CaloMan: Fast generation of calorimeter showers with density estimation on learned manifolds

Jesse Cresswell, Brendan Ross, Gabriel Loaiza-Ganem,
Humberto Reyes-González, Marco Letizia, Anthony Caterini

Layer 6 AI, University of Genoa, INFN

Dec. 3, 2022

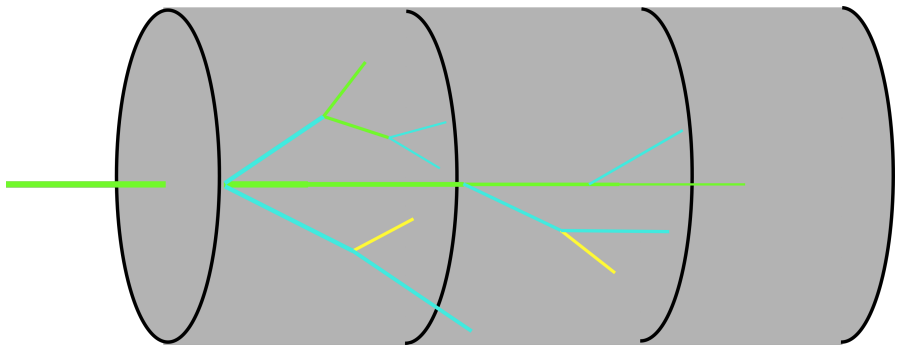
Machine Learning and the Physical Sciences Workshop at NeurIPS 2022

Fast Calorimeter Simulation Challenge 2022

LHC experiments require simulations of how particles interact with detectors, but physics-based simulations of calorimeter showers are slow.

Challenge: train a **surrogate model** that can generate realistic showers quickly and from the correct distribution.

Deep generative models trained on shower data can learn the distribution of showers, and enable fast sampling.



Manifold Hypothesis

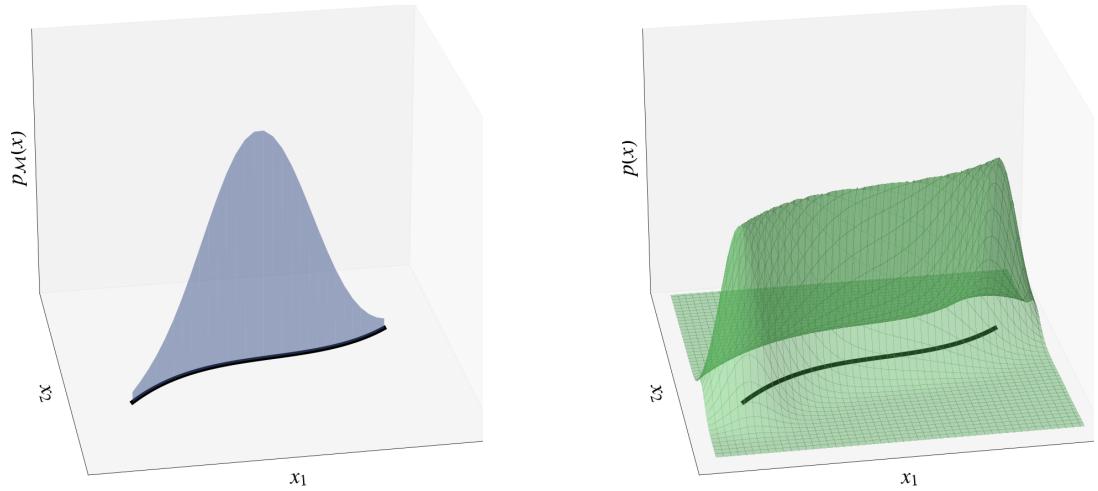
The **Manifold Hypothesis** states that high-dimensional real-world data is supported on a low-dimensional embedded submanifold $\mathcal{M} \subset \mathbb{R}^D$.
(Bengio, Courville, Vincent; IEEE TPAMI 2013)

EM calorimeter showers are highly structured.
Constraints of QED processes $\implies$ shower data has manifold structure.

Hence, the target distribution $\mathbb{P}^*$ is supported on $\mathcal{M}$, not $\mathbb{R}^D$.

What happens when we try to model $\mathbb{P}^*$ with a DGM that learns a density $p_\theta(x)$ on $\mathbb{R}^D$?

