Sampling Boltzmann distributions via normalizing flow approximation of transport maps

Anonymous authors
Paper under double-blind review

Abstract

In a celebrated paper, Noé, Olsson, Köhler and Wu (Noé et al. (2019)) introduced an efficient method for sampling high-dimensional Boltzmann distributions arising in molecular dynamics via normalizing flow approximation of transport maps. Here, we place this approach on a firm mathematical foundation. We prove the existence of a normalizing flow between the reference measure and the true Boltzmann distribution up to an arbitrarily small error in the Wasserstein distance. This result covers general Boltzmann distributions from molecular dynamics, which have low regularity due to the presence of interatomic Coulomb and Lennard-Jones interactions. The proof is based on a rigorous construction of the Moser transport map for low-regularity endpoint densities and approximation theorems for neural networks in Sobolev spaces.

Numerical simulations for a simple model system and for the alanine dipeptide molecule confirm that the true and generated distributions are close in the Wasserstein distance. Moreover we observe that the RealNVP architecture does not just successfully capture the equilibrium Boltzmann distribution but also the metastable dynamics.

1 Introduction

Our goal in this paper is to put the recent method of Noé et al. (2019) for sampling from high-dimensional Boltzmann distributions on a firm mathematical footing.

Sampling from Boltzmann distributions is of great interest in computational physics, chemistry, and biochemistry, because the conformations of a molecule are described by the dominant regions of the associated Boltzmann distribution. Sampling from this distribution is challenging for two reasons: first, the configuration space on which the distribution lives is high-dimensional, of dimension $3N$, where N is the number of atoms in the molecule. Second, trajectory methods, i.e. using long-time samples of solutions of the underlying Langevin stochastic differential equation, while in principle correct, are expensive; this is because trajectories typically get trapped for a long time in metastable regions separated by energetic or entropic barriers, leading to very long simulation times for sufficient exploration of the configuration space.

Traditional approaches to this sampling problem include Markov state models Chodera et al. (2007), metadynamics Laio & Parrinello (2002), and umbrella sampling Torrie & Valleau (1977). While useful, these methods often either require careful hand-tuning of collective variables or suffer from high computational cost in high dimensions.

A novel approach based on ideas from machine learning was introduced in a pioneering paper by Noé et al. (2019), who proposed *Boltzmann generators*—invertible neural networks (normalizing flows) trained to map an easy-to-sample reference distribution (or prior) on the high-dimensional configuration space, for instance a Gaussian, to the target Boltzmann distribution. Note that, unlike in standard approaches in machine learning for approximating high-dimensional distributions, *the domain of the prior is not dimension-reduced.* Once the flow is trained, a single forward pass through the flow of statistically independent samples drawn from the prior yields statistically independent equilibrium samples. This bypasses long-time stochastic trajectories and enables efficient exploration of metastable states. This approach builds on the rapid development of

normalizing flows Dinh et al. (2017); Kingma & Dhariwal (2018); Papamakarios et al. (2021) and deep generative modeling for physical systems Hernandez et al. (2019); Wu et al. (2018).

Despite impressive empirical successes, a mathematical foundation of flow-based Boltzmann sampling is hitherto missing. Several open questions present themselves:

- **Dealing with singularities:** Physical Boltzmann distributions vanish on the collision set where two atoms collide, due to the singularities of the underlying Coulomb and Lennard-Jones potentials. Such zero regions, in turn, prevent the existence of smooth transport maps (for a proof of this interesting phenomenon see section 3). Thus it is not even clear whether any transport map exists between the Gaussian reference and the Boltzmann distribution which belongs to the regularity classes covered by expressivity results for neural networks.

- **Indeterminacy:** Even if singularities are absent or have been removed by some controlled approximation, the transport condition is highly degenerate. Which of the infinite-dimensionally many transport maps should be targeted for theoretical approximation results (and practical training)?

- **Approximability by normalizing flows:** Because of the need to efficiently evaluate the Jacobian of the transport map, a suitable class of neural networks is given by normalizing flow architectures such as RealNVP Dinh et al. (2017). Do these architectures possess sufficient expressive power to approximate the required transport maps between low-regularity densities well enough to produce correct samples?

The first problem is overcome by a simple regularization of the underlying molecular potential in high-energy regions, which entails a controlled L^1 error on the Boltzmann distribution. See section 3.

Regarding the second problem, a natural idea would be to use the Brenier map supplied by optimal transport theory Brenier (1991); Villani (2003); Friesecke (2024). For image generation, methods based on the Brenier maps from Gaussian noise to subsets of training images have proven useful Lipman et al. (2023); Liu et al. (2023); Chemseddine et al. (2025); Tong et al. (2023); Pooladian et al. (2023). Neural network based solvers for more general optimal transport problems have been introduced in Korotin et al. (2021); Uscidda & Cuturi (2023), based on input convex eural networks (ICNNs) respectively multi-layer perceptrons (MLPs). But the Brenier map, and the underlying flow, are not easy to compute between continuous distributions in high dimensions, and known regularity results for the map Caffarelli (1991); De Philippis & Figalli (2011) require strong assumptions on the densities. We therefore focus on a related but easier-to-compute and more regular transport map between continuous distributions, namely the Dacorogna-Moser map, and its underlying flow. We generalize its classical construction Moser (1965); Dacorogna & Moser (1990) from smooth densities to low-regularity densities (see Theorem 2 a)). We remark that for a uniform reference density, the governing problem correponds to a linearization of the optimal transport problem Friesecke (2024). The practical question of how to efficiently bias the learning of the flow towards the Dacorogna-Moser flow lies beyond the scope of this paper.

Finally, the third problem, rigorous approximability, is addressed in Theorem 2 c). We prove that the Moser map between continuously differentiable, everywhere positive densities can be approximated by normalizing flow architechtures such as RealNVP in the Sobolev space $W^{1,p}$.

Combining these results yields existence of a normalizing flow between the reference measure and the true Boltzmann distribution of a molecular system up to an arbitrarily small error in the Wasserstein distance (see Theorem 4), thus providing a theoretical guarantee that the Boltzmann generator method gives correct samples.

