Equivariant Flows: Exact Likelihood Generative Learning for Symmetric Densities

Jonas Köhler [*1] **Leon Klein** [*1] **Frank Noé** [1 2 3]

Abstract

Normalizing flows are exact-likelihood generative neural networks which approximately transform samples from a simple prior distribution to samples of the probability distribution of interest. Recent work showed that such generative models can be utilized in statistical mechanics to sample equilibrium states of many-body systems in physics and chemistry. To scale and generalize these results, it is essential that the natural symmetries in the probability density – in physics defined by the invariances of the target potential – are built into the flow. We provide a theoretical sufficient criterion showing that the distribution generated by *equivariant* normalizing flows is invariant with respect to these symmetries by design. Furthermore, we propose building blocks for flows which preserve symmetries which are usually found in physical/chemical many-body particle systems. Using benchmark systems motivated from molecular physics, we demonstrate that those symmetry preserving flows can provide better generalization capabilities and sampling efficiency.

1. Introduction

Generative learning using exact-likelihood methods based on invertible transformations has had remarkable success in accurately representing distributions of images (Kingma & Dhariwal, 2018), audio (Oord et al., 2017) and 3D point cloud data (Liu et al., 2019b; Noé et al., 2019).

Recently, *Boltzmann Generators* (BG) (Noé et al., 2019) have been introduced for sampling Boltzmann type distributions $\rho'(x) \propto \exp(-u(x))$ of high-dimensional many-body problems, such as valid conformations of proteins.

This approach is widely applicable in the physical sciences, and has also been employed in the sampling of spin lattice states (Nicoli et al., 2019; Li & Wang, 2018) and nuclear physics models (Albergo et al., 2019). In contrast to typical generative learning problems, the target density $\rho'(x)$ is specified by definition of the many-body energy function $u(x)$ and the difficulty lies in learning to sample it efficiently. BGs do that by combining an exact-likelihood method that is trained to approximate the Boltzmann density $\rho'(x)$, and a statistical mechanics algorithm to reweigh the generated density to the target density $\rho'(x)$.

Physical systems of interest usually comprise symmetries, such as invariance with respect to global rotations or permutations of identical elements. As we show in experiments ignoring such symmetries in flow-based approaches to density estimation and enhanced sampling, e.g. using BGs, can lead to inferior results which can be a barrier for further progress in this domain. In our work we thus provide the following contributions:

- We show how symmetry-preserving generative models, satisfying the exact-likelihood requirements of Boltzmann generators, can be obtained via *equivariant flows*.
- We show that symmetry preservation can be critical for success by showing experiments on highly symmetric many-body particle systems. Concretely, equivariant flows are able to approximate the system's densities and generalize beyond biased data, whereas approaches based on non-equivariant normalizing flows cannot.
- We provide a numerically tractable and efficient implementation of the framework for many-body particle systems utilizing gradient flows derived from a simple mixture potential.

While this work focuses mostly on applications in the physical sciences the results could provide a takeaway towards a greater ML audience: studying symmetries of target distributions and considering them in the architecture of a density estimation / sampling mechanism can lead to better generalization and can even be critical for successful learning.

[*]Equal contribution [1]Freie Universität Berlin, Department of Mathematics and Computer Science. [2]Freie Universität Berlin, Department of Physics. [3]Rice University, Department of Chemistry. Correspondence to: Jonas Köhler <jonas.koehler@fu-berlin.de>, Frank Noé <frank.noe@fu-berlin.de>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

2. Related Work

Statistical Mechanics The workhorse for sampling Boltzmann-type distributions $p(x) \propto \exp(-u(x))$ with known energy function $u(x)$ are Molecular dynamics (MD) and Markov-Chain Monte-Carlo (MCMC) simulations. MD and MCMC take local steps in configurations x, are guaranteed to sample from the correct distribution for infinitely long trajectories, but are subject to the *rare event sampling problem*, i.e. the get stuck in local energy minima of $u(x)$ for long time. Statistical mechanics has developed many tools to speed up rare events by adding a suitable bias energy to $u(x)$ and subsequently correcting the generated distribution by reweighing or Monte-Carlo estimators using the ratio of true over generated density, e.g. (Torrie & Valleau, 1977; Bennett, 1976; Laio & Parrinello, 2002; Wu et al., 2016). These methods can all speed up MD or MCMC sampling significantly, but here we pursue sampling of the equilibrium density with flows.

Normalizing Flows Normalizing flows (NFs) are diffeomorphisms $f_\theta \colon \mathbb{R}^n \to \mathbb{R}^n$ which transform samples $z \sim \rho$ from a simple prior density ρ into samples $x = f_\theta(z)$ (Tabak et al., 2010; Tabak & Turner, 2013; Rezende & Mohamed, 2015; Papamakarios et al., 2019). Denoting the density of the transformed samples ρ_{f_θ}, we obtain the probability density of any generated point via the *change of variables* equation:

$$\rho_{f_\theta}(x) = \rho\left(f_\theta^{-1}(x)\right) \det \frac{\partial f_\theta^{-1}(x)}{\partial x}.$$

ρ_{f_θ} is also called the *push-forward* of ρ along f_θ.

While flows can be used to build generative models by maximizing the likelihood on a data sample, having access to tractable density is especially useful in variational inference (Rezende & Mohamed, 2015; Tomczak & Welling, 2016; Louizos & Welling, 2017; Berg et al., 2018) or approximate sampling from distributions given by an energy function (Oord et al., 2017), which can be made exact using importance sampling (Müller et al., 2018; Noé et al., 2019).

