LEARNING COMPOSABLE TIME-UNCONDITIONAL MOLECULAR ENERGY LANDSCAPES FOR ZERO-SHOT ENVIRONMENTAL TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion and flow-matching models are now the standard generative approaches for molecular conformational ensembles, but their learned vector fields are non-conservative and yield no scalar potential. Energy-based formulations restore conservativity by taking forces as the gradient of a learned scalar, yet the landscapes they recover remain noise-conditional or are defined only up to an arbitrary scale. In neither case can they be added to a classical force field in physical units or composed with external environmental potentials inside a molecular dynamics engine. Classical potentials are useful precisely because environmental interactions enter additively: a solute potential can be moved between vacuum, solvent, and a protein binding pocket simply by summing it with a potential for the new environment. A generative model that learns a transferable molecular energy could enable generation inside new physical environments, such as different solutes, temperatures, pressures and other simulation contexts. We introduce LEVI, a time-unconditional energy-based generative model that is trained by both score matching and an energy-regression term that aligns it to the physical energies carried by the data. We also introduce LEVI-MD, which composes the resulting potential with external environmental terms at every step of a molecular dynamics integrator and computes physical interaction energies between solute and environment atoms. In this way, a single solute model can be used to infer conformations in new environments to which it was never exposed during training. Further, in this approach the environment atoms are dynamic and can physically respond to the generated molecule during an inference trajectory. We validate this in two systems: a Langevin dynamics system where data lies on a known one-dimensional manifold, and on alanine dipeptide, where a learned potential trained on vacuum structures alone can be composed with explicit water at inference, recapitulating known conformational shifts between vacuum and water.

1 INTRODUCTION

In classical molecular dynamics, environmental interactions enter the potential energy additively. For a solute in an environment, the total potential partitions as the sum of system, environment and interaction components, $U = U^{\text{sol}} + U^{\text{env}} + U^{\text{int}}$, with forces $\mathbf{F} = -\nabla U$ (MacKerell Jr, 2004; Jorgensen & Tirado-Rives, 2005). This additivity makes classical force fields transferable. A solute potential parameterized once can be used in vacuum, in explicit water, or in a heterogeneous environment such as a protein binding pocket by simply adding it to the potential energy of the new environment.

Within this decomposition, the solute term U^{sol} is traditionally evaluated either with fixed empirical force fields, which compromise quantum-mechanical accuracy, or with *ab initio* calculations, which are prohibitively expensive for extensive thermodynamic sampling (Gkeka et al., 2020; MacKerell Jr, 2004; Marx & Hutter, 2000). Discriminative machine learning force fields bridge this accuracy-efficiency gap, but depend on *ab initio* energy labels for non-equilibrium trajectories (Behler & Parrinello, 2007; Unke et al., 2021). Deep generative models circumvent this bottleneck by learning equilibrium structural distributions directly from data without non-equilibrium force labels (Song et al., 2020; Lipman et al., 2022; Noé et al., 2019; Köhler et al., 2020).

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.

Their outputs, however, cannot be substituted into $\mathbf{F} = -\nabla U$. These architectures parameterize an unconstrained vector field $\mathbf{s}_\theta(\mathbf{x}, t) \in \mathbb{R}^{3N}$ rather than the gradient of a scalar. Deployed as surrogate thermodynamic forces, such fields are non-conservative. They perform spurious work along closed phase-space loops ($\oint_C \mathbf{s}_\theta \cdot d\mathbf{x} \neq 0$) and act as an artificial heat pump that drives energy drift and simulation instability (Chmiela et al., 2017; Arts et al., 2023; Bigi et al., 2024). Energy-based models (Zhao et al., 2016; Du & Mordatch, 2019; LeCun et al., 2006) instead learn a scalar function $E_\theta(\mathbf{x})$ and compute the vector field as its negative gradient, $-\nabla_\mathbf{x} E_\theta$, which is conservative by construction (Plainer et al., 2026; Batzner et al., 2022; Chao et al., 2022). Although inference-time composability has been exploited to enforce geometric shape constraints (Griesbacher et al., 2026) or heuristic force penalties (Lai et al., 2026), transferring a learned solute model to an unseen physical environment zero-shot has not been demonstrated without model retraining. In this work, we introduce **LEVI** (**L**earned **E**nergy **V**ia **I**terative denoising), a diffusion-inspired energy-based model whose landscape is recovered in physical units, and **LEVI-MD**, an inference-time framework that composes the learned energy with external environmental terms, such as those used in molecular dynamics simulations. For training, denoising score matching shapes the landscape while an energy-regression term aligns it to the physical energies attached to the data and fixes the scale that score matching alone leaves free. At inference, LEVI-MD evaluates $\mathbf{F}_{\text{total}} = -\nabla_\mathbf{x}(U_\theta^{\text{sol}} + U^{\text{env}} + U^{\text{int}})$ at every integrator step, adding an environment to the learned potential without retraining. LEVI never sees an environment during training. An environmental potential can nonetheless be added to it at inference, and a single trained solute model transfers zero-shot to new environments.

Our contributions in this work are as follows:

1. We add an energy-regression term when training an energy-based model. This term aligns the learned landscape to the physical energies attached to the data and reproduces thermodynamic free energies between data subpopulations. It also places the learned potential on the same energy scale as a classical force field, which allows additivity with environmental potential energies.
2. We show that an environmental term can be added to a learned potential at inference time, so that a solute model trained in vacuum alone can be used to infer structures in new environments (e.g. explicit water).
3. We develop LEVI-MD, an inference-time composition framework for learned molecular potentials, capable of coupling a single trained solute model to environments introduced at inference and recovering the conformational response they induce.

2 Related Work

Energy-Based Generative Models for Molecules. Defining forces as the negative gradient of a learned scalar, $\mathbf{F}_\theta(\mathbf{x}) = -\nabla_\mathbf{x} E_\theta(\mathbf{x})$, guarantees a curl-free field and removes the spurious work that unconstrained score networks accumulate along closed loops (LeCun et al., 2006; Arts et al., 2023; Bigi et al., 2024). Arts et al. (2023) obtain such a field from a denoising objective, showing that the gradient of a scalar energy network serves as a coarse-grained force field; trained on Boltzmann-distributed data at known temperature, these forces carry physical units, but the network is noise-conditional and the energy landscape changes as a function of the diffusion time. *ScoreMD* (Plainer et al., 2026) traces the inconsistency arising from the time-dependent energy landscape to the small-timestep regime and regularizes the learned score using a Fokker–Planck equation loss term, so that conventional diffusion inference and the dynamics driven by the learned forces agree. Both approaches retain the diffusion time as a network input, so what they define is a family of surfaces indexed by noise level rather than a single energy landscape. The case for removing that dependence has been made in the image domain, where models inspired by flow-matching (e.g. *Equilibrium Matching* (Wang & Du, 2025) and *Energy Matching* (Balcerak et al., 2026)) were used to generate images using time-independent scalar fields. Molecular instances followed: *EBMol* (Griesbacher et al., 2026) trains atom-additive scalar potentials via Restoring Field Matching, and *GPFF* (Hessmann et al., 2026) trains machine learning force fields (MLFFs) on a quadratic pseudo-PES defined relative to reference equilibria, whose minima coincide with known equilibrium structures by construction rather than being recovered from energies. In both cases the scalar field is anchored by a modeling convention, either a zero-point imposed at the data or the shape of a quadratic centered on reference equilibria, rather than by a measured energy. The energy surface is therefore

determined only up to an arbitrary scaling factor. As this is immaterial for generation and ranking applications, both works treat alignment to physical energies as a direction left open (Griesbacher et al., 2026; Hessmann et al., 2026). It becomes material once the learned term is summed with a classical one, since $E_\theta + U^{\text{env}}$ then weights solute against environment by a factor that no thermostat or integrator can recover.

