Generative AI and Diffusion Models
a Statistical Physics Perspective

Giulio Biroli

ENS Paris

ellis

European Laboratory for Learning and Intelligent Systems

PR[AI]RIE

Paris School

ENS

Département de Physique

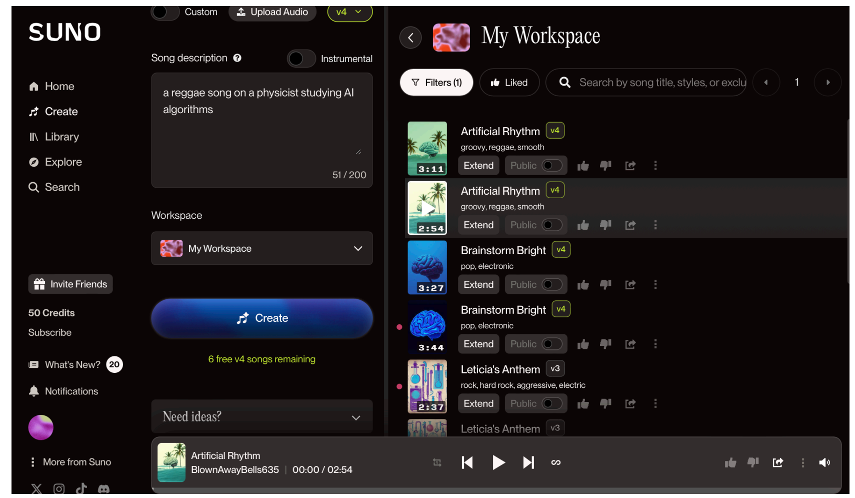
École normale supérieure

Generative AI & Diffusion Models

- Diffusion models are the state of the art in generating image, videos, audio, 3-d scenes

Images

Audios



Videos

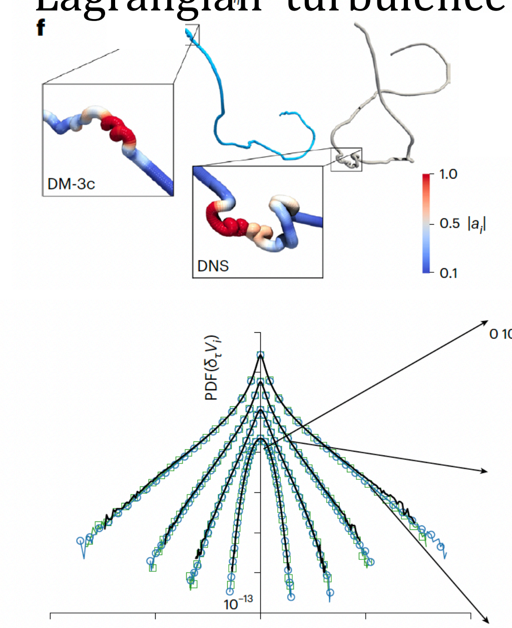
2015 Sohl-Dickstein et al. (physics paper)
2019 Yang & Ermon; 2020, Ho et al, 2021 Song et al…
2021 Dall-E,…

DALL·E 3

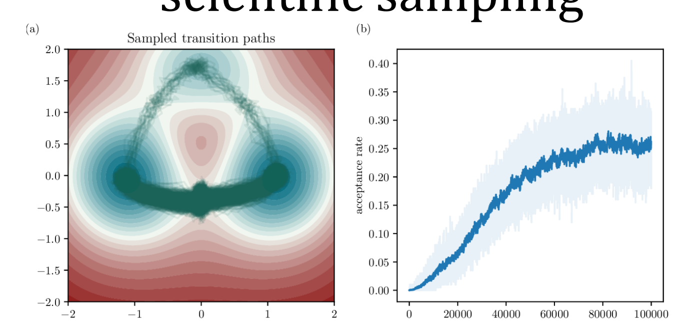
STABLE DIFFUSION

- Application to science generation & sampling

Lagrangian turbulence (Li, Biferale et al 2024)



