Machine Learning Assisted Sampling: Applications to Physics

Oct 3-7, 2022

ASSAI semester IA-PhyStat
Workshop summary

Organizers: Marylou Gabrié (CMAP, École Polytechnique),
Tony Lelièvre (ENPC, Cermics), Valentin de Bortoli (ENS, Deepmind)

Nov 29, 2023

Participants

Invited Speakers

- David Aristoff, Colorado State University
- Peter Bolhuis, University of Amsterdam
- Freddy Bouchet, École Normale Supérieure, Lyon
- Maria K. Cameron, University of Maryland
- Arnaud Doucet, University of Oxford
- Alain Durmus, ENS Paris Saclay & École Polytechnique
- Stéphane Mallat, Ecole Normale Supérieure, Paris
- Pierre Monmarché, Sorbonne Université
- Jutta Rogal, New York University
- Phiala Shanahan, Massachusetts Institute of Technology
- Gabriel Stoltz, École Nationale des Ponts et Chaussées, Paris
- Jonathan Weare, New York University
- Martin Weigt, Sorbonne Université
- Wei Zhang, Zuse Institute Berlin

Registered participants ~ 80 (50 % online)

Format: 2 open days + 3 discussion days

Monday Tuesday Wednesday Thursday Friday

Public workshop with all the talks

Organized discussion sessions among invitees

Model Reduction

Rare Event sampling

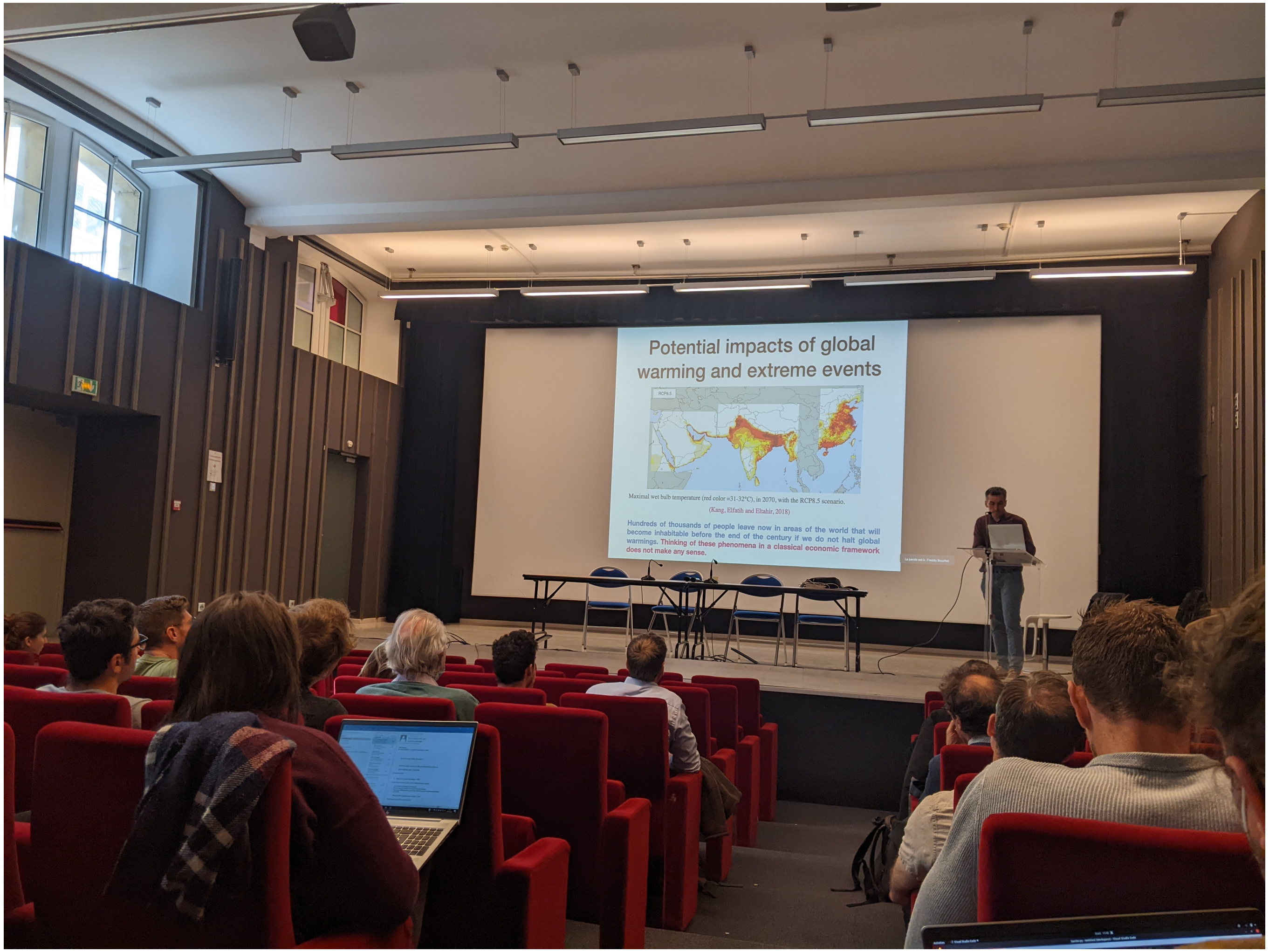
Structured Generative Models

Free discussions

On YouTube!

Freddy Bouchet's Colloquium:

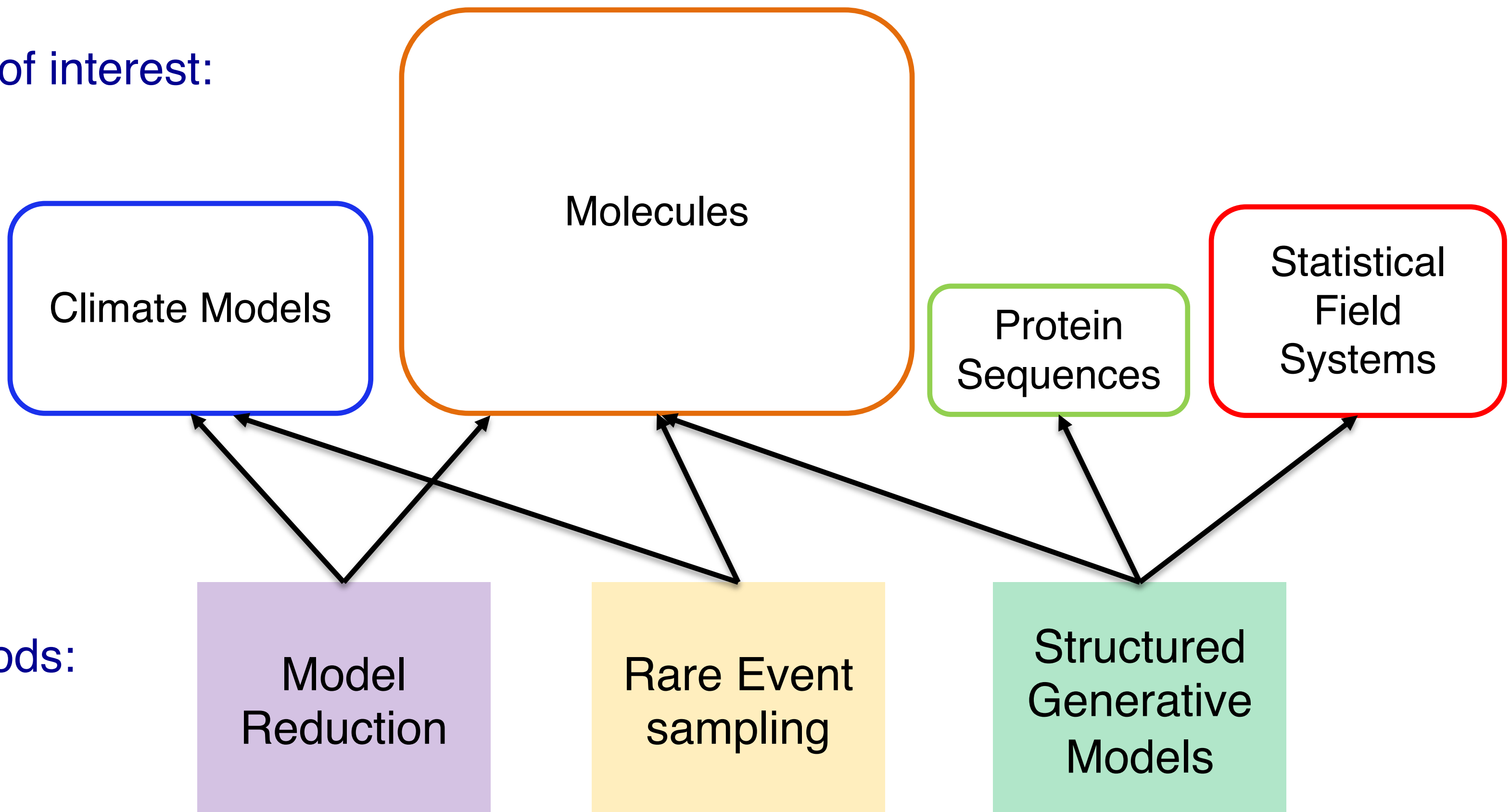
Probabilistic forecast of extreme heat waves using convolutional neural networks and rare event simulations



