DOI: 10.5281/zenodo.21358361 | Volume: 1 | Issue: 1 | ISSN: 3141-643X

Ktrend – International Journal of Computational Mathematics and Scientific Computing

Mathematical Modelling of Structure-Preserving Physics-Informed Geometric Neural Stochastic Differential Equations for Molecular Dynamics

Jeremiah U. Atsu[1], **Samuel O. Essang**[2], **Jackson E. Ante**[3,*]

[1]Department of Mathematics, University of Cross River State, Calabar, Nigeria

[2]Department of Physical Science, Landmark University, Omu-Aran, Nigeria

[3]Department of Mathematics, Topfaith University, Mkpatak, Nigeria

Corresponding Author: jackson.ante@topfaith.edu.ng

Abstract

Accurate data-driven models for molecular dynamics must generate stable long-horizon predictions while providing uncertainty estimates that remain consistent with the principles of statistical mechanics. Motivated by these requirements, this study advances the comparative analysis of uncertainty quantification in molecular dynamics through a Physics-Informed Geometric Neural Stochastic Differential Equation (PI-GNSDE) framework. The proposed formulation combines four fundamental components: an underdamped Langevin stochastic differential equation, a neural-network-based potential energy model, geometric equivariance with respect to molecular coordinates, and the fluctuation-dissipation relation, which enforces a thermodynamically consistent coupling between friction and thermal noise. The paper presents the governing equations, explicitly states the assumptions required to guarantee thermodynamic consistency, and establishes a reproducible workflow for data analysis and model validation. To ensure transparency and traceability, benchmark descriptors are compiled from publicly available Chignolin (CLN025) protein-folding datasets, including peptide properties, solvation conditions, thermostat and integrator configurations, the number and duration of simulations, integration time step, simulation box volume, solvent and ion atom count, initial thermal energy, thermostat damping timescale, and total integration-step count. These quantities are reported exclusively as reproducible dataset metadata and should not be interpreted as model training or predictive performance results. Furthermore, the study clearly distinguishes the mathematical formulation, modelling assumptions, numerical integration procedures, and empirical validation protocol. Since the raw molecular trajectory files were not available during the preparation of this work, quantitative performance measures, including Kullback-Leibler divergence, energy drift, prediction coverage, and long-term stability metrics, are intentionally excluded to avoid unsupported conclusions. Nevertheless, the proposed framework provides a rigorous, transparent,

© 2026 The Author(s). CC BY 4.0.

and reproducible foundation for future model training, calibration, uncertainty quantification, and validation once the corresponding trajectory data are acquired and processed using the prescribed pipeline.

Keywords: Molecular dynamics; mathematical modelling; stochastic differential equations; uncertainty quantification; Langevin dynamics; geometric deep learning; neural stochastic differential equations; fluctuation-dissipation theorem; BAOAB integration; Chignolin.

1. Introduction

Molecular dynamics (MD) has become one of the principal computational tools for investigating the dynamical behaviour of molecular systems at the atomic scale. By integrating the governing equations of motion, MD generates trajectories describing the temporal evolution of particle positions and momenta, thereby providing detailed insight into molecular conformations, thermodynamic properties, and kinetic processes. Owing to its broad applicability, molecular dynamics has found extensive use in computational chemistry, structural biology, materials science, and statistical physics, while simultaneously motivating significant developments in mathematical modelling, numerical analysis, stochastic processes, and machine learning.

Classical molecular dynamics is founded upon well-established physical principles. The conservative component of the dynamics is determined by the potential energy surface, whereas interactions with the surrounding thermal environment are commonly represented through stochastic thermostats. In particular, Langevin dynamics models the exchange of energy with a heat bath by introducing dissipative and stochastic forces that satisfy the fluctuation-dissipation theorem, thereby ensuring consistency with equilibrium statistical mechanics. Maintaining the correct balance between deterministic forces, dissipation, and thermal fluctuations is essential for reproducing the Gibbs-Boltzmann equilibrium distribution and obtaining physically meaningful long-term simulations. Consequently, molecular dynamics should be viewed not merely as a prediction problem, but as a structure-preserving modelling problem in which numerical stability and thermodynamic consistency are fundamental requirements.

Recent advances in scientific machine learning have introduced powerful data-driven approaches for learning continuous-time dynamical systems directly from observational data. Among these, Neural Ordinary Differential Equations (Neural ODEs) provide deterministic neural representations of vector fields, while Neural Stochastic Differential Equations (Neural SDEs) extend this framework by incorporating stochastic diffusion processes to model uncertainty and random fluctuations. These methods have demonstrated remarkable flexibility in learning complex nonlinear dynamics; however, unconstrained neural parameterizations generally do not preserve the physical structure underlying molecular systems. In particular, they may fail to respect geometric symmetries, conserve physically meaningful invariants, maintain thermodynamic equilibrium, or produce stable trajectories over long integration horizons.

For molecular dynamics, an effective learning framework should satisfy several fundamental requirements. First, it should remain numerically stable under long-time integration. Second, it must preserve the geometric structure of molecular configurations, including invariance or equivariance with respect to translations and rotations. Third, it should provide reliable and well-calibrated uncertainty quantification through an appropriate stochastic formulation. Finally, the resulting

© 2026 The Author(s). CC BY 4.0.

dynamics should satisfy the fluctuation-dissipation relation to ensure convergence towards the correct Gibbs-Boltzmann equilibrium distribution. Simultaneously satisfying these requirements remains a significant challenge for existing data-driven molecular modelling approaches.

To address these challenges, this paper develops a Physics-Informed Geometric Neural Stochastic Differential Equation (PIG-NSDE) framework for molecular dynamics. The proposed model combines four complementary components: an underdamped Langevin stochastic differential equation, a neural-network approximation of the potential energy surface, a geometrically equivariant representation of molecular coordinates, and the fluctuation-dissipation theorem, which establishes the thermodynamically consistent relationship between friction and stochastic forcing. Rather than treating deterministic and stochastic components independently, the proposed formulation integrates them within a mathematically consistent framework that preserves the essential physical principles governing molecular motion.

The primary objective of this work is to establish a rigorous mathematical modelling framework for uncertainty-aware molecular dynamics that is physically interpretable, thermodynamically consistent, and reproducible. To this end, we derive the governing equations, explicitly state the assumptions required for thermodynamic consistency, and develop a transparent protocol for reproducible data analysis and future empirical validation. Publicly available Chignolin (CLN025) protein-folding metadata are employed to define traceable benchmark descriptors, including simulation parameters, thermostat configurations, integration settings, and derived physical quantities. These metadata are presented solely to ensure transparency and reproducibility and should not be interpreted as measures of model training or predictive performance.

Accordingly, this study deliberately distinguishes between the mathematical formulation of the proposed model and its future empirical evaluation. While the present work establishes the theoretical framework, modelling assumptions, and validation protocol, quantitative performance metrics such as prediction accuracy, Kullback-Leibler divergence, energy drift, uncertainty calibration, and long-term stability require complete molecular trajectory datasets and corresponding computational experiments. By maintaining this distinction, the proposed framework provides a rigorous foundation for subsequent implementation, training, and benchmarking while ensuring that all reported quantities remain reproducible and fully supported by available data.

