Neural Thermodynamic Integration: Free Energies from Energy-based Diffusion Models

Bálint Máté [1 2 3] **François Fleuret** [2] **Tristan Bereau** [1]

Abstract

Thermodynamic integration (TI) offers a rigorous method for estimating free-energy differences by integrating over a sequence of interpolating conformational ensembles. However, TI calculations are computationally expensive and typically limited to coupling a small number of degrees of freedom due to the need to sample numerous intermediate ensembles with sufficient conformational-space overlap. In this work, we propose to perform TI along an alchemical pathway represented by a trainable neural network, which we term Neural TI. Critically, we parametrize a time-dependent Hamiltonian interpolating between the interacting and non-interacting systems, and optimize its gradient using a denoising-diffusion objective. The ability of the resulting energy-based diffusion model to sample all intermediate ensembles allows us to perform TI from a single reference calculation. We apply our method to Lennard-Jones fluids, where we report accurate calculations of the excess chemical potential, demonstrating that Neural TI is capable of coupling hundreds of degrees of freedom at once.

1. Introduction

Accurate estimation of free-energy differences is pivotal in numerous scientific disciplines, including chemistry, biology, and materials science. These estimations are essential for understanding molecular interactions, reaction mechanisms, and phase transitions (Gao et al., 2006; Mobley & Gilson, 2017; Agarwal et al., 2021). The main methodologies to estimate free-energy differences are rooted in statistical mechanics including free-energy perturbation and thermodynamic integration (TI). They estimate the ratio of partition functions of the two ensembles, typically sampling the two conformational ensembles via Monte Carlo or molecular dynamics simulations. Because of typically poor overlap between distributions, intermediate simulations are required to help interpolate the two end points. The exact pathway connecting the two ensembles can be chosen freely, because the free energy is a state function. The chosen interpolation is often unphysical, resulting in a so-called *alchemical transformation*. The necessity to sample intermediate Hamiltonians leads to significant computational expense (Chipot & Pohorille, 2007; Mey et al., 2020).

Machine learning—in particular generative models—is rapidly transforming how we model molecular systems (Sanchez-Lengeling & Aspuru-Guzik, 2018; Hermann et al., 2020; Noé et al., 2020; Fedik et al., 2022). Recently, normalizing flows have been proposed for sampling in the context of statistical physics (Nicoli et al., 2020), quantum field theory (Albergo et al., 2019), and molecules (Noé et al., 2019). Flow-based approaches are attractive as they represent exact probability densities that can be used for unbiased estimation of observables when combined with importance sampling to correct for the mismatch between the learnt and target densities. Flow-based methods have been proposed as a means to compute free-energy differences (Noé et al., 2019; Wirnsberger et al., 2020). However, flows also present challenges: coupling flows (Dinh et al., 2016) are cumbersome to work with when one needs to encode physical bias in the architecture, and continuous flows (Chen et al., 2018) suffer from the cost of computing the divergence of the model.

In this work, we instead propose to use denoising diffusion models (DDMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020) to estimate free-energy differences. To this end, we assign our data and latent distributions to the *interacting* and *non-interacting* Hamiltonians, respectively. We then train a time-dependent, interpolating potential U_t^θ between the potentials of the target, U_0, and the prior, U_1, along the finite-time interval of the DDM. We match the force exerted

[*]Equal contribution [1]Institute for Theoretical Physics, Heidelberg University, Heidelberg, Germany [2]Department of Computer Science, University of Geneva, Carouge, Switzerland [3]Department of Phyiscs, University of Geneva, Geneva, Switzerland. Correspondence to: Bálint Máté <balint.mate@unige.ch>.

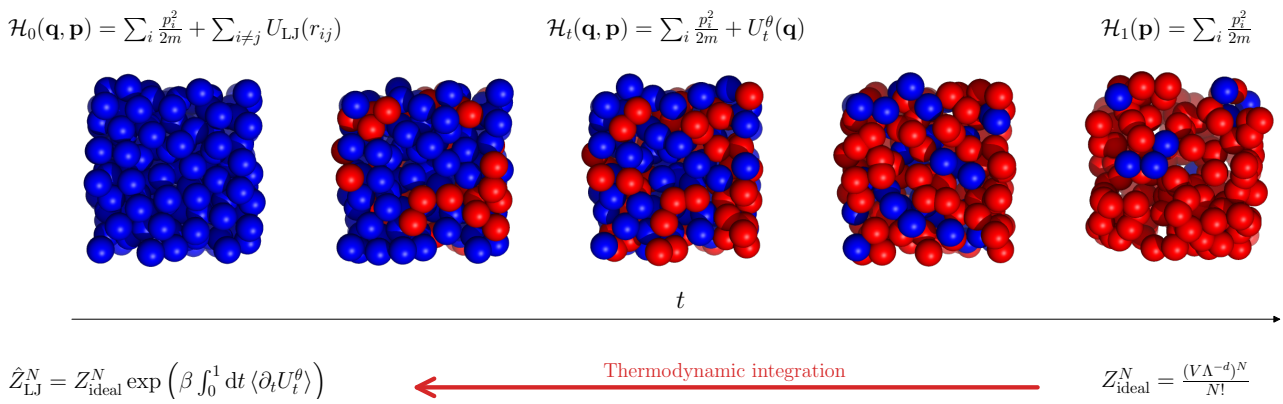

Figure 1. Schematic summary of the proposed approach. We interpolate between the target, $\mathcal{H}_0$, and latent, $\mathcal{H}_1$, Hamiltonians with a time-dependent potential U_t^θ. During sampling, the normalizing constant of the target can be estimated via thermodynamic integration. Particles whose separation from their closest neighbor is less than 0.85σ are colored red, the rest are colored blue. This color-coding illustrates that in the ideal gas (right) there are many colliding particles and as the LJ potential is turned on (left) the particles do not overlap anymore.

by the potential to the Hyvärinen score of the DDM

$$s(x,t) = \nabla_x \log \rho_t = -\beta \nabla_x U_t^\theta.$$

This parametrization of the score, while limiting its expressivity, will easily allow us to compute time derivatives of the energy—a critical component to perform TI along the diffusion process. The TI calculation efficiently estimates the ratio of partition function of turning on the interactions of the Hamiltonian. However, because we can analytically calculate the partition function of the non-interacting system as well, our methodology yields the partition function of the target Hamiltonian. We call our method Neural TI.

