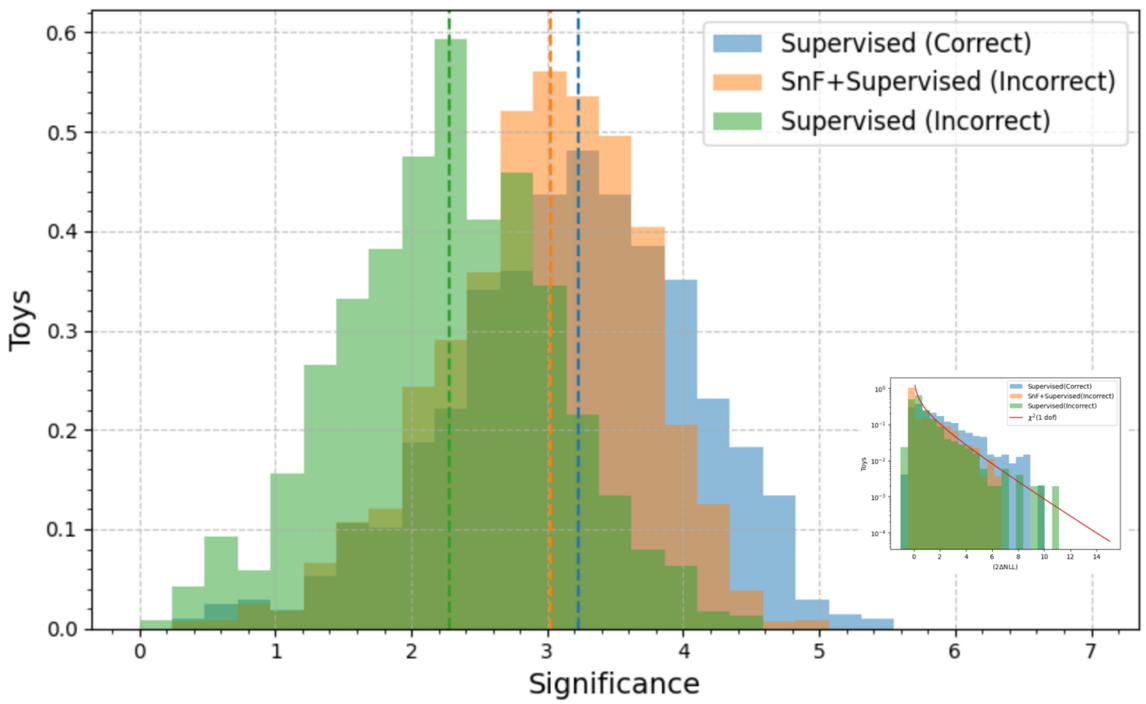








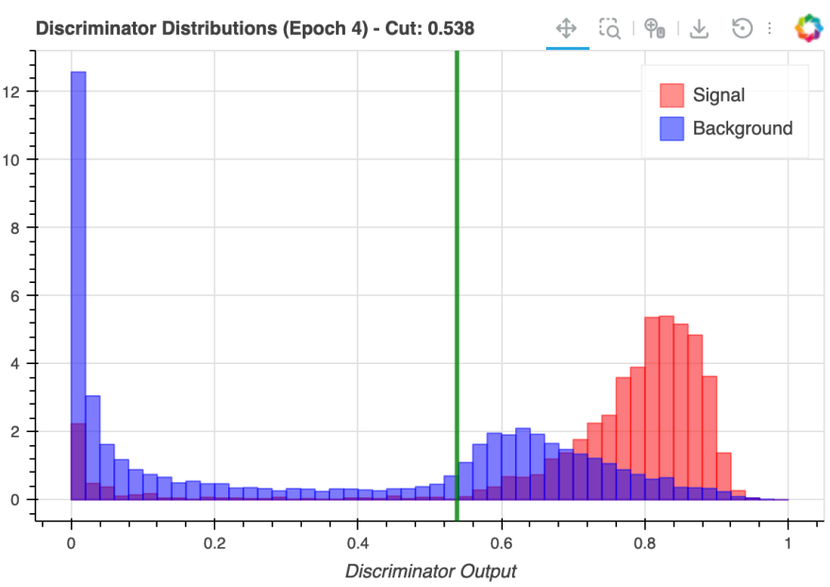
hh →bbγγ
Analysis (3ab^{-1})

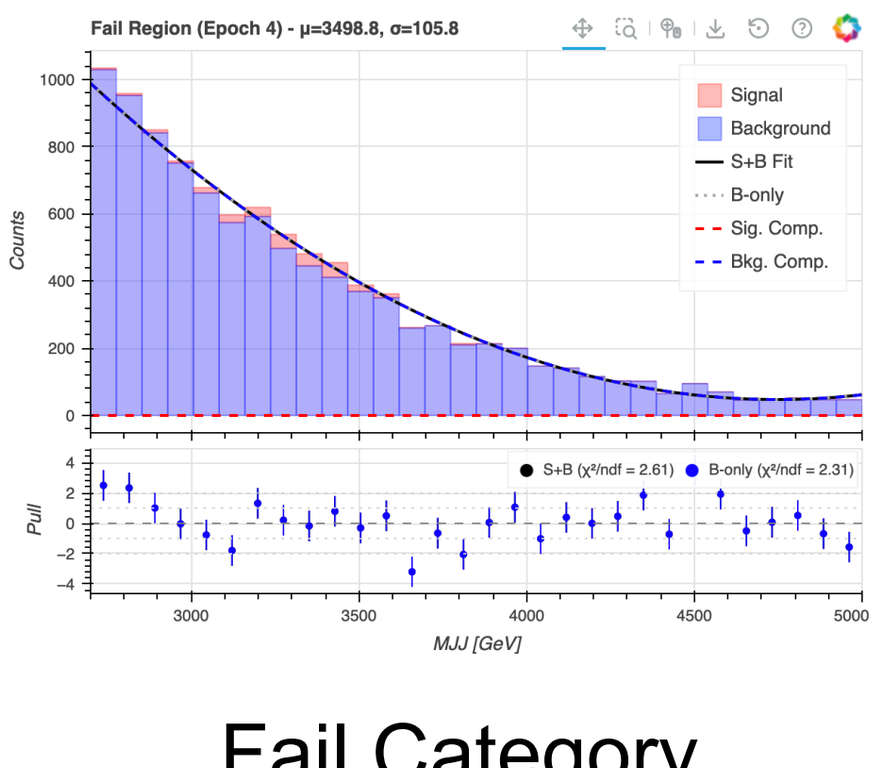


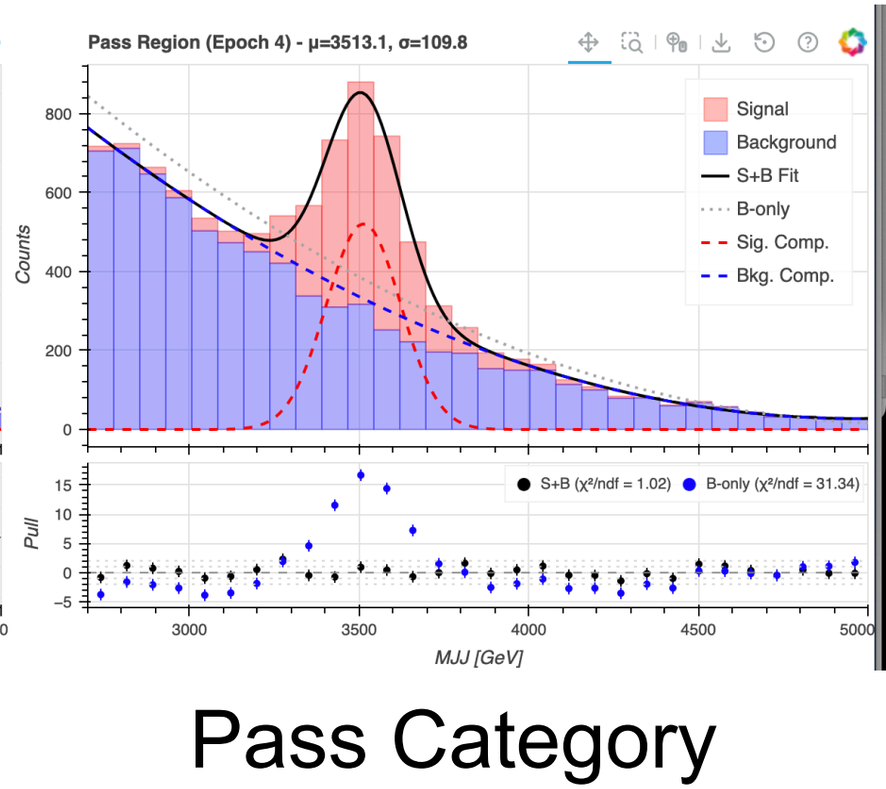
Sig w/o training on data: 2.2
Sig w/doppio on data : 3.0
Idal significance : 3.2

Pre-train on MC
Fine Tune on Data
Recovers Data/MC disagreement

Anomaly Detection





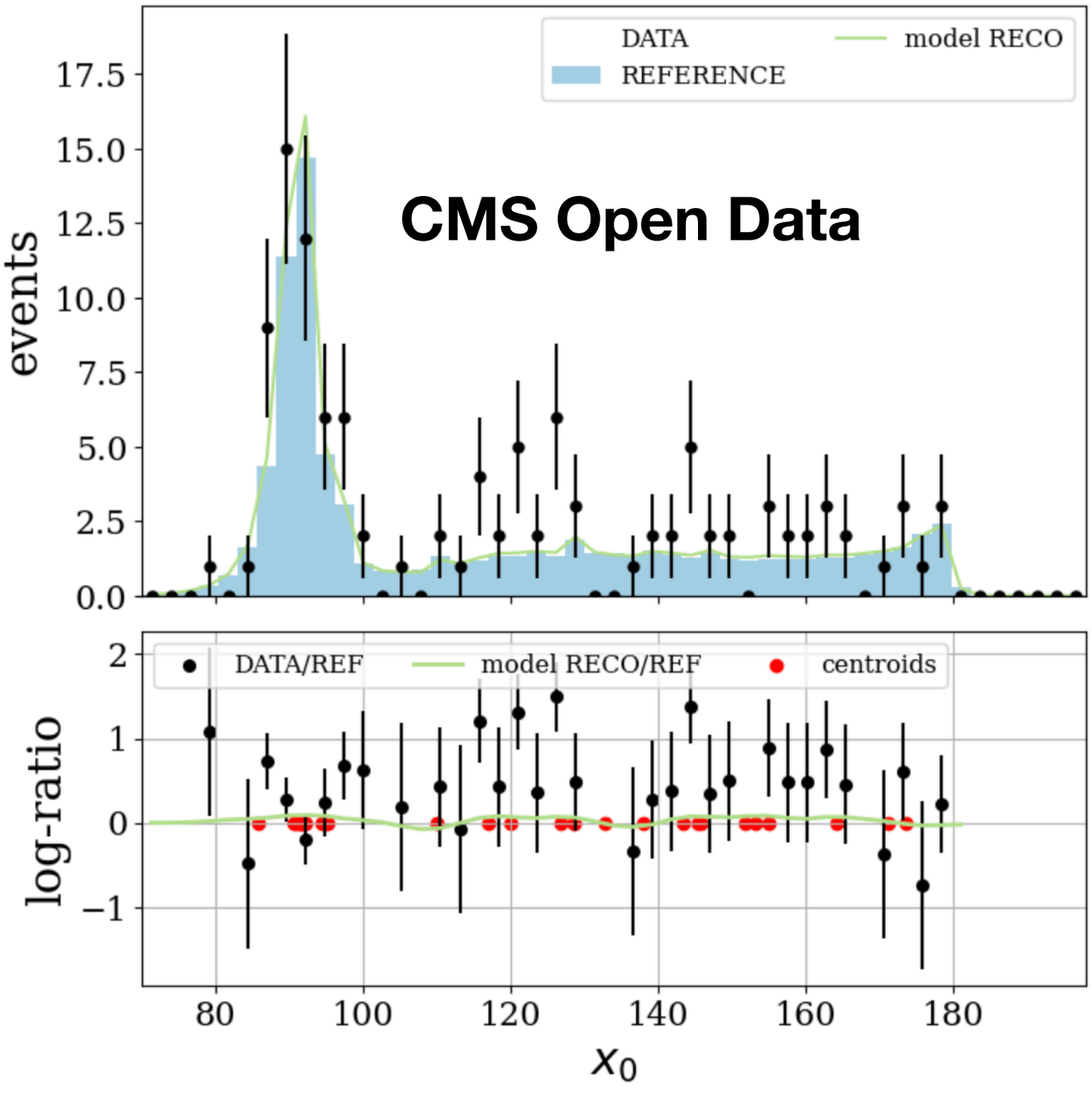


Discriminator Fail Category Pass Category

- This works in data: Here optimizing any peak
  - Just require allow the mass to move up anywhere
  - Backpropagates the signal to find a deviation

Neruips 2025

Now Find Anomalies



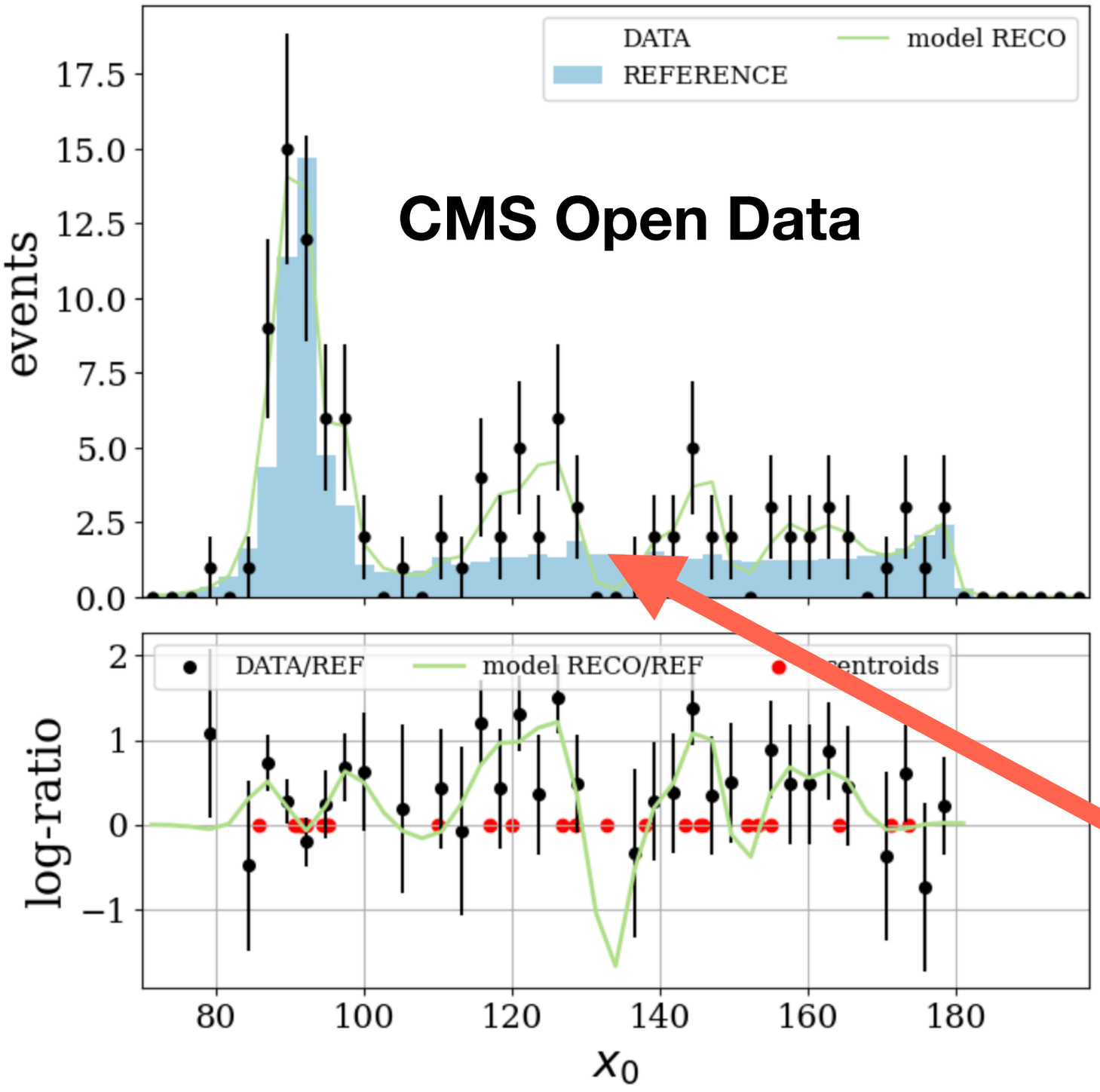
Use AI:

Find deviations

From a reference

in **N-dims**

Now Find Anomalies



Use AI:

Find deviations

From a reference

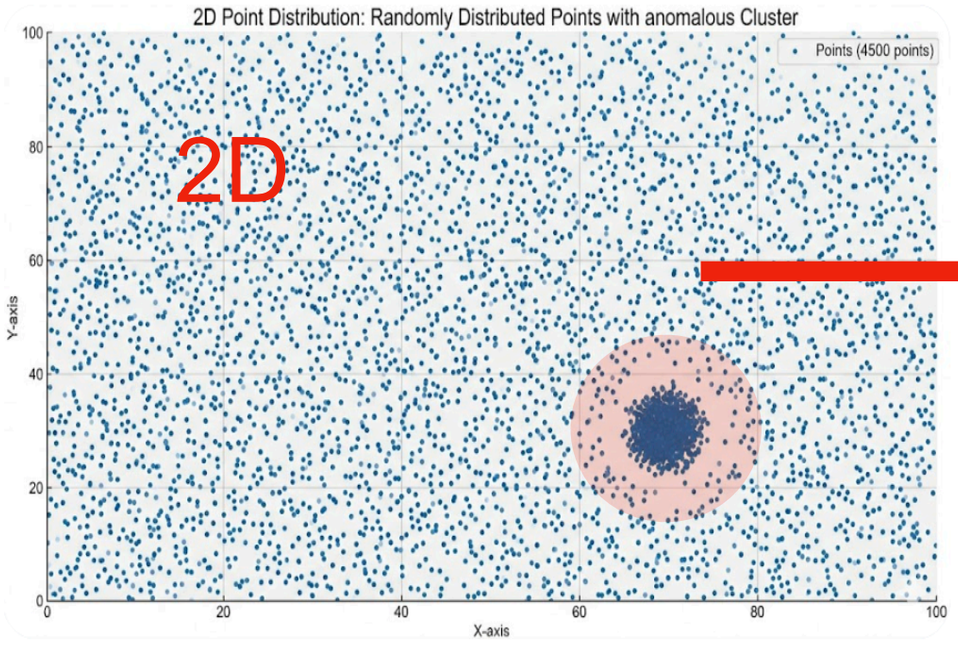
in N-dims

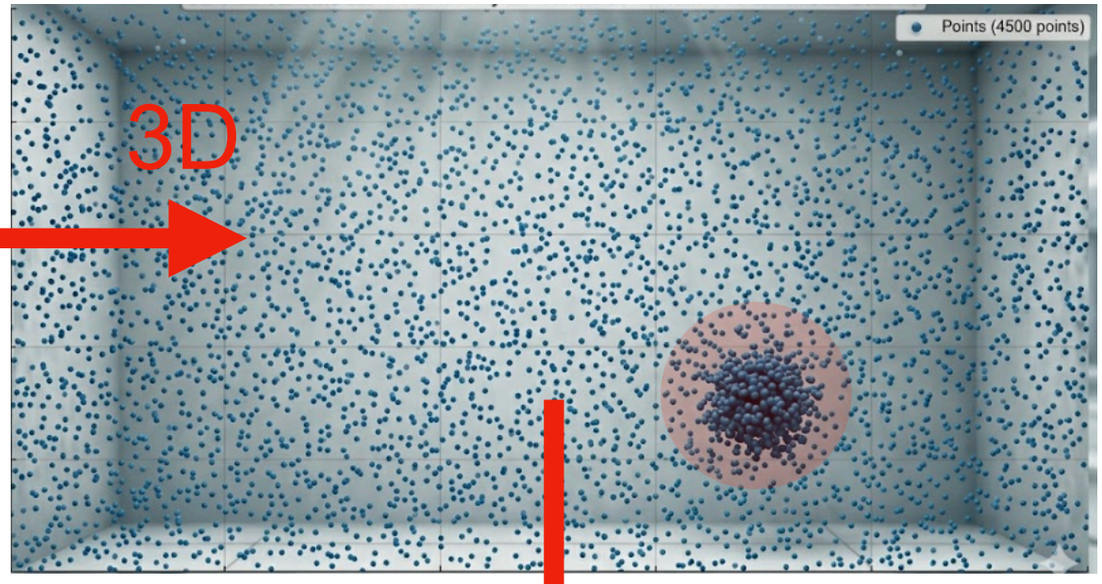
Found it!

Automatically

Fitting Gaussians

A natural inductive Bias element is a localized overdensity



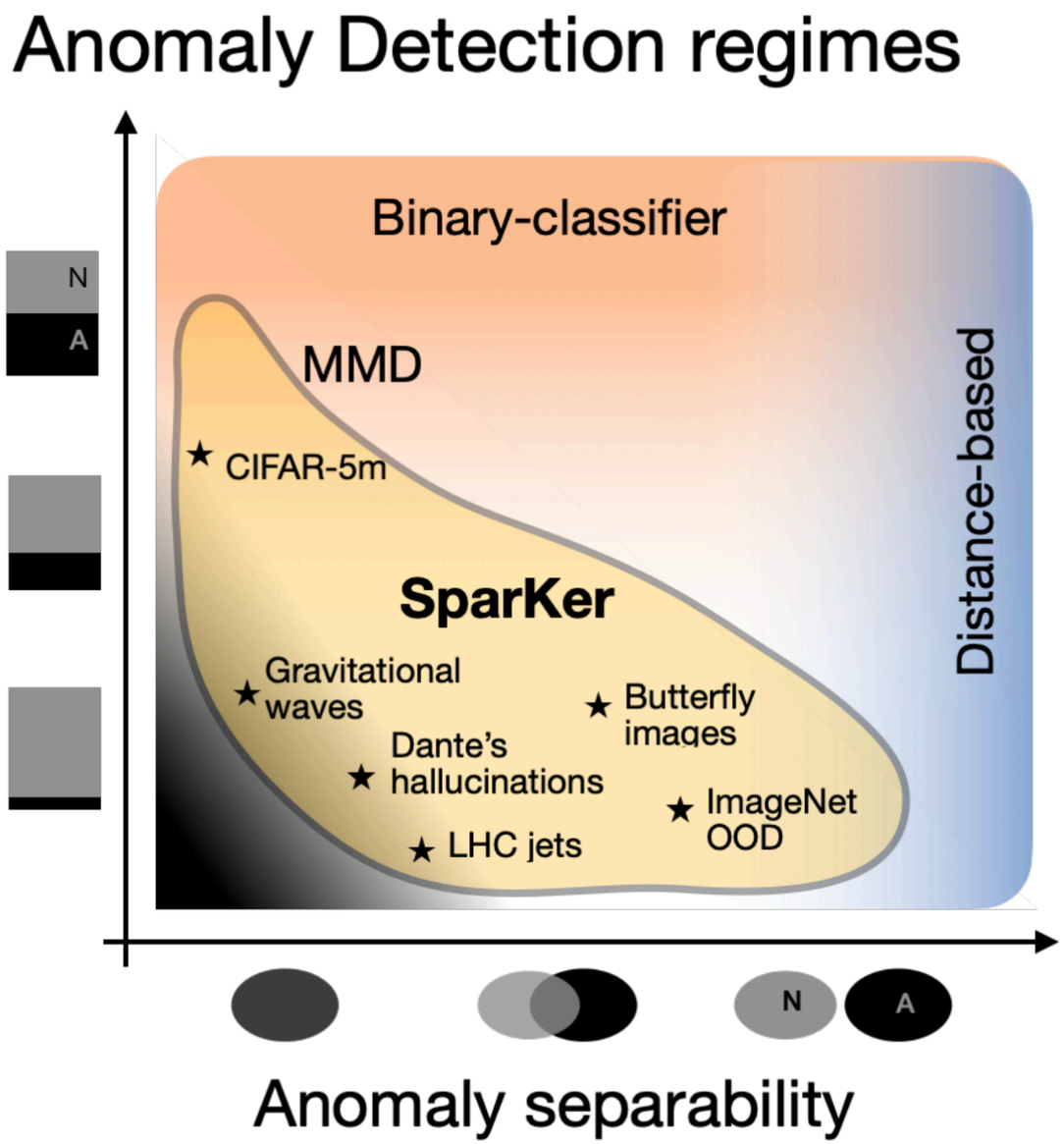


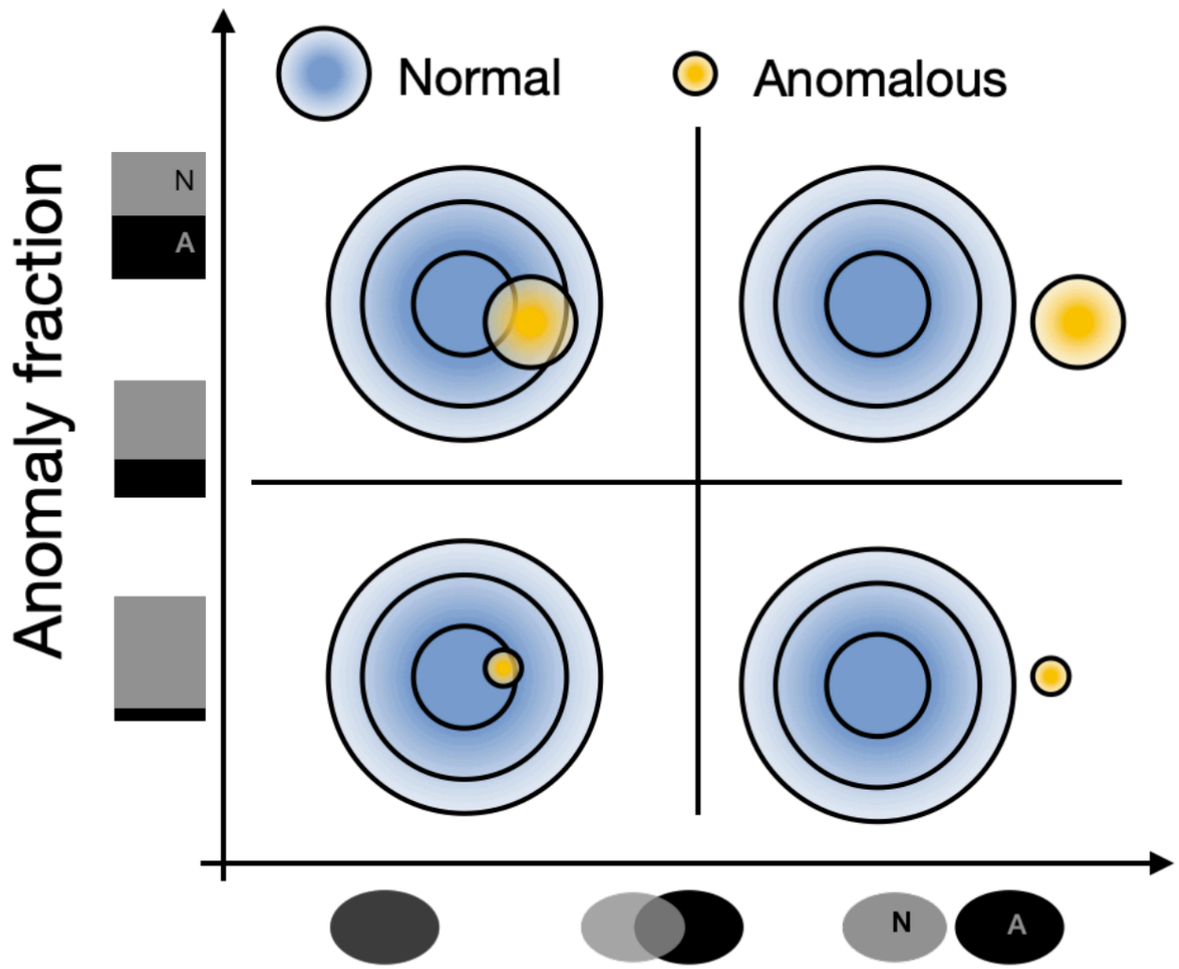
Sparse, self-organizing ensembles of local kernels (Sparker)

1000D

Sparker: fits N high dimensional Gaussians
To overdensities in the data

Fitting Gaussians

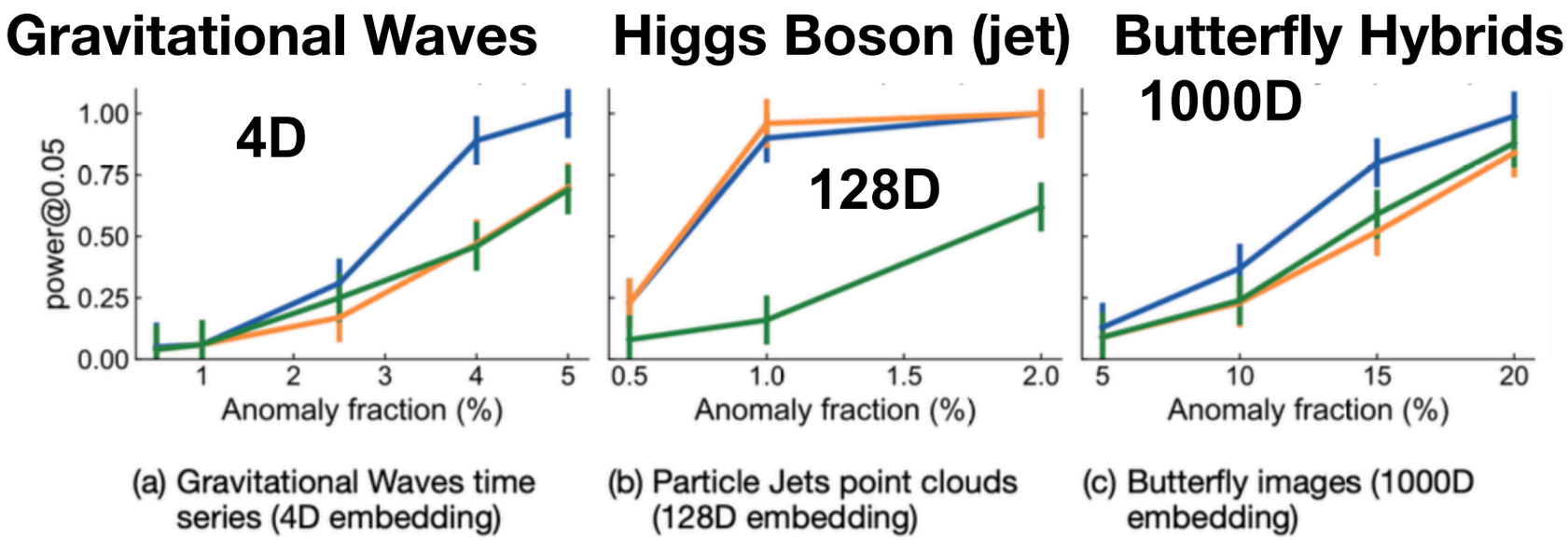




Combine high dimensional Gaussian Fit With NPLM based anomaly detection

Arxiv:2510.21935
Neruips 2025

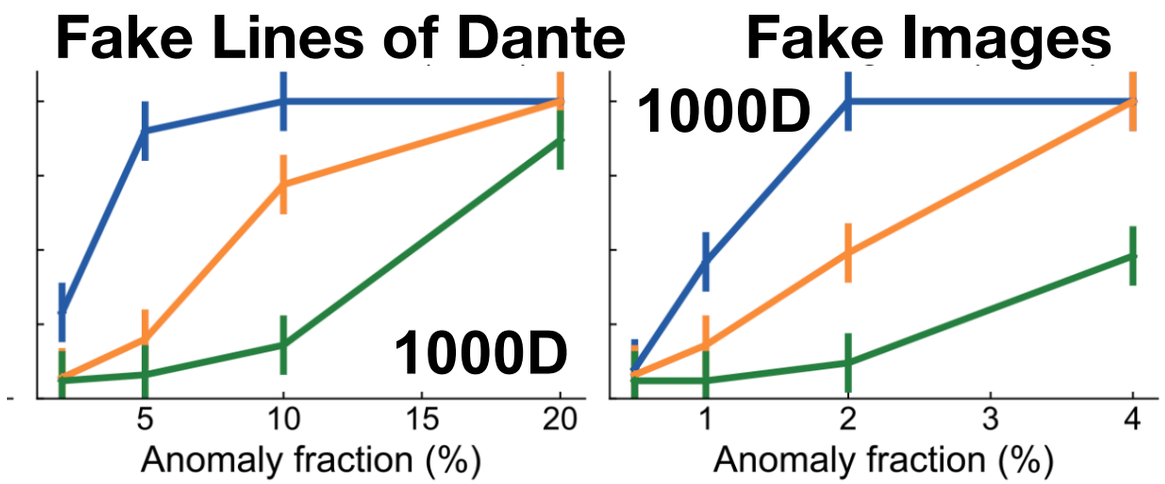
Sparker



(a) Gravitational Waves time series (4D embedding)

(b) Particle Jets point clouds (128D embedding)

(c) Butterfly images (1000D embedding)



(d) Divine Comedy lines (384D embedding)

(e) ImageNet images (1000D embedding)

Key

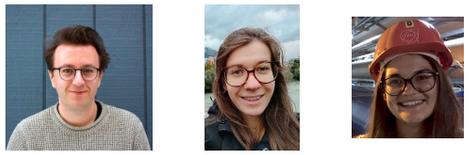
Non-Sparse Kernels

Sparse Kernels

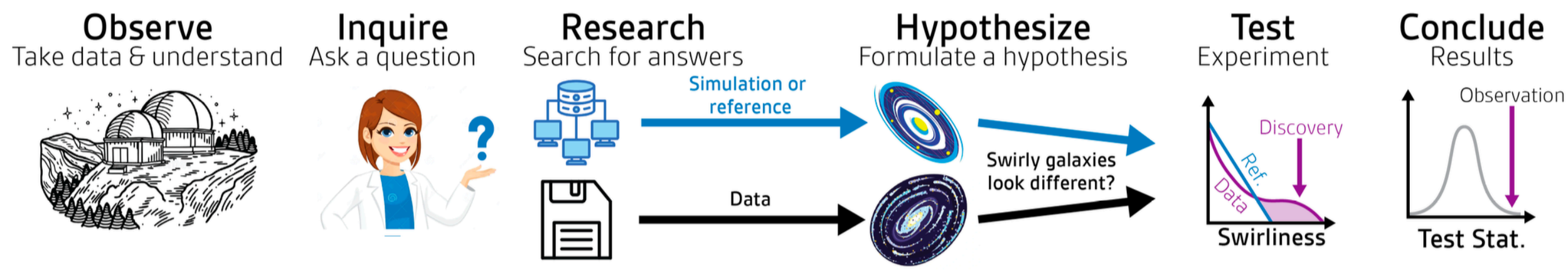
Rule/Metric Based Anomaly

Now we can look for over densities in >1000D

Arxiv:2510.21935
Neruips 2025

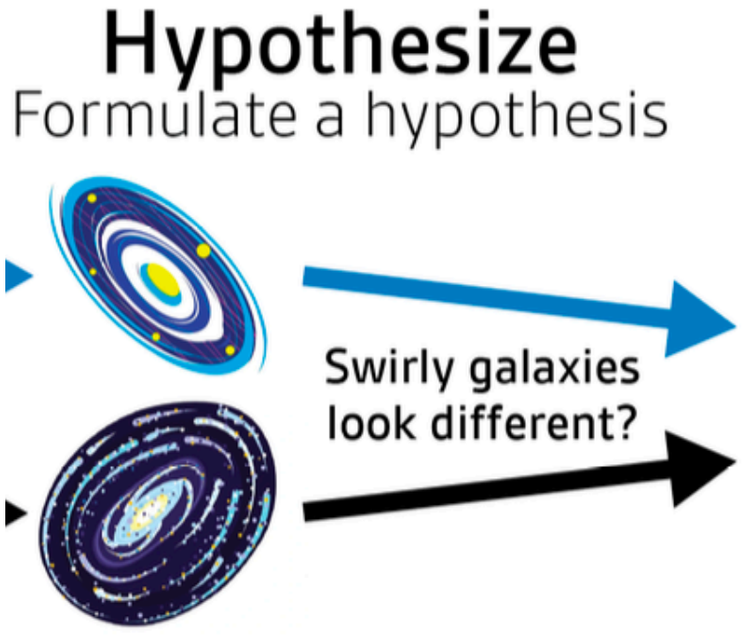


Back to the basics



The Scientific Method

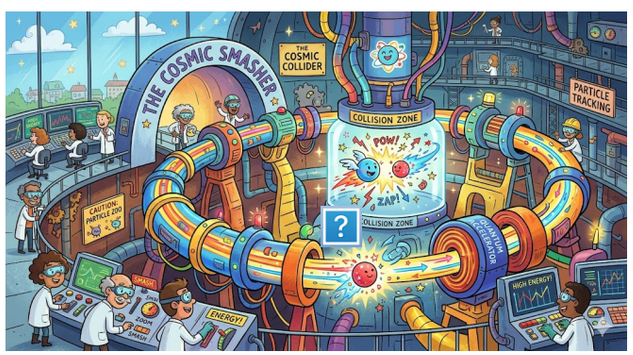
Hypothesize



- Can we compare A vs B
  - What is our data A
  - What is our reference for B
    - What are the modelling uncertainties
  - How do embed these in a hypothesis?

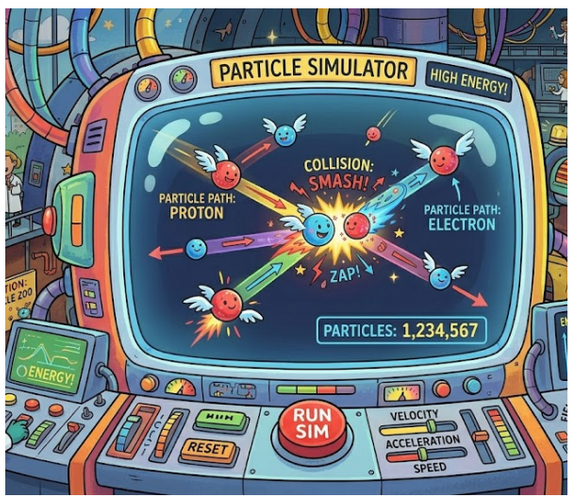
A

Data



B

Simulations



$+\sigma_{\text{model}}$

Hypothesis: Featurizing

Data

$$\begin{bmatrix} \vec{p}_1, \\ \vec{p}_2, \\ \vdots \\ \vec{p}_{1000}? \end{bmatrix}$$

Features

$$\begin{bmatrix} \text{Feature}_1, \\ \text{Feature}_2, \\ \dots \\ \text{Feature}_N \end{bmatrix}$$

Ultimately we are concerned about projecting down to the most useful scenario

Hypothesis: Featurizing

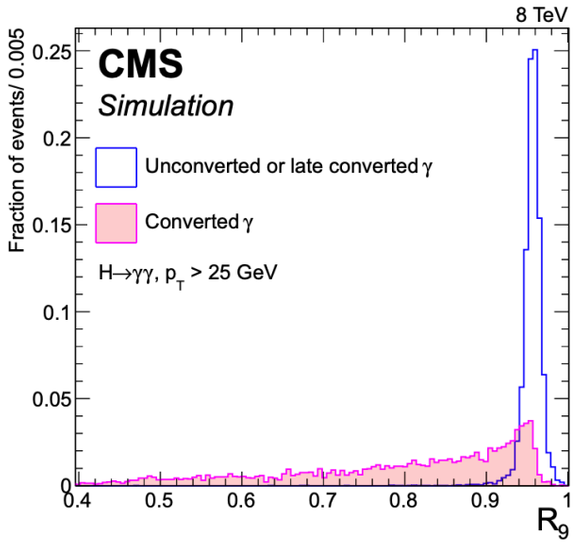
Formulating hypothesis often means featurizing

ML for
featurizer

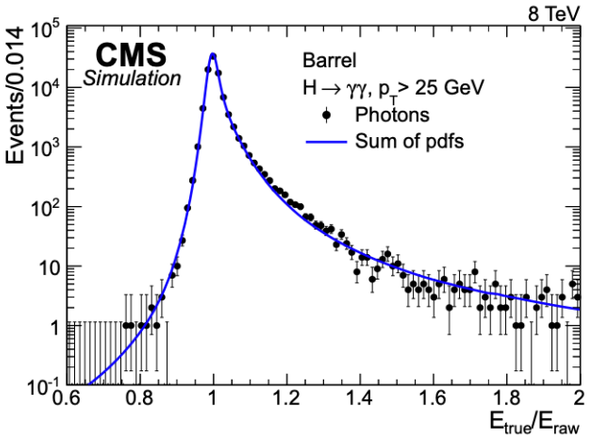
Highlight the parts that matter, ignore the rest

A 1st Result Showing This

High Level Features



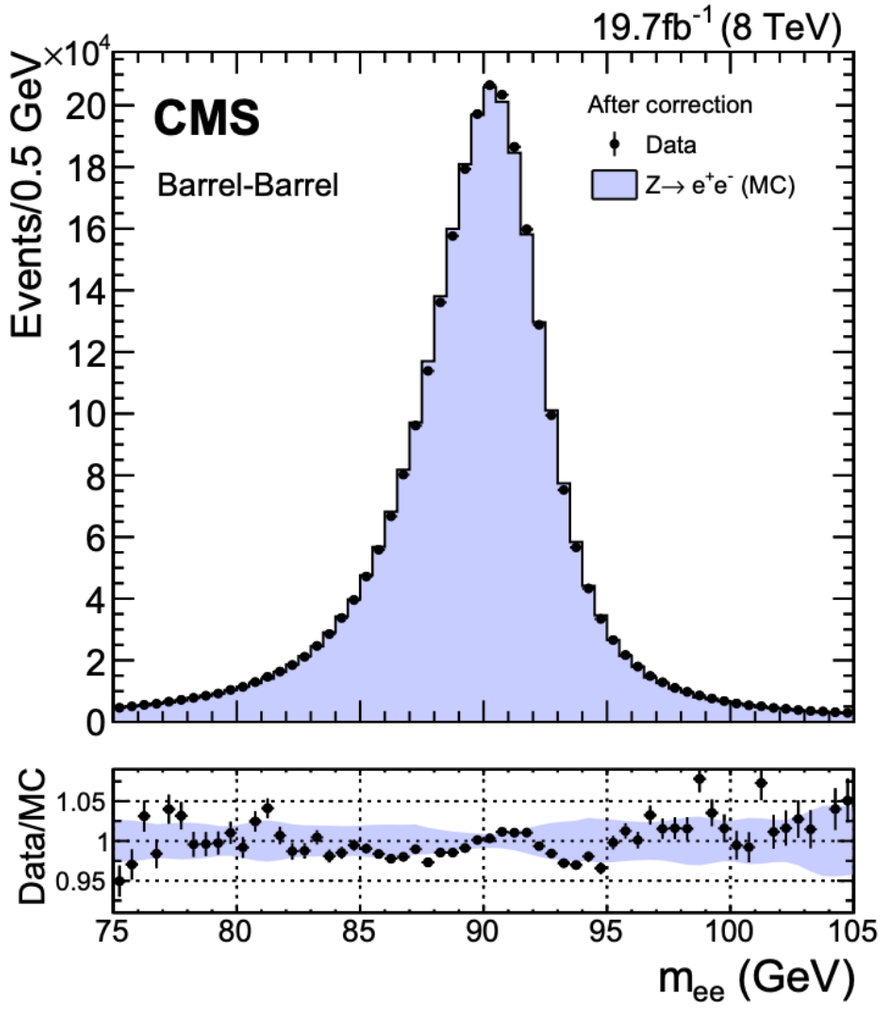
. . . .



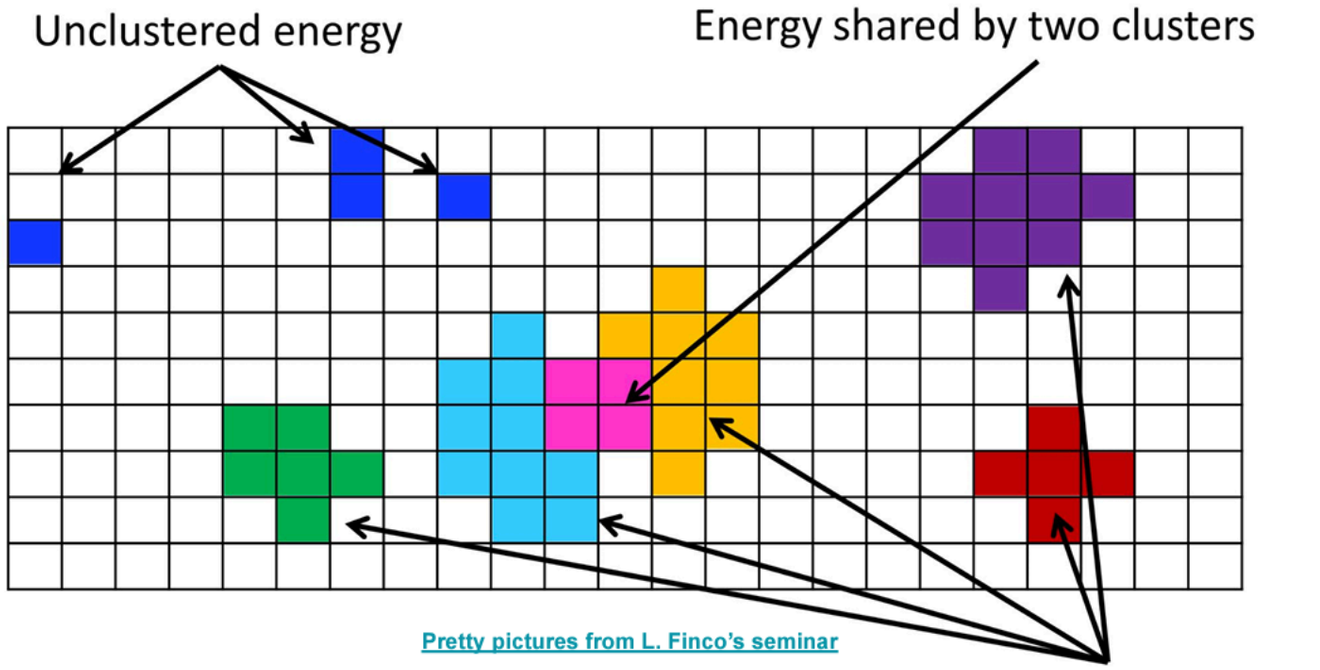
NN Regression

2012

Regressed Mass



Improved Result w/This

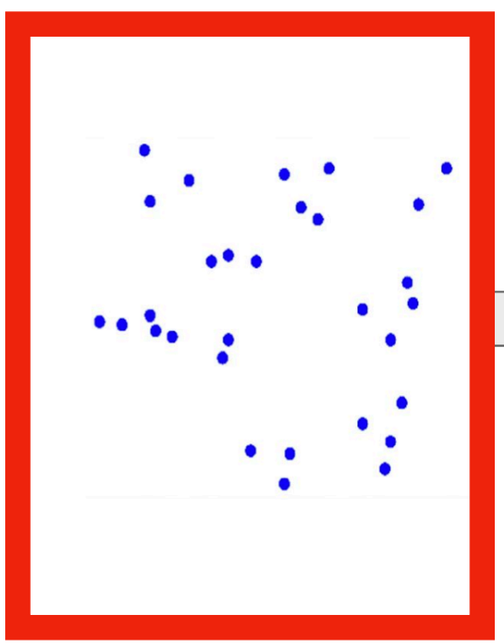


Pretty pictures from L. Finco's seminar

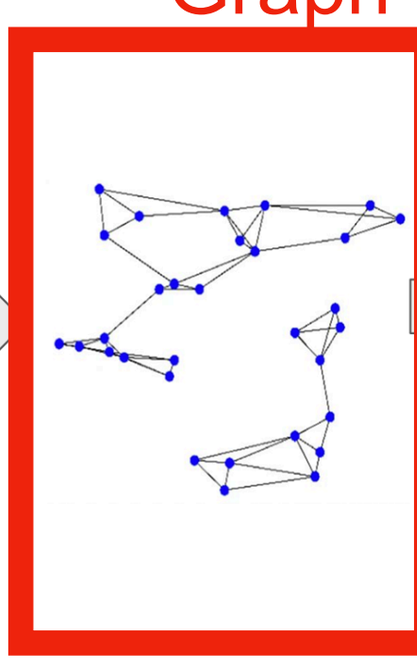
After 10 years we finally beat the high Level features

2022

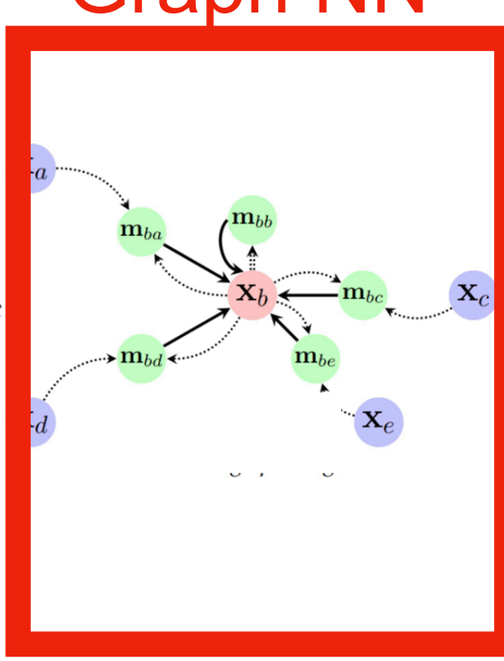
Raw Hits



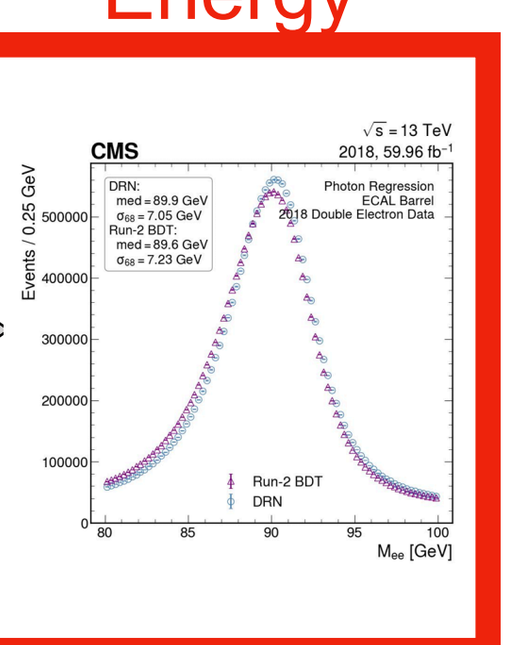
Graph



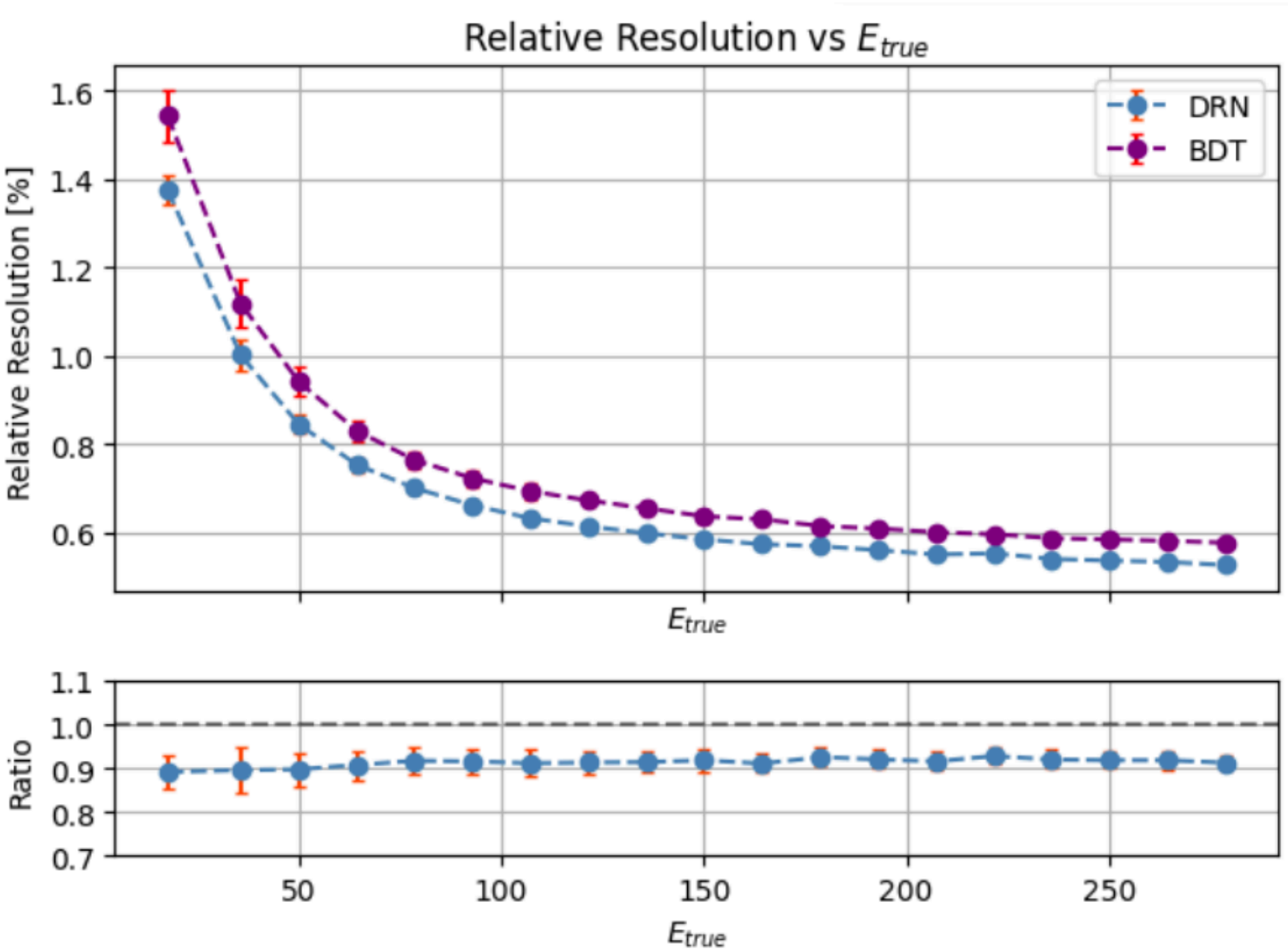
Graph NN



Energy



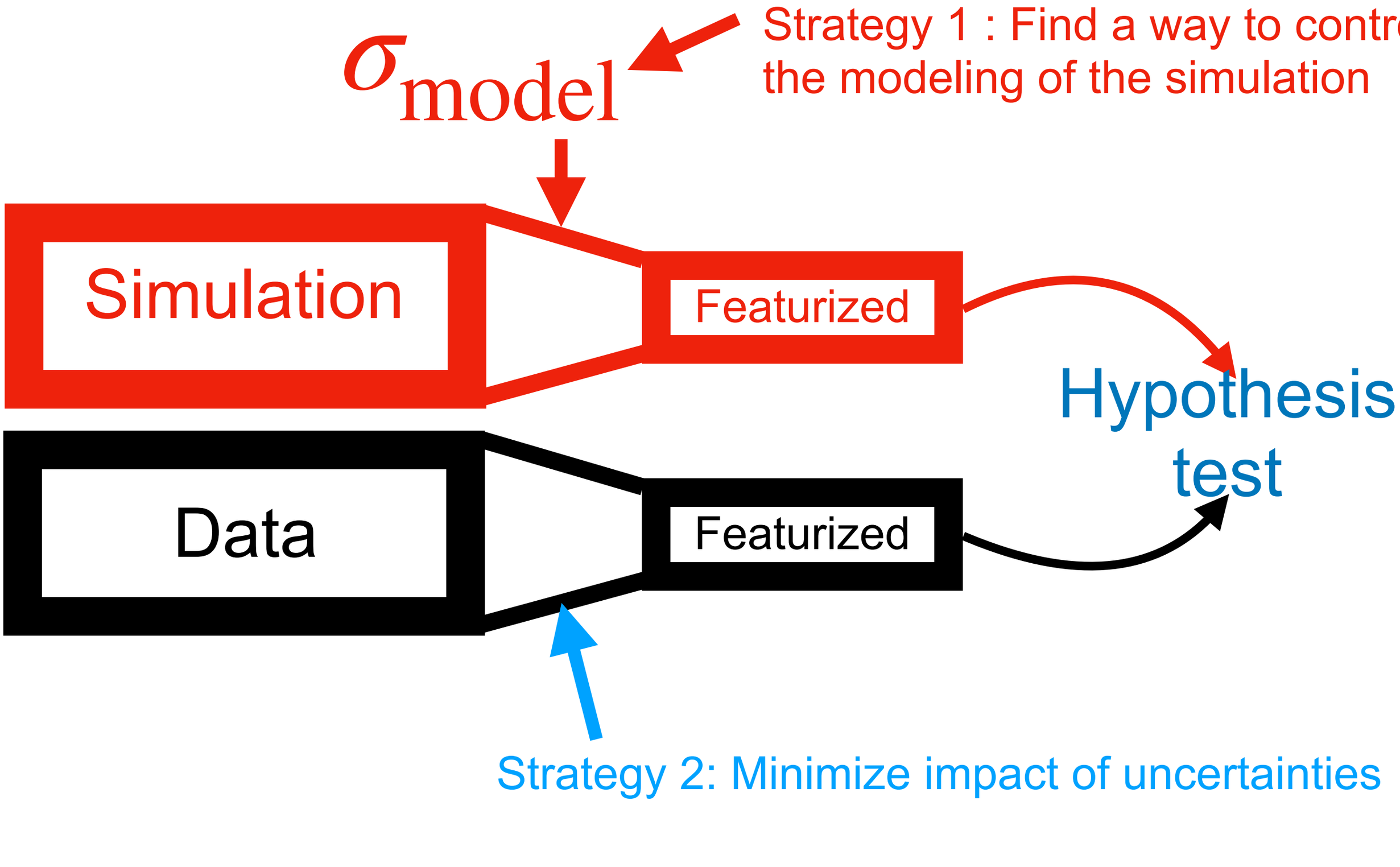
From Raw to solution



Consistent 10% improvement across the board

Choice of architecture is key insight

Improving hypothesis



Can you find the 2 differences between left and right?

Here are the differences?

What did we learn

- Can you point out which one of these paintings is Van Gogh?

What did we learn

- Can you point out which one of these paintings is Van Gogh?

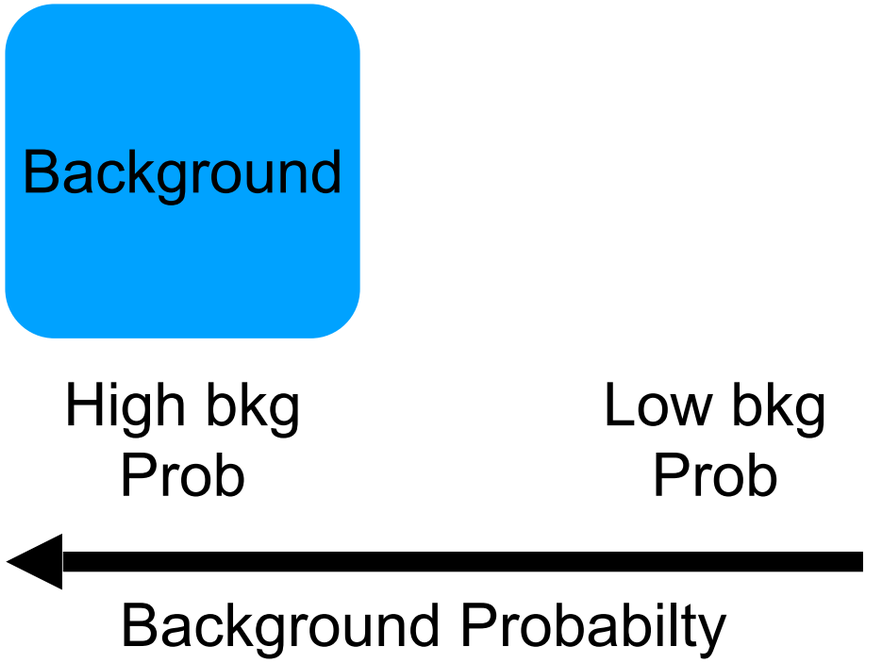
This one has characteristic features of the previous one

The analytic approach

Line them up and Pixel by Pixel difference

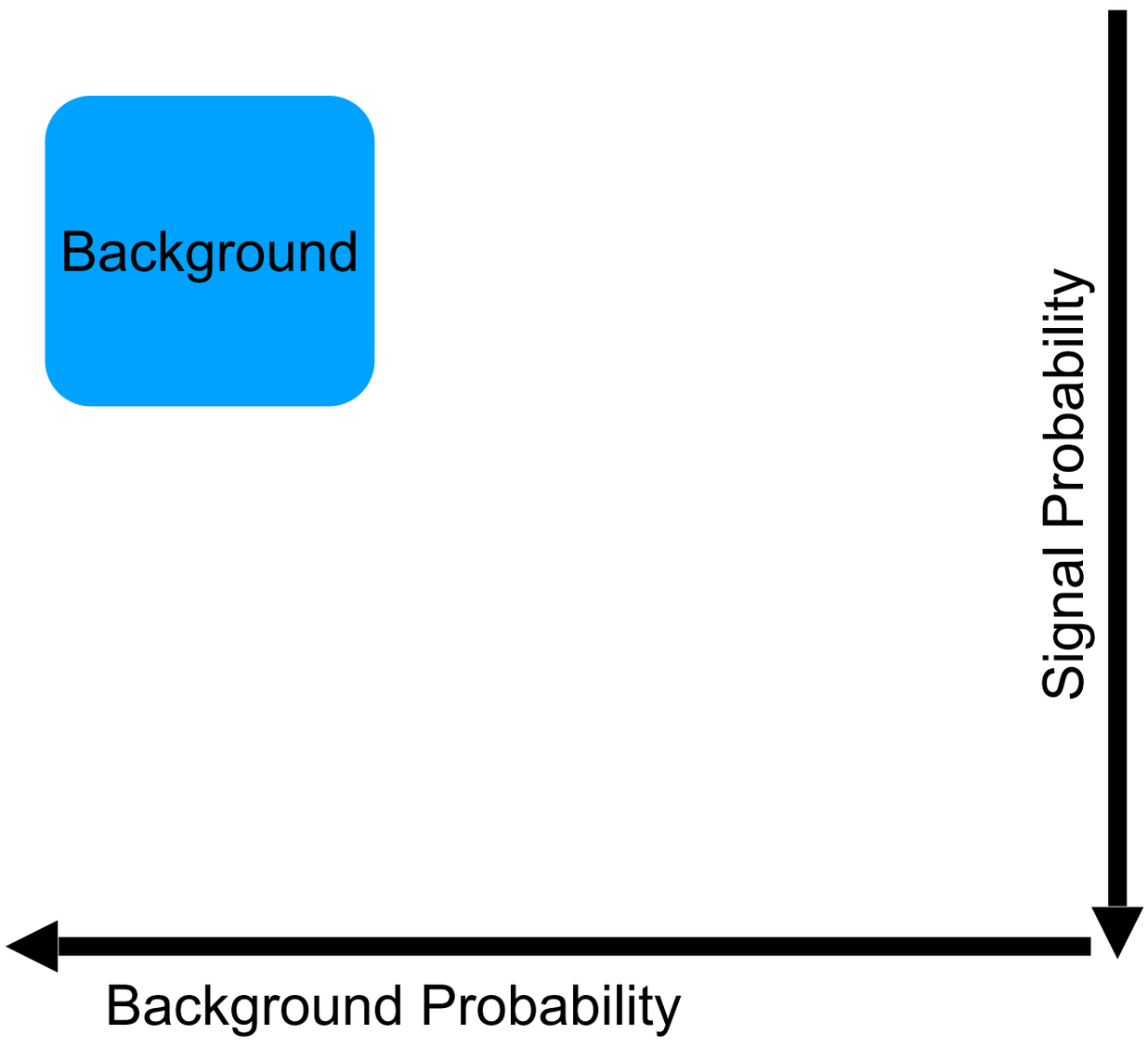
QUasi Anomalous Knowledge

Look at each sample and compute the probability distribution



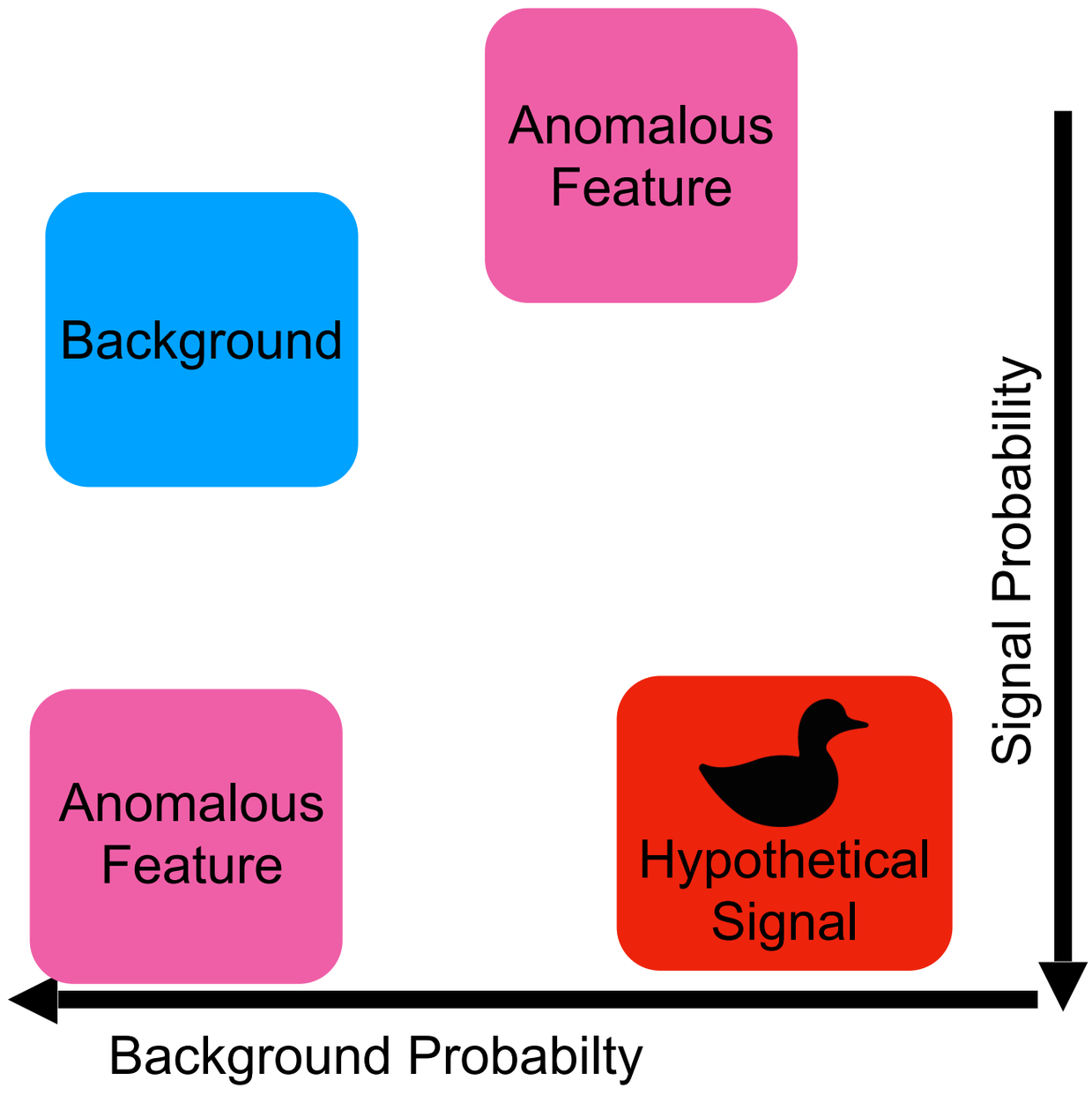
Normalizing Flows/autoencoders
Used for probability