the only picture I took ...

Main themes & domains of interests of participants

- ▷ Diverse backgrounds of physicists, applied mathematicians and chemists
- ▷ All with an interest in sampling
- ▷ Systems of interest:



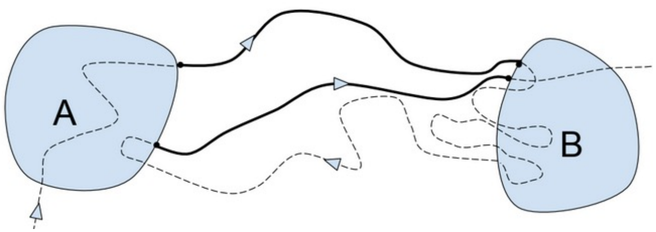
- ▷ Expertise & methods:

Molecular Systems & Climate Sciences: context

Well established field with experts thinking of opportunities in using machine/deep learning

▷ Given a known dynamic operator and/or scarce observations:

- How to sample rare transitions?



- Identifying reactive channels & maximum likelihood transition paths

- Estimating reaction rates or expected escape times from basins of attractors

e.g. Understand isomers transformations, characterize whether a storm will become a hurricane

▷ Important mathematical objects to characterize the dynamics

- Infinitesimal generator

$$\mathcal{L} = \text{generator of the SDE}$$

- Committor

$$q(x) = \mathbb{P}(\tau_A < \tau_B) = \text{committor}$$

solution of

$$\boxed{\begin{aligned} \mathcal{L}q(x) &= 0, \quad x \in (A \cup B)^c \\ q(\partial A) &= 0, \quad q(\partial B) = 1 \end{aligned}}$$

[Credit images & equations Maria Cameron]

