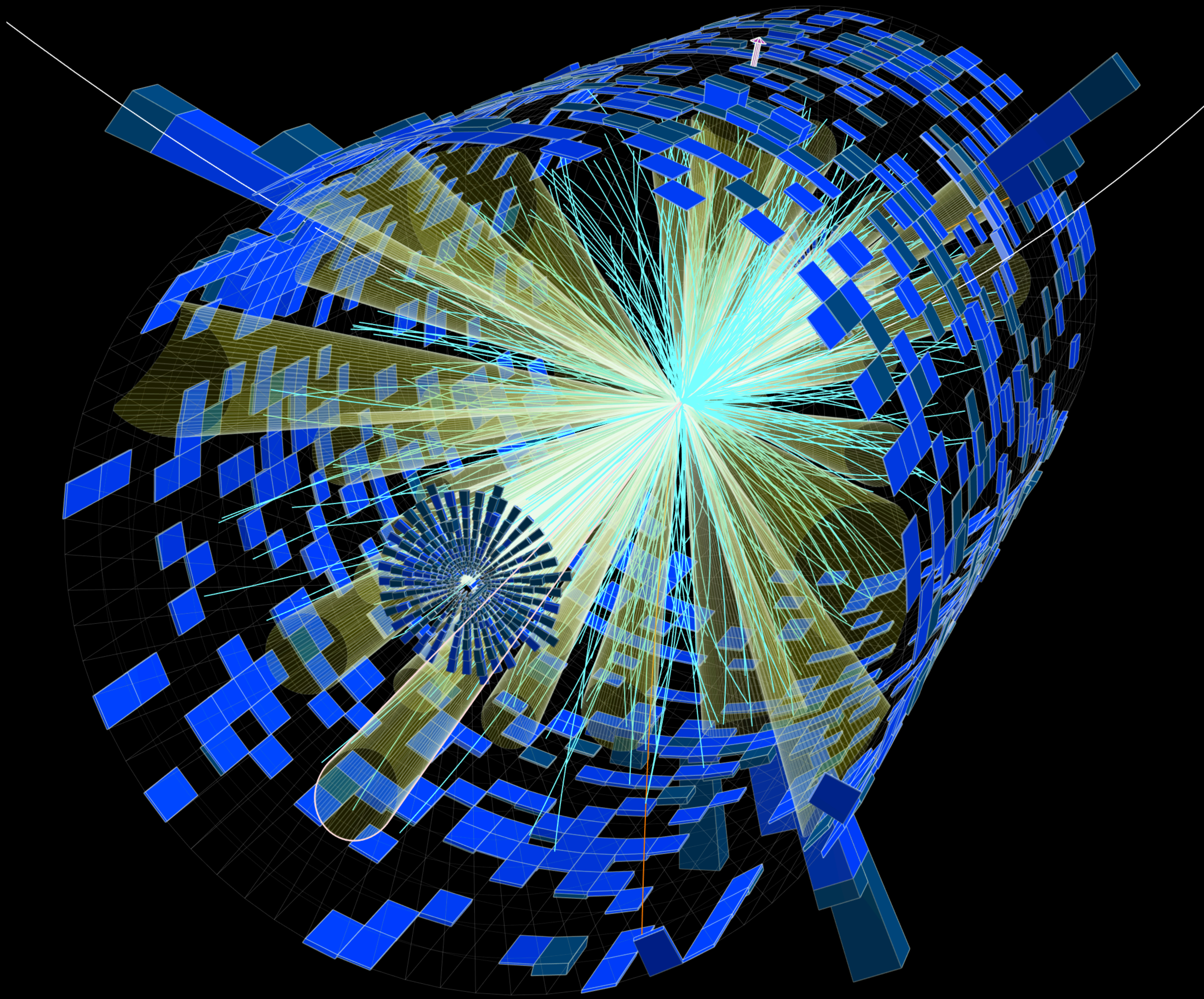
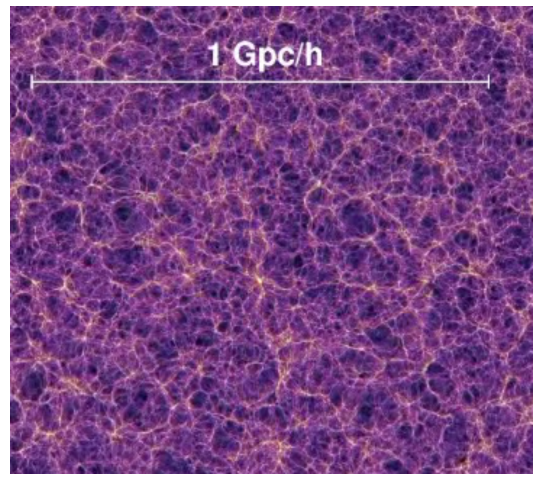
CONNECTIONS AND CROSS POLLINATION

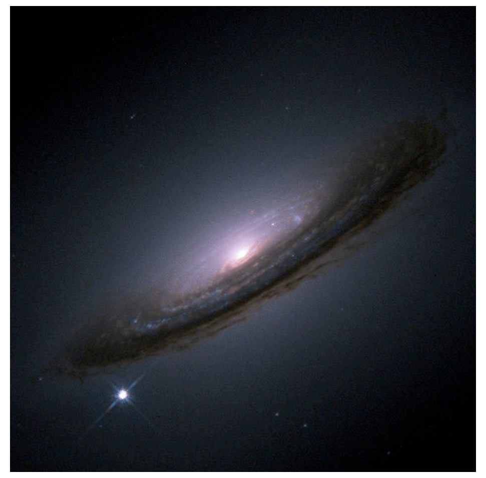
FROM QUARKS TO THE COSMOS

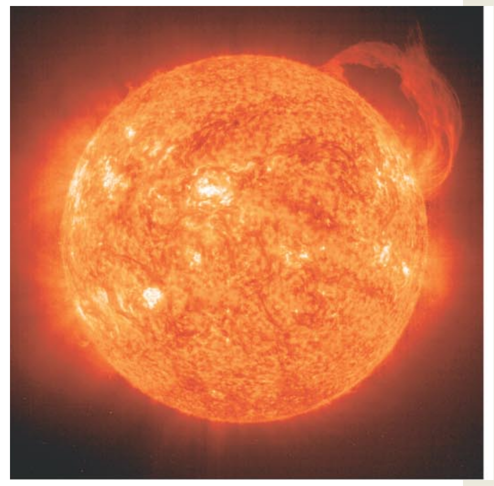


@KyleCranmer
University of Wisconsin-Madison
American Family Insurance Data Science Institute

Fundamental Particles & Interactions







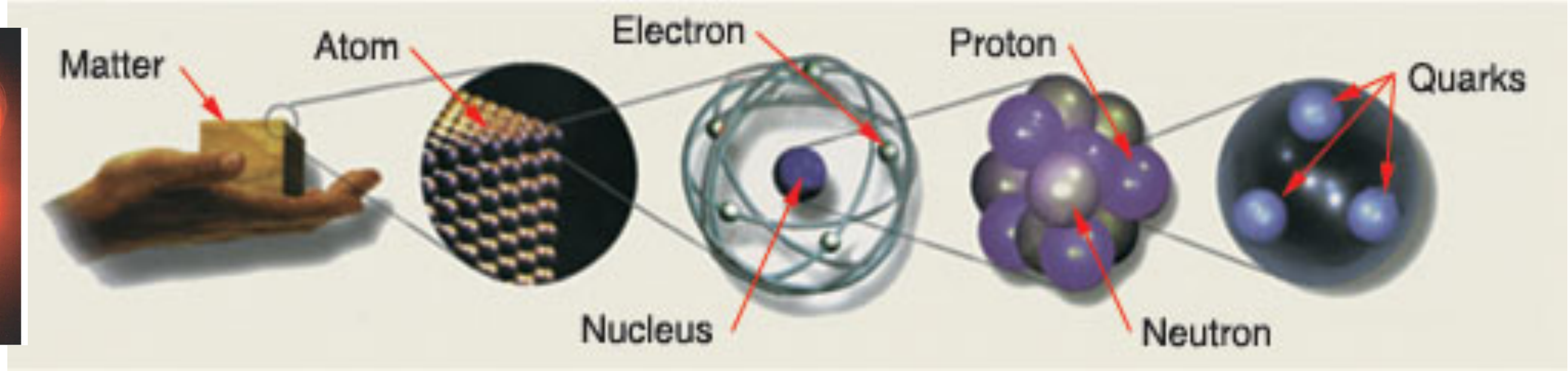
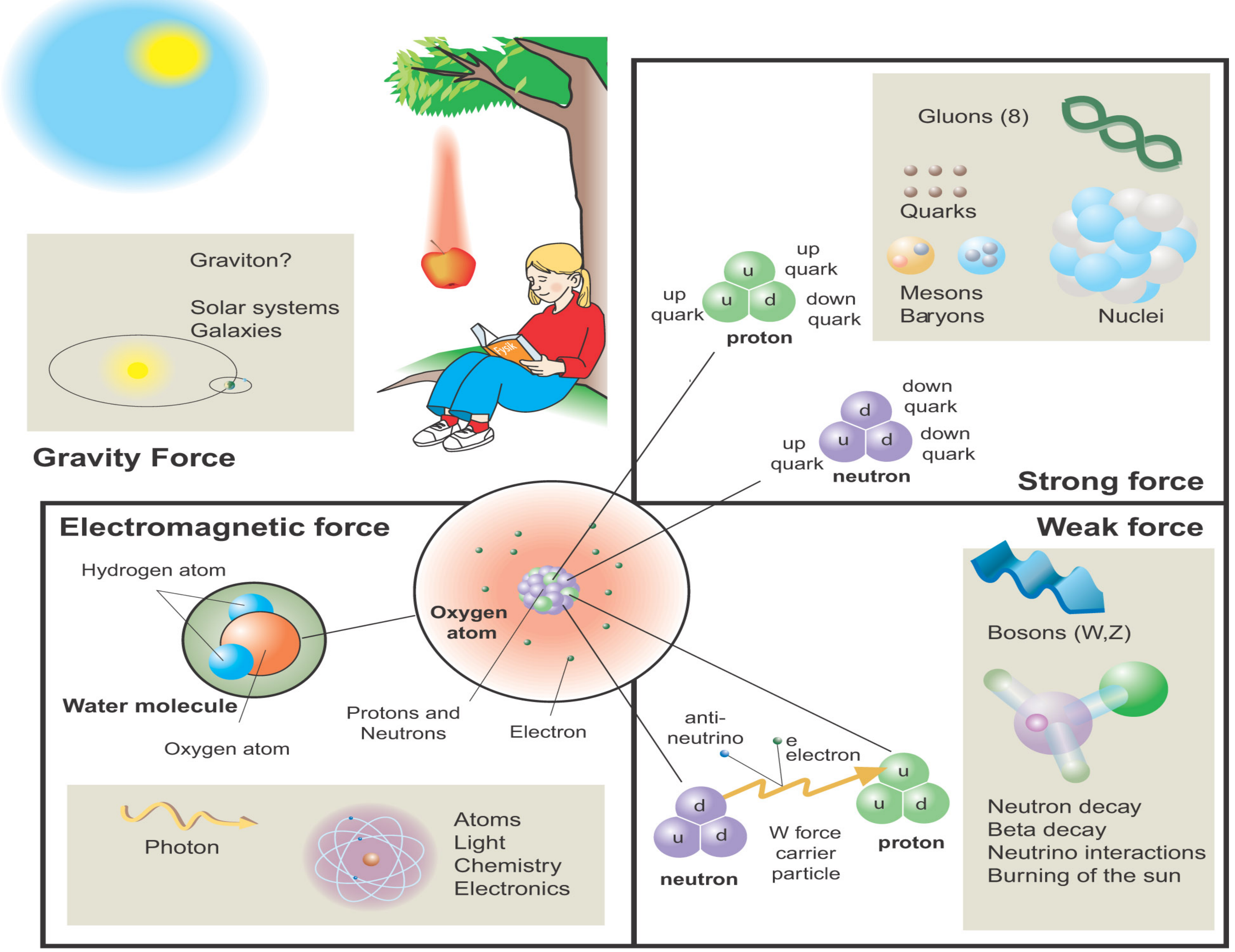
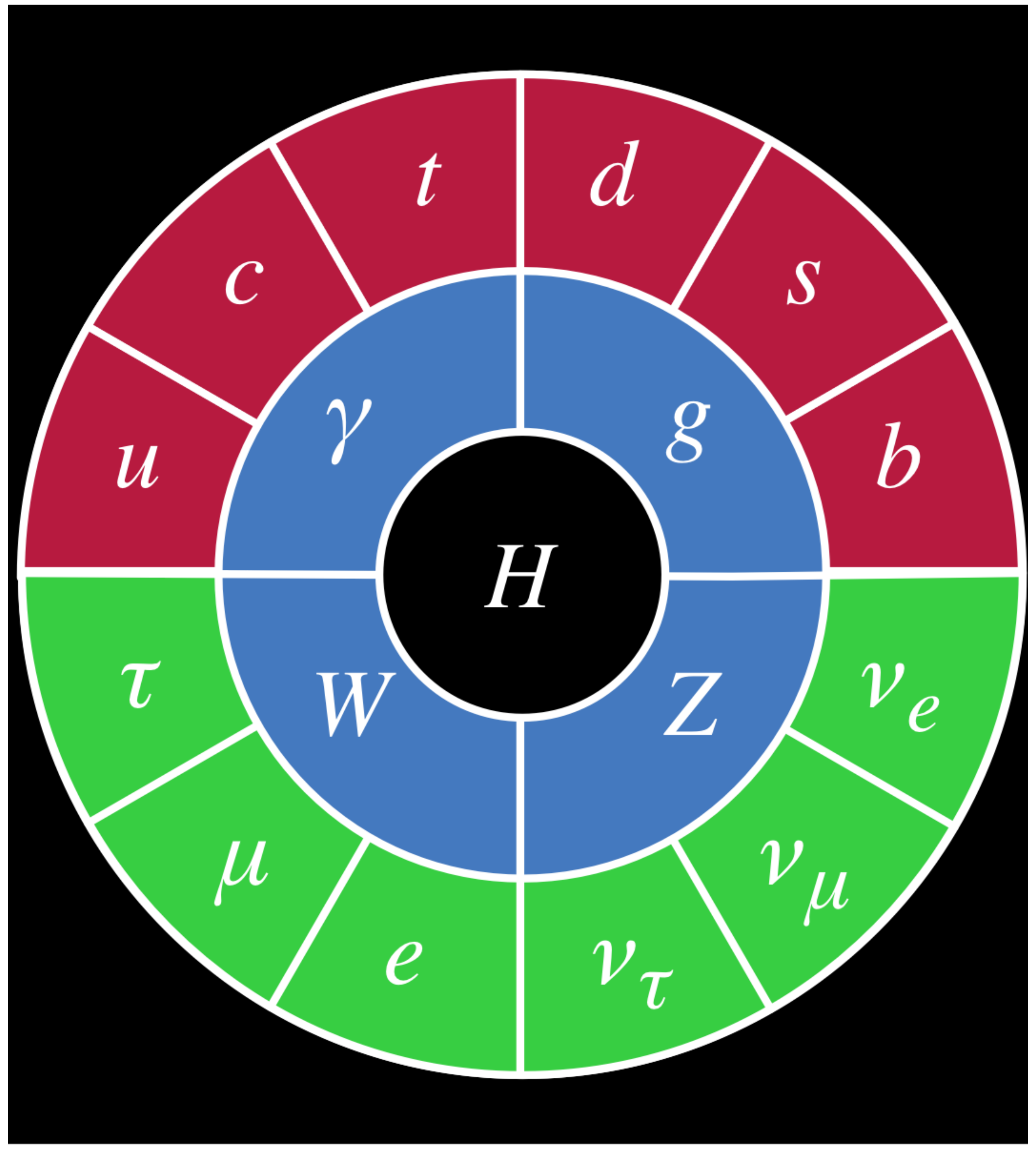
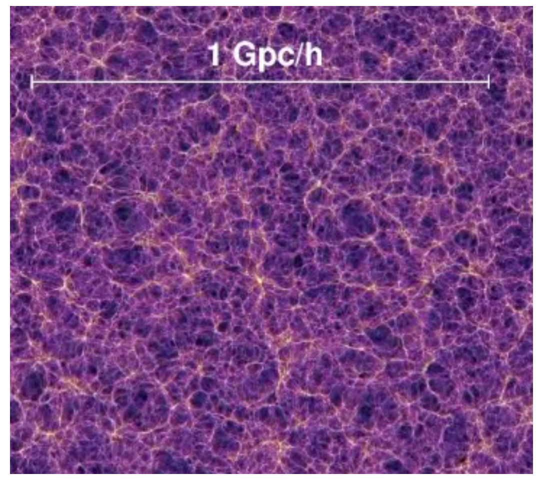


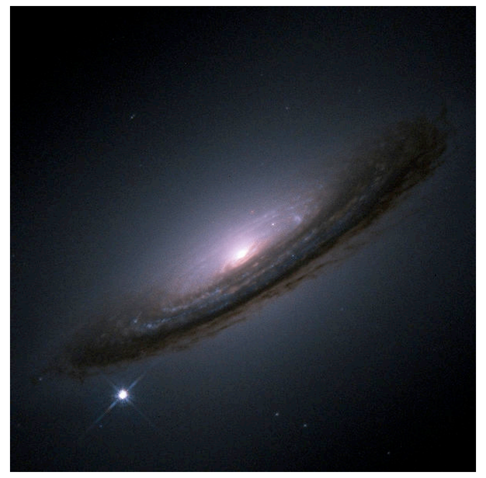
Illustration: Typoform

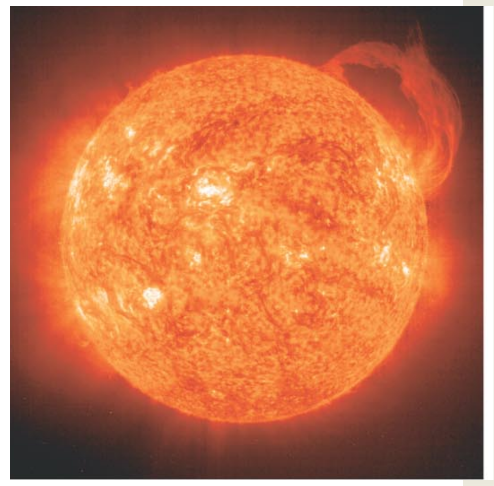


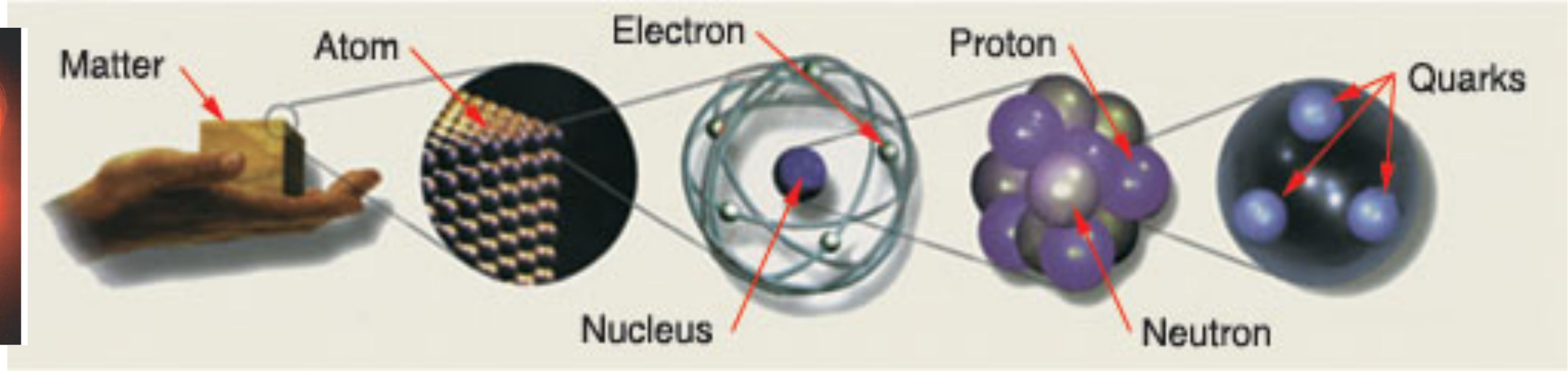


Fundamental Particles & Interactions









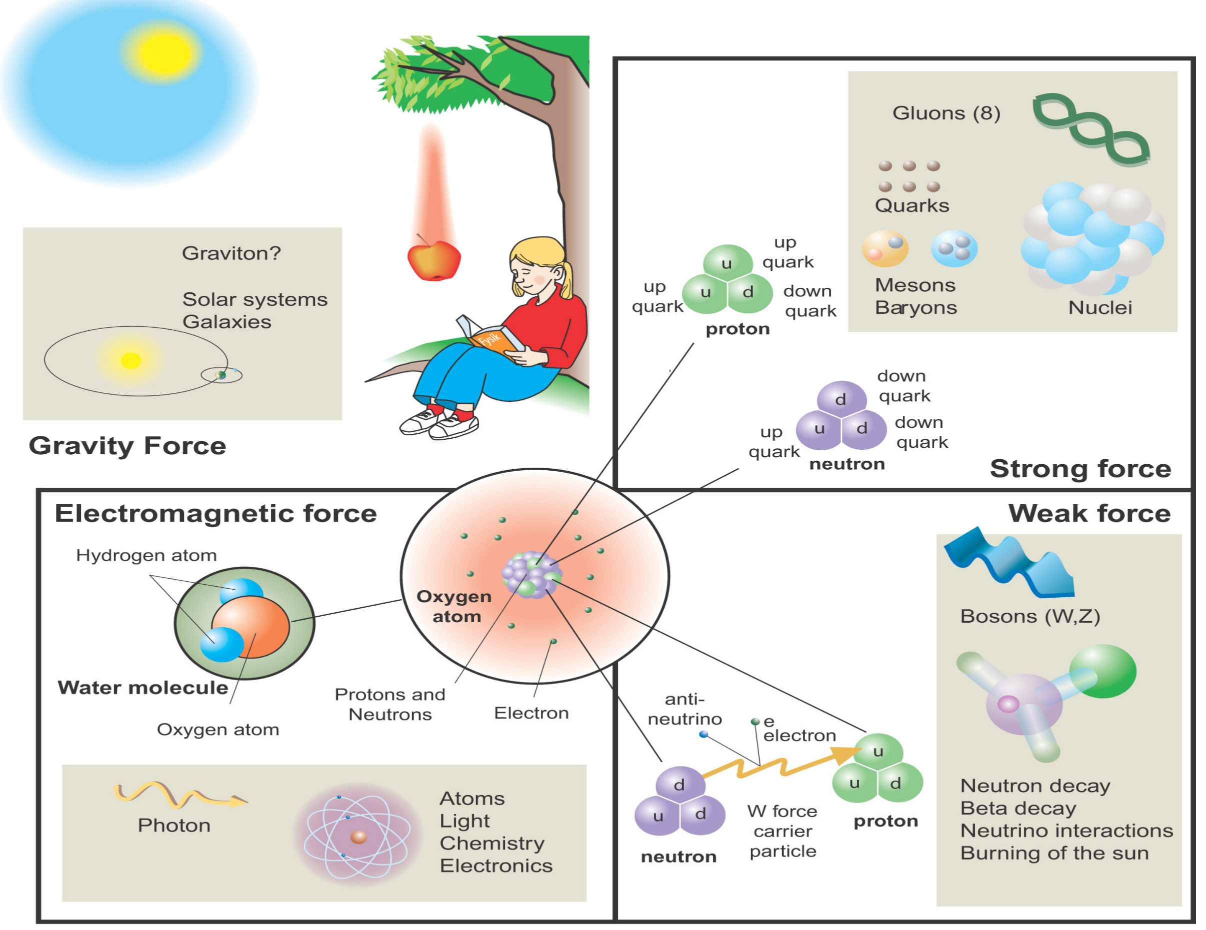
$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^a_{\mu\nu}G_a^{\mu\nu}}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$

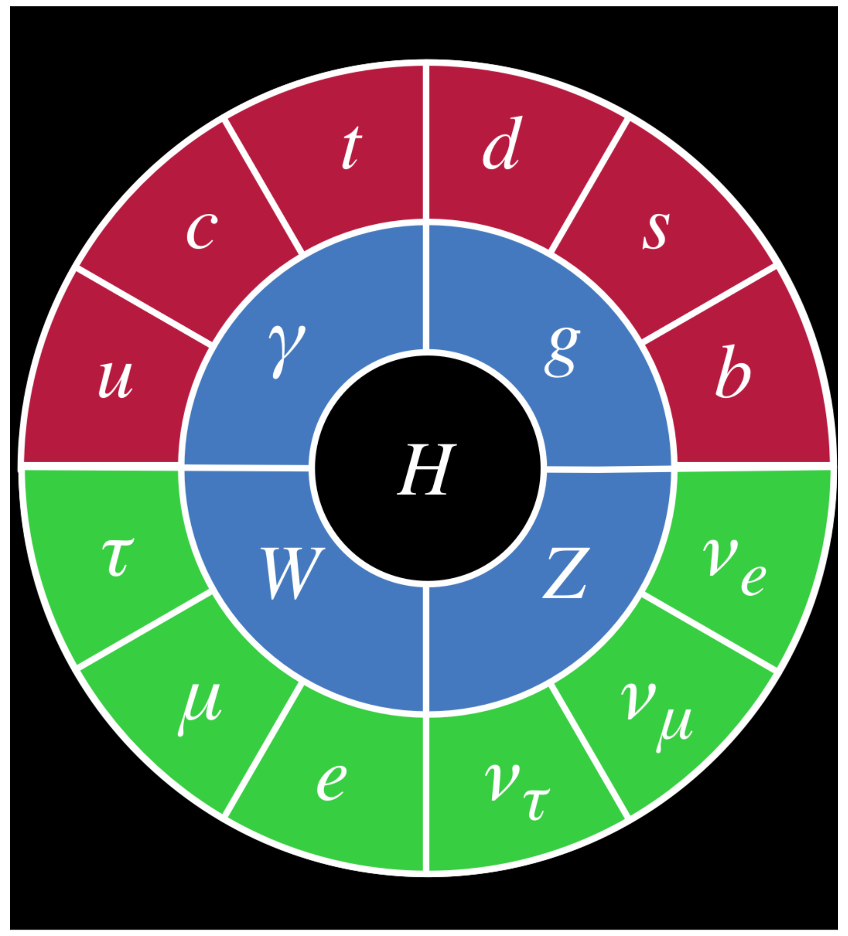
$$+ \underbrace{\bar{L}\gamma^\mu(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)L + \bar{R}\gamma^\mu(i\partial_\mu - \frac{1}{2}g'YB_\mu)R}_{\text{kinetic energies and electroweak interactions of fermions}}$$

$$+ \underbrace{\frac{1}{2}\left|(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)\,\phi\right|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{and Higgs masses and couplings}}$$

$$+ \underbrace{g''(\bar{q}\gamma^\mu T_a q)\,G^a_\mu}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1\bar{L}\phi R + G_\varsigma\bar{R}\phi_c L + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$

Illustration: Typoform

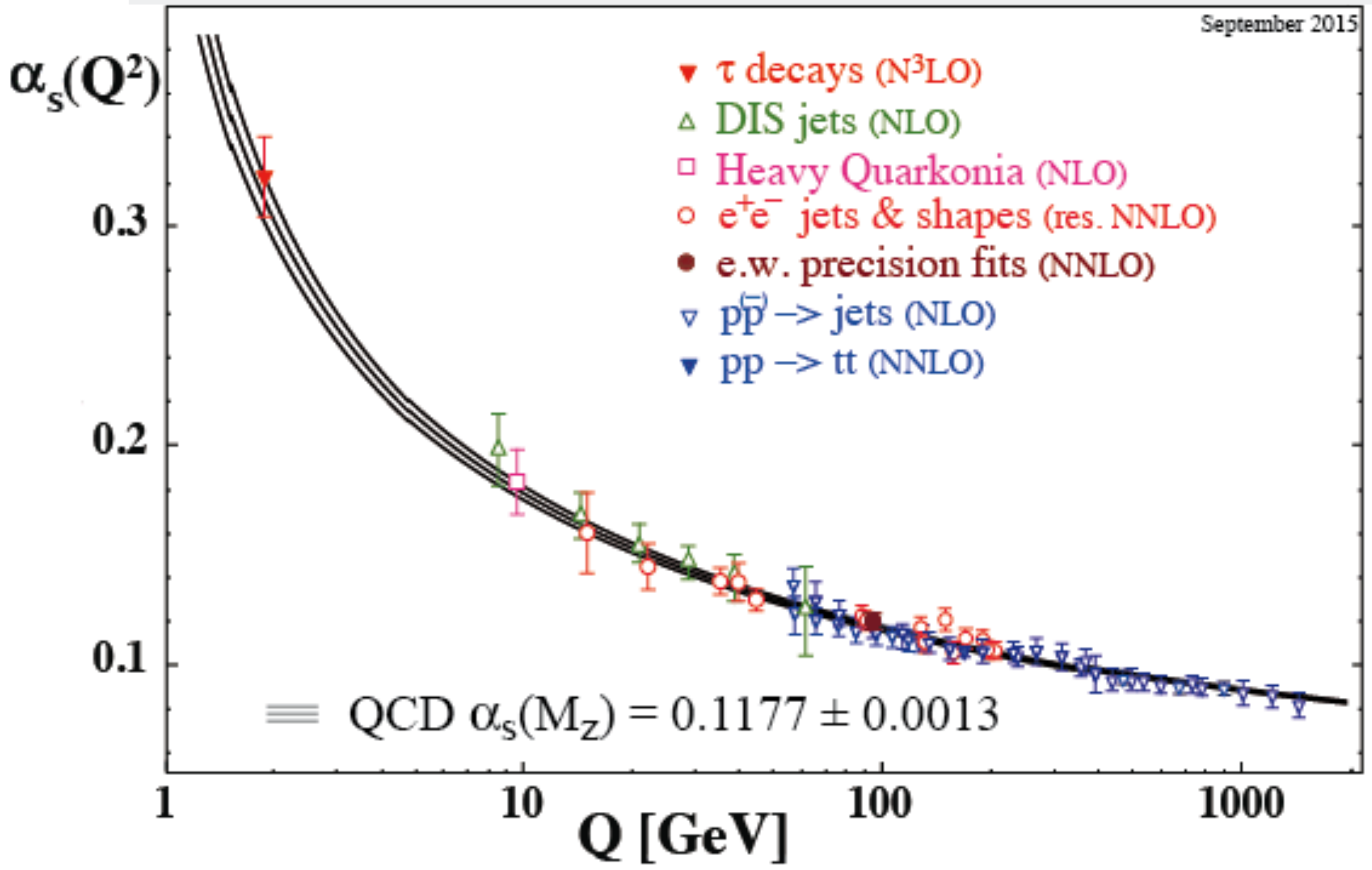




The strong force: Quantum Chromodynamics

Interaction strength depends on energy

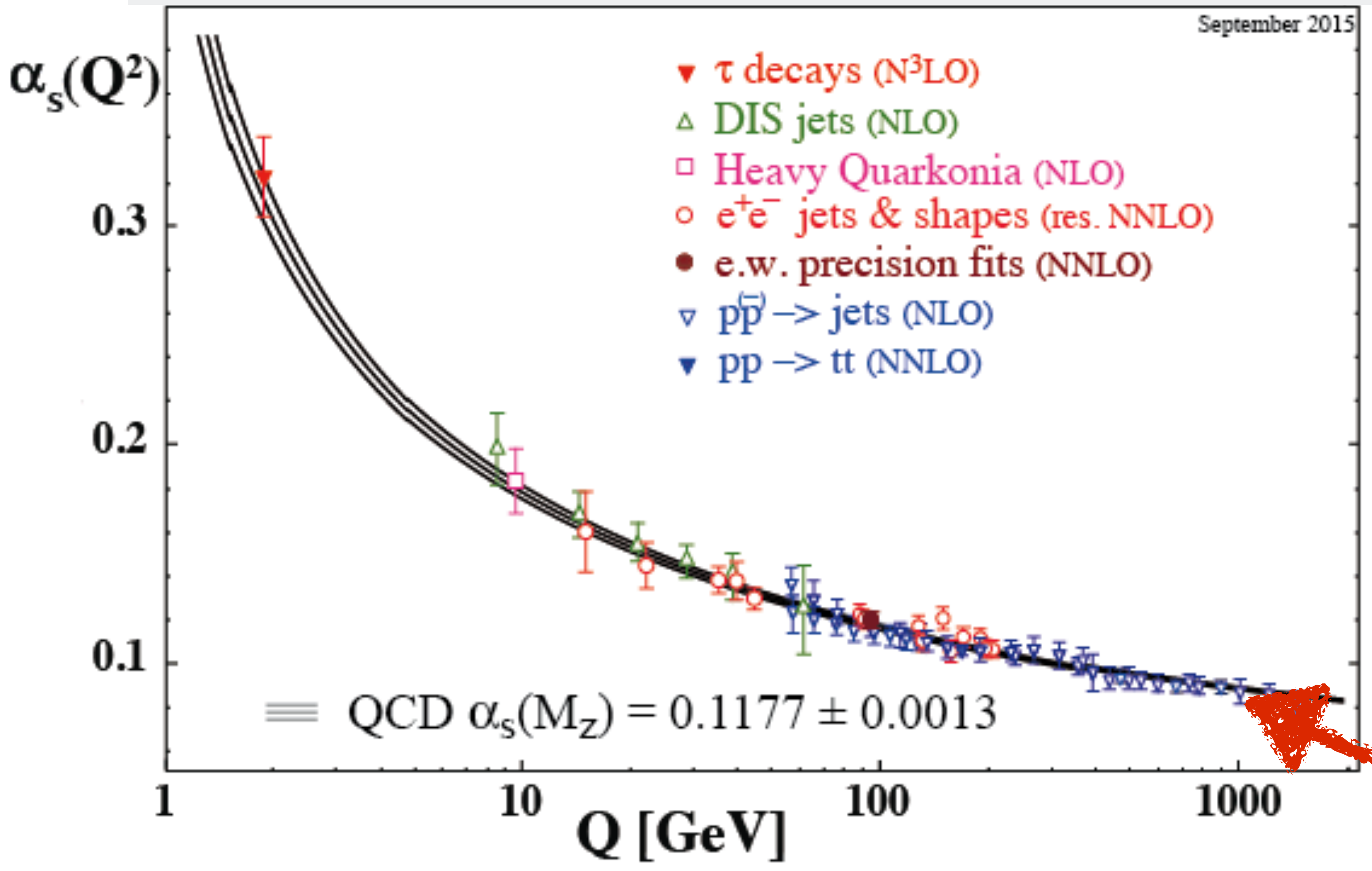
[Gross, Politzer, Wilczek, Nobel 2004]



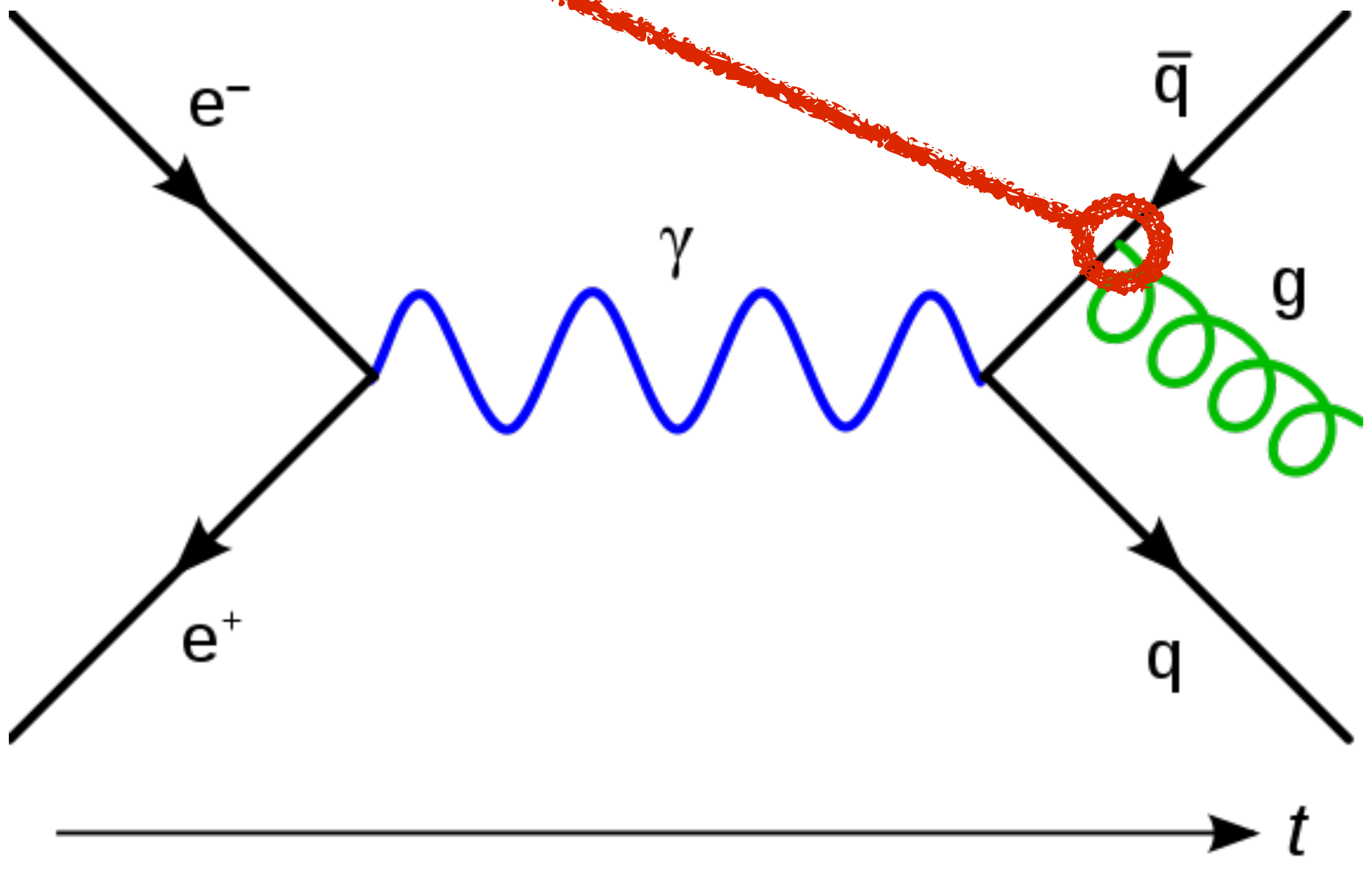
The strong force: Quantum Chromodynamics

Interaction strength depends on energy

[Gross, Politzer, Wilczek, Nobel 2004]



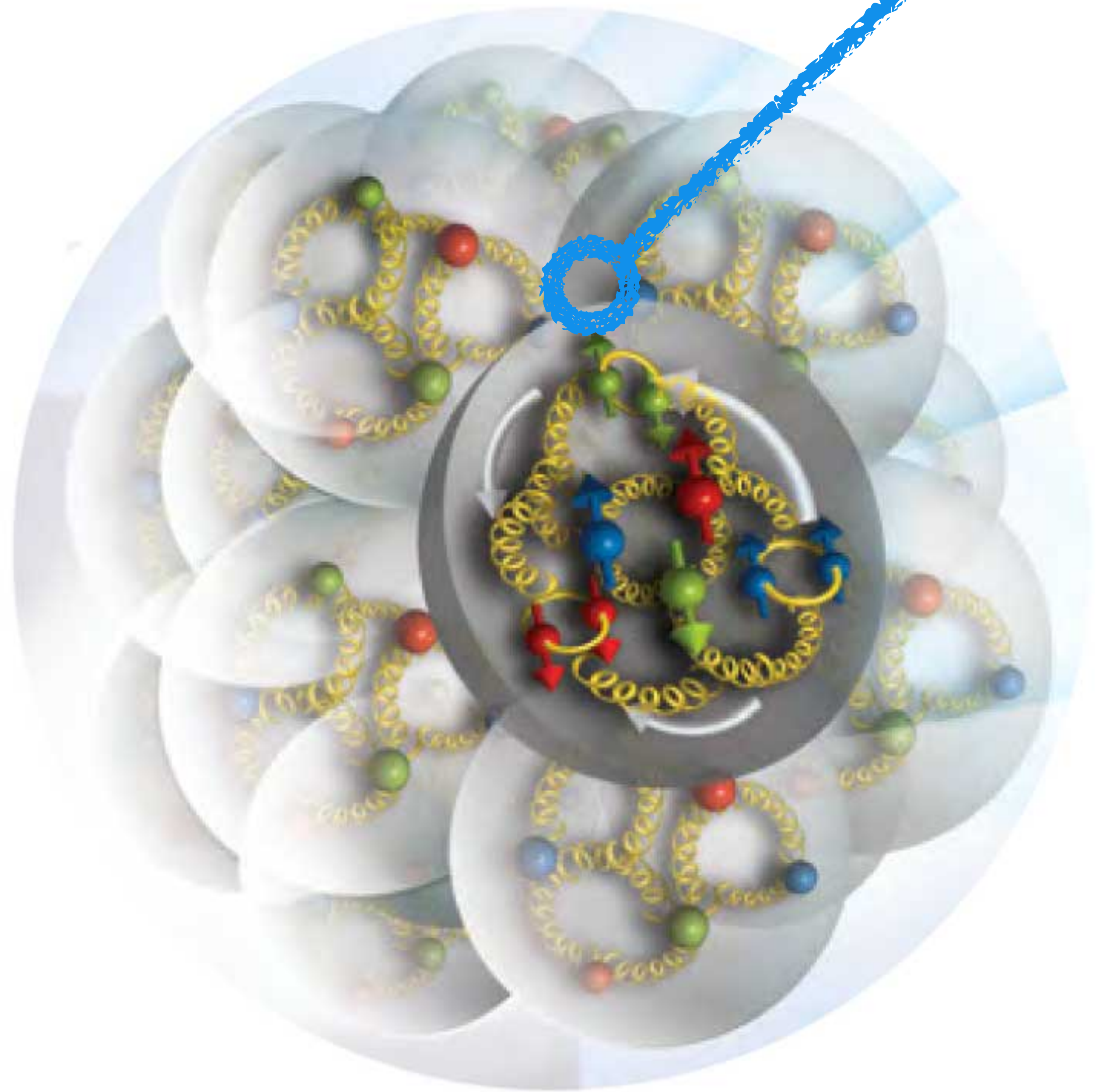
QCD is weak at at high-energies, small coupling, perturbation theory works



The strong force: Quantum Chromodynamics

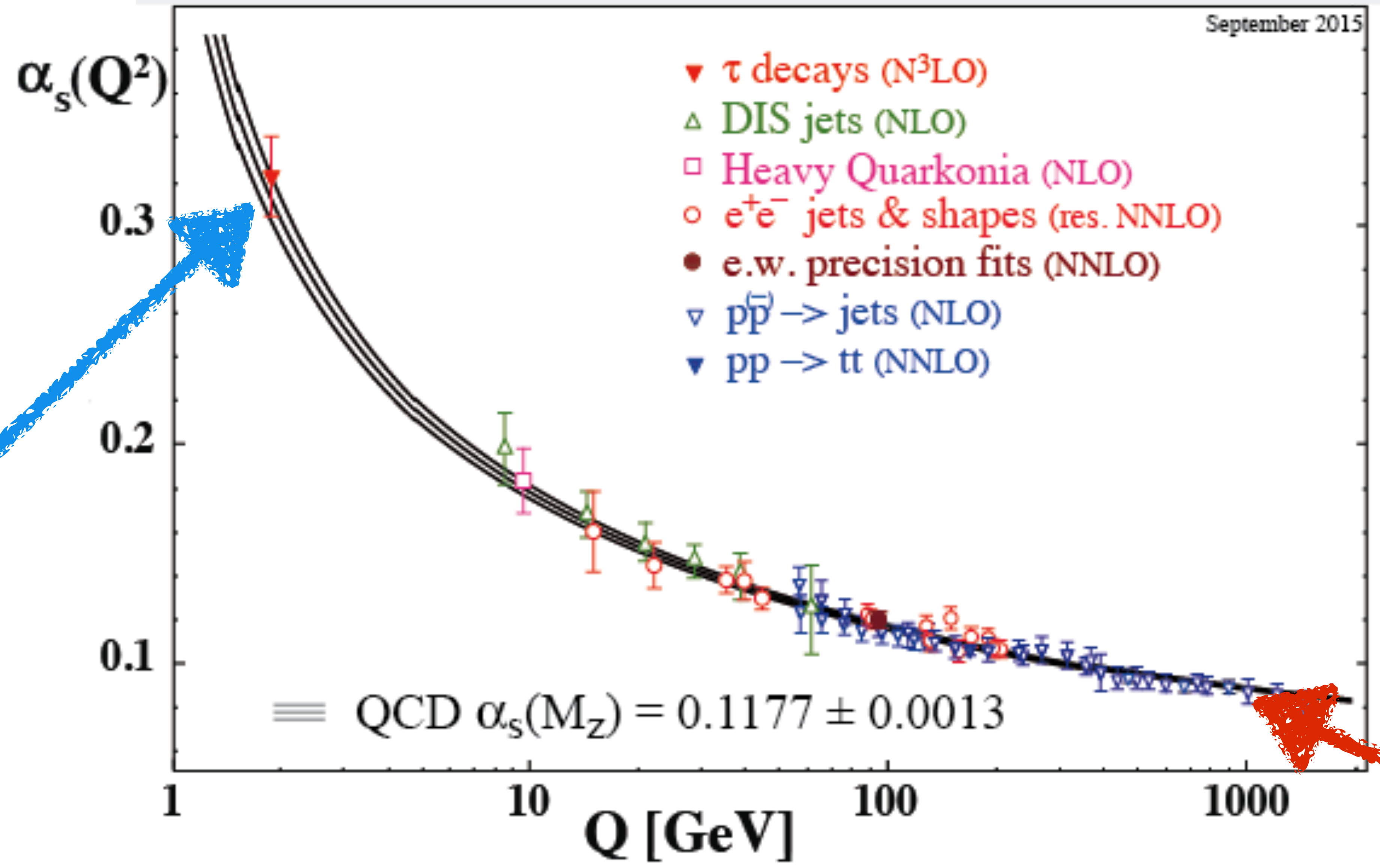
QCD is strong at at low-energies, no small coupling, perturbation theory fails.

Emergent phenomena: protons, pions, etc.

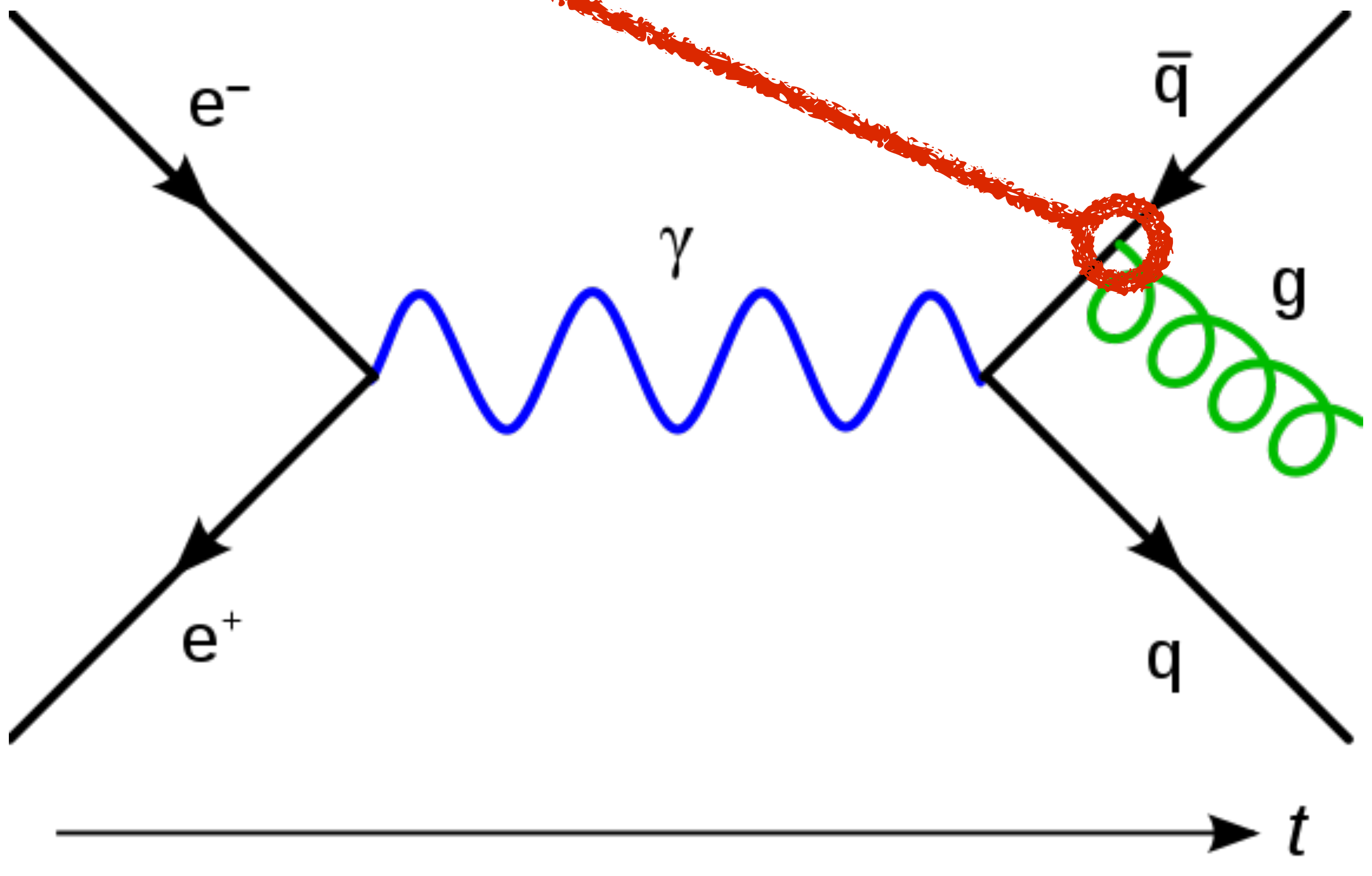


Interaction strength depends on energy

[Gross, Politzer, Wilczek, Nobel 2004]



QCD is weak at at high-energies, small coupling, perturbation theory works



Like the silicon chips of more recent years, the Feynman diagram was bringing computation to the masses.

–JULIAN SCHWINGER

Simulating particle physics processes

Theory parameters

θ

Evolution

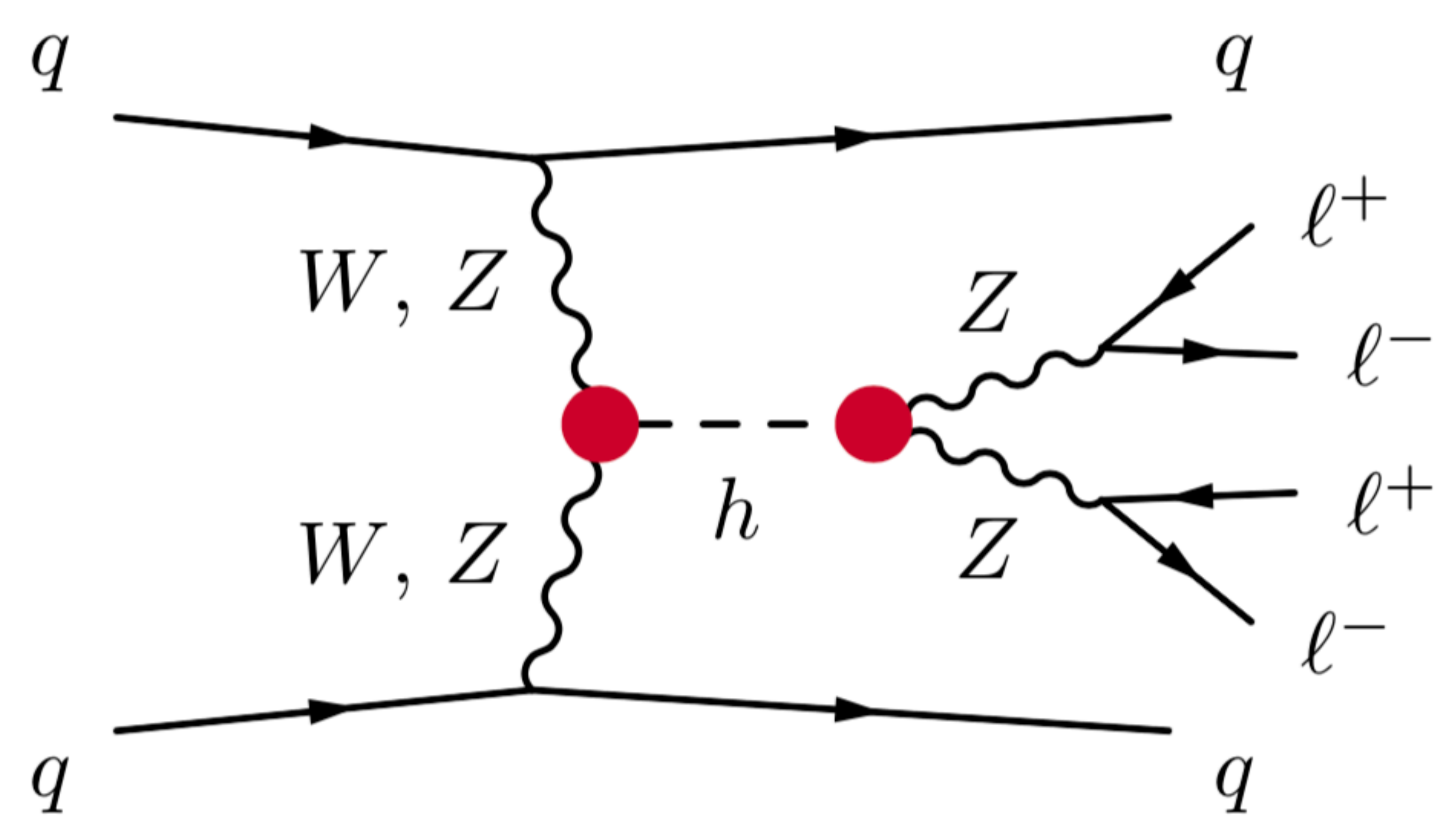
Simulating particle physics processes

Latent variables

Parton-level momenta

Theory parameters

$z_p \longleftarrow \theta$

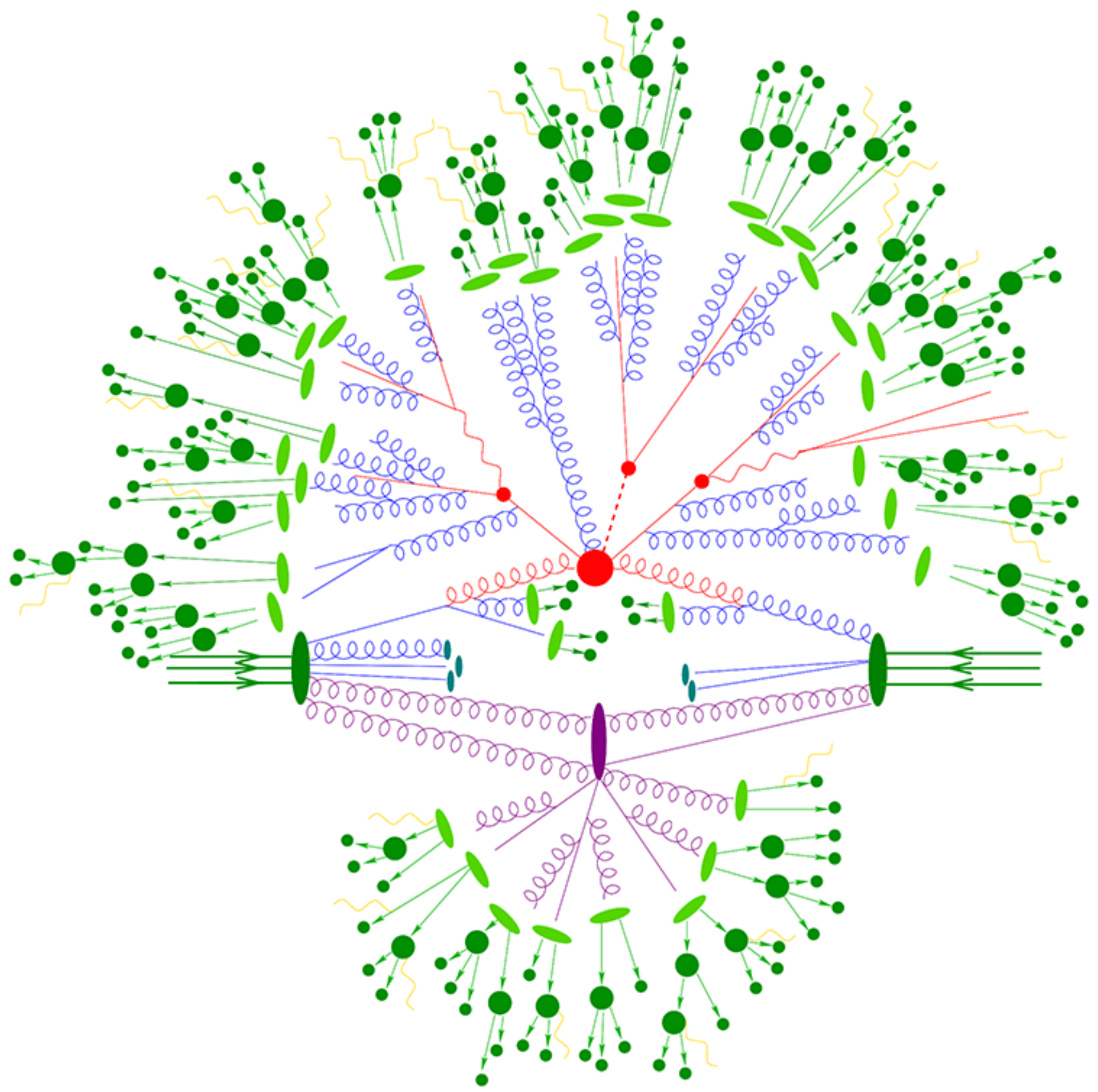


Evolution

Simulating particle physics processes

Latent variables

Shower splittings		Parton-level momenta		Theory parameters
z_s	←	z_p	←	θ



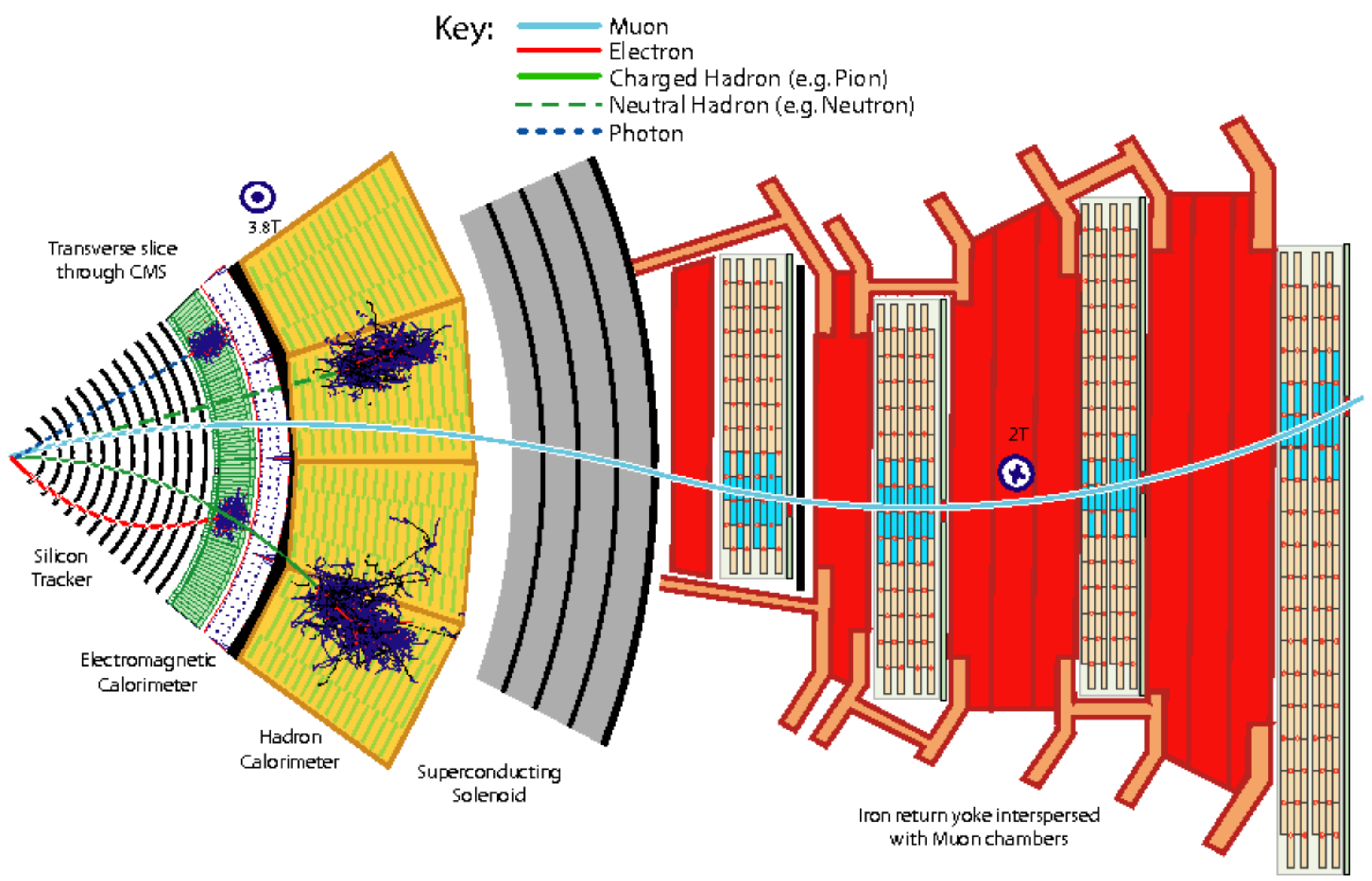
[F. Krauss]

← Evolution

Simulating particle physics processes

Latent variables

Detector interactions		Shower splittings		Parton-level momenta		Theory parameters
z_d	$\longleftarrow$	z_s	$\longleftarrow$	z_p	$\longleftarrow$	θ



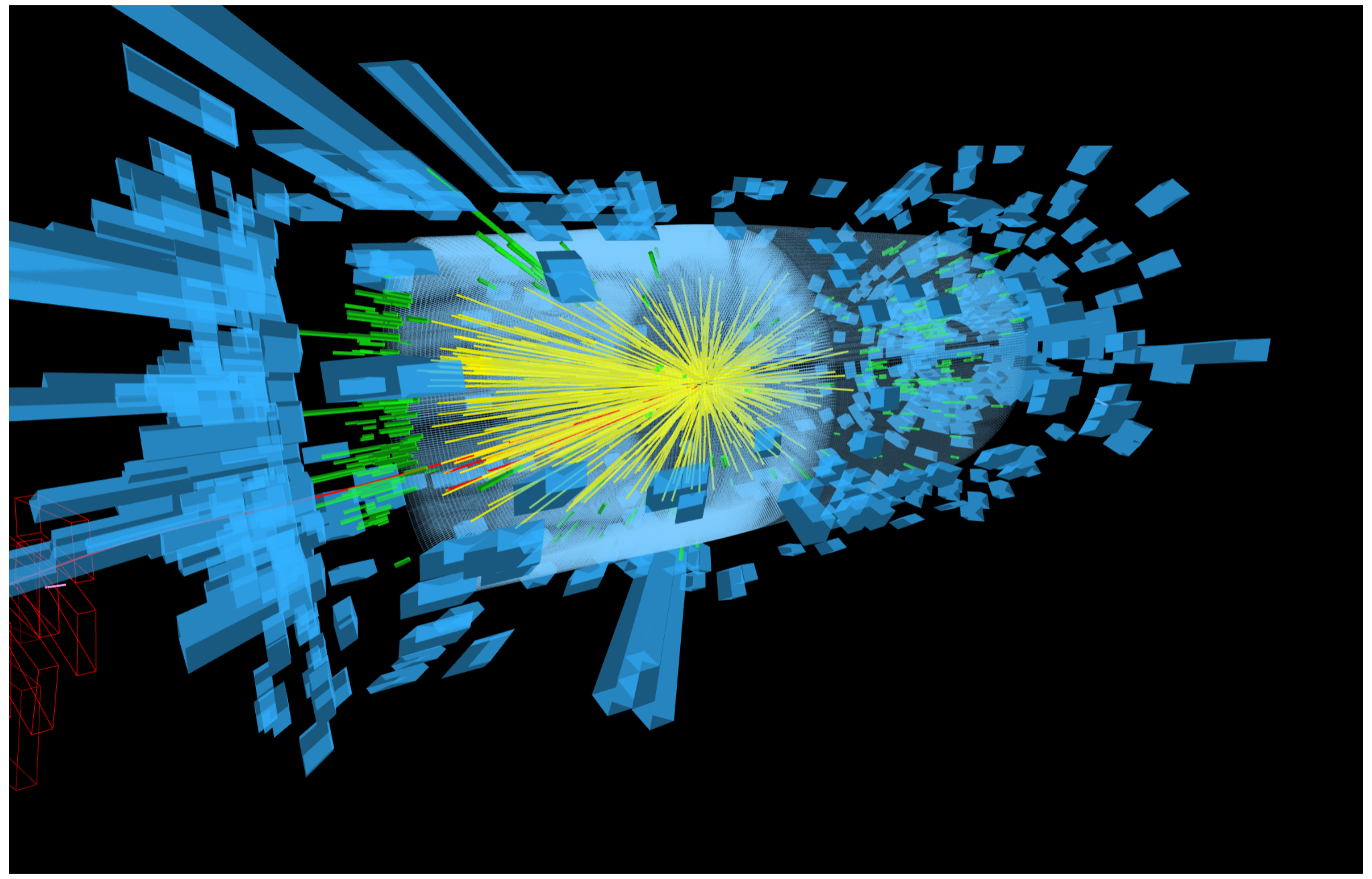
[CMS]

$\longleftarrow$ Evolution

Simulating particle physics processes

Latent variables

Observables	Detector interactions	Shower splittings	Parton-level momenta	Theory parameters
x ←	z_d ←	z_s ←	z_p ←	θ



Evolution

Simulating particle physics processes

Latent variables

Observables		Detector interactions	Shower splittings	Parton-level momenta	Theory parameters
x	←	z_d ←	z_s ←	z_p ←	θ
Sample from	$p(x\|z_d)$	$p(z_d\|z_s)$	$p(z_s\|z_p)$	$p(z_p\|\theta)$	

GEANT4 A SIMULATION TOOLKIT

MADGRAPH5_aMC@NLO

←

Prediction (simulation)

Simulating particle physics processes

Latent variables

Observables	Detector interactions	Shower splittings	Parton-level momenta	Theory parameters
x	z_d	z_s	z_p	θ

$$x \leftarrow z_d \leftarrow z_s \leftarrow z_p \leftarrow \theta$$

$$p(x|\theta) = \int \mathrm{d}z_d \int \mathrm{d}z_s \int \mathrm{d}z_p \quad p(x|z_d) \quad p(z_d|z_s) \quad p(z_s|z_p) \quad p(z_p|\theta)$$

Inference →

Simulating particle physics processes

Latent variables

Observables | Detector interactions | Shower splittings | Parton-level momenta | Theory parameters

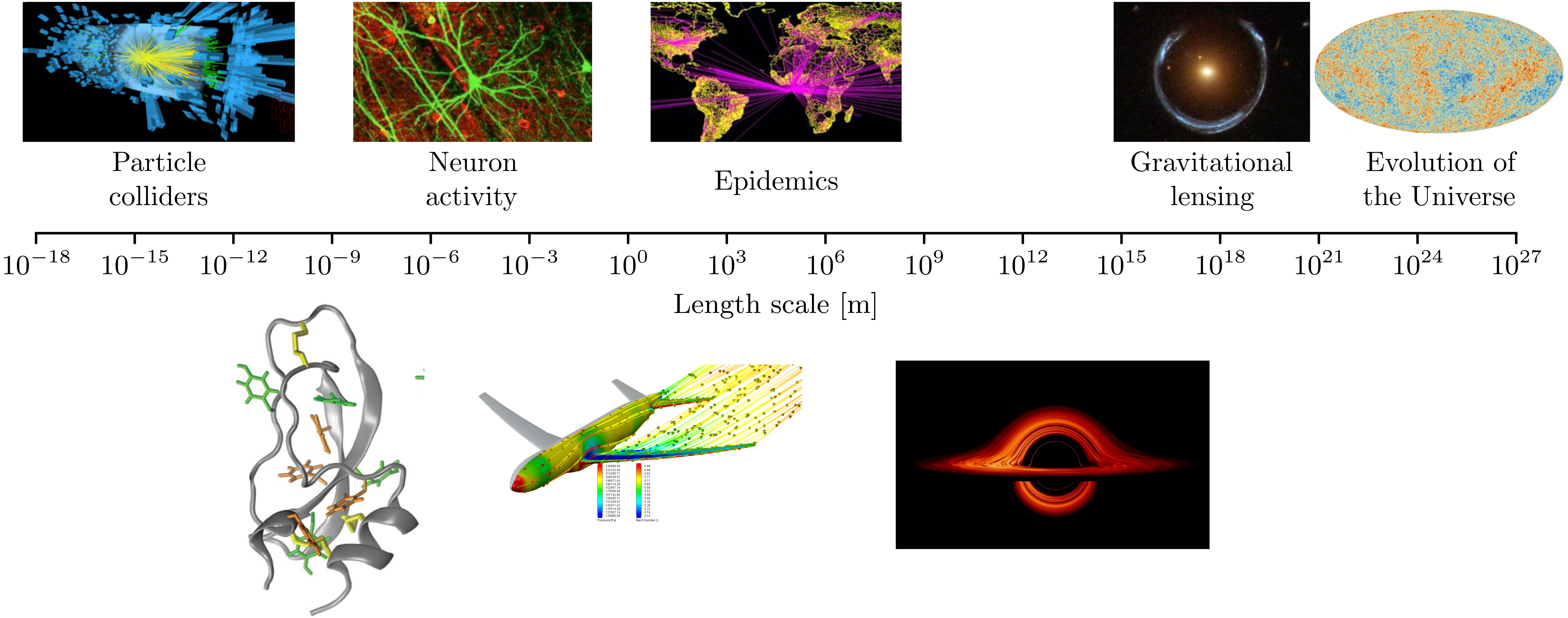
$$x \leftarrow z_d \leftarrow z_s \leftarrow z_p \leftarrow \theta$$

$$p(x|\theta) = \boxed{\int \mathrm{d}z_d \int \mathrm{d}z_s \int \mathrm{d}z_p} \; p(x|z_d) \quad p(z_d|z_s) \quad p(z_s|z_p) \quad p(z_p|\theta)$$

It's infeasible to calculate the integral over this enormous space!

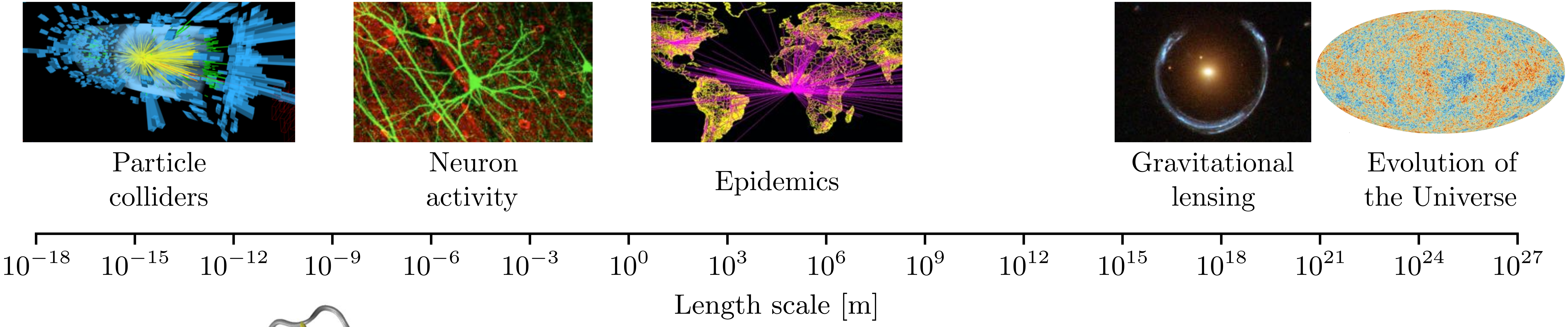
Inference →

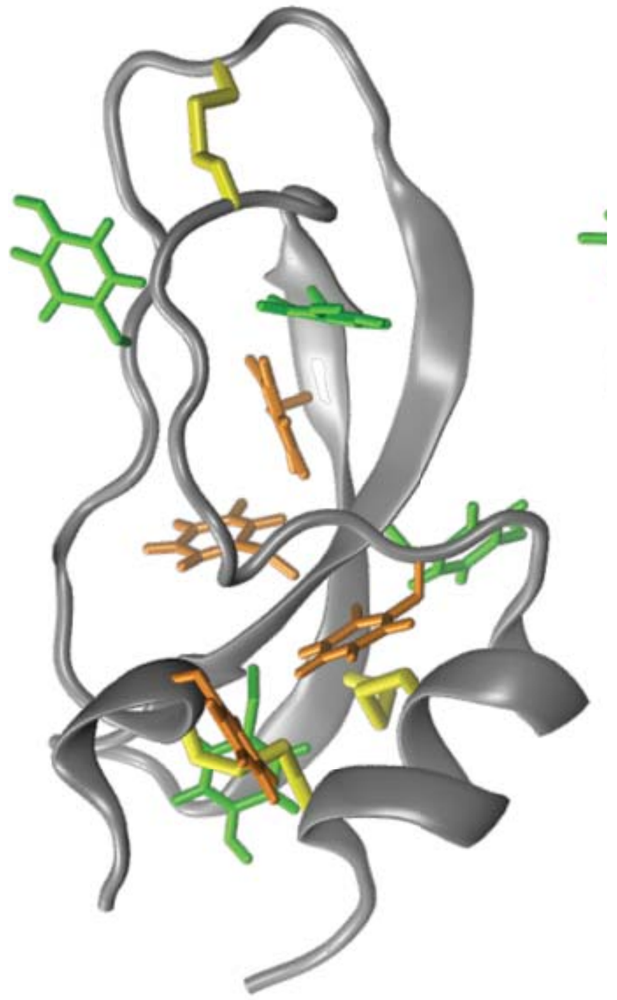
Science is replete with high-fidelity simulators

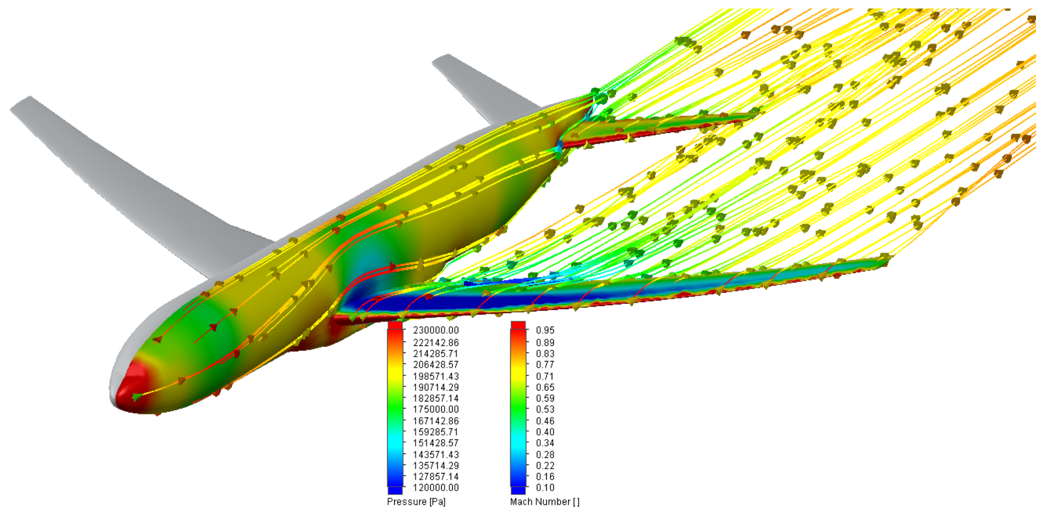


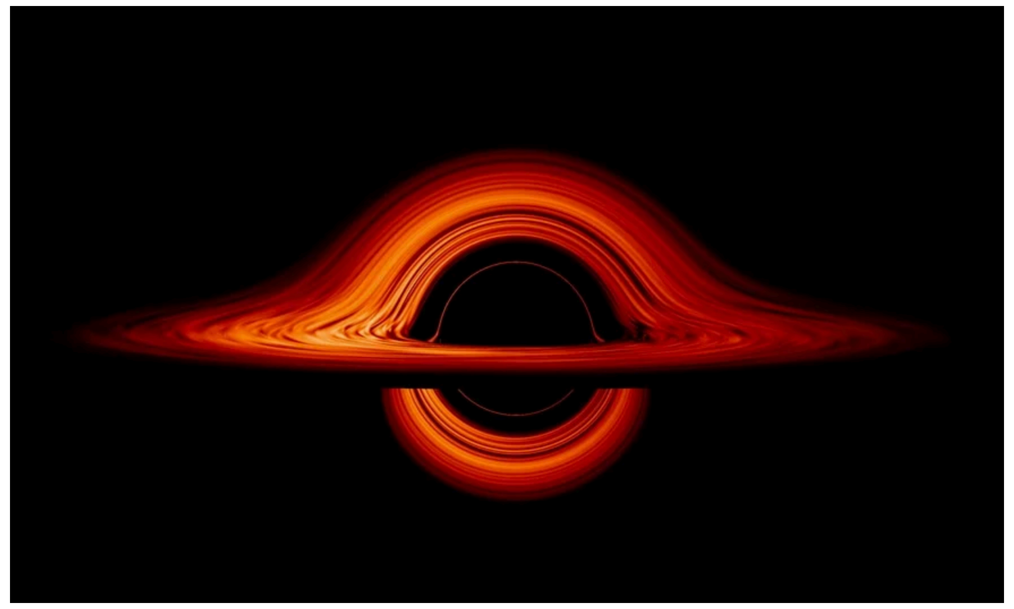
Simulators are causal, generative models of the data generating process

Science is replete with high-fidelity simulators





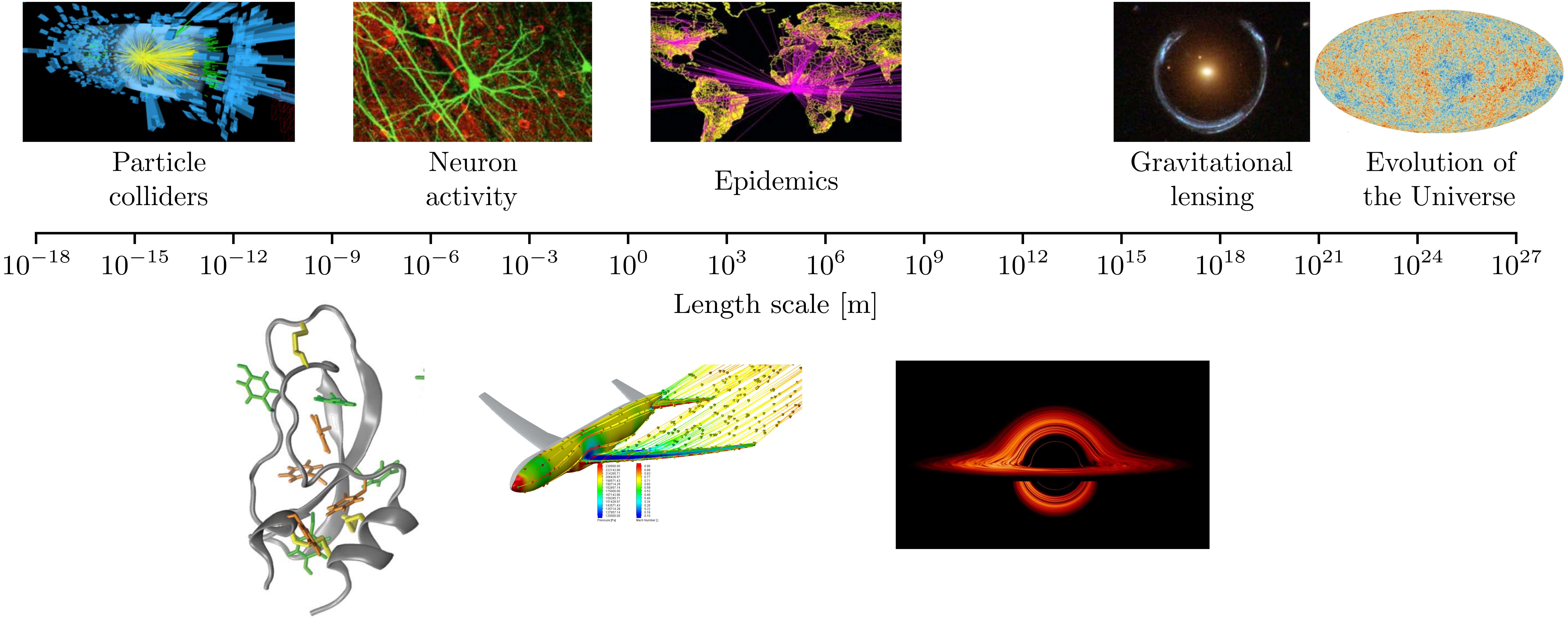




The expressiveness of programming languages facilitates the development of complex, high-fidelity simulations, and the power of modern computing provides the ability to generate synthetic data from them.

[Cranmer, Brehmer, Louppe PNAS (2020), arXiv:1911.01429]

Science is replete with high-fidelity simulators



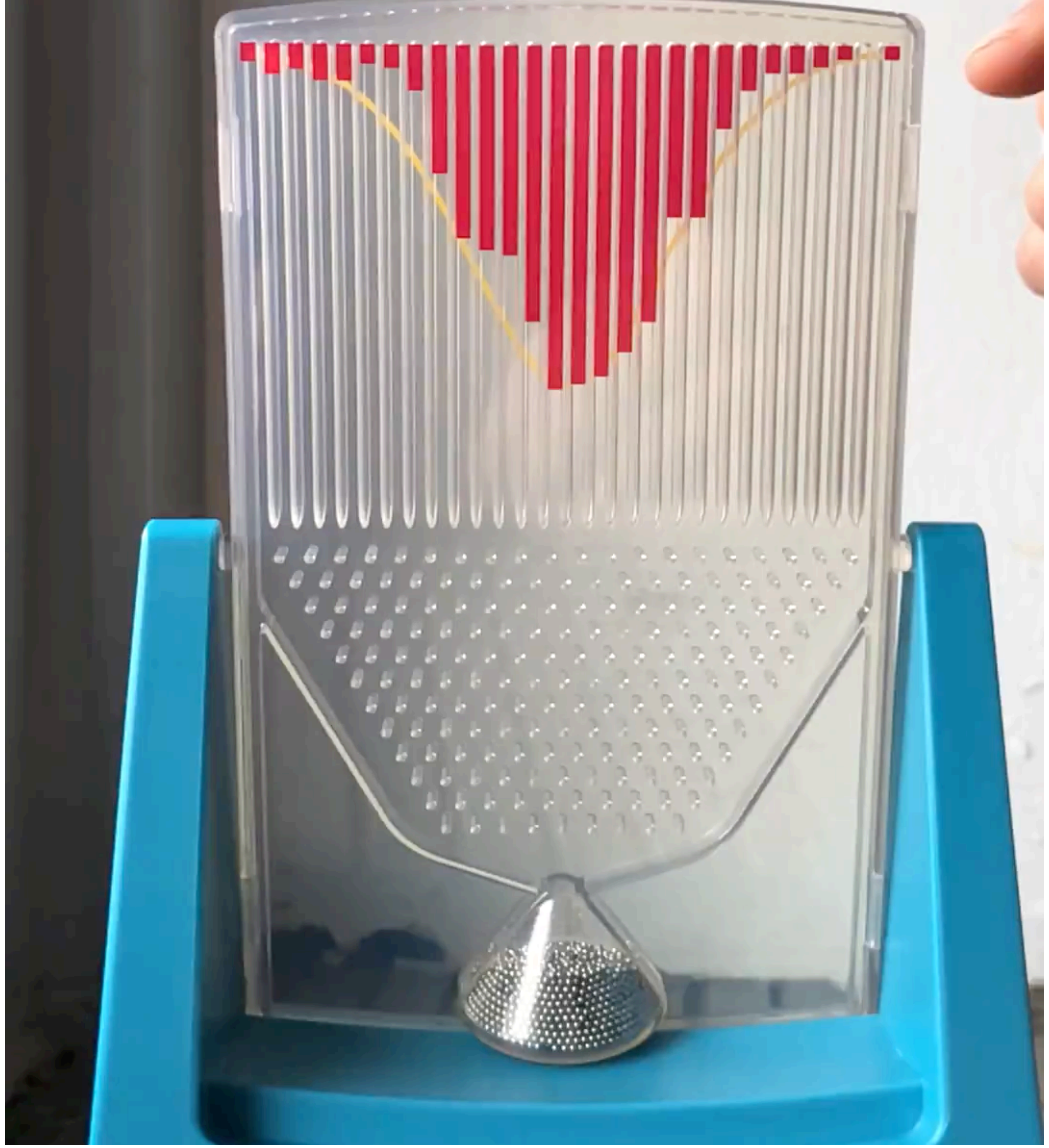
Unfortunately, these simulators are poorly suited for statistical inference.