Inference-Time Composition and Environmental Transferability. Steering a trained generative model at inference is desirable, as it avoids retraining for every new target environment, and energy parameterizations make this direct, since scalar landscapes can be summed and sampled jointly. Within molecular generation, composition has been used to bias geometry rather than energetics. Griesbacher et al. (2026) sum E_θ with a differentiable potential defined from the PCA eigenvalues of atom positions to steer samples toward linear, planar, or spherical shapes. Lai et al. (2026) add a classical potential directly to the sampler. They evaluate a molecular mechanics force field (MMFF94) whose non-bonded van der Waals and electrostatic terms are extended to act between the ligand and a protein pocket, and add its gradient to each sampling step of a pretrained flow-matching or diffusion model without retraining. The added term is a classical energy, but it is added to a velocity field rather than a force, so the two are not commensurate. Their relative weight is carried by a free hyperparameter λ, bounded above by a Lipschitz condition that guarantees non-increasing energy per step rather than by any physical consideration. The procedure biases a denoising trajectory toward energetically favorable geometries; it does not integrate an equation of motion.

LEVI-MD evaluates $\mathbf{F}_{\text{total}} = -\nabla_{\mathbf{x}}(U_\theta^{\text{sol}} + U^{\text{env}} + U^{\text{int}})$ at every step of a continuous integrator, with no relative weight to tune. To the best of our knowledge, this is the first framework in which a learned generative potential is composed additively with a classical environmental potential in physical units inside a molecular dynamics simulation, allowing a solute potential trained in vacuum to be simulated in an environment supplied only at inference.

3 METHOD

We introduce LEVI, a time-unconditional energy-based model that learns a scalar potential $E_\theta(\mathbf{x})$ over molecular configurations, and LEVI-MD, an inference-time framework that composes it with external environmental terms.

3.1 PROBLEM FORMULATION

The target of this work is a learned potential that can take the place of U^{sol} in the classical decomposition. For a solute embedded in an environment, the total potential energy separates as

$$U \;=\; U^{\text{sol}} \;+\; U^{\text{env}} \;+\; U^{\text{int}}, \tag{1}$$

where U^{sol} is internal to the solute, U^{env} is internal to the environment, and U^{int} collects the nonbonded interactions between them, evaluated from the partial charges and Lennard-Jones parameters of the atoms on both sides. LEVI supplies a surrogate (E_θ) for the first term alone, which quantitatively approximates U^{sol} at the data, but is otherwise smooth and convex, with gradients that are everywhere directed toward points in the dataset. We note that E_θ is never exposed to the other two energy terms (U^{env} and U^{int}) during training.

We parameterize a scalar potential $E_\theta(\mathbf{x}; \mathcal{G}, \mathbf{a}) \in \mathbb{R}$ over Cartesian coordinates $\mathbf{x} \in \mathbb{R}^{N\times 3}$, conditioned on the covalent topology $\mathcal{G}$, which records connectivity and bond order, and on the per-atom nonbonded attributes $\mathbf{a}_i = (q_i, \epsilon_i, \sigma_i)$, the partial charge and the Lennard-Jones well depth and diameter of atom i. Forces follow by differentiation with respect to coordinates, $\mathbf{F}_\theta = -\nabla_{\mathbf{x}} E_\theta$.

The energy is assembled in the same manner as a classical force field. Each configuration contributes an atom term, a bond term over covalently bonded pairs, an angle term over bonded triples, and a dihedral term over bonded quadruples, read respectively from interatomic distances, bond angles, and torsion angles. Every contribution is divided among the atoms it involves, so the total can be written as a sum of per-atom energies, $E_\theta(\mathbf{x}) = \sum_{i=1}^{N} E_\theta^{(i)}(\mathbf{x})$.

As the energy reads the geometry only through interatomic distances and bond and torsion angles, E_θ is invariant to global rotation and translation and $\mathbf{F}_\theta$ is correspondingly equivariant; the learned

term transforms like a physical potential. Conditioning on $\mathbf{a}$ rather than on categorical atom identities is what allows the learned term to be composed: the attributes that specify E_θ are the same ones a classical force field requires to evaluate U^{int}, so a solute described to LEVI is described to the environment in the same language. In the experiments reported here $\mathcal{G}$ and $\mathbf{a}$ are fixed and only $\mathbf{x}$ is generated.

3.2 Learning a calibrated energy landscape

Shaping an energy landscape and placing it on a physical scale are separate problems, and score matching solves only the first. The shaping objective constrains E_θ through its gradient alone and is blind to the affine map that carries a dimensionless surface onto a physical energy axis. Existing energy-based generative models have closed that freedom by convention rather than by measurement, anchoring the energy of clean data toward zero (Griesbacher et al., 2026). Such a landscape reproduces the relative shape of the distribution while carrying no physical units, and cannot be added to an environmental potential measured in physical unit. We fix it with data instead: molecular datasets carry a potential energy alongside each conformation, and we regress the energy of E_θ onto those values, fitting a learned scale and offset alongside the surface itself. A single network is therefore shaped away from the data by a score-matching term evaluated on noise-corrupted configurations, and pinned at the data by a regression term evaluated on clean structures.

Time-unconditional denoising score matching. Clean configurations are corrupted by a variance-exploding forward process: with a noise level indexed by $t \in \{0, \dots, T-1\}$, $\bar{\alpha}_t = \prod_{s \le t}(1-\beta_s)$ and $\sigma_t = \sqrt{(1-\bar{\alpha}_t)/\bar{\alpha}_t}$,

$$\mathbf{x}_t \;=\; \mathbf{x}_0 \;+\; \sigma_t\, \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

where $\mathbf{x}_0$ is centered and $\boldsymbol{\epsilon}$ is drawn and then centered in the zero-centroid subspace $\mathbb{V} = \{\mathbf{x} : \sum_i \mathbf{x}_i = \mathbf{0}\}$. Restricting the corruption to $\mathbb{V}$ is required because a translation-invariant energy has a gradient with zero net sum, so every force that E_θ can produce lies in $\mathbb{V}$, and a target with a translational component would leave a residual no conservative field could cancel. The network is trained so that the force it generates opposes the displacement that was applied,

$$\mathcal{L}_{\text{DSM}}(\theta) \;=\; \mathbb{E}_{t,\,\mathbf{x}_0,\,\boldsymbol{\epsilon}}\Big[\big\|\mathbf{F}_\theta(\mathbf{x}_t) + \boldsymbol{\epsilon}\big\|^2\Big], \qquad \mathbf{F}_\theta = -\nabla_{\mathbf{x}} E_\theta, \tag{3}$$

with t drawn uniformly over the schedule.

The forward kernel is Gaussian, so its score is available in closed form. Writing $q_t(\mathbf{x} \mid \mathbf{x}_0) \propto \exp\!\big(-\|\mathbf{x}-\mathbf{x}_0\|^2/2\sigma_t^2\big)$ and differentiating,

$$\nabla_{\mathbf{x}} \log q_t(\mathbf{x} \mid \mathbf{x}_0) \;=\; -\frac{\mathbf{x}-\mathbf{x}_0}{\sigma_t^2} \;=\; -\frac{\sigma_t \boldsymbol{\epsilon}}{\sigma_t^2} \;=\; -\frac{\boldsymbol{\epsilon}}{\sigma_t}, \tag{4}$$

so the target $-\boldsymbol{\epsilon}$ in Eq. 3 is the kernel score scaled by σ_t. Since the network sees only $\mathbf{x}_t$, the squared residual is minimized by the conditional mean of that target, substituting $-\boldsymbol{\epsilon}$ using Eq. 2 we get the minimizer of the force function:

$$\mathbf{F}^*(\mathbf{x}) \;=\; \mathbb{E}_{\mathbf{x}_0 \sim p(x_0),\, t \sim \mathcal{U}\{0, T-1\}}\left[-\boldsymbol{\epsilon} \mid \mathbf{x}_t\right]. \tag{5}$$

Note that $\mathbf{F}^*(\mathbf{x})$ does not depend on the strength of the noise applied. This implies that the magnitude of the force should be roughly constant as a function of $\mathbf{x}$, with a corresponding energy surface (E_θ) that rises linearly as we move away from the data. The relative depth of these energy wells is not yet fixed.