scientific sampling



-> Eric Vanden Eijnden

Time-reversal and generative AI for images

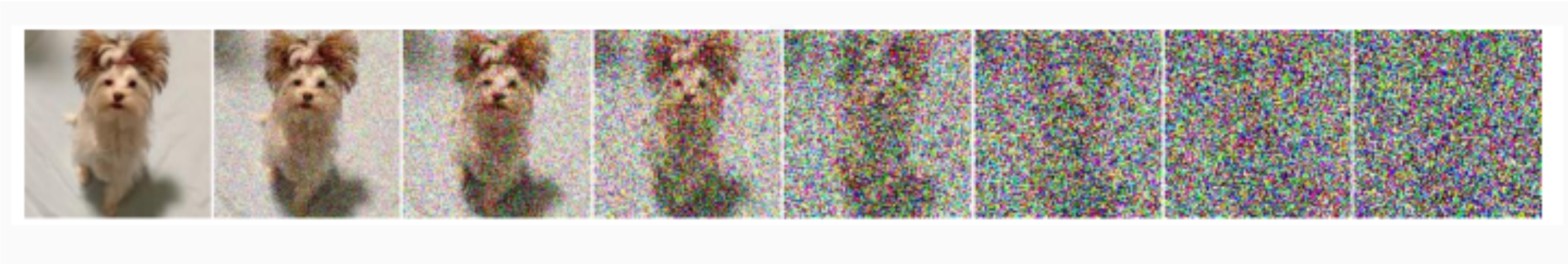
Equilibration

$$\frac{dx_i}{dt} = -\frac{\partial E}{\partial x_i} + \eta_i(t)$$

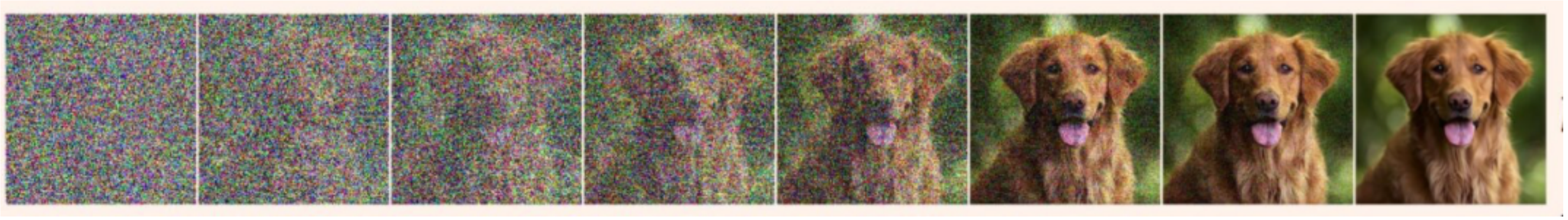
Generative diffusion models go back in time
(denoising from white noise)

Forward in time

$$\frac{dx_i}{dt} = -x_i + \eta_i(t)$$



Backward in time



Recap on Langevin Equation

$$\frac{dx_i}{dt} = -\frac{\partial E}{\partial x_i} + \eta_i(t) \qquad \langle \eta_i(t)\eta_j(t')\rangle = 2T\delta_{i,j}\delta(t-t') \qquad x \in \mathbb{R}^d$$

In math notation $dx_i = -\frac{\partial E}{\partial x_i} dt + \sqrt{2T}\, dB_t^i$ $\qquad \mathbb{E}[\cdot] = \langle \cdot \rangle$

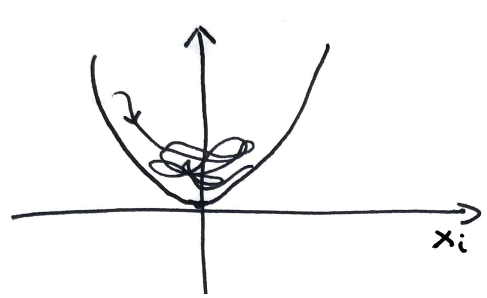
Fokker-Planck Equation & Equilibration

$$\frac{\partial}{\partial t} P(x,t) = \sum_i \frac{\partial}{\partial x_i}\left[\frac{\partial E}{\partial x_i} + T\frac{\partial}{\partial x_i}\right] P(x,t) \qquad P(x,t) \underset{t\to\infty}{\to} P_{GB}(x) = \frac{e^{-E/T}}{Z_T}$$

Forward process: equilibration in a quadratic well

$$\frac{dx_i}{dt} = -x_i + \eta_i(t) \qquad x_i(t) = x_i(0)e^{-t} + \int_0^t e^{-(t-t')}\eta_i(t')dt'$$

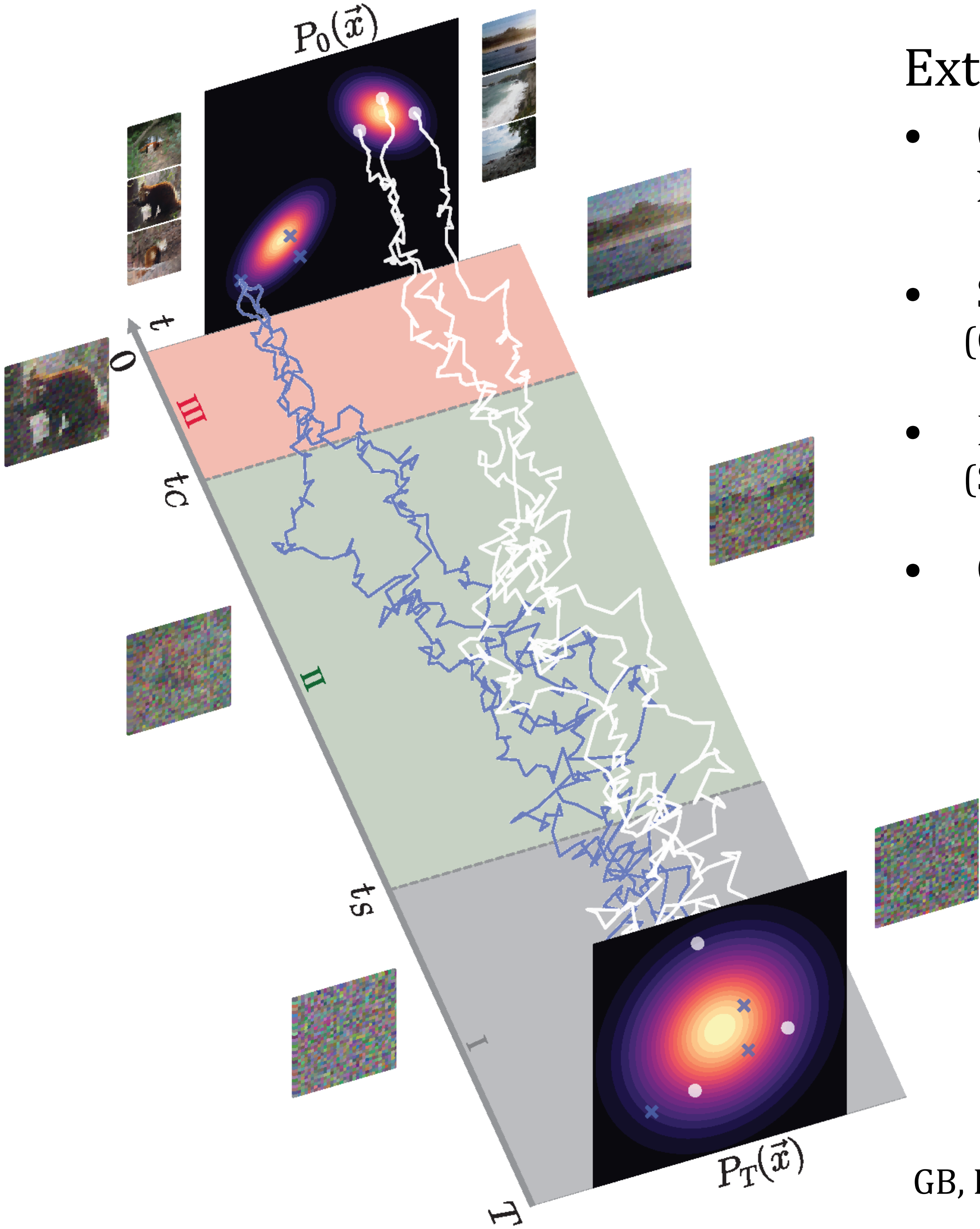
$T = 1 \qquad E = \sum_i \frac{x_i^2}{2}$



