PAPER • OPEN ACCESS

# Molecular relaxation by reverse diffusion with time step prediction

IOP Publishing Mach. Learn.: Sci. Technol. 5 (2024) 035038 https://doi.org/10.1088/2632-2153/ad652c

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
9 April 2024

REVISED
13 June 2024

ACCEPTED FOR PUBLICATION
18 July 2024

PUBLISHED
6 August 2024



PAPER

# Molecular relaxation by reverse diffusion with time step prediction

**Khaled Kahouli**[1,2,*], **Stefaan Simon Pierre Hessmann**[1,2], **Klaus-Robert Müller**[1,2,3,4], **Shinichi Nakajima**[1,2,5], **Stefan Gugler**[1,2,7] and **Niklas Wolf Andreas Gebauer**[1,2,6,7]

[1] Machine Learning Group, Technische Universität Berlin, Berlin, Germany
[2] BIFOLD—Berlin Institute for the Foundations of Learning and Data, Berlin, Germany
[3] Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea
[4] Max-Planck Institute for Informatics, Saarbrücken, Germany
[5] RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
[6] BASLEARN—TU Berlin/BASF Joint Lab for Machine Learning, Technische Universität Berlin, Berlin, Germany
[7] Equal contribution.
[*] Author to whom any correspondence should be addressed.

**E-mail:** khaled.kahouli@tu-berlin.de, stefan.gugler@tu-berlin.de and n.gebauer@tu-berlin.de

**Keywords:** geometry optimization, molecular relaxation, diffusion time prediction, diffusion models, generative modeling

Supplementary material for this article is available online

## Abstract

Molecular relaxation, finding the equilibrium state of a non-equilibrium structure, is an essential component of computational chemistry to understand reactivity. Classical force field (FF) methods often rely on insufficient local energy minimization, while neural network FF models require large labeled datasets encompassing both equilibrium and non-equilibrium structures. As a remedy, we propose MoreRed, molecular relaxation by reverse diffusion, a conceptually novel and purely statistical approach where non-equilibrium structures are treated as noisy instances of their corresponding equilibrium states. To enable the denoising of arbitrarily noisy inputs via a generative diffusion model, we further introduce a novel diffusion time step predictor. Notably, MoreRed learns a simpler pseudo potential energy surface (PES) instead of the complex physical PES. It is trained on a significantly smaller, and thus computationally cheaper, dataset consisting of solely unlabeled equilibrium structures, avoiding the computation of non-equilibrium structures altogether. We compare MoreRed to classical FFs, equivariant neural network FFs trained on a large dataset of equilibrium and non-equilibrium data, as well as a semi-empirical tight-binding model. To assess this quantitatively, we evaluate the root-mean-square deviation between the found equilibrium structures and the reference equilibrium structures as well as their energies.

## 1. Introduction

Geometry optimization is crucial for understanding reactivity in computational chemistry [1], as it allows for the study of chemical reaction networks [2–14], which are fundamental in catalysis [15–19], combustion [20, 21], polymerization [22], or atmospheric chemistry [23]. Reactivity is governed by activation energy barriers connecting two equilibrium structures via a transition state, and are necessary for microkinetic modeling [24–29]. Moreover, for generative [30], or enumerative [31, 32] explorations, e.g. in drug, battery, or catalyst design [33–38], a common approach is to use a computationally cheap method to generate a dataset, followed by a geometry optimization to obtain equilibrium structures for which most physical properties are defined. Equilibrium structures represent local minima on the Born–Oppenheimer potential energy surface (PES) [39, 40] and are identified by molecular relaxation, that is solving the electronic Schrödinger equation while varying nuclear coordinates by iteratively following the negative gradients of the energy, i.e. the forces, until they converge to zero [1, 41, 42].

Because iterative *ab initio* electronic structure calculations are computationally expensive and not feasible in high-throughput settings, methods for finding equilibrium structures need to be efficient [41, 42]. To address this limitation, numerous approaches have been developed that speed up the computation of forces

but also suffer from a loss in accuracy compared to *ab initio* methods. These include classical force field (FF) methods like MMFF94 [43], the universal FF [44], or CHARMM [45], on the one hand, and semiempirical methods such as GFN2-xTB [46], PM6-7 [47, 48], or OM2 [49] on the other hand. Furthermore, machine learning FF (MLFF) models [50–65] have emerged as promising alternatives to physical models. When trained on a sufficient amount of *ab initio* calculations, MLFF models, using kernel methods such as sGDML [66–68] or neural networks such as SchNet [69, 70], learn the physical PES and very efficiently predict the forces. Moreover, they have been shown to produce promising results in relaxation tasks [71, 72], while reducing computational cost by several magnitudes. Although MLFF models significantly accelerate gradient computations for structural relaxation, the training dataset must cover a wide range of the chemical space including equilibrium and non-equilibrium structures with accurately computed physical labels, introducing a large computational cost for generating a training dataset.

An emerging machine learning-based approach to exploring chemical space is training generative models on a dataset of equilibrium structures in order to learn to generate new molecular structures. For instance, diffusion models have been used recently in molecule generation [73–77], conformer search [78] and molecular graph generation [79, 80]. They generate samples via iterative denoising, starting from a simple prior distribution like isotropic Gaussian noise. Several other generative models exist for 3D molecular structures, but they are generally not designed for denoising or generation from arbitrary states. Typically, they generate equilibrium structures from scratch by iteratively adding new atoms [81–85] or by transforming samples from a prior distribution to a target distribution in one shot [86–89]. Furthermore, generative models have been used to sample conformations given molecular graphs as input [90–96]. A common drawback of conventional generative methods is that, unlike relaxation-based methods, equilibrium structures are generated from scratch, which makes it difficult to steer the generation towards desired structures. We define denoising as generating samples from the data manifold given arbitrarily noisy inputs. This task is fundamentally different from previous work [97–101] based on the idea of denoising autoencoders [102], where different denoising, yet not generative, techniques are used on molecular structures as an auxiliary or pre-training task for the original regression task of predicting forces and other molecular properties, aiming to improve data efficiency, generalizability and robustness. Relative to our approach, Hsu *et al* [103] use a score-based model to eliminate thermal noise or perturbations in the atomic positions of condensed materials, dealing with relatively small noise magnitudes.

In this work, we propose a conceptually novel statistical approach to molecular relaxation through reverse diffusion, which we will call MoreRed. The distortion in a non-equilibrium input structure is interpreted as a noise level, as the structure has 'diffused away' from its equilibrium state. In this setting, the molecular relaxation can be modeled as a denoising process which can be achieved by reverse diffusion. In contrast to MLFFs, MoreRed does not learn the physical PES, but a simple pseudo PES that emerges from removing Gaussian noise from distorted structures (see figure 1). This offers a significant advantage over the MLFF models: Training MoreRed requires *only* equilibrium structures *without* labels for physical properties such as energy and forces, which considerably reduces the computational costs of generating training datasets. Therefore, it potentially expands the applicability of ML-based relaxation to domains where MLFFs cannot be trained because only equilibrium structures are reported.

A key technical novelty of MoreRed is the *diffusion time step predictor*. Existing diffusion models require the time step as an input that indicates how noisy the input is. However, in molecular relaxation the noise level, i.e. how far away from the equilibrium a structure lies, is unknown. Therefore, in MoreRed we predict the appropriate time step, enabling us to denoise distorted molecular structures at arbitrary noise levels. To this end, we provide a theoretical argument for why accurate time step prediction via a neural network is possible. In contrast to existing diffusion models, which follow a fixed pre-defined time step schedule, this allows furthermore for an adaptive schedule, where the time step is dynamically increased or decreased during denoising depending on the detected noise level, potentially correcting errors in the denoising process. We demonstrate the advantage of our adaptive schedule over classical, fixed sampling.

We test the performance of our method on QM7-X [104], a dataset containing 42 000 equilibrium structures found with third-order self-consistent charge density functional tight binding [105] (DFTB3) [106–108] and many-body dispersion (MBD) [109, 110] corrections. They cover all molecular graphs in GDB13 [111] with up to 7 heavy atoms and include the elements H, C, N, O, S, and Cl. For each equilibrium structure, 100 non-equilibrium structures generated via normal-mode displacements of the equilibrium geometry are reported, including DFT calculations for energies and forces at the PBE0+MBD level [109, 112, 113] with `FHI-aims` [114, 115]. In our experiments, we employ several baselines for comparison, including the classical FF MMFF94, the semiempirical method GFN2-xTB, as well as a MLFF model with an equivalent neural network backbone architecture, and show that MoreRed performs favourably. Specifically, MoreRed accurately maps non-equilibrium structures back to the data manifold of equilibrium structures that it has been trained on. This is despite being trained on two orders of magnitude fewer structures than the MLFF.

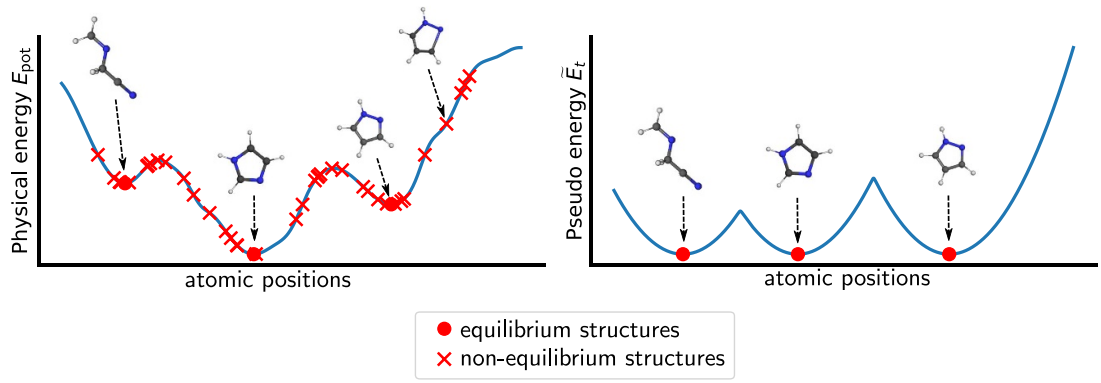


**Figure 1.** Schematic depictions of the physical PES (left) and the pseudo PES (right), which emerge from removing Gaussian noise from distorted structures. Learning the physical energy requires equilibrium (circles) and many non-equilibrium (crosses) training structures, while learning the pseudo PES only requires equilibrium training structures.

Additionally, while MLFFs can only relax structures that are covered by the distribution of training data, the inherent augmentation of the training data through the diffusion process enhances the robustness of MoreRed against variations in the noise distribution of the non-equilibrium test structures. Consequently, MoreRed successfully identifies the correct equilibrium structures for non-equilibrium inputs where MLFFs fail. We also show that the difference in DFT energies between the reference equilibrium structures in QM7-X and the structures obtained by MoreRed through molecular relaxation falls below the threshold of chemical accuracy.

## 2. Theory and methods

### 2.1. Diffusion models

Introduced by Sohl–Dickstein *et al* [116], diffusion models are latent-variable generative models that can efficiently generate samples of a complex data distribution $q_{\text{data}}(\mathbf{x}_0)$, where direct sampling is intractable, such as the distribution of equilibrium molecular structures. Instead of directly sampling from the target distribution, the idea is to obtain an initial sample $\mathbf{x}_T$ from a simple prior distribution $q_T(\mathbf{x}_T)$, often an isotropic Gaussian, and then use a learned mapping $h(\cdot)$ to transform $\mathbf{x}_T$ into a sample $\mathbf{x}_0 = h(\mathbf{x}_T)$ within $q_{\text{data}}(\mathbf{x}_0)$.

While defining and learning $h(\cdot)$ is challenging, the opposite task of transforming the complex distribution $q_{\text{data}}(\mathbf{x}_0)$ into a simple distribution $q_T(\mathbf{x}_T)$ is manageable because it only involves simplifying the data by diminishing its signal, for instance by iteratively adding noise. Diffusion models leverage this concept by using two opposite processes (see figure 2). A fixed, usually non-learned, forward diffusion process iteratively encodes $q_{\text{data}}(\mathbf{x}_0)$ into a tractable latent distribution $q_T(\mathbf{x}_T)$ over $T$ steps. A backward or reverse process parametrized by a machine learning model then learns to reverse the forward diffusion to effectively map from $q_T(\mathbf{x}_T)$ back to $q_{\text{data}}(\mathbf{x}_0)$, akin to the objective of the mapping $h(\cdot)$.

Summarized in figure 2, we focus on denoising diffusion probabilistic models (DDPM) [117] as a special class of diffusion models that defines the forward diffusion process as the fixed Markov process,

$$q(\mathbf{x}_{0:T}) = q_{\text{data}}(\mathbf{x}_0)\prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \quad q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right), \tag{1}$$

where $q_0(\mathbf{x}_0) = q_{\text{data}}(\mathbf{x}_0)$ and $\mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$ denotes that $\mathbf{x}_t$ follows an isotropic Gaussian with mean $\sqrt{1-\beta_t}\mathbf{x}_{t-1}$ and variance $\beta_t\mathbf{I}$, with the identity matrix $\mathbf{I}$. The diffusion process from equation (1) can be simulated by sampling $\mathbf{x}_0 \sim q_{\text{data}}(\mathbf{x}_0)$ from the training dataset representing the equilibrium positions of atoms and then iteratively applying $\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\varepsilon}_t$ for $t = 1, 2, \ldots, T$, with Gaussian noise $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The variance $\beta_t$ follows a fixed monotonically increasing noise schedule $\{\beta_t \in (0,1)\}_{t=1}^{T}$, thus progressively injecting Gaussian noise with variance $\beta_t$ into the atom positions while diminishing the signal with the factor $\sqrt{1-\beta_t}$. This results in the generation of increasingly noisier molecular structures, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$, from the original sample $\mathbf{x}_0$ with increasing diffusion time step $t$ (see figure 2, forward Gaussian diffusion process, from left to right). At the endpoint $t = T$, the process destroys all the signal in the sample, converging to pure Gaussian noise, i.e. $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

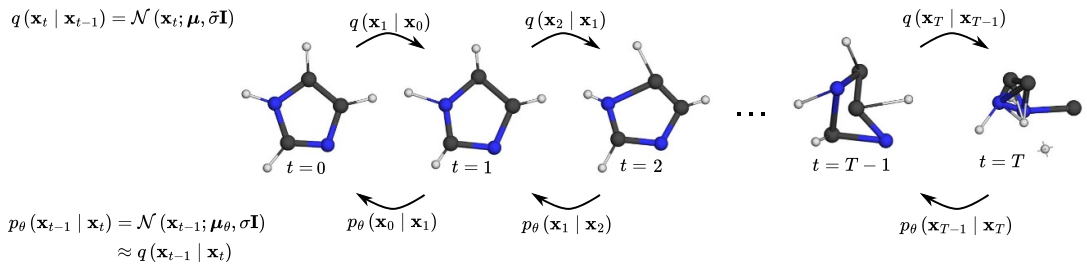


**Figure 2.** The diffusion model applied to the molecular structure of imidazole. It incorporates two stochastic processes: a forward process and a reverse process. The forward process involves a fixed diffusion kernel $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ to transform the original sample $\mathbf{x}_0 \sim q_{\text{data}}(\mathbf{x}_0)$ into complete noise sample, $\mathbf{x}_T \sim q_T(\mathbf{x}_T)$, usually isotropic Gaussian noise $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The backward process is a learned model with parameters $\theta$, which reverses the forward process, i.e. $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \approx q(\mathbf{x}_{t-1}|\mathbf{x}_t)$. It maps a noise sample $\mathbf{x}_T$ from the tractable prior distribution $p_T(\mathbf{x}_T) = q_T(\mathbf{x}_T)$ to the complex target distribution of equilibrium structures, $q_{\text{data}}(\mathbf{x}_0)$.

When generating progressively noisier samples $\{\mathbf{x}_t\}_{t=1}^T$, the diffusion process creates latent distributions $q_t(\mathbf{x}_t)$ that are increasingly smoother versions of the original data distribution $q_{\text{data}}(\mathbf{x}_0)$, such that $\mathbf{x}_t \sim q_t(\mathbf{x}_t)$. These latent distributions can be derived from equation (1) as $q_t(\mathbf{x}_t) = \int q(\mathbf{x}_t|\mathbf{x}_0) q_{\text{data}}(\mathbf{x}_0) \mathrm{d}\mathbf{x}_0$ for $t \in [1, T]$. The perturbation kernel, $q(\mathbf{x}_t|\mathbf{x}_0)$, has the closed-form solution $\mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$, where $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ and $\alpha_s = 1 - \beta_s$ [117]. With this definition, we can directly sample from the forward diffusion process at any time step $t$ using an equilibrium structure $\mathbf{x}_0$ from the training data:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2}$$

avoiding the iterative simulation through all intermediate steps $\mathbf{x}_0, \ldots, \mathbf{x}_t$.

Reversing the forward diffusion process enables the generation of new samples by mapping back from $q_T(\mathbf{x}_T)$ to $q_{\text{data}}(\mathbf{x}_0)$ using the reverse transition $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$. Given that $T$ is large enough, the reverse transition is also Gaussian [116]. However, unlike the forward process, it is not tractable. Therefore, we need to approximate the reverse process, e.g. by learning a parametrized model $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \approx q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ such that [117]:

$$p_\theta(\mathbf{x}_{0:T}) = p_T(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \quad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}\right) \tag{3}$$

where $p_T(\mathbf{x}_T) = q_T(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the endpoint of the forward process and the starting point (or latent prior) of the reverse process. The variance is $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$ and the mean, $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)\right)$, is the only unknown quantity, where $\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)$ is an estimate of the noise that was added to $\mathbf{x}_0$ to obtain $\mathbf{x}_t$. This noise is predicted by a neural network that gets the noisy structure $\mathbf{x}_t$ and the current time step $t$ as an input. To train this network, we uniformly sample a diffusion time step $t \sim \mathcal{U}(1, T)$, and perform forward diffusion to generate the noisy sample $\mathbf{x}_t$ from a data point $\mathbf{x}_0$ using the sampled noise direction $\boldsymbol{\varepsilon}_t$ as described in equation (2). Then, we minimize the mean squared error between the predicted and true noise direction, resulting in the following loss:

$$L_{\text{DDPM}} = \mathbb{E}_{t\sim\mathcal{U}(1,T), \mathbf{x}_0\sim q_{\text{data}}, \boldsymbol{\varepsilon}_t\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[||\boldsymbol{\varepsilon}_t - \boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)||^2\right], \tag{4}$$

once the noise predictor, $\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)$, is trained, we can generate new samples $\mathbf{x}_0 \sim q_{\text{data}}(\mathbf{x}_0)$ by simulating the reverse process in equation (3). We first draw a starting sample $\mathbf{x}_T \sim p_T(\mathbf{x}_T)$ from the Gaussian noise distribution and then iteratively apply

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t\bar{\boldsymbol{\varepsilon}}, \quad \bar{\boldsymbol{\varepsilon}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{5}$$

for $t = T, T-1, \ldots, 1$, which progressively removes the noise from the sample to denoise it, in the optimal case ending with a sample $\mathbf{x}_0$ from the target $q_{\text{data}}(\mathbf{x}_0)$ after $T$ reverse steps (see figure 2, learned reverse generative process, right to left). This mimics molecular structure optimization by following atomistic forces. Here, the noise prediction $\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)$ defines the opposite of the force directions that minimize the energy and the scaling terms that depend on $\alpha_t$ and $\beta_t$ determine the magnitudes and the step sizes used at each

optimization step. The noise $\bar{\varepsilon}$ added at each denoising step results in a stochastic optimizer instead of a deterministic one, which can be helpful in the case of the presence of many shallow local minima. However, because noise prediction $\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)$ requires the diffusion time step $t$ as an input, diffusion models can only be used on structures with known noise level. For data generation, they start with samples from the known noise distribution at $t = T$, i.e. input structures that are pure noise. For molecular relaxation, in contrast, the non-equilibrium input structure can have an arbitrary level of perturbation such that the suitable initial time step is unknown, making the application of the standard plain diffusion models infeasible.