2. Basic Notations

The notation below is used throughout the paper. The model assumes a finite-dimensional phase space and classical molecular mechanics.

- Symbol: q; Meaning: Position vector; Dimension or unit: $\mathbb{R}^d$.
- Symbol: p; Meaning: Momentum vector; Dimension or unit: $\mathbb{R}^d$.
- Symbol: z; Meaning: Phase-space state, $z = (q, p)$; Dimension or unit: $\mathbb{R}^{2d}$.
- Symbol: M; Meaning: Positive definite mass matrix; Dimension or unit: $d \times d$.
- Symbol: $V(q)$; Meaning: True potential energy; Dimension or unit: Energy.
- Symbol: $V_\theta(q)$; Meaning: Learned neural potential energy; Dimension or unit: Energy.

© 2026 The Author(s). CC BY 4.0.

- Symbol: $H(q, p)$; Meaning: Hamiltonian; Dimension or unit: Energy.
- Symbol: Γ; Meaning: Friction matrix; Dimension or unit: Momentum/time scale.
- Symbol: γ; Meaning: Scalar friction coefficient; Dimension or unit: Time^{-1}.
- Symbol: T; Meaning: Absolute temperature; Dimension or unit: Kelvin.
- Symbol: k_B; Meaning: Boltzmann constant; Dimension or unit: Energy per Kelvin.
- Symbol: R; Meaning: Gas constant; Dimension or unit: kJ mol^{-1} K^{-1}.
- Symbol: W_t; Meaning: Wiener process; Dimension or unit: Stochastic process.
- Symbol: $\pi(q, p)$; Meaning: Stationary phase-space density; Dimension or unit: Probability density.
- Symbol: Δt; Meaning: Integration time step; Dimension or unit: Time.
- Symbol: $\mathcal{L}$; Meaning: Training loss; Dimension or unit: Scalar.
- Symbol: D_{KL}; Meaning: Kullback-Leibler divergence; Dimension or unit: Dimensionless.

3. Mathematical Preliminaries

3.1. Phase-space representation

Let $q_t \in \mathbb{R}^d$ denote the molecular configuration at time t, and let $p_t \in \mathbb{R}^d$ denote the corresponding momentum. The phase-space state is

$$z_t = \begin{bmatrix} q_t \\ p_t \end{bmatrix} \in \mathbb{R}^{2d}. \quad (1)$$

For a classical molecular system with separable kinetic and potential energy, the Hamiltonian is

$$H(q, p) = T(p) + V(q), \qquad T(p) = \tfrac{1}{2} p^\top M^{-1} p. \quad (2)$$

The conservative deterministic Hamiltonian dynamics are

$$\dot{q} = \nabla_p H(q, p) = M^{-1} p, \qquad \dot{p} = -\nabla_q H(q, p) = -\nabla V(q). \quad (3)$$

Equation (3) is the structural core of molecular mechanics. It says that the learned force should not be an arbitrary vector field when the system is conservative. It should be the negative gradient of a scalar potential energy function.

3.2. Canonical ensemble

For a molecular system at temperature T, the canonical Gibbs-Boltzmann distribution is

$$\pi(q, p) = Z^{-1} \exp\left[-\beta H(q, p)\right], \qquad \beta = \tfrac{1}{k_B T}, \quad (4)$$

where Z is the normalising constant

© 2026 The Author(s). CC BY 4.0.

$$Z = \int_{\mathbb{R}^{2d}} \exp\left[-\beta H(q,p)\right] dq\, dp. \quad (5)$$

A learned molecular model intended for equilibrium sampling must either preserve or accurately approximate this distribution. If the model has the wrong stationary distribution, then long-term averages such as conformational populations, free energies, and transition statistics may be biased.

3.3. Underdamped Langevin dynamics

Underdamped Langevin dynamics introduces friction and random thermal forcing into the Hamiltonian system. In matrix form, the Itô SDE is

$$\begin{aligned} dq_t &= M^{-1} p_t\, dt, \\ dp_t &= -\nabla V(q_t)\, dt - \Gamma M^{-1} p_t\, dt + \sqrt{2k_B T\Gamma}\, dW_t. \end{aligned} \quad (6)$$

The square root in (6) denotes any matrix factor B satisfying

$$BB^{\top} = 2k_B T\Gamma. \quad (7)$$

Equation (7) is the fluctuation-dissipation relation in this setting. It is the mathematical condition that links the noise scale to the dissipative friction. If Γ is changed while the noise scale is held fixed, the invariant distribution changes. This is why a generic neural diffusion term can be physically dangerous in equilibrium molecular modelling.

4. Modelling Problem

The modelling problem is to learn a physically meaningful approximation $V_\theta(q)$ of the potential energy surface from molecular trajectory data, while preserving the stochastic thermodynamic structure of (6). The model is not required to learn arbitrary drift and arbitrary diffusion. It learns a scalar energy function. The conservative force is obtained by differentiation.

$$F_\theta(q) = -\nabla_q V_\theta(q). \quad (8)$$

The learned dynamics are

$$\begin{aligned} dq_t &= M^{-1} p_t\, dt, \\ dp_t &= F_\theta(q_t)\, dt - \Gamma M^{-1} p_t\, dt + \sqrt{2k_B T\Gamma}\, dW_t. \end{aligned} \quad (9)$$

The learned Hamiltonian is

$$H_\theta(q,p) = \tfrac{1}{2} p^{\top} M^{-1} p + V_\theta(q). \quad (10)$$

The target stationary density induced by the learned Hamiltonian is

$$\pi_\theta(q,p) = Z_\theta^{-1} \exp\left[-\beta H_\theta(q,p)\right]. \quad (11)$$

Remark 1: Equation (9) admits (11) as its invariant distribution since if the drift and diffusion are related as in (9), with fixed T and Γ, then the stochastic thermostat is consistent with the Gibbs-Boltzmann distribution corresponding to H_θ.

© 2026 The Author(s). CC BY 4.0.

5. Physics-Informed Geometric Neural SDE

5.1. Structure of the model

The proposed model has three layers of structure. The first layer is the stochastic differential equation itself. The second layer is the neural energy model $V_\theta(q)$. The third layer is the integration method used for numerical simulation.

The drift vector is

$$f_\theta(q,p) = \begin{bmatrix} M^{-1}p \\ -\nabla_q V_\theta(q) - \Gamma M^{-1}p \end{bmatrix}, \quad (12)$$

and the diffusion matrix is

$$G = \begin{bmatrix} 0 \\ \sqrt{2k_B T\Gamma} \end{bmatrix}. \quad (13)$$

The phase-space SDE is

$$dz_t = f_\theta(z_t)\,dt + G dW_t. \quad (14)$$

The diffusion is not an unconstrained neural network. This design choice is deliberate. In a physical thermostat, the diffusion is constrained by temperature and friction. Allowing a neural network to learn diffusion independently may increase apparent data fit, but it can destroy thermodynamic interpretability.