QUasi Anomalous Knowledge



arxiv/2011.03550

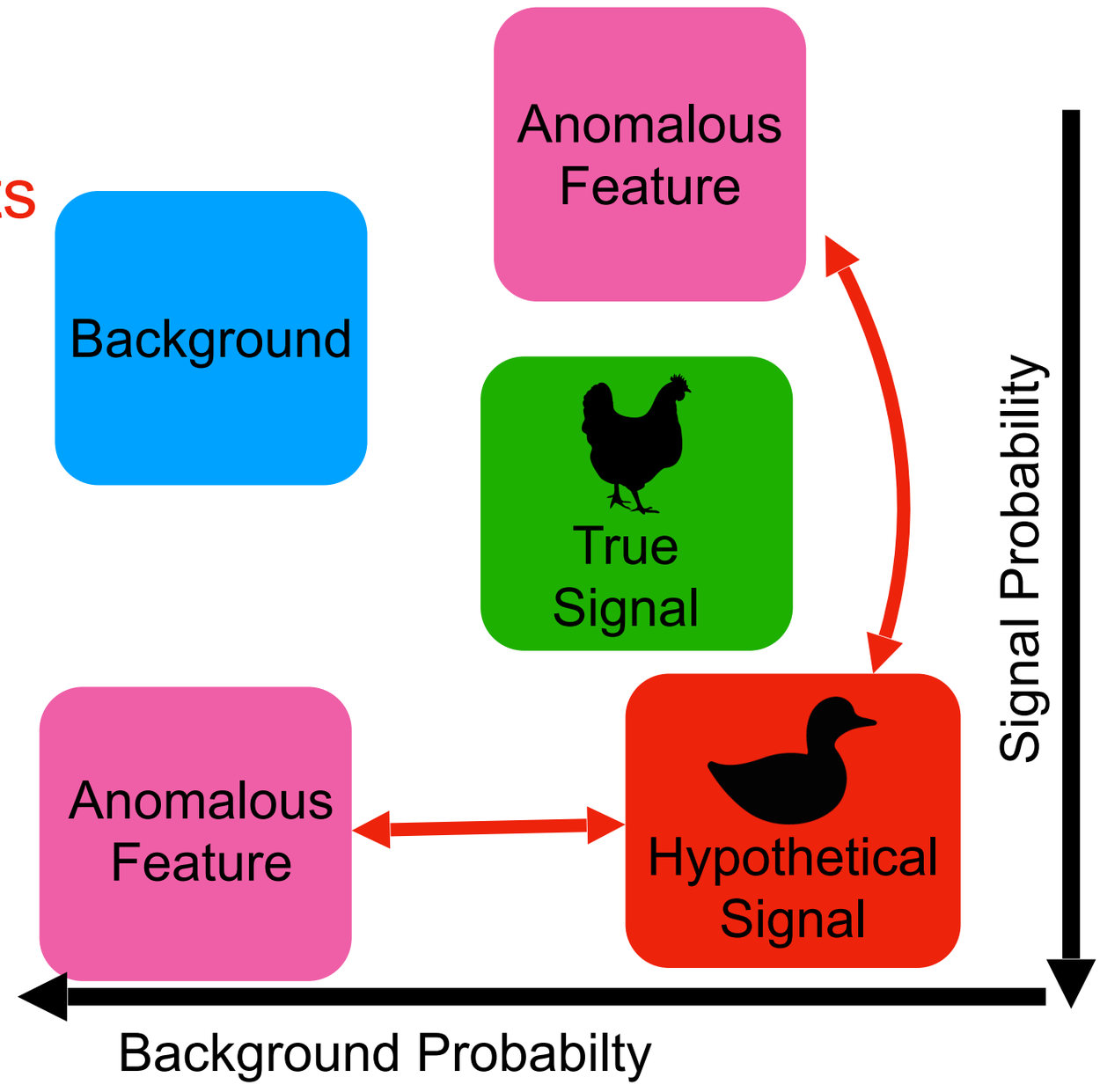
QUasi Anomalous Knowledge



2D QUAK Space

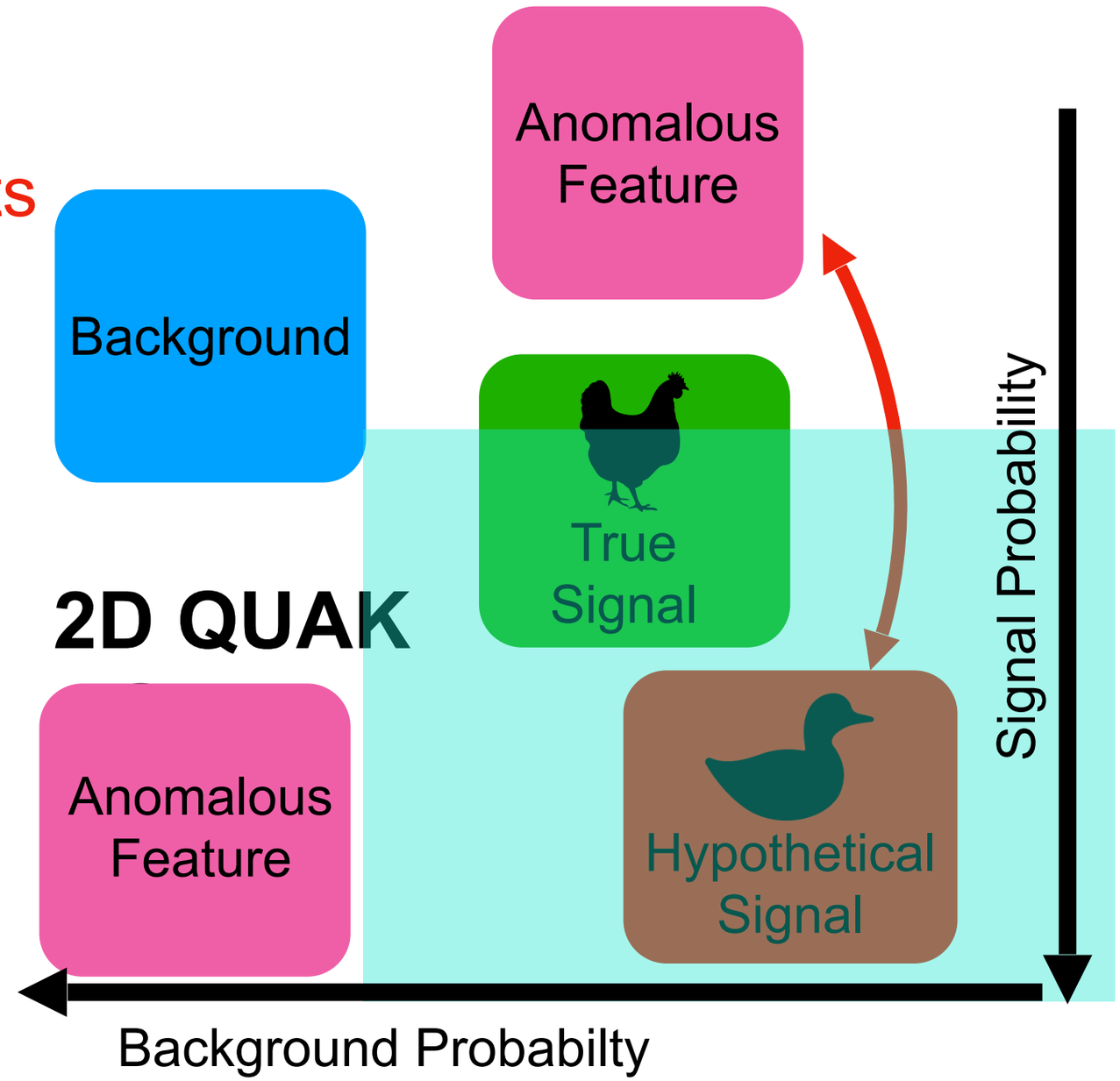
QUasi Anomalous Knowledge

Adding Signals splits
Signals through 2D

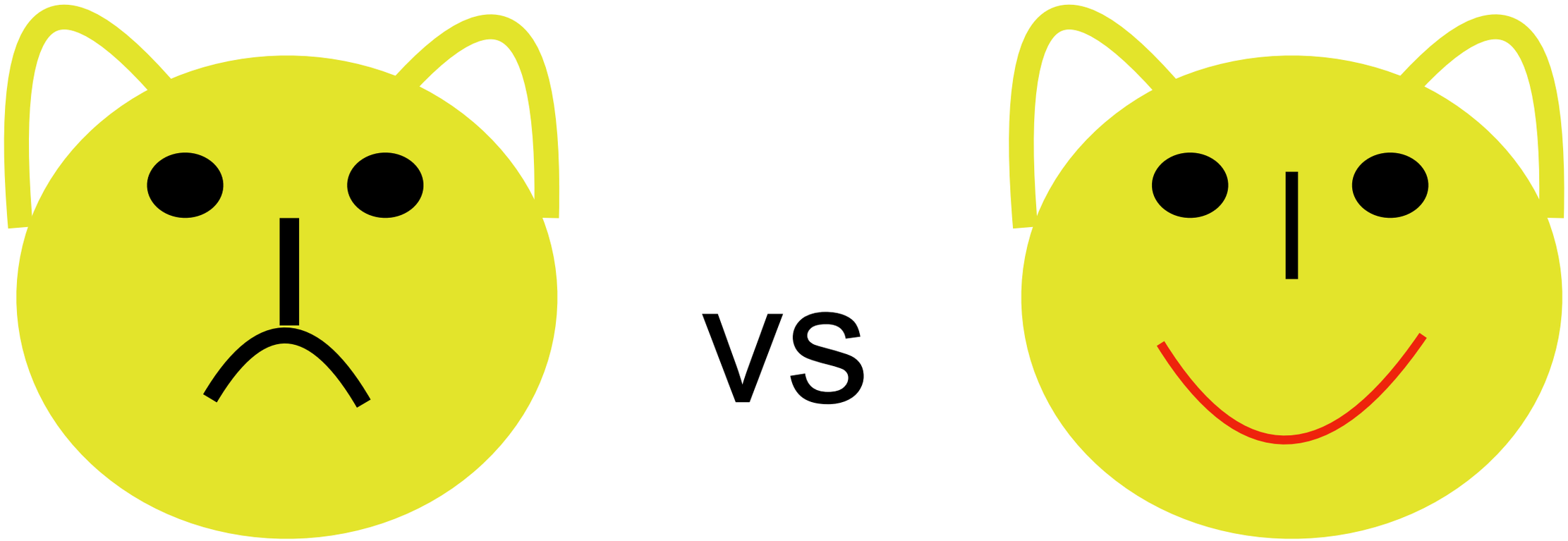


QUasi Anomalous Knowledge

Adding Signals splits Signals through 2D

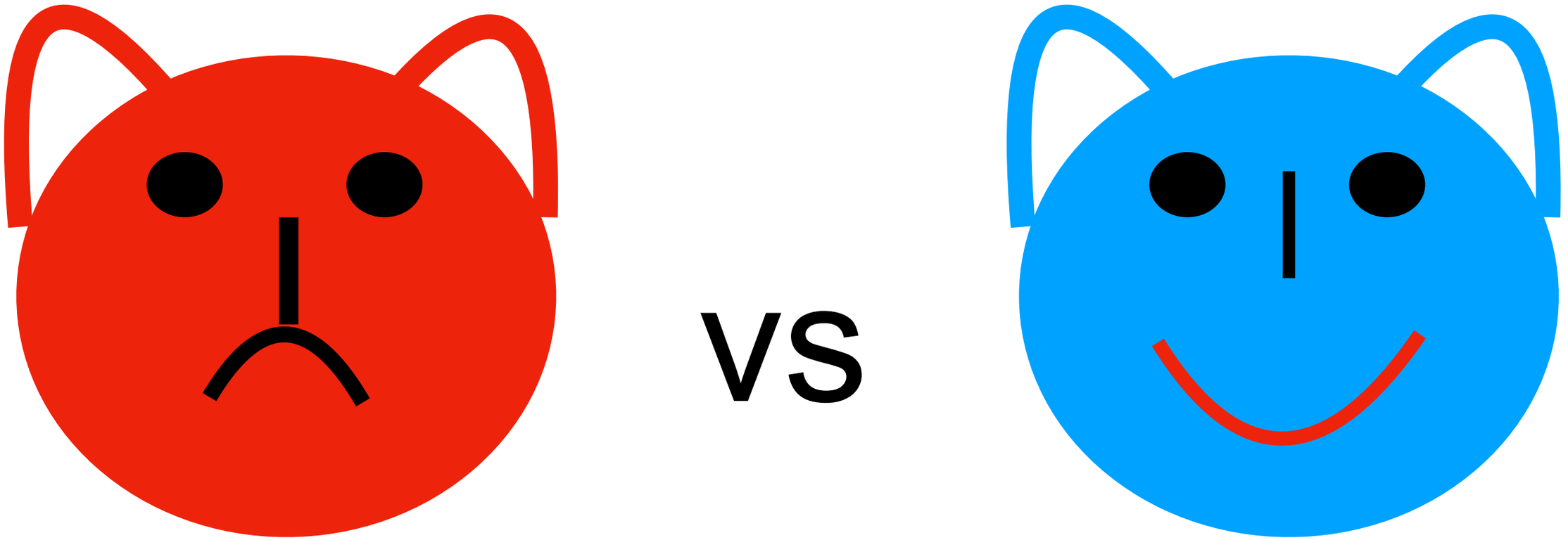


Lets illustrate a tagger

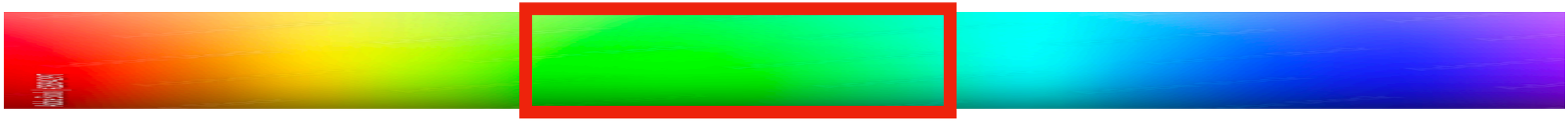


- We are going to discriminate a happy face form a sad

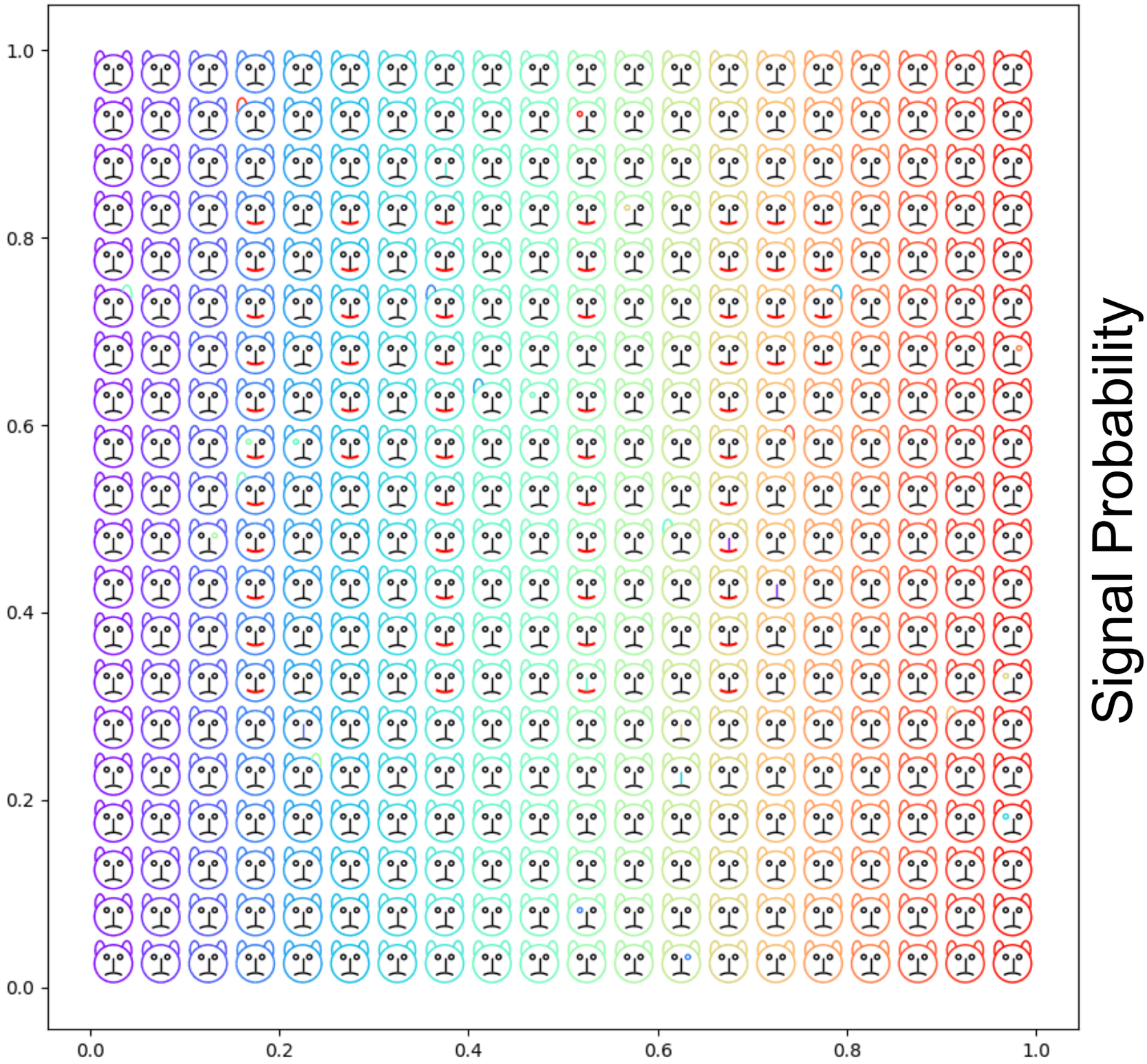
Lets illustrate a tagger



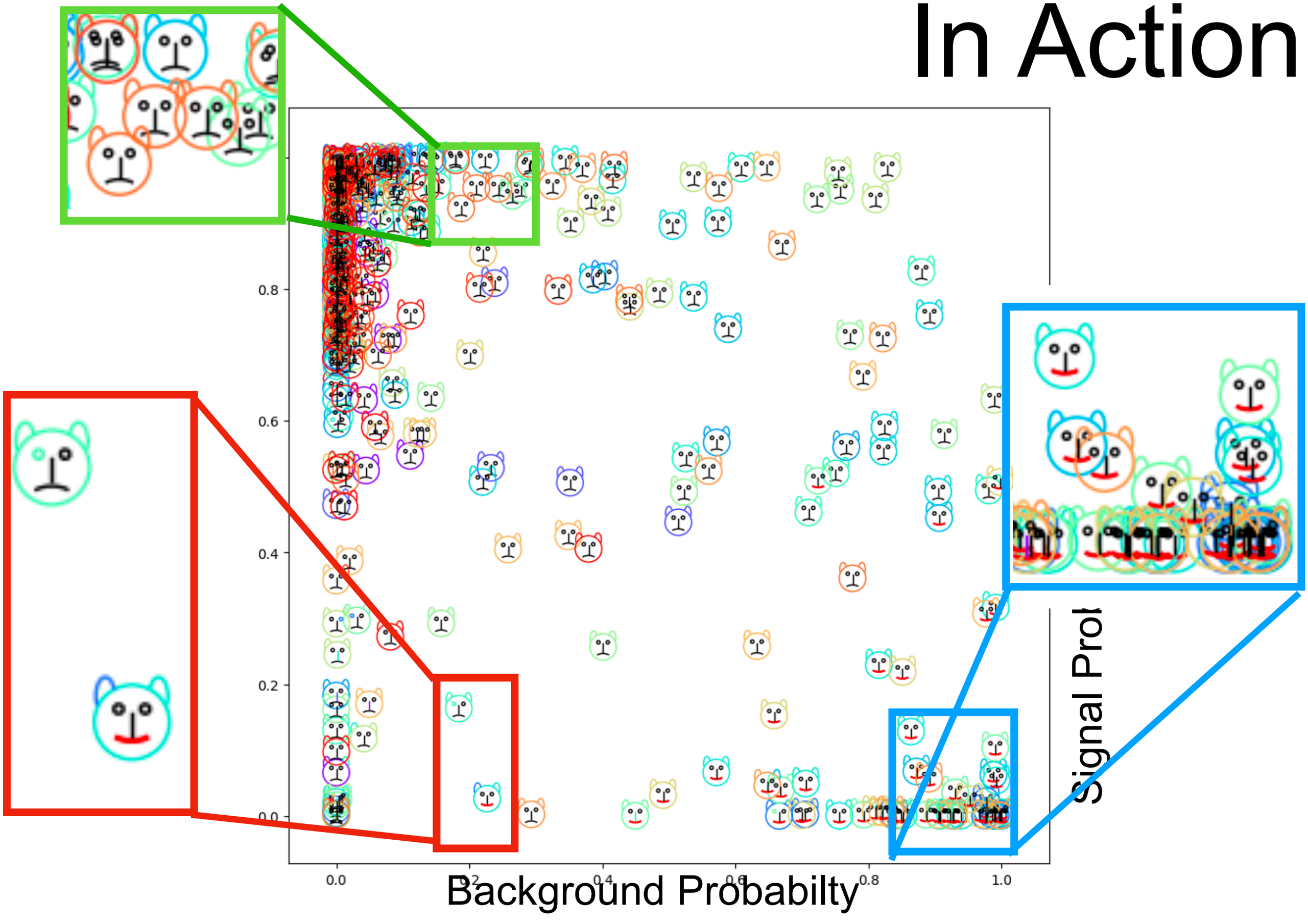
- We are going to discriminate a happy face form a sad
- They will vary in color, signal will be in center of spectrum



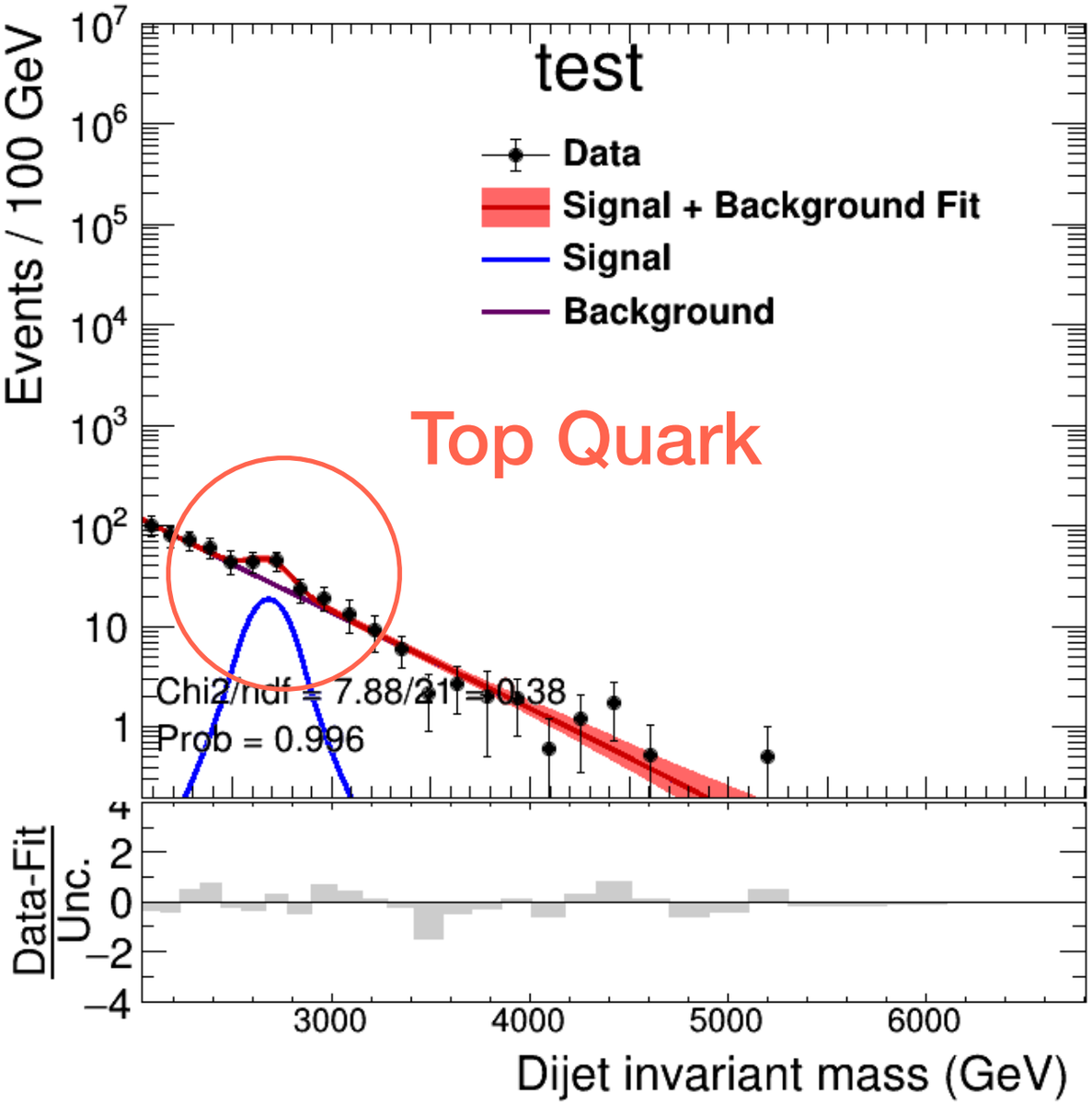
In Action


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Signal Probability
Background Probabilty

In Action


0.0
0.2
0.4
0.6
0.8
0.0
0.2
0.4
0.6
0.8
1.0
Background Probabilty
Signal Prol

Does it work?



We found a 3 sigma excess in our validation region

This was not expected, turned out to be top quarks

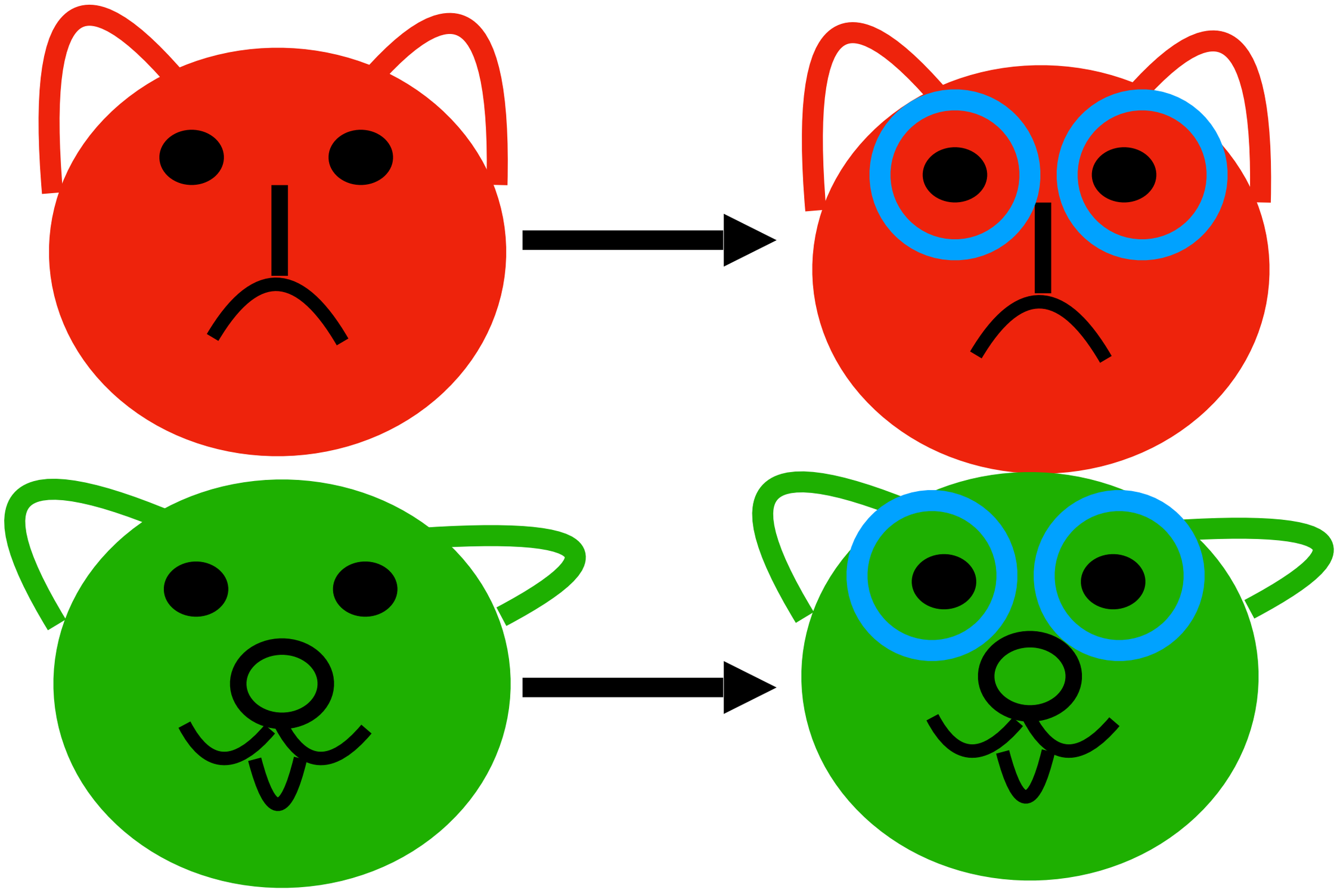
Can we generalize this?

Image

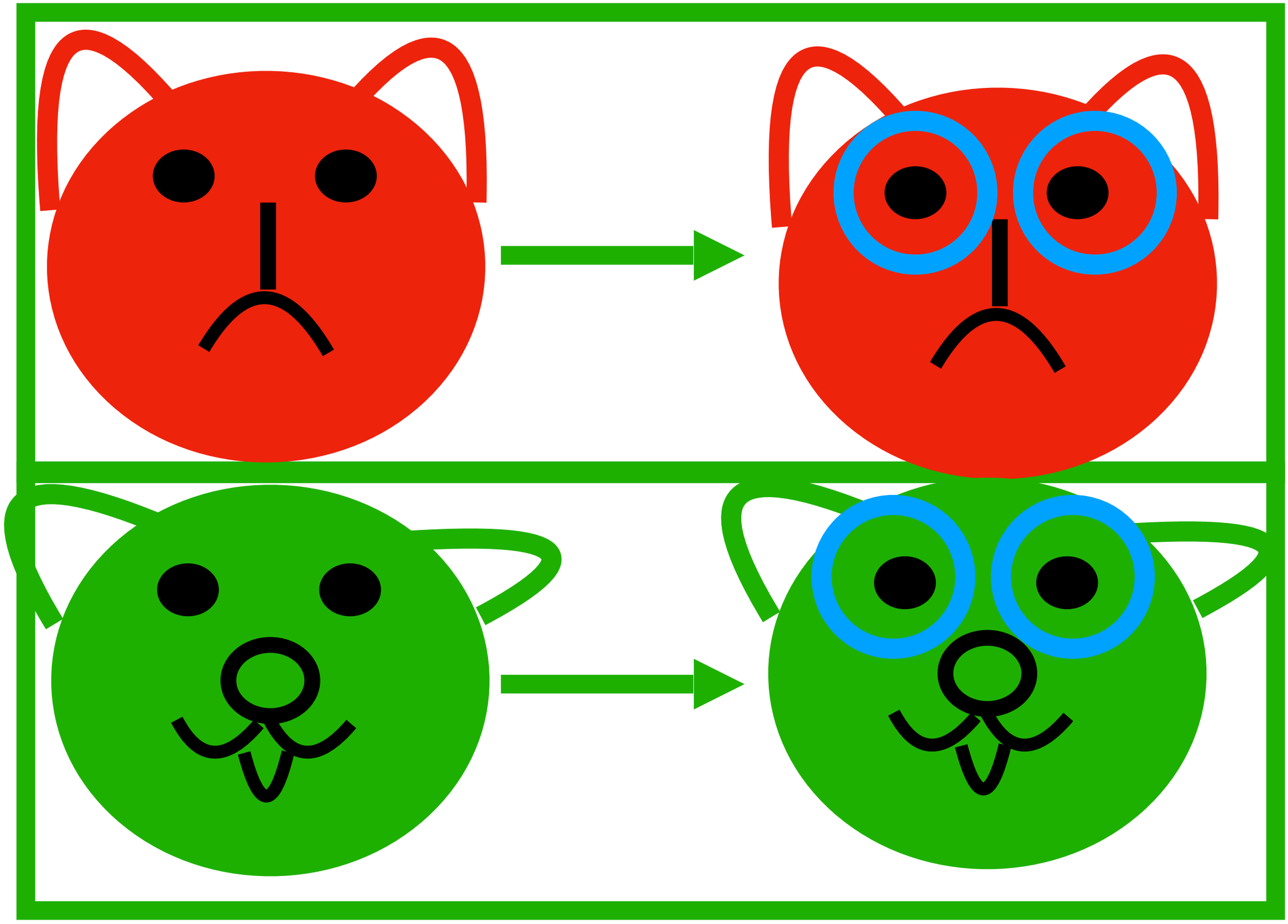
Augmented Image

Take Objects & Modify slightly

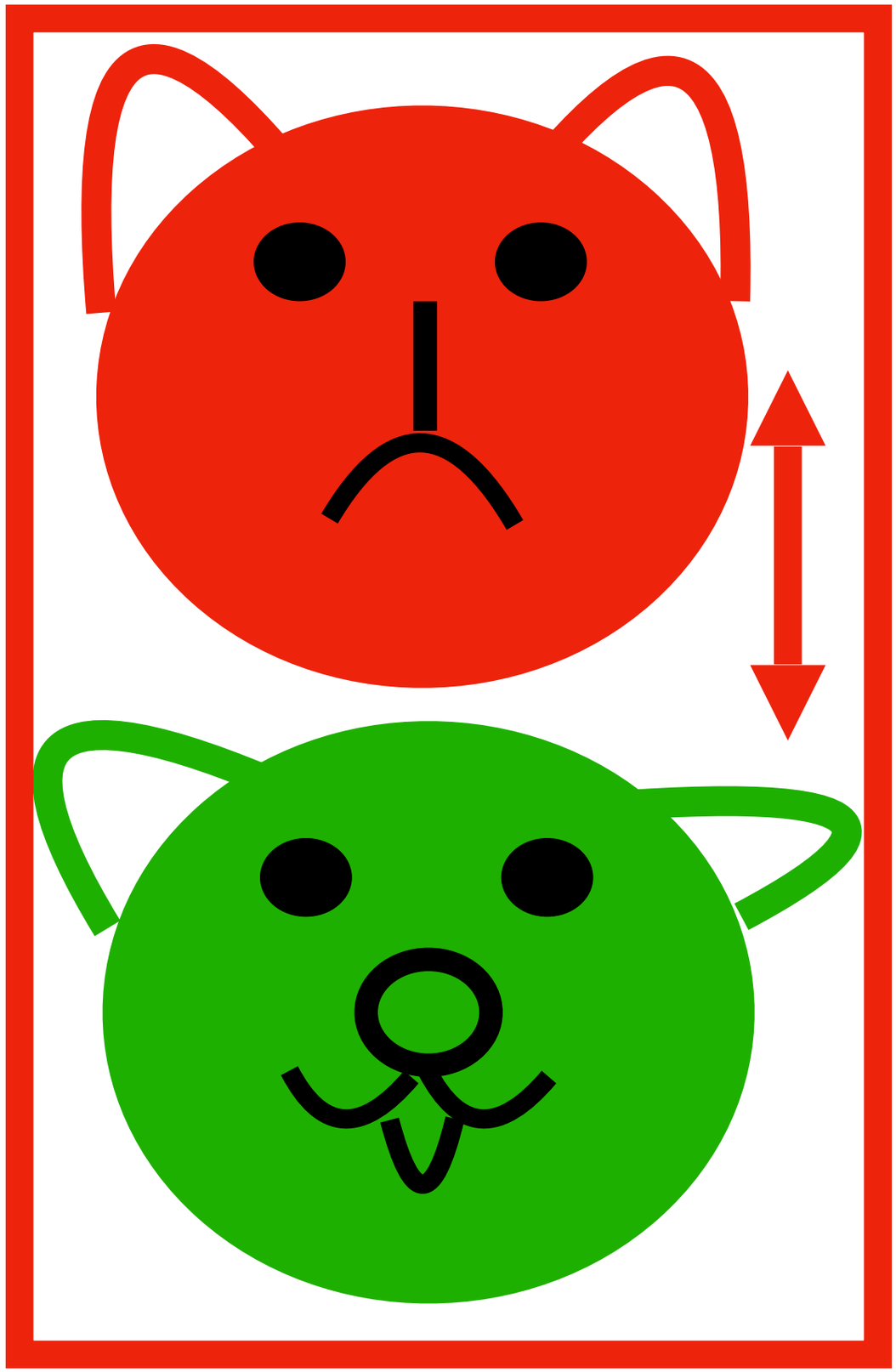
Take Objects & Modify slightly

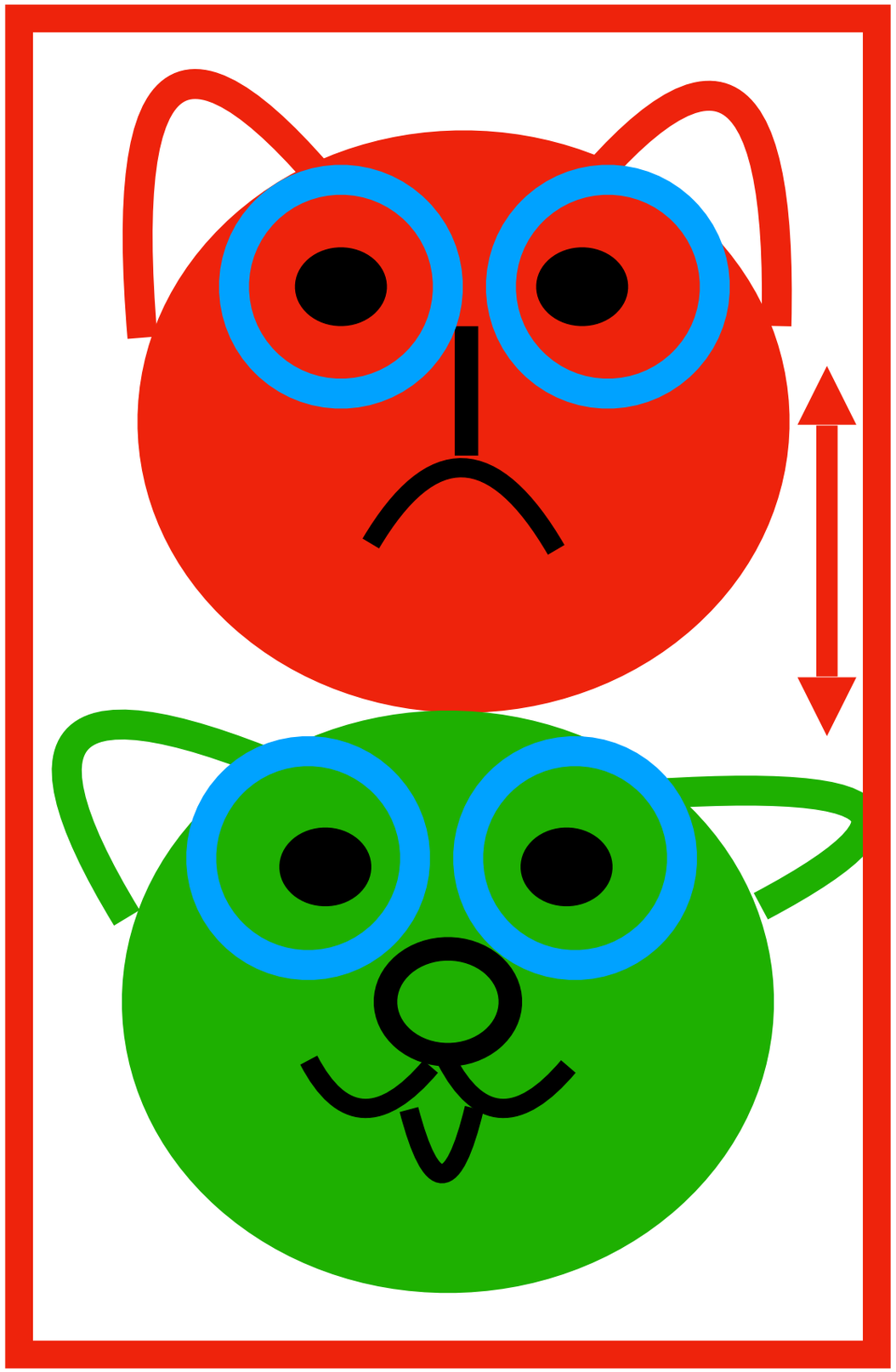


Minimize



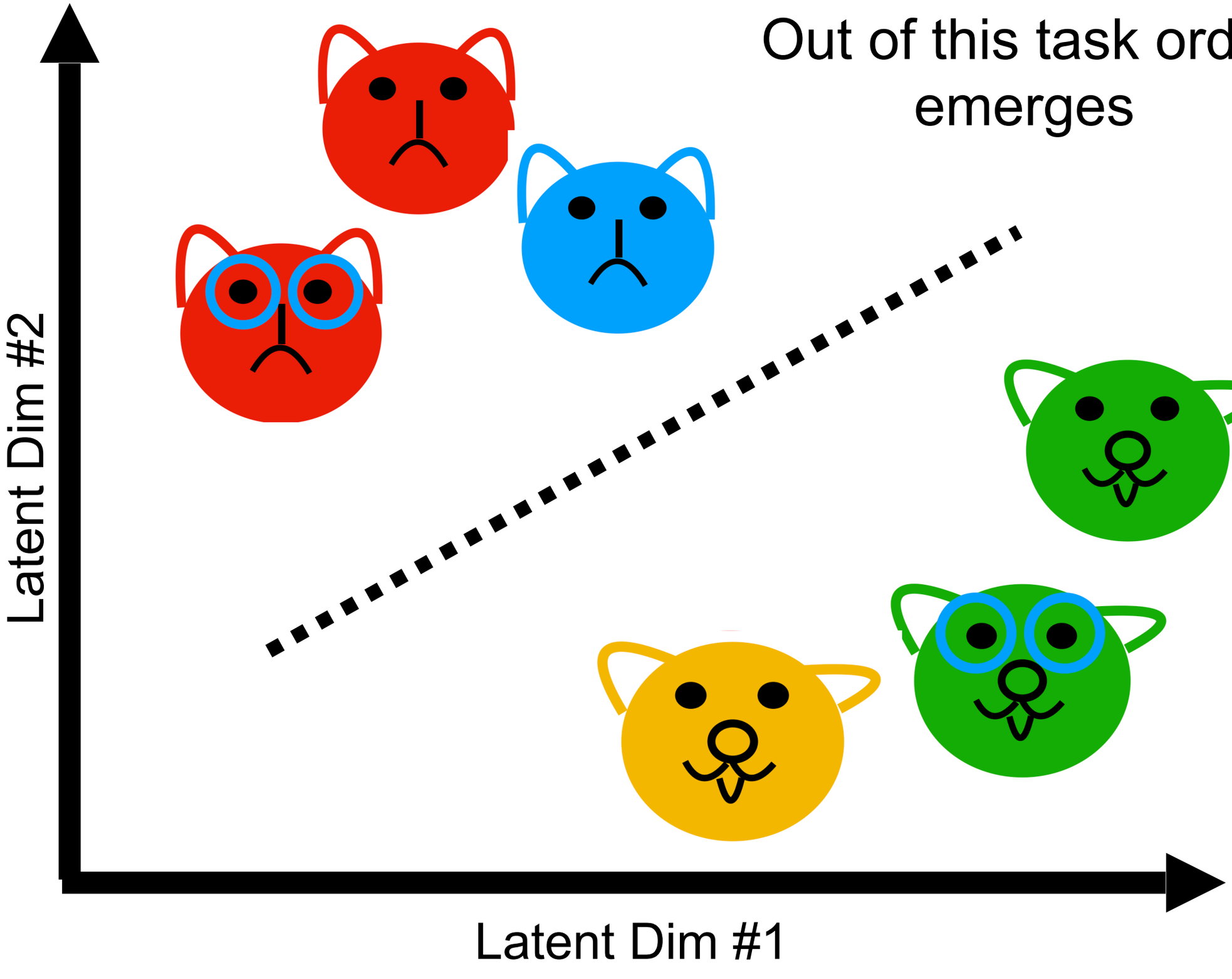
Maximize





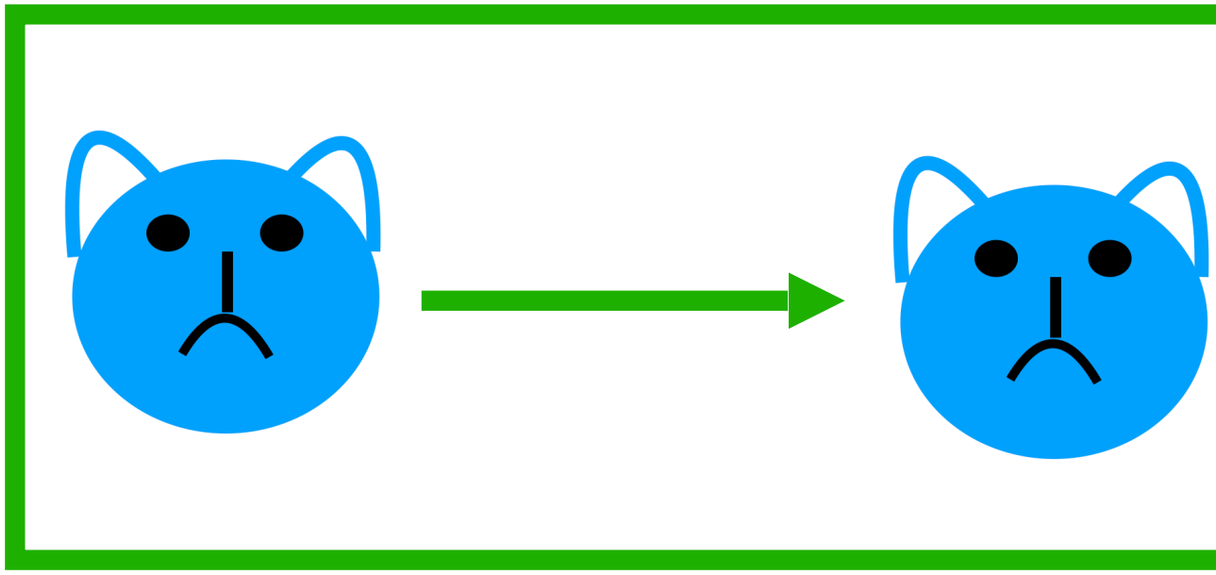
A space

Out of this task order emerges



Whats the right augmentation?

Shift in X or Y direction
In our Original Space

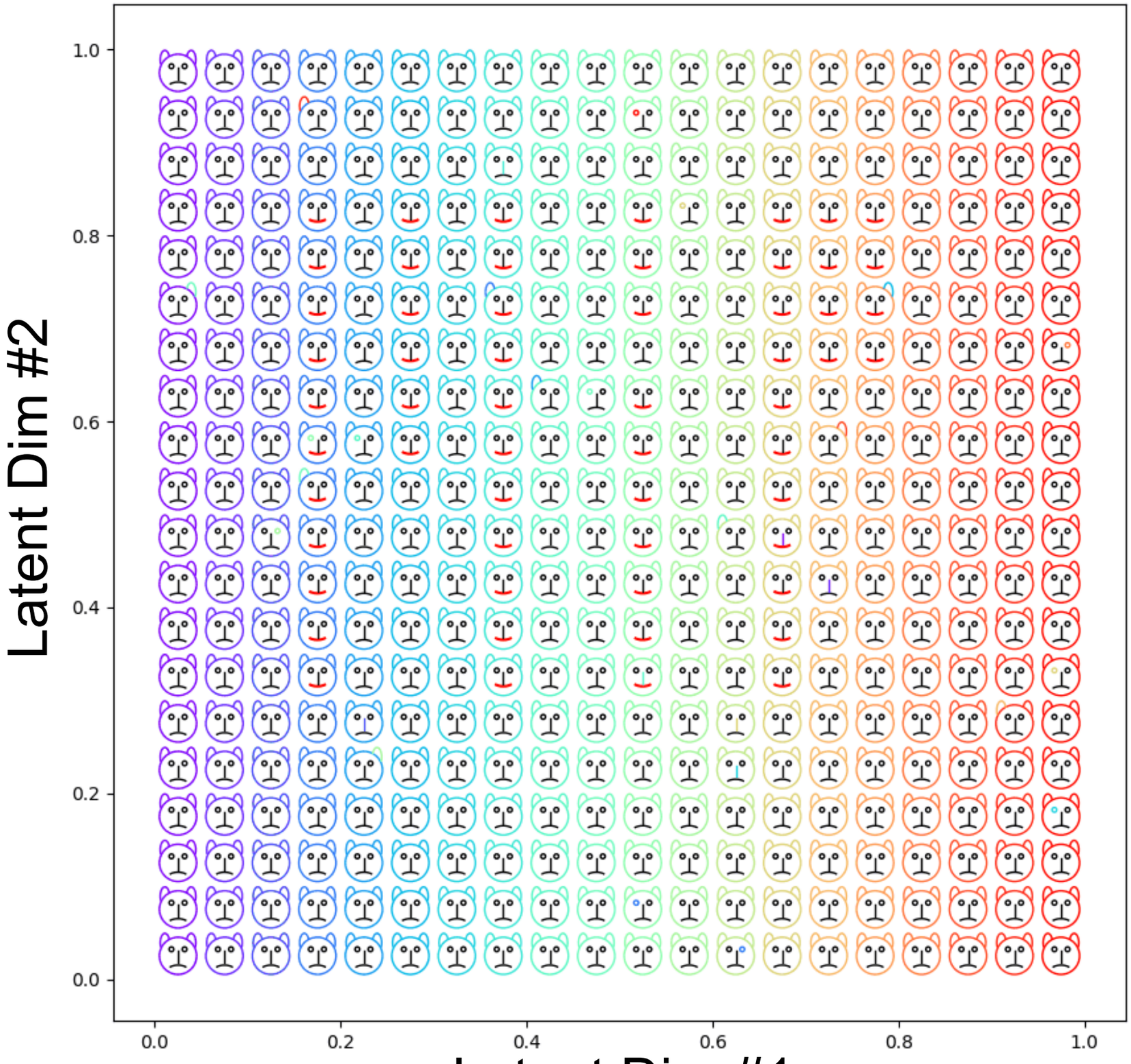


Should give same result

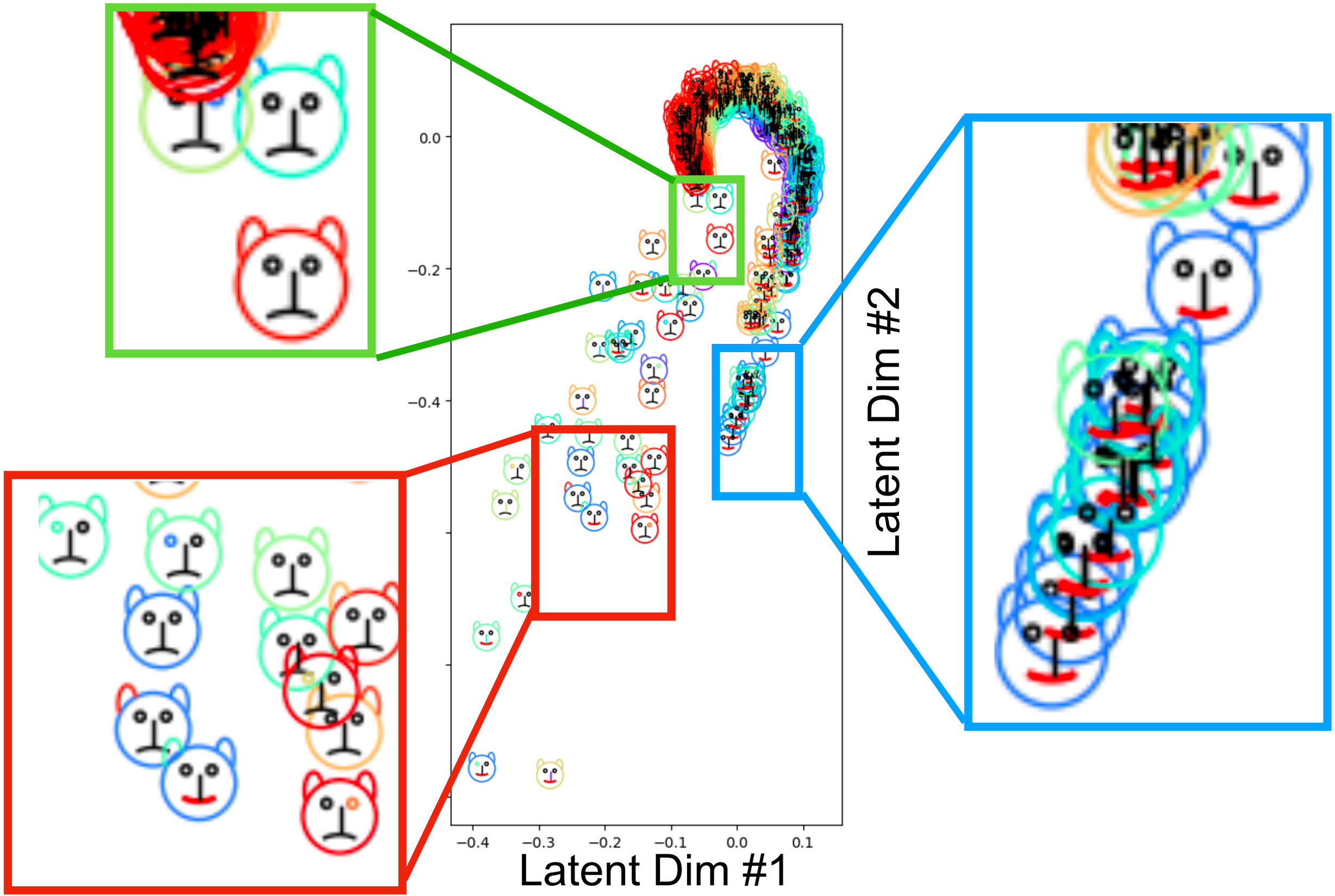
Noether’s theorem

Embedding translation **symmetry invariance**
as a learning strategy
This also makes us colorblind

In Action



In Action



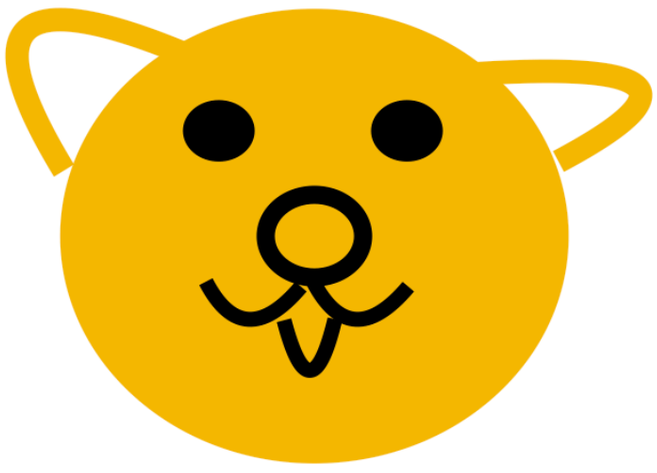
Arxiv:2403.07066

Re-simulation

Cats

Higgs

Dogs



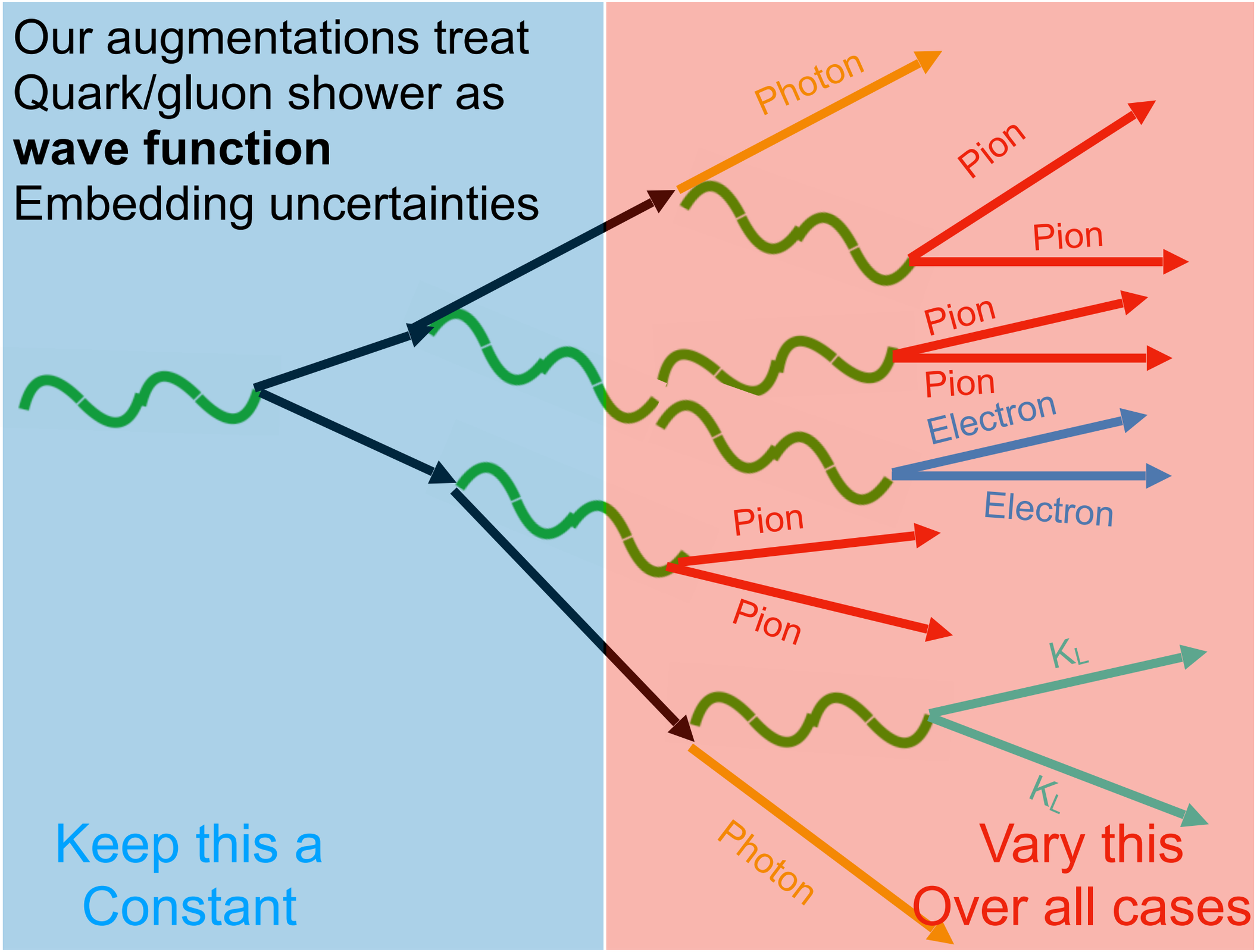
Quarks and Gluons

Arxiv:2403.07066

Embedding QM

Our augmentations treat Quark/gluon shower as **wave function**

Embedding uncertainties



Keep this a Constant

Vary this Over all cases

Augmentation

Higgs

b-quark

b-quark

Baseline

Higgs

b-quark

b-quark

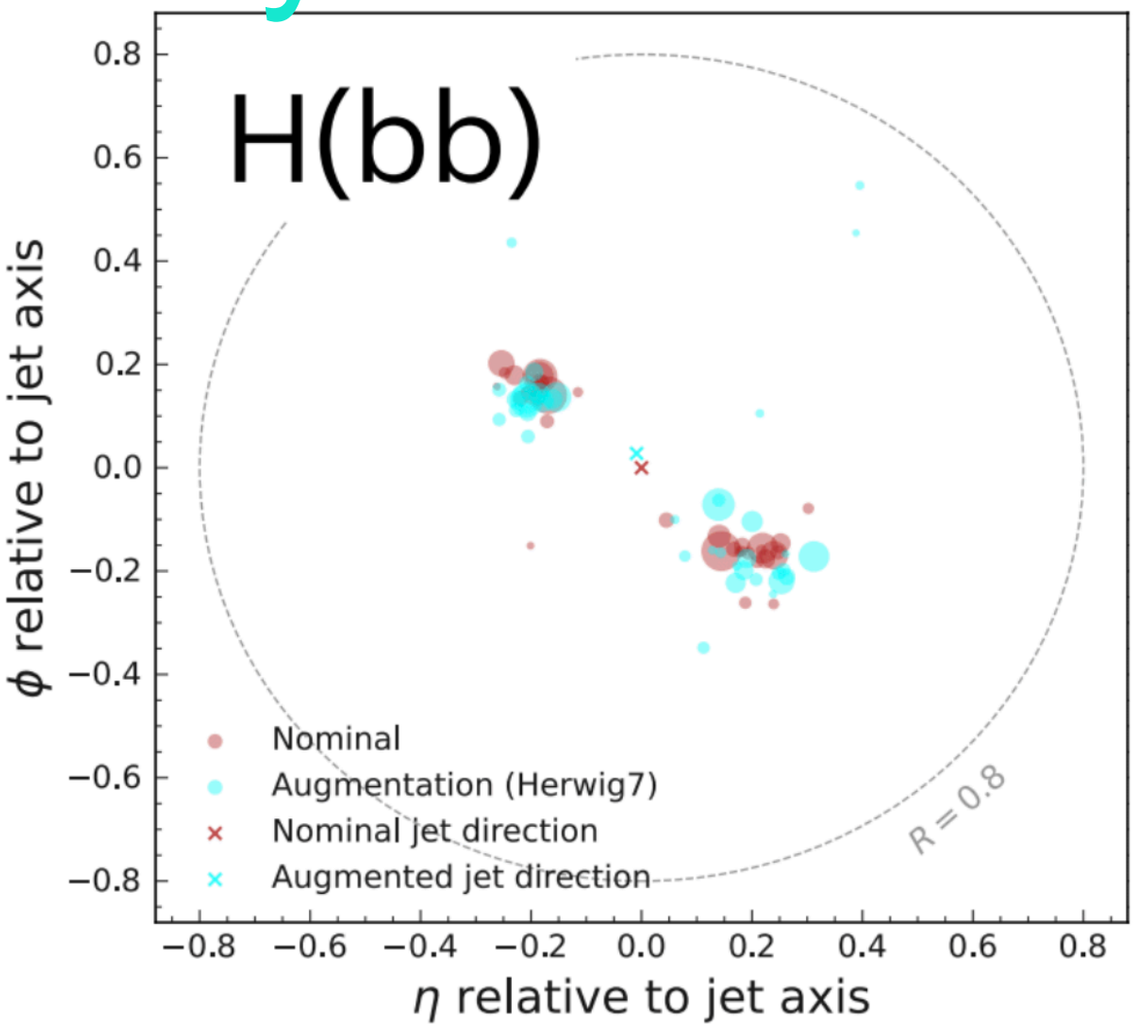
Augmented by Reshowering

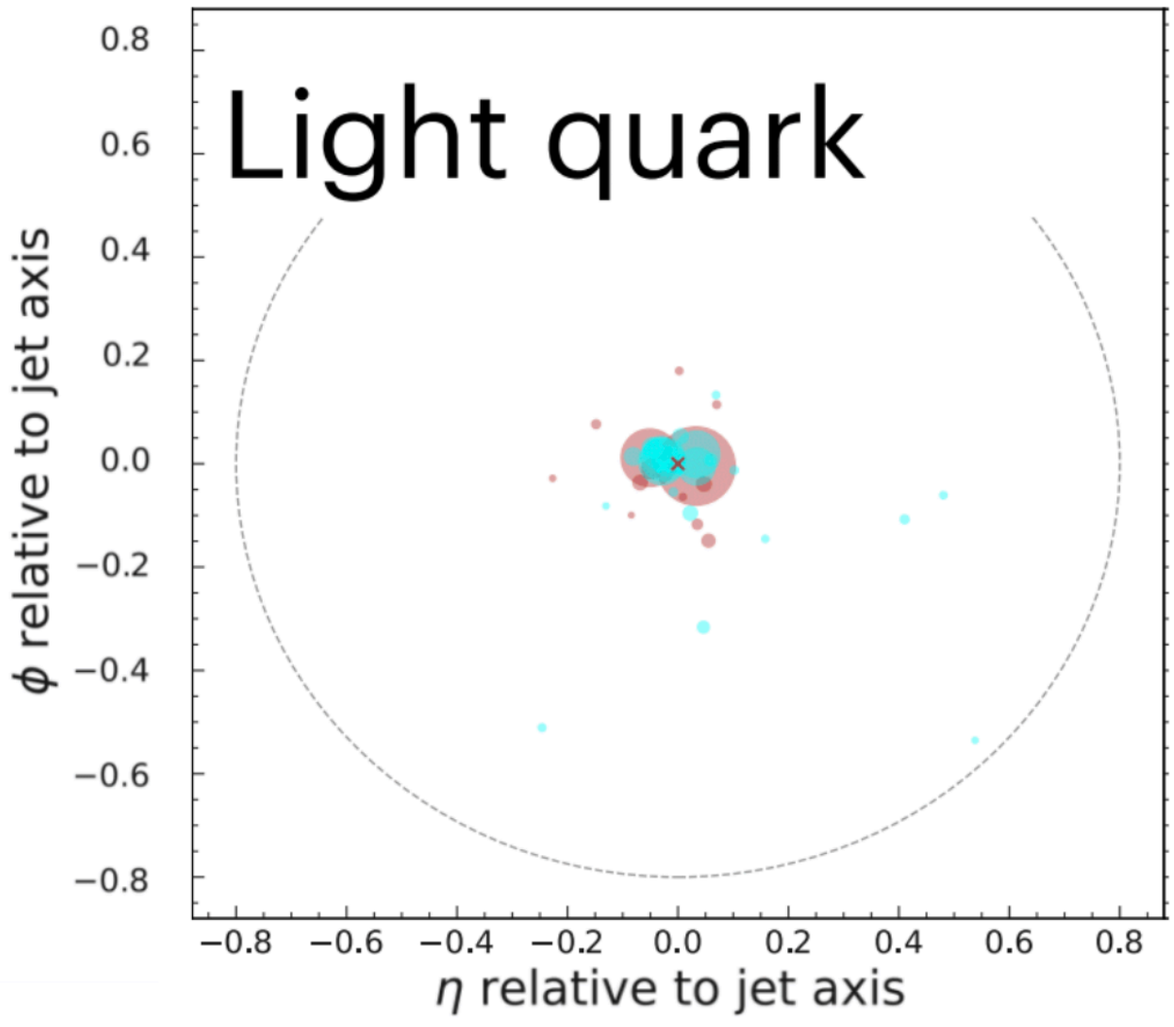
Require these to be the same in the space

Arxiv:2403.07066

Visualizing: R3sL

Cyan → Red

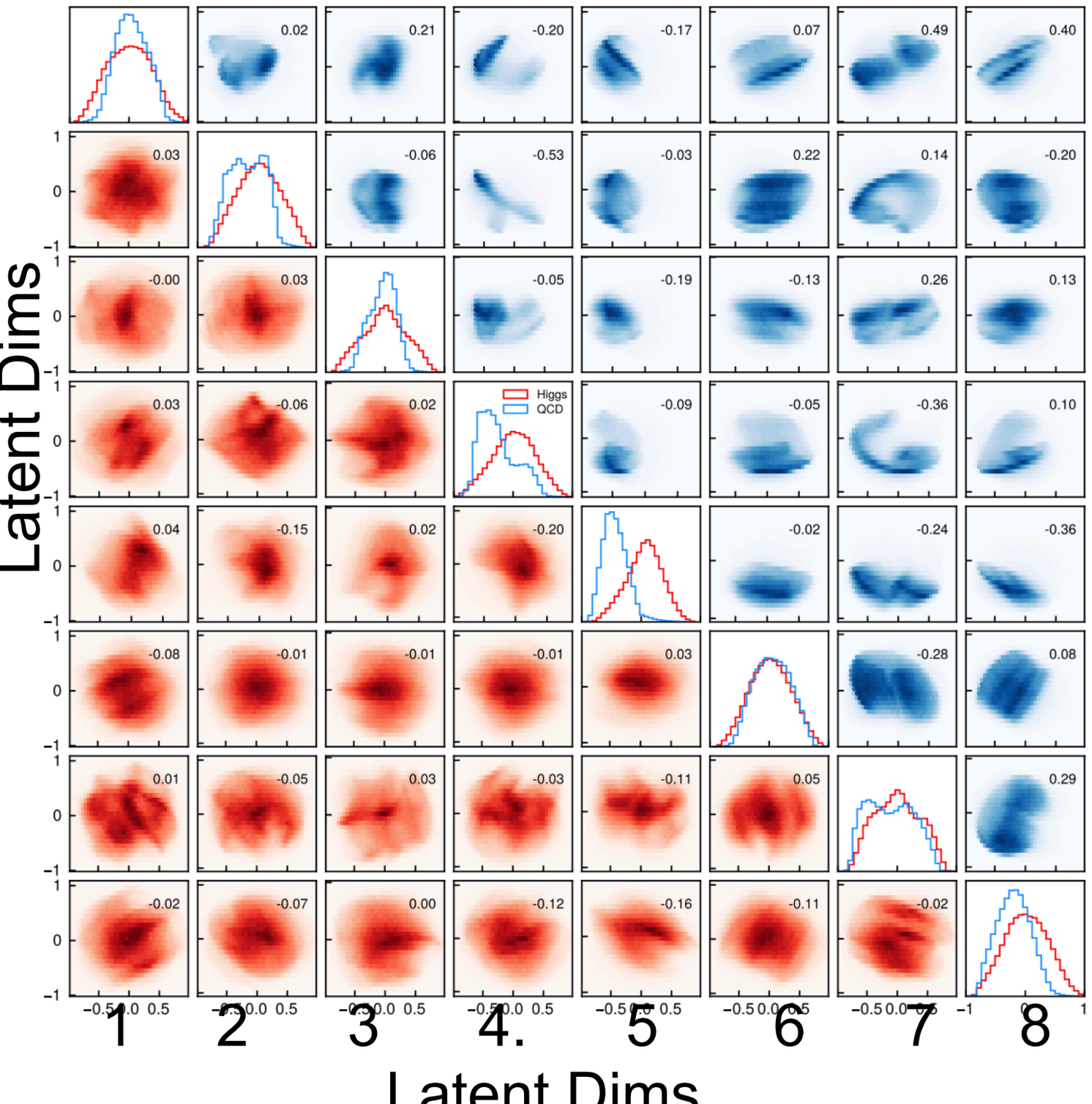




Start with a jet and reshower it

Essentially embedding a notion of Quantum. Mechanics

Embedded Space



Higgs

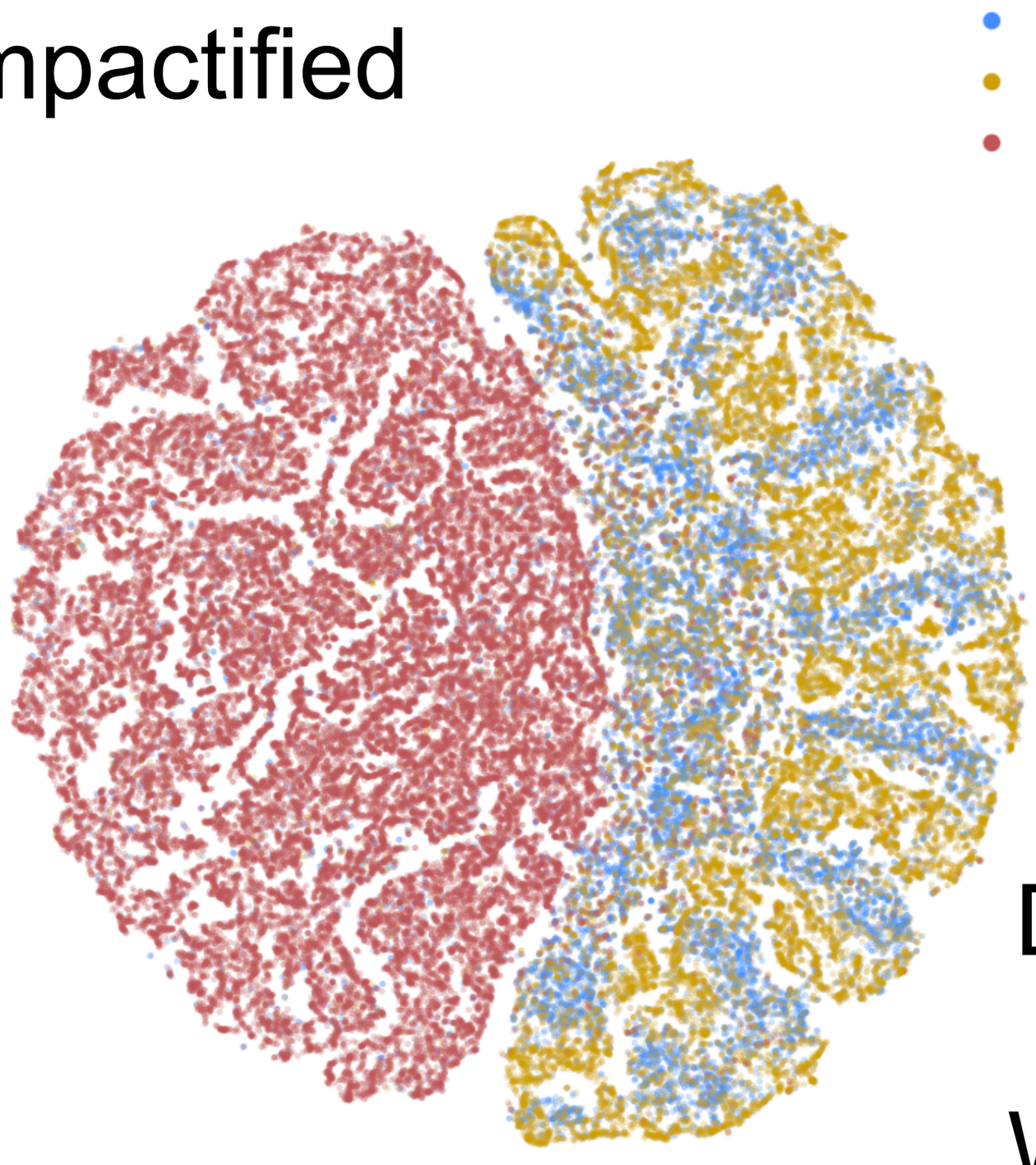
QCD

Embed in 8D

Arxiv:2403.07066

Embedded Space

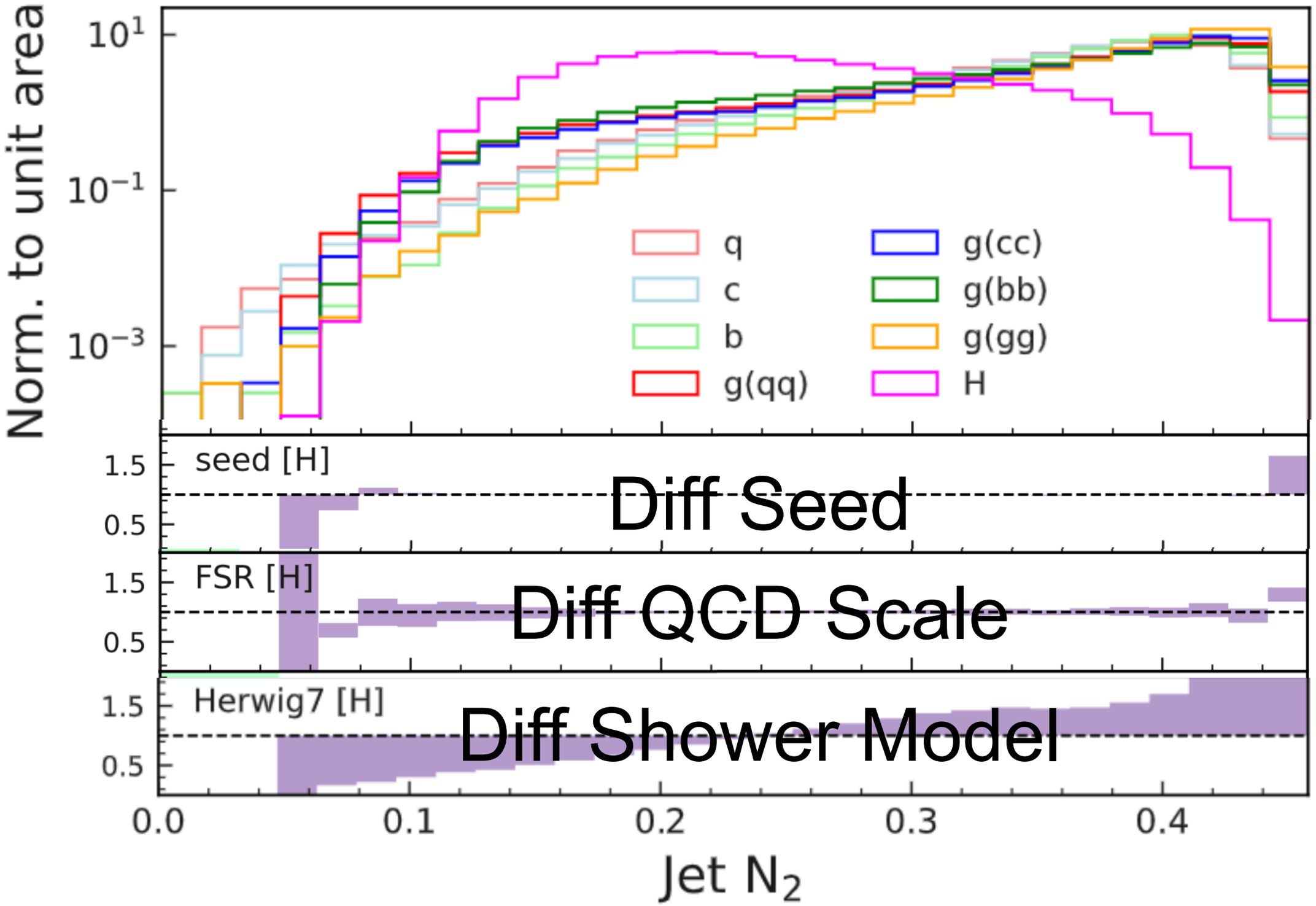
Compactified



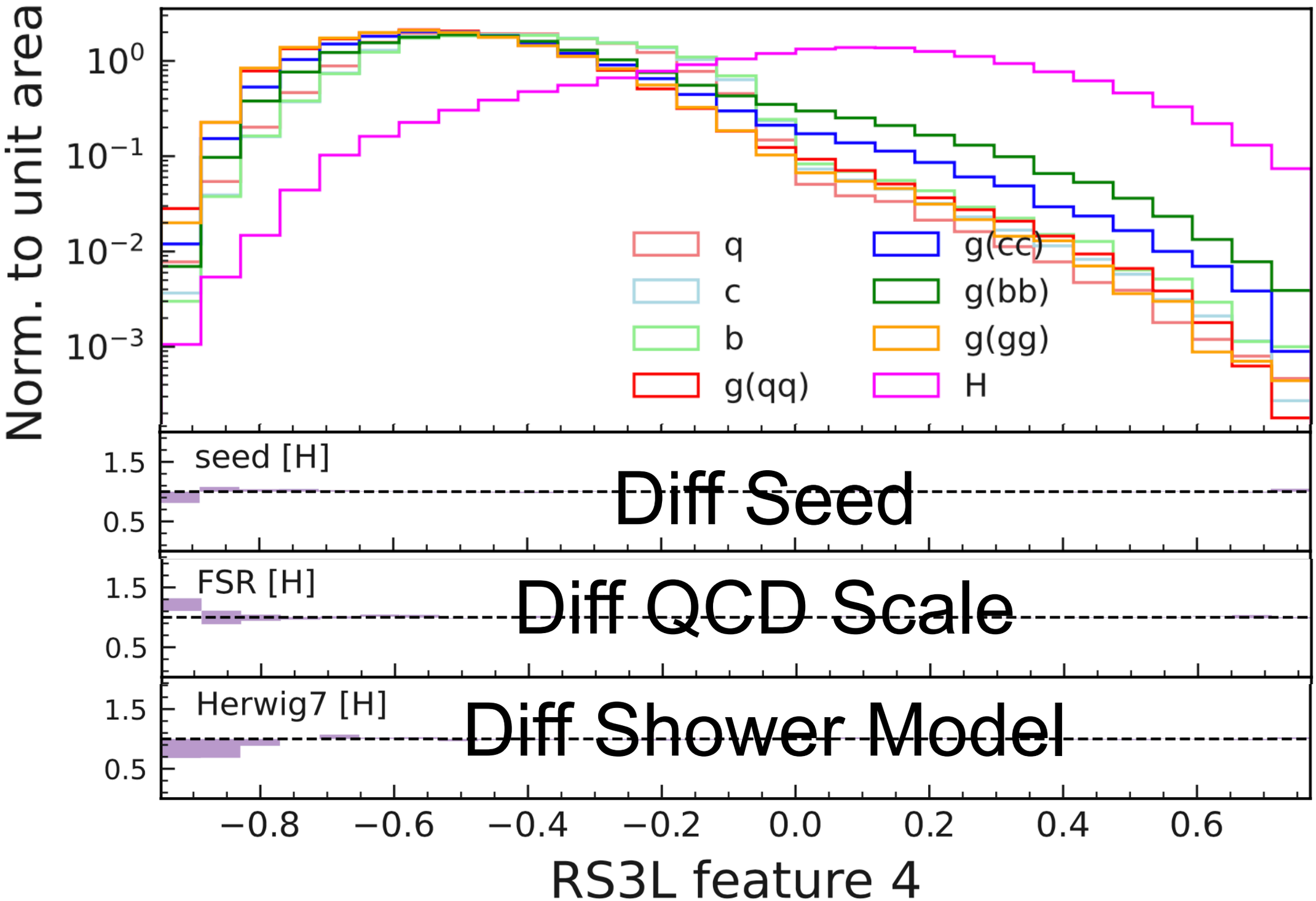
Embed in 8D

Despite Mapping
Jet to Jet
We build a space

What do we augment

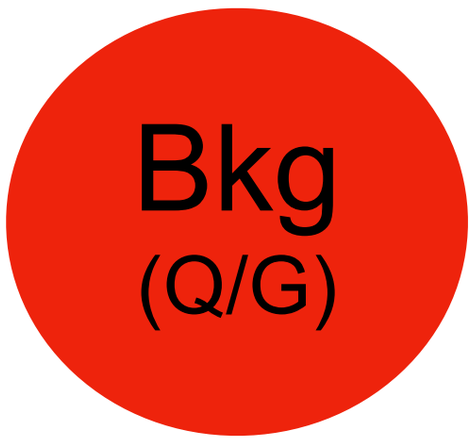


What do we augment



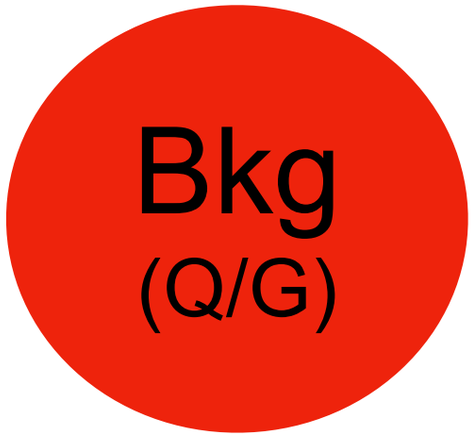
Pretraining H vs Q gives us

vs.