The majority of NFs can be categorized into two families: (1) Coupling layers (Dinh et al., 2014; 2016; Kingma & Dhariwal, 2018; Müller et al., 2018), which are a subclass of autoregressive flows (Germain et al., 2015; Papamakarios et al., 2017; Huang et al., 2018; De Cao et al., 2019; Durkan et al., 2019), and (2) residual flows (Chen et al., 2018; Zhang et al., 2018; Grathwohl et al., 2018; Behrmann et al., 2018; Chen et al., 2019).

Symmetries in flow models have been discussed in the context of permutations in graphs (Liu et al., 2019a). A preliminary account of equivariant normalizing flows has been given in two recent workshop submissions (Rezende et al., 2019; Köhler et al., 2019).

Boltzmann-Generating Flows While flows and other generative models are typically used for estimating the an unknown density ρ' from samples and then generating new samples from it, BGs know the desired target density $\rho'(x) \propto \exp(-u(x))$ up to a prefactor and aim at learning to efficiently sample it (Noé et al., 2019).

A BG combines two elements to achieve this goal:

1. An exact-likelihood generative model that generates samples x_k from a density ρ_{f_θ} that approximates the given Boltzmann-type target density ρ'.
2. An algorithm to reweigh the generated density to the target density ρ'. For example, using importance sampling the asymptotically unbiased estimator of the expectation value of observable $O(x)$ is:

$$\mathbb{E}_{x\sim\rho'}[O] \approx \frac{\sum_k w(x_k) O(x_k)}{\sum_k w(x_k)}, \quad x_k \sim \rho_{f_\theta},$$

where the importance weights

$$w(x_k) = \exp(-u(x_k))/\rho_{f_\theta}(x_k)$$

can be computed from the trained flow.

The exact likelihood model is needed in order to be able to conduct the reweighing step. When a flow is used in order to generate asymptotically unbiased samples of the target density, we speak of a Boltzmann-generating flow.

Boltzmann-generating flows are trained to match $\rho_{f_\theta} \approx \rho'$ using loss functions that also appear in standard generative learning problems, but due to the explicit availability of $\exp(-u(x))$ their functional form and interpretation changes:

1. *KL-training* We minimize the reverse Kullback-Leibler divergence $KL(\rho_{f_\theta} \| \rho')$:

$$\mathcal{L}_{\text{KL}} = \mathbb{E}_{z\sim\rho}\left[u(f_\theta(z)) - \log\left|\det \frac{\partial f_\theta(z)}{\partial z}\right|\right].$$

This approach is also known as energy-based training where the energy corresponding to the generated density is matched with $u(x)$.

2. *ML-training*: If data $\{x_n\}_{n=1\ldots N}$ from a data distribution ρ'_{data} is given that at least represents one or a few high-probability modes of ρ', we can maximize the likelihood under the model, as is typically done when performing density estimation:

$$\mathcal{L}_{\text{ML}} = \mathbb{E}_{x\sim\rho'_{\text{data}}}\Bigg[-\log\rho\left(f_\theta^{-1}(x)\right) - \log\left|\det \frac{\partial f_\theta^{-1}(x)}{\partial x}\right|\Bigg].$$

The final training loss is then obtained using a convex sum over both losses, where the mixing parameter λ may be changed from 0 to 1 during the course of training:

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{\text{ML}} + \lambda\mathcal{L}_{\text{KL}}.$$

3. Invariant Densities via Equivariant Flows

In this work we consider densities ρ, ρ' over euclidean vector spaces $\mathbb{R}^n$ which are invariant w.r.t. to symmetry transformations e.g. given by rotations and permutations of the space. In other words, we want to construct flows such that both, the prior and the target density share the same symmetries.

More precisely, let G be a group which acts on $\mathbb{R}^n$ via a representation $R\colon G \to GL(n), g \to R_g$ and assume that ρ is invariant w.r.t. G, i.e. $\forall g \in G, x \in \mathbb{R}^n\colon \rho(R_g x) = \rho(x)$. We first remark that for any $g \in G$ the matrix R_g satisfies $\det(R_g) \in \{-1, 1\}$[1]. This allows us to formulate our result:

Theorem 1. *Let ρ is a density on $\mathbb{R}^n$ which is G-invariant and $G > H$. If f is a H-equivariant diffeomorphism, i.e. $\forall h \in H, x \in \mathbb{R}^n\colon f(R_h x) = R_h f(x)$, then ρ_f is H-invariant.*

As a direct consequence if $H < O(n)$, any push-forward of an isotropic normal distribution along a H-equivariant diffeomorphism will result in a H-invariant proposal density.

4. Constructing Equivariant Flows

In general it is not clear how to define equivariant diffeomorphisms which provide tractable inverses and Jacobians. We will provide a possible implementations based on the recently introduced framework of *continuous normalizing flows* (CNFs) (Chen et al., 2017).

Equiviariant Dynamical Systems CNFs define a dynamical system via a time-dependent vector field $v\colon \mathbb{R}^n \times [0,\infty) \to \mathbb{R}^n$. If v is globally Lipschitz, we can map each $z \in \mathbb{R}^n$ onto the unique characteristic function $x_{v,z}\colon [0,\infty) \to \mathbb{R}^n$, which solves the Cauchy-problem

$$\tfrac{d}{dt}x(t) = v(x_{v,z}(t), t), \qquad x_{v,z}(0) = z.$$

This allows us to define a bijection $F_{v,T}\colon \mathbb{R}^n \to \mathbb{R}^n$ for each $T \in [0,\infty)$ by setting

$$F_{v,T}(z) = x_{v,z}(0) + \int_0^T dt\; v(x_{v,z}(t), t).$$

Given a density ρ on $\mathbb{R}^n$, each T defines a push-forward $\rho_{F_{v,T}}$ along $F_{v,T}$, which satisfies

$$\tfrac{d}{dt}\log \rho_{F_{v,t}}(x_{v,z}(t)) = -\text{div}\left(v(x_{v,z}(t), t)\right).$$

[1] All proofs and derivations can be found in the Suppl. Material.