5.2. Geometric neural parameterisation

Molecules are embedded in Euclidean space. If a molecule is translated or rotated, its energy should not change. If the coordinates are transformed by a Euclidean motion, the force field should transform equivariantly. Let R be an orthogonal matrix and a be a translation vector. The energy invariance condition is

$$V_\theta(Rq + a) = V_\theta(q). \quad (15)$$

Taking gradients gives the associated force equivariance condition

$$F_\theta(Rq + a) = RF_\theta(q). \quad (16)$$

This is why an E(n)-equivariant graph neural network is an appropriate modelling choice for V_θ. The graph representation uses atoms as nodes and interatomic relations as edges. Scalar node features may include atom type or residue features, while coordinate updates preserve Euclidean transformation rules. The purpose is not simply to improve accuracy. The purpose is to remove physically meaningless degrees of freedom from the learning problem.

© 2026 The Author(s). CC BY 4.0.

5.3. Learning objective

Suppose a dataset contains observed trajectory windows

$$\mathcal{D} = \{z_i\left(t_j\right) \mid i = 1, \ldots, N,\ j = 0, \ldots, m\}. \quad (17)$$

A trajectory-matching loss can be written as

$$\mathcal{L}_{traj}(\theta) = \frac{1}{Nm} \sum_{i=1}^{N} \sum_{j=1}^{m} \| \widehat{z}_i\left(t_j;\theta\right) - z_i\left(t_j\right) \|_2^2. \quad (18)$$

If only configurations are observed, the loss can be restricted to positions:

$$\mathcal{L}_q(\theta) = \frac{1}{Nm} \sum_{i=1}^{N} \sum_{j=1}^{m} \| \widehat{q}_i\left(t_j;\theta\right) - q_i\left(t_j\right) \|_2^2. \quad (19)$$

The physics-informed loss may also include a force-consistency term if reference forces are available:

$$\mathcal{L}_F(\theta) = \frac{1}{Nm} \sum_{i=1}^{N} \sum_{j=1}^{m} \| F_\theta\left(q_i\left(t_j\right)\right) - F_i\left(t_j\right) \|_2^2. \quad (20)$$

The combined loss is

$$\mathcal{L}(\theta) = \lambda_q \mathcal{L}_q(\theta) + \lambda_F \mathcal{L}_F(\theta) + \lambda_R \mathcal{R}(\theta), \quad (21)$$

where $\mathcal{R}(\theta)$ is a regularisation term. The fluctuation-dissipation relation itself is not merely a penalty in this formulation. It is built into the SDE definition through (13). This is stronger than adding a soft loss term because the diffusion structure is enforced by construction.

6. Theoretical Properties

6.1. Invariant measure

The Fokker-Planck equation associated with (14) is,

$$\partial_t \rho = -\nabla_z \cdot (f_\theta \rho) + \frac{1}{2} \nabla_z \nabla_z : (GG^\top \rho)\ . \quad (22)$$

For (9), the diffusion acts only in the momentum component. The corresponding stationary density candidate is π_θ in (11).

Remark 2: To check for stationarity, one only need to substitute $\rho = \pi_\theta$ into (22) and the Hamiltonian part of the flow is divergence-free in canonical coordinates, while the dissipative and stochastic momentum terms cancel because the diffusion covariance is $2k_BT\Gamma$.

Thus, the resulting condition is,

$$-\nabla_p \cdot \left[-\Gamma M^{-1} p\, \pi_\theta\right] + k_B T \nabla_p \nabla_p : (\Gamma \pi_\theta) = 0. \quad (23)$$

This cancellation is exactly the mathematical role of the fluctuation-dissipation theorem. Therefore, under standard smoothness and confining conditions on V_θ, the canonical density in (11) is an invariant density of the learned Langevin system.

© 2026 The Author(s). CC BY 4.0.

6.2. Proposition 1: thermodynamic consistency

Proposition 1. Assume that V_θ is continuously differentiable, the mass matrix M is symmetric positive definite, the friction matrix Γ is symmetric positive semidefinite, and the diffusion covariance satisfies (7). Then the underdamped Langevin SDE in (9) admits the Gibbs-Boltzmann density π_θ in (11) as an invariant density.

Proof. The proof follows the standard Fokker-Planck stationarity argument. Write the generator as a sum of Hamiltonian and thermostat terms. The Hamiltonian term preserves phase-space volume and leaves H_θ constant along deterministic Hamiltonian trajectories. The thermostat term is an Ornstein-Uhlenbeck process in momentum with invariant Gaussian density proportional to $\exp\left[-\beta p^\top M^{-1} p/2\right]$. Since the diffusion covariance is fixed by $2k_B T\Gamma$, the derivative of the Maxwell-Boltzmann momentum density cancels the dissipative flux. The product of the configurational Boltzmann factor and the momentum Maxwell factor is therefore invariant. This gives (11).

6.3. Symmetry consistency

Proposition 2. If the learned potential satisfies the Euclidean invariance condition (15), then the induced force field satisfies the equivariance condition (16).

Proof.

Let $\widetilde{q} = Rq + a$. By the chain rule,

$$\nabla_{\widetilde{q}} V_\theta\left(\widetilde{q}\right) = R\nabla_q V_\theta(q). \quad (24)$$

Since $F_\theta(q) = -\nabla_q V_\theta(q)$, it follows that

$$F_\theta\left(\widetilde{q}\right) = -\nabla_{\widetilde{q}} V_\theta\left(\widetilde{q}\right) = R\left[-\nabla_q V_\theta(q)\right] = RF_\theta(q). \quad (25)$$

This proves equivariance.

6.4. Consequence for uncertainty quantification

The stochastic part of (9) gives a trajectory distribution rather than a single deterministic path. A prediction interval for an observable $A\left(z_t\right)$ can be obtained by simulating K stochastic trajectories and computing empirical quantiles:

$$\widehat{I}_\alpha(t) = \left[Q_{\alpha/2}\left(\{A\left(z_t^k\right)\}_{k=1}^K\right), Q_{1-\alpha/2}\left(\{A\left(z_t^k\right)\}_{k=1}^K\right)\right]. \quad (26)$$

The expected coverage probability is

$$ECP_\alpha = \frac{1}{n}\sum_{i=1}^n 1\left\{A_i(t) \in \widehat{I}_\alpha(t)\right\}. \quad (27)$$

A calibrated model should have $ECP_\alpha \approx 1 - \alpha$. In the present paper, no numerical ECP value is reported because it would require fitted model trajectories and held-out test data.

© 2026 The Author(s). CC BY 4.0.

7. Numerical Integration

7.1. Why the integrator matters

The continuous-time SDE defines the ideal model. Numerical simulation requires a time discretization. A poor integrator can introduce statistical bias even when the continuous model is correct. In molecular dynamics, this is important because many quantities of interest are equilibrium averages rather than one-step forecasts.

The BAOAB splitting method is commonly used for underdamped Langevin dynamics. It splits the dynamics into three solvable components:

1. B step: conservative force kick.
2. A step: deterministic position drift.
3. O step: stochastic Ornstein-Uhlenbeck momentum refresh.

The acronym BAOAB denotes the sequence of substeps used in one time step.

