THEORETICAL LIMITS AND SAMPLING GUARANTEES FOR MACHINE-LEARNED INTERATOMIC POTENTIALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine-learned interatomic potentials (MLIPs) accelerate molecular dynamics (MD) simulations by learning the energies and forces that govern atomic motion. Recent work shows that relaxing physical inductive biases can improve computational efficiency and predictive accuracy, raising the question of whether energy-and-force matching alone ensures accurate sampling. We show that the relationship between prediction error and sampling accuracy is exponentially ill-conditioned in dimension. In particular, prediction errors that decay exponentially with dimension can still coexist with nearly maximal total variation distance between the learned and target distributions. We identify two distinct failure mechanisms. The first is kinetic trapping: relaxation to an accurate learned equilibrium requires exponentially long times, even when simulations start from training configurations whose energies and forces are matched exactly. The second is equilibrium mismatch: the learned stationary distribution itself is far from the target. Under explicit assumptions, we show that equilibrium mismatch can persist in local, symmetry-preserving models, where small per-atom errors coexist with a distributional discrepancy that approaches its maximum as system size increases. Finally, we establish global energy-control conditions that ensure equilibrium accuracy and identify additional conditions needed for rapid relaxation.

1 INTRODUCTION

Machine-learned interatomic potentials (MLIPs) have emerged as a prominent application of scientific machine learning, enabling molecular simulations at a small fraction of the cost of electronic structure methods. They are trained to match energies and forces on configurations drawn from expensive quantum mechanical calculations, and are subsequently deployed in classical molecular dynamics (MD) simulations to sample the equilibrium Boltzmann distribution associated with the underlying potential energy surface. The quantities of scientific interest, including free energies, transport coefficients, and phase behavior, are functionals of this equilibrium distribution. Consequently, the practical utility of an MLIP depends not only on its accuracy in predicting energies and forces, but also on whether the resulting dynamics sample the correct distribution.

A growing body of empirical work reports that low test-set force and energy errors do not translate into reliable simulations or accurate physical property predictions. Fu et al. (2023) find that force accuracy is poorly correlated with MD stability across state-of-the-art models, while Mannan et al. (2026) report a similar "reality gap" with respect to experimental measurements. Similar failures have been observed for phonon spectra and vibrational and thermal properties (Loew et al., 2025; Deng et al., 2025), particularly for scalable direct-force models that do not enforce energy conservation (Bigi et al., 2025; Fu et al., 2025).

These observations raise a natural question: *Does minimizing the energy-and-force matching objective determine the distribution sampled by the induced molecular dynamics?* The question is further sharpened by a recent trend showing that relaxing physically motivated architectural constraints can improve efficiency and modeling capabilities. Whether optimizing an unconstrained learning objective can be sufficient at scale without additional architectural constraints or physical priors is one of the central open questions in the field (Creed et al., 2026). Existing evidence, however, cannot resolve it, for three reasons. First, it is empirical and tied to specific architectures, datasets, and benchmark tasks, making the observed failures easy to attribute to insufficient data or limited model

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

complexity. Second, it treats energy and force errors as primary evaluation metrics and reaches mixed conclusions about their correlation with downstream performance. Third, it assesses simulations through a small collection of low-dimensional observables, which can agree across trajectories even when the underlying equilibrium distributions differ substantially. Consequently, it remains unclear whether unconstrained energy-and-force matching alone is sufficient to guarantee the correct equilibrium distribution, or whether additional constraints are required.

Our work addresses this gap by studying sampling guarantees and theoretically characterizing when energy-and-force matching ensures accurate sampling. Our main contributions are as follows:

- **Exponential ill-conditioning for sampling accuracy:** We show that the relationship between prediction error and sampling accuracy is exponentially ill-conditioned in dimension. In particular, arbitrarily small population energy and force errors can coexist with nearly maximal distributional error, establishing a fundamental limitation of unconstrained energy-and-force matching.
- **High-dimensional failure mechanisms:** We identify two sampling failure modes, kinetic trapping (Theorems 4.1 and 4.2) and equilibrium mismatch (Theorem 4.3), and prove that they can persist under exponentially small prediction errors, exact interpolation and initialization at training configurations.
- **Roles and limitations of structural priors:** We extend our analysis to conservative, local, permutation-invariant, and $E(3)$-symmetric hypothesis classes, characterizing the sampling guarantees and limitations of each structural prior (Table 2).
- **Sufficient conditions for accurate sampling:** We identify global energy-control conditions (Theorem 4.6) that guarantee accurate equilibrium sampling and stronger conditions needed for rapid relaxation.

To our knowledge, this work provides the first formal analysis of the robustness of MLIP-driven MD simulations to energy-and-force matching errors. Our results are adversarial/worst-case in nature. They do not imply that any particular MLIP fails. Rather, they show that the learning objective does not preclude such failures. Because these failures are established through explicit constructions, they reflect a limitation of energy-and-force matching rather than of any particular architecture or dataset. Our results therefore suggest that additional constraints beyond energy-and-force matching may be important for reliable MD sampling. While our work is primarily theoretical, we include a small set of controlled experiments on an Einstein crystal and FCC argon in Appendix A to illustrate the theory.

2 Related work

Force matching is a standard route from reference calculations to machine-learned potentials and underpins modern MLIPs (Behler, 2021; Zhang et al., 2018; Schütt et al., 2021; Batzner et al., 2022; Batatia et al., 2022; Musaelian et al., 2023; Unke et al., 2021). It is closely related to score matching because the score of a Boltzmann distribution is proportional to its force field. This connects our question to robustness analyses of score-based generative models, where a sharp contrast is known: diffusion models are robust to appropriately controlled L^2 errors in their time-dependent scores (Chen et al., 2023; Benton et al., 2024; Li et al., 2024; Liang et al., 2025), whereas Cao et al. (2026) recently showed that overdamped Langevin dynamics need not be robust to L^2 score error.

Our Theorems 4.1 and 4.2 extend their results (Theorems 1 and 7) to second-order (kinetic) Langevin dynamics used in molecular simulations. The proofs require a different, hypoelliptic exit-time argument because inertia fundamentally changes the short-time dynamics (see Remark D.2). Our setting also requires control of energy errors and constrains the learned force field to be the negative gradient of a scalar potential. Because MLIPs are trained on reference energies and forces, we additionally establish control of the corresponding empirical errors.

Beyond the kinetic extension, our analysis establishes several additional results. Theorem 4.3 strengthens the asymptotic separation from $1-\varepsilon$, for a constant ε fixed in advance, to $1-e^{-\Omega(d)}$, while preserving conservativity and a well-defined Gibbs equilibrium. Theorem 4.6 provides global energy-control conditions for equilibrium accuracy and additional conditions for rapid relaxation. Under explicit assumptions, Appendix H extends the equilibrium-mismatch result to a local, $E(3)$-invariant hypothesis class. The kinetic failures we construct are metastability phenomena (Bovier

et al., 2004; 2005; Menz & Schlichting, 2014), induced by estimation error rather than inherited from the target potential.

3 BACKGROUND

3.1 SETUP AND NOTATION

We consider a system of N atoms with $d = 3N$ configurational degrees of freedom. Positions and velocities are $x, v \in \mathbb{R}^d$, and $\|\cdot\|$ denotes the Euclidean norm. Throughout, we adopt unit atomic masses. Heterogeneous masses can be absorbed by the coordinate rescaling $x \mapsto M^{1/2}x$. We write I_d for the identity matrix, $A \succ 0$ ($A \succeq 0$) for positive (semi-)definiteness, $\lambda_{\min}(A)$ for the smallest eigenvalue of a symmetric matrix, and $\mathcal{N}(m, \Sigma)$ for the Gaussian with mean m and covariance Σ.

The reference potential energy is $V : \mathbb{R}^d \to \mathbb{R}$ with force field $F = -\nabla V$, and at inverse temperature β the target Boltzmann distribution is $\pi_{\mathrm{tar}} \propto e^{-\beta V}$. A machine-learned interatomic potential (MLIP) is a surrogate $\hat{V}_\theta$ with associated force field $\hat{F}_\theta = -\nabla \hat{V}_\theta$, obtained by minimizing the joint energy-and-force-matching objective

$$\theta \in \arg\min_\theta \ \mathbb{E}_{x\sim\pi_{\mathrm{tar}}}\Big[\ \Big|\hat{V}_\theta(x) - V(x)\Big|^2 + \lambda\ \Big\|\hat{F}_\theta(x) - F(x)\Big\|^2\ \Big], \tag{1}$$

where $\lambda > 0$ balances the two loss components. We suppress the subscript θ when the parameterization is immaterial. Equation (1) is the ideal population objective with respect to the target distribution. Our analysis considers surrogates with arbitrarily small population loss, including constructions that exactly interpolate the empirical training data.

The Boltzmann distribution of the learned potential is $\hat{\pi} \propto e^{-\beta \hat{V}}$, and we write $\hat{\pi}_T$ for the position marginal of the distribution of (x_t, v_t) at time T under dynamics driven by $\hat{F}$ from a specified initialization. Here $\hat{\pi}_T$ denotes the position distribution at time T, whereas $\hat{\pi}$ denotes the stationary distribution induced by the learned potential. We measure sampling error by the total variation distance between position marginals, $\mathrm{TV}(\rho, \rho') = \sup_A |\rho(A) - \rho'(A)| \in [0, 1]$, and approximation error by the L^p norm, $\left\|\hat{F} - F\right\|_{L^p} = \left(\mathbb{E}_{x\sim\pi_{\mathrm{tar}}}\left\|\hat{F}(x) - F(x)\right\|^p\right)^{1/p}$ for $p \geq 1$. Note that the training objective (1) controls the case $p = 2$, while our results hold for every finite $p \geq 1$.

3.2 ENSEMBLE SAMPLING

The thermodynamic state of an MD simulation is set by the ensemble it samples. We focus on the widely used canonical (NVT) ensemble, in which thermodynamic properties of interest are functionals of the equilibrium measure (Lelièvre & Stoltz, 2016). Many thermostat algorithms exist (Leimkuhler & Matthews, 2015). We focus on the standard Langevin thermostat for clarity. The same mechanisms are expected to arise more broadly across ergodic NVT and NPT sampling schemes, although we do not formalize those extensions here. Under Langevin thermostat and our unit-mass convention, the dynamics are governed by

$$\mathrm{d}x_t = v_t\,\mathrm{d}t, \qquad \mathrm{d}v_t = F(x_t)\,\mathrm{d}t - \gamma v_t\,\mathrm{d}t + \sqrt{2\gamma\beta^{-1}}\,\mathrm{d}B_t, \tag{2}$$

where B_t is a standard Brownian motion and $\gamma > 0$ is the friction coefficient.

Under mild regularity conditions, (2) is ergodic with unique stationary distribution given by the Gibbs measure on phase space, whose position marginal is π_{tar} and whose velocity marginal is an independent Maxwell-Boltzmann distribution $\mathcal{N}(0, \beta^{-1}I_d)$. Consequently, long-time averages of observables converge to their canonical ensemble averages. In practice, however, simulations are finite and the relevant questions are therefore whether the learned dynamics converge to the correct equilibrium and, if so, how rapidly they do so. These are quantified by the stationary distribution induced by the learned force field and its mixing time, defined as the time required for the distribution of (x_t, v_t) to approach stationarity to a prescribed accuracy.

3.3 HARMONIC CRYSTAL POTENTIAL

To isolate failures arising from the learning objective rather than from the complexity of the underlying potential energy surface, we center our analysis on the harmonic crystal and then generalize our results to a broader class of target potentials.

The harmonic crystal is the simplest nontrivial model of a crystalline solid. It is smooth, strongly convex, and exactly solvable. It is obtained by expanding a smooth potential to second order about a stable equilibrium μ:

$$V(x) = \tfrac{1}{2}(x-\mu)^\top K(x-\mu), \qquad F(x) = -K(x-\mu), \tag{3}$$

where $K = \nabla^2 V(\mu) \succ 0$ is the force-constant matrix and its eigenpairs are the normal modes (phonons). The Boltzmann distribution is Gaussian, $\pi_{\mathrm{tar}} = \mathcal{N}(\mu, (\beta K)^{-1})$, and is therefore strongly log-concave. Classical results imply that Langevin dynamics with arbitrary initialization converges to equilibrium in time polynomial in the dimension (Bakry & Émery, 1985; Villani, 2009; Bakry et al., 2014; Bobkov, 1999).

The isotropic case $K = kI_d$ is the Einstein model of a solid, in which every atom vibrates independently in an identical harmonic well (Einstein, 1907). In this case, π_{tar} factorizes into independent Gaussians $\bigotimes_i \mathcal{N}(\mu_i, (\beta k)^{-1})$. Harmonic systems are among the regimes in which learned force fields are generally expected to perform well. Observing sampling failures even in this setting suggests that similar failures may persist in more complicated potentials.

4 RESULTS

We separate three questions that are often grouped under the term "MD accuracy." First, does controlling population approximation error under the target distribution control the learned dynamics at a finite simulation time? Second, is the stationary distribution of the learned dynamics close to the target? Third, which architectural constraints and inductive biases are necessary to control either of these quantities and what failure modes they eliminate?

We answer the first two questions negatively under unconstrained energy-and-force matching and then study how additional structure changes these conclusions. Sections 4.1 and 4.2 establish kinetic trapping, including under exact interpolation and initialization from the training set. Section 4.3 establishes equilibrium mismatch for a broader class of targets and arbitrary initialization. Section H.3 shows that locality and symmetry can potentially reduce the scale of hidden errors but do not prevent their amplification with system size. Finally, Section 4.4 separates conditions for accurate stationary distributions from those required for rapid relaxation.

Table 1: Error-to-sampling structure of the main results. "Population error" is measured with respect to the target distribution. The table distinguishes predictive accuracy, finite-time dynamics, and stationary-distribution accuracy.

Result	Setting	Error / Control condition	Physical conclusion
Theorem 4.1	Harmonic target; displaced initialization	Exponentially small population error	Kinetic trapping for $e^{\Omega(d)}$ time
Theorem 4.2	Harmonic target; training-data initialization	Exponentially small population error + Exact interpolation	Kinetic trapping for $e^{\Omega(d)}$ time
Theorem 4.3	Broader class of confining targets; arbitrary initialization	Exponentially small population error	Incorrect stationary distribution at distance $1 - e^{-\Omega(d)}$
Theorem 4.6	General targets with global energy control	Uniform or one-sided energy bounds	Stationary accuracy; uniform control also ensures relaxation.
Theorem H.5	Local, invariant site-energy models	Per-atom error $e^{-\Omega(\nu)}$	Global equilibrium error $1 - e^{-\Omega(N)}$

4.1 KINETIC TRAPPING UNDER EXPONENTIALLY SMALL ENERGY AND FORCE ERROR

Our first result provides a simplified setting for Theorem 4.2 and isolates finite-time dynamical failure. It establishes a lower bound on the total-variation distance between the distribution generated from a displaced initialization and the equilibrium distribution of an Einstein crystal.

Theorem 4.1. *Fix a constant $\xi > 5$. Let $V(x) = \frac{1}{2}\|x-\mu\|^2$ with $\|\mu\| = \xi\sqrt{d}$, so that $\pi_{\mathrm{tar}} = \mathcal{N}(\mu, I_d)$. There exists $\hat{V} \in C^\infty(\mathbb{R}^d)$, with $\hat{F} := -\nabla\hat{V}$, and $\sup_x \left\|\nabla^2\hat{V}(x)\right\|_{\mathrm{op}} \le H$ for a d-independent constant H, such that:*

(i) *For every $p \ge 1$, $\exists\, d_0(p,\xi)$ such that for $d \ge d_0(p,\xi)$,*

$$\mathbb{E}_{x\sim\pi_{\mathrm{tar}}}\left[\left|\hat{V}(x)-V(x)\right|^p\right]^{1/p} \le e^{-\Omega(d)}, \qquad \mathbb{E}_{x\sim\pi_{\mathrm{tar}}}\left[\left\|\hat{F}(x)-F(x)\right\|^p\right]^{1/p} \le e^{-\Omega(d)}$$

(ii) *There are constants $c, d_1 > 0$, depending only on ξ, such that for all $d \ge d_1$ and all $T \le e^{cd}$, the distribution $\hat{\pi}_T$ induced by Langevin dynamics with $\hat{F}$, initialized at $x_0, v_0 \overset{iid}{\sim} \mathcal{N}(0, I_d)$, satisfies*

$$\mathrm{TV}\big(\hat{\pi}_T, \pi_{\mathrm{tar}}\big) \ge 1 - e^{-\Omega(d)}$$

Moreover $\hat{F}$ is Lipschitz with a constant independent of d, so (2) has a unique strong solution.

Theorem 4.1 shows that, in high dimensions, the distribution induced by MD can remain far from π_{tar} over every polynomial time horizon despite exponentially small energy and force errors. The construction (Appendix D) introduces a smooth basin far from μ, where π_{tar} assigns mass $e^{-\Omega(d)}$, and interpolates back to the target potential wherever π_{tar} places appreciable mass. The perturbation is therefore invisible to the population matching objective while substantially altering the resulting dynamics. Importantly, $\hat{V}$ is smooth and conservative, its force is globally Lipschitz, and its Hessian bound is independent of dimension. These properties ensure regularity and well-posed dynamics, but the theorem shows that they do not control the global wells and barriers that determine sampling. In particular, such hidden basins can coexist with force errors that decay exponentially with dimension.

The freedom exploited by the construction reflects a practical feature of MLIP training. Reference data can be generated only for configurations on which the electronic structure calculation remains well defined and computationally tractable. Over much of configuration space such calculations are ill-defined: strongly overlapping atoms and dissociated fragments prevent SCF convergence or demand treatments beyond standard single-reference methods. The Boltzmann density is strictly positive throughout, so these regions may be absent from the data while still having nonzero probability under the target distribution.

