Generative Modeling for LArTPC Images

Zev Imani

NPML 2024

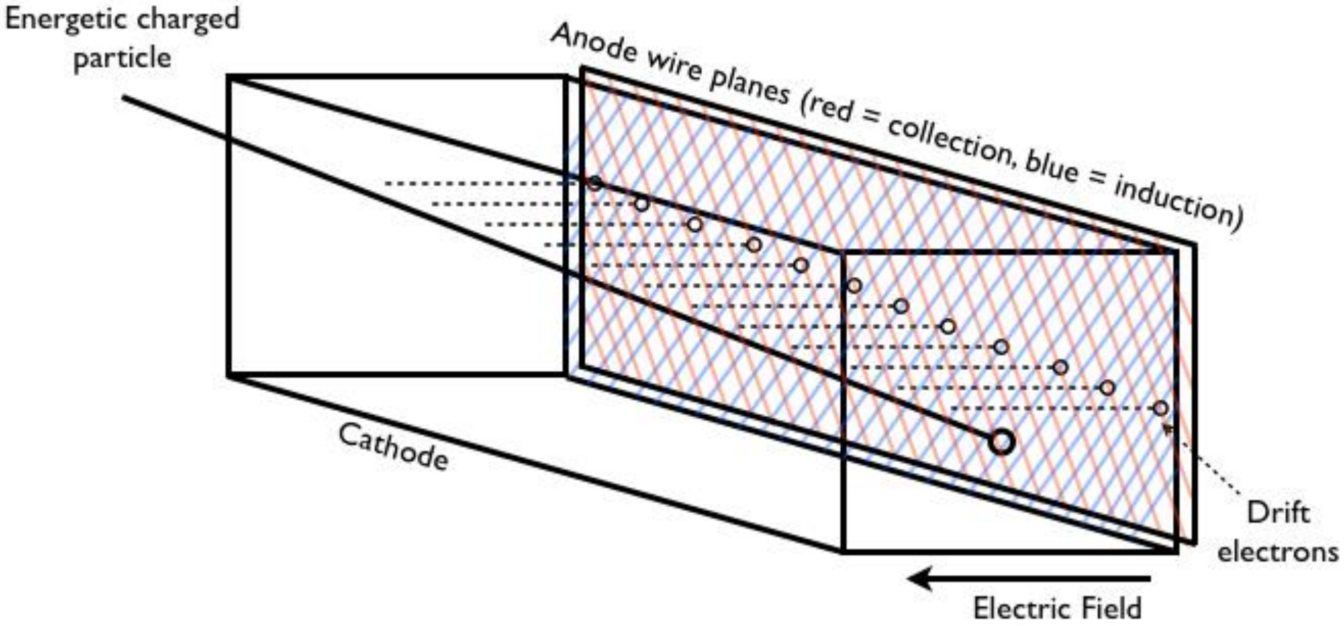
The NSF Institute for
Artificial Intelligence and
Fundamental Interactions

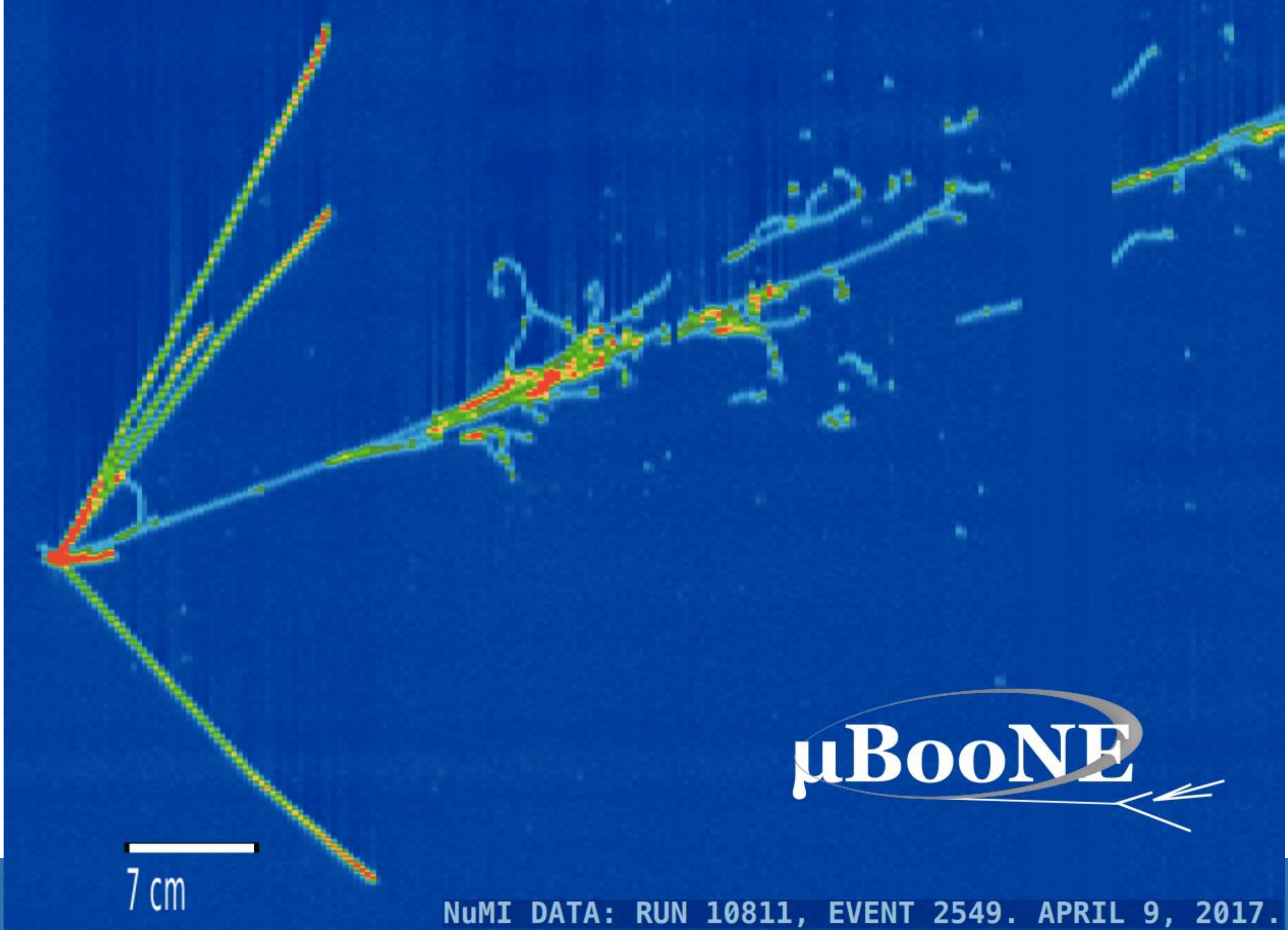
Outline

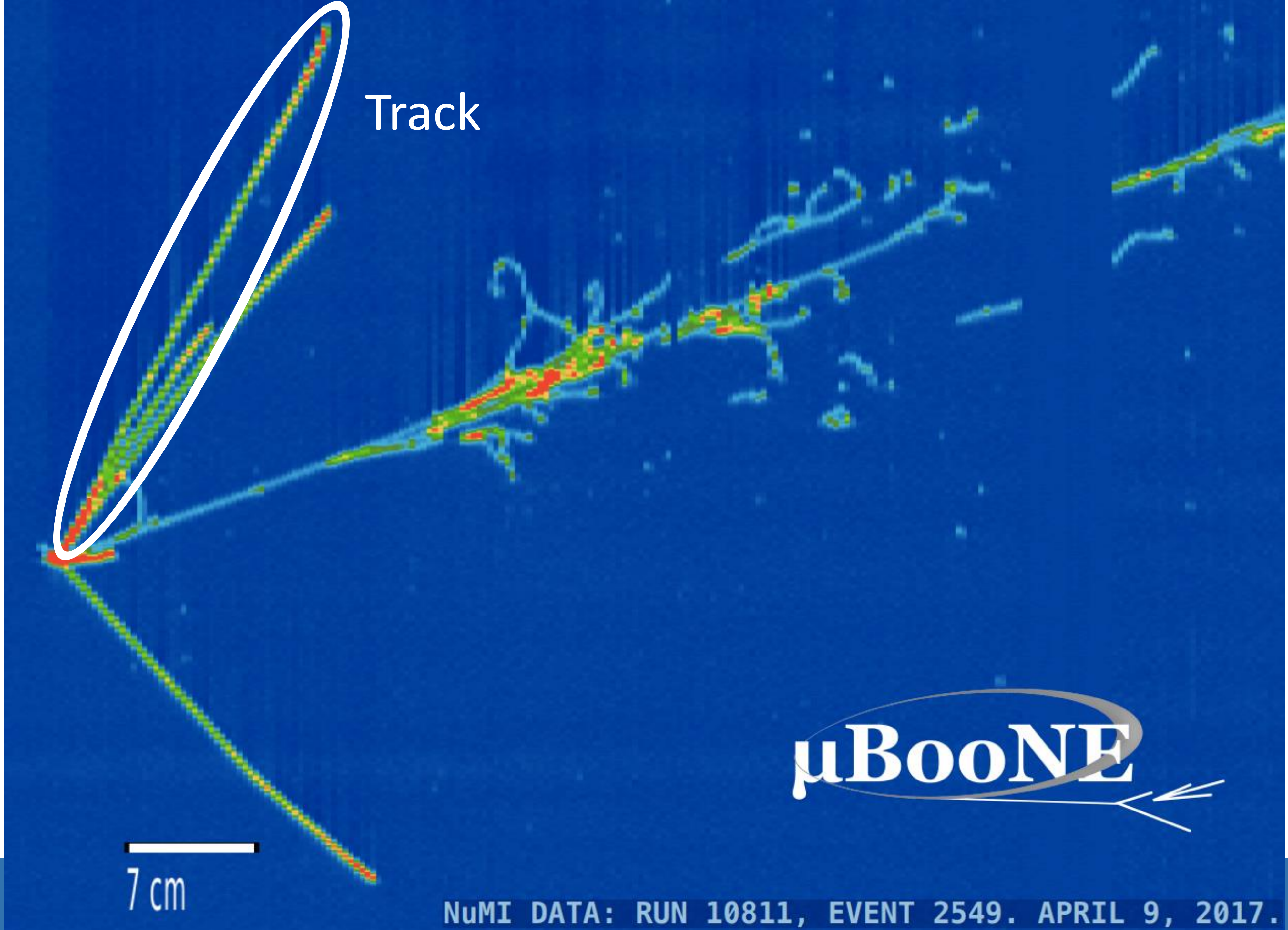
1. Data Motivation
2. LArTPC Image Generation Attempts
3. ~~Diffusion Methodology~~
4. Quality Tests (Abridged)
5. Distance Metrics
6. Takeaways

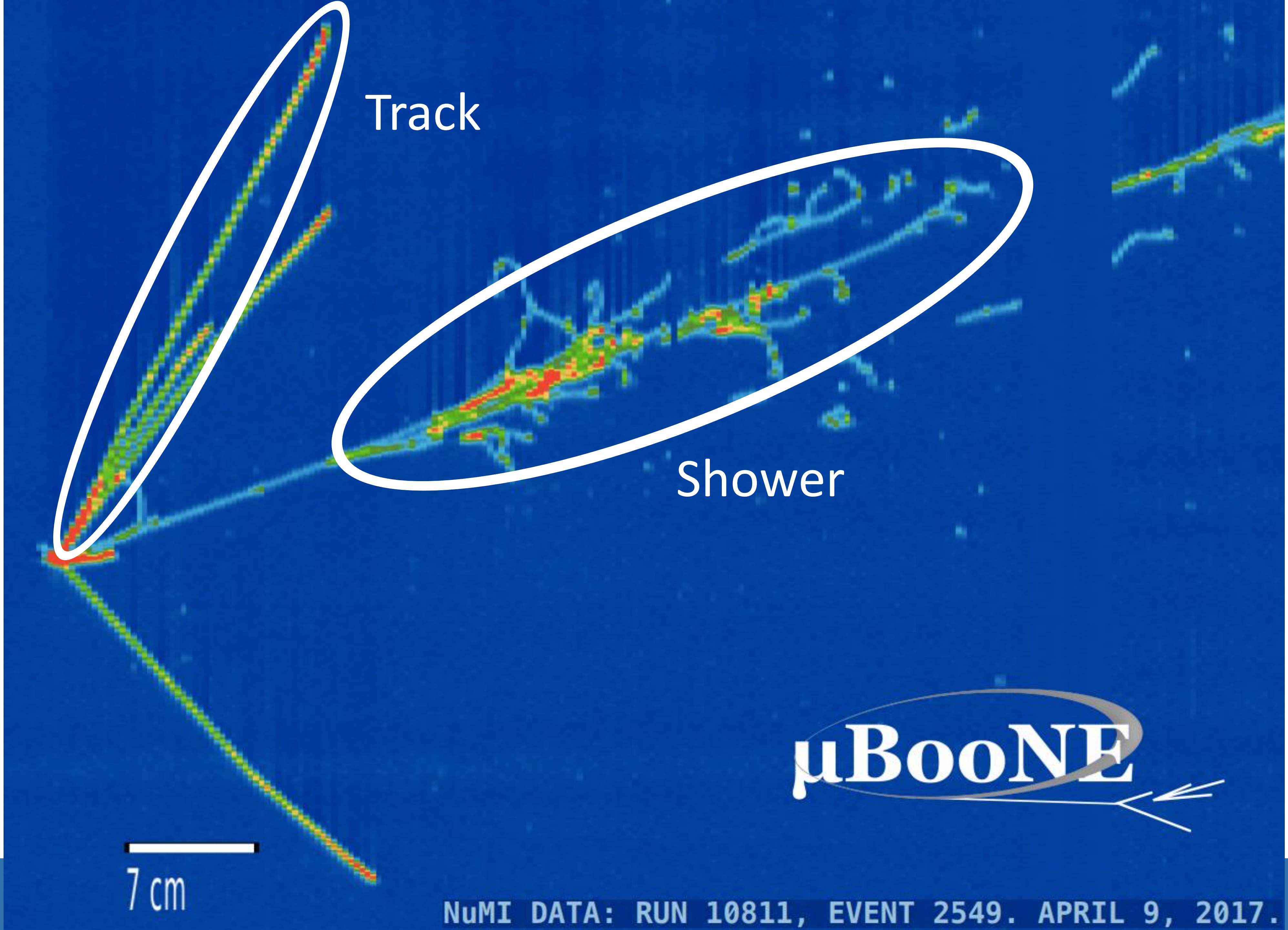
Liquid Argon Time Projection Chamber (LArTPC)

- Detector for HEP experiments
  - Ongoing neutrino research
  - Particle interaction images



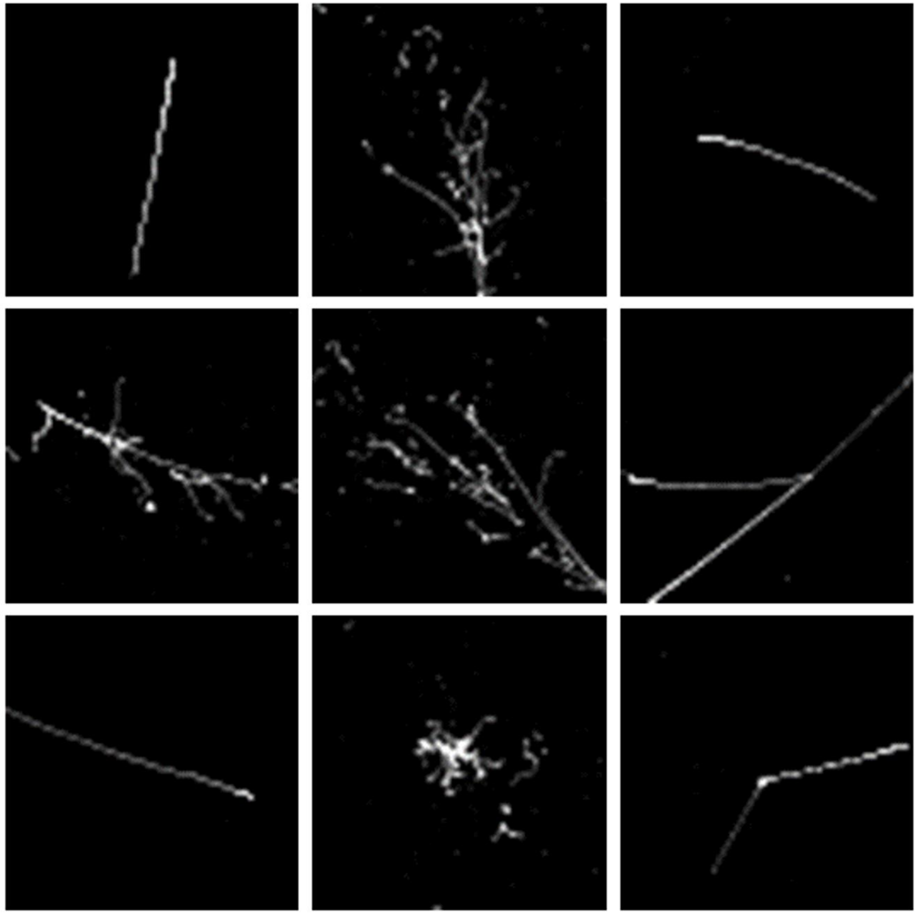

µBooNE
7 cm
NuMI DATA: RUN 10811, EVENT 2549. APRIL 9, 2017.


Track
7 cm
μBooNE
NuMI DATA: RUN 10811, EVENT 2549. APRIL 9, 2017.


Track
Shower
µBooNE
7 cm
NuMI DATA: RUN 10811, EVENT 2549. APRIL 9, 2017.

LArTPC Images

- Cropped image from detector
- Globally sparse, but locally dense

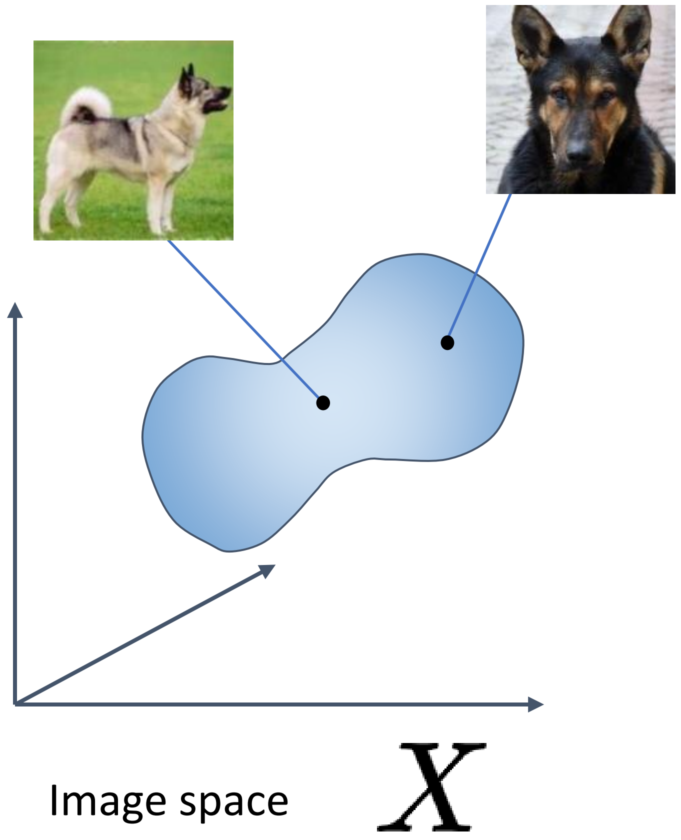


Why Generative Modeling

- Observing rare neutrino events requires analyzing large datasets
- Potential to be faster than traditional simulation methods
- New tool for reconstruction and analyses
- Another way of understanding our data
- Proof of concept ML application

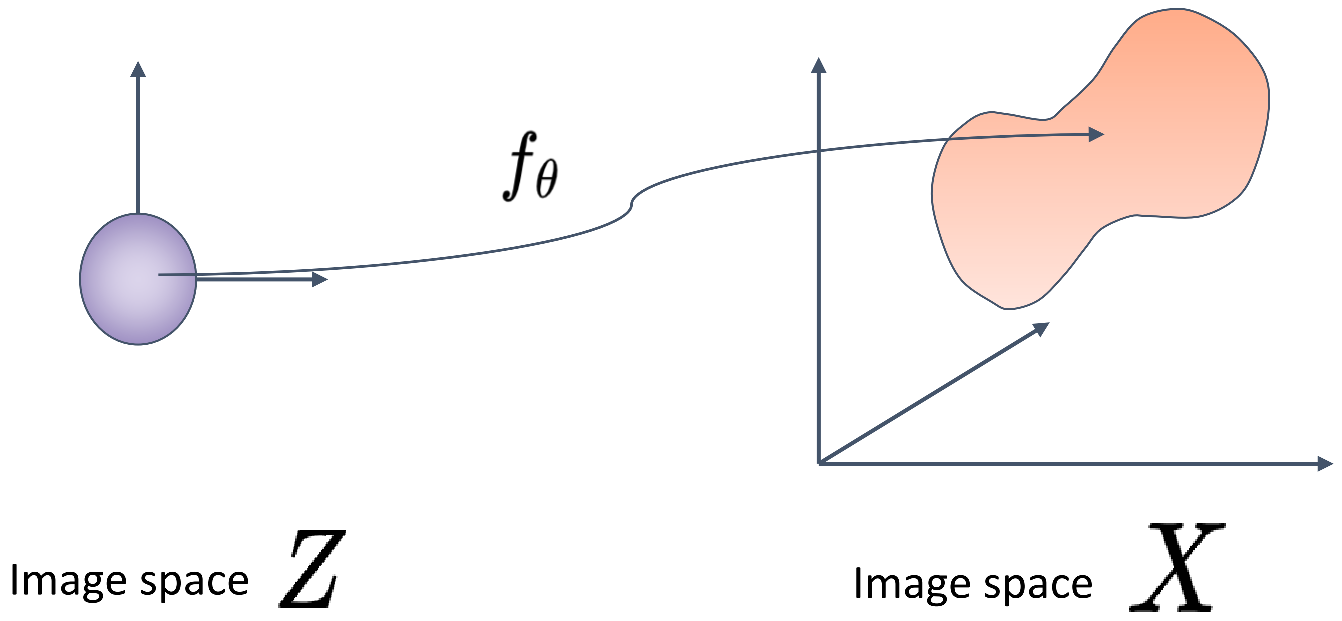
How to Generate Images

- Our data **x** is sampled from some p(**x**)
- We don't know p(**x**) directly



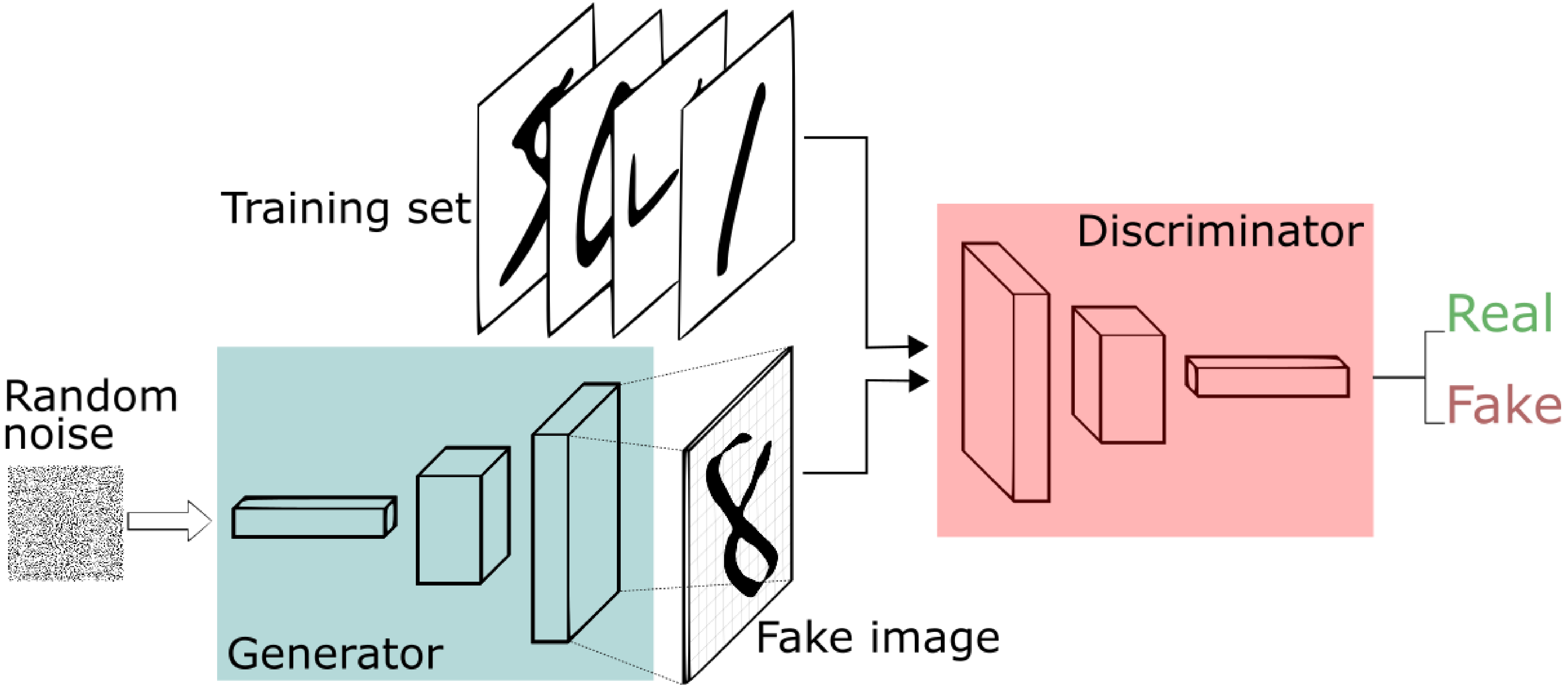
How to Generate Images

- Instead, we sample from a known distribution $z \sim \mathcal{N}(0, 1)$
- Learn a mapping $x = f_\theta(z)$