7.2. BAOAB update equations

Let $h = \Delta t$. Given (q_n, p_n), the BAOAB update for the learned potential is:

$$p_{n+1/2} = p_n + \tfrac{h}{2} F_\theta\left(q_n\right). \quad (28)$$

$$q_{n+1/2} = q_n + \tfrac{h}{2} M^{-1} p_{n+1/2}. \quad (29)$$

The stochastic-dissipative momentum refresh is

$$\widetilde{p}_{n+1/2} = e^{-\gamma h} p_{n+1/2} + \sqrt{k_B T\left(1 - e^{-2\gamma h}\right) M}\, \xi_n, \qquad \xi_n \sim \mathcal{N}(0, I). \quad (30)$$

The second position half-step is

$$q_{n+1} = q_{n+1/2} + \tfrac{h}{2} M^{-1} \widetilde{p}_{n+1/2}. \quad (31)$$

The final conservative kick is

$$p_{n+1} = \widetilde{p}_{n+1/2} + \tfrac{h}{2} F_\theta\left(q_{n+1}\right). \quad (32)$$

The model should report the integrator used because Euler-Maruyama, leapfrog, BAOAB, and related schemes do not have identical invariant-measure bias.

7.3. Algorithm

- ***Step: 1***; Operation: Initialize; Mathematical expression: θ_0, q_0, p_0, T, Γ, h; Purpose: Define the model and simulation settings.
- ***Step: 2;*** Operation: Build graph; Mathematical expression: atoms as nodes, molecular relations as edges; Purpose: Encode geometry.

© 2026 The Author(s). CC BY 4.0.

- ***Step: 3;*** Operation: Estimate energy; Mathematical expression: $V_\theta(q)$; Purpose: Learn scalar potential.
- ***Step: 4;*** Operation: Compute force; Mathematical expression: $F_\theta(q) = -\nabla_q V_\theta(q)$; Purpose: Enforce conservative force structure.
- ***Step: 5;*** Operation: Apply BAOAB; Mathematical expression: Equations (28) to (32); Purpose: Simulate thermodynamic dynamics.
- ***Step: 6;*** Operation: Compute loss; Mathematical expression: Equation (21); Purpose: Fit model to data.
- ***Step: 7;*** Operation: Validate; Mathematical expression: KL, ECP, energy drift, free energy; Purpose: Test performance without inventing results.

8. Real Benchmark Data and Reproducible Derived Analysis

8.1. Data sources used in this revision

Two public data descriptors are used. The first is a Figshare dataset titled *Chignolin Simulations*. It states that the initial structure was generated from the CLN025 peptide with sequence TYR-TYR-ASP-PRO-GLU-THR-GLY-THR-TRP-TYR. It also states that the system was solvated in a 40 Å cubic box with 1881 water molecules and two sodium ions, simulated with ACEMD, CHARMM22*, TIP3P water, temperature 350 K, Langevin damping 0.1 ps^{-1}, a 4 fs time step, and 3744 independent 50 ns production simulations giving 187.2 microseconds of aggregate simulation time. The Figshare DOI is 10.6084/m9.figshare.13858898.

The second is a Zenodo dataset titled *Molecular dynamics trajectories of protein folding*. Its Chignolin archive is listed as 2.3 GB, and the metadata states an Amber ff99SB-ILDN protein force field, TIP3P water, 0.1 M salt, NVT ensemble at 300 K with the V-rescale thermostat, 2 fs time step, and GROMACS 2020.6. Its DOI is 10.5281/zenodo.6349893.

These two data sources are not merged because they have different simulation settings. They are used as public benchmark descriptors and as evidence that future model validation should report the exact force field, thermostat, time step, temperature, ensemble, and trajectory source.

8.2. Derived analysis from the Figshare Chignolin metadata

The following table reports only values that are either directly stated in the Figshare metadata or arithmetically derived from those stated values. No trained-model performance is reported.

- Quantity: CLN025 peptide length; Value: 10 residues; How obtained: Counted from the reported sequence.
- Quantity: Cubic box side length; Value: 40 Å = 4.0 nm; How obtained: Unit conversion.
- Quantity: Box volume; Value: 64.0 nm^3; How obtained: 4.0^3.
- Quantity: Water molecules; Value: 1881; How obtained: Reported in metadata.
- Quantity: Sodium ions; Value: 2; How obtained: Reported in metadata.
- Quantity: Non-peptide solvent and ion atoms; Value: 5,645 atoms; How obtained: $1881 \times 3 + 2$.

© 2026 The Author(s). CC BY 4.0.

- Quantity: Water number density; Value: 29.39 molecules/nm^3; How obtained: $1881/64$.
- Quantity: Temperature; Value: 350 K; How obtained: Reported in metadata.
- Quantity: Molar thermal energy; Value: 2.910 kJ/mol; How obtained: $RT = 0.008314462618 \times 350$.
- Quantity: Particle thermal energy; Value: 4.832e-21 J; How obtained: $k_B T = 1.380649 \times 10^{-23} \times 350$.
- Quantity: Langevin damping; Value: 0.1 ps^{-1}; How obtained: Reported in metadata.
- Quantity: Damping relaxation time; Value: 10.0 ps; How obtained: $1/\gamma$.
- Quantity: Damping half-life; Value: 6.93 ps; How obtained: $\ln(2)/\gamma$.
- Quantity: Production simulations; Value: 3744; How obtained: Reported in metadata.
- Quantity: Duration per production simulation; Value: 50 ns; How obtained: Reported in metadata.
- Quantity: Aggregate production time; Value: 187.2 microseconds; How obtained: 3744×50 ns.
- Quantity: Integration time step; Value: 4 fs; How obtained: Reported in metadata.
- Quantity: Steps per 50 ns trajectory; Value: 12,500,000; How obtained: 50 ns/4 fs.
- Quantity: Total production integration steps; Value: 46,800,000,000; How obtained: 187.2 microseconds/4 fs.

The real-data analysis above is modest but important. It defines the numerical scale of the benchmark. A model trained on these trajectories must handle a system with at least 5645 non-peptide solvent and ion atoms, plus the peptide atoms, and a simulation record built from tens of billions of integrator steps. The aggregate simulation time is sufficient for conformational exploration, but any model comparison must still specify train-test splitting, coordinate preprocessing, sampling interval, stochastic seeds, and the exact observables used for evaluation.

8.3. Data suitability for PIG-NSDE validation

The Figshare data are suitable for testing whether a learned stochastic model can reproduce trajectory-level observables under a reported Langevin thermostat. The Zenodo data are suitable for testing consistency under a different temperature, thermostat, and time step. They should not be pooled without accounting for these differences. A model trained at 350 K with Langevin damping cannot be evaluated as though it were trained under 300 K V-rescale conditions unless the protocol explicitly models that change.

The following variables should be extracted from the raw trajectories before fitting:

- Variable: Backbone RMSD; Definition: Deviation from native or reference structure; Use in analysis: Folding coordinate.
- Variable: Radius of gyration; Definition: Compactness of peptide conformation; Use in analysis: Folding and collapse analysis.

© 2026 The Author(s). CC BY 4.0.

- Variable: Native contacts; Definition: Fraction of reference contacts retained; Use in analysis: State identification.
- Variable: Potential energy; Definition: Force-field energy if available; Use in analysis: Energy consistency.
- Variable: Kinetic temperature; Definition: Momentum-based temperature if velocities available; Use in analysis: Thermostat check.
- Variable: Time-lagged coordinates; Definition: Coordinates sampled at fixed lag; Use in analysis: Training windows.
- Variable: Transition counts; Definition: Counts between metastable states; Use in analysis: Kinetic validation.