By following the characteristic this allows to compute the total density change as

$$\log \frac{\rho_{F_{v,T}}(x_{v,z}(T))}{\rho(x_{v,z}(0))} = -\int_0^T dt\;\text{div}\left(v(x_{v,z}(t), t)\right).$$

Equivariant flows can thus be constructed very naturally:

Theorem 2. *Let v be a H-equivariant vectorfield on $\mathbb{R}^n$ (not necessarily bijective). Then for each $T \in [0,\infty)$ the bijection $F_{v,T}$ is H-equivariant.*

Consequently, if ρ is a G-invariant density on $\mathbb{R}^n$ and $G > H$, then each push-forward $\rho_{F_{v,T}}$ is H-invariant.

Equivariant Gradient Fields There has been a significant amount of work in recent years proposing G-equivariant functions for different groups acting on $\mathbb{R}^n$. A generic implementation however is given by a gradient flow: if $\Phi\colon \mathbb{R}^n \to \mathbb{R}$ is a G-invariant function, the vector $\nabla_x \Phi$ will transform G-equivariantly.

Gradient flows (not necessarily G-equivariant) can map any ρ onto any ρ' over $\mathbb{R}^n$ as long as both densities do not vanish (Benamou & Brenier, 2000; McCann, 2001) and have been discussed in the context of density estimation (Zhang et al., 2018; Papamakarios et al., 2019).

Numerical Implementations While providing an elegant solution, implementing equivariant flows using continuous gradient flows is numerically challenging due to three aspects.

First, even if $F_{v,T}$ is invertible assuming exact integration, there are no such guarantees for any discrete-time approximation of the integral, e.g. using Euler or Runge-Kutta integration. Thus, Chen et al. propose adaptive-step solvers, such as Dopri5 (Dormand & Prince, 1980), which can require hundreds of vector field evaluations to reach satisfying numerical accuracy.

Second, in order to train v via the adjoint method as suggested by Chen et al., gradients of the loss w.r.t. parameters are obtained via backward integration. However, in general, there are no guarantees that this procedure is stable, which therefore can result in very noisy gradients, leading to long training times and inferior final results (Gholami et al., 2019). In contrast to this *optimize-then-discretize* (OTD) approach, Gholami et al. suggest to unroll the ODE into a fixed-grid sequence and backpropagate the error using classic automatic differentation (AD). Such a *discretize-then-optimize* (DTO) approach will guarantee that gradients are computed correctly, but might suffer from inaccuracy due to the discretization errors as mentioned before. Throughout our experiments, we rely on the latter approach during training and show that for our presented architecture OTD and DTO will yield similar results, while the latter offers a significant speedup per iteration, more robust training and faster convergence.

Finally, computing the divergence of v using off-the-shelf AD frameworks requires $O(n)$ backpropagation passes, which would result in an infeasible overhead for high-dimensional systems (Grathwohl et al., 2018). Thus, Grathwohl et al. suggest an approximation via the Hutchinson-estimator (Hutchinson, 1989). This is an unbiased rank-1 estimator of the divergence where variance scales with $O(n)$. As we show in our experiments, even for small particle systems, relying on such an estimator will render importance weighing and thus the benefits of Boltzmann generating flows useless, e.g. when used in downstream sampling applications. Another approach relies on designing special dynamics functions, in which input dimensions are decoupled and then combine the `detach`-operator with one backpropagation pass to compute the divergence exactly (Chen & Duvenaud, 2019). For general symmetries as studied in this paper such a decoupling is not possible, without either destroying equivariance of the dynamics function, or enforcing it to be trivial. Our proposed vector field based on a simple mixture of Gaussian radial basis functions (RBF) allows computing the divergence numerically exact as one vectorized operation and without relying on AD backward passes.

Relation to Hamiltonian Flows If our space decomposes as $\mathbb{R}^n = \mathbb{R}^m \bigoplus \mathbb{R}^m$ where each element is written as $x = (q, p)$ and where we call q the generalized position and p the generalized momentum, we can define a time-dependent Hamiltonian $\mathcal{H}\colon \mathbb{R}^m \times \mathbb{R}^m \times [0, \infty) \to \mathbb{R}$, which defines the Hamiltonian system

$$v(q, p, t) = \left(\frac{\partial \mathcal{H}(q, p, t)}{\partial p}, -\frac{\partial \mathcal{H}(q, p, t)}{\partial q} \right).$$

If $\mathcal{H}$ factorizes as $\mathcal{H}(q, p, t) = V(q, t) + \frac{1}{2}\|p\|^2$ a numerically stable and finite-time invertible solution of the system is given by *Leapfrog-integration*. Furthermore, due to the symplecticity of v, each $F_{v,T}$ will be volume preserving. Unrolling the Leapfrog-integration in finite time, will result in a stack of *NICE*-layers (Dinh et al., 2014) with equivariant translation updates.

We can always create an artificial Hamiltonian version of any density estimation problem, by augmenting a density $\rho(q)$ on $\mathbb{R}^n$ to $\rho(q, p) = \rho(q) \cdot \rho(p|q)$ on $\mathbb{R}^n \times \mathbb{R}^n$. Due to the interaction between q and p within the flow, we cannot expect that both $\rho(p|q) = \rho(p)$ and $\rho'(p|q) = \rho'(q)$ within a finite number of steps. Thus, if an isotropic normal distribution is used for $\rho(p, q)$, having only access to $\rho'(q)$ will require a variational approximation of $\rho'(p|q)$ (Toth et al., 2019).