We demonstrate our methodology on computer simulations of a condensed-phase, many-body system: the Lennard-Jones (LJ) liquid. The particles are confined to a box with periodic boundary conditions, leading to topological constraints on the DDM addressed further below. The latent space consists of the ideal-gas system: particles that carry kinetic energy, but do not interact. The DDM thereby interpolates between interacting and non-interacting system, whose trajectories model the dynamics of a time-dependent force-field $-\nabla U_t^\theta$. See Figure 1 for a schematic summary of the proposed approach. We validate Neural TI on accurate calculations of the excess chemical potential. For the first time, we report accurate free-energy differences of coupling an entire liquid, totalling 600 degrees of freedom.

2. Thermodynamic integration

Suppose now that U_λ is a one-parameter family of potentials, and $Z_\lambda = \int \mathrm{d}x \, e^{-\beta U_\lambda(x)} = e^{-\beta F}$ are the corresponding normalizing constants at inverse temperature $\beta = (k_\mathrm{B}T)^{-1}$. Given samples from all $\rho_\lambda = \mathrm{e}^{-\beta U_\lambda(x)}/Z_\lambda$ the free energy difference $\Delta F_{0\to 1}$ between $\lambda = 0$ and $\lambda = 1$ can be written as

$$\begin{aligned}
\beta \Delta F_{0\to 1} &= \log Z_0 - \log Z_1 && (1)\\
&= -\int_0^1 \mathrm{d}\lambda \, \partial_\lambda \log Z_\lambda && (2)\\
&= -\int_0^1 \mathrm{d}\lambda \, \frac{1}{Z_\lambda} \partial_\lambda \left(\int \mathrm{d}x \, \mathrm{e}^{-\beta U_\lambda(x)} \right) && (3)\\
&= \beta \int_0^1 \mathrm{d}\lambda \, \frac{1}{Z_\lambda} \left(\int \mathrm{d}x \, \mathrm{e}^{-\beta U_\lambda(x)} \partial_\lambda U_\lambda(x) \right) && (4)\\
&= \beta \int_0^1 \mathrm{d}\lambda \, \langle \partial_\lambda U_\lambda \rangle_\lambda \,, && (5)
\end{aligned}$$

where $\langle \partial_\lambda U_\lambda \rangle_\lambda$ denotes the expected value of $\partial_\lambda U_\lambda$ under the density $\rho_\lambda = Z_\lambda^{-1} \mathrm{e}^{-\beta U_\lambda}$. Practically, the small phase-space overlap between ensembles requires an interpolation of many intermediate Hamiltonians parametrized by the coupling variable λ (Mey et al., 2020). Here instead we will use a DDM (§3) to learn the alchemical pathway, i.e., we replace λ by the diffusion model's time variable, t. The generative model's ability to sample all intermediate ensembles will do away with the need for intermediate reference simulations, and provide an accurate estimate of much larger free-energy differences than reported so far (Straatsma & Berendsen, 1988).

3. Denoising Diffusion models

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) are a class of generative models defined by a pair (forward and reverse) of stochastic processes. In the continuous-time formulation (Song et al., 2020), the forward

process is given by the stochastic differential equation (SDE)

$$dX = f_t X \mathrm{d}t^{\rightarrow} + g_t \mathrm{d}W_{t\rightarrow}, \tag{6}$$

where $\mathrm{d}W_{t\rightarrow}$ is a Wiener process and SDE starts from the initial condition $X_0 \sim \rho_0$. Note that $\rho_t(x_t|x_0)$ is a Gaussian for all t and x_0. In particular f_t and g_t are time-dependent and are chosen such that $\rho_t(x_1|x_0)$ is close to a standard Gaussian distribution for all x_0. The noising process then corresponds to simulating Eq. (6) from $t = 0$ to $t = 1$. The same time marginals can be reproduced by a time-reversed SDE (Anderson, 1982) from $t = 1$ to $t = 0$,

$$dY = \left[f_t Y + g_t^2 \nabla_x \log \rho_t\right] \mathrm{d}t^{\leftarrow} + g_t \mathrm{d}W_t^{\leftarrow}, \tag{7}$$

with initial condition $Y_0 \sim \mathcal{N}(0, \mathbb{I})$. Given the score $s(x,t) = \nabla_x \log \rho_t(x)$, one could use the reverse Eq. (7) to map Gaussian samples to the initial target density ρ_0.

Physically speaking, both the forward and reverse processes correspond to an overdamped Langevin dynamics driven by the time-dependent potentials $\frac{1}{2} f_t ||x||^2$ and $\frac{-1}{2} f_t ||y||^2 - g_t^2 \log \rho_t$, respectively. We can thus interpret the score, $\nabla \log \rho_t$, as a force induced by the potential $(-\log \rho_t)$ (Arts et al., 2023).

Score estimation on $\mathbb{R}^d$ To estimate the score $s(x,t) = \nabla_x \log \rho_t$, one integrates over the conditional scores

$$\nabla \log \rho_t(x) = \mathbb{E}_{x_0 \sim \rho_0, x_t \sim p(x_t|x_0)} \left[\nabla \log \rho_t(x_t|x_0)\right]. \tag{8}$$

Denoting the mean and variance of $\log \rho(x_t|x_0)$ by $\gamma_t x_0$ and σ_t^2, we can rewrite the integrand as

$$\nabla \log \rho_t(x_t|x_0) = -\frac{x_t - \gamma_t x_0}{\sigma_t^2} = -\frac{\epsilon}{\sigma_t}. \tag{9}$$

Instead of learning the score directly, it is customary to train a neural network $\epsilon_\theta(x,t)$ to predict the noise term $\epsilon = \frac{1}{\sigma_t}(x_t - \gamma_t x_0)$ instead. The score is then recovered as $s_\theta(x,t) = -\epsilon_\theta(x,t)/\sigma_t$.