Because E_θ takes no noise level as input, a single surface is fit against residuals drawn from the entire schedule at once. Equation 3 reads E_θ only through its gradient, and the residual is formed against a displacement of unit variance at every noise level, so nothing in the objective refers to an energy scale. Therefore, the landscape it produces is correct in shape but still arbitrary in magnitude, which is insufficient for composability, as the magnitude of E_θ must determine the precise balance of system and environmental forces.

Algorithm 1 One LEVI training step

1: sample $\mathbf{x}_0 \sim p_{\text{data}}$ with label $U(\mathbf{x}_0)$, and center it
2: sample $t \sim \mathcal{U}\{0, T-1\}$, set $\sigma_t = \sqrt{(1-\bar{\alpha}_t)/\bar{\alpha}_t}$
3: draw $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and center it
4: $\mathbf{x}_t \leftarrow \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}$
5: $\mathbf{F}_\theta \leftarrow -\nabla_{\mathbf{x}} E_\theta(\mathbf{x}_t; \mathcal{G}, \mathbf{a})$ {first forward pass}
6: $\mathcal{L}_{\text{DSM}} \leftarrow \|\mathbf{F}_\theta + \boldsymbol{\epsilon}\|^2$
7: $\mathcal{L}_{\text{E}} \leftarrow \big(E_\theta(\mathbf{x}_0; \mathcal{G}, \mathbf{a}) - (U(\mathbf{x}_0) - b)/s\big)^2$ {second forward pass}
8: update (θ, s, b) on $\mathcal{L}_{\text{DSM}} + \gamma_E \mathcal{L}_{\text{E}}$

Energy regression on the data. Each conformation in the training set carries the physical potential energy assigned to it, $U(\mathbf{x}_0)$. A scale s and an offset b, trained jointly with θ, relate that label to the network output, and the loss is formed as:

$$\mathcal{L}_{\text{E}}(\theta, s, b) \;=\; \mathbb{E}_{\mathbf{x}_0}\left[\Big(E_\theta(\mathbf{x}_0) - \frac{U(\mathbf{x}_0) - b}{s}\Big)^2\right], \tag{6}$$

where $E_\theta(\mathbf{x}_0) = \sum_{i=1}^{N} E_\theta^{(i)}(\mathbf{x}_0)$ sums the contributions of the N atoms. Inverting the same relation reads the trained network in physical units,

$$U_\theta^{\text{sol}}(\mathbf{x}) \;=\; s\, E_\theta(\mathbf{x}) \;+\; b, \tag{7}$$

which is the quantity we refer to as the learned energy that can be additively composed with a classical potential.

Training objective. The two residuals are evaluated on different configurations. The energy labels exist only at the clean structures, while the score-matching residual is calculated on noised data. Each training step therefore evaluates the network twice, and the total objective is their sum,

$$\mathcal{L}(\theta, s, b) \;=\; \mathcal{L}_{\text{DSM}}(\theta) \;+\; \gamma_E \mathcal{L}_{\text{E}}(\theta, s, b), \tag{8}$$

where γ_E is a scalar defining the relative strength of the two losses. Algorithm 1 shows one step.

3.3 LEVI-MD: Composing the learned potential at inference

At inference we sample solute conformations inside an environment by integrating the forces of a composed potential. A trained LEVI model supplies the solute term in physical units, so it enters the sum on the same footing as the classical terms and requires no relative weighting. Formally, for a solute with coordinates $\mathbf{x}$ and an environment with coordinates $\mathbf{y}$, LEVI-MD integrates

$$U_{\text{total}}(\mathbf{x}, \mathbf{y}) \;=\; U_\theta^{\text{sol}}(\mathbf{x}) \;+\; U^{\text{env}}(\mathbf{y}) \;+\; U^{\text{int}}(\mathbf{x}, \mathbf{y}), \qquad \mathbf{F} \;=\; -\nabla U_{\text{total}}, \tag{9}$$

with $U_\theta^{\text{sol}} = sE_\theta + b$ from Eq. 7. This recovers Eq. 1 with U^{sol} replaced by the learned surrogate. The environment term is taken unmodified from a classical force field, and the non-bonded interaction term is defined as follows:

$$U^{\text{int}} \;=\; \sum_{i \in \text{sol}} \sum_{j \in \text{env}} \left[4\epsilon_{ij} \left(\Big(\frac{\sigma_{ij}}{r_{ij}}\Big)^{12} - \Big(\frac{\sigma_{ij}}{r_{ij}}\Big)^{6} \right) \;+\; \frac{k_e\, q_i q_j}{r_{ij}} \right], \tag{10}$$

with $\sigma_{ij} = \frac{1}{2}(\sigma_i + \sigma_j)$, $\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$, and k_e the Coulomb constant. The parameters on the solute side are the same attributes $(\mathbf{a})$ that condition E_θ, so no assignment step intervenes between the learned term and the classical one.

Initialization and coupling. Inference-time simulations start from solute atoms placed at random positions inside the environment, using fixed chemical identities $\mathbf{a}$, with U_θ^{sol} introduced over a sequence of stages at the target temperature (Appendix A.3, Table 3). At the beginning its contribution is zero and the environment alone determines the motion; as the contribution grows the learned potential takes over, and all reported trajectories in experiment section are run in the final stage, where U_θ^{sol} supplies the entire intramolecular energy of the solute. A flat-bottomed harmonic restraint with force constant 10000 kJ/mol/nm^2 is applied between the nearest atoms of disconnected components to keep the particles from dispersing before they assemble; it is zero whenever they form a single connected group, with gaps no larger than 2.5 Å.

Algorithm 2 LEVI-MD inference

Require: trained (θ, s, b); attributes $\mathbf{a}$; topology $\mathcal{G}$; initial environment coordinates $\mathbf{y_0}$

1: initialize $\mathbf{x}$ at random positions inside the environment
2: **for** each step **do**
3: $\quad U_{\text{total}} \leftarrow s\, E_\theta(\mathbf{x}; \mathcal{G}, \mathbf{a}) + b \;+\; U^{\text{env}}(\mathbf{y}) \;+\; U^{\text{int}}(\mathbf{x}, \mathbf{y})$
4: $\quad$ advance $(\mathbf{x}, \mathbf{y})$ by one Langevin step at temperature T under $-\nabla U_{\text{total}}$
5: **end for**
6: **return** the trajectory

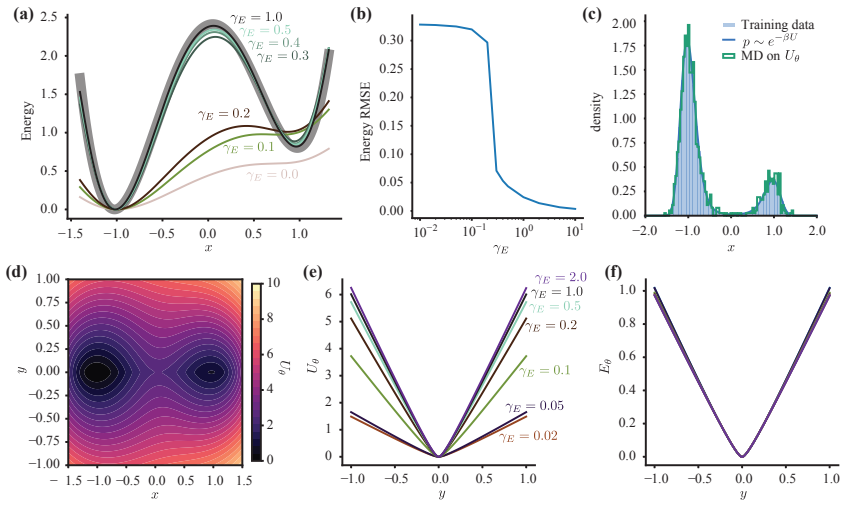

Figure 1: **Langevin system in 1D and 2D. (a)** The final trained energy surfaces as a function of x for a set of γ_E values. The surfaces approach the training potential (thick gray line) as γ_E increases. **(b)** The root mean squared error in the energy surface, evaluated over the data support, as a function of γ_E. **(c)** Density from a set of molecular dynamics simulations using the learned potential (green line) agrees well with the expected Boltzmann probabilities (blue line). **(d)** The 2D surface, obtained with $\gamma_E = 1.0$, showing the restoring force onto the $y = 0$ data manifold. **(e)** A slice across this manifold at $x = -1.0$, for a set of γ_E values. **(f)** The same slice, plotting E_θ, showing that differences in the restoring force arise solely from differences in the energy scale parameter, s.