[Cranmer, Brehmer, Louppe PNAS (2020), arXiv:1911.01429]

A toy example

Imagine the entire board is slightly tilted, which biases the probability to bounce left/right.

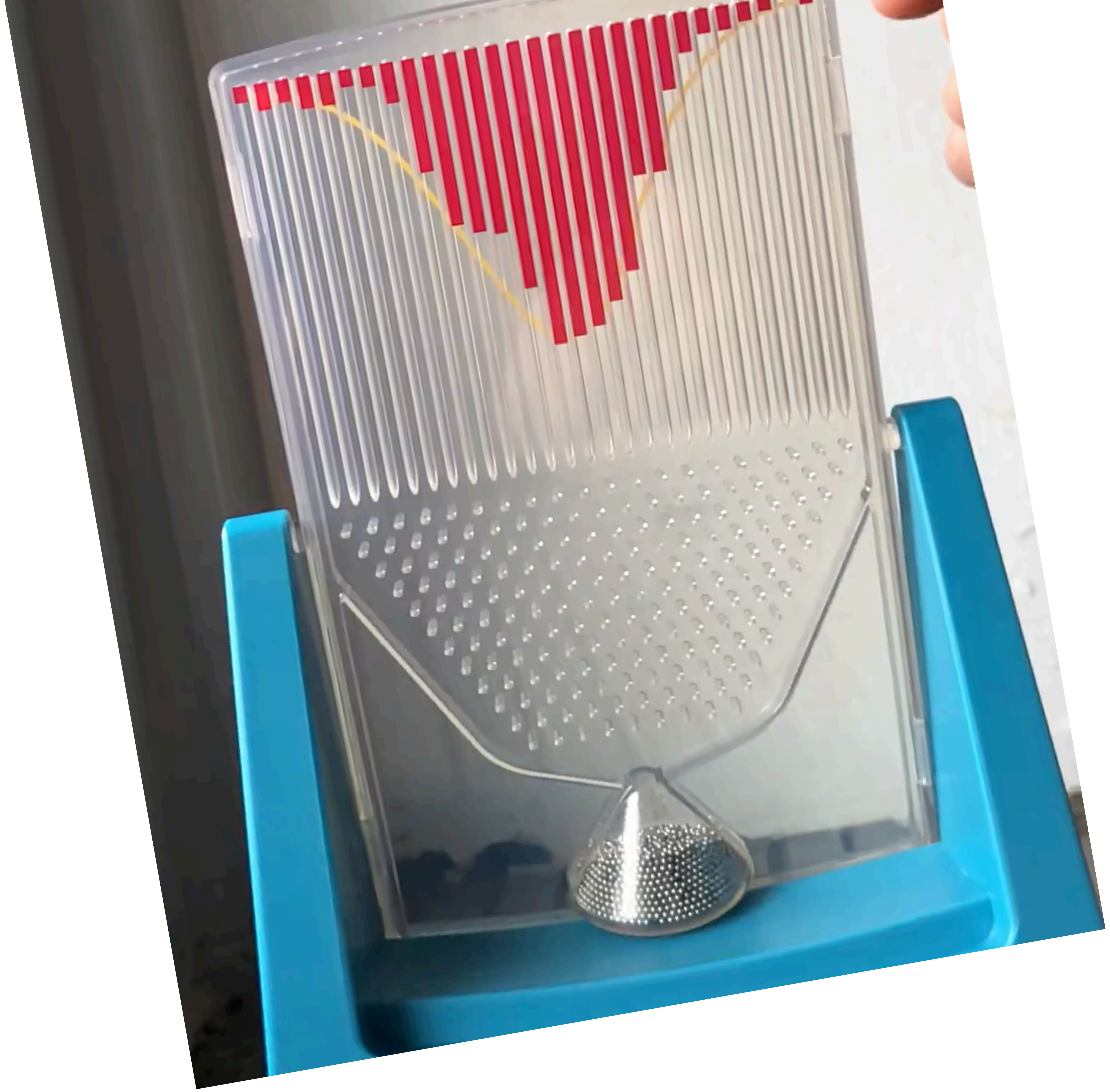
Say we want to infer θ, the probability to bounce right based on distribution of x



A toy example

Imagine the entire board is slightly tilted, which biases the probability to bounce left/right.

Say we want to infer θ, the probability to bounce right based on distribution of x



A toy example

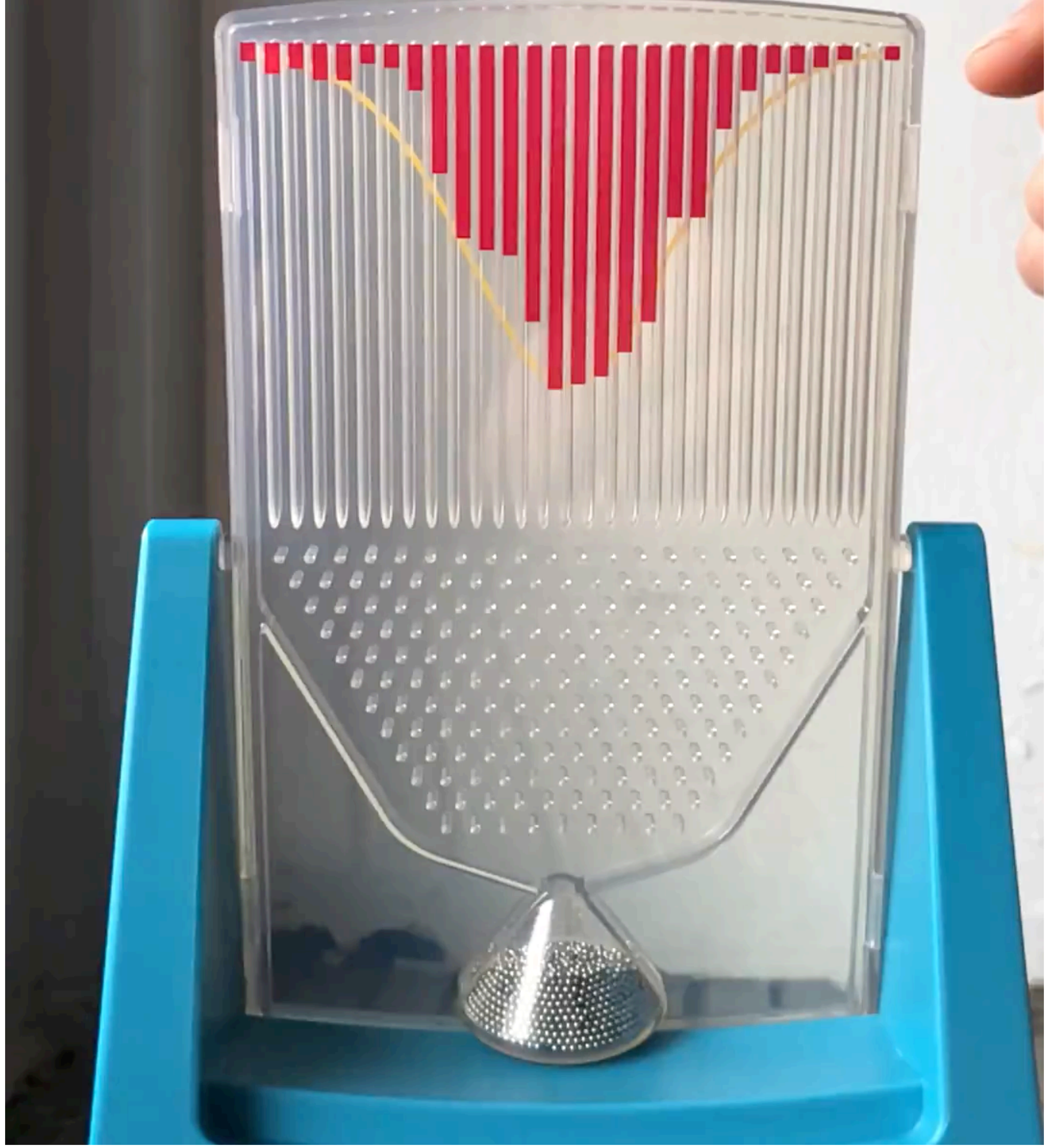
Imagine the entire board is slightly tilted, which biases the probability to bounce left/right.

Say we want to infer θ, the probability to bounce right based on distribution of x

A toy example

Imagine the entire board is slightly tilted, which biases the probability to bounce left/right.

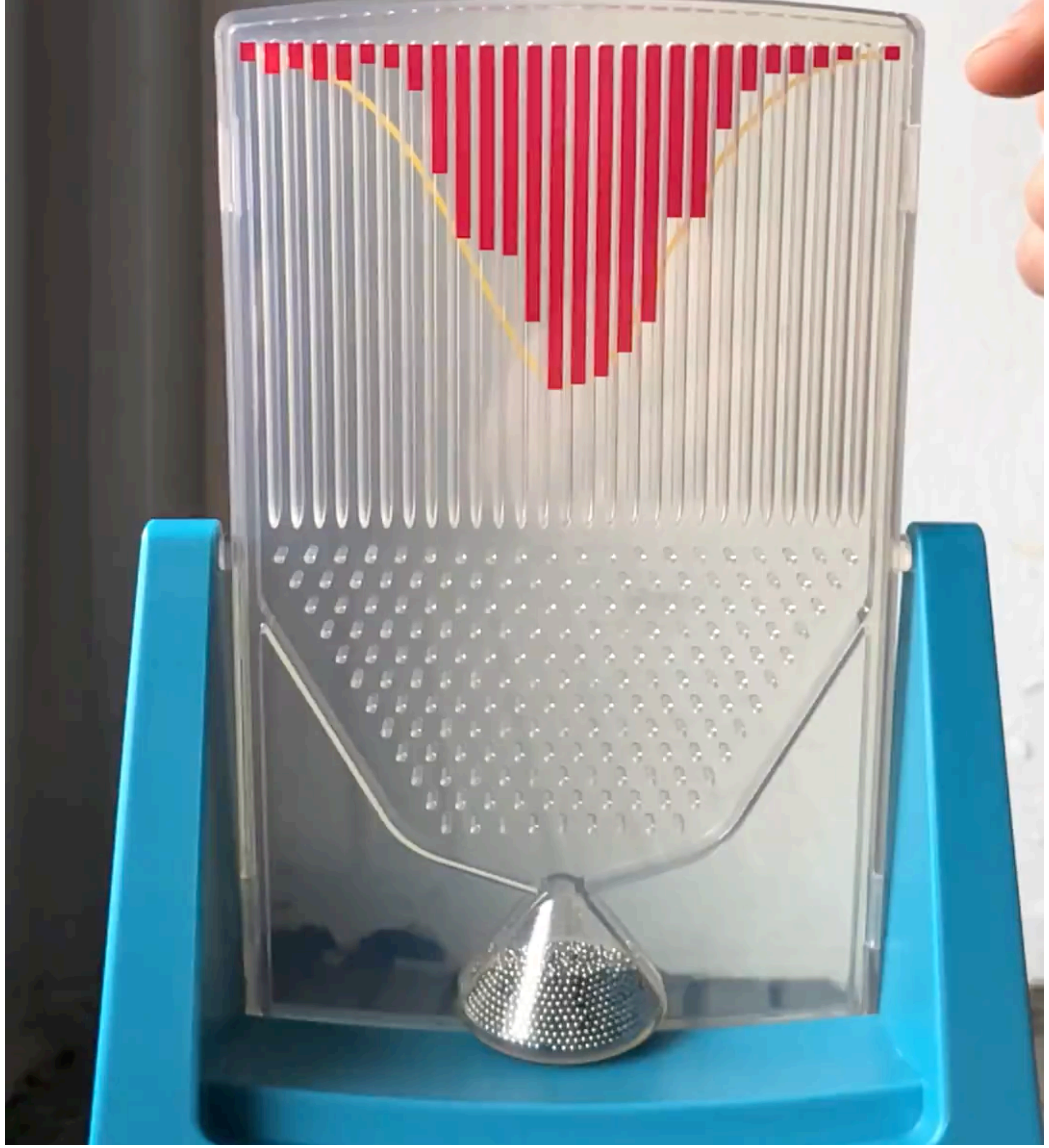
Say we want to infer θ,
the probability to bounce right
based on distribution of x



A toy example

Imagine the entire board is slightly tilted, which biases the probability to bounce left/right.

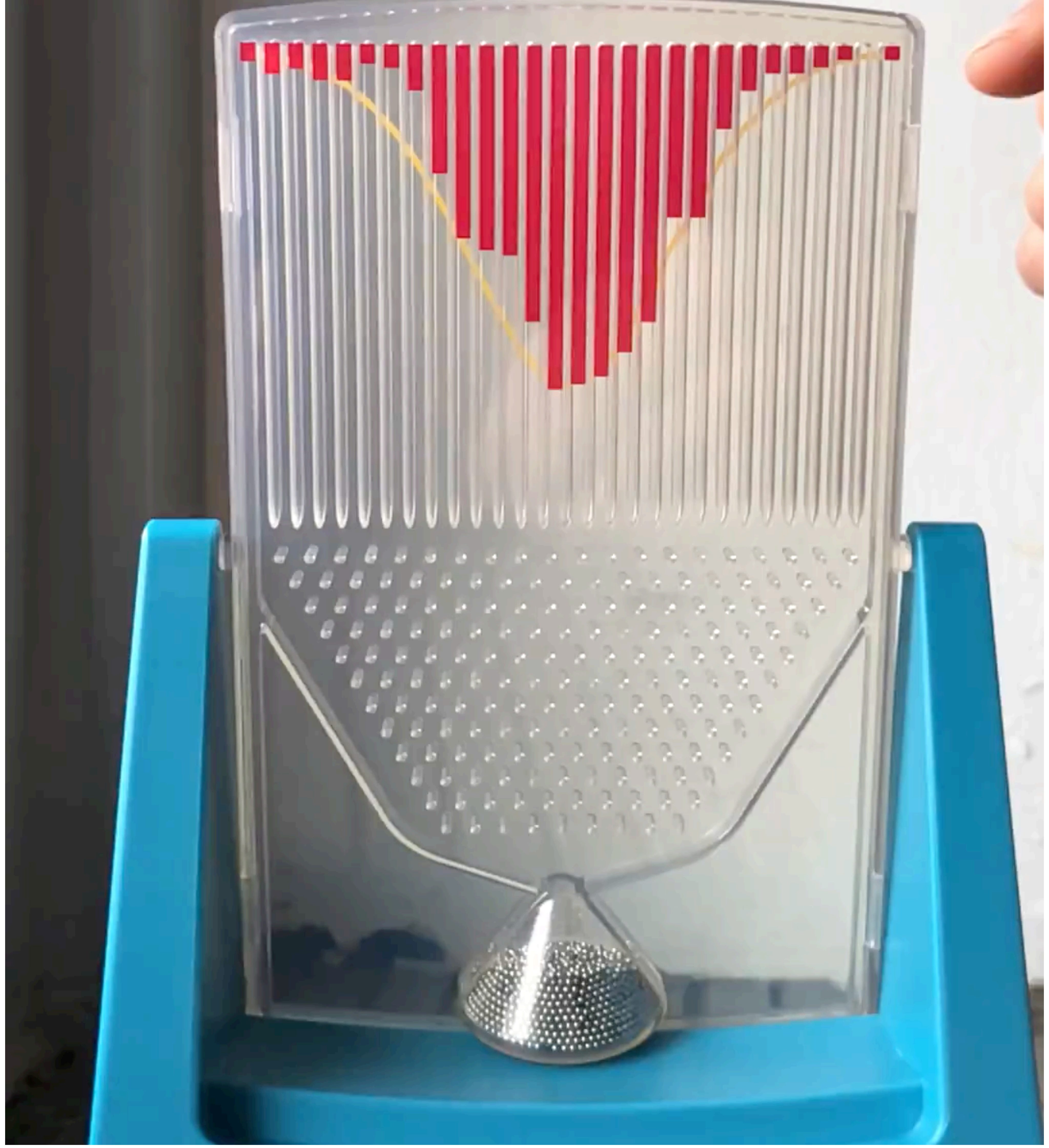
Say we want to infer θ, the probability to bounce right based on distribution of x



A toy example

Imagine the entire board is slightly tilted, which biases the probability to bounce left/right.

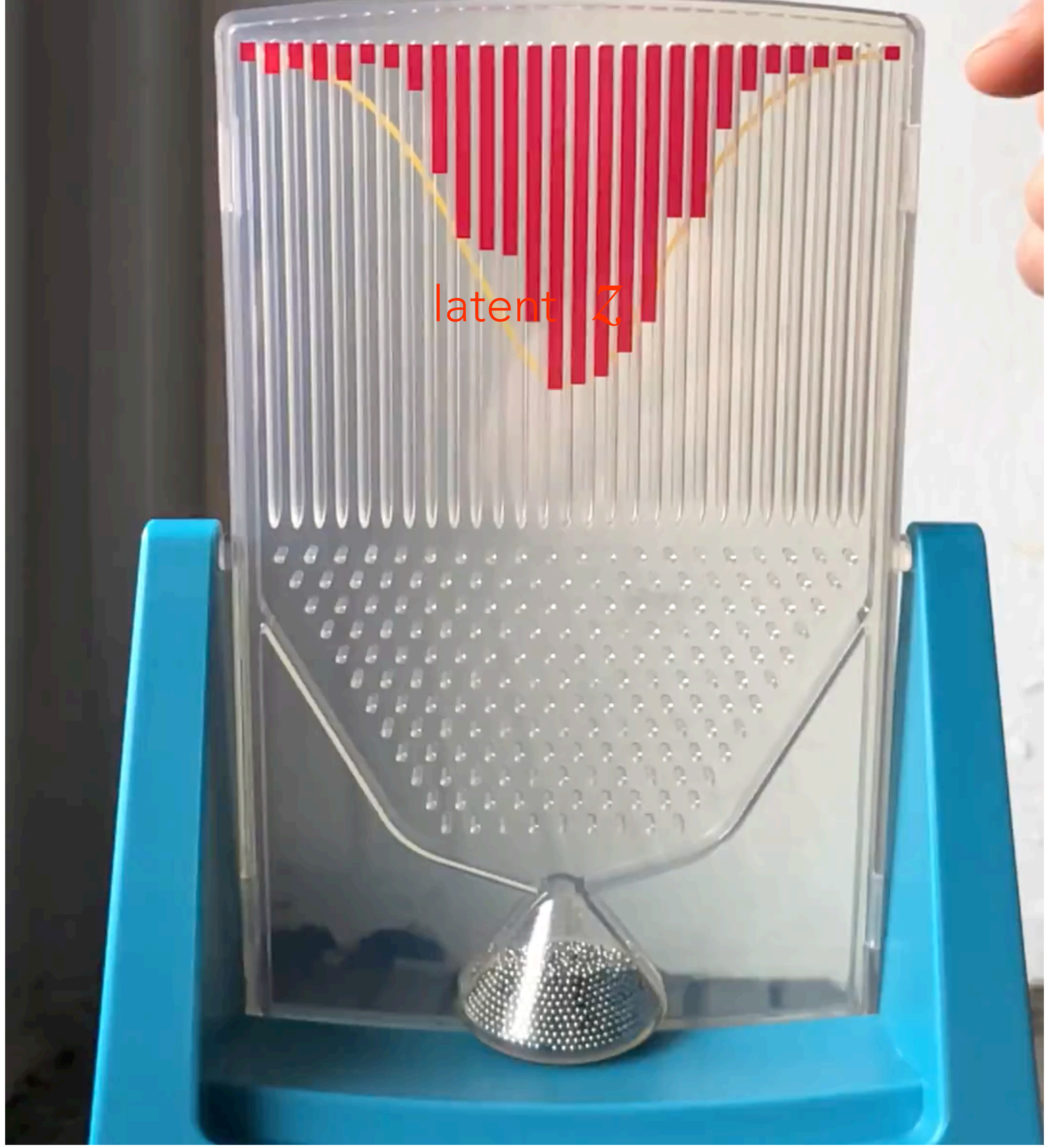
Say we want to infer θ, the probability to bounce right based on distribution of x



A toy example

Imagine the entire board is slightly tilted, which biases the probability to bounce left/right.

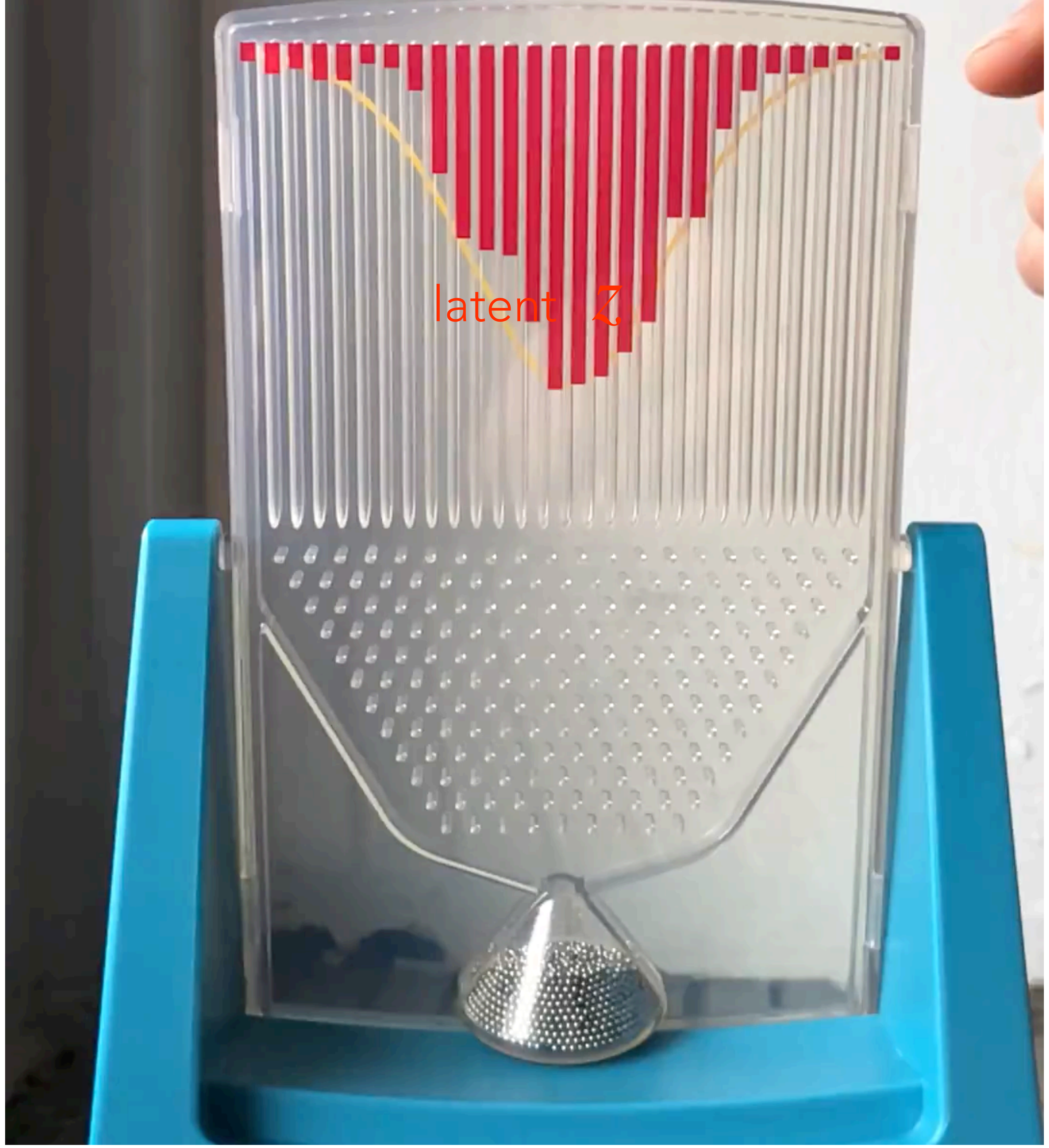
Say we want to infer θ, the probability to bounce right based on distribution of x



A toy example

Imagine the entire board is slightly tilted, which biases the probability to bounce left/right.

Say we want to infer θ,
the probability to bounce right based on distribution of x

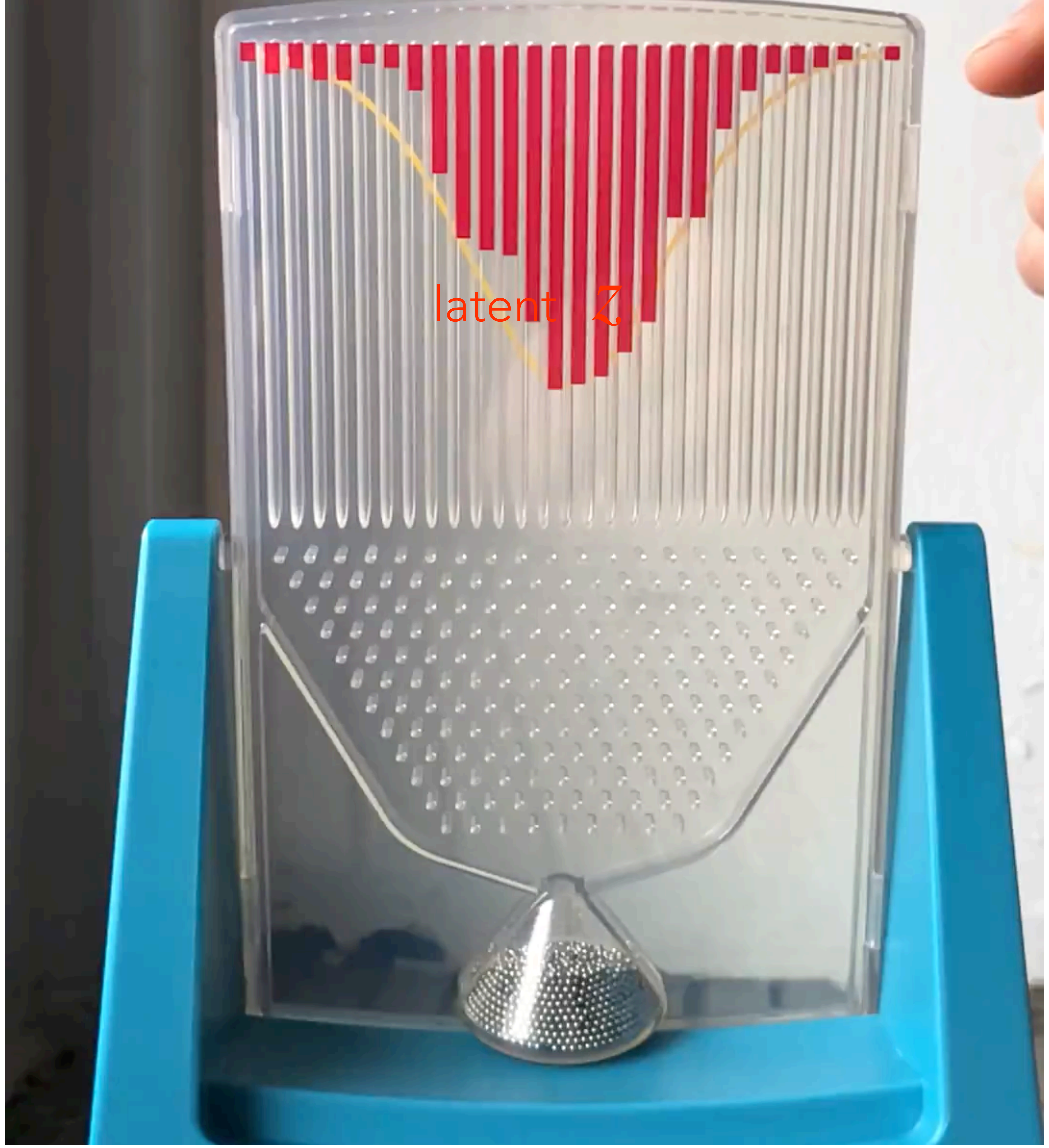


observe x

A toy example

Imagine the entire board is slightly tilted, which biases the probability to bounce left/right.

Say we want to infer θ, the probability to bounce right based on distribution of x



The probability of ending in bin x corresponds to the total probability of all the paths z from start to x.

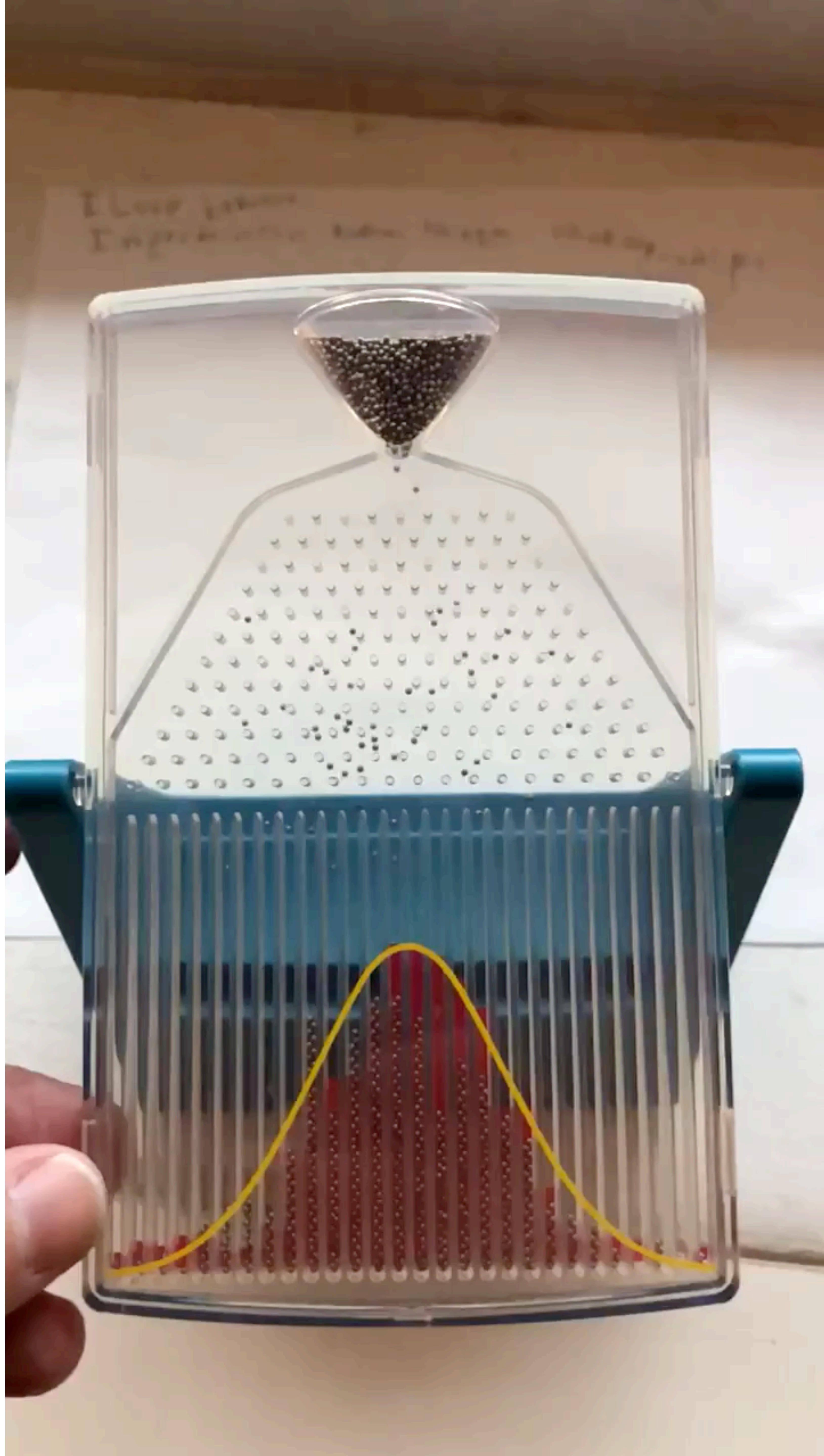
$$p(x|\theta) = \int p(x,z|\theta)dz = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

Uh oh!

The actual situation is much more complicated.

It's not a Binomial distribution!

What is it?

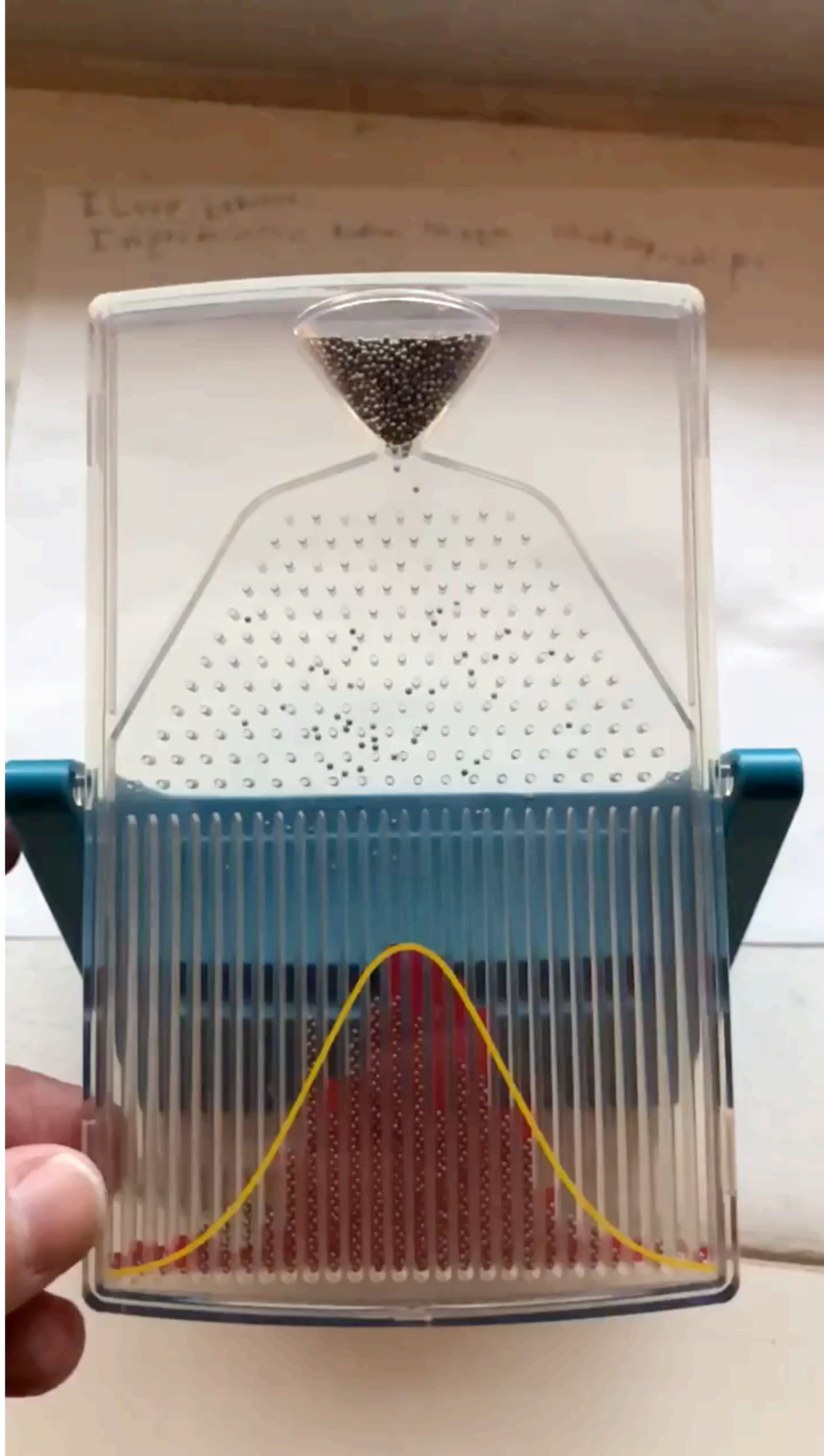


Uh oh!

The actual situation is much more complicated.

It's not a Binomial distribution!

What is it?



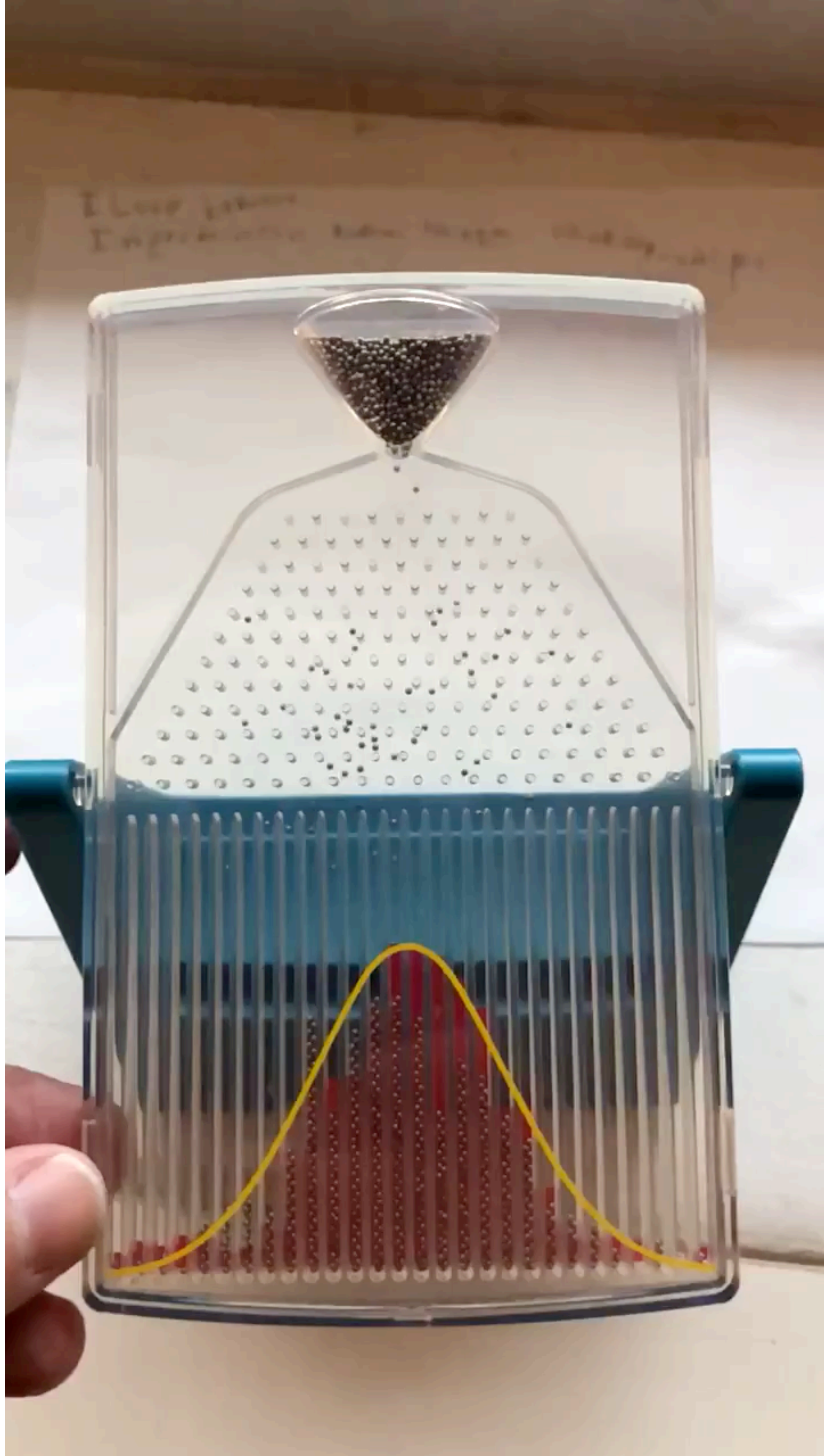
Uh oh!

The actual situation is much more complicated.

It's not a Binomial distribution!

What is it?

I have no idea, but I can **simulate it!**

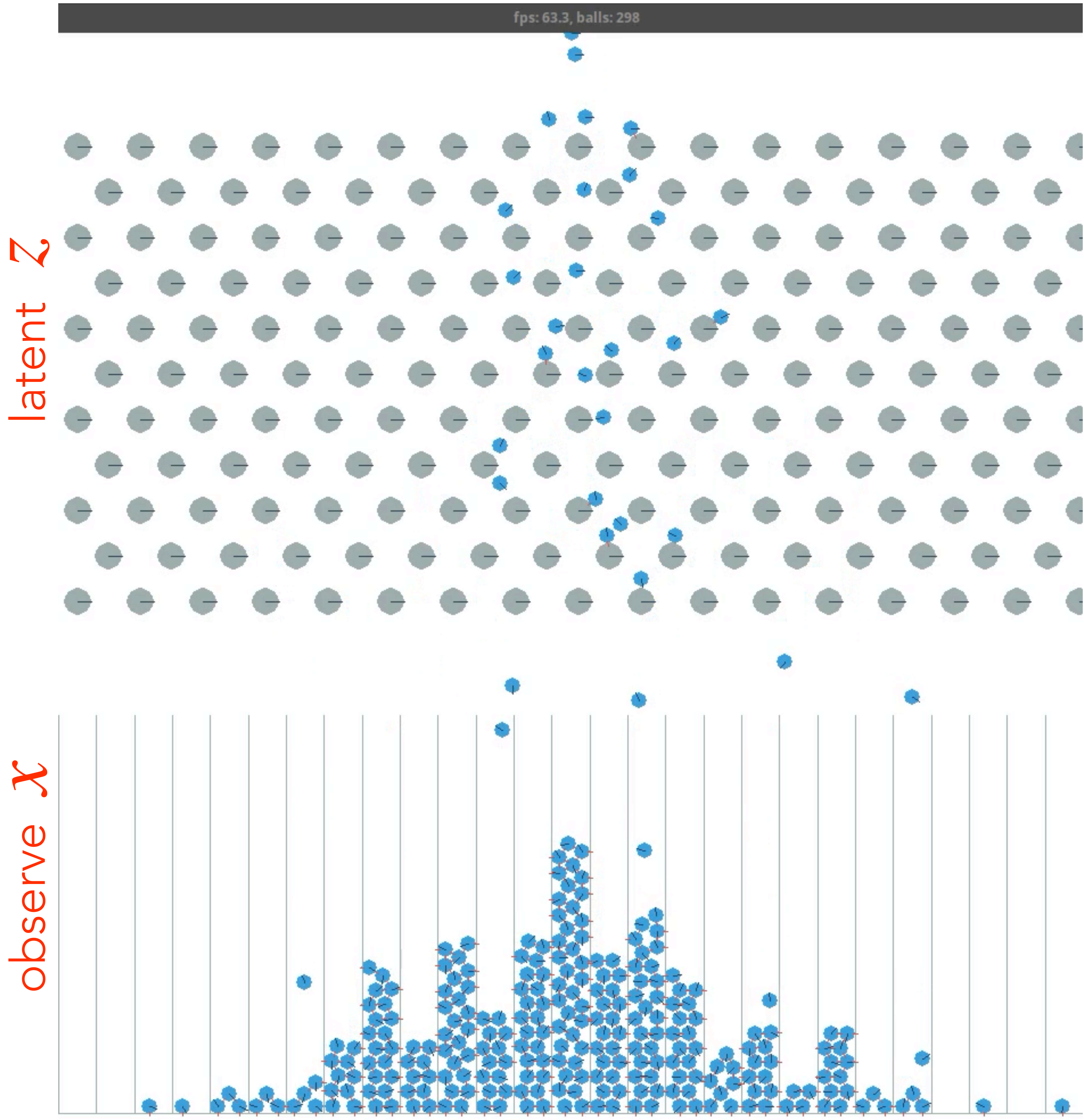


Properties of simulators

Two broad classes:

- Deterministic evolution of initial state
  - (eg. differential equations, fluid dynamics, N-body simulations, etc.)
- Stochastic evolution
  - (eg. Markov processes, molecular dynamics, Gibbs / Boltzmann distribution in statistical mechanics, stochastic differential equations, etc.)

ANIMATION BY ATILIM GÜNEŞ BAYDIN

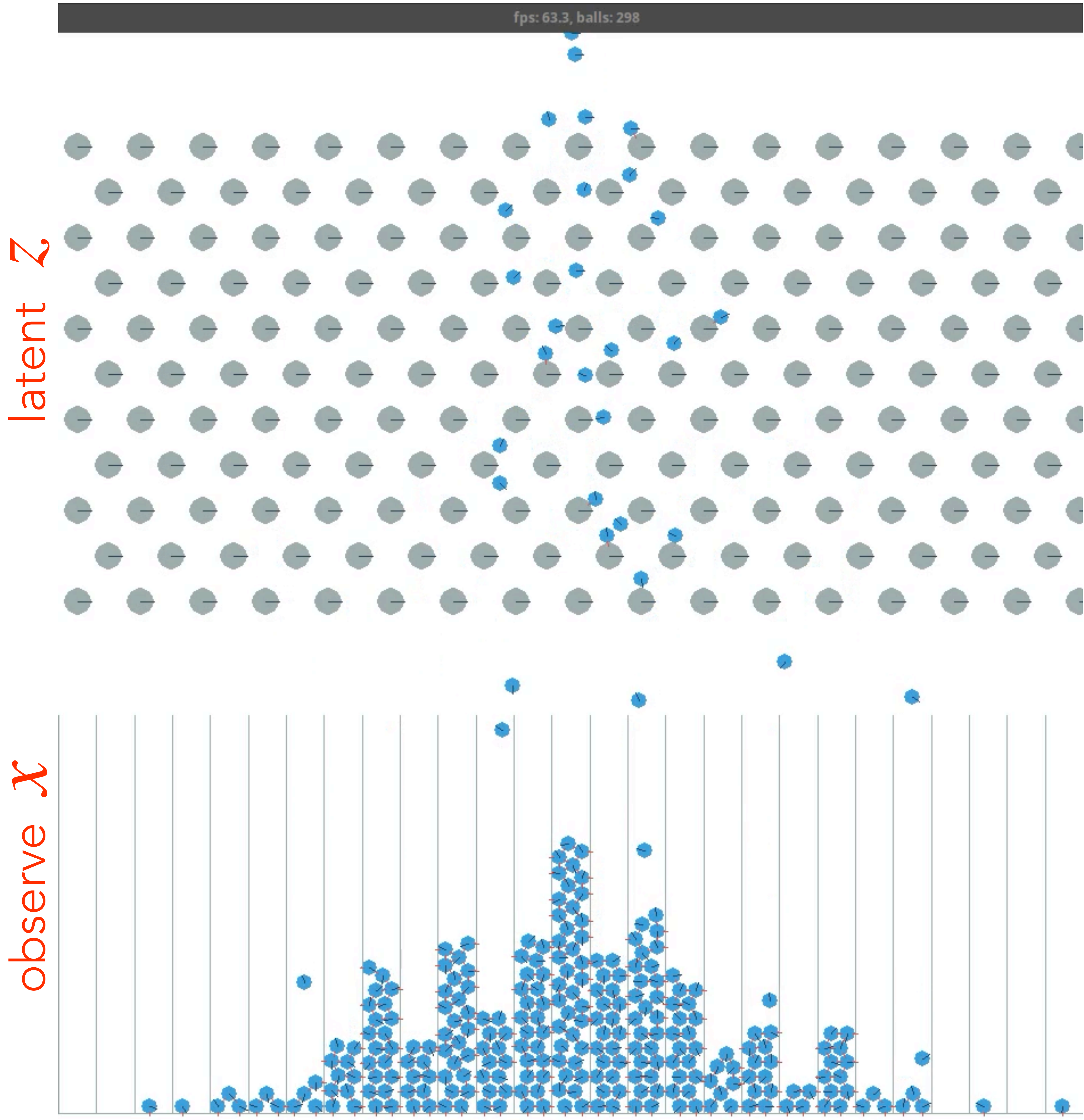


Integral over latent variables is typically **intractable** $p(x|\theta) = \int p(x, z \mid \theta) \mathrm{d}z$

Properties of simulators

Two broad classes:

- Deterministic evolution of initial state
  - (eg. differential equations, fluid dynamics, N-body simulations, etc.)
- Stochastic evolution
  - (eg. Markov processes, molecular dynamics, Gibbs / Boltzmann distribution in statistical mechanics, stochastic differential equations, etc.)



Integral over latent variables is typically **intractable** $p(x|\theta) = \int p(x, z \mid \theta) \mathrm{d}z$

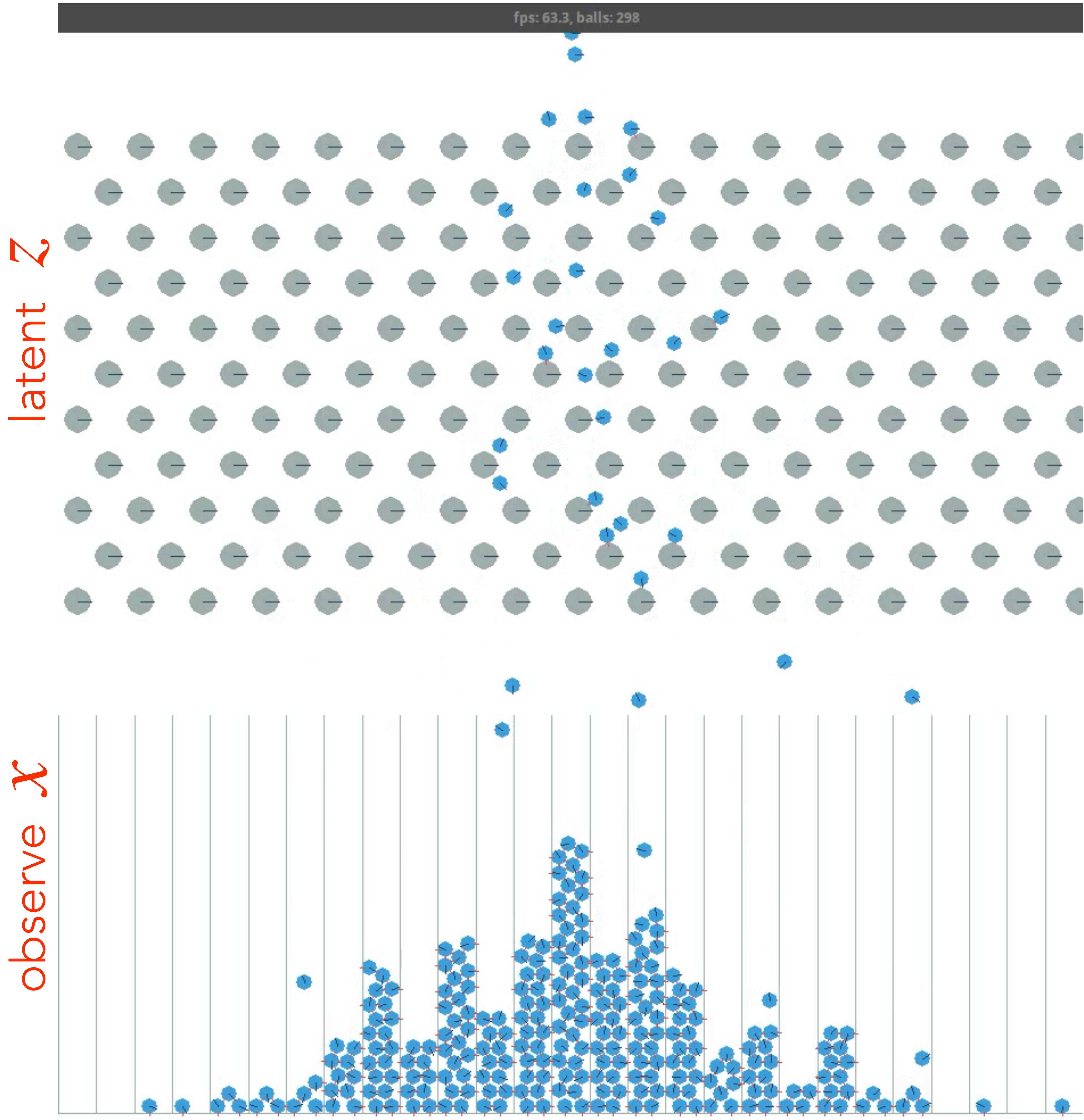
An example

The probability of landing in a bin x corresponds to cumulative probability of all the latent paths z that end in x

$$p(x|\theta) = \int p(x, z \mid \theta) \mathrm{d}z$$

- But the integral (sum) can no longer be simplified analytically
- As the latent space grows, the number of possible paths grows rapidly.
- The integral becomes **intractable**
- But generating synthetic observations remains easy

ANIMATION BY ATILIM GÜNEŞ BAYDIN



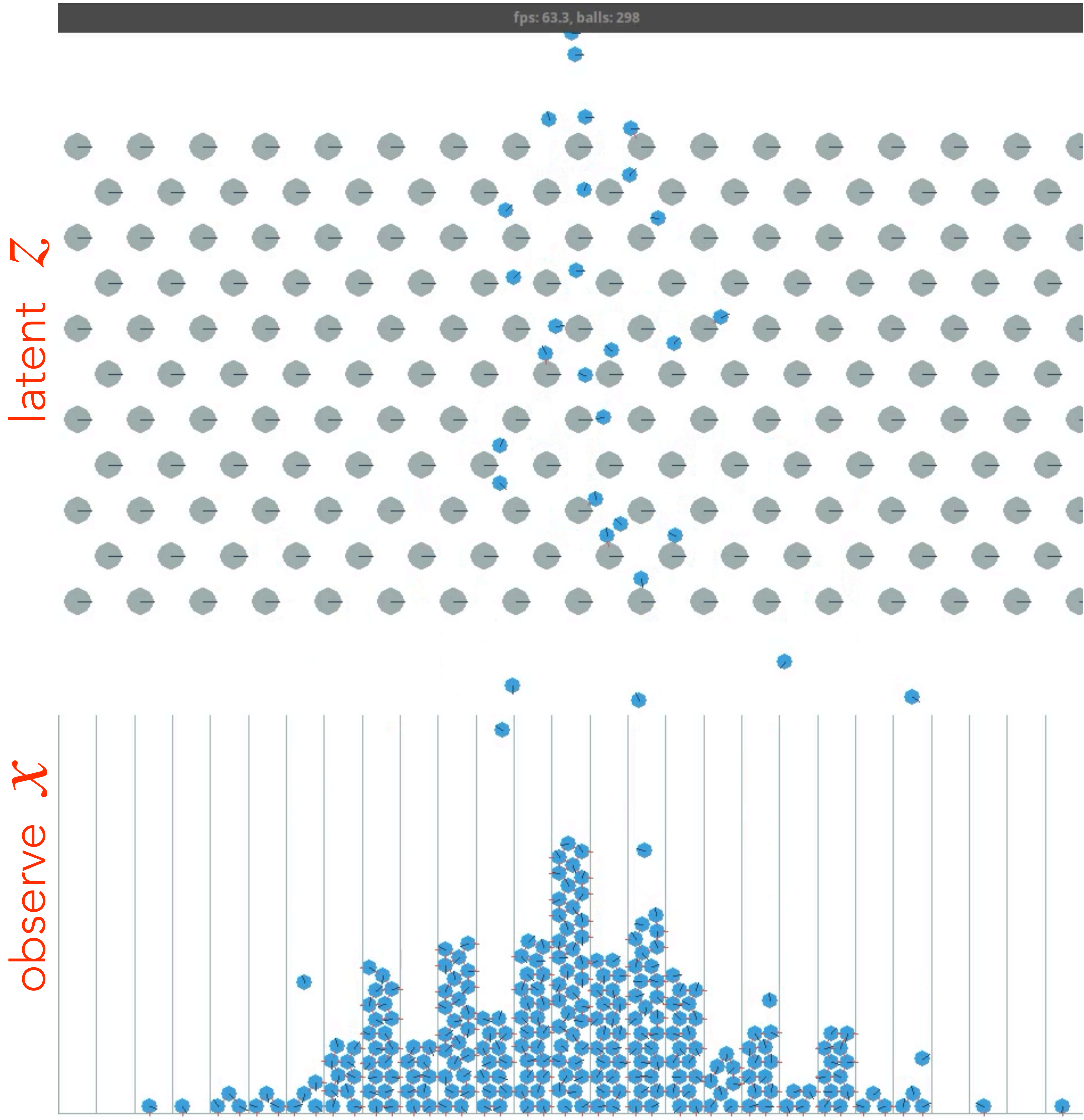
An example

The probability of landing in a bin x corresponds to cumulative probability of all the latent paths z that end in x

$$p(x|\theta) = \int p(x, z \mid \theta) \mathrm{d}z$$

- But the integral (sum) can no longer be simplified analytically
- As the latent space grows, the number of possible paths grows rapidly.
- The integral becomes **intractable**
- But generating synthetic observations remains easy

ANIMATION BY ATILIM GÜNEŞ BAYDIN



A rose by any other name

This motivates a class of inference methods for a stochastic simulator where

- evaluating the **likelihood is intractable**, but
- it is **possible to sample** synthetic data $x \sim p(x \mid \theta)$

This setting is often referred to as **likelihood-free inference**, but I prefer the term **simulation-based inference** because usually one approximates the likelihood (or likelihood ratio) and then use established inference techniques

- applies to both Bayesian or Frequentist inference

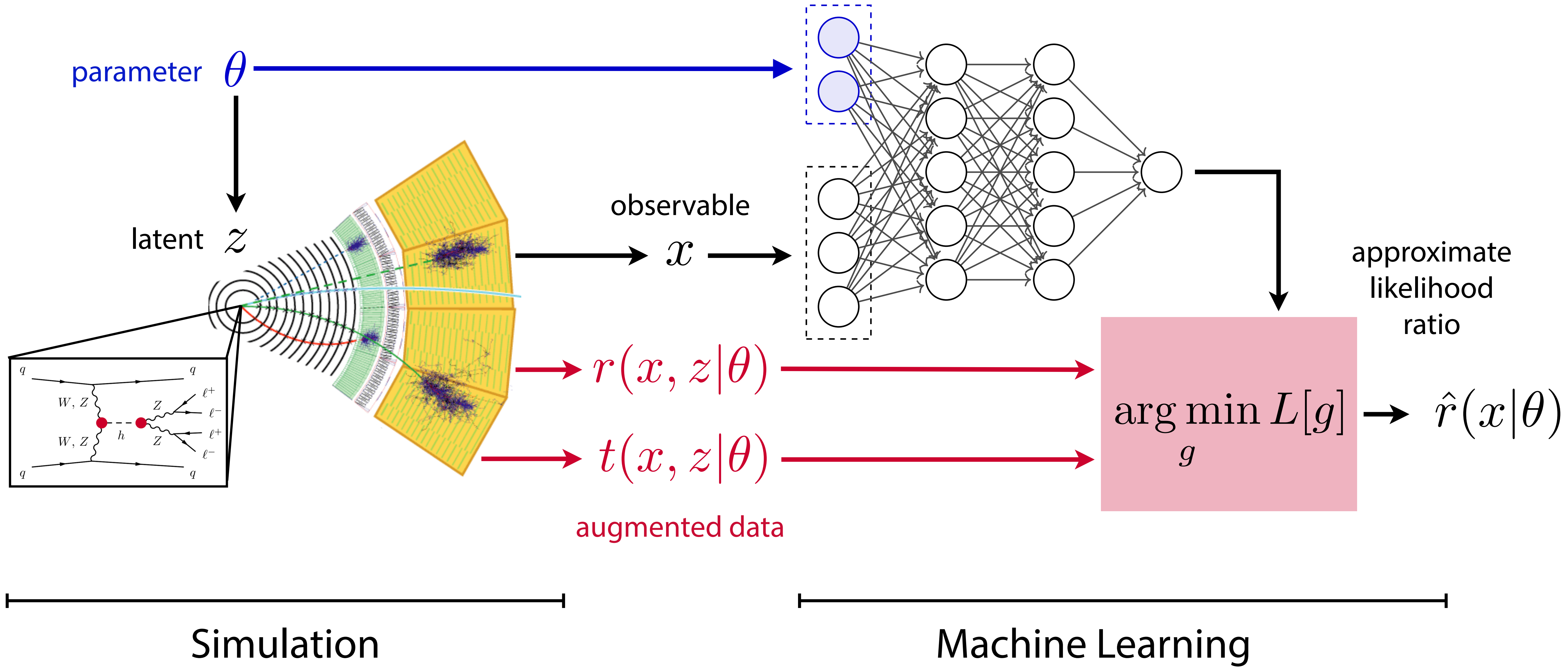
Gold mining: augmenting the training data

Sample efficiency is a major concern for these methods as many simulators are computationally expensive

Gold mining: augmenting the training data

Sample efficiency is a major concern for these methods as many simulators are computationally expensive

Recently, we realized we can **extract more from the simulator**.
We can use **augmented data** to improve training




Johann Brehmer

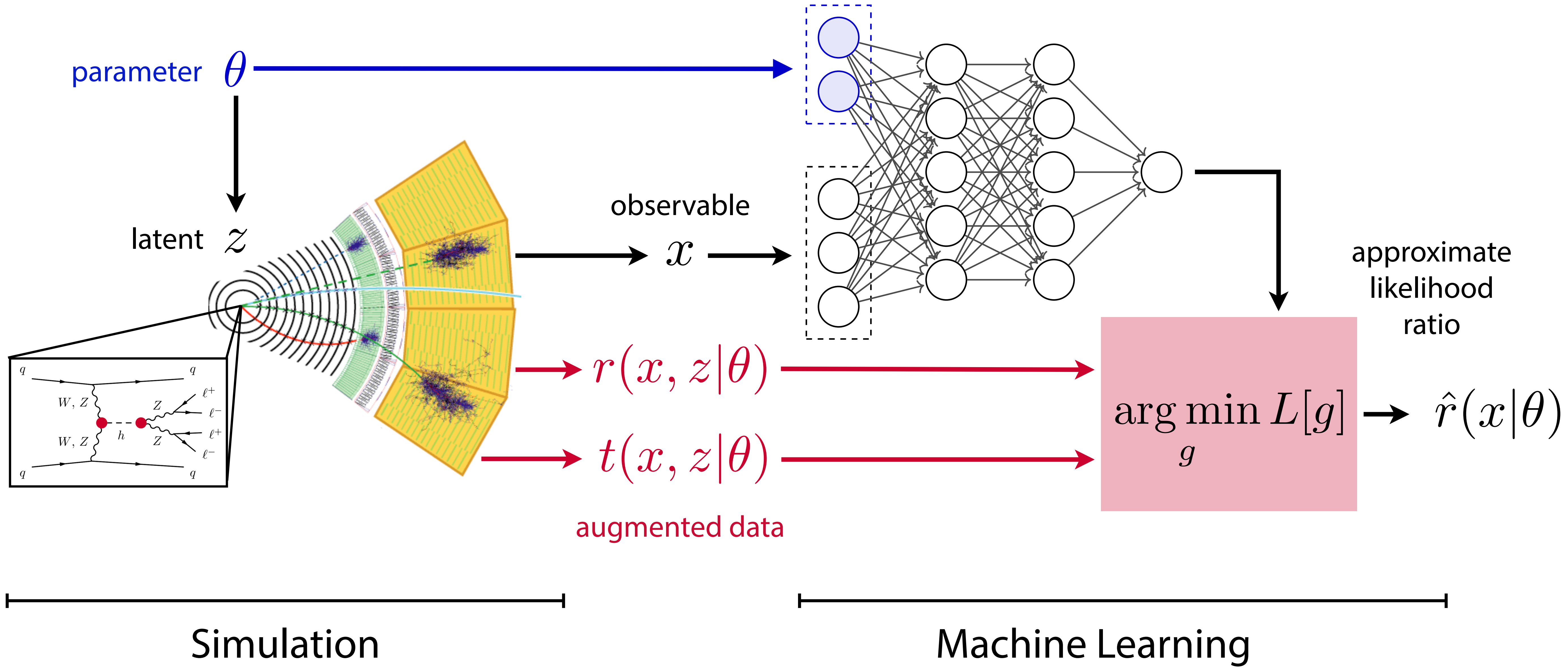

Gilles Louppe

Brehmer, Louppe, Pavez, KC, PNAS (2019), arXiv:1805.12244
See also Wenliang, Moskovitz, Kanagawa, Sahani, ICML2020

Gold mining: augmenting the training data

Sample efficiency is a major concern for these methods as many simulators are computationally expensive

Recently, we realized we can **extract more from the simulator**.
We can use **augmented data** to improve training



While implicit density is intractable

$$p(x|\theta) = \int dz p(x, z|\theta)$$

We can **augment the simulator** to calculate some quantities conditioned on latent z, which are tractable:

Joint likelihood ratio:

$$r(x, z|\theta_0, \theta_1) = \frac{p(x, z|\theta_0)}{p(x, z|\theta_1)}$$

and joint score:

$$t(x, z|\theta_0) = \frac{\nabla_\theta p(x, z|\theta)|_{\theta_0}}{p(x, z|\theta_0)} = \nabla_\theta \log p(x, z|\theta)|_{\theta_0}$$


Johann Brehmer

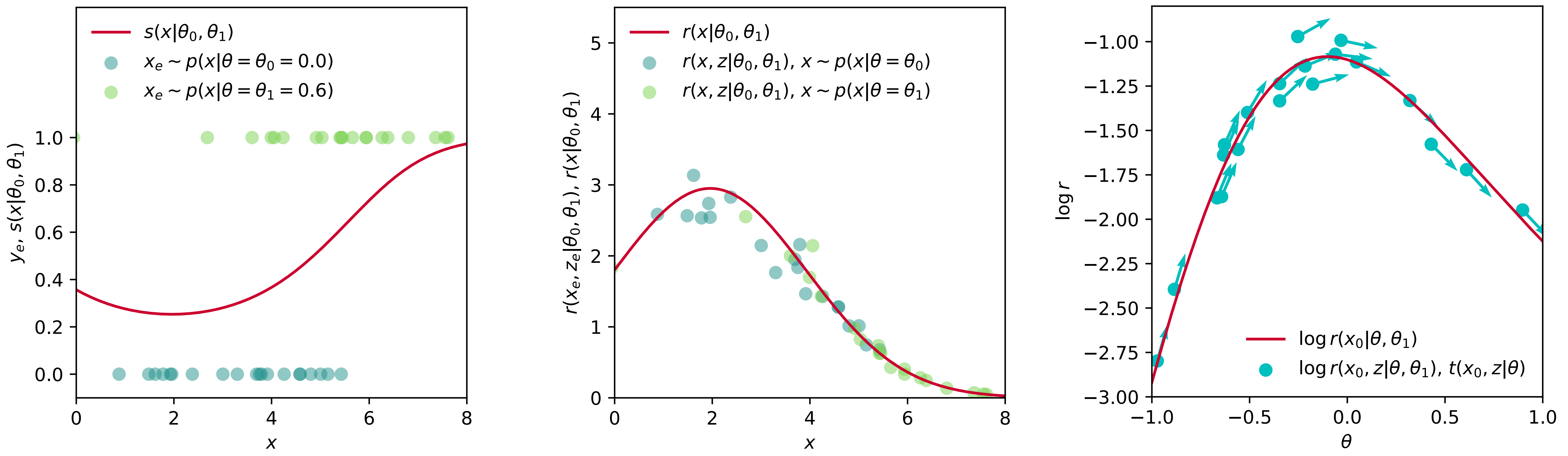

Gilles Louppe

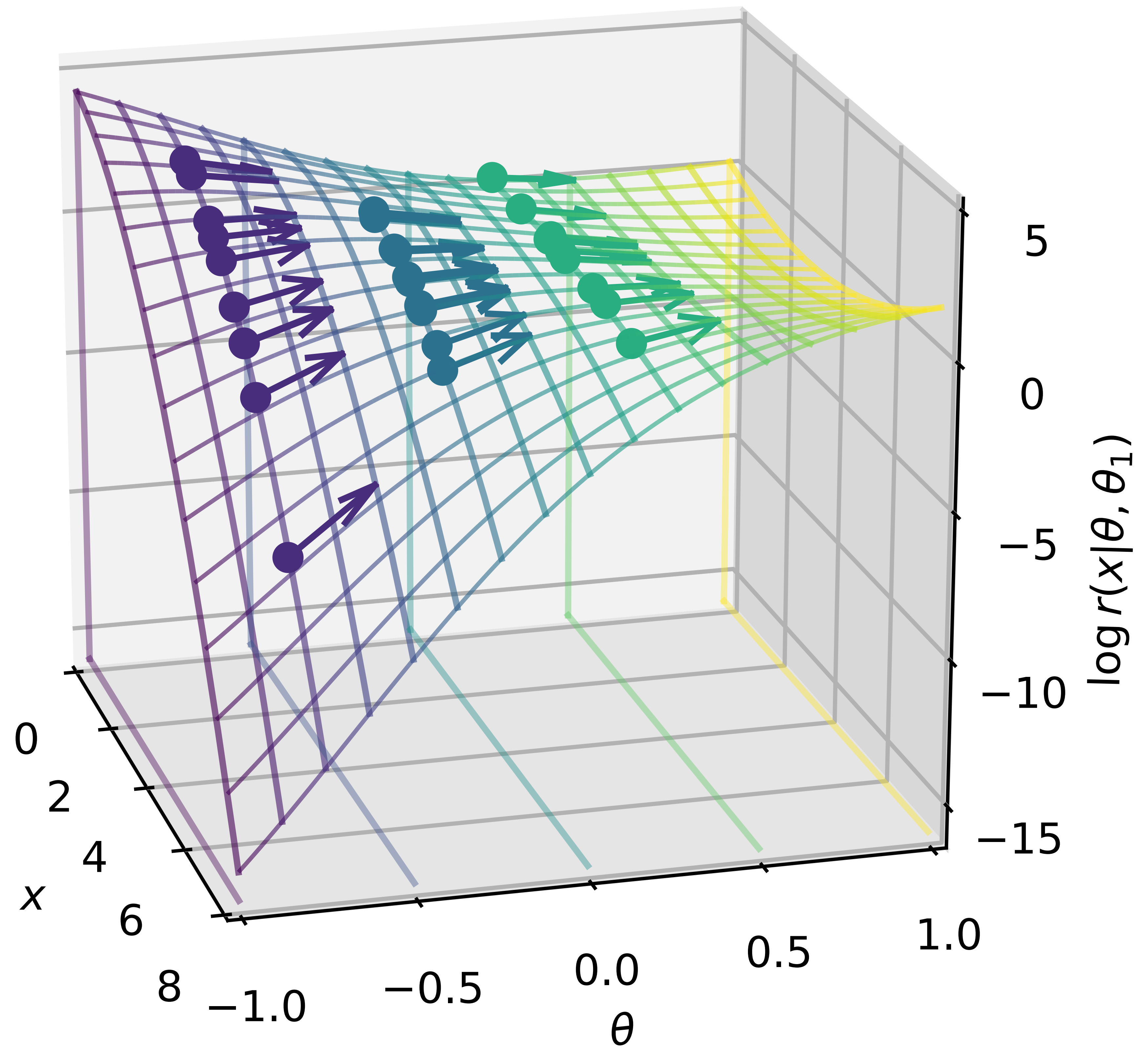
Brehmer, Louppe, Pavez, KC, PNAS (2019), arXiv:1805.12244
See also Wenliang, Moskovitz, Kanagawa, Sahani, ICML2020

Gold mining: augmenting the training data

The augmented training data converts supervised classification into supervised regression with lower variance

- improvement in training efficiency

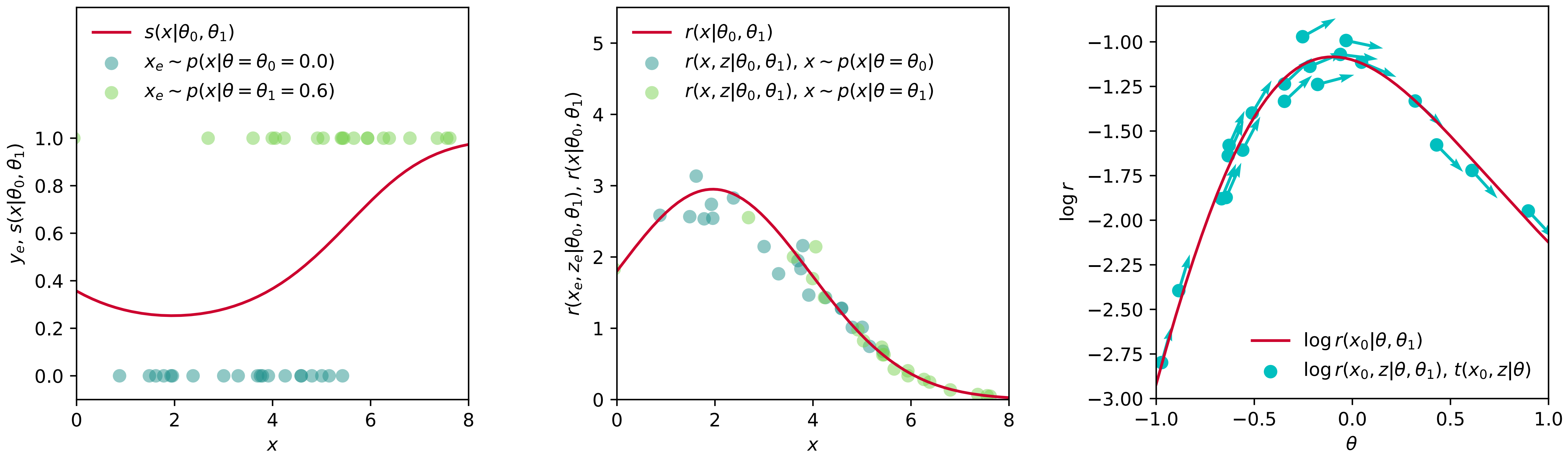


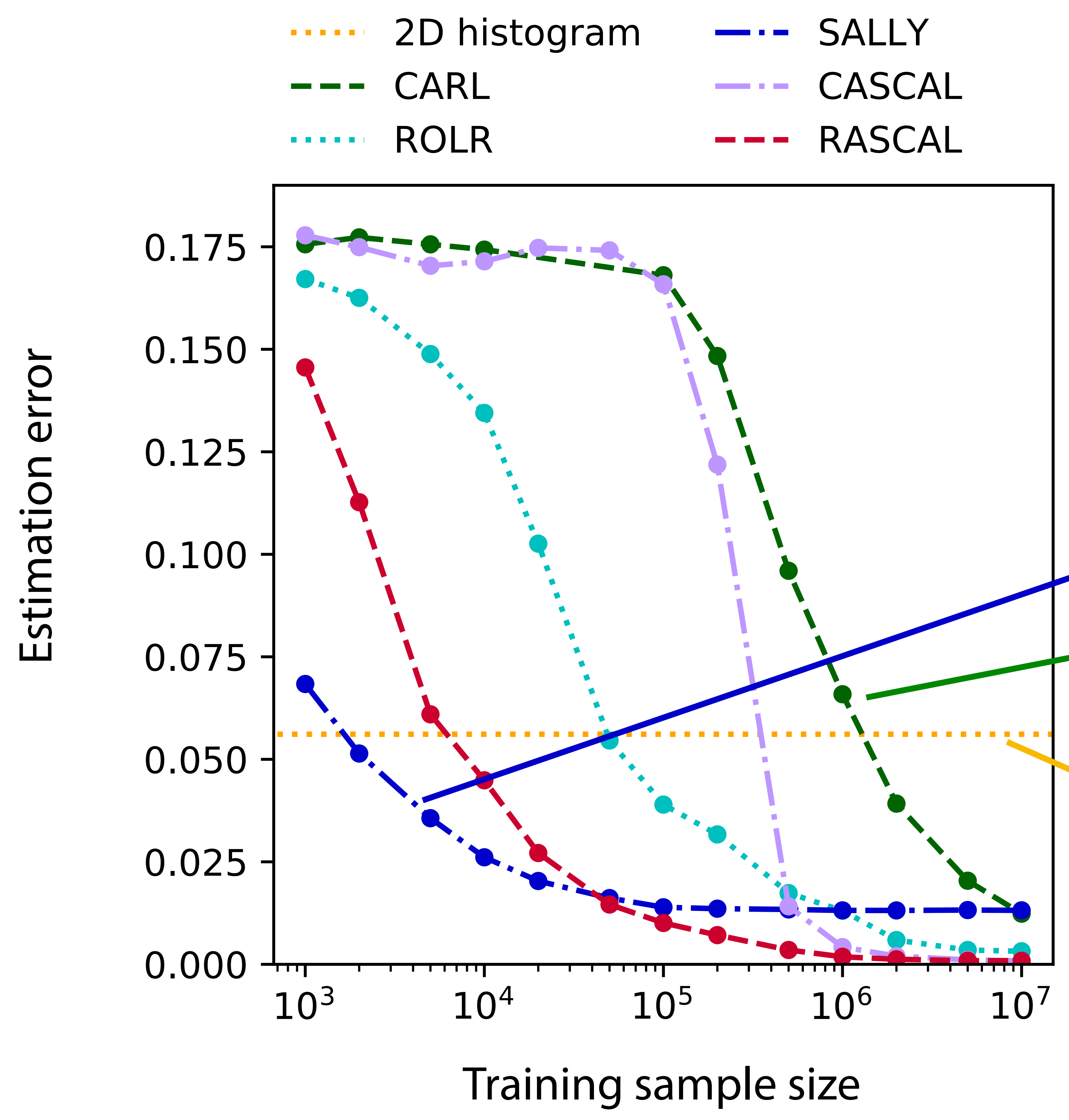


Gold mining: augmenting the training data

The augmented training data converts supervised classification into supervised regression with lower variance

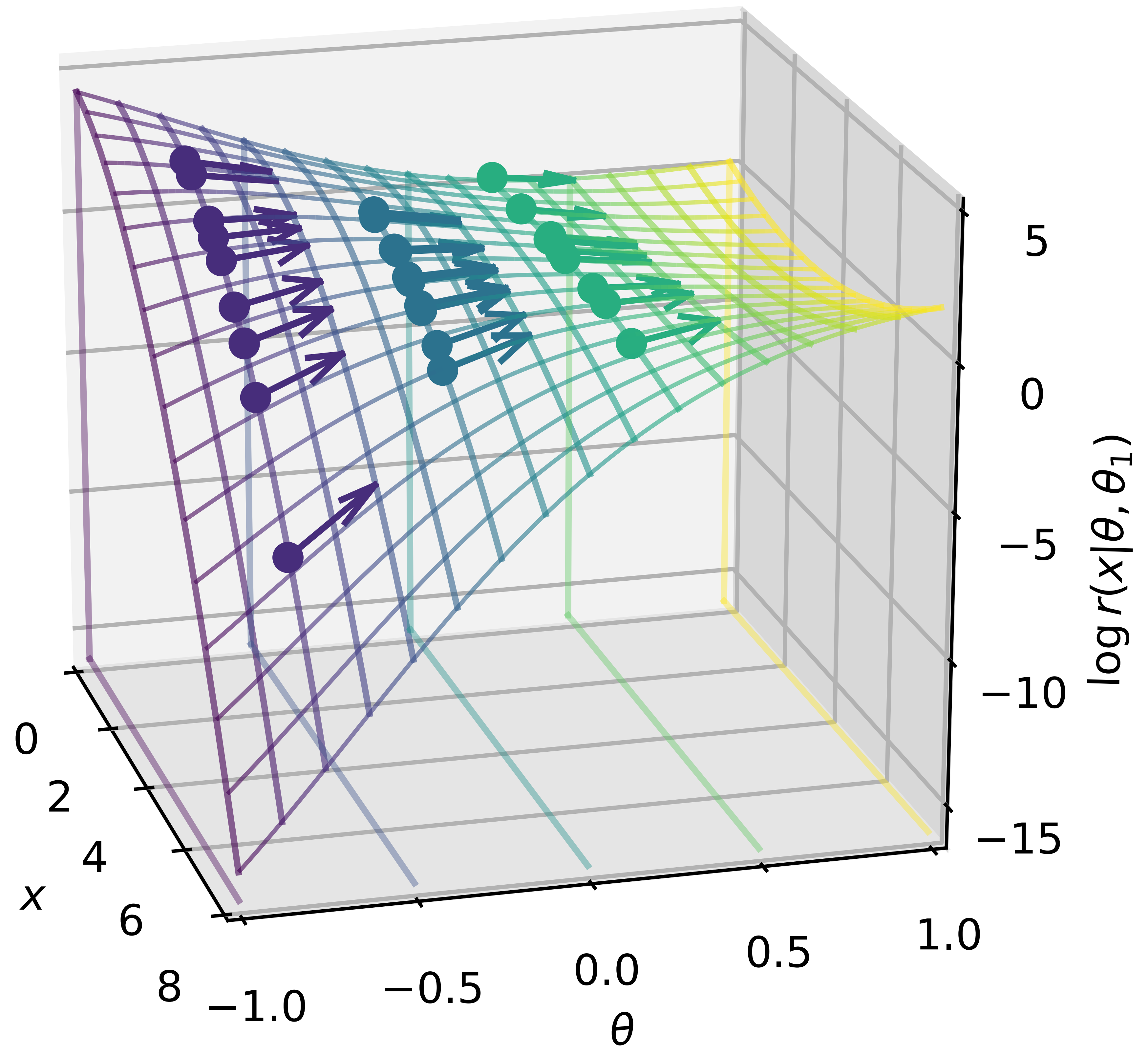
- improvement in training efficiency





New techniques require less data than without augmented data

Traditional Approach no NN



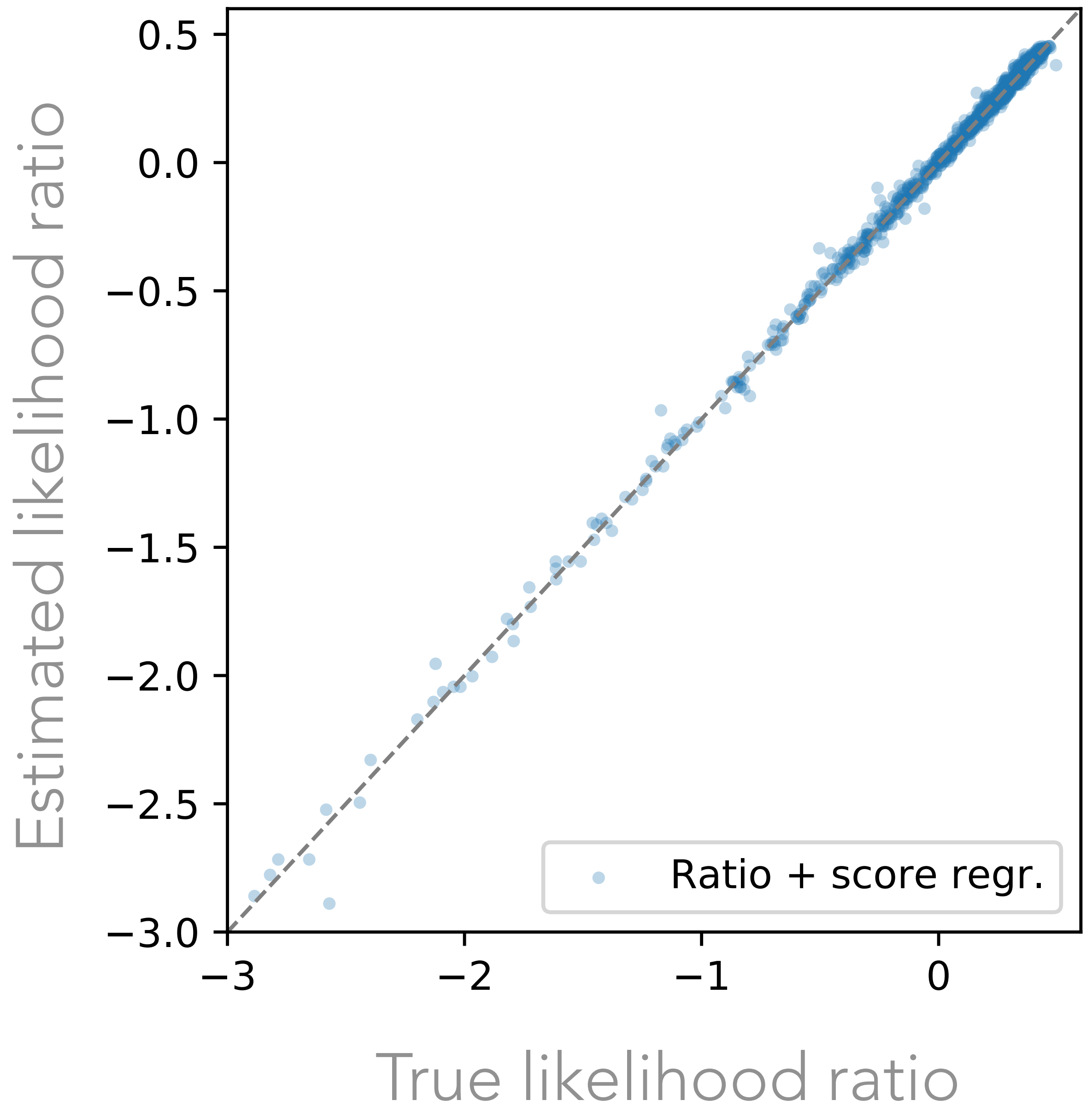
Brehmer, Louppe, Pavez, KC, PNAS (2019), arXiv:1805.12244

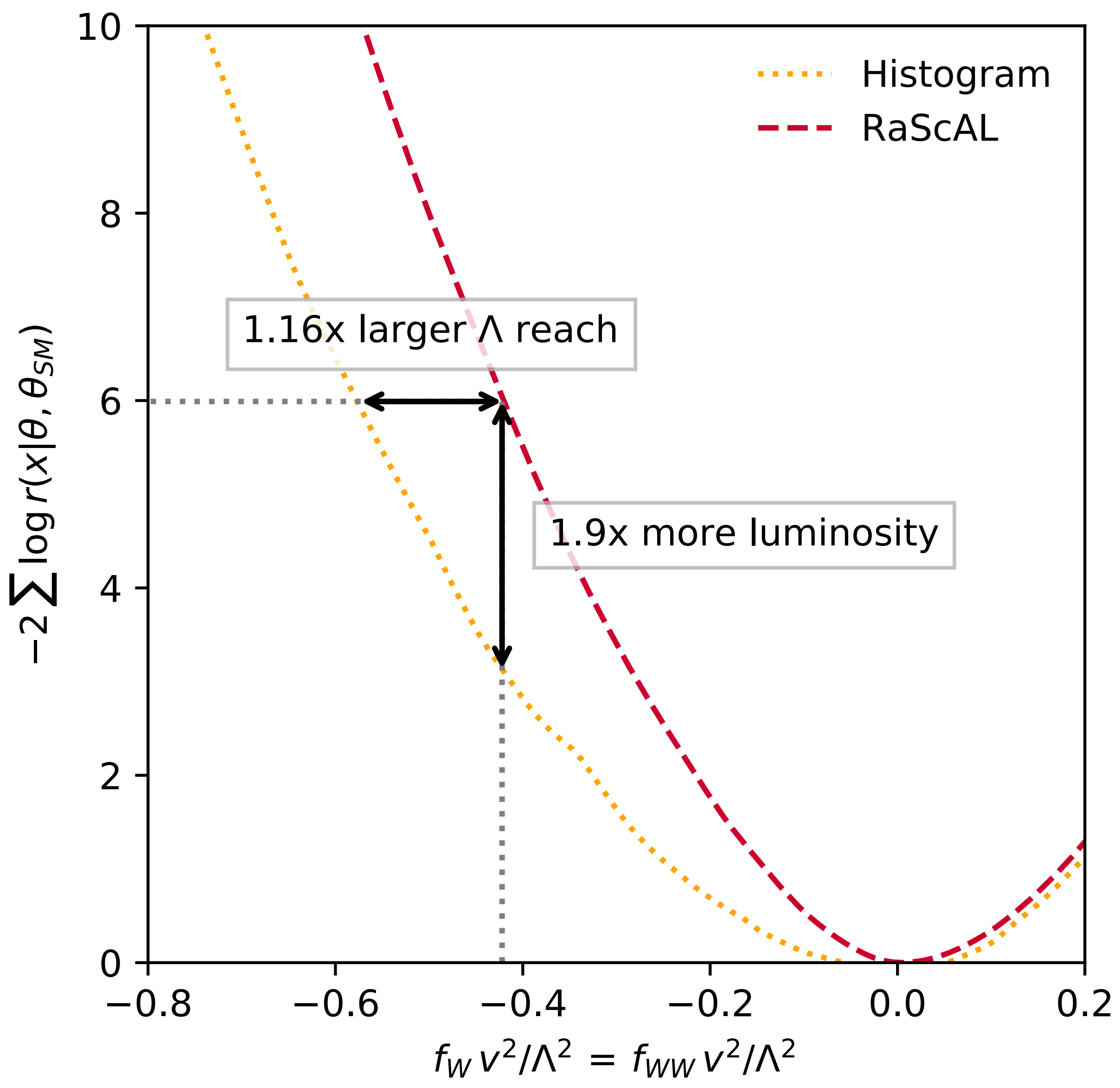
Impact on science: The Higgs boson

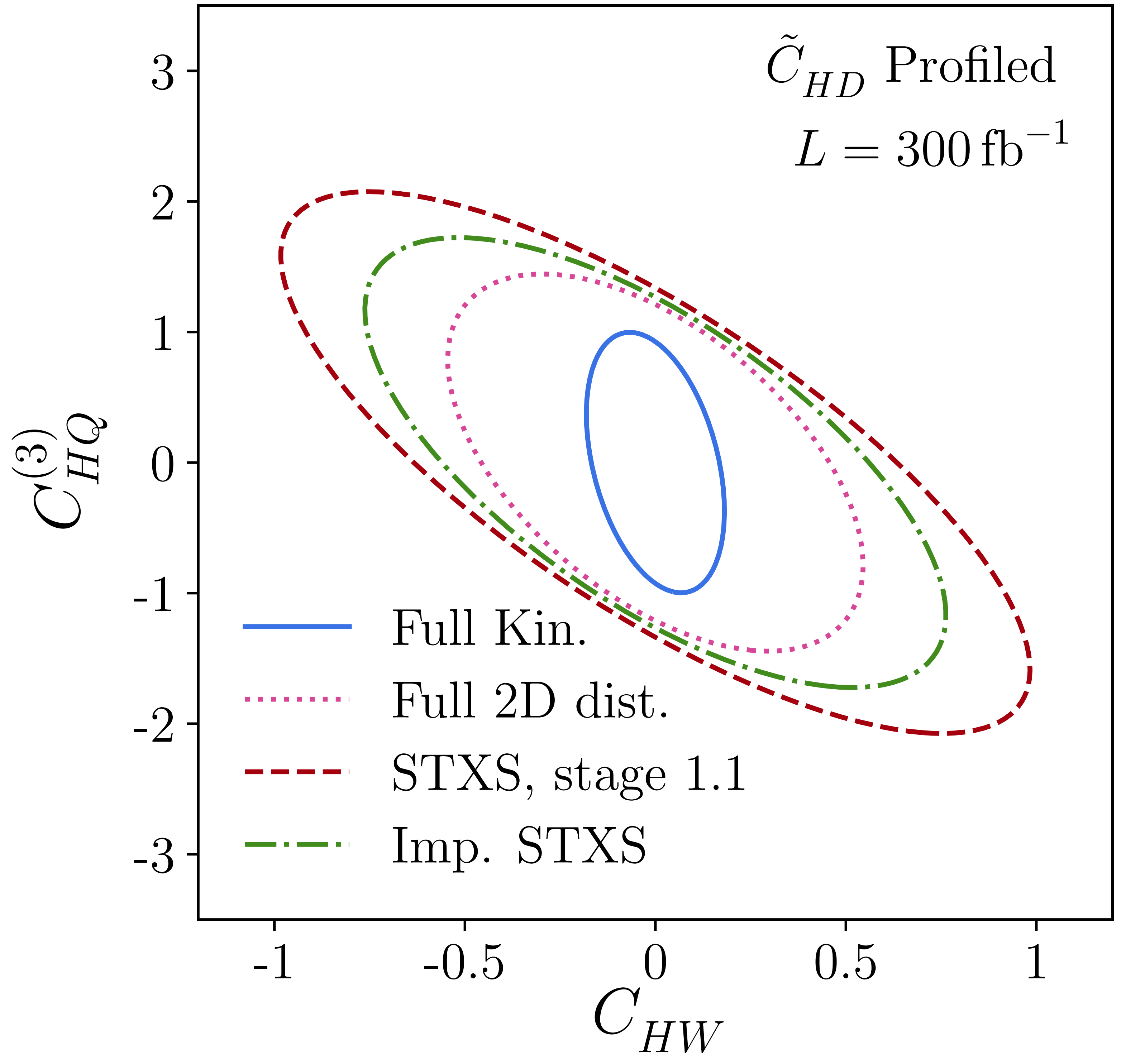
Massive gains in precision of a flagship measurement at the LHC !