### 2.2. MoreRed: molecular relaxation by reverse diffusion

We therefore introduce Molecular Relaxation by Reverse Diffusion (MoreRed) as a diffusion-based approach to finding minima on a PES. MoreRed reframes molecular relaxation as a denoising problem solved using a learned reverse diffusion process, where non-equilibrium molecular structures are considered as diffused noisy versions of their equilibrium counterparts. While diffusion models were initially designed to generate novel samples from complete noise $\mathbf{x}_T \sim p_T(\mathbf{x}_T)$, we adapt them to be applicable in this denoising framework, where we initiate the reverse process from a noisy sample $\mathbf{x}_t \sim q_t(\mathbf{x}_t)$ at an arbitrary diffusion time step $t < T$ to reconstruct the nearest $\mathbf{x}_0$. Taking figure 2 as an illustration, the objective is to initiate the reverse process from any step $t$ within the trajectory where the structure of the yet noisy sample $\mathbf{x}_t$ remains identifiable, such as the third noisy structure from the left, and perfectly reconstruct the initial noiseless structure $\mathbf{x}_0$ on the far left. In contrast, starting from the complete noise sample $\mathbf{x}_T$ on the far right would yield different relaxed structures in repeated denoising attempts because no structure is apparent in the input.

Using the setup explained in section 2.1, MoreRed learns the distribution of *equilibrium* molecular structures as the target data distribution $q_{\text{data}}(\mathbf{x}_0)$, and smoothed versions of it as the latent distributions,

$$q_t(\mathbf{x}_t) = \int q(\mathbf{x}_t|\mathbf{x}_0)\, q_{\text{data}}(\mathbf{x}_0)\, \mathrm{d}\mathbf{x}_0, \qquad \text{for } t \in [1, T].$$

This amounts to learning a pseudo PES, $\widetilde{E}_t = -\log q_t(\mathbf{x})$, that depends on $t$ when diffusion models are trained to predict the noise direction [118]. This is conceptually similar to MLFF models, which implicitly learn the PES when trained on forces, where the diffusion noise could intuitively be seen as the opposite of the forces. Yet, as depicted in figure 1, $\widetilde{E}_t$ is much simpler than the physical potential energy $E_{\text{pot}}(\mathbf{x})$ that the existing MLFF models need to learn. Furthermore, MoreRed exhibits superior data efficiency compared to MLFF models, because *it requires only equilibrium structures*. The whole input space of non-equilibrium structures, including physically non-plausible structures, is simply covered by the efficient forward diffusion process that adds Gaussian noise, as explained in section 2.1. In contrast, for MLFF models to be reliable for any possible input structure, a large training dataset with many non-equilibrium structures derived from the physical PES is required, which can be infeasible to generate. However, if trained on extensive labeled data, MLFFs become applicable for molecular dynamics simulations. We note that MoreRed, on the contrary, *only* targets molecular relaxation and is, at this stage, not intended for molecular dynamics simulations in its current form.

We design our diffusion model such that the distribution is invariant with respect to rotations $\mathcal{R}(\cdot)$ and translations $\mathcal{T}(\cdot)$, i.e. $p_\theta(\mathcal{T}(\mathcal{R}(\mathbf{x}_t)) = p_\theta(\mathbf{x}_t)$. To guarantee translational invariance, we center the atomic positions after each forward or reverse step. As proven by Xu *et al* [78], rotational invariance of the marginal distributions $p_\theta(\mathbf{x}_t)$ is achieved by using an invariant prior $p_T(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, and an equivariant transition probability $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, which amounts to using an equivariant noise model $\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)$. Therefore, we adopt the equivariant message passing architecture PaiNN [119], which allows us to directly predict equivariant tensor properties, such as the noise $\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)$, as well as invariant scalar properties.

To perform reverse diffusion starting from a non-equilibrium molecular structure $\tilde{\mathbf{x}}$ at an arbitrary noise level, i.e. not sampled from the prior noise distribution $p_T(\mathbf{x}_T)$, it is necessary to set the initial diffusion time step $t < T$ for the reverse process appropriately. A starting time step that does not match the deviation of the noisy input structure from the data manifold of equilibrium structures would lead to inaccurate predictions of the noise direction and an incorrect number of denoising steps. Consequently, successful relaxation would not be possible. To address this issue, we introduce a time step predictor as a novel extension for diffusion models in the subsequent subsection.

### 2.3. Diffusion time step prediction

To identify the noise level of non-equilibrium molecular structures that we want to relax, we train a neural network $\tau_\Theta(\mathbf{x}_t)$ parametrized by $\Theta$ to predict the diffusion time step by minimizing the following loss:

$$L_{\text{DTP}} = \mathbb{E}_{t\sim\mathcal{U}(1,T),\mathbf{x}_0\sim q_{\text{data}},\boldsymbol{\varepsilon}_t\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left(\tau_\Theta(\mathbf{x}_t) - a(t)\right)^2\right], \text{ where } \mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}_t, \tag{6}$$

$a(t)$ is a monotonic function to scale the output, e.g. $a(t) = t/T$, and $t$ is sampled uniformally between 1 and $T$. Analogous to the noise estimator $\varepsilon_\theta(\mathbf{x}_t, t)$, we again adopt the neural network architecture PaiNN [119] for the model $\tau_\Theta(\mathbf{x}_t)$. However, in this case, we use the scalar features to predict the time step, considering it as an invariant quantity similar to energy. In the following, we provide a theoretical argument on why an accurate prediction of the time step is feasible. Empirical evaluations of the time step prediction performance are found in the results section 3.1.

In equation (2), the latent distribution $q_t$ at time step $t$ is derived by applying isotropic Gaussian noise to the training data points representing equilibrium structures in the input space, denoted as $\mathbf{x}_0^{(i)} \sim q_{\text{data}}(\mathbf{x}_0)$, where $i$ represents the index of the training data points. This process transforms each $\mathbf{x}_0^{(i)}$, which is a Dirac delta function, into a Gaussian distribution, $\mathcal{N}(\mathbf{x}_t^{(i)}; \sqrt{\bar{\alpha}_t}\mathbf{x}_0^{(i)}, (1-\bar{\alpha}_t)\mathbf{I})$. Therefore, considering that the equilibrium structures, $\mathbf{x}_0 \sim q_{\text{data}}(\mathbf{x}_0)$, are isolated from each other in the input space, up to the symmetry operations, $q_t$ essentially forms a mixture of Gaussians with each Gaussian component centered around one of the training equilibrium structures, $\mathbf{x}_0^{(i)}$. Moreover, the variance term $(1-\bar{\alpha}_t)$ of these Gaussian components, which is defined by the diffusion noise schedule, increases monotonically with the diffusion time step $t$. Consequently, predicting the time step $t$ from a sample $\mathbf{x}_t^{(i)} \sim q_t(\mathbf{x}_t)$ amounts to estimating its noise level, $(1-\bar{\alpha}_t)$, or the distance from $\mathbf{x}_0^{(i)}$. This estimation is feasible when the dimension $D$ of the input space is large and the different mixture components do not overlap, due to the following reasons.

Let us transform the Gaussian distribution with variance $(1-\bar{\alpha}_t)$ from the Euclidean to the polar coordinate system. By marginalizing out the polar directions, we can compute the marginal distribution over the (scaled) radius $\tilde{r} = \frac{r}{\sqrt{D(1-\bar{\alpha}_t)}}$ as:

$$p(\tilde{r}) = \frac{D^{D/2}\tilde{r}^{D-1}}{2^{D/2-1}\Gamma(D/2)} \exp\left(-\frac{D\tilde{r}^2}{2}\right), \tag{7}$$

where $\Gamma(\cdot)$ denotes the Gamma function. As discussed in Bishop [120], for large $D$, $p(\tilde{r})$ has a sharp peak at $\hat{r} \approx 1$, as illustrated in the left plot in figure A1 in the supplementary information. This implies that each Gaussian component of $q_t$ at time step $t$ represents a sphere centered at a training sample $\mathbf{x}_0^{(i)}$, i.e. its density, represented by the set of diffused samples $\mathbf{x}_t^{(i)}$, is concentrated in a thin shell at radius $\hat{r} \approx 1$. Therefore, most of the samples $\mathbf{x}_t^{(i)}$ have similar distance from $\mathbf{x}_0^{(i)}$. Accordingly, assuming that the model can learn the data manifold, the distance, which corresponds to the noise level $(1-\bar{\alpha}_t)$, is easy to identify from a single sample $\mathbf{x}_t^{(i)}$, as long as the noise level is small such that the mixture components (spheres) do not overlap. When the noise level increases, the diffused samples $\mathbf{x}_t^{(i)}$ from different training data points $\mathbf{x}_0^{(i)}$ overlap with each other. This overlap makes the estimation of the diffusion time step difficult, as indicated by the right plot in figure A1. Further explanation and discussion are provided in section A.1 in the supplementary information, where we also show empirical evidence together with the derivation of equation (7).

### 2.4. Variants of reverse diffusion

We compare three variants of MoreRed that differ in how they handle the diffusion time step prediction (section 3.1). In the first variant, called *MoreRed initial time prediction* (MoreRed-ITP), only the initial diffusion time step, defining the start of the denoising process, is predicted. Given a non-equilibrium structure $\tilde{\mathbf{x}}$, MoreRed-ITP estimates an appropriate starting time step, $\hat{t} = \tau_\Theta(\tilde{\mathbf{x}})$, sets $\mathbf{x}_{\hat{t}} = \tilde{\mathbf{x}}$, and performs the iterative update described in equation (5) for $t = \hat{t}, \hat{t}-1, \ldots, 0$, instead of starting from $t = T$.

As a second variant, we use a more flexible process where the time step prediction is performed before *every* denoising step instead of only at the start. We call this approach *MoreRed adaptive scheduling* (MoreRed-AS). It iterates through a time-adaptive version of equation (5):

$$\hat{t} = \tau_\Theta(\tilde{\mathbf{x}}), \qquad x_{\hat{t}-1} = \frac{1}{\sqrt{\alpha_{\hat{t}}}}\left(\mathbf{x}_{\hat{t}} - \frac{\beta_{\hat{t}}}{\sqrt{1-\bar{\alpha}_{\hat{t}}}}\varepsilon_\theta(\mathbf{x}_{\hat{t}}, \hat{t})\right) + \sigma_{\hat{t}}\bar{\varepsilon}, \quad \bar{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{8}$$

In contrast to the fixed schedule $t, t-1, \ldots, 0$, which always decreases towards 0 by a one-step decrement, this adaptive approach allows the denoising process to move back and forth in the trajectory. In this way, errors in the noise prediction $\varepsilon_\theta(\mathbf{x}_t, t)$, which lead to unexpected noise levels in the subsequent structure, can be compensated. For instance, if after one denoising step the resulting sample has less noise and converges faster than expected, the prediction $\hat{t}$ will be smaller than $t-1$ to jump more than 1 step towards 0. If, on the other hand, the resulting sample has a higher noise level than expected, the prediction $\hat{t}$ will be higher than in the previous step. Similar to classical molecular relaxation methods, we define a convergence criterion for stopping the adaptive denoising process, i.e. we require time step predictions smaller than a threshold $\hat{t} \leqslant \underline{t}$.

A third variant, *MoreRed joint training* (MoreRed-JT), uses the same adaptive reverse diffusion process as described in equation (8) for MoreRed-AS but differs in the model definition and training. For MoreRed-AS we employ two separate neural networks with separate backbone representations, where one is used to predict the noise $\boldsymbol{\varepsilon}_\theta$ and the other one to predict the time step $\tau_\Theta$. For MoreRed-JT, we use one neural network as a shared backbone representation, and we add two prediction heads on top, one for the noise and one for the time step. This forces the noise and the time step heads to learn a joint molecular representation. We train this joint network by minimizing the joint loss $L_{\text{joint}} = \eta L_{\text{DDPM}} + (1-\eta)L_{\text{DTP}}$, for $\eta \in [0,1]$ defining a trade-off between the two losses and combining equations (4) and (6). In the supplementary information, we provide details for the training in algorithm 1 and for the sampling in algorithm 2 in appendix A, and for the models in section C.1.

## 3. Results and discussion

An integral part of relaxation with MoreRed is the time step predictor. It estimates how far the non-equilibrium input is away from the learned data manifold of equilibrium structures, which determines the appropriate time step for the reverse diffusion process and the number of denoising steps. Therefore, we first evaluate the time step predictor. Subsequently, we validate the relaxation performance of MoreRed with respect to the root-mean-square deviation (RMSD) and DFT energy. We compare the results against those obtained from a baseline MLFF, as well as from the FF model MMFF94 [43] and the semi-empirical model GFN2-xTB [46] in several experiments of molecular relaxation.

### 3.1. Diffusion time step prediction performance

The diffusion time step predictor determines the starting step $\hat{t}$ of the reverse diffusion for molecular relaxation. The further a non-equilibrium structure deviates from its equilibrium structure, the more denoising steps are required, which means that a higher starting step $\hat{t}$ should be predicted. As can be seen in figure 3(a), the predicted starting steps correlate well with the RMSDs between the non-equilibrium and equilibrium test structures from QM7-X. This is a notable insight, as the non-equilibrium structures in QM7-X stem from DFTB normal-mode displacements of equilibrium geometries. All non-equilibrium examples used for training the time predictors, on the other hand, stem from diffusing equilibrium structures with Gaussian noise. Nevertheless, the time step predictors reliably predict $\hat{t} > 0$ for all the 10 000 test structures, highlighting the robustness in identifying non-equilibrium structures even if they do not contain Gaussian noise. Moreover, we observe that a joint model for predicting both time step and noise, as in MoreRed-JT, leads to fewer outliers in the predictions of $\hat{t}$ and, consequently, a higher Pearson correlation coefficient ($\rho = 0.85$; top) than the separately trained time step predictor in MoreRed-ITP/AS ($\rho = 0.72$; bottom).

Usually, diffusion models follow a fixed schedule where the time step $t$ is reduced by one after each denoising step until it reaches $t = 0$. In MoreRed-ITP, we follow such a fixed schedule and only use the time step predictor once to obtain a suitable starting step $\hat{t}$. As described in equation (8) before, the other two variants, MoreRed-AS/-JT, utilize the time step predictor at *every* denoising step to obtain a new time step estimate $\hat{t}$. This results in an adaptive schedule, where the relaxation ends after a variable number of steps, as soon as $\hat{t} = 0$ is predicted. Figure 3(b) shows the merit of this approach, where the time step trajectories during geometry relaxations of three different test structures are plotted in red, orange, and green. When following the fixed schedule (dashed lines), errors can occur and accumulate [118]. If not corrected, they lead to a mismatch between the true noise level in the structure and the time step $t$. Therefore, the relaxation may end before the sample reaches the equilibrium geometry. This can be observed for all three examples in the plot: The predicted but not enacted time step values associated with the fixed schedule (dotted lines) show a high value when the denoising with the fixed schedule ends, as the dashed line reaches $t = 0$. In contrast, the adaptive schedule (solid lines), MoreRed-JT in this case, can account for such errors by adapting $\hat{t}$ at each denoising step. After converging to $\hat{t} = 0$, the relaxed structures are significantly closer to the ground truth equilibrium geometry than those obtained with the fixed schedule, despite starting from comparable initial time steps (see the RMSD values after relaxation in the right-hand side box of figure 3(b). In figure B6 in the appendix B.6.1 we present similar trajectories to figure 3(b) but for three cases where the predicted initial time steps are very different for both methods. It highlights that even when MoreRed-ITP starts from a higher initial step, it is still outperformed by MoreRed-JT.

To conclude, we find that the time step predictor accurately identifies non-equilibrium structures, where larger time steps are predicted if the RMSD from the equilibrium structure is larger. Moreover, employing an adaptive schedule that utilizes the time step predictor to determine the time step at every denoising iteration proves beneficial compared to constantly decreasing the time step at a fixed rate. We provide further experiments in B.6 in the supplementary information, where we show that the time step predictor also

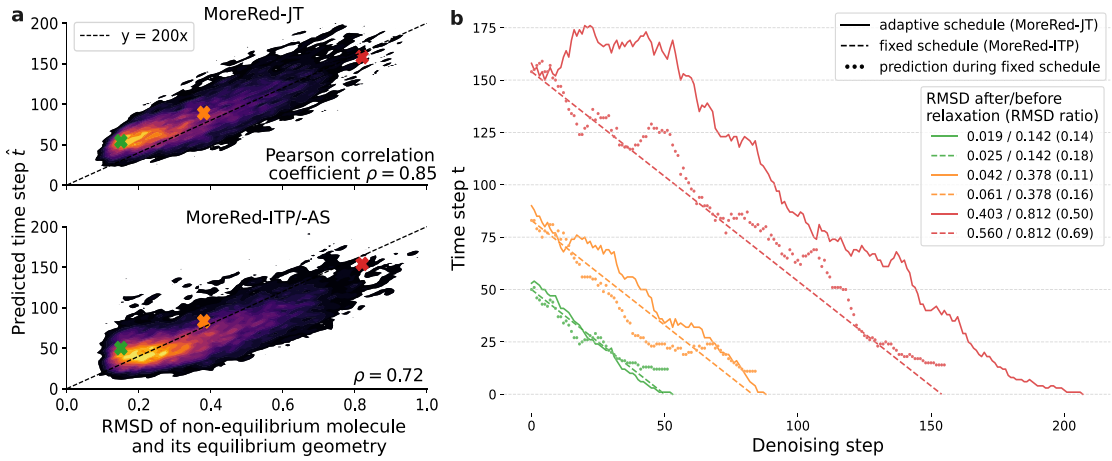


**Figure 3.** (a): Density plots of the RMSD of 10 000 non-equilibrium test structures from QM7-X and their equilibrium structures vs. the initial diffusion time step, $\hat{t}$, predicted by MoreRed-JT (top) and MoreRed-ITP/-AS (bottom). A brighter color (yellow) indicates a higher density of scatter points, a darker color (purple) indicates a lower density. For both, we show the Pearson correlation coefficient, $\rho$. The three crosses (red, orange, green) mark the structures for which a time step trajectory is shown in plot (b) in matching colors. (b): Comparison of time step trajectories for relaxation of three non-equilibrium test structures from QM7-X in red, orange, and green (matching the crosses in plot (a)). MoreRed-JT follows an adaptive schedule (the time step and noise are predicted at each denoising step; solid lines) and MoreRed-ITP follows a fixed schedule (the time step is only predicted at the initial denoising step; dashed lines). The dots show the predicted (but not utilized) time steps for MoreRed-ITP during denoising with the fixed schedule. All two methods start from a comparable initial time step for each non-equilibrium structure. The inset box shows the RMSDs between the reference equilibrium geometry and the non-equilibrium structure after and before relaxation, respectively.

significantly enhances the performance of diffusion models in the original task of novel structure generation from complete noise.