If $\mathcal{H}$ is G-invariant, i.e. $\mathcal{H}(R_g q, R_g p, t) = \mathcal{H}(q, p, t)$ for all $g \in G, (q, p) \in \mathbb{R}^m \times \mathbb{R}^m, t \in [0, \infty)$, we see that v will be G-equivariant. This results in the recently proposed framework of *Hamiltonian Equivariant Flows* (HEF) (Rezende et al., 2019), which we thus see as a special case of our framework for densities with linearly represented symmetries defined over $\mathbb{R}^n$. On the other hand, HEFs can handle more general spaces or symmetries with nonlinear representations – in contrast to the presented framework – hence the two approaches are complementary.

For completeness, we note that Hamiltonian flows do not suffer from those numerical complications in the former paragraph, due to symplectic integration and volume preservation. However, in order to compute unbiased estimates of target densities which is essential for physics applications, a variational approximation of $\rho'(p|q)$ cannot be applied.

5. Sampling of Coupled Particle Systems

We evaluate the importance of incorporating symmetry into flows when aiming to sample from symmetric densities, by applying the theoretic framework to the problem of sampling coupled many-body systems of interchangeable particles. Such systems have states $x \in \mathbb{R}^n,\ n = N \cdot D$ consisting of N particles x_i with $D \in [2, 3]$ degrees of freedom, which are coupled via a potential energy $u(x)$. In thermodynamic equilibrium such a system follows a Boltzmann-type distribution $\rho'(x) \propto \exp(-u(x))$. Assuming interchangeable particles in vacuum without external field, we obtain three symmetries (S1-3): u (and thus ρ') does not change if we permute particles (S1), rotate the system around the center of mass (CoM) (S2), or translate the CoM by an arbitrary vector (S3).

Due to the simultaneous occurrence of (S1) and (S2) no autoregressive decomposition / coupling layer can be designed to be equivariant. Either a variable split has to be performed among particles or among spatial coordinates, which will break permutation and rotation symmetry respectively. Thus, residual flows are the only class of flows which can be applied here. In this work we will rely on CNFs, design an equivariant vector field by taking the gradient field of an invariant potential function, and then combine theorems 1 and 2 to conclude the symmetry of the proposal density.

Invariant Prior Density We first start by designing an invariant prior. By only considering systems with zero CoM symmetry (S3) is easily satisfied. The set of CoM-free systems forms a $(N-1) \cdot D$-dimensional linear subspace $U < \mathbb{R}^n$. Equipping $\mathbb{R}^n$ with an isotropic normal density ρ, implicitly equips U with a normal distribution $\tilde{\rho}$. We can sample it, by sampling $z \sim \rho$ and projecting on U, and evaluate its likelihood for $z \in U$, by computing $\rho(z)$.

Equivariant Vector Field We design our vector field as the gradient field $v(x(t)) = \nabla_{x(t)}\Phi(x(t))$ of a potential $\Phi\colon \mathbb{R}^n \to \mathbb{R}$. If Φ is invariant under symmetry transformations (S1-3) it directly implies equivariance of v.

Our invariant potential Φ is given as a sum of pairwise couplings over particle distances:

$$\Phi(x(t)) = \sum_{ij} \tilde{\Phi}(d_{ij}(t), t)$$

with $r_{ij}(t) = x_i(t) - x_j(t),\ d_{ij}(t) = \|r_{ij}(t)\|$. This yields per-particle updates

$$v_i(x(t)) = \sum_j v_{ij}(x(t)).$$

For a well-chosen coupling potential $\tilde{\Phi}(d_{ij}, t)$ we can express

$$v_{ij}(x(t)) = \underbrace{R(t)^T W K(d_{ij}(t))}_{\phi(d_{ij})} \cdot r_{ij}(t), \tag{1}$$

where $K\colon \mathbb{R} \to \mathbb{R}^M$ and $R\colon \mathbb{R} \to \mathbb{R}^L$ are vector-valued functions, each component is given by a Gaussian RBF and $W \in \mathbb{R}^{T\times M}$ is a trainable weight matrix (see Figure 1).

Using this architecture, the divergence becomes:

$$\operatorname{div} \frac{\partial x(t)}{\partial t} = \sum_{ij} \frac{\partial \phi(d_{ij}(t), t)}{\partial d_{ij}(t)} d_{ij}(t) + D \cdot \phi(d_{ij}(t)).$$

Thus, the gradient and the divergence can be computed *exactly* and *as one vectorized operation* (see Suppl. Material for details).

During training we optimize W and RBF means and bandwidths simultaneously. By keeping weights small and bandwidths large we can control the complexity of the dynamics. As we show in our experiments even a small amount of weight-decay is sufficient to properly optimize the flow with a fixed-grid solver introducing a negligible amount of error during the integration.

Other Invariant Potential Functions While Φ could be modeled by any kind of invariant graph neural networks, such as *SchNet* (Schütt et al., 2017), this would require us to 1) use AD in order to compute $\nabla_{x(t)}\Phi(x(t))$ and 2) compute $\Delta_{x(t)}\Phi(x(t))$ at every function evaluation while integrating v. This implies the numerical challenges as mentioned before. As we show in the Suppl. Mat. our simple couplings are considerably faster, have a fraction of parameters while consistently outperforming neural network approaches to modeling Φ for the studied target systems.