Numerical simulations for a simple synthetic system and for the alanine dipeptide molecule (see section 6.1 and 6.2) confirm that the learned flows yield correct samples reproducing the equilibrium densities.

2 Molecular dynamics at thermal equilibrium

In equilibrium molecular dynamics, the state of a many-body system can be described by the Boltzmann probability distribution

$$d\rho(x) = \frac{1}{Z}\exp\Big(-\frac{U(x)}{\kappa_B T}\Big)\mathrm{d}x, \tag{1}$$

where $x = (x_1, ..., x_N)$ with $x_i \in \mathbb{R}^3$ denoting the position of the i-th atom, κ_B is a physical constant, T is the temperature, and U is the potential energy function of the system. One assumes that the position vector x of all atoms is confined to some bounded domain $D \subset \mathbb{R}^{3N}$, and Z is a normalization constant so that $\int_D \rho = 1$.

The potential energy function is typically taken to be of the form

$$\begin{aligned} U(x) = &\sum_{\ell_{ij} \text{ bond lengths}} b_{ij}(\ell_{ij} - \overline{\ell_{ij}})^2 + \sum_{\alpha_{ijk} \text{ bond angles}} a_{ijk}(\alpha_{ijk} - \overline{\alpha_{ijk}})^2 \\ &+ \sum_{\phi_{ijkl} \text{ torsion angles}} \sum_{n=1}^{n_{max}} \kappa_{ijkl,n}\left[1 + \cos(n\phi_{ijkl} - \overline{\phi_{ijkl,n}})\right] \\ &+ \sum_{i,j} \frac{q_i q_j}{|x_i - x_j|} + \sum_{i,j}\left(\frac{A_{ij}}{|x_i - x_j|^{12}} - \frac{B_{ij}}{|x_i - x_j|^6}\right), \end{aligned} \tag{2}$$

where the first three terms are summed, respectively, over the bond lengths, bond angles, and torsion angles in the bond graph of the molecular system and the last two terms are summed over all atoms in the system. In the above q_i, A_{ij}, B_{ij}, a_{ijk}, b_{ij}, $\kappa_{\phi_{ijkl,n}}$, $\overline{\ell_{ij}}$, $\overline{\alpha_{ijk}}$, and $\overline{\phi_{ijkl,n}}$ are constants specific to the molecular system. The constants b_{ij} and a_{ijk} are positive, $\kappa_{ijkl,n}$, A_{ij} and B_{ij} are nonnegative, and A_{ij} must be positive when $q_i q_j < 0$ or $B_{ij} > 0$, and $x_i \in \mathbb{R}^3$ is the position of the i-th atom in the system. The last two terms in the energy functional represent electrostatic and Lennard-Jones interactions.

The energy equation 2 has various important qualitative properties:

$$\begin{aligned} &U : \mathbb{R}^{3N} \to \mathbb{R} \cup \{+\infty\} \text{ is } C^1 \text{ on } \{x : x_i - x_j \neq 0 \ \forall i \neq j\}, \\ &U(x) \geq U_0 \text{ for some constant } U_0, \\ &U(x) \to \infty \text{ as } \min_{i \neq j} |x_i - x_j| \to 0. \end{aligned} \tag{3}$$

The last property in equation 3 means that the potential energy tends to infinity when two atoms approach each other. As a consequence, the Boltzmann distribution equation 1 vanishes on the collision set $\{x_i = x_j \text{ for some } i \neq j\}$.

3 Transport maps and their lack of regularity

The Boltzmann generator approach for sampling from Boltzmann distributions starts from a smooth, positive, easy-to-sample reference distribution, or prior, and seeks to compute an invertible transport map between the reference distribution and the Boltzmann distribution equation 1. We begin by recalling basic mathematical definitions and facts. We then show that in our present context, transport maps must necessarily be quite irregular.

Assume that we are given two probability measures μ and ν on $\mathbb{R}^d$. (In the molecular dynamics case, $d = 3N$.) A *transport map* between μ and ν is a measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$ satisfying

$$T_\sharp \mu = \nu. \tag{4}$$

The pushforward $T_\sharp \mu$ is defined such that for any measurable set $A \subset \mathbb{R}^d$, $(T_\sharp \mu)(A) = \mu(T^{-1}(A))$. Equivalently, for any measurable function f, we have

$$\int_{\mathbb{R}^d} f(y)\, d(T_\sharp \mu)(y) = \int_{\mathbb{R}^d} f(T(x))\, d\mu(x).$$

This means that if a random variable $X \sim \mu$ then $T(X) \sim \nu$.

Assume now that μ and ν possess densities ρ_0 and ρ_1. In this case we write condition equation 4 as

$$T_\sharp \rho_0 = \rho_1. \tag{5}$$

If the map T is invertible and its inverse is differentiable, by the change-of-variables formula eq. equation 4 is equivalent to

$$\rho_1(z) = \rho_0(T^{-1}(z)) \left|\det J_{T^{-1}}(z)\right|. \tag{6}$$

where $J_{T^{-1}}$ is the Jacobian matrix of T.

Invertible transport maps exist (in the low-regularity class of measurable mappings); a basic example is the Brenier map Brenier (1991); Villani (2003); Friesecke (2024) from optimal transport theory which exists provided the densities have finite second moments or compact support. But this map is not easy to compute in high dimension. Moreover known regularity results Caffarelli (1991); De Philippis & Figalli (2011) require the target density to be bounded away from zero, which is violated by Boltzmann distributions for molecules, due to equation 3.

Another, simpler and smoother, example of a transport map is the Dacorogna-Moser map, described in the next section; but its construction requires, among other assumptions, strict positivity of the target density, and therefore fails for Boltzmann distributions for molecules, which vanish on the collision set, due to equation 3.

In fact we have:

Proposition 1 (Nonexistence of Lipschitz transport maps)**.** *Let $\Omega \subset \mathbb{R}^{3N}$ be any open bounded set, let ρ_0 be any reference probability density on Ω which is continuous and strictly positive, and let ρ_1 be a Boltzmann distribution equation 1 on an open bounded set $D \subset \mathbb{R}^{3N}$ with U satisfying equation 3. Then there does not exist any Lipschitz map T which transports ρ_0 to the Boltzmann distribution ρ_1.*

The singular nature of (exact) transport maps revealed by this result is a challenge for theory, and also for efficient learning of such maps, on which more later.

