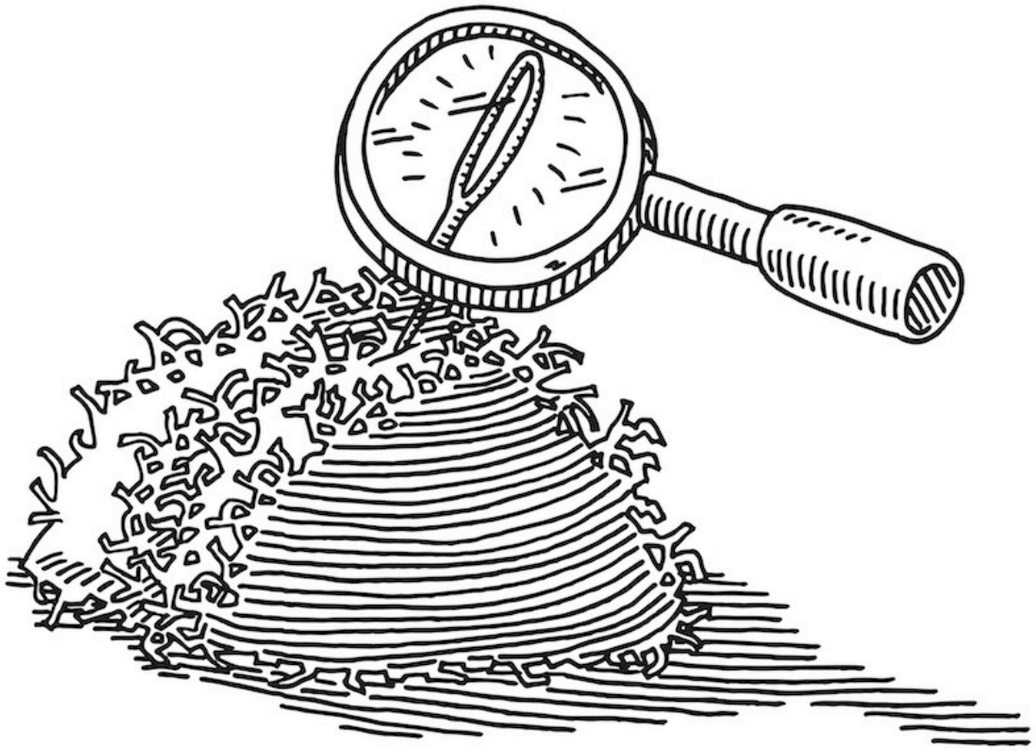
Challenges with Anomaly Detection in High Energy Physics

Based on ongoing work with
Katherine Fraser, Samuel Homiller, Rashmish Mishra, and Matthew Schwartz

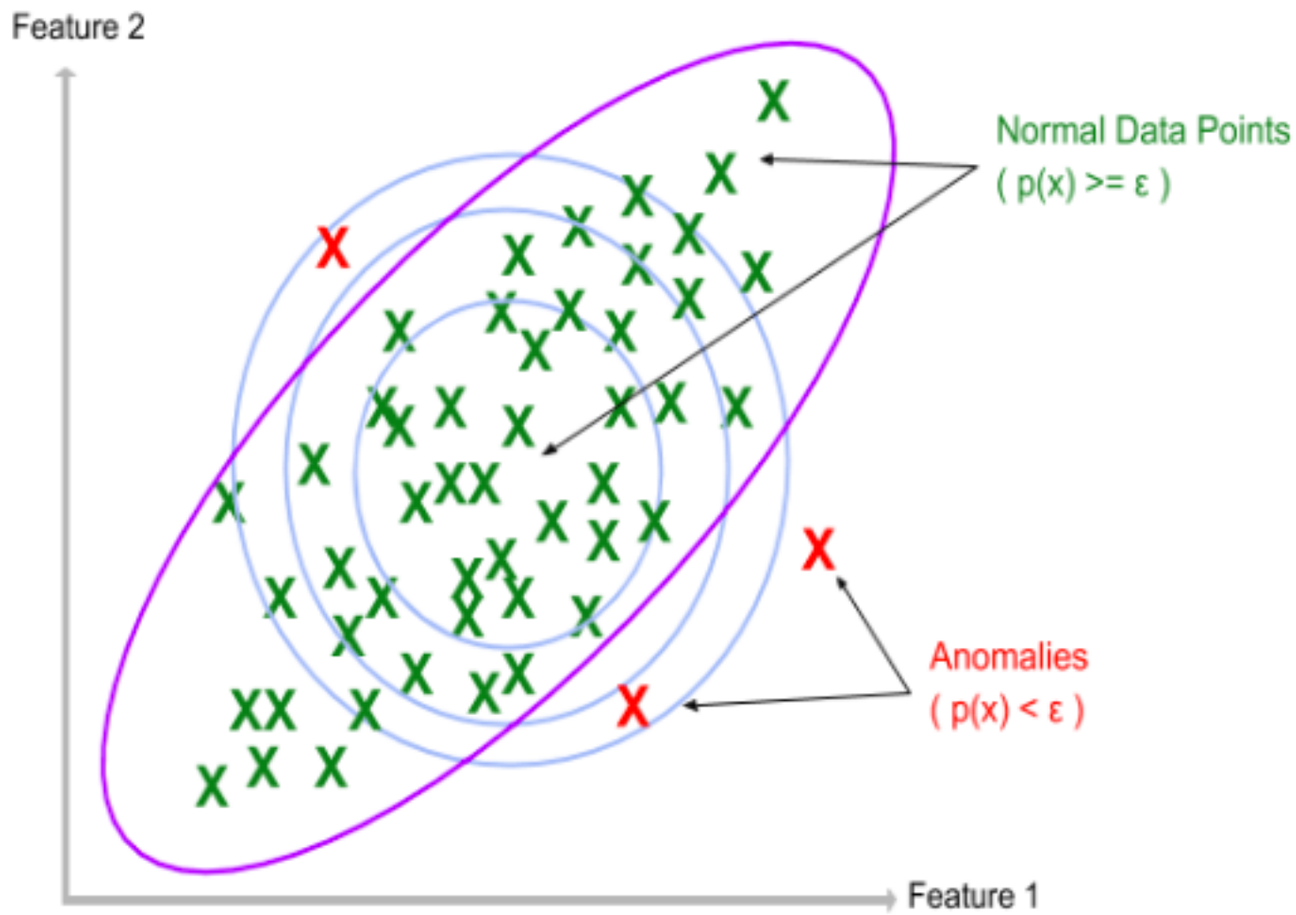
[arXiv:2110.06948]



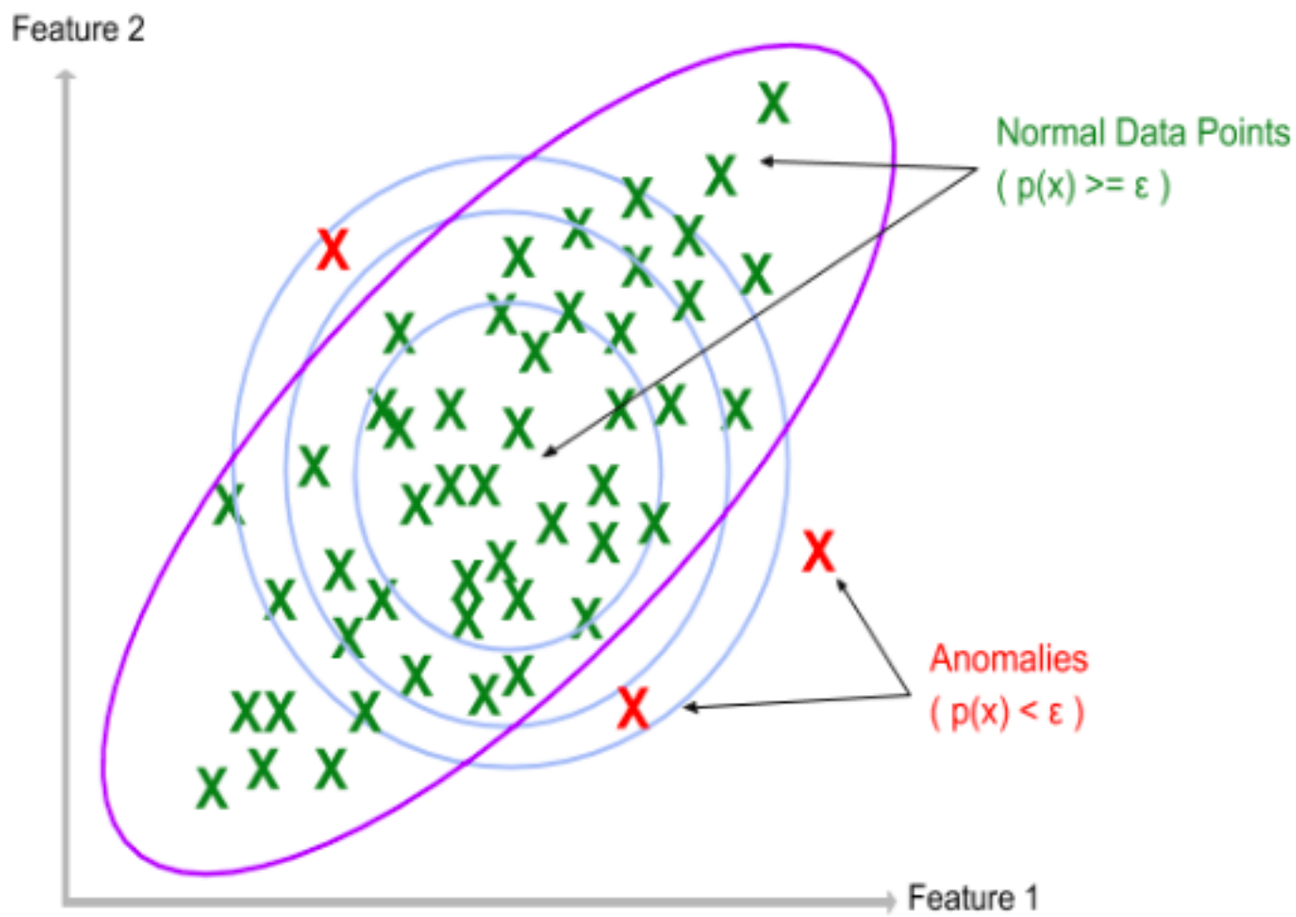
Bryan Ostdiek
Harvard University

Argonne National Laboratory
HEP Division Seminar
December 8, 2021

What is Anomaly Detection?

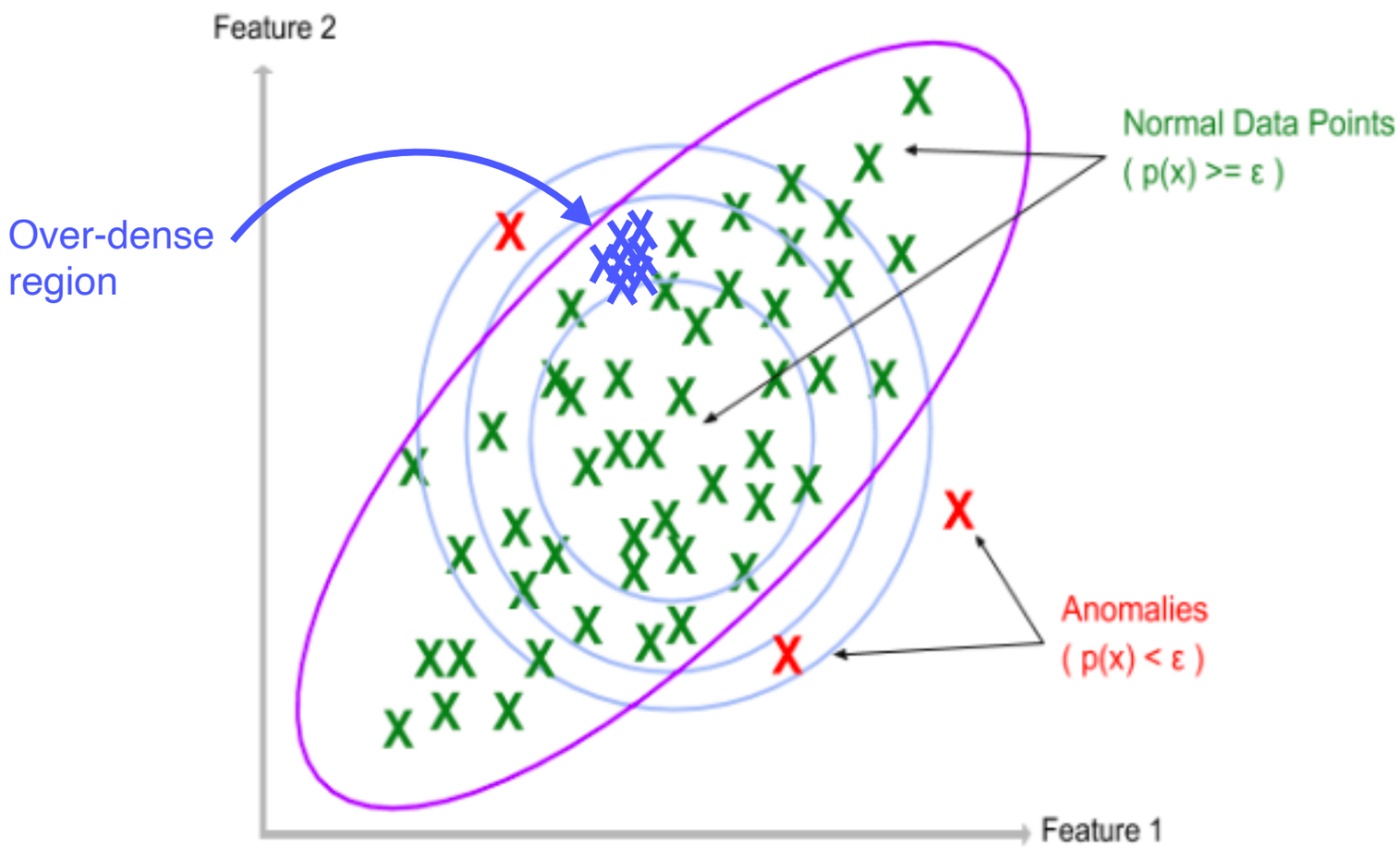


What is Anomaly Detection?



1) Events that occur with low probability (look different in some feature space)

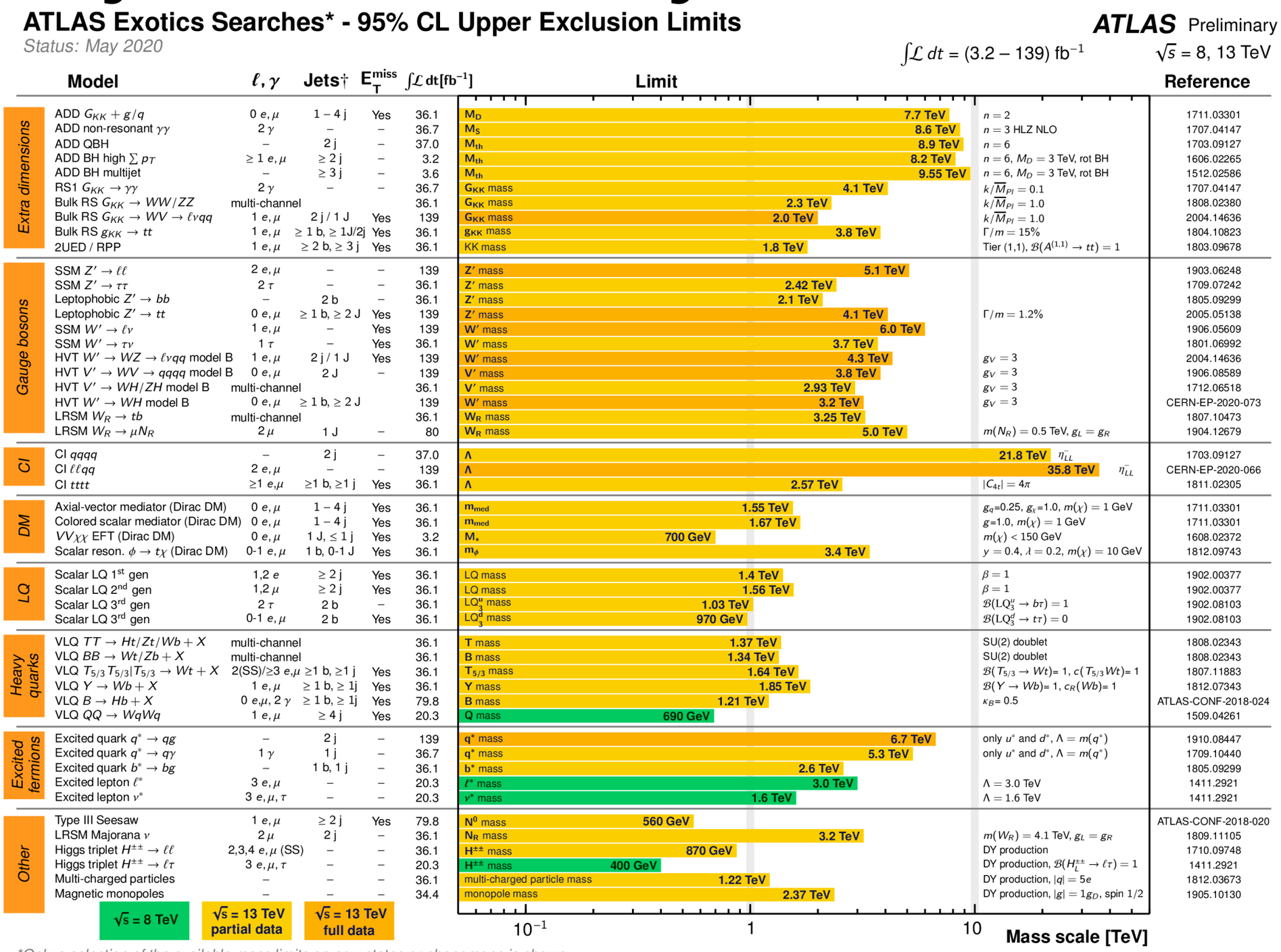
What is Anomaly Detection?



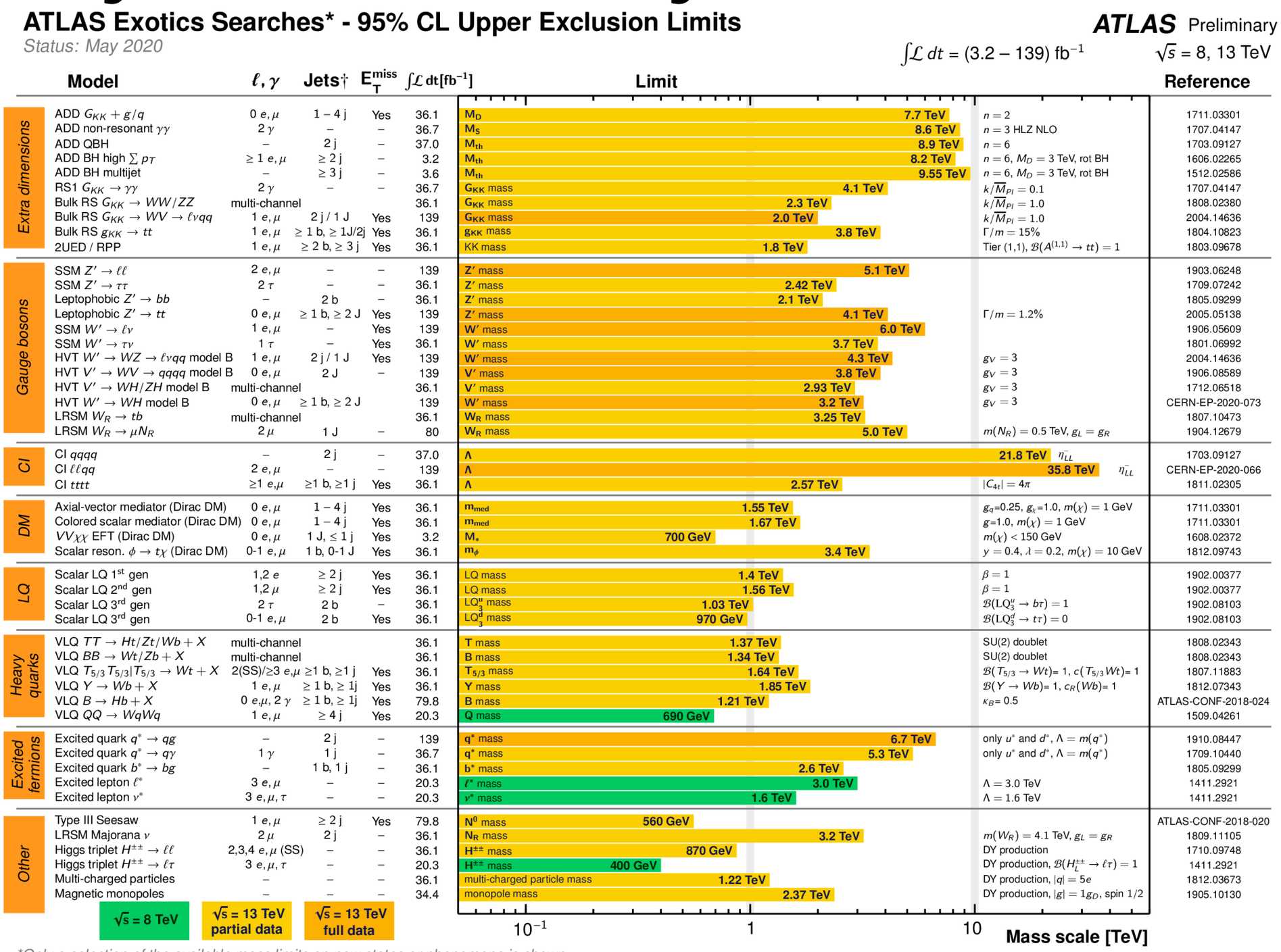
1) Events that occur with low probability (look different in some feature space)

2) Events that don't look very different, but occur at a higher rate than expected

Why Anomaly Detection?



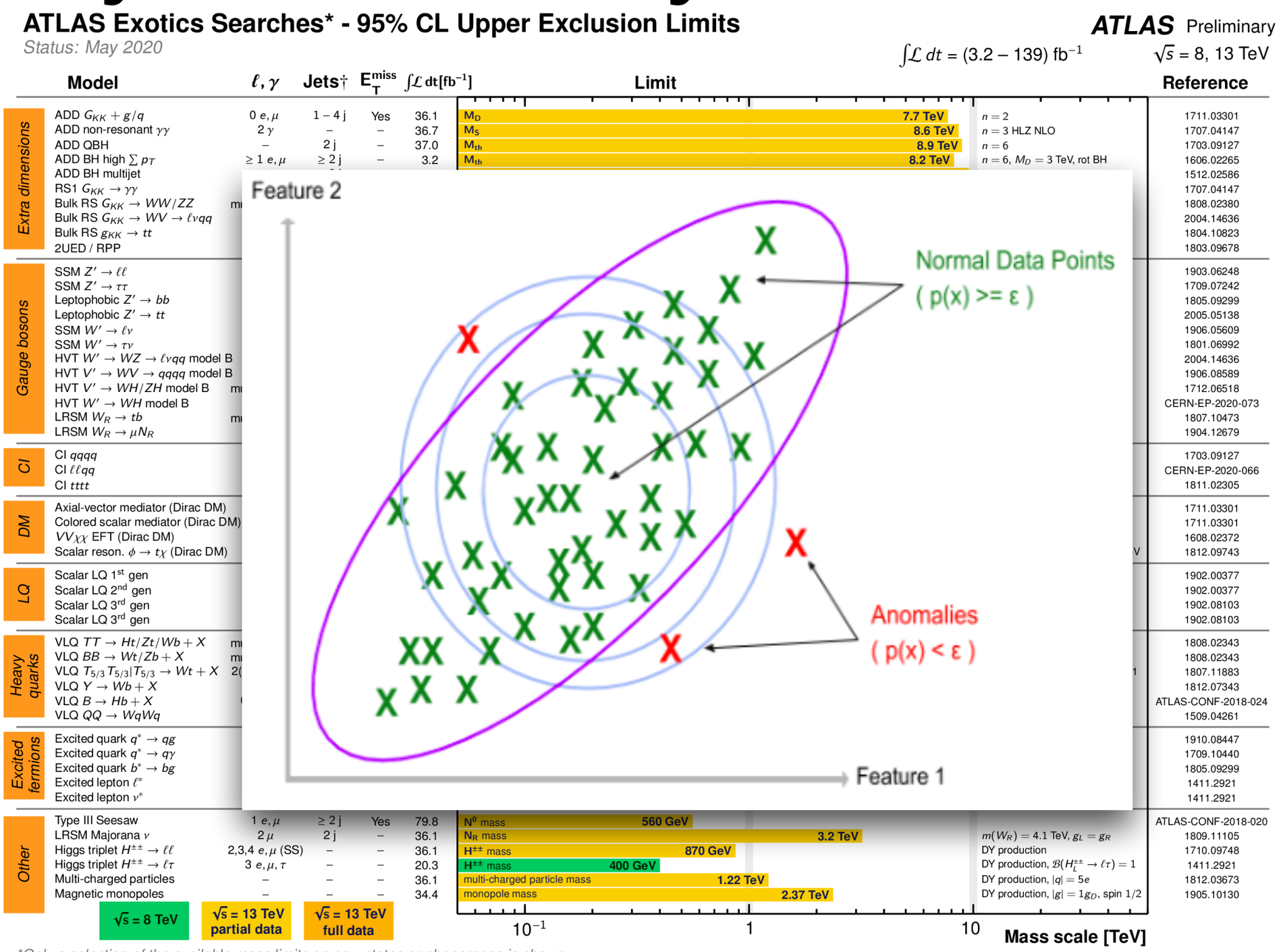
Why Anomaly Detection?