$$x_i(t) \overset{d}{=} x_i(0)e^{-t} + \sqrt{1-e^{-2t}}\, g_t^i \qquad g_t^i \sim \mathcal{N}(0,1)$$

$$P(x,t) \to P_{GB}(x) = \prod_{i=1}^{d} \frac{e^{-x_i^2/2}}{\sqrt{2\pi}}$$

Formation on main classes/features of the data from pure noise



Extensions of the theory

- GMs on low dimensional manifolds and latent spaces (Achilli et al. 2025, George et al 2025)
- StatMech Models: Curie-Weiss and 1D Ising (GB and Mezard 2023, Achilli to appear; Guth and Bruna to appear)
- Hierarchical Models (Sclocchi et 2024, Pavasovic et al 2025)
- General large-t expansion: $t_S = \frac{1}{2}\log\Lambda$ (GB et al 2024)

  Λ Largest principal component of the correlation matrix of the data

GB, Bonnaire, De Bortoli, Mézard Nature Commun. 2024

Tests in Real Images

Model

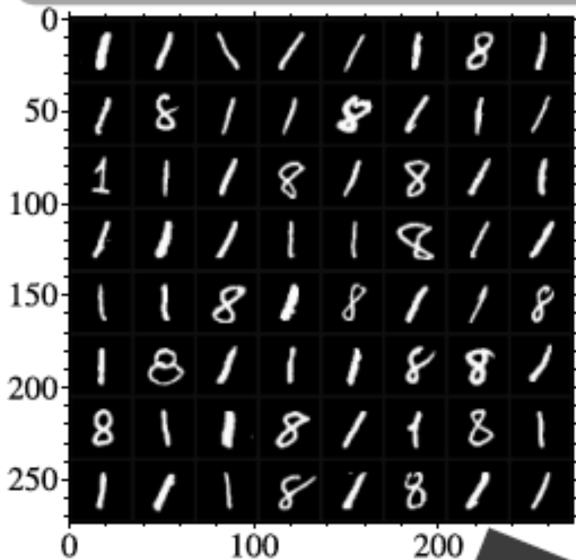
Similar model of Ho et al 2020
U-Net, 4 resolution maps
with 2 convolutional blocks
Dropout rate 0.1
25.7 millions parameters

Training

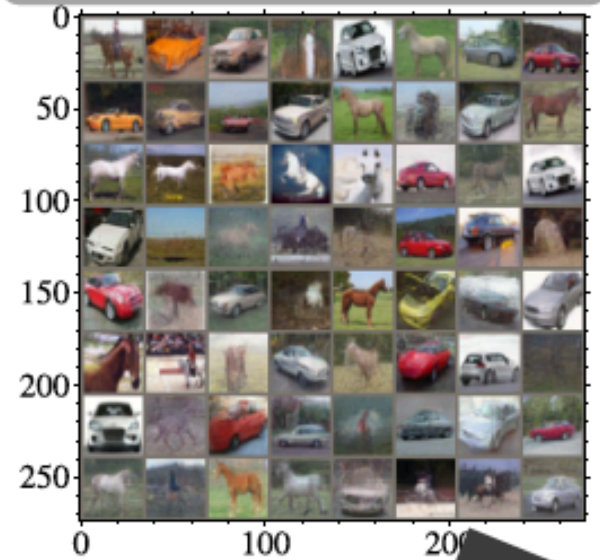
Adam optimizer
LR 10^{-4}
Multiplied by 0.98 every 50 epochs

Imagenet16 — 500k steps
- 2000 samples
- L. pandas and seashores
- $N = 16\times16\times3 = 768$

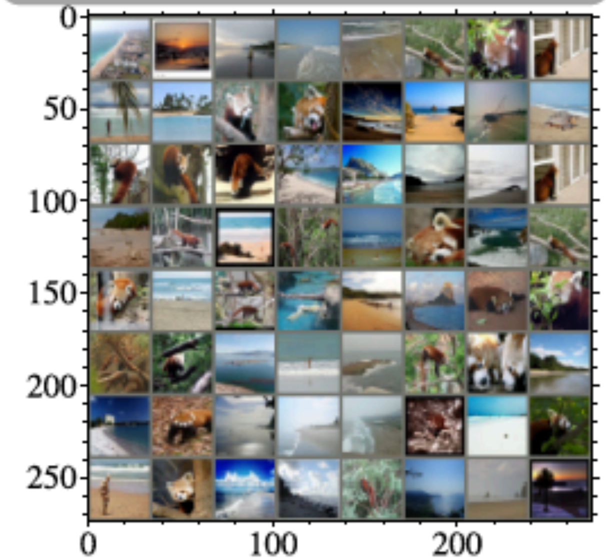
MNIST32 — 100k steps
- 10000 samples
- Classes 1 and 8
- $N = 32\times32\times1 = 1024$



