Grainger College of Engineering
Physics

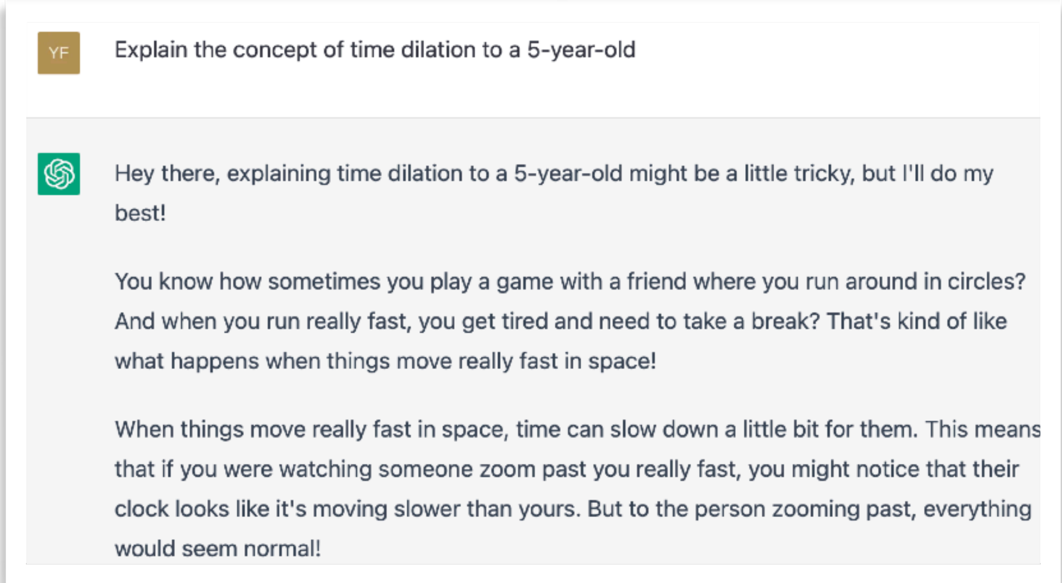
Illinois Center for Advanced Studies of the Universe

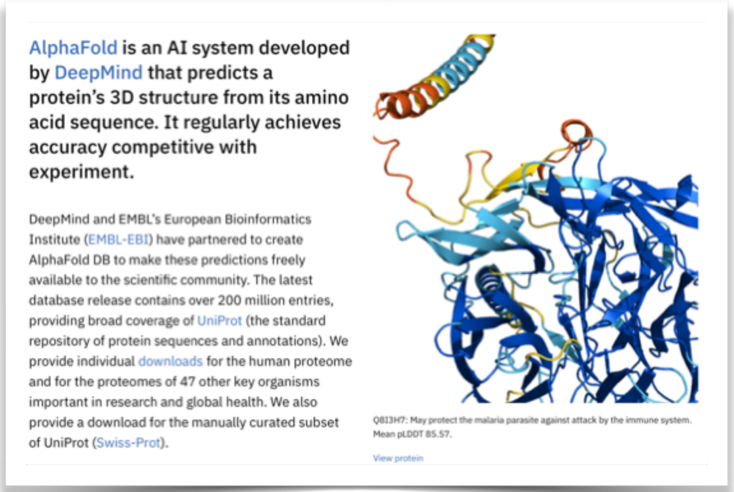
ILLINOIS NCSA
Center for Artificial Intelligence Innovation

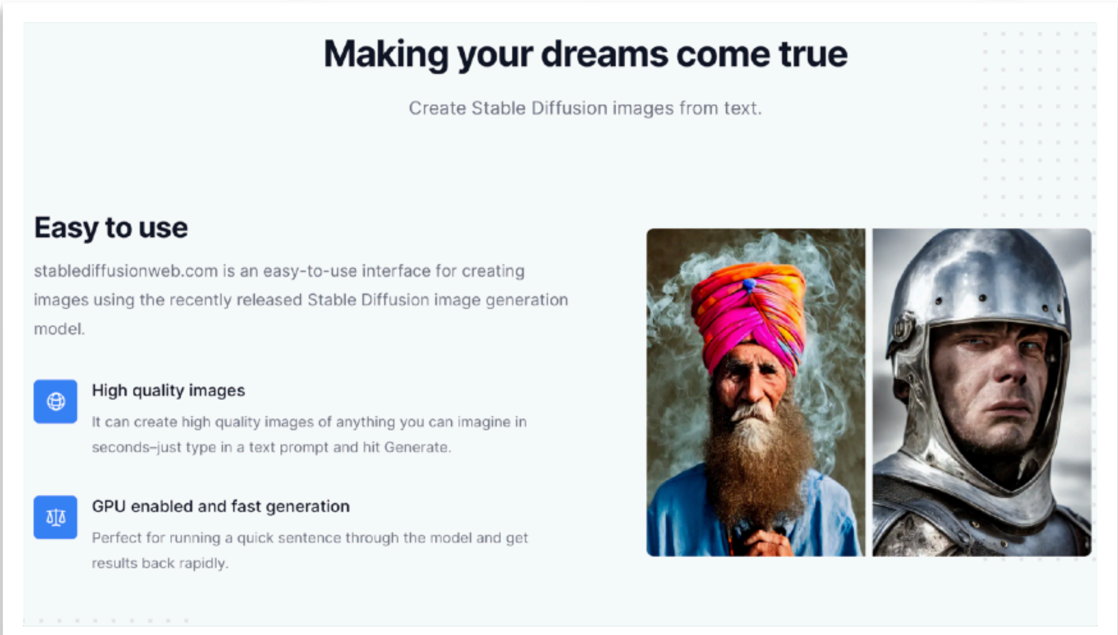
Dumb Machine Learning for Physics

Yoni Kahn
University of Illinois at Urbana-Champaign
IAIFI Colloquium, 3/10/23

The golden age of AI?





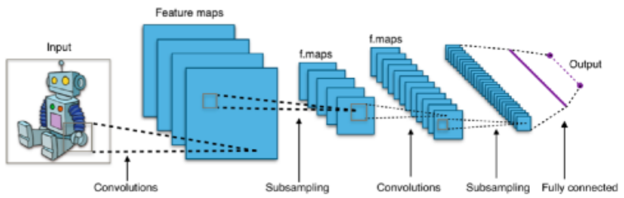


It feels (to an outsider like me) like we are on the cusp of something amazing. New tools continue to impress on an almost daily basis

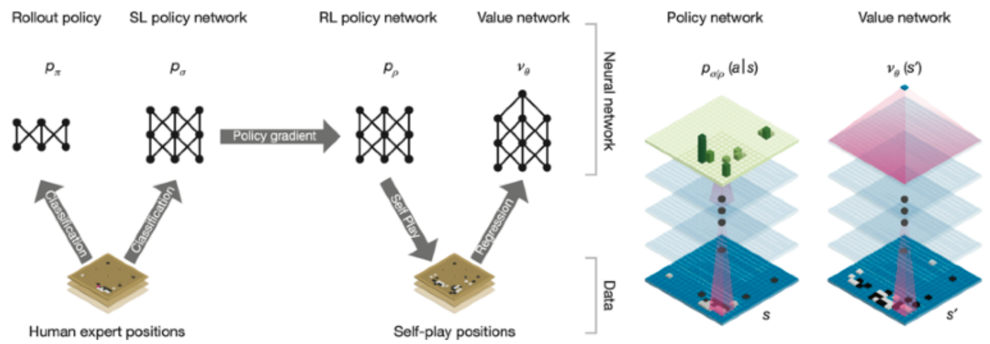
[Clockwise from top left: ChatGPT, AlphaFold, Stable Diffusion, Libratus]

Deep learning in industry, 2013-present

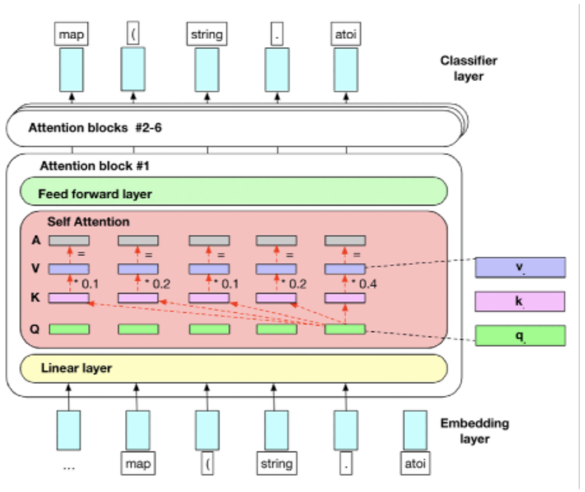
2013: AlexNet uses convolutional neural net (CNN) to win ImageNet competition w/error rate of 15.6%



2016: AlphaGo beats world experts at Go with no prior knowledge



2020: GPT-3 generates text indistinguishable from human responses in some cases

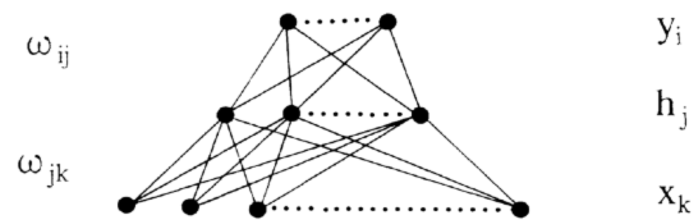


Deep learning in collider physics, 1990-present

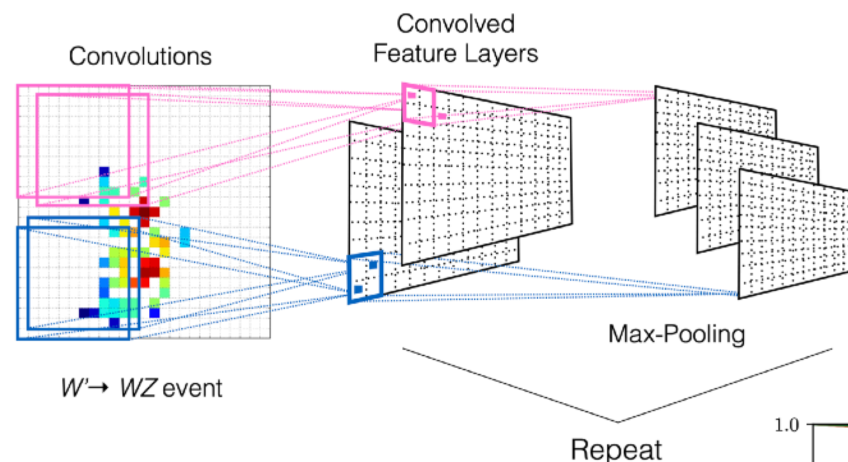
Finding Gluon Jets with a Neural Trigger

Leif Lönnblad,[a] Carsten Peterson,[b] and Thorsteinn Rögnvaldsson[c]
Department of Theoretical Physics, University of Lund, Sölvegatan 14A, S-22362 Lund, Sweden
(Received 6 April 1990)

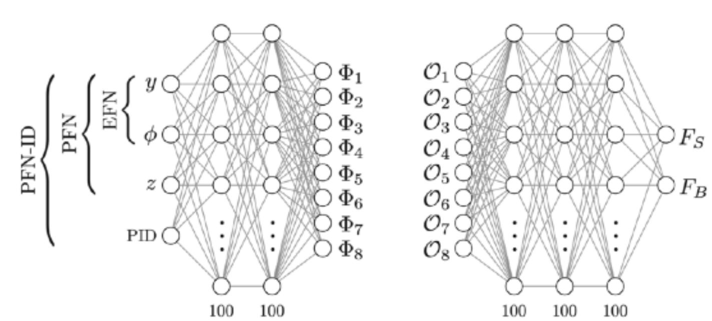
Using a neural-network classifier we are able to separate gluon from quark jets originating from Monte Carlo-generated e^+e^- events with 85%–90% accuracy.

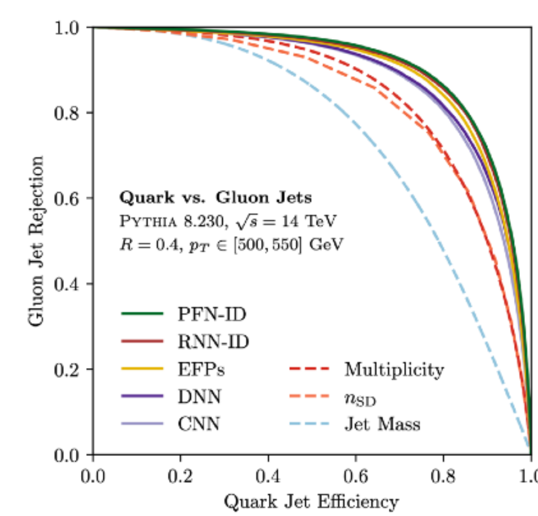


2017: CNNs for
jet image classification



2018: permutation-invariant deep sets for jet classification (just two feedforward networks!)



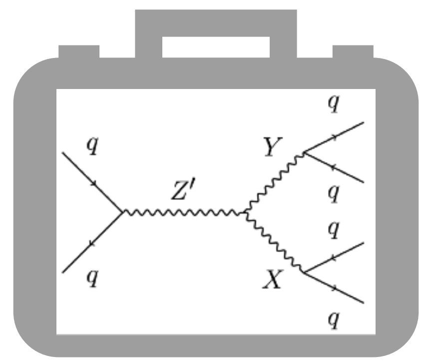


[Lönnblad, Peterson, Rögnvaldsson, PRL 1990; de Oliveira et al., JHEP 2016; Komiske, Metodiev, Thaler, JHEP 2019]

Where is our ChatGPT or AlphaGo in physics?

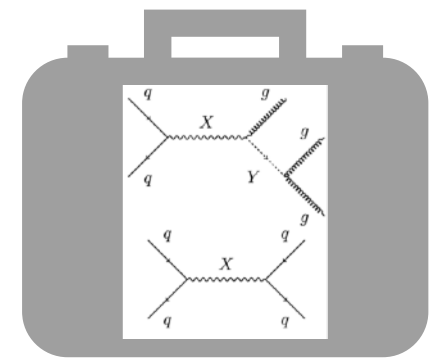
The LHC Olympics 2020

A Community Challenge for Anomaly Detection in High Energy Physics



Box 1:
2-body resonance,
same topology as R&D

Box 2:
nothing
(bg-only)



Box 3:
two decay
channels

- The black box number (1-3) corresponding to their submission.
- A short abstract describing their method.
- A p-value associated with the dataset having no new particles (null hypothesis).
- As complete a description of the new physics as possible. For example: the masses and decay modes of all new particles (and uncertainties on those parameters).
- How many signal events (with the associated uncertainty) are in the dataset (before any selection criteria).

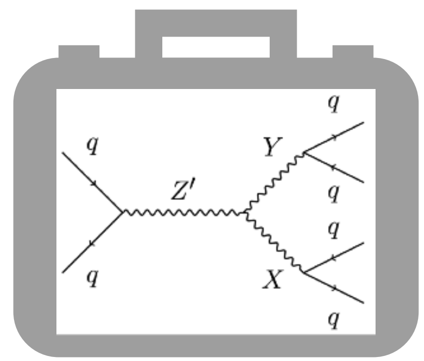
How did we do?

[Kasieczka et al., Rept. Prog. Phys. 2021]

Where is our ChatGPT or AlphaGo in physics?

The LHC Olympics 2020

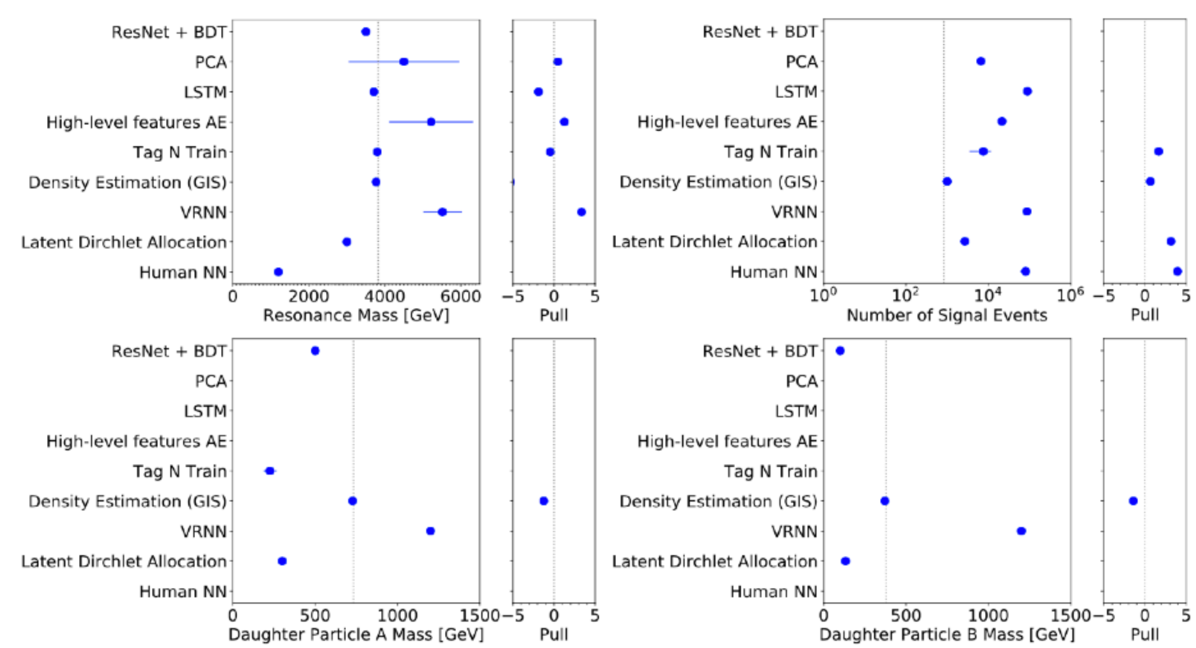
A Community Challenge for Anomaly Detection in High Energy Physics



Box 1:
2-body resonance,
same topology as R&D

- The black box number (1-3) corresponding to their submission.
- A short abstract describing their method.
- A p-value associated with the dataset having no new particles (null hypothesis).
- As complete a description of the new physics as possible. For example: the masses and decay modes of all new particles (and uncertainties on those parameters).
- How many signal events (with the associated uncertainty) are in the dataset (before any selection criteria).

Of these submissions, four approaches identified the correct resonance mass either within the claimed error (PCA) or within a window of ±200 GeV (LSTM, Tag N Train, Density Estimation). Accurate predictions for the other observables were achieved only by the Density Estimation method.