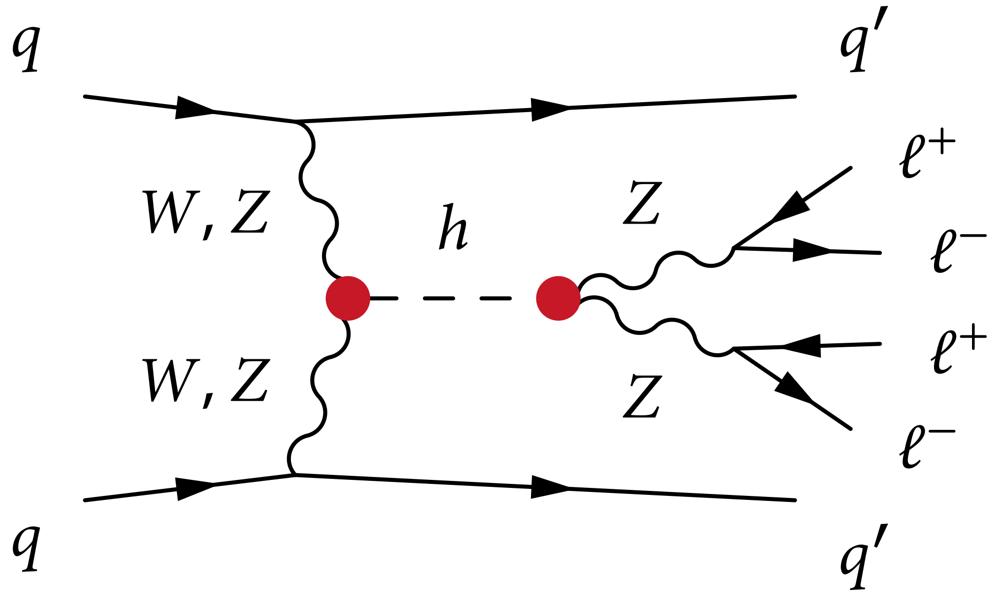
Equivalent to increasing data collected by LHC by several factors

(based on a 42-Dim observation ***x***)









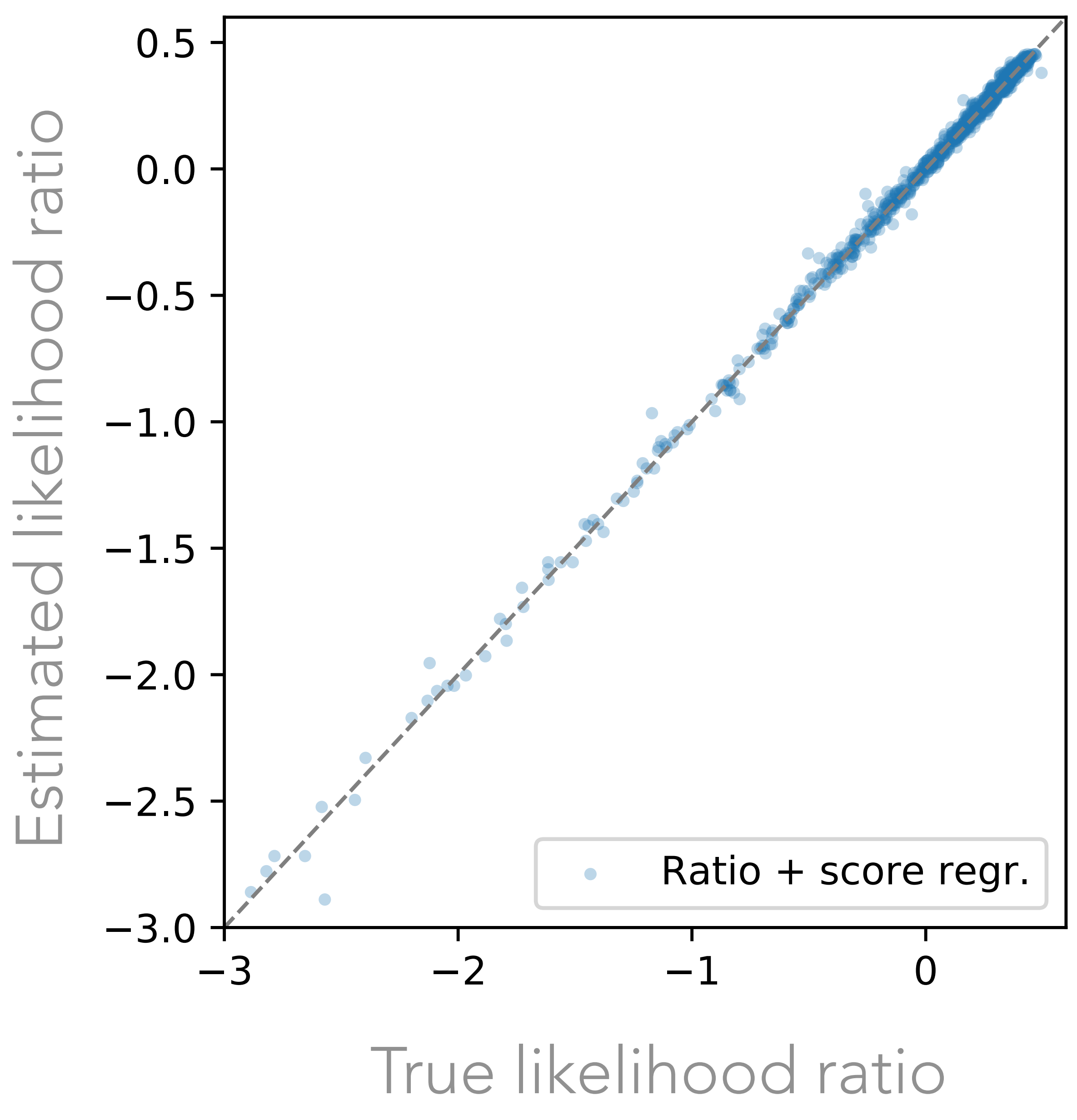
[J. Brehmer, S. Dawson, S. Homiller, F. Kling, T. Plehn 1908.06980]
J Brehmer, J Pavez, G Louppe, K.C. PRL & PRD 2018 [arXiv:1805.00013 & arXiv:1805.00020]
"Better Higgs Measurements Through Information Geometry" [arXiv:1612.05261] & CARL [arxiv:1506.02169]

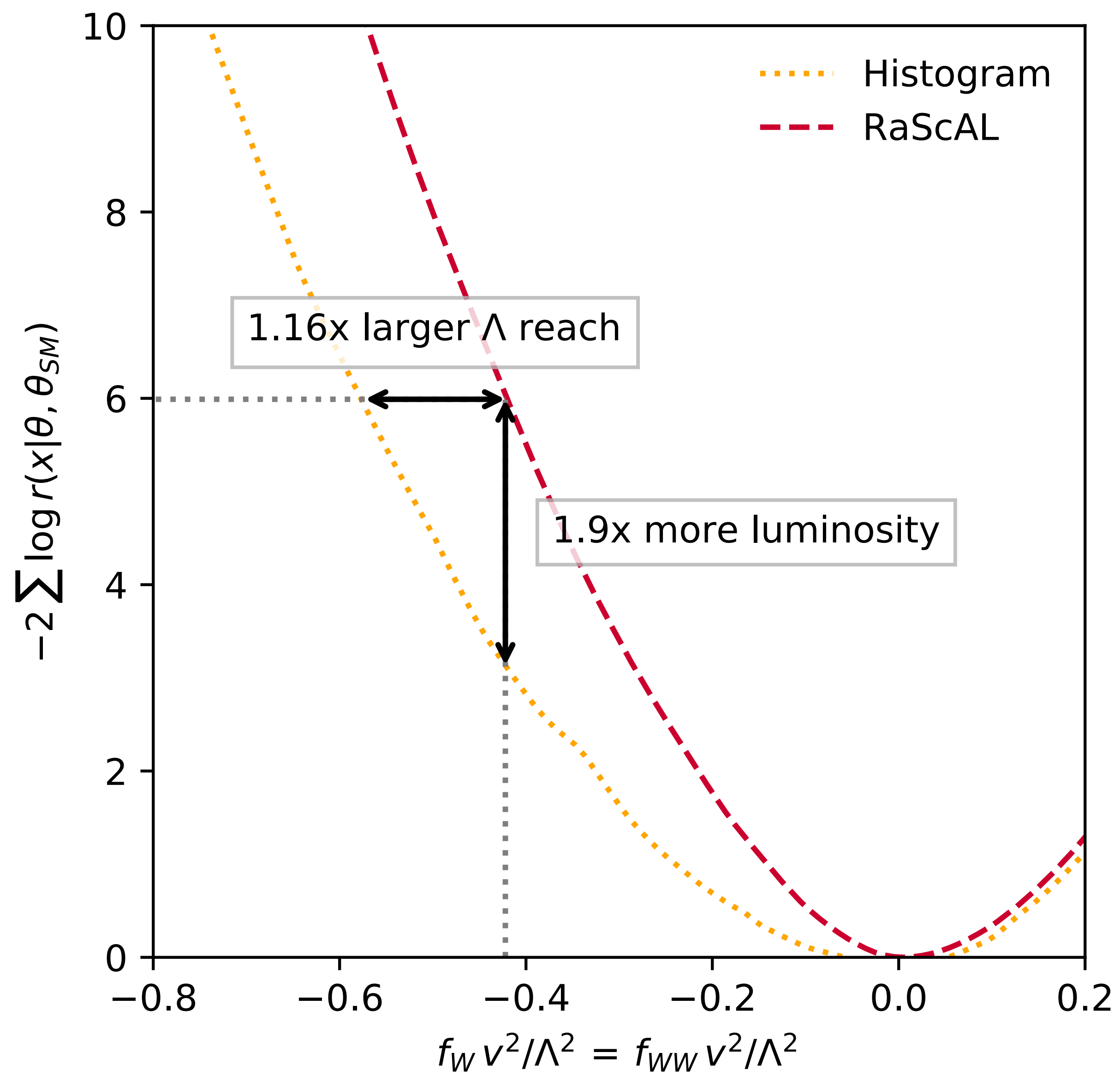
Impact on science: The Higgs boson

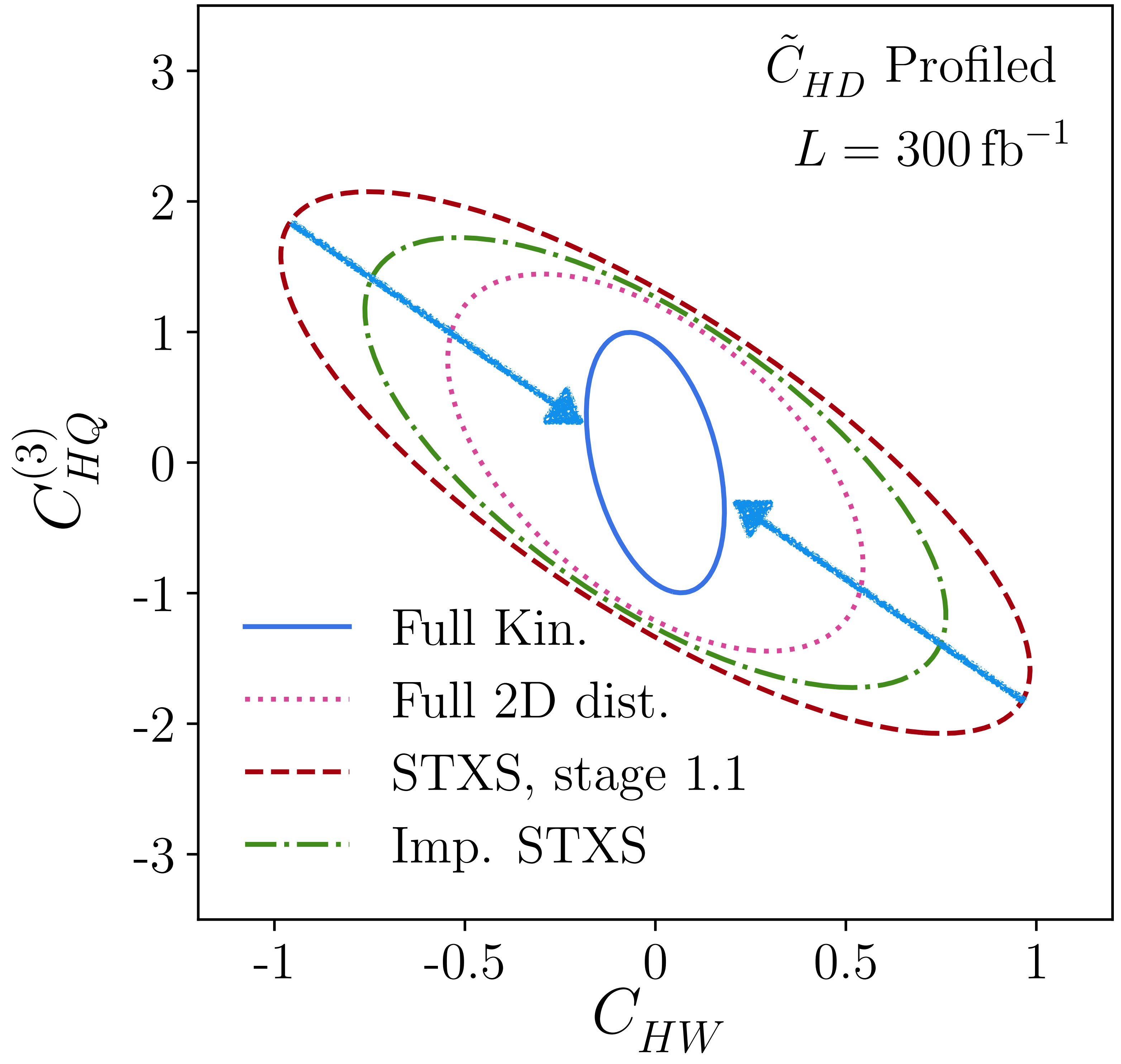
Massive gains in precision of a flagship measurement at the LHC !

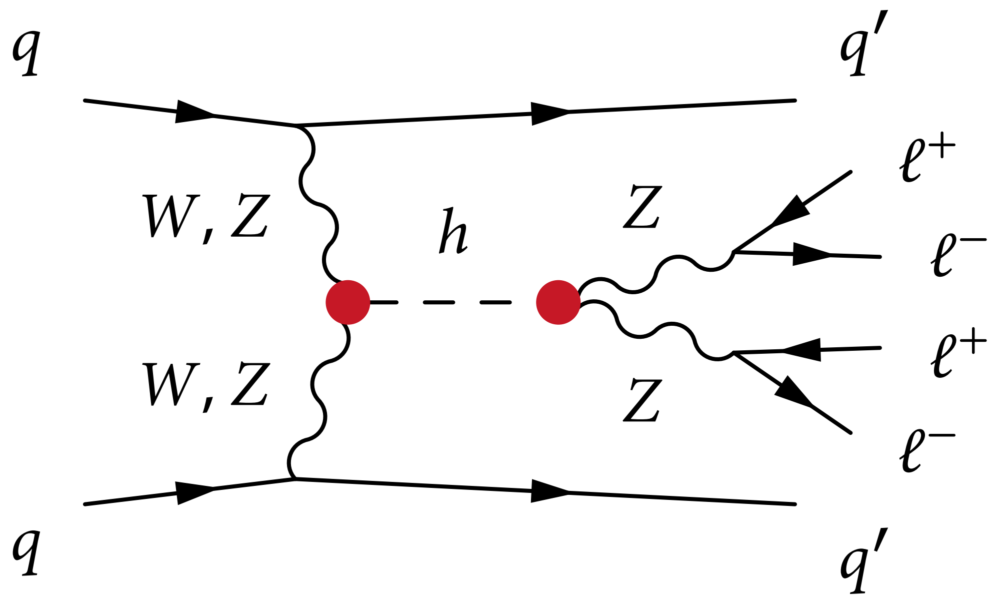
Equivalent to increasing data collected by LHC by several factors

(based on a 42-Dim observation ***x***)



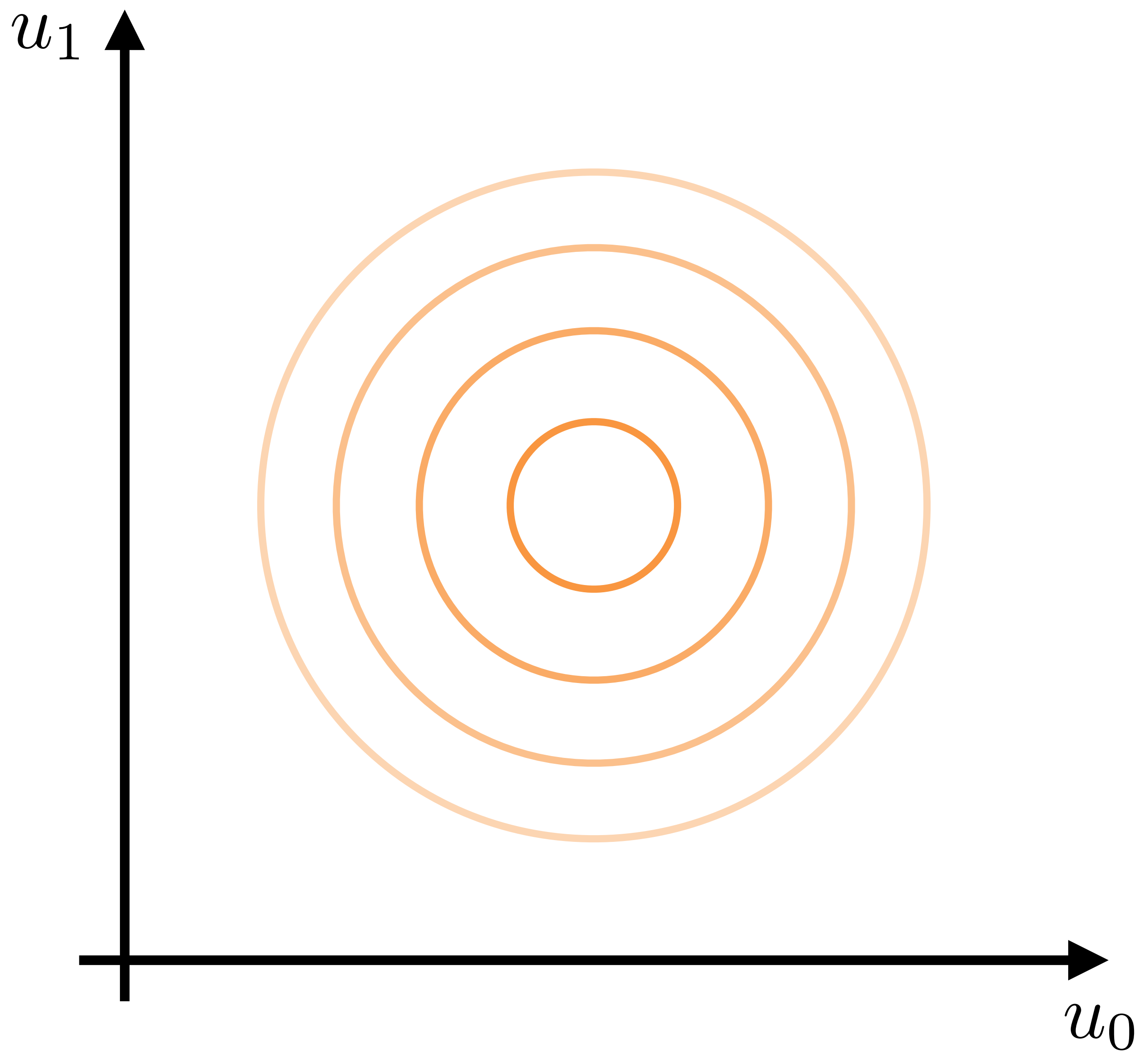


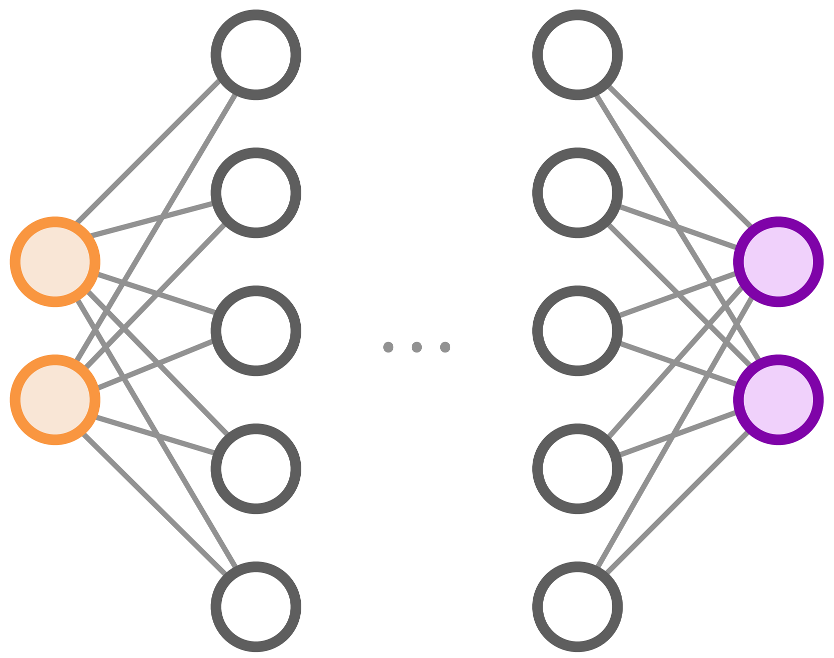


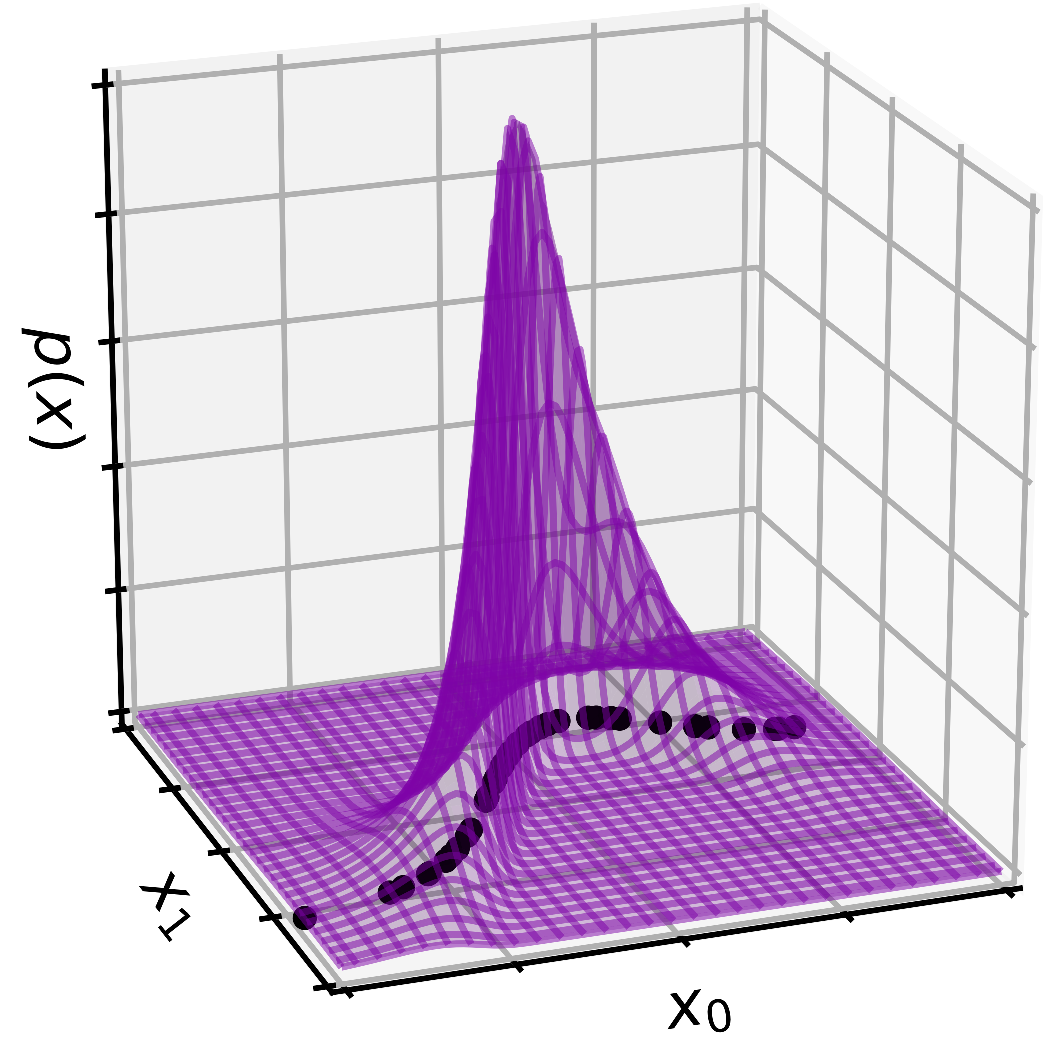


[J. Brehmer, S. Dawson, S. Homiller, F. Kling, T. Plehn 1908.06980]
J Brehmer, J Pavez, G Louppe, K.C. PRL & PRD 2018 [arXiv:1805.00013 & arXiv:1805.00020]
"Better Higgs Measurements Through Information Geometry" [arXiv:1612.05261] & CARL [arxiv:1506.02169]

Normalizing flows in the ambient data space







$u \sim p_u(u)$

f

x

d-dim. latent variables

invertible NN

tractable density over ambient data space

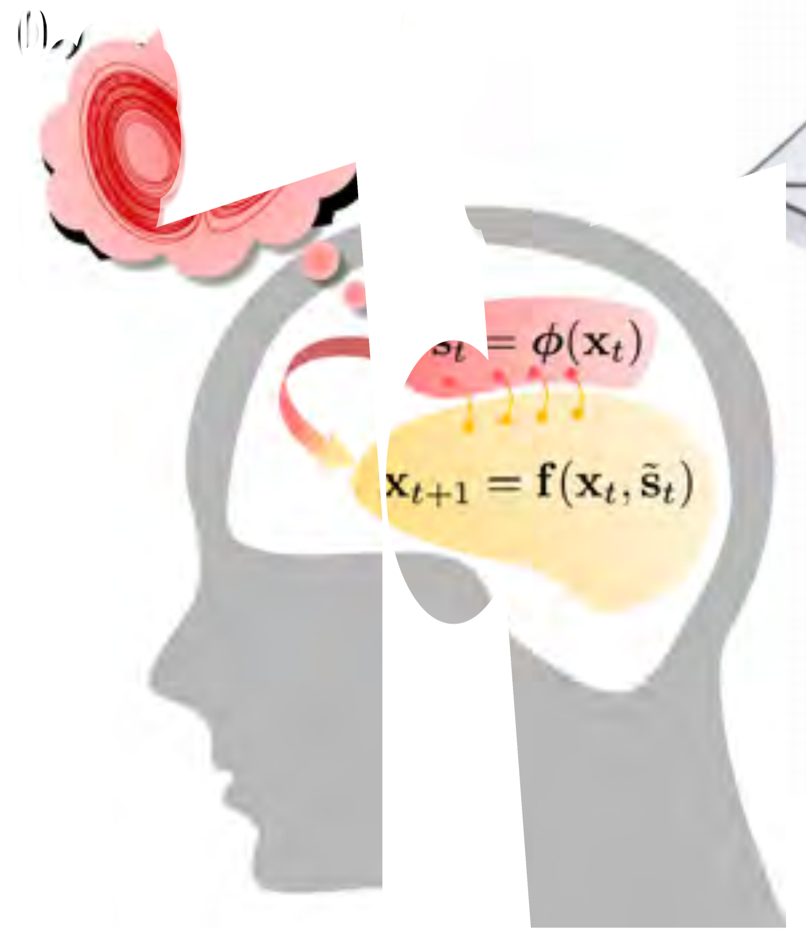
$$p_x(x) = p_u(f^{-1}(x)) \, |\det J_f(f^{-1}(x))|^{-1}$$

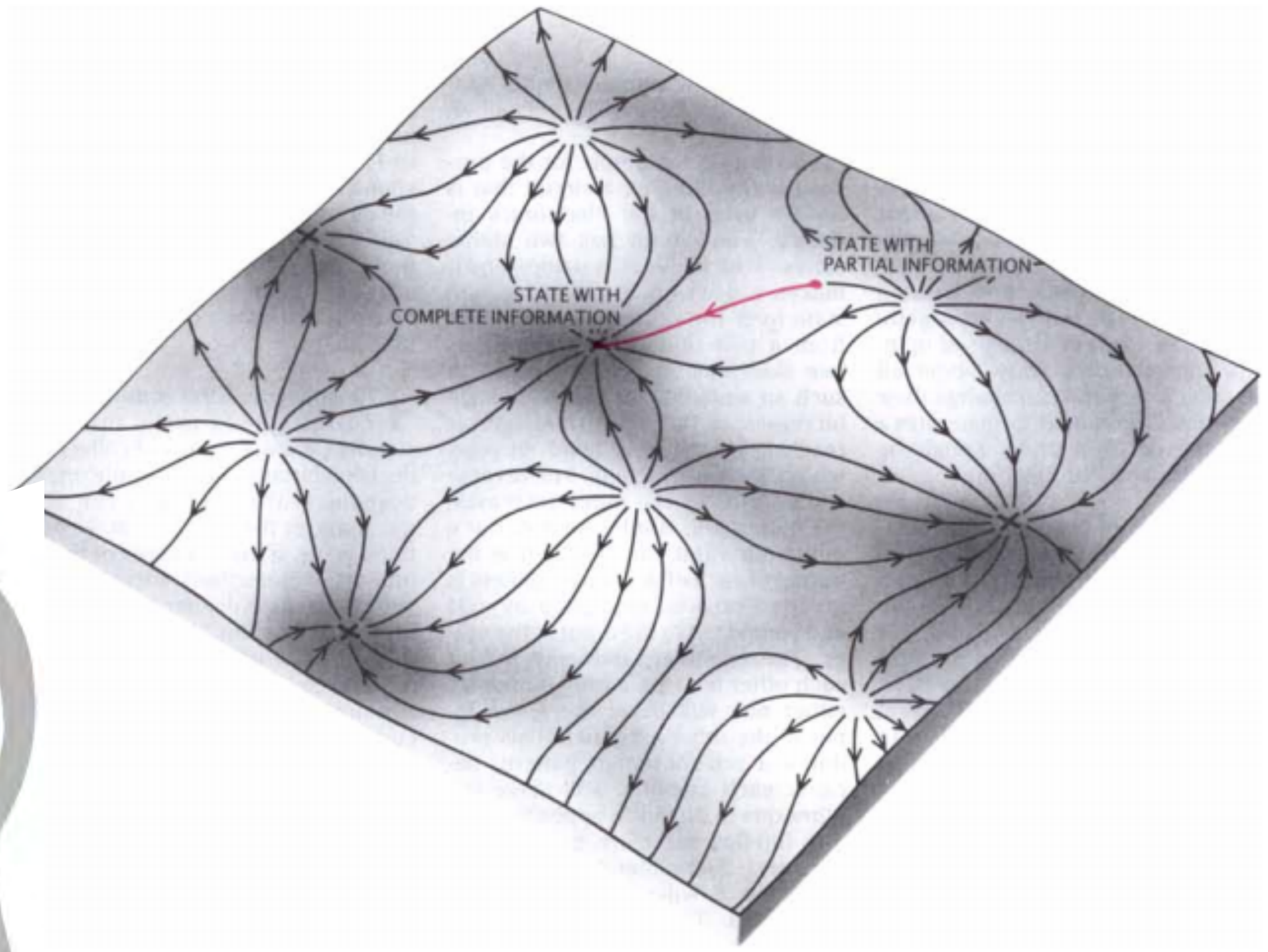
[G. Papamakarios et al 1912.02762]

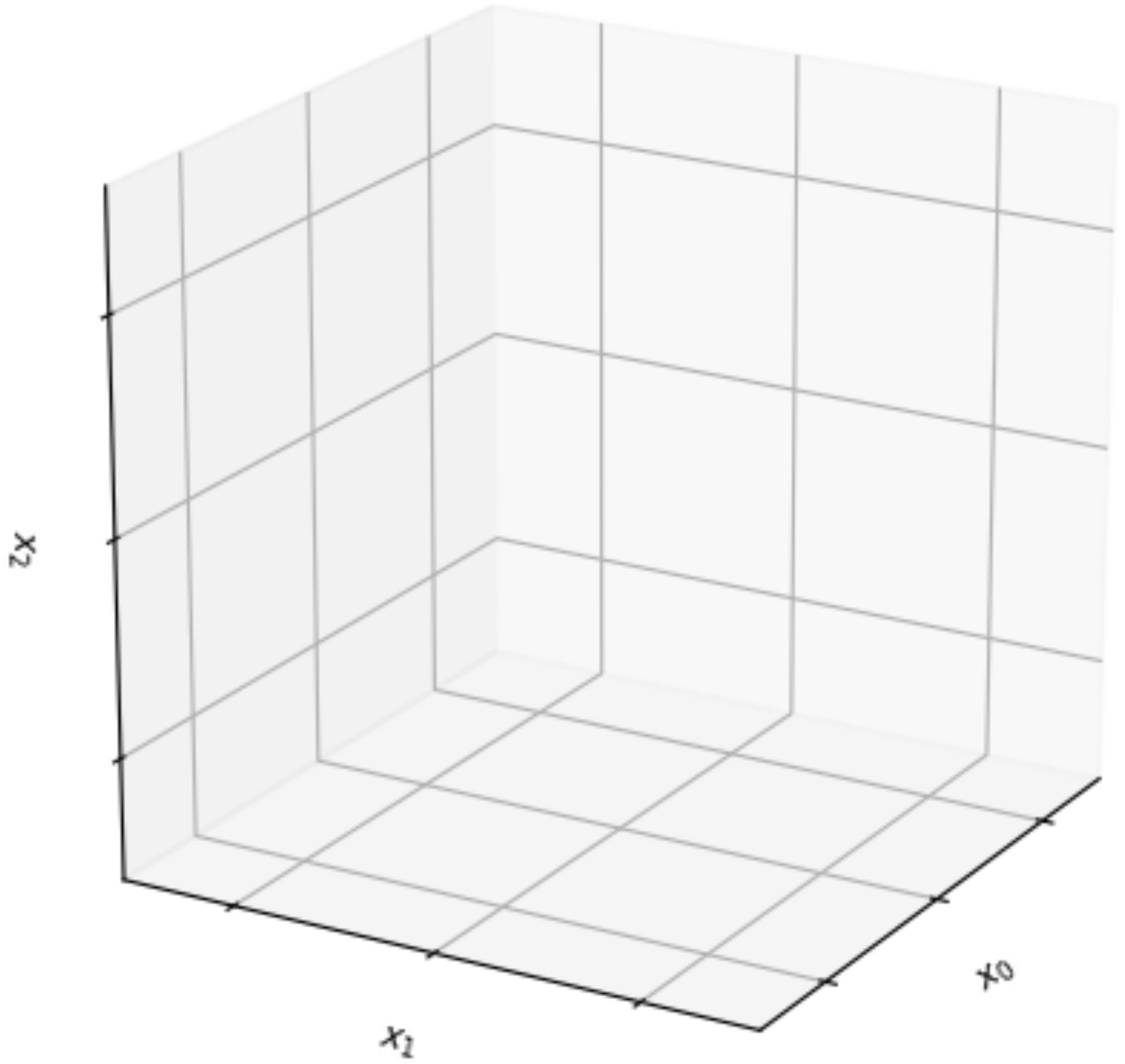
Why the data lives on a manifold

Dynamical systems like

- the Lorenz attractor
- Attractor networks in theoretical Neuroscience







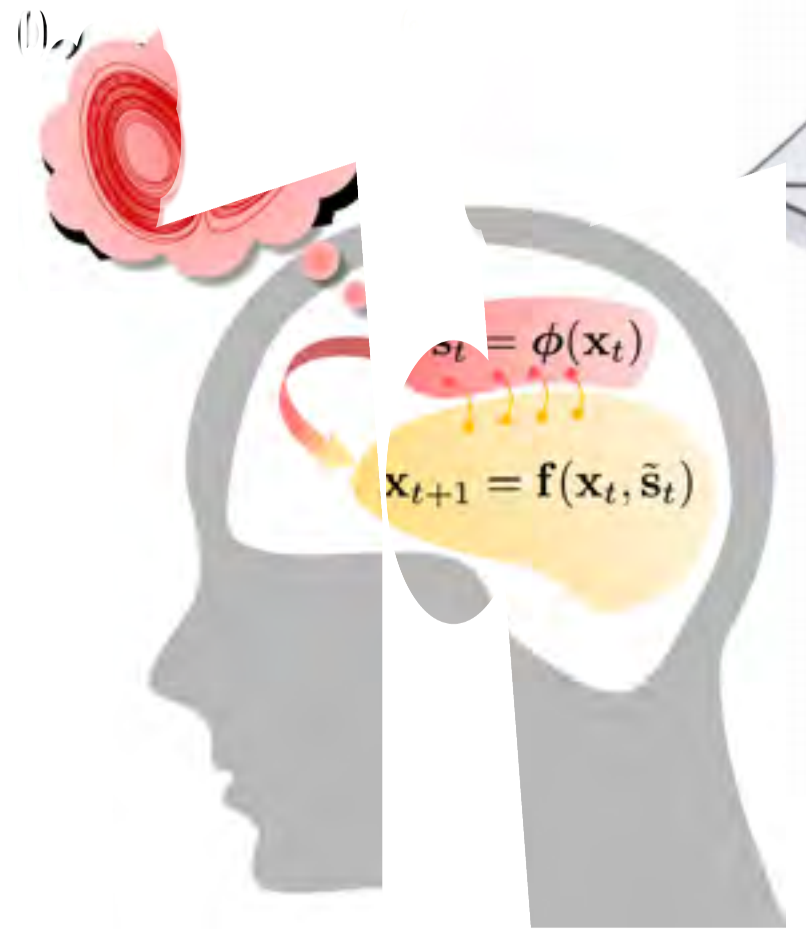
$$\frac{\mathrm{d}x_0}{\mathrm{d}t} = \sigma(x_1 - x_0)\,, \quad \frac{\mathrm{d}x_1}{\mathrm{d}t} = x_0(\rho - x_2) - x_1\,, \quad \frac{\mathrm{d}x_2}{\mathrm{d}t} = x_0 x_1 - \beta x_2\,.$$

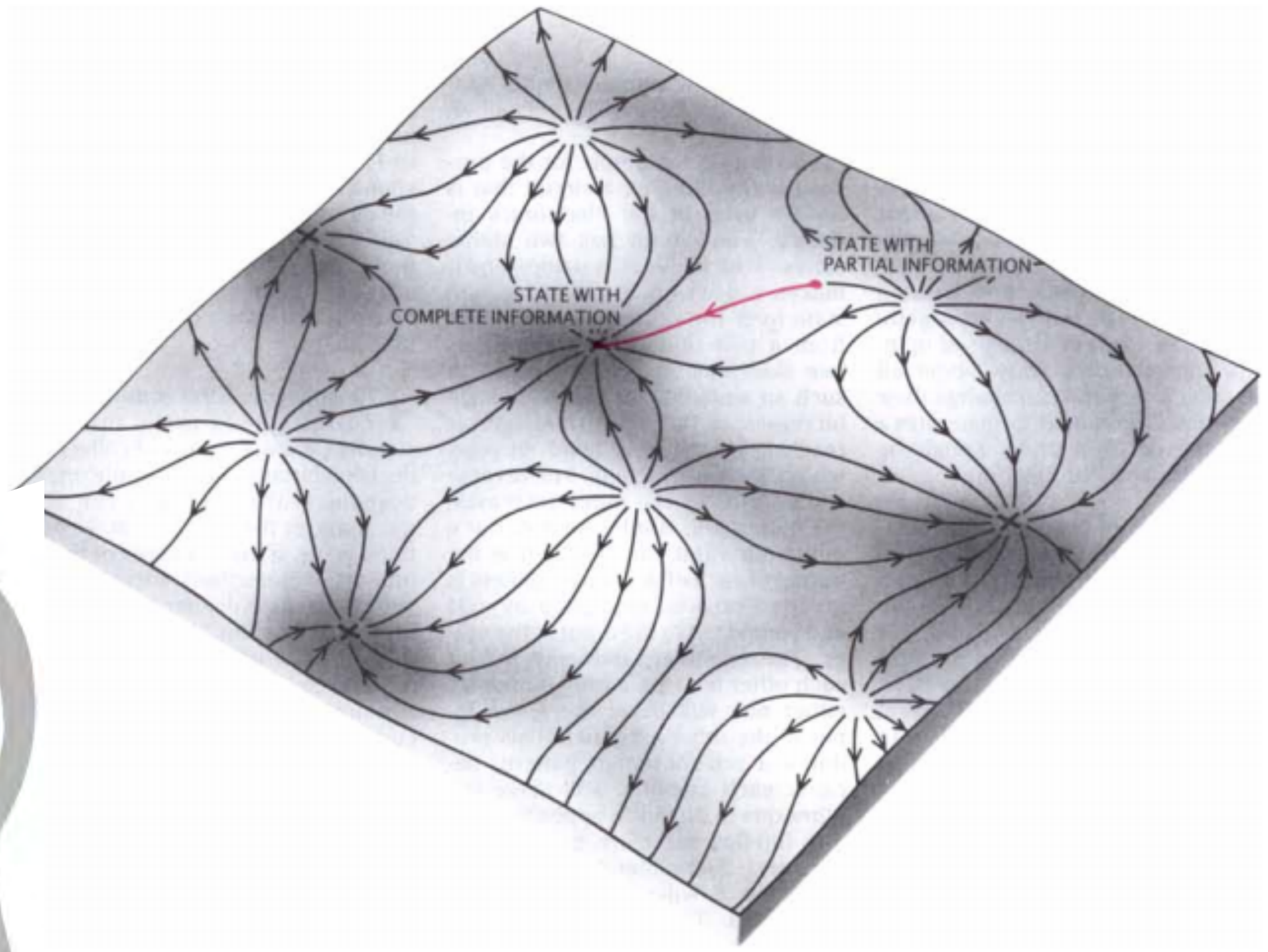
Tank and Hopfield 1987; Lu and Danielle S. Bassett arXiv:1807.05214 ; E. Lorenz 1963

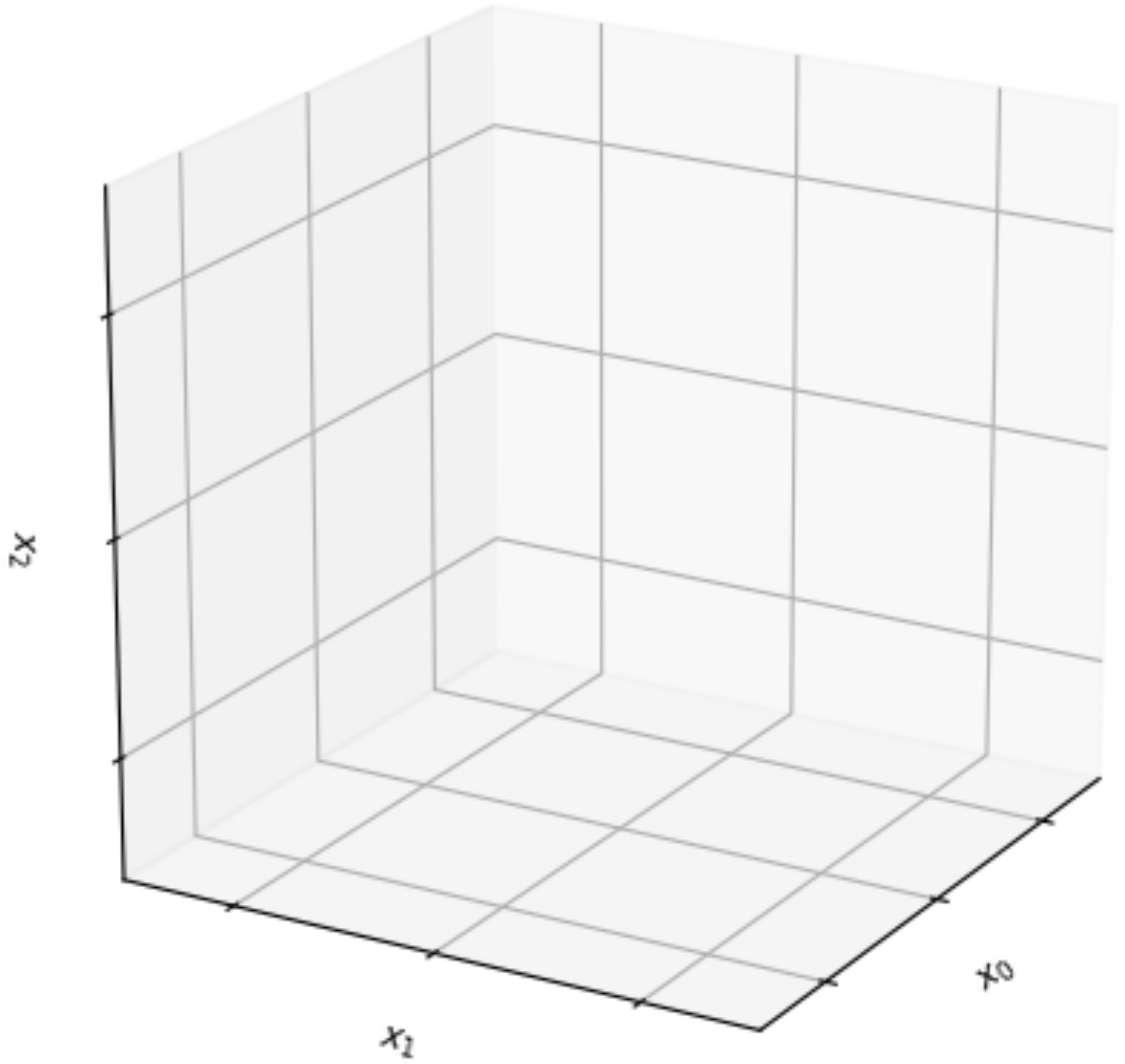
Why the data lives on a manifold

Dynamical systems like

- the Lorenz attractor
- Attractor networks in theoretical Neuroscience







$$\frac{\mathrm{d}x_0}{\mathrm{d}t} = \sigma(x_1 - x_0)\,, \quad \frac{\mathrm{d}x_1}{\mathrm{d}t} = x_0(\rho - x_2) - x_1\,, \quad \frac{\mathrm{d}x_2}{\mathrm{d}t} = x_0 x_1 - \beta x_2\,.$$

Tank and Hopfield 1987; Lu and Danielle S. Bassett arXiv:1807.05214 ; E. Lorenz 1963

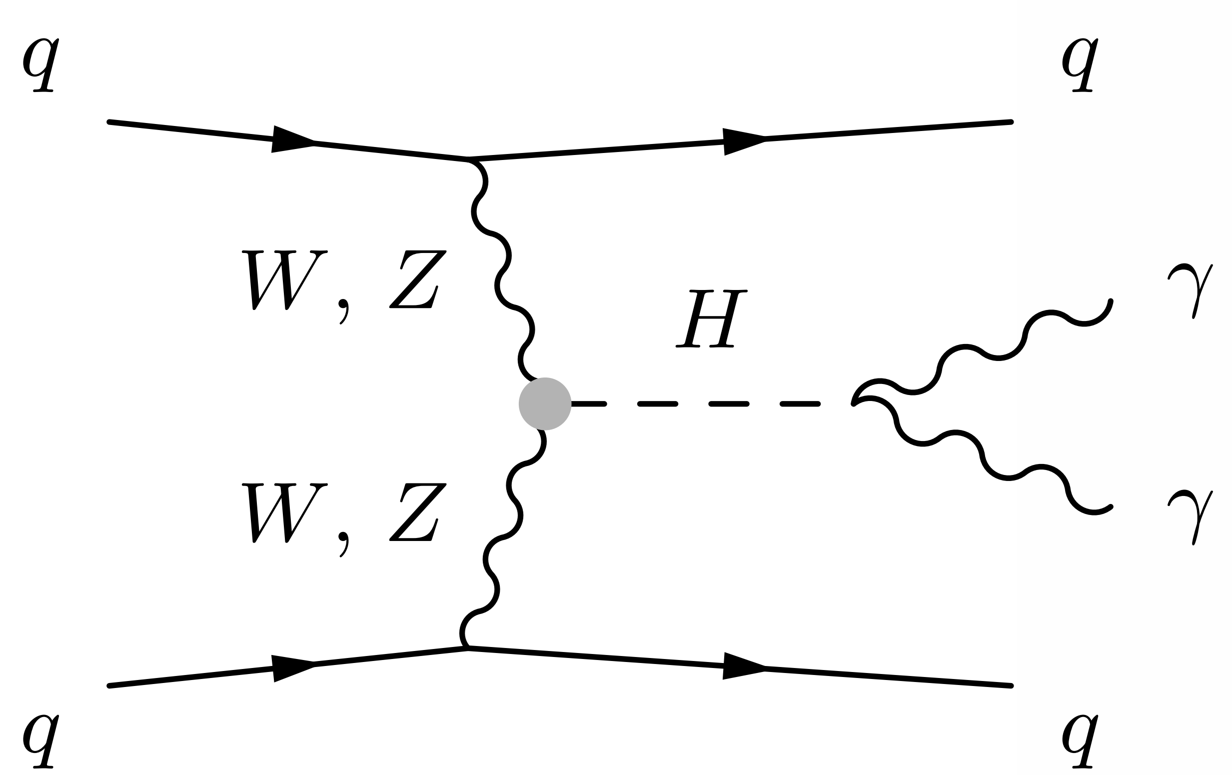
Why the data lives on a manifold

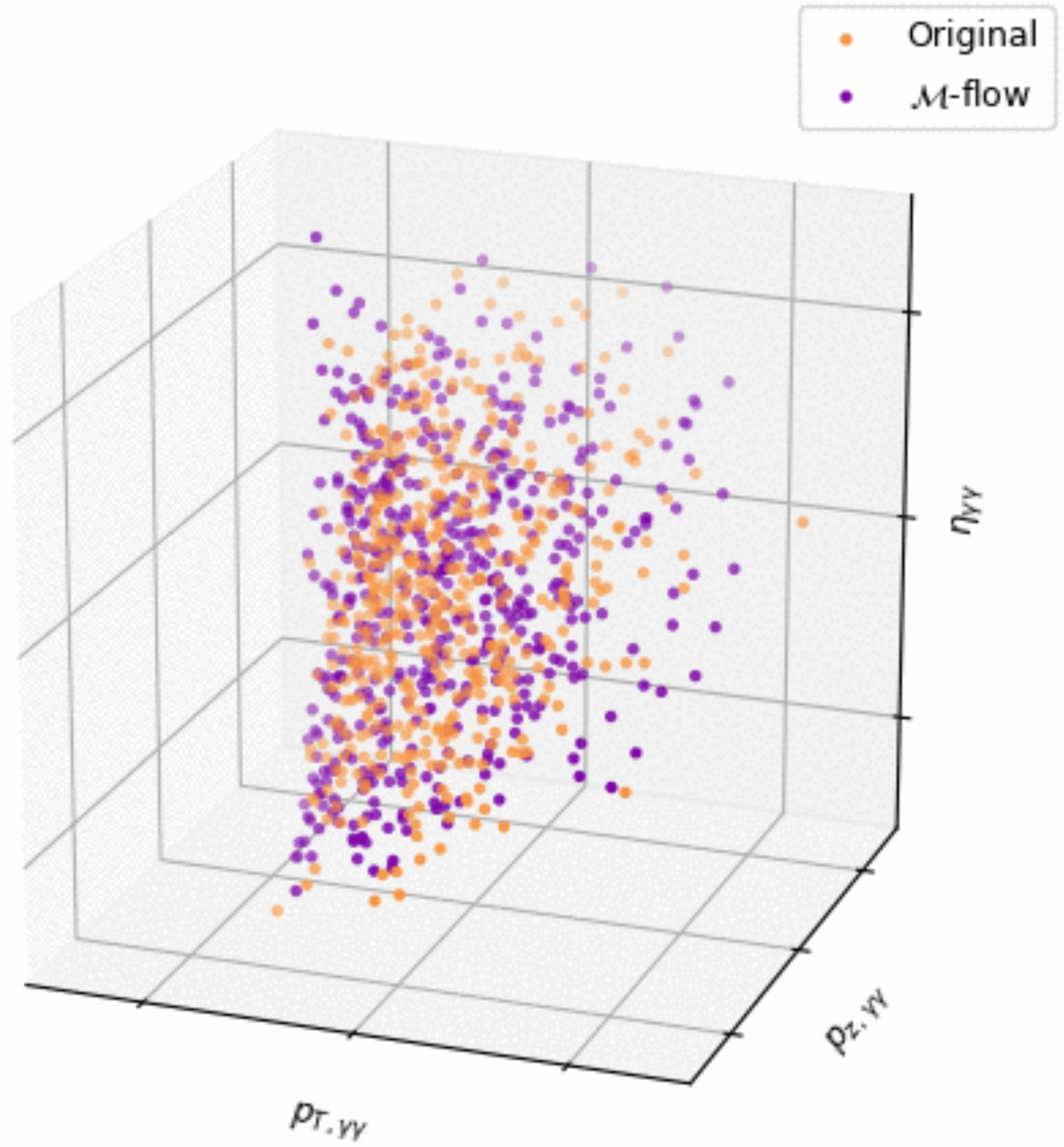
Animation by Johann Brehmer

Conservation laws

Redundant features

Other Constraints





redundant features

particle masses ("on-shell condition")

energy-momentum conservation

14-dimensional manifold embedded in 40-dimensional data space

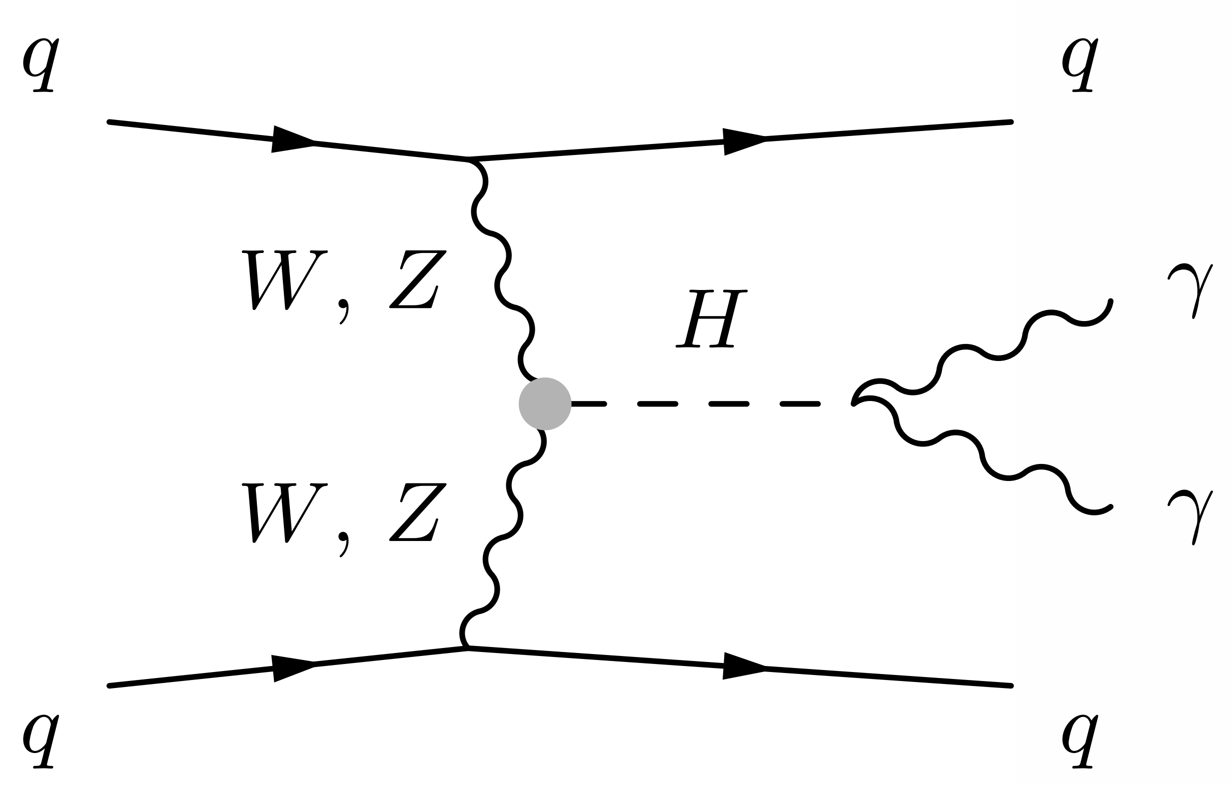
Why the data lives on a manifold

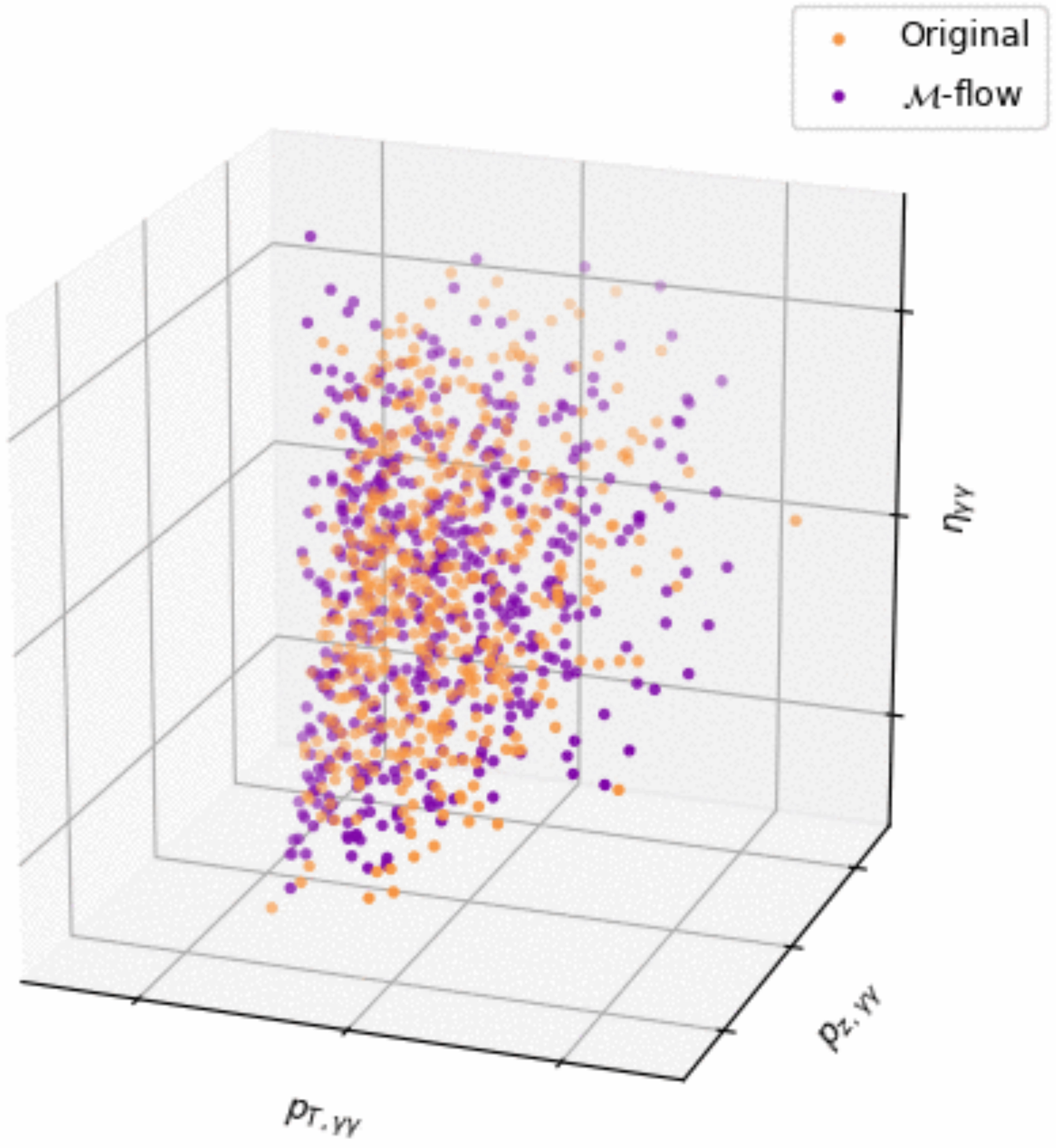
Animation by Johann Brehmer

Conservation laws

Redundant features

Other Constraints





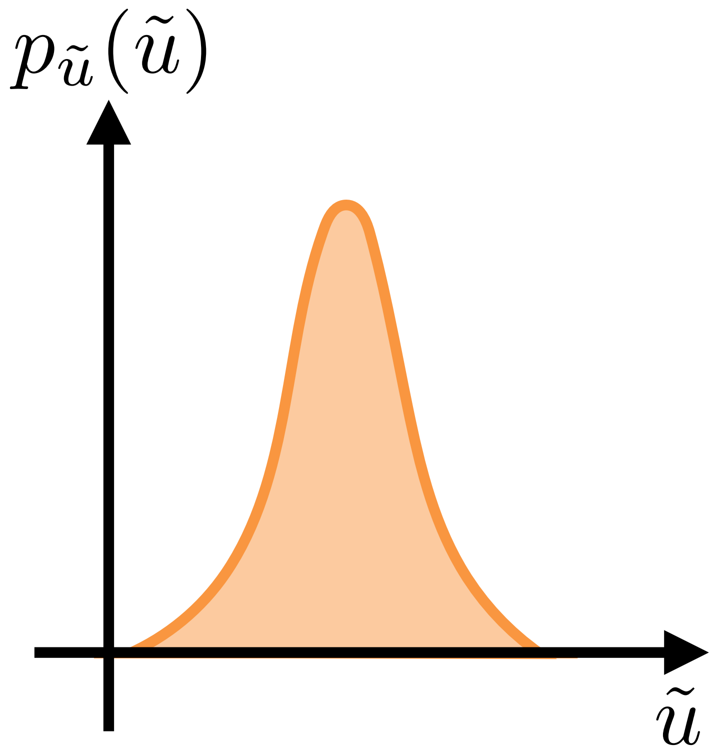
particle masses
("on-shell condition")

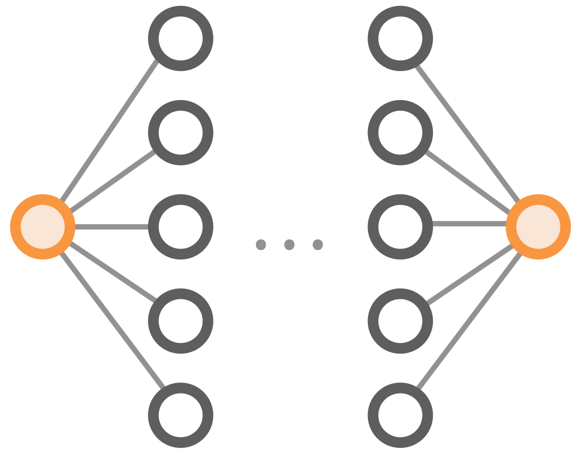
redundant
features

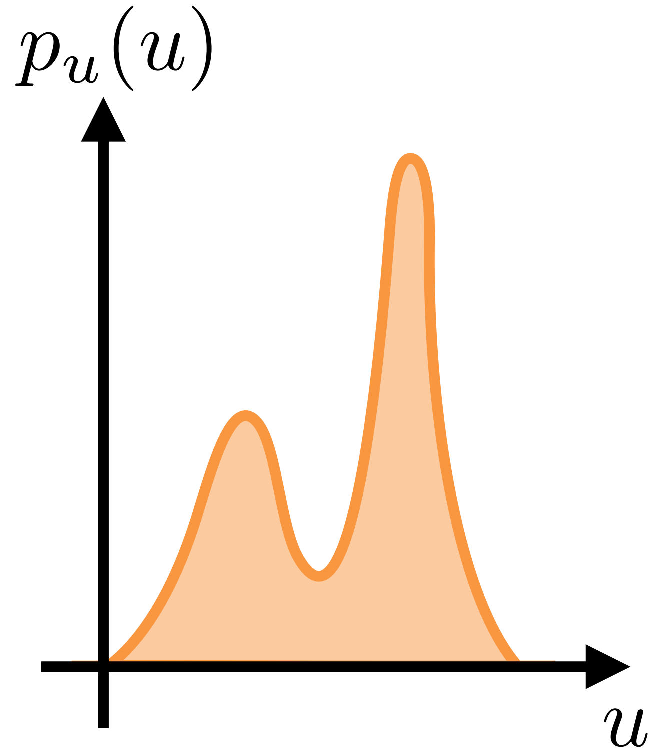
energy-momentum
conservation

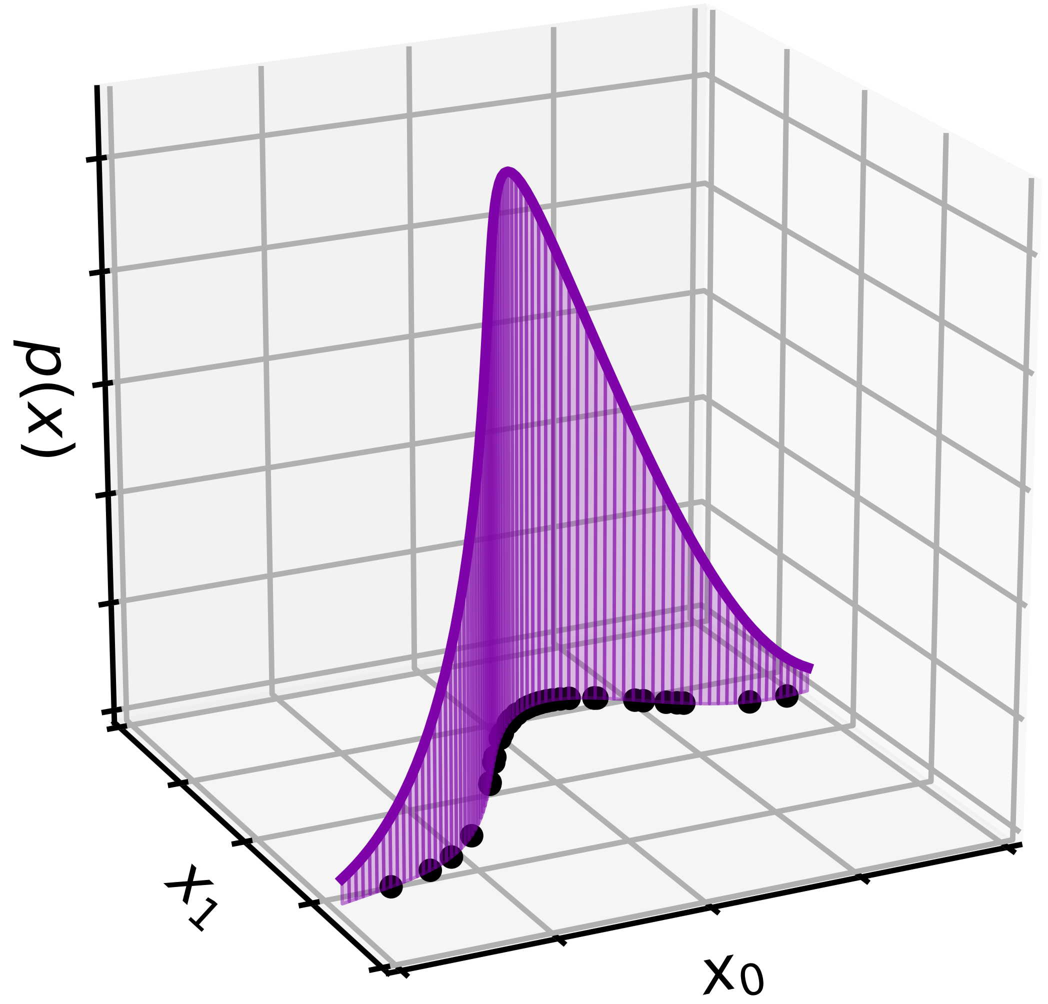
14-dimensional manifold
embedded in 40-dimensional
data space

Flows on a prescribed manifold









$\tilde{u} \sim p_{\tilde{u}}(\tilde{u})$ ⟷ h ⟷ u ⟷ g^* ⟷ x

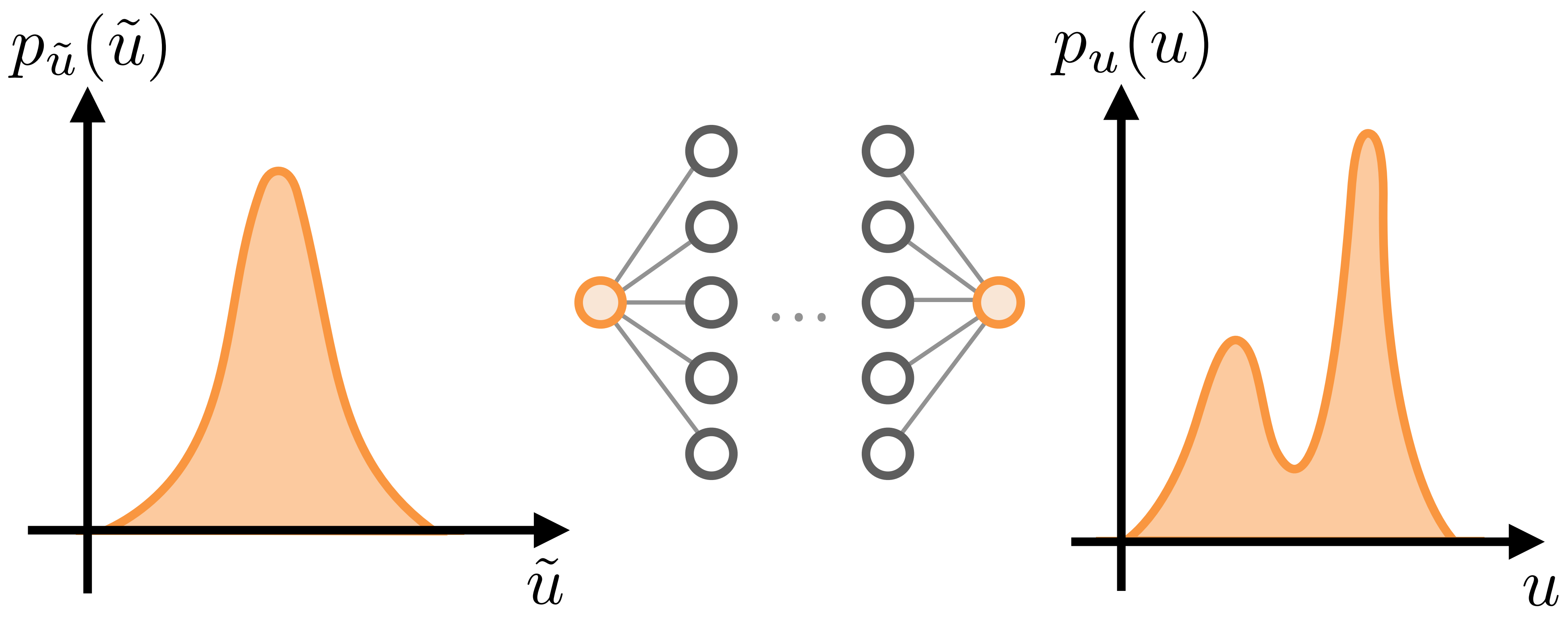
n-dim. latents | invertible NN | n-dim. latents | prescribed chart | tractable density over $\mathcal{M}^*$

$$p_{\mathcal{M}^*}(x) = p_{\tilde{u}}(\tilde{u}) \;\left|\det J_h(\tilde{u})\right|^{-1} \cdot \left|\det[J_{g^*}^T(u) J_{g^*}(u)]\right|^{-\frac{1}{2}}$$

[M. Gemici et al 1611.02304;
D. Rezende, …, M. Albergo, …, K. Cranmer 2002.02428

$\mathcal{M}$-flows

See Spotlight talk by Johann Brehmer "NOTAGAN"



$\tilde{u} \sim p_{\tilde{u}}(\tilde{u})$ ⟷ u

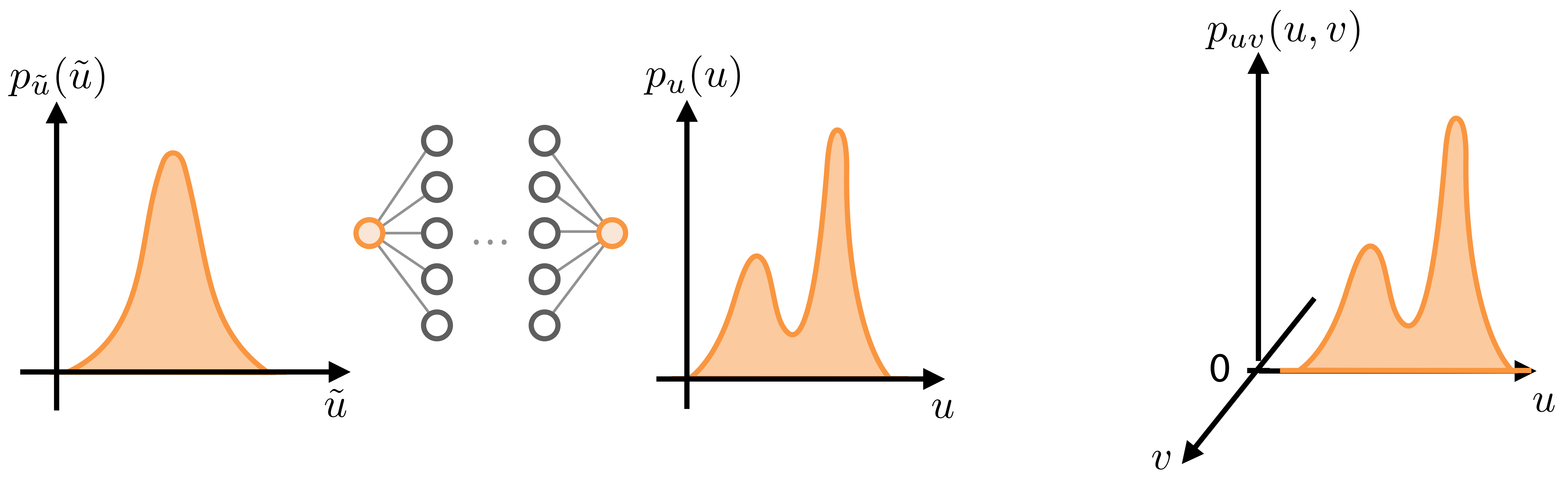
h

n-dim. latents | inv. NN | n-dim. latents

[Johann Brehmer, Kyle Cranmer 2003.13913]

$\mathcal{M}$-flows

See Spotlight talk by Johann Brehmer "NOTAGAN"