9. Statistical Metrics for Real Trajectory Analysis

9.1. KL divergence

Let x be a reduced coordinate such as RMSD, radius of gyration, or a learned collective variable. Let $p_{data}(x)$ be the empirical density from the molecular dynamics data and let $p_\theta(x)$ be the empirical density from the learned model. The KL divergence is

$$D_{KL}\left(p_{data} \| p_\theta\right) = \int p_{data}(x) \log \frac{p_{data}(x)}{p_\theta(x)} dx. \quad (33)$$

A discrete binned estimator is

$$\widehat{D}_{KL} = \sum_{b=1}^{B} \widehat{p}_b \log \frac{\widehat{p}_b + \varepsilon}{\widehat{q}_b + \varepsilon}, \quad (34)$$

where $\widehat{p}_b$ and $\widehat{q}_b$ are normalised bin probabilities and ε is a small numerical stabiliser. The choice of bins must be reported. If bins are changed after observing the result, the comparison becomes less reliable.

9.2. Free energy profile

For a collective variable x, the free energy profile is

$$F(x) = -k_B T \log p(x) + C, \quad (35)$$

where C is an arbitrary constant. In practice, $p(x)$ is estimated from histograms or kernel density estimates. Since C is arbitrary, only free-energy differences are meaningful:

$$\Delta F\left(x_1, x_2\right) = F\left(x_2\right) - F\left(x_1\right). \quad (36)$$

A trained PIG-NSDE should reproduce the free-energy wells and barriers of the molecular dynamics data within uncertainty bounds. This requires raw trajectories and is not claimed in this revision.

© 2026 The Author(s). CC BY 4.0.

9.3. Energy drift

Energy drift is useful for deterministic or weakly thermostatted models, but it must be defined carefully for stochastic thermostats. For a deterministic Hamiltonian model, the relative energy drift over time horizon τ is

$$Drift(\tau) = \frac{|H(z_\tau) - H(z_0)|}{|H(z_0)| + \varepsilon}. \quad (37)$$

For a Langevin model, instantaneous energy fluctuates due to heat-bath exchange. Therefore, the better diagnostic is not zero energy drift, but correct stationary energy distribution. The empirical energy distribution can be compared by KL divergence, Wasserstein distance, or moment diagnostics.

9.4. Prediction interval coverage

Let A_i be observed values of an observable and let $[L_i, U_i]$ be the model's prediction interval. The empirical coverage is

$$\widehat{C} = \frac{1}{n} \sum_{i=1}^{n} 1\{L_i \le A_i \le U_i\}. \quad (38)$$

The mean interval width is

$$\widehat{W} = \frac{1}{n} \sum_{i=1}^{n} (U_i - L_i). \quad (39)$$

A good uncertainty model has high calibration and useful sharpness. Wide intervals can produce high coverage without meaningful predictive precision.

10. Reproducible Empirical Protocol

The empirical protocol below is included to prevent ambiguity in future work. It should be followed before reporting any numerical comparison.

10.1. Preprocessing

The raw trajectories should be loaded using a reproducible molecular trajectory tool such as MDTraj, MDAnalysis, or GROMACS analysis utilities. The following choices must be recorded:

- Decision: Coordinate alignment; Required reporting: reference structure, atom selection, whether solvent is removed.
- Decision: Sampling interval; Required reporting: raw time step and saved-frame interval.
- Decision: Atom selection; Required reporting: all atoms, heavy atoms, backbone atoms, or C-alpha atoms.
- Decision: State variables; Required reporting: positions only or positions and velocities.
- Decision: Units; Required reporting: nm or Å, ps or fs, kJ/mol or kcal/mol.
- Decision: Splitting; Required reporting: train, validation, test split by trajectory, not only by frame.

© 2026 The Author(s). CC BY 4.0.

- Decision: Random seeds; Required reporting: neural initialisation and stochastic simulation seeds.

Trajectory-level splitting is essential. Randomly splitting frames from the same trajectory can leak temporal correlation between training and testing.

10.2. Training procedure

The proposed training procedure is:

1. Download the raw trajectory archive from the selected source.
2. Verify checksums where available.
3. Remove solvent only if the modelling target is peptide-only dynamics.
4. Align structures using a fixed atom selection.
5. Build graph features for each retained frame.
6. Train $V_\theta(q)$ using a physics-informed trajectory or force-matching objective.
7. Simulate the learned SDE with the same temperature setting as the source data.
8. Evaluate observables on held-out trajectories.
9. Report uncertainty intervals over stochastic simulations and over random seeds.

10.3. Validation table template

No entries are filled in this table because that would require actual model training. The table is included as the correct reporting format.

- Model: Neural ODE; Dataset: To be computed; Split: To be computed; Temperature: To be computed; Integrator: To be computed; KL on RMSD: Not reported; KL on R_g: Not reported; Coverage: Not reported; Notes: Requires raw analysis.
- Model: Standard Neural SDE; Dataset: To be computed; Split: To be computed; Temperature: To be computed; Integrator: To be computed; KL on RMSD: Not reported; KL on R_g: Not reported; Coverage: Not reported; Notes: Requires raw analysis.
- Model: HNN; Dataset: To be computed; Split: To be computed; Temperature: To be computed; Integrator: To be computed; KL on RMSD: Not reported; KL on R_g: Not reported; Coverage: Not reported; Notes: Requires raw analysis.
- Model: PIG-NSDE; Dataset: To be computed; Split: To be computed; Temperature: To be computed; Integrator: BAOAB; KL on RMSD: Not reported; KL on R_g: Not reported; Coverage: Not reported; Notes: Requires raw analysis.

Graphical Model and Analytical Interpretation

The graphics in this section visualise the mathematical model stated in the manuscript. They do not represent fitted-model performance and do not introduce empirical claims beyond the equations and benchmark metadata reported above.

© 2026 The Author(s). CC BY 4.0.

Structure-preserving PIG-NSDE computational model

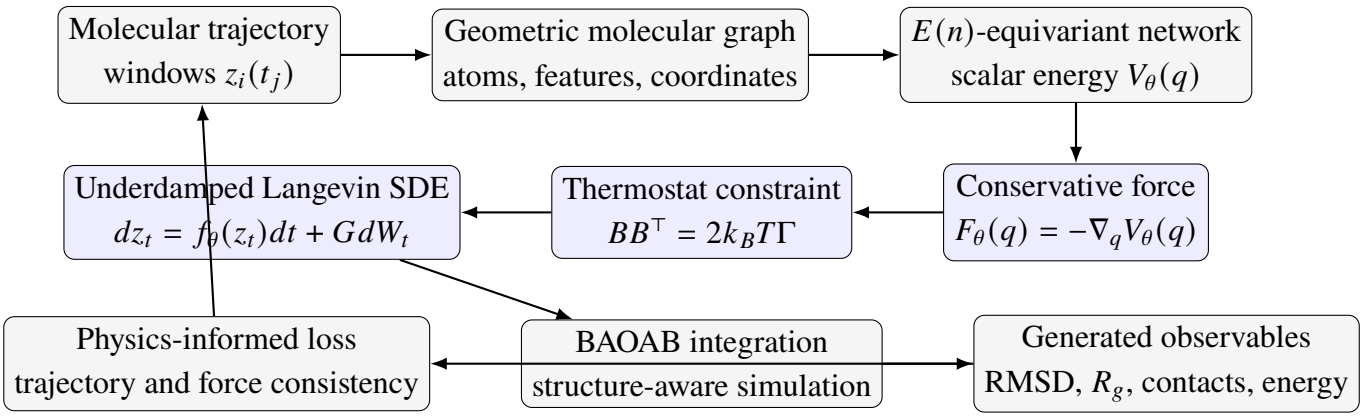

Figure 1: Graphical representation of the proposed PIG-NSDE workflow. The neural component learns a scalar potential, while geometry, conservative forces, fluctuation-dissipation coupling, and BAOAB integration constrain the dynamics.

