

Collide & Compress

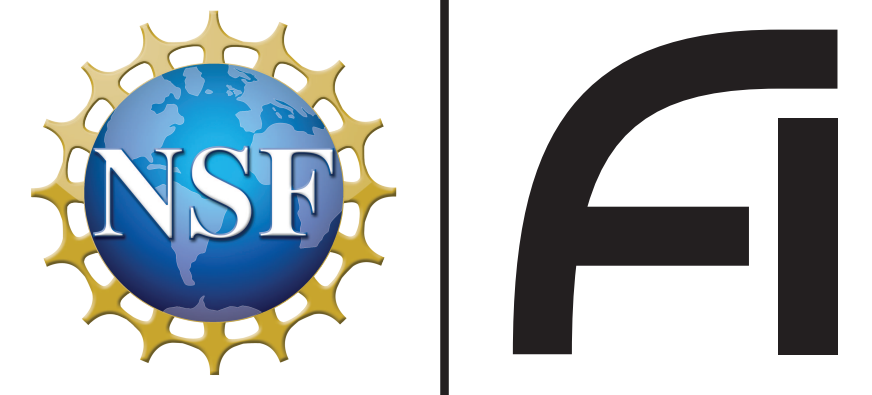
Building robust embedding spaces to simplify new physics searches at the LHC

Sam Bright-Thonney

IAIFI/MIT

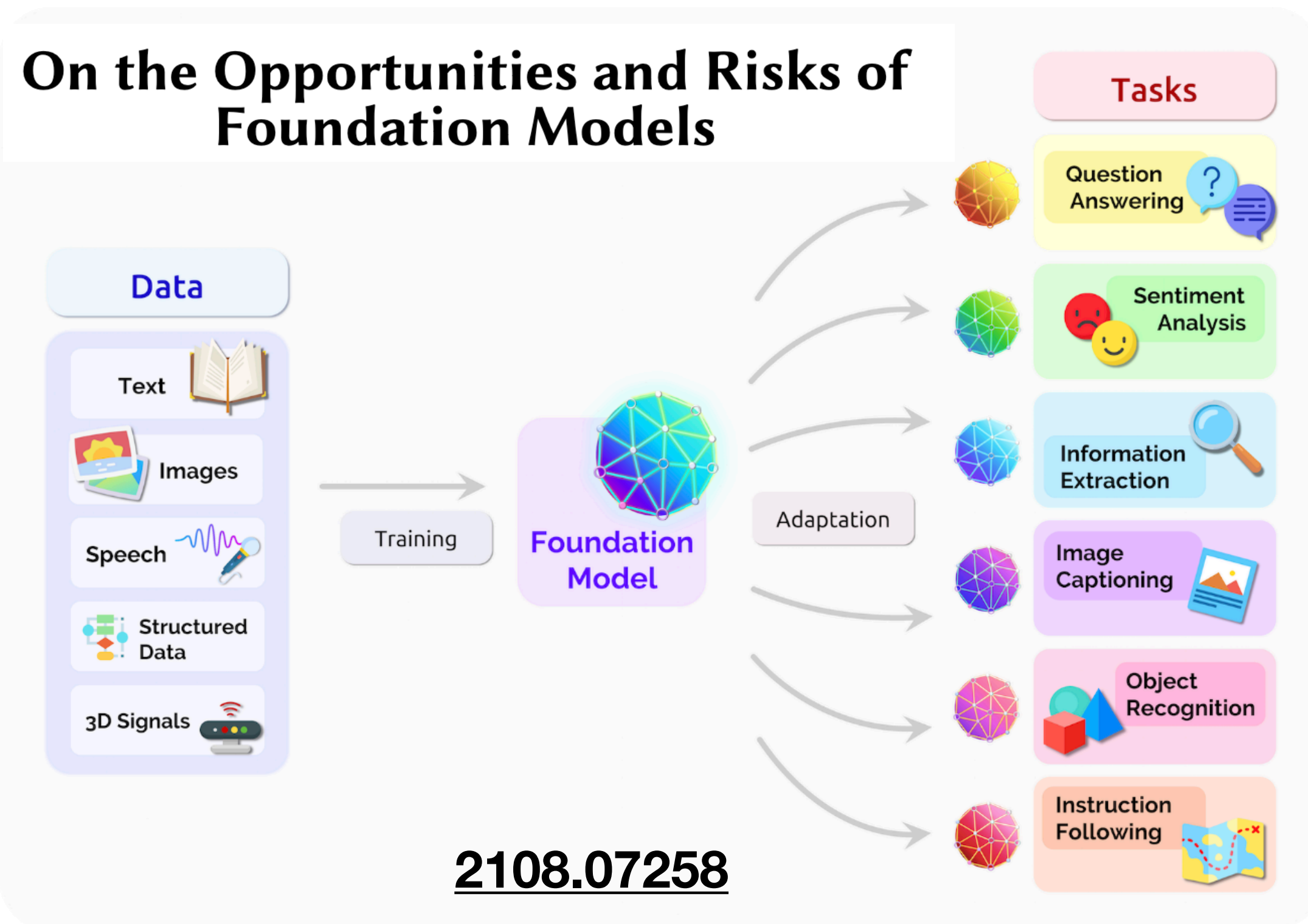
IAIFI Colloquium

February 14th, 2025



The age of foundation models

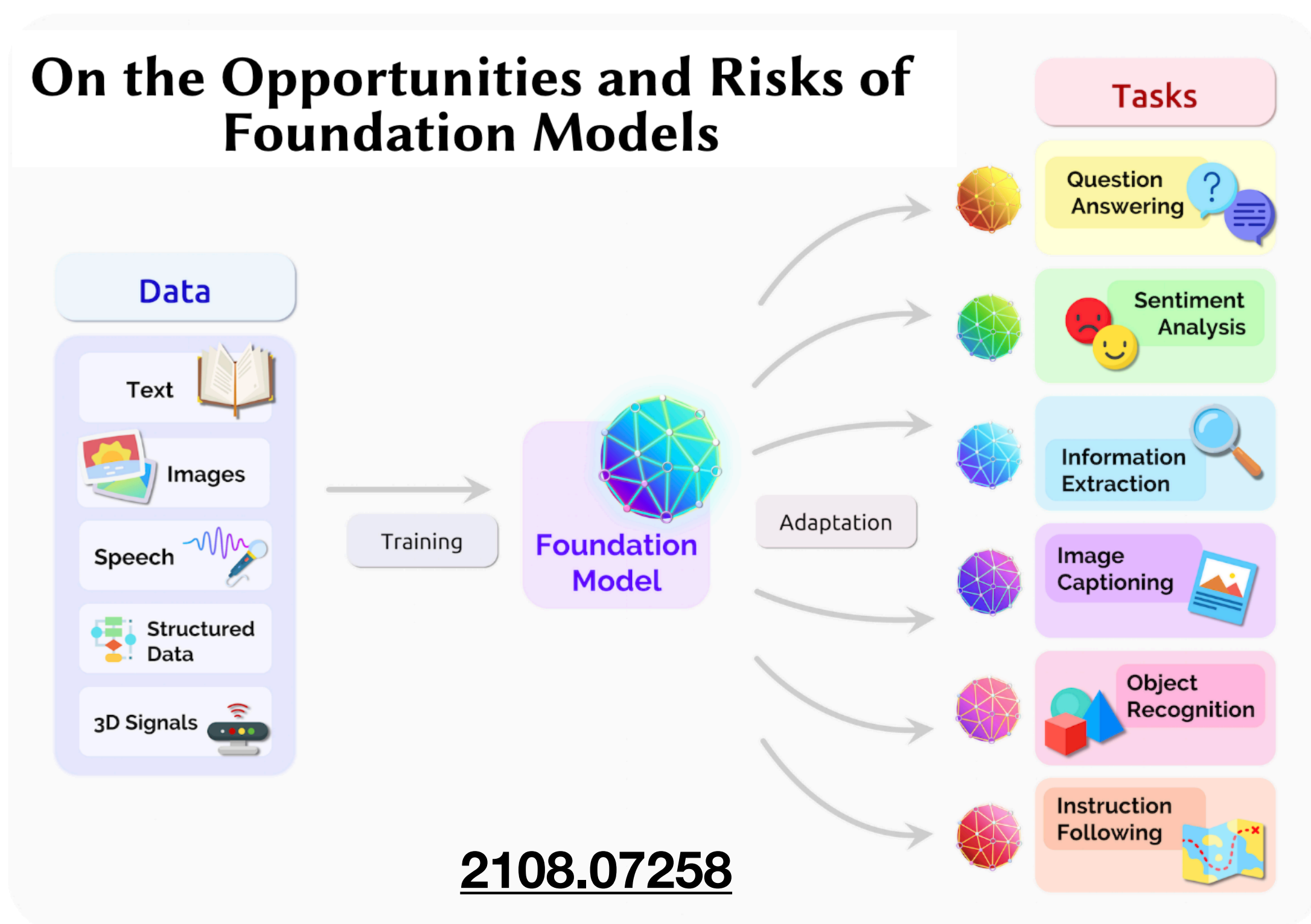
We are very much in the age of (hype about) “foundation models”



“...any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks”

The age of foundation models

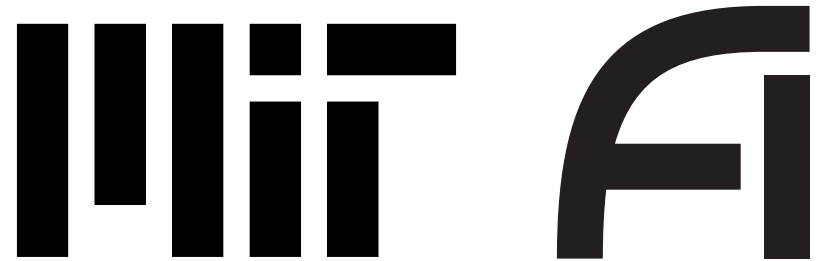
We are very much in the age of (hype about) “foundation models”



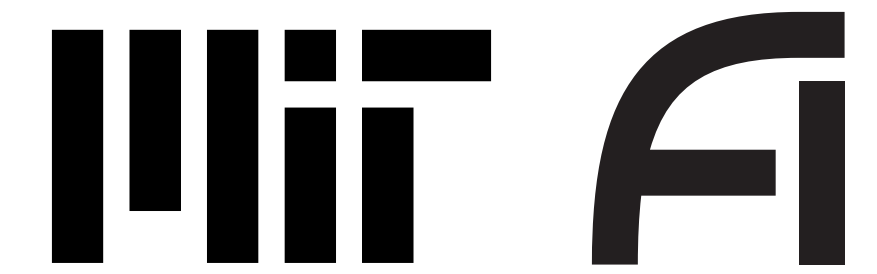
“...any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks”



Language isn't everything



Language isn't everything



- Mastering language is *viscerally* impressive, but LLMs completely dominate the conversation

Language isn't everything

- Mastering language is *viscerally* impressive, but LLMs completely dominate the conversation
- A “model that understands X” where X = a scientific domain can be **extremely valuable!**
 - ▶ Operating in complexity regimes where *our* abilities are severely limited

Article | [Open access](#) | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

[John Jumper](#) , [Richard Evans](#), [Alexander Pritzel](#), [Tim Green](#), [Michael Figurnov](#), [Olaf Ronneberger](#), [Kathryn Tunyasuvunakool](#), [Russ Bates](#), [Augustin Žídek](#), [Anna Potapenko](#), [Alex Bridgland](#), [Clemens Meyer](#), [Simon A. A. Kohl](#), [Andrew J. Ballard](#), [Andrew Cowie](#), [Bernardino Romera-Paredes](#), [Stanislav Nikolov](#), [Rishub Jain](#), [Jonas Adler](#), [Trevor Back](#), [Stig Petersen](#), [David Reiman](#), [Ellen Clancy](#), [Michal Zielinski](#), ... [Demis Hassabis](#)  [+ Show authors](#)

[Nature](#) 596, 583–589 (2021) | [Cite this article](#)

1.99m Accesses | 20k Citations | 4006 Altmetric | [Metrics](#)

Demis Hassabis

“for protein structure prediction”



© Nobel Prize Outreach. Photo: Clément Morin

John Jumper

“for protein structure prediction”



© Nobel Prize Outreach. Photo: Clément Morin



Language isn't everything

- Mastering language is *viscerally* impressive, but LLMs completely dominate the conversation
- A “model that understands X” where X = a scientific domain can be **extremely valuable!**
 - ▶ Operating in complexity regimes where *our* abilities are severely limited

Article | [Open access](#) | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

[John Jumper](#) , [Richard Evans](#), [Alexander Pritzel](#), [Tim Green](#), [Michael Figurnov](#), [Olaf Ronneberger](#), [Kathryn Tunyasuvunakool](#), [Russ Bates](#), [Augustin Žídek](#), [Anna Potapenko](#), [Alex Bridgland](#), [Clemens Meyer](#), [Simon A. A. Kohl](#), [Andrew J. Ballard](#), [Andrew Cowie](#), [Bernardino Romera-Paredes](#), [Stanislav Nikolov](#), [Rishub Jain](#), [Jonas Adler](#), [Trevor Back](#), [Stig Petersen](#), [David Reiman](#), [Ellen Clancy](#), [Michal Zieliński](#), ... [Demis Hassabis](#)  + Show authors

[Nature](#) 596, 583–589 (2021) | [Cite this article](#)

1.99m Accesses | 20k Citations | 4006 Altmetric | [Metrics](#)

Demis Hassabis

“for protein structure prediction”



© Nobel Prize Outreach. Photo: Clément Morin

John Jumper

“for protein structure prediction”





© Nobel Prize Outreach. Photo: Clément Morin



Article | [Open access](#) | Published: 29 November 2023

Scaling deep learning for materials discovery

[Amil Merchant](#) , [Simon Batzner](#), [Samuel S. Schoenholz](#), [Muratahan Aykol](#), [Gwoon Cheon](#) & [Ekin Dogus Cubuk](#) 

[Nature](#) 624, 80–85 (2023) | [Cite this article](#)

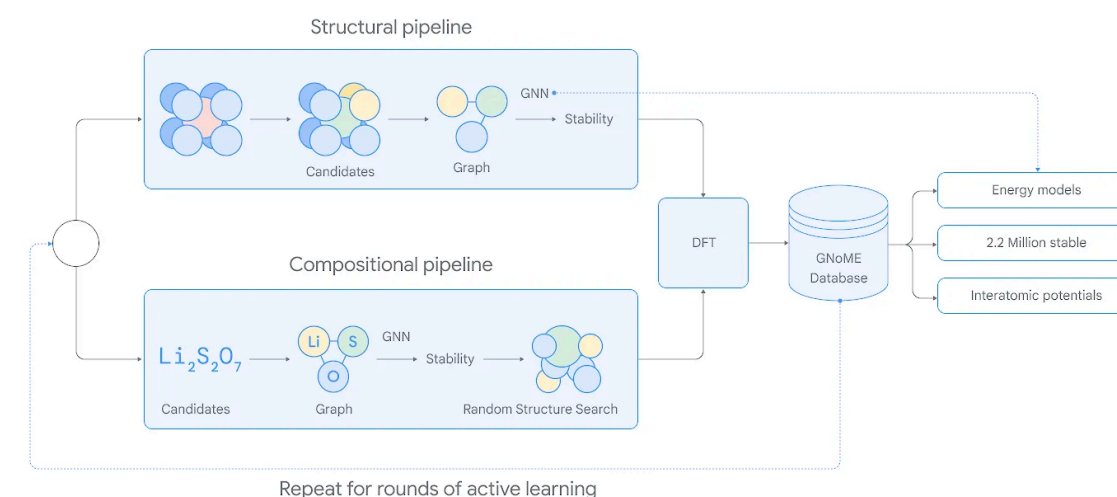
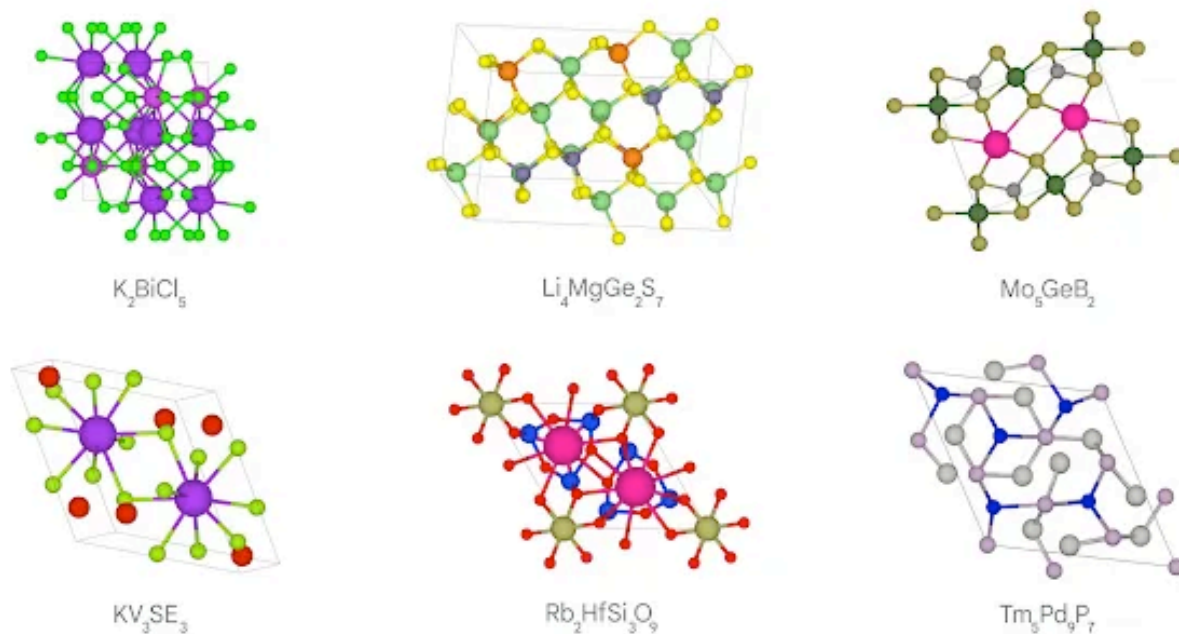



Image credit: DeepMind

Article | Published: 16 January 2025

A generative model for inorganic materials design

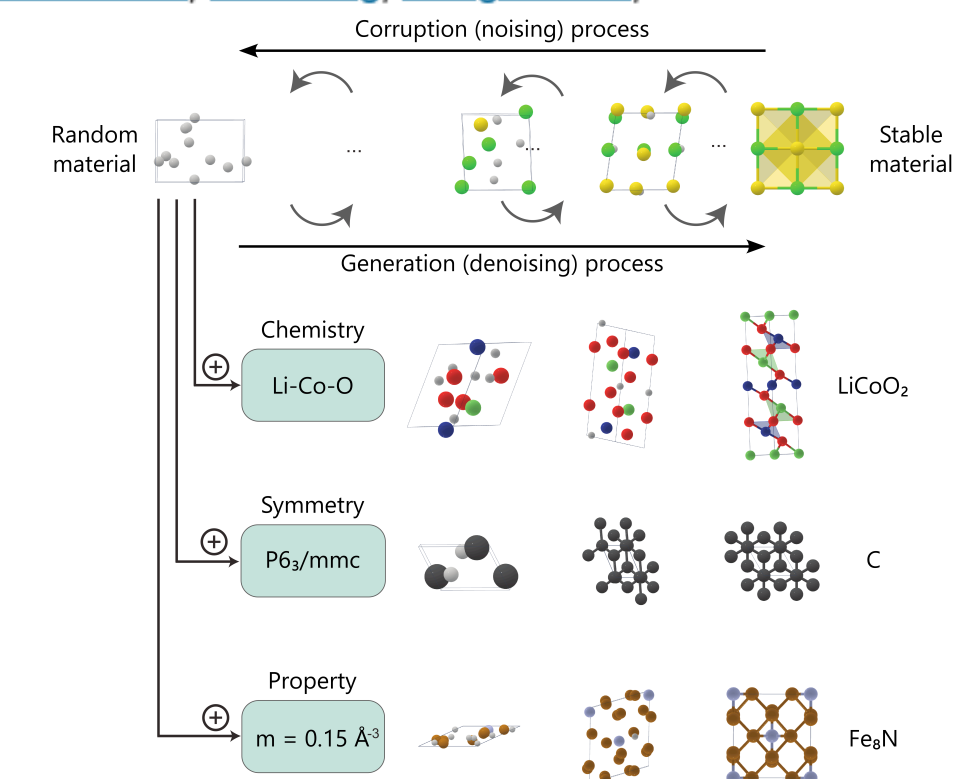
[Claudio Zeni](#), [Robert Pinsler](#), [Daniel Zügner](#), [Andrew Fowler](#), [Matthew Horton](#), [Xiang Fu](#), [Zilong Wang](#), [Aliaksandra Shysheya](#), [Jonathan Crabbé](#), [Shoko Ueda](#), [Roberto Sordillo](#), [Lixin Sun](#), [Jake Smith](#), [Bichlien Nguyen](#), [Hannes Schulz](#), [Sarah Lewis](#), [Chin-Wei Huang](#), [Ziheng Lu](#), [Yichi Zhou](#), [Han Yang](#), [Hongxia Hao](#), [Jielan Li](#), [Chunlei Yang](#), [Wenjie Li](#), ... [Tian Xie](#)  + Show authors

[Nature](#) (2025) | [Cite this article](#)

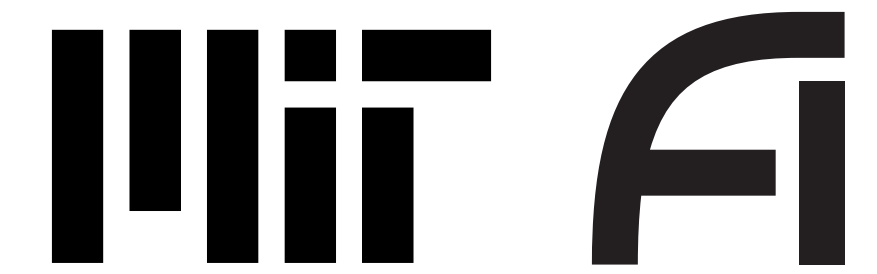
78k Accesses | 147 Altmetric | [Metrics](#)



Image credit: Microsoft



Models that understand physics



I'd like a foundation model that understands particle physics

Models that understand physics

I'd like a foundation model that understands particle physics

give me a concise but technical explanation of asymptotic freedom in QCD

Asymptotic freedom in Quantum Chromodynamics (QCD) refers to the property that the strong coupling constant α_s decreases at high energy (or short distances), causing quarks and gluons to interact **weakly** at high energies but **strongly** at low energies, leading to confinement.

Technical Explanation

1. Beta Function & Running Coupling:

The behavior of the QCD coupling $\alpha_s = g_s^2/(4\pi)$ is governed by the **renormalization group equation**:

$$\mu \frac{d\alpha_s}{d\mu} = \beta(\alpha_s),$$

where the **QCD beta function** at one-loop is:

$$\beta(\alpha_s) = -\frac{\beta_0}{2\pi} \alpha_s^2, \quad \text{with } \beta_0 = 11 - \frac{2}{3} N_f.$$

Here, N_f is the number of active quark flavors, and for $N_f \leq 16$, we get $\beta_0 > 0$, implying **asymptotic freedom**.

2. Energy Dependence of α_s :

Integrating the beta function gives the **running coupling**:

$$\alpha_s(\mu) \approx \frac{4\pi}{\beta_0 \ln(\mu^2/\Lambda_{\text{QCD}}^2)},$$

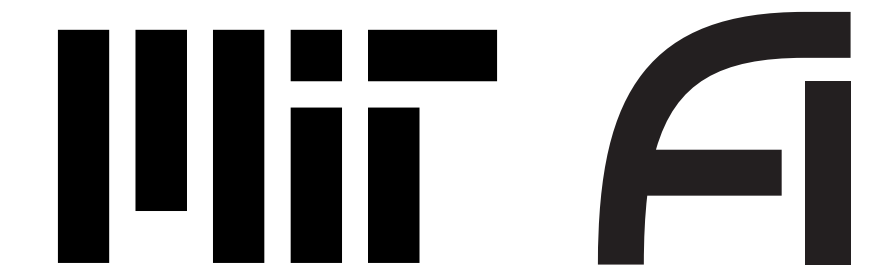
where $\Lambda_{\text{QCD}} \sim 200$ MeV is the QCD scale parameter.

- At **high energy** ($\mu \gg \Lambda_{\text{QCD}}$), α_s **decreases**, meaning quarks behave as **free particles** (asymptotic freedom).
- At **low energy** ($\mu \sim \Lambda_{\text{QCD}}$), α_s **increases**, leading to **quark confinement** into hadrons.



thanks for listening!

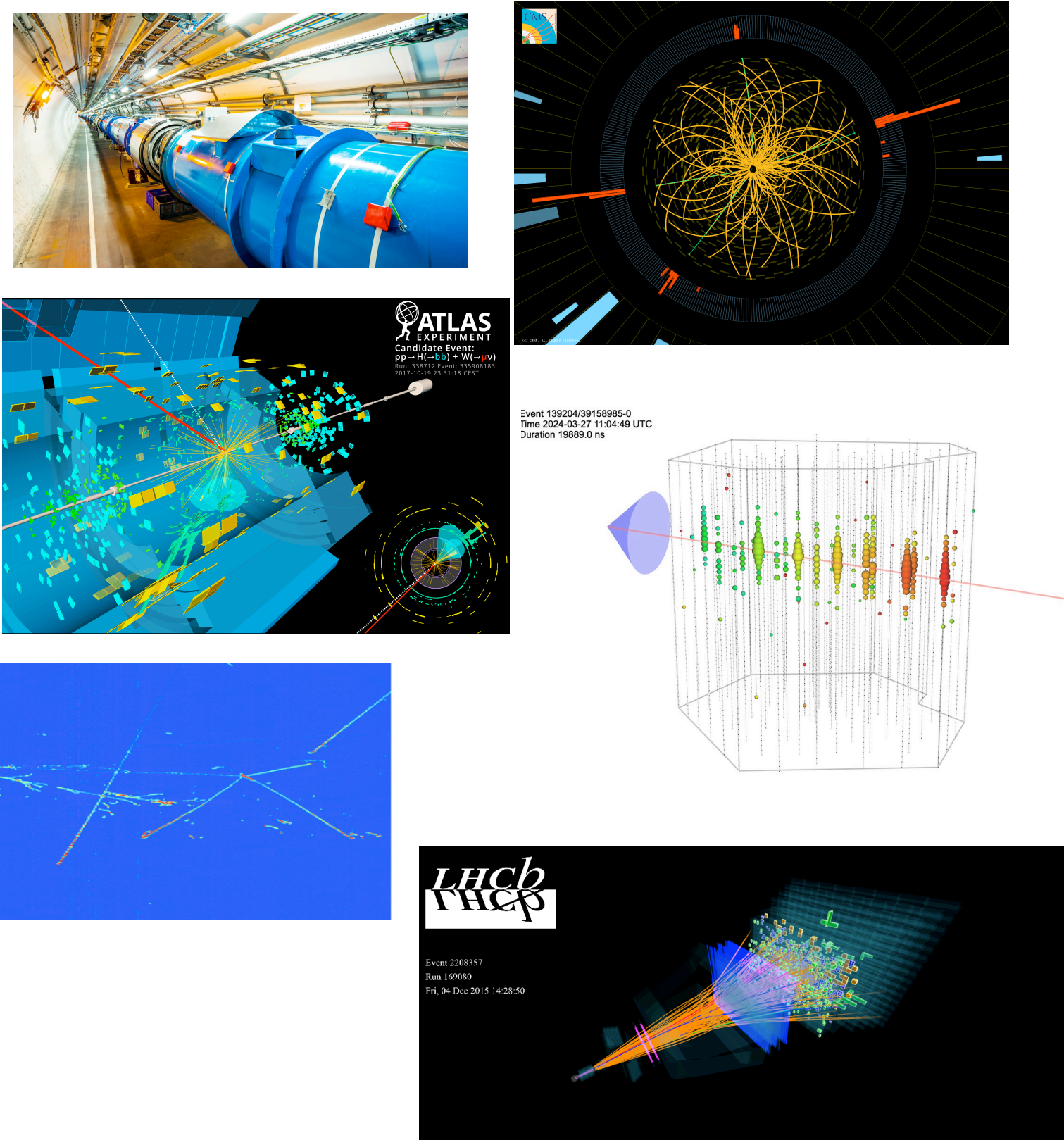
Models that understand physics



I'd like a foundation model that understands particle physics ***data!***

Models that understand physics

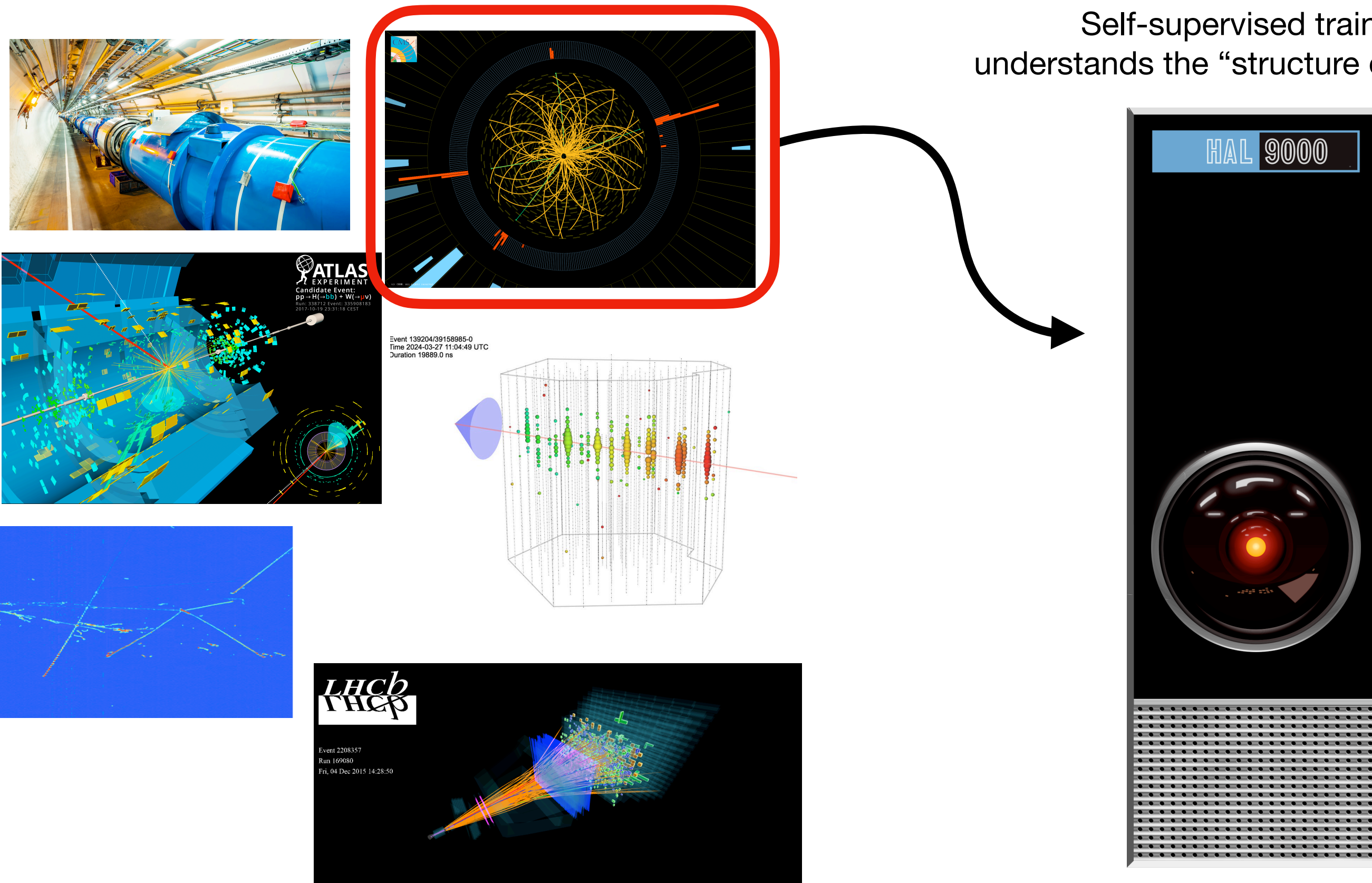
I'd like a foundation model that understands particle physics *data!*



Experimental data
(hundreds/thousands of channels)

Models that understand physics

I'd like a foundation model that understands particle physics *data!*



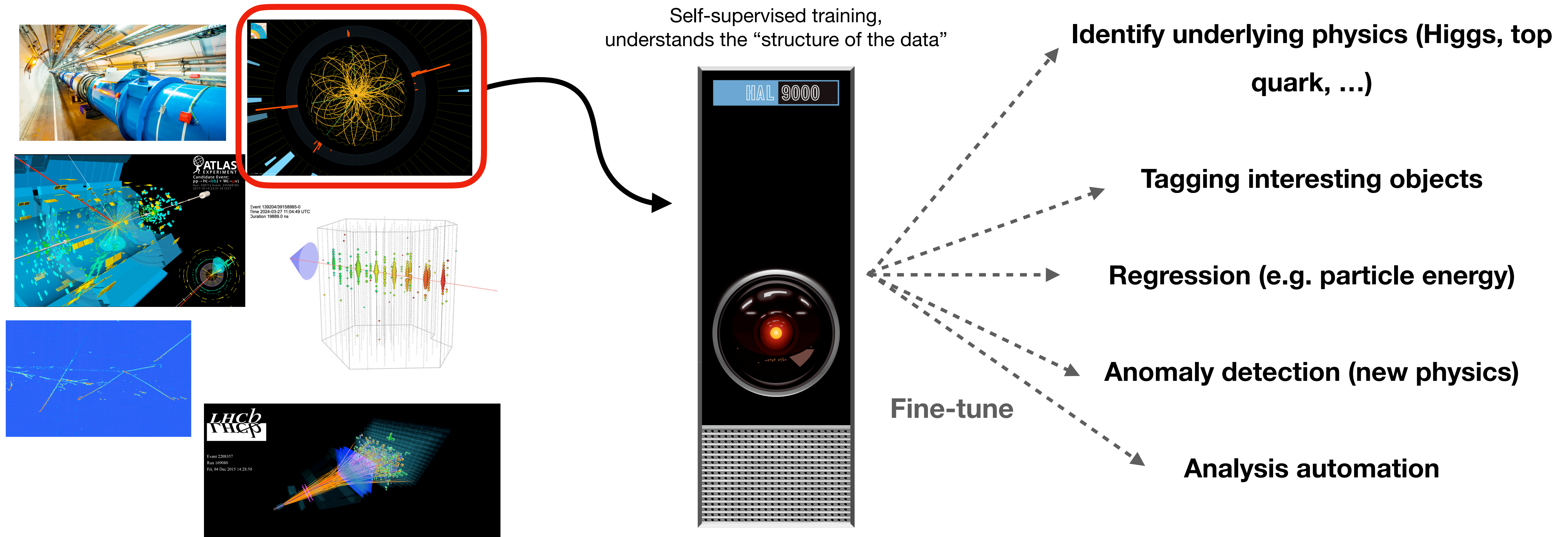
Experimental data
(hundreds/thousands of channels)

Foundation model

*experiment-specific

Models that understand physics

I'd like a foundation model that understands particle physics **data!**



Experimental data
(hundreds/thousands of channels)

Foundation model

*experiment-specific

Downstream tasks

Understanding physics data

If LLMs can write, code, and “reason”, can they also do my job?

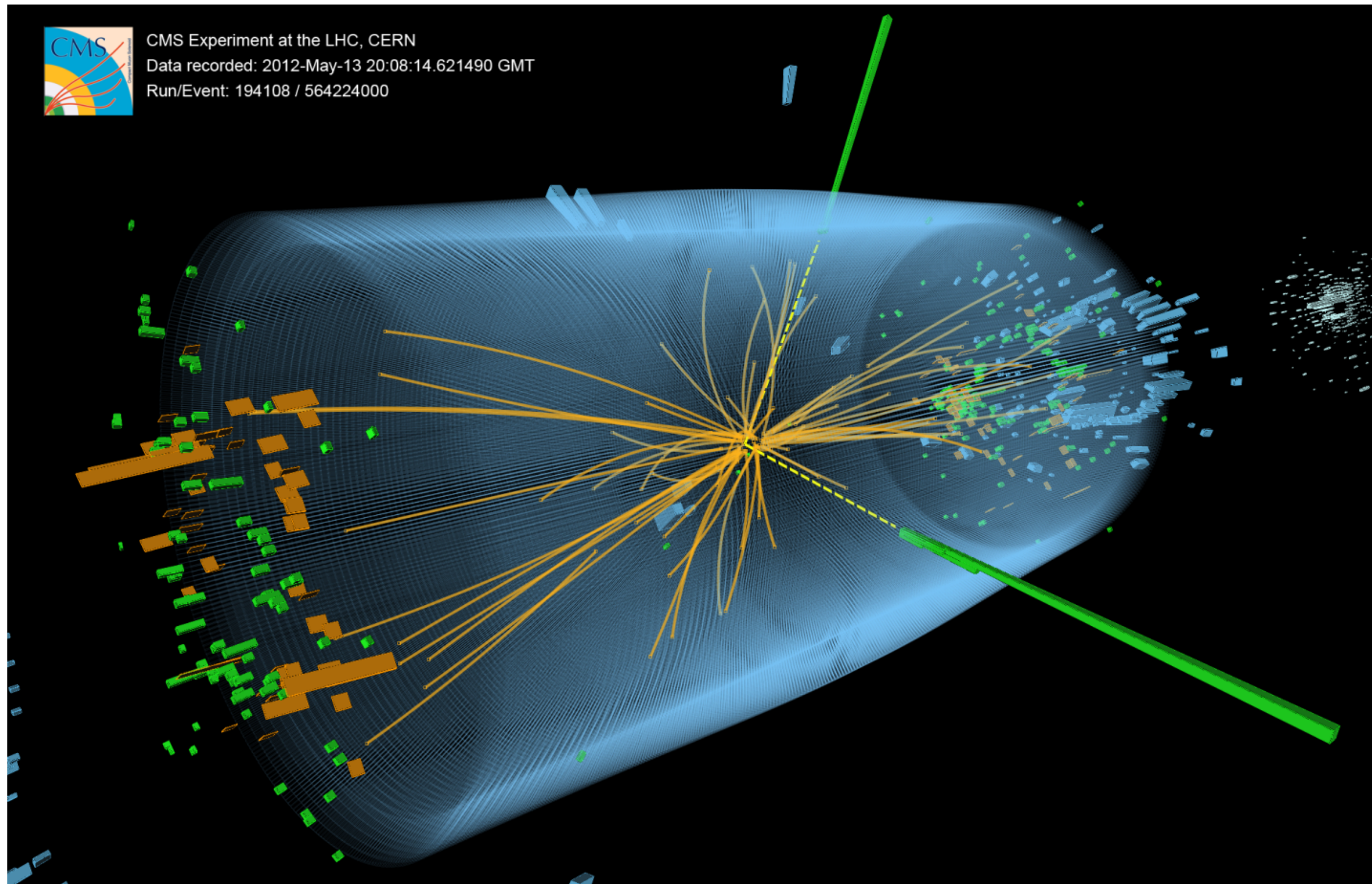


Image credit: CERN

$$H \rightarrow \gamma\gamma$$

Understanding physics data

If LLMs can write, code, and “reason”, can they also do my job?

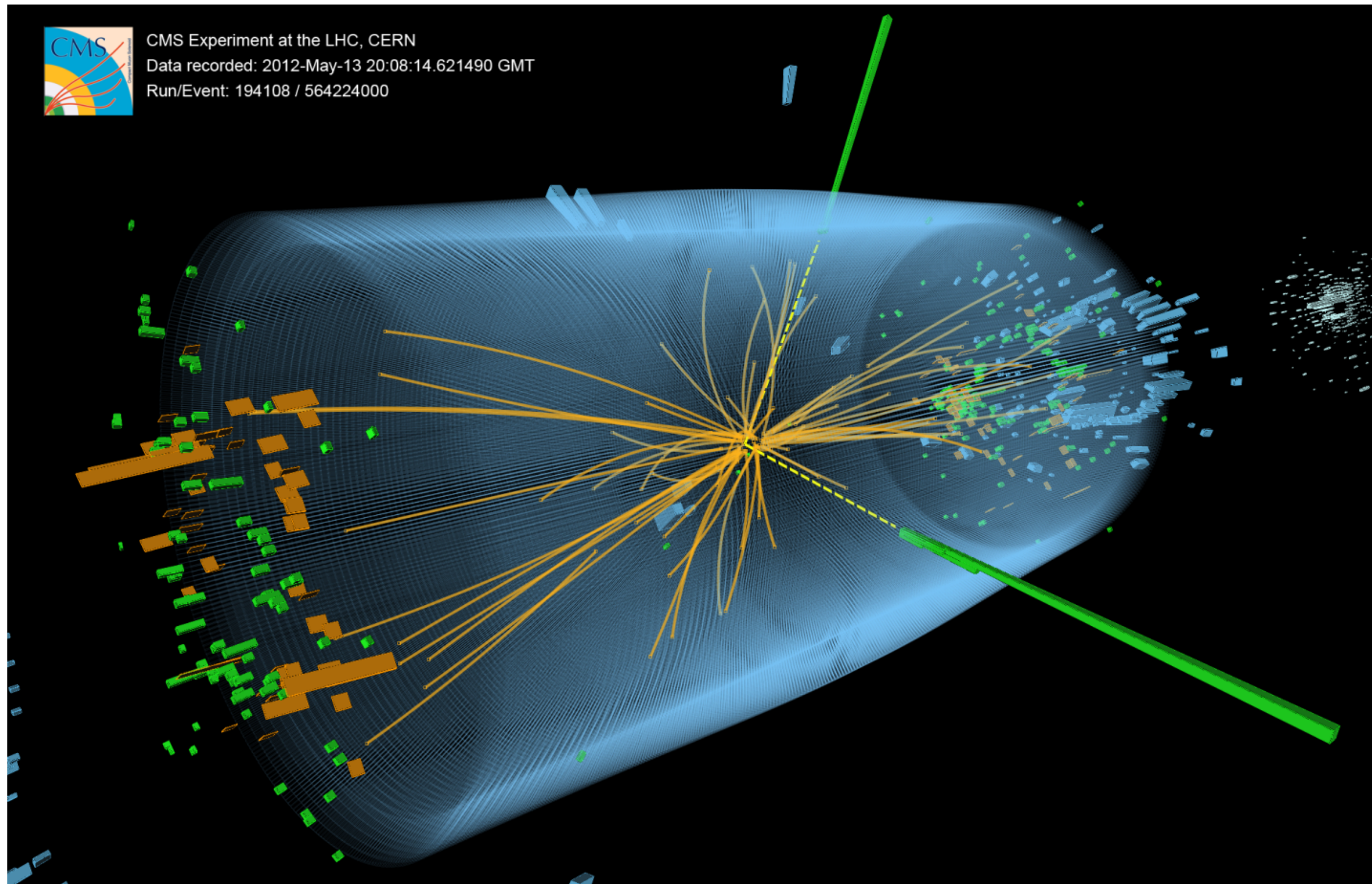
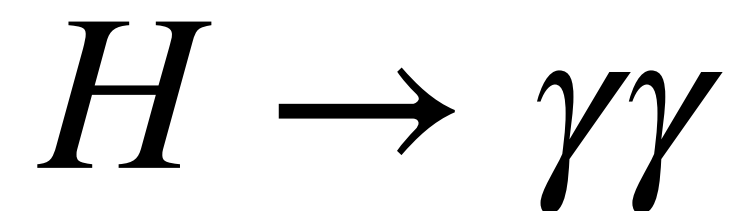


Image credit: CERN



what is going on in this picture

This is an event display from the **CMS Experiment** at the **Large Hadron Collider (LHC)**. Such displays visualize the data collected from particle collisions inside the detector, showing the trajectories of particles and energy deposits as they are reconstructed.

Key Features of the Image:

1. Collision Point (Interaction Vertex):

- The orange lines emanating from the center represent the **tracks of charged particles** produced in the collision. The point where these lines originate is the **interaction vertex**, where two protons collided.

2. Yellow Dashed Line:

- Indicates the possible trajectory or momentum imbalance associated with particles not directly detected (e.g., **neutrinos** or missing transverse energy).

3. Green and Orange Rectangles:

- These are **energy deposits** in the calorimeters:
 - **Orange regions** represent energy deposited by particles in the electromagnetic calorimeter (ECAL), typically photons or electrons.
 - **Green regions** represent energy deposited in the hadronic calorimeter (HCAL), typically from hadrons like protons, neutrons, or pions.

4. Blue Cylinder:

- Represents the **CMS detector geometry**, showing the surrounding tracking systems and calorimeters that measure particles' momenta and energy.

5. Long Green Lines:

- Likely represent **muons**, which penetrate the entire detector and are reconstructed in the outermost layers of the CMS muon chambers.

6. Purpose of the Event:

- This display could represent a **candidate event for a Higgs boson decay** or other interesting physics processes, such as electroweak interactions or rare decays. The clean central vertex and extensive particle activity suggest it is a high-energy event.

Would you like more detailed information about a specific component or the physics process being studied?

Understanding physics data

If LLMs can write, code, and “reason”, can they also do my job?

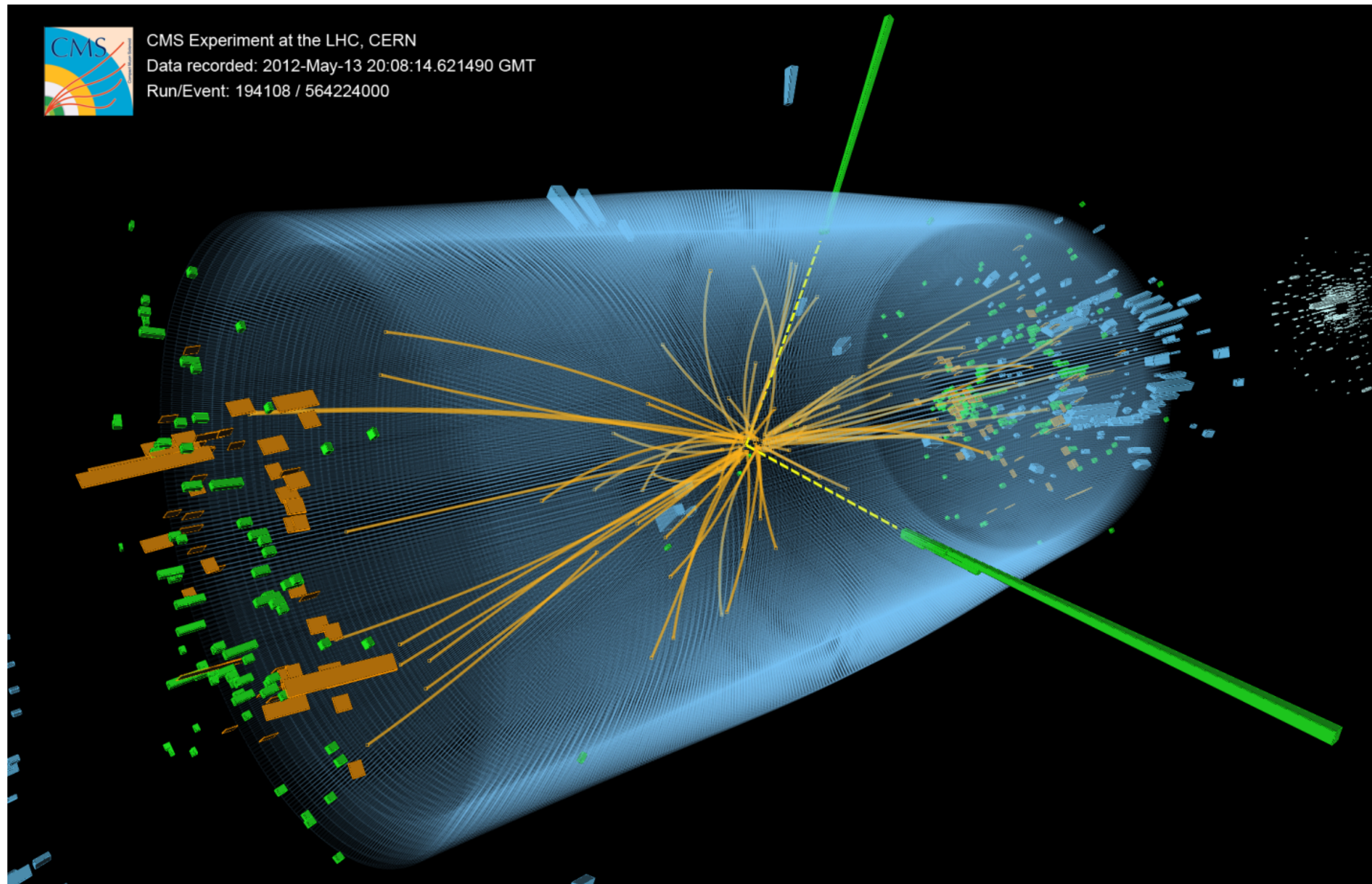
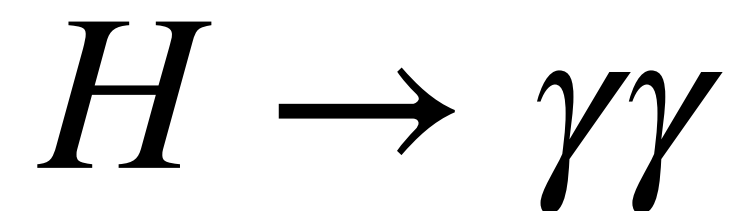


Image credit: CERN



what is going on in this picture

This is an event display from the **CMS Experiment** at the **Large Hadron Collider (LHC)**. Such displays visualize the data collected from particle collisions inside the detector, showing the trajectories of particles and energy deposits as they are reconstructed.

Key Features of the Image:

1. Collision Point (Interaction Vertex):

- The orange lines emanating from the center represent the **tracks of charged particles** produced in the collision. The point where these lines originate is the **interaction vertex**, where two protons collided.

2. Yellow Dashed Line:

- Indicates the possible trajectory or momentum imbalance associated with particles not directly detected (e.g., **neutrinos** or missing transverse energy).

3. Green and Orange Rectangles:

- These are **energy deposits** in the calorimeters:
 - **Orange regions** represent energy deposited by particles in the electromagnetic calorimeter (ECAL), typically photons or electrons.
 - **Green regions** represent energy deposited in the hadronic calorimeter (HCAL), typically from hadrons like protons, neutrons, or pions.

4. Blue Cylinder:

- Represents the **CMS detector geometry**, showing the surrounding tracking systems and calorimeters that measure particles' momenta and energy.

5. Long Green Lines:

- Likely represent **muons**, which penetrate the entire detector and are reconstructed in the outermost layers of the CMS muon chambers.

6. Purpose of the Event:

- This display could represent a **candidate event for a Higgs boson decay** or other interesting physics processes, such as electroweak interactions or rare decays. The clean central vertex and extensive particle activity suggest it is a high-energy event.

Would you like more detailed information about a specific component or the physics process being studied?

Understanding physics data

If LLMs can write, code, and “reason”, can they also do my job?

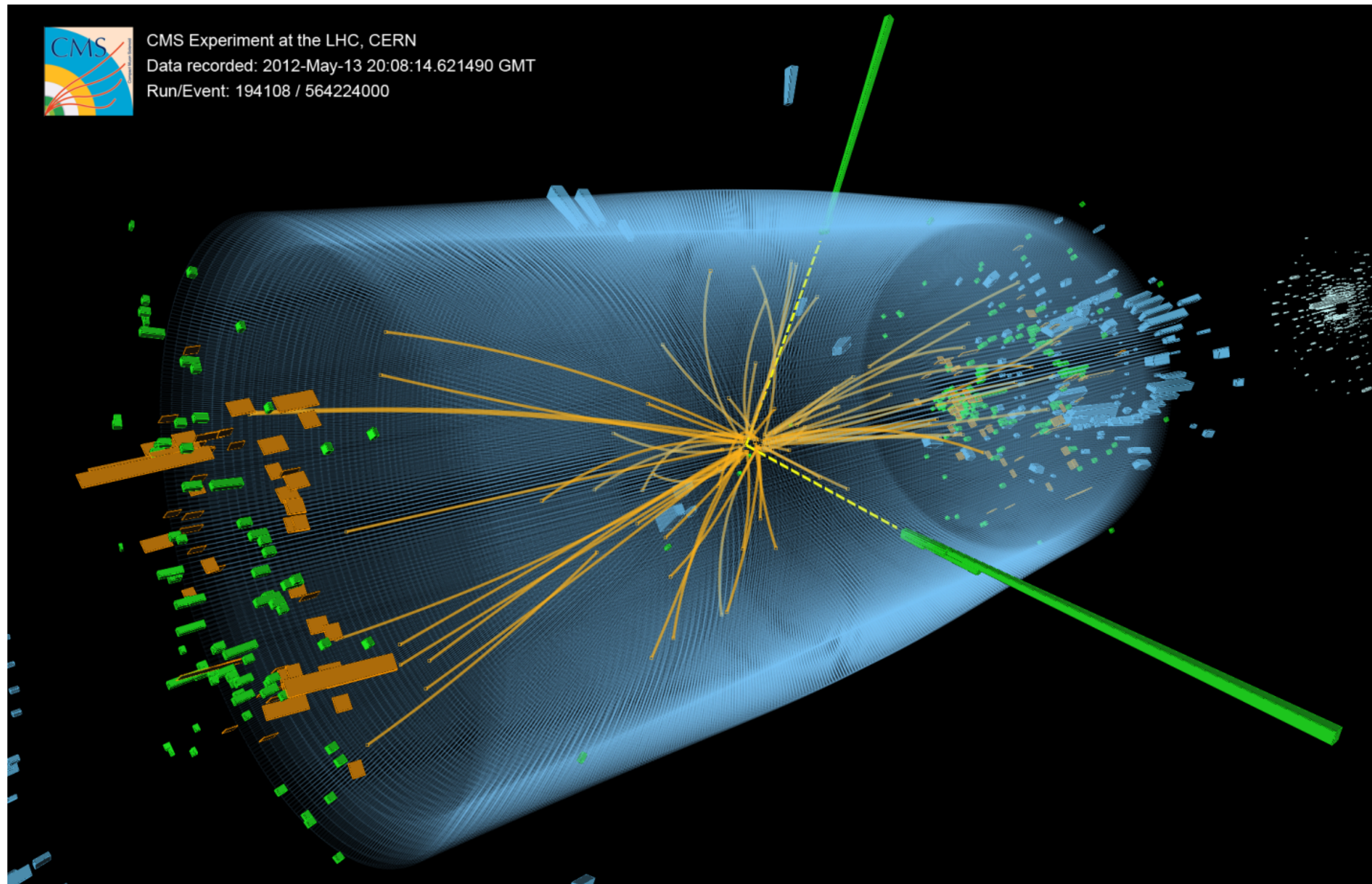
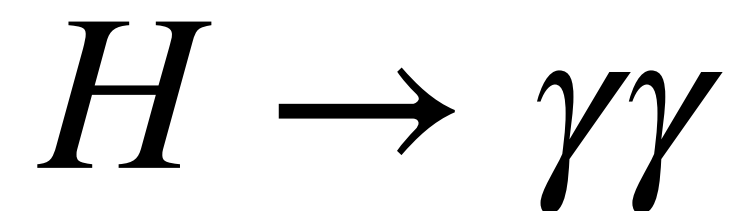


Image credit: CERN



what is going on in this picture

This is an event display from the **CMS Experiment** at the **Large Hadron Collider (LHC)**. Such displays visualize the data collected from particle collisions inside the detector, showing the trajectories of particles and energy deposits as they are reconstructed.

Key Features of the Image:

1. Collision Point (Interaction Vertex):

- The orange lines emanating from the center represent the **tracks of charged particles** produced in the collision. The point where these lines originate is the **interaction vertex**, where two protons collided.

2. Yellow Dashed Line:

- Indicates the possible trajectory or momentum imbalance associated with particles not directly detected (e.g., **neutrinos** or missing transverse energy).

3. Green and Orange Rectangles:

- These are **energy deposits** in the calorimeters:
 - **Orange regions** represent energy deposited by particles in the electromagnetic calorimeter (ECAL), typically photons or electrons.
 - **Green regions** represent energy deposited in the hadronic calorimeter (HCAL), typically from hadrons like protons, neutrons, or pions.

4. Blue Cylinder:

- Represents the **CMS detector geometry**, showing the surrounding tracking systems and calorimeters that measure particles' momenta and energy.

5. Long Green Lines:

- Likely represent **muons**, which penetrate the entire detector and are reconstructed in the outermost layers of the CMS muon chambers.

6. Purpose of the Event:

- This display could represent a **candidate event for a Higgs boson decay** or other interesting physics processes, such as electroweak interactions or rare decays. The clean central vertex and extensive particle activity suggest it is a high-energy event.

Would you like more detailed information about a specific component or the physics process being studied?

Understanding physics data

tell me about the physics process occurring here

This event display from the CMS experiment at the LHC likely represents a **high-energy proton-proton collision**. The physics process occurring here could involve a **Standard Model process**, such as **Higgs boson production**, **electroweak interactions**, or **QCD jets**, depending on the details of the event. Let me walk you through the possibilities based on the visible features:

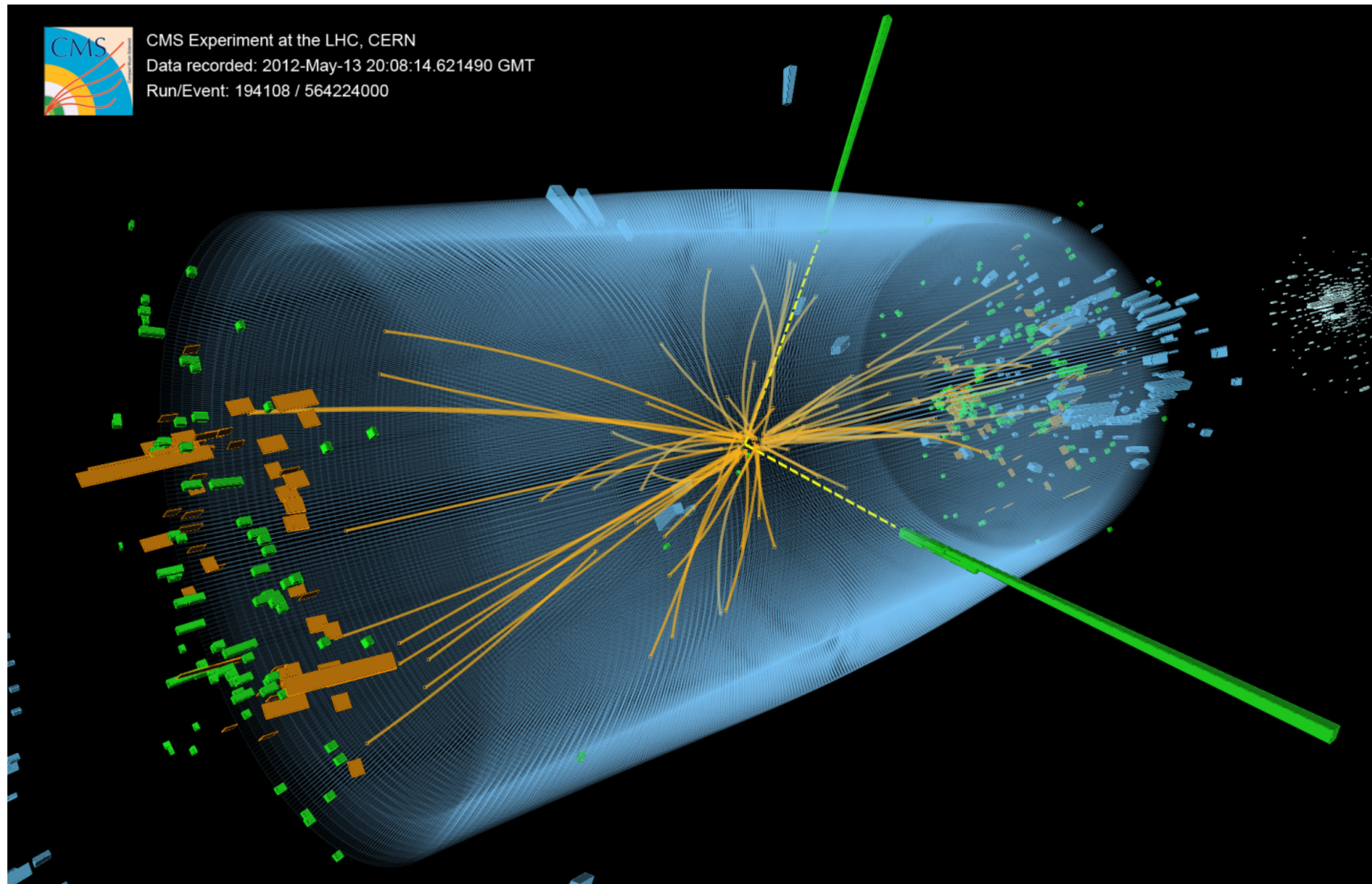
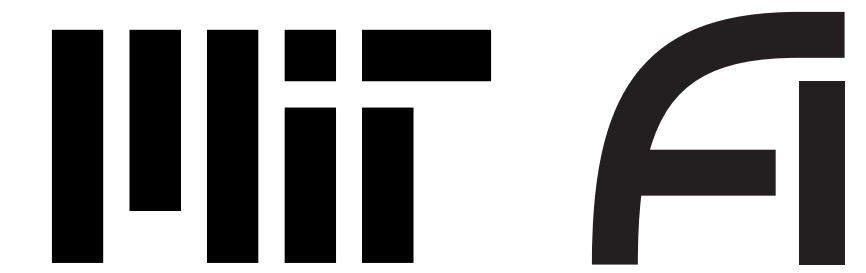


Image credit: CERN

$$H \rightarrow \gamma\gamma$$

Understanding physics data



tell me about the physics process occurring here

This event display from the CMS experiment at the LHC likely represents a **high-energy proton-proton collision**. The physics process occurring here could involve a **Standard Model process**, such as **Higgs boson production**, **electroweak interactions**, or **QCD jets**, depending on the details of the event. Let me walk you through the possibilities based on the visible features:

Possible Physics Processes

1. Higgs Boson Candidate Event

- The display shows a high-energy collision producing many particles. It could correspond to a **Higgs boson decay**, where the Higgs is produced and decays into detectable particles.
- **Potential Higgs decay channels:**
 - **H → $\gamma\gamma$ (two photons):**
 - If two large, isolated energy deposits in the electromagnetic calorimeter (orange) are visible, this could signify photons from a Higgs decay.
 - **H → ZZ → 4 leptons:**
 - Tracks corresponding to muons (long green lines) or electrons might indicate this channel, a “golden channel” for Higgs observation due to its clean signature.
 - **H → WW → leptons and neutrinos:**
 - Missing transverse energy (indicated by the yellow dashed line) suggests escaping neutrinos, a hallmark of this channel.

2. Jets and QCD Processes

- The dense spray of orange and green rectangles near the interaction vertex suggests the presence of **jets**, which are collimated showers of particles originating from quark or gluon fragmentation.
- This could be a **hard scattering process**, such as:
 - **QCD dijet production** (dominant at the LHC).
 - **Top quark pair production (t-tbar):**
 - If the event involves jets and missing energy, it could signify top quark decays into W bosons and jets.

3. Missing Transverse Energy and New Physics

- The yellow dashed line (missing transverse energy) indicates an imbalance in momentum in the transverse plane. This often occurs when particles like **neutrinos** escape the detector.
- Missing energy could also hint at **new physics** processes, such as:
 - **Supersymmetry (SUSY):**
 - Lightest supersymmetric particles (LSPs) can escape detection, leading to large missing transverse energy.
 - **Dark Matter production:**
 - Hypothetical dark matter particles produced in the collision would not interact with the detector, causing a momentum imbalance.

4. Electroweak Processes

- The event could also involve **weak bosons** (W^\pm or Z^0), which often produce leptons (e.g., muons or electrons) and neutrinos:
 - **W → lv** (lepton and neutrino decay).

How to Identify the Process

To determine the exact physics process:

1. Track and Vertex Analysis:

- Examine whether the tracks correspond to charged leptons (e.g., muons or electrons) or jets.
- Look for high-energy isolated objects (e.g., photons or leptons).

2. Calorimeter Energy Deposits:

- The patterns of green (hadronic) and orange (electromagnetic) deposits reveal whether the event is jet-dominated (QCD) or includes isolated photons/electrons (Higgs or electroweak processes).

3. Missing Energy:

- The presence of a missing transverse energy vector (yellow dashed line) often signifies neutrinos or potential new particles.

4. Monte Carlo Simulation:

- CMS physicists compare such events to simulations of known physics processes to classify them.

Most Likely Interpretation

Given the structure of this event:

- If it features isolated high-energy photons or leptons, it might be a **Higgs boson candidate**.
- If it involves large sprays of particles (jets) and missing energy, it could be **top quark production** or a **search for new physics**.
- If jets dominate with no missing energy, it is likely a **QCD dijet process**.

Would you like more specific guidance on identifying key features or understanding the event reconstruction process?

Understanding physics data

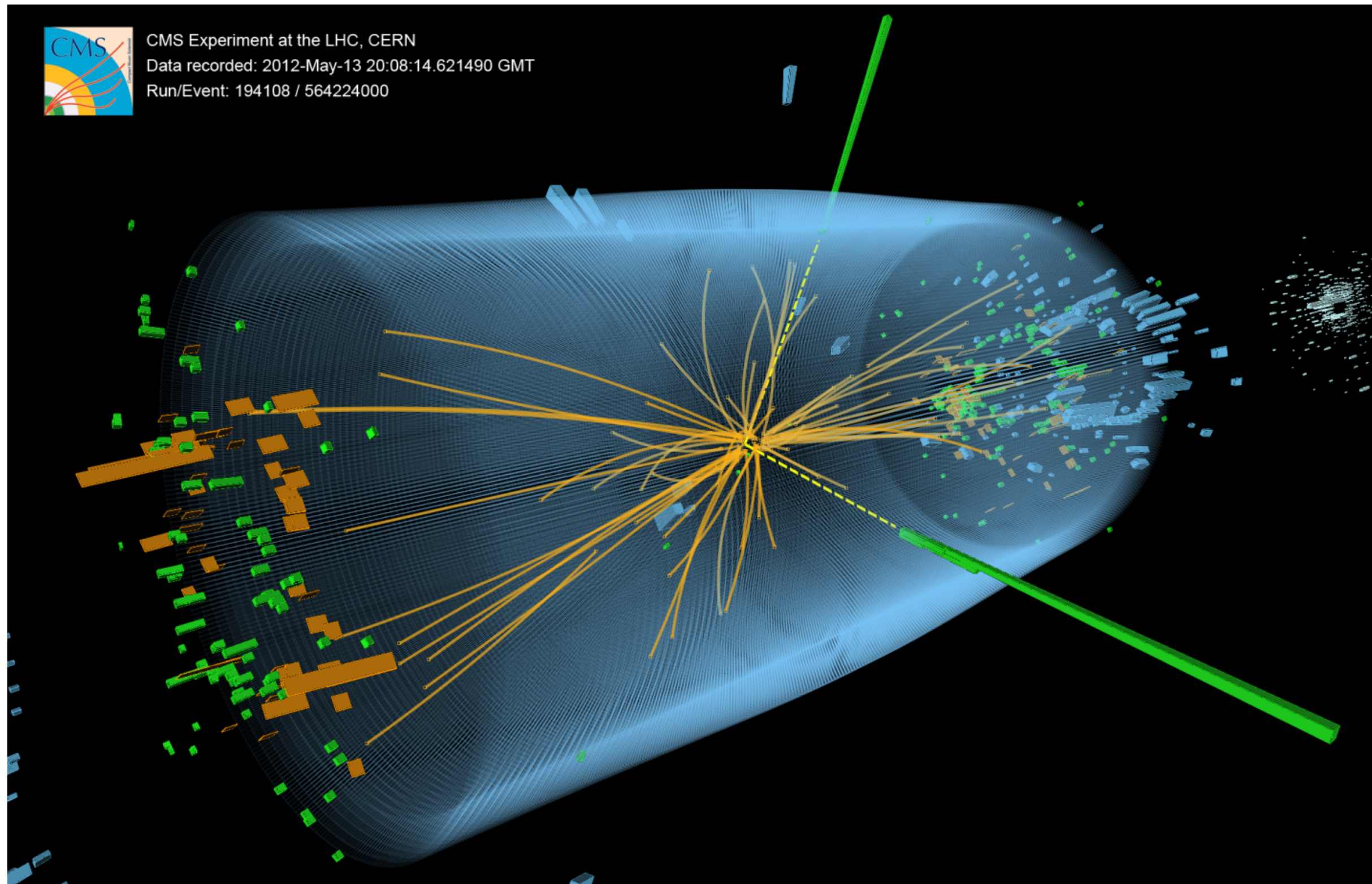


Image credit: CERN

$$H \rightarrow \gamma\gamma$$

tell me your best guess about the specific process occurring (a single answer)

My best guess is that this event represents **Higgs boson production with decay into two W bosons**, followed by $H \rightarrow WW \rightarrow \ell\nu\ell\nu$ (where ℓ represents leptons, and ν are neutrinos).

Why this process?

- **Yellow dashed line (missing transverse energy):** Suggests escaping neutrinos, characteristic of $W \rightarrow \ell\nu$ decays.
- **Spray of particles near the vertex (jets):** Consistent with activity from initial proton-proton collision and decay products.
- **Presence of long green tracks (muons):** Likely indicates leptonic decays of the W -bosons.

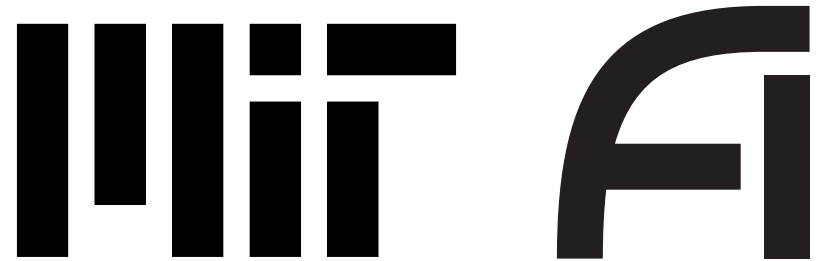
This process is a prominent channel used to study the Higgs boson because of its relatively clean leptonic signature combined with missing energy.

Not even close!

Clearly there is some work to be done...

But what do we mean by “foundation model”?

Handling physics data



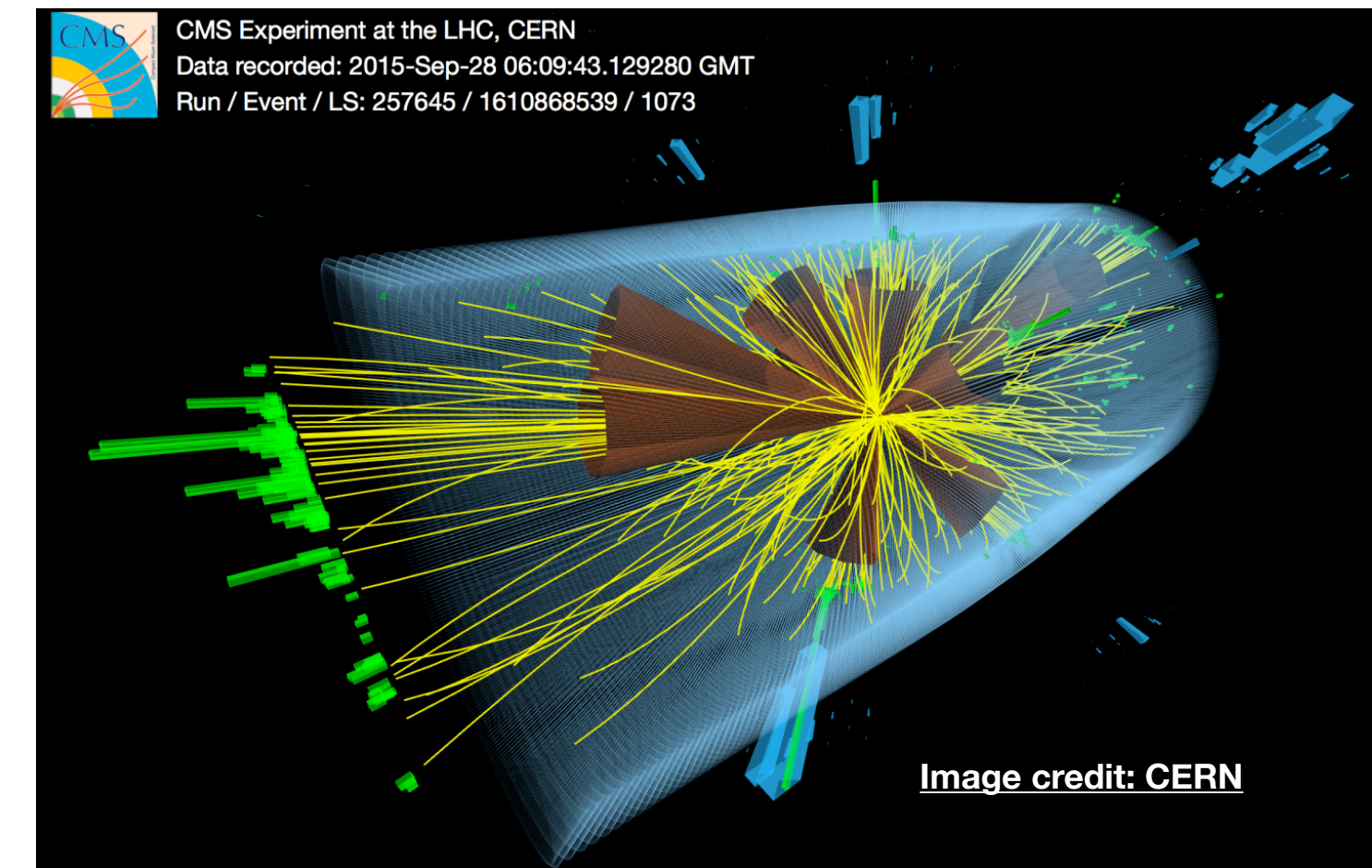
Handling physics data

Particle physics data is *really different* than
language!

Handling physics data

Particle physics data is *really different* than language!

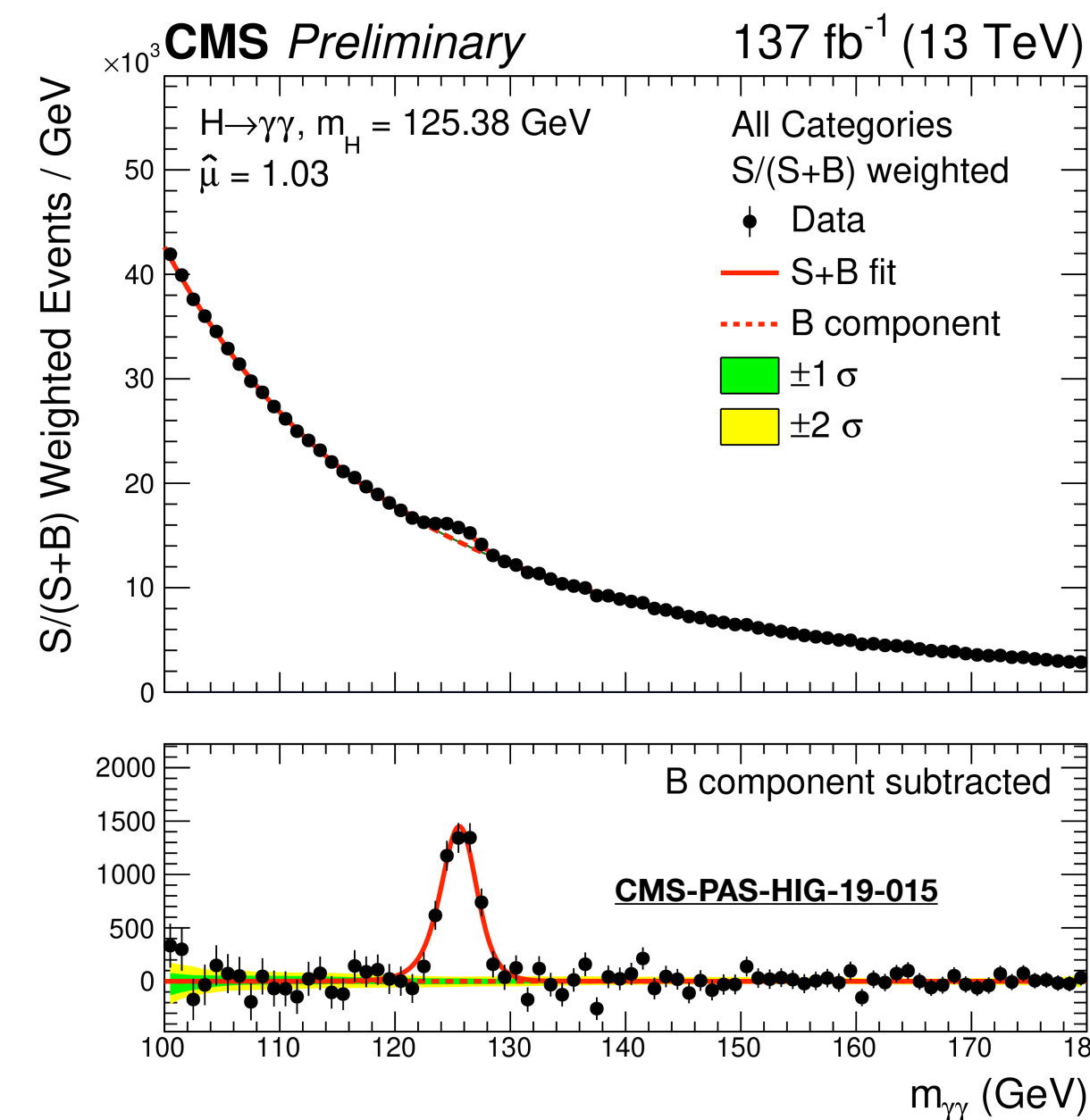
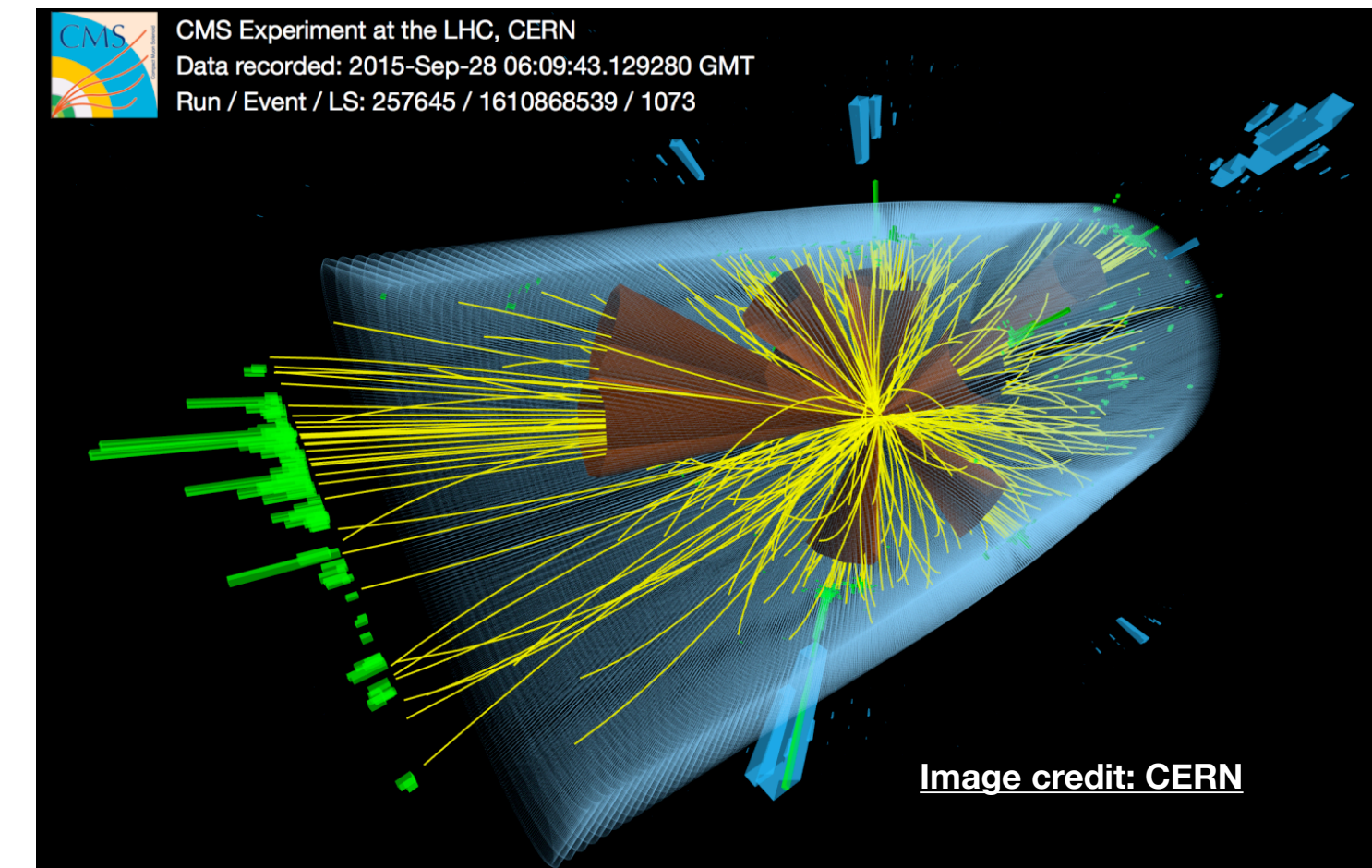
- ▶ Very high-dimensional, not tokenized, almost no “meaning” by itself
- ▶ Unordered



Handling physics data

Particle physics data is **really different** than language!

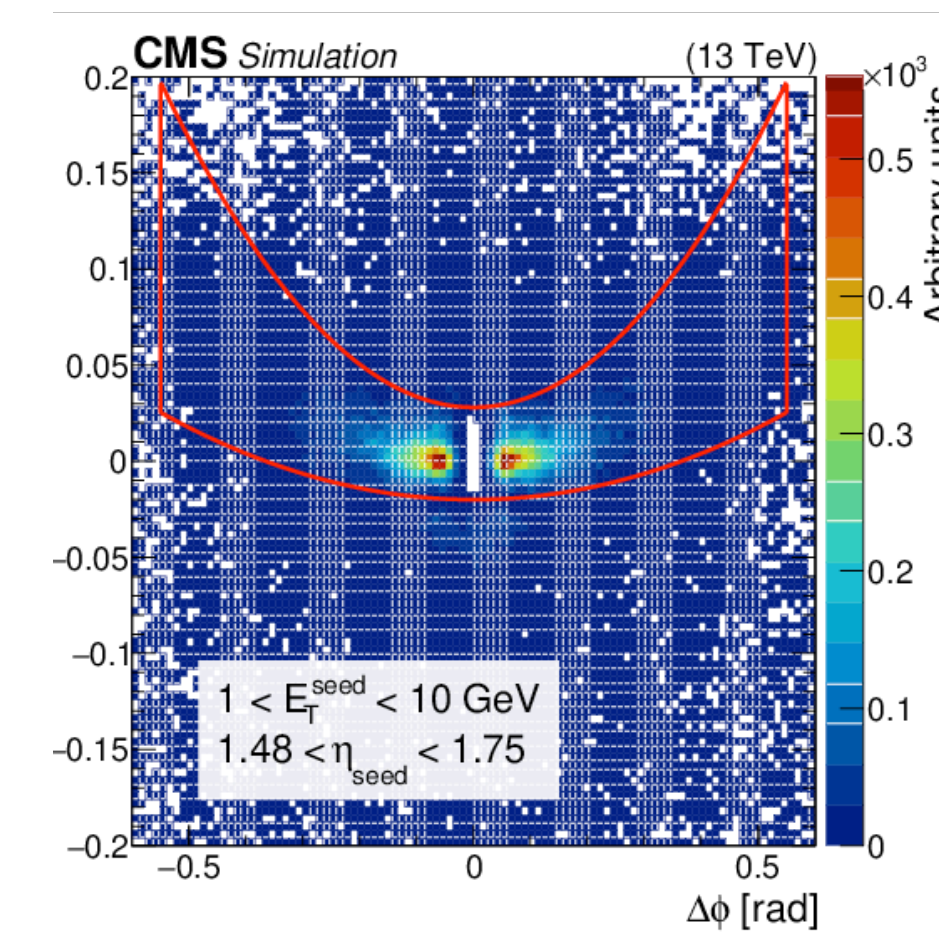
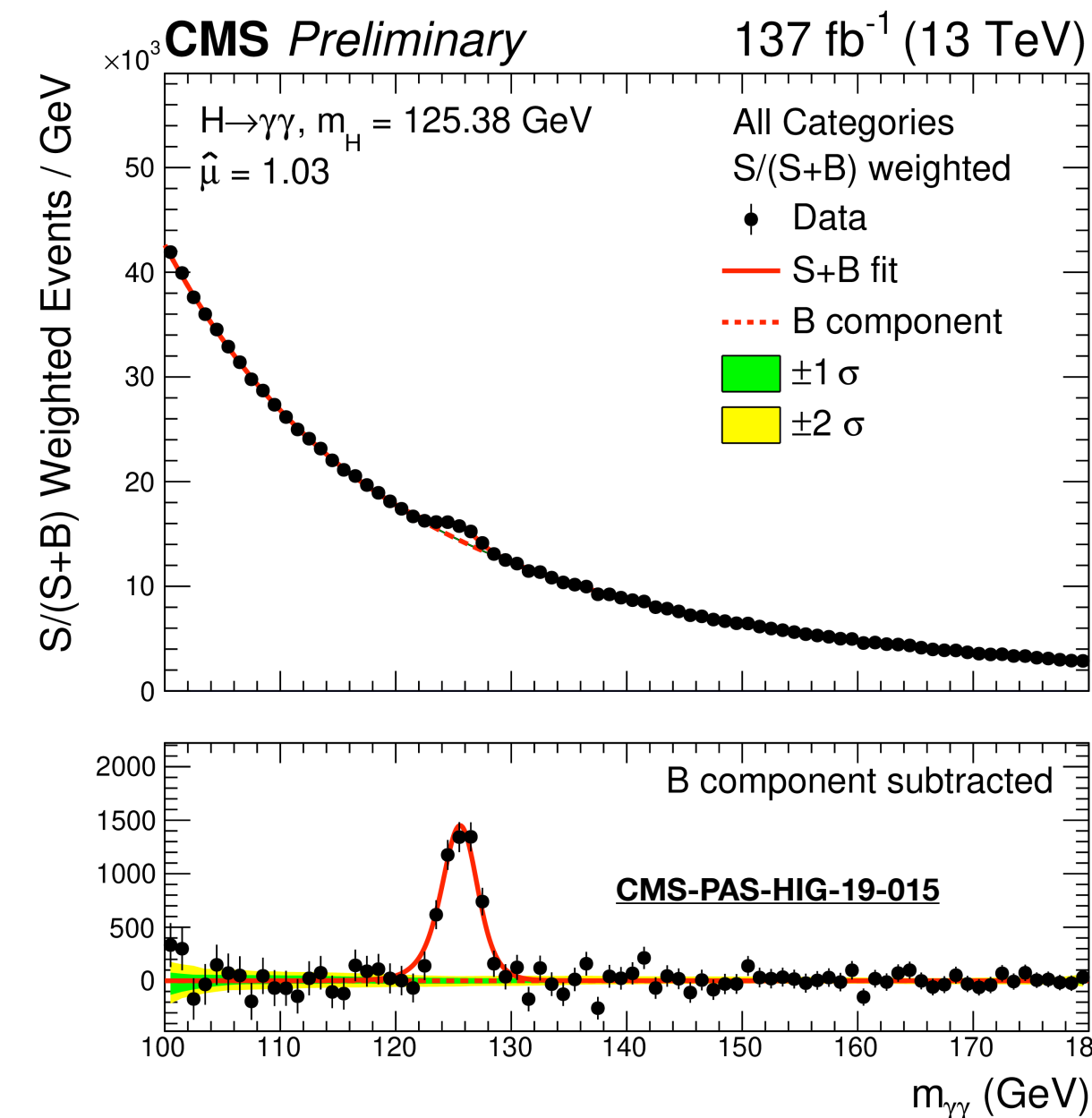
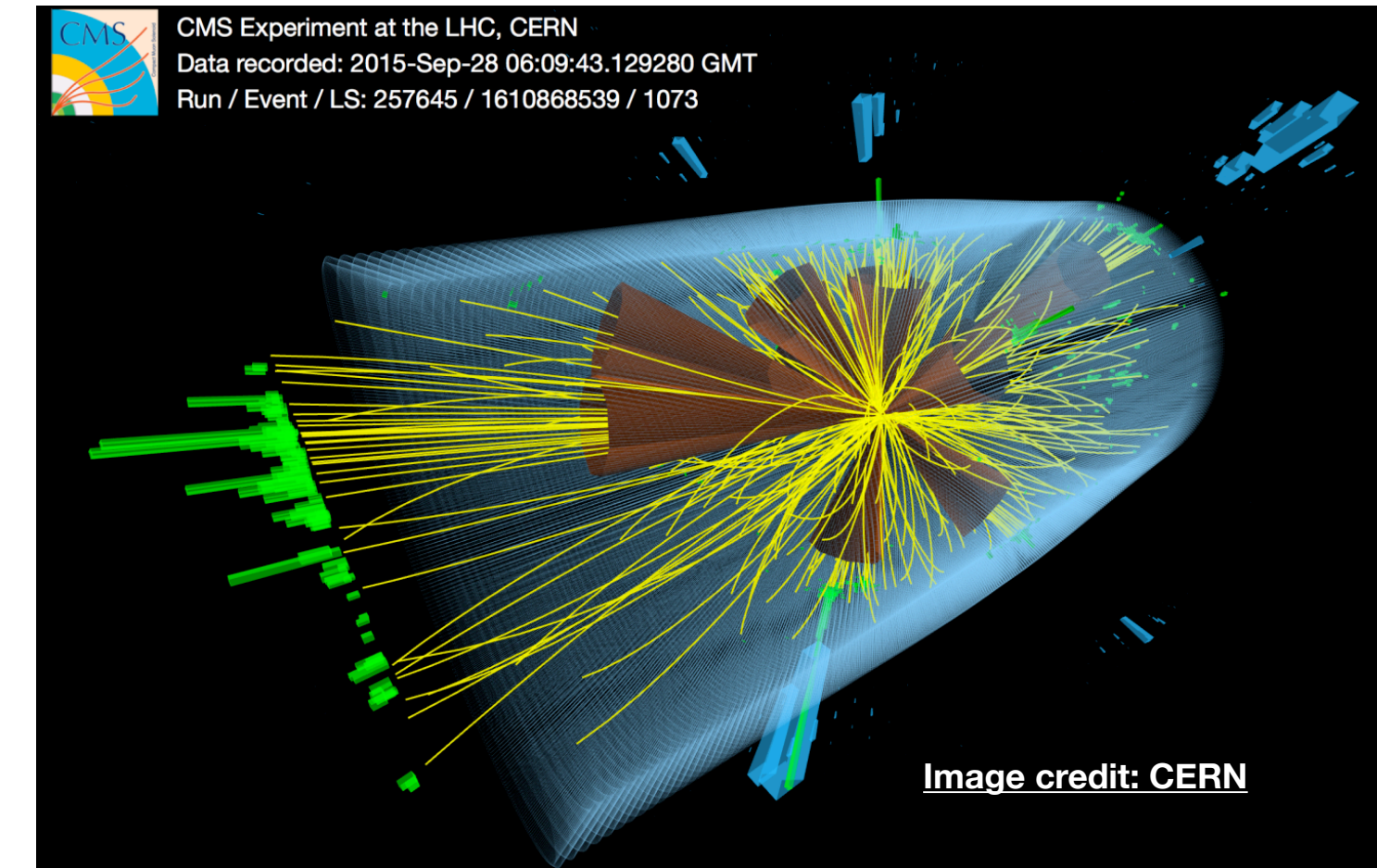
- ▶ Very high-dimensional, not tokenized, almost no “meaning” by itself
- ▶ Unordered
- ▶ Interpretation is *statistical* (no $N = 1$ discovery)
- ▶ Highly context-dependent (detector, reco)



Handling physics data

Particle physics data is *really different* than language!