### 3.2. Molecular relaxation performance

In the following, we compare the relaxation performance of MoreRed with a baseline MLFF, as well as the FF method MMFF94, [43] and the semi-empirical method GFN2-xTB [46]. As a molecular representation, all variants of MoreRed as well as the baseline MLFF use the same equivariant message passing neural network architecture PaiNN [119] as implemented in the open-source software package SchNetPack [121, 122]. Details on the models' architectures including the hyperparameters are shown in the supplementary information appendix C. While training MoreRed requires only different unlabeled equilibrium structures, MLFFs have to be trained on non-equilibrium structures as well and require the energies and forces as labels. For this reason, we use QM7-X [104], the only labeled dataset that provides both equilibrium (42 000) as well as the corresponding non-equilibrium structures (100 each) of different chemical compositions, enabling the training of both MoreRed and MLFFs. However, the dataset has a mismatch between the computational methods employed for finding the equilibrium structures (DFTB3 + MBD) and those for computing their energy and force labels (PBE0 + MBD), introducing a challenge in comparing the performance of MoreRed and the MLFF. On one hand, evaluating the geometric deviation, such as RMSD, between a structure after molecular relaxation and the corresponding equilibrium structure reported in the QM7-X is in favour of MoreRed. This is because it is trained to learn the data manifold of these reference structures, which were determined using DFTB3+MDB. On the other hand, comparing the DFT energies of the structures after relaxation and the equilibrium structures in QM7-X using PBE0+MBD reference calculations favours the MLFF because it is trained on the energies and forces resulting from PBE0+MBD calculations. Therefore, we must carefully integrate our findings on both metrics before drawing conclusions.

For our evaluation, we have reserved a test set of 6504 reference equilibrium structures from QM7-X which are not utilized for training the neural networks (see supplementary information appendix B.1 for details). To cover a wide range of test examples, we sort the 100 non-equilibrium structures of each reference structure based on the RMSD to their equilibrium geometry and choose three of them: the closest, one from the middle, and the most distant. This results in almost 20 000 non-equilibrium test inputs, $\tilde{\mathbf{x}}$, for relaxation. Note that whenever we compute the RMSD, the rotation and translation of structures are aligned. For molecular relaxation, we employ Open Babel's [123] built-in routines and optimizer for MMFF94. Additionally, we utilize the L-BFGS optimization algorithm implemented in ASE [124] for both the MLFF and the semi-empirical GFN2-xTB, where we set a convergence threshold of

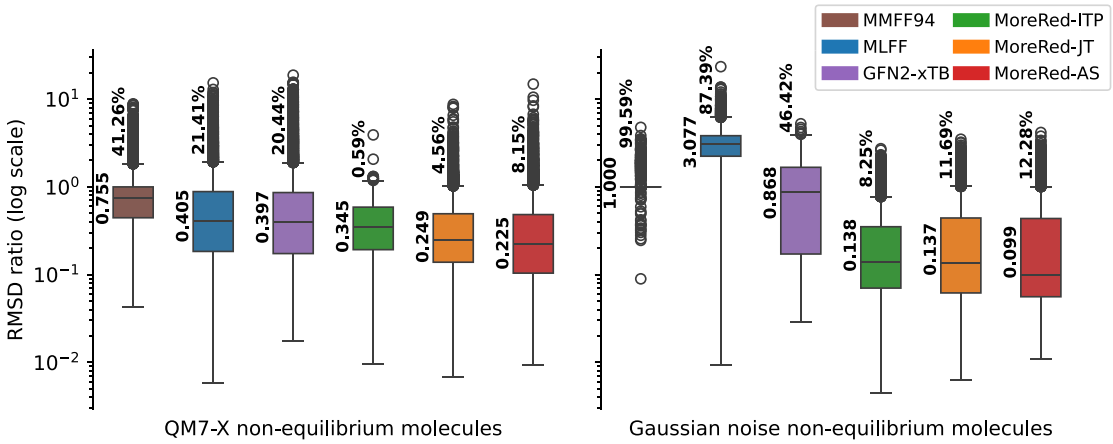


**Figure 4.** The RMSD ratios, i.e. the RMSD after relaxation divided by the RMSD of the non-equilibrium inputs, of structures relaxed with the baselines and the MoreRed variants, trained on the DFTB3+MBD equilibrium structures reported in QM7-X, for 20 000 non-equilibrium structures from the QM7-X test split (left) and 6500 non-equilibrium structures obtained by adding Gaussian noise over 250 forward diffusion steps to equilibrium structures from the QM7-X test split (right). The median values and the percentage of failure cases, i.e. the cases where the RMSD ratio $\geqslant 1$, are shown to the left of each box plot in bold. In figure 5, we provide examples of relaxed structures comparing both MoreRed and the MLFF model.

$f_{\max} = 1.15 \cdot 10^{-3}$ kcal mol$^{-1}$ Å$^{-1}$ for the forces and use the tblite[8] Python interface for the official GFN2-xTB model implementation. Besides, we set a maximum number of relaxation steps $T = 1000$ for all methods, including MoreRed, and a convergence criterion of $\hat{t} \leqslant 0$ for MoreRed-AS/-JT.

We first evaluate the geometric deviation of structures relaxed with MoreRed and the different baselines from the reference equilibrium structures in QM7-X. To this end, we calculate the RMSD ratio, which is the RMSD of the reference structure from the test structure after relaxation divided by the RMSD of the reference structure from the test structure before relaxation. It captures to which extent the non-equilibrium test structure was brought closer to the reference equilibrium structure. We define *failure cases* as cases where the RMSD ratio exceeds 1, which means that the RMSD increased during molecular relaxation. Those failures correspond to cases where the structure diverges or the relaxation converges to a different local minimum in the PES. Figure 4(a) (left) shows boxplots of the RMSD ratio for the MLFF model, the FF method MMFF94, the semi-empirical method GFN2-xTB, as well as MoreRed-ITP, MoreRed-JT, and MoreRed-AS. The lowest median RMSD ratios and the lowest percentages of failure cases are all achieved by MoreRed. There are almost no failure cases for the model with a fixed time step schedule, MoreRed-ITP, and the median RMSD ratio is particularly low for the two variants with an adaptive time step schedule, MoreRed-JT/-AS. These low ratios translate well to low absolute RMSDs between the relaxed structures and the reference structures, where the variants MoreRed-ITP/-JT/-AS achieve a median RMSD of 0.12 Å, 0.06 Å, and 0.05 Å, respectively. Further details and results based on the absolute RMSD are provided in figure B1 in the supplementary information appendix B.2. Our findings show that the MoreRed variants, especially when using an adaptive schedule, excel in reliably bringing the test structures close to the reference equilibrium structures. The classic FF method MMFF94 shows the highest number of failure cases, which is above 40%, and has the worst median RMSD ratio. Interestingly, the MLFF and GFN2-xTB show very similar performance to each other, with more than 20% failure cases and a median RMSD ratio close to 0.4. However, we note that the baseline methods might capture minima from slightly different PES than the one described by the reference structures, for instance, due to the discussed structure-label mismatch in QM7-X. For a more comprehensive understanding, we first test the robustness of all approaches and then proceed to evaluate the DFT energy levels of the relaxed structures.

We assess the robustness of the methods on synthetically generated inputs by diffusing equilibrium test structures from QM7-X with 250 forward diffusion steps. We ensure that the resulting median RMSD between the diffused configurations and the equilibrium test structures is within the range of the RMSD between the non-equilibrium structures and the equilibrium test structures from QM7-X. Figure 4(a) (right) shows the RMSD ratios after molecular relaxation of the diffused structures. For all MoreRed variants, the median RMSD ratio further improves compared to relaxing the QM7-X test structures, which is expected as our method is trained to denoise diffused structures. However, there is an increase in failure cases, which we attribute to more often ending up in equilibrium states different from the reference geometry. We

[8] https://github.com/tblite/tblite.

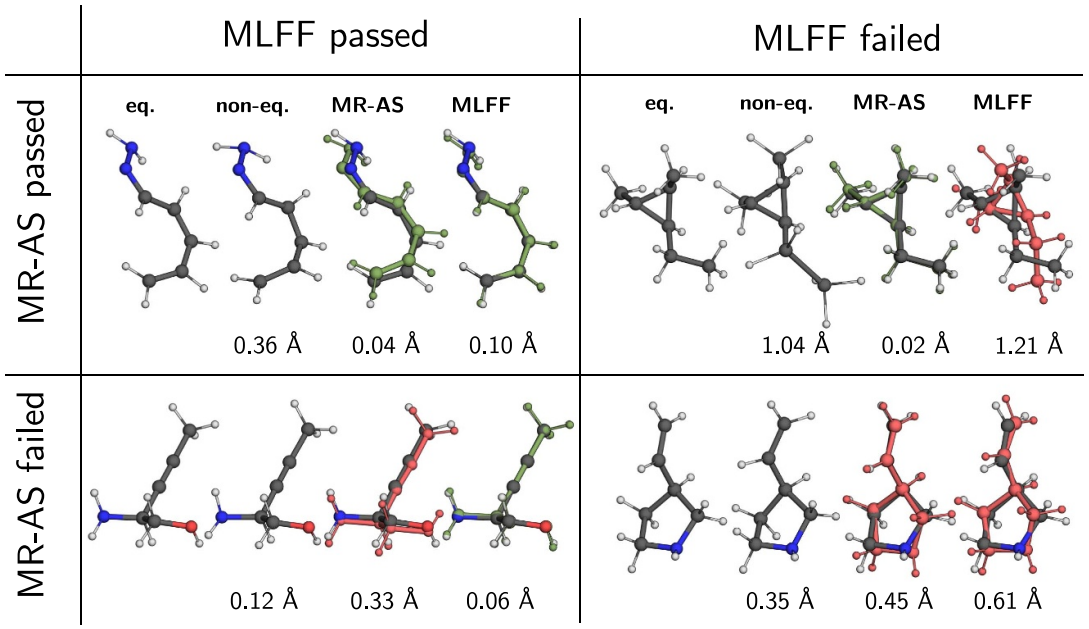


**Figure 5.** A pairwise comparison of molecular structures from QM7-X that were relaxed with both MoreRed-AS (MR-AS) and MLFF. In each block, from left to right, the equilibrium structure is shown (eq.), followed by the non-equilibrium structure (non-eq.) and the structures relaxed by MoreRed-AS and MLFF, all labelled with their respective RMSD to the equilibrium structure in angstrom (Å). The first two structures follow the CPK colouring, i.e. white for hydrogen, gray for carbon, blue for nitrogen, and red for oxygen. In the subsequent two depictions, the equilibrium structure (CPK) is superimposed with MoreRed-AS and MLFF relaxed structure. Structures with an decreased RMSD value relative to the non-equilibrium structure are shown in green, while failure cases with a higher RMSD compared to the initial structure are highlighted in salmon.

hypothesize that this is caused by the physically less plausible deviations in the diffused structures compared to the normal mode displaced structures in QM7-X. The existence of physically less plausible deviations is supported by the results of the baseline methods, where we observe a clear deterioration in performance. MMFF94 completely fails to handle input structures perturbed with Gaussian noise and, in nearly all cases, just returns the non-equilibrium input structure, leading to a median RMSD ratio of 1 and resulting in 99.6% failure cases. The median RMSD ratio of GFN2-xTB as well as its percentage of failure cases are more than doubled. Most notably, the MLFF fails to get closer to the reference geometry in almost 90% of the cases. For the MLFF, this is expected as the training data distribution of non-equilibrium structures from QM7-X does not cover all the chemical space, including the Gaussian diffused inputs. Therefore, relaxation often completely fails, leading to disconnected structures even if the input structure does not appear to be overly distorted. We show an example of this in figure B2(b), with a detailed discussion in appendix B.3, in the supplementary information. This means that, although the MLFF uses 100 times more training data than MoreRed, it cannot easily be transferred to relax the diffused structures. In contrast, MoreRed performs well on both the diffused samples and on the non-equilibrium structures from QM7-X albeit requiring only the unlabeled equilibrium structures for training. Accordingly, the diffusion training scheme leads to a more robust method for relaxation that can be used for input structures that are obtained from different sources, e.g. different datasets, various empirical FFs, or other generative models. This also means that MoreRed will oftentimes find a reasonable structure even if the input was physically not plausible, which should be considered by practitioners using the method.

For illustration, we show a series of pairwise comparisons of molecular structures from QM7-X relaxed by both MoreRed-AS and MLFF methods in figure 5. Each panel contains a sequence of depictions, beginning with the equilibrium structure (eq.), followed by the corresponding non-equilibrium structure, and then structures relaxed using MoreRed-AS and MLFF respectively. The RMSD values, relative to the equilibrium structure, are provided for each case. For both models, we show examples that were successfully relaxed as well as failure cases. Additionally, in figure B2 and appendix B.3 in the supplementary information, we provide further examples and discussion of relaxed structures using all MoreRed variants and baseline models, including MMFF94 and GFN2-xTB.

Finally, we compare the energies of relaxed structures using DFT calculations. We randomly sample 100 non-equilibrium test inputs from QM7-X and compute the energy of the corresponding equilibrium reference structure as well as of the structures resulting from the relaxation of the test inputs with all

methods. The energies are calculated with PBE0+MBD [109, 110] and a def2-TZVP [125] basis, using the PySCF [126–128] implementation. The results are reported in figure 6, where we compare the deviation of the energy of the structures resulting from relaxation with the different methods from the energy of the reference equilibrium structures reported in QM7-X. Positive energy differences occur when the relaxation method yields a structure with higher energy than the reference structure, and a negative difference indicates that the molecular relaxation yields a structure with lower energy than the reference. The structures relaxed with MMFF94 mostly have significantly larger energies than the reference structures. Hence, MMFF94 is clearly outperformed by MoreRed and the other baselines, as it results in the largest structural deviations in terms of the RMSD ratio and the worst energy levels in our DFT calculations. The MLFF, on the other hand, mostly finds structures with lower energy levels compared to the reported reference structures in QM7-X. According to the previously discussed structure-label mismatch in the dataset, this can be attributed to training the MLFF on energy and force labels calculated with the more accurate PBE0+MBD, whereas the reference equilibrium structures were found with DFTB+MBD. These reported equilibrium structures still have a mean and median force magnitude greater than 5 kcal $mol^{-1}$ $Å^{-1}$ when calculated with PBE0+MBD instead of DFTB+MBD, i.e. they are no minima on the PES defined by PBE0+MBD. The energy of structures relaxed with GFN2-xTB is higher than with the MLFF but still lower than the energy of the DFTB+MBD reference structures. This shows that the larger RMSD ratios found for the MLFF and GFN2-xTB stem from finding minima on a more accurate PES that is different from the one described by the QM7-X reference equilibrium structures. The energy levels, indicated by the darker colors in figure 6, of structures relaxed with MoreRed-ITP/-JT, trained on the reference DFTB3+MBD minima from QM7-X, match the energy levels of these reference structures the closest. For a majority of relaxed structures, the energy is less than 1 kcal $mol^{-1}$ higher, i.e. within chemical accuracy. Only for MoreRed-AS, which had the smallest median RMSD ratio, we observe higher energy levels than expected. It also performs more reverse denoising steps than the other two variants most of the time, further reducing the structural deviation while underestimating the interatomic distances, which are more strongly penalized in energy calculations than in the RMSD metric. Overall, we observe a mismatch between the RMSD results and the DFT energies of the MLFF and all MoreRed variants that were trained on equilibrium structures from QM7-X (darker colors in figure 6), where methods having lower RMSD ratios result in higher energies and vice versa. Therefore, to further investigate this mismatch and ensure a fair comparison, we further relaxed the DFTB+MBD equilibrium structures of QM7-X using the MLFF to minimize the forces and consequently the energies. We then retrained all MoreRed variants on this MLFF-relaxed dataset and used them to relax the non-equilibrium test structures from QM7-X. The resulting DFT energies, represented with light-colored boxes in figure 6, show that MoreRed models, trained on the MLFF-relaxed structures, achieve much lower PBE0+MBD energies compared to those trained on the original DFTB+MBD minima. However, they exhibit higher RMSD ratios, as demonstrated in figure B3 in the supplementary information. This outcome aligns with the results achieved by the MLFF and supports the notion that the RMSD-energy mismatch is related to the structure-label mismatch. Specifically, the MLFF finds lower PBE0+MBD energy minima that differ from the reference DFTB+MBD minima reported in QM7-X, resulting in higher RMSD ratios. Further details on this experiment are provided in appendix B.4 of the supplementary information.

In summary, we find that MoreRed accurately captures the data manifold of equilibrium structures as it outputs relaxed structures that are close to the reference structures in both structural deviation and in terms of energy difference. Moreover, using equilibrium structures with lower energy during training further improves MoreRed's energy performance. MoreRed outperforms the classical MMFF94 method in all of our experiments and metrics. While the MLFF and GFN2-xTB find structures with larger structural deviations, the obtained structures have lower energies according to reference calculations with PBE0+MBD. This was explained by the mismatch of computational methods employed in QM7-X to obtain equilibrium structures (DFTB+MBD) and to compute the energy and force labels (PBE0+MBD). Besides, MoreRed shows improved robustness to distorted inputs compared to all baseline methods, allowing it to relax structures from different sources albeit requiring only unlabeled equilibrium structures for training. When considering all metrics together, MoreRed-JT, which uses an adaptive schedule and predicts both the time step as well as the noise for reverse diffusion with the same neural network, performs better than MoreRed-ITP/-AS and is therefore recommended.

