



**HABILITATION À DIRIGER
DES RECHERCHES**

DE L'UNIVERSITÉ PSL

Présentée à l'École normale supérieure

**Leveraging generative models to assist
Monte Carlo sampling**

*An example of scientific computing assisted by
modern machine learning techniques*

Présentation des travaux par

Marylou GABRIÉ

Le 03 juillet 2026

Discipline

Physique

Composition du jury :

Julie DELON Professeure, École Normale Supérieure	<i>Présidente</i>
Arnaud DOUCET Professor, University of Oxford	<i>Rapporteur</i>
Alberto ROSSO Directeur de recherche, Université Paris Saclay	<i>Rapporteur</i>
Francesco ZAMPONI Professore, Sapienza Università di Roma	<i>Rapporteur</i>
Giulio BIROLI Professeur, École Normale Supérieure	<i>Examineur</i>
Stéphane MALLAT Professeur, École Normale Supérieure et Collège de France	<i>Examineur</i>
Jutta ROGAL Research scientist, Flatiron Institute	<i>Examinatrice</i>
Beatriz SEOANE BARTOLOME Profesora, Universidad Complutense de Madrid	<i>Examinatrice</i>

Acknowledgements

I would like to thank the members of the jury for taking the time to evaluate my work, and in particular Arnaud Doucet, Alberto Rosso, and Francesco Zamponi for agreeing to serve as rapporteurs. I very much look forward to the discussions following the presentation, given the breadth of expertise and perspectives you each bring. I am also grateful to Tony Lelièvre and Guilhem Semerjian for their insightful comments on an earlier version of this manuscript, and to Luigi Fogliani for carefully spotting a number of typos.

Looking back on my research journey since the end of my PhD, I feel deeply grateful to the collaborators and students with whom I have had the chance to work on sampling with generative models. My interest in the topic was sparked during my postdoc at NYU through collaborations with Eric Vanden-Eijnden and Grant Rotskoff, whom I thank for many stimulating discussions—on Zoom as much as in person—throughout the pandemic.

After returning to Paris, I was fortunate to recruit the ever-motivated Louis Grenioux, with whom we initiated methodological developments in close interaction with Alain Durmus and his student Maxence Noble, exploring in particular diffusion-based approaches.

On the applications side, I am grateful to Pilar Cossio, with whom we joined forces with Olga Lopez-Acevedo and Ana Molina-Taborda to study metallic nanoclusters in what became my first all-female collaboration. My work on molecular systems greatly benefited from collaborations with Tony Lelièvre and Gabriel Stoltz, driven by the meticulous work of Christoph Schönle and later joined by the enthusiastic Davide Carbone—an experience marked by both scientific depth and genuine enjoyment. Around the same time, interactions with Alessandro Laio led to another fruitful collaboration, during which we deeply felt the loss of Samuel Tamagnone, taken back by the mountains he loved.

More recently, my move to ENS opened new directions on particle systems through collaborations with Ludovic Berthier and Giulio Biroli, with continued contributions from Louis Grenioux and the rapid involvement of Leonardo Galliano. It also fostered theoretical developments: on the one hand with Bruno Loureiro and our talented students Roman Soletskyi and Luigi Fogliani on learning theory for variational inference; on the other hand with Giulio Biroli, Patrick Charbonneau, and the ever-resourceful Luca del Bono, exploring diffusion models in simplified field-theoretic settings.

Over the past three years, I have also run the SaLT reading group (“Sampling, Learning and Transport”) at École Polytechnique and later at ENS. I warmly thank all speakers and participants who made it a lively and stimulating forum for exchange.

More broadly, I am grateful to my colleagues at ENS for providing an environment that is both intellectually stimulating and genuinely supportive. I particularly thank Bruno Loureiro, Giulio Biroli, Gabriel Peyré, Olivier Cappé, Stéphane Mallat, Eddie Aamari, Kimia Nadjahi, and Julie Delon for contributing to the vibrant scientific life at CSD. In the physics department, I am especially thankful to Jean-Marc Berroir for his warm welcome, and to Giulio Biroli, Camille Scalliet, and Guilhem Semerjian for their friendship and informal guidance. In the chemistry department, I look forward to developing our collaboration with Guillaume Stirnemann and Damien Laage, currently in the good hands of Pierre Veillard.

Finally, I thank my family and friends for their support throughout my post-PhD journey in academia. My deepest gratitude goes to Raphaël, who helps me broaden my perspective every day and offers unwavering support and encouragement, both within and beyond my academic life.

Leveraging generative models to assist Monte Carlo sampling

(An example of scientific computing assisted by modern machine learning techniques)

Marylou Gabri 

March 17, 2026

Foreword

My objective here is to write something that will be useful beyond the evaluation of my “Habilitation   diriger des recherches”. Having worked consistently on the topic of sampling with generative models since 2019, I have witnessed the field evolve from a few pioneering works to a very active area of research. While I cannot hope to be exhaustive given the current pace of development, I will try to give a snapshot of the early chapters of the story of sampling with generative models, reviewing and placing in perspective the contributions I have been involved in along the way. The works I co-authored in are marked by an extra [*] in the citation label.

I was very lucky along the years to be able to collaborate with many talented researchers and students who have nourished my understanding of this field. I am greatly indebted to them and I hope that this manuscript will be a valuable resource for researchers, students and practitioners interested in this exciting area of research.

Contents

1	Introduction	6
1.1	Why sampling? Why is it hard?	6
1.2	Generative models for sampling	8
1.3	Outline of the manuscript and my journey in the field	9
2	A brief introduction to transport-based generative models	11
2.1	Normalizing flows	11
2.1.1	Original formulation	11
2.1.2	Continuous normalizing flows	12
2.2	Diffusion models	13
2.3	Stochastic interpolants a.k.a. flow matchings	15
3	Exact samplers with normalizing flows	17
3.1	Initial NF-powered samplers	17
3.1.1	Importance sampling	17
3.1.2	Independent proposal Markov chains	19
3.1.3	Reparametrization through transport maps	20
3.2	A first comparison of NF powered samplers	21
3.3	Combining NFs with tempering	24
3.3.1	Adapting the base distribution to the target	24
3.3.2	Combining NFs with AIS: stochastic normalizing flows	25
3.3.3	Combining NFs with SMC	26
3.3.4	Combining NFs with PT	27
3.3.5	Recalibrating NF samples with AIS or SMC	28
3.4	Combining NFs with coarse-granings	28
3.4.1	Partial resampling in the latent space	29
3.4.2	Collective-variable guided sampling	30

4	Training generative models without data	33
4.1	Normalizing flows	33
4.1.1	Variational inference	33
4.1.2	Adaptive strategies	35
4.1.3	Sequential tempering	36
4.2	Continuous annealed flows	37
4.3	Diffusion models	40
4.3.1	On estimating the likelihood of diffusion models	40
4.3.2	Monte Carlo estimation of the score	41
4.3.3	Revisiting variational inference with stochastic optimal control	44
5	Verifiable sampling with continuous-time generative models	48
5.1	Reweighting for continuous normalizing flows and stochastic interpolants	48
5.2	Diffusion samplers as stochastic flows	50
6	Examples of applications	52
6.1	Training energy-based models	52
6.1.1	Definition and challenges	52
6.1.2	Proposed method: flowMC-EBM	53
6.1.3	Preliminary results	54
6.2	Sampling isomers of metallic nanoclusters	55
6.2.1	Motivation and test system	55
6.2.2	Adjustments to A-flowMC	55
6.2.3	Summary of results	56
6.3	Toward sampling conformational ensembles of solvated biomolecules	58
6.3.1	Test system and CVs	58
6.3.2	Design choices for the training and sampling pipeline	59
6.3.3	Summary of results	60
7	Perspectives	62

List of Algorithms

1	Neural importance sampler (Neural-IS)	17
2	flow independent Metropolis Hastings (flow-IMH)	19
3	flow iterated sampling importance resampling (flow-iSIR)	20
4	neutra-MCMC	21
5	(Continuously Repeated) Annealed Flow Transport (AFT/CRAFT)	26
6	Learned replica exchange (LREX)	27
7	Gibbs flow-Monte Carlo (GflowMC)	29
8	Non-equilibrium update of orthogonal degrees of freedom	31
9	Collective-variable guided flow-Monte Carlo (CV-FlowMC)	32
10	Adaptive flow Monte Carlo (A-flowMC)	35
11	Sequential tempering	37
12	On the fly score estimation by Langevin dynamics	41
13	Reverse diffusion Monte Carlo (RDMC)	42
14	Stochastic localization through iterative posterior sampling (SLIPS)	43
15	Iterated score regression	44
16	flowMC-EBM training	53

List of Acronyms and Notations

Notation	Definition
I_d	Identity matrix of dimension d
0_d	Zero vector of dimension d
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and covariance Σ
$\mathcal{N}(\cdot; \mu, \Sigma)$	Gaussian probability density function with mean μ and covariance Σ
StopGrad($\cdot$)	Stop gradient: dependency on parameters is not backpropagated
Acronym	Definition
AIS	Annealed Importance Sampling
BAR	Bennett Acceptance Ratio
CNF	Continuous Normalizing Flows
CV	Collective Variable
DM	Diffusion Models
ESS	Effective Sample Size
HMC	Hamiltonian Monte Carlo
IMH	Independent Metropolis-Hastings
IS	Importance Sampling
iSIR	Iterated Sampling Importance Resampling
MALA	Metropolis-Adjusted Langevin Algorithm
MCMC	Markov Chain Monte Carlo
MH	Metropolis-Hastings
NF	Normalizing Flows
ODE	Ordinary Differential Equation
OU	Ornstein-Uhlenbeck
PT	Parallel Tempering
SDE	Stochastic Differential Equation
SI	Stochastic Interpolants
SMC	Sequential Monte Carlo
TFEP	Targeted Free Energy Perturbation
ULA	Unadjusted Langevin Algorithm
VI	Variational Inference

1 Introduction

1.1 Why sampling? Why is it hard?

Sampling and its challenges Sampling is a fundamental task in scientific computing, as it enables the efficient approximation of high-dimensional integrals. In observational and experimental sciences, for instance, sampling is essential for uncertainty quantification and model calibration through Bayesian inference. In theoretical physics, sampling is an indispensable computational tool for studying many-body systems across a broad spectrum of applications, ranging from lattice gauge theory for the strong nuclear force to quantum Monte Carlo methods for solving the Schrödinger equation or investigating the thermodynamic properties of materials.

The task of sampling consists in generating realizations of a random variable according to a given target distribution. In this manuscript, we focus on the common setting where the state space is continuous and the probability density function—denoted by π throughout—is known up to a normalization constant. Sampling from π shares certain features with optimization: both require identifying the most probable regions of the state space, typically located near local maxima of the target density. Crucially, however, sampling must also correctly capture fluctuations around these maxima and reproduce the relative weights of different regions. It is therefore a fundamentally different problem from optimization, with its own intrinsic challenges.

The first modern sampling algorithms emerged in the early 1950s, enabled by the advent of digital computers. The Metropolis algorithm [MU49; von51; Met+53] was the first Markov Chain Monte Carlo (MCMC) method and laid the foundation for a vast family of techniques that followed. Since then, an extensive literature has developed, proposing a wide variety of algorithms tailored to the diverse challenges arising across application domains and target distributions (see, e.g., [Liu04] for a comprehensive introduction).

To this day, two major sources of difficulty remain central: *scalability*, the ability to handle high-dimensional target measures, and *metastability*—also referred to as multimodality—the presence of multiple well-separated local maxima. MCMC methods can often cope with moderately high-dimensional spaces by exploring the state space through random walks. However, their essentially local exploration mechanism typically makes transitions between distant high-probability regions exponentially unlikely. As a result, the chain may remain trapped for extremely long times in so-called metastable states. A striking illustration is provided by protein folding: although the folded configuration corresponds to the global minimum of the Boltzmann free energy, locating this free-energy minimum by sampling the associated Boltzmann distribution with traditional, local molecular dynamics or MCMC methods is essentially impossible in practice, as the dynamics remains trapped for astronomically long times in long-lived metastable intermediates.

Important families of sampling algorithms to tackle metastability in high dimension To clarify why metastability is such a difficult problem, it is useful to briefly review the main families of sampling algorithms that have been developed to address it.

Non-local jump samplers. As previously mentioned, Markov Chain Monte Carlo (MCMC) methods typically rely on local proposals. While this ensures reasonable scalability with dimension, it severely limits mixing in the presence of metastability, making transitions between modes exponentially rare. To address this issue, various approaches have been proposed to incorporate non-local moves. In certain specific settings, remarkably effective solutions have been developed.

In lattice spin systems with discrete variables, cluster algorithms such as the Swendsen-Wang [SW87] and Wolff [Wol89] updates enable collective moves of large groups of spins and can dramatically accelerate sampling close to the critical region despite of a diverging correlation length. In the context of particle systems, several non-local strategies have also proven highly

effective. Swap Monte Carlo [GP01; Ber+19]—which proposes exchanges of particle identities in addition to local displacements—can accelerate equilibration of glass-forming systems by many orders of magnitude. Event-chain Monte Carlo [BKW09], a rejection-free irreversible algorithm that propagates displacements through chains of particles, has likewise led to major improvements in the simulation of dense particle systems.

The statistics literature has attempted to construct general-purpose mode-hopping proposals [TH01; PHL20]. However, designing effective non-local proposals in general is challenging, as it requires identifying high-density regions that occupy an increasingly small fraction of the state space as the dimension grows. Simple parameterizations of such proposals, such as Gaussians or mixtures of Gaussians, typically fail in high dimension, since these approximations remain too crude and lie too far from the true target distribution.

Annealing-based samplers. A powerful family of methods for sampling complex distributions is based on tempering, or annealing. The central idea is to define a sequence of intermediate distributions that bridge a simple base distribution and the complex target distribution. This bridge is typically constructed by gradually increasing an effective temperature or weakening interactions between variables until reaching an independent (product) measure. The resulting path is then exploited to progressively transport samples toward the target, which can substantially improve exploration in multimodal settings. Among the most widely used tempering-based methods are annealed importance sampling (AIS) [Nea01], parallel tempering (PT, also known as replica exchange) [SW86; Gey91; HN96], and sequential Monte Carlo (SMC, also known as population annealing) [HI03; DDJ06].

Despite their effectiveness, annealing methods present practical challenges. The choice of a suitable annealing path and of the number of intermediate distributions can have a decisive impact on performance. Self-adaptive SMC [CP20] provides a principled mechanism to adjust the number of steps dynamically and is among the most effective algorithms currently available for multimodal targets. However, annealing methods may still struggle when the reference measure differs too drastically from the target distribution. In protein conformational sampling, for example, high-temperature annealing may lead to complete denaturation, so that configurations obtained at high temperature carry little useful information about the folded state. Similarly, in statistical mechanics, the presence of a discontinuous (first-order) phase transition along the annealing path can break the connection between high- and low-temperature regimes, rendering the bridge ineffective. Thus, while annealing-based samplers represent the state of the art for addressing multimodality, alternative strategies remain necessary — in particular in situations where constructing a suitable annealing path is difficult, or where one may hope to bypass annealing altogether through more direct and potentially more efficient approaches.

Dimensionality reduction to enhance sampling. Another classical strategy to tackle high-dimensional multimodal sampling problems consists in identifying a low-dimensional projection of the full configuration space — often referred to as a set of collective variables (CVs), or equivalently a coarse-graining of the system — that resolves the different modes of the target distribution. Sampling across these modes is then driven by forcing the exploration in the CV space which is typically of much lower dimension than the full space. These approaches, particularly popular in molecular simulation, come in many variants; a recent review can be found in [Hén+22]. However, these techniques face two major challenges. First, they are typically restricted to very low-dimensional CV spaces, as they rely on non-parametric estimation of the marginal distribution in the reduced space, which becomes rapidly intractable as the dimension of the coarse-graining increases. Second, identifying an appropriate set of collective variables — that is, a suitable low-dimensional projection capable of separating the relevant metastable states — is itself a difficult task, especially under stringent dimensionality constraints.

Building on recent advances in machine learning, a new paradigm for sampling has begun to emerge, leveraging the expressive power of generative models to learn complex probability distributions and produce representative samples. In the following, we aim to present the main ideas and results in this rapidly developing field. In what sense do these methods overcome the limitations discussed above? How can they be combined with classical approaches? What challenges and open questions remain?

1.2 Generative models for sampling

Generative modelling vs sampling Let us first clarify the difference between generative modelling and sampling, as these two terms are often used interchangeably in the literature. *Generative modelling* consists in learning a probabilistic model – namely the generative model – given an extensive set of samples from an unknown distribution. It is a fundamentally data-driven task of density estimation, where the goal is to learn a model that can generate new samples that are similar to the training data, this similarity being not necessarily well-defined¹. In contrast, *sampling* is the task of generating samples from a target distribution whose probability density function is known (possibly up to a normalizing constant). The task is different in two fundamental ways: (i) instead of a large dataset, information must be extracted from the knowledge of a mathematical expression and little if any data is available; (ii) the objective is to produce samples that are exactly distributed according to the target measure, which has a well-defined mathematical meaning.

Generative models for scientific computing Generative models are highly flexible probabilistic models typically parametrized via deep neural networks that were primarily developed for generative modelling tasks on images and text generation with the incredible successes we know. They have also already proven successful in data-driven computational tasks in physics and chemistry, such as protein folding with AlphaFold [Jum+21] and predicting protein-ligand binding structures [Cor+23]. Here, we will discuss how generative models are also proving to be powerful tools for sampling.

Pioneering works in this direction [LW18; AKS19; Noé+19] have demonstrated that subclasses of generative models, namely normalizing flows and autoregressive networks, can rather easily be repurposed for exact sampling by enabling the computation of reweighting of the generated realizations with respect to the target measure. Proofs of concept have been shown for various areas of physics and chemistry, including lattice quantum field-theories [Abb+24], biomolecules [Noé+19] or, in sampling nano-clusters of heavy atoms at quantum accuracy [Mol+24*].

These early successes have also uncovered important limitations of such direct reweighting strategies, and have spurred the development of more sophisticated approaches. As a result, a broad landscape of methods has emerged, ranging from refined flow-based constructions to the most recent diffusion models adapted to sampling tasks. This diversity has created a rapidly evolving “zoo” of techniques in which conceptual connections are not always transparent and methodological differences can be subtle. In this manuscript, we survey a representative subset of these approaches with the aim of clarifying their underlying principles, highlighting their common structure, and emphasizing the trade-offs they entail. While striving to remain concise, we focus on the core ideas and main results, and refer the reader to the original works for technical details and further developments.

¹For one thing, the distribution most faithful to a training dataset is the sum of Dirac deltas at the training points.

Previous partial reviews of sampling with generative models As the field continues to gain momentum, any attempt at exhaustiveness would be futile. Nevertheless, assembling the first chapters of this rapidly developing story can be valuable, particularly for newcomers seeking an entry point into the subject. As of March 2026, the field is evolving at a fast pace, and new ideas have already appeared that are not covered in the present manuscript.

We would also like to point to several works that have already attempted to summarize specific aspects of the field. Chapter 7.2 of the textbook [Daw+25*], coordinated by Anna Dawid following a doctoral school on machine learning for quantum physics held in September 2021, reviews early developments in the use of generative models for many-body physics. The lecture notes by Eric Vanden-Eijnden for the Les Houches Summer School on Statistical Physics & Machine Learning (August 2022) introduce foundational principles of sampling with generative models [AV24]. The commentary [Cor+24] discusses the pioneering work [No  +19] on Boltzmann generators and its immediate extensions. More focused reviews include [Cra+23], which surveys advances with specific applications to lattice quantum chromodynamics, [He+25], which places different diffusion-based sampling approaches into perspective and [Ols26], which reviews the use of generative models for molecular dynamics simulations.

1.3 Outline of the manuscript and my journey in the field

The “pipeline” of sampling with generative models can be decomposed into three main stages, which will guide the structure of this manuscript. The first stage (i) consists in choosing a suitable family of generative models, typically a transport-based model such as a normalizing flow (NF) or a diffusion model.

Section 2 briefly reviews this class of generative models, both for completeness and to introduce the notations used throughout the manuscript. It also recalls how these models are traditionally trained from data in the standard “generative modelling” setting.

The next two stages in sampling with generative models are (ii) training the generative model in the specific context of sampling, where no data is available, and (iii) designing sampling algorithms that leverage the trained model to generate samples from the target distribution. These two stages are closely intertwined. In particular, as we will see, the training procedure can itself exploit the sampling capabilities of the generative model. For this reason, the exposition does not strictly follow the chronological order of the pipeline, but instead adopts an organization that we believe makes the underlying ideas clearer.

Section 3 focuses on the design of exact sampling algorithms, assuming that a generative model has already been trained. This section concentrates in particular on algorithms based on “discrete-time” NFs, which were the first generative models to be used for sampling and provide a transparent setting to understand the core principles. We move from early, straightforward reweighting strategies to more sophisticated approaches aimed at extending the scope of these methods, notably in the presence of metastability.

Section 4 then takes a step back to examine the training of generative models for sampling. Starting with procedures originally developed for “discrete-time” NFs, we discuss how these ideas extend to more recent classes of models, such as continuous normalizing flows and diffusion models.

Section 5 returns to the design of sampling algorithms, this time focusing on recent developments involving “continuous-time” normalizing flows and diffusion models.

Taken together, these sections provide a general review of the early chapters in the development of sampling with generative models. Along the way, they discuss several contributions to which I

have been directly involved, in particular in sections 3 and 4. For the purpose of this Habilitation, references to my own works are explicitly highlighted with a [*].

The final sections adopt a more personal perspective.

Section 6 presents selected applications of generative models for sampling in machine learning and chemical physics. These works, carried out in collaboration with colleagues and students, illustrate the practical potential of these methods for addressing challenging sampling problems.

Section 7 concludes with perspectives on the field and points to directions of interest that were not covered in this manuscript.

2 A brief introduction to transport-based generative models

Here, we briefly introduce the different types of generative models used to assist Monte Carlo sampling, while introducing the notation employed throughout this review. We focus on tasks lying in continuous state spaces, generically denoted by $\mathcal{X}$, which we assume to be a subset of $\mathbb{R}^d$. The models presented below belong to the class of deep latent generative models, which transform samples drawn from simple base distributions into complex outputs. Originally, deep latent generative models—such as variational autoencoders and generative adversarial networks—relied on latent variables of much lower dimensionality than the state space of interest, reflecting the manifold hypothesis, namely that high-dimensional data typically lie close to a manifold of much lower intrinsic dimension. More recent models presented here drop this assumption and instead transport base samples defined in the same space as the data to generate new samples.

2.1 Normalizing flows

2.1.1 Original formulation

Normalizing flows (NFs) were introduced by Tabak & Vanden Eijnden [TV10] as a method for density estimation on $\mathbb{R}^d$, based on learning the transport map from an empirical distribution toward a simple, typically Gaussian, base distribution, hence the name *normalizing flow*. More generally, denote ρ_0 the density of the base distribution on the state space $\mathcal{X} \subset \mathbb{R}^d$, and T_θ a diffeomorphism on $\mathcal{X}$ with parameter vector $\theta \in \Theta$. A NF is a learnable probabilistic model with realizations x sampled through,

$$z \sim \rho_0, \quad x = T_\theta(z), \quad (1)$$

and associated probability density function given by the change-of-variables formula,

$$\rho_\theta(x) = \rho_0 \left(T_\theta^{-1}(x) \right) \left| \det \nabla \left(T_\theta^{-1}(x) \right) \right|. \quad (2)$$

According to Brenier’s theorem [Bre91], provided that ρ_0 is absolutely continuous with respect to the Lebesgue measure, there exists, for any absolutely continuous target distribution π , a transport map T^* that pushes forward ρ_0 to π . This theoretical result justifies the use of transport maps as generative models, as they are, in principle, capable of representing arbitrary target distributions. In practice, however, the expressiveness of the model depends on the chosen parametrization for T_θ .

The central challenge is therefore to design parametrized maps T_θ that are sufficiently expressive while allowing efficient evaluation of the inverse map and the determinant of its Jacobian. To this end, normalizing flows are typically constructed as compositions of simpler transformations

$$T_\theta = T_{\theta_L} \circ \dots \circ T_{\theta_1}, \quad (3)$$

each T_{θ_i} interpreted as a layer involving learnable neural networks and L being the number of layers, or depth, of the NF. The so-called coupling layer trick is behind many popular architectures for these simple invertible transformations, with prominent examples including Real Non-Volume Preserving flows (RealNVP) [DSB17] and Neural Spline Flows (NSF) [Dur+19]. Comprehensive reviews of design choices for NFs can be found in [Pap+21; KPB21].

Owing to their tractable probability density, NFs can be trained for density estimation via maximum likelihood. Given a dataset of training samples $\{x^{(k)}\}_{k=1}^N$ assumed to be i.i.d. from an

unknown target probability distribution π , one learns parameters θ of the NF by minimizing the negative log-likelihood of the training set,

$$\theta \in \arg \min - \sum_{k=1}^N \log \rho_{\theta}(x^{(k)}). \quad (4)$$

While extremely convenient for density estimation, the NF architecture described above based on a finite composition of simple invertible layers can prove too restrictive when the target distribution is highly complex and multimodal. To overcome this issue, more recent models rely on continuous-time dynamics to define the transport map. By opposition to this class of models, the original NF architecture are sometimes referred to as *discrete-time* normalizing flows.

2.1.2 Continuous normalizing flows

Chen and co-authors introduced neural ordinary differential equations (NeuralODEs) [Che+18]. In this framework, samples from the base distribution are transported by the integration of a learnable velocity field $v_{\theta} : \mathcal{X} \times [0, T] \rightarrow \mathcal{X}$, from time $t = 0$ up to time $t = T$,

$$\frac{dX_t}{dt} = v_{\theta}(X_t, t), \quad X_0 \sim \rho_0. \quad (5)$$

The samples produced by this continuous normalizing flow (CNF) correspond to the final states X_T . Denoting by $\rho_{\theta,t}$ the marginal distribution of the process $(X_t)_{t \in [0, T]}$ defined above, the associated continuity equation reads

$$\frac{\partial \rho_{\theta,t}(x)}{\partial t} + \nabla \cdot (\rho_{\theta,t}(x) v_{\theta}(x, t)) = 0. \quad (6)$$

Thus, the probability density function of the generated samples can be computed by integrating, along the flow trajectory, the instantaneous change-of-variables formula derived from the continuity equation:

$$\frac{d \log \rho_{\theta,t}(X_t)}{dt} = \frac{\partial \log \rho_{\theta,t}(X_t)}{\partial t} + v_{\theta}(X_t, t) \cdot \nabla \log \rho_{\theta,t}(X_t) = -\nabla \cdot v_{\theta}(X_t, t), \quad (7)$$

$$\Rightarrow \log \rho_{\theta,T}(X_T) = \log \rho_{\theta,0}(X_0) - \int_0^T \nabla \cdot v_{\theta}(X_t, t), dt. \quad (8)$$

In practice, this formula is evaluated either along a forward trajectory during sampling (increasing time t from $X_0 \sim \rho_0$), or along a backward trajectory when estimating the likelihood of samples from a training set (decreasing time t from $X_T = X^{(k)}$).

With CNFs, both sampling and likelihood evaluation require the numerical integration of an ordinary differential equation (ODE), which is computationally costly and introduces discretization errors. In particular, likelihood computations involve evaluating the divergence of the velocity field, which can be obtained via automatic differentiation but further increases the computational cost.² Moreover, training CNFs by maximum likelihood requires the use of an adjoint method to differentiate through the trajectory followed along the likelihood integration (eq. (8)), which itself depends on the velocity field and its parameters [Che+18]. Despite these disadvantages, CNFs were shown to be remarkably expressive compared to discrete-time NFs and represent a

²The divergence $\nabla \cdot v(x)$ is the trace of the Jacobian $\text{Tr}(\nabla v(x))$. To avoid forming the full Jacobian requiring d backpropagation passes, Hutchinson’s estimator approximates this trace by using the identity $\text{Tr}(A) = \mathbb{E}_{\xi}[\xi^{\top} A \xi]$, where ξ is a random vector with zero mean and identity covariance (e.g. standard Gaussian or Rademacher) [Hut89; Gra+18]. In practice, this reduces the computation to Jacobian-vector products, which can be evaluated efficiently by one backpropagation pass. However, this stochastic approximation of the divergence only yields a stochastic approximation of the likelihood.

conceptual step toward the highly successful models discussed next—diffusion models and flow matching—which also benefit from the same underlying principle.

In the context of generative modeling for physical systems, continuous-time generative models offer an additional, non-negligible advantage: they enable the design of probabilistic models whose densities respect the invariances of the problem at hand, such as rotational or translational invariances. To describe this concisely, consider a symmetry group $\mathcal{G}$ leaving the target distribution π invariant, i.e. $\forall g \in \mathcal{G}, \pi(g(x)) = \pi(x)$. One can show that using an invariant base measure ρ_0 together with an equivariant velocity field, i.e. satisfying

$$\forall x \in \mathcal{X}, \rho_0(g(x)) = \rho_0(x), \quad \text{and} \quad \forall x \in \mathcal{X}, v_\theta(g(x), t) = g(v_\theta(x, t)), \quad (9)$$

yields an invariant marginal probability density $\rho_{\theta,t}$ at all times [KKN20; Sat+21]. Such equivariant architectures can be constructed either explicitly, for instance as gradients of invariant functions [KKN20], or by relying on dedicated libraries [SHW21; GS22].

By contrast, obtaining invariant densities with discrete-time normalizing flows is considerably more difficult. In principle, one could combine an invariant base distribution with an equivariant transport map, but this would require layers that are simultaneously invertible, equipped with a tractable Jacobian determinant, and equivariant. The coupling-layer constructions typically used to satisfy the first two requirements make it non-trivial to enforce the third.³

Diffusion models and stochastic interpolants, which we review next, inherit the flexibility and symmetry compatibility of CNFs, yet sidestep the computational burden associated with integrating the instantaneous change-of-variables formula, thanks to training objectives that differ from direct likelihood maximization.

2.2 Diffusion models

CNFs trained by maximum likelihood encourage the velocity field to match the model final time marginal $\rho_{\theta,T}$ to the data target distribution, without imposing constraints on the intermediate distributions. In contrast, diffusion models (DMs) enforce the intermediate distributions to follow a prescribed noising-denoising path. For simplicity, we consider here the case $\mathcal{X} = \mathbb{R}^d$ and an Ornstein-Uhlenbeck (OU) noising process defined by the stochastic differential equation (SDE)

$$d\tilde{X}_\tau = -\tilde{X}_\tau d\tau + \sqrt{2} d\tilde{W}_\tau, \quad \tilde{X}_0 \sim \pi, \quad (10)$$

where $\tilde{W}_\tau$ is a standard Brownian motion and π denotes the unknown target distribution from which the training set is sampled. In what follows, the marginal distribution of the noising process at time τ is denoted π_τ . The forward OU process — equivalently, an overdamped Langevin equation with quadratic confining potential — admits the Gaussian $\mathcal{N}(0_d, I_d)$ as unique invariant measure and converges to it at an exponential rate. Consequently, one typically assumes that $\pi_T \approx \mathcal{N}(0_d, I_d)$ for $T \gg 1$.

Anderson [And82] established that the process defined in backward time $t = T - \tau$ by

$$dX_t = (X_t + 2\nabla \log \pi_{T-t}(X_t)) dt + \sqrt{2} dW_t, \quad X_0 \sim \mathcal{N}(0_d, I_d), \quad (11)$$

admits π_{T-t} as its marginal density at time t , which can be readily verified using the Fokker-Planck equation. In other words, this process has the same marginal distributions as the noising process but evolves backward in time, and is therefore referred to as the denoising process.

³One possible workaround is a lifting strategy in which the state space is augmented, allowing the construction of equivariant architectures through finite compositions of coupling layers [Mid+23; SNR25]. This approach, however, comes at the cost of losing access to the exact likelihood on the original space.

As originally proposed by [Soh+15], a generative model can be constructed from this noising-denoising framework by running the backward denoising process starting from samples drawn from the limiting distribution of the noising process, playing the role of the base distribution—here the standard Gaussian $\rho_0 = \mathcal{N}(0_d, I_d)$. However, the denoising process eq. (11) involves the gradient of the logarithm of the time-marginal density, known as the *score*. Let’s examine the score closely.

The conditional distribution of the noising process at time τ given the initial state x_0 , denoted $p_{\tau|0}(\cdot | x_0)$, is Gaussian, yielding the time marginal

$$\pi_\tau(x) = \int dx_0 p_{\tau|0}(x|x_0) \pi(x_0) = \int dx_0 \mathcal{N}(x; x_0 e^{-\tau}; (1 - e^{-2\tau})I_d) \pi(x_0), \quad (12)$$

where $\mathcal{N}(\cdot; m; \Sigma)$ denotes the density of a Gaussian distribution with mean m and covariance Σ . From here, the score can be written as a conditional expectation over the initial state, using Bayes formula to recognize the conditional measure:

$$\begin{aligned} \nabla \log \pi_\tau(x) &= -\frac{1}{\pi_\tau(x)} \int dx_0 \frac{x - x_0 e^{-\tau}}{1 - e^{-2\tau}} \mathcal{N}(x; x_0 e^{-\tau}; (1 - e^{-2\tau})I_d) \pi(x_0) \\ &= -\int dx_0 \frac{x - x_0 e^{-\tau}}{1 - e^{-2\tau}} \frac{p_{\tau|0}(x | x_0) \pi(x_0)}{\pi_\tau(x)} \\ &= -\mathbb{E}_{x_0} \left[\frac{x - x_0 e^{-\tau}}{1 - e^{-2\tau}} \middle| \tilde{X}_\tau = x \right]. \end{aligned} \quad (13)$$

As a result, the score is doubly intractable: in generative modelling tasks, the target distribution π is unknown and only accessible through an empirical training dataset. Additionally, even in sampling applications when π is known, the conditional expectation remains a high-dimensional integral without a closed-form expression.