- ▶ Very high-dimensional, not tokenized, almost no “meaning” by itself
- ▶ Unordered
- ▶ Interpretation is *statistical* (no $N = 1$ discovery)
- ▶ Highly context-dependent (detector, reco)
- ▶ Defined at many different scales



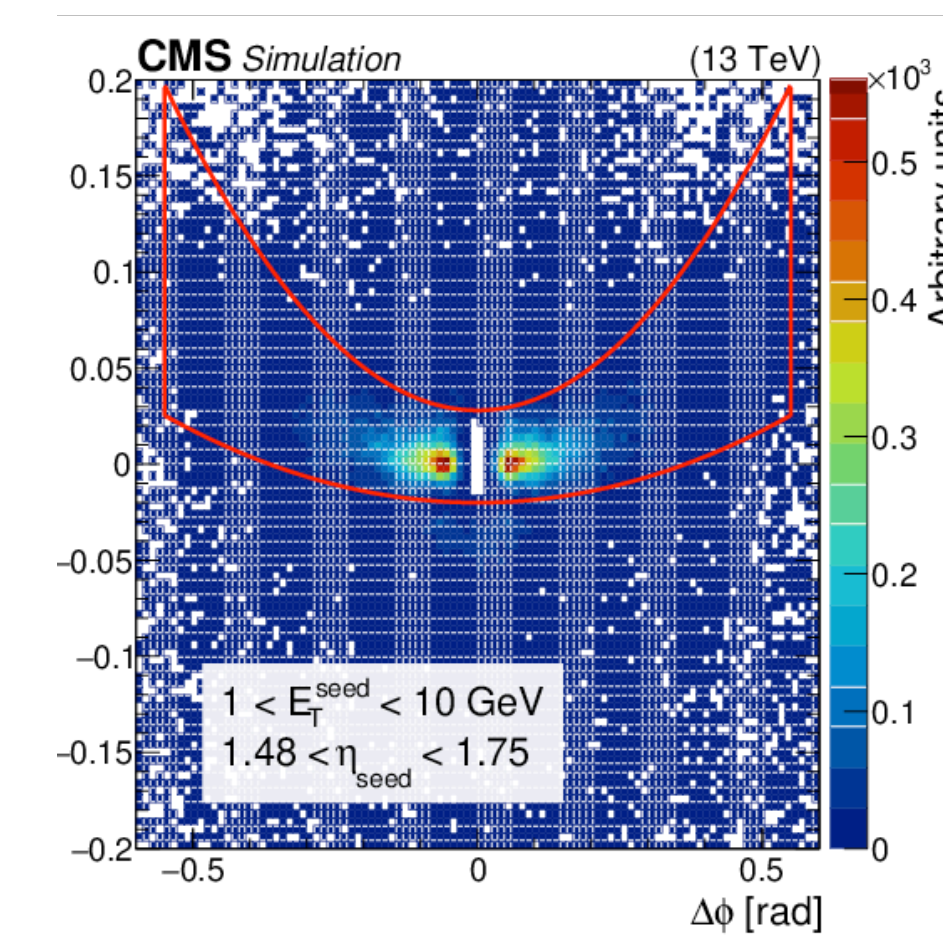
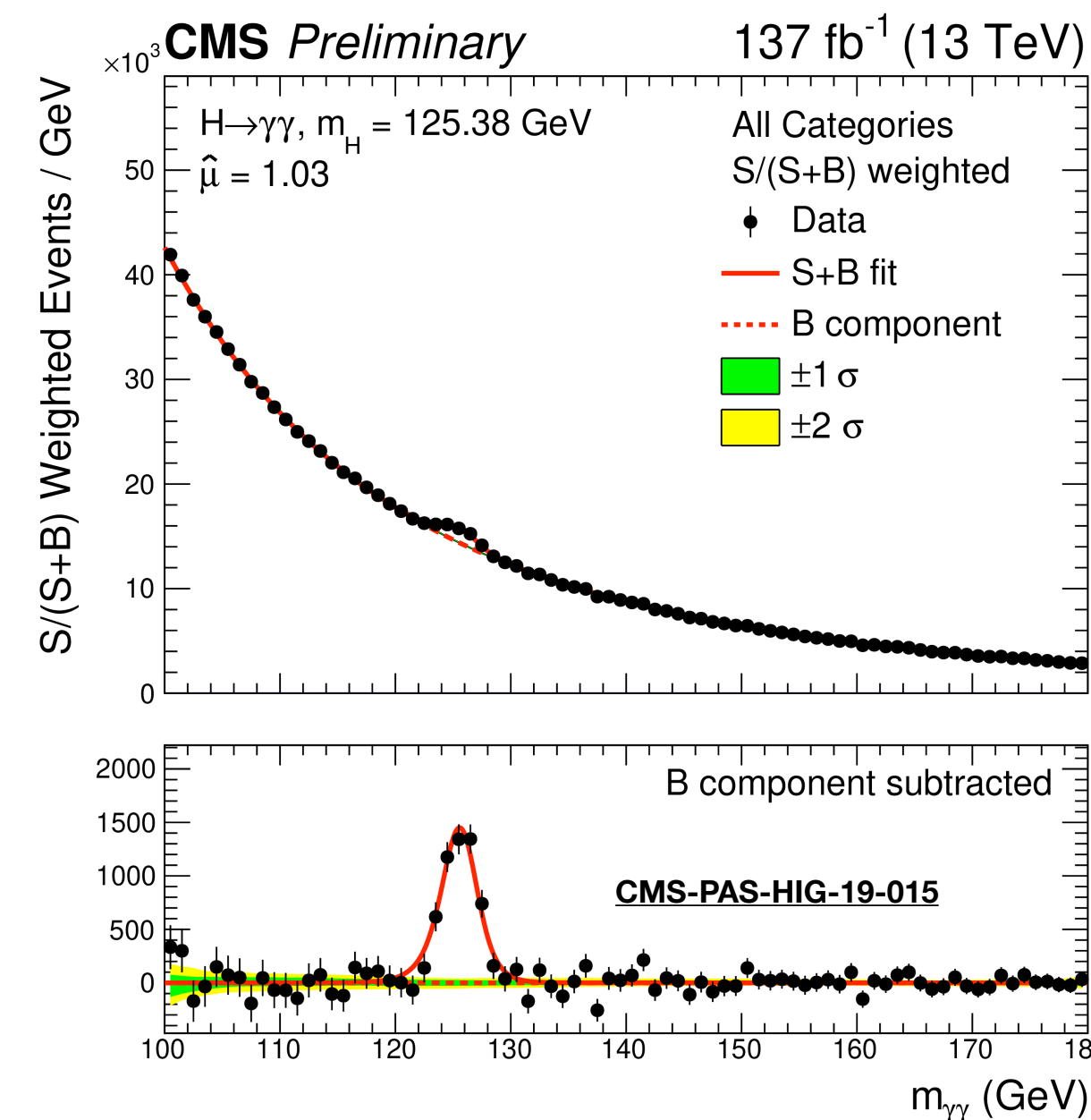
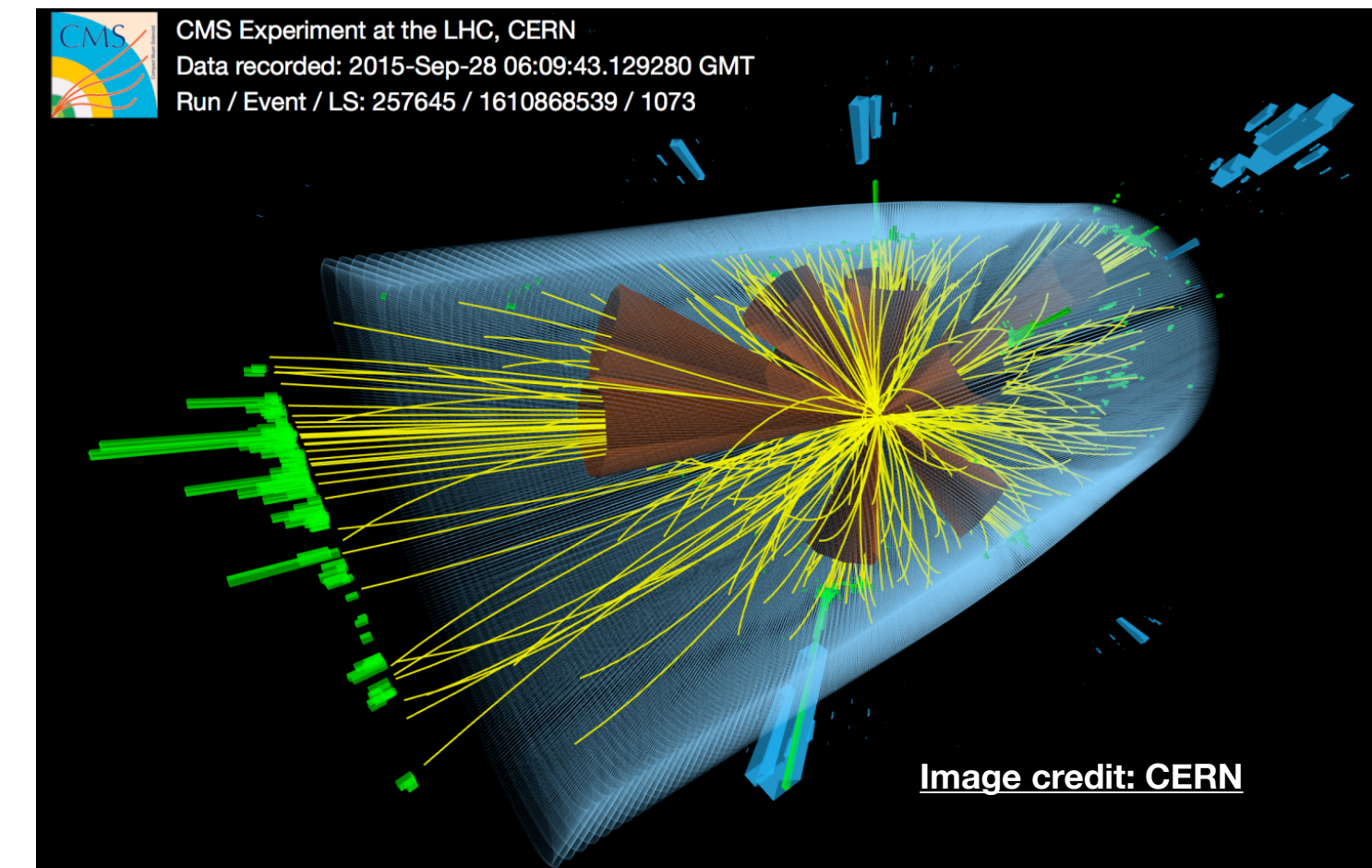
$p^\mu = (E, p_x, p_y, p_z)$

Handling physics data

Particle physics data is **really different** than language!

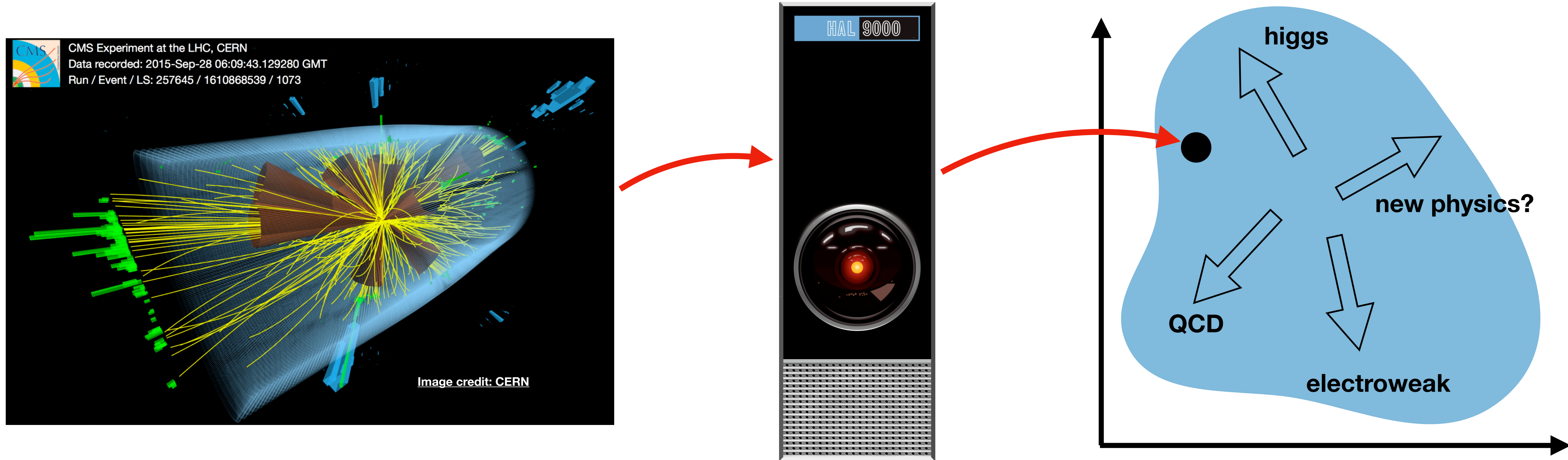
- ▶ Very high-dimensional, not tokenized, almost no “meaning” by itself
- ▶ Unordered
- ▶ Interpretation is *statistical* (no $N = 1$ discovery)
- ▶ Highly context-dependent (detector, reco)
- ▶ Defined at many different scales
- ▶ *Generative* models (e.g. GPT) are rarely necessary

What kind of “foundation model” approach is suited to this data?



$p^\mu = (E, p_x, p_y, p_z)$

The “space” view

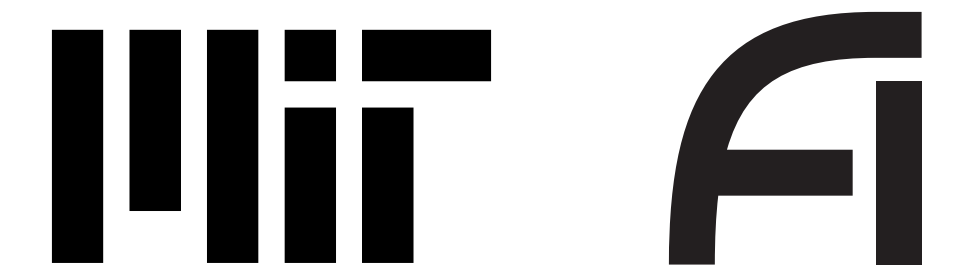


- My position in this talk: a good “FM” for collider data should **compress noisy, complex, high/variable-dimensional data into a low-dimensional, physically meaningful embedding**
- A “universal pre-processing step” for collider data — everything else can be fine-tuned downstream
- Possibility to inject inductive biases (e.g. preserving physical metrics, accounting for systematics, etc)

Outline

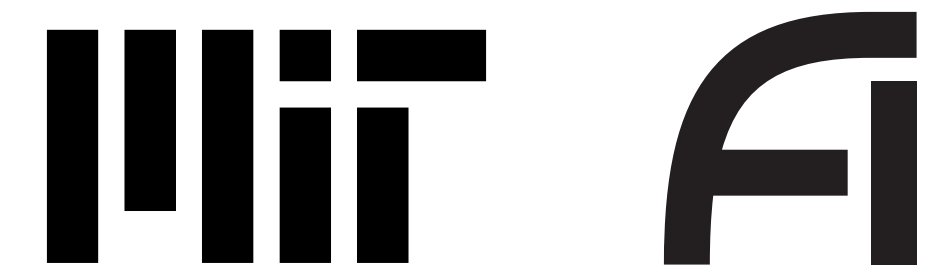


Outline



- An LHC physics primer

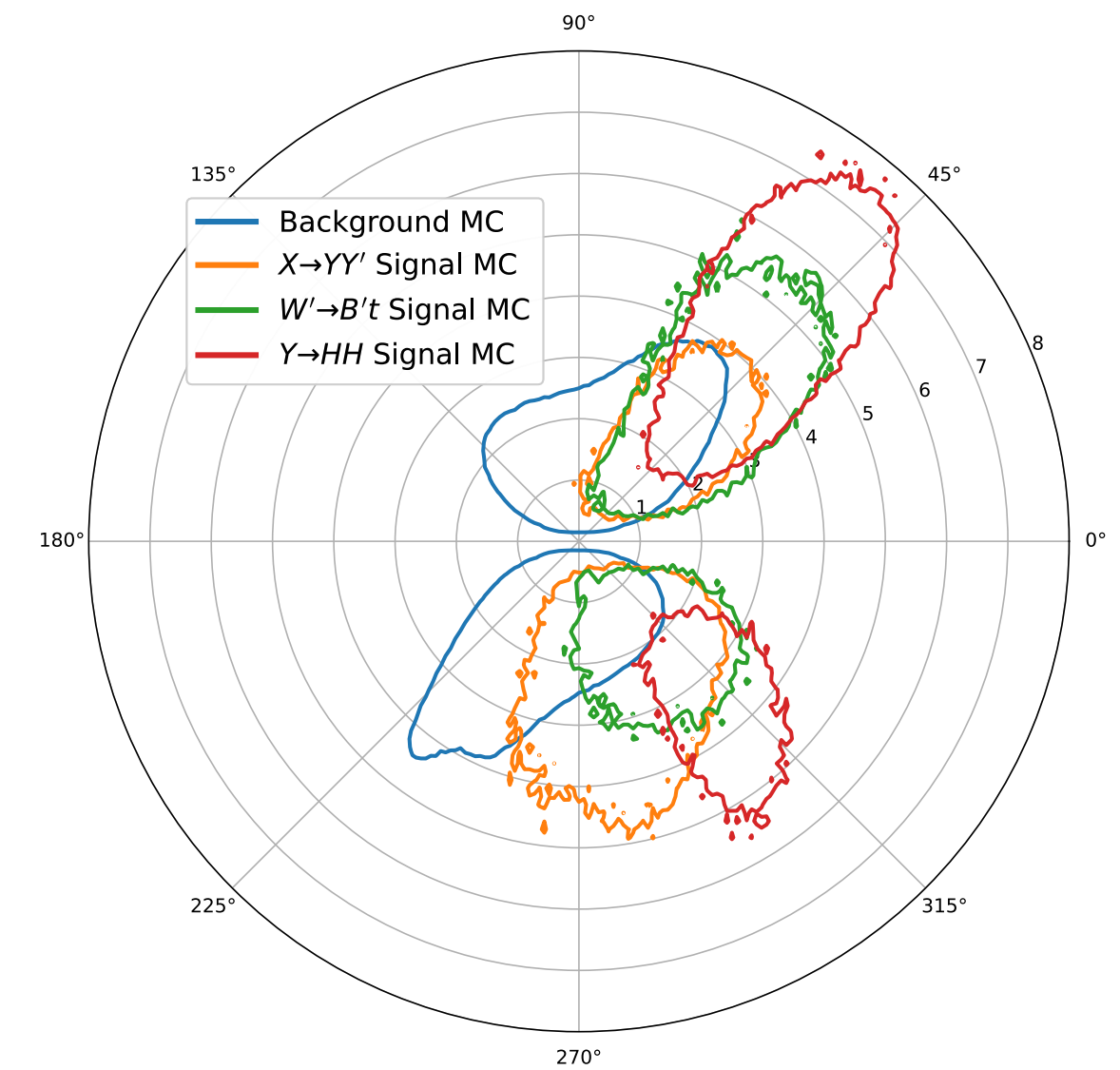
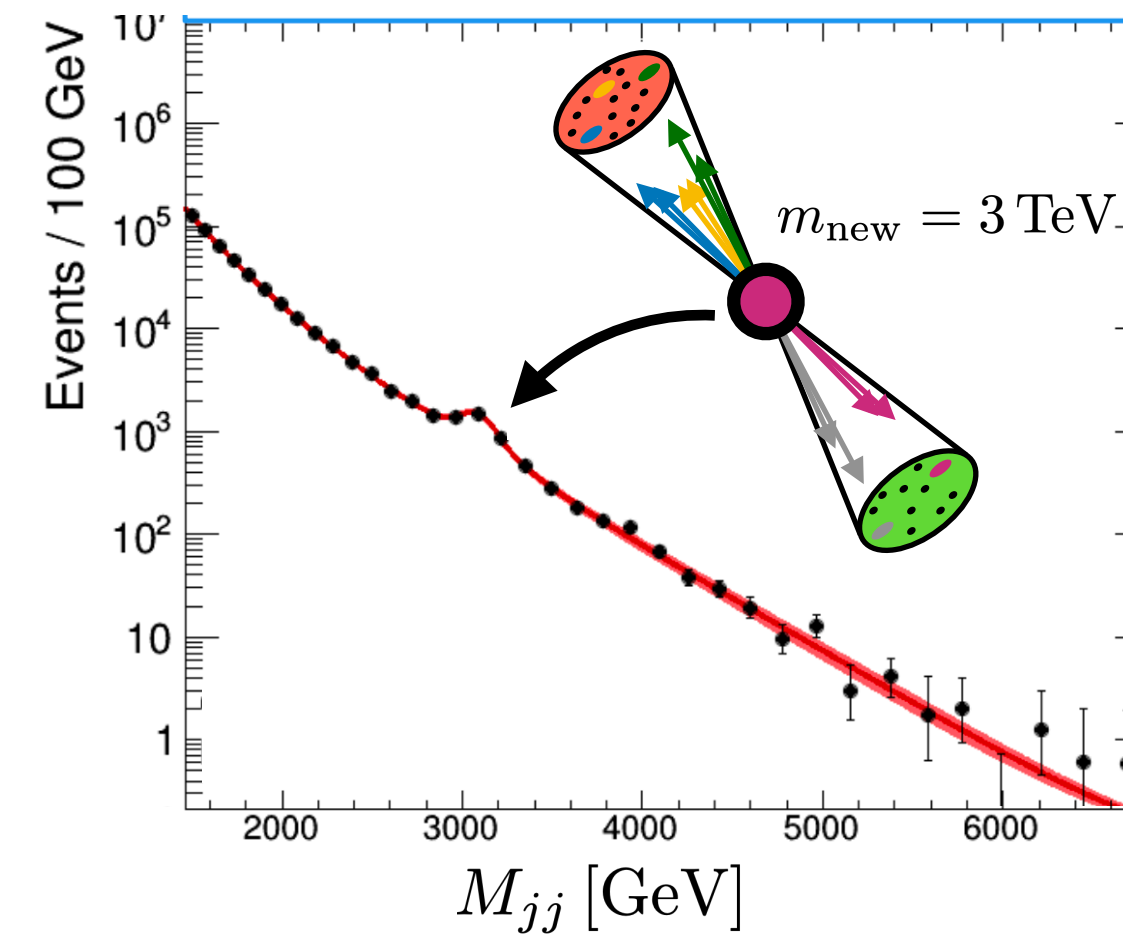
Outline



- An LHC physics primer
- Progress towards useful “spaces” for HEP

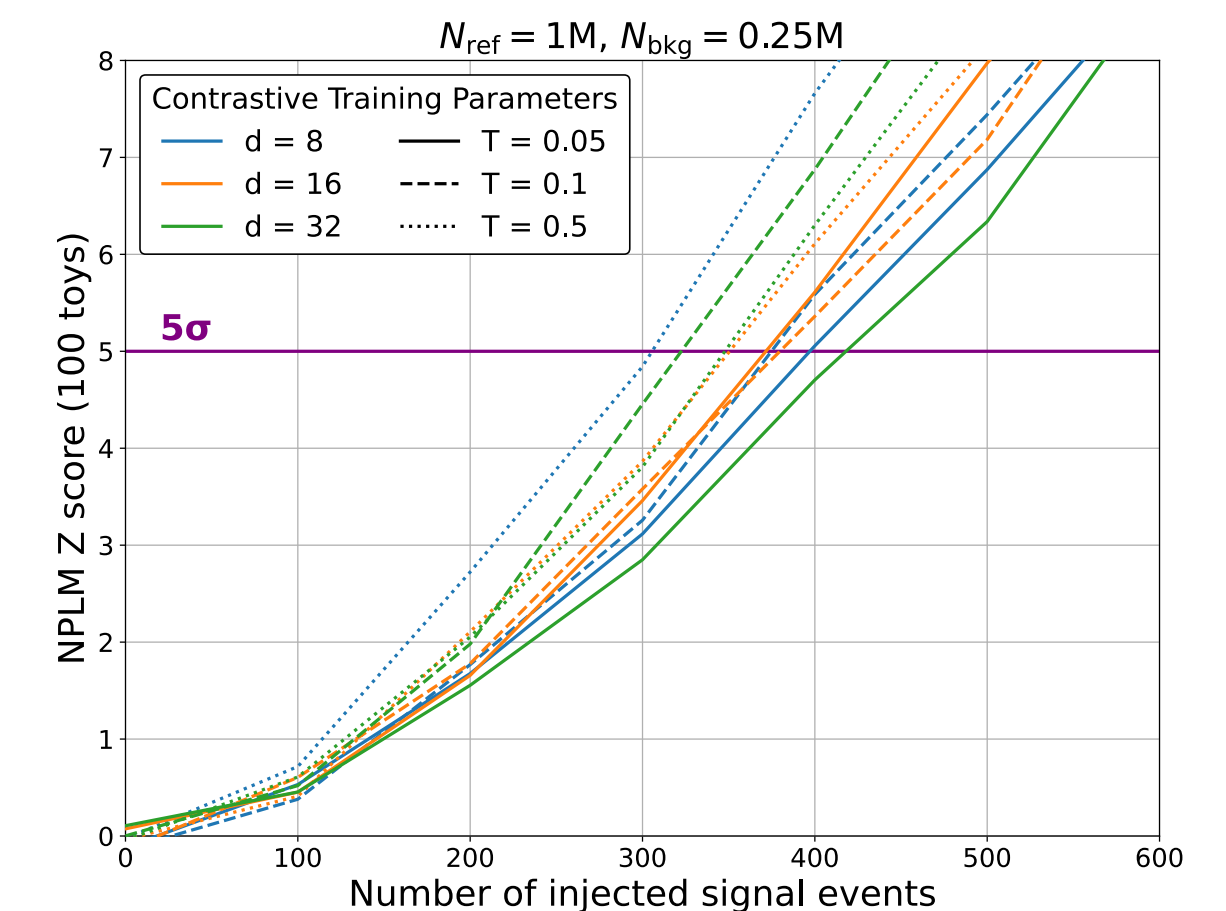
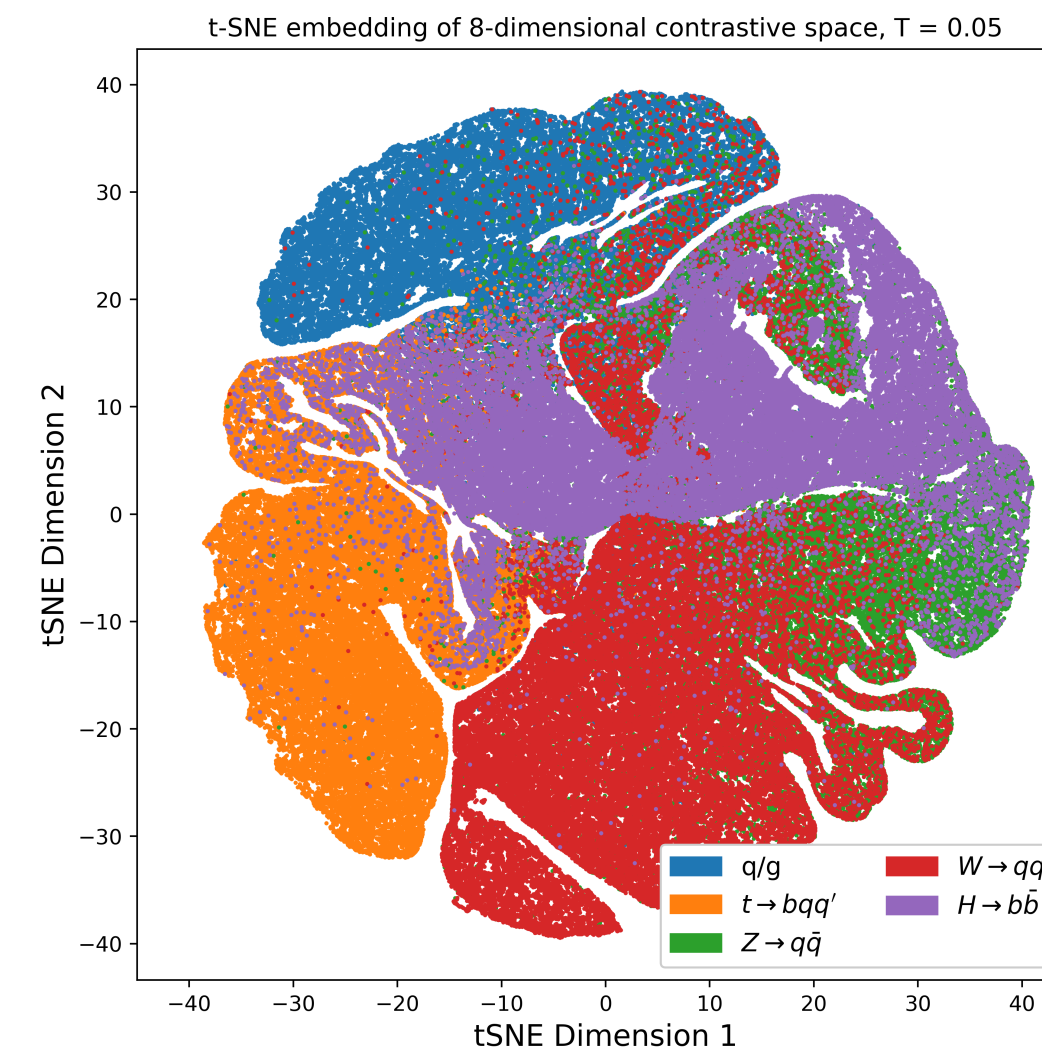
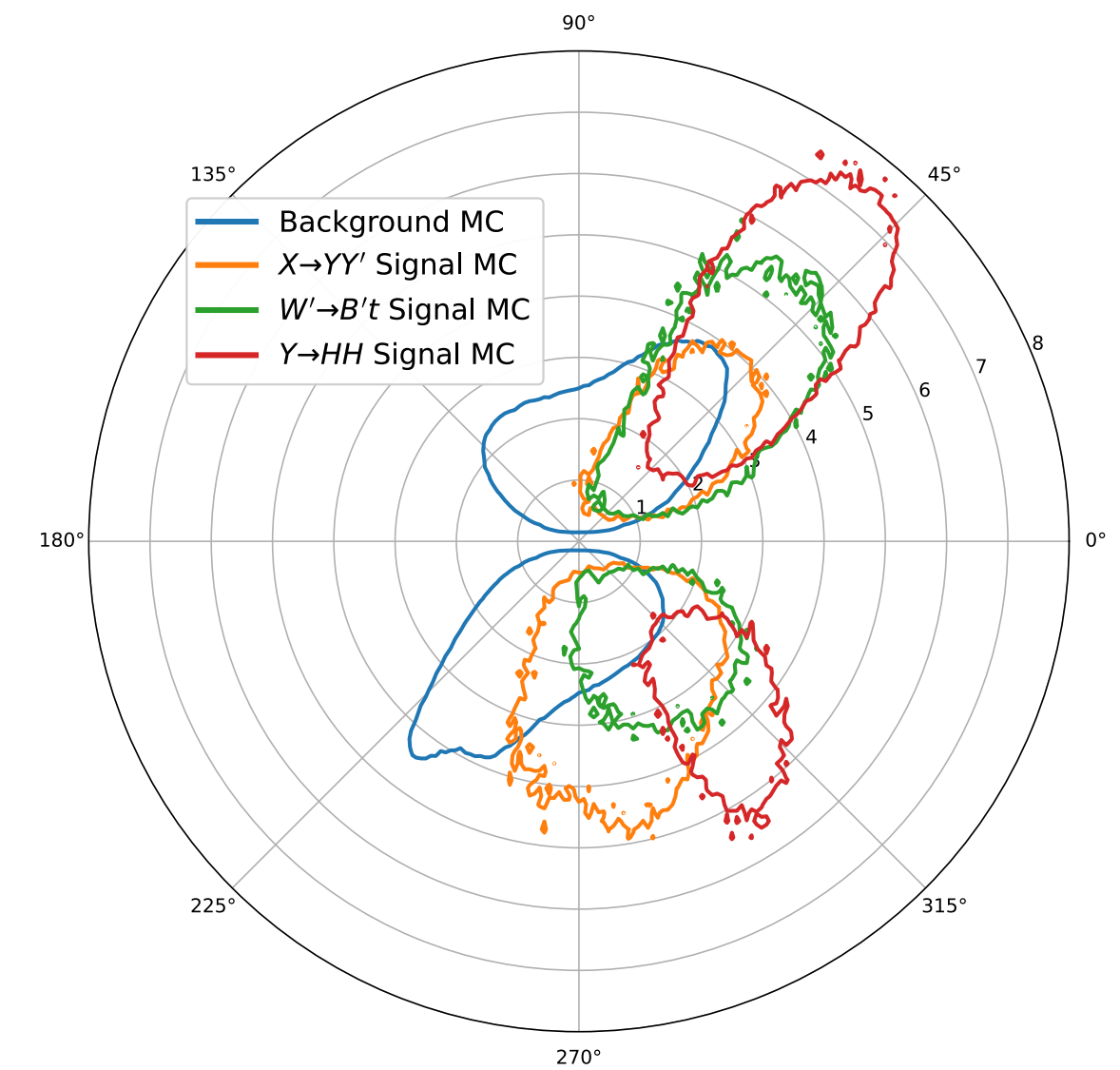
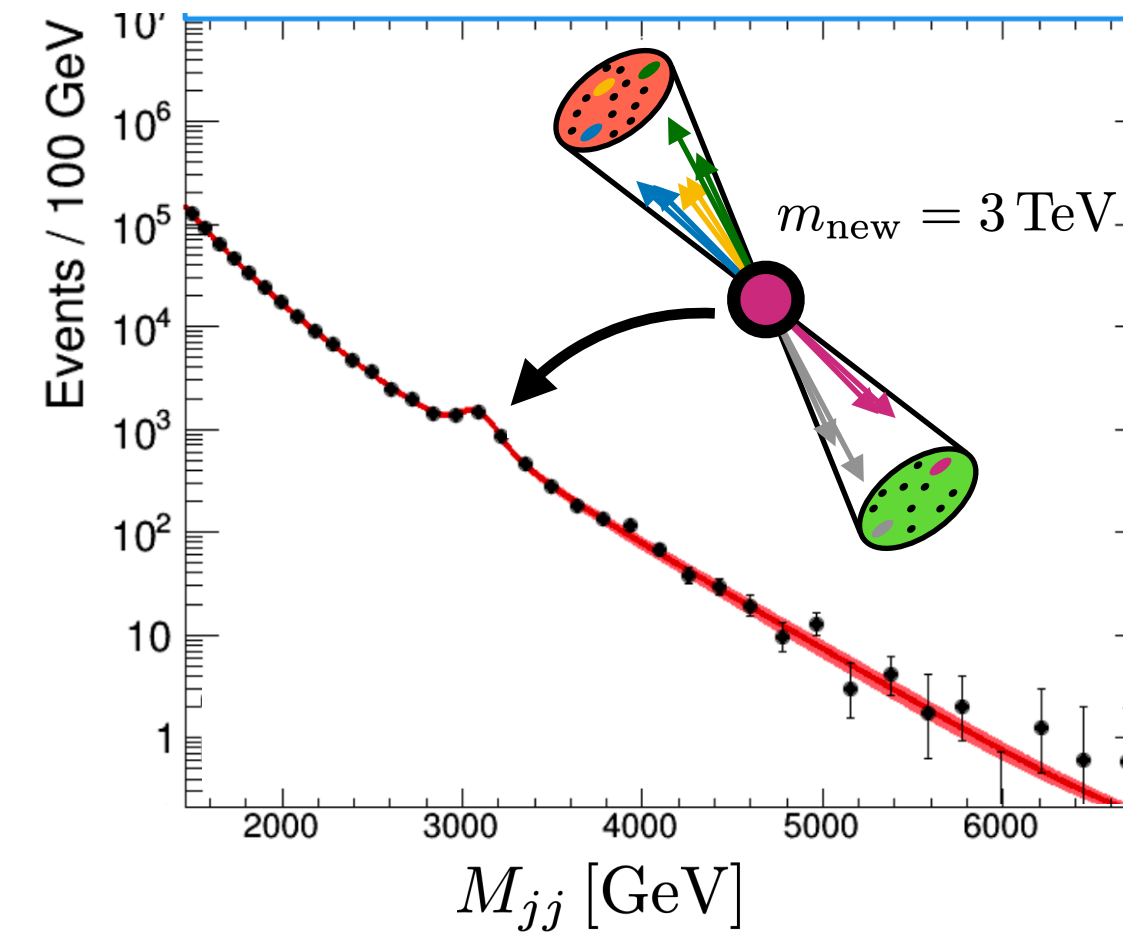
Outline

- An LHC physics primer
- Progress towards useful “spaces” for HEP
 - ▶ **Semi-supervised** spaces for anomaly detection



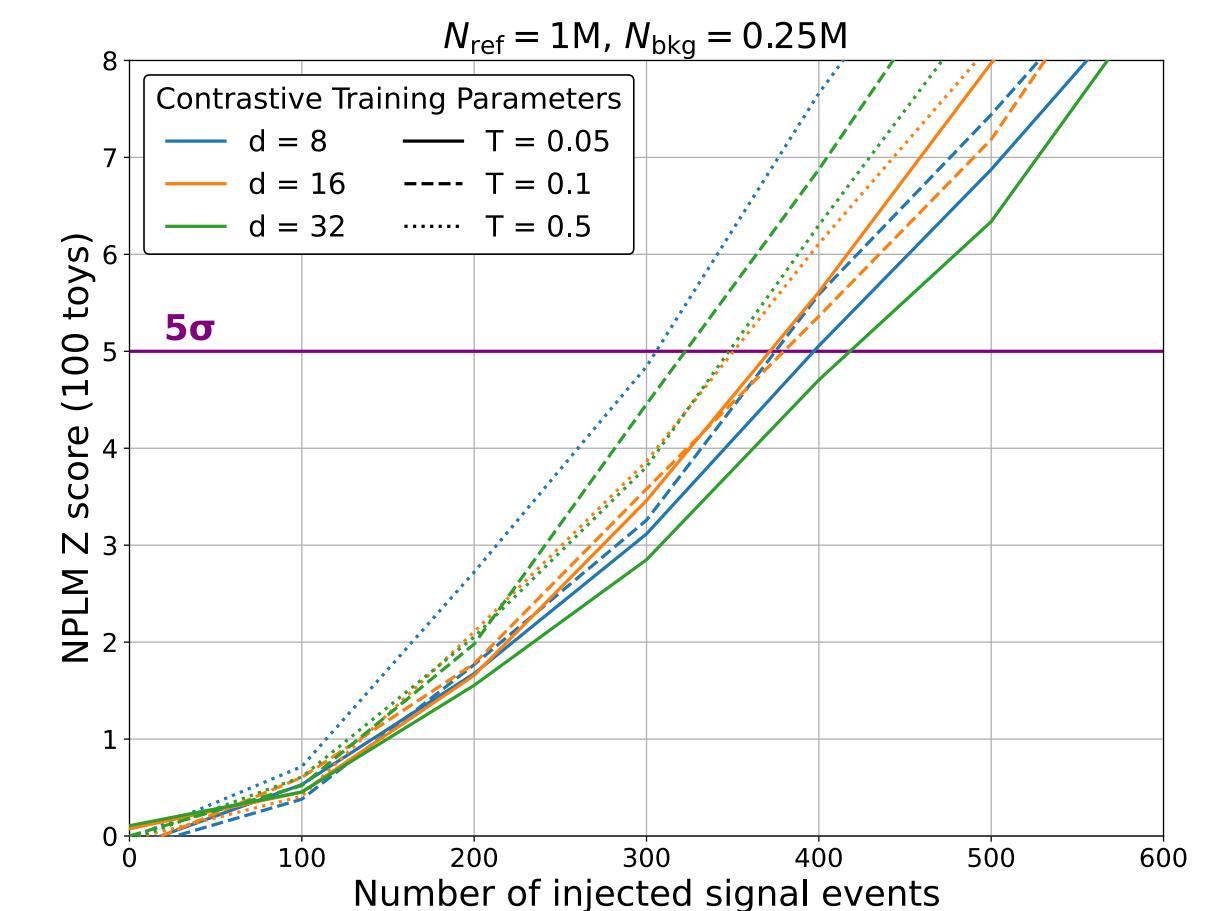
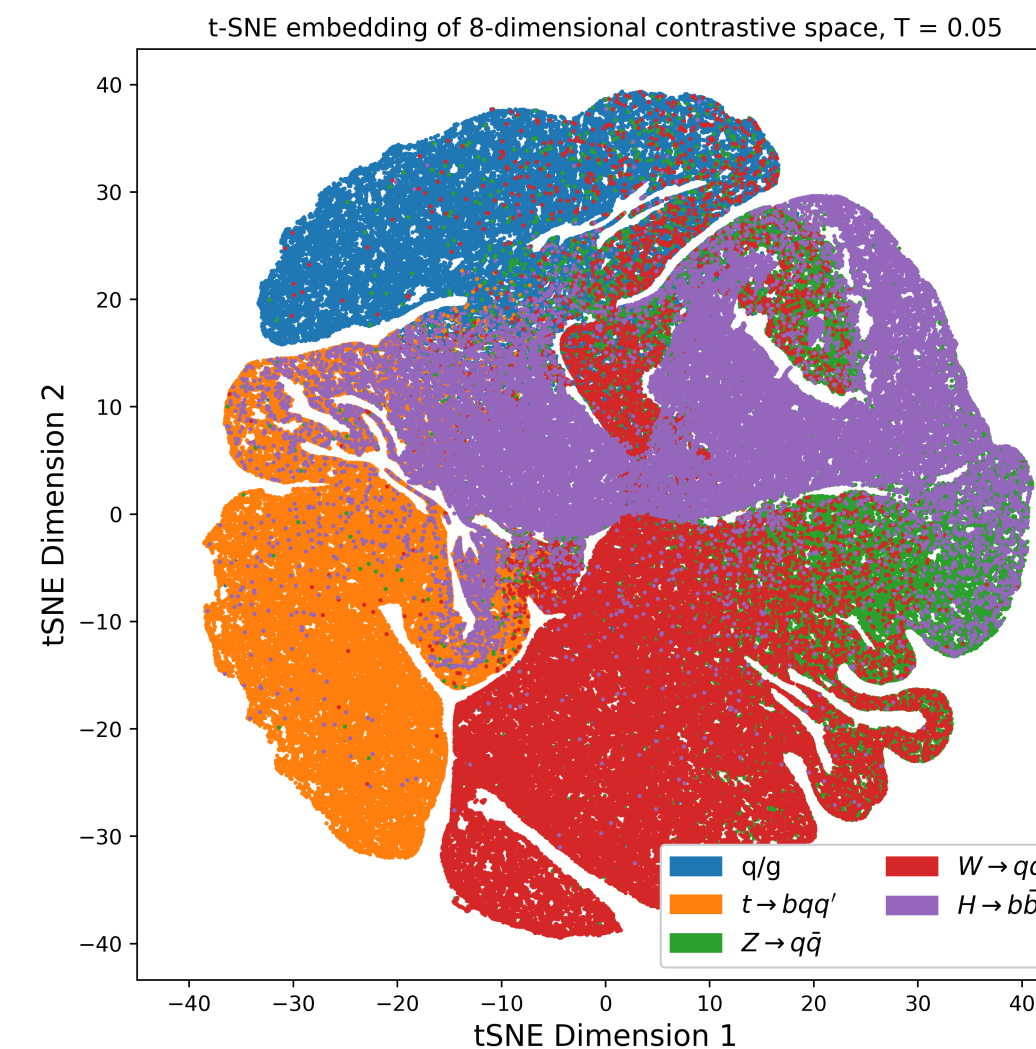
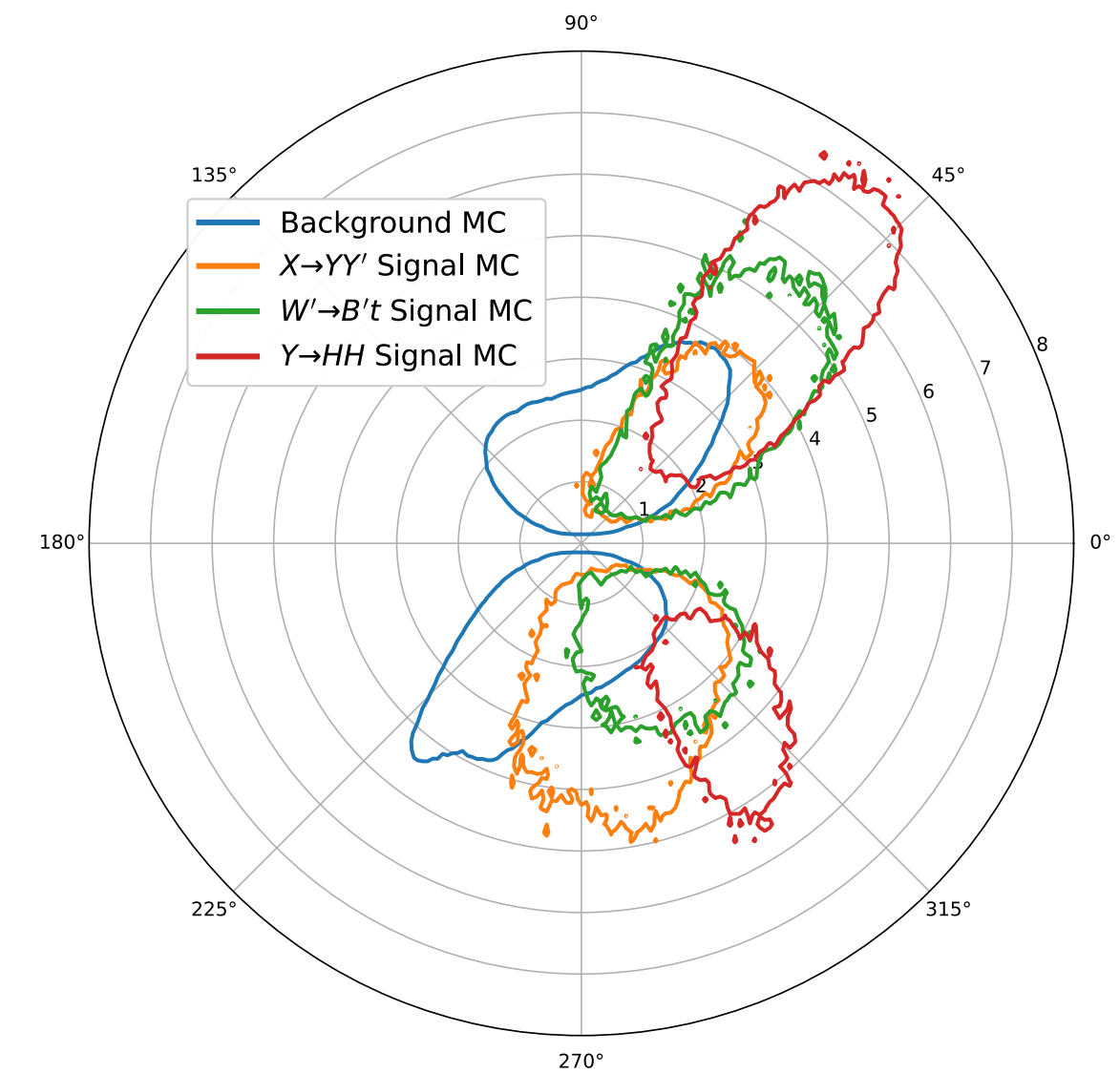
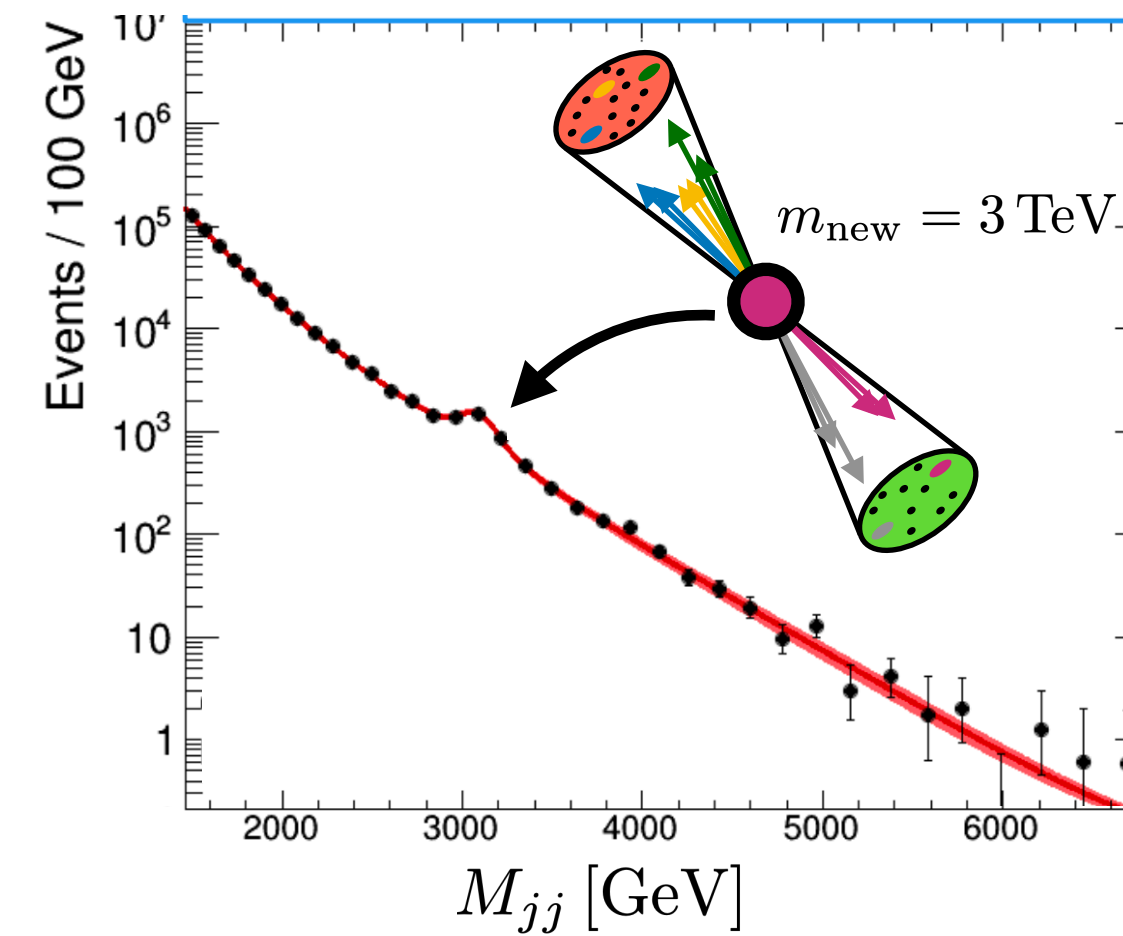
Outline

- An LHC physics primer
- Progress towards useful “spaces” for HEP
 - ▶ **Semi-supervised** spaces for anomaly detection
 - ▶ **Self-supervised** contrastive spaces for anomaly detection & more (FM-style)
 - ▶ Correcting simulation bias



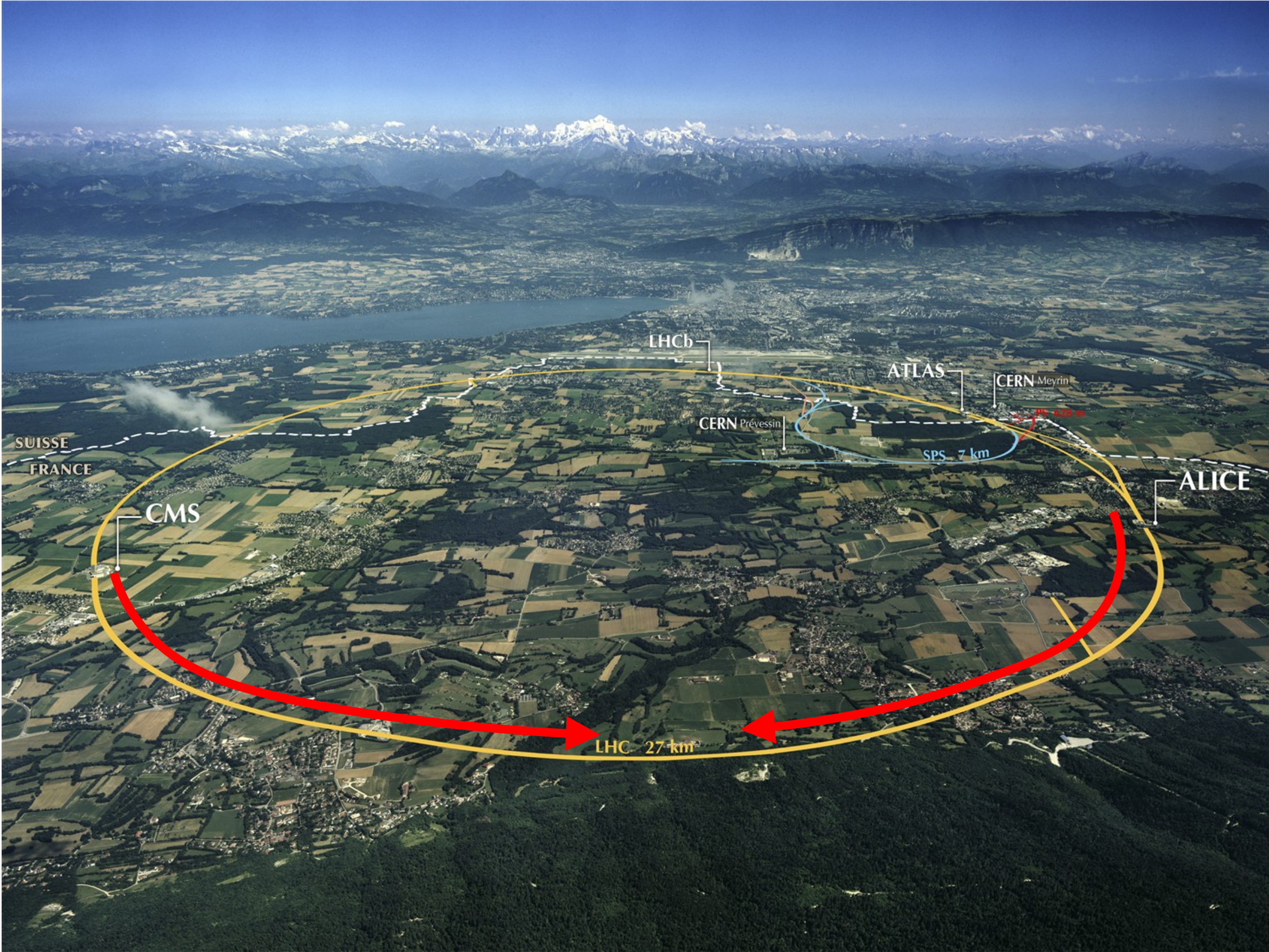
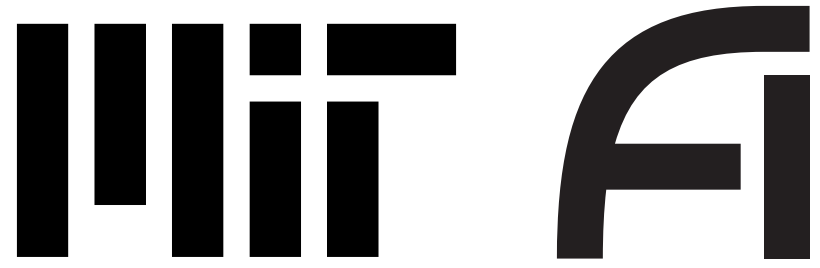
Outline

- An LHC physics primer
- Progress towards useful “spaces” for HEP
 - ▶ **Semi-supervised** spaces for anomaly detection
 - ▶ **Self-supervised** contrastive spaces for anomaly detection & more (FM-style)
 - ▶ Correcting simulation bias
- Ongoing/future work & outlook

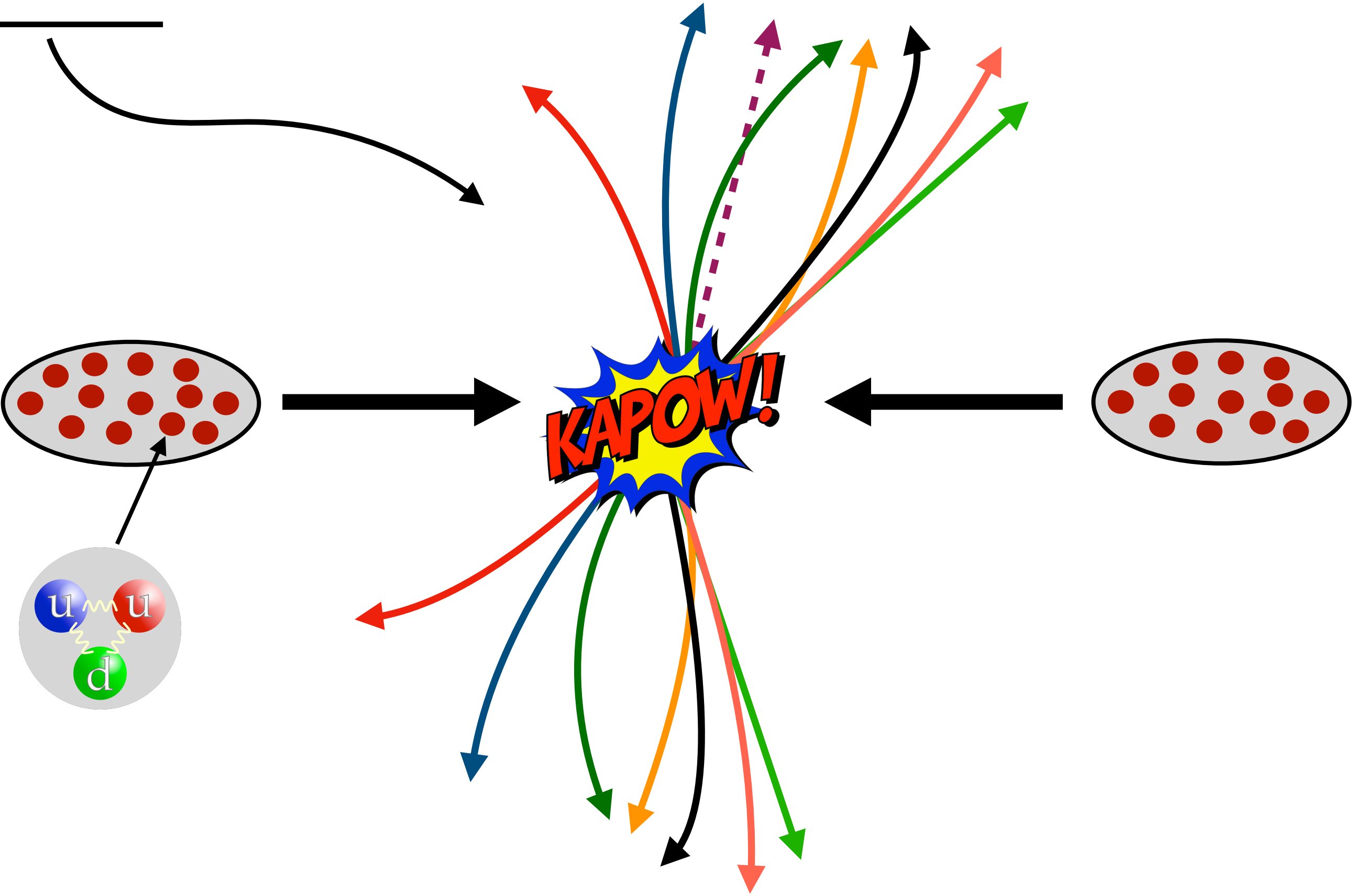


Some LHC physics basics

The Large Hadron Collider

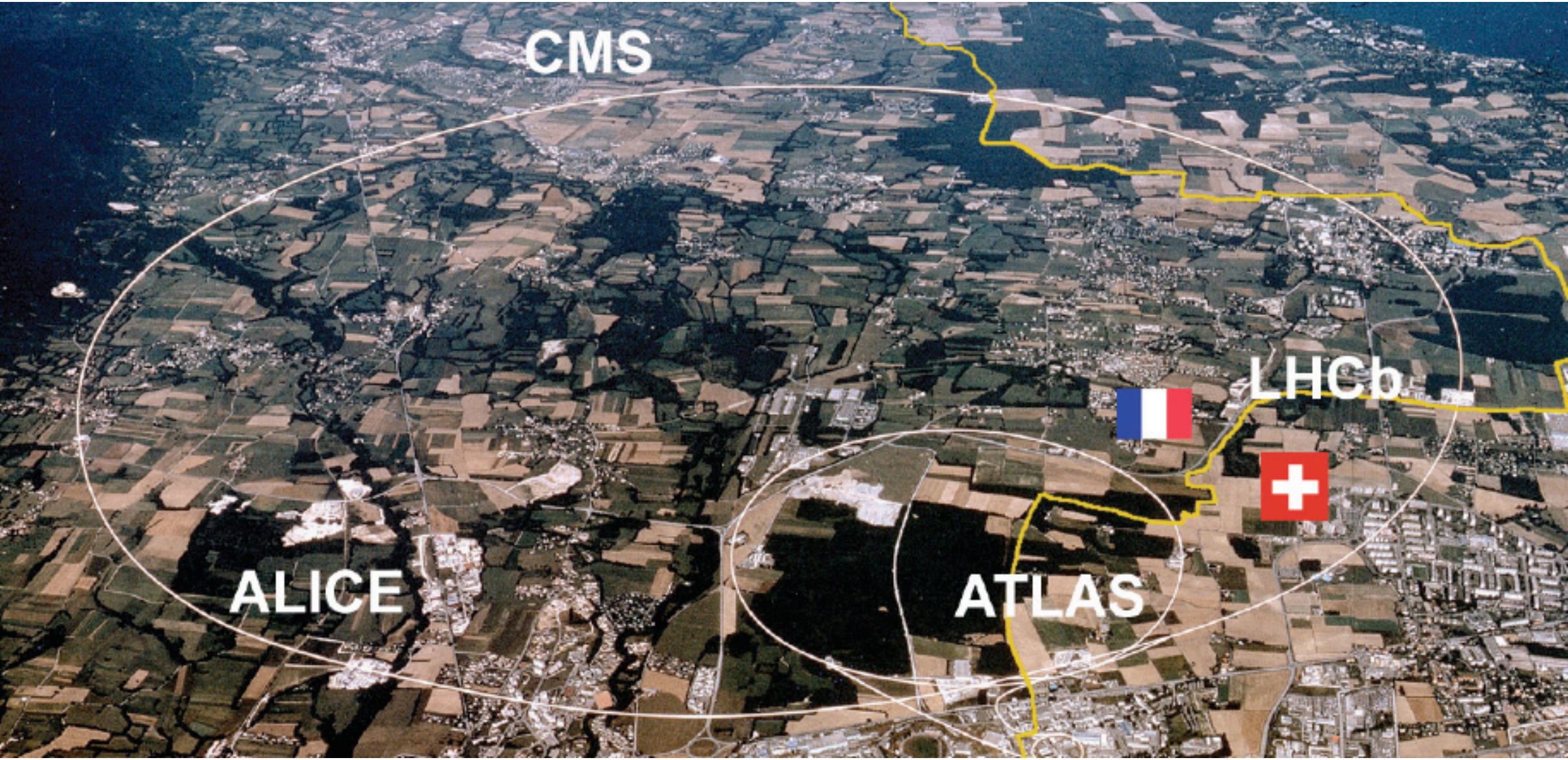


World's largest particle collider — **27 km circumference**



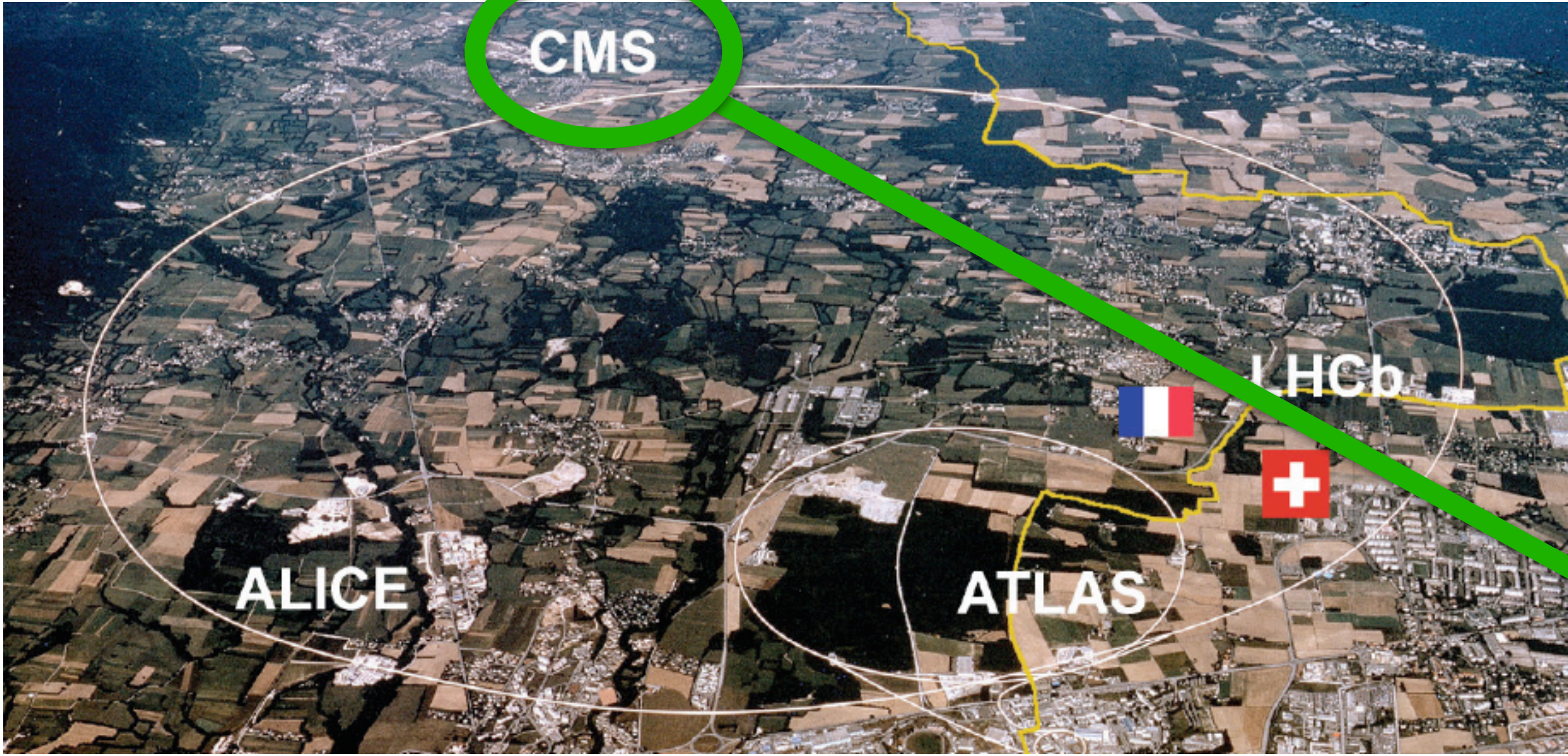
The LHC collides (bunches of) **protons** at a center of mass energy of 13 TeV

The experiments



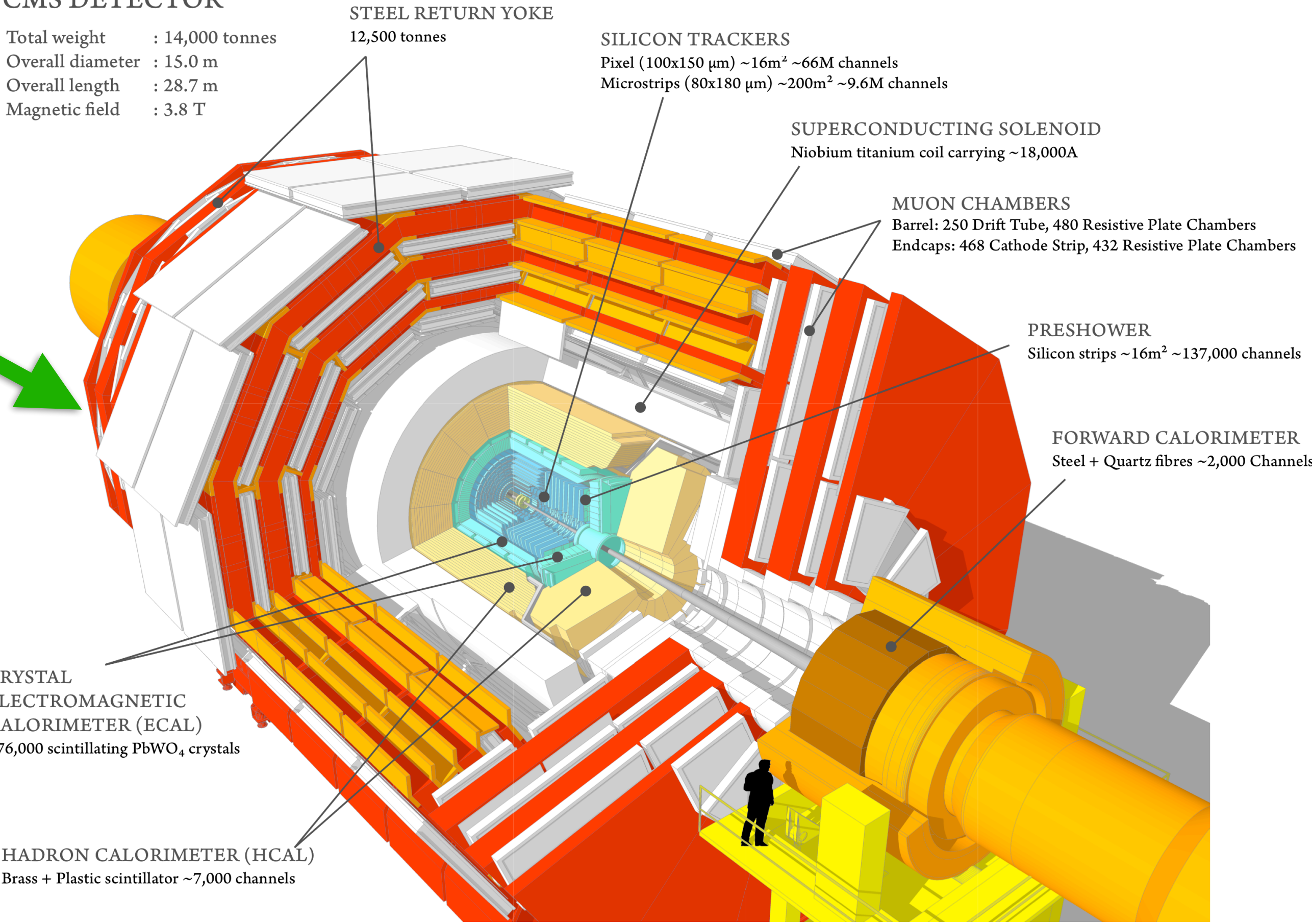
The experiments

The Compact Muon Solenoid (CMS)



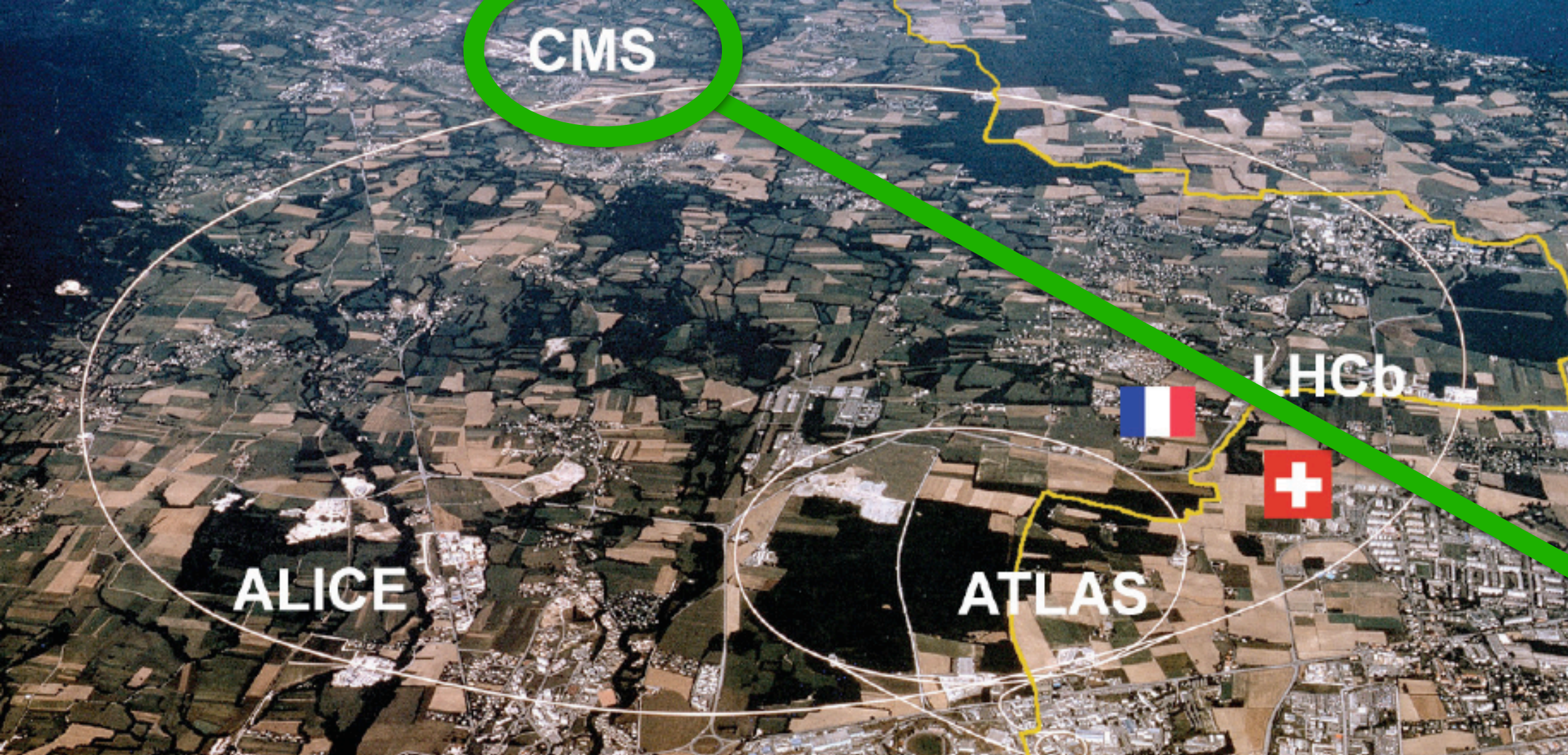
CMS DETECTOR

Total weight : 14,000 tonnes
 Overall diameter : 15.0 m
 Overall length : 28.7 m
 Magnetic field : 3.8 T



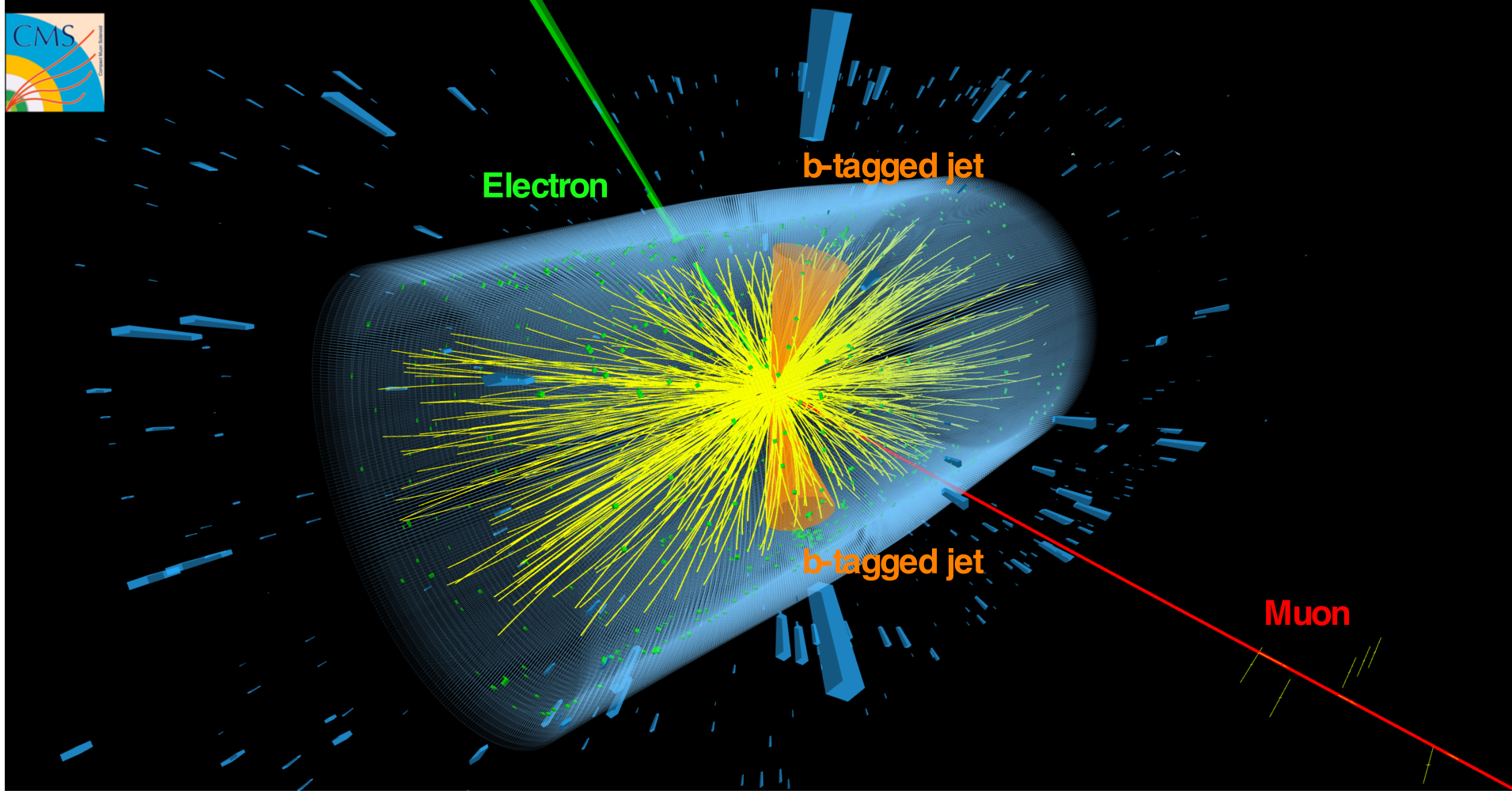
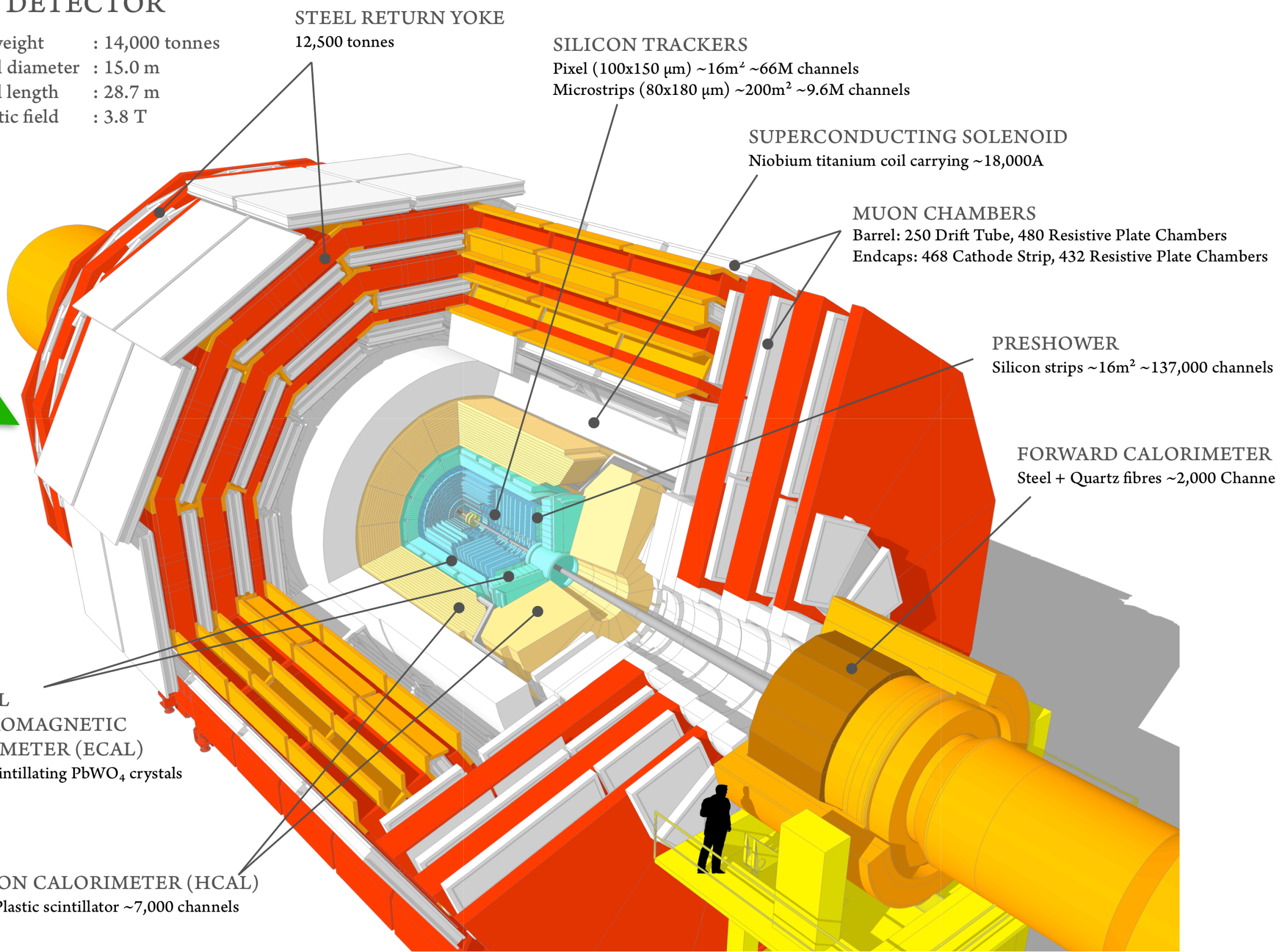
The experiments

The Compact Muon Solenoid (CMS)



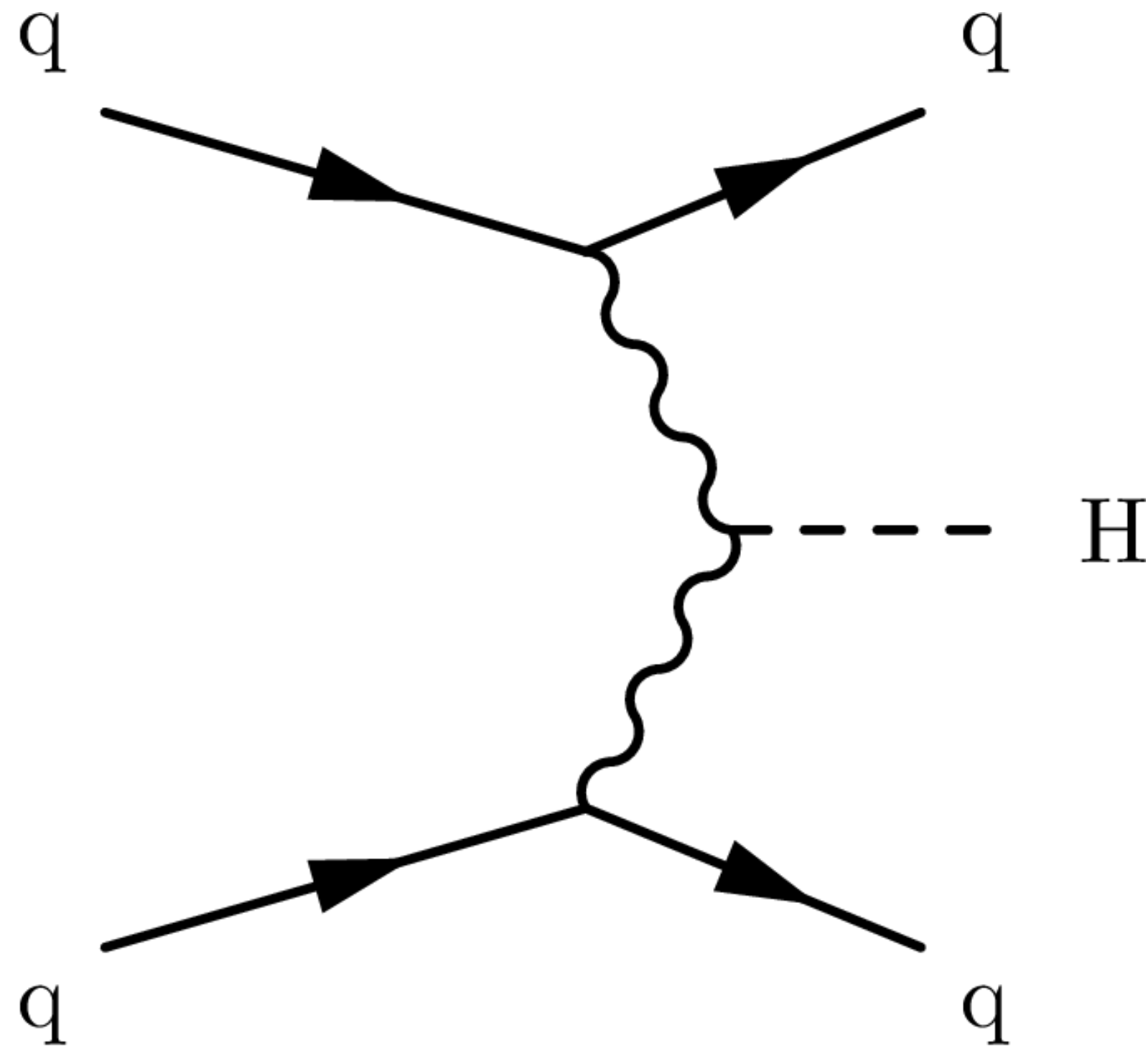
CMS DETECTOR

- Total weight : 14,000 tonnes
- Overall diameter : 15.0 m
- Overall length : 28.7 m
- Magnetic field : 3.8 T

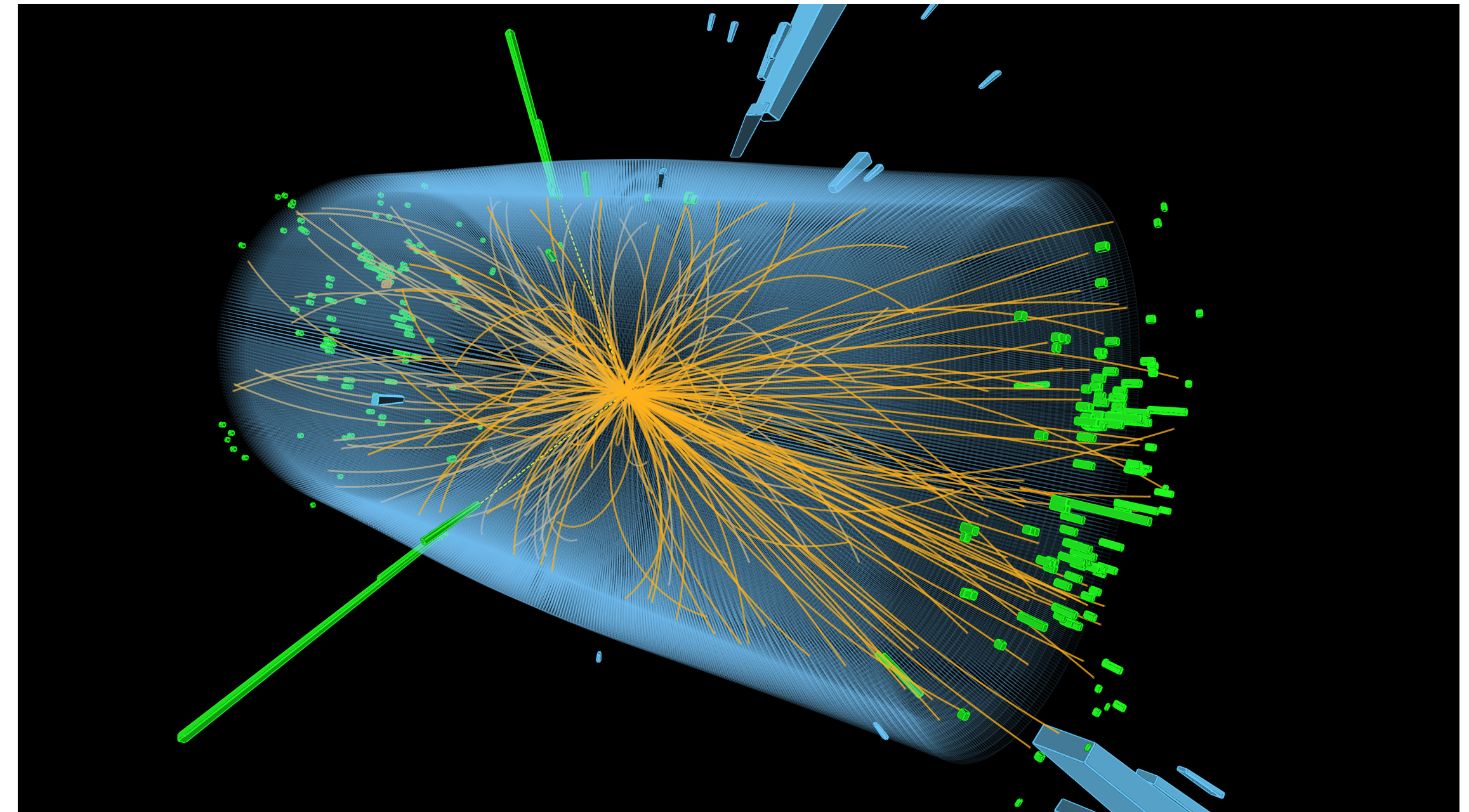


Interpreting the data

What we care about
(and can calculate)



What we can measure



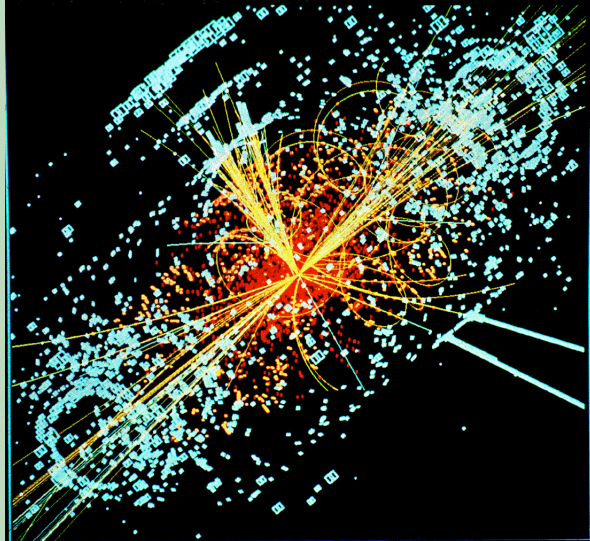
Machine learning is *extremely* well-suited to tackling HEP problems!