Where is our ChatGPT or AlphaGo in physics?

The LHC Olympics 2020

A Community Challenge for Anomaly Detection in High Energy Physics

Box 2:
nothing
(background only)

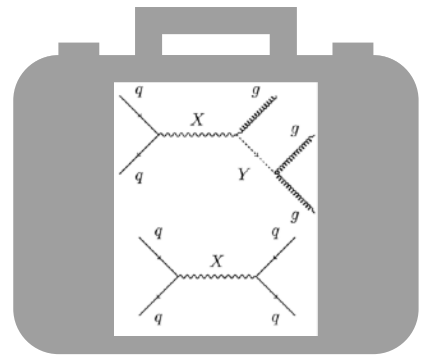
- The black box number (1-3) corresponding to their submission.
- A short abstract describing their method.
- A p-value associated with the dataset having no new particles (null hypothesis).
- As complete a description of the new physics as possible. For example: the masses and decay modes of all new particles (and uncertainties on those parameters).
- How many signal events (with the associated uncertainty) are in the dataset (before any selection criteria).

Next, Black Boxes 2 and 3 were unblinded in Summer 2020 [37]. For Black Box 2, a resonance at 4.8 TeV (PCA), at 4.2 TeV (VRNN, Sec. 3.1), 4.6 TeV (embedding clustering, Sec. 3.9), and 5 TeV (QUAK, Sec. 5.3) were predicted. For LDA (Sec. 3.6), the absence of signal as di-jet resonance was reported. As Black Box 2 did not contain any injected signal, these results highlight a possible vulnerability of anomaly detection methods in the tail of statistical distributions.

Where is our ChatGPT or AlphaGo in physics?

The LHC Olympics 2020

A Community Challenge for Anomaly Detection in High Energy Physics



Box 3: two decay channels

- The black box number (1-3) corresponding to their submission.
- A short abstract describing their method.
- A p-value associated with the dataset having no new particles (null hypothesis).
- As complete a description of the new physics as possible. For example: the masses and decay modes of all new particles (and uncertainties on those parameters).
- How many signal events (with the associated uncertainty) are in the dataset (before any selection criteria).

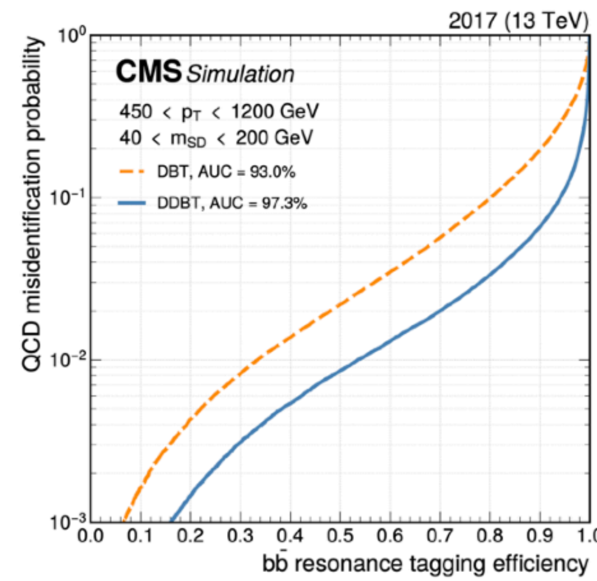
For Black Box 3 a resonance decaying to hadrons and invisible particles (PCA), a resonance with a mass between 5.4 and 6.4 TeV (LDA), at 3.1 TeV (embedding clustering), and between 5 and 5.5 TeV (QUAK) was reported. No signal was observed by one approach (VRNN). The true injected resonance with a mass of 4.2 TeV and two competing decay modes was not detected by any approach.

For Black Box two and three, no additional observations of a signal were reported after unblinding.

When has deep learning made **discoveries** in physics?

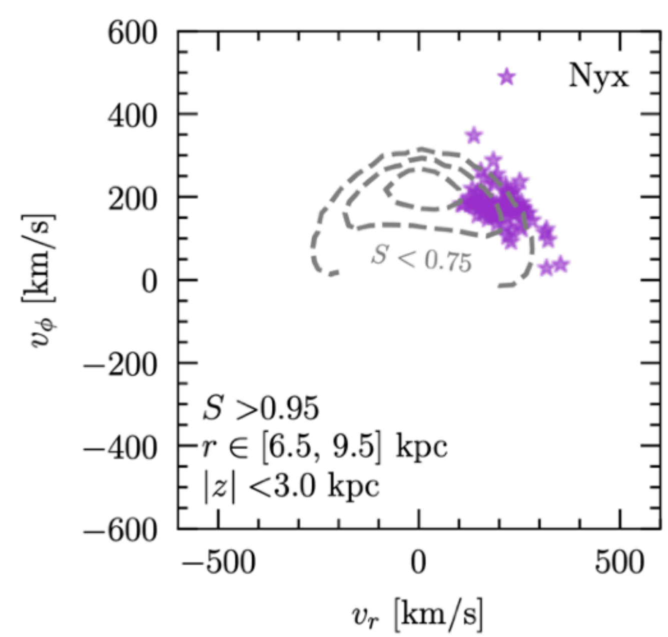
Higgs discovery in bottom quark channel:

The neural network uses four hidden layers that are fully connected, each with 100 nodes. Increasing the number of hidden layers and the number of nodes per layer had negligible effects on the performance.



Stellar stream discovery with Gaia:

All networks in this study are constructed with five layers (the input, three hidden layers, and the output). The networks take between 4 and 9 measured quantities per star as inputs — these variations will be discussed in detail below. The hidden layers consist of 100 nodes each, using a ReLU activation function, *i.e.*, $h(x) = \max(0, x)$.

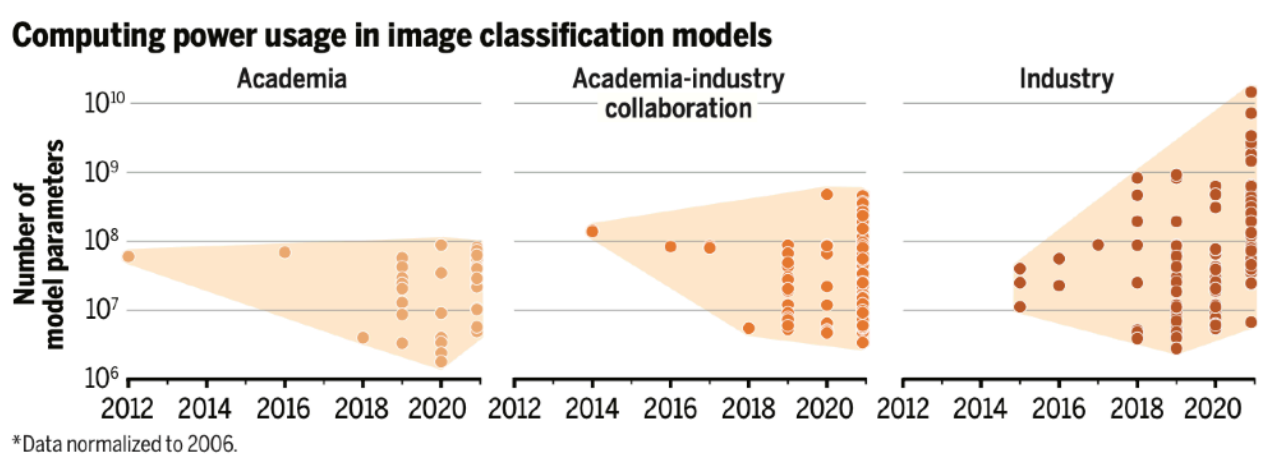


[CMS collab., JHEP 2020; Necib et al., Nature Astron. 2020]

Three possible explanations

1. Industry practitioners are smarter/more clever/ more competent than physicists.

2. Physics is compute-limited, all the best GPUs/ TPUs are at private companies.

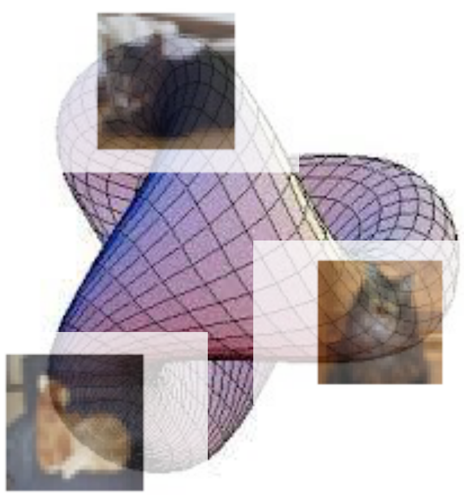


3. Physics data is qualitatively and quantitatively different than data “in the wild”

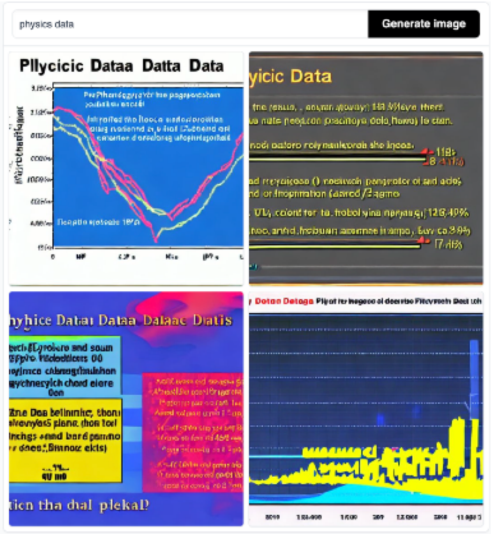
Why might physics data be different?

Physics data lives on manifolds

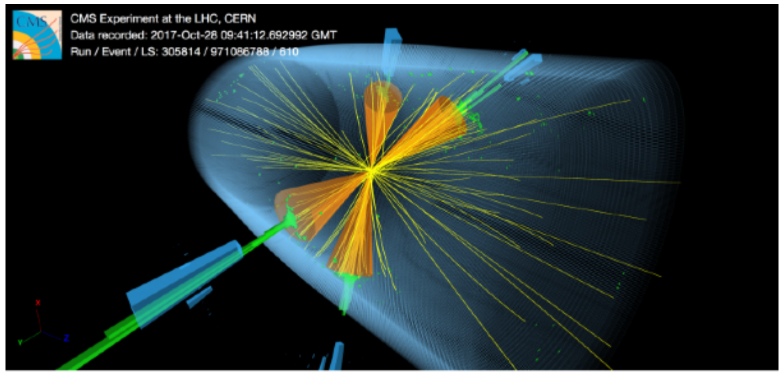
Physics cares about uncertainty quantification



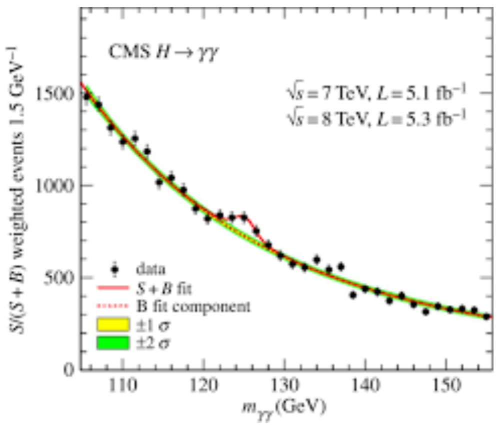
$\mathcal{M}_{\text{cats}} \hookrightarrow \mathbb{R}^{32\times 32\times 3}$?



vs.

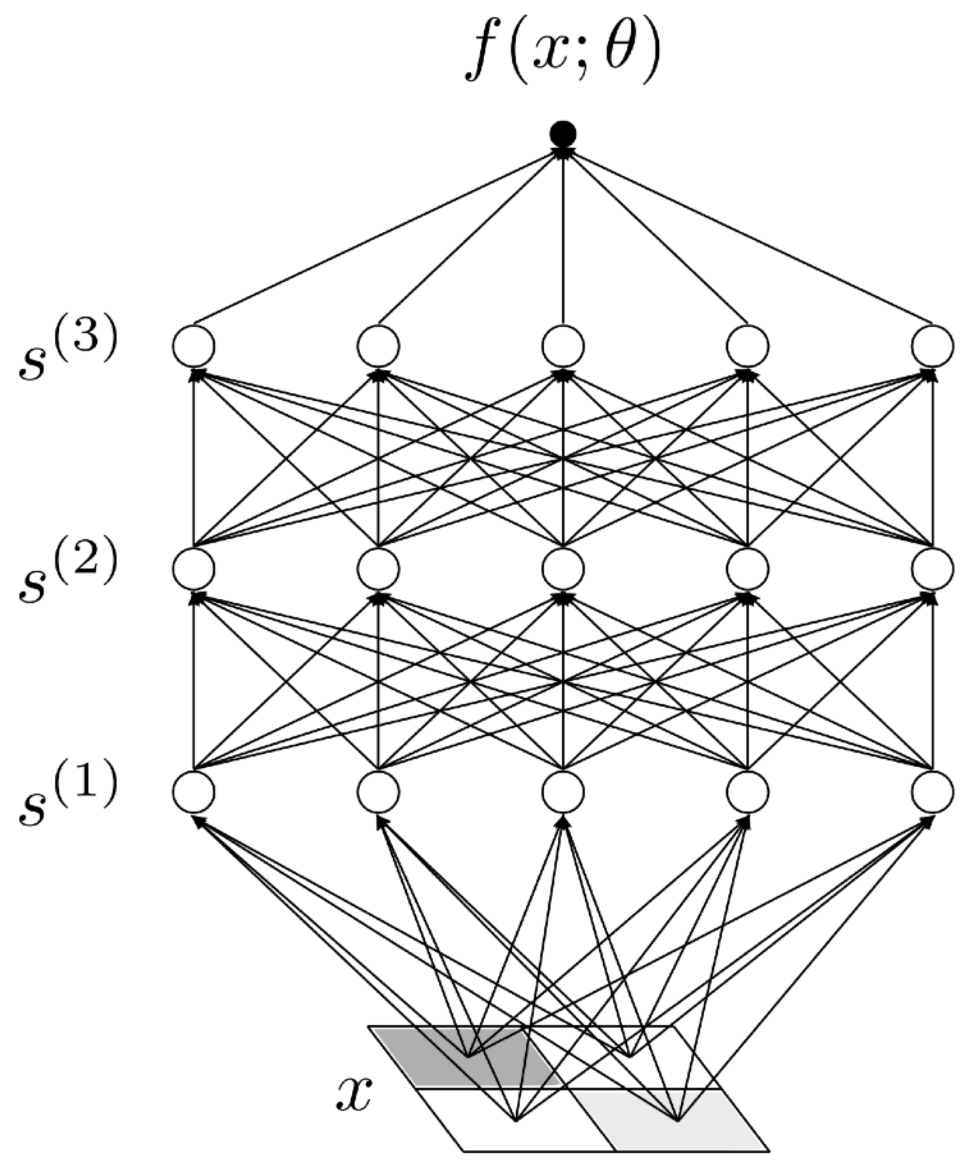


vs.



[Image credits: CMS collab., Stable Diffusion]

Dumb machine learning



A fully-connected network (FCN) or multi-layer perceptron (MLP) already has an incredible amount of structure.

Let's understand that first before jumping to fancier tools.

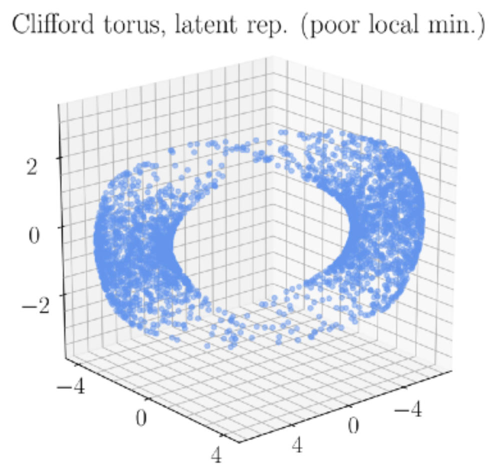
[Image credit: Roberts, Yaida, Hanin, Cambridge University Press 2022]

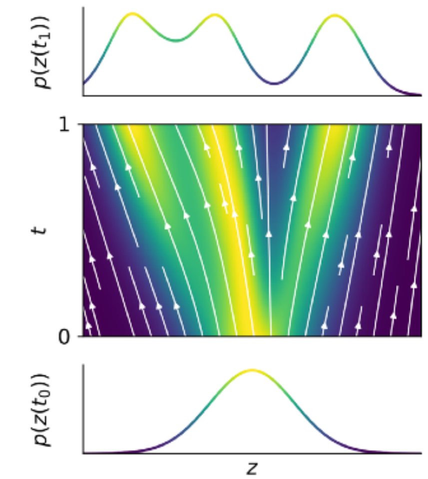
A series of illustrative examples

- How to succeed at a hard problem and fail at an easy one:

1. Data topology, autoencoders, and collider anomaly detection

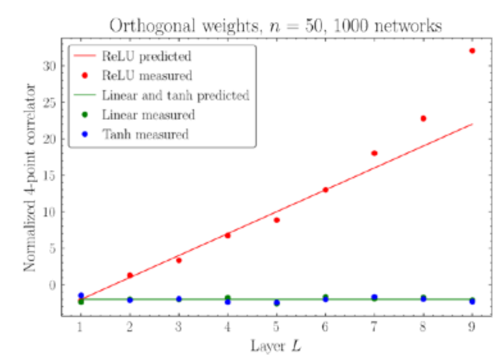
2. Learning stellar phase space with normalizing flows





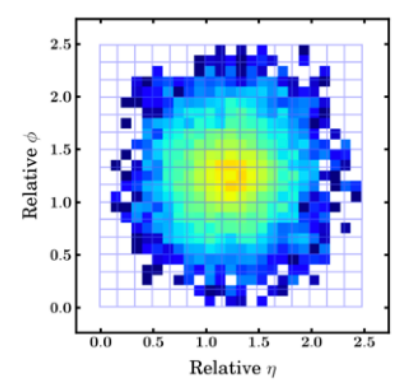
- How to control your network fluctuations:

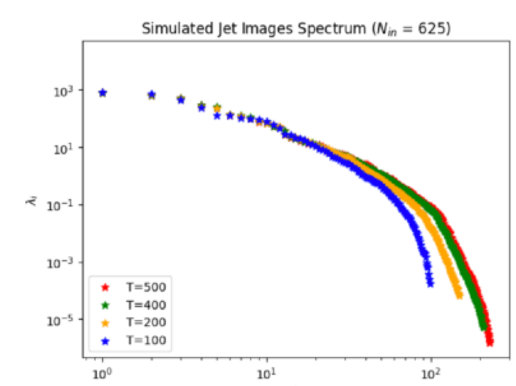
3. Orthogonal initializations and feature learning