Higher rates, but hard to distinguish from SM

Look more different, but more low rates

Why Anomaly Detection?

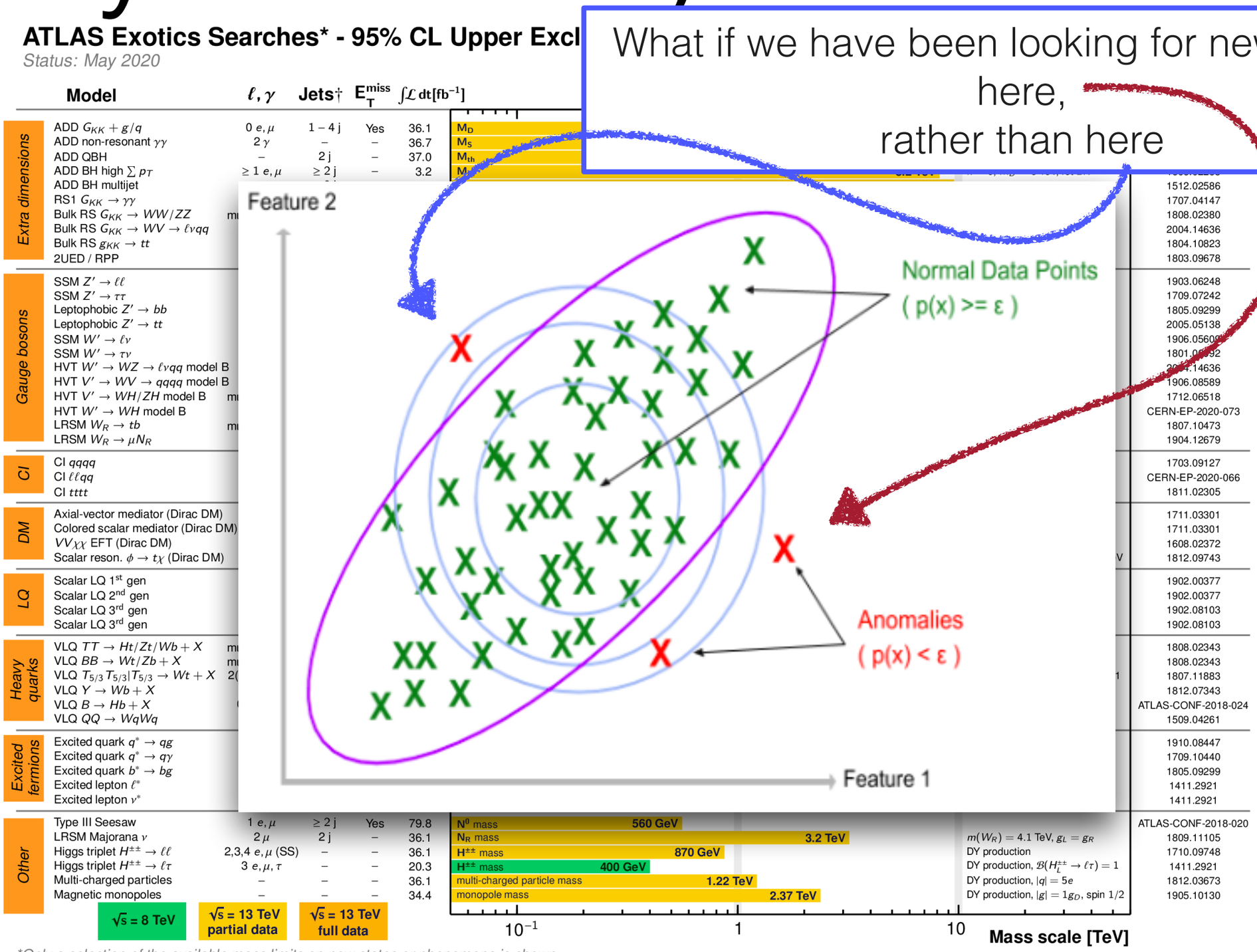


Higher rates, but hard to distinguish from SM

Look more different, but more low rates

Why Anomaly Detection?

What if we have been looking for new physics here,
rather than here



Higher rates, but hard to distinguish from SM

Look more different, but more low rates

Why Anomaly Detection?

1) Model Agnostic: We are good at looking for models we know of, but what if we don't know what we should be looking for?

2) Simulation Independent: With no signal model, it is possible to use methods directly on data from the LHC without Monte Carlo simulations.

How to detect anomalies?

Over 50 papers in HEP anomaly detection
https://iml-wg.github.io/HEPML-LivingReview/

LHC Olympics [2101.08320] focuses on finding over densities in all-hadronic events

- Black Box 1: Similar to example data: 4 methods found resonance
- Black Box 2: SM only. 4 methods claimed a resonance, 1 claimed lack-of-resonance
- Black Box 3: Correct resonance not detected by any group

Dark Machines Challenge [2105.14027] focuses on finding individual events which look different

- Train on SM-only events, apply to many different signals
- Find methods which work best for most signals
- No method finds every new physics signal

How to detect anomalies?

PHYSICAL REVIEW D **101,** 075021 (2020)

Searching for new physics with deep autoencoders

Marco Farina,[1,2] Yuichiro Nakai,[2] and David Shih[2]
[1]*C.N.Yang Institute for Theoretical Physics, Stony Brook, New York 11794, USA*
[2]*NHETC, Dept. of Physics and Astronomy Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA*

(Received 10 January 2019; accepted 27 March 2020; published 13 April 2020)

We introduce a potentially powerful new method of searching for new physics at the LHC, using autoencoders and unsupervised deep learning. The key idea of the autoencoder is that it learns to map "normal" events back to themselves, but fails to reconstruct "anomalous" events that it has never encountered before. The reconstruction error can then be used as an anomaly threshold. We demonstrate the effectiveness of this idea using QCD jets as background and boosted top jets and R-parity violating (RPV) gluino jets as signal. We show that a deep autoencoder can significantly improve signal over background when trained on backgrounds only, or even directly on data which contain a small admixture of signal. Finally, we examine the correlation of the autoencoders with jet mass and show how the jet mass distribution can be stable against cuts in reconstruction loss. This may be important for estimating QCD backgrounds from data. As a test case, we show how one could plausibly discover 400 GeV RPV gluinos using an autoencoder combined with a bump hunt in jet mass. This opens up the exciting possibility of training directly on actual data to discover new physics with no prior expectations or theory prejudice.

DOI: 10.1103/PhysRevD.101.075021

[1808.08992]

SciPost Phys. 6, 030 (2019)

QCD or what?

Theo Heimel[1], Gregor Kasieczka[2], Tilman Plehn[1⋆] and Jennifer M. Thompson[1]

1 Institut für Theoretische Physik, Universität Heidelberg, Germany
2 Institut für Experimentalphysik, Universität Hamburg, Germany

⋆ plehn@uni-heidelberg.de

Abstract

Autoencoder networks, trained only on QCD jets, can be used to search for anomalies in jet-substructure. We show how, based either on images or on 4-vectors, they identify jets from decays of arbitrary heavy resonances. To control the backgrounds and the underlying systematics we can de-correlate the jet mass using an adversarial network. Such an adversarial autoencoder allows for a general and at the same time easily controllable search for new physics. Ideally, it can be trained and applied to data in the same phase space region, allowing us to efficiently search for new physics using un-supervised learning.

[1808.08979]

1. Use *Autoecoders* for anomaly detection
2. Emphasize training directly on data, how much signal can be in training and still work?

Outline

- Review Autoencoders and their pitfalls
- Update to Variational Autoencoders (VAE)
  - Regularizes and adds structure to latent space
  - Best method for Signal A is not the best for Signal B
  - Latent space distances are correlated with “physical distances” between events
- Take distances from quintessential events
  - Faster than training VAE
  - Still hard to remain model agnostic
  - Better at “inverse problem” than VAE

The challenges are the same as the motivations: how to be model agnostic

Outline

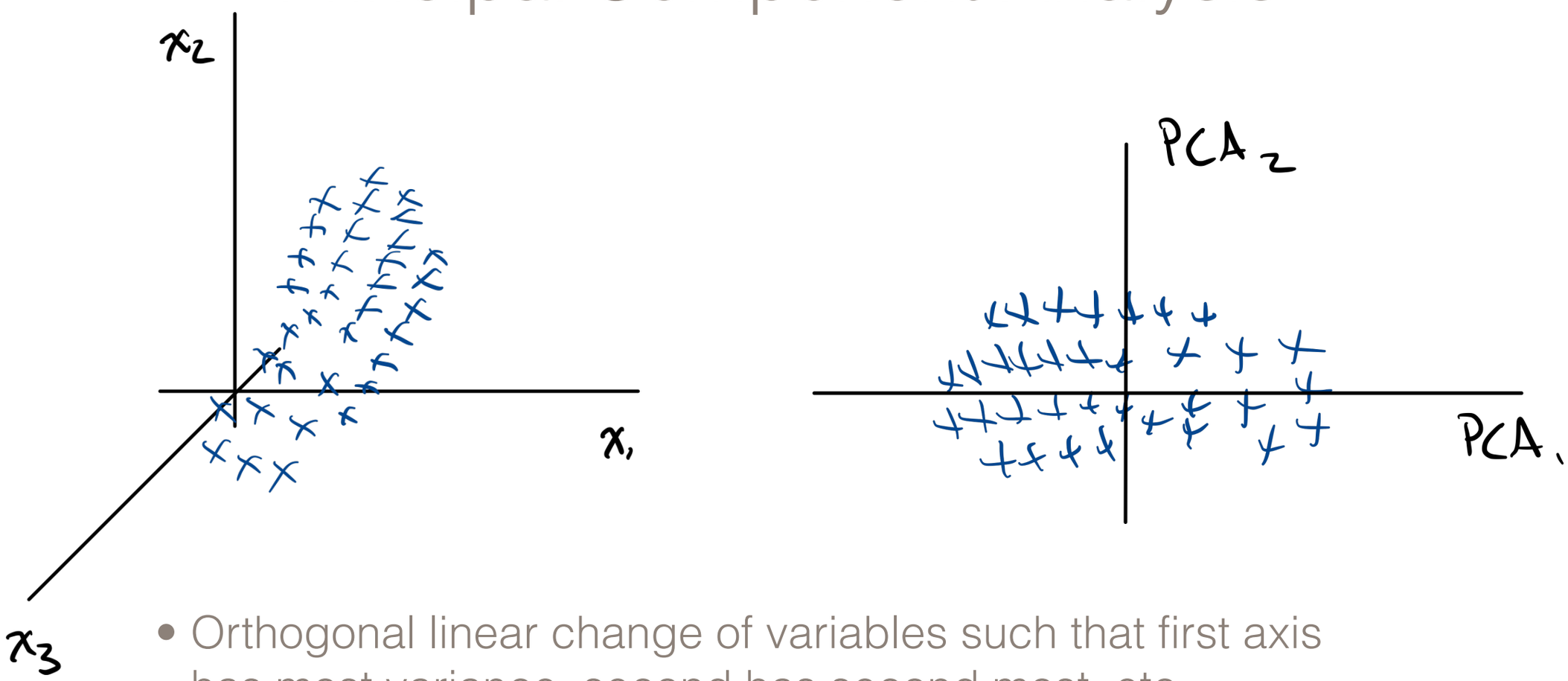
- Review Autoencoders and their pitfalls
- Update to Variational Autoencoders (VAE)
  - Regularizes and adds structure to latent space
  - Best method for Signal A is not the best for Signal B
  - Latent space distances are correlated with “physical distances” between events
- Take distances from quintessential events
  - Faster than training VAE
  - Still hard to remain model agnostic
  - Better at “inverse problem” than VAE

The challenges are the same as the motivations: how to be model agnostic

Introduction to Autoencoders

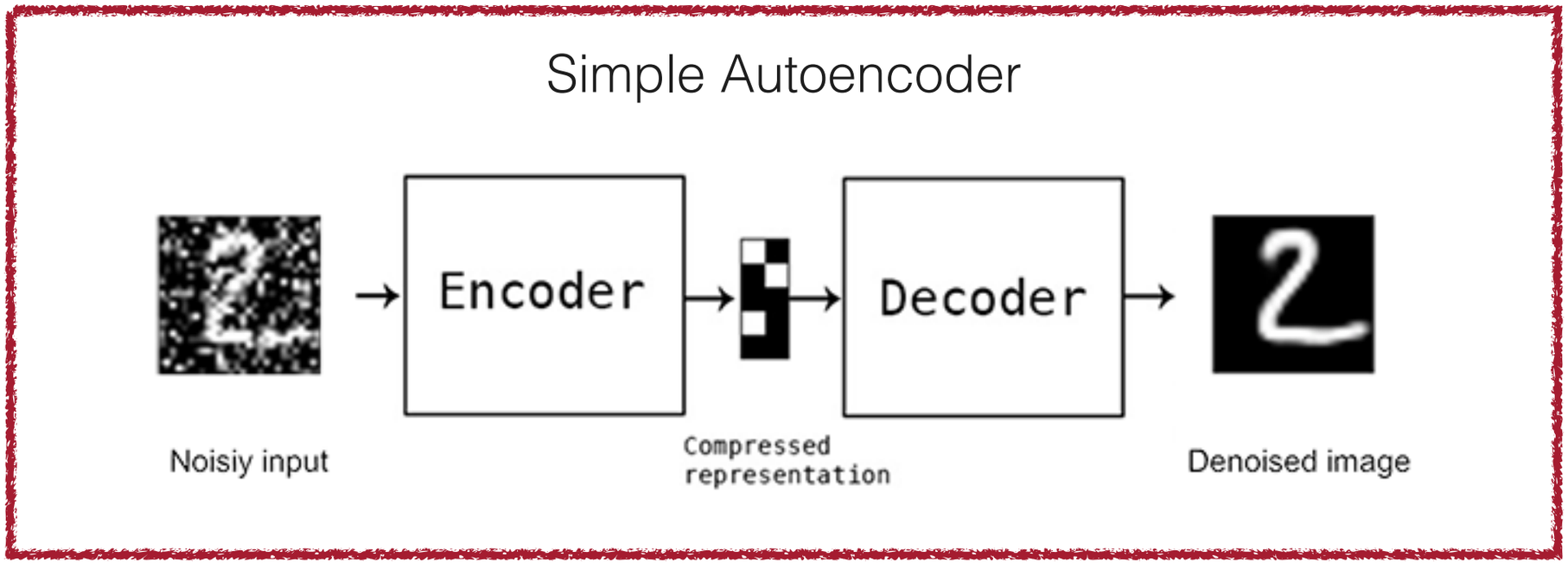
- Method for dimensionality reduction
- Focuses on important pieces of information and ignores noise

Principal Component Analysis



- Orthogonal linear change of variables such that first axis has most variance, second has second most, etc
- Which variables describe most of the variation of the data?
- Invertible (lossy if not using full basis)

Introduction to Autoencoders



- Encoding and Decoding can be non-linear
- Encoder learns what is important in the data and what is not
- Size of compressed representation chosen before training
- Compressed representation changed each training of the networks

Introduction to Autoencoders

Introduction to Autoencoders

- Unlike PCA, autoencoders need to be trained
- Need input data, output predictions, and a target
  - Target data is the same as input data

Introduction to Autoencoders

- Unlike PCA, autoencoders need to be trained
- Need input data, output predictions, and a target
  - Target data is the same as input data
- How to compare output predictions and target?
  - Use Mean Squared Error

$$d_{\mathrm{ME},2} = \mathrm{MSE} \equiv \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \left| \mathrm{D}\big(\mathrm{E}(\mathrm{x})\big)_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}} \right|^2$$

Introduction to Autoencoders

- Unlike PCA, autoencoders need to be trained
- Need input data, output predictions, and a target
  - Target data is the same as input data
- How to compare output predictions and target?
  - Use Mean Squared Error

$$d_{\mathrm{ME},2} = \mathrm{MSE} \equiv \frac{1}{n} \sum_{i=1}^{n} \left| \mathrm{D}\big(\mathrm{E}(x)\big)_i - y_i \right|^2$$

y = target (=x)
(or the input in this case)

Introduction to Autoencoders

- Unlike PCA, autoencoders need to be trained
- Need input data, output predictions, and a target
  - Target data is the same as input data
- How to compare output predictions and target?
  - Use Mean Squared Error

$$d_{\mathrm{ME},2} = \mathrm{MSE} \equiv \frac{1}{n}\sum_{i=1}^{n}\left|\mathrm{D}\big(\mathrm{E}(x)\big)_i - y_i\right|^2$$

D(E((x)) = output of decoded, encoded data
use f(x) for rest of talk

y = target (=x)
(or the input in this case)

Introduction to Autoencoders

- Unlike PCA, autoencoders need to be trained
- Need input data, output predictions, and a target
  - Target data is the same as input data
- How to compare output predictions and target?
  - Use Mean Squared Error

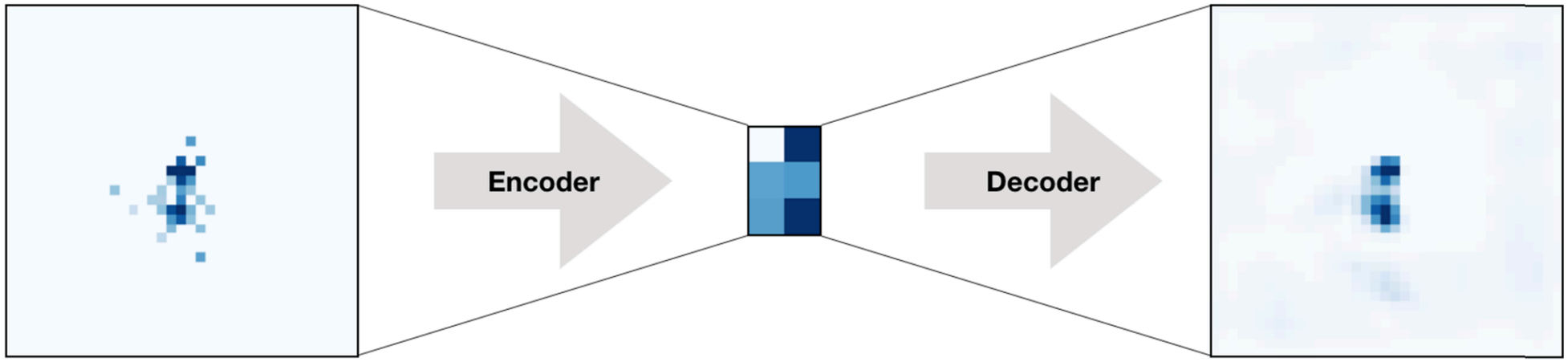
Take the average over all dimensions of the data vector

$$d_{\mathrm{ME},2} = \mathrm{MSE} \equiv \frac{1}{n}\sum_{i=1}^{n}\left|\mathrm{D}\big(\mathrm{E}(x)\big)_i - y_i\right|^2$$

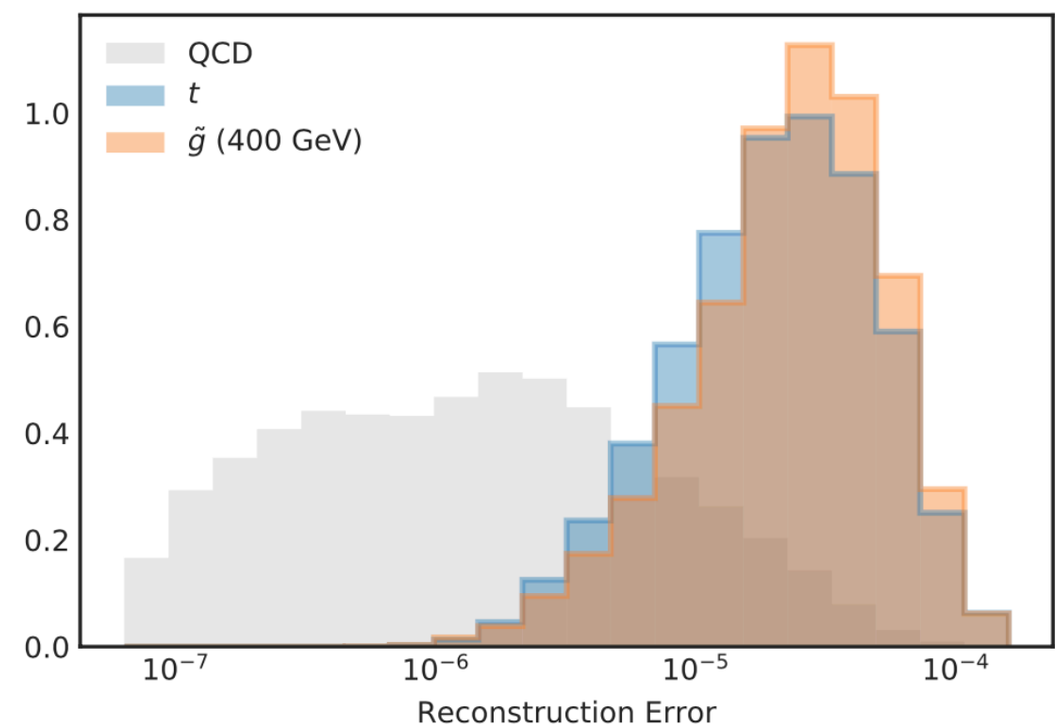
D(E((x)) = output of decoded, encoded data
use f(x) for rest of talk

y = target (=x)
(or the input in this case)

Autoencoders for Anomaly Detection



- Networks are trained to minimize the reconstruction error of events from SM background
- The encoding-decoding of BSM events will have larger reconstruction errors



[1808.08992]

Challenges for Autoencoders

- The encoding-decoding of BSM events will have larger reconstruction errors

Challenges for Autoencoders

- ~~The encoding-decoding of BSM events will have larger reconstruction errors~~

Challenges for Autoencoders

- ~~The encoding-decoding of BSM events will have larger reconstruction errors~~

"Inverse problem": The signal is less complex than the background

Ex: Use top jets a background and treat QCD jets as anomalous

Challenges for Autoencoders

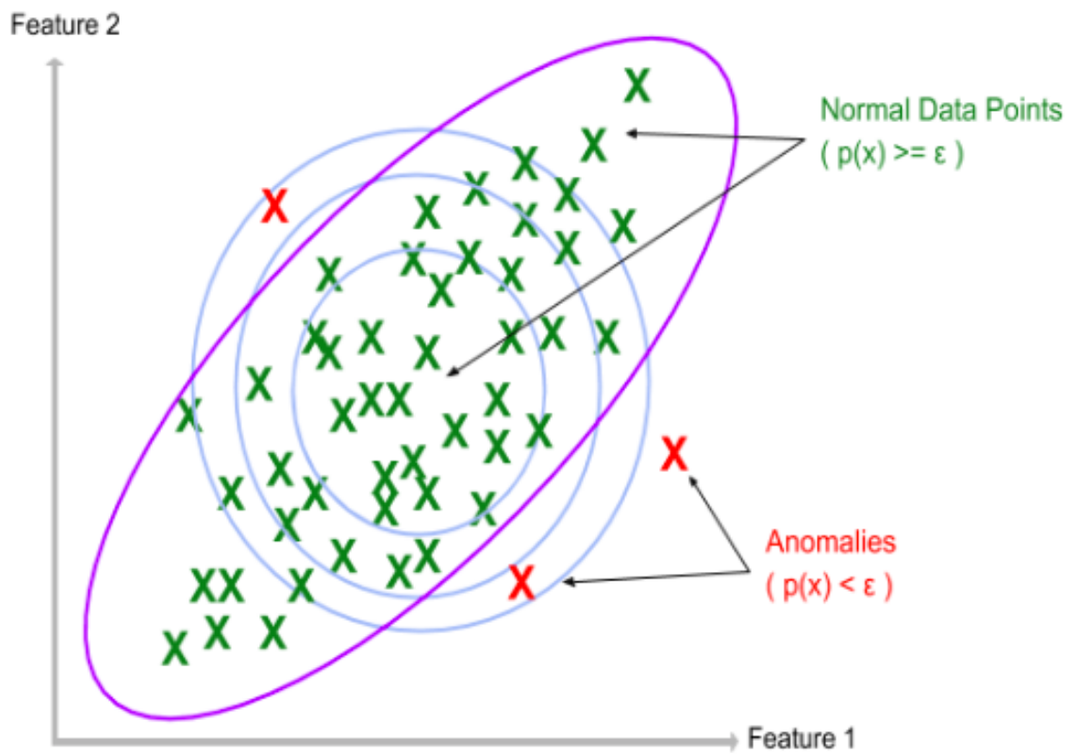
- ~~The encoding-decoding of BSM events will have larger reconstruction errors~~