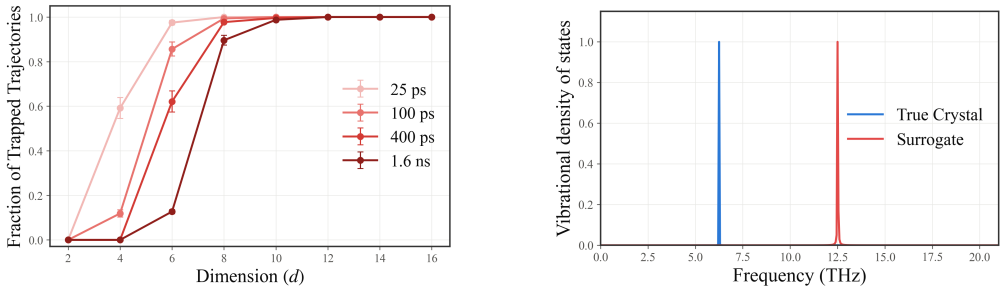
It is tempting to dismiss such regions as dynamically irrelevant. For a diatomic, this intuition is correct: the trajectory is confined by a dissociation barrier far exceeding k_BT, and no reasonable initialization escapes it. Two effects break this intuition in high dimensions. First, that barrier is a property of V, whereas the simulation is integrated on $\hat{V}$, which (1) constrains only where π_{tar} places mass. Second, independently of reachability, low π_{tar} density does not imply low $\hat{\pi}$-mass. For a target concentrating on a shell of radius $\Theta(\sqrt{d})$, the region beyond radius R may have π_{tar}-mass $e^{-\Omega(d)}$ while its volume scales as $R^{\Theta(d)}$. Suppressing $\hat{\pi}$-mass in such regions therefore requires an energy penalty that grows with d, which an L^2 bound does not guarantee. This energy-entropy competition is what Theorem 4.6 rules out, through a one-sided lower bound $\hat{V} \ge V - \delta$ that holds globally rather than only on the equilibrium region.

Proof Sketch: The proof is based on a hypoelliptic exit-time bound. The technical core is Lemma C.1 in Appendix C. Consider the linear kinetic Langevin dynamics

$$\mathrm{d}x_t = v_t\,\mathrm{d}t, \qquad \mathrm{d}v_t = (-\alpha x_t - \gamma v_t)\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t,$$

initialized with $\|x_0\| \le \rho_x\sqrt{d}$ and $\|v_0\| \le \rho_v\sqrt{d}$. The lemma shows that the exit time from the ball $\{\|x\| \le r\sqrt{d}\}$ exceeds $e^{c_\star d}$ with probability at least $1 - Ce^{-c_\star d}$, provided

$$\alpha \ge \alpha_\star = \Theta\left(\frac{1+\gamma^2\rho_x^2+\rho_v^2}{(r-\rho_x)^2}\right).$$



(a) Kinetic trapping in the Einstein crystal

(b) Phonon vibrational density of states

Figure 1: Illustration of kinetic trapping mechanism underlying Theorems 4.1 and 4.2 (details in Section A.1). (a) Escape time grows as $e^{\Omega(d)}$ with an $\mathcal{O}(1)$ rate per degree of freedom. The fully trapped plateau extends to arbitrarily large d. (b) The true Einstein crystal shows a single sharp line at the analytic $\nu_E = 6.25\,\mathrm{THz}$; the trapped surrogate's line is at $12.5\,\mathrm{THz}$.

A key feature is that the effect of the initial velocity enters only through $\rho_v/\sqrt{\alpha}$, so increasing the stiffness suppresses ballistic excursions. Moreover, although $\alpha_\star$ grows as the margin $r - \rho_x$ shrinks, the exponent $c_\star = \alpha_\star(r-\rho_x)^2/64$ remains unchanged. This scale invariance allows the same lemma to be applied at the much smaller radii required for Theorem 4.2.

Since $\|x_0\|, \|v_0\| \le 1.1\sqrt{d}$ with probability $1 - e^{-\Omega(d)}$, taking $r = 4$ and $\rho_x = \rho_v = 1.1$ confines the trajectory to $\|x_t\| \le 4\sqrt{d}$ for every $T \le e^{c_\star d}$. In contrast, π_{tar} concentrates outside this ball because $\|\mu\| = \xi\sqrt{d}$ with $\xi > 5$. Consequently, the distribution induced by the dynamics remains at total variation distance $1 - e^{-\Omega(d)}$ from π_{tar} throughout this time horizon. Meanwhile, the energy and force errors remain exponentially small because $\hat{V}$ differs from V only inside $\|x\| \le b_\xi\sqrt{d}$, where $b_\xi < \xi - 1$, a region whose π_{tar} mass is $e^{-\Omega(d)}$. The complete proof is in Appendix D.

4.2 Kinetic Trapping with Exact Interpolation and Initialization from Training Data

A standard practice is to initialize MD from a configuration within the training set. We show that even when the same samples are used both to fit an exactly interpolating $\hat{F}$ and to initialize MD, the failure mode of Theorem 4.1 persists.

Theorem 4.2. *Let $\mu \in \mathbb{R}^d$, $V(x) = \frac{1}{2}\|x-\mu\|^2$, so that $\pi_{\mathrm{tar}} = \mathcal{N}(\mu, I_d)$, and let $x_1, \dots, x_n$ with $n = \mathrm{poly}(d)$ be i.i.d. samples from π_{tar}. With probability $1 - e^{-\Omega(d)}$ over the draw of the training set, there exists $\hat{V} \in C^\infty(\mathbb{R}^d)$, depending on $x_1, \dots, x_n$, with $\hat{F} := -\nabla\hat{V}$ and $\sup_x \left\|\nabla^2\hat{V}(x)\right\|_{\mathrm{op}} \le H$ for a d-independent constant H, such that:*

(i) *$\hat{V}(x_i) = V(x_i)$ and $\hat{F}(x_i) = F(x_i)$ for every $i = 1, \dots, n$; that is, $\hat{V}$ attains zero empirical energy-and-force matching loss on the training set.*

(ii) *For every $p \ge 1$, $\exists\, d_0(p)$ such that for $d \ge d_0(p)$,*

$$\mathbb{E}_{x\sim\pi_{\mathrm{tar}}}\left[\left|\hat{V}(x) - V(x)\right|^p\right]^{1/p} \le e^{-\Omega(d)}, \qquad \mathbb{E}_{x\sim\pi_{\mathrm{tar}}}\left[\left\|\hat{F}(x) - F(x)\right\|^p\right]^{1/p} \le e^{-\Omega(d)}$$

(iii) *There are constants $c, d_1 > 0$ such that for all $d \ge d_1$ and all $T \le e^{cd}$, the distribution $\hat{\pi}_T$ induced by Langevin dynamics with $\hat{F}$, initialized at $x_0 \sim \frac{1}{n}\sum_i \delta_{x_i}$ and $v_0 \sim \mathcal{N}(0, I_d)$, satisfies*

$$\mathrm{TV}\left(\hat{\pi}_T, \pi_{\mathrm{tar}}\right) \ge 1 - e^{-\Omega(d)}$$

As before, the uniform Hessian bound makes $\hat{F}$ globally Lipschitz, so (2) has a unique strong solution. The full proof is given in Appendix E.

Theorem 4.2 strengthens Theorem 4.1 in two ways: the initialization is a training configuration drawn from π_{tar}, and the learned potential fits the training data exactly in both energy and force. The failure

therefore cannot be detected by the standard energy and force metrics, with an empirical loss driven to zero and an exponentially small test error coexisting with dynamics that remains far from the target over any $\mathrm{poly}(d)$ horizon. The proof exploits the fact that in high dimensions each training configuration occupies a neighborhood of exponentially small π_{tar}-measure, so the surrogate can place a basin around every sample while still matching energies and forces exactly there. The $\mathrm{poly}(d)$ basins together carry probability $e^{-\Omega(d)}$, leaving the population errors exponentially small while trajectories launched from the data remain confined.

4.3 Equilibrium mismatch for arbitrary initialization

Theorems 4.1 and 4.2 establish finite-horizon failures. One might therefore suspect that they arise from initialization or a limited simulation budget, and that alternative initializations or longer trajectories would recover π_{tar}. This is indeed true for the preceding constructions. By Proposition G.1, both learned potentials induce accurate stationary distributions, so the failure is purely kinetic. We next establish a stronger failure mode in which the learned stationary distribution is itself incorrect. This failure applies to a broader class of target distributions, holds for arbitrary initialization, and persists as $t \to \infty$. Throughout this section, we assume that the target satisfies Assumptions F.1 and F.2 and take $\inf V = 0$ without loss of generality.

Assumption F.1 is the standard Lyapunov condition under which (2) is hypoelliptic and geometrically ergodic with the phase-space Gibbs measure as its unique invariant measure (Hörmander, 1967; Mattingly et al., 2002). Assumption F.2 says the target concentrates on a ball of the physical radius $O(\sqrt{d})$, true for every strongly log-concave target and more generally under a dimension-free log-Sobolev inequality, and that the energy is extensive on that scale, true whenever $\nabla^2 V$ is bounded.

Theorem 4.3. *Let $\pi_{\mathrm{tar}} \propto e^{-\beta V}$ satisfy Assumptions F.1 and F.2 and fix $\kappa_2 > 0$. There exists $\hat{V} \in C^\infty(\mathbb{R}^d)$, with $\hat{F} := -\nabla \hat{V}$ and $\sup_x \left\| \nabla^2 \hat{V}(x) \right\|_{\mathrm{op}} \le L + H'$ for a d-independent constant H', such that:*

(i) *For every $p \ge 1$, $\exists\, d_0(p)$ such that for $d \ge d_0(p)$,*

$$\mathbb{E}_{x\sim\pi_{\mathrm{tar}}}\big[\,\big|\hat{V}(x) - V(x)\big|^p\,\big]^{1/p} \le e^{-\Omega(d)}, \qquad \mathbb{E}_{x\sim\pi_{\mathrm{tar}}}\big[\,\big\|\hat{F}(x) - F(x)\big\|^p\,\big]^{1/p} \le e^{-\Omega(d)}$$

(ii) *Langevin dynamics (2) driven by $\hat{F}$ is geometrically ergodic, its unique stationary distribution is the phase-space Gibbs measure of $\hat{V}$, whose position marginal is $\hat{\pi} \propto e^{-\beta \hat{V}}$, and for all d larger than a constant,*

$$\mathrm{TV}\big(\hat{\pi}, \pi_{\mathrm{tar}}\big) \;\ge\; 1 - e^{-\Omega(d)}$$

(iii) *Consequently, for* every *initial law with $\mathbb{E}\big[V(x_0) + \|x_0\|^2 + \|v_0\|^2\big] < \infty$,*

$$\lim_{t\to\infty} \mathrm{TV}\big(\hat{\pi}_t, \pi_{\mathrm{tar}}\big) \;=\; \mathrm{TV}\big(\hat{\pi}, \pi_{\mathrm{tar}}\big) \;\ge\; 1 - e^{-\Omega(d)}$$

Moreover $\hat{V} = V$ outside a ball of radius $2\theta\sqrt{d}$ carrying π_{tar}-mass $e^{-\Omega(d)}$, and $\left\|\hat{V} - V\right\|_\infty = O(d/\beta)$.

The construction (Appendix F) subtracts from V a smooth well of depth Δ, supported on a ball of radius $2\theta\sqrt{d}$ placed just beyond the region where π_{tar} concentrates. Two competing requirements determine Δ. The well must be deep enough that $e^{\beta\Delta}$ compensates for the exponentially small prior mass that $\hat{\pi}$ would otherwise assign to the ball, requiring $\beta\Delta = \Omega(d)$. At the same time, it must remain shallow and wide enough that the induced curvature, $\Delta/(\theta\sqrt{d})^2$, stays independent of d, which is achieved by taking $\beta\Delta = O(d)$ and a width of $\Theta(\sqrt{d})$. These requirements are compatible, yielding a confining potential with a well-defined Boltzmann equilibrium that is not the target equilibrium.

Remark 4.4. The construction is extensive rather than divergent. The energy offset is $\Delta = O(d/\beta)$ in total, which is $O(k_B T)$ per atom, and the force perturbation is $O(\sqrt{d})$ in norm, which is $O(1)$ per atom. Consequently, a model that is accurate to $e^{-\Omega(d)}$ on the sampled basin but underestimates the energy by only a few $k_B T$ per atom in an unsampled region can nevertheless place all equilibrium mass essentially there.

Remark 4.5. The construction is conservative, and by the confinement properties established in Theorem 4.3, the learned dynamics has a well-defined Gibbs equilibrium. Non-conservative direct-force models can exhibit additional failures, including rotational drift and non-gradient escape mechanisms. For example, the overdamped construction of Cao et al. (2026) introduces a constant drift within a narrow cone that transports the process to infinity. Because this field is not a gradient, that construction does not apply to conservative models. These mechanisms may help explain empirical failures of non-conservative models recently observed by Bigi et al. (2025) and Fu et al. (2025). However, Theorems 4.1 and 4.3 show that conservative models with well-defined equilibria need not have the correct equilibrium distribution or mix efficiently.

All of our main results are stated for smooth, conservative potentials without architectural constraints such as locality or symmetry, thereby emphasizing the need for additional structure. Standard MLIPs impose additional structure through finite-range locality, permutation invariance, and $E(3)$-invariant energies with equivariant forces. It remains unclear whether these inductive biases are sufficient to rule out the failure mechanisms identified here, whether further constraints are required, and how sampling guarantees change under these architectural constraints.

Appendix H shows that the equilibrium-mismatch mechanism of Theorem 4.3 can persist under these inductive biases when a competing structure is locally recognizable, meaning that its local atomic environments can be distinguished from those typical of the target distribution within the model's finite receptive field. Under explicit concentration and bounded-overlap assumptions, we construct a smooth, finite-cutoff, size-extensive site-energy model with a $2L$-hop receptive field whose per-atom energy and force errors are $e^{-\Omega(\nu)}$, while its stationary distribution satisfies $\mathrm{TV}(\hat{\pi}, \pi_{\mathrm{tar}}) \geq 1 - e^{-\Omega(d)}$. Here, $\nu \asymp z^{L-1}$ in the bounded-degree regime, where z is the local coordination number. Compared with the unconstrained construction of Theorem 4.3, locality could potentially raise the error that can be hidden from $e^{-\Omega(d)}$ to $e^{-\Omega(\nu)}$. However, it does not prevent an undetected local bias from accumulating extensively when a competing structure, such as a distorted lattice or competing polymorph, is locally recognizable and differs from the target by only $O(1)$ free energy per atom.

Because $\hat{V}$ is expressed as a sum of atomic energies, the trained model can be evaluated at system sizes different from those used during training, as is common practice in MLIP applications. As shown in Appendix H, the per-atom error bounds remain independent of system size, whereas the equilibrium discrepancy approaches one exponentially fast in N. Consequently, the model can exhibit unchanged per-atom validation error as system size increases while its equilibrium distribution becomes increasingly inaccurate. This separation highlights the need for caution when deploying models trained on smaller systems for inference and sampling in larger systems (Remark H.8).

4.4 Sufficient Conditions

The negative results reveal two distinct requirements for accurate sampling. Thermodynamic accuracy requires the learned invariant distribution to be close to the target, while dynamical accuracy additionally requires that this distribution be reached on the simulation timescale. The following theorem gives sufficient conditions for each.

Theorem 4.6. *Let $\hat{\pi} \propto e^{-\beta \hat{V}}$ and $\pi_{\mathrm{tar}} \propto e^{-\beta V}$.*

(i) *If $\left\| \hat{V} - V \right\|_\infty \leq \delta$, then $\frac{\mathrm{d}\hat{\pi}}{\mathrm{d}\pi_{\mathrm{tar}}} \in [e^{-2\beta\delta}, e^{2\beta\delta}]$ and $\mathrm{TV}(\hat{\pi}, \pi_{\mathrm{tar}}) \leq 1 - e^{-2\beta\delta} \leq 2\beta\delta$.*

(ii) *Suppose (a) $\hat{V} \geq V - \delta$ everywhere, and (b) there is a set S with $\pi_{\mathrm{tar}}(S) \geq 1 - \eta$ and $\hat{V} \leq V + \delta'$ on S. Then $\frac{\mathrm{d}\hat{\pi}}{\mathrm{d}\pi_{\mathrm{tar}}} \leq \frac{e^{\beta(\delta+\delta')}}{1-\eta}$ and $\mathrm{TV}(\hat{\pi}, \pi_{\mathrm{tar}}) \leq 1 - e^{-\beta(\delta+\delta')} + \eta$.*

(iii) *Suppose in addition that π_{tar} satisfies a log-Sobolev inequality with constant $\rho > 0$ and $\sup_x \left\| \nabla^2 \hat{V}(x) \right\|_{\mathrm{op}} \leq H$. Under (i), $\hat{\pi}$ satisfies a log-Sobolev inequality with constant at least $\rho\, e^{-2\beta\delta}$. Consequently, Langevin dynamics driven by $\hat{F} = -\nabla \hat{V}$ converges exponentially to its invariant distribution $\hat{\pi}$, with a hypocoercive rate depending on $\rho e^{-2\beta\delta}$, H, γ, and β. Under condition (ii) alone, no analogous uniform relaxation guarantee follows.*

Part (i) imposes a uniform error bound and thereby rules out both failure modes. Such a guarantee is stronger than what is typically obtained from unconstrained matching: an L^p bound under π_{tar} need

Table 2: Summary of guarantees provided by common assumptions and the aspects of sampling they leave uncontrolled.

Assumption	Provides	Does not control
Small L^p energy/force error	Accuracy on typical configurations	Existence of equilibrium
Conservative force + Confinement	Scalar potential and Gibbs equilibrium	Equilibrium accuracy and mixing time
Smoothness + bounded Hessian	Lipschitz force and well-posed dynamics	Global wells, barriers, sampling
Locality, permutation and $E(3)$ symmetry	Exact transformation laws, extensivity and a potential receptive-field safeguard	Accumulation of residual local bias with system size
One-sided global control + on-target accuracy	Equilibrium accuracy	High barriers and kinetic trapping
Uniform two-sided control + functional inequality	Equilibrium accuracy and exponential relaxation	A strong global condition not supplied by ordinary matching

not provide meaningful control in regions where π_{tar} places very little mass, including regions that may appear dynamically irrelevant under the target distribution. Part (ii) weakens this requirement in the direction relevant for practice, requiring accuracy only on the support of the target and replacing global accuracy with a global lower bound $\hat{V} \geq V - \delta$. This condition rules out the mechanism in Theorem 4.3, which constructs a well of depth $\Delta = \Omega(d/\beta)$ below the true potential: preventing wells deeper than δ limits the equilibrium mass that can accumulate off-support. However, it provides no control over barriers and therefore no guarantee on relaxation.

Part (ii) can provide a design target. Enforcing a lower bound on the predicted energy, for example through a repulsive baseline or a coercive residual parameterization $\widehat{V} = V_0 + f_\theta$, rules out wrong-equilibrium failures without sacrificing on-support accuracy. It does not require extrapolation, but only that $\widehat{V}$ does not fall arbitrarily below V in regions where the target places little mass.

In practice, constructing such coercive baselines with limited knowledge of the true potential energy surface from classical physics remains an important direction for future work. Physical priors, such as short-range ZBL repulsion, can provide local protection against unphysical energy collapse when atoms approach very small separations and have been incorporated into recent MLIPs (Tompa et al., 2026; Zhang et al., 2026) for this purpose. However, such priors address only a specific class of off-support failures and do not provide the global one-sided energy control required by our result. In particular, the failure mechanisms constructed here need not involve atomic collisions or anomalously short interatomic distances, and can therefore remain invisible to purely short-range repulsive corrections. Developing broadly applicable priors that provide the required energy control without imposing detailed knowledge of the true potential remains an open problem.