Efficiency-wise, we note that training MoreRed takes 1.45 days on average, while the MLFF model needs more than 7 days on NVIDIA P100. The relaxation of one single structure with MMFF94 takes 0.022 s. A single relaxation step per structure with the MLFF, GFN2-xTB, and MoreRed takes 0.02 s, 1.5 s, and 0.03 s, respectively. In contrast to the other methods that relax one structure at a time, MoreRed relaxes the

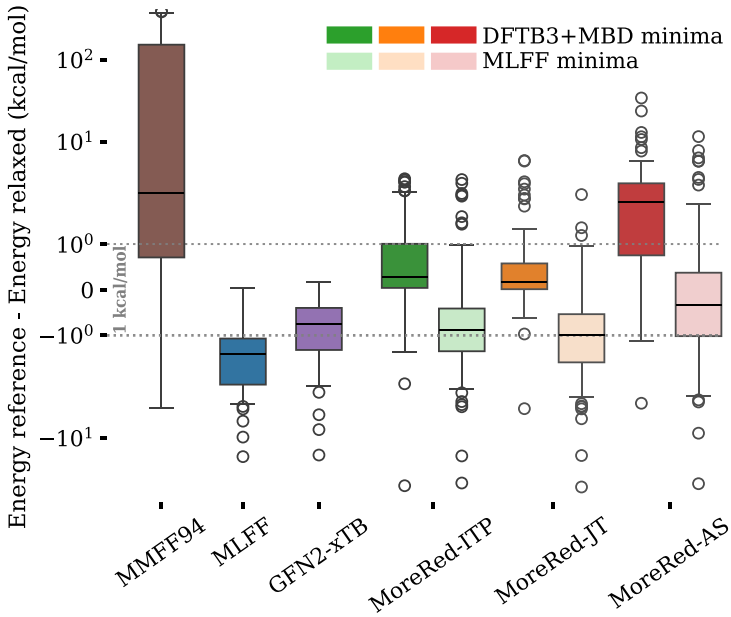


**Figure 6.** Deviation of the energy of structures resulting from relaxation with the different methods from the energy of the reference structures from QM7-X. Negative values indicate lower energy than the equilibrium reference structure. The DFT energy calculations were performed with PBE0+MBD and a def2-TZVP basis for 100 randomly sampled test structures from QM7-X. Each MoreRed variant has two versions: the darker color represents the model trained on the original DFTB3+MBD equilibrium structures from QM7-X, which have non-negligible PBE0+MBD residual forces, and uses the same relaxation results used for figure 4. On the other hand, the lighter color represents the model trained on equilibrium structures resulting from further relaxation of the QM7-X reference structures to minimize the residual forces, using the MLFF trained on the more accurate PBE0+MBD force labels.

structures batchwise, which could yield an even higher speed-up on tensor units. In our experiments, using batches of 128 structures results in 0.05 s per relaxation step per batch, i.e. 0.05 s/128 ≈ 0.0004 s per relaxation step per structure. A more detailed analysis of the computation times can be found in the supplementary information appendix D. Moreover, the effect of the size of the molecule on the performance and efficiency of MoreRed is studied in appendix E.

## 4. Conclusion

In this study, we introduced MoreRed, a conceptually novel and data-efficient approach for molecular relaxation employing reverse diffusion with a time step prediction component. MoreRed learns the data manifold of equilibrium structures and accurately maps non-equilibrium structures to equilibrium structures, without the need for forces, energies, or non-equilibrium training data. Compared to the other tested methods, its performance in relaxing non-equilibrium structures distorted by either normal modes or Gaussian noise remains robust.

A key technical novelty of our diffusion model lies in the integration of a time step predictor, which estimates the distortion level within a molecular structure. This enables the denoising of input structures with arbitrary noise levels, extending the applicability of diffusion models. Additionally, it allows for a novel adaptive schedule, enhancing MoreRed's capability to rectify accumulated errors in the reverse denoising process. To this end, we provided both theoretical arguments and empirical evidence supporting the feasibility of time step prediction in high-dimensional spaces. Three variants of MoreRed were introduced: (i) MoreRed-ITP (Initial Time Prediction), which estimates the distortion level for only the initial input structure; (ii) MoreRed-AS (Adaptive Schedule), which predicts a new time step for each denoising step of the reverse diffusion process, providing enhanced flexibility to move back and forth in time; and (iii) MoreRed-JT (Joint Training), which retains the adaptive schedule of MoreRed-AS but estimates both the diffusion noise and time step using one joint neural network instead of two separate ones. Our adaptive approach not only proves beneficial for molecular relaxation but also for molecular generation tasks.

While not directly comparable due to the mismatch in computational methods used to create the dataset, MoreRed exhibits accurate structure relaxation performance with significantly fewer training points and reduced training time compared to machine learning FFs. The data efficiency might prove beneficial for larger systems or extensive databases where generating a sufficient amount of accurately labeled training data, especially non-equilibrium structures, is challenging. The results also revealed an inherent issue when

using mismatched methods for structure relaxation and properties calculation, as observed in the QM7-X dataset. While machine learning FFs produce more accurate minima by following the PBE0+MBD-based forces, MoreRed accurately learns the data manifold of the provided DFTB+MBD-based equilibrium structures. Utilizing more accurate minima during training can further improve MoreRed's energy performance while maintaining its data efficiency advantage.

## Data availability statement

The code and its associated data are made public in Zenodo [129] and on Github at https://github.com/khaledkah/MoreRed. The datasets utilized for training the models, namely QM7-X [104] and QM9 [130, 131], are also publicly accessible.

## Acknowledgments

This work was partly funded by the German Ministry for Education and Research (BMBF) as BIFOLD—Berlin Institute for the Foundations of Learning and Data (BIFOLD24B) (under Refs 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18056A, 01IS18025A and 01IS18037A) and BBDC/BZML. Furthermore, Klaus-Robert Müller was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korean Government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University and No. 2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation). We thank Stefan Chmiela, Jonas Lederer, and Elron Pens for insightful discussions and feedback.

## Appendix A. MoreRed: details

### A.1. Diffusion time step prediction

*A.1.1. Derivation of equation (7)*

In the polar coordinate system $(r, \boldsymbol{\varphi})$, the marginal distribution of the radius $r$ of the centered isotropic Gaussian over the direction $\boldsymbol{\varphi}$ is given by

$$\begin{aligned} p(r) &= \int p(r, \boldsymbol{\varphi})\, \mathrm{d}\boldsymbol{\varphi} \\ &= \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \cdot r^{D-1} S_D \\ &= \frac{r^{D-1}}{2^{D/2-1}\Gamma(D/2)\,\sigma^D} \exp\left(-\frac{r^2}{2\sigma^2}\right), \end{aligned} \tag{A.1}$$

where $S_D = \frac{2\pi^{D/2}}{\Gamma(D/2)}$ is the surface area of the $(D-1)$-dimensional unit sphere embedded in the $D$-dimensional space, $\Gamma(\cdot)$ denotes the Gamma function [120] and $\sigma^2$ is the variance, which is equal to $(1-\bar{\alpha}_t)$ in equation (7). By changing the radius variable $r$ to the scaled version, $\tilde{r} = \frac{r}{\sqrt{D}\sigma}$, we get

$$\begin{aligned} p(\tilde{r}) &= p(r)\,\frac{\mathrm{d}r}{\mathrm{d}\tilde{r}} \\ &= \frac{\left(\sqrt{D}\sigma\tilde{r}\right)^{D-1}}{2^{D/2-1}\Gamma(D/2)\,\sigma^D} \exp\left(-\frac{\left(\sqrt{D}\sigma\tilde{r}\right)^2}{2\sigma^2}\right) \cdot \sqrt{D}\sigma \\ &= \frac{D^{D/2}\tilde{r}^{D-1}}{2^{D/2-1}\Gamma(D/2)} \exp\left(-\frac{D\tilde{r}^2}{2}\right), \end{aligned} \tag{A.2}$$

which gives equation (7).

*A.1.2. Discussion*

Our diffusion time step predictor essentially predicts the noise level of the input sample, which reduces to predicting the variance of the Gaussian component, from which the noisy input sample is drawn, using solely this *single* perturbed sample. Intuitively, this is too challenging in a low-dimensional space because the

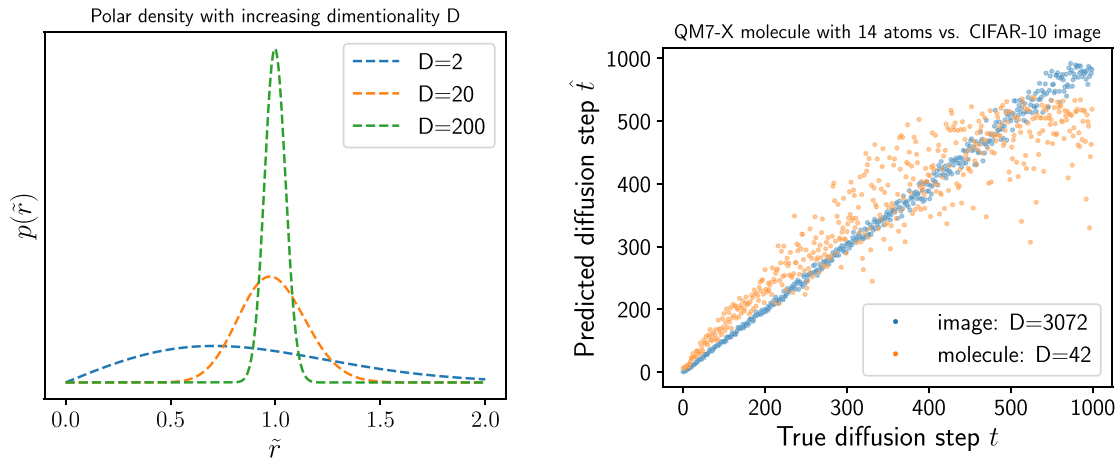


**Figure A1.** Left: distribution of the (scaled) radius of one Gaussian component in the polar coordinate system (equation (7)), illustrating sharper density with increasing dimensionality $D$. Right: diffusion time prediction performance using a molecule (orange) with 14 atoms from QM7-X, resulting in an input sample with $14 \cdot 3 = 42$ dimensions, and using an image (blue) from CIFAR-10 with $32 \cdot 32 \cdot 3 = 3072$. The time prediction becomes more accurate with increasing dimensions and/or decreasing noise variance, meaning low time step $t$.

distances between the mean and different samples from the same Gaussian are broadly distributed. This intuition does not apply to a high-dimensional space.

As discussed in section 2.3, the marginal distribution of the radius in the polar coordinate system, provided in equation (7) and depicted in the left plot in figure A1, implies that most of the Gaussian perturbed samples lie in a thin shell with an equal distance to the center of the Gaussian. This implies that a neural network, which can learn the data manifold, can predict the variance of the perturbation noise, and consequently, the diffusion time step.

One might still worry that training samples drawn from two overlapping Gaussian components will deteriorate the performance of the diffusion time prediction. Indeed, such overlapping makes the prediction harder, as can be empirically seen in the right plot in figure A1. However, with high dimensional input space, such overlapping does not significantly affect the time prediction performance for samples perturbed with small noise variance, meaning when the diffusion time step $t$ is low. Assume that there are two training molecules $\mathbf{x}_a, \mathbf{x}_b$ with the Euclidean distance $r = \|\mathbf{x}_a - \mathbf{x}_b\|$, and consider the Gaussian component centered at $\mathbf{x}_a$, which represents its noisy versions with standard deviation $\sigma = r$. Although, in this situation, $\mathbf{x}_b$ lies in the high-density shell of this Gaussian component (the bump in figure A1 left), the noisy samples of $\mathbf{x}_a$ are uniformly distributed all over the high $(D-1)$-dimensional shell. Therefore the probability that the Gaussian noise produces a sample close to $\mathbf{x}_b$ is extremely low. On the other hand, many training samples from the neighbourhood of $\mathbf{x}_b$ are fed to the diffusion time predictor as slightly noisy versions of $\mathbf{x}_b$, because of its high density. Accordingly, the network is trained to recognize the molecules close to $\mathbf{x}_b$ as low noise samples resulting from $\mathbf{x}_b$ without being disturbed by high noise samples from $\mathbf{x}_a$. This intuition can be mathematically confirmed by computing the density ratio between two Gaussian components around $\mathbf{x}_b$, i.e. $\mathcal{N}(\mathbf{x}_b + \boldsymbol{\varepsilon}; \mathbf{x}_b, \delta^2\mathbf{I}) / \mathcal{N}(\mathbf{x}_b + \boldsymbol{\varepsilon}; \mathbf{x}_a, r^2\mathbf{I})$ for $\|\boldsymbol{\varepsilon}\| \sim \delta \ll r$, which is extremely high unless $D$ is very small.

To conclude, the high dimensionality of the data space enables accurate diffusion time step prediction, especially for the samples close to one of the training equilibrium molecules. This can also be observed empirically. Figure A1(right) with orange dots shows a scatter plot of the true diffusion time vs. its prediction by our diffusion time predictor after training on the equilibrium molecules from the QM7-X dataset. We also show the performance of the diffusion time predictor trained on CIFAR10 [132]—a common image benchmark dataset—where the images have a higher dimensionality than the molecule data, as blue dots. As discussed above, the diffusion time prediction is easier when the dimension $D$ is large, and the true diffusion time, i.e. the noise level, is small.

### A.2. Algorithms

Algorithm 1 shows the training procedure for MoreRed-JT. For the other two variants, we instead train two separate architectures and use only the first part of the loss in line 7 to train the denoising model and the second part to train the time step prediction model separately. Algorithm 2 describes the sampling with the adaptive MoreRed variants (AS and JT). MoreRed-ITP uses a fixed schedule $i = \hat{t}, \hat{t}-1, \ldots, 1$ but starts from a predicted initial time step $i = \hat{t}$ instead of a fixed value.

**Algorithm 1.** Training.

**Input:** $q_{\text{data}}(\mathbf{x}_0)$, $a(t)$, $\eta$, $\theta$, $\Theta$
**Output:** $\varepsilon_\theta$, $\tau_\Theta$
**repeat**
  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
  $t \sim \mathcal{U}(1, T)$
  $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  subtract center of geometry from $\boldsymbol{\varepsilon}$
  $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}$
  Take SGD step with the gradient
  $\nabla_{(\theta,\Theta)} \Big[ \eta ||\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, \tau_\Theta(\mathbf{x}_t))||^2 + (1-\eta)||\tau_\Theta(\mathbf{x}_t) - a(t)||^2 \Big]$
**until** convergence

**Algorithm 2.** Sampling.

**Input:** $\varepsilon_\theta$, $\tau_\Theta$
**Output:** new sample $\mathbf{x}_i$, #iterations $i$
$i = 0$
$\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
**while** $\tau_\Theta(\mathbf{x}_i) \neq 0$ **do**
  $\hat{t} = \tau_\Theta(\mathbf{x}_i)$
  $\bar{\boldsymbol{\varepsilon}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  subtract center of geometry from $\bar{\boldsymbol{\varepsilon}}$
  $\boldsymbol{\varepsilon}_\theta = \boldsymbol{\varepsilon}_\theta(\mathbf{x}_i, \hat{t})$
  subtract center of geometry from $\boldsymbol{\varepsilon}_\theta$
  $\mathbf{x}_{i+1} = \frac{1}{\sqrt{\alpha_{\hat{t}}}} \left( \mathbf{x}_i - \frac{\beta_{\hat{t}}}{\sqrt{1-\bar{\alpha}_{\hat{t}}}} \boldsymbol{\varepsilon}_\theta \right) + \sigma_{\hat{t}} \bar{\boldsymbol{\varepsilon}}$
  $i = i + 1$
**end while**
**return** $\mathbf{x}_i$, $i$

## Appendix B. Further experiments and details

### B.1. Datasets

*B.1.1. QM7-X*

QM7-X [104] is a comprehensive dataset that was derived from 7000 molecular graphs sampled from the GDB13 chemical space with up to 7 heavy atoms, including types C, N, O, S, and Cl. For each SMILES string, structural and constitutional isomers were obtained using the MMFF94 FF and subsequently optimized with DFTB3+MBD computations, leading to 42 000 equilibrium structures. To capture the PES close to the equilibrium molecules, non-equilibrium molecules were generated by displacing each equilibrium molecule along a linear combination of normal mode coordinates computed with DFTB3+MBD, such that the energy difference between the non-equilibrium and equilibrium structures follow a Boltzmann distribution. For each equilibrium structure, 100 non-equilibrium configurations were generated, leading to 4200 000 non-equilibrium structures in total, where forces and energies for each structure were computed with DFT calculations at the PBE0+MBD level with FHI-aims.

For our experiments, we split the dataset into individual sets for training, validation, and testing of all methods. This is done at the molecular graph level to prevent bias leakage between different sets due to related isomers and conformations originating from the same graph. Specifically, we use the molecules resulting from 4500 graphs for training, 1250 for validation and the rest for testing. Note that MoreRed does not utilize the non-equilibrium configurations for training, effectively decreasing the training set size by a factor of 100 compared to the training set of the MLFF model.

*B.1.2. QM9*

We evaluate the molecular structure generation performance, discussed in details in appendix B.6, on the QM9 dataset [130, 131], a widely used benchmark for molecular generation tasks [73, 81, 82]. It comprises approximately 130k equilibrium organic molecules, each containing up to 9 heavy atoms of types C, O, N,

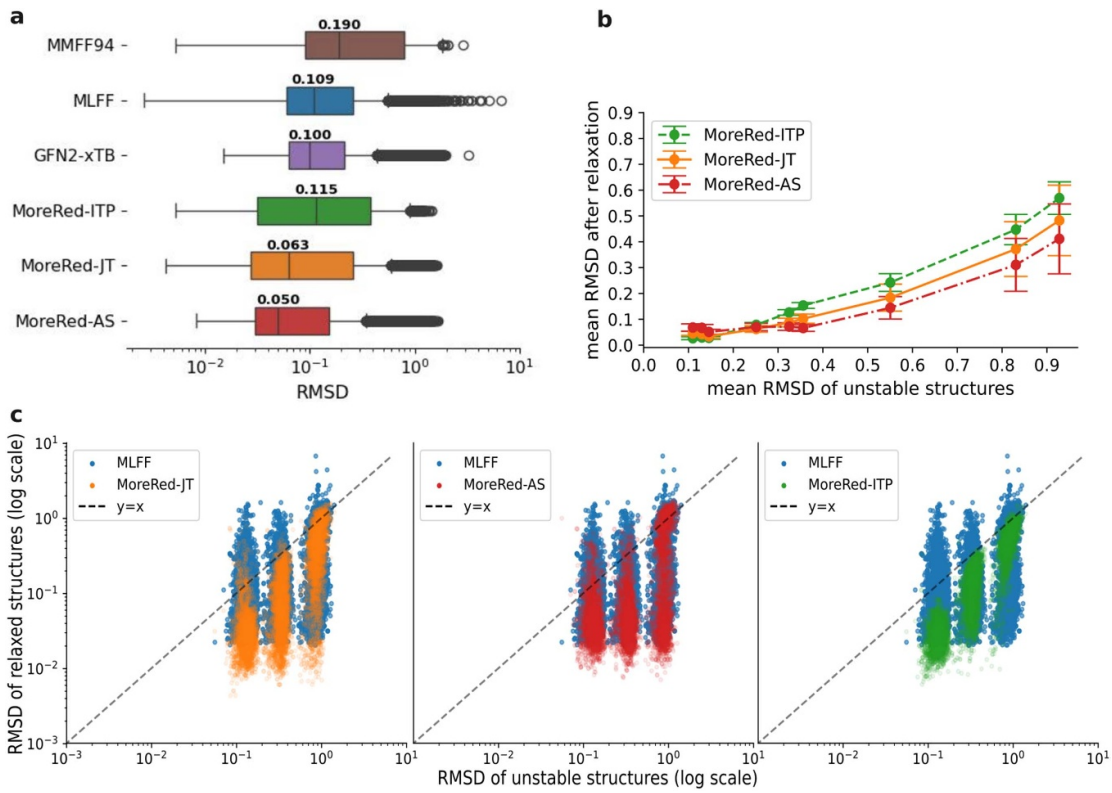


**Figure B1.** (a): The RMSD of molecules relaxed with the baselines and the MoreRed variants for 20 000 non-equilibrium structures from the QM7-X test split. (b): The mean RMSD of 20k non-equilibrium structures from the QM7-X test split vs. their mean RMSD after relaxation. The mean is taken over bins of 2k non-equilibrium molecules with increasing RMSD. The bars show the standard deviation of the RMSD after relaxation over five runs with the respective MoreRed variant. (c): RMSD of 10k non-equilibrium structures from the QM7-X test split before ($x$-axis) and after ($y$-axis) relaxation for all three variants of MoreRed and the baseline MLFF model. The three partitions occur because the test structures are sampled to cover low/intermediate/high RMSD values, where we sample 3 different non-equilibrium structures per equilibrium structure.

and F. We use 55k molecules for training, 10k for validation, e.g. for scheduling the learning rate, and define the rest as the test split.

