BOLTZMANN SAMPLING WITH STOCHASTIC INTERPOLANTS

Harry Winston Sullivan, Zichen Huang
Diffusion Delinquents
{sull1276,huan2984}@umn.edu

ABSTRACT

A foremost objective of modern research is the development and optimization of computer simulations that can accurately predict the behavior of materials from the atomic scale for any arbitrary system, and we expect such models to closely agree with experimental measurements of macroscopic behavior (Pressure-Volume diagrams, heat capacity, etc) and microscopic behavior (pair correlation functions, reaction mechanisms, etc). Such simulations are called *molecular dynamics simulations* or simply MD. In 2025 MD is central to modern computational science, enabling the study of protein folding, nucleation processes, and phase transitions that remain experimentally intractable. Unfortunately, many such systems of interest suffer from the rare event problem due to rough free energy surfaces, leading to poor sampling with physics based techniques. *This report makes the case that obtaining these samples via deep generative models can be computationally efficient amortize costs effectively solving the rare event problem,*

1 INTRODUCTION

Molecular simulations provide a practical route for connecting microscopic structure and dynamics to experimentally measurable signals. In particular, they can be used to predict scattering patterns (neutron and X-ray) as well as spectroscopic observables spanning near-infrared (Czarnecki et al., 2015), terahertz (Schmuttenmaer, 2004), sum frequency generation (SFG) (Hosseinpour et al., 2020), and nuclear magnetic resonance (NMR) (Mishkovsky & Frydman, 2009) measurements. As these experimental probes are increasingly applied to characterize hydrogen-bond networks in interfacial water (Li et al., 2022), electrolyte solutions (Wang et al., 2022), and biological environments (Meng et al., 2023), there is sustained interest in simulation methodologies that can compute such observables from first principles (Gastegger et al., 2017). Together, these applications underscore the central role of molecular dynamics across chemistry, materials science, chemical engineering, and biology.

Realizing this promise hinges on generating ensembles that are truly representative of the equilibrium (or steady-state) distribution (Bolhuis & Dellago, 2010). For many chemically and biologically relevant systems, however, the underlying free-energy landscape is rugged, with long-lived metastable basins separated by high barriers (Prinz et al., 2011). When the transitions between these basins occur on timescales beyond what is accessible to straightforward simulations, trajectories can appear well-equilibrated within a mode while still failing to explore other important regions of phase space, motivating the use of specialized sampling strategies. In this paper we will explore this phenomena and how we resolve it using deep generative modeling.

1.1 STATISTICAL PHYSICS

To understand how these complex observables are modeled we first need to consider the essential physics involved. As a hallmark example, consider a molecular/atomistic system consisting of N point particles each with their own mass $m_{(i)}$. In classical mechanics the dynamical state of the system is completely specified by the $3N$ coordinates $\boldsymbol{q}^N := \boldsymbol{q}_{(1)}, \cdots, \boldsymbol{q}_{(N)}$ and $3N$ momenta $\boldsymbol{p}^N := \boldsymbol{p}_{(1)}, \ldots \boldsymbol{p}_{(N)}$. Each coordinate and momenta vector lives $\in \mathbb{R}^3$ and as a collective are an element $\in \mathbb{R}^{6N} = \Omega$. At constant number, temperature, and volume the dynamics of such a system

is well modeled by the Klein–Kramers stochastic differential equations (SDE) (Kramers, 1940).

$$d\boldsymbol{p}_{(i)} = -\gamma \boldsymbol{p}_{(i)} dt - \nabla_{\boldsymbol{q}_{(i)}} U(\boldsymbol{q}^N) dt + \sqrt{2\gamma m_{(i)} k_b T} d\boldsymbol{W}_{(i)}, \quad d\boldsymbol{q}_{(i)} = \frac{\boldsymbol{p}_{(i)}}{m_{(i)}} dt \tag{1}$$

Under the following variable change[1] $\boldsymbol{x}_t = \left[\boldsymbol{q}^N, \boldsymbol{p}^N\right]^\top$ one can show that this SDE satisfies a corresponding Fokker-Planck partial differential equation (PDE) given by

$$\partial_t \rho(\boldsymbol{x}, t) = -\nabla \cdot (\boldsymbol{\mu}(\boldsymbol{x}) \rho(\boldsymbol{x}, t)) + \tfrac{1}{2} \nabla \cdot (\boldsymbol{\sigma}\boldsymbol{\sigma}^\top \nabla \rho(\boldsymbol{x}, t)) = -\mathcal{L}\rho(\boldsymbol{x}, t) \tag{2}$$

where $\mathcal{L}$ is the generator of motion (Risken, 1996). By setting $\partial_t \rho(\boldsymbol{x}, t) = 0$ one can show that in the long time limit ($t \to \infty$) ρ approachs the so-called *Boltzmann distribution* or its marginalized form *the configurational Boltzmann distribution*.

$$\rho(\boldsymbol{q}^N, \boldsymbol{p}^N) = Z^{-1} e^{-\beta H(\boldsymbol{q}^N, \boldsymbol{p}^N)} \tag{3}$$

$$\implies \rho(\boldsymbol{q}^N) = \int_{\mathbb{R}^{3N}} \rho(\boldsymbol{q}^N, \boldsymbol{p}^N) d\boldsymbol{p}^N = Q^{-1} e^{-\beta U(\boldsymbol{q}^N)} \tag{4}$$

where (Q) Z is the (configurational) partition function and H is the Hamiltonian a.k.a. the total energy of the system.

The Boltzmann distribution may be used to estimate physical properties of the system of interest. For example, the *heat capacity* is found with

$$C_V = \frac{\partial \mathbb{E}[H]}{\partial T} = \text{heat capacity} \tag{5}$$

which represents the amount of *heat* which must be supplied to the system to produce a change in its temperature. In fact, *all physical properties* of a system can be written as an expectations over this distribution. Some other examples are given below

$$U = \mathbb{E}\big[H\big] = \text{internal energy} \quad p = -\mathbb{E}\left[\frac{\partial H}{\partial V}\right] = \text{pressure}, \quad S = -k_B \mathbb{E}\big[\ln \rho\big] = \text{entropy} \tag{6}$$

In general, for an observable $f : \Omega \to \mathbb{R}$, the experimental observation of f is given by $\mathbb{E}[f]$. Special examples of f give rise to aforementioned scattering patterns, NMR spectra, and even the SFG. This setup gives us the fundamental problem which molecular dynamics aims to solve: *How can we sample equation 3 despite the intractability of the partition function Z.*

1.2 Molecular Dynamics and Advanced Sampling

The most common tool used for sampling equation 3 is molecular dynamics. There are many such permutations of any given scheme, however most are based around the generator $\mathcal{L}$. By discritizing time using a Trotter factorization Trotter (1959) one can solve the split partial differential equation analytically. This scheme, originally developed by Verlet (1967), can be thought of as a discretization akin to the Euler-Maruyama scheme. Formally, by applying the Metropolis correction (Metropolis et al., 1953) this gives rise to the Markov chain Monte Carlo (MCMC) scheme described in algorithm 1 (Bussi & Parrinello, 2007). By applying the Metropolis Hastings correction we can assert that the the Boltzmann distribution is stationary with respect to this sampling process. For completeness, we derive this algorithm exactly in in sections 6.1.2 and 6.1.3 and prove that it satisfies detailed balance, and hence is stationary.

Unfortunately, like most MCMC schemes, the sampling can suffer from the rare event problem when modes are separated by regions of low probability density (Bolhuis & Dellago, 2010). We expand on this point mathematically in supporting information section 6.1.5. When this issue presents itself in the context of MD simulations the results can become biased and unrealistic, defeating the point of MD entirely. In the past this has been resolved using techniques such as transition path sampling (Dellago et al., 2002), forward flux sampling (Allen et al., 2009), parallel tempering (Swendsen & Wang, 1986), among others. Such techniques tend to require a careful hand, and are not easily transferrable across chemistries and conditions. As a scientist, to use one of these techniques on a

[1] A constant number, temperature, and volume distribution is only example of the distributions in statistical physics. Details about this and the mystical varible change are provided in supporting information section 6.1.1.

brand new system of interest one has to start from scratch, rediscover relevant collective variables describing the slow dynamics, and recompute all the reaction pathways for the new system.

Additionally, Markov chain methods (such as algorithm 1) generate correlated samples through a sequence where the next state $\boldsymbol{x}_{t+1}$ depends on the previous $\boldsymbol{x}_t$. This sequential dependency fundamentally limits efficiency: obtaining N independent samples requires $\mathcal{O}(N\tau_{\text{int}})$ chain steps, where the integrated autocorrelation time τ_{int} quantifies the correlation length (Foreman-Mackey et al., 2024). For molecular systems near phase transitions or with multiple metastable states, τ_{int} can exceed 10^6 steps in the worst cases.

1.3 Sampling with Generative Models

With the challenges outlined above in mind, we ask *why not use the vast amount of data available to enhance the sampling on unseen systems?* Generative models do exactly this while (1) eliminating the bottleneck induced by autocorrelation times, (2) prevent exponentially small transition rates induced the rare event problem, and (3) allow a transfer of knowledge across chemistries. They do so by learning a direct map $\boldsymbol{T}_\theta$ which pushes samples forward from some easy-to-sample probability distribution ρ_0 into the desired Boltzmann distribution like

$$\boldsymbol{x} = \boldsymbol{T}_\theta(\boldsymbol{z}) \sim \rho(\boldsymbol{x}) \text{ where } \quad \boldsymbol{z} \overset{\text{i.i.d.}}{\sim} \rho_0(\boldsymbol{z}) \tag{7}$$

This i.i.d. sampling property means N samples require exactly N solutions of equation 20, which may be done in parallel. The computational cost then shifts from runtime sampling to one-time training, amortizing over unlimited future samples. One may be concerned with the a lack of theoretical convergence with only finite data, however if we have access to the Hamiltonian, we can easily correct the samples using the Metropolis criteria. Theoretically speaking, this implies a generative model is as good as typical MCMC scheme in terms of accuracy.

There are many such generative models which can achieve high accuracy when it comes to Boltzmann sampling. In this work we explore the stochatic interpolant framework Albergo et al. (2023). We do so for one simple reason: *Stochastic interpolants contain all diffusion and flow based models as special cases*. There are other reasons, however this is the primary reason. Outside of this Stochastic interpolants have: variable interpolant choices for a broader design space, the ability to use arbitrary base distribution ρ_0 unlike diffusion, the ability to use SDE or ODE based sampling, and theoretical connections with non-equilibrium thermodynamics (He et al., 2025) via the Jarzynksi inequality (Jarzynski, 1997).

2 Related Work

Various generative models have been applied to approximate equation 4, including generative adversarial networks (Jones et al., 2024), variational autoencoders (Monroe & Shen, 2022), and even exotic Lagrangian based methods (Du et al., 2024) to name a few. These methods leverage sequences from equation 1 as dataset a $\mathcal{D}$ and the Hamiltonian H. In particular, Boltzmann generators (BGs) (Noé et al., 2019), which is an example of an equilibrium sampling method (ESM), are able to take advantage of the known energy by introducing an auxiliary loss when compared with typical normalizing flows. Usual normalizing flows maximize the following probability

$$\log \rho_\theta(\boldsymbol{x}^N) \approx \log \rho_Z\big(\boldsymbol{T}_\theta(\boldsymbol{\Gamma})\big) + \log |J_{\boldsymbol{T}_\theta}| \tag{8}$$

where ρ_Z is an easy-to-sample prior distribution. With access to the log probability via the Hamiltonian we can also compute, up to an additive constant, the log density in the latent space variable $\boldsymbol{z}$ induced by $\boldsymbol{T}_\theta$:

$$\log \rho_Z(\boldsymbol{z}) \approx -\beta H\big(\boldsymbol{T}_\theta^{-1}(\boldsymbol{z})\big) + \log |J_{\boldsymbol{T}_\theta^{-1}}| + C. \tag{9}$$

enabling forward and backward losses for a data-free objective.

Since then the paradigm has evolved past invertible neural networks, instead using continuous normalizing flows (Klein & Noé, 2025). In these methods, rather than fitting the limiting distribution, they fit the transition probability kernel itself. Mathematically, this kernel is given by the formal solution to equation 2 $\rho(\boldsymbol{x}, t) = \exp(-\mathcal{L})\rho(\boldsymbol{x}, t)$. We denote these as non-equilibrium sampling methods (NESM). In particular, there are methods such as ITO (Schreiner et al., 2023), BoPITO

(Diez et al., 2024), and Timewarp (Klein et al., 2023). By modeling the transition kernel rather than the equilibrium distribution they obtain a modified training paradigm compared to standard flow modeling. They now require a pair of temporally related states from the dataset $\mathcal{D}$. To highlight their differences, we provide algorithms in the supporting information which describes the training process of a standard flow model (Lipman et al., 2023) with a conditional optimal transport path for both a ESM and a NESM respectively.