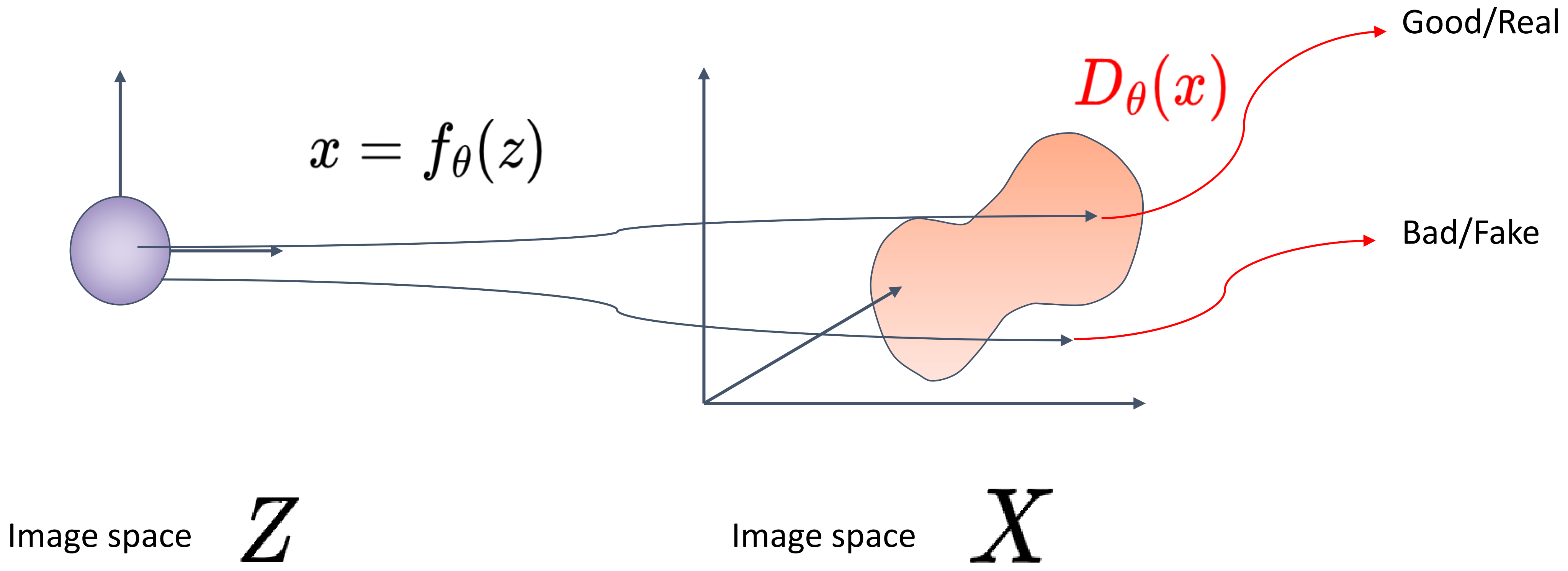


$$p(x) = p(f_\theta(z))$$

Attempt 1: Generative Adversarial Network



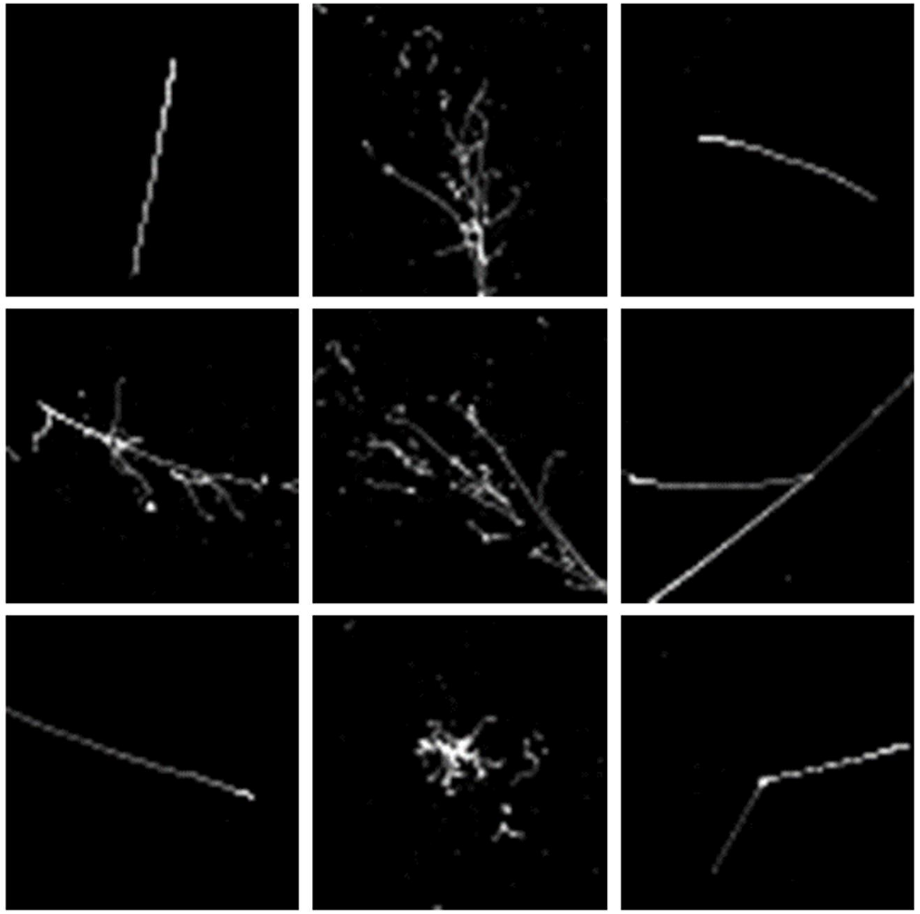
GAN Mapping



$x = f_\theta(z)$
$D_\theta(x)$
Good/Real
Bad/Fake
Image space Z
Image space X

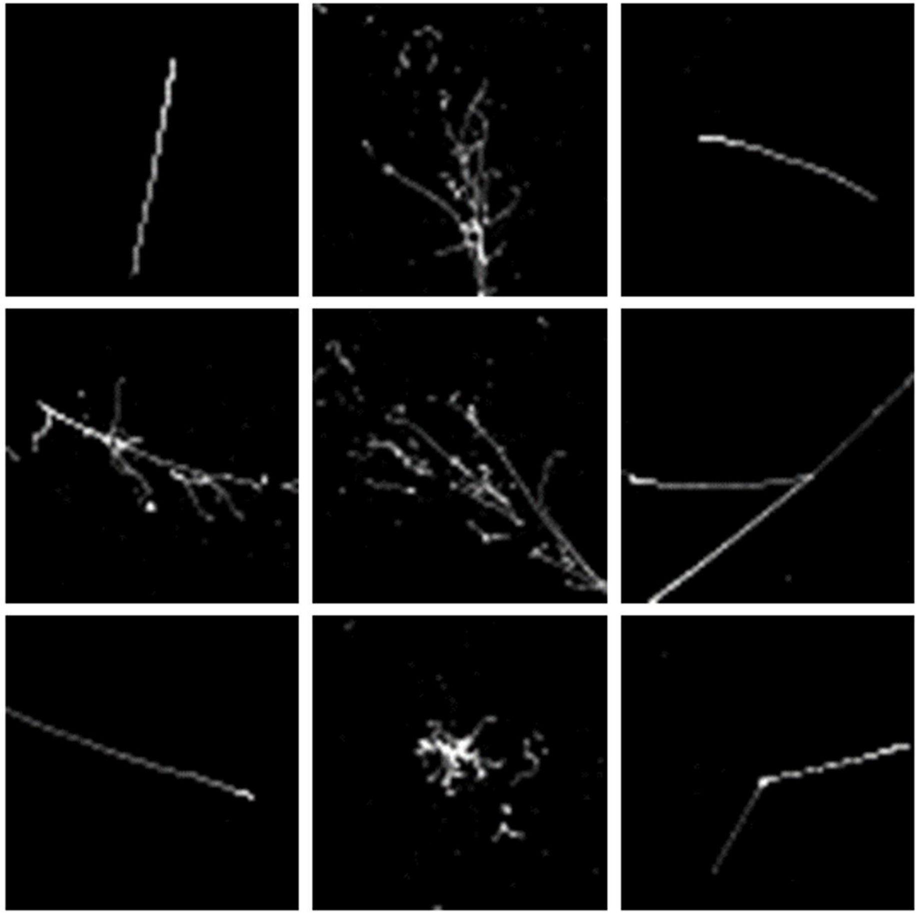
LArTPC GAN

Validation LArTPC Data

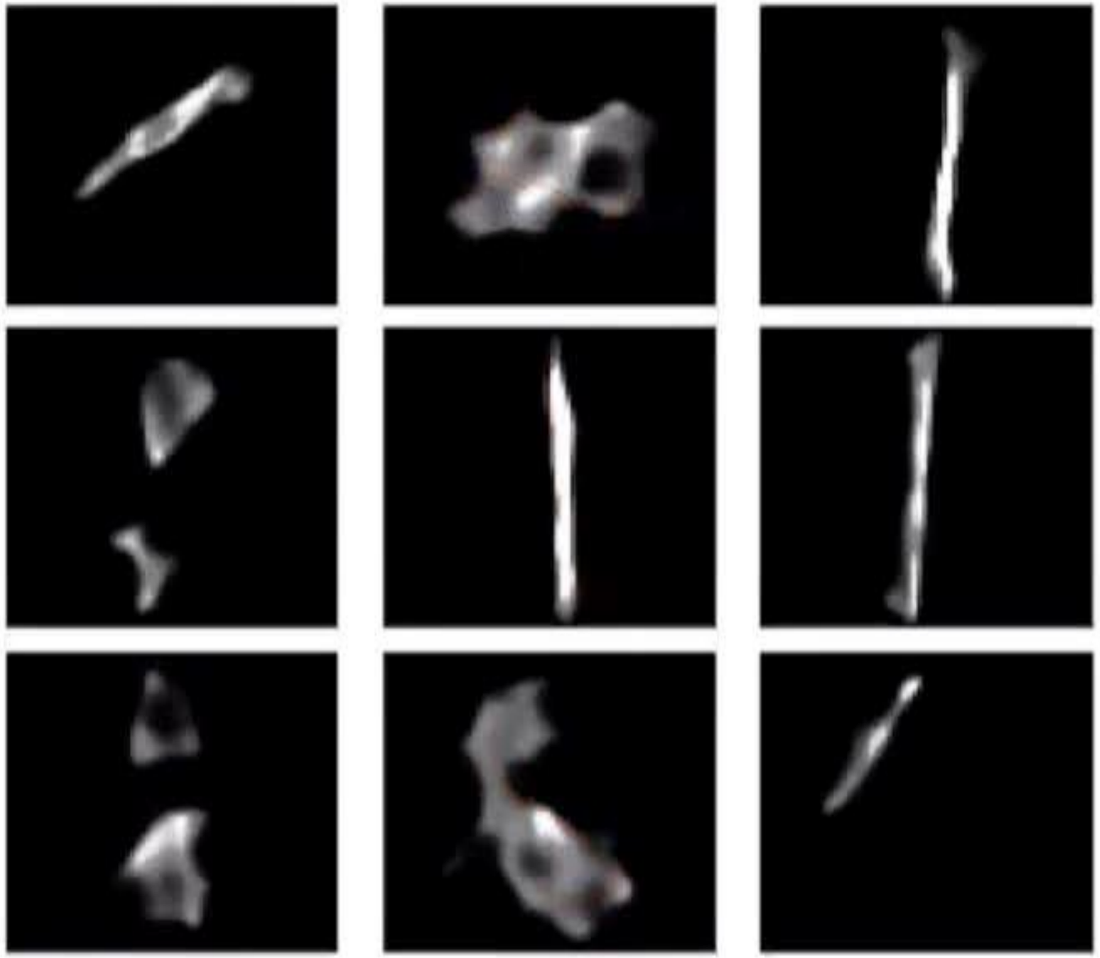


LArTPC GAN

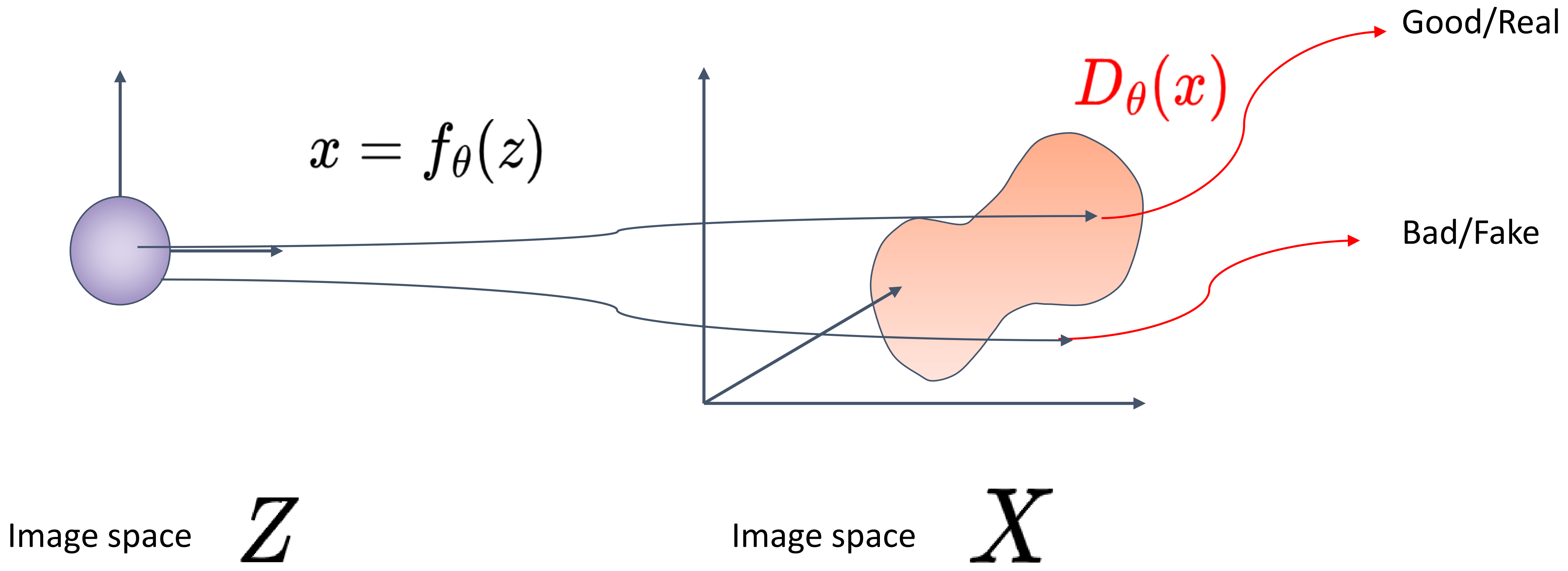
Validation LArTPC Data



GAN Generated



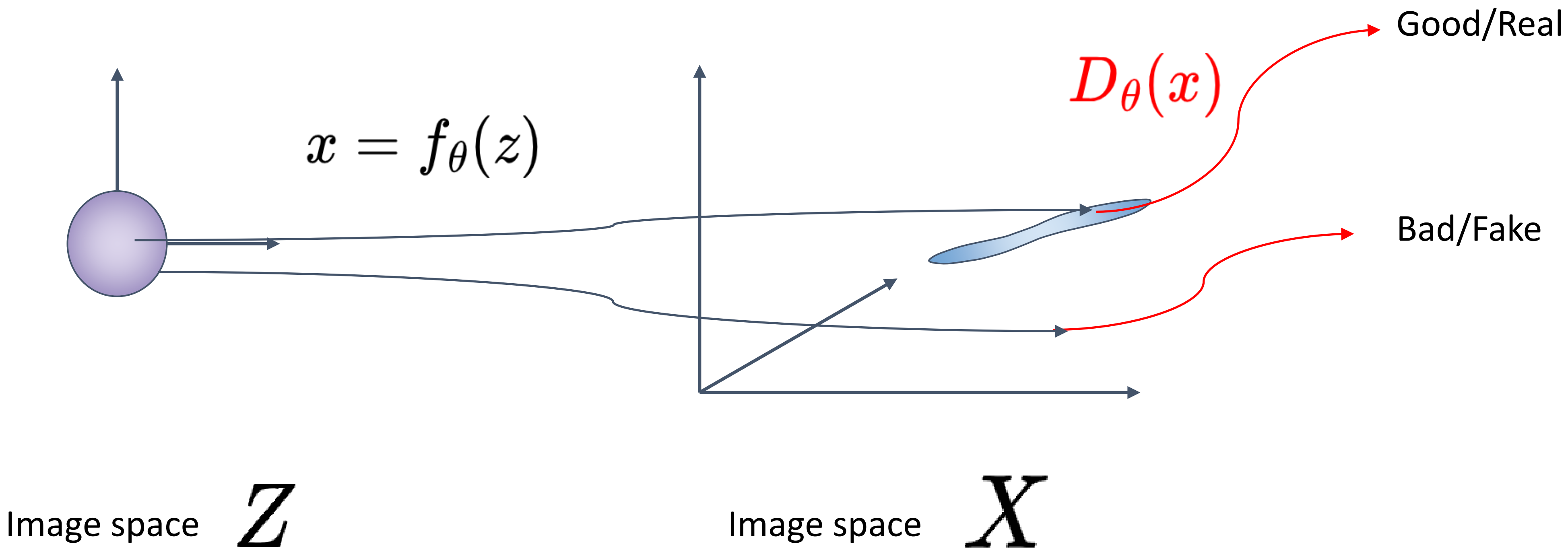
GAN Mapping



$x = f_\theta(z)$
$D_\theta(x)$
Good/Real
Bad/Fake
Image space Z
Image space X

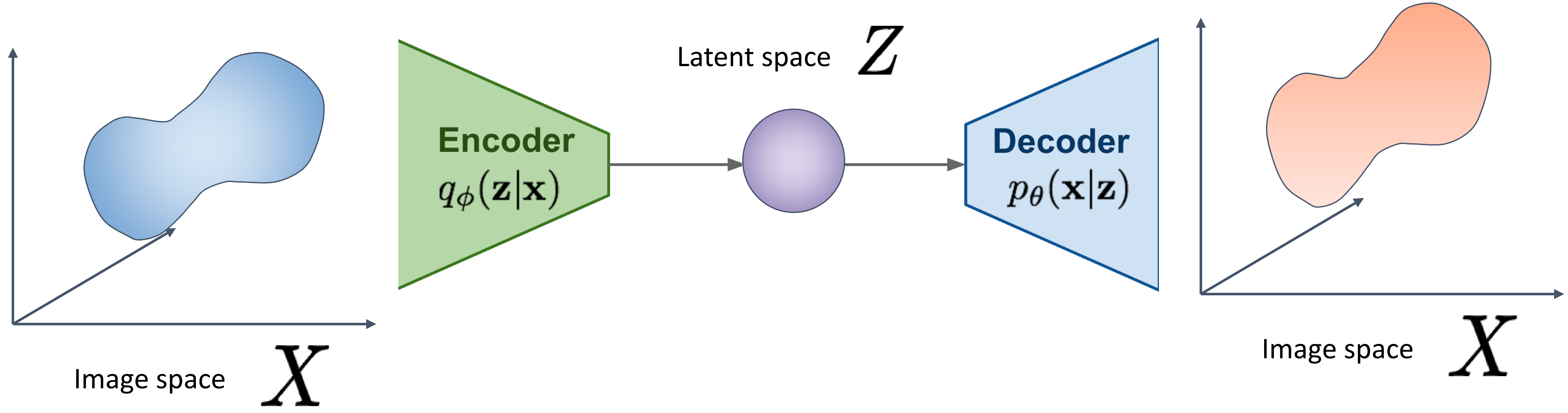
GAN Mapping

- LArTPC images exist as thin manifold in image space



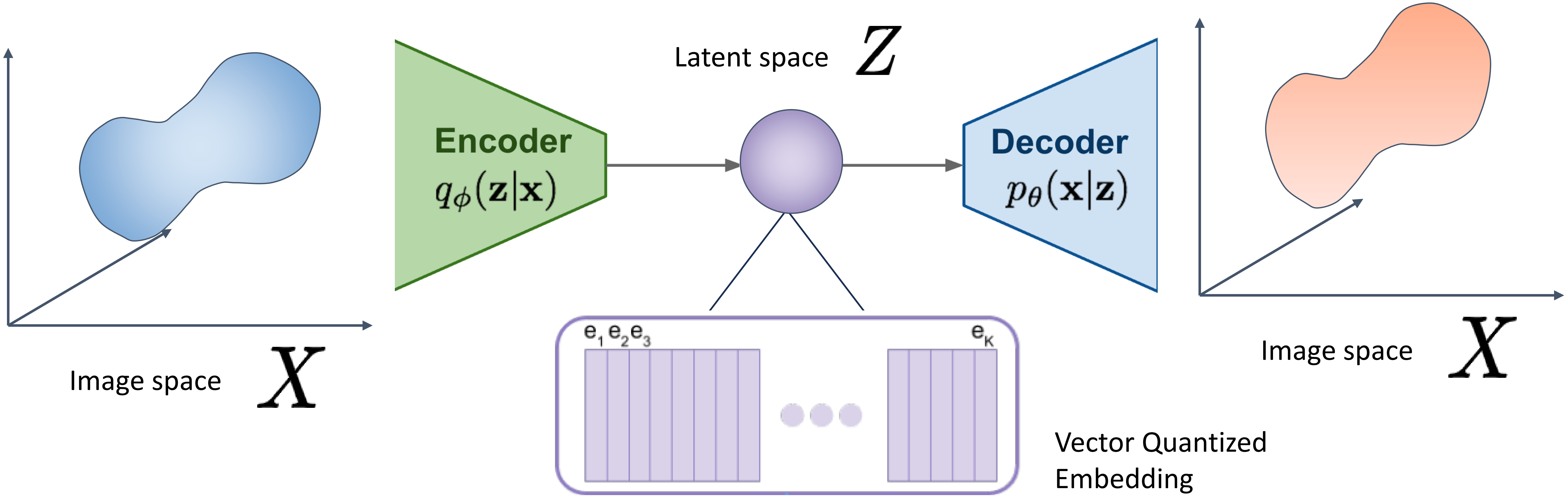
Attempt 2: VQ-VAE

- Vector Quantized Variational Autoencoder



Attempt 2: VQ-VAE

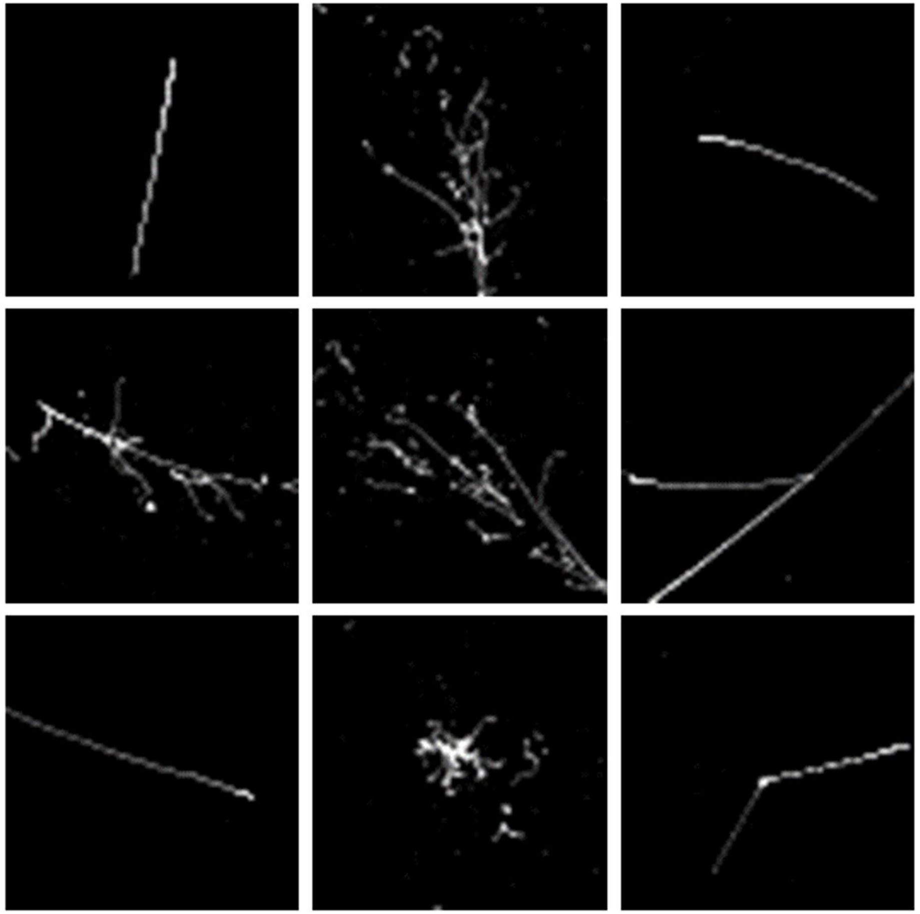
- Vector Quantized Variational Autoencoder