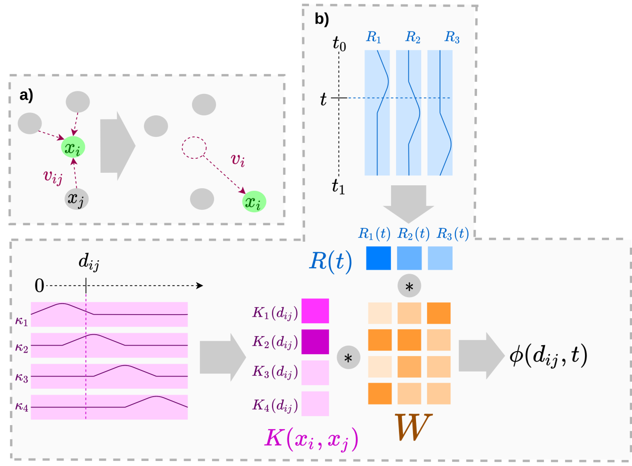

Figure 1. **a)** Each particle x_i is updated by a weighted sum of radial forces depending on distances and the integration time. **b)** Time and distances are expanded in a RBF basis and mixed together with a shared weight matrix.

6. Benchmark Systems

We study two systems where all symmetries (S1), (S2), (S3) are present (Figure 2):

DW-2 / DW-4 The first system is given by $N \in [2, 4]$ particles with a pairwise *double-well* potential acting on particle distances

$$u^{\mathtt{DW}}(x) = \frac{1}{2\tau}\sum_{i,j} a\,(d_{ij} - d_0) + b\,(d_{ij} - d_0)^2 + c\,(d_{ij} - d_0)^4$$

for $D = 2$, which produces two distinct low energy modes separated by an energy barrier. By coupling multiple particles with such double-well interactions we can create a frustrated system with multiple metastable states. Here a, b, c and d_0 are chosen design parameters of the system and τ the dimensionless temperature.

LJ-13 The second system is given by the *Lennard-Jones* (LJ) potential with $N = 13,\ D = 3$. LJ is a model for solid-state models and rare gas clusters. LJ clusters have complex energy landscapes whose energy minima are difficult to find and sample between. These systems have been extensively studied (Wales & Doye, 1997) and are good candidates for benchmarking structure generation methods. In order to prevent particles to dissociate from the cluster at the finite sampling temperature, we add a small harmonic potential to the CoM. The LJ potential with parameters ϵ and r_m at dimensionless temperature τ is defined by

$$u^{\mathtt{LJ}}(x) = \frac{\epsilon}{2\tau}\left[\sum_{i,j}\left(\left(\frac{r_m}{d_{ij}}\right)^{12} - 2\left(\frac{r_m}{d_{ij}}\right)^{6}\right)\right].$$

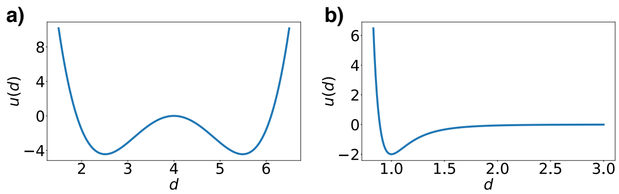

Figure 2. The two model systems: shown are the energy contributions per distance **a)** for the *double-well* and **b)** the *Lennard-Jones* potential.

7. Experiments

7.1. Computation of Divergence

In a first experiment we show that *fast* and *exact* divergence computation can be critical especially when the number of particles grows. We compare different ways to estimate the change of log-density: (1) using brute-force computation relying on AD (2) using the Hutchinson estimator described by Grathwohl et al., and (3) computing the trace exactly in close form.

Brute-force computation quickly yields a significant overhead per function evaluation during the integration, which makes it impractical for online computations (Figure 3 c), such as using the flow within a sampling procedure or just for training. If we use Hutchinson estimation, the error grows quickly with the number of particles (Figure 3 a) and renders reweighing, even for the very simple DW-2 system, impossible (Figure 3 b). By having access to an exact closed-form trace, we obtain the best of both worlds: fast computation and the possibility for exact reweighing (Figure 3 b+c).

7.2. DTO vs. OTD Optimization

In this experiment we show that by simply regularizing W, e.g. using weight decay, OTD and DTO based optimization of the flow barely shows any difference (Figure 4 a), while the former quickly results in a significant overhead due to the increasing number of function evaluations required to match the preset numerical accuracy (Figure 4 b). We compare the OTD implementation presented in (Chen et al., 2018; Grathwohl et al., 2018) using the `dopri5`-option (`atol` $= 10^{-10}$, `rtol` $= 10^{-5}$) to the DTO implementation given by Gholami et al. using a fixed grid of 20 steps and 4th-order Runge-Kutta as solver.

7.3. Statistical Efficiency for Density Estimation

We compare the proposed equivariant flow to a non-equivariant flow where $v(x(t), t)$ is given by a simple fully-connected neural network. As brute-force computation of

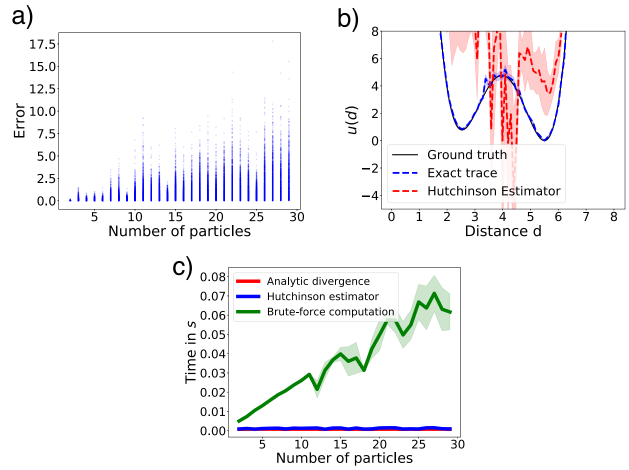

Figure 3. **a)** Error of divergence estimates introduced by the Hutchinson estimator with growing number of particles. **b)** Free-energy profile of a DW-2 potential and importance-weighed estimates. **c)** Wall-clock time of evaluating $v(x(t), t)$ with growing number of particles.