"Inverse problem": The signal is less complex than the background

Ex: Use top jets a background and treat QCD jets as anomalous

"Topological Obstructions to Autoencoding," Batson, Grace Haaf, Kahn, and Roberts [2102.08380]

Challenges for Autoencoders



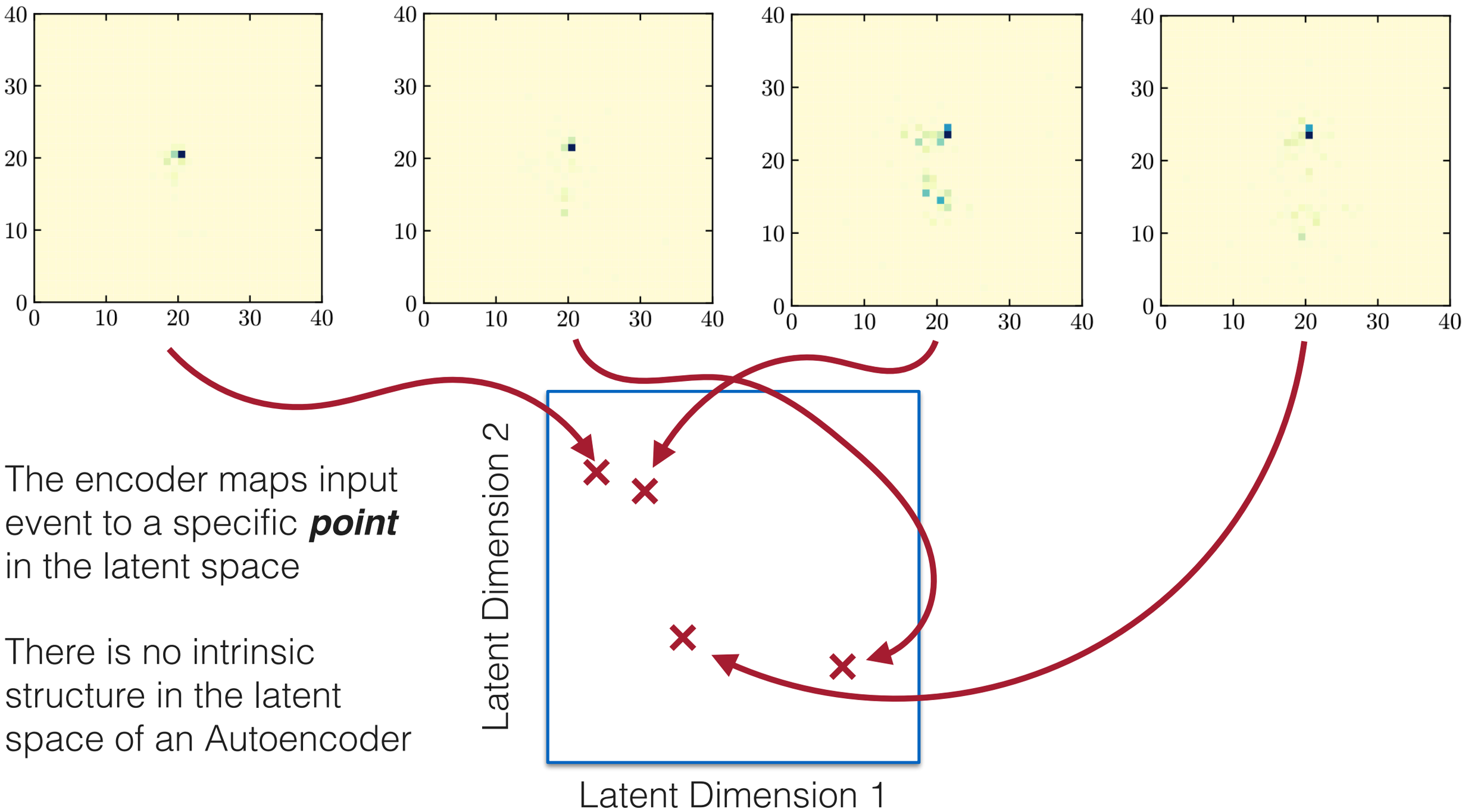
Do autoencoders fit with this picture of anomaly detection?

- Delicate balance between reconstructing SM well, but anomalous data poorly
- Latent space doesn't have structure, can't assign probabilistic interpretation
- Is MSE the right metric to use for reconstruction?

Variational Autoencoders (VAE)

- Review Autoencoders and their pitfalls
- Update to Variational Autoencoders (VAE)
    - Regularizes and adds structure to latent space
    - Best method for Signal A is not the best for Signal B
    - Latent space distances are correlated with “physical distances” between events
- Take distances from quintessential events
    - Faster than training VAE
    - Still hard to remain model agnostic
    - Better at “inverse problem” than VAE

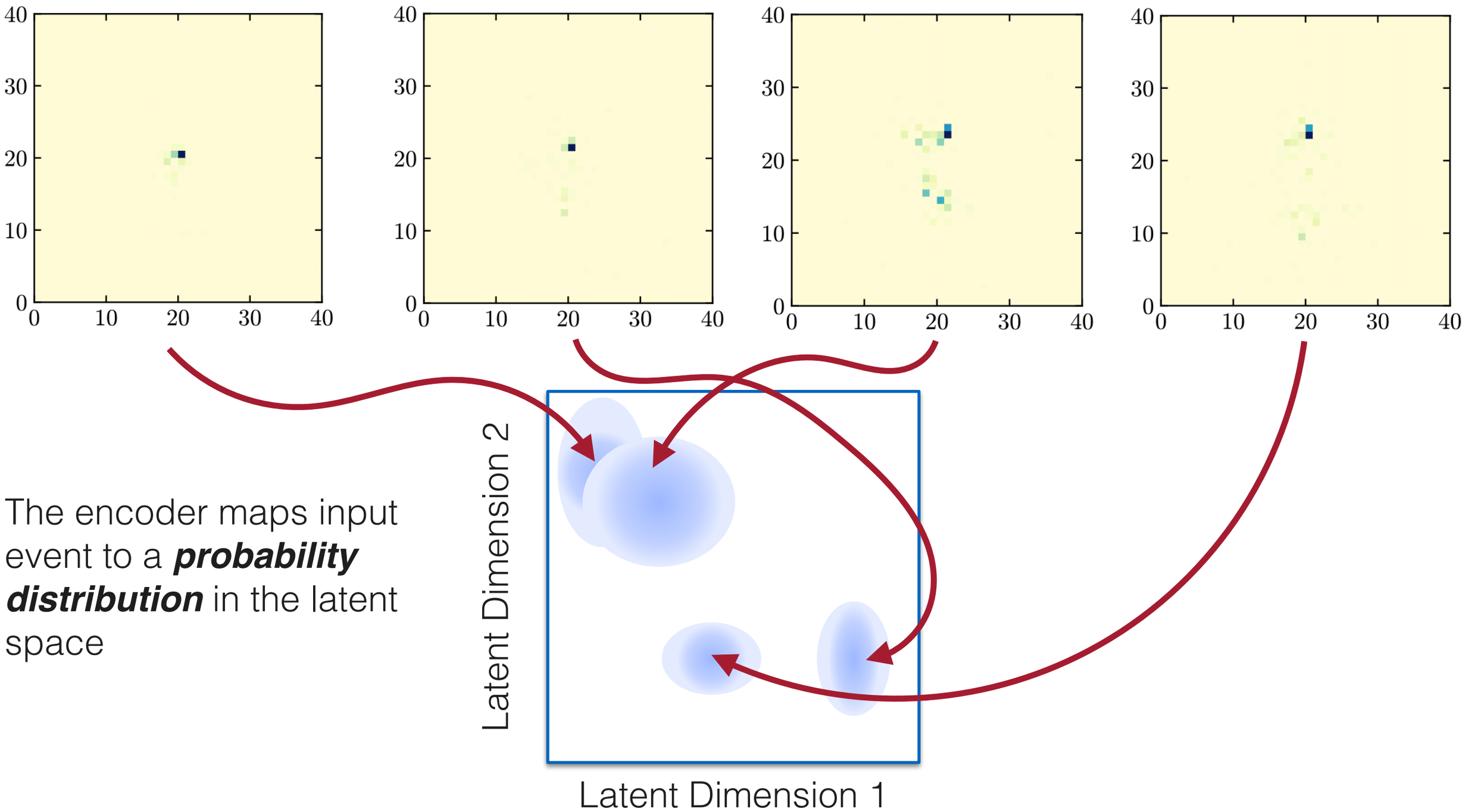
Plain Autoencoders



The encoder maps input event to a specific ***point*** in the latent space

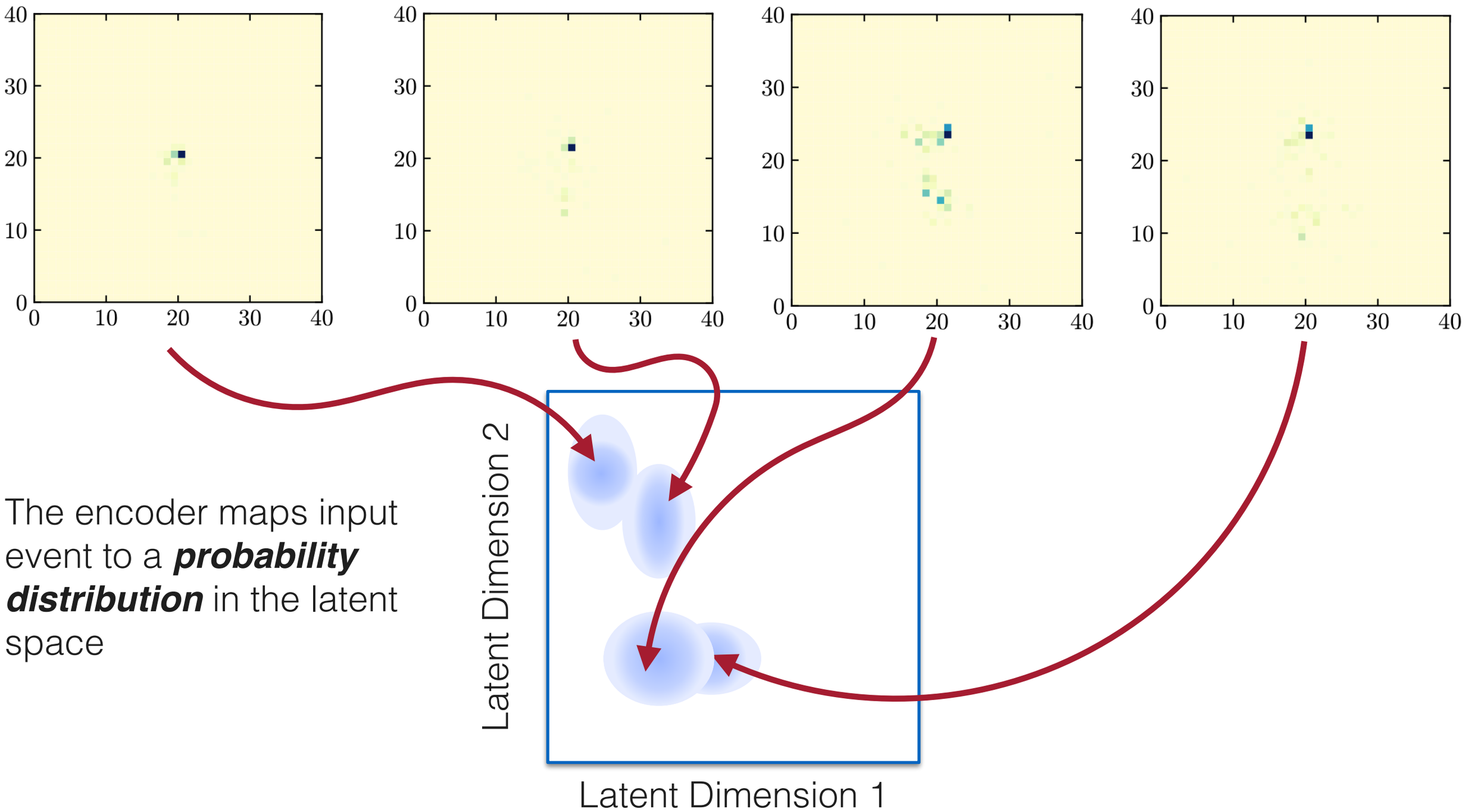
There is no intrinsic structure in the latent space of an Autoencoder

Variational Autoencoders



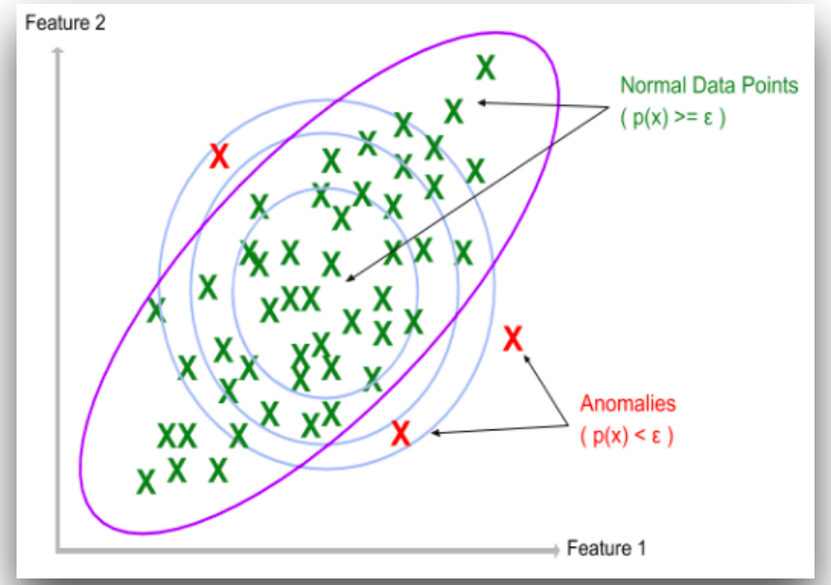
The encoder maps input event to a ***probability distribution*** in the latent space

Variational Autoencoders



The encoder maps input event to a ***probability distribution*** in the latent space

Variational Autoencoders



Add a probabilistic interpretation

How likely is a detector event, given a point in latent space?

Model each detector pixel as a Gaussian:

$$p(E|z) \propto \exp\left(\frac{-\left(E - D(z)\right)^2}{2\sigma^2}\right)$$

$$x = \{E_1, E_2, \cdots, E_N\}$$

$$p(x|z) = \prod_{i=1}^{N} p(E_i|z)$$

Probability of the observing an event

$$p(x) = \int p(x|z)p(z)dz \equiv \mathbb{E}_{p(z)}\left[p(x|z)\right]$$

What is the latent space to integrate over?

Variational Autoencoders

Probability of the observing an event

$$p(x) = \int p(x|z)p(z)dz \equiv \mathbb{E}_{p(z)}\left[p(x|z)\right]$$

What is the latent space to integrate over?

Use the encoder to learn/sample the latent space

$$p(x) = \int p(x|z)p(z)\frac{q(z|x)}{q(z|x)}dz = \mathbb{E}_{q(z|x)}\left[p(x|z)\frac{p(z)}{q(z|x)}\right]$$

Variational Autoencoders

Probability of the observing an event

$$p(x) = \int p(x|z)p(z)dz \equiv \mathbb{E}_{p(z)}\left[p(x|z)\right]$$

What is the latent space to integrate over?

Use the encoder to learn/sample the latent space

$$p(x) = \int p(x|z)p(z)\frac{q(z|x)}{q(z|x)}dz = \mathbb{E}_{q(z|x)}\left[p(x|z)\frac{p(z)}{q(z|x)}\right]$$

Take log likelihood and use Jensen's inequality to move log inside expectation value

$$\log p(x) \geq \mathbb{E}_{q(z|x)}\left[\log\left(p(x|z)\right) + \log\left(\frac{p(z)}{q(z|x)}\right)\right]$$

Variational Autoencoders

Probability of the observing an event

$$p(x) = \int p(x|z)p(z)dz \equiv \mathbb{E}_{p(z)}\left[p(x|z)\right]$$

What is the latent space to integrate over?

Use the encoder to learn/sample the latent space

$$p(x) = \int p(x|z)p(z)\frac{q(z|x)}{q(z|x)}dz = \mathbb{E}_{q(z|x)}\left[p(x|z)\frac{p(z)}{q(z|x)}\right]$$

Take log likelihood and use Jensen's inequality to move log inside expectation value

$$\log p(x) \geq \mathbb{E}_{q(z|x)}\left[\log\left(p(x|z)\right) + \log\left(\frac{p(z)}{q(z|x)}\right)\right]$$

~ MSE of encoded-decoded event

Variational Autoencoders

Probability of the observing an event

$$p(x) = \int p(x|z)p(z)dz \equiv \mathbb{E}_{p(z)}\left[p(x|z)\right]$$

What is the latent space to integrate over?

Use the encoder to learn/sample the latent space

$$p(x) = \int p(x|z)p(z)\frac{q(z|x)}{q(z|x)}dz = \mathbb{E}_{q(z|x)}\left[p(x|z)\frac{p(z)}{q(z|x)}\right]$$

Take log likelihood and use Jensen's inequality to move log inside expectation value

$$\log p(x) \geq \mathbb{E}_{q(z|x)}\left[\log\left(p(x|z)\right) + \log\left(\frac{p(z)}{q(z|x)}\right)\right]$$

~ MSE of encoded-decoded event

Kullback-Leibler Divergence (KLD) between encoded representation and latent prior

Summary of Autoencoders

Plain autoencoders: no sampling, $L = \mathrm{MSE}$

Variation autoencoders: sampling in latent space, $L = \mathrm{MSE} + \mathrm{KLD}$

Sampling adds/forces structure on the latent space.
KLD term helps to regularize and adds to
probabilistic interpretation

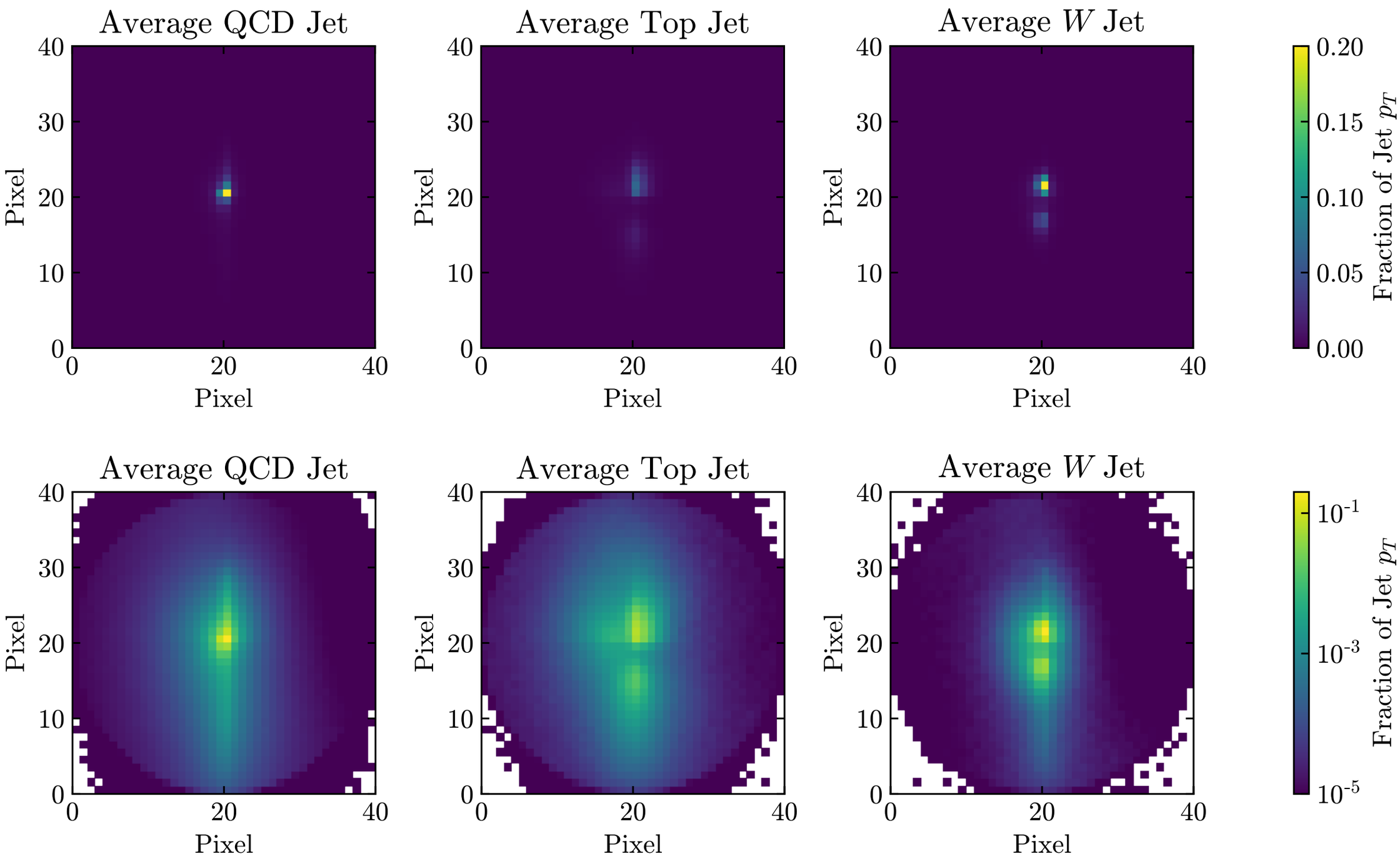
In practice, these terms may be too far apart,
introduce a scaling between them

$$L = (1 - \beta)\,\mathrm{MSE} + \beta\,\mathrm{KLD}$$

VAEs in Action

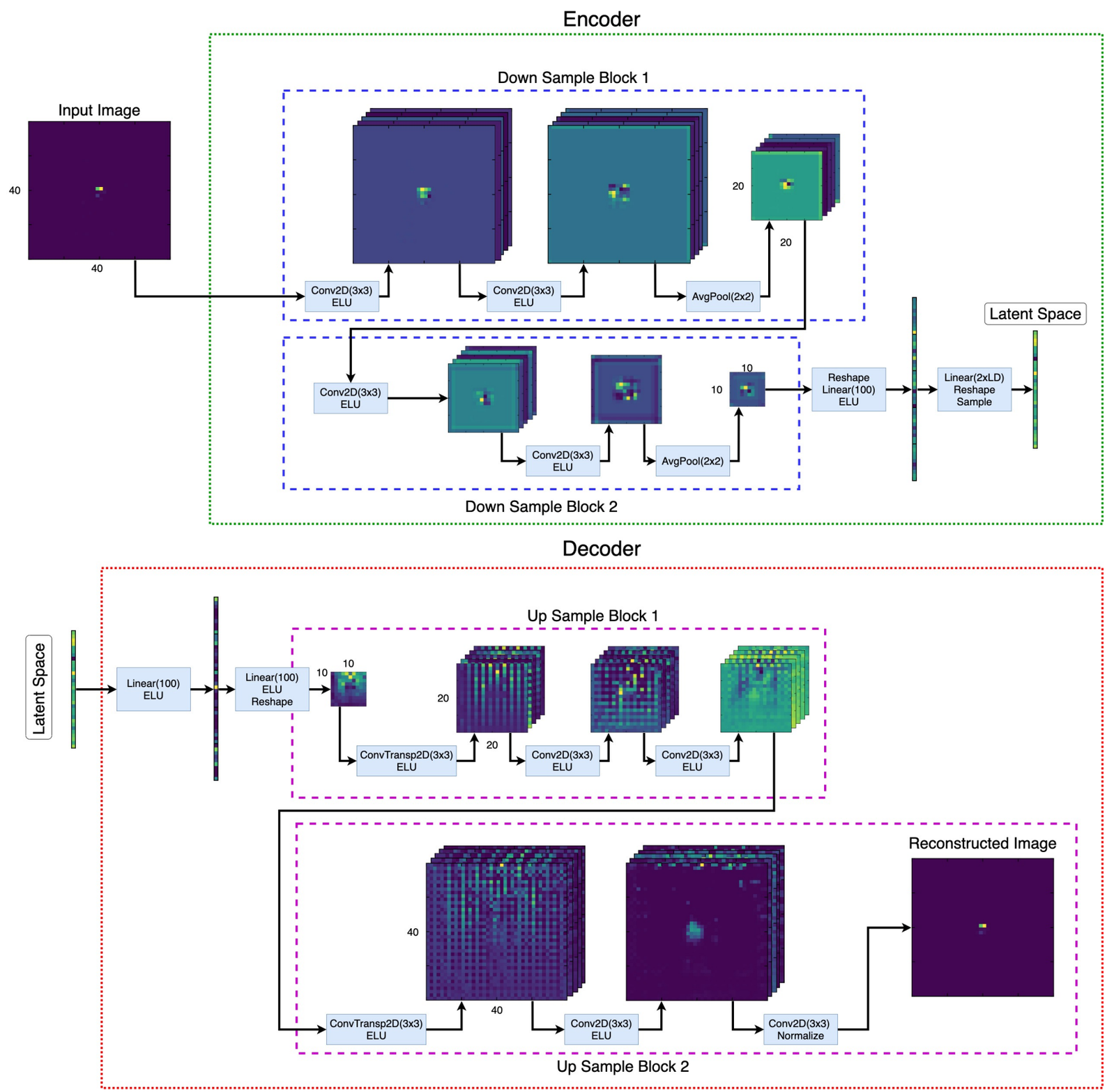
The Data

Cheng, Arguin, Leissner-Martin, Pilette, and Golling [2007.01850]



https://zenodo.org/record/4614656

The Networks



<u>Keep fixed</u>:

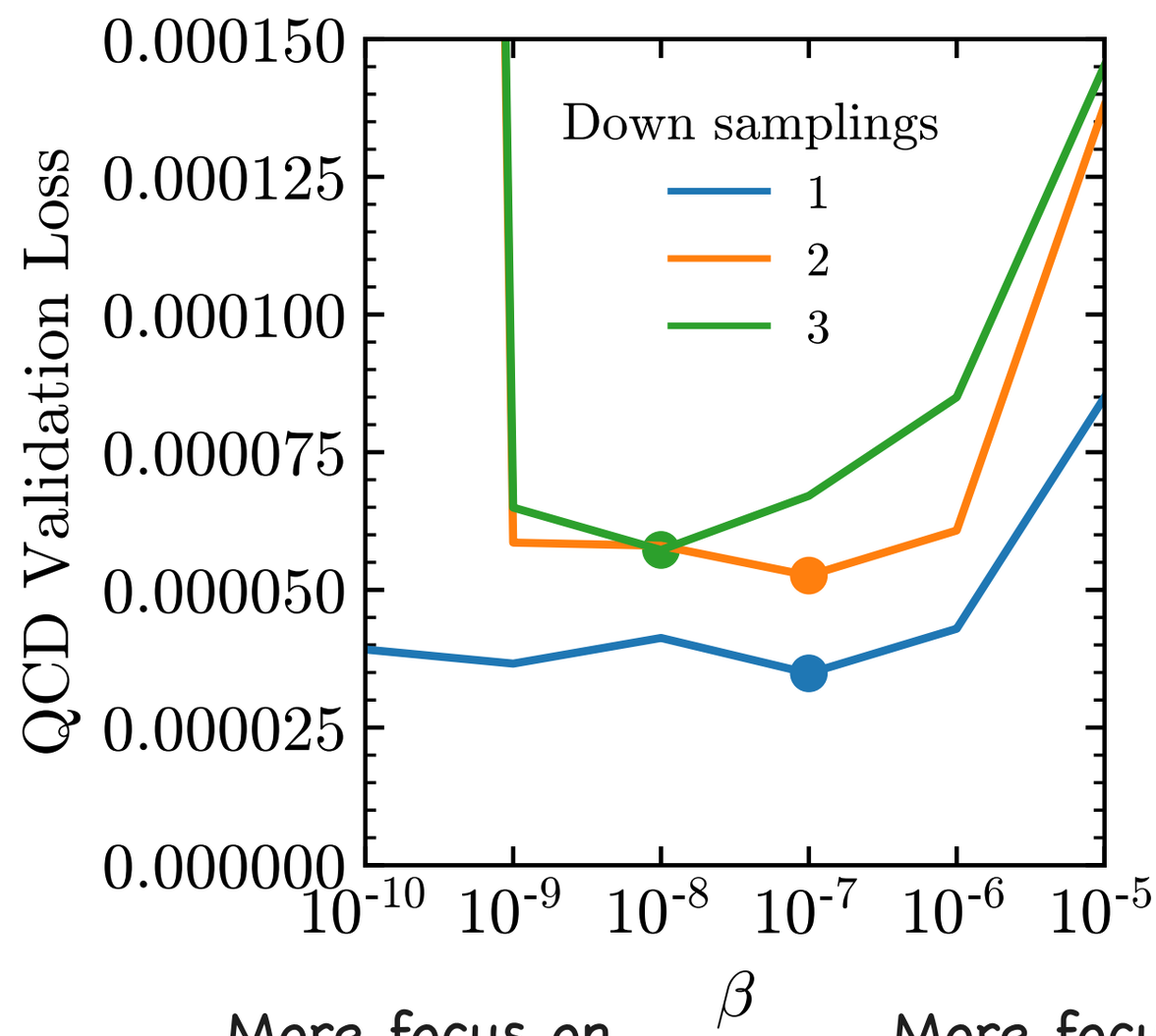
- Number of filters
- CNN Kernel size
- Latent space size

<u>Experiments</u>:

- β (relative importance of reconstruction vs. KLD)
- Number of Down Sample Blocks

Initial Results

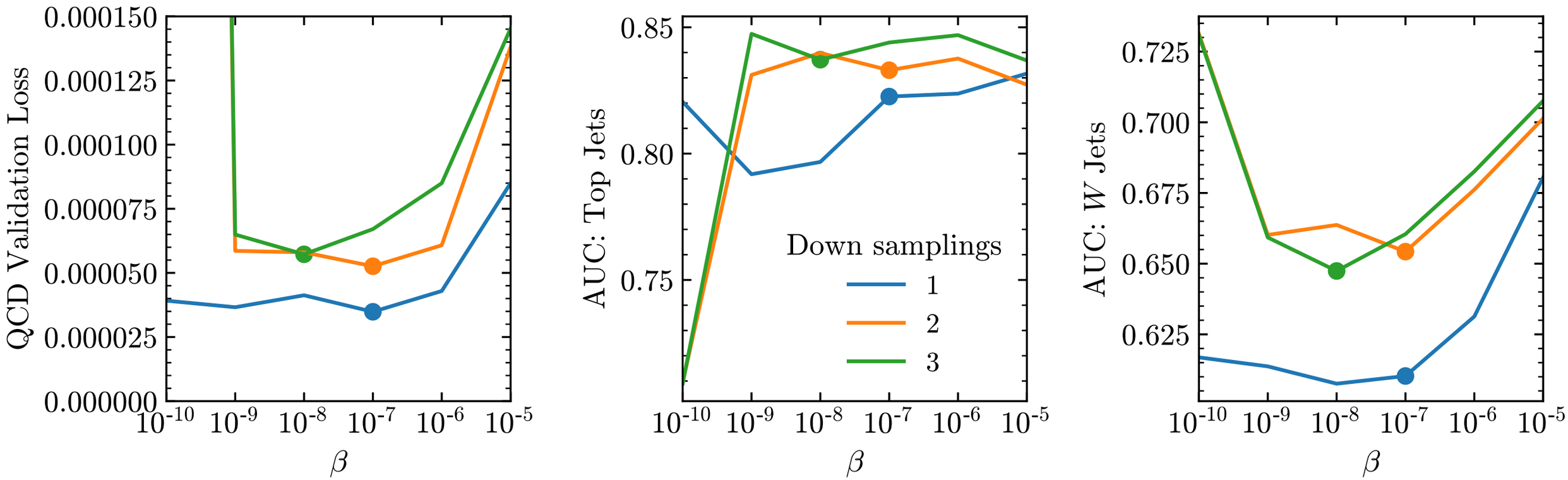
- Train on 600k QCD events
- Examine validation loss on independent 50k QCD events
- Train until validation loss stops improving
- Which setup is best?



More focus on reconstruction

More focus on latent space

Initial Results



- Apply network to 100k QCD test events and 100k Top/W jets
- AUC of 0.5~random guess, 1.0~perfect classifier
- Networks with best loss do not have the best AUC!
- Best networks for W detection are not the best for top detection

Choice of Metric

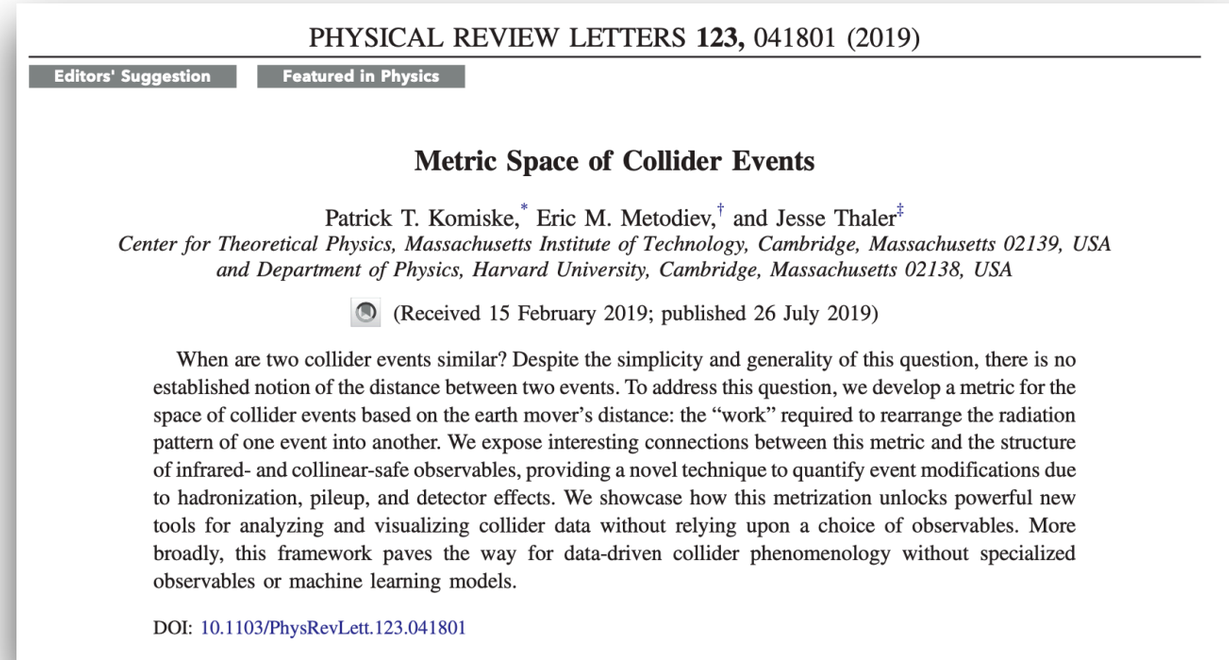
Training and inference using the same metric

$$L = (1 - \beta)\,\mathrm{MSE} + \beta\,\mathrm{KLD}$$

Helps to learn the probability distribution of QCD events, but maybe it doesn't generalize well to anomalous events

Use optimal transport distance between reconstruction and initial event?

Choice of Metric



PHYSICAL REVIEW LETTERS **123,** 041801 (2019)

Editors' Suggestion Featured in Physics

Metric Space of Collider Events

Patrick T. Komiske,* Eric M. Metodiev,† and Jesse Thaler‡
Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA and Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA

(Received 15 February 2019; published 26 July 2019)

When are two collider events similar? Despite the simplicity and generality of this question, there is no established notion of the distance between two events. To address this question, we develop a metric for the space of collider events based on the earth mover's distance: the "work" required to rearrange the radiation pattern of one event into another. We expose interesting connections between this metric and the structure of infrared- and collinear-safe observables, providing a novel technique to quantify event modifications due to hadronization, pileup, and detector effects. We showcase how this metrization unlocks powerful new tools for analyzing and visualizing collider data without relying upon a choice of observables. More broadly, this framework paves the way for data-driven collider phenomenology without specialized observables or machine learning models.

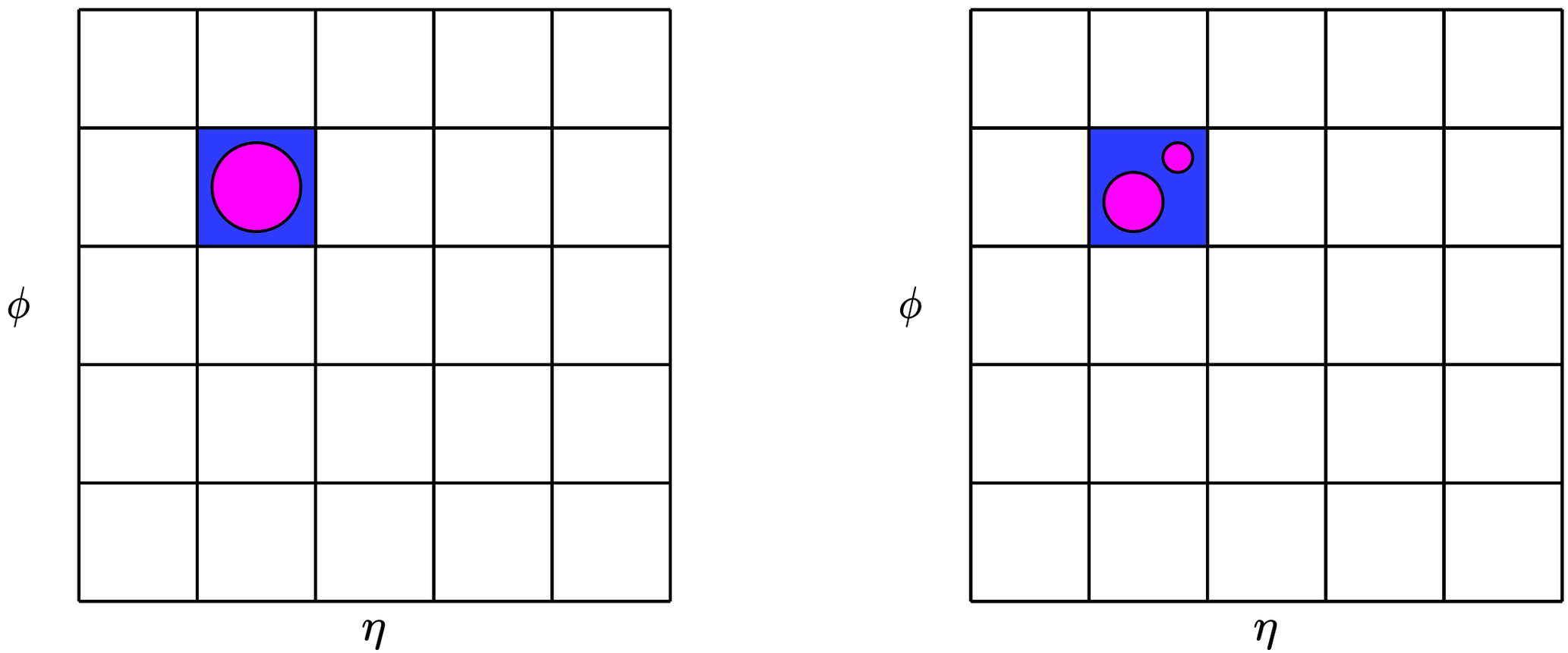
DOI: 10.1103/PhysRevLett.123.041801

$$d^{(p)}_{\text{Wass}} = \min_{f_{ij}>0} \sum_{ij} f_{ij} \frac{\theta^{(p)}_{ij}}{R} + \left| \sum_i E_i - \sum_j E'_j \right| \to 0$$

$$\sum_j f_{ij} \leq E_i \qquad \sum_i f_{ij} \leq E'_j$$

Normalized images,
"balanced" optimal transport

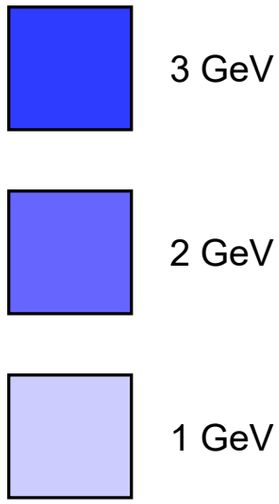
Choice of Metric



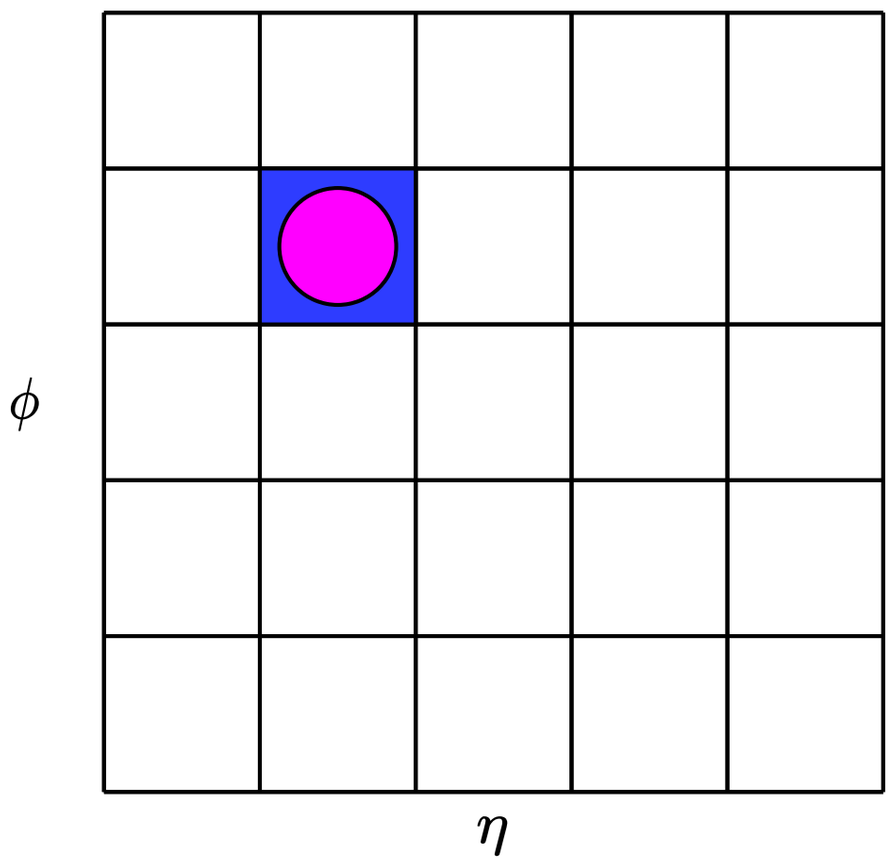
Pixelation reduces sensitivity to some splittings

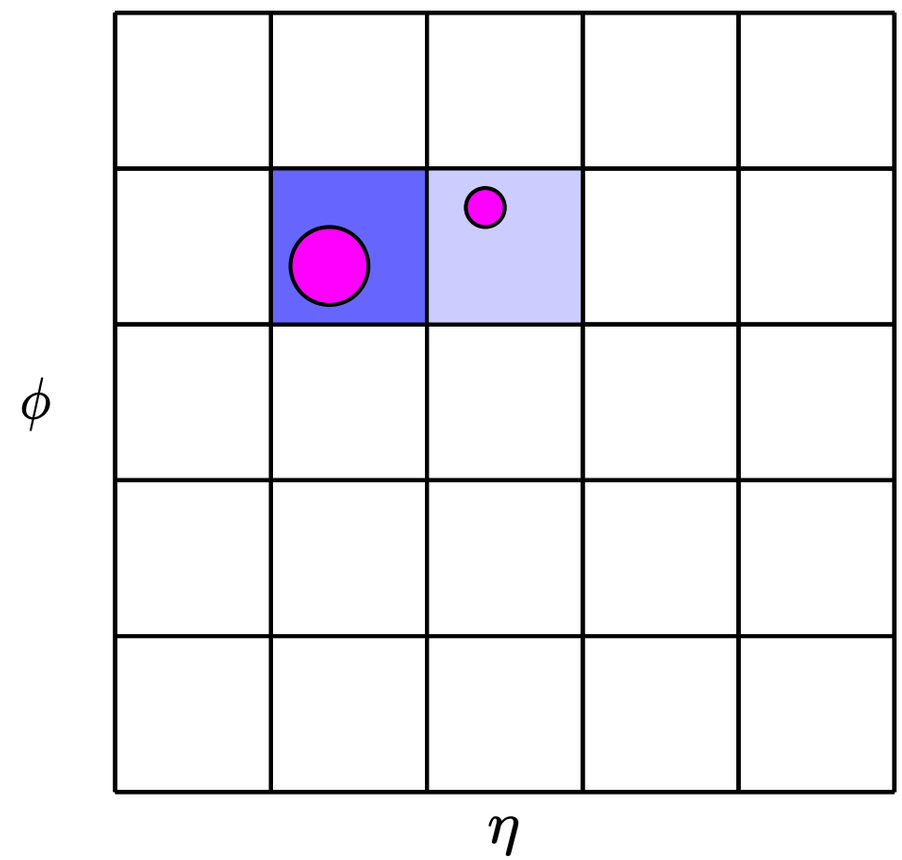
$$\mathrm{MSE} = \frac{1}{25}\left[(0-0)^2 + \cdots + (3-3)^2 + (0-0)^2 + \cdots\right]$$

$$d^{(p)}_{\mathrm{Wass}} \propto 3\mathrm{GeV}\frac{0^{(\mathrm{p})}}{\mathrm{R}}$$



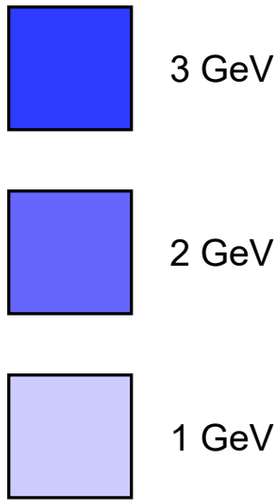
Choice of Metric



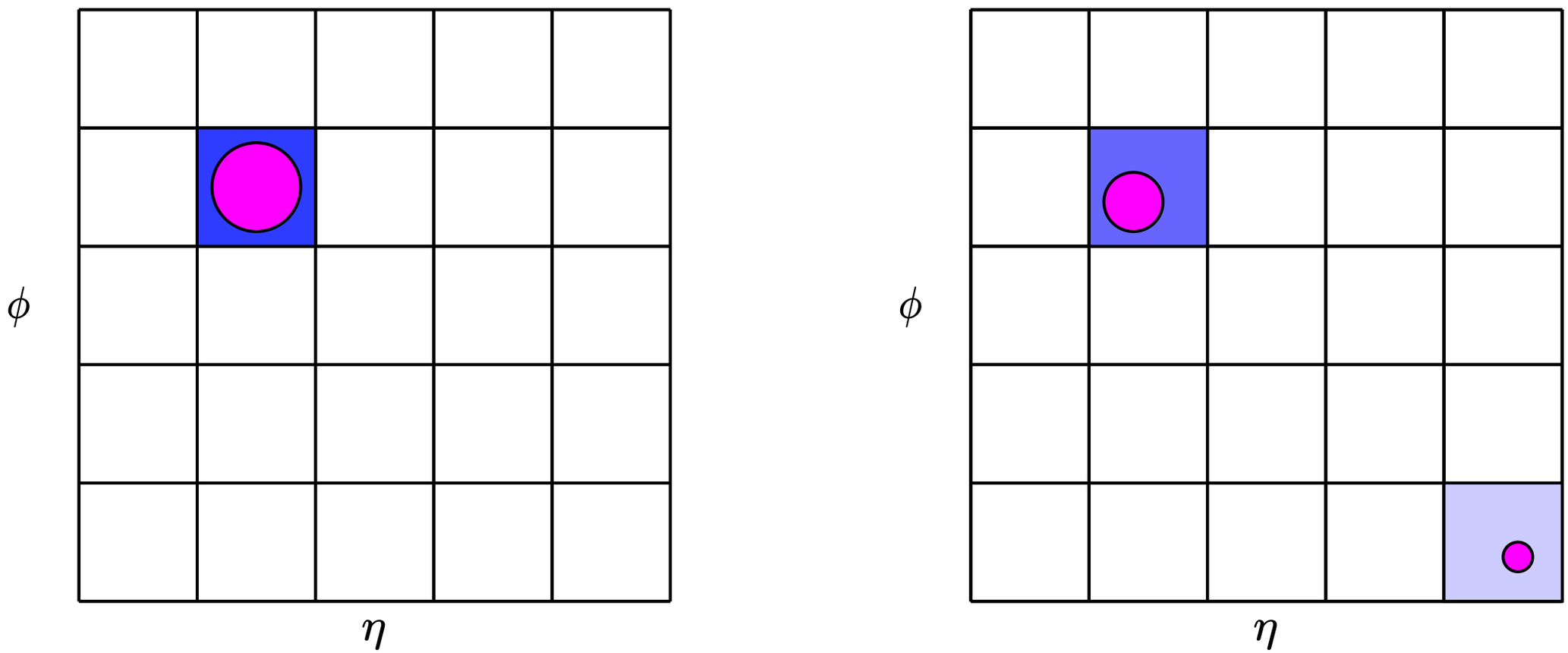


$$\mathrm{MSE} = \frac{1}{25}\left[(0-0)^2 + \cdots + (3-2)^2 + (0-1)^2 + \cdots\right]$$

$$d^{(p)}_{\mathrm{Wass}} \propto 2\,\mathrm{GeV}\frac{0^{(\mathrm{p})}}{\mathrm{R}} + 1\,\mathrm{GeV}\frac{1^{(\mathrm{p})}}{\mathrm{R}}$$