- How to balance model size and data:

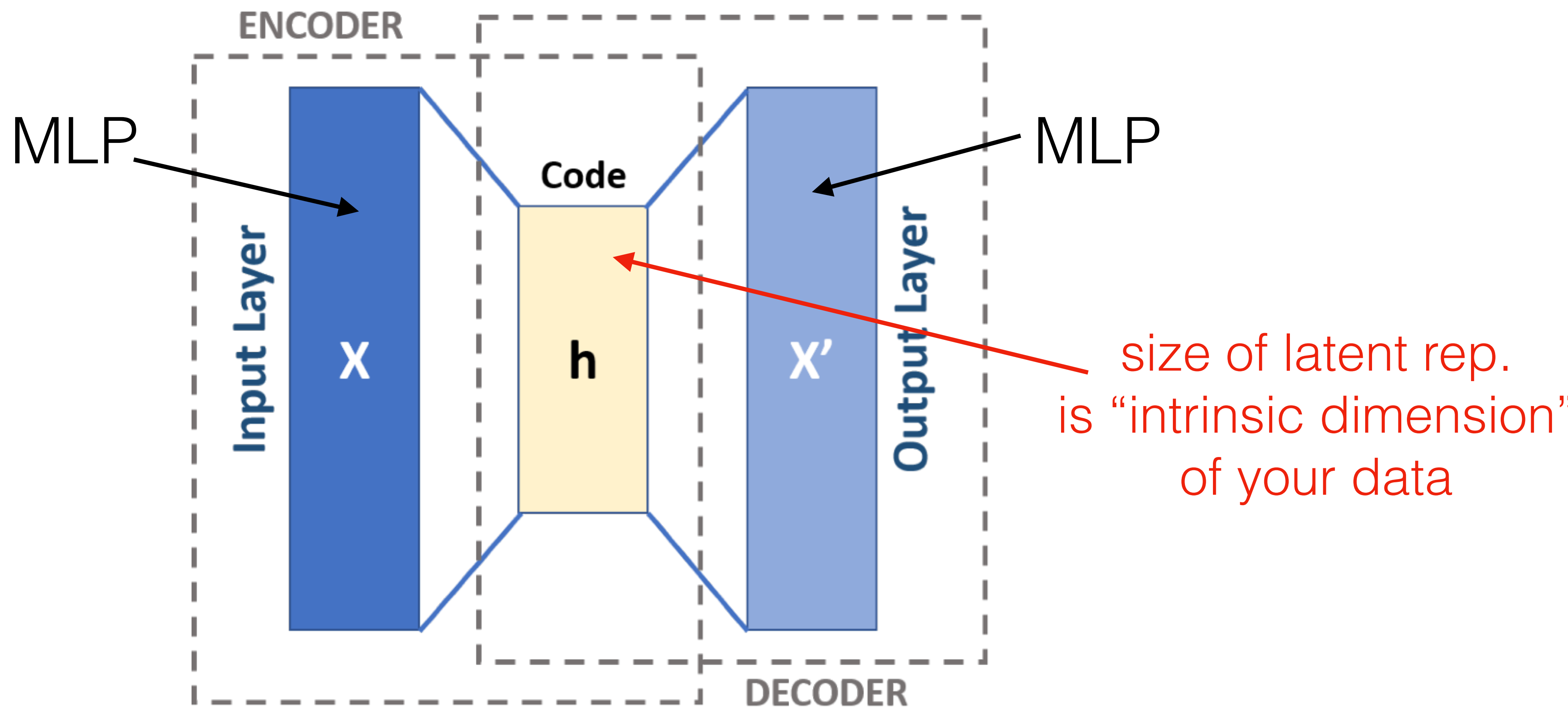
4. Data dimension and power laws in jet classification





Unifying themes: ensemble variance and data dimensionality may matter more in physics applications than in industry!

Autoencoders for anomalies



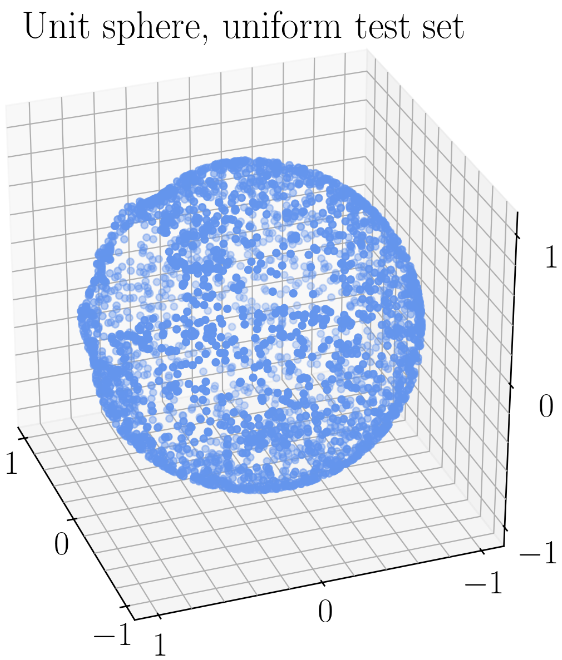
$$L = ||f(\mathbf{x}) - \mathbf{x}||^2$$

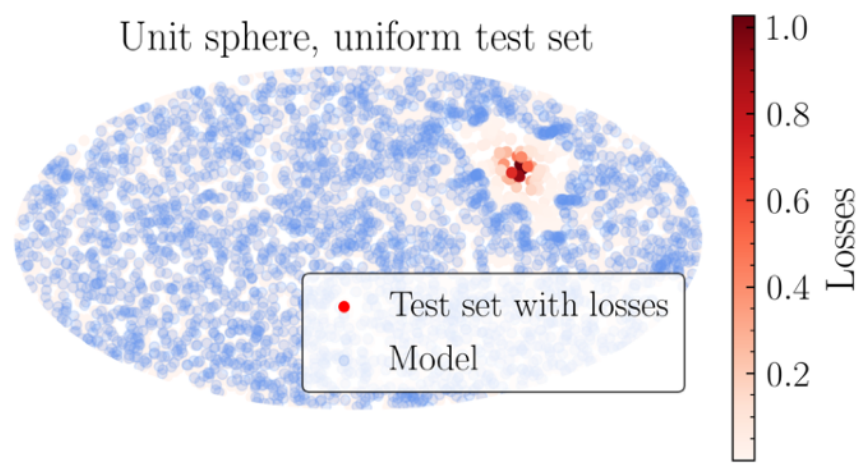
Approximates the identity function
on your data, but should be garbage on any other input

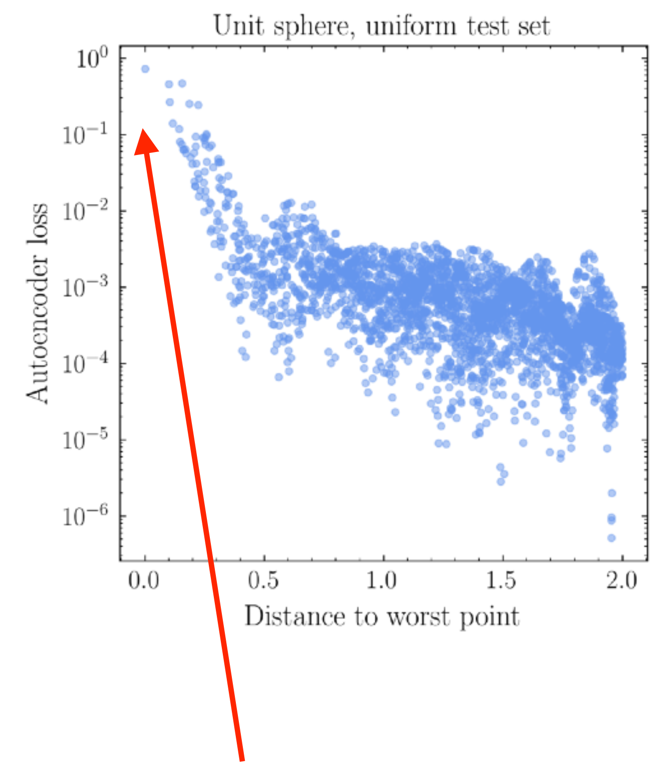
"Normal" data = good reconstruction, "anomalous" data = poor reconstruction

Many ideas in the literature for model-independent anomaly detection

Can an autoencoder learn a 2-sphere?

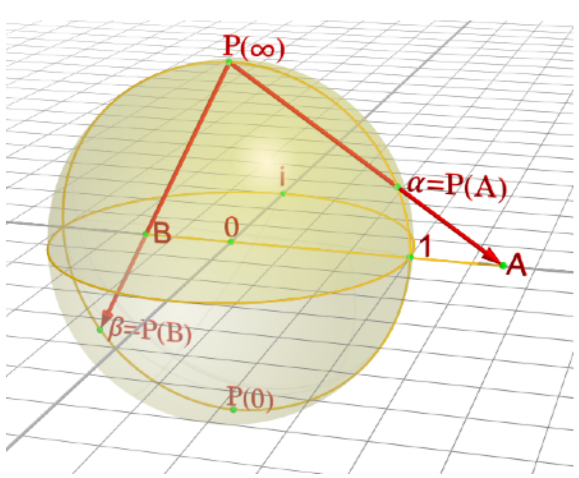






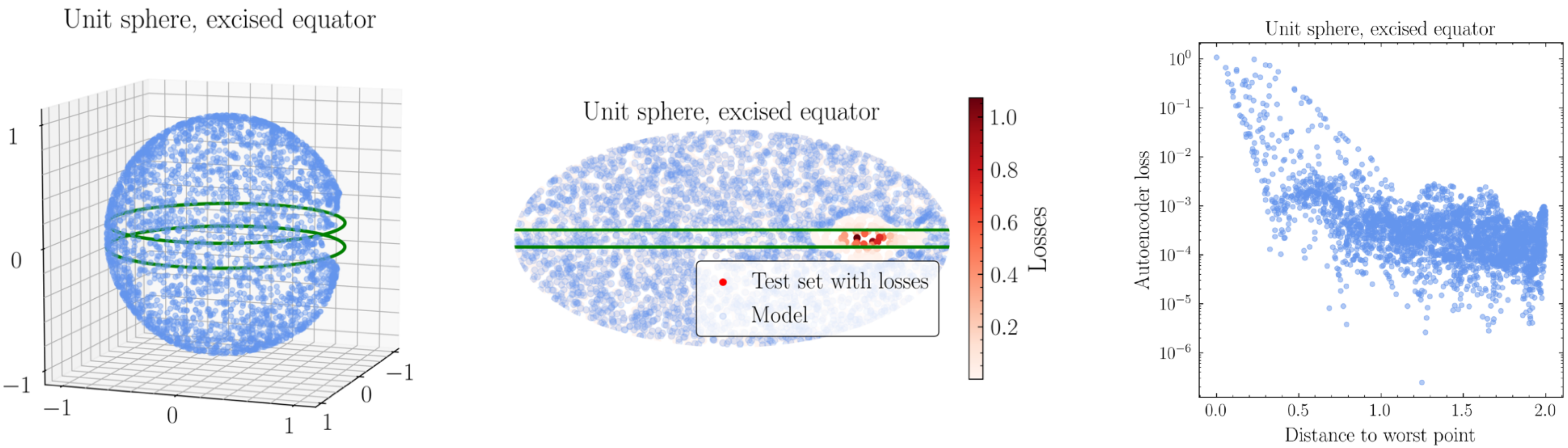
Neural nets can learn stereographic projection: map from sphere to plane must break at a single point

Loss is localized in neighborhood of a point



Bad loss from a random (non-anomalous) point, because data has topology!

“Anomalous” submanifolds



Excising a submanifold poses no problems for interpolation. The equator should be an anomaly (completely absent from training set), yet test loss is comparable to any other generic point, except at an isolated point

NN’s have an inductive bias towards interpolation, want to trivialize the topology in simplest way possible. Problem for submanifold anomaly identification with autoencoders

(3-body) phase space is a topological (5)-sphere

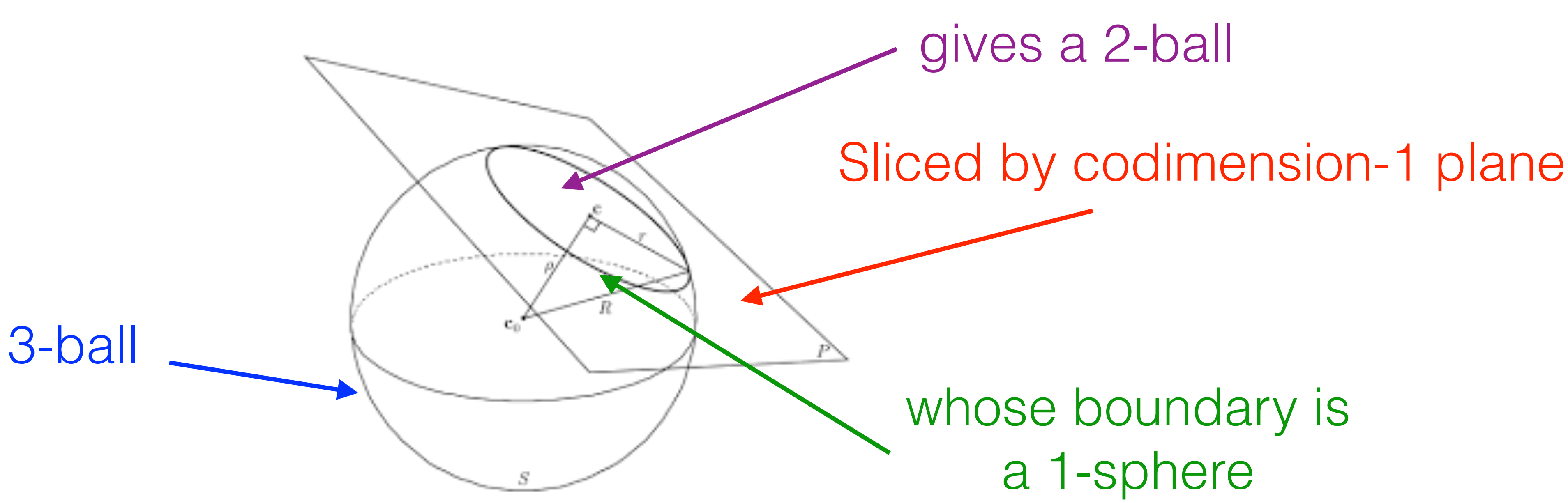
$E_i = \sqrt{m_i^2 + |\vec{p}_i|^2}$ energy is a convex function of momenta

$E_1 + E_2 + E_3 \leq E_0$ defines a convex ball in $\mathbb{R}^9$

$\vec{p}_1 + \vec{p}_2 + \vec{p}_3 = \vec{0}$ event plane slices 9-ball to form a 6-ball

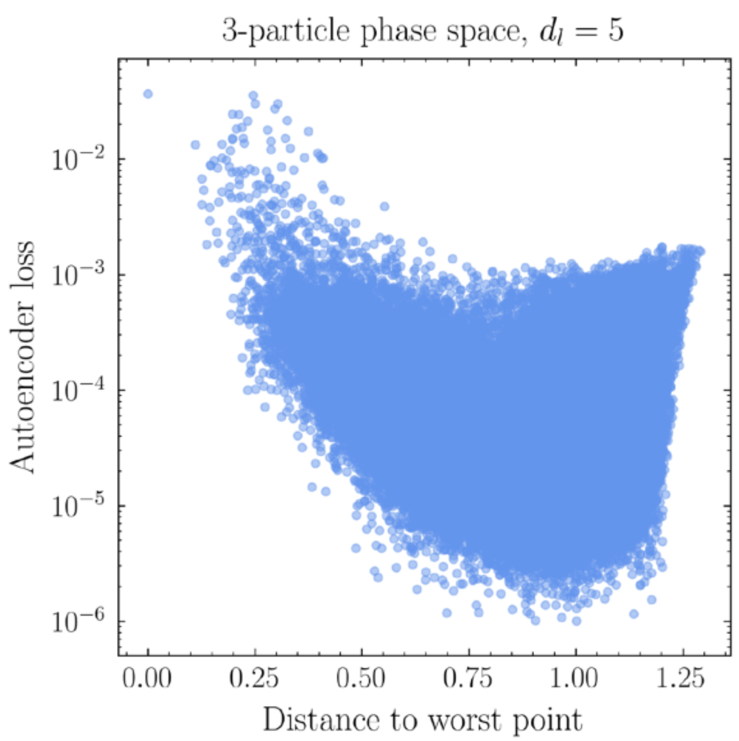
$E_0 = E_{\rm CM}$ boundary of 6-ball is 5-sphere

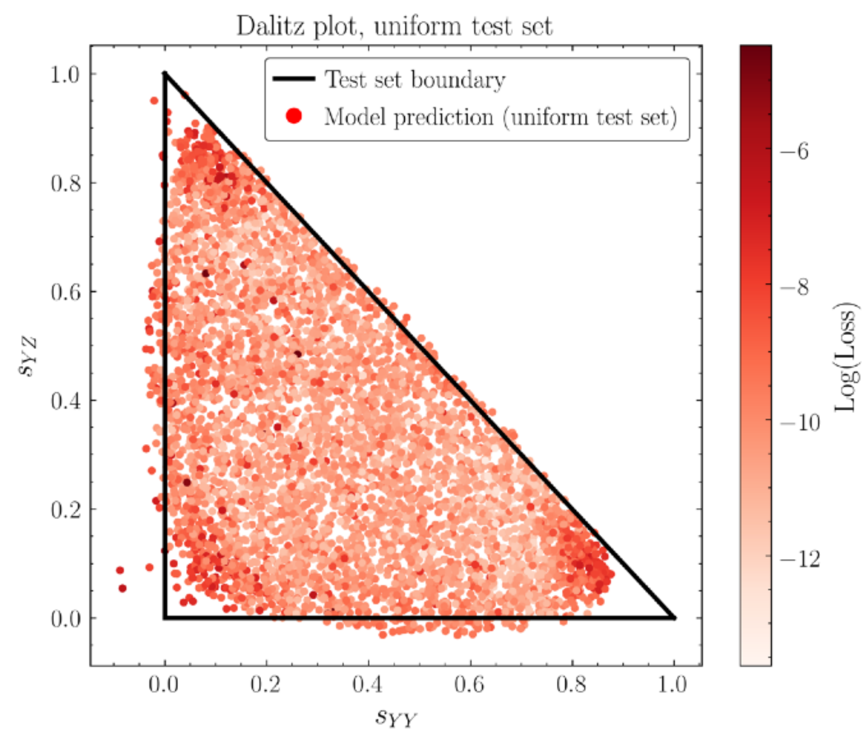
Lower-dimensional analogue for visualization:

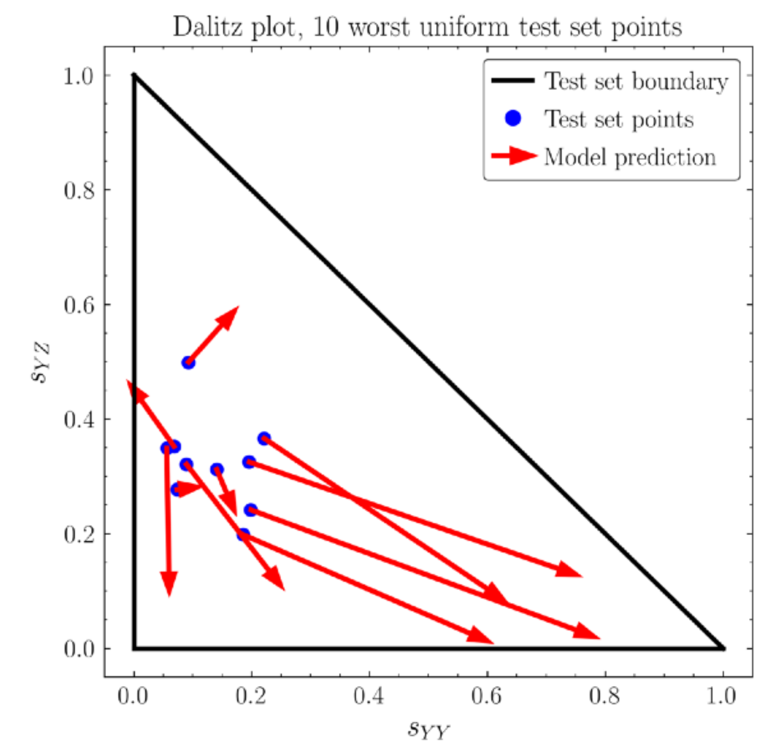


Autencoding phase space

Train an autoencoder with latent dimension 5 on
4-vectors uniformly sampled from 3-body phase space:







Looks like a sphere:
loss localized near
a single point

Haar measure on SO(3)
effectively oversamples boundaries
(collinear redundancy):
local max. of loss at corners

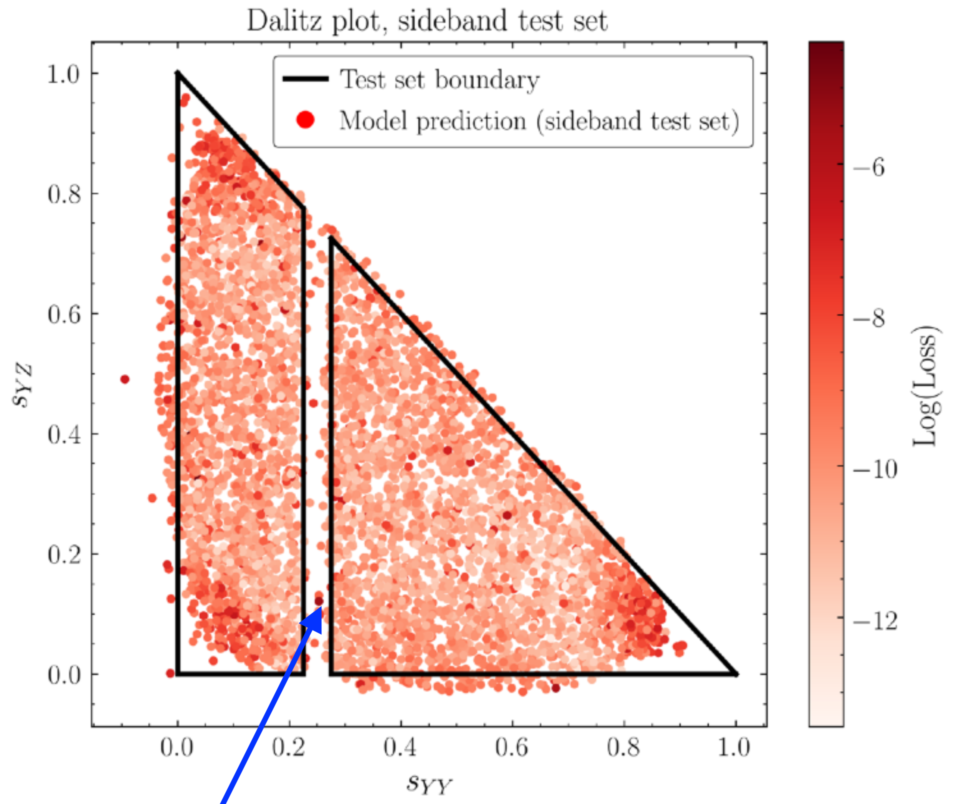
Global max. of loss
at random point
in interior

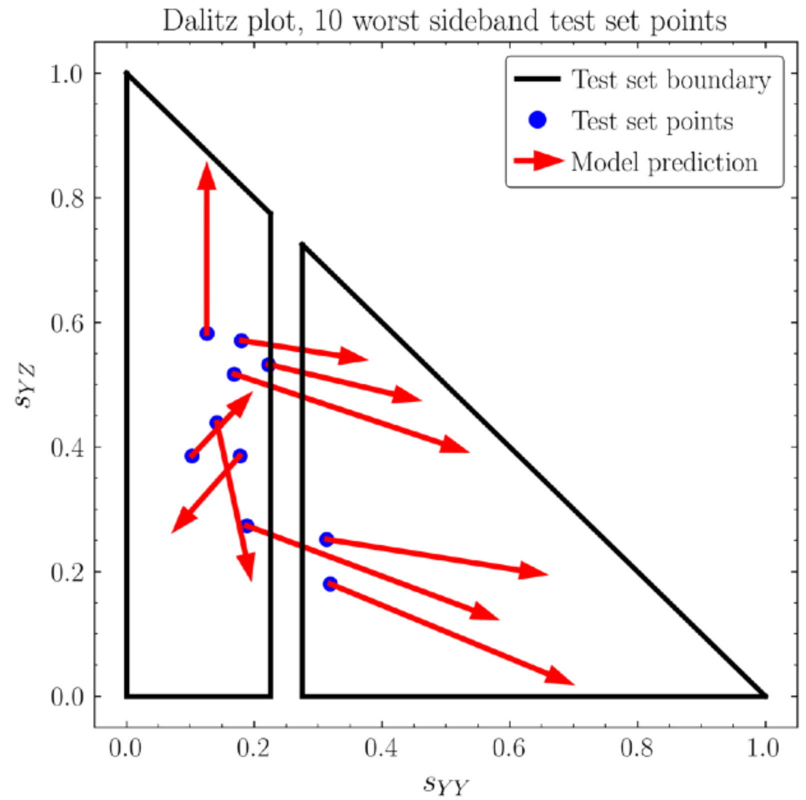
Nothing is anomalous here! Every point in phase space is as good as any other, but the latent map has to break somewhere

[Batson, Haaf, **YK,** Roberts, JHEP 2021]

Failure of a bump hunt

Let's use this phase space example as a cartoon of a bump hunt in leptons, where collider observables are 4-vectors





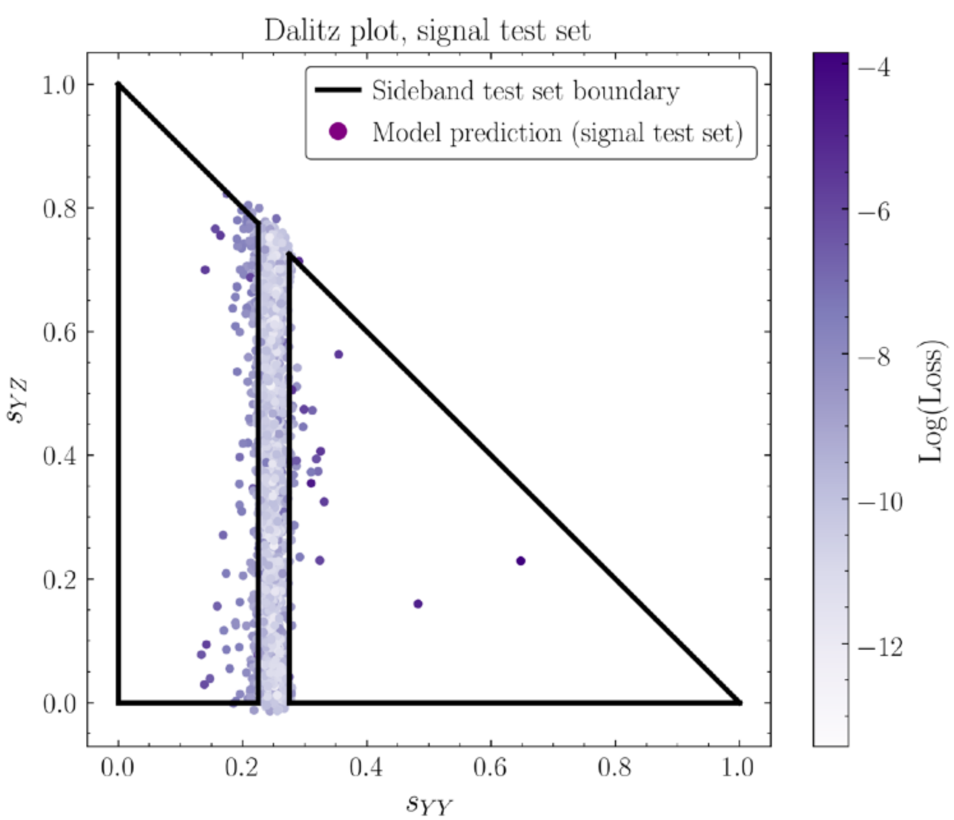
Pick a signal region, excise it from the training set, train an autoencoder

As before, latent map breaks at a random point, but close to the excised region (c.f. sphere minus equator)

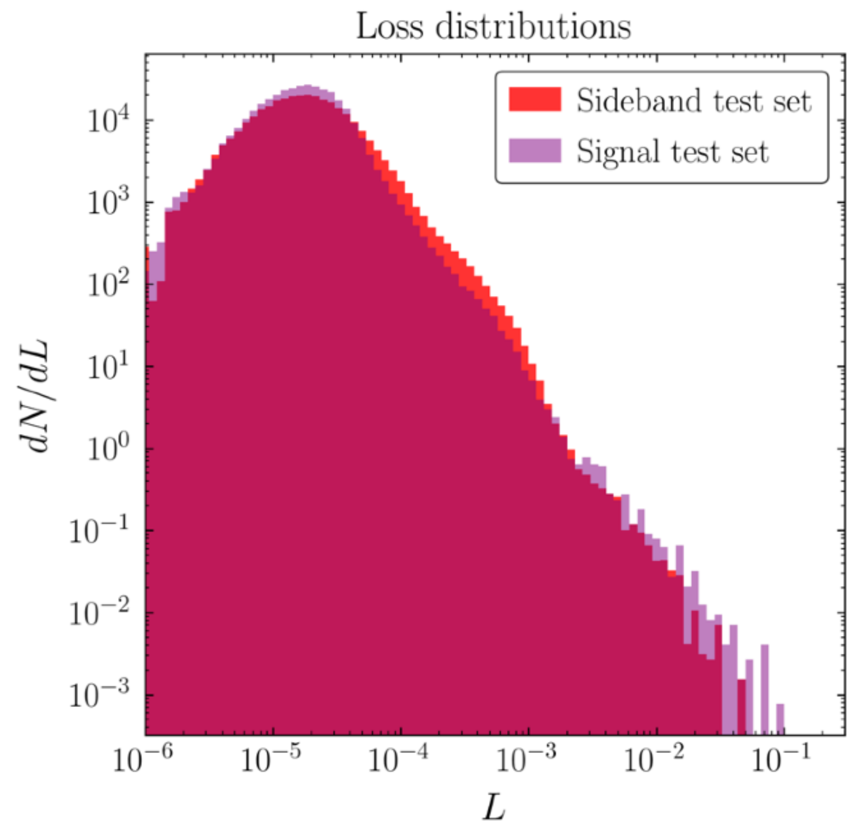
What happens if we feed pure signal into this sideband-only network?

Failure of a bump hunt

Let's use this phase space example as a cartoon of a bump hunt in leptons, where collider observables are 4-vectors



Network interpolates through excised region



Loss tails are indistinguishable

There are large-loss points which are not anomalies, and points absent from the training data are not flagged as anomalies: at most a single point in phase space will be identified as anomalous

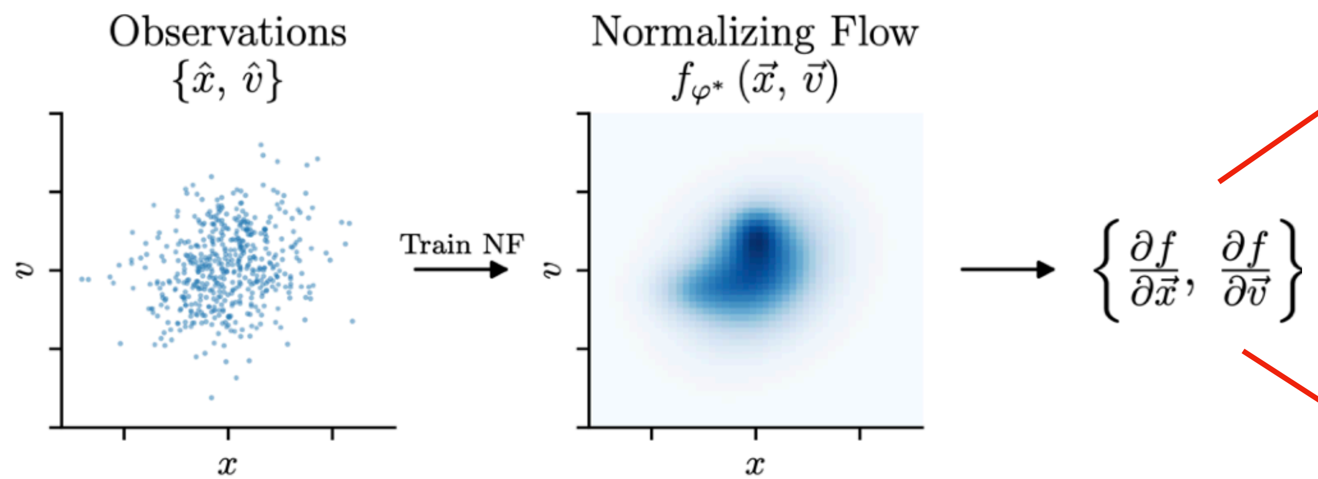
[Batson, Haaf, **YK,** Roberts, JHEP 2021]

Learning (stellar) phase space

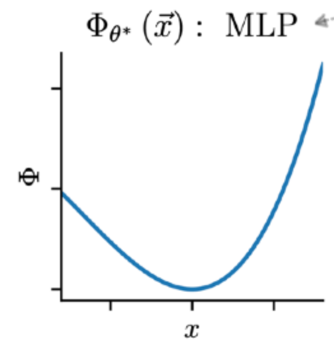
Given 6D phase space data for a bunch of stars, what's the potential?

Deep Potential: Recovering the gravitational potential from a snapshot of phase space

Gregory M. Green,[1] Yuan-Sen Ting (丁源森),[2,3] and Harshil Kamdar[4]



solve Boltzmann equation with a second MLP



solve Boltzmann equation analytically by matrix inversion

$$\mathcal{L}_{\mathrm{MSE}}(\vec{x}) = \frac{1}{N}\sum_{\beta=1}^{N}\left|\frac{df}{dt}\right|^2_{\vec{v}=\vec{v}^\beta}$$

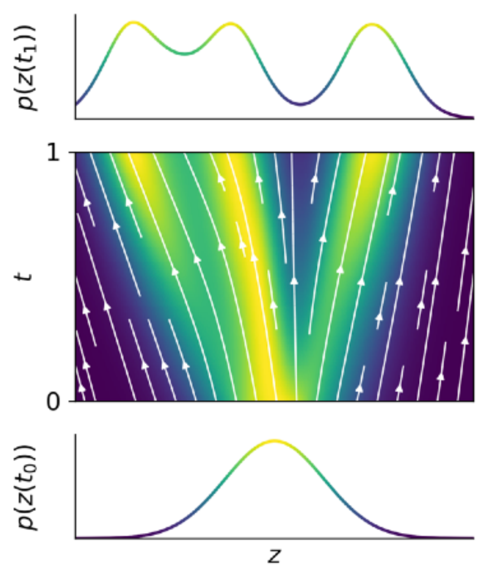
$$\frac{N}{2}\frac{\partial \mathcal{L}_{\mathrm{MSE}}}{\partial a_i} = \sum_{\beta=1}^{N}\left[\left(v_j\frac{\partial f}{\partial x_j} + a_j\frac{\partial f}{\partial v_j}\right)\frac{\partial f}{\partial v_i}\right]_{\vec{v}=\vec{v}^\beta}$$

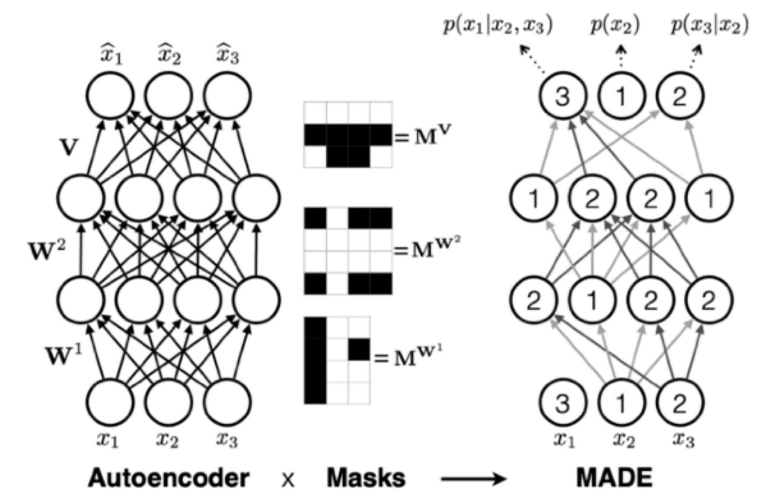
Measuring Galactic Dark Matter through Unsupervised Machine Learning