In generative modeling applications, however, the score can be regressed directly from training data, typically using a neural network parametrization $s_\theta : \mathcal{X} \times [0, T] \rightarrow \mathcal{X}$. Its expression as a conditional expectation motivates the denoising score matching objective introduced in [Son+20; HJA20], which takes the form of a mean-squared error at a fixed time τ ,

$$\min_{\theta} \mathbb{E}_{x_0} \left[\mathbb{E}_{x|x_0} \left[\left\| s_\theta(x, \tau) + \frac{x - x_0 e^{-\tau}}{1 - e^{-2\tau}} \right\|^2 \right] \right], \quad (14)$$

where the reversed order of the expectations compared to eq. (13) stems from the law of total expectation and reflects the practical implementation of the loss. Indeed, an empirical approximation of this loss is constructed using samples x_0 from the training set and samples of the form $\tilde{X}_\tau = x_0 e^{-\tau} + \sqrt{1 - e^{-2\tau}} \xi$ with $\xi \sim \mathcal{N}(0_d, I_d)$ drawn from the conditional distribution of the noising process $p_{\tau|0}(\cdot | x_0)$. A weighted integral of this loss over $\tau \in [0, T]$ is commonly employed in practice, enabling the simultaneous training of the score across all time marginals.

Presented here with the canonical illustration of the OU process, this noising-denoising construction admits many natural generalizations. Different forward diffusions can be considered, for instance the variance-exploding SDE, which relies on a purely diffusive process without drift, in contrast to the variance-preserving OU dynamics. We refer in particular to [Kar+22b], which clarifies several important design choices underlying the practical success of diffusion models, including noise scheduling, discretization schemes and neural architectures used to parametrize the score.

Beyond modifications of the forward diffusion, alternative noising processes can also be employed. In particular, Schrödinger bridges — which achieve convergence in finite time and relax the

assumption of a Gaussian base distribution — can be leveraged within the same noising-denoising paradigm for generative modelling [De +21], even if they remain comparatively less explored in practice. Motivated by this idea of finite-time constructions and increased flexibility, another class of models — stochastic interpolants or flow matchings — has recently attracted attention in the generative modelling community. They offer a flexible framework that interpolates between prescribed endpoint distributions and will be the focus of the next section.

2.3 Stochastic interpolants a.k.a. flow matchings

The idea of constraining the path of intermediate distributions, has inspired the development of another class of transport-based continuous-time generative models, proposed independently in [AV23; LGL23; Lip+23]. Here, we adopt the most general formulation of stochastic interpolants (SIs) introduced in [ABV23].

While the base distribution of diffusion models is necessarily given by the limiting distribution of a noising process, the base distribution of stochastic interpolants (SIs), denoted ρ_0 in the following, can be any distribution that is easy to sample from. The bridge of distributions is no longer defined through an SDE, but, instead, constructed using an interpolation function $I_t : [0, 1] \times \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{X}$ together with a positive scalar function $\gamma : [0, 1] \rightarrow \mathbb{R}^+$,

$$X_t = I_t(X_0, X_1) + \gamma(t)Z, \quad X_0 \sim \rho_0, \quad X_1 \sim \pi, \quad Z \sim \mathcal{N}(0_d, I_d), \quad (15)$$

Crucially, the interpolation function satisfies $I_0(x_0, x_1) = x_0$ and $I_1(x_0, x_1) = x_1$, and the noise amplitude vanishes at the extremities, $\gamma(0) = \gamma(1) = 0$, together with suitable regularity conditions on both functions (see [ABV23], Definition 2.1).

The marginal distribution of X_t at time $t \in [0, 1]$ is given by

$$\rho_t(x) = \int dx_0 dx_1 dz \rho_0(x_0) \pi(x_1) \mathcal{N}(z; 0, I_d) \delta(x - I_t(x_0, x_1) - \gamma(t)z), \quad (16)$$

where δ denotes the Dirac distribution. Although this marginal density is generally intractable, sampling from the interpolation process is straightforward, as it only requires sampling from the base distribution, the target distribution, and the Gaussian noise.

To construct a generative model, one observes that the velocity field defined as the conditional expectation

$$v^*(x, t) = \mathbb{E}_{x_0, x_1, z} [\partial_t I_t(x_0, x_1) + \dot{\gamma}z \mid X_t = x], \quad (17)$$

where $\dot{\gamma}$ denotes the time derivative of γ , satisfies the continuity equation associated with the interpolating path of distributions eq. (16)

$$\frac{\partial \rho_t(x)}{\partial t} + \nabla \cdot (\rho_t(x) v^*(x, t)) = 0. \quad (18)$$

As a consequence, the deterministic process $(X_t^*)_{t \in [0, 1]}$ defined by the ODE

$$\frac{dX_t^*}{dt} = v^*(X_t^*, t), \quad X_0^* \sim \rho_0, \quad (19)$$

admits the same marginal distributions as the interpolation process at all times. This observation enables the construction of a generative model by learning a neural network approximation $v_\theta : \mathcal{X} \times [0, 1] \rightarrow \mathcal{X}$ of v^* .

As for the score of diffusion models, the conditional-expectation form of the target velocity field leads to a mean-squared regression loss at a fixed time t ,

$$\min_\theta \mathbb{E}_{x_0, x_1, z} \left[\|v_\theta(X_t, t) - (\partial_t I_t(x_0, x_1) + \dot{\gamma}z)\|^2 \mid X_t = I_t(x_0, x_1) + \gamma(t)z \right]. \quad (20)$$

In practice, an empirical approximation of this loss is obtained by drawing samples x_1 from the training set, samples x_0 from the base distribution, Gaussian noise samples z , and averaging the mean squared error over (uniform or weighted) samples of time t in $[0, 1]$.

Once the velocity field has been learned, new samples are generated by integrating the ODE driven by v_θ from samples of the base distribution ρ_0 up to time $t = 1$. This sampling procedure boils down to the sampling procedure of CNFs, with the key difference that the velocity field is explicitly trained to match a prescribed interpolation path of distributions. Avoiding the maximum likelihood training of CNFs, this approach is significantly more efficient computationally. The generation process can also be generalized to a stochastic differential equation, at the cost of additionally regressing the score of the time marginals (see [AV23], Corollary 2.9). A commonly used choice for the interpolation process is the noiseless linear interpolation $I_t(x_0, x_1) = (1 - t)x_0 + tx_1$, as proposed in [LGL23; Lip+23] who introduced the denomination of flow matching for this particular case.

*

With this, we reviewed the main transport based generative models, from the original discrete-time normalizing flows to the most recently introduced stochastic interpolants. The notations and mathematical tools introduced in this section will be used throughout the rest of this manuscript, as we now turn to the application of these models to sampling problems.

3 Exact samplers with normalizing flows

Normalizing flows were the first generative models to be repurposed from generative modelling to the task of sampling. As we will see below, their tractable likelihood allows to use them as helpers in a variety of Monte Carlo scheme in the place of traditional simple probabilistic models such as Gaussians or factorized distributions. We review below the main proposed strategies to leverage NFs for sampling, and discuss their strengths and limitations. While our discussion will focus on continuous state spaces, we note that many of the described methods can be adapted to discrete state spaces by replacing NFs with autoregressive models [Uri+16], sharing the same tractable likelihood property.

In all described approaches, the first step is to train a flow model to approximate the target distribution of interest π on the state space $\mathcal{X}$. While this training is not trivial given the absence of direct samples from the target a priori, we postpone the discussion on training strategies to the next section. We assume access to a NF built from a base ρ_0 and a transport map T_θ such that the pushforward distribution ρ_θ (given by eq. (2)) is, loosely speaking, close to π .

Note that even if the agreement between ρ_θ and π is imperfect — as is inevitably the case in practice due to approximation and optimization errors — the samplers described below remain exact. In other words, they generate samples from π in the limit of large sample sizes or many iterations. This makes the present application of deep learning to scientific computing rather distinctive: the unavoidable imperfections introduced during the learning stage do not necessarily translate into biased results, provided that the subsequent sampling procedure is properly designed.

Below, we first describe a set of simple approaches originally proposed to leverage NFs for sampling (section 3.1), and then we discuss the advantages and limitations of these approaches in a systematic benchmark (section 3.2). We then review more recent approaches looking to push further the abilities of NF-powered samplers (sections 3.3 and 3.4).

3.1 Initial NF-powered samplers

3.1.1 Importance sampling

Elementary algorithm A simple way to leverage NFs for sampling is to use them as proposal distributions in importance sampling (IS). For a target distribution π , known up to a normalization constant, the IS estimation of the expectation of an observable $f : \mathcal{X} \rightarrow \mathbb{R}$ follows the steps:

algorithm 1: **Neural importance sampler (Neural-IS)**

```

for  $i = 1 \dots N$  do
    Sample the proposal distribution  $x_i \sim \rho_\theta(\cdot)$ ,
    Compute weight  $w_i = \frac{\pi(x_i)}{\rho_\theta(x_i)}$  (up to a multiplicative constant),
for  $i = 1 \dots N$  do
    Compute normalized weight  $\hat{w}_i = w_i / \sum_{j=1}^N w_j$ ,
return self-normalized IS estimator  $\hat{f}_{\text{SNIS}} = \sum_{i=1}^N \hat{w}_i f(x_i)$ 

```

A practical IS proposal needs (i) to be straightforward to sample from to obtain an extensive set of N i.i.d. samples easily and (ii) to have a tractable probability density function to compute weights, both conditions being satisfied by NFs.

Note that if the normalization constant of π is unknown, the IS weights can only be computed up to a multiplicative constant which however cancels in the self-normalized weights and hence has no impact on the self-normalized IS estimator $\hat{f}_{\text{SNIS}}$ returned by the algorithm. Conversely, were

the normalization constant of π known, one could use the plain IS estimator $\hat{f}_{\text{IS}} = \frac{1}{N} \sum_{i=1}^N w_i f(x_i)$ which is unbiased, unlike the self-normalized version.

Under mild conditions on the proposal distribution, $\hat{f}_{\text{SNIS}}$ is a consistent estimator of $\mathbb{E}_{\pi}[f]$ in the sense that it converges (at least in probability) to $\mathbb{E}_{\pi}[f]$ as $N \rightarrow \infty$. However, at finite N , the agreement of the proposal distribution ρ_{θ} with the target π directly impacts the quality of the IS estimator. In particular, the variance of the IS weights is minimized when ρ_{θ} is equal to π , yielding normalized weights uniformly equal to $1/N$. This observation justifies the relevance of leveraging recent progress in deep generative models, in the form of NFs, to build accurate proposals in place of usual simple tractable probabilistic models (e.g. Gaussian or factorized). In turn, the quality of the IS estimator is directly related to the quality of the NF approximation of the target distribution. Pioneering works in this direction include [Mül+19; Noé+19; Nic+20; Nic+21].

Free energy computations Building on similar ideas of IS, one can also use NFs to estimate the normalization constants, or free energies in the language of statistical mechanics. The idea is to use the NF as a reference state of known normalization constant and estimate its ratio with normalization constant of the target π using either Target Free Energy Perturbation (TFEP) [Jar02] or Bennett Acceptance Ratio (BAR) [Ben76; HT09]. This idea was simultaneously proposed with Bayesian inference in mind [JS20], and for applications to atomic solids [Wir+20; Wir+22; WIP23; SNR25]. Computations of protein-ligand binding free energies with NFs were also explored in [DSB17], as well as computations of conformal free energy differences in [Mol+24*; Ole+24].

By convention, the normalization constant of a NF is 1 due to writing ρ_{θ} directly in its normalized form. For $\pi(x) = e^{-U(x)}/\mathcal{Z}_{\pi}$, the TFEP estimator of its free energy $F_{\pi} = -\log \mathcal{Z}_{\pi}$ is build upon the equality

$$\mathcal{Z}_{\pi} = \mathbb{E}_{\rho_{\theta}} \left[e^{-U(x)} / \rho_{\theta}(x) \right] \Rightarrow F_{\pi} = -\log \mathbb{E}_{\rho_{\theta}} \left[e^{-U(x)} / \rho_{\theta}(x) \right] \quad (21)$$

and reads

$$\hat{F}_{\pi}^{\text{TFEP}} = -\log \hat{\mathcal{Z}}_{\pi}^{\text{TFEP}} = -\log \left(\frac{1}{N} \sum_{i=1}^N \frac{e^{-U(x_i)}}{\rho_{\theta}(x_i)} \right), \quad \text{with } x_i \sim \rho_{\theta}(\cdot) \text{ for } i = 1, \dots, N, \quad (22)$$

where one recognizes the IS weights w_i in the summation. This estimator is easy to implement, requiring mere samples from the flow and yields an unbiased estimator $\hat{\mathcal{Z}}_{\pi}^{\text{TFEP}}$ of $\mathcal{Z}_{\pi}$ and a consistent estimator $\hat{F}_{\pi}^{\text{TFEP}}$ of F_{π} as $N \rightarrow \infty$. However, as for IS, it can suffer from large variance at finite N if ρ_{θ} is not close enough to π .

The BAR method improves over TFEP in cases where samples from both ρ_{θ} and π are available [Ben76]. Given two sets of samples $\{x_i\}_{i=1}^N$ from ρ_{θ} and $\{x_j\}_{j=1}^M$ from π , the BAR estimator can be interpreted as the maximum likelihood estimate of the free energy difference obtained by viewing the problem as a binary classification task: given a sample x , one asks whether it was drawn from ρ_{θ} or from π . Using Bayes' rule, the posterior probability that x originates from ρ_{θ} can be written as a logistic function of the energy difference between the two models, involving the unknown free energy difference F_{π} . Maximizing the corresponding likelihood over the two sets of samples leads to a self-consistent equation whose solution defines the BAR estimator [Shi+03]. Namely, the estimator $\hat{F}_{\pi}^{\text{BAR}}$ is defined as the solution of the implicit equation

$$\sum_{i=1}^N \frac{1}{1 + \frac{N}{M} e^{U(x_i) + \log \rho_{\theta}(x_i) - \hat{F}_{\pi}^{\text{BAR}}}} = \sum_{j=1}^M \frac{1}{1 + \frac{M}{N} e^{-U(x_j) - \log \rho_{\theta}(x_j) + \hat{F}_{\pi}^{\text{BAR}}}} \quad (23)$$

This estimator is known to be asymptotically optimal, meaning that, among estimators using samples from both distributions, it has minimal variance in the limit of infinite samples. A derivation specifically in the case of bijective mapping is given in [HT09]. If BAR improves on TFEF, it strictly requires samples from the target distribution π , which are not available a priori in the context discussed here. If some references use approximate samples obtained from molecular dynamics [DZ21], we note that provided a NF approximating π is available, the NF can also be used to converge an MCMC to generate samples from π to be used in BAR [Mol+24*]. The description of NF assisted MCMC schemes is given in the next section.

3.1.2 Independent proposal Markov chains

MCMCs can also greatly benefit from a NF trained to approximate the target distribution π . A simple way to do so is to use the NF as an independent proposal distribution in a Metropolis-Hastings (MH) algorithm [AKS19; No  +19; Nic+20; GRV21*]. The resulting algorithm reads:

algorithm 2: **flow independent Metropolis Hastings (flow-IMH)**

```

Initialize  $x^{(0)}$ ,
for  $k = 1 \dots K$  do
    Sample proposal  $x' \sim \rho_\theta(\cdot)$ ,
    Compute acceptance probability  $\alpha = \min \left( 1, \frac{\pi(x')}{\pi(x^{(k-1)})} \times \frac{\rho_\theta(x^{(k-1)})}{\rho_\theta(x')} \right)$ ,
    Set  $x^{(k)} = x'$  with probability  $\alpha$ , else  $x^{(k)} = x^{(k-1)}$ ,
return samples  $\{x^{(k)}\}_{k=1}^K$ 

```

As in IS, observe that the proposal distribution of a MH algorithm needs to be (i) easy to sample from and (ii) its density needs to be tractable to evaluate the acceptance probability, both conditions being satisfied by NFs.

The resulting independent proposal MCMC is asymptotically unbiased regardless of the quality of the flow approximation, meaning that the distribution of $x^{(K)}$ converges to π as $K \rightarrow \infty$. However, at finite K , the acceptance rate of the algorithm directly depends on the agreement of ρ_θ with π , being equal to 1 when $\rho_\theta = \pi$ and dropping to zero when ρ_θ is a poor approximation of π . A high acceptance rate implies an extremely efficient sampler here as the proposal are *global moves*, uncorrelated with the current state of the chain. Attempting global moves across the state space is made possible by the flexibility of NFs to approximate complex distributions, and represents a key advantage of using NFs as proposal distributions in place of local proposals such as random-walk or Langevin dynamics. In particular, global moves can help escape from metastable states in multimodal distributions, a situation where local MCMC proposals often fail.

Local-global Markov Chains Variants around the independent Metropolis-Hastings presented above can also be of interest. For instance, one can consider a hybrid scheme alternating local MCMC moves (e.g. Langevin dynamics or MH with Gaussian jump proposals) and global NF-based proposals [GRV21*; Sam+22*]. This is particularly useful when the NF misrepresents some part of the target distribution, as the local moves can help improve exploration in such regions. The advantage of alternating local and global moves was theoretically proven for simple statistical physics models with discrete state spaces and therefore using autoregressive networks in place of NFs [DRZ25].

Multi-proposal version Another interesting variant considers more than one proposal per iteration [Sam+22*], following the iterated sampling importance resampling (iSIR) scheme [ADH10]. Given the cheap sampling cost of NFs, one can consider $M > 1$ proposals $\{x'_i\}_{i=1}^M$ sampled from ρ_θ at each iteration of the MCMC yielding the following algorithm:

algorithm 3: **flow iterated sampling importance resampling (flow-iSIR)**

```

Initialize  $x^{(0)}$ ,
for  $k = 1 \dots K$  do
  Set  $x'_0 = x^{(k-1)}$ ,
  for  $i = 1 \dots M$  do
    Sample proposal  $x'_i \sim \rho_\theta(\cdot)$ ,
  for  $i = 0 \dots M$  do
    Compute weight  $w_i = \frac{\pi(x'_i)}{\rho_\theta(x'_i)}$ ,
  for  $i = 0 \dots M$  do
    Compute normalized weight  $\hat{w}_i = w_i / \sum_{j=0}^M w_j$ ,
  Sample index  $i$  from the categorical distribution with probabilities  $\{\hat{w}_j\}_{j=0}^M$ ,
  Set  $x^{(k)} = x'_i$ ,
return samples  $\{x^{(k)}\}_{k=1}^K$ 

```

By considering multiple proposals, this scheme increases the chance to propose a state in high probability regions of the target distribution, improving the mixing of the chain per iteration. However, the computational cost per iteration also increases with M . As a result, a large M is typically advantageous if parallel computing resources are available, even if the overall computational efficiency of the sampler does not necessarily improve with M .

3.1.3 Reparametrization through transport maps

Yet another way to leverage NFs for sampling is to use them as reparametrization maps to simplify the geometry of the target distribution [PM18; Hof+19; Noé+19; CN23; SG23*]. Here, we reference this set of methods as neutra-MCMCs following [Hof+19] play of words on neural transport maps "NeuTra-lizing" the target's bad geometry.

Once a NF transport map T_θ approximating the target distribution π is available, one can consider the pullback distribution of π through T_θ defined as

$$\pi_{T_\theta}(z) = \pi(T_\theta(z)) |\det \nabla_z T_\theta(z)| \quad \text{for } z \in \mathcal{X}. \quad (24)$$

By construction, if ρ_θ is a good approximation of π , then the pullback distribution π_{T_θ} is close to the base distribution ρ_0 , which is usually chosen to be a distribution with easy geometry (e.g. standard Gaussian). Sampling from π can therefore be achieved by first sampling from the simpler distribution π_{T_θ} and then pushing forward the samples through T_θ . For instance, if indeed $\rho_0 = \mathcal{N}(0, I_d)$, following the intuition that $\pi_{T_\theta} \approx \rho_0$, a natural choice is to resort to Langevin dynamics for sampling in the latent space, either with a Metropolis accept reject (MALA for Metropolis adjusted Langevin algorithm) or without (ULA for unadjusted Langevin algorithm). However, all other types of samplers can be employed: [Hof+19] examined Hamiltonian Monte Carlo (HMC), [CN23] elliptical slice sampling and [SG23*] Metropolis-within-Gibbs sampling. Provided an unbiased sampler is used to target $\rho_0 = \mathcal{N}(0, I_d)$, transporting its output through T_θ provides unbiased samples from π . Sketching an algorithm to summarize the overall method yields:

algorithm 4: **neutra-MCMC**

```

Initialize  $z^{(0)}$ ,
for  $k = 1 \dots K$  do
     $z^{(k+1)} \sim M_{T_\theta, \pi}(\cdot | z^{(k)})$  with  $M_{T_\theta, \pi}$  Markov kernel leaving  $\pi_{T_\theta}$  invariant,
     $x^{(k+1)} = T_\theta(z^{(k+1)})$ 
return samples  $\{x^{(k)}\}_{k=1}^K$ 

```

Note that Markov kernels $M_{T_\theta, \pi}(\cdot | z^{(k)})$ based on gradient information of the target distribution – such as MALA, ULA and HMC – require computing gradients of π_{T_θ} at each iteration of the algorithm. This involves automatic differentiation through T_θ making each update significantly more costly than when directly targeting π .

3.2 A first comparison of NF powered samplers

Faced with the variety of ways of easing sampling leveraging a transport map T_θ sending ρ_0 close to the target π , a comparison is in order. A detailed numerical benchmark was conducted in [Gre+23*], considering different sources of difficulties in sampling, namely ill-conditioning, multimodality and high-dimensionality. We rediscuss the results of this study in the following paragraphs.

Neutra-MCMCs are the most robust to ill-conditioning If regions of high-probability of the target have significantly different scales in different directions of the state space, the target distribution is said to be ill-conditioned and this typically causes trouble to the convergence of MCMC based samplers. To assess the ability of the different flow-based approaches to alleviate this issue, one needs to also consider the quality of the training of T_θ . For a perfectly trained flow $\rho_\theta = \pi$, and IS and IMH produce i.i.d samples from π and are therefore optimal. The picture becomes more complicated if a residual mismatch persists between ρ_θ and π . To systematically assess the effect of an imperfect learning on sampling an ill-conditioned target, the following experiment was conducted.

The target distribution is a high-dimensional Gaussian ($d = 128$) with an ill-conditioned covariance (see appendix D.2 of [Gre+23*] for complete experimental setting). An analytical flow is defined with a scalar quality parameter $t \in [0, 1]$ as illustrated on the left panel of fig. 1. For $t = 0.5$ the flow is perfect. For $t = 0$ the flow density ρ_t is isotropic and over-concentrated, with variance matching the smallest directional variance of the target distribution. Conversely, for $t = 1$, ρ_t is isotropic underconcentrated, matching the largest directional variance of the target distribution. When $t \neq 0.5$, the flow fails at perfectly preconditioning the pullback π_{T_t} . Results of the experiment reproduced on the right panel of fig. 1 show that neutra-MCMC methods are more robust to an imperfect map than approaches based on independent proposals (neural-IS, flow-IMH, flow-iSIR) whose performances degrade quickly as the flow training parameter deviates from $t = 0.5$. This advantage is however mitigated with the cost per iteration of the method (higher than for independent proposal methods by a factor of 2 to 30 for the experiments reported in [Gre+23*]) and the challenges posed by multimodality for neutra-MCMCs, as discussed next.

Neutra-MCMCs struggle to mix between modes When the target π is multimodal, the intuition that the pullback π_{T_θ} has a simplified geometry breaks down. The reason is that common NF architecture based on transporting a standard Gaussian ρ through coupling layers have troubles separating the mass of ρ_0 into truly disconnected modes and respectively do not generate pullbacks with reunited mass. The difficulty of NFs to match multimodal distribution has been discussed in [Cor+20; Sal+22]. Figure 2 illustrates well this phenomenon on a $2d$ example.

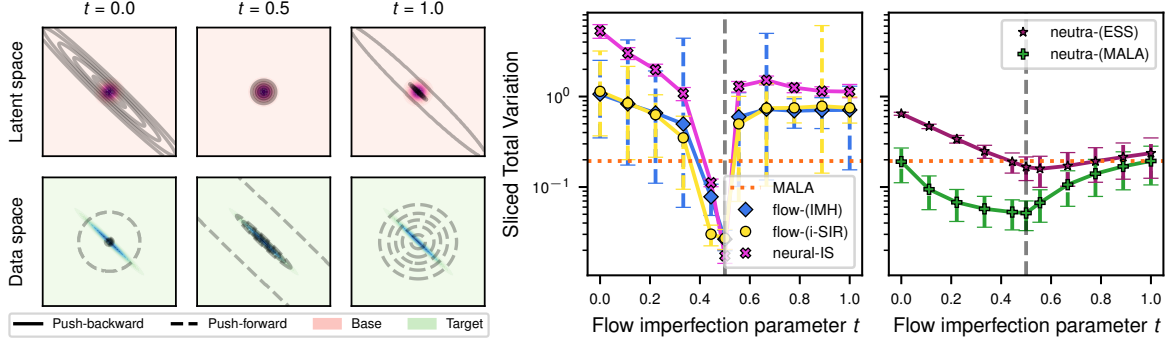


Figure 1: **(Left)** Pullback π_{T_t} and pushforward ρ_t as a function of the flow imperfection parameter t . **(Right)** Sliced TV distances to target of NF-powered Monte Carlo samples compared as a function of the quality of the flow t , using 256 chains of length 1400 initialized with draws from NF with T_t . neural-IS was evaluated with 14000 samples. Results were qualitatively unchanged for $d = 16, 32, 64, 256$. ESS stands for elliptical sliced-sampling in this figure. Reproduced from [Gre+23*].

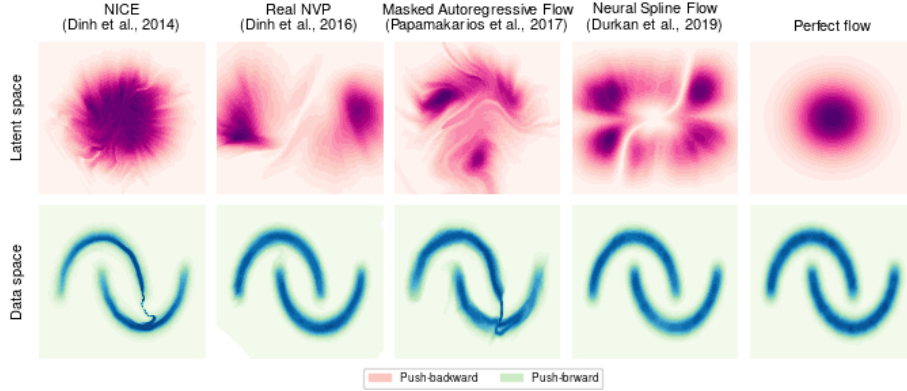


Figure 2: Pullback and pushforward of popular deep normalizing flows architectures targeting the two moons distribution. Reproduced from [Gre+23*]

As a result, neutraMCMCs relying on MCMC kernels struggling with multimodality are not out of trouble by targeting the pullback. In this context, methods relying on independent proposals such as flow-IMH are better off. This can be seen on the results presented on the left and center panel of fig. 3 considering the task of sampling from a Gaussian mixture. In particular, as the dimension increases, methods based on reparametrization increasingly fail to discover all the modes of the distribution despite the fact that the pushforward ρ_θ did have significant mass overlap with all the modes. However, this success of independent proposal methods for multimodal distribution is limited by their sensitivity as the dimension further increases, which we examine next.

Independent proposal methods are most affected by high-dimensionality Methods relying on independent proposals, either in the context of importance sampling (neural-IS) or of MCMCs (flow-IMH and flow-iSIR) suffer similarly from the increase of the dimensionality. A classical elementary calculation by [Aga+17] allows to catch the source of the difficulty considering the example case where the considered target and proposal distributions are the product of $1d$ measures. But before we cover this example, let us define an important quantity to assess the efficiency of IS and IMH that has not yet been introduced. The effective sample size (ESS) [Kis65] is defined, for a batch of N samples $\{x_i\}_{i=1}^N$ drawn from the proposal distribution ρ , with

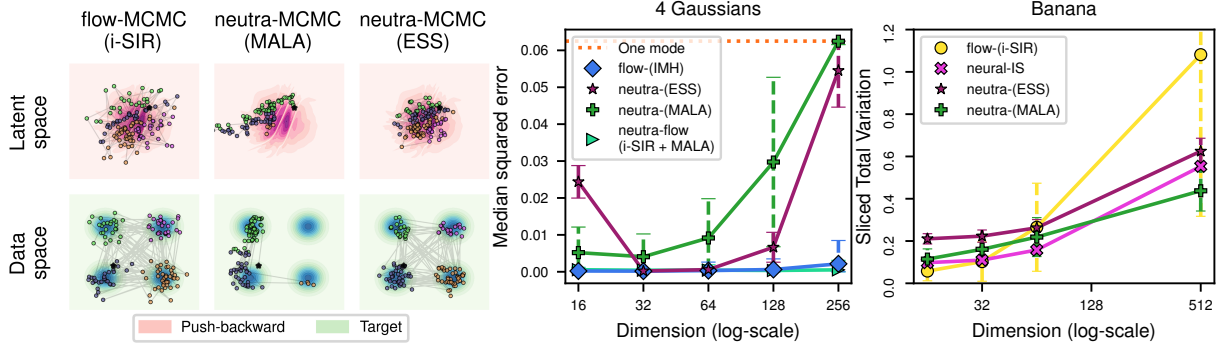


Figure 3: **(Left) Example chains of NF-enhanced walkers with a 2d target mixture of 4 Gaussians.** The 128-step MCMC chain is colored according to the closest mode in the data space (bottom row) with corresponding location in the latent space (top row). The complex geometry of the pullback π_{T_θ} hinders the mixing of local-update algorithms. MALA’s step-size was chosen to reach 75% of acceptance. **(Middle) Median squared error of the histograms of visited modes of 4 Gaussians per chain against the perfect uniform histogram as a function of dimension.** 512 chains of 1000-steps on average were used. **(Right) Sliced total variation in sampling the Banana distribution in increasing dimension** using a RealNVP. 128 chains of 1024-steps were used. Reproduced from [Gre+23*]

importance weights $w(x_i) = \pi(x_i)/\rho(x_i)$, as

$$\text{ESS} = \frac{\left(\sum_{i=1}^N w(x_i)\right)^2}{\sum_{i=1}^N w(x_i)^2}. \quad (25)$$

If all the weights are equal, meaning that the proposal perfectly matches the target, then $\text{ESS} = N$. Conversely, if one weight dominates all the others, then $\text{ESS} \approx 1$. The ESS therefore provides a measure of the number of independent samples from the target distribution that can be extracted from the weighted samples drawn from the proposal. Note that both unnormalized and self-normalized IS weights can be used to compute the ESS, as the normalization constant cancels out in the ratio.

Going back to the example of product of $1d$ distributions, consider target and proposal distributions defined on $\mathcal{X} = \mathbb{R}^d$ as

$$\pi(x) = \prod_{i=1}^d \pi_{1d}(x_i) \quad \text{and} \quad \rho(x) = \prod_{i=1}^d \rho_{1d}(x_i). \quad (26)$$

Then the importance weight function itself factorizes $w(x) = \prod_{i=1}^d w_{1d}(x_i)$ with $w_{1d}(x_i) = \pi_{1d}(x_i)/\rho_{1d}(x_i)$. In the limit of large N , thanks to the independence of all the coordinates, the ESS can be approximated as

$$\text{ESS} \approx \frac{(N \mathbb{E}_\rho[w_{1d}]^d)^2}{N \mathbb{E}_\rho[w_{1d}^2]^d} = N \left(\frac{\mathbb{E}_{\rho_{1d}}[w_{1d}]^2}{\mathbb{E}_{\rho_{1d}}[w_{1d}^2]} \right)^d \approx N e^{-d \log(\mathbb{E}_{\rho_{1d}}[w_{1d}]^2 / \mathbb{E}_{\rho_{1d}}[w_{1d}^2])} \quad (27)$$

where the logarithm in the exponent is positive unless the proposal perfectly matches the target in $1d$ in which case it is zero. This calculation shows that unless a perfect match is achieved, the ESS decays exponentially fast with the dimension d . Although this calculation is made under restrictive assumptions, it captures well the curse of dimensionality affecting IS and IMH methods. For instance, note that in statistical mechanics applications of sampling, the target distribution takes the form of a Boltzmann-Gibbs distribution $\pi(x) \propto e^{-U(x)}$ where the potential

energy $U(x)$ is often extensive in the number of degrees of freedom, meaning scaling linearly with d . In such situations, even small relative errors in the approximation of $U(x)$ by the NF transport map T_θ can lead to exponentially small ESS as d increases.