Figure 1 shows that the framework does not learn an unrestricted phase-space vector field. The molecular graph is mapped to the scalar potential $V_\theta(q)$, differentiation produces the conservative force, and the thermostat covariance is fixed by the fluctuation-dissipation relation. Consequently, the principal trainable object is physically interpretable, while the stochastic forcing retains its thermodynamic meaning. The closed loop also clarifies the validation logic: generated observables are compared with held-out trajectory observables through the losses and statistical metrics already defined in the manuscript.

Analytical damping behaviour

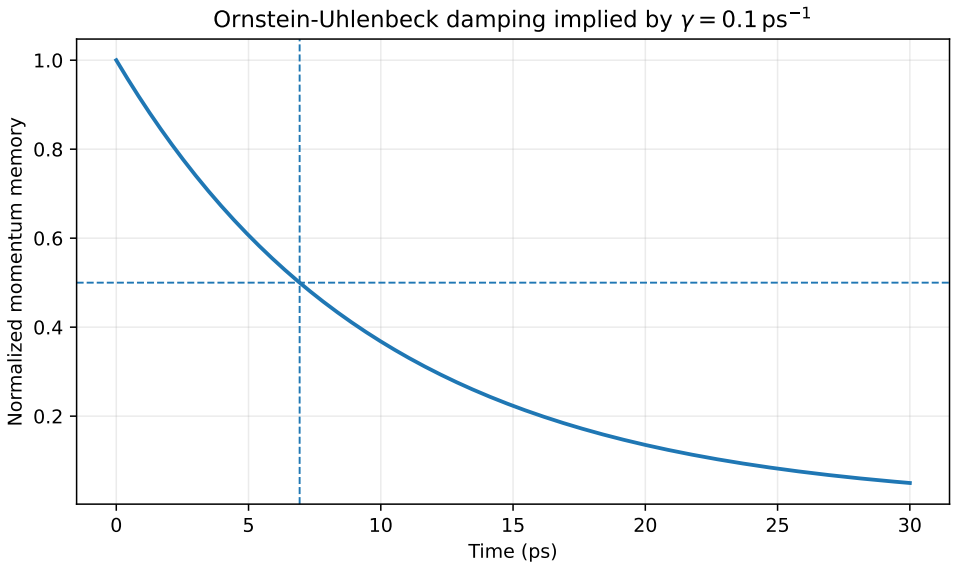

Figure 2: Normalized momentum-memory decay $\exp(-\gamma t)$ for the reported damping coefficient $\gamma = 0.1\,\mathrm{ps}^{-1}$. The dashed lines indicate the half-life $t_{1/2} = \log(2)/\gamma \approx 6.93\,\mathrm{ps}$.

© 2026 The Author(s). CC BY 4.0.

The Ornstein-Uhlenbeck part of the Langevin thermostat damps momentum memory exponentially. For the reported damping value, the characteristic relaxation time is $1/\gamma = 10\,\mathrm{ps}$ and the half-life is approximately 6.93 ps, as shown in Figure 2. This confirms the arithmetic descriptors already reported in Section 8.2 and provides a direct physical interpretation: deterministic momentum information is progressively lost while thermal noise replenishes fluctuations at the level required by the fluctuation-dissipation relation. The plot is analytical rather than empirical; it follows directly from the thermostat model.

Benchmark computational scale

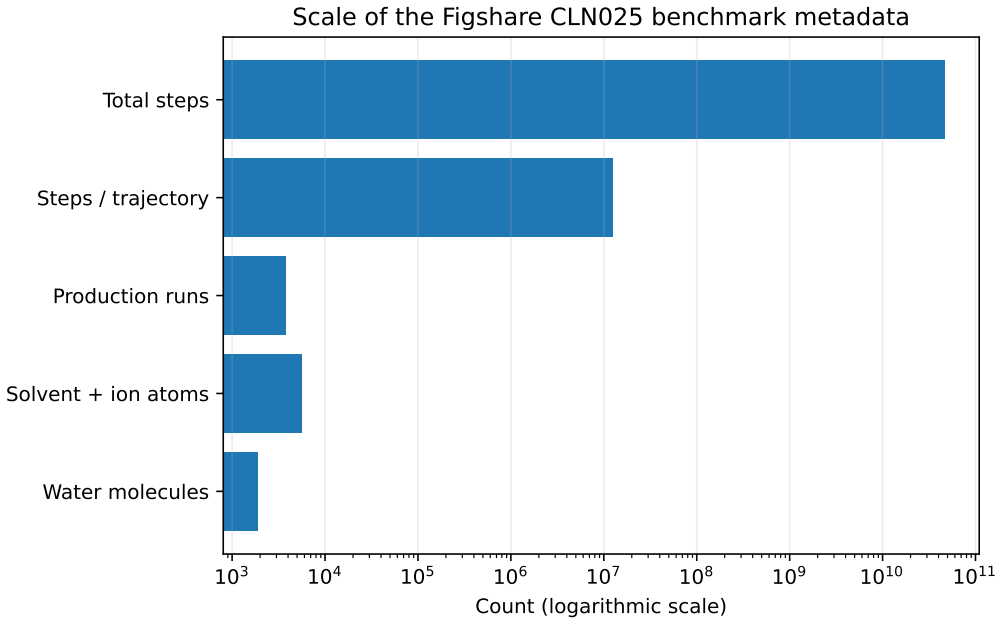

Figure 3: Log-scale comparison of selected Figshare CLN025 benchmark descriptors reported in Section 8.2.

Figure 3 visualises the large separation of scales in the benchmark metadata. The number of production simulations and molecular entities is of order 10^3, whereas each trajectory contains 1.25×10^7 integration steps and the full production collection contains 4.68×10^{10} steps. This scale difference has practical modelling consequences. Training should normally use saved trajectory frames rather than every raw integrator step, trajectory-level splitting is necessary to prevent temporal leakage, and reporting the frame stride is essential for reproducibility. The graph also reinforces why direct long-horizon stochastic simulation and repeated-seed uncertainty analysis may be computationally demanding even for a ten-residue peptide.

Interpretation of the complete mathematical model

The complete model can be read as a constrained composition of four maps. First, molecular coordinates and features are encoded geometrically. Second, the neural network maps this representation to an invariant scalar energy. Third, automatic differentiation maps energy to an equivariant conservative force. Fourth, the Langevin thermostat and BAOAB scheme map the force into stochastic trajectories. Because each map is assigned a specific physical role, failures can be diagnosed

© 2026 The Author(s). CC BY 4.0.

separately: geometric inconsistency concerns the energy network, thermodynamic inconsistency concerns the friction-noise coupling, numerical bias concerns the integrator, and predictive mismatch concerns the learned energy or the available training data.