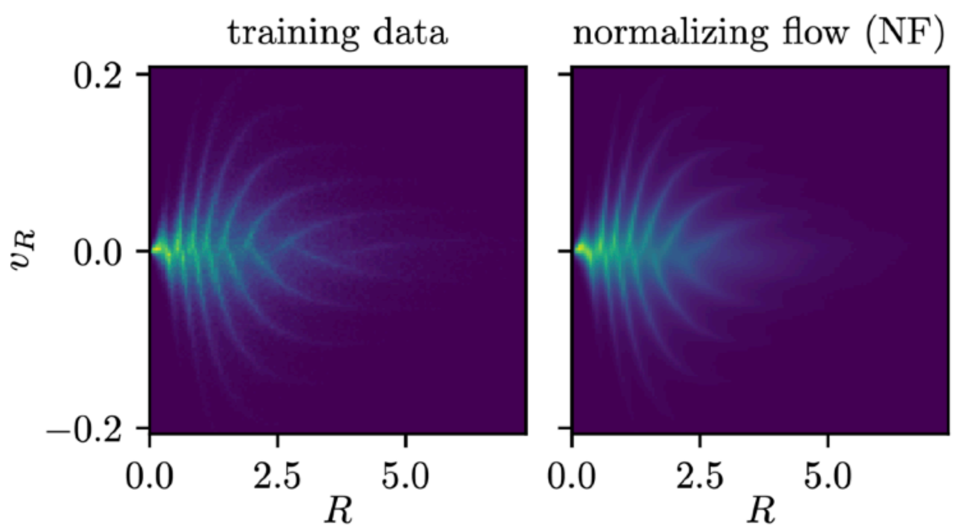
Matthew R. Buckley, Sung Hak Lim, Eric Putney, and David Shih
Department of Physics and Astronomy, Rutgers University, Piscataway, NJ 08854, USA

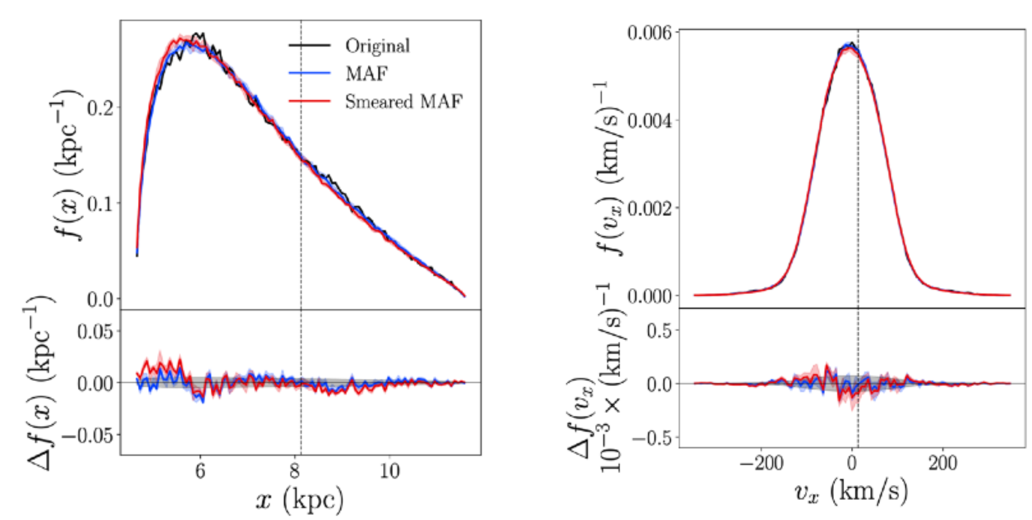
Normalizing flows

Main tool in both approaches is normalizing flow network: learn Jacobian which transforms easy-to-sample distribution into your target distribution









Both approaches seem to work pretty well…

[Green, Ting, Kamdar, NeurIPS 2020 and ApJ 2023; Buckley, Lim, Putney, Shih, arXiv:2205.01129; Grathwohl, Chen, Bettencourt, Sutskever, Duvenaud, ICLR 2019; Papamakarios, Pavlakou, Murray, NIPS 2017]

Let’s try something easier…

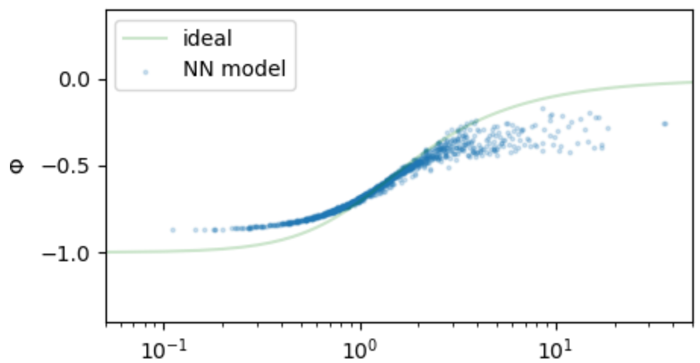
“Plummer sphere”

$$f(\vec{r},\vec{v}) \propto \begin{cases} \left[-E(\vec{r},\vec{v})\right]^{7/2}, & E<0 \\ 0, & E \geq 0 \end{cases}, \qquad \rho(r) = \frac{3}{4\pi}\left(1+r^2\right)^{-5/2}, \quad \Phi(r) = -\left(1+r^2\right)^{-1/2}$$

$$\text{where } E = \frac{1}{2}v^2 + \Phi(r)$$

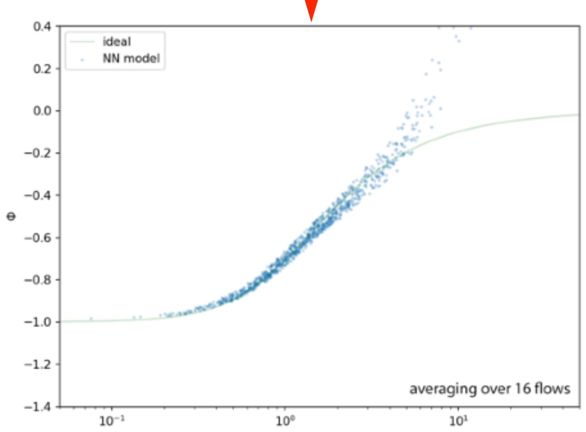
FFJORD masking

one NF

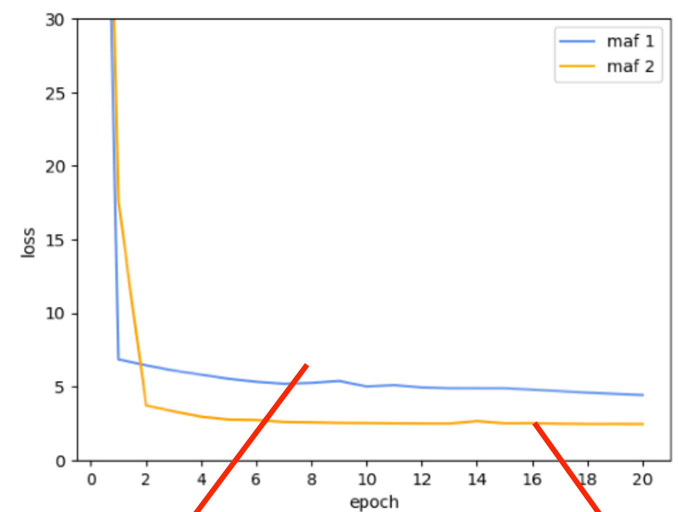


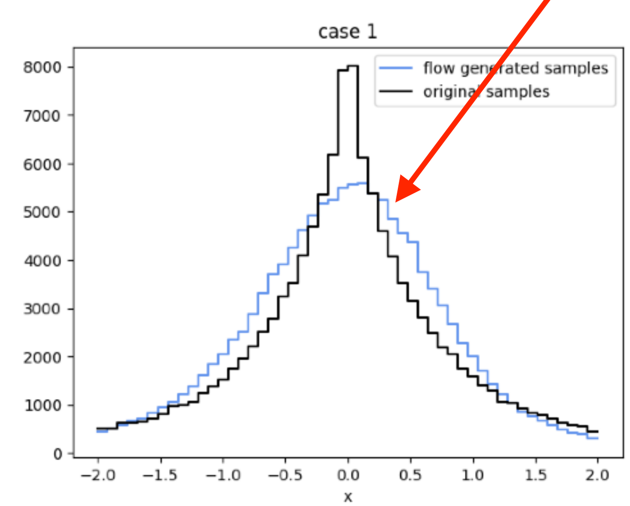
average

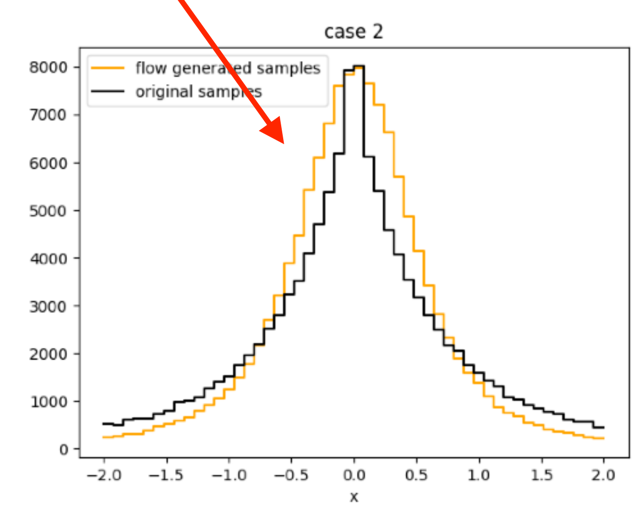
16 NF’s



MADE masking







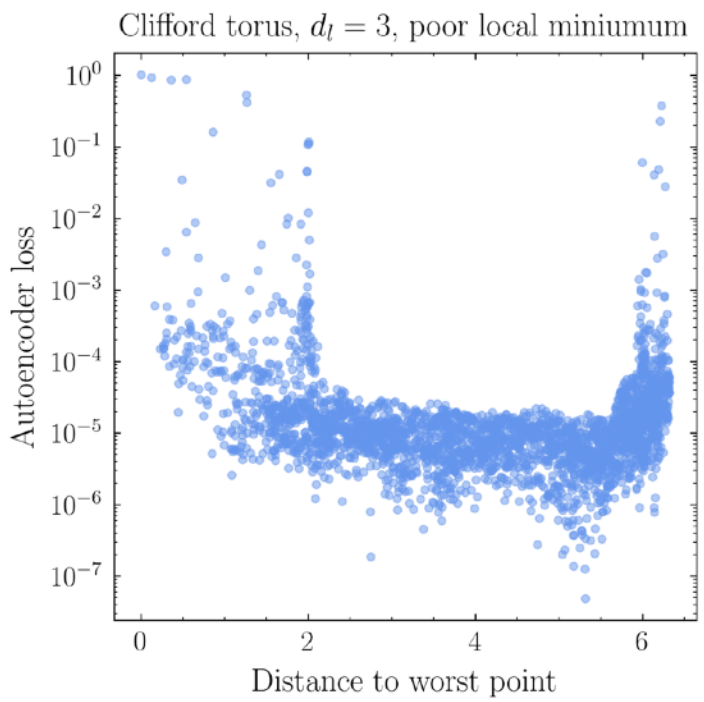
Foiled by ensemble variance

Ensemble variance: good and bad local minima

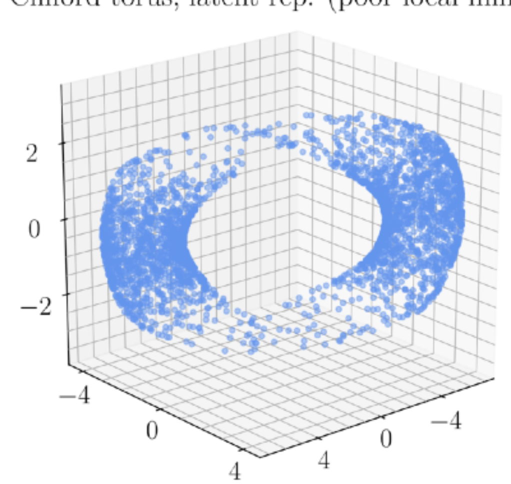
Consider the "Clifford torus" in $\mathbb{R}^4$:

$$(x, y, z, w) = (\cos\theta, \sin\theta, \cos\phi, \sin\phi)$$

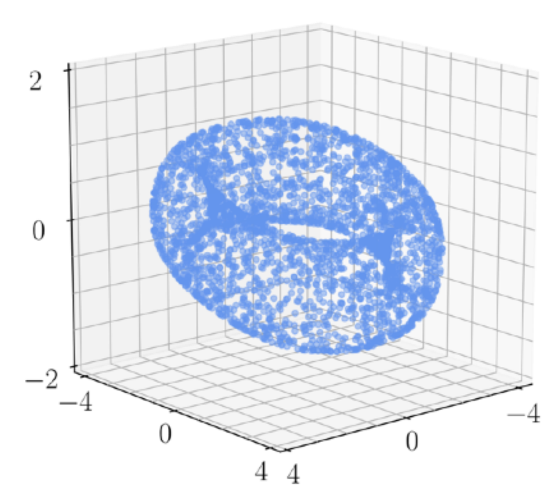
Can an autoencoder learn the correct "donut" embedding in $\mathbb{R}^3$?



Most common latent representation



The right answer (rare)



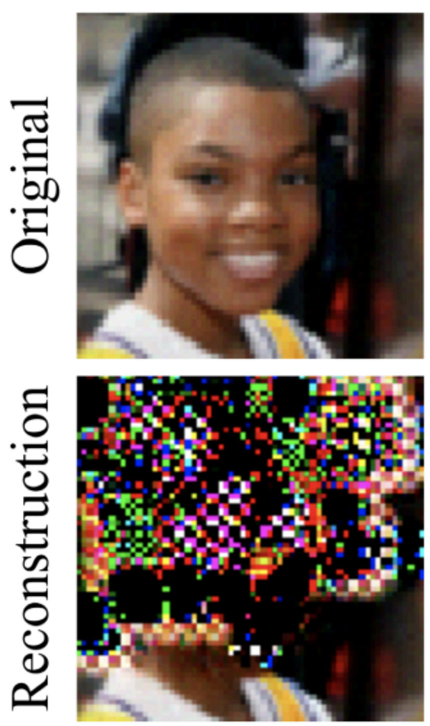
Embedding exists, but network only rarely finds it!

Output variance in invertible NNs

Normalizing flows must be invertible, otherwise Jacobian is singular

Understanding and Mitigating Exploding Inverses in Invertible Neural Networks

Jens Behrmann[*1] **Paul Vicol**[*2,3] **Kuan-Chieh Wang**[*2,3] **Roger Grosse**[2,3] **Jörn-Henrik Jacobsen**[2,3]
[1]University of Bremen [2]University of Toronto [3]Vector Institute * Equal contribution

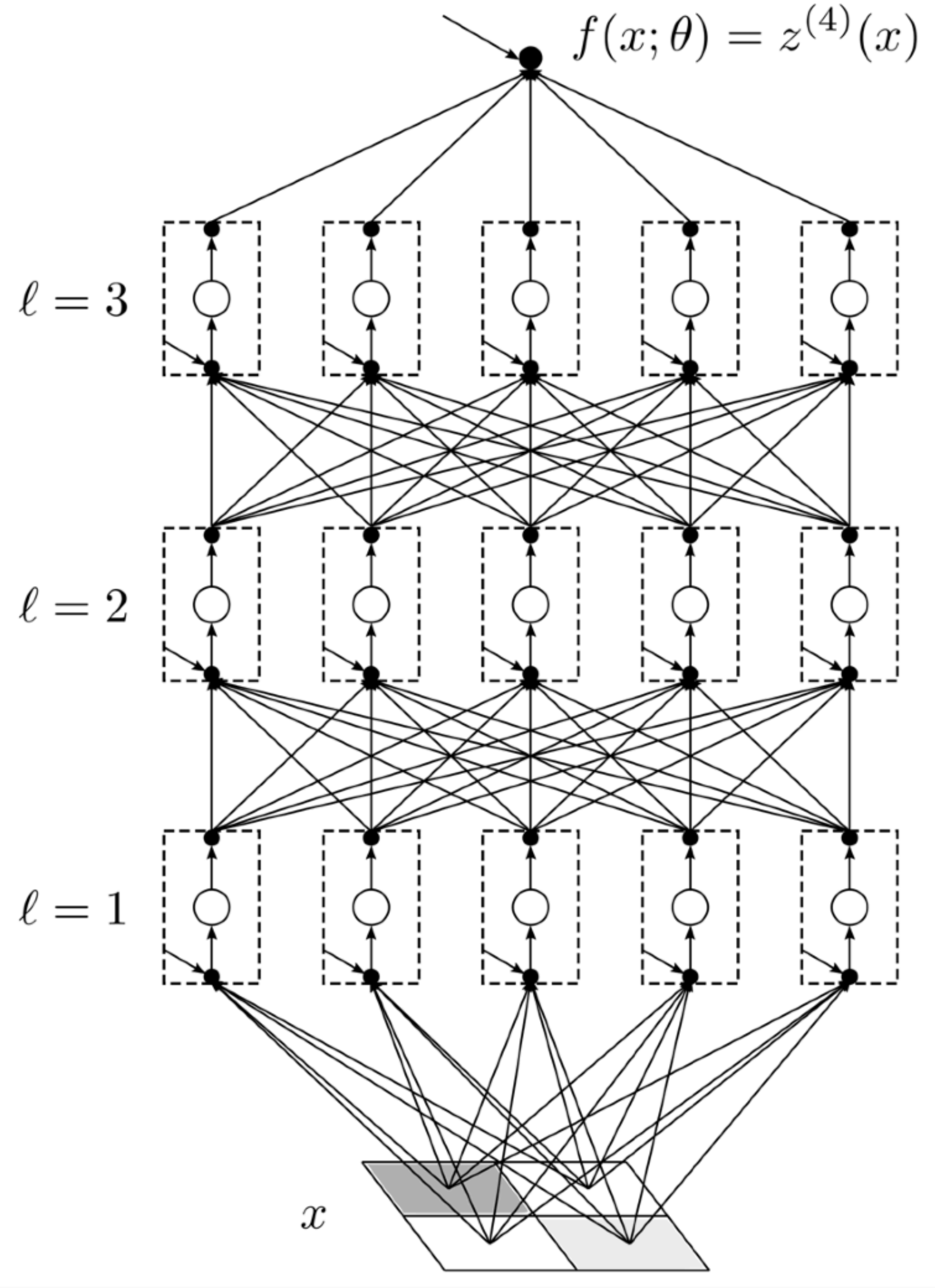


Analytically invertible network may be numerically non-invertible because of large ensemble variance

(this is probably not the source of the aforementioned variance, but still a concern)

[Berhmann, Vicol, Wang, Grosse, Jacobsen, AISTATS 2021]

Back to vanilla MLPs



Notation refresher:

$$z_i^{(\ell+1)}(x_\alpha) \equiv b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma\Big(z_j^{(\ell)}(x_\alpha)\Big)$$

To avoid index proliferation, call all parameters θ

Consider ensemble of networks w/Gaussian parameters, train with ordinary G.D.