In Domain 25% gain off state of the art

vs.



Out of Domain 10% gain

Can train on all datasets and win everywhere!

One Big Pretraining

Higgs vs. QCD

In Domain 25% gain off state of the art

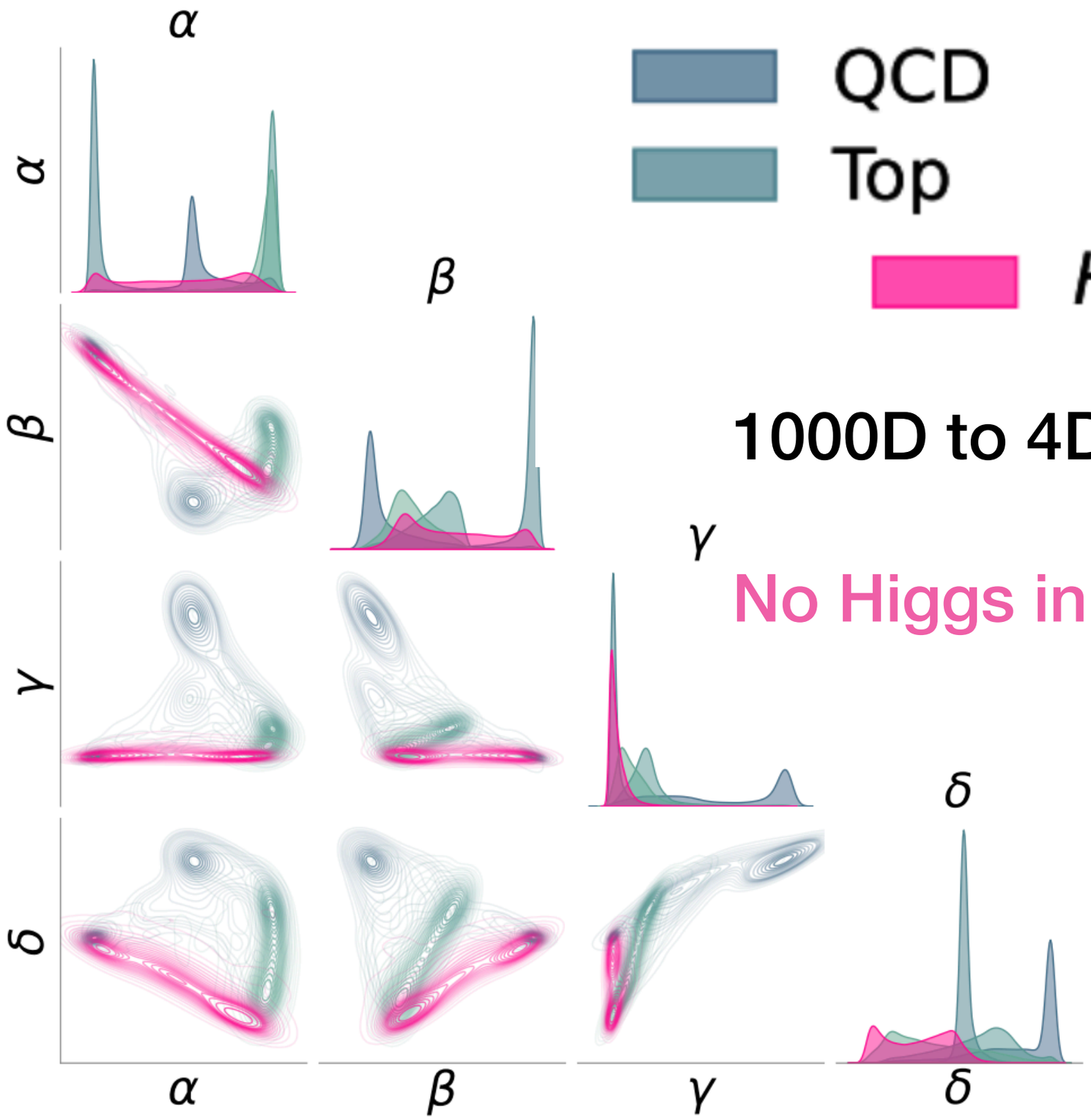
W vs. QCD

Out of Domain 10% gain

Can train on all datasets and win everywhere!

Neruips 2025

All possibilities at once



QCD
Top
$W \to q\bar{q}$
$Z \to q\bar{q}$
$H \to b\bar{b}$

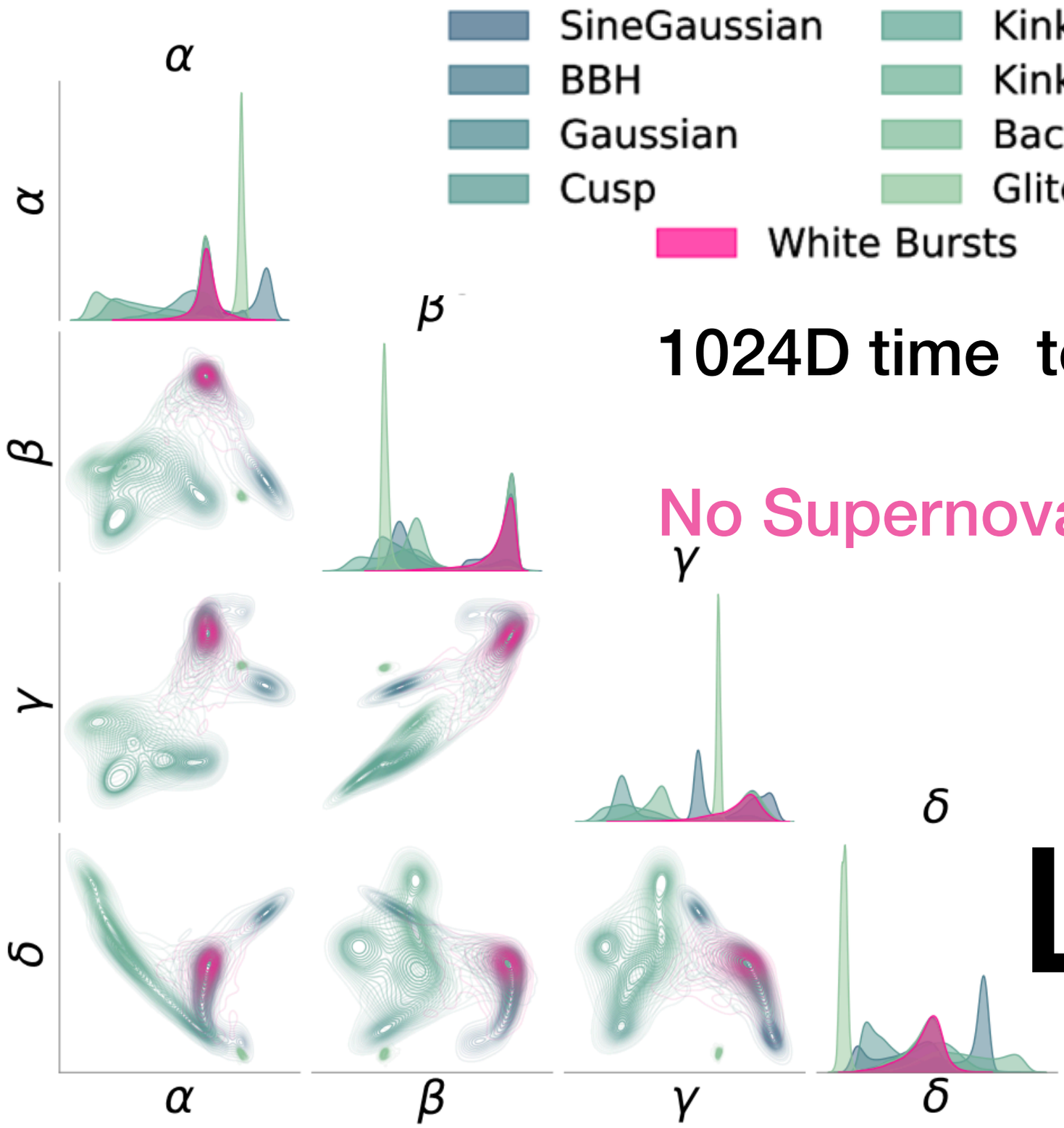
1000D to 4D as a first step

No Higgs in space, but it learns it

LHC

Neruips 2025

All possibilities at once

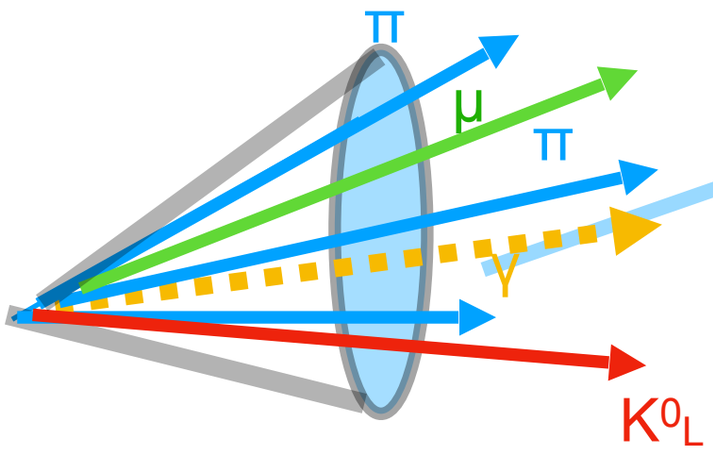


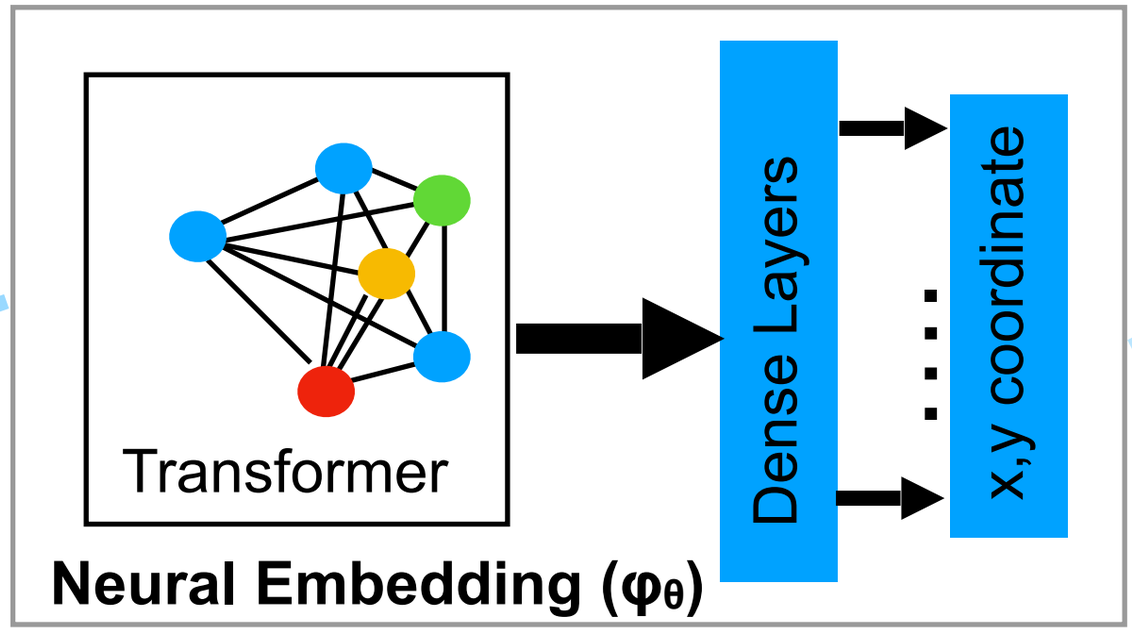
1024D time to 4D as a first step

No Supernovae, but it learns it

LIGO

Embedding





We perform the embedding with a neural network

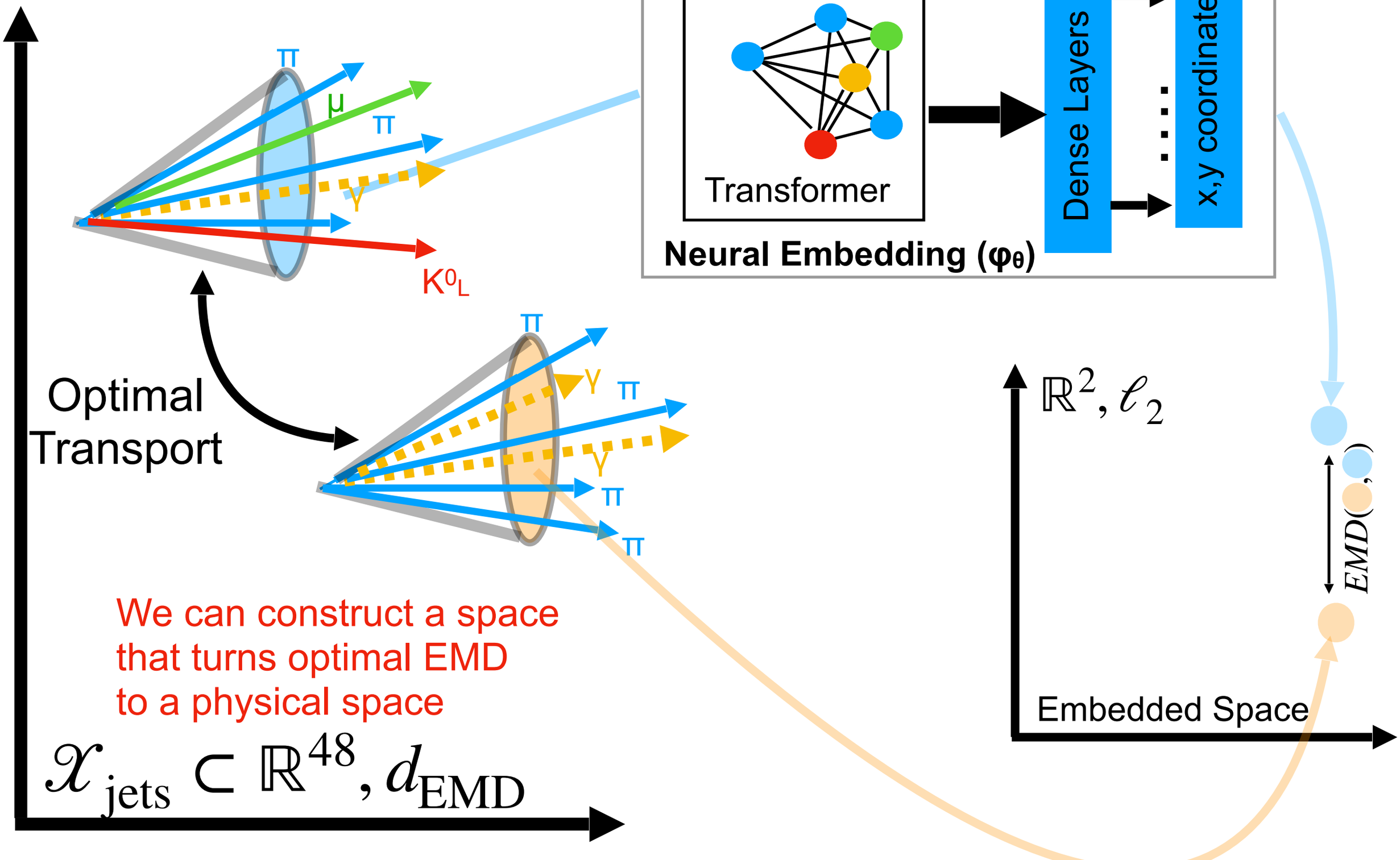
$\mathscr{X}_{\text{jets}} \subset \mathbb{R}^{48}, d_{\text{EMD}}$

$\mathbb{R}^2, \ell_2$

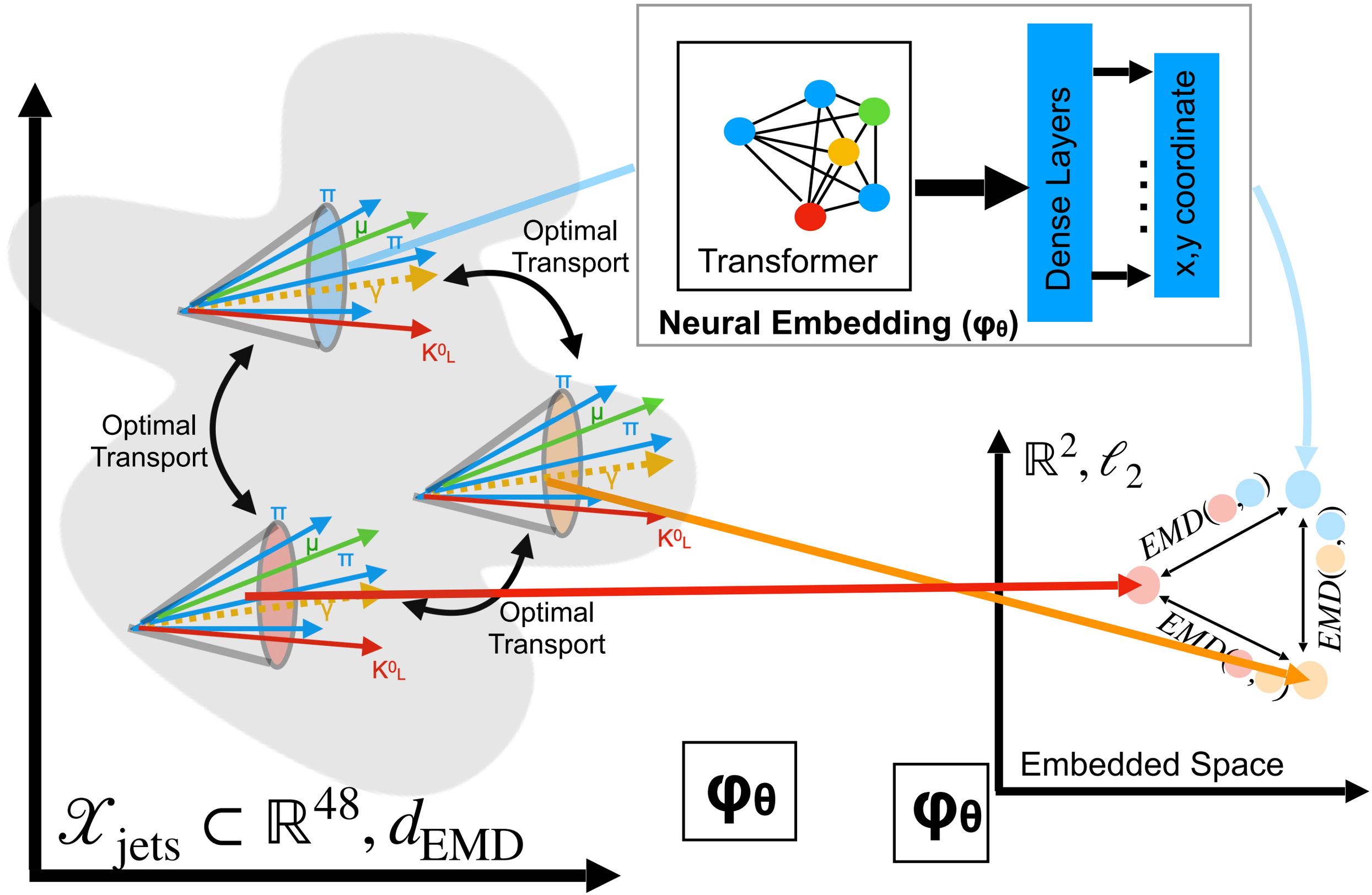
Embedded Space

Neural Embedding

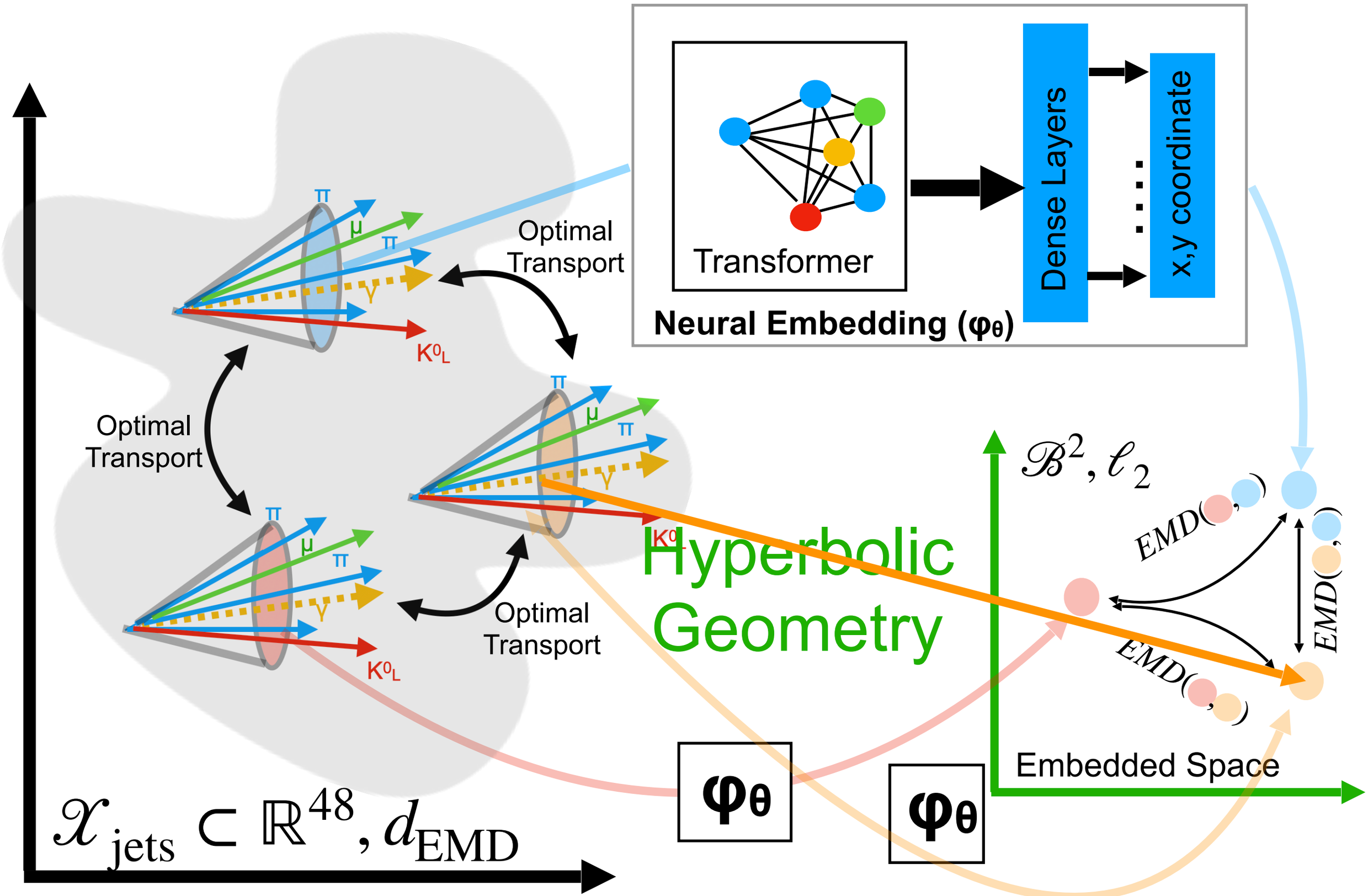
EMD = Wasserstein-1



Neural Embedding

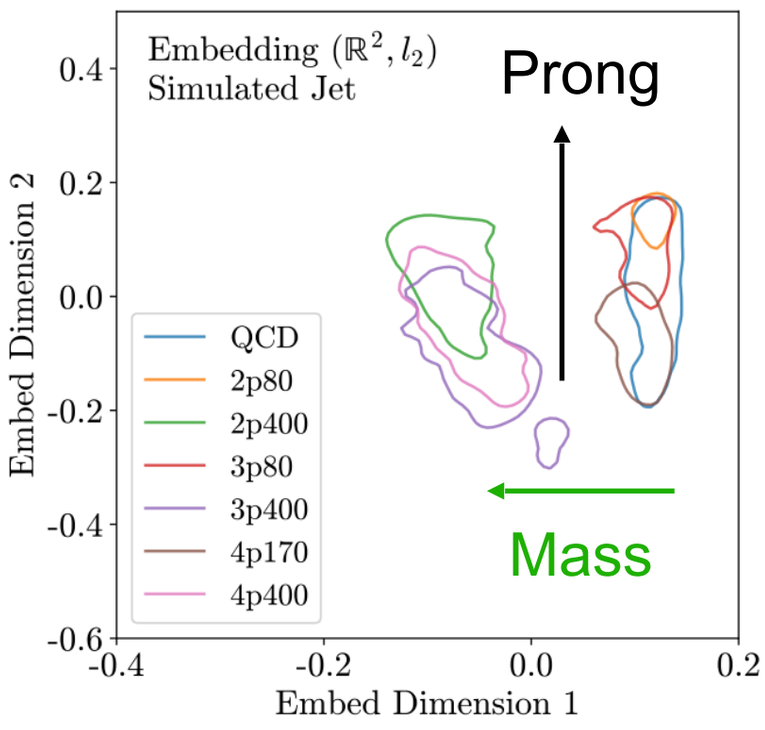


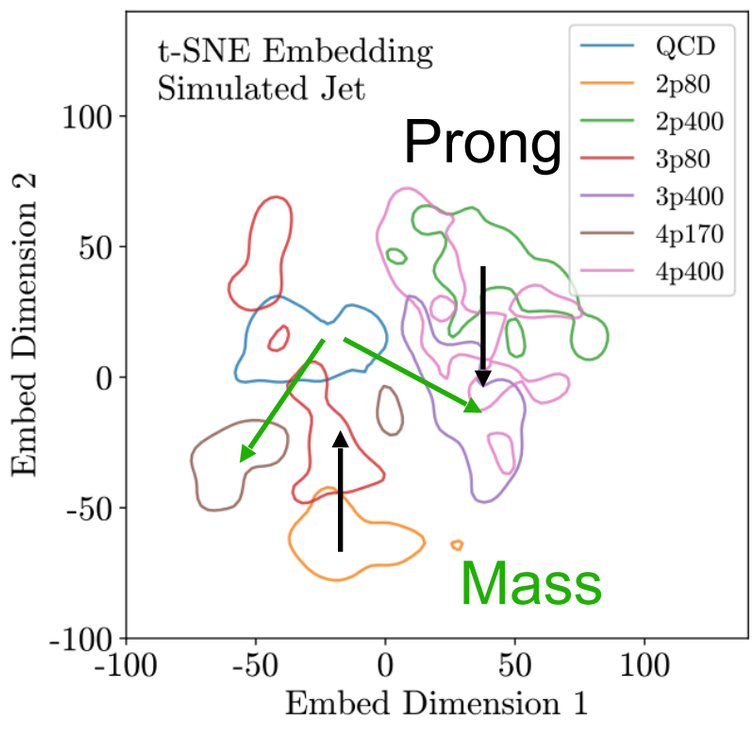
Neural Embedding

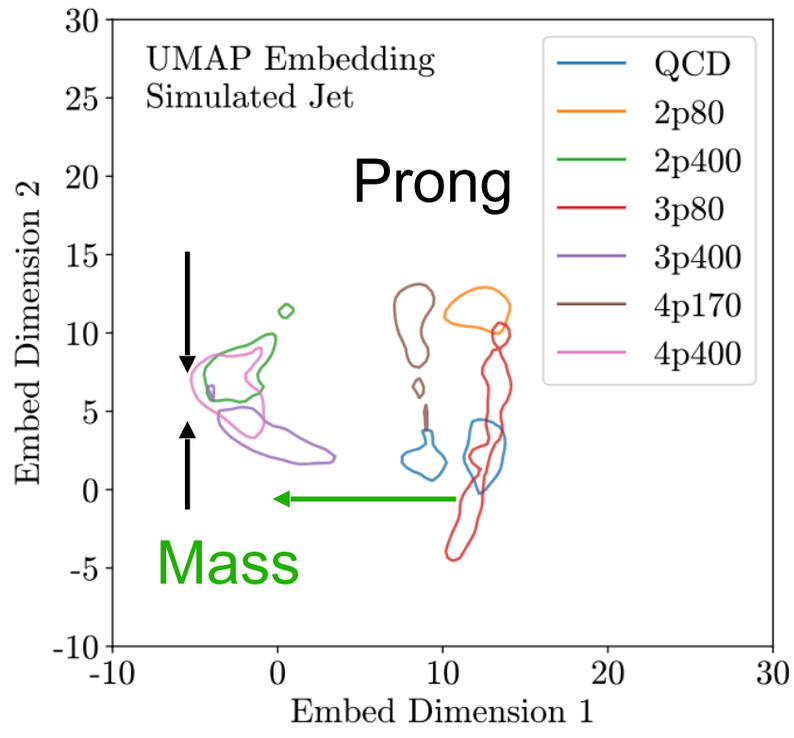


Other Approaches?

- We can consider other clustering methods

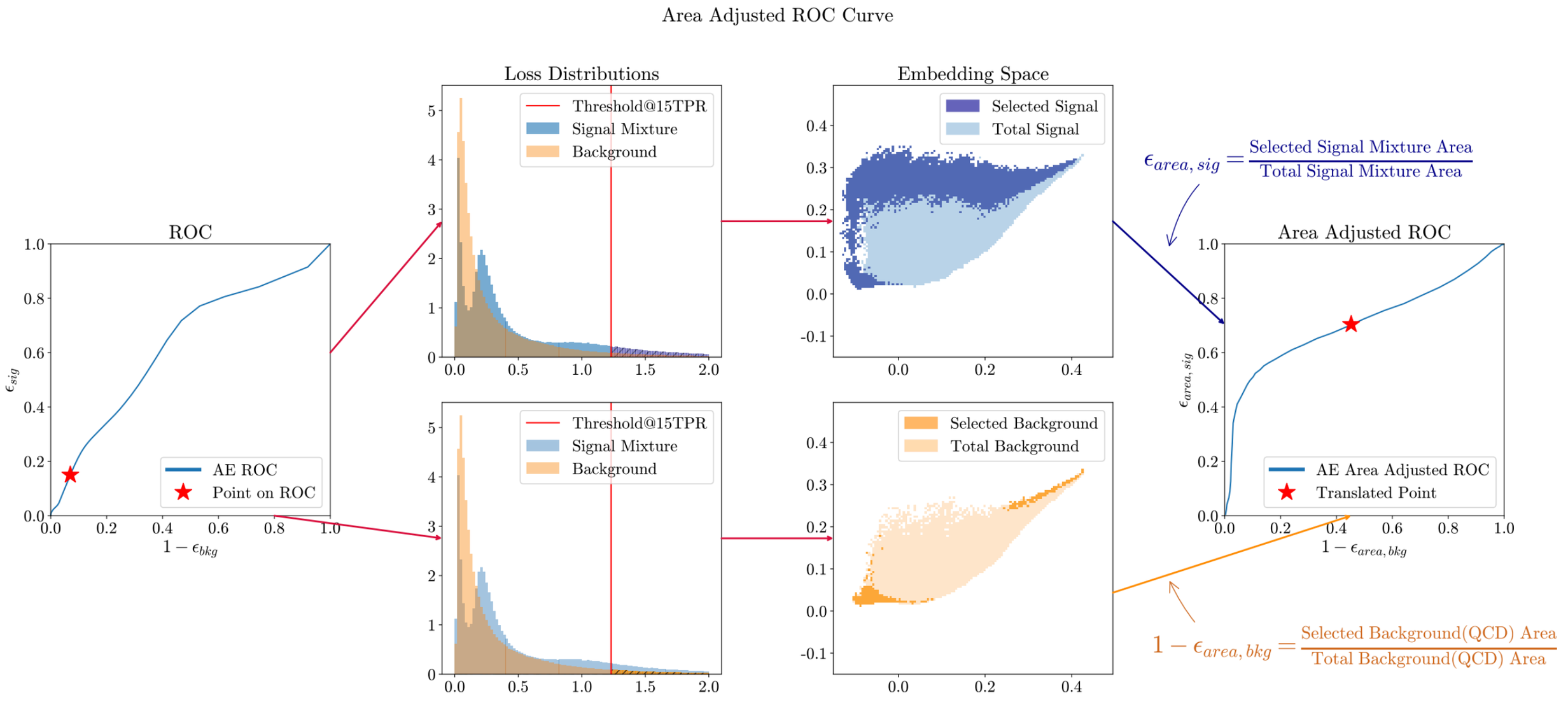




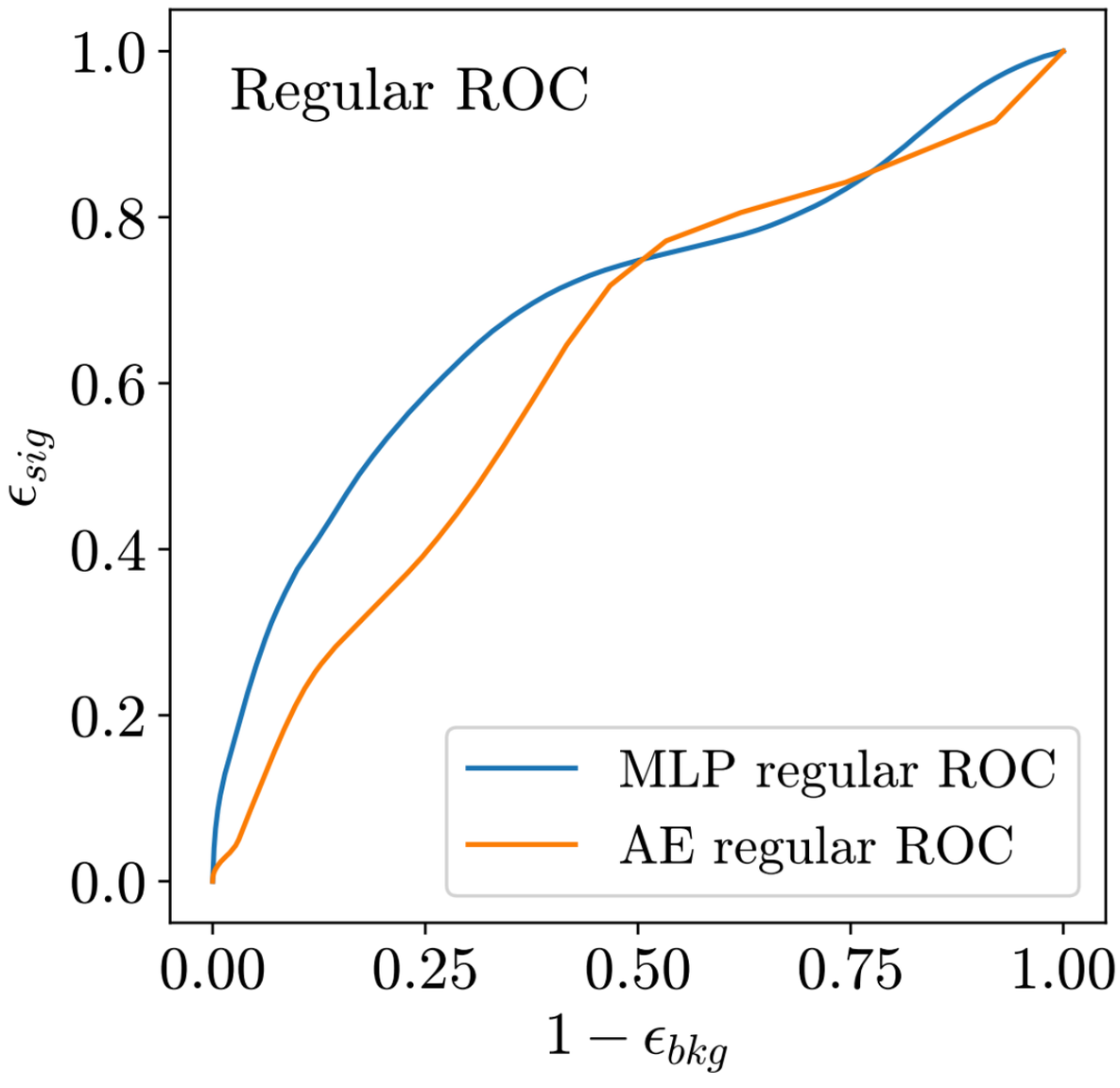


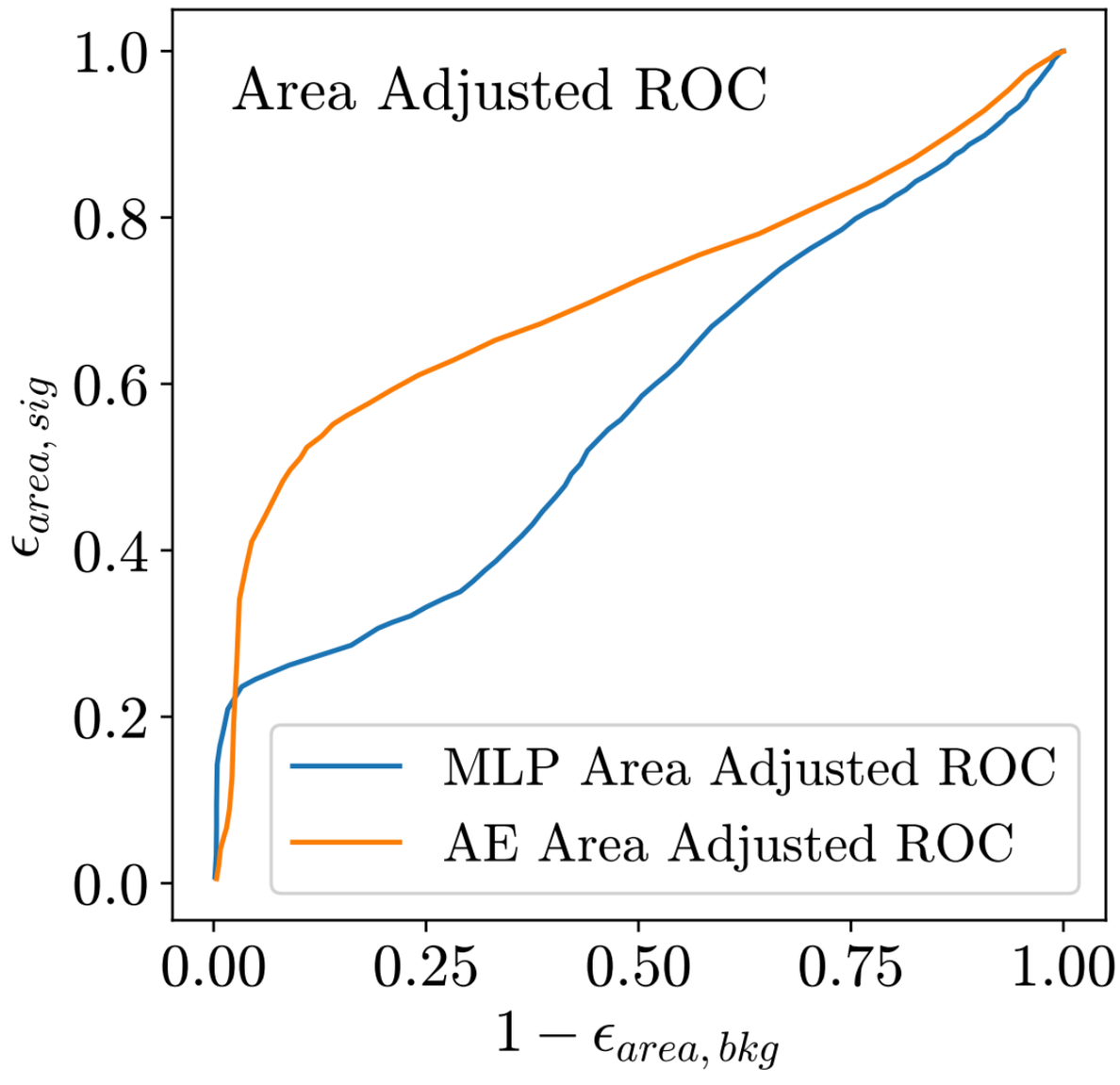
- t-SNE and UMAP both do not form a regularized space
- Embedded space leads to a physical progression

Qualifying Anomalies



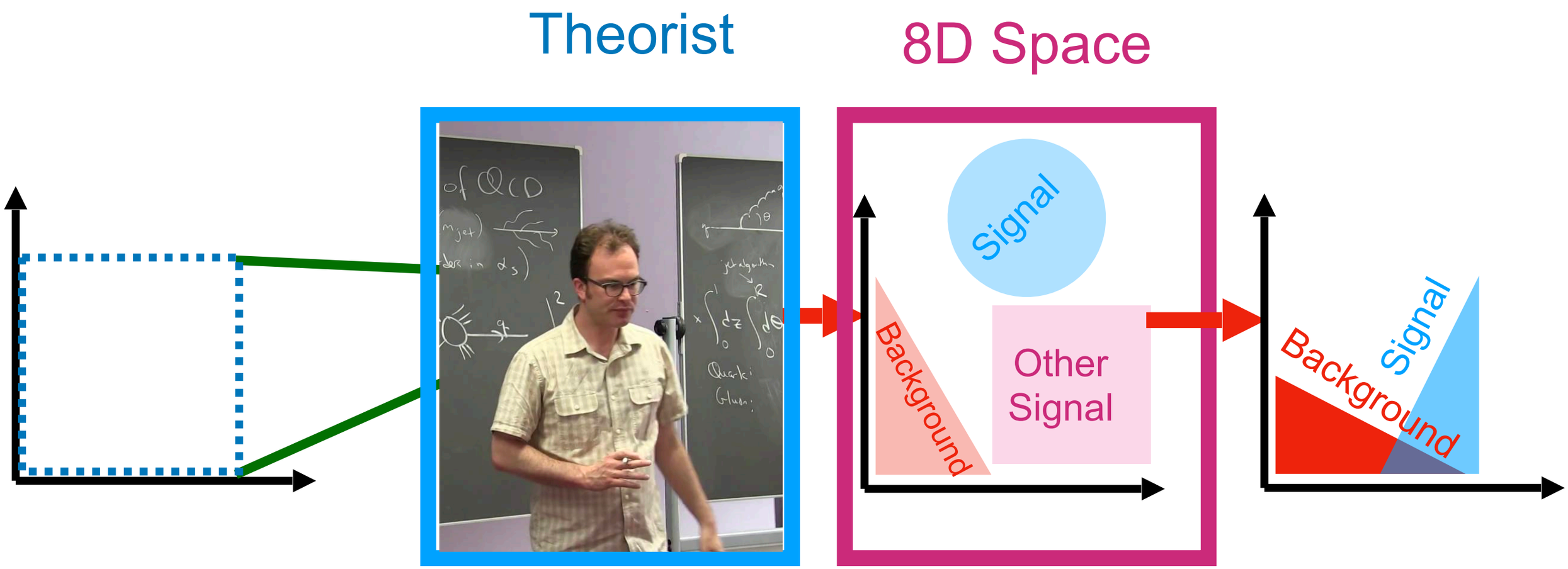
Qualifying Anomalies





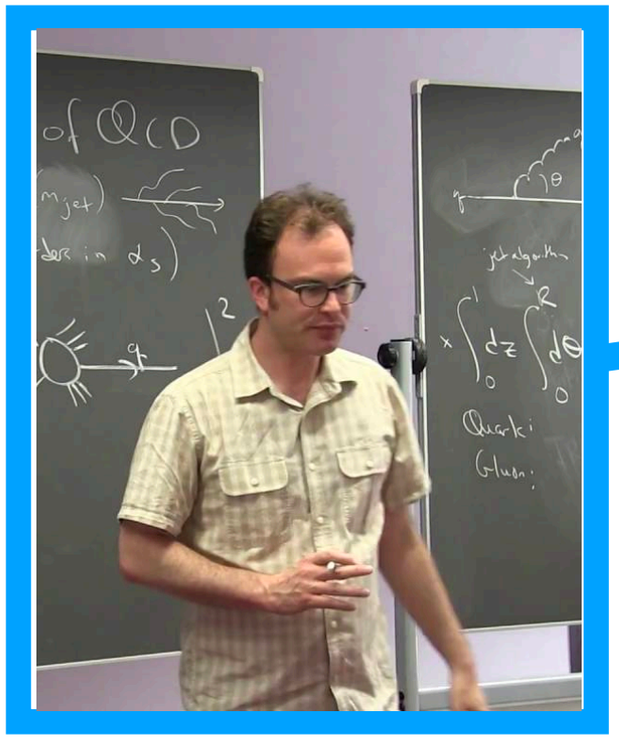
larger diversity of exploration w/Signals

Ancient Times(aka 2016)

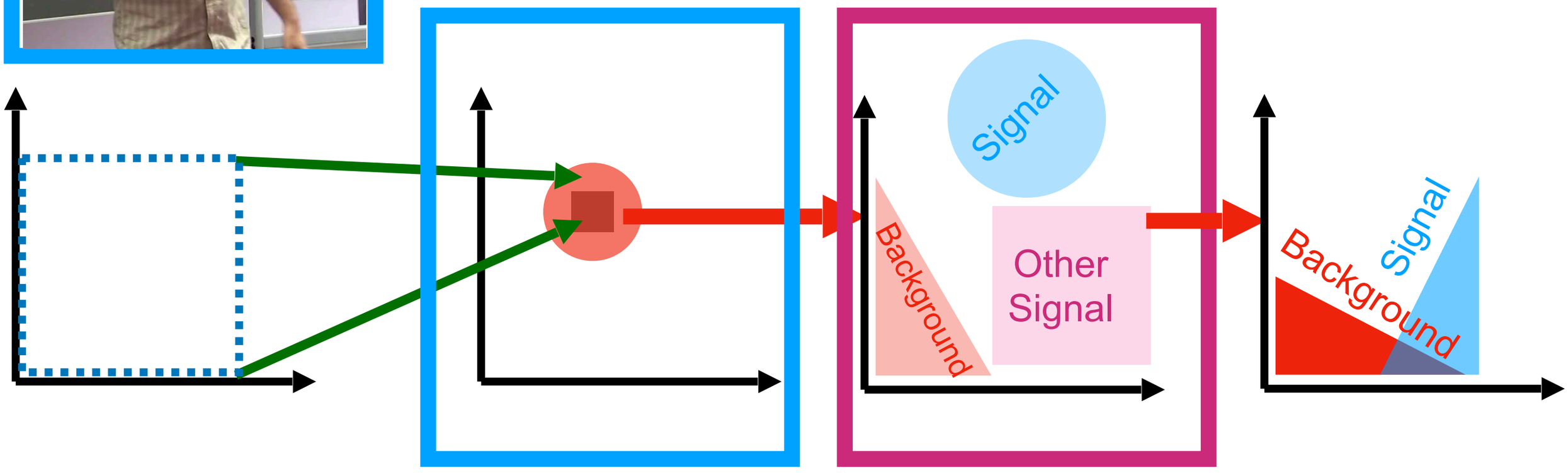


Present Times

Theorist



Gauge Invariance!



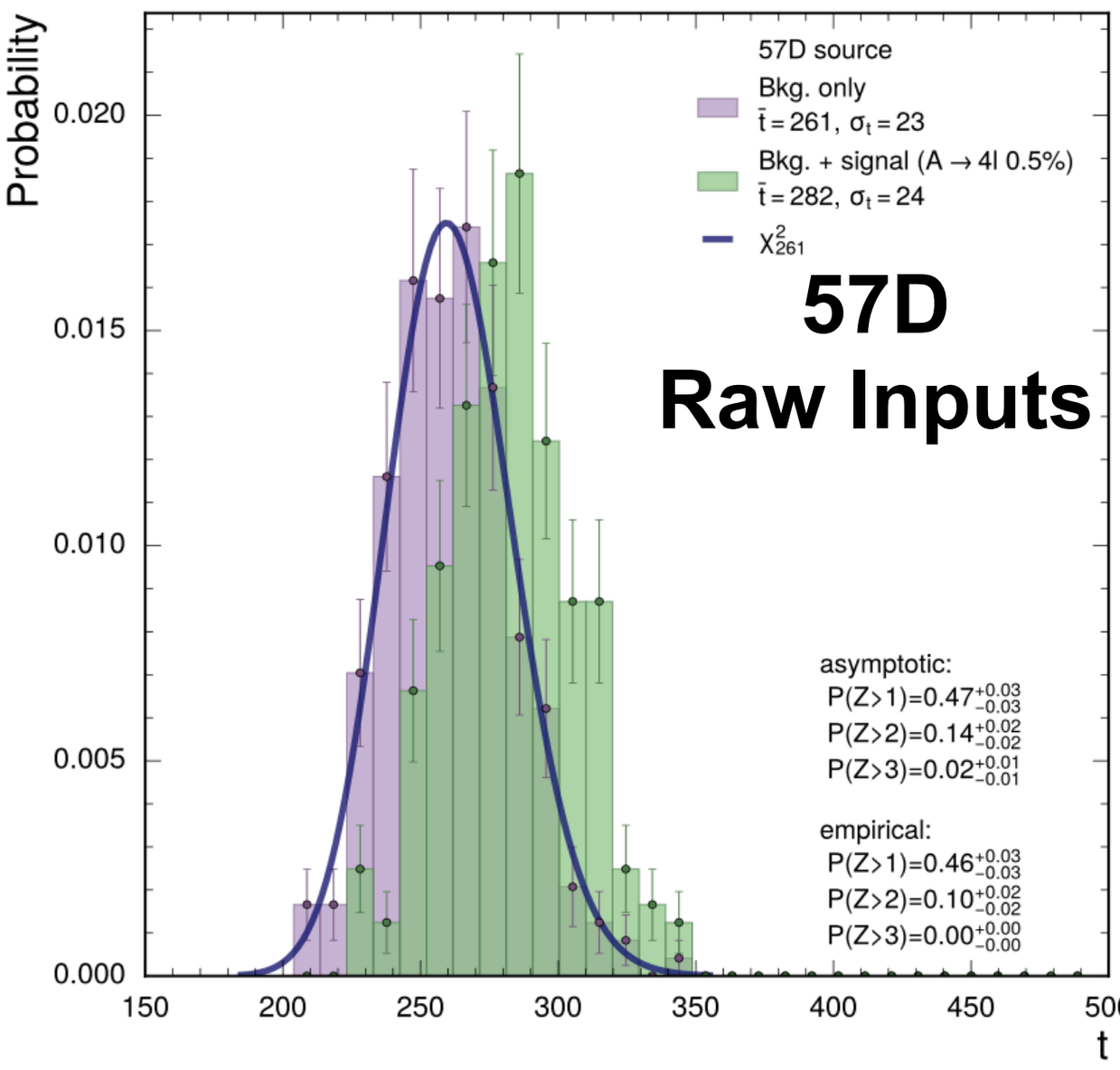
NN

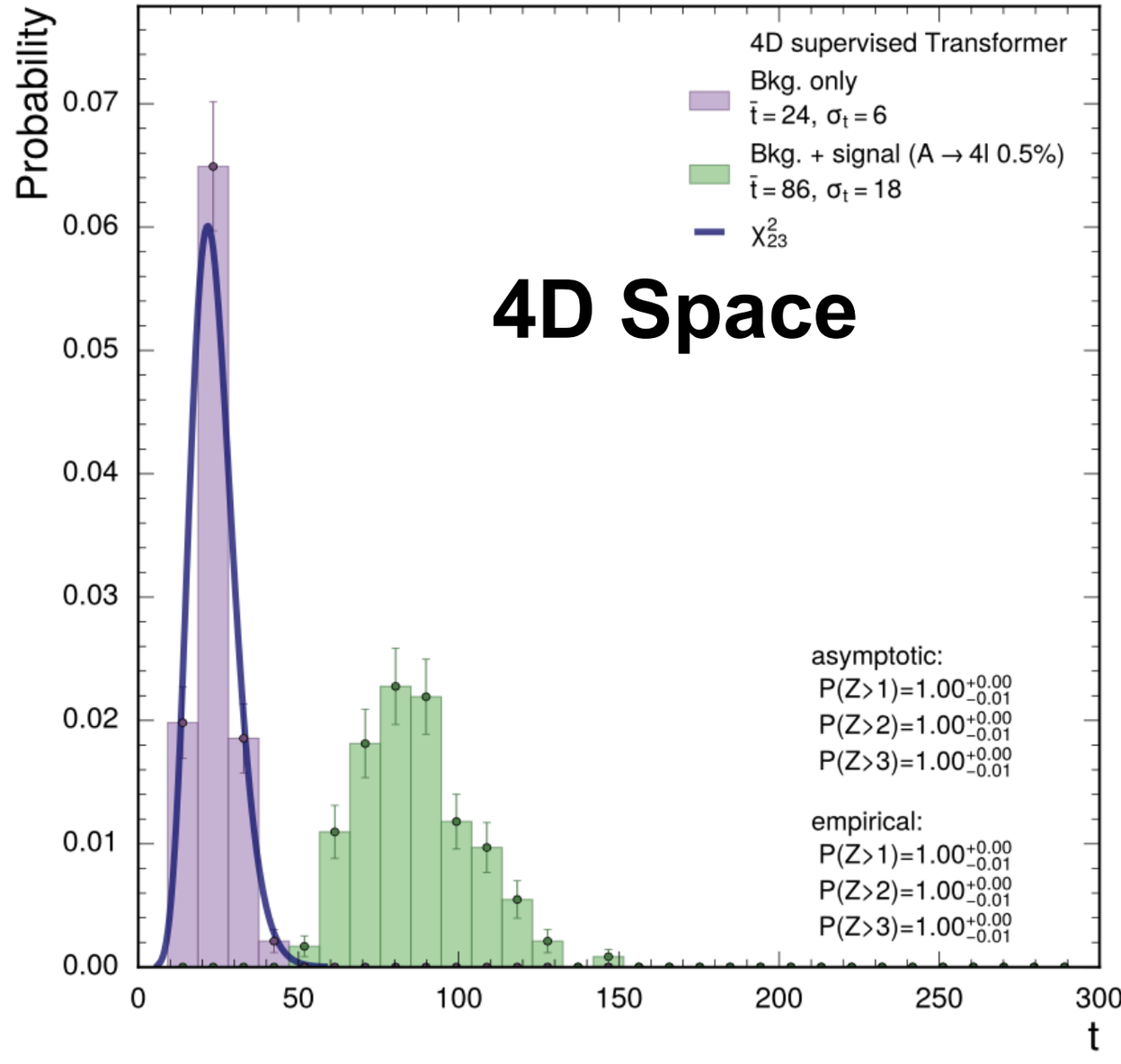
Capture
Physics

Arxiv:2502.15926

Critical Features

Can dramatically Improve our ability to find hidden signals!





The same anomaly detection algo (NPLM) on a compressed space

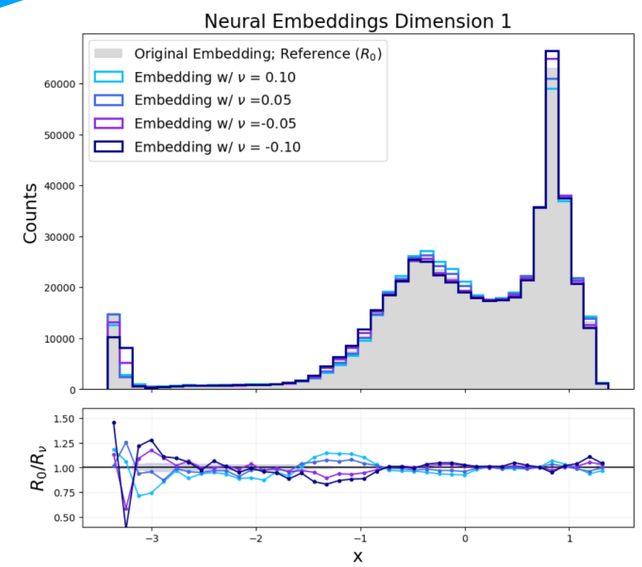
In context of Unc

Features

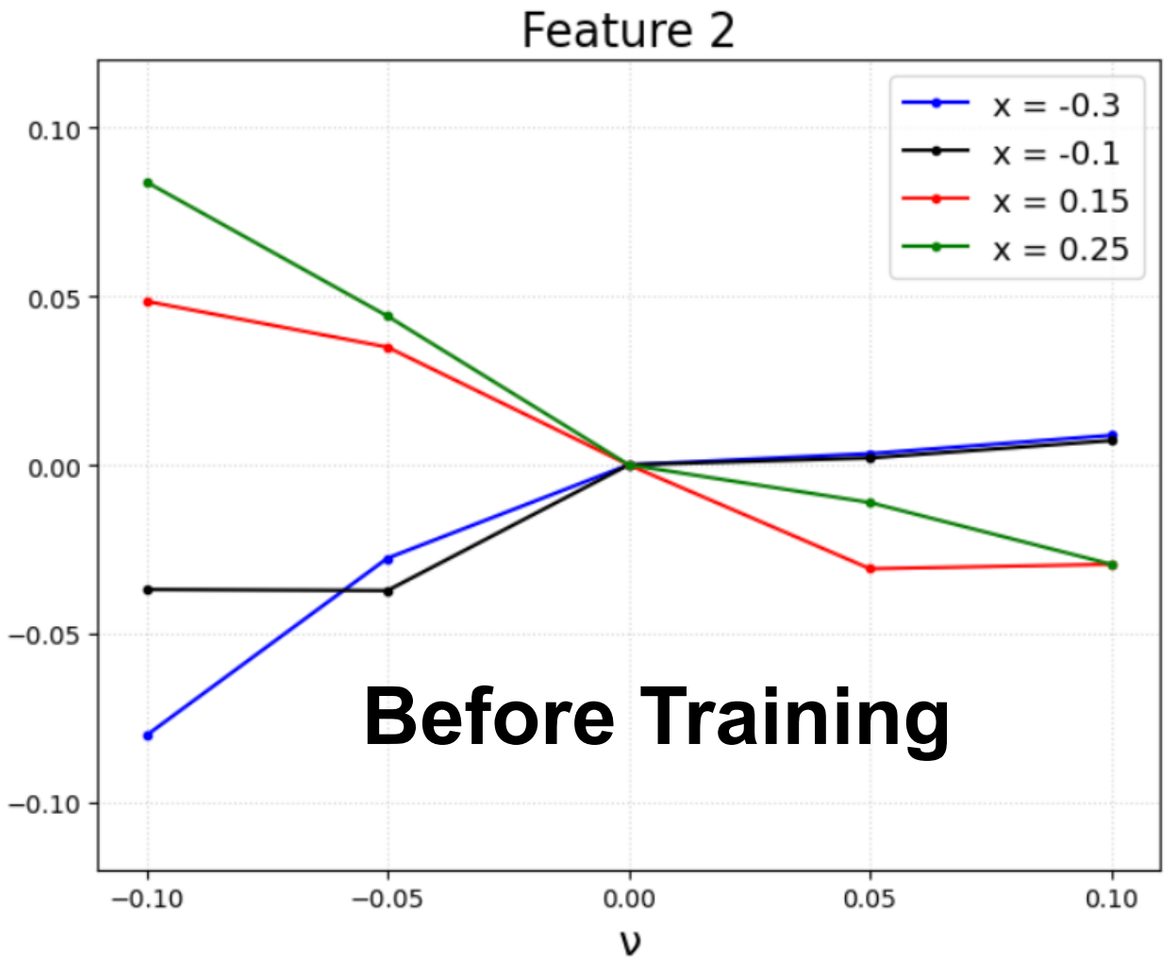
$\vec{p}_1$,
$\vec{p}_2$,
$\vdots$
$\vec{p}_{1000}$?

While Embedding Can we linearize uncertainty

Feature_1,
Feature_2,
$\cdots$
Feature_N

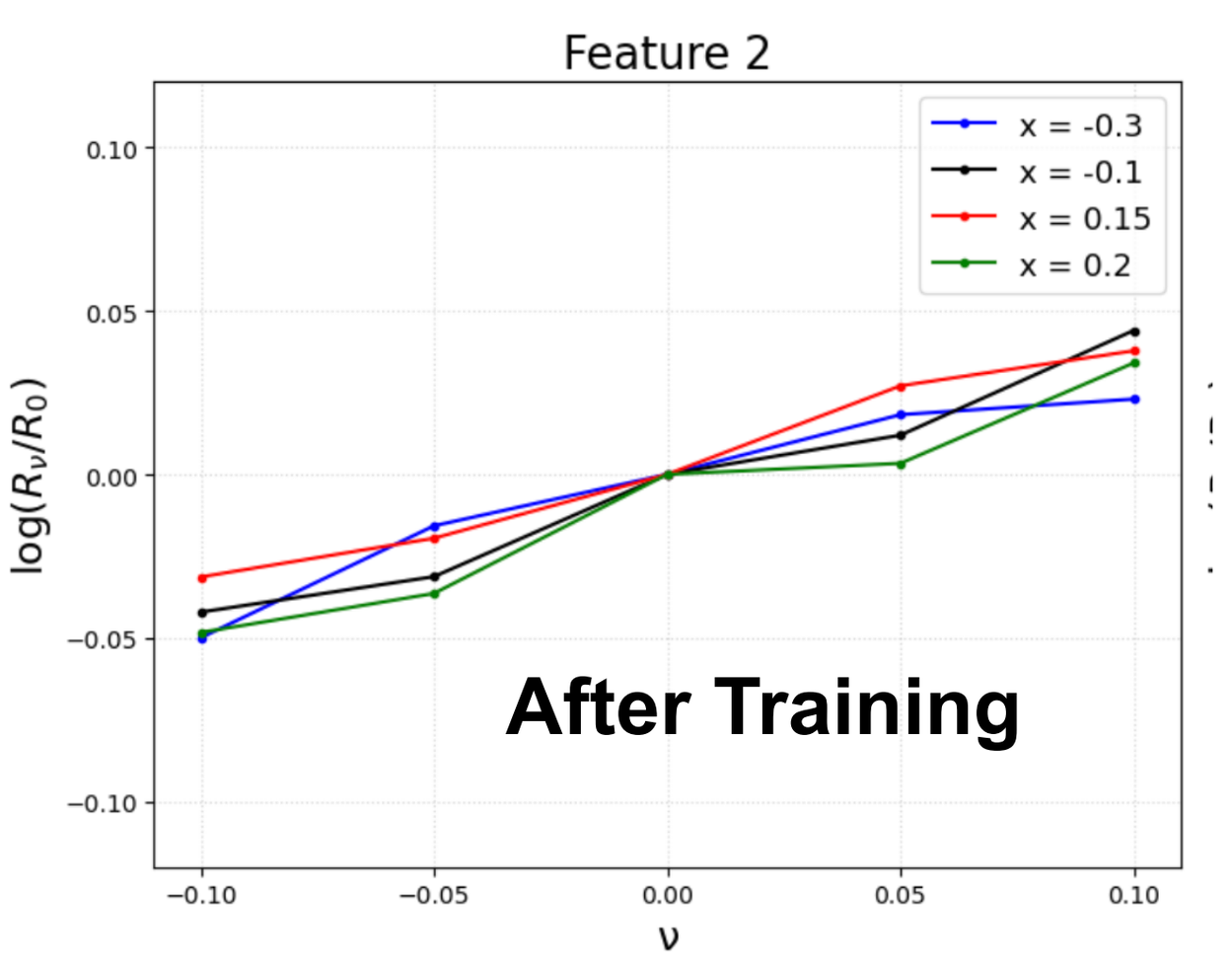


In context of Unc



Before Training

Nuisance Variation



After Training

Nuisance Variation

$$\mathbf{Loss}_{\mathbf{total}} = \mathbf{Loss}_{\mathbf{CL}} + \lambda \times \mathbf{Reg}_{\mathbf{Poly}}$$

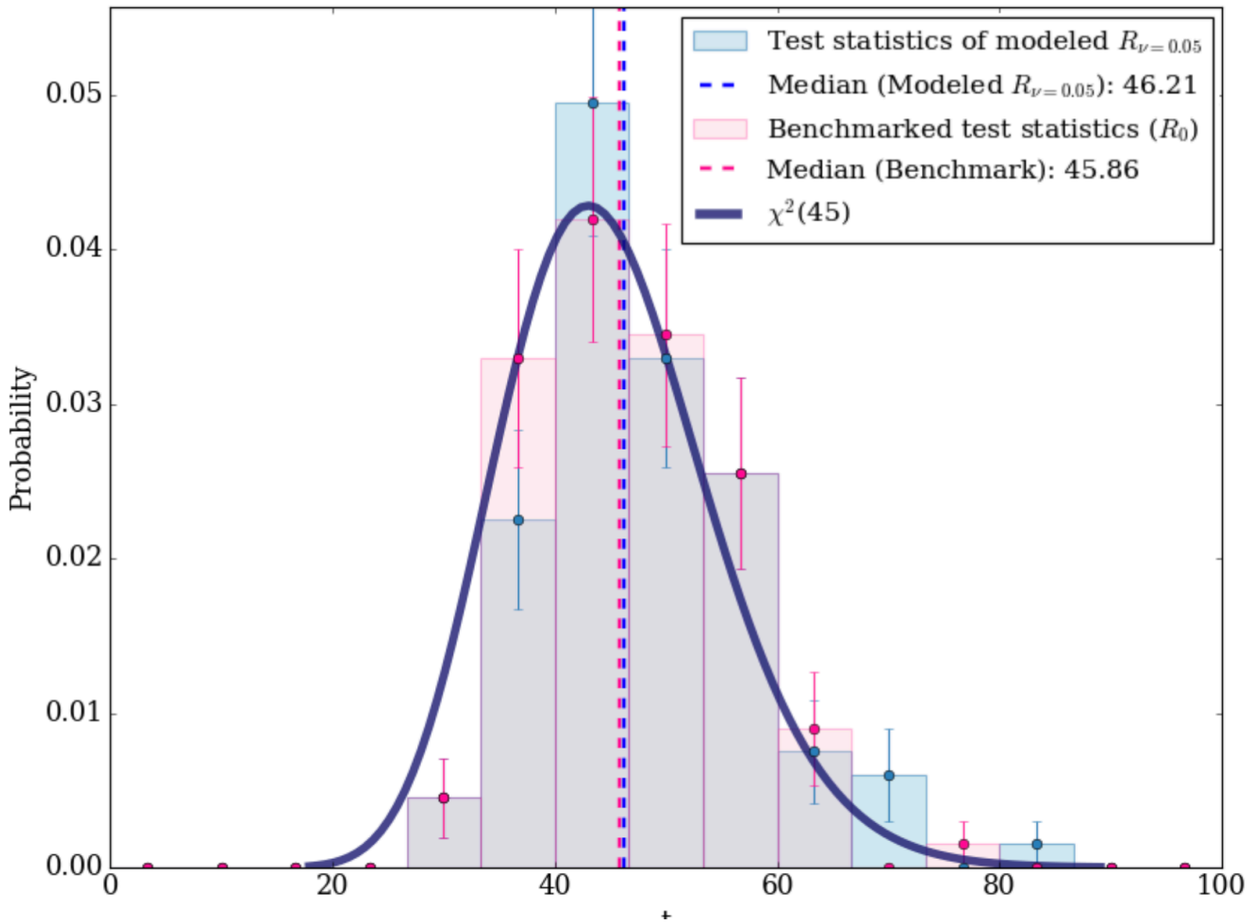
Loss from supervised contrastive learning

Factor we fine-tune

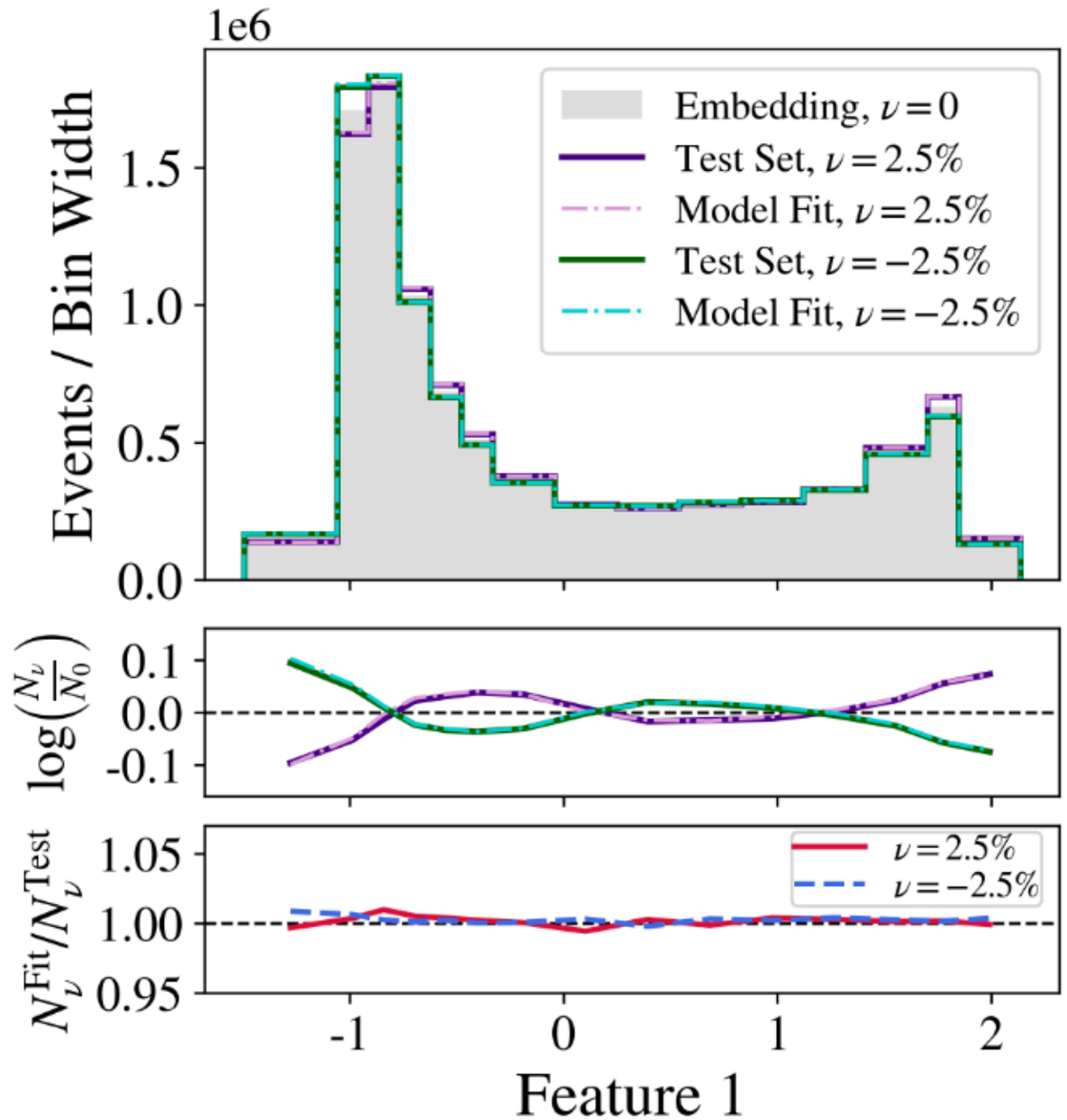
Regularization that induces polynomial behavior of the density ratio

In context of Unc

Linearized Networks Can be profiled in SBI

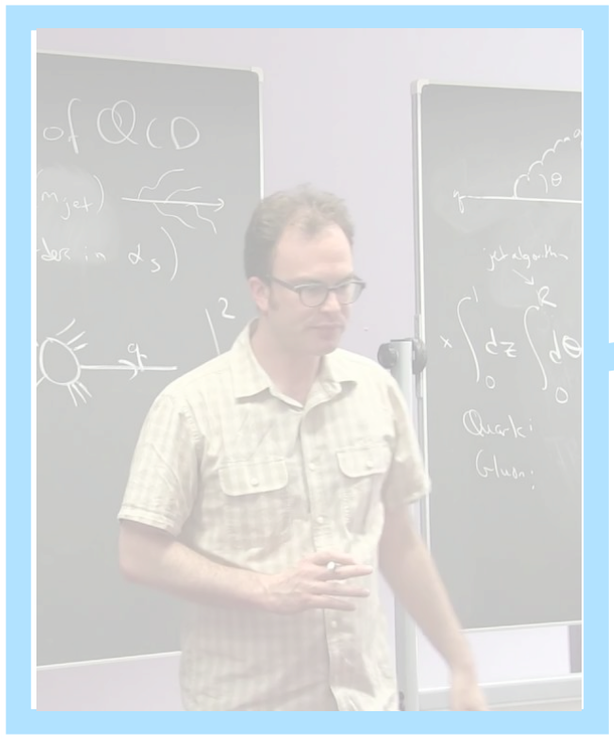


Linearized space allows us to extrpaolate nuisance



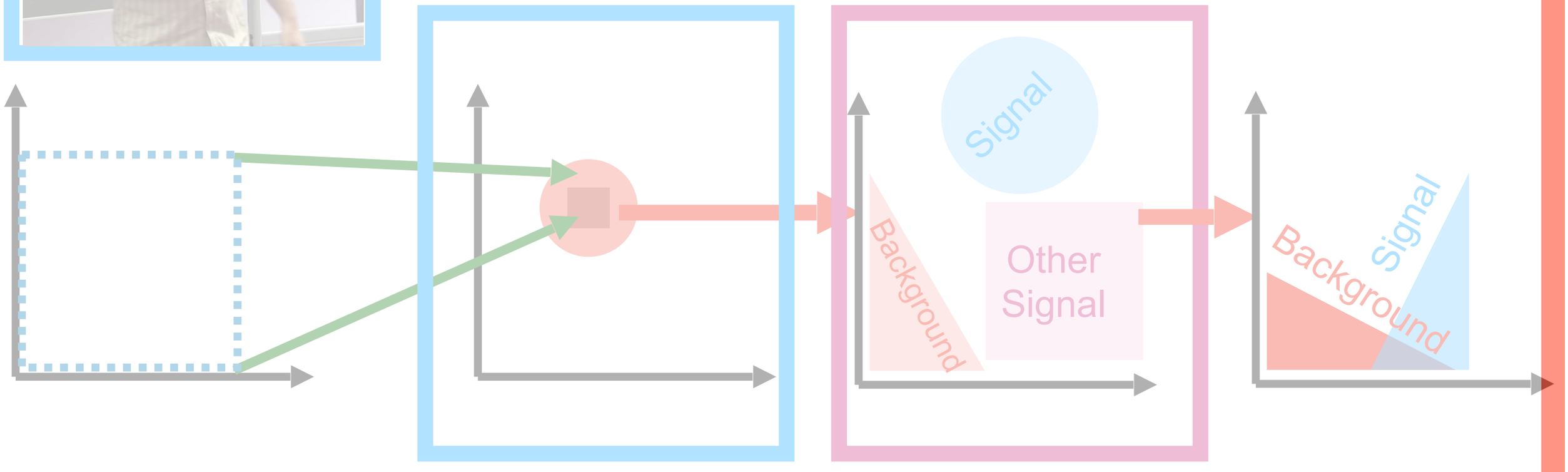
Theorist

Present Times



Gauge Invariance!