### B.2. Extended analysis of relaxation with MoreRed

Here we further analyze the three different variants of MoreRed by discussing extended results from our experiments in section 3.2 on relaxing non-equilibrium structures from the QM7-X test set. In figure B1, we analyze the RMSD values of the optimized molecules in comparison to their equilibrium structure, instead of the RMSD ratio as it is reported in the main text. First of all, in figure B1(a) we compare the RMSD values of the three MoreRed variants to the baseline models, including MMFF94, MLFF and GFN2-xTB.

Notably, the two variants with adaptive scheduling have lower median RMSDs than all baselines, while the median RMSD of MoreRed-ITP is slightly worse than that of MLFF and GFN2-xTB. The reason for this can be seen in figure B1(c), where the RMSD values after relaxation are compared to the RMSD values of the initial non-equilibrium structures for all three MoreRed variants. While the performance of MoreRed-ITP (green) is particularly good for structures that are already close to the equilibrium state, its performance is impaired for structures that initially have a high RMSD. The adaptive variants, MoreRed-JT/-AS (orange, red), show a more balanced performance, successfully relaxing structures over the whole spectrum of non-equilibrium test molecules. This suggests that the adaptive scheduling with the time step prediction improves the relaxation of molecules that are further away from the data manifold, which is in line with our findings in section 3.1. This comes at the cost of a higher number of relaxation steps for the adaptive variants and more failure cases (see section 3.2 and figure 4).

Furthermore, to investigate the stochasticity of our method, we analyze the mean RMSD values and their standard deviation from the mean after optimization, subject to the RMSD of the initial structures. For this, we created 8 bins of initial structures based on their RMDS and measured the RMSD after optimization (see figure B1(b)). It shows that not only does the mean RMSD increase based on the initial RMSD, but also the standard deviation of the RMSD after optimization increases, with MoreRed-AS having the lowest variance and MoreRed-ITP having the highest. However, the variance is still small in comparison to the mean RMSD.

This is expected, because structures with large RMSD are assigned to high time steps, resulting in higher variance values for the diffusion reverse kernel. Besides, with high time steps, MoreRed needs more optimization steps until convergence and with every step, a small amount of stochasticity is added to the positions. Considering statistics over all test structures, the deviations across multiple relaxation runs with MoreRed are very low and therefore not reported in the boxplots.

In conclusion, for initial structures with higher levels of perturbation, the RMSD of the optimized structures increases and the variants with adaptive scheduling, MoreRed-JT/-AS, via time step prediction provide the required flexibility to perform more accurate optimization. On the other hand, the fixed schedule variant, MoreRed-ITP, provides fast and very accurate equilibrium molecules for initial structures with low levels of noise.

### B.3. Examples of relaxed structures

For illustration, in figure B2 we visualize relaxation of non-equilibrium structures from different sources. Example structures are shown in equilibrium and non-equilibrium on the left-hand side. On the right-hand side, the different relaxation methods are shown where the hydrogen atoms are suppressed. In panel (b), we use the non-equilibrium molecules obtained by perturbing the equilibrium geometries using 250 steps of Gaussian diffusion. Although the diffused non-equilibrium examples are only distorted up to an extent where the reference structure is visually still recognizable and the median noisy RMSD matches that of the non-equilibrium structures from QM7-X, they become hard to relax for the baseline methods. They fail at relaxing the first example while the three MoreRed variants converge towards the ground-truth equilibrium geometry. Especially the results from the models with adaptive schedule, MoreRed-JT/-AS, match the reference structure better than MoreRed-ITP. The MLFF gives a physically implausible result. In the bottom example, where the non-equilibrium structure has a lower RMSD from the equilibrium geometry, the MLFF and the semi-empirical method also converge to the reference geometry. However, all MoreRed variants achieve a significantly lower RMSD and the simple FF baseline, MMFF94, fails at recovering the reference. In panel (a), we use the QM7-X non-equilibrium structures. In the bottom example, all methods manage to find the reference equilibrium, where it is matched most closely by the two MoreRed variants with adaptive schedule (JT/AS). While the low RMSD values are a good indicator that MoreRed has accurately captured the distribution of equilibrium reference structures reported in QM7-X, the top example shows why additional metrics should be considered in the analysis. In this case, where the non-equilibrium structure has a higher RMSD from the equilibrium geometry, none of the methods recovers the reference. Nevertheless, the obtained configurations might be local minima on the PES that structurally deviate even further from the equilibrium structure reported in QM7-X than the non-equilibrium starting point of the relaxation. Therefore, we evaluate DFT-computed energies of relaxed structures in section 3.2 in the main text to gain further insights into the performance of all methods.

### B.4. Training on low energy minima from MLFF relaxations:

We investigate the structure-label mismatch in QM7-X, where force labels were computed using a more accurate and expensive method, PBE0+MBD, than for identifying the reference equilibrium structures, DFTB3+MBD. To this end, we first used the MLFF model, which was trained on the PBE0+MBD labels, to further relax the DFTB3+MBD equilibrium structures to lower energy levels. We then retrained all MoreRed variants on the relaxed data. Subsequently, we performed relaxation with the new MoreRed models on the non-equilibrium test structures from QM7-X, and computed the DFT energies of the resulting samples and their RMSD to the QM7-X reference minima, maintaining the same setup as in the main experiments.

If the inconsistency between the RMSD and energy results observed when training on the QM7-X references is due to the structure-label mismatch in the dataset, then training MoreRed on the MLFF-relaxed structures should result in similar energy and RMSD levels to those of MLFF. This is because MoreRed, by definition, learns the data manifold described by the training data and outputs structures with similar characteristics.

We summarize the results of the RMSD ratios in figure B3 and the DFT energies in figure 6. The three MoreRed variants achieve higher RMSD ratios but lower DFT energies, comparable to MLFF, when trained on the relaxed structures from MLFF (the lighter colors in the figures) rather than the original DFTB3-MBD reference structures from QM7-X (the darker colors). Thus, we can conclude that the structure-label mismatch in QM7-X causes the inconsistency between RMSD and energy results observed in the original experiments. Specifically, MLFF found minima with lower energy states but different structures than the reference minima in QM7-X, resulting in higher RMSD but lower DFT energy, as it was trained on more accurate labels. These accurate labels exhibit non-negligible forces, indicating that the equilibrium structures identified by DFTB3-MBD still have higher energies.

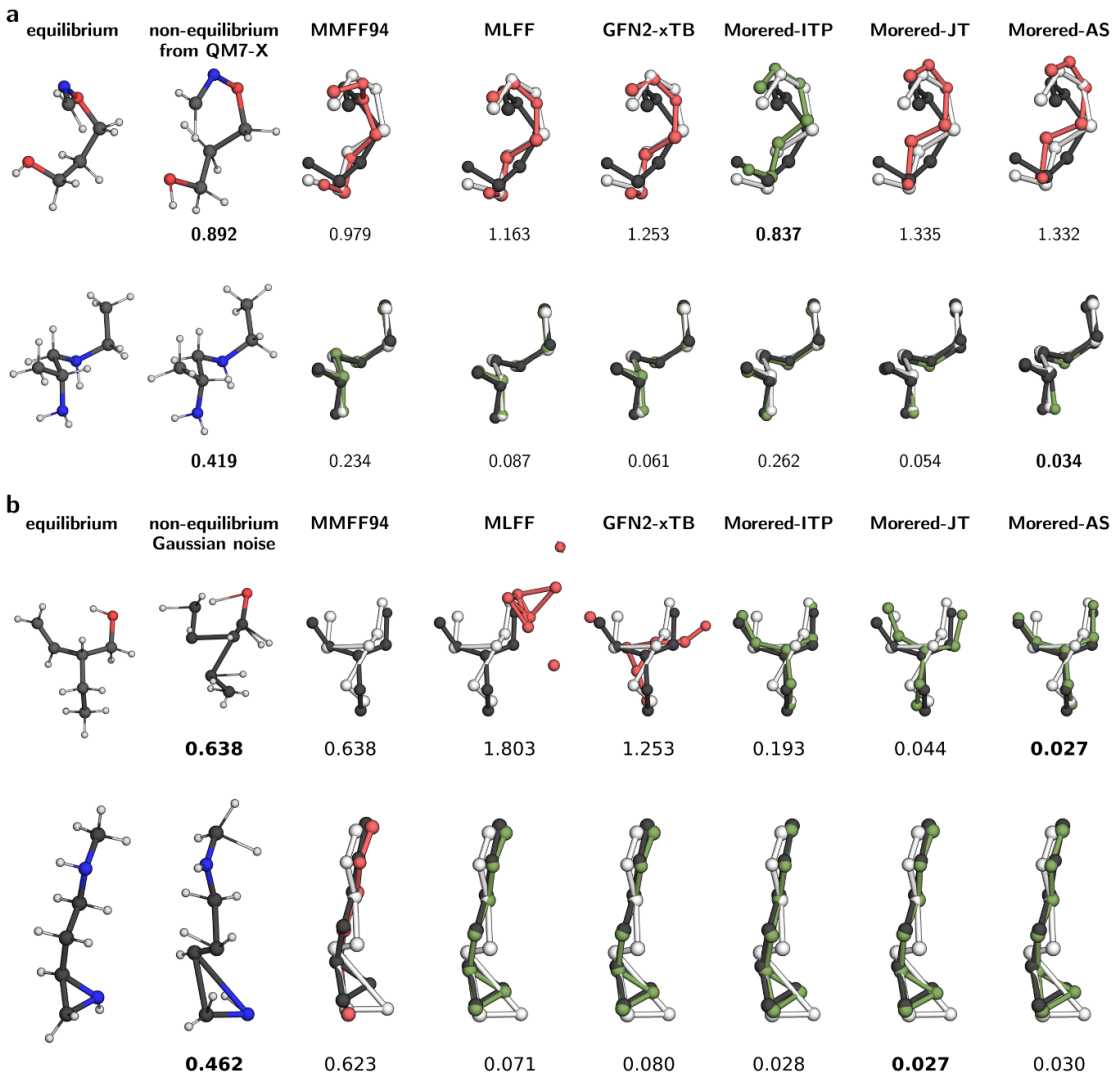


**Figure B2.** Exemplary relaxation results for all baselines and the MoreRed variants. Equilibrium geometries from the QM7-X test split and corresponding non-equilibrium structures are shown on the left-hand side with nitrogen in blue, oxygen in red, hydrogen in white, and carbon in grey. They are followed by results from the different relaxation methods, where the hydrogen atoms are suppressed for clarity and the equilibrium and the non-equilibrium structures are super-imposed in black and white respectively, together with the relaxed structure of the respective method. Failure cases are drawn in red and success cases in green. The RMSD from the equilibrium geometry is denoted below each structure. In (a), the non-equilibrium molecule is taken from QM7-x. In (b), the non-equilibrium molecule is obtained by applying the forward diffusion process, i.e. adding Gaussian noise, for 250 steps. In both panels, the top row shows a non-equilibrium structure with larger RMSD that was more difficult to converge for many methods than the example in the bottom row.

### B.5. Generalization with different equivariant representations

To further assess the robustness of our method, we conducted a set of experiments employing an alternative equivariant molecular representation to PaiNN. Specifically, we trained the MLFF model and all three variants of MoreRed using SO3Net [122] as a backbone representation. This representation incorporates spherical harmonics in the spirit of Tensor Field Networks [63] and NequIP [57] to handle SO(3)-equivariance, distinguishing it from the PaiNN architecture.

Utilizing the same data splits as in the PaiNN experiments, we tested the models on the same set of 20 000 non-equilibrium structures from the test split of QM7-X. All other experimental details align with those outlined in section 3.2 for PaiNN. Given the long training time of the MLFF model (7 d) in comparison to MoreRed, we opted to use half the number of parameters employed in PaiNN to expedite the experiments. However, to maintain fairness, we used the identical model hyperparameters for both MoreRed and MLFF. Additional hyperparameter details for SO3Net are provided in section C.2.3.

Our findings, summarized in figure B4, affirm that our approach performs comparably well with this alternative equivariant neural network backbone, consistently outperforming the MLFF model in terms of structure accuracy with MoreRed-AS and MoreRed-JT. Yet, the overall performance for all models, including the MLFF, is slightly worse than reported in figure 4 with PaiNN, and there is a subtle discrepancy in

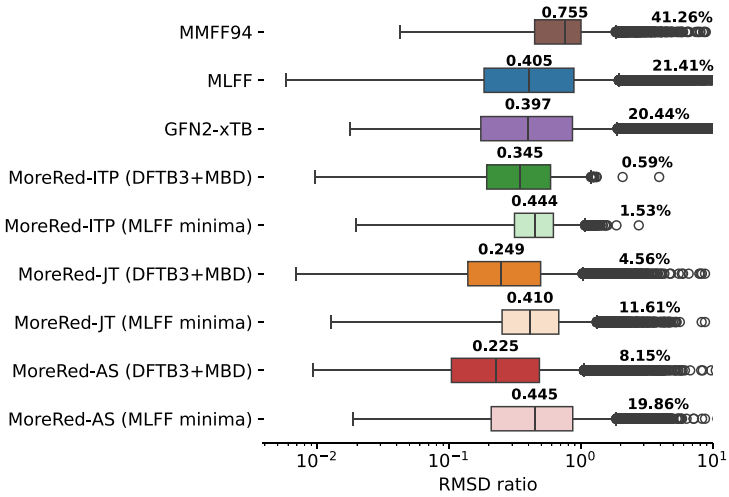


**Figure B3.** The RMSD ratios of structures relaxed with the baselines and the different MoreRed variants for 20 000 non-equilibrium structures from QM7-X, with each MoreRed variant having two versions. One version is trained on the original DFTB3+MBD equilibrium structures from QM7-X, while the other is trained on equilibrium structures resulting from further relaxation of the QM7-X reference structures using the MLFF, as in figure 6. The median values and the percentage of failure cases, defined as those with an RMSD ratio $\geqslant 1$, are displayed on top of each box plot.

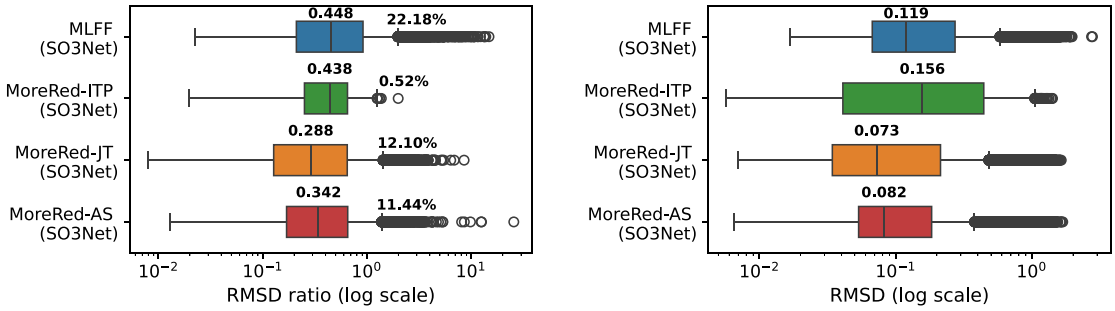


**Figure B4.** The RMSD ratios (left) and RMSD (right) of molecules relaxed with the baseline MLFF model and with our MoreRed variants for 20k non-equilibrium structures from the QM7-X test split using SO3Net implementation in SchNetPack [122] as an alternative equivariant backbone representation for PaiNN. The median values and the percentage of relaxation failure cases, i.e. the cases where the RMSD ratio exceeds 1, are shown above each box plot.

performance between MoreRed-AS and MoreRed-JT. We attribute these differences to the lack of hyperparameter tuning with SO3Net and the use of half the number of parameters employed in PaiNN.

### B.6. Molecular generation with time step prediction

*B.6.1. Improved molecule generation with adaptive schedule*

In contrast to existing diffusion models, e.g DDPM in our case, MoreRed-AS/-JT adaptively control the reverse diffusion process with the diffusion time step predictor, as observed in the relaxation task (see figure 3(b) in the main text for similar initial time step predictions and figure B6 for different initial time step predictions). This means that we follow an adaptive time step schedule where the time step at each denoising/sampling iteration is estimated by a neural network $\hat{t} = \tau_\Theta(\mathbf{x}_i)$. Sampling in existing diffusion models follows a pre-defined fixed schedule where exactly $T$ denoising steps with decreasing time step values $t = T, T-1, \ldots, 1$ are done, as illustrated by the dashed line in the left plot in figure B5. In the following, we compare the sampling performance with the adaptive schedule of our MoreRed to the standard, fixed diffusion model sampling process in the task of data generation from complete noise that diffusion models were originally designed for. To this end, given a molecular composition $\mathbf{Z}$, complete noise samples from the isotropic Gaussian prior distribution, $\mathcal{N}(\mathbf{0}, \mathbf{I})$, are used as initial molecular structures and are then denoised by both models to sample valid molecular structures. For a fair comparison, we employ the same noise model for classical sampling with a fixed schedule, denoted as DDPM, and for adaptive sampling using the time step predictor, i.e. MoreRed. In the rest of this section, we refer to MoreRed-JT as MoreRed. Examples of generated structures can be found in figures D1 and D2 for MoreRed and DDPM, respectively.