Oord, Vinyals, & Kavukcuoglu; arXiv:1711.00937

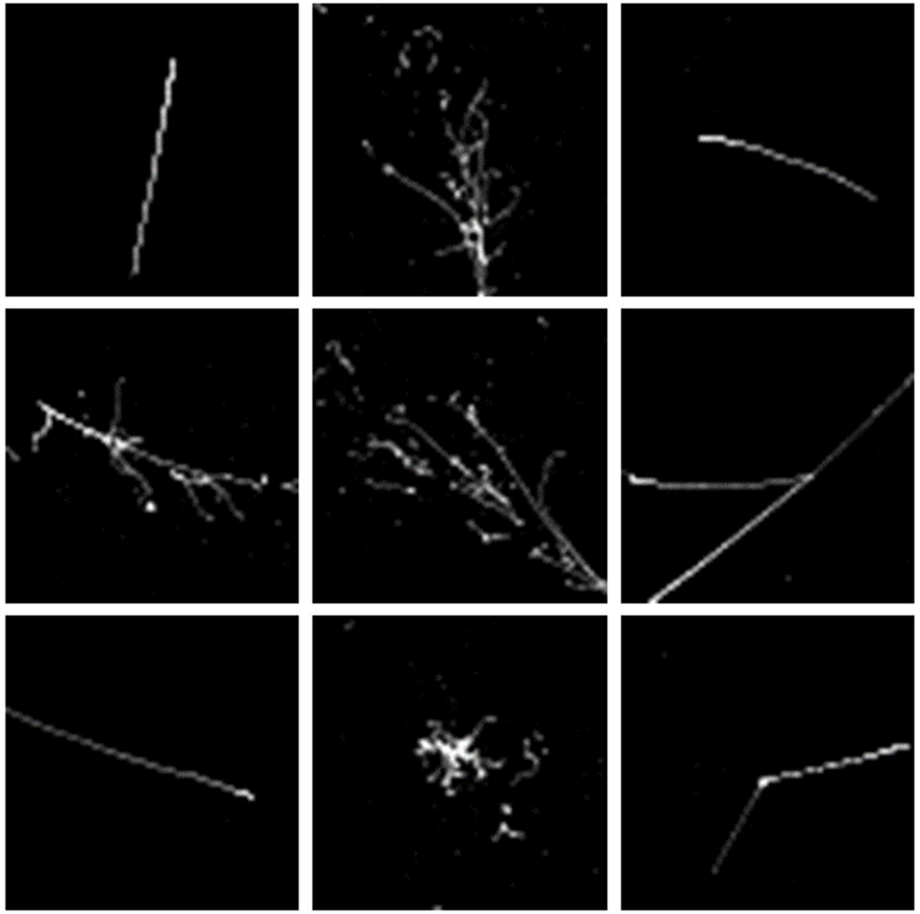
LArTPC VQ-VAE

Validation LArTPC Data

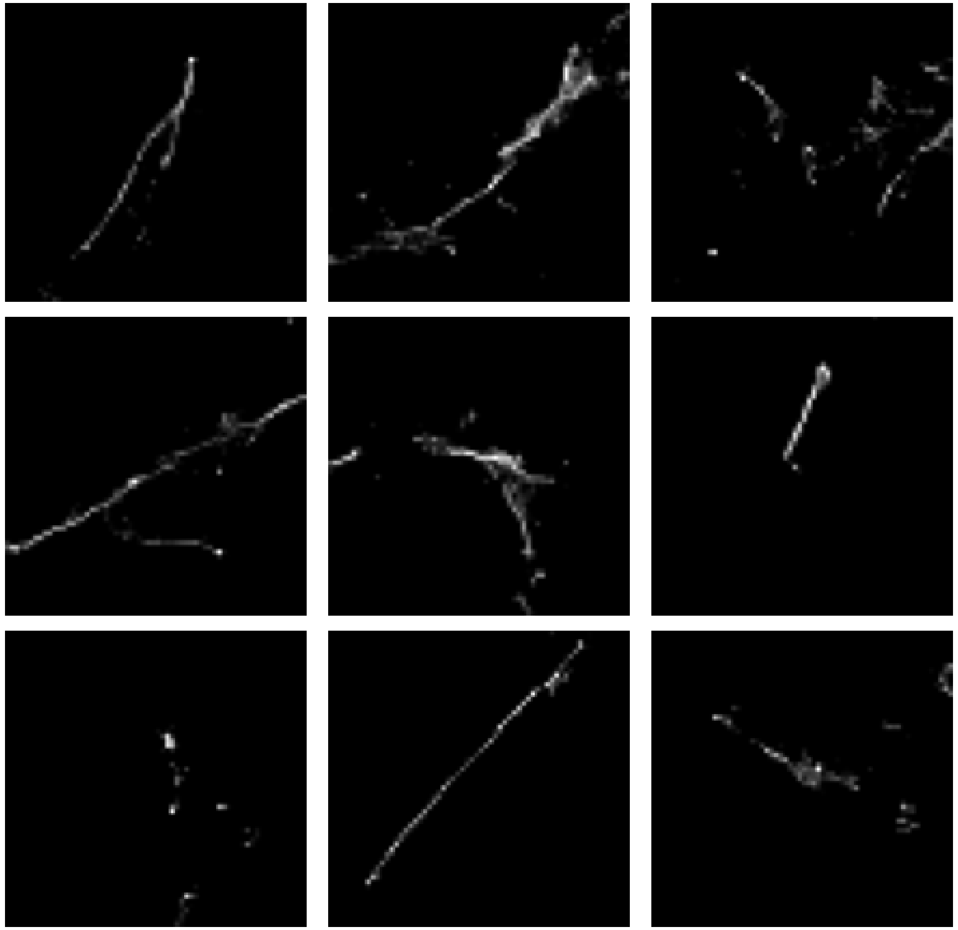


LArTPC VQ-VAE

Validation LArTPC Data



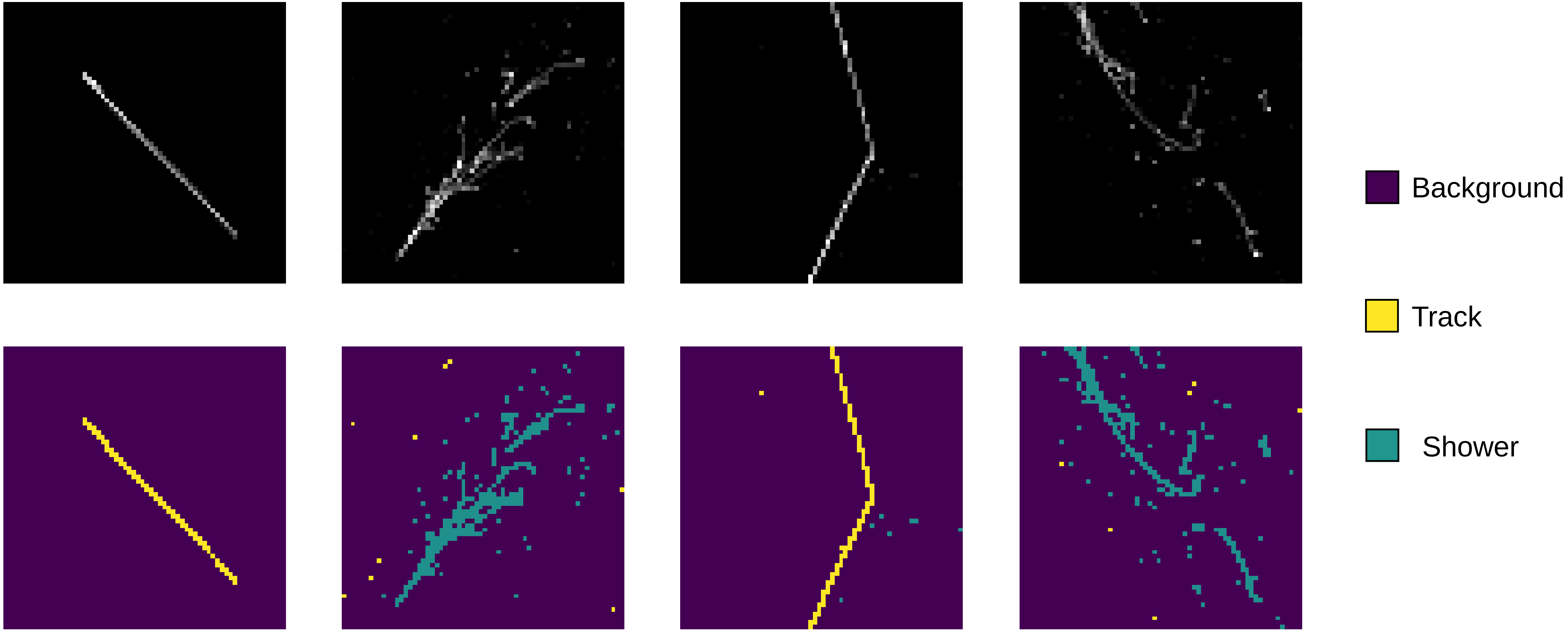
VQ-VAE Generated



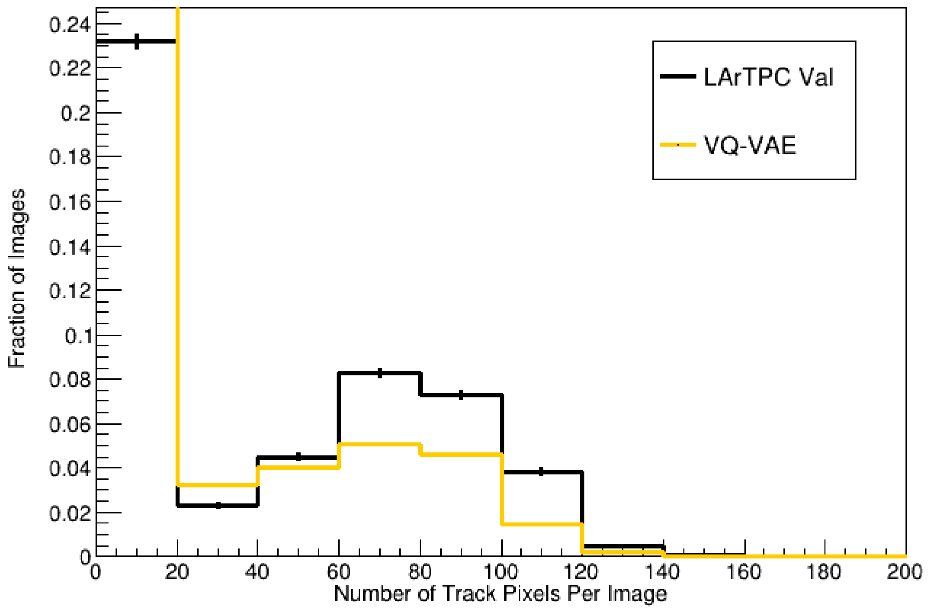
What is Good Enough?

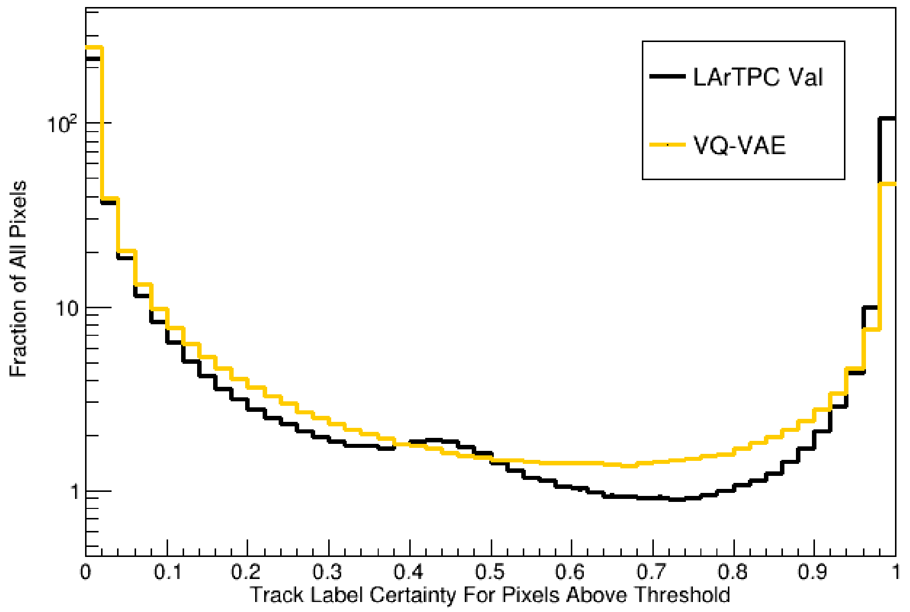
- No standard quality tests for LArTPC images
- 64x64 are too small for traditional physics analysis
- We developed several options

Semantic Segmentation Network (SSNet)

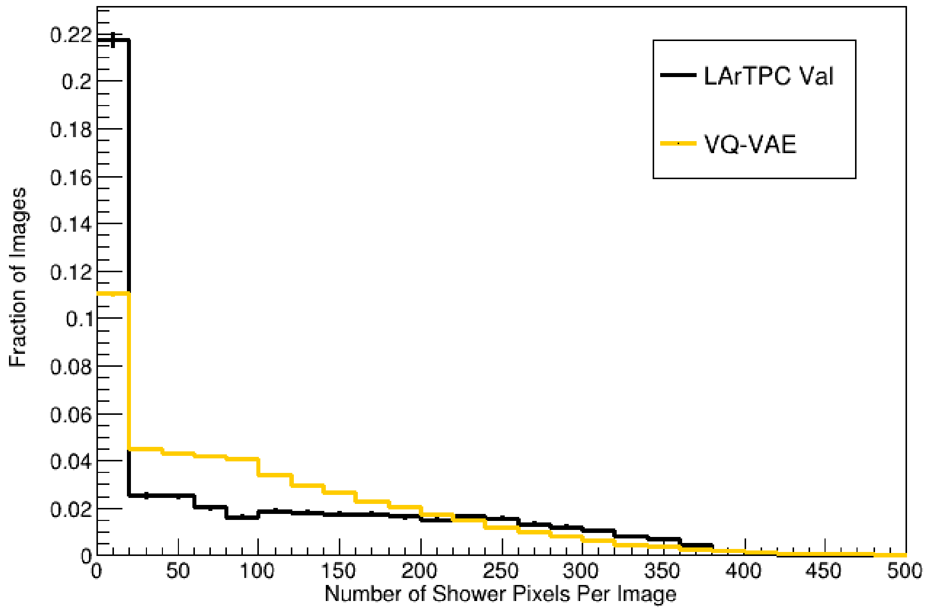


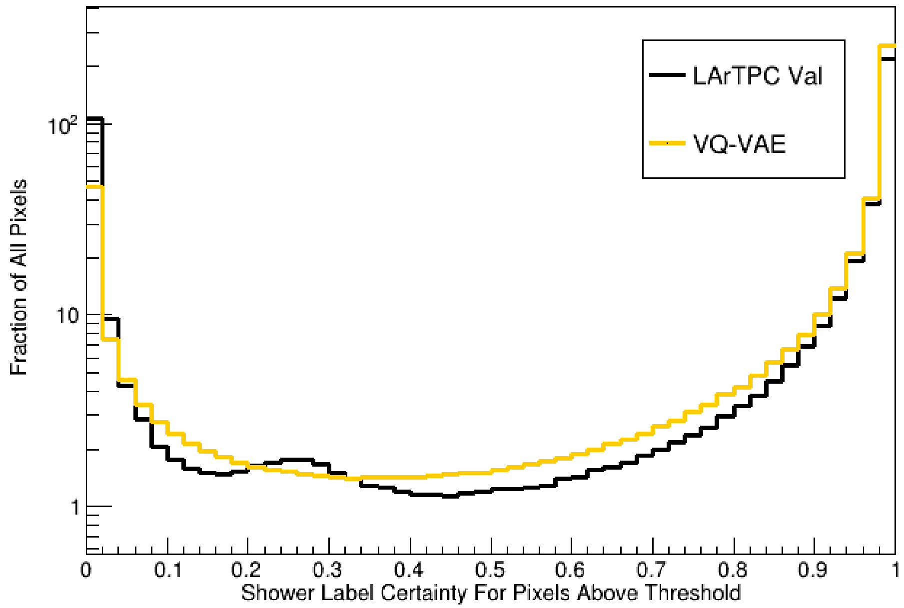
Tracks





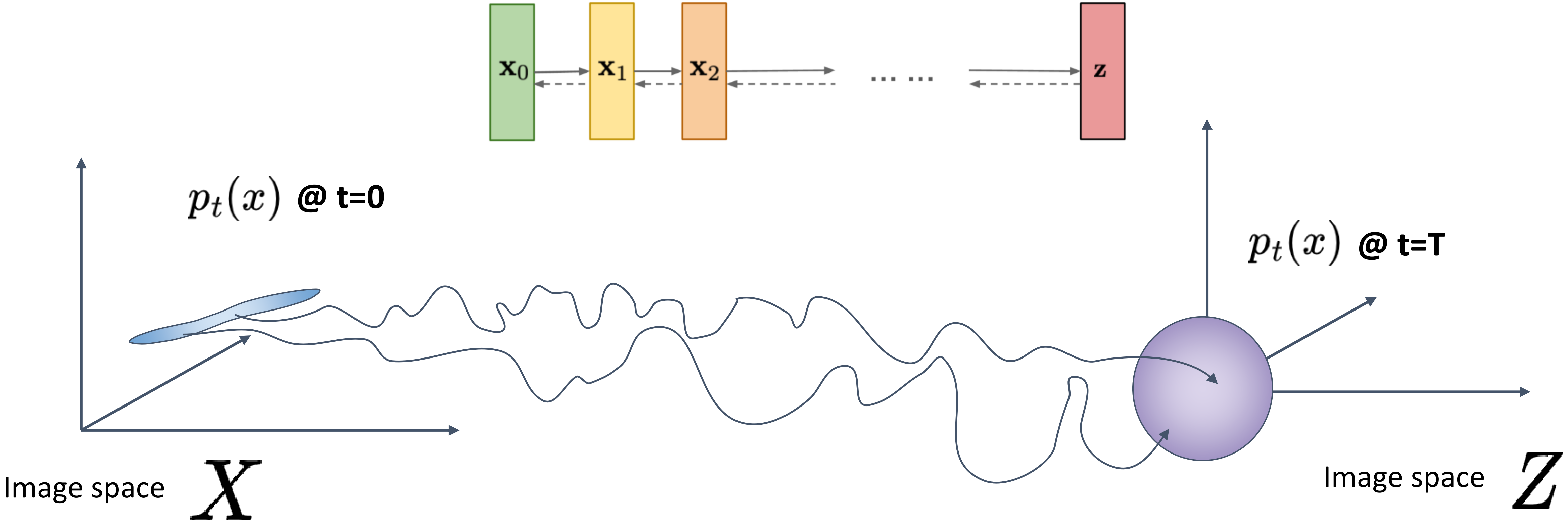
Showers





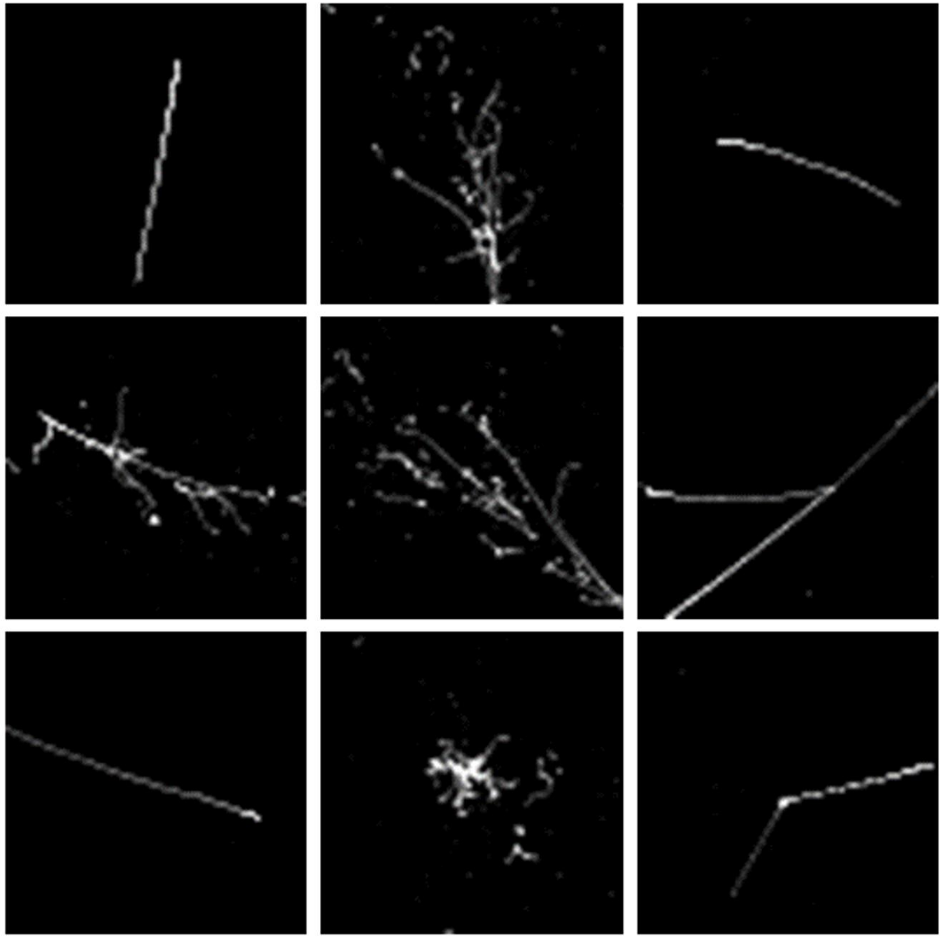
Imani, Aeron, & Wongjirad; PhysRevD.109.072011