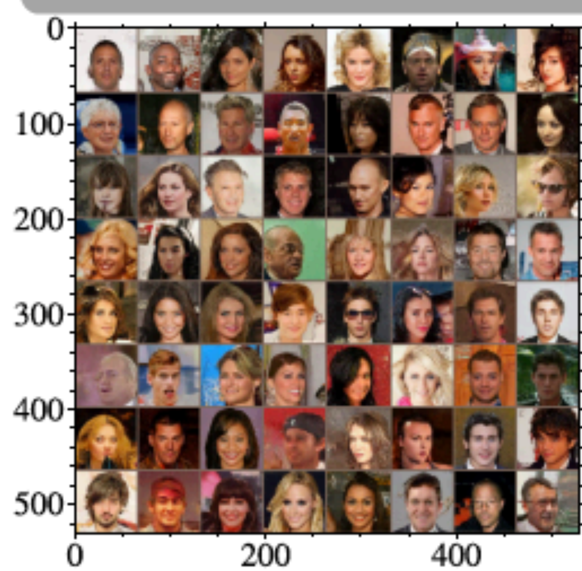
CIFAR2 — 100k steps
- 6000 samples
- Classes horses and cars
- $N = 32\times32\times3 = 3072$



Imagenet32 — 500k steps
- 2000 samples
- L. pandas and seashores
- $N = 32\times32\times3 = 3072$



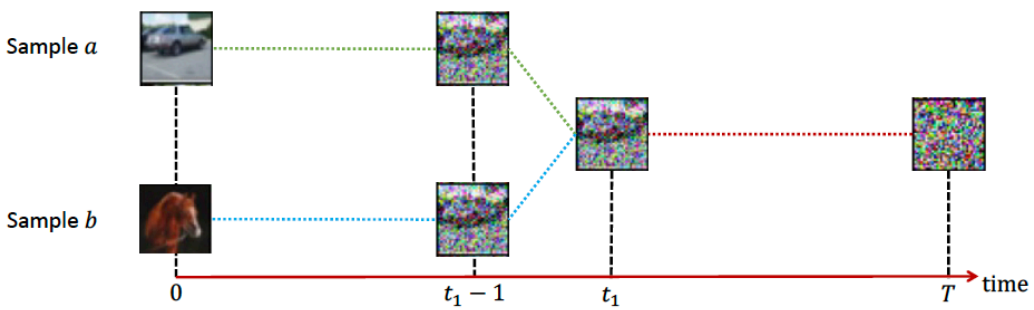
CelebA64 — 130k steps
- 40000 samples
- Classes males and females
- $N = 64\times64\times3 = 12288$



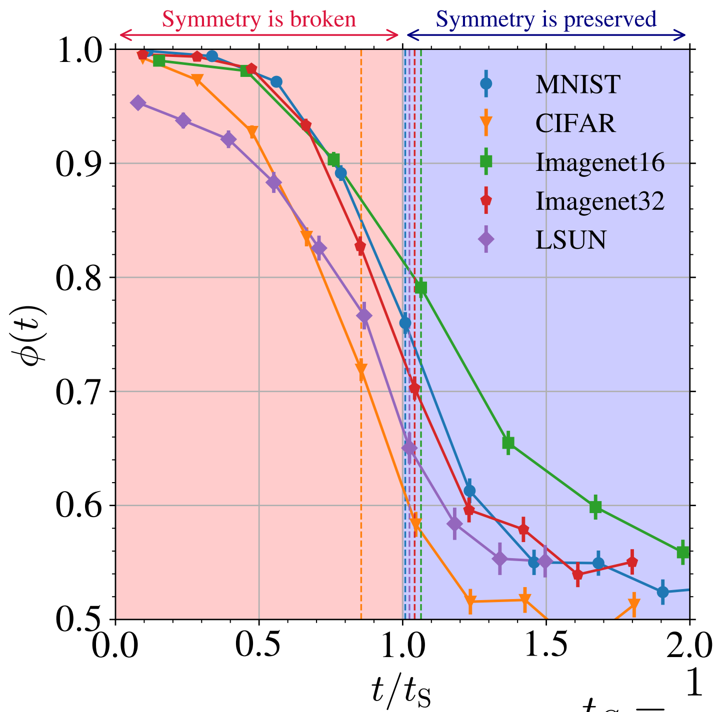
LSUN64 — 310k steps
- 40000 samples
- Conference and churches
- $N = 64\times64\times3 = 12288$

Speciation Transition in Real Images

Cloning experiment



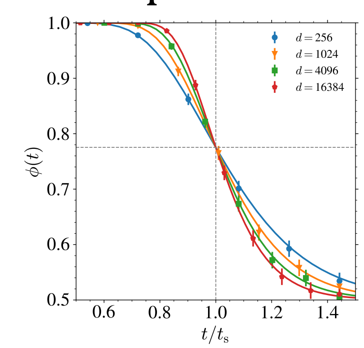
Numerical experiments



$$t_S = \frac{1}{2} \log \Lambda$$

Probability 2 clones in the same class

Analytical result for simple models



Confirm the speciation phenomenon & good estimation of the speciation time

Observed numerically in U-turns experiments

Behjoo et al 2023
Kadkhodale et al 2023
Schlocchi et al 2024

Dynamical regimes relevant for applications -> conditional diffusion & classifier free guidance
(Kynkäänniemi et al 2024 (NVIDIA), Pavasovic et al 2025)

Why Diffusion Models Don't Memorize?