Space



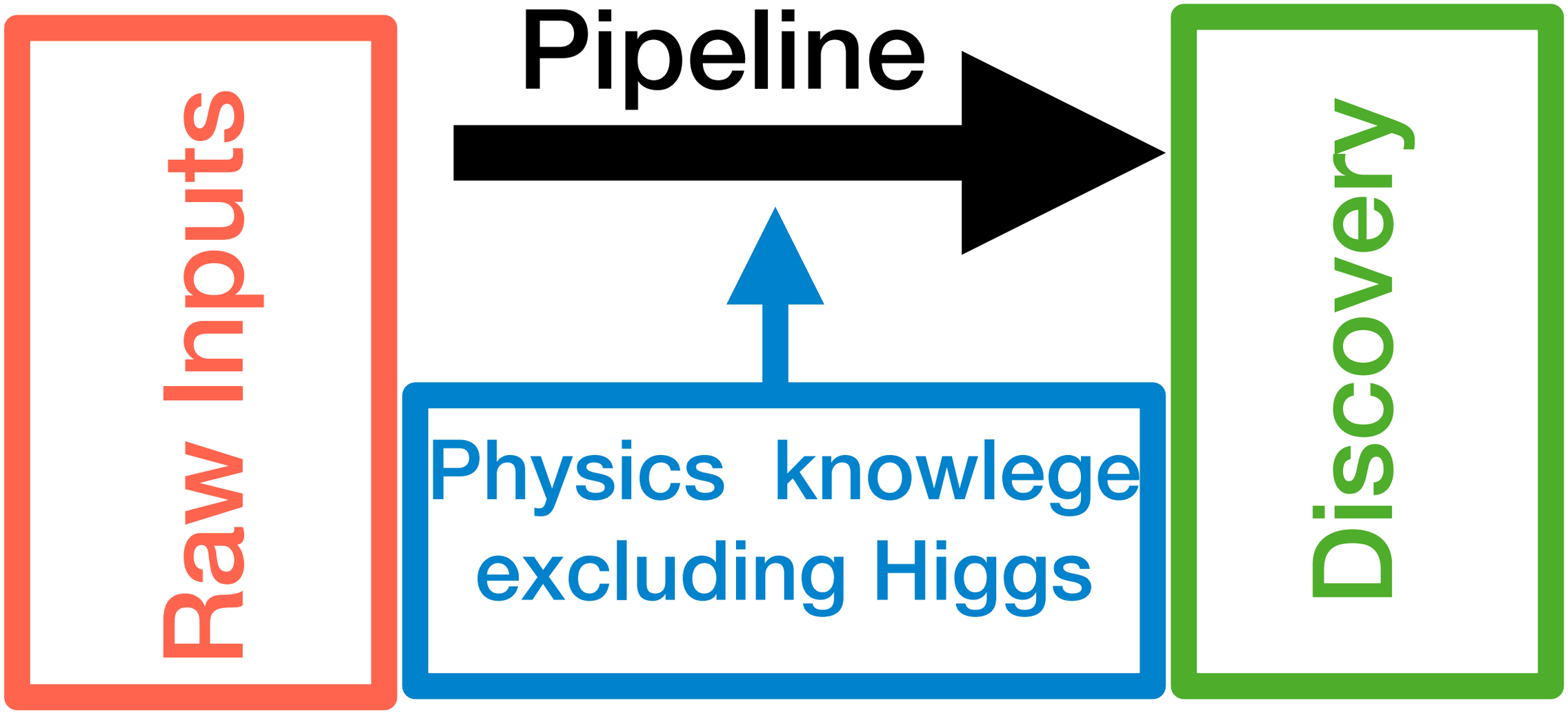
NN

Capture Physics

One Automated Step

rxiv:2510.2193
Neurips 2025

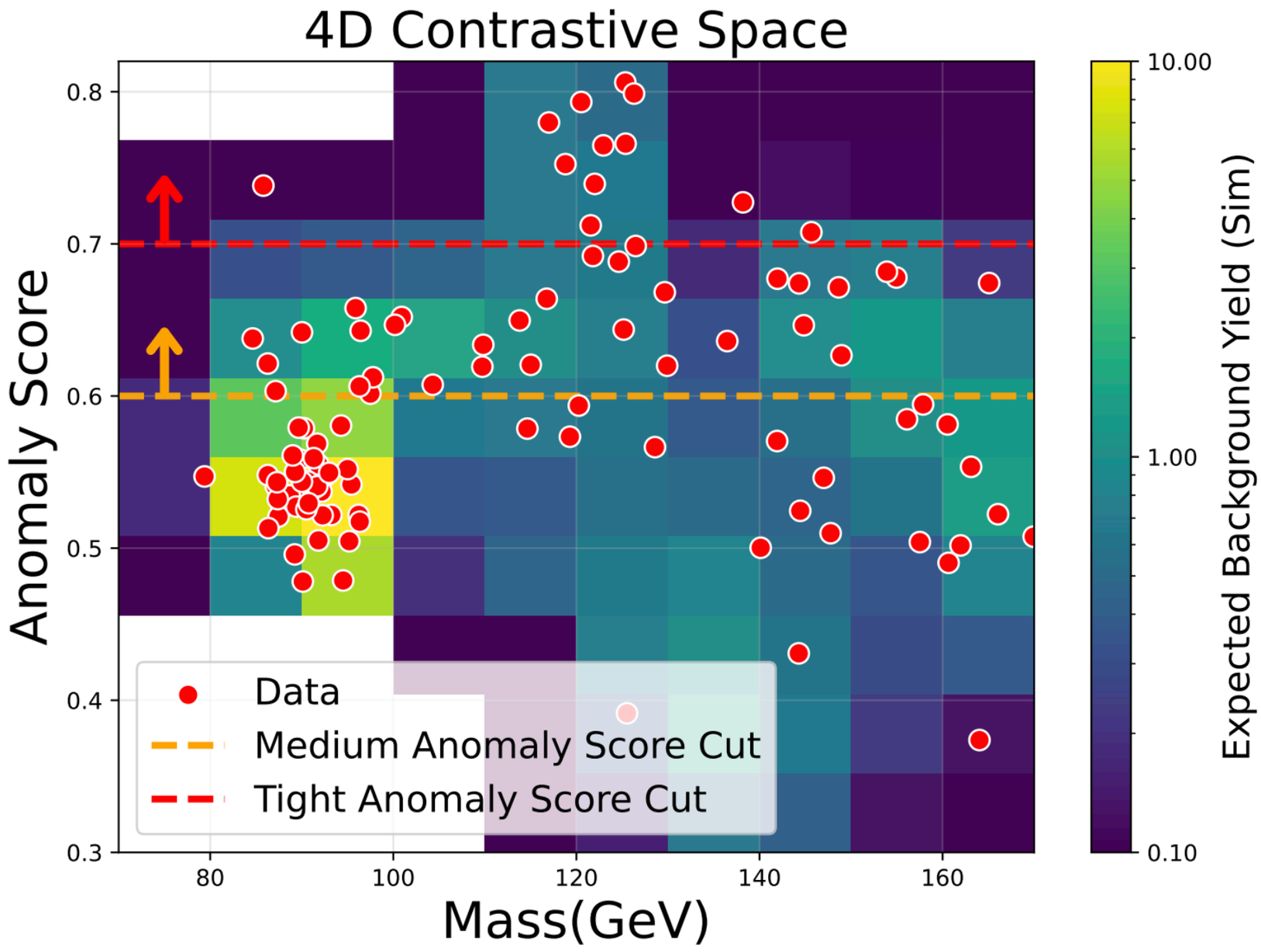
Discovering Higgs



We never told this algorithm about the Higgs

rxiv:2510.2193
Neurips 2025

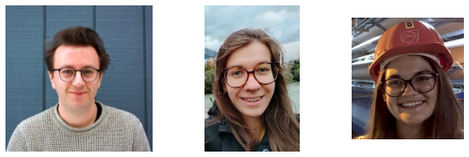
Discovering Higgs



We never told this algorithm about the Higgs

It learned Mass and Found It

Arxiv:2510.21935
Neruips 2025



Observe
Take data & understand

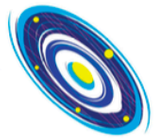
Inquire
Ask a question

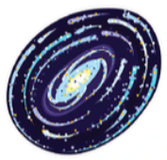
Research
Search for answers

Hypothesize
Formulate a hypothesis

Simulation or reference

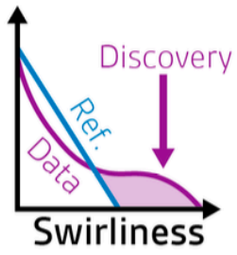
Data



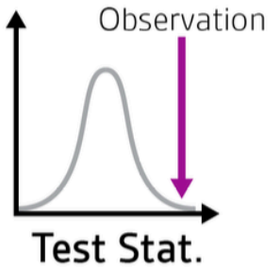


Swirly galaxies look different?

Test
Experiment

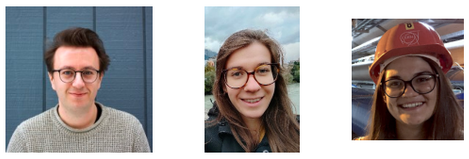


Conclude
Results



The Scientific Method

Arxiv:2510.21935
Neruips 2025



All possibilities at once

Observe
Take data & understand

Inquire
Ask a question

Research
Search for answers

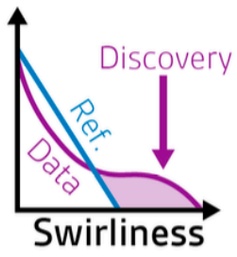
Simulation or reference

Data

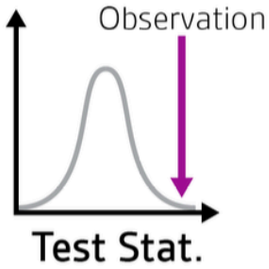
Hypothesize
Formulate a hypothesis

Swirly galaxies look different?

Test
Experiment

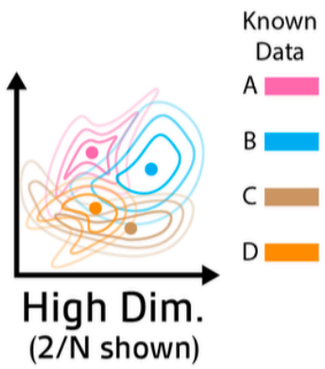


Conclude
Results



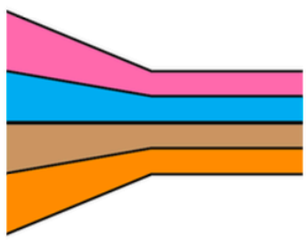
Expert / LLM Input
What to look at?
What reference?

Reduce
Simplify with symmetry



Pre-trained model / foundation model

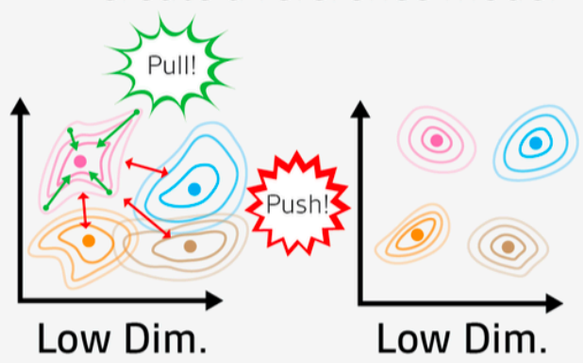
Isolate
Select a subset



Data selection

AutoSciDACT

Learn a Space
Organize known signals to create a reference model



Supervised contrastive learning

Resulting ref. model

Anomaly Detection & Test
Compare incoming data with reference & access significance with a stat. test

Search for overdensity

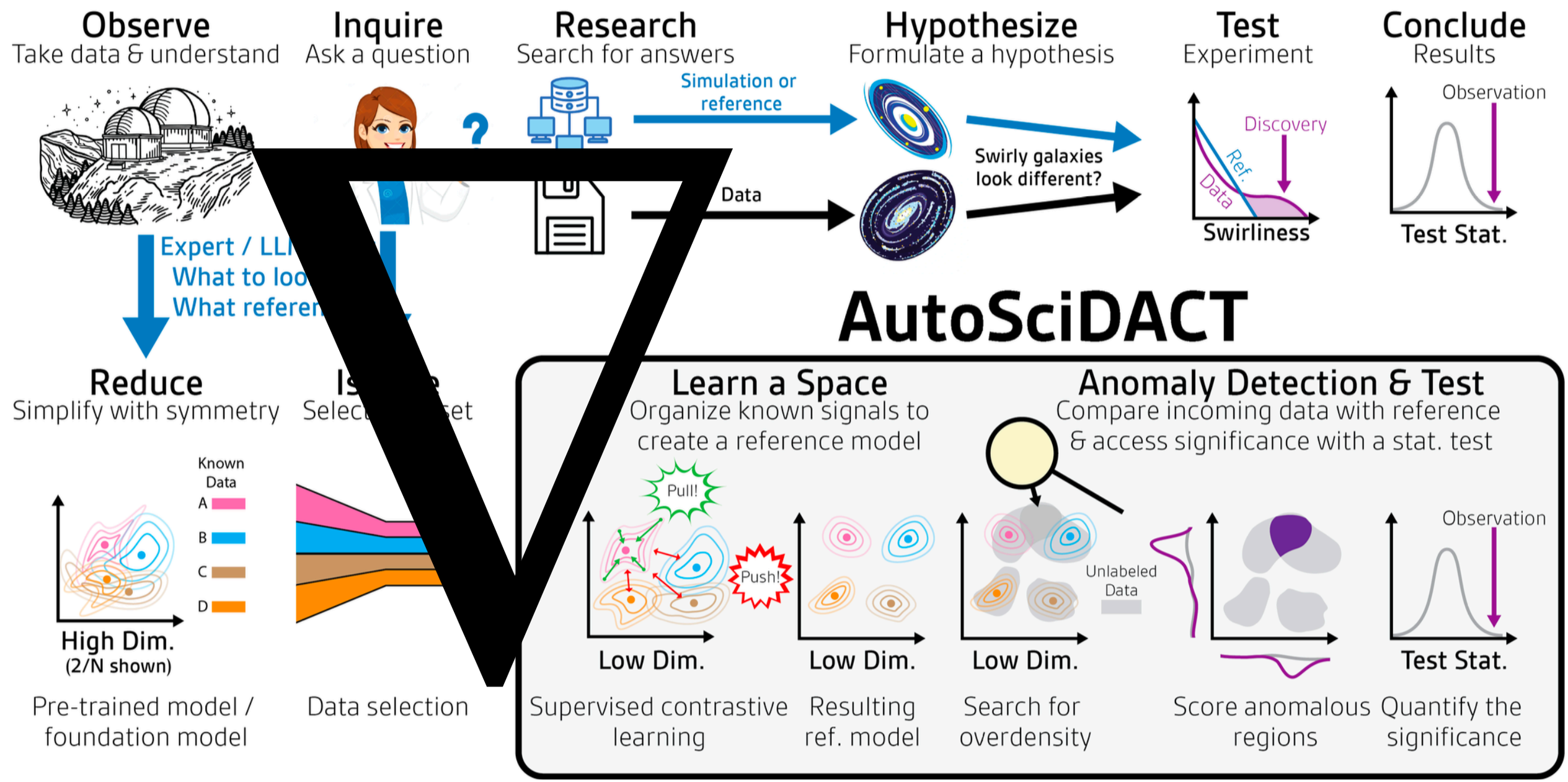
Score anomalous regions

Quantify the significance

An automatic AI toolkit for discovery

Arxiv:2510.21935
Neruips 2025

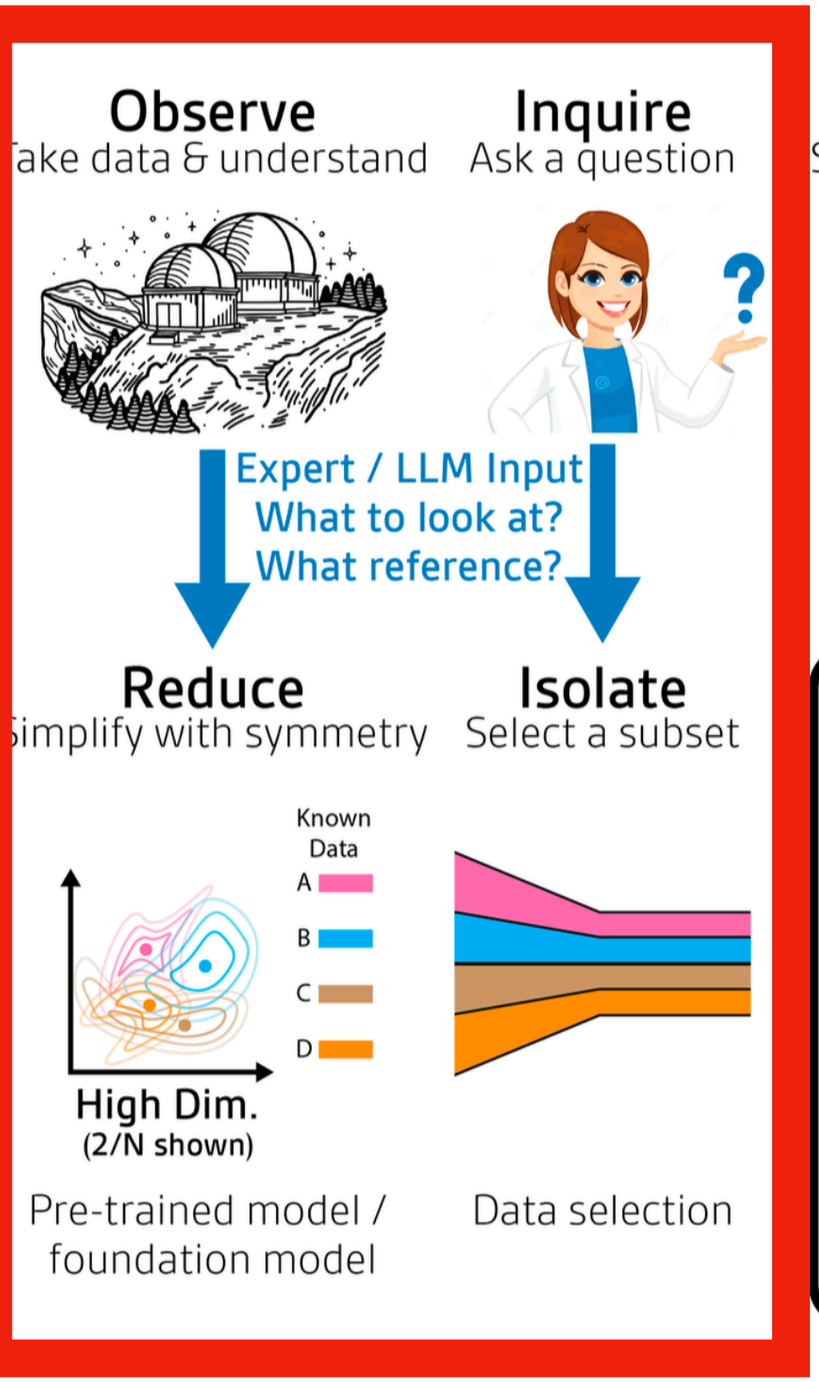
All possibilities at once



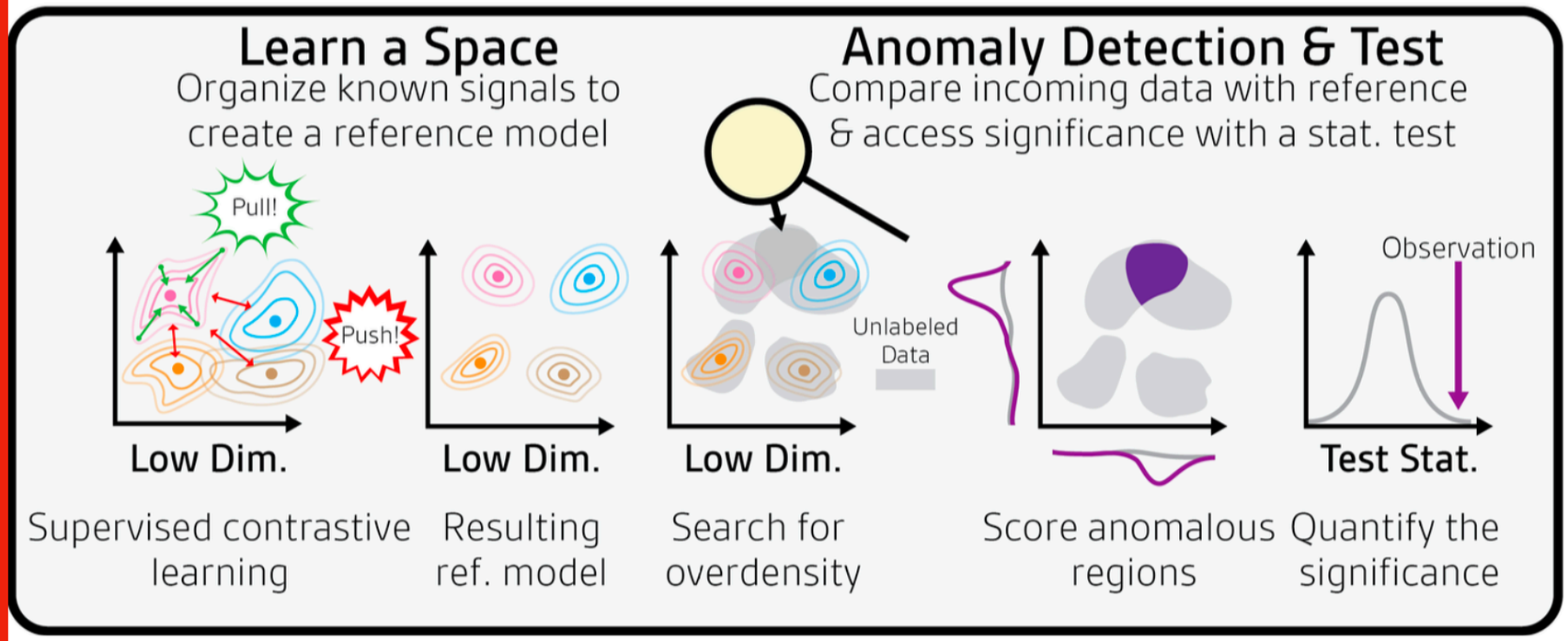
Our automated pipeline is fully differentiable

Arxiv:2510.21935
Neruips 2025

All possibilities at once



AutoSciDACT



What about this part?

Are you sick of AI Slop?

Do you like it?

Are you sick of AI Slop?

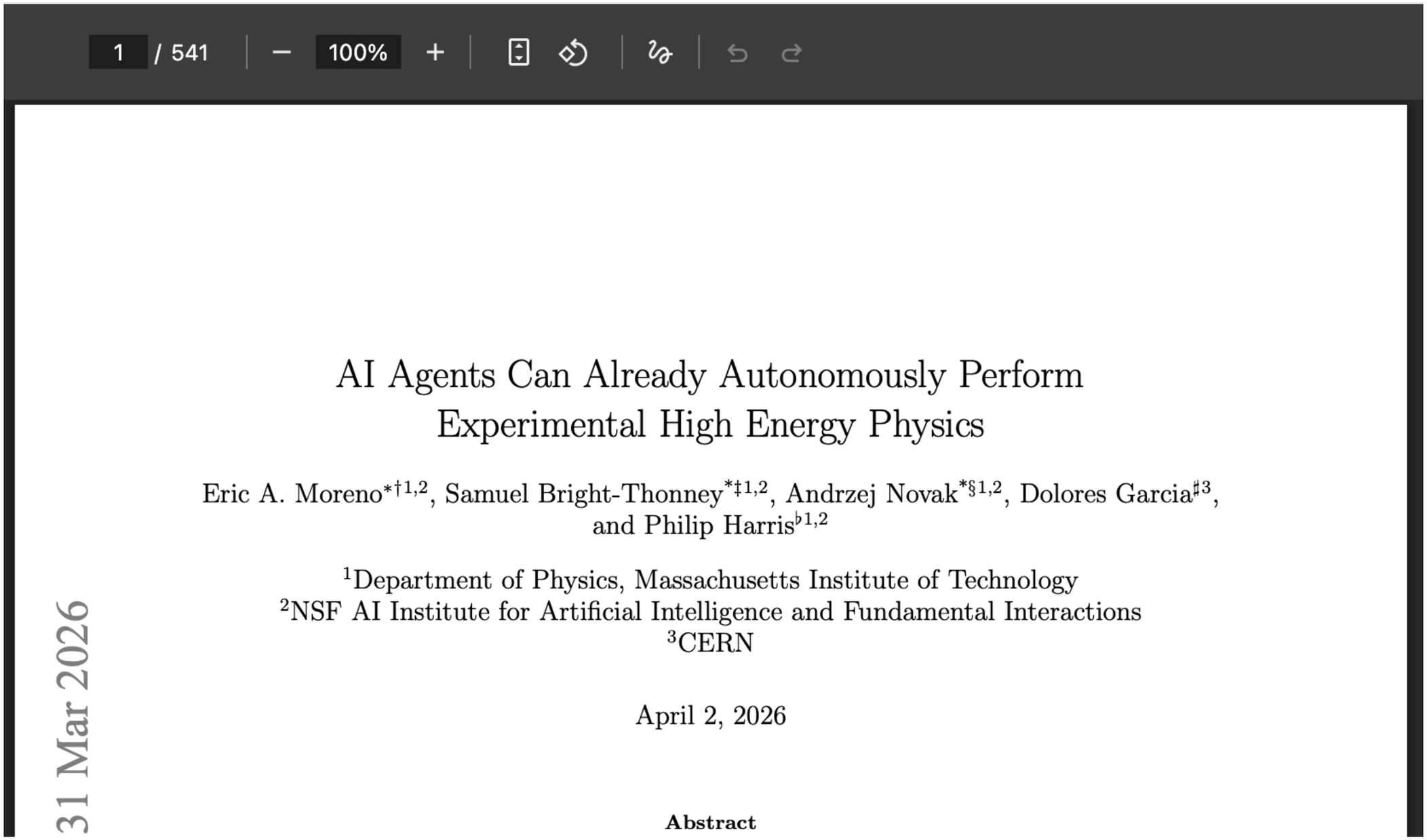
Do you like it?

Slop Until we Drop!

Automating Analyses

- If we can re-refine our neural networks fitting data
  - We can effectively train on data, with our full model
    - Following a pre-selection or dataset
      - Ultimately means tuning simulation if its differentiable
        - Can reach paradigm of a near perfect physics model
- What if we can go further and automate the analysis procedure
  - Large Language Models and Agentic AI propose a solution

Pursuing full automation of Physics Analysis



AI Agents Can Already Autonomously Perform Experimental High Energy Physics

Eric A. Moreno*†[1,2], Samuel Bright-Thonney*‡[1,2], Andrzej Novak*§[1,2], Dolores Garcia♯[3], and Philip Harris♭[1,2]

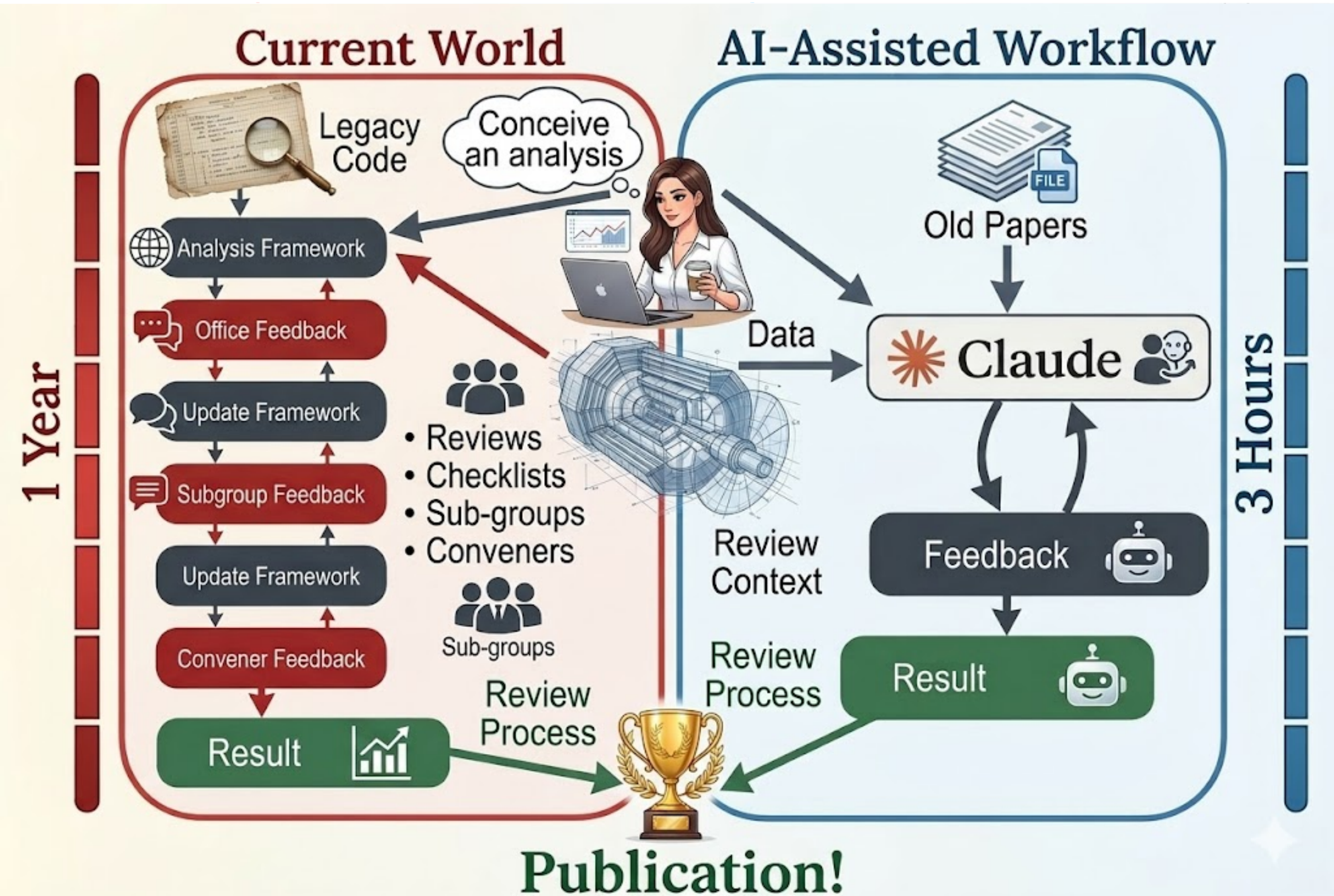
[1]Department of Physics, Massachusetts Institute of Technology
[2]NSF AI Institute for Artificial Intelligence and Fundamental Interactions
[3]CERN

April 2, 2026

31 Mar 2026

Abstract

https://arxiv.org/pdf/2603.20179


Current World
AI-Assisted Workflow
1 Year
3 Hours
Legacy Code
Conceive an analysis
Analysis Framework
Office Feedback
Update Framework
Subgroup Feedback
Update Framework
Convener Feedback
Result
• Reviews
• Checklists
• Sub-groups
• Conveners
Sub-groups
Review Process
Old Papers
FILE
Data
Claude
Review Context
Feedback
Review Process
Result
Publication!

What does a Prompt Look Like?

Scaffold and run a measurement analysis of the primary Lund jet plane density in hadronic Z decays using archived ALEPH data at $\sqrt{s} \sim 91.2$ GeV.

Setup: scaffold analyses/lund_plane as a measurement, set data_dir=[...] in .analysis_config, install the pixi environment, then begin orchestrating.

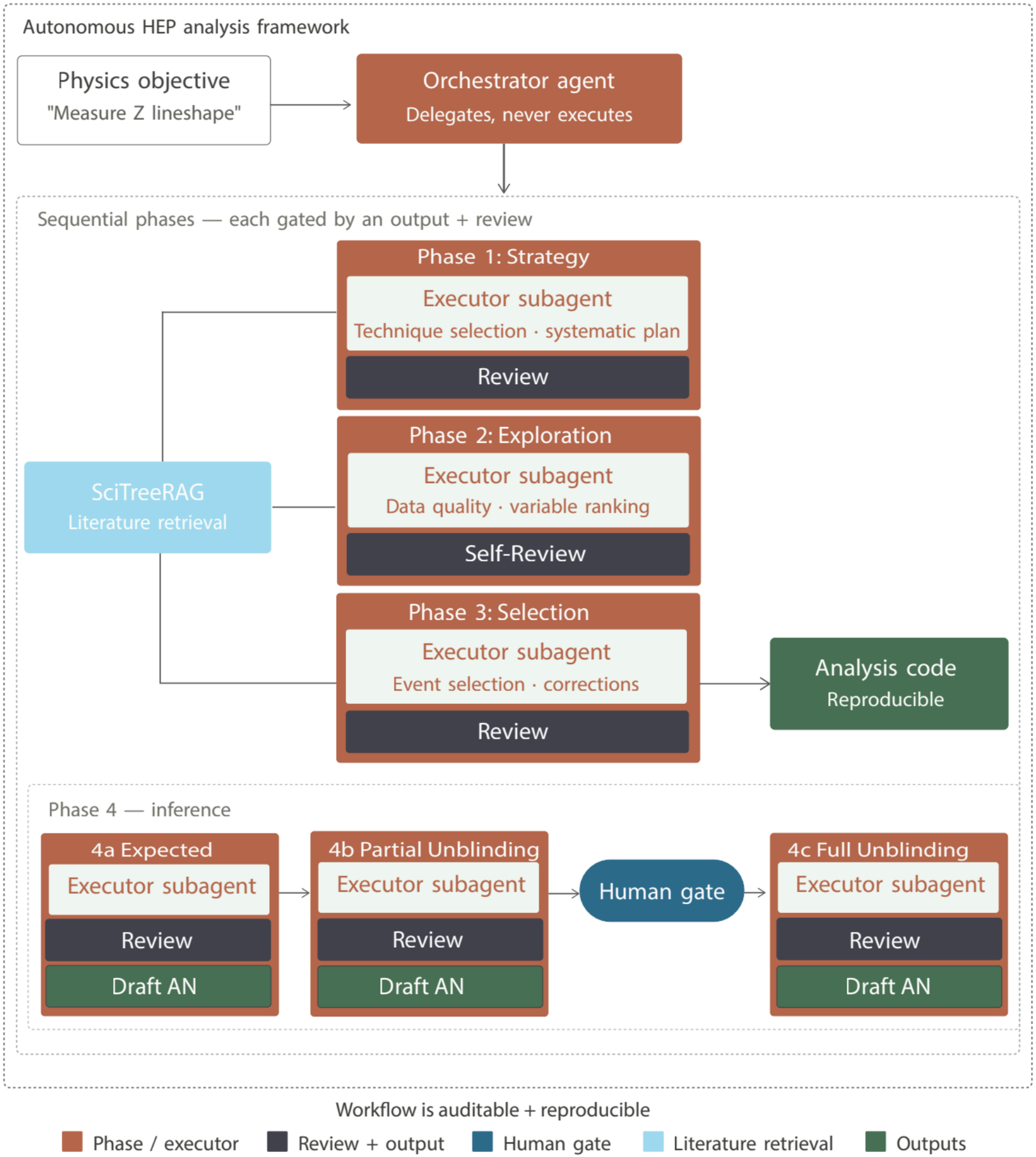
Observable: The 2D density of primary Cambridge/Aachen declusterings in each thrust hemisphere, mapped to coordinates $(\ln 1/\Delta\theta,\ \ln k_t/\mathrm{GeV})$, where $\Delta\theta$ is the emission angle and $k_t = E_{\mathrm{soft}} sin\Delta\theta$. Use charged particles only (`pwflag` == 0, `highPurity` == 1). One jet = one hemisphere.

Deliverables:

1. 2D density $\rho(\ln 1/\Delta\theta, \ln k_t)$ corrected to charged-particle level, 10–15 × 10–15 bins at least - as fine as it makes sense
2. 1D projections (k_t spectrum, angular spectrum) with covariance
3. Number of primary declusterings vs. minimum k_t threshold
4. Comparison to PYTHIA 6 MC
5. Machine-readable results (CSV/NPY)

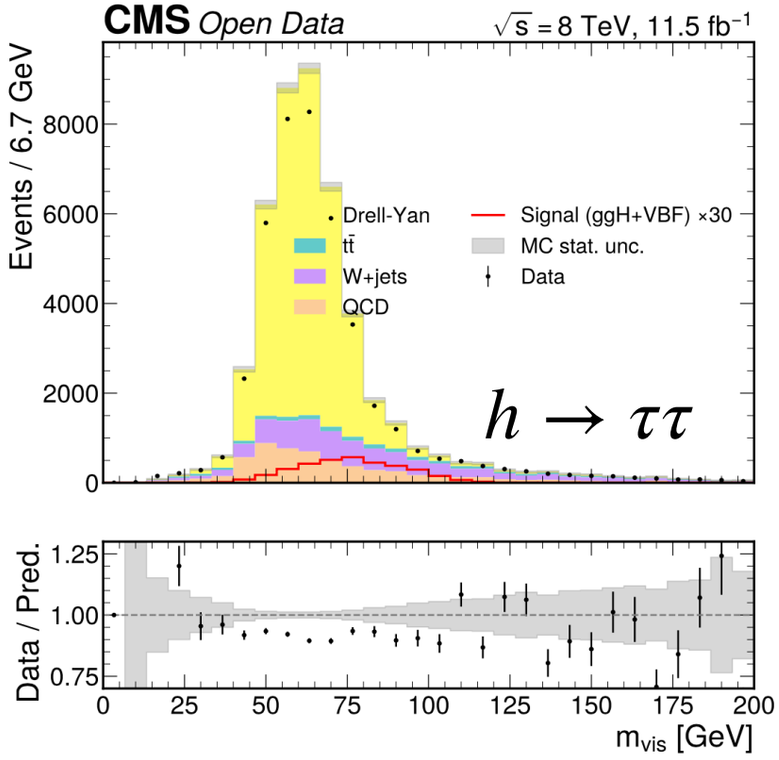
Data: [...]

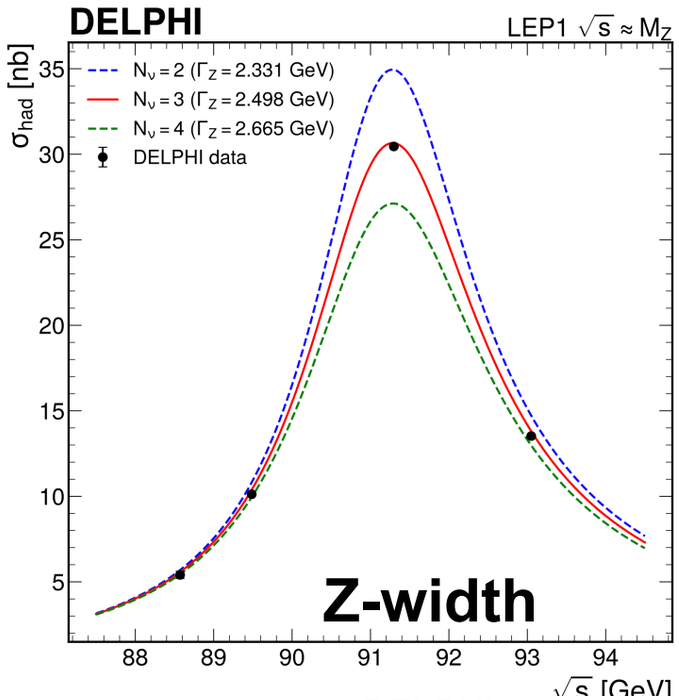
Think of each agent as human(student) that sp physics who you give instructions to in a bun text files

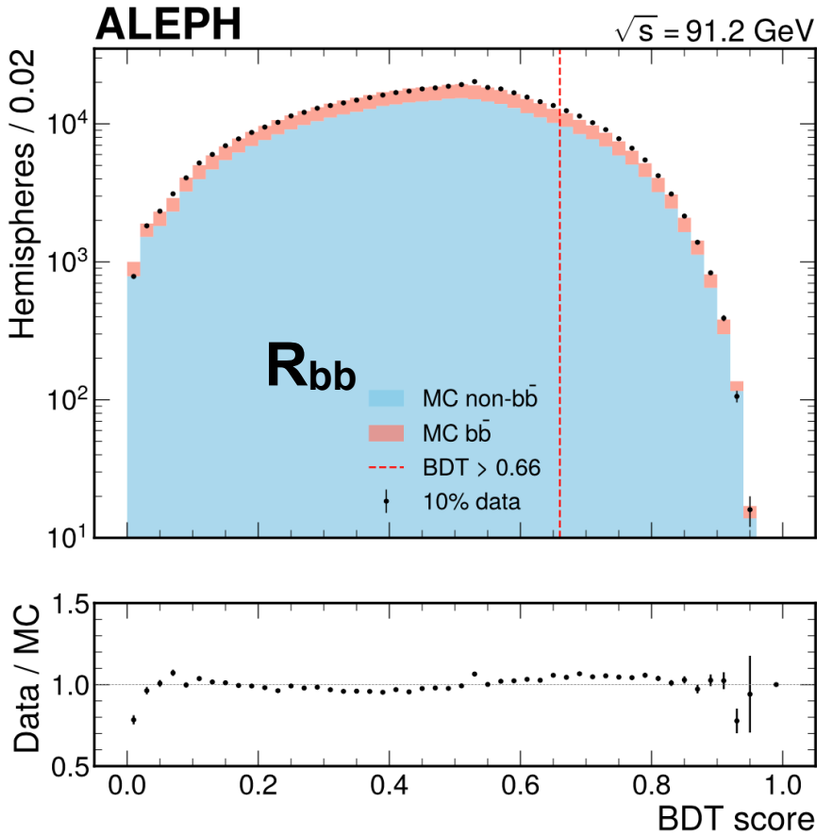


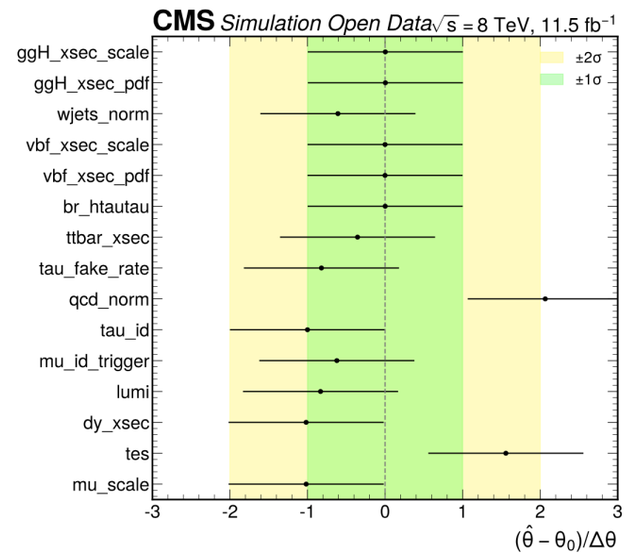
How good is it?

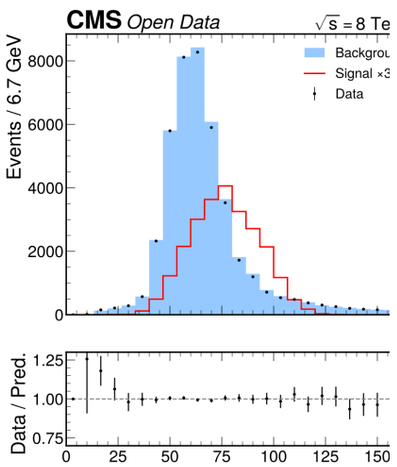
- This is just from two days

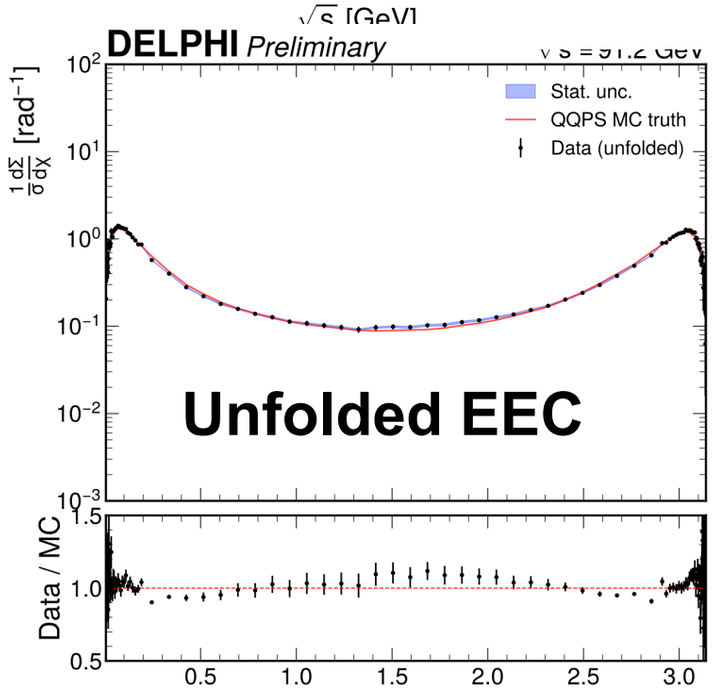


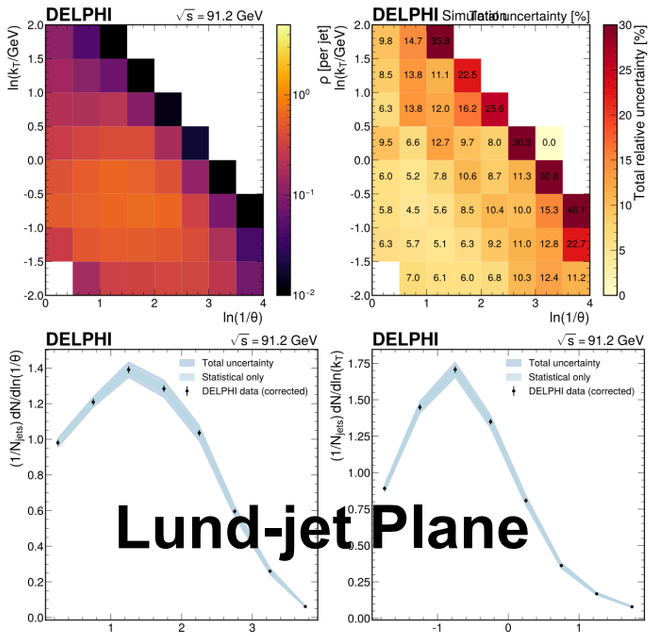




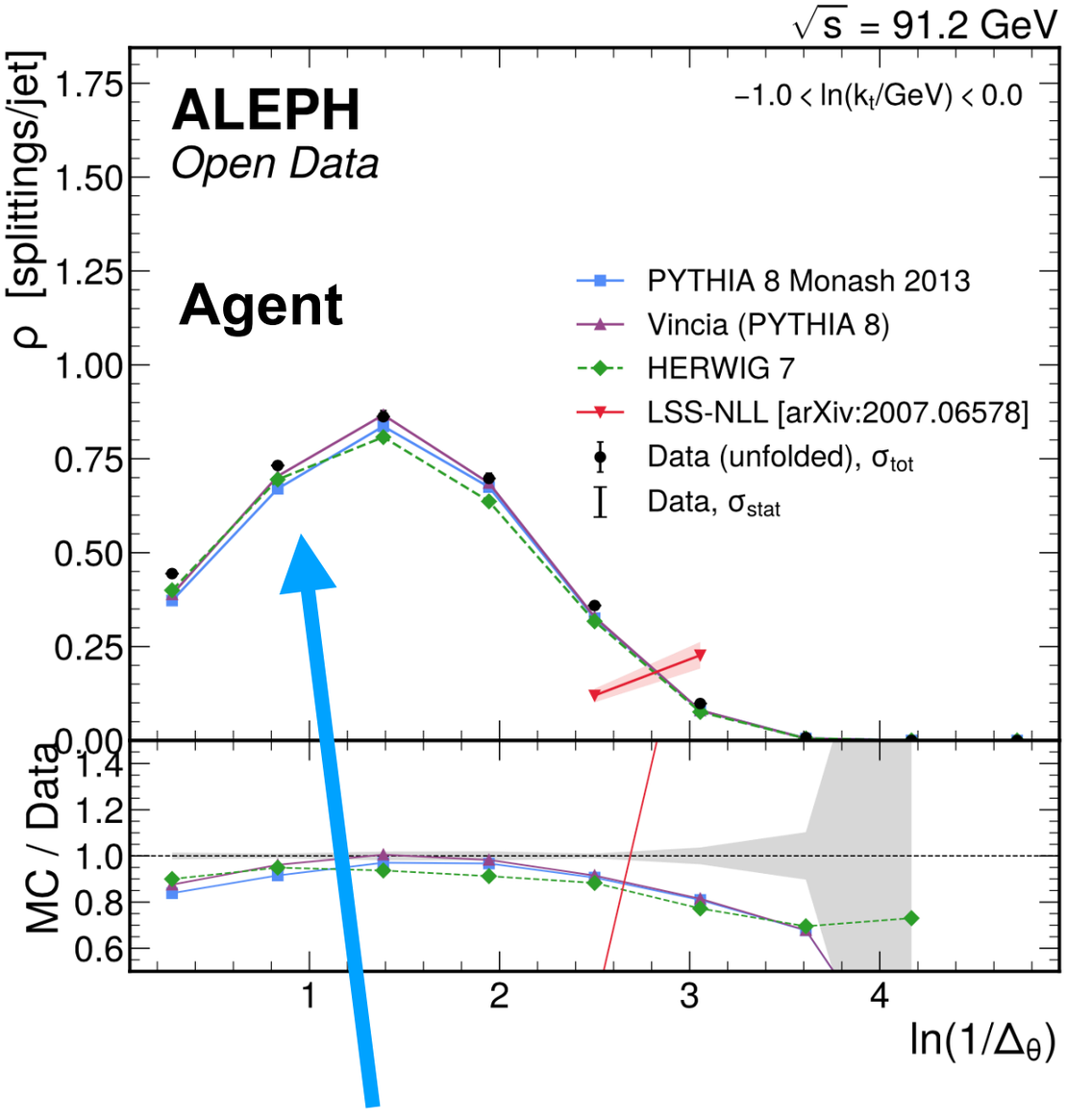


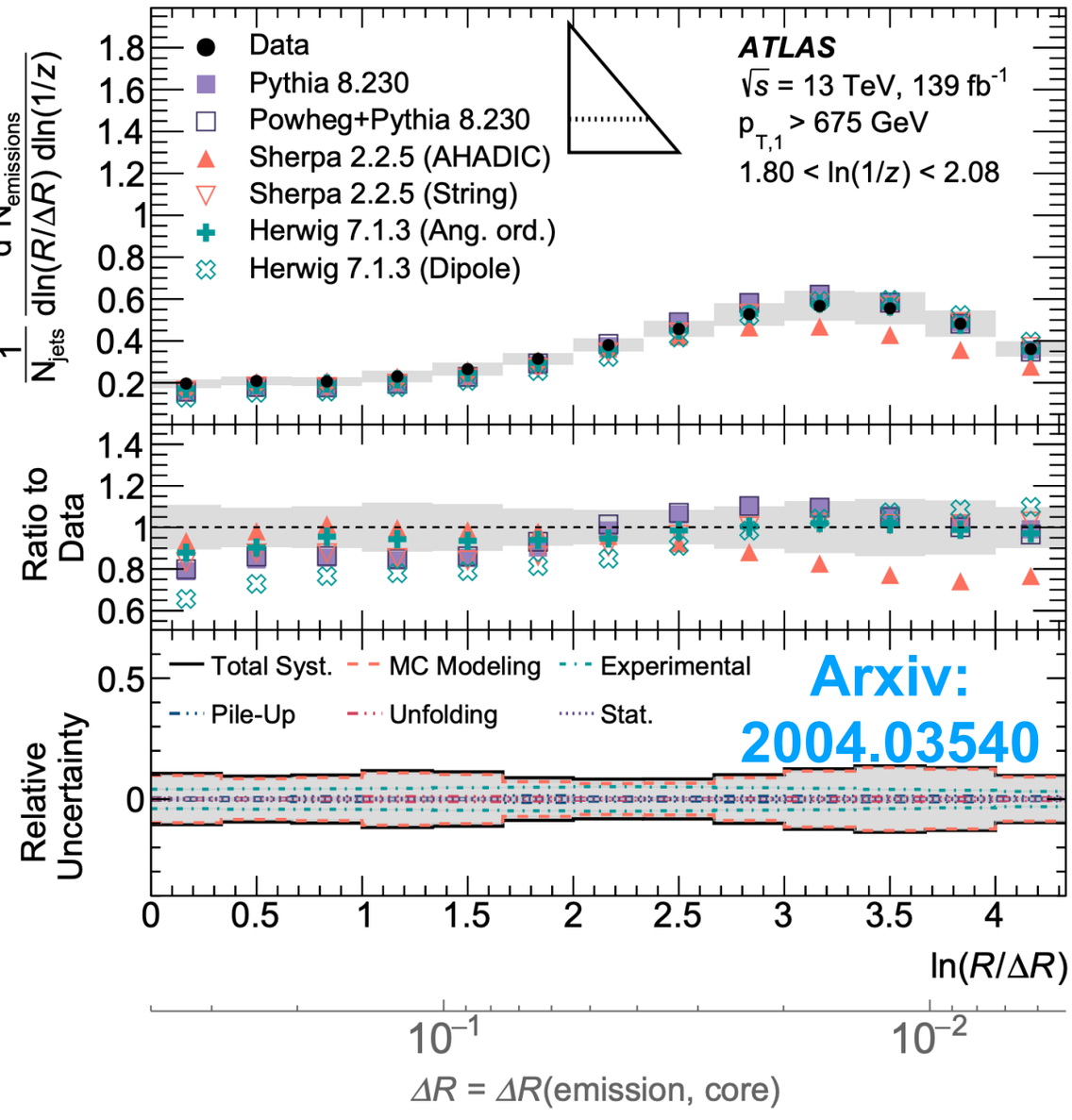






How good is it really?

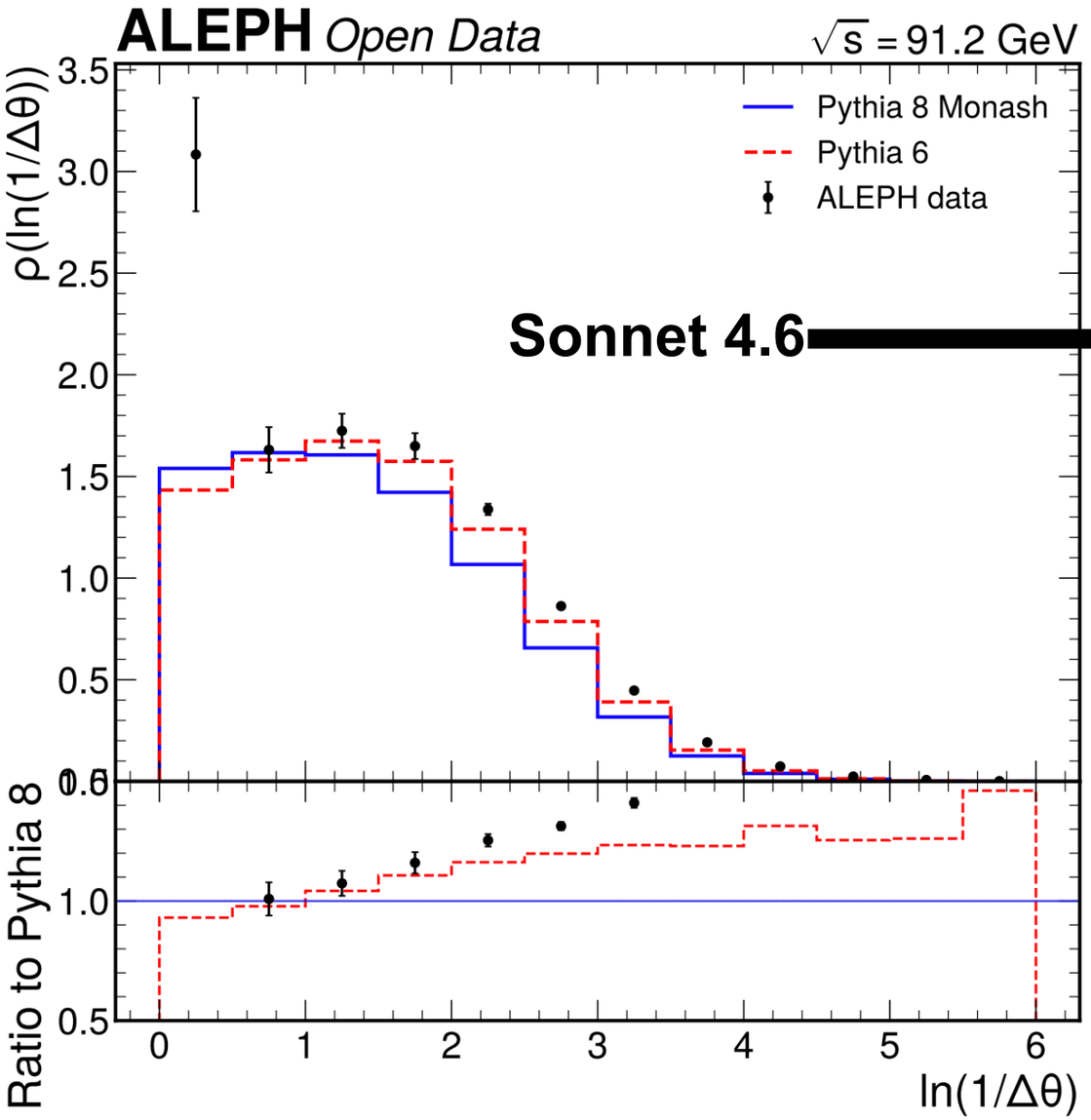




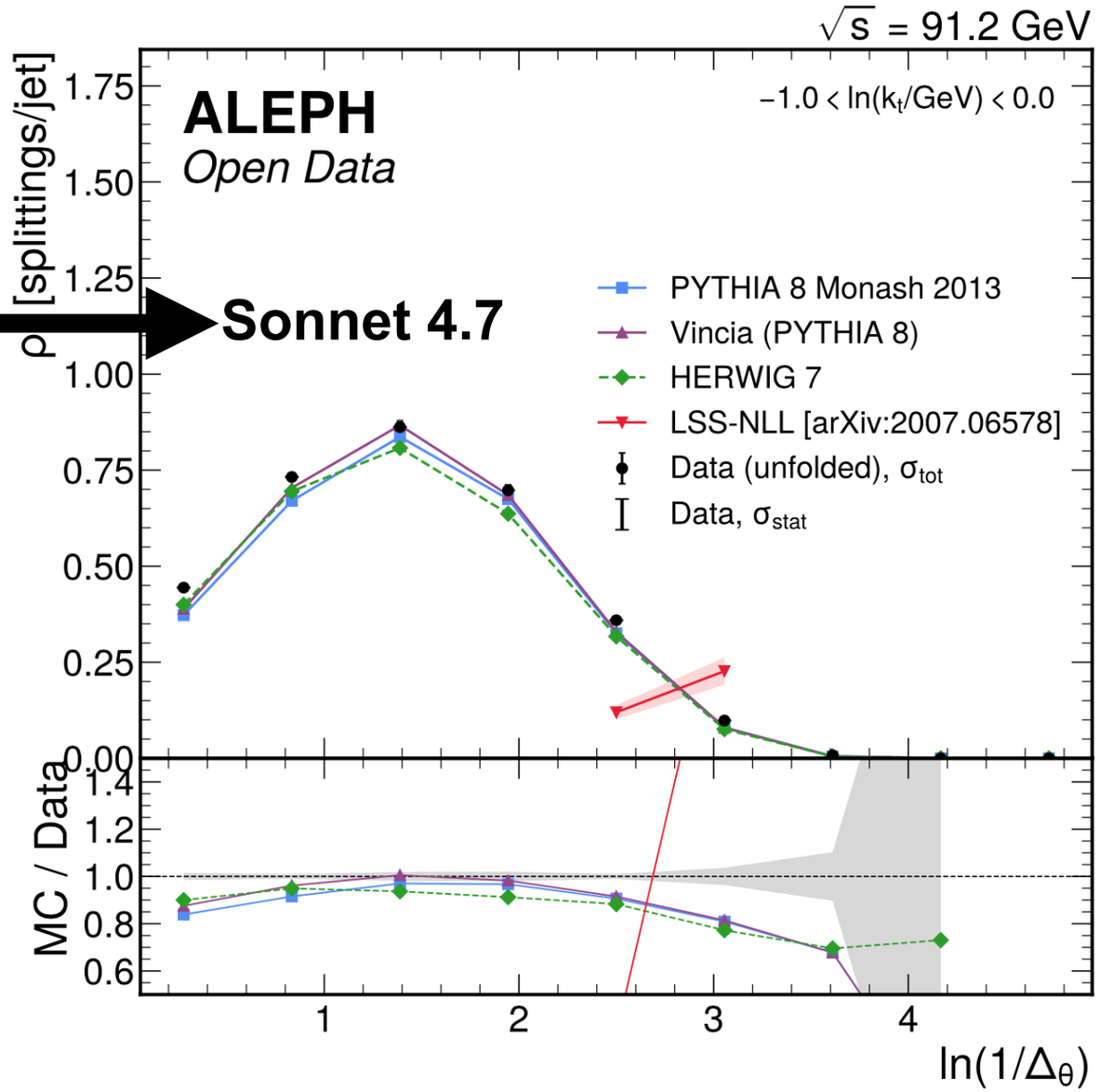
Similar trends, qualitatively good performance

This is a novel result, nobody has done it

How Fast is It Improving?



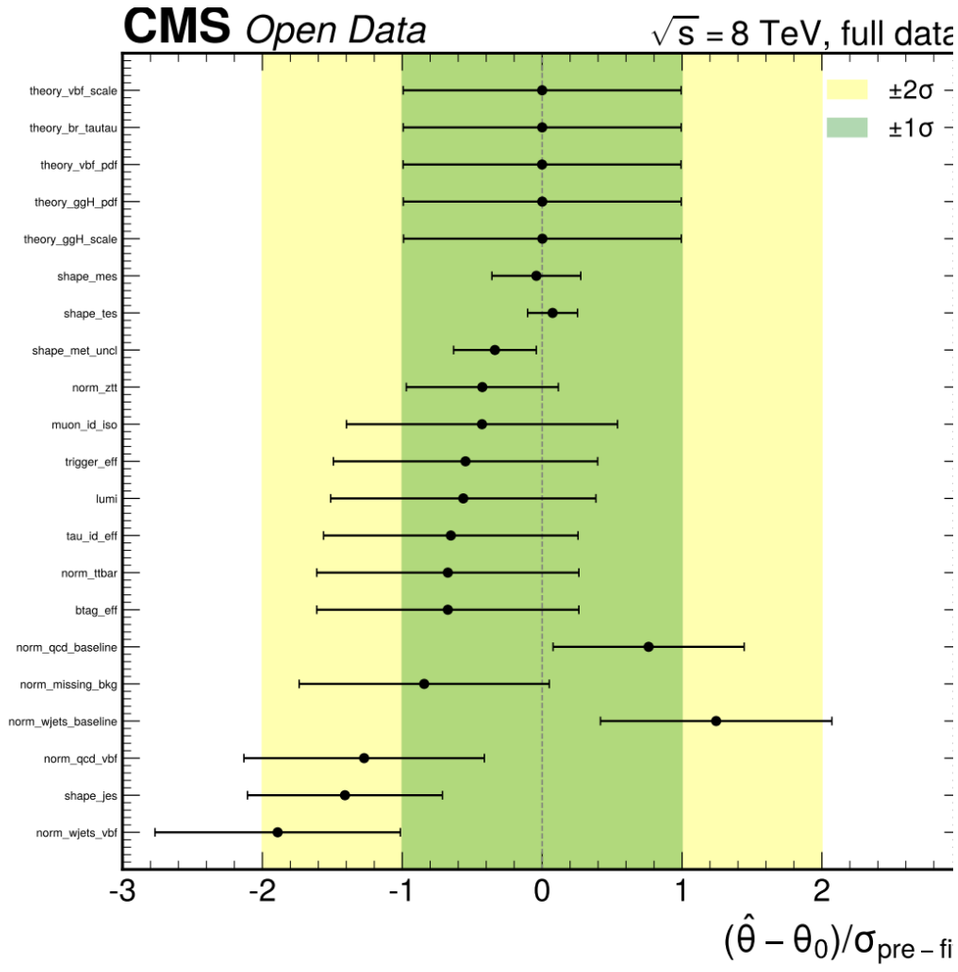
Sonnet 4.6 → **Sonnet 4.7**

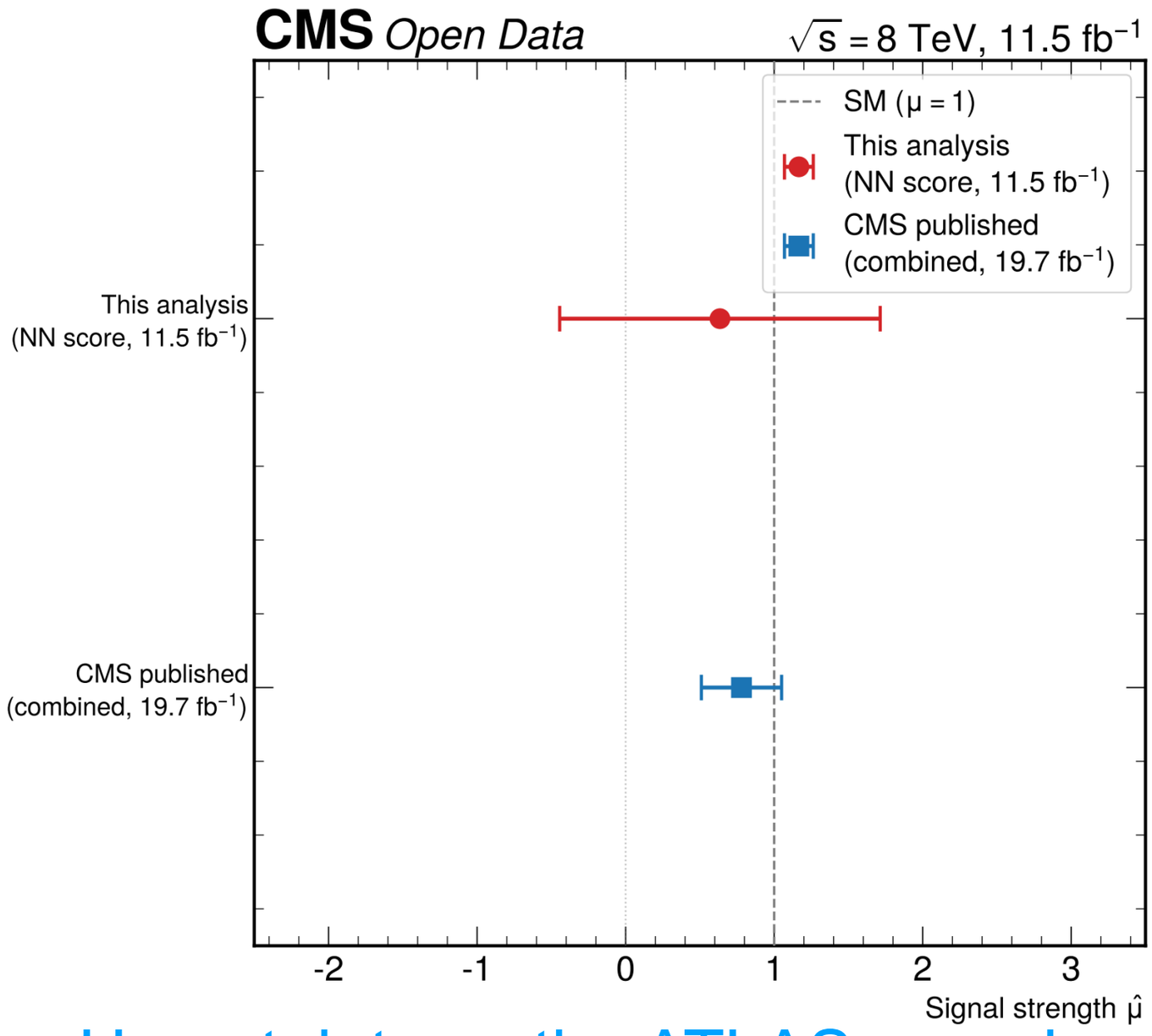


Results are in many ways better
But they are still not controllled

How good is it really?

How does it compare to a reference





LLM: 0.63 ±1.08

CMS*: 1.01±0.59

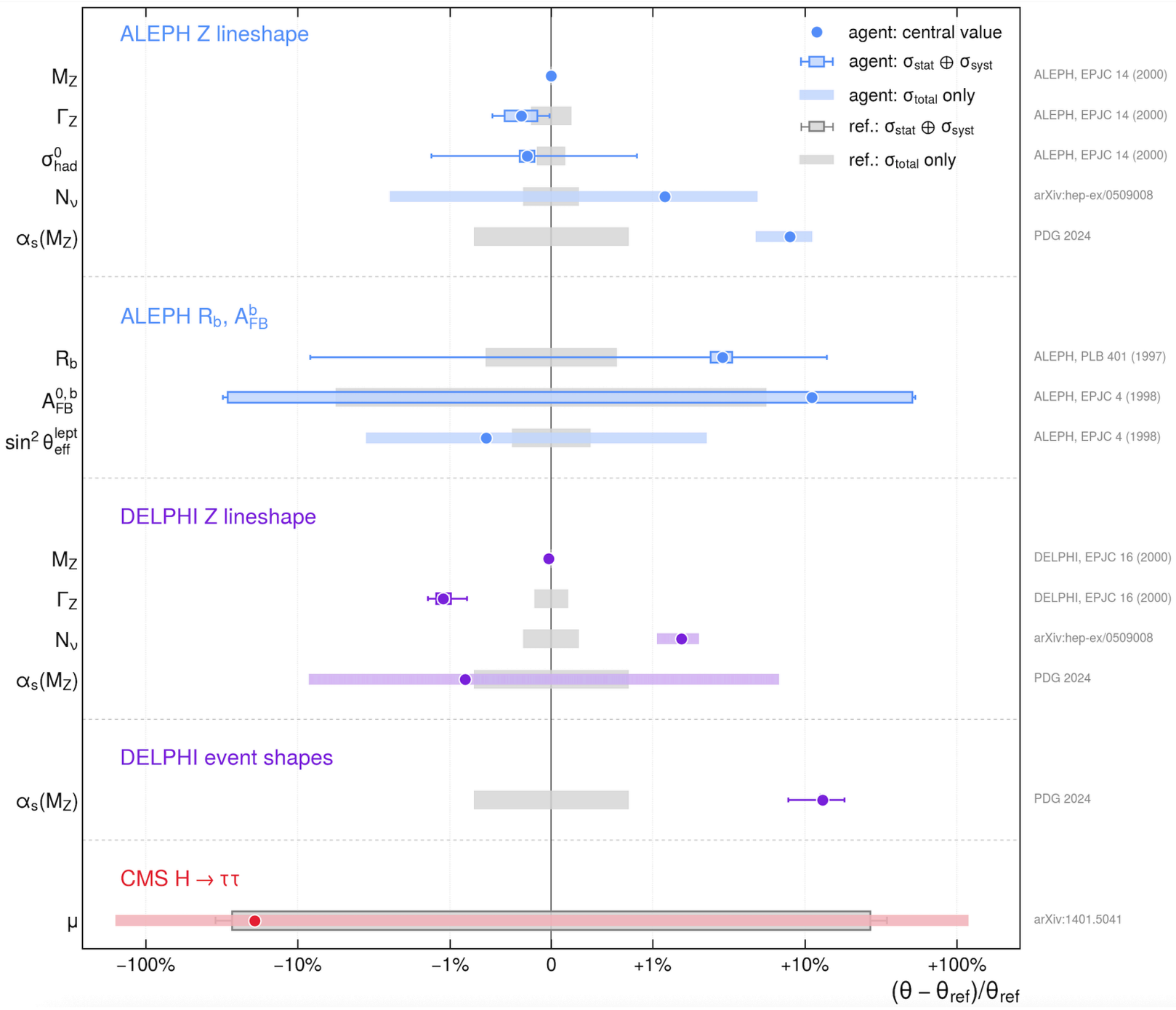
ATLAS*:1.1±1.09

Same Uncertainty as the ATLAS experiment

*Very approximate scaling

How good is it really?

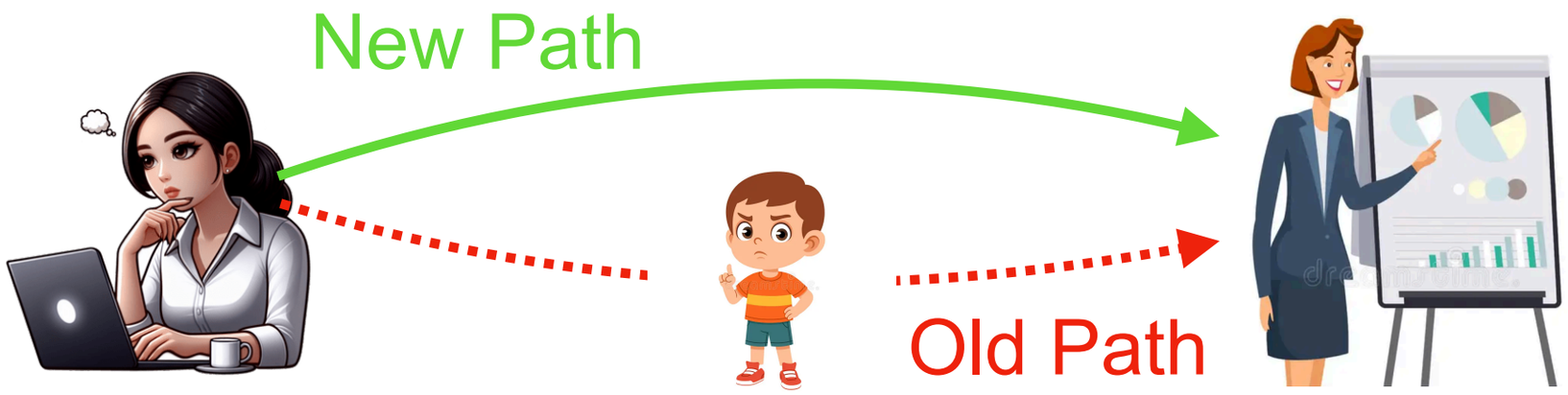
How does it compare to a reference

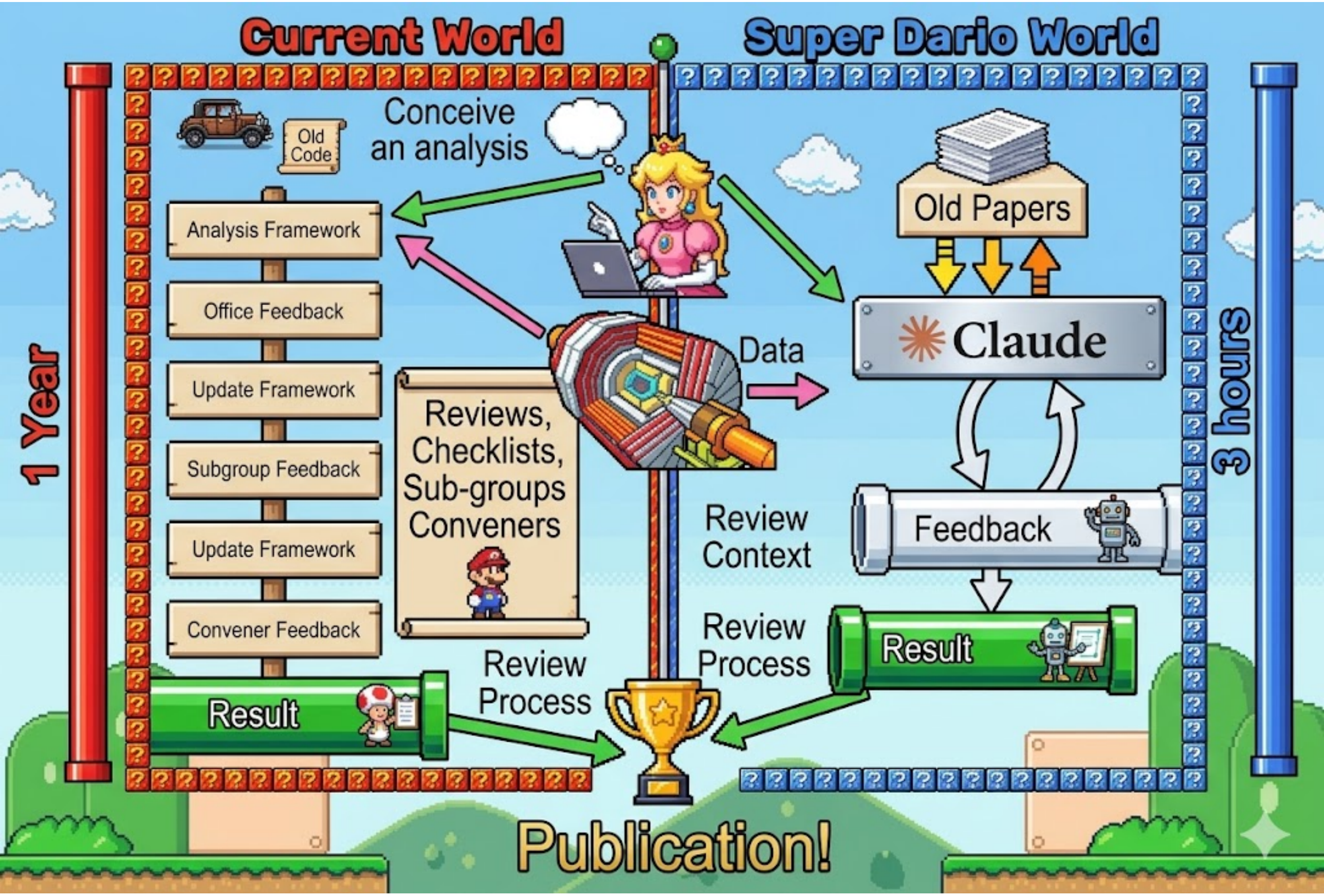


Some Things
Are consistent
Some things
deviate

What does this mean?