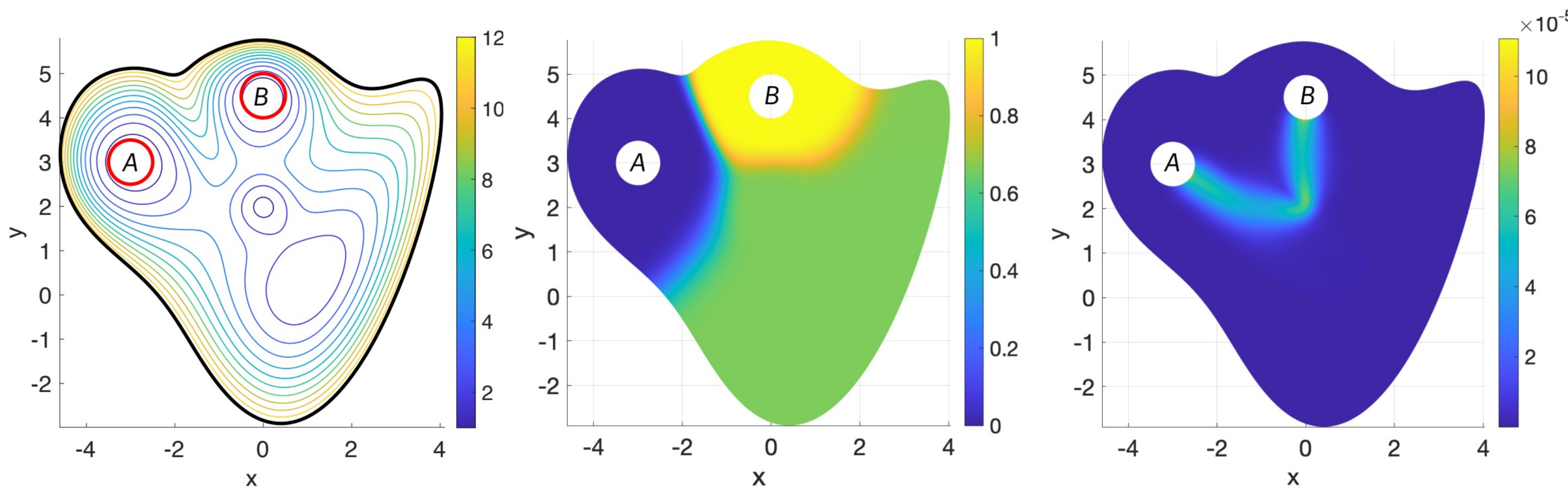
Molecular Systems & Climate Sciences: context

▷ Important mathematical objects to characterize the dynamics

- Infinitesimal generator $\mathcal{L} = \text{generator of the SDE}$
- Committor $q(x) = \mathbb{P}(\tau_A < \tau_B) = \text{committor}$

$$\boxed{\begin{array}{ll} \mathcal{L}q(x) = 0, & x \in (A \cup B)^c \\ q(\partial A) = 0, & q(\partial B) = 1 \end{array}}$$

The potential function | The committor | The reactive current

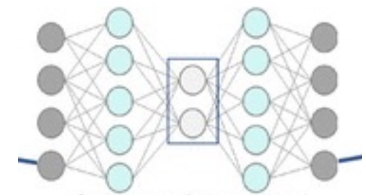


↳ The committor is the optimal “reaction coordinate” to force the transition, but is a high-dimensional function: can it be inferred from data?

[Credit images & equations Maria Cameron]

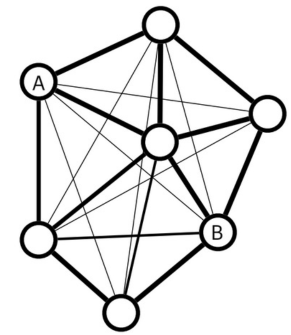
Molecular Systems & Climate Sciences: methods

▷ Dimensionality reductions



- Use **auto-encoders** to learn non-linear low-d embeddings (Bolhuis, Stoltz)

  ↳ Which features to include in the reconstruction loss?



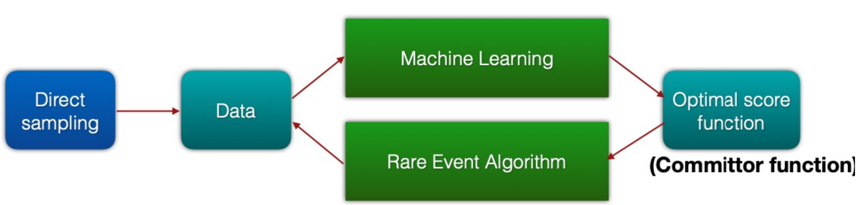
- Project the dynamics to discrete states before solving for the committor solving a **linear program** (Cameron, Bouchet)

  ↳ Is the projected dynamics precise enough?