Proof. Let $x^0 = (x_1^0, x_2^0, ..., x_N^0) \in \mathbb{R}^{3N}$ be any point with $x_1^0 = x_2^0$. For $r > 0$ and $\varepsilon > 0$, define the following neighborhood of this point:

$$B'_\varepsilon \times B'_r \times (B_r)^{N-2} := \{x \in \mathbb{R}^{3N} \,:\, |\tfrac{x_1 - x_2}{\sqrt{2}}| \le \varepsilon,\, |\tfrac{(x_1 - x_1^0) + (x_2 - x_2^0)}{\sqrt{2}}| \le r,\, |x_i - x_i^0| \le r \,\forall i = 3, ..., N\}.$$

Now assume for contradiction that there exists an invertible Lipschitz map $T : \Omega^N \to \mathbb{R}^{3N}$ such that $T_\sharp \rho_0 = \rho_1$. Introduce the quantities

$$f_{min} := \min_{T^{-1}(B'_r \times B'_r \times (B_r)^{N-2})} \rho_0, \qquad \eta(\varepsilon) := \max_{B'_\varepsilon \times B'_r \times (B_r)^{N-2}} \rho_1.$$

Note that the minimum is attained because the set over which it is taken is closed (since T is continuous) and bounded (since Ω is). The key point is that Lipschitz maps can inflate the Lebesgue measure of any measurable set $A \subset \mathbb{R}^{3N}$ only by a factor, $\mathcal{L}(T(A)) \le (\operatorname{Lip} T)^{3N} \mathcal{L}(A)$. We exploit this as follows:

$$\begin{aligned}
f_{min} \cdot \mathcal{L}\big(B'_\varepsilon \times B'_r \times (B_r)^{N-2}\big) &= f_{min} \cdot \mathcal{L}\big(T(T^{-1}(B'_\varepsilon \times B'_r \times (B_r)^{N-2}))\big) \\
&\le f_{min} \cdot (Lip\, T)^{3N} \mathcal{L}\big(T^{-1}(B'_\varepsilon \times B'_r \times (B_r)^{N-2})\big) \\
&\le (Lip\, T)^{3N} \mu\big(T^{-1}(B'_\varepsilon \times B'_r \times (B_r)^{N-2})\big) \\
&= (Lip\, T)^{3N} \nu\big(B'_\varepsilon \times B'_r \times (B_r)^{N-2}\big) \\
&= (Lip\, T)^{3N} \nu\big(B'_\varepsilon \times B'_r \times (B_r)^{N-2}\big) \\
&\le (Lip\, T)^{3N} \mathcal{L}\big(B'_\varepsilon \times B'_r \times (B_r)^{N-2}\big) \cdot \eta(\varepsilon).
\end{aligned}$$

But f_{min} is positive whereas $\eta(\varepsilon)$ tends to zero as $\varepsilon \to 0$, a contradiction. □

4 Normalizing flow approximation of transport maps

Throughout this section, we work in a general dimension d, which in the molecular dynamics setting corresponds to $d = 3N$.

Further, we assume that the map has the form of an ℓ-fold composition

$$T = G^{(\ell)} \circ \cdots \circ G^{(1)}. \tag{7}$$

This composition structure emerges naturally for (i) the Brenier transport map which can be constructed via the Benamou–Brenier dynamic formulation of optimal transport, (ii) the Moser transport map Moser (1965); Dacorogna & Moser (1990) which is based on the flow of a more explicit vector field; see section 4.1.

In practice, the underlying flow and its step maps $G^{(k)}$ in equation 7 are not available explicitly.

The normalizing flow approach approximates each $G^{(k)}$ by a trainable invertible neural network block $F^{(k)}_{\theta_k}$,

$$G^{(k)} \approx F^{(k)}_{\theta_k},$$

(which itself consists of multiple coupling layers and alternating input partitions). The overall approximated transport map is then given by

$$F_\theta = F^{(\ell)}_{\theta_\ell} \circ F^{(\ell-1)}_{\theta_{\ell-1}} \circ \cdots \circ F^{(1)}_{\theta_1},$$

forming a neural-network discretization of the continuous transport. The resulting density is

$$\rho_\theta = (F_\theta)_\sharp \rho_0,$$

which approximates the target density ρ_1. Each $F^{(k)}_{\theta_k}$ is designed to be bijective with analytically computable inverse and Jacobian determinant.

The RealNVP architecture Dinh et al. (2017) provides a widely used framework for constructing such invertible transformations. For a RealNVP flow with ℓ coupling layers, the transformation reads

$$z = F_\theta(x) = F^{(\ell)} \circ F^{(\ell-1)} \circ \cdots \circ F^{(1)}(x).$$

For each layer $k = 1, \ldots, \ell$, the forward transformation is

$$\begin{cases} z^{(k)}_a = x^{(k-1)}_a, \\ z^{(k)}_b = x^{(k-1)}_b \odot \exp\big(s^{(k)}_{\theta_k}(x^{(k-1)}_a)\big) + t^{(k)}_{\theta_k}(x^{(k-1)}_a), \end{cases}$$

with $x^{(0)} = x$ and $z = z^{(\ell)}$. The inverse mapping is explicit:

$$\begin{cases} x^{(k-1)}_a = z^{(k)}_a, \\ x^{(k-1)}_b = \big(z^{(k)}_b - t^{(k)}_{\theta_k}(z^{(k)}_a)\big) \odot \exp\big(-s^{(k)}_{\theta_k}(z^{(k)}_a)\big). \end{cases}$$

The Jacobian matrix of each layer is block-triangular:

$$J_{F^{(k)}}(x^{(k-1)}) = \begin{pmatrix} I_{d_a} & 0 \\ \dfrac{\partial z^{(k)}_b}{\partial x^{(k-1)}_a} & \operatorname{diag}\big(\exp(s^{(k)}_{\theta_k}(x^{(k-1)}_a))\big) \end{pmatrix},$$

with determinant

$$\det J_{F^{(k)}}(x^{(k-1)}) = \exp\left(\sum_{j=1}^{d_b} s^{(k)}_{\theta_k,j}(x^{(k-1)}_a)\right).$$

For the complete flow, the Jacobian determinant factorizes as

$$\det J_{F_\theta}(x) = \prod_{k=1}^{\ell} \det J_{F^{(k)}}(x^{(k-1)}).$$

This enables efficient density evaluation via the change-of-variables formula

$$\log \rho_\theta(z) = \log \rho_0(F_\theta^{-1}(z)) - \log \left| \det J_{F_\theta}(F_\theta^{-1}(z)) \right|.$$

4.1 Composition structure of the Brenier and Moser maps

We now describe how the composition structure equation 7 arises naturally for standard transport maps.