- We can think bigger and bolder about our research
  - Less analysis frameworks/Streamlined review process
  - Preservation of Physics data becomes easy
- Ultimately students/PD/Profs go back to physics
  - Students will work more like professors

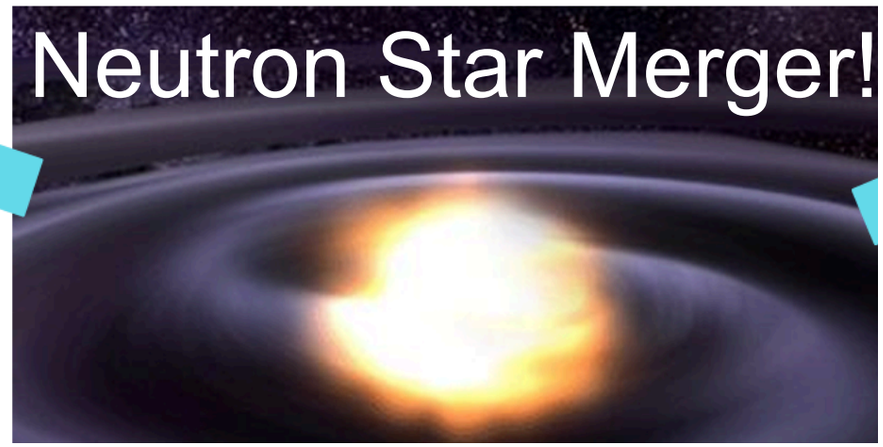



Current World
Super Dario World
1 Year
3 hours
Old Code
Conceive an analysis
Analysis Framework
Office Feedback
Update Framework
Subgroup Feedback
Update Framework
Convener Feedback
Result
Review Process
Reviews, Checklists, Sub-groups Conveners
Data
Old Papers
Claude
Review Context
Feedback
Review Process
Result
Publication!

Gravity

Multi-Messenger Astronomy

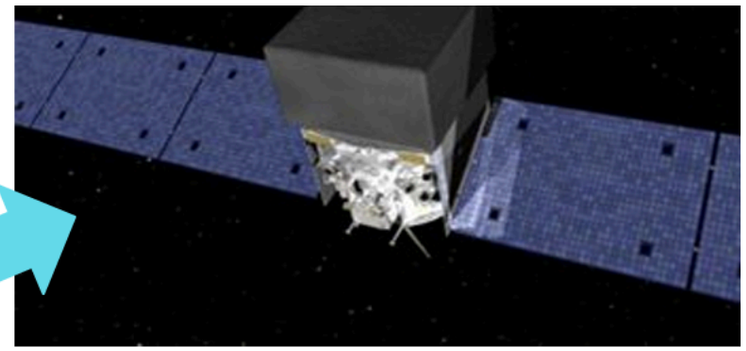
Gravitational waves



A gravitational wave will see a Neutron star before meger

Multi-Messenger Astronomy

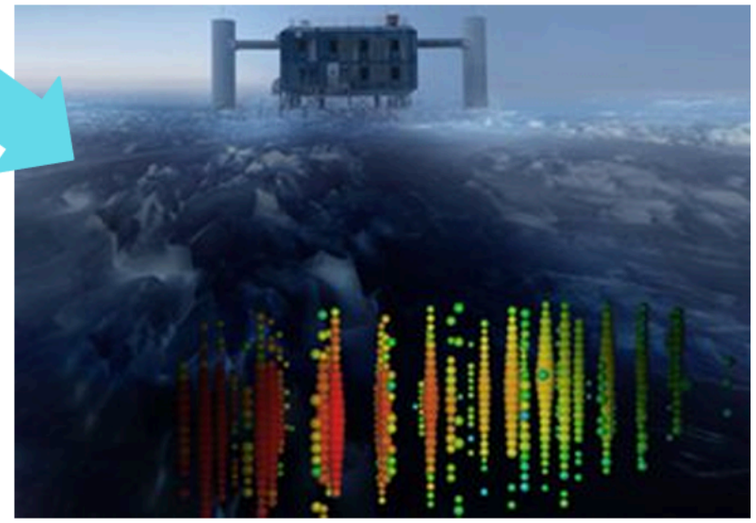
Gravitational waves



X-rays/Gamma-rays

Visible/infrared light

Radio waves

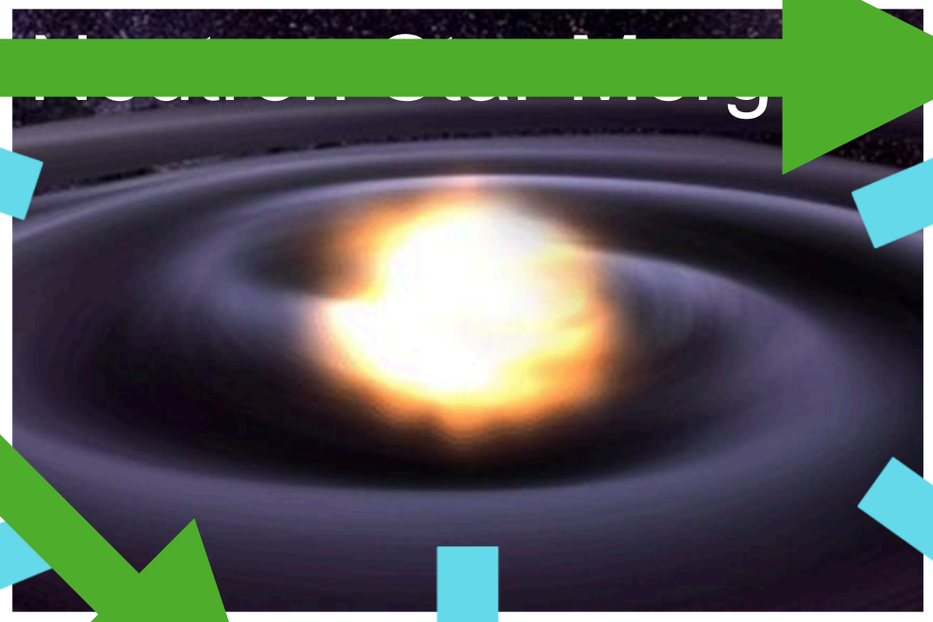


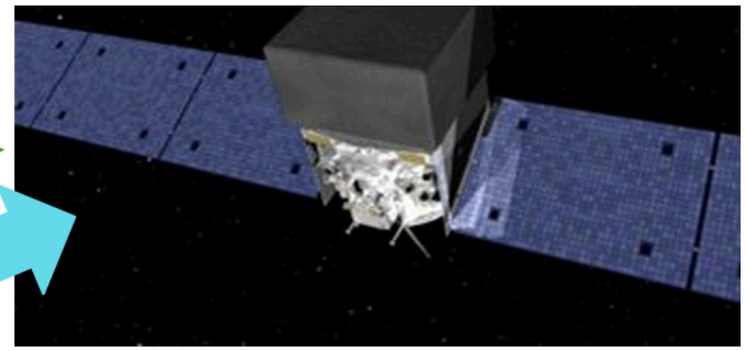
Neutrinos

Everybody Else will see it much further after

Multi-Messenger Astronomy

Gravitational waves

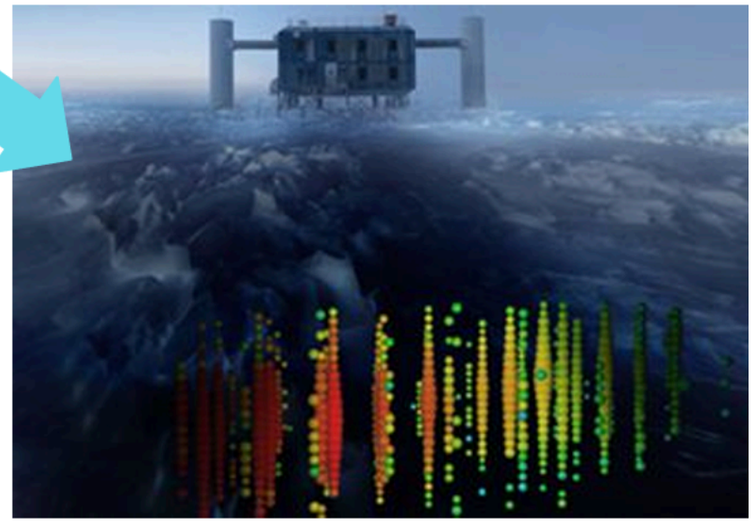




X-rays/Gamma-rays

Visible/infrared light

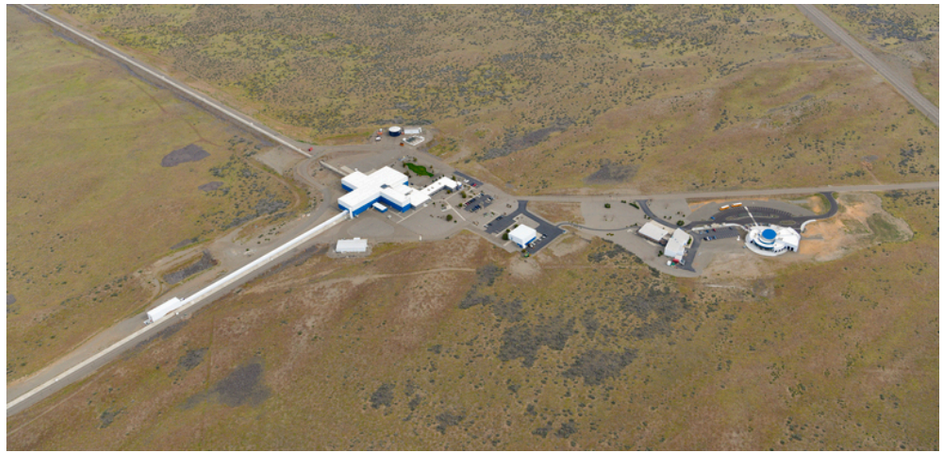
Radio waves

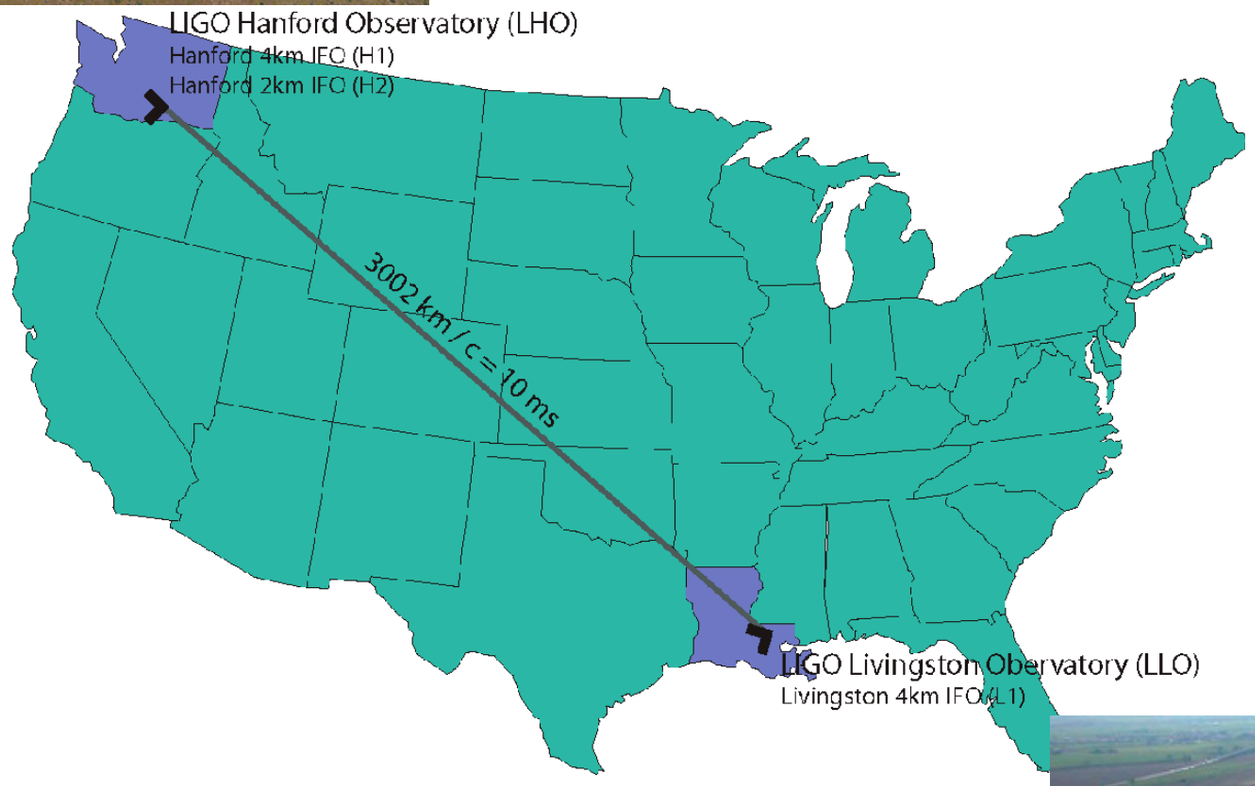


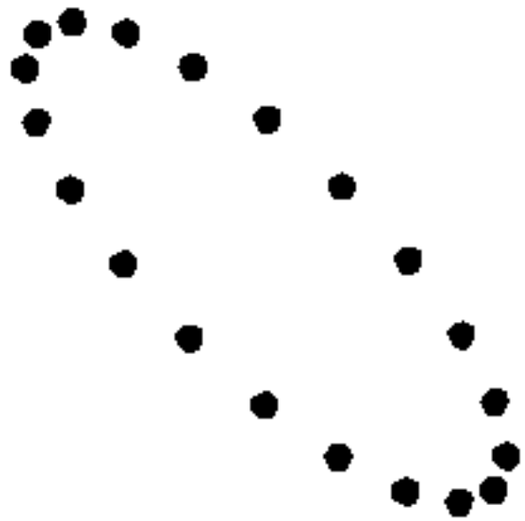
Neutrinos

Performing fast identification of GWs critical to alerting world!

LIGO

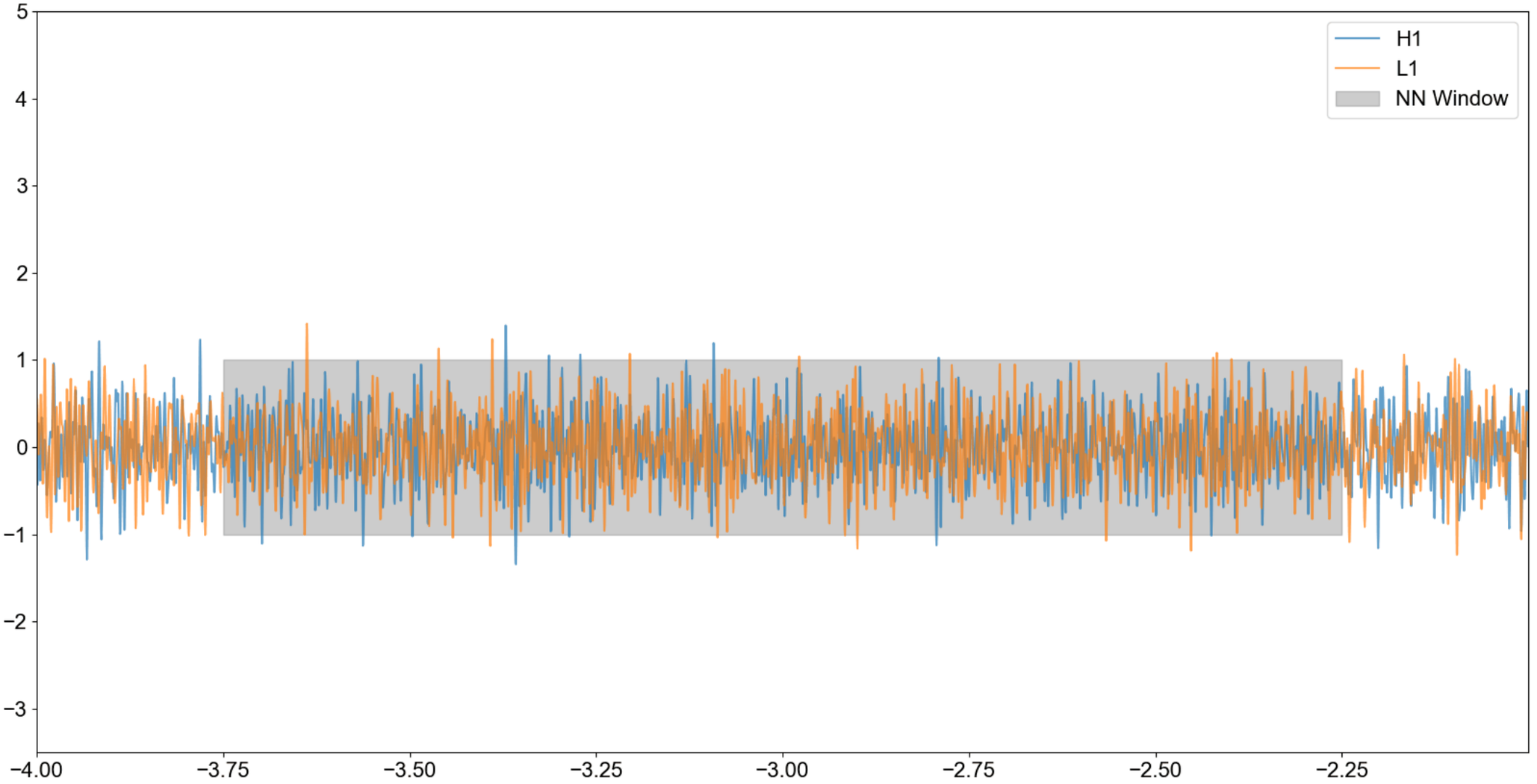






Two detectors that are sensitive to GW

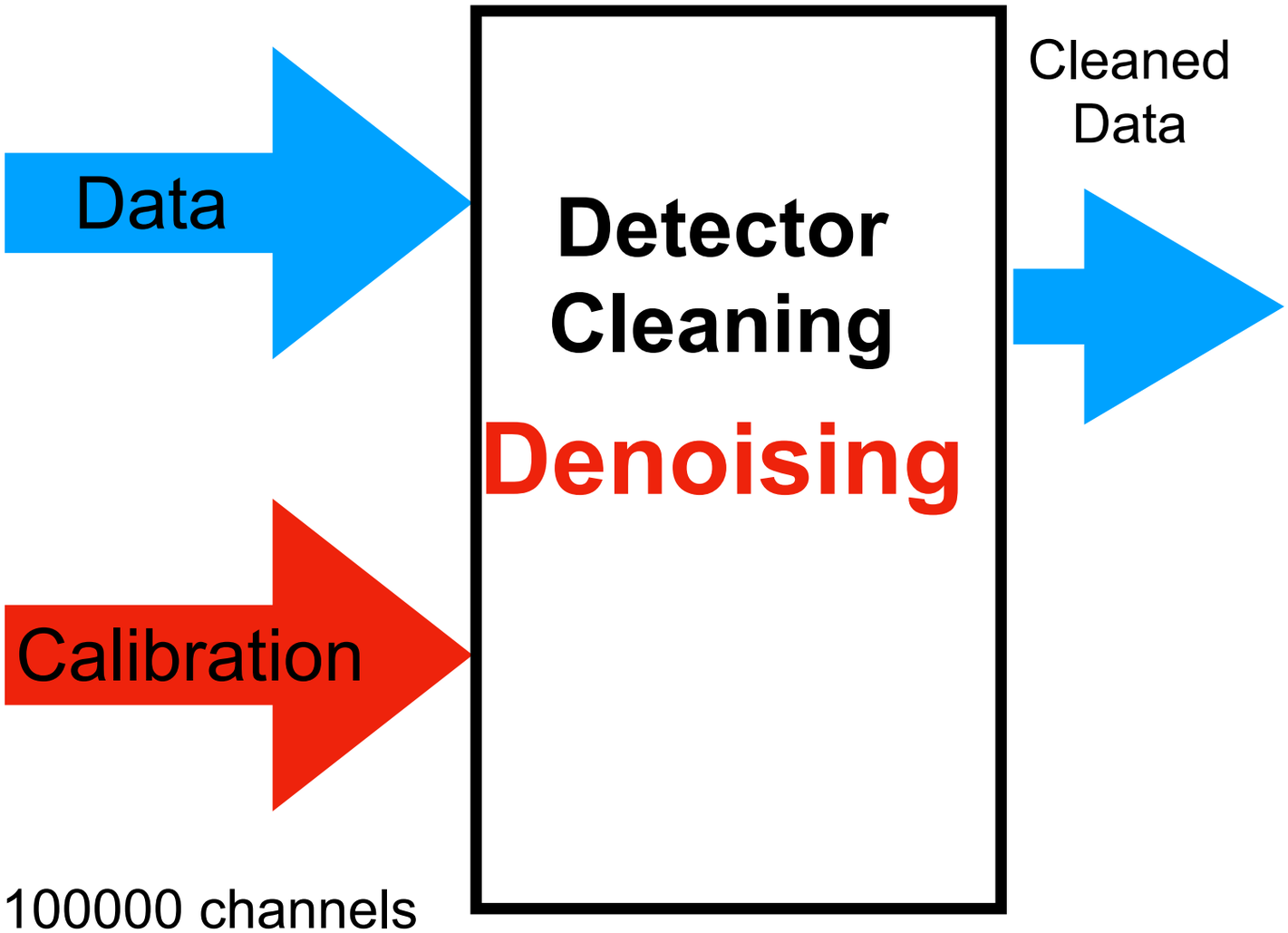
Gravitational Wave Data



Data is time series from at least two detectors

We aim for a correlated signal

GW Data Workflow



GW Data Workflow

Data

Calibration

100000 channels

Detector Cleaning

Denoising

Cleaned Data

Event Detection

Black Holes/ Neutron Stars/ Anomalies

Event

GW Data Workflow

Data

Calibration

100000 channels

Detector Cleaning

Denoising

Cleaned Data

Event Detection

Black Holes/ Neutron Stars/ Anomalies

Event

Event Analysis

Parameter Estimation/ Sky Localization

Alert

GW Data Workflow

Data

Calibration

100000 channels

Detector Cleaning

Denoising

Cleaned Data

Event Detection

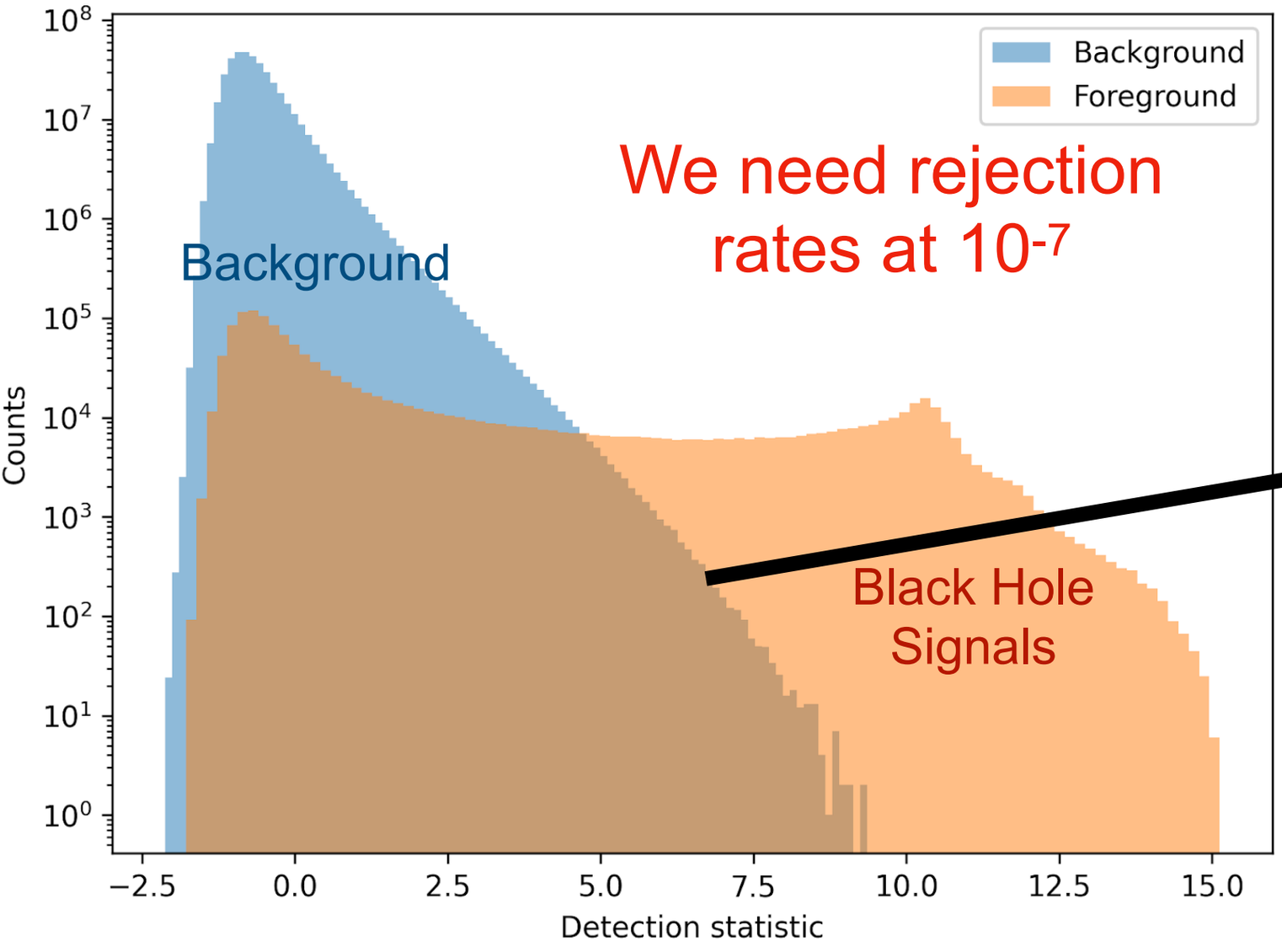
Black Holes/ Neutron Stars/ Anomalies

Event Analysis

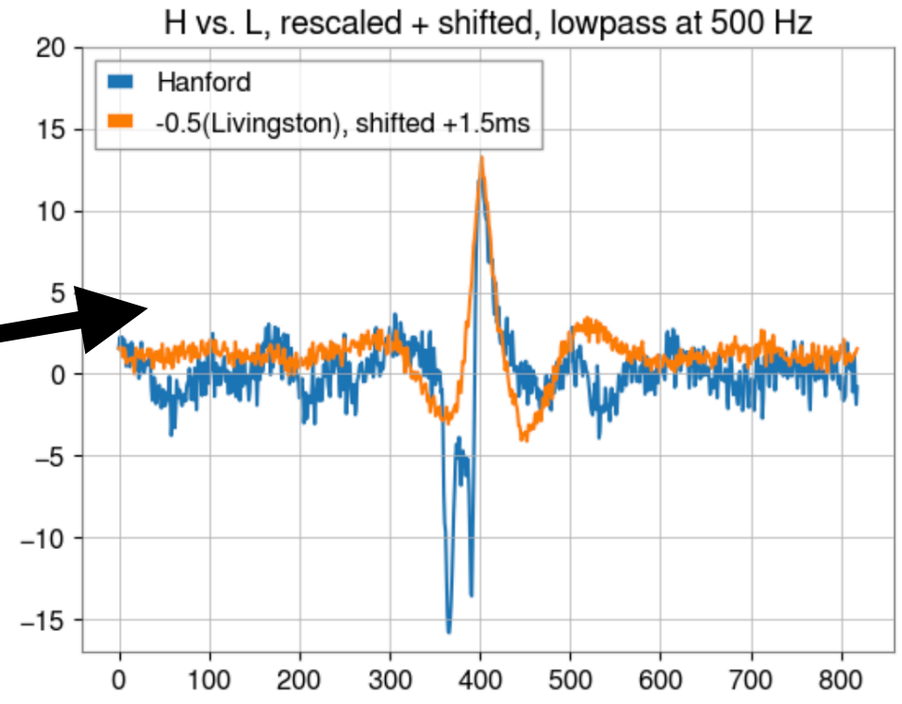
Parameter Estimation/ Sky Localization

We replaced this pipeline with a chain of deep learning algos

Why is this so hard?

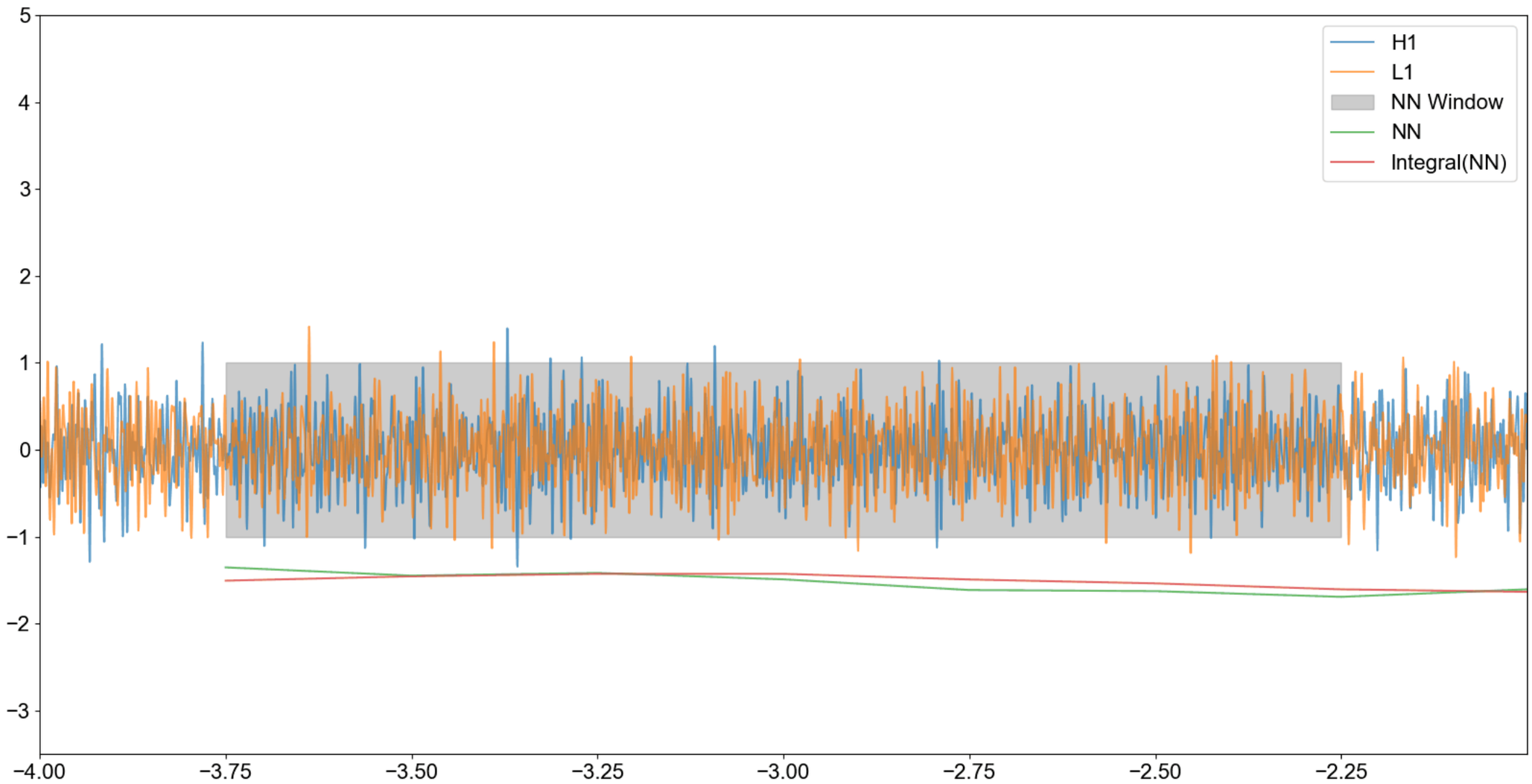


Tail of the Background is all glitches



- Noise source in the bulk are random fluctuations (Markovian)
- Noise source in the tail is all glitches (Non-Markovian)

A-Frame(Black Hole Mergers)

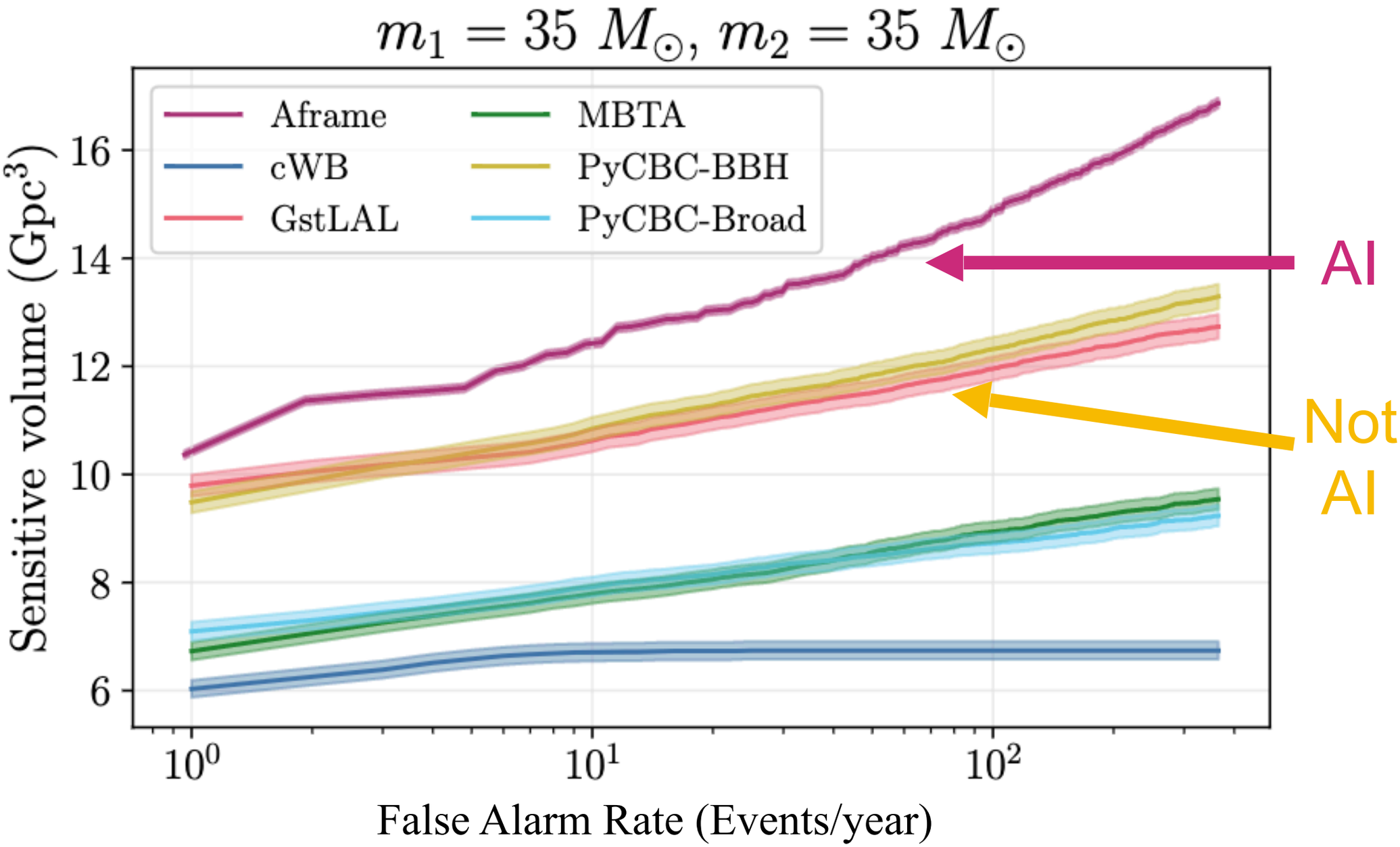


Key Insight: Build a massive training sample on data

Through time, sliding the data between detectors to break correlation

arXiv:2403.18661

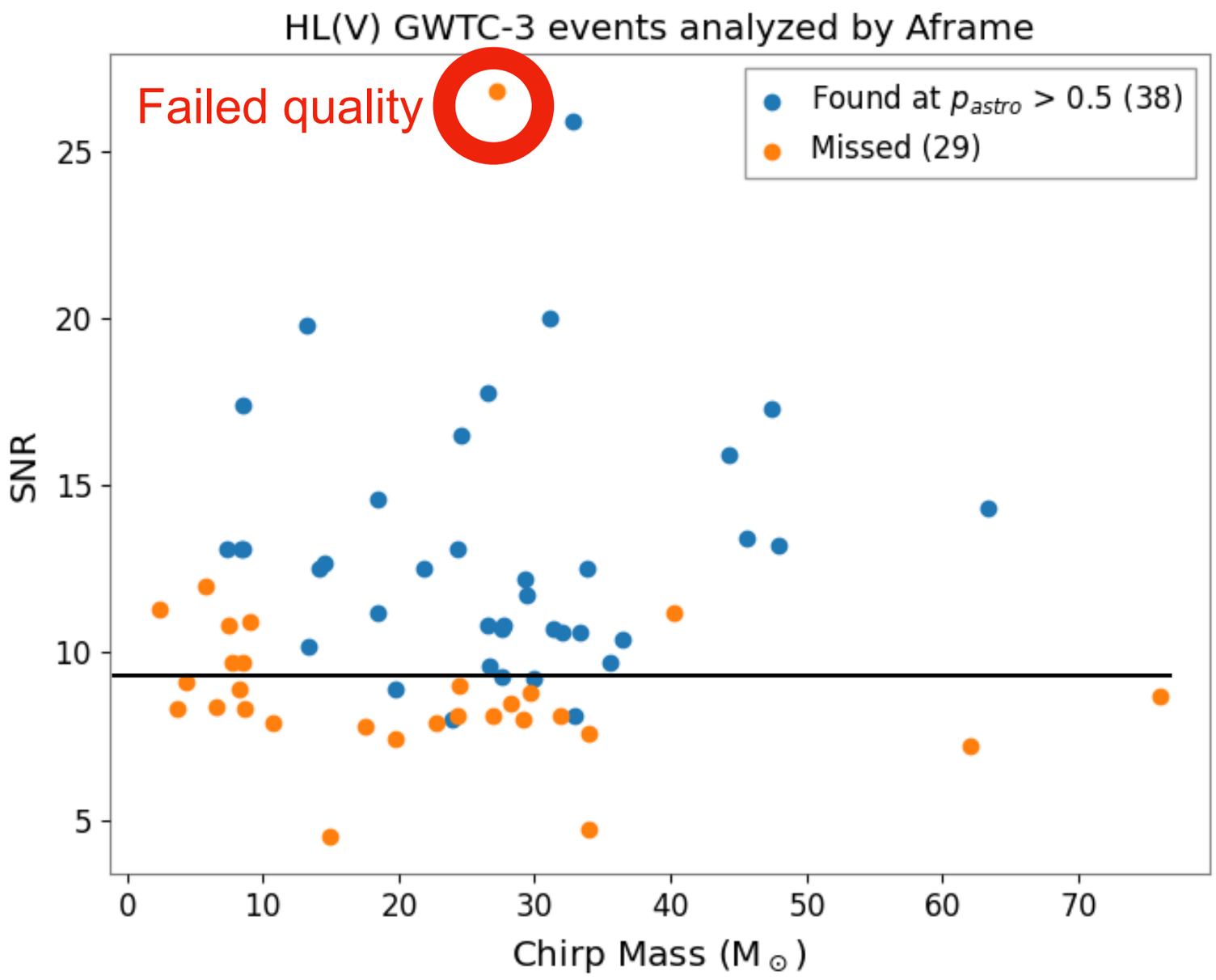
A-Frame(Black Hole Mergers)



Probes 10% of the known universe for Heavy Mergers

arXiv:2403.18661

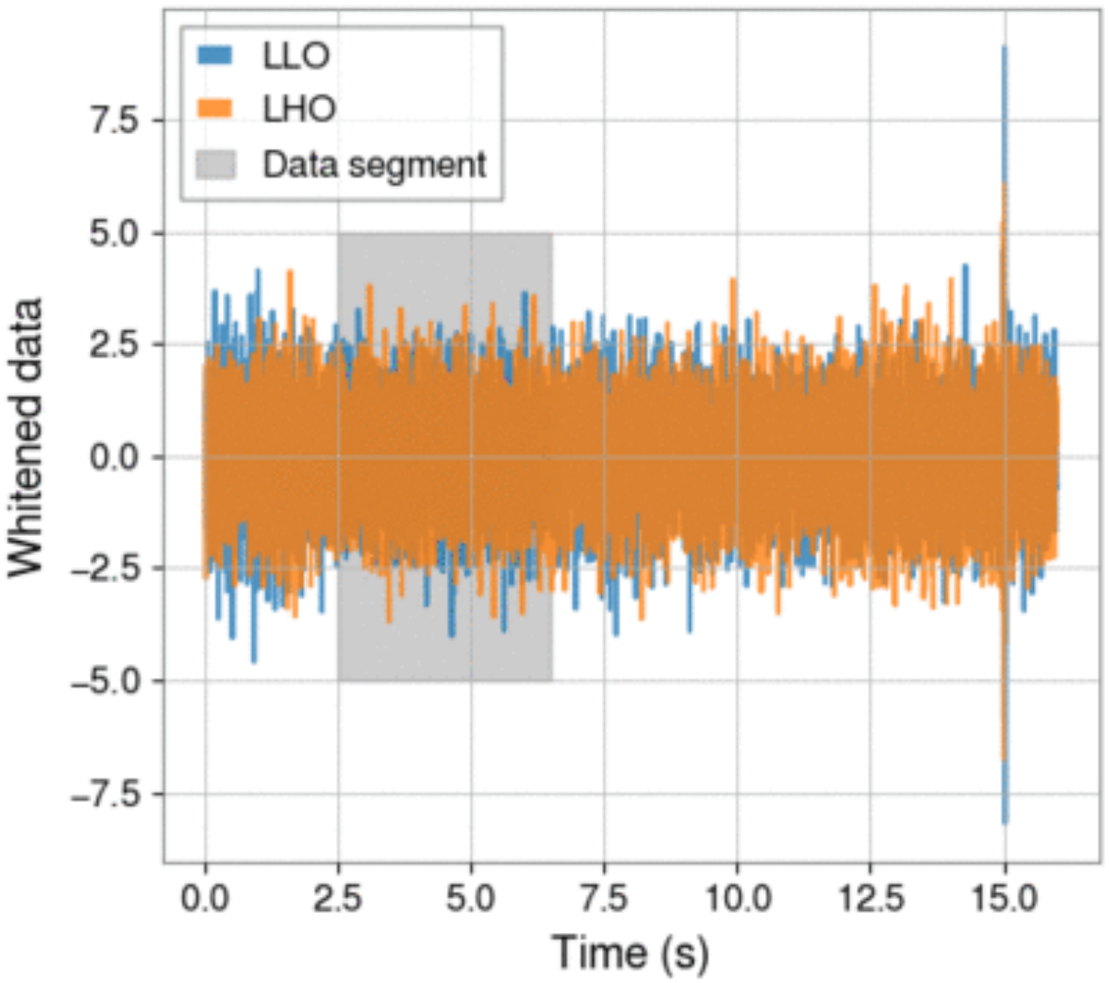
A-Frame(Black Hole Mergers)

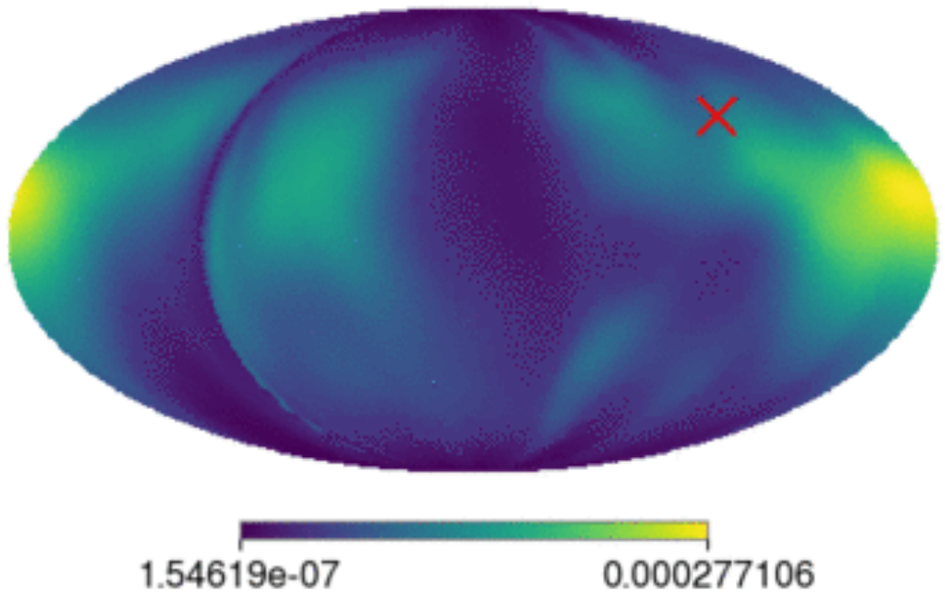


We found all the high signal Black hole mergers

arXiv:2407.19048

Parameter Estimation





SBI for Sky Localization

100,000x times faster than conventional parameter estimation

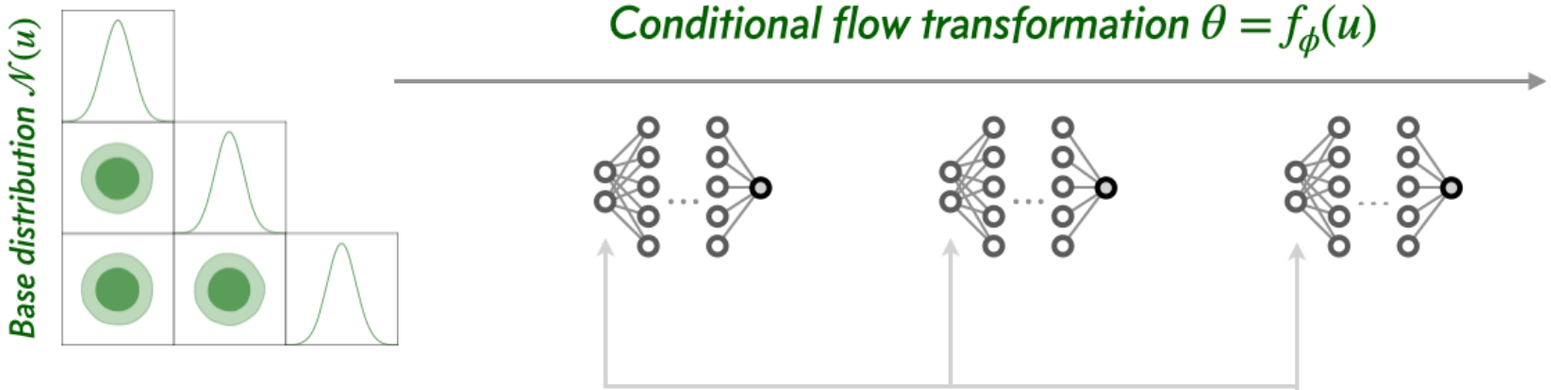
Optimized to be deployed in a real-time system

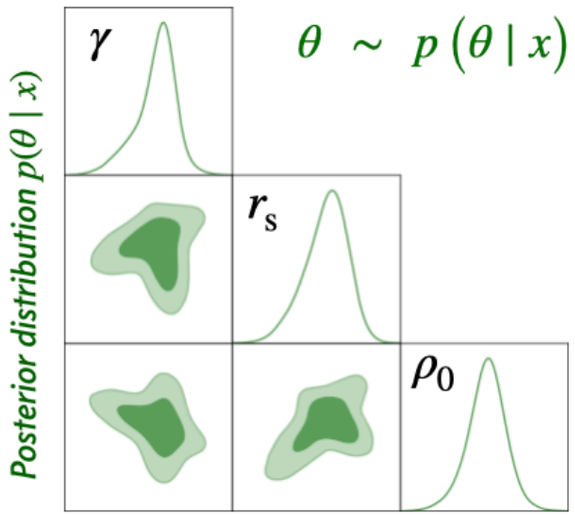
(Not like DINGO)

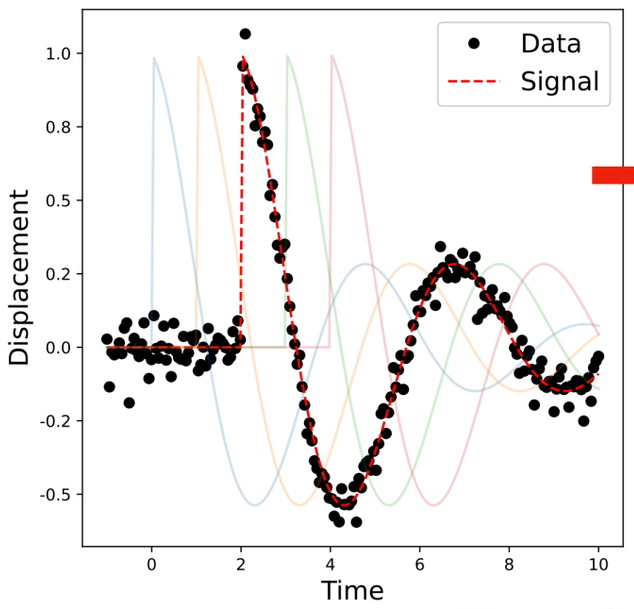
arXiv:2407.19048

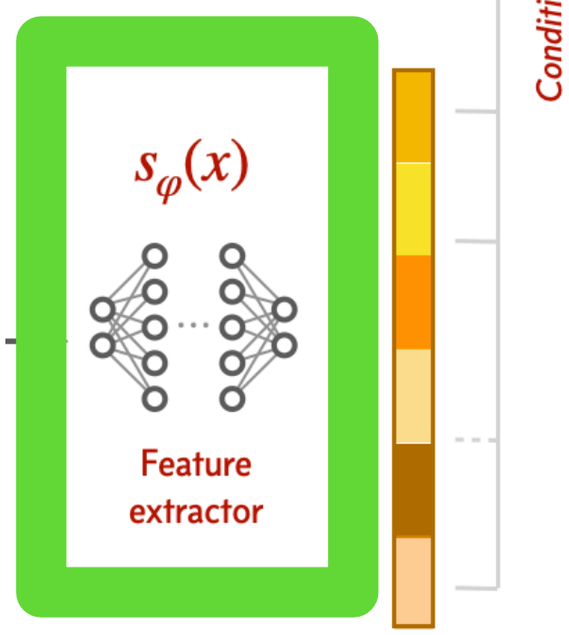
Parameter Estimation

Conditional posterior density estimation







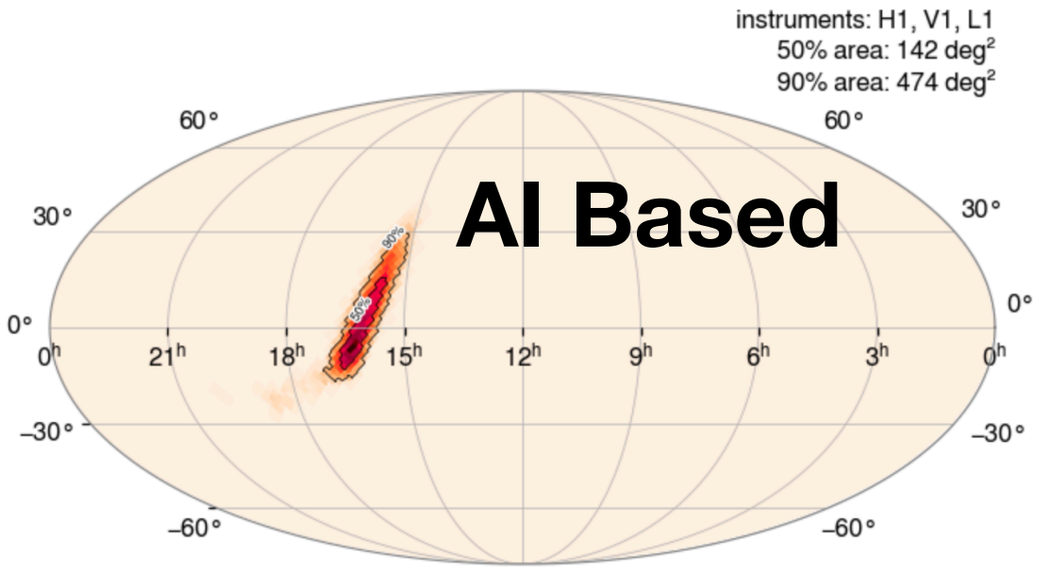


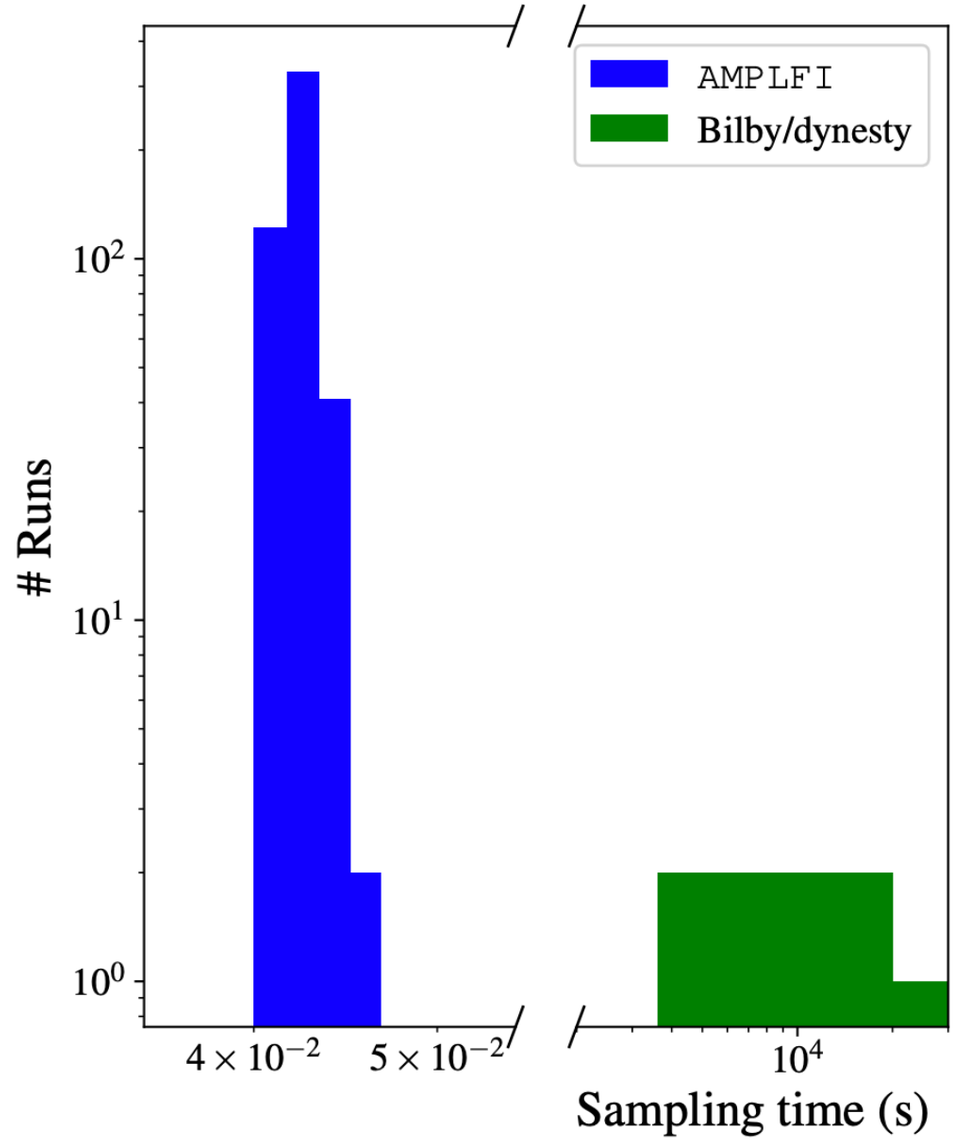
Time Invariant Contrastive Embedding

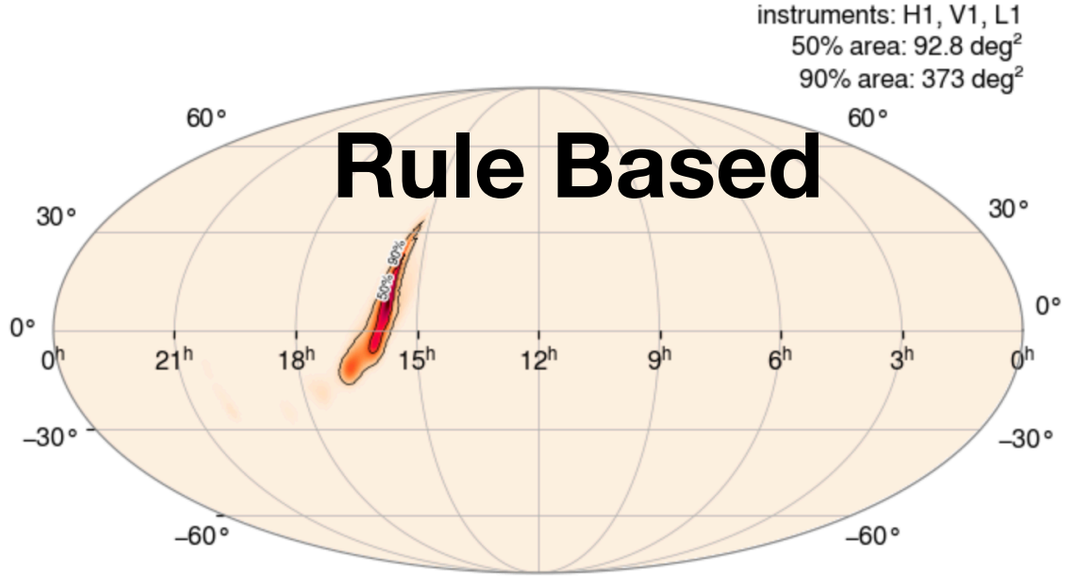
Project away time shift invariance through a constrastive embedding

arXiv:2407.19048

Parameter Estimation





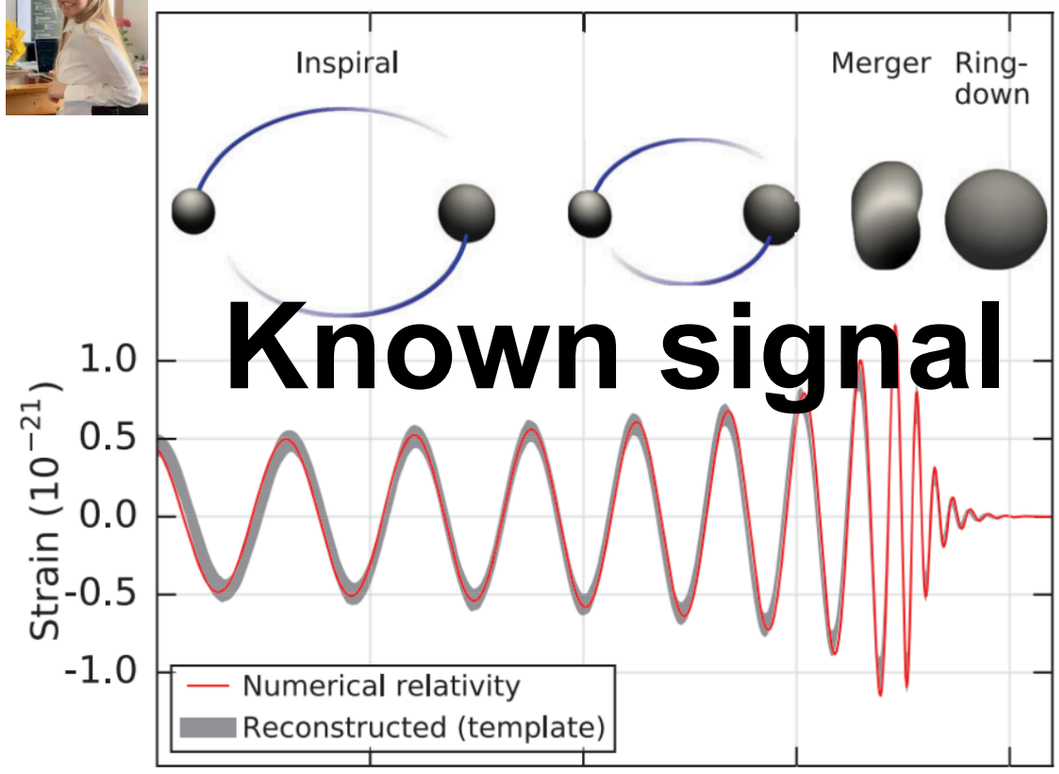


Neural Network Parameter estimation

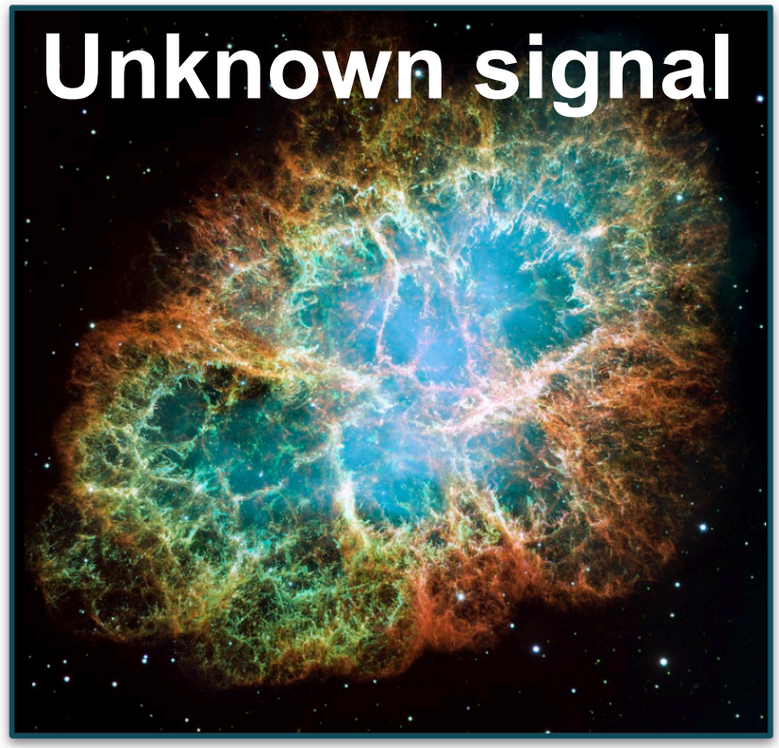
>10,000x times faster than conventional parameter estimation

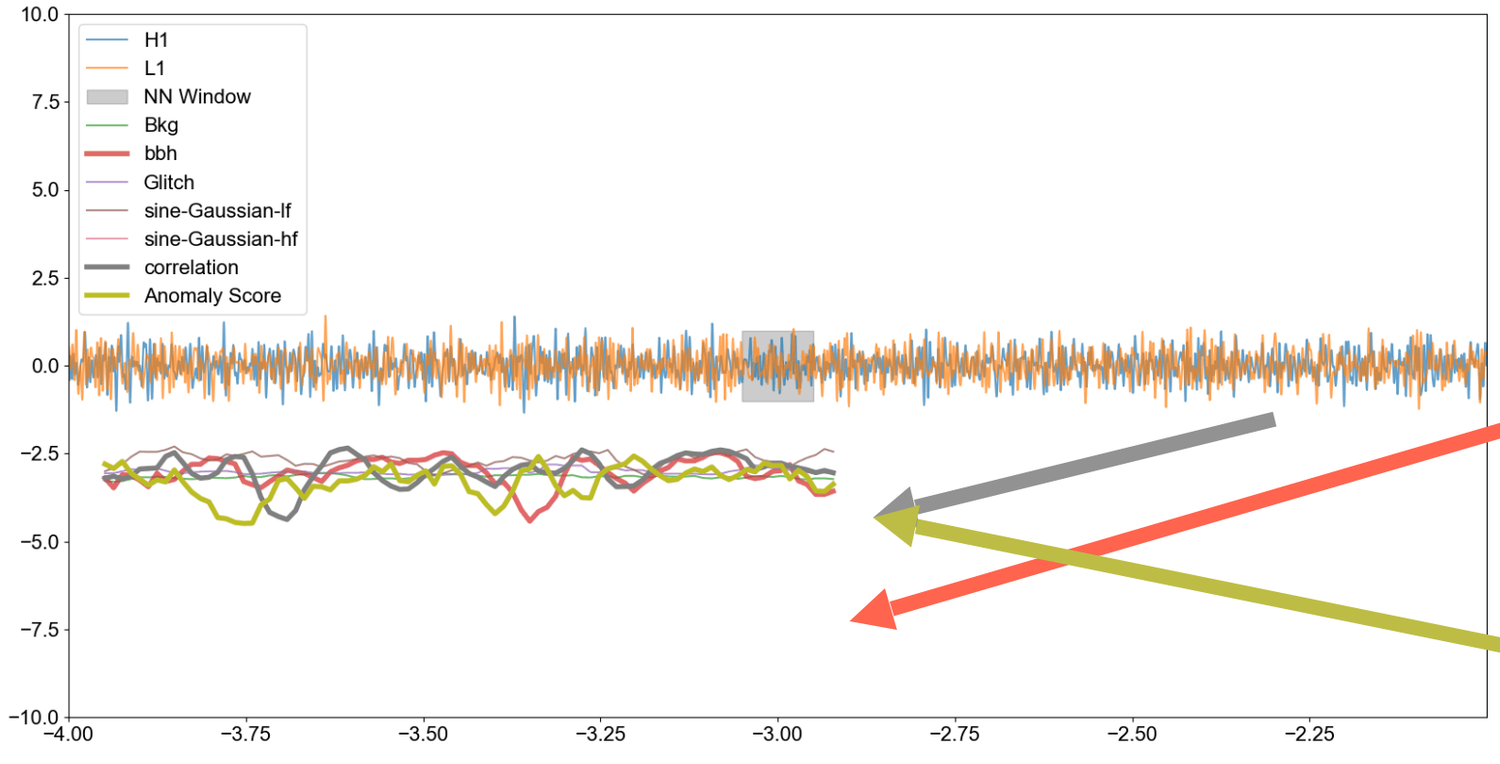
Further optimized larger model takes < 1s (current default)

GWAK:Anomaly Detection



Core-collapse supernova (CCSN)





Key Insights

11D space

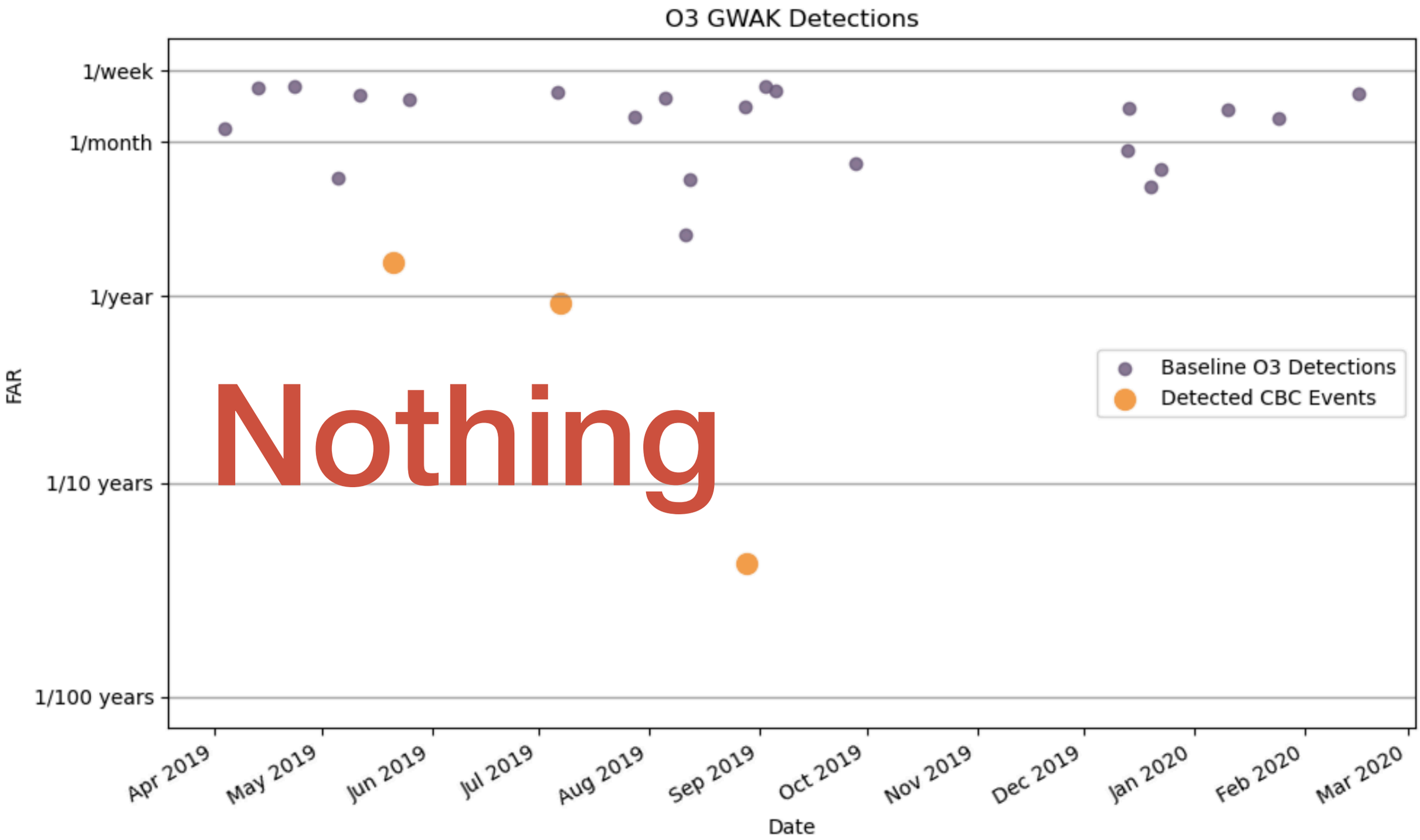
Correlation

BBH Prob

Others

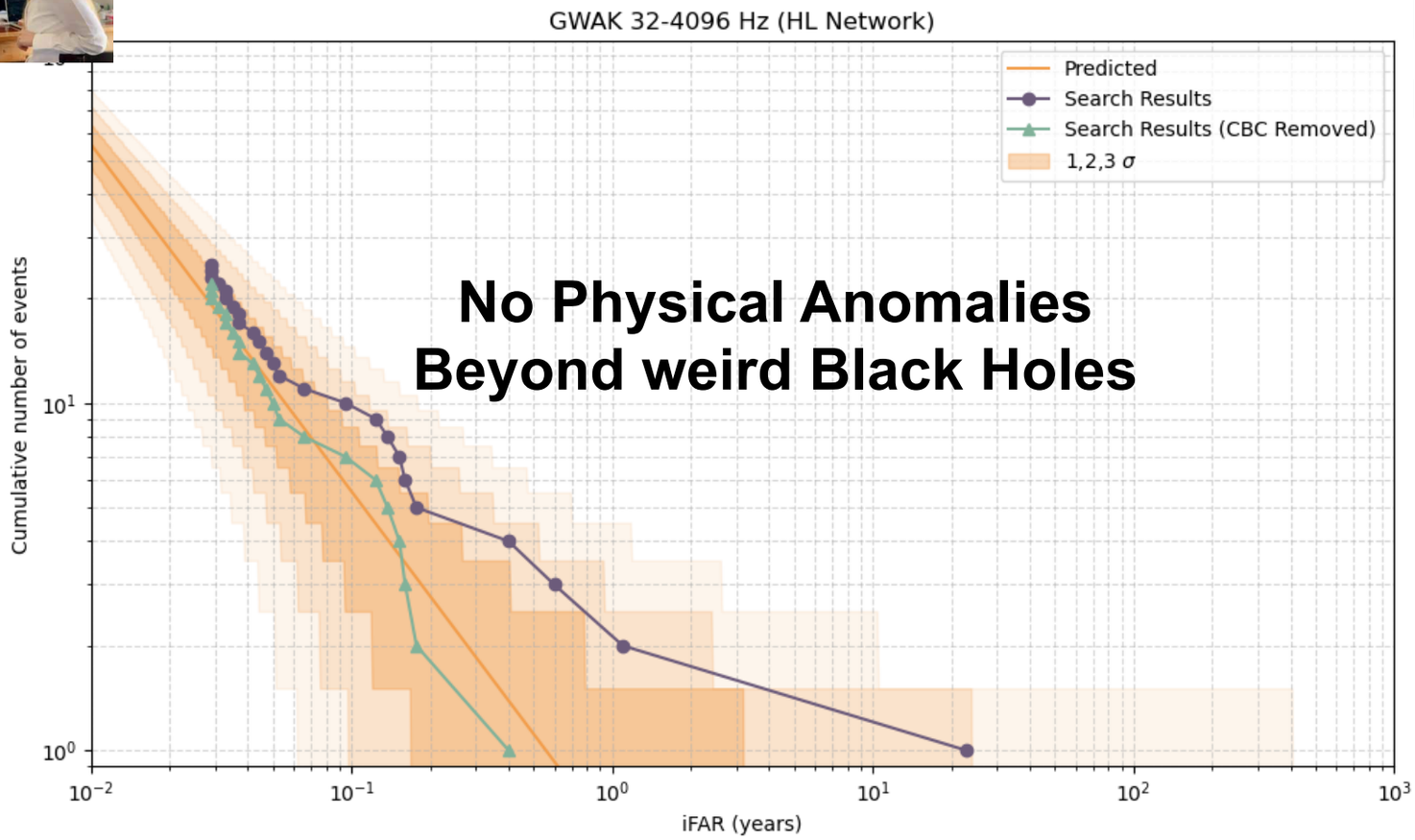
Total Score

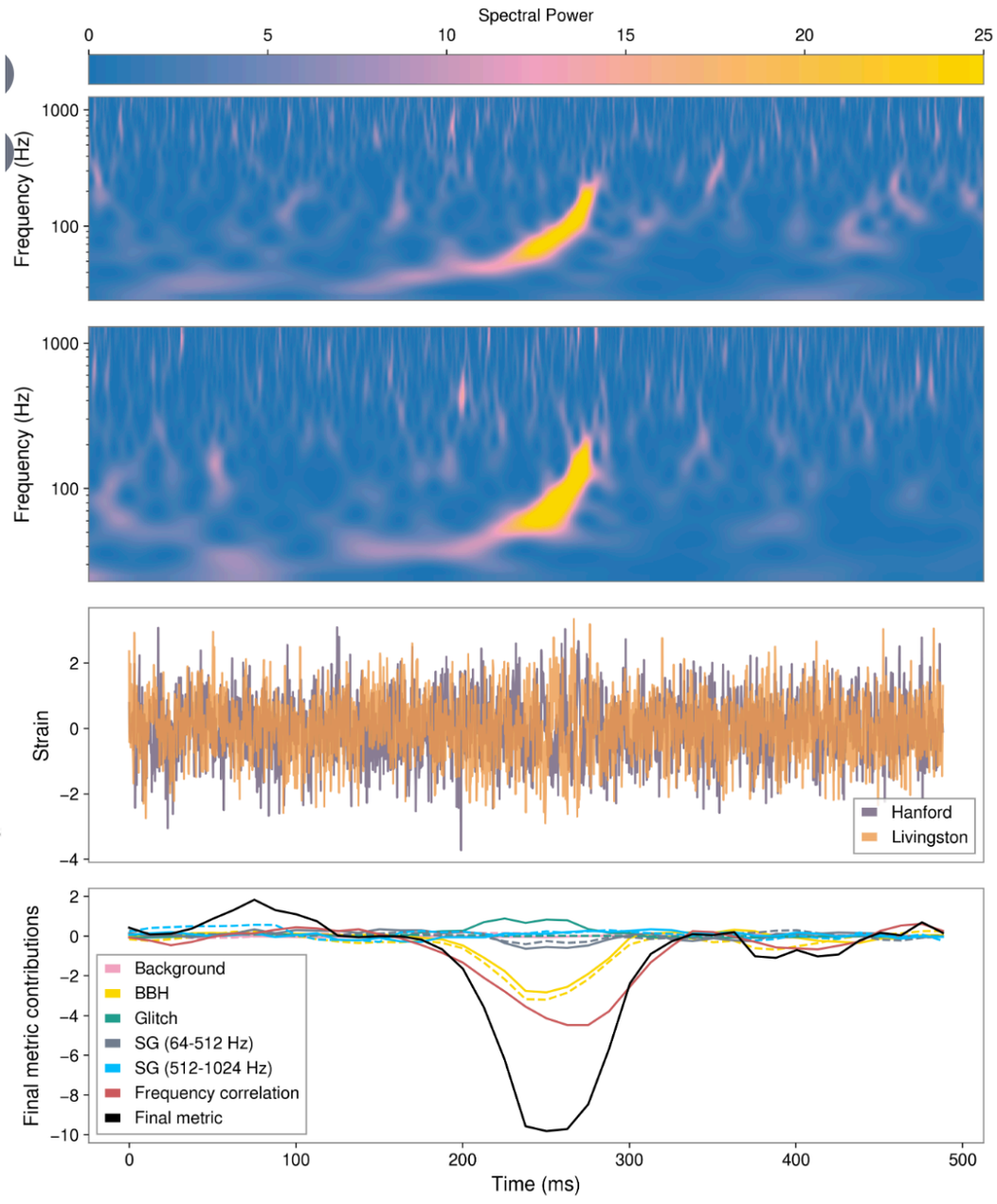
GWAK On Data


O3 GWAK Detections
1/week
1/month
1/year
1/10 years
1/100 years
FAR
Nothing
Baseline O3 Detections
Detected CBC Events
Apr 2019
May 2019
Jun 2019
Jul 2019
Aug 2019
Sep 2019
Oct 2019
Nov 2019
Dec 2019
Jan 2020
Feb 2020
Mar 2020
Date

Arxiv:2412.19883

GWAK On Data



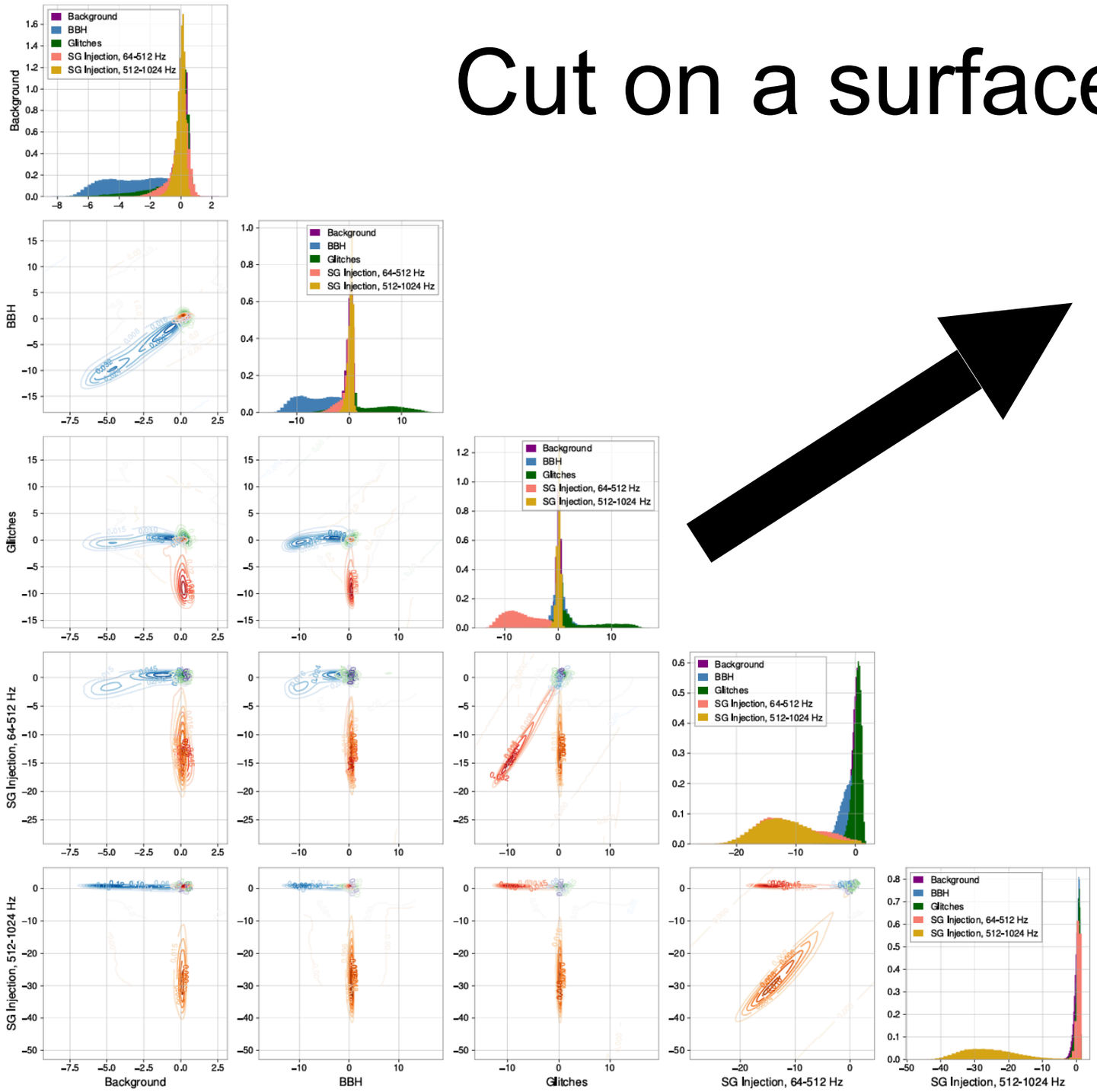


No glaring true anomalies

We found weird black hole mergers + weird glitches

GWAK Space

Cut on a surface in the 5D Space


Background
BBH
Glitches
SG Injection, 64-512 Hz
SG Injection, 512-1024 Hz

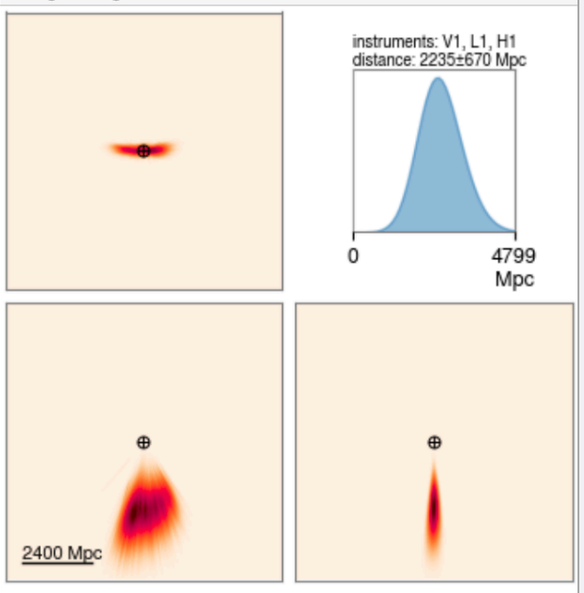
Our System is Live!