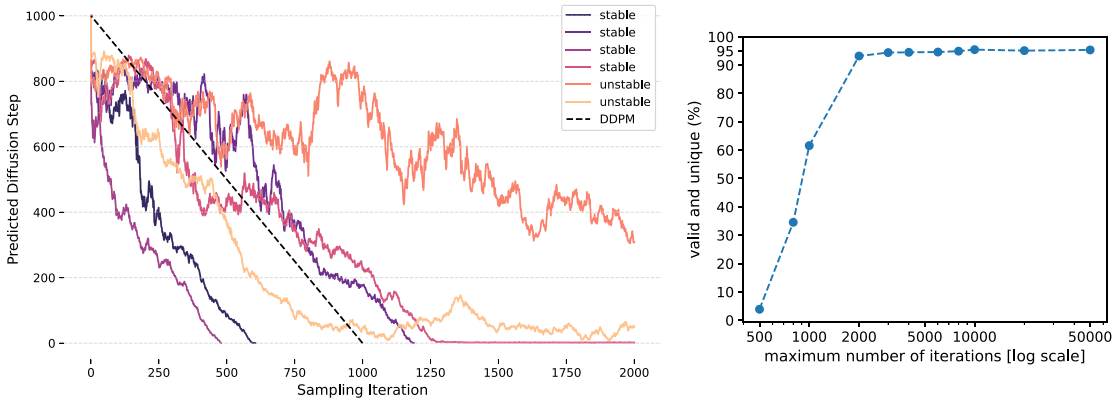


**Figure B5.** Left: examples of time step trajectories from sampling molecular structures from complete noise using MoreRed-JT. The solid lines indicate the predicted time steps by MoreRed through the sampling iterations. For reference, the dashed line indicates the fixed decreasing time step schedule used in the usual reverse diffusion from plain DDPM. Each colour indicates a different sample. In the box to the top-right, we note whether the sampled molecule passes the validity check explained in appendix B.6.1 (stable) or not (unstable). Right: the evolution of validity and uniqueness when increasing the maximum number of iterations for MoreRed-large (short for MoreRed-JT-Large) in the molecule generation task on QM9.

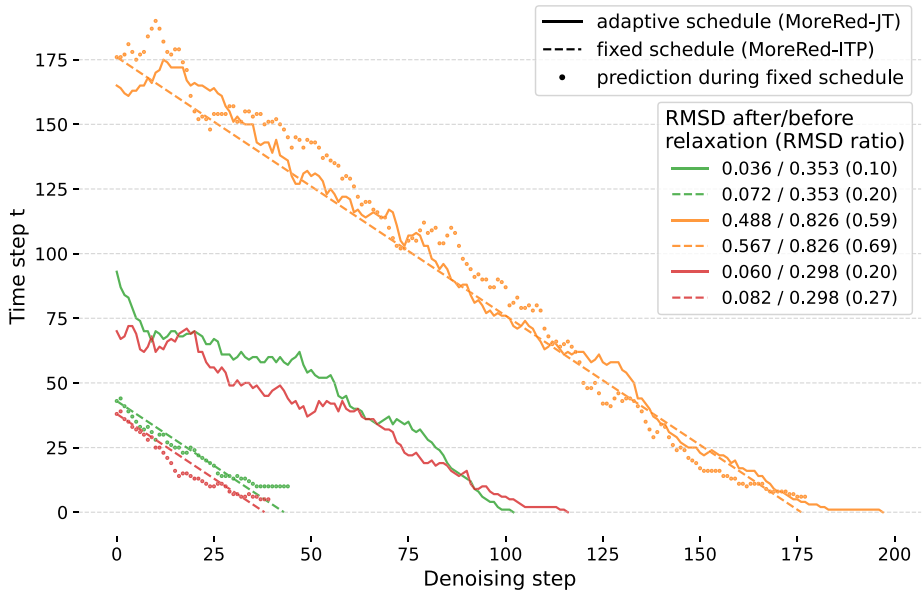


**Figure B6.** Comparison of time step trajectories for relaxation of three non-equilibrium test structures from QM7-X in red, orange, and green, similar to figure 3(b) but with significantly different initial time step predictions. MoreRed-JT follows an adaptive schedule (the time step and noise are predicted at each denoising step; solid lines) and MoreRed-ITP follows a fixed schedule (the time step is only predicted at the initial denoising step; dashed lines). The dots show the predicted (but not utilized) time steps for MoreRed-ITP during denoising with the fixed schedule. The inset box shows the RMSDs between the reference equilibrium geometry and the non-equilibrium structure after and before relaxation, respectively. MoreRed-JT outperforms MoreRed-ITP in all three cases, including the orange trajectory, where MoreRed-ITP starts from a higher initial step.

We evaluate the generation performance on the QM9 dataset [130, 131], which is a widely used benchmark for molecular generation tasks and described in appendix B.1. For data generation with MoreRed, we set the convergence criteria to $\hat{t} \leqslant 0$ or a maximum number of sampling iterations equal to 2000. For the standard DDPM [117], we use a fixed schedule with $T = 1000$ sampling iterations (same as during training). For our evaluation, we generate 10 000 structures starting from the latent prior distribution where the atomic compositions $\mathbf{Z}$ of molecules are randomly drawn from the QM9 test split.

As metrics, we adopt validity, uniqueness and novelty as proposed by Gebauer *et al* [81], using their publicly available analysis script for comparability. It translates the generated structures to canonical SMILES [133] encodings, which is a string representation of molecular graphs. A molecule is considered valid if all its atoms are connected and possess the proper valency in that encoding. Furthermore, unique and novel molecules are identified by comparing the canonical SMILES strings of all generated structures to each other and those of all molecules in QM9, respectively. Table B1 summarizes the results. We observe that

**Table B1.** Quality of 10 000 generated molecules after training on QM9. DDPM-large and MoreRed-large are larger neural networks that use 4 times more learnable parameters than DDPM and MoreRed, respectively. $V$ = valid, $U$ = unique, $N$ = novel.

| Model | $V$ (%) | $V+U$ (%) | $V+U+N$ (%) |
|---|---|---|---|
| DDPM | 78.2 | 77.3 | 62.5 |
| MoreRed | **89.3** | **88.0** | **68.6** |
| DDPM-large | 86.6 | 85.3 | 63.8 |
| MoreRed-large | **94.7** | **92.4** | **66.7** |

MoreRed, i.e. adaptive scheduling, performs better than DDPM, i.e. fixed scheduling, in all criteria. The same tendency can also be observed for architectures with more parameters, i.e. MoreRed-large. This confirms our hypothesis that our adaptive reverse diffusion procedure based on the time step prediction is beneficial for unconditional sampling from complete Gaussian noise and is not restricted to relaxation from noisy non-equilibirum structures.

MoreRed can dynamically adapt the time step at each sampling iteration to match the current noise level in the sample, correcting for the errors caused by the noise predictor, $\varepsilon_\theta$, as can be seen in some exemplary sampling trajectories in figure B5 left. This adaptive scheduling dynamically determines the number of reverse (sampling) iterations, providing a dynamic solution to the issue identified in Song *et al* [118] of one step noise prediction with fixed reverse diffusion leading to potential sample deviation from the optimal reverse trajectory. Namely, MoreRed mitigates this problem by automatically adjusting the time steps, offering a promising solution without manual hyperparameter tuning, which is necessary for previously proposed solutions that use correction steps after each reverse diffusion step, such as using second-order SDE/ODE solvers or running Langevin dynamics iterations [118, 134].

*B.6.2. Maximum number of sampling iterations*

The results presented in table B1 are obtained using a maximum number of sampling steps of 2000. However, to gain a more comprehensive understanding of the model's behaviour, we conduct experiments with varying maximum numbers of sampling steps using the large version of MoreRed-JT, i.e. MoreRed-large, and summarize the validity and uniqueness results in the right plot in figure B5.

We observe that the model can generate around 4% valid and unique molecules with less or equal to 500 sampling iterations and up to 35% with no more than 800 steps, which is less than 1000 steps during training. Interestingly, MoreRed-large yields around 25% fewer valid and unique molecules compared to the standard diffusion model DDPM-large when using 1000 steps. Yet, it outperforms it by approximately 7% using 2000 sampling steps at most. This suggests that the error introduced by the stochastic predictor of the diffusion step may slow down convergence in certain cases but ultimately lead to finer samples. Furthermore, by observing the evolution of the curve, we deduce that using 2000 steps is sufficient to achieve results close to the best performance. Beyond 2000 steps and up to 50k sampling steps, it exhibits only marginal improvements compared to the significant progress observed between 500 and 2000 steps. The latter aligns with the findings of Song *et al* [118], who showed that using exactly 2000 steps of a predictor-corrector sampler with manually tuned hyperparameters instead of only 1000 steps of a predictor-only sampler, like DDPM, to sample from a diffusion model trained on 1000 steps enhances performance in images. Nevertheless, in our method, we do not fix the exact number of iterations to 2000 for all the samples and we do not need a manual tuning of hyperparameters but MoreRed dynamically sets the number of steps by iteratively predicting the time step, eventually ending sampling after less than 2000 steps if the convergence criteria, $\hat{t} \leqslant 0$, is met.

## Appendix C. Model architectures and hyperparameters

### C.1. Architectures

*C.1.1. MLFF*

The architecture of MLFF is illustrated in figure C1. Using the atomwise invariant features $\mathbf{s}$ from PaiNN, an atomwise multi-layer perceptron (MLP) predicts the atom-wise energies $\{E_{\text{pot}}^{(i)}\}_{i=1}^{N}$, which are then aggregated to form a permutation-invariant potential energy $E_{\text{pot}} = \sum_{i=1}^{N} E_{\text{pot}}^{(i)}$, where $N$ is the number of atoms in the molecule. The gradients, i.e. the interatomic forces, are computed as the derivative of the potential energy $E_{\text{pot}}$ with respect to the atom positions $\mathbf{x}$, which ensures energy conservation and equivariant predictions of interatomic forces.

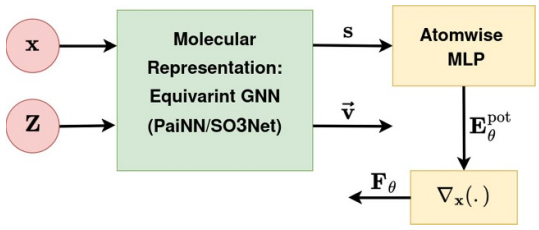


**Figure C1.** MLFF architecture: **x** are the atom positions, and **Z** are the atom types. An equivariant graph neural network, e.g. PaiNN, is used to learn an equivariant molecular representation that extracts one vector of invariant features **s** and one vector of equivariant vectorial features $\vec{\mathbf{v}}$ per atom. The scalar features $s$ are used to estimate the energy $E_{\text{pot}} = \sum_{i=1}^{N} E_{\text{pot}}^{(i)}$ using an atomwise MLP $E_\theta^{\text{pot}} = E_{\text{pot}}^{(i)}$, where $\theta$ represents the neural network weights. The derivatives with respect to the positions **x** represents the forces prediction, i.e. $F_\theta(\mathbf{x}) = -\nabla_{\mathbf{x}} E_{\text{pot}}$.

*C.1.2. MoreRed variants, including DDPM*

The architectures of the different variants are illustrated in figure C2. On the right side is the architecture of MoreRed-JT, where the time and noise head share the same backbone molecular representation, e.g. PaiNN. Similar to the invariant energy in MLFF, the time $\hat{t} = \tau_\Theta(\mathbf{x}_i)$ is predicted by an atomwise MLP using the invariant features as input. In contrast to the energy-conservative forces in MLFF, the diffusion noise is directly predicted as an equivariant tensorial quantity using an equivariant gated MLP on top of the equivariant vectorial features $\vec{\mathbf{v}}$. For MoreRed-ITP and MoreRed-AS, we use the same model architecture as for MoreRed-JT, with the only difference being that the time head $\tau_\Theta(\mathbf{x}_i)$ and the noise head $\varepsilon_\theta(\mathbf{x}_i, \hat{t})$ use two separate, but identical, molecular representation networks that are trained separately, instead of using one joint network. On the left side is the architecture of the plain DDPM [117] (section 2.1) used as a baseline model in the molecular generation experiments in appendix B.6. The overall architecture is similar to MoreRed, except that the diffusion time step $t$ is explicitly provided by the user as input to the noise head, rather than being dynamically predicted by a neural network.

### C.2. Hyperparameters

The models used in our experiments with molecules are implemented and trained using SchNetPack [122]. For all our experiments except the generalization experiments in section B.5, we use PaiNN [119] with 3 interaction blocks and 20 Gaussian radial basis functions with a cosine cutoff of 5 Å as molecular representation for all the models. After computing the molecular representation, the number of atomic features is halved in each layer of the output heads, with a total of 3 layers for each head for all the models. We use the AdamW [135] optimizer for all the models and train them until complete convergence. Moreover, we use the exponential moving average (EMA) of the model parameters with a decay of 0.999 for all models across all training epochs during validation, testing and inference rather than using the most recent parameter updates. Additionally, we use a learning schedule that halves the learning rate during training if the validation loss stagnates for a predefined number of epochs, allowing for finer steps near the local minima and avoiding fluctuating around them. We use early stopping to stop the training process when the validation loss stops decreasing after some epochs instead of using a fixed number of epochs and we use the model checkpoint with the lowest validation loss for testing and inference. The specific details for the different models are listed below.

*C.2.1. MLFF*

Overall, for MLFF training, we follow the hyperparameters and the training details reported in the original work [119], but we further tuned the batch size on $\{10, 64, 128\}$, the learning rate on $\{10^{-3}, 10^{-4}\}$ and the atomic features on $\{64, 128, 256\}$. We found that a batch size of 10, learning rate of $10^{-3}$ and 128 atomic features achieve the lowest loss, which aligns with the results from previous work using PaiNN [72, 119]. We found that using more atomic features than 128, i.e. more parameters, for MLFF hurt the performance. Additionally, we use a patience of 15 epochs for the learning rate schedule and 30 epochs for early stopping. The resulting baseline MLFF achieves a mean square error of 0.376 kcal mol$^{-1}$ for energy and 0.519 kcal mol$^{-1}$ Å$^{-1}$ for forces after training, which is within chemical accuracy of 1 kcal mol$^{-1}$ and on par with the benchmarks on QM7-X as reported in Unke *et al* [72].

*C.2.2. MoreRed variants, including DDPM*

For all diffusion models used in our work, we employ the polynomial approximation of the cosine noise schedule [73] with $T = 1000$ discretization steps, and a precision parameter of $s = 10^{-5}$ to prevent the atoms

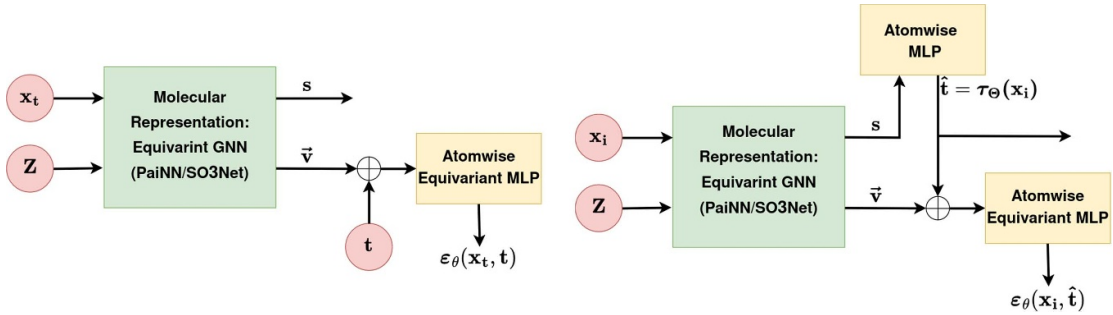


**Figure C2.** $\mathbf{x}_t$ and $\mathbf{x}_i$ are the (noisy) atom positions at true time step $t$ or unknown step $i$ respectively, and $\mathbf{Z}$ is a vector of integers representing the atom types. $\oplus$ denotes the elementwise vector concatenation operation. Both architectures use an equivariant graph neural network, e.g. PaiNN, to learn an equivariant molecular representation that extracts one vector of invariant features $\mathbf{s}$ and one vector of equivariant vectorial features $\vec{\mathbf{v}}$ per atom. Left: the model architecture used to predict the diffusion noise in the plain DDPM (section 2.1). The diffusion step $t \in \{0, 1, \ldots, T\}$ is normalized by dividing by the maximum time step $T$, i.e. $t/T \in [0, 1]$, and concatenated to the vectorial features $\vec{\mathbf{v}}$ of each atom to directly predict one equivariant noise vector $\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)$ per atom using an equivariant gated MLP. Right: the model architecture for MoreRed-JT. It is identical to the DDPM architecture except that the unknown time step of $\mathbf{x}_i$ is first predicted by an atomwise MLP $\hat{t} = \tau_\Theta(\mathbf{x}_i)$ on top of the invariant scalar features $\mathbf{s}$, similar to predicting the invariant potential energy in MLFF. MoreRed-ITP/-AS use the same architecture as MoreRed-JT except that $\tau_\Theta(\mathbf{x}_i)$ and $\boldsymbol{\varepsilon}_\theta(\mathbf{x}_i, \hat{t})$ use two separate GNN molecular representations, i.e. two separate green boxes in the diagram. In our experiments, we use two identical PaiNN architectures to compute the seperate features for the time and noise head when using MoreRed-ITP/-AS.

from undergoing large unrealistic movements during the initial sampling steps. We use a large batch size of 128 molecules to improve the accuracy of the loss estimation, as it involves uniformly sampling a single diffusion step $t$ per molecule per batch instead of using the whole trajectory for each molecule in the batch. We set the number of atomic features to 256 for all variants of MoreRed, including the plain DDPM model used in molecular generation experiments in section B.6. In these molecular generation experiments, we also employ architectures with more parameters, namely MoreRed-large and DDPM-large. Here we use 512 atomic features and 5 layers for the noise head instead of 3, resulting in circa 10M parameters instead of 2.5M. For MoreRed-JT, we found that setting $\eta$ to 0.9 works well because the noise prediction provides more signals (3N per molecule with N atoms) compared to diffusion step prediction (one step per molecule). While we use a separate time predictor with a separate representation for MoreRed-AS and MoreRed-ITP, all hyperparameters are kept consistent across all MoreRed variants for all experiments and datasets. As depicted in figure C3, uniformly sampling one time step $t$ per molecule per batch results in a noisy training loss since the model uses different diffusion steps at each training iteration instead of the entire diffusion trajectory. To mitigate this issue, we used EMA of the parameters with a decay of 0.999 across all training epochs during validation, testing and inference rather than using the most recent parameter updates. This approach yielded smoother learning evolution, as reflected in the less noisy validation loss in figure C3 because the EMA of the parameters better maintains the previously learned signal from the different diffusion steps seen per molecule. We use a patience of 300 epochs for early stopping and 150 epochs instead of 15 for the learning rate schedule because MoreRed uses only non-equilibrium molecules resulting in 100 times fewer data and fewer iterations per epoch.