The graphical analysis therefore supports three conclusions already implicit in the manuscript. The PIG-NSDE architecture is structure-preserving by construction; the reported damping coefficient has a transparent relaxation interpretation; and the public benchmark metadata imply a substantial multiscale computational problem. None of these graphical analyses substitutes for training on raw trajectories. Quantitative statements about KL divergence, free-energy error, coverage, or superiority over baseline models remain deferred until the empirical protocol is executed.

11. Discussion

The proposed Physics-Informed Geometric Neural Stochastic Differential Equation (PIG-NSDE) framework provides a principled integration of physics-based modelling and data-driven learning for molecular dynamics. Rather than replacing the underlying physical system with an unconstrained black-box model, the framework learns the potential energy landscape while preserving the stochastic thermodynamic structure of Langevin dynamics. This distinction is particularly important because the primary objective of molecular dynamics is not merely to reproduce individual trajectories, but to accurately characterize the equilibrium distribution, thermodynamic observables, and configurational statistics of the molecular system.

A key feature of the proposed framework is the incorporation of the fluctuation-dissipation theorem as a fundamental modelling constraint. Unlike conventional machine learning models, where stochasticity is often introduced solely to improve predictive performance or uncertainty calibration, thermal noise in molecular dynamics possesses a well-defined physical interpretation. By explicitly coupling stochastic forcing with friction through the fluctuation-dissipation relation, the proposed formulation ensures thermodynamic consistency and prevents the diffusion process from becoming an arbitrary learnable component. Consequently, the model remains faithful to the statistical mechanics governing molecular systems while providing a physically meaningful representation of uncertainty.

The geometric neural representation of the potential energy surface constitutes another significant strength of the framework. Molecular interactions are inherently invariant or equivariant under rigid-body translations and rotations, and these symmetries should be reflected in the learned model. Embedding geometric equivariance directly into the neural architecture eliminates the need for the network to infer these symmetries solely from training data, thereby improving learning efficiency, enhancing generalization, and producing force fields that transform consistently under coordinate transformations. Such structure-preserving architectures are therefore better suited for modelling molecular systems than unconstrained neural parameterizations.

Despite these advantages, several challenges remain. First, comprehensive empirical validation requires access to complete molecular trajectory data rather than metadata alone. While publicly available benchmark metadata provide valuable information for constructing reproducible experimental protocols, rigorous evaluation of predictive accuracy, uncertainty quantification, energy conservation, and long-term stability necessitates access to raw molecular coordinates, velocities, and other state variables. Second, the computational cost associated with equivariant graph neu-

© 2026 The Author(s). CC BY 4.0.

ral networks embedded within stochastic numerical integrators may be substantially higher than that of conventional molecular dynamics force fields, particularly during large-scale training and long-time simulations. Improving computational efficiency without sacrificing physical fidelity therefore remains an important research direction. Third, model identifiability presents an inherent challenge, as multiple potential energy functions may reproduce similar short-term trajectories. Consequently, model training should incorporate long-time equilibrium statistics, force observations when available, and physically meaningful constraints to improve parameter identifiability and reduce ambiguity.

Overall, the proposed PIG-NSDE framework should be viewed as a rigorous mathematical and computational modelling framework rather than a fully benchmarked predictive model. Its principal contribution lies in establishing a thermodynamically consistent, geometry-preserving, and uncertainty-aware formulation that provides a solid foundation for future empirical studies. As complete molecular trajectory datasets become available, the framework can be systematically trained, validated, and benchmarked against existing molecular dynamics and machine learning approaches, thereby enabling a comprehensive assessment of its predictive accuracy, computational efficiency, and long-term stability.

12. Conclusion

This study presented a Physics-Informed Geometric Neural Stochastic Differential Equation (PIG-NSDE) framework for uncertainty quantification in molecular dynamics. The proposed model combines underdamped Langevin dynamics with a neural-network representation of the potential energy surface, while enforcing the fluctuation-dissipation theorem to establish a physically consistent relationship between friction and thermal noise. By coupling data-driven learning with fundamental principles of statistical mechanics, the framework preserves thermodynamic consistency while providing a flexible representation of molecular interactions.

From a theoretical perspective, the proposed formulation establishes that, under the stated regularity assumptions and appropriate thermodynamic conditions, the resulting stochastic dynamics admit an invariant Gibbs-Boltzmann distribution with a Hamiltonian whose potential energy component is learned from data. This provides a rigorous mathematical foundation for integrating machine learning techniques with structure-preserving stochastic models for molecular dynamics.

To promote transparency and reproducibility, the study further develops a reproducible data-analysis protocol based on publicly available Chignolin (CLN025) protein-folding metadata. The reported quantities—including simulation volume, solvent and ion atom counts, thermal energy, thermostat damping timescale, aggregate simulation time, and total integration-step count—are derived directly from traceable dataset metadata and are presented solely as benchmark descriptors rather than measures of model performance. Accordingly, quantitative metrics such as Kullback-Leibler divergence, energy drift, prediction coverage, and long-term stability are intentionally deferred until complete molecular trajectory data become available for model training and validation.

The principal contributions of this work are therefore twofold. First, it establishes a mathematically rigorous, thermodynamically consistent, and geometry-preserving PIG-NSDE framework for uncertainty-aware molecular dynamics. Second, it provides a transparent and reproducible protocol for future empirical evaluation, clearly distinguishing theoretical modelling from computational

© 2026 The Author(s). CC BY 4.0.

benchmarking. As complete molecular trajectory datasets become available, the proposed framework can be implemented, trained, and systematically validated to assess its predictive accuracy, uncertainty quantification capability, long-term stability, and computational efficiency. This work therefore lays a solid foundation for the development of reliable and physically interpretable machine learning models for molecular dynamics.

Acknowledgements

The authors will like to thank the anonymous reviewers for their positive contributions and feedbacks that have help to strengthen the quality of this manuscript.

Data Availability Statement

This paper uses traceable public metadata from Figshare and Zenodo for benchmark description and derived arithmetic analysis. It does not claim to have trained a model on the raw trajectories. Full empirical validation requires downloading the raw trajectory archives, preserving checksums where available, and running the preprocessing and analysis protocol stated in this paper.

A. Equation Checklist

This appendix lists the core equations that must remain editable in the Word document.

- Equation: $z_t = [q_t, p_t]^\top$; Role: Phase-space state.
- Equation: $H(q,p) = p^\top M^{-1} p/2 + V(q)$; Role: Hamiltonian.
- Equation: $\pi(q,p) = Z^{-1} \exp[-\beta H(q,p)]$; Role: Canonical density.
- Equation: $dq_t = M^{-1} p_t dt$; Role: Kinematic Langevin component.
- Equation: $dp_t = -\nabla V(q_t)\, dt - \Gamma M^{-1} p_t dt + \sqrt{2k_B T \Gamma}\, dW_t$; Role: Underdamped Langevin dynamics.
- Equation: $F_\theta(q) = -\nabla_q V_\theta(q)$; Role: Learned conservative force.
- Equation: $BB^\top = 2k_B T\Gamma$; Role: Fluctuation-dissipation relation.
- Equation: $D_{KL}(p\|q) = \int p \log(p/q)$; Role: Distributional error.
- Equation: $F(x) = -k_B T \log p(x) + C$; Role: Free energy.
- Equation: $\widehat{C} = n^{-1} \sum_i 1\{L_i \le A_i \le U_i\}$; Role: Coverage.