With a trained NESM or ESM one has access to many additional results beyond plain sampling. Raja et al. demonstrated that pre-trained generative models can be repurposed for transition path sampling in a zero-shot manner. Their method interprets candidate transition paths as trajectories over the data manifold that minimize the Onsager-Machlup action functional (Onsager & Machlup, 1953) under the learned score function. By leveraging pre-trained models without task-specific retraining, this approach obtains diverse, physically realistic transition pathways and generalizes beyond the original training distribution.

Furthermore, Arts et. al. have proven that training score-based generative models on coarse grained (CG) data implicitly approximates a CG force field $-\nabla U$, which can be exported to directly simulate CG MD Arts et al. (2023). This allows the score-matching framework to connect with broader statistical mechanical theory and compete with standard CG methods, such as relative entropy minimization Shell (2008). These works exemplify the trend toward re-purposing large-scale, general-purpose generative models for specialized sampling tasks (Liu et al., 2023).

3 METHODS

In this methodological development we will describe three primary prongs: (1) Symmetry in physical systems and the consequences for generative models. (2) Generative modeling with stochastic interpolants. (3) Equivariant architectures to assert accurate sampling. While each of these topics are not novel on their own, we hope the discussion, free to access implementation, and unique combination is a valuable contribution to the research corpus. Other generative Boltzmann sampling models account for equivariance and apply flow/diffusion concepts, however, we believe the exposition and provided python package will allow for hastened development of deep learning methods and can serve as a jumping off point for future researchers.

3.1 SYMMETRY AND EQUIVARIANCE

From vast amounts of experimental evidence we know that equation 3 exhibits symmetry with respect to rotations and translations (which constitute a mathematical group). The reasoning behind such symmetry is that the laws of classical physics do not change depending on your reference frame, this is akin to choosing a basis on a vector space. If at any point the result of calculation depends on the reference frame then the calculation must be wrong. The group in question is known as the special Euclidean group[2] $SE(3)$. Consider the case when $N = 1$, the action of a group element $g \in SE(3)$, which may be represented by a translation $\boldsymbol{\tau} \in \mathbb{R}^3$ and a rotation matrix $\boldsymbol{Q} \in SO(3)$, is applied to the components in the following way

$$\boldsymbol{q}_{(i)} \to \boldsymbol{Q} \cdot (\boldsymbol{q}_{(i)} + \boldsymbol{\tau}) = \tilde{\boldsymbol{q}_{(i)}}, \quad \boldsymbol{p}_{(i)} \to \boldsymbol{Q} \cdot \boldsymbol{p}_{(i)} = \tilde{\boldsymbol{p}_{(i)}}, \tag{10}$$

Under such a group action the Hamiltonian is known to remain *invariant* $H(\boldsymbol{q}^N, \boldsymbol{p}^N) = H(\tilde{\boldsymbol{q}}^N, \tilde{\boldsymbol{p}}^N)$, which in turn makes the distribution in equation 3 invariant to the same such action $\implies \rho(\boldsymbol{q}^N, \boldsymbol{p}^N) = \rho(\tilde{\boldsymbol{q}}^N, \tilde{\boldsymbol{p}}^N)$. This feature of the distribution must be accounted for in any model of it, otherwise we will be violating the laws of physics. Many generative models work under the assumption of a differentiable bijection $\boldsymbol{T}_\theta$ acting on a d-dimensional real vector $\boldsymbol{q}$ between some easy-to-sample base distribution ρ_Z at flow time (not physical time) $t = 0$ and the distribution of interest at $s = 1$. The bijection can then be applied to obtain a model distribution via the pushforward $\boldsymbol{T}_{\boldsymbol{\theta},\#}[\,\cdot\,]$ like

$$\log \rho_\theta(\boldsymbol{q}, t = 1) = \boldsymbol{T}_{\boldsymbol{\theta},\#}\left[\log \rho_Z\right](\boldsymbol{q}) = \log \rho_Z\big(\boldsymbol{T}_\theta(\boldsymbol{q})\big) + \log |J_{\boldsymbol{T}_\theta}|. \tag{11}$$

[2]The definition of $SE(3)$ is given by $SE(3) \simeq \mathbb{R}^3 \rtimes SO(3)$. It is the semidirect product of translations and rotations, constituting the set of reference frames.

where $J_{\boldsymbol{T}_\theta}$ is the determinant of the Jacobian. In the context of flow based ESM model the bijection is commonly written in terms of a velocity $\boldsymbol{v}_\theta$ like

$$\boldsymbol{\phi}_t(\boldsymbol{q}) = \int_0^s \boldsymbol{v}_\theta\left(\boldsymbol{\phi}_{s'}(\boldsymbol{q}), s\right) ds' \implies \boldsymbol{T}_\theta(\boldsymbol{q}) = \boldsymbol{\phi}_1(\boldsymbol{q}). \tag{12}$$

Which in turn implies a family of distributions over time.

The question then remains, what properties does the velocity $\boldsymbol{v}_\theta$ giving rise to $\boldsymbol{T}_\theta$ need to have in order to ensure that all symmetries $g \in G$ leave the density $\rho_\theta(\boldsymbol{q}, t)$ unchanged? We require:

$$\forall g \in G : \rho_\theta(g \circ \boldsymbol{q}, t) = \rho_\theta(\boldsymbol{q}, t). \tag{13}$$

This problem was investigated by Köhler et al. (2020), in that work they proved the following statement: Let G and H be groups with representations on $\mathbb{R}^N$. Let ρ_Z be a density on $\mathbb{R}^n$ which is G-invariant where G is a subgroup of H (i.e. $G < H$). If $\boldsymbol{f}$ is a H-equivariant diffeomorphism, i.e. $\forall h \in H, \boldsymbol{q} \in \mathbb{R}^n : \boldsymbol{f}(h \circ \boldsymbol{q}) = h \circ \boldsymbol{f}(\boldsymbol{q})$, then $\boldsymbol{f}_\#[\rho_Z]$ is H-invariant. This implies the following statement in our context: *in order to obey the laws of physics our velocity function $\boldsymbol{b}_\theta$ must rotate if its input $\boldsymbol{q}$ is rotated, i.e. it must be equivariant.*.

This bounds the set of all models to the subset of only those models which commute with rotation. As described in the proposal, one method to achieve this is data augmentation, however this provably less efficient in some cases (Mei et al., 2021) and empirically shown to be worse (Schütt et al., 2021). Although there is evidence pointing towards built in equivariance, it is still debated which is the correct method in practice. Recently Plainer et al. (2025) have opted for data augmentation for approximate equivariance in the context of equilibrium Boltzmann sampling and have shown competitive performance in terms of both accuracy and computational effort.

3.2 Interpolant Design

In order to sample equation 4 we built a stochastic interpolant (Albergo et al., 2023), establishing us as an ESM. To do so, we made the choice of base and target distribution

$$\rho(\boldsymbol{q}^N, t=0) = \mathcal{N}(\boldsymbol{q}^N | \boldsymbol{0}^N, \sigma^2 \,\mathrm{id}_{\mathbb{R}^{3N}}), \quad \rho(\boldsymbol{q}^N, t=1) = Q^{-1} e^{-\beta U(\boldsymbol{q}^N)} \tag{14}$$

along with the choice of interpolant

$$\boldsymbol{q}_t = I(t, \boldsymbol{q}_0, \boldsymbol{q}_1) + \gamma(t)\boldsymbol{z}, \quad t \in [0,1], \tag{15}$$

where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \mathrm{id}_{\mathbb{R}^3})$. This formula is understood to be applied to each atom in the N body system. In particular, we further restricted the interpolant to be temporally and spatially linear with a relatively simple volatility scale $\gamma(s)$, meaning.

$$I(t, \boldsymbol{q}_0, \boldsymbol{q}_1) = (1-t)\boldsymbol{q}_0 + t\boldsymbol{q}_1, \quad \gamma(s) = a\sqrt{2t(1-t)}. \tag{16}$$

Leaving only two free interpolant hyperparameters as a and σ.

This setup creates a stochastic process connecting the base and target distribution. There exists two key functions which describe $\rho(\boldsymbol{q}, t)$, namely *the velocity* $\boldsymbol{b}$ and *the denoiser*[3] $\boldsymbol{\eta}$. These are defined as

$$\boldsymbol{b}(t, \boldsymbol{q}) = \mathbb{E}\big[\partial_t I(t, \boldsymbol{q}_0, \boldsymbol{q}_1) + \dot{\gamma}(t)\boldsymbol{z} \mid \boldsymbol{q}_t = \boldsymbol{q}\big] \quad \boldsymbol{\eta}_z(t, \boldsymbol{q}) = \mathbb{E}\left[\boldsymbol{z} \middle| \boldsymbol{q}_t = \boldsymbol{q}\right]. \tag{17}$$

where $\boldsymbol{z}$ is a normally distributed random variable independent of the end points. These functions then describe three partial differential equations: (1) the transport equation, (2) the forward Fokker-Planck equation, and (3) the backward Fokker-Planck equation.

To train such an interpolant we utilize the following losses

$$\mathcal{L}_b(\theta) = \mathbb{E}\left[\tfrac{1}{2}\|\boldsymbol{b}_\theta(t, \boldsymbol{q}_t)\|^2 - \left(\partial_t I(t, \boldsymbol{q}_0, \boldsymbol{q}_1) + \dot{\gamma}(t)\boldsymbol{z}\right) \cdot \boldsymbol{b}_\theta(t, \boldsymbol{q}_t)\right] \tag{18}$$

$$\mathcal{L}_\eta(\theta) = \mathbb{E}\left[\tfrac{1}{2}\|\boldsymbol{\eta}_z(t, \boldsymbol{q}_t)\|^2 - \boldsymbol{z} \cdot \boldsymbol{\eta}_z(t, \boldsymbol{q}_t)\right] \tag{19}$$

where the expectation is taken over uniformly distributed times $t \in [0,1]$, $\boldsymbol{q}_0 \sim \rho_0$, and $\boldsymbol{q}_1 \in \mathcal{D}$ where $\mathcal{D}$ is the dataset sampled from ρ_1 which is the configurational Boltzmann distribution in

[3]The denoiser $\boldsymbol{\eta}_z$ is related to the score $\boldsymbol{\mathcal{S}}$ via $\boldsymbol{\mathcal{S}}(s, \boldsymbol{q}) = -\gamma^{-1}(t)\boldsymbol{\eta}_z(t, \boldsymbol{q}) = \nabla_{\boldsymbol{q}} \log \rho(\boldsymbol{q}^N, t)$

equation 4. If you look closely at the loss you can notice that it is similar to the loss of a flow model (Lipman et al., 2023). The only difference is the introduction of a latent noise parameter $\boldsymbol{z}$, which provably improves the upper bound on the KL divergence between the model pushforward and the data distribution (Albergo & Vanden-Eijnden, 2023).

The sampling procedure can then be done two ways. The first given by numerically solving the deterministic probability flow (left) SDE (right) defined with

$$\frac{d\boldsymbol{q}_t}{dt} = \boldsymbol{b}_\theta(t, \boldsymbol{q}_t), \quad d\boldsymbol{q}_t = \left(\boldsymbol{b}_\theta(t, \boldsymbol{q}_t) - \tfrac{\epsilon(t)}{\gamma(t)}\boldsymbol{\eta}_z(t, \boldsymbol{q}_t)\right) dt + \sqrt{2\epsilon(t)} d\boldsymbol{W}_{\boldsymbol{t}}. \tag{20}$$

where $\epsilon(t) \geq 0$ is an inference time hyperparameter controlling the diffusivity of the sampling. In principle any interpolant satisfying the axioms outlined by Albergo & Vanden-Eijnden (2023) could be chosen. We opted for the simple choice of a temporally and spatially linear set with $a = 0.1$ and $\sigma = 0.5$ as an initial test case. At the time of this report we did not have the opportunity to ablate these parameters alongside the underlying interpolant functional form.