5 Conclusion

We have shown that the relationship between energy-and-force matching error and sampling accuracy is exponentially ill-conditioned in dimension. Smooth, conservative surrogates with dimension-independent Hessian bounds can interpolate the training data exactly, achieve exponentially small population energy and force errors, and still induce dynamics whose distribution remains far from the target, either because the target equilibrium cannot be reached in sub-exponential time (kinetic trapping) or because the learned equilibrium distribution itself is incorrect (equilibrium mismatch). Equilibrium mismatch can persist under locality, permutation invariance, and $E(3)$-equivariance, although the local construction may incur a receptive-field-dependent approximation error. A one-sided global lower bound on the learned energy rules out the equilibrium-mismatch mechanism, while rapid relaxation additionally requires control of barriers and global geometry. Together, our impossibility results for unconstrained MLIPs provide theoretical support for imposing structure beyond energy-and-force matching and identify design targets for such constraints. Our results are existential and worst case and therefore do not establish how frequently these failure modes arise in practice. Characterizing their prevalence under realistic training procedures and data distributions is an important direction for future work.

AI USE STATEMENT

In this work, generative AI tools have been used to improve clarity and readability, provide suggestions for the organization of tables and appendices, give feedback on the correctness and presentation of proofs, and assist with coding for numerical experiments. We also used generative AI tools, including those provided through Google's Paper Assistant Tool (PAT) and tools used independently by the authors, to review the final version of the manuscript and provide feedback on its presentation, clarity and technical claims. In response to questions raised by the LLM review, we added a small number of remarks in the appendix to clarify proofs, delineate the scope of technical claims, and address those questions. We did not use generative AI tools for any other tasks requiring disclosure, and the remaining disclosure categories are not applicable to this work. We reviewed all AI-assisted content and take full responsibility for the final content of this work, including all text, claims, proofs, code, and other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

Full proofs of all theoretical claims are provided in the appendix, with proof sketches included in the main text. The appendices state the assumptions used in each result and give the explicit constructions and intermediate lemmas needed to verify the claimed approximation and sampling bounds. For the numerical experiments, Appendix A provides the details of the physical systems, construction of the target distributions, training setup for the learned potentials, molecular-dynamics integration procedure, and evaluation metrics used in the figures.

REFERENCES

Robert J. Adler and Jonathan E. Taylor. *Random Fields and Geometry*. Springer, 2007.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84*, pp. 177–206. Springer, 1985.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014.

Ilyes Batatia, Dávid P. Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, 2022.

Jörg Behler. Four generations of high-dimensional neural network potentials. *Chemical Reviews*, 121 (16):10037–10072, 2021.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *International Conference on Learning Representations (ICLR)*, 2024.

Filippo Bigi, Marcel F. Langer, and Michele Ceriotti. The dark side of the forces: assessing non-conservative force models for atomistic machine learning. In *Forty-second International Conference on Machine Learning*, 2025.

Sergey G. Bobkov. Isoperimetric and analytic inequalities for log-concave probability measures. *The Annals of Probability*, 27(4):1903–1921, 1999.

Anton Bovier, Michael Eckhoff, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes I: Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 6(4):399–424, 2004.

Anton Bovier, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes II: Precise asymptotics for small eigenvalues. *Journal of the European Mathematical Society*, 7(1): 69–99, 2005.

Daniel Yiming Cao, August Y. Chen, Karthik Sridharan, and Yuchen Wu. On the robustness of Langevin dynamics to score function error. In *Forty-third International Conference on Machine Learning*, 2026.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023.

Isabel Creed, Tim Rein, Ingvars Vitenburgs, Wojciech G. Stark, Viktor Ellingsson, Ahmed Y. Ismail, Guangyu Liu, Yuchen Lou, Bradley A. A. Martin, Cyprien Bone, Matthew A. H. Walker, Mueen Taj, Shirui Wang, Kelvin Wong, Ruiqi Wu, Prakriti Kayastha, Bingqing Cheng, Aditi Krishnapriyan, Michele Ceriotti, Marcel F. Langer, Jarvist Moore Frost, Alex M. Ganose, Venkat Kapil, and Keith T. Butler. Six open questions in machine-learned interatomic potential foundation models. *arXiv preprint arXiv:2606.07327*, 2026.

Bowen Deng, Yunyeong Choi, Peichen Zhong, Janosh Riebesell, Shashwat Anand, Zhuohan Li, KyuJung Jun, Kristin A. Persson, and Gerbrand Ceder. Systematic softening in universal machine learning interatomic potentials. *npj Computational Materials*, 11:9, 2025.

Albert Einstein. Die Plancksche Theorie der Strahlung und die Theorie der spezifischen Wärme. *Annalen der Physik*, 22:180–190, 1907.

Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gomez-Bombarelli, and Tommi Jaakkola. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *Transactions on Machine Learning Research*, 2023.

Xiang Fu, Brandon M Wood, Luis Barroso-Luque, Daniel S. Levine, Meng Gao, Misko Dzamba, and C. Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction. In *Forty-second International Conference on Machine Learning*, 2025.

Richard Holley and Daniel Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46(5):1159–1194, 1987.

Lars Hörmander. Hypoelliptic second order differential equations. *Acta Mathematica*, 119:147–171, 1967.

Benedict Leimkuhler and Charles Matthews. *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*. Springer, 2015.

Tony Lelièvre and Gabriel Stoltz. Partial differential equations and stochastic methods in molecular dynamics. *Acta Numerica*, 25:681–880, 2016.

Gen Li, Yu Huang, Timofey Efimov, Yuting Wei, Yuejie Chi, and Yuxin Chen. Accelerating convergence of score-based diffusion models, provably. In *International Conference on Machine Learning (ICML)*, 2024.

Jiadong Liang, Zhihan Huang, and Yuxin Chen. Low-dimensional adaptation of diffusion models: convergence in total variation. In *Conference on Learning Theory (COLT)*, 2025.

Antoine Loew, Dewen Sun, Hai-Chen Wang, Silvana Botti, and Miguel A. L. Marques. Universal machine learning interatomic potentials are ready for phonons. *npj Computational Materials*, 11: 178, 2025.

Sajid Mannan, Vaibhav Bihani, Carmelo Gonzales, Kin Long Kelvin Lee, Nitya Nand Gosvami, Sayan Ranu, Santiago Miret, and N. M. Anoop Krishnan. UniFFBench: Evaluating universal machine learning force fields against experimental measurements. *Nature Computational Science*, 6:894–906, 2026. doi: 10.1038/s43588-026-01019-4.

J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101(2): 185–232, 2002.

Georg Menz and André Schlichting. Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. *Annals of Probability*, 42(5):1809–1884, 2014.

Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J. Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14:579, 2023.

Bernt Øksendal. *Stochastic Differential Equations*. Springer, 2003.

Kristof T. Schütt, Oliver T. Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning (ICML)*, 2021.

Tamás Lajos Tompa, Eszter Varga-Umbrich, Ilyes Batatia, Alin M. Elena, Noam Bernstein, and Gábor Csányi. Fine-tuning MLIP foundation models: strategies for accuracy and transferability. *arXiv preprint arXiv:2606.12704*, 2026.

Oliver T. Unke, Stefan Chmiela, Huziel E. Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T. Schütt, Alexandre Tkatchenko, and Klaus-Robert Müller. Machine learning force fields. *Chemical Reviews*, 121(16):10142–10186, 2021. doi: 10.1021/acs.chemrev.0c01111.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Cédric Villani. Hypocoercivity. *Memoirs of the American Mathematical Society*, 202, 2009.

Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical Review Letters*, 120(14):143001, 2018.

Wentao Zhang, Xingxing Wu, Chen Wang, Siyu Hu, Yueyang Liu, and Lin-Wang Wang. Constructing machine learning interatomic potentials with minimum amount of ab initio data. *npj Computational Materials*, 12(1):174, 2026.

A Experiments

We include a small set of controlled simulations to illustrate the sampling failures predicted by our theory. These simulations are not intended to reproduce realistic MLIP training pipelines or support claims about practical model performance. Their purpose is to provide numerical demonstrations of the limitations of unconstrained energy-and-force matching. Although our theory concerns worst-case constructions, simulation collapse has been empirically observed in state-of-the-art MLIPs and reported in the literature (Fu et al., 2023; Mannan et al., 2026), suggesting that related failure mechanisms may arise in practice (see Figure 3). In realistic MLIP training pipelines, however, isolating and verifying the precise mechanism responsible for a given simulation failure is generally impractical, since multiple sources of model and sampling error may interact.

We first construct an Einstein-crystal model that directly exhibits the kinetic-trapping mechanism of Theorems 4.1 and 4.2. We then study a harmonic argon crystal and show that neural networks trained by energy-and-force matching can produce inaccurate molecular dynamics despite small held-out force errors. These experiments are intended as finite-time illustrations of the phenomena captured by the theory rather than numerical realizations of the theorem constructions.

A.1 Einstein crystal: explicit construction

We begin with an Einstein crystal for which both the equilibrium distribution and the relevant thermodynamic observables are available analytically. We use an fcc Al lattice with lattice constant $a_0 = 4.05$ Å and a $2 \times 2 \times 2$ supercell containing $N = 32$ atoms ($d = 3N = 96$). Each atom is tethered to its lattice site s_i by an isotropic harmonic potential with force-constant matrix $K = kI$, where $k = m\omega_E^2$. We set the Einstein temperature to $\theta_E = h\nu_E/k_B = 300$ K and the atomic mass to $m = 26.98$ u, which gives $k \simeq 4.3$ eV/Å^2 and $\nu_E = \omega_E/(2\pi) = 6.25$ THz. The corresponding classical predictions are $\langle u^2 \rangle = 3k_BT/k = 0.0180$ Å^2, $C_v = 3k_B$ per atom, and a single vibrational-density-of-states (VDOS) peak at ν_E.

Table 3: Analytic predictions and molecular-dynamics estimates for the Einstein-crystal construction. The surrogate trajectory is kinetically trapped (Figure 1). Its heat capacity remains apparently correct even though its displacement statistics and mean structure are incorrect.

Physical property	Analytic	True MD	Surrogate MD
Debye–Waller $\langle u^2 \rangle$ (Å^2)	0.0180	0.0180	**0.0045**
Structure error (Å)	0	0.004	**1.06**
Heat capacity C_v (k_B/atom)	3.00	3.03	**3.03**

To illustrate kinetic trapping, we modify the surrogate potential by introducing an additional basin with local stiffness $\alpha = 4k$. The basin is centered at μ, with $\|\mu - s\| = 6$ Å, and is gated by a C^1 smoothstep that is equal to one within $r_a = 2.0$ Å and vanishes beyond $r_c = 4.0$ Å. A trajectory initialized inside this basin remains trapped there, reproducing the mechanism underlying Theorems 4.1 and 4.2.

We generate Langevin trajectories with ASE at $T = 300$ K. Initial velocities are drawn from the Maxwell–Boltzmann distribution with center-of-mass motion removed. We use a 3 fs timestep, 10 ps of equilibration, and 40 ps of NVT sampling with friction 0.02 fs^{-1}. The VDOS is estimated from the velocity power spectrum of a subsequent 20 ps NVE trajectory integrated with velocity-Verlet using a 2 fs timestep.

We report the mean-square displacement $\langle u^2 \rangle = \langle \|r_i - \bar{r}_i\|^2 \rangle_{i,t}$, the structural error $\left[N^{-1} \sum_i \|\bar{r}_i - s_i\|^2\right]^{1/2}$, the heat capacity $C_v = \frac{\mathrm{Var}(E_{\mathrm{tot}})}{k_B T^2 N}$, and the VDOS,

$$\mathrm{VDOS}(\nu) \propto \sum_i |\hat{\mathbf{v}}_i(\nu)|^2.$$

True MD reproduces the analytic predictions, whereas the trapped surrogate gives an incorrect mean-square displacement, mean structure, and phonon frequency. The heat capacity remains close to $3k_B$ per atom because the classical harmonic heat capacity does not depend on the spring stiffness. The numerical comparison is given in Table 3.

We also examine how the trapping behavior changes with dimension. We consider $d \in \{2, \ldots, 16\}$ independent Einstein oscillators with spring constant $k = 4.3$ eV/Å^2, mass $m = 26.98$ u, and temperature $T = 300$ K. A near-marginal well with stiffness $\alpha = 0.066k$ is introduced along one collective direction, with its offset and gate radii scaled proportionally to $L_0\sqrt{d}$, where $L_0 = \sqrt{k_B T/k}$ is the thermal length per degree of freedom.

For each of four seeds, $M = 120$ walkers are initialized inside the basin and propagated for up to 1800 ps using BAOAB Langevin dynamics ($\Delta t = 3$ fs, $\gamma = 0.02$ fs^{-1}). A walker is considered to have escaped once $\|u - u_t\|$ exceeds the outer gate radius, where u_t is the basin center. Figure 1a reports the fraction of walkers whose escape time exceeds $t \in \{25, 100, 400, 1600\}$ ps. At each fixed observation time, the trapped fraction increases toward one with dimension, while longer observation times shift the onset of trapping to larger d. This behavior is consistent with the increasing escape times predicted by the theoretical construction.

A.2 FCC Argon

We next consider a harmonic face-centred-cubic argon crystal. This setting is useful because the target equilibrium distribution is known exactly, allowing us to compare trajectories generated by a learned potential against a closed-form reference.

We model argon using a Lennard–Jones potential with $\sigma = 3.40$ Å and $\epsilon = 0.0104$ eV, smoothly truncated at 8.5 Å. Minimizing the Lennard–Jones energy under periodic boundary conditions gives a relaxed lattice constant $a_0 = 5.26$ Å and a nearest-neighbor distance of 3.72 Å. All simulations are performed at $T = 30$ K ($0.36T_{\mathrm{melt}}$), where the Lindemann ratio is approximately 5%.

The force-constant matrix K is obtained by central finite differences at the relaxed geometry, symmetrized, and adjusted to satisfy the acoustic sum rule. After removing the three translational

zero modes, the system has $3N - 3$ vibrational modes with an upper band edge of approximately $\nu_{\max} = 2.0$ THz. For the harmonic potential

$$V(u) = \tfrac{1}{2} u^\top K u,$$

the equilibrium distribution on the vibrational subspace is $\pi_{\mathrm{tar}} = \mathcal{N}(0, k_B T K^+)$. We draw both the training and test configurations independently from this distribution.

Surrogate model and training: The learned potential $\hat{V}_\theta(u)$ is represented by an unconstrained scalar-output $\tanh$ MLP with eight width-512 hidden blocks and skip connections. We deliberately impose no symmetry, locality, or coercivity. For any fixed network with finite weights, the smooth $\tanh$ activations give spatially bounded derivatives, although we do not impose a Hessian bound that is uniform across system sizes. We use Xavier initialization with the $\tanh$ gain, skip connections, and gradient clipping at norm 1. Forces are obtained by automatic differentiation, $\hat{F}_\theta = -\nabla \hat{V}_\theta$.

We minimize the empirical energy-and-force matching objective in (1), normalizing the force term by the force scale and including a per-degree-of-freedom energy anchor with weight 0.02. The network is trained with Adam for 3×10^4 steps using a batch size of 4096 and a cosine-annealed learning rate initialized at 2×10^{-3}.

Molecular dynamics: We integrate (2) with BAOAB (Leimkuhler & Matthews, 2015), using $\Delta t = 10$ fs and $\gamma = 10^{-3}$ fs^{-1}. For each trained model, 256 walkers are initialized independently from π_{tar} and propagated for 25 ns, with observables computed over the second half of each trajectory.

We consider system sizes $N \in \{16, 32, \ldots, 128\}$ while keeping the training density fixed at $n/N = 625$. Thus, the number of training configurations increases from $n = 10^4$ at $N = 16$ to $n = 8 \times 10^4$ at $N = 128$. We separately study the effect of additional data at $N = 128$, increasing the training set to $n = 5 \times 10^5$. Network initialization, training samples, and Langevin noise are varied independently across four random seeds.

Sampling diagnostics: To quantify the discrepancy between the learned dynamics and the target, we project MD configurations onto the exact phonon modes and compare their empirical covariance with the target covariance $\mathrm{diag}(k_B T/\lambda_j)$. We denote the resulting Gaussian moment-based KL diagnostic by $\widehat{\mathrm{KL}}$. If the exact moments were known, the corresponding Gaussian divergence would lower-bound the KL divergence to the Gaussian target. Because we estimate the moments from finite, temporally correlated trajectories, $\widehat{\mathrm{KL}}$ need not retain this lower-bound property. We therefore use it only as a finite-trajectory diagnostic and do not assume that the learned dynamics converge to a stationary distribution.

We also compute the Debye–Waller factor $\langle u^2 \rangle$ using the same per-atom mean-square-displacement estimator as in the Einstein-crystal experiment. Although the trajectories themselves are generated using classical MD, we estimate the heat capacity from the phonon spectrum using the quantum harmonic expression

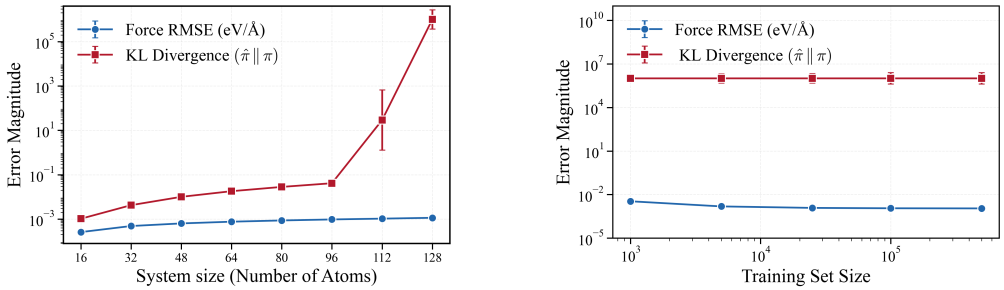
$$C_v = \frac{k_B}{N} \sum_j \frac{x_j^2 e^{x_j}}{(e^{x_j} - 1)^2}, \qquad x_j = \frac{h\nu_j}{k_B T},$$

where $\{\nu_j\}$ denotes either the true or learned phonon spectrum. We use this spectrum-based estimate because argon is partially quantum at 30 K, so its heat capacity is below the classical value of $3k_B$ per atom.

As a reference, a well-trained surrogate at $N = 32$ gives $\widehat{\mathrm{KL}} = 0.004$, while $\langle u^2 \rangle$, C_v, and $\nu_{\max}$ agree with the analytic values to within 0.5%.