Attempt 3: Diffusion



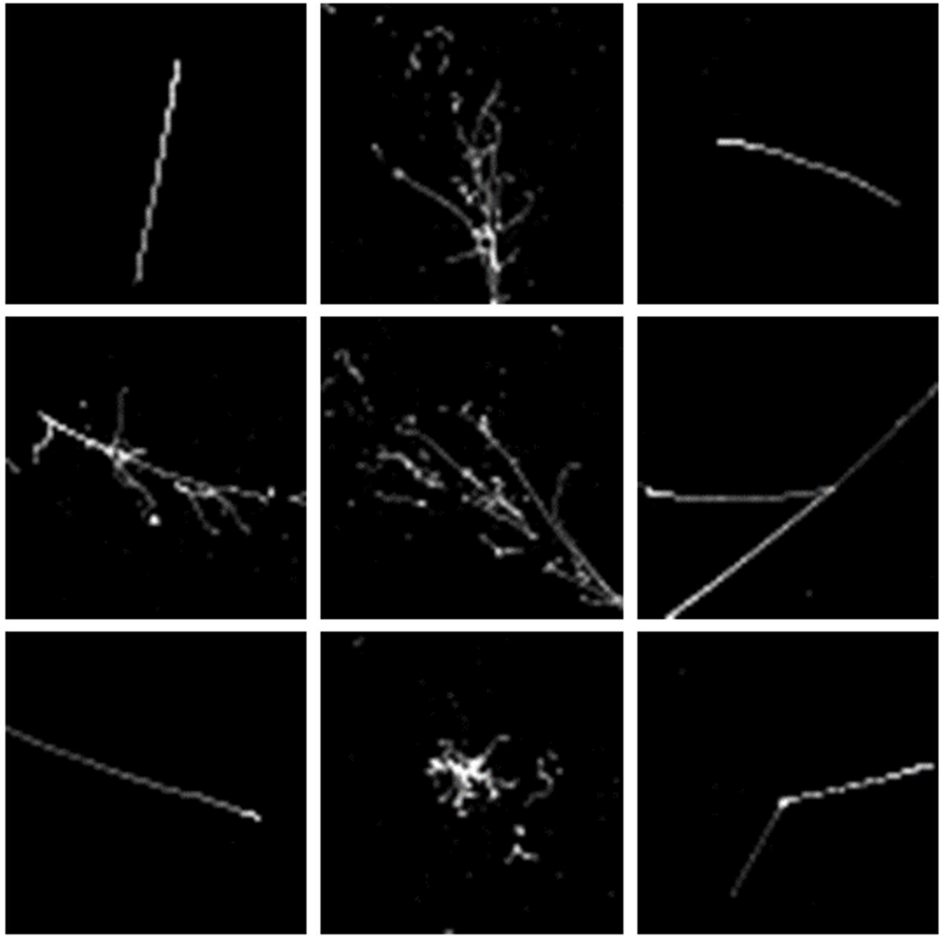
Attempt 3: Diffusion

Validation LArTPC Data

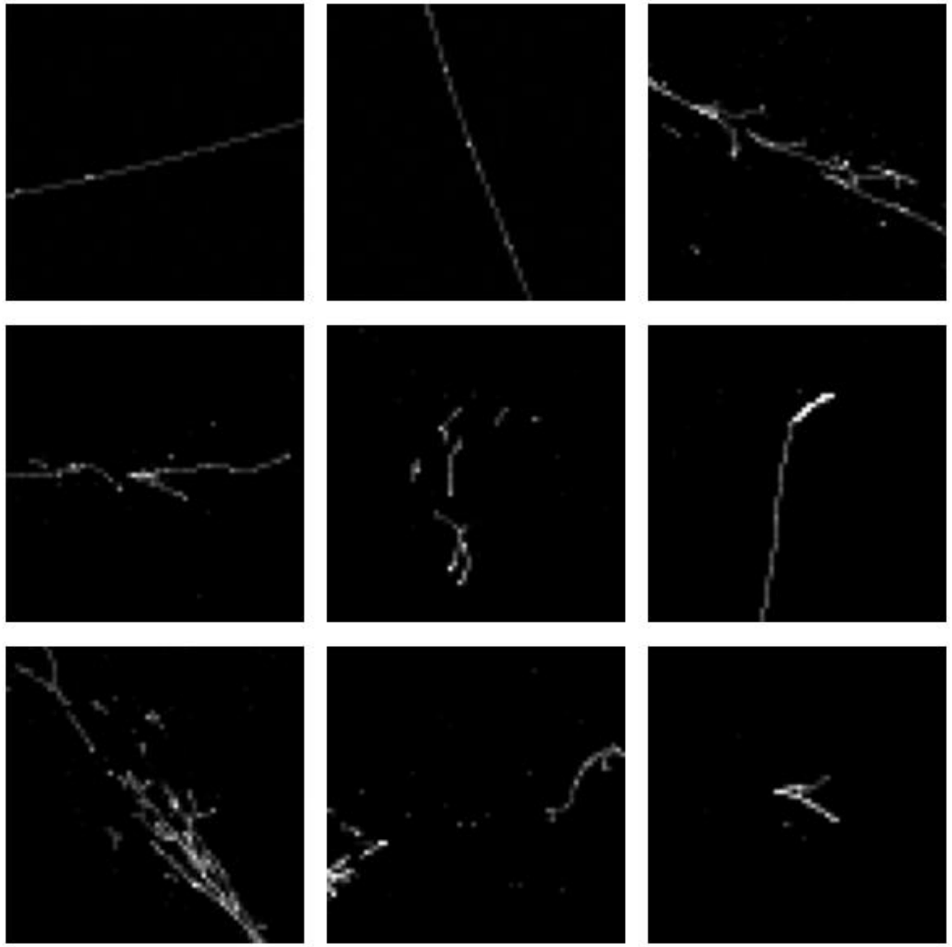


Attempt 3: Diffusion

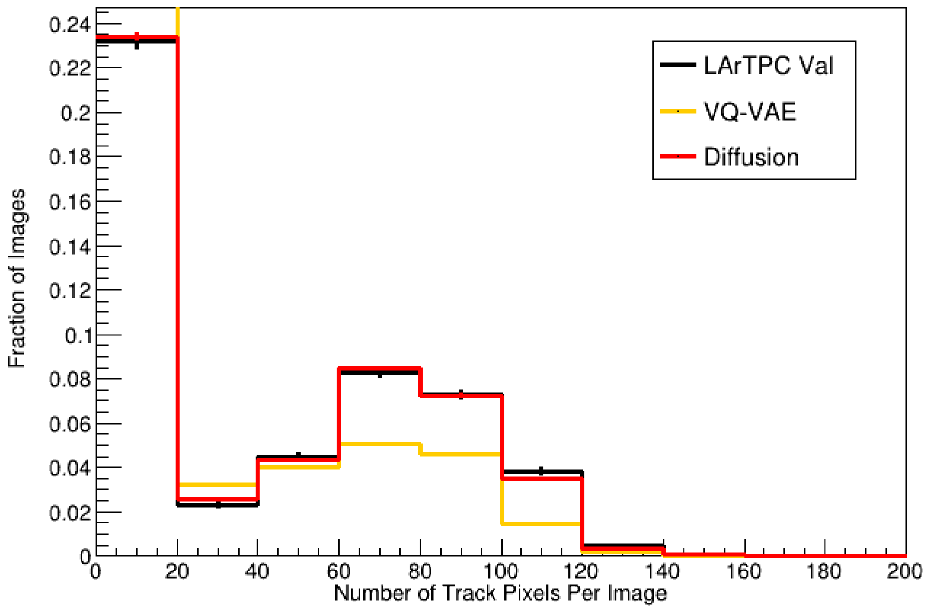
Validation LArTPC Data

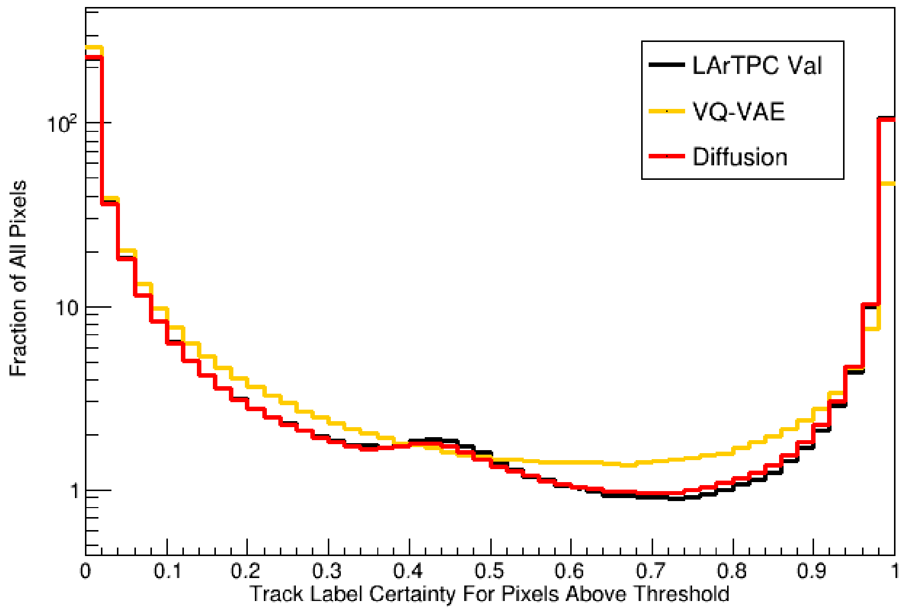


Diffusion Generated

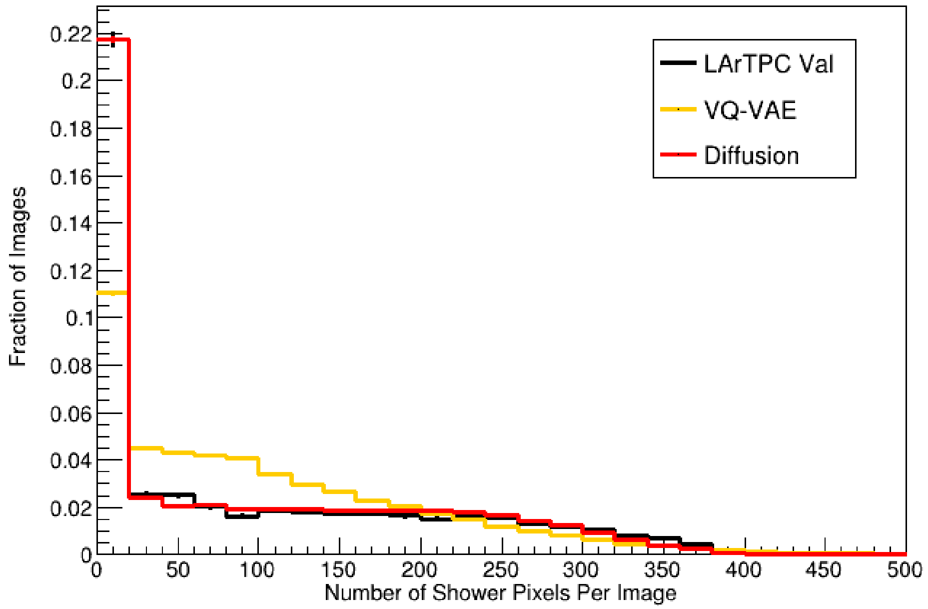


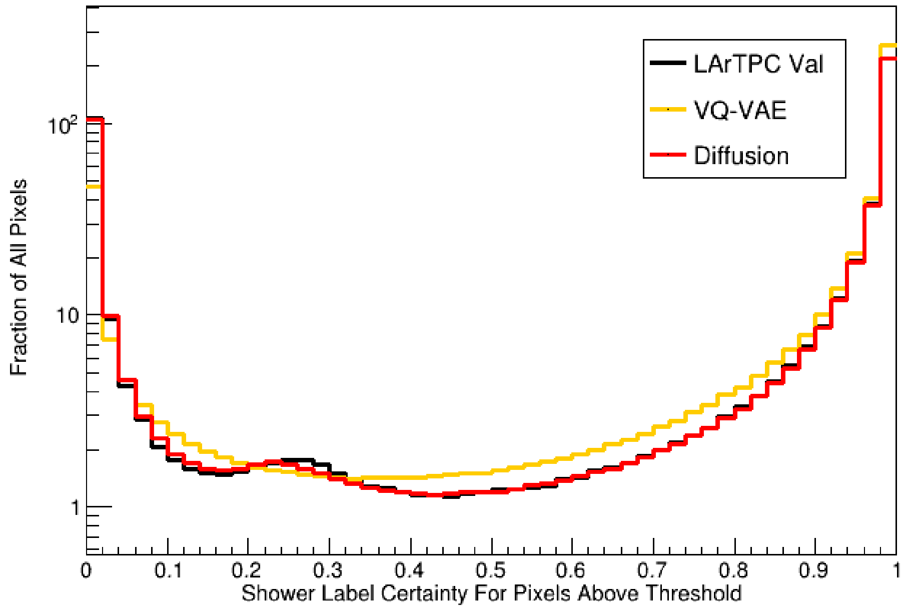
Tracks



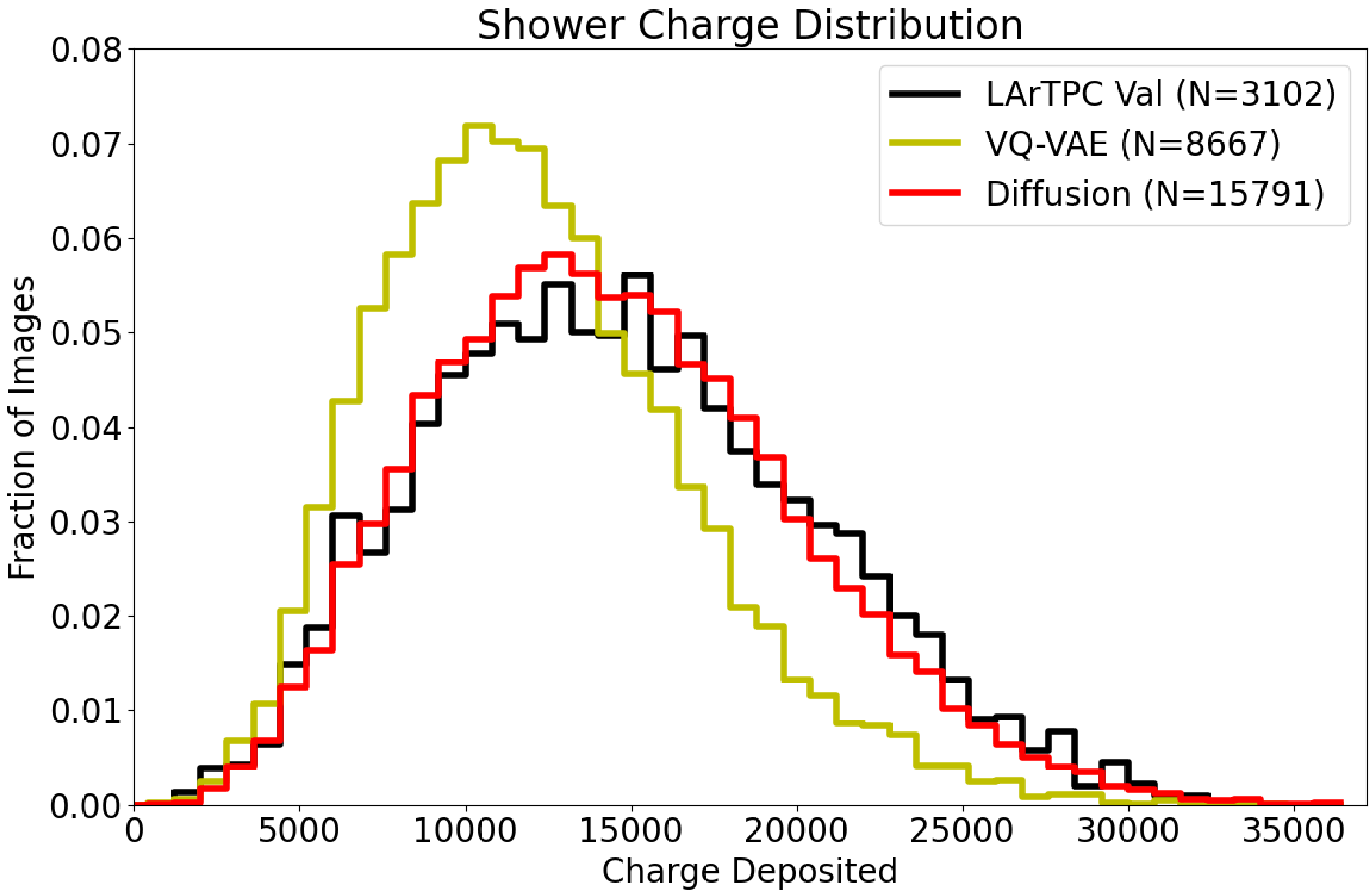


Showers

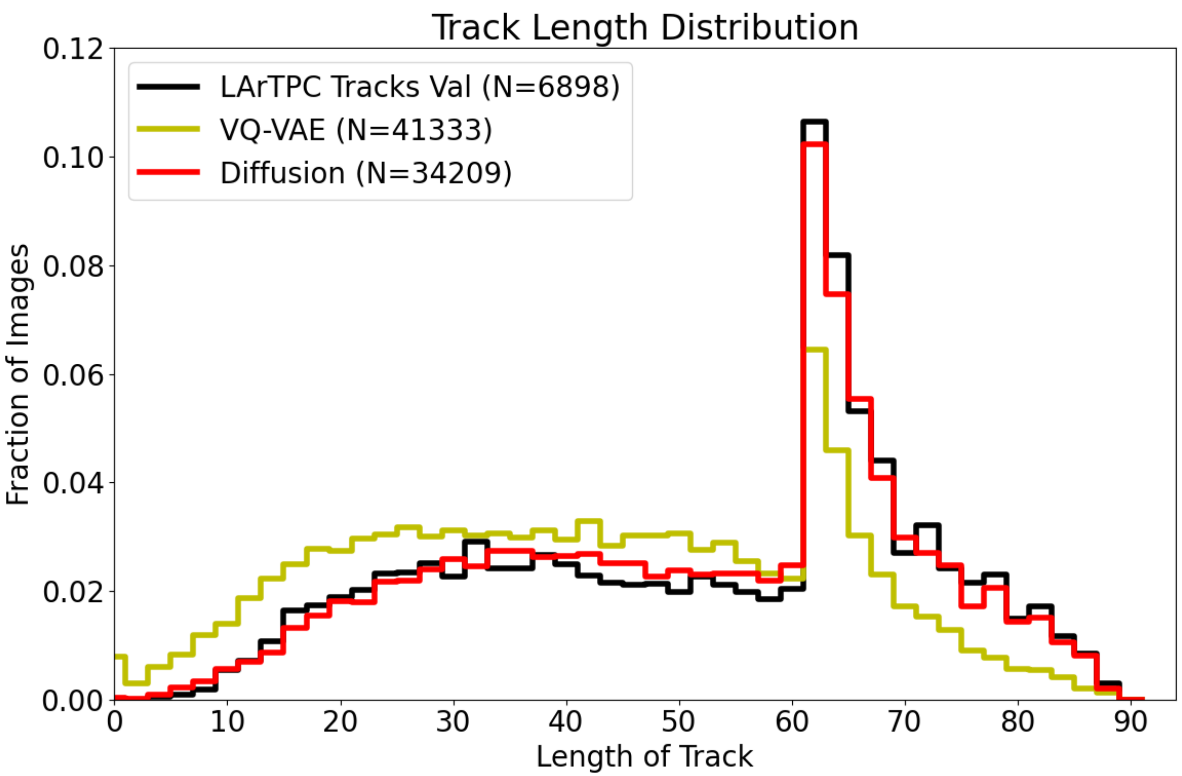


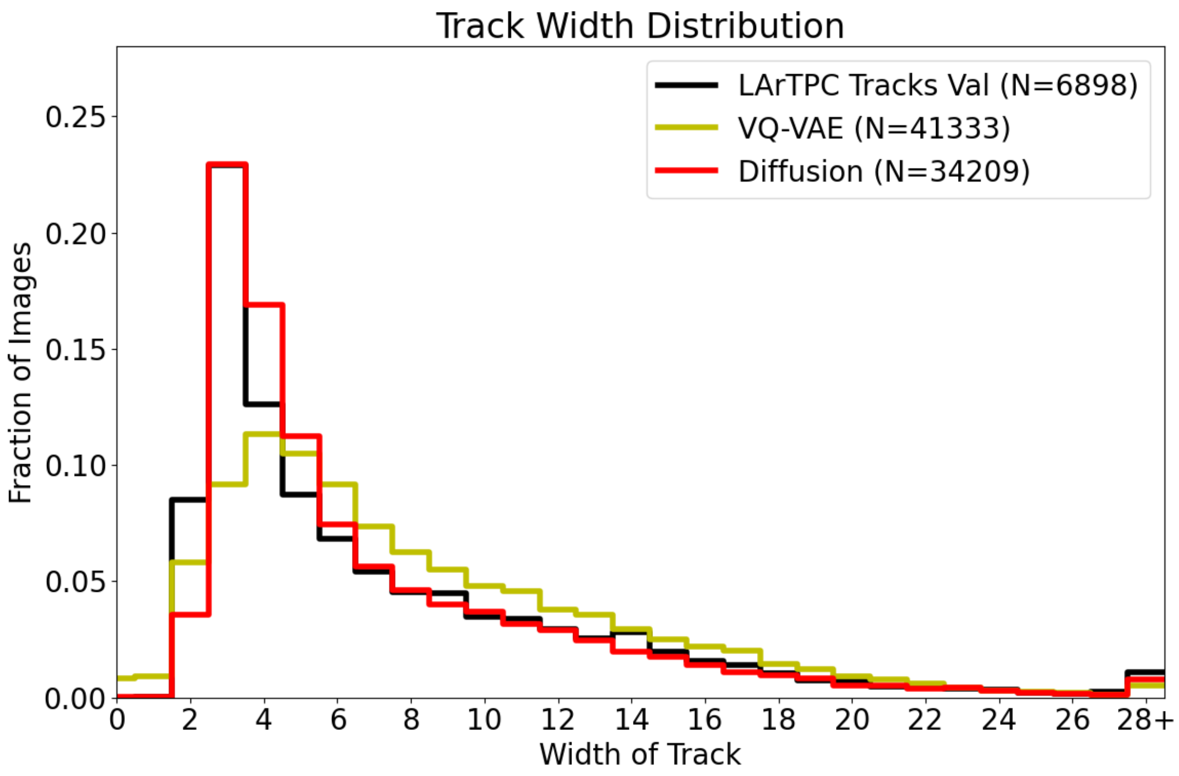


Physics Quality Tests: Showers



Physics Quality Tests: Tracks





Additional Quality Tests

- High dimensional goodness of fit tests
  - Maximum Mean Discrepancy (MMD)
  - Sinkhorn divergence
  - Wasserstein-1 (EMD)

- SSNet-FID

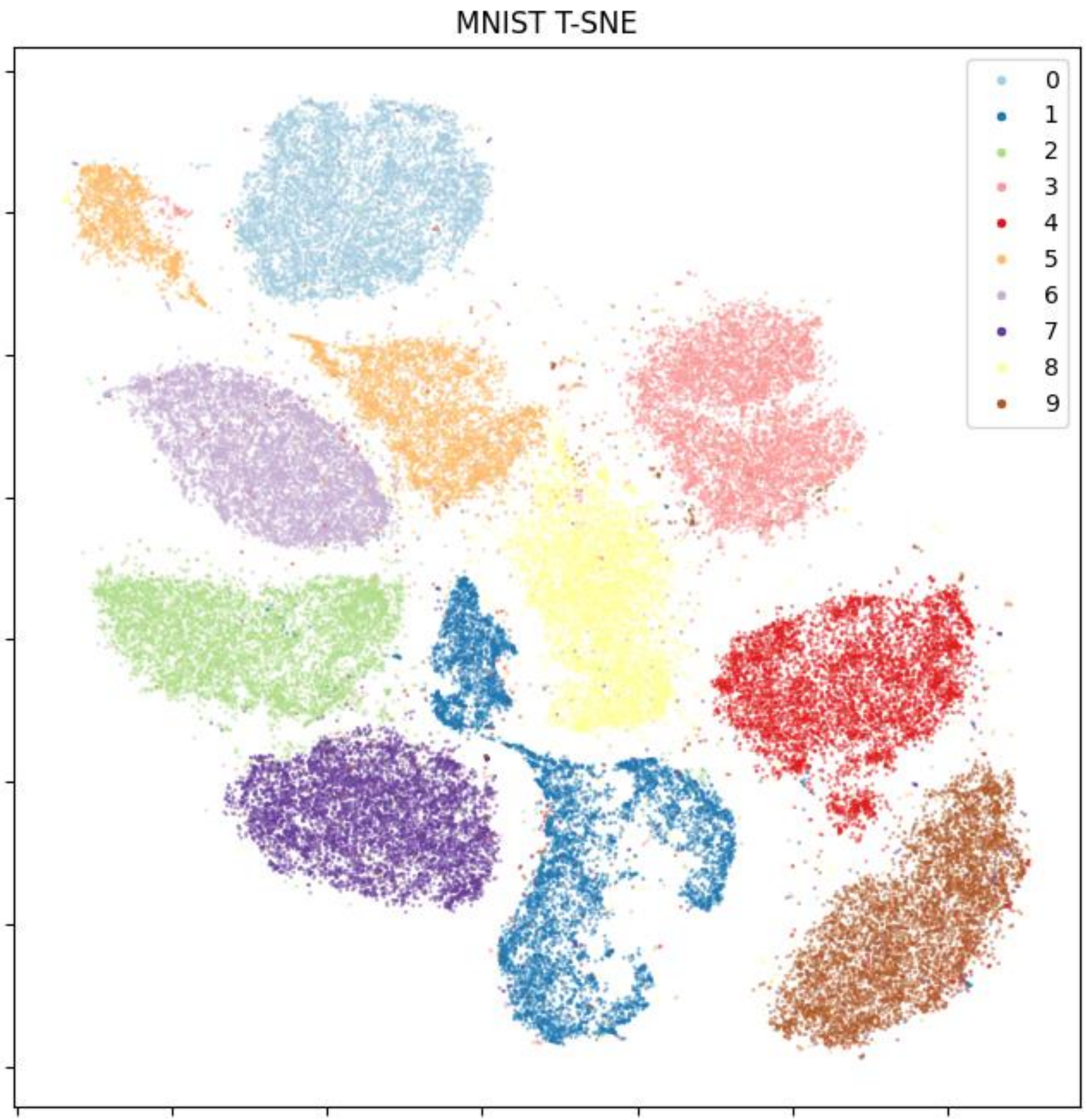
- Turing test survey

Next Steps

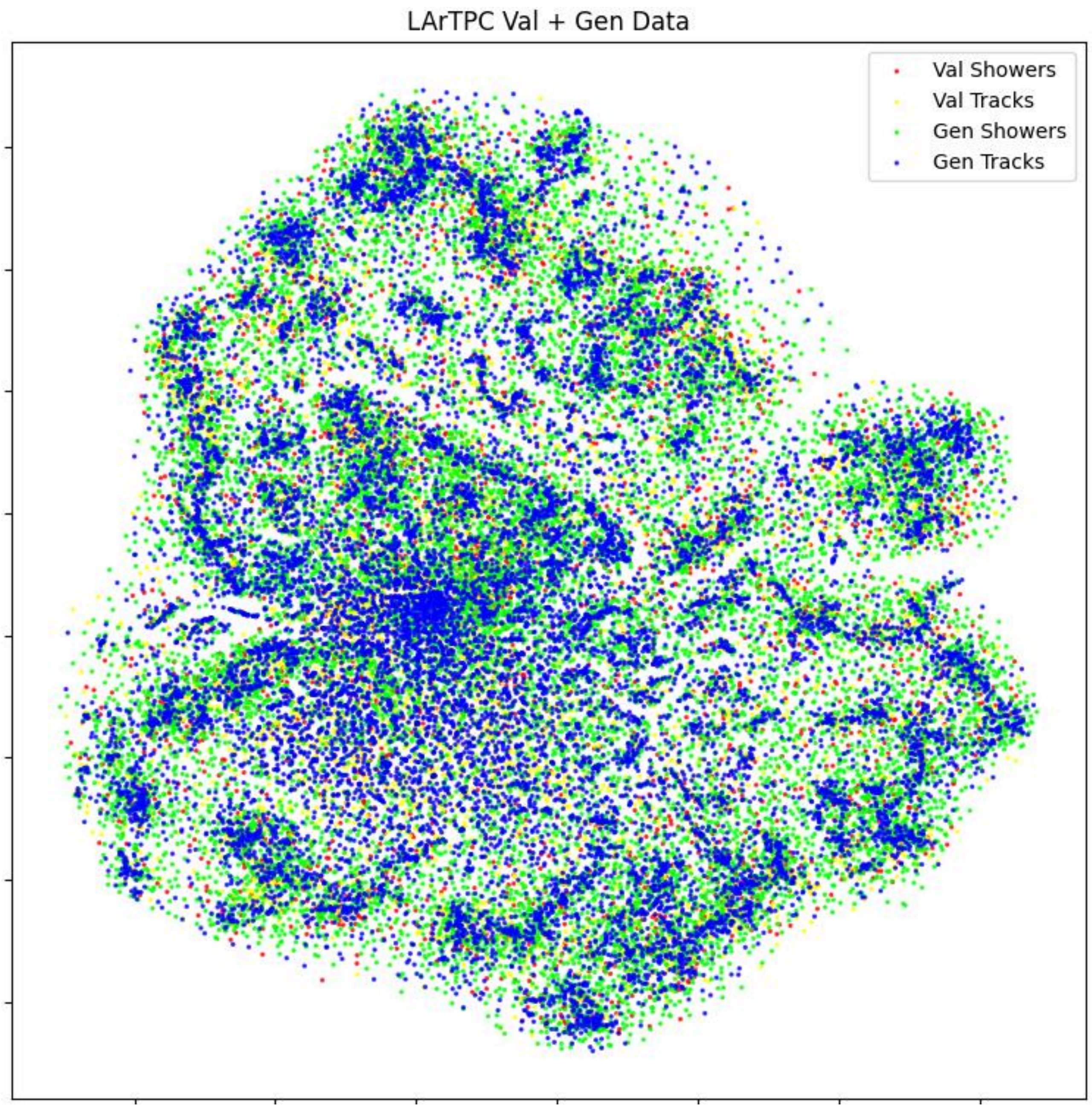
- Scale up to larger images
  - Goal of 512x512 image size to do physics analyses
  - Use latent diffusion to overcome scaling issue
- Conditional generation on energy and particle type
- Improve generation speed and efficiency

Visualizing Distributions

- T-distributed Stochastic Neighbor Embedding (T-SNE)
- Nonlinear dimensionality reduction, maintains relative distance



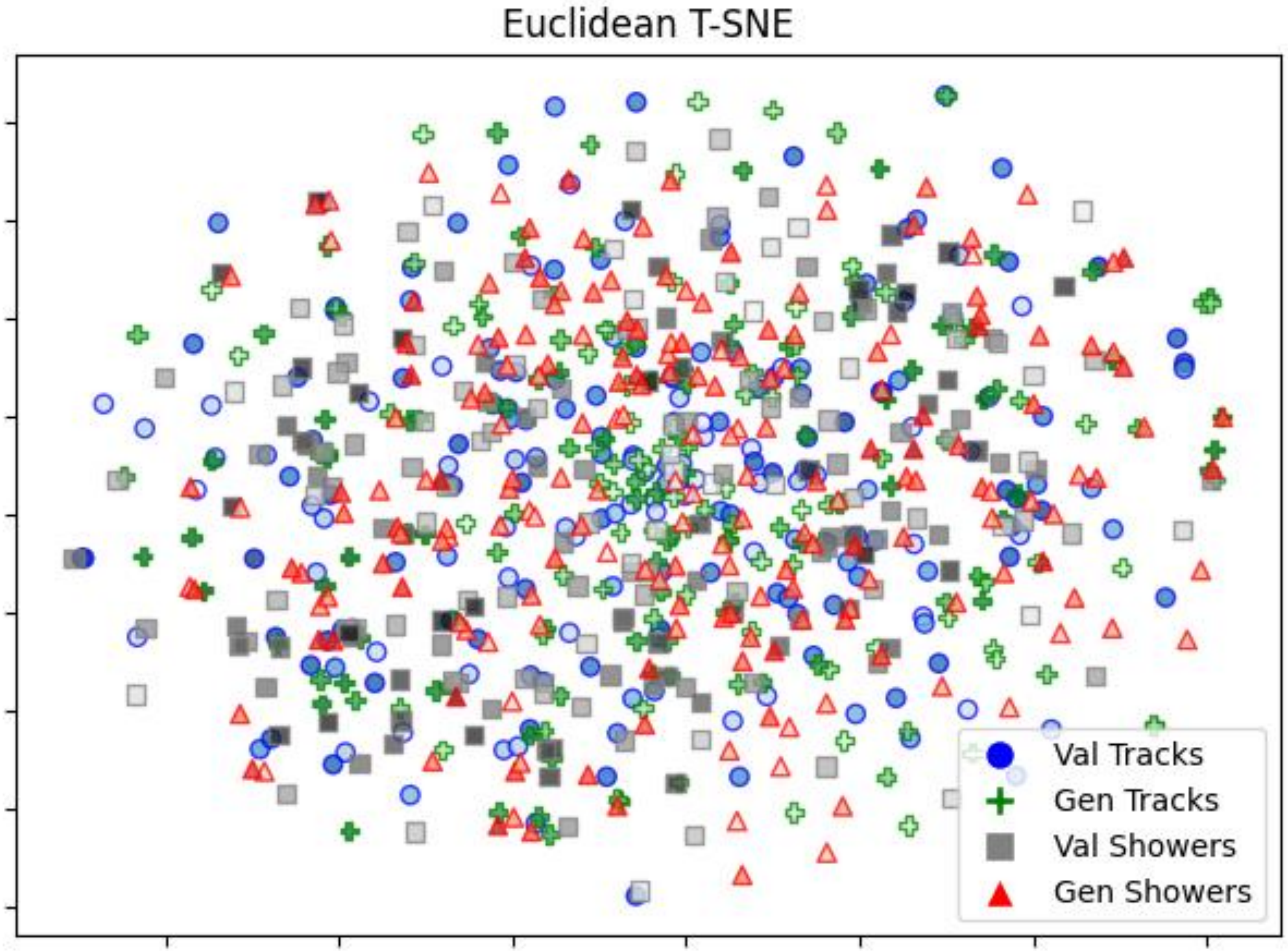
T-SNE on LArTPC



- Pretty, but no clear structure

T-SNE on LArTPC

- Darker points = longer/more charge



Digression: Distance Metrics

- Euclidian distance (L2 norm) $\|\boldsymbol{x}\|_2 := \sqrt{x_1^2 + \cdots + x_n^2}$

Digression: Distance Metrics

- Euclidian distance (L2 norm)

$$\|\boldsymbol{x}\|_2 := \sqrt{x_1^2 + \cdots + x_n^2}$$

- Earth Mover's Distance (EMD)
  - Wasserstein-1 distance
  - 'Natural' metric for particle physics

$$\mathrm{EMD}(P, Q) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{y \sim Q}[f(y)]$$

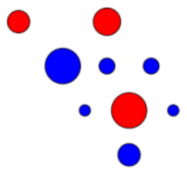
$$\min_F \sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} d_{i,j}$$

Digression: Distance Metrics

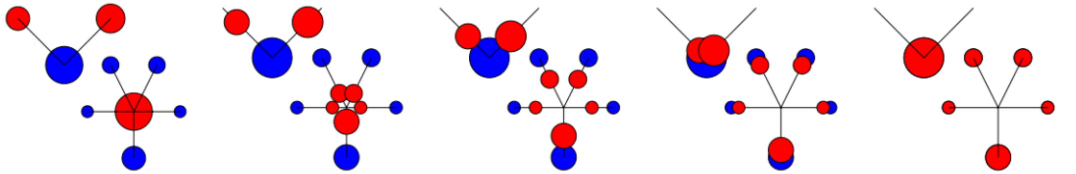
- Euclidian distance (L2 norm)

$$\|\boldsymbol{x}\|_2 := \sqrt{x_1^2 + \cdots + x_n^2}$$

- Earth Mover's Distance (EMD)
  - Wasserstein-1 distance
  - 'Natural' metric for particle physics

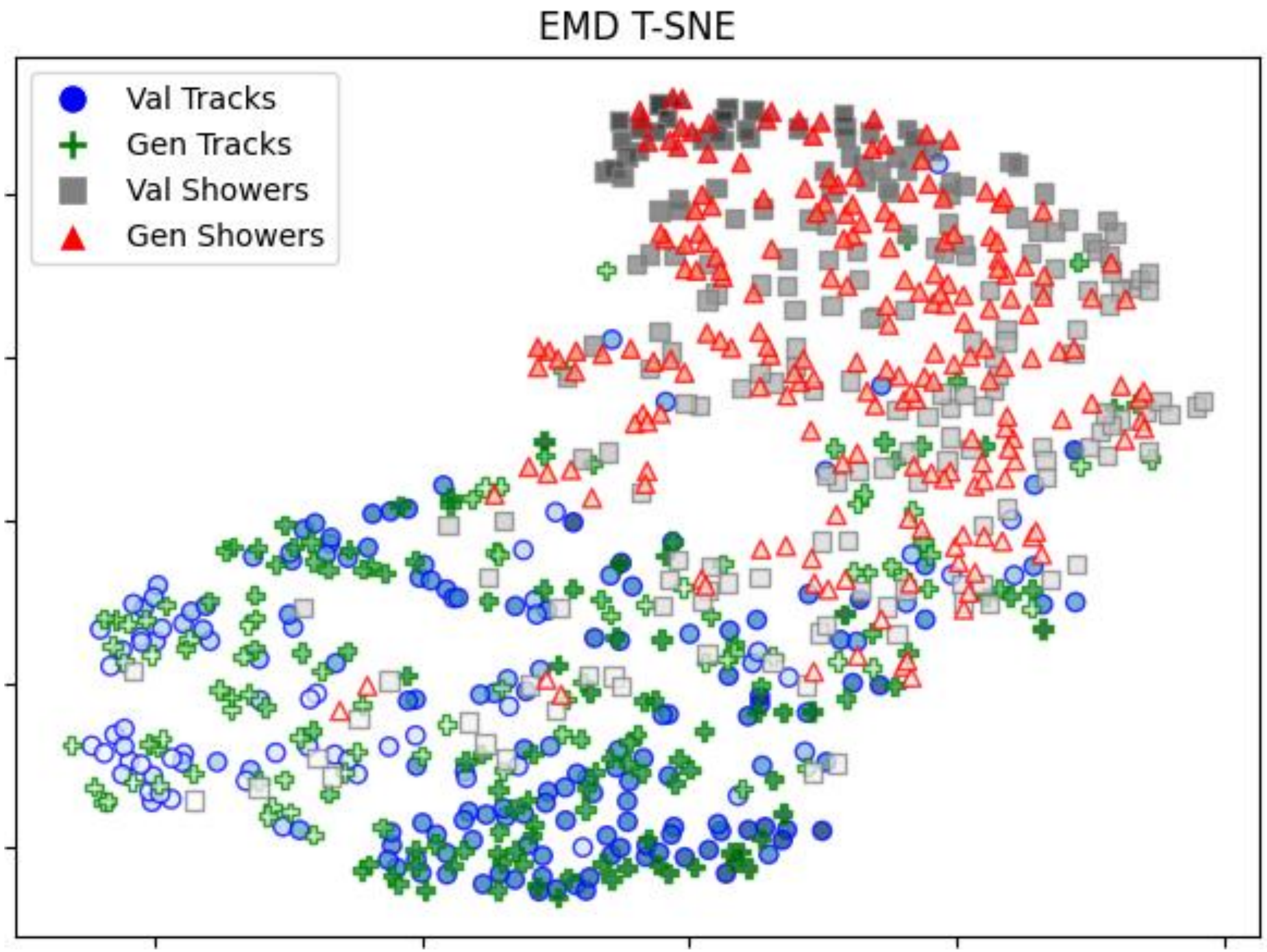


- red distribution: "dirt"
- blue distribution: "holes"



T-SNE EMD

- Separation of track and shower events
- Ongoing exploration of this data representation



Key Takeaways

1. LArTPC data differs from natural images
   - Globally sparse, but locally dense

2. Diffusion is a versatile method of data generation
   - Can handle our LArTPC data

3. Development of some quality metrics for LArTPC images

4. Earth Mover’s Distance is a useful metric for particle event data

Score-based Diffusion Models for Generating Liquid Argon Time Projection Chamber Images

By Zeviel Imani, Shuchin Aeron, & Taritree Wongjirad

PhysRevD.109.072011

Questions?