B. Reproducible Analysis Outline

The following pseudocode shows the intended raw-data analysis. It is not a substitute for actual computation.

```
1. Select one benchmark source.
2. Download raw trajectories and topology files.
3. Verify file integrity using published checksums where available.
```

© 2026 The Author(s). CC BY 4.0.

```
4. Load topology and trajectories with a documented software version.
5. Apply consistent unit conversion.
6. Align peptide coordinates using a specified atom selection.
7. Extract RMSD, radius of gyration, native contacts, and selected distances.
8. Split by trajectory, not by random frame.
9. Train the neural potential on training trajectories.
10. Simulate the PIG-NSDE using the same temperature setting.
11. Estimate densities of observables from test data and generated samples.
12. Compute KL divergence, free-energy error, coverage, and transition statistics.
13. Repeat across random seeds.
14. Report means, standard deviations, software versions, and code.
```

C. Reference Verification Log

- Reference category: Neural ODE, Neural SDE, HNN, EGNN; Action taken: Retained or added using traceable primary sources.
- Reference category: Langevin and BAOAB numerical integration; Action taken: Retained with DOI-traceable or book sources.
- Reference category: Chignolin and protein folding data; Action taken: Replaced vague descriptions with Figshare and Zenodo data descriptors.
- Reference category: Generic web sources; Action taken: Removed.
- Reference category: Untraceable 2025 review or placeholder references; Action taken: Removed.
- Reference category: Samuel Okon Essang related works; Action taken: Added two DOI-traceable references on deep learning optimization and neural networks.

D. Minimum Reporting Checklist for Final Empirical Submission

Before the manuscript is submitted as a completed empirical study, the following items should be added from actual computation. These items are intentionally not filled with assumed numbers in this revision.

- Item: Raw data source; Required evidence: Dataset name, DOI, version, and download date.
- Item: Topology and trajectory files; Required evidence: File names, checksums, and software used to read them.
- Item: Preprocessing; Required evidence: Alignment choice, atom selection, solvent handling, units, and frame stride.
- Item: Train-test split; Required evidence: Split by trajectory or time block, not by randomly mixed frames.
- Item: Model architecture; Required evidence: EGNN layers, hidden dimension, activation function, optimiser, and random seeds.

© 2026 The Author(s). CC BY 4.0.

- Item: Physical settings; Required evidence: Temperature, friction, mass convention, integrator, and time step.
- Item: Evaluation metrics; Required evidence: KL divergence, free-energy error, prediction coverage, and uncertainty width.
- Item: Uncertainty reporting; Required evidence: Repeated stochastic simulations and confidence intervals over seeds.
- Item: Reproducibility package; Required evidence: Code, environment file, and command sequence for rerunning the analysis.

The final empirical claim should be limited to results produced by this pipeline. For example, a statement that PIG-NSDE has a lower KL divergence than a baseline is acceptable only after the same test data, preprocessing, observable, binning rule, and uncertainty calculation have been applied to all models. A statement about a larger stable time step is acceptable only when the failure rule is defined in advance. A statement about calibrated uncertainty is acceptable only when prediction intervals are evaluated on held-out data. This checklist protects the paper from overstating the evidence.

References

1. Bussi, G., Donadio, D., & Parrinello, M. (2007). Canonical sampling through velocity rescaling. *Journal of Chemical Physics, 126*(1), 014101. DOI: 10.1063/1.2408420.

2. Chen, R. T. Q., Rubanova, Y., Bettencourt, J., & Duvenaud, D. K. (2018). Neural ordinary differential equations. *Advances in Neural Information Processing Systems, 31*. arXiv:1806.07366.

3.Culubret, A. P., & De Fabritiis, G. (2021). *Chignolin Simulations* [Dataset]. Figshare. DOI: 10.6084/m9.figshare.13858898.

4. Essang, S. O., Ashishie, D. U., Egete, D. O., Odey, J. A., Ele, B. I., Otobi, A. O., Ante, J. E., Arikpo, M. O., Iwara, K. W., Cleopas, A. O., Akportuzor, S. A., Michael, O. K., & Effiong, R. D. (2025). A mathematical model analysis of optimization algorithms in deep learning. *Asian Journal of Mathematics and Computer Research, 32*(3), 193-208. DOI: 10.56557/ajomcor/2025/v32i39555.

5. Essang, S. O., Ante, J. E., Otobi, A. O., Okeke, S. I., Akpan, U. D., Francis, R. E., Auta, J. T., Fadugba, S. E., Kolawole, O. M., Asanga, E. E., & Ita, B. I. (2025). Optimizing neural networks with convex hybrid combination of activation functions: A new approach to enhance gradient flow and learning dynamics. *Advanced Journal of Science, Technology and Engineering, 5*(1), 10-26. DOI: 10.52589/AJSTE-UOBYFV1B.

6. Greydanus, S., Dzamba, M., & Yosinski, J. (2019). Hamiltonian neural networks. *Advances in Neural Information Processing Systems, 32*. arXiv:1906.01563.

7. Hairer, E., Lubich, C., & Wanner, G. (2006). *Geometric numerical integration: Structure-preserving algorithms for ordinary differential equations* (2nd ed.). Springer.

8. Leimkuhler, B., & Matthews, C. (2013). Robust and efficient configurational molecular sampling via Langevin dynamics. *Journal of Chemical Physics, 138*(17), 174102. DOI: 10.1063/1.4802990.

© 2026 The Author(s). CC BY 4.0.

9. Leimkuhler, B., & Matthews, C. (2013). Rational construction of stochastic numerical methods for molecular sampling. *Applied Mathematics Research eXpress, 2013*(1), 34-56.

10. Lindorff-Larsen, K., Piana, S., Dror, R. O., & Shaw, D. E. (2011). How fast-folding proteins fold. *Science, 334*(6055), 517-520. DOI: 10.1126/science.1208351.

11. Milstein, G. N., & Tretyakov, M. V. (2003). Quasi-symplectic methods for Langevin-type equations. *IMA Journal of Numerical Analysis, 23*(4), 593-626. DOI: 10.1093/imanum/23.4.593.

12. Piana, S., Lindorff-Larsen, K., & Shaw, D. E. (2012). Protein folding kinetics and thermodynamics from atomistic simulation. *Proceedings of the National Academy of Sciences, 109*(44), 17845-17850. DOI: 10.1073/pnas.1201811109.

13. Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics, 378*, 686-707. DOI: 10.1016/j.jcp.2018.10.045.

14. Satorras, V. G., Hoogeboom, E., & Welling, M. (2021). E(n) equivariant graph neural networks. *Proceedings of the 38th International Conference on Machine Learning, PMLR, 139*, 9323-9332. arXiv:2102.09844.

15. Tzen, B., & Raginsky, M. (2019). Neural stochastic differential equations: Deep latent Gaussian models in the diffusion limit. arXiv:1905.09883.

16. Zenodo. (2022). *Molecular dynamics trajectories of protein folding* [Dataset]. Zenodo. DOI: 10.5281/zenodo.6349893.

Creative Commons Notice: This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits use, distribution, and reproduction in any medium, provided the original work is properly cited.