Sampling error with increasing system size: Figure 2a shows a sharp separation between predictive and dynamical accuracy. The held-out force error remains of order 10^{-3} or smaller throughout the dimension sweep, while the sampling discrepancy increases by several orders of magnitude. At $N = 128$, all four learned trajectories eventually leave the harmonic crystal basin. This occurs even though the target distribution is Gaussian and strongly log-concave on the vibrational subspace.

The experiment therefore illustrates that small held-out force error need not translate into accurate finite-time sampling as the number of degrees of freedom increases. Although our asymptotic results



(a) Sampling error grows with dimension

(b) Additional data does not improve sampling

Figure 2: Molecular dynamics of the harmonic argon crystal. Error bars show the standard deviation across four random seeds. (a) As the system size increases, the held-out force RMSE remains of order 10^{-3} or smaller, while the finite-trajectory Gaussian KL diagnostic increases with dimension. At $N = 128$, trajectories from all four seeds leave the crystal basin. (b) At fixed system size $N = 128$, increasing the training-set size does not substantially reduce the sampling discrepancy under the tested training protocol.

apply only beyond a fixed dimension d_0, the same qualitative behavior appears here at system sizes routinely accessible in molecular simulation.

Increasing the amount of training data does not substantially change this behavior. As shown in Figure 2b, the sampling discrepancy at $N = 128$ remains large as the training set is increased, even though the surrogate continues to predict forces accurately on held-out equilibrium configurations. The escaped trajectories enter regions that carry negligible probability under π_{tar}, where the training objective provides little control over the learned energy landscape. This behavior is consistent with the off-distribution freedom used in our theoretical constructions. Related behavior was reported by Fu et al. (2023), who observed that, for certain architectures, increasing the training-set size from 1,000 to 90,000 configurations reduced force errors but did not eliminate simulation failures.

Coercive baselines: To examine whether simple global control of the learned potential can mitigate these failures, we consider two proof-of-concept coercive baselines, C1 and C2, motivated by Theorem 4.6. In both cases, the learned potential takes the form

$$\widehat{U}(u) = U_0(u) + f_\theta(u),$$

where U_0 is a fixed confining potential. The quadratic or quartic growth of U_0 ensures that $\widehat{U}$ is confining and has a finite partition function. These constructions do not, however, guarantee the stronger global comparison $\widehat{U} \geq V - \delta$ required by Theorem 4.6.

For C1, we use an isotropic harmonic prior

$$U_0(u) = \frac{1}{2}\kappa\|u\|^2,$$

where κ is the mean force constant, and train the network on the residual $V(u) - U_0(u)$.

For C2, we instead use a repulsive boundary potential

$$U_0(u) = \varepsilon \left[\mathrm{ReLU}\left(\frac{u^\top \Sigma^{-1} u}{k} - c\right)\right]^2,$$

where Σ is the empirical covariance of the training configurations. The prior vanishes inside the ellipsoid $u^\top \Sigma^{-1} u/k < c$ and therefore leaves the learned potential unchanged near the training distribution. Outside this region, it grows quartically with displacement.

The results in Table 4 suggest that even simple confining tails can substantially change the behavior of the learned dynamics. C1 reduces the large-N discrepancy but still shows appreciable displacement inflation. C2 keeps the mean-square displacement close to the target across the dimension sweep and produces substantially smaller KL diagnostics at the largest system sizes. These results are consistent with the role of global energy control in Theorem 4.6, although neither baseline certifies the assumptions of the theorem.

Table 4: Debye–Waller inflation, finite-trajectory Gaussian KL diagnostic, and phonon-based quantum harmonic heat capacity as functions of system size. C1 reduces the large-N discrepancy, while C2 keeps the displacement statistics close to the target and substantially reduces the KL diagnostic. C_v is computed from the phonon spectrum and averaged over seeds that remain within the harmonic basin. A dash indicates that all seeds left the basin, so a basin-based phonon spectrum is not reported.

	$\langle u^2\rangle/\langle u^2\rangle_{\mathrm{t}}$			$\widehat{\mathrm{KL}}$			C_v (k_B/atom)			
N	Unconst.	C1	C2	Unconst.	C1	C2	Analytic	Unconst.	C1	C2
16	1.00	1.00	1.00	0.001	0.001	0.001	1.901	1.901	1.901	1.901
32	1.00	1.00	1.00	0.004	0.004	0.004	1.980	1.980	1.980	1.980
48	1.00	1.00	1.00	0.010	0.010	0.010	2.004	2.005	2.004	2.005
64	1.00	1.00	1.00	0.018	0.018	0.018	2.019	2.020	2.019	2.020
80	1.00	1.00	1.00	0.028	0.027	0.028	2.028	2.030	2.028	2.030
96	1.01	1.00	1.01	0.041	0.040	0.041	2.034	2.037	2.034	2.037
112	8.90	3.30	1.01	29.1	7.20	0.057	2.039	2.041	2.033	2.041
128	641	2.25	1.02	1.5×10^6	5.70	0.590	2.042	–	2.085	2.048

Implementation: The experiments were implemented in PyTorch, NumPy, and ASE using the Lennard–Jones calculator. Each run used a single NVIDIA A100 GPU. The computational cost was dominated by the $N = 128$ simulations, for which propagating 256 walkers for 2.5×10^6 integration steps required approximately one hour per run.

A.3 Limitations

These experiments are restricted to controlled harmonic systems and unconstrained neural networks. They therefore do not establish how often analogous failures occur in practical MLIP training or deployment. The sampling discrepancies reported above are Gaussian moment-based diagnostics estimated from finite and temporally correlated trajectories. They need not measure the full distributional divergence. In particular, the unconstrained neural network need not define a normalizable Boltzmann distribution, so trajectory escape should not be interpreted as a direct numerical verification of Theorem 4.3.

The training-set-size experiment likewise describes the behavior of the specific architecture and optimization procedure studied here rather than a general impossibility result about additional data. Similarly, the coercive baselines demonstrate the effect of introducing confining tails but do not guarantee the global energy comparison required by Theorem 4.6. The experiments should therefore be viewed as controlled illustrations of the mechanisms suggested by the theory. The formal impossibility and sampling guarantees follow from the theoretical results themselves.

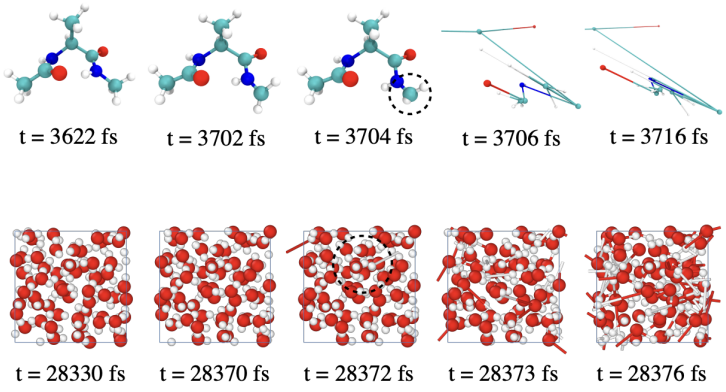


Figure 3: Figure adapted from Fu et al. (2023) (Figure 6) depicting simulation collapse when applying (a) NequIP to alanine dipeptide and (b) GemNet-T to water.

B TECHNICAL PRELIMINARIES

We first introduce several key technical preliminaries, beginning with two concentration-of-measure results in high dimensions.

B.1 CONCENTRATION OF MEASURE

Lemma B.1 (Gaussian norm concentration; Vershynin, 2018)**.** *There are universal constants $c, C > 0$ such that for $X \sim \mathcal{N}(0, I_d)$ and all $t > 0$,*

$$\mathbb{P}\Big(\Big| \|X\| - \sqrt{d} \Big| \geq t\sqrt{d} \Big) \;\leq\; Ce^{-ct^2 d}.$$

Lemma B.2 (Lipschitz concentration; Bakry et al., 2014)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz. If $X \sim \mathcal{N}(0, I_d)$ then $\mathbb{P}\big(|f(X) - \mathbb{E}f(X)| \geq t \big) \leq 2e^{-t^2/(2L^2)}$ for all $t > 0$. More generally, if the law of X is α-strongly log-concave the same bound holds with $e^{-\alpha t^2/(2L^2)}$ in place of $e^{-t^2/(2L^2)}$; this yields the strongly-log-concave extension of Remark D.3.*

We note two consequences that will be used later in our analysis.

First, if $x_0 \sim \mathcal{N}(0, I_d)$, then Lemma B.1 with $t = 0.1$ gives

$$\mathbb{P}\big(\|x_0\| \leq 1.1\sqrt{d}\big) \geq 1 - Ce^{-cd},$$

and likewise for $v_0 \sim \mathcal{N}(0, I_d)$.

Second, let $\pi_{\mathrm{tar}} = \mathcal{N}(\mu, \mathrm{Id})$ with $\|\mu\| = \xi\sqrt{d}$ for $\xi > 5$, and fix $t \in (0, \xi - 1 - b_\xi)$. By Lemma B.1, we have

$$\|x - \mu\| \leq (1 + t)\sqrt{d} \qquad \text{with probability } 1 - e^{-\Omega(d)},$$

and on this event,

$$\|x\| \geq \|\mu\| - \|x - \mu\| \geq (\xi - 1 - t)\sqrt{d} > b_\xi\sqrt{d}.$$

Hence, we obtain the bounds

$$\pi_{\mathrm{tar}}\big(\|x\| \geq b_\xi\sqrt{d}\big) \geq 1 - e^{-\Omega(d)} \qquad \text{and, in particular} \qquad \pi_{\mathrm{tar}}\big(\|x\| \geq 4\sqrt{d}\big) \geq 1 - e^{-\Omega(d)}.$$

B.2 BUMP FUNCTIONS

Lemma B.3. *There is a universal constant C_0 with the following property. For all $0 < a < b$ and all $d \geq 1$, there is a radial $\psi_{a,b} \in C^\infty(\mathbb{R}^d; [0, 1])$ with $\psi_{a,b} \equiv 1$ on $\{\|y\| \leq a\}$, $\psi_{a,b} \equiv 0$ on $\{\|y\| \geq b\}$, and*

$$\|\nabla\psi_{a,b}\|_\infty \leq \frac{C_0}{b - a}, \qquad \|\nabla^2\psi_{a,b}\|_{\mathrm{op},\infty} \leq \frac{C_0}{(b - a)^2} + \frac{C_0}{a(b - a)}. \tag{B.1}$$

In particular, these bounds are independent of d.

Proof. Let

$$f(u) = e^{-1/u}\mathbf{1}_{\{u>0\}} \qquad \text{and} \qquad g_0(u) = \frac{f(1 - u)}{f(1 - u) + f(u)}.$$

The denominator is strictly positive for every $u \in \mathbb{R}$, so $g_0 \in C^\infty(\mathbb{R}; [0, 1])$ with $g_0 \equiv 1$ on $(-\infty, 0]$, $g_0 \equiv 0$ on $[1, \infty)$, and universal bounds

$$\|g_0'\|_\infty,\ \|g_0''\|_\infty \leq C_0.$$

Put

$$g(u) := g_0\left(\frac{u - a}{b - a}\right) \qquad \text{and} \qquad \psi_{a,b}(y) := g(\|y\|).$$

Since $\psi_{a,b} \equiv 1$ on a neighborhood of the origin and $y \mapsto \|y\|$ is smooth on $\mathbb{R}^d \setminus \{0\}$, we have $\psi_{a,b} \in C^\infty(\mathbb{R}^d)$. Writing $r = \|y\|$ and $\hat{y} = y/r$,

$$\nabla\psi_{a,b}(y) = g'(r)\,\hat{y}, \qquad \nabla^2\psi_{a,b}(y) = g''(r)\,\hat{y}\hat{y}^\top + \frac{g'(r)}{r}\big(\mathrm{Id} - \hat{y}\hat{y}^\top\big), \tag{B.2}$$

and g', g'' vanish outside $\{a \le r \le b\}$, where $r \ge a$. The bounds follow directly from

$$\|g'\|_\infty \le \frac{C_0}{b-a} \qquad \text{and} \qquad \|g''\|_\infty \le \frac{C_0}{(b-a)^2}.$$

The eigenvalues of $\nabla^2 \psi_{a,b}$ are $g''(r)$ and $g'(r)/r$ irrespective of dimension, so no estimate depends on d. □

Theorem 4.1 uses $\psi_{4\sqrt{d},\, b_\xi\sqrt{d}}$ with $b_\xi := \frac{1}{2}(3+\xi)$, which satisfies $4 < b_\xi < \xi - 1$ for every $\xi > 5$. Since $b_\xi - 4 = \frac{1}{2}(\xi - 5)$, the gradient and Hessian bounds are $C_\xi/\sqrt{d}$ and C_ξ/d, with C_ξ depending only on ξ (of order $(\xi-5)^{-1}$ and $(\xi-5)^{-2}$ respectively). Theorem 4.2 uses $\psi_{0.15\sqrt{d},\, 0.16\sqrt{d}}(\cdot - x_i)$, with bounds $O(1/\sqrt{d})$ and $O(1/d)$, and Theorem 4.3 uses $\psi_{\theta\sqrt{d},\, 2\theta\sqrt{d}}(\cdot - x_\star)$, with bounds $C_0/(\theta\sqrt{d})$ and $2C_0/(\theta^2 d)$.

C THE KINETIC EXIT-TIME LEMMA

The following lemma quantifies how long a stiff underdamped oscillator remains inside a high-dimensional ball. It is the dynamical input used to turn the localized potential constructions in Appendices D and E into exponentially long trapping statements.

Lemma C.1. *Fix $\gamma > 0$ and constants $r > \rho_x \ge 0$, $\rho_v \ge 0$, and write $m := r - \rho_x$. Let $(x_t, v_t)_{t\ge 0}$ solve the linear SDE*

$$\mathrm{d}x_t = v_t\,\mathrm{d}t, \qquad \mathrm{d}v_t = (-\alpha x_t - \gamma v_t)\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t,$$

where B_t is a standard d-dimensional Brownian motion and the deterministic initial condition satisfies $\|x_0\| \le \rho_x\sqrt{d}$ and $\|v_0\| \le \rho_v\sqrt{d}$. Let $\tau = \inf\{t \ge 0 : \|x_t\| \ge r\sqrt{d}\}$. There is a universal constant C' such that, setting

$$\alpha_\star := \max\left\{\frac{\gamma^2}{2},\, \frac{C'\big(1+\gamma^2\rho_x^2+\rho_v^2\big)}{m^2}\right\}, \qquad c_\star := \frac{\alpha_\star m^2}{64},$$

the following holds: there are $C, K_\star > 0$, depending only on $\gamma, r, \rho_x, \rho_v$, such that for all $\alpha \ge \alpha_\star$, $d \ge K_\star$, and $T \le e^{c_\star d}$,

$$\mathbb{P}(\tau \le T) \;\le\; Ce^{-c_\star d}.$$

The bound is uniform over all deterministic initial data satisfying the stated norm constraints.

Proof. Throughout, we write

$$m := r - \rho_x > 0 \tag{C.1}$$

for the available margin. Consider one Cartesian coordinate; the d coordinates are independent and identically distributed. The scalar process solves $\ddot{x} = -\alpha x - \gamma\dot{x} + \sqrt{2\gamma}\,\dot{B}$.

In the underdamped regime $\alpha > \gamma^2/4$, we set $\omega = \sqrt{\alpha - \gamma^2/4}$. The Green's function of $\ddot{G} + \gamma\dot{G} + \alpha G = \delta$ is given and bounded by

$$G(u) = \frac{1}{\omega}e^{-\gamma u/2}\sin(\omega u) \quad (u \ge 0),$$
$$|G(u)| \le \frac{1}{\omega}e^{-\gamma u/2} \tag{C.2}$$

Solving by variation of parameters, and writing h_0, h_1 for the position responses to a unit initial position and velocity, we have

$$x_t = h_0(t)x_0 + h_1(t)v_0 + \sqrt{2\gamma}\underbrace{\int_0^t G(t-s)\,\mathrm{d}B_s}_{=:W_t}, \tag{C.3}$$

$$h_0(t) = e^{-\gamma t/2}\left(\cos\omega t + \frac{\gamma}{2\omega}\sin\omega t\right),$$
$$h_1(t) = G(t),$$

where $W_t \in \mathbb{R}^d$ has i.i.d. coordinates.

Stationary Scale: Using $\gamma^2+4\omega^2 = 4\alpha$ and the identity $\sin^2(\theta) = \frac{1}{2}(1-\cos 2\theta)$, we can explicitly compute the integral:

$$\begin{aligned}\int_0^\infty G(r)^2\,\mathrm{d}r &= \frac{1}{2\omega^2}\left(\frac{1}{\gamma} - \frac{\gamma}{\gamma^2+4\omega^2}\right)\\ &= \frac{1}{2\omega^2}\cdot\frac{4\omega^2}{4\alpha\gamma}\\ &= \frac{1}{2\alpha\gamma}\end{aligned} \tag{C.4}$$

Therefore, each coordinate of $\sqrt{2\gamma}W_t$ has variance $2\gamma\int_0^t G^2 \le 1/\alpha$, uniformly in t. Equivalently, applying the equipartition theorem $\frac{1}{2}\alpha\langle x^2\rangle = \frac{1}{2}$ yields $\langle x^2\rangle = 1/\alpha$.

Homogeneous Part: Assume $\alpha \ge \gamma^2/2$, so that $\omega \ge \sqrt{\alpha/2}$, which implies $\frac{\gamma}{2\omega} \le \frac{\gamma}{\sqrt{2\alpha}}$ and $\frac{1}{\omega} \le \sqrt{2/\alpha}$.

With the initial bounds $\|x_0\| \le \rho_x\sqrt{d}$, $\|v_0\| \le \rho_v\sqrt{d}$, and $e^{-\gamma t/2} \le 1$, we obtain the following bound on the homogeneous evolution:

$$\begin{aligned}\|h_0(t)x_0 + h_1(t)v_0\| &\le \left(1+\frac{\gamma}{\sqrt{2\alpha}}\right)\rho_x\sqrt{d} + \sqrt{\frac{2}{\alpha}}\,\rho_v\sqrt{d}\\ &= (\rho_x+\delta_\alpha)\sqrt{d},\end{aligned} \tag{C.5}$$

where we have defined

$$\delta_\alpha := \frac{\gamma\rho_x}{\sqrt{2\alpha}} + \frac{\sqrt{2}\,\rho_v}{\sqrt{\alpha}}.$$

Since $\delta_\alpha \to 0$ as $\alpha\to\infty$, we may fix a stiffness threshold α_1:

$$\alpha_1 := \frac{16}{m^2}\left(\frac{\gamma\rho_x}{\sqrt{2}} + \sqrt{2}\,\rho_v\right)^2, \tag{C.6}$$

$$\text{so that}\qquad \delta_\alpha \le \frac{m}{4}\qquad \text{for all } \alpha \ge \alpha_1.$$

The homogeneous part is then at most $(\rho_x + \frac{m}{4})\sqrt{d}$ for every $t \ge 0$. Note that the velocity enters only through $\rho_v/\sqrt{\alpha}$. A large initial velocity is admissible provided α is taken correspondingly large, which is precisely the mechanism behind the choice of (C.6).