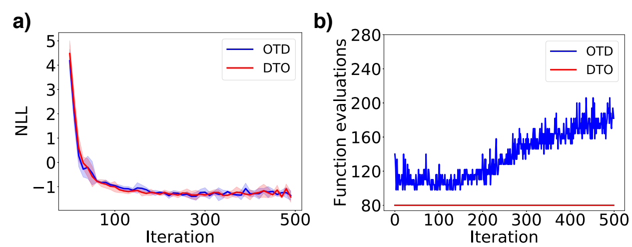

Figure 4. **a)** Log-likelihood on test data after training with DTO/OTD for the DW-4 system **b)** Number of function evaluations increase significantly during training using the OTD approach. The curve for DTO remains flat.

the divergence quickly becomes prohibitively slow for the LJ-13 system, we rely on Hutchinson-estimation during training and compute the exact divergence only during evaluation.

The training data is generated by taking 10 / 100 / $1,000$ / $10,000$ samples from a long MCMC trajectory (throwing away $1,000$ burn-in samples to enforce equilibration). After training we evaluate the likelihood of the model on an independent 10,000 trajectory. We train both flows using Adam with weight decay (Kingma & Ba, 2014; Loshchilov & Hutter) until convergence. For the non-equivariant flow we tested both: data augmentation by applying random rotations and permutations, and no data augmentation.

Our results show that an equivariant flow generalizes well to the unseen trajectory even in the low data regime. When applying data augmentation, the non-equivariant flow significantly performs worse (DW-4) or even fails to fit the data at all and remains close to the prior distribution (LJ-13). Without data augmentation yet using strong regularization we observe strong over-fitting behavior: the DW-4 system can only be fitted if trained on amount of data that is close to the full equilibrium distribution, the LJ-13 system cannot be fitted sufficiently at all (Figure 5). It is worth to remark that the equivariant flow only requires 620 trainable parameters in order to achieve this result compared to the 5256 (DW-4) / 21671 (LJ-13) parameters of the black-box model.

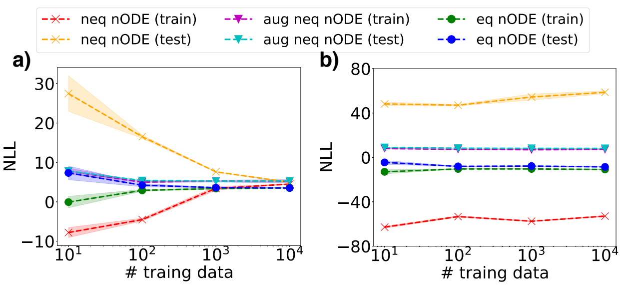

Figure 5. Log-likelihood on train and test data for both **a)** the DW-4 and **b)** the LJ-13 system after training on an increasing number of data points. *eq nODE*: proposed equivariant flow, *neq nODE*: non-equivariant baseline without data augmentation, *aug neq nODE*: non-equivariant baseline with data augmentation.

7.4. Equivariance in Boltzmann-Generating Flows

In a fourth experiment we compare how equivariance affects normalizing flows when used in the context of Boltzman generators (see section 2 or (Noé et al., 2019)). For the DW-4 system we compare our equivariant flow to a non-equivariant one when being trained using both: maximizing likelihood on data and minimizing reverse KL-divergence w.r.t. the target density (for details see Suppl. Material). For the non-equivariant flow we tested both: data augmentation and no data augmentation.

The equivariant flow achieves a significant overlap with the target distribution. This allows the target energies to be reweighed to the ground-truth distribution (see Figure 6 c) and thus to draw asymptotically unbiased samples. The non-equivariant flow without data augmentation quickly samples low-energy states. However, as indicated by the reweighted distribution and the high train and test likelihood (10.85 and 11.40 respectively), this is due to collapsing to one mode of the distribution (see Figure 6 a). As a result asymptotically unbiased sampling will not be possible. The non-equivariant flow after being trained with data augmentation falls short in both: producing accurate low energy states and thus reweighing to the ground-truth (see Figure 6 b).

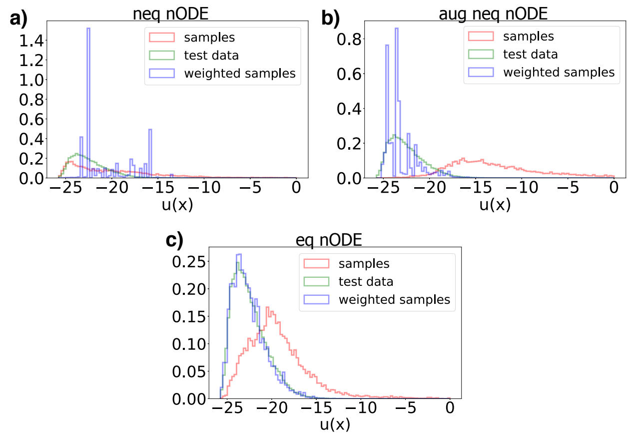

Figure 6. Energy histograms for samples from the DW-4 system with different models. **a)** non equivariant nODE, **b)** non equivariant nODE with data augmentation, and **c)** proposed equivariant flow.

7.5. Discovery of Meta-Stable States

In our final experiment, we evaluate to which extend these models help discovering new meta-stable states, which have not been observed in the training data set. Here we characterize metastable states as the set of configurations x that minimize to the same local minimum on the energy surface. Finding new meta-stable states is especially non-trivial for LJ systems with many particles.