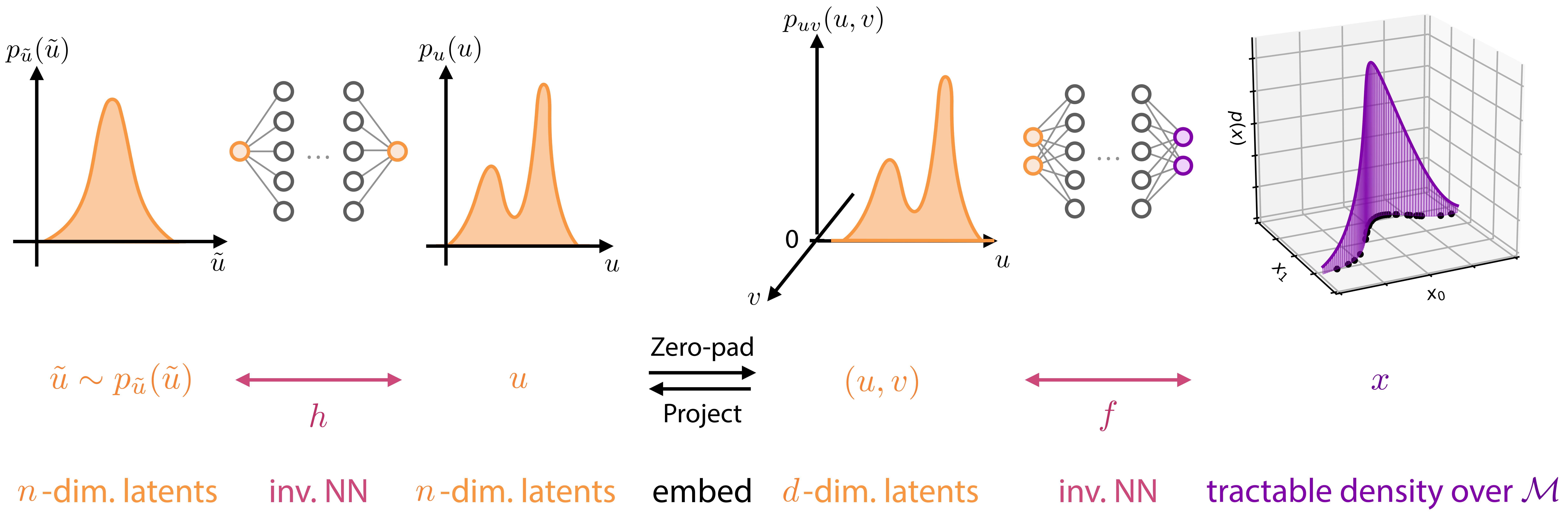


$\tilde{u} \sim p_{\tilde{u}}(\tilde{u})$	$\longleftrightarrow$ h	u	Zero-pad / Project	(u, v)
n-dim. latents	inv. NN	n-dim. latents	embed	d-dim. latents

[Johann Brehmer, Kyle Cranmer 2003.13913]

$\mathcal{M}$-flows

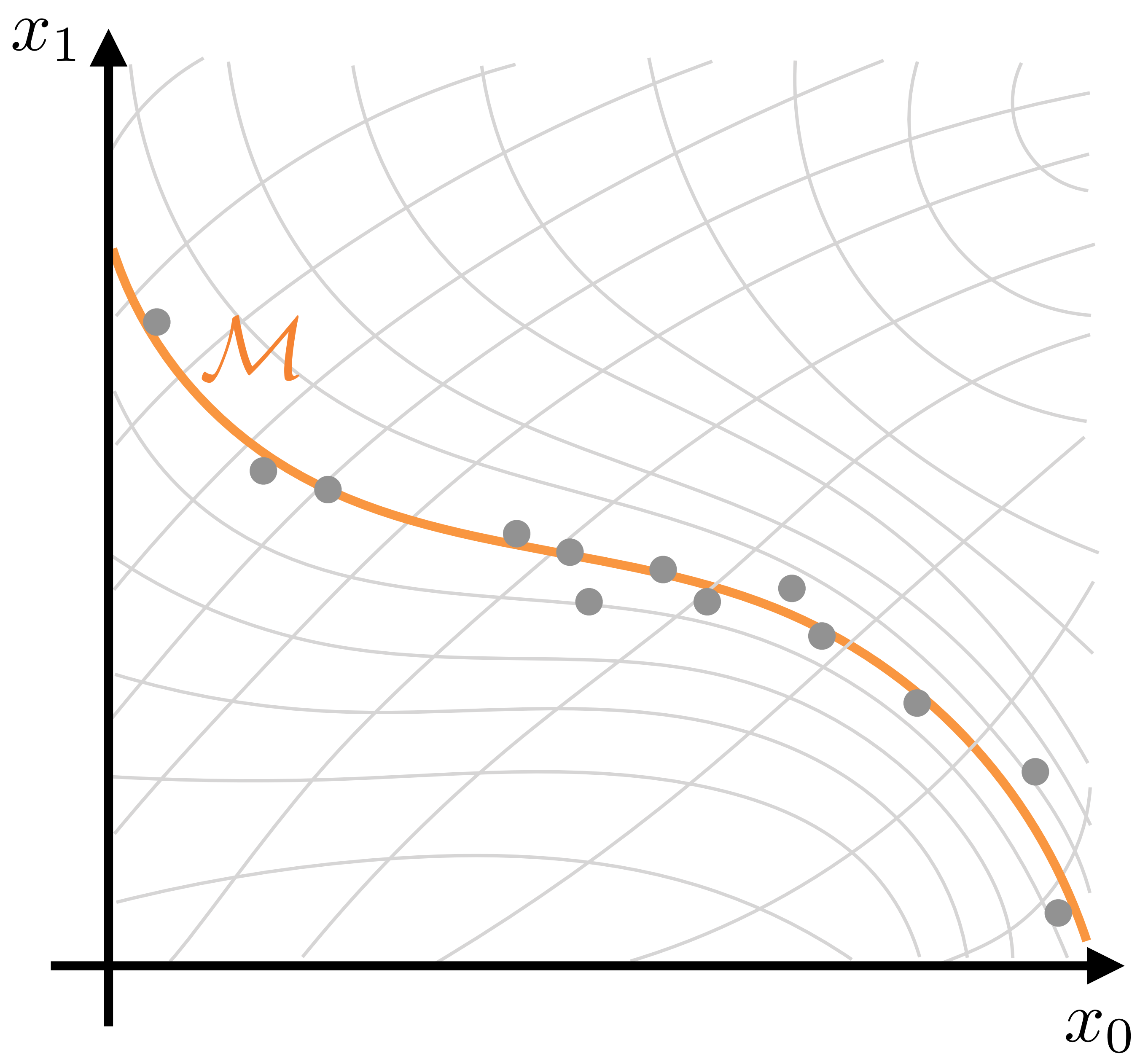
See Spotlight talk by Johann Brehmer “NOTAGAN"

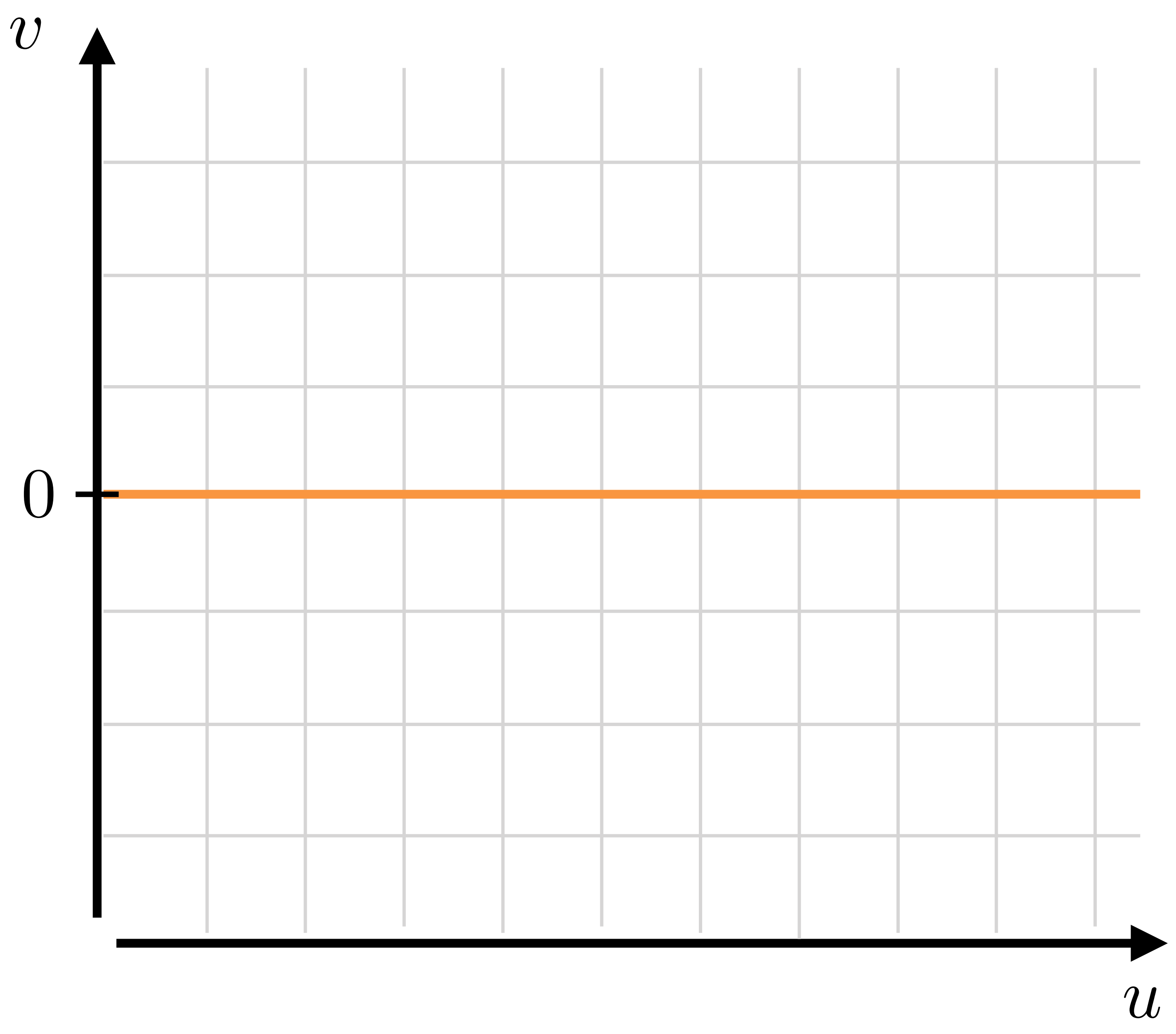


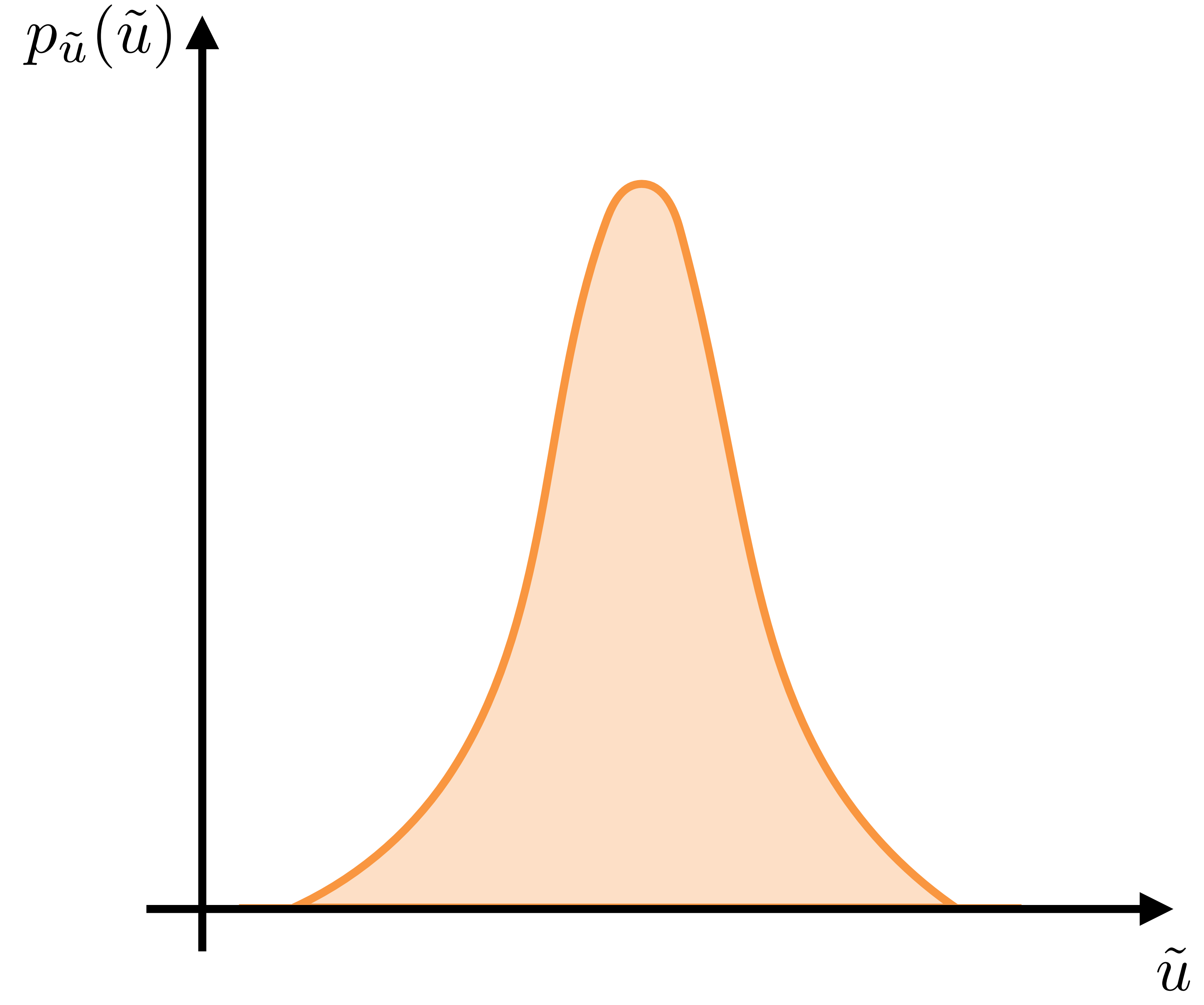
$$p_{\mathcal{M}}(x) = p_{\tilde{u}}(\tilde{u}) \, |\det J_h(\tilde{u})|^{-1} \cdot \left| \det \left[(\mathbb{1} \ 0) J_f(u)^T J_f(u) \begin{pmatrix} \mathbb{1} \\ 0 \end{pmatrix} \right] \right|^{-\frac{1}{2}}$$

[Johann Brehmer, Kyle Cranmer 2003.13913]

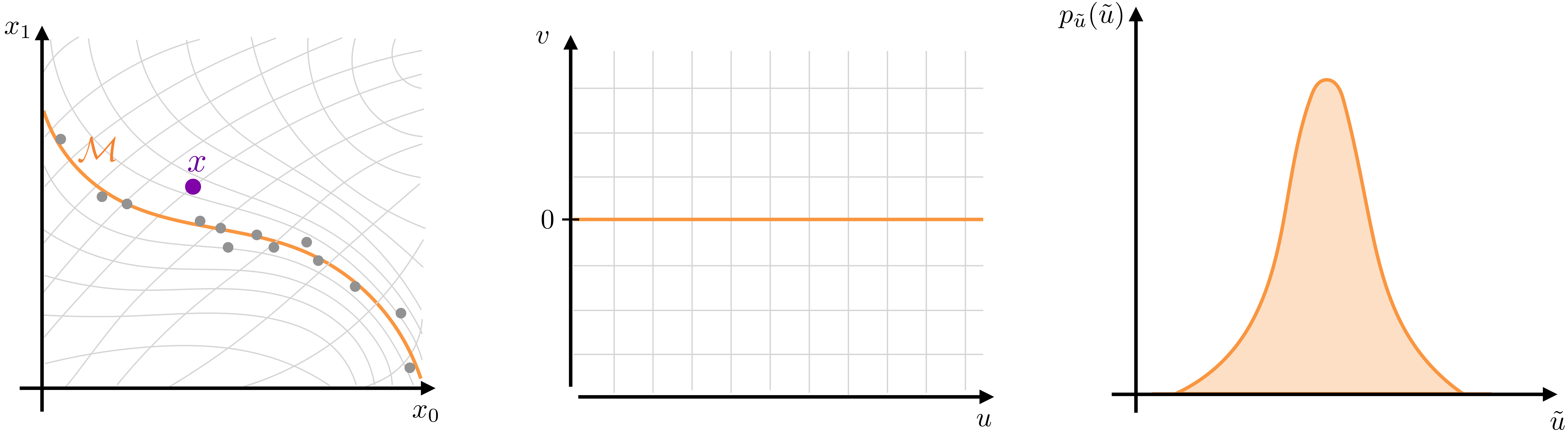
Evaluating data on or off the manifold



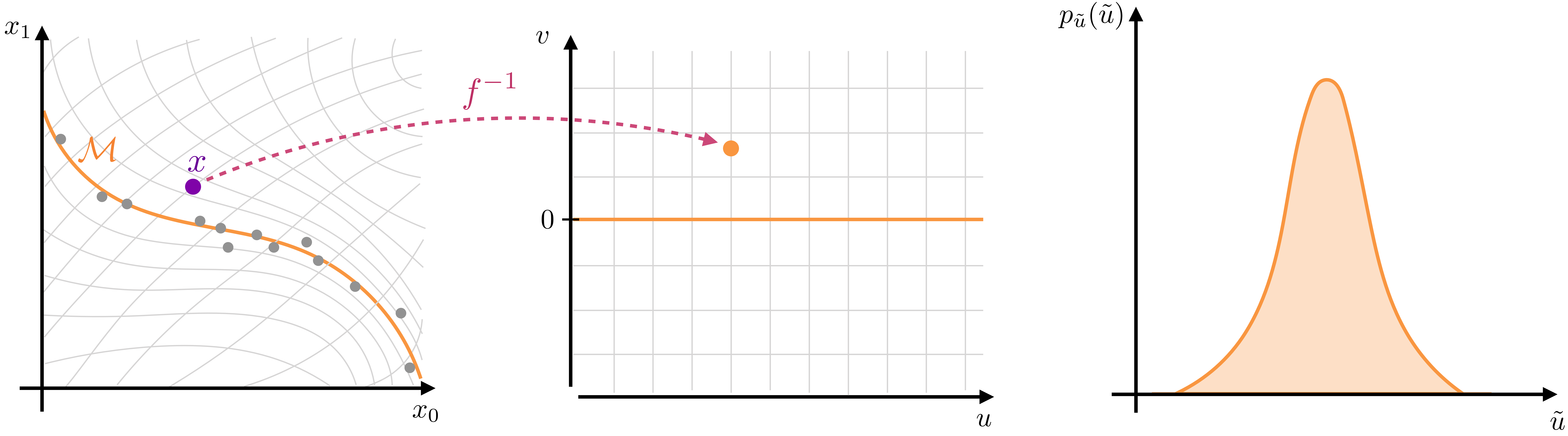




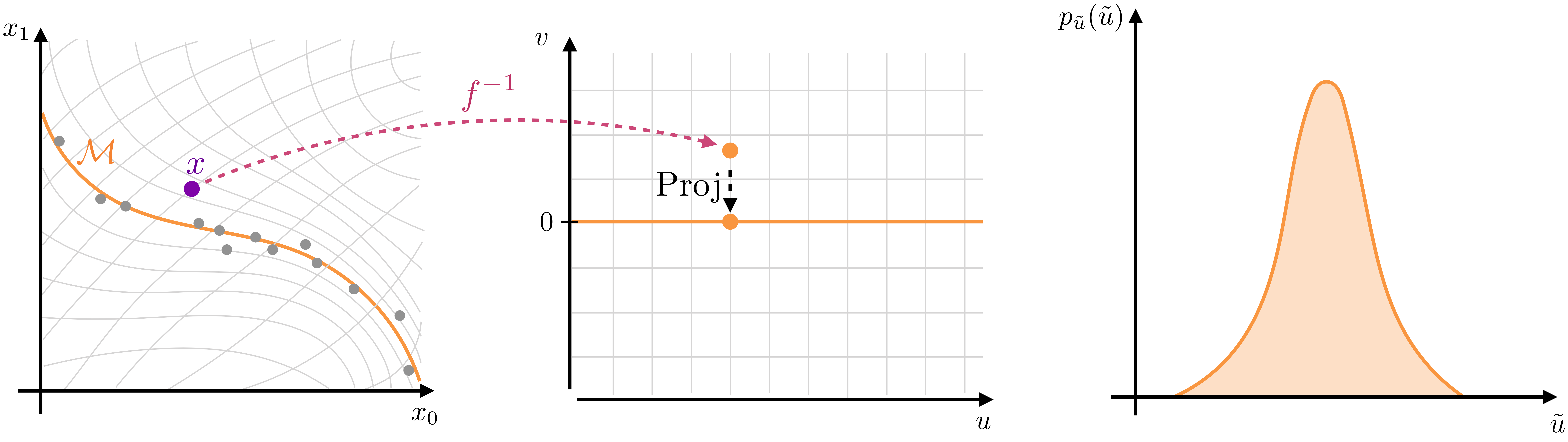
Evaluating data on or off the manifold



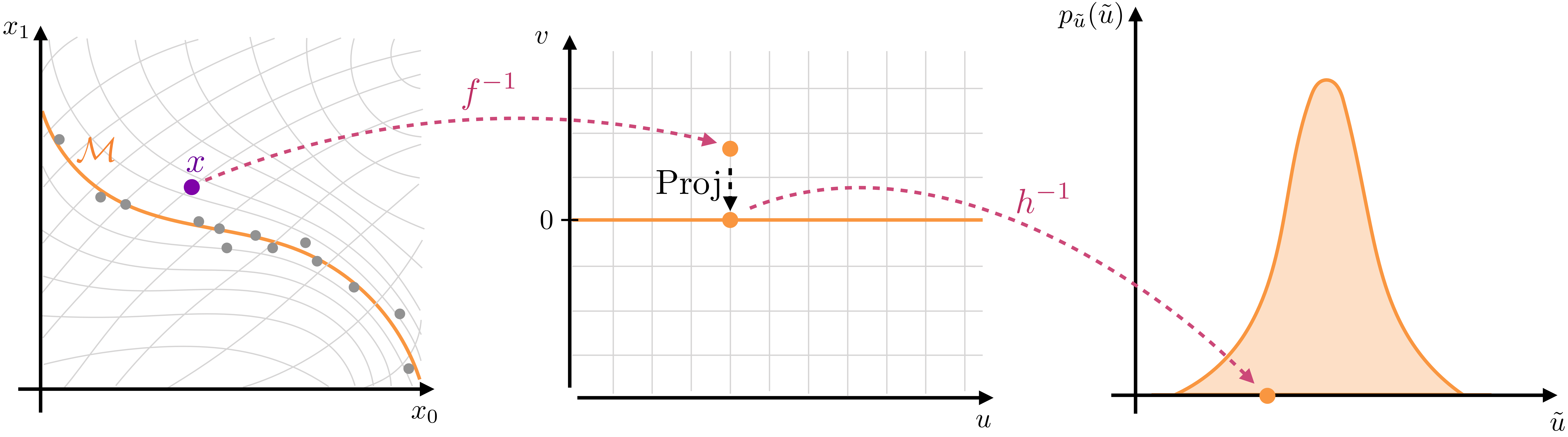
Evaluating data on or off the manifold



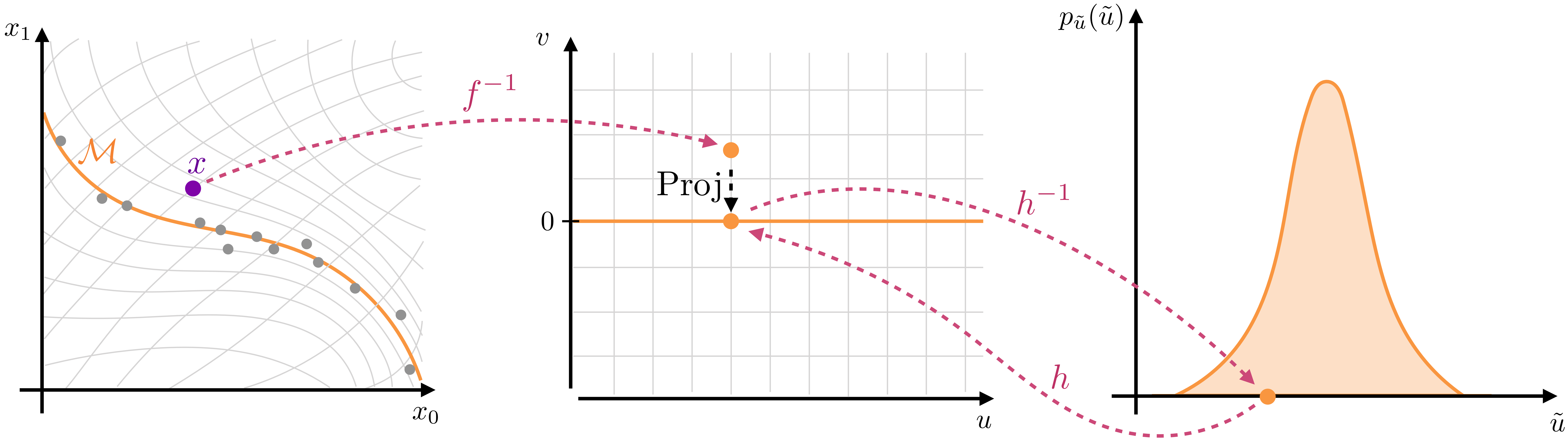
Evaluating data on or off the manifold



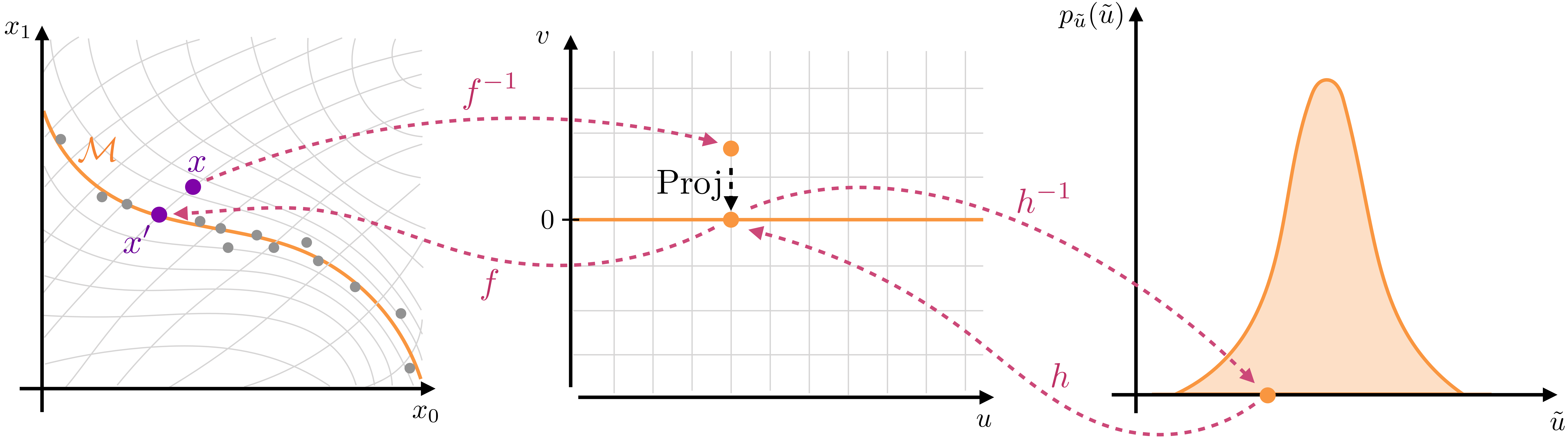
Evaluating data on or off the manifold



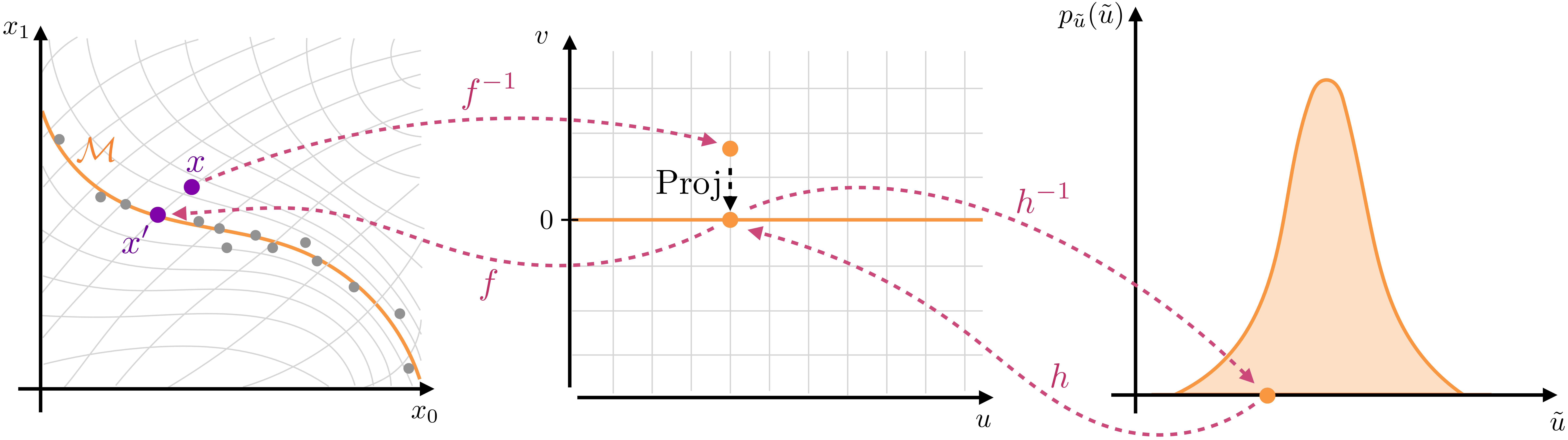
Evaluating data on or off the manifold



Evaluating data on or off the manifold



Evaluating data on or off the manifold



Input x ⇝ Representation $\tilde{u}$ (dimensionality reduction)

⇝ Projection to manifold x' (denoising)

⇝ Reconstruction error $\|x - x'\|$ (training, OOD detection)

⇝ Likelihood after projection $p_{\mathcal{M}}(x')$ (training, inference)

Generative models vs. the data manifold

Model	Manifold	Chart	Generative	Tractable density	Restr. to manifold
Ambient flow (AF)	no	no	✓	✓	no
Flow on prescr. manifold	prescribed	prescribed	✓	✓	✓
GAN	learned	no	✓	no	✓
VAE	learned	no	✓	only ELBO	(no)
$\mathcal{M}$-flow	learned	learned	✓	✓ (potentially slow)	✓

The likelihood is not what it seems

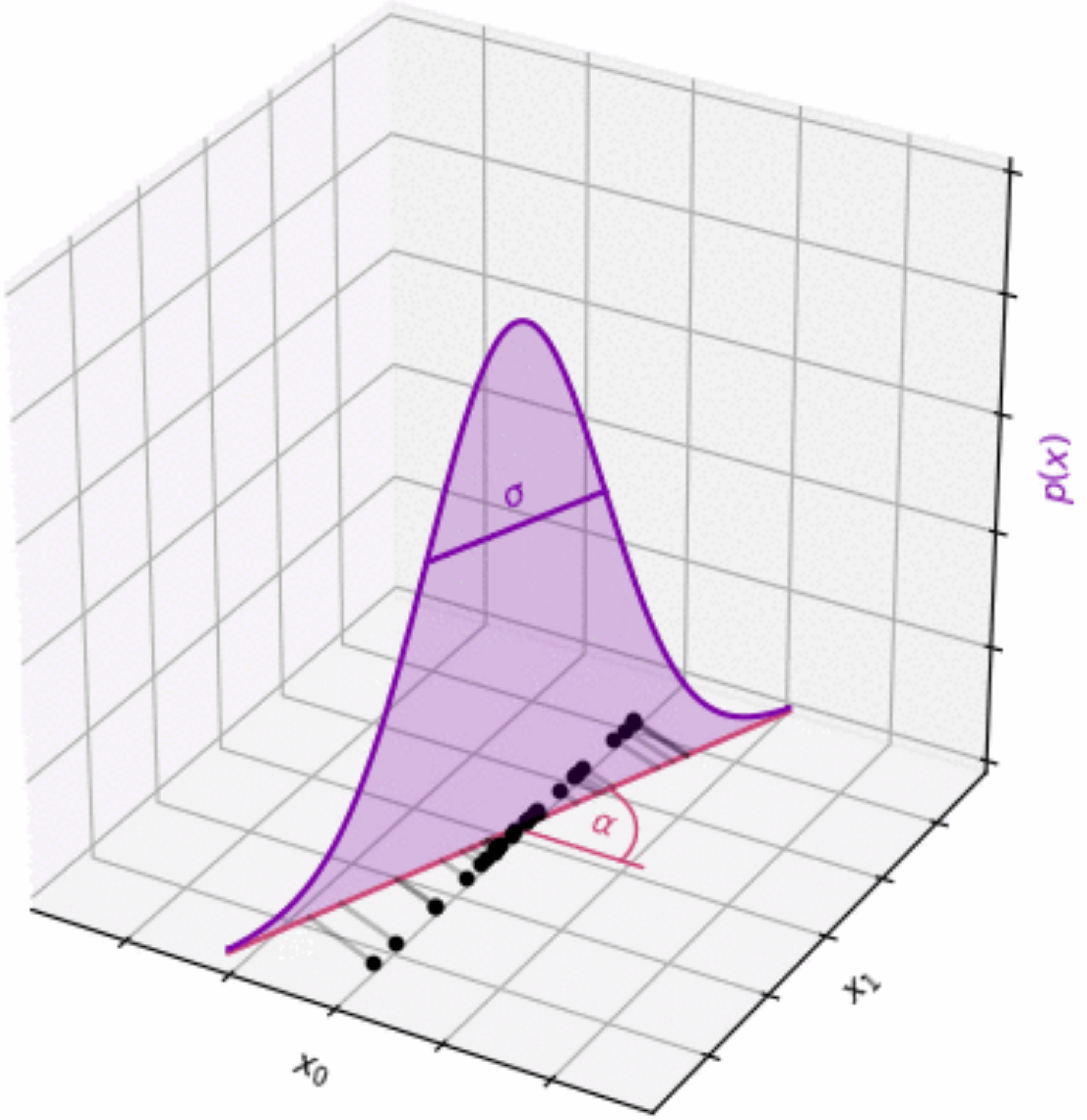
Likelihood defined after projection to $\mathcal{M}$, which is defined through NN weights ϕ_f

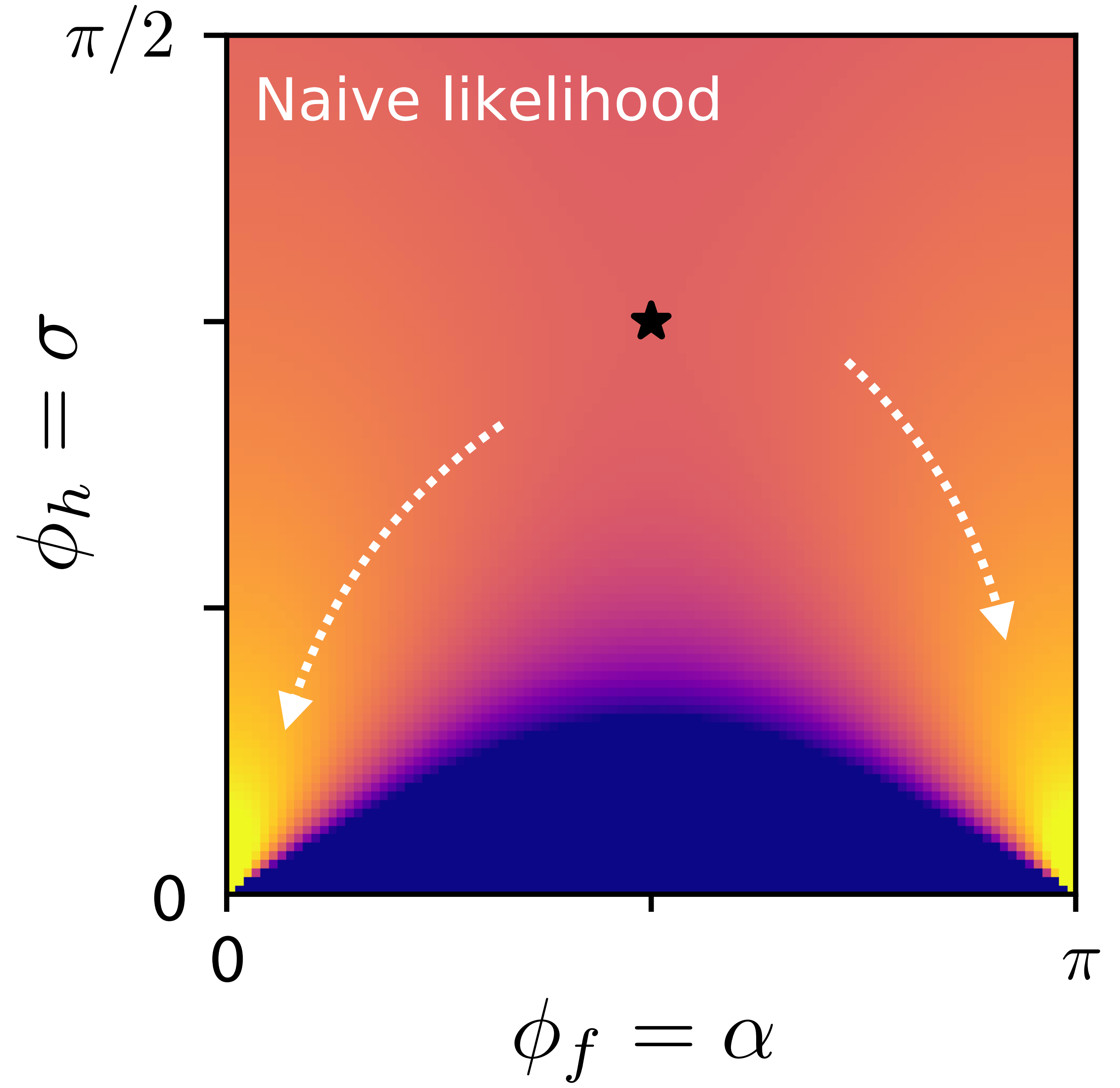
Family of likelihoods $p_{\phi_f}(x|\phi_h)$
rather than one likelihood $p(x|\phi_f, \phi_h)$

⇒ Learning ϕ_f by maximizing
$p_{\phi_f}(x|\phi_h)$ is unstable

$p_{\phi_f}(x|\phi_h)$ is not really a likelihood function in the parameter ϕ_f

We call it the "naive likelihood"





The likelihood is not what it seems

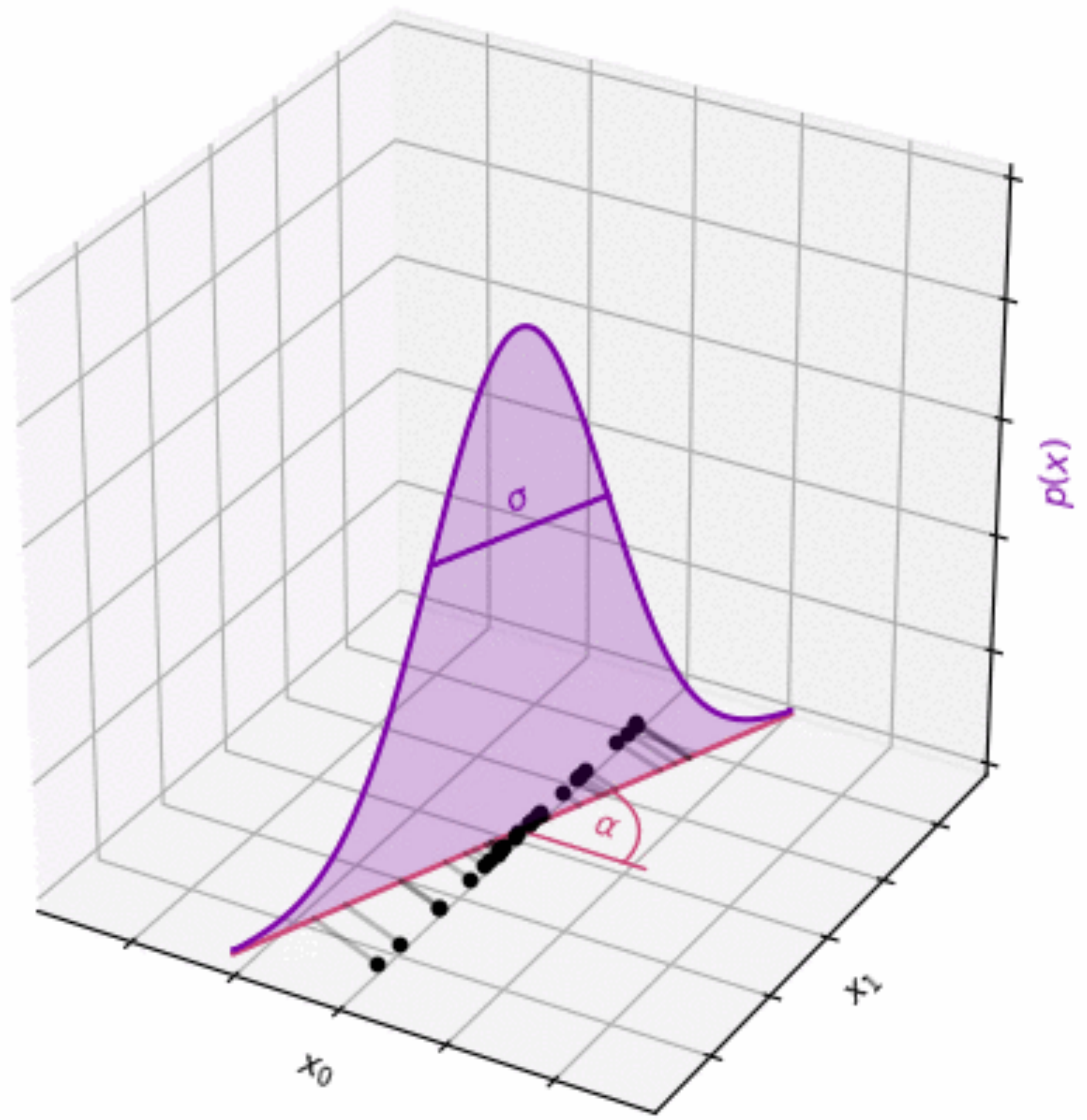
Likelihood defined after projection to $\mathcal{M}$, which is defined through NN weights ϕ_f

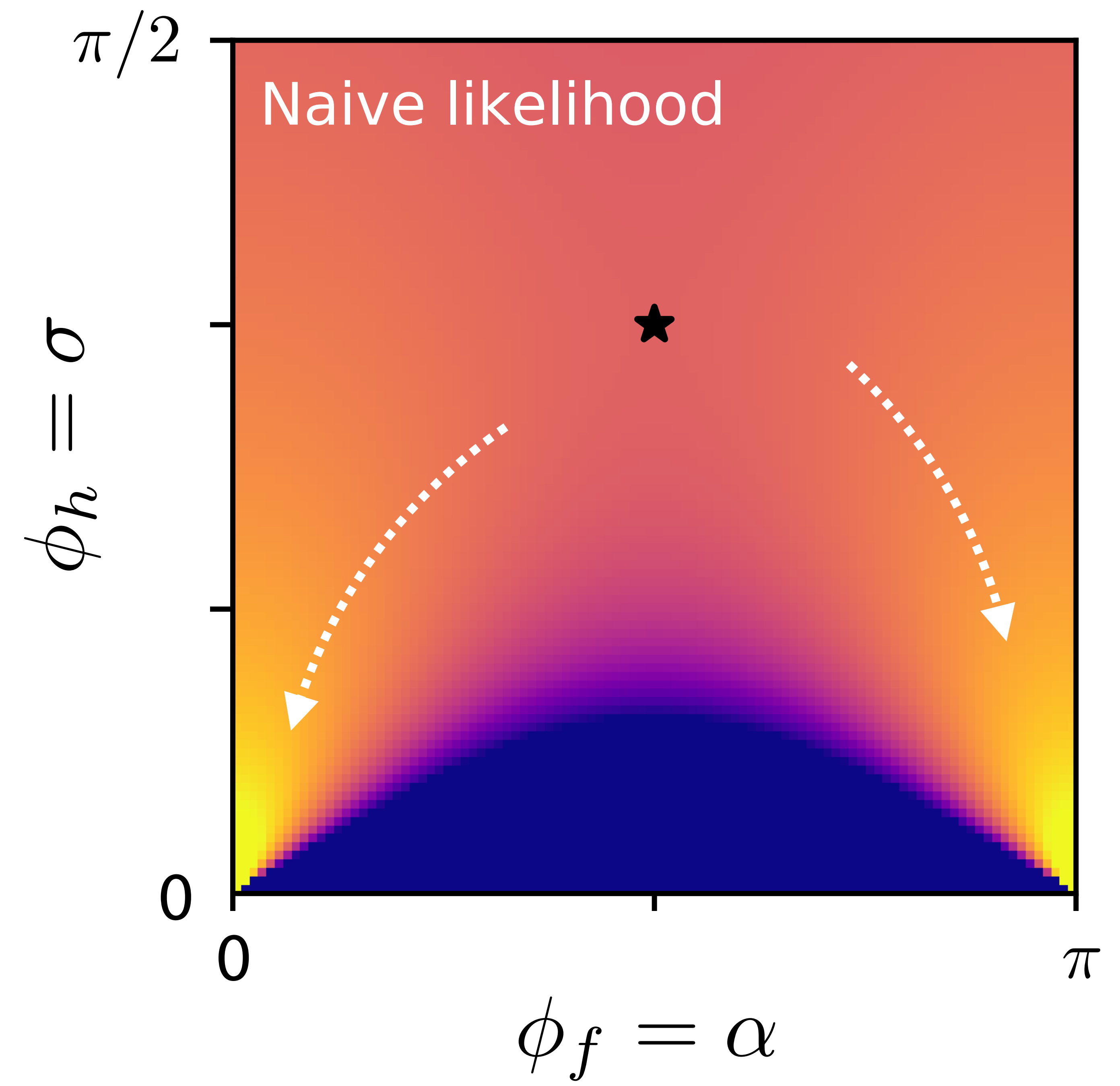
Family of likelihoods $p_{\phi_f}(x|\phi_h)$
rather than one likelihood $p(x|\phi_f, \phi_h)$

⇒ Learning ϕ_f by maximizing
$p_{\phi_f}(x|\phi_h)$ is unstable

$p_{\phi_f}(x|\phi_h)$ is not really a likelihood
function in the parameter ϕ_f

We call it the “naive likelihood”

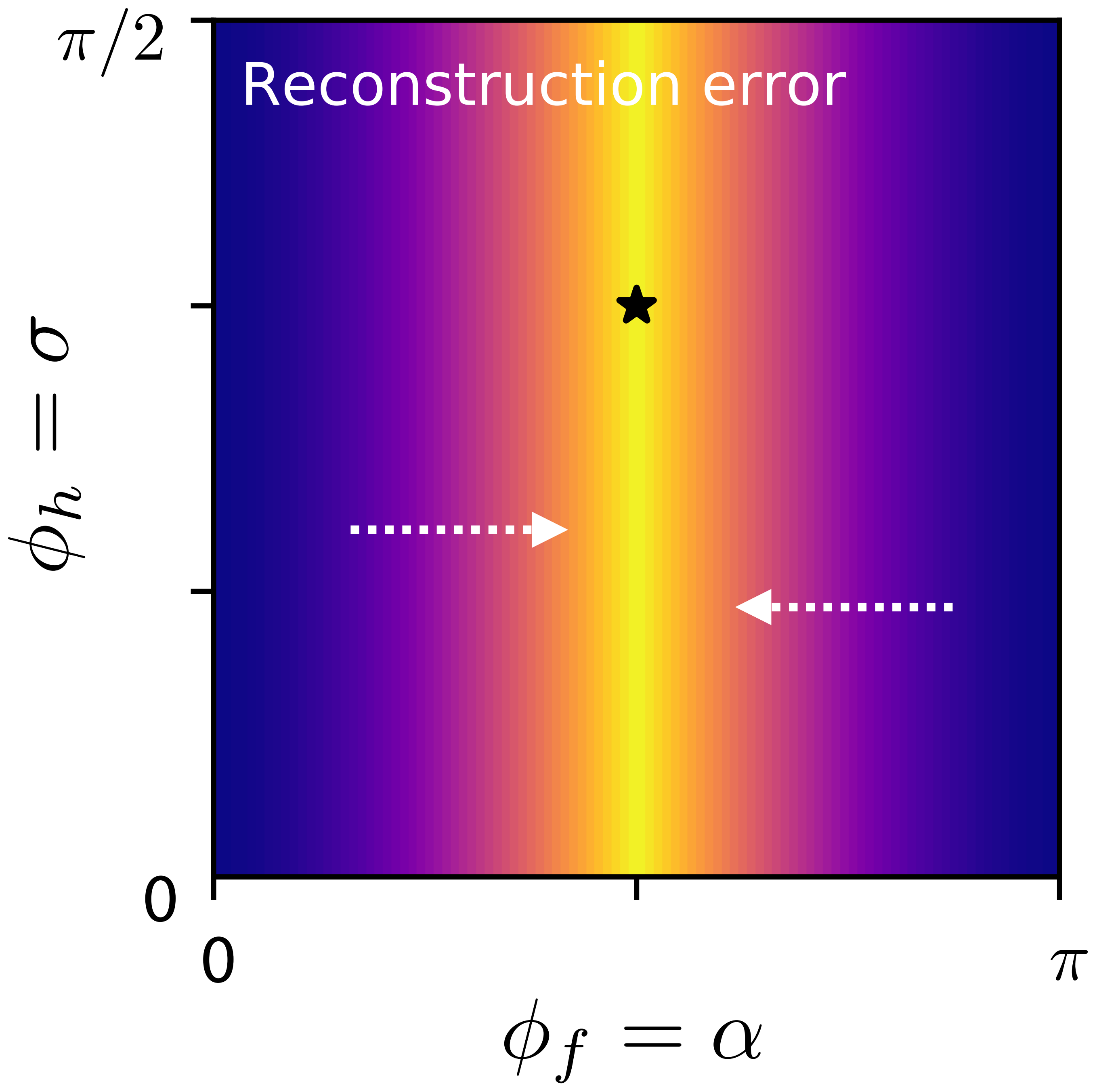




M/D training

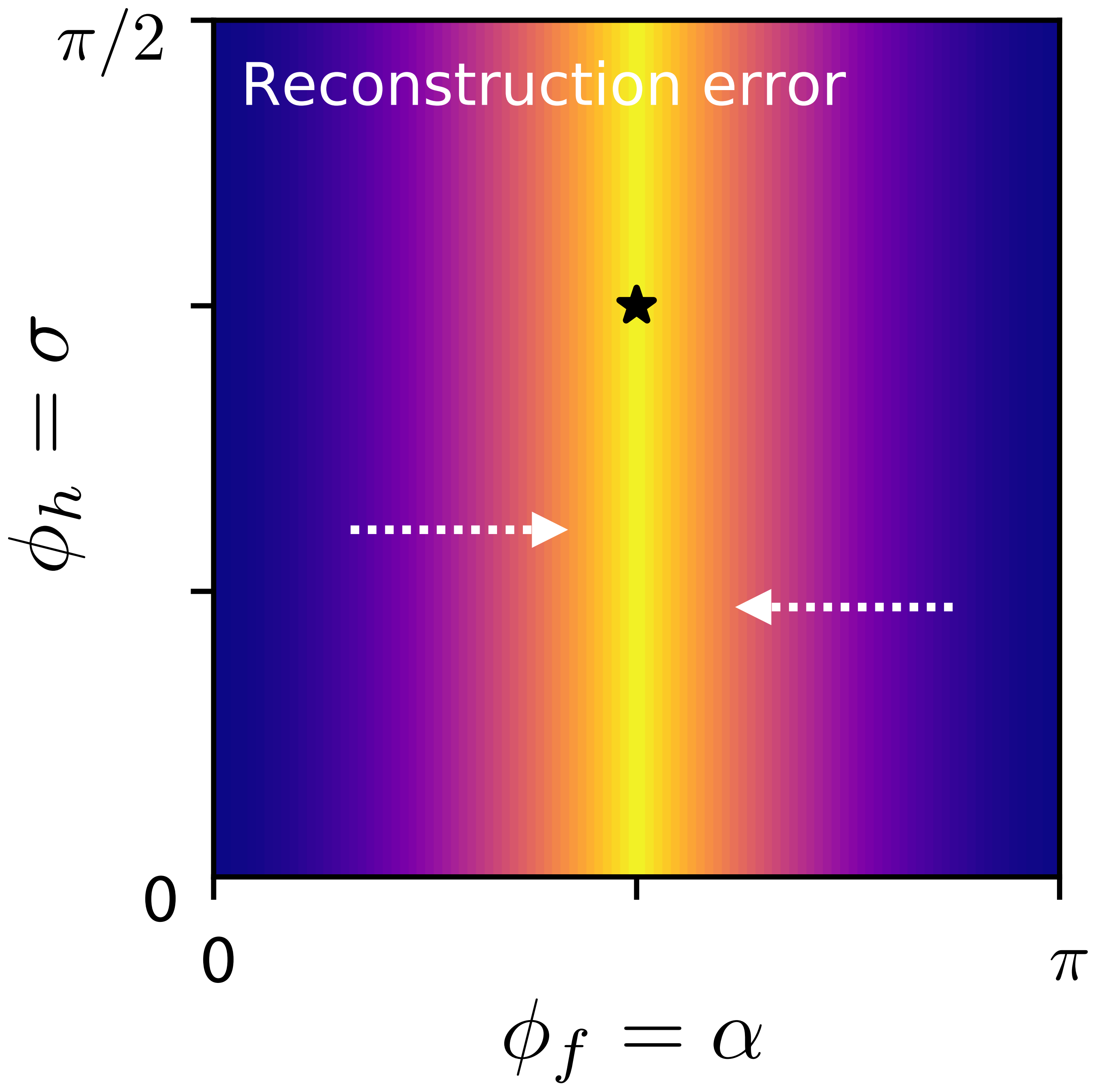
Solution: separate training in two phases!

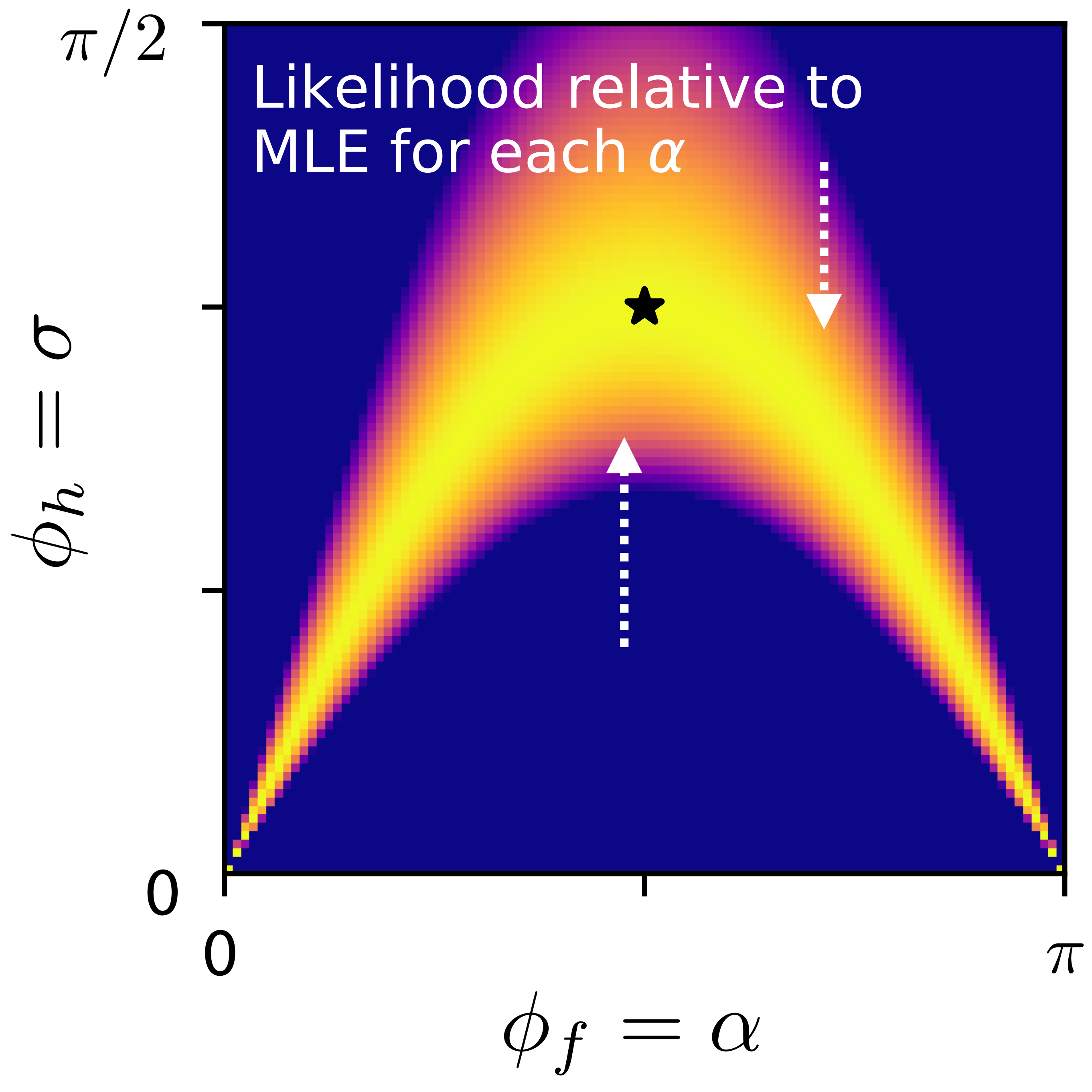
- Manifold phase:
  update ϕ_f (and thus $\mathcal{M}$) by minimizing $\|x - x'\|$



Solution: separate training in two phases!

- **Manifold phase**: update ϕ_f (and thus $\mathcal{M}$) by minimizing $\|x - x'\|$
- **Density phase**: update ϕ_h (and thus $p_{\mathcal{M}}(x)$) by maximum likelihood (keeping $\mathcal{M}$ fixed)





Quantum Field Theory

AI for Lattice Field Theory

Shanahan

Massachusetts Institute of Technology

Abbott

Massachusetts Institute of Technology

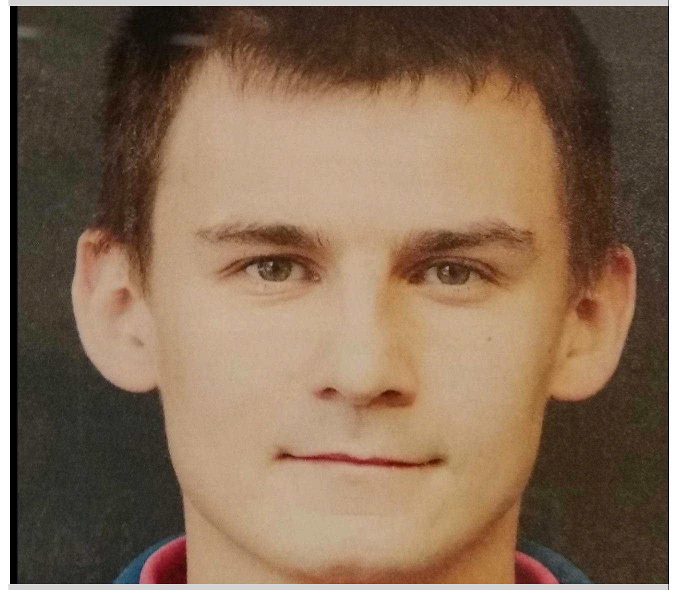
Hackett

Massachusetts Institute of Technology

Romero-López

Massachusetts Institute of Technology

Boyda

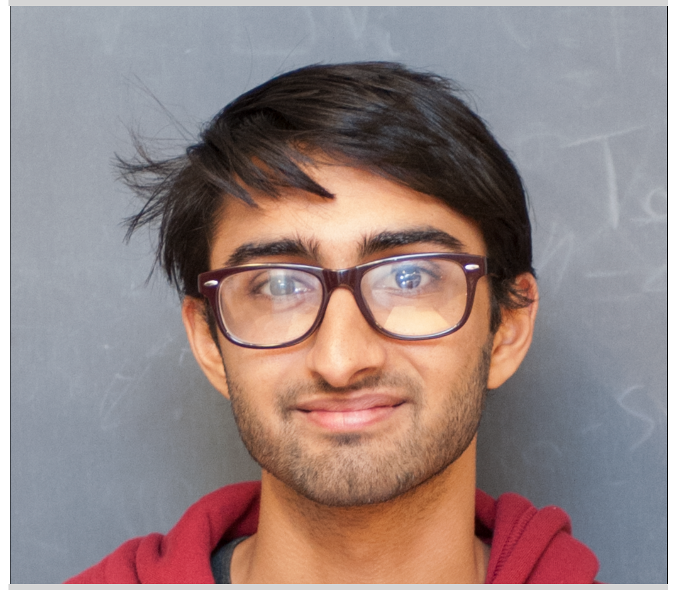


Argonne National Laboratory → MIT

Urban

Universität Heidelberg → MIT

Kanwar



MIT → u^b Bern

Botev

DeepMind

Matthews

DeepMind

Rezende

DeepMind

Racanière

DeepMind

Razavi

DeepMind

Albergo

NYU

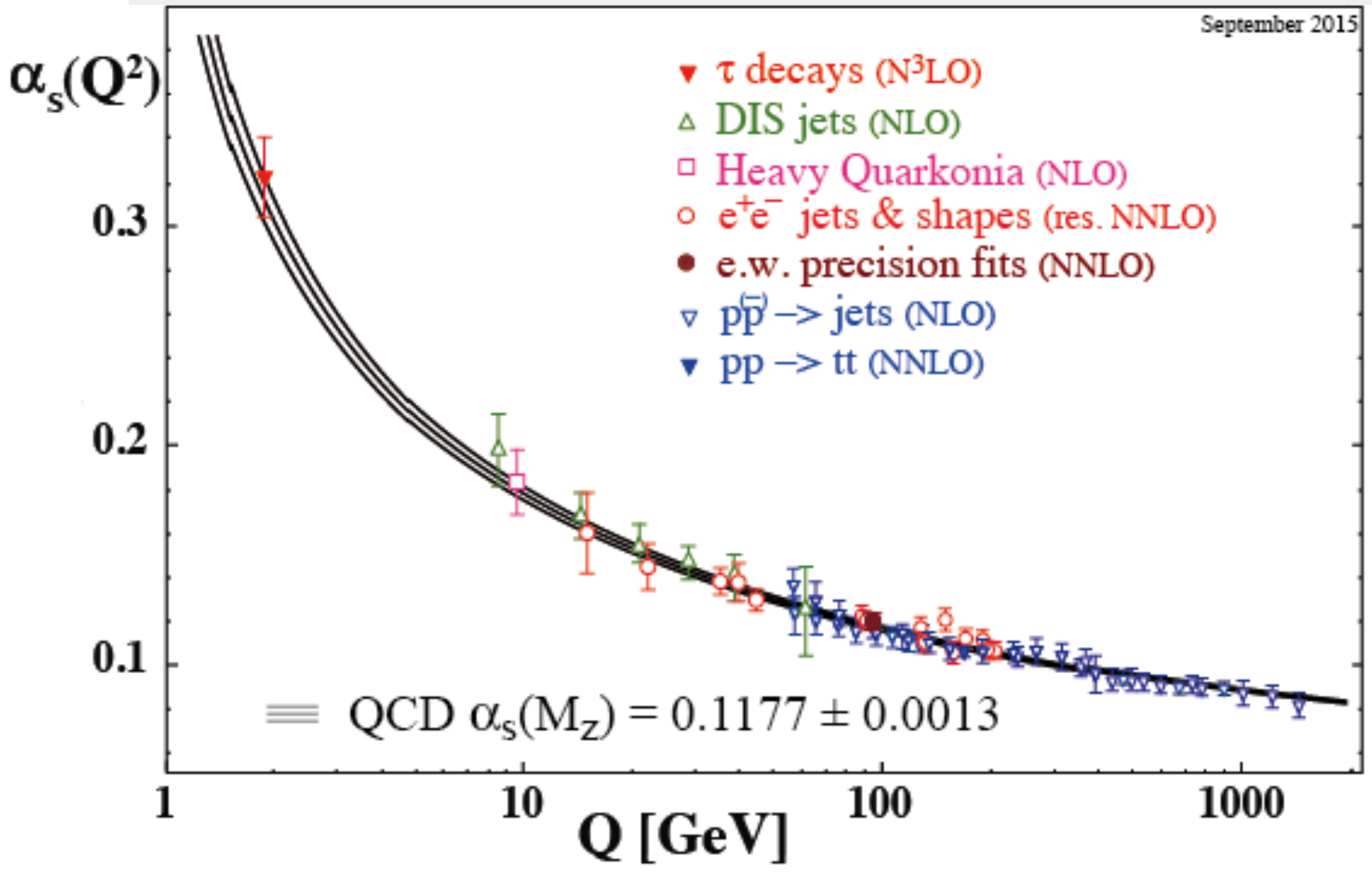
Cranmer

Wisconsin
University of Wisconsin-Madison

The strong force: Quantum Chromodynamics

Interaction strength depends on energy

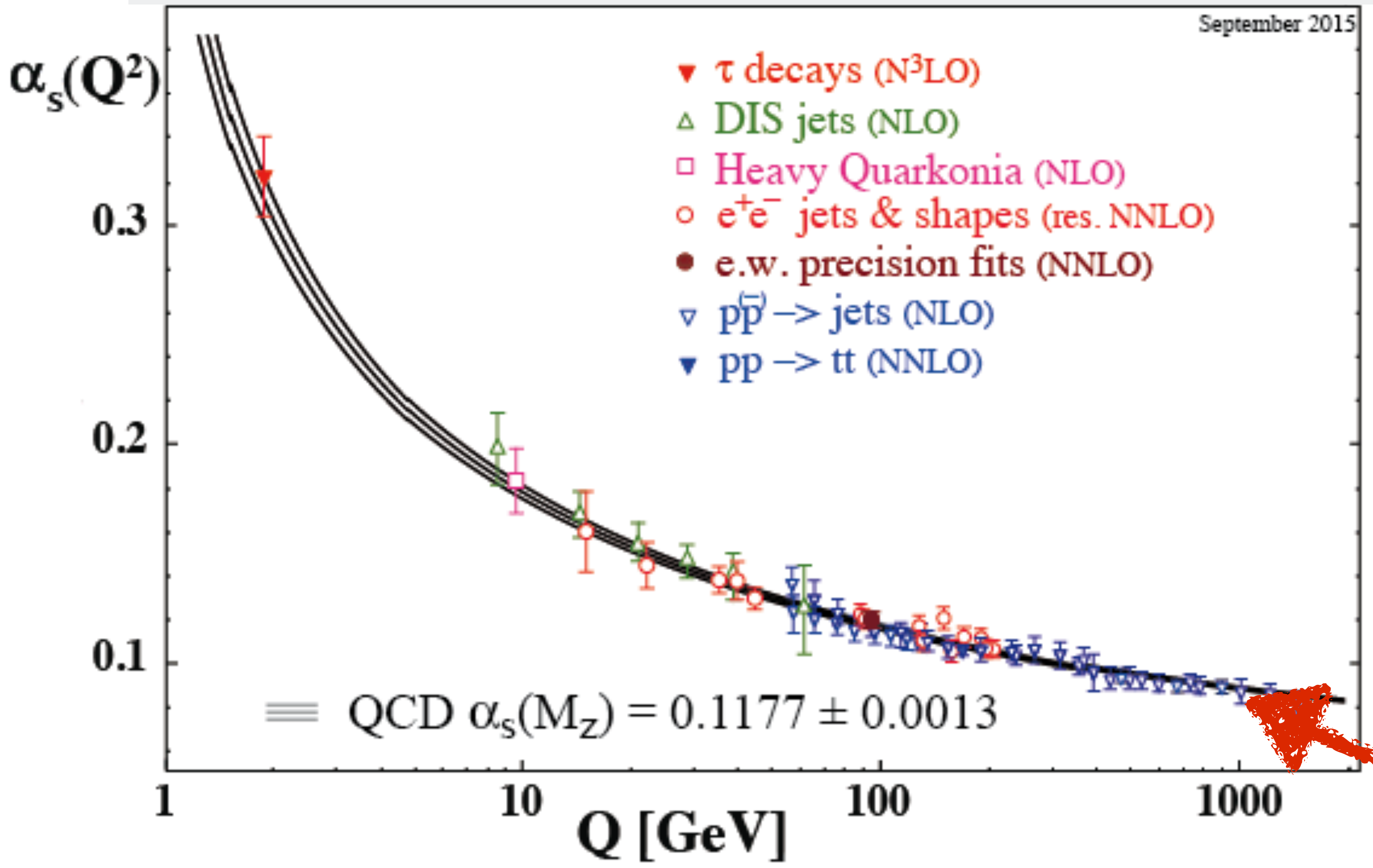
[Gross, Politzer, Wilczek, Nobel 2004]



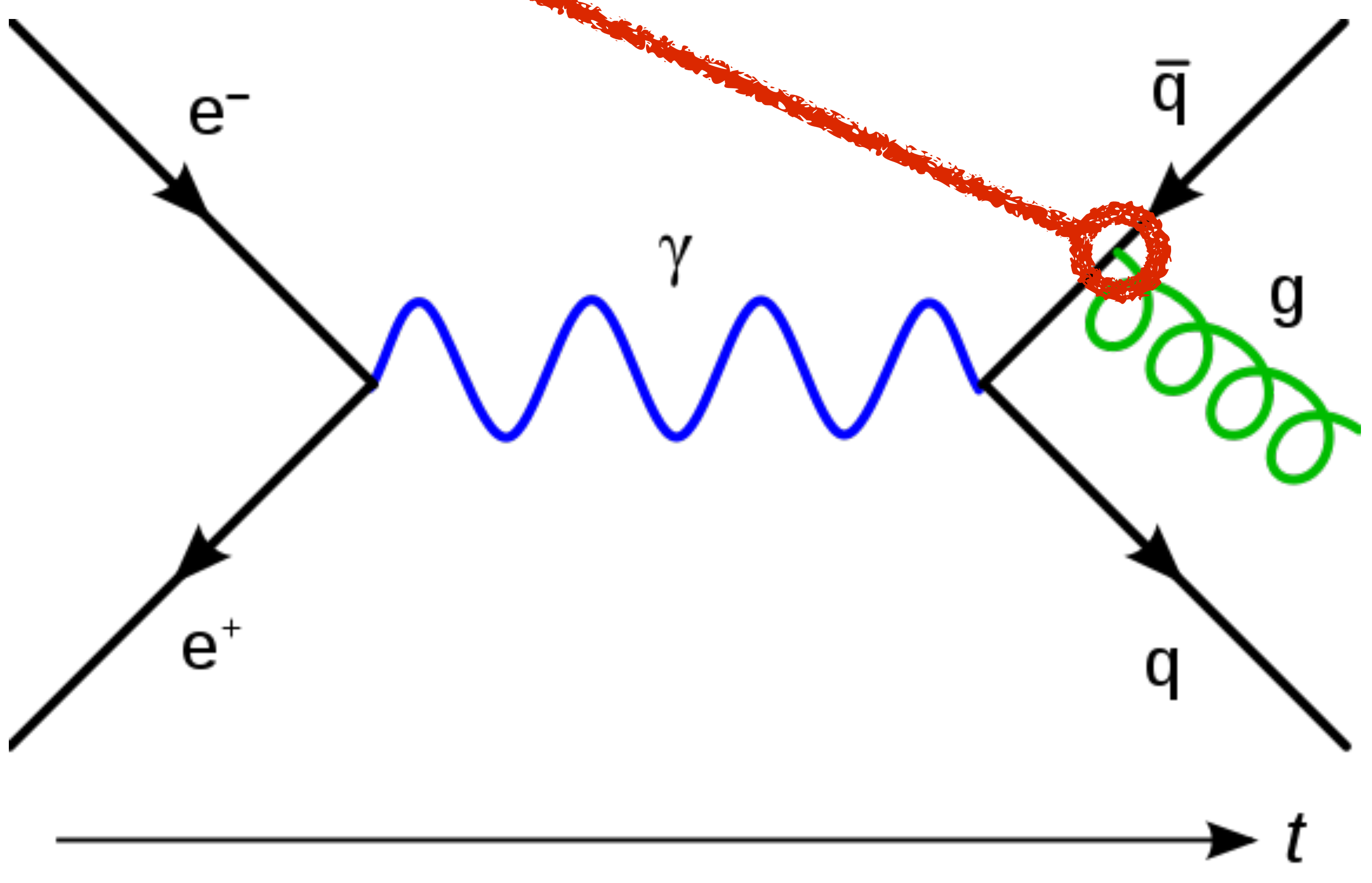
The strong force: Quantum Chromodynamics

Interaction strength depends on energy

[Gross, Politzer, Wilczek, Nobel 2004]



QCD is weak at at high-energies, small coupling, perturbation theory works



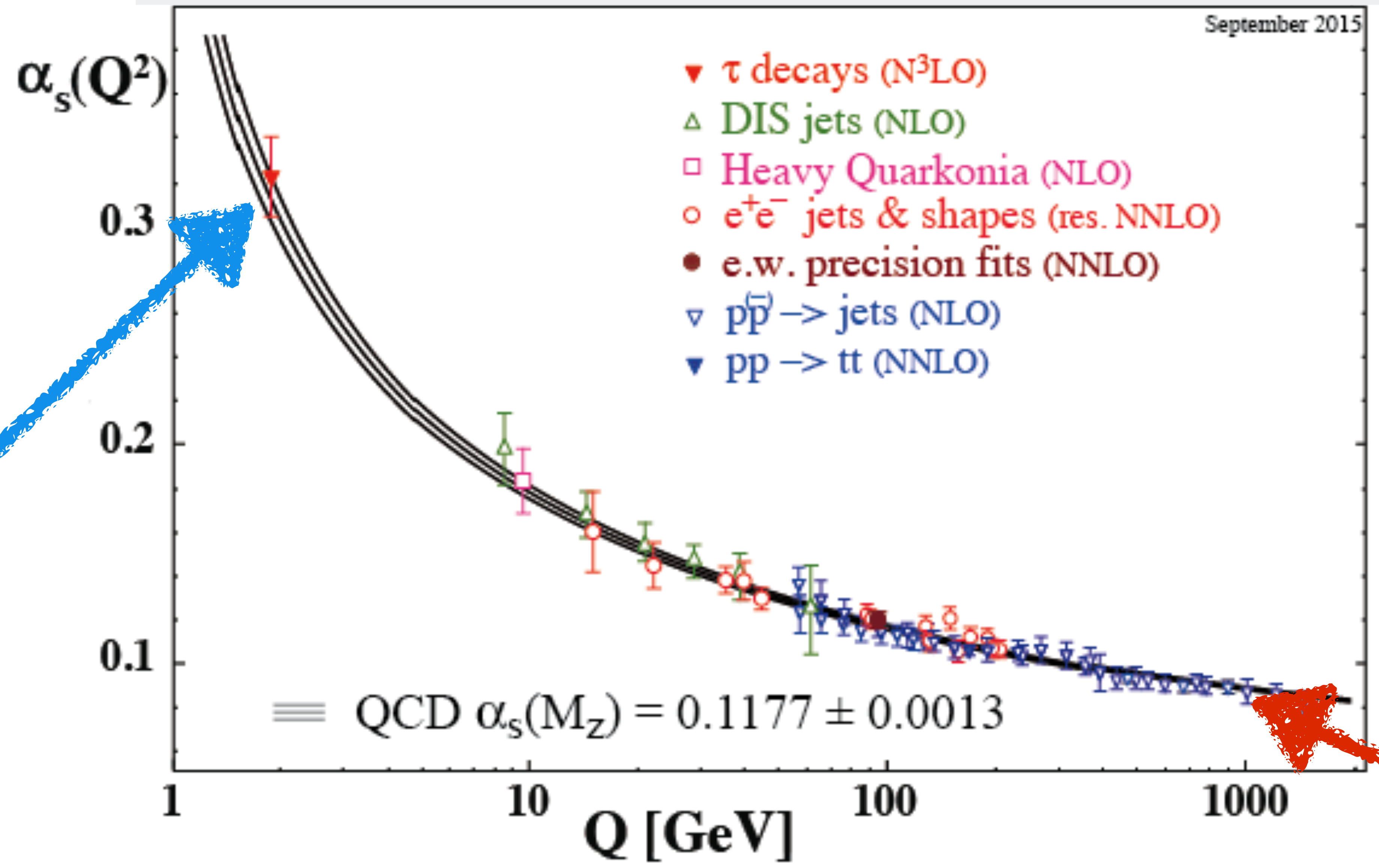
The strong force: Quantum Chromodynamics

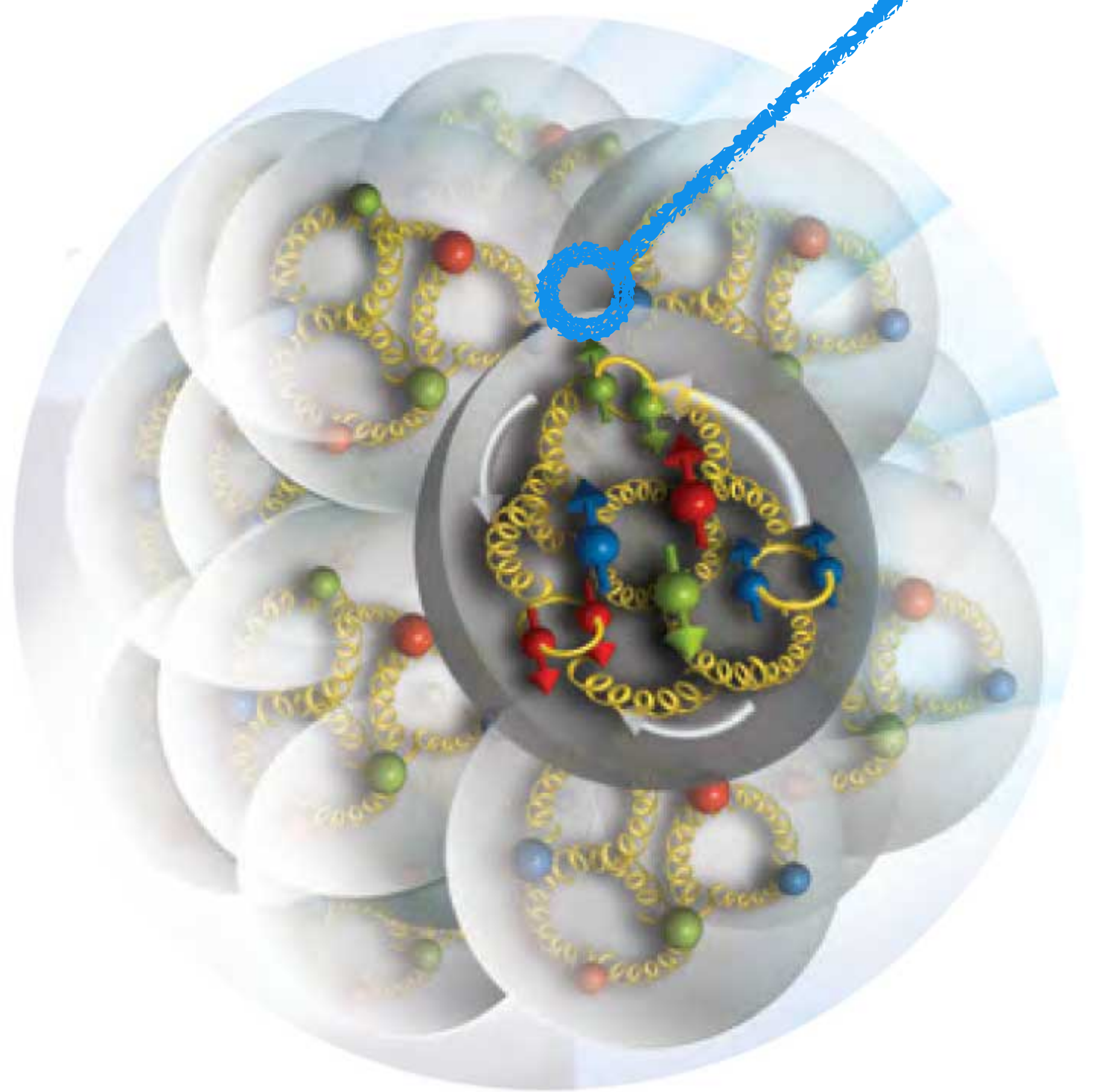
QCD is strong at at low-energies, no small coupling, perturbation theory fails.

Emergent phenomena: protons, pions, etc.