3.3 Geometric Graphs and Equivariant Message Passing Neural Networks

For our use case of approximating equation 4, we require an equivariant velocity $\boldsymbol{b}_\theta$. To this end, we apply equivariant message passing graph neural networks. Such a model is based on a geometric graph $\mathcal{G}$ representing the state of the N body system with coordinates $\boldsymbol{q}^N$. Mathematically such a geometric graph is defined by the set $\mathcal{G} := \left\{(\boldsymbol{q}_{(i)}, \mathbf{f}_{(i)}\right\}$ which represents the location $\boldsymbol{q}_{(i)}$ and state $\mathbf{f}_{(i)}$ of node i in the graph. Importantly, this is a *set*, meaning order does not matter. We choose the features $\mathbf{f}_i$ to be coefficients of the spherical harmonics, meaning they are spherical tensors. Formally speaking, a tensor of rank p is a multilinear map from p copies of some vector space to the real numbers Schrödinger (1985).

When someone lists out the components of a tensor they must realize that it is tied to a given basis. The components come from an evaluation of the multilinear map at all possible combinations of the basis vectors. This implies under a rotation $\boldsymbol{Q}$ the components transform as the basis vectors also do. Such a transformation is given by $f_m^\ell \to \sum_{n=1}^{2\ell+1} D_{mn}^{(\ell)}(\boldsymbol{Q}) f_n^\ell$ where $D_{mn}^{(\ell)}(Q)$ is the mnth component of the block diagonal Wigner-D matrix of rank ℓ. Each rank ℓ contains $2\ell + 1$ components. The block diagonal property of $\mathbf{D}$ is due to the spherical tensors being the *irreducible representations* of the group $O(3) < E(3)$.

To process a geometric graph we require a neighborhood rule. In order to preserve equivariance such a rule must be only dependent scalar properties, such as relative distances $\|\boldsymbol{q}_i - \boldsymbol{q}_j\|$. We do exactly this, implying $j \in \mathcal{N}_i \iff \|\boldsymbol{q}_j - \boldsymbol{q}_i\| \leq r_{\text{max}}$ .If r_{max} is sufficiently large this corresponds to a fully connected graph. We opted for $r_{\text{max}} \to \infty$ in our implementation.

Given $\mathcal{G}$ along with the neighborhoods $\mathcal{N}_i$ we then require equivariant modules which preserve rotations. The allowed operations are outlined by Duval et al. (2024), however we recap the important concepts here briefly.

1. **Scalarization**: Scalarization can be done through the use of a tensor contraction. This involves a conversion from spherical to cartesian tensors and then summation over tensor dimensions using a Dirac delta symbol δ_{ij} or a Levi-Civita symbol ϵ_{ijk}. This process lowers the rank of the tensor.
2. **Linear Transformation** : Tensors may be summed with one another as long as they share their tensor rank. They may be combined linearly in the typical way with scalar multiplication followed by summation. Some call this a *n-mode (matrix) product of a tensor*. However, there is no reason for such complication. It is just a linear map.
3. **Tensor Products**: The tensor product takes a tensor T of rank p and a tensor S of rank q to a tensor of rank $p + q$. The resulting tensor $S \otimes T$ is again a tensor.
4. **Scalar Non-Nonlinearities**: Any operation may be done on the scalars. Typically these are just multilayer perceptrons.

With these operations in mind, we use a message passing neural network that takes interpolant time t, noisy interpolant positions $\boldsymbol{q}_t^N$, and per-particle physical features (charge, mass, particle size, well depth, and atom type). Time and physics features are embedded using sin-cos encoding (Ho et al., 2020) and MLPs respectively. A neighborhood graph is then built directly from the geometry

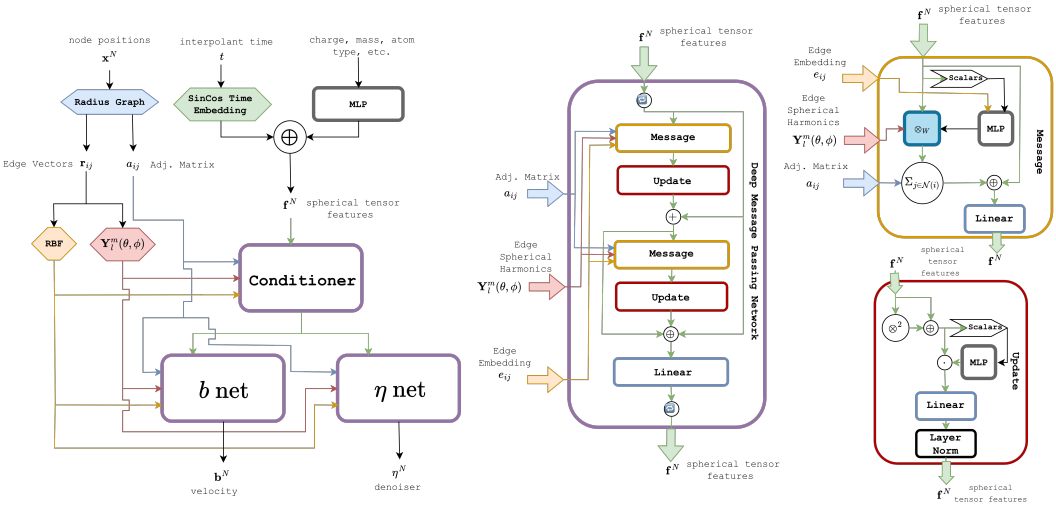

Figure 1: Our message passing neural network. (Left) the entire network, (Center) deep message passing module, (Upper Right) a message module, (Lower Right) an update module.

of $\boldsymbol{q}_t^N$ (e.g., a radius graph). For each neighbor pair, the model computes the edge length and direction from the relative displacement, expands the edge length with an RBF basis, and expands the edge direction with spherical harmonics. These geometric expansions, together with the embedded time/physics context, are passed through a conditioner network, whose outputs are then fed into two heads: an η-net and a b-net. The conditioner, η-net, and b-net share the same architectural template (same layer types and connectivity) but use separate parameters, and each is implemented as a deep message passing network.

Each deep message passing network consists of a repeated sequence of message/update blocks with two skip connections. Starting from state (a), message and update functions produce intermediate state (b). This intermediate state is then combined with point (a) via a residual summation skip to form point (b), after which a second message/update block is applied. Finally, a concatenation skip forms $[(a), (b), \text{current}]$, which is mapped back to the hidden dimension by a linear layer. This full pattern is repeated twice, with the two repetitions using different parameters.

In the message layer, we split each node's scalar and tensor features into two streams. The scalar stream, together with the edge RBF embedding, is passed through an MLP to produce mixing weights. These weights parameterize a weighted tensor product between the node's spherical tensor features and the edge spherical harmonics, yielding edge messages. The edge messages are summed over each node's neighborhood, concatenated with the node's original state, and mapped back to the hidden dimension with a linear layer. In the update layer, the tensor stream is enhanced by concatenating the tensor-square with the original tensor features, then scalar stream is split to produce scalars used for scalar multiplication with the tensor features. The resulting tensor features are then passed through a linear layer followed by a geometric layer normalization.

4 Results

We evaluate our equivariant stochastic interpolant on an alanine dipeptide system, a 22-atom benchmark dataset found on mdshare from the Markov Modeling Group (2024). The model was trained on 25K randomly subsampled trajectory frames and rescaled to have zero mean and unit standard deviation. 5K additional frames were chosen as a testing set to gauge overfitting during training. We opted hidden spherical tensor features up to rank $\ell = 1$, details can be seen in the provided code within the EquivariantMINTModule object instantiation. In total we have 13.5 M trainable parameters. We run our code on MSI on the sarupria group account using interactive GPUs. We trained our network for 13k steps with a batch size of 128. Training and testing losses are shown in figure 2. This process took approximately 9h 58m 47s and completed a total of 120 epochs. A GitHub repo is provided `https://github.com/WinstonWinstonWinston/mint`.

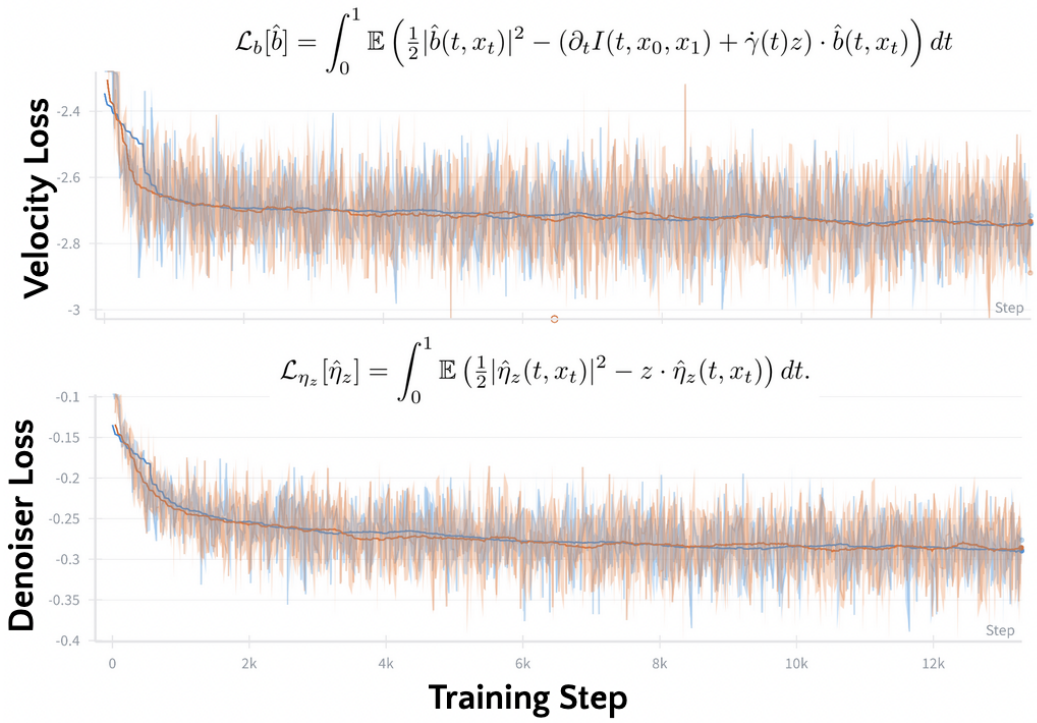

Figure 2: Loss curves logged with Weights & Biases (wandb). The top panel shows the velocity $\boldsymbol{b}_\theta$ loss and the bottom panel shows the denoiser $\boldsymbol{\eta}_z$ loss. Training is shown in blue and validation in orange. Faint traces show the raw (per-step) values, while solid traces show a simple moving average with window size 250 steps.

Initially, to evaluate performance we computed the equivariance error. In particular we use *the commutator*. The commutator between a neural network NN_θ and a rotation $\hat{R}$ which both act on data $\mathcal{D}$ is given by

$$[\hat{R}, NN_\theta](\mathcal{D}) = \hat{R}(NN_\theta(\mathcal{D})) - NN_\theta(\hat{R}(\mathcal{D})). \tag{21}$$

A neural network is considered equivariant if the commutator is zero. Applying this to our network results very low errors due to the exactly equivariant structure. These are summarized in table 1.

Tensor	Status	Mean (Å)	Std. (Å)	Max (Å)	Tol. (Å)	$\|\cdot\|_{\text{before}}$	$\|\cdot\|_{\text{after}}$
f_{cond}	OK	8.543×10^{-7}	1.168×10^{-6}	2.575×10^{-5}	1×10^{-3}	10.46	10.46
$\boldsymbol{b}_\theta$	OK	1.175×10^{-6}	1.504×10^{-6}	1.621×10^{-5}	1×10^{-3}	1.573	1.573
$\boldsymbol{\eta}_z$	OK	7.61×10^{-7}	1.331×10^{-6}	2.05×10^{-5}	1×10^{-3}	0.4525	0.4525

Table 1: Summary statistics of equivariance error in equation 21 (in Å) computed over 128 samples from the test dataset.