Stochastic Part: Write $X_t := \sqrt{2\gamma}W_t$ and $Z_{t,u} := \langle X_t, u\rangle$ for $u \in \mathbb{S}^{d-1}$. This defines a centered Gaussian process on $\mathcal{T} := [0,T]\times\mathbb{S}^{d-1}$ with variance bounded by

$$\sigma^2 := \sup_{t,u}\mathbb{E}Z_{t,u}^2 \le \frac{1}{\alpha} \tag{C.7}$$

by (C.4), and with $\sup_{t\le T}\|X_t\| = \sup_{\mathcal{T}} Z_{t,u}$. We bound the supremum over a unit window $\mathcal{T}_1 := [0,1]\times\mathbb{S}^{d-1}$ and then tile the space.

For $0 \le t_1 \le t_2$ with $h := t_2 - t_1 \le 1$ and $u_1, u_2 \in \mathbb{S}^{d-1}$, Itô's isometry gives

$$\begin{aligned}\mathbb{E}\big(Z_{t_1,u_1} - Z_{t_2,u_2}\big)^2 &= 2\gamma\int_0^{t_1}\|G(t_1-s)u_1 - G(t_2-s)u_2\|^2\,\mathrm{d}s + 2\gamma\int_{t_1}^{t_2} G(t_2-s)^2\,\mathrm{d}s\\ &\le \underbrace{4\gamma\int_0^\infty G^2}_{=2/\alpha}\|u_1-u_2\|^2 + 4\gamma\int_0^\infty\big(G(\rho)-G(\rho+h)\big)^2\,\mathrm{d}\rho + 2\gamma\int_0^h G^2.\end{aligned} \tag{C.8}$$

Since $G'(u) = e^{-\gamma u/2}\left(\cos\omega u - \frac{\gamma}{2\omega}\sin\omega u\right)$ satisfies $|G'(u)| \le 2e^{-\gamma u/2}$ for $\alpha \ge \gamma^2/2$, we have $\int_0^\infty (G')^2 \le 4/\gamma$ and hence, by Cauchy–Schwarz and Fubini, $\int_0^\infty (G(\rho)-G(\rho+h))^2\,\mathrm{d}\rho \le$

$h^2 \int_0^\infty (G')^2 \le 4h^2/\gamma$. Moreover, $\int_0^h G^2 \le h/\omega^2 \le 2h/\alpha$. Substituting these into (C.8) yields the metric bounds:

$$d\big((t_1,u_1),(t_2,u_2)\big)^2 \le \frac{2}{\alpha}\|u_1-u_2\|^2 + 16h^2 + \frac{4\gamma h}{\alpha}, \tag{C.9}$$
$$\text{and globally,} \qquad d \le \frac{2}{\sqrt{\alpha}}$$

by (C.7). Consequently, $d \le \varepsilon$ holds whenever $\|u_1-u_2\| \le \varepsilon\sqrt{\alpha}/\sqrt{6}$ and $h \le c\min\{\varepsilon, \varepsilon^2\alpha/\gamma\}$. So with $\operatorname{diam}(\mathcal{T}_1) \le 2/\sqrt{\alpha}$, the covering numbers obey

$$\begin{aligned} N\big(\mathcal{T}_1, d, \varepsilon\big) &\le N\big(\mathbb{S}^{d-1}, \|\cdot\|, c\varepsilon\sqrt{\alpha}\big) \cdot N\big([0,1], |\cdot|, c\min\{\varepsilon, \varepsilon^2\alpha/\gamma\}\big) \\ &\le \left(\frac{C}{\varepsilon\sqrt{\alpha}}\right)^d \left(1 + \frac{C}{\varepsilon} + \frac{C\gamma}{\varepsilon^2\alpha}\right). \end{aligned} \tag{C.10}$$

The second factor is the contribution of the time direction. Unlike the sphere, it carries no exponential factor in d.

Dudley's chaining with the substitution $\varepsilon = w/\sqrt{\alpha}$ gives

$$\begin{aligned} \mathbb{E}\sup_{\mathcal{T}_1} Z_{t,u} &\le C\int_0^{2/\sqrt{\alpha}} \sqrt{\log N(\mathcal{T}_1, d, \varepsilon)}\,\mathrm{d}\varepsilon \\ &\le \frac{C}{\sqrt{\alpha}}\int_0^2 \left(\sqrt{d\log\frac{C}{w}} + \sqrt{\log\left(1 + \frac{C\sqrt{\alpha}}{w} + \frac{C\gamma}{w^2}\right)}\right)\mathrm{d}w \\ &\le \frac{C_D\sqrt{d}}{\sqrt{\alpha}} + R(\alpha,\gamma), \end{aligned} \tag{C.11}$$

where

$$R(\alpha,\gamma) := \frac{C}{\sqrt{\alpha}}\int_0^2 \sqrt{\log\left(1 + \frac{C\sqrt{\alpha}}{w} + \frac{C\gamma}{w^2}\right)}\,\mathrm{d}w.$$

An elementary bound gives

$$R(\alpha,\gamma) \le \frac{C}{\sqrt{\alpha}}\left(1 + \sqrt{\log(1+\alpha)} + \sqrt{\log(1+\gamma)}\right),$$

so $R(\alpha,\gamma) \to 0$ as $\alpha \to \infty$.

We fix a second stiffness threshold:

$$\alpha_2 := \frac{64C_D^2}{m^2}, \tag{C.12}$$
$$\text{so that} \qquad \frac{C_D}{\sqrt{\alpha}} \le \frac{m}{8} \qquad \text{for all } \alpha \ge \alpha_2.$$

Let

$$R_\star := \sup_{\alpha \ge \max\{\gamma^2/2, \alpha_2\}} R(\alpha,\gamma) < \infty, \qquad K_\star := \max\left\{1, \frac{64R_\star^2}{m^2}\right\}.$$

Then for every $\alpha \ge \max\{\gamma^2/2, \alpha_2\}$ and $d \ge K_\star$,

$$\mathbb{E}\sup_{\mathcal{T}_1} Z_{t,u} \le \frac{m}{4}\sqrt{d}.$$

By the translation invariance of (C.9) in t and the uniform variance bound (C.7), the same bound holds with $[0,1]$ replaced by any unit window $[k, k+1]$.

The Borell–TIS inequality (Adler & Taylor, 2007), applied on each window with $\sigma^2 \le 1/\alpha$ and deviation $s = \frac{m}{4}\sqrt{d}$, gives

$$\mathbb{P}\left(\sup_{[k,k+1]\times\mathbb{S}^{d-1}} Z_{t,u} \ge \frac{m}{2}\sqrt{d}\right) \le \exp\left(-\frac{\alpha m^2 d}{32}\right). \tag{C.13}$$

Conclusion: Set the overall required stiffness and corresponding exponential rate as:

$$\alpha_\star := \max\left\{\frac{\gamma^2}{2},\, \alpha_1,\, \alpha_2\right\} = \max\left\{\frac{\gamma^2}{2},\, \frac{C(1+\gamma^2\rho_x^2+\rho_v^2)}{(r-\rho_x)^2}\right\}, \tag{C.14}$$
$$c_\star := \frac{\alpha_\star m^2}{64}.$$

Let $\alpha \geq \alpha_\star$, $d \geq K_\star$, and $T \leq e^{c_\star d}$. A union bound over the $\lceil T \rceil \leq e^{c_\star d} + 1$ unit windows bounds the stochastic term:

$$\begin{aligned}\mathbb{P}\left(\sup_{t\leq T}\|X_t\| \geq \frac{m}{2}\sqrt{d}\right) &\leq (e^{c_\star d}+1)e^{-\alpha m^2 d/32} \\ &\leq 2e^{c_\star d}e^{-2c_\star d} \\ &= 2e^{-c_\star d},\end{aligned} \tag{C.15}$$

using the fact that $\alpha m^2/32 \geq \alpha_\star m^2/32 = 2c_\star$.

On the complementary event, combining this stochastic bound with the homogeneous bound from (C.5) and (C.6), we find that the total position is strictly bounded:

$$\begin{aligned}\sup_{t\leq T}\|x_t\| &\leq \left(\rho_x + \frac{m}{4}\right)\sqrt{d} + \frac{m}{2}\sqrt{d} \\ &= \left(\rho_x + \frac{3m}{4}\right)\sqrt{d} \\ &< (\rho_x + m)\sqrt{d} \\ &= r\sqrt{d}.\end{aligned} \tag{C.16}$$

Because the trajectory strictly remains inside radius $r\sqrt{d}$, we have $\tau > T$. Hence, $\mathbb{P}(\tau \leq T) \leq Ce^{-c_\star d}$. □

Remark C.2. Whenever the second term in (C.14) attains the maximum, $\alpha_\star$ scales as $(r-\rho_x)^{-2}$, while $c_\star = \alpha_\star(r-\rho_x)^2/64$ does not depend on the radii: at the threshold stiffness, $c_\star$ is a constant determined by $\gamma\rho_x$ and ρ_v. The exponential time horizon $e^{c_\star d}$ therefore does not deteriorate as the exit radius shrinks, provided α is increased accordingly. This is what allows the lemma to be applied at $r = 4$ in Theorem 4.1 and at $r = 0.10$ in Theorem 4.2 without losing the exponential-in-d time horizon.

D Proof of the kinetic-trapping theorem

Proof of Theorem 4.1. Recall $V(x) = \frac{1}{2}\|x-\mu\|^2$ with $\|\mu\| = \xi\sqrt{d}$ and $\xi > 5$, so that $\pi_{\text{tar}} = \mathcal{N}(\mu, I_d)$. We set

$$b_\xi := \frac{1}{2}(3+\xi), \qquad \text{so that} \qquad 4 < b_\xi < \xi - 1, \tag{D.1}$$

both inequalities being equivalent to $\xi > 5$. We fix

$$\alpha \;\geq\; \max\left\{\alpha_\star(\gamma, 4, 1.1, 1.1),\; \tfrac{1}{16}(b_\xi+\xi)^2\right\}, \tag{D.2}$$

with $\alpha_\star$ as in Lemma C.1; both terms are constants depending only on γ and ξ. The second is used only in Appendix G.

Construction: Let $\chi := \psi_{4\sqrt{d},\, b_\xi\sqrt{d}}$ be the radial cutoff of Lemma B.3, so that $\chi \equiv 1$ on $\{\|x\| \leq 4\sqrt{d}\}$ and $\chi \equiv 0$ on $\{\|x\| \geq b_\xi\sqrt{d}\}$. We put

$$U(x) := \frac{\alpha}{2}\|x\|^2, \qquad \hat{V}(x) := V(x) + \chi(x)\big(U(x) - V(x)\big), \qquad \hat{F} := -\nabla\hat{V}. \tag{D.3}$$

Then $\hat{V} \in C^\infty(\mathbb{R}^d)$ and $\hat{F}$ is conservative by construction. Moreover, $\hat{F}(x) = -\alpha x$ on the region $\{\|x\| \leq 4\sqrt{d}\}$, while $\hat{V} = V$ and $\hat{F} = F$ on $\{\|x\| \geq b_\xi\sqrt{d}\}$.

Regularity and well-posedness: On the set $B := \{\|x\| \le b_\xi\sqrt{d}\}$, we have $\|x-\mu\| \le (b_\xi+\xi)\sqrt{d}$. Hence, the pointwise gaps are bounded by

$$|U(x)-V(x)| = O_\xi(\alpha d), \qquad \|\nabla U(x)-\nabla V(x)\| = O_\xi(\alpha\sqrt{d}) \qquad \text{on } B, \tag{D.4}$$

while by Lemma B.3 we have $\|\nabla\chi\| = O_\xi(1/\sqrt{d})$ and $\|\nabla^2\chi\|_{\mathrm{op}} = O_\xi(1/d)$. Differentiating (D.3) twice yields

$$\begin{aligned}\nabla^2\hat{V} = \nabla^2 V &+ \chi\big(\nabla^2 U - \nabla^2 V\big)\\ &+ \nabla\chi\,(\nabla U-\nabla V)^\top + (\nabla U-\nabla V)\nabla\chi^\top\\ &+ (U-V)\nabla^2\chi.\end{aligned}$$

Each term in this expansion is $O_\xi(\alpha)$ by (D.4). Hence, $\sup_x\|\nabla^2\hat{V}(x)\|_{\mathrm{op}} \le H$ for a constant $H = O_\xi(\alpha)$ that is independent of d. Consequently, $\hat{F}$ is globally Lipschitz, and (2) has a unique strong solution by Øksendal (2003, Thm. 5.2.1).

Target mass of the modified region: Write $x = \mu + Z$ with $Z \sim \mathcal{N}(0, I_d)$. If $\|x\| \le b_\xi\sqrt{d}$, then

$$\|Z\| \ge \|\mu\| - \|x\| \ge (\xi - b_\xi)\sqrt{d},$$

and we know $\xi - b_\xi = \frac{1}{2}(\xi-3) > 1$ by (D.1). Applying Lemma B.1 with $t = \frac{1}{2}(\xi-5) > 0$, we bound the mass of the modified region:

$$\begin{aligned}\pi_{\mathrm{tar}}(B) &\le \mathbb{P}\left(\|Z\| \ge \left(1+\frac{\xi-5}{2}\right)\sqrt{d}\right)\\ &\le Ce^{-c(\xi-5)^2 d/4}\\ &= e^{-\Omega(d)}\end{aligned} \tag{D.5}$$

The exact same argument, but with 4 in place of b_ξ and using the fact that $\xi - 4 > 1$, gives

$$\pi_{\mathrm{tar}}(A) \ge 1 - e^{-\Omega(d)}, \qquad \text{where} \qquad A := \{\|x\| \ge 4\sqrt{d}\}. \tag{D.6}$$

Population L^p errors: Both $\hat{V}-V$ and $\hat{F}-F$ vanish outside of B. On B, they are bounded by (D.4) together with the $\nabla\chi$ contribution $|U-V|\|\nabla\chi\| = O_\xi(\alpha\sqrt{d})$. Hence, for any fixed $p \ge 1$, we can bound the population error for the forces:

$$\begin{aligned}\mathbb{E}_{x\sim\pi_{\mathrm{tar}}}\big[\|\hat{F}(x)-F(x)\|^p\big]^{1/p} &\le O_\xi(\alpha\sqrt{d})\,\pi_{\mathrm{tar}}(B)^{1/p}\\ &\le O_\xi(\alpha\sqrt{d})\,e^{-\Omega(d)/p}\\ &= e^{-\Omega(d)}\end{aligned} \tag{D.7}$$

Identically, we find $\mathbb{E}_{x\sim\pi_{\mathrm{tar}}}\big[|\hat{V}(x)-V(x)|^p\big]^{1/p} \le O_\xi(\alpha d)\,e^{-\Omega(d)/p} = e^{-\Omega(d)}$. In both cases, the bounds hold for all $d \ge d_0(p,\xi)$, as the polynomial prefactors are naturally absorbed once d is large relative to p. This concludes part (i).

Total-variation lower bound: Define the high-probability initialization event

$$G_0 := \big\{\|x_0\| \le 1.1\sqrt{d},\ \|v_0\| \le 1.1\sqrt{d}\big\}.$$

By Lemma B.1, $\mathbb{P}(G_0) \ge 1 - e^{-\Omega(d)}$. While $\|x_t\| \le 4\sqrt{d}$, we have $\hat{F}(x_t) = -\alpha x_t$. Therefore, up to the exit time $\tau = \inf\{t : \|x_t\| \ge 4\sqrt{d}\}$, the process coincides in law with the linear kinetic Langevin dynamics analyzed in Lemma C.1.

That bound holds uniformly over any deterministic initial data within the ball of (C.5). Hence, it also holds after conditioning on G_0 and integrating over (x_0, v_0). Applying Lemma C.1 with $r = 4$ and $\rho_x = \rho_v = 1.1$ gives

$$\mathbb{P}(\tau \le T \mid G_0) \le Ce^{-c_\star d} \qquad \text{for every } T \le e^{c_\star d},$$

and therefore $\mathbb{P}(x_T \in A \mid G_0) \le Ce^{-c_\star d}$. Consequently, we can bound the probability of the process being in A at time T:

$$\begin{aligned}\hat{\pi}_T(A) &\le \mathbb{P}(x_T \in A \mid G_0) + \mathbb{P}(G_0^c) \\ &\le e^{-\Omega(d)}.\end{aligned} \tag{D.8}$$

Combining this with (D.6), we obtain the total variation bound for every $T \le e^{c_\star d}$:

$$\begin{aligned}\mathrm{TV}\big(\hat{\pi}_T, \pi_{\mathrm{tar}}\big) &\ge \pi_{\mathrm{tar}}(A) - \hat{\pi}_T(A) \\ &\ge 1 - e^{-\Omega(d)},\end{aligned} \tag{D.9}$$

which completes part (ii). □

Corollary D.1. *For the potential of Theorem 4.1, the conclusion of part (ii) also holds for any deterministic initialization with $\|x_0\|, \|v_0\| \le 1.1\sqrt{d}$. Consequently, the mixing time of Langevin dynamics with $\hat{F}$ to any distribution within constant* TV *of π_{tar} is at least e^{cd}.*

The proof is immediate because the argument depends on the initialization only through the event G_0, and Lemma C.1 applies to any deterministic initialization (x_0, v_0) satisfying $\|x_0\|, \|v_0\| \le 1.1\sqrt{d}$.

Remark D.2. The proof differs from the overdamped setting of Cao et al. (2026) because inertia changes the short-time dynamics. In overdamped Langevin, the trapped process is an Ornstein-Uhlenbeck process, so the position is Gaussian and can be controlled directly. In kinetic Langevin, by contrast, the position is the time integral of an OU-type velocity process with Green's function

$$G(u) = \omega^{-1} e^{-\frac{\gamma u}{2}} \sin(\omega u), \qquad \omega = \sqrt{\alpha - \tfrac{\gamma^2}{4}},$$

and large initial velocities can produce ballistic excursions before friction dominates. The proof is therefore based on a hypoelliptic exit-time bound.

Remark D.3. The construction extends verbatim to strongly log-concave targets $\pi_{\mathrm{tar}} \propto e^{-V}$ with ∇V Lipschitz, using concentration of Lipschitz functions of log-concave measures (Bakry et al., 2014). General β follows by rescaling $x \mapsto \sqrt{\beta}\, x$, $t \mapsto t$, which maps (2) to the $\beta = 1$ equation with a rescaled potential.