Tufts UNIVERSITY

The NSF Institute for Artificial Intelligence and Fundamental Interactions

NSF

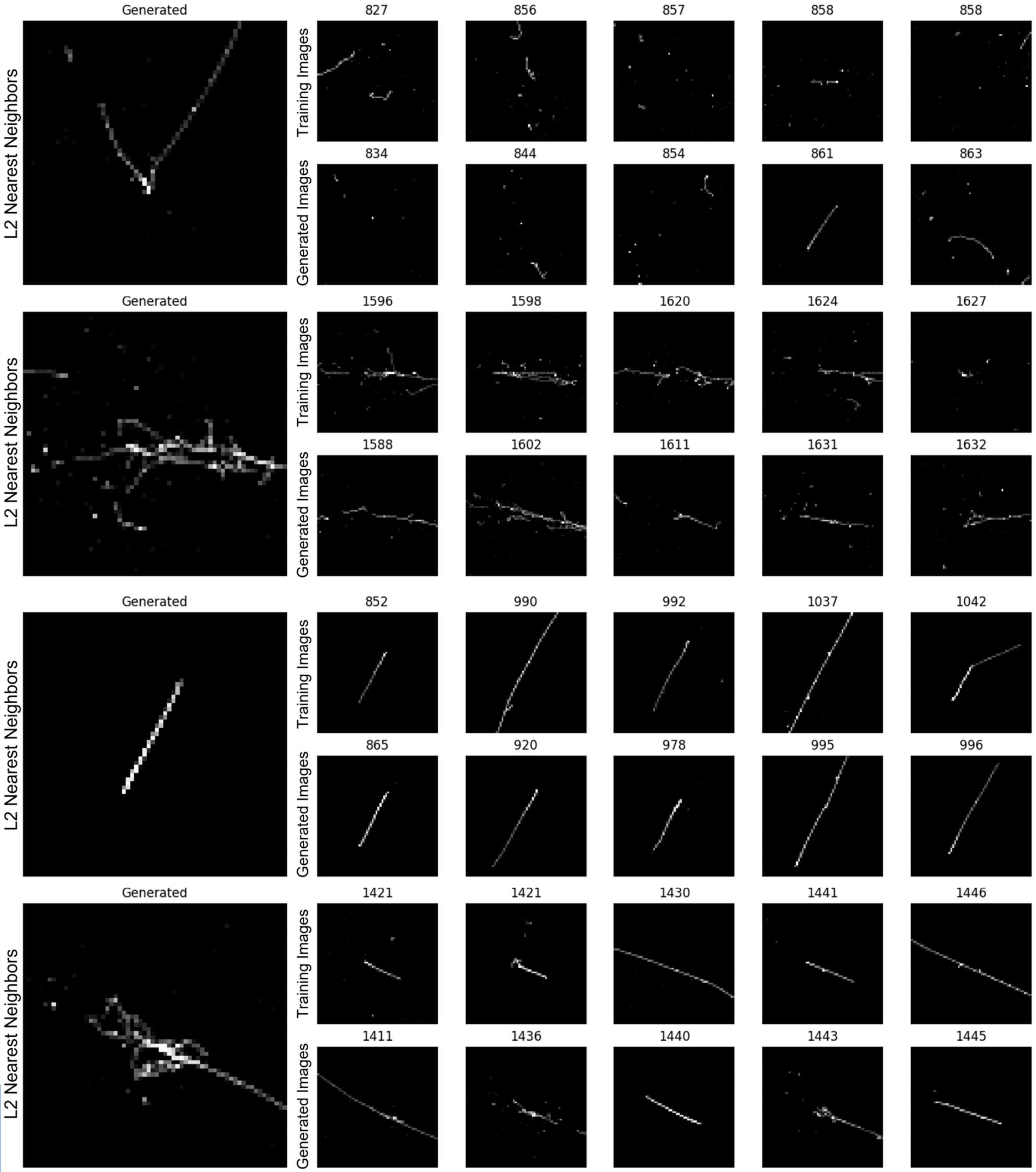
DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA

Backup Slides

(and skipped sections)

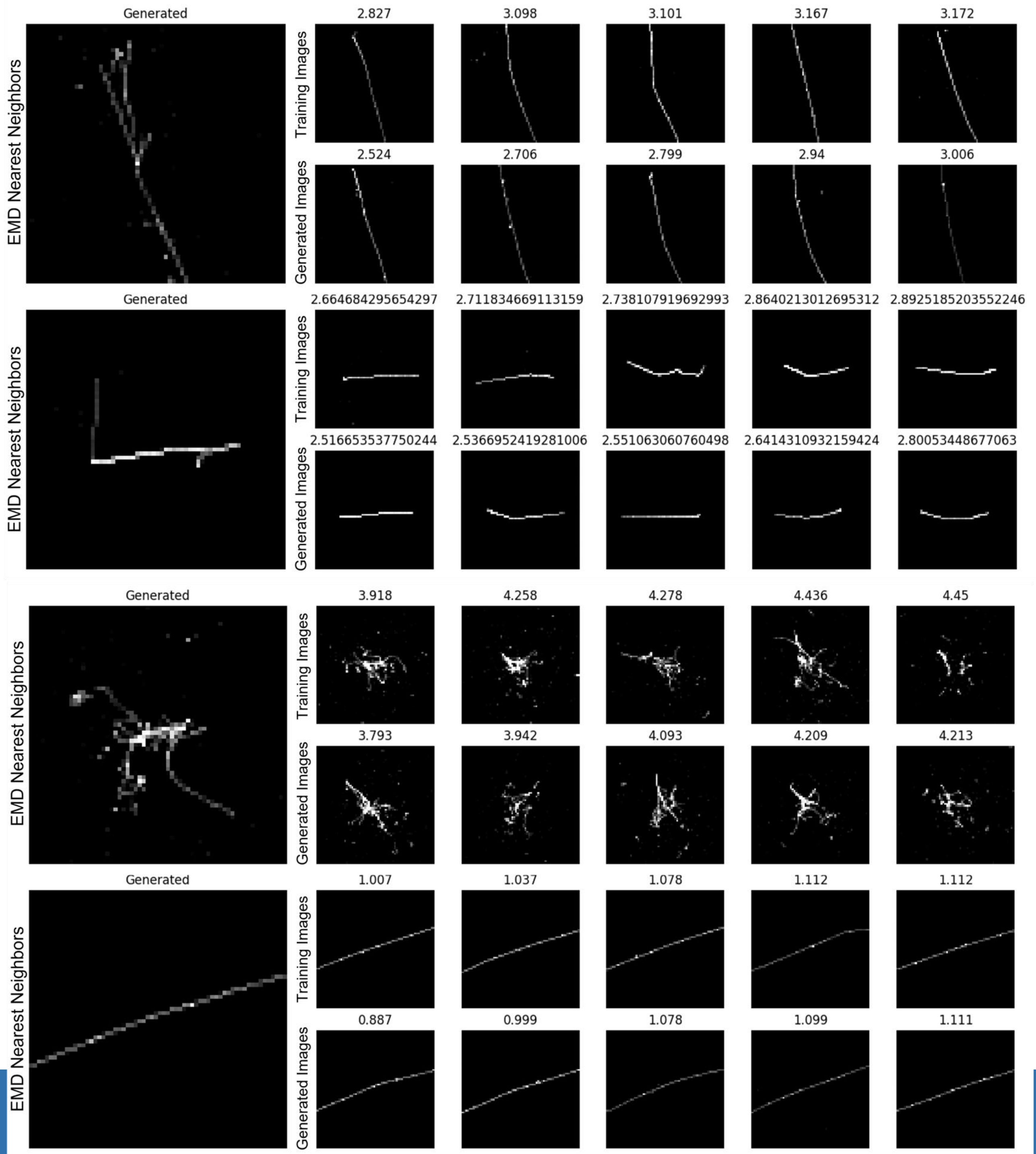
Mode Collapse

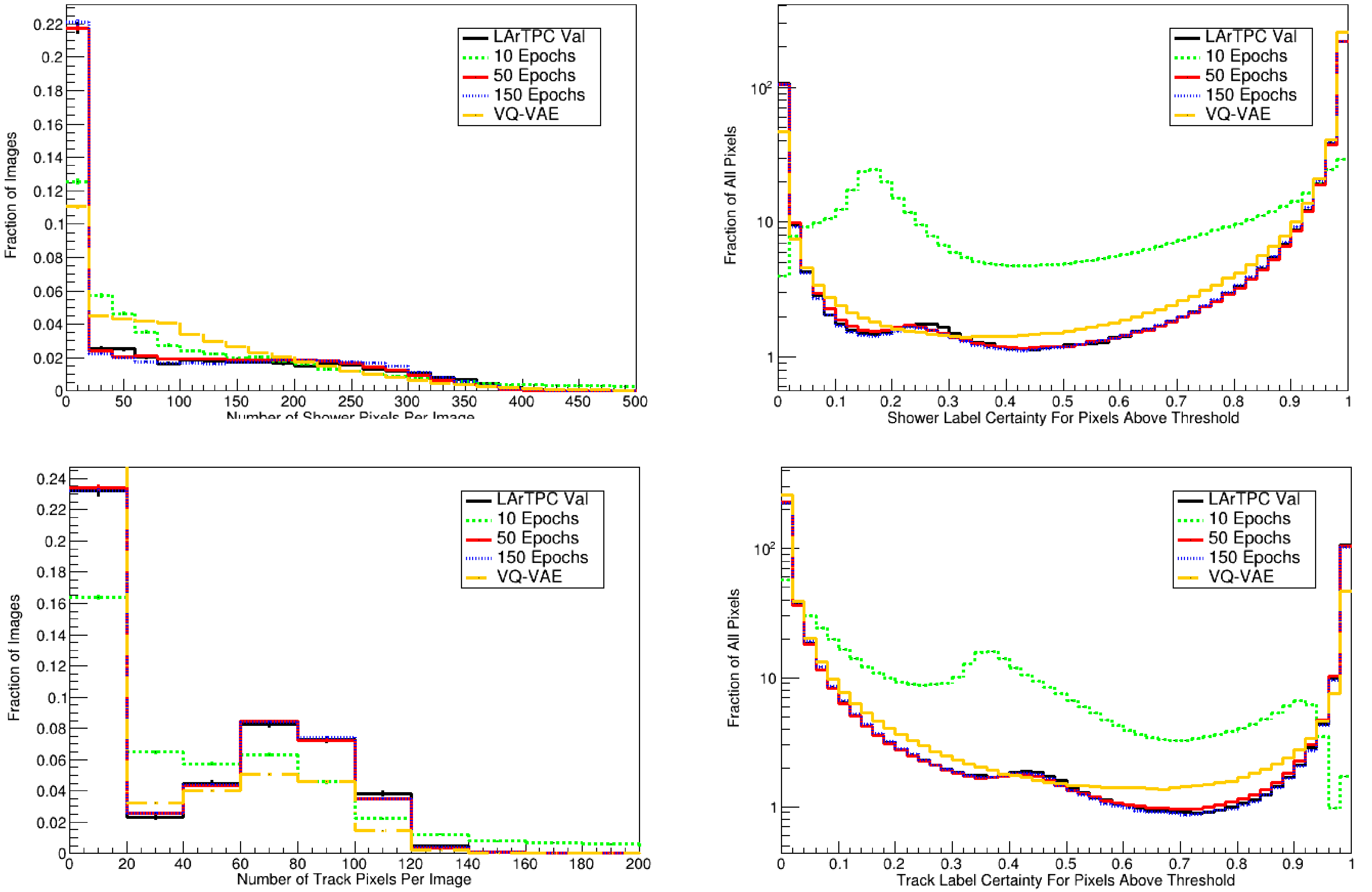
- Nearest neighbors using L2 Euclidian Norm distance



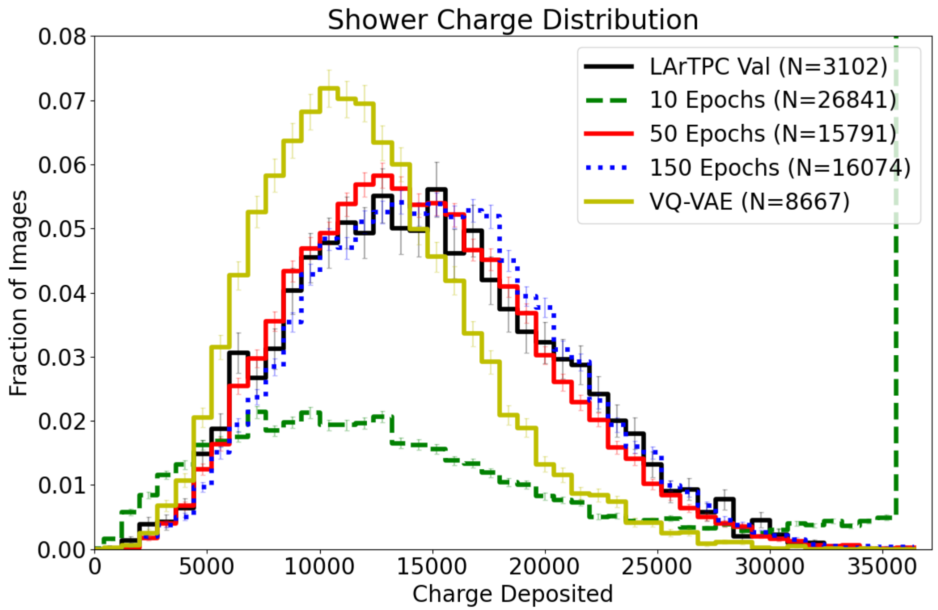
Mode Collapse

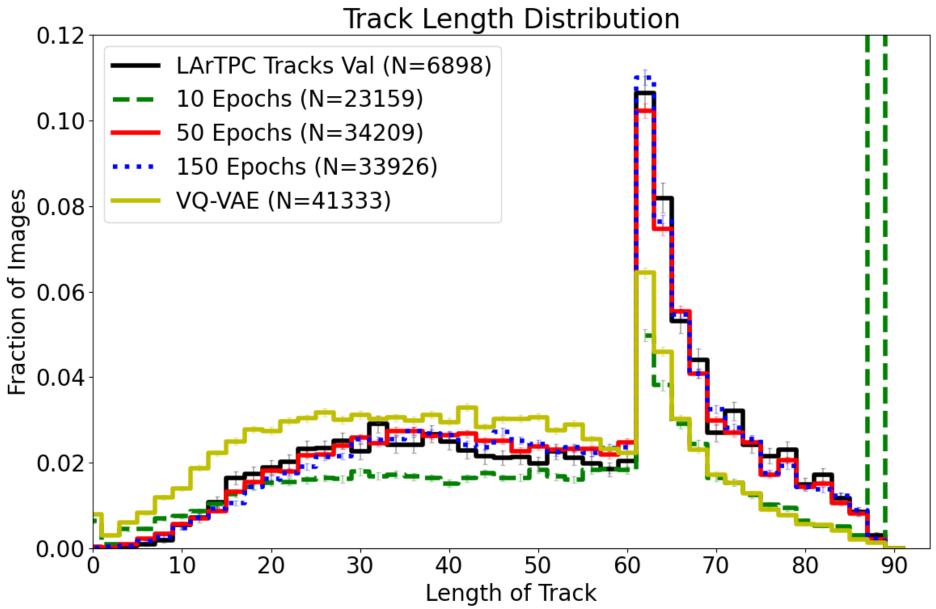
- Nearest neighbors using Earth Mover's Distance (EMD)

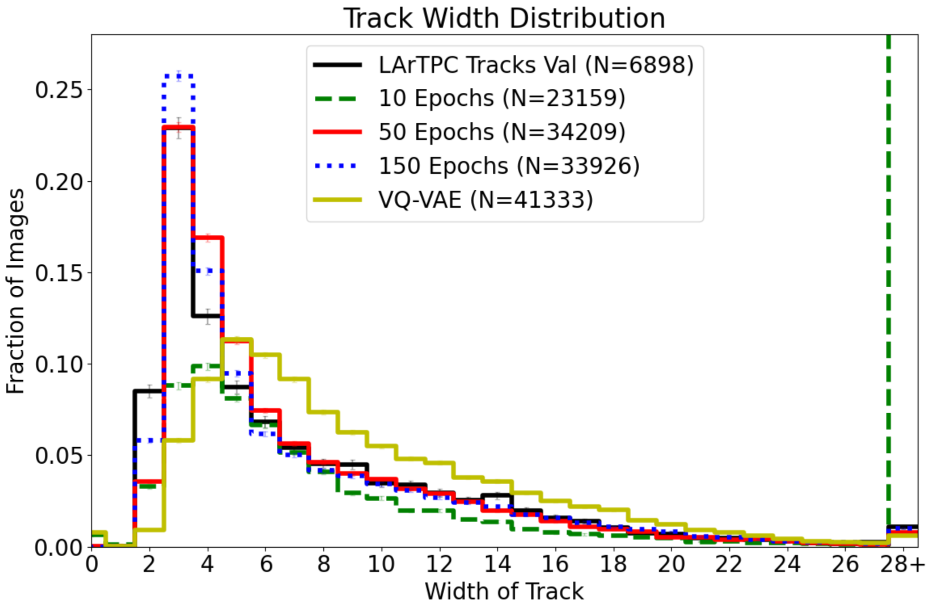




Fraction of Images
Number of Shower Pixels Per Image
LArTPC Val
10 Epochs
50 Epochs
150 Epochs
VQ-VAE
Fraction of All Pixels
Shower Label Certainty For Pixels Above Threshold
LArTPC Val
10 Epochs
50 Epochs
150 Epochs
VQ-VAE
Fraction of Images
Number of Track Pixels Per Image
LArTPC Val
10 Epochs
50 Epochs
150 Epochs
VQ-VAE
Fraction of All Pixels
Track Label Certainty For Pixels Above Threshold
LArTPC Val
10 Epochs
50 Epochs
150 Epochs
VQ-VAE

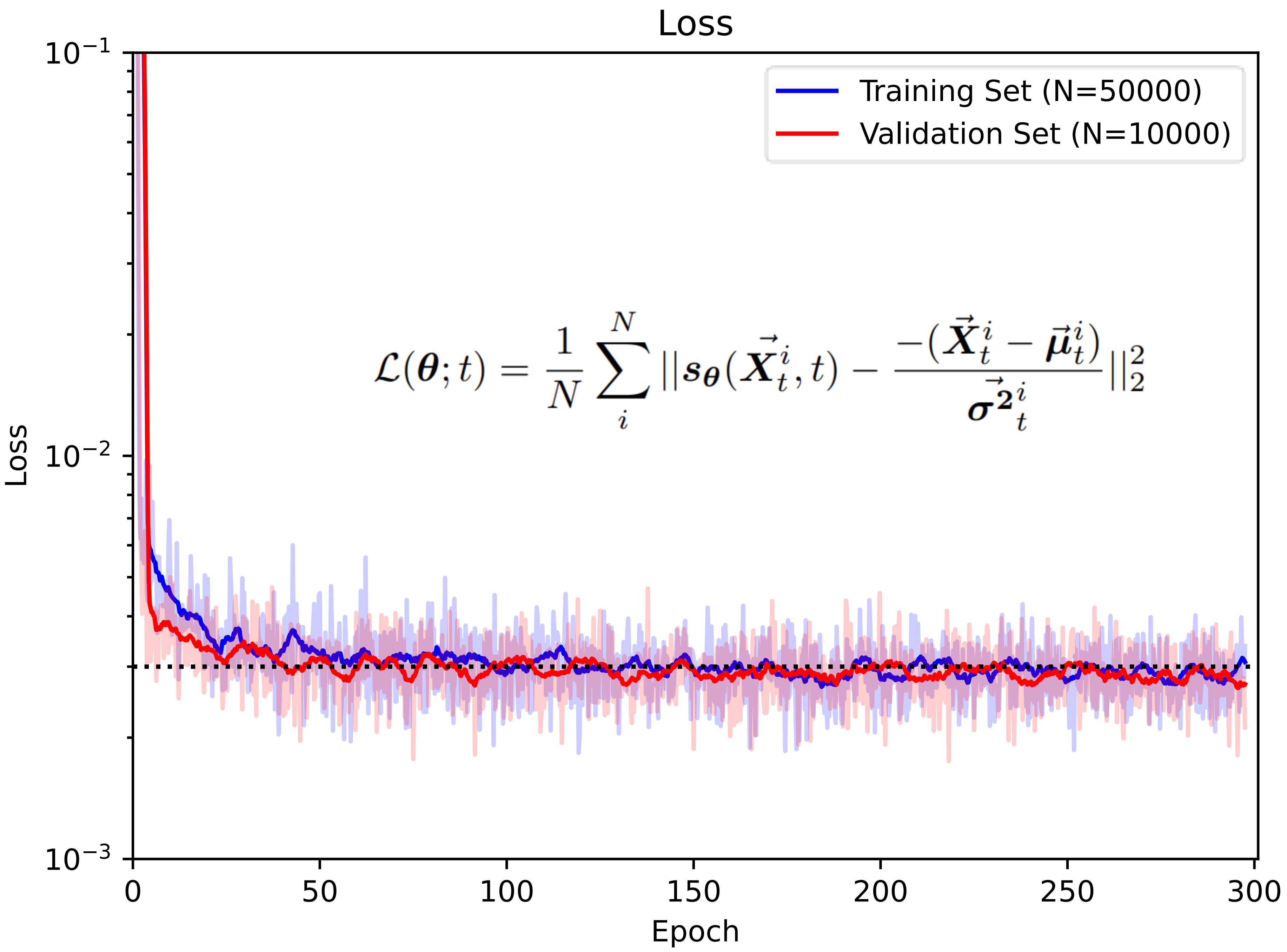

Shower Charge Distribution
LArTPC Val (N=3102)
10 Epochs (N=26841)
50 Epochs (N=15791)
150 Epochs (N=16074)
VQ-VAE (N=8667)
Fraction of Images
Charge Deposited
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0
5000
10000
15000
20000
25000
30000
35000


Track Length Distribution
LArTPC Tracks Val (N=6898)
10 Epochs (N=23159)
50 Epochs (N=34209)
150 Epochs (N=33926)
VQ-VAE (N=41333)
Fraction of Images
Length of Track
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
10
20
30
40
50
60
70
80
90

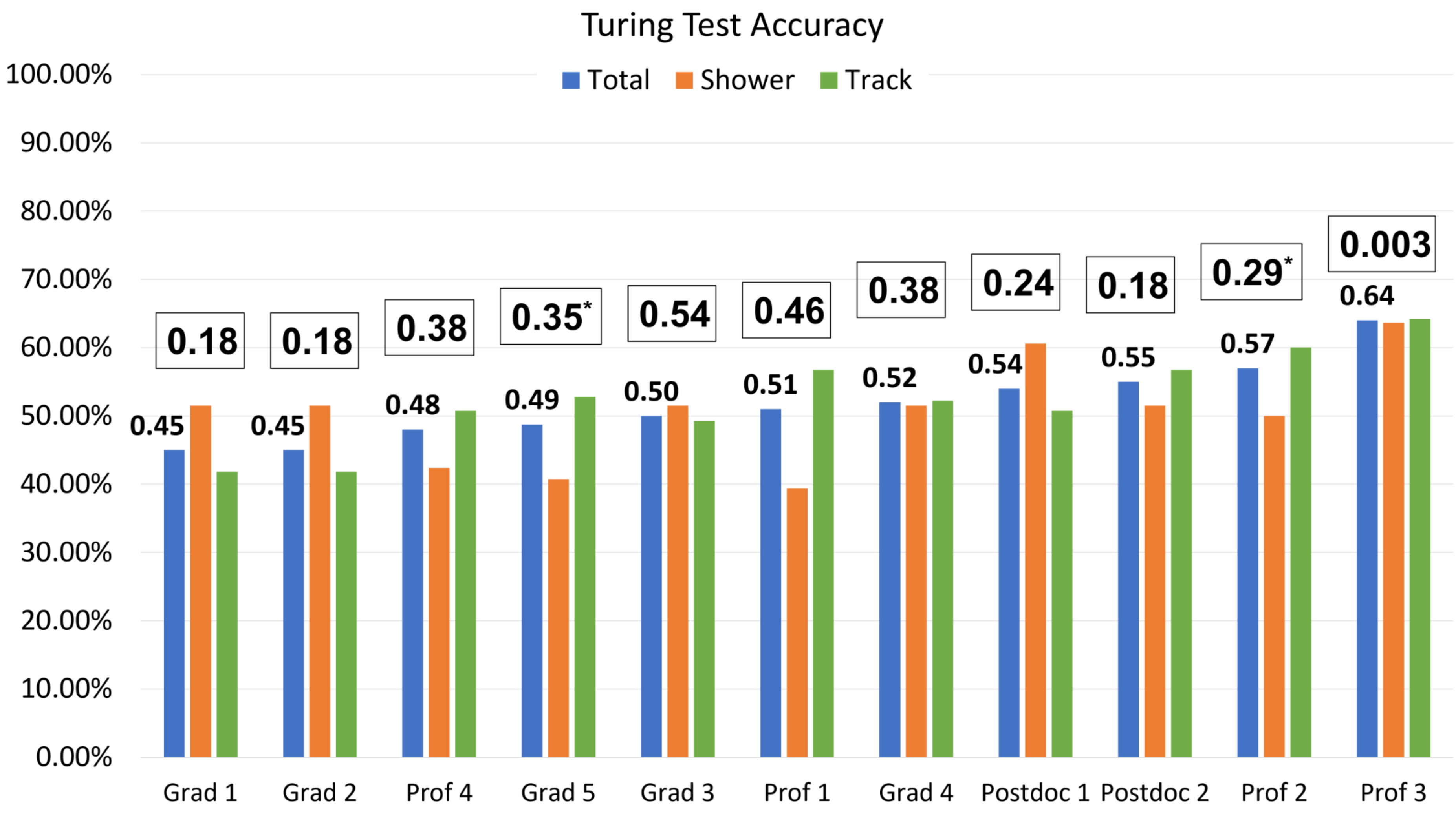

Track Width Distribution
LArTPC Tracks Val (N=6898)
10 Epochs (N=23159)
50 Epochs (N=34209)
150 Epochs (N=33926)
VQ-VAE (N=41333)
Fraction of Images
Width of Track
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28+