 [Roberts, Yaida, Hanin, Cambridge University Press 2022]

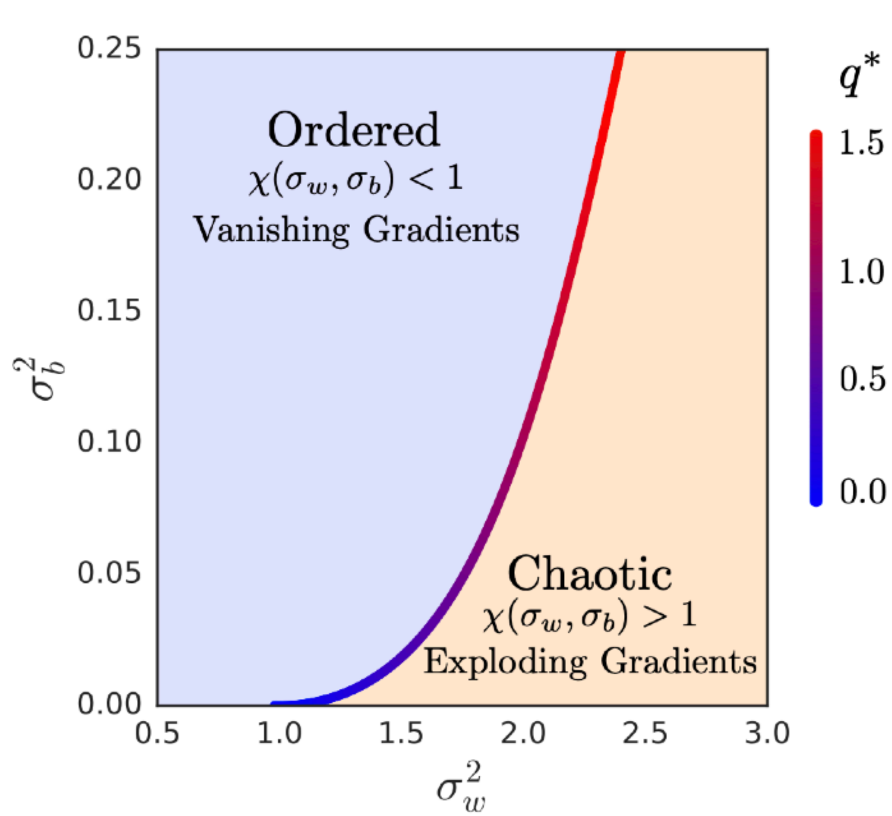
Tuning to criticality

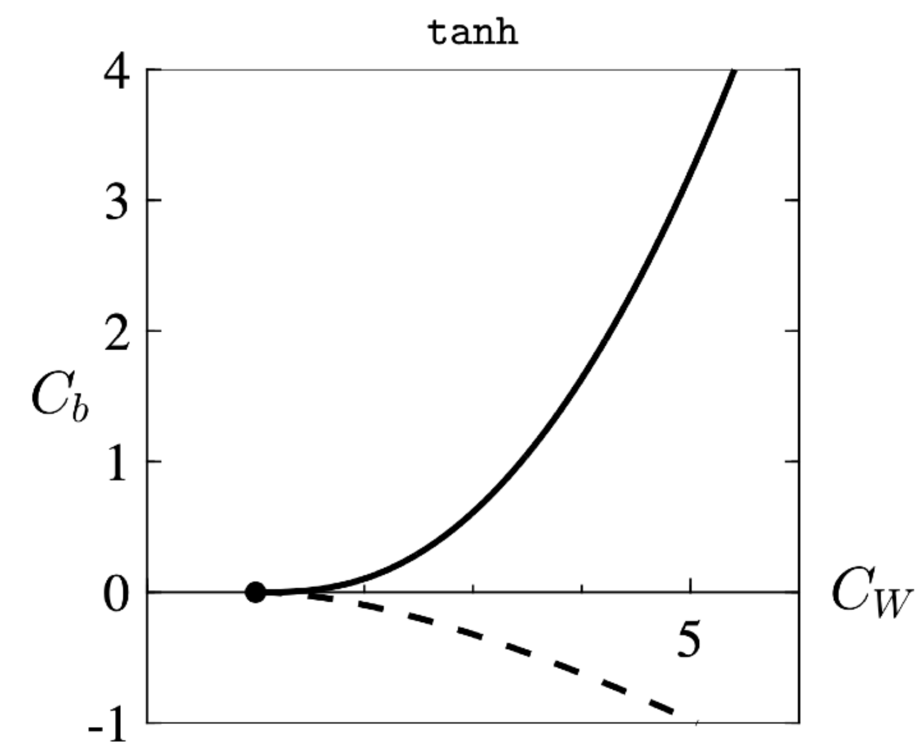
$$\mathbb{E}\left[\frac{dz^{(\ell)}}{d\theta}\frac{dz^{(\ell)}}{d\theta}\right] \sim \Theta^{(\ell)} + \mathcal{O}\left(\frac{1}{n}\right)$$

controls gradient descent dynamics

$$\mathbb{E}[z^{(\ell)}z^{(\ell)}] \sim K^{(\ell)} + \mathcal{O}\left(\frac{1}{n}\right)$$

controls typical output norm





For some activations, can tune initialization distribution so that neither blows up exponentially as a function of depth

[Pennington, Schoenholtz, Ganguli, NIPS 2017; Roberts, Yaida, Hanin, Cambridge University Press 2022]

Fluctuations and feature learning

Roberts/Yaida/Hanin: all leading non-Gaussianities scale as ℓ/n .

Some are good: $z \to -\eta\epsilon\left(\frac{dz}{d\theta}\right)^2 + \frac{\eta^2}{2}\epsilon^2\left(\frac{d^2z}{d\theta^2}\left(\frac{dz}{d\theta}\right)^2\right)$

$\mathbb{E}\left[\left(\frac{dz}{d\theta}\right)^2\right] = \Theta$ fixed at init, but $\frac{\mathbb{E}\left[z\frac{d^2z}{d\theta^2}\left(\frac{dz}{d\theta}\right)^2\right]}{\Theta^2} \propto \frac{\ell}{n}$

representation learning at finite width

Some just add noise:

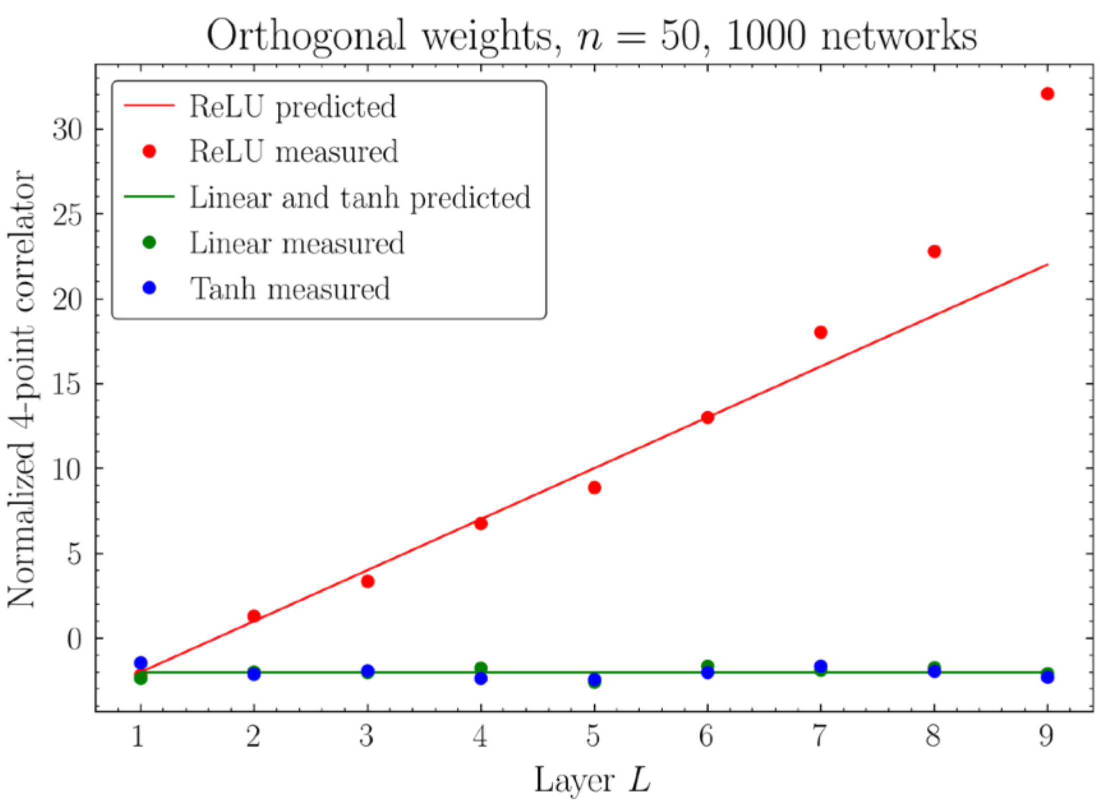
$$\frac{\mathbb{E}[z^4]_{\text{conn.}}}{K^2} \propto \frac{\ell}{n}$$

ensemble variance grows with depth

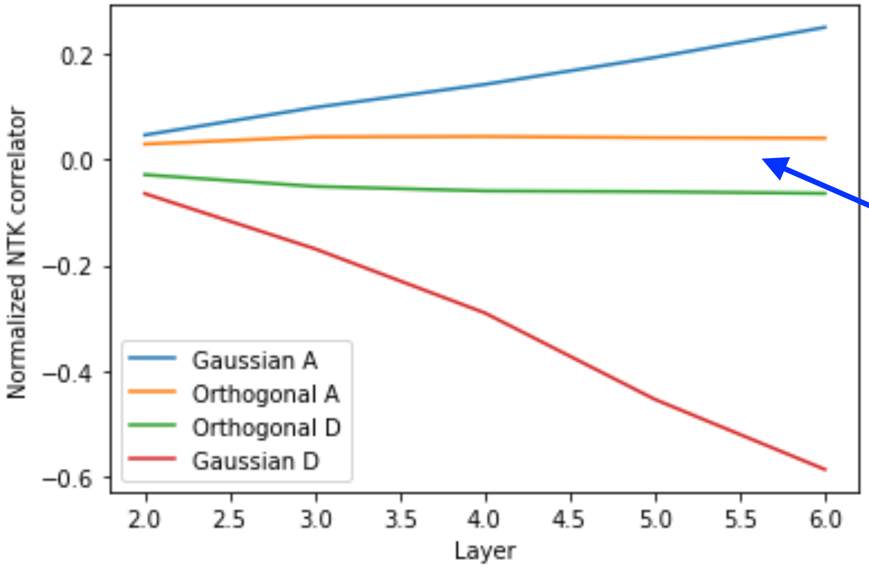
[Roberts, Yaida, Hanin, Cambridge University Press 2022]

Orthogonal initializations

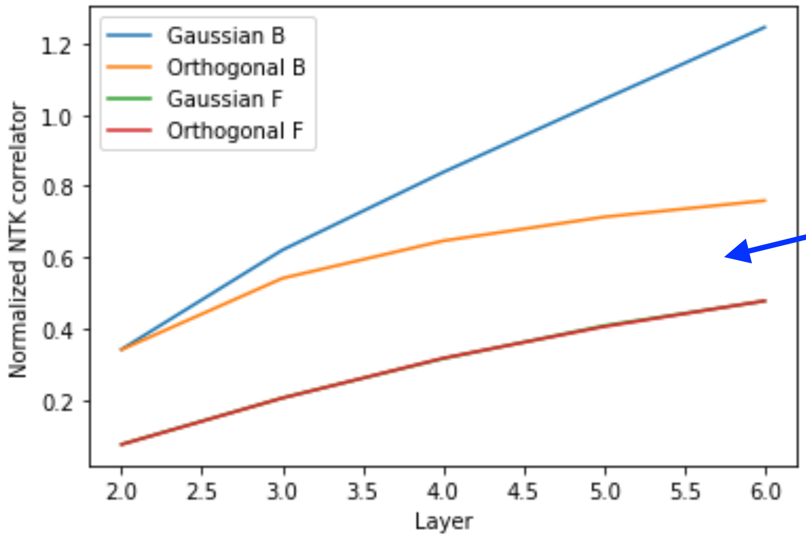
If instead we draw weights from Haar-distributed orthogonal matrices:



for certain activations, variance is depth-independent



"bad" NTK correlators are also depth-independent…



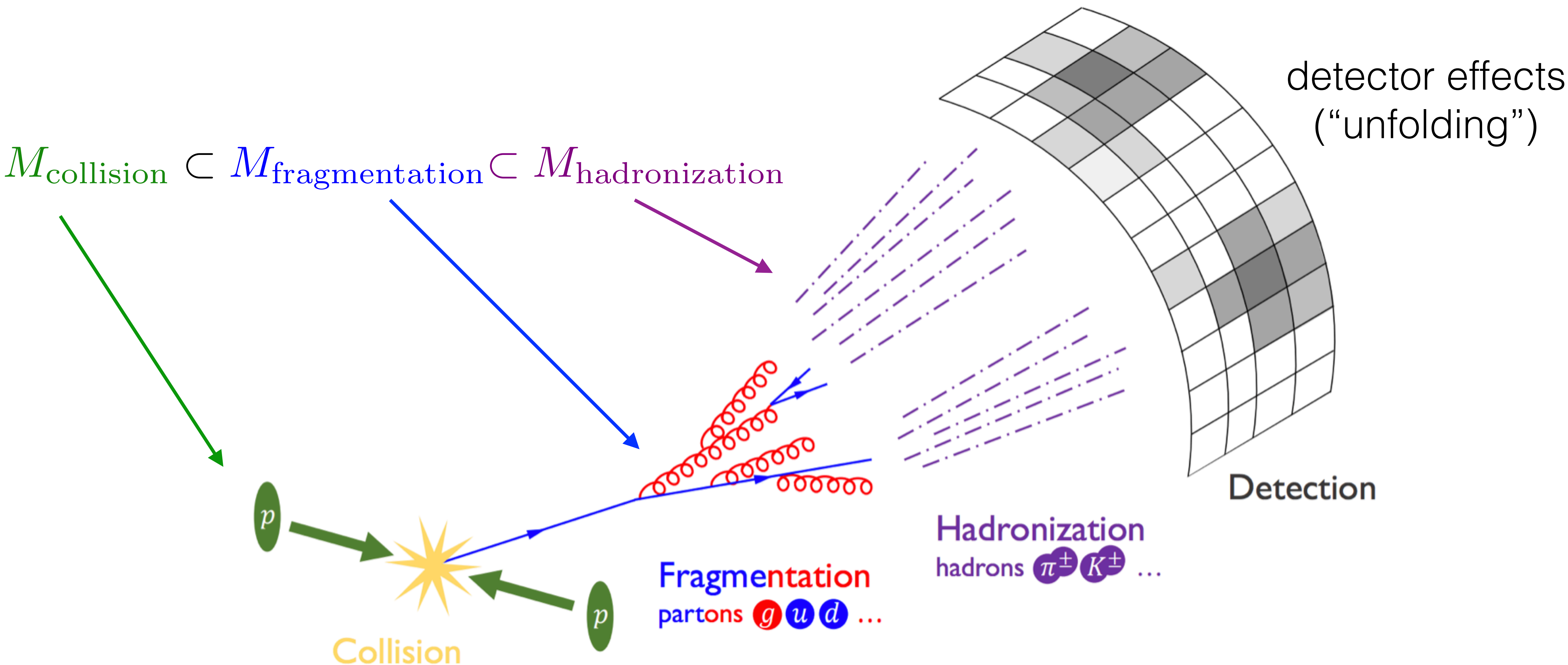
but "good" NTK correlators grow linearly!

Might this be a better init to keep variance under control?

[Day, **YK**, Roberts, to appear;
Pennington, Schoenholtz, Ganguli, NIPS 2017]

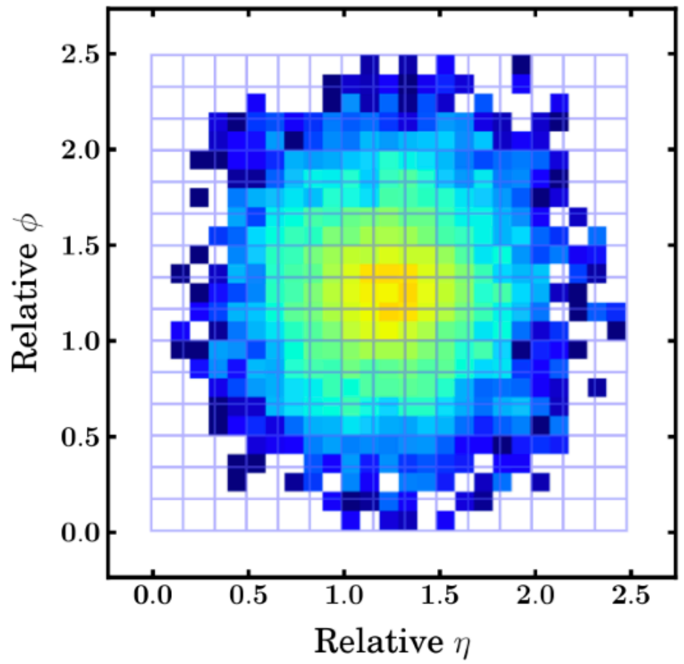
Data dimension at colliders

N particles live on a phase space manifold of dimension 3N-4.
But hadronization makes more particles, so dimension changes



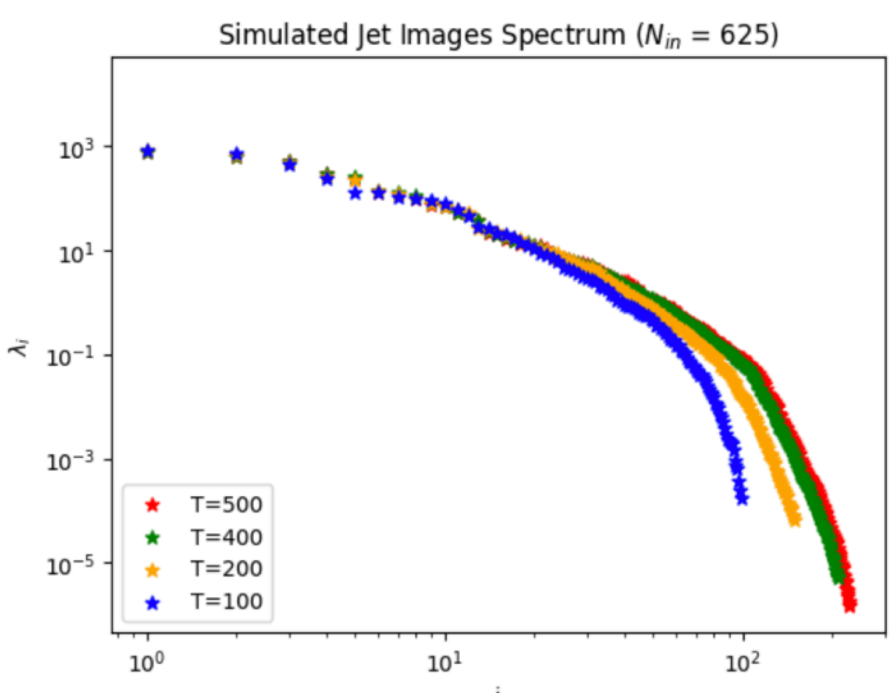
Data dimension depends on energy scale!