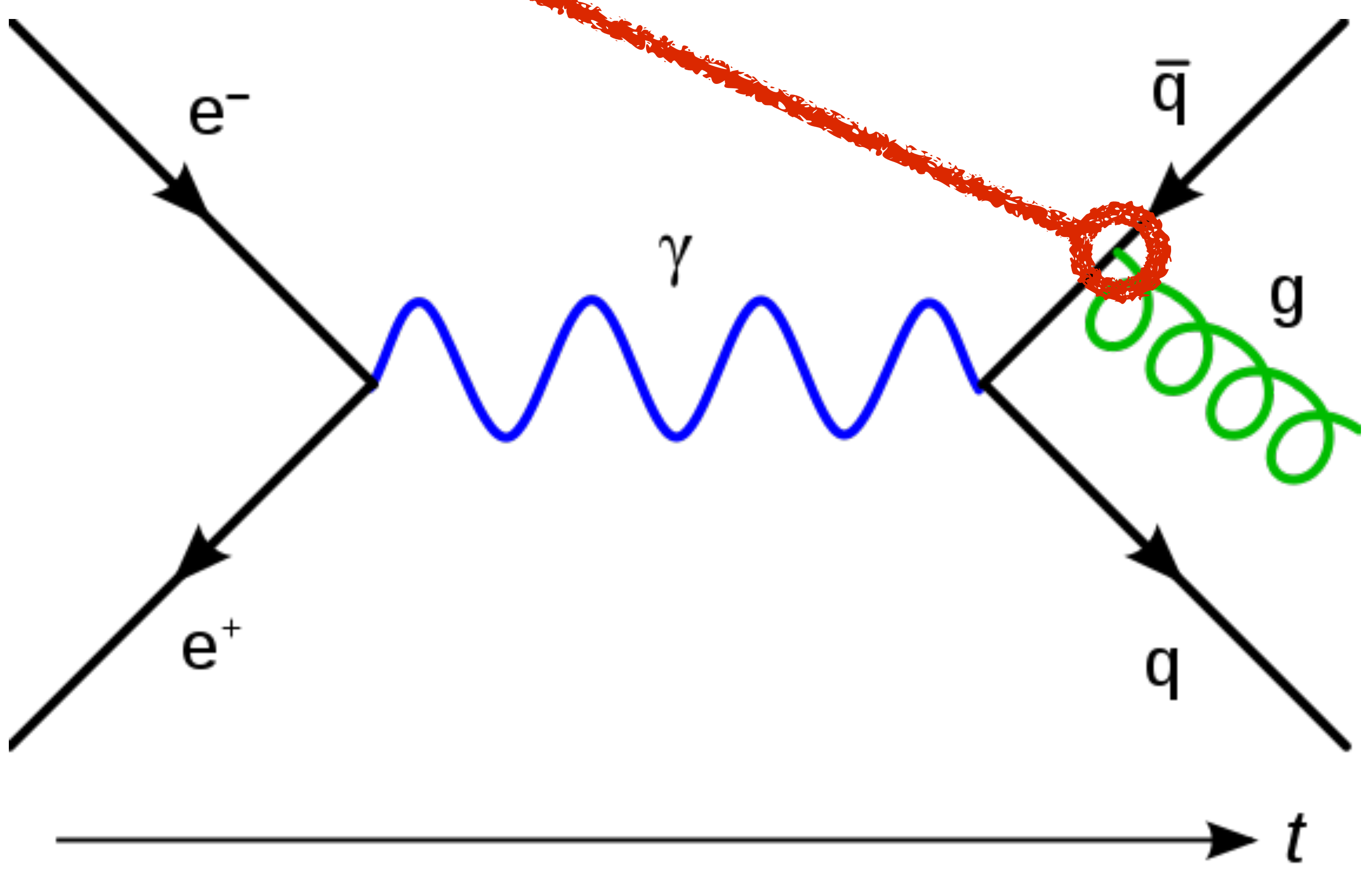
Interaction strength depends on energy

[Gross, Politzer, Wilczek, Nobel 2004]





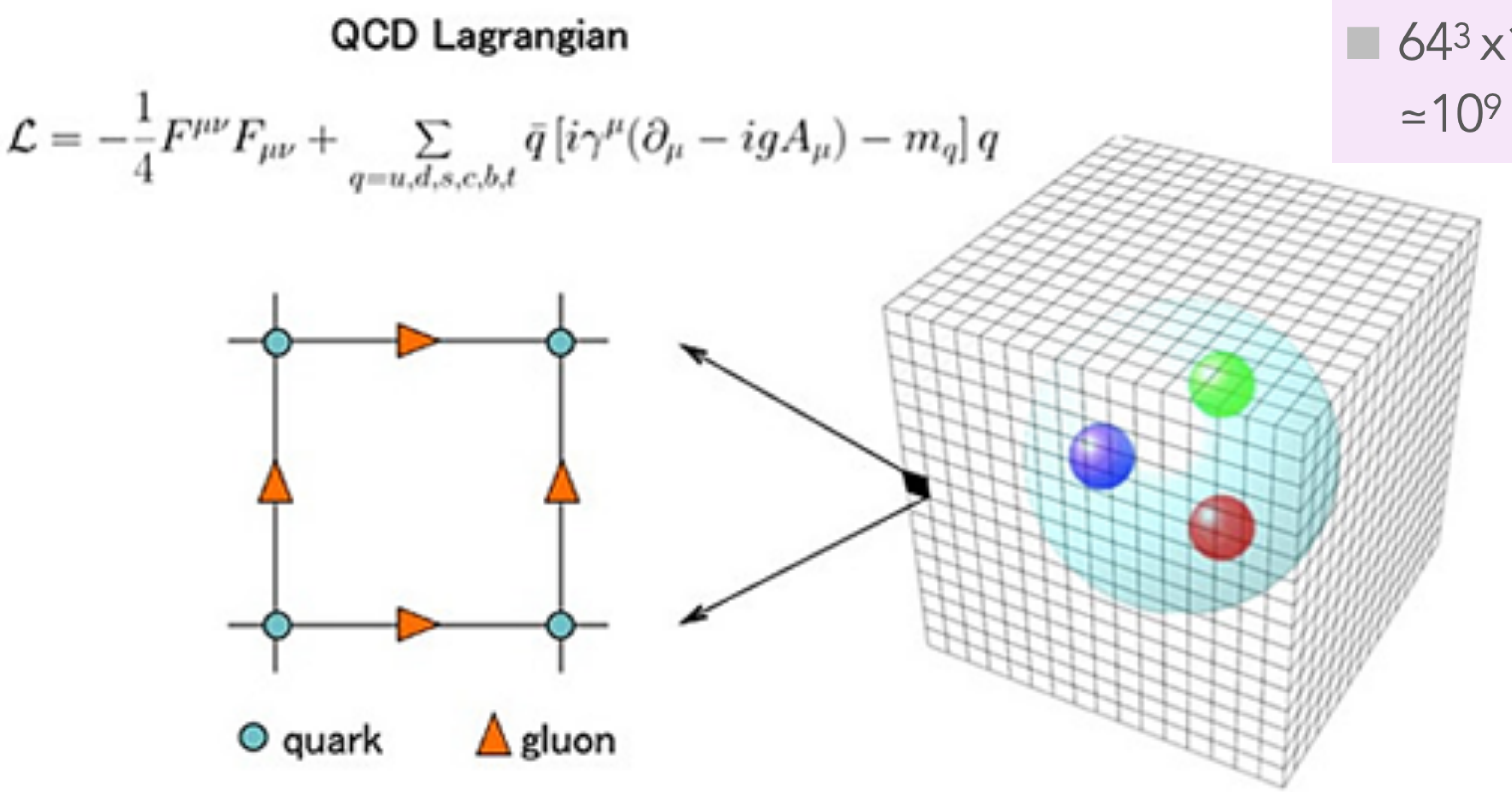
QCD is weak at at high-energies, small coupling, perturbation theory works



Lattice Field Theory

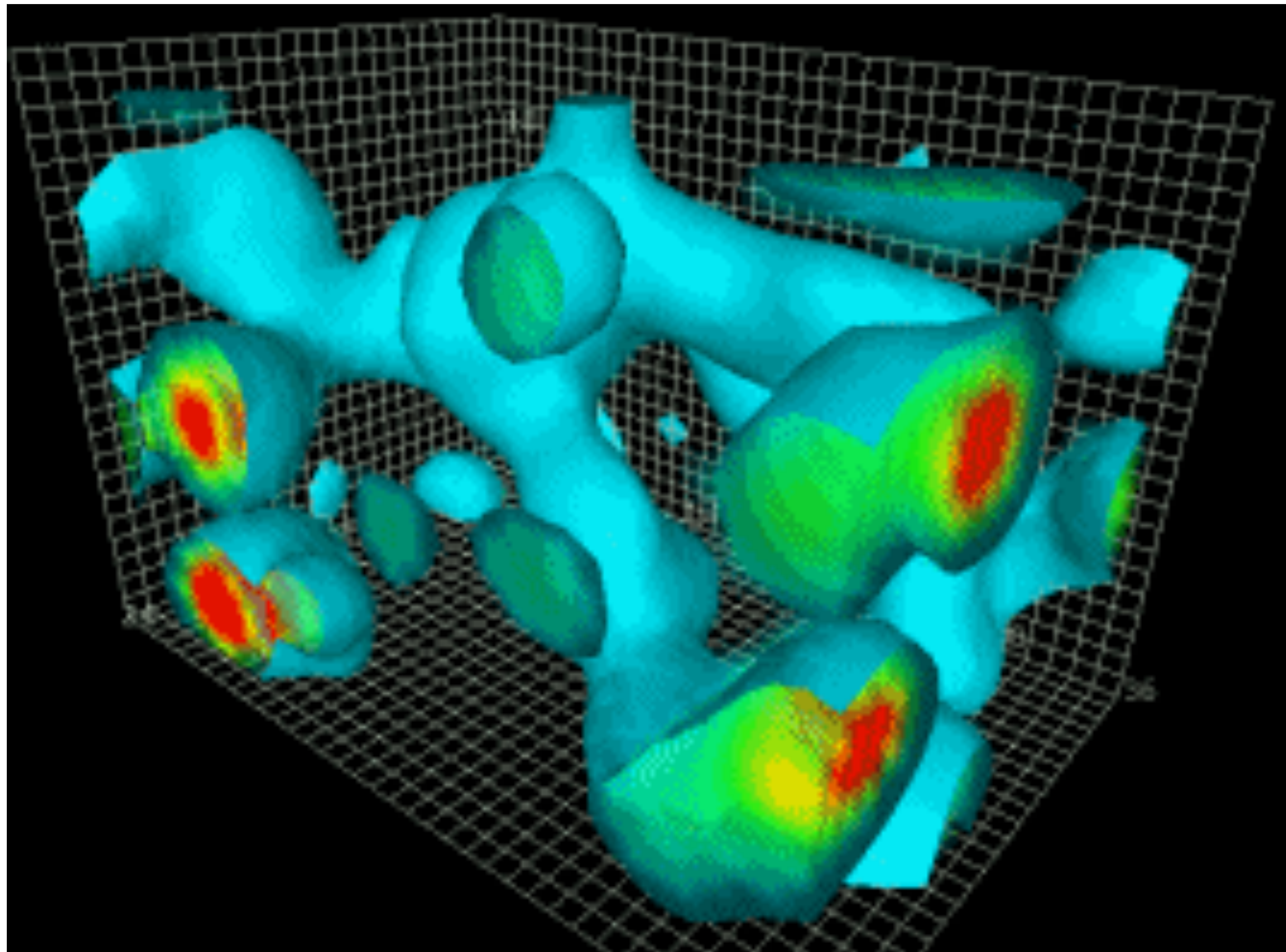
Lattice field theory is a computational approach to studying interacting field theory on a discretized space-time lattice.

Each link on the lattice has data corresponding to the symmetry group of the theory. For the strong force (QCD) each link has a 3x3 unitary matrix.



- $64^3 \times 128 \times 4 \times 9 \times 2$ $\simeq 10^9$ numbers

This animation is a single configuration of the lattice. Think of a 4-d image playing like a movie.

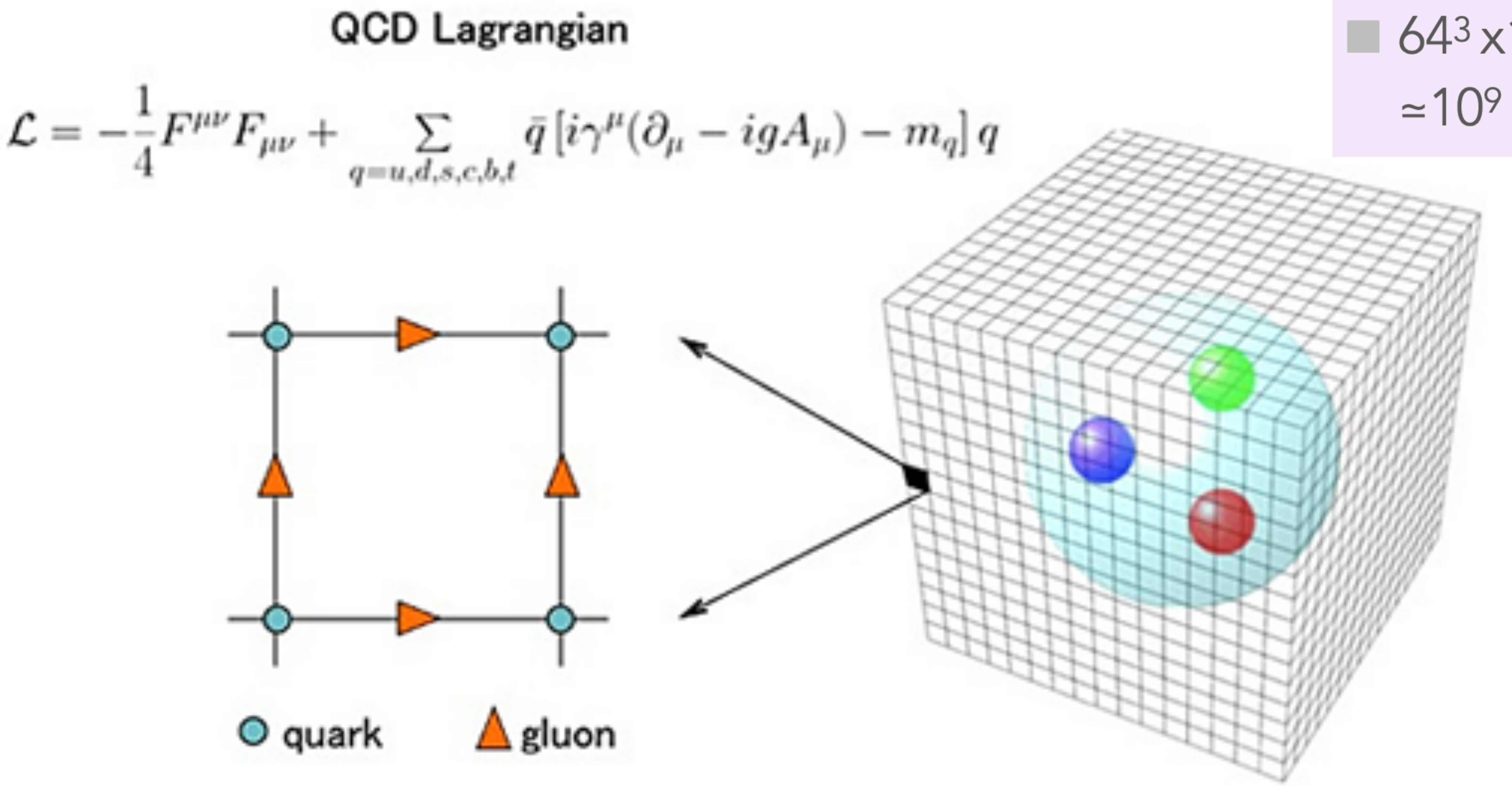


Lattice Field Theory

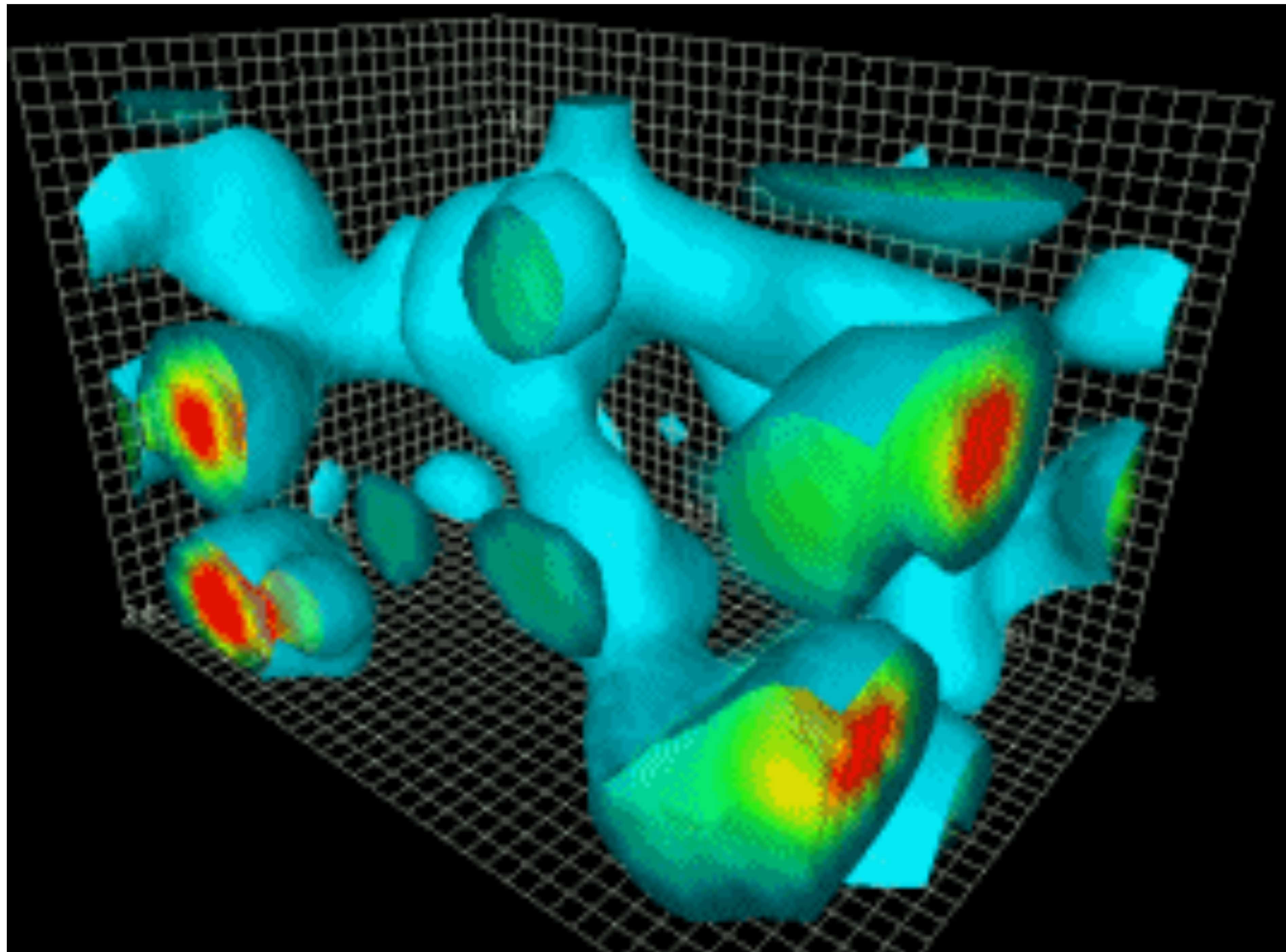
Lattice field theory is a computational approach to studying interacting field theory on a discretized space-time lattice.

Each link on the lattice has data corresponding to the symmetry group of the theory. For the strong force (QCD) each link has a 3x3 unitary matrix.

This animation is a single configuration of the lattice. Think of a 4-d image playing like a movie.



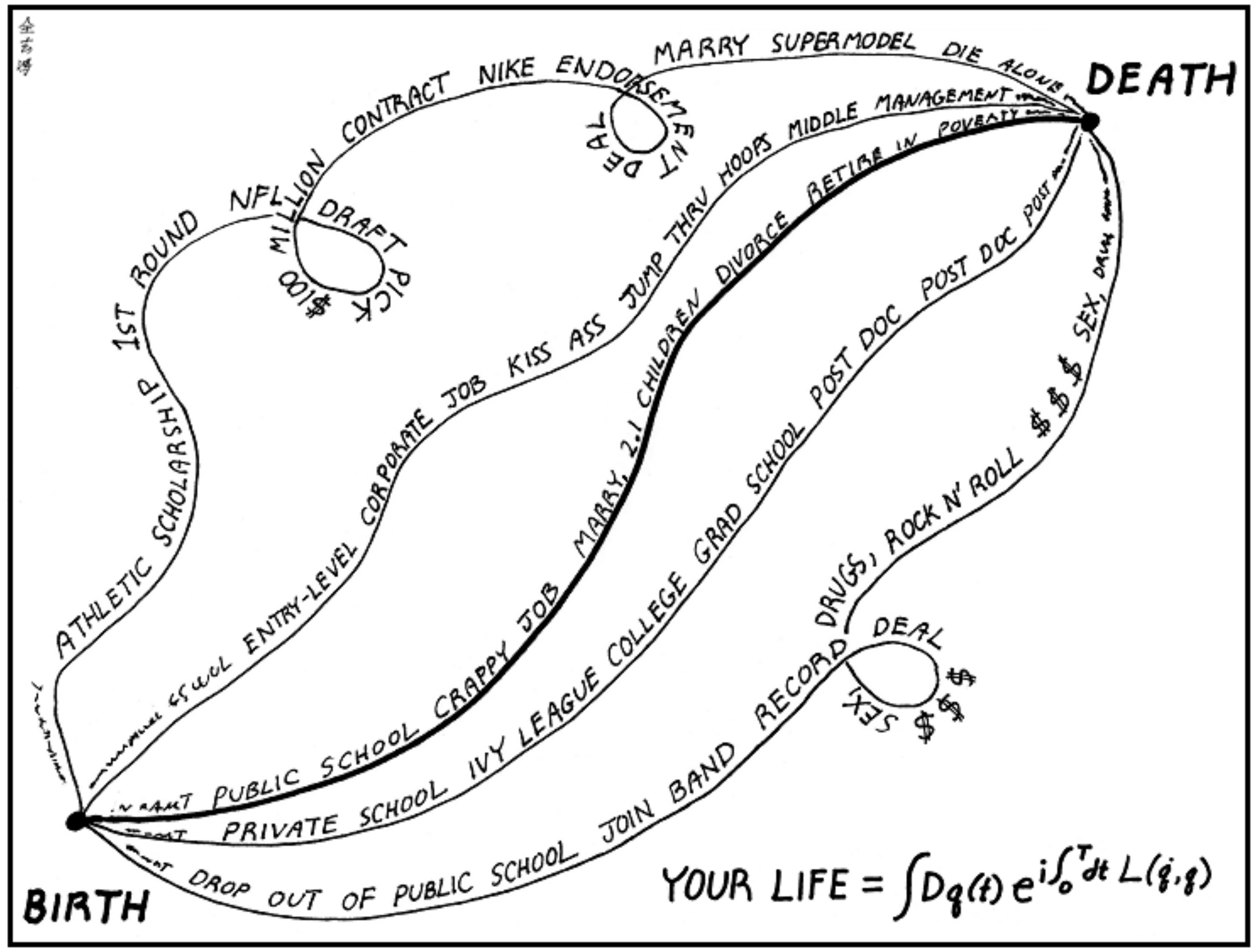
- 64^3 x128 x 4 x 9 x 2 $\simeq 10^9$ numbers



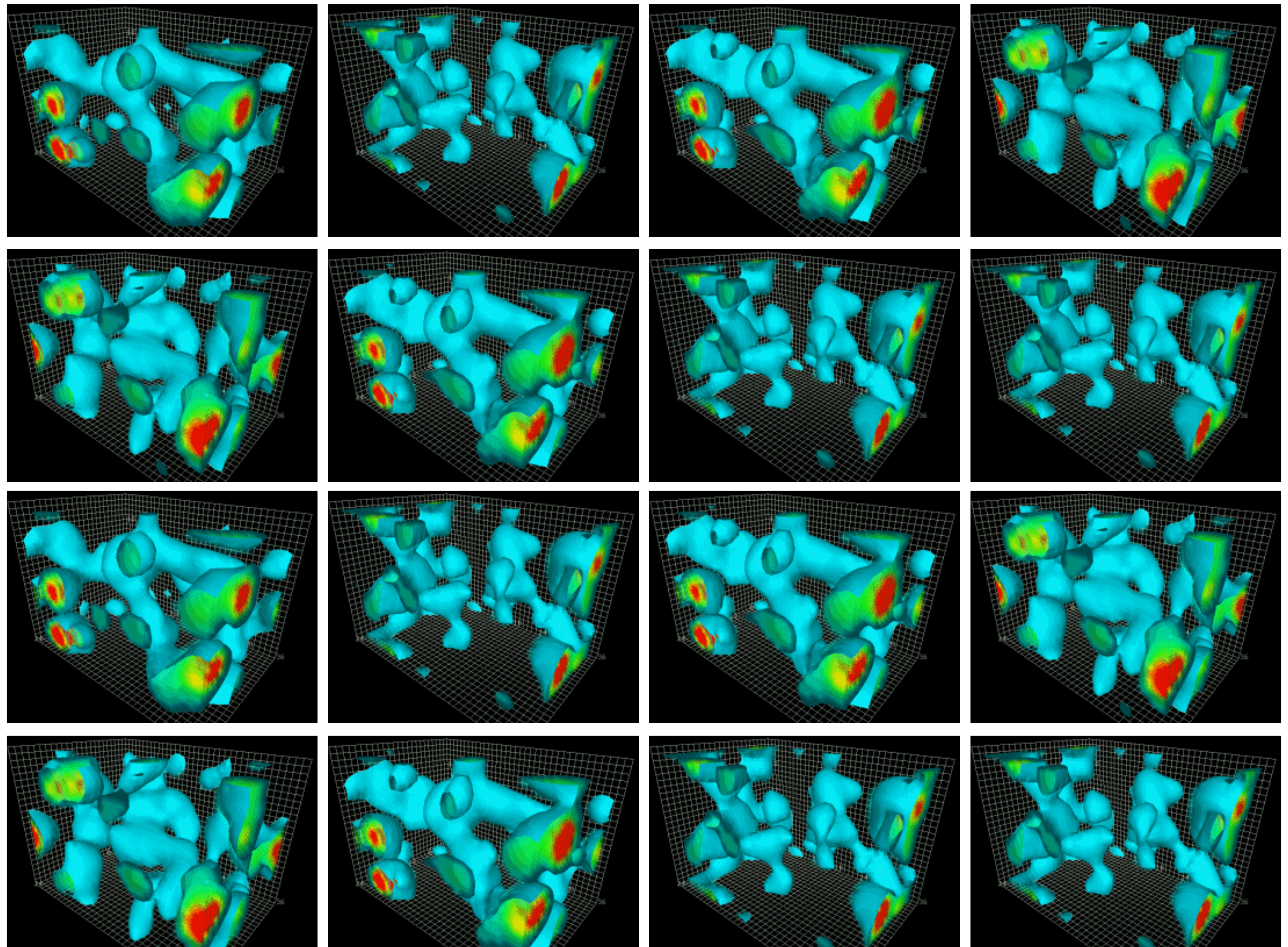
Distribution over Configurations

We don't want just a single "image" (lattice configuration), we want to sample the high-dimensional distribution of configurations predicted by the theory.

- **Path integral**: each "path" is a sample from distribution of lattice configurations path ~ exp(-Action[path]).
- Predictions are expectations of quantum operators w.r.t. this distribution.
- That integral is intractable. Typically people use Hamiltonian Monte Carlo for this, but it has limitations.



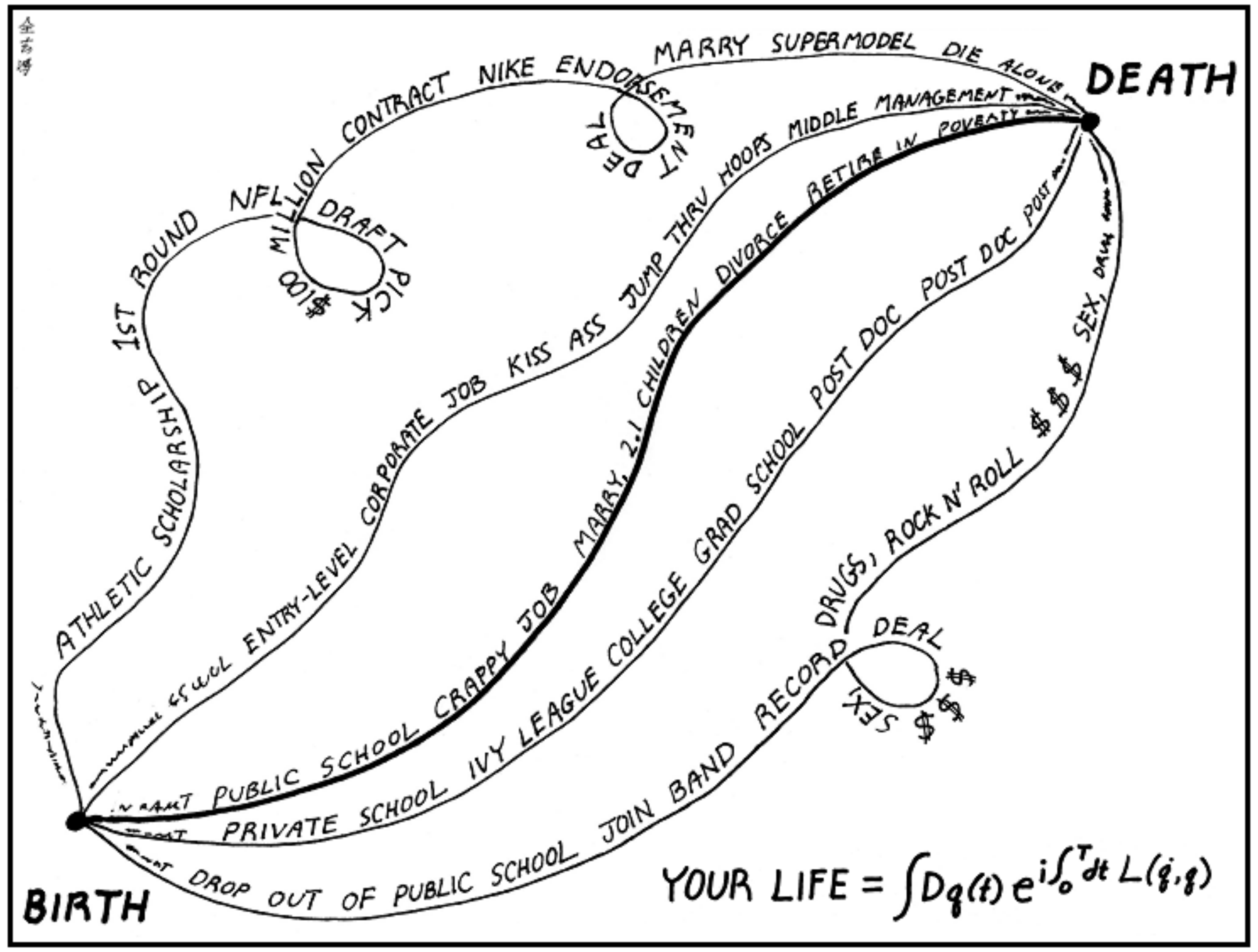
The Path Integral Formulation of Your Life



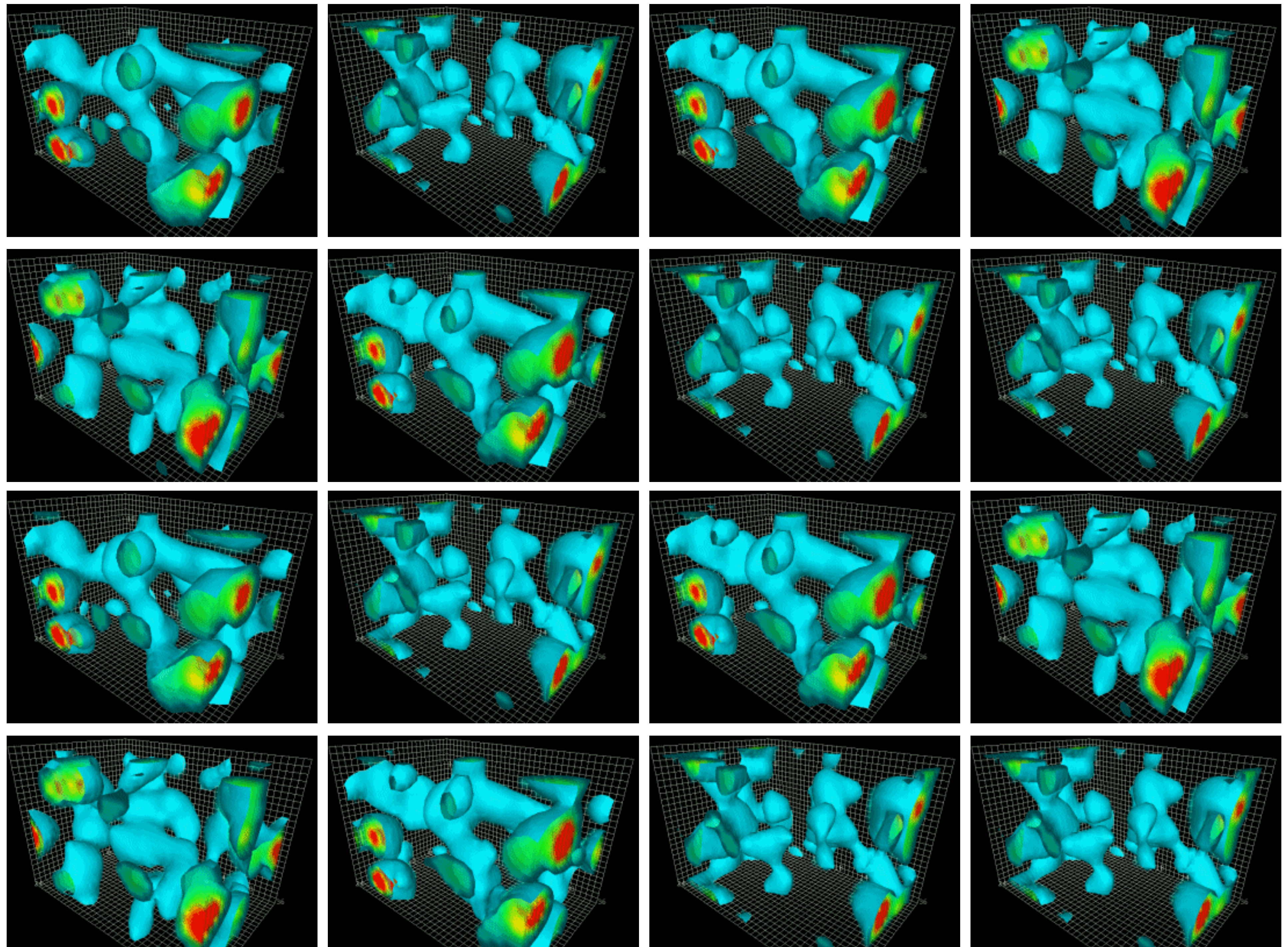
Distribution over Configurations

We don't want just a single "image" (lattice configuration), we want to sample the high-dimensional distribution of configurations predicted by the theory.

- **Path integral**: each "path" is a sample from distribution of lattice configurations path ~ exp(-Action[path]).
- Predictions are expectations of quantum operators w.r.t. this distribution.
- That integral is intractable. Typically people use Hamiltonian Monte Carlo for this, but it has limitations.

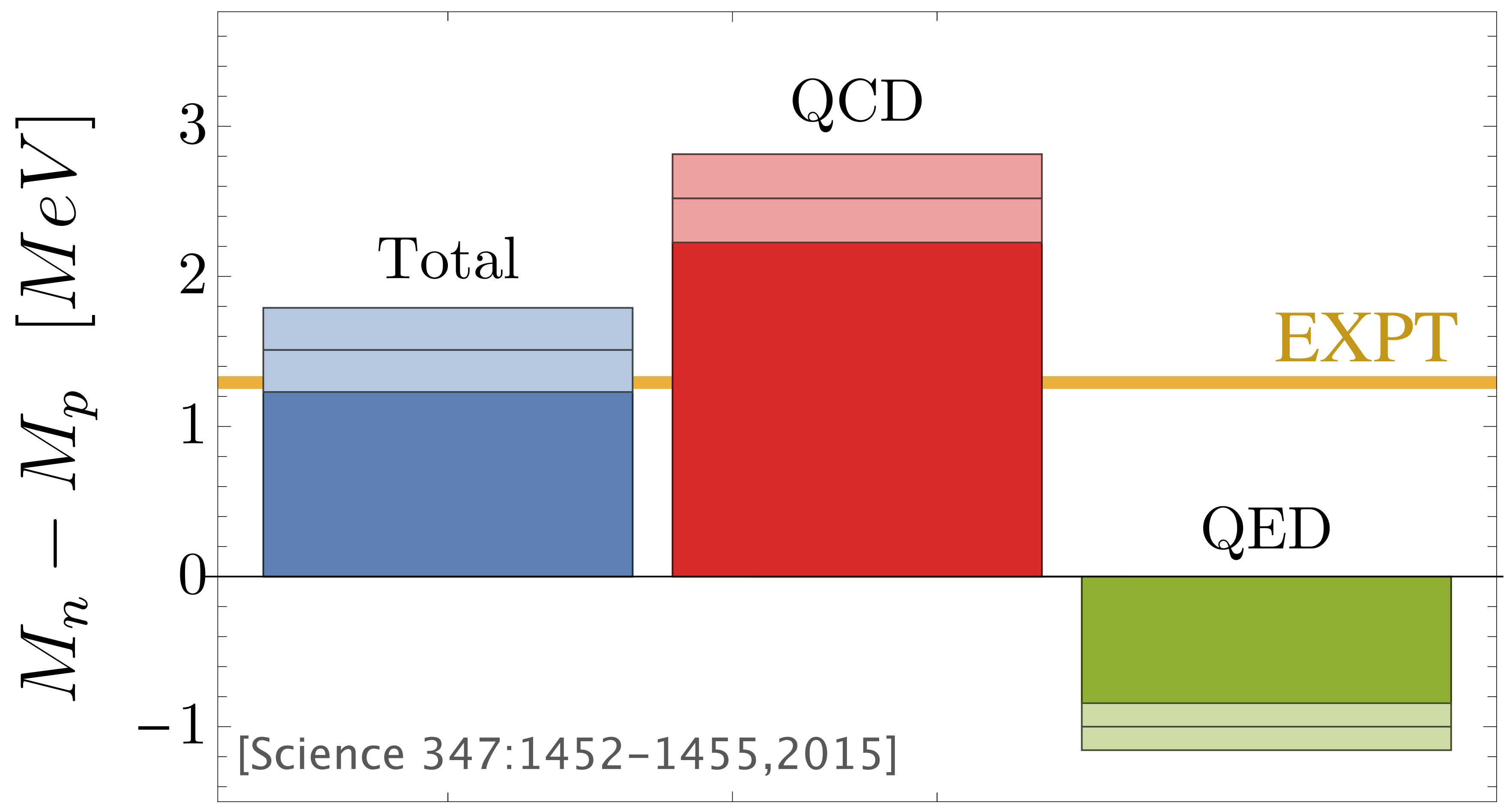


The Path Integral Formulation of Your Life

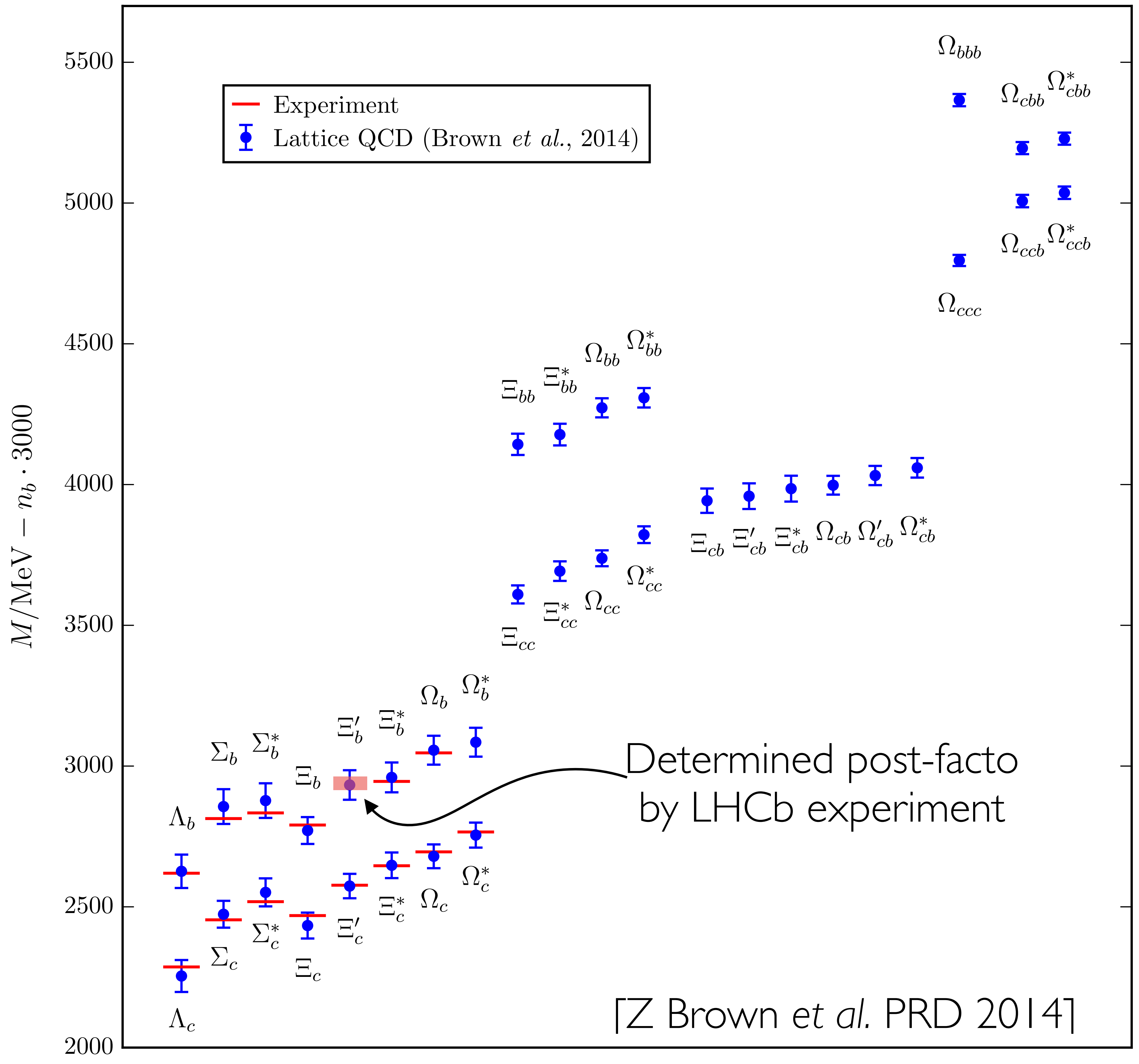


Lattice QCD works

- Ground state hadron spectrum reproduced
- p-n mass splitting reproduced
- …

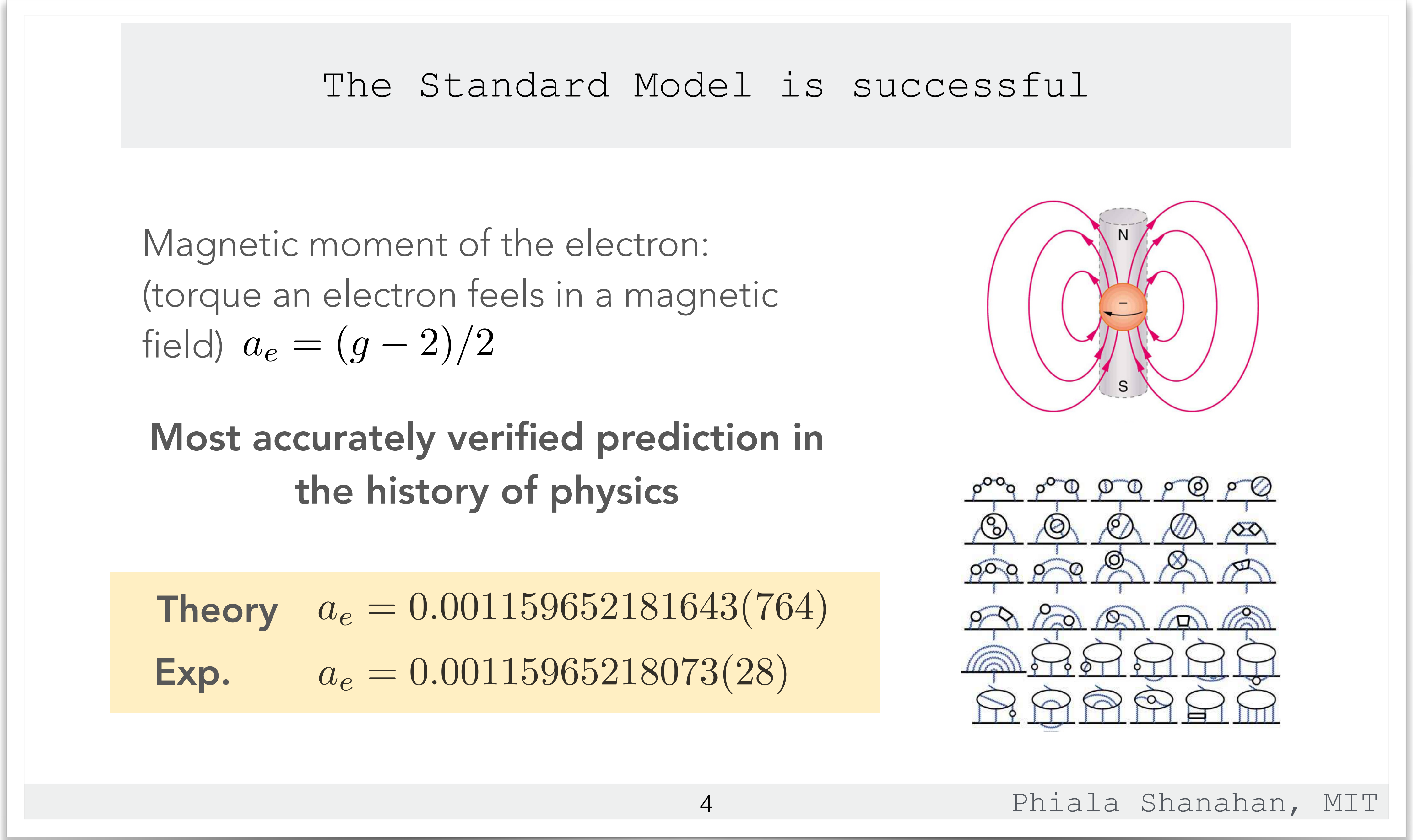


- Predictions for new states with controlled uncertainties



Phiala Shanahan, MIT

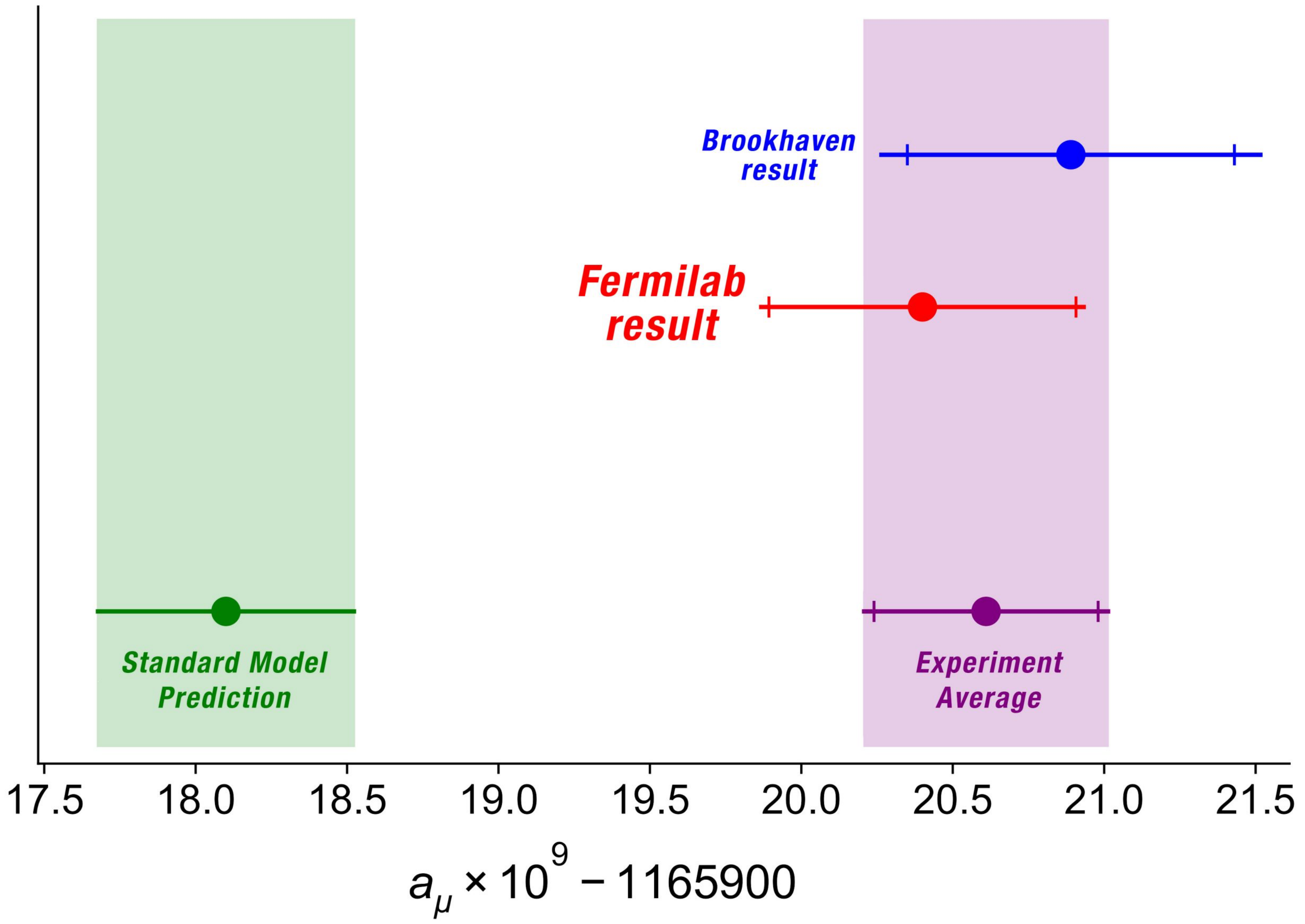
Predictions are taken seriously

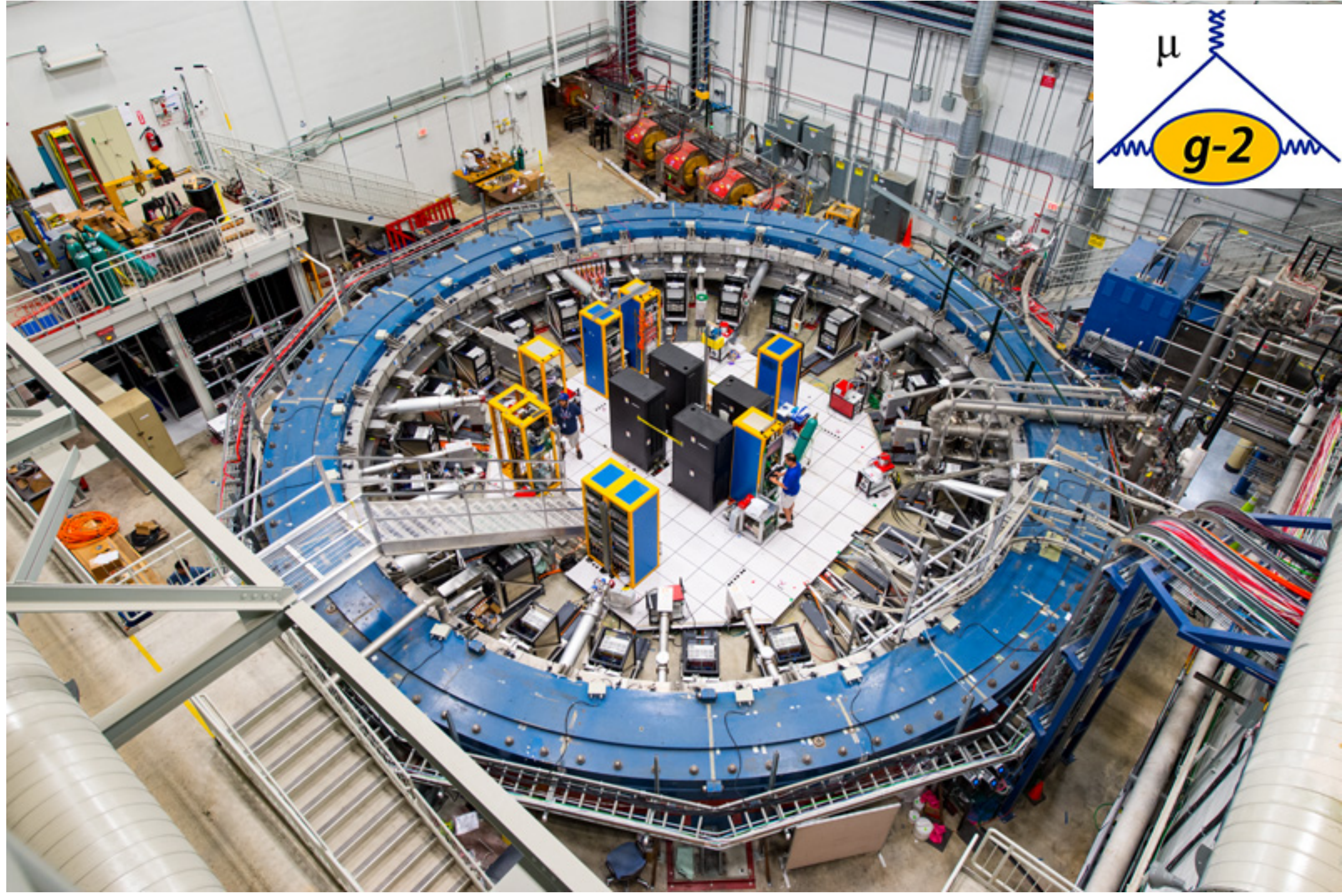

The Standard Model is successful

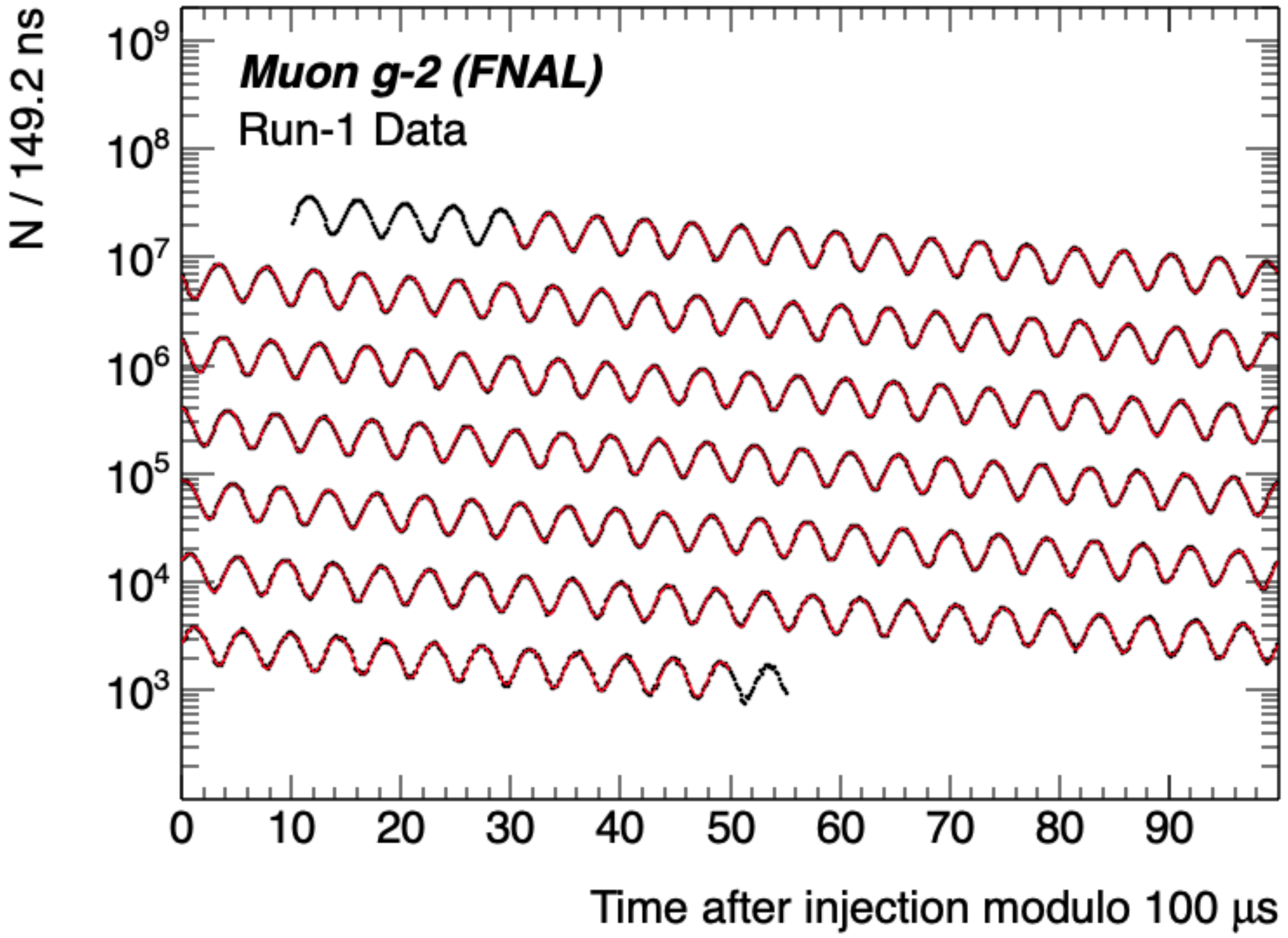
Magnetic moment of the electron:
(torque an electron feels in a magnetic field) $a_e = (g-2)/2$

Most accurately verified prediction in the history of physics

Theory	$a_e = 0.001159652181643(764)$
Exp.	$a_e = 0.00115965218073(28)$



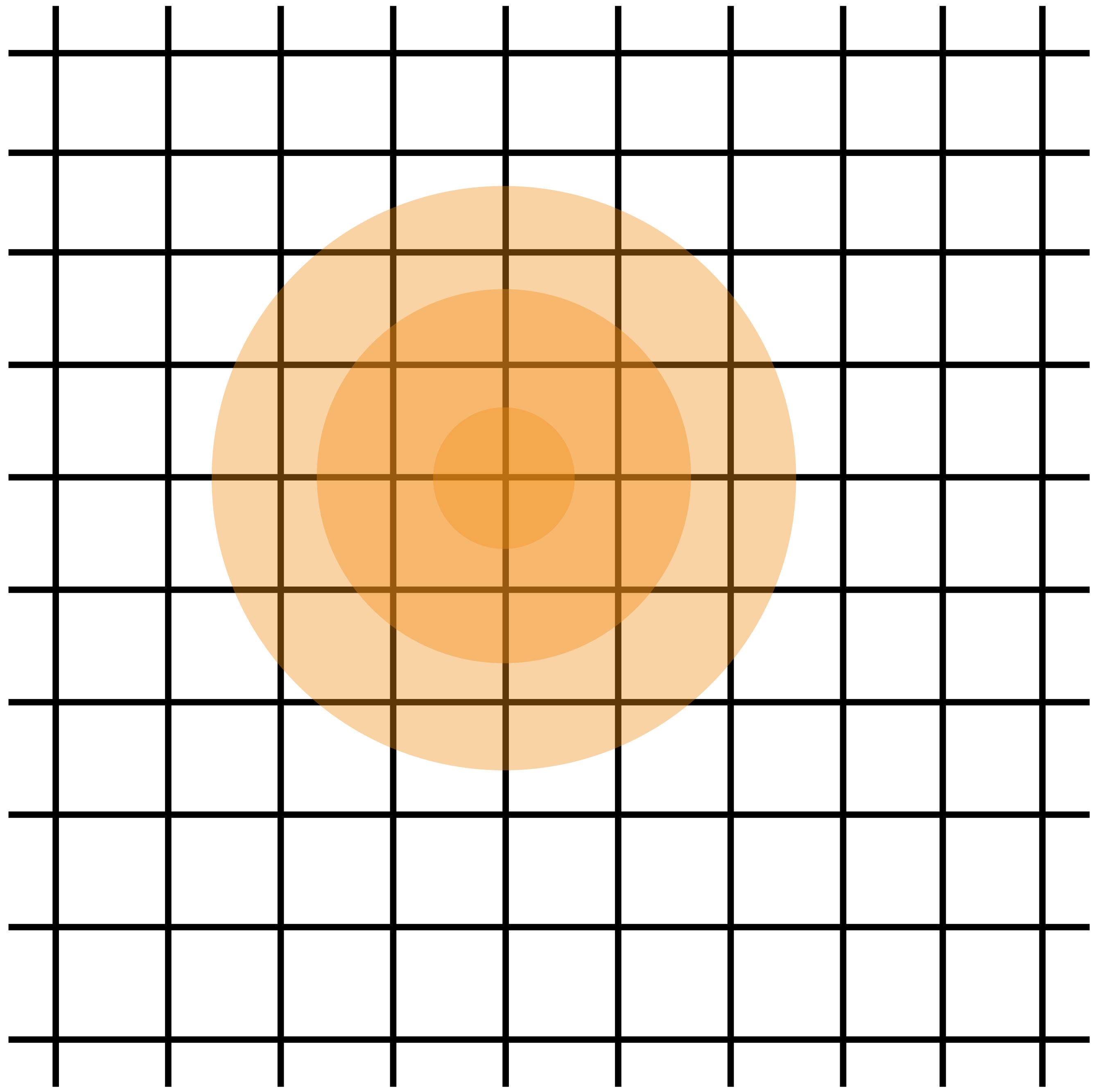




So what's the problem?

QCD gauge field configurations sampled via

Hamiltonian dynamics + Markov Chain Monte Carlo



Updates diffusive

Lattice spacing → 0

Number of updates to change fixed physical length scale → ∞

"Critical slowing-down" of generation of uncorrelated samples

Flows for LQCD

Albergo, Kanwar, Shanahan, PRD (2019) arXiv:1904.12072

Flow-based generative models for Markov chain Monte Carlo in lattice field theory

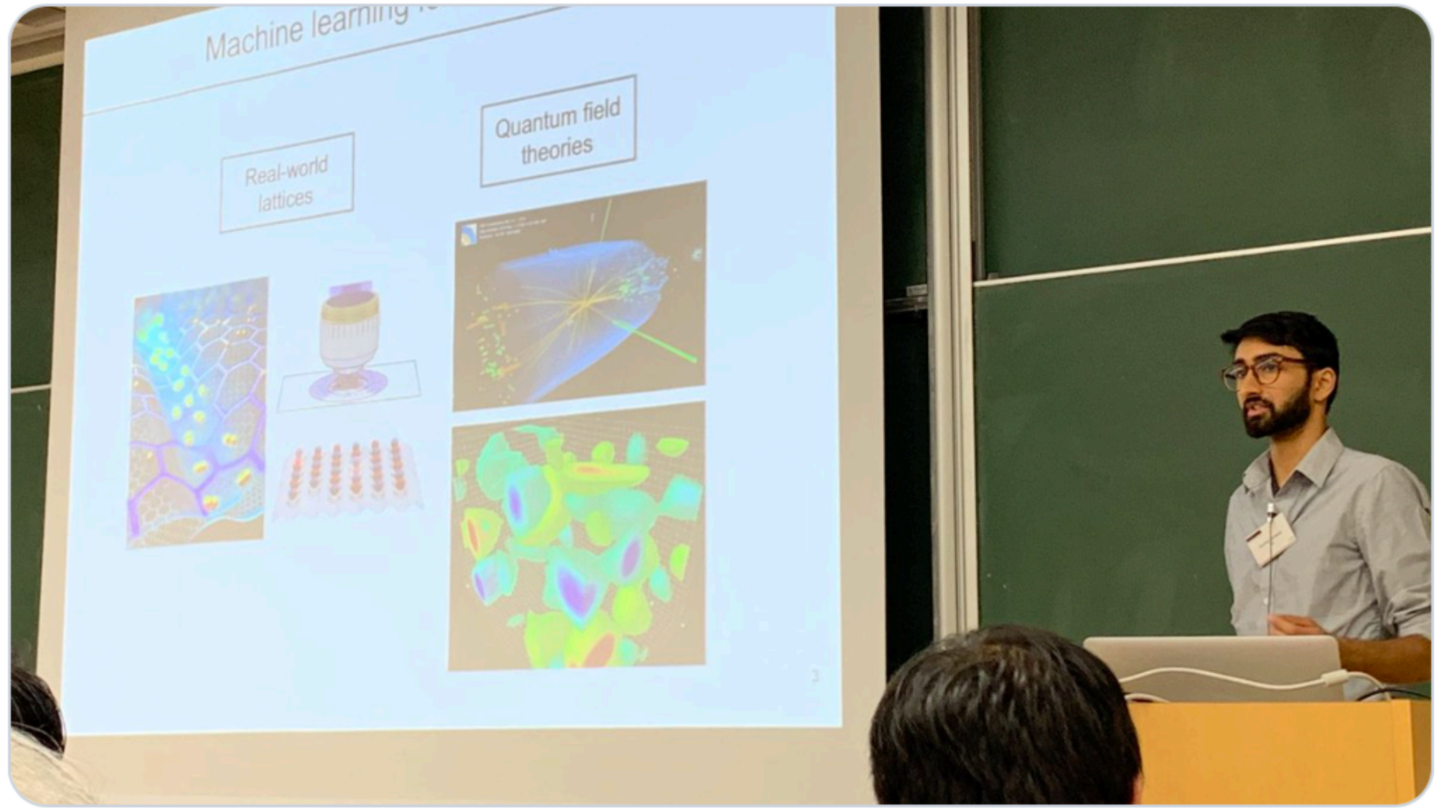
M. S. Albergo,[1,2,3] G. Kanwar,[4] and P. E. Shanahan[4,1]

[1] *Perimeter Institute for Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada*
[2] *Cavendish Laboratories, University of Cambridge, Cambridge CB3 0HE, U.K.*
[3] *University of Waterloo, Waterloo, Ontario N2L 3G1, Canada*
[4] *Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.*

A Markov chain update scheme using a machine-learned *flow-based generative model* is proposed for Monte Carlo sampling in lattice field theories. The generative model may be optimized (trained) to produce samples from a distribution approximating the desired Boltzmann distribution determined by the lattice action of the theory being studied. Training the model systematically improves autocorrelation times in the Markov chain, even in regions of parameter space where standard Markov chain Monte Carlo algorithms exhibit critical slowing down in producing decorrelated updates. Moreover, the model may be trained without existing samples from the desired distribution. The algorithm is compared with HMC and local Metropolis sampling for ϕ^4 theory in two dimensions.

Enrico Rinaldi @enricesena · Nov 1

Yesterday **Gurtej Kanwar** told us about machine learning for lattice field theories and exciting progress in Generative Models for gauge theories (collaboration with @DeepMindAI) at #DLAP2019 Today is the last day of this great conference!



5 15

RESEARCH Noé *et al.*, *Science* **365**, 1001 (2019) 6 September 2019

RESEARCH ARTICLE SUMMARY

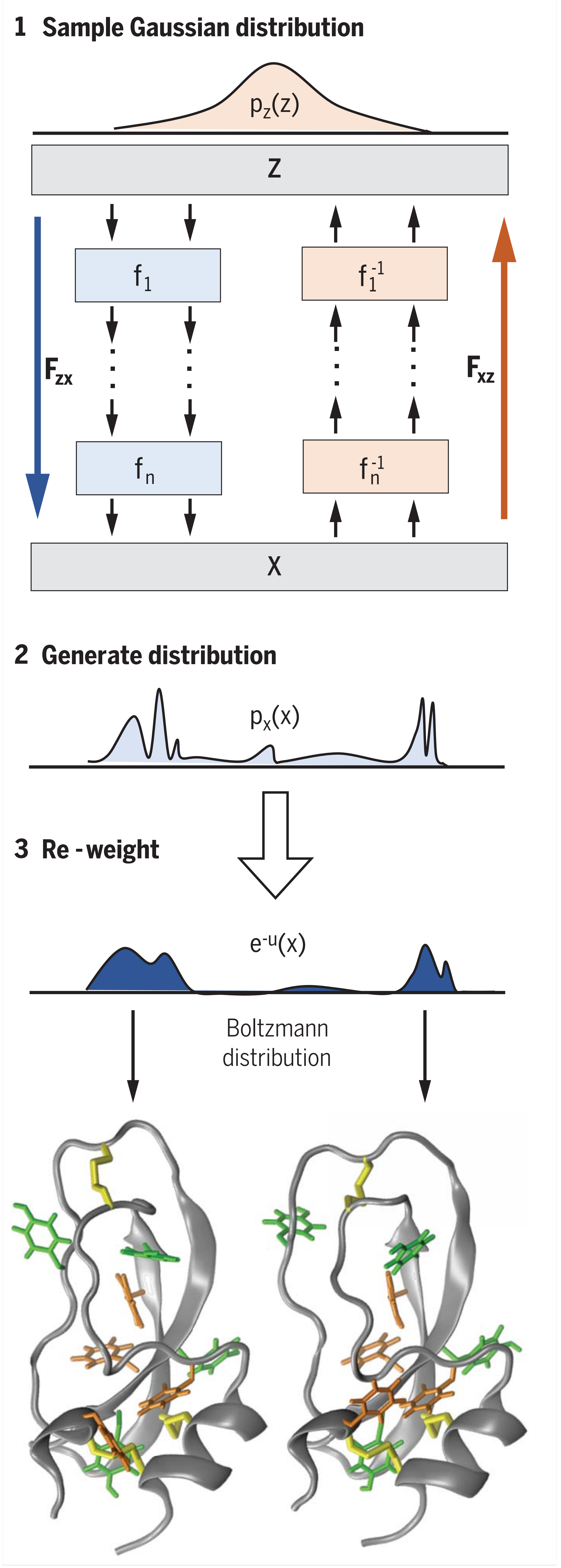
MACHINE LEARNING

Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning

Frank Noé*†, Simon Olsson*, Jonas Köhler*, Hao Wu

The main approach is thus to start with one configuration, e.g., the folded protein state, and make tiny changes to it over time, e.g., by using Markov-chain Monte Carlo or molecular dynamics (MD). However, these simulations get trapped in metastable (long-lived) states: For example, sampling a single folding or unfolding event with atomistic MD may take a year on a supercomputer.

Boltzmann generators overcome sampling problems between long-lived states.

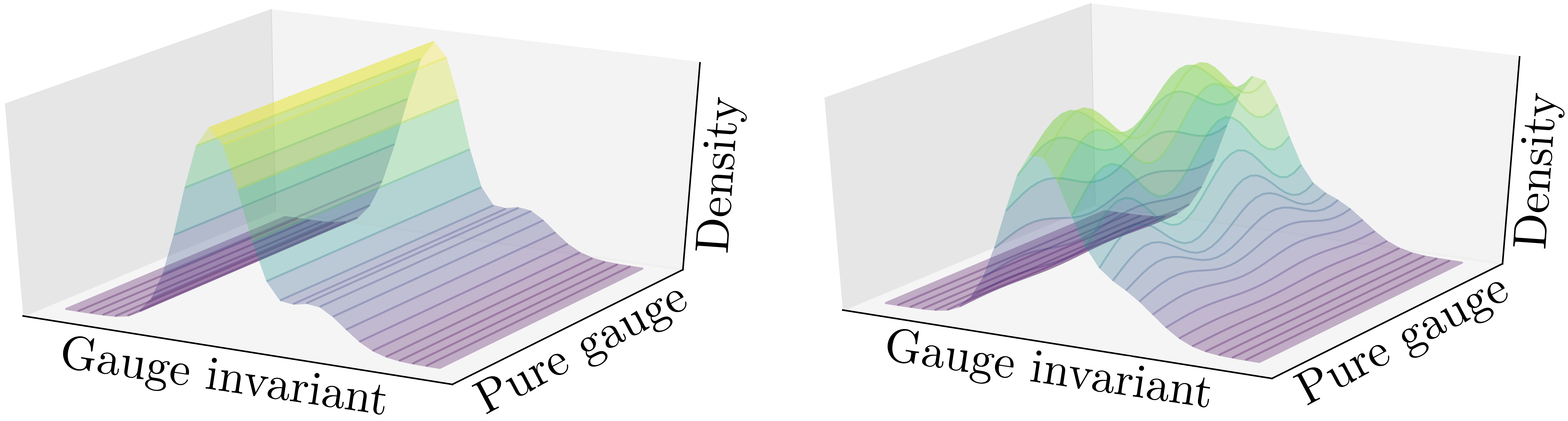


Space-time & Local Gauge Symmetry

The action is invariant to **local** gauge transformations, so the distribution is constant in those directions. It's a huge product group!

Many more pure gauge degrees of freedom than physical ones

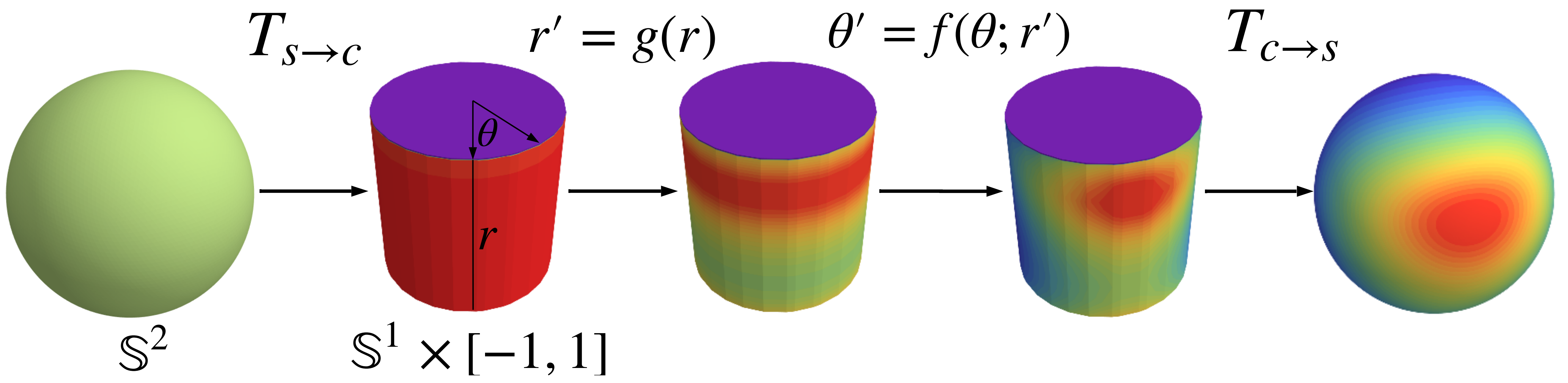
We would like to enforce this symmetry in the network, and not have to learn it.



Step 1: Flows on Spheres and Tori

[arXiv:2002.02428, ICML2020]

We designed flows on compact manifolds like Spheres and Tori that correspond to Lie groups:



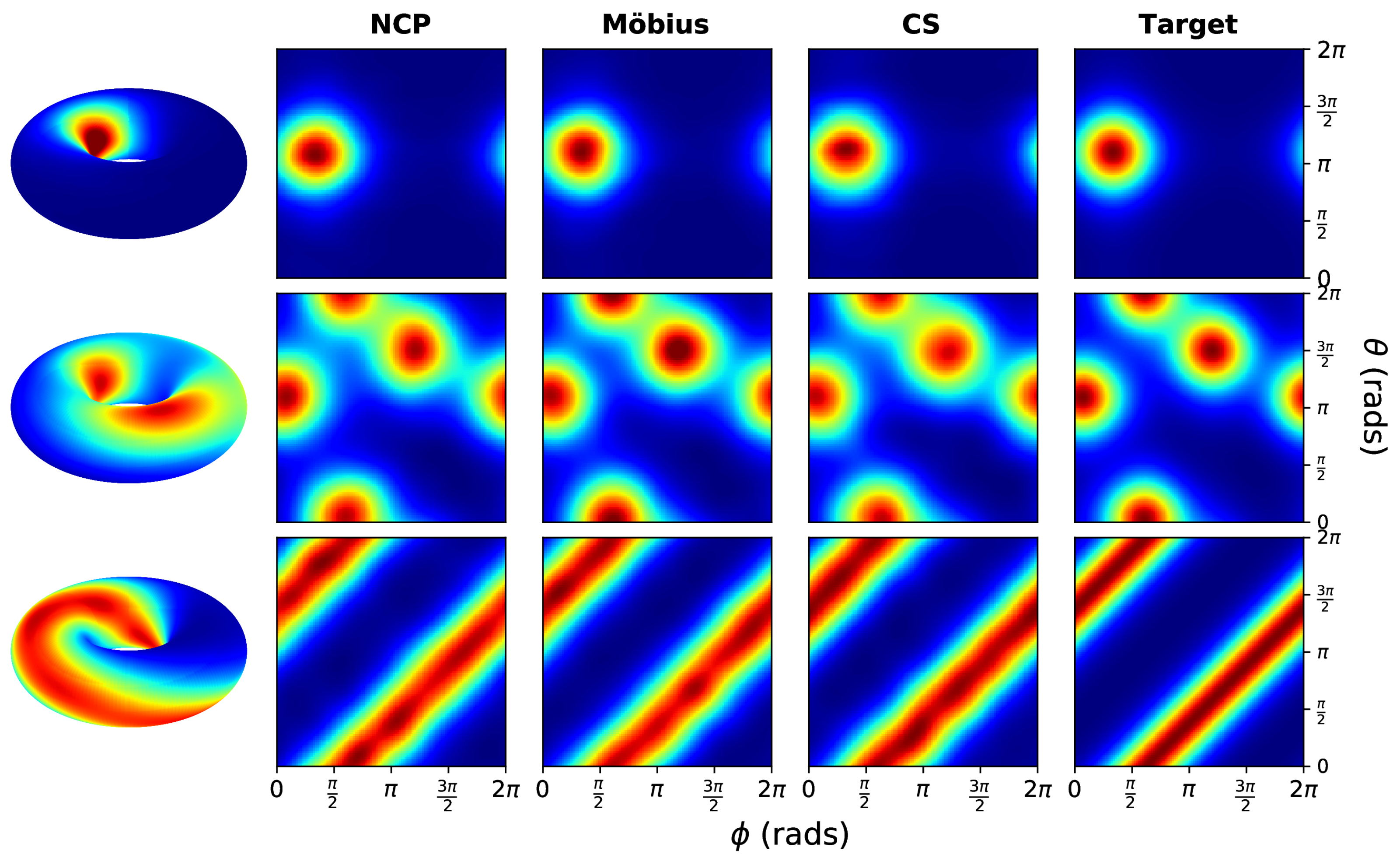


Figure 3. Learned densities on $\mathbb{T}^2$ using NCP, Möbius and CS flows. Densities shown on the torus are from NCP.