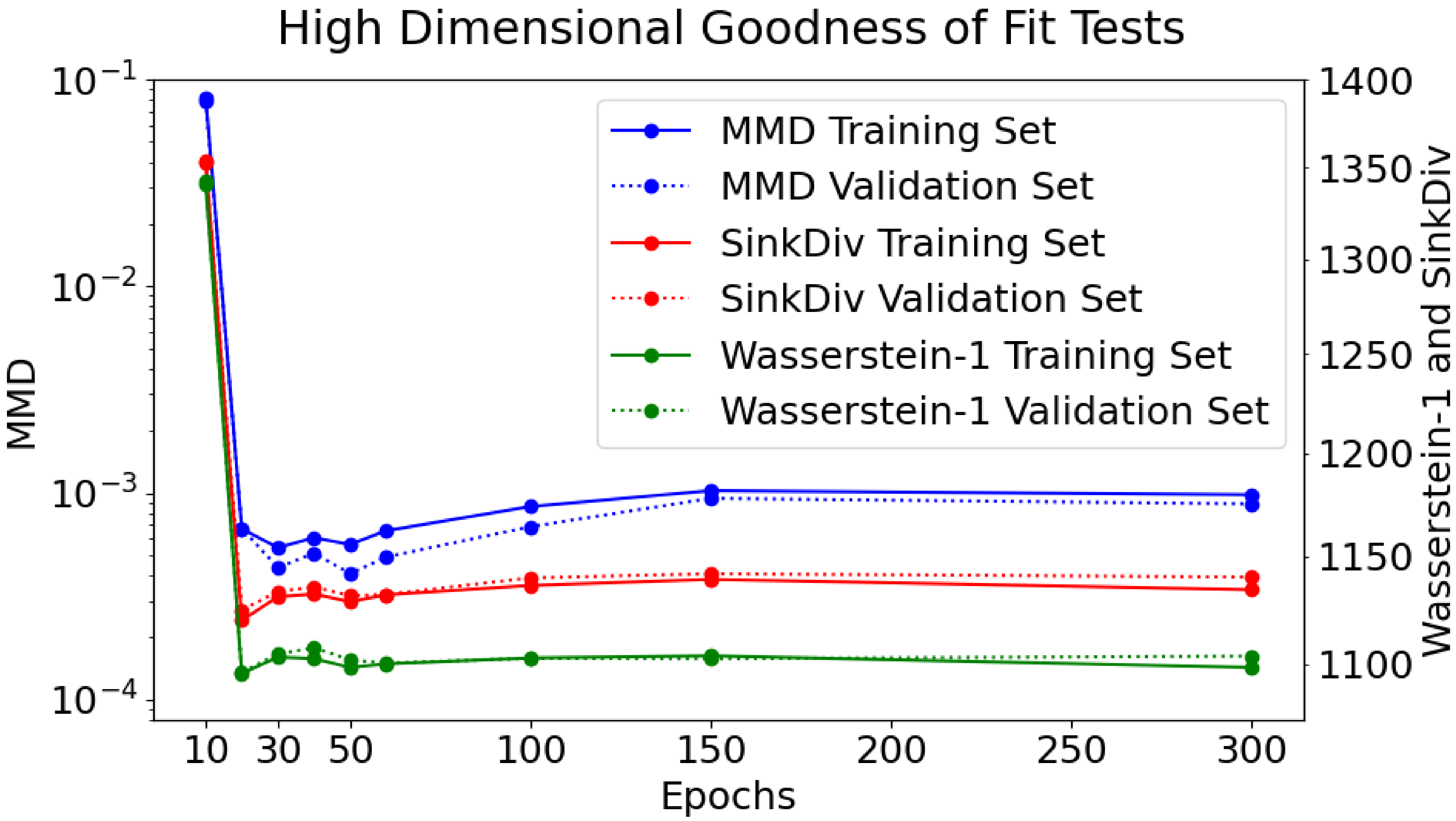
Physics Metrics: Chi-Squared

χ^2 Test	Track Length	Track Width	Shower Charge
10 Epochs	206	825	6458
50 Epochs	**126**	418	**228**
150 Epochs	130	**175**	382


Loss
Training Set (N=50000)
Validation Set (N=10000)
$\mathcal{L}(\boldsymbol{\theta};t) = \frac{1}{N}\sum_{i}^{N}||\boldsymbol{s_\theta}(\vec{\boldsymbol{X}}_t^i,t) - \frac{-(\vec{\boldsymbol{X}}_t^i - \vec{\boldsymbol{\mu}}_t^i)}{\vec{\boldsymbol{\sigma^2}}_t^i}||_2^2$
Loss
10^{-1}
10^{-2}
10^{-3}
0
50
100
150
200
250
300
Epoch

Turing Test


Turing Test Accuracy
Total
Shower
Track
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
0.18
0.18
0.38
0.35*
0.54
0.46
0.38
0.24
0.18
0.29*
0.003
0.45
0.45
0.48
0.49
0.50
0.51
0.52
0.54
0.55
0.57
0.64
Grad 1
Grad 2
Prof 4
Grad 5
Grad 3
Prof 1
Grad 4
Postdoc 1
Postdoc 2
Prof 2
Prof 3

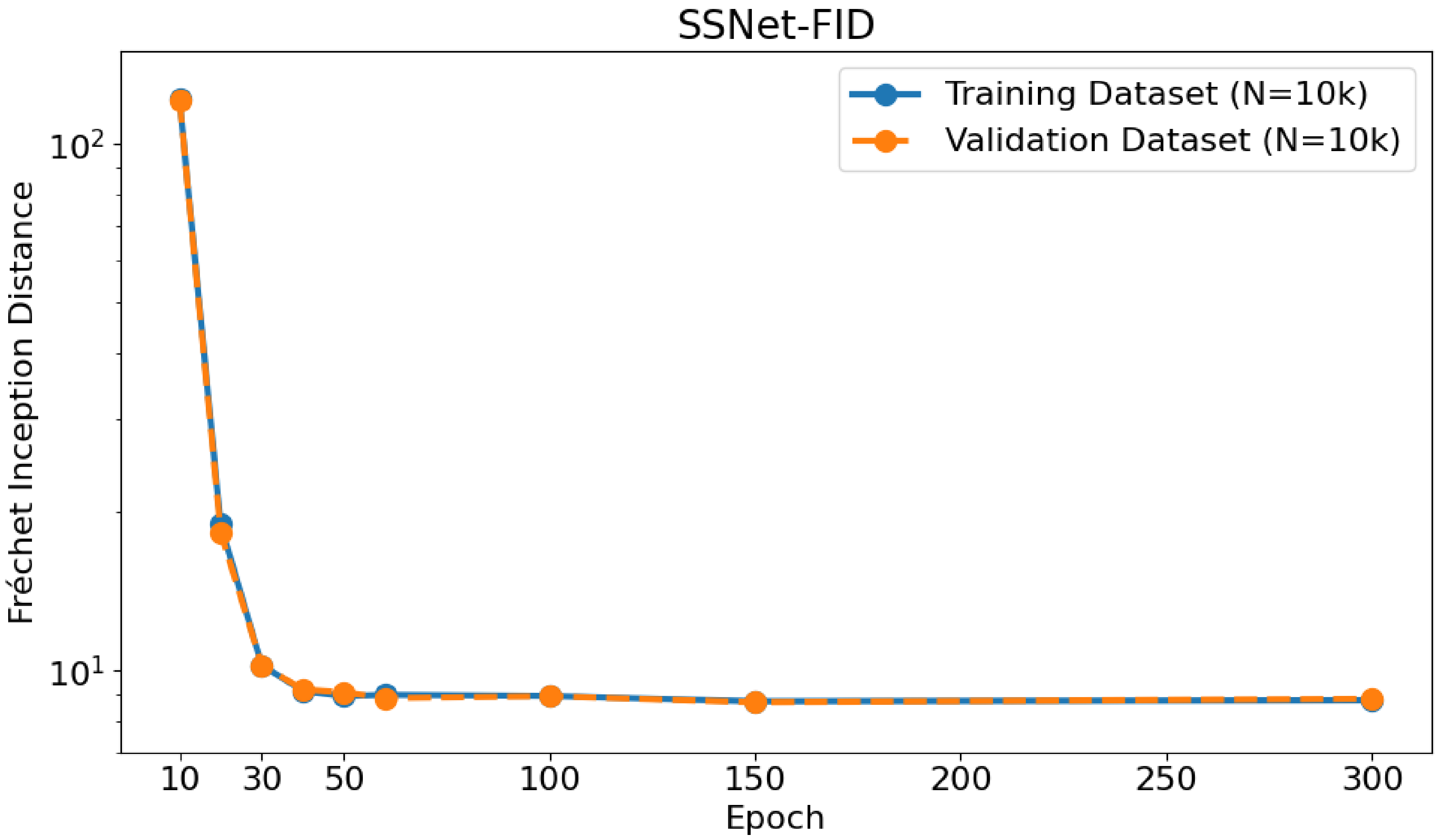

High Dimensional Goodness of Fit Tests
10^{-1}
10^{-2}
10^{-3}
10^{-4}
MMD
1400
1350
1300
1250
1200
1150
1100
Wasserstein-1 and SinkDiv
MMD Training Set
MMD Validation Set
SinkDiv Training Set
SinkDiv Validation Set
Wasserstein-1 Training Set
Wasserstein-1 Validation Set
10
30
50
100
150
200
250
300
Epochs

Fréchet Inception Distance (FID)

- Process:
  1. Get **layer activations** from classifier
     - Typically use Google's Inception v3 deepest activation layer (pool3)
       - 2048-dimensional activation vector
  2. Fit activations to multidimensional Gaussian distribution
  3. Find Wasserstein-2 distance between the Gaussians

- We can use activations from SSNet instead

SSNet-FID


SSNet-FID
Training Dataset (N=10k)
Validation Dataset (N=10k)
Fréchet Inception Distance
10^2
10^1
10
30
50
100
150
200
250
300
Epoch

Conditional 1: Statistical Reframe

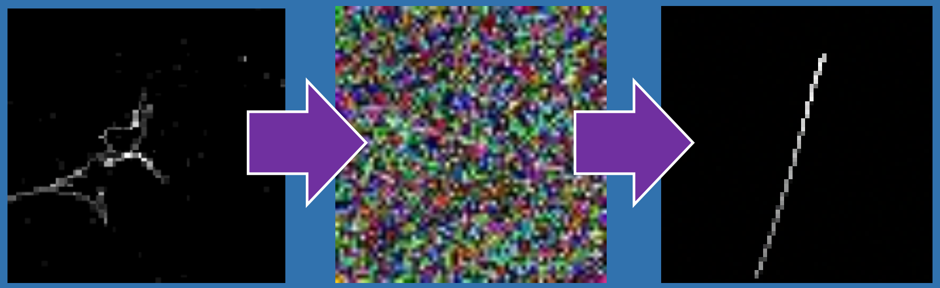
- Given random variables **x** (LArTPC image) and **y** (energy) we want to sample from p(**x** | **y**)
- Approach 1) Extend score: $s_\theta(\mathbf{x}, t) \rightarrow s_\theta(\mathbf{x}, t, \mathbf{y})$
- Or…

Conditional 2: Inverse Problem

- We know how to get **y** (energy) from **x** (LArTPC image)
- Bayes' Rule: $p(\mathbf{x}|\mathbf{y}) = \dfrac{p(\mathbf{x})p(\mathbf{y}|\mathbf{x})}{\int p(\mathbf{x})p(\mathbf{y}|\mathbf{x})\mathrm{d}\mathbf{x}}$
- Take gradient: $\nabla_{\mathbf{x}} \log p(\mathbf{x} \mid \mathbf{y}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x})$

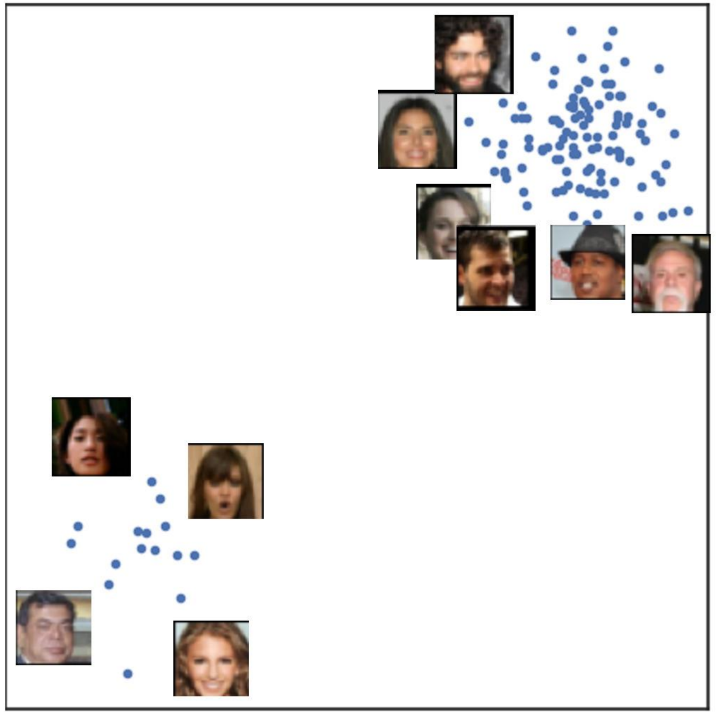
score **classifier**

Score-based Diffusion Model



Y. Song, S. Ermon,
arXiv:1907.05600

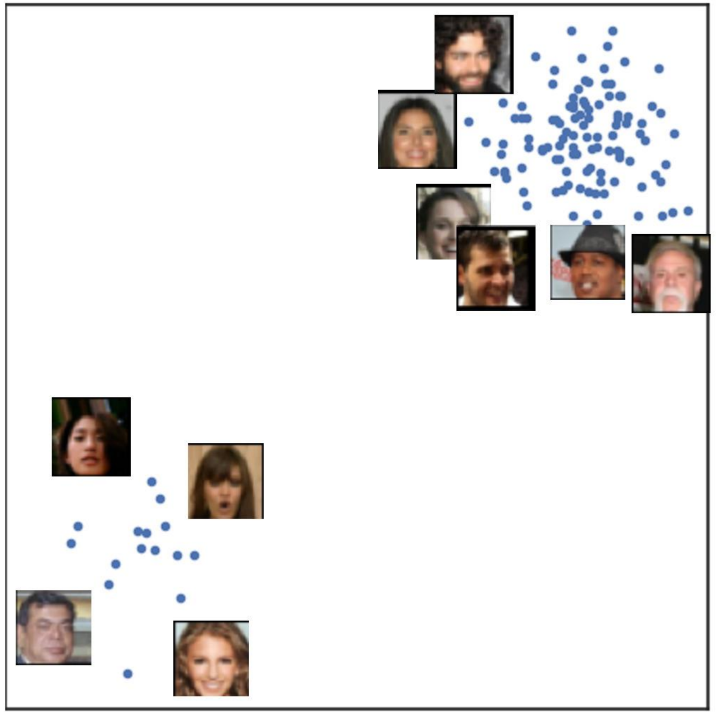
How to Generate Images



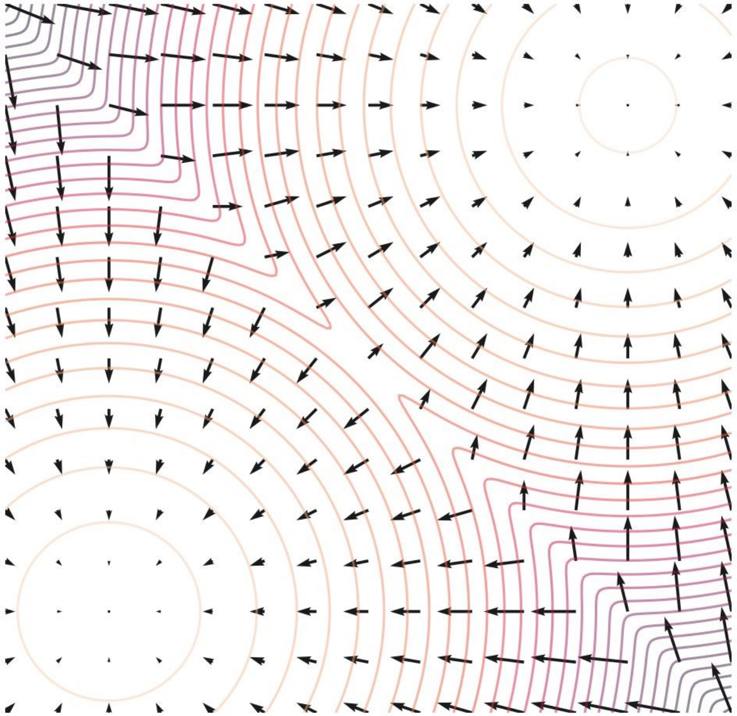
Data samples

$$\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \overset{\text{i.i.d.}}{\sim} p(\mathbf{x})$$

How to Generate Images



score matching



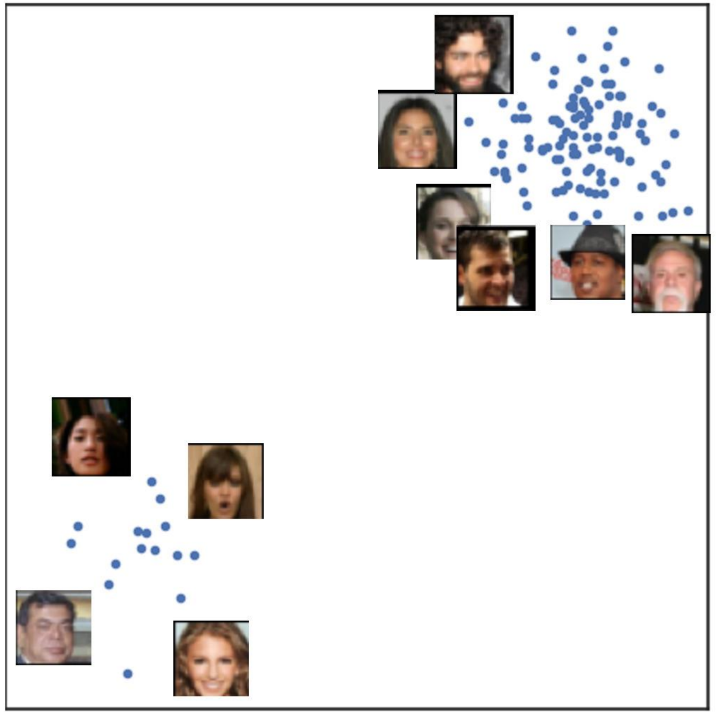
Data samples

$$\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \overset{\text{i.i.d.}}{\sim} p(\mathbf{x})$$

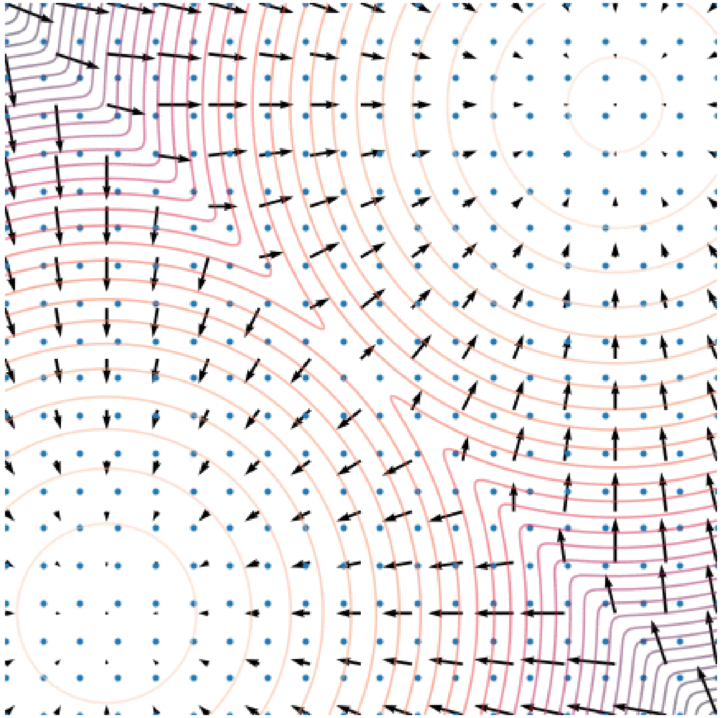
Scores

$$\mathbf{s}_\theta(\mathbf{x}) \approx \nabla_\mathbf{x} \log p(\mathbf{x})$$

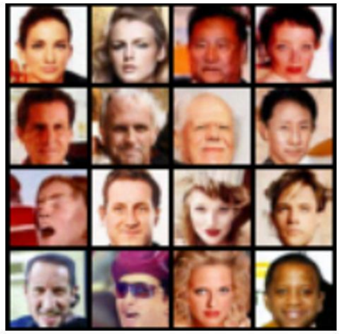
How to Generate Images



score matching



Langevin dynamics



Data samples

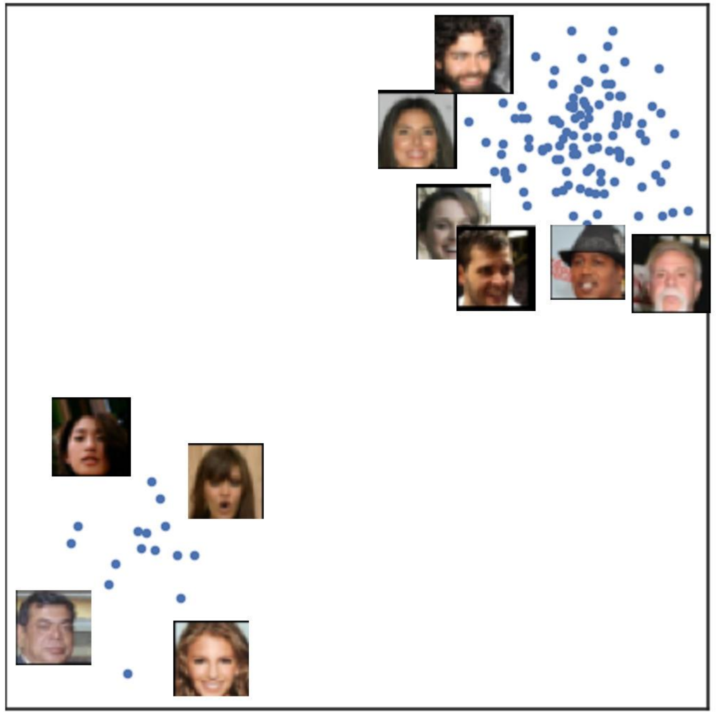
$$\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \overset{\text{i.i.d.}}{\sim} p(\mathbf{x})$$

Scores

$$\mathbf{s}_\theta(\mathbf{x}) \approx \nabla_\mathbf{x} \log p(\mathbf{x})$$

New samples

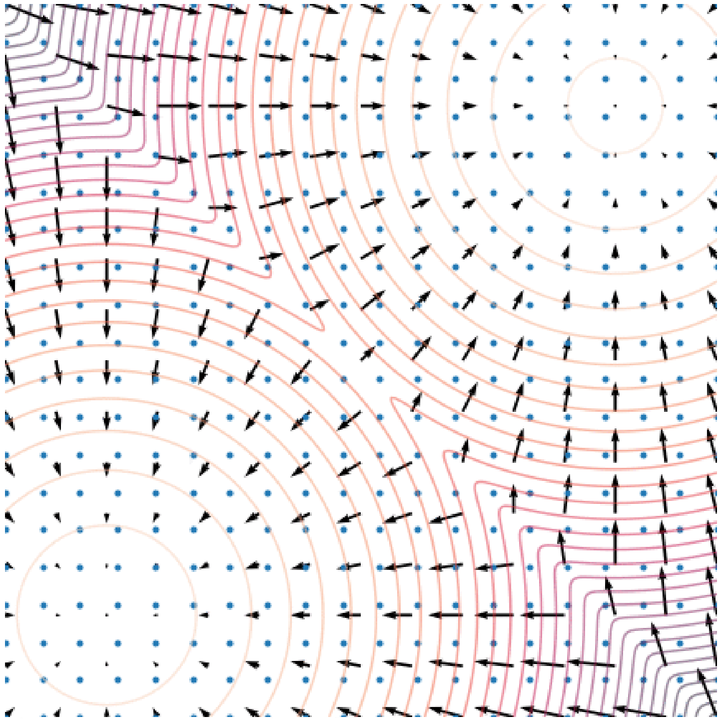
How to Generate Images



Data samples

$$\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \overset{\text{i.i.d.}}{\sim} p(\mathbf{x})$$

score matching



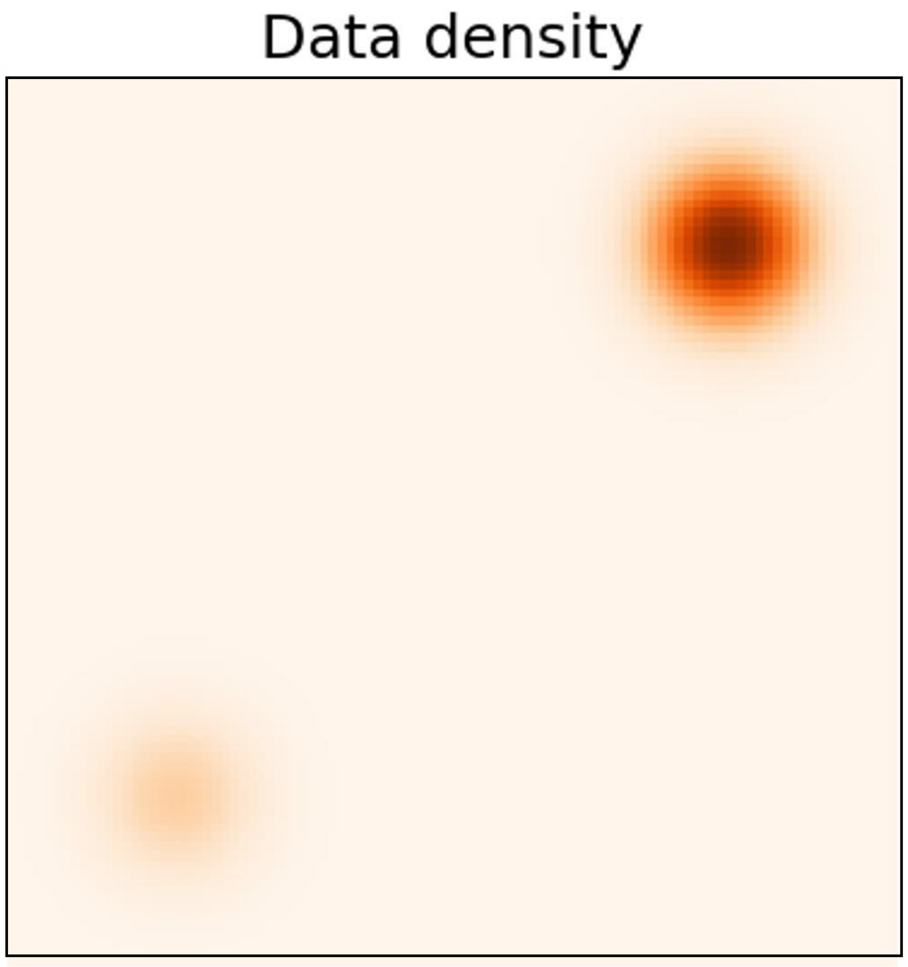
Scores

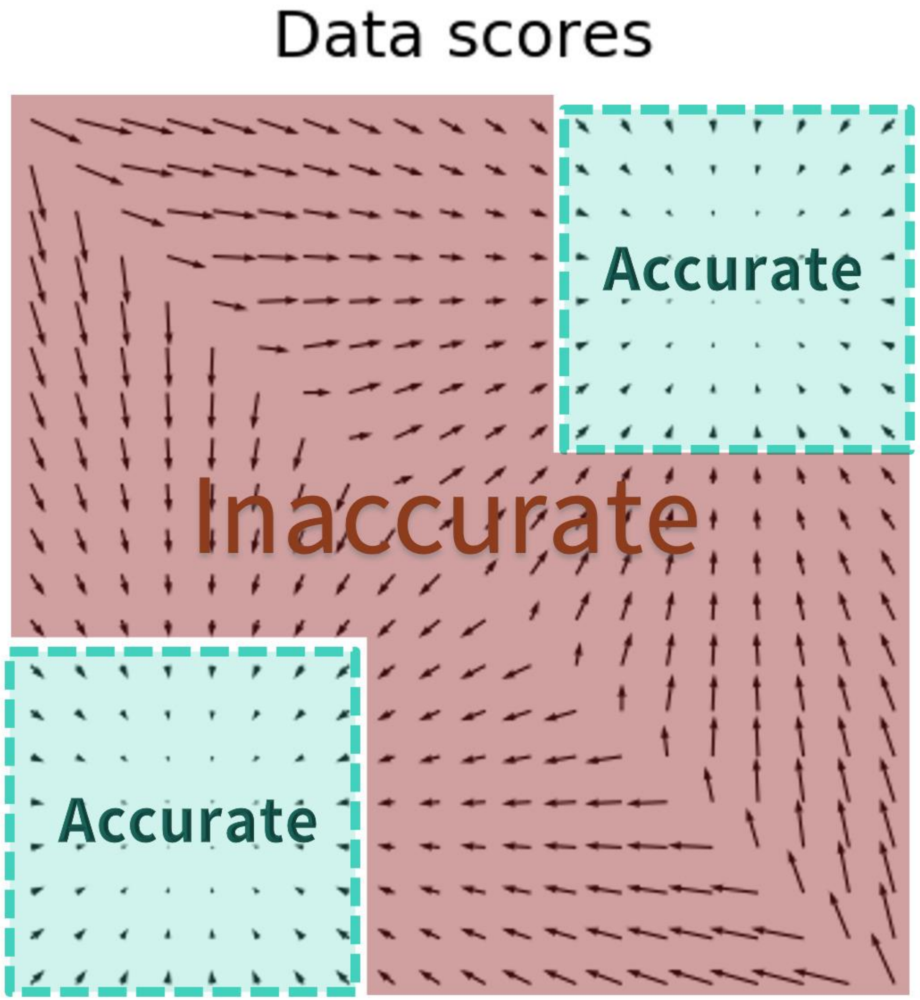
$$\mathbf{s}_\theta(\mathbf{x}) \approx \nabla_\mathbf{x} \log p(\mathbf{x})$$