To further evaluate the effectiveness of our model even further we computed the *free energy marginal*, or Ramachandran plot. It is defined as

$$F(\psi', \phi') = -k_b T \ln \mathbb{E}[\delta(\phi' - \phi(\boldsymbol{q}^N))\delta(\psi' - \psi(\boldsymbol{q}^N))] \tag{22}$$

where the functions ϕ and ψ measure the dihedral angles of backbone of alanine dipeptide. These angles are shown at the top of figure 3. In alanine dipeptide, the Ramachandran plot is a compact benchmark for whether a model samples backbone conformations correctly. Matching the locations and relative populations of the major basins indicates that the underlying energy landscape and backbone torsional energetics are realistic. If the plot is mismatched, it usually signals errors in the balance of local sterics, electrostatics, and solvation that will propagate to larger peptides and

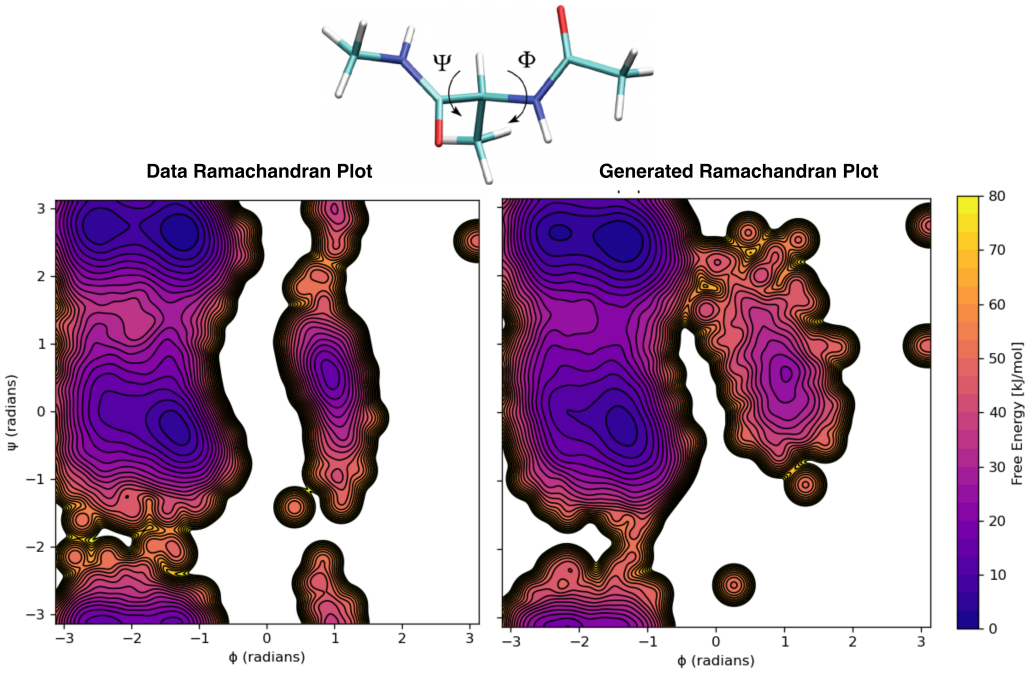

Figure 3: Ramachandran free-energy surface (kJ/mol) for alanine dipeptide: (left) reference dataset, (right) generated model. The alanine dipeptide molecule is shown at the top.

proteins. Because alanine dipeptide is small and well-studied, deviations are easy to diagnose and directly tie to sampling deficiencies.

5 Discussion and Conclusions

Across experiments, the implemented exactly equivariant message passing network consistently generated physically plausible configurations. In particular, the Ramachandran distributions of generated samples closely matched those of the reference data, indicating that the model captures the underlying conformational physics rather than merely reproducing superficial statistics. A central practical outcome is that inference-time sampling is substantially faster than conventional molecular dynamics while remaining faithful to the same physical constraints observed in the dataset. Because the model produces independent and identically distributed samples, it also directly mitigates the rare-event bottleneck of MD. However, we want it to be quicker. The dominant failure mode observed was runtime speed during training and/or generation compared with non-equivariant neural networks. While this architectural choice improved fidelity, it imposed a substantial performance penalty that limits scalability.

To address this, we implemented a data augmentation routine paired with a plain graph transformer (non-equivariant). This alternative delivered an approximately 10× speedup, confirming that much of the bottleneck stems from the equivariant operations themselves. Despite the improved throughput, the approach did not reach the quality level of the equivariant model, with reduced agreement in the physics-sensitive metrics (including the conformational statistics reflected in Ramachandran space and the commutator errors). These results highlight a key limitation: removing equivariance can recover speed, but preserving the same physical accuracy is nontrivial. We believe with more time we could have gotten this method to work and that it is the future of the field. For completeness, the Ramachandran analysis for the graph transformer is provided in the supporting information section 6.3. The results suggest a clear direction: steer away from strictly equivariant networks if the goal is scalability to full liquids, while compensating for the lost structure through other means. Furthermore, we hope to apply the aforementioned specialized sampling methods, in particular the force field extraction introduced by Arts et al. (2023).

6 SUPPORTING INFORMATION

6.1 STATISTICAL PHYSICS

6.1.1 THE DISTRIBUTIONS OF STATISTICAL PHYSICS

Consider a molecular/atomistic system consisting of N point particles each with their own mass $m_{(i)}$. In classical mechanics the dynamical state of the system is completely specified by the $3N$ coordinates $\boldsymbol{q}^N := \boldsymbol{q}_{(1)}, \cdots, \boldsymbol{q}_{(N)}$ and $3N$ momenta $\boldsymbol{p}^N := \boldsymbol{p}_{(1)}, \dots \boldsymbol{p}_{(N)}$. Each coordinate and momenta vector lives $\in \mathbb{R}^3$ and as a collective are an element $\in \mathbb{R}^{6N} = \Omega$. Colloquially these are known as *phase points*. We call the function $U : \mathbb{R}^{3N} \to \mathbb{R}$ the *potential energy* and its derivative $-\nabla_{\boldsymbol{q}^N} U : \mathbb{R}^{3N} \to \mathbb{R}^{3N}$ the *force field*. It will be useful later on to introduce the Hamiltonian H of a phase point. It is given by

$$H(\boldsymbol{q}^N, \boldsymbol{p}^N) = \tfrac{1}{2}(\boldsymbol{p}^N)^\top M^{-1} \boldsymbol{p}^N + U(\boldsymbol{q}^N) \tag{23}$$

where M is the mass matrix given by $M = \operatorname{diag}(m_{(1)}\boldsymbol{I}_3, \dots, m_{(N)}\boldsymbol{I}_3) \in \mathbb{R}^{3N \times 3N}$.

Classically we care about the dynamics of such a system. Newtons equation $\boldsymbol{F} = m\ddot{\boldsymbol{x}}$ does the job (Newton, 1687), however when dealing with many body systems and statistical averages it tends to be easier to describe the motion using Hamilton's equations. These are completely equivalent to Newtons formulation and are the basis for nearly all algorithms of molecular dynamics. Hamilton's equations are a set of coupled ODEs, for each particle these are given by

$$\frac{d}{dt}\boldsymbol{x}_{(i)} = \frac{\partial H}{\partial \boldsymbol{p}_{(i)}}, \quad \frac{d}{dt}\boldsymbol{p}_{(i)} = -\frac{\partial H}{\partial \boldsymbol{x}_{(i)}}. \tag{24}$$

The resulting ensemble/distribution in the long time limit is known as the microcanonical distribution and it is given by

$$\rho(\boldsymbol{x}^N, \boldsymbol{p}^N) = \frac{1}{\Omega(E, V, N)} \delta\big(E - H(\boldsymbol{x}^N, \boldsymbol{p}^N)\big) \tag{25}$$

where the delta form is actually due to the conservation of energy inherent to Hamilton's equations. The *microcanonical partition function* is given by

$$\Omega = \frac{1}{h^{3N} N!} \int \delta\big(E - H(\boldsymbol{x}^N, \boldsymbol{p}^N)\big) d\boldsymbol{x}^N d\boldsymbol{p}^N \tag{26}$$

where h is Planck's constant. This set of equations models a system with constant number N, volume V, and energy E.

Alternatively, the motion of a phase point in the canonical ensemble, with constant temperature T as opposed to energy, is determined by the second order Langevin equation (Lemons & Gythiel, 1997), in particular the stochastic differential equation (SDE) introduced by Kramers (1940).

$$d\boldsymbol{p}_{(i)} = -\gamma \boldsymbol{p}_{(i)} dt - \nabla_{\boldsymbol{q}_{(i)}} U(\boldsymbol{q}^N) dt + \sqrt{2\gamma m_{(i)} k_b T} d\boldsymbol{W}_{(i)}, \quad d\boldsymbol{q}_{(i)} = \frac{\boldsymbol{p}_{(i)}}{m_{(i)}} dt \tag{27}$$

where T is the temperature, γ is the friction coefficient with the heat bath, k_b is Boltzmann's constant, and $d\boldsymbol{W}_{(i)}$ is a delta correlated, zero mean, 3-dimensional Gaussian process. These SDEs are essentially stochastic versions of Hamilton's equations. Note it is sometimes easier to utilize $\beta = (k_b T)^{-1}$.

It can be helpful to change variables from $\boldsymbol{q}^N, \boldsymbol{p}^N \to \boldsymbol{x}_t$ where the subscript is meant to emphasize the time dependence. The subscript on $\boldsymbol{x}$ may be omitted in cases where it is the space variable of a PDE. In particular, let $\boldsymbol{x}_t = \left[\boldsymbol{q}^N, \boldsymbol{p}^N\right]^\top$ which lets us define the *drift* $\boldsymbol{\mu} : \mathbb{R}^{6N} \to \mathbb{R}^{6N}$ and the *volatility* $\boldsymbol{\sigma} \in \mathbb{R}^{6N \times 6N}$ like

$$\boldsymbol{\mu}(\boldsymbol{x}_t) = \begin{bmatrix} M^{-1}\boldsymbol{p}^N \\ -\gamma \boldsymbol{p}^N - \nabla_{\boldsymbol{q}^N} U(\boldsymbol{q}^N) \end{bmatrix}, \quad \boldsymbol{\sigma} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & M^{1/2}\sqrt{2\gamma k_b T} \end{bmatrix} \tag{28}$$

This converts eq. 1 into the standard SDE form with the $6N$ dimensional Wiener process given by $d\boldsymbol{x}_t = \boldsymbol{\mu}(\boldsymbol{x}_t)dt + \boldsymbol{\sigma} d\boldsymbol{W}_t$. SDEs in the standard form are known to satisfy the Fokker-Planck

equation

$$\partial_t \rho(\boldsymbol{x}, t) = -\nabla \cdot (\boldsymbol{\mu}(\boldsymbol{x})\rho(\boldsymbol{x}, t)) + \tfrac{1}{2}\nabla \cdot (\boldsymbol{\sigma}\boldsymbol{\sigma}^\top \nabla \rho(\boldsymbol{x}, t)) = -\mathcal{L}\rho(\boldsymbol{x}, t) \tag{29}$$

which describes the evolution of the probability density of the system being at a given phase point $\boldsymbol{x}$ at time t. The quantity $\mathcal{L}$ is known as the *generator* or the *Fokker Planck operator*. The formal solution to equation 2 for a finite timestep is given by

$$\rho(\boldsymbol{x}, t) = \exp(-t\mathcal{L})\rho(\boldsymbol{x}, 0) = U(t)\rho(\boldsymbol{x}, 0). \tag{30}$$

Often $U(t)$ is called the *time evolution operator* or the *propagator*.

The Markov chain generated by eq. 1 can be shown to approach the so-called *Boltzmann distribution*, $\rho(\boldsymbol{x}^N, \boldsymbol{p}^N)$ in the limit as $t \to \infty$.

$$\rho(\boldsymbol{x}^N, \boldsymbol{p}^N) = \frac{1}{Z(\beta, V, N)} \exp\big[-\beta H(\boldsymbol{x}^N, \boldsymbol{p}^N)\big] \tag{31}$$

where Z is *partition function* given by

$$Z = \frac{1}{h^{3N} N!} \int \exp\big[-\beta H(\boldsymbol{x}^N, \boldsymbol{p}^N)\big] d\boldsymbol{x}^N d\boldsymbol{p}^N. \tag{32}$$

To prove that this is the case simply set $\partial_t \rho = 0$ in equation 2 and use the Boltzmann distribution as an ansatz.

The momentum is typically marginalized over as it is a Gaussian and easy to sample in closed form. This finally gives the configurational distribution, which we care about, as

$$\rho(\boldsymbol{x}^N) = \int \rho(\boldsymbol{x}^N, \boldsymbol{p}^N) d\boldsymbol{p}^N = Q^{-1} e^{-\beta U(\boldsymbol{x}^N)} \tag{33}$$

where Q is the configurational partition function. As described in the proposal, this has many uses when it comes to modeling proteins, fluids, solids, and really any material.