The right panel of fig. 3 illustrates this curse of dimensionality affecting independent proposal methods, considering the task of sampling from a banana-shaped distribution in increasing dimension. Similar results have been reported by other works from which the systematic study of [DRW21] stands out. This work focuses on sampling from the Boltzmann distribution of a lattice field theory, the 2d ϕ^4 model, of which the size of the lattice directly controls the dimension of the state space. Additionally, a temperature parameter controls the difficulty of the sampling task, the model going through a phase transition where the correlation length between lattice sites diverges, making the sampling particularly challenging. Despite an extensive hyperparameter search and the use of advanced NF architectures, the authors find that the acceptance rate of flow-IMH samplers decreases invariably as the lattice size increases and as the temperature approaches the critical point.

Where the elementary NF enhanced samplers stand? Summarizing the above discussion, we can sketch the state of affairs after initial research efforts on leveraging NFs for sampling. Independent proposal methods such as neural-IS and flow-IMH are well suited for sampling multimodal distributions thanks to their ability to explore globally the state space. However, they suffer from the curse of dimensionality, with performances degrading quickly as the dimension of the state space increases, unless a near perfect flow approximation is available. To provide an order of magnitude, the highest dimension admissible for independent proposals lies somewhere between a hundred to a few thousands depending on the regularity of the target distribution. This range covers many practical applications such as Bayesian inference for medium size models or molecular systems with a few tens of atoms, but falls short for larger applications such as simulating liquids, large biomolecules or statistical mechanics field theories toward the continuum limit.

Conversely, neutra-MCMCs leveraging local samplers in the latent space are more robust to high-dimensional settings and ill-conditioning. However, they struggle to mix between modes when the target distribution is multimodal, as the pullback distribution often retains a complex geometry hindering the convergence of local samplers.

Taken together, these observations motivate the development of more advanced NF-powered samplers that leverage modern advances in the sampling literature, which has progressed considerably beyond the elementary tools of IS and MH. In the next two sections we review how NF can be combined with the ideas of annealing and of exploiting coarse-grainings to enhance sampling.

3.3 Combining NFs with tempering

Recall that annealed based samplers, such as AIS, PT and SMC already mentioned in the introduction, rely on a bridging path between a simple reference measure and the target measure. The philosophy behind combining tempering with NFs is no longer to expect that a single NF transport map can directly reach the target distribution, but instead to delegate part of the sampling effort to tempering-based schemes. Below, we highlight several key examples, without attempting an exhaustive survey given the rapid pace of recent developments in the literature.

3.3.1 Adapting the base distribution to the target

A first step in the direction of tempering is to adapt the base distribution ρ_0 of the NF to be closer to the target distribution π , easing the training of the NF and improving its final accuracy.

The choice of ρ_0 is highly task dependent as it needs to remain relatively easy to sample. The idea was used on statistical physics fields systems using a Gaussian base with matching small scale correlations [GRV22*], in sampling liquids using a less dense and/or higher temperature fluid as base [Sch+24; JBB24; Cor+25], as well as for sampling crystals and materials using Einstein’s harmonic crystal as base [Wir+20; Wir+22; SNR25]. This is a natural way to incorporate physics knowledge into NF-based sampling schemes that can considerably ease the learning task. It can be further combined with tempering schemes described next.

3.3.2 Combining NFs with AIS: stochastic normalizing flows

[WKN20] introduced a stochastic type of NFs amenable to enhanced sampling using the AIS framework. Instead of relying on a deterministic transport map T_θ , the flow alternates deterministic NF layers and stochastic Markovian layers. The advantage of stochastic NFs lies in the ability of stochastic layers to represent a multimodal target distribution given an unimodal base distribution ρ_0 , which traditional deterministic coupling layers struggle to achieve.

A stochastic NF with L layers is defined as a sequence of random variables $(x^0, \dots, x^L)$ with $x^0 \sim \rho_0$ and x^L being the output of the flow. The transition from layer $\ell - 1$ to layer ℓ is defined by a transition kernel $q_{\theta_\ell}(x^\ell | x^{\ell-1})$ parametrized by θ_ℓ . Deterministic NF layers are implemented via a bijective map T_{θ_ℓ} such that $q_{\theta_\ell}(x^\ell | x^{\ell-1}) = \delta(x^\ell - T_{\theta_\ell}(x^{\ell-1}))$. Stochastic layers are characterized through a $q_{\theta_\ell}(x^\ell | x^{\ell-1})$ with explicit probability density and easy to sample from. The overall joint probability density over $(x_0, \dots, x_L)$ and the marginal output distribution read respectively:

$$\rho_\theta(x^0, \dots, x^L) = \rho_0(x^0) \prod_{\ell=1}^L q_{\theta_\ell}(x^\ell | x^{\ell-1}), \quad \rho_\theta(x^L) = \int \prod_{\ell=0}^{L-1} dx^\ell \rho_\theta(x^0, \dots, x^L). \quad (28)$$

The marginal output distribution $\rho_\theta(x^L)$ is intractable as it involves marginalizing over all stochastic layers. As such, proposals from the stochastic NF cannot be used directly in Neural-IS (algorithm 1) or flow-IMH (algorithm 2) to estimate expectations under the target distribution π . Resorting instead to AIS-like scheme, a set of backward transition kernels needs to be introduced. The backward kernels, with parameters $\tilde{\theta}_\ell$, are denoted by $\tilde{q}_{\tilde{\theta}_\ell}(x^{\ell-1} | x^\ell)$ and allow to define a backward joint distribution with π as last marginal:

$$\tilde{\rho}_{\tilde{\theta}}(x^0, \dots, x^L) = \pi(x^L) \prod_{\ell=1}^L \tilde{q}_{\tilde{\theta}_\ell}(x^{\ell-1} | x^\ell), \quad \pi(x^L) = \int \prod_{\ell=1}^L dx^{\ell-1} \tilde{\rho}_{\tilde{\theta}}(x^0, \dots, x^L). \quad (29)$$

An importance sampling estimator is constructed at the level of the joint distributions based on the equality

$$\mathbb{E}_\pi[f(x)] = \mathbb{E}_{\rho_\theta} \left[f(x^L) \times w(x^0, \dots, x^L) \right], \quad \text{with } w(x^0, \dots, x^L) = \frac{\tilde{\rho}_{\tilde{\theta}}(x^0, \dots, x^L)}{\rho_\theta(x^0, \dots, x^L)}, \quad (30)$$

for any test function $f : \mathcal{X} \rightarrow \mathbb{R}$. The resulting self-normalized importance sampling estimator of $\mathbb{E}_\pi[f(x)]$ reads

$$\hat{f}_{\text{SNF}} = \frac{\sum_{i=1}^N f(x_i^L) w(x_i^0, \dots, x_i^L)}{\sum_{i=1}^N w(x_i^0, \dots, x_i^L)} \quad \text{with } (x_i^0, \dots, x_i^L) \sim \rho_\theta(\cdot). \quad (31)$$

The efficiency of the resulting estimator relies on the agreement between the forward and backward joint distributions ρ_θ and $\tilde{\rho}_{\tilde{\theta}}$, which in turn depends on the specification of the forward and backward kernels. In practice, [WKN20] suggests designing deterministic and stochastic layers such that the marginal distributions at the successive layers follow as closely as

possible an interpolation path between the base distribution ρ_0 and the target distribution π (e.g. $\pi_\ell = \rho_0^{(L-\ell)/L} \pi^{\ell/L}$), as traditionally done in AIS.

This approach has also strong connections with non-equilibrium statistical mechanics equalities by Jarzynski and Crooks [Jar97; Cro98], and their repurposing for Monte Carlo [Nil+11]. The advantage of learning NF layers on top of vanilla non-equilibrium Monte Carlo moves was validated in particular in high-energy physics applications [Cas+22; CCN24; CCN25]. In the next section, we discuss a very related line of work combining NFs with SMC.

3.3.3 Combining NFs with SMC

The additional ingredient of SMC methods compared to AIS is the intermediate resampling among parallel draws (referred to as particles). For this, SMC relies on a sequence of intermediate distributions $(\pi_\ell)_{\ell=0}^L$ with $\pi_0 = \rho_0$ and $\pi_L = \pi$ and $\pi_\ell = e^{-U_\ell}/Z_\ell$. The resampling steps aim at guaranteeing a good agreement between the distribution of the particles and the current target distribution π_ℓ .

Combining SMC with NFS, [AMD21; Mat+22] assume to have a set of NF transport maps $(T_{\theta_\ell})_{\ell=1}^L$ respectively trained to approximately transport π_ℓ to $\pi_{\ell+1}$. The SMC procedures then combines these maps, with the standard SMC steps of mutation, weighting and resampling. The algorithm can be summarized as follows:

algorithm 5: **(Continuously Repeated) Annealed Flow Transport (AFT/CRAFT)**

```

Initialize particles  $\{x_i^{(0)}\}_{i=1}^N \sim \pi_0$ ,
Initialize weights  $\hat{w}_i^{(0)} = 1/N$  for  $i = 1 \dots N$ ,
Initialize normalization constant of  $\rho_0$   $\hat{Z}_0 = 1$ ,
for  $\ell = 1 \dots L$  do
    for  $i = 1 \dots N$  do
        Transport particles  $y_i^{(\ell)} = T_{\theta_\ell}(x_i^{(\ell-1)})$ ,
        Update weights  $w_i^{(\ell)} = \hat{w}_i^{(\ell-1)} \left| \det \nabla_x T_{\theta_\ell}(x_i^{(\ell-1)}) \right| e^{-U_\ell(y_i^{(\ell)})} / e^{-U_{\ell-1}(x_i^{(\ell-1)})}$ ,
    Estimate normalization constant  $\hat{Z}_\ell = \hat{Z}_{\ell-1} \sum_{i=1}^N w_i^{(\ell)}$ ,
    Normalize weights  $\hat{w}_i^{(\ell)} = w_i^{(\ell)} / \sum_{j=1}^N w_j^{(\ell)}$  for  $i = 1 \dots N$ ,
    Compute the effective sample size  $\text{ESS}^{(\ell)}$  for  $\{\hat{w}_i^{(\ell)}\}_{i=1}^N$  as defined in eq. (25)
    if  $\text{ESS}^{(\ell)} < \alpha N$  then
        for  $i = 1 \dots N$  do
            Resample index  $a_i$  from categorical distribution with probabilities  $\{\hat{w}_j^{(\ell)}\}_{j=1}^N$ 
            Reset weight  $\hat{w}_i^{(\ell)} = 1/N$ 
        for  $i = 1 \dots N$  do
             $\tilde{y}_i^{(\ell)} = y_{a_i}^{(\ell)}$ ,
            “Mutate”  $x_i^{(\ell)} \sim M_\ell(\cdot | \tilde{y}_i^{(\ell)})$  with  $M_\ell$  a  $\pi_\ell$ -invariant Markov kernel,
    return particles  $\{x_i^{(L)}\}_{i=1}^N$  and normalization constant estimate  $\hat{Z}_L$ .
```

As is now common practice in SMC, the resampling step is triggered only when the effective sample size (ESS) drops below a threshold α (e.g. $\alpha = 0.5$). Also note that the method allows to estimate the normalizing constant of the target distribution π as a by-product (which is also the case for AIS but was omitted in the previous section for brevity).

By dividing the task of sampling from π into smaller subtasks of sampling sequentially from successive intermediate distributions π_ℓ , AFT/CRAFT eases the learning of each transport map T_{θ_ℓ} compared to elementary NF-powered samplers. Additionally, the resampling step helps focus computational efforts on particles in high probability regions of the target distribution, which is

particularly useful when the target is multimodal. Finally, the mutation step relying on local MCMC kernels M_ℓ helps improve the diversity of the particles and correct for residual errors in the flow approximations. While combining all these ingredients increases the computational cost of the sampler compared to flow-IMH or Neural-IS, it also logically improves its robustness to challenging sampling tasks. Compared to regular SMC, the addition of transport maps allows to use a coarser interpolation ladder: the requirement that $\pi_{\ell-1}$ and π_ℓ be close enough, to avoid weights degeneracy, becomes that the pushforward of $\pi_{\ell-1}$ through T_ℓ be close enough to π_ℓ . As a result, fewer intermediate steps are necessary to reach a good sampling performance.

In a related proposition, [Kar+22a] combines NFs with SMC using a single transport map that is trained progressively to transport the base distribution ρ_0 to the current intermediate target π_ℓ at each iteration of SMC. In this work, the mutation step takes also advantage of the NF transport map to employ a neutra-MCMC kernel as described in algorithm 4. The performance of this approach was validated on cosmological and astronomical Bayesian inference tasks.

3.3.4 Combining NFs with PT

Following the propositions made for AIS and SMC, NFs can also be used to ease sampling in the PT framework. As a reminder PT relies on a similar interpolation path $(x_\ell)_{\ell=0}^L$ and consists in simulating in parallel a set of $L + 1$ replicas $(x_\ell)_{\ell=0}^L$ sampling respectively from the intermediate distribution π_ℓ . Periodic swaps are proposed between replicas at adjacent levels ℓ and $\ell + 1$. To ensure a reasonable acceptance rate for swap proposals, neighboring distributions π_ℓ and $\pi_{\ell+1}$ must be close enough, which often leads to the necessity of maintaining many replicas at high computational cost.

In this context, [Inv+22] proposed to leverage NFs to skip all the intermediate steps altogether. The idea is to approximately transport an easy replica x_0 sampling from $\pi_0 = \rho_0$ to the target $\pi_L = \pi$ while transporting a target replica x_L to ρ_0 , before attempting their swap. Thereby, only two replicas are strictly necessary. Following standard practice in PT, the base distribution is typically adapted to the target distribution, choosing $\pi_0 \propto \pi^\beta$, with $\beta < 1$, a high-temperature version of the target distribution π . As a result, $\rho_0 = \pi_0$ is not as trivial to sample from than when choosing a standard Gaussian, but the assumption is that a local MCMC sampler can still efficiently explore π_0 .

Denoting by T_θ a NF transport map trained to transport ρ_0 to π , and M_0 and M_π local Markov kernels leaving respectively ρ_0 and π invariant, the PT procedure is modified as follows:

algorithm 6: **Learned replica exchange (LREX)**

```

Initialize replicas  $x_0^{(0)}$  and  $x_\pi^{(0)}$  at random,
for  $k=1 \dots K$  do
     $x_0^{(k)} \sim M_0(\cdot | x_0^{(k-1)})$ ,
     $x_\pi^{(k)} \sim M_\pi(\cdot | x_\pi^{(k-1)})$ ,
    Draw  $u \sim \text{Uniform}(0, 1)$ 
    if  $u < \gamma$  then
        Attempt swap:
         $x'_0 = T_\theta^{-1}(x_\pi^{(k)})$ 
         $x'_\pi = T_\theta(x_0^{(k)})$ ,
        Compute  $\alpha = \min \left( 1, \frac{\pi(x'_\pi) \rho_\theta(T_\theta(x'_0))}{\pi(x_\pi^{(k)}) \rho_\theta(T_\theta(x_0^{(k)}))} \right) = \min \left( 1, \frac{\pi(x'_\pi) \rho_0(x'_0)}{\pi(x_\pi^{(k)}) \rho_0(x_0^{(k)})} \times \left| \frac{\det \nabla T_\theta(x_0^{(k)})}{\det \nabla T_\theta(x'_0)} \right| \right)$ ,
        Set  $x_0^{(k)} \leftarrow x'_0$ ,  $x_\pi^{(k)} \leftarrow x'_\pi$  with probability  $\alpha$ 

```

The swap is attempted with probability γ at each iteration and ρ_θ denotes the density of the pushforward distribution induced by T_θ , given by the change-of-variables formula (eq. (2)). The

acceptance probability α ensures that the overall Markov chain over the joint state space of the two replicas leaves the joint distribution $\rho(x_0, x_\pi) = \rho_0(x_0)\pi(x_\pi)$ invariant.

In spirit, the proposition of [Inv+22] is similar to flow-IMH (algorithm 2) except that the base distribution is not sampled independently and the non-local moves are mediated by swaps between the base and target replicas. As a result, one can expect similar performances of this approach compared to flow-IMH.

A subsequent proposal by [Zha+25] places weaker requirements on the NF and retains a ladder of intermediate distributions $(\pi_\ell)_{\ell=1}^L$. As in SMC-based (CR)AFT, this ladder can be chosen coarser than in traditional PT, thanks to a collection of transport maps $(T_\ell)_{\ell=1}^L$ that approximately push forward $\pi_{\ell-1}$ to π_ℓ . Extending the two-replica version of algorithm 6, local MCMC kernels M_ℓ targeting each π_ℓ are run for every replica x_ℓ , and swaps between neighboring levels are periodically proposed using the corresponding transport maps. Although this strategy requires maintaining more replicas, it is expected to handle more complex target distributions than the two-replica scheme, as it decomposes the global transport from ρ to π into a sequence of simpler transitions from $\pi_{\ell-1}$ to π_ℓ .

3.3.5 Recalibrating NF samples with AIS or SMC

Returning to the setting where a single transport map T_θ is trained to push forward ρ_0 to π , it has also been proposed to leverage AIS or SMC in order to recalibrate the NF samples. The idea is to introduce a sequence of intermediate distributions that, here, gradually transports ρ_θ to π , for instance $\pi_\ell = \rho_\theta^{(L-\ell)/L} \pi^{\ell/L}$. Then, an AIS or SMC procedure is run targeting π , initialized from samples drawn from ρ_θ . This strategy was first proposed by [Mid+22] in the context of AIS and more recently revisited by [Tan+25] using SMC.

In contrast to elementary NF-based samplers, this approach explicitly compensates for imperfections in the learned transport map. In particular, since ρ_θ may not be sufficiently close to π to serve directly as a proposal distribution for IS or IMH, the tempering procedure corrects for this mismatch. As with other combinations of tempering and normalizing flows, the computational cost is higher than that of flow-IMH or Neural-IS, but this additional effort is expected to translate into improved sampling performance. Although other innovations are presented in [Mid+22] and [Tan+25], the former reports good performance of the AIS recalibration strategy on sampling from the Boltzmann measure of clusters of repulsive particles and the latter validates the SMC recalibration strategy on the sampling of small biomolecules.

*

Overall, the various proposals for combining NFs with tempering-based sampling schemes expand the capabilities of traditional tempering methods and ease the learning burden compared to the simple NF-powered samplers discussed above. However, a systematic comparison of these algorithms is still lacking. Moreover, while these approaches typically moderate the demands placed on learning—and are therefore likely to handle more complex targets than simple NF-enhanced samplers—it remains unclear whether they exhibit more favorable scaling with dimension. Indeed, transport is still performed in the full ambient space. The alternative strategy introduced next addresses this limitation directly by reducing the effective dimensionality of the problem.

3.4 Combining NFs with coarse-grainings

One emerging limitation of samplers relying on NF-based proposals is their poor scaling with dimension. Expecting a generative model to learn how to resample an entire high-dimensional variable in a coherent manner is, in a sense, overly ambitious. This observation motivates another

line of research seeking to leverage NFs while avoiding a full resampling. This is the underlying principle of the two methods described below.

3.4.1 Partial resampling in the latent space

A first strategy to mitigate the deterioration of acceptance rates with increasing dimension is to perform partial updates informed by the generative model trained to approximate π . In discrete state spaces, where NFs are replaced by autoregressive networks, this operation is straightforward provided that one updates variables appearing last in the autoregressive ordering [WRC21]. [SG23*] proposed an adaptation of this idea to the continuous setting where a NF is trained to transport ρ_0 to π via T_θ . Following the classical Metropolis-within-Gibbs scheme, one performs a partial resampling in latent space based on the conditional distributions of ρ_0 , and subsequently accepts or rejects the proposal according to a Metropolis acceptance rule.

More precisely, let $d_u \leq d$ denote the number of degrees of freedom to be updated, $S = (i_1, \dots, i_{d_u})$ a selected set of distinct indices, and $q(S)$ a probability distribution over such subsets. Denote by z_S and $z_{\setminus S}$ the components of the base variable z indexed by S and by its complement, respectively. The partial resampling MCMC procedure of [SG23*] consists in:

algorithm 7: **Gibbs flow-Monte Carlo (GflowMC)**

```

Initialize  $x^{(0)}$ ,
Compute  $z^{(0)} = T_\theta^{-1}(x^{(0)})$ ,
for  $k = 1 \dots K$  do
    Sample set  $S \sim q(S)$ ,
    Propose partial refreshing  $z'_S \sim \rho_0(z_S | z_{\setminus S}^{(k-1)})$ ,
    Set  $z'_{\setminus S} = z_{\setminus S}^{(k-1)}$ ,

    Compute acceptance rate  $\alpha = \min \left( 1, \frac{\pi_{T_\theta}(z')}{\pi_{T_\theta}(z^{(k-1)})} \times \frac{\rho_0(z_S | z_{\setminus S}^{(k-1)})}{\rho_0(z'_S | z_{\setminus S}^{(k-1)})} \right)$ ,

    Set  $z^{(k)} = z'$  with probability  $\alpha$ , else  $z^{(k)} = z^{(k-1)}$ ,
return samples  $\{x^{(k)} = T_\theta(z^{(k)})\}_{k=1}^K$ 

```

Here the pullback π_{T_θ} of π through T_θ , previously introduced in eq. (24), is again needed in the acceptance ratio. The conditional base distributions of the form $\rho_0(z_S | z_{\setminus S})$ are assumed to be explicit and easy to sample from. For a factorized ρ_0 , this is trivially the case as the conditional distribution $\rho_0(z_S | z_{\setminus S})$ is equal to the marginal $\rho_0(z_S)$, itself factorized – e.g. for the common choice $\rho_0 = \mathcal{N}(0_d, I_d)$, $\rho_0(z_S | z_{\setminus S}) = \mathcal{N}(0_{d_u}, I_{d_u})$.

The algorithm shares similarities with neutra-MCMC (algorithm 4), but replaces the local MCMC updates in the latent space with partial global updates that can more easily jump across modes, as illustrated by the two-dimensional example in fig. 4. In the context of statistical field theories, [SG23*] report that GFlowMC can outperform both neutra-MCMC and flow-IMH. In particular, the number of degrees of freedom updated, d_u , acts as a tunable parameter for optimizing sampler performance by trading off acceptance rate against the decorrelation in a single move.

In an application to biomolecular systems, [PG25] exploit a similar idea using continuous normalizing flows (CNFs). They report that updating a single latent variable at a time ($d_u = 1$) is optimal, and demonstrate a significant acceleration in equilibrating samples of the small, fast-folding protein Chignolin.

In the strategy described in the above paragraphs, the NF is still trained to approximate the full target distribution π . An alternative approach consists in leveraging generative models to

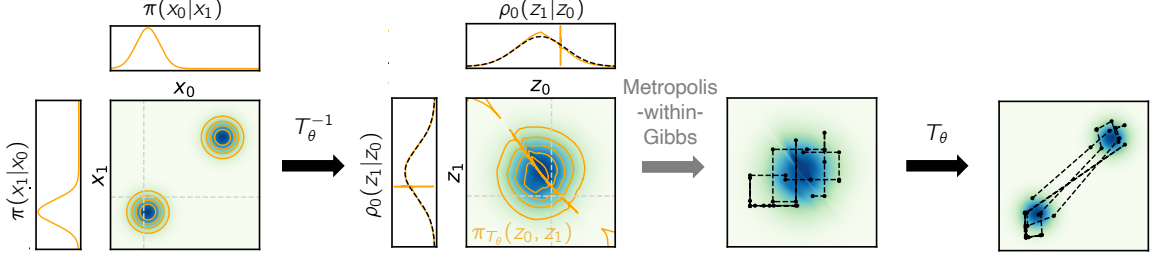


Figure 4: **Illustration of the Gibbs flow-Monte Carlo (GFlowMC) algorithm.** The target distribution π is a mixture of two Gaussians (left panel). The NF inverse transport map T_θ^{-1} approximately transports π to a standard Gaussian base distribution ρ_0 (middle left panel). The GFlowMC algorithm (middle right panel) performs partial resampling in the latent space of the NF targeting the pullback π_{T_θ} (in orange), here updating one variable at a time ($d_u = 1$). Sending the samples back to the original space (right panel) evidence the ability of GFlowMC to jump between the two modes of π . Adapted from the poster of C. Schönle on [SG23*].

sample a low-dimensional coarse-grained representation of the system, which is the idea developed next.

3.4.2 Collective-variable guided sampling

As already mentioned in the introduction, apart from annealing-based methods, another key strategy to address metastability in high dimension relies on the existence of a low-dimensional representation (or coarse-graining), referred to as collective variables (CVs), that effectively separates the different modes of the target measure. In this context, generative models provide a flexible and scalable framework to estimate the marginal law of the CVs. In contrast to traditional non-parametric approaches — which typically restrict CV-guided algorithms to very low-dimensional CV spaces due to the difficulty of estimating their marginal density — generative models make it possible to handle significantly higher-dimensional CV spaces. They therefore extend the range of admissible dimensions for the CV representation in coarse-graining-based sampling strategies.

Here, we review the approach proposed by [TLG24*; Sch+25a*; Sch+25b*], revisiting the concept of CV-guided sampling using a NF trained to sample the distribution of the CVs. The approach is based on non-equilibrium samplers previously discussed without resorting to NFs, notably in [Nea96; Ath02; Nil+11; CR15].

For simplicity, we discuss the separable case where the CVs are a subset of the full variables, i.e. $x = (z, x^\perp) \in \mathbb{R}^d$, with $z \in \mathbb{R}^{d_z}$ the CVs and $x^\perp \in \mathbb{R}^{d-d_z}$ the remaining degrees of freedom. The strategy presented below readily generalizes to an arbitrary CV mapping $z = \xi(x)$ following [Sch+25b*]. Returning to the separable case, the marginal distribution over CVs of the target distribution $\pi(x)$ is given by

$$\pi_z(z) = \int \pi(z, x^\perp) dx^\perp. \quad (32)$$

The sampling algorithm then relies on two main ingredients:

- (i) a NF trained to sample approximately from the marginal distribution of the CVs, its probability density function is noted $\rho_\theta(z)$,
- (ii) a non-equilibrium sampling procedure updating the value of $x^\perp$ conditional on the current value of the CVs z .

Let us describe the second ingredient.

Given the current state $x = (z, x^\perp)$ and a proposed new value z' for the CVs, we introduce a T -step path $S(z, z') = (z_t)_{t=0}^T$ connecting z to z' , with $z_0 = z$ and $z_T = z'$. A simple choice is the linear interpolation $z_t = z + (z' - z)t/T$. Starting from $x_0^\perp = x^\perp$, the orthogonal degrees of freedom are then updated through a sequence of T Markov kernels associated with the intermediate values z_t :

algorithm 8: Non-equilibrium update of orthogonal degrees of freedom

Input: Current state $x = (z, x^\perp)$, target CVs z' , Markov kernels $(M_z)_{z \in \mathbb{R}^{d_z}}$,
number of steps T , schedule $S(z, z') = (z_t)_{t=0}^T$
Initialize $x_0^\perp = x^\perp$,
for $t = 0 \dots K - 1$ **do**
 | Sample $x_{t+1}^\perp \sim M_{z_t}(\cdot | x_t^\perp)$,
return Final chain of states $(z_t, x_t^\perp)_{t=0}^T$

This elementary algorithm defines a procedure for proposing a new value of the full state $x' = (z', x^{\perp'}) = (z_T, x_T^\perp)$ given a proposed CV update $z \rightarrow z'$. Although, in principle, any Markov kernel M_z may be used, efficiency of the algorithm described below requires choosing M_z such that the resulting proposal $(z', x^{\perp'})$ is likely under the target distribution π . A common and effective choice is a single step of overdamped Langevin dynamics targeting the conditional distribution $\pi(x^\perp | z)$ at fixed z :

$$M_z(x_{t+1}^\perp | x_t^\perp) = \mathcal{N}\left(x_{t+1}^\perp; x_t^\perp - \delta t \nabla_{x^\perp} U(z, x_t^\perp), 2\delta t I_{d-d_z}\right), \quad (33)$$

where δt is a small time-step and $U(x) = -\log \pi(x)$ (up to an additive constant) is the potential energy function associated with the target distribution.

Using a Markov kernel Q to propose an update of the CV, an accept-reject criterion can be written à la Metropolis-Hastings to design an MCMC with stationary distribution π . For this, the CV-schedule needs to be reversible, meaning that $S(z, z') = (z_t)_{t=0}^T$ implies that $S(z', z) = (z_{T-t'})_{t'=0}^T$. The probability density of the generated path and the probability density of following the reverse path with the same protocol are respectively given by

$$Q(z'|z) \prod_{t=0}^{T-1} M_{z_t}(x_{t+1}^\perp | x_t^\perp), \quad \text{and} \quad Q(z|z') \prod_{t'=0}^{T-1} M_{z_{T-t'}}(x_{T-t'-1}^\perp | x_{T-t'}^\perp). \quad (34)$$

Then a sufficient criterion for ensuring that the overall Markov chain leaves π invariant is to accept the proposed move $(z, x^\perp) \rightarrow (z', x^{\perp'})$ with probability:

$$\alpha = \min\left(1, \frac{\pi(z', x^{\perp'})}{\pi(z, x^\perp)} \times \frac{Q(z|z')}{Q(z'|z)} \times \frac{\prod_{t'=0}^{T-1} M_{z_{T-t'}}(x_{T-t'-1}^\perp | x_{T-t'}^\perp)}{\prod_{t=0}^{T-1} M_{z_t}(x_{t+1}^\perp | x_t^\perp)}\right). \quad (35)$$

This acceptance rule satisfies a property stronger than detailed balance, sometimes referred to as super-detailed balance [Fre04; Nil+11]. By using a NF trained to approximate π_z to propose updates in the CV space — that is, by taking $Q = \rho_\theta$ — one can promote transitions between metastable modes. In turn, this mechanism yields an efficient sampler for complex multimodal distributions. The complete algorithm reads as follows:

algorithm 9: **Collective-variable guided flow-Monte Carlo (CV-FlowMC)**

```

Initialize  $x^{(0)} = (z^{(0)}, x^{\perp(0)})$ ,
for  $k = 1 \dots K$  do
    Propose new CVs  $z' \sim \rho_{\theta}(z)$ ,
    Generate non-eq. path  $(z_t, x_t^{\perp})_{t=0}^T$  using algorithm 8 with input  $(z^{(k-1)}, x^{\perp(k-1)})$  and  $z'$ ,
    Set  $x' = (z', x_T^{\perp})$ ,
    Compute acceptance rate  $\alpha$  using eq. (35),
    Set  $x^{(k)} = x'$  with probability  $\alpha$ , else  $x^{(k)} = x^{(k-1)}$ ,
return samples  $\{x^{(k)}\}_{k=1}^K$ 

```

Many variants of this algorithm can be envisioned, ranging from technical choices in the definition of the reverse protocol to different selections of the Markov kernels M_z and associated hyperparameters (e.g., the number of steering steps T , or the time step δt when using overdamped Langevin dynamics). We refer to [Sch+25a*; Sch+25b*] for a detailed analysis of these design choices and their impact on sampler performance. As mentioned above, the method can moreover be extended to arbitrary CV mappings $z = \xi(x)$ by evolving the orthogonal degrees of freedom $x^{\perp}$ at fixed CV values via constrained dynamics, as described in [Sch+25b*].

The core idea, however, is the use of a NF to propose updates in the CV space that traverse low-probability regions and connect metastable modes. Crucially, the flow is trained only in a coarse-grained space ($d_z < d$), whose marginal distribution is typically smoother than the full target distribution π . This substantially alleviates the curse of dimensionality affecting NF-based samplers in the full space, while preserving their ability to generate large, mode-crossing moves. In particular, the admissible dimensionality of the CV space is substantially higher than in traditional CV-based enhanced sampling methods: values on the order of $d_z \approx 10\text{s} - 100\text{s}$ appear feasible, whereas standard approaches often struggle beyond $d_z \approx 3\text{--}4$ due to the difficulty of estimating the marginal distribution in higher dimensions with non-parametric methods. Preliminary results reported in [Sch+25b*] on toy molecular systems indicate indeed improved performance as d_z increases, which is encouraging for applications to more complex systems.

*

Finally, here, as for all the exact samplers leveraging NFs discussed in this section 3, we have so far assumed that the flow is trained beforehand to approximate either the full target distribution or its marginal over the CV space. The question of how to train such flows efficiently for sampling purposes is, however, a substantial research direction in its own right, to which we now turn.

4 Training generative models without data

In their conventional setting, generative models are trained on large datasets consisting of samples drawn from an unknown target distribution. In scientific computing, the situation is markedly different: the target distribution π is typically known—at least up to a normalization constant—and the central challenge is precisely to generate samples from it. This fundamental distinction calls for a rethinking of the training paradigm. Rather than learning from empirical data, the model must instead exploit direct access to the unnormalized density of the target distribution.