Jet images vs. cat images

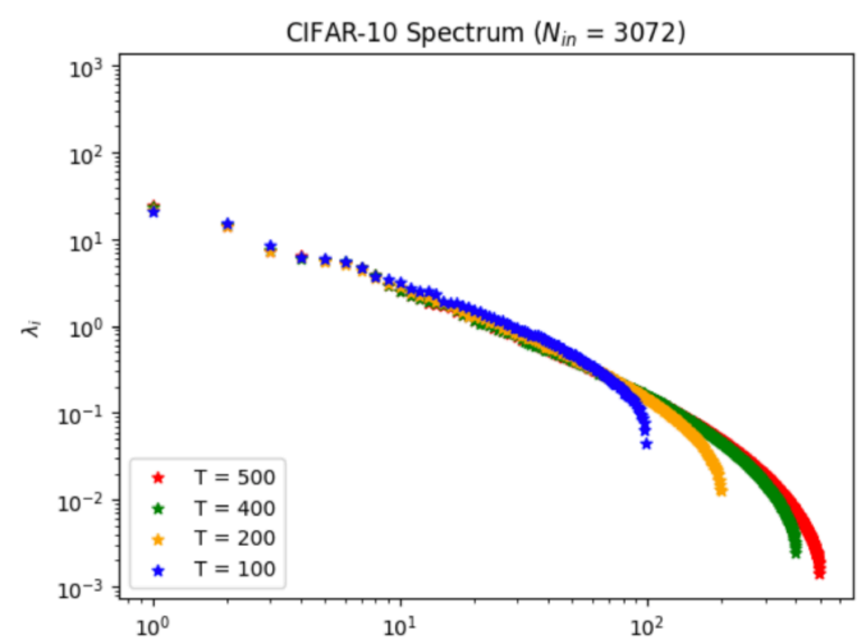


$$\frac{1}{T}\sum_{\alpha=1}^{T} x_{i;\alpha}x_{j;\alpha}$$

cat



power laws
in data
covariance
eigenvalues:
important
features
at all scales

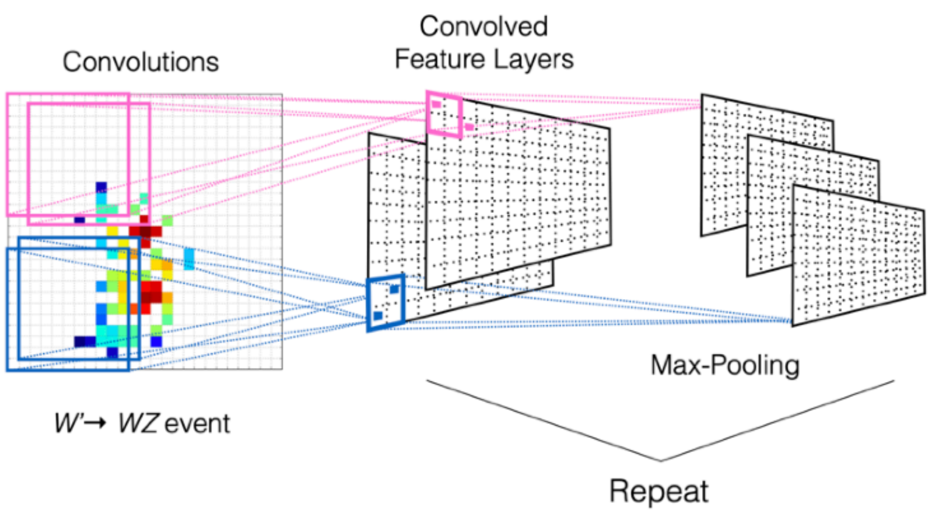


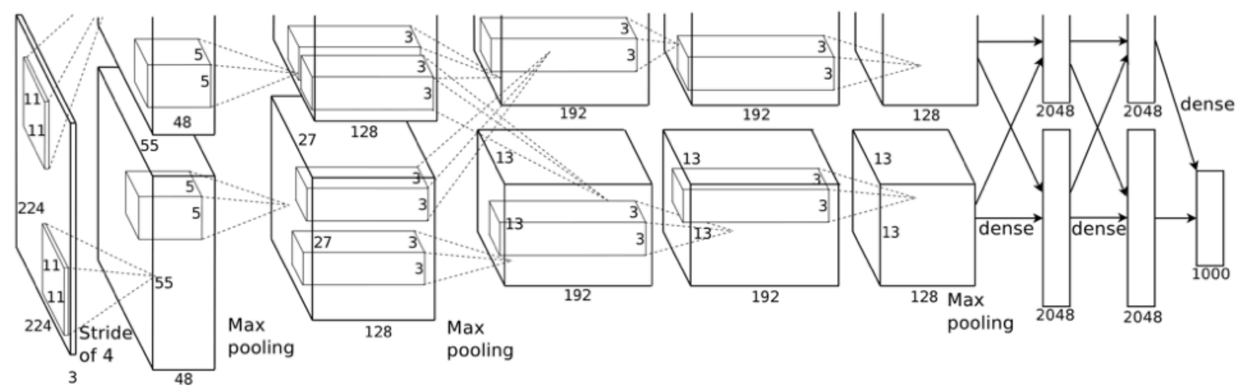
[Kogan, Kagan, Strauss, Schwartzman, JHEP 2015; CIFAR-10 dataset;
Maloney, Roberts, Sully, 2210.16859; Batson, **YK**, Mande, Roberts, to appear]

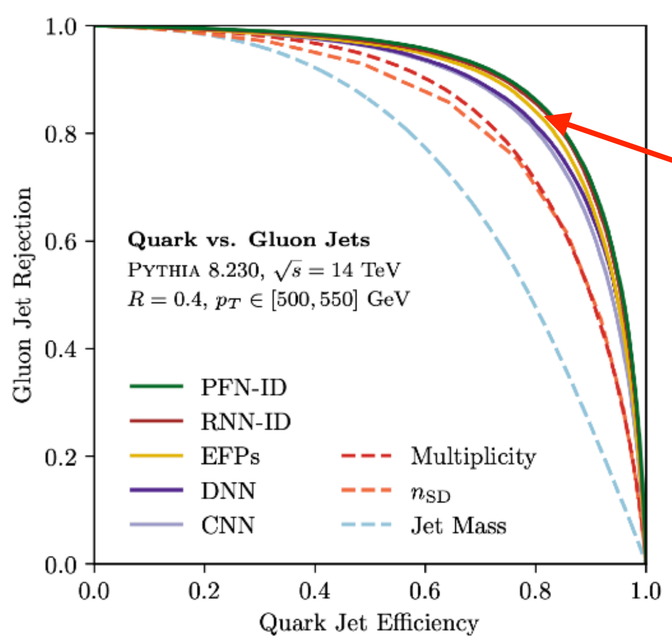
Jet images vs. cat images

Quark vs. gluon
jet classification task:

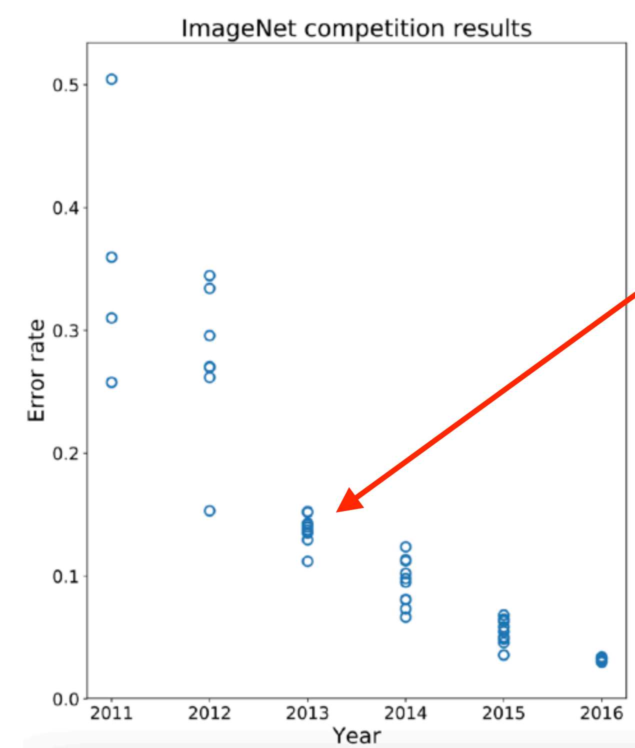
ImageNet image
classification task:







linear
regression
(on engineered
features)
does better!?

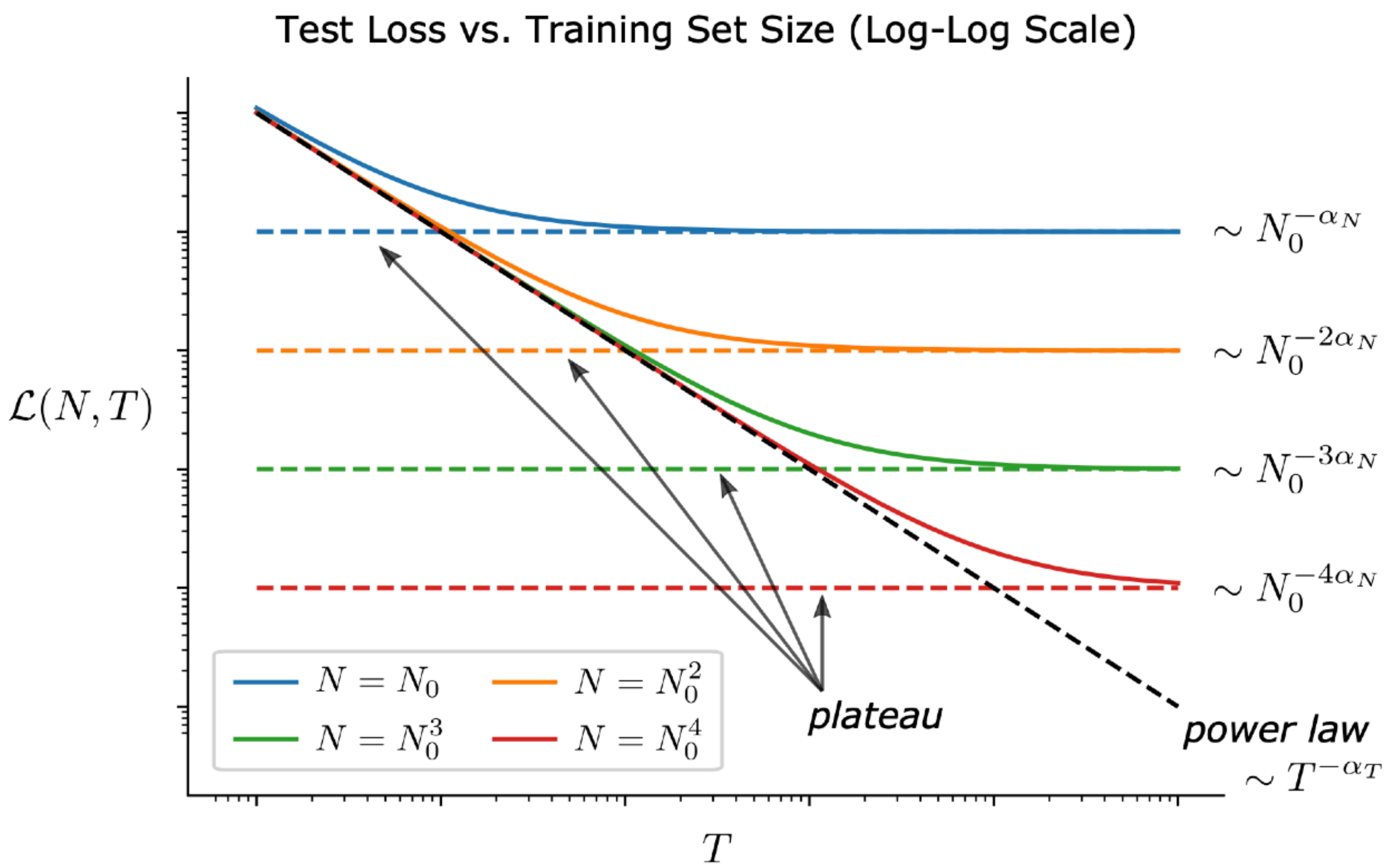


ALL
state-of-
the-art
networks
now use
CNNs

[de Oliveira et al., JHEP 2016; Komiske, Metodiev, Thaler, JHEP 2019;
Krizhevsky, Sustkever, Hinton, NeurIPS 2012]

Data dimension in “big data”

Empirical observations from large language models:

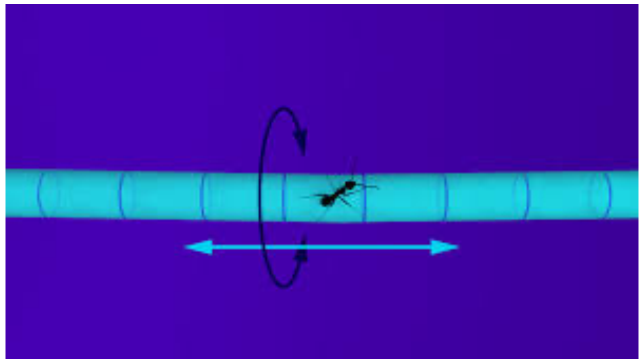


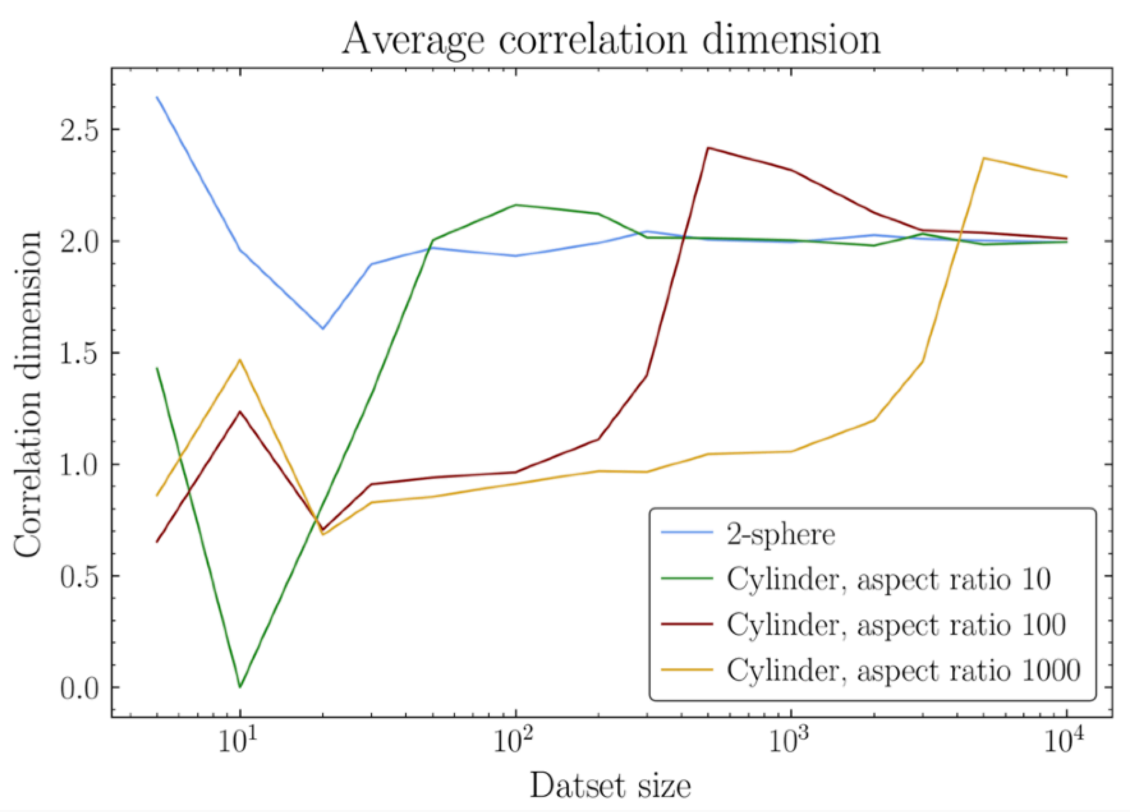
For some tasks, loss slope and data covariance slope can measure data dimension, but only if data manifold is roughly isotropic: nearest neighbors in all directions

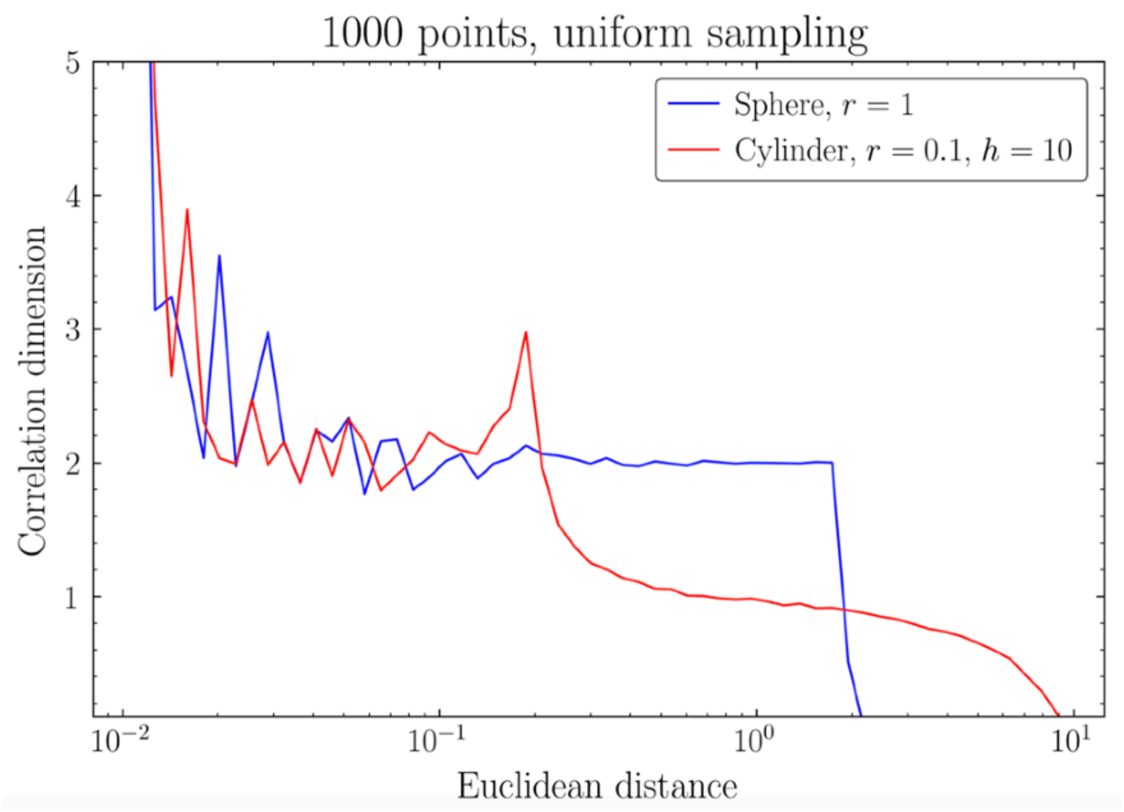
[Kaplan, McCandlish, et al., 2001.08361; Sharma and Kaplan, 2004.10802; Bahri, Dyer, Kaplan, Lee, Sharma, 2102.06701; Maloney, Roberts, Sully, 2210.16859]

Data dimension and isotropy

vs.