3.1. TI with diffusion models

Salimans & Ho (2021) raise the point that although the score $s(x,t)$ is the gradient of the log-density $\log \rho_t(x)$, it is usually modelled with a free-form neural network that does not necessarily learn a conservative vector field. Our proposal is to approximate the score as the force of a time-dependent, parametric potential U_t^θ,

$$s(x,t) = \nabla_x \log \rho_t = -\beta \nabla_x U_t^\theta.$$

This in turn means that U_t^θ itself serves as an approximation of the negative log-likelihood up to an additive constant. Since U_t^θ is a neural network, its time-derivative can be computed with automatic differentiation and the ensemble average $\langle \partial_t U_t^\theta \rangle_t$ can be estimated from samples either from the forward or from the learned reverse process.

3.2. Diffusion models on tori

In what follows we will be interested in particles living in a d-dimensional box with periodic boundary conditions. To work with such systems the usual framework of diffusion models needs to be slightly adjusted to accommodate for the different topology of the configurational space. Topologically speaking, the position of a single particle is specified by a point on a hypertorus, $\mathbb{T}^N$, of dimension d, and the configurational space of N particles is then a dN-dimensional hypertorus, $\mathbb{T}^{dN}$.

Although there are works generalizing diffusion models to non-Euclidean geometries (De Bortoli et al., 2022; Huang et al., 2022), for our case it is sufficient to derive a model for the simplest manifold with non-trivial topology, the circle $\mathbb{S}^1 = \mathbb{R}/\mathbb{Z}$. To do this, we set the forward process to be an unbiased random walk on $\mathbb{S}^1$ converging to the uniform distribution. Explicitly, the forward and backward processes take the following form,

$$dX = g_t \mathrm{d}W_{t\rightarrow} \qquad X_0 \sim \rho_0 \tag{10}$$

$$dY = g_t^2 \nabla \log \rho_t \mathrm{d}t^{\leftarrow} + g_t \mathrm{d}W_{t\leftarrow} \qquad Y_0 \sim U(\mathbb{S}^1). \tag{11}$$

Score estimation on $\mathbb{S}^1$ Note that on the circle the time-marginals $\rho_t(x_t|x_0)$ are wrapped Gaussians,

$$\rho_t(x_t|x_0) = \sum_{k \in \mathbb{Z}} \mathcal{N}(x_t + k; x_0, \sigma_t^2). \tag{12}$$

Jing et al. (2022) compute the score of the wrapped Gaussian in Eq. (12) by truncating both the numerator and denominator of the logarithmic derivative $\nabla \log \rho_t(x_t|x_0) = \frac{\nabla \rho_t(x_t|x_0)}{\rho_t(x_t|x_0)}$. Alternatively, one could write the score at x_t of the wrapped Gaussian as a weighted average of the scores of the unwrapped Gaussian over the fiber above x_t

$$\nabla \log \rho_t(x_t|x_0) = \frac{\nabla \rho_t(x_t|x_0)}{\rho_t(x_t|x_0)} \tag{13}$$

$$= \frac{\sum_{k \in \mathbb{Z}} \nabla \mathcal{N}_k}{\sum_{k \in \mathbb{Z}} \mathcal{N}_k} \tag{14}$$

$$= \frac{\sum_{k \in \mathbb{Z}} \mathcal{N}_k \frac{\nabla \mathcal{N}_k}{\mathcal{N}_k}}{\sum_{k \in \mathbb{Z}} \mathcal{N}_k} \tag{15}$$

$$= \frac{\sum_{k \in \mathbb{Z}} \mathcal{N}_k \nabla \log \mathcal{N}_k}{\sum_{k \in \mathbb{Z}} \mathcal{N}_k}, \tag{16}$$

where $\mathcal{N}_k = \mathcal{N}(x_t + k; x_0, \sigma_t^2)$ are the evaluations of the unwrapped Gaussian on the fiber and $\sigma_t^2 = \int_0^t \mathrm{d}\tau\, g_\tau^2$. This in particular means that the noise prediction model ϵ_θ can be trained by sampling $x_t = [(x_0 + \sigma_t \epsilon) \mod 1]$ where $\epsilon \sim \mathcal{N}(0,1)$ and taking a gradient step on $||\epsilon_\theta(x_t,t) - \epsilon||^2$.

4. Statistical ensembles

For the rest of the paper we exclusively consider systems of indistinguishable particles confined to a d-dimensional box of volume V with periodic boundary conditions.

4.1. Canonical ensemble (NVT)

If we assume that a system has a fixed number of particles N and is in thermal equilibrium with a reservoir at a fixed inverse temperature $\beta = (k_\mathrm{B}T)^{-1}$, then the likelihood of a particular microstate $(\mathbf{q}, \mathbf{p}) = (q_1, ...q_N, p_1, ..., p_N)$ is

$$\rho(\mathbf{q}, \mathbf{p}) = \frac{1}{Z_N} \frac{1}{h^{dN} N!} \mathrm{e}^{-\beta \mathcal{H}(\mathbf{q}, \mathbf{p})}, \tag{17}$$

where h is Planck's constant, d is the dimensionality of the system, $\mathcal{H}$ is the Hamiltonian and $Z_N = \int \frac{\mathrm{d}\mathbf{p}\mathrm{d}\mathbf{q}}{h^{dN} N!} \mathrm{e}^{-\beta \mathcal{H}(\mathbf{q}, \mathbf{p})}$ is the canonical partition function, i.e., the normalizing constant of the density $\frac{\mathrm{e}^{-\beta \mathcal{H}(\mathbf{q}, \mathbf{p})}}{h^{dN} N!}$.

Ideal and interacting gases In the ideal gas particles do not interact with each other and the system is described by a Hamiltonian only containing a kinetic term $\mathcal{H}_\mathrm{ideal}(\mathbf{p}) = \sum_i \frac{p_i^2}{2m}$. In this case Z_N can be analytically computed

$$Z_N^\mathrm{ideal} = \int \frac{\mathrm{d}\mathbf{p}}{h^{dN} N!} \mathrm{e}^{-\beta \mathcal{H}_\mathrm{ideal}(\mathbf{p})} = \frac{(V \Lambda^{-d})^N}{N!}, \tag{18}$$

where V is the volume of the box and $\Lambda = h/\sqrt{2\pi m \beta^{-1}}$ is the thermal wavelength. Since particle positions in the ideal gas are independently and uniformly distributed over the box, the latent space of the toroidal diffusion model (§3.2) describes the positions $\mathbf{q}$ of an ideal gas in the canonical ensemble.