In this section, we review several strategies that have been proposed to train generative models in this setting, organized according to the main classes of transport-based generative models introduced earlier. Throughout, we use the notation $\pi = e^{-U}/\mathcal{Z}$ for the target distribution, where U denotes the potential energy function and $\mathcal{Z}$ the (unknown) normalization constant.

4.1 Normalizing flows

4.1.1 Variational inference

Minimizing the reverse KL divergence Given their tractable likelihood, NFs are typically trained for generative modeling by maximizing the likelihood of the training data. Equivalently, this corresponds to minimizing the Kullback-Leibler (KL) divergence between the data distribution p_{data} and the NF distribution ρ_{θ} :

$$D_{\text{KL}}(p_{\text{data}}\|\rho_{\theta}) = \mathbb{E}_{p_{\text{data}}} \left[\log \frac{p_{\text{data}}(x)}{\rho_{\theta}(x)} \right] = -\mathcal{H}(p_{\text{data}}) - \mathbb{E}_{p_{\text{data}}} [\log \rho_{\theta}(x)], \quad (36)$$

where $\mathcal{H}(p_{\text{data}})$ is the entropy of the data distribution, independent of the model parameters θ . In the context of sampling, however, the “data” distribution is replaced by a target distribution π , and computing expectations with respect to π is typically challenging. There, the *reverse* KL divergence can serve as a practical training objective instead:

$$D_{\text{KL}}(\rho_{\theta}\|\pi) = \mathbb{E}_{\rho_{\theta}} \left[\log \frac{\rho_{\theta}(x)}{\pi(x)} \right] = \mathbb{E}_{\rho_{\theta}} [\log \rho_{\theta}(x) + U(x)] + \log \mathcal{Z}. \quad (37)$$

Expectations with respect to ρ_{θ} are cheap to estimate by drawing independent samples from the flow, $\log \rho_{\theta}$ is available in closed form, and $\log \mathcal{Z}$ is unknown but independent from θ . Minimizing this objective corresponds to variational inference (VI), a well-established technique in Bayesian statistics [Jor+99; BKM17].

Variational families for ρ_{θ} must maintain tractable sampling and density evaluation. Traditional choices, such as factorized or Gaussian distributions, are often poorly suited to approximate complex target distributions because they lack flexibility. NFs, in contrast, are highly expressive while preserving the required ease of sampling and density evaluation. Using NFs for VI was first proposed by [RM15], who leveraged the reparameterization trick to avoid backpropagating through the sampling process. Writing $x \sim \rho_{\theta}$ as $x = T_{\theta}(z)$ with $z \sim \rho_0$ the base distribution, the effective loss for a batch of N samples $z_i \sim \rho_0$ i.i.d. becomes

$$\ell_N(\theta) = \frac{1}{N} \sum_{i=1}^N [\log \rho_0(z_i) + \log \det |\nabla_z T_{\theta}(z_i)| + U(T_{\theta}(z_i))]. \quad (38)$$

Following this foundational work, simultaneous applications in statistical physics emerged: [AKS19] applied NFs to field theories, [Noé+19] to molecular systems – where reverse-KL training is called “training by energy” –, and [WWZ19] to spin systems (using autoregressive models instead of flows). However, this seemingly data-free training approach has an important limitation when applied to multimodal target distributions, a phenomenon often referred to as mode collapse.

Mode collapse in VI Mode collapse refers to the tendency of the learned distribution ρ_θ to cover only a subset of the modes of the target distribution π . This phenomenon stems from the well-known mode-seeking behavior of the reverse KL divergence: assigning probability mass to regions where π has low density is penalized much more strongly than failing to cover some modes of π altogether [Jor+99; Min05]. Crucially, this behavior is intrinsic to the objective function itself and is therefore not resolved by increasing model expressiveness. As a result, even highly flexible models such as NFs can suffer from mode collapse, despite their capacity to represent multimodal distributions, as reported by several authors [Nic+23; Ble+24; GNG25*]. This limitation is particularly severe in sampling applications of generative models, since multimodal target distributions are precisely those for which traditional sampling methods struggle the most.

A number of heuristic strategies have been proposed to alleviate mode collapse in VI when training NFs for sampling purposes. For example, [Noé+19] introduced a hybrid training procedure that augments the reverse KL objective with a maximum-likelihood term based on short MCMC trajectories initialized near distinct modes. Alternatively, [Mid+22] proposed estimating the χ^2 divergence between ρ_θ and π using AIS, and minimizing this estimate instead. Another broadly applicable approach consists in combining VI with annealing, which we discuss next.

Annealed VI to mitigate mode collapse A natural strategy to mitigate mode collapse when training NFs is to introduce annealing, in a manner reminiscent of annealed sampling schemes. The key idea is to define a sequence of intermediate distributions $(\pi_k)_{k=0}^K$ interpolating between a simple unimodal distribution π_0 and the target distribution $\pi_K = \pi$. The NF is then trained sequentially: at each stage k , variational inference (VI) is used to approximate π_k , initializing the parameters of ρ_{θ_k} from those obtained at stage $k-1$.

In practice, the intermediate distributions are often defined through tempering of the target distribution, $\pi_k \propto e^{-\beta_k U}$, with a monotonically increasing sequence of inverse temperatures $\beta_0 < \beta_1 < \dots < \beta_K = 1$ or through a geometric interpolation toward a chosen π_0 , $\pi_K \propto \pi^{k/K} \pi_0^{(K-k)/K}$. The progressive training strategy along the interpolation path enables the NF to gradually discover and cover distinct modes of the target distribution, thereby reducing the risk of mode collapse.

Annealed VI has demonstrated empirical success across a range of applications: the exploration of complex Bayesian posteriors [Hua+18; WCD25], the sampling of Boltzmann distributions in statistical mechanics [WWZ19; Hac+21; SF25], and in optimization tasks ranging from locating multiple modes [LeM+25] to identifying global minima [Hib+21; KIH25; BRZ25]. A closely related strategy is employed in the previously discussed (CR)AFT samplers [AMD21; Mat+22], where the flow architecture itself is expanded at each stage by adding new layers, rather than retraining a fixed model.

Despite these encouraging results, performance remains highly sensitive to the annealing schedule—most notably the choice of the initial distribution π_0 and the rate at which the transition is operated toward the target distribution. To date, theoretical guidance on these choices within the VI framework remains limited. Recent work [SGL25*; FLG26*], however, has begun to analyze the behavior of (annealed) VI in simplified settings, in particular Gaussian mixture models.

In [SGL25*], the authors identify a critical threshold for the separation between modes—dependent on their relative weights but independent of the ambient dimension—that determines whether mode collapse occurs. This behavior is also observed empirically when training NFs on Gaussian mixtures. Building on this analysis, [FLG26*] derive a closed-form expression for the probability of mode collapse in the bimodal Gaussian mixture case, enabling a more detailed study of the impact of the annealing schedule. Their results confirm the mitigating role of annealing and

provide concrete guidance for its practical implementation: starting from a sufficiently high temperature is not, by itself, sufficient—one must also ensure that the annealing schedule is slow enough. These theoretical predictions are again supported by experiments with NFs trained on Gaussian mixtures.

These works constitute an important first step toward a more rigorous understanding of (annealed) VI with expressive generative models, such as NFs, and offer valuable insight for its use in challenging sampling applications.

4.1.2 Adaptive strategies

While training NFs via the reverse KL divergence requires explicit safeguards against mode collapse—such as annealing—an alternative line of work has focused on retaining the standard and notably stable maximum-likelihood training paradigm. Because direct samples from the target distribution π are not available, these approaches rely on *adaptive strategies* in which data collection and model training are interleaved. This idea was first introduced by [PM18] in the context of triangular transport maps, independently extended to general variational families by [NLB20], and later rediscovered for modern NFs in [GRV22*; GRV21*]. We summarize the latter approach below.

The adaptive training procedure runs an MCMC chain targeting π , which alternates between two types of Markov update kernels:

- (i) A *local* MCMC kernel (e.g., Langevin or HMC) that leaves π invariant and explores the target distribution, denoted M_π below;
- (ii) A *global* NF-based kernel that also leaves π invariant – such as the flow-IMH kernel described in algorithm 2 –and enables long-range, mode-crossing updates, denoted M_{ρ_θ} below.

The NF is periodically retrained by maximum likelihood on the samples collected so far, leading to the following overall scheme:

algorithm 10: Adaptive flow Monte Carlo (A-flowMC)

Input: Initial NF parameters θ , learning rate η , frequency of global jumps K_g , frequency of retraining K_t
Initialize $x^{(0,m)}$ for $m = 1 \dots M$ in basins of attraction of the distinct modes of π
for $k = 1 \dots K$ **do**
 if $k \bmod K_g = 0$ **then**
 Global NF-based move $x^{(k,m)} \sim M_{\rho_\theta}(\cdot | x^{(k-1,m)})$
 else
 Local MCMC move $x^{(k,m)} \sim M_\pi(\cdot | x^{(k-1,m)})$
 if $k \bmod K_t = 0$ **then**
 $\theta \leftarrow \theta + \frac{\eta}{Mk} \sum_{m=1}^M \sum_{\ell=1}^k \nabla \log \rho_\theta(x^{(\ell,m)})$
return trained flow ρ_θ , samples $(x^{(k,m)})_{m,k=1}^{M,K}$

Variants of this basic scheme include discarding the oldest samples to limit memory usage, subsampling mini-batches from previously collected samples, performing multiple gradient steps at each retraining stage, and using more advanced optimizers such as Adam instead of simple stochastic gradient descent (SGD).

The key feature of this approach is that it simultaneously achieves two objectives: it produces unbiased samples from the target distribution π , and it trains a NF that can subsequently be used within an efficient standalone sampler from section 3. From a theoretical standpoint, however,

proving convergence of such adaptive MCMC schemes—where the transition kernel evolves over time based on the history of the chain and thus violates the Markov property—is non-trivial [Bro+22*]. In particular, convergence of the Markov chain to the target measure π typically requires the adaptation to vanish asymptotically. In practice, this issue can be sidestepped by stopping the training once the NF is deemed sufficiently accurate, and then continuing the simulation with a fixed kernel to generate exact samples from π .

The benefit of alternating between local and global moves is that the two components reinforce each other. Local MCMC updates explore individual modes in detail, providing training data that allow the NF to capture local structure, while the NF enables global proposals that help the chain escape metastable states. Together, these mechanisms facilitate accurate estimation of the relative weights of the different modes. Rigorous analyses demonstrating the advantages of combining local and global sampling have been established for NFs in [Sam+22*] and for autoregressive models in [DRZ25; BRZ25].

Provided that the MCMC chains can be initialized in all modes of interest of the target distribution, the maximum-likelihood objective naturally prevents mode collapse, since all modes present in the training data must be represented by the NF. A notable exception arises for modes that are significantly rarer than others: in this case, some training batches may contain no samples from these rare modes, causing the NF to accidentally forget them. This issue can be mitigated by enforcing that training batches include samples from all relevant modes, or by using a mixture of NFs in which each component is trained on a specific mode and the mixture weights are lower bounded [GRV21*; Hac+21; Mol+24*]. Aside from these rare-mode scenarios—which require special care—the main limitation of adaptive strategies based on maximum-likelihood training remains the need to initialize chains in all relevant modes. As for mode collapse, this limitation can be alleviated by resorting to annealing, as discussed next.

4.1.3 Sequential tempering

The idea of adaptive maximum-likelihood training can be easily combined with annealing in order to relax the assumption that all relevant modes of the target distribution are known in advance. This sequential tempering is therefore built upon an annealing schedule $(\pi_k)_{k=0}^K$ defined, as before, with π_0 a simple, typically unimodal distribution and $\pi_K = \pi$ the target distribution and follows an iterative procedure.

The algorithm is initialized by sampling from π_0 using a suitable MCMC method, a task that is assumed to be straightforward. A NF is then trained by maximum likelihood on the resulting samples from π_0 . Once trained, the NF is used to facilitate sampling from the slightly more challenging distribution π_1 , for instance via flow-IMH (algorithm 2). The samples collected at this stage are used in turn to retrain the NF by maximum likelihood, initializing the optimization from the parameters learned at the previous level. This procedure is repeated iteratively along the annealing path until the target distribution π_K is reached, leading to the following overall scheme:

algorithm 11: **Sequential tempering**

Input: Initial NF parameters θ , learning rate η
 $x^{(0,m)} \sim \pi_0$ using MCMC for $m = 1 \dots M$
for $k = 0 \dots K - 1$ **do**
 repeat
 $\theta \leftarrow \theta + \frac{\eta}{M} \sum_{m=1}^M \nabla \log \rho_\theta(x^{(k,m)})$
 until convergence
 $x^{(k+1,m)} \sim \text{flow-IMH targeting } \pi_{k+1} \text{ using current } \rho_\theta \text{ for } m = 1 \dots M$
return trained flow ρ_θ , samples $(x^{(K,m)})_{m=1}^M$

Training the flow to convergence at each annealing stage robustly prevents mode collapse, provided that the annealing schedule is sufficiently fine. This conservative strategy was termed “adiabatic retraining” by [Hac+21]. In practice, however, it may be computationally more efficient to perform only a limited number of training steps at each stage, in the same spirit as adaptive training without annealing (algorithm 10).

This sequential tempering strategy has been primarily explored in the context of sampling disordered spin systems using autoregressive models, beginning with the work of [McN+20] and subsequently adopted in [Cia+23; DRZ25; BRZ25]. Such settings are particularly well suited to sequential tempering, as the number and location of modes are typically unknown. Nevertheless, the approach may still fail to produce exact samples when the target distribution is too complex to be faithfully represented by the chosen model class. This limitation was illustrated in [Cia+23], where autoregressive models were found to be inadequate for sampling solutions of a graph-coloring constraint satisfaction problem.

Going further, [Cia+23] analyze the mechanisms underlying these failures by comparing different training objectives, namely VI and maximum-likelihood training with tempering. VI tends to produce approximations ρ_θ that underestimate the entropy of π , in line with the mode collapse phenomenon, while still generating configurations with appropriate energies. In contrast, maximum-likelihood training accurately captures the entropy but often yields configurations with energies higher than those of the true equilibrium distribution. While this study focused on autoregressive models, similar phenomena are likely to arise when training NFs under either objective. To move beyond these limitations, one can consider alternative classes of transport-based generative models amenable to different training objectives, as detailed in the next sections.

4.2 Continuous annealed flows

CNFs based on ODEs are typically more expressive than their discrete counterparts. However, evaluating their likelihood (see eq. (8)) is computationally demanding and inevitably subject to discretization error. Consequently, the training methods described in the previous sections for normalizing flows are not well suited to the continuous setting.

Instead, one can exploit the intrinsic continuity of CNFs and train them for sampling by matching the trajectories of the flow to a prescribed annealing path of distributions. Here we consider a path $(\pi_t)_{t \in [0,1]}$ indexed by a continuous time $t \in [0,1]$, with π_0 a simple distribution and $\pi_1 = \pi$ the target. This approach was first developed for deterministic continuous flows (as CNFs presented in section 2.1.2). It was subsequently extended to a stochastic variant, which we discuss afterward.

Deterministic flows following an annealing path The goal is to learn a velocity field $v_\theta(x, t)$ such that the time marginal $\rho_{\theta, t}$ —i.e., the instantaneous distribution of the solution to

$$\frac{dX_t}{dt} = v_\theta(X_t, t), \quad X_0 \sim \rho_0 = \pi_0, \quad (39)$$

matches the prescribed family distributions π_t for all $t \in [0, 1]$. Following [MF23; TPL24; Fan+24], one can derive a learning objective as follows. Write π_t as a Boltzmann distribution $\pi_t(x) = e^{-U_t(x)}/\mathcal{Z}_t$, where U_t is a known time-dependent energy (e.g. $U_t = tU_1 + (1-t)U_0$) and $\mathcal{Z}_t$ is the associated (intractable) partition function. A velocity field v that exactly transports π_0 to π_t at time t , satisfies the continuity equation

$$\partial_t \pi_t + \nabla \cdot (\pi_t v) = 0 \Rightarrow \partial_t \pi_t = -\nabla \cdot (\pi_t v) = -\pi_t (\nabla \cdot v - \nabla U_t \cdot v), \quad (40)$$

where in the last equality we used $\nabla \pi_t = -\pi_t \nabla U_t$. On the other hand, we also have

$$\partial_t \pi_t = -\pi_t (\partial_t U_t + \partial_t \log \mathcal{Z}_t). \quad (41)$$

Comparing eqs. (40) and (41) shows that any velocity field v that perfectly matches the reference path must satisfy on the support of π_t for all t ,

$$\partial_t U_t + \partial_t \log \mathcal{Z}_t = \nabla \cdot v - \nabla U_t \cdot v. \quad (42)$$

To turn eq. (42) into a tractable training objective for v_θ , we must address the fact that $\log \mathcal{Z}_t$ is unknown. However, one can show that any triplet $(v, U, \mathcal{Z})$ satisfying eq. (42) for all x and t necessarily also satisfies $\mathcal{Z}_t = \int e^{-U_t(x)} dx$ for all t (see, e.g., Lemma 1 in [MF23]). This motivates treating $\log \mathcal{Z}_t$ as an additional time-dependent function to be learned; here we parameterize it as $-F_\theta(t)$, in analogy with the free energy.⁴ Enforcing eq. (42) by minimizing the squared residual in expectation under a user-chosen sampling distribution ρ_t yields the loss

$$L(\theta) = \int_0^1 dt \int dx \rho_t(x) |\partial_t U_t(x) - \partial_t F_\theta(t) - \nabla \cdot v_\theta(x, t) + \nabla U_t(x) \cdot v_\theta(x, t)|^2. \quad (43)$$

This objective corresponds to a so-called physics-informed neural network (PINN) loss. It involves parameterizing both the velocity field v_θ and the free energy F_θ with neural networks, and evaluating the loss by computing the required derivatives in the residual via automatic differentiation.

In practice, the choice of where to evaluate the loss is critical, since the velocity field only needs to be accurate in regions where π_t has significant probability mass and which are therefore likely to be explored during sampling. A common choice is to set $\rho_t = \text{StopGrad}(\rho_{\theta, t})$,⁵ focusing efforts on regions discovered during training. It has also been observed that training is more stable when one begins with concentrating on small times t and progressively extends the time horizon, in a manner reminiscent of sequential tempering or of the AFT/CRAFT algorithms previously discussed for discrete-time NFs. Another strategy to avoid starting without guidance about where the probability mass lies is to complement the learned transport with an annealed Langevin dynamics, as discussed next.

Langevin-driven stochastic flows Given the reference path $(\pi_t)_{t \in [0, 1]}$, consider the annealed Langevin dynamics with time-dependent diffusion coefficient γ_t :

$$dX_t = -\gamma_t \nabla U_t(X_t) dt + \sqrt{2\gamma_t} dW_t, \quad X_0 \sim \pi_0. \quad (44)$$

⁴Note that [MF23] instead approximate $\partial_t \log \mathcal{Z}_t$ directly.

⁵The dependency of the argument of StopGrad on trainable parameters is ignored during training.

Because the potential U_t evolves continuously in time, this dynamics is out- of-equilibrium and does not in general follow the prescribed path π_t . However, a remarkable result from stochastic thermodynamics shows that one can modify the dynamics by adding a control term v to the drift so that the resulting controlled process tracks the reference path π_t exactly at all times [VJ08]. This idea was revisited by [Var+24; AV25] using modern neural networks to learn the control drift for sampling purposes.

The derivation relies on the Fokker-Planck equation. In addition to eqs. (40) and (41), the density π_t satisfies

$$\nabla \pi_t = -\nabla U_t \pi_t \quad \Rightarrow \quad 0 = \gamma_t (\nabla \pi_t + \nabla U_t \pi_t) \quad \Rightarrow \quad 0 = \gamma_t \nabla \cdot (\nabla \pi_t + \nabla U_t \pi_t). \quad (45)$$

Adding this identity to the continuity equation eq. (40), satisfied by any velocity field v that perfectly generates π_t , yields

$$\partial_t \pi_t = -\nabla \cdot (\pi_t v) + \gamma_t \nabla \cdot (\nabla \pi_t + \nabla U_t \pi_t) = -\nabla \cdot (\pi_t (v - \gamma_t \nabla U_t)) + \gamma_t \Delta \pi_t, \quad (46)$$

which is precisely the Fokker-Planck equation associated with the controlled Langevin dynamics

$$dX_t = (v(X_t, t) - \gamma_t \nabla U_t(X_t)) dt + \sqrt{2\gamma_t} dW_t, \quad X_0 \sim \pi_0. \quad (47)$$

Therefore, learning a control v_θ by minimizing the PINN loss eq. (43) produces a controlled Langevin dynamics (47) that approximately follows the reference path π_t . At the beginning of training, the control is expected to be inaccurate, so the dynamics will not yet track π_t precisely. Nevertheless, thanks to the intrinsic Langevin drift $-\gamma_t \nabla U_t$, at the same stage of training, it remains closer to the reference path than the purely deterministic ODE discussed in the previous paragraph.

Learning the annealing path jointly with the dynamics Finally, we note that within this framework it has also been proposed to learn the annealing path $(\pi_t)_{t \in [0,1]}$ itself, in addition to the velocity field and the free energy [MF23; AV25]. This additional layer of learning is motivated by the observation that, as for any annealing based sampling or training approach, the choice of path can have a substantial impact on sampling performance. In particular, *phase transitions* along the path—namely abrupt changes in the localization or relative weights of the modes—are expected to severely hinder sampling efficiency.

In [MF23], the energy U_t is treated as a trainable function of the form

$$U_{\theta,t} = tU_1 + (1-t)U_0 + t(1-t)\hat{U}_{\theta,t} \quad (48)$$

where $\hat{U}_{\theta,t}$ is parameterized by a neural network. The authors minimize the same PINN loss eq. (43) jointly over v_θ , F_θ , and U_θ . The resulting objective is likely to admit many solutions, and it remains unclear how to systematically favor those leading to favorable sampling properties. Nevertheless, this approach constitutes a first step toward learning annealing paths automatically, and encouraging numerical results are reported in simple settings where traditional annealing schemes of geometrical averaging or temperature scheduling struggle.

In contrast, [AV25] mention the possibility of learning the energy function associated with the marginal distribution of a stochastic interpolant (see eq. (16)), although without providing theoretical guarantees or numerical evaluation. The appeal of this choice is not accidental: stochastic interpolant bridges are typically smooth and regular, thereby avoiding the sharp transitions and mode switching that plague many traditional annealing paths. This regularity is also characteristic of diffusion models and partly underlies the considerable attention they have received for sampling applications, as discussed in the next section. Namely, we now turn to examine how diffusion models can be trained for sampling purposes.

4.3 Diffusion models

Due to the stochastic nature of the generative process, likelihood evaluation for diffusion models is even less tractable than for continuous NFs, for which the variational and adaptive training methods presented in section 2.1.2 are already of limited practical suitability. Moreover, diffusion models inherently define their own bridging path between a simple reference distribution and the target, and therefore cannot be trained to follow a prescribed annealing path as CNFs. For these reasons, diffusion models have prompted the development of a distinct family of training objectives tailored to sampling.

The reader interested in disordered statistical physics system and high-dimensional statistics should refer to [AMS22; Mon23; MW24; Ghi+24]. In these special instances, the score can be computed using message passing algorithms. We do not discuss these approaches here and focus instead on general purpose strategies.

We begin by clarifying the challenges associated with likelihood estimation in diffusion models (section 4.3.1) before turning to alternative training approaches that circumvent explicit likelihood computation (sections 4.3.2 and 4.3.3).

4.3.1 On estimating the likelihood of diffusion models

We adopt the notation introduced in section 2.2 and consider a diffusion model based on the OU noising process (eqs. (10) and (11)). The instantaneous probability density of the process is denoted by π_t . After learning a score model $s_\theta(t, x) \approx \nabla \log \pi_t(x)$, the generative procedure approximates the denoising dynamics via the reverse-time SDE:

$$dX_t = (X_t + 2s_\theta(T - t, X_t)) dt + \sqrt{2} dW_t, \quad X_0 \sim \mathcal{N}(0, I), \quad (49)$$

with instantaneous density denoted by $\rho_{\theta t}$. One seeks to estimate $\rho_{\theta T}$. The Fokker-Planck equation governing the evolution of the density $\rho_{\theta t}$,

$$\partial_t \rho_{\theta t} = -\nabla \cdot ((x + 2s_\theta(T - t, x))\rho_{\theta t}) + \Delta \rho_{\theta t} \quad (50)$$

is not tractable to solve directly. To circumvent this difficulty, one can instead consider a deterministic process that shares the same time marginals as the exact noising-denoising dynamics. Such a process can be derived by examining the Fokker-Planck equation associated with the OU SDE in forward time τ :

$$\begin{aligned} \partial_\tau \pi_\tau &= -\nabla \cdot (-x \pi_\tau) + \Delta \pi_\tau \\ &= -\nabla \cdot (-x \pi_\tau) + \nabla \cdot (\nabla \log \pi_\tau(x) \pi_\tau) \\ &= -\nabla \cdot ((-x - \nabla \log \pi_\tau(x)) \pi_\tau), \end{aligned}$$

which can be interpreted as a continuity equation with velocity field $v_\tau(x) = -x - \nabla \log \pi_\tau(x)$. Replacing the true score $\nabla \log \pi_\tau$ with its learned approximation $s_\theta(\tau, x)$ and switching to reverse time $t = T - \tau$ yields the deterministic denoising process:

$$dX_t^d = (X_t^d + s_\theta(T - t, X_t^d)) dt, \quad X_0^d \sim \mathcal{N}(0_d, I_d), \quad (51)$$

whose instantaneous density is denoted by $\rho_{\theta t}^d$. At this stage, the diffusion model can be view as a CNF, for which the likelihood along a trajectory $(X_t^d)_{t \in [0, T]}$ can be estimated by the instantaneous change of variables formula (eqs. (7) and (8)),

$$\log \rho_{\theta T}^d(X_T^d) = \log \rho_{\theta 0}^d(X_0^d) - \int_0^T \nabla \cdot (X_t^d + s_\theta(T - t, X_t^d)) dt. \quad (52)$$

As discussed earlier for CNFs, estimating this integral is challenging. The divergence term requires backpropagating through the score model, with a computational cost that scales with the ambient dimension. In practice, the divergence is therefore often approximated using Hutchinson’s trace estimator, introducing an additional source of error that compounds with the unavoidable discretization error of the time integral. Finally, this change-of-variables formula only yields the likelihood of the deterministic denoising process $\rho_{\theta_T}^d$, which itself merely approximates the density ρ_{θ_T} associated with the stochastic denoising SDE.

Several works have sought to improve the accuracy of likelihood estimation for diffusion models, motivated by different objectives. A detailed discussion of these methods is beyond the scope of this manuscript; we therefore briefly highlight two representative research directions. The approach proposed in [Cho+22] introduces an auxiliary model to estimate so-called time scores $\partial_t \log \pi_t(x)$, which can be used to bypass the explicit computation of the divergence term in the change-of-variables formula, while remaining focused on the deterministic denoising process. In contrast, [Skr+24] applies Itô’s lemma to derive an instantaneous change-of-variables formula directly along the stochastic denoising dynamics. Although both approaches lead to improved likelihood estimates, their accuracy remains fundamentally limited by the increasing dimensionality of the problem. We therefore turn next to likelihood-free (and data-free) strategies for training diffusion models in the context of sampling.

4.3.2 Monte Carlo estimation of the score

A first class of methods has sought to exploit Monte Carlo estimators of the score function $\nabla \log \pi_t$ in order to obtain diffusion models from the target distribution π without access to a training dataset. As already mentioned in the introduction on diffusion models (section 2.2), the score function can be expressed as a conditional expectation with respect to the noising process. For the OU process, recalling eq. (13), we have

$$\nabla \log \pi_\tau(x) = -\mathbb{E}_{x_0} \left[\frac{x - x_0 e^{-\tau}}{1 - e^{-2\tau}} \middle| \tilde{X}_\tau = x \right]. \quad (53)$$

Constructing a Monte Carlo estimator of the score therefore amounts to sampling from $q_{0|\tau}(\cdot|x)$, the conditional distribution of the noiseless state x_0 given that the noising process reaches x at time τ . This distribution is obtained via Bayes rule and, again for the OU process, is given by $q_{0|\tau}(\cdot|x) \propto \pi(x_0) \mathcal{N}(x; x_0 e^{-\tau}, (1 - e^{-2\tau})I_d)$. Several estimation strategies have been proposed.

On-the-fly estimation The sampler proposed by [Hua+24] runs the denoising dynamics of a diffusion model starting from $t = 0$ and $X_0 \sim \mathcal{N}(0, I)$, but replaces the learned score function with an on-the-fly Monte Carlo estimate of the score. This estimate is obtained by running short Langevin trajectories targeting $q_{0|T-t}(\cdot|x)$ at each time step. At denoising time t , the algorithm to estimate the score $\nabla \log \pi_{T-t}(x)$ proceeds as follows:

algorithm 12: **On the fly score estimation by Langevin dynamics**

Input: Number of Langevin steps L , number of Langevin chains N , Langevin step size η , time t , conditioning point x
Initialize Langevin chains $x_0^{(0,i)}$ for $i = 1 \dots N$
for $\ell = 1 \dots L$ **do**
 for $i = 1 \dots N$ **do**
 $x_0^{(\ell,i)} = x_0^{(\ell-1,i)} + \eta \nabla \log q_{0|T-t}(x_0^{(\ell-1,i)}|x) + \sqrt{2\eta} \xi_{i,\ell}$, with $\xi_{i,\ell} \sim \mathcal{N}(0, I)$
return Score estimate $\hat{s}(t, x) = -\frac{1}{N} \sum_{i=1}^N \frac{x - x_0^{(L,i)} e^{-(T-t_k)}}{1 - e^{-2(T-t_k)}}$

Using a simple Euler-Maruyama discretization of the denoising SDE, the resulting sampling algorithm for the target distribution π is:

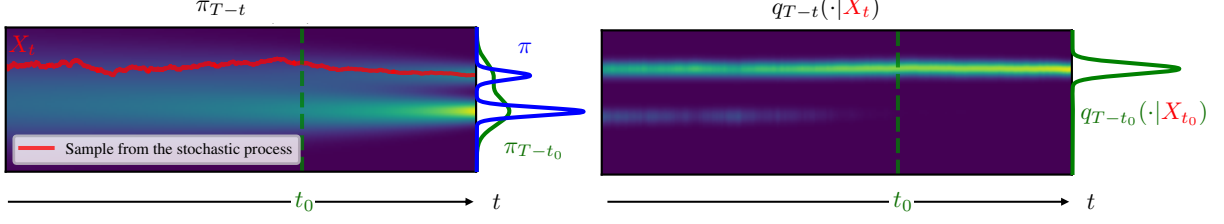


Figure 5: Illustration on a 1d multimodal target distribution π of the transition of the marginal ρ_{T-t} from being unimodal at small times t to multimodal at larger times, and the reciprocal transition of the conditional distribution $q_{0|T-t}(\cdot|x)$ from being multimodal at small times to unimodal at larger times. Adapted from [Gre+24*].

algorithm 13: Reverse diffusion Monte Carlo (RDMC)

Input: Number of denoising steps K , time horizon T , time scheduling $\{t_k\}_{k=1}^K$ with $t_K = T$
Initialize $X_0 \sim \mathcal{N}(0, I)$
for $k = 1 \dots K$ **do**
 $\Delta t_k = t_k - t_{k-1}$
 Estimate the score $\hat{s}(t_k, X_{k-1})$ by running algorithm 12
 Update $X_k = X_{k-1} + (X_{k-1} + 2\hat{s}(t_k, X_{k-1}))\Delta t_k + \sqrt{2\Delta t_k} \xi_k$, with $\xi_k \sim \mathcal{N}(0, I)$
return X_K

The inner loop of Langevin dynamics at each time step of the outer denoising process estimates the score on the fly, without requiring a training dataset. However, the quality of the score estimation is inherently limited by the ability of Langevin dynamics to sample accurately from $q_{0|T-t}(\cdot|x)$. When π is multimodal, $q_{0|T-t}(\cdot|x)$ typically undergoes a transition: it is multimodal close to $t = 0$ (since $q_{0|T} \approx \pi$) and becomes unimodal as t increases (since the conditional distribution approaches a Gaussian centered at the conditioning x rescaled by e^{T-t} , see illustration in fig. 5). Since Langevin dynamics is known to mix poorly when targeting multimodal distributions, this algorithm must be modified when π is multimodal.