We begin with the Brenier map. Its basic static construction is as the optimizer $T : \mathbb{R}^d \to \mathbb{R}^d$ of $\int |x - T(x)|^2 \rho_0(x)\, dx$ subject to the constraint $T_\sharp \rho_0 = \rho_1$. Alternatively, it is the time-1 flow map of the time-dependent vector field $(v_t)_{t\in[0,1]}$, $v_t : \mathbb{R}^d \to \mathbb{R}^d$ which minimizes the action

$$\int_0^1 |v_t(x)|^2 \rho_t(x)\, dx\, dt$$

over velocity fields and density fields $(\rho_t)_{t\in[0,1]}$ subject to the continuity equation

$$\frac{\partial \rho_t}{\partial t} + \nabla \cdot (\rho_t v_t) = 0 \tag{8}$$

and the endpoint conditions $\rho_t|_{t=0} = \rho_0, \rho_t|_{t=1} = \rho_1$. The optimal density field is the displacement interpolation $\rho_t = \Phi_{t\sharp}\rho_0$, where Φ_t is the flow map of the optimal vector field. The flow map is $\Phi_t(x) = (1-t)x + tT(x)$, that is to say it moves mass along linear trajectories from a point x to its image $T(x)$ under the Brenier map. In particular, $T = \Phi_1$, that is, it is the time-one map of a continuous flow.

The Moser map has the advantage that it is more explicit, but the price to pay is that the densities ρ_0 and ρ_1 must be assumed to be sufficiently smooth and strictly positive on a bounded domain with smooth boundary. The construction goes as follows (see Moser (1965); Dacorogna & Moser (1990), and see Friesecke (2024) for a textbook account). Define the velocity field

$$v_t(x) = \frac{\nabla u(x)}{(1-t)\rho_0(x) + t\rho_1(x)}, \tag{9}$$

where u solves the following Poisson equation with Neumann boundary conditions

$$-\Delta u = \rho_1 - \rho_0 \text{ in } \Omega, \quad \nabla u \cdot n = 0 \text{ on } \partial\Omega. \tag{10}$$

Here $n(x)$ is the outward unit normal to $\partial\Omega$ at x. This construction guarantees that the linear density interpolation $\rho_t = (1-t)\rho_0 + t\rho_1$ satisfies the continuity equation equation 8. The velocity field v_t gives rise to a flow map $\Phi_t : \Omega \to \Omega$, by solving the ordinary differential equation

$$\frac{d}{dt}\Phi_t(x) = v_t(\Phi_t(x)), \qquad \Phi_0(x) = x. \tag{11}$$

The map $T = \Phi_1$ is then a smooth diffeomorphism satisfying $(\Phi_1)_\sharp \mu = \nu$.

Finally, for both the Brenier map and the Moser map, the composition structure equation 7 is obtained by discretizing the time interval $[0, 1]$ into ℓ uniform subintervals, each of length $1/\ell$, and introducing the flow maps $\Phi_{s,t}$ by replacing the initial condition in equation 11 by $\Phi_{s,t}(x)|_{t=s} = x$. The flow map over the k-th timestep is then, letting $t_k = k/\ell$,

$$G^{(k)}(x) = \Phi_{t_{k-1},t_k}(x) \approx x + \frac{1}{\ell} v_{t_k}(x), \qquad k = 1, \ldots, \ell,$$

providing the decomposition

$$T = G^{(\ell)} \circ G^{(\ell-1)} \circ \cdots \circ G^{(1)}.$$

5 Theoretical results

In the following section, we provide a rigorous theoretical foundation for the Boltzmann generator method Noé et al. (2019), by combining a regularization argument, a rigorous construction of Moser transport for low-regularity endpoint densities, and known approximation results for RealNVP flows.

5.1 Regularization

Because of the nonexistence of smooth transport maps revealed by Proposition 1, we begin by introducing a regularization of the potential such that the error in the generated samples can be made arbitrarily small, depending on the choice of the regularization parameter.

Theorem 1 (Convergence of regularized Boltzmann distributions)**.** *For an arbitrary number N of atoms, consider the configuration space Ω, where $\Omega \subset \mathbb{R}^{3N}$ is open and bounded. Let $\beta > 0$, and let $U : \Omega \to \mathbb{R} \cup \{+\infty\}$ be a potential satisfing equation 2.*

a) *There exists a family of regularized potentials $\{U_\varepsilon\}_{\varepsilon>0}$, $U_\varepsilon : \Omega \to \mathbb{R}$, satisfying:*

- *(i)* ***Regularity:*** *U_ε is C^1 and Lipschitz.*
- *(ii)* ***Uniform lower bound:*** *There exists $M > 0$ such that $U_\varepsilon(x) \geq -M$ for all $x \in \Omega$ and all $\varepsilon > 0$.*
- *(iii)* ***Pointwise convergence:*** *For all $x \in \Omega$, $\lim_{\varepsilon\to 0} U_\varepsilon(x) = U(x)$.*
- *(iv)* ***No change except at high energy:*** *$U_\varepsilon(x) = U(x)$ for all x with $U(x) \leq \frac{1}{\varepsilon}$.*

b) *Define the normalized Boltzmann densities*

$$\rho_\varepsilon(x) = \frac{F_\varepsilon(x)}{Z_\varepsilon}, \qquad \rho(x) = \frac{F(x)}{Z},$$

with the partition functions

$$Z_\varepsilon = \int_\Omega F_\varepsilon(x)\,dx, \qquad Z = \int_\Omega F(x)\,dx.$$

Then ρ_ε is C^1 and Lipschitz, and

$$\lim_{\varepsilon\to 0} \|\rho_\varepsilon - \rho\|_{L^1(\Omega)} = 0.$$

A important feature of the above regularization is property (iv). Thus the potential, the molecular dynamics trajectories, and the Boltzmann distribution are unchanged except in regions of very high energy. At room temperature, trajectories will in practice not reach such energies and the above regularization is only needed for rigorous theory but not for practical learning of normalized flows.