Integration. The system is integrated with Langevin dynamics at temperature T and friction $\zeta = 10\ \text{ps}^{-1}$. Each step of size $\Delta t = 2$ fs uses one network evaluation to obtain $-\nabla U_\theta^{\text{sol}}$; the remaining terms are evaluated classically using the OpenMM software Eastman et al. (2017). Relative to a conventional simulation the only modification is the substitution of these forces for $-\nabla U^{\text{sol}}$, shown in Algorithm 2.

4 Experiments

4.1 Langevin system

We test the training objective on a one-dimensional system with closed-form potential, force, and equilibrium distribution: a tilted double well $V(x) = 2(x^2 - 1)^2 + 0.4x$ at $k_B T = 0.5$. Training data are x values taken from independent overdamped Langevin trajectories. E_θ is a scalar MLP over x; training and optimizer details are in Appendix A.1. We examine relative energy weightings γ_E ranging from 0.01 to 10.0 (Figure 1a,b) and find that $\gamma_E \geq 0.3$ is necessary to accurately model the forces and energies, as quantified by the energy RMSE. Running dynamics under this learned potential recreates Boltzmann probability distributions at the given temperature. Figure 1c shows the probability density from 1000 trajectories of 20000 steps each. To demonstrate composability, we also combine this learned potential with different biasing potentials unseen during training (Appendix B). Dynamics under the combined potential also recreate the appropriate ensembles, as long as the biases do not push sampling beyond the data support.

The 1D system lacks a key component, which is a learned relaxation back towards the data manifold. To demonstrate this we augment our training data with another dimension, y, which is set to zero for each point. Noise is applied to both dimensions during training, so U_θ must learn to be convex, with a global minimum at $y = 0$. Figure 1c shows $U_\theta(x, y)$, with the same potential wells along x and a landscape that smoothly descends toward $y = 0$. Figures 1d and 1e show that this restoring force is roughly constant as a function of y as predicted by Equation 5. As the score matching objective is independent of s, each potential learns the same E_θ, although differences in s lead to different strengths of the U_θ restoring forces.

4.2 Alanine Dipeptide

We next evaluate LEVI and LEVI-MD on alanine dipeptide, a system whose conformational preferences are known to change substantially between vacuum and water. Three questions are separable and we address them in order: whether the regression term places the learned surface on the scale of the reference energies, whether the calibrated potential reproduces the ensemble it was trained on when integrated on its own, and whether composing it with an explicit solvent it never saw recovers the conformational shift that solvent induces.

System. Alanine dipeptide (N-acetyl-L-alanine-N'-methylamide) is the smallest peptide with a complete backbone, and its conformational state is described almost entirely by the dihedral angles ϕ and ψ (Chekmarev et al., 2004; Parchansky et al., 2013; Wang et al., 2014). In the gas phase, structures with localized intramolecular hydrogen bonds are the most stable, where the $C_{7\text{eq}}$ conformer ($\phi \approx -83°, \psi \approx 76°$), the C_5 conformer ($\phi \approx -161°, \psi \approx 157°$), and the $C_{7\text{ax}}$ conformer ($\phi \approx 74°, \psi \approx -54°$) constitute the predominant energetic minima. In water solution, this landscape is significantly altered; interaction with the polar solvent shifts the equilibrium to favor extended and helical architectures. This environmental change makes the extended PPII structure ($\phi \approx -59°, \psi \approx 152°$) the most favorable conformer, followed closely by the helical α_R state (shifting near $\phi \approx -73°, \psi \approx -18°$), the α_L state ($\phi \approx 55°, \psi \approx 38°$), and a slightly shifted C_5 region ($\phi \approx -153°, \psi \approx 155°$) or a shifted $C_{7\text{eq}}$ region ($\phi \approx -82°, \psi \approx 86°$) (Garcia-Prieto et al., 2011; Weise & Weisshaar, 2003). Because of this environmental shift, alanine dipeptide provides a suitable test of our framework, namely whether a vacuum-trained energy surface can reconstruct the solvated ensemble without retraining, simply by adding an explicit solvent potential at inference.

Dataset. We generated the datasets by running molecular dynamics (MD) simulations in OpenMM (Eastman et al., 2017) using the CHARMM36m force field (Huang et al., 2017) at 300 K. We simulated the peptide in vacuum (five independent 100 ns trajectories) and in an explicit periodic water box (four independent 100 ns trajectories), saving coordinates every 2 ps to yield $50,000$ frames per trajectory. LEVI is trained exclusively on the vacuum data. We allocate four vacuum trajectories ($200,000$ frames) for training and one ($50,000$ frames) for validation. The four solvated trajectories are strictly unobserved during training and serve only as an evaluation reference. The MD columns of Figure 2a detail the conformational distributions of the reference ensembles. The two-dimensional probability density of the (ϕ, ψ) dihedral angles is displayed in the first column. The sampled frames are projected onto a 90×90 grid ($4°$ bins), where raw counts are normalized by both the total frame count and the bin area ($16\,\text{deg}^2$), and subsequently smoothed with a periodic $14°$ Gaussian kernel wrapping at the $\pm 180°$ boundaries.

In the third column, the sampled frames are integrated into discrete populations using the fixed $120°$ Ramachandran partition defined by Rubio-Martinez et al. (2017), established *a priori* so the resulting populations reflect intrinsic ensemble properties rather than clustering artifacts. The vacuum environment concentrates most of its probability mass in the hydrogen-bonded C_5 and $C_{7\text{eq}}$ basins (Figure 2a, third column), while explicit solvent shifts the majority of the population to the extended PPII conformer and substantially increases the helical α_R population. Predicting this environmental distribution shift from a vacuum-trained surface is the zero-shot challenge for LEVI. The target energy distribution of the vacuum training frames, characterized in Figure 2b, aligns with these basin populations. The dominant C_5 and $C_{7\text{eq}}$ minima correspond to the lowest median energies, showing LEVI should regress a landscape where rare conformations map to high-energy regions, on average.

Training setup. The energy network embeds each atom from its attributes $\mathbf{a}_i$ alone, with no element identity supplied, and refines those features with six stacks of SchNet (Schütt et al., 2017)