In principle, there are many such permutations of such a probability distributions over phase space. In fact, if we allow the particle number N to vary then it is given by

$$\rho(\boldsymbol{x}^N, \boldsymbol{p}^N, N) = \frac{1}{\Theta(\beta, \mu, V)} \frac{1}{h^{3N} N!} \exp\big[-\beta\big(H(\boldsymbol{x}^N, \boldsymbol{p}^N) - \mu N\big)\big] \tag{34}$$

where μ is the so called *chemical potential* and Θ is the *grand canonical partition function* given by

$$\Theta = \sum_{N=0}^{\infty} \frac{e^{\beta \mu N}}{h^{3N} N!} \int \exp\big[-(\boldsymbol{x}^N, \boldsymbol{p}^N)\big] d\boldsymbol{x}^N d\boldsymbol{p}^N \tag{35}$$

This distribution as a probability measure on the disjoint union of the N-particle phase spaces, which is very similar to the *Fock space* of quantum statistical mechanics. We could keep listing distributions in this manner, holding some thermodynamic variable constant, creating an SDE/ODE, and writing the limiting distribution.

6.1.2 The Standard Numerical Sampling Algorithm

Tuckerman et al. (1992) were able to produce samples of ρ by taking advantage of Trotter's factorization (Trotter, 1959) of the propagator. The reversible reference system propagator algorithm (RESPA) is still in use to this day and is the basis for most modern integrators, such as the BAOAB/ABOB/Gromacs stochastic dynamics variants (Leimkuhler & Matthews, 2012) (Kieninger & Keller, 2022), SIN(R) (Leimkuhler et al., 2013), and the basic velocity Verlet (Verlet, 1967). RESPA avoids the need for discretization in space and only does so in time, with the caveat that it assumes a delta function initial condition for ρ if you want to apply their method. To understand their method, first consider an S stage decomposition with P time steps such that $\sum_{i=1}^{S} \mathcal{L}_i$ and

$t = P\Delta t$. Utilizing Trotters formula to the solution in eq. 30 gives

$$U(t) \approx \left(\prod_{j=1}^{S} \exp\left(-\frac{\Delta t}{2} \mathcal{L}_{S+1-j} \right) \prod_{k=1}^{S} \exp\left(-\frac{\Delta t}{2} \mathcal{L}_k \right) \right)^P = G(\Delta t)^P \tag{36}$$

where the first product may be thought of as writing the sum $\sum_{i=1}^{S} \mathcal{L}_i$ from largest to smallest index and the second vice versa. By distributing half the time in a palindromic way a time reversible approximation is obtained $G(\Delta t)^{-1} = G(-\Delta t)$. This is simply due to the fact that $(AB)^{-1} = B^{-1}A^{-1}$. Note this is only true if each term in $\mathcal{L}_i$ has a proper inverse. Tuckerman's strategy is very similar to the so-called Strang splitting method (Strang, 1968).

6.1.3 Langevin Thermostat

Bussi et al.'s integrator, known as the Langevin thermostat, can be considered a special case of Tuckerman's integrator with $S = 3$ stages (Bussi & Parrinello, 2007). The splitting of $\mathcal{L}$ is given by

$$\begin{aligned} \mathcal{L}_{\boldsymbol{p}} = -\nabla_{\boldsymbol{q}^N} U(\boldsymbol{q}^N) \cdot \nabla_{\boldsymbol{p}^N}, \quad \mathcal{L}_{\boldsymbol{q}} = M^{-1} \boldsymbol{p}^N \cdot \nabla_{\boldsymbol{q}^N}, \\ \mathcal{L}_\gamma = -\gamma \left(\nabla_{\boldsymbol{p}^N} \cdot \boldsymbol{p}^N + \frac{1}{\beta} \nabla_{\boldsymbol{p}^N} \cdot \left(M \nabla_{\boldsymbol{p}^N} \right) \right). \end{aligned} \tag{37}$$

Such a scheme then gives discretized operator as

$$G(\Delta t) = \exp\left(-\frac{\Delta t}{2} \mathcal{L}_\gamma \right) \underbrace{\exp\left(-\frac{\Delta t}{2} \mathcal{L}_p \right) \exp\left(-\Delta t \mathcal{L}_q \right) \exp\left(-\frac{\Delta t}{2} \mathcal{L}_p \right)}_{=\text{velocity Verlet}} \exp\left(-\frac{\Delta t}{2} \mathcal{L}_\gamma \right) \tag{38}$$

which corresponds conceptually to the stepping scheme: half thermal step $\rightarrow$ velocity Verlet $\rightarrow$ half a thermal step. This is the OBABO ordering in the notation of Leimkuhler & Matthews (2012). An important note is that this operator is not strictly reversible like typical RESPA in the sense $G(\Delta t)^{-1} \neq G(-\Delta t)$, which can be attributed to the dissipative property of $\mathcal{L}_\gamma$.

With a properly discretized operator, we can consider the time evolution of some density ρ in the following way $\rho(\boldsymbol{x}, \Delta t) = G(\Delta t)\rho(\boldsymbol{x}, 0)$. To solve the system we may apply each stage sequentially, first consider the thermal $\mathcal{L}_\gamma$ operator subject to the initial condition $\rho(\boldsymbol{x}, 0) = \delta(\boldsymbol{x} - \boldsymbol{x}_0)$

$$\partial_t \rho(\boldsymbol{x}, t) = -\mathcal{L}_\gamma \rho(\boldsymbol{x}, t) = -\gamma \left(\nabla_{\boldsymbol{p}^N} \boldsymbol{p}^N + \frac{1}{\beta} \nabla_{\boldsymbol{p}^N} \cdot \left(M \nabla_{\boldsymbol{p}^N} \right) \right) \rho(\boldsymbol{x}, t). \tag{39}$$

The Green's function of this differential operator is known (Risken, 1996).

$$\rho(\boldsymbol{x}, \Delta t/2) = Z_{\gamma, \Delta t/2}^{-1} \exp\left(-\frac{(\boldsymbol{p}^N - \boldsymbol{p}_0^N e^{-\gamma \Delta t/2})^\top M^{-1} (\boldsymbol{p}^N - \boldsymbol{p}_0^N e^{-\gamma \Delta t/2})}{2 k_b T (1 - e^{-\gamma \Delta t})} \right) \delta(\boldsymbol{q}^N - \boldsymbol{q}_0^N) \tag{40}$$

$$\text{where } Z_{\gamma, \Delta t/2} = \left(2\pi k_b T (1 - e^{-\gamma \Delta t}) \right)^{\frac{3N}{2}} \sqrt{\det M}. \tag{41}$$

This probability distribution may then be sampled using the reparameterization trick with $\boldsymbol{z} \in \mathbb{R}^{3N}$ which is distributed normally (Kingma & Welling, 2022). Holding the configuration constant this gives the γ update rule as

$$\boldsymbol{p}_{\text{New}}^N = \boldsymbol{p}_{\text{Old}}^N e^{-\gamma \Delta t/2} + M^{1/2} \sqrt{k_b T (1 - e^{-\gamma \Delta t})} \boldsymbol{z} \qquad (\gamma \text{ update rule}). \tag{42}$$

By sampling we then produce another delta density at the next timestep, implying we can apply the Green's function of $\mathcal{L}_p$ rather fully solving it for a normal density. This is the trick which permits an avoidance of the full PDE solution.

To solve the next PDE in the Strang splitting chain consider the fact that both $\mathcal{L}_p$ and $\mathcal{L}_q$ are transport equations of the form $\partial_t \rho + \nabla \cdot (\boldsymbol{b}\rho) = 0$ where

$$\boldsymbol{b}_p = \begin{bmatrix} \mathbf{0}^N \\ -\nabla_{\boldsymbol{q}^N} U(\boldsymbol{q}^N) \end{bmatrix}, \quad \boldsymbol{b}_q = \begin{bmatrix} M^{-1}\boldsymbol{p}^N \\ \mathbf{0}^N \end{bmatrix}, \implies \nabla \cdot \boldsymbol{b}_p = 0, \quad \nabla \cdot \boldsymbol{b}_q = 0 \tag{43}$$

The divergence free property is due to this being Hamiltonian flow, which is volume preserving (Hansen & McDonald, 2013). This permits simplification of the PDE into $\partial_t \rho + \boldsymbol{b} \cdot \nabla \rho = 0$ The solution to a PDE of this form is a just advection. The initial condition and the form asserts the solution must be of the form $\rho(\boldsymbol{x}, \Delta t/2) = \delta(\boldsymbol{x} - (\boldsymbol{x}_0 + \boldsymbol{b}\Delta t/2))$ (Evans, 2022). Applying this gives the following update rules, noting that during the update the conjugate variable is held constant.

$$\boldsymbol{p}^N_{\text{New}} = \boldsymbol{p}^N_{\text{Old}} - \nabla_{\boldsymbol{q}^N} U(\boldsymbol{q}^N_{\text{Old}}) \tfrac{\Delta t}{2} \qquad (p \text{ update rule}) \tag{44}$$

$$\boldsymbol{q}^N_{\text{New}} = \boldsymbol{q}^N_{\text{Old}} + M^{-1}\boldsymbol{p}^N_{\text{Old}} \Delta t \qquad (q \text{ update rule}) \tag{45}$$

This fully specifies an algorithm for sampling ρ.

It can be helpful to introduce the transition densities derived in this section as *kernels* $K_{\Delta t}(\boldsymbol{x}|\boldsymbol{x}')$. This is the same object as the Green's function. In this particular case we have the final state $\boldsymbol{x} = (\boldsymbol{q}^N, \boldsymbol{p}^N)$ and initial state $\boldsymbol{x}' = (\boldsymbol{q}'^N, \boldsymbol{p}'^N)$, and three kernels total

$$K^{(\gamma)}_{\Delta t}(\boldsymbol{x}|\boldsymbol{x}') = f(\boldsymbol{p}^N_1, \boldsymbol{p}^N_0)\delta(\boldsymbol{q}^N - \boldsymbol{q}'^N), \tag{46}$$

$$K^{(p)}_{\Delta t}(\boldsymbol{x}|\boldsymbol{x}') = \delta(\boldsymbol{q}^N - \boldsymbol{q}'^N)\delta\left(\boldsymbol{p}^N - \left(\boldsymbol{p}'^N + \boldsymbol{F}(\boldsymbol{q}'^N)\Delta t\right)\right), \tag{47}$$

$$K^{(q)}_{\Delta t}(\boldsymbol{x}|\boldsymbol{x}') = \delta\left(\boldsymbol{q}^N - \left(\boldsymbol{q}'^N + M^{-1}\boldsymbol{p}'^N \Delta t\right)\right)\delta(\boldsymbol{p}^N - \boldsymbol{p}'^N), \tag{48}$$

where $\boldsymbol{F}(\boldsymbol{q}^N) = -\nabla_{\boldsymbol{q}^N} U(\boldsymbol{q}^N)$ is the force and the function f is

$$f(\boldsymbol{p}^N, \boldsymbol{p}'^N) = Z^{-1}_{\gamma, \Delta t/2} \exp\left(-\frac{(\boldsymbol{p}^N - \boldsymbol{p}'^N e^{-\gamma \Delta t/2})^\top M^{-1} (\boldsymbol{p}^N - \boldsymbol{p}'^N e^{-\gamma \Delta t/2})}{2k_b T(1 - e^{-\gamma \Delta t})}\right). \tag{49}$$

Which then lets the total kernel for $G(\Delta t)$ be

$$\begin{aligned} &K^{(G)}_{\Delta t}(\boldsymbol{x}_5|\boldsymbol{x}_0) \\ &= \int_\Omega K^{(\gamma)}_{\Delta t/2}(\boldsymbol{x}_1|\boldsymbol{x}_0) K^{(p)}_{\Delta t/2}(\boldsymbol{x}_2|\boldsymbol{x}_1) K^{(q)}_{\Delta t}(\boldsymbol{x}_3|\boldsymbol{x}_2) K^{(p)}_{\Delta t/2}(\boldsymbol{x}_4|\boldsymbol{x}_3) K^{(\gamma)}_{\Delta t/2}(\boldsymbol{x}_5|\boldsymbol{x}_4) d\boldsymbol{x}_{1\ldots4} \qquad (50) \\ &= \frac{\det M}{(\Delta t)^{3N}} f\left(\boldsymbol{p}^N_5, M\frac{\boldsymbol{q}^N_5 - \boldsymbol{q}^N_0}{\Delta t} + \boldsymbol{F}(\boldsymbol{q}^N_5)\frac{\Delta t}{2}\right) f\left(M\frac{\boldsymbol{q}^N_5 - \boldsymbol{q}^N_0}{\Delta t} - \boldsymbol{F}(\boldsymbol{q}^N_0)\frac{\Delta t}{2}, \boldsymbol{p}^N_0\right). \qquad (51) \end{aligned}$$

This expression can be derived by executing the integral and applying the sifting property of the delta function.