Proof. For $\lambda > 0$, define the set $K_\lambda = \{x \in \Omega : U(x) > \lambda\}$. Choose a cutoff function $\zeta \in C^1(\mathbb{R}^{3N})$ with $0 \leq \zeta \leq 1$ which is 1 on $K_{2/\varepsilon}$ and 0 outside $K_{1/\varepsilon}$, and let $U_\varepsilon(x) = (1 - \zeta(x))U(x) + \zeta(x) \cdot \frac{2}{\varepsilon}$. Then (i) and (iv) are satisfied. Moreover U_ε is everywhere finite, C^1, and globally Lipschitz, and so is ρ_ε. Since the U_ε are uniformly bounded below and converge pointwise to U, the Boltzmann factors $F_\varepsilon = e^{-\beta U_\varepsilon}$ converge pointwise to $F = e^{-\beta U}$. The bound $F_\varepsilon \leq e^{\beta M}$ allows application of the dominated convergence theorem, implying $Z_\varepsilon \to Z$ and $F_\varepsilon \to F$ in $L^1(\Omega)$. Hence $\rho_\varepsilon = F_\varepsilon/Z_\varepsilon \to \rho = F/Z$ in $L^1(\Omega)$, completing the argument. □

5.2 Moser transport map: existence and approximation by RealNVP flow

We now prove that the Moser transport map exists between low-regularity endpoint densities, thereby extending the classical construction in a smooth setting Moser (1965); Dacorogna & Moser (1990). We then show Sobolev approximability of the map by invertible neural networks.

Theorem 2 (Existence and RealNVP approximation of the Moser transport map)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded C^3 domain. Let ρ_0 and ρ_1 be Lipschitz continuous probability densities on Ω which are strictly positive with* $\inf_\Omega \rho_0 > 0$, $\inf_\Omega \rho_1 > 0$.

a) *There exists a solution $u \in W^{2,2}(\Omega)$ to the Neumann problem equation 10, and this solution belongs to $W^{3,p}(\Omega)$ for all $p > 1$. The corresponding Moser vector field equation 9 is Lipschitz. The associated Moser transport map $T = \Phi_1$ defined by equation 11 is an invertible bilipschitz map and transports ρ_0 to ρ_1, that is to say it satisfies equation 4.*

If in addition ρ_0 and ρ_1 are C^1, then:

b) *The Moser transport map $T = \Phi_1$ is a C^1-diffeomorphism.*

c) *For every $p > d$ and every $\varepsilon > 0$, there exists an invertible RealNVP neural network $F_\theta : \Omega \to \Omega$ such that*

$$\|F_\theta - T\|_{W^{1,p}(\Omega)} < \varepsilon.$$

Proof. We begin with a). Existence of a $W^{2,2}$ solution to the Neumann problem equation 10 is a standard fact of PDE theory, noting that $\rho_1 - \rho_0$ belongs to $L^2(\Omega)$ and integrates to 0. Higher $W^{3,p}$ regularity is proved in Grisvard (2011) Lemma 2.4.2.2 and Theorem 2.5.1.1, noting that $\rho_1 - \rho_0$ belongs to $W^{1,p}(\Omega)$. This higher regularity means $\nabla u \in W^{2,p}$, and the Sobolev embedding theorem for $p > d$ gives $\nabla u \in C^1(\overline{\Omega})$. In particular, ∇u is Lipschitz. Together with the assumptions on ρ_0 and ρ_1 this yields that the vector field v_t defined in equation 9 is Lipschitz in x and continuous in t. It is then a standard fact of ODE theory that the associated flow map $\Phi_t : \Omega \to \Omega$ exists and is an invertible bilipschitz map for any t. By construction T transports ρ_0 to ρ_1. b) follows from the fact that under the additional assumption made, the vector field equation 9 is C^1. As regards c), RealNVP coupling flows are dense with respect to the $W^{1,p}$ topology in the class of C^1-diffeomorphisms on compact domains Ishikawa et al. (2023). For an earlier, related approximation result in Sobolev spaces (namely approximability of Lipschitz functions by ReLu networks in the $W^{s,p}$ topology for all $s < 1$) see Gühring et al. (2020). Thus T can be approximated arbitrarily well in $W^{1,p}(\Omega)$ by a RealNVP map F_θ. Finally, being a RealNVP map, F_θ is invertible. □

5.3 Convergence of push-forward density

Next we prove convergence of the push-forward density under the RealNVP flow, in the narrow sense. This relies on the $W^{1,p}$ convergence result from Theorem 2, the Sobolev embedding theorem, and simple estimates.

Recall that a sequence of probability distributions ρ_j on any closed subset of $\mathbb{R}^d$ converges narrowly to a probability distribution ρ on Ω if

$$\int_\Omega f(y) d\rho_j(y) \to \int_\Omega f(y) d\rho(y) \text{ for all } f \in C_b(\Omega), \tag{12}$$

where $C_b(\Omega)$ is the space of bounded continuous functions on Ω.

Theorem 3 (Convergence of push-forward densities)**.** *Let Ω, ρ_0, ρ_1 be as in Theorem 2* b)*. Then there exists a sequence of invertible realNVP neural networks $F_{\theta_j} : \Omega \to \Omega$ such that*

$$F_{\theta_j \sharp} \rho_0 \to \rho_1 \text{ narrowly as } j \to \infty.$$

The theorem says that bounded continuous observables are approximated correctly by the RealNVP coupling flow.