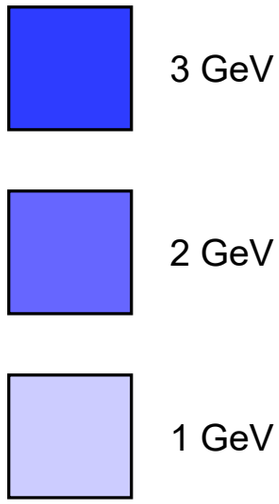


Choice of Metric



$$\mathrm{MSE} = \frac{1}{25}\left[(0-0)^2 + \cdots + (3-2)^2 + (0-0)^2 + \cdots + (0-1)^2\right]$$

$$d^{(p)}_{\mathrm{Wass}} \propto 2\,\mathrm{GeV}\frac{0^{(\mathrm{p})}}{\mathrm{R}} + 1\,\mathrm{GeV}\frac{(3\sqrt{2})^{(\mathrm{p})}}{\mathrm{R}}$$



Choice of Metric

Use same networks from previous training, but now use different anomaly score

Signal	Top jet			W jet		
Training Metric	Down Sampling	Anomaly Metric	AUC	Down Sampling	Anomaly Metric	AUC
MSE	1	Loss	0.823	**1**	**Loss**	**0.610**
	1	MSE	0.820	1	MSE	0.603
	1	MAE	0.793	1	MAE	0.480
	1	EMD(0.5)	0.819	1	EMD(0.5)	0.446
	1	**EMD(1)**	**0.828**	1	EMD(1)	0.411
	1	EMD(2)	0.807	1	EMD(2)	0.388
	2	**Loss**	**0.833**	**2**	**Loss**	**0.654**
	2	MSE	0.832	2	MSE	0.652
	2	MAE	0.802	2	MAE	0.529
	2	EMD(0.5)	0.816	2	EMD(0.5)	0.509
	2	EMD(1)	0.815	2	EMD(1)	0.506
	2	EMD(2)	0.809	2	EMD(2)	0.543
	3	**Loss**	**0.837**	**3**	**Loss**	**0.647**
	3	MSE	0.837	3	MSE	0.647
	3	MAE	0.808	3	MAE	0.533
	3	EMD(0.5)	0.828	3	EMD(0.5)	0.522
	3	EMD(1)	0.836	3	EMD(1)	0.518
	3	EMD(2)	0.817	3	EMD(2)	0.539

Using metric that network was trained on works best

Choice of Metric

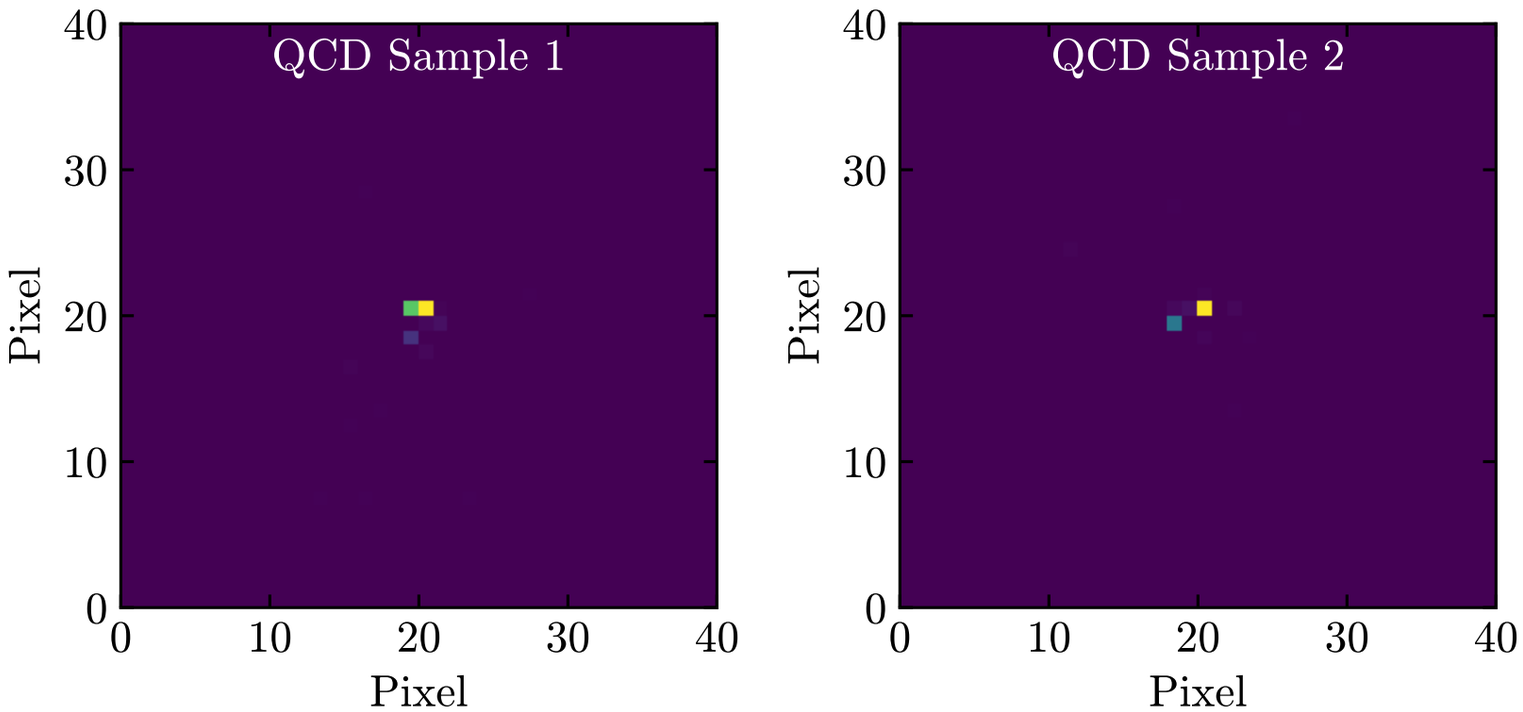
Signal	Top jet			W jet		
Training Metric	Down Sampling	Anomaly Metric	AUC	Down Sampling	Anomaly Metric	AUC
$d^{(1)}_{\text{Wass}}$	1	Loss	0.776	1	Loss	0.436
	1	MSE	0.711	**1**	**MSE**	**0.572**
	1	MAE	0.718	1	MAE	0.486
	1	EMD(0.5)	0.751	1	EMD(0.5)	0.469
	1	**EMD(1)**	**0.776**	1	EMD(1)	0.436
	1	EMD(2)	0.756	1	EMD(2)	0.389
	2	**Loss**	**0.790**	2	Loss	0.457
	2	MSE	0.757	**2**	**MSE**	**0.608**
	2	MAE	0.750	2	MAE	0.516
	2	EMD(0.5)	0.772	2	EMD(0.5)	0.494
	2	EMD(1)	0.790	2	EMD(1)	0.457
	2	EMD(2)	0.770	2	EMD(2)	0.401
	3	Loss	0.788	3	Loss	0.412
	3	**MSE**	**0.788**	**3**	**MSE**	**0.604**
	3	MAE	0.774	3	MAE	0.513
	3	EMD(0.5)	0.785	3	EMD(0.5)	0.473
	3	EMD(1)	0.788	3	EMD(1)	0.412
	3	EMD(2)	0.720	3	EMD(2)	0.358

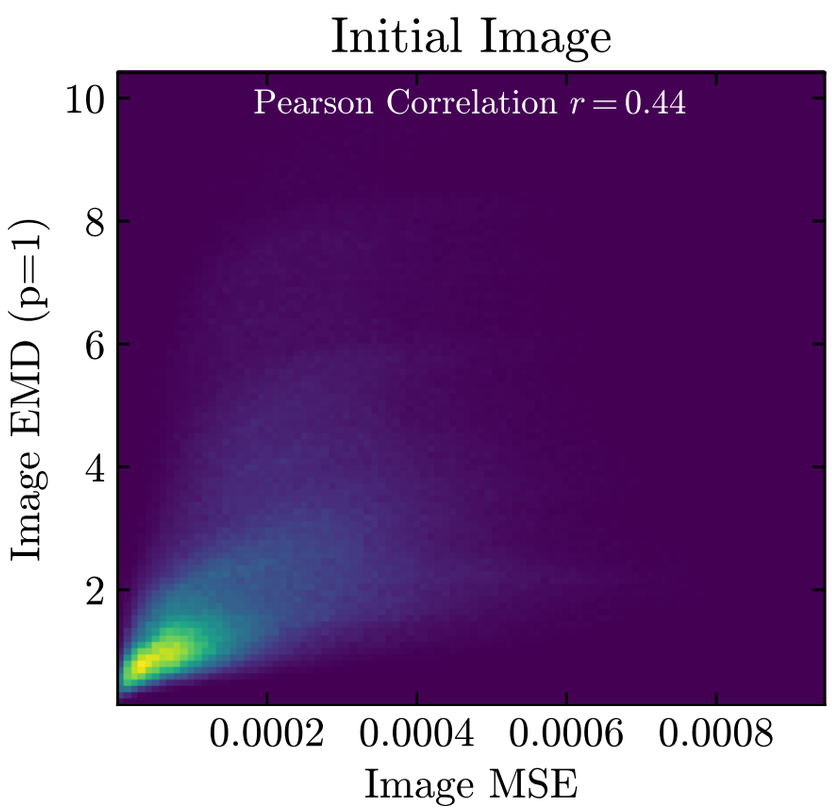
Train using EMD in loss function

Optimal transport not fast/easy for back propagation in training

When trained with EMD, MSE gives best anomaly detection

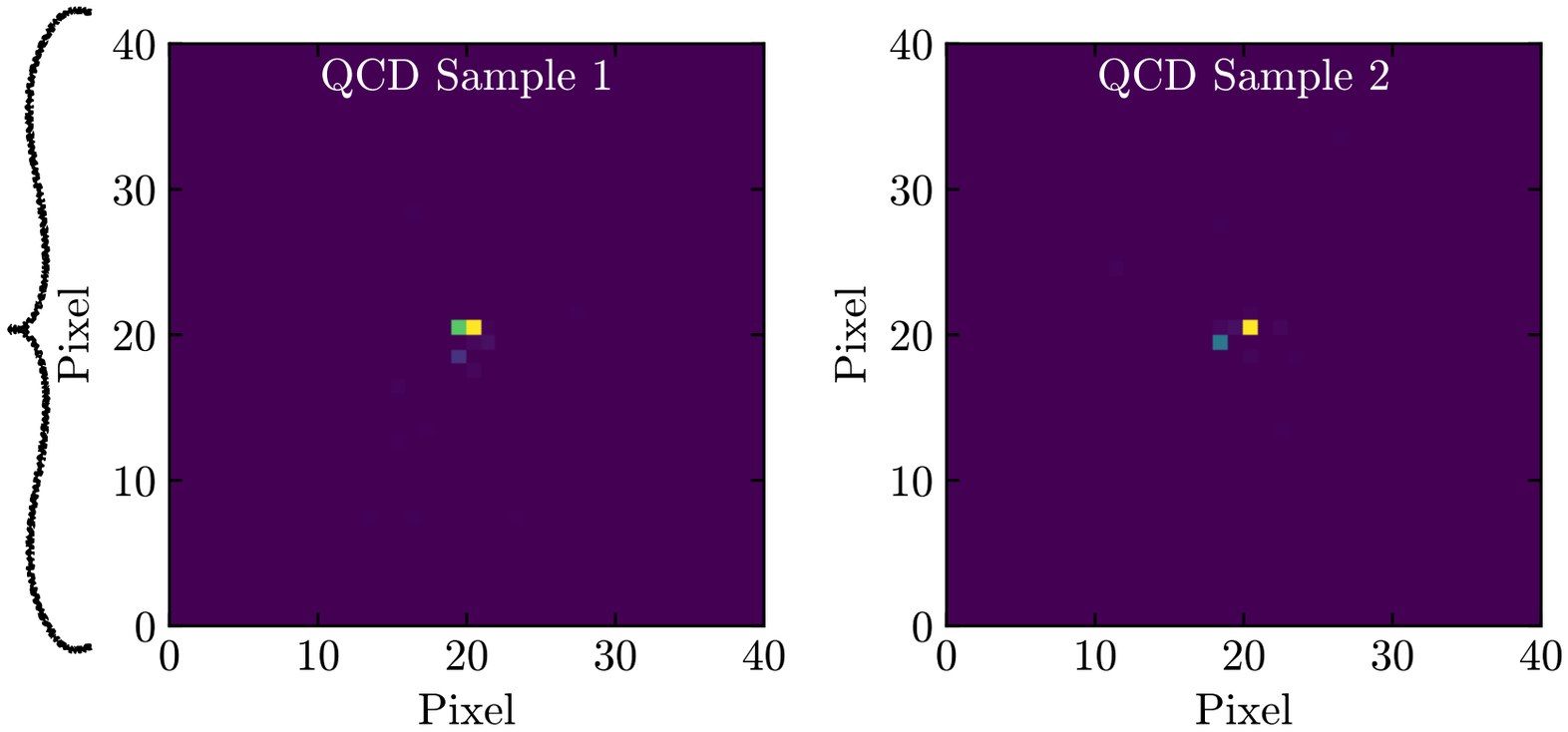
Pairwise Image Distances

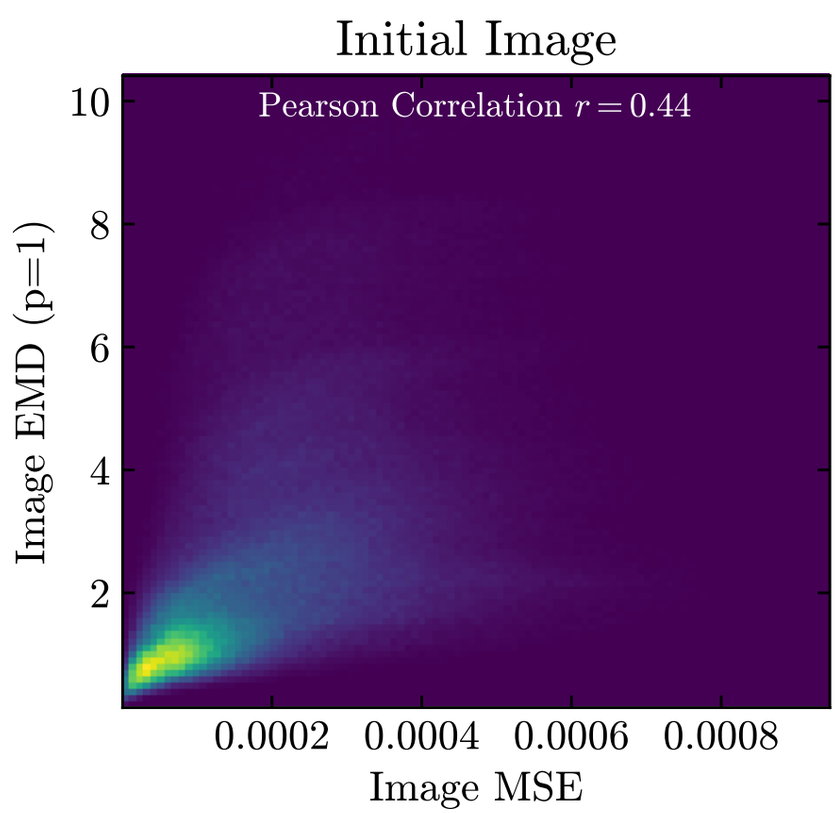




Pairwise Image Distances

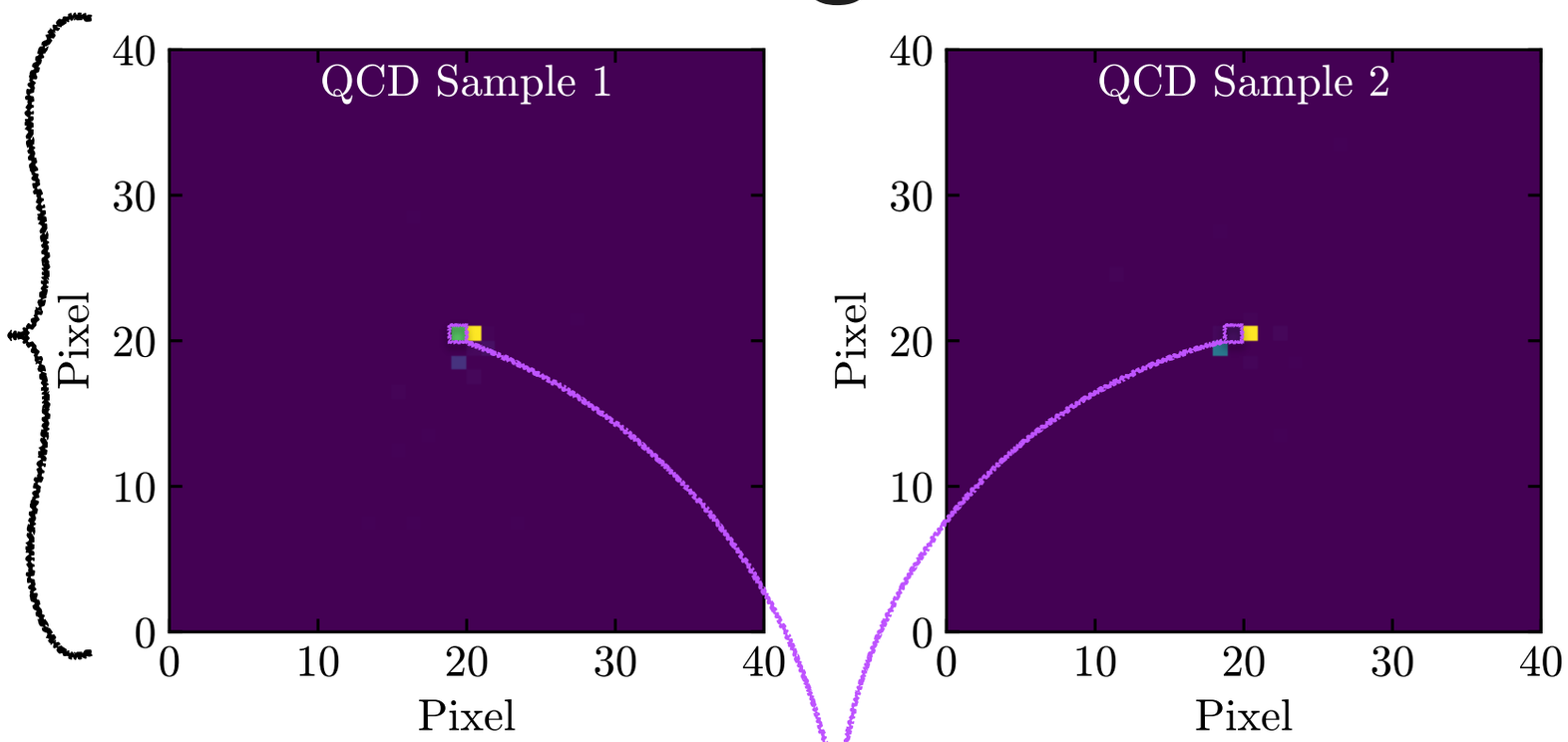
EMD distance between pairs of events



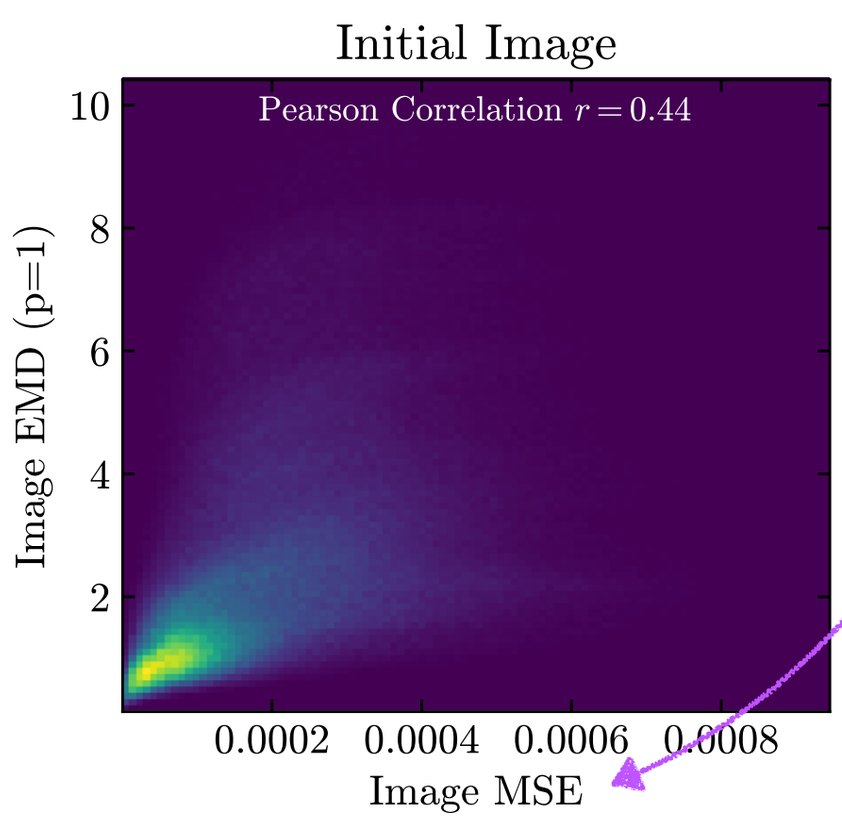


Pairwise Image Distances

EMD distance between pairs of events

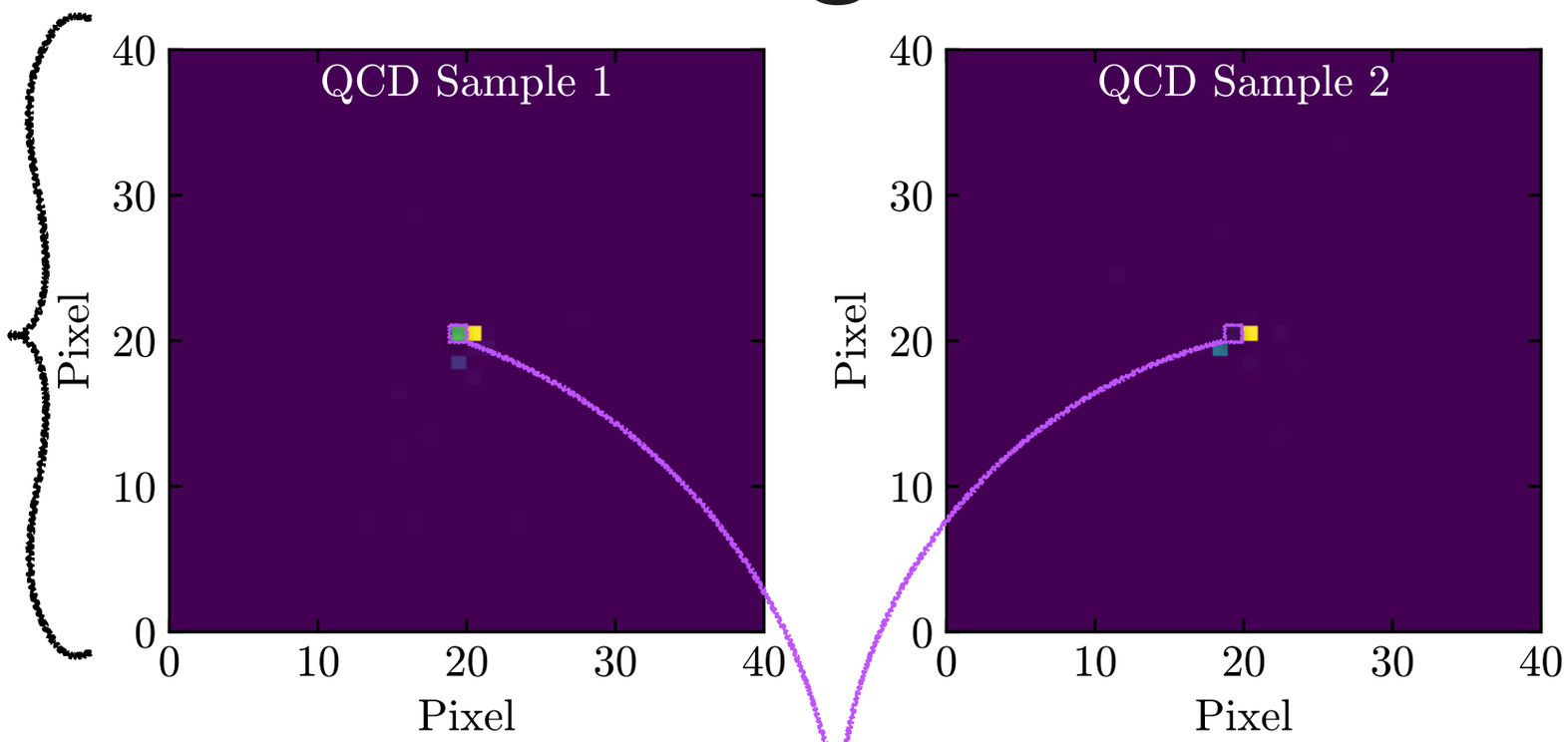


Pixel-by-pixel Mean Squared Error (pairwise between events)