Counting Distinct Meta-Stable States Let ψ be the function mapping a state x onto its next meta-stable state $\psi(x)$. We implement it by minimizing x w.r.t. $u(x)$ using a non-momentum optimizer until convergence and filtering out saddle-points. Then we equate two minima $\psi(x) \sim \psi(x')$, whenever they are identical up to rotations and permutations. To avoid computing the orthogonal Procrustes problem between all minimized structures, we compute the all-distance matrix $M_d(\psi(x))$ of each minimum state, sort it in ascending order to obtain $M_{d,\text{sorted}}(\psi(x))$ and equate two structures $\psi(x) \sim_{\text{approx}} \psi(x')$, whenever $\|M_{d,\text{sorted}}(x) - M_{d,\text{sorted}}(x')\| < \epsilon$, where $\epsilon \ll 1$ is a threshold depending on the system. This ensures that $\psi(x) \sim \psi(x') \implies \psi(x) \sim_{\text{approx}} \psi(x')$, however the inverse direction might not hold. Thus, reported numbers on the count of unique minima found remain a lower bound.

DW-4 For this system, we can fully enumerate those five meta-stable minima between which the system jumps in equilibrium. We train both an equivariant flow and a non-equivariant flow on a single minimum state perturbed by a tiny amount of Gaussian noise until convergence. Then we sample $10,000$ structures from both models and compute the set of unique minima. While the non-equivariant flow model can only reproduce the minimum state it has been trained on, the equivariant flow discovers all minimum states of the system (see Figure 7 a).

LJ-13 Finding meta-stable minima with low energies is a much more challenging task for the LJ system. Here we compare the proposed equivariant flow to standard sampling by (1) training on a short equilibrium MCMC trajectory consisting of $1,000$ samples, (2) sampling $1,000$ samples from the generator distribution after training, and (3) counting the amount of unique minima states found according to the procedure described above. The amount of unique minima found is compared to sampling an independent equilibrium MCMC trajectory having the same amount of samples as the training set and a long trajectory with $100,000$ samples.

As can be seen from Table 1 the equivariant flow model clearly outperforms naive sampling in finding low-energy meta-stable states compared to the short MCMC trajectory which had access to the same amount of target energy evaluations. Furthermore, in contrast to the latter, it consistently finds the global minimum state, which has not been present in the training trajectory. It performs closely as good as the long trajectory which had access to 100x more evaluations of the target function. Figure 7 b shows structures of low-energy minima generated by the equivariant flow.

Table 1. Count of unique minima states discovered: displayed are means and standard deviation over 10 independent rounds.

		$u(x)$	
METHOD	$(-70, -60)$	$(-80, -70)$	$(-\infty, -80)$
TRAINING	0	7	0
SHORT	2.70 ± 3.80	7.70 ± 3.23	0.90 ± 0.30
LONG	64.60 ± 6.11	48.60 ± 4.13	1.00 ± 0.00
EQ-FLOW	38.30 ± 2.49	41.50 ± 2.50	1.00 ± 0.00

8. Discussion

We presented a construction principle to incorporate symmetries of densities defined over $\mathbb{R}^n$ into the structure of normalizing flows. We further demonstrated the superior generalization capabilities of such symmetry-preserving flows compared to non-symmetry-preserving ones on two physics-motivated particle systems, which are difficult to sample with classic methods. Our proposed equivariant gradient field utilizing a simple mixture potential has several structural advantages over black box CNFs, such as an analytically computable divergence, explicit handling of numerical stability and very few parameters.

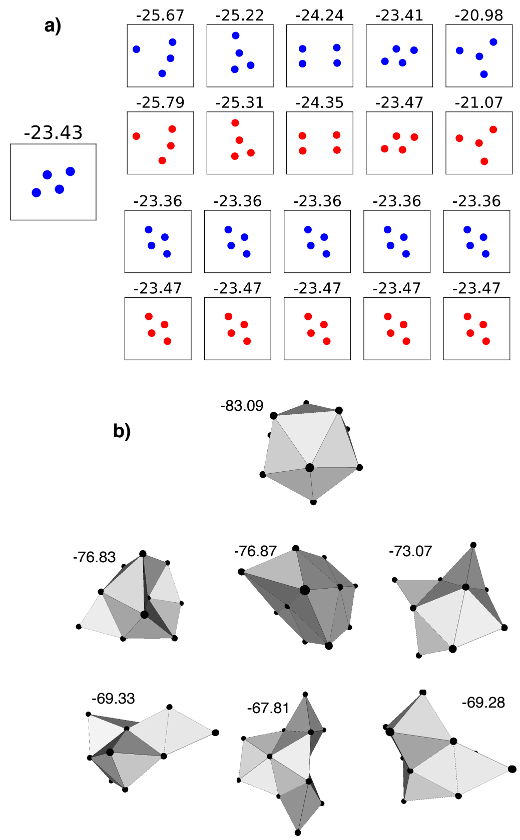

Figure 7. **a)** On the left: Minimum state used for training in the DW-4 system. *Upper rows*: samples from equivariant flow (blue) and corresponding minimum states (red). *Bottom rows*: samples from non-equivariant flow (blue) and corresponding minimum states (red)., **b)** Exemplary unique minima states from the LJ-13 system generated within the three given energy intervals. The top state marks the global minimum, which consists of a perfect icosahedron with one particle in the center.

Acknowledgements

We acknowledge funding from European Commission (ERC CoG 772230), Deutsche Forschungsgemeinschaft (SFB1114/A04, DAEDALUS/P04), MATH+ The Berlin Mathematics Research Center (AA1-6, EF1-2).

We further thank Moritz Hoffmann (FU Berlin), Pim de Haan (UvA Amsterdam / Qualcomm) and Onur Caylak (TU Eindhoven) for helpful remarks and discussions.

References

Albergo, M. S., Kanwar, G., and Shanahan, P. E. Flow-based generative models for markov chain monte carlo in lattice field theory. *Phys. Rev. D*, 100:034515, 2019.

Behrmann, J., Grathwohl, W., Chen, R. T., Duvenaud, D., and Jacobsen, J.-H. Invertible residual networks. *arXiv preprint arXiv:1811.00995*, 2018.