We conjecture that such an approximation result is also possible in the strong L^1 sense. This would follow if the RealNVP map F_θ from Theorem 2 could be constructed in such a way that, in addition, its inverse F_θ^{-1} approximates T^{-1} arbitrarily well in $W^{1,p}(\Omega)$. Such a result with $p = d$ would imply that the inverse Jacobian appearing in equation 6 is approximated in $L^1(\Omega)$.

Proof. Let $T : \Omega \to \Omega$ be the Moser map between ρ_0 and ρ_1 from Theorem 2 a), b). By Theorem 2 c) and the Sobolev embedding $W^{1,p}(\Omega) \hookrightarrow C^0(\overline{\Omega})$ for $p > d$, there exists a sequence of invertible RealNVP neural network F_{θ_j} such that

$$\sup_{x\in\Omega} |F_{\theta_j}(x) - T(x)| \to 0 \ (j \to \infty). \tag{13}$$

For any given $f \in C_b(\Omega)$ we now calculate using the change-of-variables formula for the push-forward

$$\begin{aligned}
\Big| \int f \, dF_{\theta_j \sharp}\rho_0 - \int f \, d\rho_1 \Big| &= \Big| \int f \, dF_{\theta_j \sharp}\rho_0 - \int f \, dT_\sharp \rho_0 \Big| \\
&= \Big| \int f(F_{\theta_j}(x)) d\rho_0(x) - \int f(T(x)) \, d\rho_0(x) \Big| \\
&\leq \sup_{x\in\Omega} \big| f(F_{\theta_j}(x)) - f(T(x)) \big| \int 1 \, d\rho_0 \ \to 0 \ (j \to \infty),
\end{aligned}$$

where the convergence in the last line is due to the fact the first factor converges to zero thanks to equation 13 and the continuity of f, and the second factor is equal to 1 since ρ_0 is a probability measure. Since f was arbitrary, this establishes the asserted narrow convergence equation 12. □

5.4 Main result

Finally we are in a position to state our main result.

Theorem 4 (Theoretical justification of the Boltzmann generator method)**.** *For an arbitrary number N of atoms, any bounded domain $\Omega \subset \mathbb{R}^{3N}$ of class C^3, any inverse temperature $\beta > 0$, any potential $U : \Omega \to \mathbb{R} \cup \{+\infty\}$ satisfing equation 2, any C^1 and Lipschitz reference probability density on Ω which is strictly positive with $\inf_\Omega \rho_0 > 0$, and any $\epsilon > 0$, there exists an invertible RealNVP neural network $F_\theta : \Omega \to \Omega$ such that*

$$W_2(F_{\theta\sharp}\rho_0, \, \rho_1) < \varepsilon,$$

where ρ_1 is the Boltzmann distribution equation 1 and W_2 denotes the Wasserstein-2 distance.

Proof. This follows by combining Theorems 1, 2, 3 and the fact (see e.g. Villani (2003); Friesecke (2024)) that the Wasserstein-2 distance metrizes narrow convergence of probability measures on compact sets. □

6 Numerical Experiments

In this section, we illustrate the Boltzmann generator method for two models: (1) overdamped Langevin dynamics for a particle in a two-dimensional double-well potential, (2) full molecluar dynamics for the alanine dipeptide molecule. We also check how close the true and generated distributions are in the Wasserstein-2 distance and find close agreement, in line with what is expected from our theoretical findings (Theorem 4).

6.1 2D Double-well potential

We simulated the Boltzmann distribution equation 1 for a 2D double-well potential using overdamped Langevin dynamics

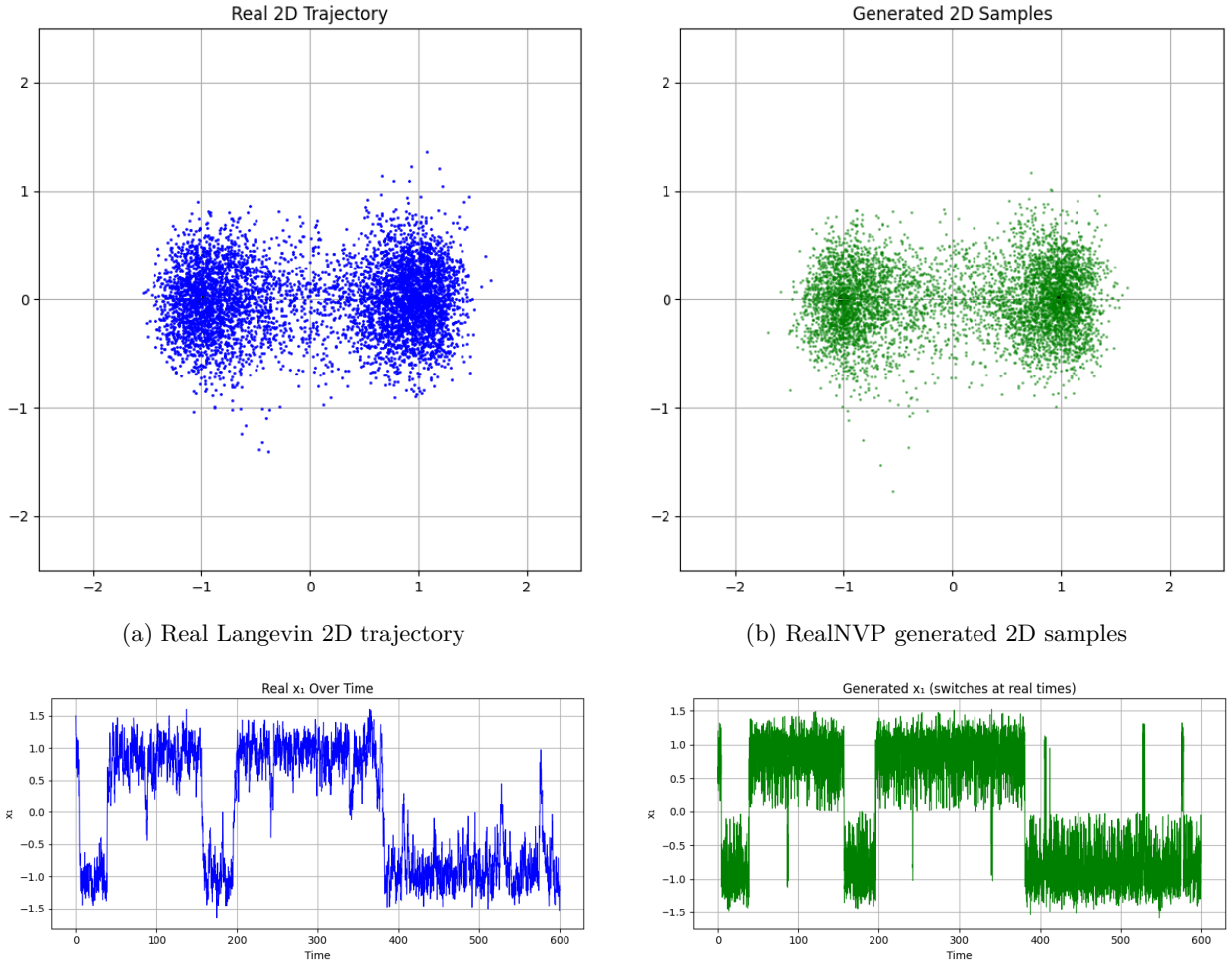
$$dx = -\nabla U(x) \, dt + \sqrt{\tfrac{2}{\beta}} dW$$