More generally, a many-body system will also consist of interactions, as described by a potential $U(\mathbf{q})$,

$$\mathcal{H}(\mathbf{q}, \mathbf{p}) = \sum_i \frac{p_i^2}{2m} + U(\mathbf{q}). \tag{19}$$

The separation between the positions and momenta offers a factorization of the partition function

$$Z_N = Z_N^\mathrm{ideal} \int \mathrm{d}\mathbf{q}\, \mathrm{e}^{-\beta U(\mathbf{q})}. \tag{20}$$

Estimating the partition function Suppose now that we have trained a DDM on the positions $\mathbf{q}$ with its score parametrized as the force of a time-dependent potential, $s_\theta(\mathbf{q}, t) = -\beta \nabla U_t^\theta(\mathbf{q})$, between the target and non-interacting Hamiltonians, i.e., $U_0^\theta(\mathbf{q}) = U(\mathbf{q})$ and $U_1^\theta(\mathbf{q}) = U_\mathrm{ideal}(\mathbf{q}) \equiv 0$. As the toroidal DDM maps configurations of the ideal gas to the interacting system, the setup is adequate to perform TI over the coupling of interactions. In addition, because Z_N^ideal is known analytically, we obtain an estimate of the partition function of the full, target Hamiltonian

$$\hat{Z}_N = Z_N^\mathrm{ideal} \exp\left(\beta \int_0^1 \mathrm{d}t \, \langle \partial_t U_t^\theta \rangle_t\right). \tag{21}$$

4.2. Grand Canonical ensemble (μVT)

Let us now assume that a system is both in thermal and chemical equilibrium with the reservoir at fixed temperature T and chemical potential μ. The likelihood of a microstate $(N, \mathbf{q}, \mathbf{p}) = (N, q_1, ...q_N, p_1, ..., p_N)$ is

$$\rho(N, \mathbf{q}, \mathbf{p}) = \frac{1}{\mathcal{Z}} \frac{1}{h^{dN} N!} \mathrm{e}^{\beta(\mu N - \mathcal{H}(\mathbf{q}, \mathbf{p}))}, \tag{22}$$

where $\mathcal{Z} = \sum_N \int \frac{\mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}}{h^{dN} N!} \mathrm{e}^{\beta(\mu N - \mathcal{H}(\mathbf{q}, \mathbf{p}))}$ is the grand canonical partition function. The grand canonical partition function is a weighted sum of the canonical partition functions according to the chemical potential,

$$\mathcal{Z}(\mu) = \sum_N \mathrm{e}^{\beta \mu N} Z_N. \tag{23}$$

The excess chemical potential The choice of a Hamiltonian $\mathcal{H}$ and chemical potential μ defines a marginal distribution, $p(N)$, of the number of particles present in the system

$$p(N) = \frac{1}{\mathcal{Z}} \mathrm{e}^{\beta \mu N} Z_N. \tag{24}$$

In the case of the ideal gas, $p(N) = \frac{1}{\mathcal{Z}} \frac{(\mathrm{e}^{\beta\mu} V \Lambda^{-d})^N}{N!}$ is Poisson distributed with expected value

$$\langle N \rangle(\mathcal{H}^\mathrm{ideal}, \mu) = \mathrm{e}^{\beta \mu} V \Lambda^{-d}. \tag{25}$$

For interacting systems, the expected number of particles $\langle N \rangle(\mathcal{H}, \mu)$ can be used to decompose the chemical potential as $\mu = \mu_\mathrm{ideal} + \mu_\mathrm{ex}$, where μ_ideal is the chemical potential of the ideal gas of the same density

$$\mu_\mathrm{ideal} = \beta^{-1} \log \frac{\langle N \rangle \Lambda^d}{V}. \tag{26}$$

Intuitively, the excess chemical potential $\mu_\mathrm{ex} = \mu - \mu_\mathrm{ideal}$ measures the deviation in chemical potential due to the interaction term of the Hamiltonian at fixed density.

The straightforward way of estimating the excess chemical potential of an interacting system is to perform a grand-canonical Monte Carlo (GCMC) simulation on the distribution in Eq. (22) at a prescribed value of μ. Grand-canonical sampling of the number of particles allows us to estimate the empirical mean $\langle N \rangle_\mathrm{GCMC}$, from which we extract μ_ideal via Eq. (26). From μ and μ_ideal, one can simply compute the excess chemical potential, $\mu_\mathrm{ex} = \mu - \mu_\mathrm{ideal}$.

In this work we estimate the excess chemical potential from several canonical simulations. We train a generative model on several canonical ensembles and then estimate the canonical partition function Z_N for each N. The collection $\{Z_N\}$ is used to compute the distribution of the number of particles, $p(N)$, see Eq. (24), for a predefined choice of the chemical potential, μ. In combination with Eq. (26), we compute the excess chemical potential, μ_{ex}.

5. Application: Lennard-Jones fluid

We consider a crowded condensed-matter system: a collection of three-dimensional particles confined in a box, interacting via a pairwise Lennard-Jones (LJ) potential

$$U_{\text{LJ}}(\mathbf{q}) = \sum_{i \neq j} 4\varepsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right], \tag{27}$$

where r_{ij} represents the inter-particle distance between particles i and j.

To demonstrate the accuracy of our DDM-based Neural TI scheme, we perform the following

1. Generate training configurations via canonical Monte Carlo simulations at different particle numbers $N_1, \ldots, N_i$;
2. Train a *single* diffusion model on the positions $\mathbf{q}$ of the training data, making the model transferable between different N_i. This transferability property is crucial since we would like to estimate Z_N also at those values of N that did not belong to the training set;
3. Estimate the partition function, Z_N, by generating configurations at a given N from the DDM and computing $\hat{Z}_N$ from TI, see Eq. (21);
4. Make grand-canonical estimates from a collection of canonical $\hat{Z}_N$ across values of N, $\{\hat{Z}_{N_1}, \hat{Z}_{N_2}, \ldots\}$, and compare against reference GCMC simulations.