Remark D.4. Since $\hat{F}$ is Lipschitz, by Girsanov's theorem any discretization $\hat{x}_T$ of the SDE (e.g. BAOAB with suitable step size; Leimkuhler & Matthews, 2015) satisfying

$$\mathrm{TV}\big(\mathrm{Law}(\hat{x}_T), \mathrm{Law}(x_T)\big) \le c$$

inherits $\mathrm{TV}(\mathrm{Law}(\hat{x}_T), \pi_{\mathrm{tar}}) \ge 1 - c - e^{-\Omega(d)}$ by the triangle inequality.

E Proof of the data-interpolation theorem

Proof of Theorem 4.2. Throughout, we write $z_i := x_i - \mu$ for the centered training samples and $Z := x - \mu$ for a centered draw $x \sim \pi_{\mathrm{tar}}$, so that

$$z_1, \dots, z_n, Z \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, I_d).$$

Recall the definitions of the unperturbed energy and force:

$$V(x) = \frac{1}{2}\|x - \mu\|^2 \qquad \text{and} \qquad F(x) = -(x - \mu).$$

Fix the stiffness parameter to be sufficiently large:

$$\alpha \ge \alpha_0 := \max\big\{40,\ \alpha_\star(\gamma, 0.10, 0.05, 1.1)\big\}, \tag{E.1}$$

with $\alpha_\star$ as in Lemma C.1; note that α_0 is a constant independent of d and n.

Let $\psi \in C^\infty(\mathbb{R})$ satisfy $\psi \equiv 1$ on $(-\infty, 0.0225]$, $\psi \equiv 0$ on $[0.0256, \infty)$, and $0 \le \psi \le 1$. We set

$$\chi_i(x) := \psi\left(\frac{\|x - x_i\|^2}{d}\right),$$

so that the cutoff function satisfies

$$\chi_i \equiv 1 \quad \text{on } \{\|x - x_i\| \le 0.15\sqrt{d}\} \qquad \text{and} \qquad \chi_i \equiv 0 \quad \text{on } \{\|x - x_i\| \ge 0.16\sqrt{d}\}.$$

Define the well centers and offsets

$$w_i := x_i - \frac{1}{\alpha} z_i, \qquad c_i := \frac{1}{2}\left(1 - \frac{1}{\alpha}\right)\|z_i\|^2, \qquad U_i(x) := \frac{\alpha}{2}\|x - w_i\|^2 + c_i, \tag{E.2}$$

and set the modified global energy and force as

$$\hat{V}(x) := V(x) + \sum_{i=1}^{n} \chi_i(x)\big(U_i(x) - V(x)\big), \qquad \hat{F} := -\nabla \hat{V}. \tag{E.3}$$

The offset w_i displaces the well center from x_i toward μ by a $1/\alpha$ fraction, and c_i is the unique constant making $\hat{V}$ agree with V at x_i; these two choices are what give the exact fit asserted in part (i).

General-position event: For $i \neq j$, we have $z_i - z_j \sim \mathcal{N}(0, 2I_d)$, so $\mathbb{E}\|z_i - z_j\| = (1 + o(1))\sqrt{2d}$. By Lemma B.1, we obtain the high-probability bounds:

$$\|x_i - x_j\| = \|z_i - z_j\| \ge 0.4\sqrt{d} \qquad \text{with probability } 1 - e^{-\Omega(d)},$$
$$\|z_i\| \in [0.5\sqrt{d}, 2\sqrt{d}] \qquad \text{with probability } 1 - e^{-\Omega(d)}.$$

A union bound over the $\binom{n}{2} + n = \mathrm{poly}(d)$ events gives, with probability $1 - e^{-\Omega(d)}$,

$$\min_{i \neq j} \|x_i - x_j\| \ge 0.4\sqrt{d} \qquad \text{and} \qquad 0.5\sqrt{d} \le \|z_i\| \le 2\sqrt{d} \quad (1 \le i \le n). \tag{E.4}$$

Since $0.4\sqrt{d} > 2 \cdot 0.16\sqrt{d}$, the balls $B_i := \{\|x - x_i\| \le 0.16\sqrt{d}\}$ are pairwise disjoint on this event, so at most one summand in (E.3) is nonzero at any given x. In particular, $\hat{V} \in C^\infty(\mathbb{R}^d)$ and $\hat{F}$ is conservative by construction. We work on the event (E.4) for the remainder of the proof and record the consequence

$$\|x_i - w_i\| = \frac{1}{\alpha}\|z_i\| \le \frac{2}{\alpha}\sqrt{d} \le 0.05\sqrt{d}, \tag{E.5}$$

where the last inequality uses $\alpha \ge 40$.

Interpolation: Fix i. Since $\chi_i(x_i) = 1$ and $\chi_j(x_i) = 0$ for $j \neq i$, we have $\hat{V} = U_i$ and $\hat{F} = -\nabla U_i$ in a neighborhood of x_i, whence

$$\hat{F}(x_i) = -\alpha(x_i - w_i) = -\alpha \cdot \frac{1}{\alpha} z_i = -z_i = F(x_i),$$

and, using $\|x_i - w_i\|^2 = \|z_i\|^2/\alpha^2$, we can compute the energy match directly:

$$\begin{aligned}
\hat{V}(x_i) &= \frac{\alpha}{2} \cdot \frac{\|z_i\|^2}{\alpha^2} + c_i \\
&= \frac{1}{2\alpha}\|z_i\|^2 + \frac{1}{2}\left(1 - \frac{1}{\alpha}\right)\|z_i\|^2 \\
&= \frac{1}{2}\|z_i\|^2 = V(x_i).
\end{aligned}$$

The empirical energy-and-force matching loss of $\hat{V}$ therefore vanishes identically, which is part (i).

Regularity and well-posedness: On B_i, we have the distance bounds

$$\|x - \mu\| \le \|x - x_i\| + \|z_i\| \le 2.16\sqrt{d} \qquad \text{and, by (E.5)}, \qquad \|x - w_i\| \le 0.21\sqrt{d},$$

so that the point-wise gaps are bounded by

$$|U_i(x) - V(x)| = O(\alpha d), \qquad \|\nabla U_i(x) - \nabla V(x)\| = O(\alpha\sqrt{d}) \qquad \text{on } B_i. \tag{E.6}$$

Moreover, since χ_i is a fixed smooth function of $\|x - x_i\|^2/d$ evaluated where $\|x - x_i\| = O(\sqrt{d})$, we have

$$\|\nabla \chi_i\| = O(1/\sqrt{d}) \qquad \text{and} \qquad \|\nabla^2 \chi_i\|_{\mathrm{op}} = O(1/d).$$

Differentiating (E.3) twice and combining with (E.6), every resulting term is $O(\alpha)$. Hence,

$$\sup_x \|\nabla^2 \hat{V}(x)\|_{\mathrm{op}} \le H$$

for a constant $H = O(\alpha)$ independent of d and n. In particular, $\hat{F}$ is globally Lipschitz and (2) admits a unique strong solution.

Target mass of the modified region: Let $A := \{x : \min_i \|x - x_i\| \ge 0.16\sqrt{d}\}$, so that $A^c = \bigcup_i B_i$ and $\hat{V} = V$ on A. Fix i and consider $g(Z) := \langle Z, z_i\rangle / \|z_i\|$, which is $\mathcal{N}(0,1)$ under π_{tar}. If $x \in B_i$, then

$$|g(Z) - \|z_i\|| = \frac{|\langle Z - z_i, z_i\rangle|}{\|z_i\|} \le \|Z - z_i\| \le 0.16\sqrt{d}.$$

So by (E.4), we are bounded away from the center:

$$g(Z) \ge \|z_i\| - 0.16\sqrt{d} \ge 0.34\sqrt{d}.$$

The standard Gaussian tail bound therefore gives

$$\pi_{\mathrm{tar}}(B_i) \le \mathbb{P}\big(\mathcal{N}(0,1) \ge 0.34\sqrt{d}\big) \le e^{-0.05d},$$

and a union bound over the $n = \mathrm{poly}(d)$ balls yields

$$\pi_{\mathrm{tar}}(A^c) \le n e^{-0.05d} = e^{-\Omega(d)}, \qquad \text{equivalently,} \qquad \pi_{\mathrm{tar}}(A) \ge 1 - e^{-\Omega(d)}. \tag{E.7}$$

Population L^p errors: Both $\hat{V} - V$ and $\hat{F} - F$ vanish on A. On A^c, they are bounded by (E.6) together with the corresponding bounds on the annuli, which are of the same order. Hence, for fixed $p \ge 1$, we obtain the bounds

$$\mathbb{E}_{x\sim\pi_{\mathrm{tar}}}\big[\|\hat{F}(x) - F(x)\|^p\big]^{1/p} \le O(\alpha\sqrt{d}) \cdot \pi_{\mathrm{tar}}(A^c)^{1/p} \le O(\alpha\sqrt{d})\, e^{-\Omega(d)/p} = e^{-\Omega(d)},$$
$$\mathbb{E}_{x\sim\pi_{\mathrm{tar}}}\big[|\hat{V}(x) - V(x)|^p\big]^{1/p} \le O(\alpha d)\, e^{-\Omega(d)/p} = e^{-\Omega(d)}.$$

In both cases, these hold for all $d \ge d_0(p)$, as the polynomial prefactors are absorbed once d is large relative to p. This is part (ii).

Total-variation lower bound: Condition on $x_0 = x_i$ and write Π_i for the corresponding law. Since $v_0 \sim \mathcal{N}(0, I_d)$, Lemma B.1 gives

$$\|v_0\| \le 1.1\sqrt{d} \qquad \text{with probability } 1 - e^{-\Omega(d)},$$

and we work on this event. In the shifted coordinates $y_t := x_t - w_i$, we have $\hat{F}(x_t) = -\alpha y_t$ for as long as $x_t \in \{\|x - x_i\| \le 0.15\sqrt{d}\}$. Thus, (y_t, v_t) solves the linear SDE of Lemma C.1 with the confining center at the origin. By (E.5),

$$\|y_0\| = \|x_i - w_i\| \le 0.05\sqrt{d}, \qquad \text{and} \qquad \big\{\|y\| \le 0.10\sqrt{d}\big\} \subseteq \big\{\|x - x_i\| \le 0.15\sqrt{d}\big\},$$

so Lemma C.1 applies with $(r, \rho_x, \rho_v) = (0.10, 0.05, 1.1)$, which is legitimate by the choice (E.1) of α. The exit time $\tau = \inf\{t : \|y_t\| \ge 0.10\sqrt{d}\}$ therefore satisfies

$$\mathbb{P}_{\Pi_i}(\tau \le T) \le C e^{-c_\star d} \qquad \text{for every } T \le e^{c_\star d}.$$

Since $\{\|x - x_i\| \le 0.15\sqrt{d}\} \subseteq B_i \subseteq A^c$, we obtain

$$\mathbb{P}_{\Pi_i}(x_T \in A) \le C e^{-c_\star d} + e^{-\Omega(d)} = e^{-\Omega(d)} \qquad \text{uniformly in } i,$$

where the second term accounts for the velocity event. Averaging over the initialization $x_0 \sim \frac{1}{n}\sum_i \delta_{x_i}$ gives

$$\hat{\pi}_T(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{P}_{\Pi_i}(x_T \in A) \le e^{-\Omega(d)}.$$

Combining this with (E.7), we conclude

$$\begin{aligned}\mathrm{TV}\big(\hat{\pi}_T, \pi_{\mathrm{tar}}\big) &\geq \pi_{\mathrm{tar}}(A) - \hat{\pi}_T(A)\\ &\geq 1 - e^{-\Omega(d)}\end{aligned}$$

for all $T \leq e^{c_\star d}$ and all d larger than a suitable constant, which completes part (iii). □

Remark E.1. The quadratic form in Theorem 4.2 allows for an explicit exit-time calculation. A smooth annular-barrier construction can likewise coincide with V in a neighbourhood of every training configuration preserving all local derivatives while exhibiting exponentially long metastable confinement.

Remark E.2. The trapping mechanism extends beyond the target training distribution. Suppose a training distribution ν assigns mass $e^{-\Omega(d)}$ to an admissible region B_d, and $\pi_{\mathrm{tar}}(B_d) = e^{-\Omega(d)}$. A perturbation supported on B_d then has exponentially small energy and force errors under ν. If the learned dynamics is initialized inside a trap placed in B_d, the exit-time argument of Lemma C.1 yields $\mathrm{TV}(\hat{\pi}_T, \pi_{\mathrm{tar}}) \geq 1 - e^{-\Omega(d)}$ for $T \leq e^{\Omega(d)}$.

F Proof of the equilibrium-mismatch theorem

Assumption F.1 (Confinement)**.** $V \in C^\infty(\mathbb{R}^d)$, $\sup_x \left\|\nabla^2 V(x)\right\|_{\mathrm{op}} \leq L$, *and there exist* $a \in (0,1)$ *and* $b \geq 0$ *such that*

$$\tfrac{1}{2}\langle x, \nabla V(x)\rangle \;\geq\; aV(x) + \tfrac{\gamma^2 a(2-a)}{8(1-a)}\,\|x\|^2 - b$$

for all $x \in \mathbb{R}^d$.

Assumption F.2 (Target concentration and extensivity)**.** *There exist* $\bar{x} \in \mathbb{R}^d$ *and constants* $R \geq \theta > 0$, $\kappa, \Lambda > 0$*, all independent of* d*, such that*

$$\pi_{\mathrm{tar}}\big(B(\bar{x}, R\sqrt{d})\big) \;\geq\; 1 - e^{-\kappa d}, \qquad \sup\big\{V(x) :\; \|x - \bar{x}\| \leq (R + 3\theta)\sqrt{d}\big\} \;\leq\; \Lambda d.$$

Under Assumptions F.1 and F.2, we place an extensive but curvature-controlled well just outside the target's concentration region. We first establish its geometry and regularity, then show that it is invisible to target-weighted errors but captures nearly all learned equilibrium mass, and finally verify ergodicity and the arbitrary-initialization limit.

Proof of Theorem 4.3. Recall the normalization $\inf V = 0$, and write $Z := \int_{\mathbb{R}^d} e^{-\beta V}$, so that $\pi_{\mathrm{tar}} = Z^{-1} e^{-\beta V}$. Let ω_d denote the volume of the unit ball in $\mathbb{R}^d$. We assume throughout that d is large enough that $e^{-\kappa d} \leq \frac{1}{2}$.

Construction: Let $\bar{x}, R, \theta, \kappa, \Lambda$ be as in Assumption F.2. Choose any $x_\star \in \mathbb{R}^d$ with $\|x_\star - \bar{x}\| = (R + 2\theta)\sqrt{d}$ and set

$$B_{\mathrm{in}} := B\big(x_\star, \theta\sqrt{d}\big), \qquad B_{\mathrm{out}} := B\big(x_\star, 2\theta\sqrt{d}\big), \qquad \chi := \psi_{\theta\sqrt{d},\, 2\theta\sqrt{d}}(\cdot - x_\star), \tag{F.1}$$

with ψ the radial cutoff of Lemma B.3, so that $\chi \equiv 1$ on B_{in} and $\chi \equiv 0$ off B_{out}. Define the well depth and the learned potential by

$$\Delta := \frac{(\kappa_1 + \kappa_2)d + \log 2}{\beta}, \qquad \hat{V} := V - \Delta\chi, \qquad \hat{F} := -\nabla\hat{V}, \tag{F.2}$$

where

$$\kappa_1 := \log(R/\theta) + \beta\Lambda \geq 0 \tag{F.3}$$

is a constant independent of d (nonnegative because $\theta \leq R$). Then $\hat{V} \in C^\infty(\mathbb{R}^d)$, the field $\hat{F}$ is conservative, $\hat{V} = V$ off B_{out}, and $\|\hat{V} - V\|_\infty = \Delta = O(d/\beta)$.

Geometry of the pocket: If $\|x - x_\star\| \leq 2\theta\sqrt{d}$, then $\|x - \bar{x}\| \geq \|x_\star - \bar{x}\| - 2\theta\sqrt{d} = R\sqrt{d}$, so

$$B_{\mathrm{out}} \subseteq \big\{x : \|x - \bar{x}\| \geq R\sqrt{d}\big\} \qquad \Longrightarrow \qquad \pi_{\mathrm{tar}}(B_{\mathrm{out}}) \leq e^{-\kappa d} \tag{F.4}$$

by Assumption F.2. Conversely, every $x \in B_{\mathrm{in}}$ satisfies $\|x - \bar{x}\| \leq (R + 3\theta)\sqrt{d}$, so $V \leq \Lambda d$ on B_{in}, again by Assumption F.2.

Regularity and well-posedness: By Lemma B.3 applied with $(a, b) = (\theta\sqrt{d}, 2\theta\sqrt{d})$, we have

$$\|\nabla\chi\|_\infty \le \frac{C_0}{\theta\sqrt{d}} \qquad \text{and} \qquad \|\nabla^2\chi\|_{\mathrm{op},\infty} \le \frac{C_0}{\theta^2 d} + \frac{C_0}{\theta^2 d} = \frac{2C_0}{\theta^2 d}.$$

Hence, $\nabla^2\hat{V} = \nabla^2 V - \Delta\,\nabla^2\chi$ obeys

$$\begin{aligned}\sup_x \|\nabla^2\hat{V}(x)\|_{\mathrm{op}} &\le L + \Delta \cdot \frac{2C_0}{\theta^2 d} \\ &\le L + \underbrace{\frac{2C_0(\kappa_1+\kappa_2+1)}{\beta\theta^2}}_{=:\,H'},\end{aligned} \tag{F.5}$$

which is independent of d. In particular, $\hat{F}$ is globally Lipschitz, so (2) driven by $\hat{F}$ has a unique strong solution (Øksendal, 2003, Thm. 5.2.1). This is the step at which the width $\Theta(\sqrt{d})$ of the pocket is essential: it converts the extensive depth $\Delta = \Theta(d/\beta)$ into an intensive curvature.

Population errors (Part (i)): Both $\hat{V} - V = -\Delta\chi$ and $\hat{F} - F = \Delta\,\nabla\chi$ vanish off B_{out}, and on B_{out} are bounded in magnitude by Δ and $C_0\Delta/(\theta\sqrt{d})$, respectively. Combined with (F.4), we obtain

$$\begin{aligned}\mathbb{E}_{x\sim\pi_{\mathrm{tar}}}\big[|\hat{V}(x) - V(x)|^p\big]^{1/p} &\le \Delta\,\pi_{\mathrm{tar}}(B_{\mathrm{out}})^{1/p} \le O(d)\,e^{-\kappa d/p} = e^{-\Omega(d)}, \\ \mathbb{E}_{x\sim\pi_{\mathrm{tar}}}\big[\|\hat{F}(x) - F(x)\|^p\big]^{1/p} &\le \frac{C_0\Delta}{\theta\sqrt{d}}\,\pi_{\mathrm{tar}}(B_{\mathrm{out}})^{1/p} \le O(\sqrt{d})\,e^{-\kappa d/p} = e^{-\Omega(d)},\end{aligned}$$

for all $d \ge d_0(p)$, as the polynomial prefactors are absorbed once d is large relative to p.