to obtain stochastic trajectories whose position distribution approaches the Boltzmann distribution at long time. The potential is defined as

$$U(x_1, x_2) = \frac{1}{4}(x_1^2 - 1)^2 + \frac{1}{2}x_2^2,$$

which forms two stable wells along the horizontal axis. The random thermal fluctuations allow the particle to cross the energy barrier, producing transitions between metastable basins. This behavior is prototypical for real molecules. For the real data shown in Figure 1 we used 600 timesteps.

We trained a RealNVP flow with 6 coupling layers and 128 hidden units, resulting in a total of 7,704 trainable parameters, to learn the equilibrium distribution from Langevin trajectories. As a loss function we used negative log likelihood. After training, we generated 10,000 independent samples from the flow.



(a) Real Langevin 2D trajectory

(b) RealNVP generated 2D samples

(c) Real x_1 component over time

(d) Generated x_1 component over time

Figure 1: Comparison of real Langevin dynamics and RealNVP generated samples for the 2D double-well potential.

To evaluate the accuracy of the trained model, we computed the Wasserstein-2 distance W_2 between the samples generated by the RealNVP and the reference Langevin trajectories. We used the Python Optimal Transport (POT) library Flamary et al. (2021)to compute the 2D Wasserstein distance. The 2D joint distribution yields a W_2 distance of 0.0993, while the per-coordinate distances are 0.0101 for x_1 and 0.0036 for x_2, with an average of 0.0069.

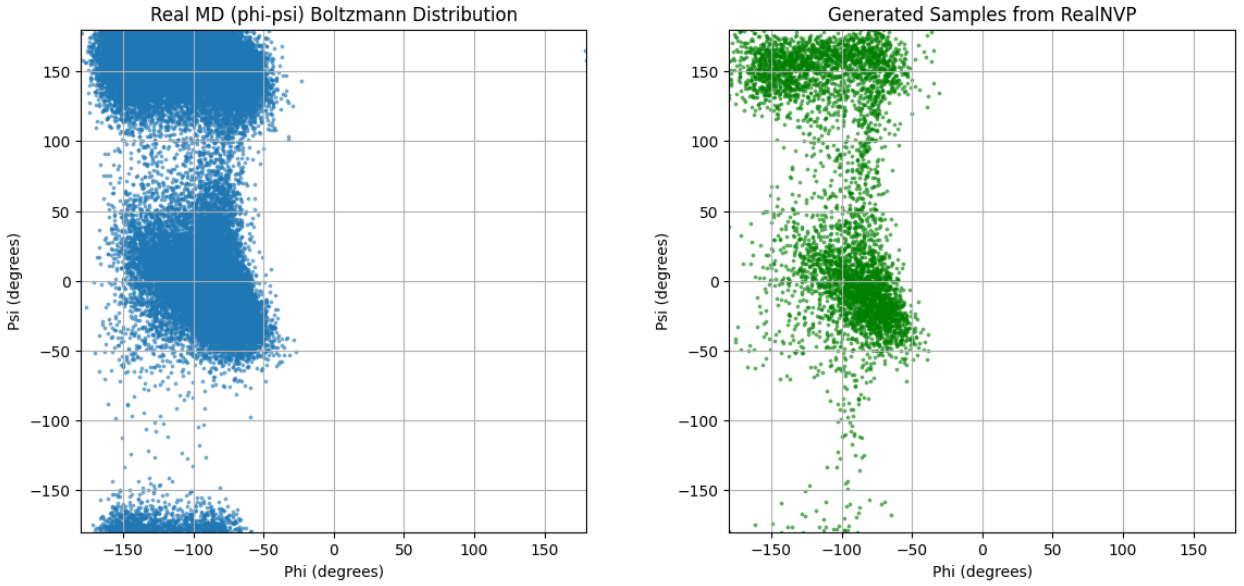
The model also exactly reproduces the switching dynamics, with both real and generated trajectories exhibiting exactly 41 transitions between the two wells. These results demonstrate that the RealNVP does not just successfully capture the equilibrium Boltzmann distribution, but also the metastable dynamics of the double-well system.

6.2 Alanine Dipeptide Molecular System

We validated the Boltzmann generator method on realistic molecular data by applying it to alanine dipeptide, a standard benchmark system for studying conformational dynamics Chodera et al. (2007).

We used a data set with 6 independent 10 ns molecular dynamics (MD) trajectories (60 ns, total frames 59,994) simulated in an implicit solvent at $T = 300$ K using the AMBER99SB-ILDN force field Lindorff-Larsen et al. (2010). The backbone dihedral angles, Φ and Ψ, which range over $[-\pi, \pi]$, serve as the slow collective variables of the system and capture the primary conformational changes. The empirical free energy surface $F(\Phi, \Psi)$ is estimated from the long MD trajectories.