Parametrization of the score To ensure that the boundary conditions at $t \in \{0, 1\}$ of the energy interpolation are met, we follow the interpolation proposed by Máté & Fleuret (2023) and parametrize the score as

$$s(\mathbf{q}, t) = -\beta \nabla \left[t(1-t) U_t^{\theta}(\mathbf{q}) + (1-t) U_t^{\text{LJ}}(\mathbf{q}) \right], \tag{28}$$

where U_t^{LJ} is a soft-core LJ-potential (Beutler et al., 1994) with a time-dependent softening parameter

$$U_t^{\text{LJ}}(\mathbf{q}) = \sum_{i \neq j} 4\varepsilon \left[\left(\frac{\sigma^2}{t\sigma^2 + r_{ij}^2} \right)^{6} - \left(\frac{\sigma^2}{t\sigma^2 + r_{ij}^2} \right)^{3} \right]. \tag{29}$$

This soft-core potential ensures that the numerically unstable region of the LJ-potential is not evaluated on noisy samples but is slowly introduced as more and more noise is removed from the samples (Figure 2). The only trainable component, U_t^{θ} can then be parametrized using ideas from the machine learning force field literature (Schütt et al., 2017; Gasteiger et al., 2019; Batatia et al., 2022; Batzner et al., 2022). Strictly speaking, our model learns a time-dependent potential, and thus the above architectures are only applicable once conditioned on the diffusion time, t.

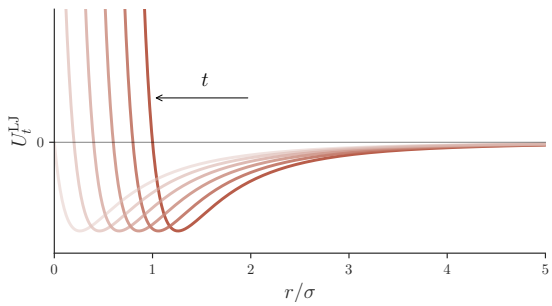

Figure 2. The soft-core LJ potential U_t^{LJ} in Eq. (29) for various values of $t \in [0, 1]$. Note that for larger values of t, particles can get closer to each other without experiencing strong repulsive forces. This is necessary since in the diffusion process it is inevitable that particles get close to each other as t increases.

Results We set the dimension of the system to $d = 3$, the volume of the box to $V = 216$, the inverse temperature to $\beta = 1$, the mass of the particles to $m = 1$, Planck's constant to $h = 10^{-4}$, and the parameters of the LJ-potential to $\varepsilon = 0.8, \sigma = 1$. We train a *single* DDM on samples from canonical Monte Carlo simulations at particle numbers $N \in \{40, 80, 120, 160, 200\}$ (i.e. densities $\rho = N/V \in \{0.19, 0.37, 0.56, 0.74, 0.93\}$). We refer the reader to Appendix B for details on the Monte Carlo simulations and on the architecture.

To evaluate the generative performance of these models we compare the radial distribution function (RDF), $g(r)$, between Monte Carlo samples and samples from the trained DDM (Figure 3). We find accurate reconstructions across a broad range of densities, $\rho = 0.19$ to $\rho = 0.93$. The shapes of the RDF clearly indicate a transition from gas to liquid as we increase particle density.

The accuracy of the Neural TI estimated free energy differences is assessed first in Figure 4. The different panels monitor estimates of the particle-number distribution, $p(N)$, and show significant changes in the distribution as we change the chemical potential. They highlight the transferability of our methodology across the phase transition.

Further, we also evaluate the functional relationship between

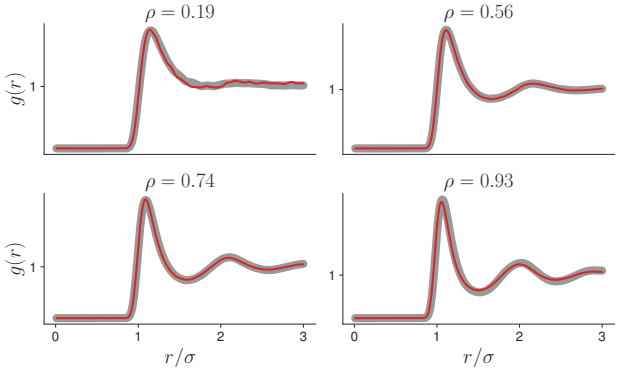

Figure 3. Radial distribution functions as predicted by Monte Carlo simulations (gray) and a diffusion model (red) trained on densities $\rho \in \{0.19, 0.37, 0.56, 0.74, 0.93\}$. Note that the model reconstructs $g(r)$ across the the gas-liquid phase transition. A Gaussian kernel with $\sigma = 0.03$ was applied to all curves in this plot.

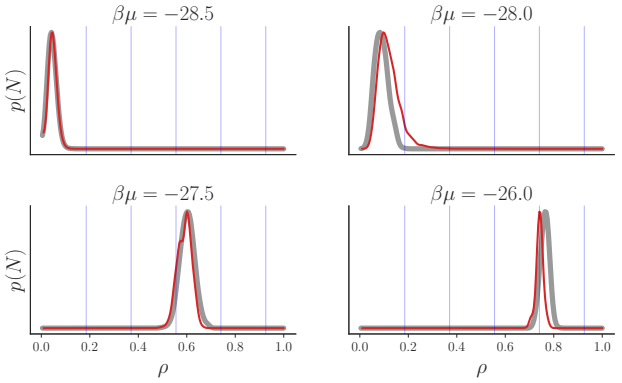

Figure 4. Distribution of the number of particles in the grand canonical ensemble at different chemical potentials from GCMC simulations (gray) and estimated by thermodynamic integration with a trained diffusion model (red). The vertical blue lines denote the canonical ensembles that the diffusion model was trained on. A Gaussian kernel with $\sigma = 0.015$ was applied to all curves in this plot.

ρ and μ in Figure 5. The left subfigure shows the average density as a function of the chemical potential, as calculated by both reference GCMC and our proposed TI methodology. Though the reference calculations are run in the grand-canonical ensemble, we make use of TI from a set of *canonical* simulations to estimate μ. The TI calculations estimate the partition functions, $\hat{Z}_N$, at different values of N. From those, we compute the particle-number distribution via Eq.(24), extract the mean, and deduce the average particle density, $\langle\rho\rangle$. We find excellent agreement across the gas–liquid phase transition—visually illustrated by the jump in density. The right subfigure shows the excess chemical potential as a function of particle density. Our methodology allows us to smoothly interpolate across the phase transition. At high density, the slight underestimation of the particle-number distribution leads a small discrepancy in the excess chemical potential. The main workhorse-method to compute excess chemical potentials from canonical simulations, the Widom particle-insertion method, shows significant difficulties in this regime due to its perturbative nature (Widom, 1963; Frenkel & Smit, 2023).