AI tools are deeply embedded in LHC analysis ecosystem

Where AI comes in

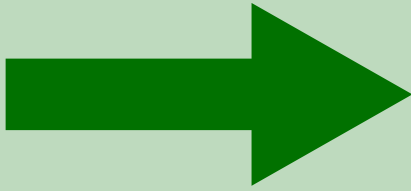


LHC physics workflow



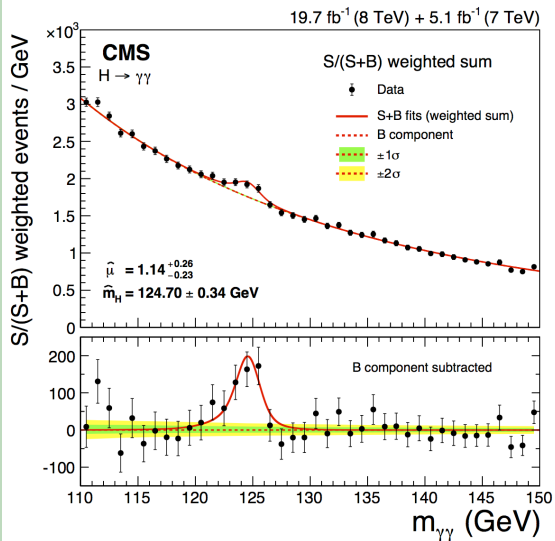
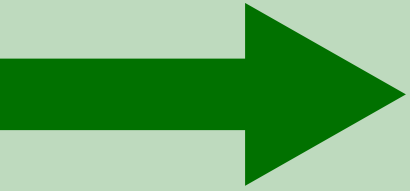
Collide

40 MHz
Overwhelming data rate!



Trigger

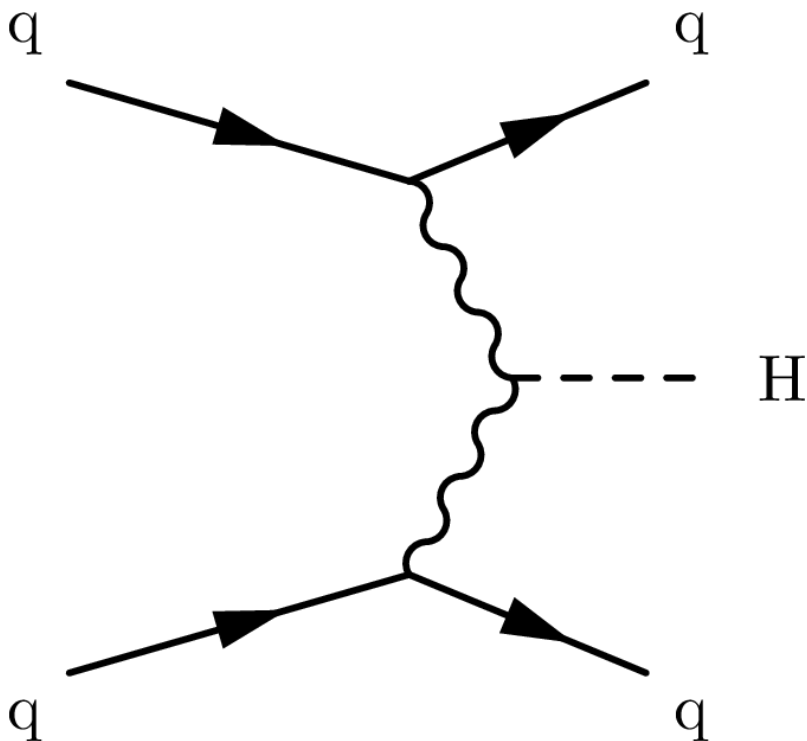
Select only the “interesting” events to save.
Fast, on-chip data reconstruction & inference



Analyze

Reconstruct data, make SM measurements, search for new physics

Nature



Emerging AI frontier

AI is everywhere ...

- Signal vs background discrimination
- **Taggers** (e.g. Higgs)
- **Anomaly detection** (new physics)

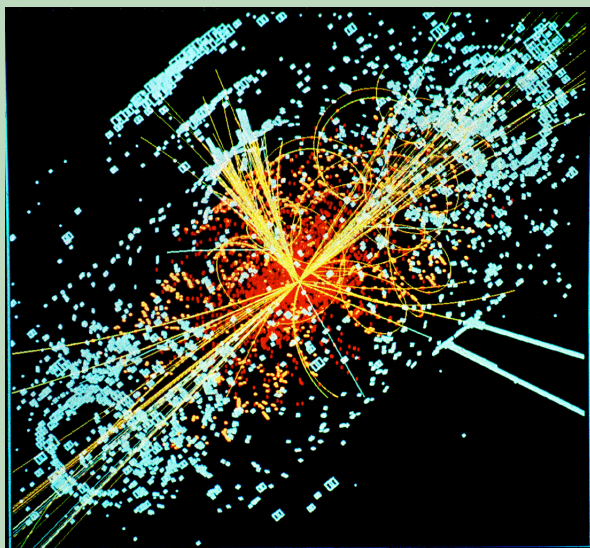
Huge, rapidly evolving field!

[Hep ML Living review](#)
(~1.4k papers & counting)

Where AI comes in

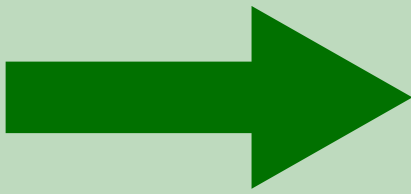


LHC physics workflow



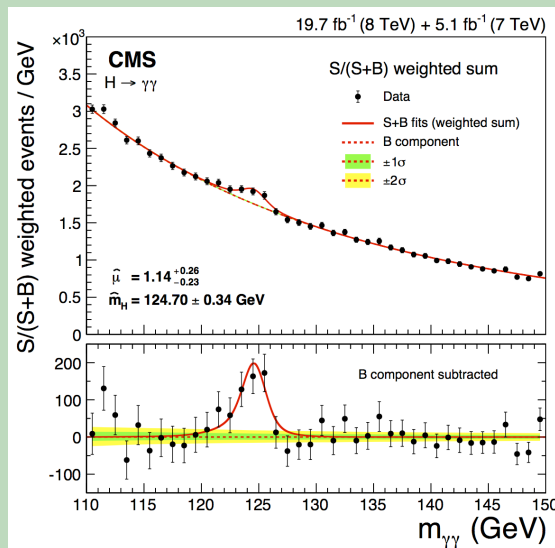
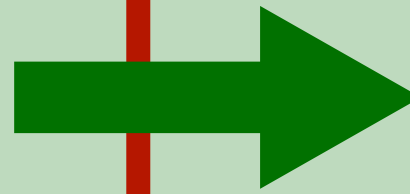
Collide

40 MHz
Overwhelming data rate!



Trigger

Select only the “interesting” events to save.
Fast, on-chip data reconstruction & inference

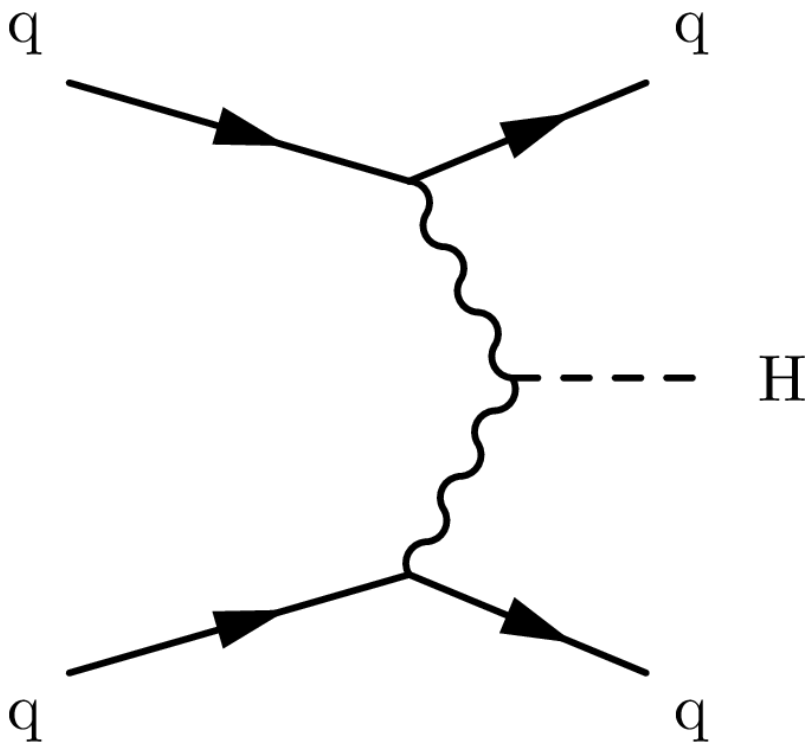


This talk

Analyze

Reconstruct data, make SM measurements, search for new physics

Nature



Emerging AI frontier

hls 4 ml
2103.05579

Anomaly trigger

FastML

AI is everywhere ...

- Signal vs background discrimination
- Taggers (e.g. Higgs)
- Anomaly detection (new physics)

Huge, rapidly evolving field!

[Hep ML Living review](#)
(~1.4k papers & counting)

Jets

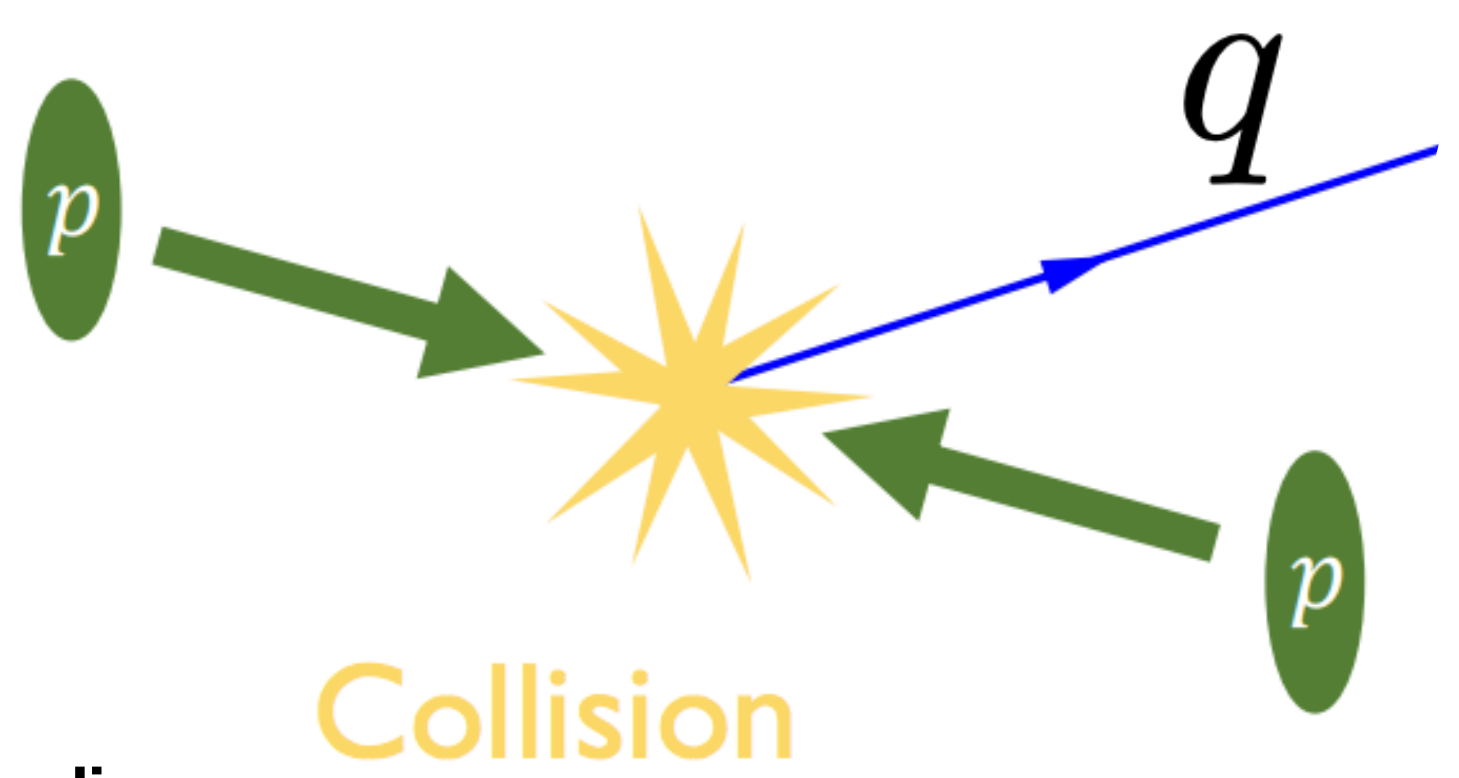


Image credit: E. Metodiev



Jets

- We can never observe “bare” quarks/gluons due to a QCD phenomenon called **color confinement**

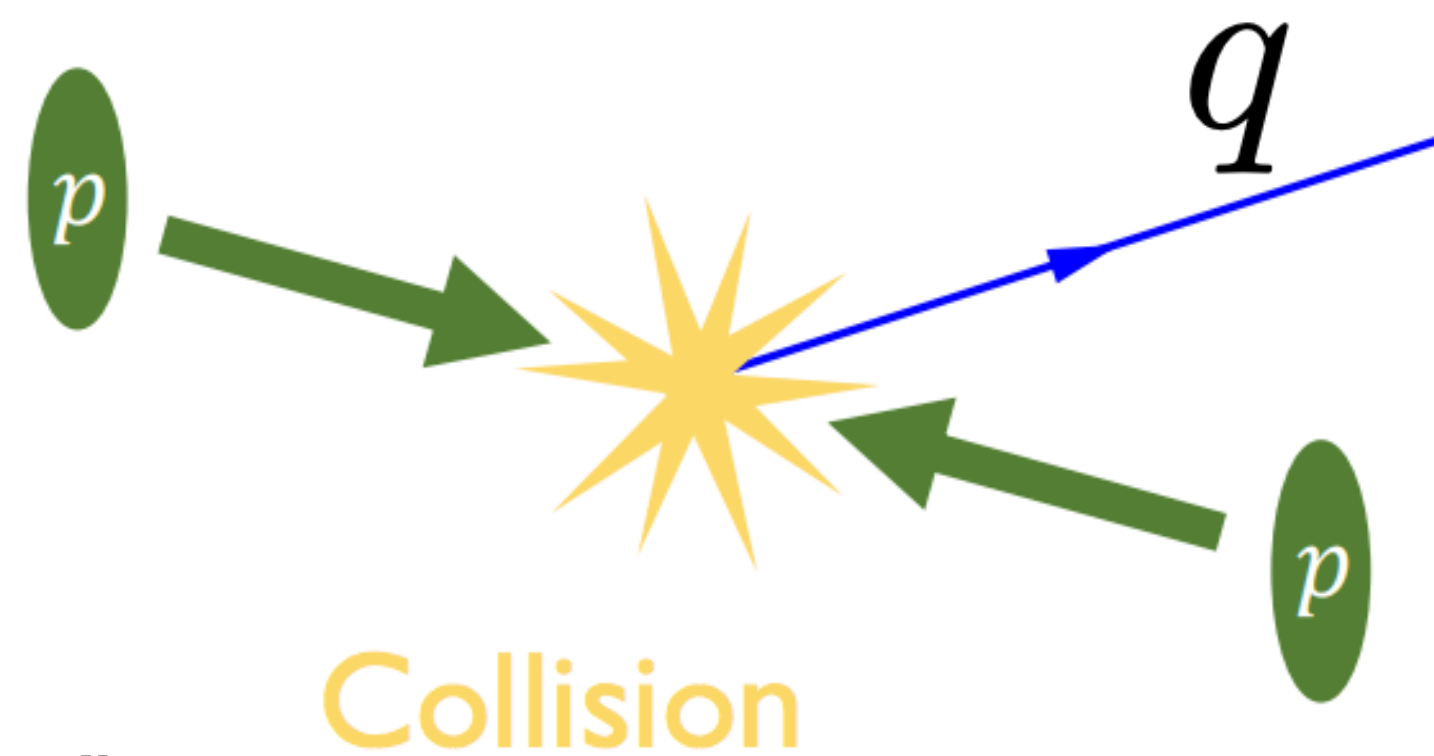


Image credit: E. Metodiev



Jets

- We can never observe “bare” quarks/gluons due to a QCD phenomenon called **color confinement**

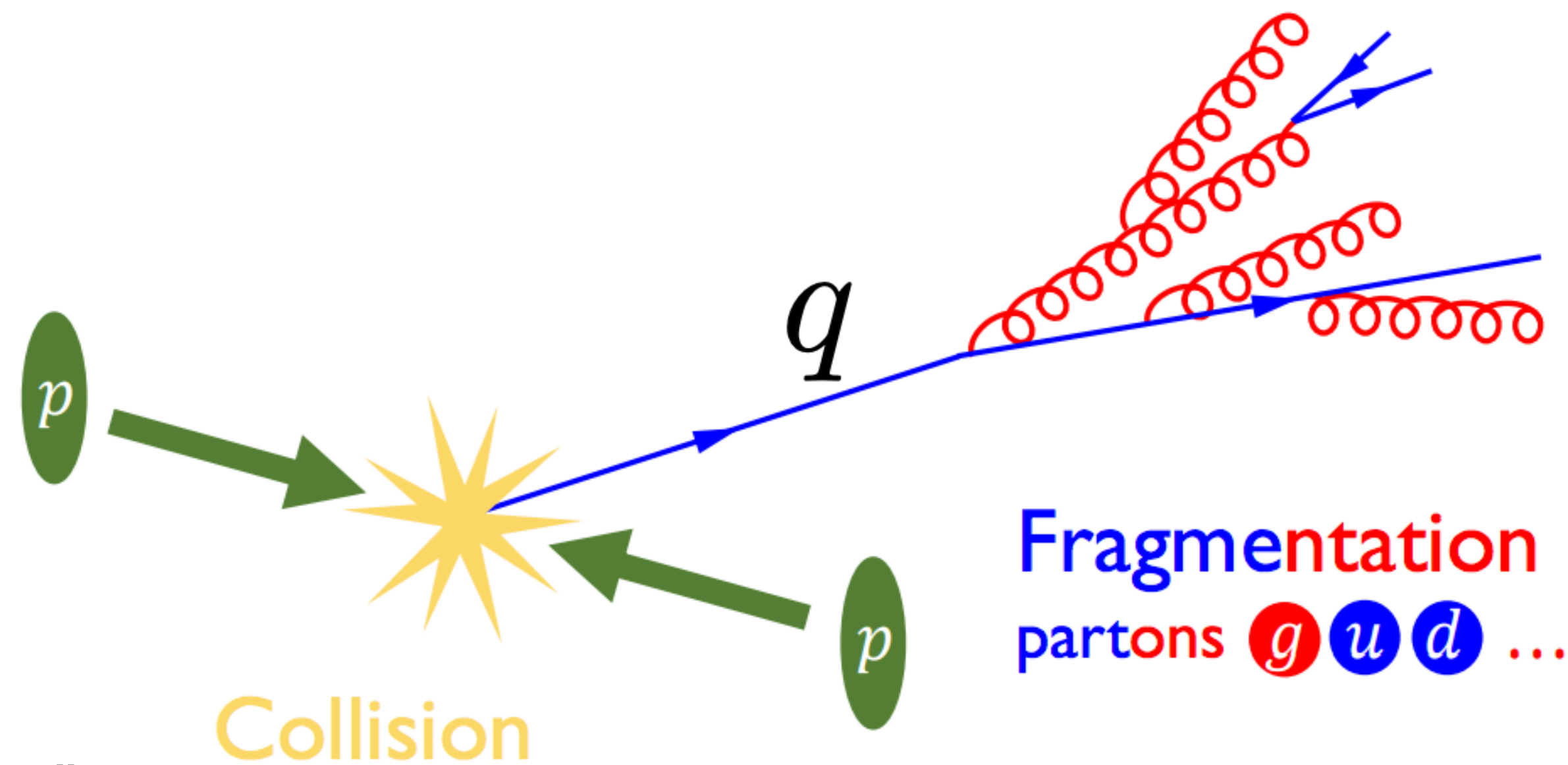


Image credit: E. Metodiev

Jets



- We can never observe “bare” quarks/gluons due to a QCD phenomenon called **color confinement**

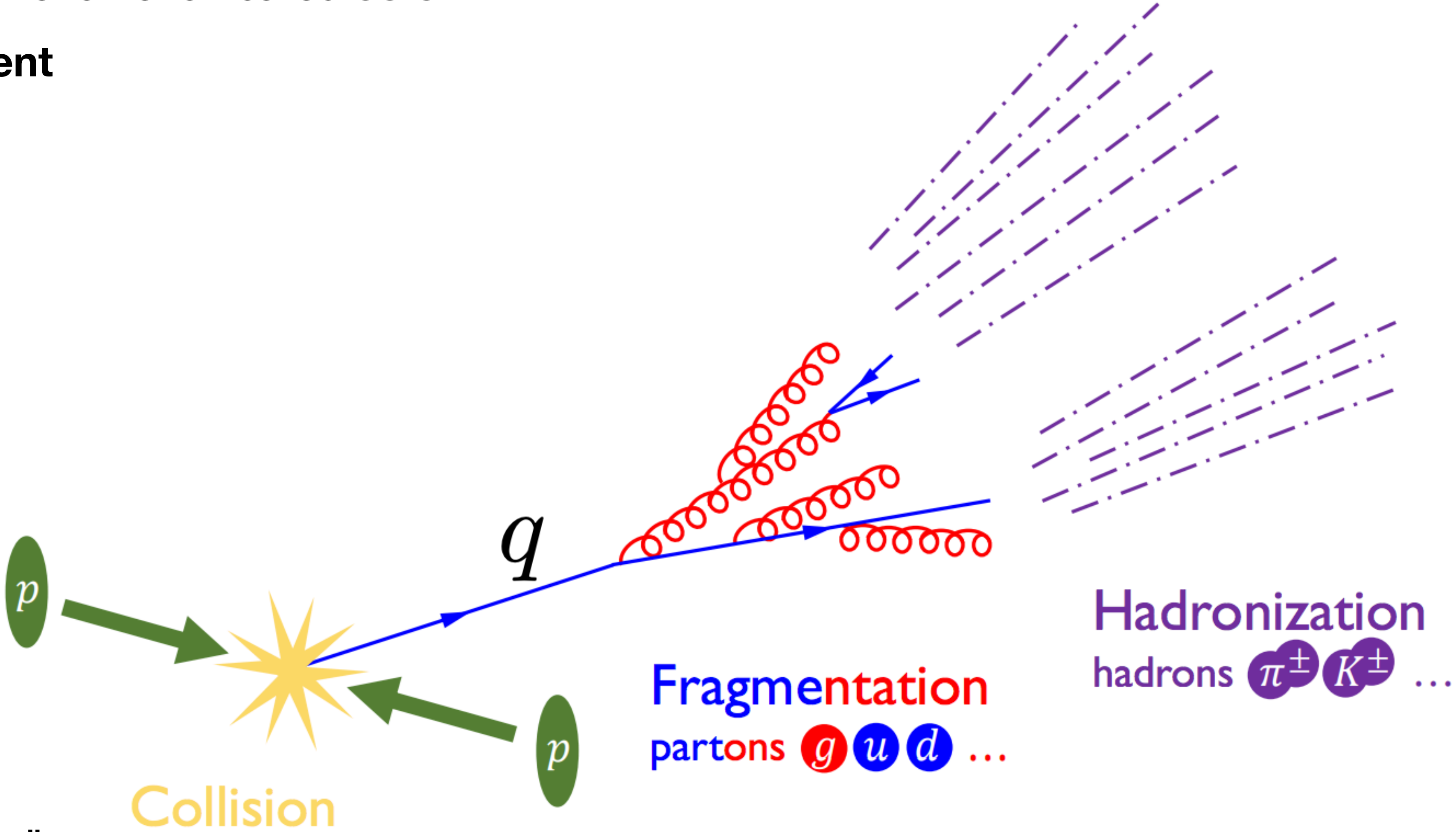


Image credit: E. Metodiev

Jets

- We can never observe “bare” quarks/gluons due to a QCD phenomenon called **color confinement**

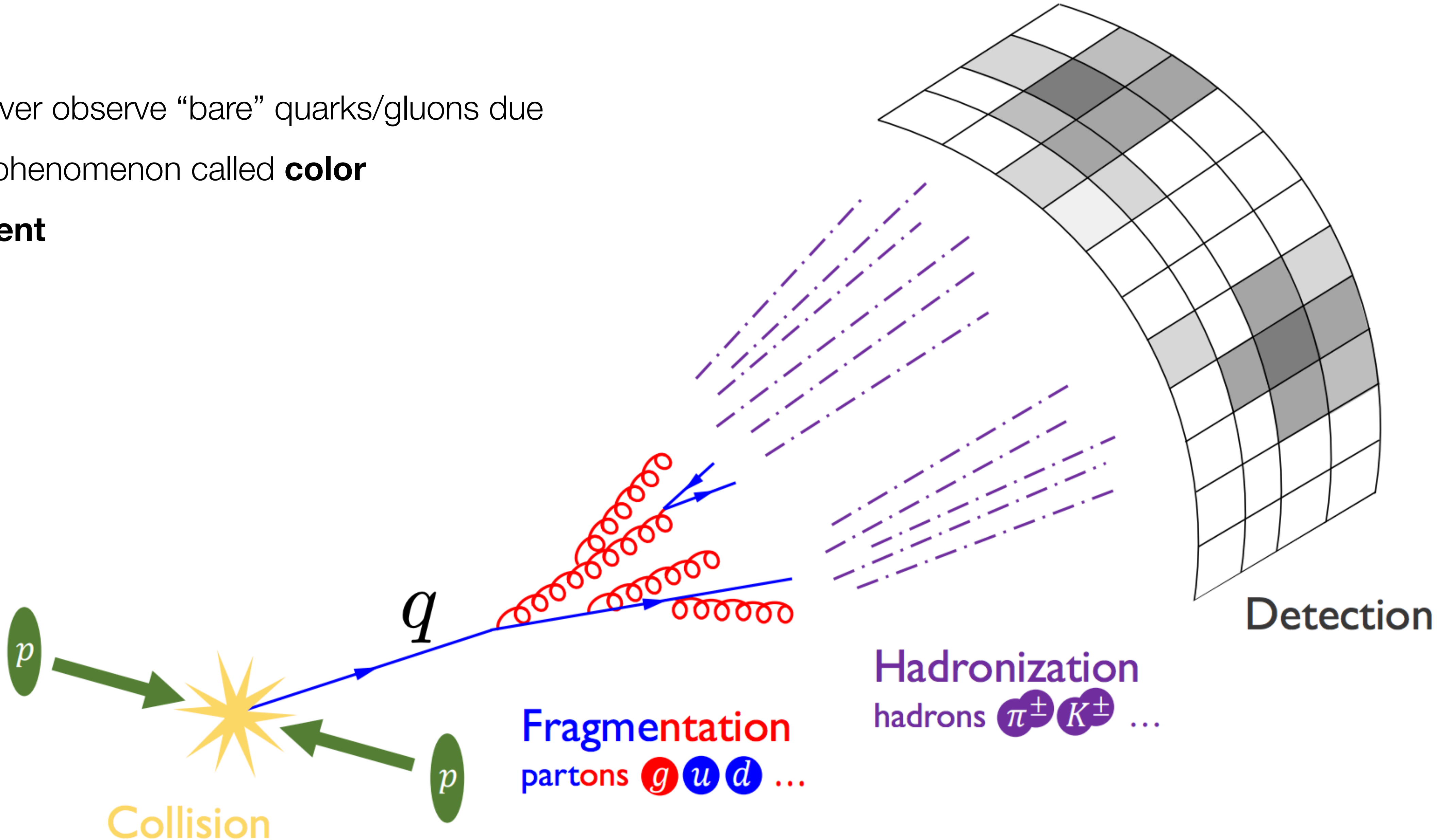


Image credit: E. Metodiev

Jets

- We can never observe “bare” quarks/gluons due to a QCD phenomenon called **color confinement**
- We detect *composite* objects called **jets** as a proxy for the underlying q/g

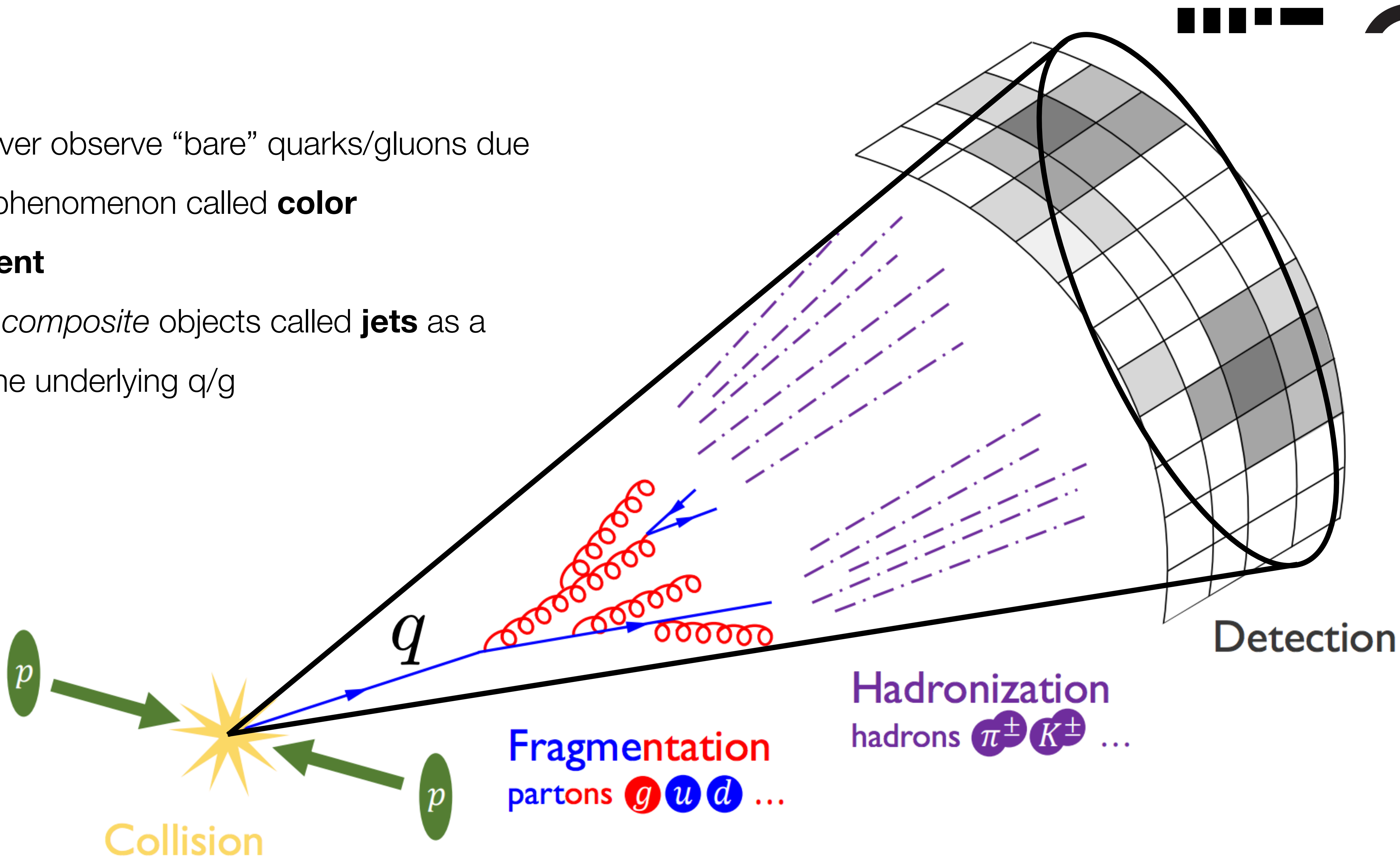


Image credit: E. Metodiev

Jets

- We can never observe “bare” quarks/gluons due to a QCD phenomenon called **color confinement**
- We detect *composite* objects called **jets** as a proxy for the underlying q/g
- Jet **substructure** can tell us about the initiating particle

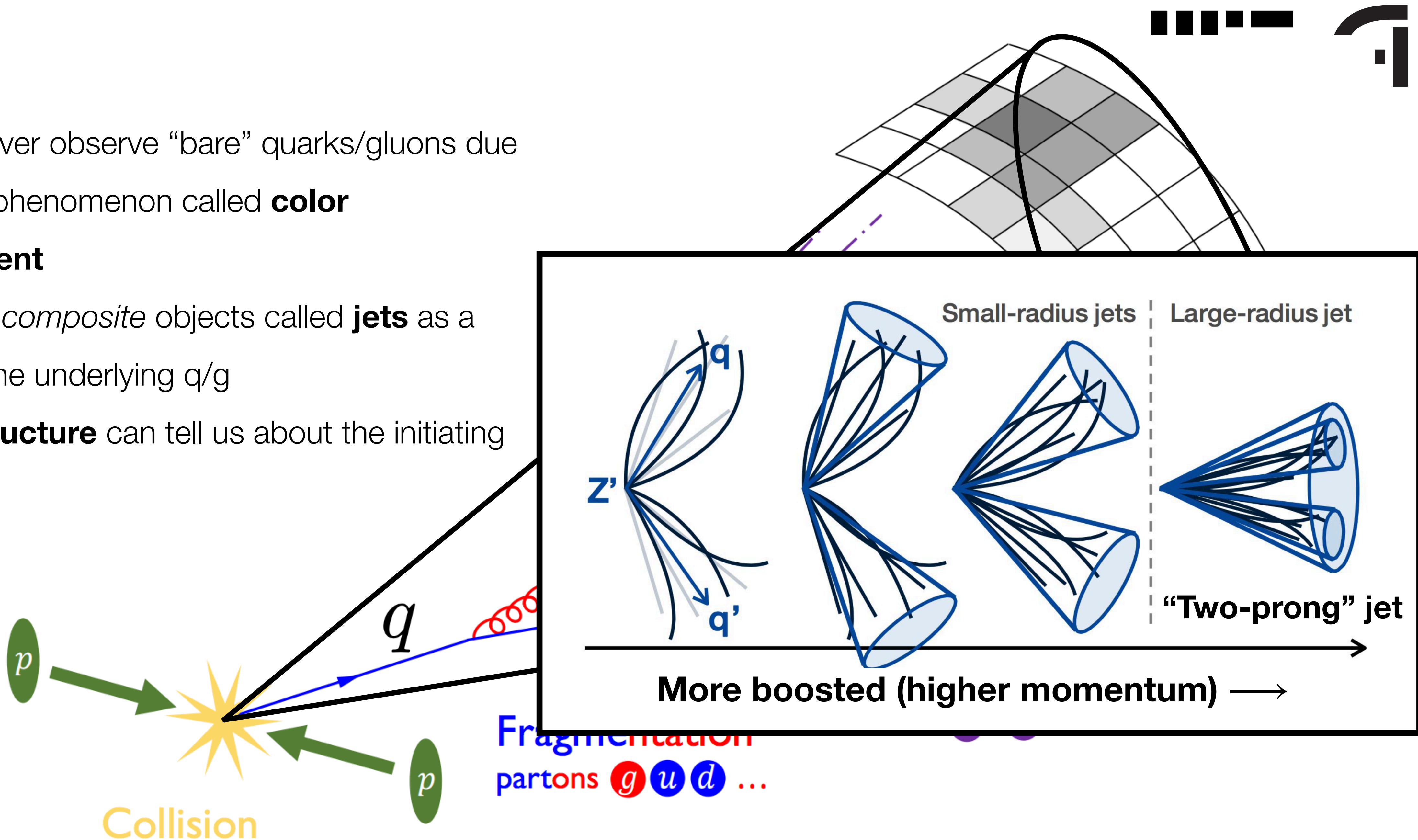


Image credit: E. Metodiev

Semi-supervised spaces

Building a space “by hand” for anomaly detection

Some motivation

After ~15 years of LHC physics, no clear experimental target for new physics



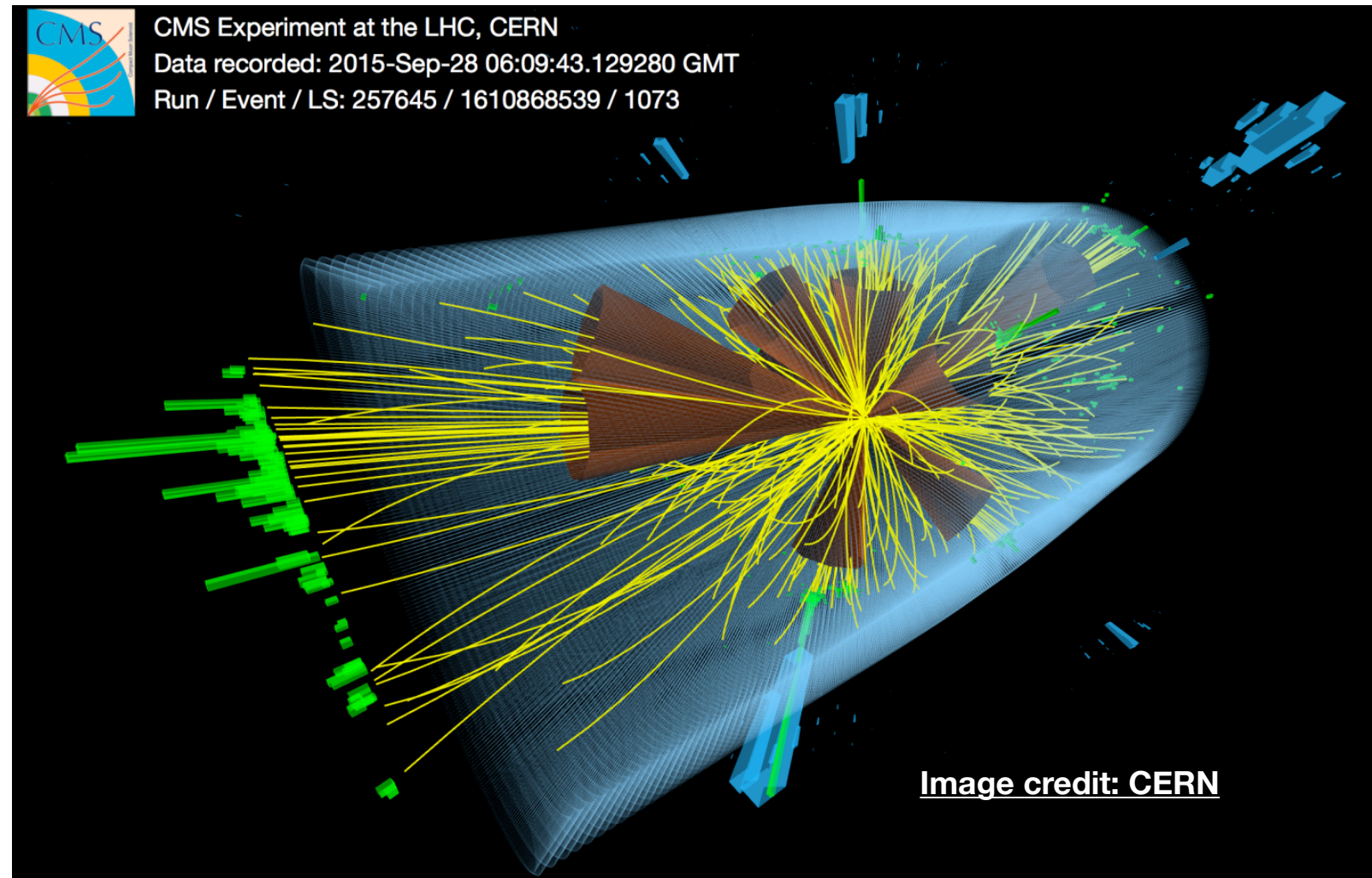
“Theory Space”



An ideal CMS search

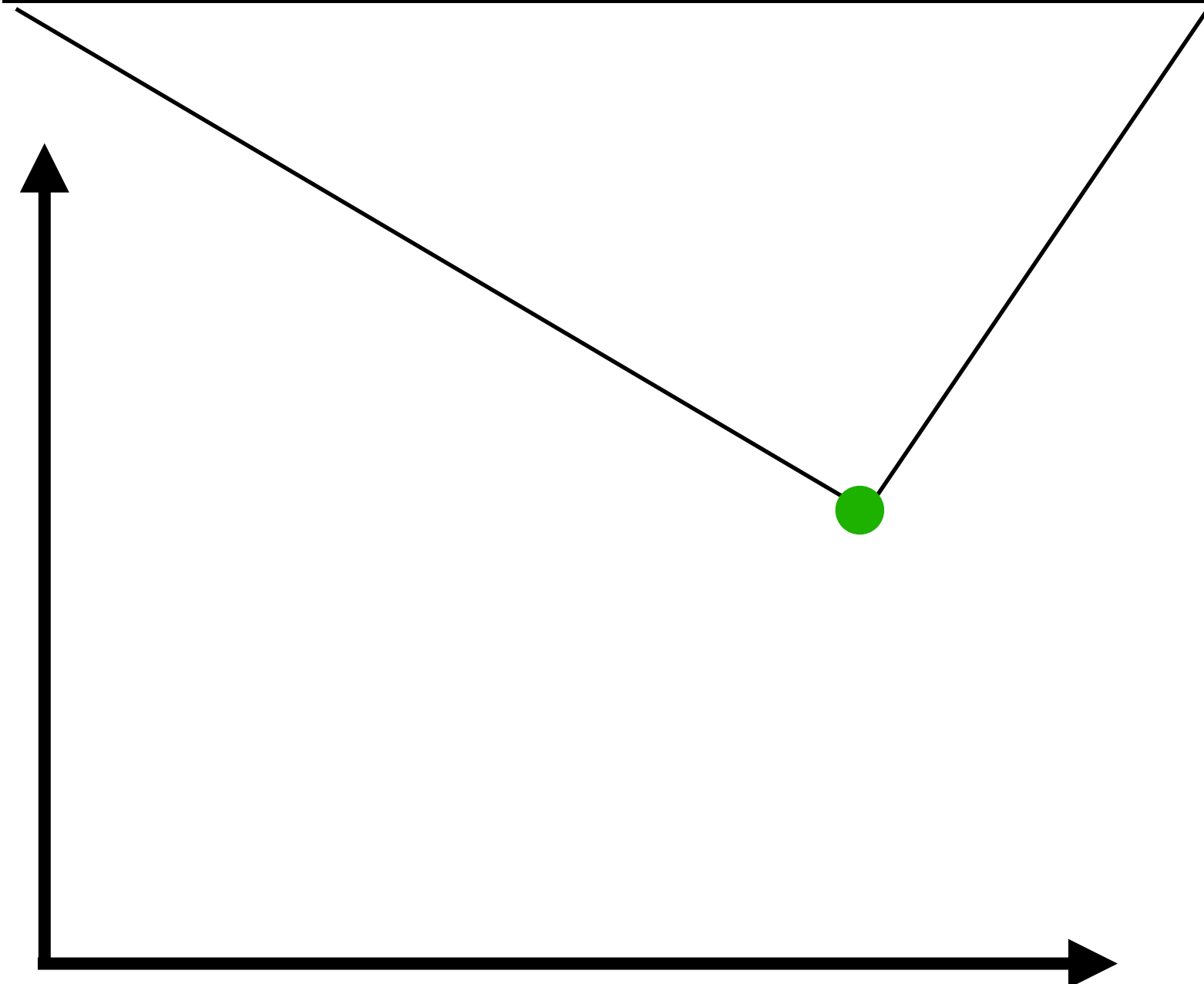
We should make our searches as broad as possible – *model agnostic*

How an embedding might help

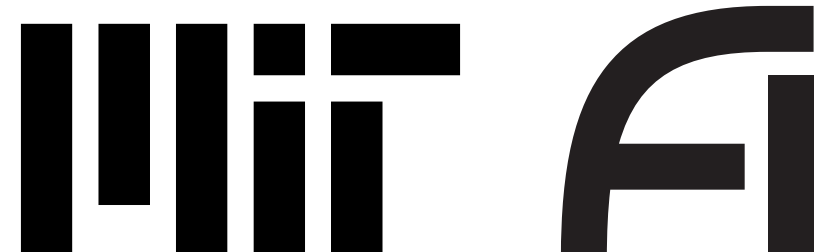


Why would an embedding be helpful for these searches?

1. Anomaly detection in high-dim datasets is **fundamentally hard**: difficult to model, more “tails”, sparse
2. Partially alleviate “performance tradeoff” btwn low-level particle info & hand-constructed observables?
3. Interpretation *might* be easier, depending on how interpretable you can make your space



Designing a search



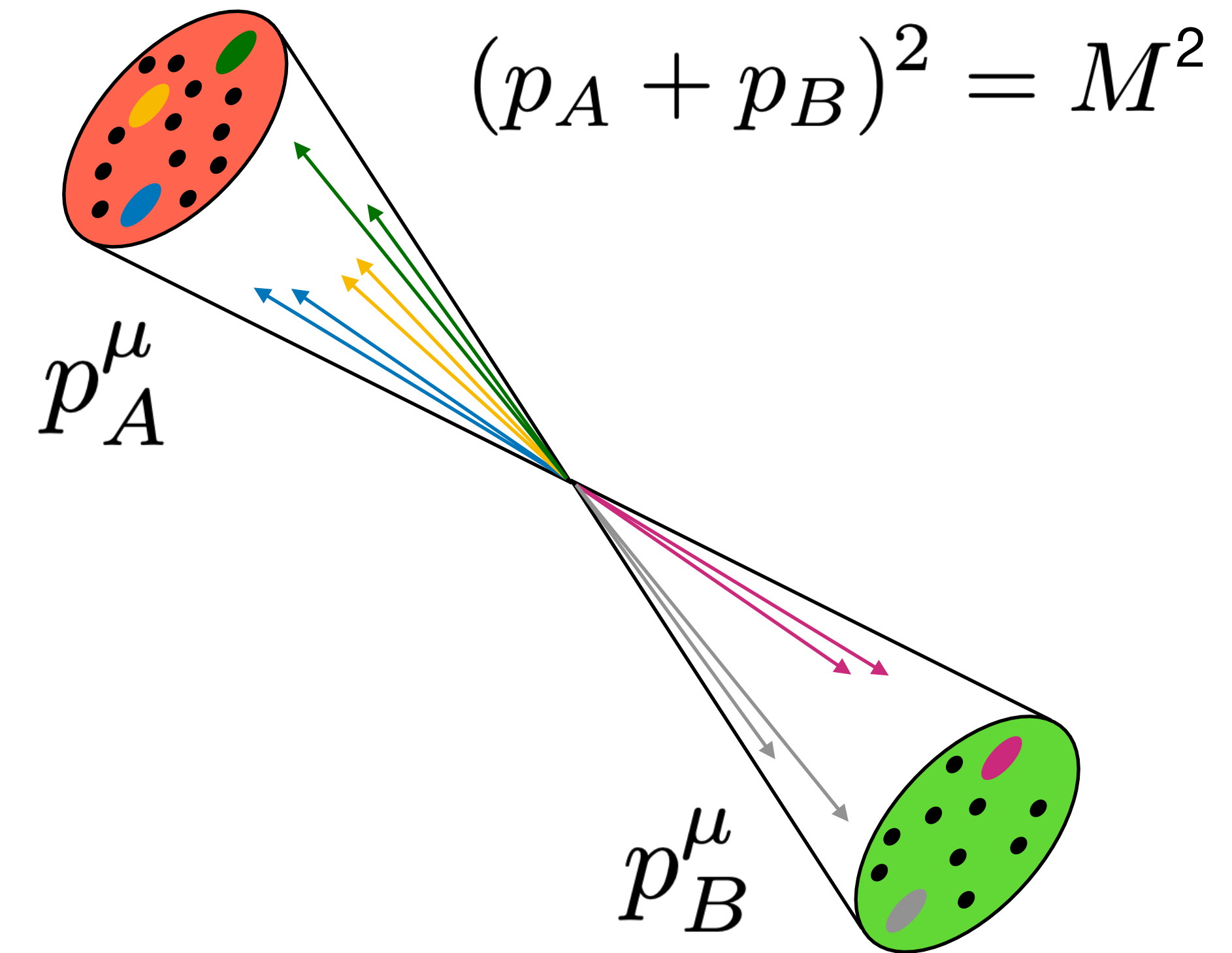
Designing a search

- What's the most generic thing we can look for?
 - ▶ A **new particle** (anything)



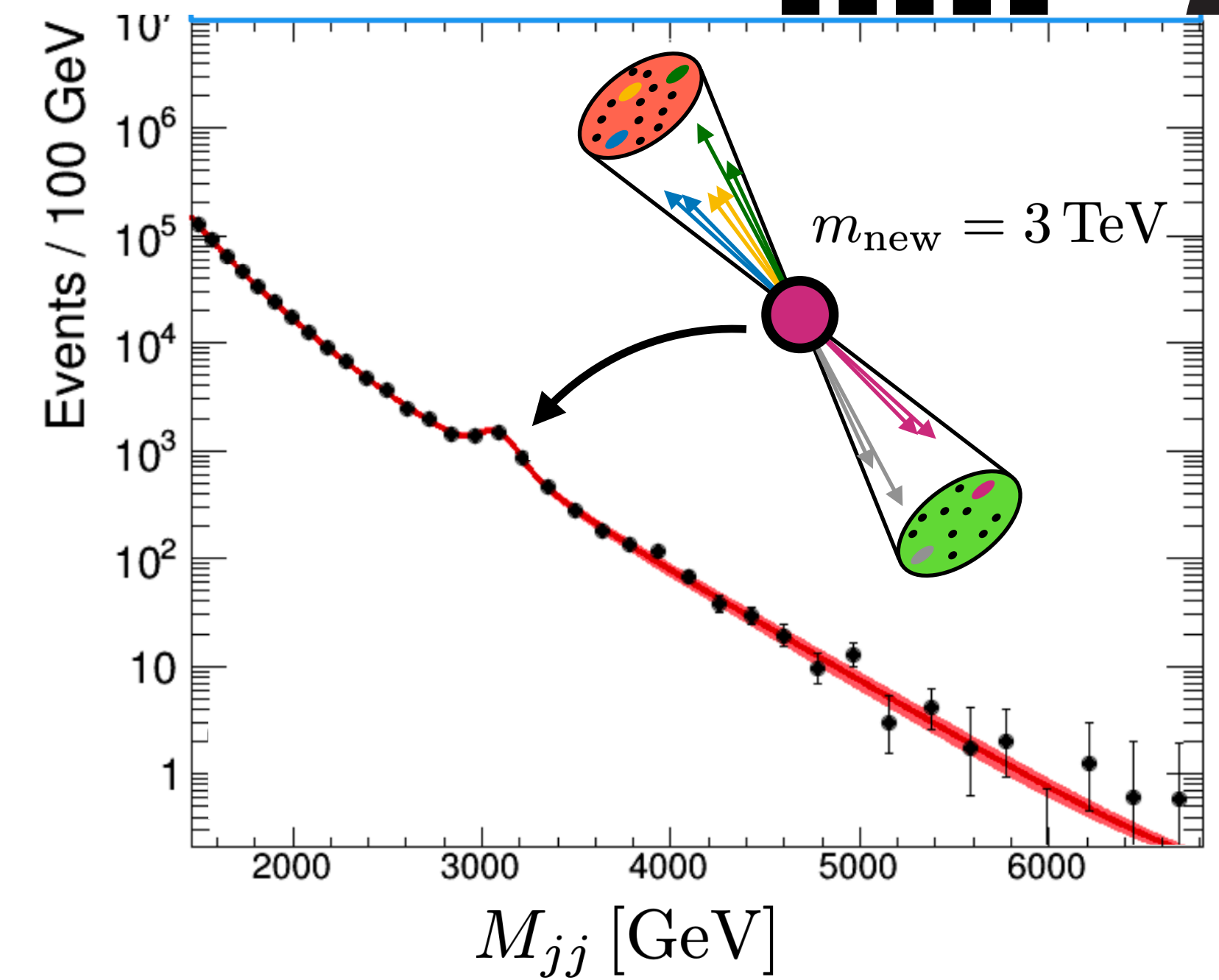
Designing a search

- What's the most generic thing we can look for?
 - ▶ A **new particle** (anything)
- What **physical signature**?
 - ▶ Two-body decay, i.e. a **resonance**
 - ▶ Assume only **hadronic decays** (jets)

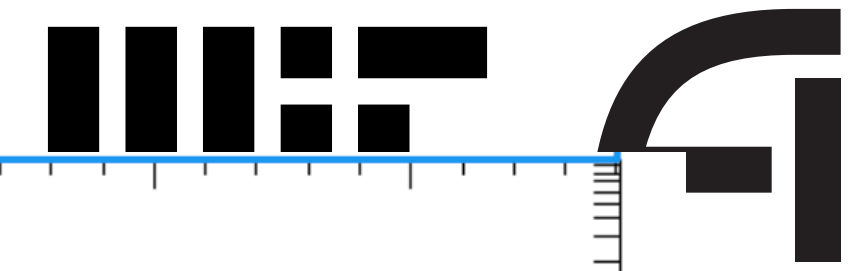


Designing a search

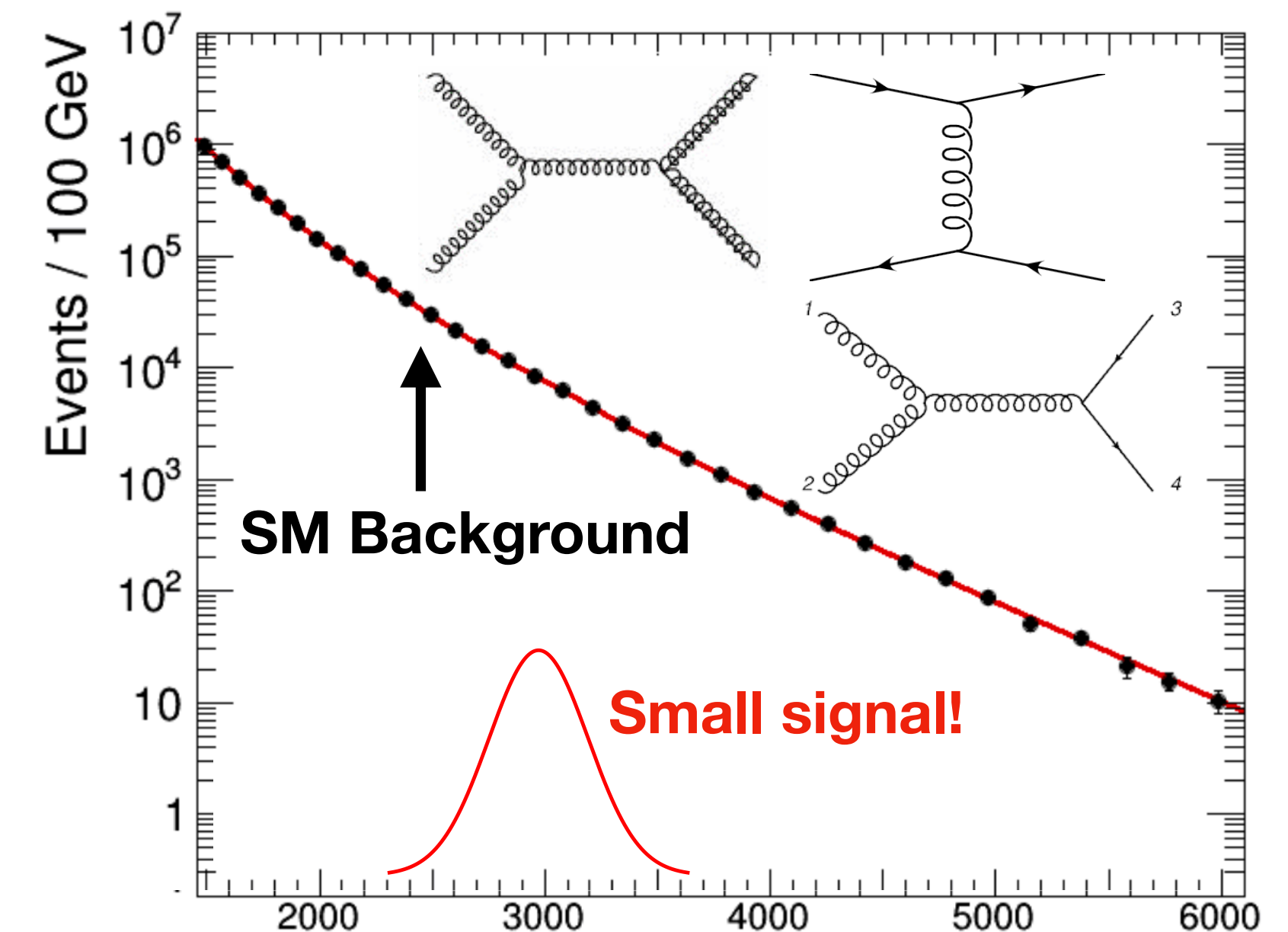
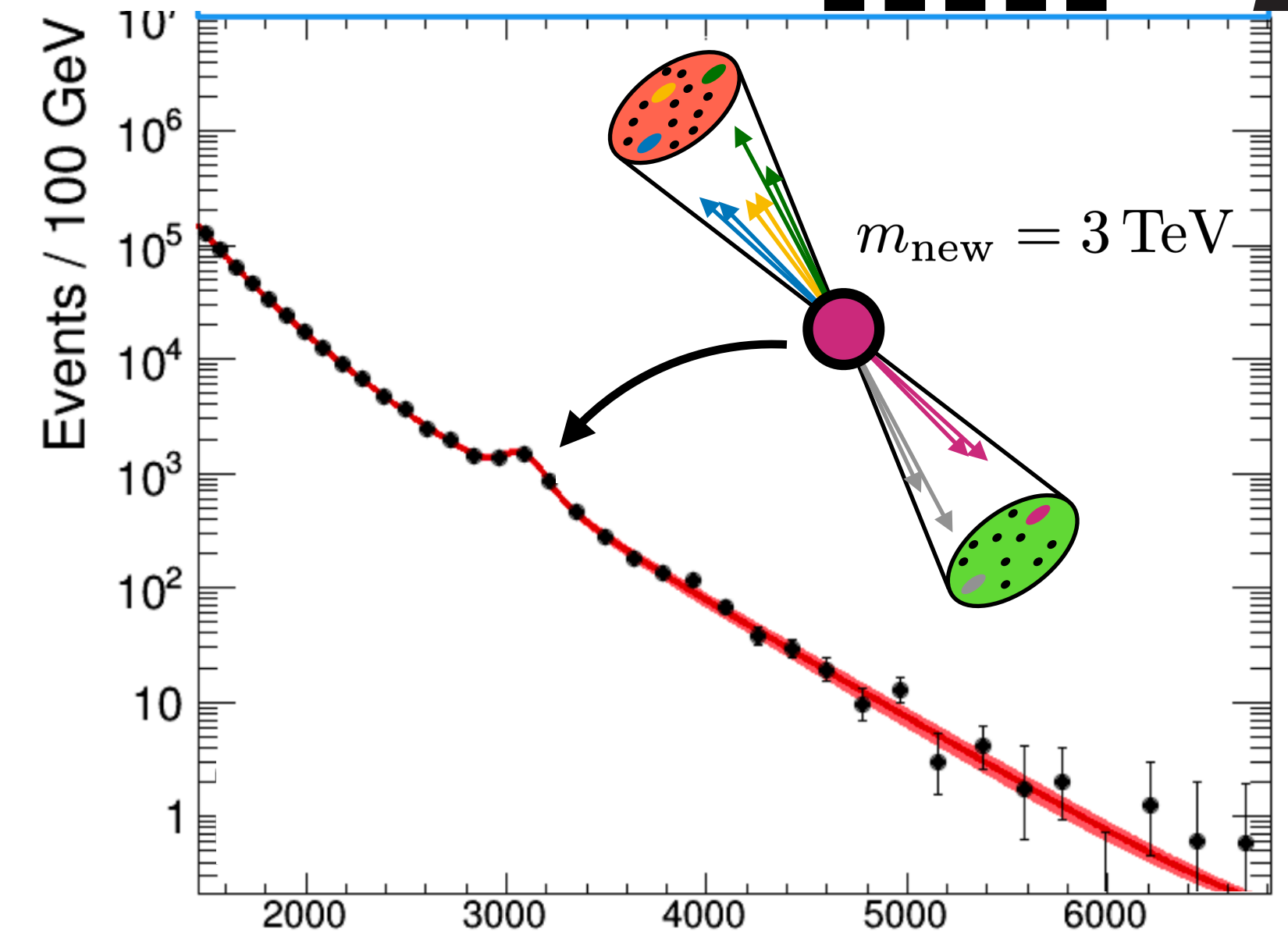
- What's the most generic thing we can look for?
 - ▶ A **new particle** (anything)
- What **physical signature**?
 - ▶ Two-body decay, i.e. a **resonance**
 - ▶ Assume only **hadronic decays** (jets)
- We call this a “**dijet resonance**” — performing a **bump hunt**



Designing a search



- What's the most generic thing we can look for?
 - ▶ A **new particle** (anything)
- What **physical signature**?
 - ▶ Two-body decay, i.e. a **resonance**
 - ▶ Assume only **hadronic decays** (jets)
- We call this a “**dijet resonance**” — performing a **bump hunt**
 - ▶ Easier said than done — **QCD** (boring stuff) is a huge background

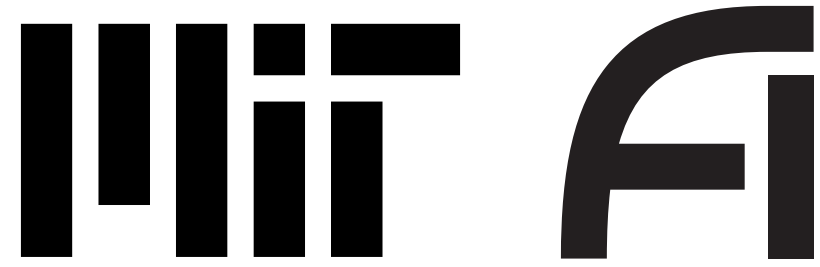


Semi-supervised searches with QUAK

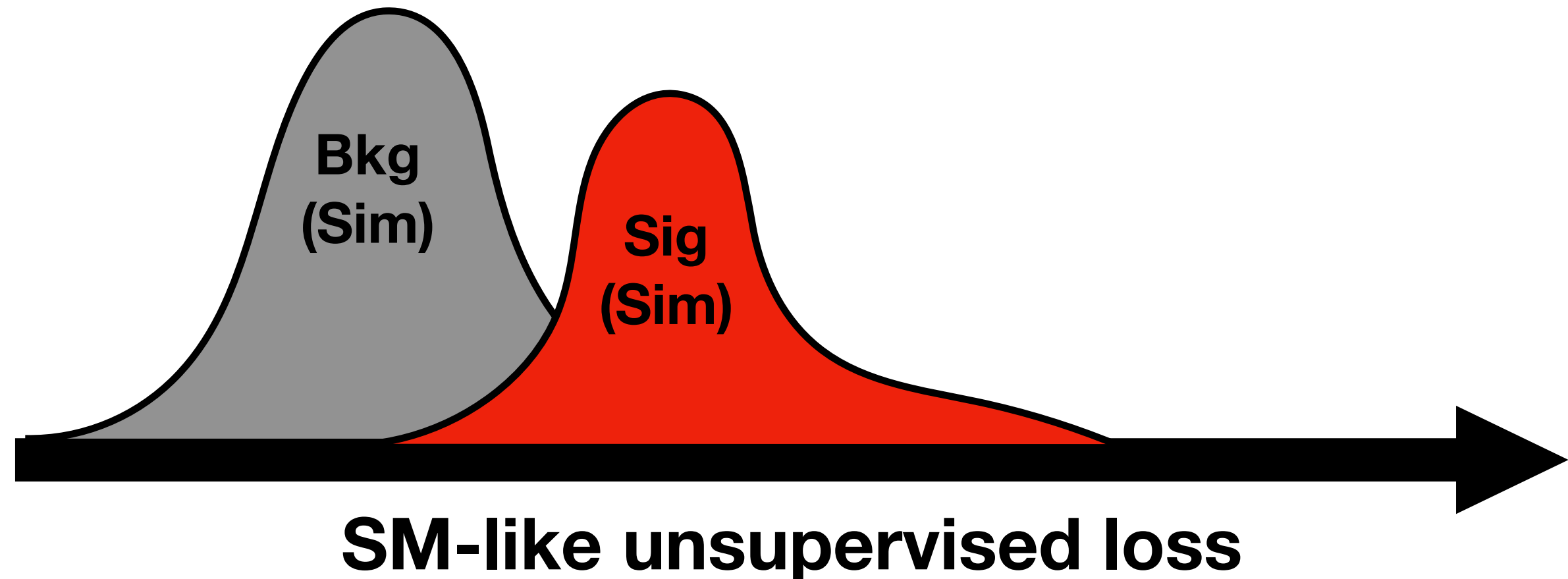


SM-like unsupervised loss

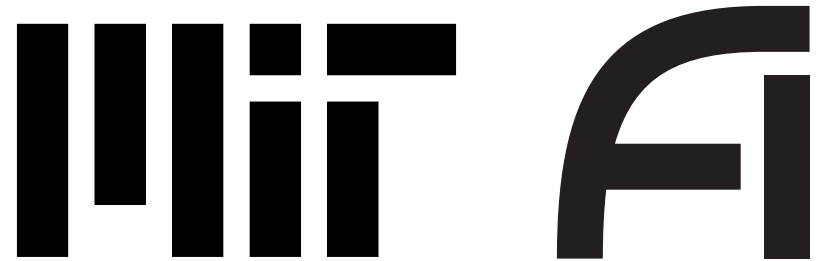
Semi-supervised searches with QUAK



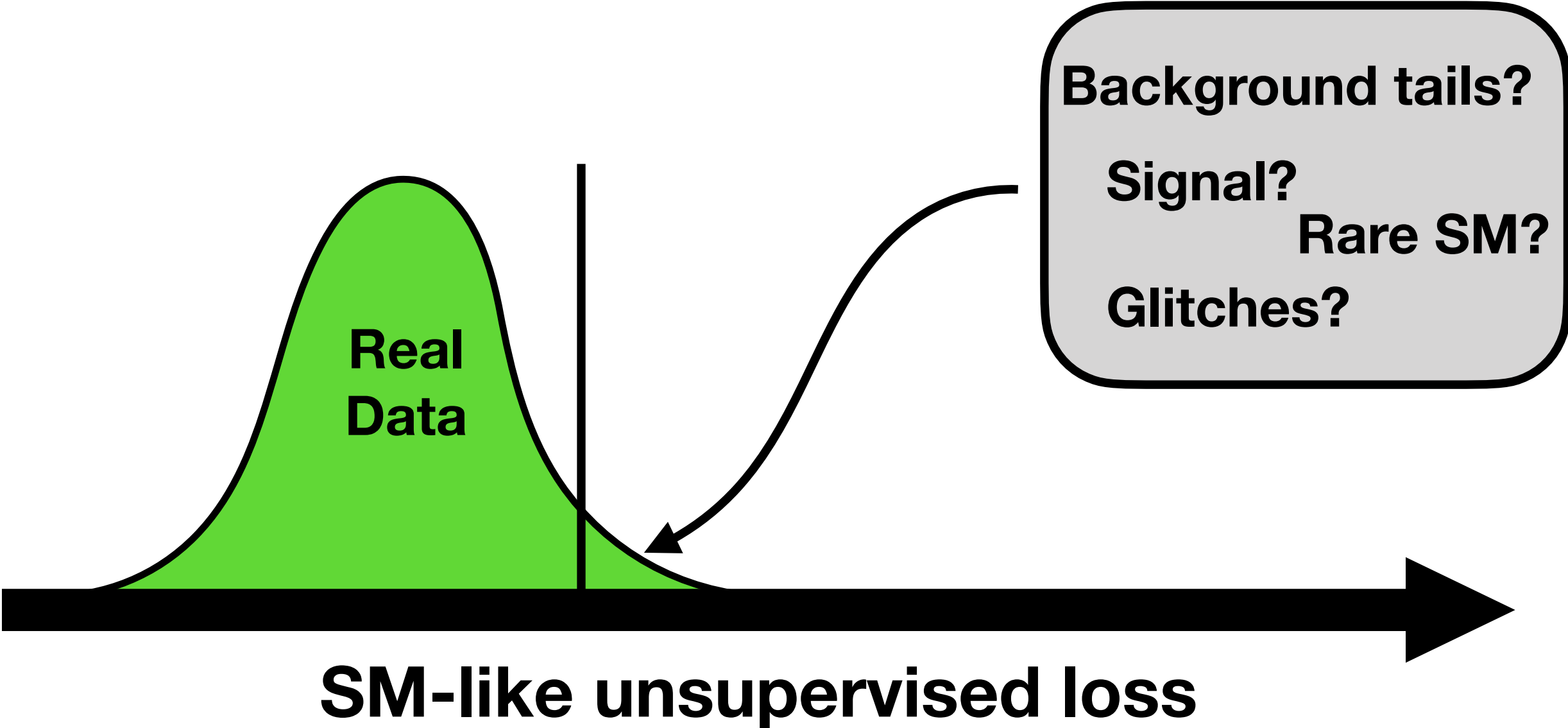
- Imagine you train an unsupervised model on SM backgrounds
 - ▶ The “high loss” region is where we expect signal — potentially very sensitive!