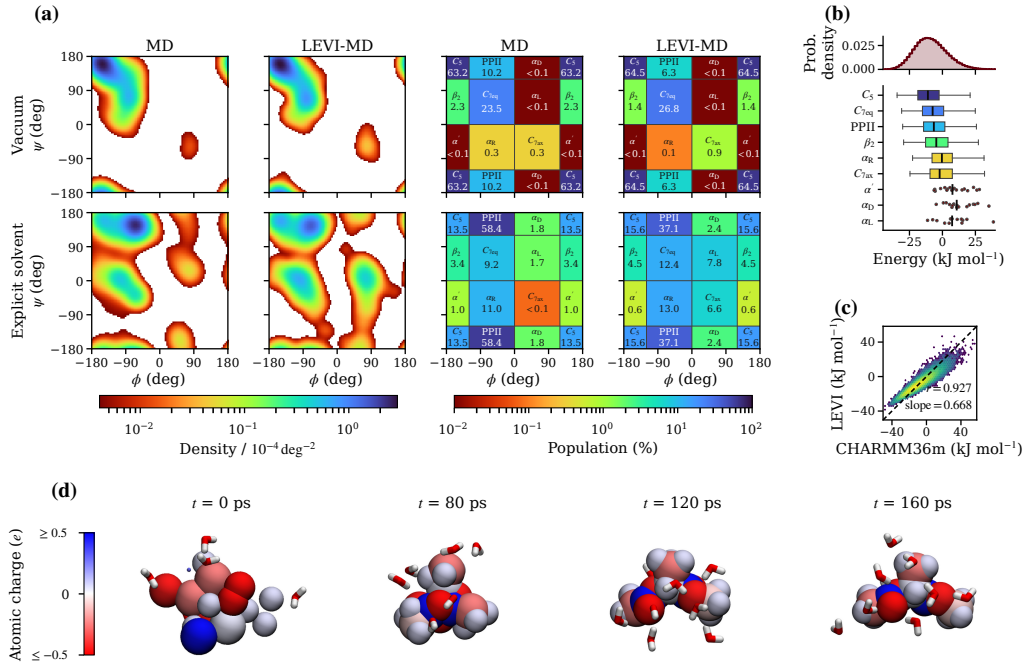

Figure 2: **Alanine dipeptide in vacuum and in explicit solvent.** **(a)** Rows denote the environment. The MD columns represent 200,000 frames per ensemble and the LEVI-MD columns 1,000 independent simulations. The first two columns show the sampled probability density of the (ϕ, ψ) dihedral angles for the reference MD and for LEVI-MD, from a 4° histogram smoothed with a periodic 14° Gaussian kernel. The last two columns show the same frames integrated into discrete conformational basins using the fixed 120° partition of Rubio-Martinez et al. (2017); the outer ϕ columns and outer ψ rows are single continuous regions wrapped across the periodic $\pm 180^\circ$ boundaries, with values repeated for clarity. LEVI is trained on the vacuum MD ensemble alone; the solvent is supplied only at inference. **(b)** Distribution of the CHARMM36m potential energy over the vacuum training frames, which is the regression target, overall (top) and within each basin (bottom). **(c)** LEVI energy against the CHARMM36m reference on the 50,000 held-out vacuum structures, shaded by density, with the dashed line marking exact agreement. **(d)** Snapshots from one LEVI-MD trajectory in explicit water, showing the solute and the water molecules within 2.3 Å of it. Atoms are colored by partial charge and scaled by Lennard-Jones radius.

interaction blocks in the modified form introduced by AGDIFF (Vieira Wyzykowski et al., 2025). The atom, bond, angle, and dihedral contributions of Section 3.1 are produced by MLPs over the resulting node and edge features and aggregated onto the atoms involved. The scale s and offset b of Eq. 7 are free parameters optimized alongside the network weights against the CHARMM36m energies of the training frames. Training details and hyperparameters are given in Appendix A.2.

Inference simulation setup. Composed simulations place the solute in explicit water and integrate Eq. 9 with Langevin dynamics at 300 K, matching the reference trajectories. U_θ^{sol} is introduced over a sequence of stages as described in Section 3.3, and all reported frames are sampled at full coupling. Each reported ensemble comprises 1,000 independent simulations, each initialized from random solute coordinates, so the sampled frames are uncorrelated. Vacuum runs are compared against the training trajectory and solvated runs against the reference solvated set, analyzed on the same grid and partition as Figure 2a. Simulation parameters and analysis details are listed in Appendix A.3.

4.3 Results

The regression term places the surface on the reference scale. Evaluated on the 50,000 validation vacuum frames, the learned energy tracks the CHARMM36m value of each structure with $r = 0.927$ (Figure 2c). The least-squares slope is 0.668. Conformations are ordered correctly and

in the right units, but the learned surface is compressed and places the lowest and highest energy structures closer to the mean.

The calibrated surface in vacuum reproduces the density of training ensemble. Run without an environment term, LEVI-MD samples the vacuum ensemble from random initial coordinates (Figure 2a, top row). The basin rank ordering of the reference is recovered, with C_5 at 64.5% against 63.2%, $C_{7\text{eq}}$ at 26.8% against 23.5%, PPII at 6.3% against 10.2%, and the minor basins remain sparsely populated. This establishes that the learned potential alone supports stable dynamics and reproduces data probability densities found in the training set. Bond lengths and intramolecular distances for both environments are compared with the reference in Appendix C.

Adding an environment term at inference models the solvated ensemble. The same U_θ^{sol}, trained on vacuum data, is composed with explicit water through Eq. 9 (Figure 2a, bottom row). The conformational preferences invert as they do in the reference. PPII rises from 6.3% in vacuum to 37.1% and becomes the dominant basin, α_{R} from 0.1% to 13.0%, while C_5 falls from 64.5% to 15.6% and $C_{7\text{eq}}$ from 26.8% to 12.4%. Three of these land within a few points of the reference solvated populations of 13.5%, 9.2%, and 11.0% for C_5, $C_{7\text{eq}}$, and α_{R}; PPII reaches the top of the rank order but at 37.1% against 58.4%, with the shortfall distributed over α_{L} and $C_{7\text{ax}}$, which the reference visits at 1.7% and below 0.1%. These populations come from $1{,}000$ independent simulations rather than from a continuous trajectory, so they are free of the correlation that makes the effective sample size of the reference far smaller than its frame count. The change in its sampled distribution comes entirely from the environmental terms calculated throughout the inference trajectory. Importantly, the solvent atoms respond to the solute over the trajectory, forming polar interactions with the solute (Figure 2d) that precisely rebalance the conformational energy landscape of alanine dipeptide conformations.

5 Conclusion, Limitations and Future Work

In this work, we addressed the gap between learning a molecular energy landscape and using it as a physical potential for zero-shot environmental transferability. Score matching constrains the landscape only through its gradient, which fixes its shape but leaves its scale free, and a landscape without a scale cannot be added to a classical force field. We closed that freedom with the data by regressing the network output onto the potential energies attached to the training conformations, fitting a scale and offset alongside the surface. The calibrated potential can then be composed additively with environmental terms inside a molecular dynamics integrator, which we implemented as LEVI-MD. On alanine dipeptide, a potential trained on vacuum trajectories alone and composed with explicit water at inference reproduced the conformational rank order of the solvated reference and recovered the rise of PPII and α_{R} and the loss of C_5 and $C_{7\text{eq}}$ populations from $1{,}000$ independent inference trajectories.

A couple of limitations remain. The regression places the surface on the reference axis but compresses it, with a least-squares slope of 0.668, so energy differences between conformations are systematically underestimated. The composed ensembles show the corresponding effect, with more population in rarely-visited basins, and PPII falls short of its reference share. Only the regression term carries physical units, and it is evaluated only at clean structures, so the landscape is calibrated near the training data and extrapolated elsewhere. Additionally, the energy label is available only at clean structures, so each training step requires a second forward pass.

The composition framework is not what limits these results. LEVI-MD requires training only the solute term and it accepts any environment expressible as a classical potential, so a protein binding pocket could be supplied in the same manner as explicit water. LEVI could also be trained for de novo generation, by incorporating attribute displacements in the score-matching framework. A model of that kind, composed through the LEVI-MD framework, could generate candidates inside a target pocket without ever having seen that target during training, with the ligand assembled while considering the interactions with binding pocket atoms. Importantly, the binding pocket atoms would in turn be able to respond to the presence of the ligand atoms, allowing for ligand-induced conformational change during de novo generation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing methods, providing feedback on research methodology or experiments, and interpreting results. We have not used generative AI tools for generating synthetic data sets, developing theoretical models or conceptual frameworks, proposing or refining hypotheses, cleaning and reformatting datasets, or supporting qualitative and thematic data analysis. Other tasks such as assisting in the writing of proofs, providing critical ingredients for proving mathematical claims, and assisting with translation are not applicable to this work. Additionally, we used an LLM to draft and edit sections of the manuscript, improve its readability, identify relevant literature, and create or modify figures. We have reviewed all AI-assisted work. Every citation was checked against the full text of the cited paper, and literature suggested by the LLM was included only after the authors had read it. AI-assisted code was reviewed by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The source code for LEVI and LEVI-MD is provided as anonymous supplementary material. The LEVI code contains the energy model, the training script, the configuration file with the hyperparameters of Table 2, the trained checkpoint evaluated in Section 4.2, a script that exports the trained network to TorchScript, and a random subset of the training data (1,000 training and 100 validation frames) for running the training pipeline end to end. It also contains the scripts that generate the alanine dipeptide training datasets and reference conformational ensembles from the vacuum and explicit-solvent MD simulations. The LEVI-MD code contains the OpenMM plugins that combine the learned potential with other energy terms and the connectivity restraint force inside the integrator, the simulation script, the configuration files of the vacuum and explicit-solvent runs with the parameters of Table 3, and the analysis script that applies the structure filter and chirality correction and computes the (ϕ, ψ) distributions. One training step and the inference procedure are given in Algorithms 1 and 2. Section 4.2 describes how the alanine dipeptide datasets were generated and how the dihedral densities and basin populations are computed, and Appendices A.2 and A.3 report the compute used for training and inference. Training details for the Langevin systems of Figure 1 are given in Appendix A.1. A Jupyter notebook for training a LEVI model on Langevin data and reproducing the figures herein is also attached as supplementary material.