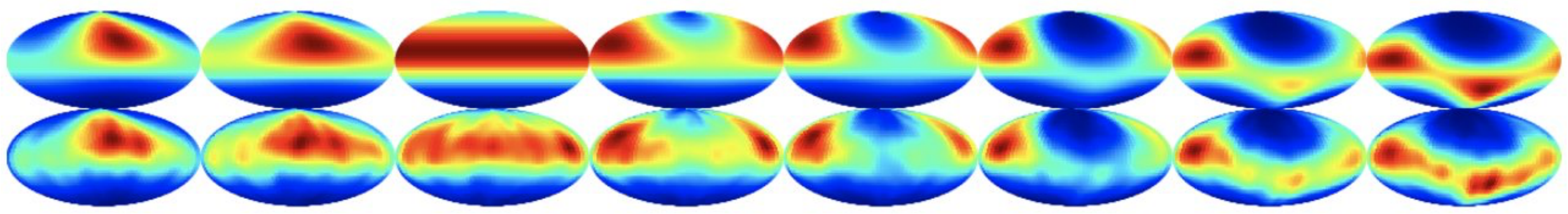
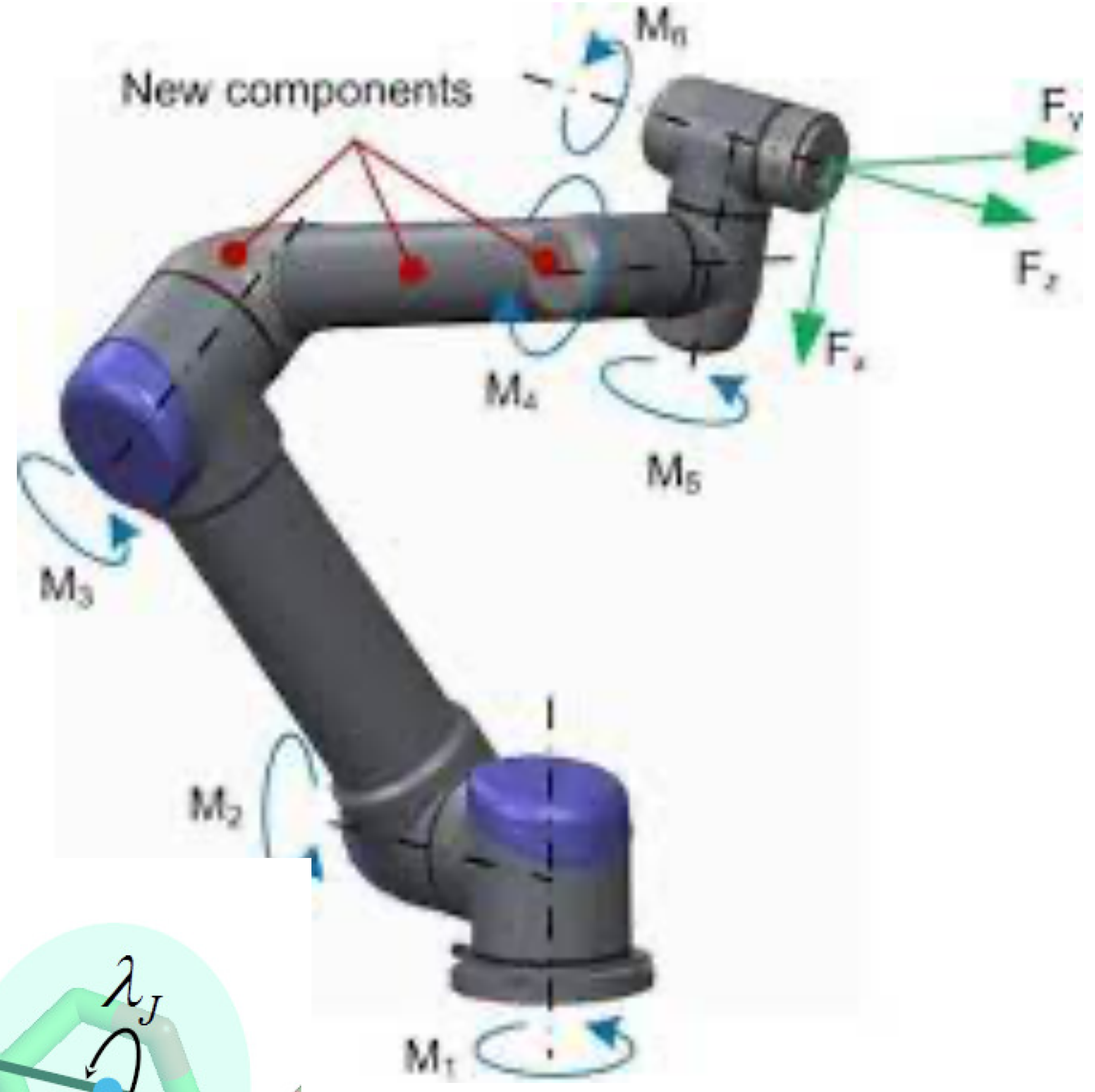
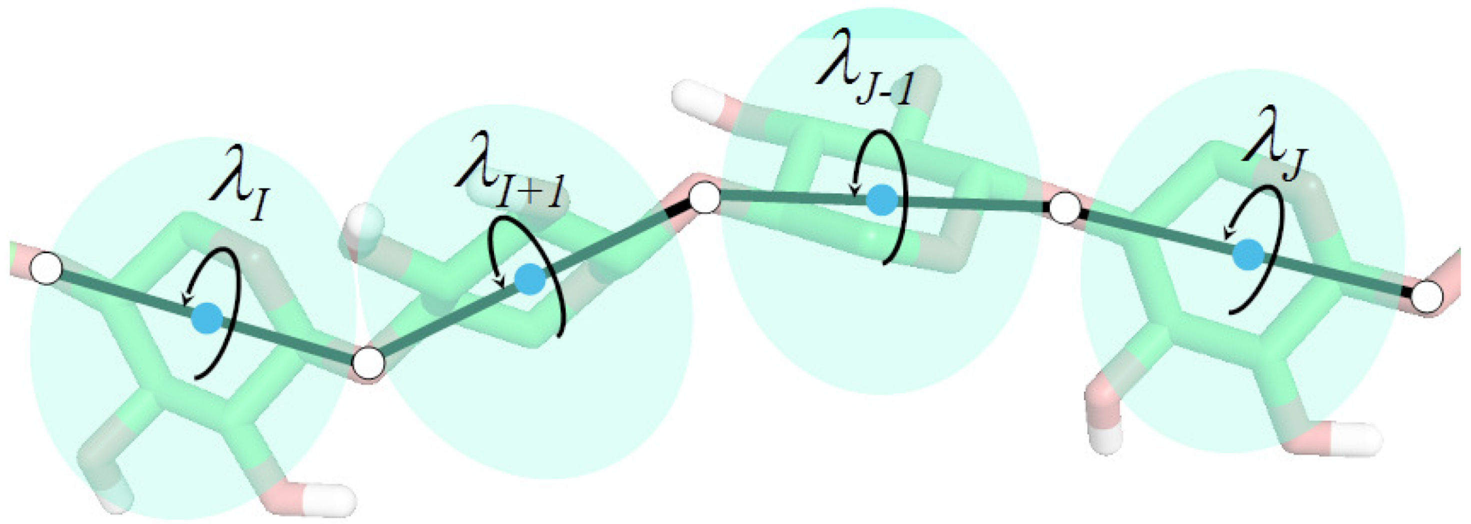


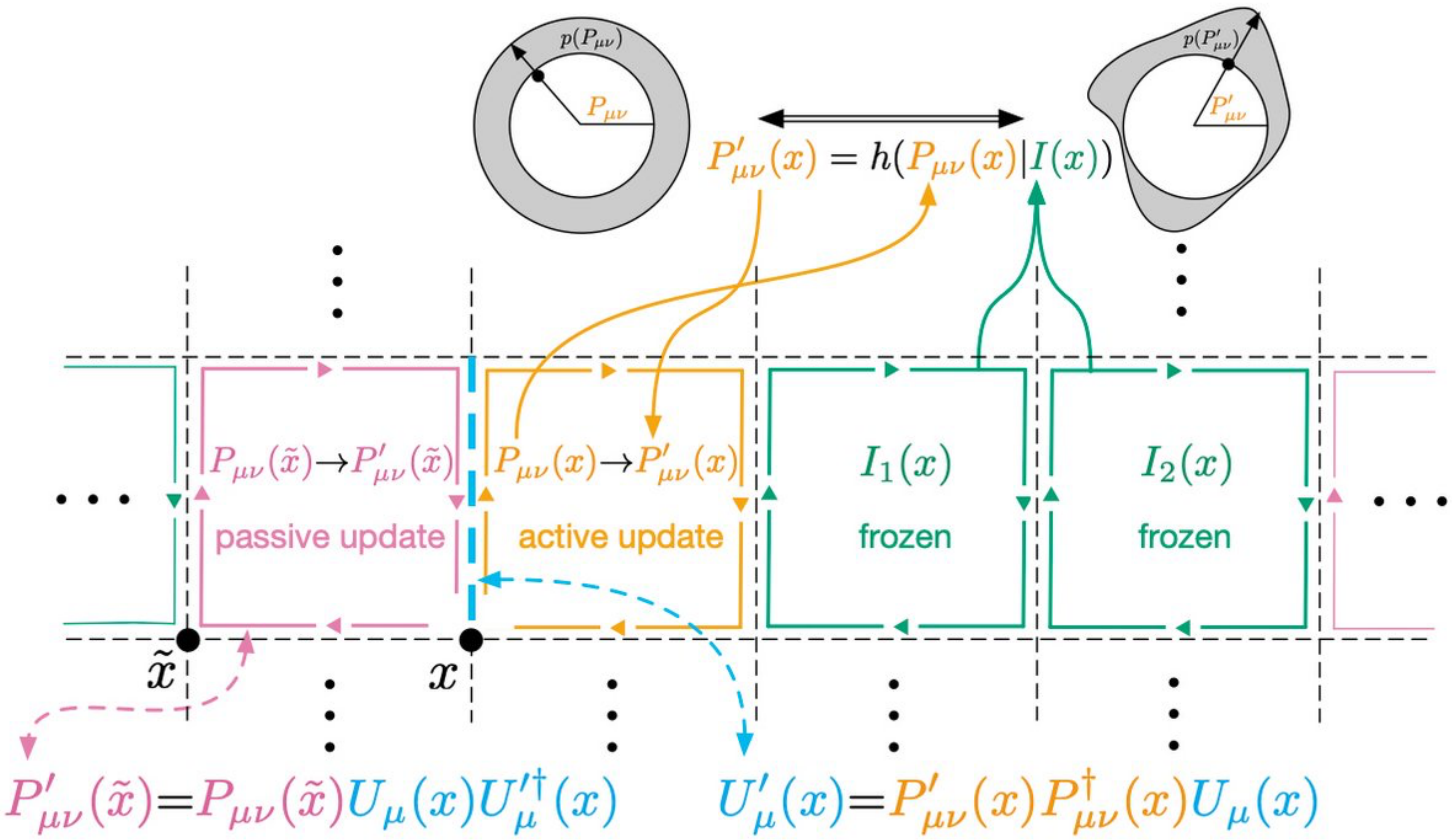
Figure 5. Learned multi-modal density on $SU(2) \equiv \mathbb{S}^3$ using the recursive flow. Each column shows an $\mathbb{S}^2$ slice of the $\mathbb{S}^3$ density





[arXiv:2003.06413]

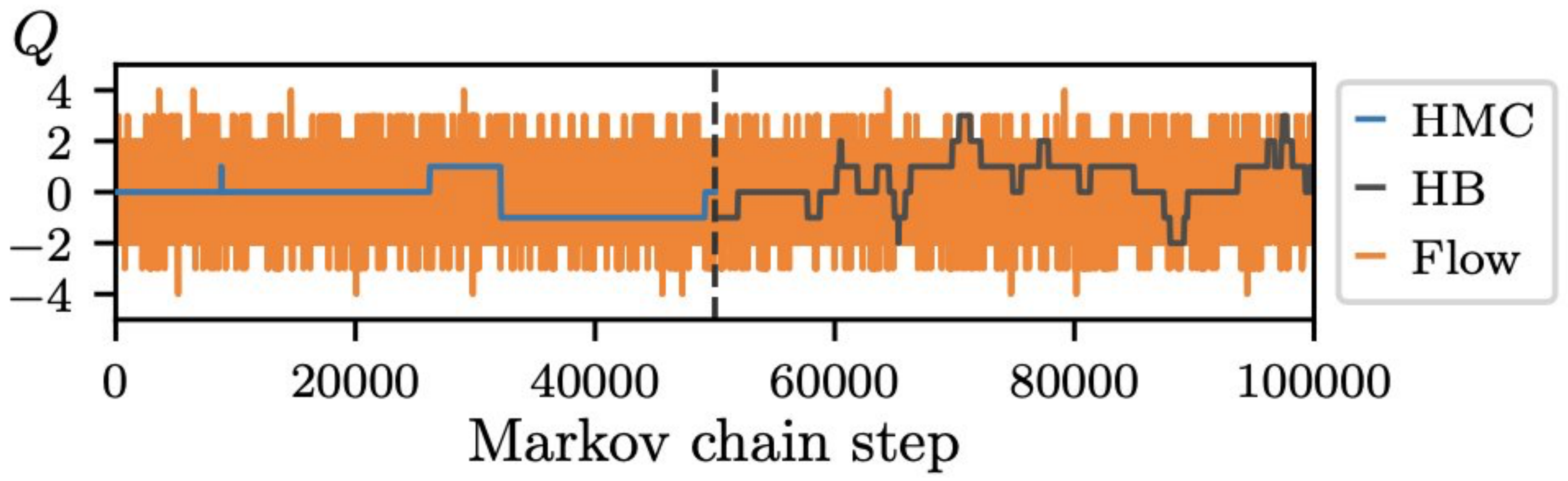
We came up with a way to build flows that are equivariant to space-time translations and local gauge transformations



Profit

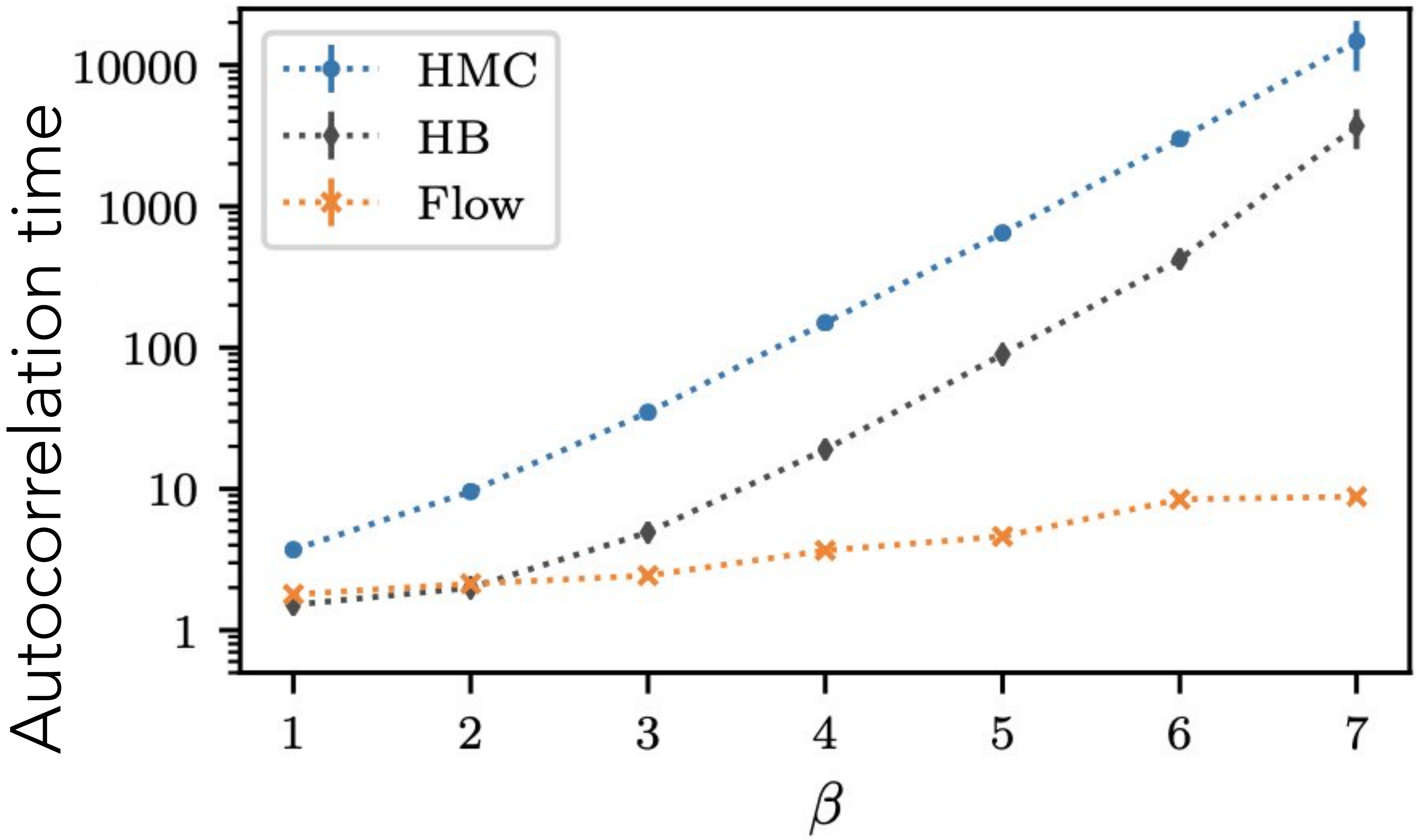
Essentially, MCMC can get stuck for a while in a certain mode.

- Our new "flow-based" proposal does much better!
- It learns to propose configurations that look like our target distribution.
- 1000x reduction in autocorrelation time

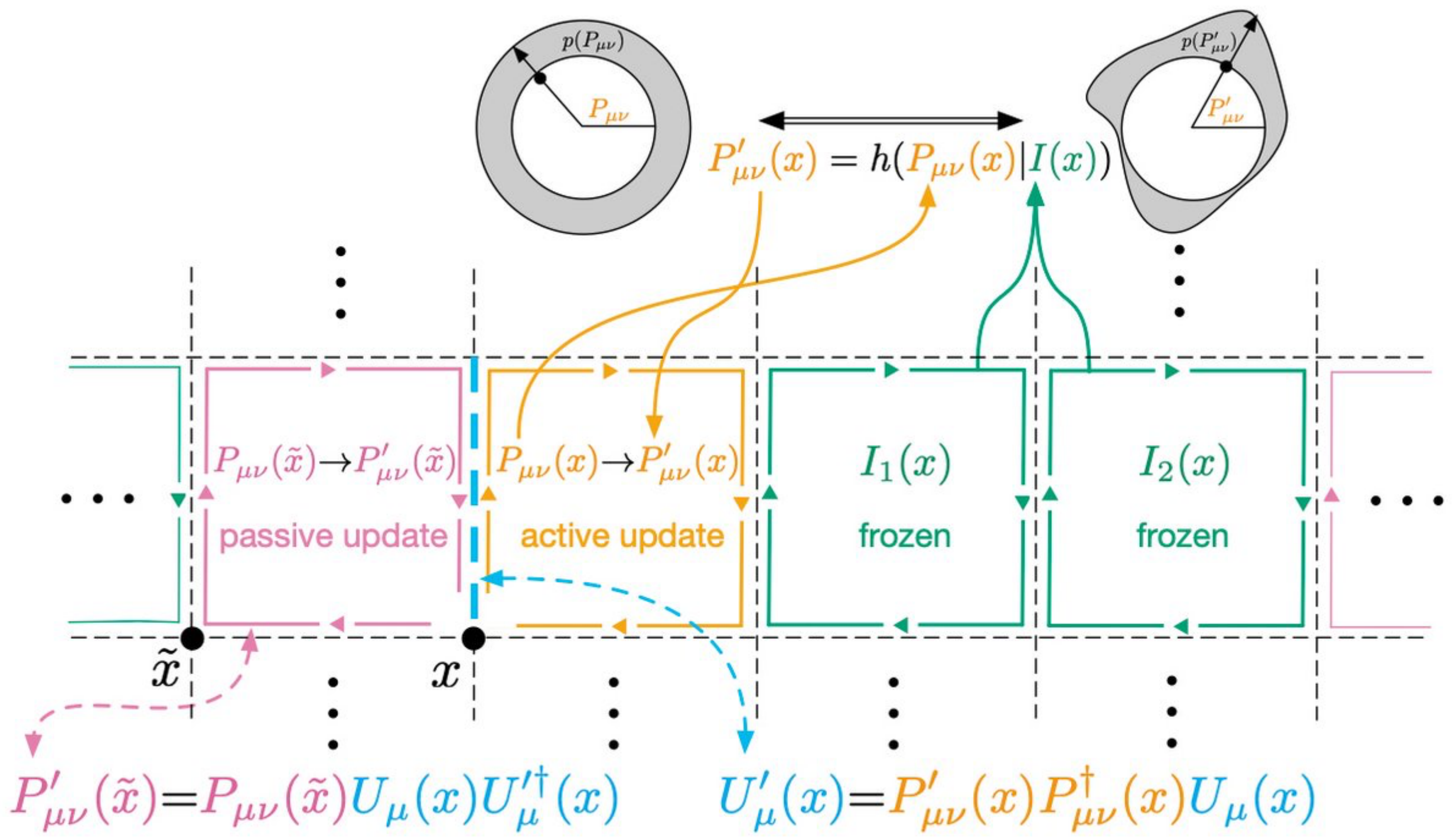


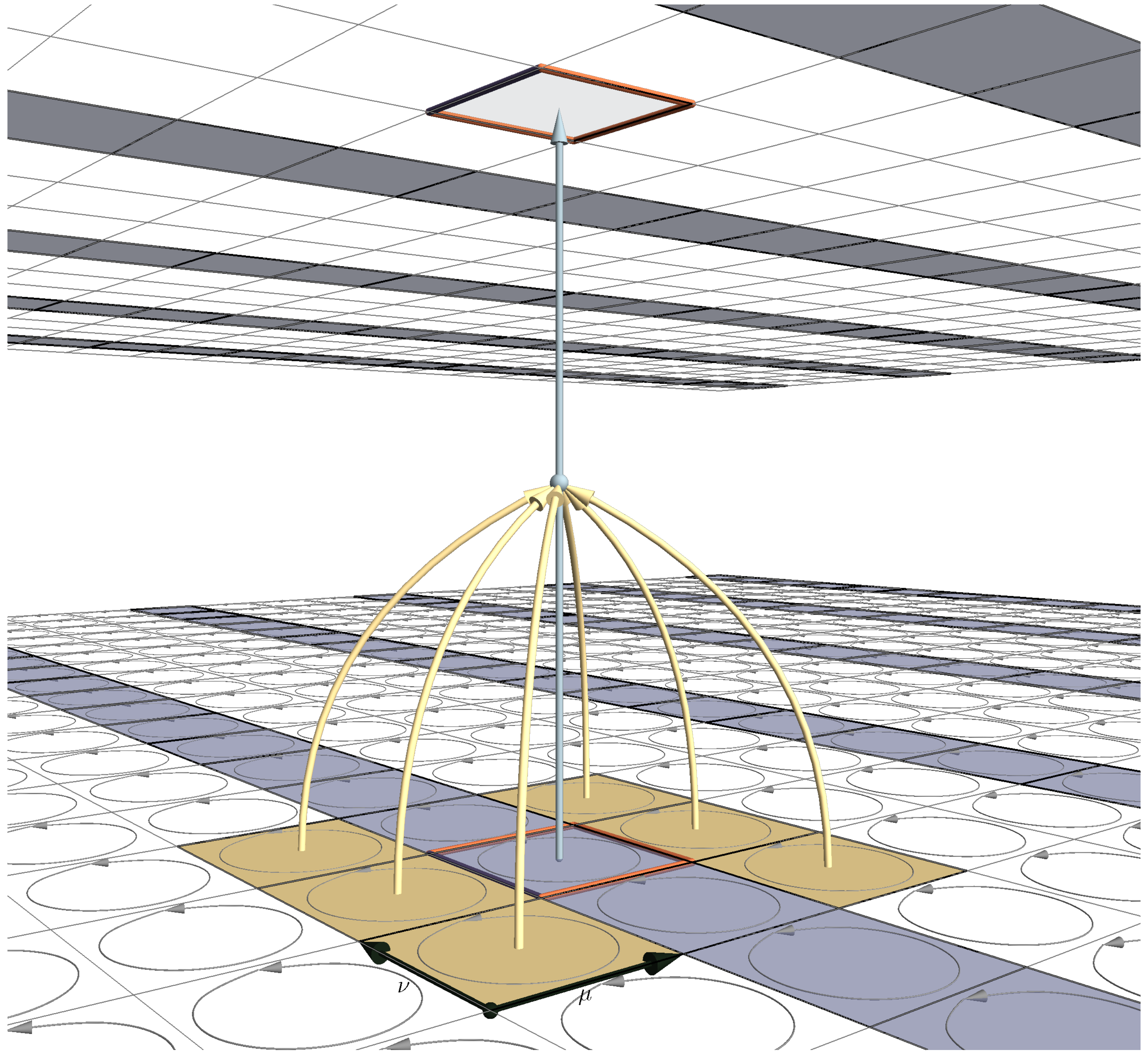
The topological charge Q will be constant for thousands of MCMC steps.

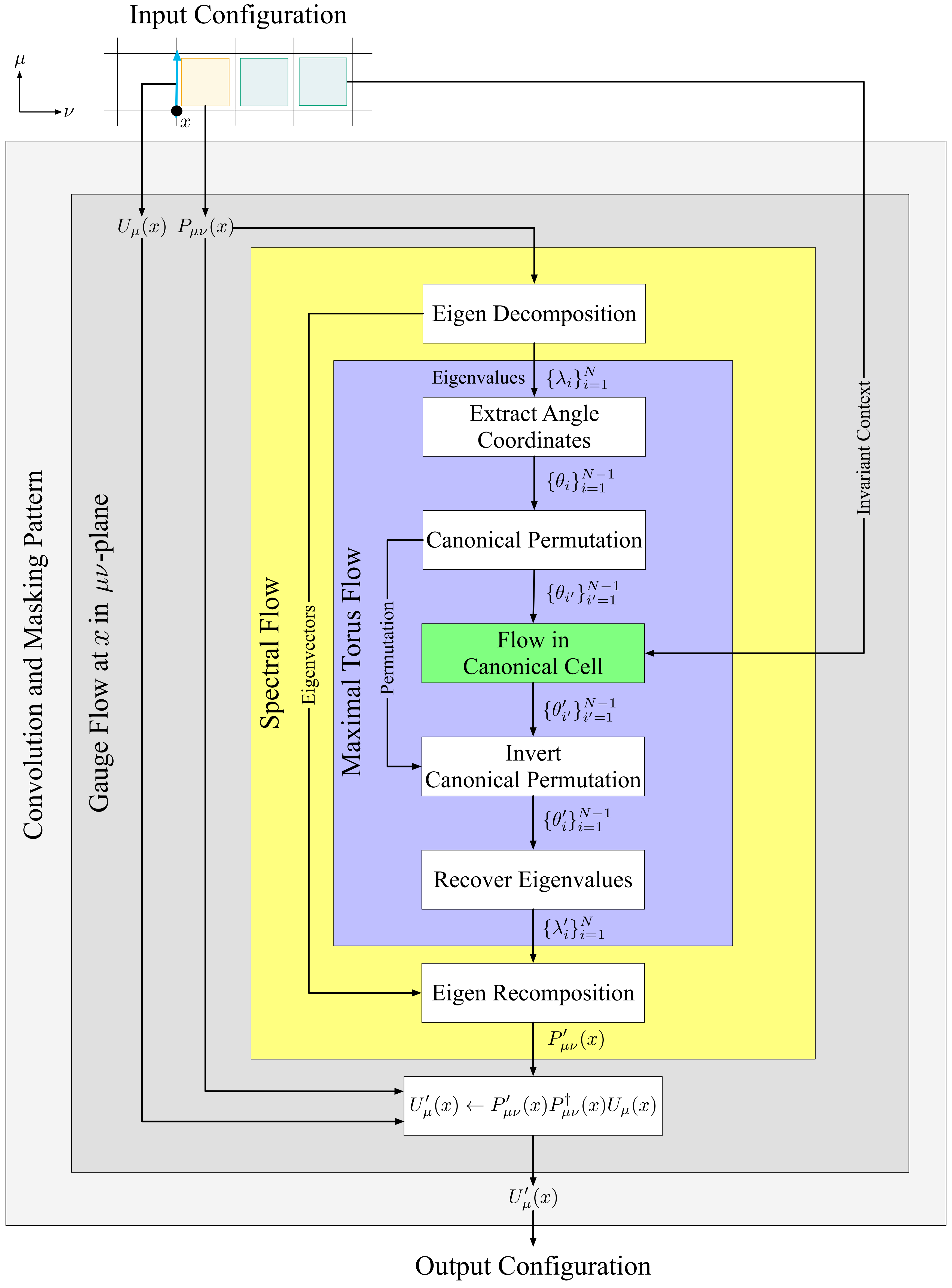
2D U(1) model

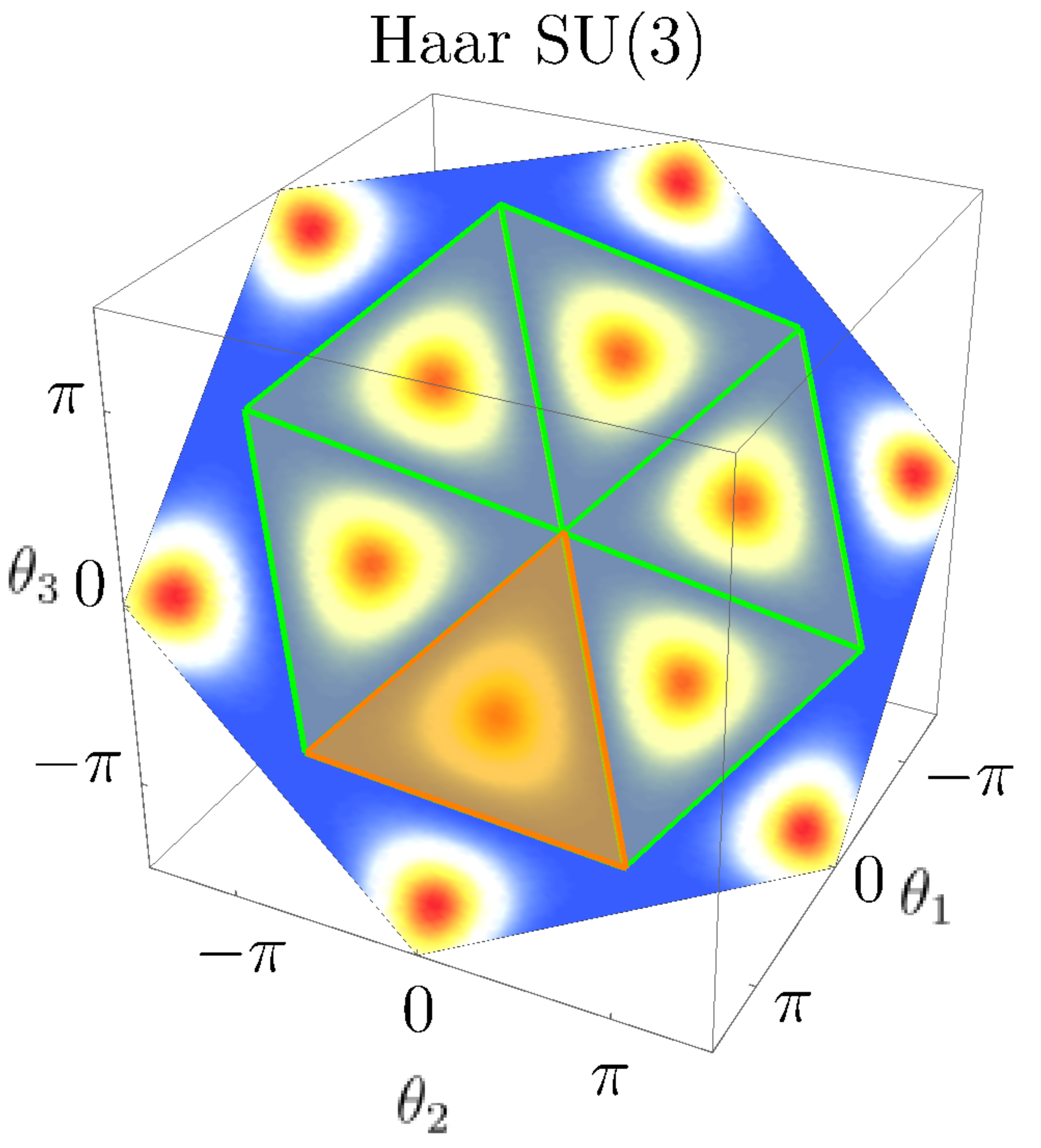


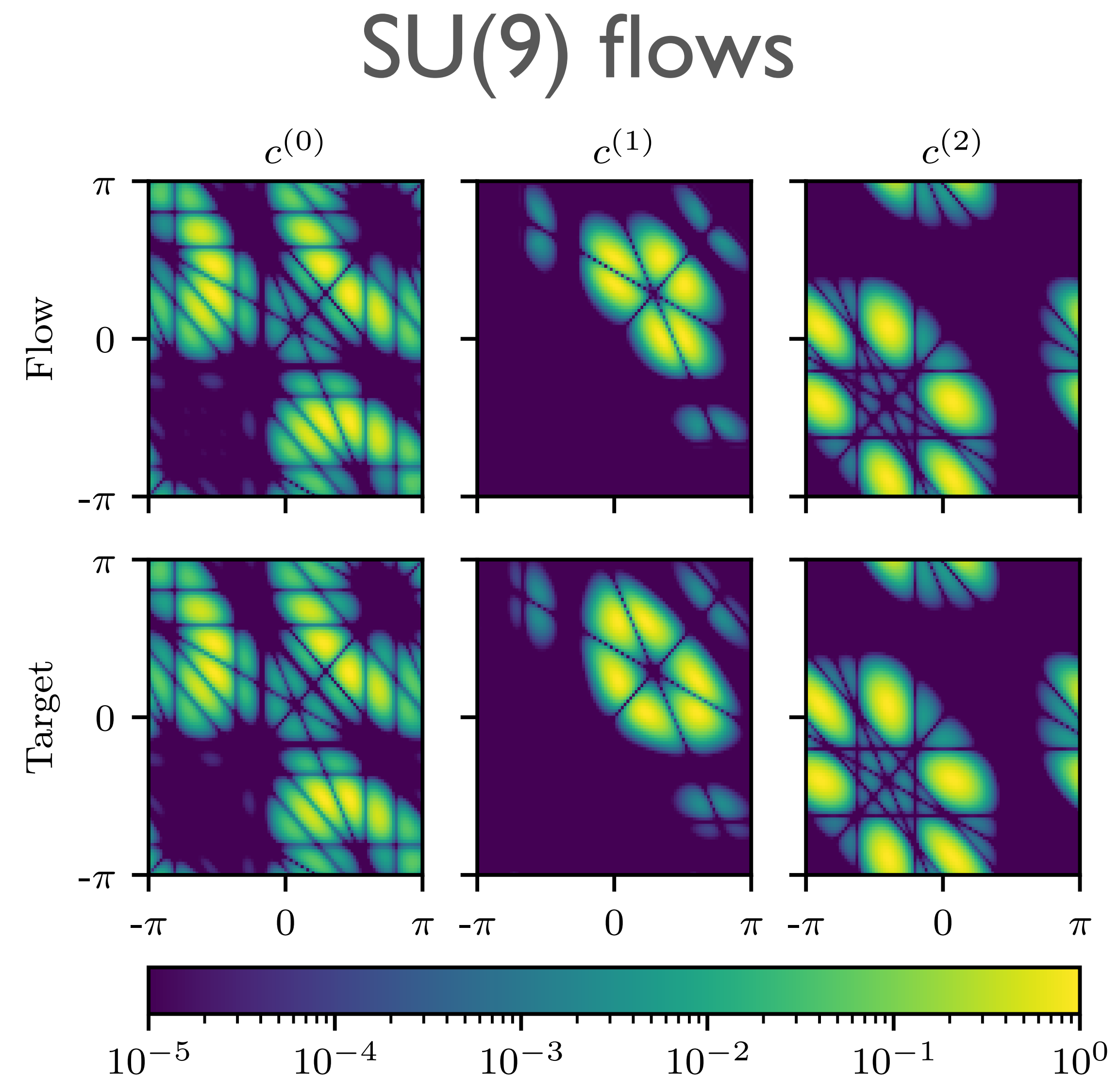
Space-time & Local, Non-Abelian Gauge Symmetry



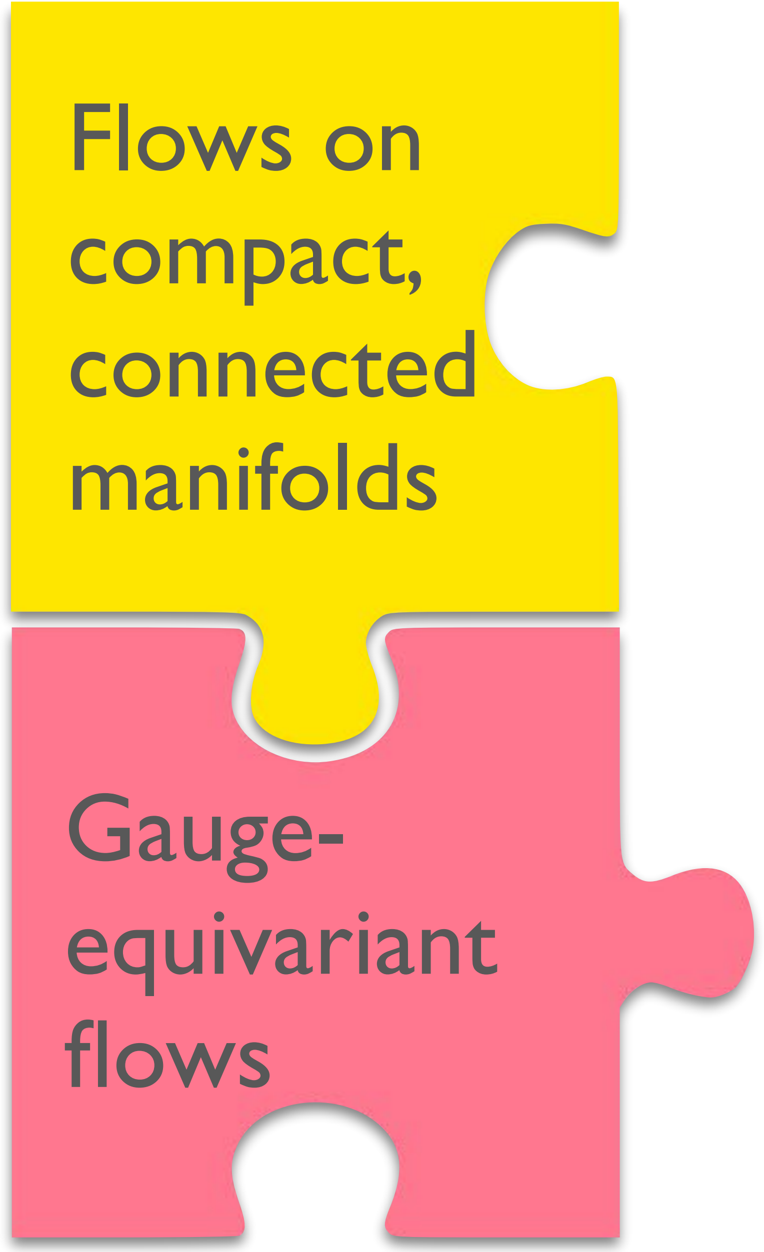






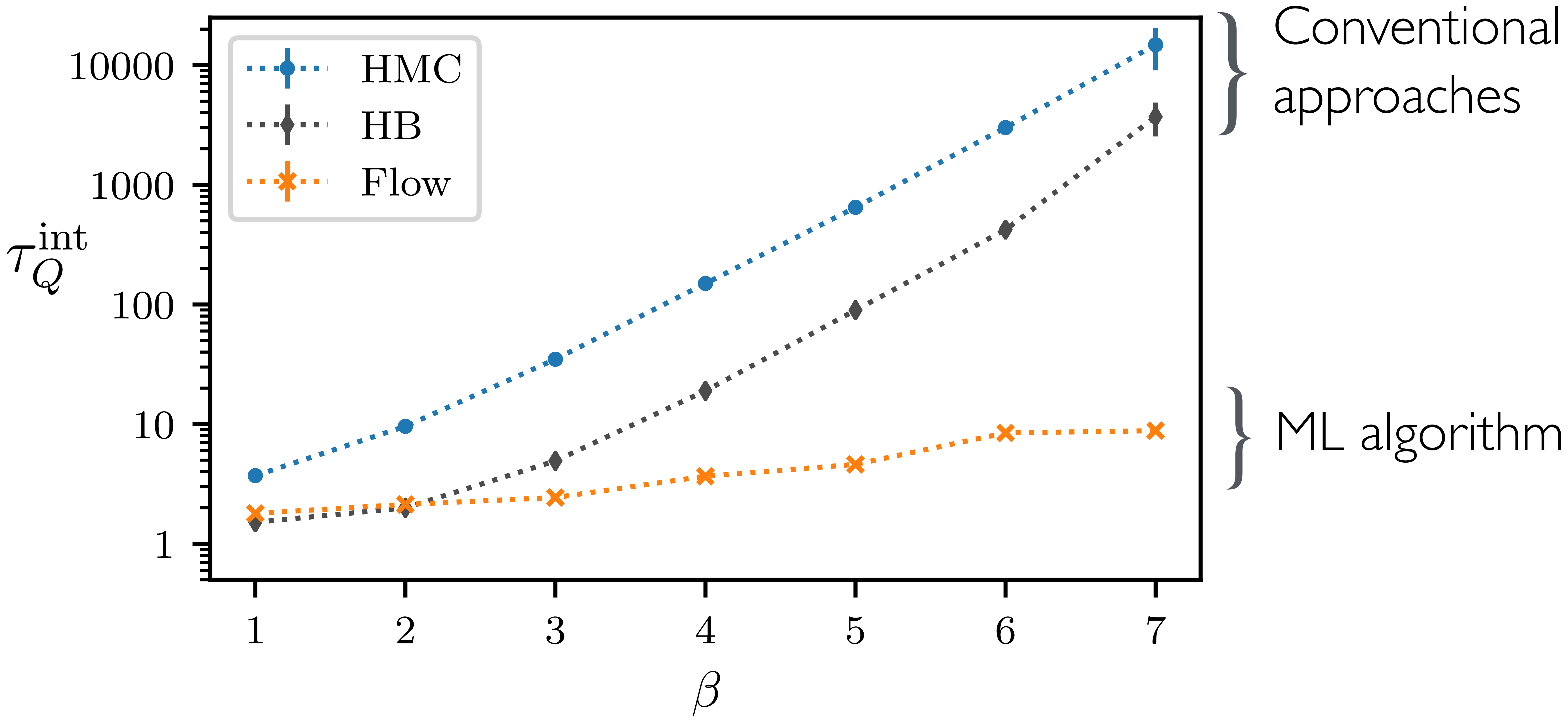


Flow models for QCD



First gauge theory application:
2D U(1) field theory

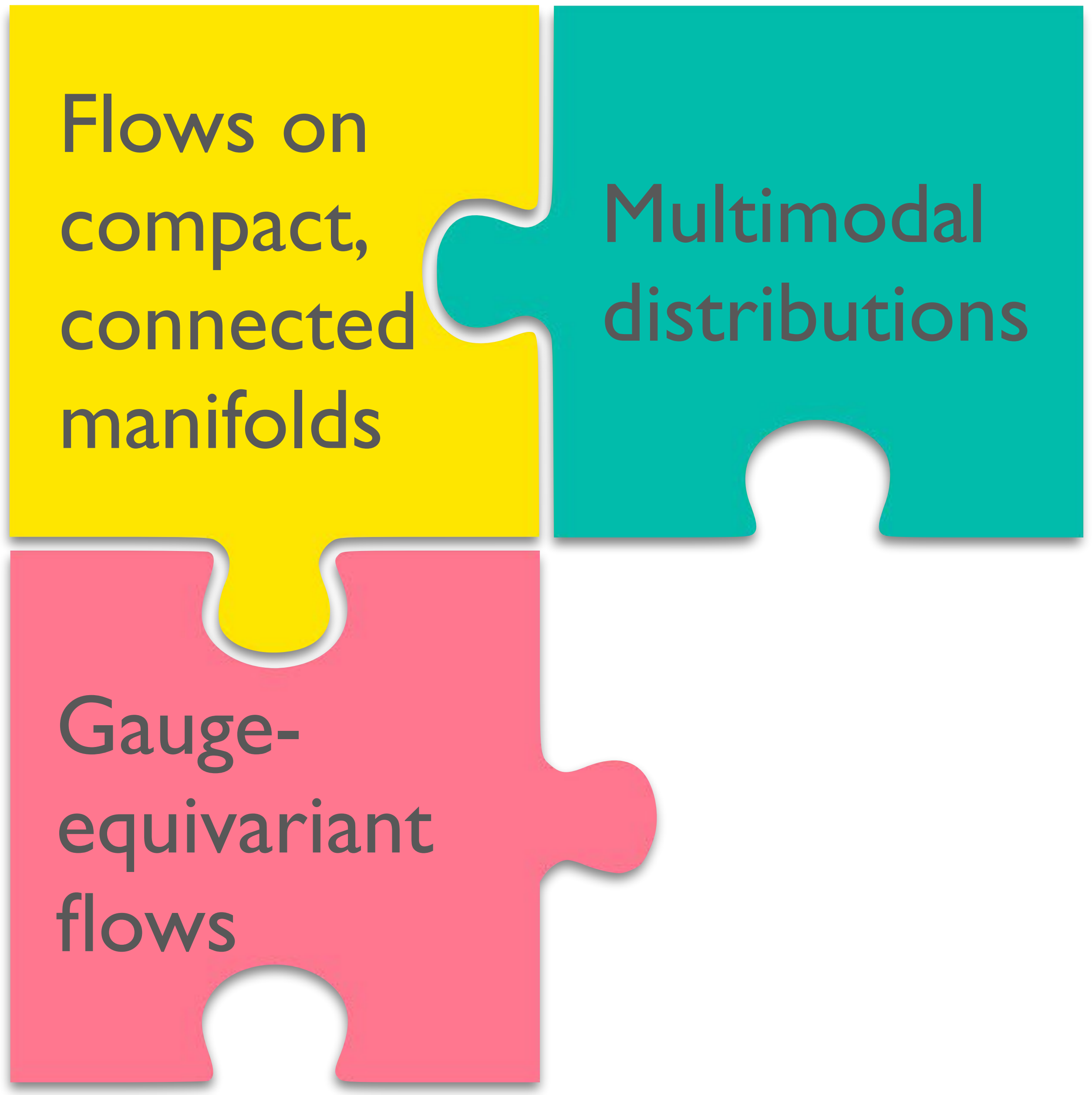
Cost per independent sample



Parameter of theory

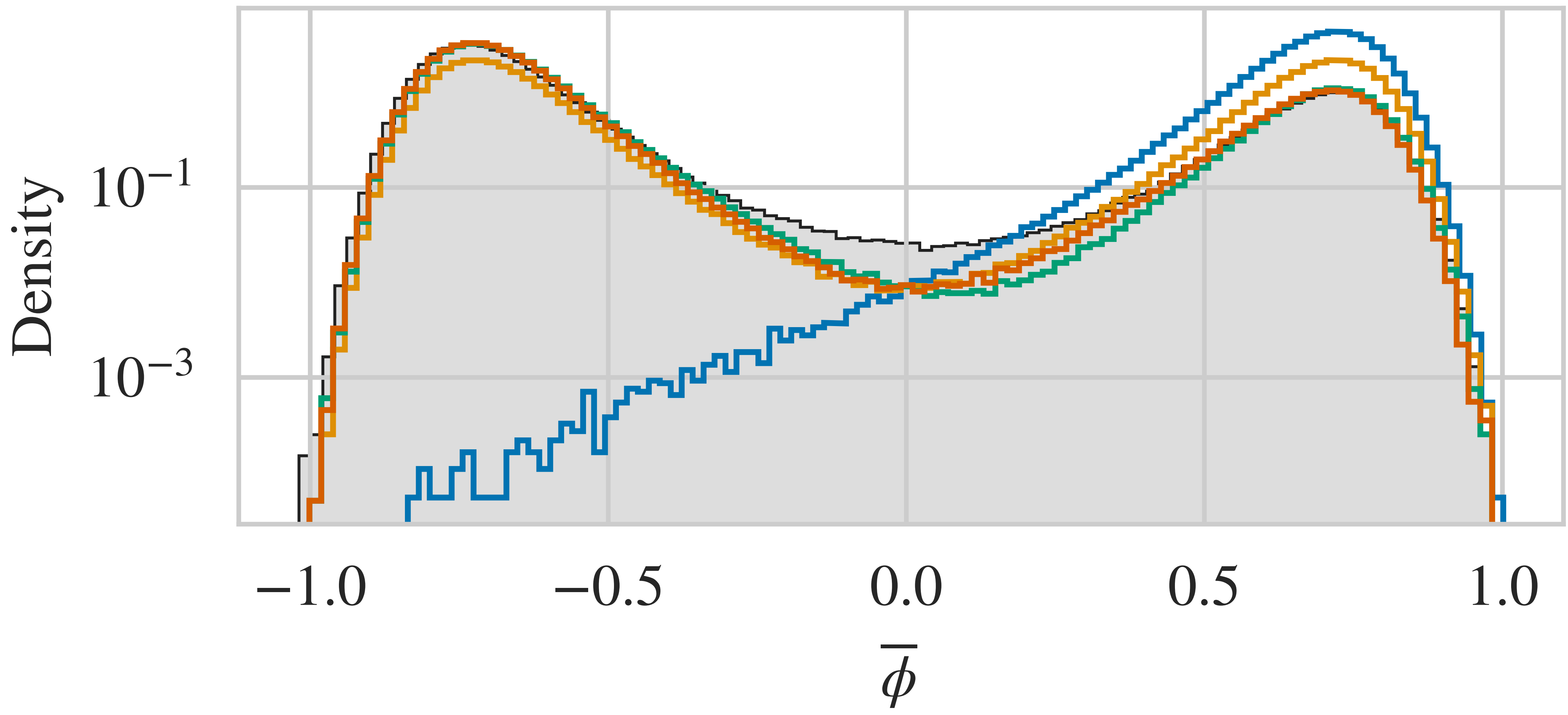
[*Phys.Rev.Lett.* 125, 121601 (2020)]

Flow models for QCD



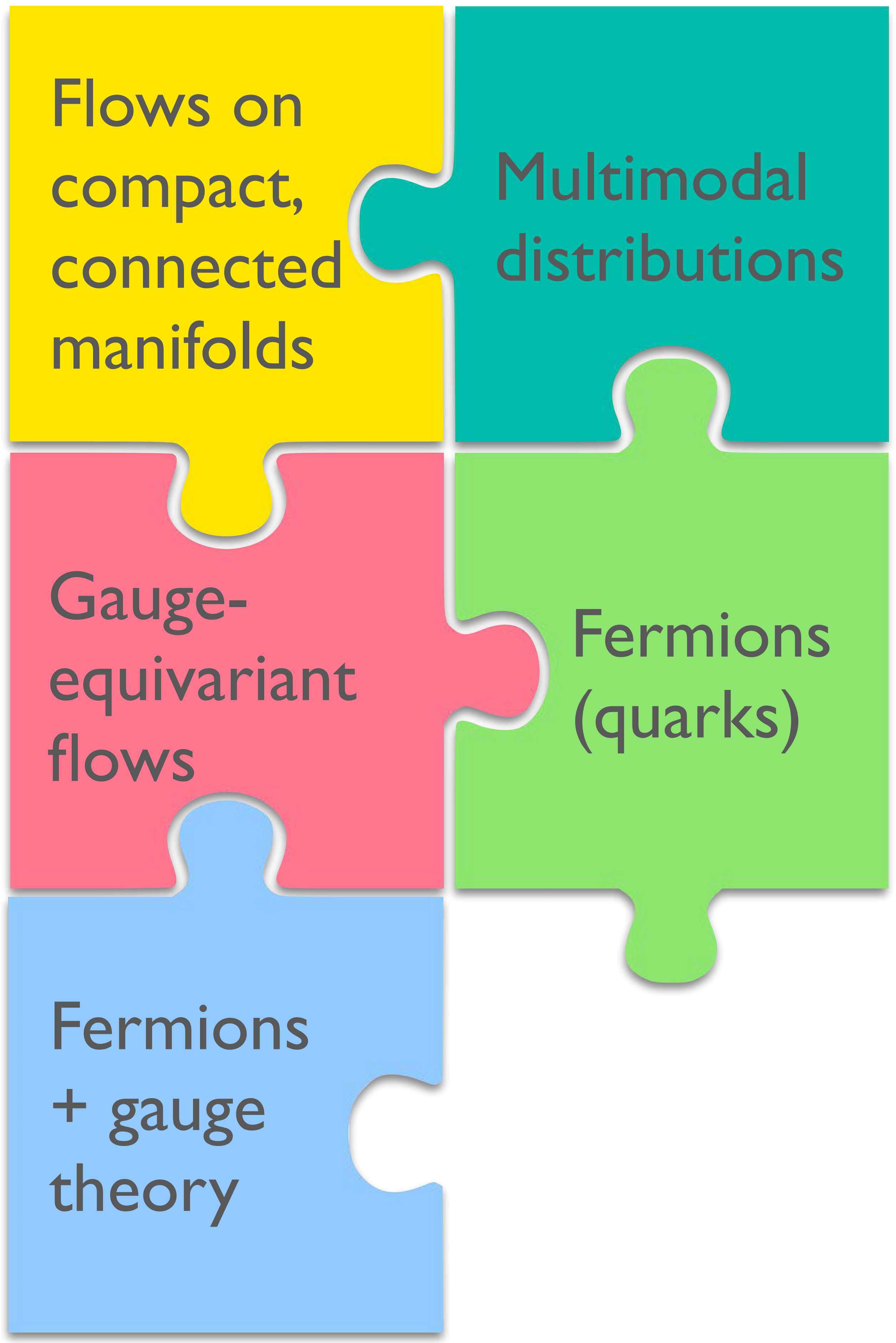
Systems with complex topologies

Need: Unbiased sampling from multi-modal distributions



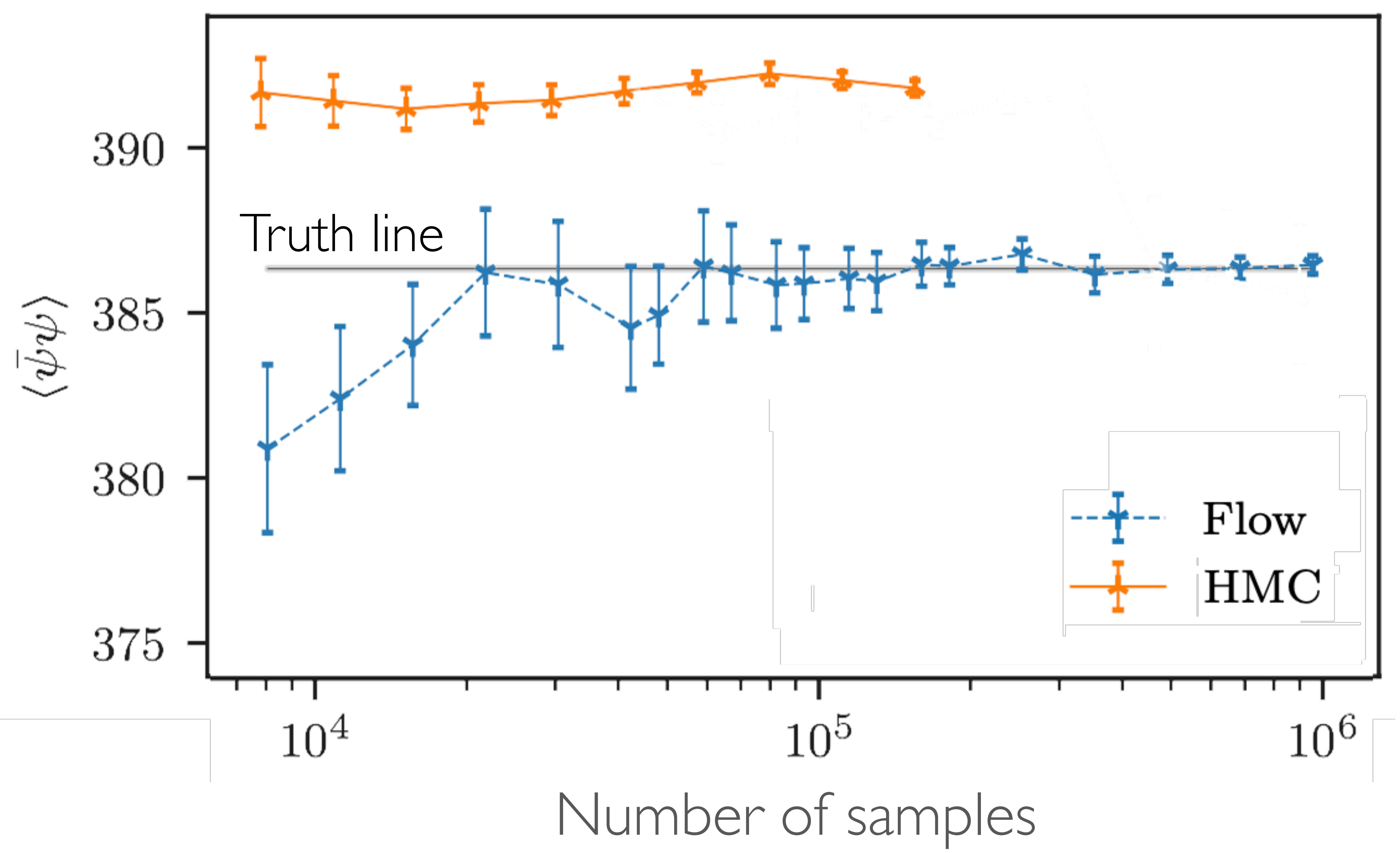
[2107.00734 (2021)]

Flow models for QCD



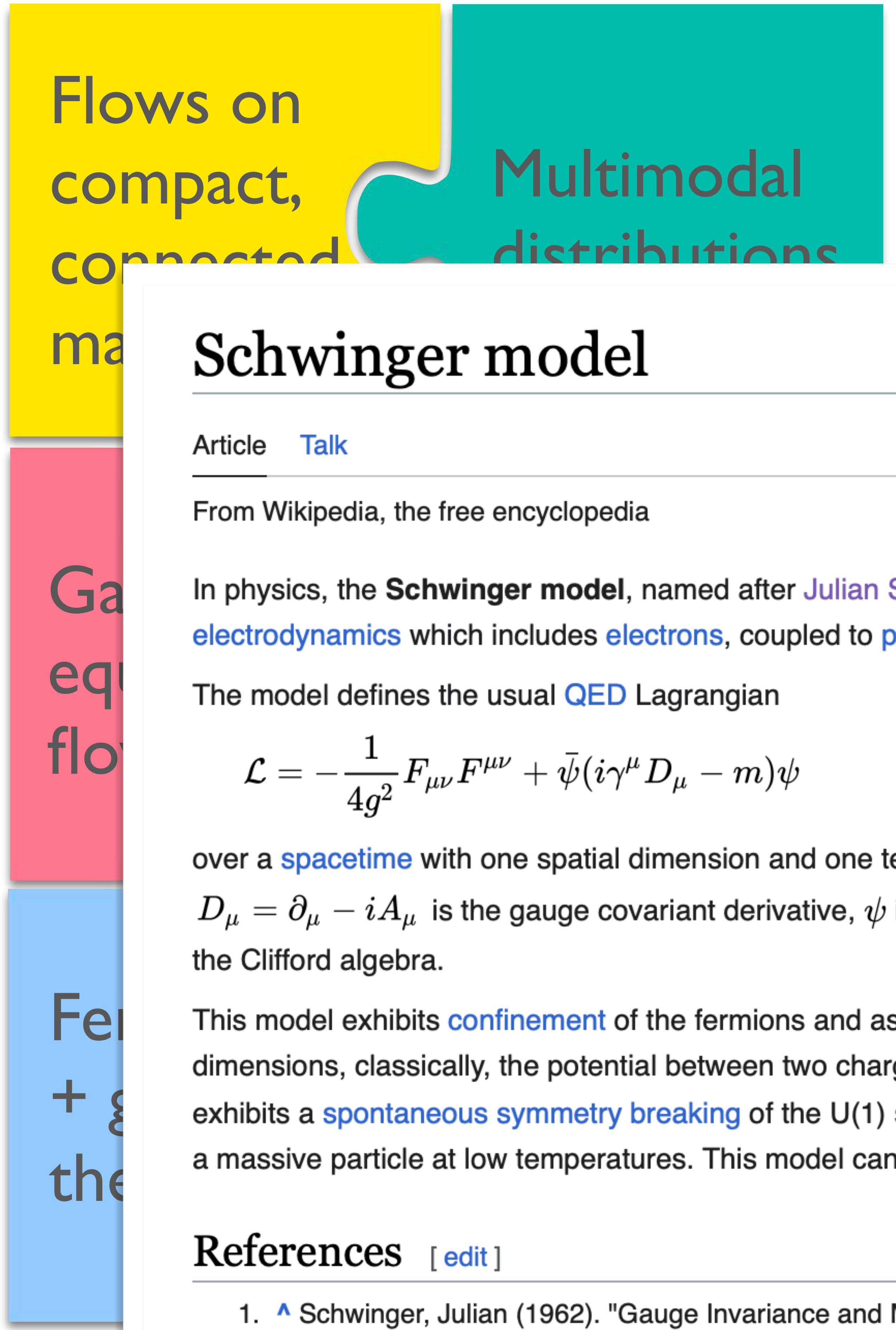
First gauge + fermion theory application:
2D Schwinger model

Measured value of observable



[*Phys.Rev.D* 104 (2021), 114507, arXiv:2202:11712]

Flow models for QCD



First gauge + fermion theory application:

2D Schwinger model

Schwinger model

3 languages

Article Talk Read Edit View history

From Wikipedia, the free encyclopedia

In physics, the **Schwinger model**, named after Julian Schwinger, is the model[1] describing 1+1D (1 spatial dimension + time) *Lorentzian* quantum electrodynamics which includes electrons, coupled to photons.

The model defines the usual QED Lagrangian

$$\mathcal{L} = -\frac{1}{4g^2}F_{\mu\nu}F^{\mu\nu} + \bar{\psi}(i\gamma^\mu D_\mu - m)\psi$$

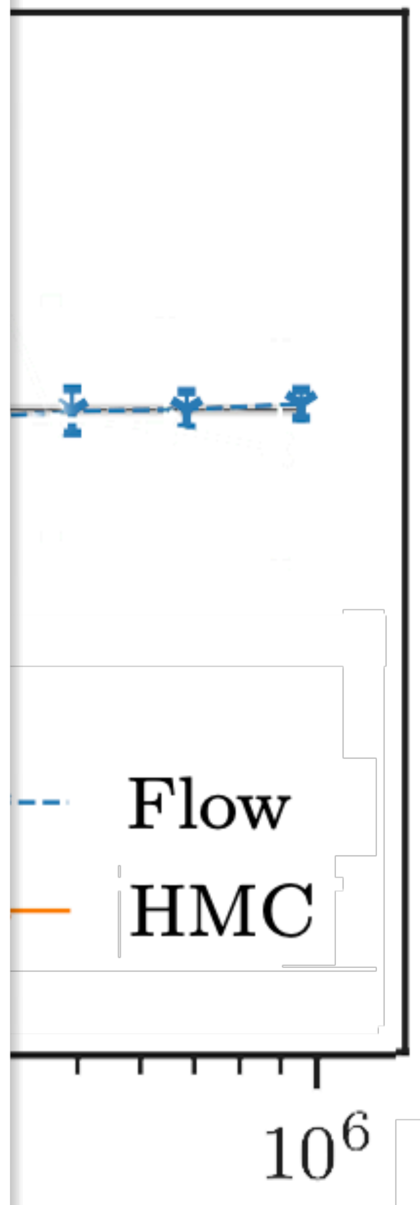
over a spacetime with one spatial dimension and one temporal dimension. Where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the $U(1)$ photon field strength, $D_\mu = \partial_\mu - iA_\mu$ is the gauge covariant derivative, ψ is the fermion spinor, m is the fermion mass and γ^0, γ^1 form the two-dimensional representation of the Clifford algebra.

This model exhibits confinement of the fermions and as such, is a toy model for QCD. A handwaving argument why this is so is because in two dimensions, classically, the potential between two charged particles goes linearly as r, instead of $1/r$ in 4 dimensions, 3 spatial, 1 time. This model also exhibits a spontaneous symmetry breaking of the U(1) symmetry due to a chiral condensate due to a pool of instantons. The photon in this model becomes a massive particle at low temperatures. This model can be solved exactly and is used as a toy model for other more complex theories.[2][3]

References [edit]

1. ^ Schwinger, Julian (1962). "Gauge Invariance and Mass. II". *Physical Review*. Physical Review, Volume 128. **128** (5): 2425–2429. Bibcode:1962PhRv..128.2425S. doi:10.1103/PhysRev.128.2425.
2. ^ Schwinger, Julian (1951). "The Theory of Quantized Fields I". *Physical Review*. Physical Review, Volume 82. **82** (6): 914–927. Bibcode:1951PhRv...82..914S. doi:10.1103/PhysRev.82.914. S2CID 121971249.
3. ^ Schwinger, Julian (1953). "The Theory of Quantized Fields II". *Physical Review*. Physical Review, Volume 91. **91** (3): 713–728. Bibcode:1953PhRv...91..713S. doi:10.1103/PhysRev.91.713.

able

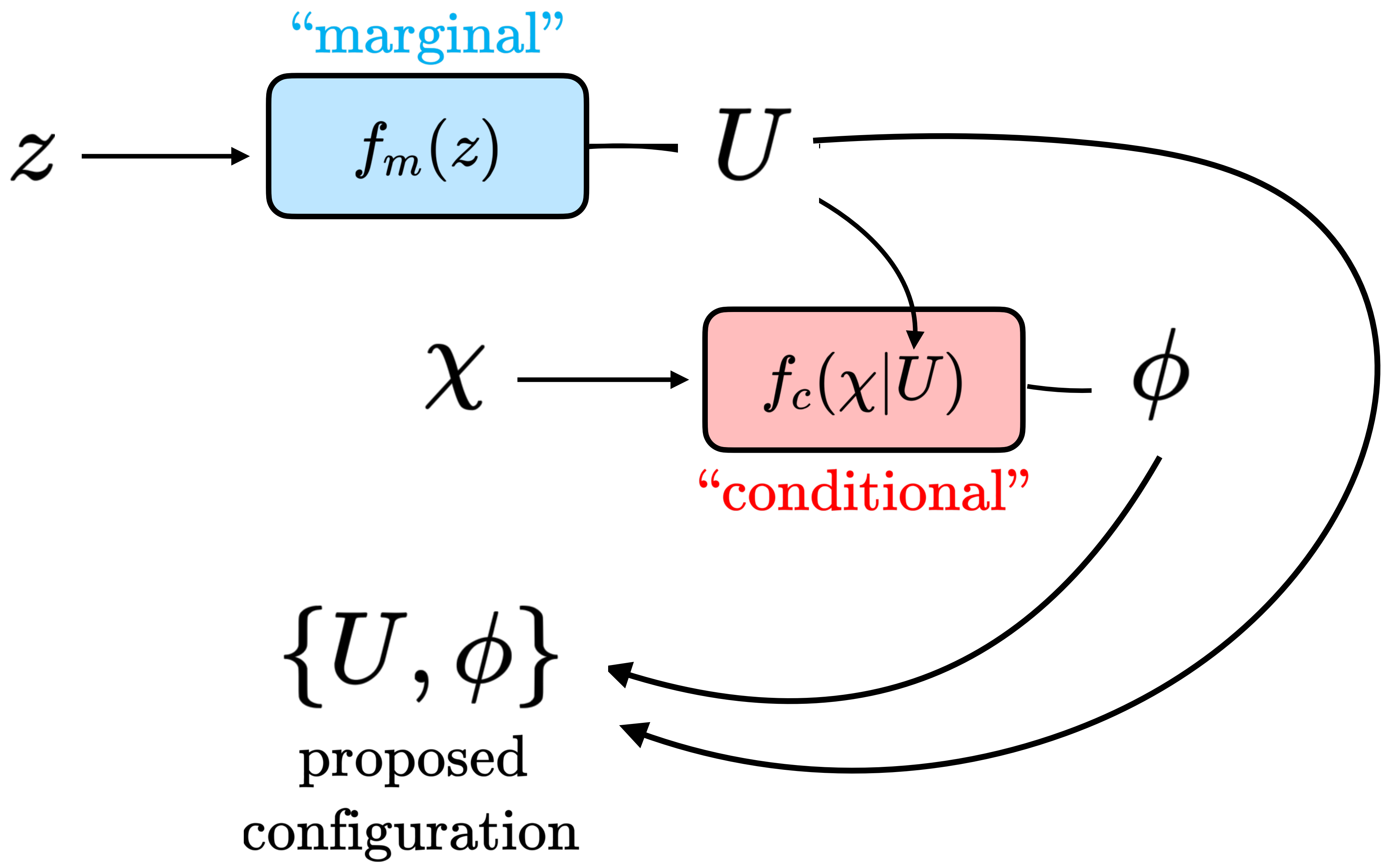


, arXiv:2202:11712]

Flow models for QCD in 4D

Initial QCD demonstration [this talk +upcoming manuscripts on scaling and 4D]

- Direct combination of published results on gauge-equivariant flows and pseudofermions [Boyda et al., 2008.05456, Abbott et al., 2207.08945]
- Illustration at straightforward parameters $V=4^4$, $N_f=2$, $\beta=1$, $\kappa=0.1$
- Observables from flow ensemble in precise agreement with HMC at high statistics (65k samples)
- Development and scaling of QCD-specific architectures in full swing — stay tuned!



Marginal:

- Haar-uniform base distribution
- 48 gauge-equivariant spline coupling layers
- Spatially separated convolutions in spectral flow to define spline parameters

Conditional:

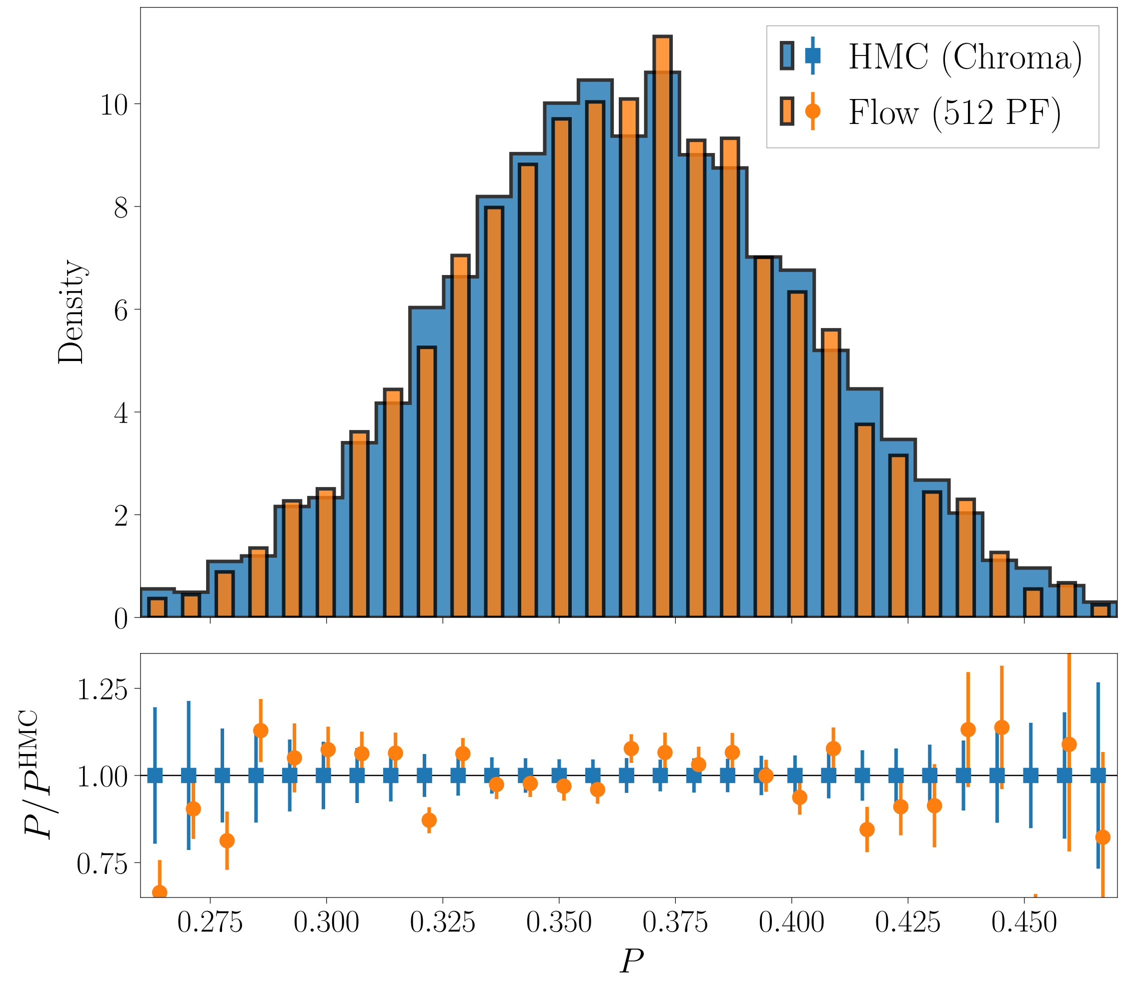
- Gaussian base distribution
- 36 pseudofermion coupling layers built from parallel transport convolutional networks
- Alternating spin and spatial masking pattern

Flow models for QCD in 4D

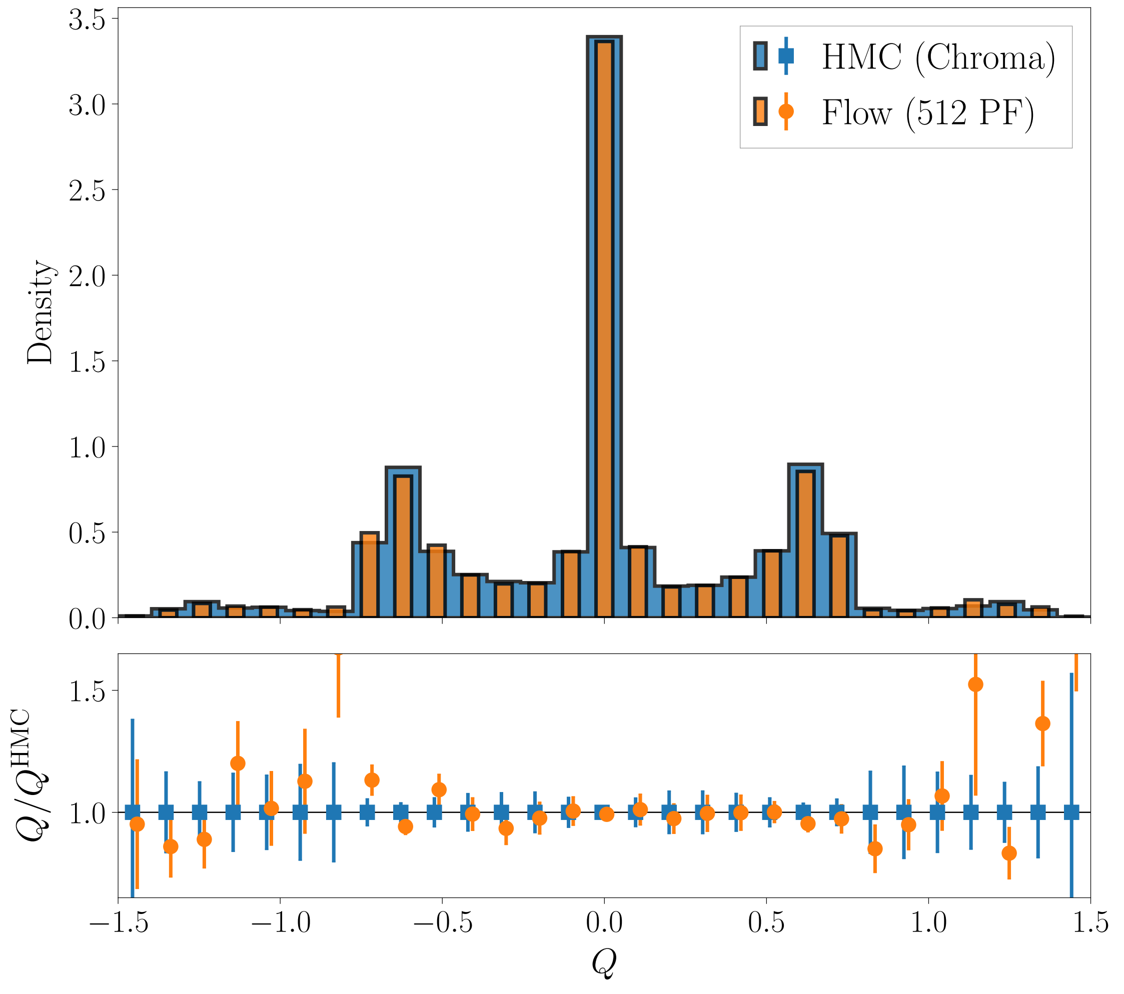
Initial QCD demonstration [this talk +upcoming manuscripts on scaling and 4D]

- Direct combination of published results on gauge-equivariant flows and pseudofermions [Boyda et al., 2008.05456, Abbott et al., 2207.08945]
- Illustration at straightforward parameters $V=4^4$, $N_f=2$, $\beta=1$, $\kappa=0.1$
- Observables from flow ensemble in precise agreement with HMC at high statistics (65k samples)
- Development and scaling of QCD-specific architectures in full swing — stay tuned!

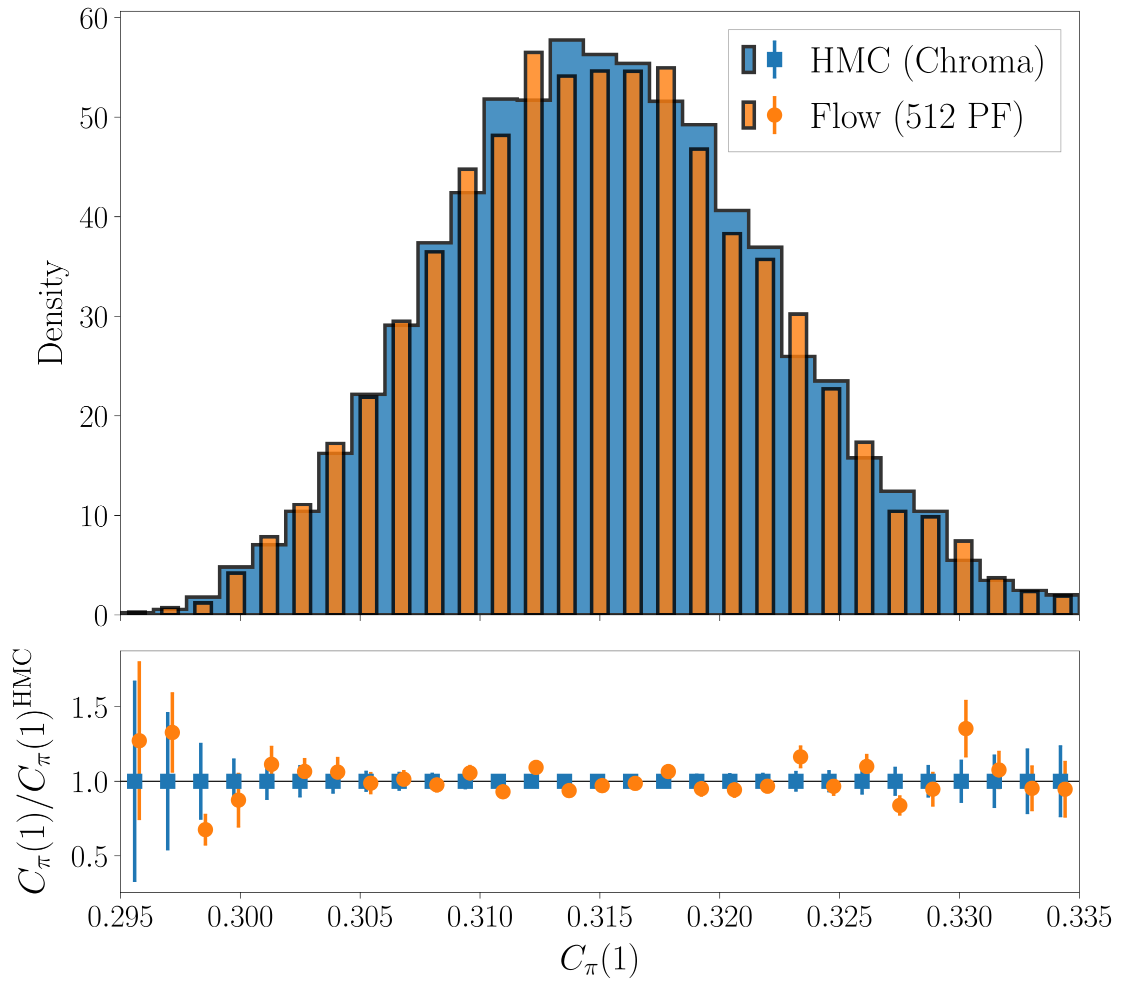
Plaquette



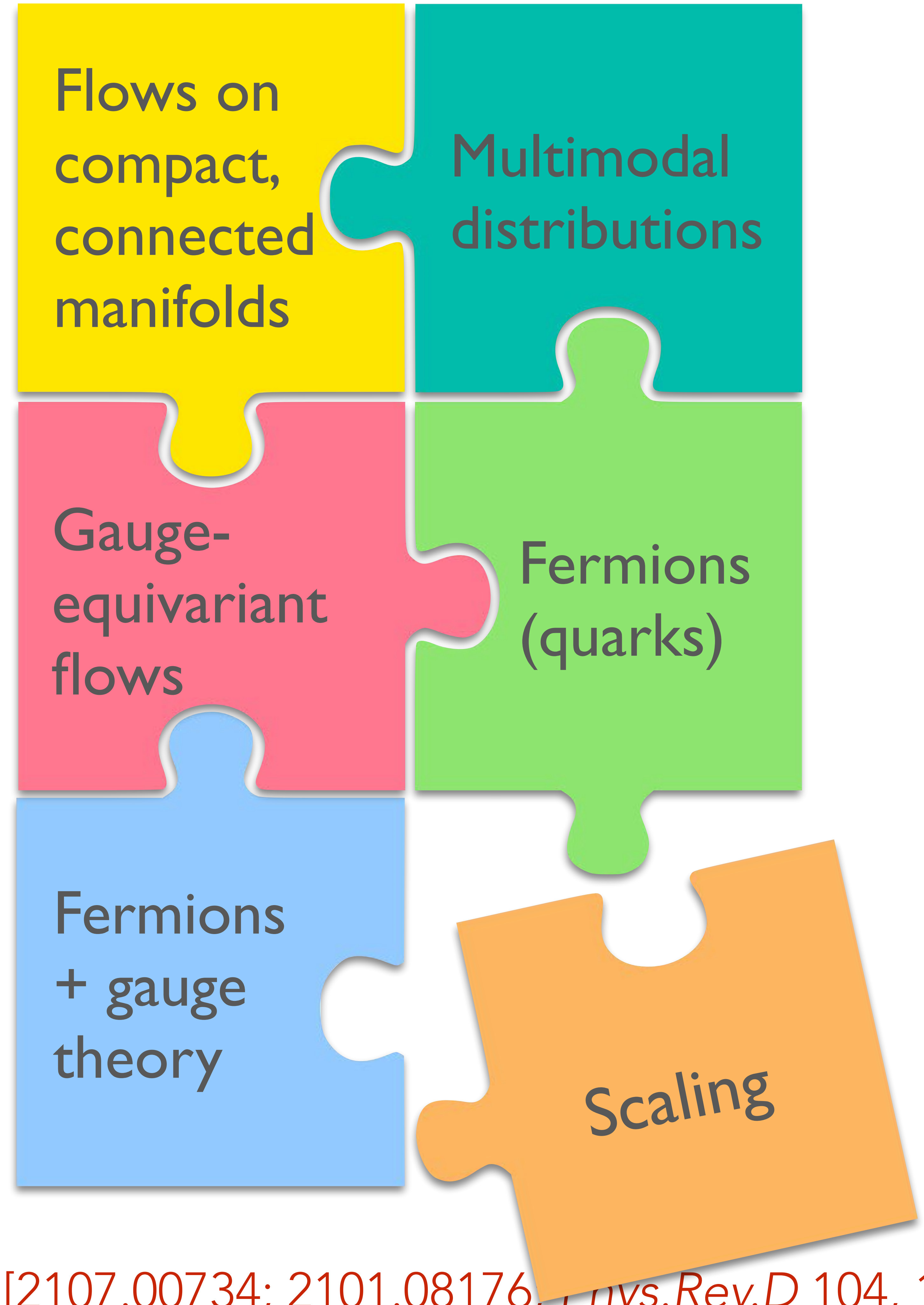
Topological charge (Wilson flowed)



Pion correlation function (t=1)



Flow models for QCD



Machine learning for QCD

- **Provably-exact** machine-learning-accelerated sampling algorithm
- Orders of magnitude more **efficient** than conventional algorithms overcoming critical slowing-down
- **Unbiased** results where traditional approaches fail

Deployment for state-of-the-art QCD scheduled for Aurora 2023 first science time

[2107.00734; 2101.08176; *Phys.Rev.D* 104, 114507; *Phys.Rev.D* 103, 074504 (2021); *Phys.Rev.Lett.* 125, 121601; PMLR 8083-8092 (2020); *Phys.Rev.D* 100, 034515 (2019); *Phys.Rev.D* 97, 094506 (2018)]

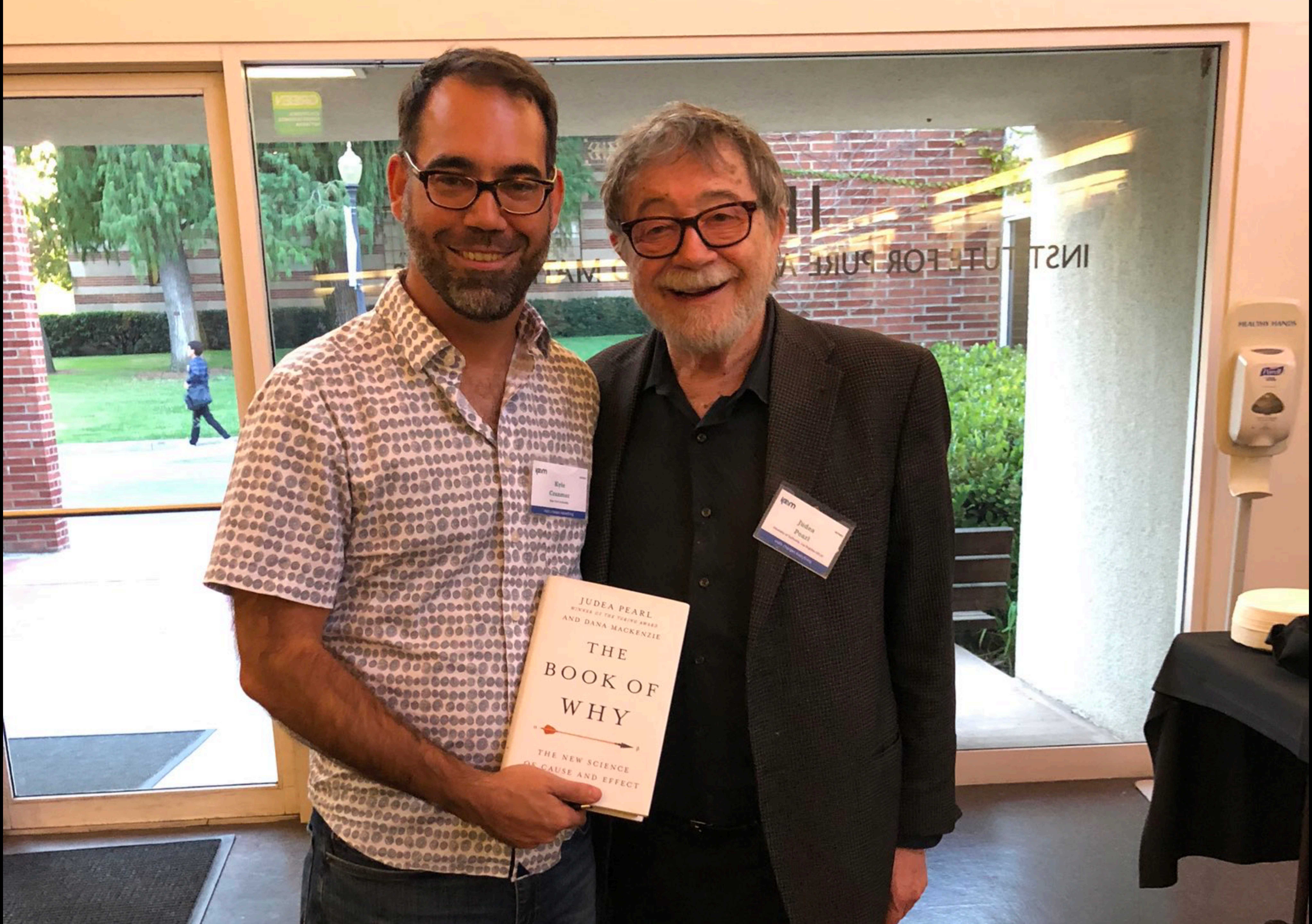
Inductive Bias

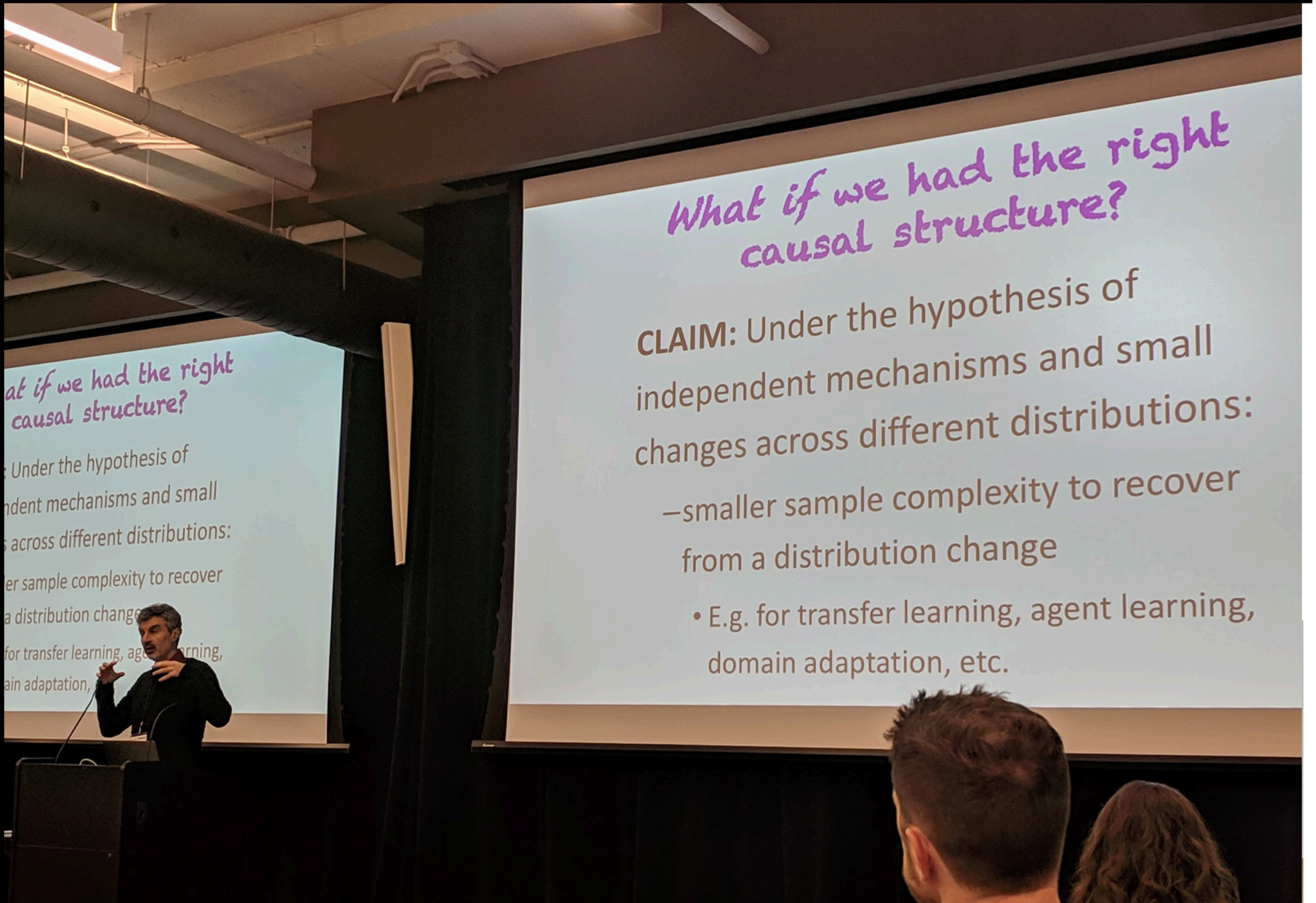
Compositionality

Symmetry

Relationships

Causality


JUDEA PEARL
WINNER OF THE TURING AWARD
AND DANA MACKENZIE
THE BOOK OF WHY
THE NEW SCIENCE OF CAUSE AND EFFECT



Max Welling Isn't this what Bernhard Schoelkopf has been saying for a while?