Memorization-Generalization Transition

Memorisation vs Generalisation

Relevant for theory and practice
(copyright problems and differential privacy)

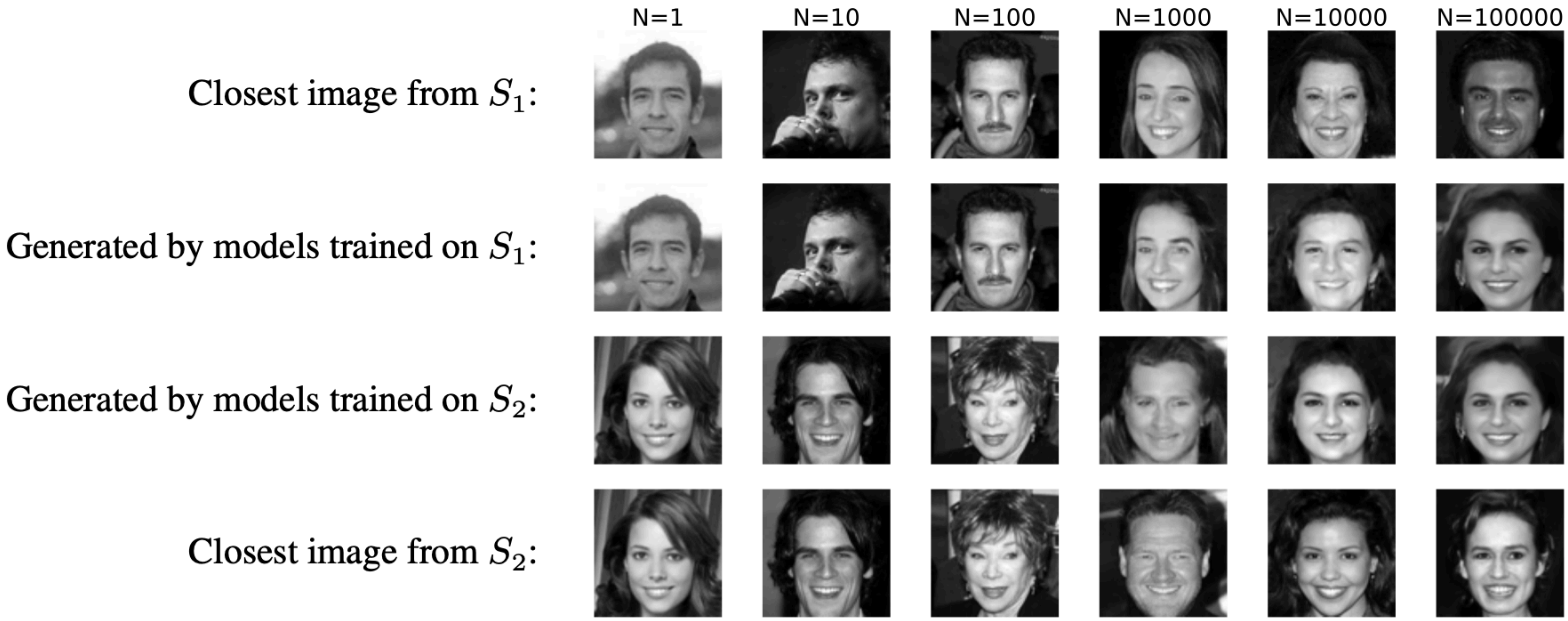
n=200 n=40000

Bonnaire, et al Nat. Comm. 2024

See also Kadkhodale, Guth, Simoncelli, Mallat 2023: experiment with two models on two training sets

Generalisation for large enough training sets
(generation of new images and independence on the training set)

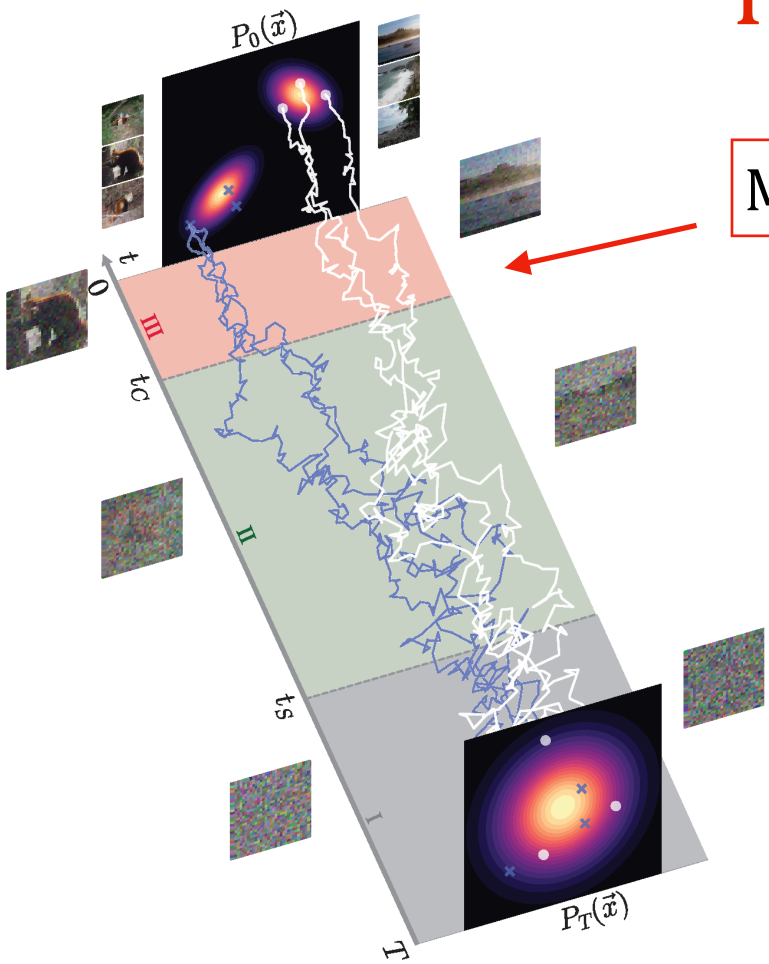
Memorisation vs Generalisation



Kadkhodale, Guth, Simoncelli, Mallat 2023

Generalisation for large enough training sets
(generation of new images and independence on the training set)