REFERENCES

Marloes Arts, Victor Garcia Satorras, Chin-Wei Huang, Daniel Zugner, Marco Federici, Cecilia Clementi, Frank Noé, Robert Pinsler, and Rianne van den Berg. Two for one: Diffusion models and force fields for coarse-grained molecular dynamics. *Journal of Chemical Theory and Computation*, 19(18):6151–6159, 2023.

Michal Balcerak, Tamaz Amiranashvili, Antonio Terpin, Suprosanna Shit, Lea Bogensperger, Sebastian Kaltenbach, Petros Koumoutsakos, and Bjoern Menze. Energy matching: Unifying flow matching and energy-based models for generative modeling. *Advances in Neural Information Processing Systems*, 38:8583–8609, 2026.

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 and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14):146401, 2007.

Filippo Bigi, Marcel Langer, and Michele Ceriotti. The dark side of the forces: assessing non-conservative force models for atomistic machine learning. *arXiv preprint arXiv:2412.11569*, 2024.

Chen-Hao Chao, Wei-Fang Sun, Bo-Wun Cheng, and Chun-Yi Lee. On investigating the conservative property of score-based generative models. *arXiv preprint arXiv:2209.12753*, 2022.

Dmitriy S Chekmarev, Tateki Ishida, and Ronald M Levy. Long-time conformational transitions of alanine dipeptide in aqueous solution: Continuous and discrete-state kinetic models. *The Journal of Physical Chemistry B*, 108(50):19487–19495, 2004.

Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019.

Peter Eastman, Jason Swails, John D Chodera, Robert T McGibbon, Yutong Zhao, Kyle A Beauchamp, Lee-Ping Wang, Andrew C Simmonett, Matthew P Harrigan, Chaya D Stern, et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7):e1005659, 2017.

Matthias Fey, Jinu Sunil, Akihiro Nitta, Rishi Puri, Manan Shah, Blaž Stojanovič, Ramona Bendias, Alexandria Barghi, Vid Kocijan, Zecheng Zhang, et al. Pyg 2.0: Scalable learning on real world graphs. *arXiv preprint arXiv:2507.16991*, 2025.

Francisco F Garcia-Prieto, Ignacio Fdez Galván, Manuel A Aguilar, and M Elena Martín. Study on the conformational equilibrium of the alanine dipeptide in water solution by using the averaged solvent electrostatic potential from molecular dynamics methodology. *The Journal of chemical physics*, 135(19), 2011.

Paraskevi Gkeka, Gabriel Stoltz, Amir Barati Farimani, Zineb Belkacemi, Michele Ceriotti, John D Chodera, Aaron R Dinner, Andrew L Ferguson, Jean-Bernard Maillet, Hervé Minoux, et al. Machine learning force fields and coarse-grained variables in molecular dynamics: application to materials and biological systems. *Journal of chemical theory and computation*, 16(8):4757–4775, 2020.

Christoph Griesbacher, Lea Bogensperger, Andreas Habring, and Thomas Pock. Generating physically consistent molecules with energy-based models. *arXiv preprint arXiv:2605.18381*, 2026.

Stefaan Simon Pierre Hessmann, Khaled Kahouli, Stefan Gugler, Michael Plainer, Frank Noé, Klaus-Robert Müller, and Niklas Wolf Andreas Gebauer. Generative pseudo-force fields for molecular generation. *arXiv preprint arXiv:2605.19050*, 2026.

Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L De Groot, Helmut Grubmüller, and Alexander D MacKerell Jr. Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nature methods*, 14(1):71–73, 2017.

William L Jorgensen and Julian Tirado-Rives. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proceedings of the National Academy of Sciences*, 102(19):6665–6670, 2005.

Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: exact likelihood generative learning for symmetric densities. In *International conference on machine learning*, pp. 5361–5370. PMLR, 2020.

Helen Lai, Tingyu Wang, Hassan Sirelkhatim, Joe Eaton, Howard Huang, Brad Rees, Ola Engkvist, Jon Paul Janet, Xiaoyun Wang, and Alessandro Tibo. Improving protein-ligand complex generation with force field guidance. *Journal of Cheminformatics*, 18(1):55, 2026.

Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8(31.10):5281, 2013.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Alexander D MacKerell Jr. Empirical force fields for biological macromolecules: overview and issues. *Journal of computational chemistry*, 25(13):1584–1604, 2004.

Dominik Marx and Jurg Hutter. Ab initio molecular dynamics: Theory and implementation. *Modern methods and algorithms of quantum chemistry*, 1(301-449):141, 2000.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

Vaclav Parchansky, Josef Kapitan, Jakub Kaminsky, Jaroslav Sebestik, and Petr Bour. Ramachandran plot for alanine dipeptide as determined from raman optical activity. *The Journal of Physical Chemistry Letters*, 4(16):2763–2768, 2013.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Michael Plainer, Hao Wu, Leon Klein, Stephan Günnemann, and Frank Noé. Consistent sampling and simulation: Molecular dynamics with energy-based diffusion models. *Advances in Neural Information Processing Systems*, 38:24460–24505, 2026.

Jaime Rubio-Martinez, M Santos Tomas, and Juan J Perez. Effect of the solvent on the conformational behavior of the alanine dipeptide deduced from md simulations. *Journal of Molecular Graphics and Modelling*, 78:118–128, 2017.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Oliver T Unke, Stefan Chmiela, Huziel E Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T Schutt, Alexandre Tkatchenko, and Klaus-Robert Muller. Machine learning force fields. *Chemical reviews*, 121(16):10142–10186, 2021.

Andre Brasil Vieira Wyzykowski, Fatemeh Fathi Niazi, and Alex Dickson. Agdiff: attention-enhanced diffusion for molecular geometry prediction. *Journal of Chemical Information and Modeling*, 65(4):1798–1811, 2025.

Han Wang, Christof Schutte, Giovanni Ciccotti, and Luigi Delle Site. Exploring the conformational dynamics of alanine dipeptide in solution subjected to an external electric field: A nonequilibrium molecular dynamics simulation. *Journal of Chemical Theory and Computation*, 10(4):1376–1386, 2014.

Runqian Wang and Yilun Du. Equilibrium matching: Generative modeling with implicit energy-based models. *arXiv preprint arXiv:2510.02300*, 2025.

Christoph F Weise and James C Weisshaar. Conformational analysis of alanine dipeptide from dipolar couplings in a water-based liquid crystal. *The Journal of Physical Chemistry B*, 107(14): 3265–3277, 2003.

Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

A HYPERPARAMETERS AND EXPERIMENT DETAILS

A.1 LEVI TRAINING ON LANGEVIN SYSTEM

Langevin data. 300 independent overdamped Langevin trajectories of 4.2×10^5 steps at $\Delta t = 10^{-3}$ (diffusion constant $\mu = 1$), 2×10^4-step burn-in discarded, configurations retained every 2×10^3 steps, giving 6×10^4 training configurations (200 per trajectory); 50 further trajectories held out for validation, generated the same way.

Implementation. For the Langevin system, E_θ is a 4-hidden-layer MLP of width 64 with GELU activations (input dimension 1 for the Langevin experiment, 2 for the off-manifold experiment). Training uses Adam (learning rate 10^{-3}) with `ReduceLROnPlateau` (factor 0.5, patience 6 evaluation intervals), batch size 4096, for 2×10^4 steps. The complete set of parameters for the Langevin model is given in Table 1.