Motivated by this observation, [Gre+24*] proposed a modification of the previous algorithm. Since $q_{0|T-t}(\cdot|x)$ becomes increasingly log-concave as t increases, the denoising process can be initialized at a time $t_0 > 0$ for which $q_{0|T-t_0}(\cdot|x)$ is already unimodal. However, this new starting time brings a difficulty. While starting the denoising process at $t = 0$ only requires sampling from the standard Gaussian base distribution, initializing at $t_0 > 0$ instead requires sampling from the intermediate noised marginal

$$\pi_{T-t_0}(x) = \int dx_0 \pi(x_0) \mathcal{N}(x; x_0 e^{-(T-t_0)}, (1 - e^{-2(T-t_0)}) I), \quad (54)$$

which involves an intractable convolution of the target distribution π with a Gaussian kernel. Nevertheless, the score of this distribution, $\nabla \log \pi_{T-t_0}(x)$, is given by the same conditional expectation as in eq. (53) with $\tau = T - t_0$, and can therefore be estimated using algorithm 12. Since the score of a distribution is sufficient to sample from it using Langevin dynamics, [Gre+24*] proposes to run a Langevin-within-Langevin dynamics to initialize $X_{t_0} \sim \pi_{T-t_0}$, before proceeding with the same denoising procedure as in [Hua+24]. The resulting algorithm named stochastic localization through iterative posterior sampling (SLIPS) is summarized in algorithm 14.

The original formulation of SLIPS uses the stochastic localization process as the outer denoising dynamics. This process can be viewed as a variant of the denoising dynamics that starts from a Dirac distribution at $t = 0$ rather than from a Gaussian. Its name originates from the stochastic localization framework introduced in [Eld13; Eld20; Eld22; CE22], where it was developed as a tool to establish results in geometric measure theory. However, the same idea can also be applied

to the standard denoising dynamics used in diffusion models, leading to the algorithm described in algorithm 14.

algorithm 14: Stochastic localization through iterative posterior sampling (SLIPS)

Input: Initialization time t_0 , number of denoising steps K , time horizon T , time scheduling $\{t_k\}_{k=1}^K$ with $t_K = T$, number of Langevin steps L , number of Langevin chains N , Langevin step size η
Initialize X_{t_0} randomly
for $\ell = 1 \dots L$ **do**
 Estimate the score $\hat{s}(t_0, X_{t_0})$ by running algorithm 12
 $X_{t_0} \leftarrow X_{t_0} + \eta(X_{t_0} + 2\hat{s}(t_0, X_{t_0})) + \sqrt{2\eta} \xi_\ell$, with $\xi_\ell \sim \mathcal{N}(0, I)$
for $k = 1 \dots K$ **do**
 $\Delta t_k = t_k - t_{k-1}$
 Estimate the score $\hat{s}(t_k, X_{k-1})$ by running algorithm 12
 Update $X_k = X_{k-1} + (X_{k-1} + 2\hat{s}(t_k, X_{k-1}))\Delta t_k + \sqrt{2\Delta t_k} \xi_k$, with $\xi_k \sim \mathcal{N}(0, I)$
return X_K

Starting at $t_0 > 0$ allows the algorithm to bypass the multimodality of $q_{0|T-t}(\cdot|x)$ at earlier times $t < t_0$, but it also introduces an additional source of error associated with the initialization of the denoising process. The quality of this initialization now depends on the ability of Langevin dynamics to sample from π_{T-t_0} . In contrast to the conditional distribution $q_{0|T-t}(\cdot|x)$, the distribution π_{T-t_0} is unimodal for t_0 small enough (since $\pi_T \approx \mathcal{N}(0, I)$, see again illustration of fig. 5). Hence, the approach of [Gre+24*] accommodates multimodal target distributions π at the cost of identifying a suitable starting time t_0 at which both π_{T-t_0} and $q_{0|T-t_0}(\cdot|x)$ are unimodal, or more conservatively log-concave. This requirement is referred to as the *duality of concavity*. Such a regime is not guaranteed to exist for all target distributions π . [Gre+24*] identifies a class of distributions for which the duality of log-concavity condition can be satisfied and demonstrated in numerical experiments accurate sampling of multimodal distributions beyond this class. However, determining the appropriate starting time t_0 in general remains an open question.

As an alternative to the standard denoising score matching loss used to train diffusion models from data, [Hua+24; Gre+24*] discussed above propose Monte Carlo estimators of the score. These approaches eliminate the need to train a neural network altogether by performing score estimation on the fly during the denoising sampling process, thereby yielding non-parametric methods. While this strategy avoids issues related to model misspecification, it precludes any form of amortization of the score estimation. A natural extension in this direction, discussed in the following paragraphs, is to regress a parametric score model onto the Monte Carlo estimates, enabling faster resampling once training is completed.

Adaptive regression of a Monte Carlo estimator of the score Let $\hat{s}(x, \tau)$ be a Monte Carlo estimator of the score $\nabla \log \pi_\tau(x)$. In order to amortize the score estimation, one can train a parametric model $s_\theta(x, \tau)$ by regressing it onto $\hat{s}(x, \tau)$. Since it is unrealistic to expect $\hat{s}(x, \tau)$ to be accurate everywhere in the state space, a key question is how to select the points at which the regression loss should be evaluated. Ideally, the learned score should be accurate in regions of high probability under π_τ that are likely to be visited during the denoising process. In the absence of training data, however, achieving such localization is extremely challenging

In a philosophy closely related to the adaptive training strategies for NFs described in section 4.1.2, the approach proposed by [Akh+24] leverages the availability of a small number of low-quality target samples, which in particular provide information about the location of the different modes. These initial samples are used to initialize a buffer of datapoints. Points from this buffer are

used as initial conditions for noising trajectories to generate an initial set of points on which the regression loss is evaluated. The learned score model is subsequently used to generate new denoised samples, which are added to the buffer, and the procedure is then repeated. This iterative scheme allows the score model to progressively improve its accuracy in regions of high probability under π_t that are most relevant for sampling from π . Using a mean squared regression loss, and a simple Euler-Maruyama discretization of the noising process, the resulting algorithm can be sketched as follows:

algorithm 15: Iterated score regression

Input: Initial score model parameters θ , learning rate η , noising time step $\Delta\tau$, number of iterations K , size of batch of trajectories per iteration B , $x^{(0,m)}$ for $m = 1 \dots M$ a few low-quality samples from π
Initialize buffer $\mathcal{B} = \{x^{(0,m)}\}_{m=1}^M$
for $k = 1 \dots K$ **do**:
 Sample $x^{(i)} \sim \mathcal{B}$ for $i = 1 \dots B$
 Initialize $\tilde{x}_0^{(i)} = x^{(i)}$ for $i = 1 \dots B$
 for $\tau = 0 \dots T$ **do**
 $\tilde{x}_\tau^{(i)} = (1 - \Delta\tau)\tilde{x}_{\tau-1}^{(i)} + \sqrt{2\Delta\tau}\xi_i$ with $\xi_i \sim \mathcal{N}(0, I)$ for $i = 1 \dots B$ (noising trajectories)
 Evaluate regression loss $\mathcal{L}(\theta) = \frac{1}{B(T+1)} \sum_{i=1}^B \sum_{t=0}^T \|s_\theta(t, \tilde{x}_t^{(i)}) - \hat{s}(t, \tilde{x}_t^{(i)})\|^2$
 $\theta \leftarrow \theta - \eta \nabla \mathcal{L}(\theta)$
 Sample $x^{(k,m)}$ from the diffusion model using the current score model s_θ for $m = 1 \dots M$
 $x^{(k,m)} \leftarrow \text{StopGrad}(x^{(k,m)})$ for $m = 1 \dots M$ (prevent backprop. through sampling process)
 Add new samples to the buffer $\mathcal{B} \leftarrow \mathcal{B} \cup \{x^{(k,m)}\}_{m=1}^M$
return trained score model s_θ

In their approach, [Akh+24] employ an importance sampling estimator of the score derived from the conditional expectation formula in eq. (53). In practice, however, importance sampling is not expected to scale well to high-dimensional settings. Nevertheless, the idea of amortizing score estimations through a regression model evaluated on an adaptive buffer of samples is appealing. Overall, the remaining open question in this line of research estimating the score by Monte Carlo concerns the choice of the sampler for the estimator, as both Langevin dynamics and importance sampling suffer from intrinsic limitations, as discussed above.

4.3.3 Revisiting variational inference with stochastic optimal control

Another line of work, which bypasses both the need to estimate likelihoods and the need for training samples when learning diffusion models, exploits the connection between diffusion models and stochastic optimal control problems [TR19; ZC21; BRU23; VGD23; RB23; Zha+23; Nob+25*; Hav+25]. This connection can also be interpreted through the lens of variational inference. Without aiming for full mathematical rigor, we provide below a high-level sketch of this relationship.

For simplicity, we restrict attention to the OU noising process of the target distribution π , denoted $\tilde{X}_\tau$ for $\tau \in [0, T]$, and to the associated exact denoising process X_t for $t \in [0, T]$, respectively defined by the SDEs

$$d\tilde{X}_\tau = -\tilde{X}_\tau d\tau + \sqrt{2}dW_\tau, \quad \tilde{X}_0 \sim \pi \quad (55)$$

$$dX_t = (X_t + 2\nabla \log \pi_{T-t}(X_t)) dt + \sqrt{2}dW_t, \quad X_0 \sim \mathcal{N}(0, I). \quad (56)$$

The trajectories $(X_t)_{t \in [0, T]}$ are distributed according to a path measure $\mathbb{P}_\pi$. The goal of sampling from π can then be rephrased as seeking to sample from $\mathbb{P}_\pi$. Following VI, one may seek to

approximate $\mathbb{P}_\pi$ by a variational path measure $\mathbb{P}_\theta$, generated by an SDE similar to the denoising process but parametrized by a score model s_θ :

$$\mathbb{P}_\theta \stackrel{\text{law}}{=} dX_t = (X_t + 2s_\theta(X_t, T-t)) dt + \sqrt{2} dW_t, \quad X_0 \sim \mathcal{N}(0, I). \quad (57)$$

The parameters of the score s_θ are learned by minimizing the KL divergence between the two path measures $\mathbb{P}_\pi$ and $\mathbb{P}_\theta$. A tractable expression of this KL divergence can be derived by introducing a reference path measure $\mathbb{P}_{\text{ref}}$. This path measure is the law of the trajectories $(X_t^{\text{ref}})_{t \in [0, T]}$ of the time-reverse of the OU noising process $\tilde{X}_\tau^{\text{ref}}$ initialized at $t = 0$ with a distribution ρ_{ref} :

$$d\tilde{X}_\tau^{\text{ref}} = -\tilde{X}_\tau^{\text{ref}} d\tau + \sqrt{2} dW_\tau, \quad \tilde{X}_0^{\text{ref}} \sim \rho_{\text{ref}} \quad (\text{with time marginal } \rho_{\text{ref}, \tau}) \quad (58)$$

$$dX_t^{\text{ref}} = (X_t^{\text{ref}} + 2\nabla \log \rho_{\text{ref}, T-t}(X_t^{\text{ref}})) dt + \sqrt{2} dW_t, \quad X_0^{\text{ref}} \sim \mathcal{N}(0, I). \quad (59)$$

The reference distribution ρ_{ref} is chosen to be sufficiently simple so that the score of its noised marginals $\rho_{\text{ref}, \tau}$ is tractable at all times, which is the case, for instance, when ρ_{ref} is Gaussian. Reparametrizing the score of the variational path measure as $s_\theta(x, t) = \nabla \log \rho_{\text{ref}, T-t}(x) + u_\theta(x, t)$, with u_θ being the residual learnable term, the KL divergence between $\mathbb{P}_\pi$ and $\mathbb{P}_\theta$ can then be expressed as

$$\begin{aligned} D_{\text{KL}}(\mathbb{P}_\theta \| \mathbb{P}_\pi) &= \mathbb{E}_{\mathbb{P}_\theta} \left[\log \frac{d\mathbb{P}_\theta}{d\mathbb{P}_\pi} \right] = D_{\text{KL}}(\mathbb{P}_\theta \| \mathbb{P}_{\text{ref}}) - \mathbb{E}_{\mathbb{P}_\theta} \left[\log \frac{\rho_{\text{ref}}(X_T)}{\pi(X_T)} \right] \\ &= \mathbb{E}_{\mathbb{P}_\theta} \left[\int_0^T \sqrt{2} \|u_\theta(X_t, T-t)\|^2 dt - \log \frac{\rho_{\text{ref}}(X_T)}{e^{-U(X_T)}} \right] + \text{const.} \end{aligned} \quad (60)$$

The second equality follows from Girsanov's theorem. The resulting training objective eq. (60) directly reads as a stochastic optimal control problem, with u_θ playing the role of the control. The control is learned so as to minimize a cost functional balancing a running cost $\int_0^T \|u_\theta(X_t, T-t)\|^2 dt$, which penalizes deviations from the reference process, and a terminal cost $\log \frac{\rho_{\text{ref}}(X_T)}{e^{-U(X_T)}}$, which encourages the terminal state X_T to follow the target distribution π .

This objective can be optimized by sampling trajectories from the variational path measure $\mathbb{P}_\theta$ and evaluating the cost functional along these trajectories. The tractability of the scores of the reference process ensures that $\mathbb{P}_\theta$ can be sampled efficiently, while the terminal cost only requires evaluating the unnormalized density of the target distribution π . This formulation provides a principled way to train diffusion models for sampling without access to a training dataset or to the likelihood of the model.

However, this training objective is computationally expensive, as it requires re-sampling, and backpropagating through, trajectories from the variational path measure $\mathbb{P}_\theta$ at each training iteration to evaluate the gradients of the cost functional. To alleviate this cost, [RB23] proposes an alternative objective, termed the log-variance loss. In place of the KL—that is, the expectation under the variational measure of the logarithm of the Radon-Nikodym derivative between the variational and target path measures—, the log-variance loss is the variance of this same log-derivative with respect to an arbitrary path measure $\tilde{\mathbb{P}}$:

$$D_{\text{LV}}(\mathbb{P}_\theta \| \mathbb{P}_\pi) = \mathbb{V}_{\tilde{\mathbb{P}}} \left[\log \frac{d\mathbb{P}_\theta}{d\mathbb{P}_\pi} \right]. \quad (61)$$

This quantity can also be re-expressed into a tractable form using the reference process. The possibility to use an arbitrary path measure $\tilde{\mathbb{P}}$ to evaluate the variance provides additional flexibility and circumvent the need to backpropagate through trajectories. In practice, a common choice remains $\tilde{\mathbb{P}} = \text{StopGrad}(\mathbb{P}_\theta)$ ⁶ as in the original KL formulation, albeit potentially re-using

⁶The dependence in the trainable parameters of the argument of StopGrad is ignored during learning.

a buffer of trajectories generated at previous values of θ . The log-variance loss is shown to exhibit lower variance and, to some extent, to be less prone to mode collapse than the traditional KL objective.

Among other attempts to reduce computational cost, [Zha+23] uses partial trajectories and [Hav+25] leverages the adjoint method to compute gradients of the KL divergence with respect to θ , showing that under some conditions the adjoint-base objective can be heuristically simplified to avoid systematic resampling of trajectories at each training iteration.

Beyond these computational considerations, two additional aspects of this variational approach warrant careful examination: the parametrization of the control and the choice of the reference distribution ρ_{ref} . Concerning the control, [He+25] show that a so-called *Langevin preconditioning*, first introduced by [ZC21], is necessary for the resulting generative process to properly cover the target distribution. This preconditioning consists in informing the score of the variational path measure $\mathbb{P}_\theta$ with the score of the target distribution π following:

$$s_\theta(x, t) = s_{\theta_1, \theta_2}(x, t) = \text{NN}_{\theta_1}(x, t) + \text{NN}_{\theta_2}(t) \nabla \log \pi(X), \quad (62)$$

with $\text{NN}_{\theta_1} : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}^d$ and $\text{NN}_{\theta_2} : [0, 1] \rightarrow \mathbb{R}$ neural networks initialized such that $\text{NN}_{\theta_1} \approx 0_d$ and $\text{NN}_{\theta_2} \approx 1$ at initialization. This parametrization ensures that, in the early stages of training, the denoising process closely resembles a Langevin dynamics targeting the distribution π , and therefore will progressively locate high-probability regions of the target.

Concerning the path measure, in principle, any $\mathbb{P}_{\text{ref}}$ can be used within the variational framework, provided that the scores of the noised marginals $\rho_{\text{ref}, \tau}$ remain tractable. In practice, however, the choice of ρ_{ref} has a significant impact on training outcomes, particularly when targeting multimodal distributions, as evidenced by [Nob+25*]. In that work, the prescriptions of [ZC21] and [VGD23] to use a Gaussian reference distribution are shown to recover the correct mode weights of a bimodal target distribution only when the reference corresponds to the best Gaussian approximation of the target, even when using the more favorable log-variance training objective. This observation motivates an alternative strategy.

Acknowledging the need to adapt the reference distribution to the target in order to address the challenges posed by multimodality, [Nob+25*] propose a pretraining strategy that learns the reference distribution ρ_{ref} itself. As in adaptive methods for NFs and in the iterated score regression approach discussed earlier, this learning procedure leverages a small number of low-quality samples from the target distribution π , which in particular provide information about the location of the modes. These initial samples can be used either to fit a Gaussian mixture model or to train a time-dependent energy-based model $\rho_{\text{ref}, \tau}$, for both of which the scores of the noised marginals are tractable. The resulting learned reference distribution is then employed to define the variational path measure $\mathbb{P}_\theta$ and to train the control u_θ , as described above. This approach is shown to significantly improve the performance of the resulting sampler on multimodal targets.

*

We have now reviewed the main approaches proposed to train NFs—both in their discrete and continuous formulations—as well as diffusion models, in the sampling setting, where no training data from the target distribution are available a priori.

For discrete-time NFs, the conceptual picture is essentially complete. As discussed in section 3, once the model has been trained to approximate the target distribution, it can be embedded within sampling schemes that enjoy rigorous convergence guarantees.

For continuous-time models—namely CNFs and diffusion models—more remains to be said. The exact sampling strategies presented in section 3 rely critically on the ability to evaluate the model likelihood, a property that is readily available for discrete-time NFs but generally challenging in

continuous time. This fundamental obstacle prevents a direct transposition of those constructions. We now turn to the first proposals in the literature that aim to overcome this difficulty and to construct exact samplers leveraging the generative capabilities of continuous-time models.

5 Verifiable sampling with continuous-time generative models

5.1 Reweighting for continuous normalizing flows and stochastic interpolants

As discussed previously, a key advantage of continuous-time generative models is that they can be constructed to preserve the invariances of the target distribution (see the discussion in section 2.1.2 around eq. (9)). This property is particularly important in physics applications, where the distribution of interest is often invariant under groups of transformations such as rotations, translations, or permutations. In addition, unlike diffusion-based models, the initial distribution of CNFs can be chosen freely, which—as already noted for discrete-time NFs (see paragraph Adapting the base distribution in section 3.3)—can significantly facilitate sampling. This freedom in choosing the base distribution is also shared by stochastic interpolants (SIs) (see section 2.3), which differ in their training objective while inducing essentially the same generative dynamics as CNFs. Finding ways to leverage CNFs and SIs for sampling is therefore a key question in the field.

Deterministic models A first, conceptually straightforward approach to using CNFs and SIs for verifiable sampling consists in accepting the cost of (approximately) evaluating the likelihood by integrating the instantaneous change-of-variables formula eq. (8). Denoting v_θ the velocity field of the CNF or SI, the logarithm of the importance weight of a sample X_T , stemming from the computation of its likelihood along the generating trajectory from a base sample X_0 , is

$$\log w(X_T) = \log \frac{\pi(X_T)}{\rho_0(X_0)} + \int_0^T \nabla \cdot v_\theta(X_t, t) dt. \quad (63)$$

This strategy was first explored by [KKN20], who used a CNF to sample interacting particle systems with pairwise interactions. Their velocity field architecture relies on a decomposition into radial basis functions of pairwise distances and is specifically designed to be equivariant and so that the divergence of the vector field can be computed efficiently via automatic differentiation (requiring a single backward pass rather than as many passes as the dimension of the field in a naive implementation). Thereby, [KKN20] were able to successfully reweight CNF-generated configurations to the target Boltzmann distribution for simple interacting particle systems. In contrast, they showed that replacing the exact divergence with a Hutchinson estimator—commonly used in the CNF literature to reduce computational cost—yields inaccurate importance weights and significantly degrades sampling performance.

Building on the same architecture, [JBB24] considered more challenging systems, in particular amorphous particle systems at decreasing temperatures approaching the glass transition. To address the increased complexity, they selected as base distribution a higher-temperature Boltzmann distribution rather than the centered Gaussian used in [KKN20]. Their promising results showed performance comparable to PT and SMC in the difficult regime near the glass transition.

Turning to SIs, [Gre+25*] also investigated amorphous particle systems, examining in detail how to make SIs invariant to the symmetries of multispecies systems in periodic boxes. Compared to CNFs, SIs rely on a different training objective that is more computationally efficient as it does not require likelihood evaluation. This removes the architectural constraints imposed in [KKN20] to speed up divergence computation during training and allows for more expressive equivariant velocity fields. However, the absence of the architectural constraints renders the weight computation during reweighting more costly. Accepting this additional cost, [Gre+25*] successfully sampled amorphous particle systems with SIs. Moreover, the increased expressivity of the velocity field allowed them to use a uniform distribution on a periodic box as a base measure.

This avoids the need to generate base samples via MCMC when using a high-temperature Boltzmann distribution as in [JBB24].

Reweighting and resampling for (stochastic) annealed flows The importance reweighting strategy described above can be revisited in the context of annealed continuous flows, presented in section 4.2, trained to follow a path $(\pi_t)_{t \in [0,1]}$ between the base distribution ρ_0 and the target distribution π . This is the proposition of [TPL24; Var+24; AV25].

Adopting again the Boltzmann representation $\pi_t = e^{-U_t}/Z_t$, recall that the generative dynamics of such a (possibly stochastic) CNF with control-velocity field v_θ are given by (see eq. (47))

$$dX_t = v_\theta(X_t, t) dt - \gamma_t \nabla U_t(X_t) dt + \sqrt{2\gamma_t} dW_t. \quad (64)$$

The time-dependent diffusion coefficient γ_t vanishes for deterministic CNFs and is strictly positive for stochastic ones. The importance weight associated with X_t from a trajectory $(X_s)_{s \in [0,t]}$ for sampling from π_t can then be expressed as

$$\log w_t \left((X_s)_{s \in [0,t]} \right) = - \int_0^t \left(\partial_s U_s(X_s) + v_\theta(X_s, s) \cdot \nabla U_s(X_s) + \nabla \cdot v_\theta(X_s, s) \right) ds. \quad (65)$$

This expression can be derived by manipulating the Fokker-Planck equation associated with the controlled Langevin dynamics (see, e.g., the proof of Proposition 2.4 in Appendix A of [AV25]). Importantly, the weight now depends explicitly on the entire trajectory, which is itself a random realization of the stochastic dynamics, rather than solely on the terminal configuration. This contrasts with direct reweighting for deterministic CNFs or SIs, where the trajectory is fully determined by its endpoint once the velocity field is fixed. In the deterministic case $\gamma_t = 0$, the expression above reduces to the previously obtained formula eq. (63), as the first two terms in the integrand combine into the material derivative of the potential energy function along the trajectory.

Conceptually, this annealed approach is closely related to the SMC-inspired (CR)AFT algorithm previously discussed for discrete-time flows [AMD21; Mat+22]. Whereas (CR)AFT relies on a finite composition of transport maps interleaved with MCMC steps, continuous-time annealed flows incorporate both ingredients directly into their dynamics: a transport term, $v_\theta(X_t, t) dt$, and a Langevin diffusion term, $-\gamma_t \nabla U_t(X_t) dt + \sqrt{2\gamma_t} dW_t$. In direct analogy with (CR)AFT and more generally with SMC, the importance weights can be used to perform resampling at intermediate times along the trajectory. Resampling may be triggered when the ESS falls below a prescribed threshold, thereby mitigating weight degeneracy and improving overall sampling performance.

Nevertheless, the need to evaluate the divergence of the velocity field to compute the weight remains in this annealed formulation. This is the key computational bottleneck for CNFs and SIs. Restrictive architectures enable efficient divergence computation but may lack sufficient flexibility for complex targets, while more expressive models incur a significant computational cost during reweighting. Moreover, an uncontrolled source of bias in these sampling strategies using continuous-time formulation is the eventual discretization of the time integrations. A couple of recent works have proposed to circumvent both of these issues by reinterpreting the generative process of diffusion models as a stochastic normalizing flow, which allows for a principled construction of exact samplers without any likelihood evaluation, even at the reweighting stage. We present these approaches next.

5.2 Diffusion samplers as stochastic flows

To avoid computing likelihoods or importance weights involving the divergence of the velocity field, one possible route to debias diffusion samples is to reinterpret the generative process as a discrete-time stochastic normalizing flow (see paragraph Combining NF with AIS: stochastic normalizing flows in section 3.3). This perspective underlies the approaches of [Phi+24] and [ZMH25], which differ slightly in their construction of the stochastic flow.

We present here the main idea, focusing on the backward SDE of the OU process. Using the approximate score function s_θ learned by the diffusion model, the backward dynamics reads

$$dX_t = (X_t + 2s_\theta(X_t, T - t)) dt + \sqrt{2} dW_t. \quad (66)$$

Discretizing the time interval $[0, T]$ with a grid $0 = t_0 < t_1 < \dots < t_K = T$, the joint distribution of the states $(x_k)_{k=0}^K$ visited by the Markovian backward process factorizes as

$$\rho_\theta(x_0, \dots, x_K) = \rho_0(x_0) \prod_{k=1}^K q_{\theta,k}(x_k | x_{k-1}). \quad (67)$$

Although the transition kernel $q_{\theta,k}$ is non-trivial due to the non-linear score function, it can be approximated by a Gaussian distribution obtained from a discretization of the backward SDE. Going beyond the naive Euler-Maruyama scheme, [Phi+24] propose using an exponential integrator yielding the Gaussian transition kernel

$$q_{\theta,k}(x_k | x_{k-1}) = \mathcal{N}\left(x_k; e^{-\Delta t_k} x_{k-1} + \mu_k, (1 - e^{-2\Delta t_k}) I_d\right), \quad (68)$$

with $\Delta t_k = t_k - t_{k-1}$ and $\mu_k = 2(1 - e^{-\Delta t_k})(x_{k-1} + s_\theta(x_{k-1}, t_{k-1}))$.⁷ This construction completes the definition of the stochastic flow defined in eq. (67).

To debias the samples generated by this stochastic flow, one considers the forward OU process started from the target distribution π , discretized on the same time grid. This forward process is exactly tractable and defines the joint distribution

$$\pi(x_0, \dots, x_K) = \pi(x_K) \prod_{k=1}^K p_k(x_{k-1} | x_k), \quad (69)$$

where

$$p_k(x_{k-1} | x_k) = \mathcal{N}\left(x_{k-1}; e^{-\Delta t_k} x_k, (1 - e^{-2\Delta t_k}) I_d\right). \quad (70)$$

Importance weights are then computed as the ratio between these two joint path distributions

$$w(x_0, \dots, x_K) = \frac{\pi(x_0, \dots, x_K)}{\rho_\theta(x_0, \dots, x_K)}. \quad (71)$$

Crucially, this ratio only involves Gaussian transition kernels and the target density $\pi(x_K)$. Therefore, it avoids any explicit likelihood computation or evaluation of divergence terms. In the overwhelmingly common case where π is only known up to a normalizing constant, the self-normalized weights and estimators are used in practice (see e.g. algorithm 1.)

Following this paradigm, [Phi+24] further develop a SMC version of the stochastic flow, incorporating intermediate resampling steps along the trajectory. For this, they not only learn the score

⁷This expression follows from rewriting the backward SDE as $dX_t = -X_t dt + 2(X_t + s_\theta(X_t, T - t)) dt + \sqrt{2} dW_t$, applying an exponential integrator to the linear term $-X_t$, and treating $X_t + s_\theta(X_t, T - t)$ as frozen over the time step. This decomposition is consistent with the parametrization adopted in [Phi+24], where the network directly models $x + s_\theta(x, t)$.

of the noising process but also the time marginal densities π_t along the trajectory, which allows them to compute the importance weights before the end of the trajectory and thus to perform resampling at intermediate times. As usual, resampling is triggered whenever the effective sample size (ESS) of the importance weights falls below a prescribed threshold.

In contrast, [Zha+25] sticks to the vanilla AIS version of the reweighting in the joint space and considers a variance-exploding diffusion (i.e., a Wiener process) instead of the OU process. This modeling choice leads to a different approximation of the denoising transition kernels. Importantly, they propose to learn the variance of these kernels so as to directly minimize the variance of the resulting importance weights, yielding substantial improvements in practice.

Overall, interpreting diffusion models through the lens of stochastic flows provides a principled framework for exact sampling without requiring costly divergence computations. However, in order to maintain importance weights with controlled variance, the joint laws over the discretized forward and backward processes must be sufficiently close. For this, the time discretization typically needs to become finer as the dimensionality of the problem increases. This refinement can significantly raise the computational cost, potentially limiting scalability in high-dimensional settings.

*

This concludes our review of emerging methods for verifiable sampling with continuous-time generative models. We have seen that while direct reweighting of CNF and SI samples is conceptually straightforward, it faces a fundamental computational bottleneck due to the need to evaluate the divergence of the velocity field. This bottleneck can be circumvented by reinterpreting continuous-time models as stochastic flows, as was proposed for diffusion models. However, this approach introduces its own challenges, particularly in terms of controlling the variance of importance weights in high-dimensional settings.

6 Examples of applications

After this effort to rationalize the emerging—yet already dense—literature on sampling assisted by transport-based generative models, I showcase in this section several applications of these methods that I have been involved in. These examples provide a glimpse into their practical implementation, highlighting the problem-specific design choices that must be made and the types of results that can be obtained.

6.1 Training energy-based models

6.1.1 Definition and challenges

We first turn to an application in generative modelling, namely the training of energy-based models (EBMs). EBMs are a class of generative models that define a probability distribution through an unnormalized density,

$$p_\phi(x) = \frac{e^{-E_\phi(x)}}{\mathcal{Z}_\phi}, \quad (72)$$

where E_ϕ is a learnable energy function and $\mathcal{Z}_\phi$ the associated normalization constant. The appeal of EBMs lies in their flexibility: by parameterizing the energy function with neural networks, they can model complex distributions thanks to the universal approximation capabilities of these architectures. Allowing the energy to take infinite values, any distribution can be represented as an EBM, including distributions with disconnected support, which are difficult to capture with latent generative models such as VAEs and NFs.

However, EBMs are difficult to sample from and therefore challenging to train. Given a training data distribution p_{data} , the log-likelihood $\ell_{\text{EBM}}(\phi) = \mathbb{E}_{p_{\text{data}}}[\log p_\phi(x)]$ has gradient

$$\nabla_\phi \ell_{\text{EBM}}(\phi) = \mathbb{E}_{p_{\text{data}}}[-\nabla_\phi E_\phi(x)] + \mathbb{E}_{p_\phi}[\nabla_\phi E_\phi(x)]. \quad (73)$$

The first term can be computed directly from the training data, whereas the second term is an expectation under the model distribution p_ϕ . Since this distribution does not admit a direct sampling procedure, the expectation must be estimated using MCMC methods. This constitutes a major bottleneck in the training of EBMs, as the MCMC chain must in principle be run to convergence at each training step, which is computationally expensive and can lead to biased gradients if not done properly.

Despite this challenge, simple heuristics have been shown to work well in practice, and in particular made restricted Boltzmann machines among the first generative models successfully applied to real-world problems in the early 2000s. These heuristics rely on short MCMC chains initialized either from the data distribution (contrastive divergence [Hin02]) or from a replay buffer of past samples (persistent contrastive divergence [Tie08]). This idea was revisited in the early 2020s using modern neural network architectures, highlighting that an important ingredient is to use matching, if approximate, sampling procedures during training and evaluation in order to obtain good sample quality [Nij+19; Nij+20; SK21; Ago+25].

Yet, this does not guarantee that the resulting model is a good EBM, in the sense that it can be sampled with a convergent MCMC and reproduce the statistics of the data distribution. In particular, if this is not the case, the learned energy function cannot be interpreted in a straightforward manner. This issue is especially problematic when dealing with multimodal distributions, where the MCMC is likely to get stuck in a single mode and fail to explore the others, leading to biased gradients and poor sample quality.

This difficulty motivated the development of training methods relying on more sophisticated samplers inspired by the statistical mechanics literature [Bér+23; Ber+24]. Another line of

work attempted to combine latent generative models with EBMs in order to leverage their complementary strengths. [Xie+18; XZL21] suggested using a Variational Autoencoder (VAE), and [Xie+22] a NF (CoopFlow algorithm), to provide initial sampling points before running short MCMC chains during maximum likelihood training, reporting good sample quality but no guarantee of statistical accuracy.

Closer to our concern of relying on a converged sampler, the NT-EBM algorithm [XYA20; Nij+22] leverages a NF-assisted sampler belonging to the family of NeutraMCMCs (algorithm 4), which however struggles to mix between modes in multimodal distributions (see discussion in section 3.2). [Xia+21] explored a similar idea using VAEs. Finally, [AZG21] combined an EBM with a Generative Adversarial Network using a Langevin sampler in latent space, although statistical performance on multimodal datasets was not reported.

Below, we discuss our proposal to leverage the fast mixing properties of NFs to train EBMs with a converged sampler [GMG23*].

6.1.2 Proposed method: flowMC-EBM

We propose to train a NF and an EBM cooperatively. The NF provides a fast-mixing sampler that is used to train the EBM by maximum likelihood on the training data, while the NF itself is trained by maximum likelihood on samples from the EBM so as to maintain a good overlap between its pushforward distribution and the EBM distribution.