Memorization vs Generalization for a "perfect machine"



Memorization

$$\mathcal{L}_{emp} = \frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_{noise}\left(S^{\theta_t}(x_\nu) + \frac{x_\nu - a_\nu e^{-t}}{1-e^{-2t}}\right)^2$$

Global minimum: $S^{emp}(x,t) = \nabla \log\left(\frac{1}{n}\sum_{\nu=1}^{n}\frac{e^{-\frac{(x-x^{\nu}e^{-t})^2}{2\Delta_t}}}{(2\pi\Delta_t)^{d/2}}\right)$

Curse of dimensionality

(Exponential number of data to decrease the memorization phase)

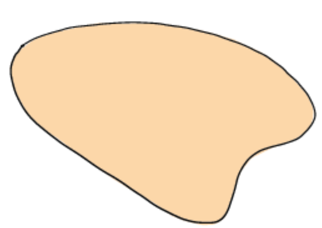
$n = e^{\alpha d}\ ;\ t_C = f(\alpha)$ $\qquad t_C \to 0$ for $\alpha \to \infty$

(Mapping to disordered systems)

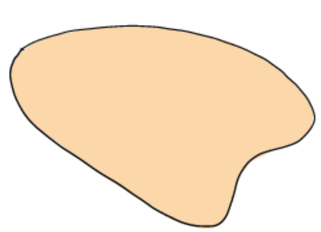
A perfect and perfectly trained machine would lead to memorization

Regime II

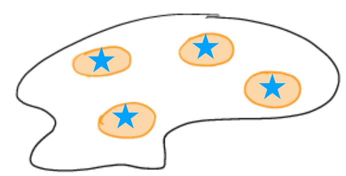
$P_{t_1}(\vec{x})$



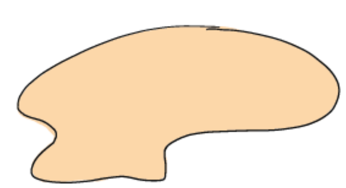
$P^{true}_{t_1}(\vec{x})$



Regime III



$$P_{t_2}(\vec{x}) = \frac{1}{n}\sum_{\nu=1}^{n}\frac{e^{-\frac{(x-x^{\nu}e^{-t})^2}{2\Delta_t}}}{(2\pi\Gamma_t)^{d/2}}$$



$$P^{true}_{t_2}(\vec{x}) = \int P_0(x^{\nu})\frac{e^{-\frac{(x-x^{\nu}e^{-t})^2}{2\Delta_t}}}{(2\pi\Gamma_t)^{d/2}}dx^{\nu}$$

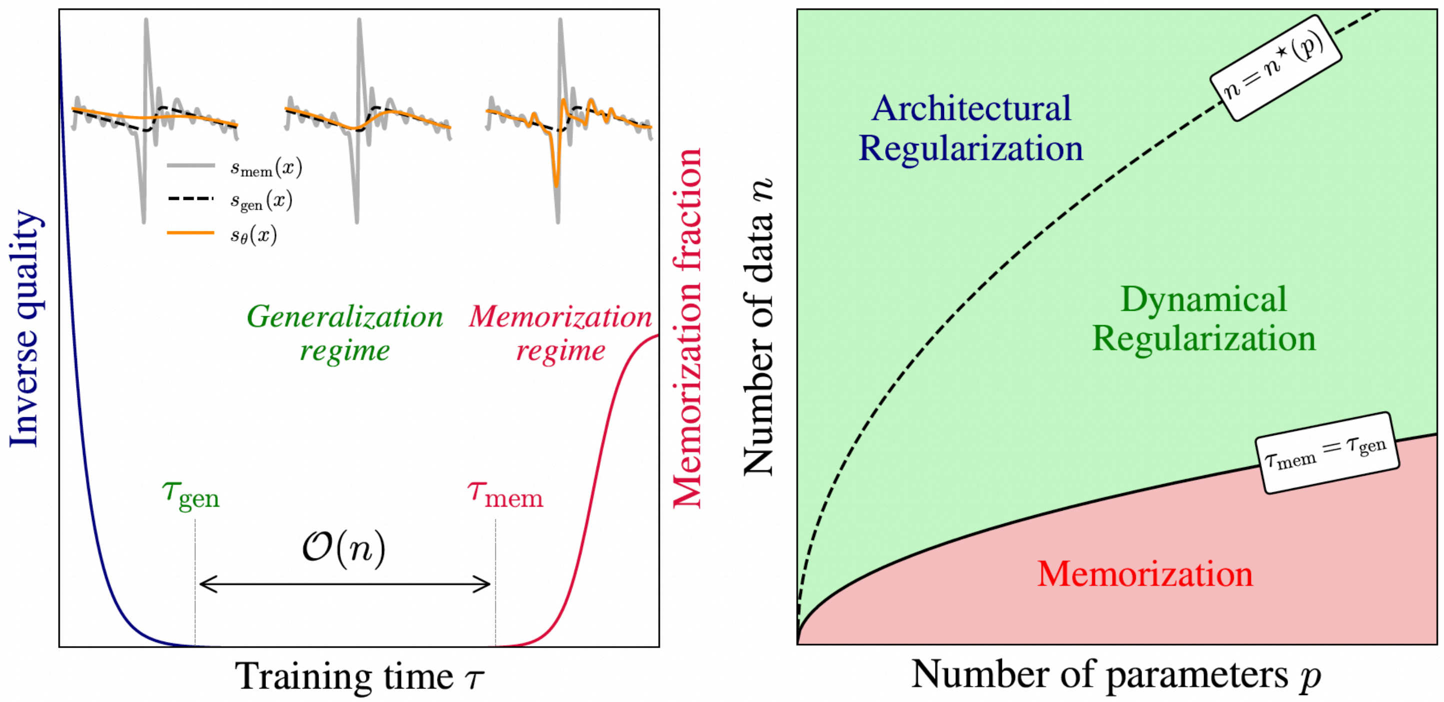
Bonnaire, et al Nat. Comm. 2024

How Diffusion Models Avoid Memorization in Practice?