Table 1: Langevin system hyperparameters.

PARAMETER	VALUE
Network	
Architecture	MLP, GELU activations
Hidden layers	4
Hidden width	64
Input dimension	1 (Langevin) / 2 (off-manifold)
Output	scalar E_θ, plus learned s, b (Eq. 7)
Total parameters	12,675 (Langevin) / 12,739 (off-manifold)
Forward process	
β schedule	Sigmoid
β_0 / β_{T-1}	10^{-7} / 2×10^{-3}
Diffusion steps T	5000
Timestep sampling	$\mathrm{Unif}\{0, \ldots, T-1\}$
Energy regression	
γ_E	$\{0, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0\}$
Regression target	$V(x)$ (Langevin) / $V_{2d}(x, y) = V(x)$ (off-manifold)
Optimization	
Optimizer	Adam
Learning rate	10^{-3}
LR schedule	Plateau (factor 0.5, patience 6 eval. intervals)
Batch size	4096
Training steps	5×10^4

A.2 LEVI TRAINING ON ALANINE DIPEPTIDE

Implementation. Our implementation uses PyTorch (Paszke et al., 2019) and PyTorch Geometric (Fey et al., 2025), with the SchNet-based backbone of AGDIFF (Vieira Wyzykowski et al., 2025). The model hyperparameters are shown in detail in Table 2.

Compute. The alanine dipeptide model was trained for 50 hours on a single NVIDIA H200. Validation over the 50,000 held-out frames runs every 1000 steps and accounts for 13 of those hours.

A.3 LEVI-MD INFERENCE ON ALANINE DIPEPTIDE

Implementation. The trained network is exported with TorchScript and evaluated inside OpenMM (Eastman et al., 2017) through a custom force plugin implemented by LibTorch, which returns $-\nabla_{\mathbf{x}} U_\theta^{\mathrm{sol}}$ to the integrator alongside the classical forces at every step. The solute is introduced

Table 2: LEVI hyperparameters for alanine dipeptide.

PARAMETER	VALUE
Network	
SchNet interaction blocks	6
Interaction cutoff	10.0 Å
Embedding dimension	128
Energy MLP hidden dimension	128
Energy MLP layers (atom / bond / angle / dihedral)	5 / 2 / 2 / 2
Total parameters	1.40M
Training	
Forward process	
β schedule	Sigmoid
β_{start} / β_{end}	1e-7 / 2e-3
Diffusion steps T	5000
Timestep sampling	$\mathcal{U}\{0, T-1\}$
Energy regression	
Energy loss weight γ_E	1.0
Regression target	CHARMM36m potential energy (kJ mol^{-1})
Optimization	
Optimizer	Adam
Adam β_1 / β_2	0.95 / 0.999
Learning rate	5e-4
LR schedule	Plateau (factor 0.6, patience 10)
Batch size	2500 atoms
Training steps	417,000
Validation interval	1000 steps

as a set of free particles at random positions carrying the fixed attributes $\mathbf{a}$, so the environment and interaction terms of Eq. 9 are ordinary OpenMM forces built from the same $(q_i, \epsilon_i, \sigma_i)$ that condition E_θ. A flat-bottomed centroid bond force is used to keep the particles centered in the middle of the simulation box. If the centroid of the particles exceeds $d_0 = 0.9$ nm from the center, all particles experience a harmonic restoring force with force constant $k = 5000$ kJ mol^{-1} nm^{-2}.

Coupling schedule. Each simulation begins from random solute coordinates and, following an energy minimization step, introduces U_θ^{sol} over eight stages of increasing coupling λ. Only the final stage, at $\lambda = 1$, is sampled. The solvated runs carry two auxiliary terms that help the solute assemble against the water: a repulsive solute–solute nonbonded term scaled by $1 - \lambda$, which is zero in the sampled stage, and a flat-bottomed restraint holding the atoms within 2.5 Å of a single connected group. The vacuum runs use neither.

Compute. With the parameters in Table 3, a single LEVI-MD simulation for alanine dipeptide takes approximately one hour on one NVIDIA V100S, in vacuum and in explicit solvent alike. The vacuum runs do not require the staged introduction of U_θ^{sol}, since there is no environment to assemble the solute against; the same schedule is used so that the two settings share identical dynamics parameters.

Analysis. Each simulation contributes only its final configuration, so the reported ensembles are uncorrelated by construction. We perceive a bond graph from the generated coordinates and atom types with RDKit (Landrum et al., 2013) and require its canonical SMILES to match that of the reference alanine dipeptide, which every structure satisfies in both environments. Alanine dipeptide has a single chiral center and the attributes $\mathbf{a}$ carry no stereochemistry, so the two enantiomers are

Table 3: LEVI-MD inference parameters for alanine dipeptide. All parameters are shared between the vacuum and solvent runs except those listed under *Solvent only*. Force constants are in kJ $\text{mol}^{-1}\text{nm}^{-2}$.

PARAMETER	VALUE
Solute	
Atoms	22
Particle mass	20 amu
Initialization	Random, 0.9 nm, min. separation 0.2 nm
Centroid bond force	$d_0 = 0.9$ nm, $k = 5 \times 10^3$ kJ mol^{-1} nm^{-2}
Coupling schedule	
Stages	8
Coupling λ	0, 0.001, 0.01, 0.1, 0.2, 0.5, 0.8, 1.0
Steps per stage	10,000 (20 ps)
Total per simulation	80,000 steps (160 ps)
Minimization before each stage	Tolerance 100, max. 1000 iterations
Integration	
Integrator	Langevin
Temperature	300 K
Friction ζ	10 ps^{-1}
Timestep Δt	2 fs
Solvent only	
Water model	TIP3P
Box	2.92 nm cubic, periodic
Solute–solute nonbonded	Repulsive, $\propto 1 - \lambda$ (zero at $\lambda = 1$)
Connectivity restraint	0.25 nm, 10^4

generated with equal probability; every R structure is reflected onto its S enantiomer before (ϕ, ψ) is computed, since mirror images carry opposite dihedral signs.

B LANGEVIN SYSTEM COMPOSITION WITH BIASING FORCES

To demonstrate composability for the Langevin system, we create two biasing potentials that are added at inference time: a linear bias that increases the relative probability of the $x = -1$ basin, and a harmonic potential centered at $x = 0$ (Figure 3). The linear bias is of the form: $E_{\text{bias}}^{\text{lin}}(x) = 2.0x$. The harmonic bias is of the form: $E_{\text{bias}}^{\text{harm}}(x) = 6.0x^2$. Probability distributions were computed with U_θ trained using $\gamma_E = 1.0$ summed with E_{bias}, using a set of 6000 trajectories, 20000 steps each. Excellent agreement is observed in each case.

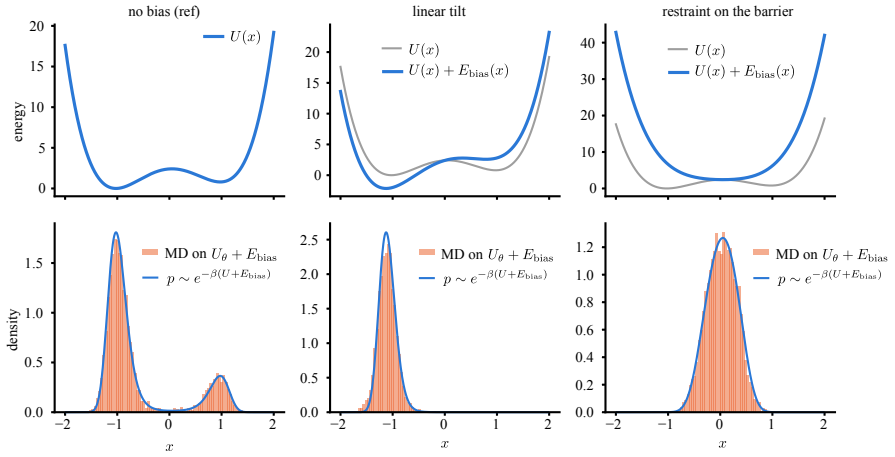

Figure 3: **Demonstration of composability for 1D Langevin system. Top row:** Energy curves for the unbiased potential (left), linearly tilted potential (middle) and a harmonic restraint potential (right). **Bottom row:** Histrograms of MD trajectories run on the unbiased (left) or composed potentials (middle and right), compared with their respective Boltzmann probability distributions (blue lines).