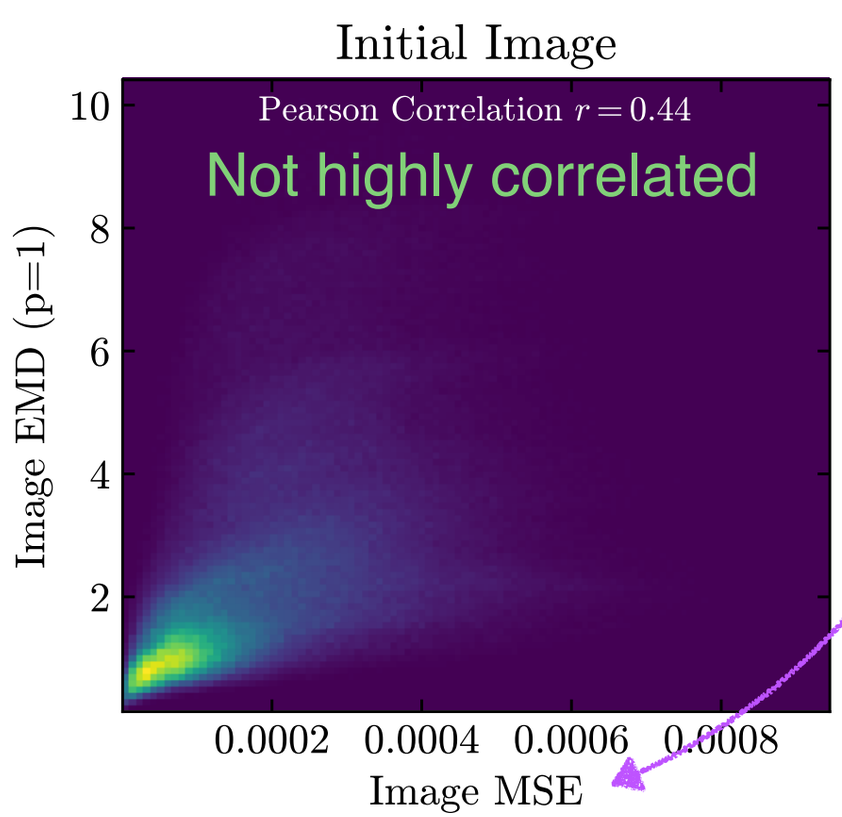


Pairwise Image Distances

EMD distance between pairs of events

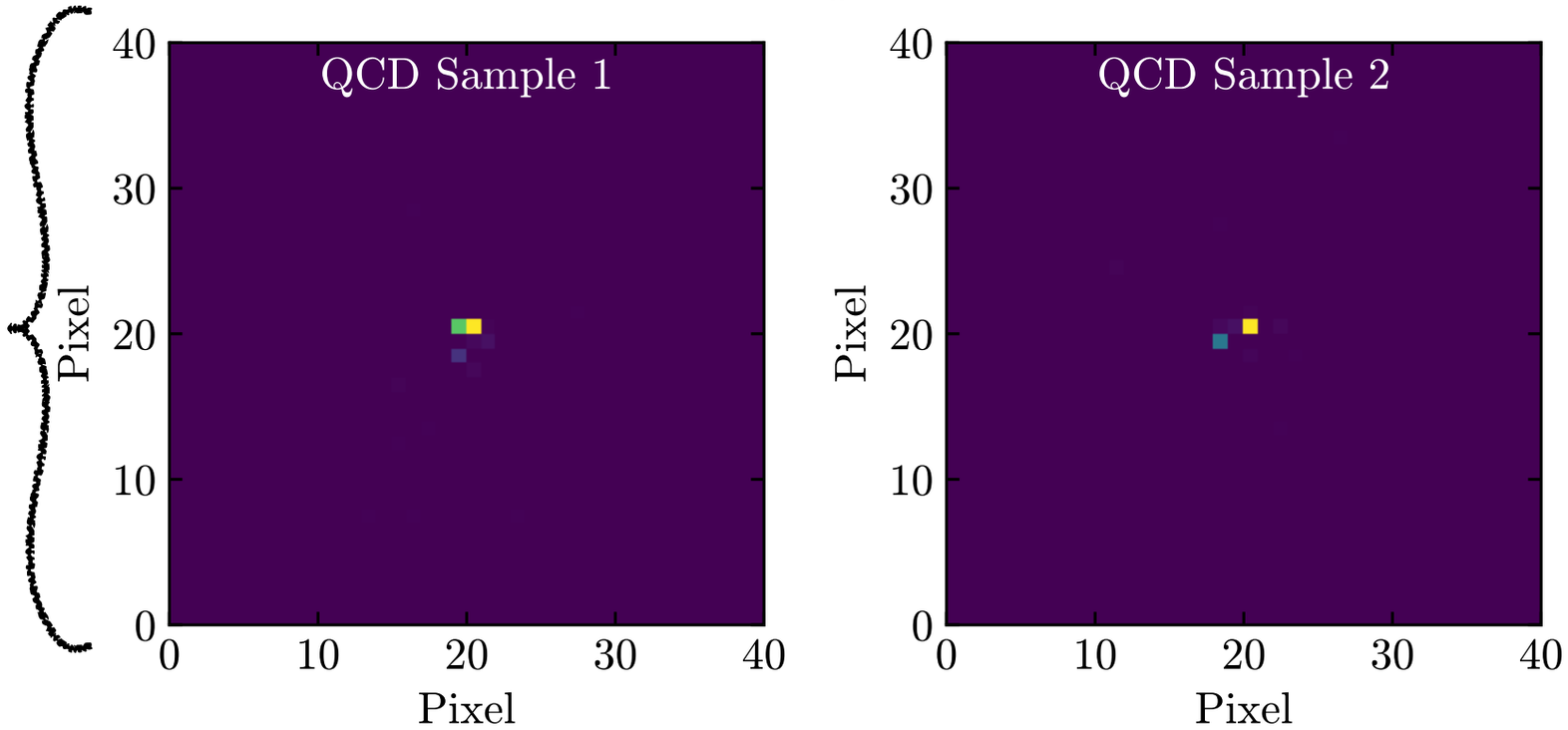


Pixel-by-pixel Mean Squared Error (pairwise between events)

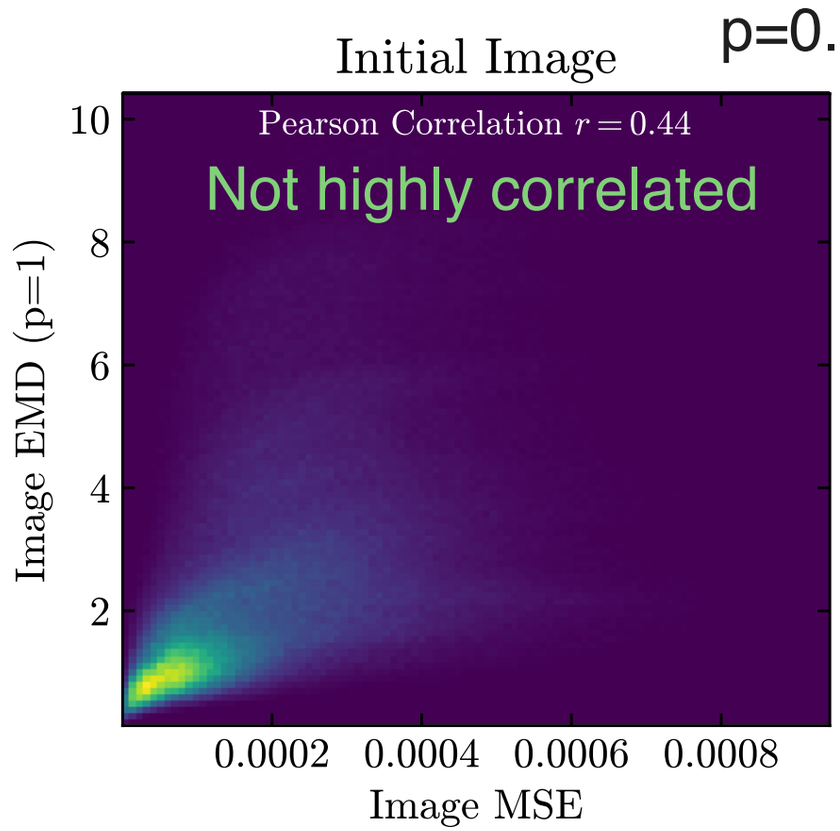


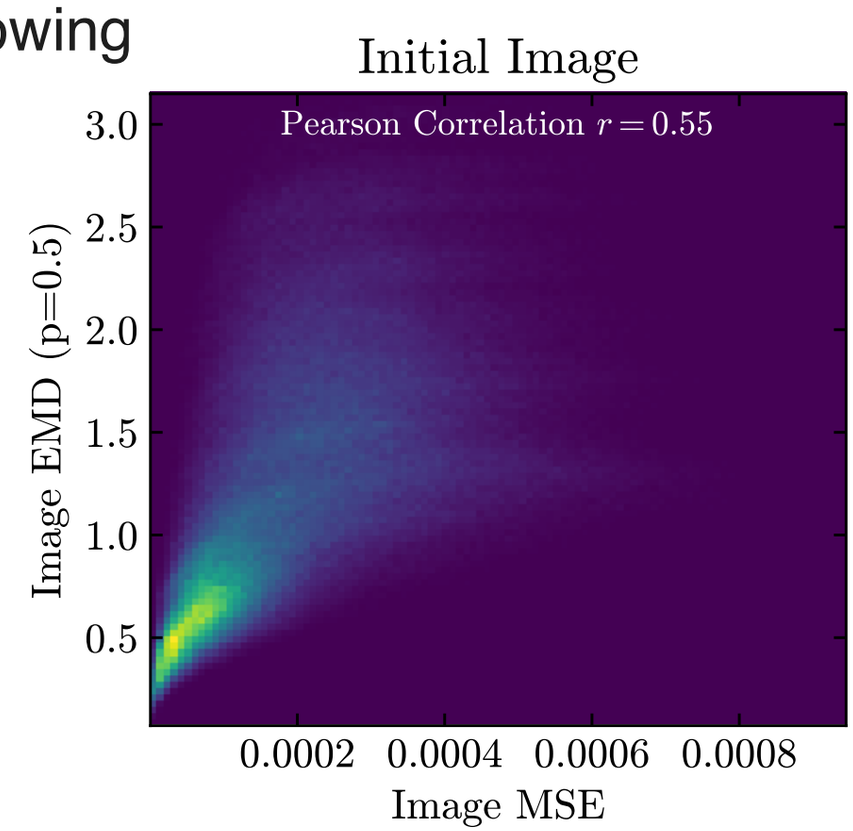
Pairwise Image Distances

EMD distance between pairs of events

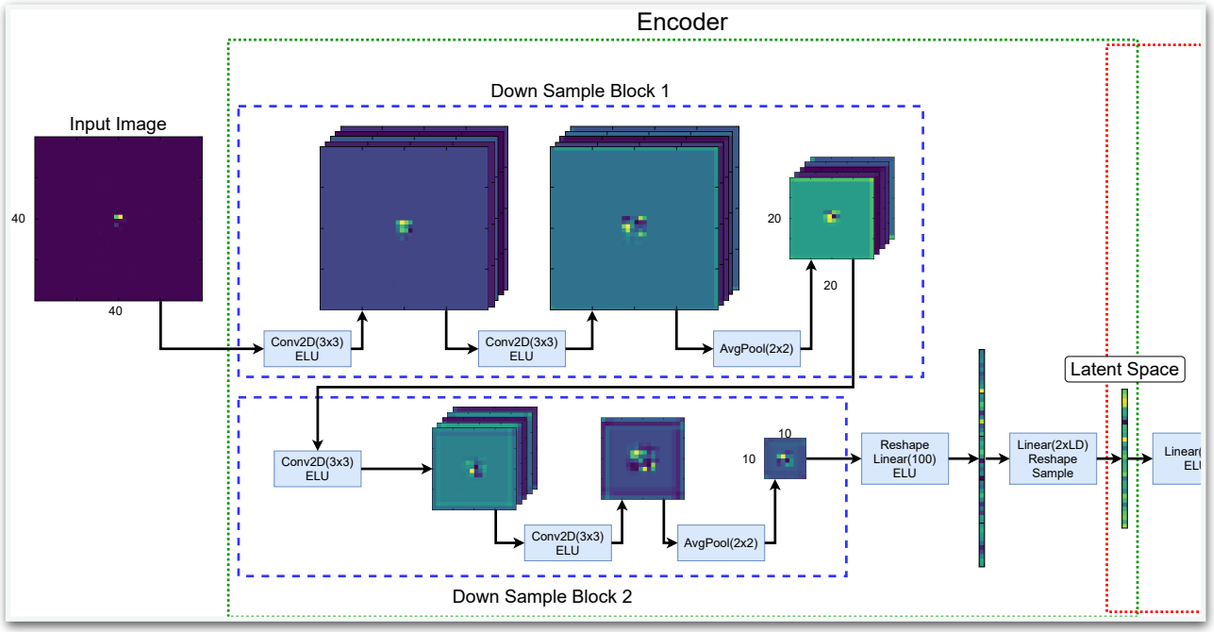


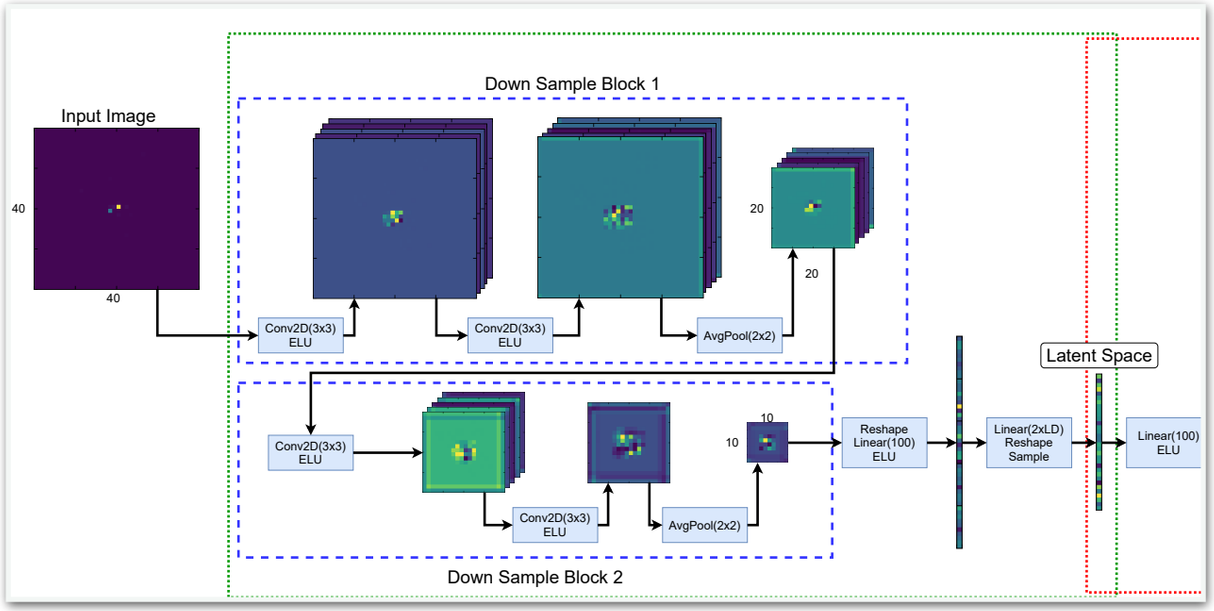
Correlation slightly larger for EMD with non-standard power, use p=0.5 for following