Accelerated AI Algorithms for Data-Driven Discovery

Per-Pipeline Event Information

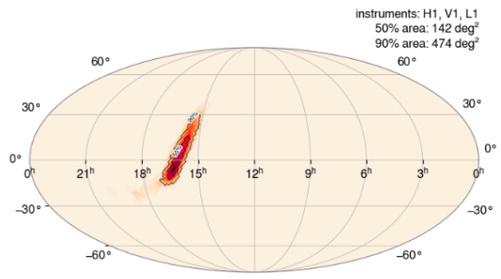
UID	Group	Pipeline	Search	gpstime	FAR (Hz)
G597361	CBC	pycbc	AllSky	1441016546.022	3.168e-10
G597350	CBC	aframe	AllSky	1441016545.998	3.154e-10
G597364	CBC	CWB	BBH	1441016546.020	6.083e-11
			AllSky	1441016545.995	1.164e-14
			AllSky	1441016546.002	1.596e-25

Log Image



Volume rendering of amplfi.multiorder.fits,0
— Submitted by read-cvmfs-emfollow robot on September 4, 2025 10:22:37 UTC

Log Image



Mollweide projection of amplfi.multiorder.fits,0
— Submitted by read-cvmfs-emfollow robot on September 4, 2025 10:22:36 UTC

We are publicly running since August 30th 2025 Delivering robust alerts

Our System is running Live!

GCN Circular 41606

Subject LIGO/Virgo/KAGRA S250830bp: Identification of a GW compact binary merger candidate
Event LIGO/Virgo/KAGRA S250830bp
Date 2025-08-30T10:55:06Z (a month ago)
From Pedro Jeronimo Santos da Silva <pedro.jeronimo@unesp.br>
Via Web form

The LIGO Scientific Collaboration, the Virgo Collaboration, and the KAGRA Collaboration report:

We identified the compact binary merger candidate S250830bp during real-time processing of data from LIGO Hanford Observatory (H1), LIGO Livingston Observatory (L1), and Virgo Observatory (V1) at 2025-08-30 10:24:18.852 UTC (GPS time: 1440584676.852). The candidate was found by the Aframe [1] cWB [2], cWB BBH [3], GstLAL [4], MBTA [5], MLy [6], PyCBC Live [7], and SPIIR [8] analysis pipelines.

S250830bp is an event of interest because its false alarm rate, as estimated by the online analysis, is 3.2e-10 Hz, or about one in 1e2 years. The event's properties can be found at this URL:

https://gracedb.ligo.org/superevents/S250830bp

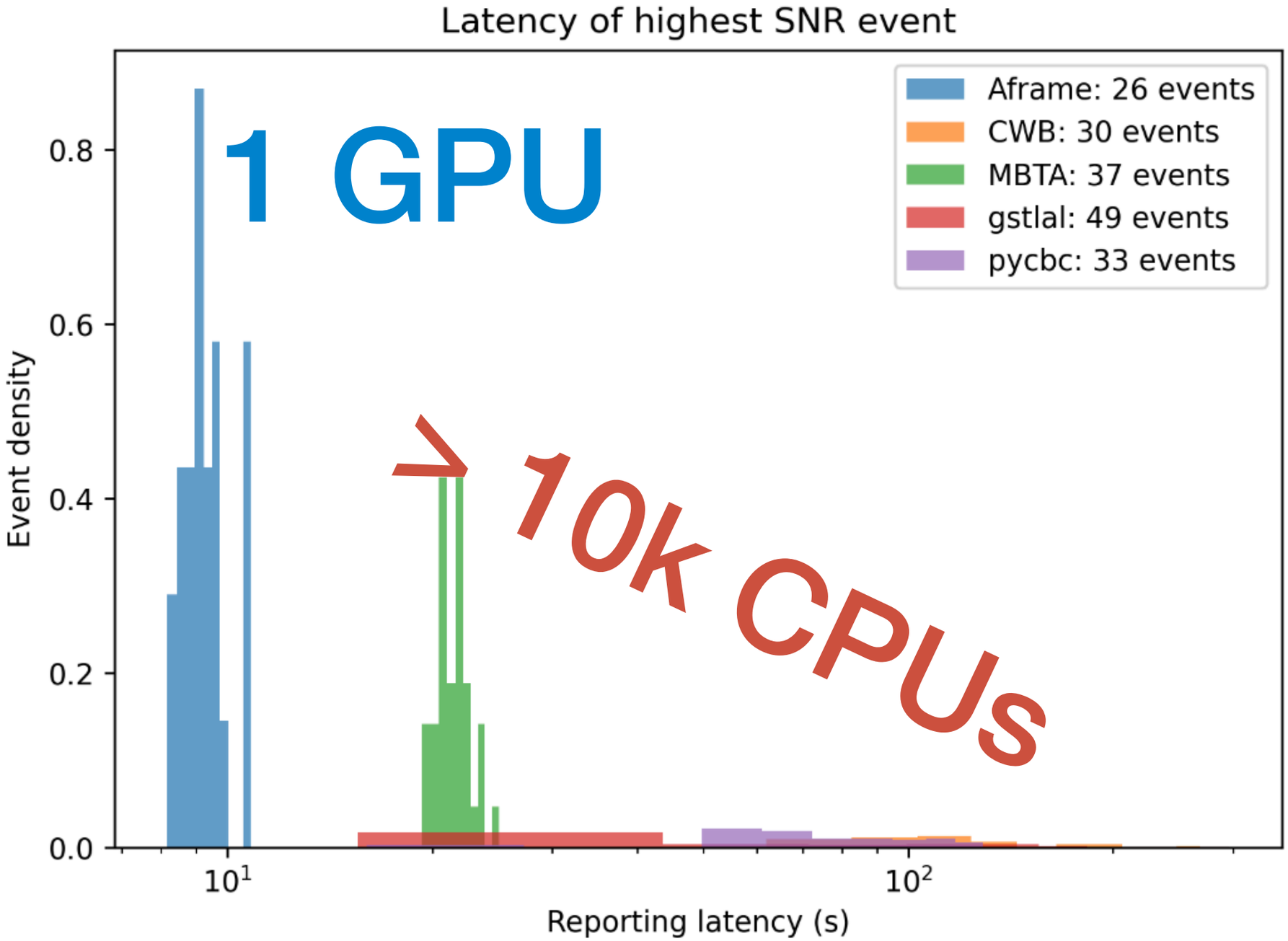
The classification of the GW signal, in order of descending probability, is BBH (84%), NSBH (16%), Terrestrial (<1%), or BNS (<1%).

Two sky maps are available at this time and can be retrieved from the GraceDB event page:
* amplfi.multiorder.fits,0, an initial localization generated by AMPLFI [10] distributed via GCN and SCiMMA notices about 24 seconds after the candidate event time.
* bayestar.multiorder.fits,0, an initial localization generated by BAYESTAR [11], distributed via GCN and SCiMMA notices about 5 minutes after the candidate event time.

Even with a system not designed for speed we are 10x faster

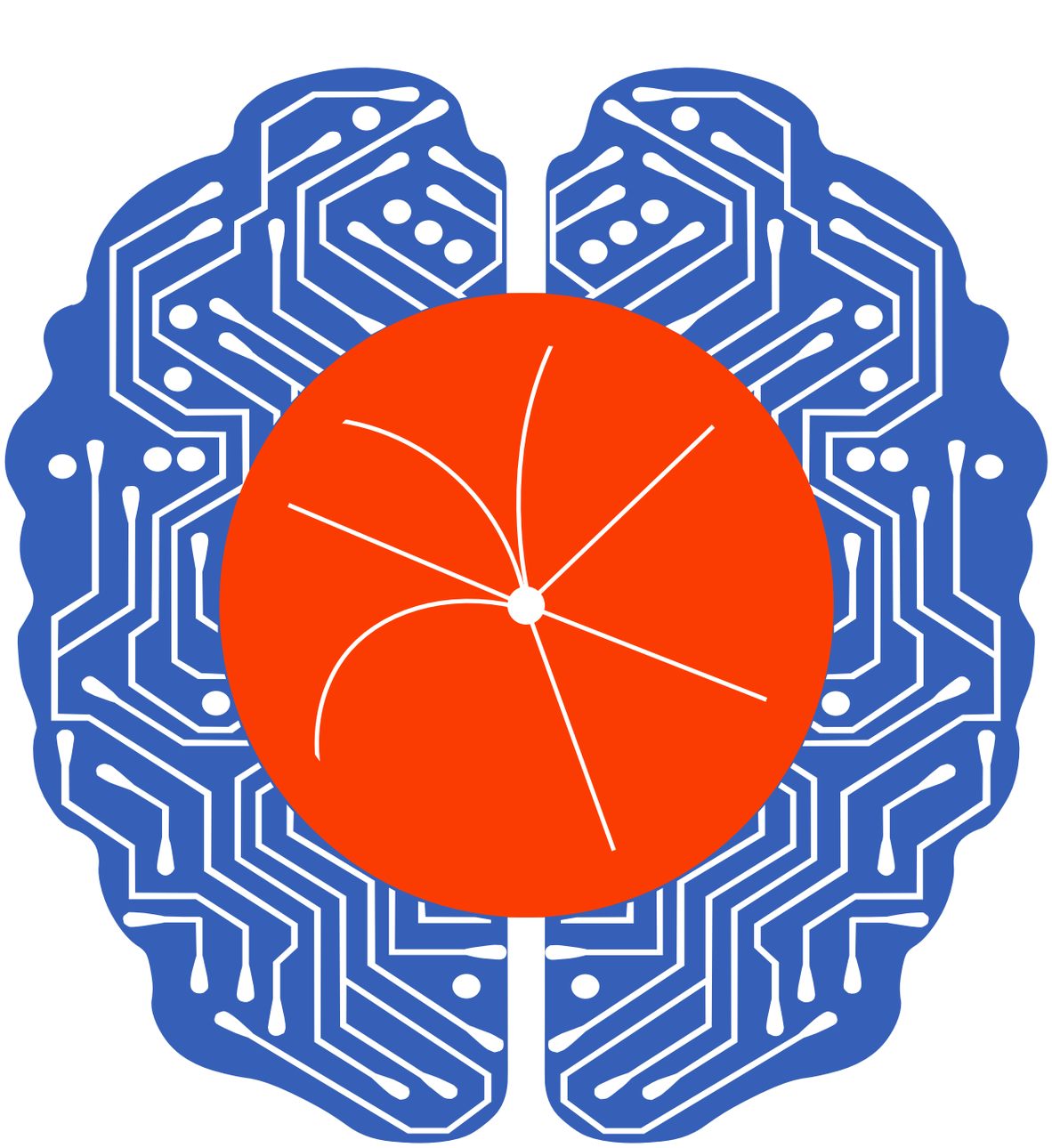
https://gcn.nasa.gov/circulars/41606?query=S250830bp

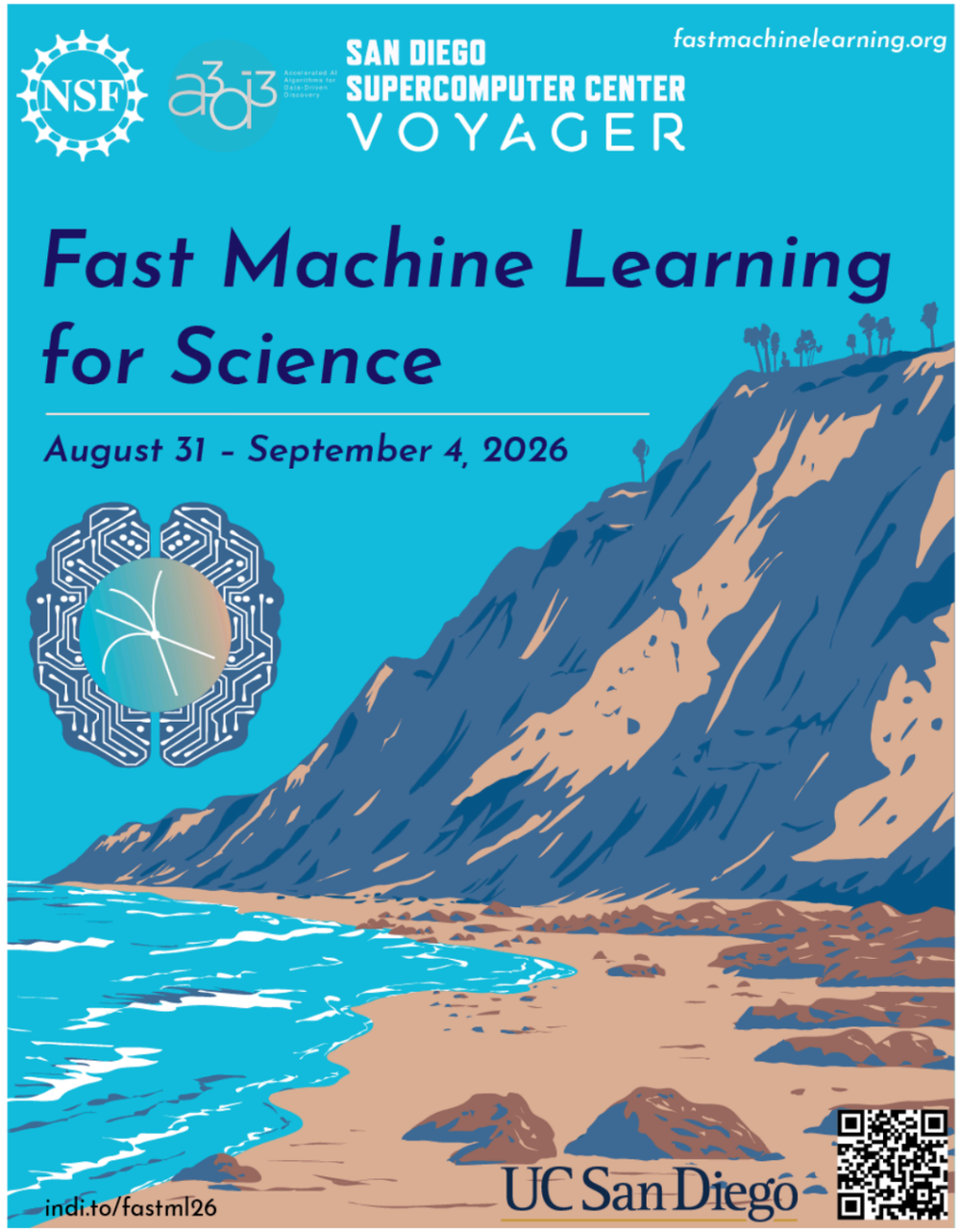
Impact Already



Our pipeline found all the large signal events
Our pipeline is 10-100x faster than others

Come and Join FastML !



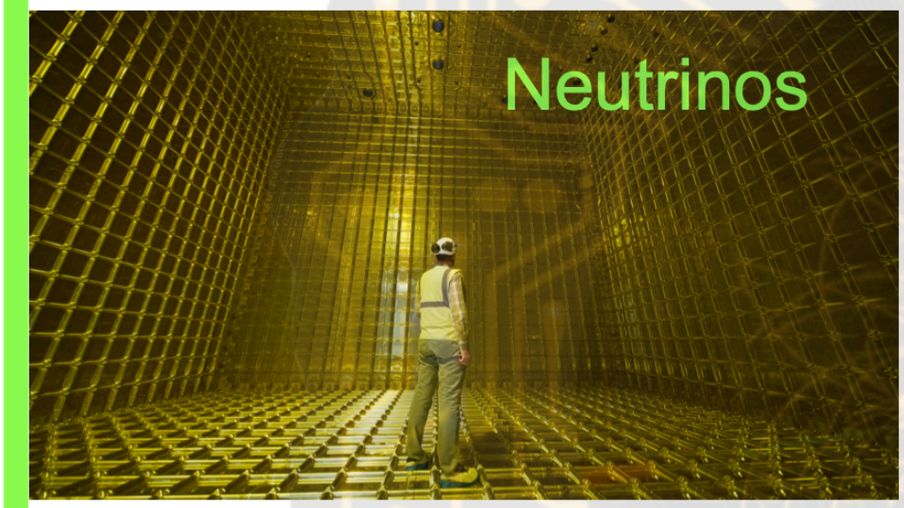


Opening doors for real-time AI in HEP and many other fields

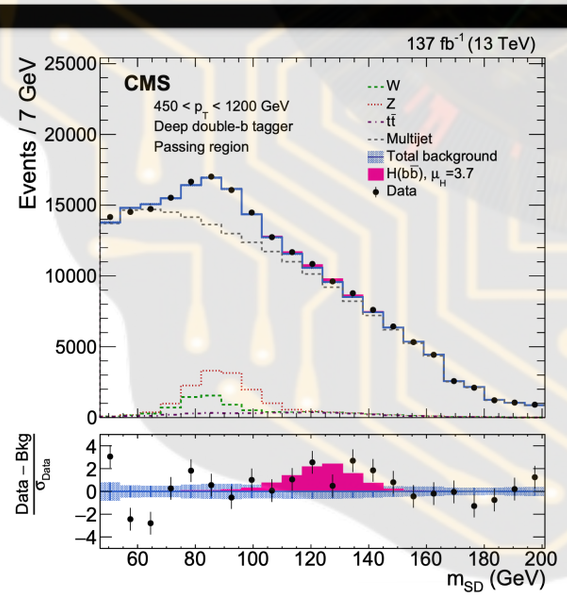
An Institute: **A3D3**

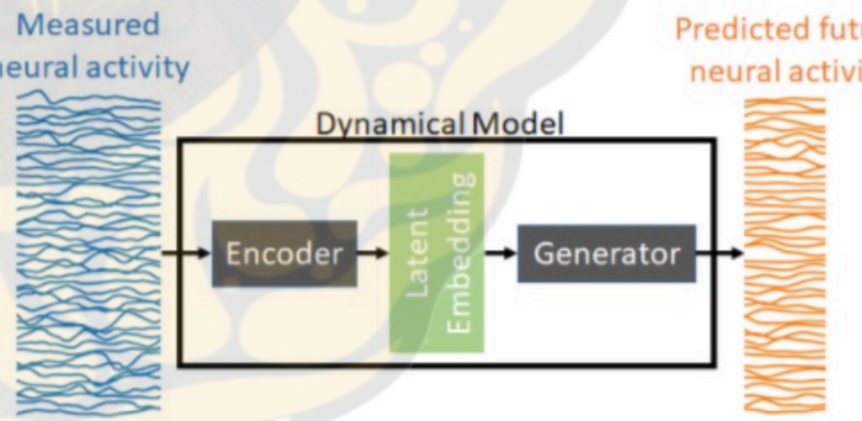
- We have been awarded a new institute to explore real-time AI
  - Accelerated AI Algorithms for Data Driven Discovery (A3D3)

New Types of Computing



LHC
Physics





NeuroScience

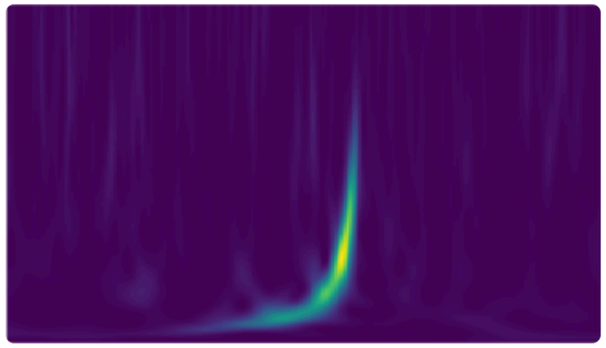
nsfhdr.org/mlchallenge

NSF HDR ML Challenge

‹ Catalog / Course

Computational Data Science in Physics I

Explore realistic, contemporary examples of how computational methods apply to physics research. In this first module, you will analyze LIGO data, detect a gravitational wave signal, and fit this signal within a physical model, among other objectives, using Jupyter notebooks.

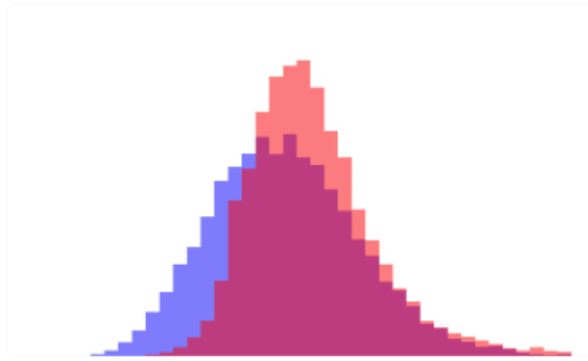


Enroll Now

‹ Catalog / Course

Computational Data Science in Physics II

This course provides realistic, contemporary examples of how computational methods apply to physics research. Topics include hypothesis testing, semi-parameteric methods, and deep learning. In the Final Project, you will analyze LHC data to measure properties of the W boson and Z boson.



Access Course Materials

About | What you'll learn | Prerequisites

This course is no longer active, but you can still access selected content. Learn More

Computational Data Science in Physics III

This course provides realistic, contemporary examples of how computational methods apply to physics research. In this third module, learners will learn about how to simulate physics processes. This will culminate in simulations of lattice models including the Ising model, with relevance to contemporary lattice QCD.

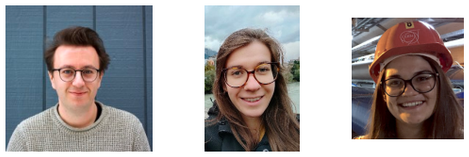
Access Course Materials

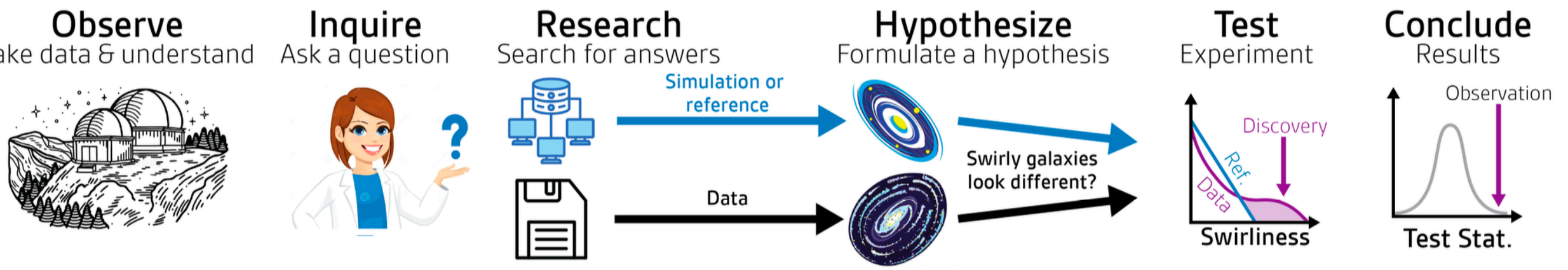
About | What you'll learn | Prerequisites | Instructors

No sessions of this course are currently open for enrollment. More sessions may be added in

MIT 8.316 8.316

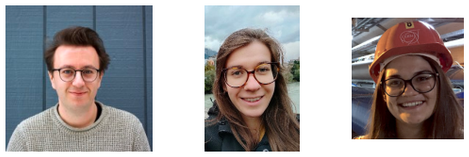
Arxiv:2510.21935
Neruips 2025





The Scientific Method

Arxiv:2510.21935
Neruips 2025



Observe
Take data & understand

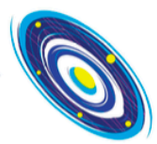
Inquire
Ask a question

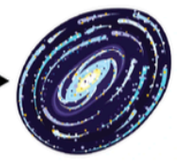
Research
Search for answers

Simulation or reference

Data

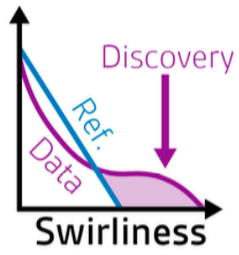
Hypothesize
Formulate a hypothesis



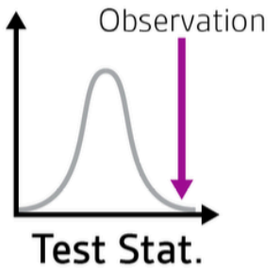


Swirly galaxies look different?

Test
Experiment



Conclude
Results



The Scientific Method Can be automated

Conclusions

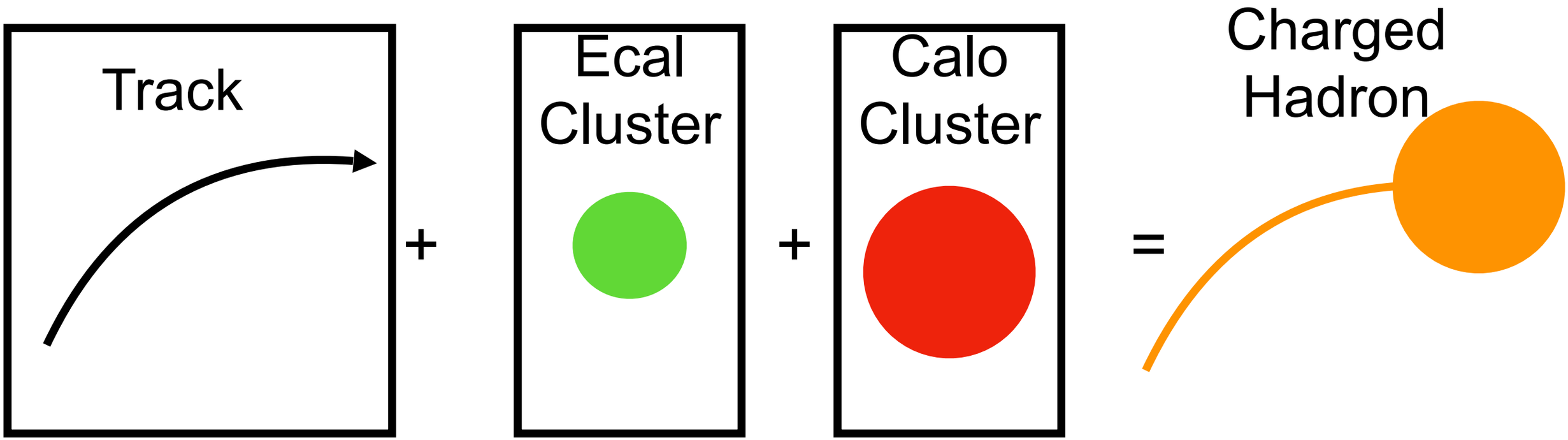
- Things are changing very quickly at all levels
  - We are getting more out of our data and automating a lot
    - However, we need to think about compute
  - We can now do truly optimal measurements
    - Mixing data & ML and uncertainties & Inductive Bias
- Things are quickly become automatic
  - We can think bigger and bolder about how we do research
    - AI can allow us to be more thorough and detailed
    - Sometimes this means just generating code
- Bar for AI is getting lower! You too! Can use AI in your research!

Team Effort

Phil Harris Lab@IIIT

Thanks!

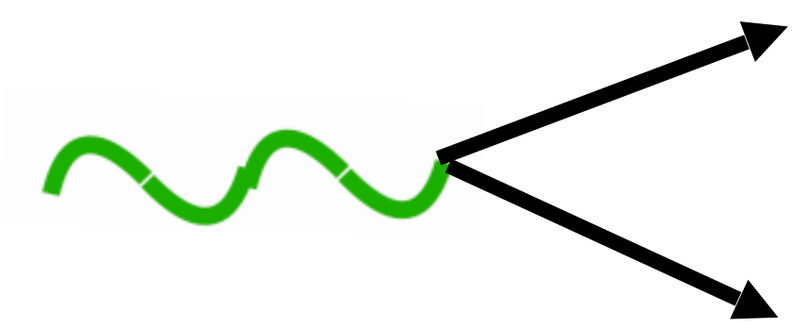
What Are Particles



Particle are linked detector elements

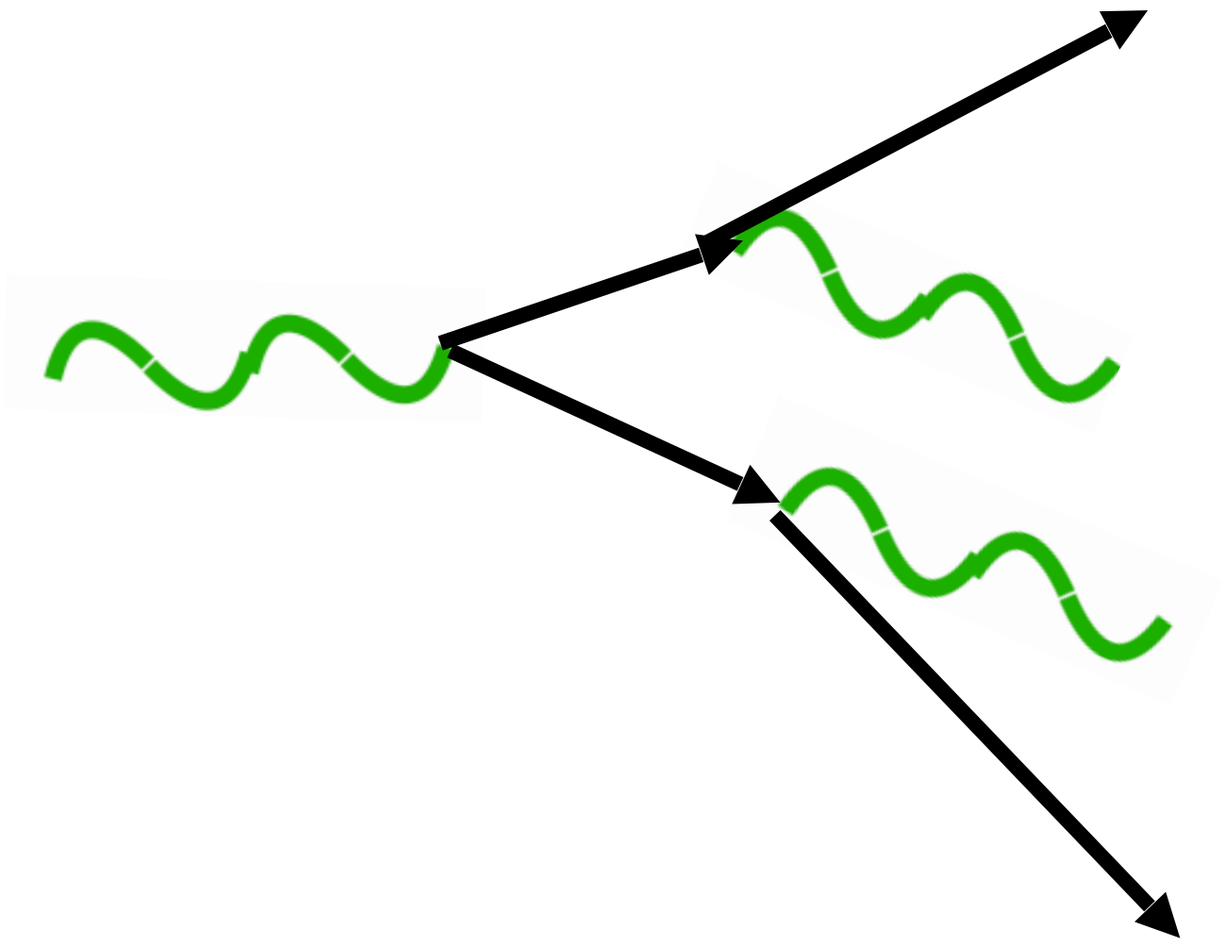
How do we use them?

A Top Down View



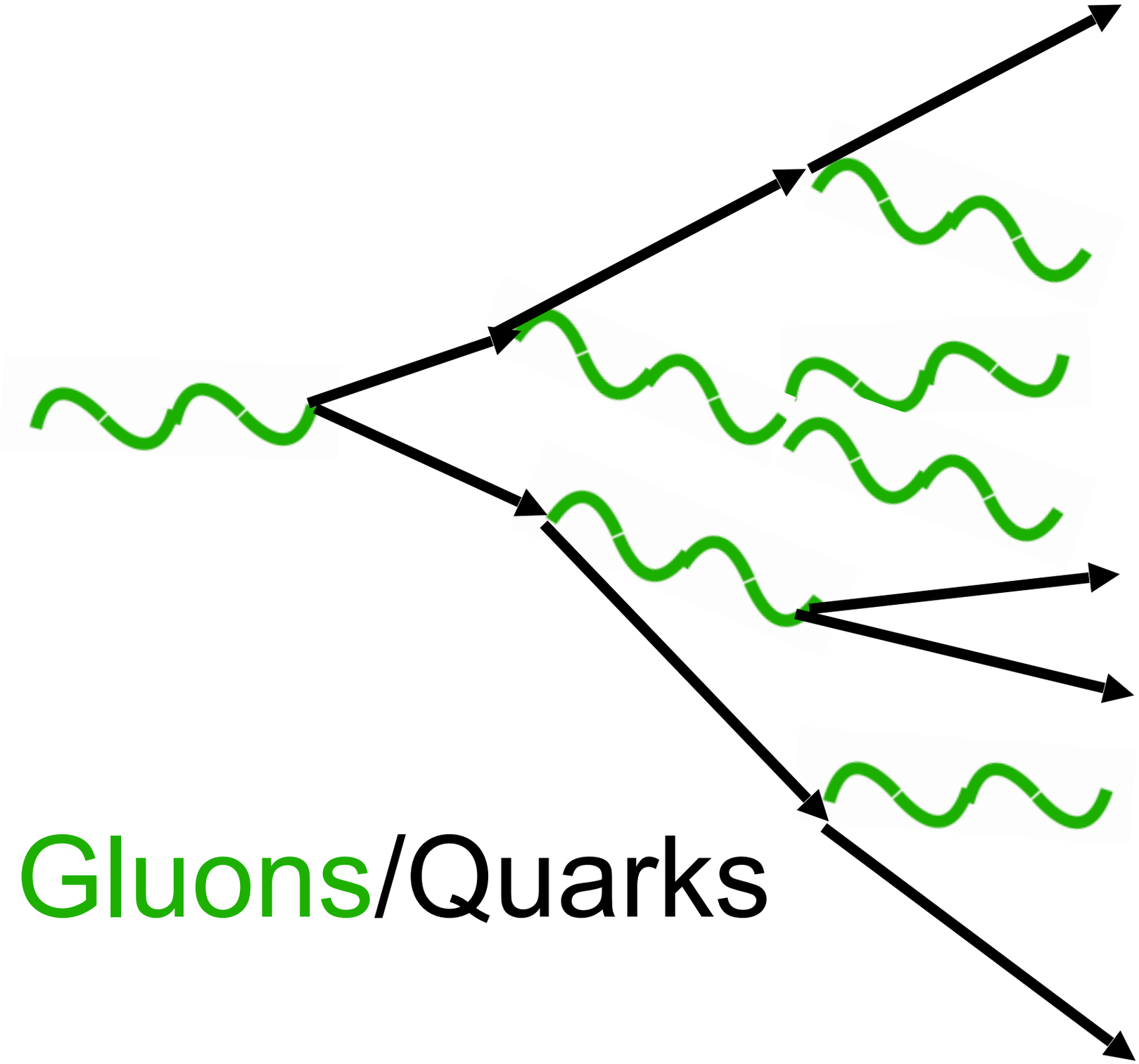
Gluons/Quarks

How Do we do this?



Gluons/Quarks

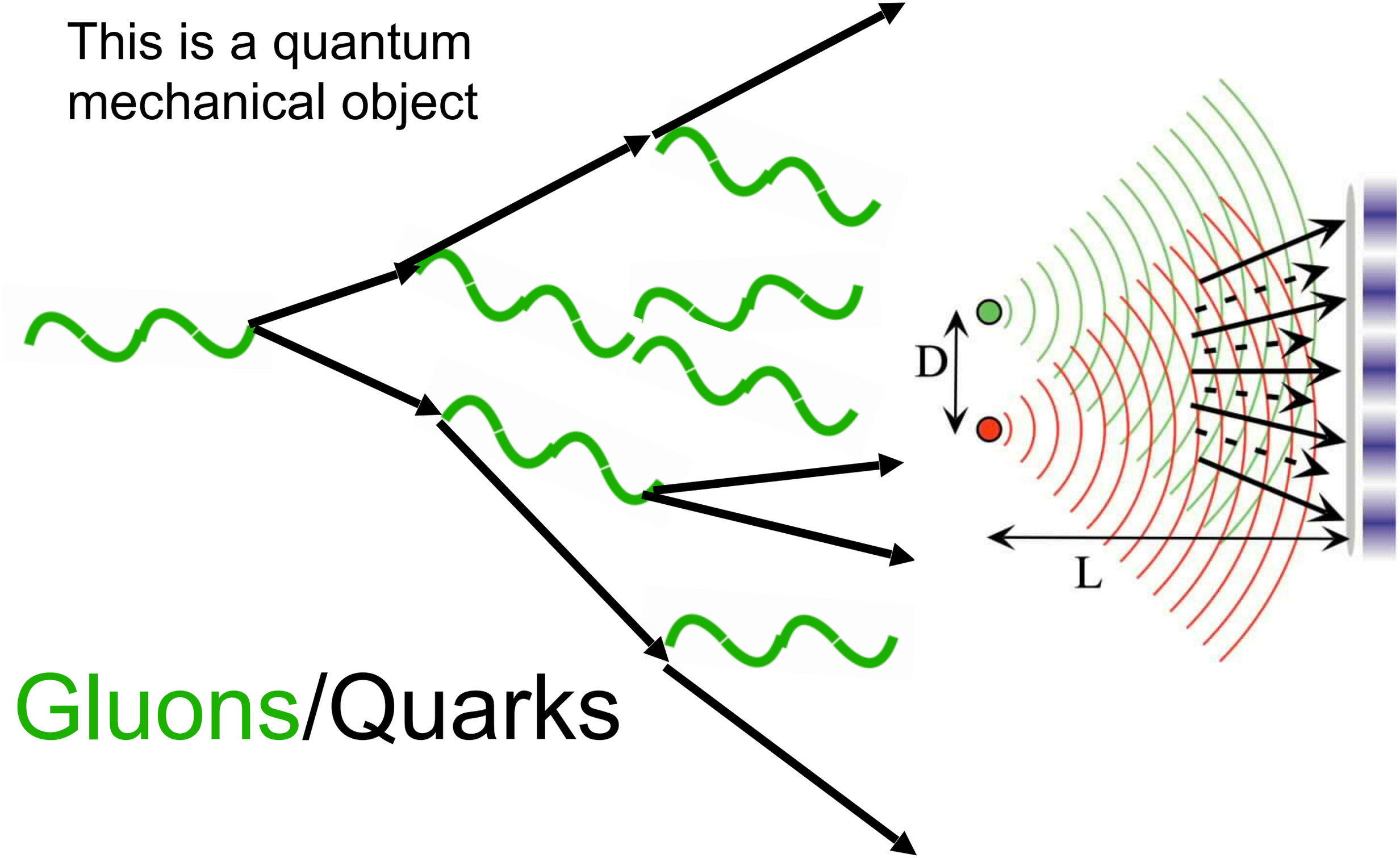
How Do we do this?



Gluons/Quarks

How Do we do this?

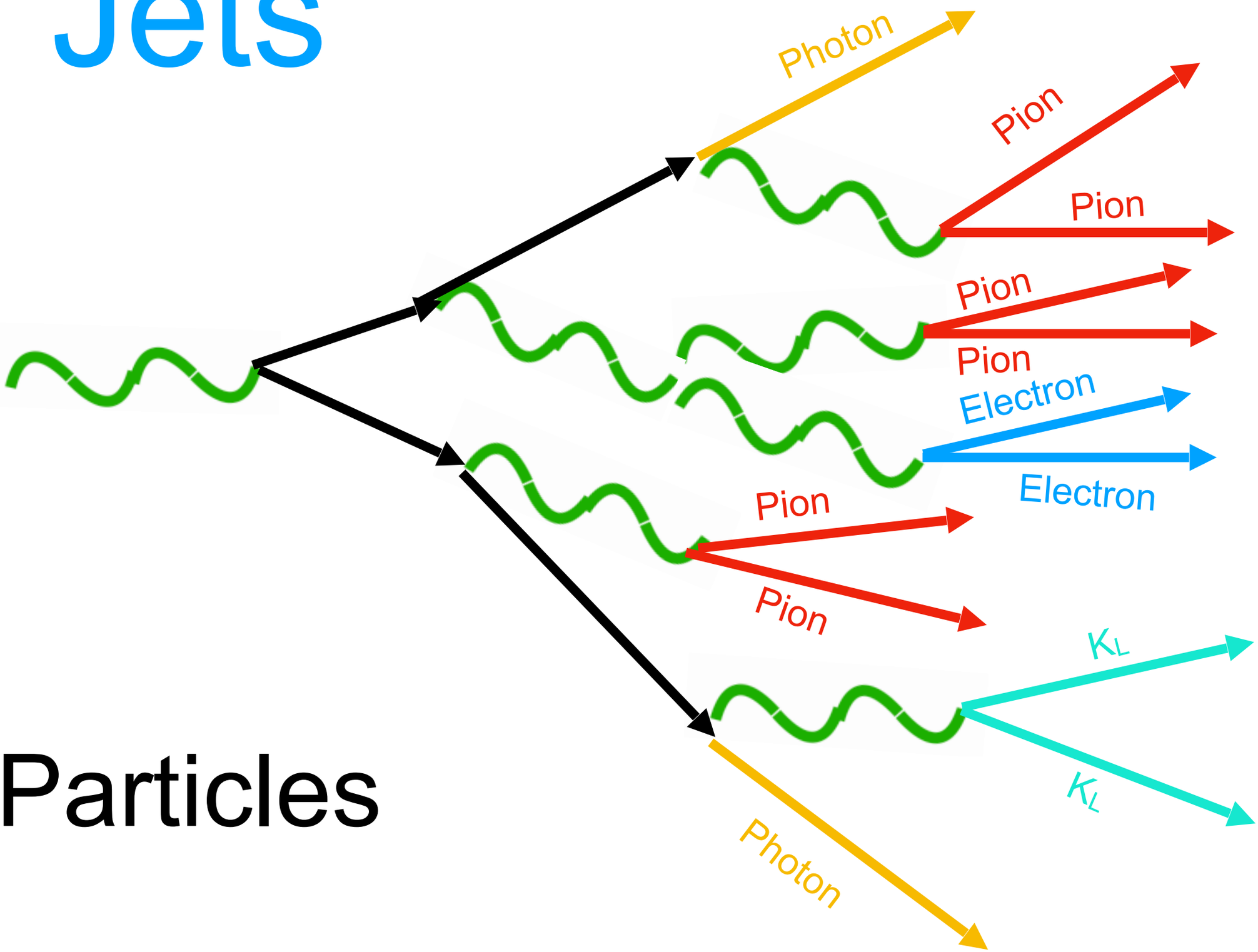
This is a quantum mechanical object



Gluons/Quarks

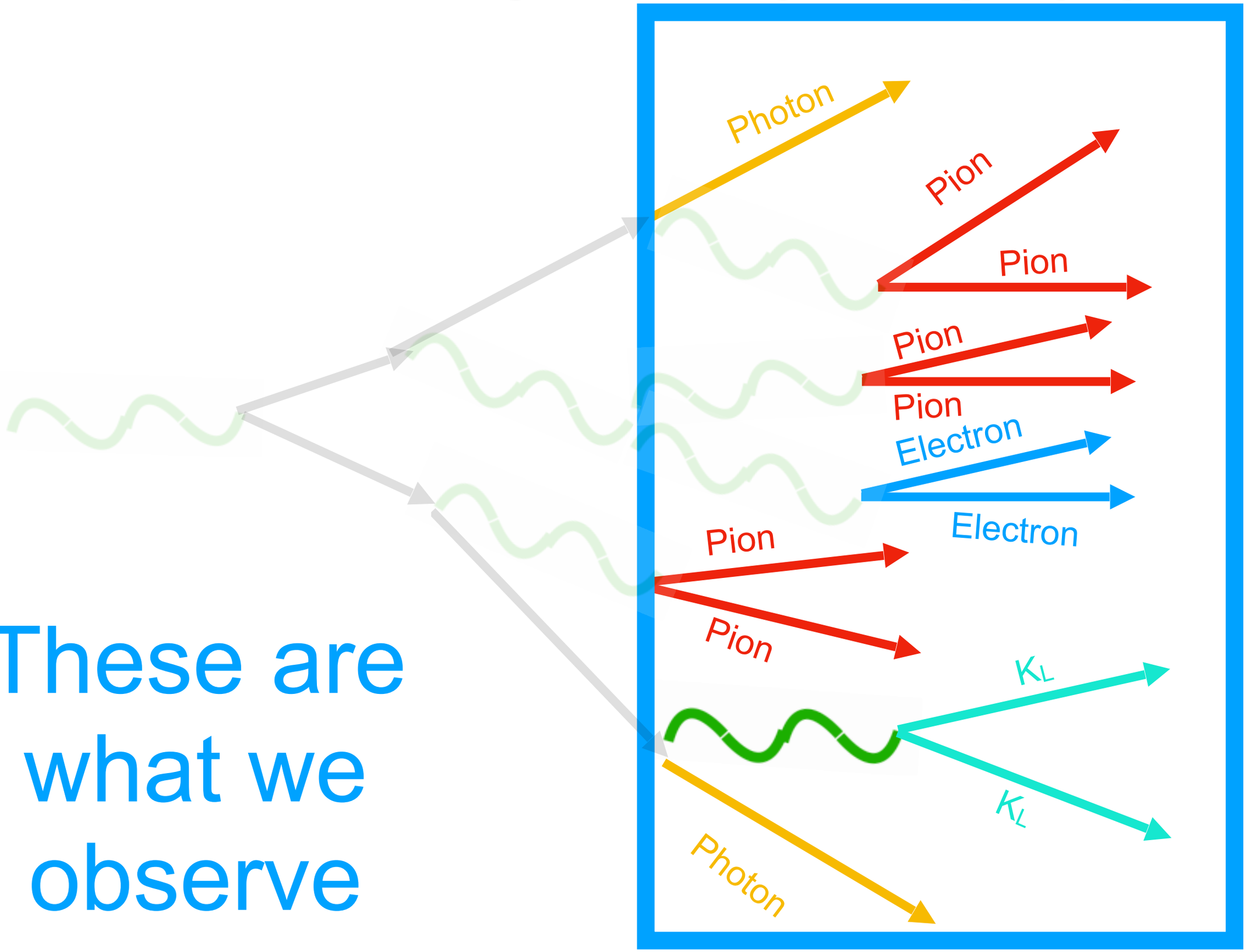
How Do we do this?

Jets



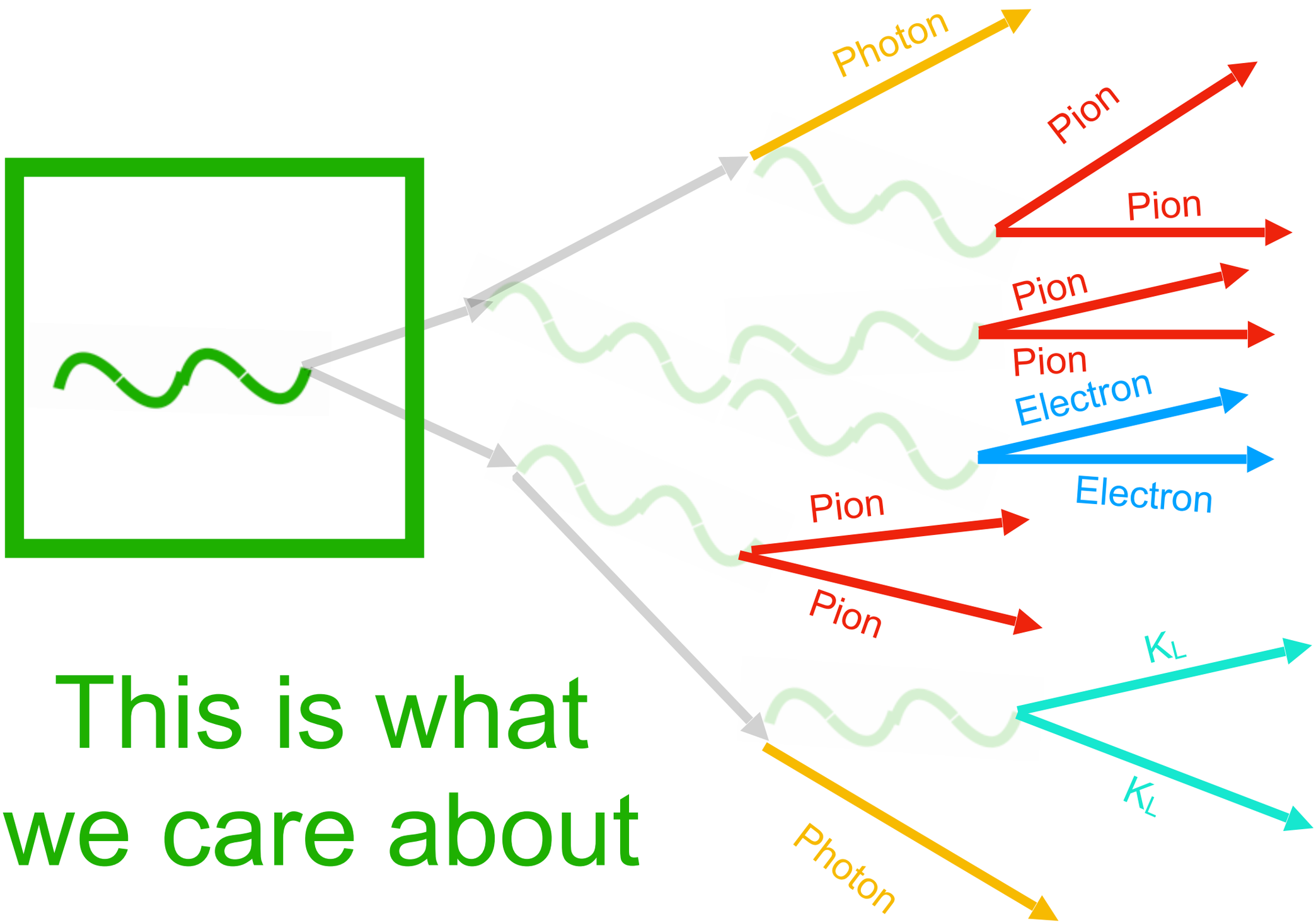
Particles

How Do we do this?



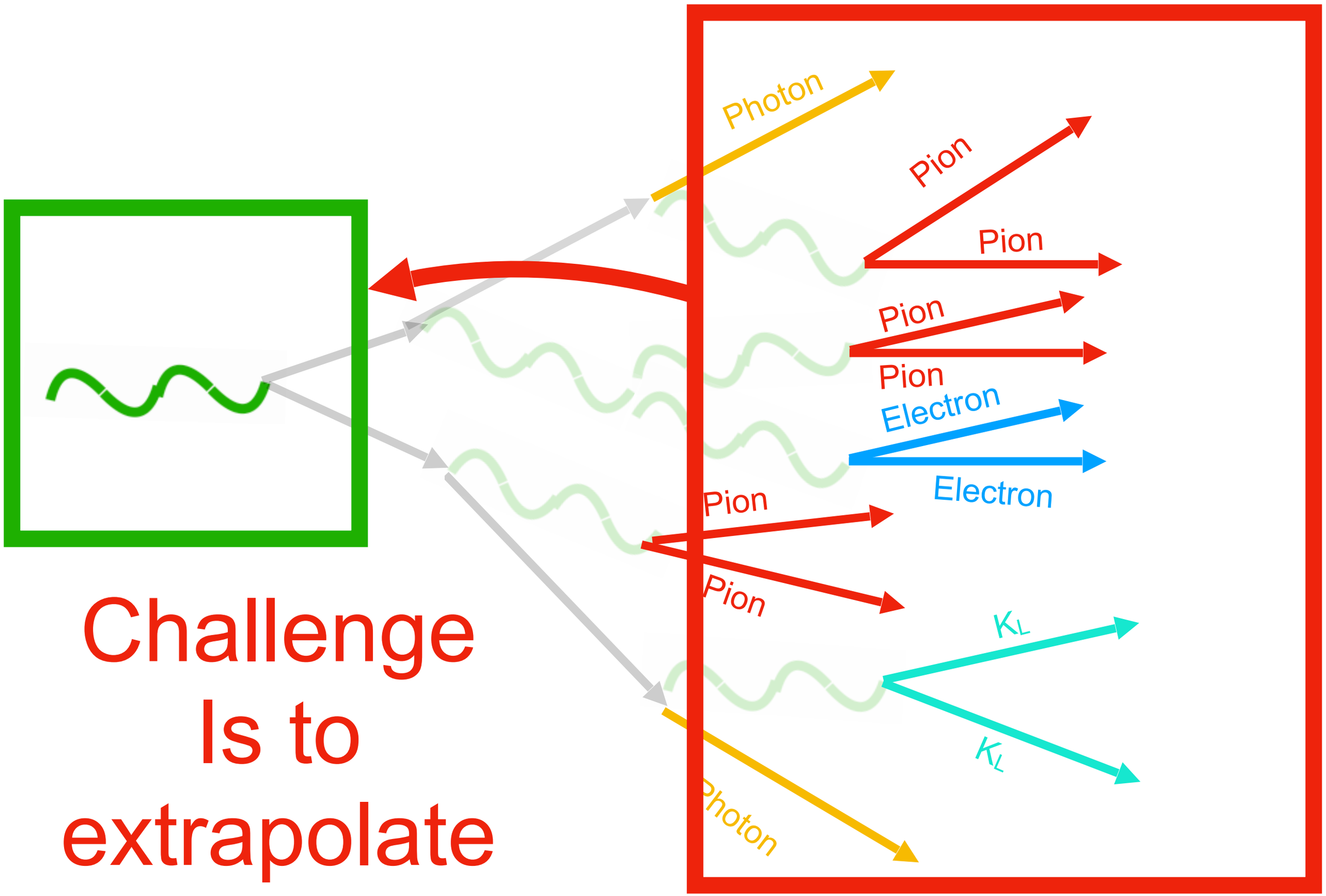
These are what we observe

How Do we do this?



This is what we care about

How Do we do this?



Challenge
Is to
extrapolate

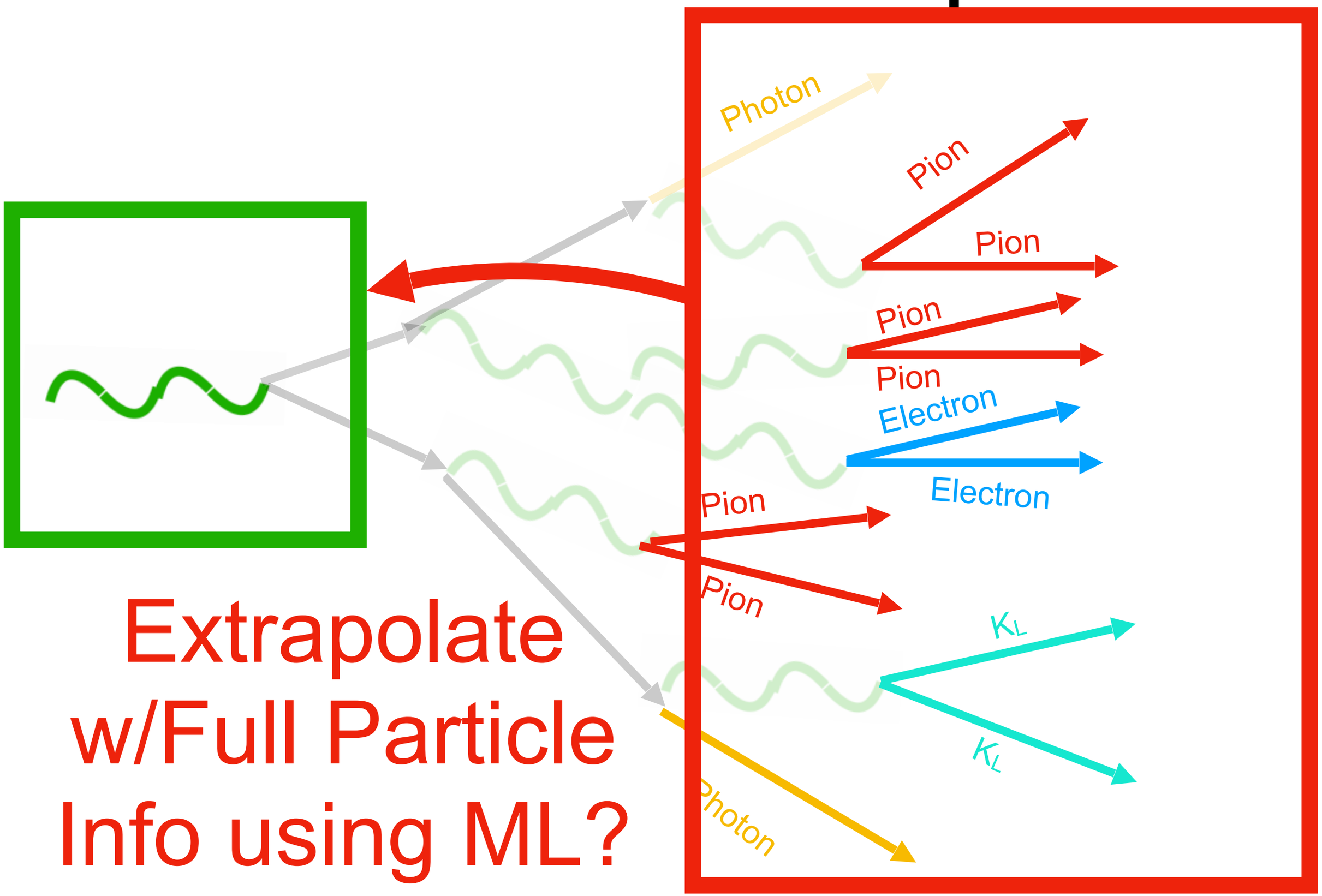
Simple Way

Vector in
one side

Vector in
other side

Extrapolate
w/few Summary
Variables

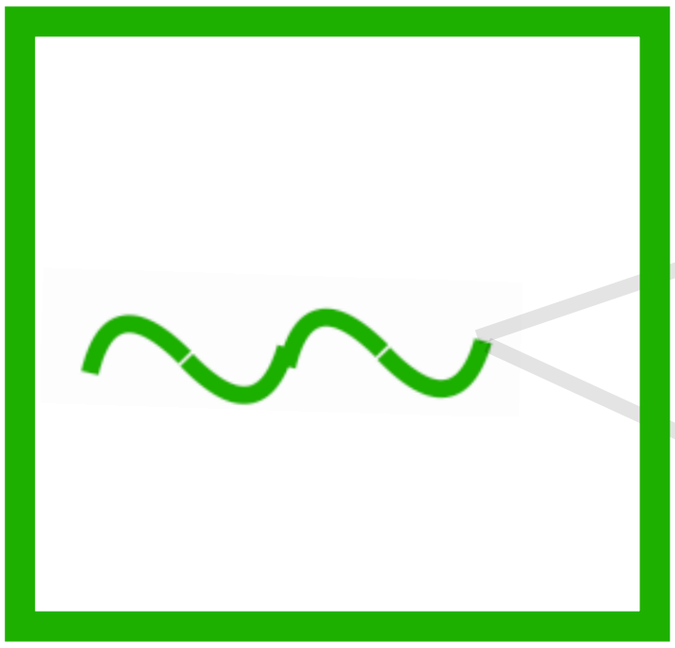
Complex Way



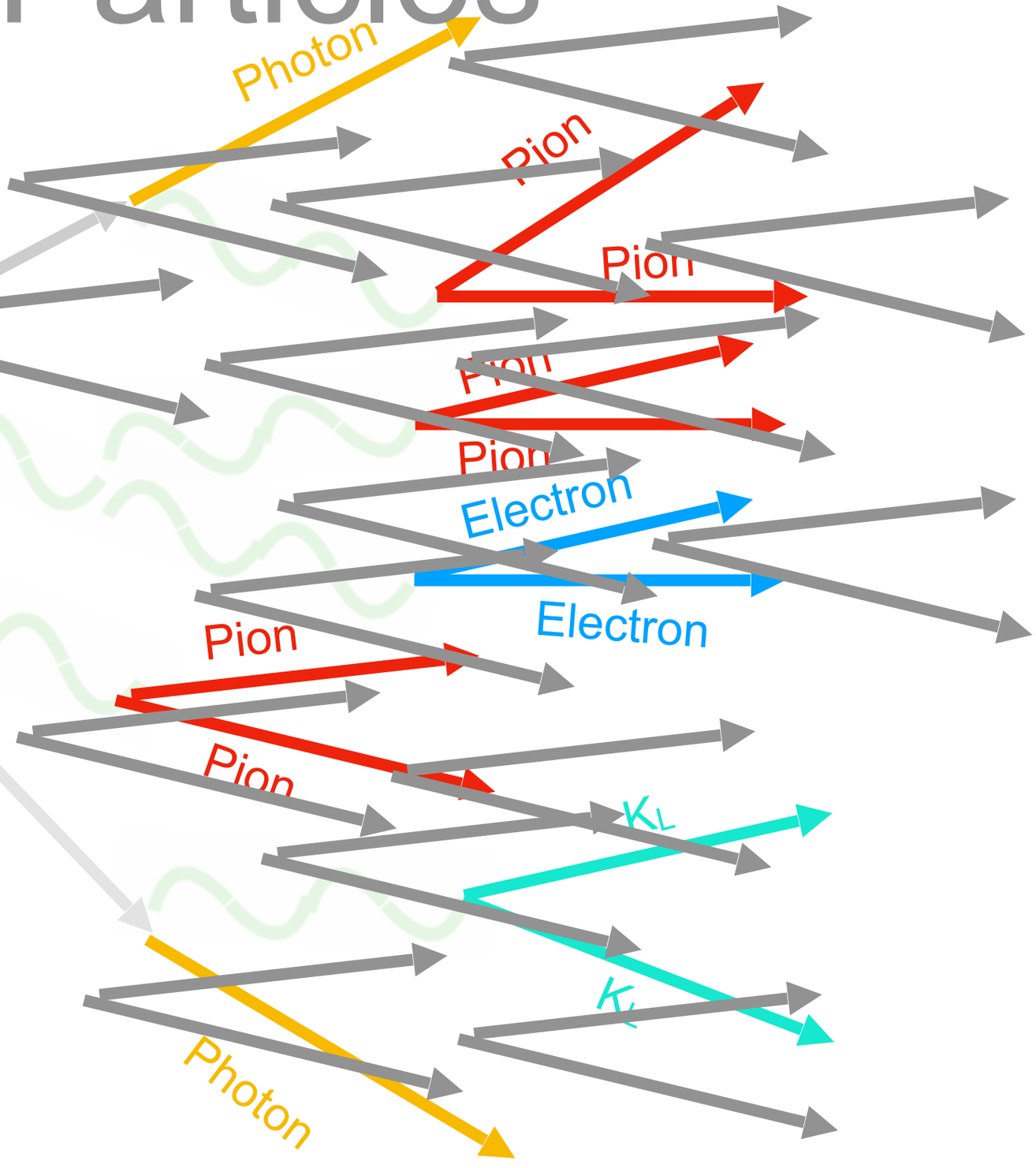
Extrapolate w/Full Particle Info using ML?