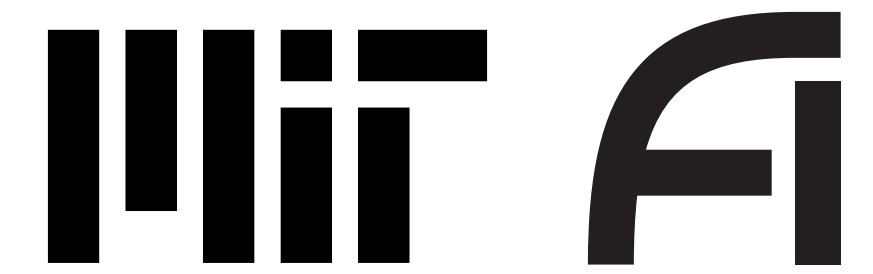
Semi-supervised searches with QUAK



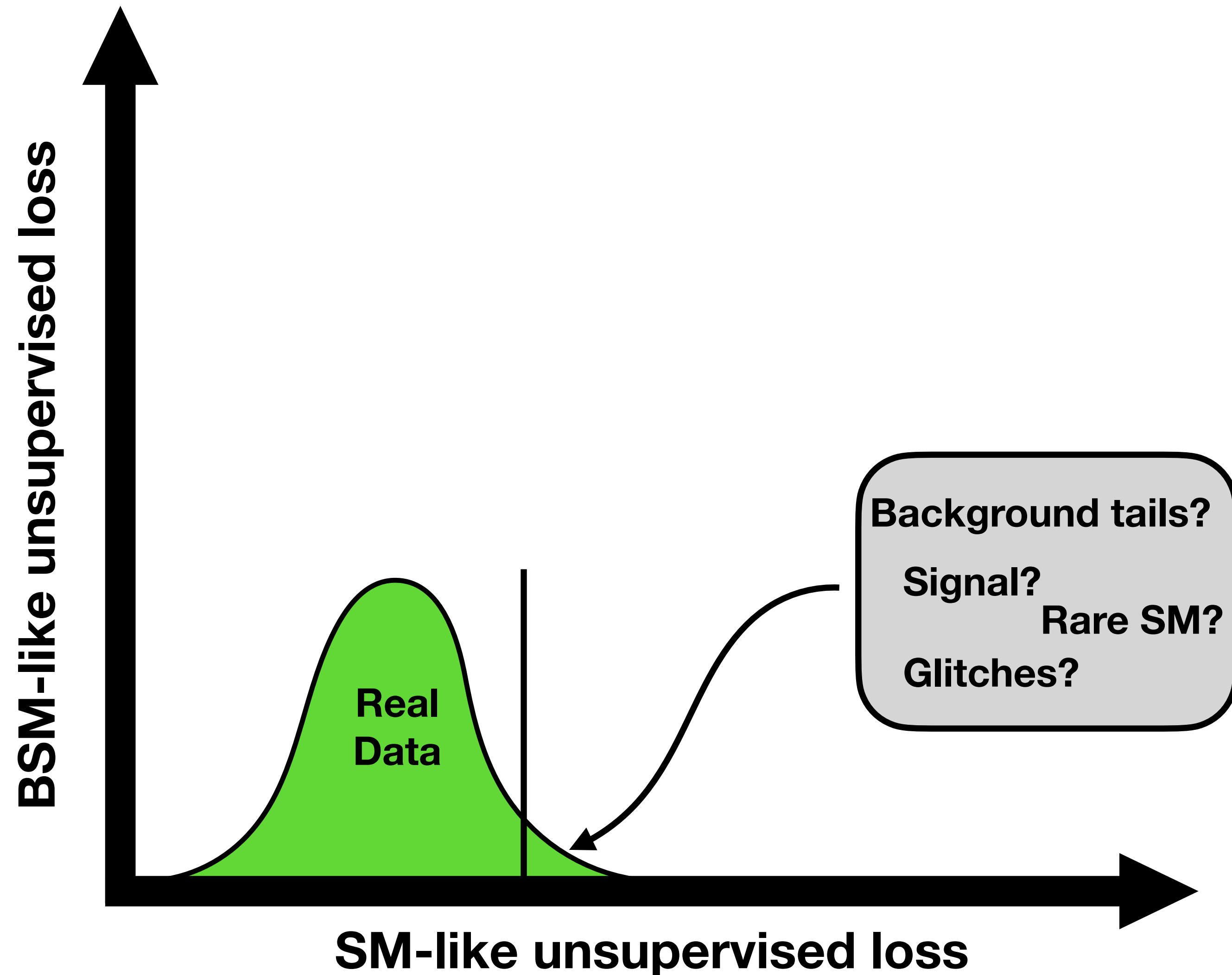
- Imagine you train an unsupervised model on SM backgrounds
 - ▶ The “high loss” region is where we expect signal — potentially very sensitive!
 - ▶ But in reality, tails are populated by all kinds of *uninteresting* anomalies



Semi-supervised searches with QUAK

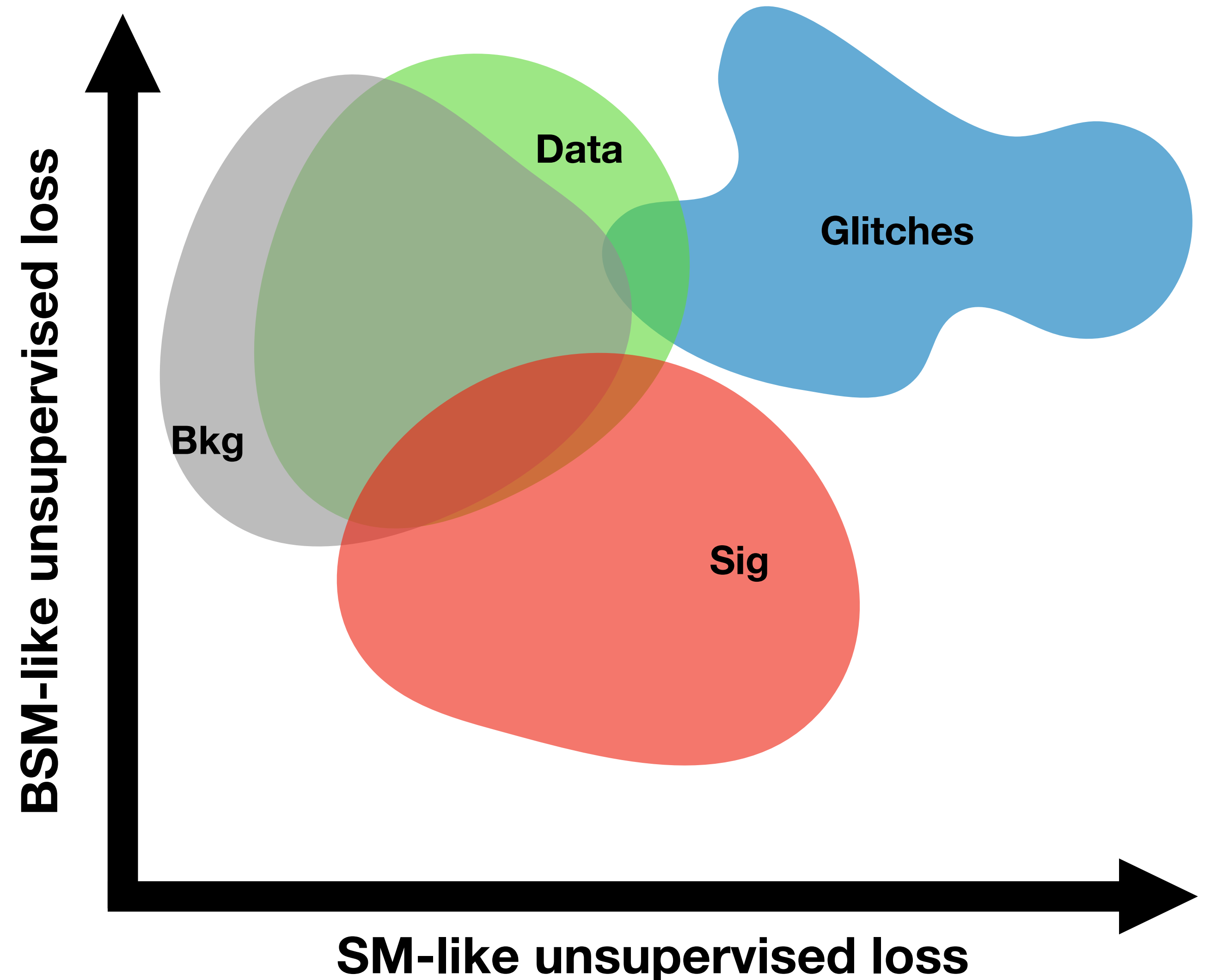


- Imagine you train an unsupervised model on SM backgrounds
 - ▶ The “high loss” region is where we expect signal — potentially very sensitive!
 - ▶ But in reality, tails are populated by all kinds of *uninteresting* anomalies
- **Semi-supervised** : We “know” what new physics could/would look like
 - ▶ Add a new “signal-like” axis or axes trained on benchmark BSM signals



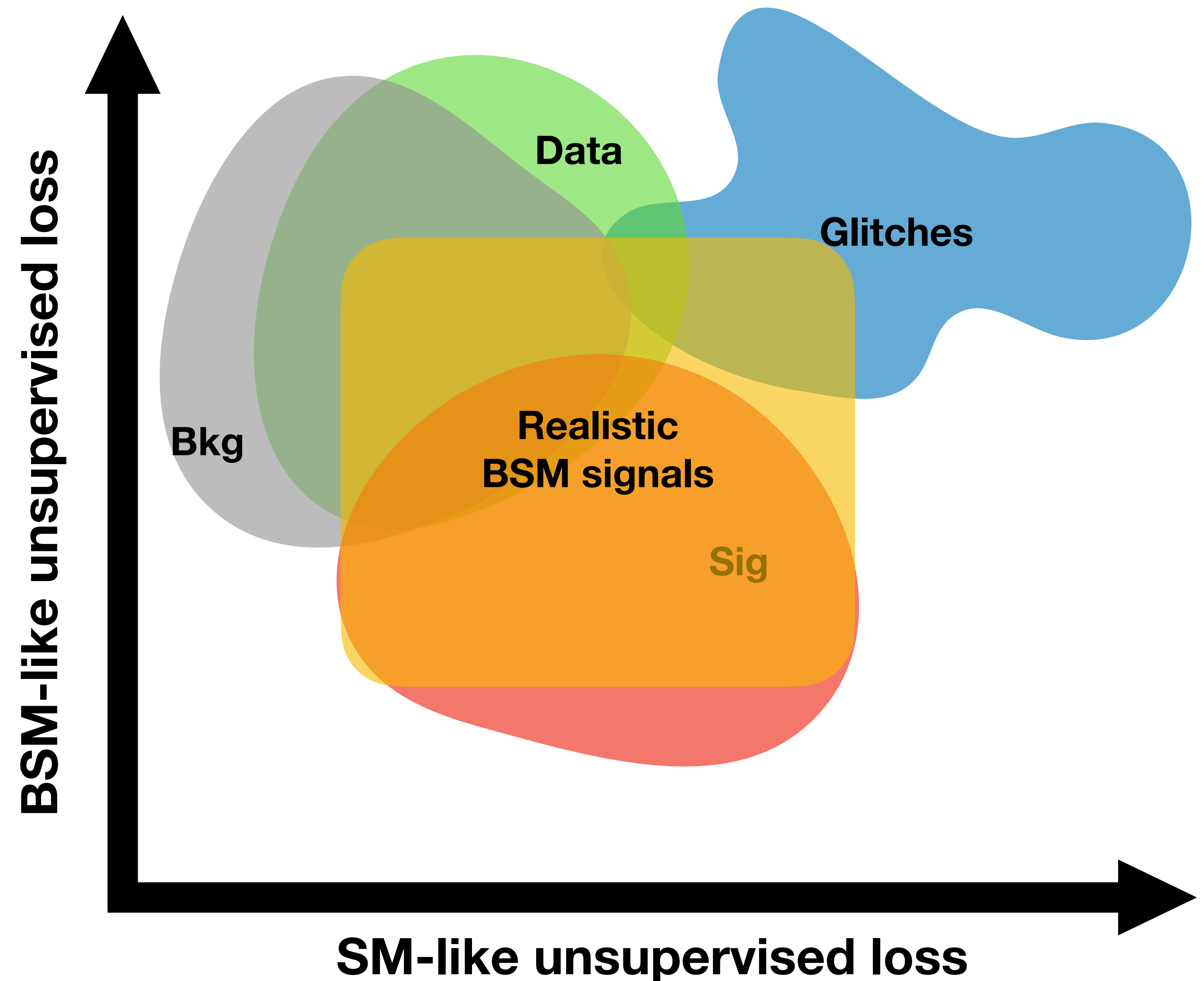
Semi-supervised searches with QUAK

- Imagine you train an unsupervised model on SM backgrounds
 - ▶ The “high loss” region is where we expect signal — potentially very sensitive!
 - ▶ But in reality, tails are populated by all kinds of *uninteresting* anomalies
- **Semi-supervised** : We “know” what new physics could/would look like
 - ▶ Add a new “signal-like” axis or axes trained on benchmark BSM signals
 - ▶ Provides a “**physics prior**” — include wide enough set of priors to enable “interpolation”



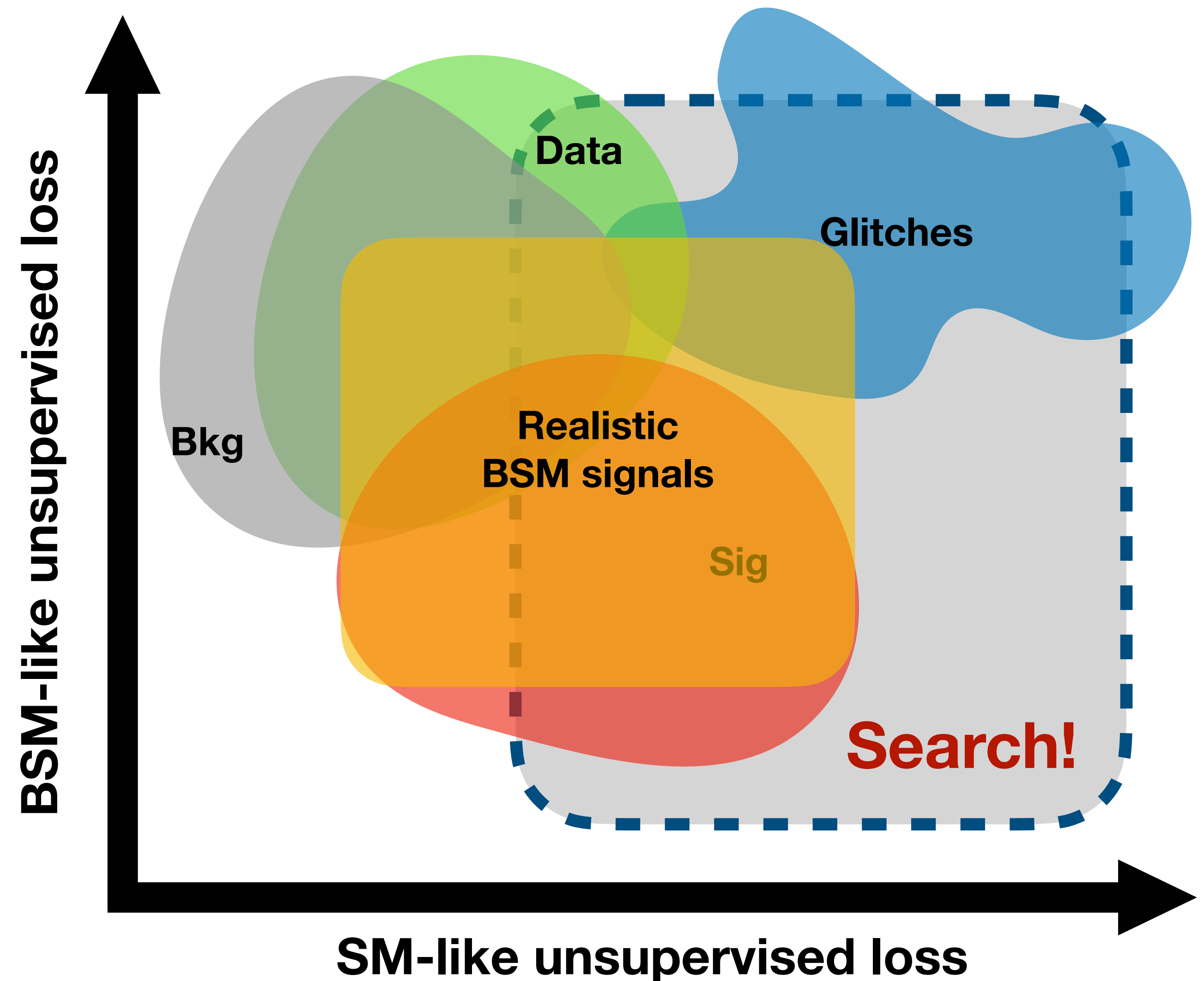
Semi-supervised searches with QUAK

- Imagine you train an unsupervised model on SM backgrounds
 - ▶ The “high loss” region is where we expect signal — potentially very sensitive!
 - ▶ But in reality, tails are populated by all kinds of *uninteresting* anomalies
- **Semi-supervised** : We “know” what new physics could/would look like
 - ▶ Add a new “signal-like” axis or axes trained on benchmark BSM signals
 - ▶ Provides a “**physics prior**” — include wide enough set of priors to enable “interpolation”



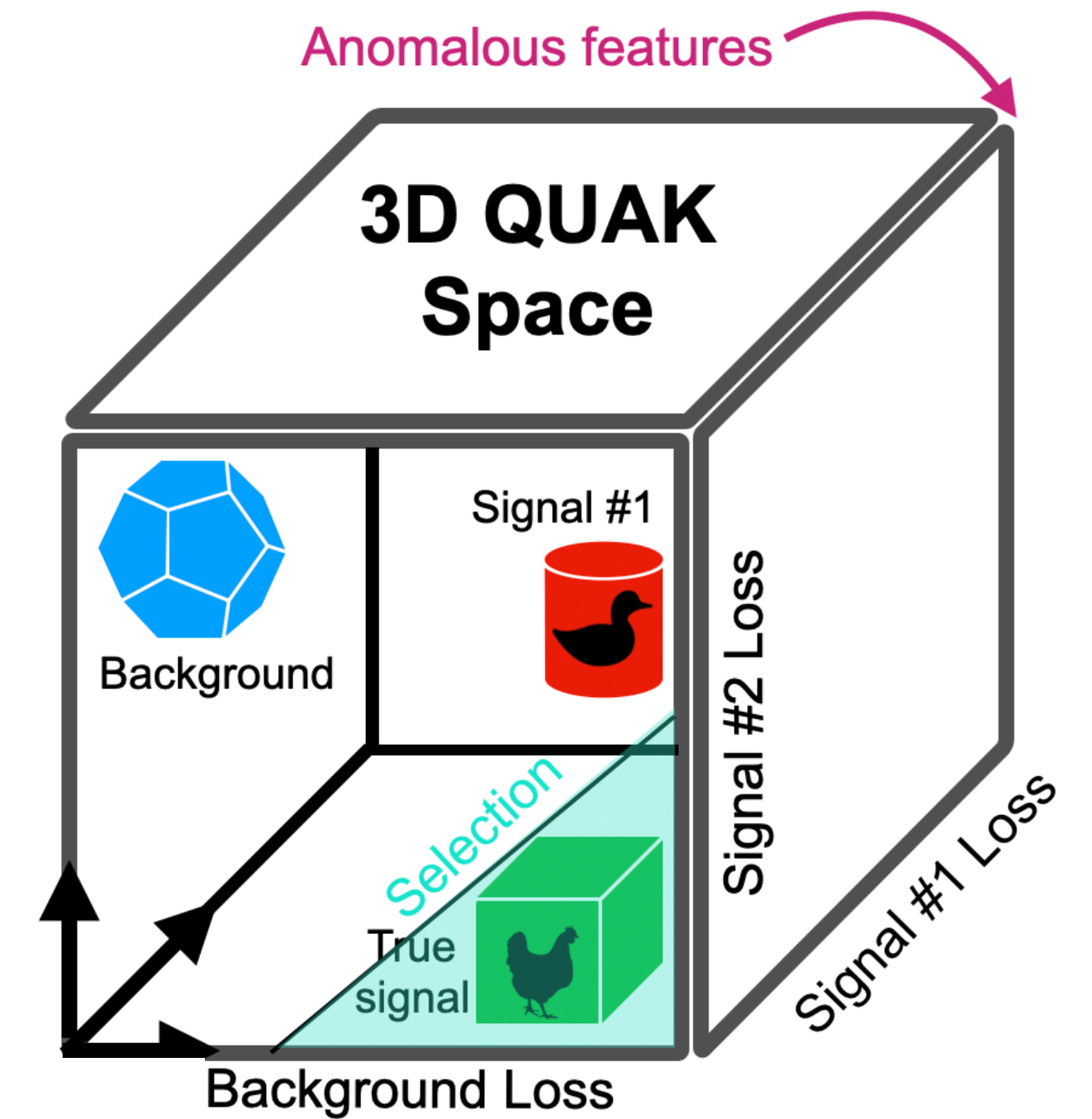
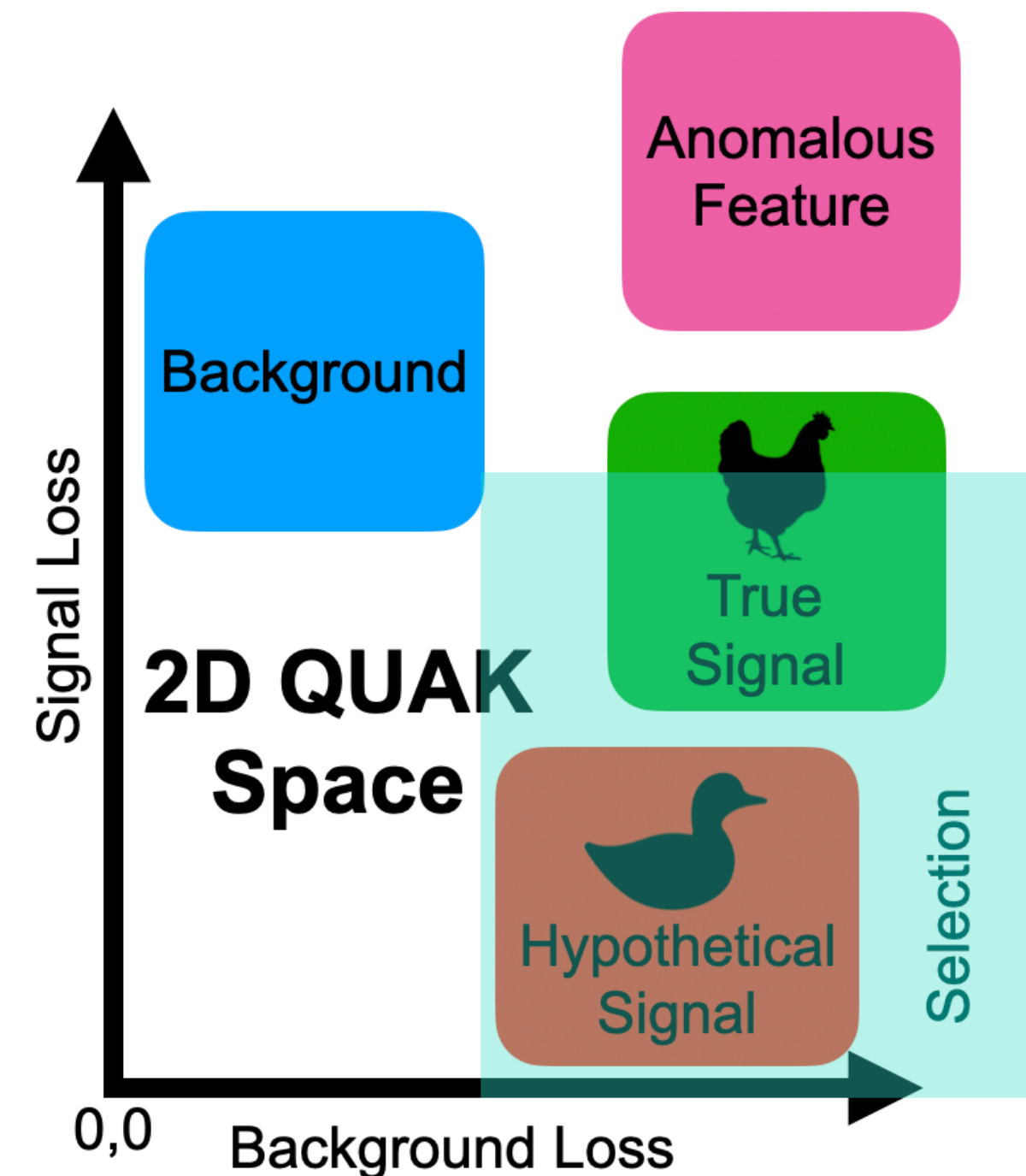
Semi-supervised searches with QUAK

- Imagine you train an unsupervised model on SM backgrounds
 - ▶ The “high loss” region is where we expect signal — potentially very sensitive!
 - ▶ But in reality, tails are populated by all kinds of *uninteresting* anomalies
- **Semi-supervised** : We “know” what new physics could/would look like
 - ▶ Add a new “signal-like” axis or axes trained on benchmark BSM signals
 - ▶ Provides a “**physics prior**” — include wide enough set of priors to enable “interpolation”



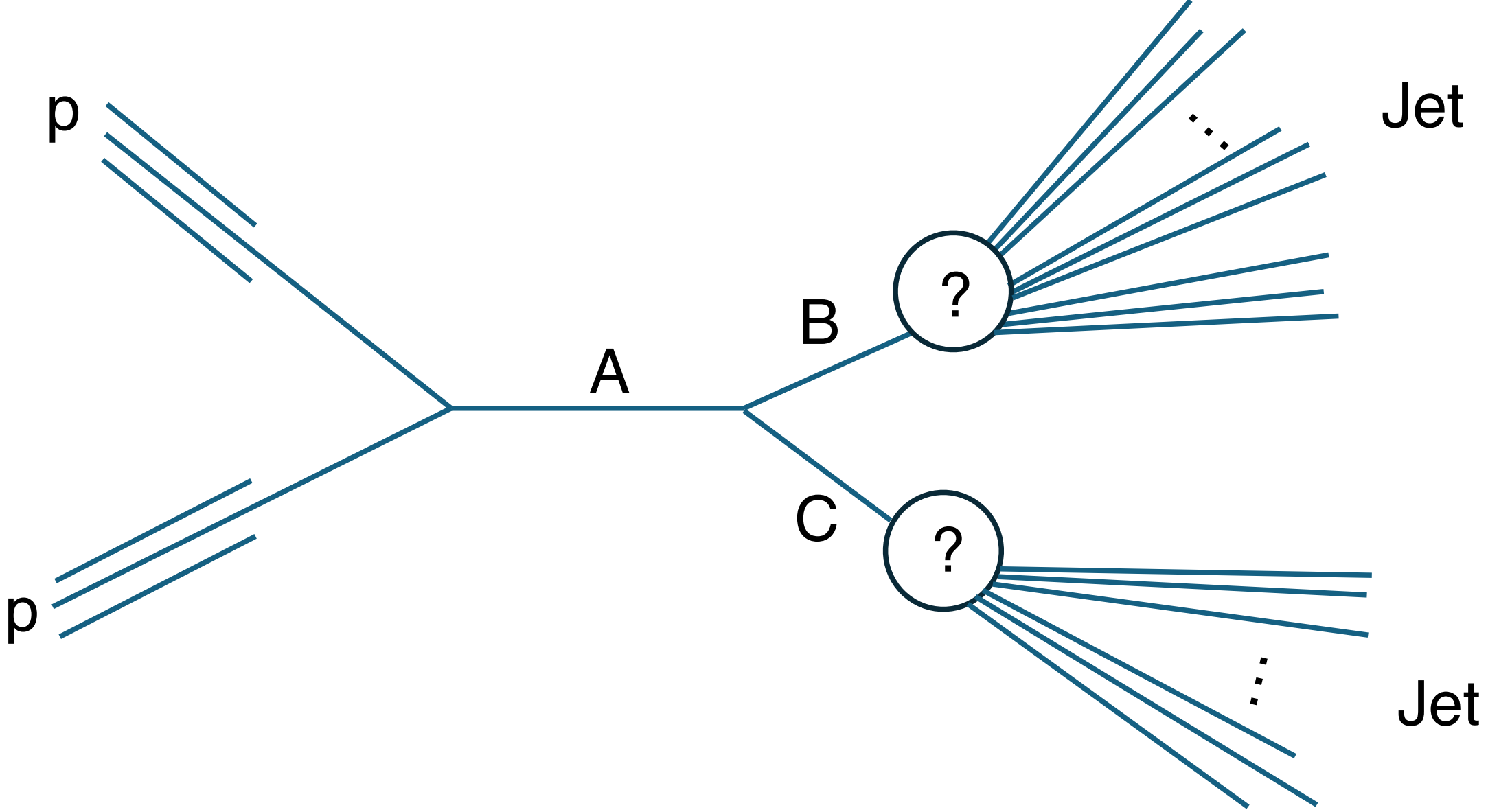
Semi-supervised searches with QUAKE

- Imagine you train an unsupervised model on SM backgrounds
 - ▶ The “high loss” region is where we expect signal — potentially very sensitive!
 - ▶ But in reality, tails are populated by all kinds of *uninteresting* anomalies
- **Semi-supervised** : We “know” what new physics could/would look like
 - ▶ Add a new “signal-like” axis or axes trained on benchmark BSM signals
 - ▶ Provides a “**physics prior**” — include wide enough set of priors to enable “interpolation”
- “**Quasi Anomalous Knowledge**” technique — QUAKE!



JHEP 06 (2021) 030

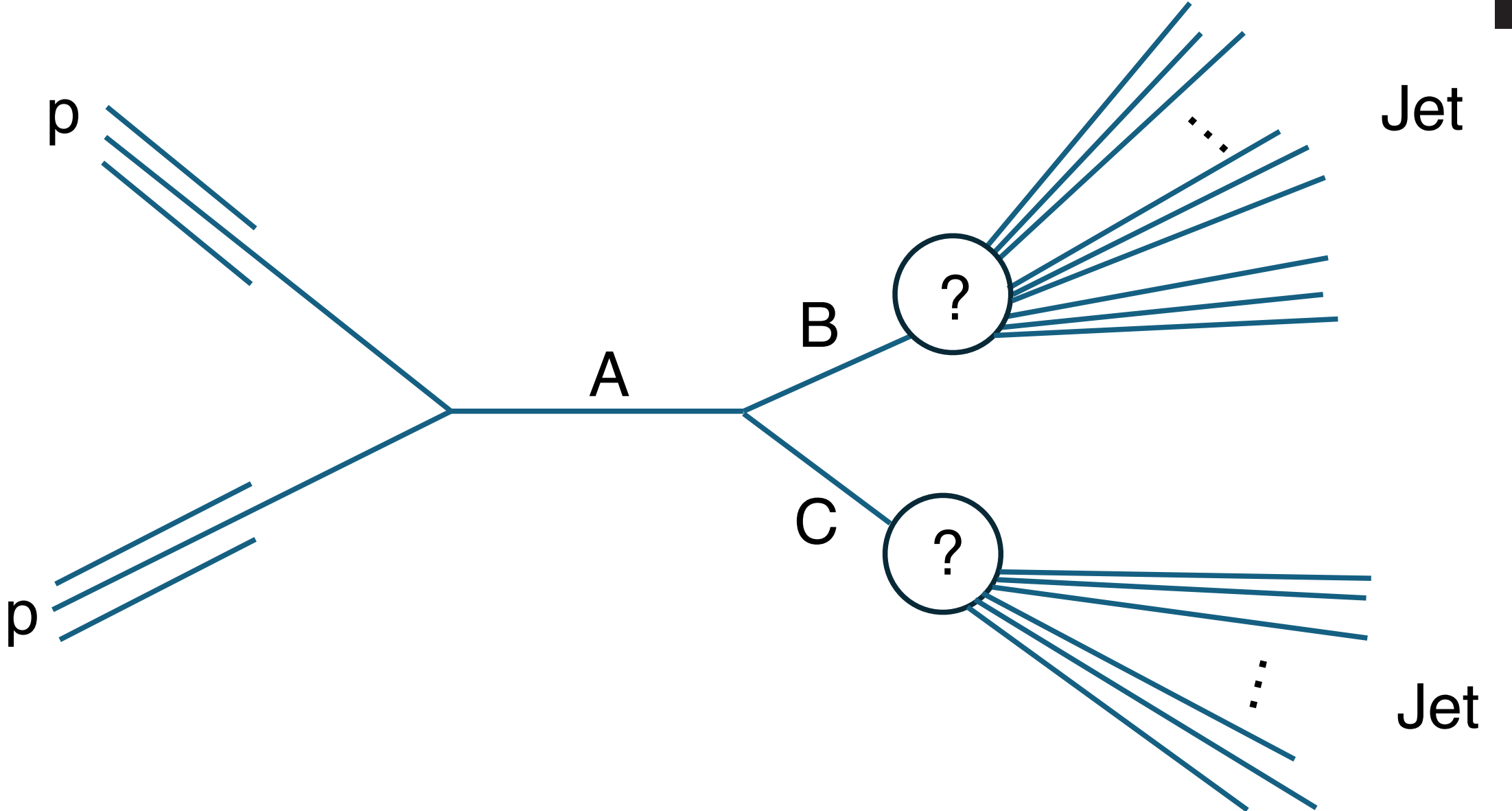
Choosing physics priors



Choosing physics priors

- Relevant physical parameters:

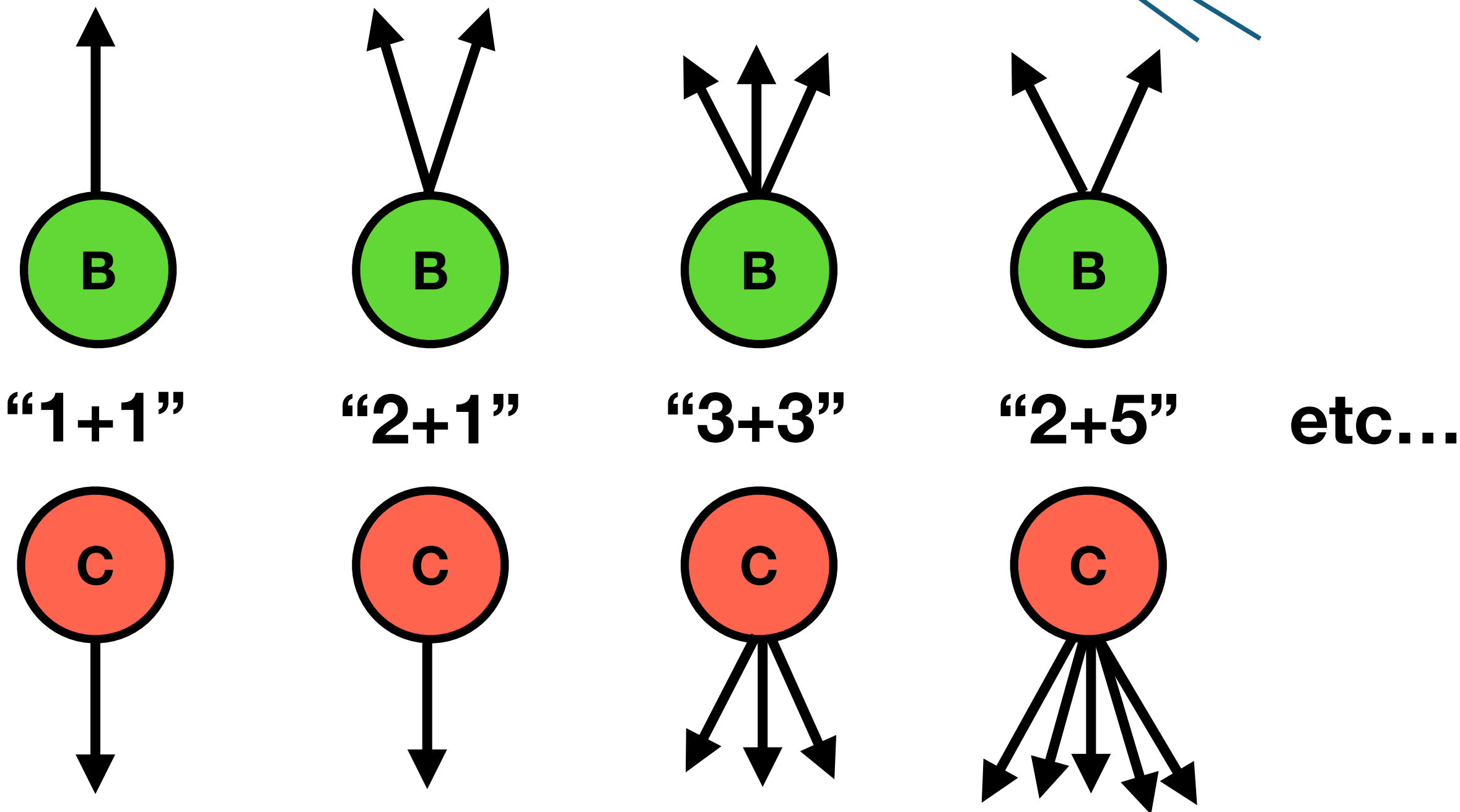
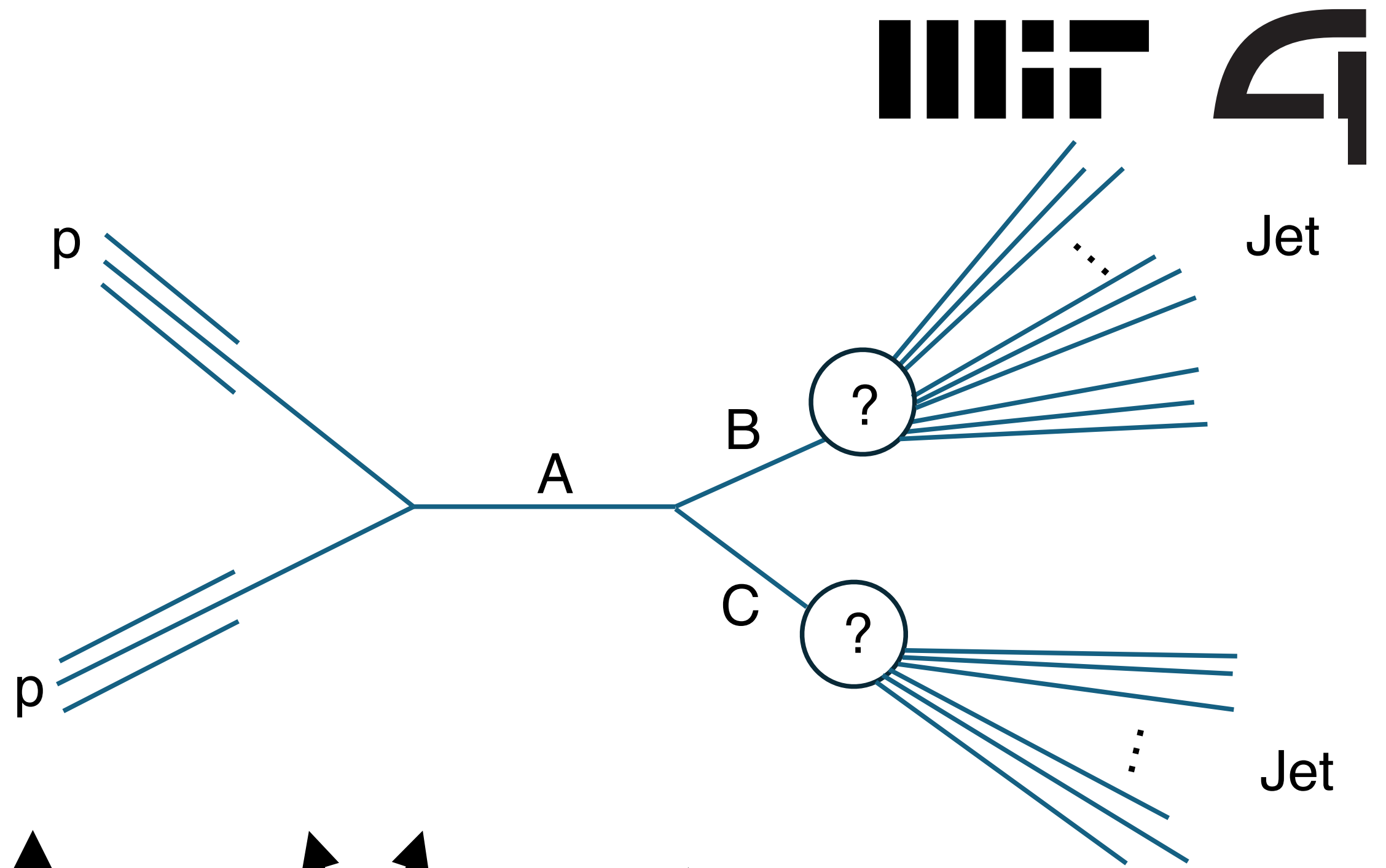
1. m_A — should be as agnostic as possible about this!
2. m_B, m_C — also agnostic, but “aware” that they are resonant & much lighter than A



Choosing physics priors

- **Relevant physical parameters:**

1. m_A — should be as agnostic as possible about this!
2. m_B, m_C — also agnostic, but “aware” that they are resonant & much lighter than A
3. B & C decay chains — “substructure topology” of an event; should be sensitive to many possibilities!

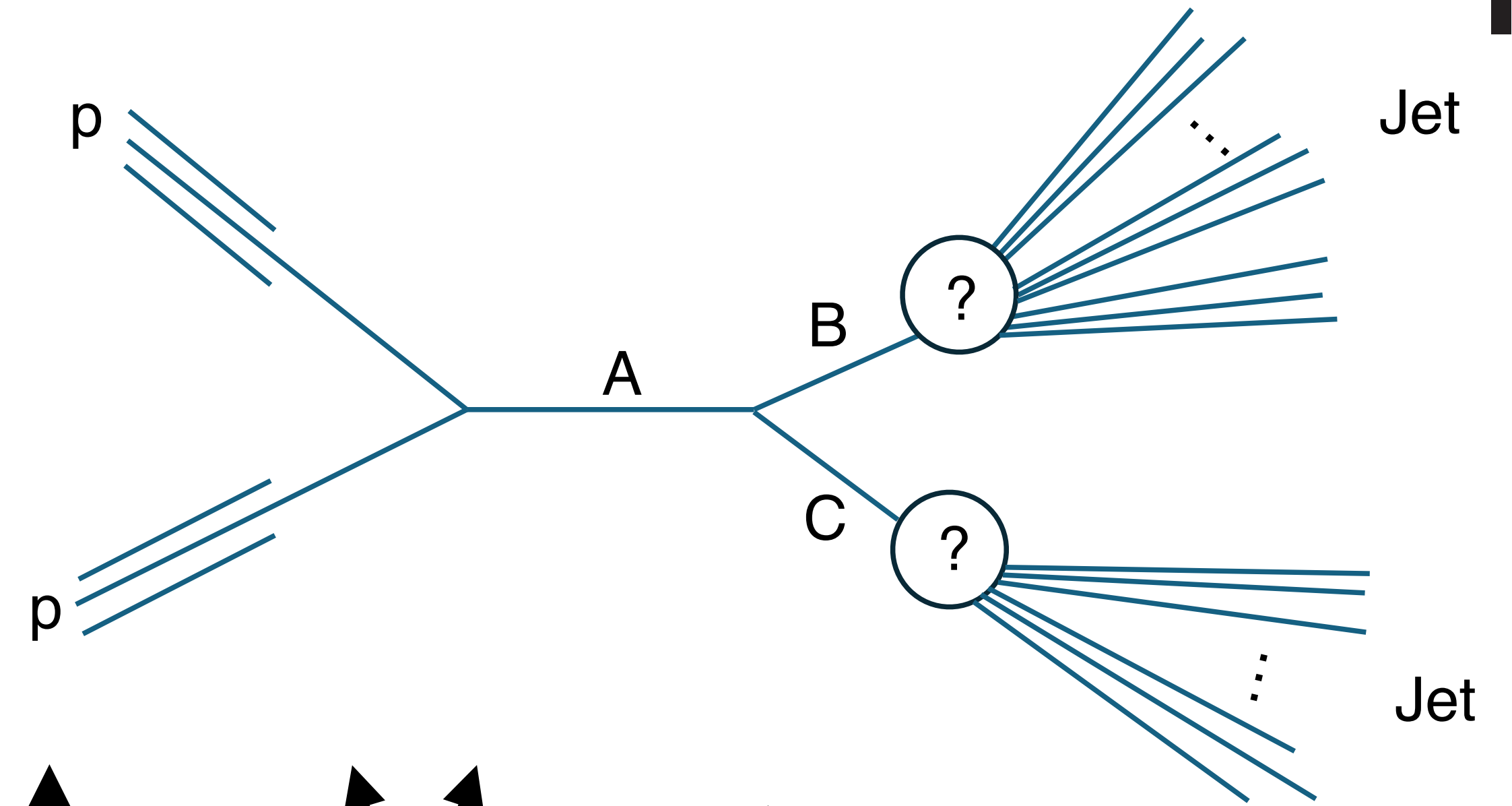


Choosing physics priors

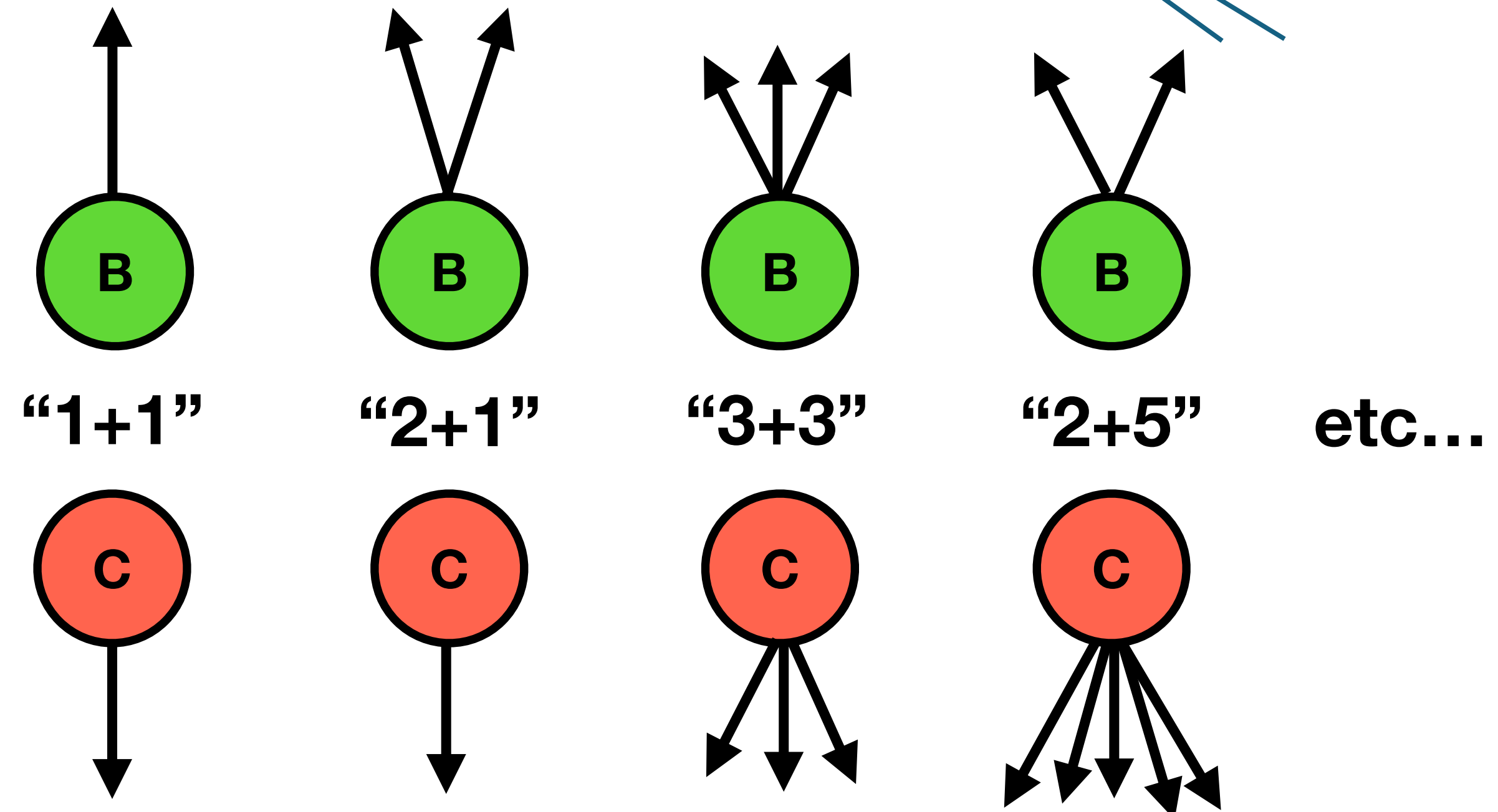


- **Relevant physical parameters:**

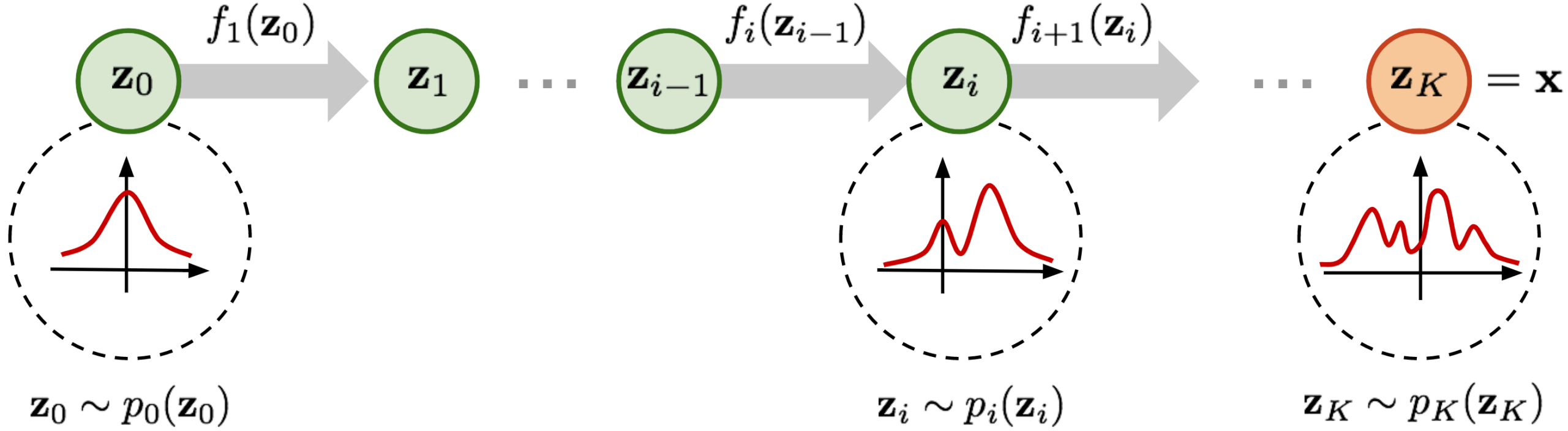
1. m_A — should be as agnostic as possible about this!
2. m_B, m_C — also agnostic, but “aware” that they are resonant & much lighter than A
3. B & C decay chains — “substructure topology” of an event; should be sensitive to many possibilities!



Sample	Prongs	B and C mass points (GeV)
$G \rightarrow gg$	1+1	-
$Q^* \rightarrow qW'$	1+2	$M_{W'} = 25, 80, 170, 400$
$X \rightarrow YY', Y/Y' \rightarrow q\bar{q}$	2+2	$M_{Y/Y'} = 25, 80, 170, 400$
$W_{KK} \rightarrow WR, R \rightarrow WW$	2+4	$M_R = 170, 400$
$W' \rightarrow B't, B' \rightarrow bZ$	3+3	$M_{B'} = 25, 80, 170, 400$
$Z' \rightarrow T'T', T' \rightarrow tZ$	5+5	$M_{T'} = 400$
$Y \rightarrow HH, H \rightarrow tt$	6+6	$M_H = 400$



Training the axes

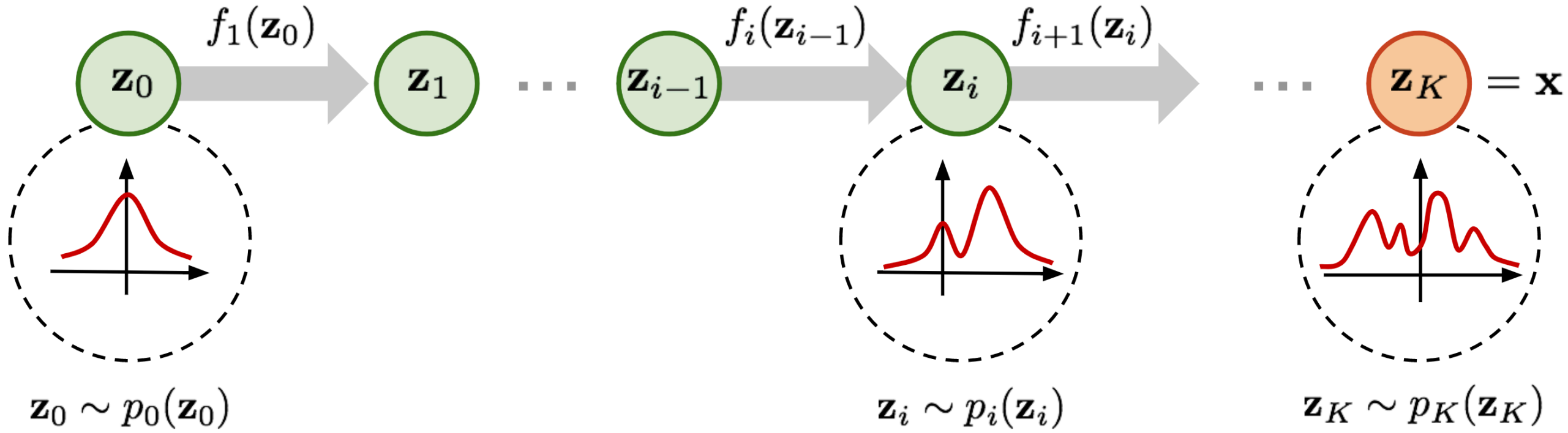


Inputs (each jet)

$$\rho = \frac{m_{SD,j}}{p_{T,j}}, \quad \tau_{21}, \quad \tau_{32}, \quad \tau_{43}, \quad \tau_S = \frac{\sqrt{\tau_{21}}}{\tau_1}, \quad P_b, \quad n_{PF}$$

Training the axes

- QUAK space axes implemented with **normalizing flows**
 - ▶ Density estimation models (unsupervised)

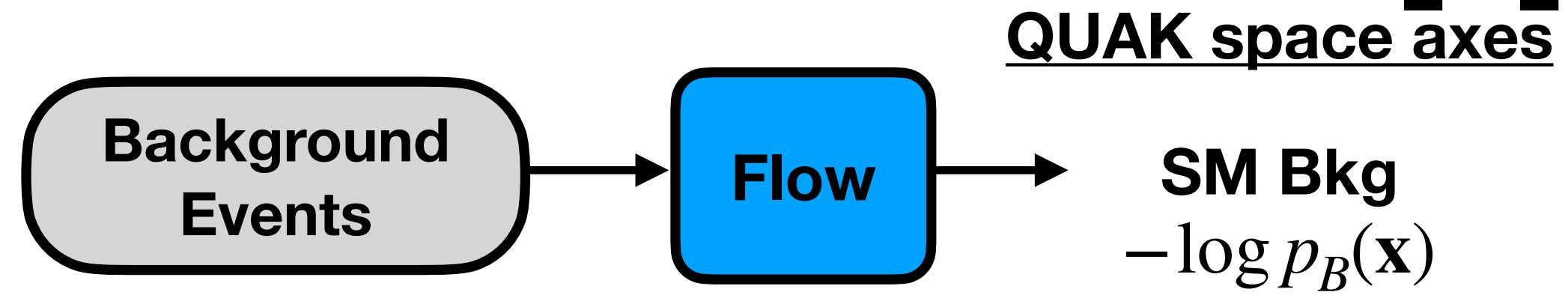


Inputs (each jet)

$$\rho = \frac{m_{SD,j}}{p_{T,j}}, \quad \tau_{21}, \quad \tau_{32}, \quad \tau_{43}, \quad \tau_S = \frac{\sqrt{\tau_{21}}}{\tau_1}, \quad P_b, \quad n_{PF}$$

Training the axes

- QUAK space axes implemented with **normalizing flows**
 - ▶ Density estimation models (unsupervised)
 - ▶ Evaluate $-\log p(\mathbf{x})$ for each event — **signal/bkg-like losses**



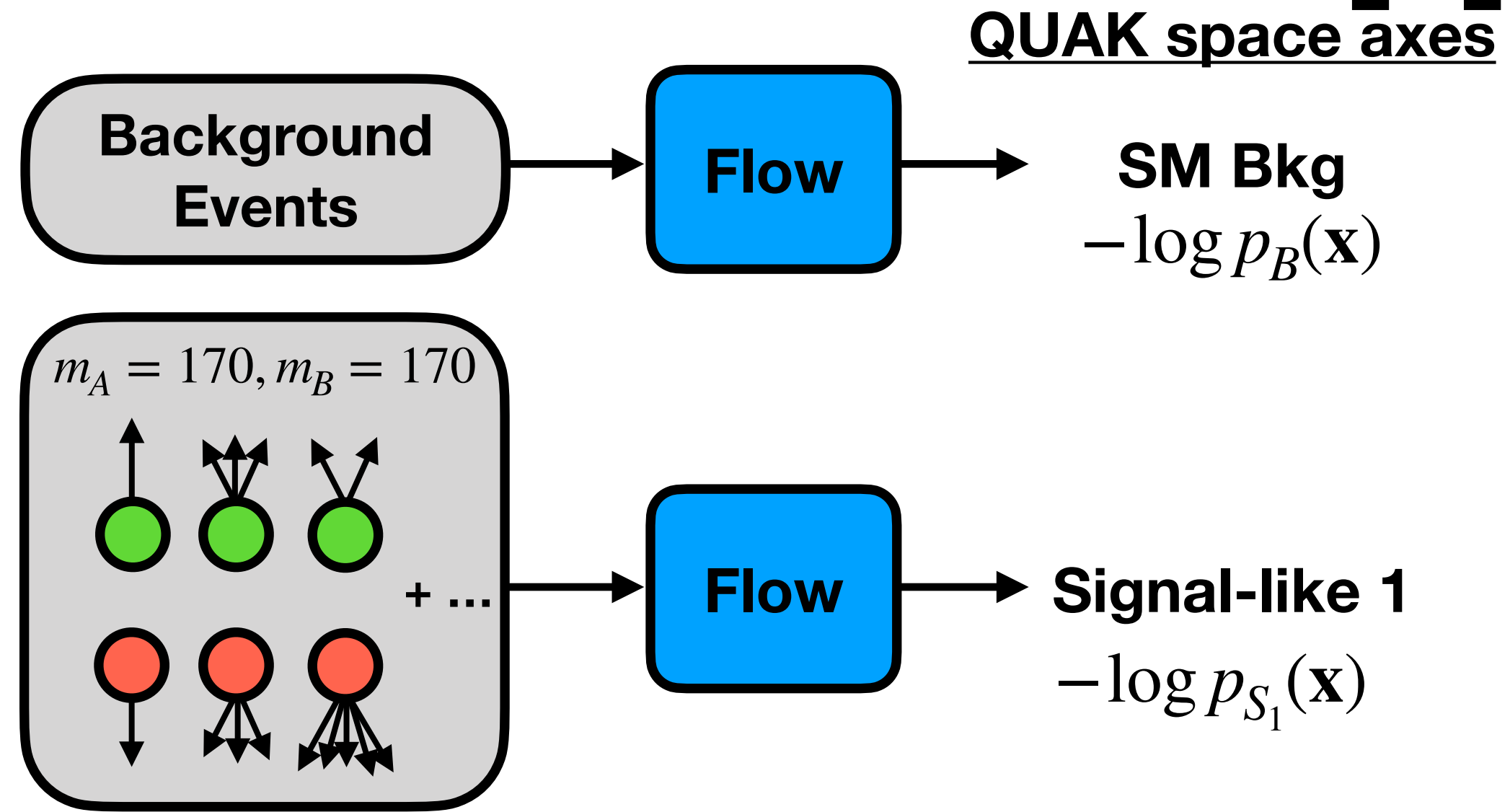
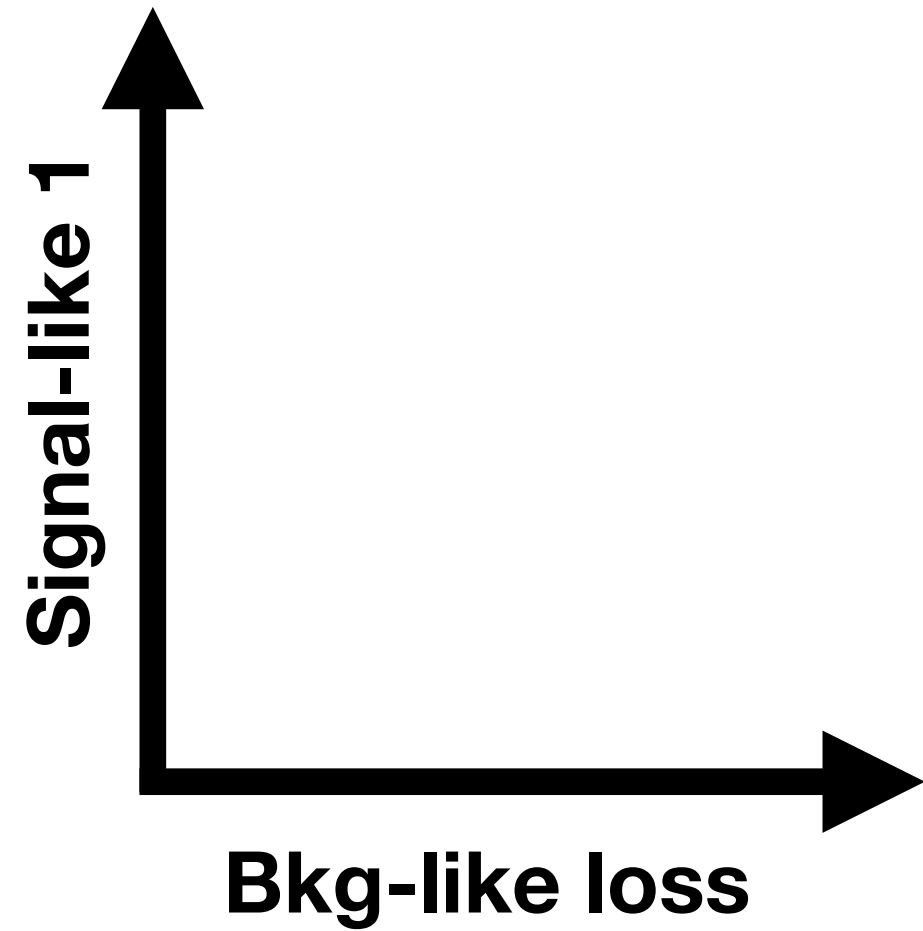
→
Bkg-like loss

Inputs (each jet)

$$\rho = \frac{m_{SD,j}}{p_{T,j}}, \quad \tau_{21}, \quad \tau_{32}, \quad \tau_{43}, \quad \tau_S = \frac{\sqrt{\tau_{21}}}{\tau_1}, \quad P_b, \quad n_{PF}$$

Training the axes

- QUAK space axes implemented with **normalizing flows**
 - ▶ Density estimation models (unsupervised)
 - ▶ Evaluate $-\log p(\mathbf{x})$ for each event — **signal/bkg-like losses**

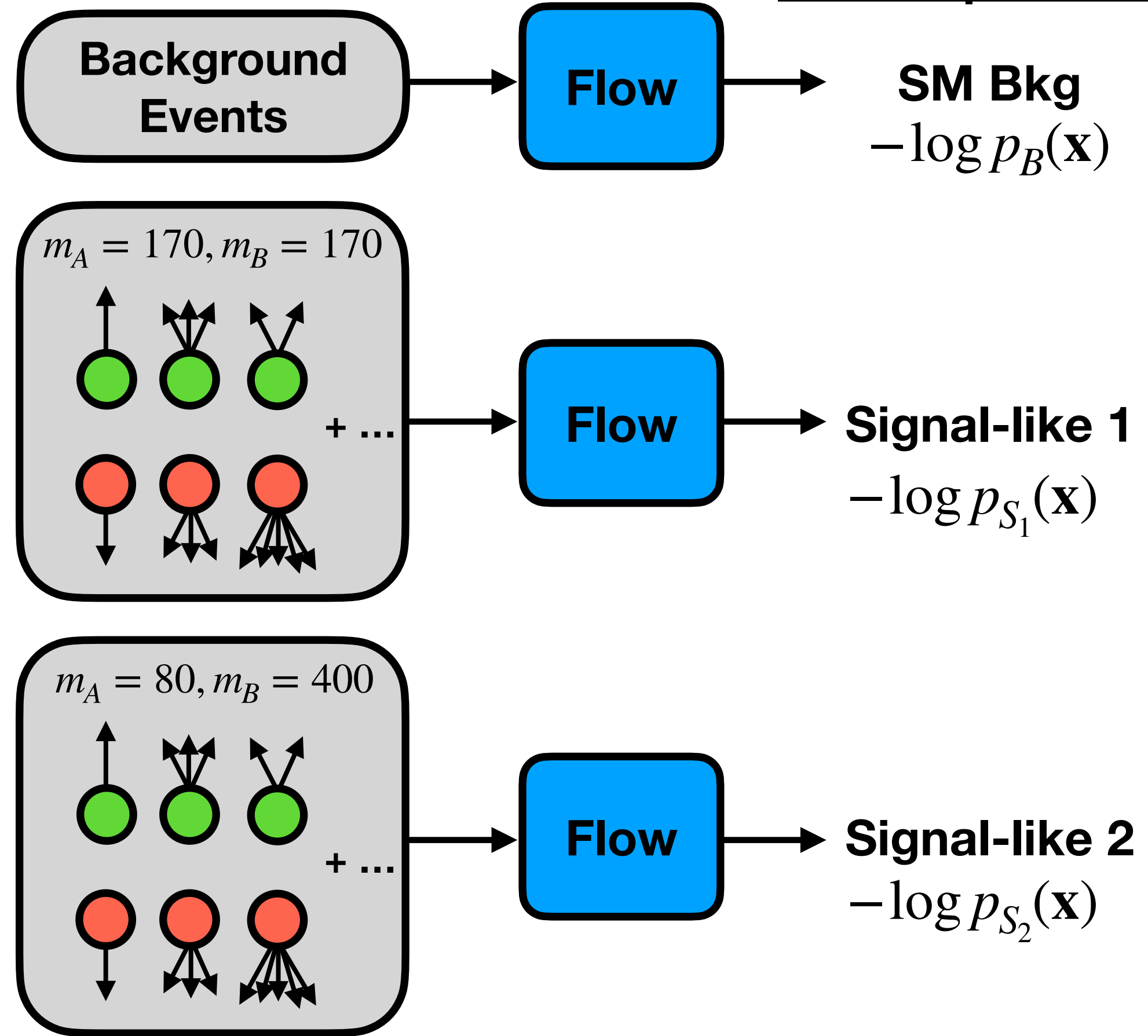
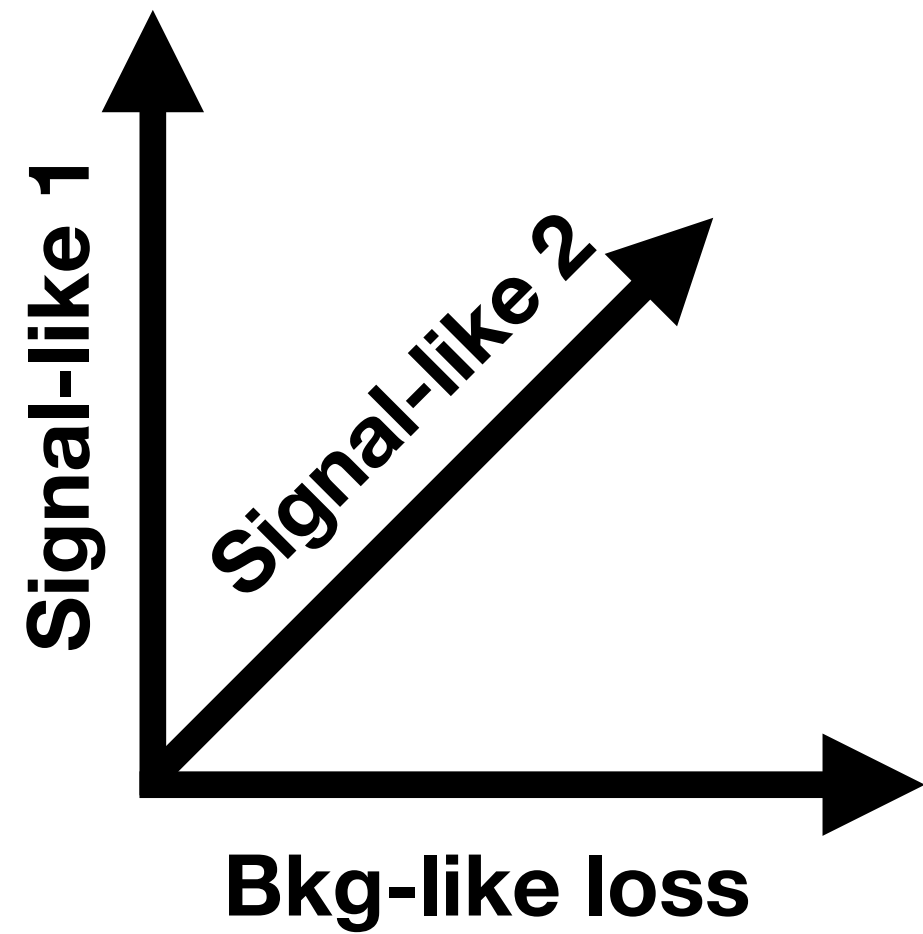


Inputs (each jet)

$$\rho = \frac{m_{SD,j}}{p_{T,j}}, \quad \tau_{21}, \quad \tau_{32}, \quad \tau_{43}, \quad \tau_S = \frac{\sqrt{\tau_{21}}}{\tau_1}, \quad P_b, \quad n_{PF}$$

Training the axes

- QUAK space axes implemented with **normalizing flows**
 - ▶ Density estimation models (unsupervised)
 - ▶ Evaluate $-\log p(\mathbf{x})$ for each event — **signal/bkg-like losses**

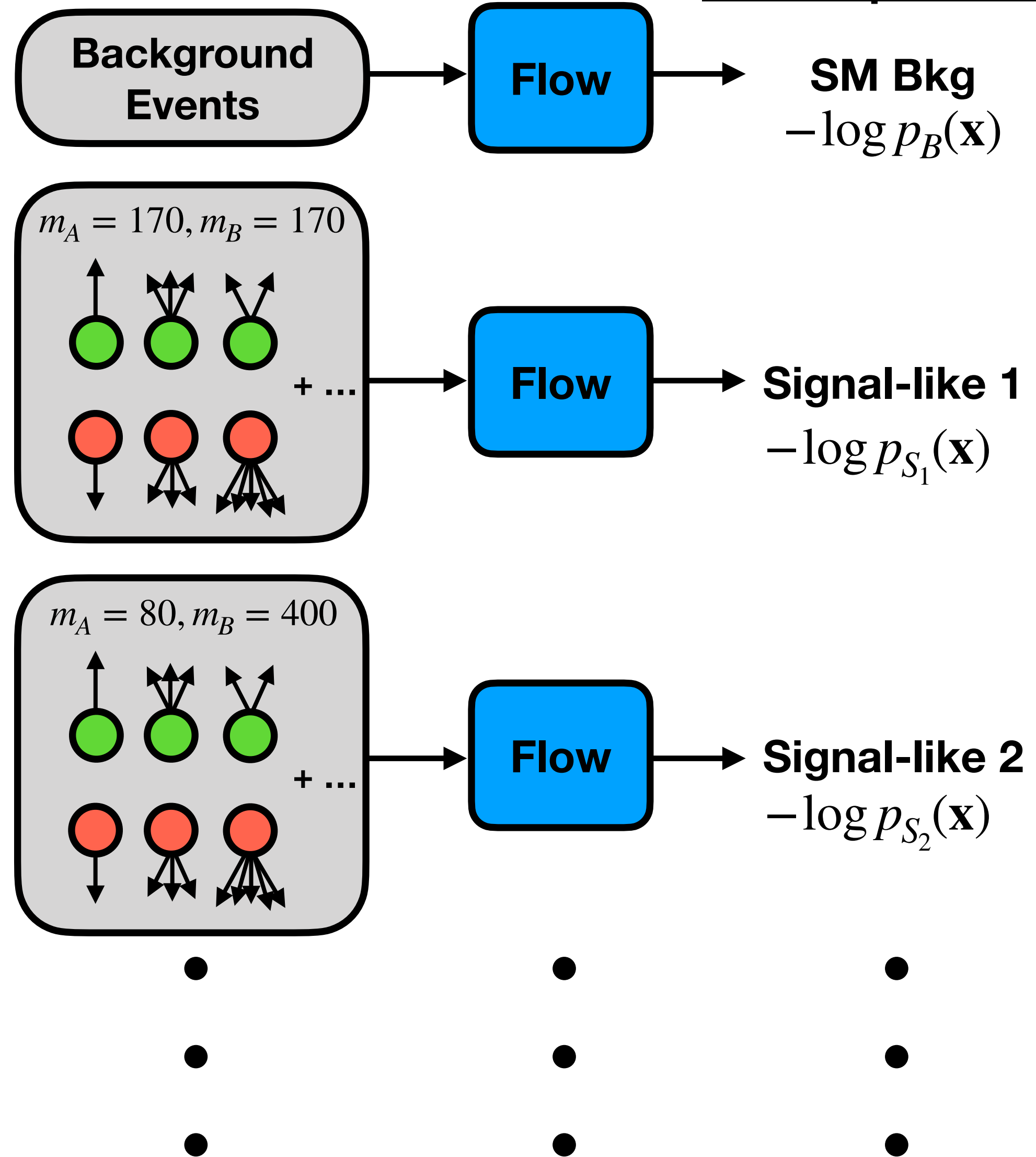
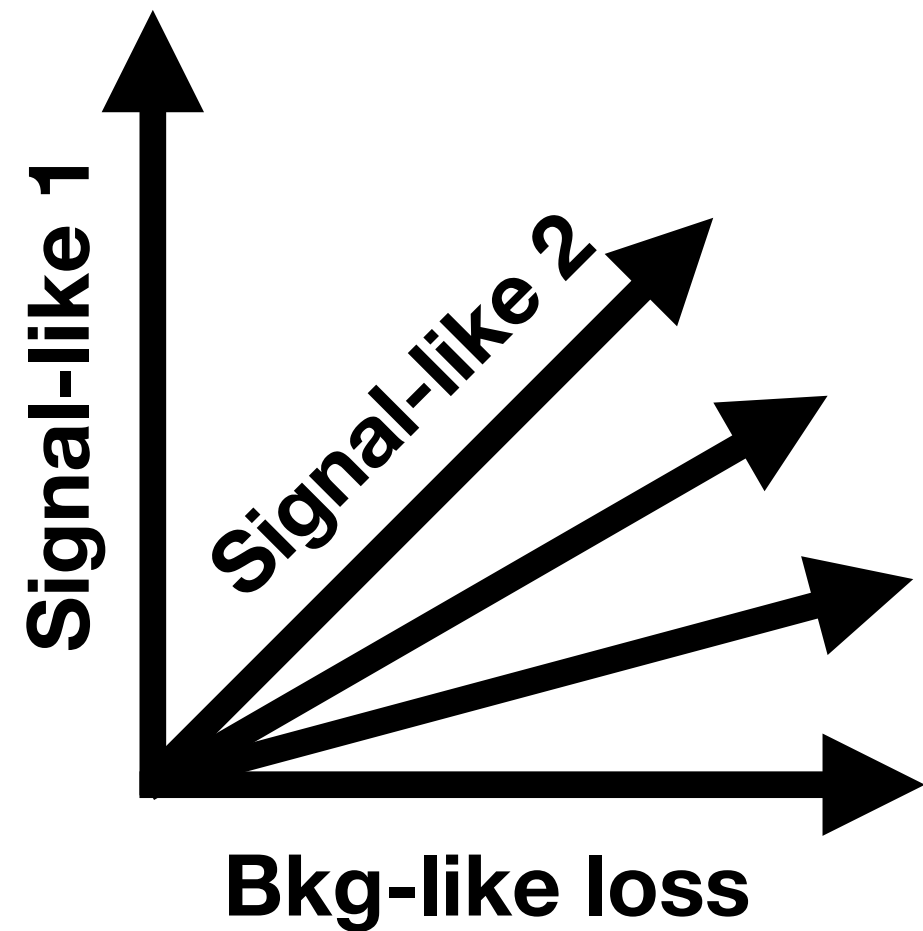


Inputs (each jet)

$$\rho = \frac{m_{SD,j}}{p_{T,j}}, \quad \tau_{21}, \quad \tau_{32}, \quad \tau_{43}, \quad \tau_S = \frac{\sqrt{\tau_{21}}}{\tau_1}, \quad P_b, \quad n_{PF}$$

Training the axes

- QUAK space axes implemented with **normalizing flows**
 - ▶ Density estimation models (unsupervised)
 - ▶ Evaluate $-\log p(\mathbf{x})$ for each event — **signal/bkg-like losses**

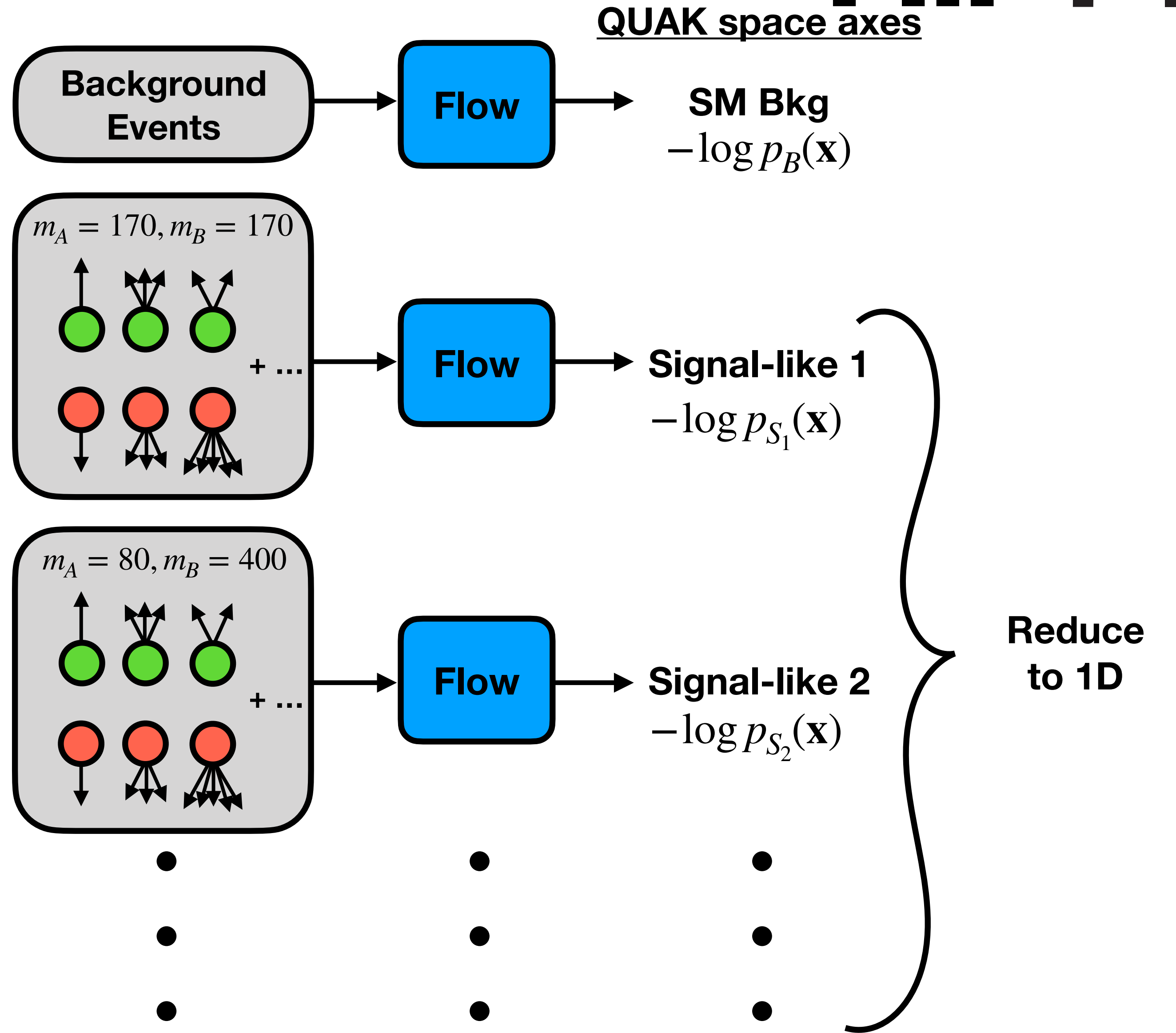
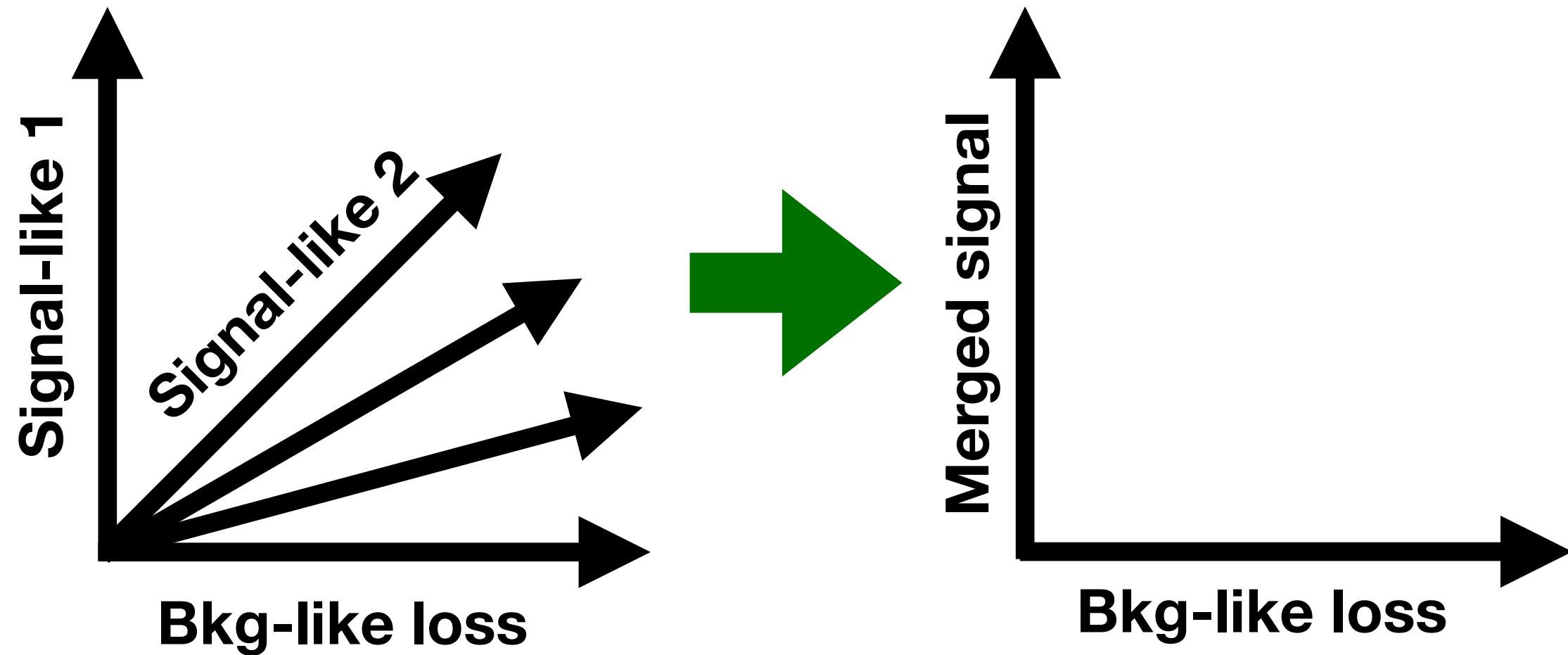


Inputs (each jet)

$$\rho = \frac{m_{SD,j}}{p_{T,j}}, \quad \tau_{21}, \quad \tau_{32}, \quad \tau_{43}, \quad \tau_S = \frac{\sqrt{\tau_{21}}}{\tau_1}, \quad P_b, \quad n_{PF}$$

Training the axes

- QUAK space axes implemented with **normalizing flows**
 - ▶ Density estimation models (unsupervised)
 - ▶ Evaluate $-\log p(\mathbf{x})$ for each event — **signal/bkg-like losses**

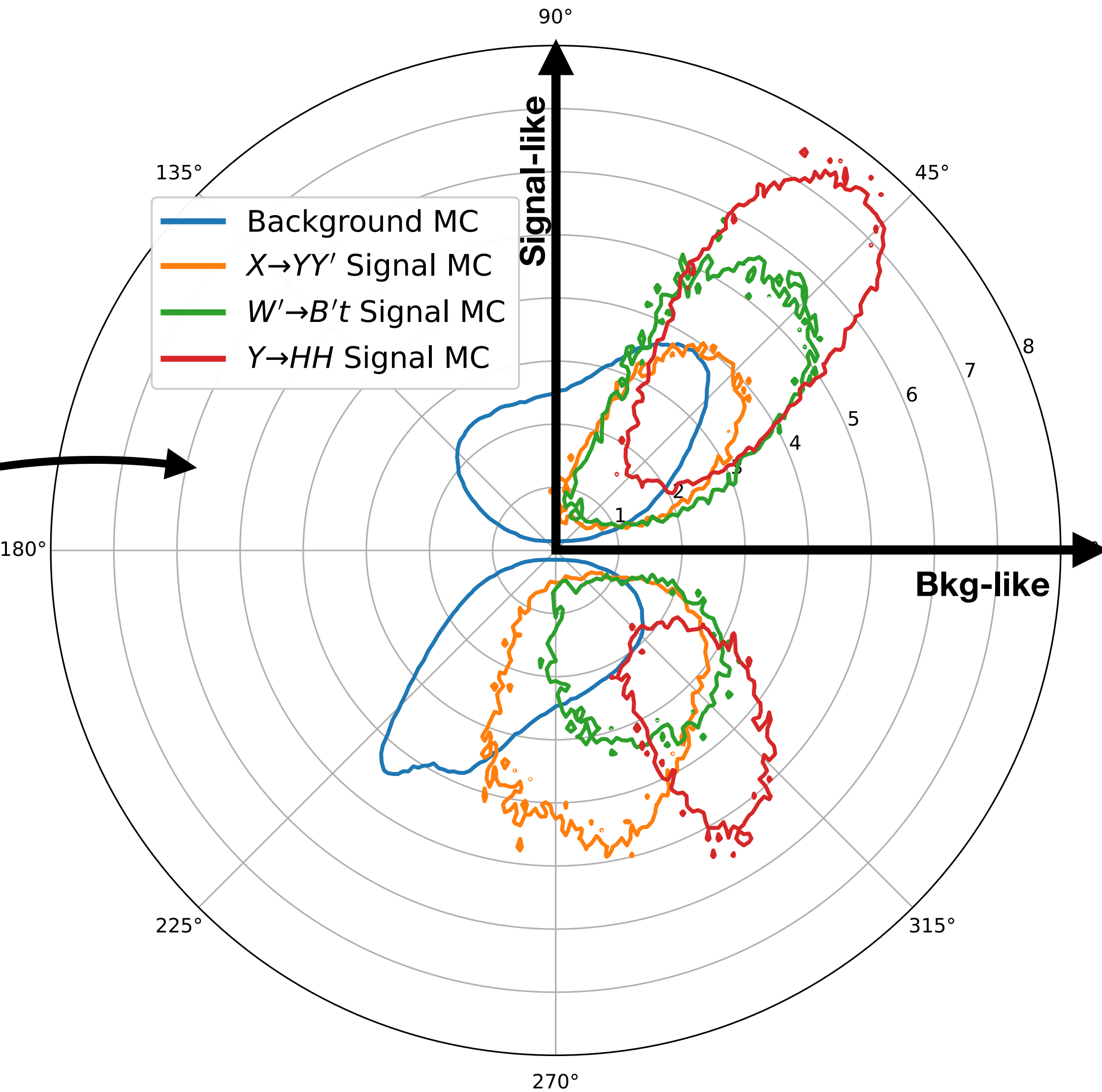
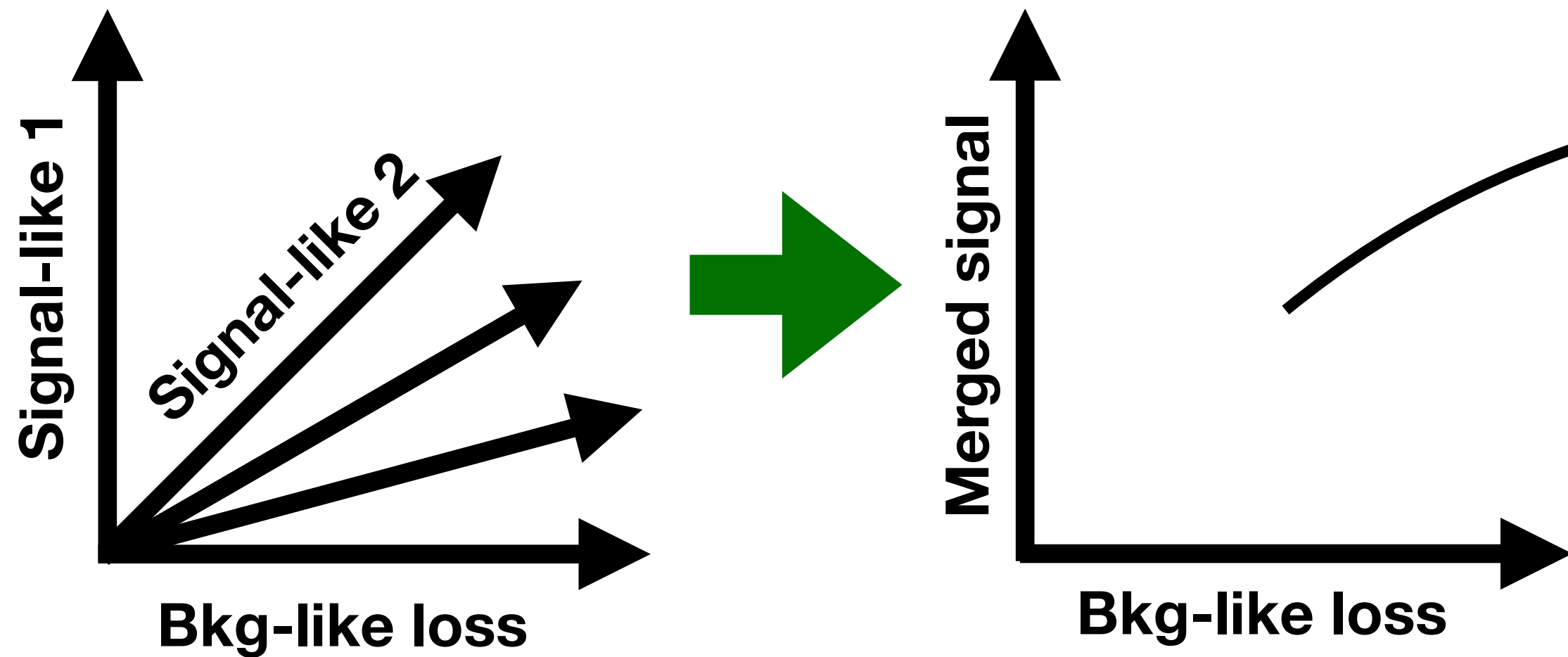


Inputs (each jet)

$$\rho = \frac{m_{SD,j}}{p_{T,j}}, \quad \tau_{21}, \quad \tau_{32}, \quad \tau_{43}, \quad \tau_S = \frac{\sqrt{\tau_{21}}}{\tau_1}, \quad P_b, \quad n_{PF}$$

Training the axes

- QUAK space axes implemented with **normalizing flows**
 - ▶ Density estimation models (unsupervised)
 - ▶ Evaluate $-\log p(\mathbf{x})$ for each event — **signal/bkg-like losses**

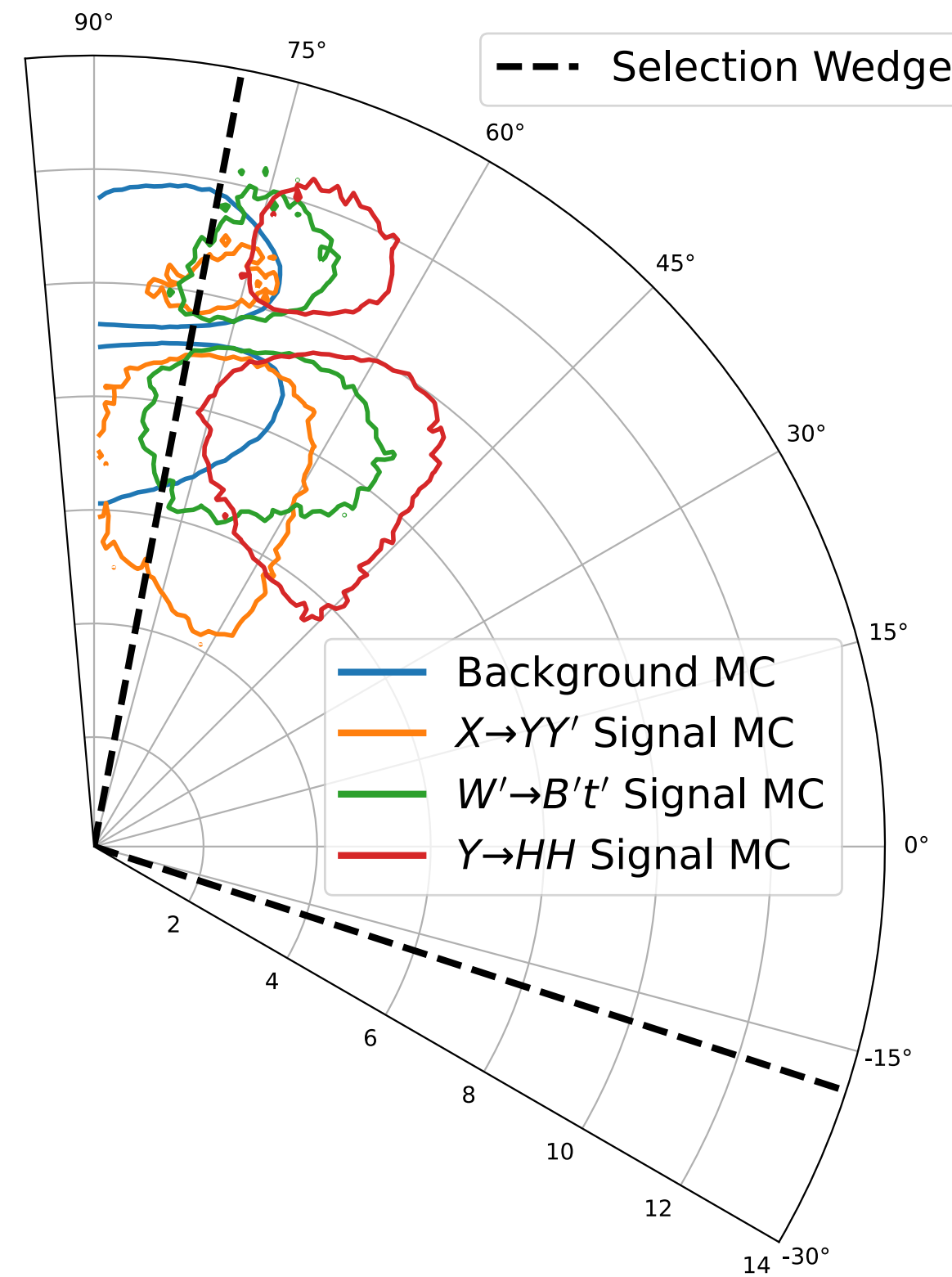


Inputs (each jet)

$$\rho = \frac{m_{SD,j}}{p_{T,j}}, \quad \tau_{21}, \quad \tau_{32}, \quad \tau_{43}, \quad \tau_S = \frac{\sqrt{\tau_{21}}}{\tau_1}, \quad P_b, \quad n_{PF}$$