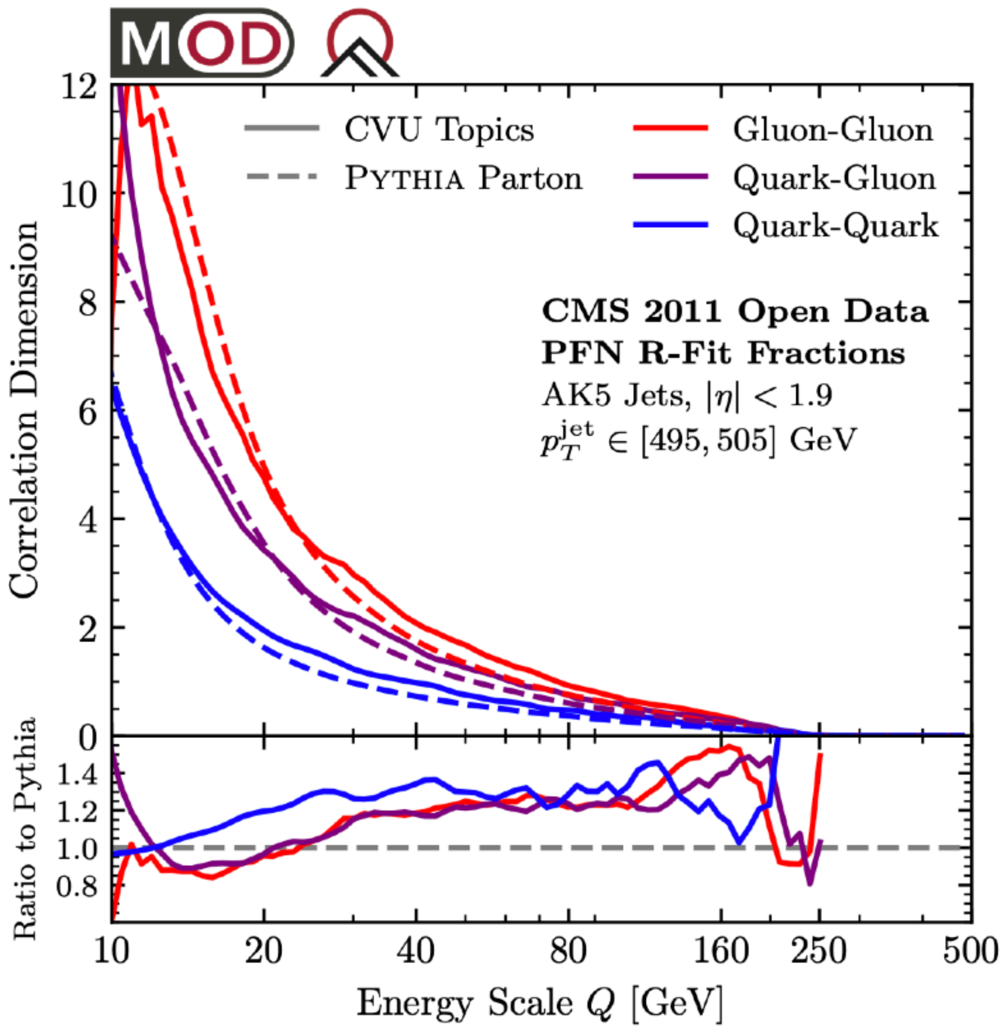






Correlation dimension: $\dim(Q) = \dfrac{\partial}{\partial \log Q} \ln \sum_{i<j} \Theta(d_{ij} < Q)$

Data dimension for jet classification

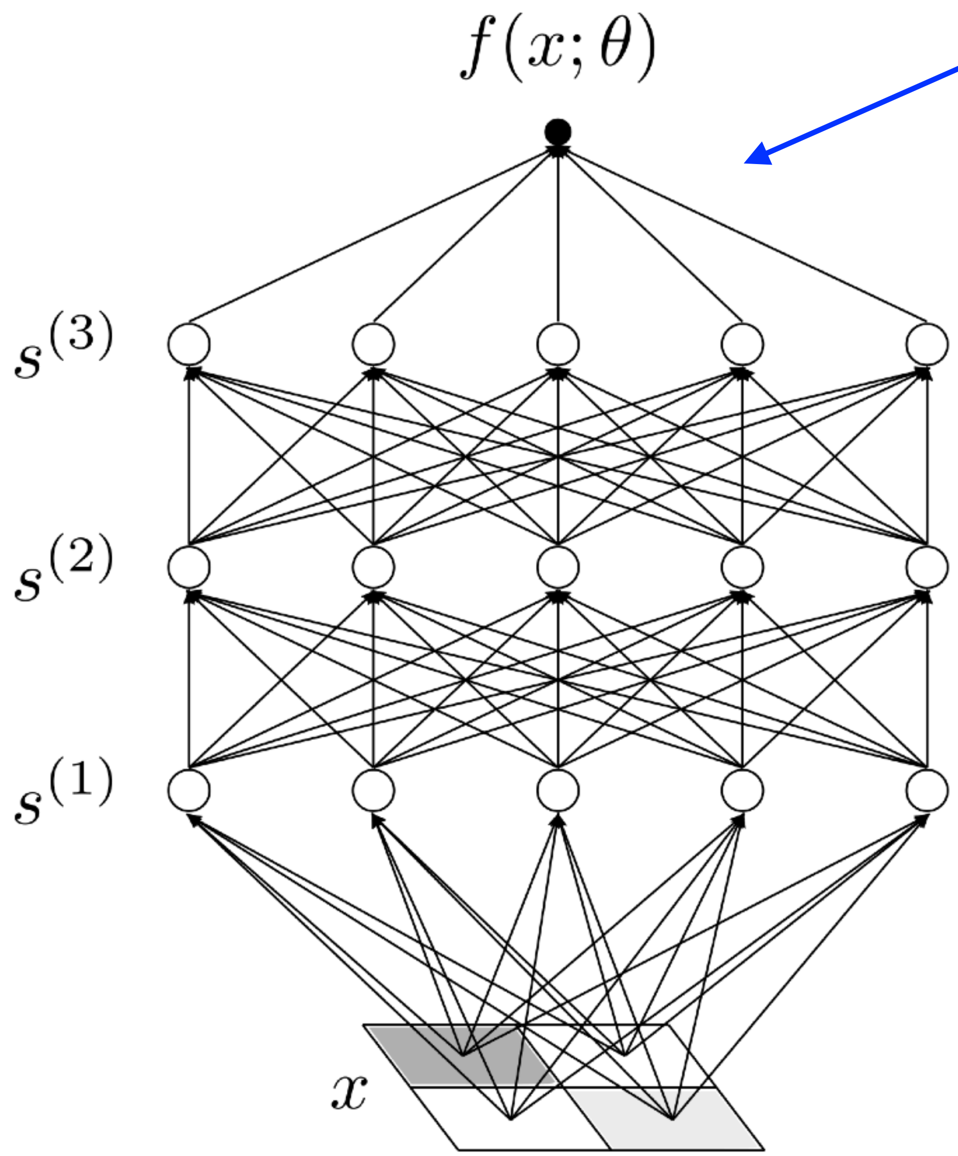


- Does underlying geometry govern scaling law in the loss?
- For a fixed training set size, what is the best classifier?
- What is the best way to jointly scale model and training data?
- Is “dumb machine learning” actually most efficient for quark/gluon discrimination?

[Komiske, Kryhin, Thaler, PRD 2022; Batson, **YK**, Mande, Roberts, to appear]

Outlook

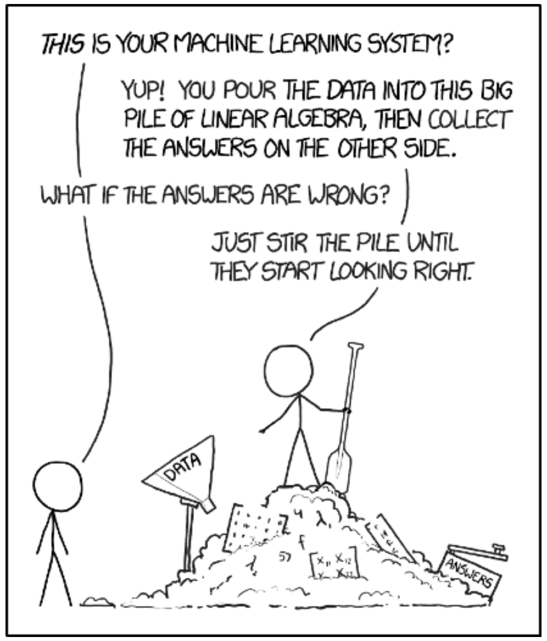
In this talk I have asked many questions, and answered few.



We can get a lot of mileage out of this simple architecture:

- Stellar stream-finders
- Jet tagging and classification
- Particle ID
- …

Physicists have the tools to understand this structure and use it reliably. Let's do so!

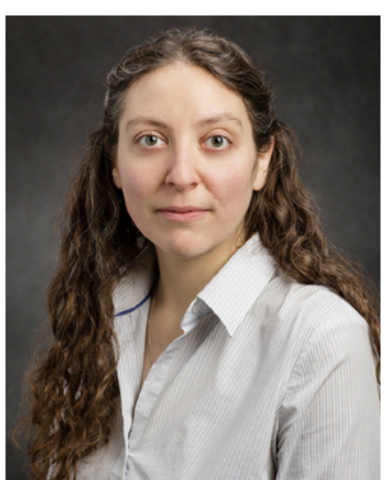


Thank you!

Victoria Tiki (grad, UIUC)

Hannah Day (grad, UIUC)

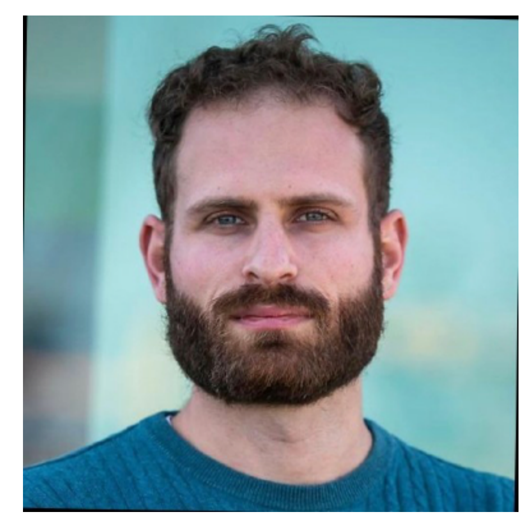
Aarav Mande
(undergrad, UIUC)



Jessie Shelton (UIUC)

Gil Holder (UIUC)

Dan Roberts (IAIFI)



Joshua Batson

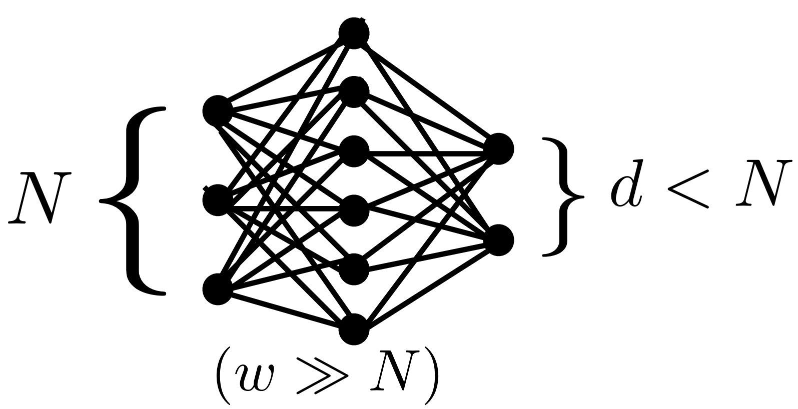
Backup

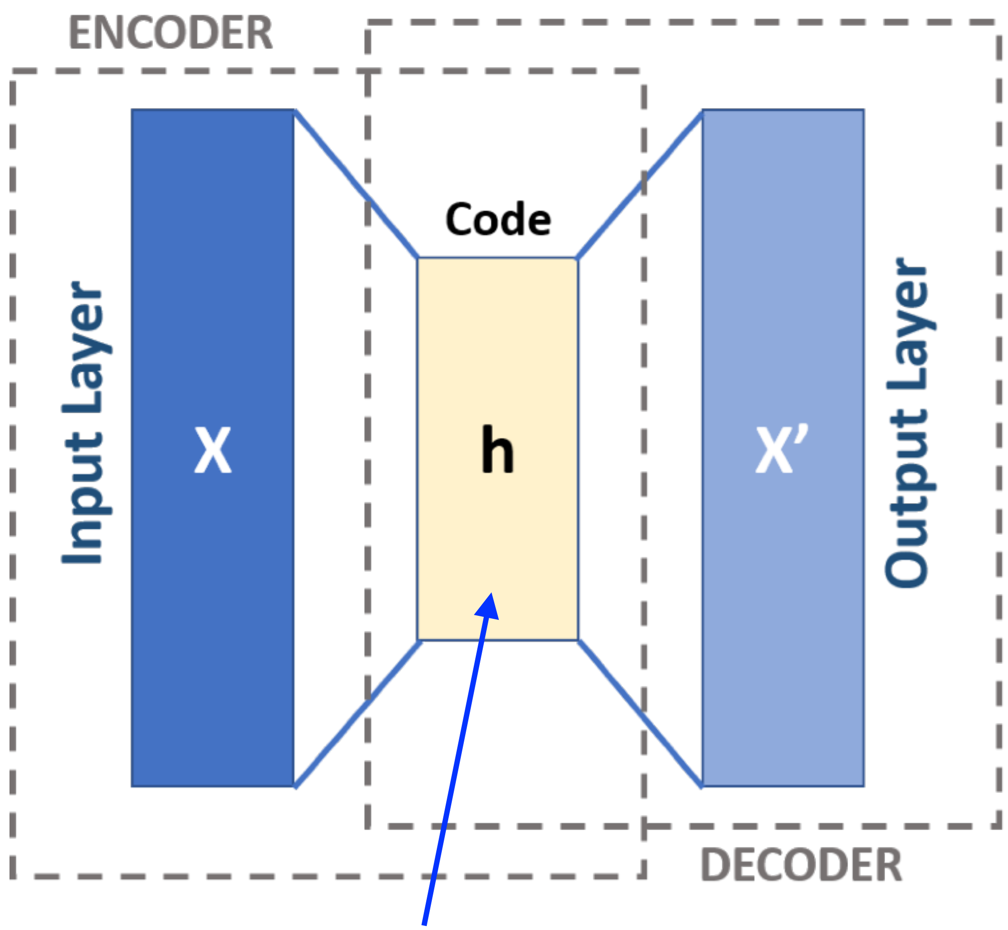
Autoencoder architecture

Encoder:

$f^{\text{enc}} : \mathbb{R}^N \mapsto \mathbb{R}^d$

1- or 2-hidden-layer fully-connected net, tanh activations





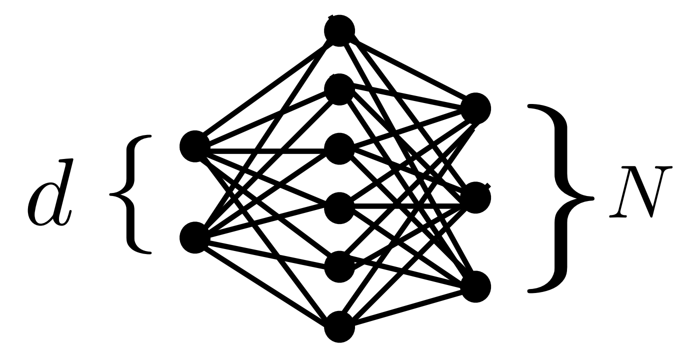
d-dimensional latent representation

$f^{\text{enc}}(\mathbf{x})$

Decoder:

$f^{\text{dec}} : \mathbb{R}^d \mapsto \mathbb{R}^N$

same architecture as encoder but no nonlinearity on output



Loss function $||f^{\text{dec}}(f^{\text{enc}}(\mathbf{x})) - \mathbf{x}||^2$, train with stochastic gradient descent