We trained a RealNVP flow on dihedral angle samples projected to (Φ, Ψ) space to learn the equilibrium densities. We used $\ell = 6$ coupling layers with 64 hidden units for each (s_θ, t_θ) subnetwork, resulting in 2,316 trainable parameters in total. Again we used negative log likelihood as a loss. The model successfully captured the main stable conformations of alanine dipeptide, corresponding to the well-known metastable regions of the (Φ, Ψ) free energy landscape.



(a) True Boltzmann distribution

(b) RealNVP generated distribution

Figure 2: Comparison of true and generated Φ–Ψ distributions for alanine dipeptide

The generated (Φ, Ψ) samples closely follow the true Boltzmann distribution obtained from the MD simulation. The main metastable regions appear at similar positions in both plots, indicating that the RealNVP successfully captures the preferred conformations. The small differences seen near transition regions are expected due to sampling noise.

To quantitatively assess the accuracy of the generated samples, as in the previous example we computed the Wasserstein-2 distance between the true and generated distributions using the Python optimal transport library Flamary et al. (2021). Following standard practice, we subsampled 1000 points from each distribution. The resulting distance of 9.21° confirms very good agreement between the generated and true Boltzmann distributions.

7 Discussion and Conclusions

We have shown that RealNVP-based normalizing flows are capable in theory of producing correct samples from Boltmann distributions of molecules despite their low regularity. Moreover our numerical simulations

add further confirmation that these flows reproduce key features of such distributions in practice as first observed in Noé et al. (2019), and quantify the error between the generated and true distributions via the Wasserstein distance.

Our theoretical work shows that a transport map with very nice theoretical properties, e.g. simultaneously controlled size of gradients and inverse gradients, is provided by a simplified version of optimal transport, namely Moser transport. Current RealNVP training protocols provide more complicated transport maps which are expected to exhibit uncontrollably large Lipschitz constants for the inverse map as discussed in other contexts in Behrmann et al. (2021); Kirichenko et al. (2020). This calls into question numerical invertibility of the flow, and robustness for unseen initial data. An interesting question left open by our work is whether good numerical invertibility in high dimensions can be achieved by a training bias towards "nice" transport maps like the Moser map without introducing too much additional computational complexity.

Acknowledgements

This work was funded by DFG within the priority program SPP 2298 *Theoretical Foundations of Deep Learning*, project number 543965508.

References

Jens Behrmann, Paul Vicol, Kuan-Chieh Wang, Roger Grosse, and Jörn-Henrik Jacobsen. Understanding and mitigating exploding inverses in invertible neural networks. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1792–1800. PMLR, 2021. URL `https://proceedings.mlr.press/v130/behrmann21a.html`.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

Luis A. Caffarelli. Some regularity properties of solutions of monge ampère equation. *Communications on Pure and Applied Mathematics*, 44(8-9):965–969, 1991.

J. Chemseddine, P. Hagemann, G. Steidl, and C. Wald. Conditional wasserstein distances with applications in bayesian ot flow matching. *arXiv preprint arXiv:2403.18705*, 2025. URL `https://arxiv.org/pdf/2508.18122`.

John D. Chodera, William C. Swope, Jed W. Pitera, and Ken A. Dill. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *The Journal of Chemical Physics*, 126(15):155101, 2007.

Bernard Dacorogna and Jürgen Moser. On a partial differential equation involving the Jacobian determinant. *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, 7(1):1–26, 1990.

Guido De Philippis and Alessio Figalli. $W^{2,1}$ regularity for solutions of the monge–ampère equation. *Inventiones mathematicae*, 192(1):55–69, 2011.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations (ICLR)*, 2017.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL `http://jmlr.org/papers/v22/20-451.html`.

Gero Friesecke. *Optimal Transport: A Comprehensive Introduction to Modeling, Analysis, Simulation, Applications*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2024. ISBN 978-1-61197-808-7.

Pierre Grisvard. *Elliptic Problems in Nonsmooth Domains*, volume 69 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011. ISBN 978-1-611972-02-3. Reprint of the 1985 original [Pitman].

Ingo Gühring, Gitta Kutyniok, and Philipp Petersen. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Analysis and Applications*, 18(5):803–859, 2020.

Quentin Hernandez, Alberto Badias, David Gonzalez, Francisco Chinesta, and Elias Cueto. Structure-preserving neural networks. *Journal of Computational Physics*, 396:400–415, 2019.

Isao Ishikawa, Takeshi Teshima, Koichi Tojo, Kenta Oono, Masahiro Ikeda, and Masashi Sugiyama. Universal approximation property of invertible neural networks. *Journal of Machine Learning Research*, 24(287): 1–68, 2023.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, 2018.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Why normalizing flows fail to detect out-of-distribution data. *arXiv preprint arXiv:2006.08545*, June 2020. doi: 10.48550/arXiv.2006.08545. URL `https://arxiv.org/abs/2006.08545`.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, Alexander Filippov, and Evgeny Burnaev. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021. doi: 10.48550/ARXIV.2106.01954.

Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1950–1958, 2010. doi: 10.1002/prot.22711.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *International Conference on Learning Representations (ICLR)*, 2023. URL `https://arxiv.org/abs/2210.02747`.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023. URL `https://openreview.net/forum?id=XVjTT1nw5z`.

Jürgen Moser. On the volume elements on a manifold. *Transactions of the American Mathematical Society*, 120:286–294, 1965.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Aram-Alexandre Pooladian, Kirill Neklyudov, Benjamin Schär, Nikola Vuckovic, Octavian-Eugen Ganea, and Dimitrios Milios. Multisample flow matching: Straightening flows with minibatch couplings. *arXiv preprint arXiv:2304.14772*, April 2023. doi: 10.48550/arXiv.2304.14772. URL `https://arxiv.org/abs/2304.14772`.

Alexander Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yan Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. Multisample flow matching: Straightening flows with minibatch couplings. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, Honolulu, Hawaii, July 2023. URL `https://frontiers4lcd.github.io/`.

G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.

Théo Uscidda and Marco Cuturi. The Monge gap: A regularizer to learn all transport maps. *International Conference on Machine Learning (ICML)*, 2023.

Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003. ISBN 082183312X.

Hao Wu, Andreas Mardt, Luca Pasquali, and Frank Noé. Deep generative Markov state models. *arXiv preprint arxiv:1805.07601*, 2018.