Bump hunting

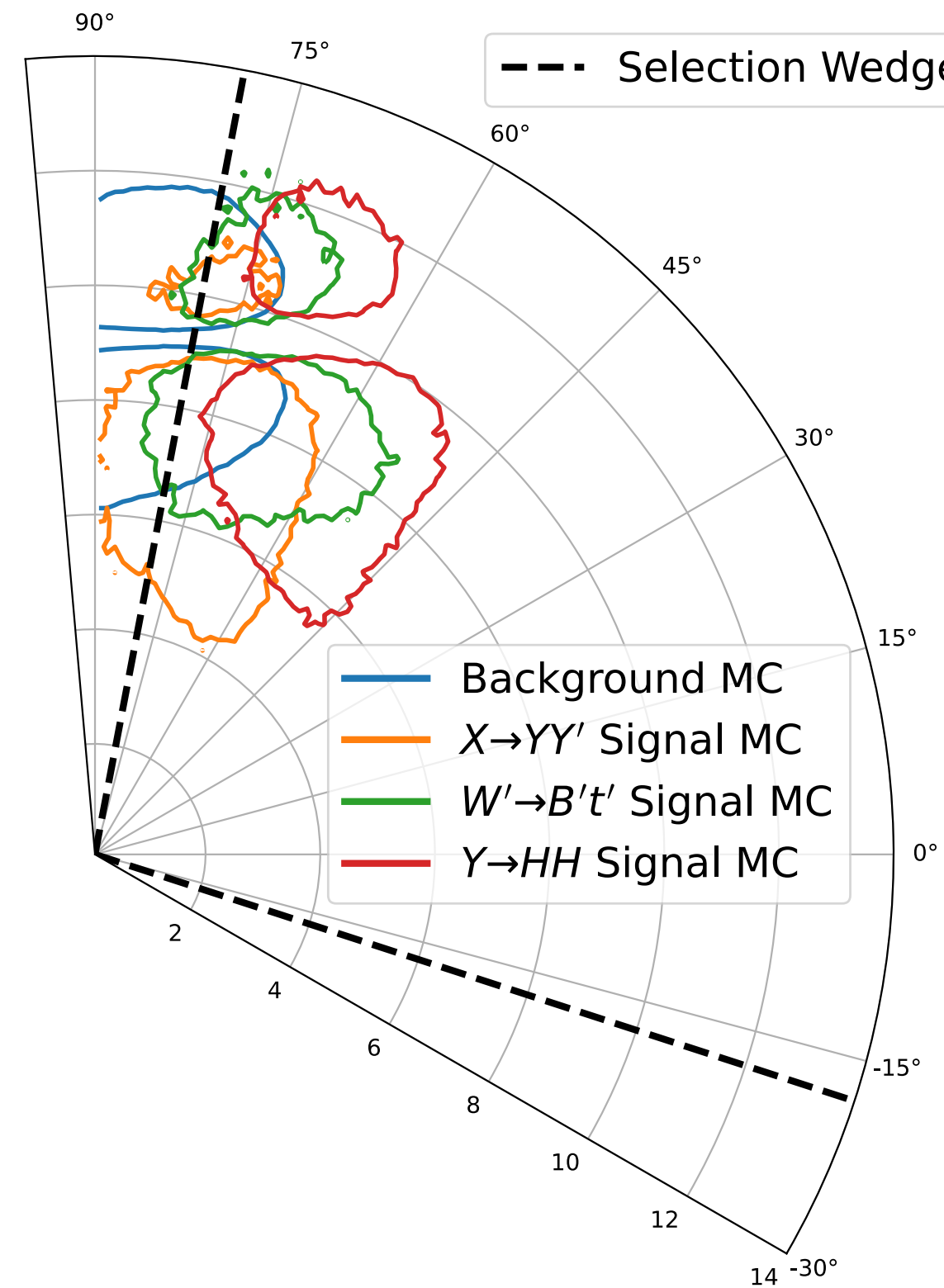
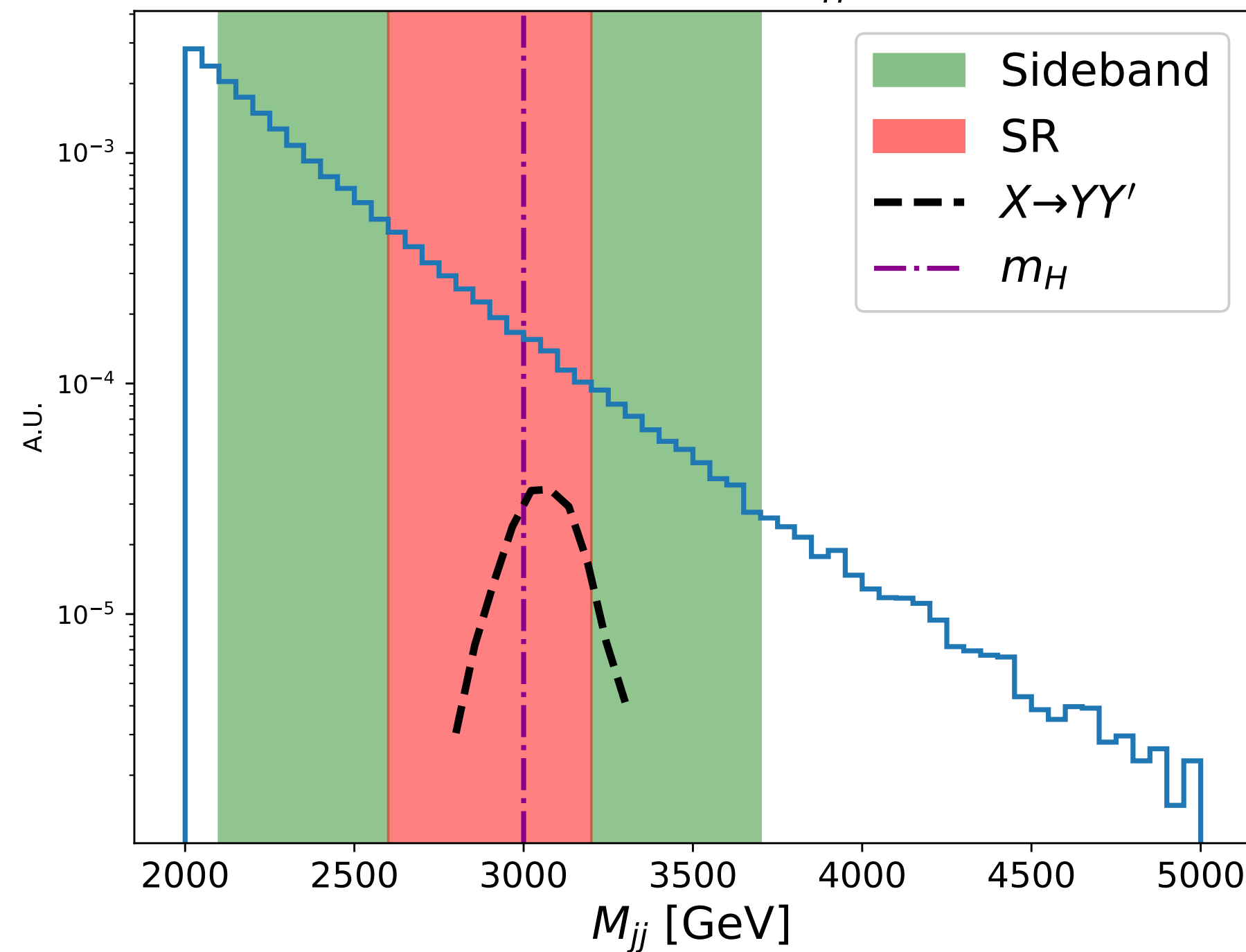
To search for a resonance with mass m_H , we need a consistent & unbiased procedure for selecting anomalous events



Bump hunting

To search for a resonance with mass m_H , we need a consistent & unbiased procedure for selecting anomalous events

SR & Sideband for $m_H = 3$ TeV

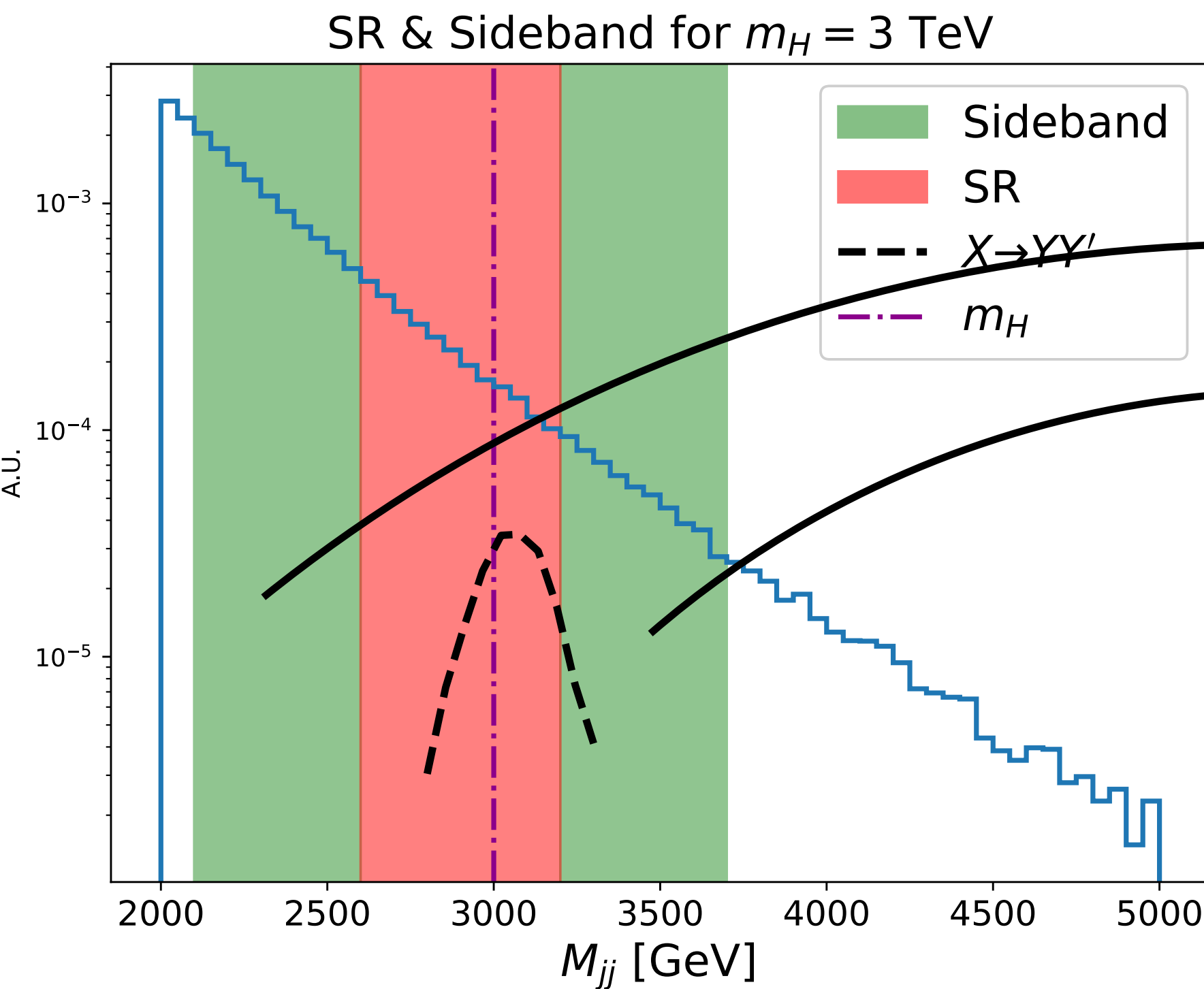


Step 1

Define a **signal region** and **sideband** around m_H — SB should be bkg-dominated

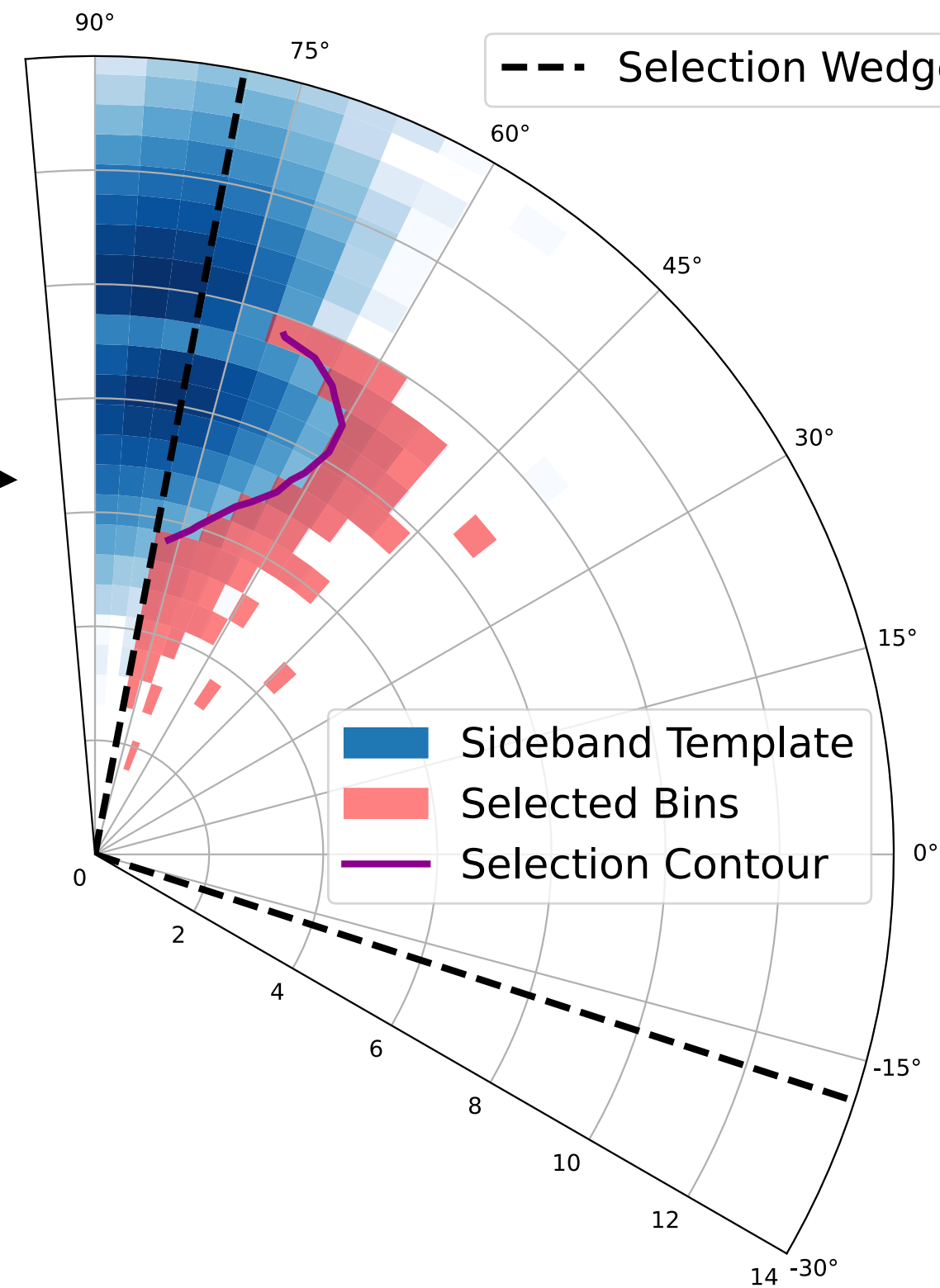
Bump hunting

To search for a resonance with mass m_H , we need a consistent & unbiased procedure for selecting anomalous events



Step 1

Define a **signal region** and **sideband** around m_H — SB should be bkg-dominated

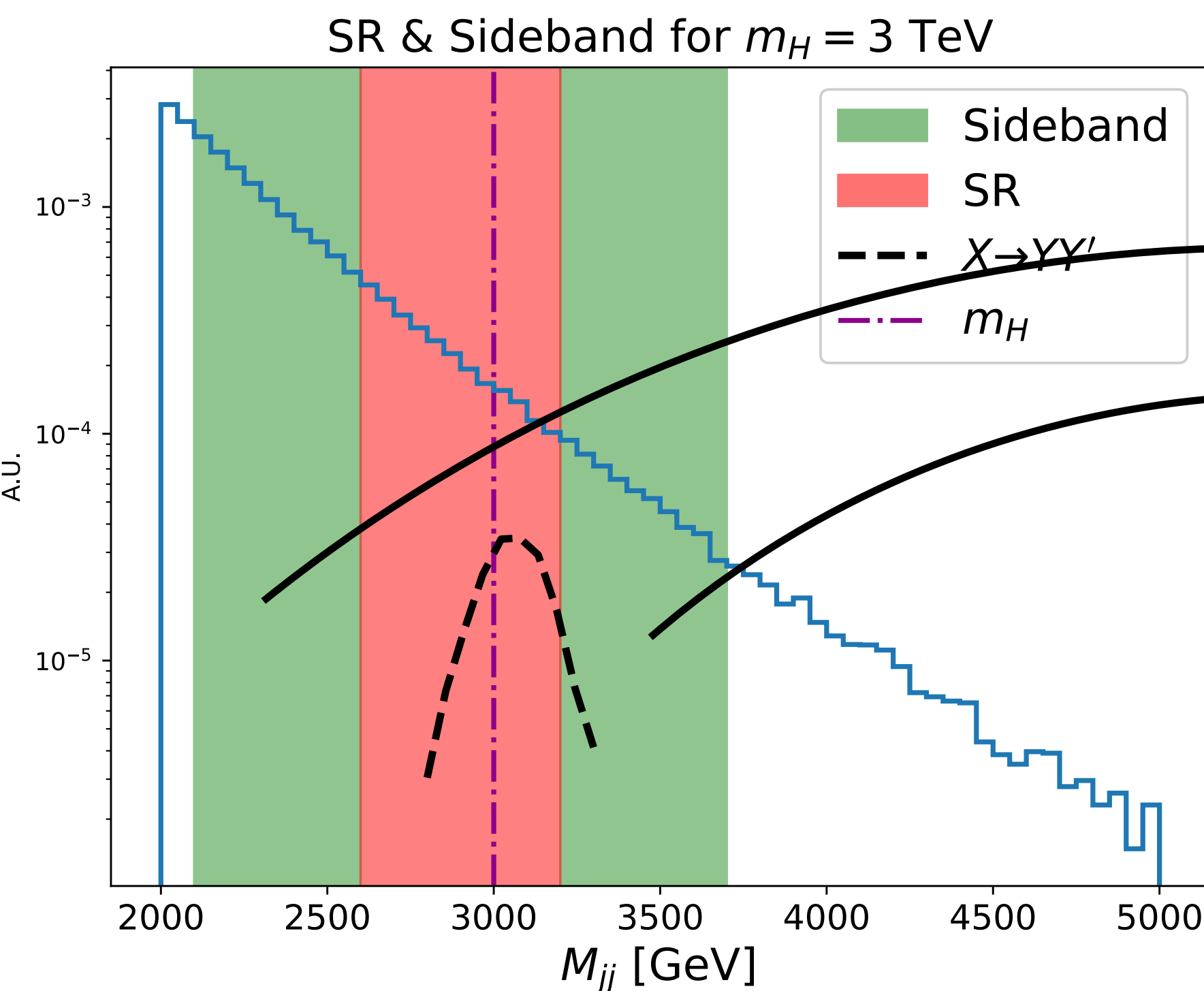


Step 2

Sideband events define a **template** in QUAK space; sparsest bins define a **selection contour** for anomalous events

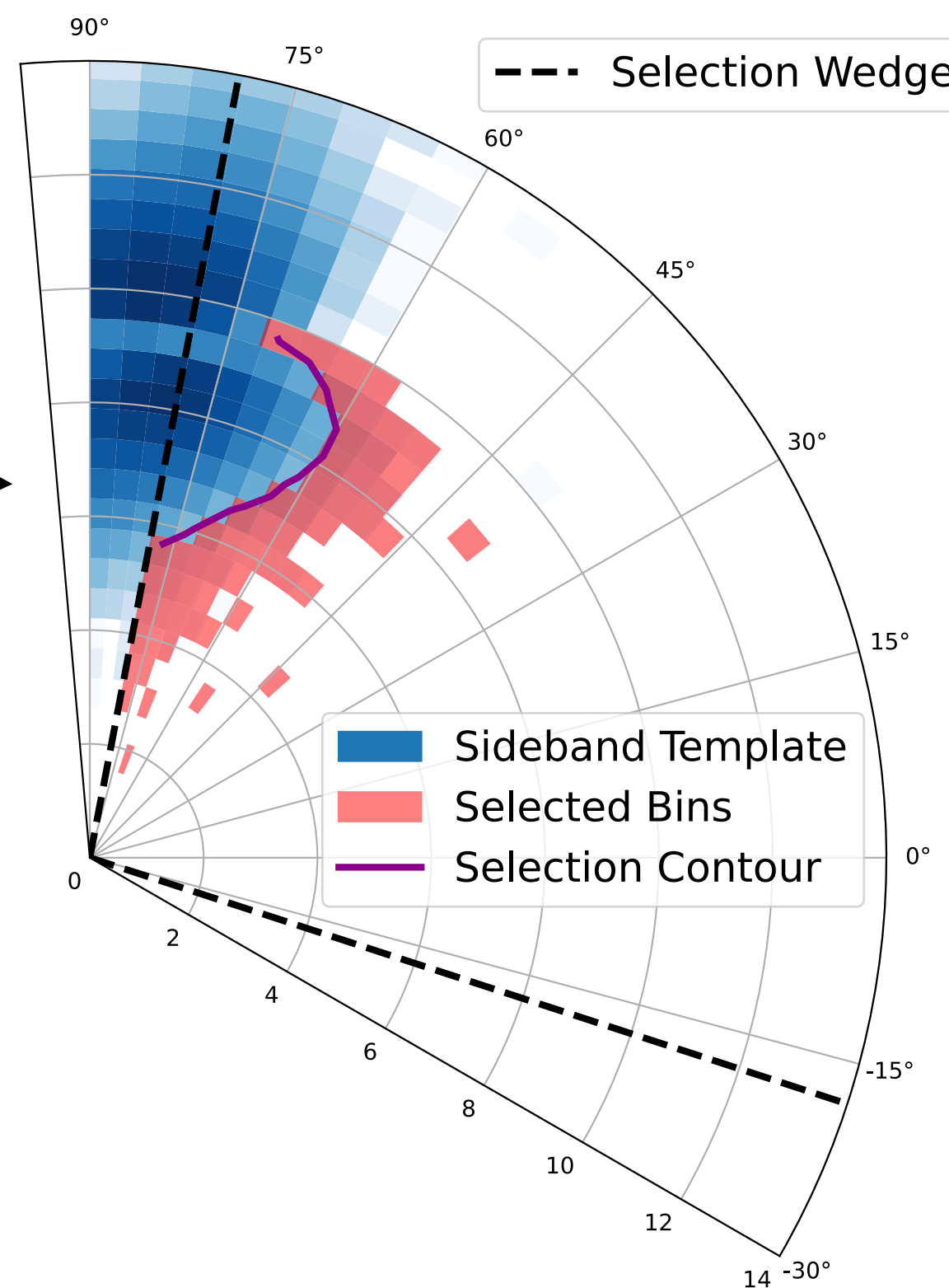
Bump hunting

To search for a resonance with mass m_H , we need a consistent & unbiased procedure for selecting anomalous events



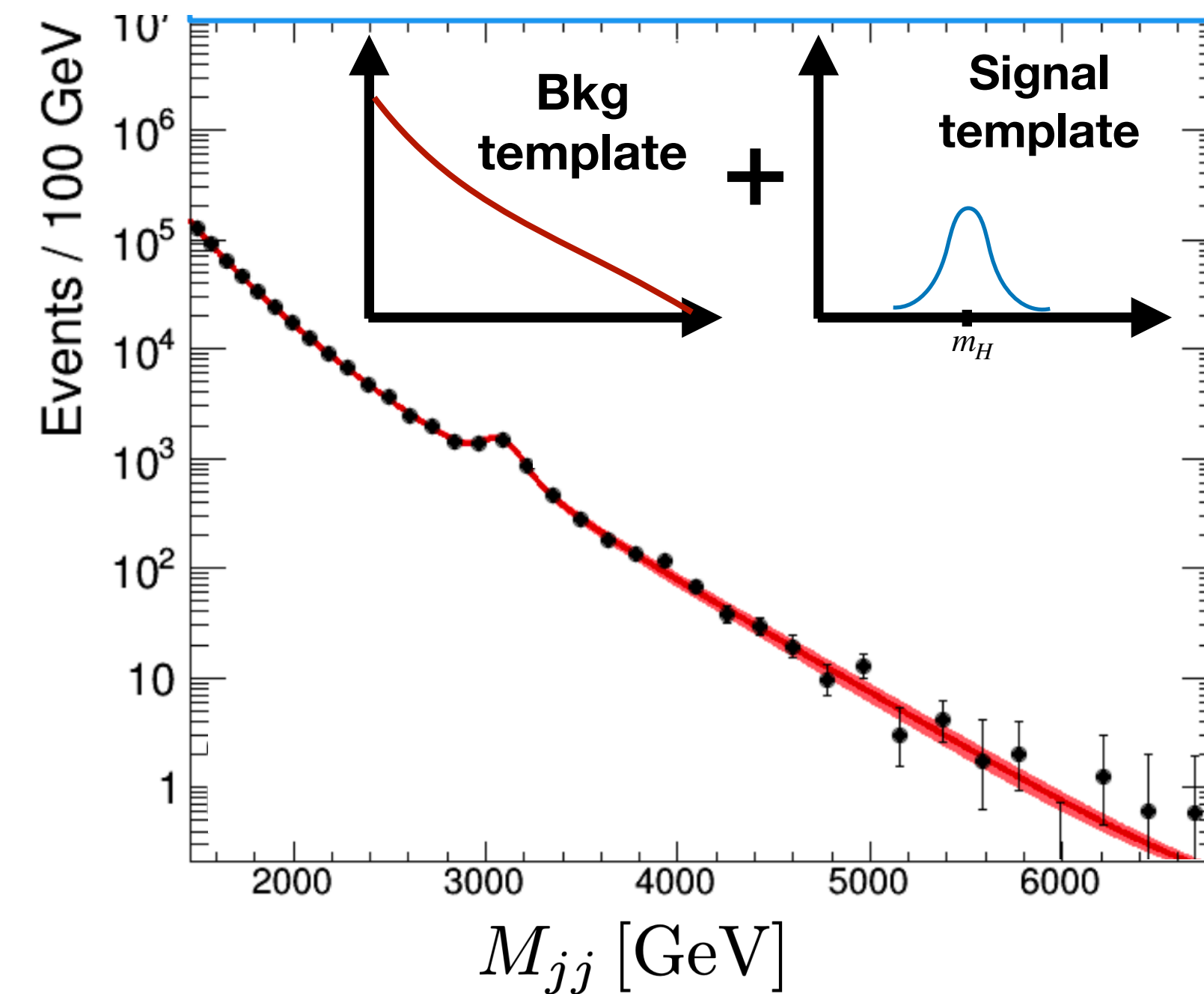
Step 1

Define a **signal region** and **sideband** around m_H — SB should be bkg-dominated



Step 2

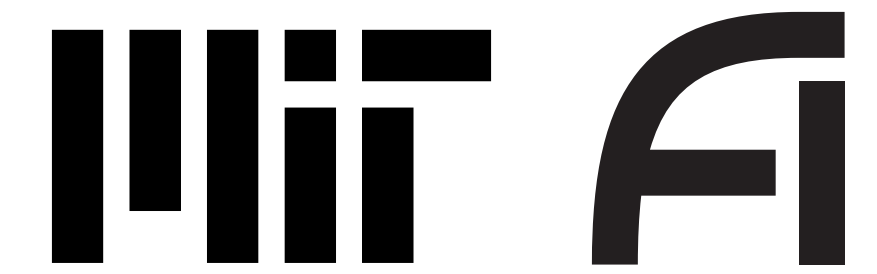
Sideband events define a **template** in QUAK space; sparsest bins define a **selection contour** for anomalous events



Step 3

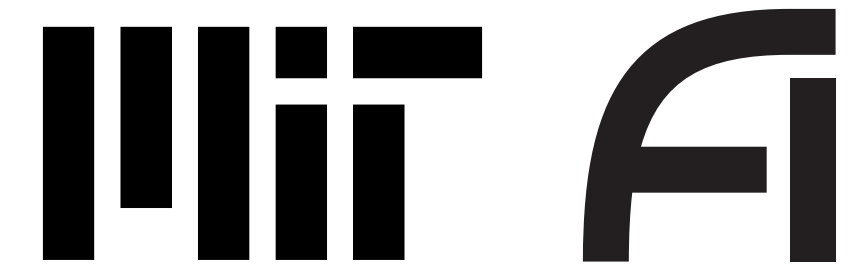
Select events based on contour & construct dijet mass spectrum
Run a fit to look for a bump

Results

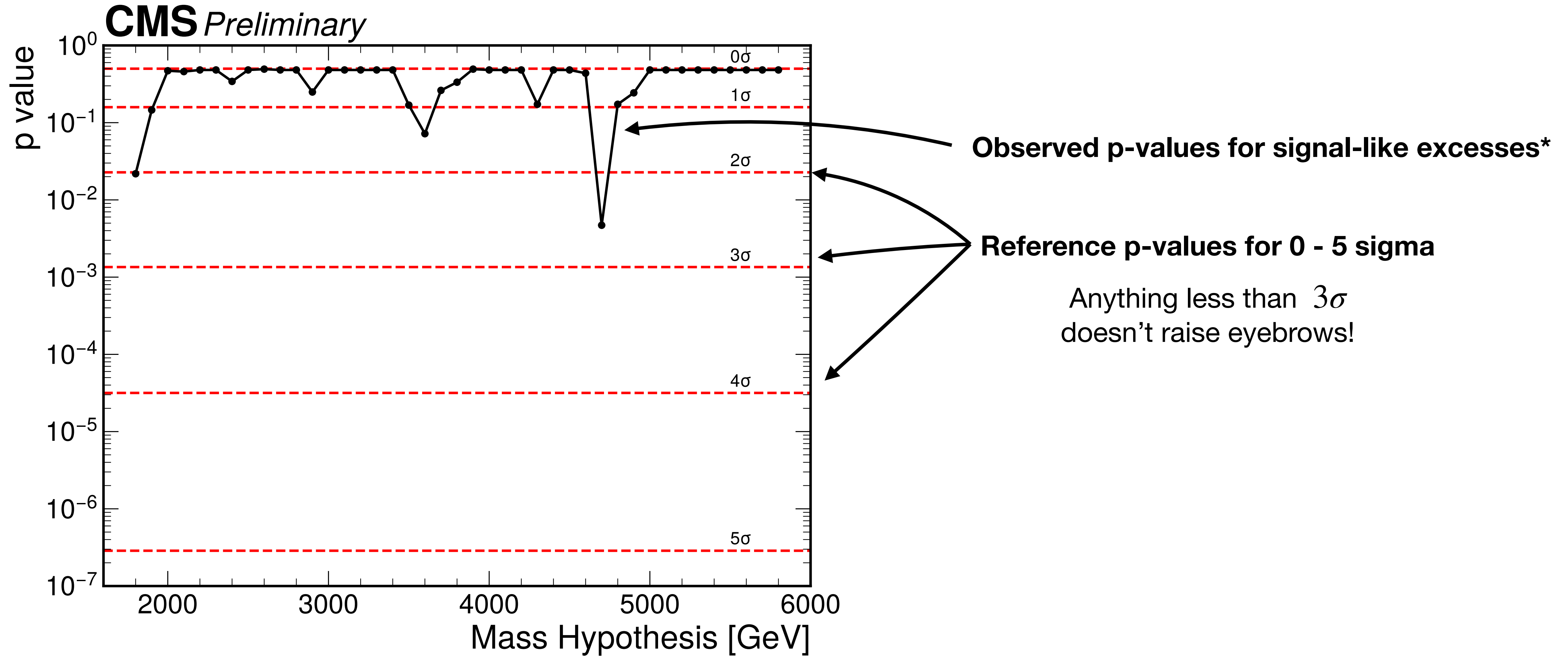


We repeat this procedure on real data for m_H between 1.8 and 6 TeV, and find...

Results

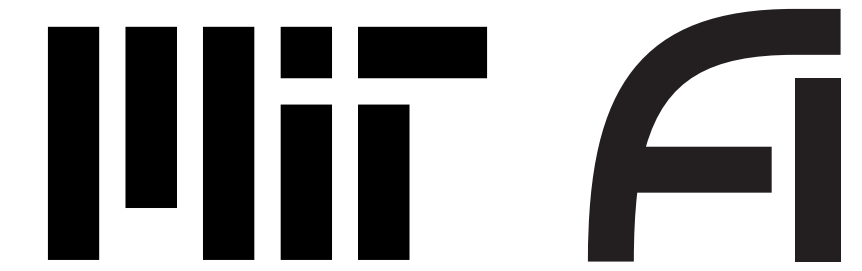


We repeat this procedure on real data for m_H between 1.8 and 6 TeV, and find...

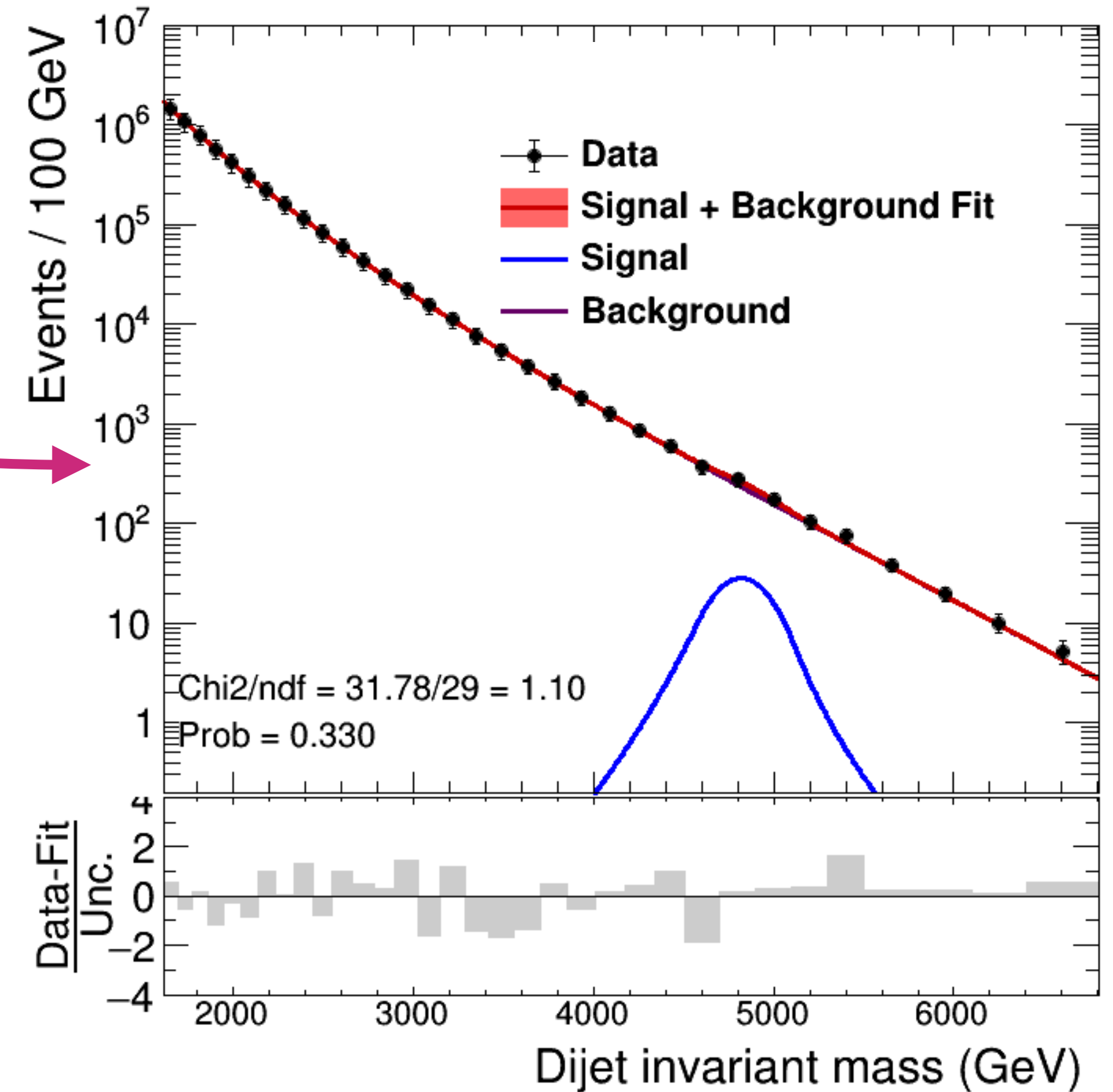
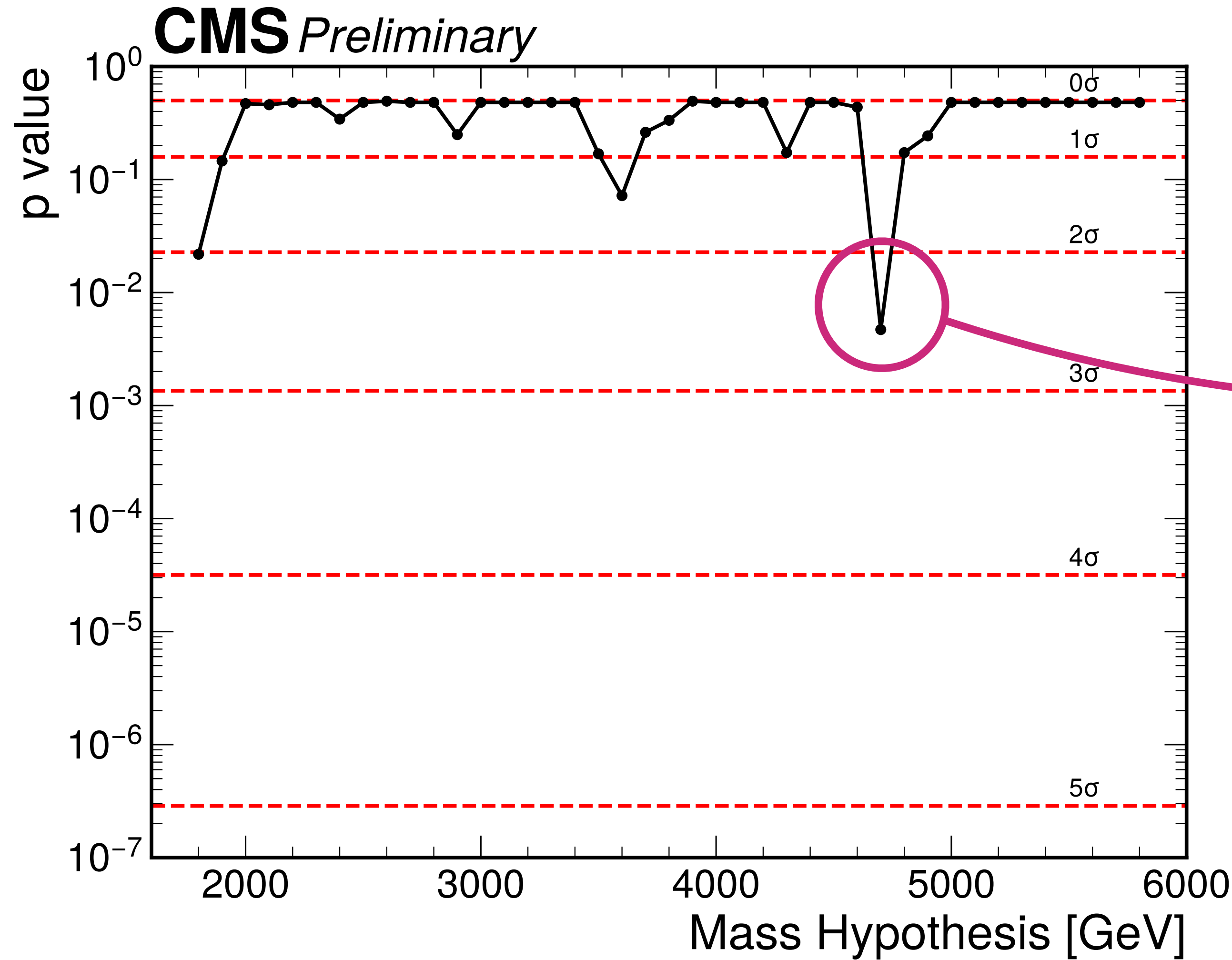


*local significance

Results



We repeat this procedure on real data for m_H between 1.8 and 6 TeV, and find...

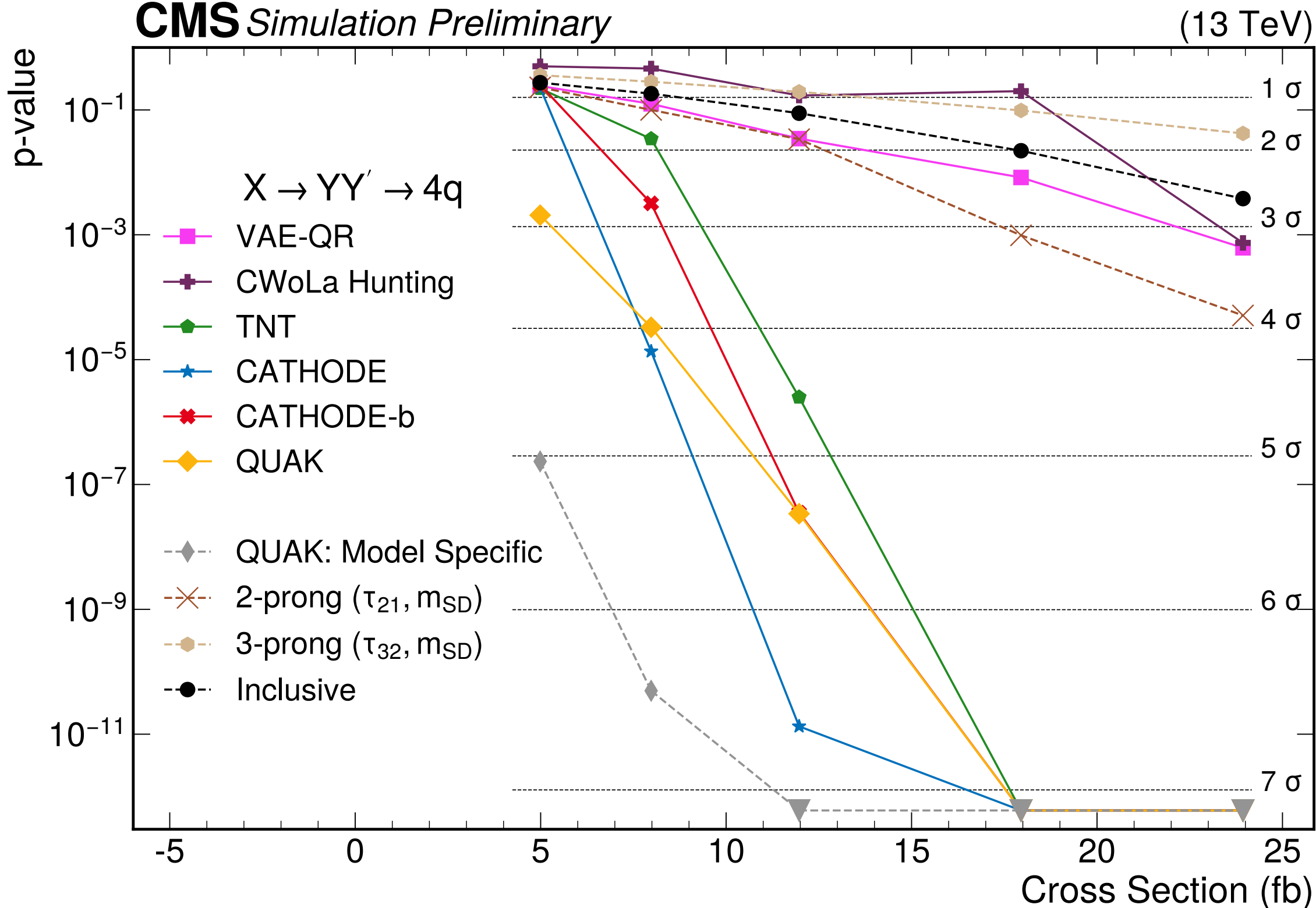


*local significance

Sensitivity to new signals



No new physics in the data, but *more importantly* we can demonstrate our sensitivity to potential signals with **injections**



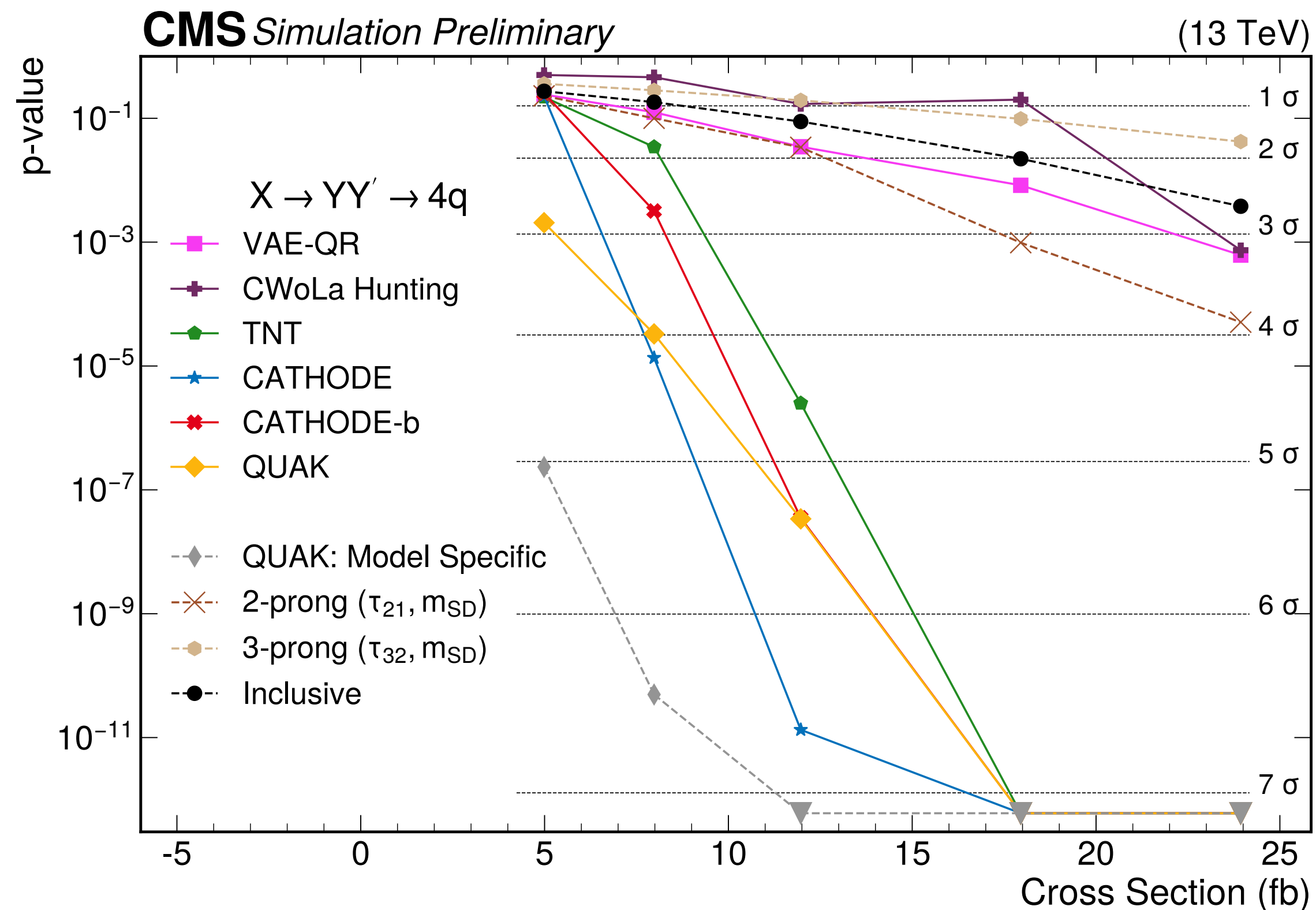
$$X \rightarrow YY', Y \rightarrow 2q, Y' \rightarrow 2q$$

2+2-prong, $m_X = 3$ TeV, $m_Y = m_{Y'} = 170$ GeV

Sensitivity to new signals

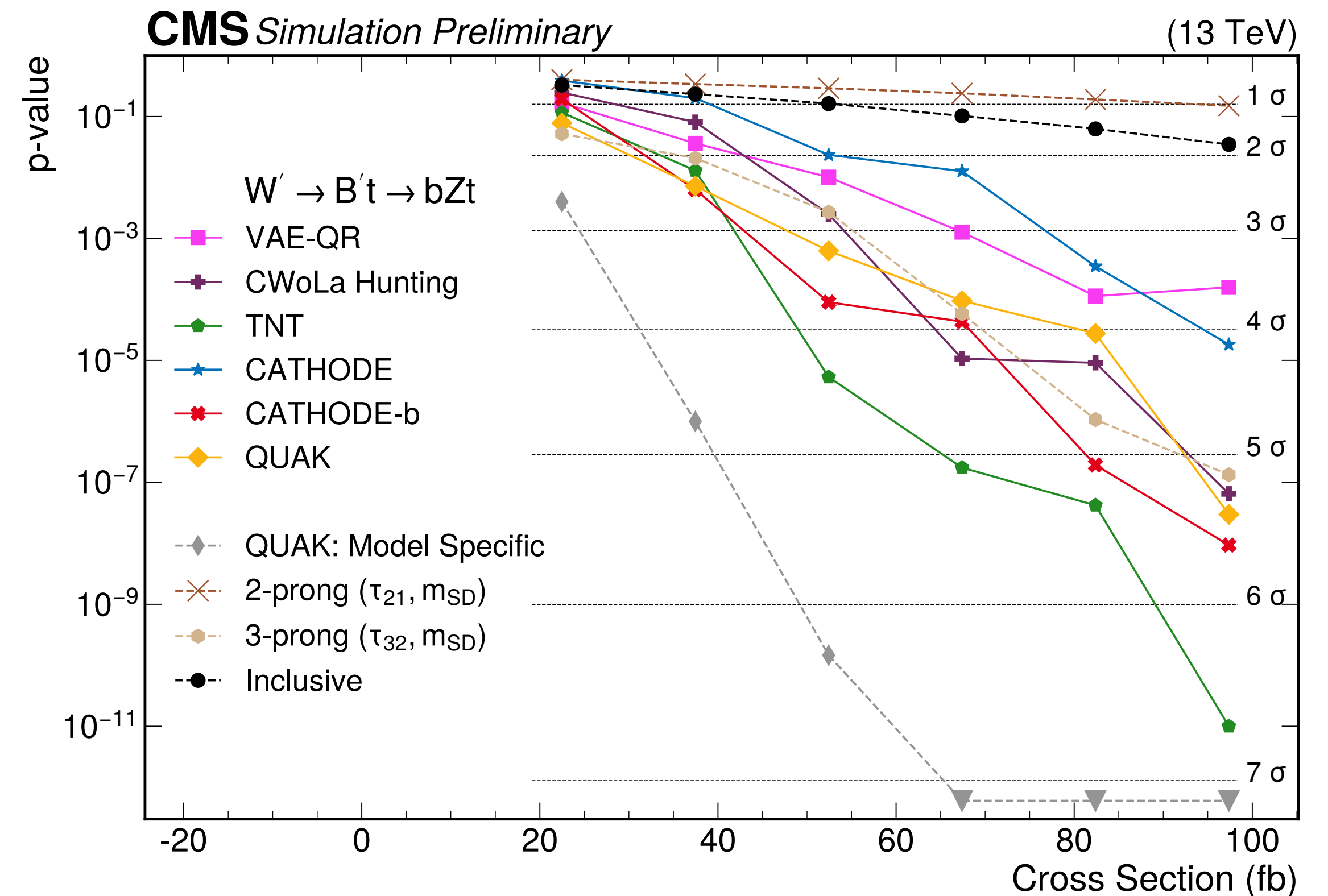


No new physics in the data, but *more importantly* we can demonstrate our sensitivity to potential signals with **injections**



$$X \rightarrow YY', Y \rightarrow 2q, Y' \rightarrow 2q$$

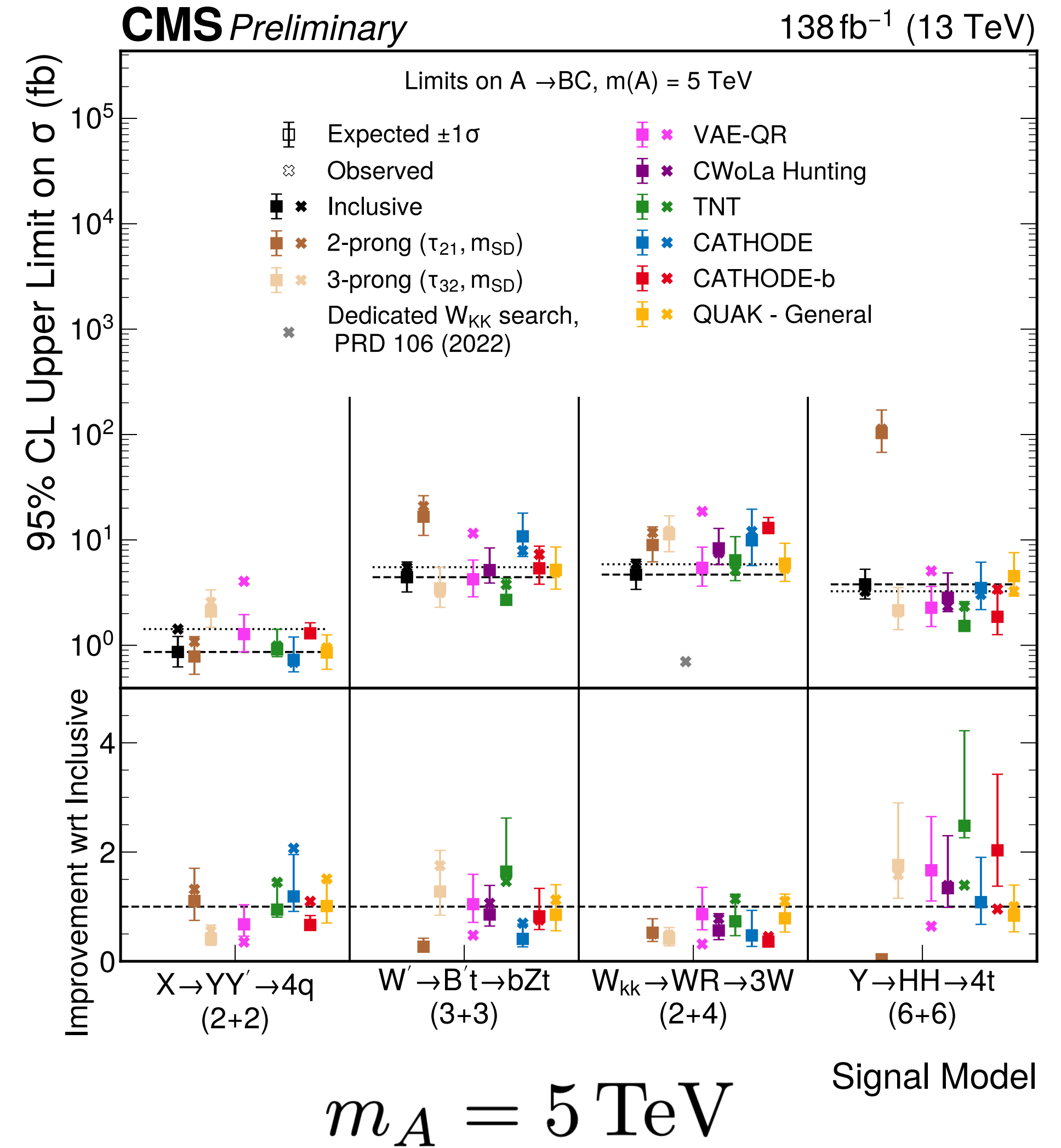
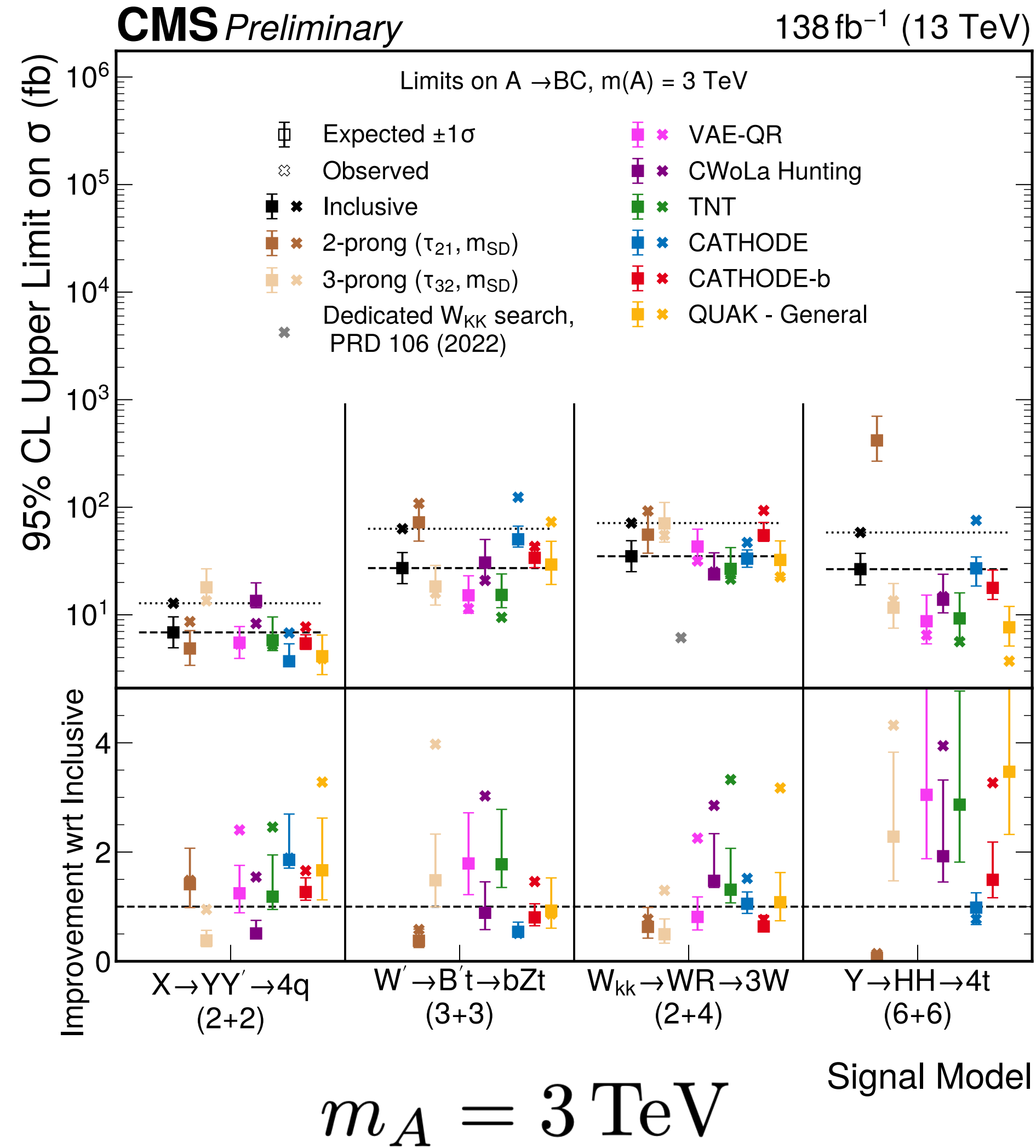
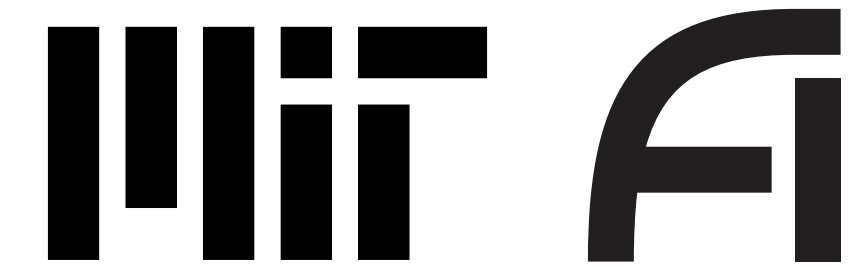
2+2-prong, $m_X = 3 \text{ TeV}$, $m_Y = m_{Y'} = 170 \text{ GeV}$



$$W' \rightarrow B't, B' \rightarrow bZ$$

3+3-prong, $m_{W'} = 3 \text{ TeV}$, $m_{B'} = 400 \text{ GeV}$

Limits

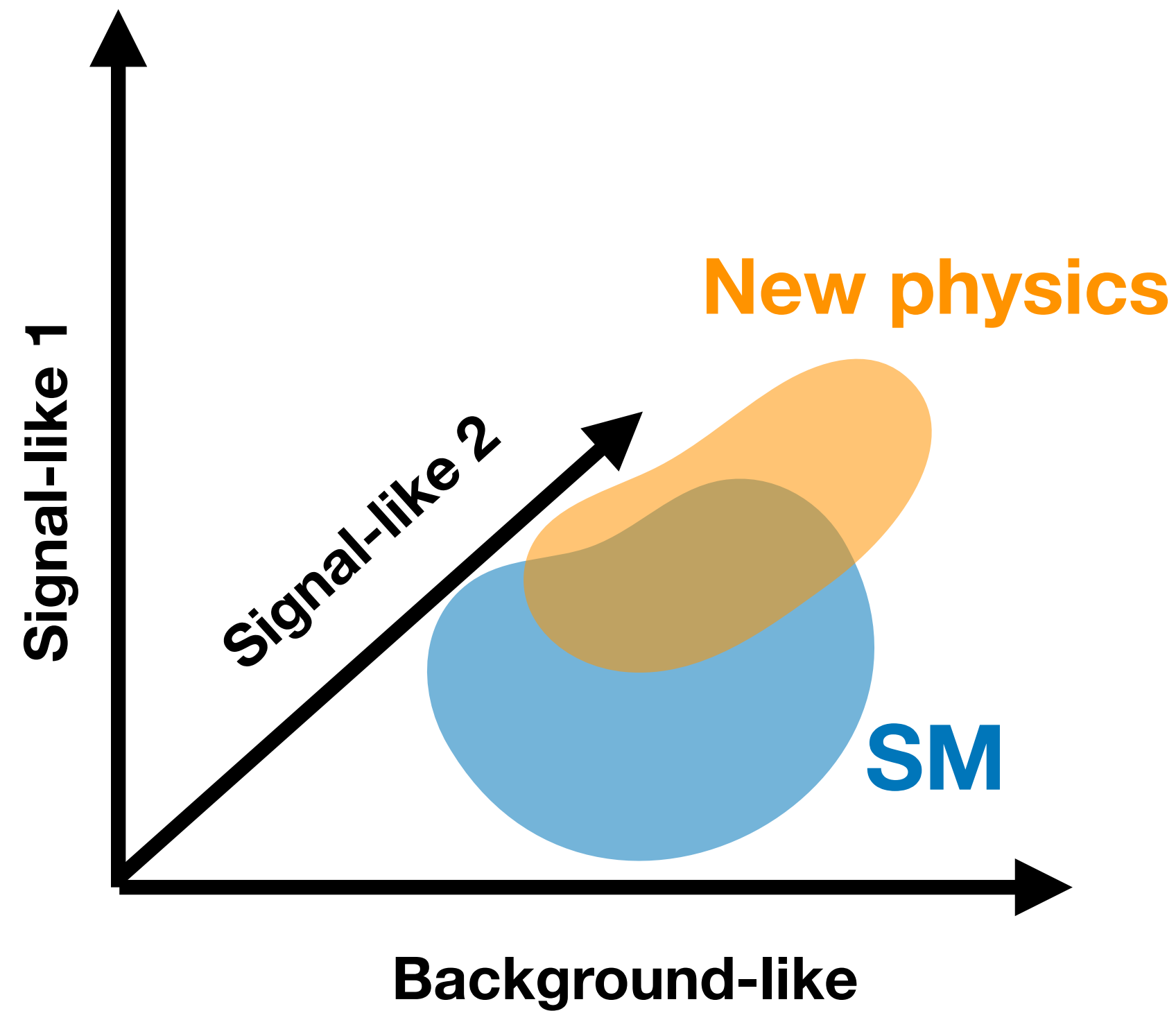


First collider limits for nearly all of these benchmark signals! All with a *single* analysis framework

Self-supervised spaces

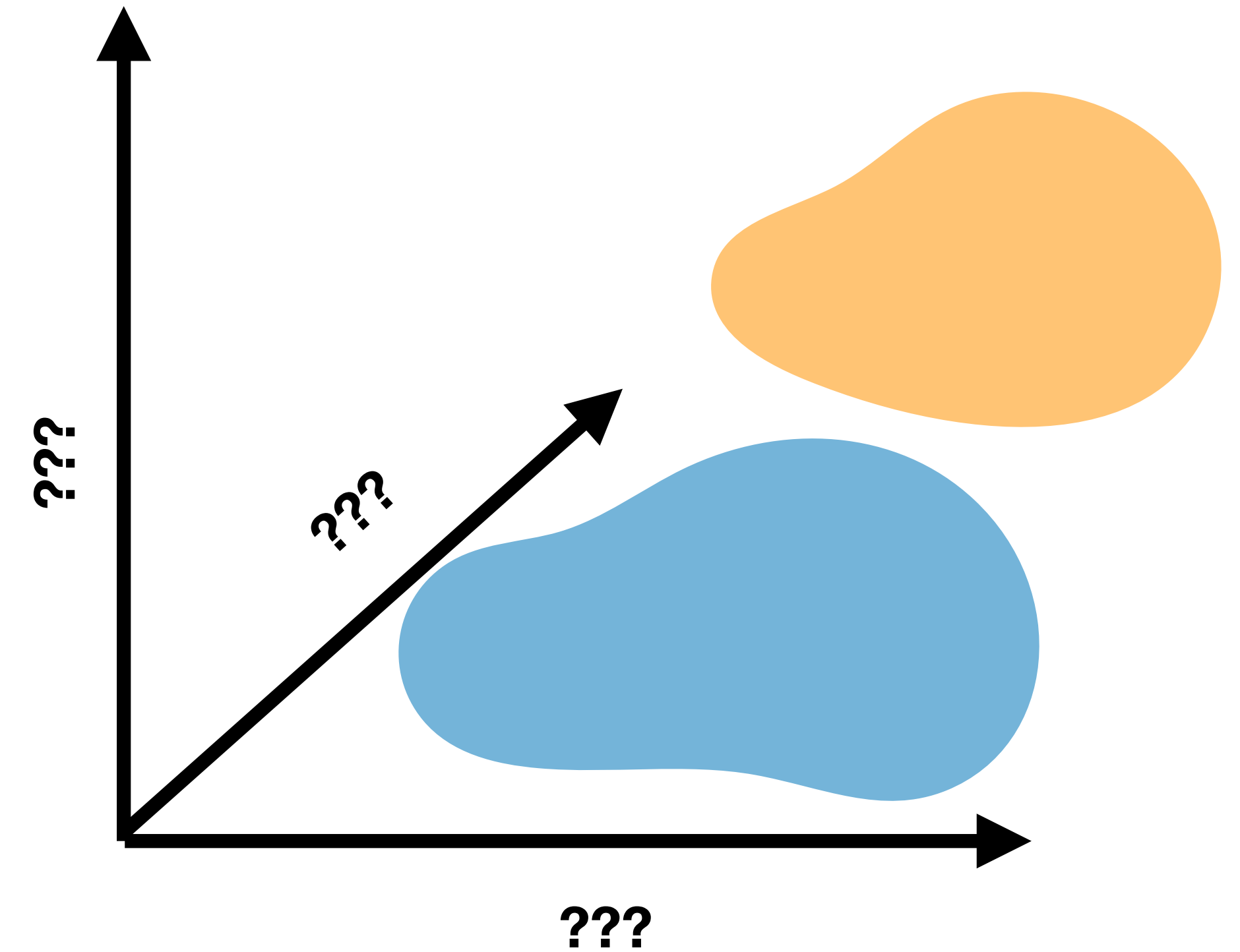
Built “by the machine”

Self-supervised approach



QUAK-style

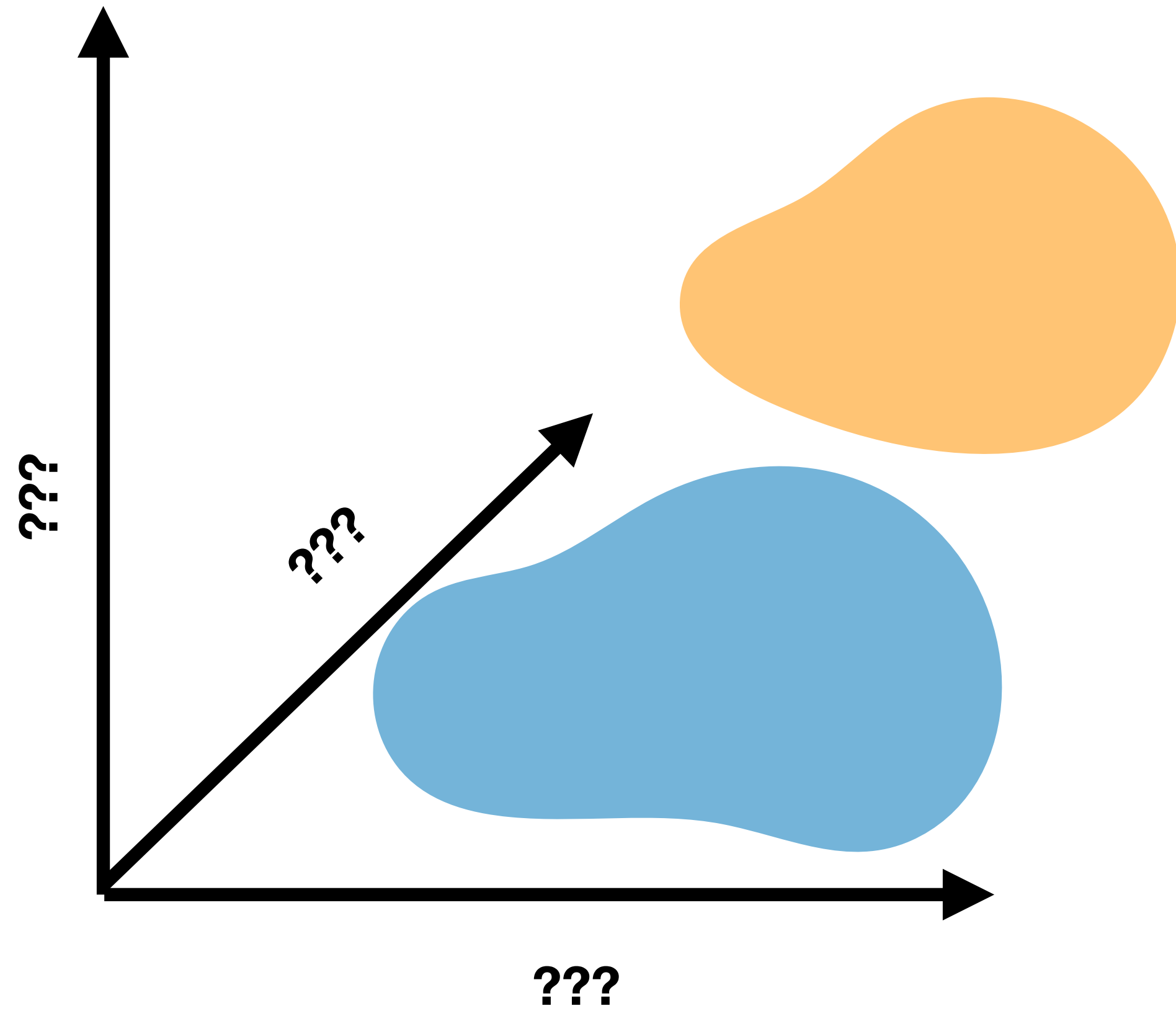
- Axes defined by *our choice* of priors
- Space is *interpretable* but is it *optimal*?



FM-style

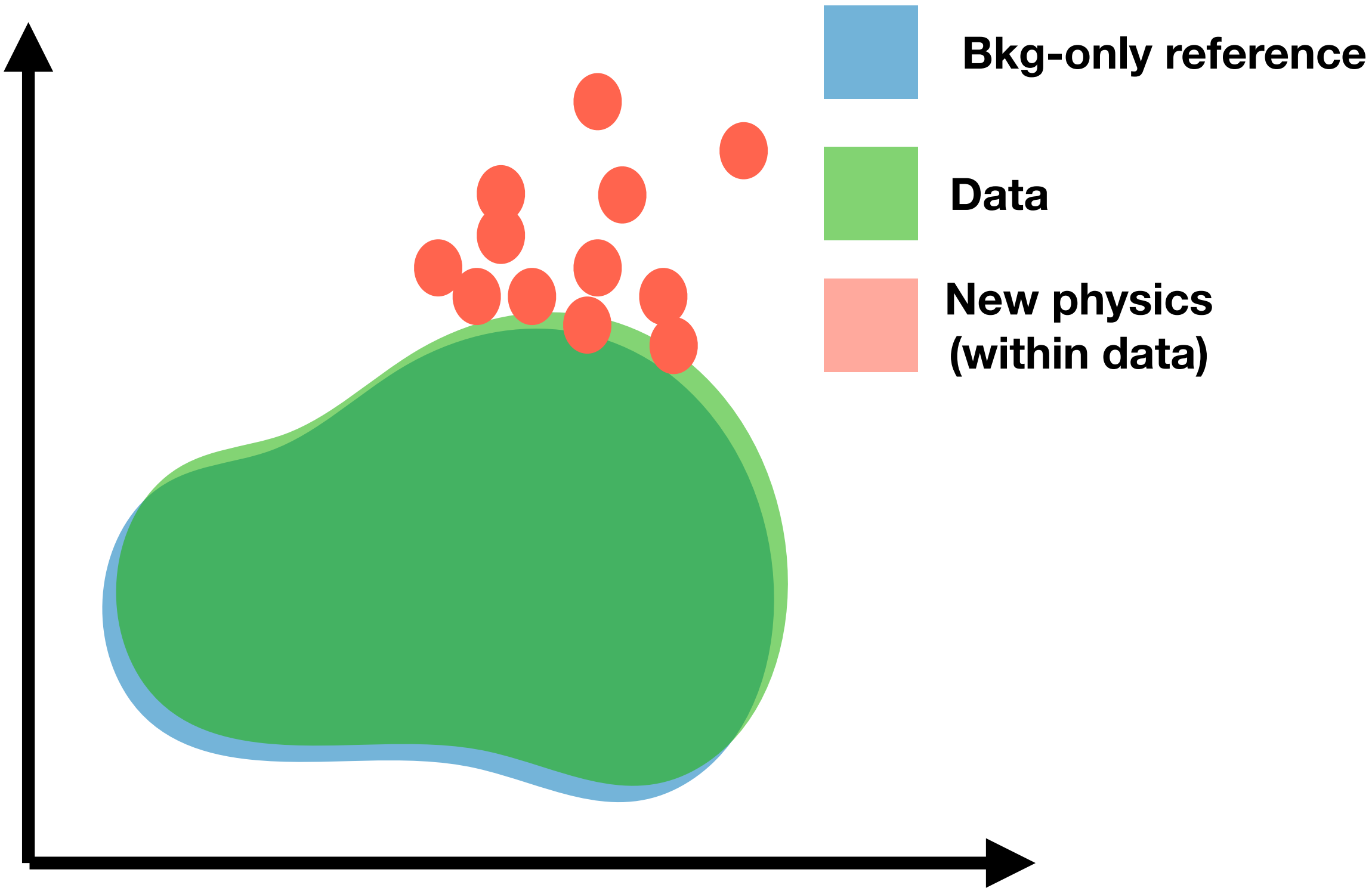
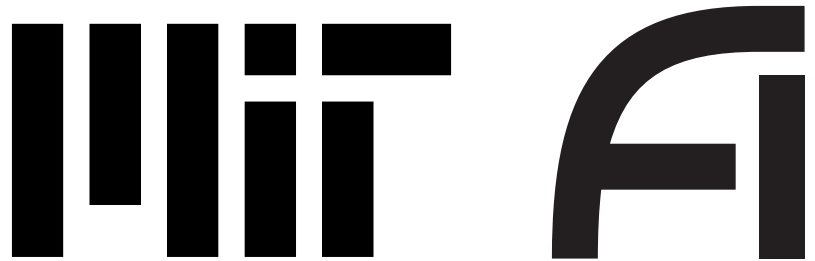
- Axes are *learned* via self-supervision
- Not interpretable, but perhaps the model can learn/encode very sophisticated things

Anomaly detection



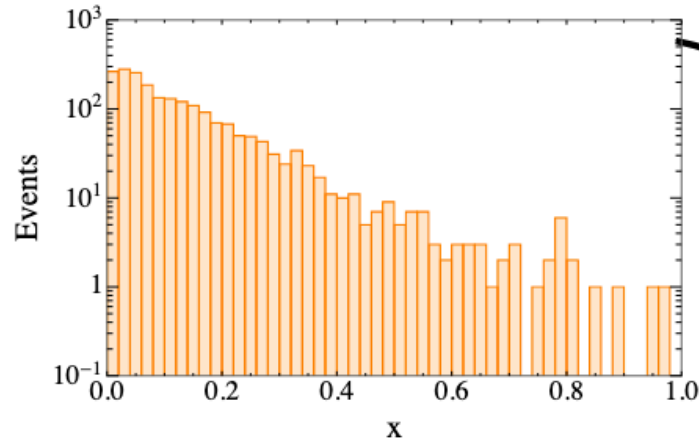
Our space no longer has human-interpretable structure - how do we do anomaly detection?