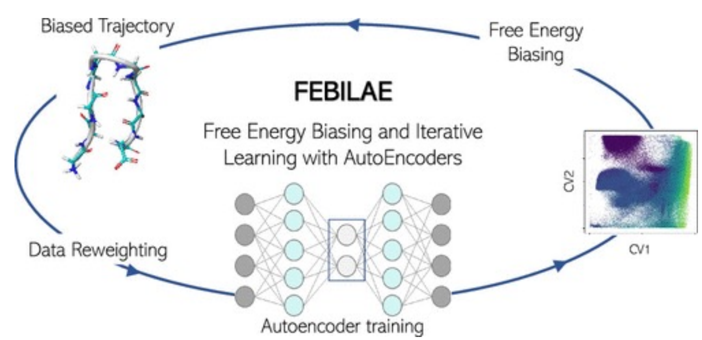
- Identify slow modes associated with metastability using a **feed-forward NN to solve variationally the eigenvalue problem** of the infinitesimal generator (Zhang)

  ↳ How do NN ansatz compare to usual basis function expansions?

▷ Adaptive methods to create data as you go (Bouchet, Stoltz, Gabrié, Weare)



[Credit F. Bouchet]



[Belkacemi et al JCTC 2021]

Generative models for protein sequence data & statisfical field systems

Experts from different fields that have found applications of generative probabilistic models

▷ Perform density estimation with a structured model and resample data

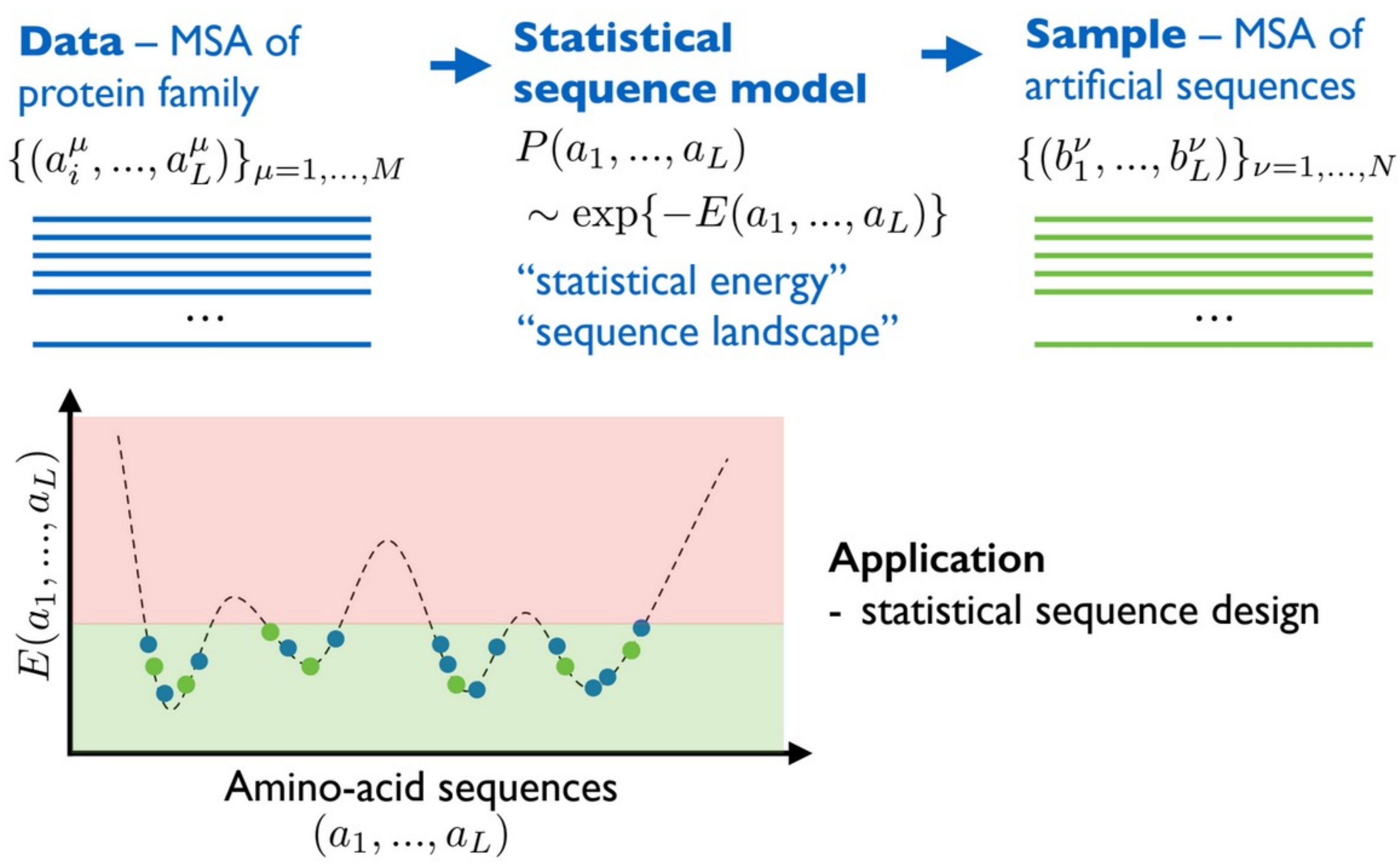
- Predict mutational effects, forecast protein evolutions, design proteins (Weigt)
- Avoid critical slowing down at resampling time (Mallat)