C GEOMETRY OF THE LEVI-MD ENSEMBLES

C.1 BOND LENGTHS OF THE LEVI-MD ENSEMBLES

We compare the four backbone bond lengths of the vacuum and solvated LEVI-MD ensembles with a reference from vacuum MD, the 200,000 frames on which LEVI was trained. Both ensembles are compared against this reference; the solvated MD bond lengths differ from it by at most 0.005 Å, so the two references are interchangeable here. Each LEVI-MD ensemble holds the final configuration of each independent simulation, one frame per simulation. The LEVI-MD standard deviations run from 0.062 to 0.069 Å in vacuum and 0.062 to 0.068 Å in water, against a reference range of 0.028 to 0.034 Å. The vacuum means agree with the reference to within 0.01 Å on all four bonds. The solvated means are shorter by 0.01 to 0.03 Å.

C.2 INTRAMOLECULAR DISTANCES

Figure 5 and Table 5 compare three distances in the molecule: $\mathrm{C}\cdots\mathrm{C}_\alpha$, $\mathrm{N}\cdots\mathrm{N}$, and $\mathrm{CH}_3\cdots\mathrm{CH}_3$, the two terminal methyl carbons. In vacuum the three means agree with the reference to 0.001, 0.018 and 0.033 Å, and the two states of the $\mathrm{CH}_3\cdots\mathrm{CH}_3$ distance are both reproduced. Putting the same potential in a water box moves all three distances in the direction as the reference, by $+0.018$, -0.185 and -0.355 Å against $+0.016$, -0.028 and -0.207 Å. LEVI-MD was trained only on the vacuum trajectory and saw none of these solvated conformations, so the response comes from the composition with the environment.

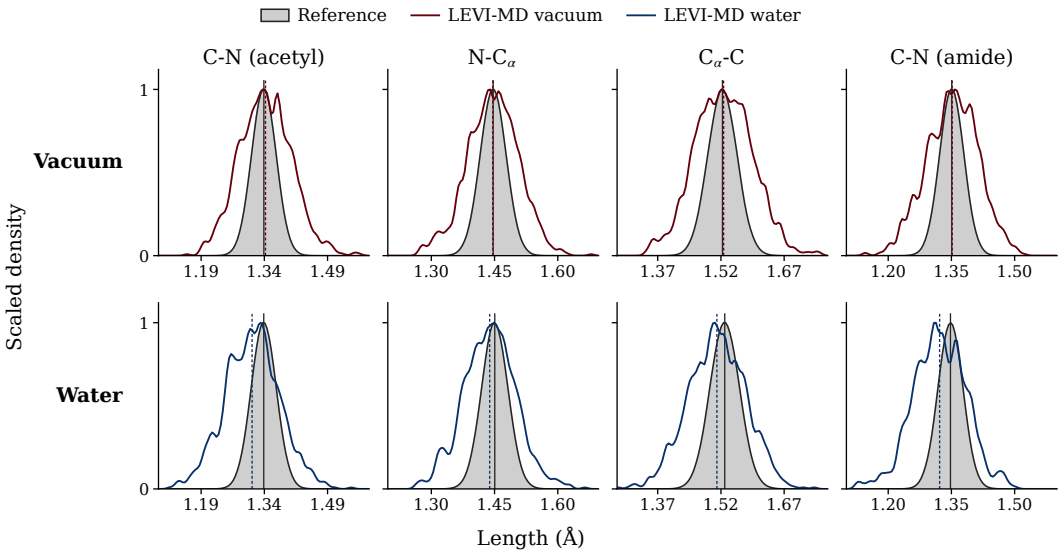

Figure 4: Backbone bond lengths of the LEVI-MD ensembles and of their MD references, in vacuum (top) and in explicit solvent (bottom). Each column shares its window between the two rows, which spans 0.50 Å centred on the reference vacuum mean. Vertical rules mark the reference and LEVI-MD means. All curves are scaled to unit peak rather than unit area.

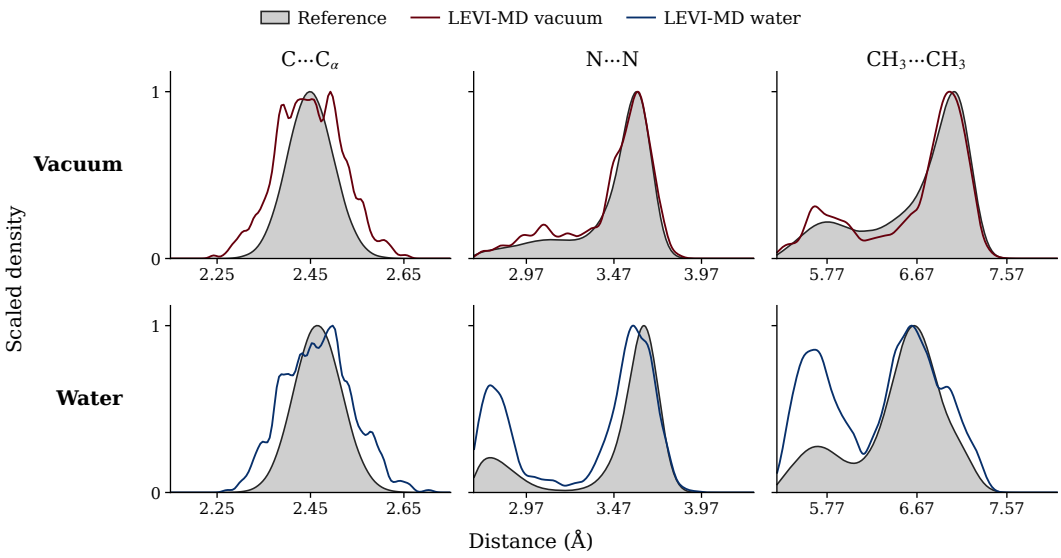

Figure 5: Intramolecular distances of the LEVI-MD ensembles, in vacuum (top) and in explicit solvent (bottom). Each row is shown against its own reference, and each column shares its window between the two rows. All curves are scaled to unit peak rather than unit area.

Table 4: Backbone bond lengths of the LEVI-MD ensembles against the reference vacuum MD, in Å.

BOND	μ_{Ref}	$\mu_{\text{LEVI-MD (vac.)}}$	$\mu_{\text{LEVI-MD (wat.)}}$	σ_{Ref}	$\sigma_{\text{LEVI-MD (vac.)}}$	$\sigma_{\text{LEVI-MD (wat.)}}$
C–N (acetyl)	1.34	1.34	1.31	0.028	0.062	0.066
N–C_α	1.45	1.45	1.44	0.030	0.062	0.062
C_α–C	1.52	1.53	1.51	0.034	0.069	0.068
C–N (amide)	1.35	1.35	1.32	0.028	0.062	0.063

Table 5: Intramolecular distances of the LEVI-MD ensembles against the MD references, in Å. The reference is the vacuum MD in the upper block and the solvated MD in the lower.

DISTANCE	μ_{Ref}	$\mu_{\text{LEVI-MD}}$	σ_{Ref}	$\sigma_{\text{LEVI-MD}}$
Vacuum				
$C\cdots C_\alpha$	2.450	2.449	0.049	0.071
$N\cdots N$	3.469	3.451	0.242	0.254
$CH_3\cdots CH_3$	6.668	6.635	0.524	0.551
Water				
$C\cdots C_\alpha$	2.466	2.467	0.051	0.074
$N\cdots N$	3.441	3.266	0.355	0.410
$CH_3\cdots CH_3$	6.461	6.280	0.481	0.576