- Generalisation due to architectural regularization
  - Kamb, Ganguli 2024; Kadkhodale et al. 2023 -> convolutional architecture
  - George, Veiga, Macris 2025: "Denoising Score Matching with Random Features: Insights on Diffusion Models from Precise Learning Curves" -> Analytical study on Random Feature Score Models
- Generalisation due to dynamical regularisation
  - Wu, Marion, Biau, Boyer 2025: "Taking a big step: Large learning rates in denoising score matching prevent memorization." - > learning rate
  - Li,Li, Zhang, Bian 2025: "On the generalisation properties of diffusion models" -> early stopping

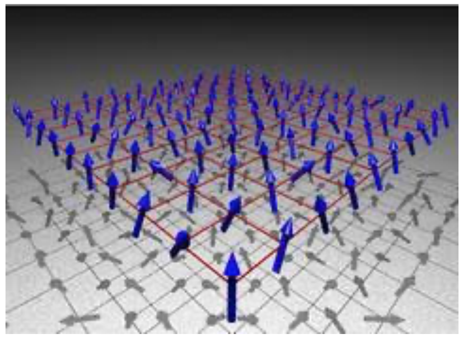
Implicit dynamical regularisation

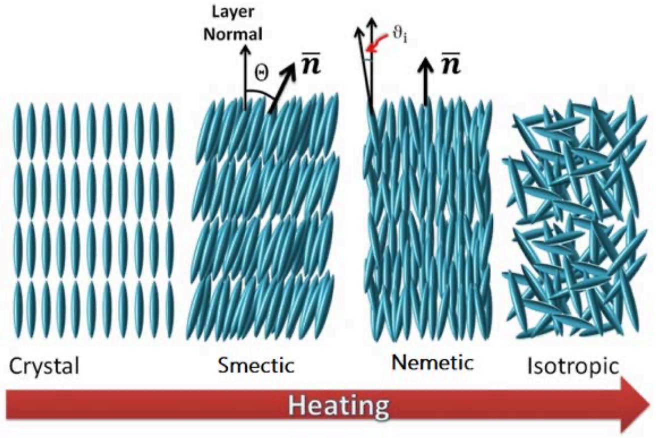
Bonnaire, Urfin, GB, Mézard 2025
See also Favero,Schlocchi, Wyart 2025

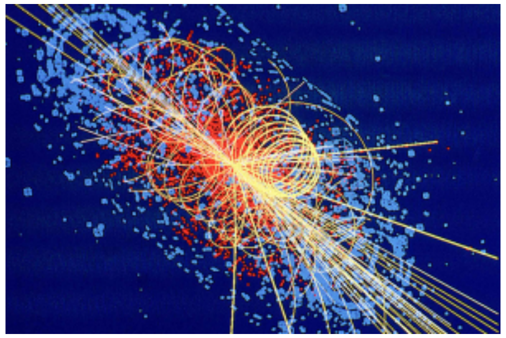


A Recap (or Crash Course) in Renormalization Group

One of the most important conceptual framework in physics -> theory of phase transitions, high-energy physics, multi scale phenomena