In initial experiments with MoreRed, we observed that the model outputs occasionally explode, resulting in exploding gradients and divergence of the training, as illustrated by the high peaks in the validation loss in figure C3 after 400 epochs. Subsequent analysis revealed that this phenomenon occurs only for high diffusion steps above $t = 850$, which was associated with highly noisy and unrealistically dense molecules where all atoms were tightly packed with tiny atomic distances, as shown in figure C3. Hence, we added gradient clipping with a global norm of 0.5 to mitigate this explosive behaviour.

### *C.2.3. Details for SO3Net*

In the generalization experiments, in section B.5, using SO3Net instead of PaiNN, we maintain a consistent configuration with 128 atomic features across the entire architecture and utilize 2 hidden layers for all property prediction heads for all models, including the MLFF. The latter results in identical model sizes for both MLFF and MoreRed, with MoreRed using fewer parameters than in the main experiments. Additionally, we set $l_{\text{max}} = 1$ for the maximum degree of the spherical harmonics features. To speed up training, a large batch size of 512 and a learning rate of $2 \cdot 10^{-3}$ is utilized for all models. All other experimental details remain unchanged from those employed in the main experiments with PaiNN.

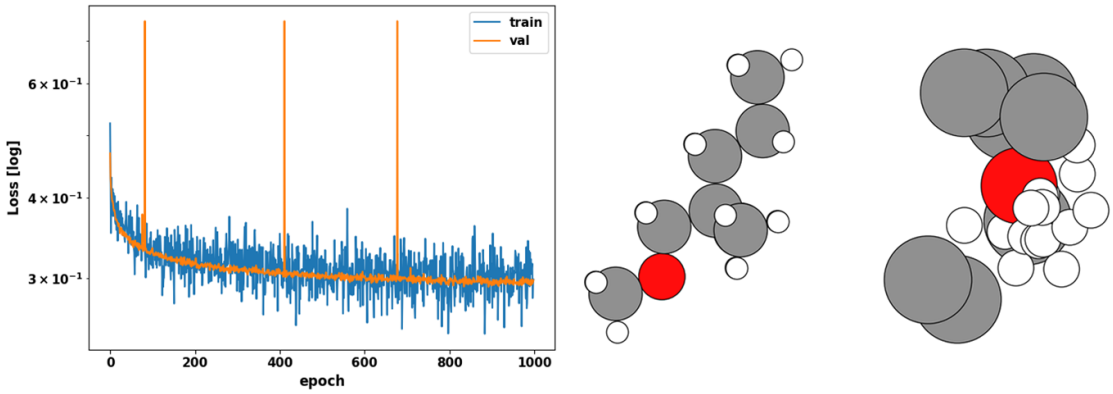


**Figure C3.** Left: plot showing the evolution of the training loss in blue and validation loss in orange in the log scale. The validation loss has fewer fluctuations than the training loss, showcasing the effect of applying EMA on the model's parameters. The validation loss shows three high peaks due to exploding model outputs. Subsequent analysis revealed that this explosive behaviour occurs only for high diffusion steps above $t = 850$, which was associated with highly noisy and unrealistically dense molecules where all atoms were tightly packed with tiny atomic distances, as can be seen in the rightmost figure. Right: example of a degenerated molecule after 916 diffusion steps. On the left is the original molecule. On the right is the diffused molecule, where all the hydrogens (white nodes) are tightly condensed, resulting in tiny atomic distances and explosive behaviour of the model. However, this issue does not occur frequently during training.

## Appendix D. Computation time

Here we give a more detailed analysis of the computation time for the three MoreRed variants and the MLFF model that have been used for the experiments in section 3.2. The median number of steps until convergence as well as the accuracy of the equilibrium structures strongly depend on the MoreRed variant, ranging from 53 steps for the fastest MoreRed-ITP to 1000 steps for the most accurate MoreRed-AS. For MoreRed-AS, we observe many trajectories where the model predicts many consecutive low-time steps until reaching the maximum number of allowed steps, but the optimization does not converge due to the strict convergence criterion of $\hat{t} \leqslant 0$. Applying less strict convergence criteria might decrease the number of steps per structure optimization significantly, which is a direction for future work. Furthermore, the computation time per structure during inference dramatically improves if many structures are evaluated in batches instead of sequentially. While batch-wise optimization can be done straightforwardly with all MoreRed variants, batch-wise structure optimization utilizing the MLFF model is not trivial, due to the dependency on the L-BFGS optimizer. In the following, we give a rough estimate of the inference time by either using sequential optimization or batch-wise optimization. For the analysis, we neglect any computational cost that is not directly related to model inference.

### D.1. Sequential optimization

Comparing MoreRed variants to the MLFF model, we observed an average inference time for a single structure, not a batch, of 0.03 s for MoreRed and 0.02 s for the MLFF model. To compute the mean inference time per structure optimization performed sequentially, we need to use not the median, as reported above, but the mean. The mean number of optimization steps until convergence was measured as 64 steps for MoreRed-ITP, 489 steps for MoreRed-JT, 992 for MoreRed-AS, and 122 for the MLFF model, which results in an average total inference time per structure optimization of $0.03\ \text{s} \cdot 64 = 1.92$ s for MoreRed-ITP, 14.67 s for MoreRed-JT, 29.76 s for MoreRed-As and 2.44 s for the FF.

### D.2. Parallelized optimization

For efficient optimization of a large number of structures, as is usually the case in many applications, model inference is preferably done in batches. Assuming batch-wise relaxation with the MLFF model is possible, we observed an inference time for evaluating a batch of 128 molecules of 0.03 s for the FF and 0.05 s for the MoreRed variants. Since the batch-wise relaxation is done until all the structures in the batch have converged, in a worst-case scenario, which is more likely to happen the larger the batch is, both methods need the maximum number of allowed steps, which is 1000. This would result in an average inference time per structure optimization of $\frac{0.03\ \text{s} \cdot 1000}{128} = 0.23$ s for the FF and $\frac{0.05\ \text{s} \cdot 1000}{128} = 0.39$ s for the MoreRed variants.

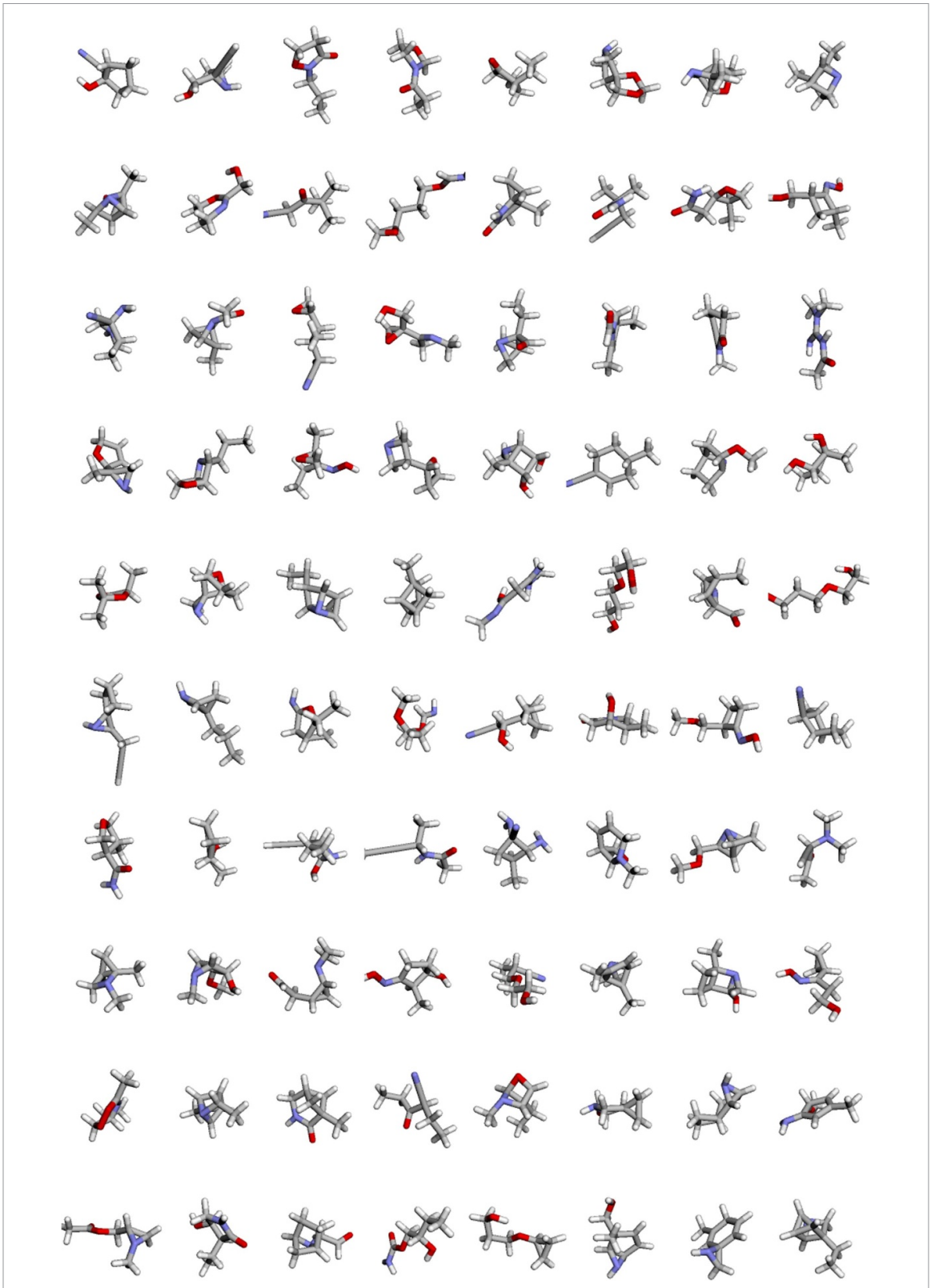

**Figure D1.** Batch of generated molecules with MoreRed-JT trained on QM9.

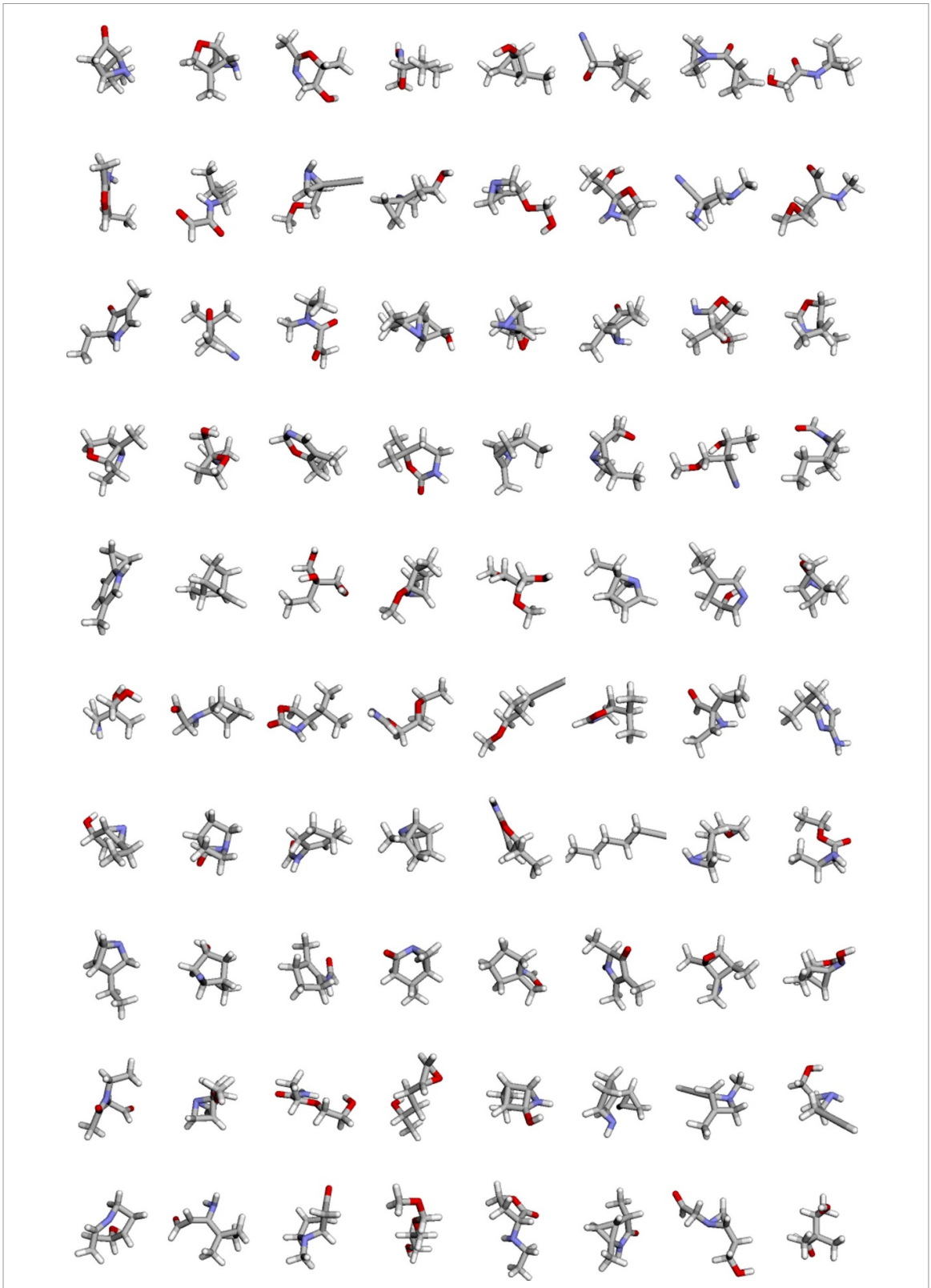

**Figure D2.** Batch of generated molecules with DDPM trained on QM9.

## Appendix E. Effect of molecular size on performance and efficiency

For many chemistry applications, it is required to perform molecular relaxation on large systems with many atoms. To evaluate how MoreRed performs on molecules with different numbers of atoms, we use the results of section 3.2, where molecular relaxation with MoreRed-JT is performed on 20 000 non-equilibrium structures of QM7-X, and categorize the results by the number of heavy atoms per molecule. In figure E1 we report the RMSD ratio, the number of denoising steps per relaxation trajectory and the predicted initial

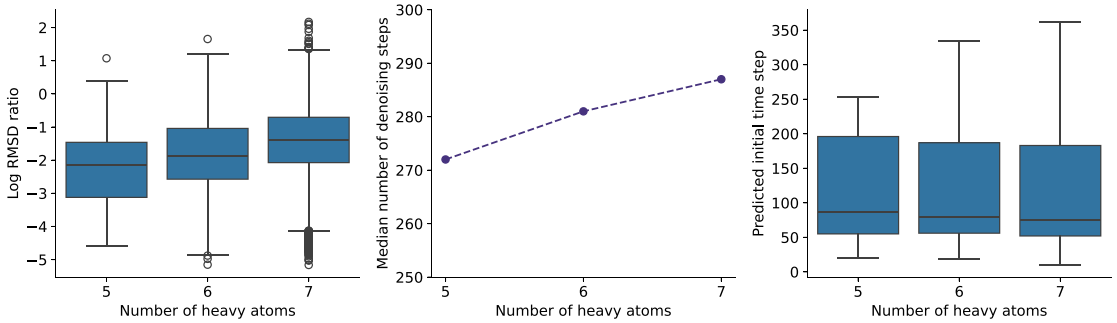


**Figure E1.** Analysis of molecular relaxation outcomes categorized by the count of heavy atoms, spanning from 5 to 7. The graph illustrates the RMSD ratio (left), the count of denoising steps required for convergence (mid), and the estimated initial timestep (right). The initial structures are derived from the non-equilibrium configurations of QM7-X and we use MoreRed-JT for the denoising.

timestep for all structures with 5–7 heavy atoms. We find that, excluding some outliers, the predicted initial time step remains largely unaffected by the number of heavy atoms. Conversely, as the number of atoms grows, the median number of denoising steps needed for convergence increases by up to 10 steps, and the RMSD ratios also rise. However, using the RMSD ratios as a metric is in favor of small structures because local inaccuracies tend to affect the whole molecule, significantly impacting the atom positions at the molecule's periphery. Therefore, similar inaccuracies at one point can lead to significantly larger RMSD of positions if the molecule is larger. Furthermore, the number of possible local minima increases with the number of heavy atoms.