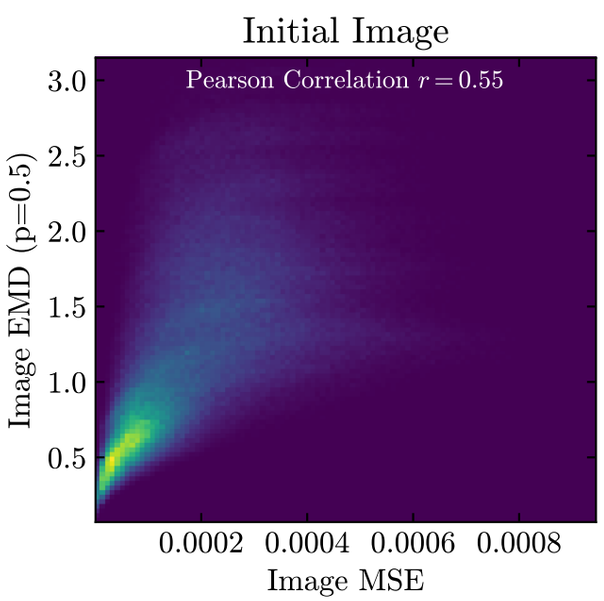




Pairwise Image Distances

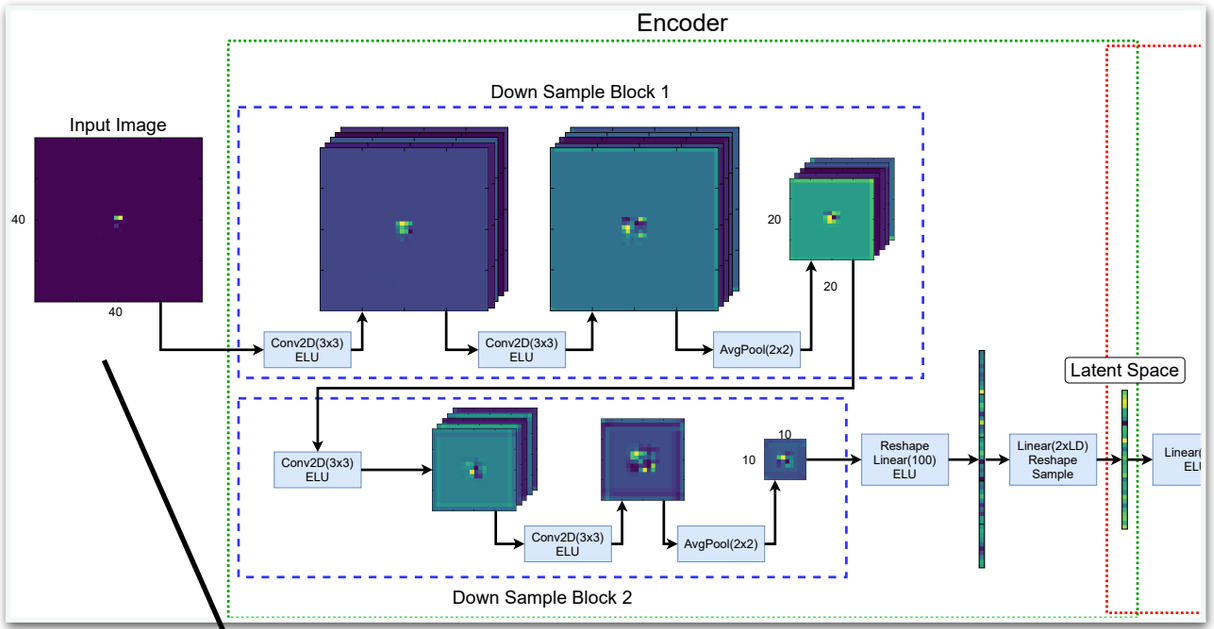


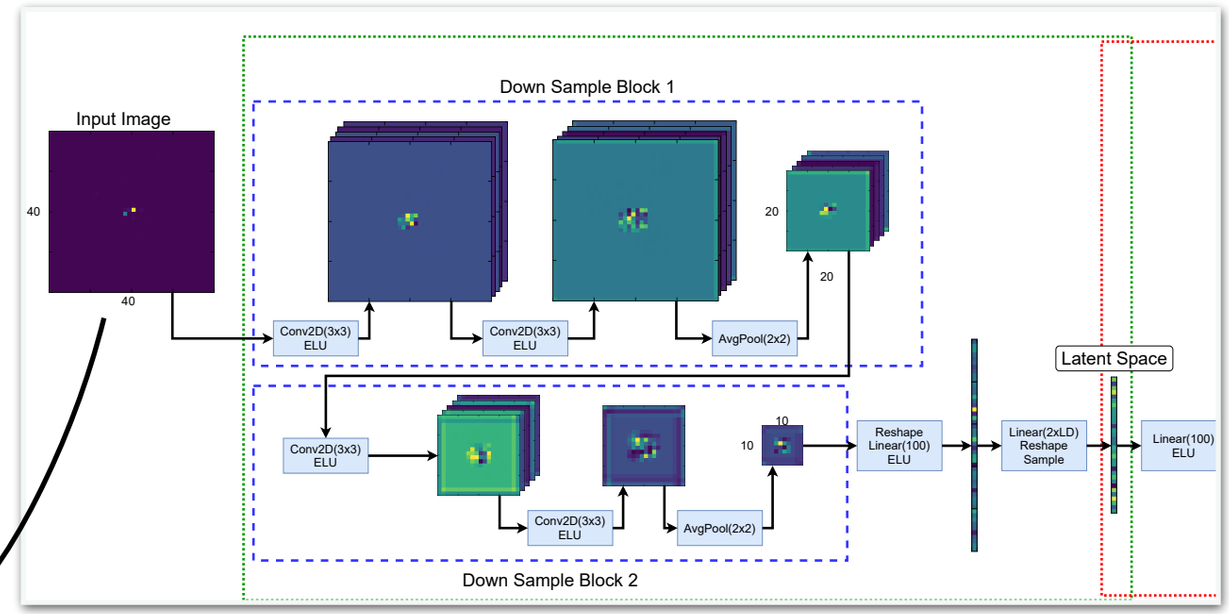


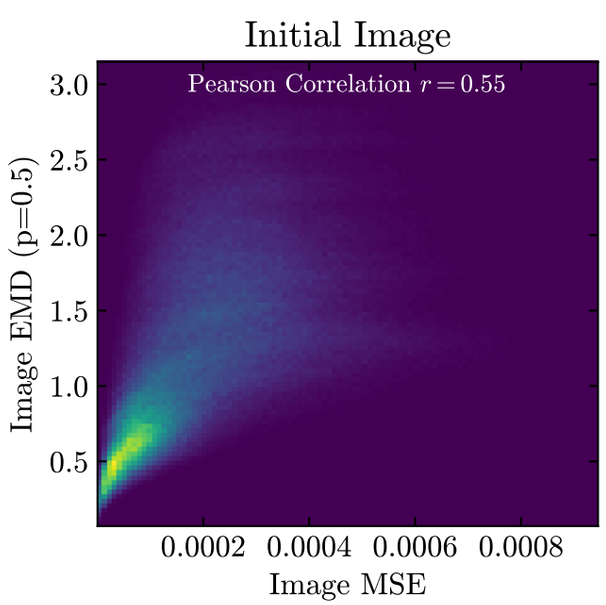


Layer activations become more correlated with EMD

Pairwise Image Distances

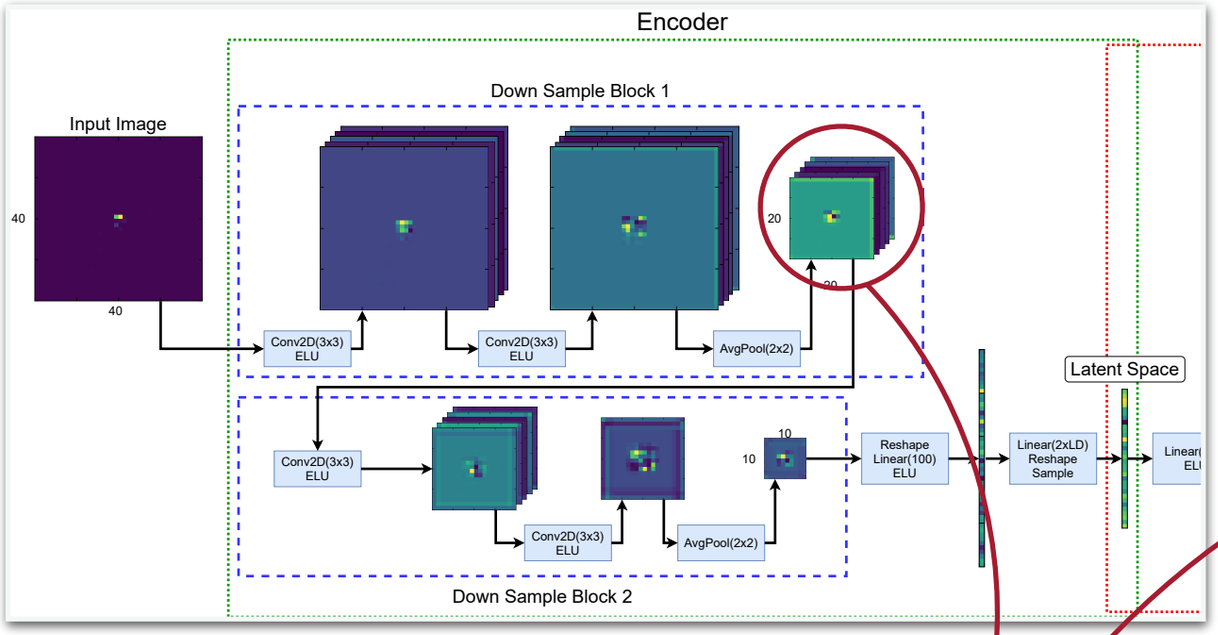


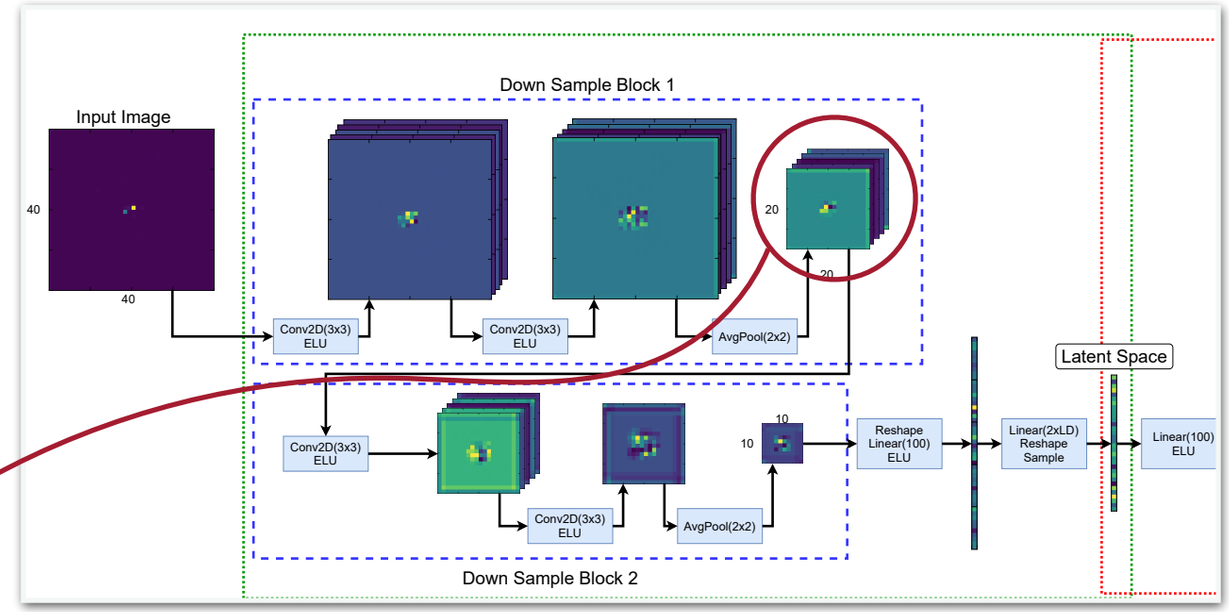


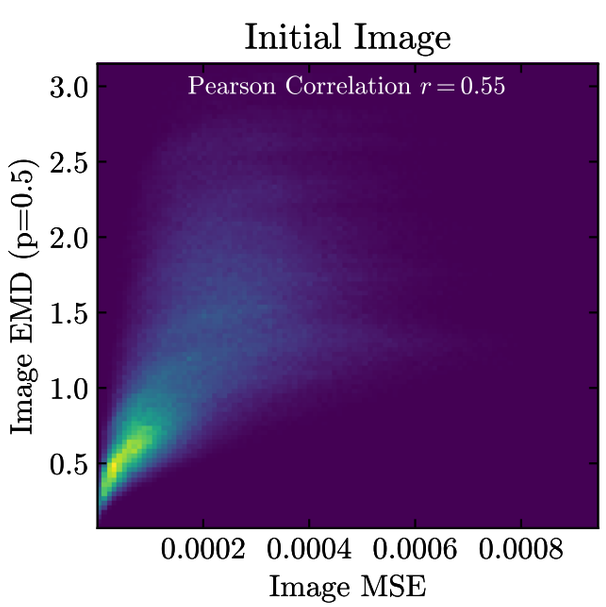


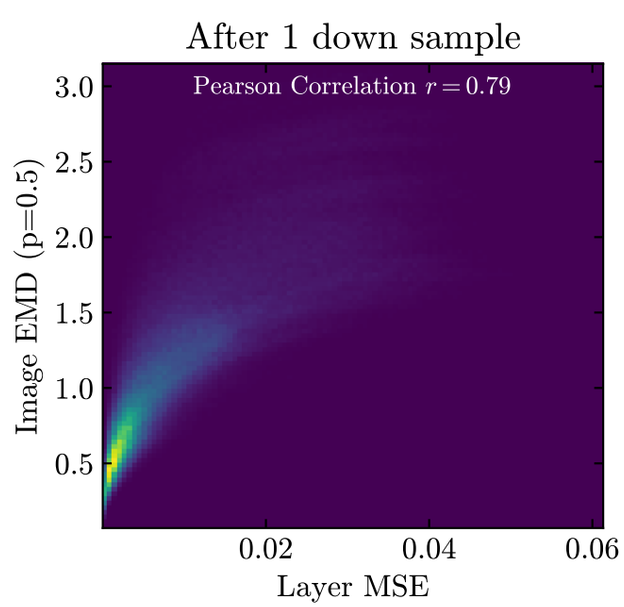
Layer activations become more correlated with EMD

Pairwise Image Distances



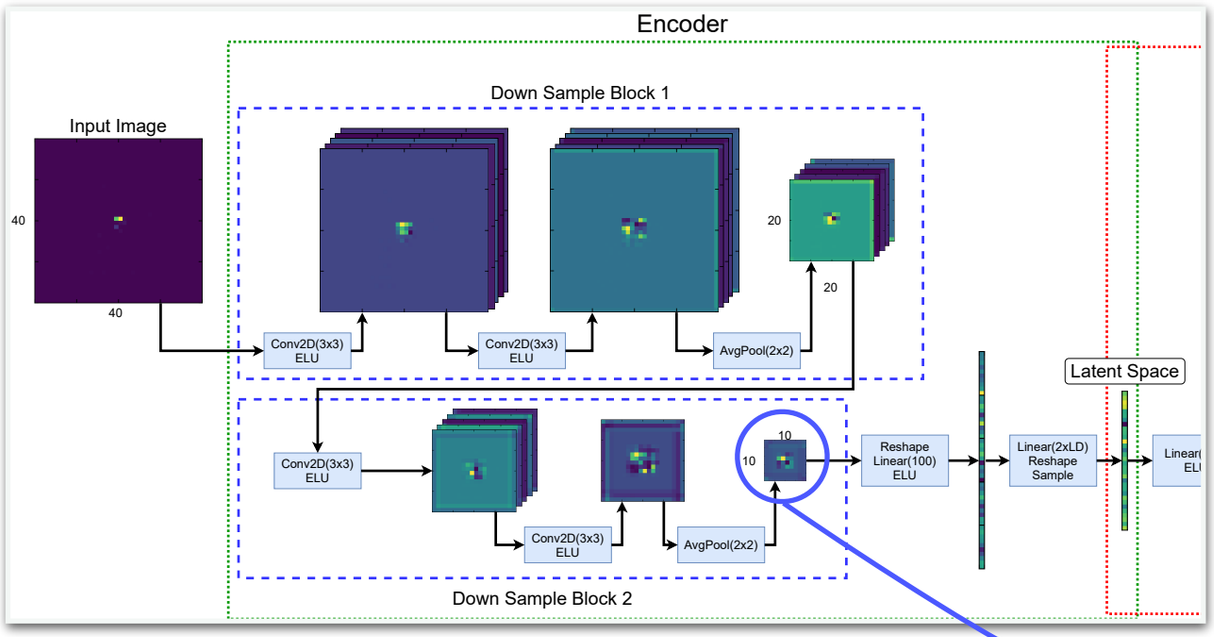


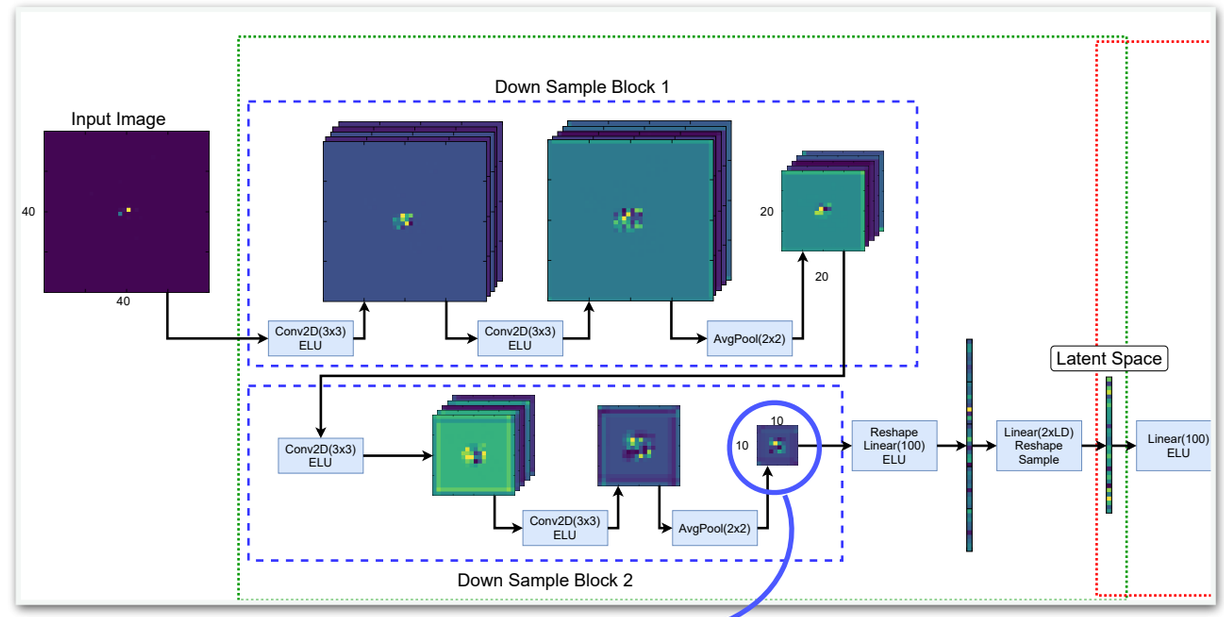


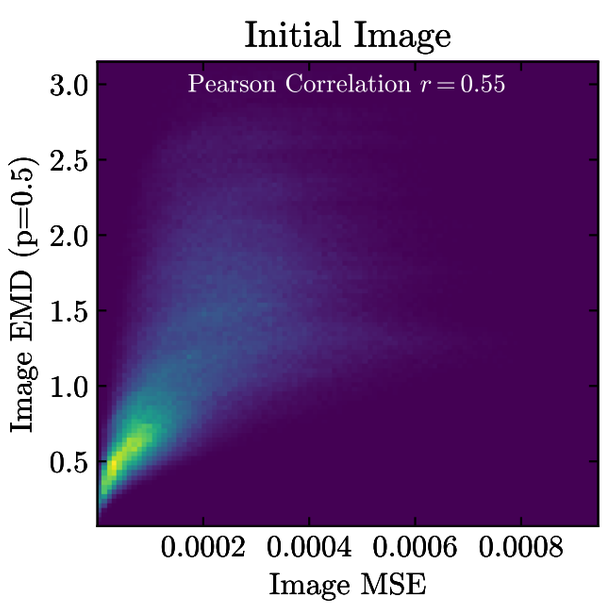


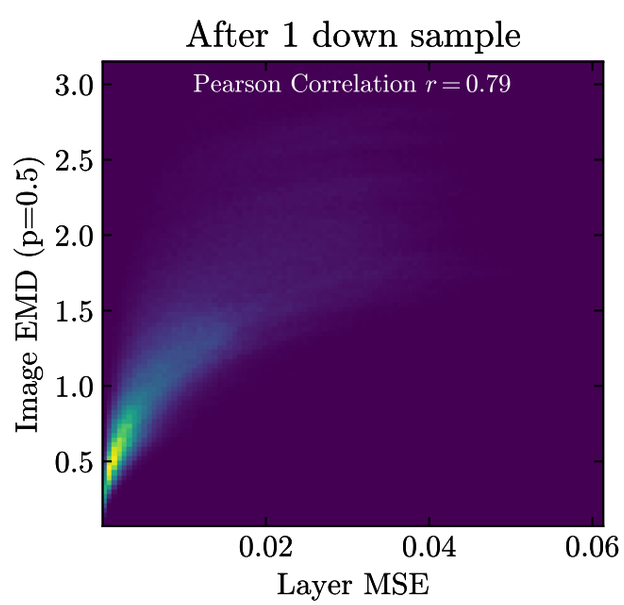
Layer activations become more correlated with EMD

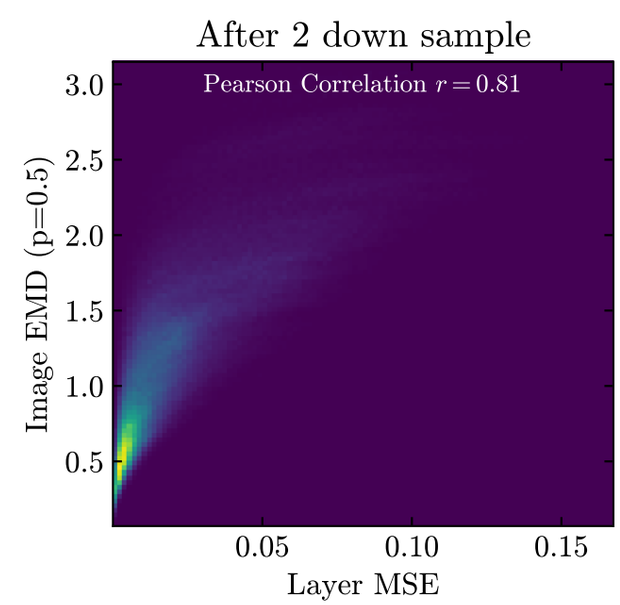
Pairwise Image Distances





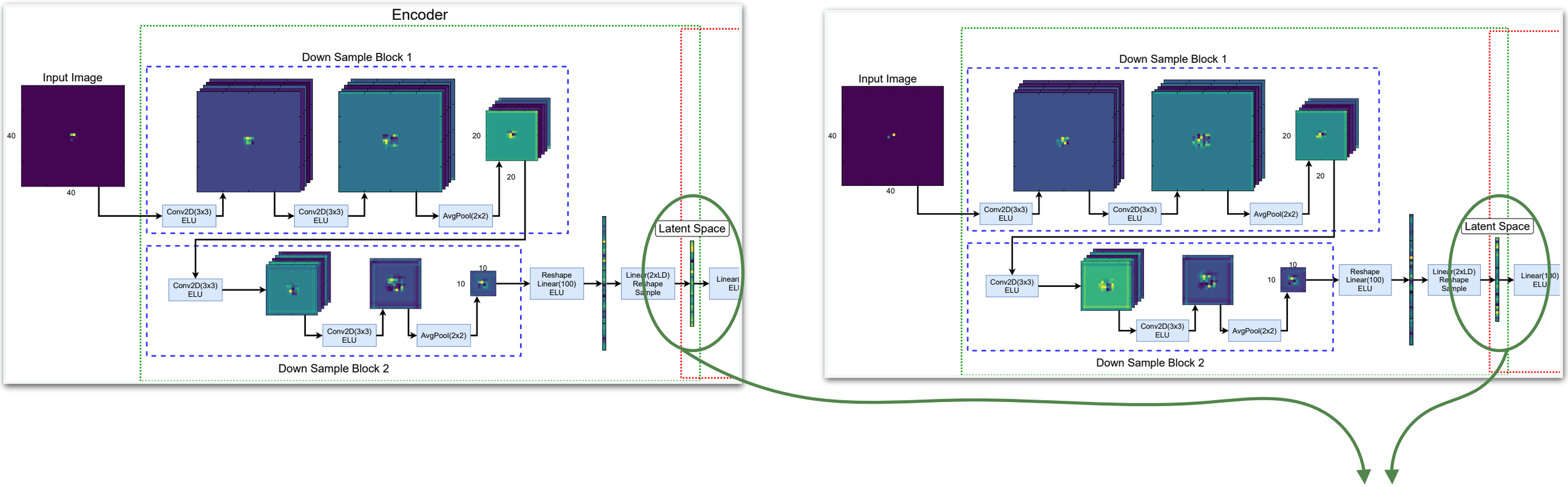


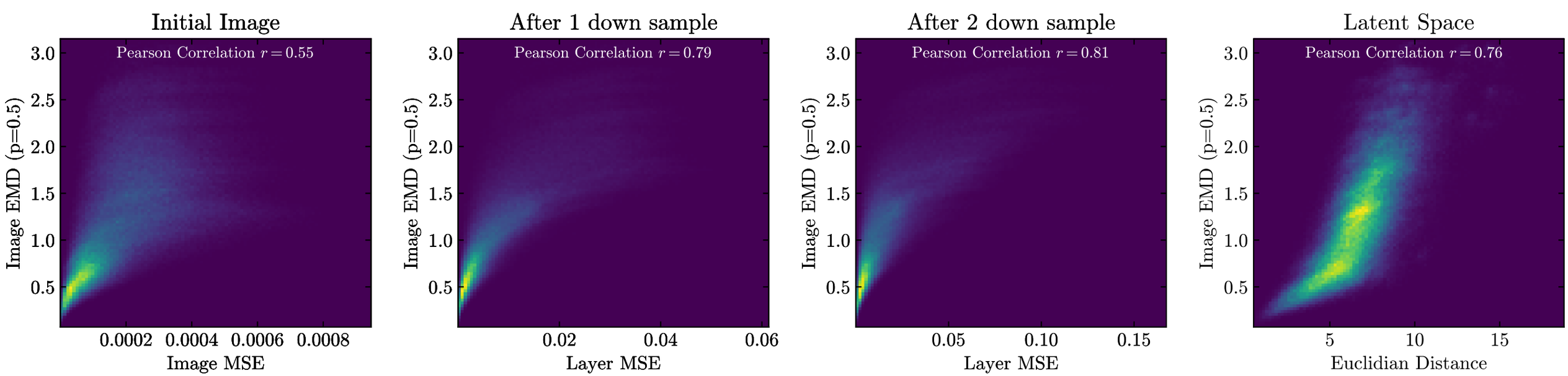




Layer activations become more correlated with EMD

Pairwise Image Distances



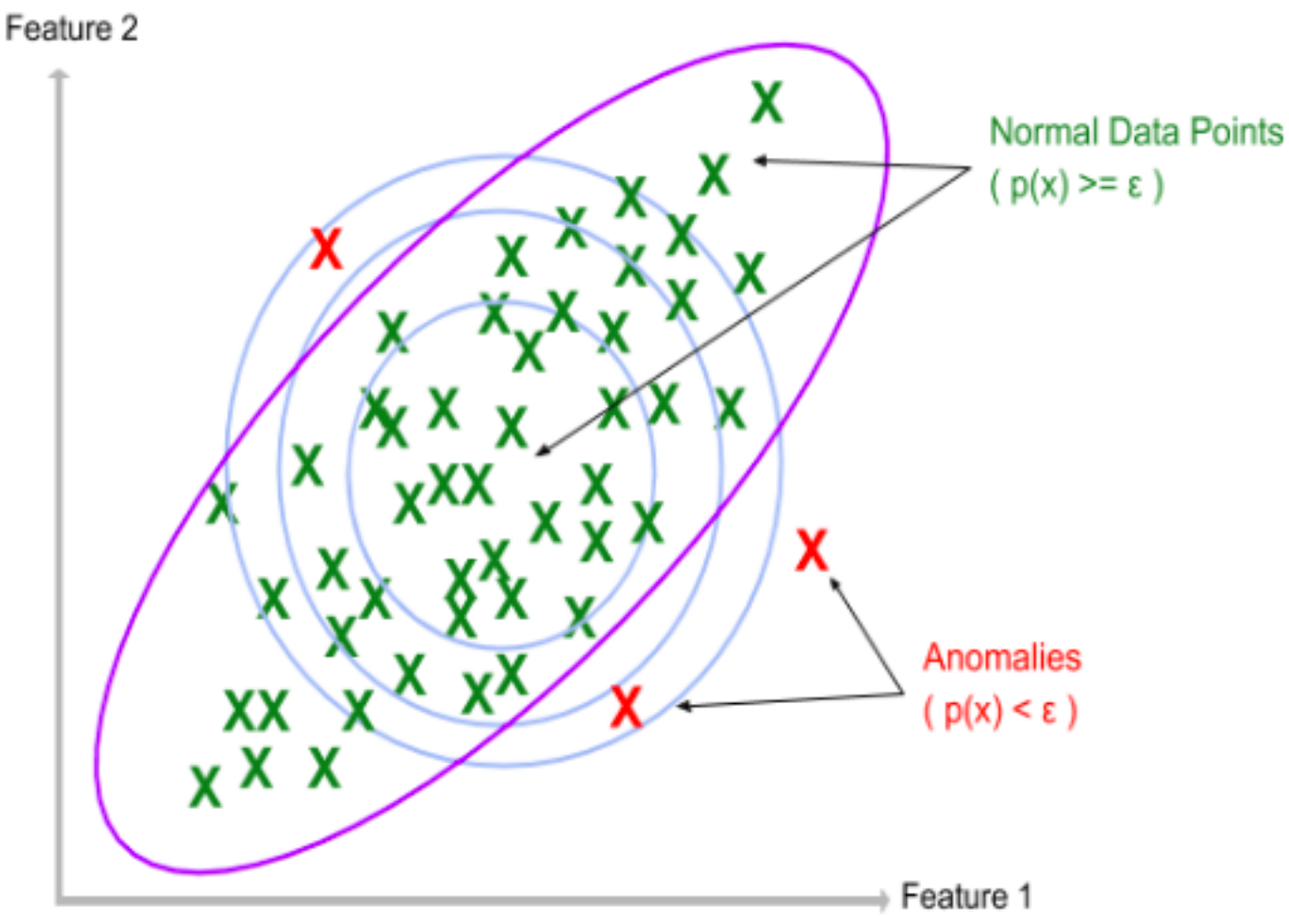


Layer activations become more correlated with EMD

Transition

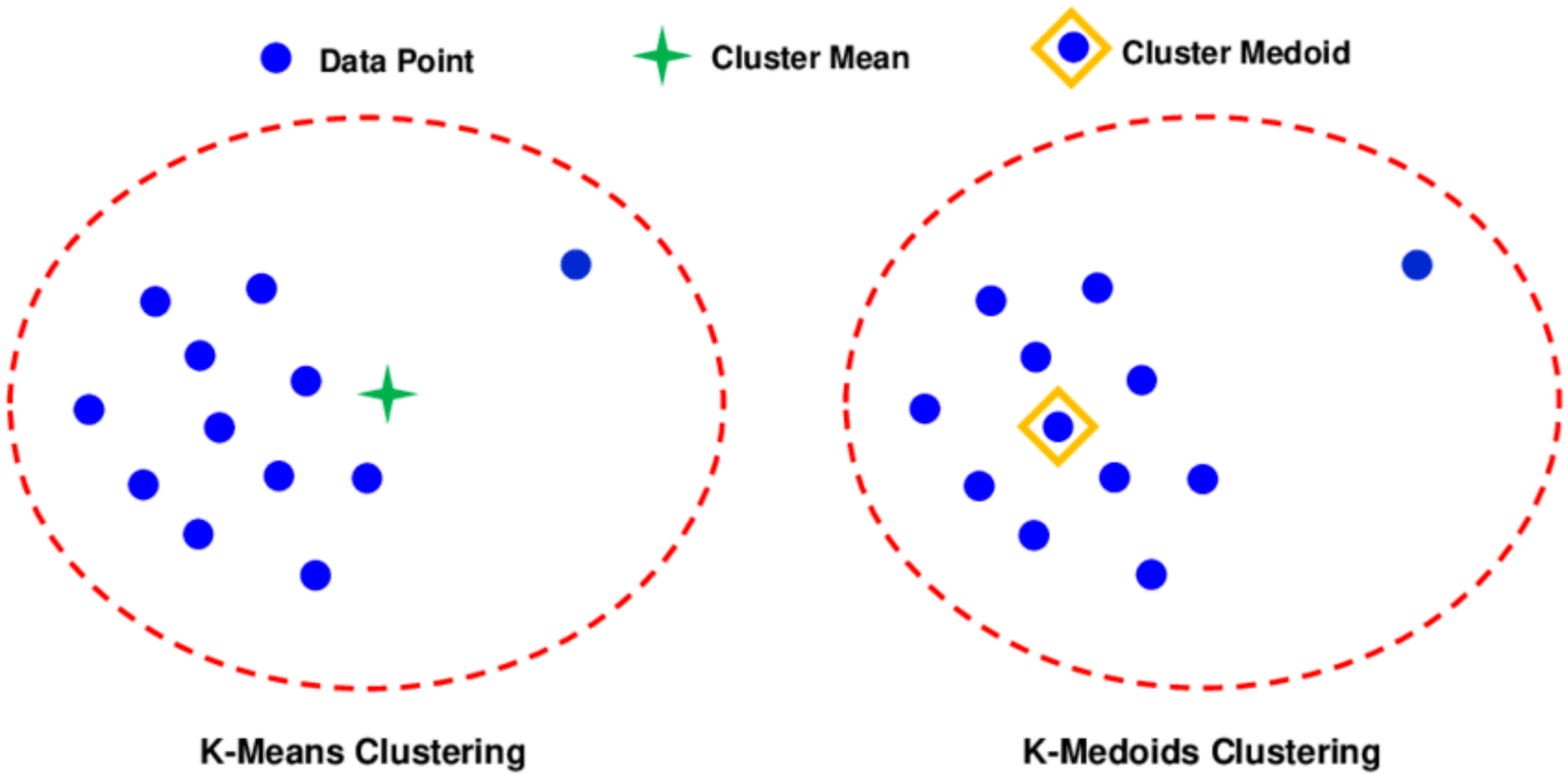
- Review Autoencoders and their pitfalls
- Update to Variational Autoencoders (VAE)
  - Regularizes and adds structure to latent space
  - Best method for Signal A is not the best for Signal B
  - Latent space distances are correlated with “physical distances” between events
- Take distances from quintessential events
  - Faster than training VAE
  - Still hard to remain model agnostic
  - Better at “inverse problem” than VAE

What is the goal?