Since $\pi_{\mathrm{tar}}\big(B(\bar{x}, R\sqrt{d})\big) \ge \frac{1}{2}$ and $V \ge 0$, we can bound Z from above:

$$Z \le 2\int_{B(\bar{x},R\sqrt{d})} e^{-\beta V} \le 2\,\mathrm{vol}\big(B(\bar{x}, R\sqrt{d})\big) = 2\,\omega_d(R\sqrt{d})^d.$$

Meanwhile, $V \le \Lambda d$ on B_{in} gives $\int_{B_{\mathrm{in}}} e^{-\beta V} \ge \omega_d(\theta\sqrt{d})^d e^{-\beta\Lambda d}$. Therefore,

$$\pi_{\mathrm{tar}}(B_{\mathrm{in}}) = \frac{1}{Z}\int_{B_{\mathrm{in}}} e^{-\beta V} \ge \frac{1}{2}\left(\frac{\theta}{R}\right)^d e^{-\beta\Lambda d} = \frac{1}{2}e^{-\kappa_1 d}, \tag{F.6}$$

with κ_1 as defined in (F.3).

Stationary mismatch (Part (ii)): Write $\widehat{Z} := \int e^{-\beta\hat{V}}$. Since $\hat{V} = V - \Delta$ on B_{in},

$$\widehat{Z} \ge \int_{B_{\mathrm{in}}} e^{-\beta\hat{V}} = e^{\beta\Delta}\int_{B_{\mathrm{in}}} e^{-\beta V} = e^{\beta\Delta} Z\,\pi_{\mathrm{tar}}(B_{\mathrm{in}}).$$

Since $\hat{V} = V$ off B_{out} and $B_{\mathrm{out}}^c \subseteq B_{\mathrm{in}}^c$, we have

$$\begin{aligned}\hat{\pi}(B_{\mathrm{out}}^c) &= \frac{1}{\widehat{Z}}\int_{B_{\mathrm{out}}^c} e^{-\beta V} \\ &= \frac{Z\,\pi_{\mathrm{tar}}(B_{\mathrm{out}}^c)}{\widehat{Z}} \\ &\le \frac{1}{e^{\beta\Delta}\,\pi_{\mathrm{tar}}(B_{\mathrm{in}})} \\ &\le \frac{2e^{\kappa_1 d}}{e^{(\kappa_1+\kappa_2)d+\log 2}} \\ &= e^{-\kappa_2 d},\end{aligned}$$

by (F.6) and the choice (F.2) of Δ. Combining this with $\pi_{\mathrm{tar}}(B_{\mathrm{out}}^c) \ge 1 - e^{-\kappa d}$ from (F.4), we obtain

$$\mathrm{TV}(\hat{\pi}, \pi_{\mathrm{tar}}) \ge \pi_{\mathrm{tar}}(B_{\mathrm{out}}^c) - \hat{\pi}(B_{\mathrm{out}}^c) \ge 1 - e^{-\kappa d} - e^{-\kappa_2 d} = 1 - e^{-\Omega(d)}. \tag{F.7}$$

It remains to identify $\hat{\pi}$ as the position marginal of the unique stationary law. The diffusion matrix of (2) is degenerate, acting only on velocities, so ellipticity is unavailable; the drift and diffusion fields do, however, satisfy Hörmander's bracket condition, so the generator is hypoelliptic (Hörmander, 1967). It therefore suffices to verify the Lyapunov condition of Assumption F.1 for $\hat{V}$.

First, $\hat{V} \geq V - \Delta \geq -\Delta$ is bounded below, and $e^{-\beta\hat{V}} \leq e^{\beta\Delta}e^{-\beta V}$ is integrable. Second, since $\chi \geq 0$, we have $\hat{V} \leq V$, hence $aV \geq a\hat{V}$, and therefore

$$\begin{aligned}\frac{1}{2}\langle x, \nabla\hat{V}(x)\rangle &= \frac{1}{2}\langle x, \nabla V(x)\rangle - \frac{\Delta}{2}\langle x, \nabla\chi(x)\rangle \\ &\geq a\hat{V}(x) + \frac{\gamma^2 a(2-a)}{8(1-a)}\|x\|^2 - b - \frac{\Delta M}{2},\end{aligned}$$

where $M := \sup_{x\in B_{\text{out}}} \|x\|\|\nabla\chi\|_\infty < \infty$ because χ has compact support. Thus, Assumption F.1 holds for $\hat{V}$ with the same parameter a and with b replaced by $b + \frac{1}{2}\Delta M + a\Delta$ (the last term absorbing the shift needed to renormalize $\inf \hat{V}$ to zero).

By Mattingly et al. (2002, Thm. 3.2), the process is geometrically ergodic with unique invariant measure

$$\hat{\pi}(\mathrm{d}x\,\mathrm{d}v) \propto e^{-\beta(\hat{V}(x)+\frac{1}{2}\|v\|^2)}\,\mathrm{d}x\,\mathrm{d}v,$$

whose position marginal is $\hat{\pi} \propto e^{-\beta\hat{V}}$. Moreover, there exist constants $C, \lambda > 0$ and a Lyapunov function $\mathcal{L}(x, v) \asymp \hat{V}(x) + \|x\|^2 + \|v\|^2$ such that

$$\mathrm{TV}\big(\mathrm{Law}(x_t, v_t), \hat{\pi}\big) \leq C\big(1 + \mathbb{E}\,\mathcal{L}(x_0, v_0)\big)e^{-\lambda t}. \tag{F.8}$$

Arbitrary initialization (Part (iii)): Fix any initial law with $\mathbb{E}[V(x_0) + \|x_0\|^2 + \|v_0\|^2] < \infty$. Since $|\hat{V} - V| \leq \Delta$, this is equivalent to $\mathbb{E}\,\mathcal{L}(x_0, v_0) < \infty$. Total variation between position marginals is dominated by total variation on phase space, so (F.8) gives $\mathrm{TV}(\hat{\pi}_t, \hat{\pi}) \to 0$. By the triangle inequality,

$$|\mathrm{TV}(\hat{\pi}_t, \pi_{\text{tar}}) - \mathrm{TV}(\hat{\pi}, \pi_{\text{tar}})| \leq \mathrm{TV}(\hat{\pi}_t, \hat{\pi}),$$

Hence by (F.7),

$$\lim_{t\to\infty} \mathrm{TV}(\hat{\pi}_t, \pi_{\text{tar}}) = \mathrm{TV}(\hat{\pi}, \pi_{\text{tar}}) \geq 1 - e^{-\Omega(d)}$$

□

Remark F.3. The rate λ in (F.8) is typically exponentially small in d for $\hat{V}$, since $\hat{\pi}$ is bimodal with an exponentially deep well. Part (iii) is accordingly a purely asymptotic statement, and it is complementary to, rather than implied by, Theorems 4.1 and 4.2: those bound finite horizons for potentials whose equilibrium is correct, whereas this one bounds the equilibrium itself.

Remark F.4. The failures exhibited by the construction are also reflected in physical observables beyond total variation which can be a stringent metric in high dimensions. Theorems 4.1 and 4.2 produce trapping and exponentially slow mixing, while Theorem 4.3 yields $W_2(\hat{\pi}, \pi_{\text{tar}}) = \Omega(\sqrt{d})$, corresponding to an $\Omega(1)$ per-atom RMSD and hence a structurally distinct equilibrium state.

G Proof of the sampling-guarantee theorem

The first two parts of the proof compare Gibbs densities through their normalization constants. The third transfers a target log-Sobolev inequality through a bounded perturbation and then invokes hypocoercivity. We finally return to the two kinetic constructions to show that one-sided equilibrium control does not imply rapid mixing.

Proof of Theorem 4.6. Part (i): Uniform two-sided control. Write the normalization constants as

$$Z = \int e^{-\beta V} \qquad \text{and} \qquad \widehat{Z} = \int e^{-\beta\hat{V}}.$$

From the uniform bound $|\hat{V} - V| \le \delta$, we immediately obtain $e^{-\beta\delta} Z \le \widehat{Z} \le e^{\beta\delta} Z$. Thus, the density ratio satisfies

$$\frac{\mathrm{d}\hat{\pi}}{\mathrm{d}\pi_{\mathrm{tar}}}(x) = e^{-\beta(\hat{V}(x)-V(x))} \frac{Z}{\widehat{Z}} \in \left[e^{-2\beta\delta},\, e^{2\beta\delta}\right].$$

Consequently, we can bound the total variation distance:

$$\mathrm{TV}(\hat{\pi}, \pi_{\mathrm{tar}}) = \int (\pi_{\mathrm{tar}} - \hat{\pi})_+ \le \left(1 - e^{-2\beta\delta}\right) \int \pi_{\mathrm{tar}} = 1 - e^{-2\beta\delta} \le 2\beta\delta.$$

Part (ii): One-sided global control. By condition (a), we have $\hat{V} - V \ge -\delta$, so $e^{-\beta(\hat{V}-V)} \le e^{\beta\delta}$. By condition (b),

$$\widehat{Z} \ge \int_S e^{-\beta\hat{V}} \ge e^{-\beta\delta'} \int_S e^{-\beta V} = e^{-\beta\delta'} Z\, \pi_{\mathrm{tar}}(S) \ge e^{-\beta\delta'} Z(1-\eta)$$

From this, we deduce

$$\frac{Z}{\widehat{Z}} \le \frac{e^{\beta\delta'}}{1-\eta} \qquad \text{and} \qquad \frac{\mathrm{d}\hat{\pi}}{\mathrm{d}\pi_{\mathrm{tar}}} \le \frac{e^{\beta(\delta+\delta')}}{1-\eta},$$

which proves the first claim.

For the total variation bound, note that condition (a) also implies $\widehat{Z} \le e^{\beta\delta} Z$, hence $Z/\widehat{Z} \ge e^{-\beta\delta}$. Meanwhile, condition (b) gives $e^{-\beta(\hat{V}-V)} \ge e^{-\beta\delta'}$ on the set S. Together, these imply $\mathrm{d}\hat{\pi}/\mathrm{d}\pi_{\mathrm{tar}} \ge e^{-\beta(\delta+\delta')}$ on S. Splitting the positive part of the total variation distance over S and S^c, we find:

$$\begin{aligned}
\mathrm{TV}(\hat{\pi}, \pi_{\mathrm{tar}}) &= \int_S (\pi_{\mathrm{tar}} - \hat{\pi})_+ + \int_{S^c} (\pi_{\mathrm{tar}} - \hat{\pi})_+ \\
&\le \left(1 - e^{-\beta(\delta+\delta')}\right) \pi_{\mathrm{tar}}(S) + \pi_{\mathrm{tar}}(S^c) \\
&\le 1 - e^{-\beta(\delta+\delta')} + \eta
\end{aligned}$$

Part (iii): Functional inequalities and relaxation. Under condition (i), we may write the density as

$$\mathrm{d}\hat{\pi} = \mathcal{Z}^{-1} e^{-W}\, \mathrm{d}\pi_{\mathrm{tar}}, \qquad \text{where} \qquad W = \beta(\hat{V} - V),$$

so that its oscillation is bounded by

$$\mathrm{osc}(W) := \sup W - \inf W \le 2\beta\delta.$$

The Holley–Stroock perturbation principle (Holley & Stroock, 1987) therefore implies that $\hat{\pi}$ satisfies a log-Sobolev inequality with a constant of at least

$$\rho e^{-\mathrm{osc}(W)} \ge \rho e^{-2\beta\delta}.$$

Since $\nabla^2 \hat{V}$ is bounded, the hypocoercivity theory of Villani (2009) then yields exponential convergence of the kinetic Langevin dynamics driven by $\hat{F} = -\nabla\hat{V}$ to its invariant phase-space Gibbs measure. In particular, the convergence rate can be bounded in terms of the log-Sobolev constant of $\hat{\pi}$, the Hessian bound H, the friction parameter γ, and the inverse temperature β. Using the bound above, this gives an exponential relaxation rate depending only on $\rho e^{-2\beta\delta}$, H, γ, and β. Thus, condition (i) guarantees exponential relaxation of the learned dynamics in addition to controlling its stationary distribution. The same conclusion does not follow from condition (ii) alone, as shown below. □

Proposition G.1. *The potentials of Theorems 4.1 and 4.2 both satisfy* $\mathrm{TV}(\hat{\pi}, \pi_{\mathrm{tar}}) \le e^{-\Omega(d)}$*; that of Theorem 4.2 additionally satisfies* $\hat{V} \ge V$ *everywhere, so it meets the hypotheses of Theorem 4.6(ii) with* $\delta = \delta' = 0$, $\eta = e^{-\Omega(d)}$. *Both have mixing time* $e^{\Omega(d)}$.

Proof. Data-interpolation construction: Fix i and set $q := U_i - V$. This is a quadratic function with Hessian $(\alpha-1)\mathrm{Id} \succ 0$, and is therefore strictly convex. Its unique minimizer solves $\alpha(x - w_i) = x - \mu$, which rearranges to

$$(\alpha - 1)x = \alpha w_i - \mu.$$

By (E.2), we have $w_i = \mu + (1 - 1/\alpha)z_i$, so we can compute

$$\begin{aligned}\alpha w_i - \mu &= \alpha\mu + (\alpha - 1)z_i - \mu \\ &= (\alpha - 1)(\mu + z_i) \\ &= (\alpha - 1)x_i.\end{aligned}$$

Thus, the minimizer is exactly x_i, where $q(x_i) = \hat{V}(x_i) - V(x_i) = 0$ by the exact-fit computation in Appendix E. Hence, $U_i \geq V$ everywhere. Since $\chi_i \geq 0$ and the supports are disjoint, it follows that $\hat{V} \geq V$ on all of $\mathbb{R}^d$.

Theorem 4.6(ii) therefore applies with $\delta = 0$, $S = A$, $\delta' = 0$, and $\eta = \pi_{\mathrm{tar}}(A^c) \leq e^{-\Omega(d)}$ from (E.7). This yields the bound

$$\mathrm{TV}(\hat{\pi}, \pi_{\mathrm{tar}}) \leq \eta = e^{-\Omega(d)}$$

Kinetic-trapping construction: Here, the condition $\hat{V} \geq V$ fails near the origin, so we argue directly. Split $\mathbb{R}^d$ into the core $\{\|x\| \leq 4\sqrt{d}\}$, the annulus $\mathcal{A} := \{4\sqrt{d} \leq \|x\| \leq b_\xi\sqrt{d}\}$, and the exterior $\{\|x\| \geq b_\xi\sqrt{d}\}$.

On $\mathcal{A}$, by the choice (D.2) of α and the fact that $\|x - \mu\| \leq (b_\xi + \xi)\sqrt{d}$, we have

$$U(x) \geq 8\alpha d \geq \frac{1}{2}(b_\xi + \xi)^2 d \geq V(x) \tag{G.1}$$

Since $\chi \in [0, 1]$ and $\hat{V} = V + \chi(U - V)$, this implies $\hat{V} \geq V$ on $\mathcal{A}$. Thus, by (D.5), we can bound the partition function integral over the annulus:

$$\int_{\mathcal{A}} e^{-\beta\hat{V}} \leq Z\,\pi_{\mathrm{tar}}(\mathcal{A}) \leq Ze^{-\Omega(d)}. \tag{G.2}$$

On the core $\{\|x\| \leq 4\sqrt{d}\}$, we have $\hat{V} = U = \frac{\alpha}{2}\|x\|^2$, so

$$\begin{aligned}\int_{\|x\|\leq 4\sqrt{d}} e^{-\beta\hat{V}} &\leq \int_{\mathbb{R}^d} e^{-\beta\alpha\|x\|^2/2} \\ &= \left(\frac{2\pi}{\beta\alpha}\right)^{d/2} \\ &= \alpha^{-d/2}\left(\frac{2\pi}{\beta}\right)^{d/2} \\ &= \alpha^{-d/2}Z,\end{aligned}$$

which is $Ze^{-\Omega(d)}$ because $\alpha \geq 40$.

Finally, $\hat{V} = V$ on the exterior $\{\|x\| \geq b_\xi\sqrt{d}\}$, which is a set of π_{tar}-mass $1 - e^{-\Omega(d)}$ by (D.5).

Therefore, $\widehat{Z} = Z(1 \pm e^{-\Omega(d)})$. The two densities are proportional on a set of π_{tar}-mass $1 - e^{-\Omega(d)}$ with a ratio of $1 \pm e^{-\Omega(d)}$, giving

$$\mathrm{TV}(\hat{\pi}, \pi_{\mathrm{tar}}) \leq e^{-\Omega(d)}.$$

Mixing-time lower bounds: In both cases, Corollary D.1 and Theorem 4.2(iii) exhibit initial conditions from which the law remains at a total variation distance of $1 - e^{-\Omega(d)}$ from π_{tar}. Therefore, by the two bounds just proved and the triangle inequality, the law also remains at a distance of $1 - e^{-\Omega(d)}$ from $\hat{\pi}$ for every $T \leq e^{cd}$. The mixing time is thus $e^{\Omega(d)}$. □

H SYMMETRY- AND LOCALITY-AWARE HYPOTHESIS CLASSES

We consider one atomic species in a periodic cell Λ_N at fixed density. A configuration is $x = (x_1, \ldots, x_N)$. To remove the irrelevant non-normalizability associated with a freely translating center

of mass, probability measures may be understood either on the periodic configuration space modulo diagonal translations or in a fixed center-of-mass gauge. This changes the dimension by only three, so $d = 3N + O(1)$ and exponential statements in N and d are equivalent. Relative vectors use the periodic convention. The local rules below are invariant under orthogonal transformations of all relative vectors, equivariant under relabeling of identical atoms, and translation invariant. Equivalently, when the cell is transformed together with the atoms, they have the usual $E(3)$ invariance of an interatomic potential.

H.1 The local invariant hypothesis class

Choose a C^∞ cutoff kernel that vanishes, together with all of its derivatives, at the cutoff radius r_c. Let $\mathcal{E}_i^{(L)}(x)$ denote a smooth representation of the information reaching site i after L local aggregation steps. We require the family $(\mathcal{E}_i^{(L)})_{i=1}^N$ to be permutation covariant, translation invariant, and orthogonally covariant. Smooth cutoff weights, rather than a hard neighbor-list indicator, are part of the environment map. This convention is essential since a graph defined by the discontinuous rule $\|x_i - x_j\| \le r_c$ does not by itself define a globally smooth hypothesis class.