Anomaly detection

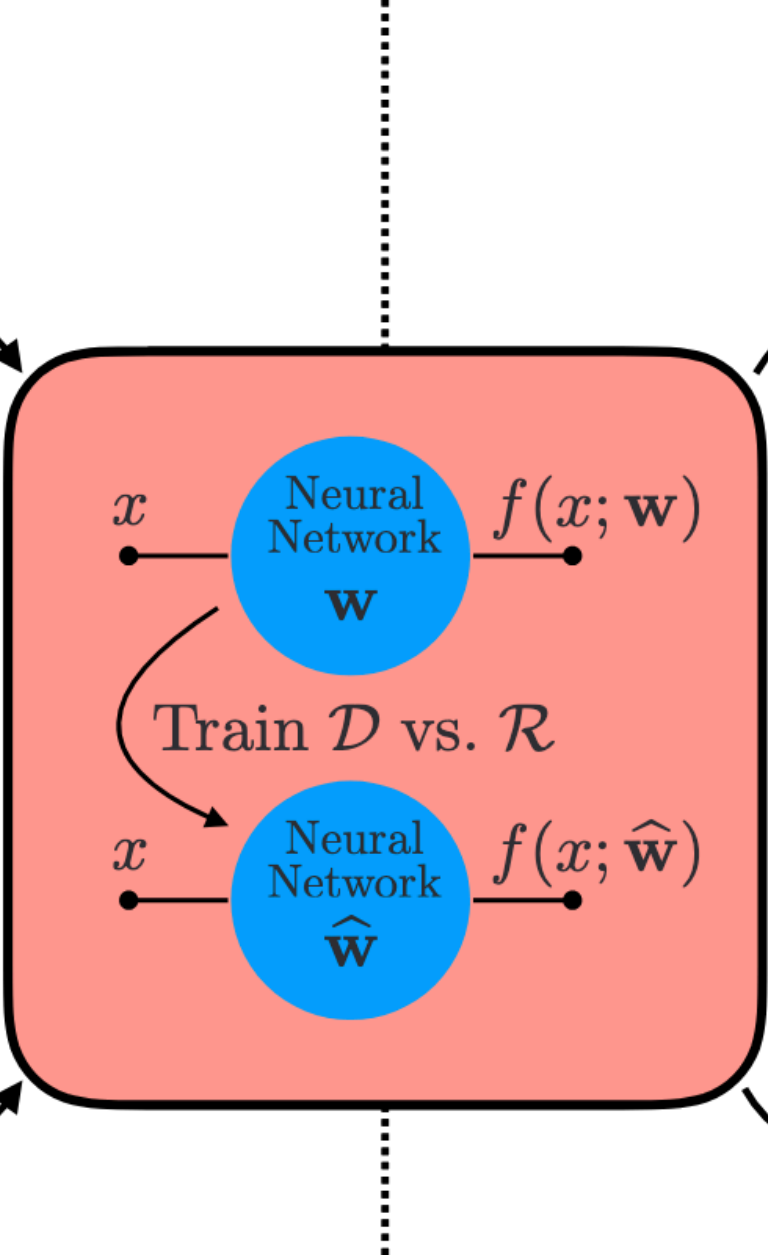
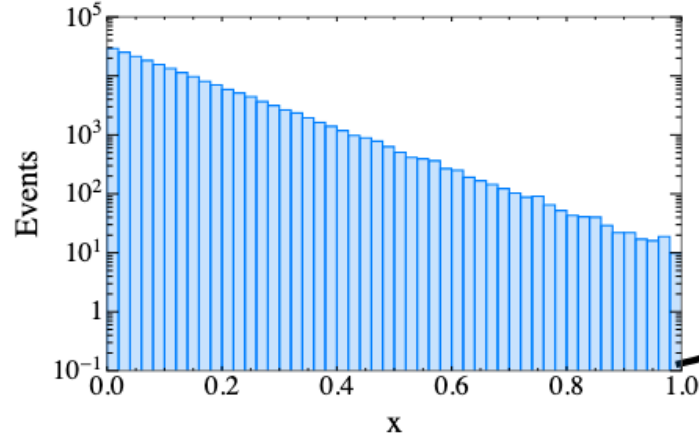


INPUT

Data sample \mathcal{D}

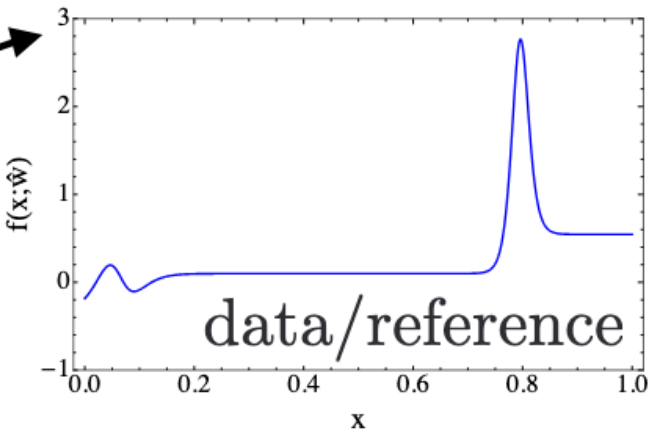


Reference sample \mathcal{R}



OUTPUT

Dist. log ratio



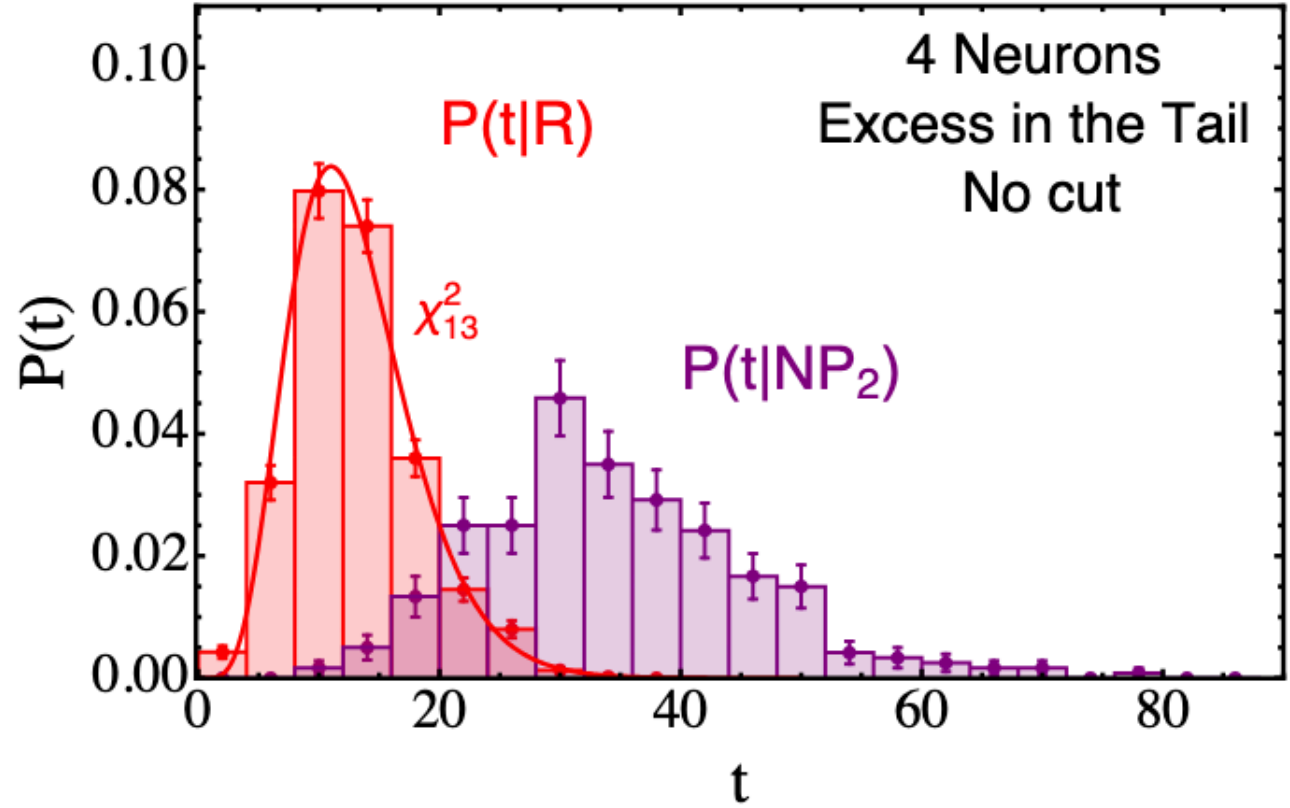
$$f(x; \hat{\mathbf{w}}) \simeq \log \left[\frac{n(x|\mathcal{T})}{n(x|\mathcal{R})} \right]$$

Test statistic t computed on the data sample \mathcal{D}

$$t(\mathcal{D}) = -2 \text{Min}_{\{\mathbf{w}\}} L[f]$$

Answer: *statistical* method driven by ML

NPLM (Ask Gaia all about this)



Building the space: contrastive learning

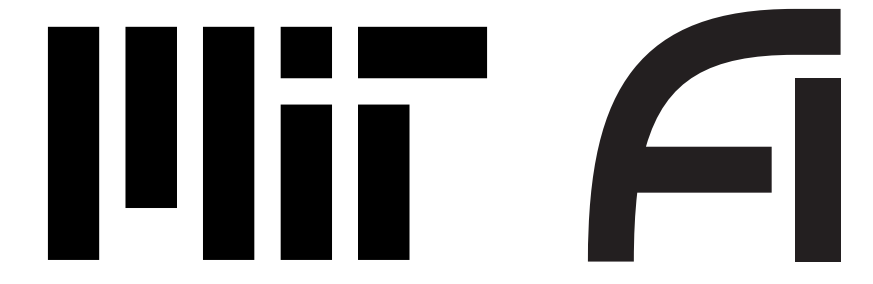


Building the space: contrastive learning

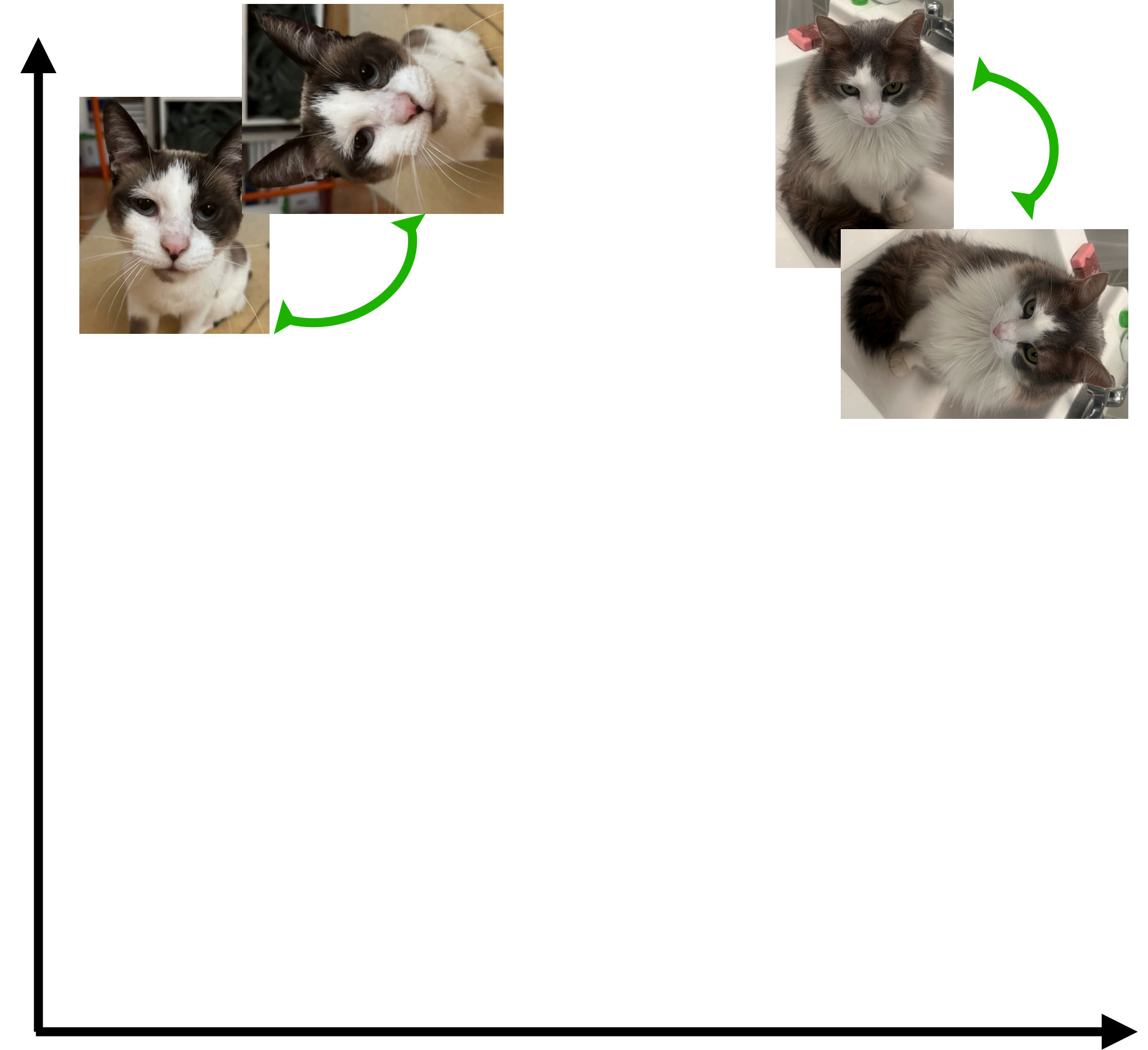
- Want to build this space in a **self-supervised** manner —
we opt for **contrastive learning**



Building the space: contrastive learning

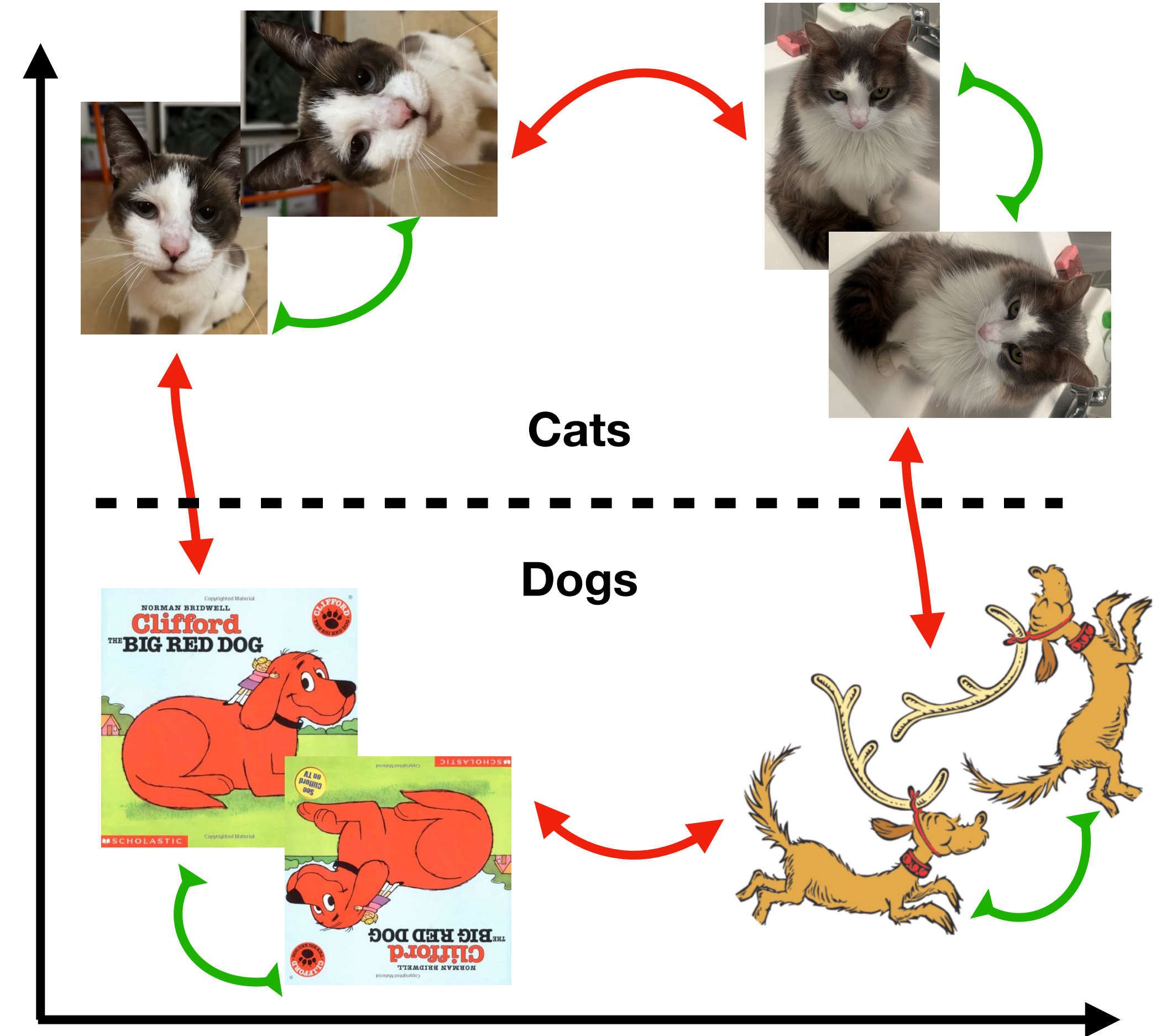


- Want to build this space in a **self-supervised** manner — we opt for **contrastive learning**
- Fundamentally based on **data augmentations**
 - ▶ For an input x , create an “augmented” version x' and teach a neural network to treat x and x' in the same way

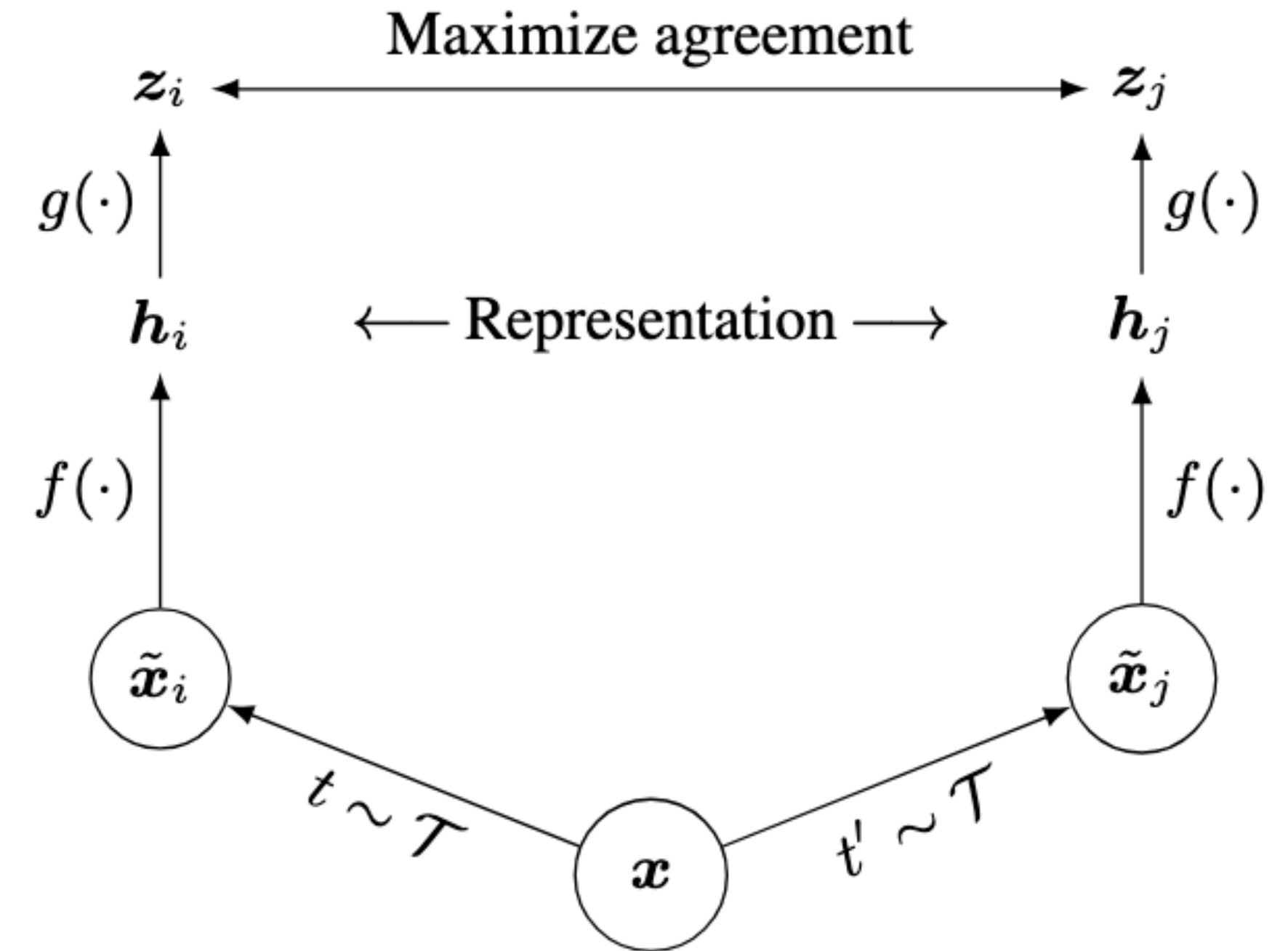
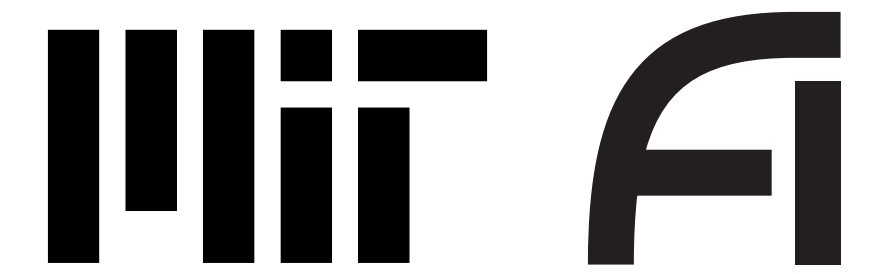


Building the space: contrastive learning

- Want to build this space in a **self-supervised** manner — we opt for **contrastive learning**
- Fundamentally based on **data augmentations**
 - ▶ For an input x , create an “augmented” version x' and teach a neural network to treat x and x' in the same way
 - ▶ All other data points are **negative samples** and should be pushed apart



(Supervised) SimCLR

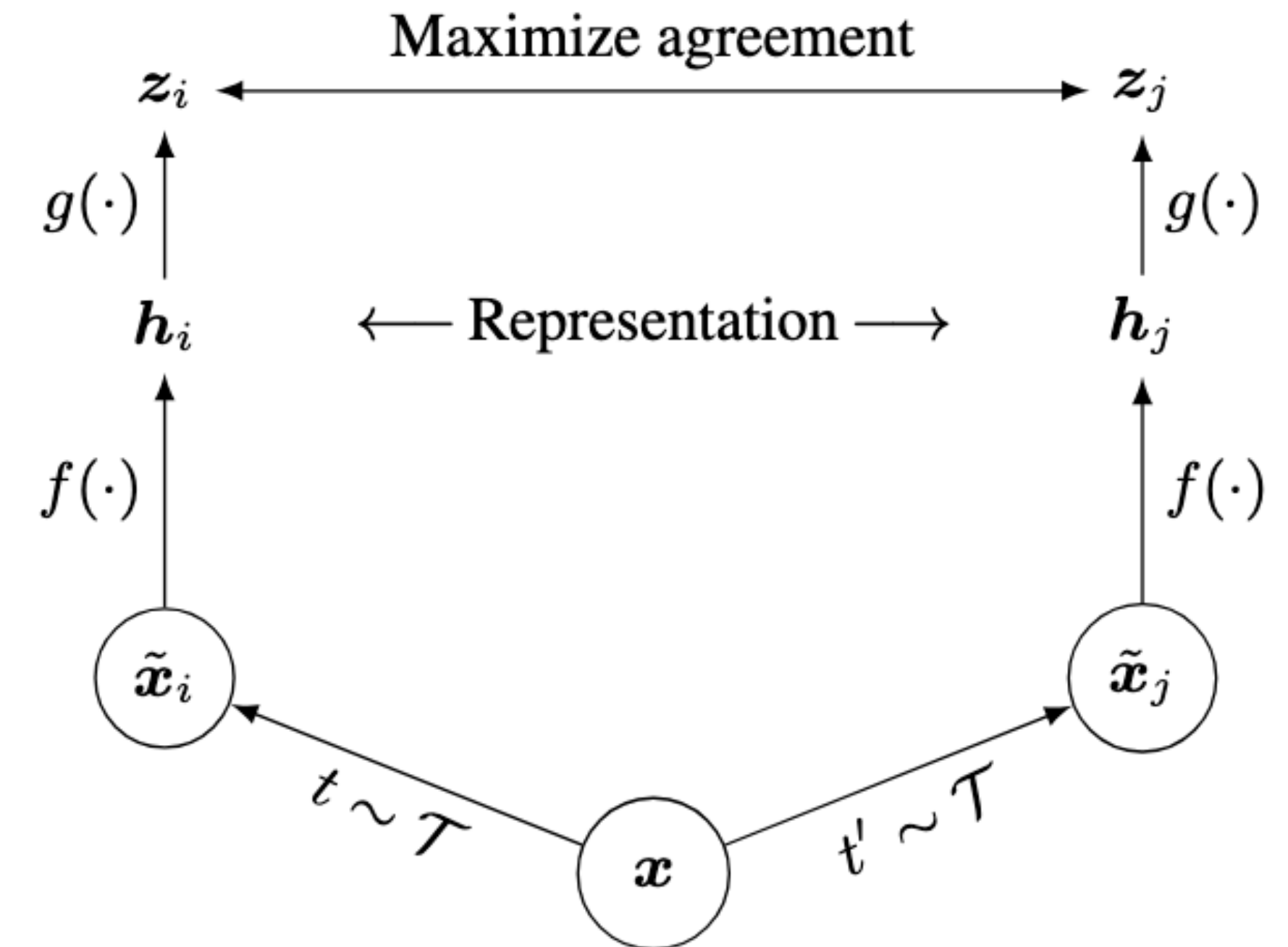


$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$$

[2002.05709](#)

(Supervised) SimCLR

- SimCLR is a popular contrastive learning setup
 - ▶ **Cosine similarity** $\hat{\mathbf{z}}_i \cdot \hat{\mathbf{z}}_j$ measures alignment
- But what augmentations are “best” for particle physics?
 - ▶ Boosts? Rotations? Energy smearing? ...

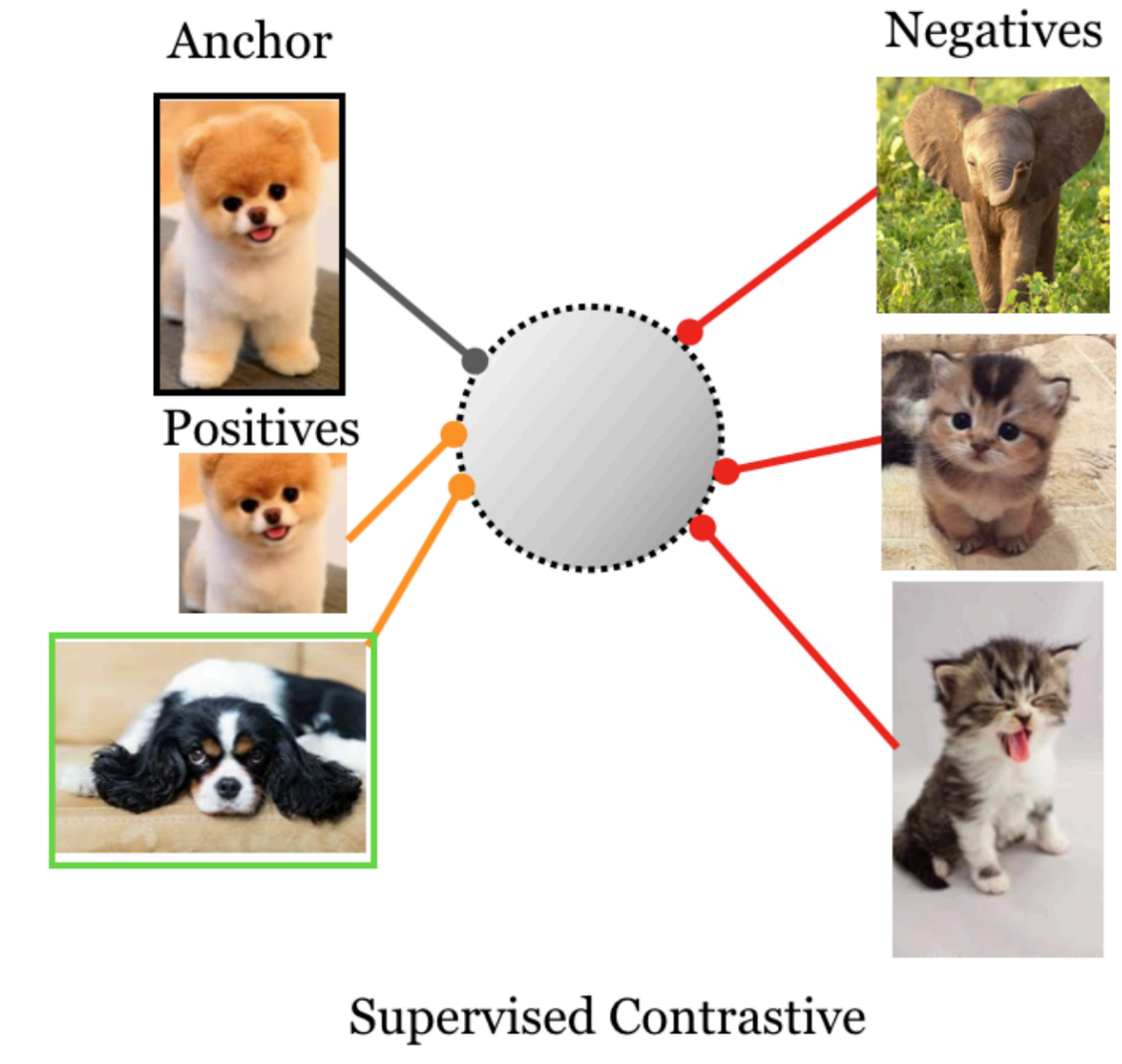
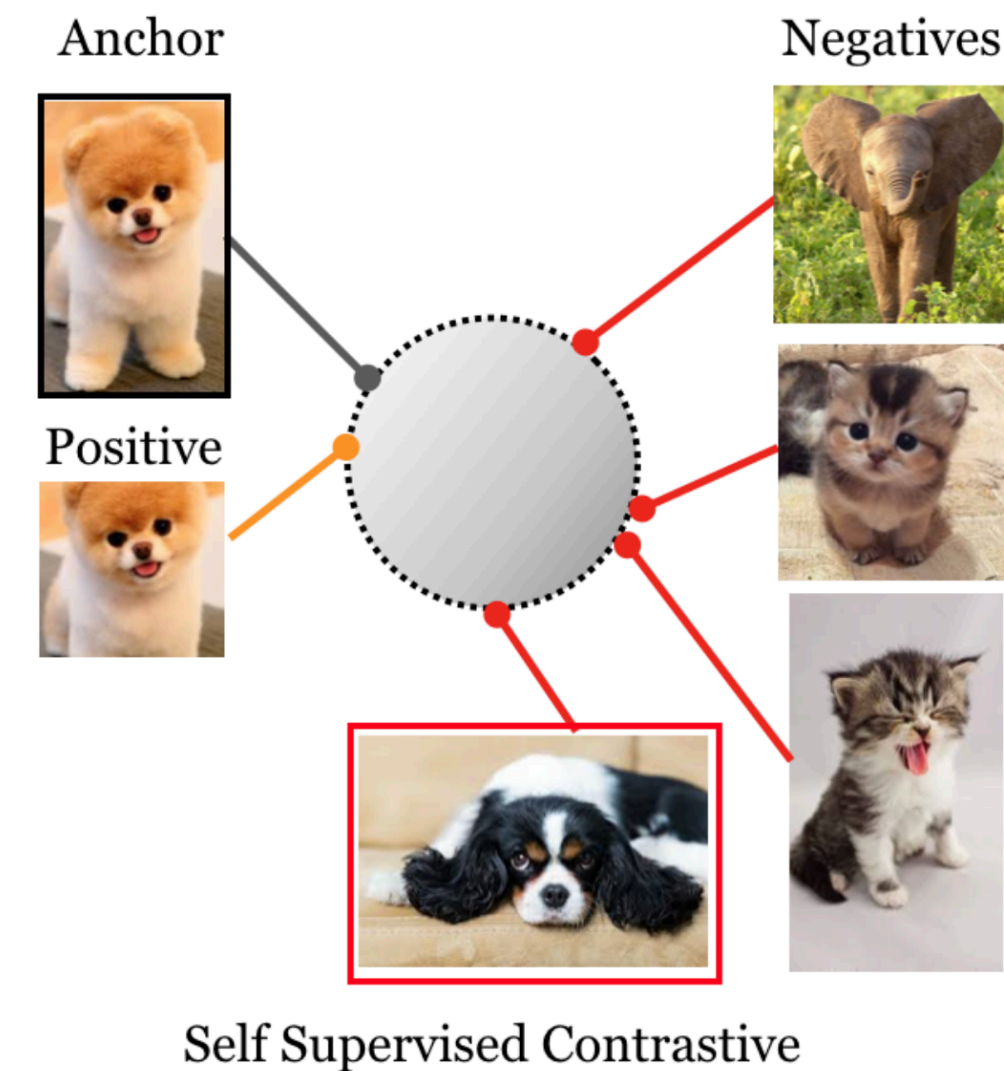


$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}$$

[2002.05709](#)

(Supervised) SimCLR

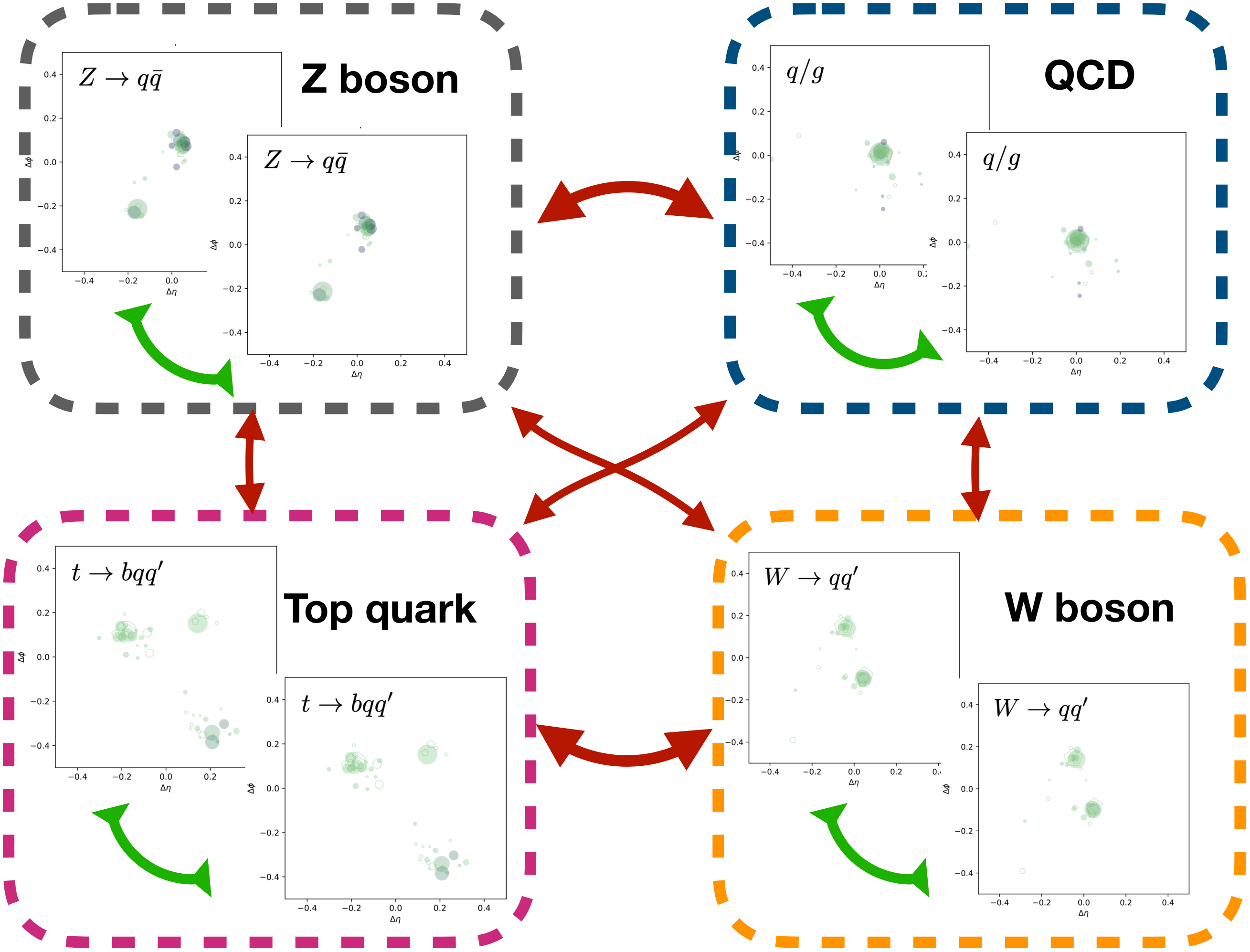
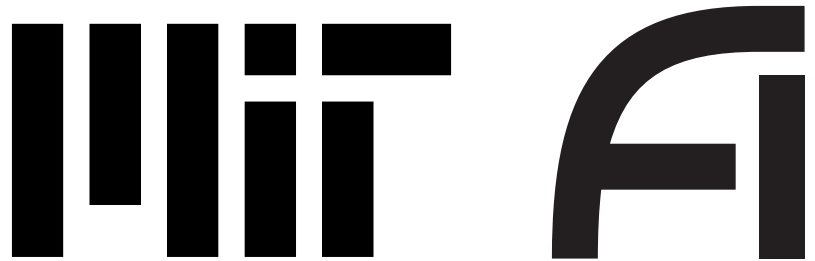
- SimCLR is a popular contrastive learning setup
 - ▶ **Cosine similarity** $\hat{\mathbf{z}}_i \cdot \hat{\mathbf{z}}_j$ measures alignment
- But what augmentations are “best” for particle physics?
 - ▶ Boosts? Rotations? Energy smearing? ...
- We *do* have an **abundance of labeled data** from simulations
 - ▶ Motivates a **supervised** approach — positive pairs are drawn from the same *class* of objects



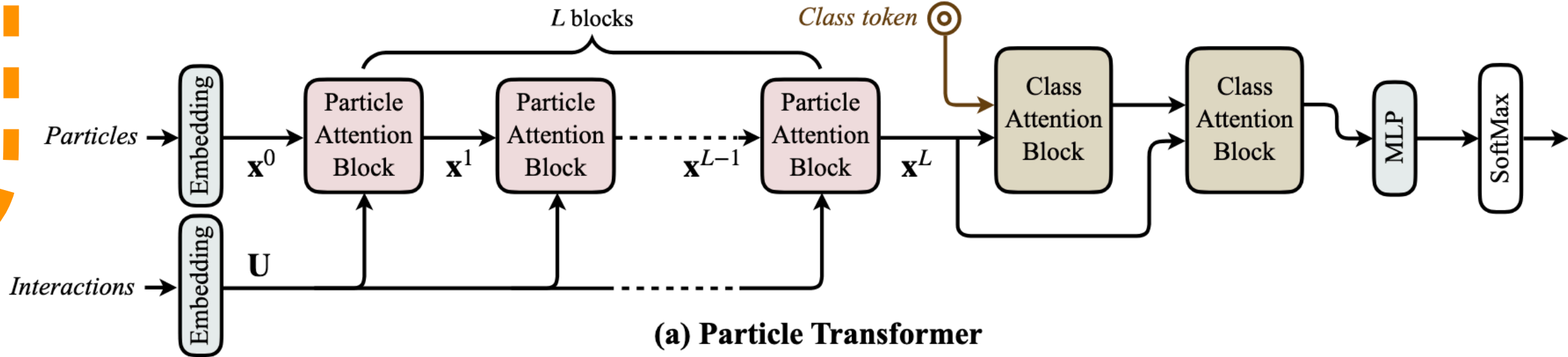
$$\mathcal{L}_{out}^{sup} = \sum_{i \in I} \mathcal{L}_{out,i}^{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)}$$

[2004.11362](#)

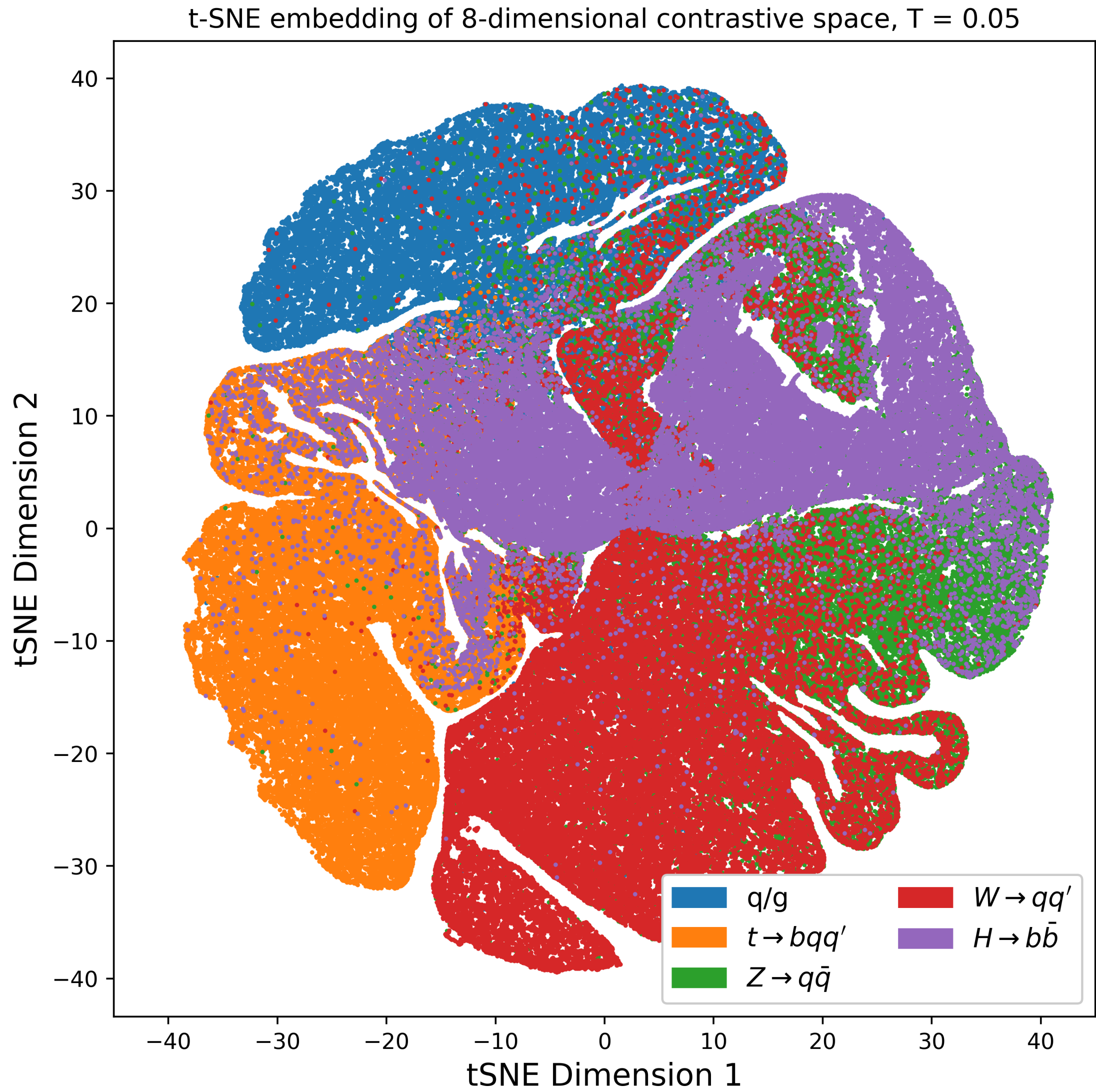
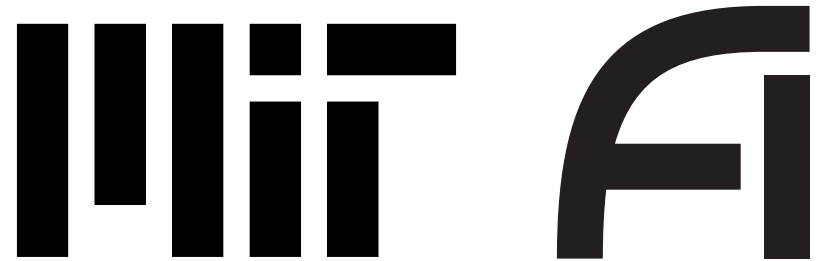
Training on jets



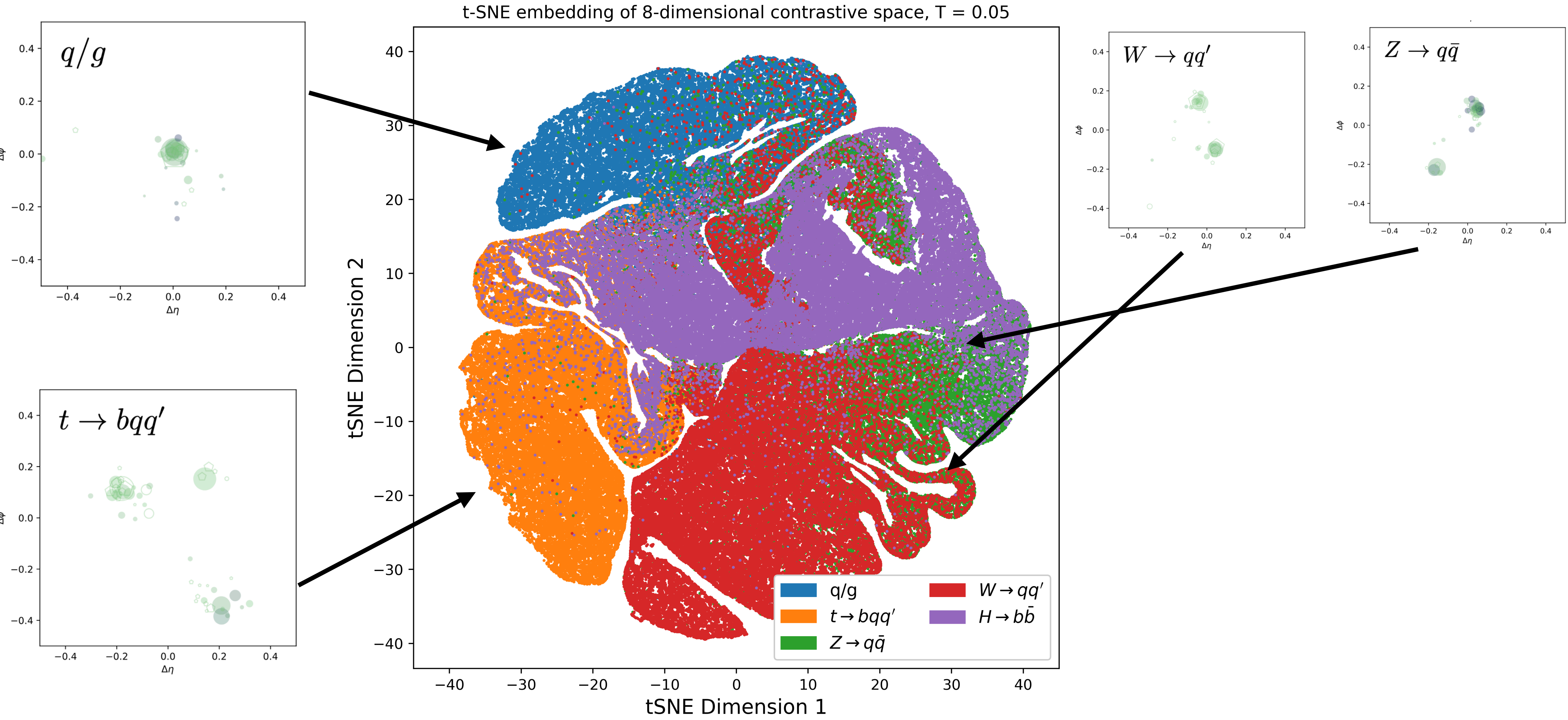
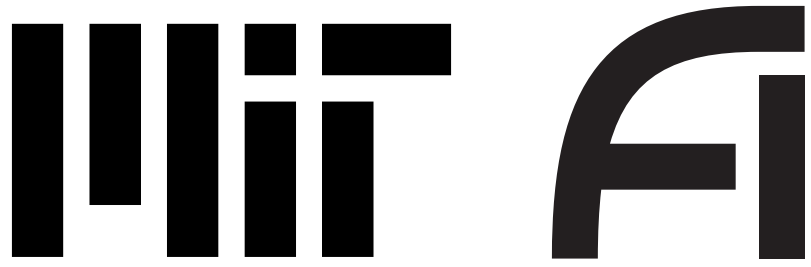
- Train supervised SimCLR embeddings of **individual jets** from four different classes (common SM processes)
- Use the **particle transformer** architecture
 - ▶ Full attention between all particles in a jet (up to 128)
 - ▶ 17 input features for each particle (kinematic, ID, trajectory)
 - ▶ Variable-dim embedding spaces



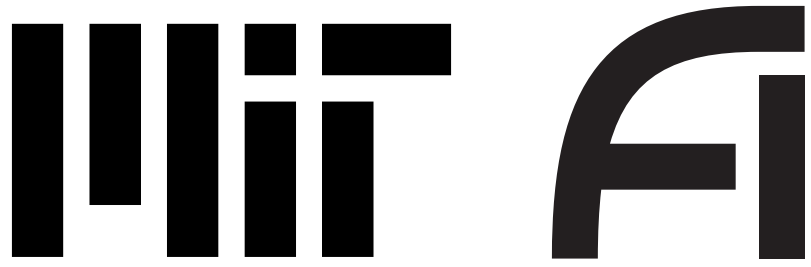
The learned space



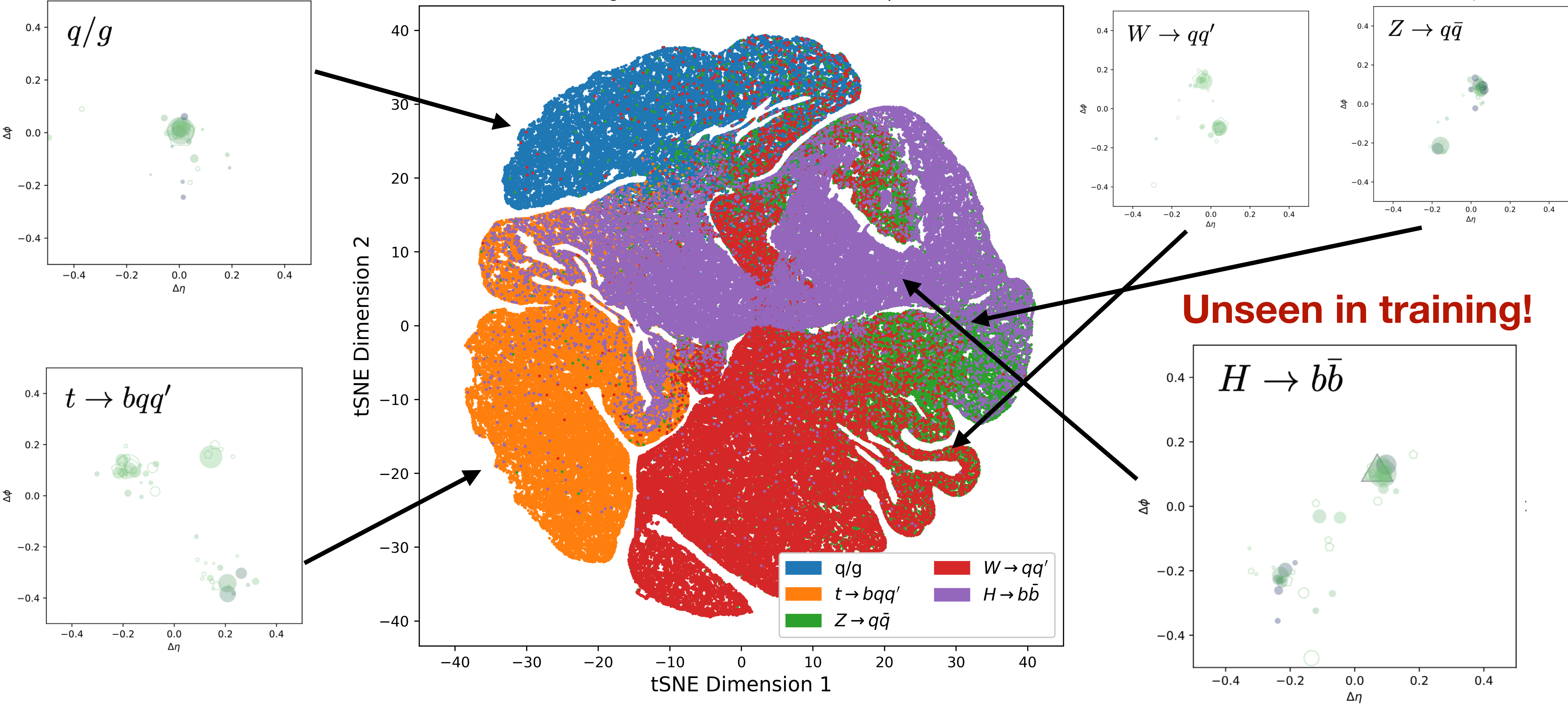
The learned space



The learned space



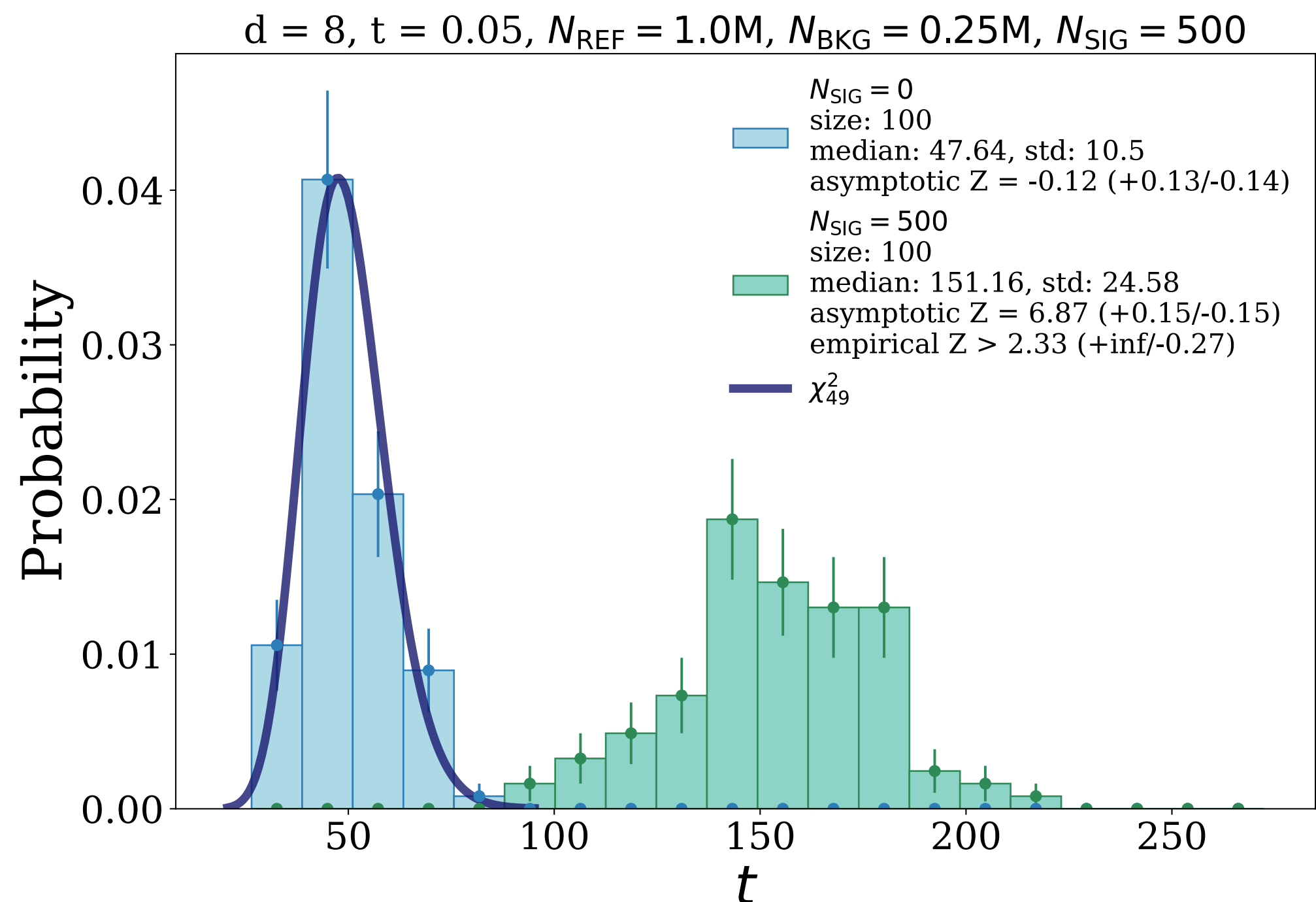
t-SNE embedding of 8-dimensional contrastive space, $T = 0.05$



“Discovering” the Higgs

Can we use this space to “discover” $H \rightarrow b\bar{b}$?

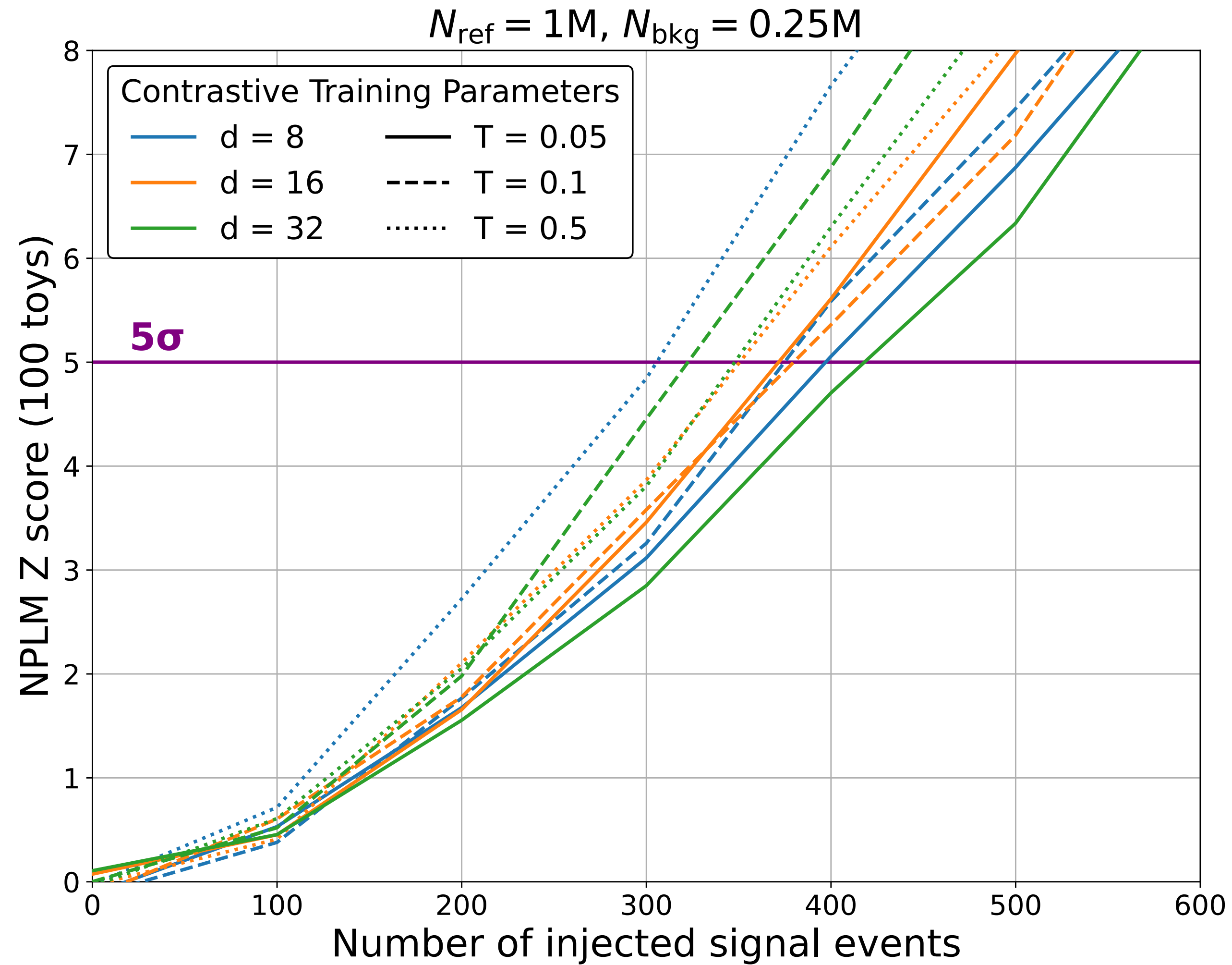
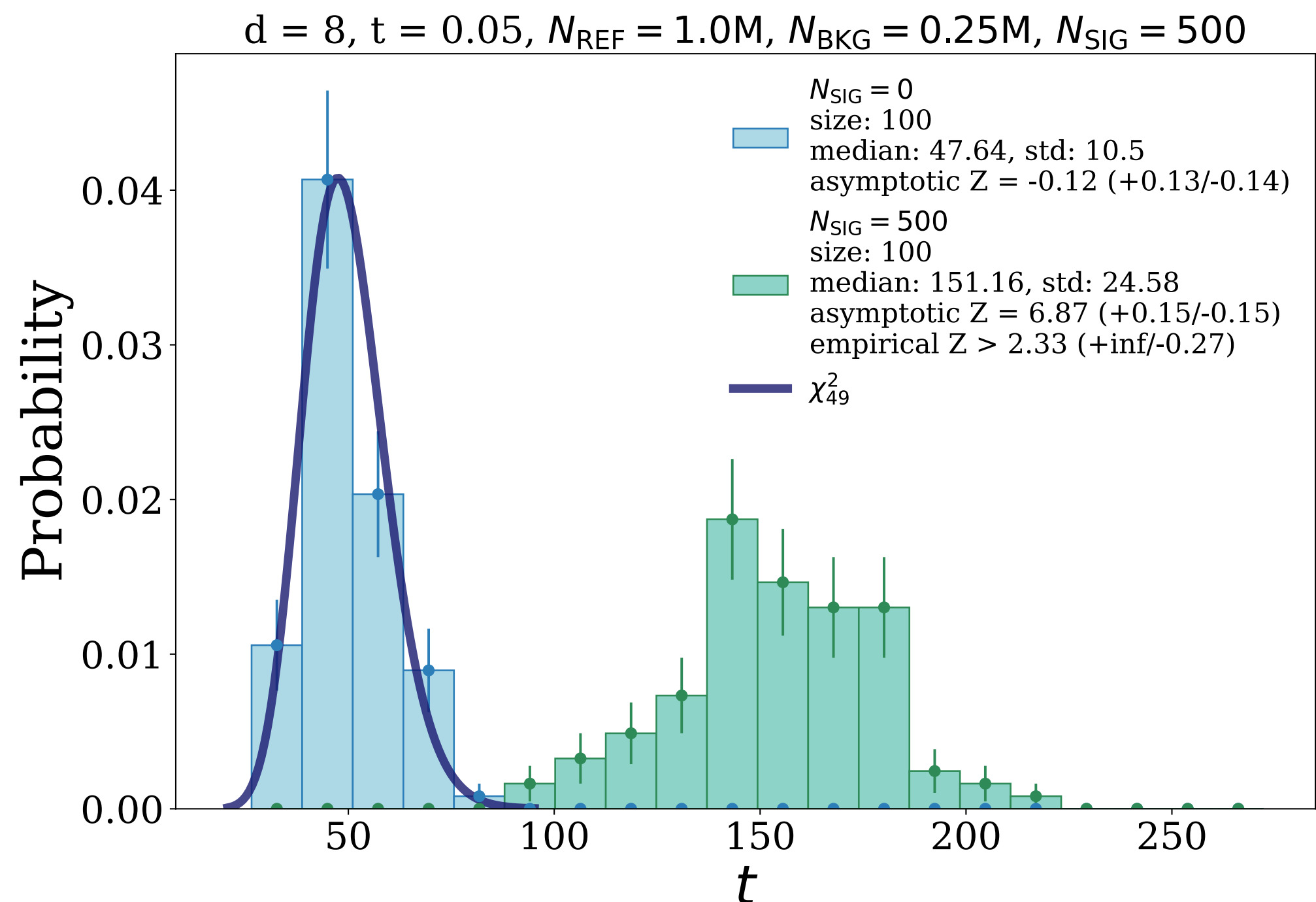
- ▶ Create “reference” datasets using a combination of QCD, W, Z, and top quark backgrounds
- ▶ Inject increasing numbers of $H \rightarrow b\bar{b}$ jets into “test” datasets with the same composition as the reference



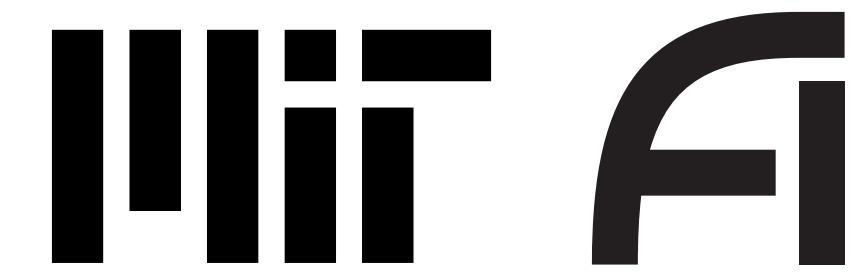
“Discovering” the Higgs

Can we use this space to “discover” $H \rightarrow b\bar{b}$?

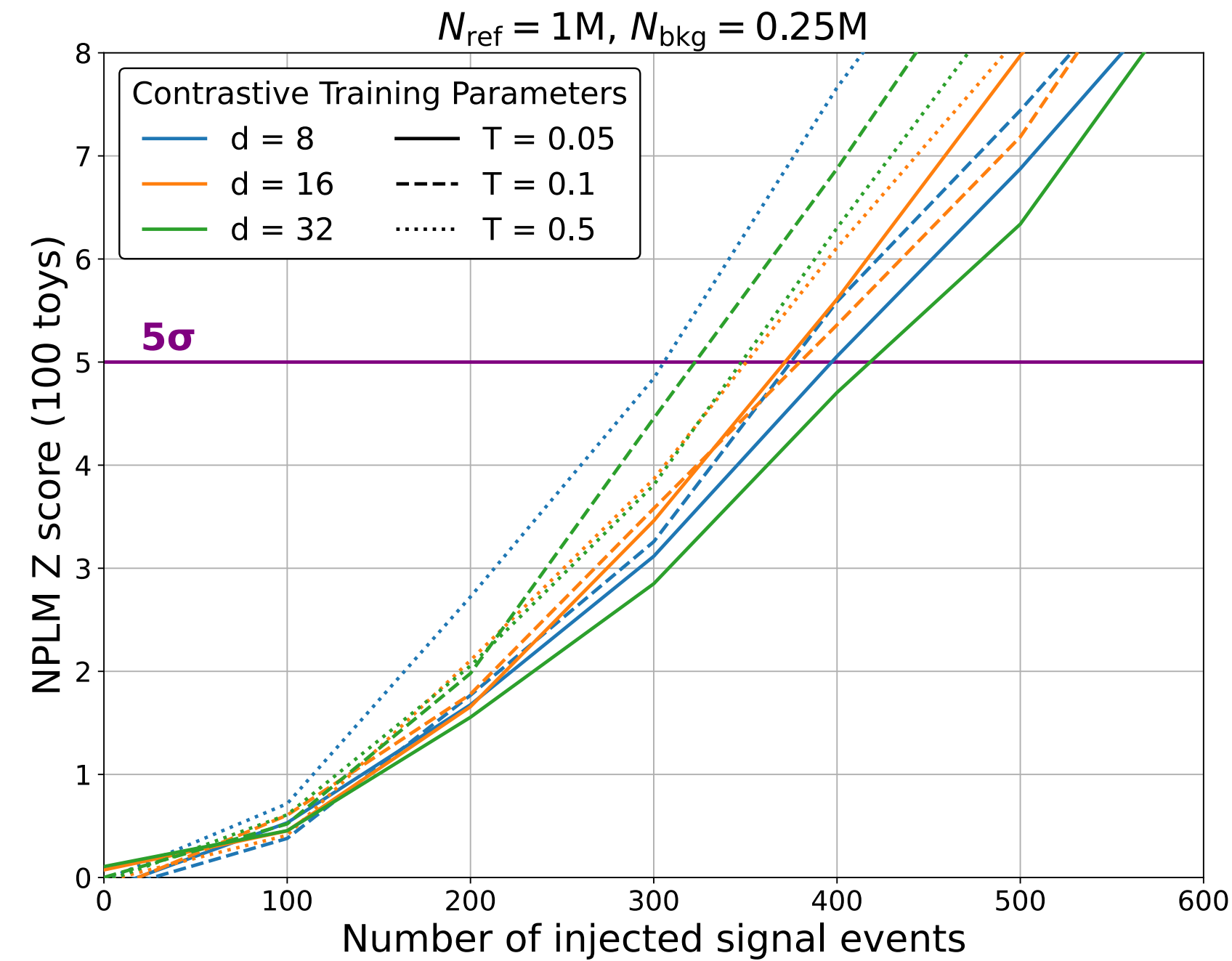
- ▶ Create “reference” datasets using a combination of QCD, W, Z, and top quark backgrounds
- ▶ Inject increasing numbers of $H \rightarrow b\bar{b}$ jets into “test” datasets with the same composition as the reference



How does this measure up?



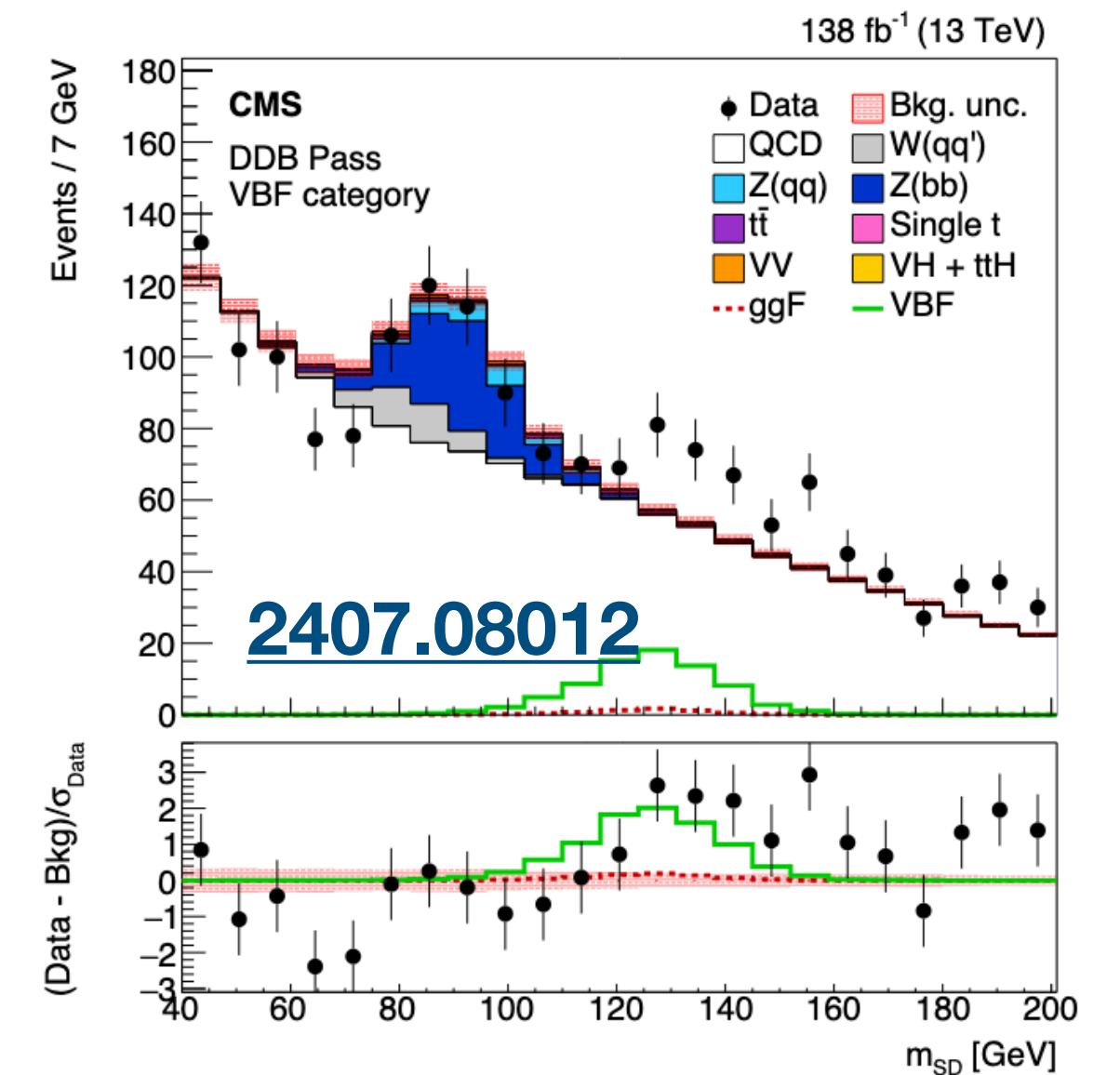
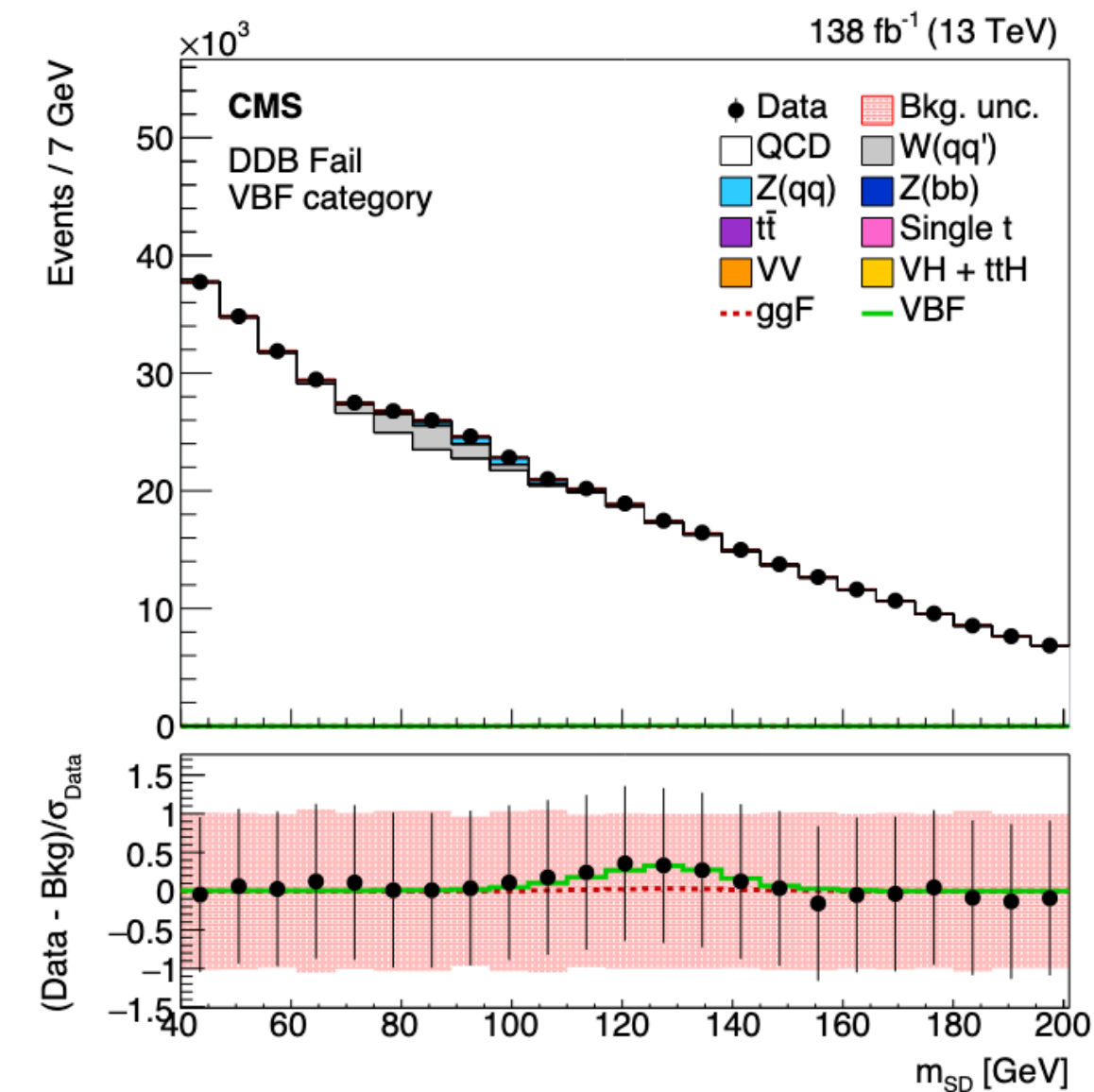
Anomaly detection w/ contrastive space



Assuming $Z \propto \text{lumi}$, we'd expect 4σ at 0.0056% signal-to-bkg for CMS Run 2 luminosity

One contrastive training + NPLM

Dedicated CMS search



Observed $H \rightarrow b\bar{b}$ in ggF + VBF at 4σ on CMS Run 2 dataset, estimated 0.0053% signal-to-bkg

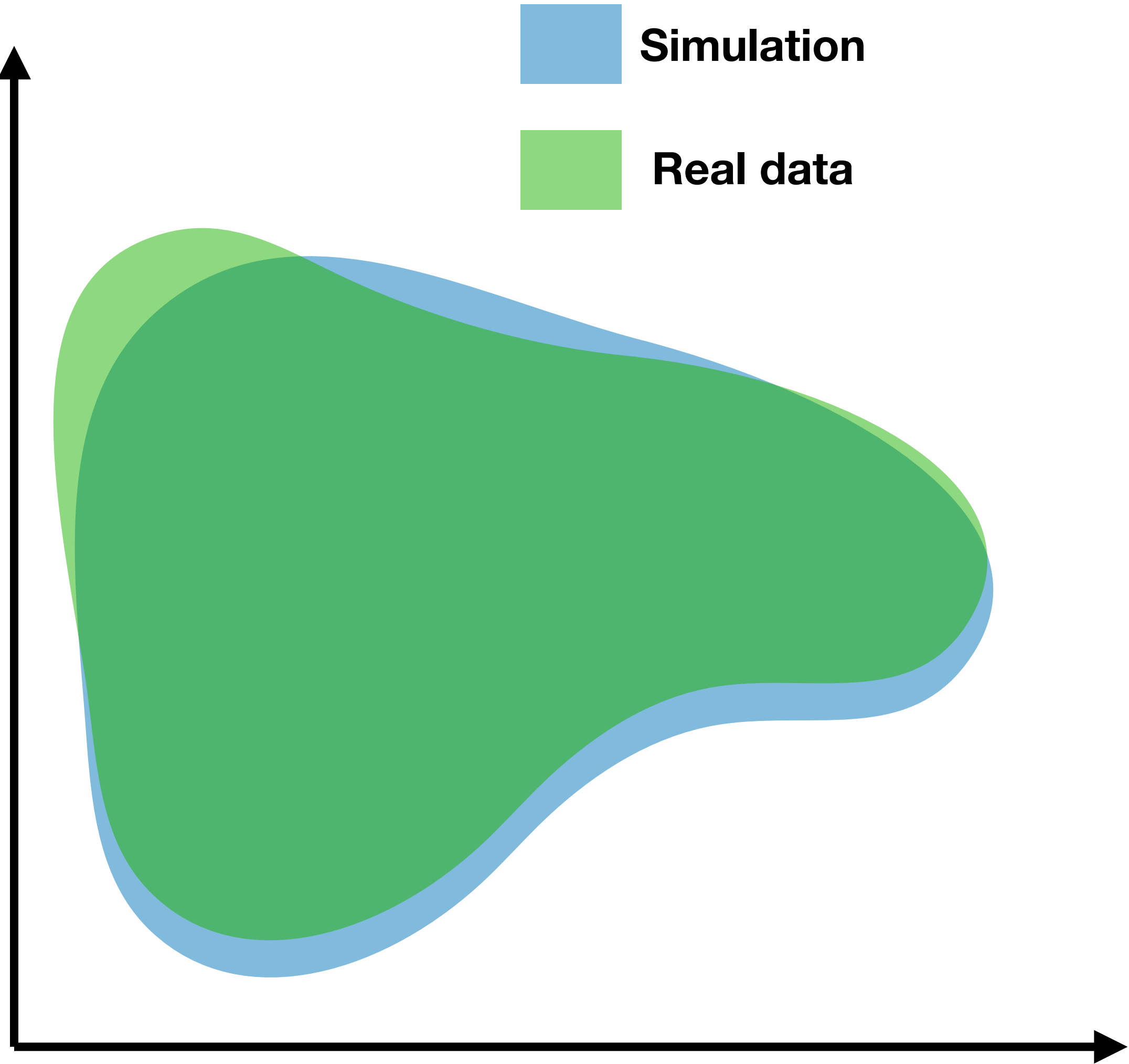
Dedicated search: hand-tuned selection, ggF/VBF categorization, dedicated jet tagger, etc.

NB: these two setups are only very approximately comparable!

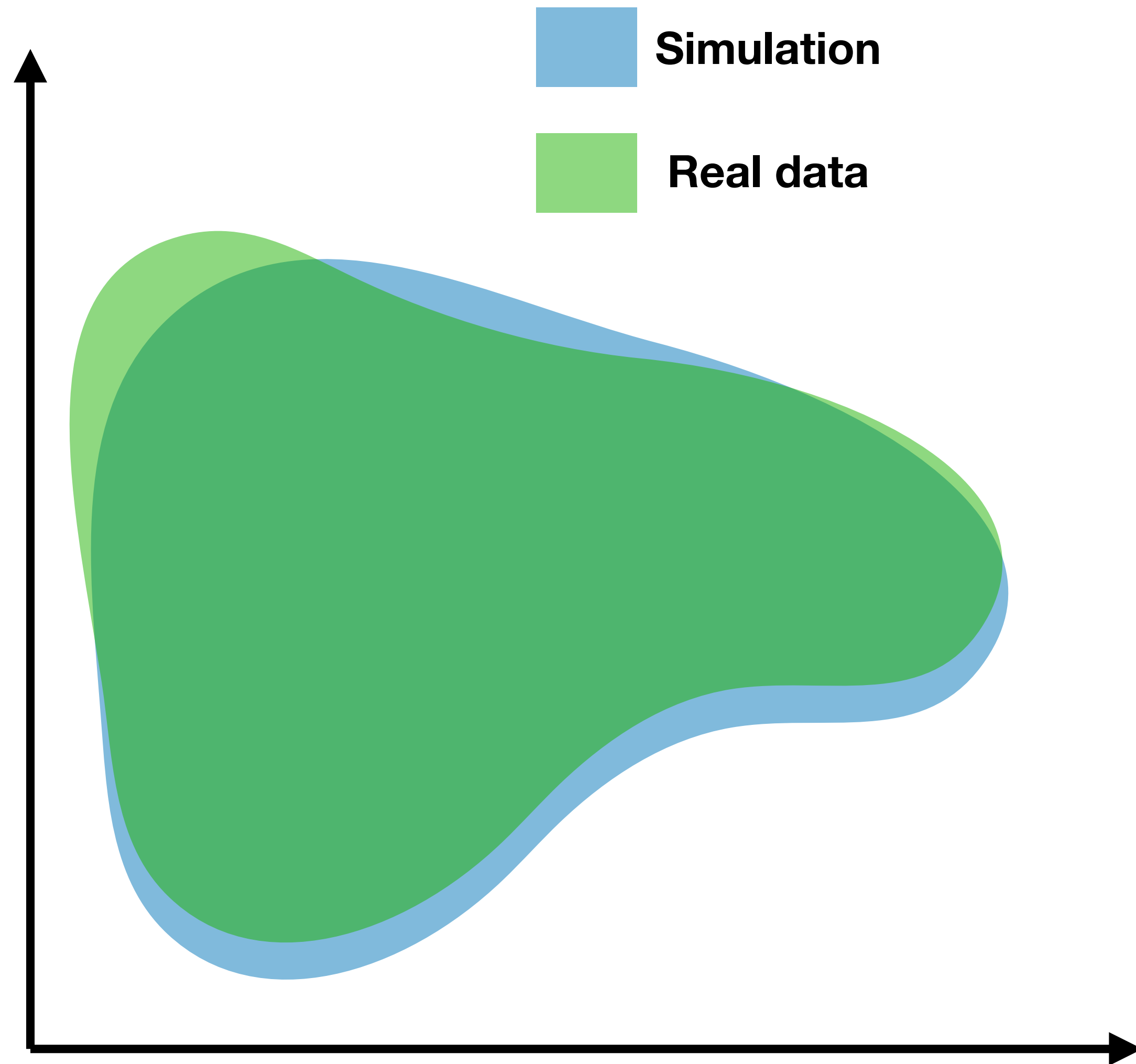
Robustness

Simulations are not to be trusted

A problematic possibility

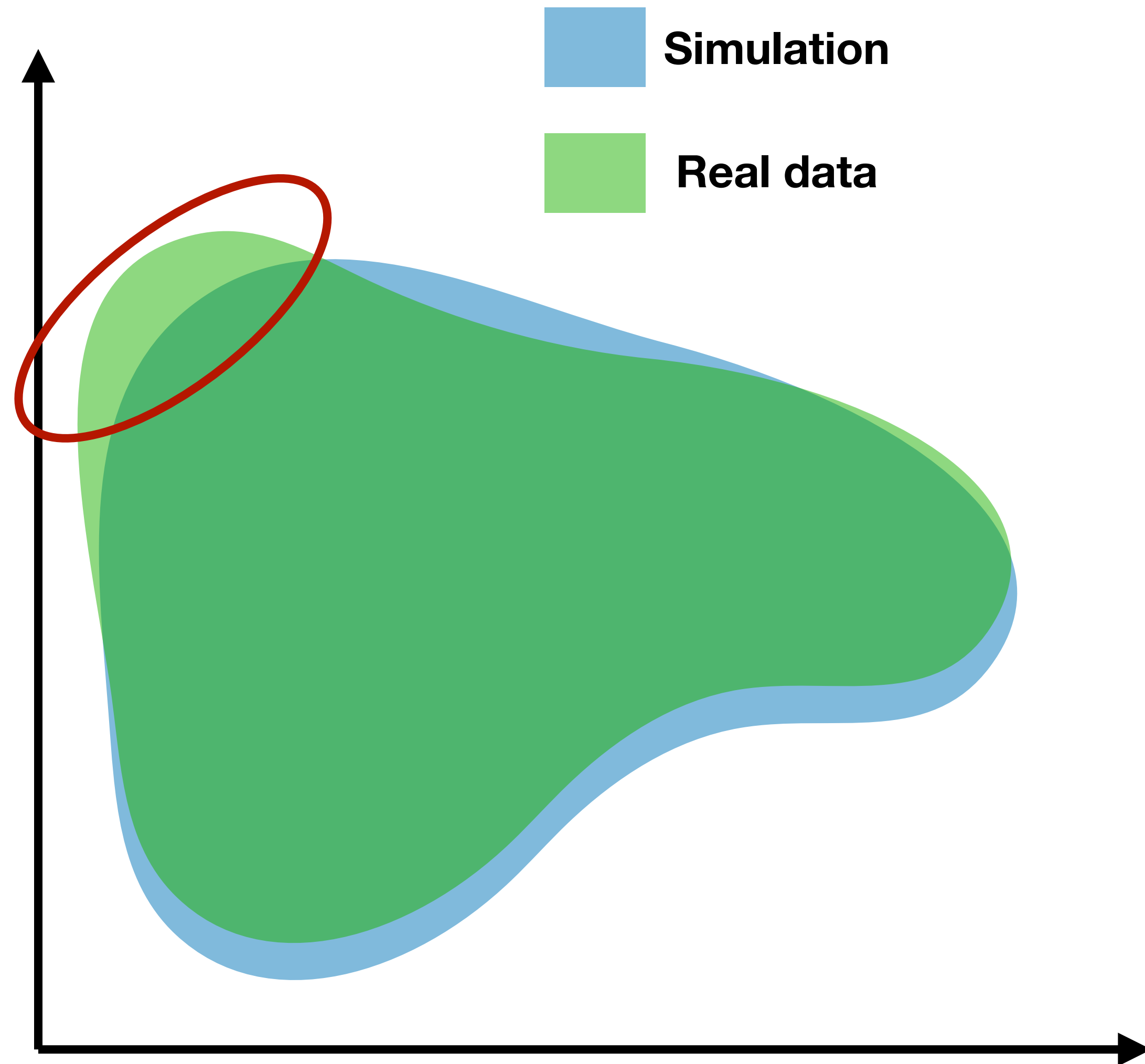


A problematic possibility



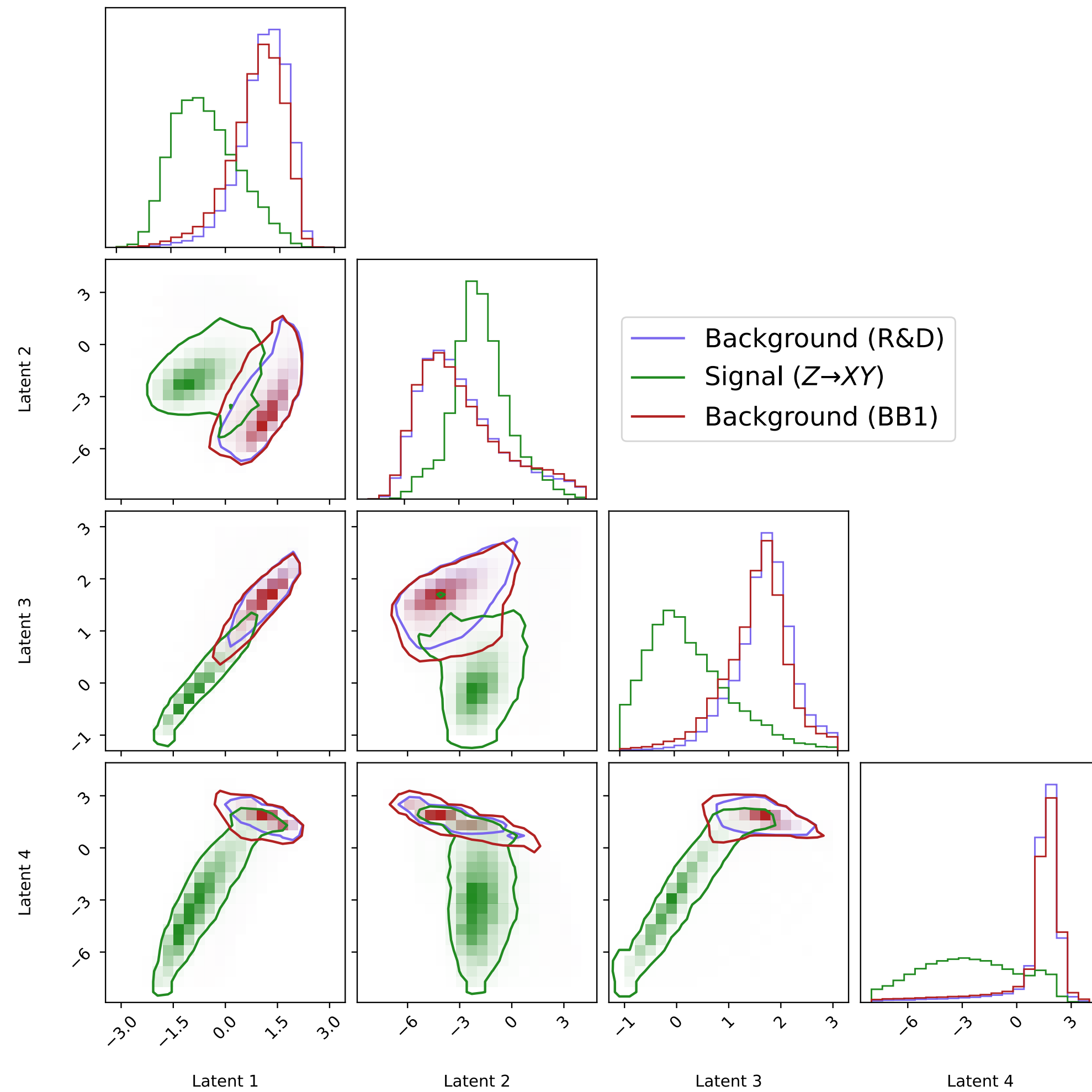
- Easiest to train an embedding on simulation — **what if our simulations are wrong?**
 - ▶ Even subtle/small systematic differences can have an impact
 - ▶ Needs to be accounted for; might reduce the sensitivity of downstream tasks (e.g. NPLM)

A problematic possibility



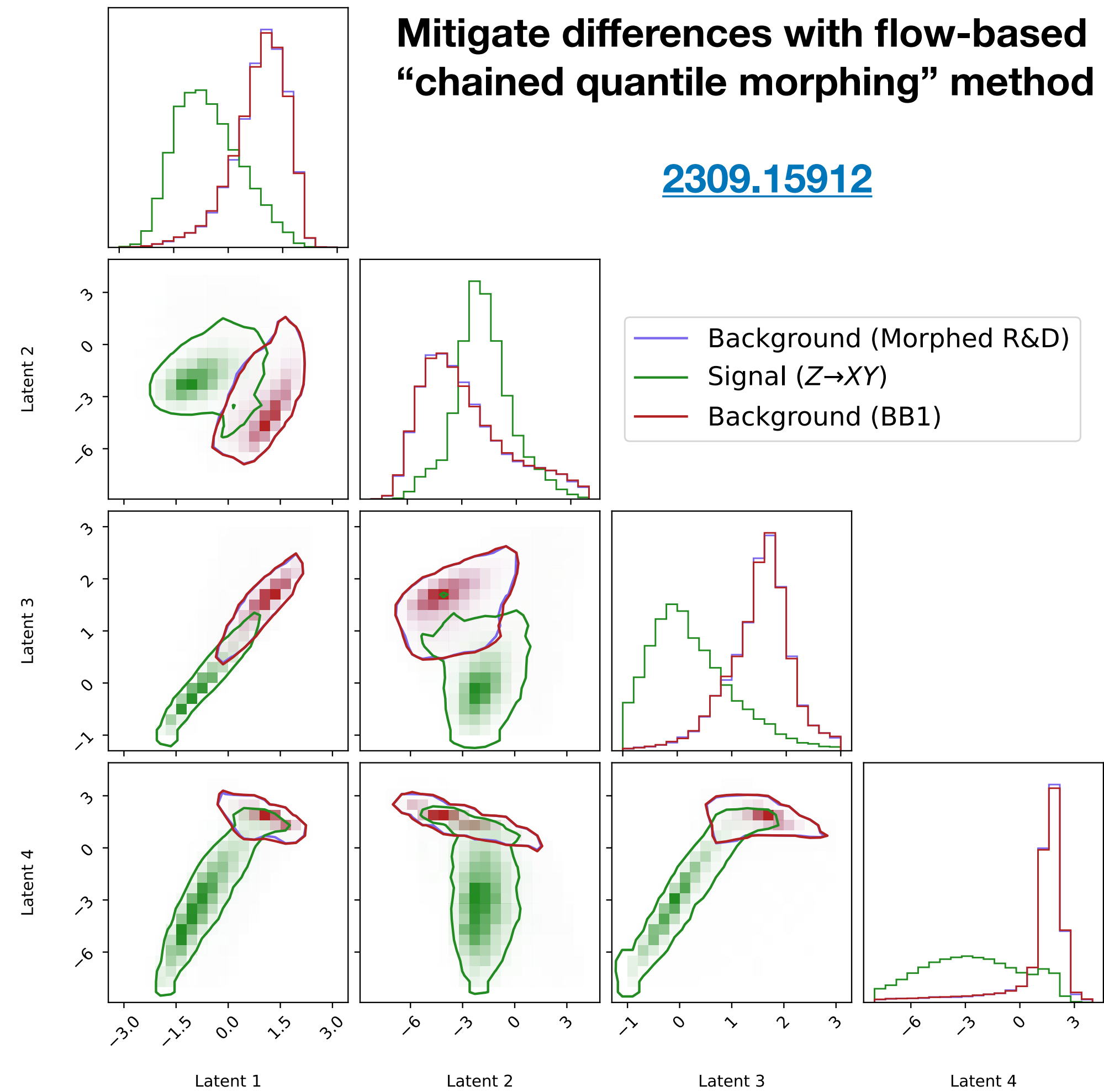
- Easiest to train an embedding on simulation — **what if our simulations are wrong?**
 - ▶ Even subtle/small systematic differences can have an impact
 - ▶ Needs to be accounted for; might reduce the sensitivity of downstream tasks (e.g. NPLM)
- One option is **reweighting** simulation
 - ▶ $w(\mathbf{x}) = p_D(\mathbf{x})/p_{\text{sim}}(\mathbf{x})$
 - ▶ Works well enough, but **cannot account for discrepancies in the tail!**