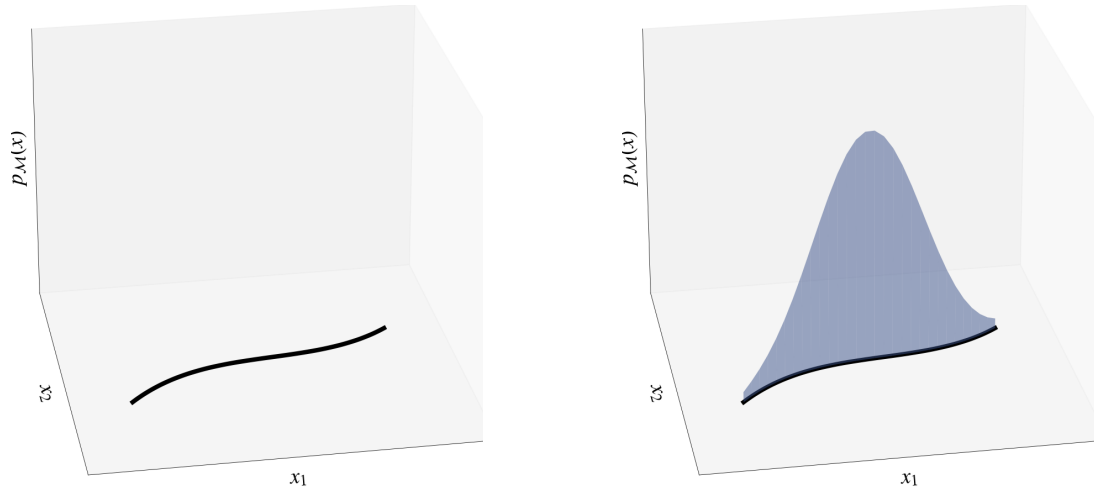
Maximum likelihood estimation can fail when the dimensionalities of $p_\theta(x)$ and $\mathbb{P}^*$ differ. **Manifold overfitting** can occur where $\mathcal{M}$ is learned but not the distribution $\mathbb{P}^*$ on it. (Loaiza-Ganem, Ross, Cresswell, Caterini; TMLR 2022)



To maximize the likelihood of the data, the density is sent to infinity around $\mathcal{M}$, where $\mathcal{M}$ is a set of measure zero wrt Lebesgue measure.

This does not happen when $p_\theta(x)$ and $\mathbb{P}^*$ have the same dimensionality because $p_\theta(x)$ must remain normalized.

The simple solution is a **two-step approach**: first learn the data manifold, then estimate the distribution on it.



1) Learn $\mathcal{M}$ with a *generalized autoencoder* - any model that constructs a low-dimensional encoding $z = g(x)$, and can reconstruct data with a decoder $x = G(z)$.

2) Perform density estimation on the manifold, obtaining the low-dimensional density $p(z)$.

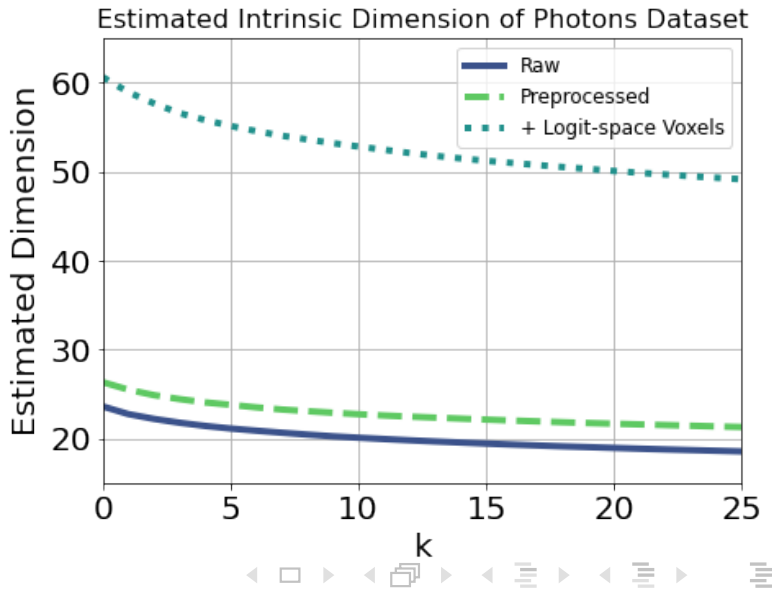
Calorimeter Shower Manifolds

Use a statistical estimator of intrinsic dimension (Levina & Bickel; NeurIPS 2004)

$$\hat{d}_k = \left(\frac{1}{n(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{k-1} \log \frac{T_k(x_i)}{T_j(x_i)} \right)^{-1}, \tag{1}$$

$T_k(x_i)$ - Euclidean distance between x_i and its kth nearest neighbour.
k - scale at which the manifold is probed.