▷ Sample from a distribution known up to a normalization constant leveraging a surrogate generative model

- Bayesian posterior (Doucet, Gabrié)
- Boltzmann distributions (Shanahan, Gabrié)

Energy Based and Autoregressive Models for Protein Modelling

▷ Motivations for inferring a sequence lanscape



[Credit M. Weigt]

Scale by scale modelling of field systems

complicated distribution

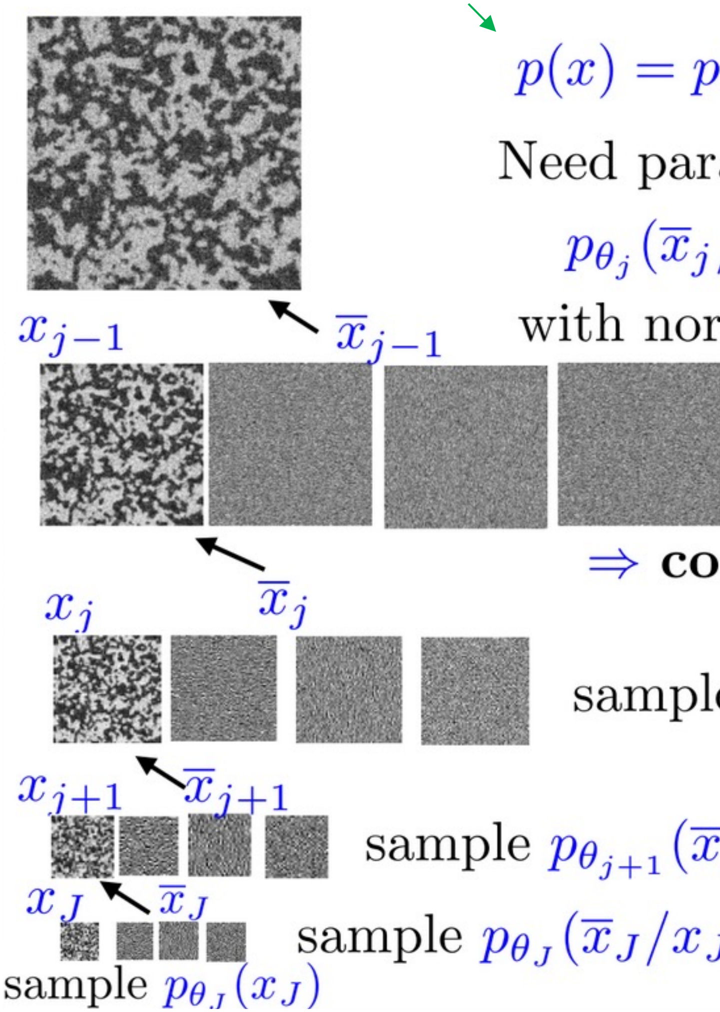
$$p(x) = p(x_J)\prod_{j=1}^{J} p(\overline{x}_j / x_j)$$