Other Challenges

Additional Particles

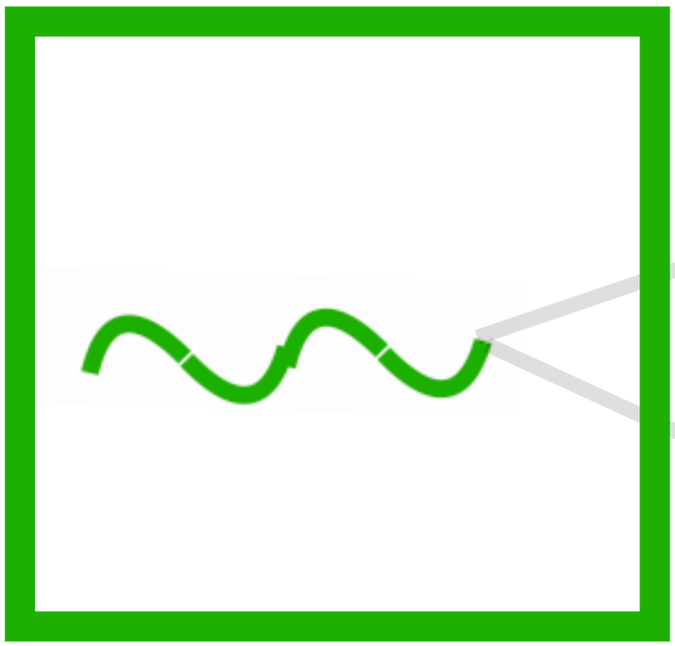


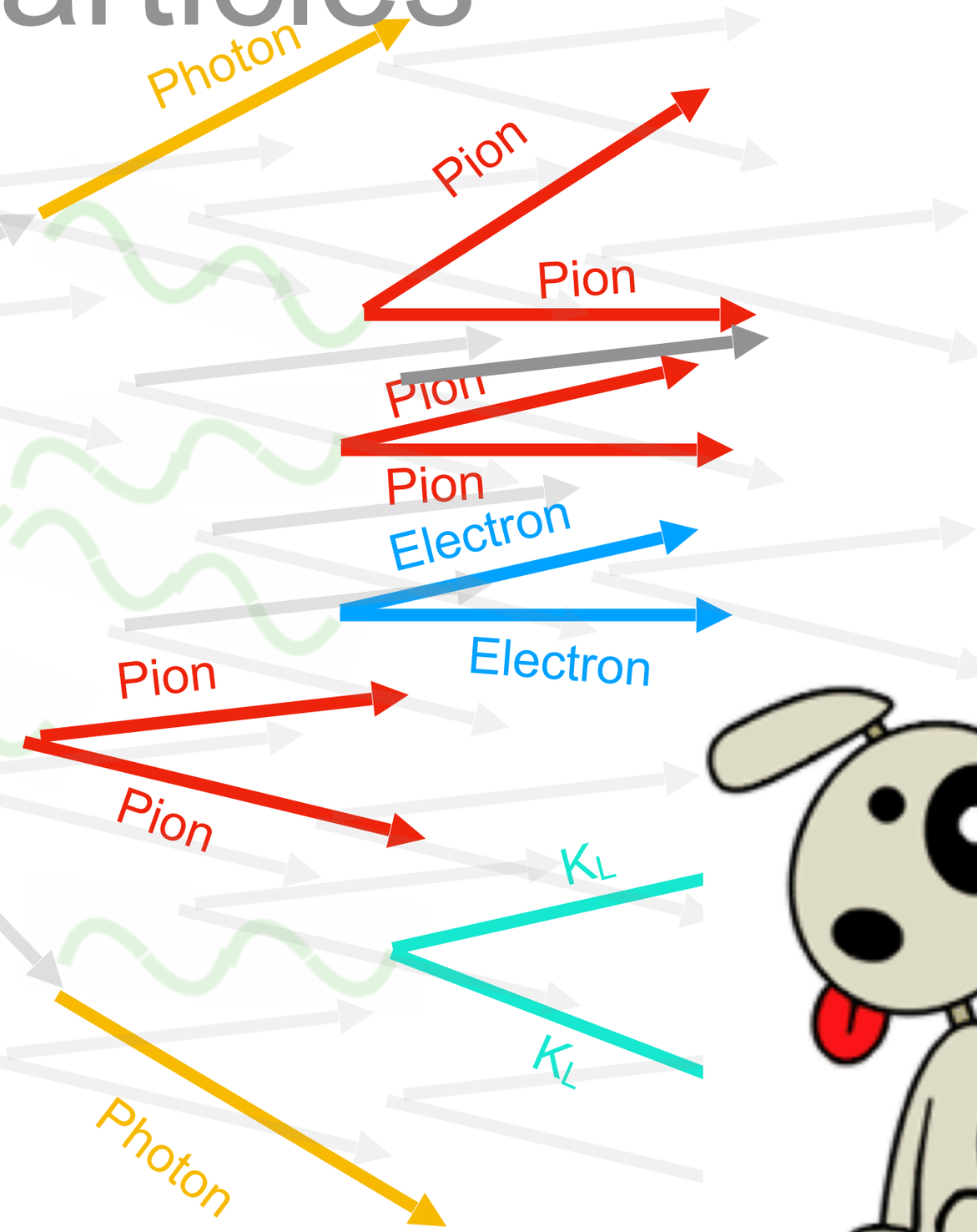
Actually 60+ overlapping collisions(pileup)

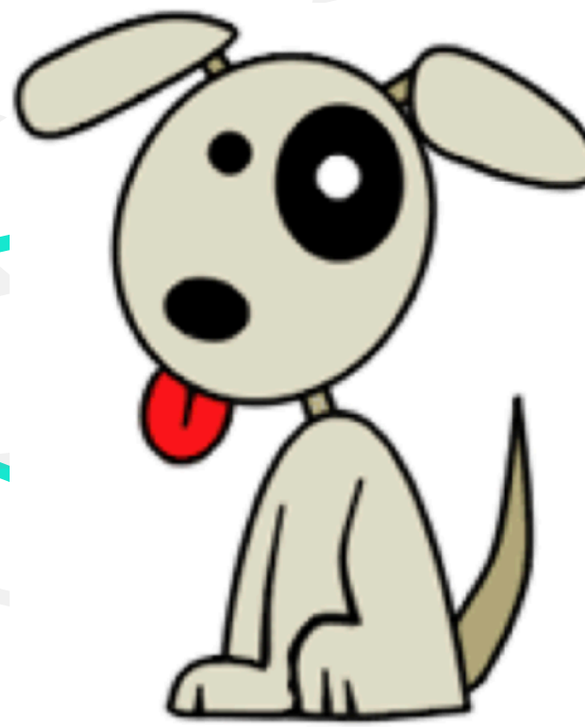


Other Challenges

Additional Particles

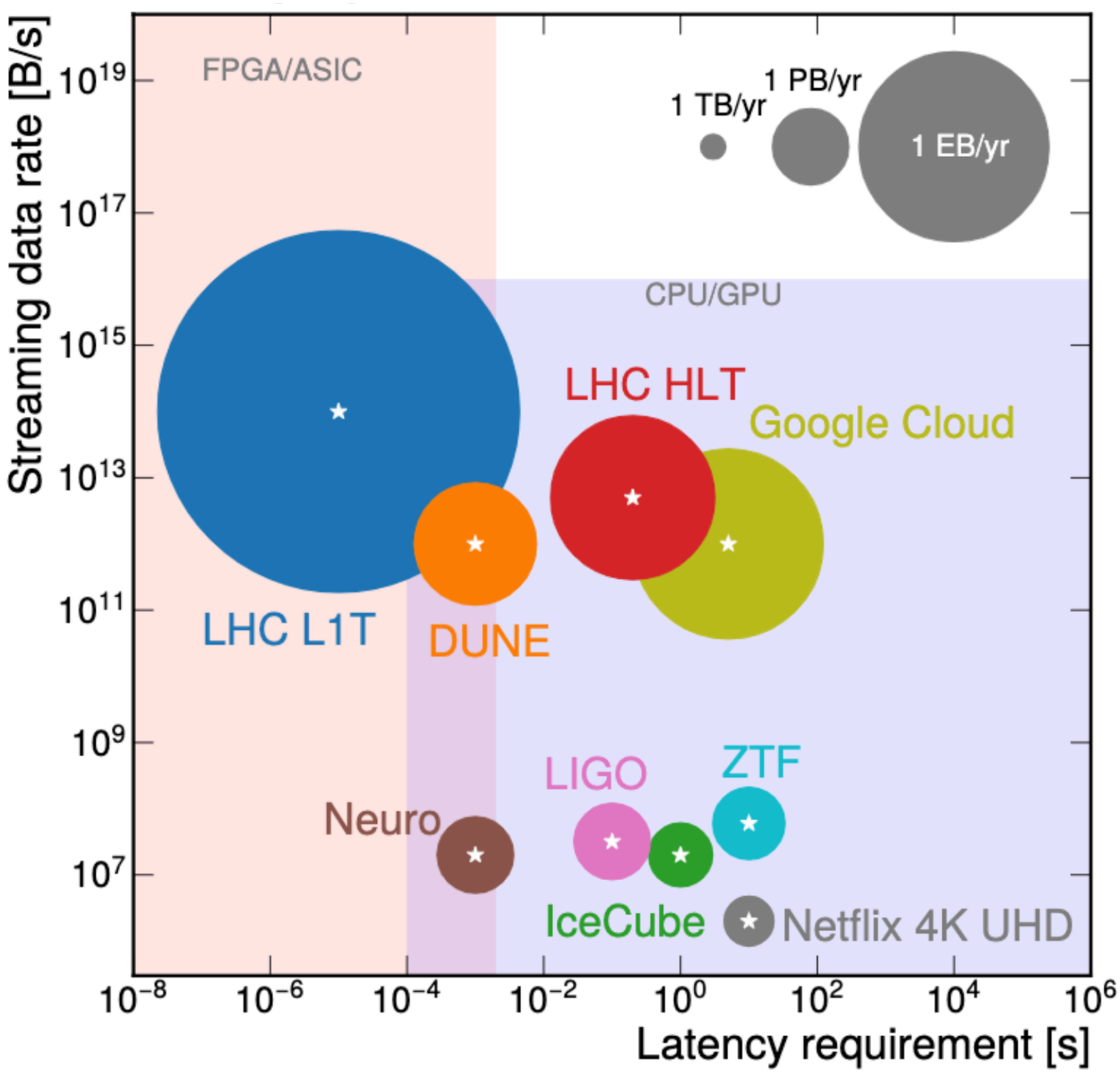




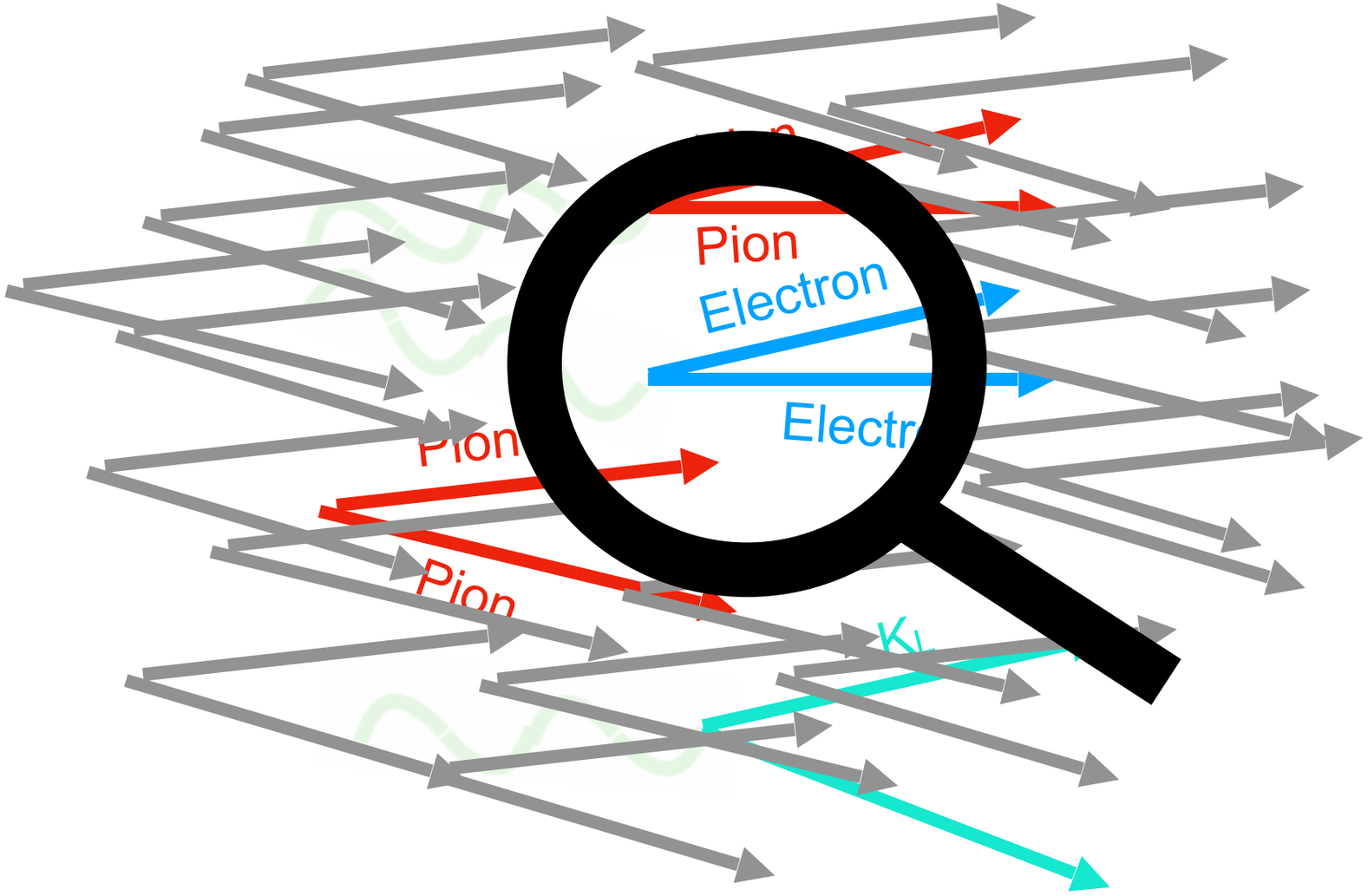


PUPPI Algo gets rid of it

Substantial Challenge@LHC



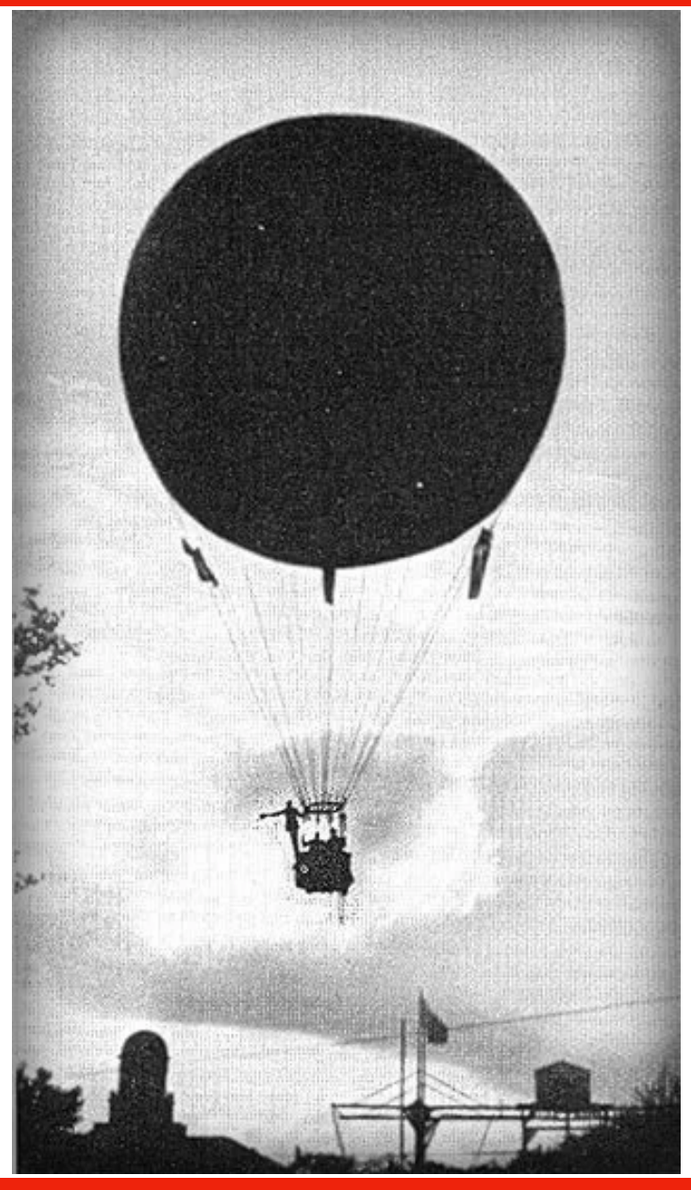
Theme



- Like IBM succeed we can take very complex ideas
  - Simplify to things we understand
  - AI is a key building block to making this happen
- Ultimately this leads to many important physics results

Flow of talk

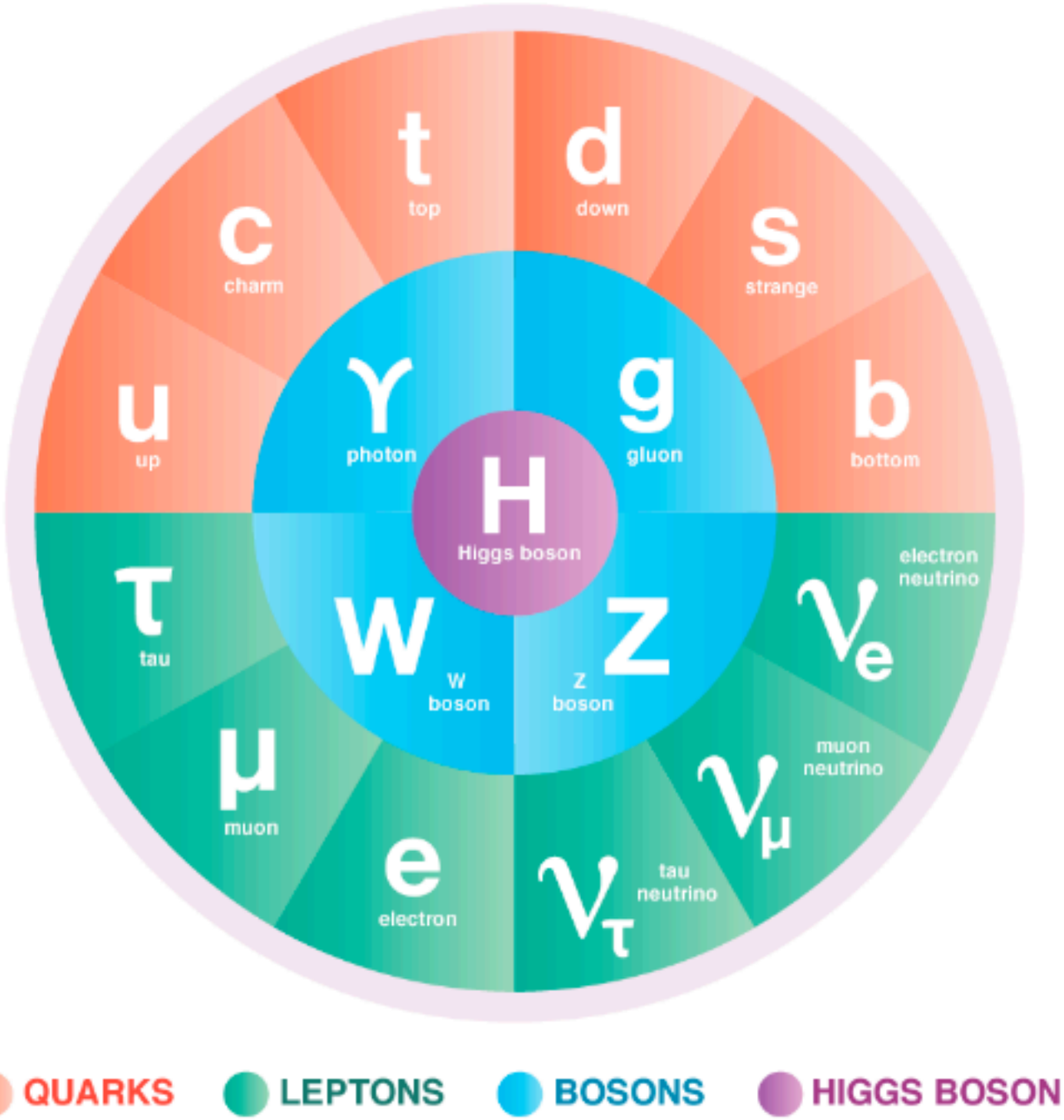
Talk will get faster and Faster as we go through



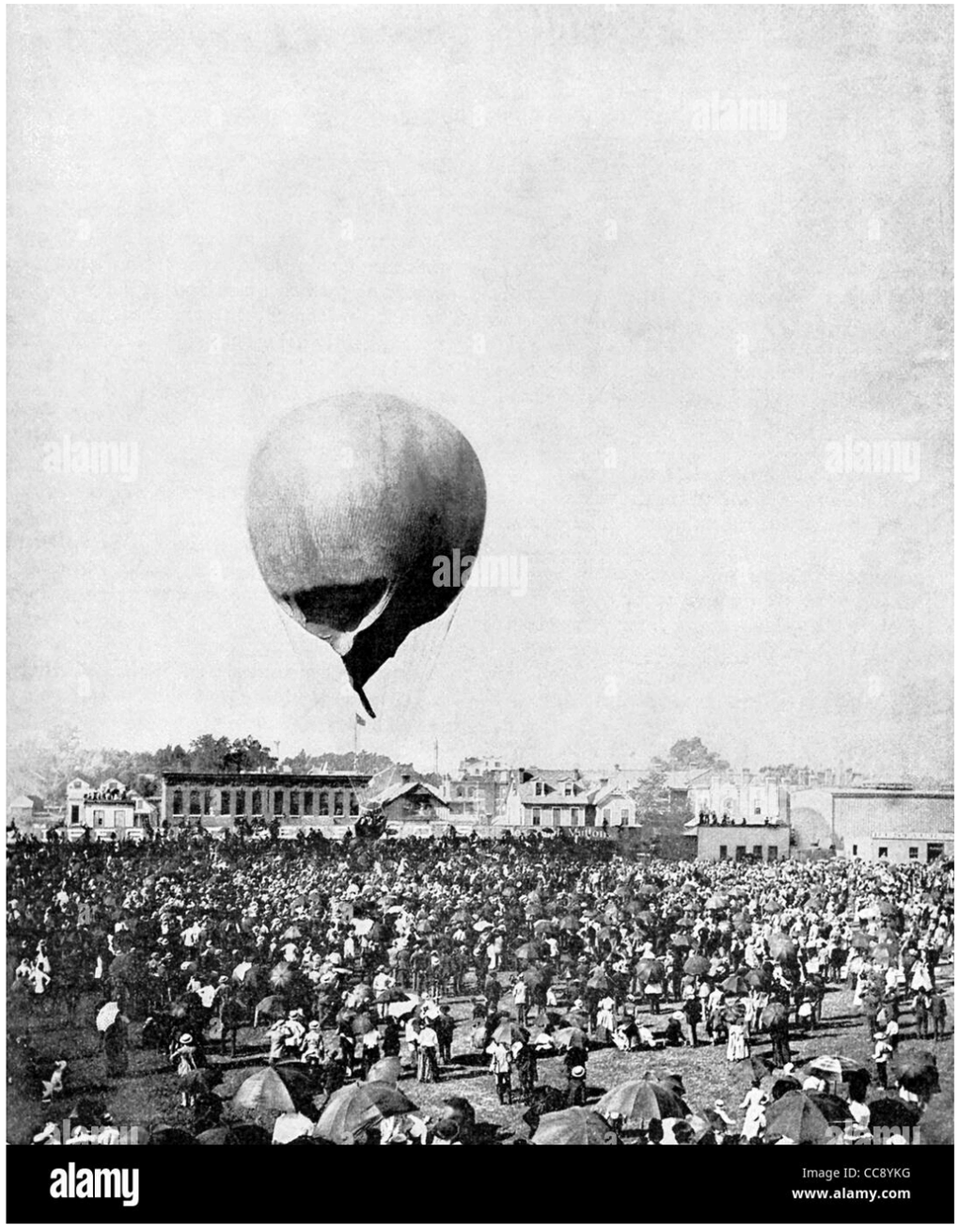
Fast → Slow

Journey Of Forces

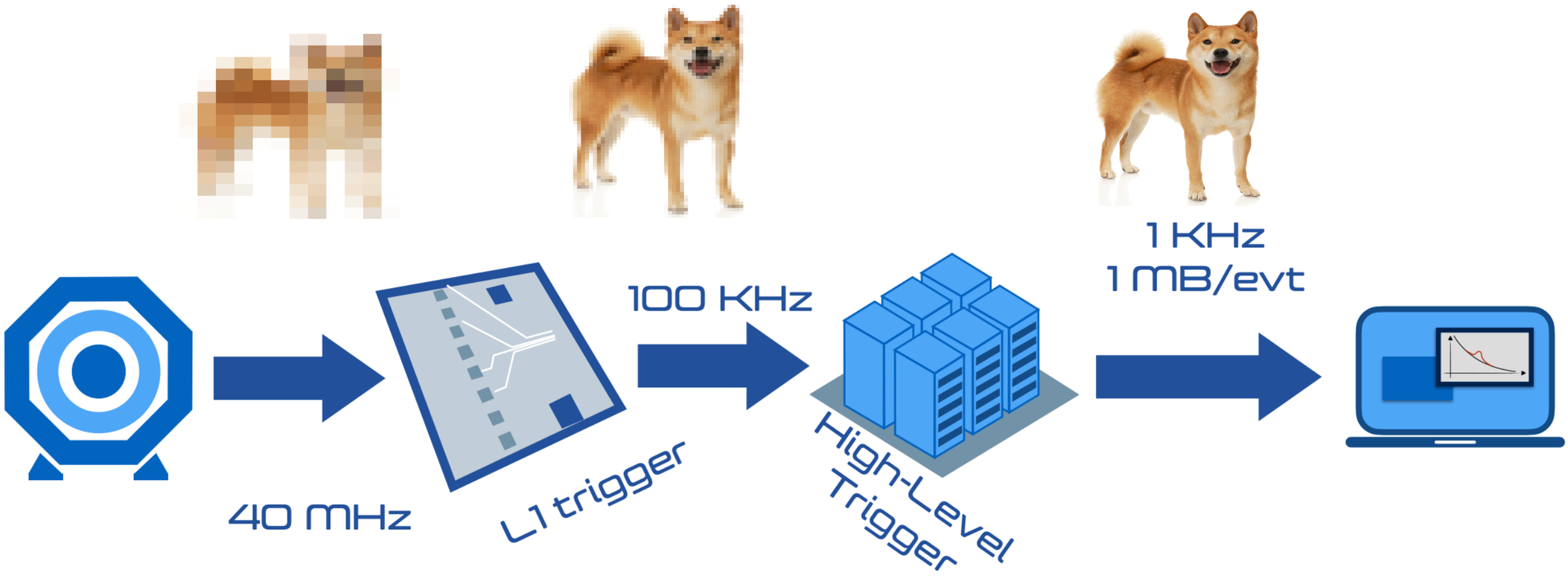
- Chapter 1 : Higgs Boson
- Chapter 2 : Weak Force
- Chapter 3 : New Forces
- Chapter 4 : Strong Force
- Chapter 5 : Electromagnetic
- Chapter 6 : Gravity Forces



This is going to be a whirlwind tour

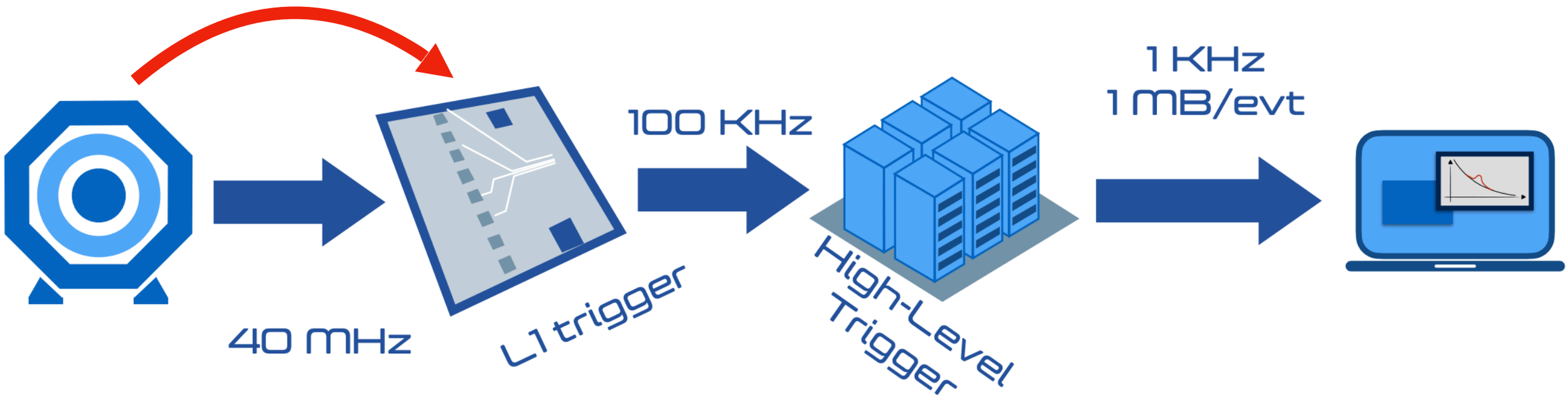


Fast Machine Learning +Higgs



CMS has multiple tiers of data

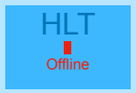
The first tier streams 1 Pb/s of data



1 Pb/s is 17 Yottabytes/year
>100 internets worth of data

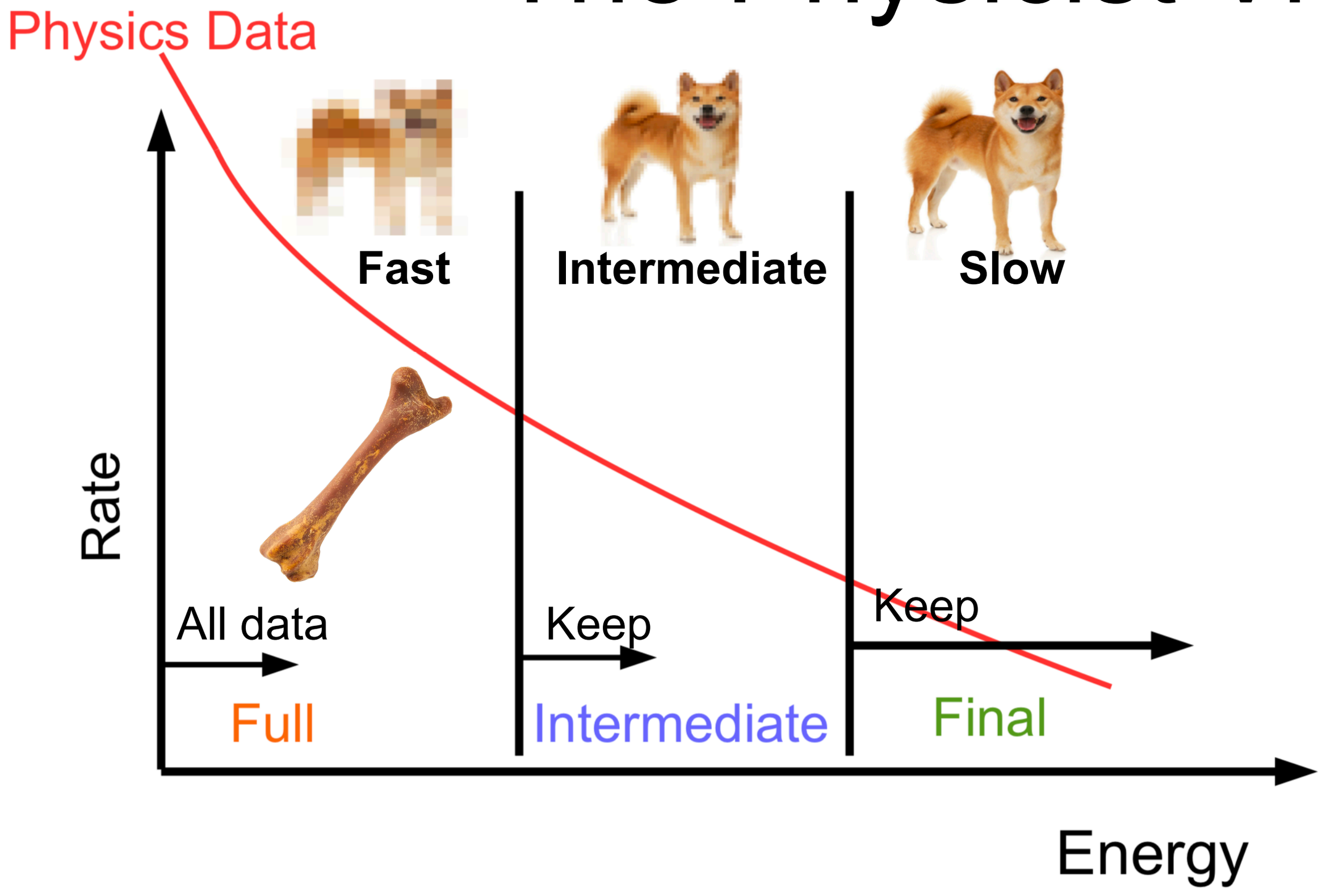
A Vision For Trigger

Everything

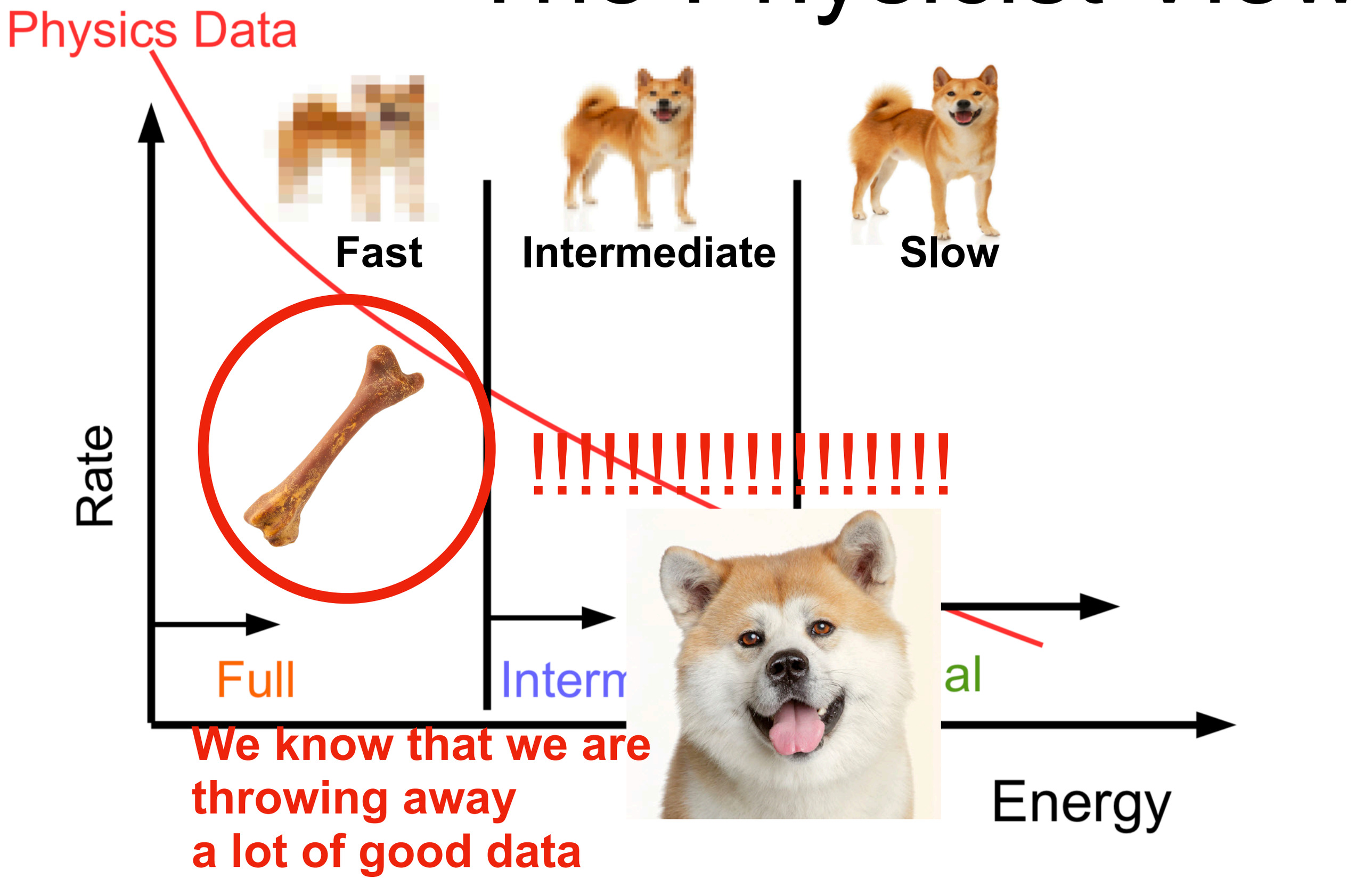


Can we interpret Everything?

The Physicist View



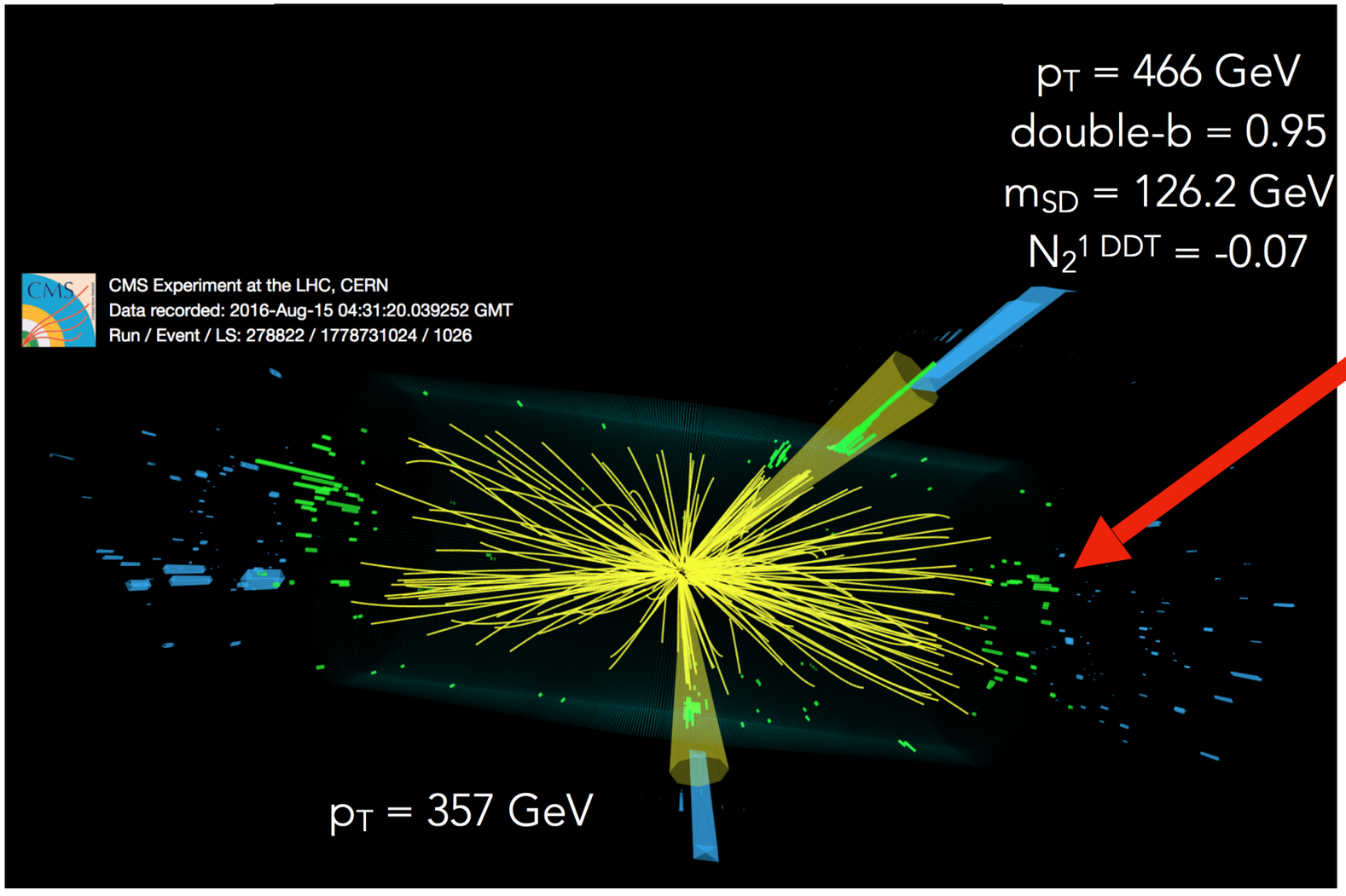
The Physicist View



We know that we are throwing away a lot of good data

Hidden gems?

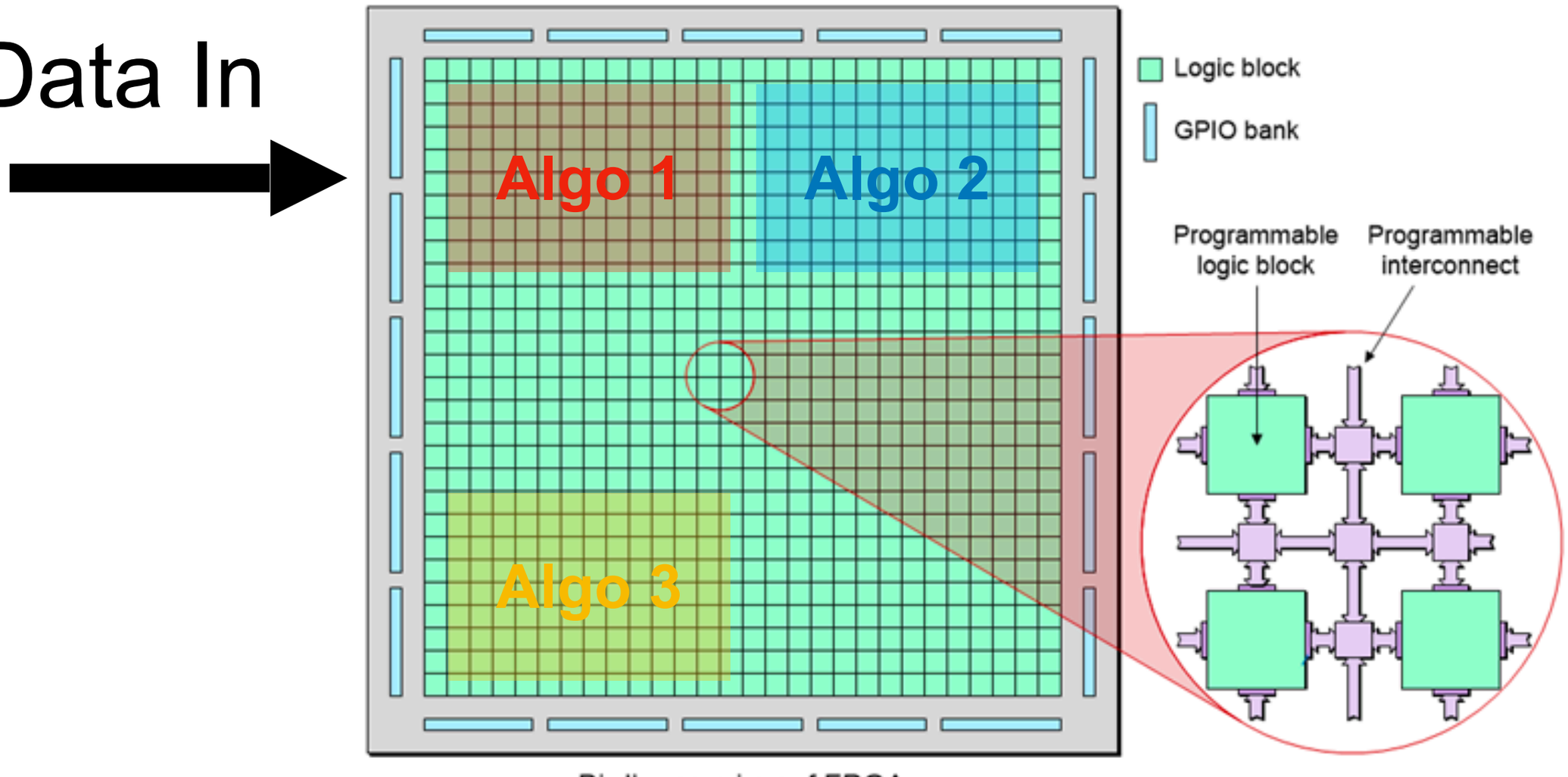
- There is a plethora of physics that we throw out



Higgs boson right on the cusp of being thrown out

Processing Tech

To deal with data rate, we use Field Programmable Gate Arrays

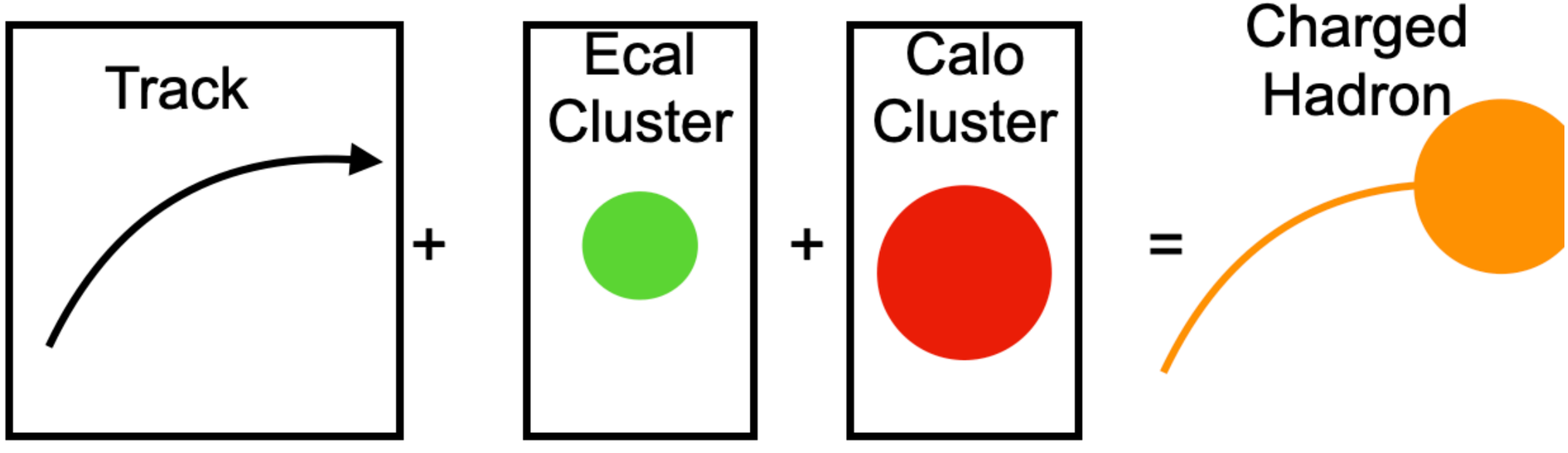


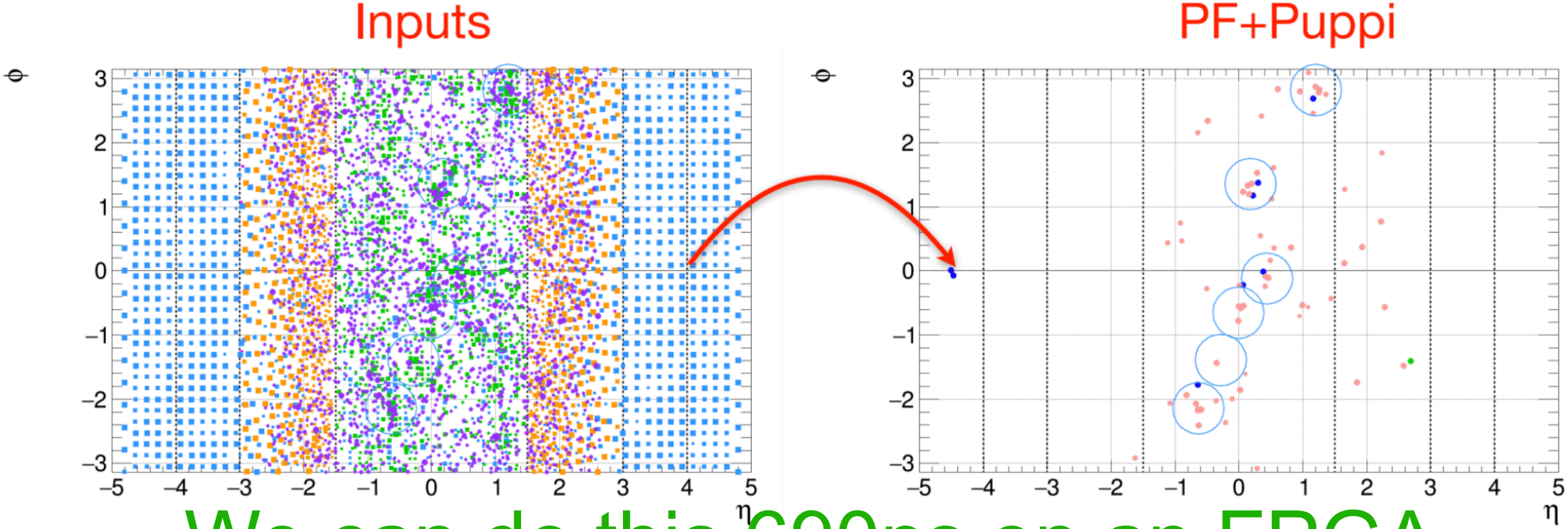
Processing Tech

CPU	GPU	FPGA
1 player	A few at same time	The whole chip
A soloist	A group	An orchestra
Whatever	Main theme and some freedom to improv	Score has to be known perfectly beforehand

Next Generation System

On Chip



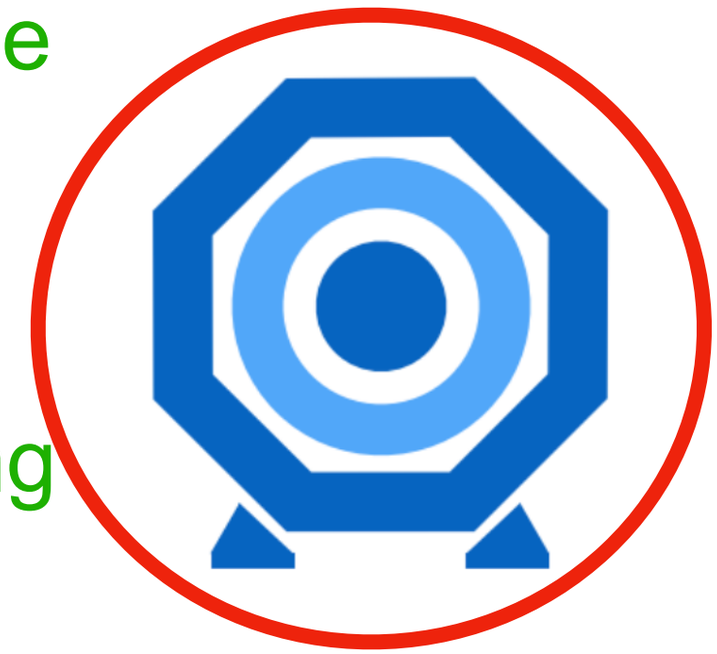


We can do this 690ns on an FPGA

On Every LHC collision

Real-Time Deep Learning

Human Scale Response times
> 1 ms
(Ex. speaking Is 5-10 Hz)



LHC Response times
25 ns
Collision rate is 40 MHz

Bringing Deep Learning to the LHC required a complete rethinking of how we program it on a chip

This paradigm for deployment of AI algorithms didn’t exist before

BU Alumni

Penn Arts & Sciences
Department of Physics and Astronomy

About People Undergraduate Graduate Courses

Dylan Rankin

Standing Faculty

Assistant Professor

Research Areas: Particle Physics

dsrankin@sas.upenn.edu

DRL 4N16A

The next many slides done w/BU Ph.D Dylan Rankin

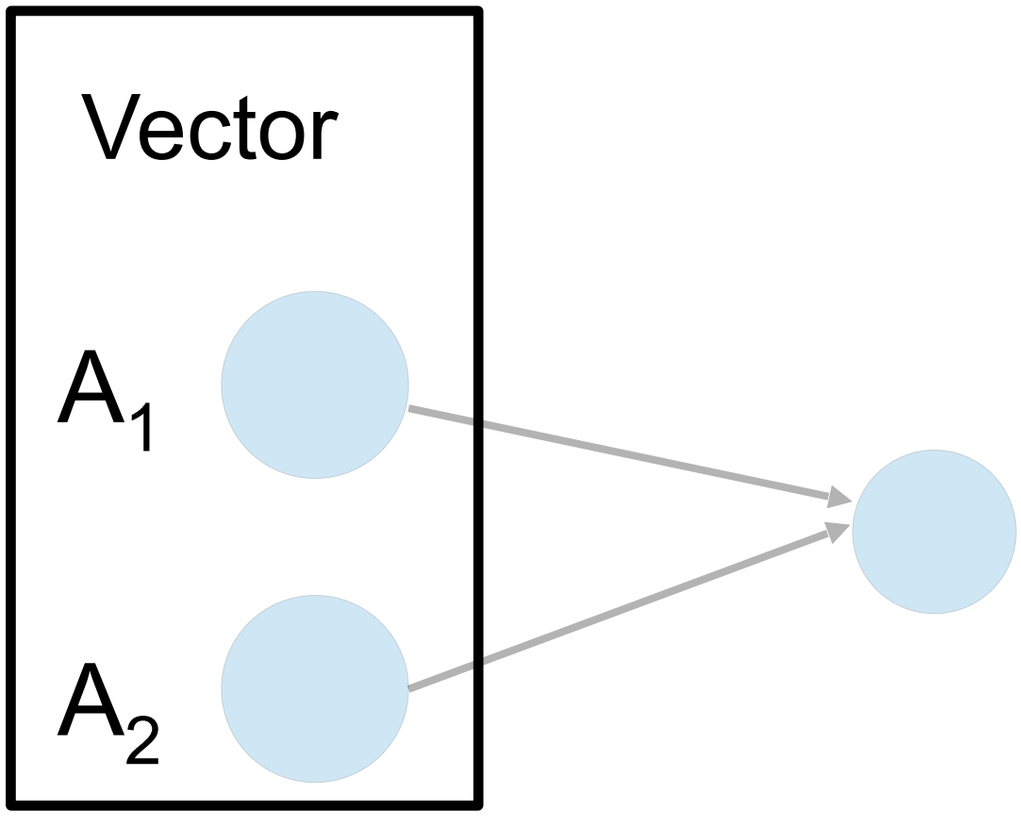
Now faculty at Penn

Note the education

Education

Ph.D. Physics, Boston University (2018)
Sc.B., Physics, Massachusetts Institute of Technology (2012)

Matrix Mult in Math



How can we parallelize this?

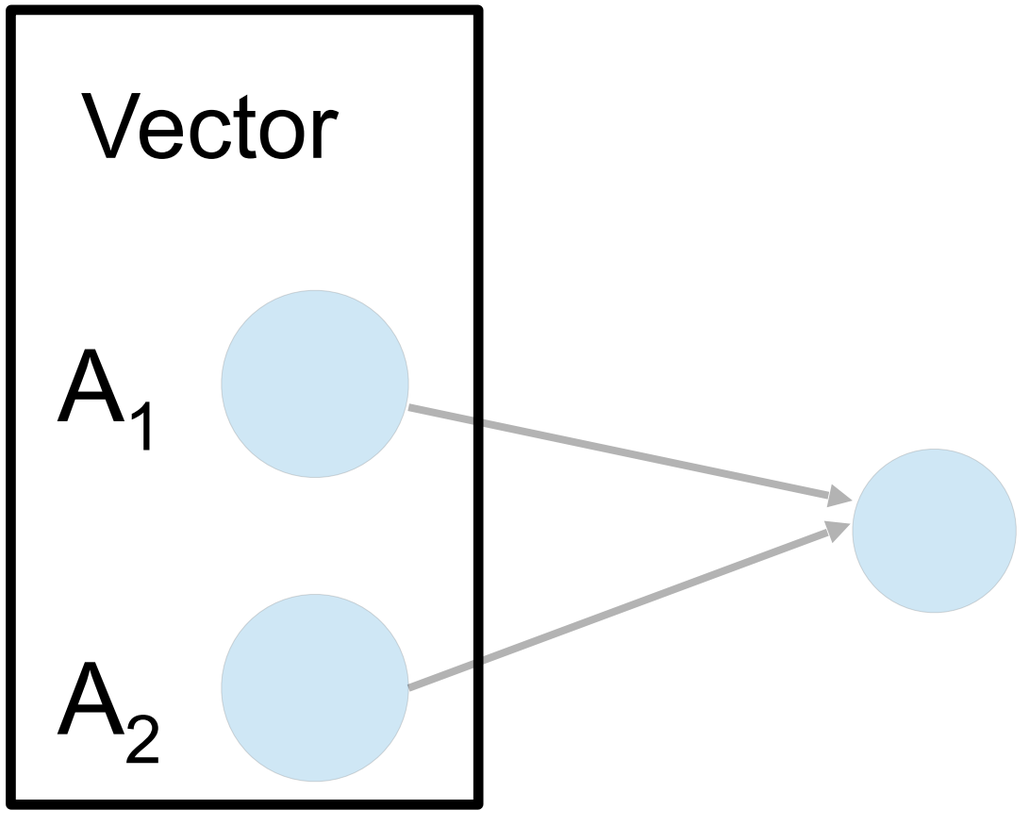
$$\varphi(A_1W_{11}+A_2W_{21}+B_1)=O_1$$

Activation function

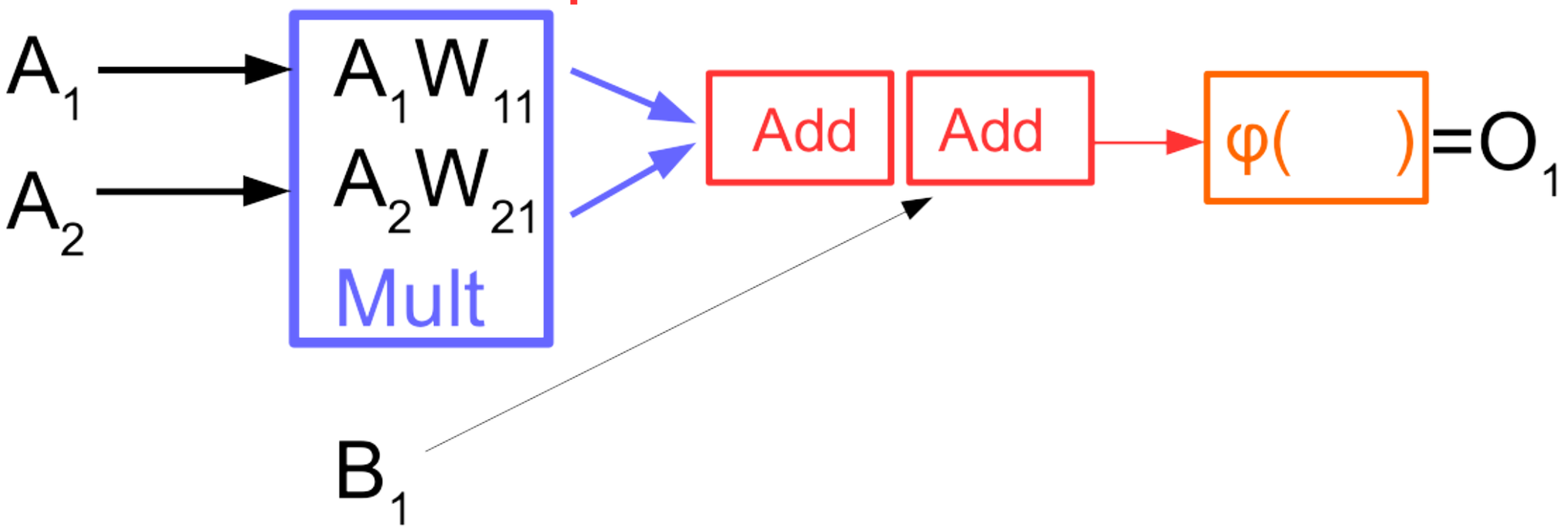
Matrix Multiplication

Vector Addition

Matrix Mult in Math



How can we parallelize this?



Matrix Mult in an FPGA

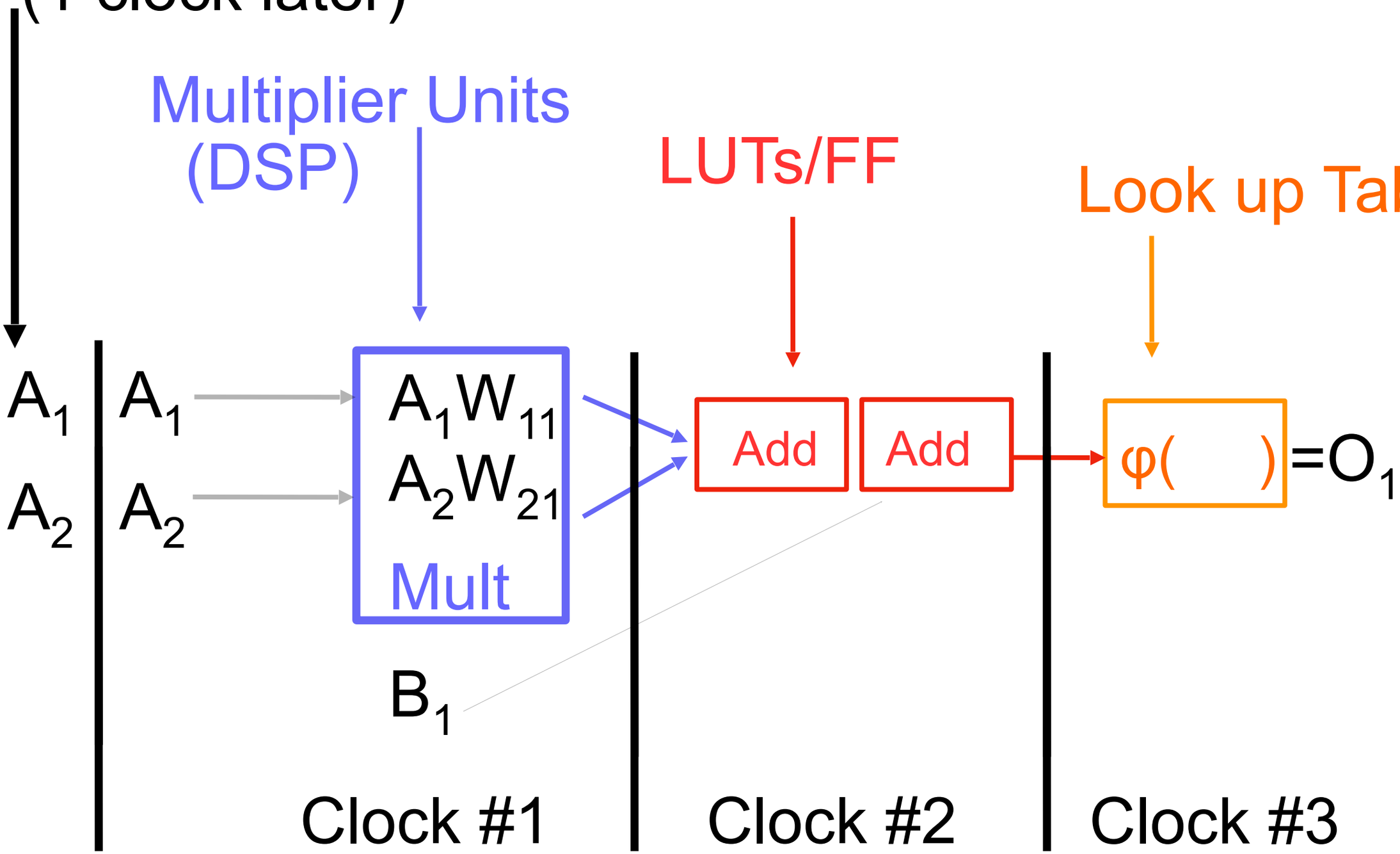
Next vector of inputs
(1 clock later)

3 Clock algorithm

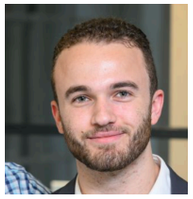
Multiplier Units
(DSP)