The flow probability density ρ_θ , as defined above in this manuscript, is the pushforward of the base distribution ρ_0 by the flow map T_θ . To ensure good initial agreement between the flow and the EBM, T_θ is initialized to the identity map and the definition of the EBM is modified as follows:

$$p_\phi(x) = \frac{1}{Z_\phi} e^{-E_\phi(x)} \rho_0(x), \quad (74)$$

with $E_\phi \approx 0$ at initialization. Note that this modification does not change the formal expression of the gradient of the EBM log-likelihood (73). The resulting joint training algorithm is:

algorithm 16: flowMC-EBM training

Input: training data $\{x^{(i)}\}_{i=1}^N$, number of training steps K , learning rates η_{EBM} and η_{NF} , batch size B .
Initialize $T_\theta(\cdot) = \text{Id}(\cdot)$
Initialize $E_\phi(\cdot) \approx 0$.
Initialize B Markov chains $\{\tilde{x}^{(i)}\}_{i=1}^B$ with samples from ρ_θ .
for $k = 1 \dots K$ **do**
 Sample a batch of data $\{x^{(i)}\}_{i=1}^B$ from the training set.
 Update $\tilde{x}^{(i)}$ with flow-IMH (algorithm 2) using ρ_θ as proposal and p_ϕ as target.
 Estimate EBM gradient: $\nabla_\phi \widehat{\ell_{\text{EBM}}}(\phi) = \frac{1}{B} \sum_{i=1}^B \left[-\nabla_\phi E_\phi(x^{(i)}) + \nabla_\phi E_\phi(\tilde{x}^{(i)}) \right]$
 Update EBM parameters: $\phi \leftarrow \phi + \eta_{\text{EBM}} \widehat{\nabla_\phi \ell_{\text{EBM}}}(\phi)$.
 Update NF parameters: $\theta \leftarrow \theta + \eta_{\text{NF}} \nabla_\theta \left(\frac{1}{B} \sum_{i=1}^B \log \rho_\theta(\tilde{x}^{(i)}) \right)$
return EBM parameters ϕ and NF parameters θ .

Here, the NF-assisted sampler targets the evolving EBM distribution. The NF is therefore trained on the fly to maintain good overlap with this target. In practice, the matching initialization of the EBM and the NF is indispensable for success, and the learning rates η_{EBM} and η_{NF} must be co-adjusted so that this matching is approximately maintained throughout training. Provided this condition holds, the fast-mixing NF-assisted sampler yields converged samples from the EBM

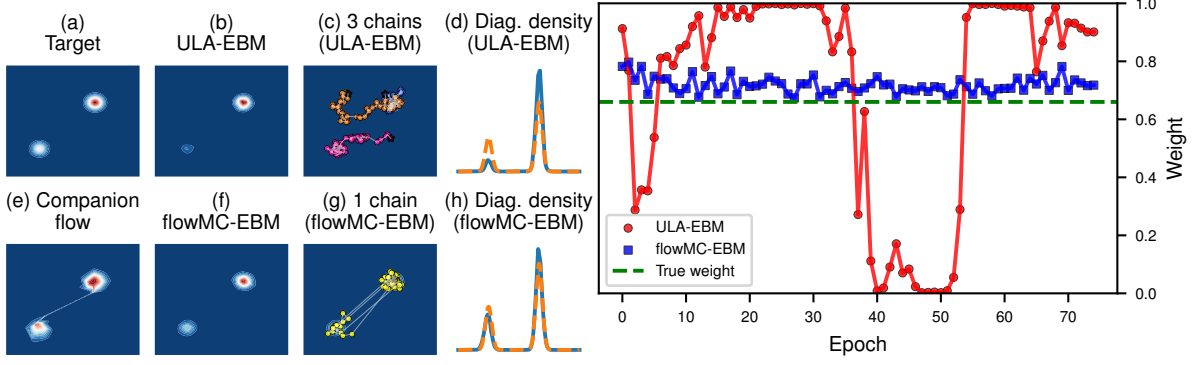


Figure 6: **(Left) Comparison of ULA persistent contrastive divergence and flowMC-EBM training on a toy 2D mixture with a 2:1 weight ratio.** The EBM learned with flowMC (f) correctly captures the relative weights, unlike the EBM trained with ULA (b), as also shown by the conditional densities along the axis passing through the centroids (d,h). This statistical accuracy is promoted by the fast-mixing NF-assisted sampler (g). **(Right) Estimated weight of the top-right mode during the training of ULA-EBM and flowMC-EBM.** The weight is computed by summing the probability mass in the top-right and bottom-left corners of the target distribution. The improved accuracy comes at the cost of training a NF additionally to the EBM for flowMC-EBM compared to ULA-EBM. Adapted from [GMG23*].

in only a few MCMC steps at each training iteration, which promotes the statistical accuracy of the resulting EBM.

6.1.3 Preliminary results

To illustrate the potential of this approach and the importance of using a converged sampler, we consider a dataset drawn from a Gaussian mixture distribution in $\mathbb{R}^2$ with two modes of weight ratio 2:1. We train an EBM on this dataset using the flowMC-EBM algorithm described above and compare it with a more standard training procedure based on persistent contrastive divergence with the Unadjusted Langevin Algorithm (ULA) as the MCMC sampler, which is known to struggle to mix between modes in multimodal distributions. Results are shown in fig. 6.

As ULA gets trapped in the modes ((c) versus (g)), the gradient estimates become biased. As a result, the EBM trained with ULA oscillates between over- and under-estimating the mode weights multiple times during training (fig. 6, Right), and the final value of the mode weight is essentially random across different runs.

In contrast, during flowMC-EBM training, the use of converged MCMC samples from the EBM ((f) and (h)) leads to a stable estimation of the mode weights. Note that the companion flow, which is more constrained in its parametrization, does not achieve a fit as accurate as the EBM. Nevertheless, its match with the EBM remains sufficient to enable the fast-mixing MCMC that is key to the success of the method.

*

Additional experiments and comparisons with other methods are provided in [GMG23*], confirming the importance of using a converged sampler to train well-calibrated EBMs. This approach nevertheless has two main limitations. First, it requires training two models instead of one, increasing the computational cost and requiring careful tuning of the learning rates to maintain a good match between them during training. Second, while the algorithm performs well in

moderately high dimensions, it is expected to struggle in very high dimensions, as the NF-assisted flow-IMH sampler is unlikely to scale well with dimension.

The method is therefore particularly promising for problems of moderate dimension where the functional form of the energy is interpretable and where statistical accuracy of the EBM is essential for extracting meaningful information from the learned energy function.

6.2 Sampling isomers of metallic nanoclusters

6.2.1 Motivation and test system

Next, we turn to an application in computational chemistry. Metallic nanoclusters have a wide range of promising applications, including catalysis, sensing, and drug delivery. They are small aggregates of atoms that can adopt different structural arrangements, called isomers, which can exhibit very different properties. Understanding the thermodynamic properties of these isomers can therefore aid in the design of nanoclusters with desired functionalities.

While, in principle, the equilibrium distribution of isomers can be obtained by running long molecular dynamics (MD) simulations, the computational cost of such simulations is often prohibitive for two reasons. First, the potential energy surface of systems composed of heavy atoms is complex and must be evaluated with high accuracy, which can only be achieved with expensive quantum chemistry methods, also referred to as *ab initio* calculations. Second, metastable isomers are typically separated by high energy barriers that trap local MD simulations.

In [Mol+24*], we propose leveraging NF-assisted MCMC samplers to efficiently explore the state space of metallic nanoclusters and obtain accurate estimates of the equilibrium distribution. We focus on the case of sampling clusters of six silver atoms. This system has been reported to exhibit two coexisting isomers at room temperature—a planar one and a pyramidal one—whose relative occurrence we characterize precisely.

Two main adjustments were necessary to apply the adaptive joint sampling and training scheme A-flowMC (algorithm 10) in this setting.

6.2.2 Adjustments to A-flowMC

A mixture of flows The first adjustment was to employ two separate NFs, one for each isomer, and to use a mixture of these two flows as the proposal distribution in the MCMC sampler. This practical choice stemmed from two observations.

First, the two isomers are well separated in state space. While NFs are capable of learning multimodal distributions, doing so is typically more difficult—and requires larger batches and more training iterations—than learning two unimodal distributions separately. Given the high cost of the adaptive training procedure, which requires querying a quantum force field, we opted for the latter approach. A flow was learned adaptively for each isomer, yielding two unimodal proposal distributions, ρ_{θ_0} and ρ_{θ_1} . The resulting proposal distribution is then a mixture of these two flows, $\rho_{\theta} = w\rho_{\theta_0} + (1 - w)\rho_{\theta_1}$, where w controls the relative weight of each flow.

Second, the use of a mixture of flows allows one to maintain good sampling efficiency even when one of the isomers is much less populated than the other. In this setting, a single flow would struggle to learn the less populated mode, leading to poor sampling efficiency and inaccurate estimates of its population. In contrast, by using a mixture of flows and enforcing a minimum probability for each mode, both modes remain well represented in the proposal distribution, resulting in efficient sampling and accurate estimates of their populations. This strategy has also been mentioned for applications in other physical contexts by [GRV22*; Hac+21].

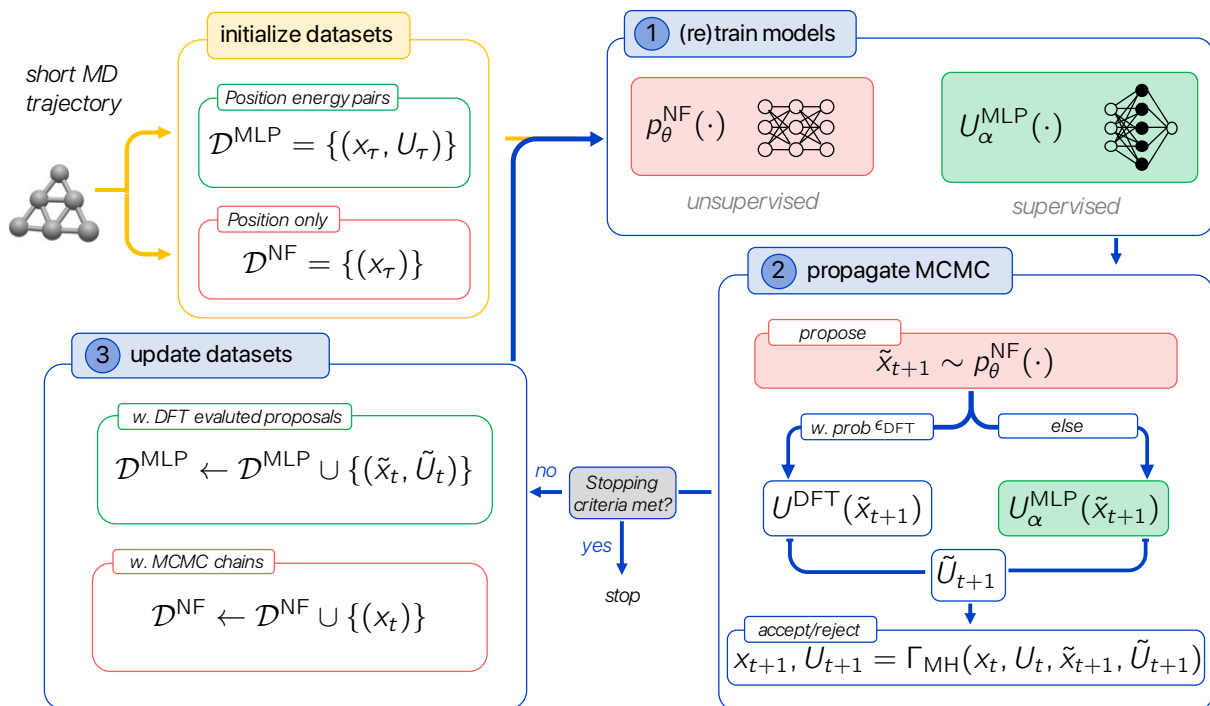


Figure 7: *Ab initio* flowMC workflow. Starting from short MD trajectories, a cycle of three steps is repeated. Step 1, the normalizing flow (NF) and the machine learning potential (MLP) are trained. Step 2, leveraging the NF and MLP models, MCMC chains are generated by restarting them from the last state of the previous cycle. Step 3, the training datasets of the ML models are augmented with the newly created MCMC samples and with the position-energy pairs for which DFT was evaluated. The function Γ_{MH} implements the Metropolis-Hastings update rule. The stopping criteria evaluates the convergence of the Markov Chain using the scale reduction factor $\hat{R}$ [Veh+21]. Reproduced from [Mol+24*].

Simultaneously training a machine learning potential (MLP) Following current trends in computational chemistry, the second adjustment consisted of simultaneously training a machine learning potential (MLP) to approximate the quantum force field. This MLP is regressed on the fly during the adaptive training and sampling procedure for each isomer. The samples generated by the MCMC sampler are used to augment the MLP training dataset. At each iteration, a coin is flipped to decide whether the quantum force field or the MLP is evaluated on the newly generated samples. The probability of evaluating the quantum force field decreases with the number of iterations, so that the MLP is used increasingly often as training progresses. This strategy reduces the number of expensive quantum force field evaluations while maintaining good sampling efficiency. The resulting workflow, named *ab initio* flowMC, is illustrated in fig. 7.

6.2.3 Summary of results

We first verify that the single-isomer adaptive training scheme is able to learn a good proposal distribution for each isomer and that the generated MCMC samples are in good agreement with the reference distribution obtained from long MD simulations (results not shown here, both with and without the MLP, see [Mol+24*]). In all cases, the NF proposals are accepted with a probability above 40% at the end of the adaptive training, and the resulting MCMC samples agree well with the reference distribution. This confirms that the NF-assisted sampler efficiently explores the state space of each isomer and generates accurate samples from their respective distributions.

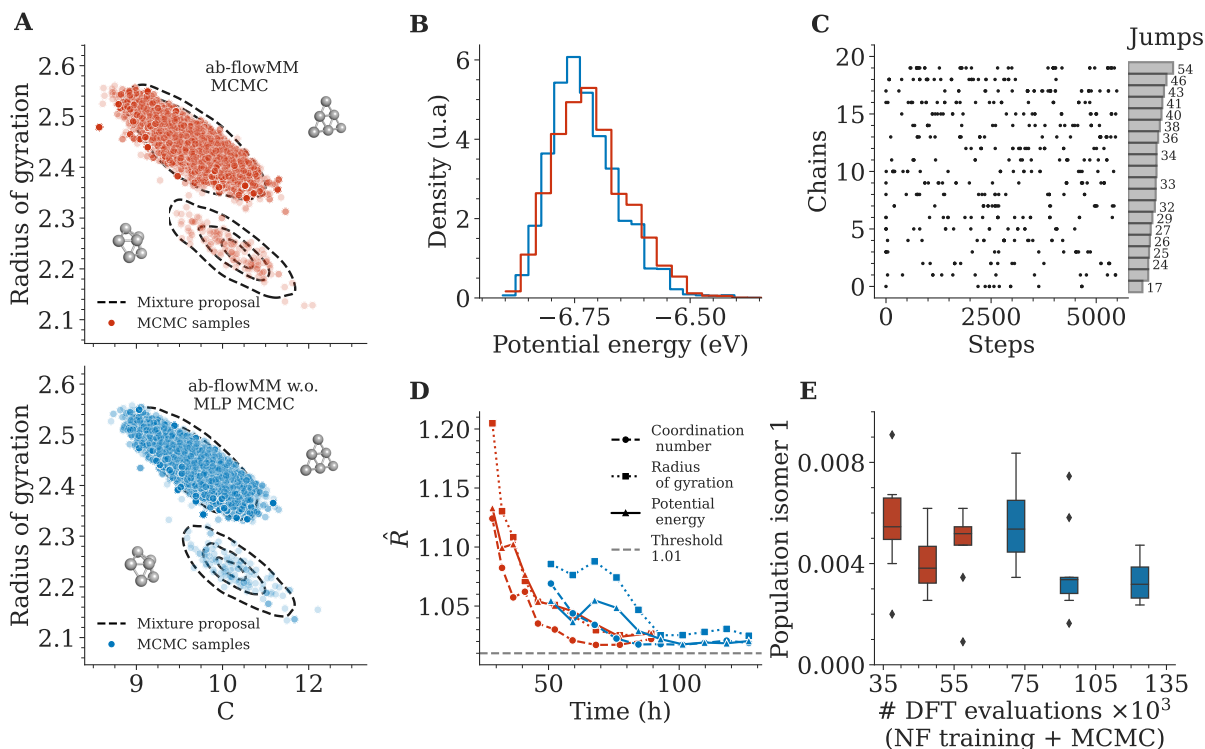


Figure 8: MCMC across isomers with the mixture model built using the NFs and MLPs from ab-flowMC. A) Mixture proposal (black dash-lines), and MCMC samples with MLP (top, red) and w.o. MLP using 100% DFT energy evaluations (bottom, blue). B) Potential energy distributions of the samples from the two settings. C) MCMC isomer state visited at each step. Visits to the bipyramidal isomer 1 are marked with a black dot, and to isomer 0 are left in white. Chains are sorted by total number of visits to isomer 1, which is reported in the histogram on the right. D) Scale reduction factor $\hat{R}$ as a function of wall clock time. E) Average population of isomer 1 as a function of the number of DFT evaluations used during flow training and the ab-flowM simulation using flows from different cycles of ab-flowMC. Box plot displays the variability across batches of 20 chains using the last 2750 steps. Reproduced from [Mol+24*].

We next turn to the joint sampling of the two isomers using the mixture of flows (see fig. 8). The resulting MCMC sampler efficiently mixes between the two isomers (panels A and C). Consistent estimates of the relative populations of the two isomers are obtained both with and without the MLP (panel E), and are further confirmed through BAR computations of the free energy difference between the two isomers (not shown here, see [Mol+24*] and the paragraph Free energy computations in section 3.1.1). We also verify that the MCMC chains are well converged, as assessed by the scale reduction factor $\hat{R}$ [Veh+21], which stabilizes around 1.02. Moreover, this convergence is achieved significantly faster in wall-clock time when using the MLP than without it (panel D).

*

This collaboration produced a first example of the use of NF-assisted samplers to obtain accurate estimates of the equilibrium distribution of a real molecular system. It required not only the methodological adjustments discussed above, but also engineering efforts to interface PyTorch with molecular simulation software. While the system considered remains modest in size, this work provides a valuable proof of concept.

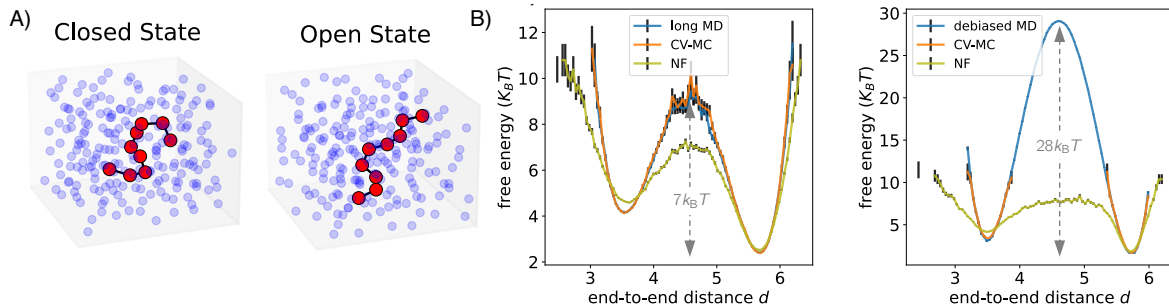


Figure 9: **A)** Representative configurations of the polymer surrounded by its Lennard-Jones solvent in its open and closed states. **B)** Free energy profile along the end-to-end distance between the first and the last bead of the polymer. The reference value (in blue) is obtained from long MD simulations. The value obtained with the CV-FlowMC sampler (in orange) is in good agreement with it, even if the marginal distribution of the CVs is not perfectly captured by the NF proposal (in green), especially around the barrier between the two states. Adapted from [TLG24*; Sch+25b*]

6.3 Toward sampling conformational ensembles of solvated biomolecules

Biomolecules are long polymers that can adopt a wide variety of conformations whose stability is strongly influenced by the surrounding solvent. Because their function is closely tied to their structure, characterizing the conformational ensemble of a biomolecule is a central problem in structural biology. As in many other molecular systems, transitions between distinct conformations are typically rare events that are difficult to observe using straightforward molecular dynamics (MD) simulations based on local updates. For this reason, a vast literature has been devoted to enhanced sampling strategies for biomolecular systems, including tempering methods, collective-variable-based techniques, and path sampling approaches. Here, we take a first step toward applying NF-assisted samplers to this class of problems.

As discussed in section 3.2, a key limitation of early NF-assisted samplers is their difficulty in scaling to high-dimensional systems. Consequently, directly applying these methods to a fully atomistic biomolecular system—including the surrounding solvent, which often involves thousands of atoms—remains unrealistic in the near term. Instead, we leverage a coarse-grained representation of the system and employ the CV-FlowMC algorithm introduced in section 3.4 (algorithm 9). In essence, the approach consists in training a NF to sample a set of coarse-grained variables, commonly referred to as collective variables (CVs), and using this learned model as a proposal distribution within an MCMC sampler targeting the full system. The remaining degrees of freedom are then sampled through a non-equilibrium dynamics that steers the system toward the NF-proposed update in CV space. We refer the reader to the previous section 3.4 for a detailed description of the sampling method, and focus here on the design choices specific to this application as well as on the preliminary results obtained.

6.3.1 Test system and CVs

We study a simplified molecular model consisting of a nine-bead polymer placed in a three-dimensional box together with 200 solvent particles. All particles, including both the polymer beads and the solvent molecules, interact through a Lennard-Jones (LJ) potential. Within the polymer chain, the distances between nearest and second-nearest neighboring beads—thereby indirectly constraining the bond angles—are weakly enforced through harmonic potentials. Additionally, the end-to-end distance of the polymer is subject to a bimodal potential, which induces two metastable conformations corresponding to an open and a closed configuration (see panel A of fig. 9).

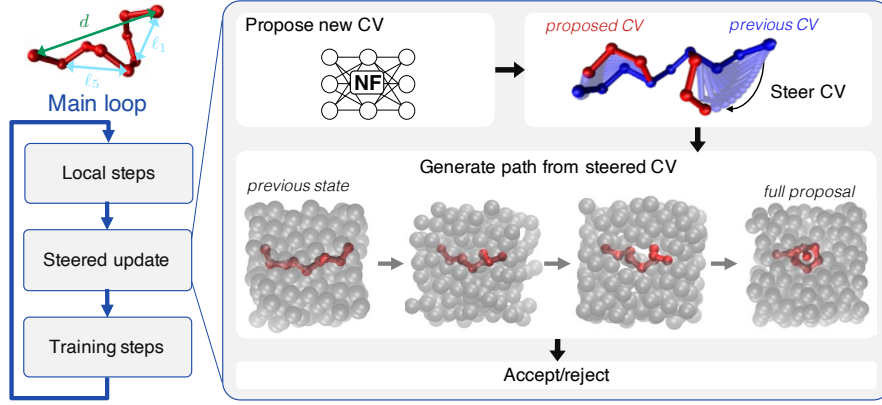


Figure 10: Schematic illustration of the algorithm. At each iteration, the system is first updated through local moves using a local sampler (MALA here). A non-local steered update is then proposed: the NF suggests new values for the polymer internal distances, the polymer is stretched or compressed via steered MD, and the solvent relaxes accordingly. The resulting configuration is accepted or rejected based on the work performed during steering and the ratio of NF proposal probabilities. After these updates, the NF is further trained using the generated samples. Reproduced from [TLG24*].

Let $\{q_i\}_{i=1,\dots,9}$ denote the coordinates of the polymer beads and $\{q_i\}_{i=10,\dots,N}$ those of the solvent particles. The total potential energy of the system can then be written as

$$\begin{aligned}
 V(q) = & \sum_{i=1}^8 V_1(|q_i - q_{i+1}|) + \sum_{i=1}^7 V_2(|q_i - q_{i+2}|) + \sum_{\substack{1 \leq i \leq 6 \\ i+3 \leq j \leq 9}} V_{\text{LJ}}(|q_i - q_j|) + V_e(|q_1 - q_9|) \\
 & + \sum_{\substack{1 \leq i \leq 9 \\ 10 \leq j \leq N}} V_{\text{LJ}}(|q_i - q_j|) + \sum_{10 \leq i < j \leq N} V_{\text{LJ}}(|q_i - q_j|),
 \end{aligned} \tag{75}$$

where V_{LJ} denotes the Lennard-Jones interaction, $V_\kappa(r) = \frac{k_\kappa}{2}(r - d_\kappa)^2$ ($\kappa \in 1, 2$) are harmonic potentials enforcing the polymer geometry, and the end-to-end interaction is described by the double-well potential $V_e(r) = \frac{k_e}{4}(\frac{r-d_e}{a})^4 - k_e(\frac{r-d_e}{a})^2$ which has two local minima located at $r = d_e \pm a\sqrt{2}$. Complete details on the system and the potential parameters can be found in [Sch+25b*].

In [TLG24*], hard constraints were imposed on the nearest and second-nearest neighbor distances within the polymer. Under these constraints, the configuration of the polymer is almost entirely determined by the six internal distances between fourth-neighboring beads, which were therefore used as collective variables (CVs) (the blue distances ℓ_k in the top-left panel of fig. 10).

In contrast, [Sch+25b*] retains the soft constraints on the internal polymer distances described above. In that work, experiments are performed using two different choices of CVs: the one-dimensional end-to-end distance of the polymer and the 27-dimensional vector formed by the full Cartesian coordinates of the polymer beads.

6.3.2 Design choices for the training and sampling pipeline

Non-equilibrium dynamics As mentioned previously, the proposal kernel in the CV space is chosen to be a NF. For simplicity, the CV schedule is fixed to a linear interpolation between the current CV value and the value proposed by the NF. The remaining design choice concerns how to sample the other degrees of freedom of the system.

The non-equilibrium CV-driven dynamics described in algorithm 8 can accommodate different Markov kernels for this purpose. In [TLG24*], overdamped Langevin dynamics with a Metropolis-Hastings correction (MALA) is used. In [Sch+25b*], both overdamped and underdamped Langevin dynamics are considered. In the latter case, an auxiliary momentum variable is introduced and resampled at each steered update.

In all cases, the hyperparameters of the non-equilibrium dynamics are tuned to maximize the acceptance probability (see eq. (35) in the detailed description of the algorithm in section 3.4) or, in this bimodal example, to minimize the average cost of accepting barrier-crossing proposals. In particular, the integration time step is chosen as large as possible while maintaining the numerical stability of the integrator. The number of integration steps—corresponding to the number of intermediate configurations along the CV schedule—is taken to be proportional to the distance between the current and proposed CV values, with the proportionality speed parameter optimized accordingly. Finally, the friction coefficient and, in the underdamped Langevin setting, the mass associated with the auxiliary momentum variable are also tuned.

Adaptive training The NF is trained using the adaptive procedure described in algorithm 10. The process begins with a pretraining stage on a dataset of configurations generated from short MD simulations that have not yet reached convergence, in particular with respect to the relative populations of the two metastable states. During sampling, the large steered updates proposed by the NF are interleaved with local updates performed using the MALA sampler. The resulting algorithm is illustrated in fig. 10. It consists of an outer loop corresponding to the adaptive training procedure and an inner loop implementing the non-equilibrium dynamics associated with each steered update.

6.3.3 Summary of results

Impact of the energy barrier We first focus on the results reported in [TLG24*], obtained using the 6-dimensional CV representation together with hard constraints on the polymer internal distances. In fig. 9 (panel B), we compare the free-energy profile along the end-to-end distance obtained with the CV-FlowMC sampler to a reference profile computed from long MD simulations, for two different values of the energy barrier imposed by the double-well potential V_e .

In both cases, the CV-FlowMC sampler successfully captures the relative weights of the two metastable states, even though the NF proposal does not perfectly match the target distribution of CVs, particularly in the barrier region separating the two states. This discrepancy is especially pronounced for the higher barrier, where the NF significantly underestimates the barrier height. Nevertheless, the resulting MCMC sampler remains able to efficiently explore the state space within the two relevant basins. This observation highlights that a perfect match between the NF proposal and the target distribution is not required to obtain accurate estimates of thermodynamic properties, provided that the generative model adequately represents the thermodynamically dominant regions of the state space.

At leading order, the NF effectively learns an approximation of the free-energy landscape along the six CVs (see [TLG24*] for projections not shown here). This result underscores the potential of NFs to learn coarse-grained representations of complex systems, in regimes where traditional CV-based methods struggle due to the difficulty of faithfully representing the free-energy surface with non-parametric models in such dimensions.

Finally, it is worth noting that the sampling efficiency of the CV-FlowMC sampler is not significantly affected by the height of the energy barrier acting purely on the CVs. In particular, the acceptance rate of the steered updates remains similar in the two cases, averaging around 20%, with a comparable number of steps in the non-equilibrium dynamics ($\sim 35,000$). This

behavior contrasts sharply with CV-guided methods based on biased local dynamics, where the algorithm must explicitly visit thermodynamically subleading barrier regions in order to accurately estimate thermodynamic properties. When the barrier is high, this requirement can lead to very long convergence times. The present results therefore illustrate the ability of CV-FlowMC to efficiently sample systems with large energy barriers in CV space by relying on non-local updates that effectively jump across these barriers, without requiring extensive sampling of the barrier region itself.

Optimization of non-equilibrium dynamics In [Sch+25b*], returning to the case of a $7k_B T$ barrier in V_e , we further compare the performance of the CV-FlowMC sampler for different choices of non-equilibrium dynamics, in particular with and without auxiliary momentum variables. We consistently observe that using underdamped Langevin dynamics in the deterministic limit (i.e., without friction) leads to a significantly lower average cost for accepting barrier-crossing proposals than both overdamped Langevin dynamics and underdamped Langevin dynamics with friction.

When using the one-dimensional end-to-end distance as the CV, the gain between overdamped dynamics and frictionless momentum-augmented deterministic dynamics exceeds two orders of magnitude. For the 27-dimensional CV, the gain could not be precisely quantified because no barrier-crossing proposal was accepted when using overdamped Langevin dynamics. Strikingly, in the deterministic limit the steered updates are accepted about 99% of the time. Although the origin of this advantage remains unclear, the same behavior was consistently observed in other simple physical systems studied in [Sch+25b*], suggesting an interesting direction for future investigation.

CV choice Finally, these experiments on the solvated polymer also highlight the importance of the choice of collective variables. In practice, the one-dimensional end-to-end distance CV did not require training a NF; instead, a Gaussian mixture model was fitted to construct the proposal distribution. While this provides a cheaper proposal mechanism, the observed average cost of mode switching was about one order of magnitude higher when using the 27-dimensional CV, for which a NF was required to learn an accurate proposal.

These results suggest an advantage in using higher-dimensional CV representations and therefore confirm the relevance of introducing NFs in CV-guided sampling. Confirming this observation in more complex biomolecular systems will be the focus of future work.

*

While this polymer system remains relatively simple, it illustrates how the NF-assisted sampling framework must be adapted and refined when applied to more complex systems. Many questions remain open regarding the application of these methods to realistic biomolecular systems, including the choice of collective variables, the design of the non-equilibrium dynamics, and the construction of the CV schedule. If even simple, programmatic design choices already lead to significant improvements in sampling efficiency, further gains may be expected from more sophisticated strategies. This therefore represents a promising direction for future research.

7 Perspectives

Where do we stand? Generative-model-assisted samplers have not yet led to scientific discoveries that would have been unattainable with existing methodologies. Nevertheless, they introduce a fundamentally different mechanism from traditional enhanced sampling techniques and therefore offer a complementary perspective that deserves careful exploration. In particular, these methods often propose configurations that effectively jump across low-probability regions rather than traversing them through a sequence of intermediate states. This qualitative difference in how state space is explored may open new opportunities for tackling sampling problems characterized by pronounced metastability.

As discussed throughout this manuscript, the field is evolving rapidly, with many recent proposals extending the earliest ideas. A particularly promising direction consists in combining generative models with established sampling strategies, such as tempering approaches or collective-variable-guided sampling methods. In addition, the noising–denoising paradigm introduced by diffusion models provides a novel way of constructing annealing paths that may circumvent the phase transitions that often hinder traditional tempering strategies. More broadly, the rapid progress in the design and training of generative models naturally fuels optimism about their future impact on scientific sampling problems.

A central question, however, is whether the computational cost associated with training a generative model is ultimately offset by the gains in sampling efficiency. The answer is necessarily problem-dependent and remains difficult to assess in general. Importantly, such an evaluation should also account for the design cost of a sampling strategy. While deep generative models involve numerous hyperparameters and architectural choices, designing an efficient traditional sampler is also often highly problem-specific and may require substantial expert knowledge. One may therefore hope that machine-learning-based samplers could eventually offer more automated and less system-dependent solutions. Whether this promise will materialize remains an open question and is closely related to the issue of transferability discussed below.