product of easy conditionals - to be learned

Need parameterised models

$$p_{\theta_j}(\overline{x}_j / x_j) = \mathcal{Z}_j^{-1}\, e^{-U_{\theta_j}(\bar{x}_j / x_j)}$$

with normalised variances $\bar{x}_j \to \bar{x}_j / \sigma_j$



$\Rightarrow$ **coarse to fine sampling**

sample $p_{\theta_j}(\overline{x}_j / x_j)$

sample $p_{\theta_{j+1}}(\overline{x}_{j+1} / x_{j+1})$

sample $p_{\theta_J}(\overline{x}_J / x_J)$

Log-concave conditionals allow fast sampling of learned factorization

[Credit S. Mallat]

Generative models in for protein sequence data & statisfical field systems

▷ Perform density estimation with a structured model and resample data

- Predict mutational effects, forecast protein evolutions, design proteins (Weigt)
- Avoid critical slowing down at resampling time (Mallat)

↳ Quality of the retrieved model hard to measure

▷ Sample from a distribution known up to a normalization constant leveraging a surrogate generative model

- Bayesian posterior (Doucet, Gabrié)
- Boltzmann distributions (Shanahan, Gabrié)

↳ Generative models incorporated in algorithms with performance guarantees

Assisting sampling with surrogate generative models

No data a priori, only a density of probability $\rho_*(x)$ (Bayesian posterior, Boltzamnn distribution)

▷ **Architecture strategies:** Design generative models to incorporate known symmetries to ease the learning of a surrogate $\rho_\theta \approx \rho_*$ (e.g. Lattice QCD gauge invariances)

▷ **Training strategies:**

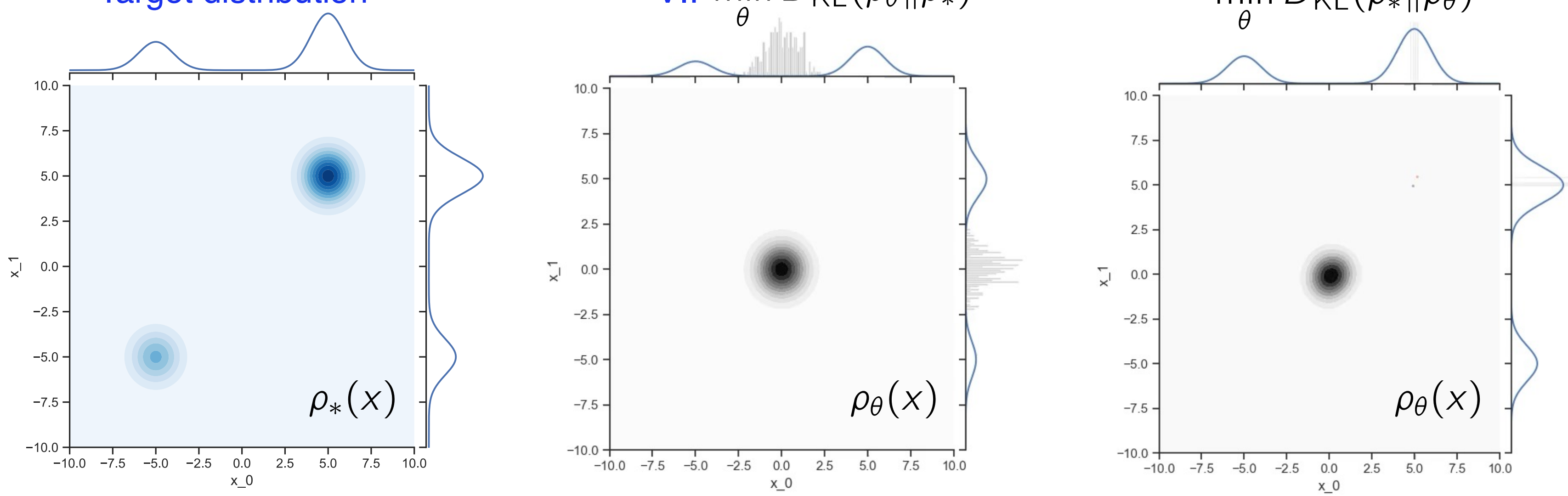
- Variational inference (VI)
- Adaptive training to create data as you go

MCMC convergence garantees!