Definition H.1 (Local invariant class)**.** Let $\mathcal{I}$ be the set of smooth site functions that are invariant under orthogonal transformations and permutations of the neighbors in a local environment. The depth-L local invariant class is

$$\mathcal{H}_{\mathrm{loc}}(r_c, L) := \left\{ \hat{V}(x) = \sum_{i=1}^{N} \varepsilon\Big(\mathcal{E}_i^{(L)}(x)\Big) \,:\, \varepsilon \in \mathcal{I} \right\}. \tag{H.1}$$

The same site function ε is used at every site and for every system size. We assume a bounded-receptive-field regime: one aggregation step uses at most z atomic coordinate blocks and an L-step environment uses at most m blocks, where z and m do not grow with N.

On a bounded-degree interaction graph, $m \le 1+z+\cdots+z^L = O(z^L)$. A uniform bound of this kind is essential since fixed density alone does not prevent arbitrarily many atoms from entering a cutoff ball. It can be justified, for example, on an admissible configuration space with a hard-core separation condition. Every member of $\mathcal{H}_{\mathrm{loc}}$ is smooth and conservative, its force is equivariant, and its energy is size extensive and transferable across system sizes. The class includes the structural form used by local invariant and equivariant MLIPs. We assume that the target is realizable, $V \in \mathcal{H}_{\mathrm{loc}}(r_c, L)$.

H.2 Invariant local descriptors and concentration

The constructions in Appendices D to F gate a perturbation with a global distance. Here we replace that gate by a smooth scalar descriptor at each site. A typical example starts from a smooth one-hop scalar

$$q_j(x) = \sum_{k \ne j} f_c(x_k - x_j),$$

where f_c is a smooth, periodic, radial cutoff kernel, and averages these quantities with smooth L-hop aggregation weights.

Definition H.2 (Averaged local descriptor)**.** Let $w_{ij}^{(L)}(x) \ge 0$ be smooth, permutation-covariant and rigid-motion invariant weights, supported only when information can pass from j to i in at most L aggregation steps. Assume their normalizing sums are positive. Set

$$s_i(x) := \frac{\sum_{j=1}^N w_{ij}^{(L)}(x) q_j(x)}{\sum_{j=1}^N w_{ij}^{(L)}(x)}. \tag{H.2}$$

The family $(s_i)_{i=1}^N$ is permutation covariant, while each s_i is a rigid-motion-invariant scalar. Consequently, $\sum_i \varphi(s_i)$ is a valid invariant site-energy sum for every smooth scalar function φ.

The averaging formula motivates the derivative scaling used below, but does not imply it without assumptions on the aggregation weights and correlations. We therefore state the analytic properties needed by the construction explicitly.

Lemma H.3 (Receptive-field concentration)**.** *Suppose that each s_i depends on a coordinate-block set S_i contained in a $2L$-hop receptive field. Assume that there are constants $C_z, C_z^{(2)}, C_{\mathrm{ov}} > 0$, independent of N, such that, pointwise,*

$$\|\nabla s_i\| \le C_z\sqrt{\frac{z}{m}}, \qquad \|\nabla^2 s_i\|_{\mathrm{op}} \le C_z^{(2)}\frac{z}{m}, \tag{H.3}$$

and each atomic coordinate block belongs to at most

$$\kappa_{\mathrm{ov}} \le C_{\mathrm{ov}} m \tag{H.4}$$

of the sets S_i.

Assume also that the target measure satisfies the following dimension-free Lipschitz concentration inequality: for every smooth K-Lipschitz scalar function g and every $t > 0$,

$$\mathbb{P}_{x\sim\pi_{\mathrm{tar}}}\big(|g(x) - \mathbb{E}_{\pi_{\mathrm{tar}}} g| \ge t\big) \le 2\exp\left(-\frac{\alpha t^2}{2K^2}\right)$$

for a constant $\alpha > 0$ independent of N. Then, for every i,

$$\mathbb{P}_{x\sim\pi_{\mathrm{tar}}}\left(|s_i(x) - \mathbb{E}_{\pi_{\mathrm{tar}}} s_i| \ge t\right) \le 2\exp\left(-\frac{\alpha m t^2}{2C_z^2 z}\right). \tag{H.5}$$

Moreover, if

$$n(x) := \sum_{i=1}^{N} \varphi(s_i(x))$$

and $\varphi : \mathbb{R} \to [0,1]$ is Lipschitz, then

$$\mathbb{P}_{x\sim\pi_{\mathrm{tar}}}\left(n(x) - \mathbb{E}_{\pi_{\mathrm{tar}}} n \ge tN\right) \le 2\exp\left(-\frac{\alpha t^2 N}{2C_{\mathrm{ov}}C_z^2 z\|\varphi'\|_\infty^2}\right). \tag{H.6}$$

Proof. The first bound in (H.3) makes s_i $C_z\sqrt{z/m}$-Lipschitz, so (H.5) follows directly from the assumed concentration inequality.

For the second claim, set $u_i := \varphi'(s_i)\nabla s_i$, which is supported on S_i. The overlap bound gives

$$\left\|\sum_i u_i\right\|^2 \le \kappa_{\mathrm{ov}} \sum_i \|u_i\|^2.$$

Hence

$$\begin{aligned}\|\nabla n\|^2 &\le \kappa_{\mathrm{ov}}\|\varphi'\|_\infty^2 \sum_{i=1}^{N}\|\nabla s_i\|^2 \\ &\le C_{\mathrm{ov}}C_z^2 z N\|\varphi'\|_\infty^2.\end{aligned}$$

Thus n is $\sqrt{C_{\mathrm{ov}}C_z^2 zN}\|\varphi'\|_\infty$-Lipschitz. Applying the concentration inequality at deviation tN proves (H.6). □

For a descriptor separation $\Delta s > 0$, define

$$\nu := \frac{\alpha m(\Delta s)^2}{2C_z^2 z}. \tag{H.7}$$

When $m \asymp z^L$ and the remaining constants are fixed, $\nu \asymp z^{L-1}$.

Assumption H.4. *There are constants $a_1 < a_2$, $\Delta s > 0$, and $\kappa > 0$, independent of N, and a sequence of invariant Borel sets $\mathcal{D}_N$ such that*

(a) $\mathrm{dist}([a_1, a_2], \mathbb{E}_{\pi_{\mathrm{tar}}} s_i) \ge \Delta s$ *for every i;*

(b) $s_i(x) \in [a_1, a_2]$ *for every $x \in \mathcal{D}_N$ and every i;*

(c) $\pi_{\mathrm{tar}}(\mathcal{D}_N) \geq e^{-\kappa N}$.

Condition (c) says that the restricted free energy of the decoy set is at most $\kappa N/\beta$ above the target free energy. It prevents the decoy from being superexponentially rare. Conditions (a)-(b), together with the concentration lemma, make the same set exponentially atypical for the target once a positive fraction of its sites must lie in the descriptor window. Competing polymorphs, uniformly strained cells, and defect-rich structures are possible physical realizations, but the theorem uses only the three stated properties.

H.3 Failure in the local invariant class

Theorem H.5. *Let $V \in \mathcal{H}_{\mathrm{loc}}(r_c, L)$ and let $\pi_{\mathrm{tar}} \propto e^{-\beta V}$ be a probability measure. Suppose that the descriptor family satisfies Lemma H.3, Assumption H.4 holds, and*

$$\sup_x \|\nabla^2 V(x)\|_{\mathrm{op}} \leq H_0 \tag{H.8}$$

for a constant H_0 independent of N.

Let $\varphi \in C^\infty(\mathbb{R}; [0,1])$ satisfy $\varphi \equiv 1$ on $[a_1, a_2]$ and $\varphi \equiv 0$ outside $[a_1 - h, a_2 + h]$, where $0 < h < \Delta s/2$, with $\|\varphi^{(k)}\|_\infty \leq C_k h^{-k}$. Define

$$\nu_h := \frac{\alpha m(\Delta s - h)^2}{2C_z^2 z} = \left(1 - \frac{h}{\Delta s}\right)^2 \nu. \tag{H.9}$$

Choose a constant Δ_1 such that

$$\beta\Delta_1 > 2\kappa, \qquad \nu_h \geq \log\left(\frac{4\beta\Delta_1}{\kappa}\right), \tag{H.10}$$

and set

$$\hat{V}(x) := V(x) - \Delta_1 \sum_{i=1}^{N} \varphi(s_i(x)), \qquad \hat{F} := -\nabla\hat{V}. \tag{H.11}$$

Then $\hat{V} \in \mathcal{H}_{\mathrm{loc}}(r_c, 2L)$, $\hat{V}$ is invariant, $\hat{F}$ is equivariant, and

$$\sup_x \|\nabla^2 \hat{V}(x)\|_{\mathrm{op}} \leq H \tag{H.12}$$

for a constant H independent of N. Moreover:

(i) *For every fixed $p \geq 1$,*

$$\frac{1}{N}\|\hat{V} - V\|_{L^p(\pi_{\mathrm{tar}})} \leq \Delta_1 (2e^{-\nu_h})^{1/p}, \tag{H.13}$$

and

$$\left(\frac{1}{3N}\mathbb{E}_{\pi_{\mathrm{tar}}}\|\hat{F} - F\|^2\right)^{1/2} \leq \frac{C\Delta_1 C_z}{h}\sqrt{\frac{\kappa_{\mathrm{ov}} z}{m}}\, e^{-\nu_h/2}. \tag{H.14}$$

In particular, since $\kappa_{\mathrm{ov}} \leq C_{\mathrm{ov}} m$,

$$\left(\frac{1}{3N}\mathbb{E}_{\pi_{\mathrm{tar}}}\|\hat{F} - F\|^2\right)^{1/2} \leq \frac{C\Delta_1 C_z\sqrt{C_{\mathrm{ov}} z}}{h} e^{-\nu_h/2}. \tag{H.15}$$

The displayed per-atom bounds do not grow with N.

(ii) *Write $\widehat{\pi} := \widehat{Z}^{-1} e^{-\beta\hat{V}}$. This measure is well defined because $-\Delta_1 N \leq \hat{V} - V \leq 0$ and the target is normalizable. There are constants $c_1, c_2 > 0$, independent of N, such that*

$$\mathrm{TV}(\widehat{\pi}, \pi_{\mathrm{tar}}) \geq 1 - e^{-c_1 N} - 2e^{-c_2 N} = 1 - e^{-\Omega(N)} = 1 - e^{-\Omega(d)}. \tag{H.16}$$

Normalizability is automatic on a periodic configuration space. If the kinetic Langevin process driven by $\hat{F}$ is ergodic, then its position marginal converges to $\widehat{\pi}$, so the same discrepancy is its long-time sampling error. On a connected compact position space, smoothness is sufficient for the usual hypoelliptic ergodicity conclusion. On $\mathbb{R}^d$, one may instead impose the confining hypothesis of Assumption F.1 on $\hat{V}$.

(iii) $\hat{V} \le V$ *everywhere, and*

$$\inf_x (\hat{V}(x) - V(x)) = -\Delta_1 N. \tag{H.17}$$

Consequently, the global lower bound $\hat{V} \ge V - \delta$ *in Theorem 4.6 can hold only if* $\delta \ge \Delta_1 N$. *Since (H.10) implies* $\Delta_1 = \Omega(\beta^{-1})$, *it fails for every* $\delta = o(N/\beta)$.

Proof. Because s_i depends on at most a $2L$-hop environment, $\varphi(s_i)$ is an admissible depth-$2L$ site energy. The covariance of the family $(s_i)_i$ implies that its sum is invariant, which proves the class and symmetry claims.

For the Hessian, write

$$\nabla^2 \hat{V} = \nabla^2 V - \Delta_1 \sum_{i=1}^{N} B_i, \qquad B_i := \varphi''(s_i) \nabla s_i \nabla s_i^\top + \varphi'(s_i) \nabla^2 s_i. \tag{H.18}$$

By (H.3),

$$\|B_i\|_{\mathrm{op}} \le C \frac{z}{m} \left(\frac{C_z^2}{h^2} + \frac{C_z^{(2)}}{h} \right).$$

Each B_i is supported on $S_i \times S_i$. Hence, for every vector u,

$$\left| u^\top \left(\sum_i B_i \right) u \right| \le \max_i \|B_i\|_{\mathrm{op}} \sum_i \|P_{S_i} u\|^2 \le \kappa_{\mathrm{ov}} \max_i \|B_i\|_{\mathrm{op}} \|u\|^2.$$

Combining this inequality with (H.4) and (H.8) gives

$$\sup_x \|\nabla^2 \hat{V}(x)\|_{\mathrm{op}} \le H_0 + C \Delta_1 C_{\mathrm{ov}} z \left(\frac{C_z^2}{h^2} + \frac{C_z^{(2)}}{h} \right) =: H,$$

which is independent of N.

Local energy and force errors: Let $I_i := \mathbf{1}\{s_i \in \operatorname{supp} \varphi\}$. Assumption H.4(a) and the support of φ imply

$$I_i = 1 \quad \Longrightarrow \quad |s_i - \mathbb{E}_{\pi_{\mathrm{tar}}} s_i| \ge \Delta s - h.$$

Therefore, by (H.5),

$$\mathbb{P}_{\pi_{\mathrm{tar}}}(I_i = 1) \le 2e^{-\nu_h}. \tag{H.19}$$

Since

$$|\hat{V} - V| \le \Delta_1 \sum_{i=1}^{N} I_i, \tag{H.20}$$

Minkowski's inequality yields

$$\|\hat{V} - V\|_{L^p(\pi_{\mathrm{tar}})} \le \Delta_1 \sum_{i=1}^{N} \|I_i\|_{L^p(\pi_{\mathrm{tar}})} \le \Delta_1 N (2e^{-\nu_h})^{1/p},$$

which proves (H.13).

For the force error, $\hat{F} - F = \Delta_1 \sum_i \varphi'(s_i) \nabla s_i$. The overlap bound gives

$$\|\hat{F} - F\|^2 \le \kappa_{\mathrm{ov}} \Delta_1^2 \|\varphi'\|_\infty^2 \sum_{i=1}^{N} I_i \|\nabla s_i\|^2. \tag{H.21}$$

Taking expectations and using (H.19) and (H.3) gives

$$\mathbb{E}_{\pi_{\mathrm{tar}}} \|\hat{F} - F\|^2 \le C \kappa_{\mathrm{ov}} \Delta_1^2 h^{-2} C_z^2 \frac{z}{m} N e^{-\nu_h}.$$

Dividing by $3N$ and taking square roots proves (H.14). Substituting $\kappa_{\mathrm{ov}} \le C_{\mathrm{ov}} m$ proves (H.15).

Global equilibrium mismatch: Set $n(x) := \sum_i \varphi(s_i(x)) \in [0, N]$. Exponential tilting gives, for every measurable A,

$$\widehat{\pi}(A) = \frac{\mathbb{E}_{\pi_{\mathrm{tar}}}[\mathbf{1}_A e^{\beta\Delta_1 n}]}{\mathbb{E}_{\pi_{\mathrm{tar}}}[e^{\beta\Delta_1 n}]}. \tag{H.22}$$

On $\mathcal{D}_N$, Assumption H.4(b) gives $n = N$, so

$$\mathbb{E}_{\pi_{\mathrm{tar}}}[e^{\beta\Delta_1 n}] \geq e^{(\beta\Delta_1 - \kappa)N}. \tag{H.23}$$

Let $\eta := \kappa/(\beta\Delta_1) < 1/2$ and $\mathcal{T}_\eta := \{n \leq \eta N\}$. From (H.22)–(H.23),

$$\widehat{\pi}(\mathcal{T}_\eta) \leq e^{-(\beta\Delta_1 - 2\kappa)N} = e^{-c_1 N}, \qquad c_1 := \beta\Delta_1 - 2\kappa > 0. \tag{H.24}$$

On the other hand, (H.19) and the second inequality in (H.10) imply

$$\mathbb{E}_{\pi_{\mathrm{tar}}} n \leq 2Ne^{-\nu_h} \leq \frac{\eta N}{2}. \tag{H.25}$$

Applying (H.6) with $t = \eta/2$ gives

$$\pi_{\mathrm{tar}}(\mathcal{T}_\eta) \geq 1 - 2e^{-c_2 N}, \qquad c_2 := \frac{\alpha\eta^2}{8C_{\mathrm{ov}}C_z^2 z\|\varphi'\|_\infty^2} > 0. \tag{H.26}$$

Therefore

$$\mathrm{TV}(\widehat{\pi}, \pi_{\mathrm{tar}}) \geq \pi_{\mathrm{tar}}(\mathcal{T}_\eta) - \widehat{\pi}(\mathcal{T}_\eta) \geq 1 - e^{-c_1 N} - 2e^{-c_2 N},$$

which proves (H.16). The statement about the long-time law is then exactly the definition of ergodic convergence to $\widehat{\pi}$.

Extensive depth of the perturbation: Because $0 \leq \varphi \leq 1$,

$$-\Delta_1 N \leq \hat{V} - V = -\Delta_1 \sum_i \varphi(s_i) \leq 0.$$

Assumption H.4(c) makes $\mathcal{D}_N$ nonempty, and on this set every gate equals one. Equality is therefore attained at the lower endpoint, which proves (H.17) and the remaining claims. □

The construction is invariant under translations, identical-particle permutations, and orthogonal transformations of the local relative vectors. Its force is equivariant under the same actions. Higher body order or angular resolution changes which scalar descriptor families satisfy the hypotheses, and hence changes the constants in (H.7), but does not invalidate the tilting argument once those hypotheses hold.

Remark H.6. For fixed p, the construction in the unconstrained class has errors $e^{-\Omega(d)}$, whereas (H.13) and (H.15) give respectively $e^{-\Omega(\nu_h/p)}$ and $e^{-\Omega(\nu_h)}$. Since ν_h is independent of N for a fixed local architecture, locality can potentially impose an N-independent floor on how accurately this particular decoy can be hidden. The global equilibrium error nevertheless remains $1 - e^{-\Omega(N)}$.

Remark H.7. In the regime $m \asymp z^L$ with all other constants fixed, achieving either displayed per-atom error at most ε requires $\nu_h = \Omega(\log(1/\varepsilon))$, and hence $L = 1 + \Omega(\log\log(1/\varepsilon)/\log z)$.

Remark H.8. Because $\hat{V}$ is specified by one site rule, the same model can be evaluated at different system sizes. Its per-atom bounds in part (i) do not grow with N, whereas the equilibrium discrepancy in part (ii) approaches one exponentially fast in N. Consequently, accurate validation on a small cell does not by itself certify equilibrium sampling at deployment scale.