6.1.4 The Metropolis Correction

In the limit of infinitesimal Δt the scheme derived in section 6.1.3 will produce accurate samples as asserted by Trotter's factorization being valid in the limit of $P \to \infty$. However, in practice this is never the case as we always choose finite Δt when running a computer simulation. With finite, but small, Δt the sampling can be quite accurate in the sense that it approximately satisfies the so-called *detailed balance* (Levin & Peres, 2017). Mathematically, detailed balance implies the following

$$\forall \boldsymbol{x} \in \Omega, \forall \boldsymbol{y} \in \Omega: \ \rho(\boldsymbol{x}) K_{\Delta t}(\boldsymbol{y}|\boldsymbol{x}) = \rho(\boldsymbol{y}) K_{\Delta t}(\boldsymbol{x}|\boldsymbol{y}) \tag{52}$$

To see why satisfying eq. 52 is important, consider integrating both sides with respect to $\boldsymbol{y}$

$$\int_\Omega \rho(\boldsymbol{x}) K_{\Delta t}(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{y} = \int_\Omega \rho(\boldsymbol{y}) K_{\Delta t}(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{y} \tag{53}$$

The left hand side clearly gives just $\rho(\boldsymbol{x})$ due to $K_{\Delta t}(\boldsymbol{y}|\boldsymbol{x})$ being normalized. The right hand side is the same as computing the time evolution of the density

$$\int_\Omega \rho(\boldsymbol{y}) K_{\Delta t}(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{y} = U(\Delta t)\rho(\boldsymbol{x}). \tag{54}$$

This implies any density satisfying eq. 52 satisfies $U(\Delta t)\rho(\boldsymbol{x}) = \rho(\boldsymbol{x})$, meaning any ρ satisfying detailed balance is stationary under the dynamic. In particular the Boltzmann distribution should satisfy detailed balance. Production of a stationary density is a necessary but not sufficient assertion for an integrator to be considered good, it is also required to be *regular* (Manousiouthakis & Deem, 1999).

In our case we do not have the exact kernel $K_{\Delta t}$. Rather we have its approximation $K_{\Delta t}^{(G)}$ from the OBABO split defined in 6.1.3. We may account for this discrepancy by rewriting the kernel as a Metropolis Hastings kernel (Metropolis et al., 1953) which utilizes a new function $A(\boldsymbol{x}, \boldsymbol{y})$ which is an acceptance probability. If we accept a move then we make the transition $\boldsymbol{x} \to \boldsymbol{y}$. If we reject the move then we simply stay at $\boldsymbol{x}$. Effectively this splits the kernel into two portions (Tierney, 1998).

$$K_{\Delta t}(\boldsymbol{y}|\boldsymbol{x}) = \underbrace{K_{\Delta t}^{(G)}(\boldsymbol{y}|\boldsymbol{x})A(\boldsymbol{x},\boldsymbol{y})}_{\text{Make the transition}} + \underbrace{\delta(\boldsymbol{y}-\boldsymbol{x})\overbrace{\int_\Omega (1 - A(\boldsymbol{x},\boldsymbol{z}))K_{\Delta t}(\boldsymbol{x}|\boldsymbol{z})d\boldsymbol{z}}^{:=r(\boldsymbol{x})}}_{\text{Stay at } \boldsymbol{x}} \tag{55}$$

This then gives the detailed balance condition as

$$\rho(\boldsymbol{x})\left(K_{\Delta t}^{(G)}(\boldsymbol{y}|\boldsymbol{x})A(\boldsymbol{x},\boldsymbol{y}) + \delta(\boldsymbol{y}-\boldsymbol{x})r(\boldsymbol{x})\right) = \rho(\boldsymbol{y})\left(K_{\Delta t}^{(G)}(\boldsymbol{x}|\boldsymbol{y})A(\boldsymbol{y},\boldsymbol{x}) + \delta(\boldsymbol{x}-\boldsymbol{y})r(\boldsymbol{y})\right) \tag{56}$$

This then gives us the freedom to choose the function $A(\boldsymbol{x}, \boldsymbol{y})$ such that detailed balance is asserted.

Following the proof given by Andrieu et al. (2020), consider integrating 56 w.r.t. test functions g and h on both $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively. Each term involved is strictly positive, this allows the splitting of the single balance condition into two simpler ones. Firstly

$$\int_\Omega g(\boldsymbol{x})h(\boldsymbol{y})\delta(\boldsymbol{y}-\boldsymbol{x})r(\boldsymbol{x})\rho(\boldsymbol{x})d\boldsymbol{x}d\boldsymbol{y} = \int_\Omega g(\boldsymbol{x})h(\boldsymbol{y})\delta(\boldsymbol{x}-\boldsymbol{y})r(\boldsymbol{y})\rho(\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y} \tag{57}$$

It is clear that the delta functions make this true no matter the choice of $r(\boldsymbol{x})$, and hence no matter the choice of $A(\boldsymbol{x}, \boldsymbol{y})$. This leaves the term involving the discretized kernel.

$$\int_\Omega g(\boldsymbol{x})h(\boldsymbol{y})K_{\Delta t}^{(G)}(\boldsymbol{y}|\boldsymbol{x})A(\boldsymbol{x},\boldsymbol{y})\rho(\boldsymbol{x})d\boldsymbol{x}d\boldsymbol{y} = \int_\Omega g(\boldsymbol{x})h(\boldsymbol{y})K_{\Delta t}^{(G)}(\boldsymbol{x}|\boldsymbol{y})A(\boldsymbol{y},\boldsymbol{x})\rho(\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y} \tag{58}$$

The difficulty now is choosing $A(\boldsymbol{x}, \boldsymbol{y})$ such that the above is true.

The Metropolis Hastings ansatz for A is given by

$$A(\boldsymbol{x},\boldsymbol{y}) = \min\left(1, \frac{\rho(\boldsymbol{y})K_{\Delta t}^{(G)}(\boldsymbol{x}|\boldsymbol{y})}{\rho(\boldsymbol{x})K_{\Delta t}^{(G)}(\boldsymbol{y}|\boldsymbol{x})}\right) \tag{59}$$

This means we have

$$\begin{aligned}\int_\Omega g(\boldsymbol{x})h(\boldsymbol{y})K_{\Delta t}^{(G)}(\boldsymbol{y}|\boldsymbol{x})\rho(\boldsymbol{x})\min\left(1, \frac{\rho(\boldsymbol{y})K_{\Delta t}^{(G)}(\boldsymbol{x}|\boldsymbol{y})}{\rho(\boldsymbol{x})K_{\Delta t}^{(G)}(\boldsymbol{y}|\boldsymbol{x})}\right)d\boldsymbol{x}d\boldsymbol{y} \\ = \int_\Omega g(\boldsymbol{x})h(\boldsymbol{y})K_{\Delta t}^{(G)}(\boldsymbol{x}|\boldsymbol{y})\rho(\boldsymbol{y})\min\left(1, \frac{\rho(\boldsymbol{x})K_{\Delta t}^{(G)}(\boldsymbol{y}|\boldsymbol{x})}{\rho(\boldsymbol{y})K_{\Delta t}^{(G)}(\boldsymbol{x}|\boldsymbol{y})}\right)d\boldsymbol{x}d\boldsymbol{y},\end{aligned} \tag{60}$$

but the minimum function is linear. This implies

$$\begin{aligned}\int_\Omega g(\boldsymbol{x})h(\boldsymbol{y})\min\left(K_{\Delta t}^{(G)}(\boldsymbol{y}|\boldsymbol{x})\rho(\boldsymbol{x}), \rho(\boldsymbol{y})K_{\Delta t}^{(G)}(\boldsymbol{x}|\boldsymbol{y})\right)d\boldsymbol{x}d\boldsymbol{y} \\ = \int_\Omega g(\boldsymbol{x})h(\boldsymbol{y})\min\left(K_{\Delta t}^{(G)}(\boldsymbol{x}|\boldsymbol{y})\rho(\boldsymbol{y}), \rho(\boldsymbol{x})K_{\Delta t}^{(G)}(\boldsymbol{y}|\boldsymbol{x})\right)d\boldsymbol{x}d\boldsymbol{y},\end{aligned} \tag{61}$$

which is true for all test functions due to the minimum being symmetric. Therefore we have proven that the Metropolis Hastings kernel decomposition satisfies detailed balance. This setup implies a natural 8 step forward simulation scheme given by algorithm 1.

Algorithm 1 One integration/Metropolis step of molecular dynamics simulation

Require: Current state $\boldsymbol{x} = (\boldsymbol{q}^N, \boldsymbol{p}^N)$, step size Δt, friction γ, mass matrix M, temperature T, potential energy function U
Ensure: Updated state $\boldsymbol{x}$
$\boldsymbol{y} \leftarrow \boldsymbol{x}$ ▷ save current state
Half thermal step:

$$\boldsymbol{p}^N \leftarrow \boldsymbol{p}^N e^{-\gamma \Delta t/2} + M^{1/2}\sqrt{k_B T\big(1 - e^{-\gamma \Delta t}\big)}\, \boldsymbol{z}, \qquad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, I)$$

Half momentum step:

$$\boldsymbol{p}^N \leftarrow \boldsymbol{p}^N - \nabla_{\boldsymbol{q}^N} U(\boldsymbol{q}^N)\, \frac{\Delta t}{2}$$

Full configurational step:

$$\boldsymbol{q}^N \leftarrow \boldsymbol{q}^N + M^{-1}\boldsymbol{p}^N\, \Delta t$$

Half momentum step:

$$\boldsymbol{p}^N \leftarrow \boldsymbol{p}^N - \nabla_{\boldsymbol{q}^N} U(\boldsymbol{q}^N)\, \frac{\Delta t}{2}$$

Half thermal step:

$$\boldsymbol{p}^N \leftarrow \boldsymbol{p}^N e^{-\gamma \Delta t/2} + M^{1/2}\sqrt{k_B T\big(1 - e^{-\gamma \Delta t}\big)}\, \boldsymbol{z}, \qquad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, I)$$

Draw $\alpha \sim \mathcal{U}(0,1)$
Compute acceptance probability $A(\boldsymbol{x}, \boldsymbol{y})$
if $\alpha > A(\boldsymbol{x}, \boldsymbol{y})$ **then**
    $\boldsymbol{x} \leftarrow \boldsymbol{y}$ ▷ reject
end if

6.1.5 Failures of the Langevin Integrator and the Rare Event Problem

The question then remains. Is this enough? Why is the problem of sampling not "solved" with this algorithm? We clearly proved samples the correct density, so what is left? Unfortunately, this algorithm is only accurate in the limit of infinite steps. To better understand the issue we can analyze an example. Suppose there is a two mode probability distribution, the modes named A and B, both modes equally probable. In order to transition from mode $A \to B$ we must overcome an energy barrier, I denote this as ΔE.

The motion induced by momentum and configurational steps approximately conserve energy when Δt is small. This implies the energy change can only be due to the thermal γ updates, implying it is purely kinetic. This implies the energy change due to a sampling is given by

$$\Delta K(\boldsymbol{p}^N_{\text{New}}) = \tfrac{1}{2}(\boldsymbol{p}^N_{\text{New}} - \boldsymbol{p}^N_{\text{Old}})^\top M^{-1}(\boldsymbol{p}^N_{\text{New}} - \boldsymbol{p}^N_{\text{Old}}) \tag{62}$$

The question then is, what is the probability of such a given energy change, i.e. $\rho(\Delta E_\gamma)$, in a single half thermal step? This may be computed via an expression like

$$\rho(\Delta E_\gamma) = \int_{\mathbb{R}^{3N}} \delta(\Delta E_\gamma - \Delta K(\boldsymbol{p}^N_{\text{New}}))\mathcal{N}(\boldsymbol{p}^N_{\text{New}} | \boldsymbol{p}_{\text{Old}} e^{-\gamma \Delta t/2}, M k_b T(1 - e^{-\gamma \Delta t}))d\boldsymbol{p}^N_{\text{New}}. \tag{63}$$

Looking closely we may re-parameterize like

$$\boldsymbol{y} = f(\boldsymbol{p}_{\text{New}}) = \lambda^{-1/2} M^{-1/2}(\boldsymbol{p}^N_{\text{New}} - \boldsymbol{p}^N_{\text{Old}}) \text{ where } \lambda = k_b T(1 - e^{-\gamma \Delta t}) \tag{64}$$

$$\implies \boldsymbol{p}^N_{\text{New}} = \boldsymbol{p}^{3N}_{\text{Old}} + \lambda^{1/2} M^{1/2}\boldsymbol{y}, \quad \det\left|\frac{\partial f^{-1}}{\partial \boldsymbol{y}}\right| = \lambda^{3N/2} \det|M|^{1/2} \tag{65}$$

Then this gives

$$\rho(\Delta E_\gamma) = \int_{\mathbb{R}^{3N}} \delta(\Delta E_\gamma - \tfrac{\lambda}{2}\boldsymbol{y}^2)\mathcal{N}(\boldsymbol{y} | \boldsymbol{\mu}, \boldsymbol{I}_{3N})d\boldsymbol{y} \tag{66}$$

where $\boldsymbol{\mu} = (e^{-\gamma\Delta t/2} - 1)\lambda^{-1/2}M^{-1/2}\boldsymbol{p}_{\text{Old}}$. Which clearly gives the non-central χ'^2 statistic with $3N$ degrees of freedom. This means we have

$$\frac{2\beta\Delta E_\gamma}{1 - e^{-\gamma\Delta t}} \sim \chi'^2_{3N}\left(\beta\frac{1 - e^{-\gamma\Delta t/2}}{1 + e^{-\gamma\Delta t/2}}\boldsymbol{p}_{\text{Old}}^\top M^{-1}\boldsymbol{p}_{\text{Old}}\right). \tag{67}$$

We then want to investigate the limit as $\Delta t \to 0$, which is the limit of accurate sampling. In such a limit one can show that $\mathbb{E}[\Delta E_\gamma] \to 0$ with variance Δt^2. Such a limit implies that the ability to obtain energy and the ability to have accurate sampling are in direct competition with one another.