Langevin dynamics

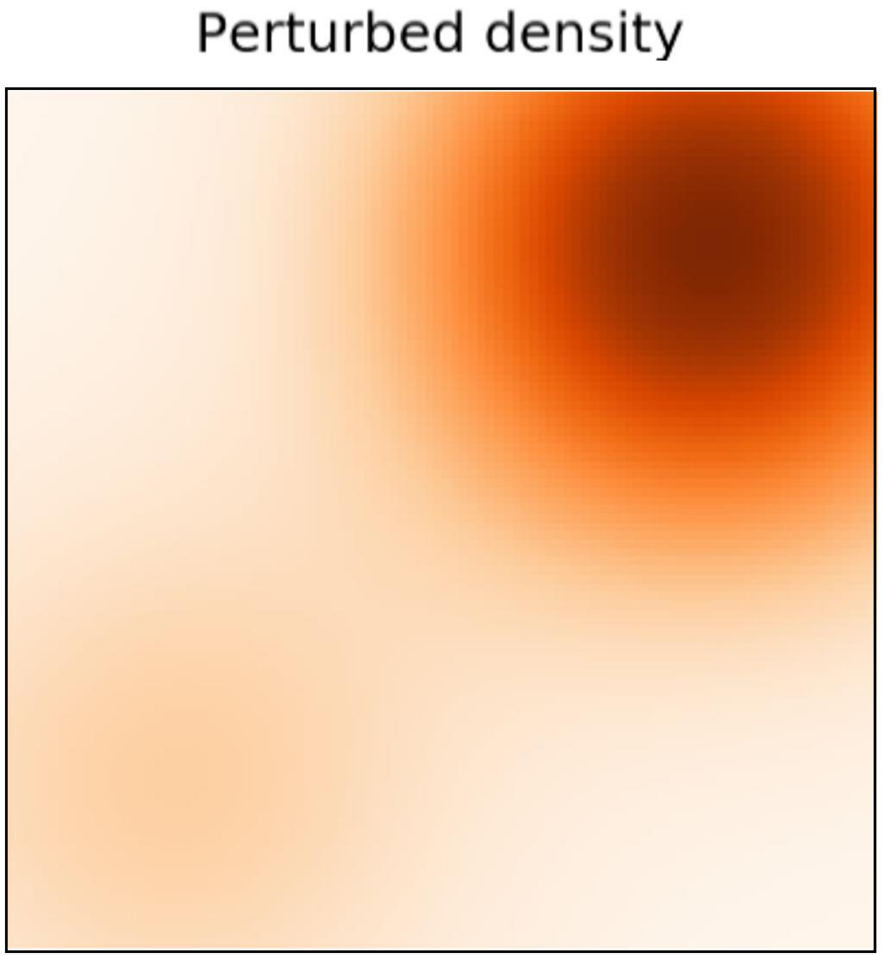
New samples

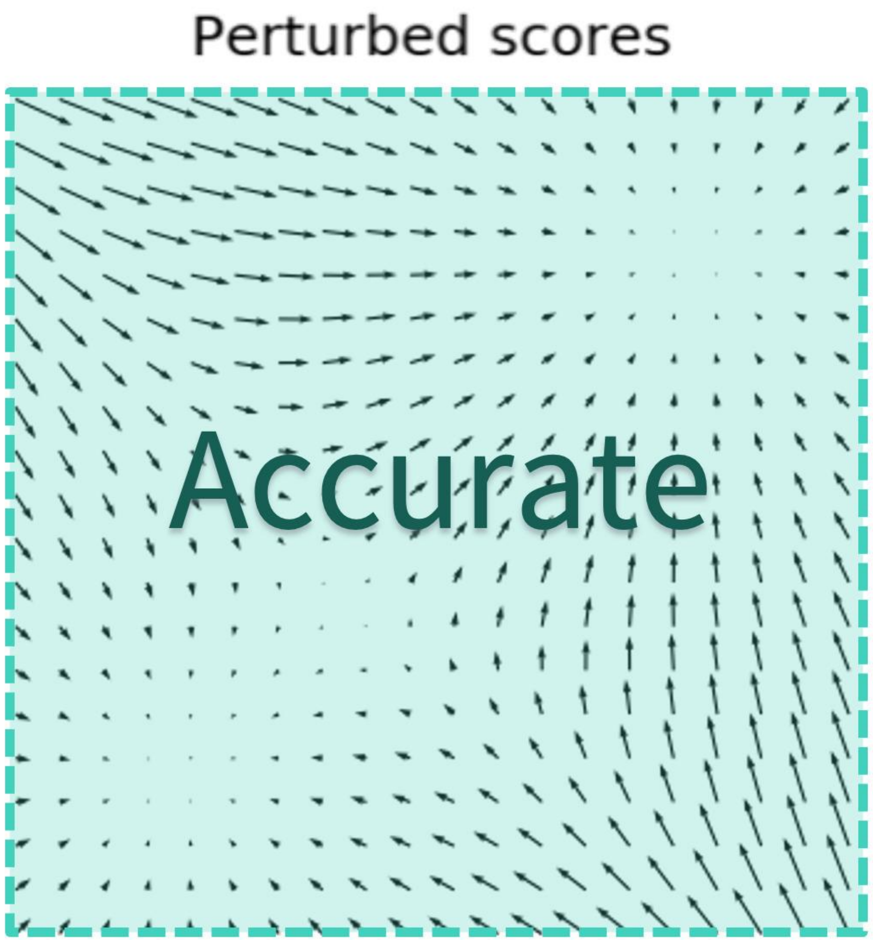
Manifold Hypothesis



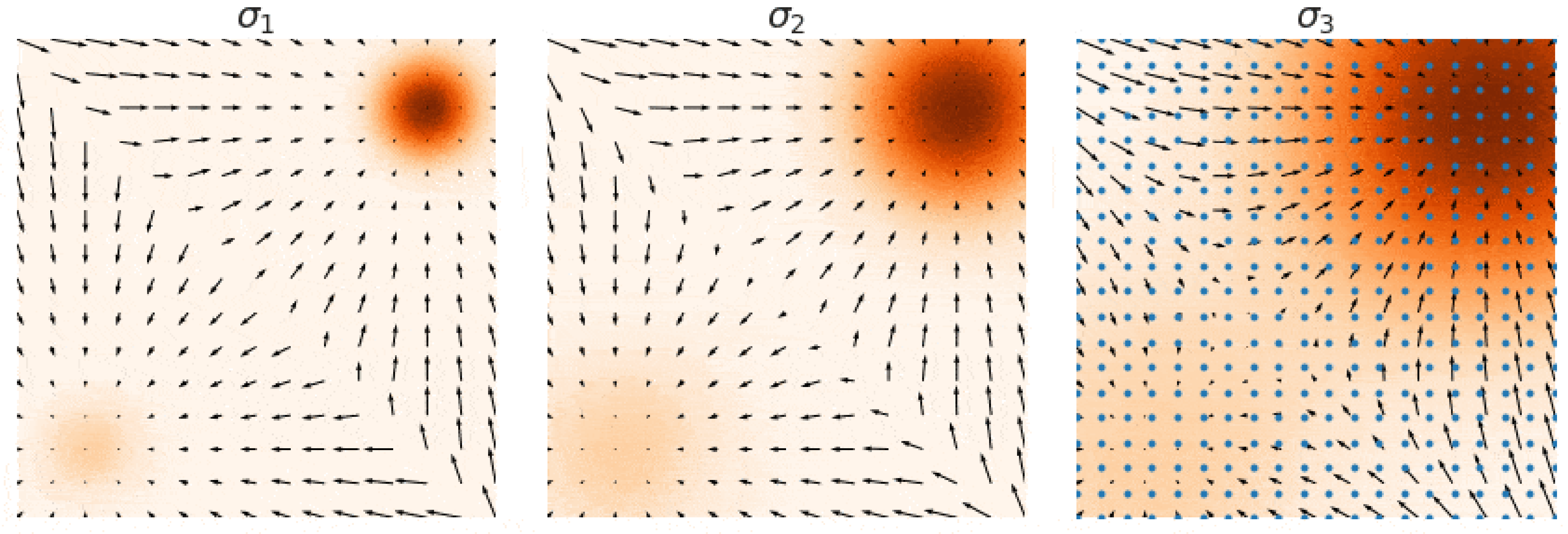


Add Diffusion



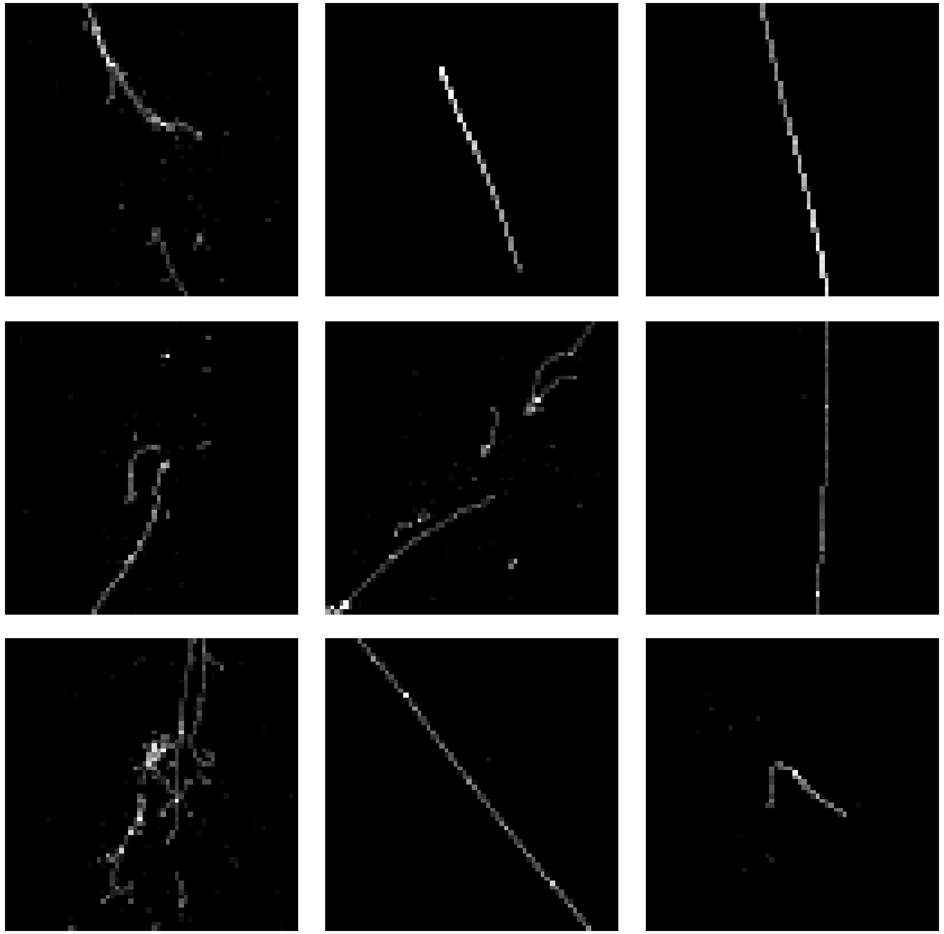


Annealed Langevin Sampling

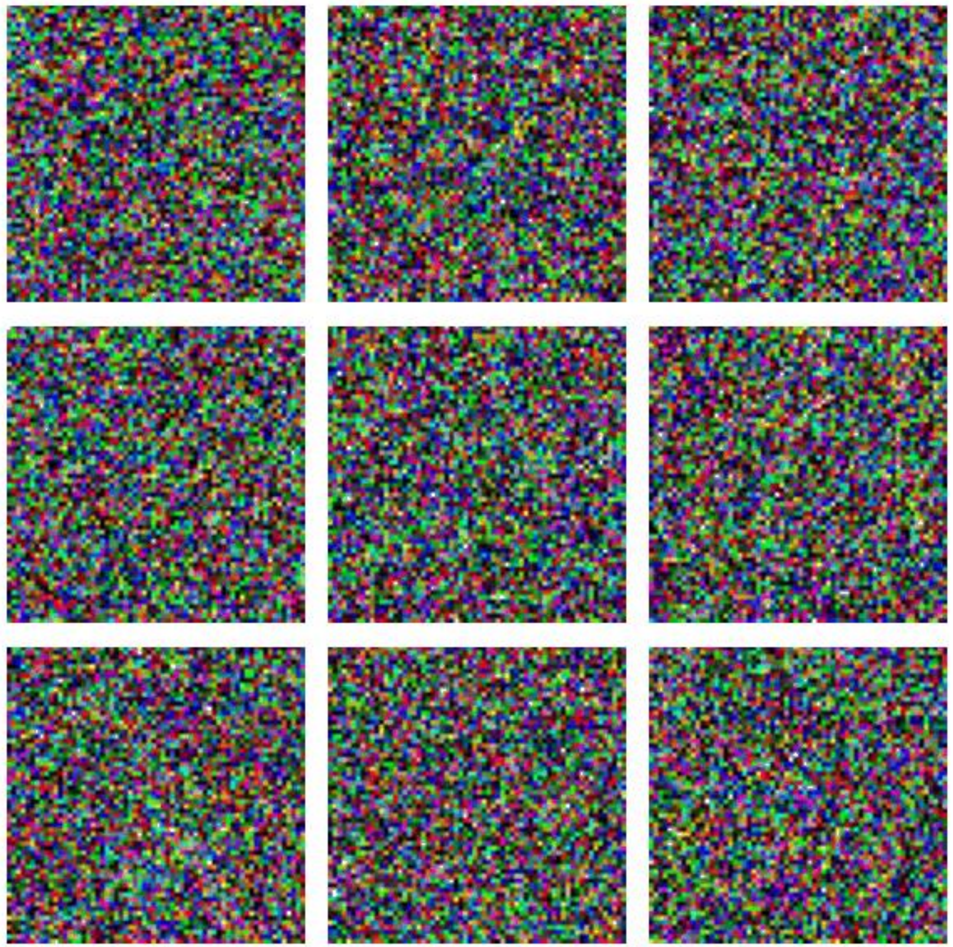


LArTPC Image Generation

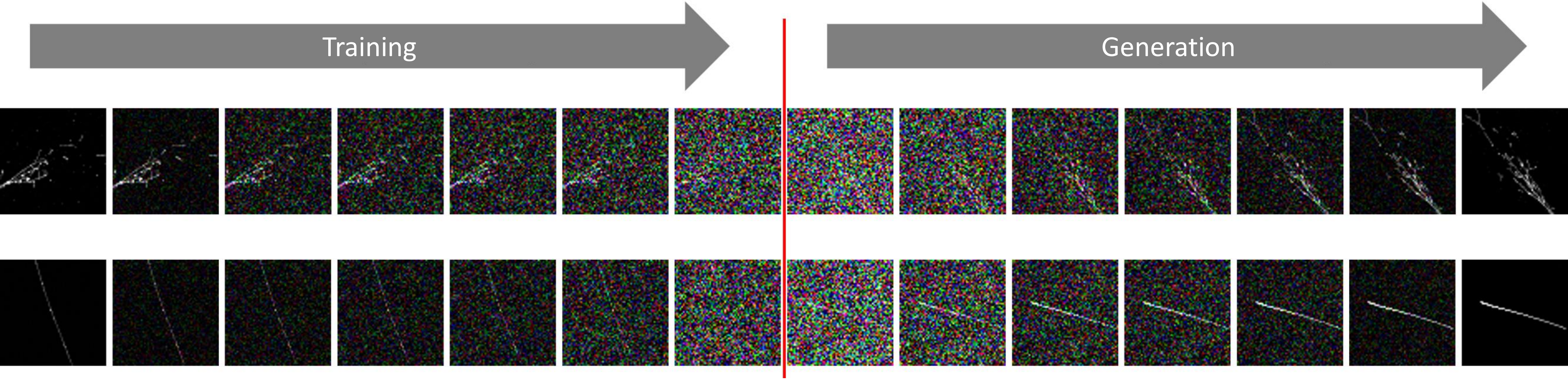
Training Images



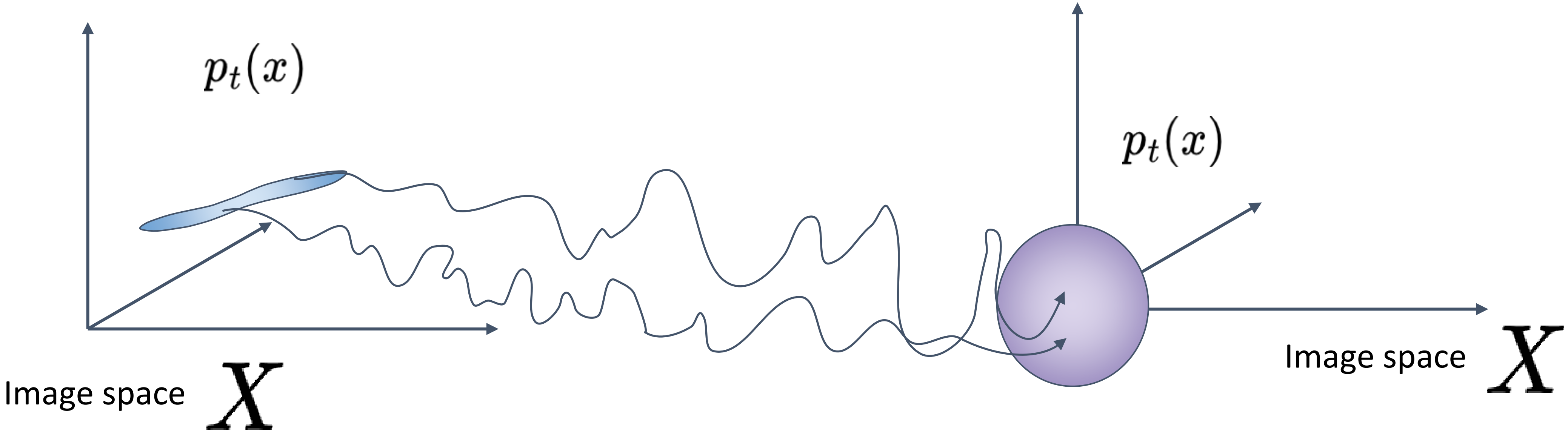
Generated Images



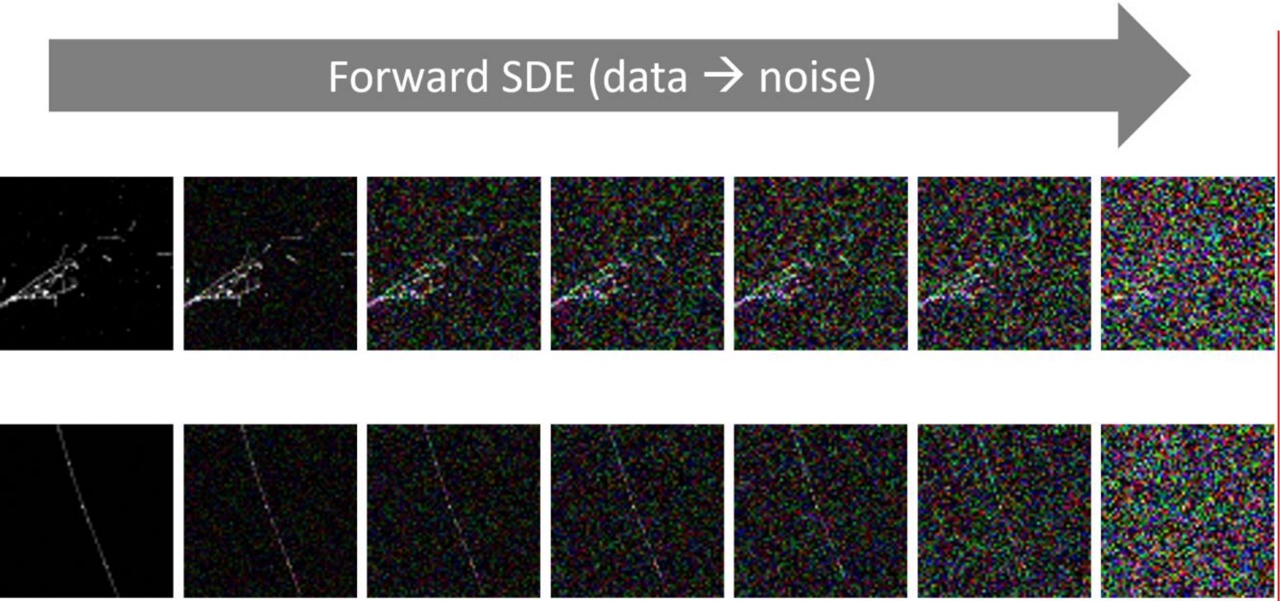
All Together Now



Where is the mapping?



Forward Stochastic Differential Equation (SDE)



$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$$

Drift $\mathbf{f}(\mathbf{x}, t)\mathrm{d}t$

Deterministic evolution

$\mathbf{f}(\mathbf{x}, t) = -\mathbf{x}\frac{1}{2}\beta_t$

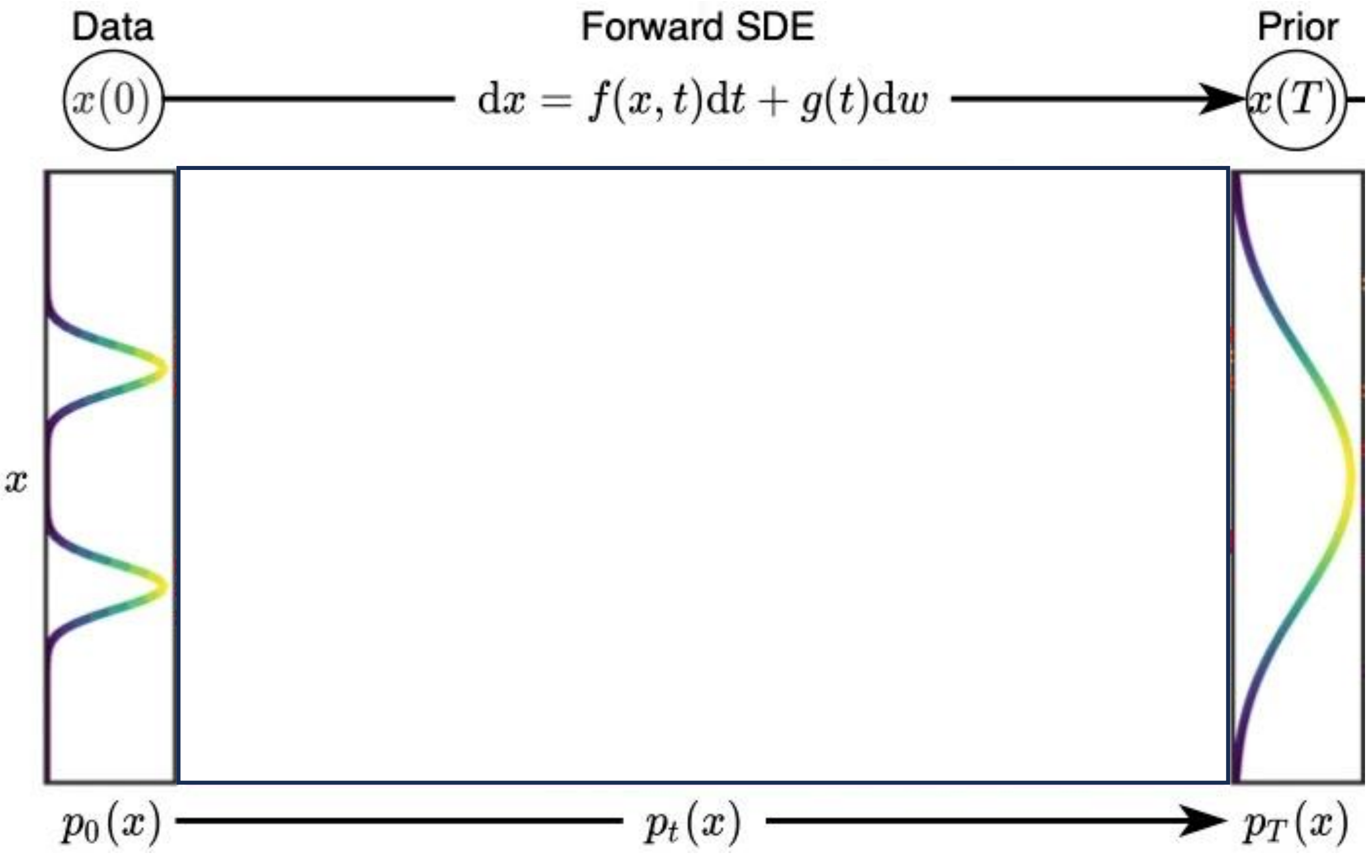
dt = time increment

Diffusion $g(t)\mathrm{d}\mathbf{w}$

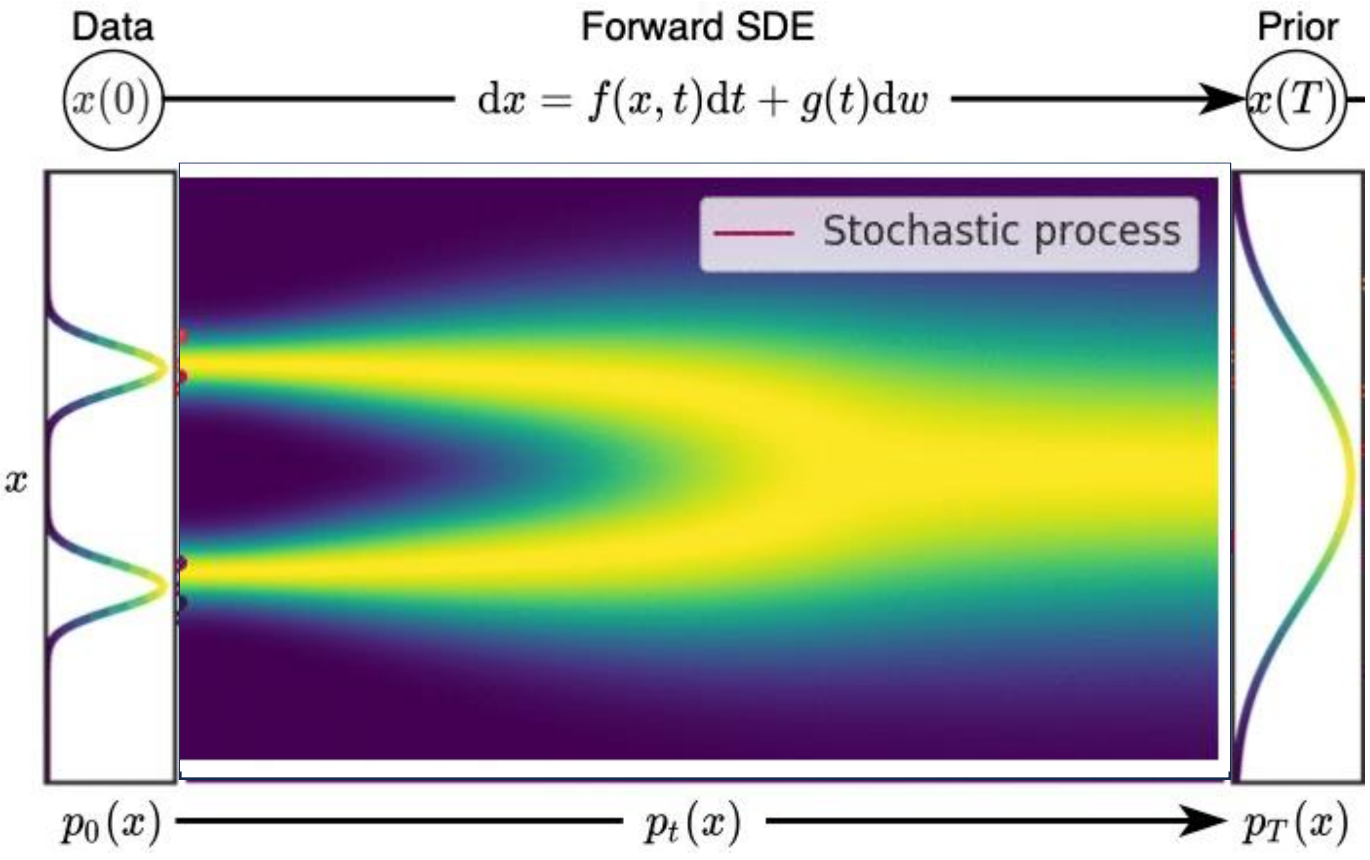
Scale factor $g(t) = \sqrt{\beta_t}$

d**w** = Brownian motion (Random walk)