Kenneth Wilson 1982
Nobel Prize in Physics

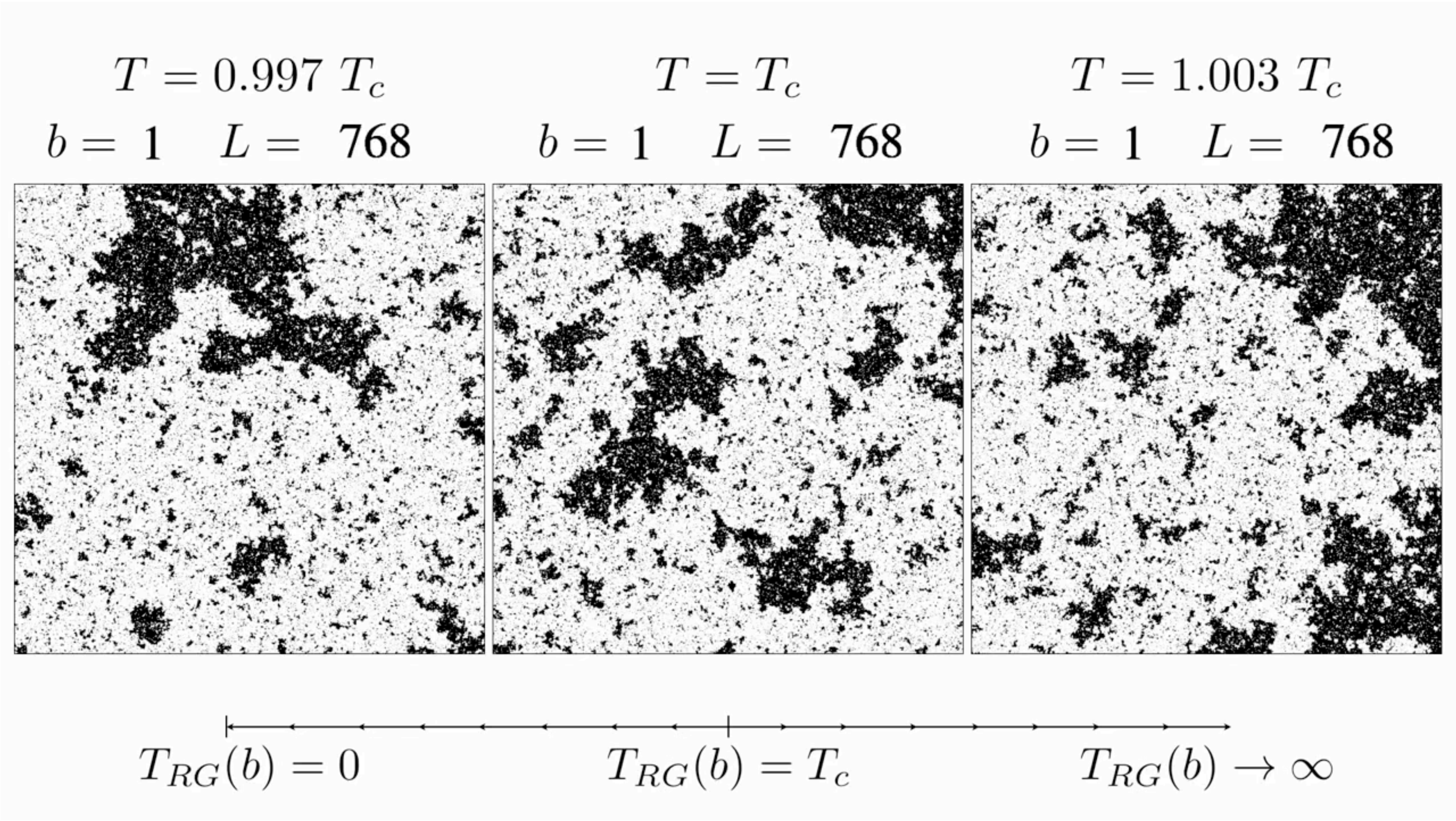
- RG: Hierarchical coarse grain of the probability distribution from small to large scale. From small scale properties to large scale physics.

- After more than 50 years of works on RG, it turns out that there is a new dynamical formulation of RG and that is strongly connected to diffusion models!

Bauerschmidt, Bodineau, Dagallier 2023 (and before); Clothler, Rezchikov 2023; Masuki, Ashida 2025

Renormalisation group in a nutshell

RG for
the Ising Model



- Integrate out the “fast” (or local) degrees of freedom and rescale

$$\ell_{j-1} = 2^{j-1} \to \ell_j = 2^j \qquad \{\varphi_{j-1}(i)\} \to (\{\varphi_j(i)\}, \{\psi_j(i)\})$$

Coarse-grained field

Small scale fluctuations
“Fast degrees of freedom”

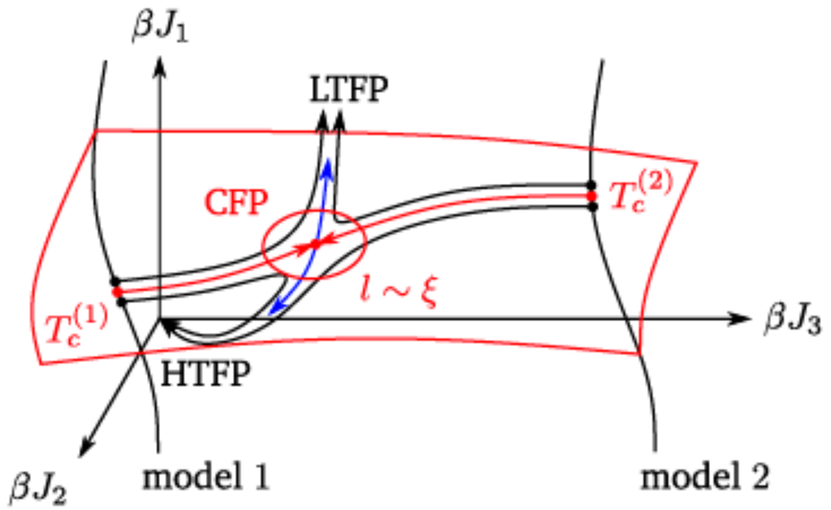
$$P_j(\varphi_j) = \int d\psi_j P_{j-1}(\varphi_{j-1})$$

$$P_j(\varphi_j) = \frac{1}{Z_j} e^{-\mathcal{S}_j(\varphi_j)}$$

$$\mathcal{S}_{j-1}(\varphi_{j-1}) \to \mathcal{S}_j(\varphi_j)$$

Renormalisation group in a nutshell

- Integrate out the “fast” (or local) degrees of freedom and rescale
- RG leads to a flow in energy functions (or probability distributions)
- Second order phase transition associated to non-trivial fixed points



Crucial ingredient
RG always works on short-scale (or“fast”) degrees of freedom scale by scale

$$\{\varphi_{j-1}(i)\} \to (\{\varphi_j(i)\}, \{\psi_j(i)\})$$

The probability distribution is singular (phase transition + multiscale)
Perturbation theory fails but approximating the RG flow is fine

Approximation -> no singular behaviour (divergencies), no instability.

$$P_j(\varphi_j) = \int d\psi_j P_{j-1}(\varphi_{j-1})$$

$$\mathcal{H}_{j-1}(\varphi_{j-1}) \to \mathcal{H}_j(\varphi_j)$$

Renormalisation group in a nutshell

- Obtaining the RG flow is a crucial for many physical systems -> major problem in physics
- Many methods to implement RG approximatively (Kadanoff real-space, Wilson-Fisher Fourier space, Operator expansions,…)
- Exact and non-perturbative RG by Polchinsky (and later Wetterich)