Like · Reply · 6w

Yann LeCun ...and Leon Bottou ?

Like · Reply · 6w

Leon Bottou Yoshua's paper says: if you observe a distribution change that comes from a causal effect, then you'll adapt faster if your generative model matches the causal model.

Another way of seeing it is : the right causal graph suggests a particular factorization of the joint distribution (a directed bayesian network). A causal intervention means that you only change one of these factors (or a few factors) while leaving the other ones unchanged. Therefore if your generative model is the right causal model, meaning that it factorizes the joint in the same way, it will be easy to adapt it to the change because only a few parameters need changing (those associated with the factors that actually changed).

Max Welling Dan Roy I am, and I think most of us, are keenly aware that Josh has been the big proponent of this view. And I think most people agree with him on this view. Integrating this view with deep learning for more narrowly defined tasks seems to me an interesting intellectual pursuit though. I think that's what's happening here but I was not at the talk 😉


The message from human cognition:
Richly structured models of objects and their relations are a powerful tool for reasoning about, and interacting with, the world.
Objects and relations reflect decisions made by evolution, experience, and task demands about how to represent the world in an efficient and useful way
Intelligence is about model-building, beyond just recognizing patterns (Tenenbaum)
Combinatorial generalization via abstraction and compositionality ("infinite use of finite means")
Peter Battaglia

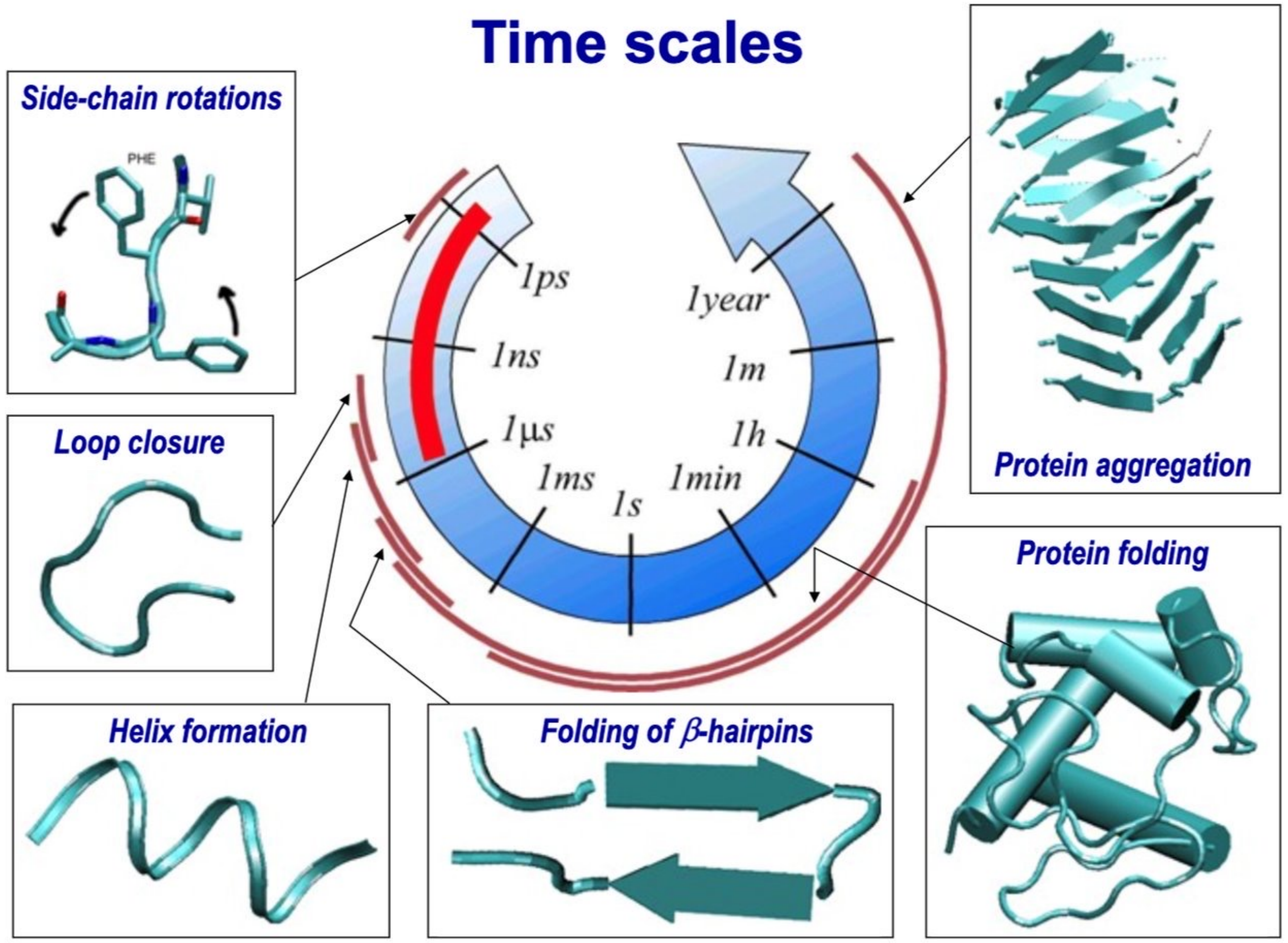


Figure from J.E. Shea

Emergence: Philosophical Musings

Scale separation & emergence

Scale separation can lead to different effective descriptions & ontologies that describe the phenomena that emerge at different scales

- Identifying and naming the relevant objects / concepts already significant
- Understanding how they interact and developing an effective law or theory at that scale is even more significant
- Understanding how these objects and interactions emerge from a more fundamental scale is profound

This has generally been done by humans, and there is an opportunity for AI to assist / accelerate / automate this process.

Scale separation & emergence

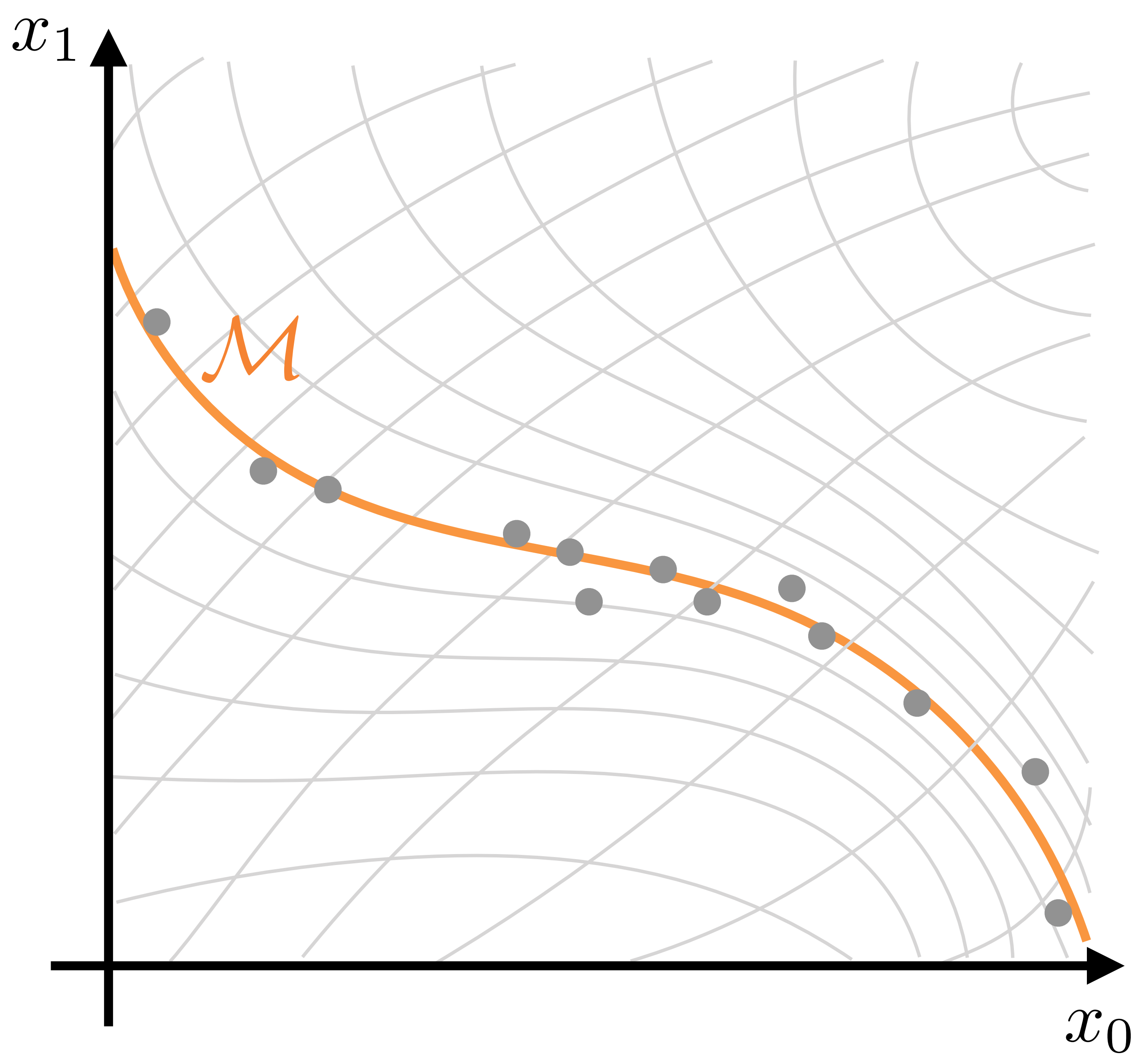
Questions:

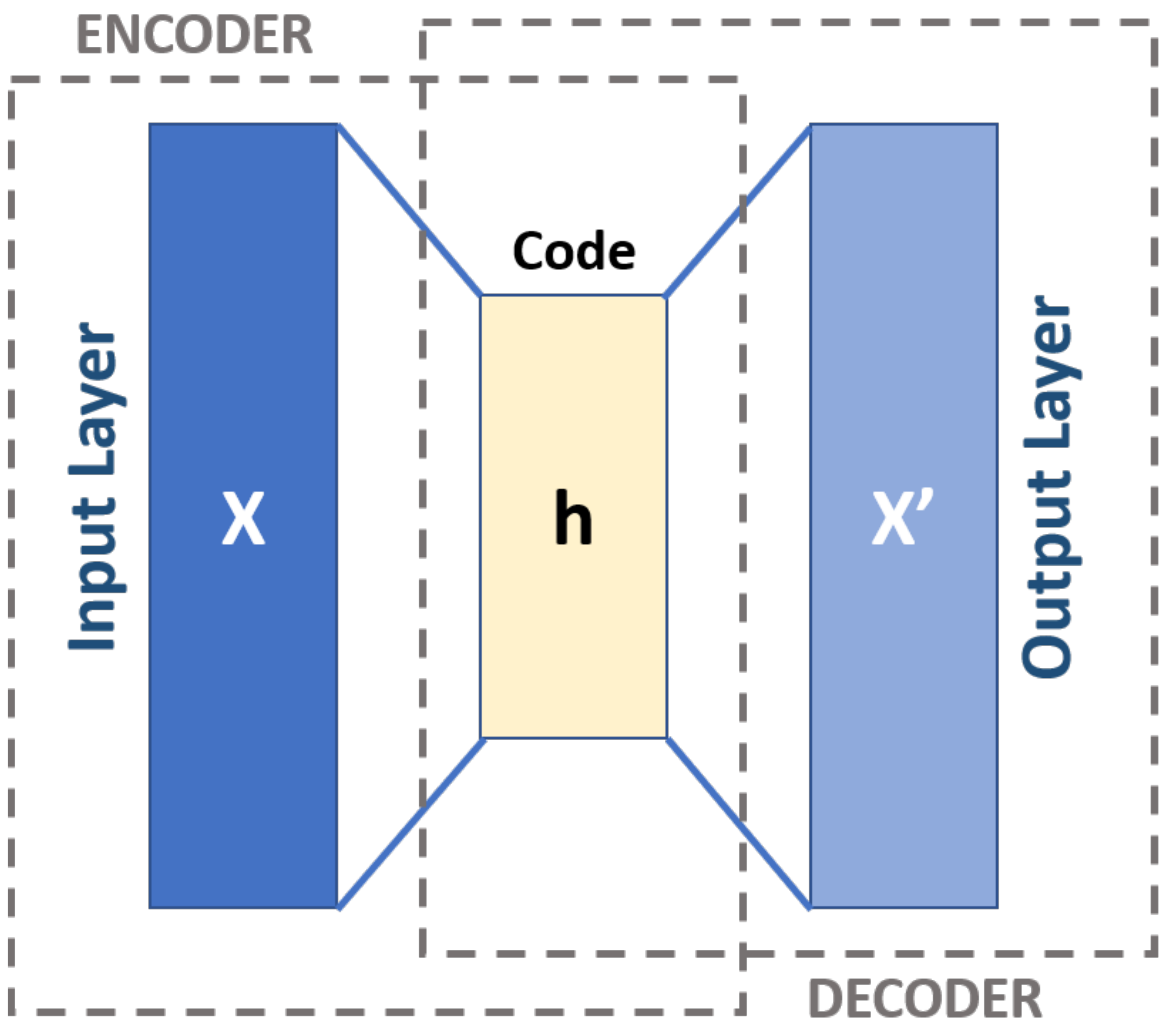
- How arbitrary or unambiguous are:
  - the scales where the "right" effective description applies?
  - the right objects / degrees of freedom in the effective description?
  - the laws that describe the interactions among those objects?
- Is there a principle that can help guide us or allow us to judge or rank different approaches?

Observation:

- Dynamics of lower-dimensional coarse-grained model sweep out a manifold in the state space of the fine-grained model
  - Coarse graining and emergence can be seen as geometrical structure of the "data manifold"
  - Useful insight for generative models, up-sampling, denoising etc.

Evaluating data on or off the manifold

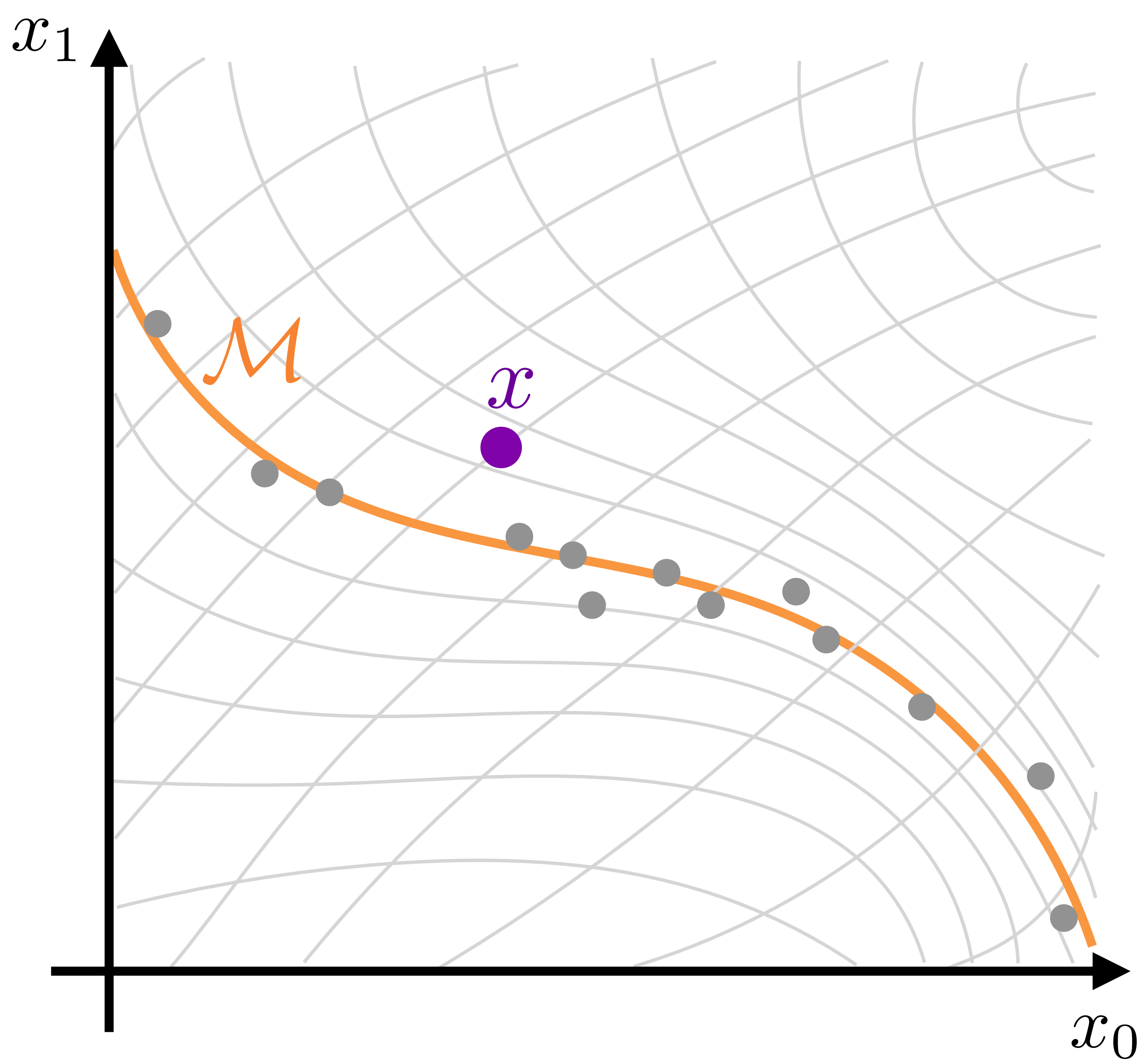


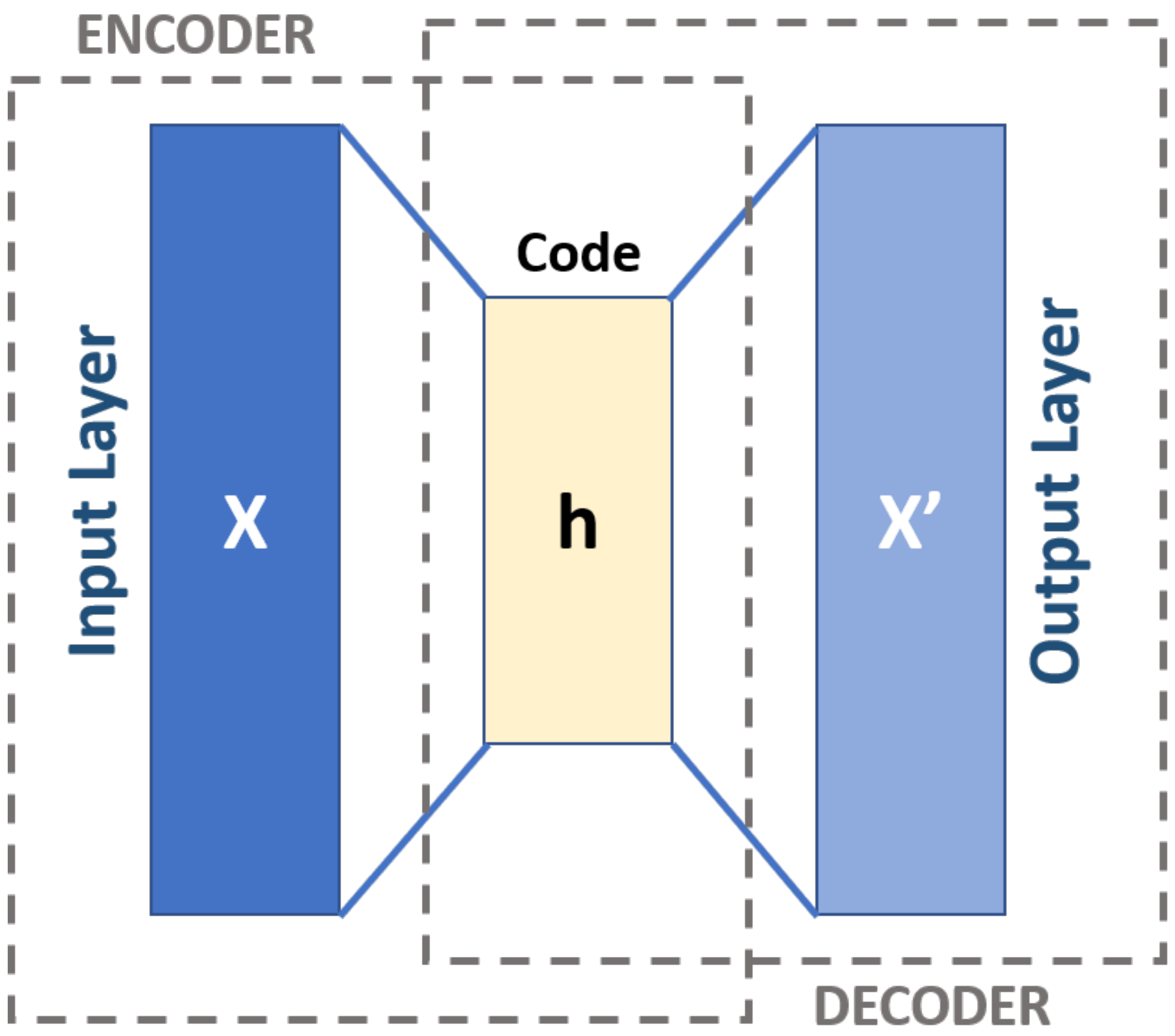


Vanilla autoencoder acting general-purpose like compression.

- When trained on L2 loss, not specialized for any particular down-stream task

Evaluating data on or off the manifold

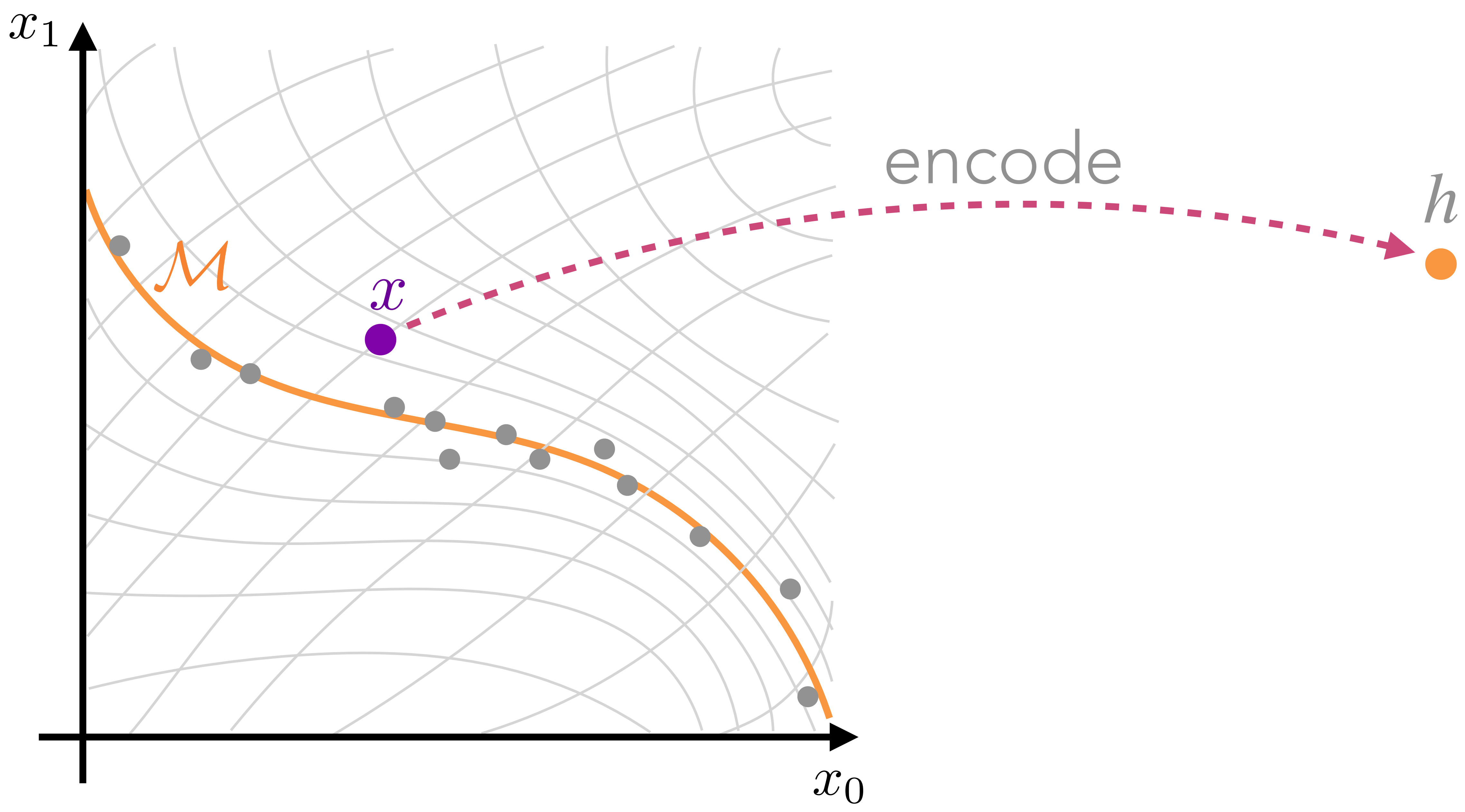


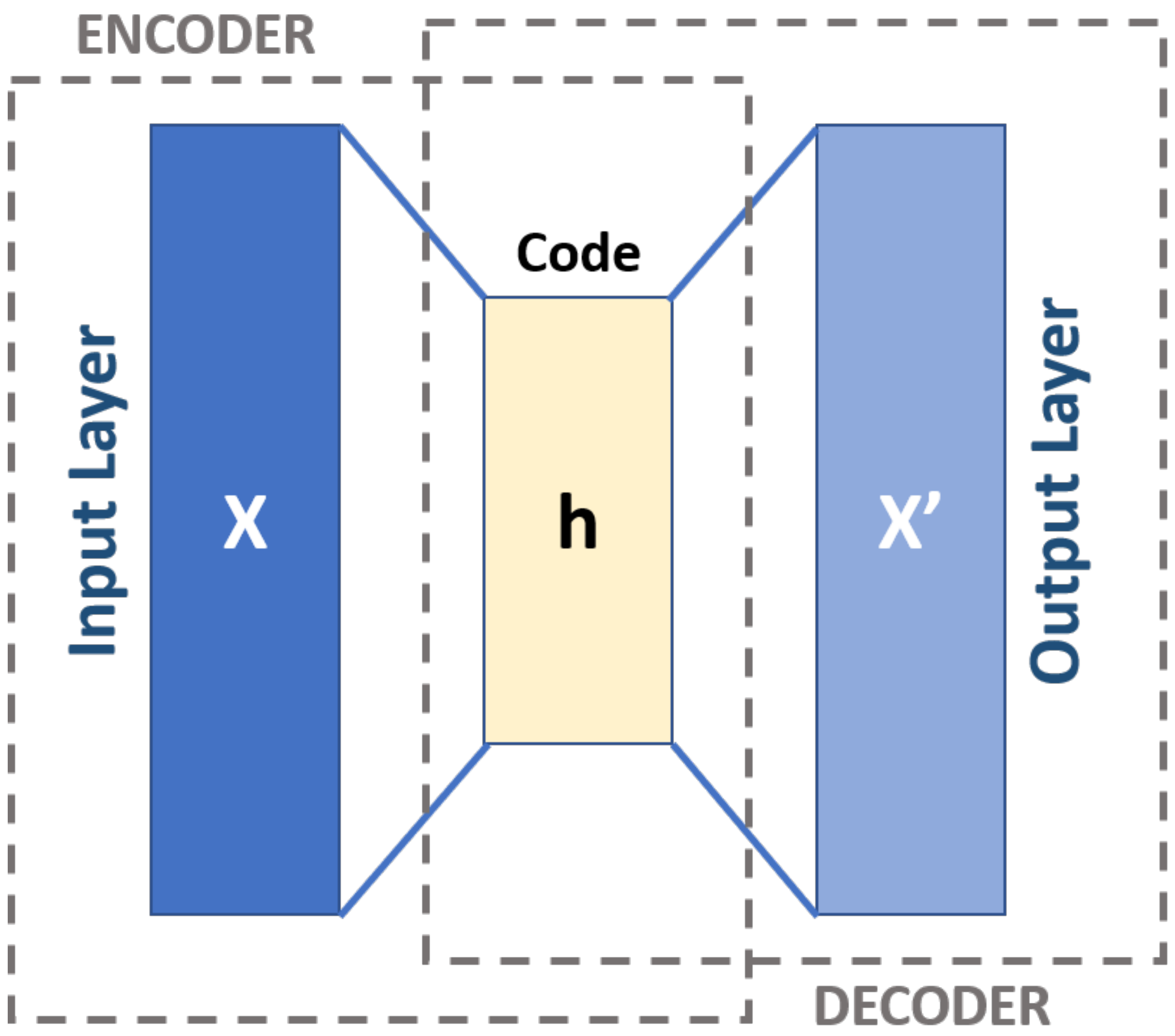


Vanilla autoencoder acting general-purpose like compression.

- When trained on L2 loss, not specialized for any particular down-stream task

Evaluating data on or off the manifold

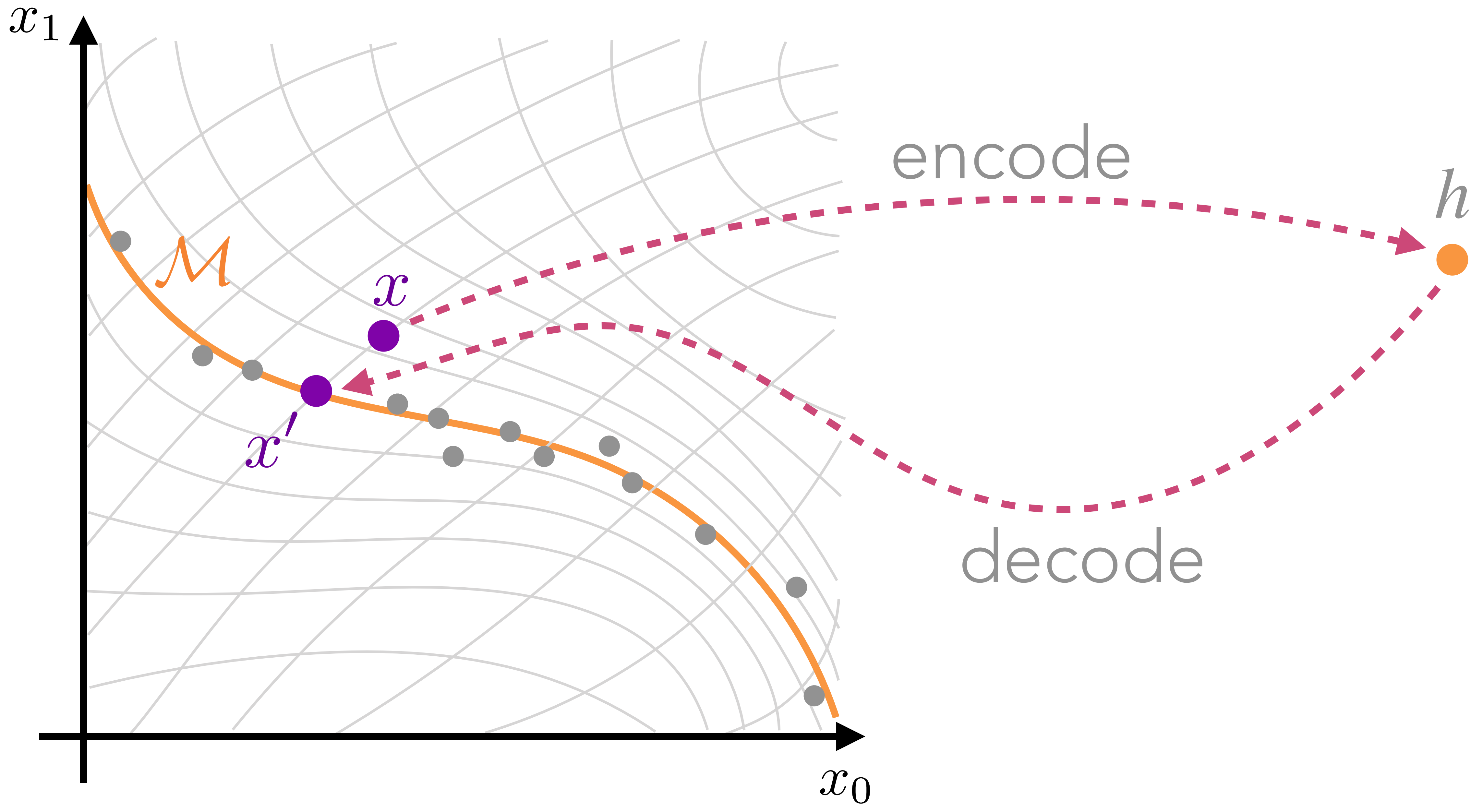


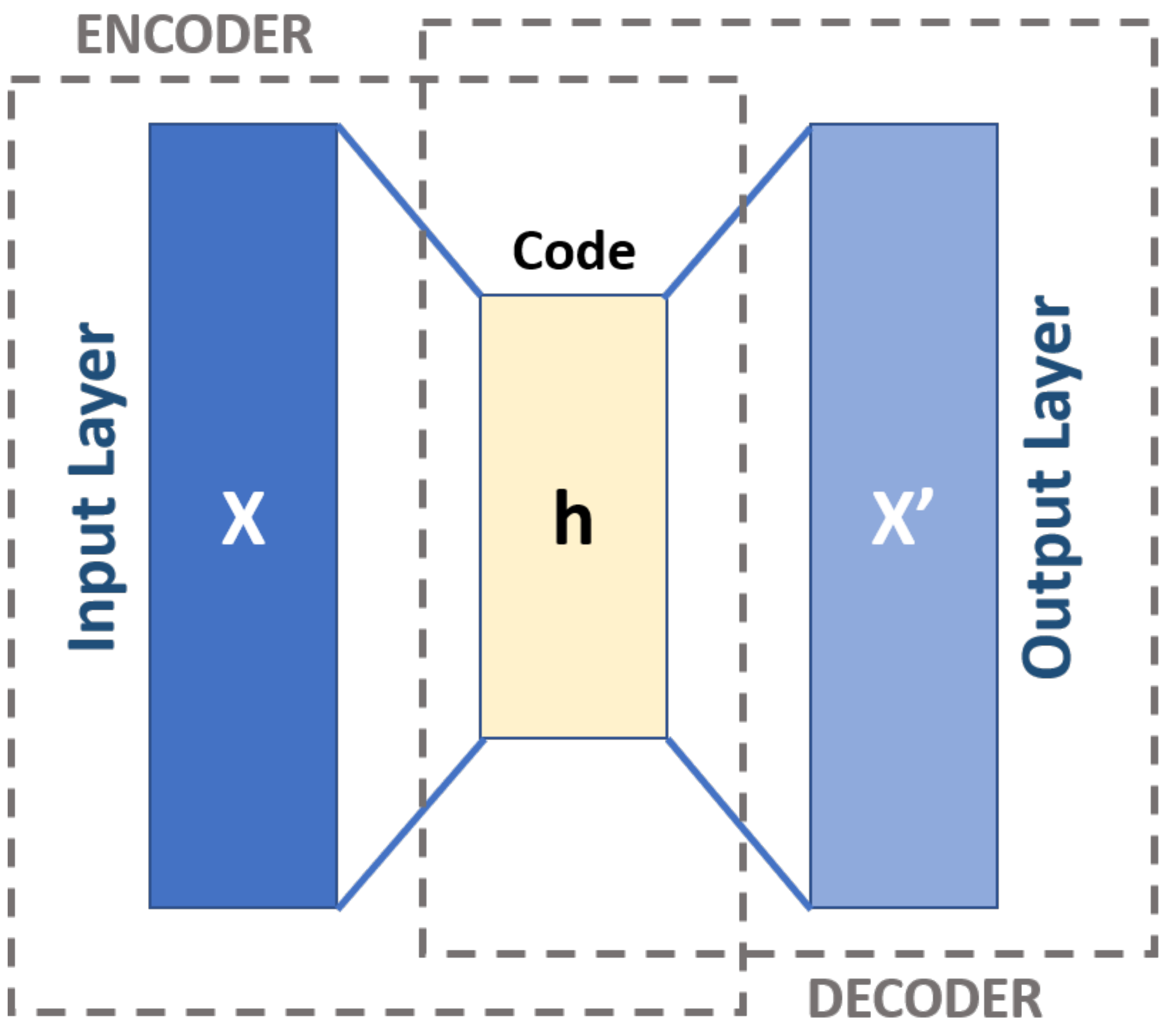


Vanilla autoencoder acting general-purpose like compression.

- When trained on L2 loss, not specialized for any particular down-stream task

Evaluating data on or off the manifold





Vanilla autoencoder acting general-purpose like compression.

- When trained on L2 loss, not specialized for any particular down-stream task

Supervised learning and sufficiency

In contrast, say we have some down-stream task, then some of the information in x will be useful, but other information can be thrown away without any significant loss in performance.

- This happens automatically in learned representations for supervised learning tasks

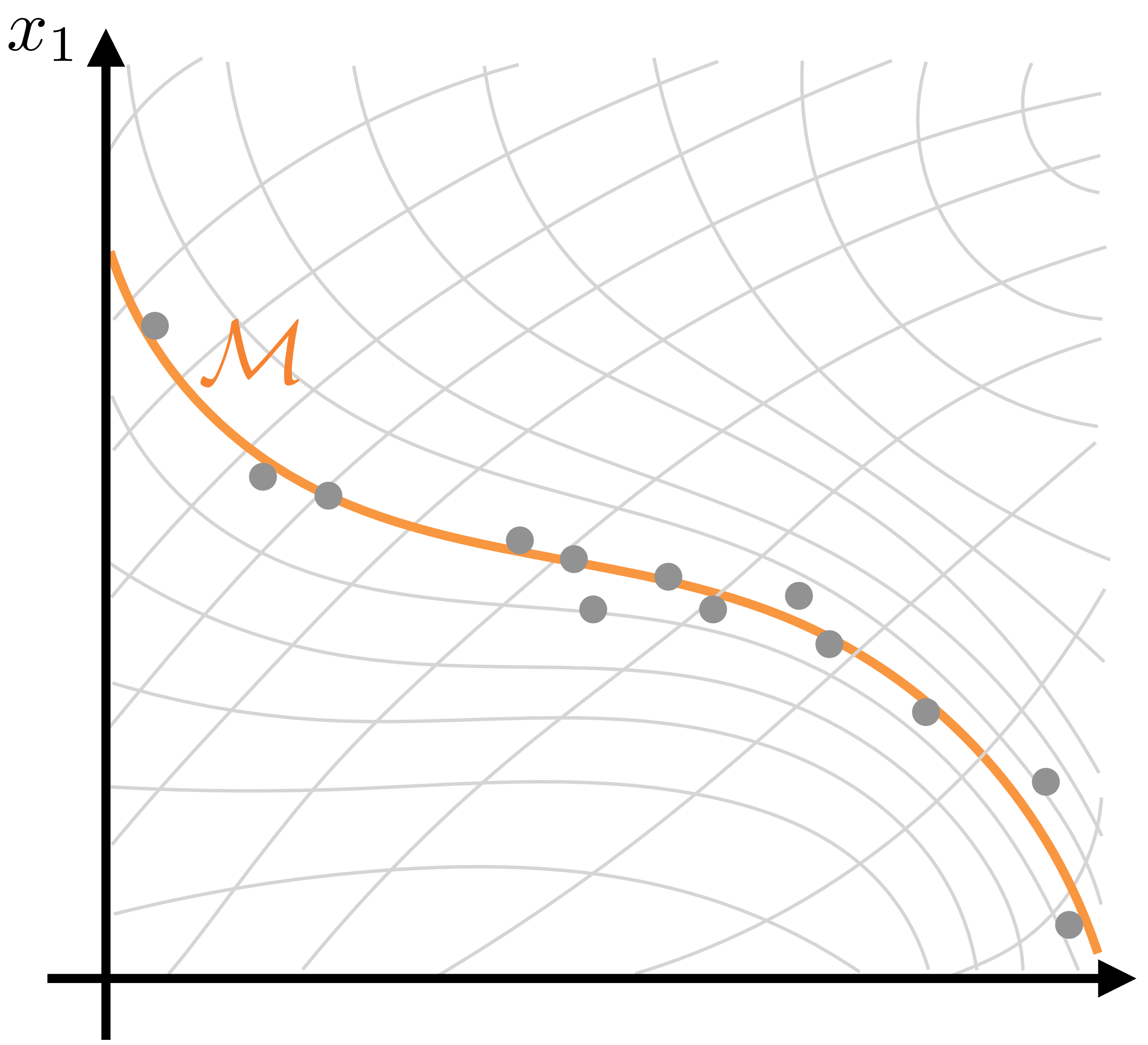
For example, say θ represents some property that is useful for my downstream task, and abstractly I can think about the joint $p(x, \theta)$ or conditional $p(x \mid \theta)$ — example: think of θ as a reaction coordinate

- A function (encoder) $T(x)$ is called a sufficient statistic for θ if it can be factorized as
  - $p(x \mid \theta) = g(T \mid \theta)\, h(x)$
- Equivalently
  - $I\big(\theta; T(X)\big) = I(\theta; X)$
  - $p(\theta \mid X = x) = p(\theta \mid T(X) = t(x))$

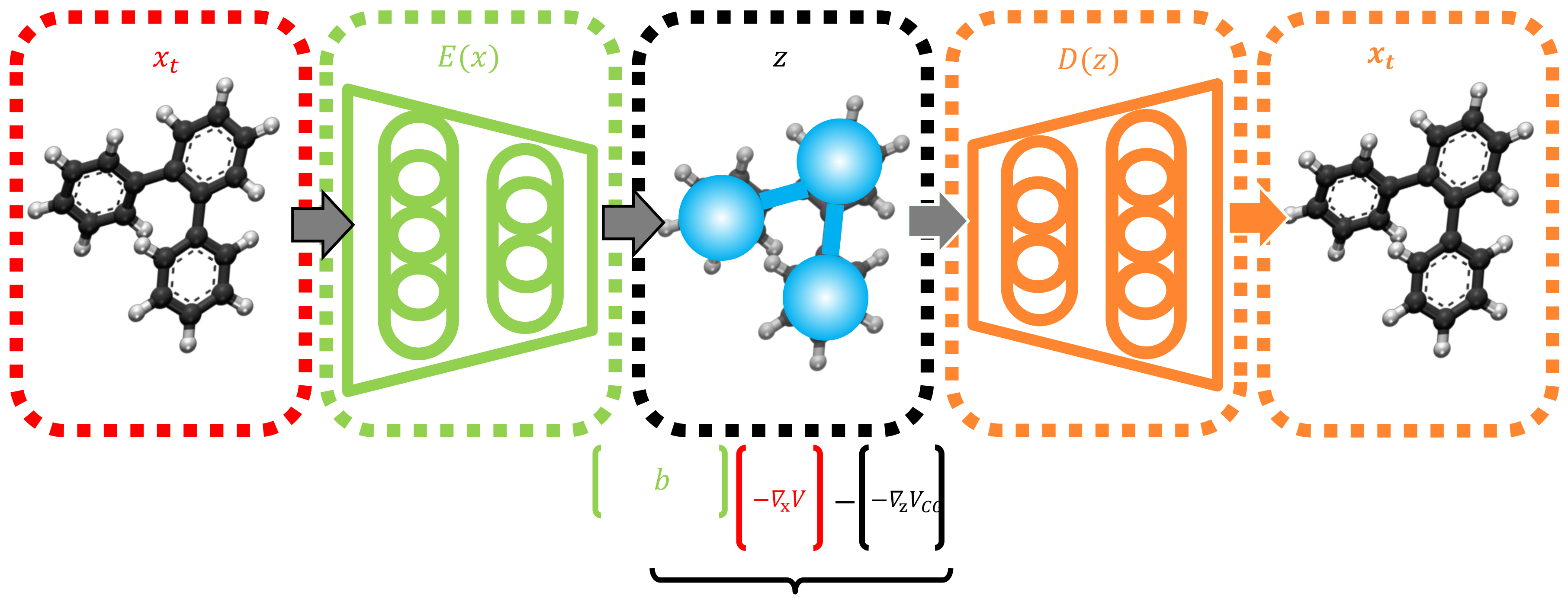
Closely related to collective variables, order parameters, etc.

- Exact sufficient statistics don't usually exist, but approximate sufficient globally
- $t(x \mid \theta') = \nabla_\theta \log p(x \mid \theta)\,|_{\theta'}$ is "locally sufficient"

Geometrical Picture



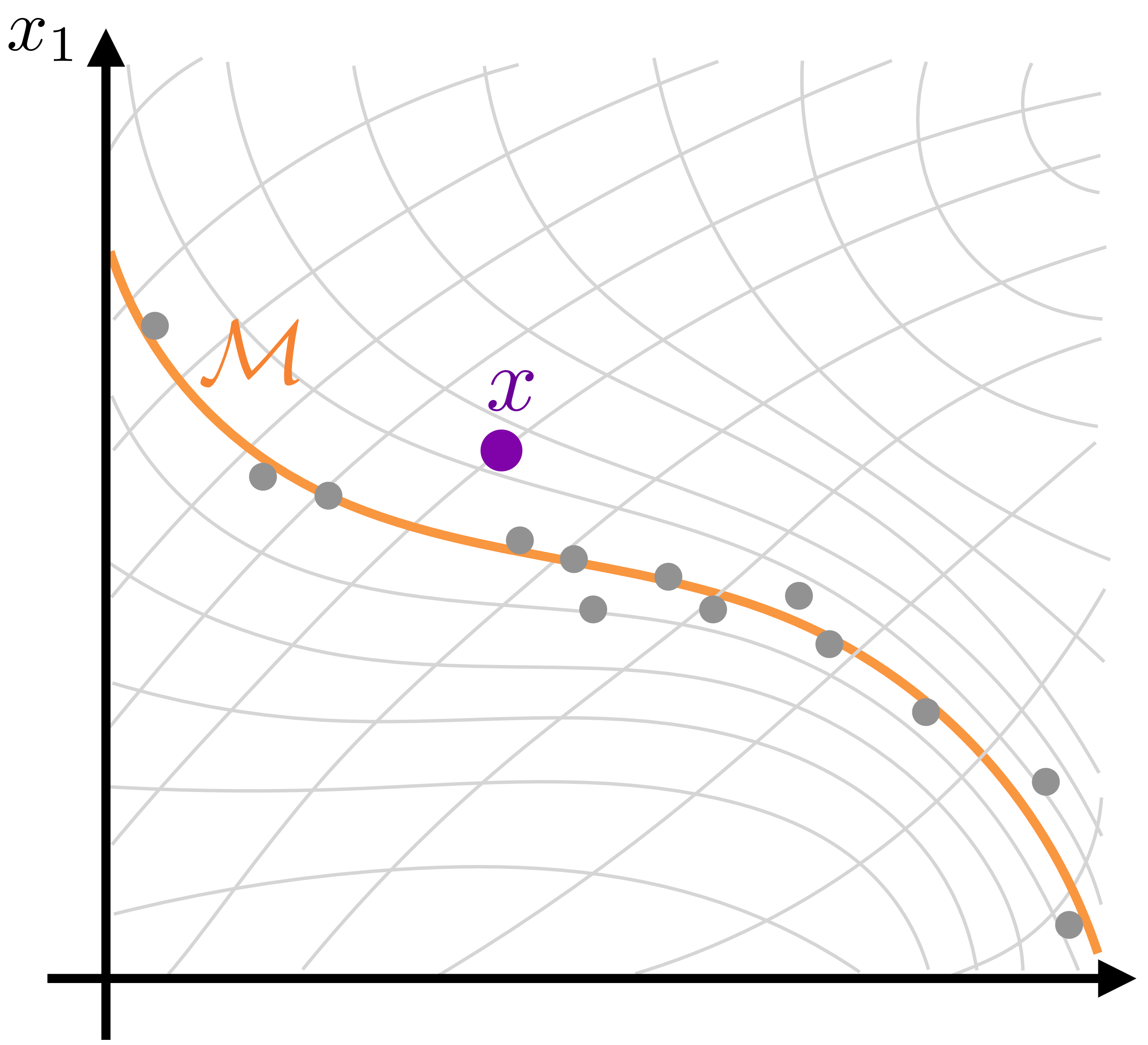
Coarse Graining Auto-Encoding Framework



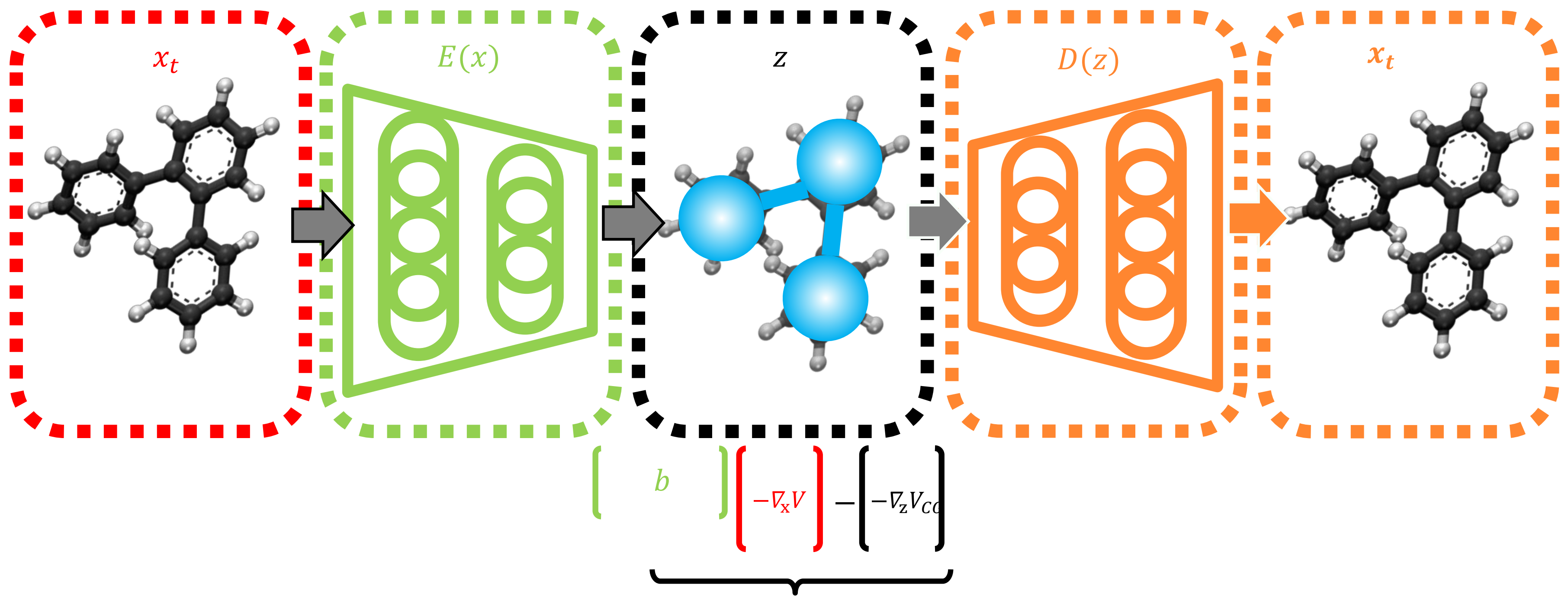
$$L_{VCGE} = |-b \cdot \nabla_x V + \nabla_z V_{CG}(E(x))|^2 + |D\,E\,(x_t) - x_t|^2 + \rho F_{ins}\left(E(x)\right)^2$$

- **AutoEncoder** automatically coarse-grains atomistic coordinates to CG coordinates in a data-driven way
- **Force matching** also helps to shape the learning of CG and obtain $V_{CG}(z = E(x))$ for CG simulations

Geometrical Picture



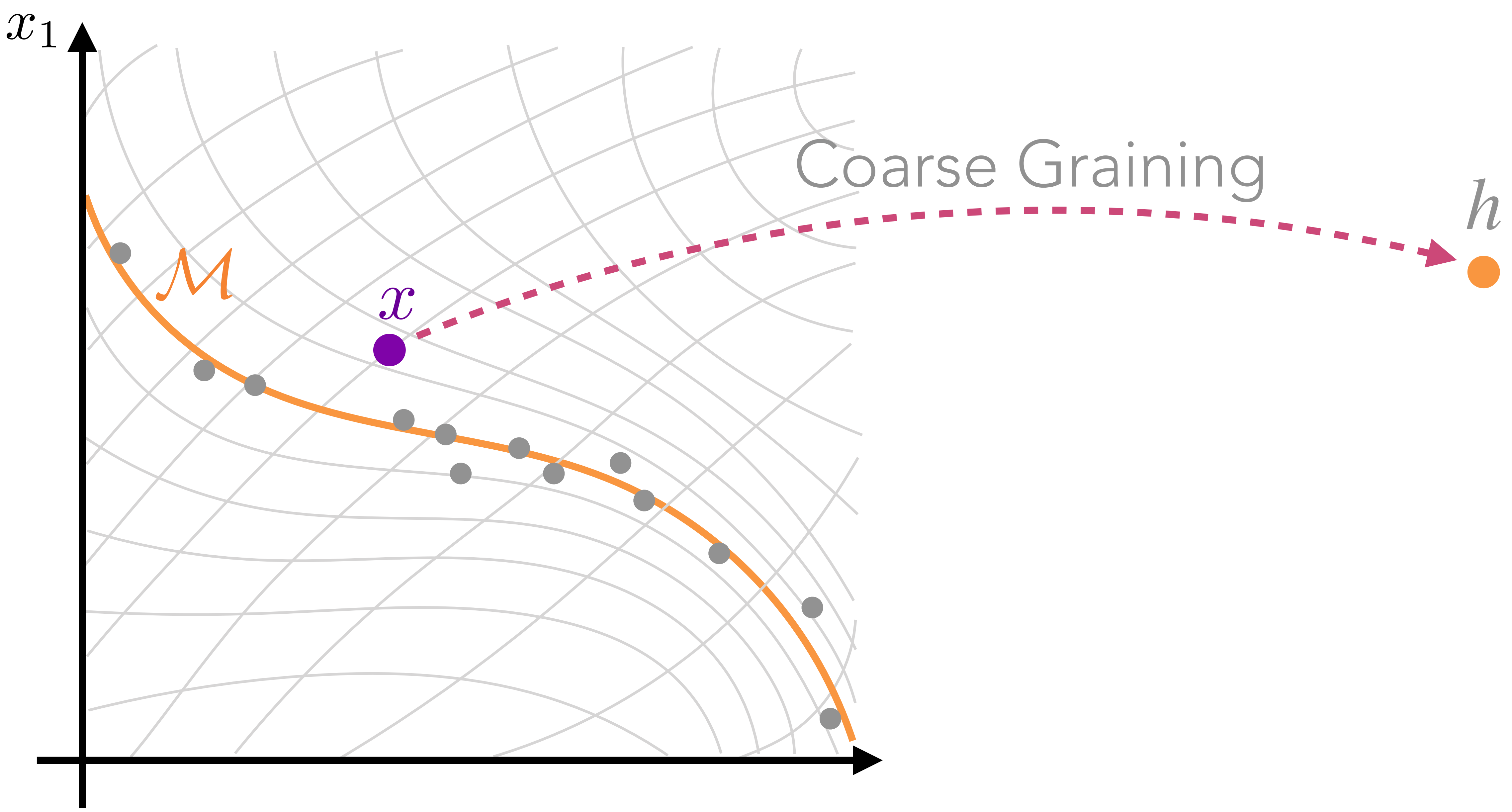
Coarse Graining Auto-Encoding Framework



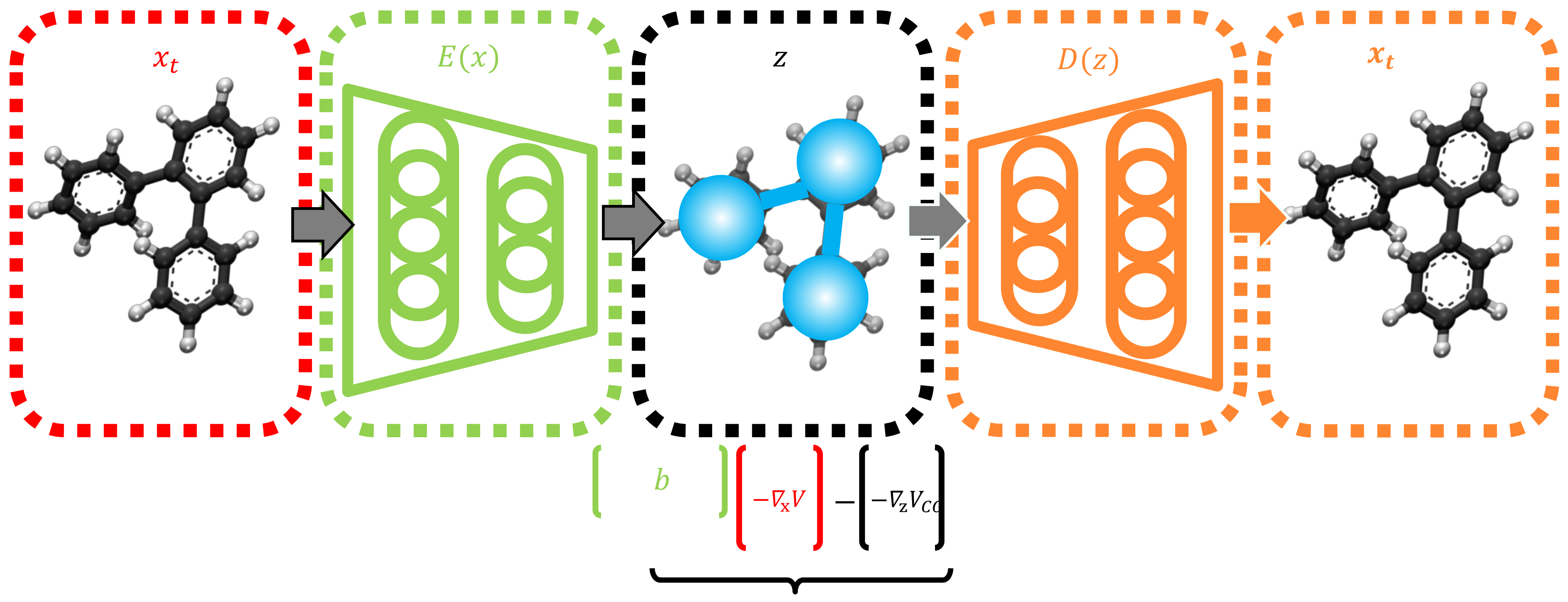
$$L_{VCGE} = |-b \cdot \nabla_x V + \nabla_z V_{CG}(E(x))|^2 + |D\,E\,(x_t) - x_t|^2 + \rho F_{ins}\,(E(x))^2$$

- **AutoEncoder** automatically coarse-grains atomistic coordinates to CG coordinates in a data-driven way
- **Force matching** also helps to shape the learning of CG and obtain $V_{CG}(z = E(x))$ for CG simulations

Geometrical Picture



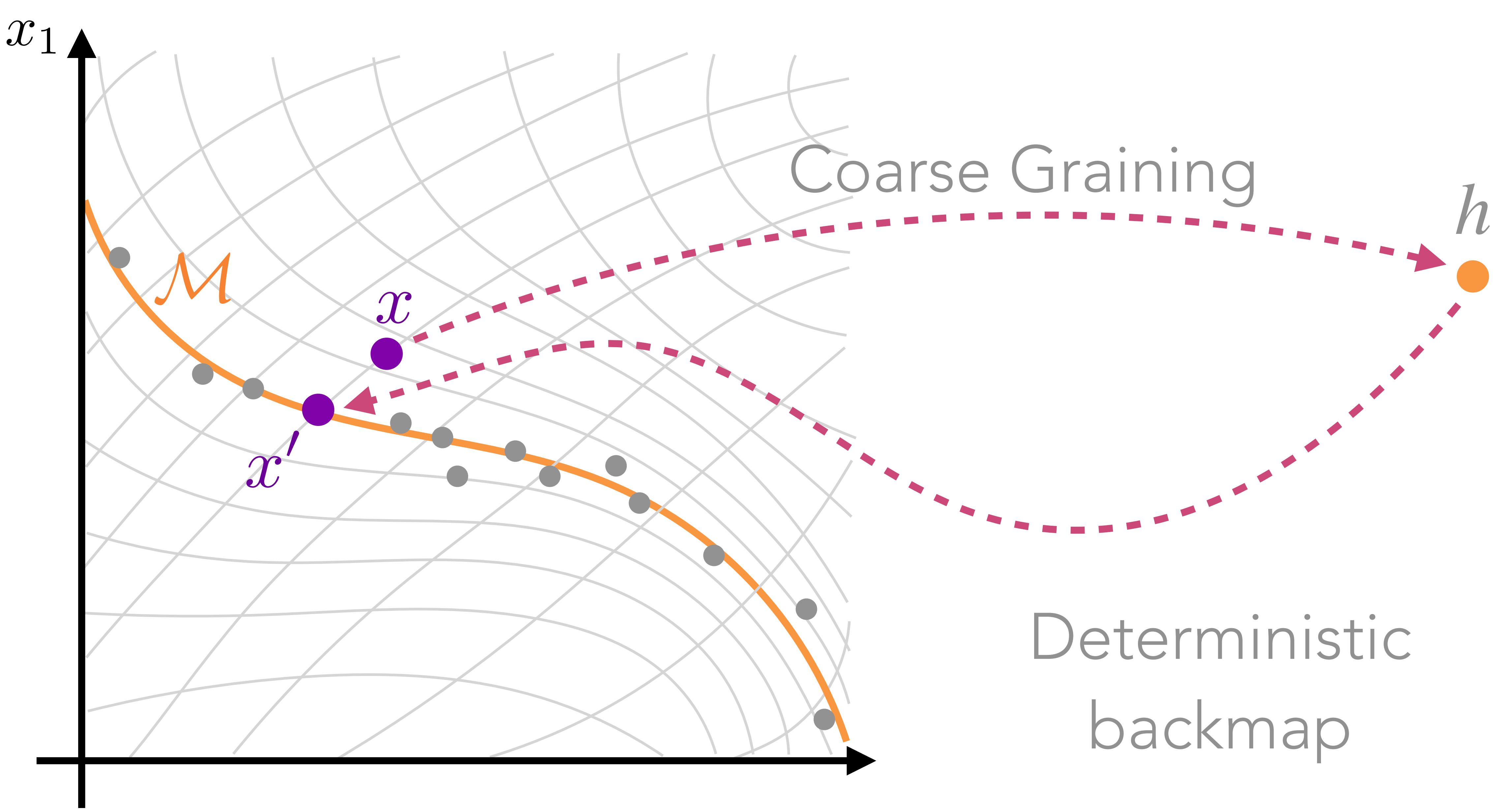
Coarse Graining Auto-Encoding Framework



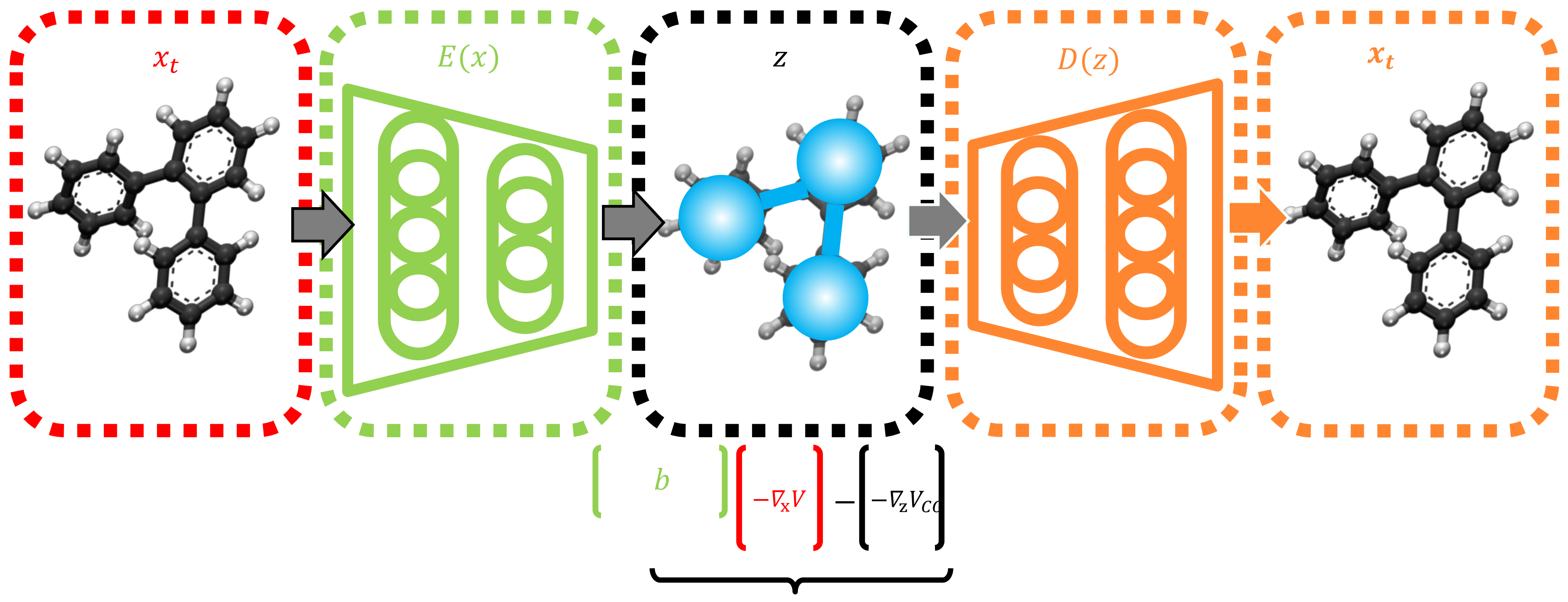
$$L_{VCGE} = |-b \cdot \nabla_x V + \nabla_z V_{CG}(E(x))|^2 + |D\,E(x_t) - x_t|^2 + \rho F_{ins}(E(x))^2$$

- **AutoEncoder** automatically coarse-grains atomistic coordinates to CG coordinates in a data-driven way
- **Force matching** also helps to shape the learning of CG and obtain $V_{CG}(z = E(x))$ for CG simulations

Geometrical Picture



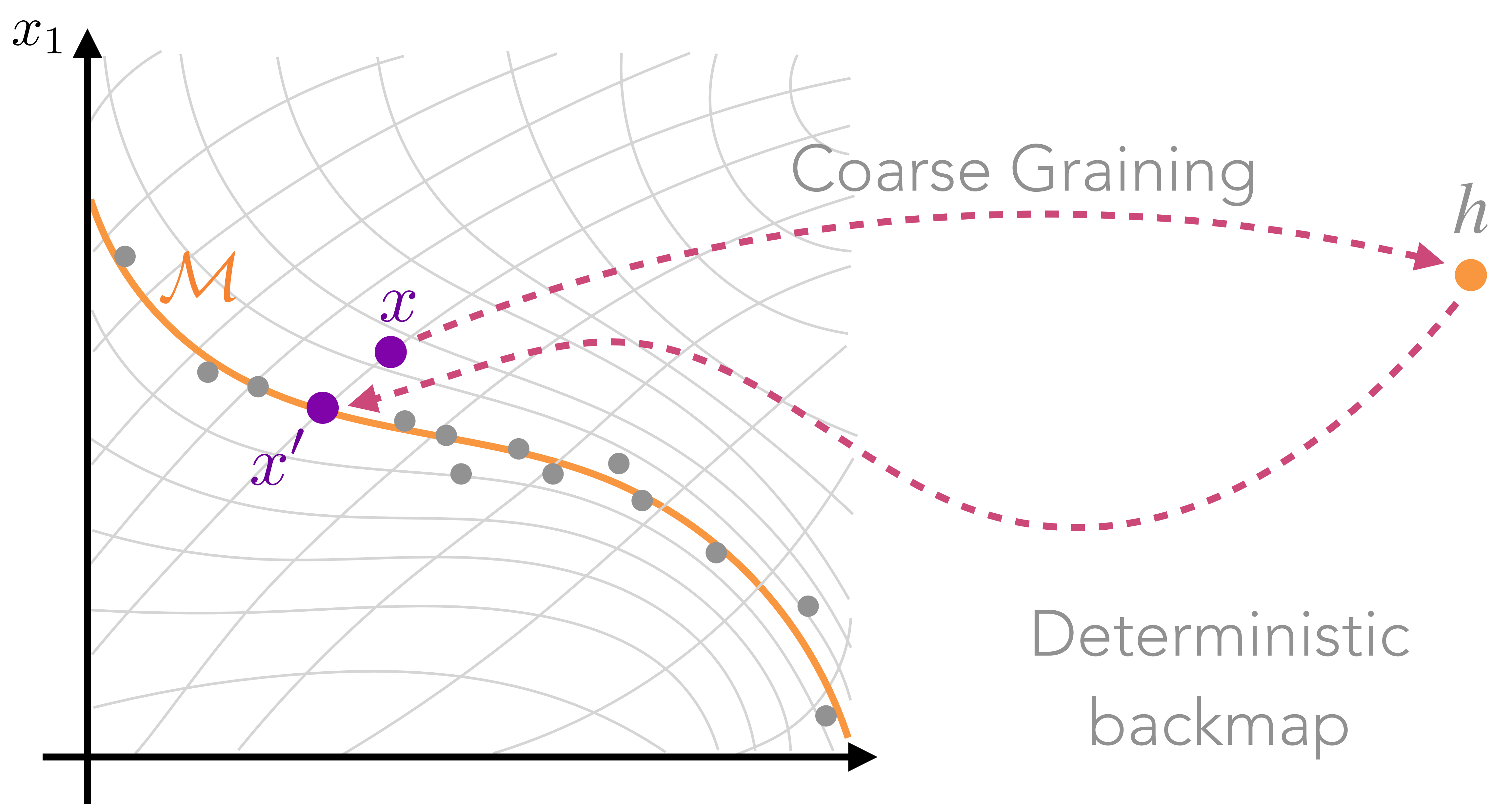
Coarse Graining Auto-Encoding Framework



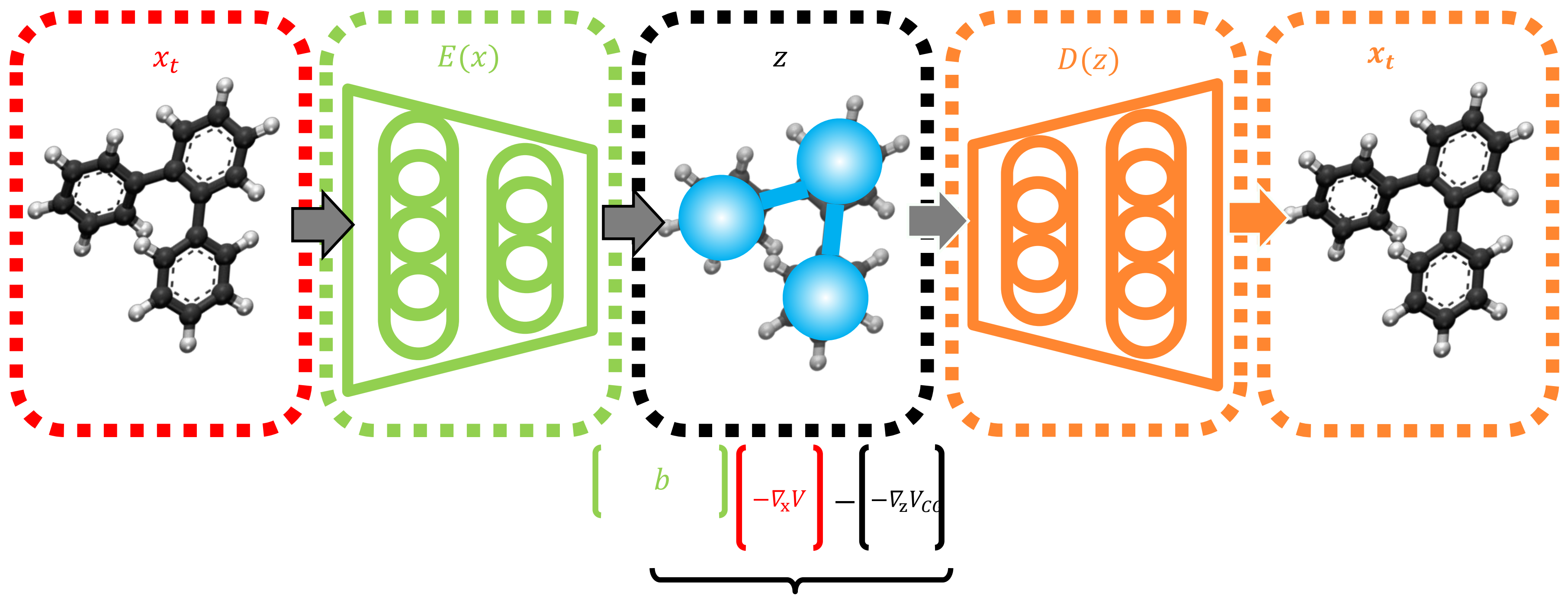
$$L_{VCGE} = |-b \cdot \nabla_x V + \nabla_z V_{CG}(E(x))|^2 + |D\,E(x_t) - x_t|^2 + \rho F_{ins}(E(x))^2$$

- **AutoEncoder** automatically coarse-grains atomistic coordinates to CG coordinates in a data-driven way
- **Force matching** also helps to shape the learning of CG and obtain $V_{CG}(z = E(x))$ for CG simulations

Geometrical Picture



Coarse Graining Auto-Encoding Framework



$$L_{VCGE} = |-b \cdot \nabla_x V + \nabla_z V_{CG}(E(x))|^2 + |D\,E(x_t) - x_t|^2 + \rho F_{ins}\,(E(x))^2$$

- **AutoEncoder** automatically coarse-grains atomistic coordinates to CG coordinates in a data-driven way
- **Force matching** also helps to shape the learning of CG and obtain $V_{CG}(z = E(x))$ for CG simulations

Equivariant generative decoder

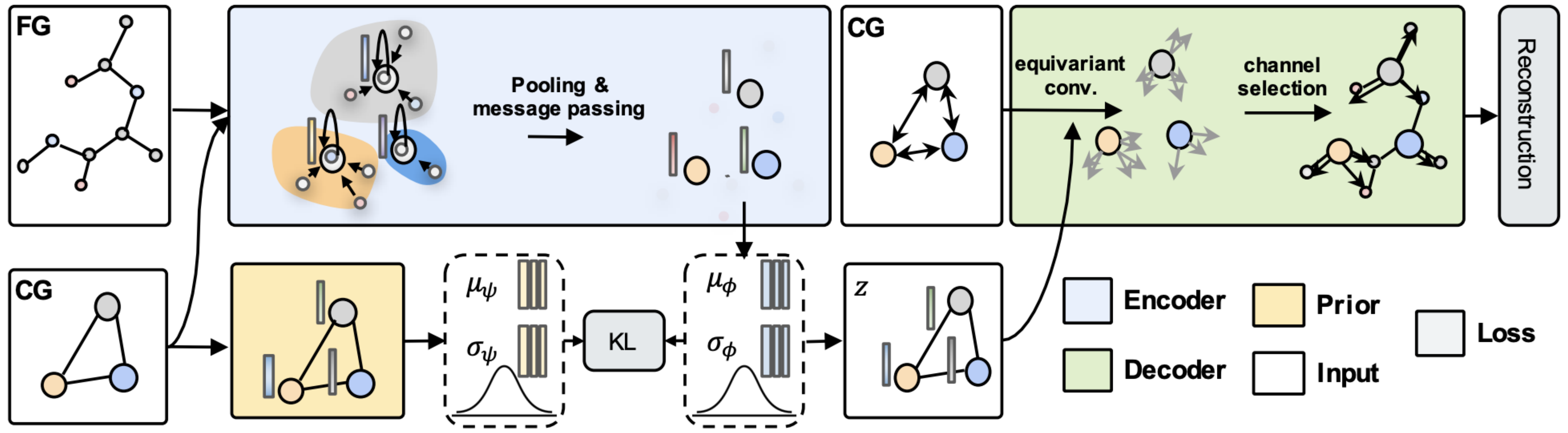
Information is always lost in CG – needs to be recovered statistically.

All atom to CG is surjective, a generative (non-deterministic) model is needed

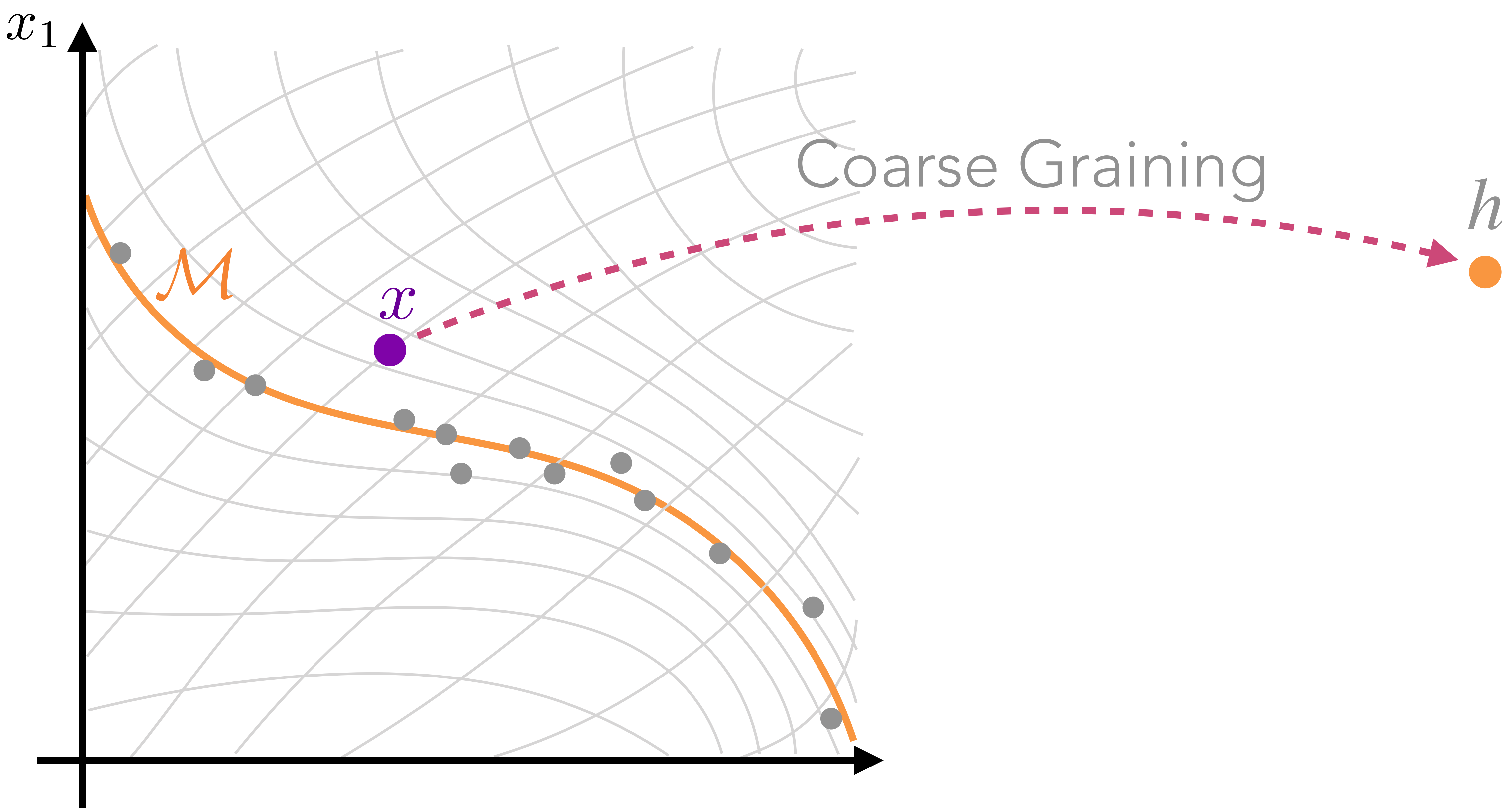
Avoid FF refinement.

Create latent variable to hold info for decoding (depends on x and X at train and only on X during inference.

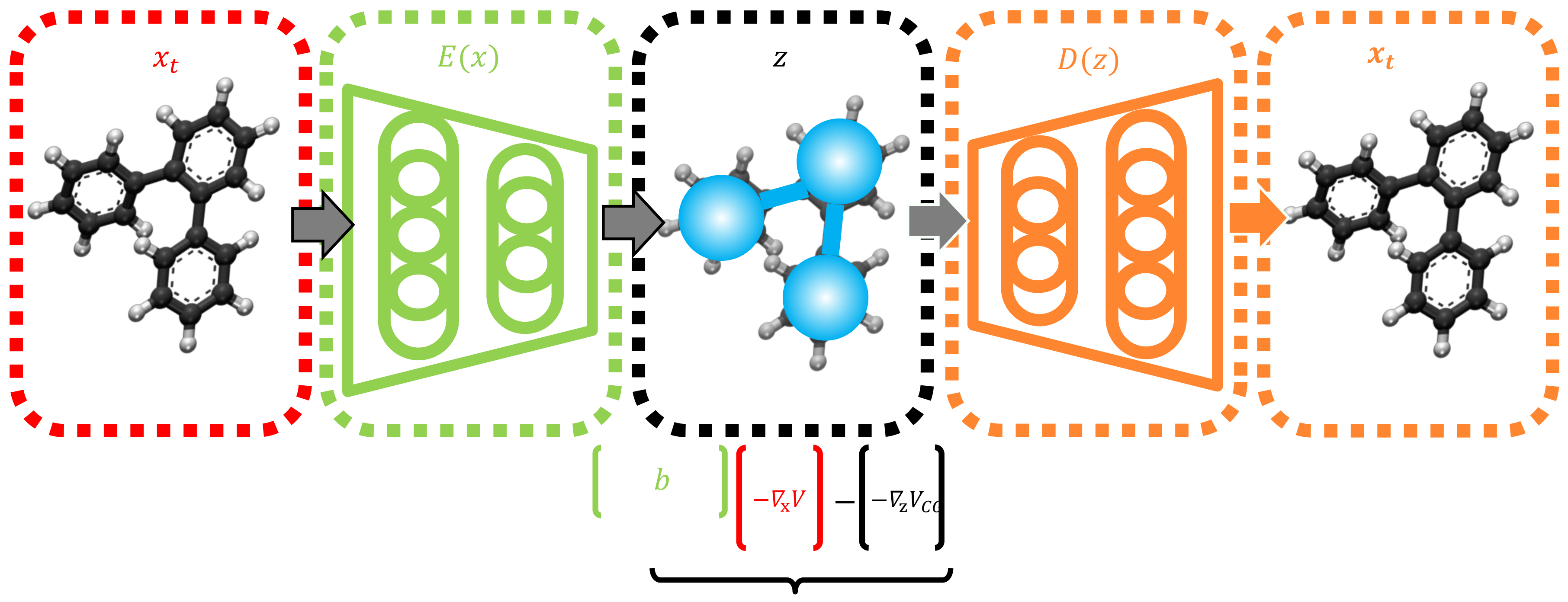
Equivariant decoding through inter-bead vectors.