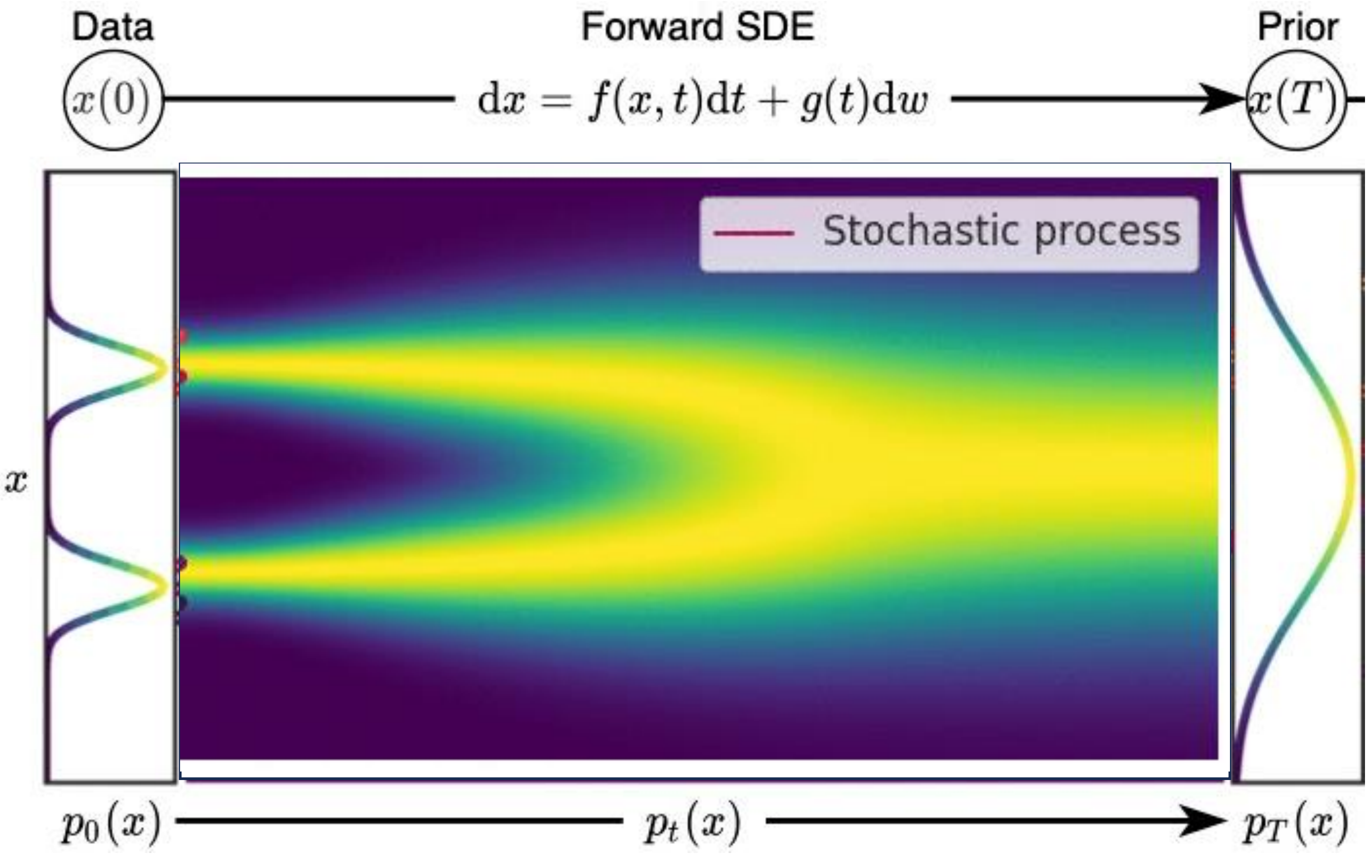
Forward SDE



Forward SDE


Data
Forward SDE
Prior
$x(0)$
$dx = f(x,t)dt + g(t)dw$
$x(T)$
Stochastic process
x
$p_0(x)$
$p_t(x)$
$p_T(x)$

Forward SDE


Data
Forward SDE
Prior
$x(0)$
$\mathrm{d}x = f(x,t)\mathrm{d}t + g(t)\mathrm{d}w$
$x(T)$
Stochastic process
x
$p_0(x)$
$p_t(x)$
$p_T(x)$

Reverse Stochastic Differential Equations (SDE)

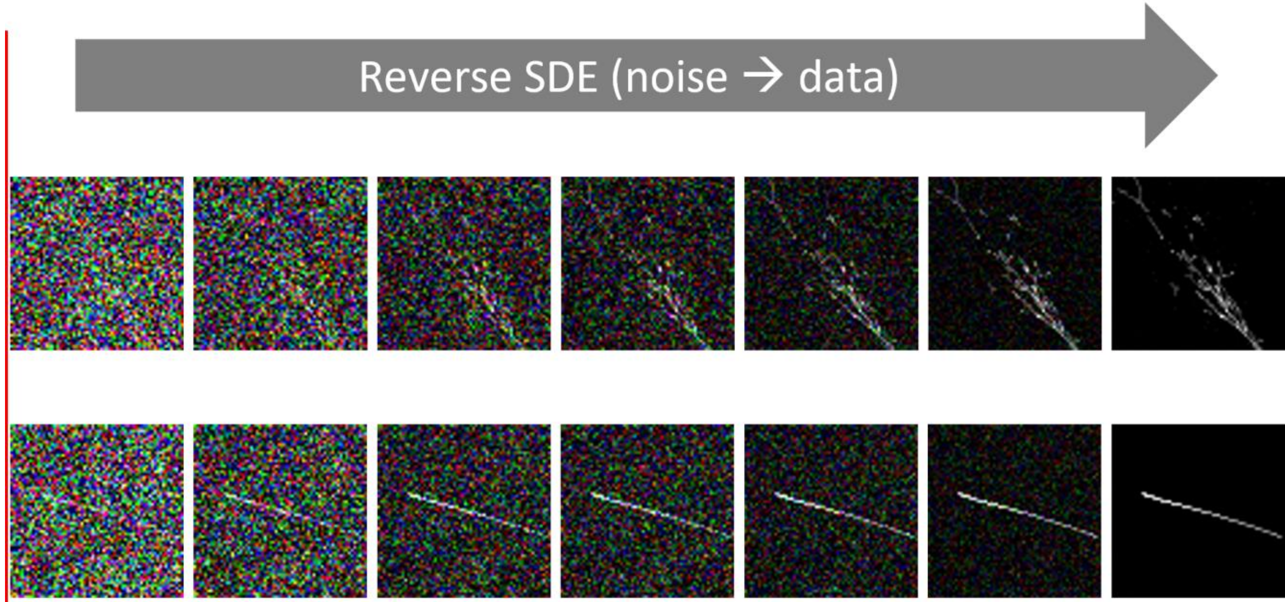
Drift (Reverse) $\mathbf{f}(\mathbf{x}, t)\mathrm{d}t$

Diffusion (Reverse) $g(t)\mathrm{d}\bar{\mathbf{w}}$

score function

$$g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})]$$

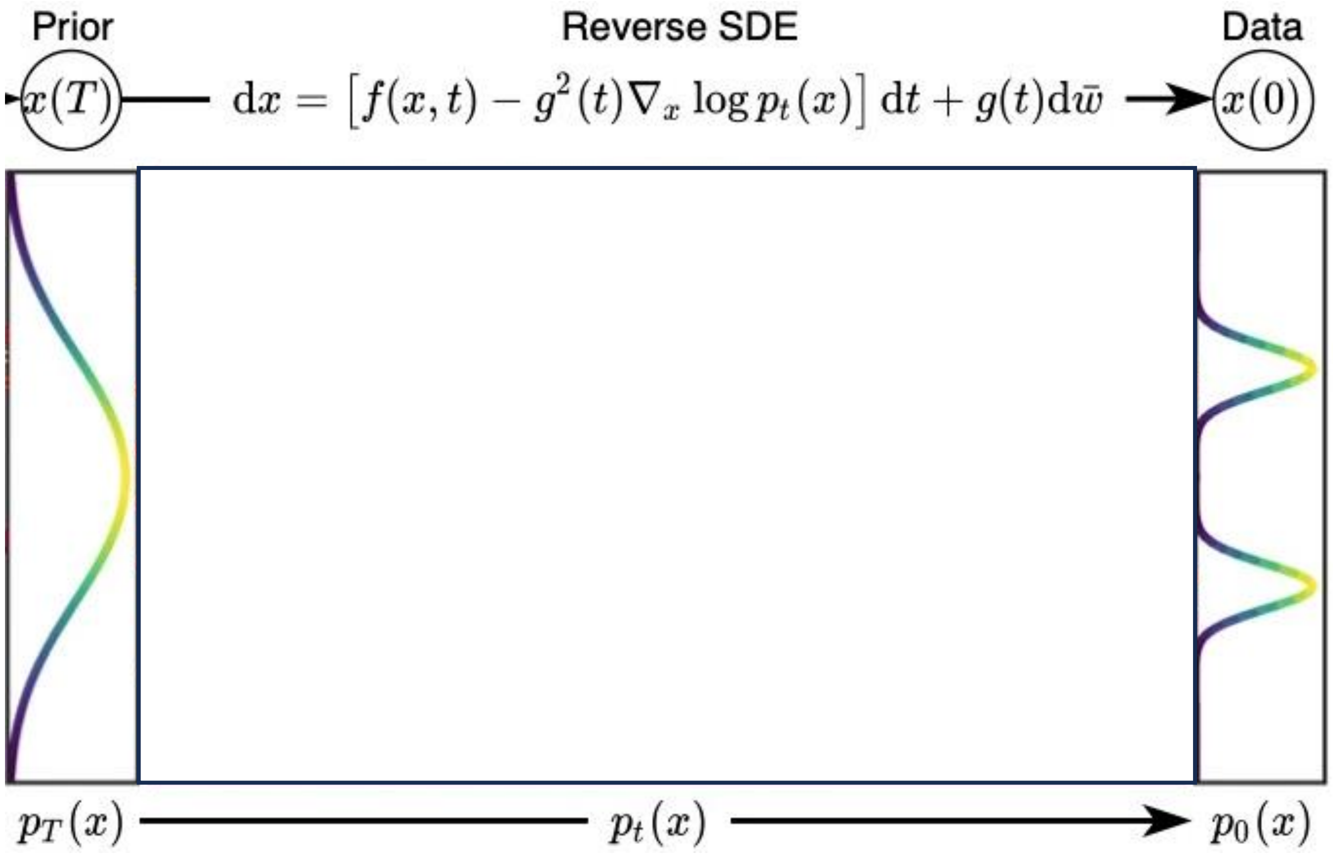
Scale factor $g^2(t) = \beta_t$



score function

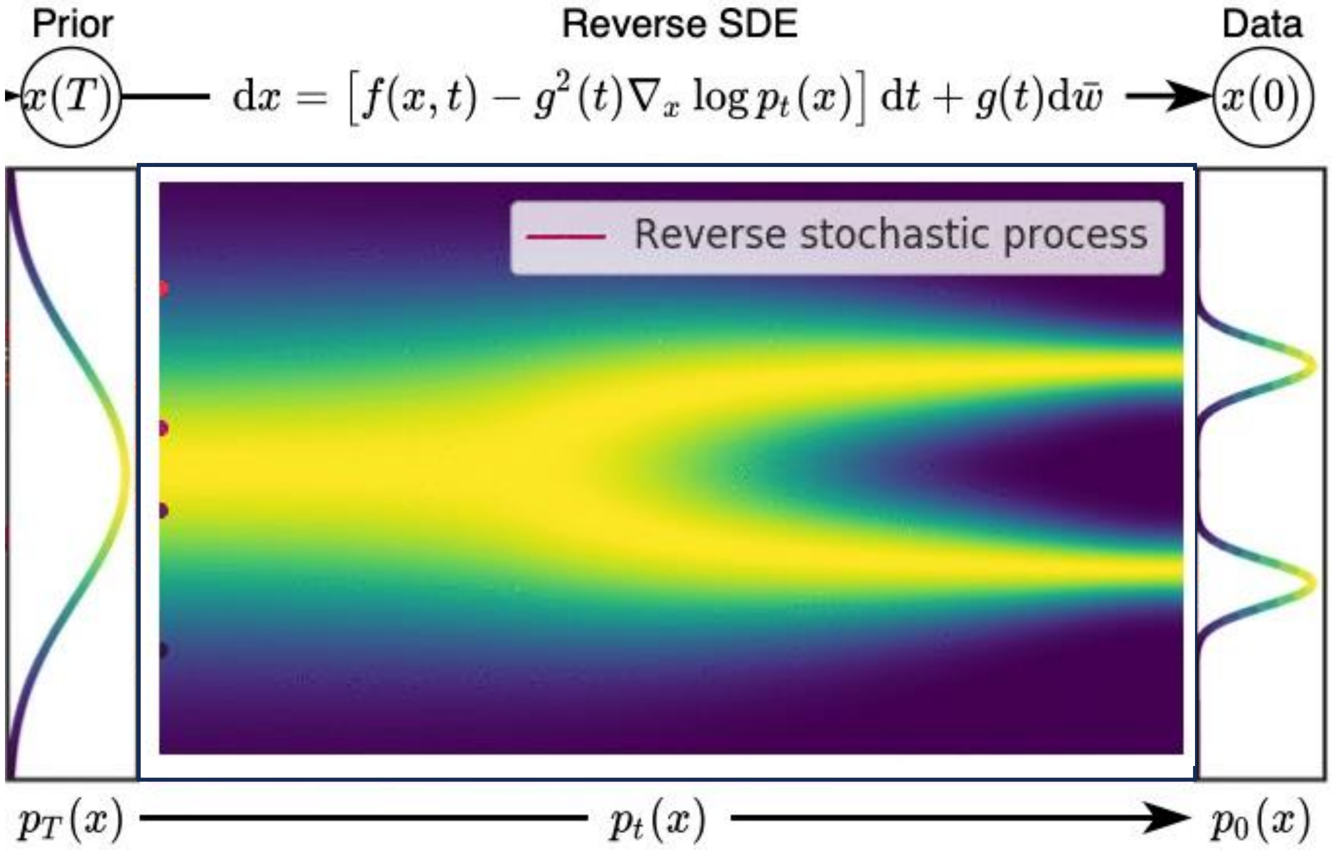
$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$$

Reverse SDE

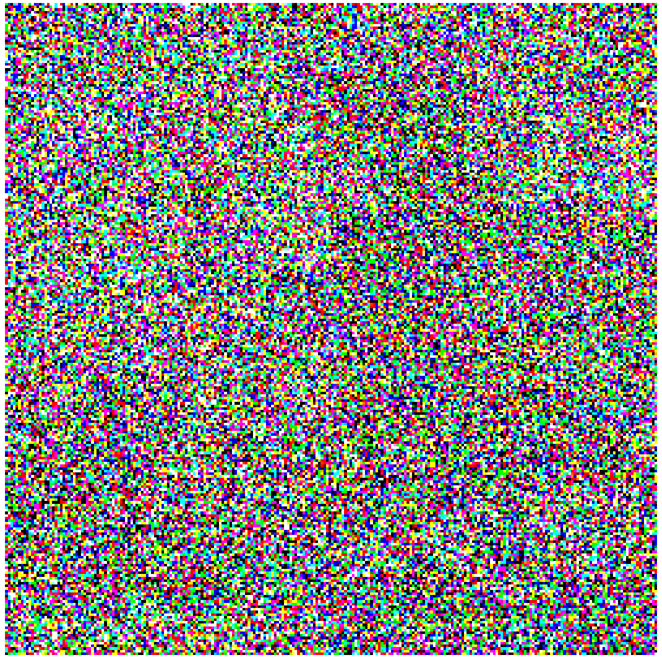


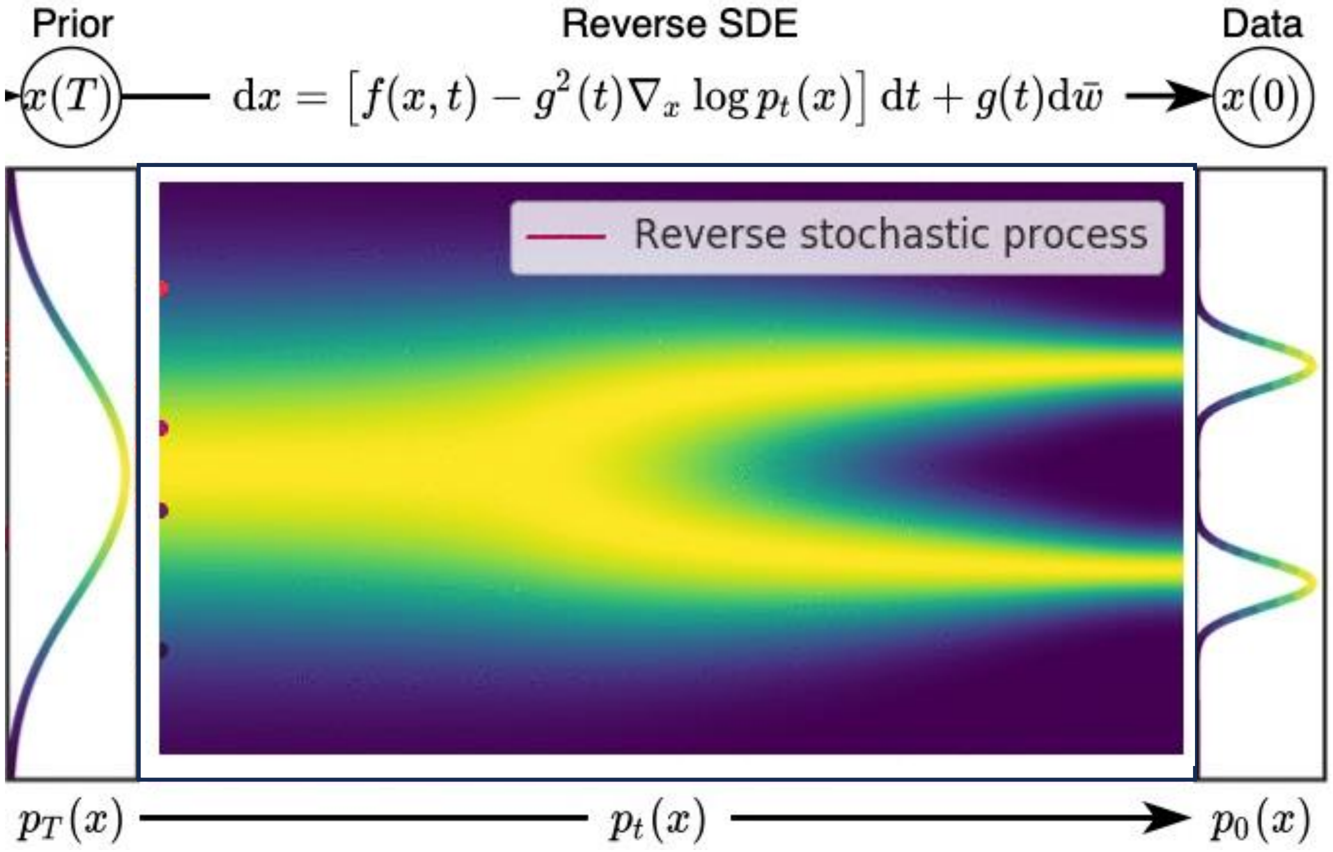
Prior
Reverse SDE
Data
$x(T)$
$\mathrm{d}x = \left[f(x,t) - g^2(t)\nabla_x \log p_t(x)\right] \mathrm{d}t + g(t)\mathrm{d}\bar{w}$
$x(0)$
$p_T(x)$
$p_t(x)$
$p_0(x)$

Reverse SDE


Prior
Reverse SDE
Data
$x(T)$
$\mathrm{d}x = \left[f(x,t) - g^2(t)\nabla_x \log p_t(x)\right]\mathrm{d}t + g(t)\mathrm{d}\bar{w}$
$x(0)$
Reverse stochastic process
$p_T(x)$
$p_t(x)$
$p_0(x)$

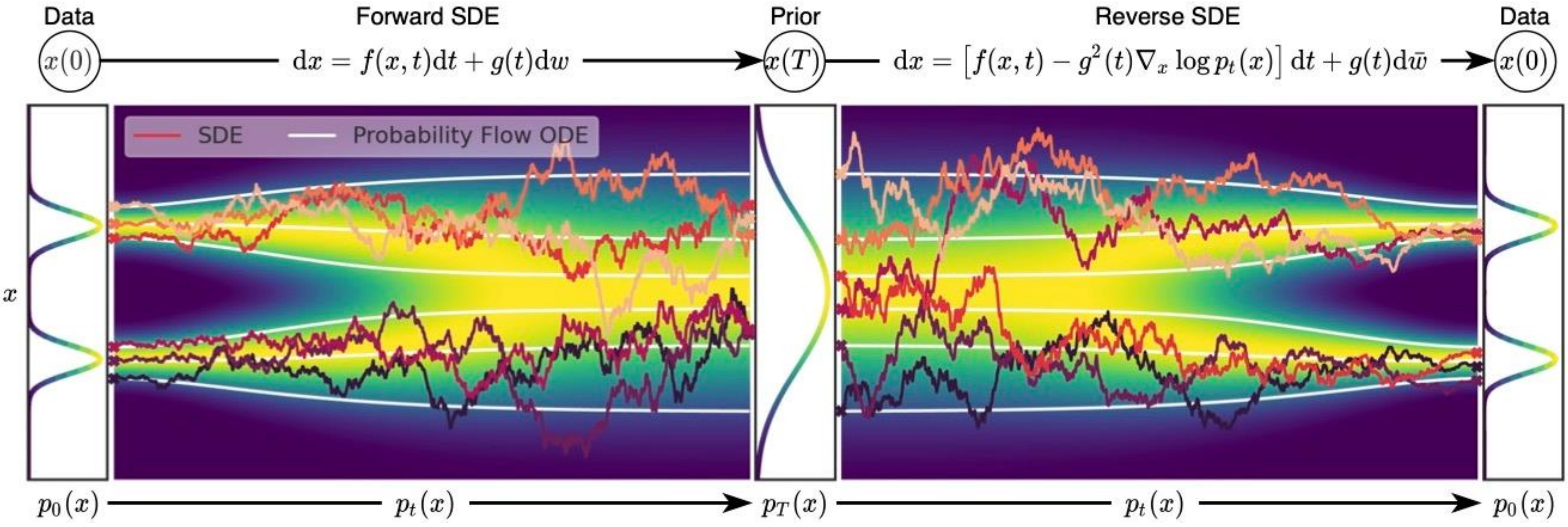
Reverse SDE



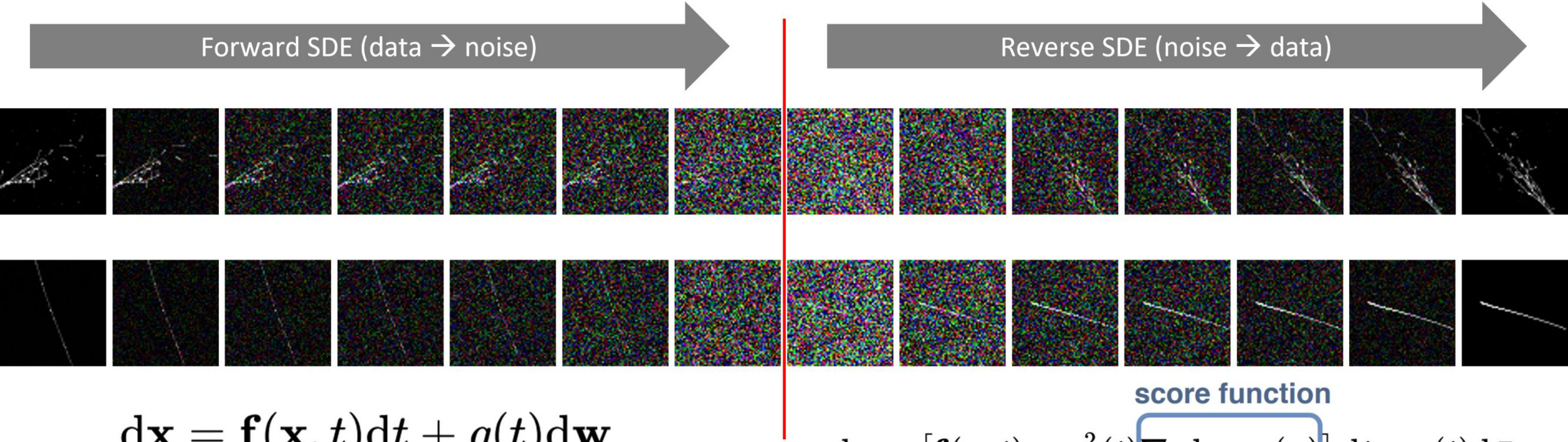


Yang Song et al., arXiv:2011.13456

Full Process



Full Process



$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}$$

score function

$$d\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\boxed{\nabla_{\mathbf{x}} \log p_t(\mathbf{x})}\right] dt + g(t)d\bar{\mathbf{w}}$$