Adaptive MCMC: $\min_\theta D_{\mathrm{KL}}(\rho_* \| \rho_\theta)$

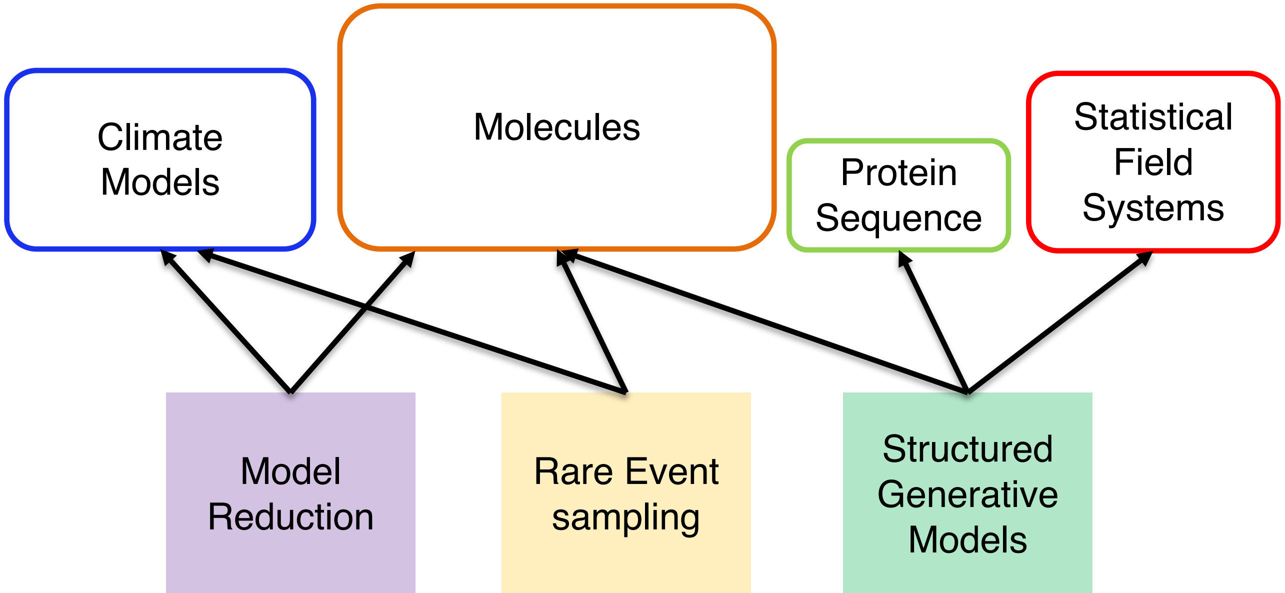
Target distribution

VI: $\min_\theta D_{\mathrm{KL}}(\rho_\theta \| \rho_*)$



Learning a well covering generative model requires minimum knowledge of modes before-hand

Conclusion: how/where has machine learning helped?



- ▷ Autoencoders to learn non-linear dimensionality reductions
- ▷ Variational principles to solve eigenvalue problems & partial differential equations
- ▷ Normalized generative models to accelerate sampling
- ▷ Structured generative models to extract/exploit structure from data

Format: 2 open days + 3 discussion days

Monday Tuesday Wednesday Thursday Friday

Public workshop
with all the talks

On YouTube!

Organized discussion sessions
among invitees

Model
Reduction

Rare Event
sampling

Structured
Generative
Models

Free discussions

Machine Learning Assisted Sampling: Applications to Physics

Oct 3-7, 2022

ASSAI semester IA-PhyStat
Workshop summary

Organizers: Marylou Gabrié (CMAP, École Polytechnique),
Tony Lelièvre (ENPC, Cermics), Valentin de Bortoli (ENS, Deepmind)

Nov 29, 2023