Morphing a contrastive space



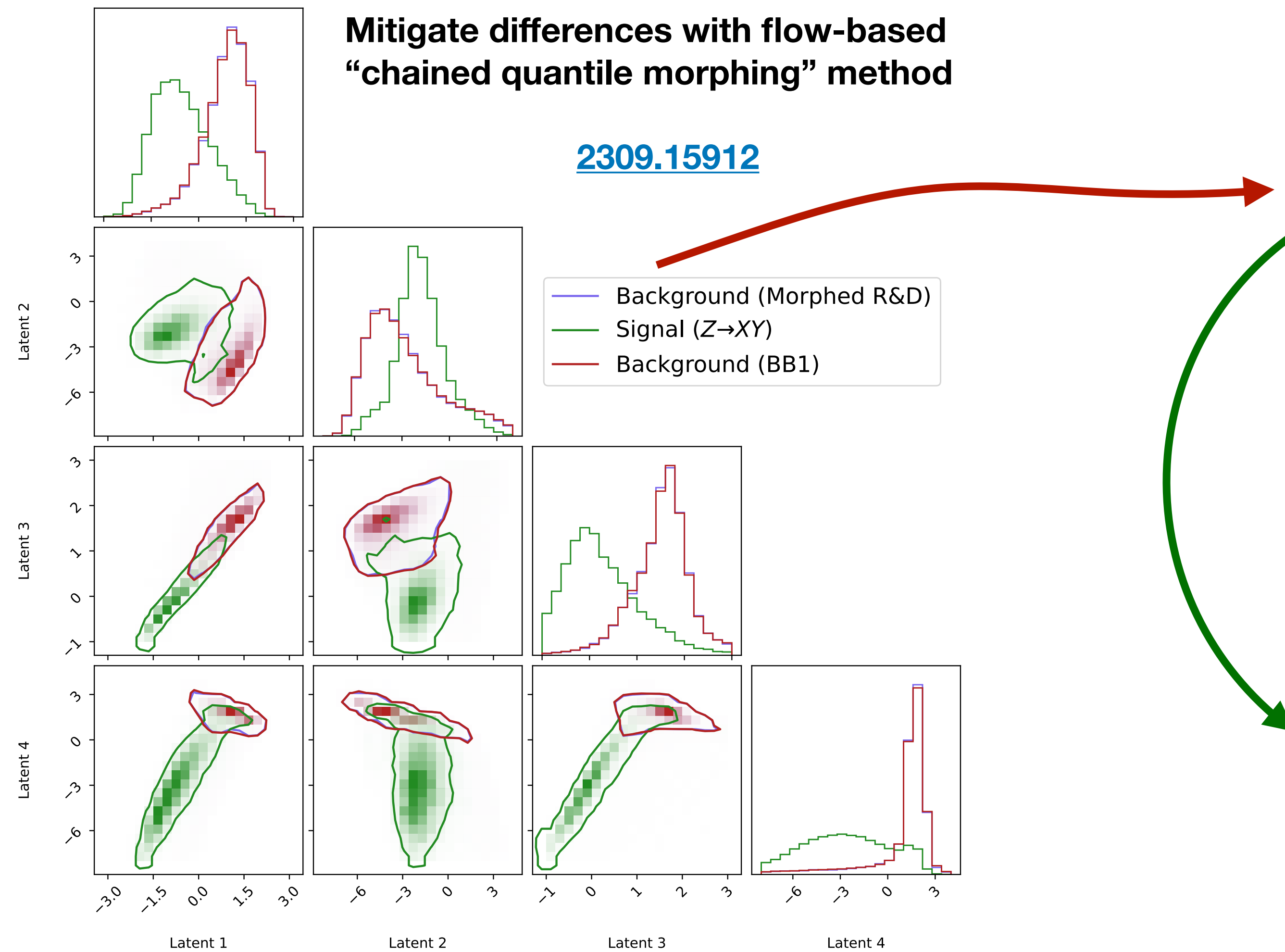
Contrastive space trained on “MC” treats “data” differently — discrepancies in learned space

Morphing a contrastive space

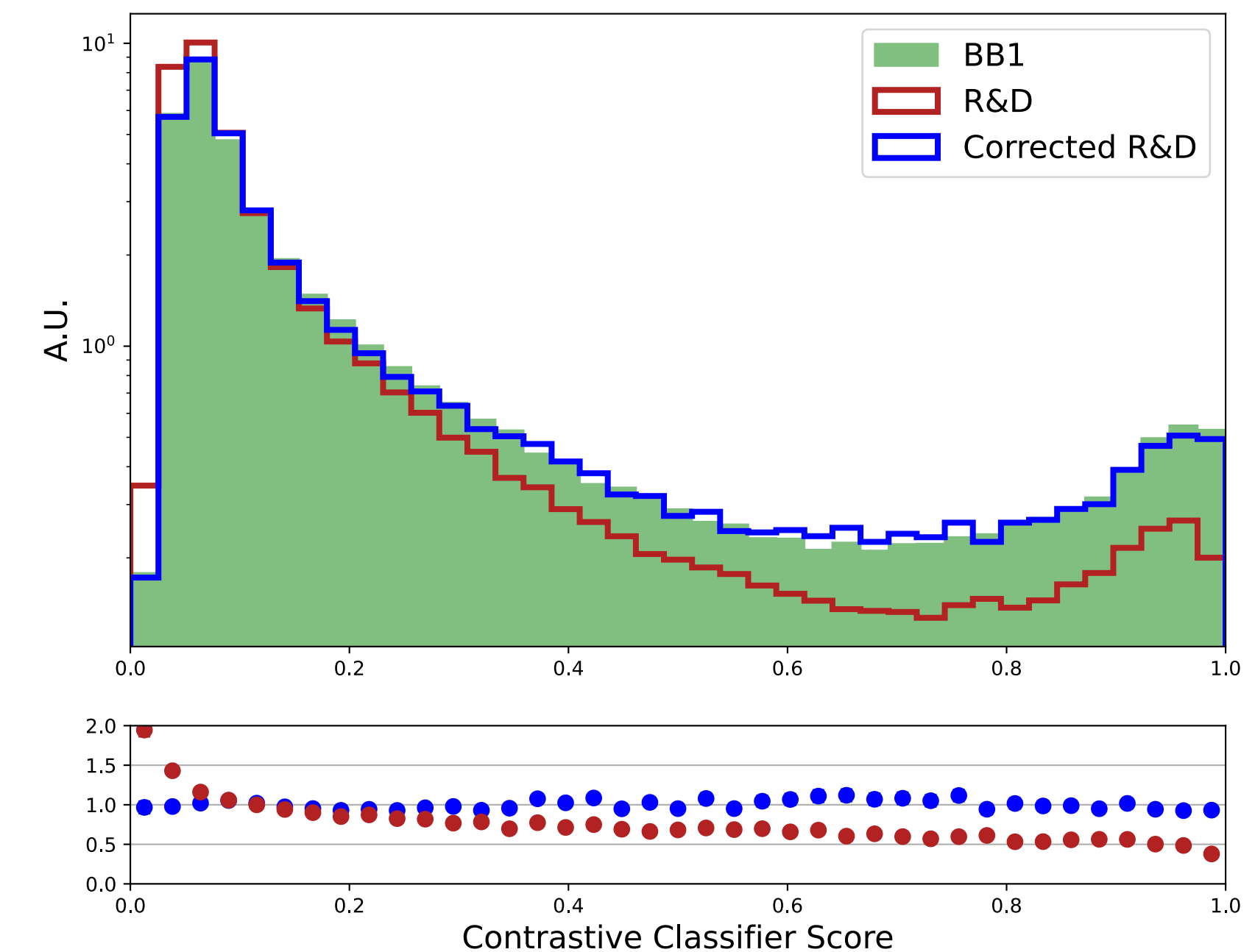


Contrastive space trained on “MC” treats “data” differently — discrepancies in learned space

Morphing a contrastive space



Classifier	Sample	AUC
ParticleNet Tagger	R&D	0.954
	BB1	0.931
ParticleNet Latent Tagger	R&D	0.952
	BB1	0.928
	Morphed R&D	0.928
High-Level Tagger	R&D	0.910
	BB1	0.872
	Morphed R&D	0.873

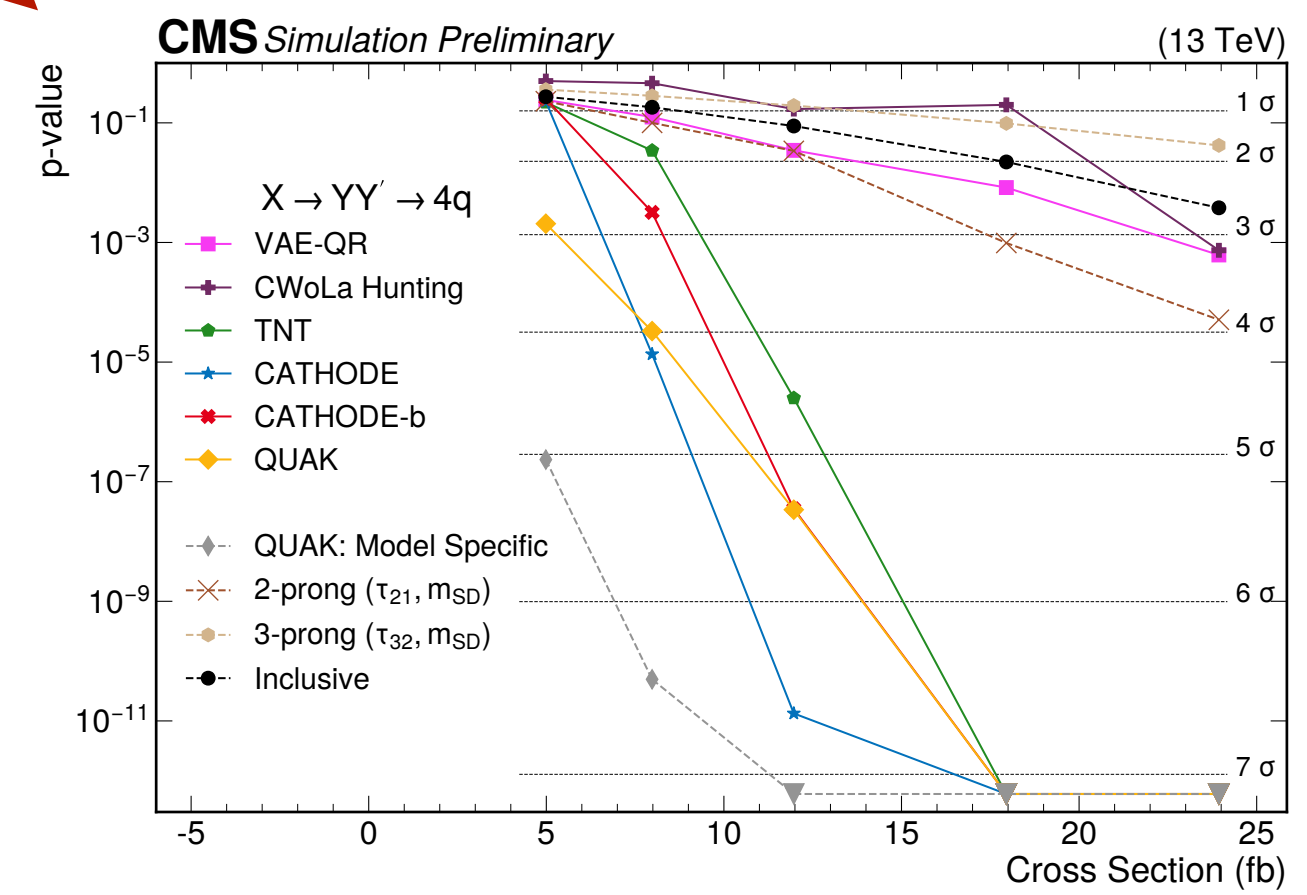
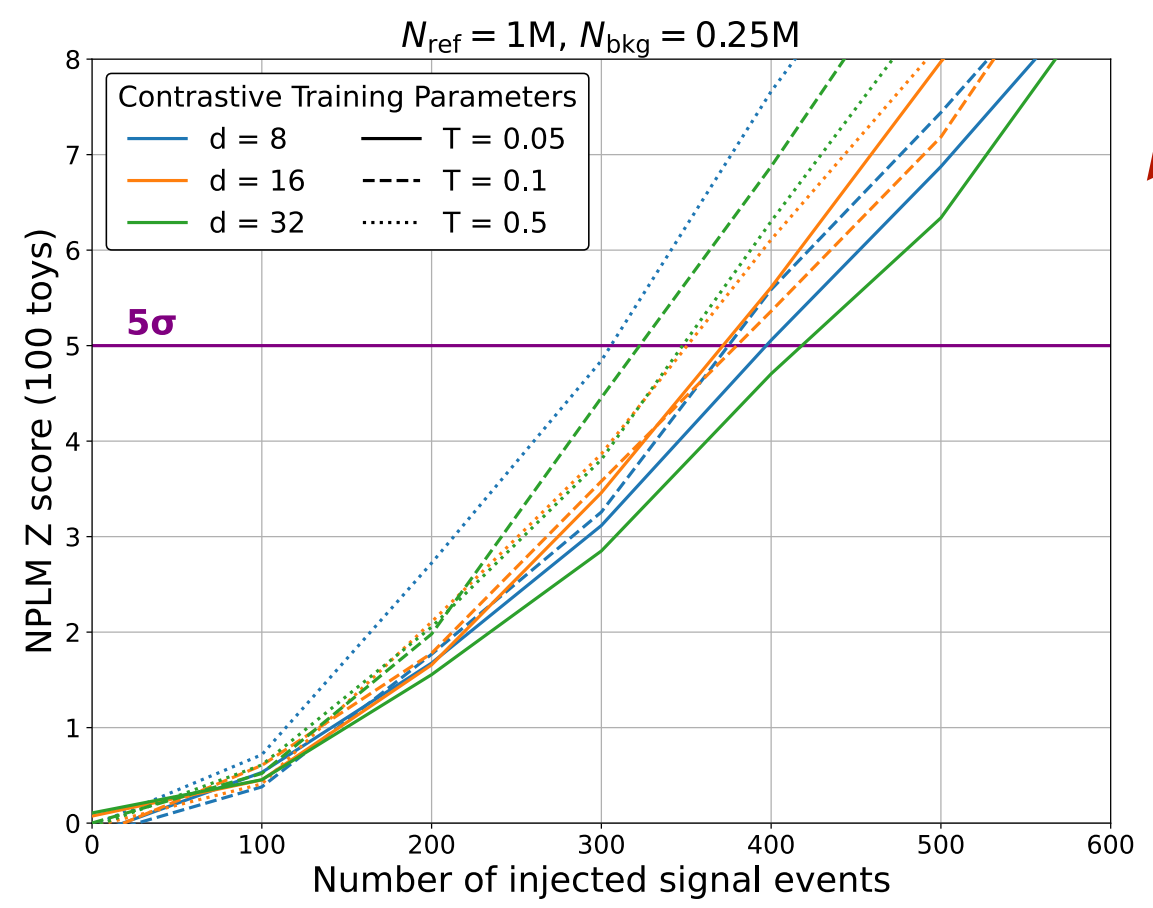
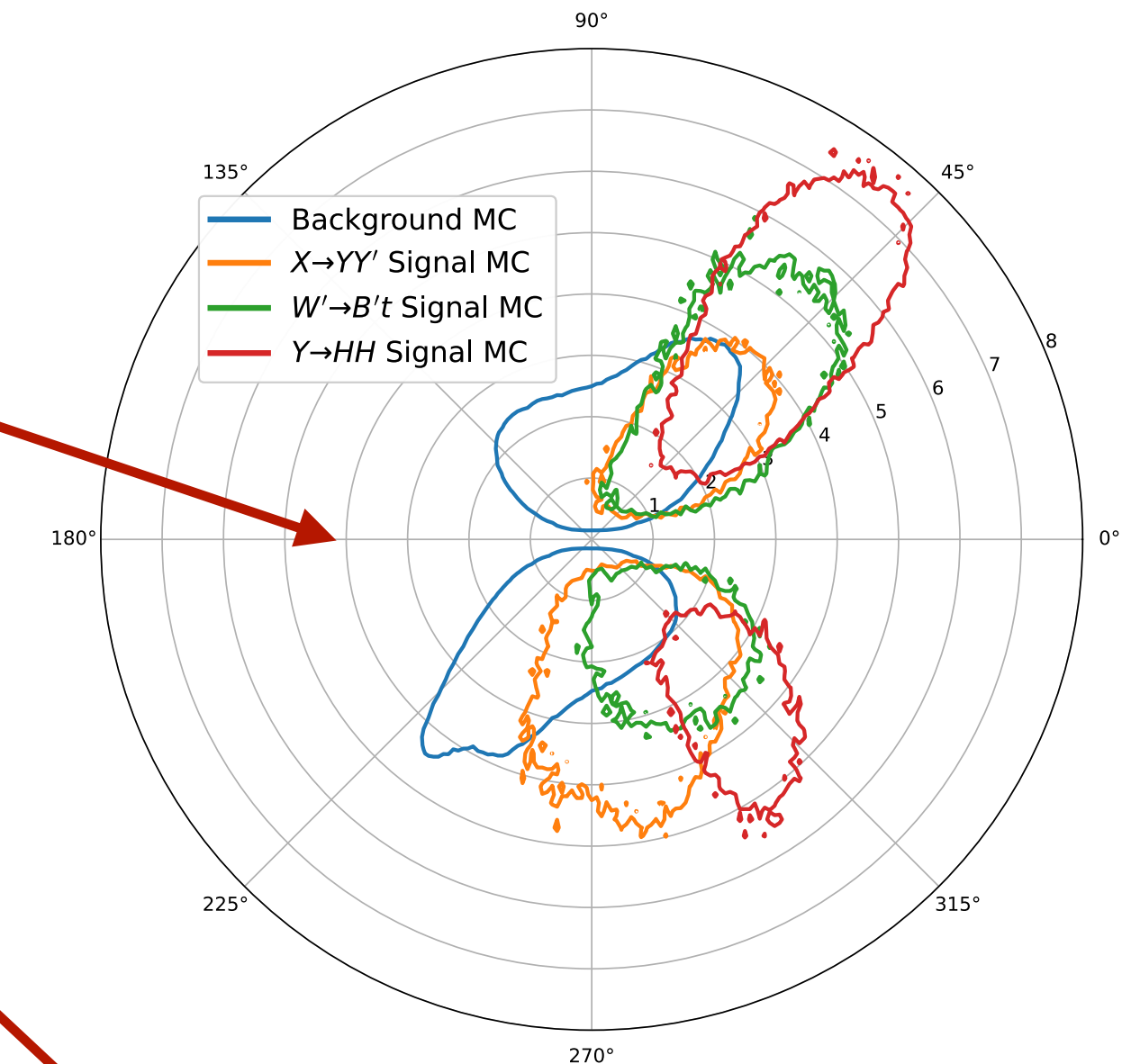
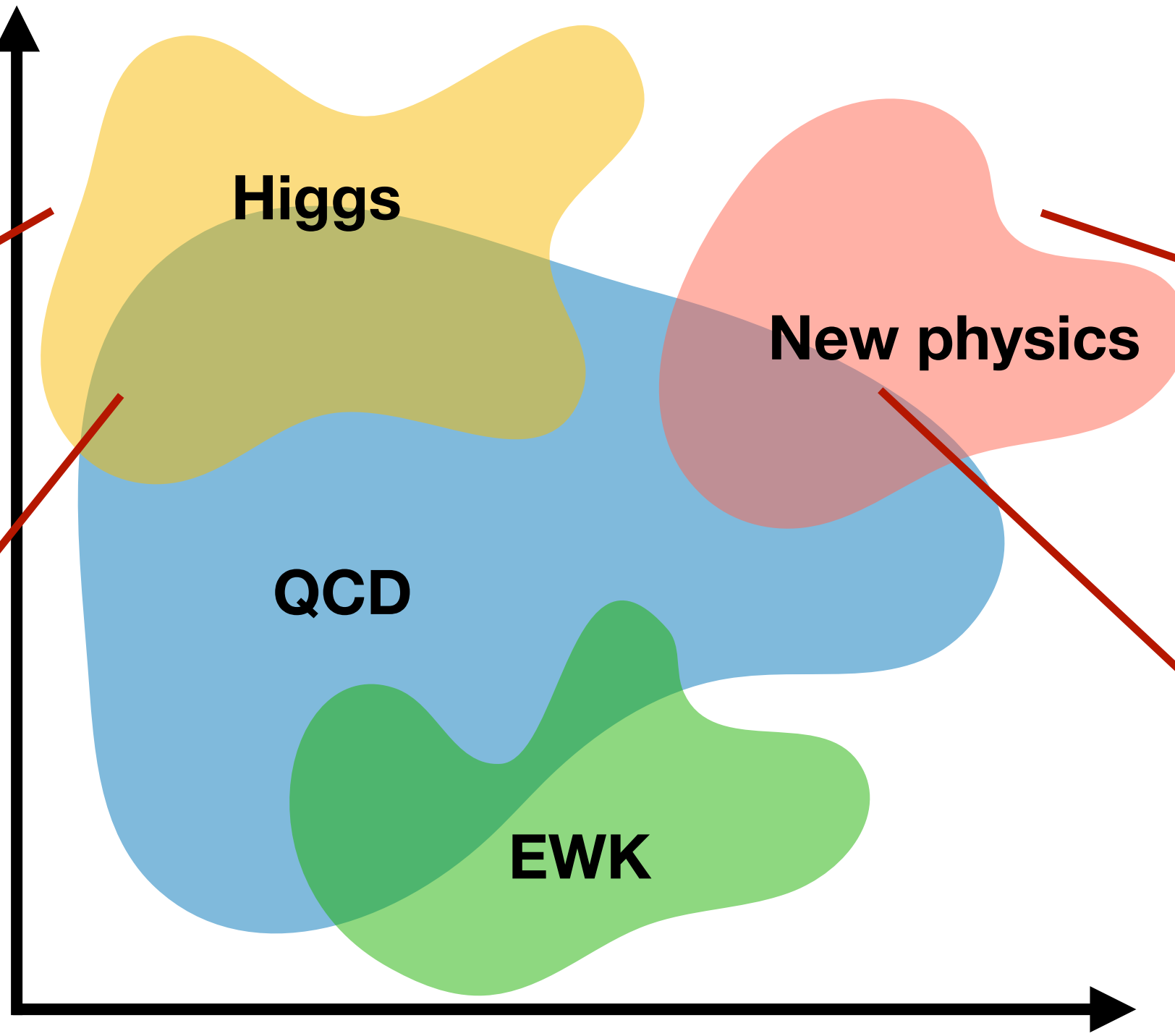
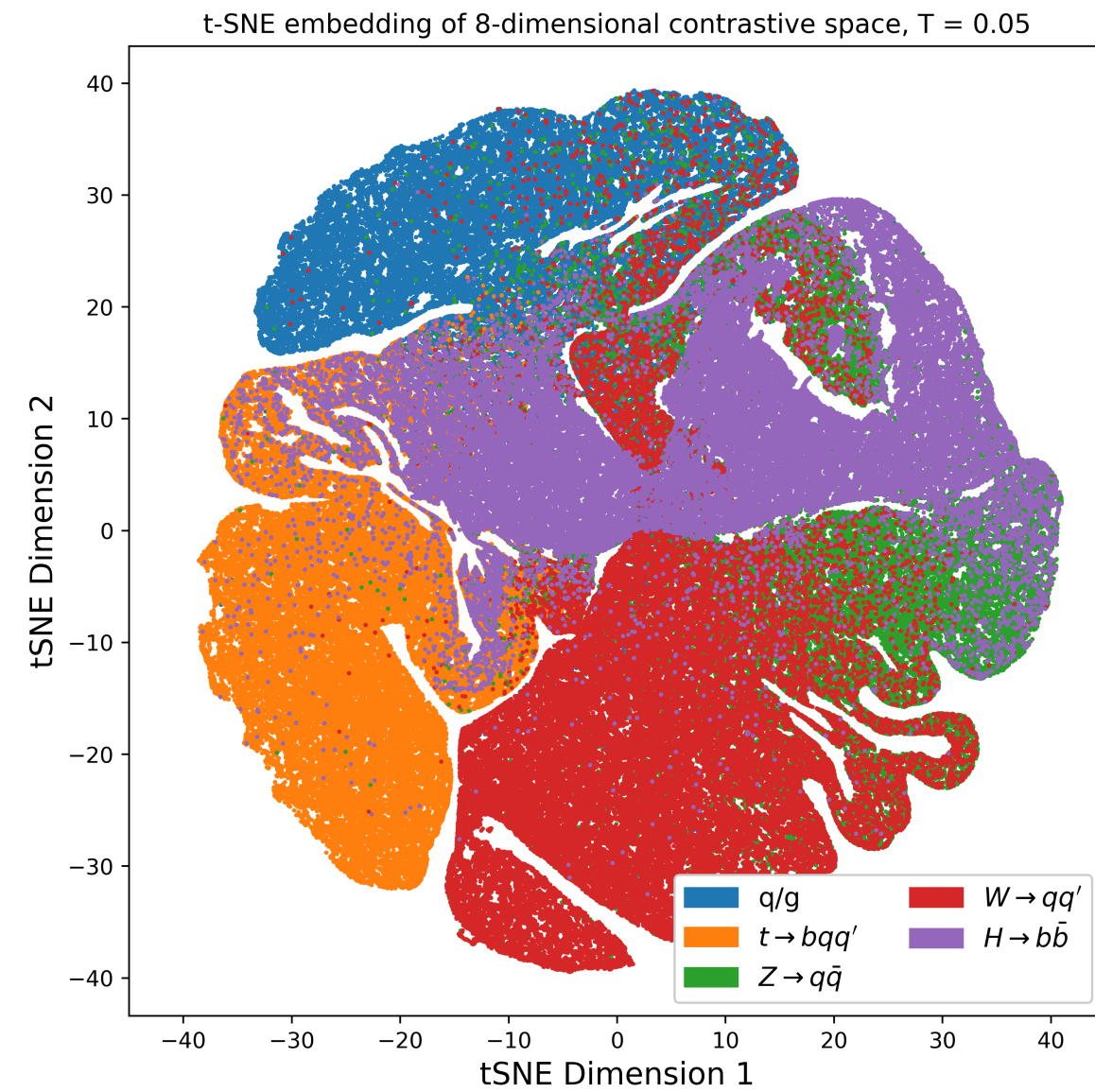


Contrastive space trained on “MC” treats “data” differently — discrepancies in learned space

The future

Where do we go from here?

One space to rule them all



One space, many downstream tasks!

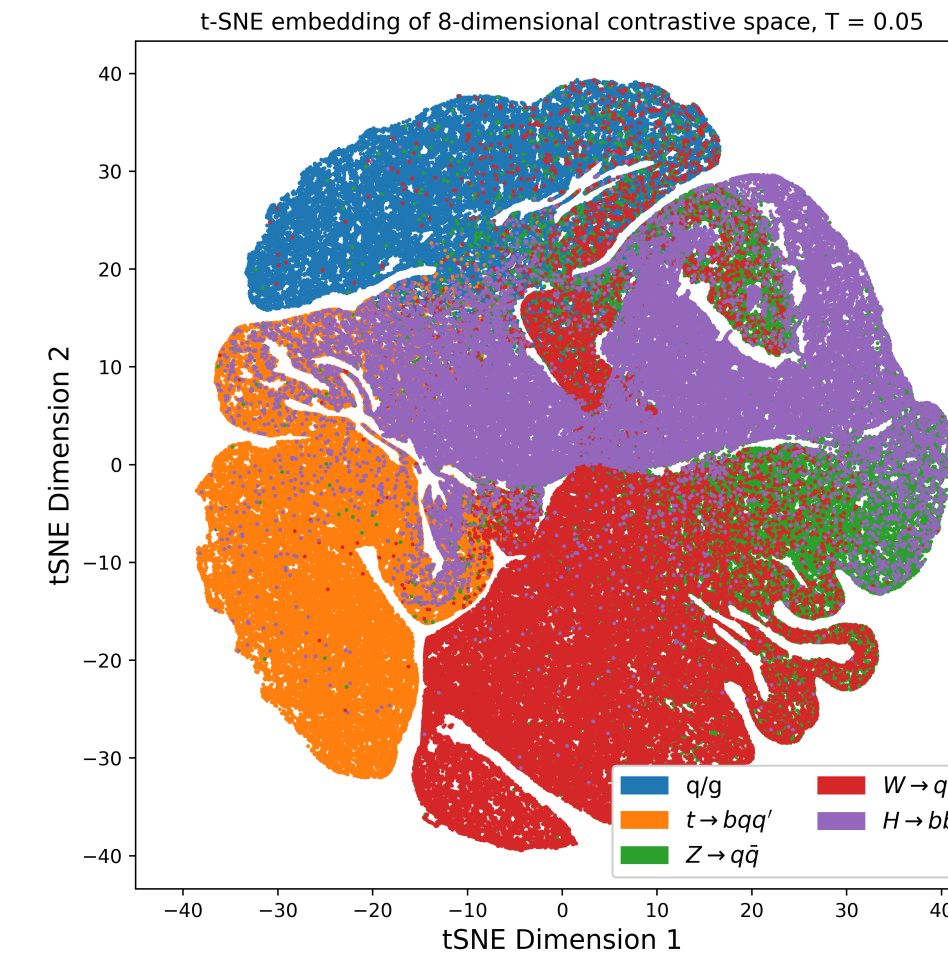
Early evidence that embeddings are a promising tool for collider physics!

Lots of exciting work to come & ongoing!

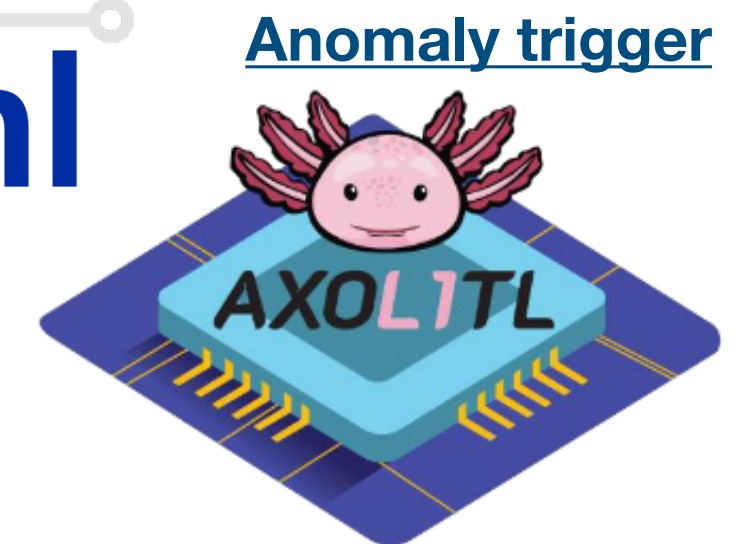
Some future directions

A few things we're hoping to pursue

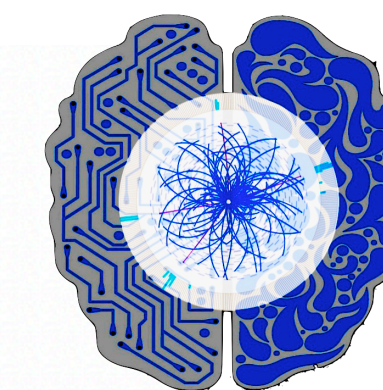
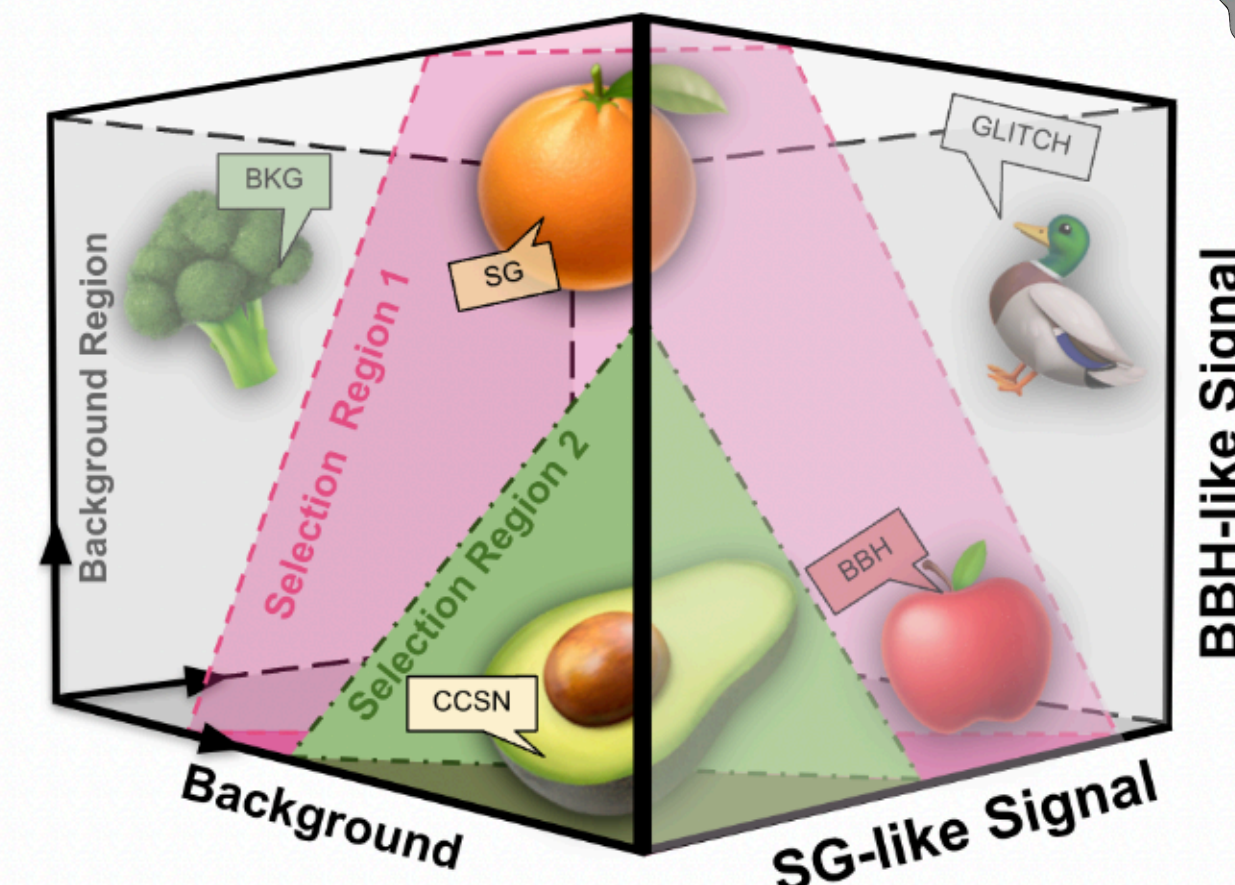
1. Develop “supervised” contrastive embedding into full-scale demonstration for a realistic LHC search/measurement
2. Learn embeddings for a **“Level 1 foundation model”** — isolate anomalies & rare SM in real time!
3. Contrastive space for GWAK v2 at LIGO - online anomaly detection & more



Emerging AI frontier

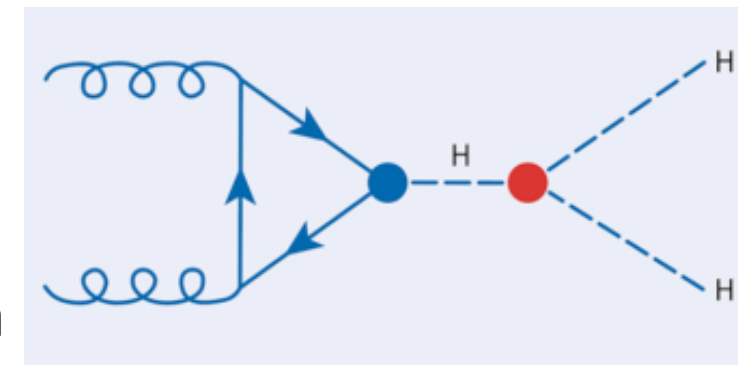
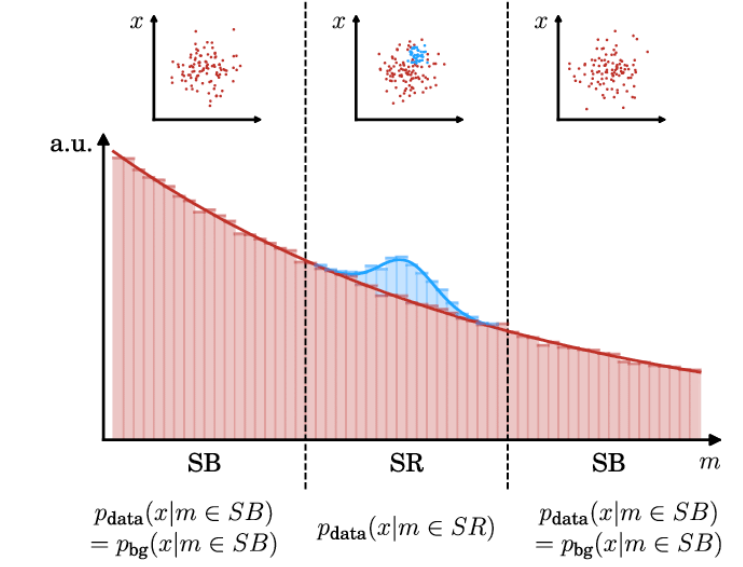
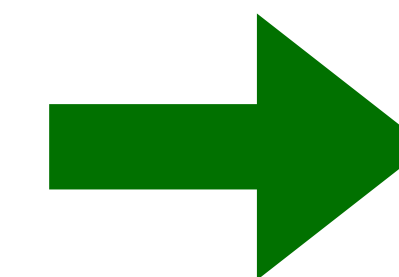
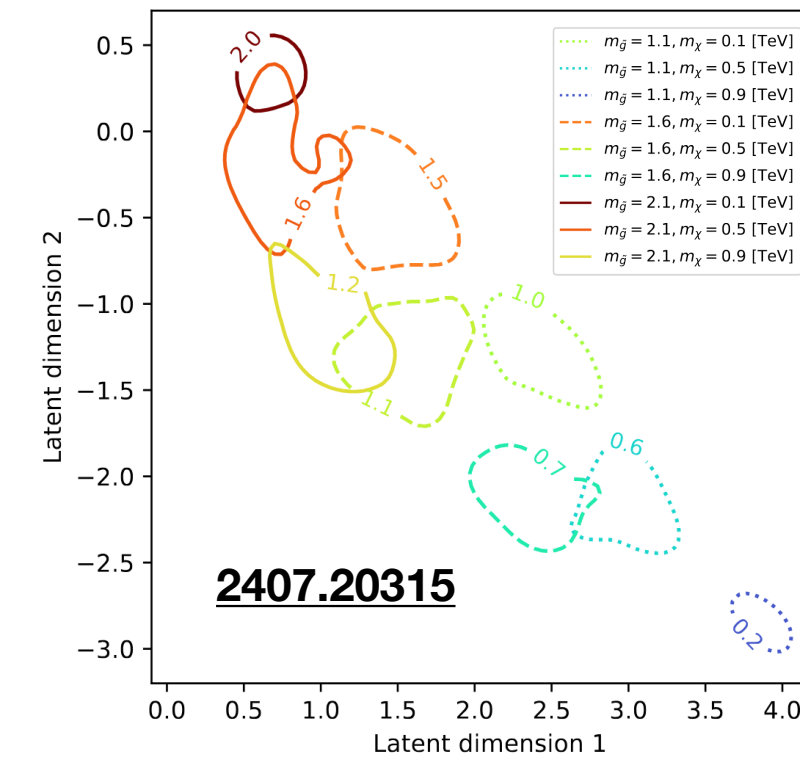
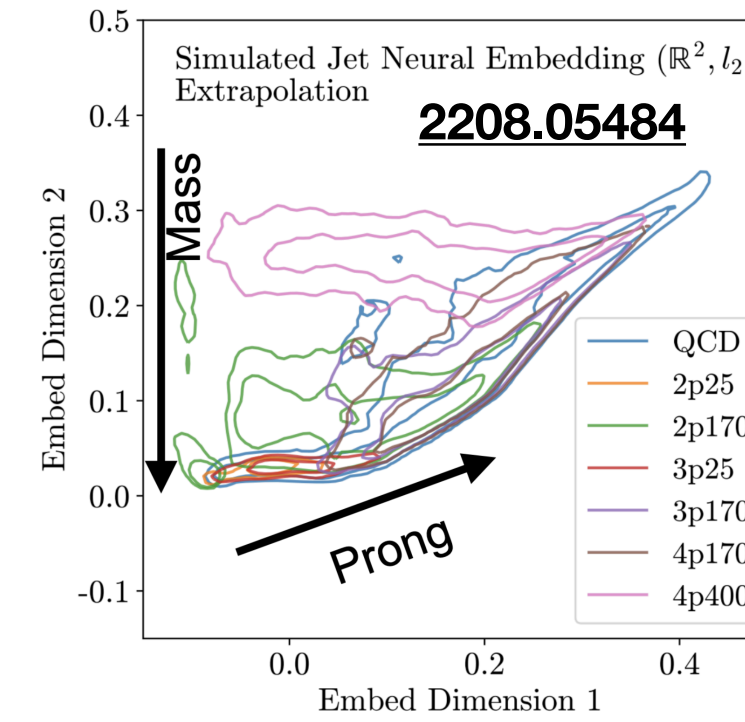
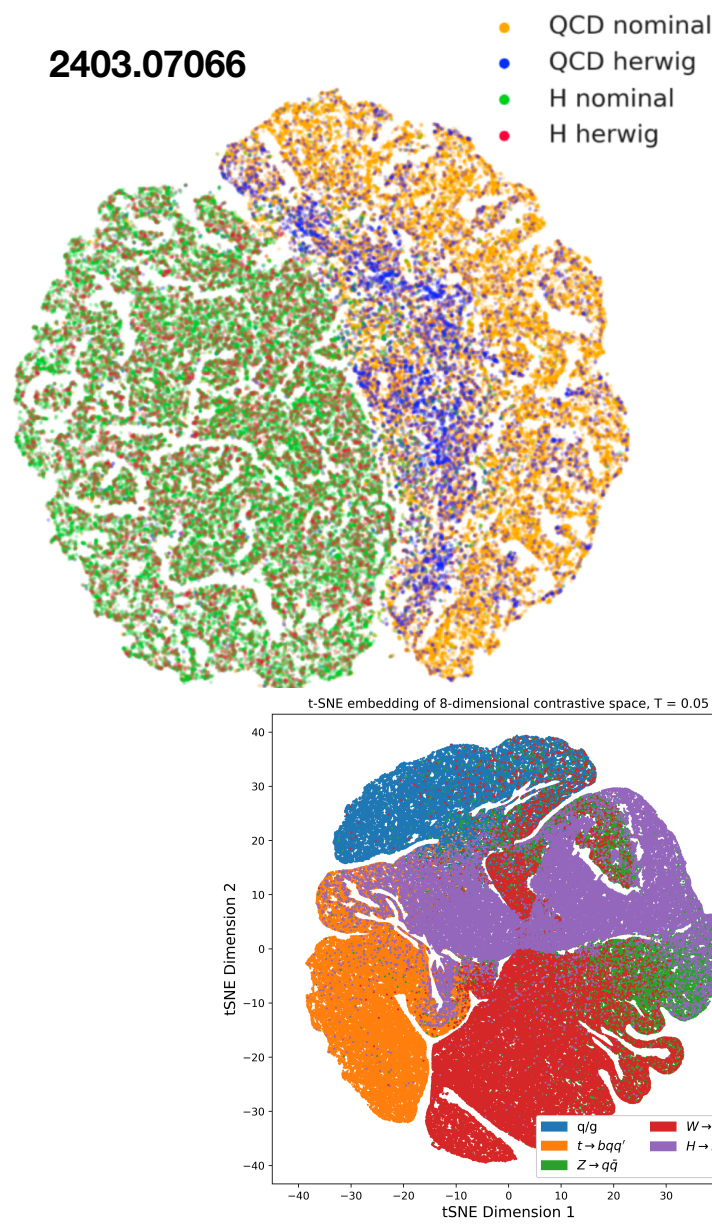
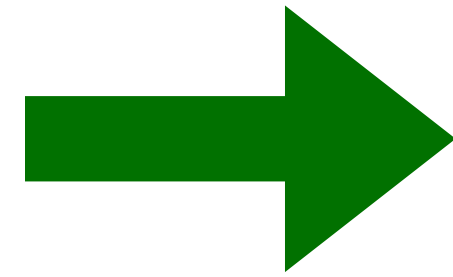
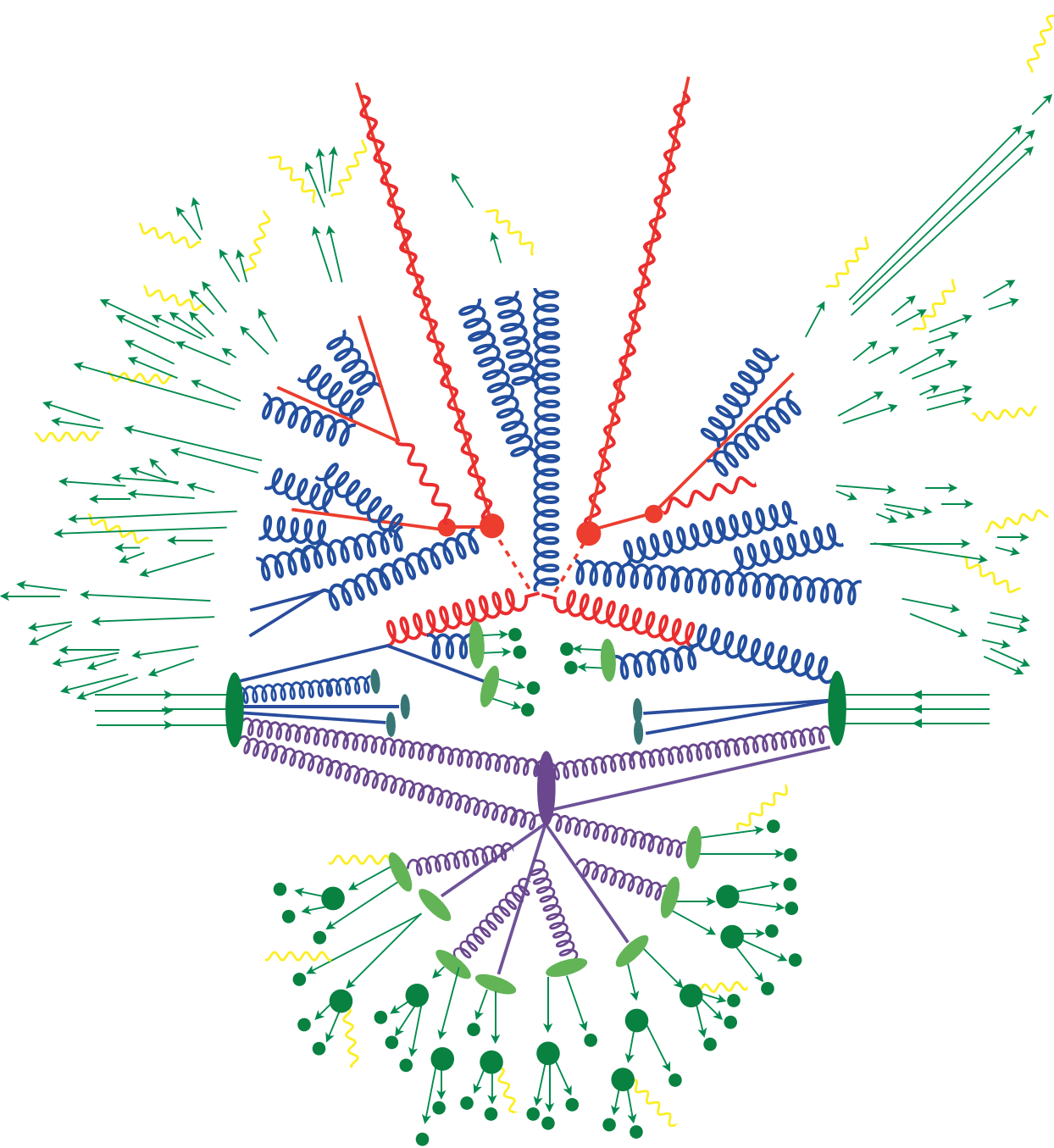


3D GWAK Space



FastML

A new workflow for HEP



Anomaly detection

Precision measurements

BSM physics searches

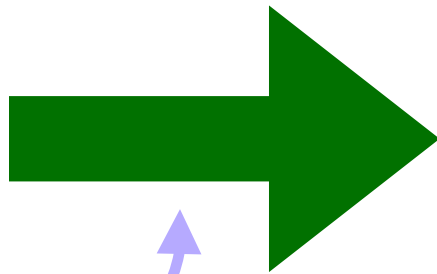
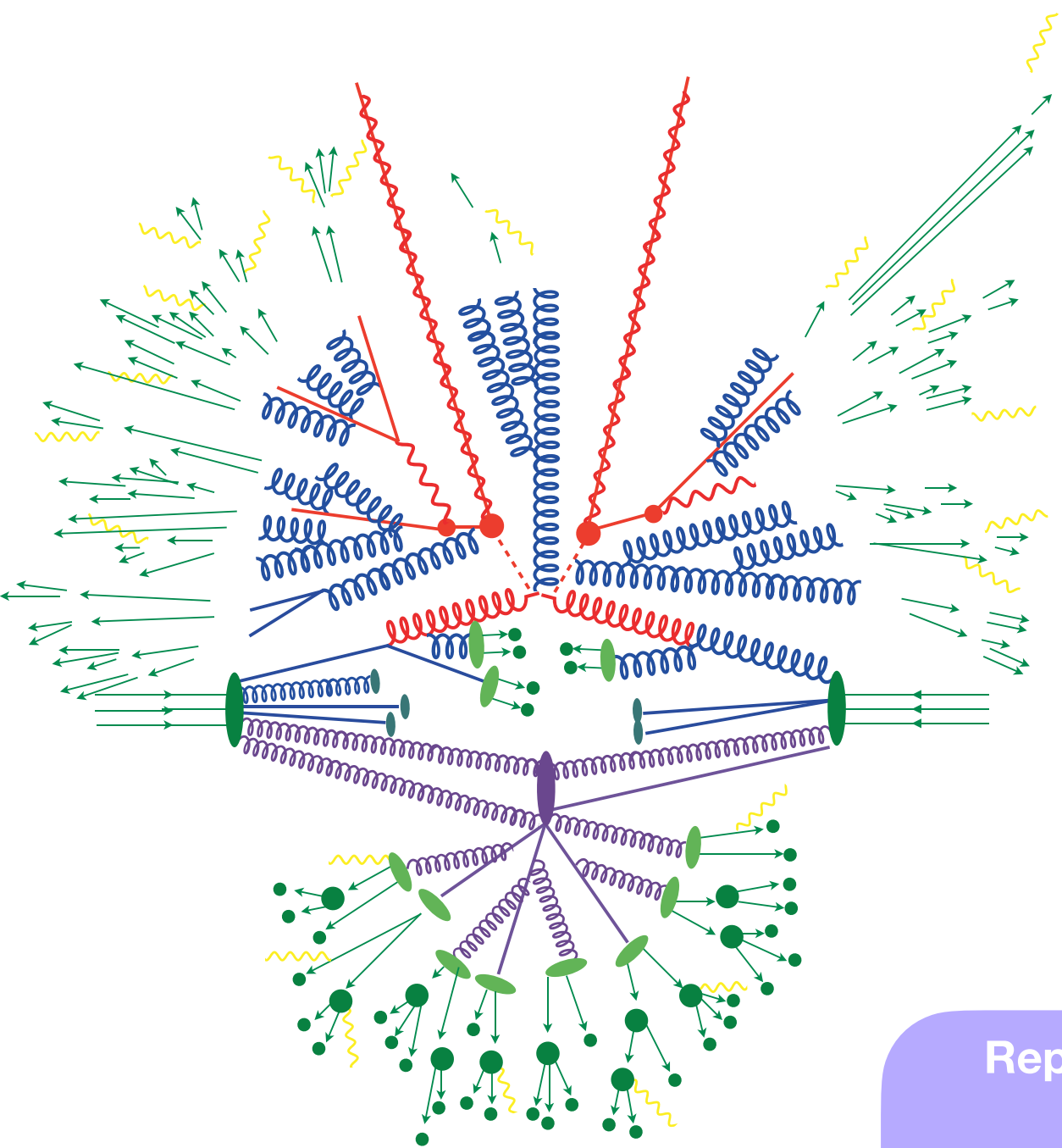
Any analysis we want,

faster & easier

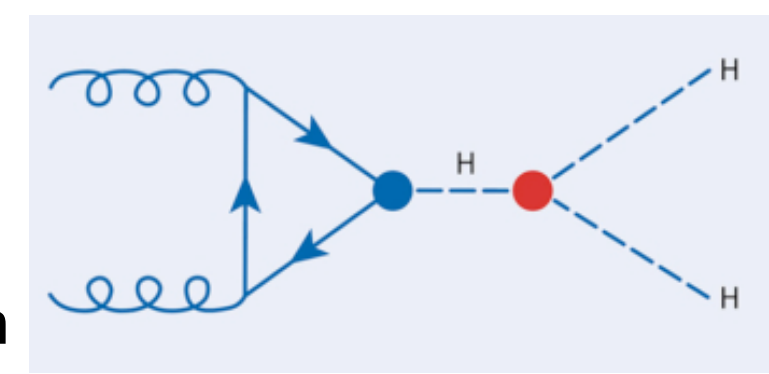
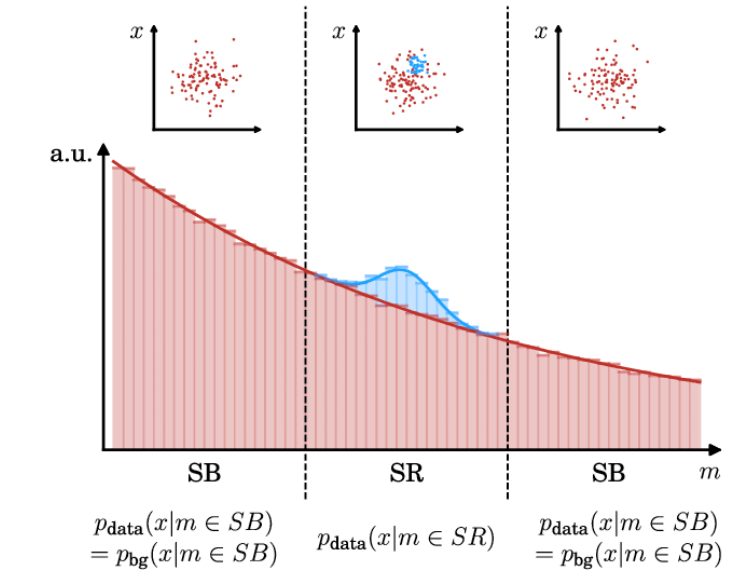
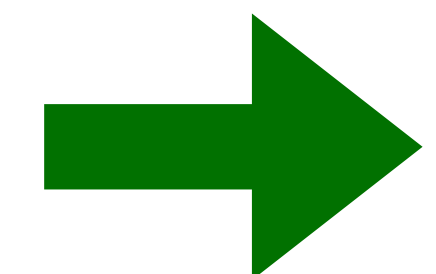
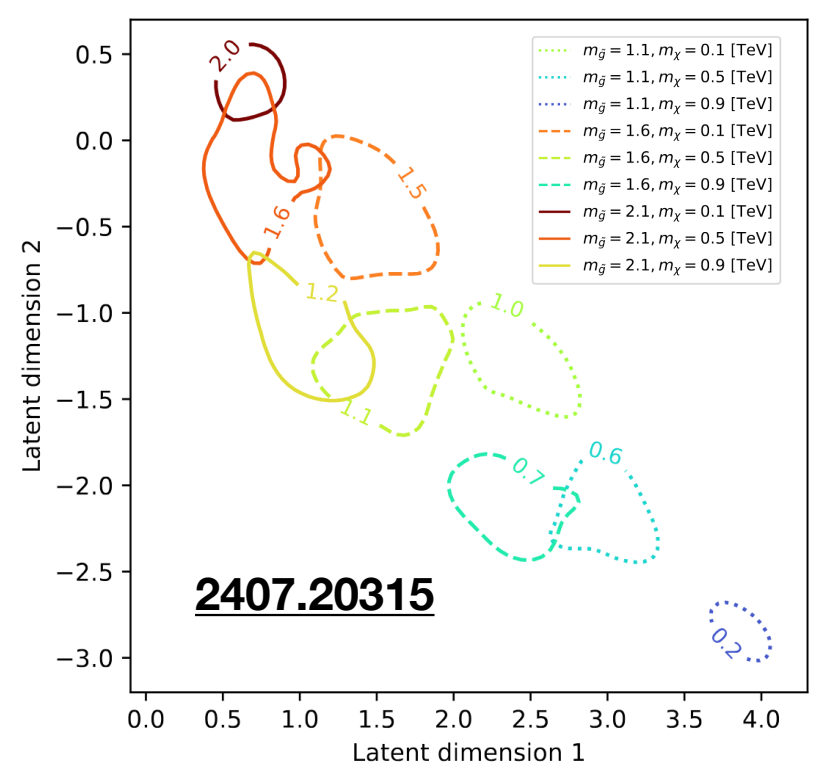
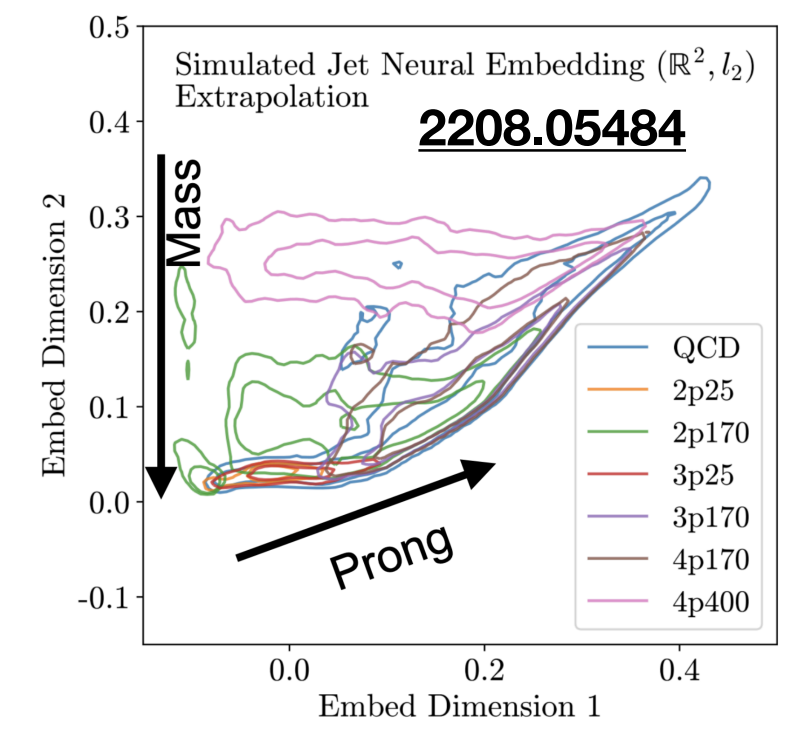
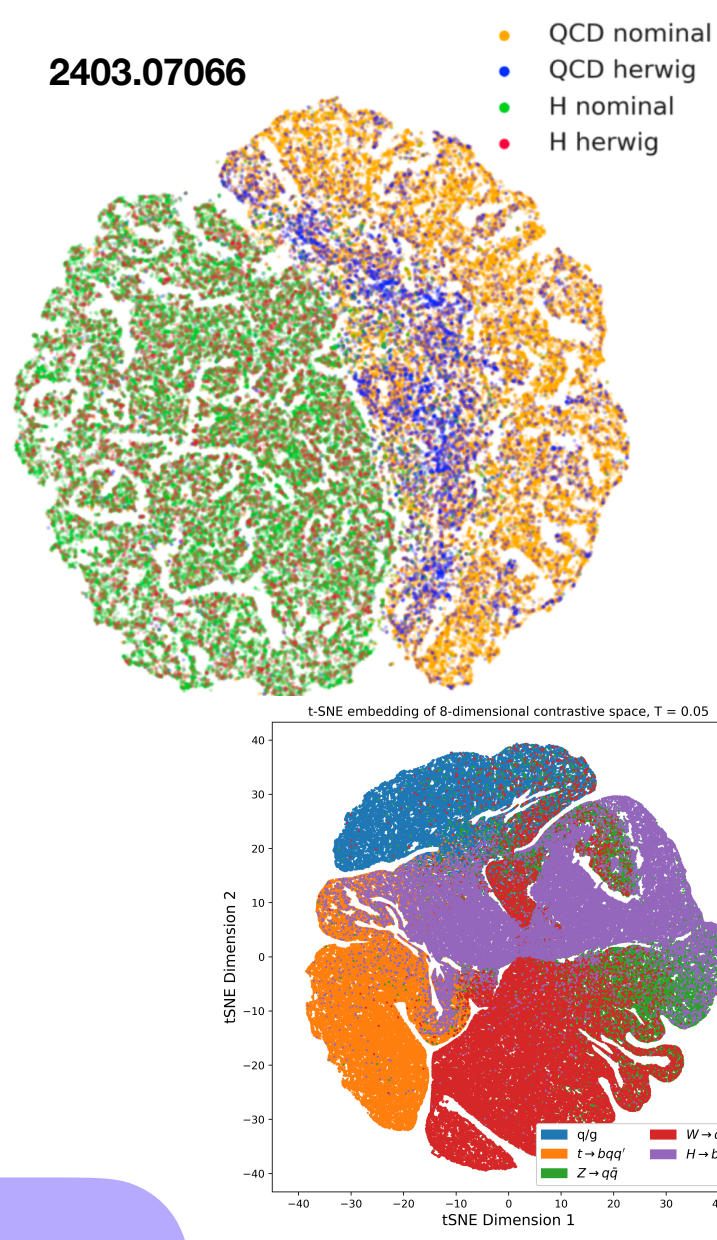
Noisy, high-dim,
complex physics inputs

Self-organized, physically meaningful,
and expressive latent features

A new workflow for HEP



Representation learning
 Self-supervised learning
 Foundation models
 Physics-informed architectures?
 Physics-aware training?
 Interpretable ML?



Anomaly detection
 Precision measurements

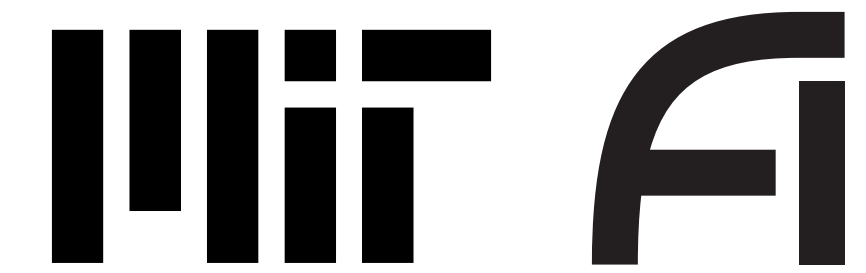
BSM physics searches

Any analysis we want,
 faster & easier

Self-organized, physically meaningful,
 and expressive latent features

Lots of room for
 new ideas!

Talk to me about...



Physics things

- **Representation learning** for collider physics and/or LIGO
- Training/deploying “**Foundation models**” for experiments
- **Robustness/bias** of AI models in science
 - uncertainty, generalizability, etc.

AI things

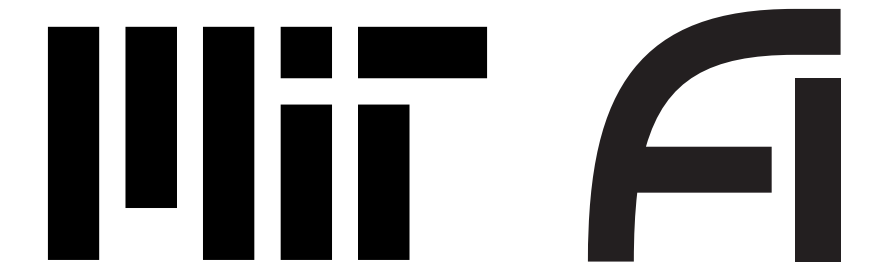
- Fundamental questions of the “how does this work” variety (toy models, etc.)
- Self-supervised learning; how training objective shapes a model’s “understanding”
- Generalization capabilities, particularly for models trained on scientific data

These lists aren't mutually exclusive!

Thanks for listening!

Backup

CASE



Phil Harris¹, Patrick McCormack¹, Sang Eon Park¹, Tobias Quadfasel², Manuel Sommerhalder², Louis Moureaux², Gregor Kasieczka², Oz Amram⁴, Petar Maksimovic³, Nadya Chernyavskaya⁵, Benedikt Maier⁷, Maurizio Pierini⁵, Kinga Wozniak⁵, Thea Aarrestad⁶, Jennifer Ngadiuba⁸, Irene Zoi⁸, Samuel Bright-Thonney⁹, David Shih¹⁰, and Aritra Bal¹¹

¹ MIT

² Universität Hamburg

³ Johns Hopkins

⁴ Johns Hopkins and Fermilab

⁵ CERN

⁶ CERN and ETH Zurich

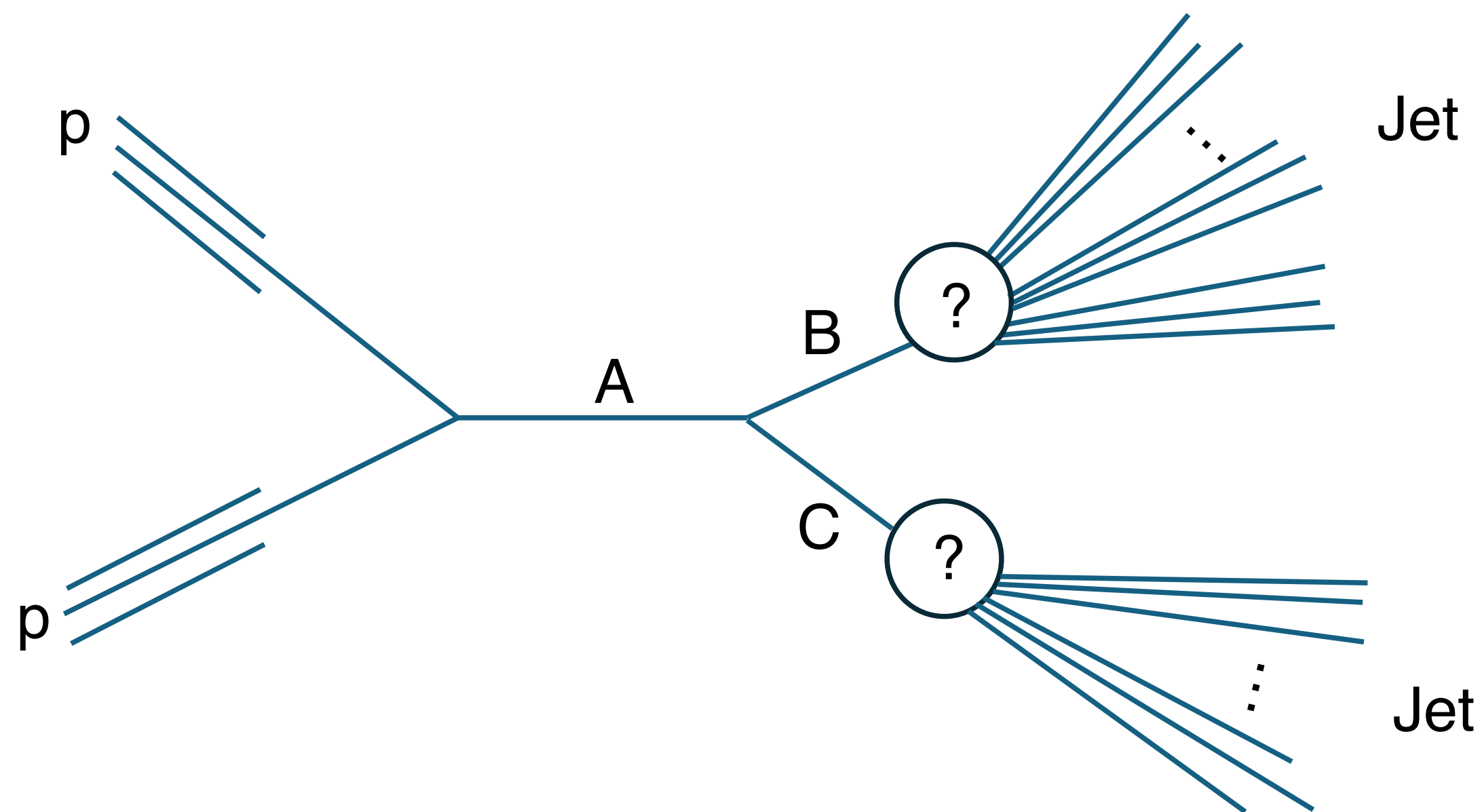
⁷ CERN and KIT

⁸ Fermilab

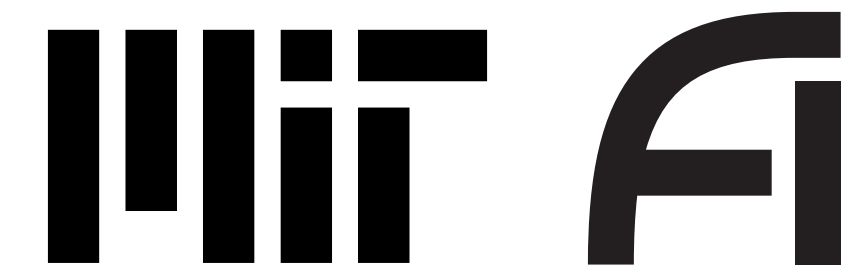
⁹ Cornell

¹⁰ Rutgers

¹¹ KIT

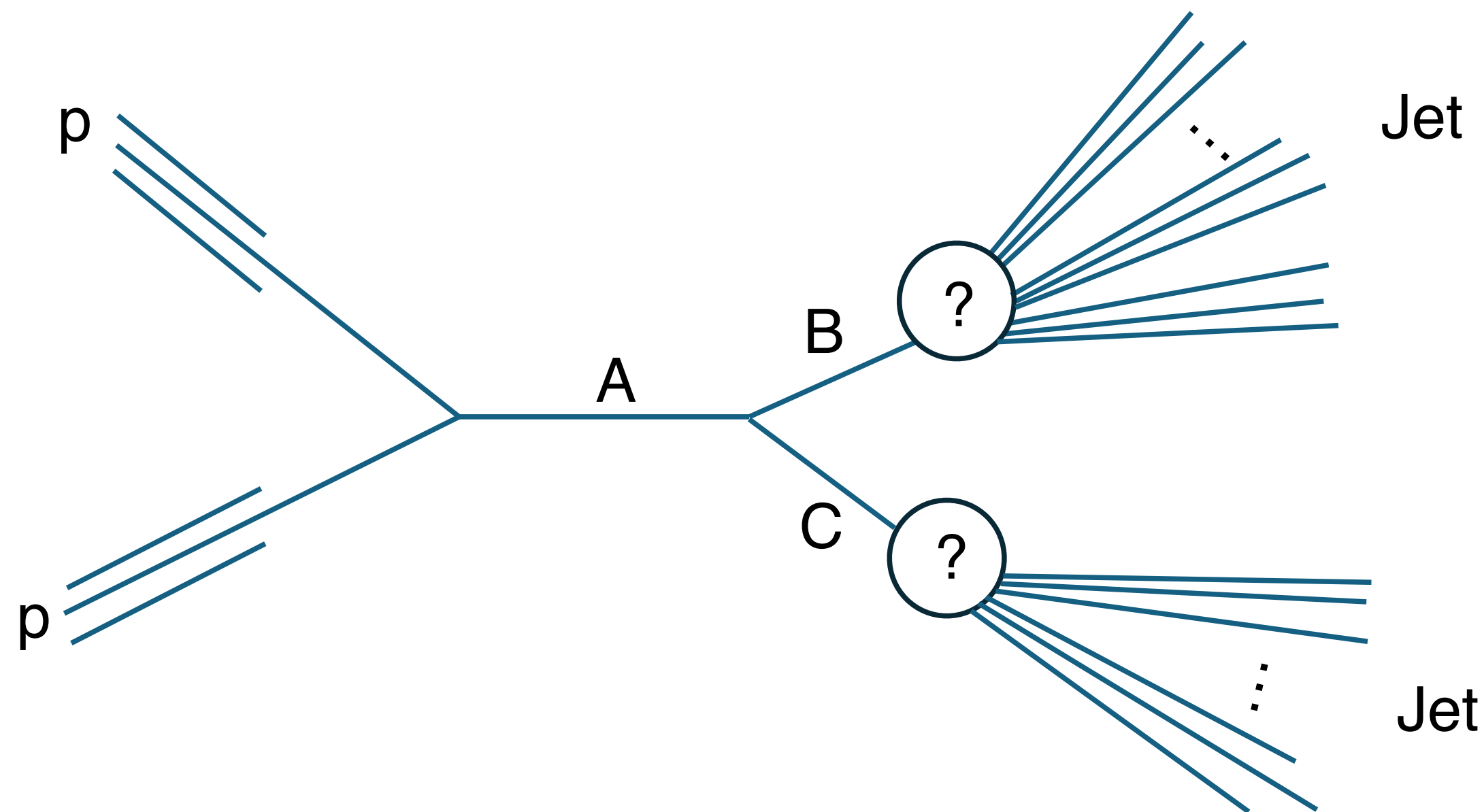


CASE



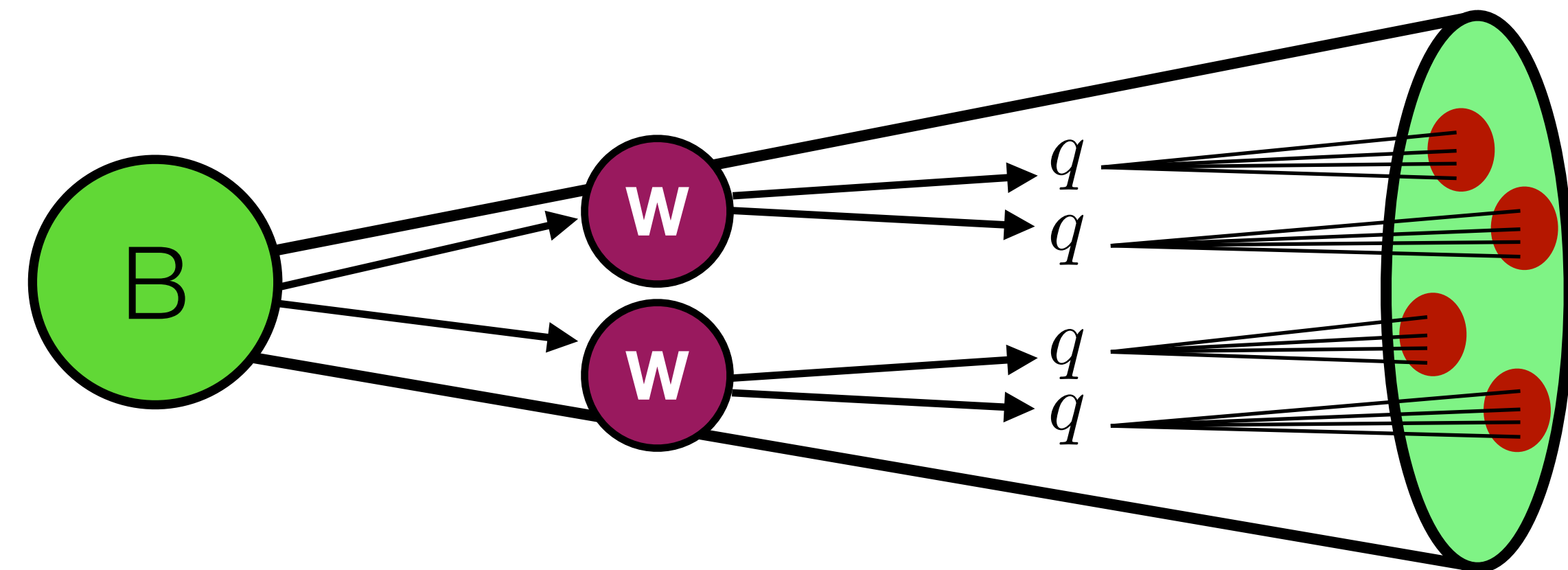
CASE = The **CMS Anomaly Search** (Effort)

- Looking for small resonant signatures using **anomaly detection** techniques
- Focusing on **jet substructure**

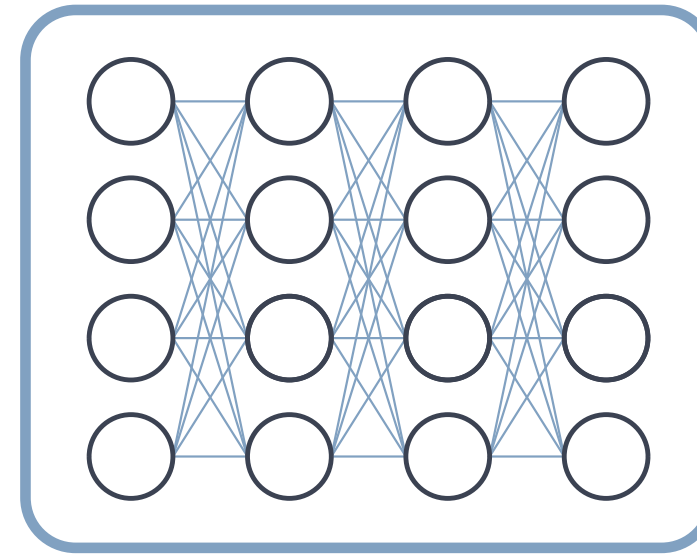
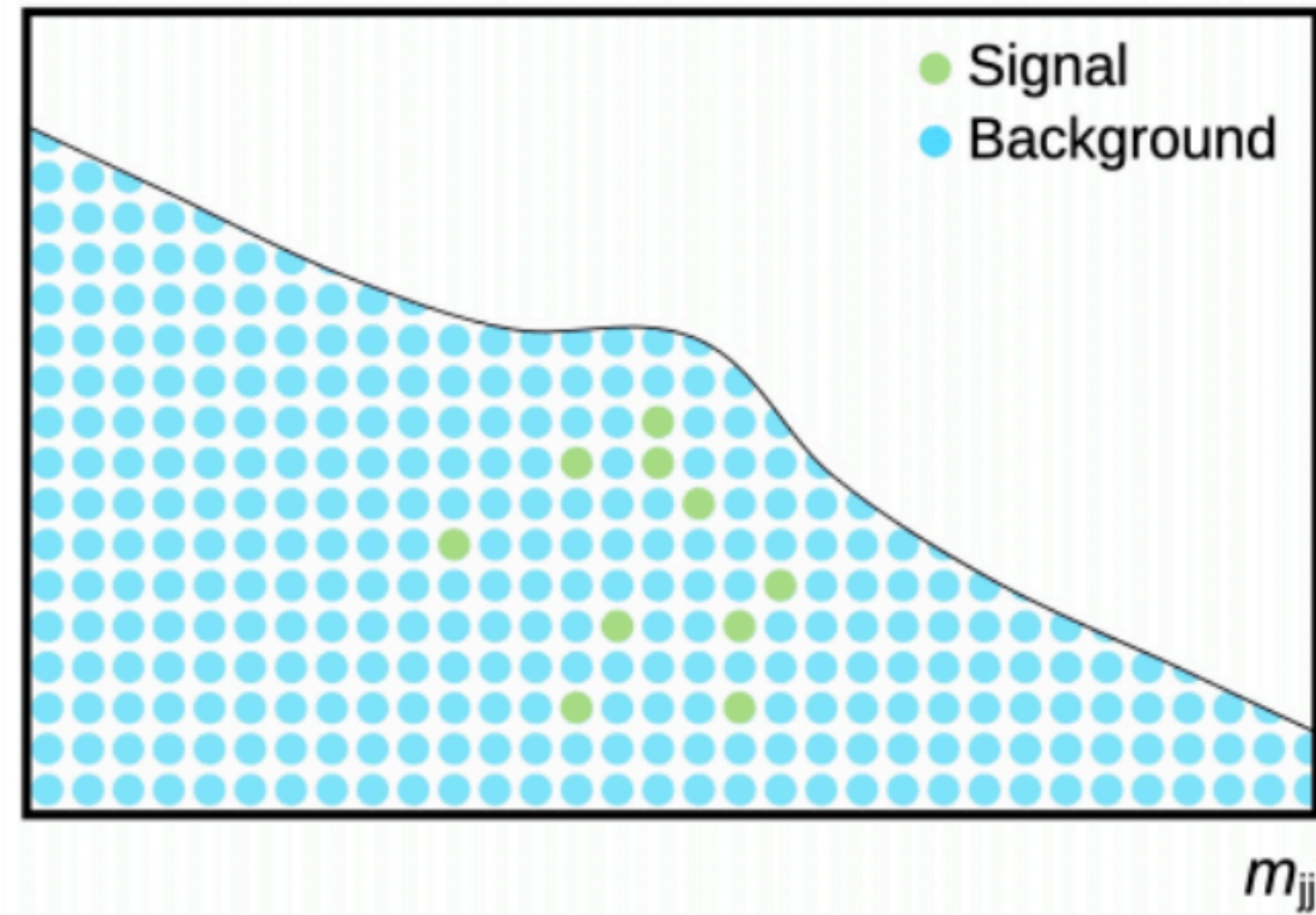


Phil Harris¹, Patrick McCormack¹, Sang Eon Park¹, Tobias Quadfasel², Manuel Sommerhalder², Louis Moureaux², Gregor Kasieczka², Oz Amram⁴, Petar Maksimovic³, Nadya Chernyavskaya⁵, Benedikt Maier⁷, Maurizio Pierini⁵, Kinga Wozniak⁵, Thea Aarrestad⁶, Jennifer Ngadiuba⁸, Irene Zoi⁸, Samuel Bright-Thonney⁹, David Shih¹⁰, and Aritra Bal¹¹

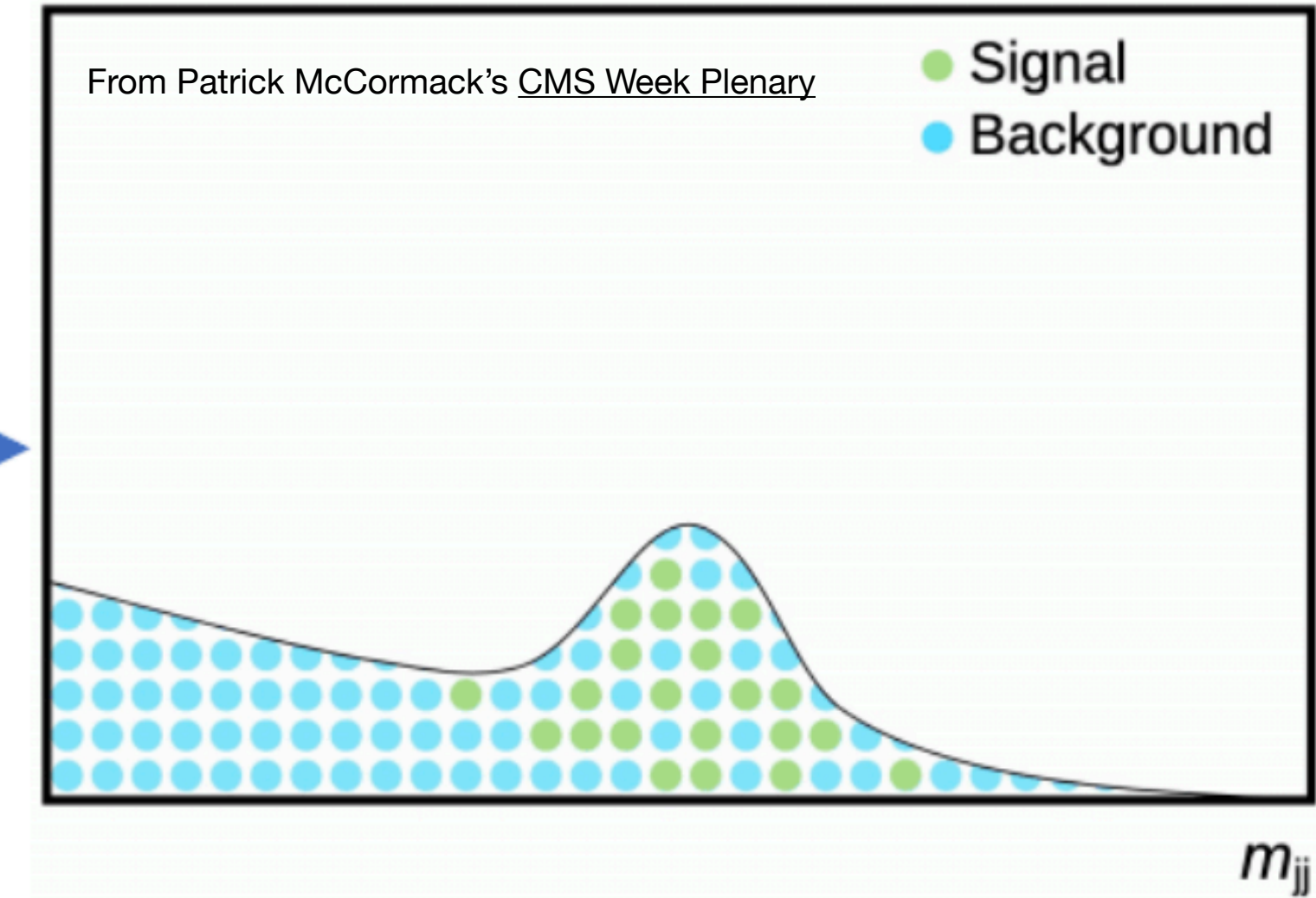
- ¹ MIT
- ² Universität Hamburg
- ³ Johns Hopkins
- ⁴ Johns Hopkins and Fermilab
- ⁵ CERN
- ⁶ CERN and ETH Zurich
- ⁷ CERN and KIT
- ⁸ Fermilab
- ⁹ Cornell
- ¹⁰ Rutgers
- ¹¹ KIT