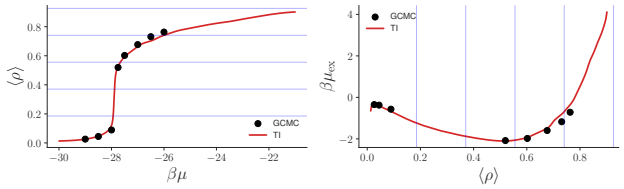

Figure 5. Expected density as a function of the chemical potential (left) and estimates of μ_{ex} as a function of the expected density (right). The left plot suggests a that a gas-liquid phase transition takes place at $\beta\mu \approx -28$. The blue lines denote the canonical ensembles that the diffusion model was trained on.

Neural TI offers unique insight in the equilibrium properties of liquids by directly reporting the free energy of coupling the *entire liquid*. Figure 6 shows the coupling free energy from the ideal gas to the target Hamiltonian of the LJ particle configurations at different densities. On the left, densities close to 0 lead to virtually no free energy due to the ideal-gas limit. As density increases close to $\rho \approx 0.75$, we observe a maximum—interestingly quite a bit higher in density compared to where the excess chemical potential attains its minimum, i.e., $\langle\rho\rangle \approx 0.55$. Instead, we hypothesize that the peak occurs where significant particle-overlap starts severely hampering the configurational space available—this is exactly the regime where perturbative methods, e.g., Widom insertion, become challenging. At a density of 1.0, the available configurational integral reduces significantly—as can be seen by the drop to low $\Delta\hat{F}$ values.

For comparison, we perform the same estimate using standard TI with $5, 10, 20, 50$ and 100 evenly spaced intermediate simulations. The convergence of the standard TI curves with the number of intermediate simulations validates Neural TI as a free-energy estimator. At high densities the slight discrepancy between Neural TI and standard TI with many simulations is likely both due to convergence issues of the reference simulations and limited expressivity of our time-dependent ML force-field architecture.

We finally point at the sheer amplitude of the estimated free-energy differences: up to 200 $k_{\text{B}}T$ in free energies are reported. Traditional TI methodologies require tens of interpolating Hamiltonians, such that each reference simulation contributes on the order of $k_{\text{B}}T$ each. Neural TI's ability to reach two orders of magnitude more demonstrates the appeal to *learn* the alchemical pathway. While traditional alchemical-transformation methods focus on minimizing the change in Hamiltonian to a handful of degrees of freedom, here we report free-energy differences

of coupling an entire box of LJ particles, totaling 600 degrees of freedom.

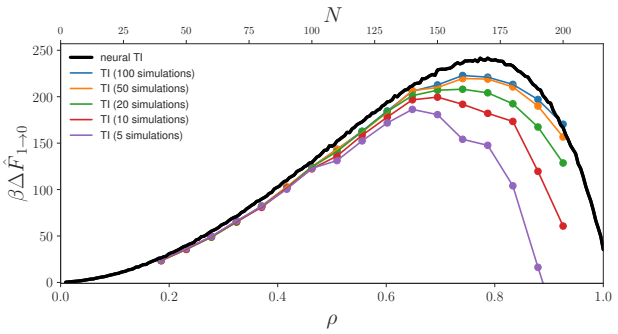

Figure 6. Estimated free energy of turning on the LJ interactions in the canonical ensemble at different densities (number of particles at the top). The subscript $1 \to 0$ corresponds to an interpolation from the non-interacting latent space to the target Hamiltonian, i.e., turning on interactions in the system. Estimate by Neural TI shown in thick black. For comparison, we also display various standard TI estimates with increasing numbers of reference simulations (various colors).

6. Conclusion

We present Neural TI: a generative machine-learning approach to perform thermodynamic integration in molecular systems. Our work shows that energy-based denoising diffusion models (DDMs) are particularly well suited to calculate the free-energy difference of turning on interactions in a many-body Hamiltonian. We find that by associating the latent space to the ideal-gas system, and further parametrizing the score as the derivative of an energy function, we can efficiently and accurately perform TI. Unlike conventional applications, our approach does *not* require reference Monte Carlo or molecular dynamics simulations at intermediate couplings between the end points. Instead, DDMs integrate the conformational ensemble along the finite-time interval of the diffusion process. Critically, we demonstrate for the first time the accurate estimation of free energies of coupling all interactions for up to 600 degrees of freedom.

We demonstrate the applicability of Neural TI on Lennard-Jones fluids. We show our DDM model to transfer across densities around the gas–liquid transition of the system. The structural accuracy of the configurations is illustrated by the radial distribution functions. More importantly, we show that our TI calculations are accurate for varying numbers of particles: the excess chemical potentials and particle-number distributions result from weighted averages of TI-estimated canonical partition functions. We report free-energy differences of coupling Hamiltonians of up to 200 $k_\mathrm{B}T$ from a *single* reference simulation—the fully interacting Hamiltonian alone. We expect DDM-based TI to significantly improve alchemical transformations of molecular systems by coupling many more degrees of freedom than what was possible until now.

Acknowledgements

We thank Fred Hamprecht and Daniel Nagel for fruitful discussions. BM acknowledges financial support by the Swiss National Science Foundation under grant number CR - SII5 - 193716 (RODEM). TB acknowledges support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1 - 390900948 (the Heidelberg STRUCTURES Excellence Cluster).

References

Agarwal, R. G., Coste, S. C., Groff, B. D., Heuer, A. M., Noh, H., Parada, G. A., Wise, C. F., Nichols, E. M., Warren, J. J., and Mayer, J. M. Free energies of proton-coupled electron transfer reagents and their applications. *Chemical reviews*, 122(1):1–49, 2021.

Albergo, M. S., Kanwar, G., and Shanahan, P. E. Flow-based generative models for markov chain monte carlo in lattice field theory. *Physical Review D*, 100(3):034515, 2019.

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. ISSN 0304-4149. doi: https://doi.org/10.1016/0304-4149(82)90051-5. URL https://www.sciencedirect.com/science/article/pii/0304414982900515.