Photon dataset showers have 368 voxels, but $\hat{d}_{10} = 20$.

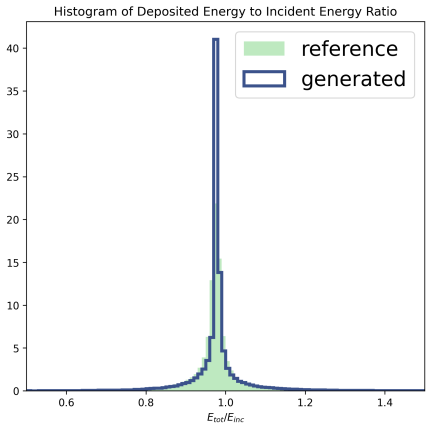


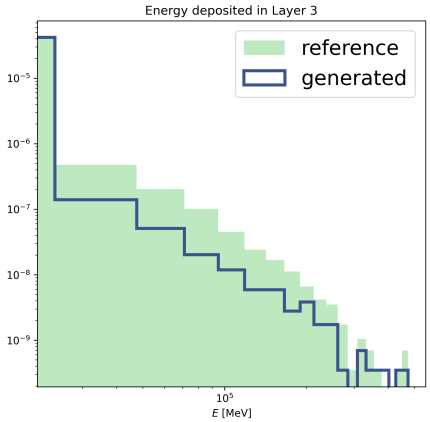
Photon Dataset - Results

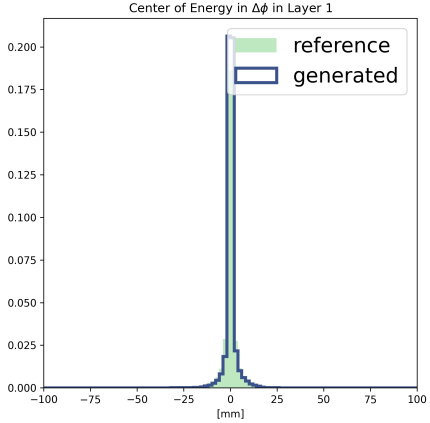
1) VAE parameterizes the encoder and decoder as MLP networks with 3 hidden layers of 512 units - output the parameters of diagonal Gaussians in 20-dimensional latent space.

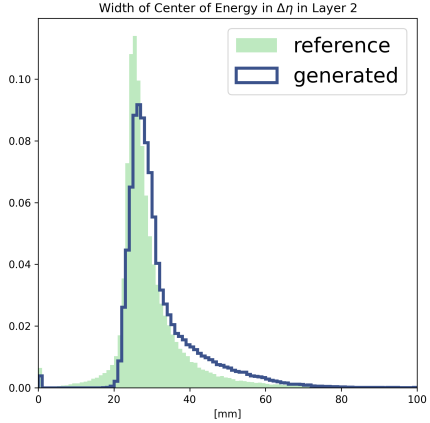
2) NF trained on 20-dimensional latent space is a 4-layer rational-quadratic neural spline flow.

Comparison of histograms between test set, and generated samples:









Conclusion

Calorimeter showers have **low-dimensional structure** dictated by physics.

Learning the manifold, then estimating the density on it is a **more principled** approach that avoids *manifold overfitting*.

Two-step models **reduce dimension**, so that training and sampling are **extremely fast** compared to full-dimensional models.

Learning topologically non-trivial manifolds without prior knowledge is also possible (Ross, Loaiza-Ganem, Caterini, Cresswell; 2206.11267).


Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2013) .


G. Loaiza-Ganem, B. L. Ross, J. C. Cresswell, and A. L. Caterini, "Diagnosing and Fixing Manifold Overfitting in Deep Generative Models," *Transactions on Machine Learning Research* (2022) .


E. Levina and P. Bickel, "Maximum likelihood estimation of intrinsic dimension," *NeurIPS* (2004) .


B. L. Ross, G. Loaiza-Ganem, A. L. Caterini, and J. C. Cresswell, "Neural implicit manifold learning for topology-aware generative modelling," *2206.11267* .