LUTs/FF

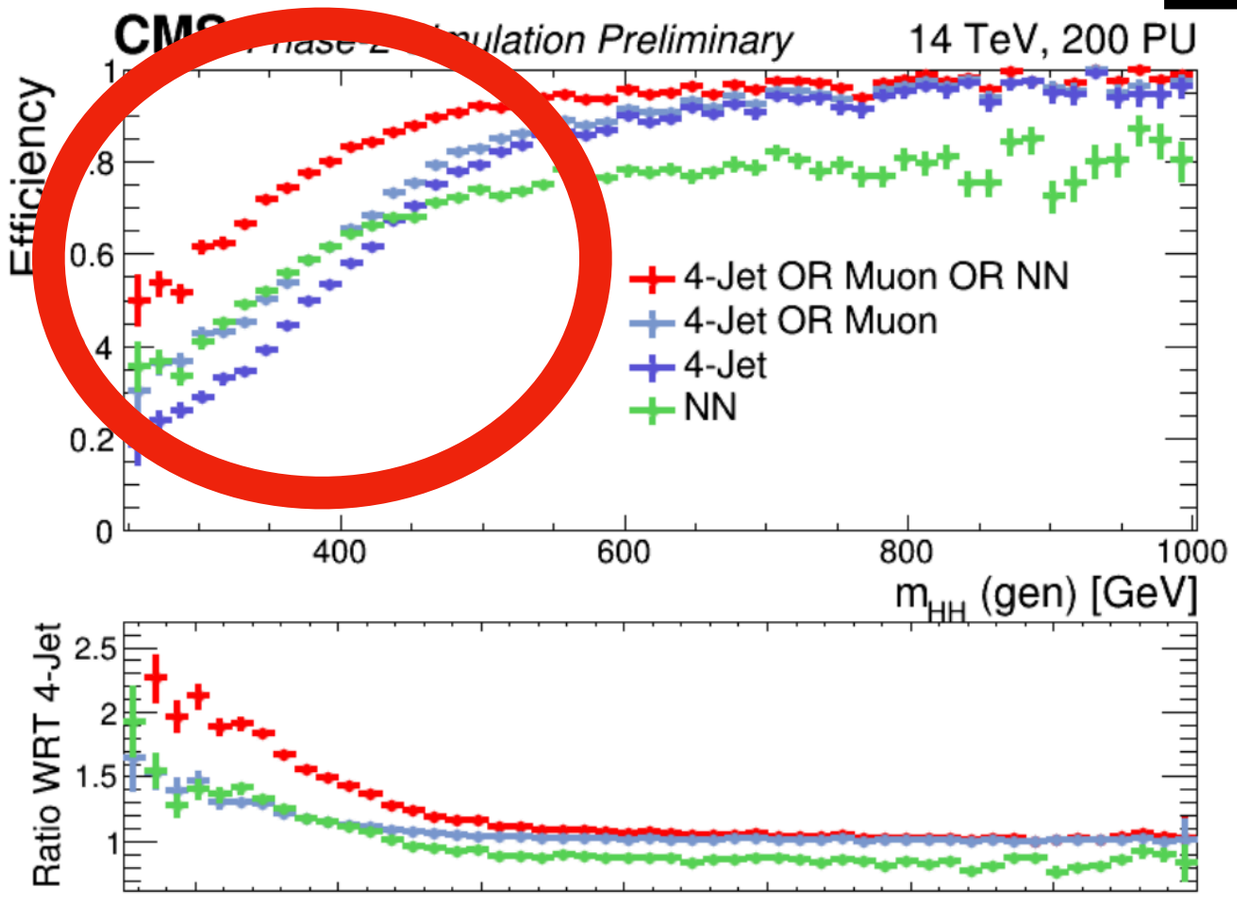
Look up Table



Results subject to precision outputs



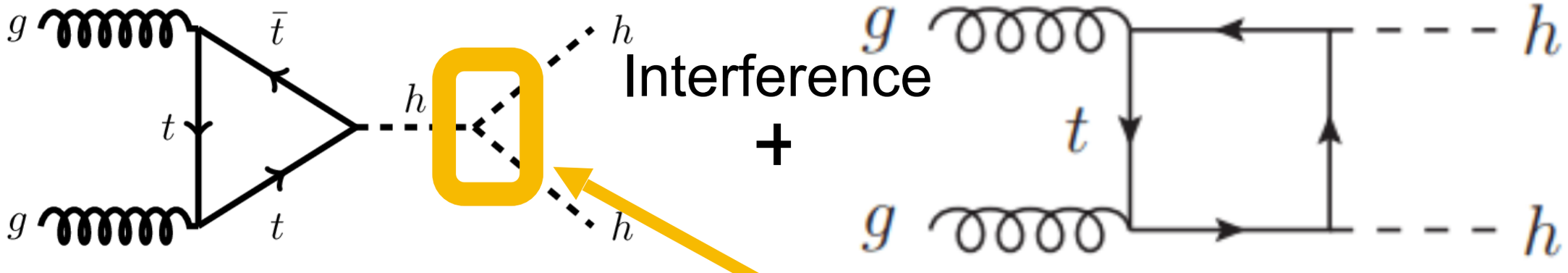
Example:B-Tag



NN for b-quark tagging
For (h + h) $\rightarrow \bar{b}\bar{b}\bar{b}\bar{b}$

Yields 30-40% improvement in Higgs self coupling

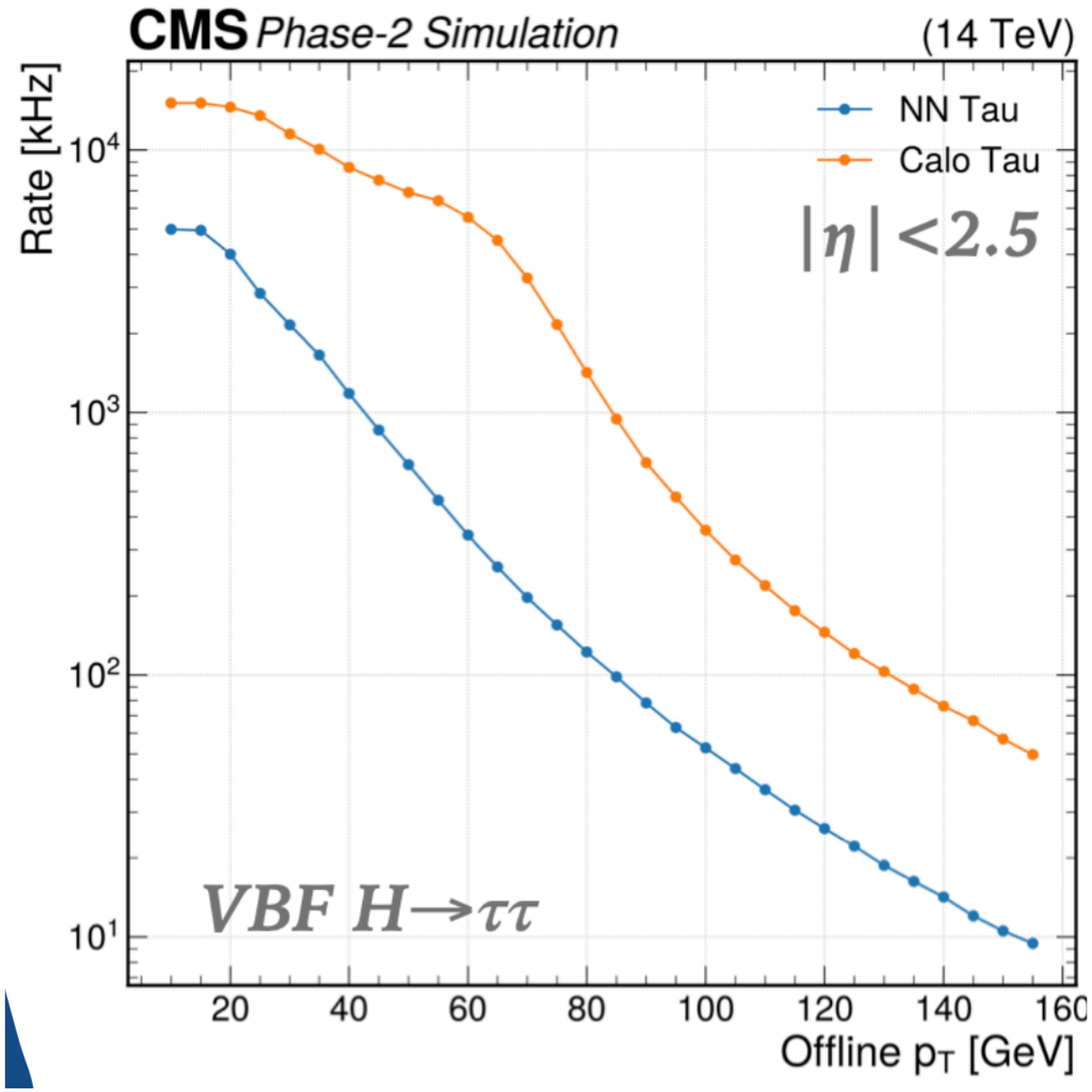
Network runs in 80 nanosecond

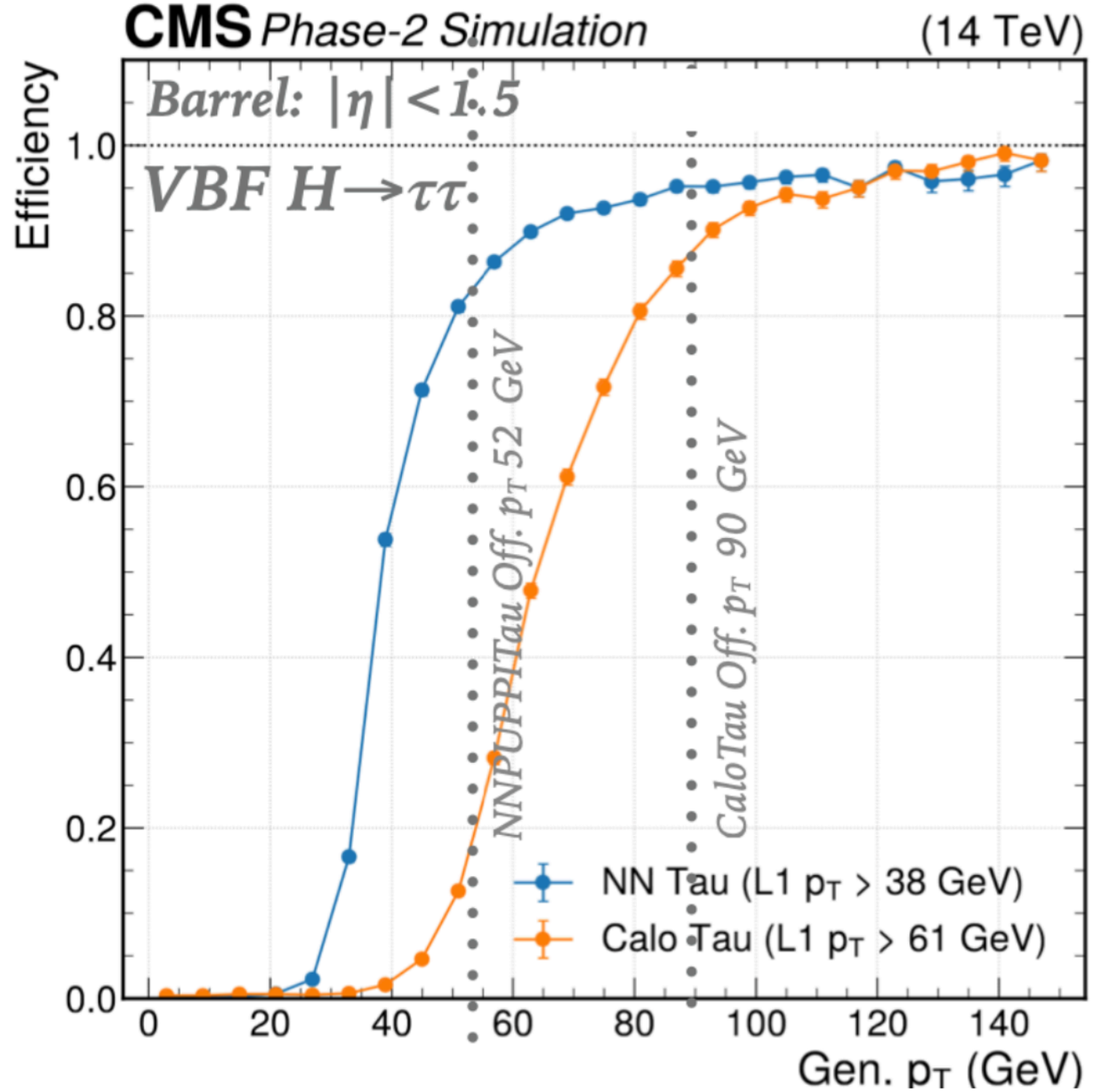


Di Higgs Boson Production

Higgs Self Coupling Term

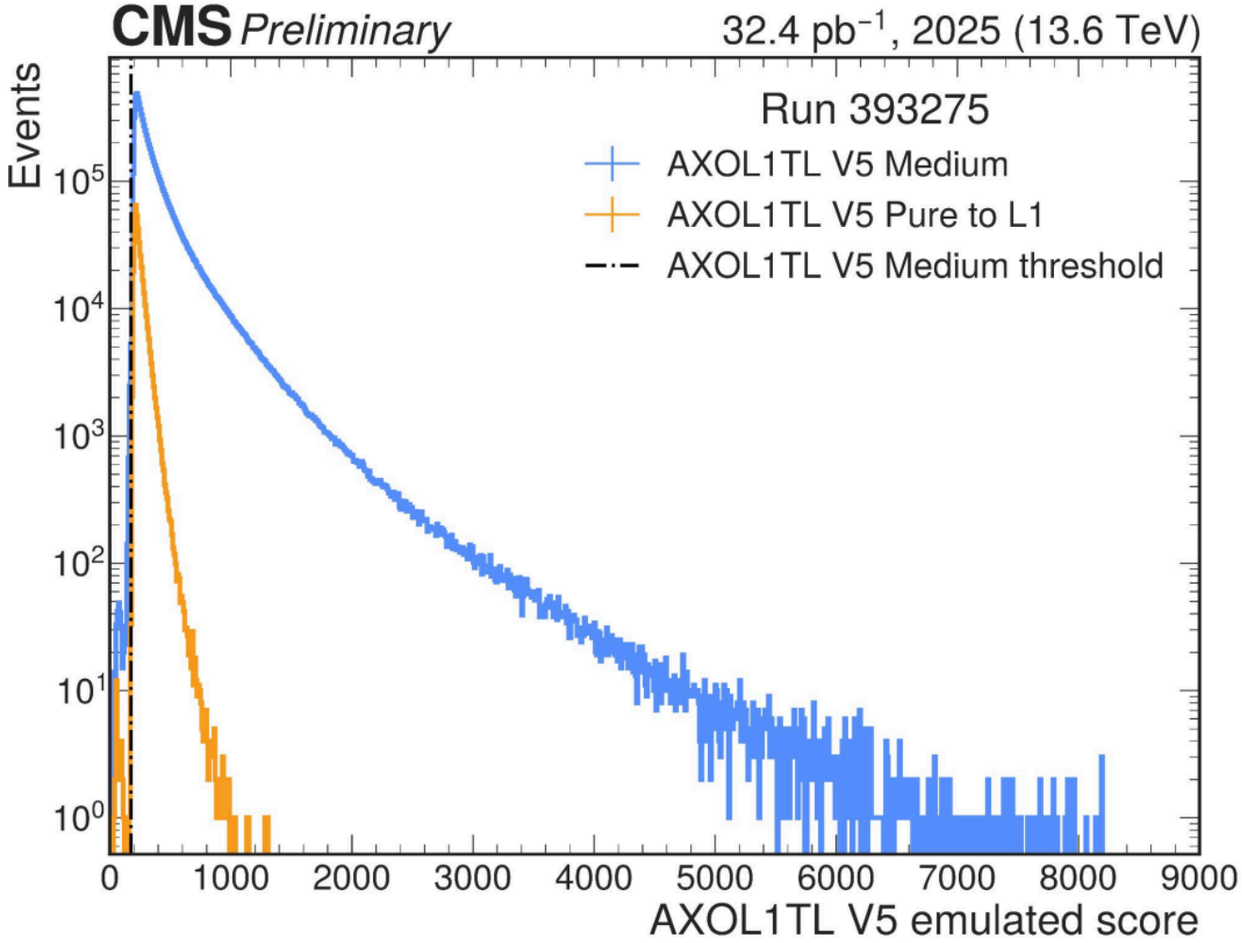
Another Example are Taus



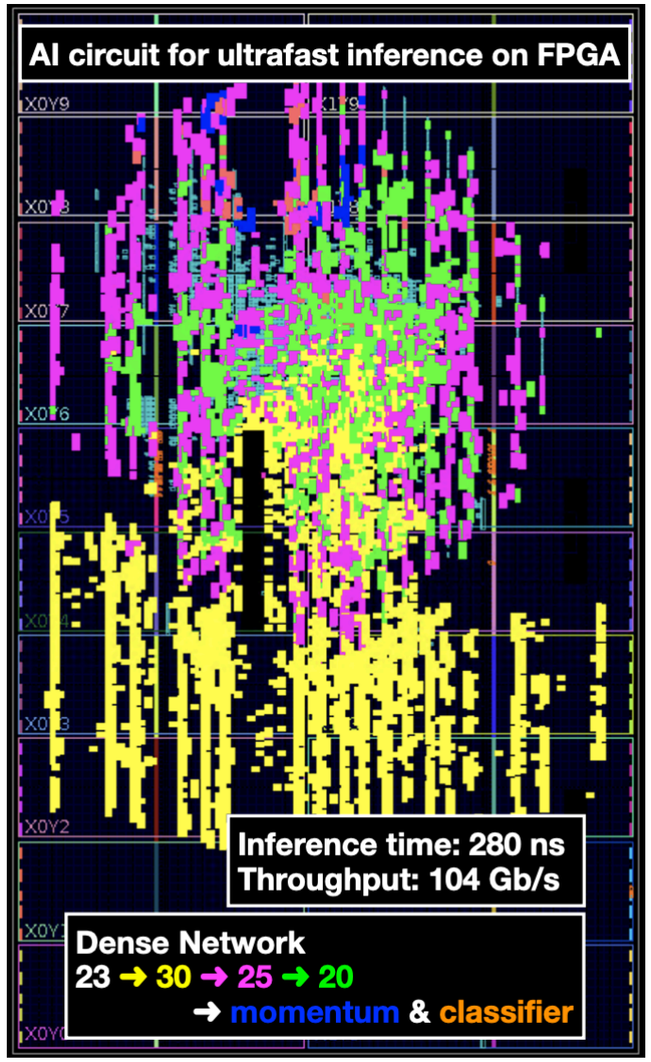


https://fastmachinelearning.org/hls4ml/

ML in Trigger Running Now

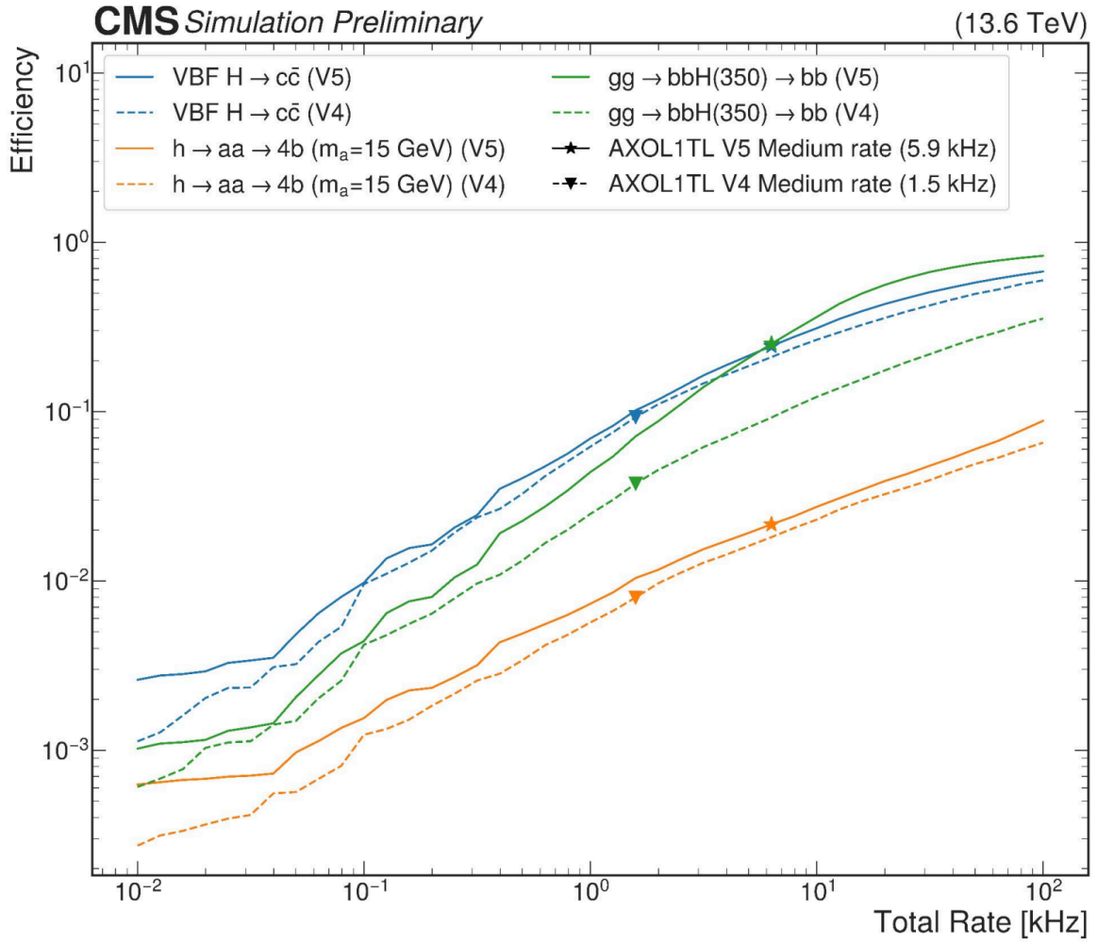


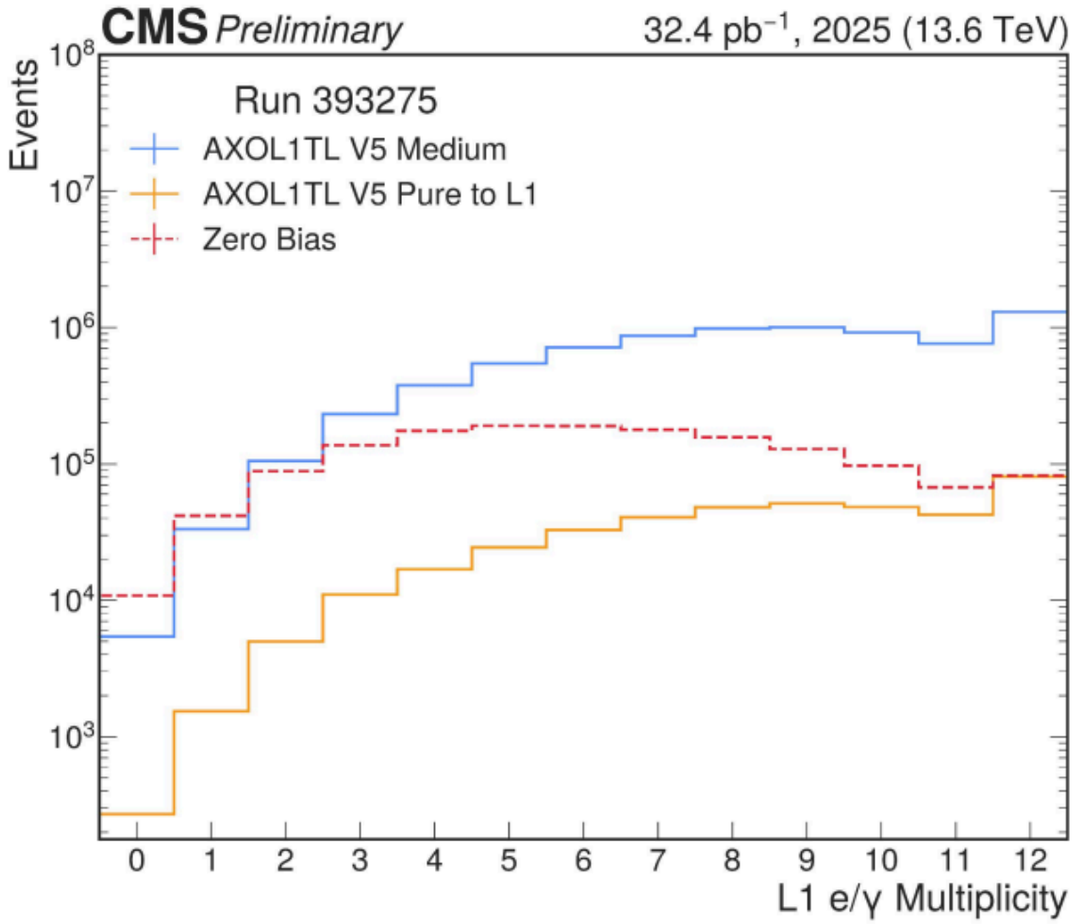
AI Anomaly Detection



Displaced Muon ID
Muon Momentum regression
(x2 reduction in rate)

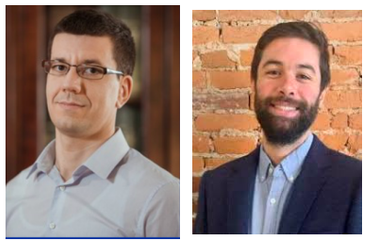
What do you get?





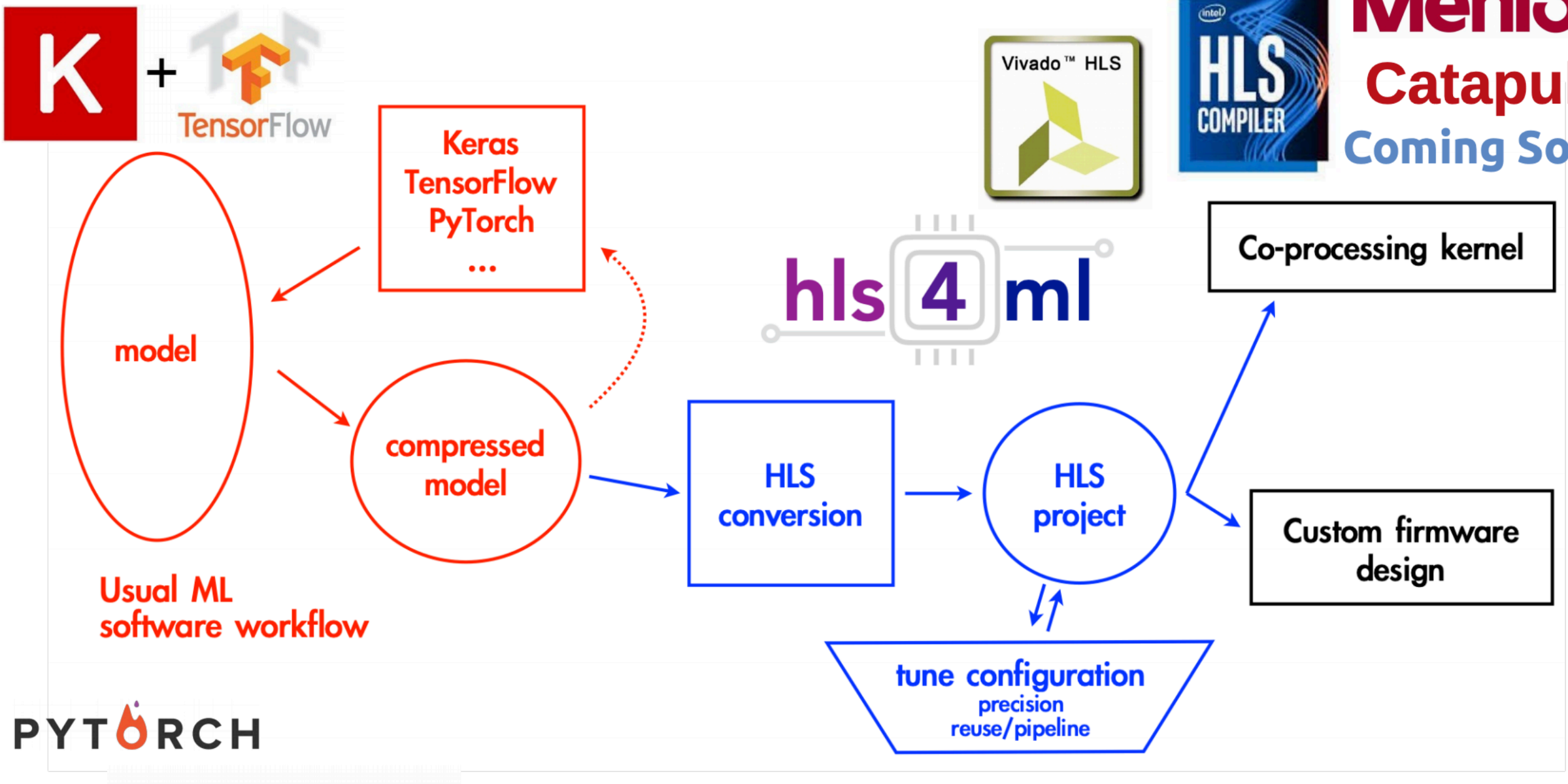
AI is building a space and scraping all the weird events

We know its interesting, but still to analyze



HLS4ML Software Tool

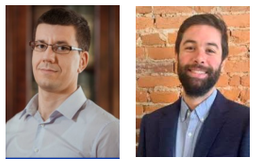
```
python keras-to-hls.py -c keras-config.yml
```



https://fastmachinelearning.org/hls4ml/

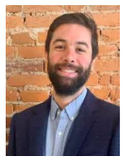
Shrinking Network

- We often want to shrink things as much as possible



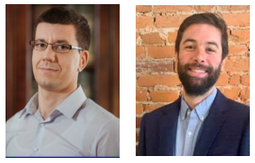
Quantize

- Reducing bit precision of algorithm helps a lot



Sparsify

- Keeping only the infor we care about



Clever Sparsify

- Embedding only the most critical info

AI inpainting

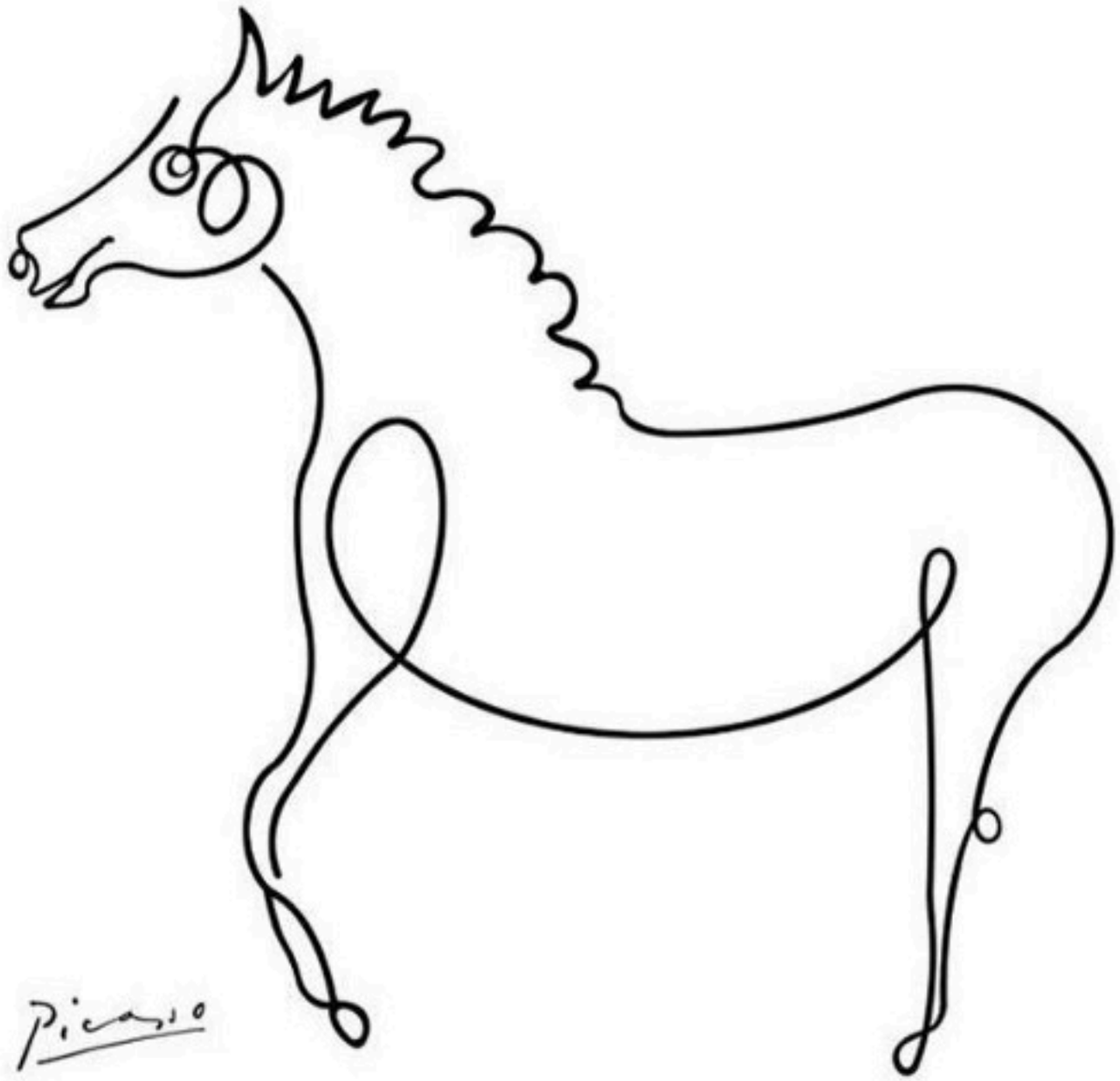
- Using AI to recover our info

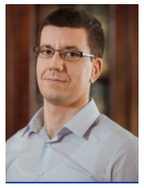
Fastification Success?

The Original

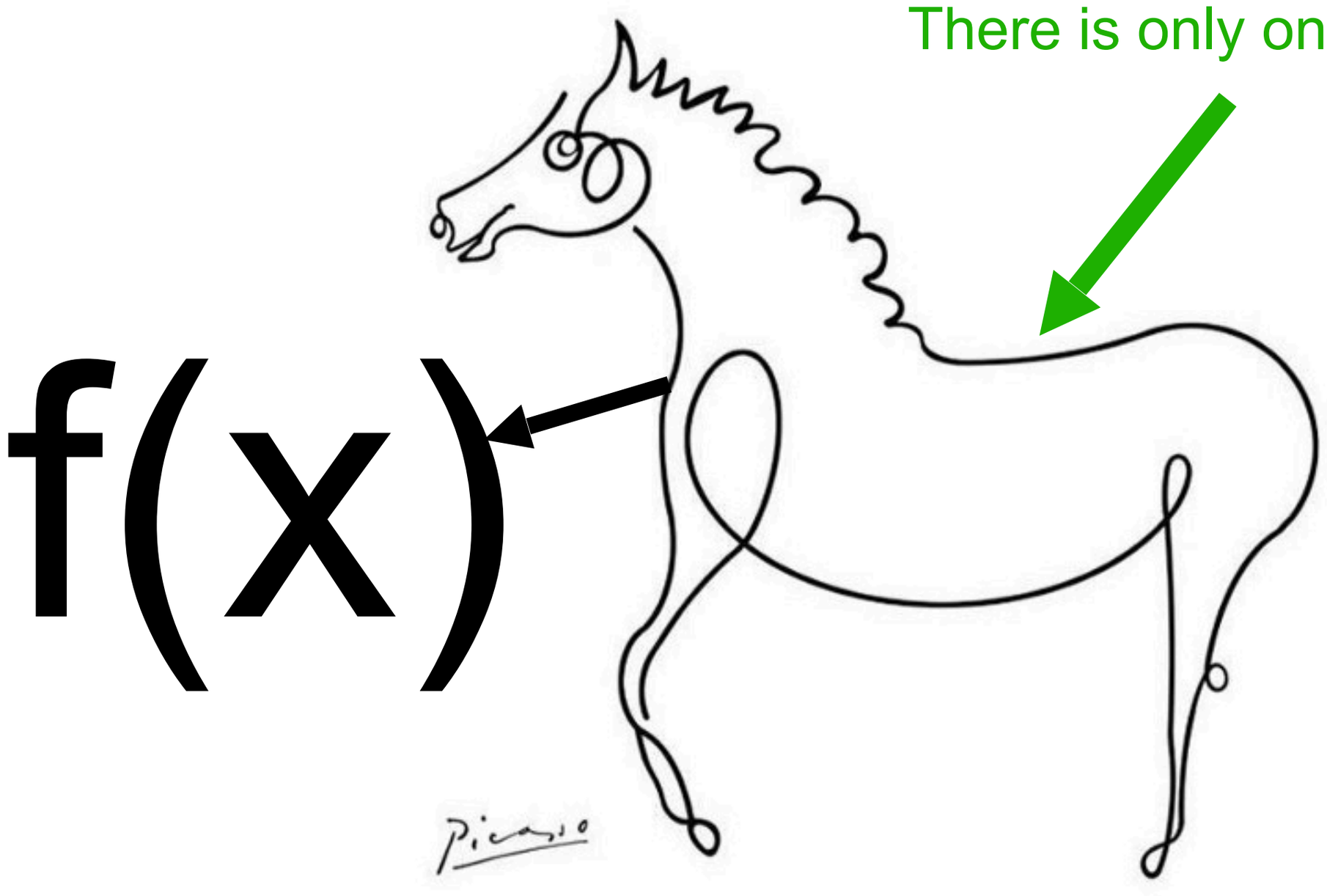
AI in-painting

Can we do better?





Can we do better?



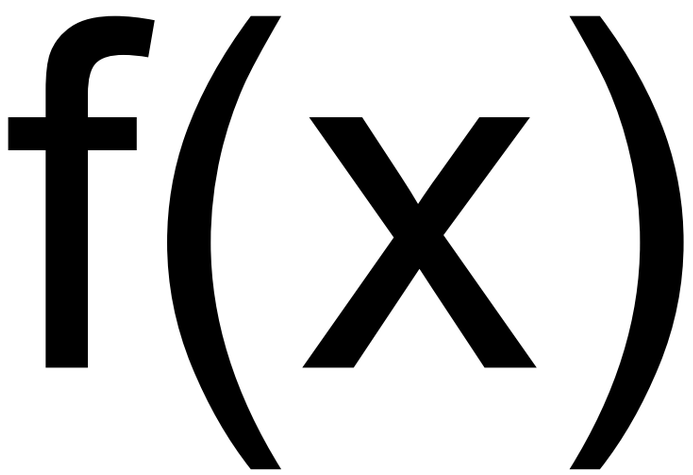
You can get to the answer very fast, but creativity is the key

Symbolic Methods

NN Problem

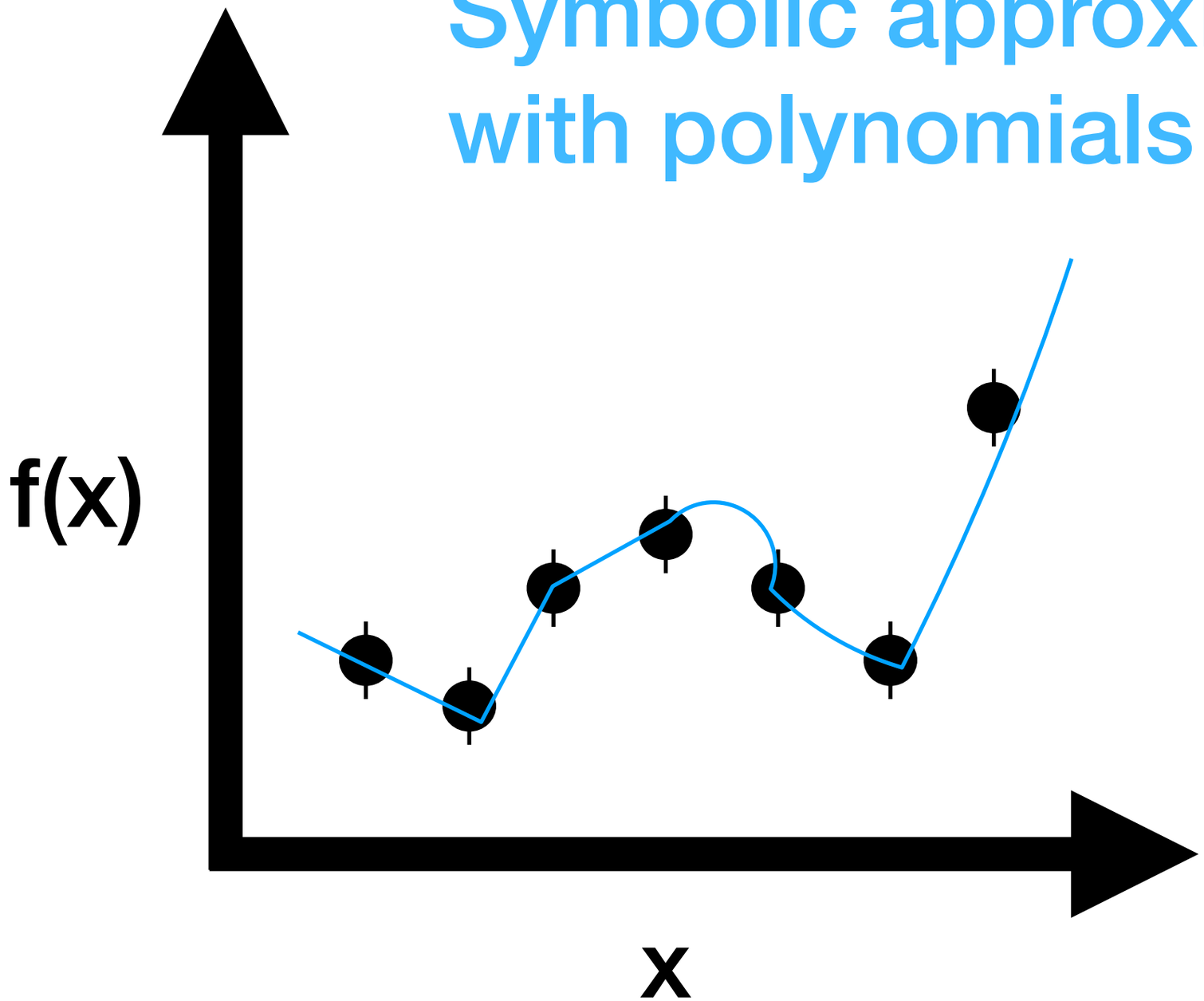
Take an NN
Compress it down
To a functional form

(With Physics)



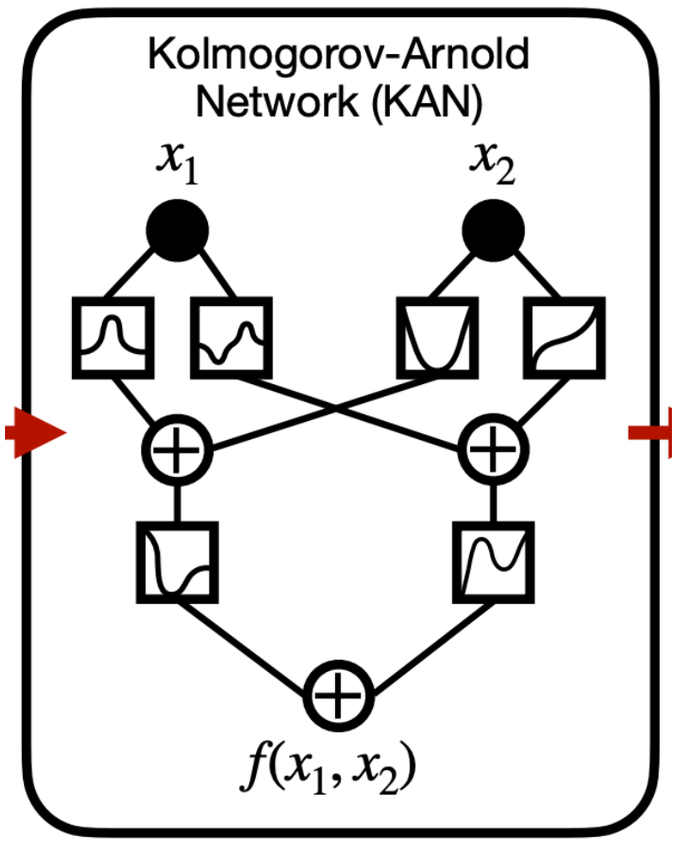
Symbolic Methods

Symbolic approximation with polynomials (Spline)



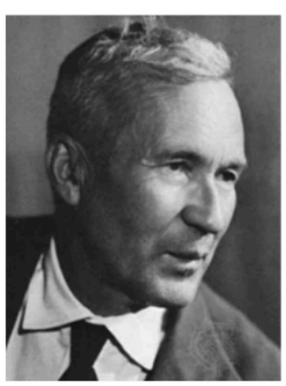
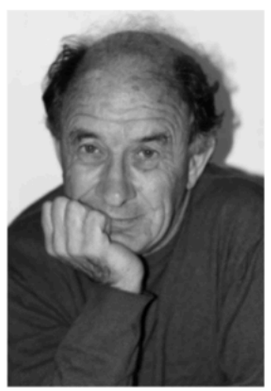
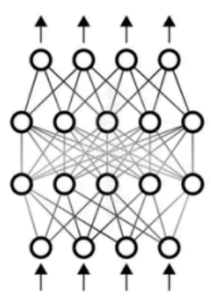
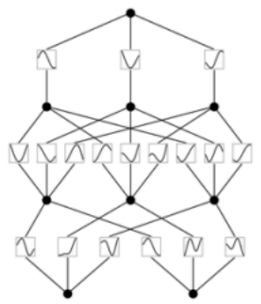
Symbolic Methods

Kolmogorov Arnold Networks



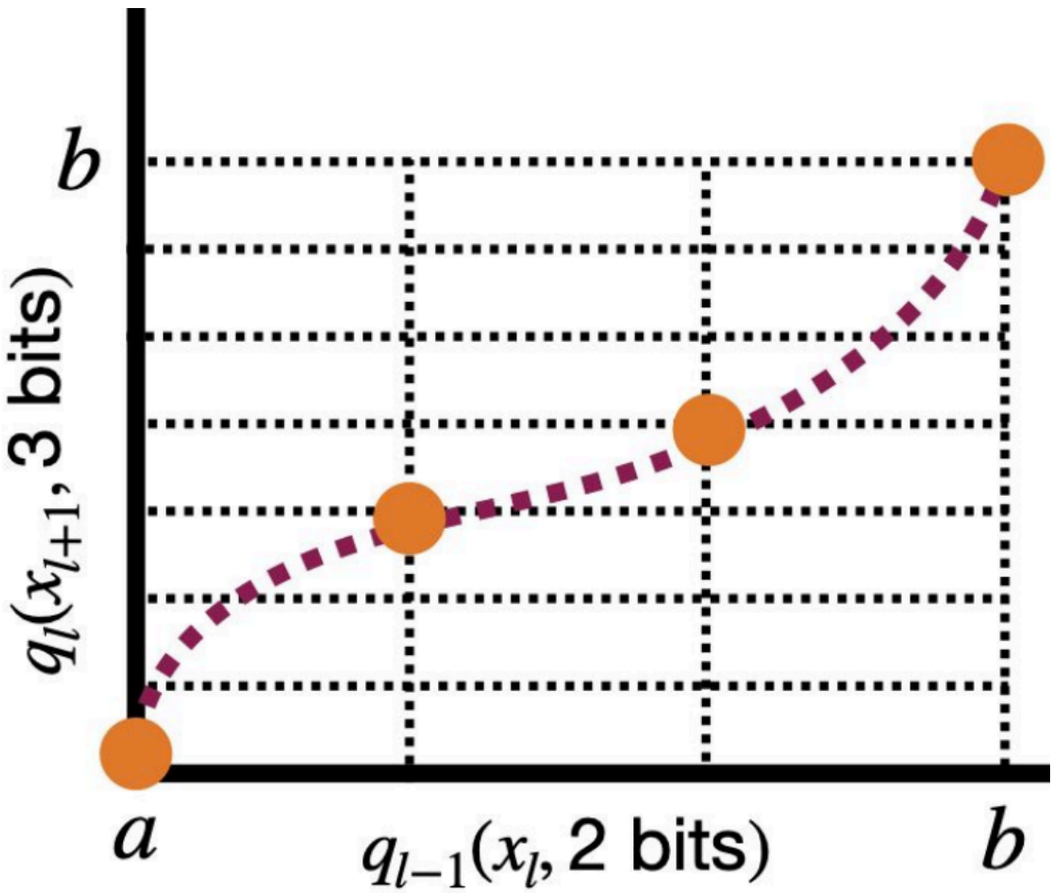
All N-dimensional functions can be represented by a sum of many 1D functions

Kolmogorov + **A**rnold + **N**etwork = **KAN**

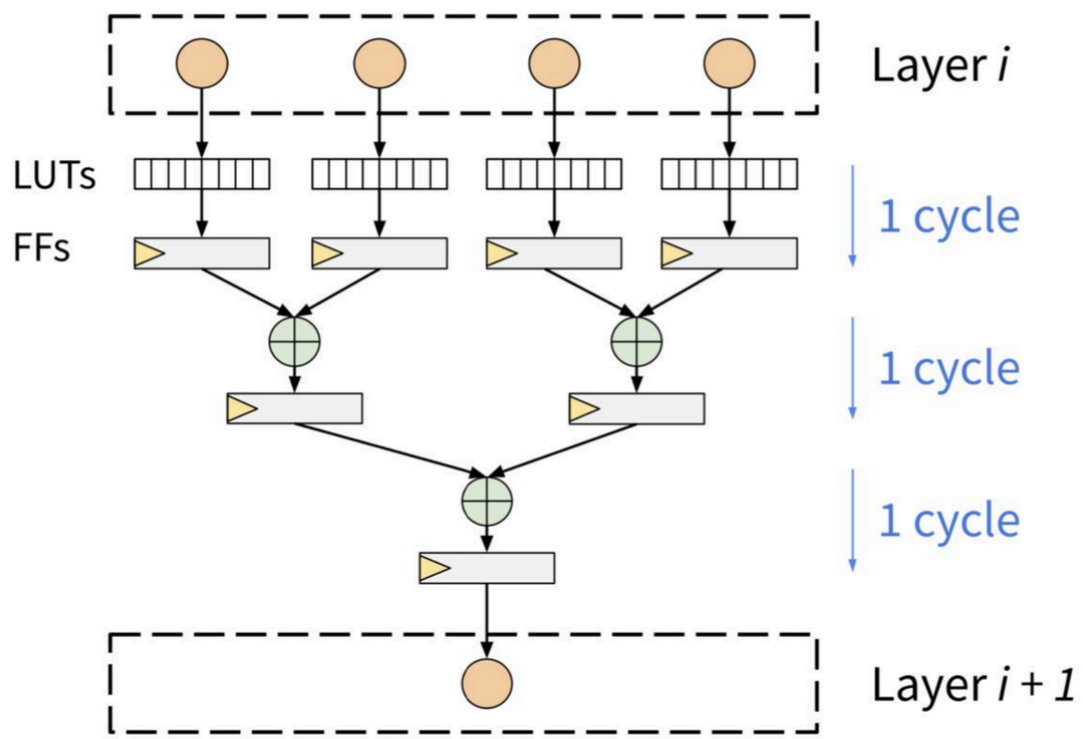





KAN for FPGA: KANELÉ

Sum over (Quantized) Splines



Compute splines

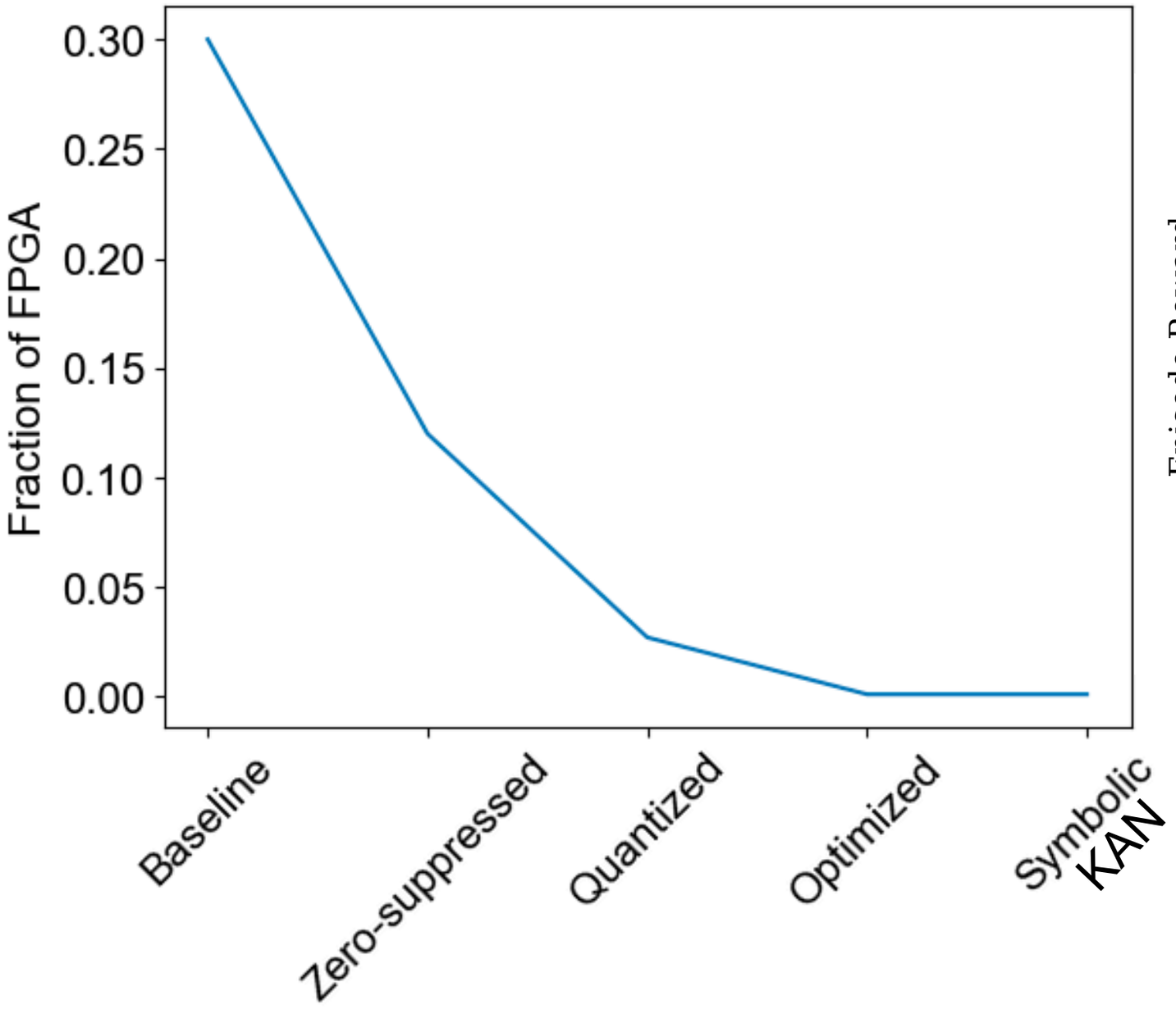


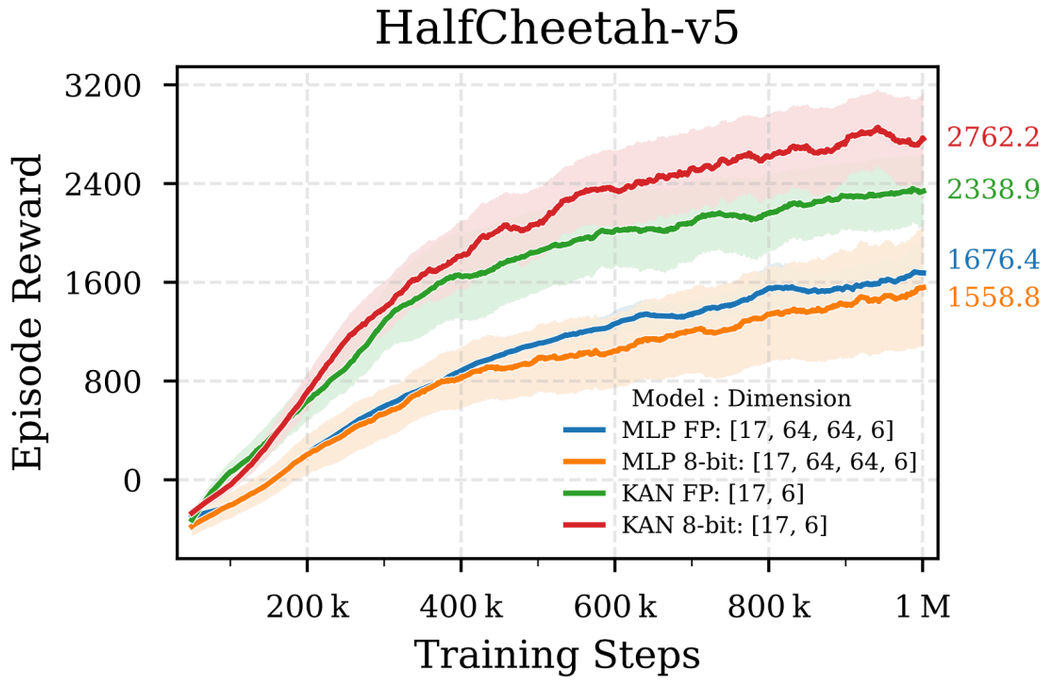
FPGA logic makes architecture compelling

Impact

Latency < 70ns

Reduction using quark/gluon tagger





Better Physics means better overall performance everywhere

Work recognized by whole FPGA community

Symbolic Methods

KANELÉ vastly outperforms *hls4ml* on MLPerf Tiny, AD task

Model	AUC	BRAM (36kb)	DSP	FF	LUT	LUTRAM	II (clocks)	Throughput (inf/s)	Latency (μs)	Energy/inf. (μJ)
KANELÉ	**0.83**	**0**	**0**	**17,643**	**29,981**	**0**	**1**	**228 M**	**0.07**	**0.01**
`hls4ml` (MLPerf Tiny v0.7) [10]	**0.83**	22.5	207	61,639	51,429	5,780	144	694 k	45	98.4

≦1/2 FF and LUTs

>10^3 x throughput, latency, energy improvements

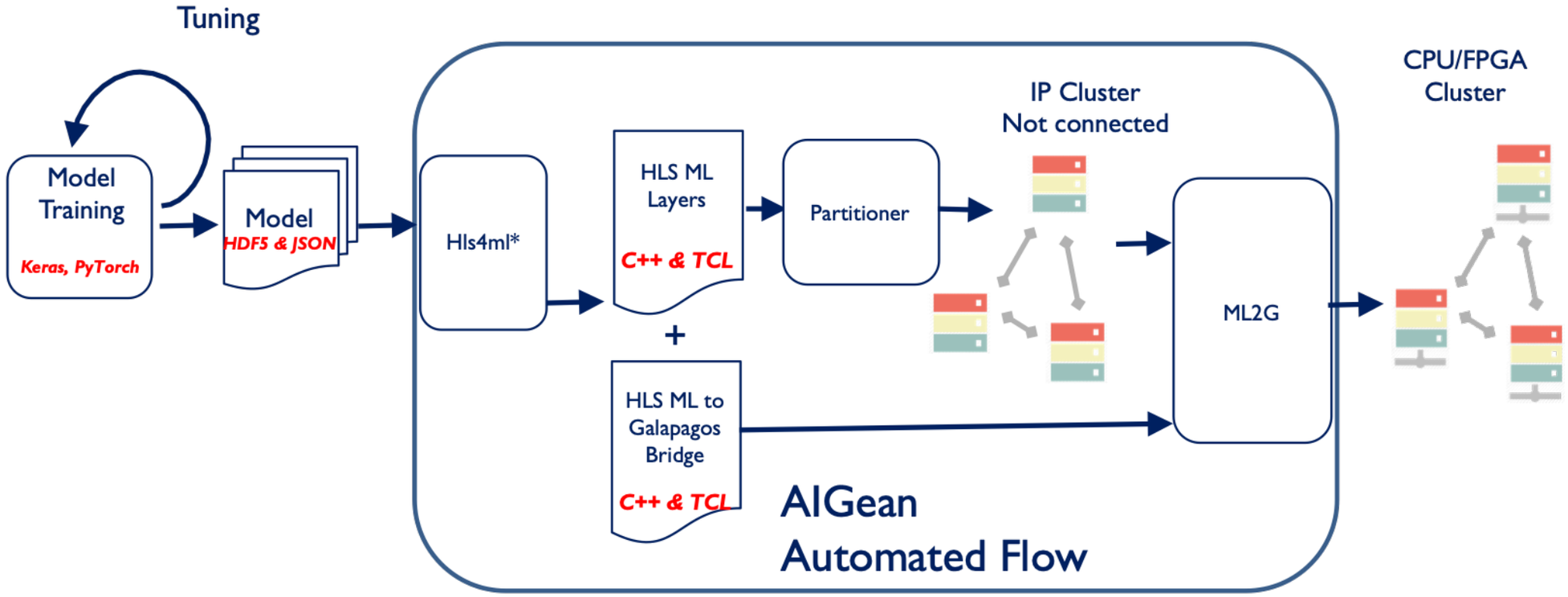
Context: ToyCar (MLPerf Tiny) is a complex audio anomaly detection task

Symbolic form (Physics) dramatically outproduces conventional ML architectures
This is the lowest power NN on any chip!

HLS4ML & The World

Paradigm of unrolling NN on chip makes it Fast

HLS4ML been adapted to solve many problems



50 Million Parameter NN deployed on 20 FPGAs

Fastest batch 1 Image algo in world

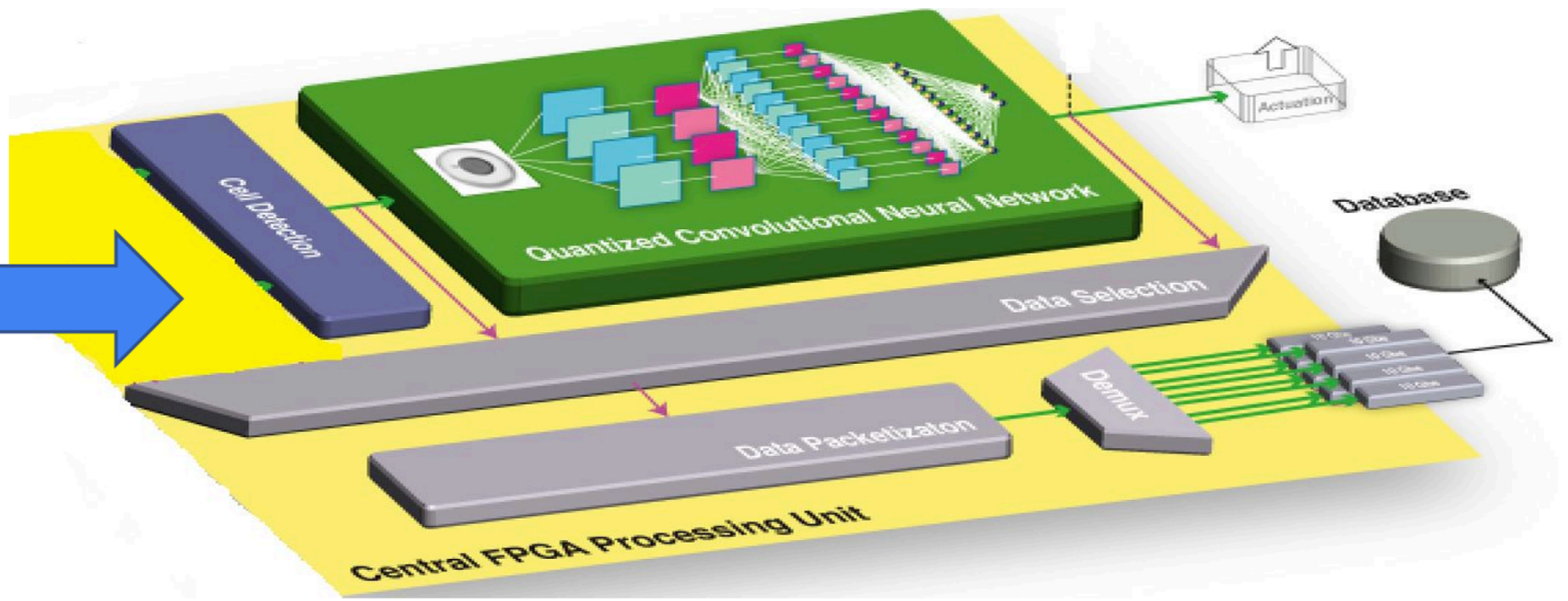
HLS4ML & The World

Paradigm of unrolling NN on chip makes it Fast

HLS4ML been adapted to solve many problems

Industry sized NNs distributing across chips

Cell Therapy

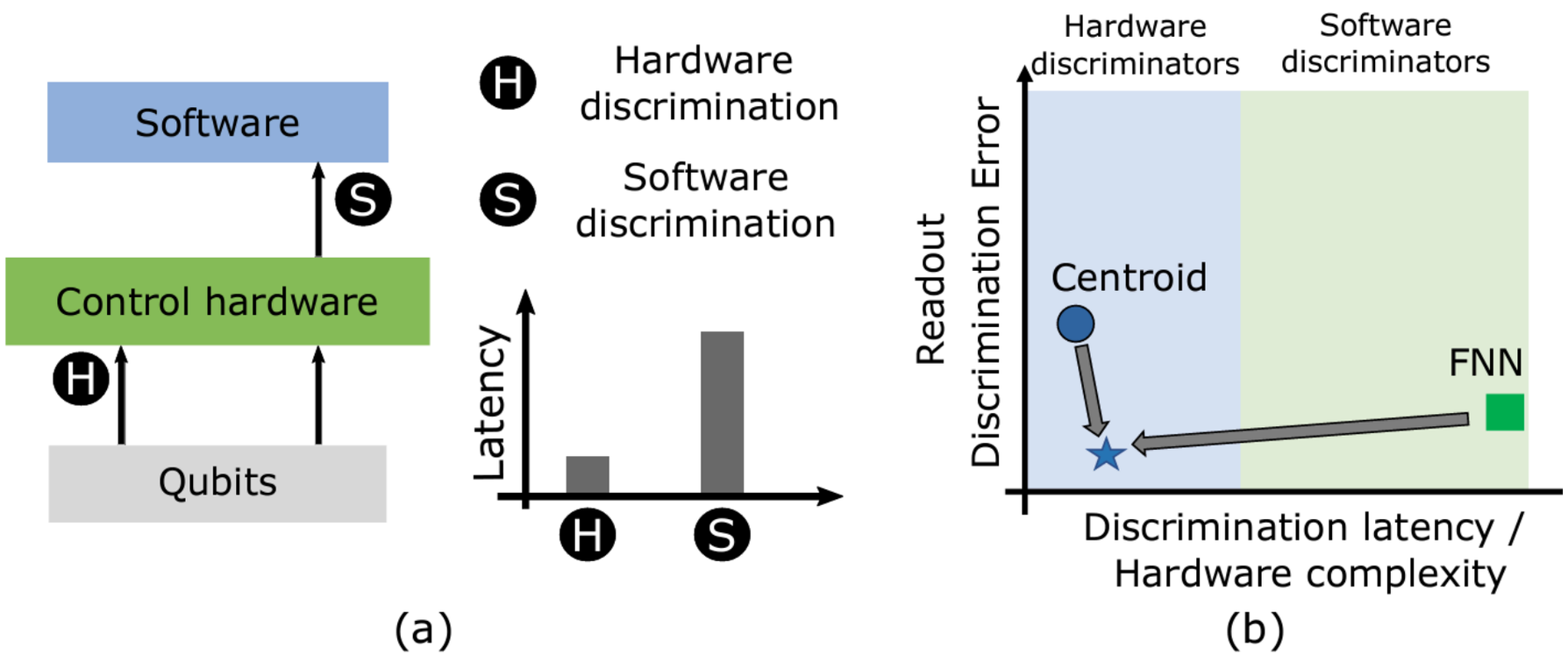


HLS4ML & The World

Paradigm of unrolling NN on chip makes it Fast

HLS4ML been adapted to solve many problems

Quantum Computing

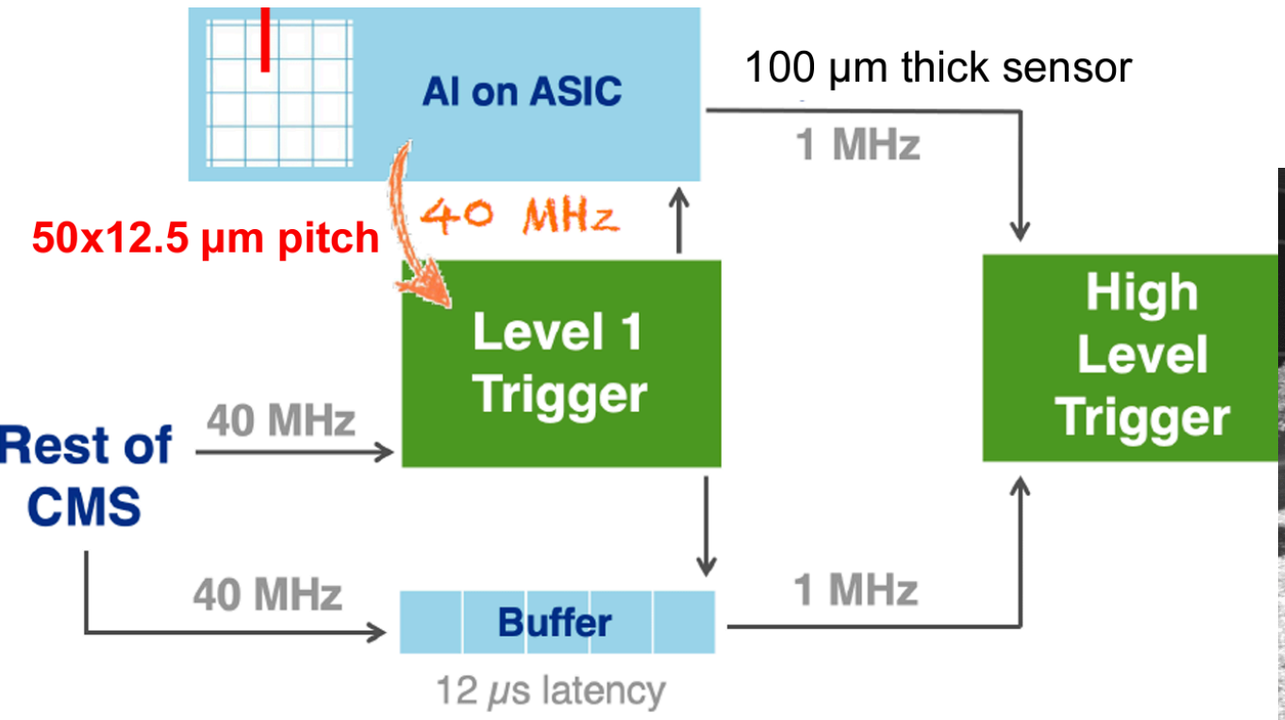


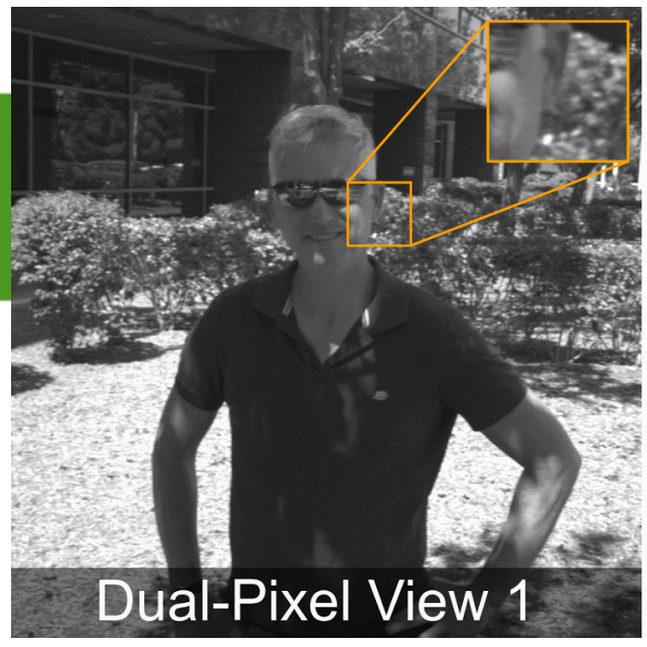
HLS4ML & The World

Paradigm of unrolling NN on chip makes it Fast

HLS4ML been adapted to solve many problems

Silicon Pixels



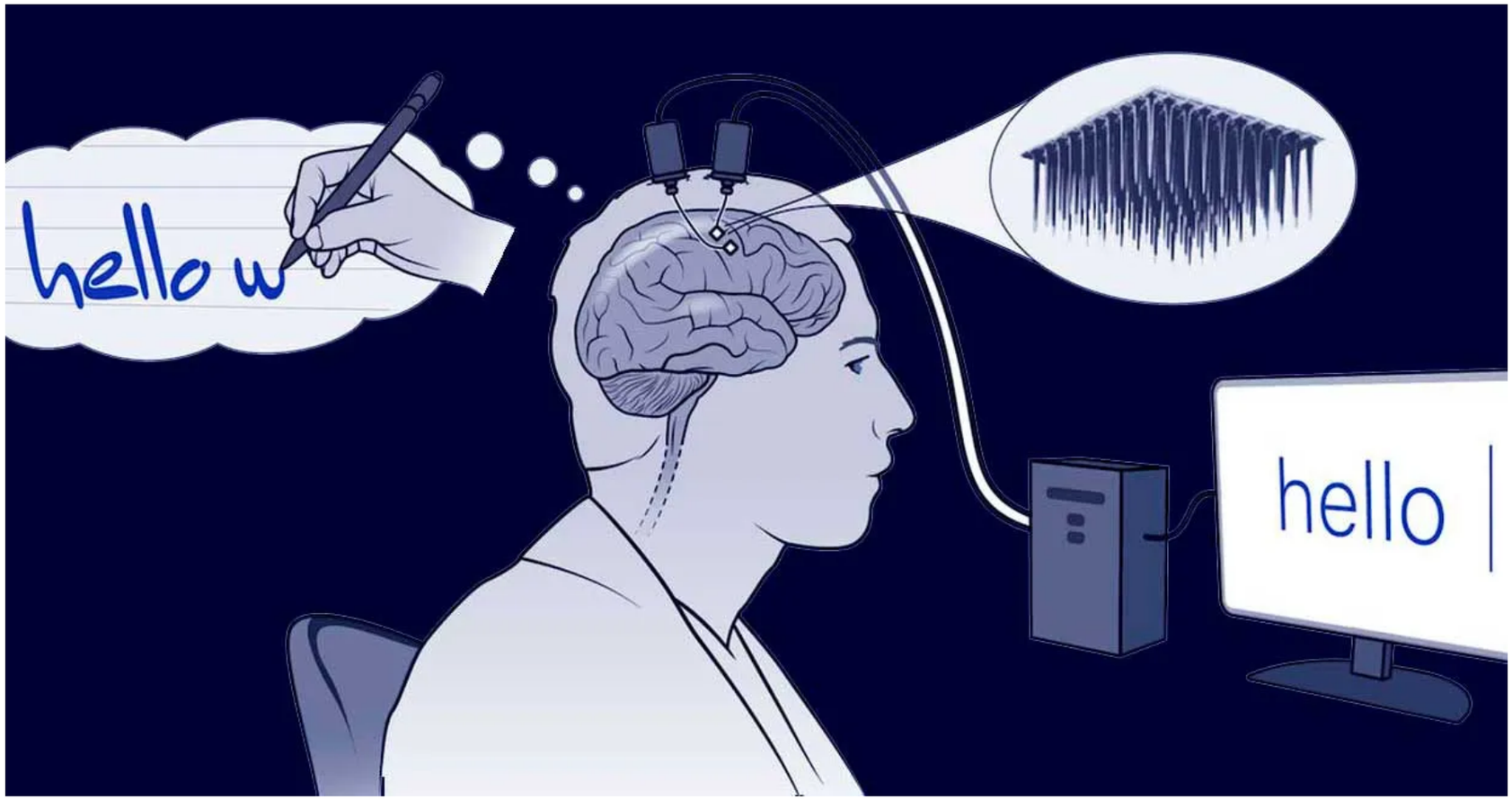


HLS4ML & The World

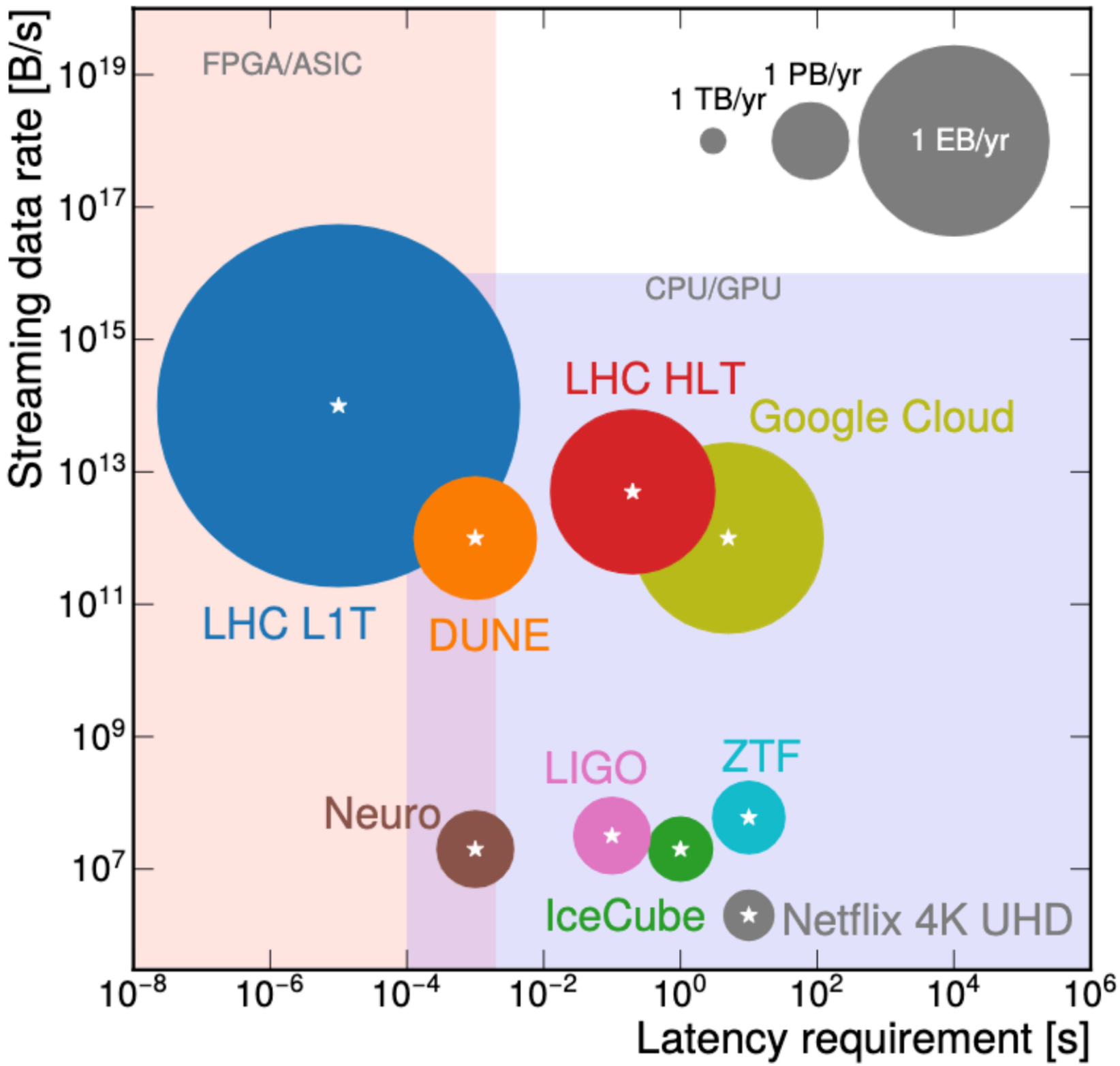
Paradigm of unrolling NN on chip makes it Fast

HLS4ML been adapted to solve many problems

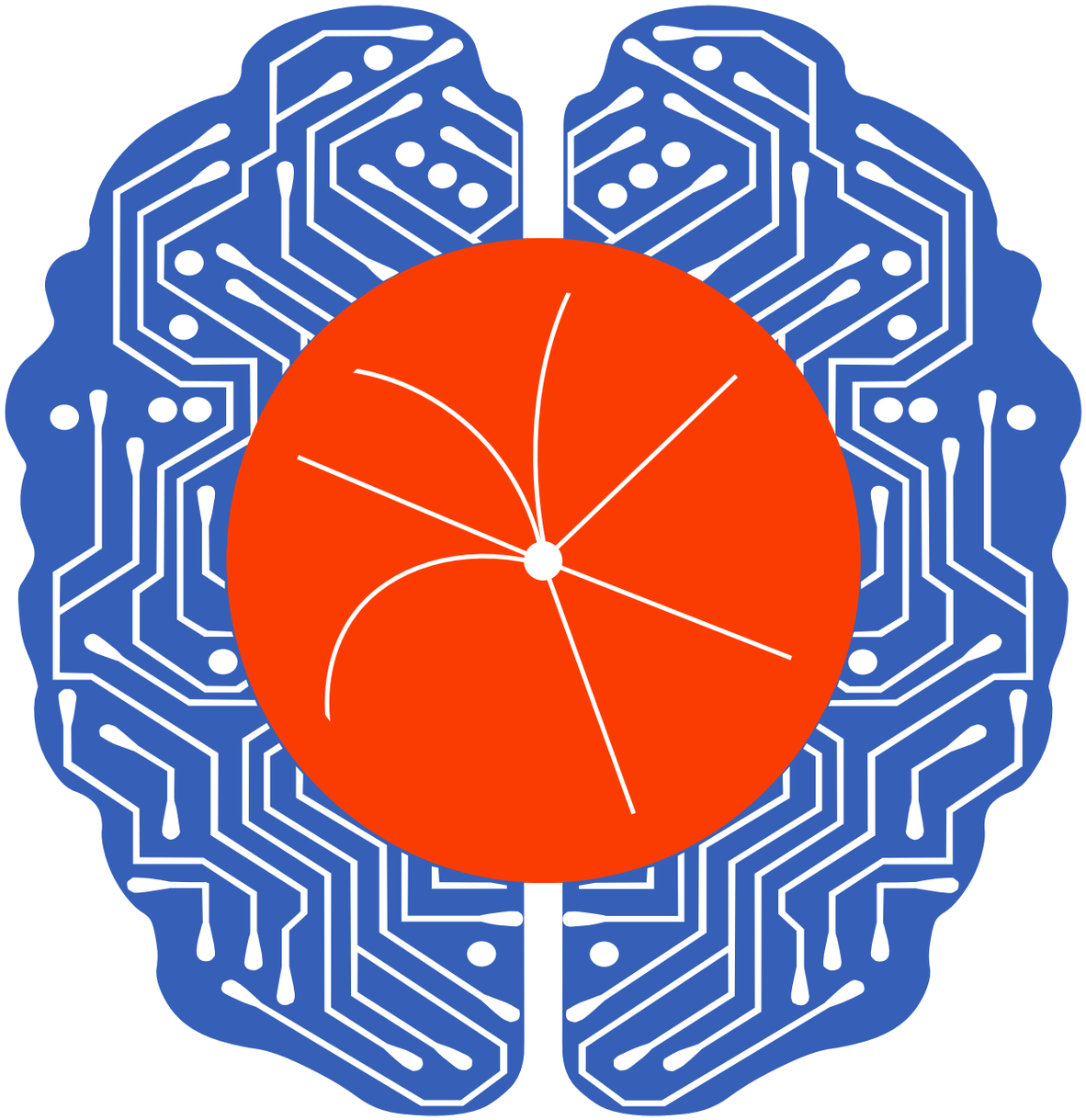
Brain Readout

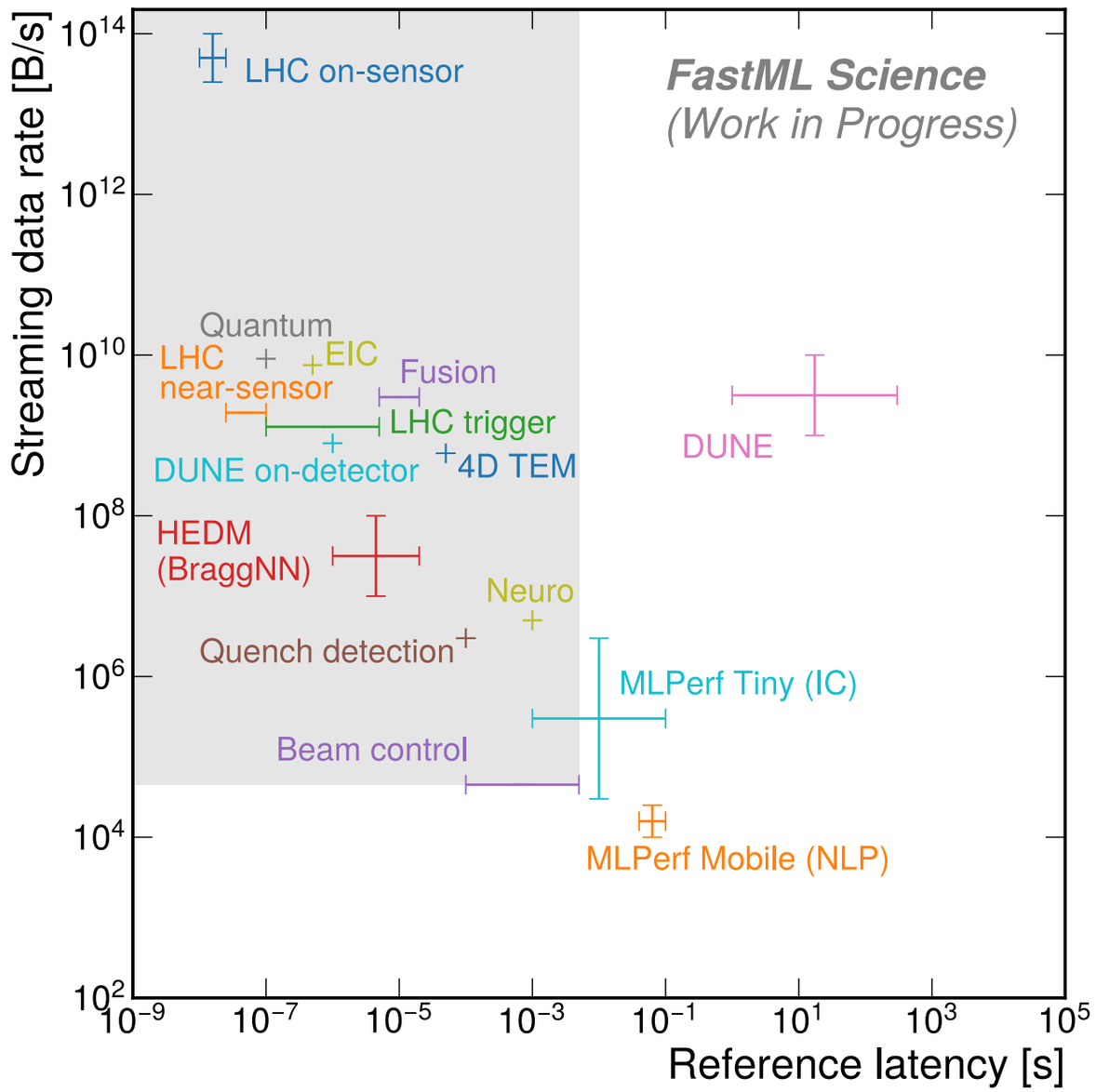


Preparing for the future



Come and Join FastML !





Opening doors for real-time AI in HEP and many other fields