Albeit, we do not (and should not) be expecting the sampler to overcome a barrier in one step. Hence, this analysis is flawed in that we should allow for a chain of states rather than just one. Even so, as a demonstrative example this emphasizes interplay between the timestep and the energy. A full analysis of this phenomena has been well studied by Eyring (1935) and Kramers (1940). Their analysis yields the exit rate k in the small temperature regime $T \to 0$ as

$$k = v\exp(-\beta\Delta E) \tag{68}$$

where v is a geometric factor involving the curvature of the barrier between $A \to B$, hence is an entropic factor. In particular (Lelièvre, 2018), shows their results again with modern rigor. This poses the rare event problem on firm mathematical ground, implying the need for methodology which is not limited by factors like Δt.

6.2 ESM AND NESM TRAINING ALGORITHMS

Algorithm 2 Training a ESM with a velocity flow model.

Require: noise $\boldsymbol{z}$, dataset $\mathcal{D}$, velocity model $\boldsymbol{v}_\theta(\boldsymbol{x}, t)$

for training step $k = 1, 2, \ldots$ **do**
  Sample $\{\boldsymbol{x}\}_{i=1}^B \sim \mathcal{D}$
  Sample $\{\boldsymbol{z}\}_{i=1}^B \sim \mathcal{N}(\mathbf{0}, I)$
  Sample $\{t\}_{i=1}^B \sim \mathcal{U}(0, 1)$
  $\boldsymbol{x}_t \leftarrow (1 - t)\boldsymbol{z} + t\,\boldsymbol{x}$
  $\boldsymbol{u} \leftarrow \dfrac{\boldsymbol{x} - \boldsymbol{x}_t}{1 - t}$ ▷ target velocity
  $\mathcal{L} \leftarrow \frac{1}{B}\sum_{i=1}^B \|\boldsymbol{v}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{u}\|_2^2$
  $\theta \leftarrow \textsc{Opt}(\theta, \nabla_\theta\mathcal{L})$
end for

Algorithm 3 Training a NESM with conditional velocity flow model.

Require: noise $\boldsymbol{z}$, dataset $\mathcal{D}$, conditional velocity model $\boldsymbol{v}_\theta(\boldsymbol{x}_{t_{\text{Flow}}} t_{\text{Flow}} | t_{\text{Phys}}, \boldsymbol{x}_0)$

for training step $k = 1, 2, \ldots$ **do**
  Sample $\{\boldsymbol{x}_0\}_{i=1}^B \sim \mathcal{D}$
  Sample $\{t_{\text{phys}}\}_{i=1}^B \sim \mathcal{U}(0, T_{\max})$
  Simulate equation 1 for t_{phys} units of time.
  $\boldsymbol{x} \leftarrow$ final simulation state.
  Sample $\{\boldsymbol{z}\}_{i=1}^B \sim \mathcal{N}(\mathbf{0}, I)$
  Sample $\{t_{\text{Flow}}\}_{i=1}^B \sim \mathcal{U}(0, 1)$
  $\boldsymbol{x}_{t_{\text{Flow}}} \leftarrow (1 - t_{\text{Flow}})\boldsymbol{z} + t_{\text{Flow}}\,\boldsymbol{x}$
  $\boldsymbol{u} \leftarrow \dfrac{\boldsymbol{x} - \boldsymbol{x}_{t_{\text{Flow}}}}{1 - t_{\text{Flow}}}$ ▷ target velocity
  $\mathcal{L} \leftarrow \frac{1}{B}\sum_{i=1}^B \|\boldsymbol{v}_\theta(\boldsymbol{x}_{t_{\text{Flow}}} t_{\text{Flow}} | t_{\text{Phys}}, \boldsymbol{x}_0) - \boldsymbol{u}\|_2^2$
  $\theta \leftarrow \textsc{Opt}(\theta, \nabla_\theta\mathcal{L})$
end for

6.3 NON-EQUIVARIANT GRAPH TRANSFORMER RESULTS

Tensor	Status	Mean (Å)	Std. (Å)	Max (Å)	Tol. (Å)	$\|\cdot\|_{\text{before}}$	$\|\cdot\|_{\text{after}}$
f_{cond}	FAIL	0.2716	0.3227	2.291	1×10^{-3}	4.855	4.81
$\boldsymbol{b}$	FAIL	0.07963	0.06793	0.4004	1×10^{-3}	1.323	1.309
$\boldsymbol{\eta}$	FAIL	0.02949	0.03209	0.2044	1×10^{-3}	0.2444	0.2479

Table 2: Summary statistics of equivariance error in equation 21 (in Å) computed over 128 samples from the test dataset. This particular table was for a non-equivariant graph transformer.

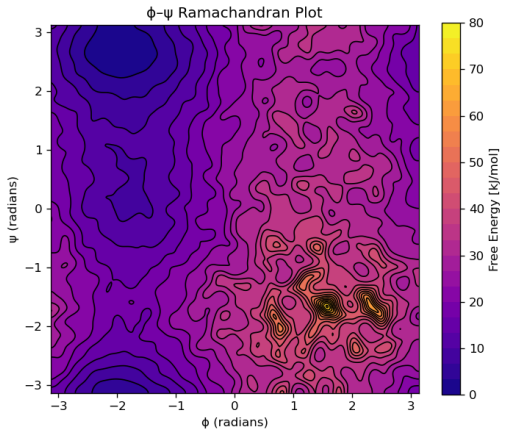

Figure 4: Ramachandran free-energy surface (kJ/mol) for alanine dipeptide generated by a non-equivariant graph transformer.

CONTRIBUTIONS

Harry Winston Sullivan: Added SE(3) transformer, Equiformer, new message passing network, and Lightning module codebase; created training, sampling, and equivariance testing scripts; implemented ADP dataset/dataloader, interpolant class, and prior sampler; created architecture figure and README; wrote SI.

Zichen Huang: Implemented the non-equivariant graph transformer; helped debug the dihedral indexing issue and the O(3) vs SO(3) bug (very difficult to spot); coded dihedral calculation and a Hungarian-algorithm filter for results (ultimately unused); attempted an Onsager–Machlup path sampler; modified experiment configurations; ran experiments (training/sampling/equivariance checks); monitored training runs and collected logs.

Both: Analysis of results; wrote and refined the report.

REFERENCES

Michael Albergo and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08739*, 2023.

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic Interpolants: A Unifying Framework for Flows and Diffusions, November 2023. URL http://arxiv.org/abs/2303.08797. arXiv:2303.08797 [cs].

Rosalind J Allen, Chantal Valeriani, and Pieter Rein Ten Wolde. Forward flux sampling for rare event simulations. *Journal of Physics: Condensed Matter*, 21(46):463102, November 2009. ISSN 0953-8984, 1361-648X. doi: 10.1088/0953-8984/21/46/463102. URL https://iopscience.iop.org/article/10.1088/0953-8984/21/46/463102.

Christophe Andrieu, Anthony Lee, and Sam Livingstone. A general perspective on the Metropolis-Hastings kernel, December 2020. URL http://arxiv.org/abs/2012.14881. arXiv:2012.14881 [stat].

Marloes Arts, Victor Garcia Satorras, Chin-Wei Huang, Daniel Zuegner, Marco Federici, Cecilia Clementi, Frank Noé, Robert Pinsler, and Rianne van den Berg. Two for One: Diffusion Models and Force Fields for Coarse-Grained Molecular Dynamics, September 2023. URL http://arxiv.org/abs/2302.00600. arXiv:2302.00600 [cs].

Peter G. Bolhuis and Christoph Dellago. *Trajectory-Based Rare Event Simulations*, chapter 3, pp. 111–210. John Wiley Sons, Ltd, 2010. ISBN 9780470890905. doi: https://doi.org/10.1002/9780470890905.ch3. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470890905.ch3.

Giovanni Bussi and Michele Parrinello. Accurate sampling using Langevin dynamics. *Physical Review E*, 75(5):056707, May 2007. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.75.056707. URL http://arxiv.org/abs/0803.4083. arXiv:0803.4083 [physics].

Mirosław Antoni Czarnecki, Yusuke Morisawa, Yoshisuke Futami, and Yukihiro Ozaki. Advances in Molecular Structure and Interaction Studies Using Near-Infrared Spectroscopy. *Chem. Rev.*, 115(18):9707–9744, September 2015. ISSN 0009-2665. doi: 10.1021/cr500013u. URL https://doi.org/10.1021/cr500013u.

Christoph Dellago, Peter G. Bolhuis, and Phillip L. Geissler. Transition Path Sampling. In I. Prigogine and Stuart A. Rice (eds.), *Advances in Chemical Physics*, volume 123, pp. 1–78. Wiley, 1 edition, July 2002. ISBN 978-0-471-21453-3 978-0-471-23150-9. doi: 10.1002/0471231509.ch1. URL https://onlinelibrary.wiley.com/doi/10.1002/0471231509.ch1.

Juan Viguera Diez, Mathias Schreiner, Ola Engkvist, and Simon Olsson. Boltzmann priors for Implicit Transfer Operators, October 2024. URL http://arxiv.org/abs/2410.10605. arXiv:2410.10605 [physics].

Yuanqi Du, Michael Plainer, Rob Brekelmans, Chenru Duan, Frank Noé, Carla P. Gomes, Alán Aspuru-Guzik, and Kirill Neklyudov. Doob's Lagrangian: A Sample-Efficient Variational Approach to Transition Path Sampling, December 2024. URL http://arxiv.org/abs/2410.07974. arXiv:2410.07974 [cs].

Alexandre Duval, Simon V. Mathis, Chaitanya K. Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D. Malliaros, Taco Cohen, Pietro Liò, Yoshua Bengio, and Michael Bronstein. A Hitchhiker's Guide to Geometric GNNs for 3D Atomic Systems, March 2024. URL http://arxiv.org/abs/2312.07511. arXiv:2312.07511 [cs].

Lawrence C. Evans. *Partial differential equations*. Number 19 in Graduate studies in mathematics. American Mathematical Society, Providence, Rhode Island, second edition edition, 2022. ISBN 978-1-4704-6942-9.

Henry Eyring. The Activated Complex in Chemical Reactions. *The Journal of Chemical Physics*, 3 (2):107–115, February 1935. ISSN 0021-9606. doi: 10.1063/1.1749604. URL https://doi.org/10.1063/1.1749604.

Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman. emcee: The mcmc hammer - autocorrelation analysis tutorial, 2024. URL https://emcee.readthedocs.io/en/stable/tutorials/autocorr/. Accessed: 2024.

Michael Gastegger, Jörg Behler, and Philipp Marquetand. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.*, 8(10):6924–6935, 2017. doi: 10.1039/C7SC02267K. URL https://pubs.rsc.org/en/content/articlelanding/2017/sc/c7sc02267k.

Jean-Pierre Hansen and I. R. McDonald. *Theory of Simple Liquids: with Applications to Soft Matter*. Academic Press, San Diego, August 2013. ISBN 978-0-12-387033-9.