Benamou, J.-D. and Brenier, Y. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Bennett, C. H. Efficient estimation of free energy differences from monte carlo data. *J. Chem. Phys.*, 22:245–268, 1976.

Berg, R. v. d., Hasenclever, L., Tomczak, J. M., and Welling, M. Sylvester normalizing flows for variational inference. *arXiv preprint arXiv:1803.05649*, 2018.

Chen, C., Li, C., Chen, L., Wang, W., Pu, Y., and Carin, L. Continuous-time flows for efficient inference and density estimation. *arXiv preprint arXiv:1709.01179*, 2017.

Chen, T. Q. and Duvenaud, D. K. Neural networks with cheap differential operators. In *Advances in Neural Information Processing Systems*, pp. 9961–9971, 2019.

Chen, T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In *Advances in neural information processing systems*, pp. 6571–6583, 2018.

Chen, T. Q., Behrmann, J., Duvenaud, D. K., and Jacobsen, J.-H. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems*, pp. 9913–9923, 2019.

De Cao, N., Titov, I., and Aziz, W. Block neural autoregressive flow. *arXiv preprint arXiv:1904.04676*, 2019.

Dinh, L., Krueger, D., and Bengio, Y. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Dormand, J. R. and Prince, P. J. A family of embedded runge-kutta formulae. *Journal of computational and applied mathematics*, 6(1):19–26, 1980.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. Neural spline flows. In *Advances in Neural Information Processing Systems*, pp. 7509–7520, 2019.

Germain, M., Gregor, K., Murray, I., and Larochelle, H. Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pp. 881–889, 2015.

Gholami, A., Keutzer, K., and Biros, G. Anode: Unconditionally accurate memory-efficient gradients for neural odes. *arXiv preprint arXiv:1902.10298*, 2019.

Grathwohl, W., Chen, R. T., Bettencourt, J., Sutskever, I., and Duvenaud, D. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. Neural autoregressive flows. *arXiv preprint arXiv:1804.00779*, 2018.

Hutchinson, M. F. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. 1989.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pp. 10215–10224, 2018.

Köhler, J., Klein, L., and Noé, F. Equivariant flows: sampling configurations for multi-body systems with symmetric energies. *arXiv preprint arXiv:1910.00753*, 2019.

Laio, A. and Parrinello, M. Escaping free energy minima. *Proc. Natl. Acad. Sci. USA*, 99:12562–12566, 2002.

Li, S.-H. and Wang, L. Neural network renormalization group. *Phys. Rev. Lett.*, 121:260601, 2018.

Liu, J., Kumar, A., Ba, J., Kiros, J., and Swersky, K. Graph normalizing flows. In *Advances in Neural Information Processing Systems*, pp. 13556–13566, 2019a.

Liu, X., Qi, C. R., and Guibas, L. J. Flownet3d: Learning scene flow in 3d point clouds. *arXiv:1806.01411*, 2019b.

Loshchilov, I. and Hutter, F. Fixing weight decay regularization in adam, corr, abs/1711.05101. In *Proceedings of the ICLR 2018 Conference Blind Submission, Vancouver, BC, Canada*, volume 30.

Loshchilov, I. and Hutter, F. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 2017.

Louizos, C. and Welling, M. Multiplicative normalizing flows for variational bayesian neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2218–2227. JMLR. org, 2017.

McCann, R. J. Polar factorization of maps on riemannian manifolds. *Geometric & Functional Analysis GAFA*, 11 (3):589–608, 2001.

Müller, T., McWilliams, B., Rousselle, F., Gross, M., and Novák, J. Neural importance sampling. *arXiv preprint arXiv:1808.03856*, 2018.

Nicoli, K. A., Nakajima, S., Strodthoff, N., jciech Samek, W., Müller, K.-R., and Kessel, P. Asymptotically unbiased estimation of physical observables with neural samplers. *arXiv:1910.13496*, 2019.

Noé, F., Olsson, S., Köhler, J., and Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

Oord, A. v. d., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G. v. d., Lockhart, E., Cobo, L. C., Stimberg, F., et al. Parallel wavenet: Fast high-fidelity speech synthesis. *arXiv preprint arXiv:1711.10433*, 2017.

Papamakarios, G., Pavlakou, T., and Murray, I. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pp. 2338–2347, 2017.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*, 2019.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

Rezende, D. J., Racanière, S., Higgins, I., and Toth, P. Equivariant hamiltonian flows. *arXiv preprint arXiv:1909.13739*, 2019.

Schütt, K., Kindermans, P.-J., Felix, H. E. S., Chmiela, S., Tkatchenko, A., and Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, pp. 991–1001, 2017.

Tabak, E. G. and Turner, C. V. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

Tabak, E. G., Vanden-Eijnden, E., et al. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

Tomczak, J. M. and Welling, M. Improving variational auto-encoders using householder flow. *arXiv preprint arXiv:1611.09630*, 2016.

Torrie, G. M. and Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comp. Phys.*, 23:187–199, 1977.

Toth, P., Rezende, D. J., Jaegle, A., Racanière, S., Botev, A., and Higgins, I. Hamiltonian generative networks. *arXiv preprint arXiv:1909.13789*, 2019.

Wales, D. J. and Doye, J. P. K. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 100 atoms. *J. Phys. Chem. A*, 101:5111–5116, 1997.

Wu, H., Paul, F., Wehmeyer, C., and Noé, F. Multiensemble markov models of molecular thermodynamics and kinetics. *Proc. Natl. Acad. Sci. USA*, 113:E3221–E3230, 2016. doi: 10.1073/pnas.1525092113.

Zhang, L., Wang, L., et al. Monge-ampere flow for generative modeling. *arXiv preprint arXiv:1809.10188*, 2018.