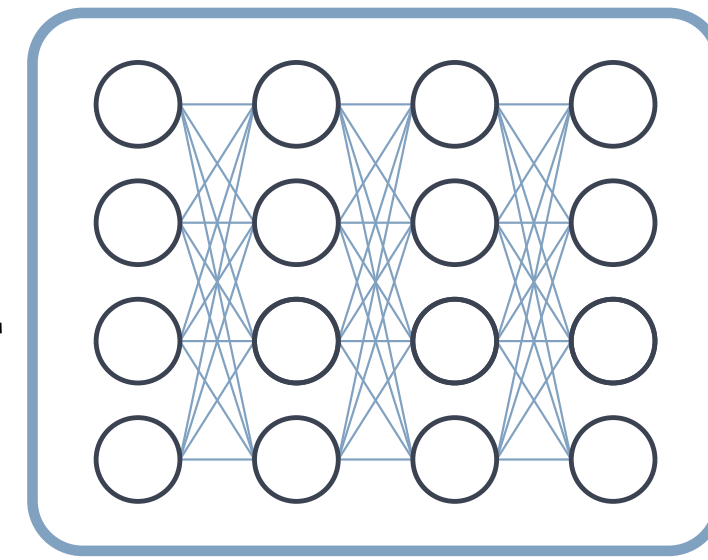
Bump hunting



**Machine learning
*anomaly detection***



Bump hunting

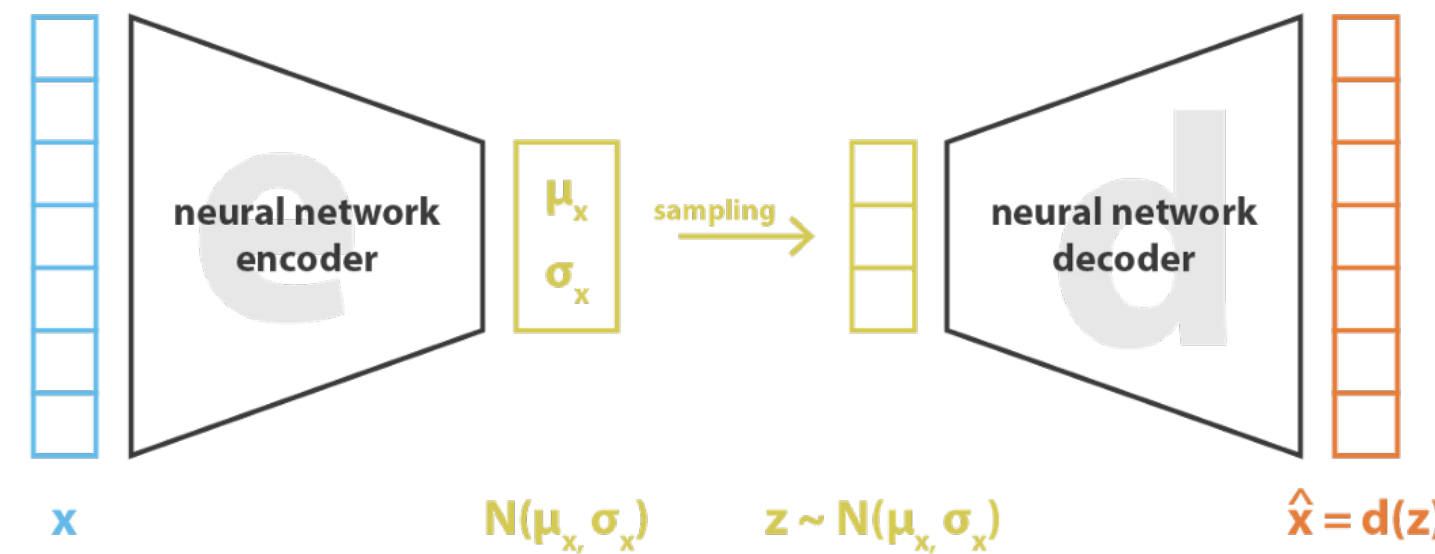
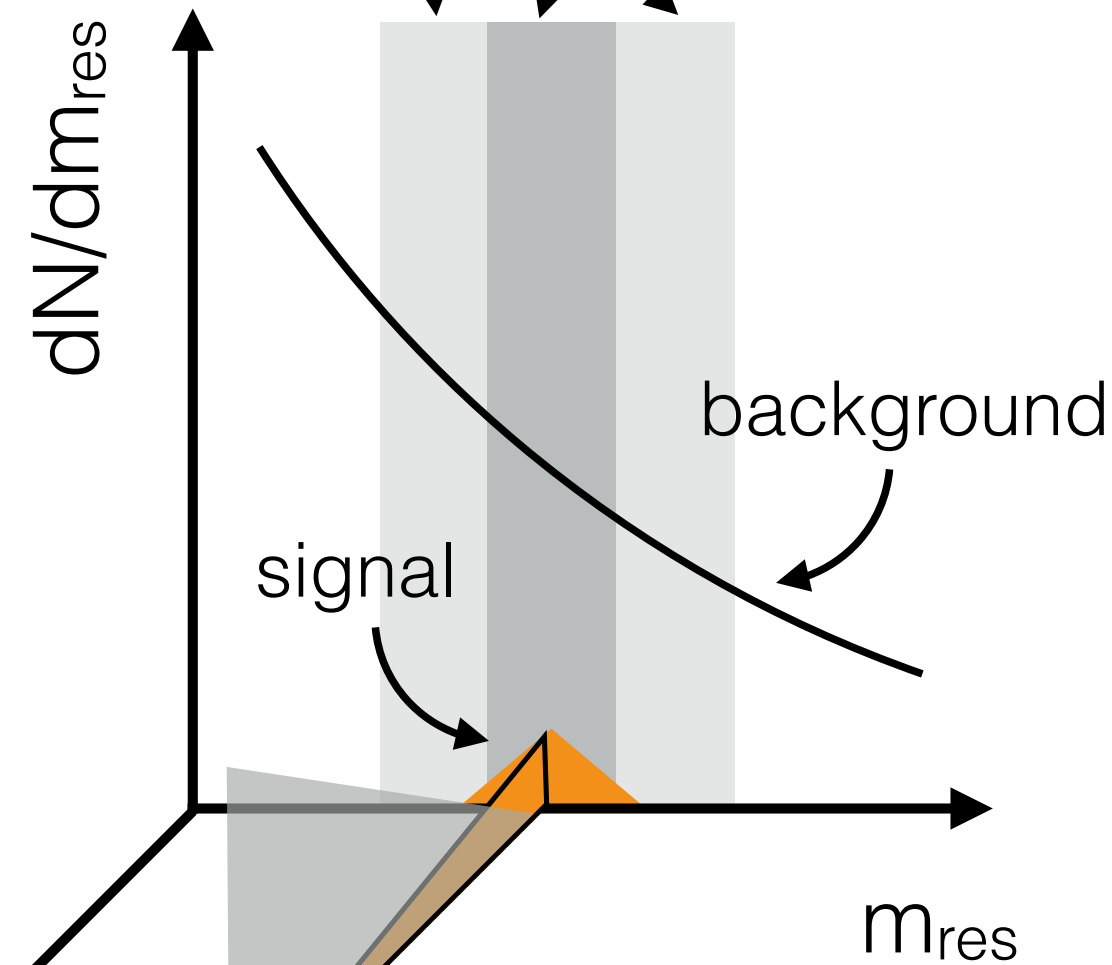
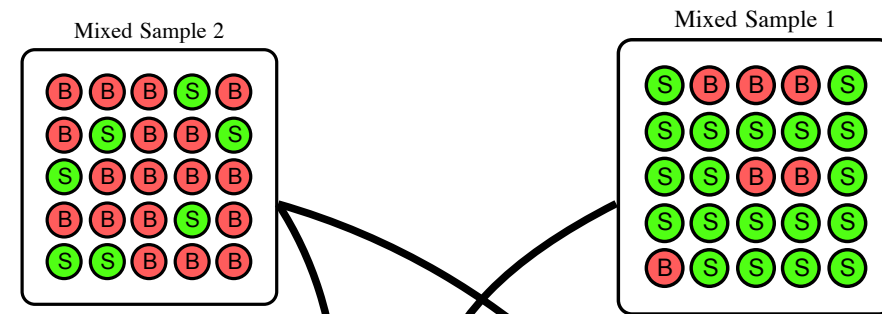


Machine learning
anomaly detection

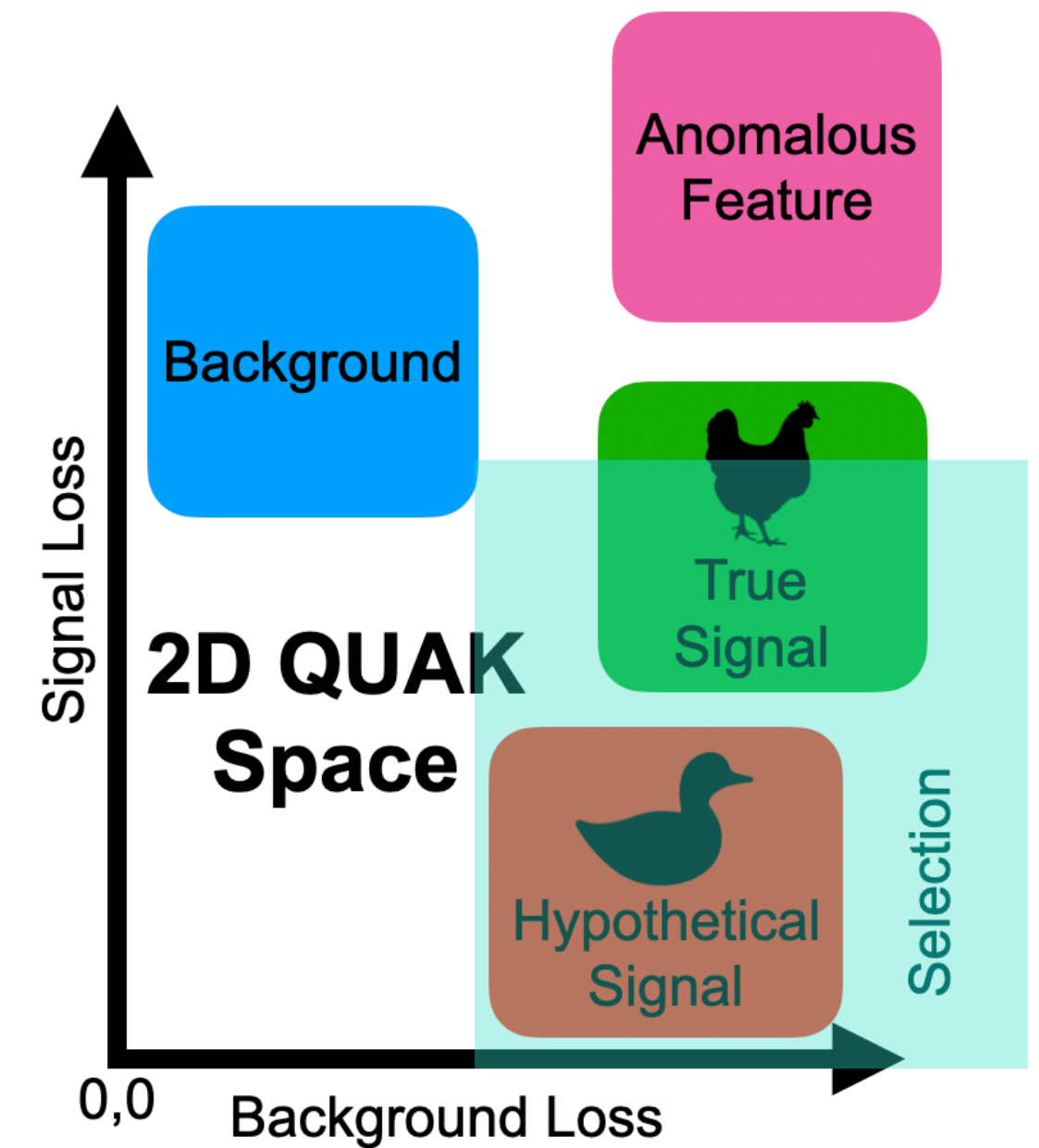
“Weakly Supervised”

“Semi-Supervised”

“Unsupervised”

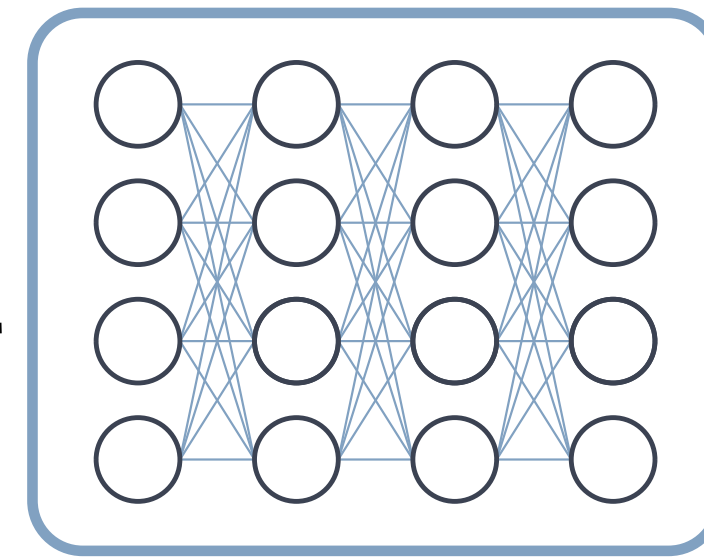


$$\text{loss} = \|x - \hat{x}\|^2 + \text{KL}[N(\mu_x, \sigma_x), N(0, I)] = \|x - d(z)\|^2 + \text{KL}[N(\mu_x, \sigma_x), N(0, I)]$$



Other features

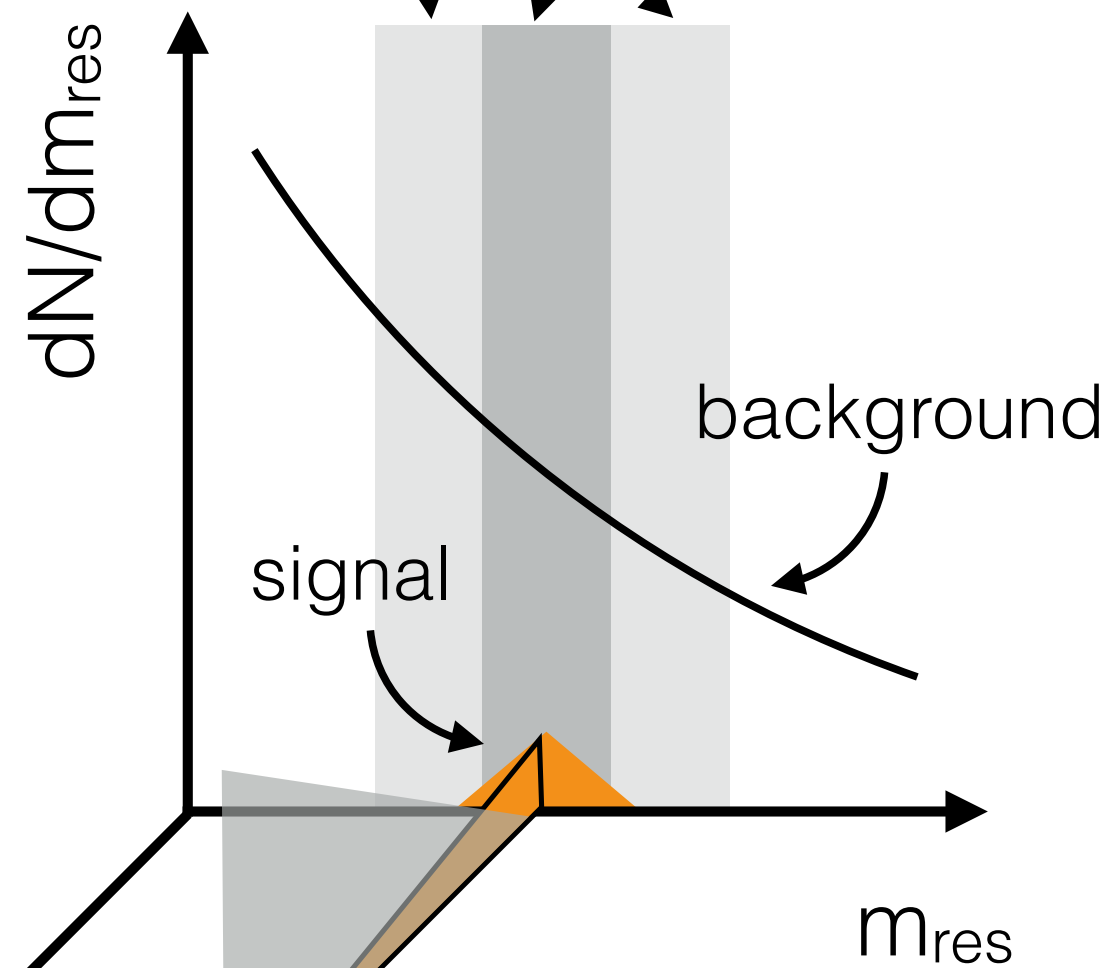
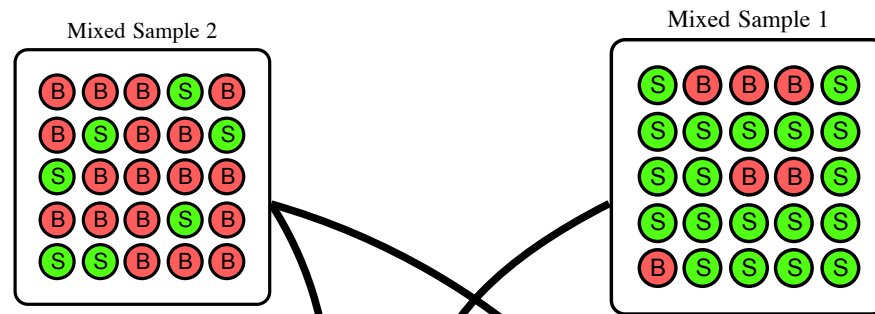
Bump hunting



Machine learning
anomaly detection

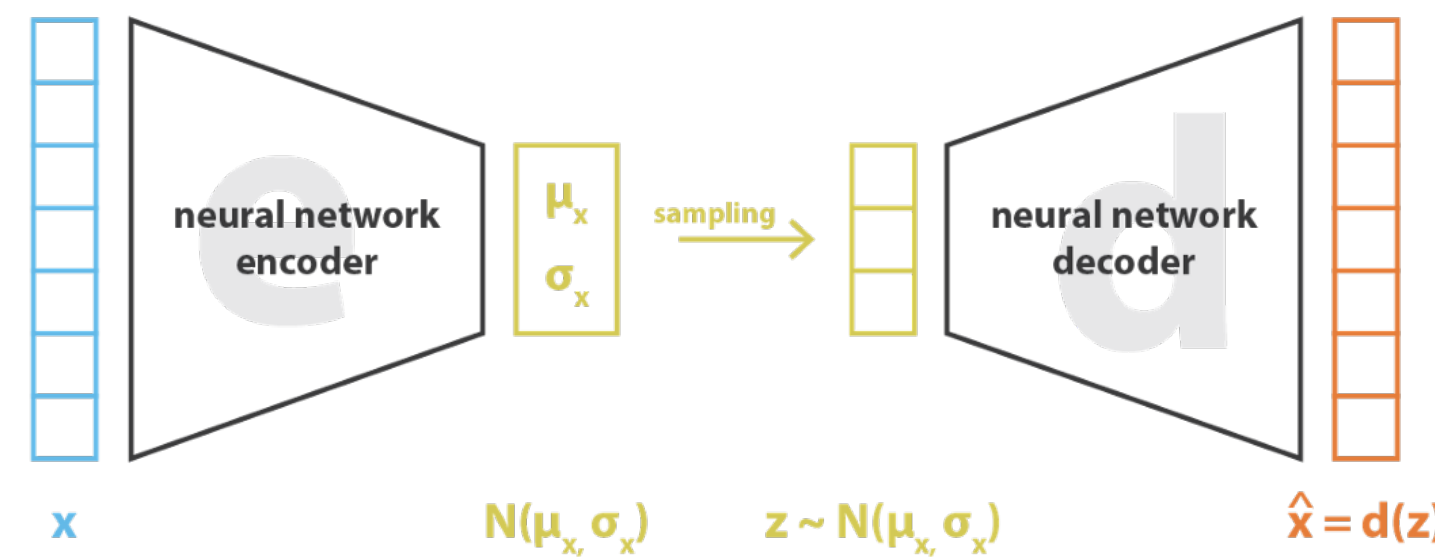
“Weakly Supervised”

“Semi-Supervised”

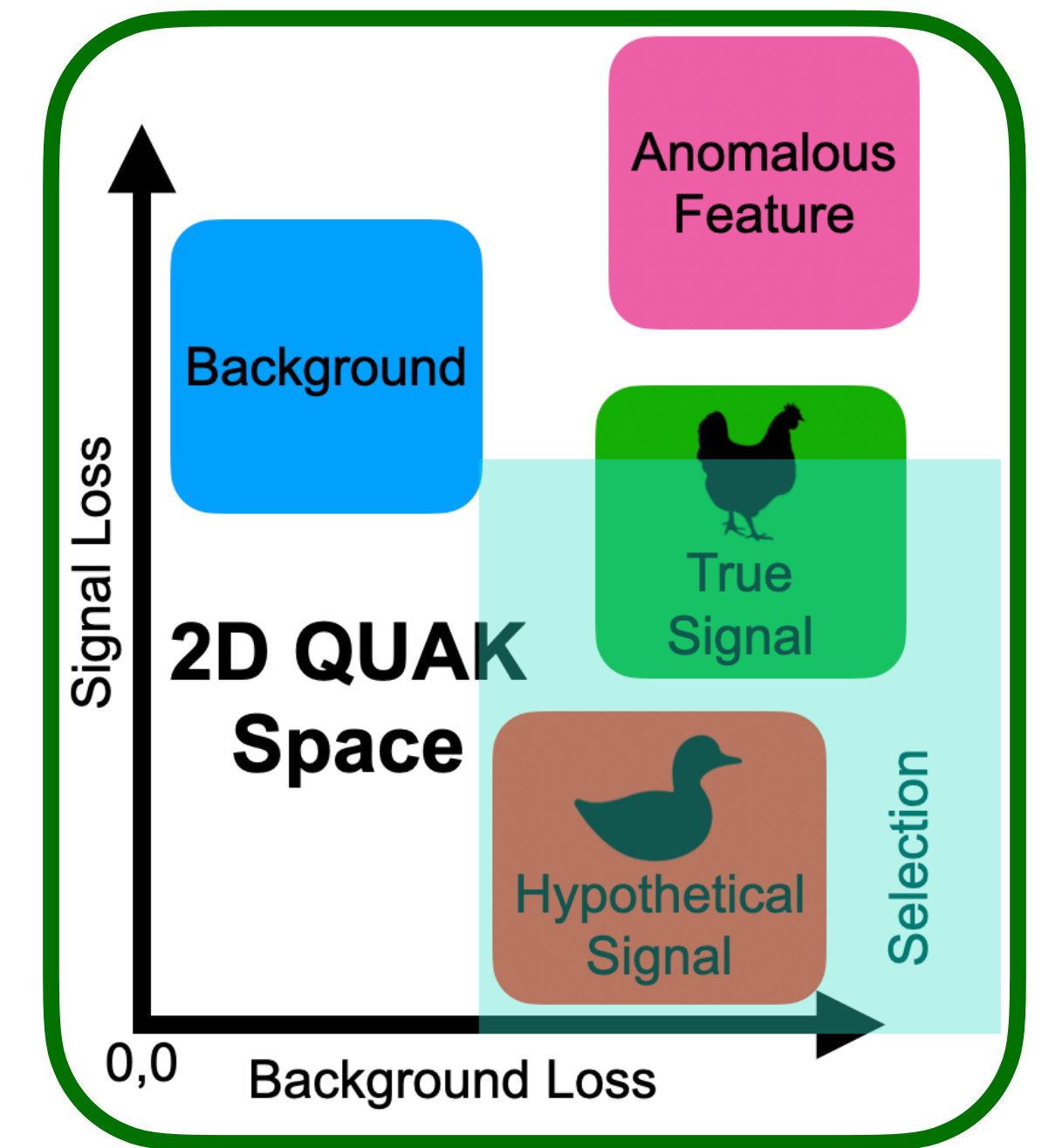


Our focus

“Unsupervised”

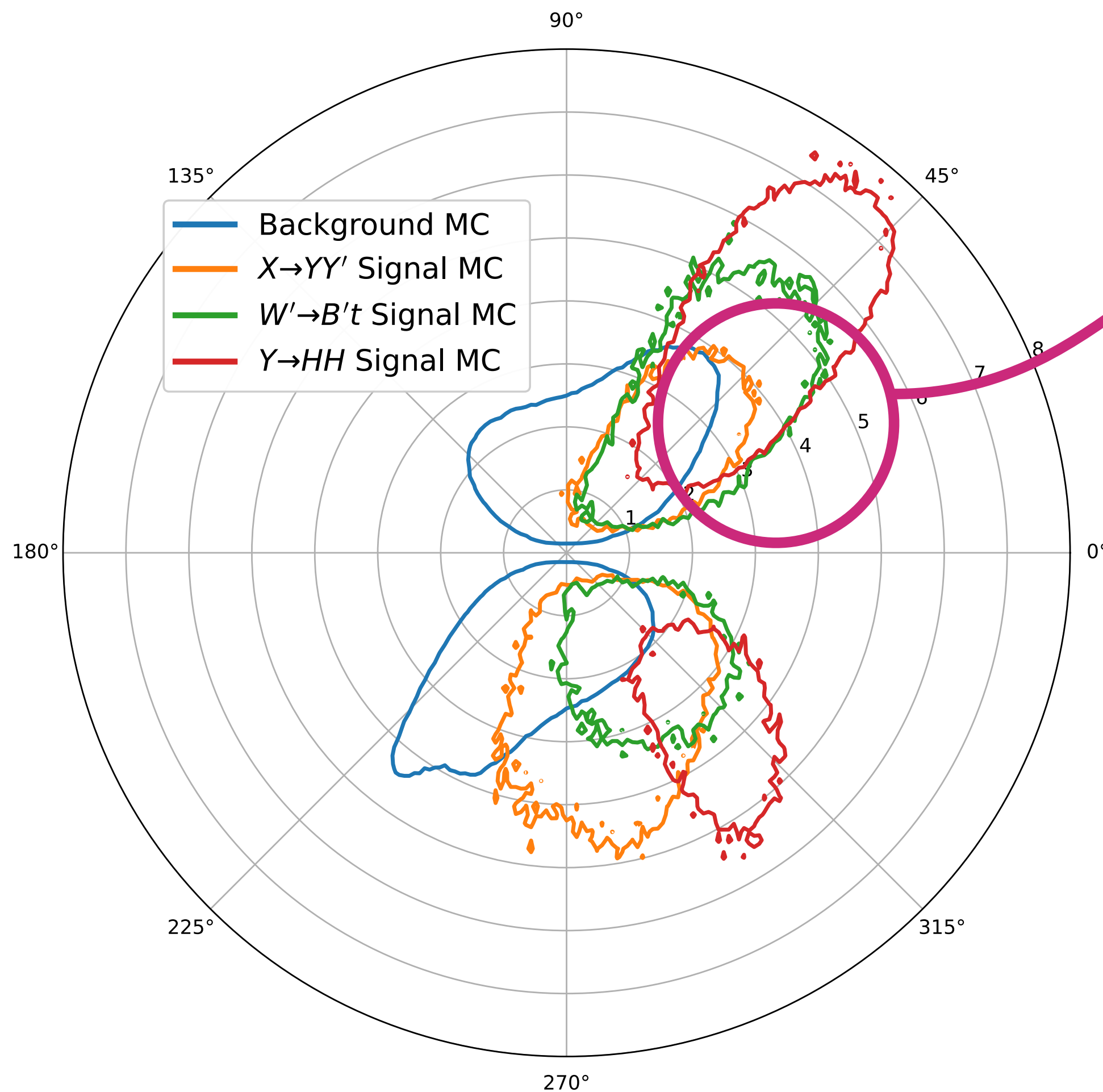


$$\text{loss} = \|x - \hat{x}\|^2 + \text{KL}[N(\mu_x, \sigma_x), N(0, I)] = \|x - d(z)\|^2 + \text{KL}[N(\mu_x, \sigma_x), N(0, I)]$$

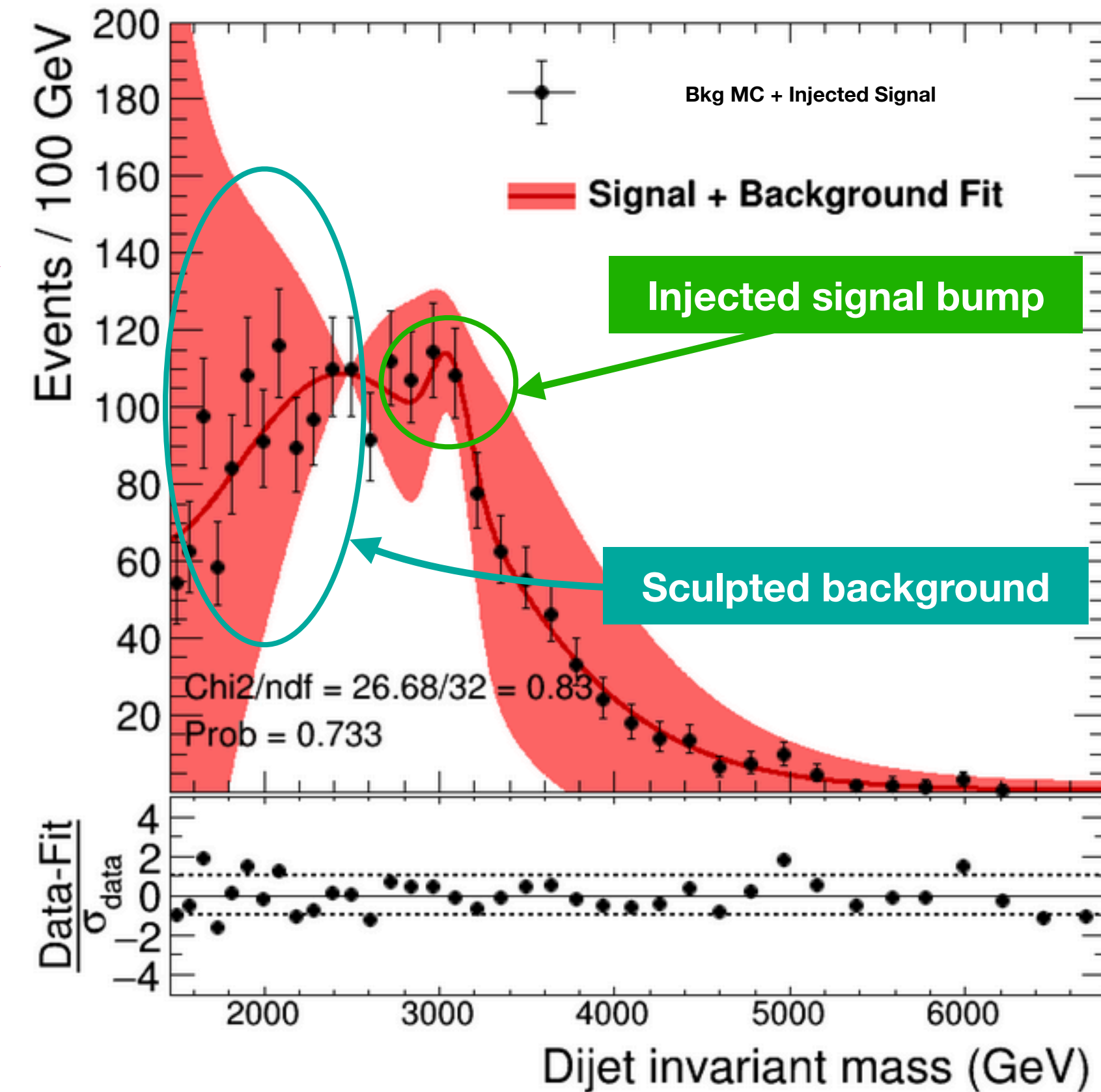


Selecting events

How should we select events for the bump hunt?



Naive choice: "bottom right"



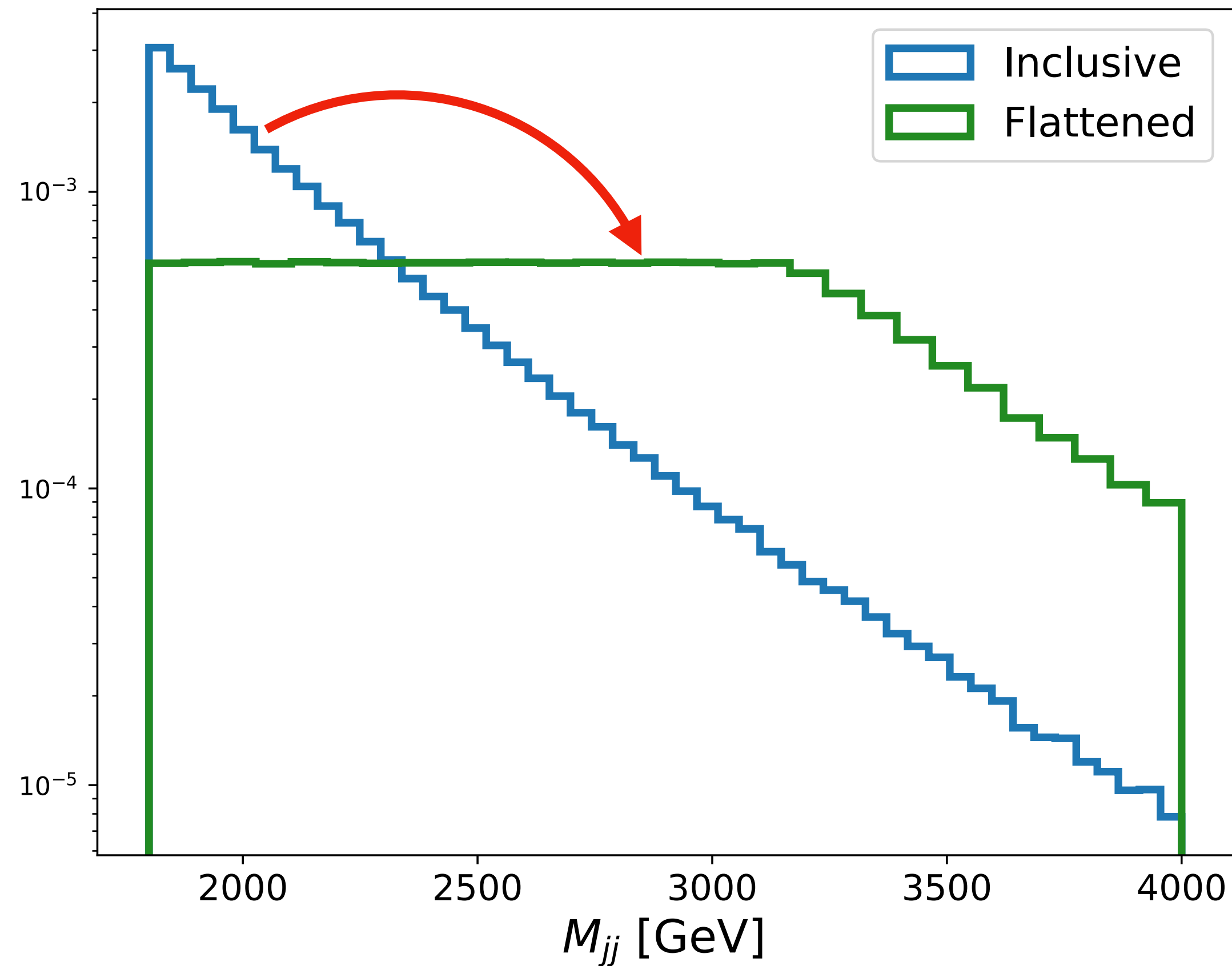
This significantly sculpts the background distribution — no good!

Driven by inverse correlation between M_{jj} & bkg-like loss

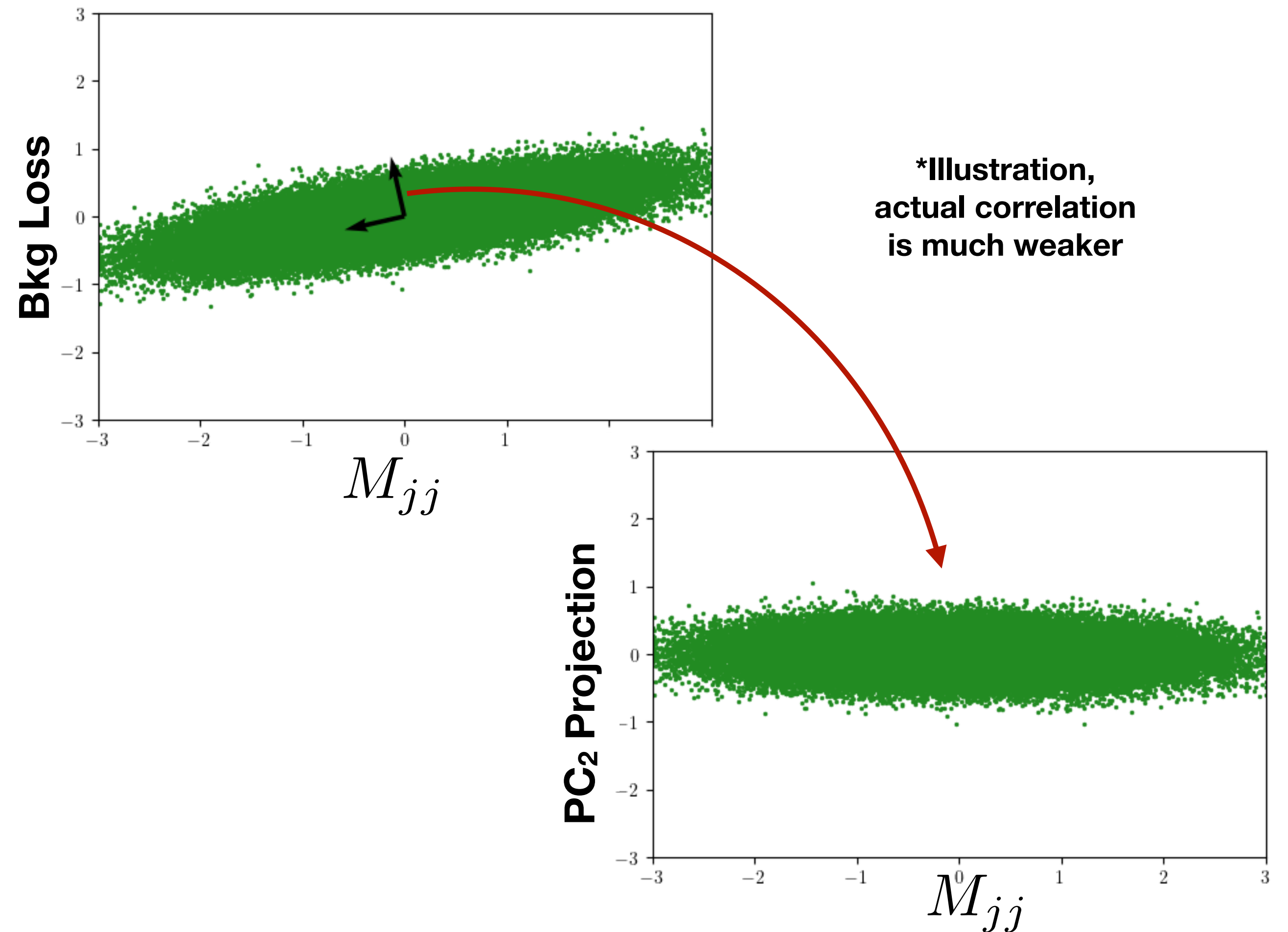
Mass decorrelation

Devised a **two step approach** to decorrelate each ML model's output from M_{jj}

Step 1: Flatten Bkg Training Sample

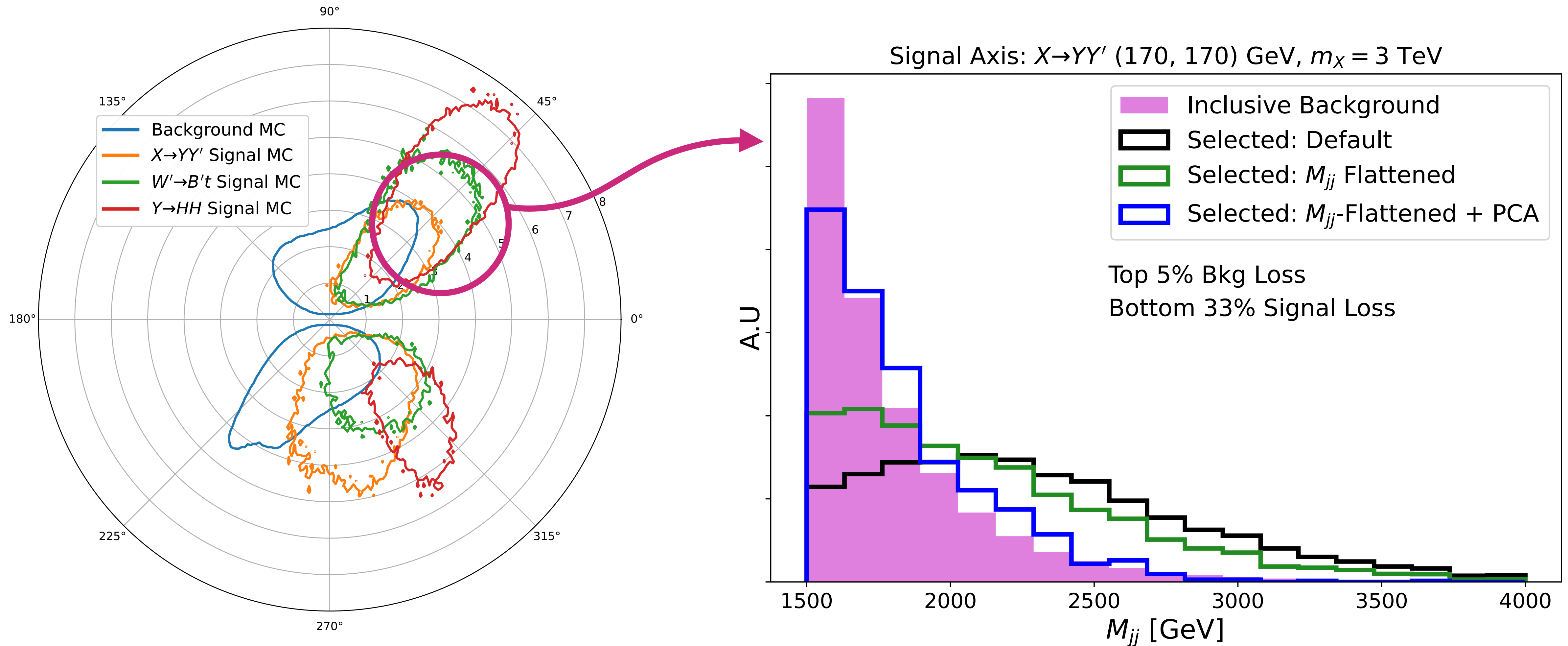


Step 2: PCA to account for residual correlation

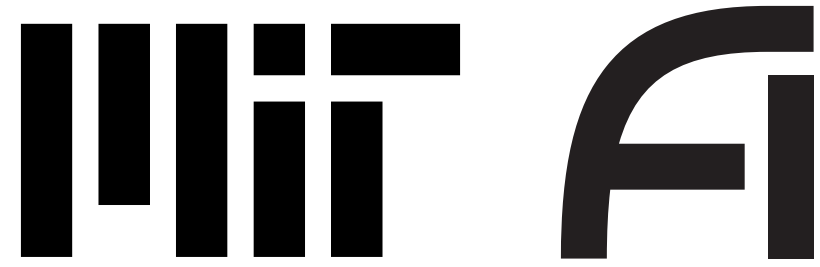


Mass decorrelation

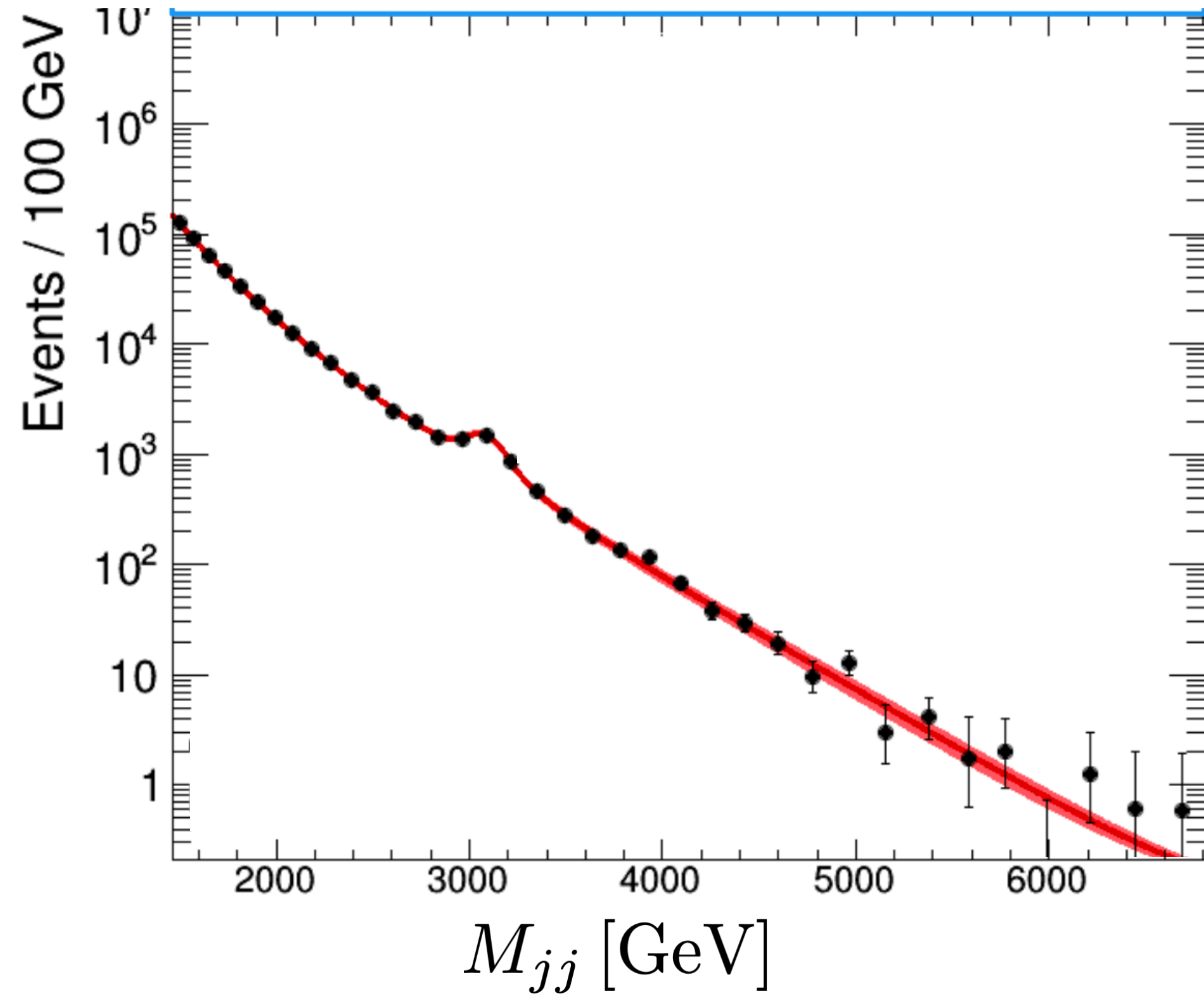
Decorrelation measures successfully “un-sculpt” background enough to use standard fit functions



Bump hunting

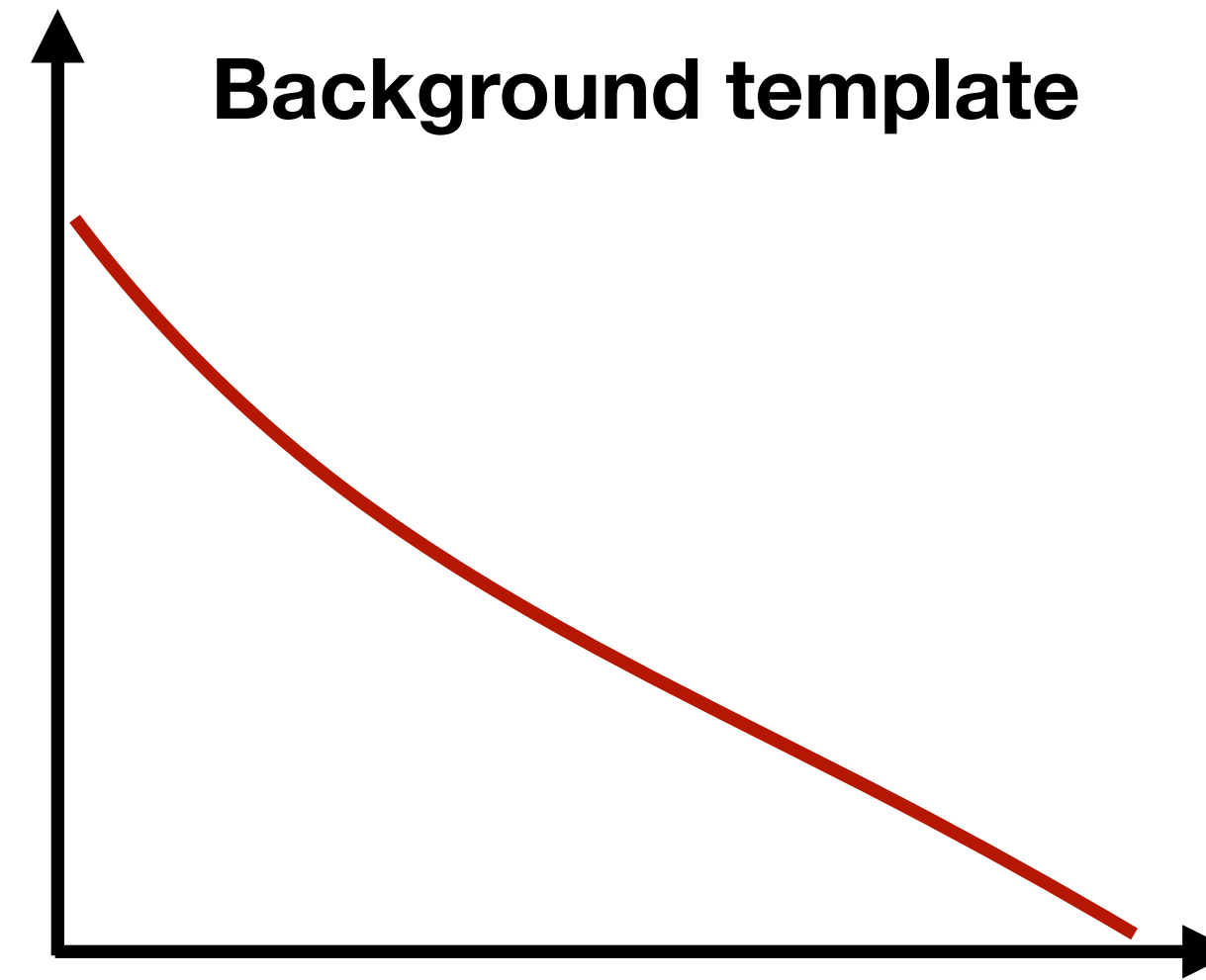


Bump hunting



Step 4

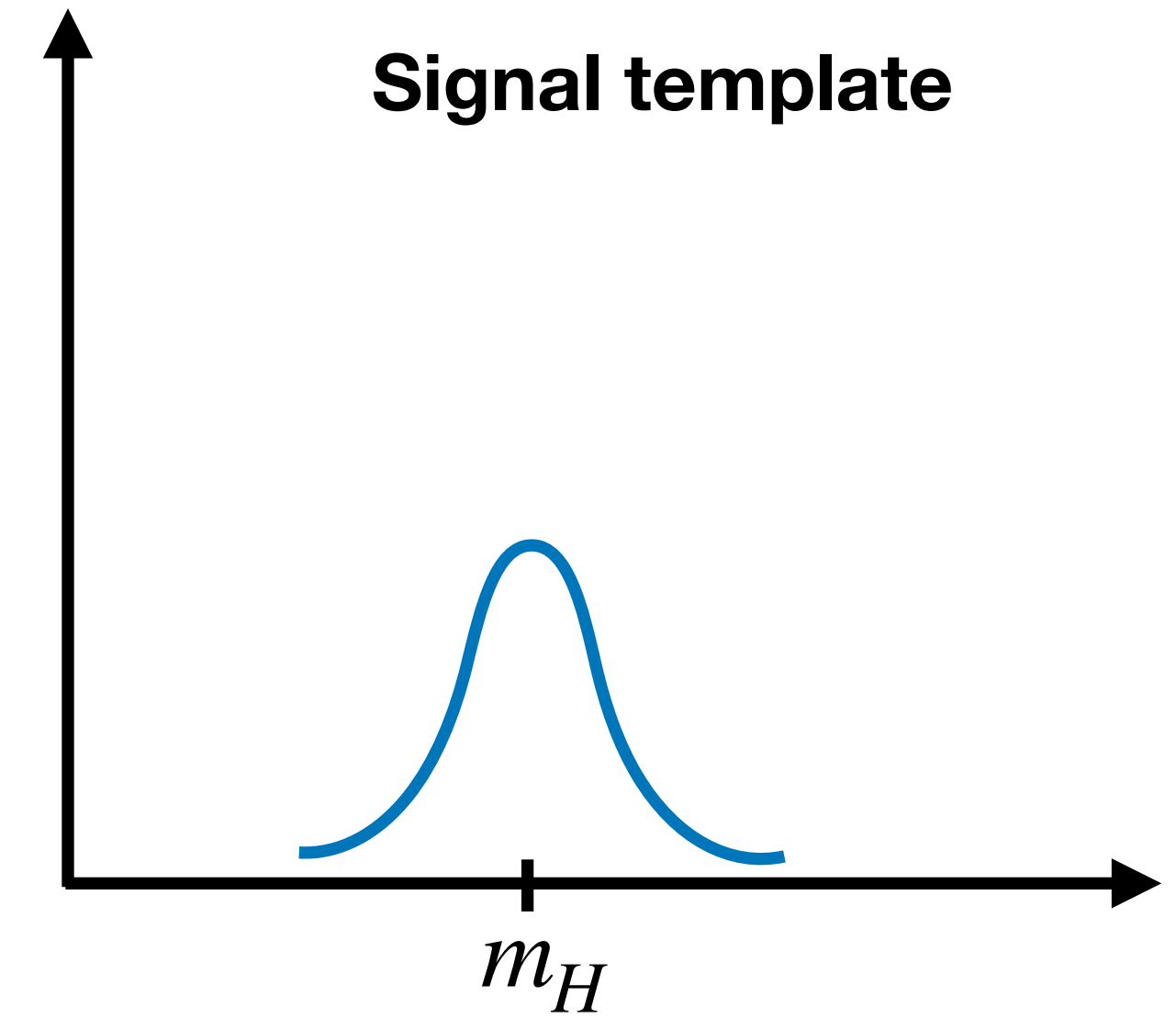
Fit the resulting spectrum to a background (smoothly falling) + signal (bump) function



$$\frac{dN}{dM_{jj}} = \frac{P_0(1-x)^{P_1}}{x^{P_2+P_3 \log(x)+P_4 \log^2(x)}}$$

$$x = M_{jj} / \sqrt{s}$$

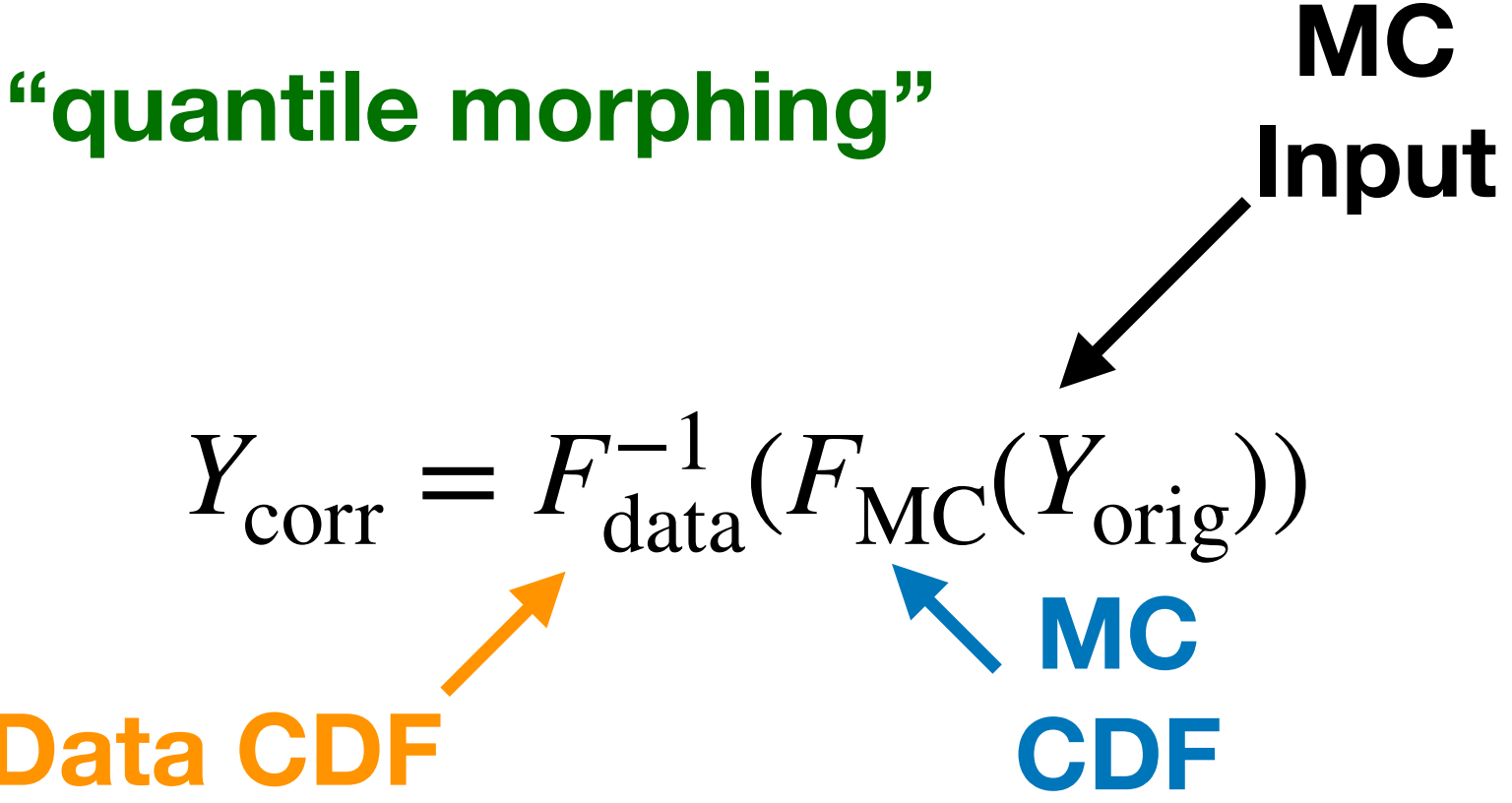
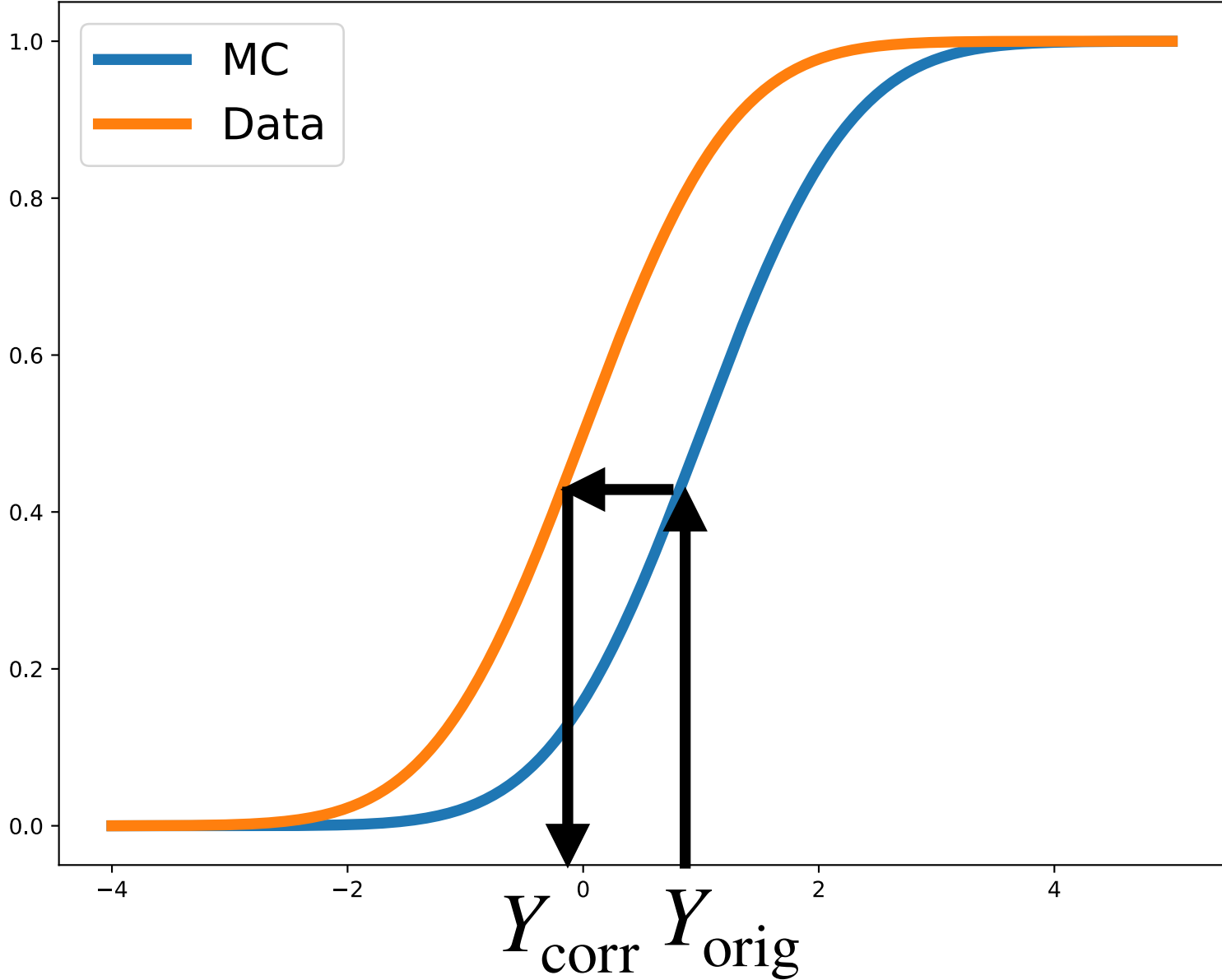
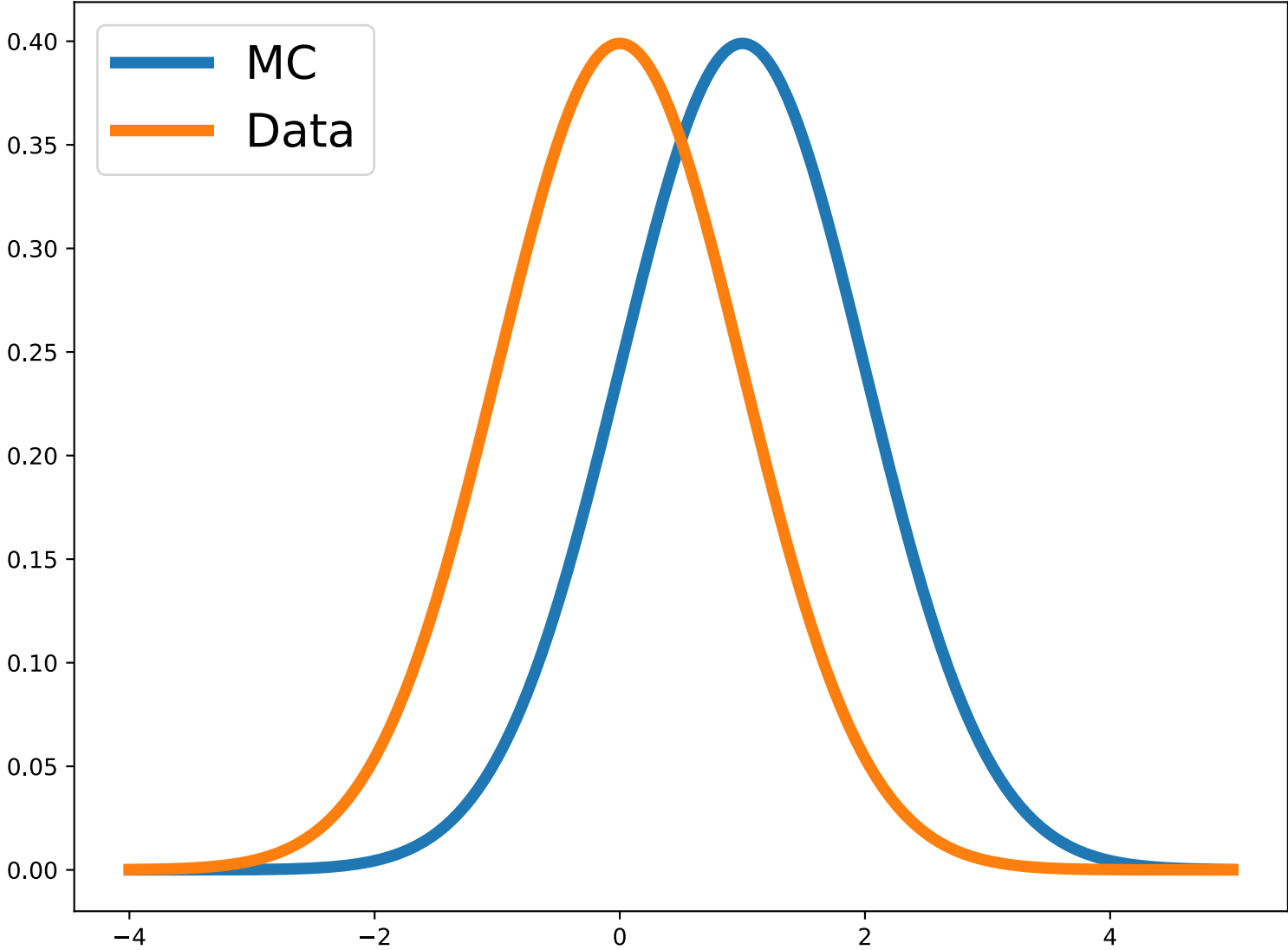
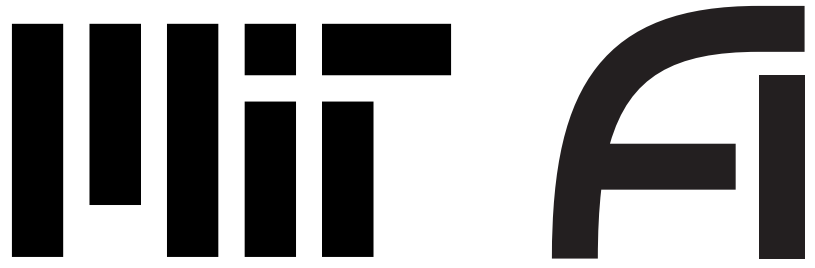
+



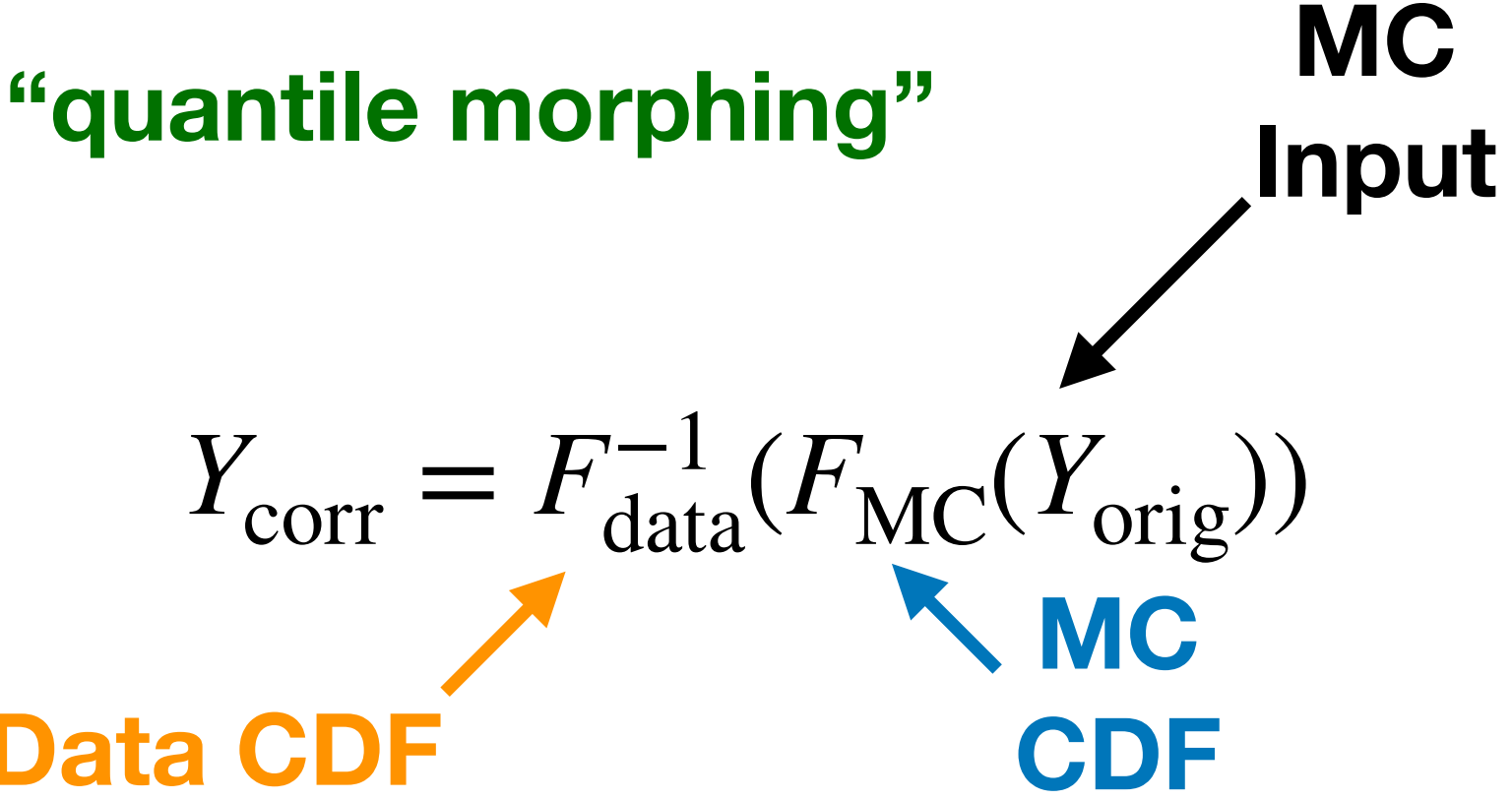
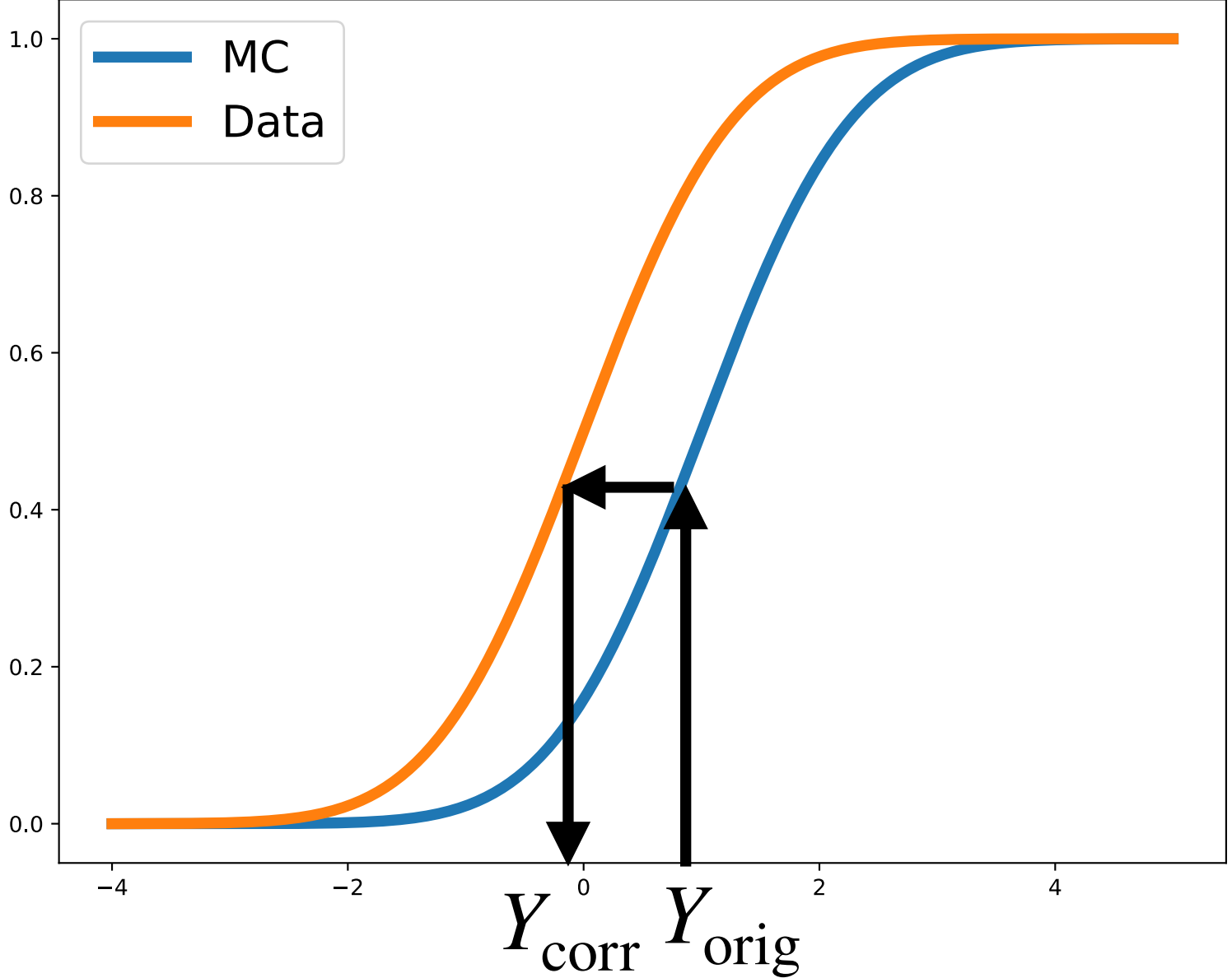
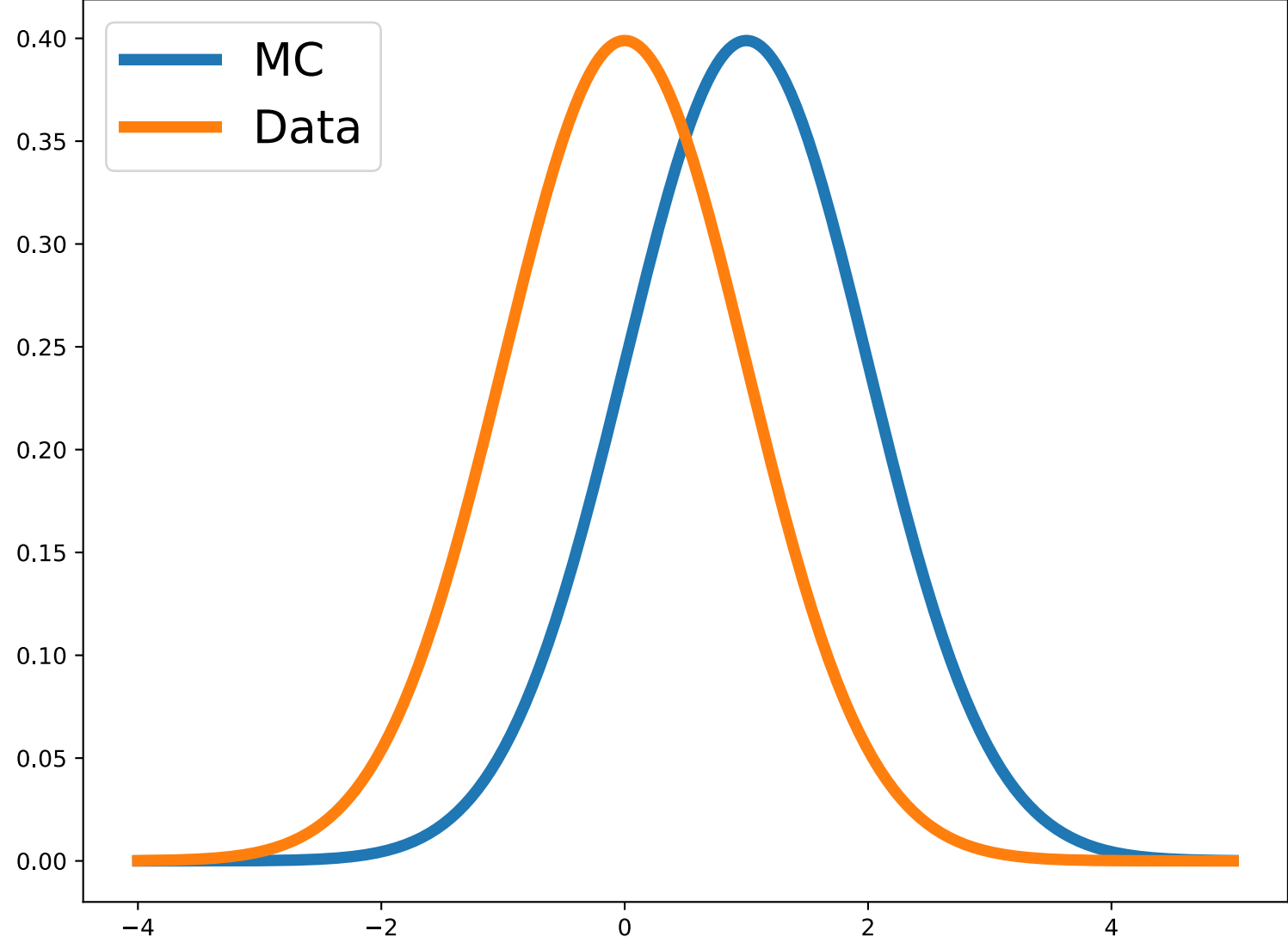
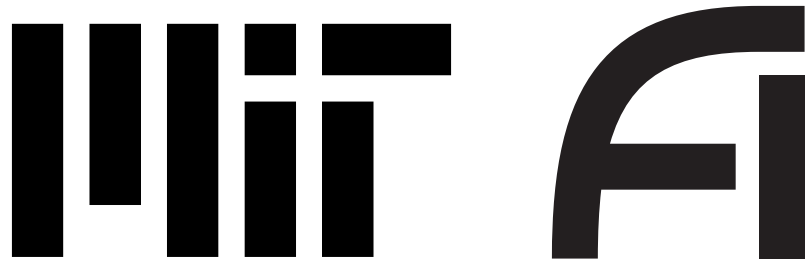
“Double-sided crystal ball”

Gaussian core + power law tails

A different solution: morphing



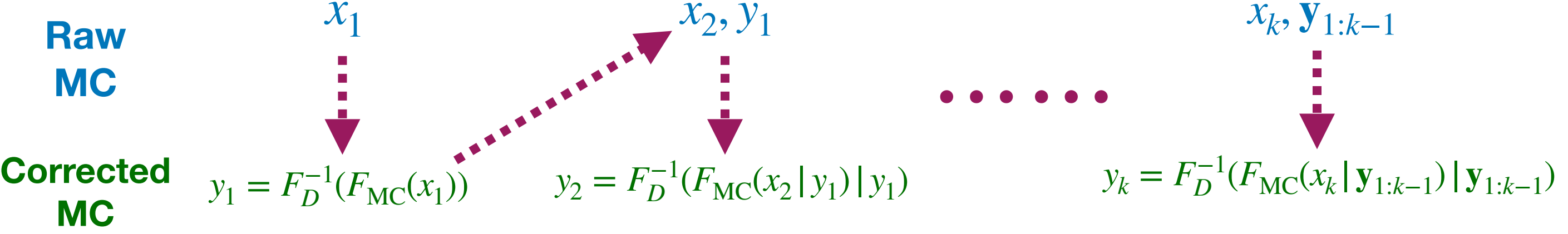
A different solution: morphing



N-D Case

$$p(\mathbf{x}) = p(x_1)p(x_2 | x_1) \cdots p(x_k | \mathbf{x}_{1:k-1})$$

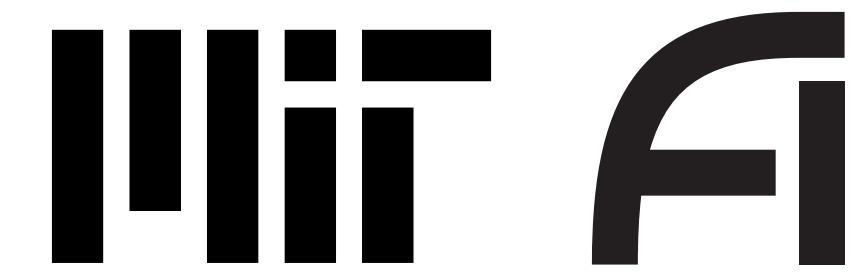
$F_1(x_1)$ $F_2(x_2 | x_1)$ $F_k(x_k | \mathbf{x}_{1:k-1})$



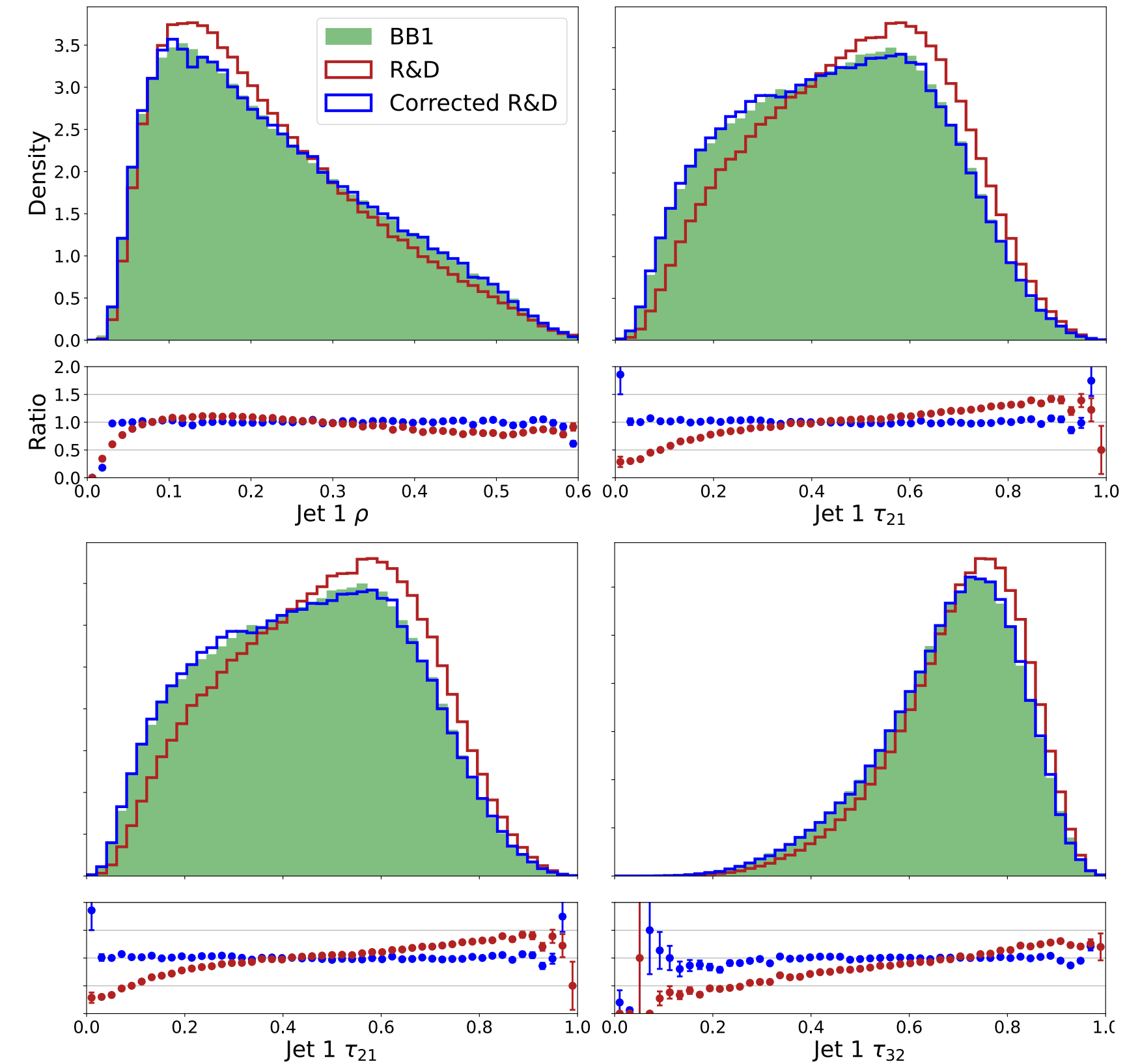
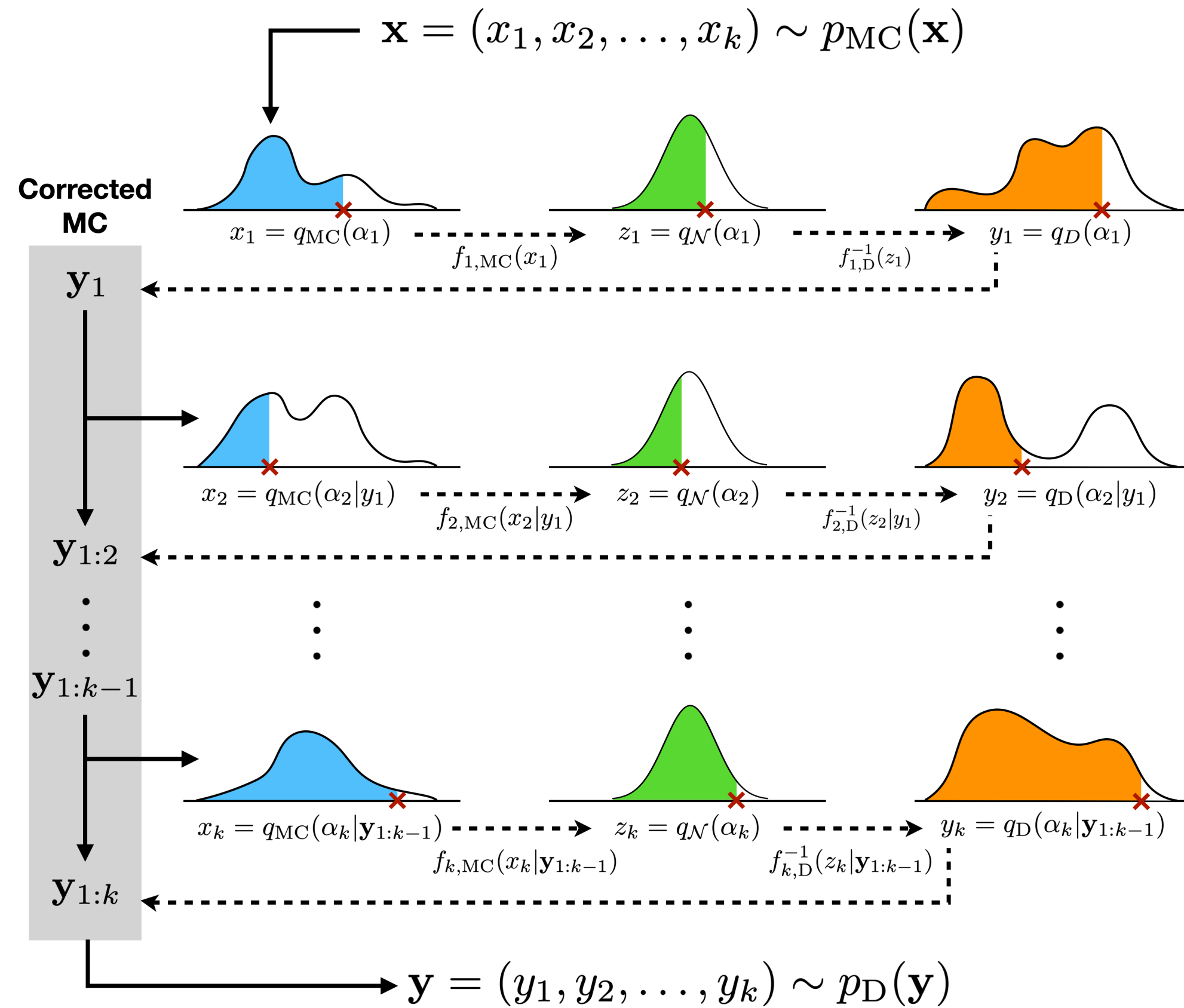
“Chained quantile morphing”

Approach used in [2208.12279](#) using BDTs

Chained quantile morphing with flows



[2309.15912](#)



Use **flows** to model the (conditional) densities & morph between data-like & simulation-like samples

Successfully applied to LHC olympics dataset; transformed high-level jet features