Finally, an important challenge lies in transferring the many methodological advances proposed in recent years—often developed and evaluated on toy benchmarks—to realistic scientific applications. Achieving this goal requires sustained efforts and close collaborations between machine-learning researchers and domain scientists. The applications reviewed in section 6 represent only initial steps in this direction, and considerable work remains to demonstrate the practical impact of these methods on challenging sampling problems arising in the physical sciences.

What we did not cover Several important directions have not been covered in detail in this manuscript. Let us briefly discuss some of them here.

We have focused primarily on sampling problems defined on continuous state spaces. A substantial body of work, however, addresses discrete systems using autoregressive models. While we briefly mentioned some ideas originating from this literature, a comprehensive discussion was beyond the scope of this manuscript. Historically, several pioneering works on generative-model-based sampling were in fact developed for discrete statistical physics problems [LW18; WWZ19; McN+20]. More recent studies have demonstrated promising advantages of autoregressive models for sampling and optimization in discrete systems [BRZ25].

The ability of generative models to characterize equilibrium states without explicitly resolving the transition regions between them is often presented as a strength. However, many scientific questions concern dynamical properties rather than equilibrium statistics alone. Several recent works have begun to address this challenge, particularly in the context of molecular simulations. Proposed approaches include learning transition-path ensembles with normalizing flows [Asg+24],

learning transition-state distributions to initialize transition path sampling [Fal+23], and learning effective transition operators to accelerate molecular dynamics simulations [Kle+23; SWO23; DSO25].

Finally, returning to the question of amortizing training costs, an important objective is to design models that do not need to be retrained for every new target distribution. Instead, one would ideally train a generative model once and then reuse it across a family of related systems. This idea, often referred to as *transferability*, has recently begun to receive attention, particularly in molecular modeling applications [Kle+23; KN24; DSO25; Lew+25]. Developing transferable generative samplers capable of generalizing across related systems remains a key open challenge for the field.

References

References to my work carry an extra *.

- [Abb+24] Ryan Abbott, Aleksandar Botev, Denis Boyda, Daniel C. Hackett, Gurtej Kanwar, Sébastien Racanière, Danilo J. Rezende, Fernando Romero-López, Phiala E. Shanahan, and Julian M. Urban. “Applications of Flow Models to the Generation of Correlated Lattice QCD Ensembles”. In: *Physical Review D* 109.9 (May 2024), p. 094514. DOI: [10.1103/PhysRevD.109.094514](https://doi.org/10.1103/PhysRevD.109.094514).
- [ABV23] Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. *Stochastic Interpolants: A Unifying Framework for Flows and Diffusions*. Mar. 2023. arXiv: [2303.08797](https://arxiv.org/abs/2303.08797) [cond-mat].
- [ADH10] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. “Particle Markov Chain Monte Carlo Methods”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 72.3 (June 2010), pp. 269–342. ISSN: 1369-7412. DOI: [10.1111/j.1467-9868.2009.00736.x](https://doi.org/10.1111/j.1467-9868.2009.00736.x).
- [Aga+17] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. “Importance Sampling: Intrinsic Dimension and Computational Cost”. In: *Statistical Science* 32.3 (2017), pp. 405–431. ISSN: 0883-4237. JSTOR: [26408299](https://www.jstor.org/stable/26408299).
- [Ago+25] Elisabeth Agoritsas, Giovanni Catania, Aurélien Decelle, and Beatriz Seoane. “Explaining the Effects of Non-Convergent Sampling in the Training of Energy-Based Models*”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2025.6 (July 2025), p. 064001. ISSN: 1742-5468. DOI: [10.1088/1742-5468/adde41](https://doi.org/10.1088/1742-5468/adde41).
- [Akh+24] Tara Akhound-Sadegh, Jarrid Rector-Brooks, Joey Bose, Sarthak Mittal, Pablo Lemos, Cheng-Hao Liu, Marcin Sendera, Siamak Ravanbakhsh, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin, and Alexander Tong. “Iterated Denoising Energy Matching for Sampling from Boltzmann Densities”. In: *Forty-First International Conference on Machine Learning*. June 2024.
- [AKS19] M. S. Albergo, G. Kanwar, and P. E. Shanahan. “Flow-Based Generative Models for Markov Chain Monte Carlo in Lattice Field Theory”. In: *Physical Review D* 100.3 (Aug. 2019), p. 034515. ISSN: 2470-0010, 2470-0029. DOI: [10.1103/PhysRevD.100.034515](https://doi.org/10.1103/PhysRevD.100.034515).
- [AMD21] Michael Arbel, Alex Matthews, and Arnaud Doucet. “Annealed Flow Transport Monte Carlo”. In: *Proceedings of the 38th International Conference on Machine Learning*. PMLR, July 2021, pp. 318–330.

- [AMS22] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. “Sampling from the Sherrington-Kirkpatrick Gibbs Measure via Algorithmic Stochastic Localization”. In: *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. Oct. 2022, pp. 323–334. DOI: [10.1109/FOCS54457.2022.00038](https://doi.org/10.1109/FOCS54457.2022.00038).
- [And82] Brian D.O. Anderson. “Reverse-Time Diffusion Equation Models”. In: *Stochastic Processes and their Applications* 12.3 (1982), pp. 313–326. ISSN: 0304-4149. DOI: [10.1016/0304-4149\(82\)90051-5](https://doi.org/10.1016/0304-4149(82)90051-5).
- [Asg+24] Solomon Asghar, Qing-Xiang Pei, Giorgio Volpe, and Ran Ni. “Efficient Rare Event Sampling with Unsupervised Normalizing Flows”. In: *Nature Machine Intelligence* 6.11 (Nov. 2024), pp. 1370–1381. ISSN: 2522-5839. DOI: [10.1038/s42256-024-00918-3](https://doi.org/10.1038/s42256-024-00918-3).
- [Ath02] M. Athènes. “Computation of a Chemical Potential Using a Residence Weight Algorithm”. In: *Physical Review E* 66.4 (Oct. 2002), p. 046705. DOI: [10.1103/PhysRevE.66.046705](https://doi.org/10.1103/PhysRevE.66.046705).
- [AV23] Michael Samuel Albergo and Eric Vanden-Eijnden. “Building Normalizing Flows with Stochastic Interpolants”. In: *The Eleventh International Conference on Learning Representations*. 2023.
- [AV24] Michael Samuel Albergo and Eric Vanden-Eijnden. “Learning to Sample Better”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2024.10 (Oct. 2024), p. 104014. ISSN: 1742-5468. DOI: [10.1088/1742-5468/ad363c](https://doi.org/10.1088/1742-5468/ad363c).
- [AV25] Michael Samuel Albergo and Eric Vanden-Eijnden. “NETS: A Non-equilibrium Transport Sampler”. In: *Forty-Second International Conference on Machine Learning*. June 2025.
- [AZG21] Michael Arbel, Liang Zhou, and Arthur Gretton. “Generalized Energy Based Models”. In: *International Conference on Learning Representations*. 2021. URL: <https://openreview.net/forum?id=OPtUPB9z6qK>.
- [Ben76] Charles H Bennett. “Efficient Estimation of Free Energy Differences from Monte Carlo Data”. In: *Journal of Computational Physics* 22.2 (1976), pp. 245–268. ISSN: 0021-9991. DOI: [10.1016/0021-9991\(76\)90078-4](https://doi.org/10.1016/0021-9991(76)90078-4).
- [Ber+19] Ludovic Berthier, Elijah Flenner, Christopher J Fullerton, Camille Scalliet, and Murari Singh. “Efficient Swap Algorithms for Molecular Dynamics Simulations of Equilibrium Supercooled Liquids”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.6 (June 2019), p. 064004. ISSN: 1742-5468. DOI: [10.1088/1742-5468/ab1910](https://doi.org/10.1088/1742-5468/ab1910).
- [Bér+23] Nicolas Béréux, Aurélien Decelle, Cyril Furtlehner, and Beatriz Seoane. “Learning a Restricted Boltzmann Machine Using Biased Monte Carlo Sampling”. In: *SciPost Physics* 14.3 (Mar. 2023), p. 032. ISSN: 2542-4653. DOI: [10.21468/SciPostPhys.14.3.032](https://doi.org/10.21468/SciPostPhys.14.3.032).
- [Bér+24] Nicolas Bereux, Aurélien Decelle, Cyril Furtlehner, Lorenzo Rosset, and Beatriz Seoane. “Fast Training and Sampling of Restricted Boltzmann Machines”. In: *The Thirteenth International Conference on Learning Representations*. Oct. 2024.
- [BKM17] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. “Variational Inference: A Review for Statisticians”. In: *Journal of the American Statistical Association* 112.518 (Apr. 2017), pp. 859–877. ISSN: 0162-1459, 1537-274X. DOI: [10.1080/01621459.2017.1285773](https://doi.org/10.1080/01621459.2017.1285773). arXiv: [1601.00670](https://arxiv.org/abs/1601.00670) [cs, stat].
- [BKW09] Etienne P. Bernard, Werner Krauth, and David B. Wilson. “Event-Chain Monte Carlo Algorithms for Hard-Sphere Systems”. In: *Physical Review E* 80.5 (Nov. 2009), p. 056704. DOI: [10.1103/PhysRevE.80.056704](https://doi.org/10.1103/PhysRevE.80.056704).

- [Ble+24] Denis Blessing, Xiaogang Jia, Johannes Esslinger, Francisco Vargas, and Gerhard Neumann. “Beyond ELBOs: A Large-Scale Evaluation of Variational Methods for Sampling”. In: *Proceedings of the 41st International Conference on Machine Learning*. PMLR, July 2024, pp. 4205–4229.
- [Bre91] Yann Brenier. “Polar Factorization and Monotone Rearrangement of Vector-Valued Functions”. In: *Communications on Pure and Applied Mathematics* 44.4 (1991), pp. 375–417. ISSN: 1097-0312. DOI: [10.1002/cpa.3160440402](https://doi.org/10.1002/cpa.3160440402).
- [Bro+22*] James Brofos, Marylou Gabri  , Marcus A. Brubaker, and Roy R. Lederman. “Adaptation of the Independent Metropolis-Hastings Sampler with Normalizing Flow Proposals”. In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. PMLR, May 2022, pp. 5949–5986. (Visited on 08/16/2023).
- [BRU23] Julius Berner, Lorenz Richter, and Karen Ullrich. “An Optimal Control Perspective on Diffusion-Based Generative Modeling”. In: *Transactions on Machine Learning Research* (Oct. 2023). ISSN: 2835-8856.
- [BRZ25] Luca Maria Del Bono, Federico Ricci-Tersenghi, and Francesco Zamponi. *Demonstrating Real Advantage of Machine-Learning-Enhanced Monte Carlo for Combinatorial Optimization*. Oct. 2025. DOI: [10.48550/arXiv.2510.19544](https://doi.org/10.48550/arXiv.2510.19544). arXiv: [2510.19544](https://arxiv.org/abs/2510.19544) [cond-mat].
- [Cas+22] Michele Caselle, Elia Cellini, Alessandro Nada, and Marco Panero. “Stochastic Normalizing Flows as Non-Equilibrium Transformations”. In: *Journal of High Energy Physics* 2022.7 (July 2022), p. 15. ISSN: 1029-8479. DOI: [10.1007/JHEP07\(2022\)015](https://doi.org/10.1007/JHEP07(2022)015).
- [CCN24] Michele Caselle, Elia Cellini, and Alessandro Nada. “Sampling the Lattice Nambu-Goto String Using Continuous Normalizing Flows”. In: *Journal of High Energy Physics* 2024.2 (Feb. 2024), p. 48. ISSN: 1029-8479. DOI: [10.1007/JHEP02\(2024\)048](https://doi.org/10.1007/JHEP02(2024)048).
- [CCN25] Michele Caselle, Elia Cellini, and Alessandro Nada. “Numerical Determination of the Width and Shape of the Effective String Using Stochastic Normalizing Flows”. In: *Journal of High Energy Physics* 2025.2 (Feb. 2025), p. 90. ISSN: 1029-8479. DOI: [10.1007/JHEP02\(2025\)090](https://doi.org/10.1007/JHEP02(2025)090).
- [CE22] Yuansi Chen and Ronen Eldan. “Localization schemes: A framework for proving mixing bounds for Markov chains”. In: *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2022, pp. 110–122.
- [Che+18] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. “Neural Ordinary Differential Equations”. In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018.
- [Cho+22] Kristy Choi, Chenlin Meng, Yang Song, and Stefano Ermon. “Density Ratio Estimation via Infinitesimal Classification”. In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. PMLR, May 2022, pp. 2552–2573.
- [Cia+23] Simone Ciarella, Jeanne Trinquier, Martin Weigt, and Francesco Zamponi. “Machine-Learning-Assisted Monte Carlo Fails at Sampling Computationally Hard Problems”. In: *Machine Learning: Science and Technology* 4.1 (Mar. 2023), p. 010501. ISSN: 2632-2153. DOI: [10.1088/2632-2153/acbe91](https://doi.org/10.1088/2632-2153/acbe91).
- [CN23] Alberto Cabezas and Christopher Nemeth. “Transport Elliptical Slice Sampling”. In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. PMLR, Apr. 2023, pp. 3664–3676.

- [Cor+20] Rob Cornish, Anthony Caterini, George Deligiannidis, and Arnaud Doucet. “Relaxing Bijectivity Constraints with Continuously Indexed Normalising Flows”. In: *Proceedings of the 37th International Conference on Machine Learning* PMLR 119 (2020).
- [Cor+23] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi S. Jaakkola. “DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking”. In: *The Eleventh International Conference on Learning Representations*. 2023.
- [Cor+24] Alessandro Coretti, Sebastian Falkner, Jan Weinreich, Christoph Dellago, and O. Anatole von Lilienfeld. “Boltzmann Generators and the New Frontier of Computational Sampling in Many-Body Systems”. In: *KIM Review* 2.3 (Apr. 2024), pp. 1–8. ISSN: 2998-2421. DOI: [10.25950/bfa99422](https://doi.org/10.25950/bfa99422).
- [Cor+25] Alessandro Coretti, Sebastian Falkner, Phillip L. Geissler, and Christoph Dellago. “Learning Mappings between Equilibrium States of Liquid Systems Using Normalizing Flows”. In: *The Journal of Chemical Physics* 162.18 (May 2025), p. 184102. ISSN: 0021-9606. DOI: [10.1063/5.0253034](https://doi.org/10.1063/5.0253034).
- [CP20] Nicolas Chopin and Omiros Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. 1st ed. 2020. Springer Series in Statistics. Cham: Springer International Publishing, 2020. ISBN: 978-3-030-47847-6 978-3-030-47845-2. DOI: [10.1007/978-3-030-47845-2](https://doi.org/10.1007/978-3-030-47845-2).
- [CR15] Yunjie Chen and Benoît Roux. “Generalized Metropolis Acceptance Criterion for Hybrid Non-Equilibrium Molecular Dynamics—Monte Carlo Simulations”. In: *The Journal of Chemical Physics* 142.2 (Jan. 2015), p. 024101. ISSN: 0021-9606, 1089-7690. DOI: [10.1063/1.4904889](https://doi.org/10.1063/1.4904889).
- [Cra+23] Kyle Cranmer, Gurtej Kanwar, Sébastien Racanière, Danilo J. Rezende, and Phiala E. Shanahan. “Advances in Machine-Learning-Based Sampling Motivated by Lattice Quantum Chromodynamics”. In: *Nature Reviews Physics* 5.9 (Sept. 2023), pp. 526–535. ISSN: 2522-5820. DOI: [10.1038/s42254-023-00616-w](https://doi.org/10.1038/s42254-023-00616-w).
- [Cro98] Gavin E. Crooks. “Nonequilibrium Measurements of Free Energy Differences for Microscopically Reversible Markovian Systems”. In: *Journal of Statistical Physics* 90.5 (Mar. 1998), pp. 1481–1487. ISSN: 1572-9613. DOI: [10.1023/A:1023208217925](https://doi.org/10.1023/A:1023208217925).
- [Daw+25*] Anna Dawid, Julian Arnold, Borja Requena, Alexander Gresch, Marcin Płodzień, Kaelan Donatella, Kim A. Nicoli, Paolo Stornati, Rouven Koch, Miriam Büttner, Robert Okuła, Gorka Muñoz-Gil, Rodrigo A. Vargas-Hernández, Alba Cervera-Lierta, Juan Carrasquilla, Vedran Dunjko, Marylou Gabrié, Patrick Huembeli, Evert van Nieuwenburg, Filippo Vicentini, Lei Wang, Sebastian J. Wetzel, Giuseppe Carleo, Eliška Greplová, Roman Krems, Florian Marquardt, Michał Tomza, Maciej Lewenstein, and Alexandre Dauphin. *Machine Learning in Quantum Sciences*. Cambridge: Cambridge University Press, 2025. ISBN: 978-1-009-50493-5. DOI: [10.1017/9781009504942](https://doi.org/10.1017/9781009504942).
- [DDJ06] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. “Sequential Monte Carlo Samplers”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68.3 (June 2006), pp. 411–436. ISSN: 1369-7412, 1467-9868. DOI: [10.1111/j.1467-9868.2006.00553.x](https://doi.org/10.1111/j.1467-9868.2006.00553.x).
- [De +21] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. “Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling”. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 17695–17709.

- [DRW21] Luigi Del Debbio, Joe Marsh Rossney, and Michael Wilson. “Efficient Modelling of Trivializing Maps for Lattice ϕ^4 Theory Using Normalizing Flows: A First Look at Scalability”. In: *Physical Review D* 104.9 (Nov. 2021), p. 094507. ISSN: 2470-0010, 2470-0029. DOI: [10.1103/PhysRevD.104.094507](https://doi.org/10.1103/PhysRevD.104.094507). arXiv: [2105.12481](https://arxiv.org/abs/2105.12481) [hep-lat].
- [DRZ25] Luca Maria Del Bono, Federico Ricci-Tersenghi, and Francesco Zamponi. “Performance of Machine-Learning-Assisted Monte Carlo in Sampling from Simple Statistical Physics Models”. In: *Physical Review E* 112.4 (Oct. 2025), p. 045307. DOI: [10.1103/s1rm-29zx](https://doi.org/10.1103/s1rm-29zx).
- [DSB17] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. “Density Estimation Using Real NVP”. In: *International Conference on Learning Representations*. 2017.
- [DSO25] Juan Viguera Diez, Mathias Schreiner, and Simon Olsson. *Transferable Generative Models Bridge Femtosecond to Nanosecond Time-Step Molecular Dynamics*. Oct. 2025. DOI: [10.48550/arXiv.2510.07589](https://doi.org/10.48550/arXiv.2510.07589). arXiv: [2510.07589](https://arxiv.org/abs/2510.07589) [physics].
- [Dur+19] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. “Neural Spline Flows”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.
- [DZ21] Xinqiang Ding and Bin Zhang. “DeepBAR: A Fast and Exact Method for Binding Free Energy Computation”. In: *The Journal of Physical Chemistry Letters* 12.10 (Mar. 2021), pp. 2509–2515. ISSN: 1948-7185, 1948-7185. DOI: [10.1021/acs.jpcllett.1c00189](https://doi.org/10.1021/acs.jpcllett.1c00189).
- [Eld13] Ronen Eldan. “Thin shell implies spectral gap up to polylog via a stochastic localization scheme”. In: *Geometric and Functional Analysis* 23.2 (2013), pp. 532–569.
- [Eld20] Ronen Eldan. “Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation”. In: *Probability Theory and Related Fields* 176.3-4 (2020), pp. 737–755.
- [Eld22] Ronen Eldan. “Analysis of high-dimensional distributions using pathwise methods”. In: *Proceedings of ICM, to appear* (2022).
- [Fal+23] Sebastian Falkner, Alessandro Coretti, Salvatore Romano, Phillip L. Geissler, and Christoph Dellago. “Conditioning Boltzmann Generators for Rare Event Sampling”. In: *Machine Learning: Science and Technology* 4.3 (Sept. 2023), p. 035050. ISSN: 2632-2153. DOI: [10.1088/2632-2153/acf55c](https://doi.org/10.1088/2632-2153/acf55c).
- [Fan+24] Mingzhou Fan, Ruida Zhou, Chao Tian, and Xiaoning Qian. “Path-Guided Particle-based Sampling”. In: *Proceedings of the 41st International Conference on Machine Learning*. PMLR, July 2024, pp. 12916–12934.
- [FLG26*] Luigi Fogliani, Bruno Loureiro, and Marylou Gabri  . *Annealing in Variational Inference Mitigates Mode Collapse: A Theoretical Study on Gaussian Mixtures*. Feb. 2026. DOI: [10.48550/arXiv.2602.12923](https://doi.org/10.48550/arXiv.2602.12923). arXiv: [2602.12923](https://arxiv.org/abs/2602.12923) [stat].
- [Fre04] Daan Frenkel. “Speed-up of Monte Carlo Simulations by Sampling of Rejected States”. In: *Proceedings of the National Academy of Sciences* 101.51 (Dec. 2004), pp. 17571–17575. DOI: [10.1073/pnas.0407950101](https://doi.org/10.1073/pnas.0407950101).
- [Gey91] Charles J Geyer. “Markov Chain Monte Carlo Maximum Likelihood”. In: *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*. Fairfax: Interface Foundation, 1991, p. 8.

- [Ghi+24] Davide Ghio, Yatin Dandi, Florent Krzakala, and Lenka Zdeborová. “Sampling with Flows, Diffusion, and Autoregressive Neural Networks from a Spin-Glass Perspective”. In: *Proceedings of the National Academy of Sciences* 121.27 (July 2024), e2311810121. DOI: [10.1073/pnas.2311810121](https://doi.org/10.1073/pnas.2311810121).
- [GMG23*] Louis Grenioux, Eric Moulines, and Marylou Gabrié. “Balanced Training of Energy-Based Models with Adaptive Flow Sampling”. In: *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*. July 2023. (Visited on 03/29/2024).
- [GNG25*] Louis Grenioux, Maxence Noble, and Marylou Gabrié. “Improving the evaluation of samplers on multi-modal targets”. en. In: *ICLR workshop on Frontiers in Probabilistic Inference: Learning meets Sampling*. Apr. 2025. URL: <https://openreview.net/forum?id=d91E9RhVFU> (visited on 05/27/2025).
- [GP01] Tomás S. Grigera and Giorgio Parisi. “Fast Monte Carlo Algorithm for Supercooled Soft Spheres”. In: *Physical Review E* 63.4 (Mar. 2001), p. 045102. DOI: [10.1103/PhysRevE.63.045102](https://doi.org/10.1103/PhysRevE.63.045102).
- [Gra+18] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. “FFJORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models”. In: *International Conference on Learning Representations*. Sept. 2018.
- [Gre+23*] Louis Grenioux, Alain Oliviero Durmus, Eric Moulines, and Marylou Gabrié. “On Sampling with Approximate Transport Maps”. In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, July 2023, pp. 11698–11733. (Visited on 03/29/2024).
- [Gre+24*] Louis Grenioux, Maxence Noble, Marylou Gabrié, and Alain Oliviero Durmus. “Stochastic Localization via Iterative Posterior Sampling”. en. In: *Proceedings of the 41st International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, July 2024, pp. 16337–16376. URL: <https://proceedings.mlr.press/v235/grenioux24a.html>.
- [Gre+25*] Louis Grenioux, Leonardo Galliano, Ludovic Berthier, Giulio Biroli, and Marylou Gabrié. *Riemannian Stochastic Interpolants for Amorphous Particle Systems*. Dec. 2025. DOI: [10.48550/arXiv.2512.16607](https://doi.org/10.48550/arXiv.2512.16607). arXiv: [2512.16607](https://arxiv.org/abs/2512.16607) [stat].
- [GRV21*] Marylou Gabrié, Grant M. Rotskoff, and Eric Vanden-Eijnden. “Efficient Bayesian Sampling Using Normalizing Flows to Assist Markov Chain Monte Carlo Methods”. In: *Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models (ICML Workshop)*. 2021. URL: <https://arxiv.org/abs/2107.08001>.
- [GRV22*] Marylou Gabrié, Grant M. Rotskoff, and Eric Vanden-Eijnden. “Adaptive Monte Carlo augmented with normalizing flows”. In: *Proceedings of the National Academy of Sciences* 119.10 (Mar. 2022). ISSN: 0027-8424. DOI: [10.1073/pnas.2109420119](https://doi.org/10.1073/pnas.2109420119). arXiv: [2105.12603](https://arxiv.org/abs/2105.12603). URL: <http://arxiv.org/abs/2105.12603%20https://pnas.org/doi/full/10.1073/pnas.2109420119>.
- [GS22] Mario Geiger and Tess Smidt. *E3nn: Euclidean Neural Networks*. July 2022. DOI: [10.48550/arXiv.2207.09453](https://doi.org/10.48550/arXiv.2207.09453). arXiv: [2207.09453](https://arxiv.org/abs/2207.09453) [cs].
- [Hac+21] Daniel C. Hackett, Chung-Chun Hsieh, Michael S. Albergo, Denis Boyda, Jiunn-Wei Chen, Kai-Feng Chen, Kyle Cranmer, Gurtej Kanwar, and Phiala E. Shanahan. *Flow-Based Sampling for Multimodal Distributions in Lattice Field Theory*. July 2021. arXiv: [2107.00734](https://arxiv.org/abs/2107.00734) [cond-mat, physics:hep-lat].

- [Hav+25] Aaron J. Havens, Benjamin Kurt Miller, Bing Yan, Carles Domingo-Enrich, Anuroop Sriram, Daniel S. Levine, Brandon M. Wood, Bin Hu, Brandon Amos, Brian Karrer, Xiang Fu, Guan-Horng Liu, and Ricky T. Q. Chen. “Adjoint Sampling: Highly Scalable Diffusion Samplers via Adjoint Matching”. In: *Forty-Second International Conference on Machine Learning*. June 2025.
- [He+25] Jiajun He, Yuanqi Du, Francisco Vargas, Dinghui Zhang, Shreyas Padhy, RuiKang OuYang, Carla P. Gomes, and José Miguel Hernández-Lobato. “No Trick, No Treat: Pursuits and Challenges Towards Simulation-free Training of Neural Samplers”. In: *Frontiers in Probabilistic Inference: Learning Meets Sampling - ICLR Workshop*. Apr. 2025.
- [Hén+22] Jérôme Hénin, Tony Lelièvre, Michael R. Shirts, Omar Valsson, and Lucie Delemotte. “Enhanced Sampling Methods for Molecular Dynamics Simulations [Article v1.0]”. In: *Living Journal of Computational Molecular Science* 4.1 (Dec. 2022), pp. 1583–1583. ISSN: 2575-6524. DOI: [10.33011/livecoms.4.1.1583](https://doi.org/10.33011/livecoms.4.1.1583).
- [HI03] K. Hukushima and Y. Iba. “Population Annealing and Its Application to a Spin Glass”. In: *AIP Conference Proceedings* 690.1 (Nov. 2003), pp. 200–206. ISSN: 0094-243X. DOI: [10.1063/1.1632130](https://doi.org/10.1063/1.1632130).
- [Hib+21] Mohamed Hibat-Allah, Estelle M. Inack, Roeland Wiersema, Roger G. Melko, and Juan Carrasquilla. “Variational Neural Annealing”. In: *Nature Machine Intelligence* 3.11 (Nov. 2021), pp. 952–961. ISSN: 2522-5839. DOI: [10.1038/s42256-021-00401-3](https://doi.org/10.1038/s42256-021-00401-3).
- [Hin02] Geoffrey E. Hinton. “Training products of experts by minimizing Contrastive divergence”. In: *Neural computation* 14 (2002), pp. 1771–1800.
- [HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising Diffusion Probabilistic Models”. In: *Neural Information Processing Systems 2020*. arXiv, Dec. 2020. arXiv: [2006.11239](https://arxiv.org/abs/2006.11239) [cs, stat].
- [HN96] Koji Hukushima and Koji Nemoto. “Exchange Monte Carlo Method and Application to Spin Glass Simulations”. In: *Journal of the Physical Society of Japan* 65.6 (June 1996), pp. 1604–1608. ISSN: 0031-9015. DOI: [10.1143/JPSJ.65.1604](https://doi.org/10.1143/JPSJ.65.1604).
- [Hof+19] Matthew Hoffman, Pavel Sountsov, Joshua V. Dillon, Ian Langmore, Dustin Tran, and Srinivas Vasudevan. *NeuTra-lizing Bad Geometry in Hamiltonian Monte Carlo Using Neural Transport*. Mar. 2019. arXiv: [1903.03704](https://arxiv.org/abs/1903.03704) [stat].
- [HT09] A. M. Hahn and H. Then. “Using bijective maps to improve free-energy estimates”. In: *Phys. Rev. E* 79 (1 Jan. 2009), p. 011113. DOI: [10.1103/PhysRevE.79.011113](https://doi.org/10.1103/PhysRevE.79.011113). URL: <https://link.aps.org/doi/10.1103/PhysRevE.79.011113>.
- [Hua+18] Chin-Wei Huang, Shawn Tan, Alexandre Lacoste, and Aaron C Courville. “Improving Explorability in Variational Inference with Annealed Variational Objectives”. In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018.
- [Hua+24] Xunpeng Huang, Hanze Dong, Yifan Hao, Yian Ma, and Tong Zhang. “Reverse Diffusion Monte Carlo”. In: *The Twelfth International Conference on Learning Representations*. 2024.
- [Hut89] M.F. Hutchinson. “A Stochastic Estimator of the Trace of the Influence Matrix for Laplacian Smoothing Splines”. In: *Communications in Statistics - Simulation and Computation* 18.3 (Jan. 1989), pp. 1059–1076. ISSN: 0361-0918. DOI: [10.1080/03610918908812806](https://doi.org/10.1080/03610918908812806).

- [Inv+22] Michele Invernizzi, Andreas Krämer, Cecilia Clementi, and Frank Noé. “Skipping the Replica Exchange Ladder with Normalizing Flows”. In: *The Journal of Physical Chemistry Letters* 13.50 (Dec. 2022), pp. 11643–11649. DOI: [10.1021/acs.jpclett.2c03327](https://doi.org/10.1021/acs.jpclett.2c03327).
- [Jar02] C. Jarzynski. “Targeted Free Energy Perturbation”. In: *Physical Review E* 65.4 (Apr. 2002), p. 046122. DOI: [10.1103/PhysRevE.65.046122](https://doi.org/10.1103/PhysRevE.65.046122).
- [Jar97] C. Jarzynski. “Nonequilibrium Equality for Free Energy Differences”. In: *Physical Review Letters* 78.14 (Apr. 1997), pp. 2690–2693. ISSN: 0031-9007, 1079-7114. DOI: [10.1103/PhysRevLett.78.2690](https://doi.org/10.1103/PhysRevLett.78.2690).
- [JBB24] Gerhard Jung, Giulio Biroli, and Ludovic Berthier. “Normalizing Flows as an Enhanced Sampling Method for Atomistic Supercooled Liquids”. In: *Machine Learning: Science and Technology* 5.3 (Aug. 2024), p. 035053. ISSN: 2632-2153. DOI: [10.1088/2632-2153/ad6ca0](https://doi.org/10.1088/2632-2153/ad6ca0).
- [Jor+99] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. “An Introduction to Variational Methods for Graphical Models”. In: *Machine Learning* 37.2 (Nov. 1999), pp. 183–233. ISSN: 1573-0565. DOI: [10.1023/A:1007665907178](https://doi.org/10.1023/A:1007665907178).
- [JS20] He Jia and Uros Seljak. “Normalizing Constant Estimation with Gaussianized Bridge Sampling”. In: *Proceedings of The 2nd Symposium on Advances in Approximate Bayesian Inference*. PMLR, Feb. 2020, pp. 1–14.
- [Jum+21] John Jumper et al. “Highly Accurate Protein Structure Prediction with AlphaFold”. In: *Nature* 596.7873 (Aug. 2021), pp. 583–589. ISSN: 0028-0836, 1476-4687. DOI: [10.1038/s41586-021-03819-2](https://doi.org/10.1038/s41586-021-03819-2).
- [Kar+22a] Minas Karamanis, Florian Beutler, John A. Peacock, David Nabergoj, and Uros Seljak. “Accelerating Astronomical and Cosmological Inference with Preconditioned Monte Carlo”. In: *Monthly Notices of the Royal Astronomical Society* 516.2 (Sept. 2022), pp. 1644–1653. ISSN: 0035-8711, 1365-2966. DOI: [10.1093/mnras/stac2272](https://doi.org/10.1093/mnras/stac2272). arXiv: [2207.05652](https://arxiv.org/abs/2207.05652) [astro-ph, physics:physics].
- [Kar+22b] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. “Elucidating the Design Space of Diffusion-Based Generative Models”. In: *Advances in Neural Information Processing Systems* 35 (Dec. 2022), pp. 26565–26577.
- [KIH25] Shoummo A Khandoker, Estelle M Inack, and Mohamed Hibat-Allah. “Lattice Protein Folding with Variational Annealing”. In: *Machine Learning: Science and Technology* 6.3 (Aug. 2025), p. 035023. ISSN: 2632-2153. DOI: [10.1088/2632-2153/adf376](https://doi.org/10.1088/2632-2153/adf376).
- [Kis65] Leslie Kish. “Sampling Organizations and Groups of Unequal Sizes On”. In: *American Sociological Review* 30.4 (1965), pp. 564–572. DOI: [10.2307/2091346](https://doi.org/10.2307/2091346). JSTOR: [2091346](https://www.jstor.org/stable/2091346).
- [KKN20] Jonas Köhler, Leon Klein, and Frank Noe. “Equivariant Flows: Exact Likelihood Generative Learning for Symmetric Densities”. In: *Proceedings of the 37th International Conference on Machine Learning*. PMLR, Nov. 2020, pp. 5361–5370.
- [Kle+23] Leon Klein, Andrew Foong, Tor Fjelde, Bruno Mlodozieniec, Marc Brockschmidt, Sebastian Nowozin, Frank Noe, and Ryota Tomioka. “Timewarp: Transferable Acceleration of Molecular Dynamics by Learning Time-Coarsened Dynamics”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 52863–52883.