Geometrical Picture



Coarse Graining Auto-Encoding Framework



$$L_{VCGE} = |-b\cdot\nabla_x V + \nabla_z V_{CG}(E(x))|^2 + |D\,E(x_t) - x_t|^2 + \rho F_{ins}\,(E(x))^2$$

- **AutoEncoder** automatically coarse-grains atomistic coordinates to CG coordinates in a data-driven way
- **Force matching** also helps to shape the learning of CG and obtain $V_{CG}(z = E(x))$ for CG simulations

Equivariant generative decoder

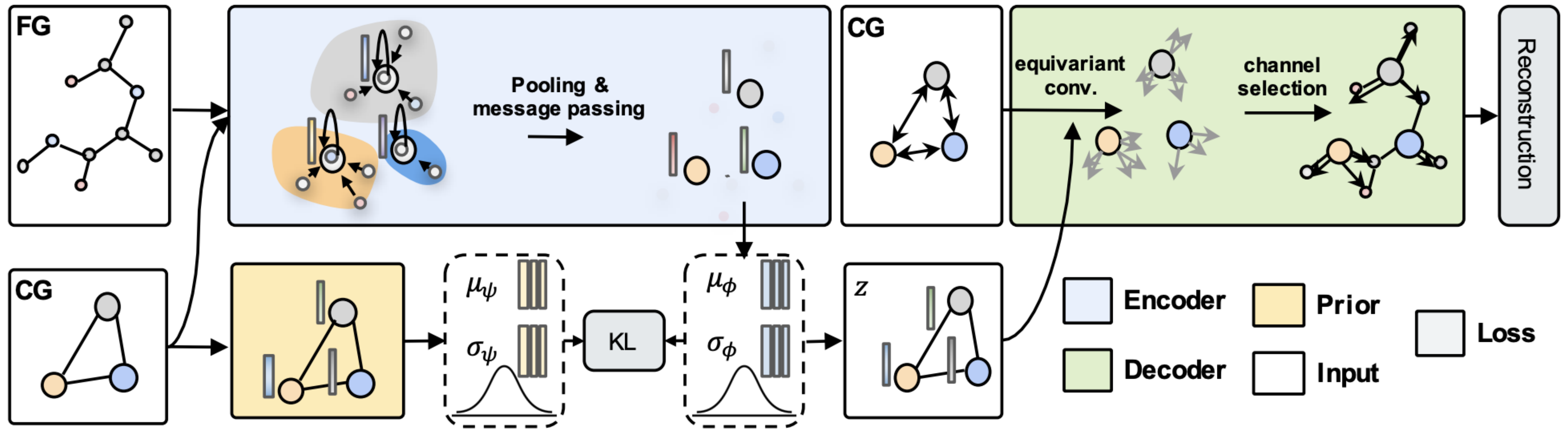
Information is always lost in CG – needs to be recovered statistically.

All atom to CG is surjective, a generative (non-deterministic) model is needed

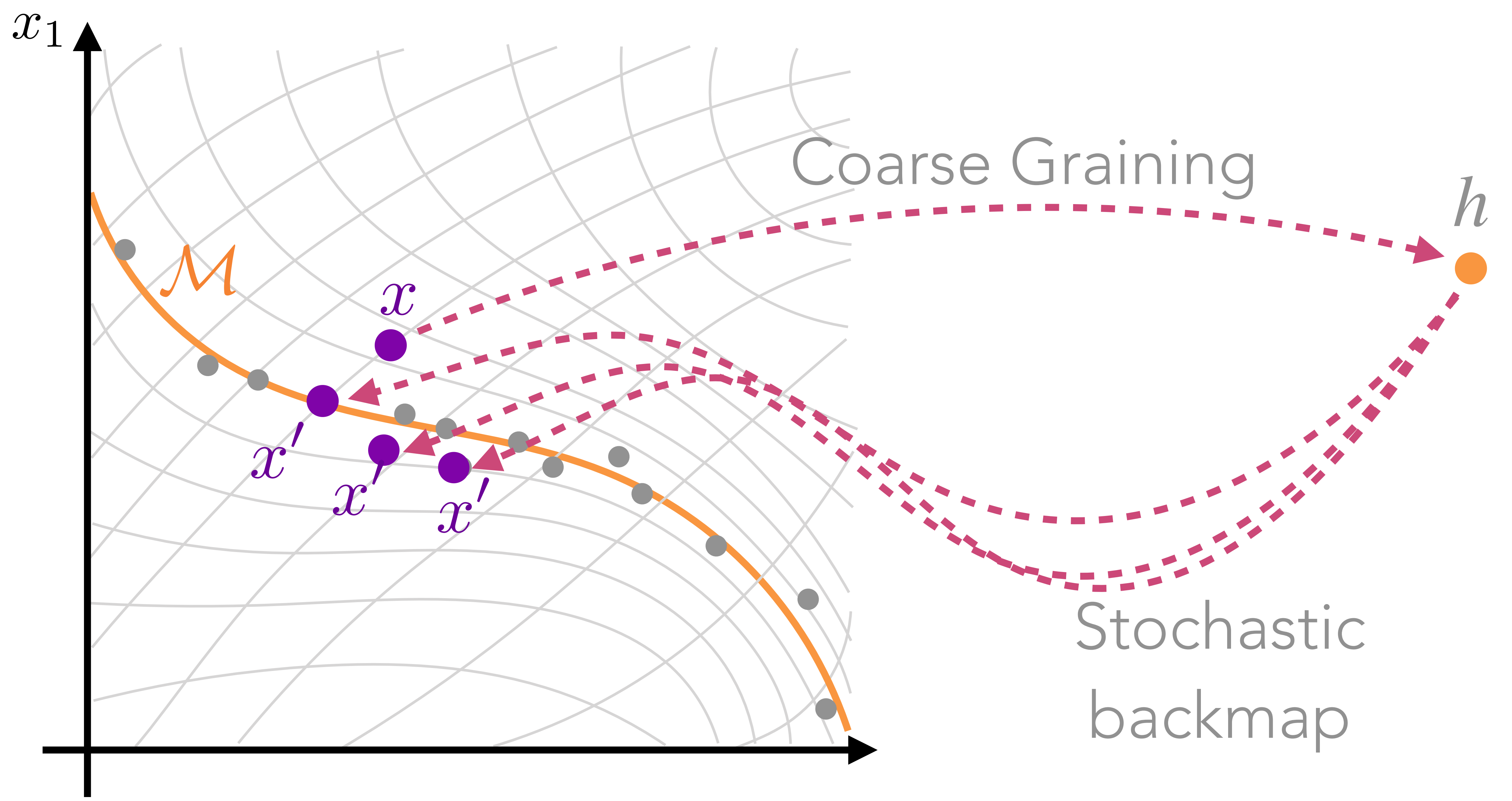
Avoid FF refinement.

Create latent variable to hold info for decoding (depends on x and X at train and only on X during inference.

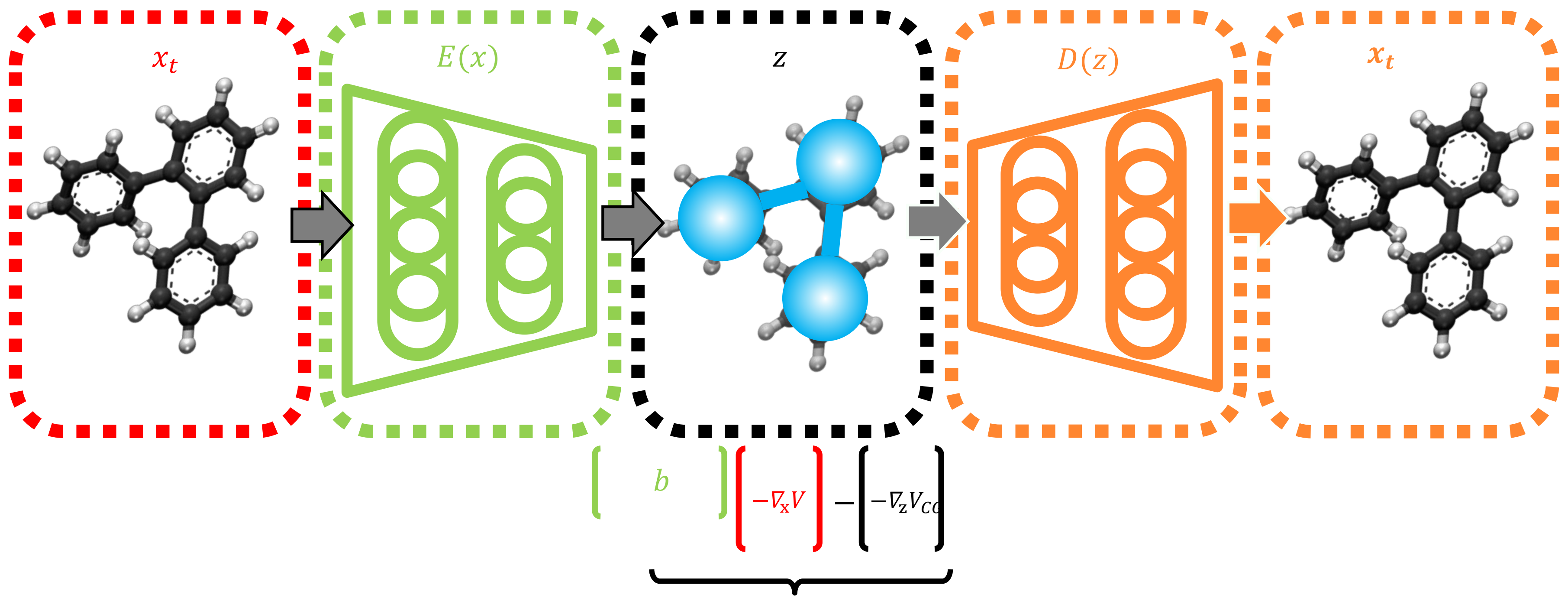
Equivariant decoding through inter-bead vectors.



Geometrical Picture



Coarse Graining Auto-Encoding Framework



$$L_{VCGE} = |-b \cdot \nabla_x V + \nabla_z V_{CG}(E(x))|^2 + |D\,E\,(x_t) - x_t|^2 + \rho F_{ins}\,(E(x))^2$$

- **AutoEncoder** automatically coarse-grains atomistic coordinates to CG coordinates in a data-driven way
- **Force matching** also helps to shape the learning of CG and obtain $V_{CG}(z = E(x))$ for CG simulations

Equivariant generative decoder

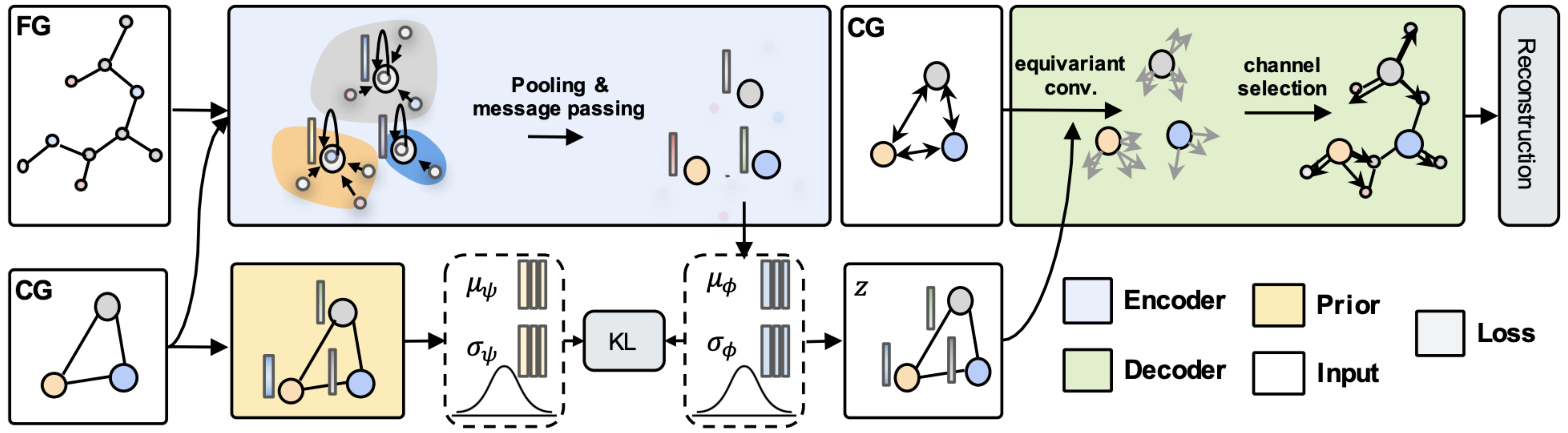
Information is always lost in CG – needs to be recovered statistically.

All atom to CG is surjective, a generative (non-deterministic) model is needed

Avoid FF refinement.

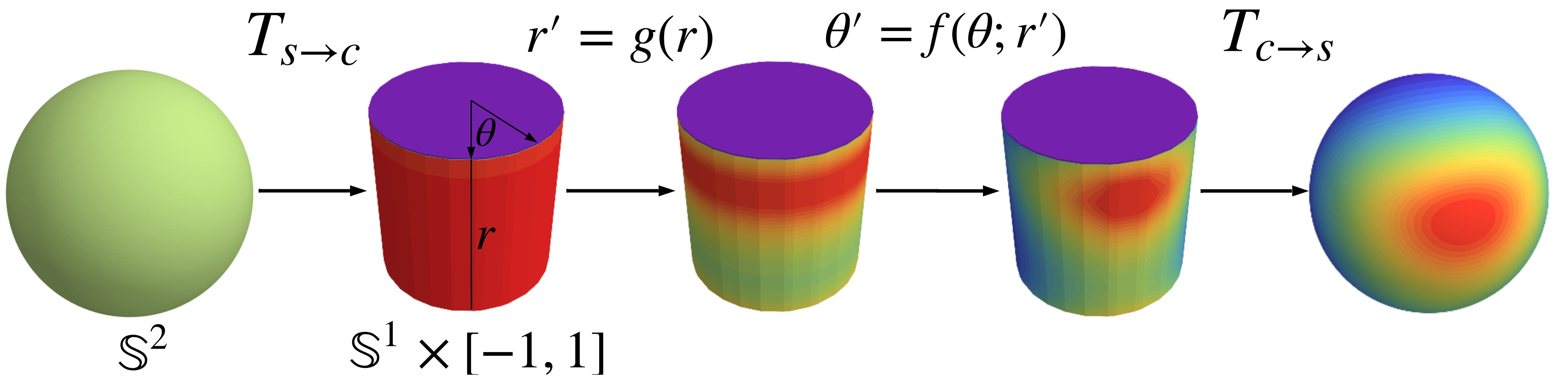
Create latent variable to hold info for decoding (depends on x and X at train and only on X during inference.

Equivariant decoding through inter-bead vectors.



[arXiv:2002.02428, ICML2020]

We designed flows on compact manifolds like Spheres and Tori



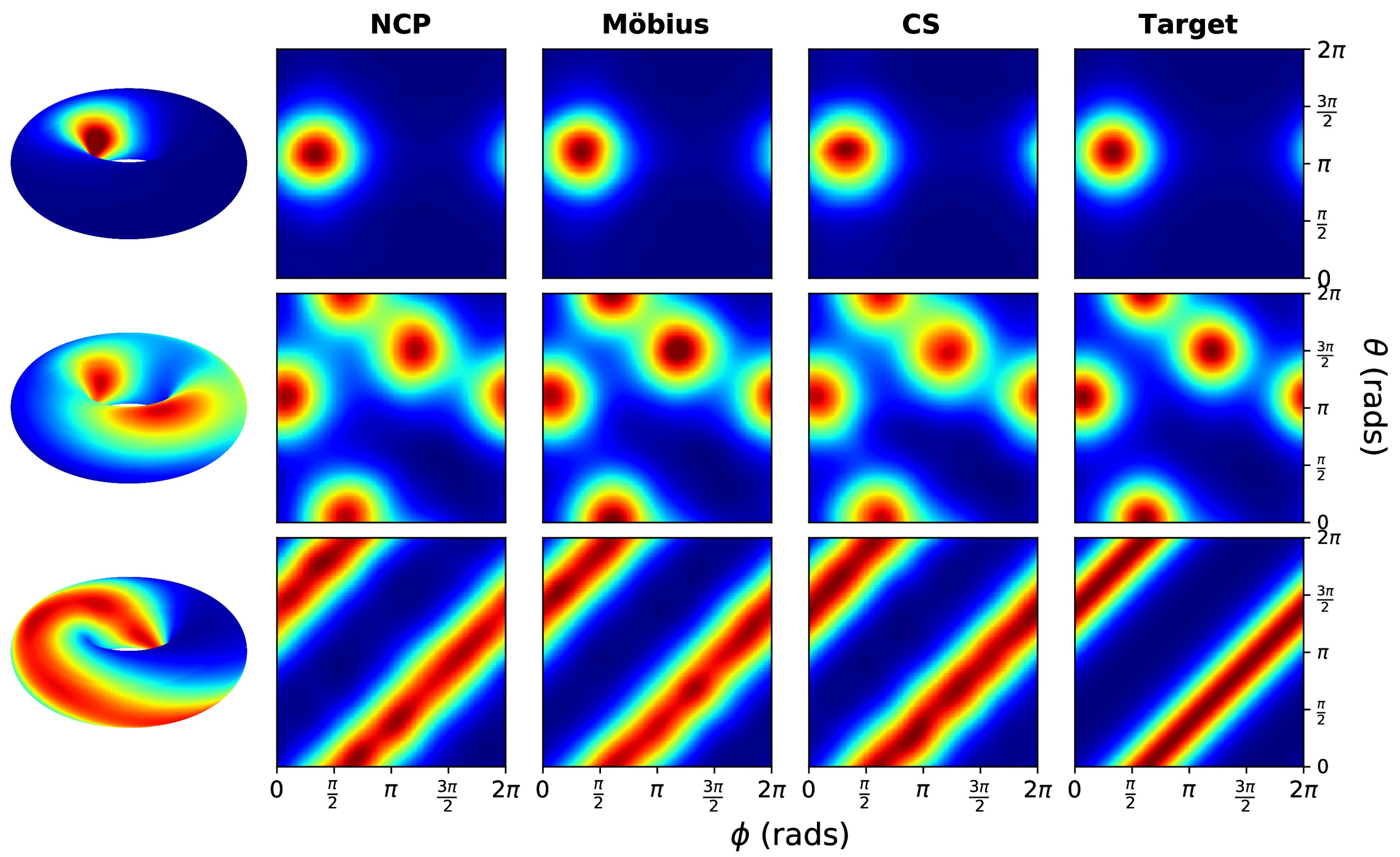
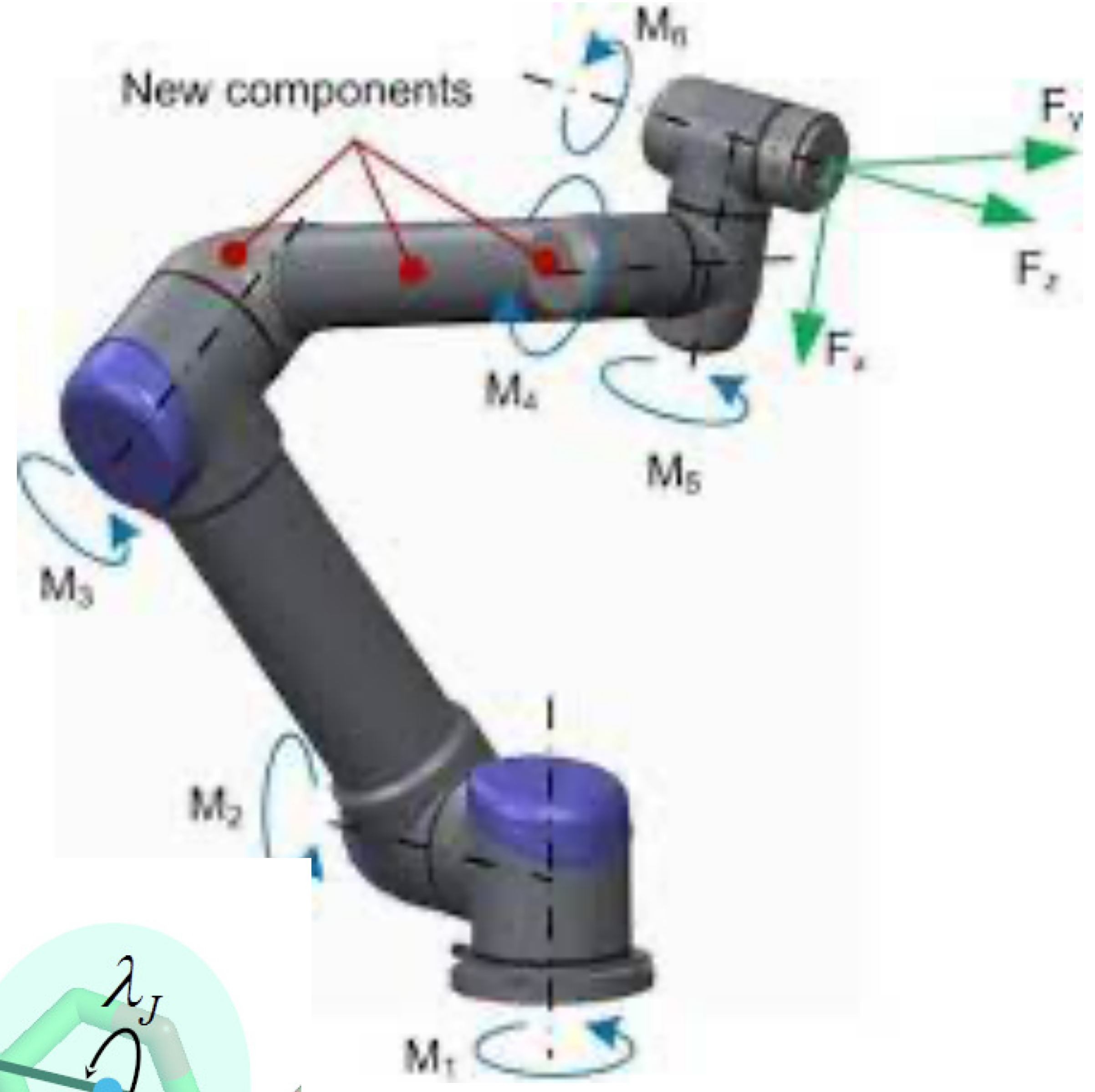


Figure 3. Learned densities on $\mathbb{T}^2$ using NCP, Möbius and CS flows. Densities shown on the torus are from NCP.



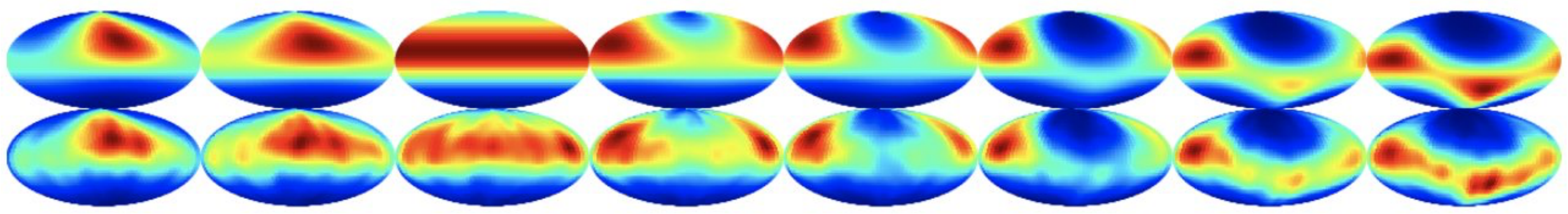
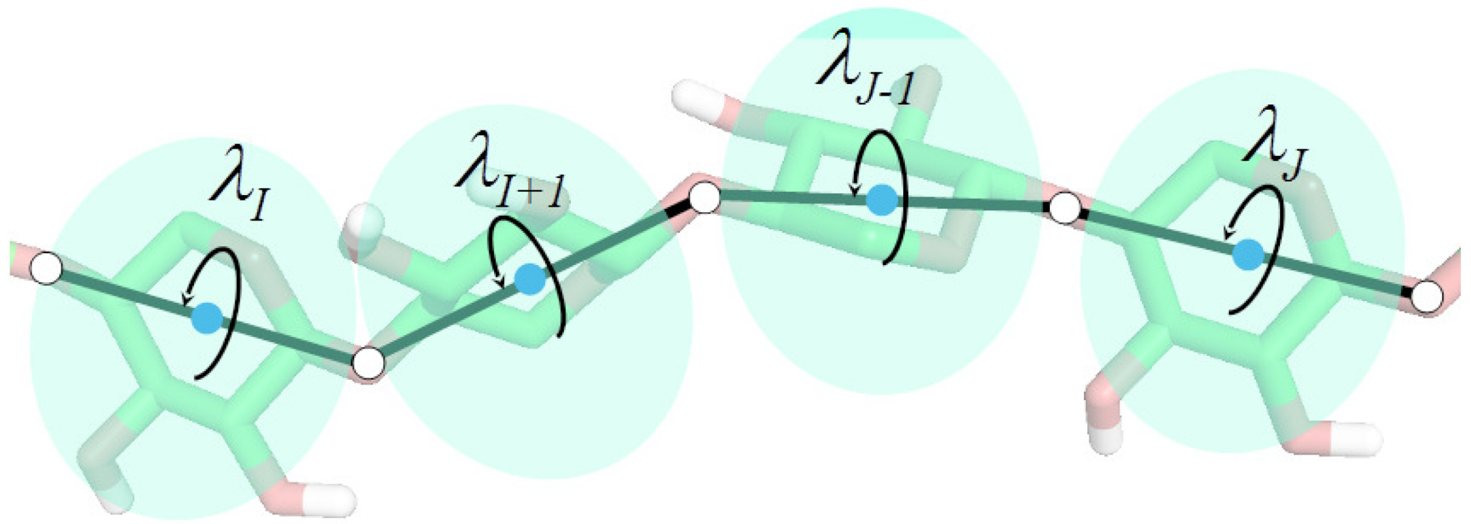
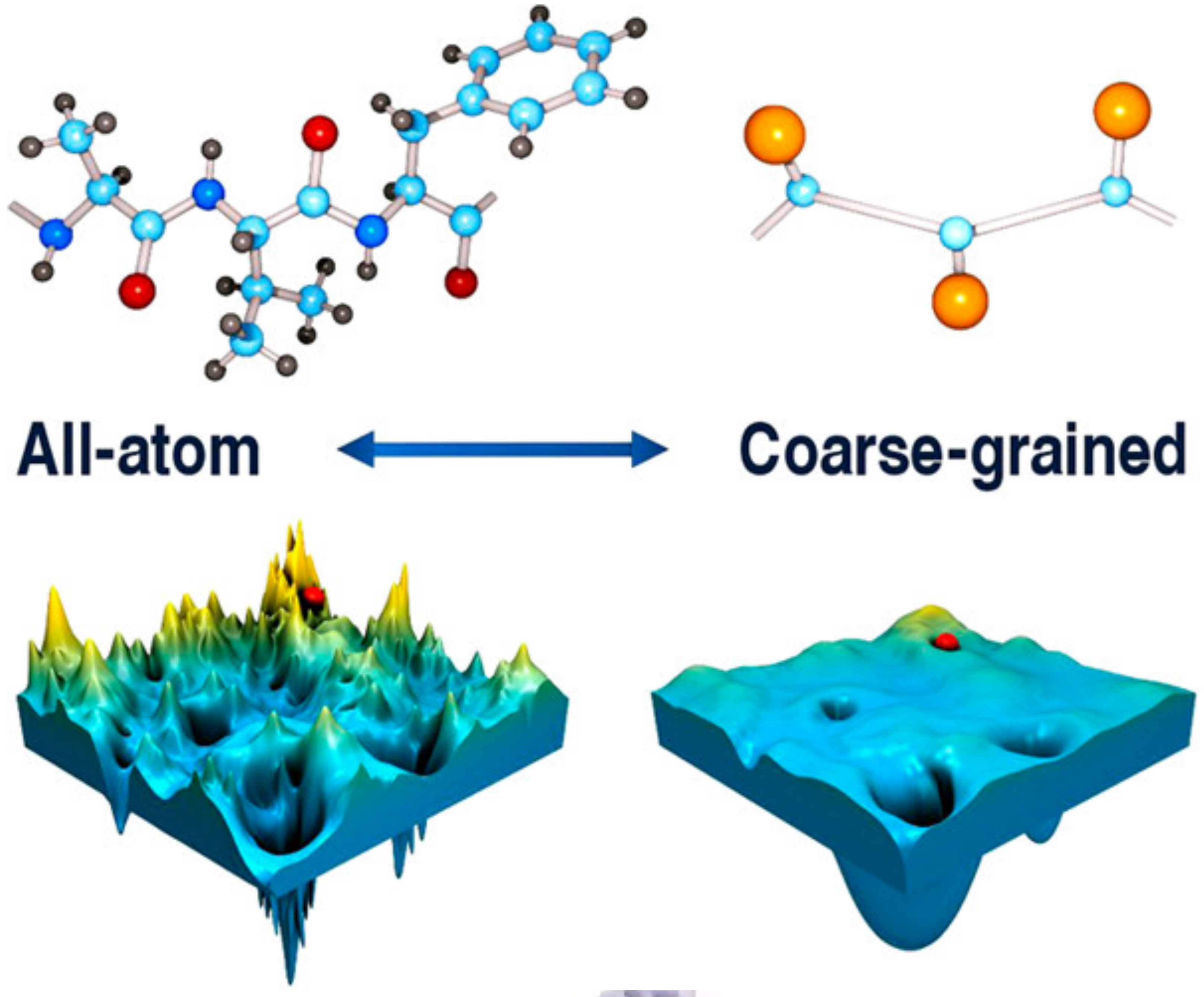
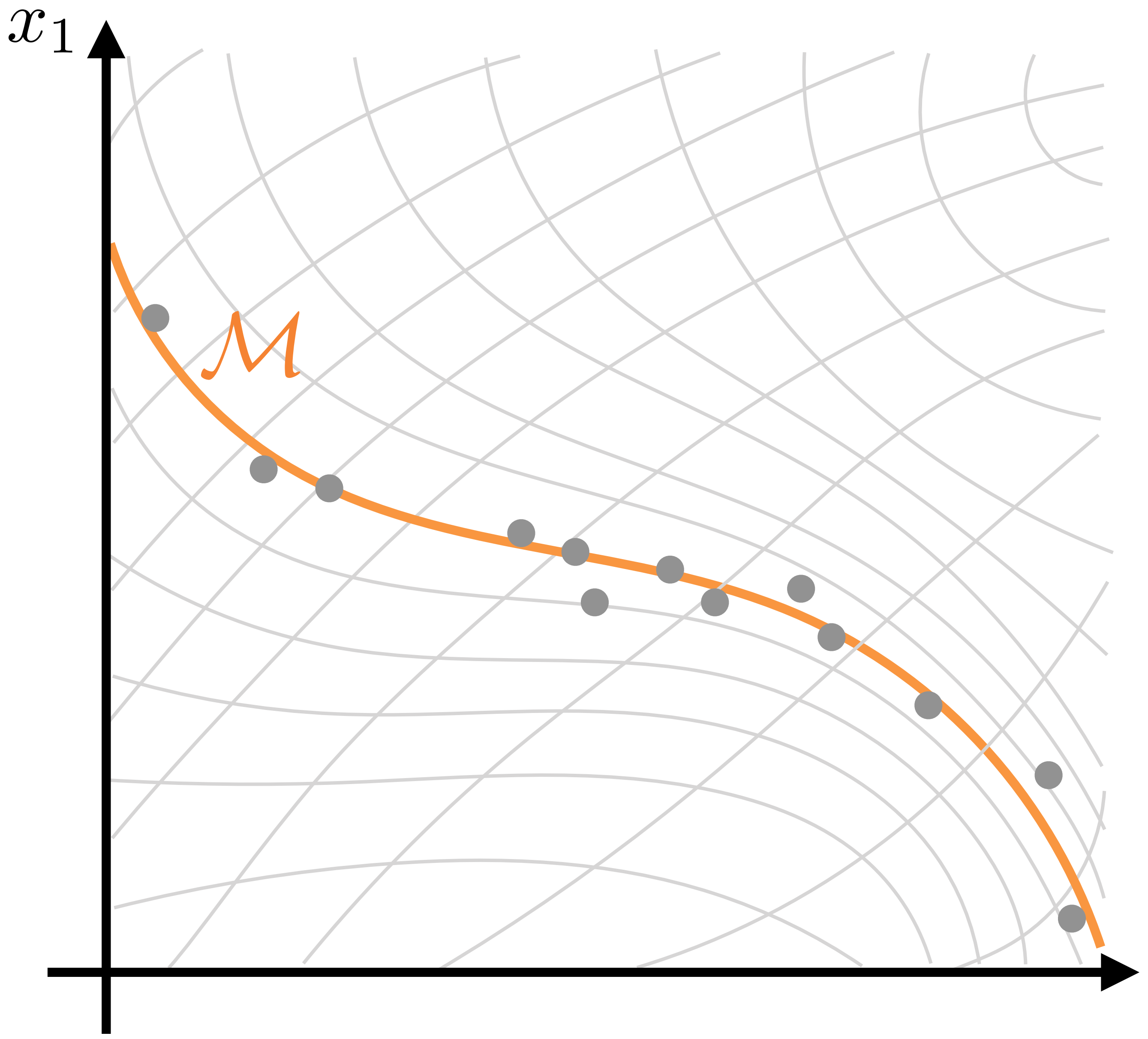
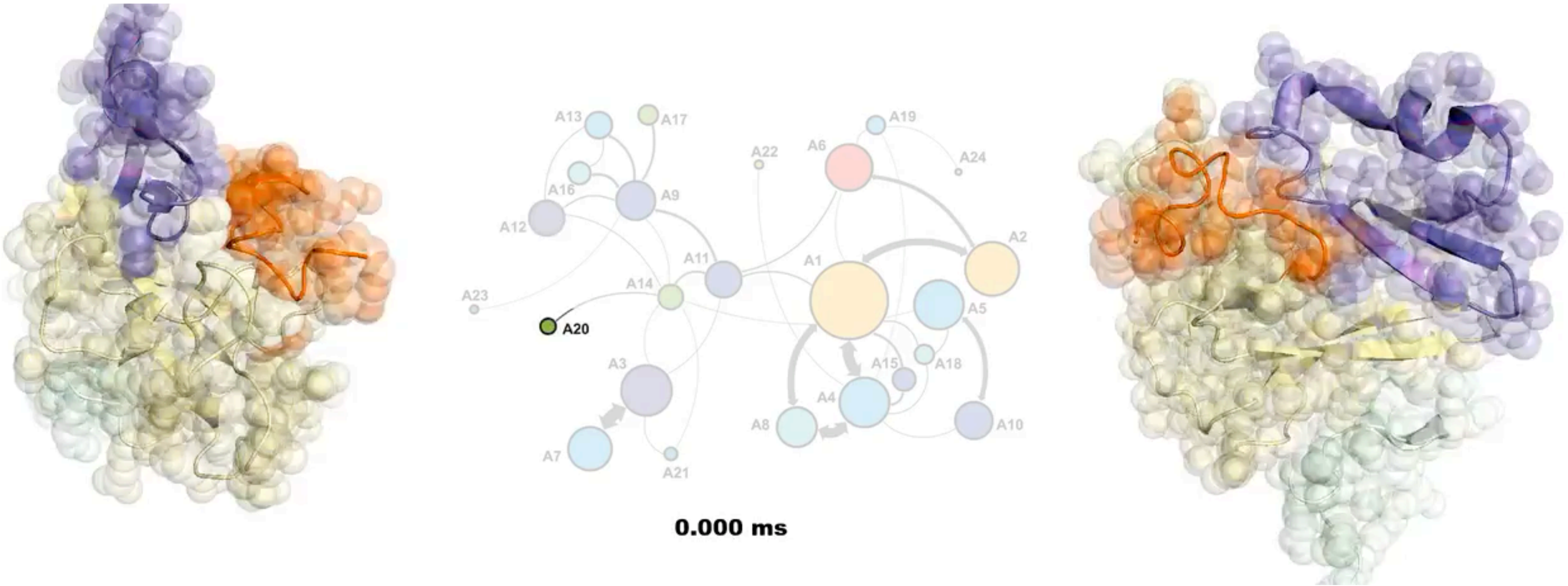


Figure 5. Learned multi-modal density on $SU(2) \equiv \mathbb{S}^3$ using the recursive flow. Each column shows an $\mathbb{S}^2$ slice of the $\mathbb{S}^3$ density

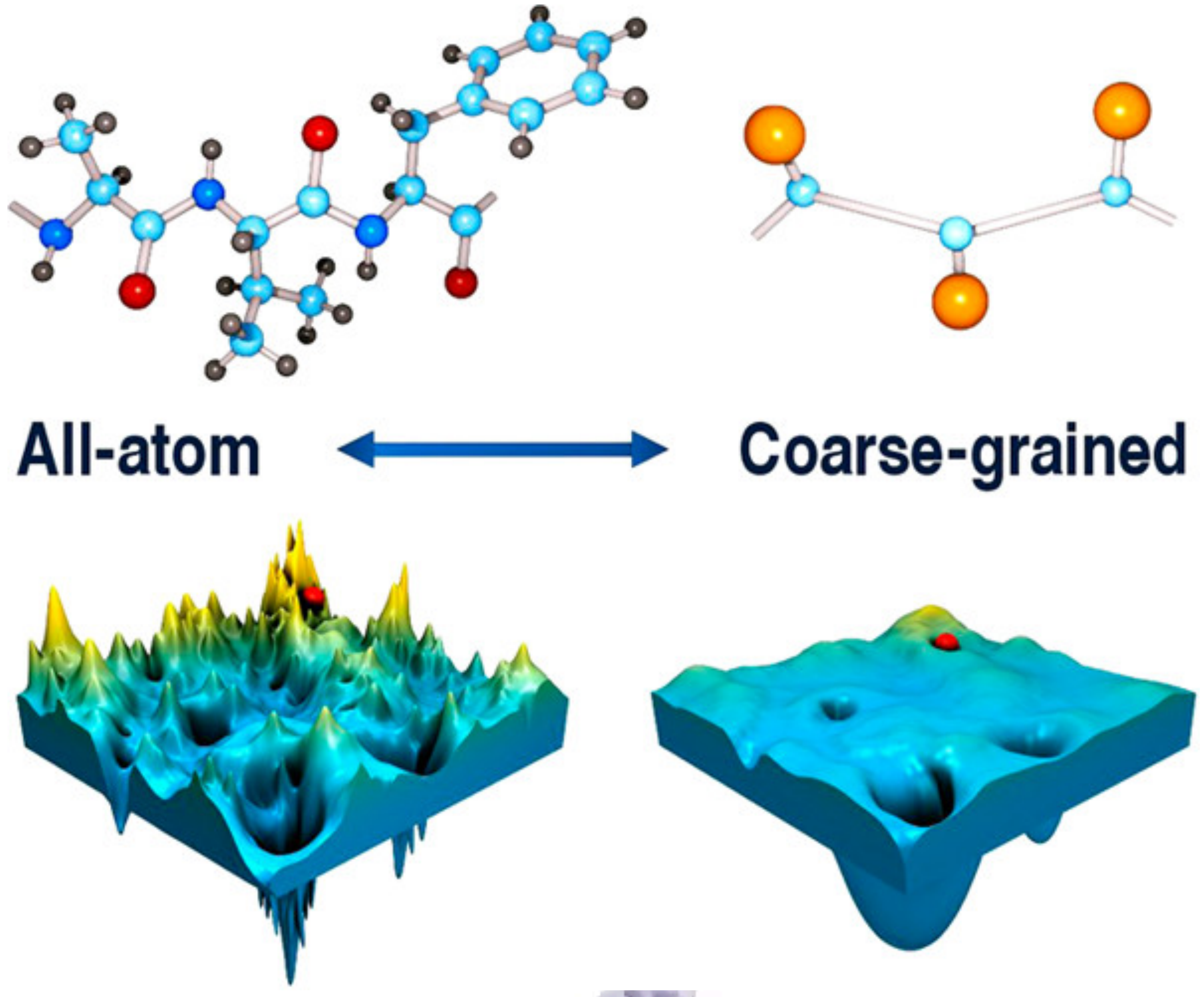


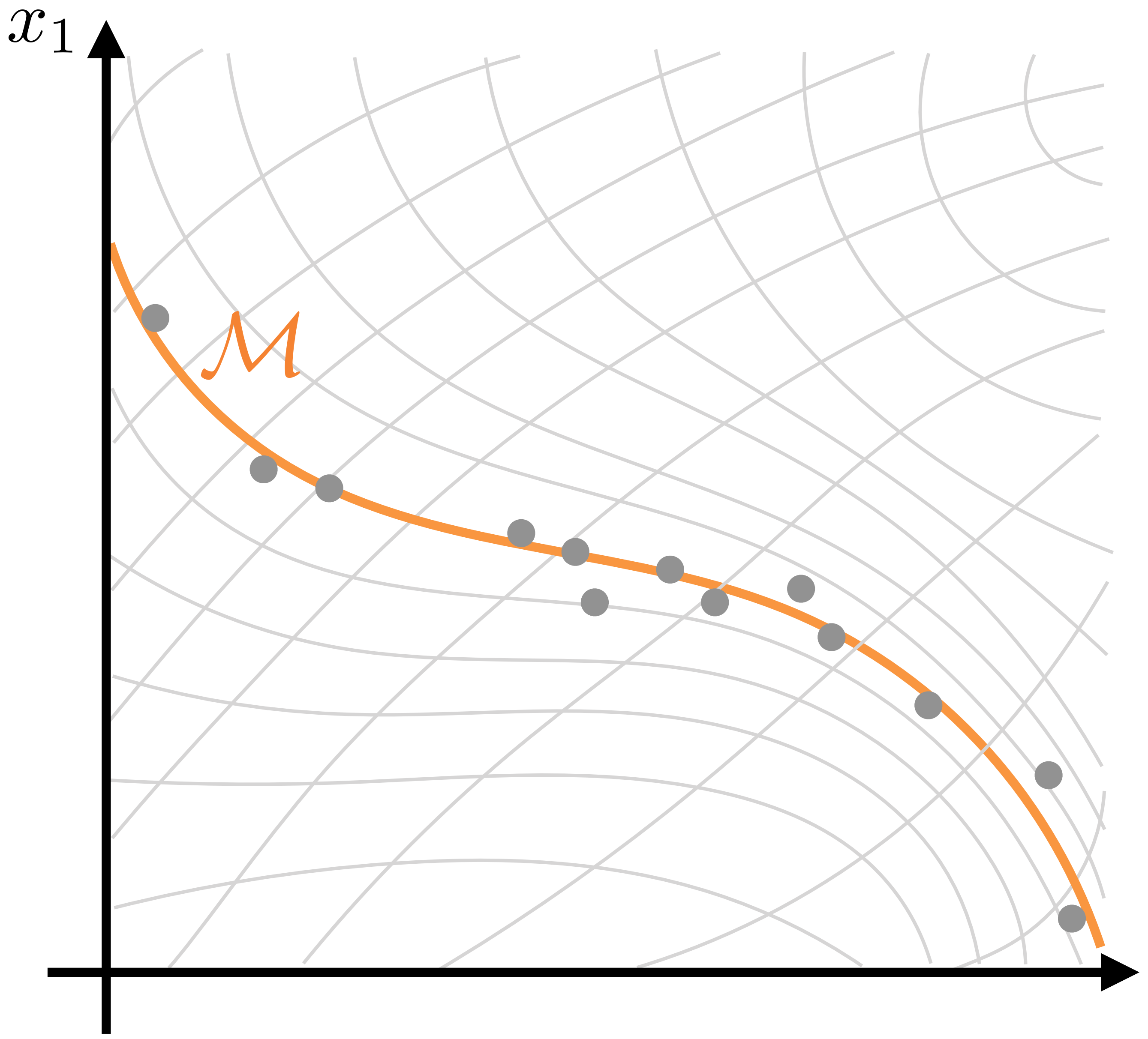


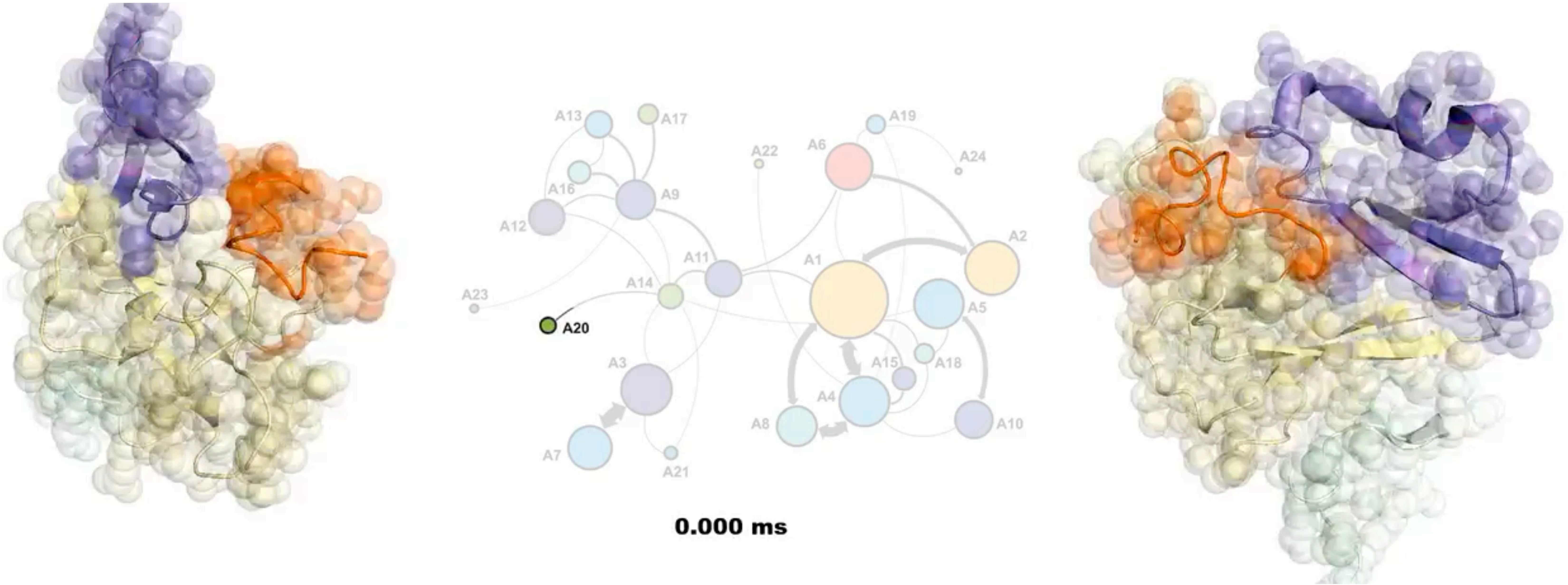




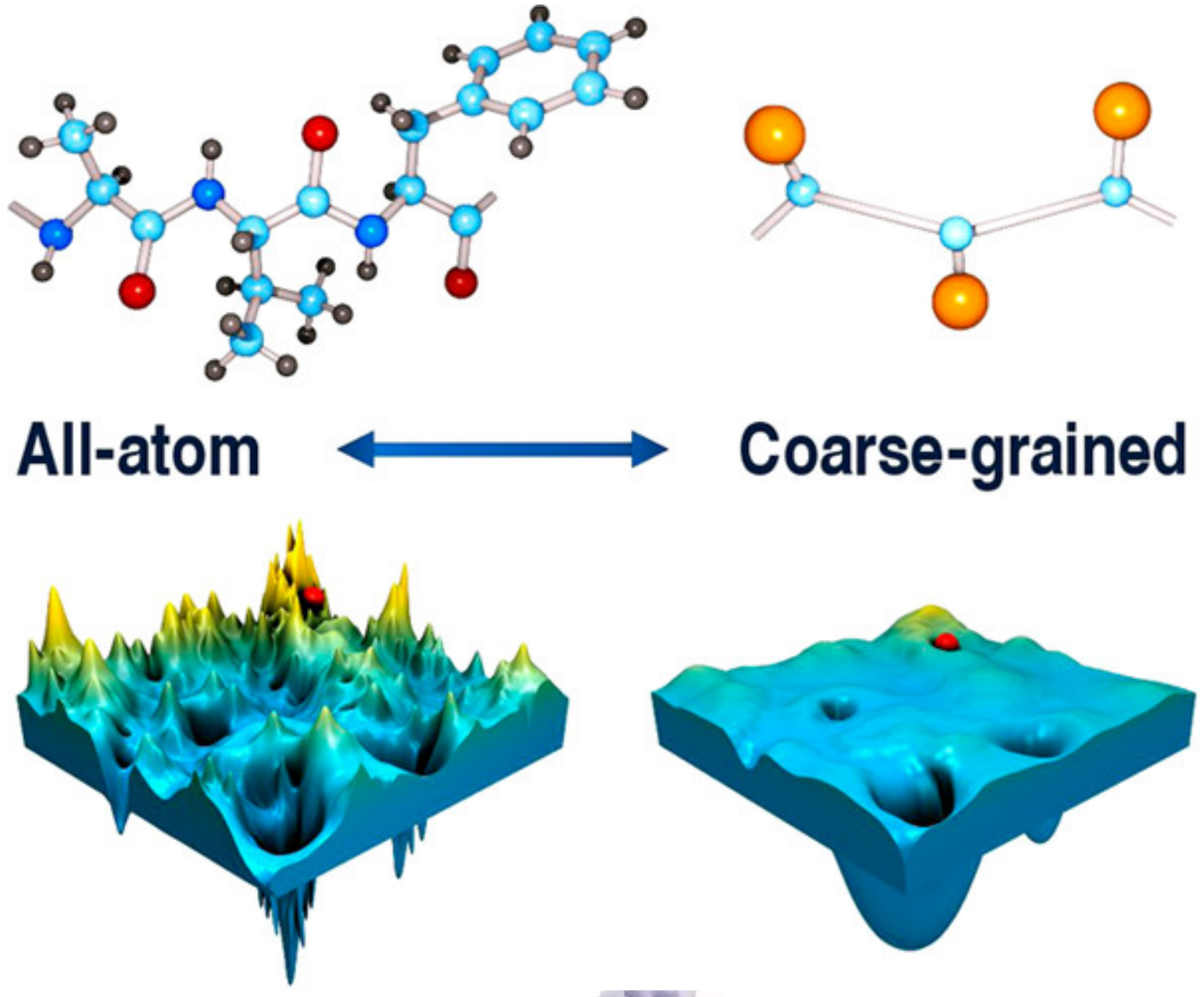
https://doi.org/10.7554/eLife.45403.032

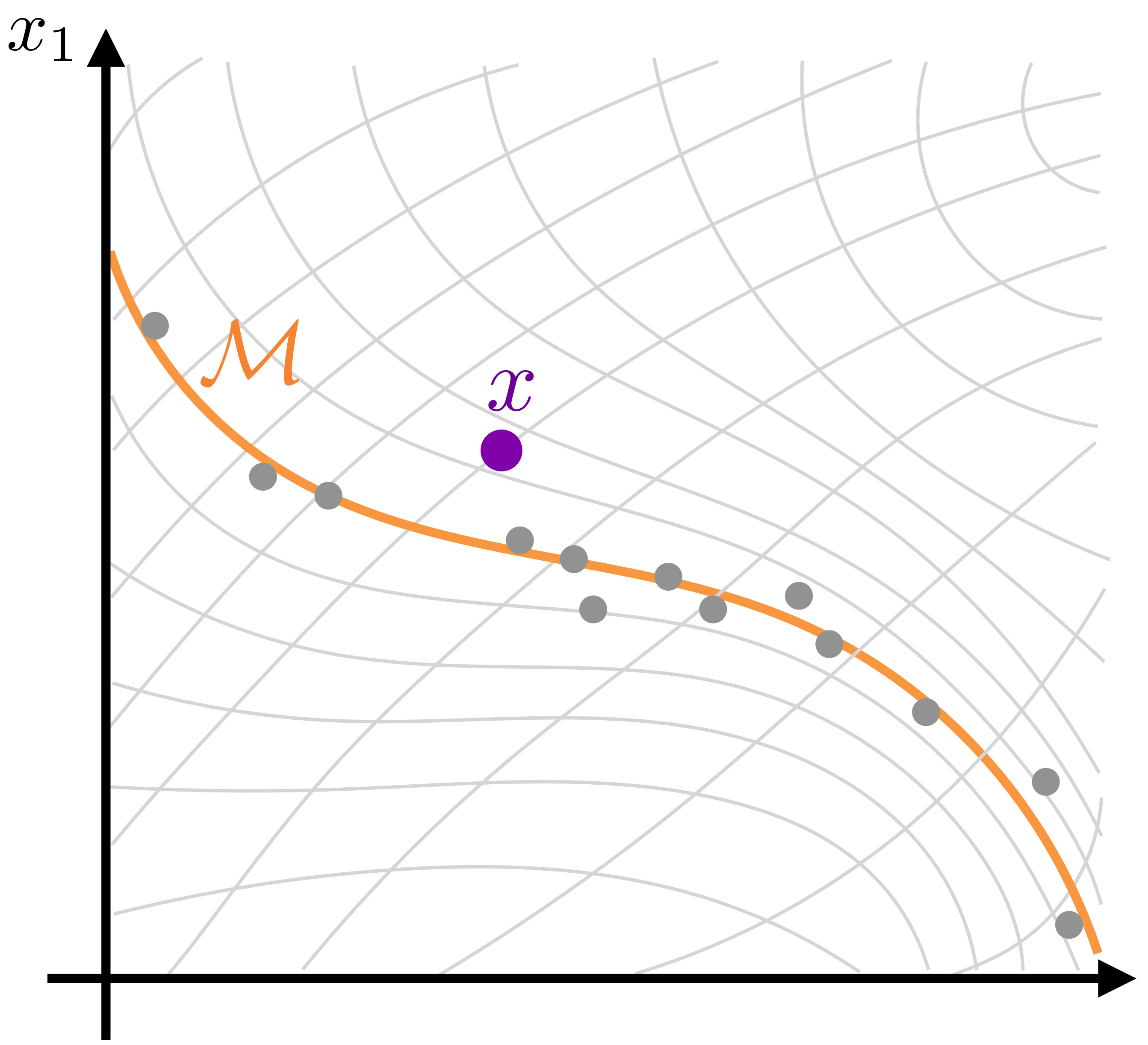


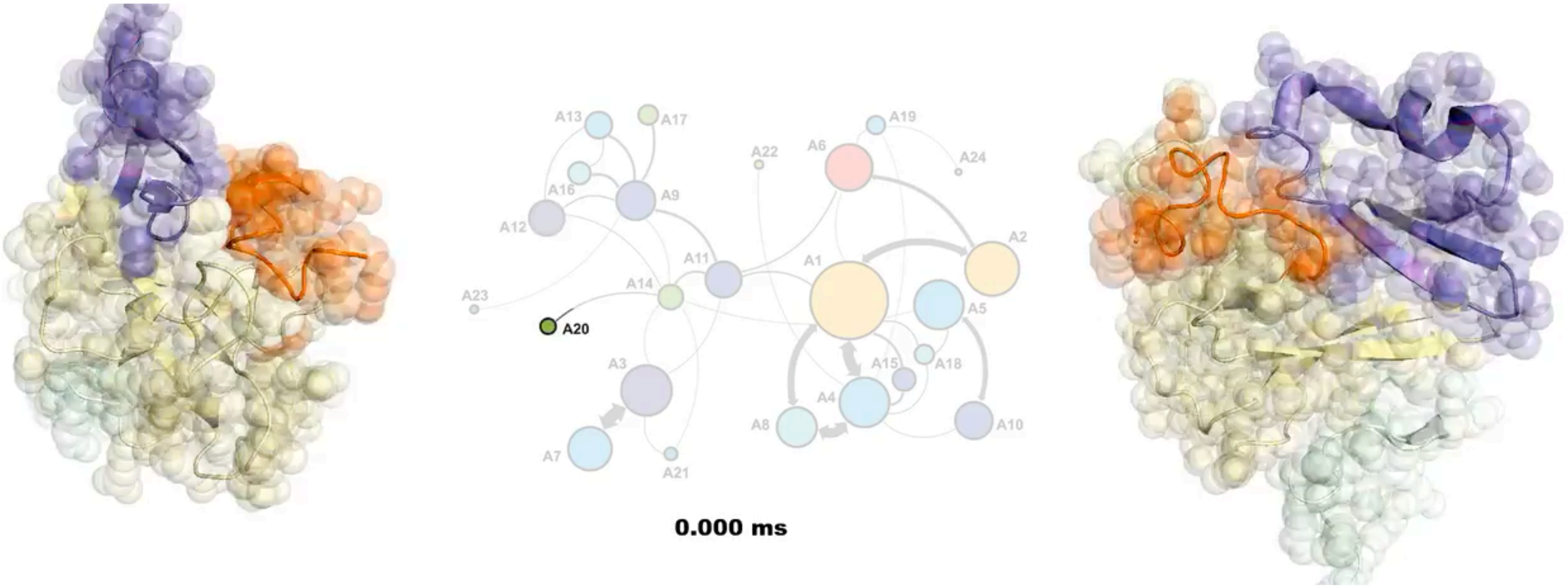




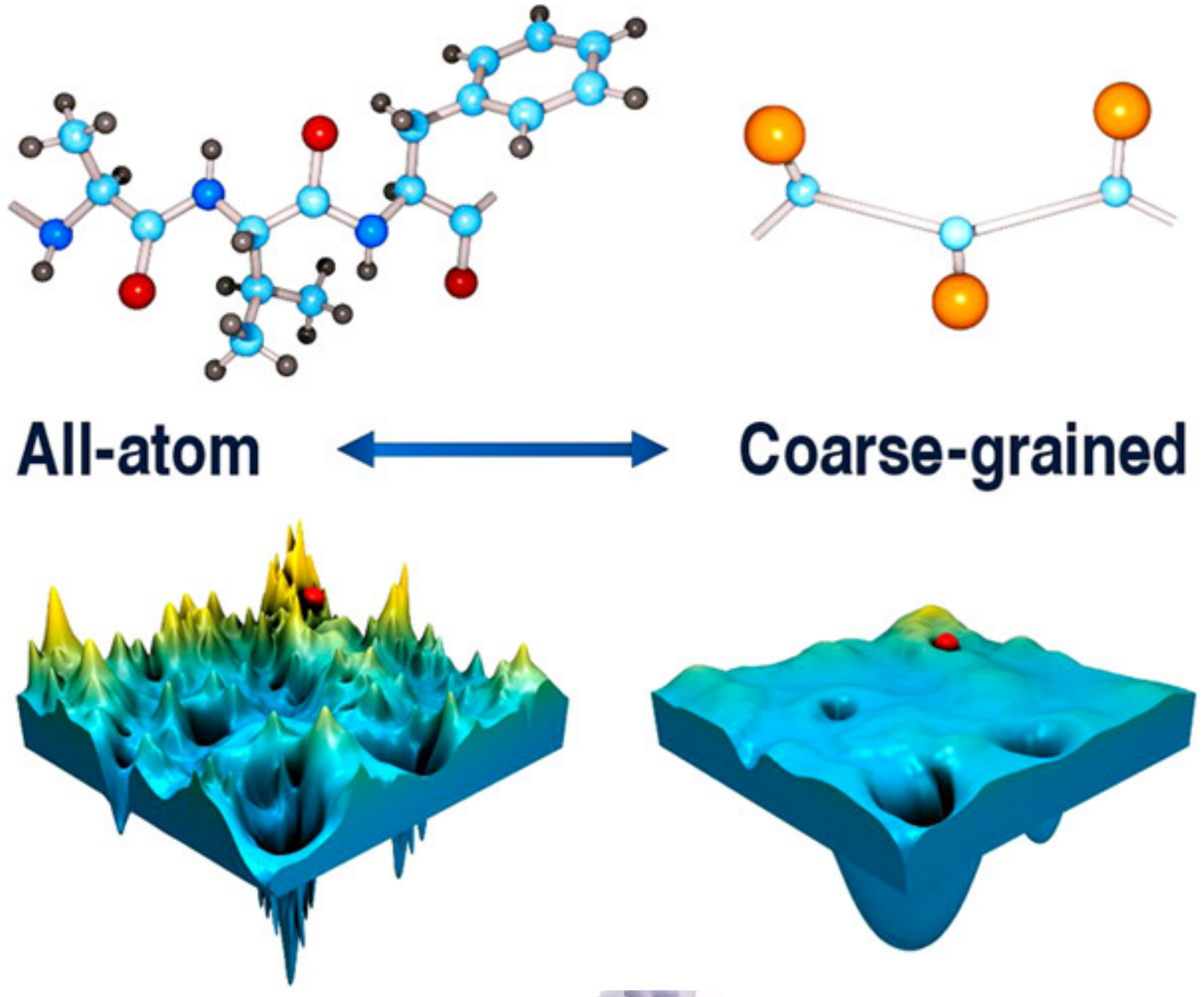
https://doi.org/10.7554/eLife.45403.032

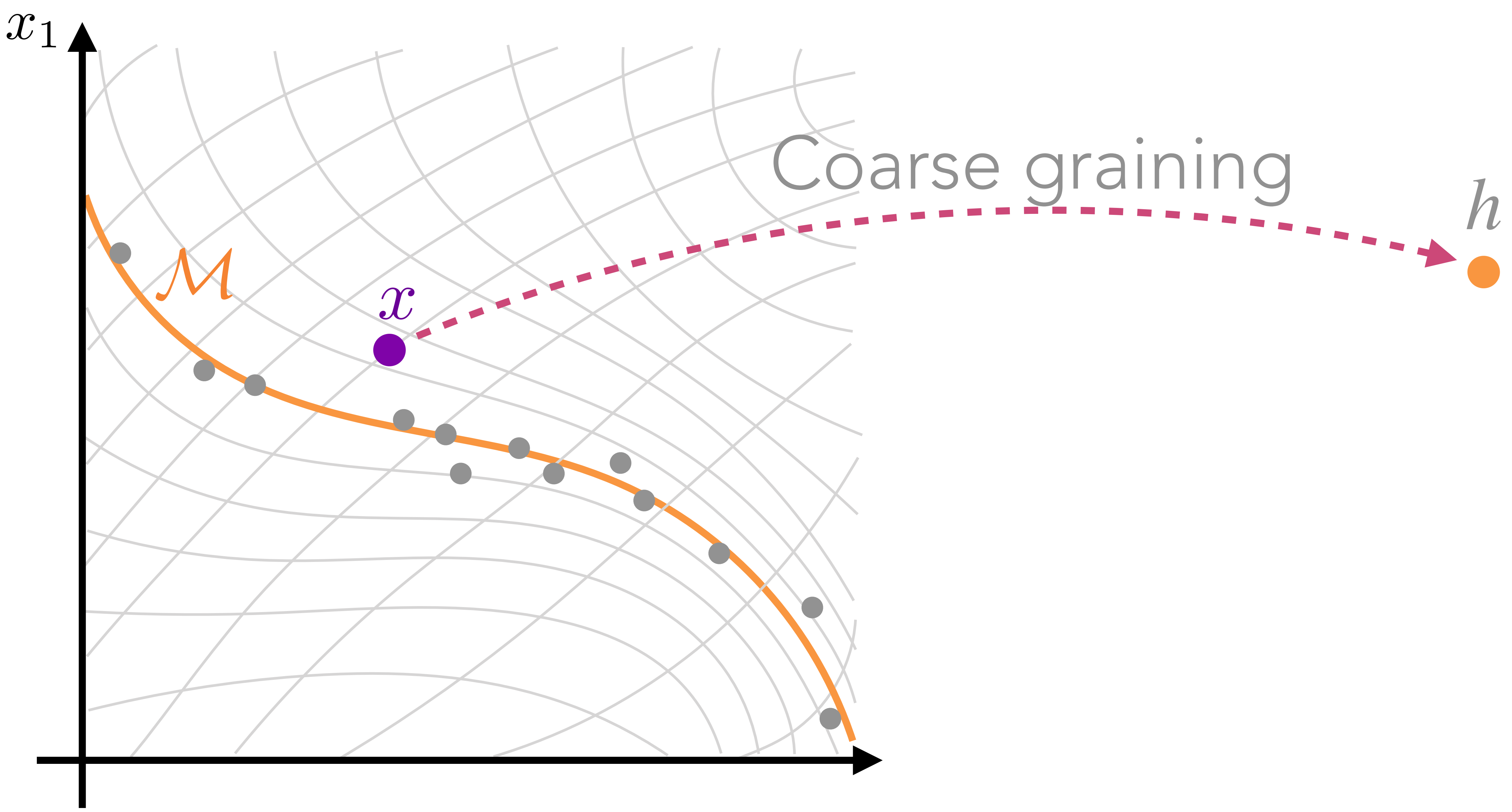


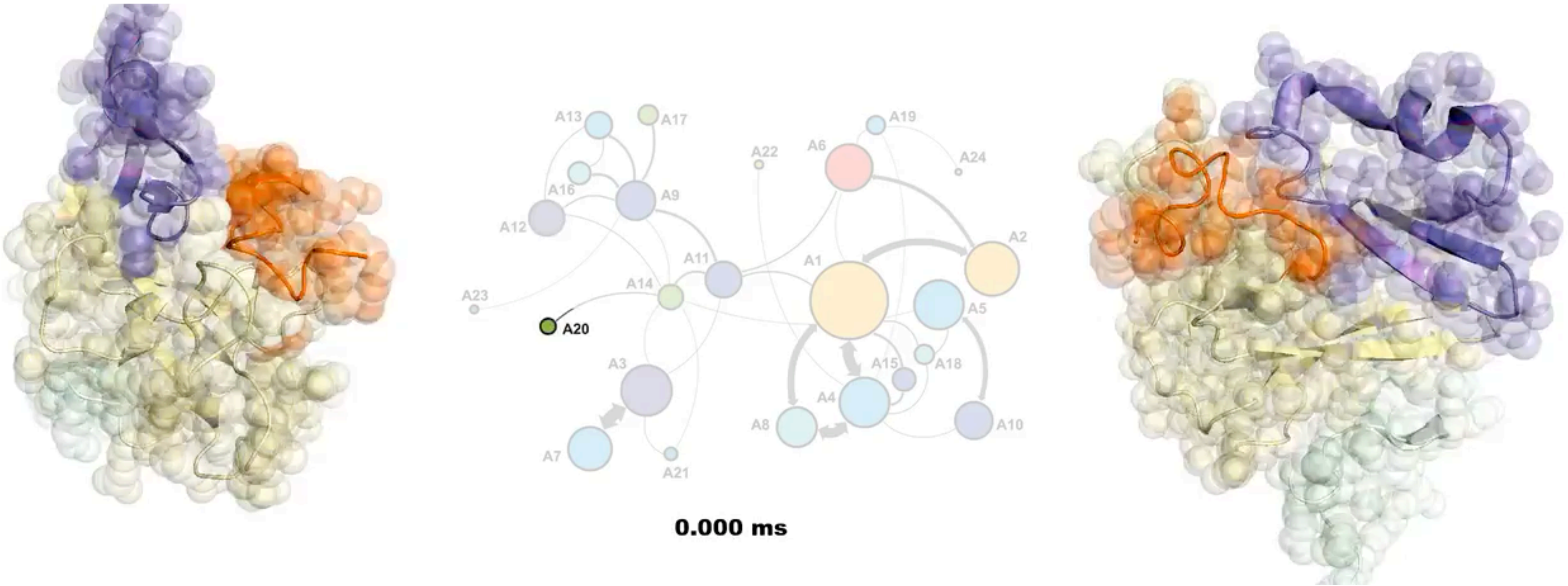




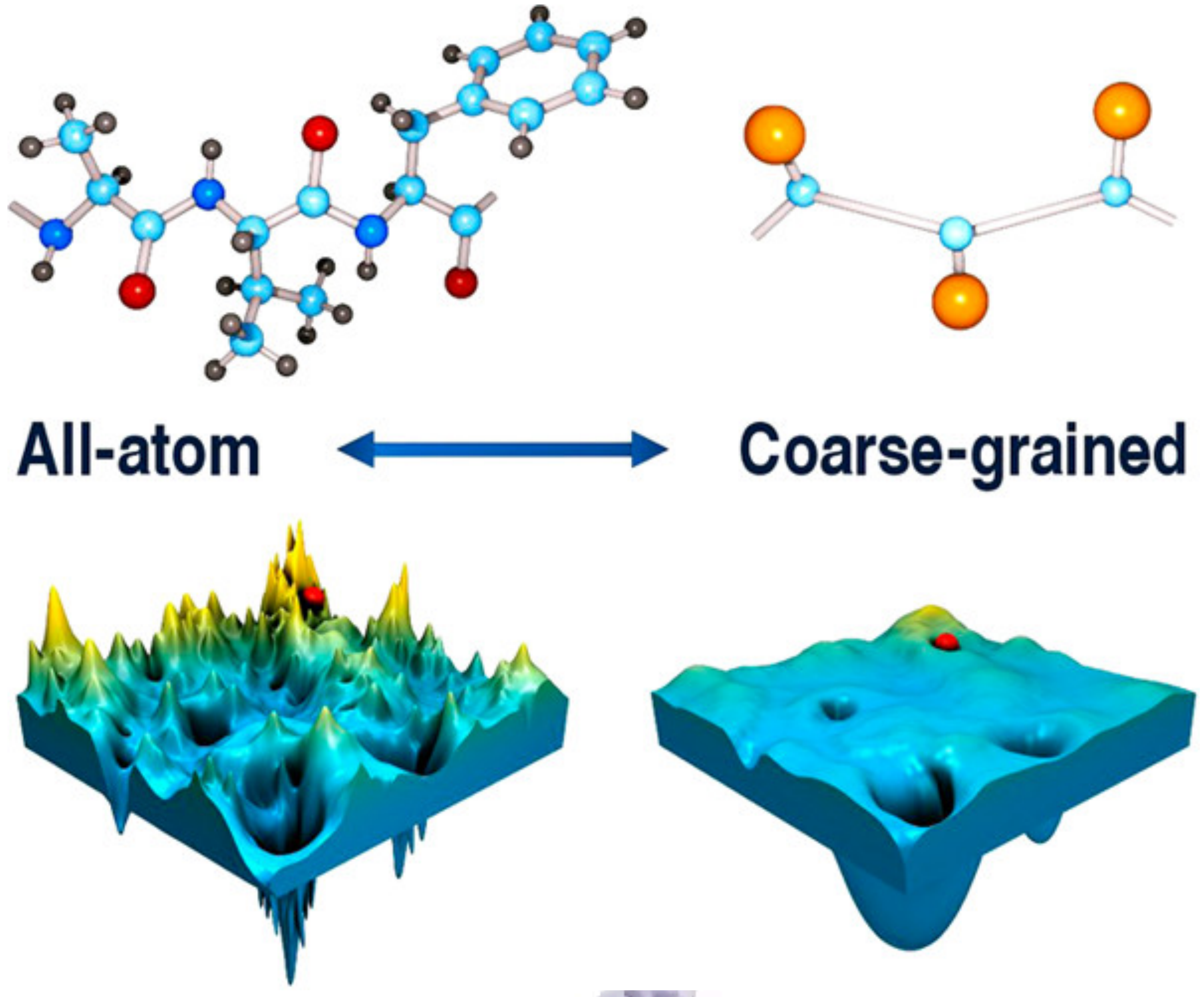
https://doi.org/10.7554/eLife.45403.032

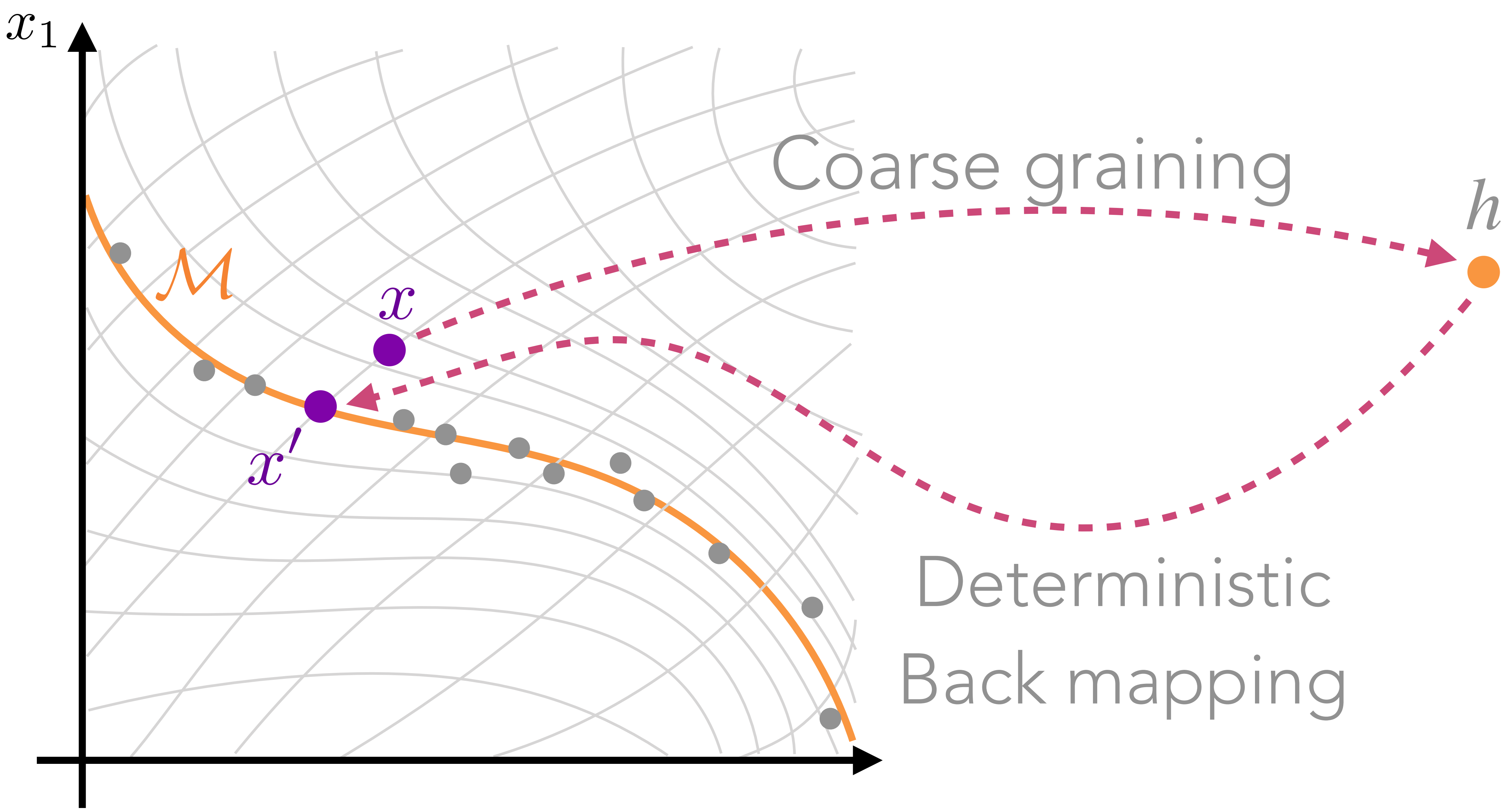


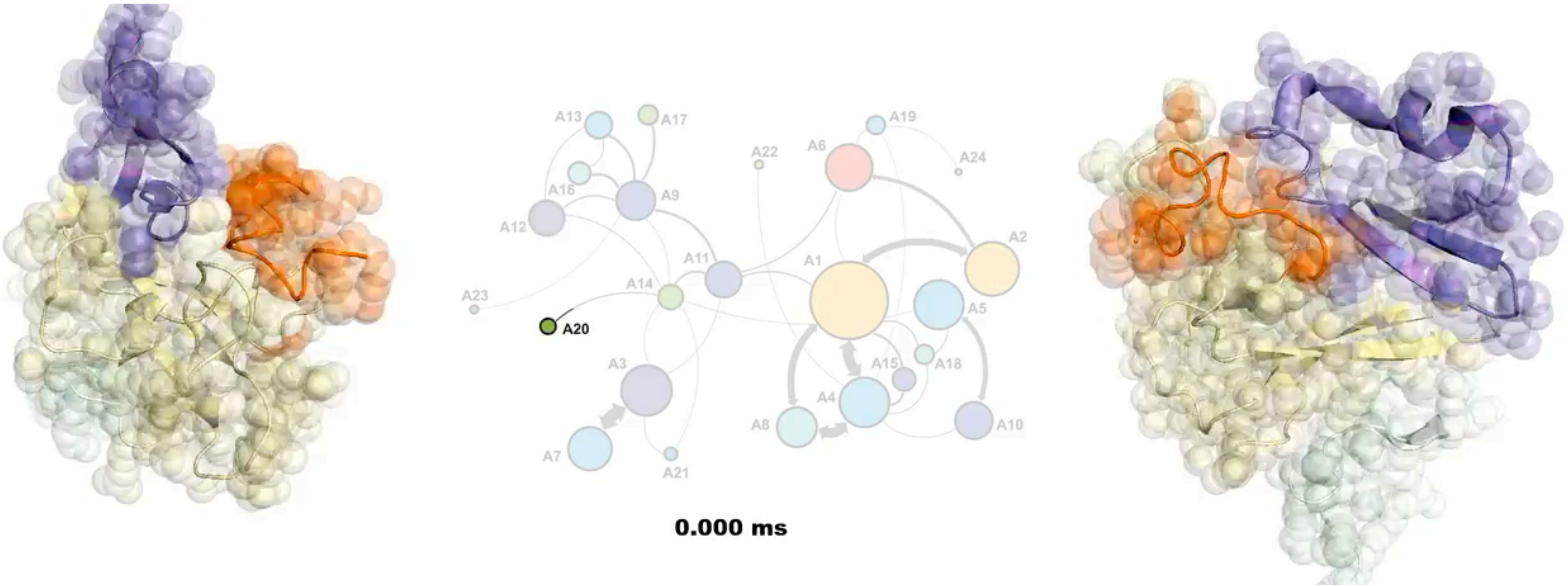




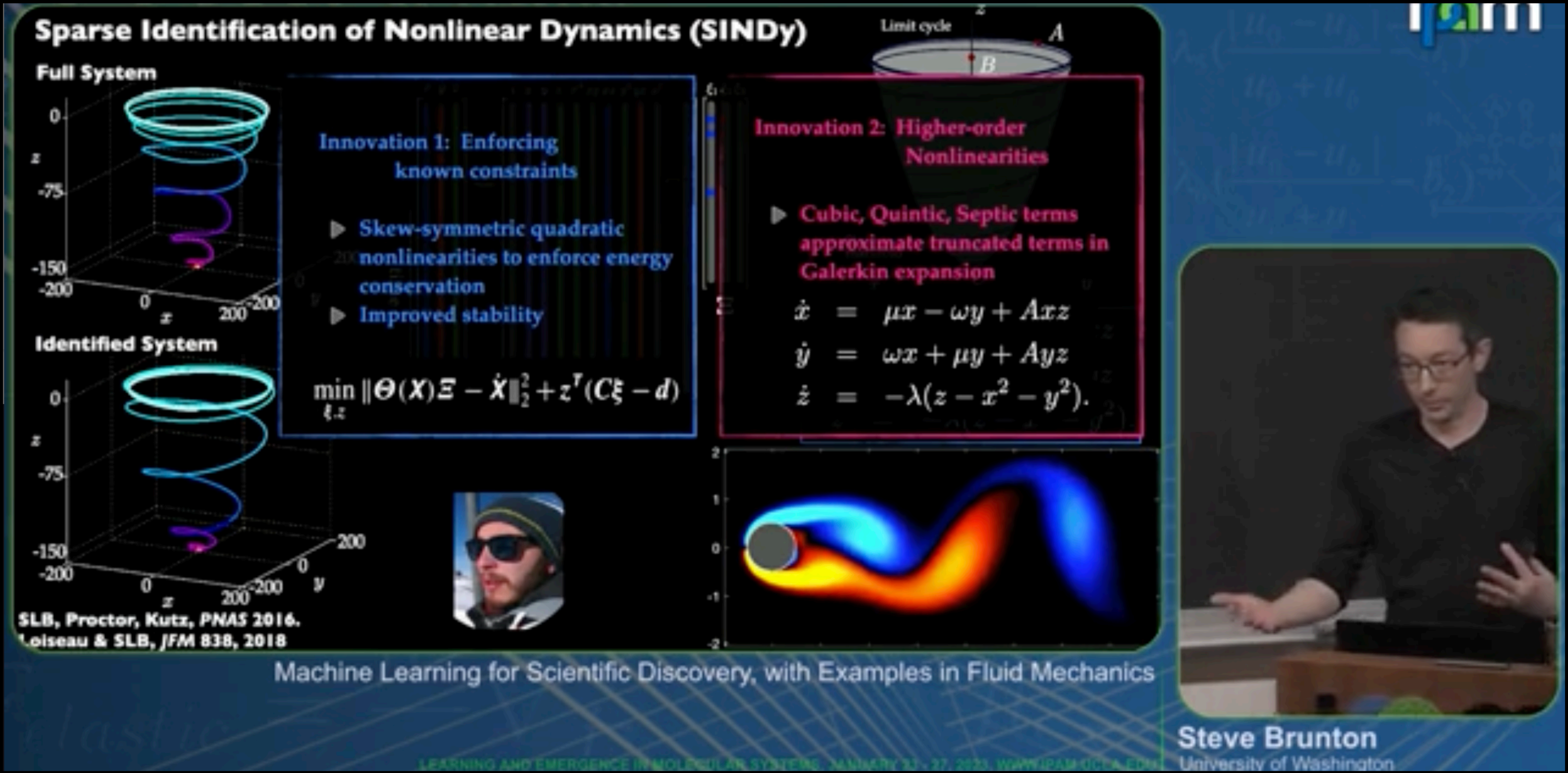
https://doi.org/10.7554/eLife.45403.032







https://doi.org/10.7554/eLife.45403.032


Sparse Identification of Nonlinear Dynamics (SINDy)
Full System
Identified System
Innovation 1: Enforcing known constraints
Skew-symmetric quadratic nonlinearities to enforce energy conservation
Improved stability
min ‖Θ(X)Ξ − Ẋ‖²₂ + zᵀ(Cξ − d)
Innovation 2: Higher-order Nonlinearities
Cubic, Quintic, Septic terms approximate truncated terms in Galerkin expansion
ẋ = μx − ωy + Axz
ẏ = ωx + μy + Ayz
ż = −λ(z − x² − y²).
Limit cycle
SLB, Proctor, Kutz, PNAS 2016.
Loiseau & SLB, JFM 838, 2018
Machine Learning for Scientific Discovery, with Examples in Fluid Mechanics
Steve Brunton
University of Washington

Various strategies

Take advantage of already known multi-scale, emergent phenomena

- Enhanced sampling, coarse graining, …
- Engineered features, inductive bias of models, …

Add the coarse graining by hand and learn the dynamics

- learned force fields, force matching, etc.
- Learn Markov transitions between fixed clustering of states

Add the coarse graining by hand, and learn the effective "dynamics" & how to map back to fine-grained representation

- Steve Brunton's talk: Reduced models, SINDy
- AlphaFold / OpenFold etc. Sequence $\Rightarrow$ structure (not really dynamics)

Simultaneously learn a (latent) coarse-grained representation and "dynamics" & how to map back to fine-grained representation

- VAEs, Diffusion Models, $\mathscr{M}$-flows

Simultaneously learn a coarse-grained representation and dynamics (discovery emergence)

- Learned Koopman operators, learned dynamics of latent space
- SSL techniques like VicREG, Barlow Twins, etc. where encoder, but no decoder.
- Much of ML does this, but interpretability of latent state is a challenge. **When would we call this "emergence"?**
  - Need a way to "operationalize" the latent space representation for some down-stream task

Physical reasoning

Humans are remarkable at being able to have a library of mental models at different levels of abstraction and finding which is most appropriate to use for a given task.

- In my work as a particle physicist I switch between ~5 mental models

Finding the right level of abstraction / coarse graining is key and depends on task

Eventually AI / ML systems may develop causal representations needed to efficiently design experiments, generate hypotheses, etc.

- It may be a foreign ontology, but I suspect that it will need to be causal to be effective