Arts, M., Garcia Satorras, V., Huang, C.-W., Zügner, D., Federici, M., Clementi, C., Noé, F., Pinsler, R., and van den Berg, R. Two for one: Diffusion models and force fields for coarse-grained molecular dynamics. *Journal of Chemical Theory and Computation*, 19(18):6151–6159, 2023.

Batatia, I., Kovacs, D. P., Simm, G., Ortner, C., and Csányi, G. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems*, 35: 11423–11436, 2022.

Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., and Kozinsky, B. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

Beutler, T. C., Mark, A. E., van Schaik, R. C., Gerber, P. R., and van Gunsteren, W. F. Avoiding singularities and numerical instabilities in free energy calculations

based on molecular simulations. *Chemical Physics Letters*, 222(6):529–539, 1994. ISSN 0009-2614. doi: https://doi.org/10.1016/0009-2614(94)00397-1. URL https://www.sciencedirect.com/science/article/pii/0009261494003971.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Chipot, C. and Pohorille, A. *Free energy calculations*, volume 86. Springer, 2007.

De Bortoli, V., Mathieu, E., Hutchinson, M., Thornton, J., Teh, Y. W., and Doucet, A. Riemannian score-based generative modelling. *Advances in Neural Information Processing Systems*, 35:2406–2422, 2022.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. In *International Conference on Learning Representations*, 2016.

Fedik, N., Zubatyuk, R., Kulichenko, M., Lubbers, N., Smith, J. S., Nebgen, B., Messerly, R., Li, Y. W., Boldyrev, A. I., Barros, K., et al. Extending machine learning beyond interatomic potentials for predicting molecular properties. *Nature Reviews Chemistry*, 6(9): 653–672, 2022.

Frenkel, D. and Smit, B. *Understanding molecular simulation: from algorithms to applications*. Elsevier, 2023.

Gao, J., Ma, S., Major, D. T., Nam, K., Pu, J., and Truhlar, D. G. Mechanisms and free energies of enzymatic reactions. *Chemical reviews*, 106(8):3188–3209, 2006.

Gasteiger, J., Groß, J., and Günnemann, S. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2019.

Hermann, J., Schätzle, Z., and Noé, F. Deep-neural-network solution of the electronic schrödinger equation. *Nature Chemistry*, 12(10):891–897, 2020.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Huang, C.-W., Aghajohari, M., Bose, J., Panangaden, P., and Courville, A. C. Riemannian diffusion models. *Advances in Neural Information Processing Systems*, 35: 2750–2761, 2022.

Jing, B., Corso, G., Chang, J., Barzilay, R., and Jaakkola, T. Torsional diffusion for molecular conformer generation. *Advances in Neural Information Processing Systems*, 35: 24240–24253, 2022.

Máté, B. and Fleuret, F. Learning interpolations between boltzmann densities. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=TH6YrEcbth.

Mey, A. S., Allen, B. K., Macdonald, H. E. B., Chodera, J. D., Hahn, D. F., Kuhn, M., Michel, J., Mobley, D. L., Naden, L. N., Prasad, S., et al. Best practices for alchemical free energy calculations [article v1. 0]. *Living journal of computational molecular science*, 2(1), 2020.

Mobley, D. L. and Gilson, M. K. Predicting binding free energies: frontiers and benchmarks. *Annual review of biophysics*, 46:531–558, 2017.

Nicoli, K. A., Nakajima, S., Strodthoff, N., Samek, W., Müller, K.-R., and Kessel, P. Asymptotically unbiased estimation of physical observables with neural samplers. *Physical Review E*, 101(2):023304, 2020.

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.

Noé, F., Tkatchenko, A., Müller, K.-R., and Clementi, C. Machine learning for molecular simulation. *Annual review of physical chemistry*, 71:361–390, 2020.

Salimans, T. and Ho, J. Should EBMs model the energy or the score? In *Energy Based Models Workshop - ICLR 2021*, 2021. URL https://openreview.net/forum?id=9AS-TF2jRNb.

Sanchez-Lengeling, B. and Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400): 360–365, 2018.

Schütt, K., Kindermans, P.-J., Sauceda Felix, H. E., Chmiela, S., Tkatchenko, A., and Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

Straatsma, T. and Berendsen, H. Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations. *The Journal of chemical physics*, 89(9):5876–5886, 1988.

Widom, B. Some topics in the theory of fluids. *The Journal of Chemical Physics*, 39(11):2808–2812, 1963.

Wirnsberger, P., Ballard, A. J., Papamakarios, G., Abercrombie, S., Racanière, S., Pritzel, A., Jimenez Rezende, D., and Blundell, C. Targeted free energy estimation via learned mappings. *The Journal of Chemical Physics*, 153(14), 2020.

A. Experiments on a 1D Lennard-Jones system

In this experiment we work with a $d = 1$-dimensional system. We set the volume (length) of the box to $V = 100$, the inverse temperature $\beta = 1$, the mass of the particles to $m = 1$, Planck's constant to $h = 10^{-4}$, and the parameters of the LJ-potential to $\varepsilon = 0.8, \sigma = 1$.

We train a *single* diffusion model on samples from canonical Monte Carlo simulations at particle numbers $N \in \{50, 70, 90\}$ (i.e. densities $\rho \in \{0.5, 0.7, 0.9\}$). The architecture of the potential network is a SchNet-like (Schütt et al., 2017) architecture with time-dependent RBF kernels. We refer the reader to Appendix B for details on the Monte Carlo simulations and on the architecture.

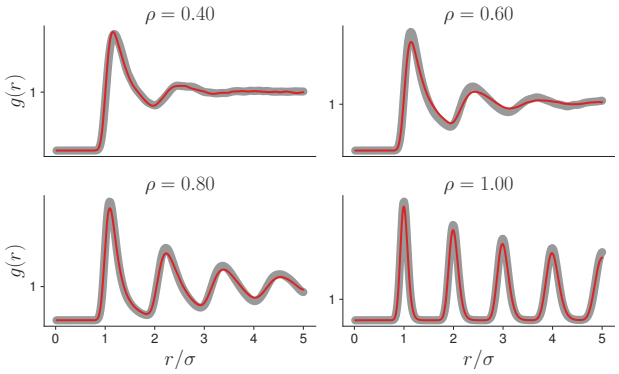

Figure 7. Radial distribution functions as predicted by Monte Carlo simulations (gray) and a diffusion model (red) trained on densities $\rho \in \{0.5, 0.7, 0.9\}$. Note that the model reconstructs $g(r)$ on a wide range of densities starting from $\rho = 0.4$, close to the gas phase to $\rho = 1.0$ close to the solid phase. A Gaussian kernel with $\sigma = 0.05$ was applied to all curves in this plot.