- [KN24] Leon Klein and Frank Noe. “Transferable Boltzmann Generators”. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. Nov. 2024.
- [KPB21] Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. “Normalizing Flows: An Introduction and Review of Current Methods”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.11 (Nov. 2021), pp. 3964–3979. ISSN: 0162-8828, 2160-9292, 1939-3539. DOI: [10.1109/TPAMI.2020.2992934](https://doi.org/10.1109/TPAMI.2020.2992934).
- [LeM+25] Tâm LeMinh, Julyan Arbel, Thomas Möllenhoff, Mohammad Emtiyaz Khan, and Florence Forbes. *Natural Variational Annealing for Multimodal Optimization*. Dec. 2025. DOI: [10.48550/arXiv.2501.04667](https://doi.org/10.48550/arXiv.2501.04667). arXiv: [2501.04667 \[stat\]](https://arxiv.org/abs/2501.04667).
- [Lew+25] Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew Y. K. Foong, Victor García Satorras, Osama Abdin, Bastiaan S. Veeling, Iryna Zaporozhets, Yaoyi Chen, Soojung Yang, Adam E. Foster, Arne Schneuing, Jigyasa Nigam, Federico Barbero, Vincent Stimper, Andrew Campbell, Jason Yim, Marten Lienen, Yu Shi, Shuxin Zheng, Hannes Schulz, Usman Munir, Roberto Sordillo, Ryota Tomioka, Cecilia Clementi, and Frank Noé. “Scalable Emulation of Protein Equilibrium Ensembles with Generative Deep Learning”. In: *Science* 389.6761 (July 2025), eadv9817. DOI: [10.1126/science.adv9817](https://doi.org/10.1126/science.adv9817).
- [LGL23] Xingchao Liu, Chengyue Gong, and Qiang Liu. “Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow”. In: *The Eleventh International Conference on Learning Representations*. 2023.
- [Lip+23] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. “Flow Matching for Generative Modeling”. In: *ICLR*. Apr. 2023.
- [Liu04] Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. New York, NY: Springer, 2004. ISBN: 978-0-387-76369-9 978-0-387-76371-2. DOI: [10.1007/978-0-387-76371-2](https://doi.org/10.1007/978-0-387-76371-2).
- [LW18] Shuo-Hui Li and Lei Wang. “Neural Network Renormalization Group”. In: *Physical Review Letters* 121.26 (Dec. 2018), p. 260601. ISSN: 0031-9007, 1079-7114. DOI: [10.1103/PhysRevLett.121.260601](https://doi.org/10.1103/PhysRevLett.121.260601).
- [Mat+22] Alex Matthews, Michael Arbel, Danilo Jimenez Rezende, and Arnaud Doucet. “Continual Repeated Annealed Flow Transport Monte Carlo”. In: *Proceedings of the 39th International Conference on Machine Learning*. PMLR, June 2022, pp. 15196–15219.
- [McN+20] B. McNaughton, M. V. Milošević, A. Perali, and S. Pilati. “Boosting Monte Carlo Simulations of Spin Glasses Using Autoregressive Neural Networks”. In: *Physical Review E* 101.5 (May 2020), p. 053312. ISSN: 2470-0045, 2470-0053. DOI: [10.1103/PhysRevE.101.053312](https://doi.org/10.1103/PhysRevE.101.053312). arXiv: [2002.04292 \[cond-mat, physics:physics\]](https://arxiv.org/abs/2002.04292).
- [Met+53] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. “Equation of State Calculations by Fast Computing Machines”. In: *The Journal of Chemical Physics* 21.6 (June 1953), pp. 1087–1092. ISSN: 0021-9606. DOI: [10.1063/1.1699114](https://doi.org/10.1063/1.1699114).
- [MF23] Bálint Máté and François Fleuret. “Learning Interpolations between Boltzmann Densities”. In: *Transactions on Machine Learning Research* (Jan. 2023). ISSN: 2835-8856.
- [Mid+22] Laurence Illing Midgley, Vincent Stimper, Gregor N. C. Simm, Bernhard Schölkopf, and José Miguel Hernández-Lobato. “Flow Annealed Importance Sampling Bootstrap”. In: *The Eleventh International Conference on Learning Representations*. Sept. 2022.

- [Mid+23] Laurence Midgley, Vincent Stimper, Javier Antorán, Emile Mathieu, Bernhard Schölkopf, and José Miguel Hernández-Lobato. “SE(3) Equivariant Augmented Coupling Flows”. In: *Advances in Neural Information Processing Systems* 36 (Dec. 2023), pp. 79200–79225.
- [Min05] Tom Minka. *Divergence Measures and Message Passing*. Tech. rep. MSR-TR-2005-173. Technical report, Microsoft Research, 2005.
- [Mol+24*] Ana Molina-Taborda, Pilar Cossio, Olga Lopez-Acevedo, and Marylou Gabrié. “Active Learning of Boltzmann Samplers and Potential Energies with Quantum Mechanical Accuracy”. In: *Journal of Chemical Theory and Computation* (Oct. 2024). Publisher: American Chemical Society. ISSN: 1549-9618. DOI: [10.1021/acs.jctc.4c00506](https://doi.org/10.1021/acs.jctc.4c00506). URL: <https://doi.org/10.1021/acs.jctc.4c00506>.
- [Mon23] Andrea Montanari. *Sampling, Diffusions, and Stochastic Localization*. May 2023. arXiv: [2305.10690](https://arxiv.org/abs/2305.10690) [cs].
- [MU49] Nicholas Metropolis and S. Ulam. “The Monte Carlo Method”. In: *Journal of the American Statistical Association* 44.247 (Sept. 1949), pp. 335–341. ISSN: 0162-1459. DOI: [10.1080/01621459.1949.10483310](https://doi.org/10.1080/01621459.1949.10483310).
- [Mül+19] Thomas Müller, Brian McWilliams, Fabrice Rousselle, Markus Gross, and Jan Novák. “Neural Importance Sampling”. In: *ACM Transactions on Graphics* 38.5 (Oct. 2019), pp. 1–19. ISSN: 0730-0301, 1557-7368. DOI: [10.1145/3341156](https://doi.org/10.1145/3341156).
- [MW24] Andrea Montanari and Yuchen Wu. *Posterior Sampling in High Dimension via Diffusion Processes*. Aug. 2024. DOI: [10.48550/arXiv.2304.11449](https://doi.org/10.48550/arXiv.2304.11449). arXiv: [2304.11449](https://arxiv.org/abs/2304.11449) [math].
- [Nea01] Radford M. Neal. “Annealed Importance Sampling”. In: *Statistics and Computing* 11.2 (Apr. 2001), pp. 125–139. ISSN: 1573-1375. DOI: [10.1023/A:1008923215028](https://doi.org/10.1023/A:1008923215028).
- [Nea96] Radford M. Neal. “Sampling from Multimodal Distributions Using Tempered Transitions”. In: *Statistics and Computing* 6.4 (Dec. 1996), pp. 353–366. ISSN: 1573-1375. DOI: [10.1007/BF00143556](https://doi.org/10.1007/BF00143556).
- [Nic+20] Kim A. Nicoli, Shinichi Nakajima, Nils Strodthoff, Wojciech Samek, Klaus-Robert Müller, and Pan Kessel. “Asymptotically Unbiased Estimation of Physical Observables with Neural Samplers”. In: *Physical Review E* 101.2 (Feb. 2020), p. 023304. ISSN: 2470-0045, 2470-0053. DOI: [10.1103/PhysRevE.101.023304](https://doi.org/10.1103/PhysRevE.101.023304). arXiv: [1910.13496](https://arxiv.org/abs/1910.13496) [cond-mat, stat].
- [Nic+21] Kim A. Nicoli, Christopher J. Anders, Lena Funcke, Tobias Hartung, Karl Jansen, Pan Kessel, Shinichi Nakajima, and Paolo Stornati. “Estimation of Thermodynamic Observables in Lattice Field Theories with Deep Generative Models”. In: *Physical Review Letters* 126.3 (Jan. 2021), p. 032001. ISSN: 0031-9007, 1079-7114. DOI: [10.1103/PhysRevLett.126.032001](https://doi.org/10.1103/PhysRevLett.126.032001).
- [Nic+23] Kim A. Nicoli, Christopher J. Anders, Tobias Hartung, Karl Jansen, Pan Kessel, and Shinichi Nakajima. “Detecting and Mitigating Mode-Collapse for Flow-Based Sampling of Lattice Field Theories”. In: *Physical Review D* 108.11 (Dec. 2023), p. 114501. DOI: [10.1103/PhysRevD.108.114501](https://doi.org/10.1103/PhysRevD.108.114501).
- [Nij+19] Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. “Learning Non-Convergent Non-Persistent Short-Run MCMC Toward Energy-Based Model”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/2bc8ae25856bc2a6a1333d1331a3b7a6-Paper.pdf.

- [Nij+20] Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, and Ying Nian Wu. “On the Anatomy of MCMC-Based Maximum Likelihood Learning of Energy-Based Models”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.04 (Apr. 2020), pp. 5272–5280. DOI: [10.1609/aaai.v34i04.5973](https://ojs.aaai.org/index.php/AAAI/article/view/5973). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/5973>.
- [Nij+22] Erik Nijkamp, Ruiqi Gao, Pavel Sountsov, Srinivas Vasudevan, Bo Pang, Song-Chun Zhu, and Ying Nian Wu. “MCMC Should Mix: Learning Energy-Based Model with Neural Transport Latent Space MCMC”. In: *International Conference on Learning Representations*. 2022. URL: <https://openreview.net/forum?id=4C93Qvn-tz>.
- [Nil+11] Jerome P. Nilmeier, Gavin E. Crooks, David D. L. Minh, and John D. Chodera. “Nonequilibrium Candidate Monte Carlo Is an Efficient Tool for Equilibrium Simulation”. In: *Proceedings of the National Academy of Sciences* 108.45 (Nov. 2011), E1009–E1018. DOI: [10.1073/pnas.1106094108](https://doi.org/10.1073/pnas.1106094108).
- [NLB20] Christian Naesseth, Fredrik Lindsten, and David Blei. “Markovian Score Climbing: Variational Inference with $KL(P \setminus \text{vert} \setminus \text{vert } q)$ ”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 15499–15510.
- [Nob+25*] Maxence Noble, Louis Grenioux, Marylou Gabri  , and Alain Oliviero Durmus. “Learned Reference-based Diffusion Sampler for multi-modal distributions”. en. In: *The Thirteenth International Conference on Learning Representations*. 2025. URL: <https://openreview.net/forum?id=fmJUYgmMbL>.
- [No  +19] Frank No  , Simon Olsson, Jonas K  hler, and Hao Wu. “Boltzmann Generators: Sampling Equilibrium States of Many-Body Systems with Deep Learning”. In: *Science* 365.6457 (Sept. 2019), eaaw1147. ISSN: 0036-8075, 1095-9203. DOI: [10.1126/science.aaw1147](https://doi.org/10.1126/science.aaw1147).
- [Ole+24] Edgar Olehnovics, Yifei Michelle Liu, Nada Mehio, Ahmad Y. Sheikh, Michael R. Shirts, and Matteo Salvalaglio. “Assessing the Accuracy and Efficiency of Free Energy Differences Obtained from Reweighted Flow-Based Probabilistic Generative Models”. In: *Journal of Chemical Theory and Computation* 20.14 (July 2024), pp. 5913–5922. ISSN: 1549-9618. DOI: [10.1021/acs.jctc.4c00520](https://doi.org/10.1021/acs.jctc.4c00520).
- [Ols26] Simon Olsson. “Generative Molecular Dynamics”. In: *Current Opinion in Structural Biology* 96 (Feb. 2026), p. 103213. ISSN: 0959-440X. DOI: [10.1016/j.sbi.2025.103213](https://doi.org/10.1016/j.sbi.2025.103213).
- [Pap+21] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. “Normalizing Flows for Probabilistic Modeling and Inference”. In: *Journal of Machine Learning Research* 22.57 (2021), pp. 1–64. ISSN: 1533-7928.
- [PG25] Xin Peng and Ang Gao. “Flow Perturbation to Accelerate Boltzmann Sampling”. In: *Nature Communications* 16.1 (July 2025), p. 6604. ISSN: 2041-1723. DOI: [10.1038/s41467-025-62039-8](https://doi.org/10.1038/s41467-025-62039-8).
- [Phi+24] Angus Phillips, Hai-Dang Dau, Michael John Hutchinson, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. “Particle Denoising Diffusion Sampler”. In: *Proceedings of the 41st International Conference on Machine Learning*. PMLR, July 2024, pp. 40688–40724.
- [PH  20] Emilia Pompe, Chris Holmes, and Krzysztof   atuszy  ski. “A Framework for Adaptive MCMC Targeting Multimodal Distributions”. In: *The Annals of Statistics* 48.5 (Oct. 2020), pp. 2930–2952. ISSN: 0090-5364, 2168-8966. DOI: [10.1214/19-AOS1916](https://doi.org/10.1214/19-AOS1916).

- [PM18] Matthew D. Parno and Youssef M. Marzouk. “Transport Map Accelerated Markov Chain Monte Carlo”. In: *SIAM/ASA Journal on Uncertainty Quantification* 6.2 (Jan. 2018), pp. 645–682. ISSN: 2166-2525. DOI: [10.1137/17M1134640](https://doi.org/10.1137/17M1134640).
- [RB23] Lorenz Richter and Julius Berner. “Improved Sampling via Learned Diffusions”. In: *The Twelfth International Conference on Learning Representations*. Oct. 2023.
- [RM15] Danilo Rezende and Shakir Mohamed. “Variational Inference with Normalizing Flows”. In: *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, June 2015, pp. 1530–1538.
- [Sal+22] Antoine Salmona, Valentin De Bortoli, Julie Delon, and Agnes Desolneux. “Can Push-forward Generative Models Fit Multimodal Distributions?”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 10766–10779.
- [Sam+22*] Sergey Samsonov, Evgeny Lagutin, Marylou Gabri , Alain Durmus, Alexey Naumov, and Eric Moulines. “Local-Global MCMC kernels: the best of both worlds”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho. 2022. URL: <https://openreview.net/forum?id=zb-xfApk4ZK>.
- [Sat+21] Victor Satorras, Emiel Hoogeboom, Fabian Fuchs, Ingmar Posner, and Max Welling. “E(n) Equivariant Normalizing Flows”. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 4181–4192.
- [Sch+24] Maximilian Schebek, Michele Invernizzi, Frank No , and Jutta Rogal. “Efficient Mapping of Phase Diagrams with Conditional Boltzmann Generators”. In: *Machine Learning: Science and Technology* 5.4 (Nov. 2024), p. 045045. ISSN: 2632-2153. DOI: [10.1088/2632-2153/ad849d](https://doi.org/10.1088/2632-2153/ad849d).
- [Sch+25a*] C. Sch nle, Marylou Gabri , T. Leli vre, and G. Stoltz. “Sampling metastable systems using collective variables and Jarzynski–Crooks paths”. In: *Journal of Computational Physics* 527 (Apr. 2025), p. 113806. ISSN: 0021-9991. DOI: [10.1016/j.jcp.2025.113806](https://doi.org/10.1016/j.jcp.2025.113806). URL: <https://www.sciencedirect.com/science/article/pii/S0021999125000890> (visited on 02/21/2025).
- [Sch+25b*] Christoph Sch nle, Davide Carbone, Marylou Gabri , Tony Leli vre, and Gabriel Stoltz. *Efficient Monte-Carlo Sampling of Metastable Systems Using Non-Local Collective Variable Updates*. Dec. 2025. DOI: [10.48550/arXiv.2512.16812](https://doi.org/10.48550/arXiv.2512.16812). arXiv: [2512.16812](https://arxiv.org/abs/2512.16812) [cond-mat].
- [SF25] Henrik Schopmans and Pascal Friederich. “Temperature-Annealed Boltzmann Generators”. In: *Forty-Second International Conference on Machine Learning*. June 2025.
- [SG23*] Christoph Sch nle and Marylou Gabri . “Optimizing Markov Chain Monte Carlo Convergence with Normalizing Flows and Gibbs Sampling”. en. In: *NeurIPS 2023 AI for Science Workshop*. Oct. 2023. URL: <https://openreview.net/forum?id=iazHKyT4YK> (visited on 02/29/2024).
- [SGL25*] Roman Soletskyi, Marylou Gabri , and Bruno Loureiro. “A theoretical perspective on mode collapse in variational inference”. en. In: *Machine Learning: Science and Technology* 6.2 (June 2025). Publisher: IOP Publishing, p. 025056. ISSN: 2632-2153. DOI: [10.1088/2632-2153/adde2a](https://doi.org/10.1088/2632-2153/adde2a). URL: <https://dx.doi.org/10.1088/2632-2153/adde2a> (visited on 06/11/2025).

- [Shi+03] Michael R. Shirts, Eric Bair, Giles Hooker, and Vijay S. Pande. “Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods”. In: *Physical Review Letters* 91.14 (Oct. 2003), p. 140601. DOI: [10.1103/PhysRevLett.91.140601](https://doi.org/10.1103/PhysRevLett.91.140601).
- [SHW21] Víctor Garcia Satorras, Emiel Hoogetboom, and Max Welling. “E(n) Equivariant Graph Neural Networks”. In: *Proceedings of the 38th International Conference on Machine Learning*. PMLR, July 2021, pp. 9323–9332.
- [SK21] Yang Song and Diederik P. Kingma. *How to Train Your Energy-Based Models*. 2021. arXiv: [2101.03288](https://arxiv.org/abs/2101.03288) [[cs.LG](#)].
- [Skr+24] Marta Skreta, Lazar Atanackovic, Joey Bose, Alexander Tong, and Kirill Neklyudov. “The Superposition of Diffusion Models Using the Itô Density Estimator”. In: *The Thirteenth International Conference on Learning Representations*. Oct. 2024.
- [SNR25] Maximilian Schebek, Frank Noé, and Jutta Rogal. *Scalable Boltzmann Generators for Equilibrium Sampling of Large-Scale Materials*. Oct. 2025. DOI: [10.48550/arXiv.2509.25486](https://doi.org/10.48550/arXiv.2509.25486). arXiv: [2509.25486](https://arxiv.org/abs/2509.25486) [[cond-mat](#)].
- [Soh+15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. “Deep Unsupervised Learning Using Nonequilibrium Thermodynamics”. In: *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, June 2015, pp. 2256–2265.
- [Son+20] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. “Score-Based Generative Modeling through Stochastic Differential Equations”. In: *International Conference on Learning Representations*. Oct. 2020.
- [SW86] Robert H. Swendsen and Jian-Sheng Wang. “Replica Monte Carlo Simulation of Spin-Glasses”. In: *Physical Review Letters* 57.21 (Nov. 1986), pp. 2607–2609. DOI: [10.1103/PhysRevLett.57.2607](https://doi.org/10.1103/PhysRevLett.57.2607).
- [SW87] Robert H. Swendsen and Jian-Sheng Wang. “Nonuniversal Critical Dynamics in Monte Carlo Simulations”. In: *Physical Review Letters* 58.2 (Jan. 1987), pp. 86–88. DOI: [10.1103/PhysRevLett.58.86](https://doi.org/10.1103/PhysRevLett.58.86).
- [SWO23] Mathias Schreiner, Ole Winther, and Simon Olsson. “Implicit Transfer Operator Learning: Multiple Time-Resolution Models for Molecular Dynamics”. In: *Advances in Neural Information Processing Systems* 36 (Dec. 2023), pp. 36449–36462.
- [Tan+25] Charlie B. Tan, Joey Bose, Chen Lin, Leon Klein, Michael M. Bronstein, and Alexander Tong. “Scalable Equilibrium Sampling with Sequential Boltzmann Generators”. In: *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, Oct. 2025, pp. 58467–58498.
- [TH01] Hakon Tjelmeland and Bjorn Kare Hegstad. “Mode Jumping Proposals in MCMC”. In: *Scandinavian Journal of Statistics* 28.1 (2001), pp. 205–223. ISSN: 1467-9469. DOI: [10.1111/1467-9469.00232](https://doi.org/10.1111/1467-9469.00232).
- [Tie08] Tijmen Tieleman. “Training restricted Boltzmann machines using approximations to the likelihood gradient”. en. In: *Proceedings of the 25th international conference on Machine learning - ICML '08*. Helsinki, Finland: ACM Press, 2008, pp. 1064–1071. ISBN: 978-1-60558-205-4. DOI: [10.1145/1390156.1390290](https://doi.org/10.1145/1390156.1390290). URL: <http://portal.acm.org/citation.cfm?doid=1390156.1390290> (visited on 05/05/2023).

- [TLG24*] Samuel Tamagnone, Alessandro Laio, and Marylou Gabri . “Coarse-Grained Molecular Dynamics with Normalizing Flows”. In: *Journal of Chemical Theory and Computation* (Sept. 2024). Publisher: American Chemical Society. ISSN: 1549-9618. DOI: [10.1021/acs.jctc.4c00700](https://doi.org/10.1021/acs.jctc.4c00700). URL: <https://pubs.acs.org/doi/abs/10.1021/acs.jctc.4c00700>.
- [TPL24] Yifeng Tian, Nishant Panda, and Yen Ting Lin. “Liouville Flow Importance Sampler”. In: *Proceedings of the 41st International Conference on Machine Learning*. PMLR, July 2024, pp. 48186–48210.
- [TR19] Belinda Tzen and Maxim Raginsky. “Theoretical Guarantees for Sampling and Inference in Generative Models with Latent Diffusions”. In: *Proceedings of the Thirty-Second Conference on Learning Theory*. PMLR, June 2019, pp. 3084–3114.
- [TV10] Esteban G. Tabak and Eric Vanden-Eijnden. “Density Estimation by Dual Ascent of the Log-Likelihood”. In: *Communications in Mathematical Sciences* 8.1 (2010), pp. 217–233. ISSN: 15396746, 19450796. DOI: [10.4310/CMS.2010.v8.n1.a11](https://doi.org/10.4310/CMS.2010.v8.n1.a11).
- [Uri+16] Benigno Uri , Marc-Alexandre C  t , Karol Gregor, Iain Murray, and Hugo Larochelle. “Neural Autoregressive Distribution Estimation”. In: *Journal of Machine Learning Research* 17.205 (2016), pp. 1–37. ISSN: 1533-7928.
- [Var+24] Francisco Vargas, Shreyas Padhy, Denis Blessing, and Nikolas N sken. “Transport Meets Variational Inference: Controlled Monte Carlo Diffusions”. In: *The Twelfth International Conference on Learning Representations*. 2024.
- [Veh+21] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian B rkner. “Rank-normalization, Folding, and Localization: An Improved $\hat{R}$ for Assessing Convergence of MCMC”. en. In: *Bayesian Analysis* 16.2 (June 2021). arXiv:1903.08008 [stat]. ISSN: 1936-0975. DOI: [10.1214/20-BA1221](https://doi.org/10.1214/20-BA1221). URL: <http://arxiv.org/abs/1903.08008> (visited on 11/07/2022).
- [VGD23] Francisco Vargas, Will Sussman Grathwohl, and Arnaud Doucet. “Denoising Diffusion Samplers”. In: *The Eleventh International Conference on Learning Representations*. Sept. 2023.
- [VJ08] Suriyanarayanan Vaikuntanathan and Christopher Jarzynski. “Escorted Free Energy Simulations: Improving Convergence by Reducing Dissipation”. In: *Physical Review Letters* 100.19 (May 2008), p. 190601. ISSN: 0031-9007. DOI: [10.1103/PhysRevLett.100.190601](https://doi.org/10.1103/PhysRevLett.100.190601).
- [von51] John von Neumann. “Various Techniques Used in Connection With Random Digits”. In: *Journal of Research of the National Bureau of Standards* 3 (1951), pp. 36–38.
- [WCD25] Yihang Wang, Chris Chi, and Aaron R. Dinner. *Mitigating Mode Collapse in Normalizing Flows by Annealing with an Adaptive Schedule: Application to Parameter Estimation*. May 2025. DOI: [10.48550/arXiv.2505.03652](https://doi.org/10.48550/arXiv.2505.03652). arXiv: [2505.03652](https://arxiv.org/abs/2505.03652) [cs].
- [WIP23] Peter Wirnsberger, Borja Ibarz, and George Papamakarios. “Estimating Gibbs Free Energies via Isobaric-Isothermal Flows”. In: *Machine Learning: Science and Technology* 4.3 (Sept. 2023), p. 035039. ISSN: 2632-2153. DOI: [10.1088/2632-2153/acefa8](https://doi.org/10.1088/2632-2153/acefa8).
- [Wir+20] Peter Wirnsberger, Andrew J. Ballard, George Papamakarios, Stuart Abercrombie, S bastien Racani re, Alexander Pritzel, Danilo Jimenez Rezende, and Charles Blundell. “Targeted Free Energy Estimation via Learned Mappings”. In: *The Journal of Chemical Physics* 153.14 (Oct. 2020), p. 144112. ISSN: 0021-9606, 1089-7690. DOI: [10.1063/5.0018903](https://doi.org/10.1063/5.0018903).

- [Wir+22] Peter Wirnsberger, George Papamakarios, Borja Ibarz, Sébastien Racanière, Andrew J. Ballard, Alexander Pritzel, and Charles Blundell. “Normalizing Flows for Atomic Solids”. In: *Machine Learning: Science and Technology* 3.2 (June 2022), p. 025009. ISSN: 2632-2153. DOI: [10.1088/2632-2153/ac6b16](https://doi.org/10.1088/2632-2153/ac6b16). arXiv: [2111.08696](https://arxiv.org/abs/2111.08696) [[cond-mat](#), [physics:physics](#), [stat](#)].
- [WKN20] Hao Wu, Jonas Köhler, and Frank Noe. “Stochastic Normalizing Flows”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 5933–5944.
- [Wol89] Ulli Wolff. “Collective Monte Carlo Updating for Spin Systems”. In: *Physical Review Letters* 62.4 (Jan. 1989), pp. 361–364. DOI: [10.1103/PhysRevLett.62.361](https://doi.org/10.1103/PhysRevLett.62.361).
- [WRC21] Dian Wu, Riccardo Rossi, and Giuseppe Carleo. “Unbiased Monte Carlo Cluster Updates with Autoregressive Neural Networks”. In: *Physical Review Research* 3.4 (Nov. 2021), p. L042024. ISSN: 2643-1564. DOI: [10.1103/PhysRevResearch.3.L042024](https://doi.org/10.1103/PhysRevResearch.3.L042024). arXiv: [2105.05650](https://arxiv.org/abs/2105.05650) [[cond-mat](#), [stat](#)].
- [WWZ19] Dian Wu, Lei Wang, and Pan Zhang. “Solving Statistical Mechanics Using Variational Autoregressive Networks”. In: *Physical Review Letters* 122.8 (Feb. 2019), p. 080602. ISSN: 0031-9007, 1079-7114. DOI: [10.1103/PhysRevLett.122.080602](https://doi.org/10.1103/PhysRevLett.122.080602). arXiv: [1809.10606](https://arxiv.org/abs/1809.10606) [[cond-mat](#), [stat](#)].
- [Xia+21] Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat. “VAEBM: A Symbiosis between Variational Autoencoders and Energy-based Models”. In: *International Conference on Learning Representations*. 2021. URL: <https://openreview.net/forum?id=5m3SEcz0V8L>.
- [Xie+18] Jianwen Xie, Yang Lu, Ruiqi Gao, and Ying Nian Wu. “Cooperative Learning of Energy-Based Model and Latent Variable Model via MCMC Teaching”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 2018). DOI: [10.1609/aaai.v32i1.11834](https://doi.org/10.1609/aaai.v32i1.11834). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/11834>.
- [Xie+22] Jianwen Xie, Yaxuan Zhu, Jun Li, and Ping Li. “A Tale of Two Flows: Cooperative Learning of Langevin Flow and Normalizing Flow Toward Energy-Based Model”. In: *International Conference on Learning Representations*. 2022. URL: <https://openreview.net/forum?id=31d5RLCUuXC>.
- [XYA20] Zhisheng Xiao, Qing Yan, and Yali Amit. *Exponential Tilting of Generative Models: Improving Sample Quality by Training and Sampling from Latent Energy*. 2020. arXiv: [2006.08100](https://arxiv.org/abs/2006.08100) [[cs.LG](#)].
- [XZL21] Jianwen Xie, Zilong Zheng, and Ping Li. “Learning Energy-Based Model with Variational Auto-Encoder as Amortized Sampler”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.12 (May 2021), pp. 10441–10451. DOI: [10.1609/aaai.v35i12.17250](https://doi.org/10.1609/aaai.v35i12.17250). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/17250>.
- [ZC21] Qinsheng Zhang and Yongxin Chen. “Path Integral Sampler: A Stochastic Control Approach For Sampling”. In: *International Conference on Learning Representations*. Oct. 2021.
- [Zha+23] Dinghuai Zhang, Ricky T. Q. Chen, Cheng-Hao Liu, Aaron Courville, and Yoshua Bengio. “Diffusion Generative Flow Samplers: Improving Learning Signals through Partial Trajectory Optimization”. In: *The Twelfth International Conference on Learning Representations*. Oct. 2023.

- [Zha+25] Leo Zhang, Peter Potapchik, Jiajun He, Yuanqi Du, Arnaud Doucet, Francisco Vargas, Hai-Dang Dau, and Saifuddin Syed. *Accelerated Parallel Tempering via Neural Transports*. Sept. 2025. DOI: [10.48550/arXiv.2502.10328](https://doi.org/10.48550/arXiv.2502.10328). arXiv: [2502.10328](https://arxiv.org/abs/2502.10328) [stat].
- [ZMH25] Fengzhe Zhang, Laurence Illing Midgley, and José Miguel Hernández-Lobato. “Efficient and Unbiased Sampling from Boltzmann Distributions via Variance-Tuned Diffusion Models”. In: *Transactions on Machine Learning Research* (Aug. 2025). ISSN: 2835-8856.

RÉSUMÉ

L'échantillonnage de distributions de probabilité en grande dimension constitue une tâche centrale en calcul scientifique, avec des applications allant de l'inférence bayésienne à la physique statistique et à la simulation moléculaire. Malgré plusieurs décennies de développements méthodologiques, deux difficultés majeures persistent : le passage à l'échelle pour les problèmes en grande dimension et l'exploration efficace de distributions multimodales. Les approches classiques, telles que les méthodes de Monte Carlo par chaînes de Markov, les techniques de tempérage ou encore les méthodes d'échantillonnage amélioré fondées sur des variables collectives, ont connu des succès importants, mais elles rencontrent aussi des limitations intrinsèques. Ce manuscrit explore un nouveau paradigme récemment apparu à l'interface entre apprentissage automatique et physique statistique computationnelle : l'utilisation de modèles génératifs comme outils d'échantillonnage. Dans ce contexte, des modèles tels que les normalizing flows et les modèles de diffusion ne sont plus utilisés dans leur cadre traditionnel, entraînés sur de grands ensembles de données, mais comme des transports probabilistes flexibles permettant d'assister l'échantillonnage de distributions connues à une constante de normalisation près. Le manuscrit passe en revue les premiers développements de ce domaine en rapide évolution et présente plusieurs directions méthodologiques, notamment des algorithmes d'échantillonnage exact reposant sur des modèles génératifs et des stratégies pour entraîner ces modèles en l'absence de données. Il présente également plusieurs études applicatives menées avec des collaborateurs, illustrant à la fois le potentiel de ces méthodes et les nombreux défis qui demeurent.

ABSTRACT

Sampling high-dimensional probability distributions is a central task in scientific computing, with applications ranging from Bayesian inference to statistical physics and molecular simulation. Despite decades of methodological developments, two major challenges remain: scaling to high dimensions and efficiently exploring multimodal distributions characterized by metastable states. Classical approaches such as Markov chain Monte Carlo, tempering methods, or enhanced sampling based on collective variables have achieved major successes, but they also face intrinsic limitations in complex energy landscapes. This manuscript explores a new paradigm that has recently emerged at the interface of machine learning and computational statistical physics: the use of generative models as tools for sampling. In this context, models such as normalizing flows and diffusion models are not used in their traditional data-driven setting, but rather as flexible probabilistic transports that can assist the sampling of distributions known only up to a normalization constant. The manuscript reviews the early development of this rapidly evolving field and discusses several methodological directions, including exact samplers based on generative models and strategies to train such models in the absence of data. It also presents several application-driven studies conducted with collaborators, illustrating both the promise of these methods and the many open challenges that remain.