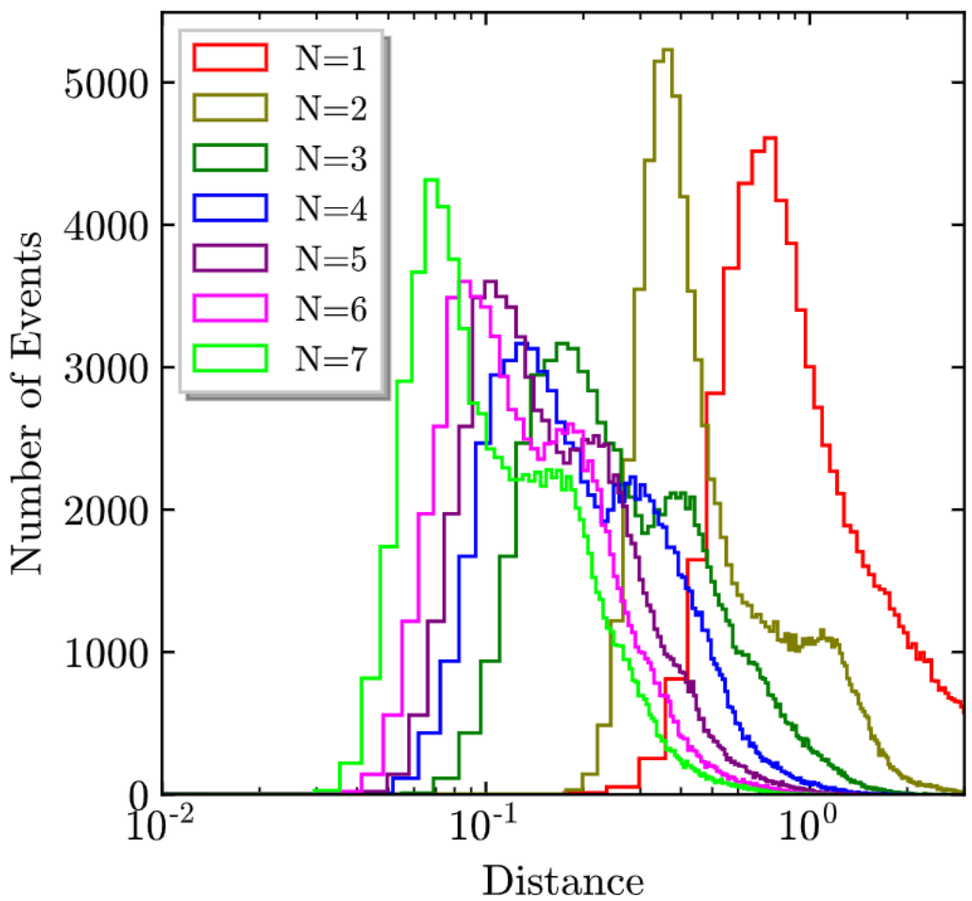


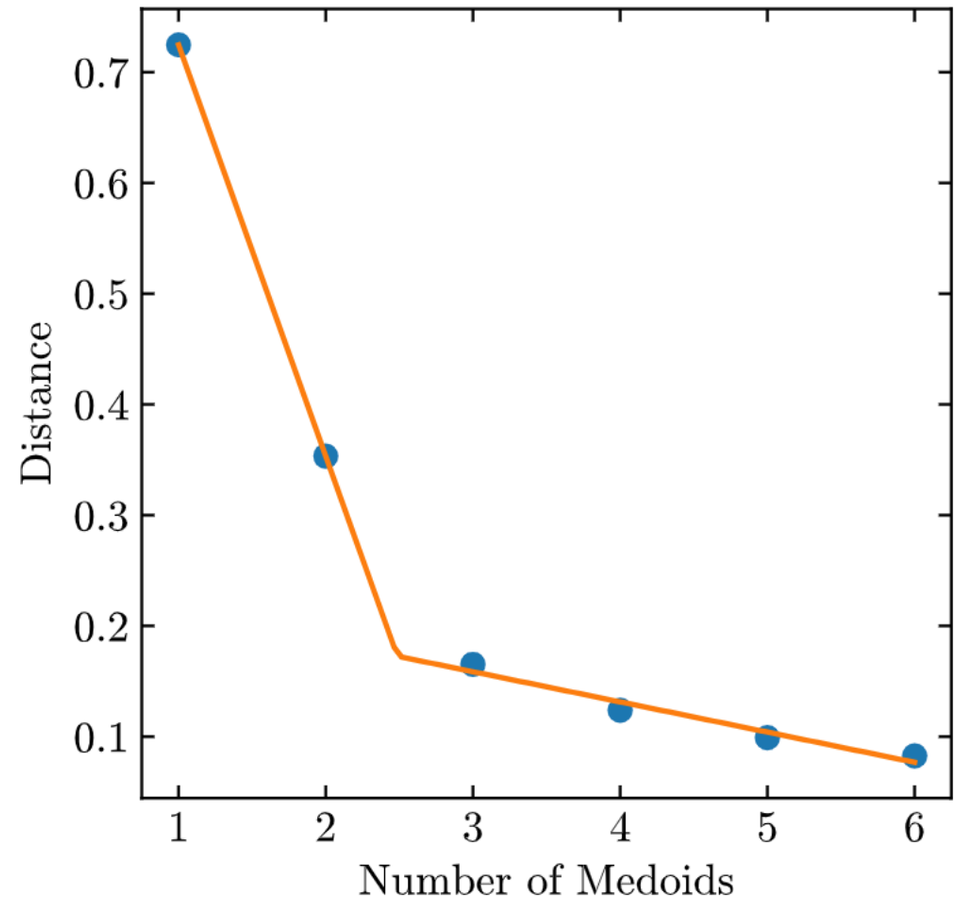
- Probabilities ~ distance from the distribution
- Describe the distribution by finding the quintessential events
- Distance from these events can be used as anomaly score

Describing the distribution



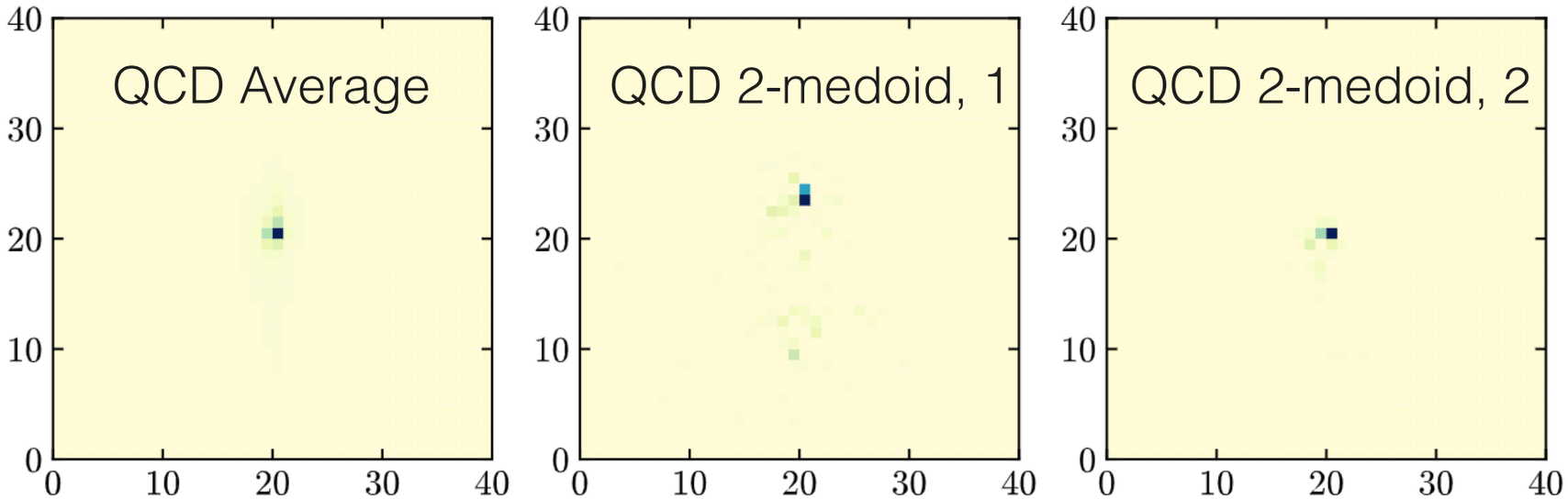
Describing the distribution





Using 1-Wasserstein balanced optimal transport, 2-3 medoids describes most of distribution

Event-to-Ensemble



				Top jet	W jet
Number of Images	Reference Sample	Combination Method	Metric	AUC	AUC
1	Avg		EMD p=1	0.806	0.618
1	Medoids		EMD p=1	0.831	0.656
3	Medoids	min	EMD p=1	0.853	0.675
5	Medoids	min	EMD p=1	0.866	0.596
7	Medoids	min	EMD p=1	0.865	0.612

VAE “best”: Top: 0.837, W: 0.654

Event-to-Ensemble

Number of Images	Reference Sample	Combination Method	Metric	Top jet AUC	W jet AUC
1	Avg		EMD p=1	0.806	0.618
1	Medoids		EMD p=1	0.831	0.656
3	Medoids	min	EMD p=1	0.853	0.675
5	Medoids	min	EMD p=1	0.866	0.596
7	Medoids	min	EMD p=1	0.865	0.612
1	Avg		EMD p=5	0.533	0.598
1	Medoids		EMD p=5	0.680	0.360
3	Medoids	min	EMD p=5	0.659	0.411
5	Medoids	min	EMD p=5	0.708	0.426
7	Medoids	min	EMD p=5	0.714	0.460
1	Avg		MSE p=1	0.832	0.712
1	Medoids		MSE p=1	0.820	0.711
3	Medoids	min	MSE p=1	0.819	0.613
5	Medoids	min	MSE p=1	0.829	0.668
7	Medoids	min	MSE p=1	0.832	0.650

$$d_{\text{Wass}}^{(p)} = \min_{f_{ij}>0} \sum_{ij} f_{ij} \frac{\theta_{ij}^{(p)}}{R}$$

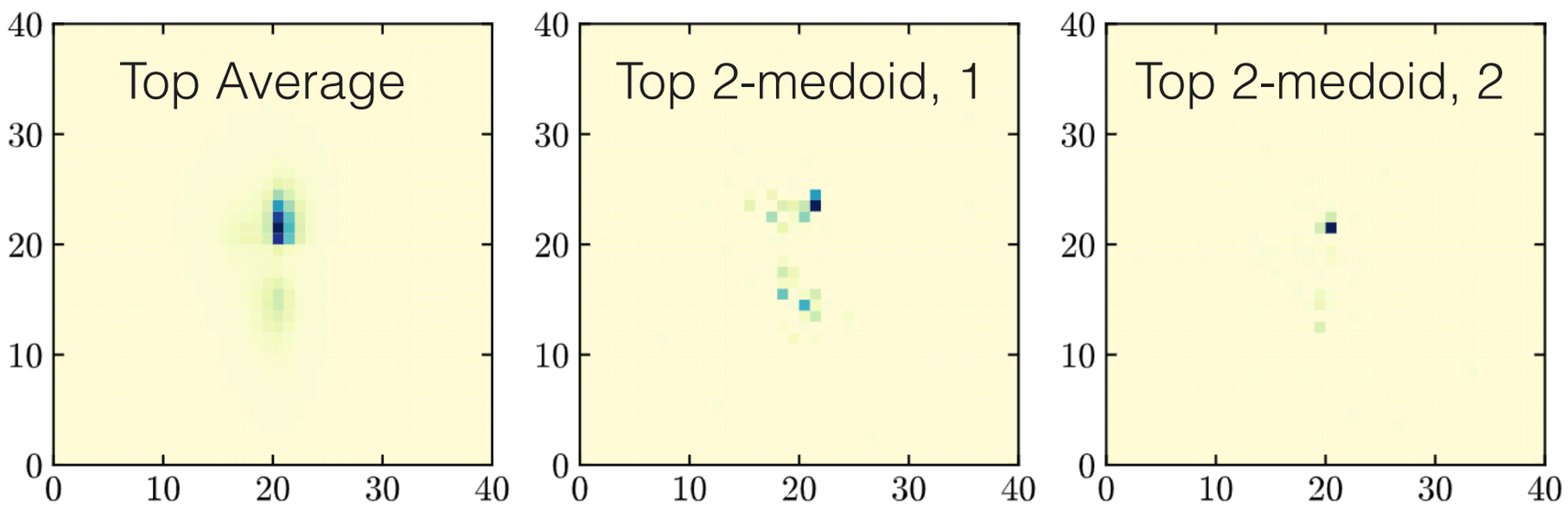
Using p=1 is fairly arbitrary, why not use any other metric

Optimizing for one signal is not optimal for the other

“Inverse Problem”

(V)AEs rely on not being able to reconstruct BSM events as well as the background events

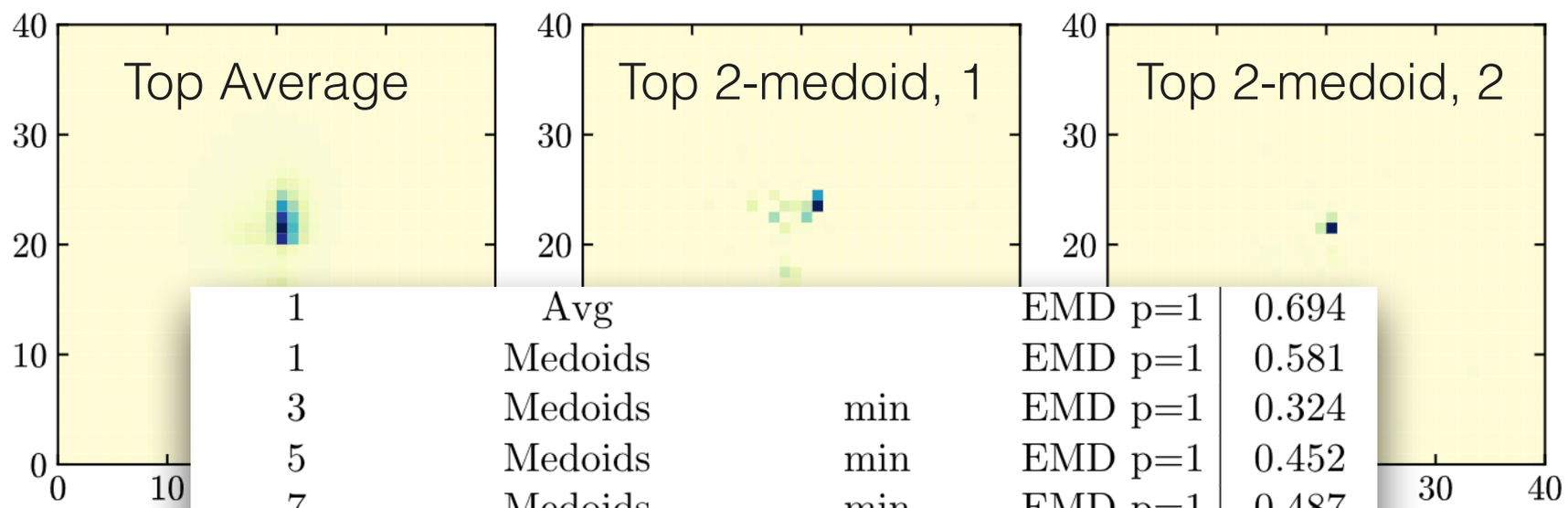
Treating tops jets as signal does not tag QCD as anomalous



“Inverse Problem”

(V)AEs rely on not being able to reconstruct BSM events as well as the background events

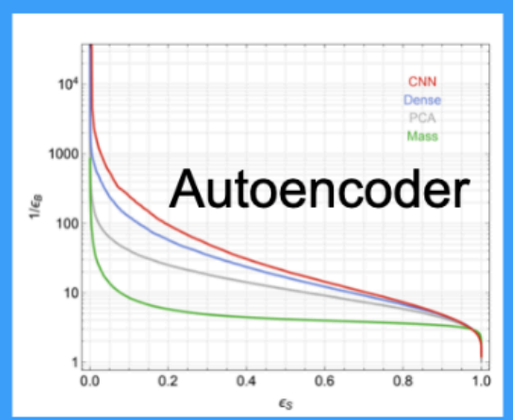
Treating tops jets as signal does not tag QCD as anomalous



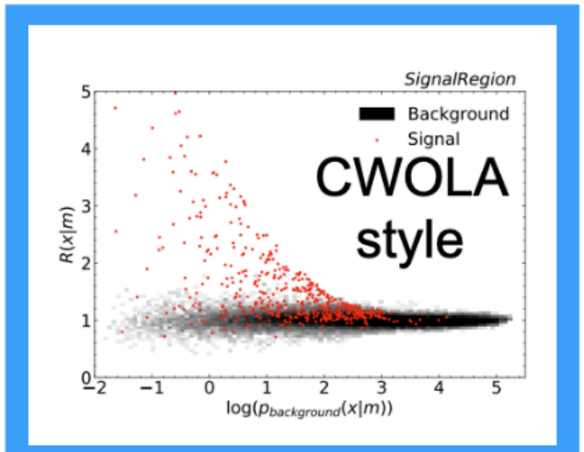
1	Avg		EMD p=1	0.694
1	Medoids		EMD p=1	0.581
3	Medoids	min	EMD p=1	0.324
5	Medoids	min	EMD p=1	0.452
7	Medoids	min	EMD p=1	0.487
1	Avg		EMD p=5	0.721
1	Medoids		EMD p=5	0.525
3	Medoids	min	EMD p=5	0.664
5	Medoids	min	EMD p=5	0.612
7	Medoids	min	EMD p=5	0.630
3	Medoids	sum	EMD p=5	0.659
5	Medoids	sum	EMD p=5	0.732
7	Medoids	sum	EMD p=5	0.747

Similar Signals?

Prior Free → Fully Supervised



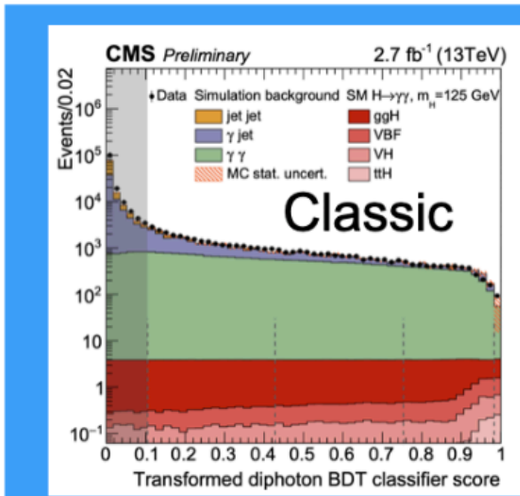
Knowledge of Background



Signal in (Mass) Window

Semi-Supervised Approaches

If it quaks like a duck?



Fully Supervised

"Quasi Anomalous Knowledge: Searching for new physics with embedded knowledge," Park, Rankin, Udrescu, Yunus, and Harris [2011.03550]

Similar Signals?

				Top jet	W jet
Reference Sample	Number of Images	Reference Sample	Metric	AUC	AUC
Top Reference	1	Avg	EMD p=1	0.69	0.69
	1	Medoid	EMD p=1	0.58	0.79
	3	min Medoids	EMD p=1	0.32	0.79
	5	min Medoids	EMD p=1	0.45	0.84
	7	min Medoids	EMD p=1	0.49	0.83
	1	Avg	EMD p=5	0.72	0.40
	1	Medoid	EMD p=5	0.53	0.52
	3	min Medoids	EMD p=5	0.66	0.61
	5	min Medoids	EMD p=5	0.61	0.54
	7	min Medoids	EMD p=5	0.63	0.53
	3	sum Medoids	EMD p=5	0.66	0.66
	5	sum Medoids	EMD p=5	0.73	0.58
	7	sum Medoids	EMD p=5	0.75	0.60
	1	Avg	MSE p=1	0.48	0.57
	1	Medoids	MSE p=1	0.29	0.64
	3	min Medoids	MSE p=1	0.25	0.36
	5	min Medoids	MSE p=1	0.32	0.58
	7	min Medoids	MSE p=1	0.33	0.59

W jets seem to be closer to tops than QCD

Somehow this doesn't work for tops?

Conclusion

Anomaly detection aims to:

- Be model agnostic
- Train directly on LHC data

Conclusion

Anomaly detection aims to:

- Be model agnostic
- Train directly on LHC data

How do you optimize network hyper parameters without looking at a signal?

Conclusion

Anomaly detection aims to:

- Be model agnostic
- Train directly on LHC data

How do you optimize network hyper parameters without looking at a signal?

Want a way to describe the distribution and how “far” an event is from the background

Conclusion

Anomaly detection aims to:

- Be model agnostic
- Train directly on LHC data

How do you optimize network hyper parameters without looking at a signal?

Want a way to describe the distribution and how “far” an event is from the background

Use k-medoids of optimal transport metric:

- No training
- Similar (better) performance to VAE

Conclusion

Anomaly detection aims to:

- Be model agnostic
- Train directly on LHC data

How do you optimize network hyper parameters without looking at a signal?

Want a way to describe the distribution and how “far” an event is from the background

Use k-medoids of optimal transport metric:

- No training
- Similar (better) performance to VAE
- Still has arbitrary choices to make without looking at signal?

Backup

Supervised Classification