Jiajun He, Yuanqi Du, Francisco Vargas, Yuanqing Wang, Carla P. Gomes, José Miguel Hernández-Lobato, and Eric Vanden-Eijnden. FEAT: Free energy Estimators with Adaptive Transport, April 2025. URL http://arxiv.org/abs/2504.11516. arXiv:2504.11516 [stat].

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models, December 2020. URL http://arxiv.org/abs/2006.11239. arXiv:2006.11239 [cs].

Saman Hosseinpour, Steven J. Roeters, Mischa Bonn, Wolfgang Peukert, Sander Woutersen, and Tobias Weidner. Structure and Dynamics of Interfacial Peptides and Proteins from Vibrational Sum-Frequency Generation Spectroscopy. *Chem. Rev.*, 120(7):3420–3465, April 2020. ISSN 0009-2665. doi: 10.1021/acs.chemrev.9b00410. URL https://doi.org/10.1021/acs.chemrev.9b00410.

C. Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78:2690–2693, Apr 1997. doi: 10.1103/PhysRevLett.78.2690. URL https://link.aps.org/doi/10.1103/PhysRevLett.78.2690.

Michael S. Jones, Kirill Shmilovich, and Andrew L. Ferguson. Tutorial on Molecular Latent Space Simulators (LSSs): Spatially and Temporally Continuous Data-Driven Surrogate Dynamical Models of Molecular Systems. *The Journal of Physical Chemistry A*, 128(47):10299–10317, November 2024. ISSN 1089-5639, 1520-5215. doi: 10.1021/acs.jpca.4c05389. URL https://pubs.acs.org/doi/10.1021/acs.jpca.4c05389.

Stefanie Kieninger and Bettina G. Keller. GROMACS Stochastic Dynamics and BAOAB Are Equivalent Configurational Sampling Algorithms. *Journal of Chemical Theory and Computation*, 18 (10):5792–5798, October 2022. ISSN 1549-9618, 1549-9626. doi: 10.1021/acs.jctc.2c00585. URL https://pubs.acs.org/doi/10.1021/acs.jctc.2c00585.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes, December 2022. URL http://arxiv.org/abs/1312.6114. arXiv:1312.6114 [stat].

Leon Klein and Frank Noé. Transferable Boltzmann Generators, February 2025. URL http://arxiv.org/abs/2406.14426. arXiv:2406.14426 [stat].

Leon Klein, Andrew Y. K. Foong, Tor Erlend Fjelde, Bruno Mlodozeniec, Marc Brockschmidt, Sebastian Nowozin, Frank Noé, and Ryota Tomioka. Timewarp: Transferable Acceleration of Molecular Dynamics by Learning Time-Coarsened Dynamics. In *NeurIPS 2023*, February 2023.

H. A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, April 1940. ISSN 0031-8914. doi: 10.1016/S0031-8914(40)90098-2. URL https://www.sciencedirect.com/science/article/pii/S0031891440900982.

Jonas Köhler, Leon Klein, and Frank Noé. Equivariant Flows: Exact Likelihood Generative Learning for Symmetric Densities, October 2020. URL http://arxiv.org/abs/2006.02425. arXiv:2006.02425 [stat].

B. Leimkuhler and C. Matthews. Rational Construction of Stochastic Numerical Methods for Molecular Sampling. *Applied Mathematics Research eXpress*, pp. abs010, June 2012. ISSN 1687-1200, 1687-1197. doi: 10.1093/amrx/abs010. URL https://academic.oup.com/amrx/article-lookup/doi/10.1093/amrx/abs010.

Ben Leimkuhler, Daniel T. Margul, and Mark E. Tuckerman. Stochastic, resonance-free multiple time-step algorithm for molecular dynamics with very large time steps. *Molecular Physics*, 111(22-23):3579–3594, December 2013. ISSN 0026-8976, 1362-3028. doi: 10.1080/00268976.2013.844369. URL http://www.tandfonline.com/doi/abs/10.1080/00268976.2013.844369.

Tony Lelièvre. Mathematical foundations of Accelerated Molecular Dynamics methods, January 2018. URL http://arxiv.org/abs/1801.05347. arXiv:1801.05347 [math].

Don Lemons and Anthony Gythiel. Paul Langevin's 1908 paper "On the Theory of Brownian Motion". *American Journal of Physics - AMER J PHYS*, 65:1079–1081, January 1997.

David Asher Levin and Yuval Peres. *Markov chains and mixing times*. Number v. 107 in AMS Non-Series Monographs. American Mathematical Society, Providence, Rhode Island, second edition edition, 2017. ISBN 978-1-4704-2962-1 978-1-4704-4232-3.

Peng Li, Yaling Jiang, Youcheng Hu, Yana Men, Yuwen Liu, Wenbin Cai, and Shengli Chen. Hydrogen bond network connectivity in the electric double layer dominates the kinetic pH effect in hydrogen electrocatalysis on Pt. *Nat. Catal*, 5(10):900–911, October 2022. ISSN 2520-1158. doi: 10.1038/s41929-022-00846-8. URL https://www.nature.com/articles/s41929-022-00846-8.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling, February 2023. URL http://arxiv.org/abs/2210.02747. arXiv:2210.02747 [cs].

Yikai Liu, Ming Chen, and Guang Lin. Backdiff: a diffusion model for generalized transferable protein backmapping, November 2023. URL http://arxiv.org/abs/2310.01768. arXiv:2310.01768 [q-bio].

Vasilios I. Manousiouthakis and Michael W. Deem. Strict Detailed Balance is Unnecessary in Monte Carlo Simulation. *The Journal of Chemical Physics*, 110(6):2753–2756, February 1999. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.477973. URL http://arxiv.org/abs/cond-mat/9809240. arXiv:cond-mat/9809240.

Markov Modeling Group. mdshare: Molecular dynamics data repository. https://github.com/markovmodel/mdshare, 2024. See also https://markovmodel.github.io/mdshare/. Accessed: 2025-10-16.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models, February 2021. URL http://arxiv.org/abs/2102.13219. arXiv:2102.13219 [stat].

Wenting Meng, Hao-Che Peng, Yuanhao Liu, Allison Stelling, and Lu Wang. Modeling the Infrared Spectroscopy of Oligonucleotides with 13C Isotope Labels. *J. Phys. Chem. B*, 127(11):2351–2361, March 2023. ISSN 1520-6106. doi: 10.1021/acs.jpcb.2c08915. URL https://doi.org/10.1021/acs.jpcb.2c08915.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.1699114. URL https://pubs.aip.org/jcp/article/21/6/1087/202680/Equation-of-State-Calculations-by-Fast-Computing.

Mor Mishkovsky and Lucio Frydman. Principles and Progress in Ultrafast Multidimensional Nuclear Magnetic Resonance. *Annu. Rev. Phys. Chem.*, 60(1):429–448, 2009. doi: 10.1146/annurev.physchem.040808.090420. URL https://doi.org/10.1146/annurev.physchem.040808.090420.

Jacob I. Monroe and Vincent K. Shen. Learning Efficient, Collective Monte Carlo Moves with Variational Autoencoders. *Journal of Chemical Theory and Computation*, 18(6):3622–3636, June 2022. ISSN 1549-9618. doi: 10.1021/acs.jctc.2c00110. URL https://doi.org/10.1021/acs.jctc.2c00110. Publisher: American Chemical Society.

Isaac Newton. *Philosophiæ Naturalis Principia Mathematica*. Jussu Societatis Regiæ, Londini, 1687. Typis Josephi Streater.

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, September 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaw1147. URL https://www.science.org/doi/10.1126/science.aaw1147.

L. Onsager and S. Machlup. Fluctuations and irreversible processes. *Phys. Rev.*, 91:1505–1512, Sep 1953. doi: 10.1103/PhysRev.91.1505. URL https://link.aps.org/doi/10.1103/PhysRev.91.1505.

Michael Plainer, Hao Wu, Leon Klein, Stephan Günnemann, and Frank Noé. Consistent Sampling and Simulation: Molecular Dynamics with Energy-Based Diffusion Models, November 2025. URL http://arxiv.org/abs/2506.17139. arXiv:2506.17139 [cs].

Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D. Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *The Journal of Chemical Physics*, 134 (17):174105, May 2011. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.3565032. URL https://pubs.aip.org/jcp/article/134/17/174105/699460/Markov-models-of-molecular-kinetics-Generation-and.

Sanjeev Raja, Martin Šípka, Michael Psenka, Tobias Kreiman, Michal Pavelka, and Aditi S Krishnapriyan. Action-Minimization Meets Generative Modeling: Efficient Transition Path Sampling with the Onsager-Machlup Functional.

Hannes Risken. *The Fokker-Planck Equation: Methods of Solution and Applications*, volume 18 of *Springer Series in Synergetics*. Springer, Berlin, Heidelberg, 1996. ISBN 978-3-540-61530-9 978-3-642-61544-3. doi: 10.1007/978-3-642-61544-3. URL https://link.springer.com/10.1007/978-3-642-61544-3.

Charles A. Schmuttenmaer. Exploring Dynamics in the Far-Infrared with Terahertz Spectroscopy. *Chem. Rev.*, 104(4):1759–1780, April 2004. ISSN 0009-2665. doi: 10.1021/cr020685g. URL https://doi.org/10.1021/cr020685g.

Mathias Schreiner, Ole Winther, and Simon Olsson. Implicit Transfer Operator Learning: Multiple Time-Resolution Surrogates for Molecular Dynamics, October 2023. URL http://arxiv.org/abs/2305.18046. arXiv:2305.18046 [physics, stat].

Erwin Schrödinger. *Space-Time Structure*. Cambridge University Press, 1 edition, October 1985. ISBN 978-0-521-31520-3 978-0-511-58644-6. doi: 10.1017/CBO9780511586446. URL https://www.cambridge.org/core/product/identifier/9780511586446/type/book.

Kristof T. Schütt, Oliver T. Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra, June 2021. URL http://arxiv.org/abs/2102.03150. arXiv:2102.03150 [physics].

M. Scott Shell. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *The Journal of chemical physics*, 129(14), 2008. URL https://pubs.aip.org/aip/jcp/article/129/14/144108/187242. Publisher: AIP Publishing.

Gilbert Strang. On the Construction and Comparison of Difference Schemes. *SIAM Journal on Numerical Analysis*, 5(3):506–517, September 1968. ISSN 0036-1429, 1095-7170. doi: 10.1137/0705041. URL http://epubs.siam.org/doi/10.1137/0705041.

Robert H. Swendsen and Jian-Sheng Wang. Replica monte carlo simulation of spin-glasses. *Phys. Rev. Lett.*, 57:2607–2609, Nov 1986. doi: 10.1103/PhysRevLett.57.2607. URL https://link.aps.org/doi/10.1103/PhysRevLett.57.2607.

Luke Tierney. A Note on Metropolis-Hastings Kernels for General State Spaces. *The Annals of Applied Probability*, 8(1):1–9, 1998. ISSN 1050-5164. URL https://www.jstor.org/stable/2667233. Publisher: Institute of Mathematical Statistics.

H. F. Trotter. On the product of semi-groups of operators. In *Proceedings of the American Mathematical Society*, volume 10, pp. 545–551, August 1959. doi: 10.1090/S0002-9939-1959-0108732-6. URL https://www.ams.org/proc/1959-010-04/S0002-9939-1959-0108732-6/. ISSN: 0002-9939, 1088-6826 Issue: 4 Journal Abbreviation: Proc. Amer. Math. Soc.

M. Tuckerman, B. J. Berne, and G. J. Martyna. Reversible multiple time scale molecular dynamics. *The Journal of Chemical Physics*, 97(3):1990–2001, August 1992. ISSN 0021-9606. doi: 10.1063/1.463137. URL https://doi.org/10.1063/1.463137.

Loup Verlet. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review*, 159(1):98–103, July 1967. doi: 10.1103/PhysRev.159.98. URL https://link.aps.org/doi/10.1103/PhysRev.159.98. Publisher: American Physical Society.

Tao Wang, Zhongliang Tian, Zihan You, Zheng Li, Hao Cheng, Wenzhang Li, Yahui Yang, Yangen Zhou, Qifan Zhong, and Yanqing Lai. Hydrogen-bond network manipulation of aqueous electrolytes with high-donor solvent additives for Al-air batteries. *Energy Storage Mater.*, 45:24–32, March 2022. ISSN 2405-8297. doi: 10.1016/j.ensm.2021.11.030. URL https://www.sciencedirect.com/science/article/pii/S240582972100547X.