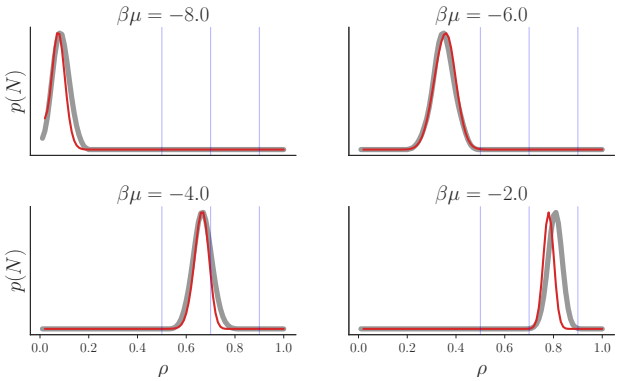

Figure 8. Distribution of the number of particles in the grand canonical ensemble at different chemical potentials as predicted by GCMC simulations (gray) and by thermodynamic integration with a trained diffusion model (red). The blue lines denote the canonical ensembles that the diffusion model was trained on. A Gaussian kernel with $\sigma = 0.015$ was applied to all curves in this plot.

To evaluate the generative performance of these models we compare the radial distribution function $g(r)$ between Monte Carlo samples, and samples from the trained diffusion model (Figure 7).

The accuracy of the thermodynamic integration along the trained diffusion model is gauged by comparing its estimates of $p(N)$ (Figure 8) and of the relation between ρ, μ and μ_{ex} (Figure 9) to grand canonical Monte Carlo simulations.

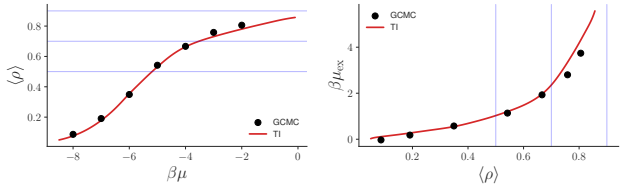

Figure 9. Expected density as a function of the chemical potential (left) and estimates of μ_{ex} as a function of the expected density (right). The vertical blue lines denote the canonical ensembles that the diffusion model was trained on.

B. Experimental details

Monte Carlo sampling

Canonical MC To generate the training dataset in the canonical ensemble, we run Monte Carlo simulations on the positions $\mathbf{q}$. The variance of the proposal distributions are tuned so that the acceptance rate is between 0.10 and 0.30. We run the simulations until 60000 samples are accepted, discard the first 10000 as the burn-in phase, and train our models on the remaining 50000.

Grand Canonical MC To generate reference grand canonical data, we perform GCMC simulations until 15000 moves that change the particle count are accepted. From the generated trajectories we discard the first 20% as the burn-in phase, and use the remaining 80% as the ground truth.

TI in Figure 6 We run simulations at $\lambda \in \{0.00, 0.01, 0.02, ..., 1.00\}$ at particle counts $10, 20, ..., 200$ with the intermediate potential

$$U_\lambda^{\text{LJ}}(\mathbf{q}) = \sum_{i \neq j} 4(1-\lambda)\varepsilon \left[\left(\frac{\sigma^2}{\lambda\sigma^2 + r_{ij}^2} \right)^6 - \left(\frac{\sigma^2}{\lambda\sigma^2 + r_{ij}^2} \right)^3 \right].$$

We collect 12000 samples from which discard the first 2000 as the burn-in phase.

Force fields

D=1 For the 1-dimensional box (Appendix A), we use a SchNet-like (Schütt et al., 2017) architecture. To make the

architecture time-dependent, we predict the parameters of the RBF kernels from the diffusion time using a small MLP. The network consists of 3 layers each of which performs a message passing step and an atom-wise update. The final readout of the energy is the sum of the features over all nodes and channels. The number of channels is 32 and the MLPs in the RBF-parameter prediction and in the atom-wise update both have two hidden layers with 64 neurons. Since the number of particles is reasonably low (≤ 100), we do not use a cutoff radius, and perform message passing on the fully connected graph between the nodes.

D=3 For the 3-dimensional box (§5) we find that the SchNet-like architecture that worked in the one-dimensional case, can not reconstruct $g(r)$ above a density of 0.30. We thus include directional information (Gasteiger et al., 2019) to the potential network. In our architecture nodes and edges are equipped with both scalar and vectorial features. For this discussion we denote these features by $h^{\text{node}}_{\text{scalar}}$, $h^{\text{node}}_{\text{vec}}$, $h^{\text{edge}}_{\text{scalar}}$, $h^{\text{edge}}_{\text{vec}}$. Our network consists of 3 layers each of which performs the following sequence of steps.

1. Update $h^{\text{edge}}_{\text{scalar}}$ ($h^{\text{edge}}_{\text{vec}}$) as a linear combination of the current value of $h^{\text{edge}}_{\text{scalar}}$ ($h^{\text{edge}}_{\text{vec}}$) and the values of $h^{\text{node}}_{\text{scalar}}$ ($h^{\text{node}}_{\text{vec}}$) of the source and target nodes.
2. Compute $g^{\text{node}}_{\text{scalar}}$ ($g^{\text{node}}_{\text{vec}}$) by aggregating the edge features at the target nodes with time-dependent RBF weights (see previous paragraph).
3. Update $h^{\text{node}}_{\text{vec}}$ as a linear combination of $h^{\text{node}}_{\text{vec}}$ and $g^{\text{node}}_{\text{vec}}$.
4. Update $h^{\text{node}}_{\text{scalar}}$ as a MLP whose input is $h^{\text{node}}_{\text{scalar}}$, $g^{\text{scalar}}_{\text{vec}}$, and the channel-wise inner product between $h^{\text{node}}_{\text{vec}}$ and $g^{\text{node}}_{\text{vec}}$.

In our experiments we used 64 channels for all scalar and vectorial features. The final readout is the sum of scalar node features over all nodes and channels. The cutoff radius when building the graph was chosen to be two times the σ parameter of the LJ-potential.