## ORCID iDs

Khaled Kahouli https://orcid.org/0009-0002-5702-0021
Stefaan Simon Pierre Hessmann https://orcid.org/0000-0002-8399-2193
Klaus-Robert Müller https://orcid.org/0000-0002-3861-7685
Shinichi Nakajima https://orcid.org/0000-0003-3970-4569
Stefan Gugler https://orcid.org/0000-0001-6257-1923
Niklas Wolf Andreas Gebauer https://orcid.org/0000-0002-9149-7424

## References

[1] Schlegel H B 2011 Geometry optimization *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **1** 790–809
[2] Broadbelt L J, Stark S M and Klein M T 1994 Computer generated pyrolysis modeling: on-the-fly generation of species, reactions and rates *Ind. Eng. Chem. Res.* **33** 790–9
[3] Broadbelt L J, Stark S M and Klein M T 1996 Computer generated reaction modelling: decomposition and encoding algorithms for determining species uniqueness *Comput. Chem. Eng.* **20** 113–29
[4] Broadbelt L J and Pfaendtner J 2005 Lexicography of kinetic modeling of complex reaction networks *AIChE J.* **51** 2112–21
[5] Fialkowski M, Bishop K J M, Chubukov V A, Campbell C J and Grzybowski B A 2005 Architecture and evolution of organic chemistry *Angew. Chem., Int. Ed.* **44** 7263–9
[6] Gothard C M, Soh S, Gothard N A, Kowalczyk B, Wei Y, Baytekin B and Grzybowski B A 2012 Rewiring chemistry: algorithmic discovery and experimental validation of one-pot reactions in the network of organic chemistry *Angew. Chem., Int. Ed.* **51** 7922–7
[7] Kowalik M, Gothard C M, Drews A M, Gothard N A, Weckiewicz A, Fuller P E, Grzybowski B A and Bishop K J M 2012 Parallel optimization of synthetic pathways within the network of organic chemistry *Angew. Chem., Int. Ed.* **51** 7928–32
[8] Sameera W M C, Maeda S and Morokuma K 2016 Computational catalysis using the artificial force induced reaction method *Acc. Chem. Res.* **49** 763–73
[9] Dewyer A L, Argüelles A J and Zimmerman P M 2017 Methods for exploring reaction space in molecular systems *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **8** e1354
[10] Maeda S, Komagawa S, Uchiyama M and Morokuma K 2010 Finding reaction pathways for multicomponent reactions: the Passerini reaction is a four-component reaction *Angew. Chem., Int. Ed.* **50** 644–9
[11] Feinberg E N, Sur D, Wu Z, Husic B E, Mai H, Li Y, Sun S, Yang J, Ramsundar B and Pande V S 2018 Potentialnet for molecular property prediction *ACS Cent. Sci.* **4** 1520–30
[12] Simm G N, Vaucher A C and Reiher M 2018 Exploration of reaction pathways and chemical transformation networks *J. Phys. Chem.* A **123** 385–99
[13] Unsleber J P and Reiher M 2020 The exploration of chemical reaction networks *Annu. Rev. Phys. Chem.* **71** 121–42
[14] Baiardi A *et al* 2022 qcscine/utilities: release 5.0.0 *Zenodo* https://zenodo.org/record/6694755
[15] Deutschmann O and Schmidt L D 1998 Modeling the partial oxidation of methane in a short-contact-time reactor *AIChE J.* **44** 2465–77
[16] Zhu H, Kee R J, Janardhanan V M, Deutschmann O and Goodwin D G 2005 Modeling elementary heterogeneous chemistry and electrochemistry in solid-oxide fuel cells *J. Electrochem. Soc.* **152** A2427

[17] Gossler H, Maier L, Angeli S, Tischer S and Deutschmann O C 2019 An improved computer-aided method for developing catalytic reaction mechanisms *Catalysts* **9** 227
[18] Ulissi Z W, Medford A J, Bligaard T and Nørskov J K 2017 To address surface reaction network complexity using scaling relations machine learning and dft calculations *Nat. Commun.* **8** 14621
[19] Steiner M and Reiher M 2022 Autonomous reaction network exploration in homogeneous and heterogeneous catalysis *Top. Catal.* **65** 6–39
[20] Sankaran R, Hawkes E R, Chen J H, Lu T and Law C K 2007 Structure of a spatially developing turbulent lean methane–air Bunsen flame *Proc. Combust. Inst.* **31** 1291–8
[21] Harper M R, Van Geem K M, Pyl S P, Marin G B and Green W H 2011 Comprehensive reaction mechanism for n-butanol pyrolysis and combustion *Combust. Flame* **158** 16–41
[22] Vinu R and Broadbelt L J 2012 Unraveling reaction pathways and specifying reaction kinetics for complex systems *Annu. Rev. Chem. Biomol. Eng.* **3** 29–54
[23] Vereecken L, Glowacki D R and Pilling M J 2015 Theoretical chemical kinetics in tropospheric chemistry: methodologies and applications *Chem. Rev.* **115** 4063–114
[24] Proppe J and Reiher M 2017 Reliable estimation of prediction uncertainty for physicochemical property models *J. Chem. Theory Comput.* **13** 3297–317
[25] Proppe J and Reiher M 2018 Mechanism deduction from noisy chemical reaction networks *J. Chem. Theory Comput.* **15** 357–70
[26] Suleimanov Y V and Green W H 2015 Automated discovery of elementary chemical reaction steps using freezing string and Berny optimization methods *J. Chem. Theory Comput.* **11** 4248–59
[27] Gao C W, Allen J W, Green W H and West R H 2016 Reaction mechanism generator: automatic construction of chemical kinetic mechanisms *Comput. Phys. Commun.* **203** 212–25
[28] Susnow R G, Dean A M, Green W H, Peczak P and Broadbelt L J 1997 Rate-based construction of kinetic models for complex systems *J. Phys. Chem.* A **101** 3731–40
[29] Han K, Green W H and West R H 2017 On-the-fly pruning for rate-based reaction mechanism generation *Comput. Chem. Eng.* **100** 1–8
[30] Arús-Pous J, Blaschke T, Ulander S, Reymond J-L, Chen H and Engkvist O 2019 Exploring the GDB-13 chemical space using deep generative models *J. Cheminform.* **11** 20
[31] Gugler S, Janet J P and Kulik H J 2020 Enumeration of *de novo* inorganic complexes for chemical discovery and machine learning *Mol. Syst. Des. Eng.* **5** 139–52
[32] Reymond J-L 2015 The chemical space project *Acc. Chem. Res.* **48** 722–30
[33] Hajduk P J and Greer J 2007 A decade of fragment-based drug design: strategic advances and lessons learned *Nat. Rev. Drug Discovery* **6** 211–9
[34] Hautier G, Jain A, Chen H, Moore C, Ong S P and Ceder G 2011 Novel mixed polyanions lithium-ion battery cathode materials predicted by high-throughput *ab initio* computations *J. Mater. Chem.* **21** 17147–53
[35] Bhowmik A, Castelli I E, Garcia-Lastra J M, Jørgensen P B, Winther O and Vegge T 2019 A perspective on inverse design of battery interphases using multi-scale modelling, experiments and generative deep learning *Energy Storage Mater.* **21** 446–56
[36] Freeze J G, Kelly H R and Batista V S 2019 Search for catalysts by inverse design: artificial intelligence, mountain climbers and alchemists *Chem. Rev.* **119** 6595–612
[37] Gantzer P, Creton B and Nieto-Draghi C 2020 Inverse-QSPR for *de novo* design: a review *Mol. Inform.* **39** 1900087
[38] von Lilienfeld O A, Müller K-R and Tkatchenko A 2020 Exploring chemical compound space with quantum-based machine learning *Nat. Rev. Chem.* **4** 347–58
[39] Born M and Oppenheimer R 1927 Zur quantentheorie der molekeln *Ann. Phys., Lpz.* **389** 457–84
[40] Sutcliffe B T 1992 *The Born-Oppenheimer Approximation* (Springer) pp 19–46
[41] Jensen F 2017 *Introduction to Computational Chemistry* 3rd edn (Wiley)
[42] Cramer C J 2004 *Essentials of Computational Chemistry: Theories and Models* 2nd edn (Wiley)
[43] Halgren T A 1996 Merck molecular force field. I. Basis, form, scope, parameterization and performance of MMFF94 *J. Comput. Chem.* **17** 490–519
[44] Rappé A K, Casewit C J, Colwell K, Goddard W A and Skiff W M 1992 UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations *J. Am. Chem. Soc.* **114** 10024–35
[45] Vanommeslaeghe K *et al* 2010 CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields *J. Comput. Chem.* **31** 671–90
[46] Bannwarth C, Ehlert S and Grimme S 2019 GFN2-xTB—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions *J. Chem. Theory Comput.* **15** 1652–71
[47] Stewart J J P 2007 Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements *J. Mol. Model.* **13** 1173–213
[48] Stewart J J P 2013 Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters *J. Mol. Model.* **19** 1–32
[49] Weber W and Thiel W 2000 Orthogonalization corrections for semiempirical methods *Theor. Chem. Acc.* **103** 495–506
[50] Rupp M, Tkatchenko A, Müller K-R and Von Lilienfeld O A 2012 Fast and accurate modeling of molecular atomization energies with machine learning *Phys. Rev. Lett.* **108** 058301
[51] De S, Bartók A P, Csányi G and Ceriotti M 2016 Comparing molecules and solids across structural and alchemical space *Phys. Chem. Chem. Phys.* **18** 13754–69
[52] Behler J and Parrinello M 2007 Generalized neural-network representation of high-dimensional potential-energy surfaces *Phys. Rev. Lett.* **98** 146401
[53] Faber F A, Christensen A S, Huang B and Von Lilienfeld O A 2018 Alchemical and structural distribution based representation for universal quantum machine learning *J. Chem. Phys.* **148** 241717
[54] Schütt K T, Arbabzadah F, Chmiela S, Müller K R and Tkatchenko A 2017 Quantum-chemical insights from deep tensor neural networks *Nat. Commun.* **8** 13890
[55] Smith J S, Isayev O and Roitberg A E 2017 ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost *Chem. Sci.* **8** 3192–203
[56] Gasteiger J, Groß J and Günnemann S 2020 Directional message passing for molecular graphs *Int. Conf. on Learning Representations*

[57] Batzner S, Musaelian A, Sun L, Geiger M, Mailoa J P, Kornbluth M, Molinari N, Smidt T E and Kozinsky B 2022 E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials *Nat. Commun.* **13** 2453
[58] Satorras V G, Hoogeboom E and Welling M 2021 E(n) equivariant graph neural networks *Proc. 38th Int. Conf. on Machine Learning (18–24 July 2021)* (*Proc. Machine Learning Research* vol 139) ed M Meila and T Zhang (PMLR) pp 9323–32
[59] Frank T, Unke O and Müller K-R 2022 So3krates: equivariant attention for interactions on arbitrary length-scales in molecular systems *Advances in Neural Information Processing Systems* vol 35, ed S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho and A Oh (Curran Associates, Inc.) pp 29400–13
[60] Batatia I, Kovacs D P, Simm G N C, Ortner C and Csanyi G 2022 MACE: higher order equivariant message passing neural networks for fast and accurate force fields *Advances in Neural Information Processing Systems* ed A H Oh, A Agarwal, D Belgrave and K Cho
[61] Unke O T *et al* 2024 Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments *Sci. Adv.* **10** eadn4397
[62] Musaelian A, Batzner S, Johansson A, Sun L, Owen C J, Kornbluth M and Kozinsky B 2023 Learning local equivariant representations for large-scale atomistic dynamics *Nat. Commun.* **14** 579
[63] Thomas N, Smidt T, Kearnes S, Yang L, Li L, Kohlhoff K and Riley P 2018 Tensor field networks: rotation-and translation-equivariant neural networks for 3D point clouds (arXiv:1802.08219)
[64] Noé F, Tkatchenko A, Müller K-R and Clementi C 2020 Machine learning for molecular simulation *Annu. Rev. Phys. Chem.* **71** 361–90
[65] Unke O T, Chmiela S, Sauceda H E, Gastegger M, Poltavsky I, Schütt K T, Tkatchenko A and Müller K-R 2021 Machine learning force fields *Chem. Rev.* **121** 10142–86
[66] Chmiela S, Tkatchenko A, Sauceda H E, Poltavsky I, Schütt K T and Müller K-R 2017 Machine learning of accurate energy-conserving molecular force fields *Sci. Adv.* **3** e1603015
[67] Chmiela S, Sauceda H E, Poltavsky I, Müller K-R and Tkatchenko A 2019 sGDML: constructing accurate and data efficient molecular force fields using machine learning *Comput. Phys. Commun.* **240** 38–45
[68] Chmiela S, Vassilev-Galindo V, Unke O T, Kabylda A, Sauceda H E, Tkatchenko A and Müller K-R 2023 Accurate global machine learning force fields for molecules with hundreds of atoms *Sci. Adv.* **9** eadf0873
[69] Schütt K, Kindermans P-J, Sauceda Felix H E, Chmiela S, Tkatchenko A and Müller K-R 2017 SchNet: a continuous-filter convolutional neural network for modeling quantum interactions *Advances in Neural Information Processing Systems* vol 30, ed I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan and R Garnett (Curran Associates, Inc.) pp 991–1001
[70] Schütt K T, Sauceda H E, Kindermans P-J, Tkatchenko A and Müller K-R 2018 SchNet—a deep learning architecture for molecules and materials *J. Chem. Phys.* **148** 241722
[71] Unke O T and Meuwly M 2019 PhysNet: a neural network for predicting energies, forces, dipole moments and partial charges *J. Chem. Theory Comput.* **15** 3678–93
[72] Unke O T, Chmiela S, Gastegger M, Schütt K T, Sauceda H E and Müller K-R 2021 SpookyNet: learning force fields with electronic degrees of freedom and nonlocal effects *Nat. Commun.* **12** 7273
[73] Hoogeboom E, Satorras V G, Vignac C and Welling M 2022 Equivariant diffusion for molecule generation in 3D *Proc. 39th Int. Conf. on Machine Learning (17–23 July 2022)* (*Proc. Machine Learning Research* vol 162) ed K Chaudhuri, S Jegelka, L Song, C Szepesvari, G Niu and S Sabato (PMLR) pp 8867–87
[74] Wu L, Gong C, Liu X, Ye M and Liu Q 2022 Diffusion-based molecule generation with informative prior bridges *Advances in Neural Information Processing Systems* vol 35, ed S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho and A Oh (Curran Associates, Inc.) pp 36533–45
[75] Huang L, Zhang H, Xu T and Wong K-C 2023 MDM: molecular diffusion model for 3D molecule generation *Proc. AAAI Conf. on Artificial Intelligence* vol 37 pp 5105–12
[76] Xu M, Powers A S, Dror R O, Ermon S and Leskovec J 2023 Geometric latent diffusion models for 3D molecule generation *Proc. 40th Int. Conf. on Machine Learning (23–29 July 2023)* (*Proc. Machine Learning Research* vol 202) ed A Krause, E Brunskill, K Cho, B Engelhardt, S Sabato and J Scarlett (PMLR) pp 38592–610
[77] Peng X, Guan J, Liu Q and Ma J 2023 MolDiff: addressing the atom-bond inconsistency problem in 3D molecule diffusion generation *Proc. 40th Int. Conf. on Machine Learning (23–29 July 2023)* (*Proc. Machine Learning Research* vol 202) ed A Krause, E Brunskill, K Cho, B Engelhardt, S Sabato and J Scarlett (PMLR) pp 27611–29
[78] Xu M, Yu L, Song Y, Shi C, Ermon S and Tang J 2022 GeoDiff: a geometric diffusion model for molecular conformation generation *Int. Conf. on Learning Representations*
[79] Vignac C, Krawczuk I, Siraudin A, Wang B, Cevher V and Frossard P 2023 Digress: discrete denoising diffusion for graph generation *The 11th Int. Conf. on Learning Representations*
[80] Kong L, Cui J, Sun H, Zhuang Y, Prakash B A and Zhang C 2023 Autoregressive diffusion model for graph generation *Proc. 40th Int. Conf. on Machine Learning (23–29 July 2023)* (*Proc. Machine Learning Research* vol 202) ed A Krause, E Brunskill, K Cho, B Engelhardt, S Sabato and J Scarlett (PMLR) pp 17391–408
[81] Gebauer N, Gastegger M and Schütt K 2019 Symmetry-adapted generation of 3D point sets for the targeted discovery of molecules *Advances in Neural Information Processing Systems* vol 32, ed H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox and R Garnett (Curran Associates, Inc.) pp 7566–78
[82] Gebauer N W A, Gastegger M, Hessmann S S P, Müller K-R and Schütt K T 2022 Inverse design of 3D molecular structures with conditional generative neural networks *Nat. Commun.* **13** 973
[83] Simm G, Pinsler R and Hernandez-Lobato J M 2020 Reinforcement learning for molecular design guided by quantum mechanics *Proc. 37th Int. Conf. on Machine Learning (13–18 July 2020)* (*Proc. Machine Learning Research* vol 119) ed H Daumé and A Singh (PMLR) pp 8959–69
[84] Simm G N C, Pinsler R, Csányi G and Hernández-Lobato J M 2021 Symmetry-aware actor-critic for 3D molecular design *Int. Conf. on Learning Representations*
[85] Meldgaard S A, Köhler J, Mortensen H L, Christiansen M-P V, Noé F and Hammer B 2022 Generating stable molecules using imitation and reinforcement learning *Mach. Learn.: Sci. Technol.* **3** 015008
[86] Noé F, Olsson S, Köhler J and Wu H 2019 Boltzmann generators: sampling equilibrium states of many-body systems with deep learning *Science* **365** eaaw1147
[87] Köhler J, Klein L and Noe F 2020 Equivariant flows: exact likelihood generative learning for symmetric densities *Proc. 37th Int. Conf. on Machine Learning (13–18 July 2020)* (*Proc. Machine Learning Research* vol 119) ed H Daumé and A Singh (PMLR) pp 5361–70

[88] Garcia Satorras V, Hoogeboom E, Fuchs F, Posner I and Welling M 2021 E(n) equivariant normalizing flows *Advances in Neural Information Processing Systems* vol 34, ed M Ranzato, A Beygelzimer, Y Dauphin, P Liang and J W Vaughan (Curran Associates, Inc.) pp 4181–92
[89] Klein L, Foong A Y K, Fjelde T E, Mlodozeniec B K, Brockschmidt M, Nowozin S, Noe F and Tomioka R 2023 Timewarp: transferable acceleration of molecular dynamics by learning time-coarsened dynamics *37th Conf. on Neural Information Processing Systems*
[90] Mansimov E, Mahmood O, Kang S and Cho K 2019 Molecular geometry prediction using a deep generative graph neural network *Sci. Rep.* **9** 20381
[91] Simm G and Hernandez-Lobato J M 2020 A generative model for molecular distance geometry *Proc. 37th Int. Conf. on Machine Learning (13–18 July 2020))* (*Proc. Machine Learning Research* vol 119) ed H Daumé and A Singh (PMLR) pp 8949–58
[92] Gogineni T, Xu Z, Punzalan E, Jiang R, Kammeraad J, Tewari A and Zimmerman P 2020 TorsionNet: a reinforcement learning approach to sequential conformer search *Advances in Neural Information Processing Systems* vol 33, ed H Larochelle, M Ranzato, R Hadsell, M Balcan and H Lin (Curran Associates, Inc.) pp 20142–53
[93] Ganea O, Pattanaik L, Coley C, Barzilay R, Jensen K, Green W and Jaakkola T 2021 GeoMol: torsional geometric generation of molecular 3D conformer ensembles *Advances in Neural Information Processing Systems* vol 34, ed M Ranzato, A Beygelzimer, Y Dauphin, P Liang and J W Vaughan Curran Associates, Inc., pp 13757–69
[94] Xu M, Wang W, Luo S, Shi C, Bengio Y, Gomez-Bombarelli R and Tang J 2021 An end-to-end framework for molecular conformation generation via bilevel programming *Proc. 38th Int. Conf. on Machine Learning (18–24 July 2021)* (*Proc. Machine Learning Research* vol 139) ed M Meila and T Zhang (PMLR) pp 11537–47
[95] Lemm D, von Rudorff G F and von Lilienfeld O A 2021 Machine learning based energy-free structure predictions of molecules, transition states and solids *Nat. Commun.* **12** 4468
[96] Jing B, Corso G, Chang J, Barzilay R and Jaakkola T S 2022 Torsional diffusion for molecular conformer generation *Advances in Neural Information Processing Systems* ed A H Oh, A Agarwal, D Belgrave and K Cho
[97] Wang Y, Xu C, Li Z and Barati Farimani A 2023 Denoise pretraining on nonequilibrium molecules for accurate and transferable neural potentials *J. Chem. Theory Comput.* **19** 5077–87
[98] Feng S, Ni Y, Lan Y, Ma Z-M and Ma W-Y 2023 Fractional denoising for 3D molecular pre-training *Proc. 40th Int. Conf. on Machine Learning (23–29 July 2023)* (*Proc. Machine Learning Research* vol 202) ed A Krause, E Brunskill, K Cho, B Engelhardt, S Sabato and J Scarlett (PMLR) pp 9938–61
[99] Zaidi S, Schaarschmidt M, Martens J, Kim H, Teh Y W, Sanchez-Gonzalez A, Battaglia P, Pascanu R and Godwin J 2023 Pre-training via denoising for molecular property prediction *The 11th Int. Conf. on Learning Representations*
[100] Liu S, Guo H and Tang J 2023 Molecular geometry pretraining with SE(3)-invariant denoising distance matching *The 11th Int. Conf. on Learning Representations*
[101] Godwin J, Schaarschmidt M, Gaunt A L, Sanchez-Gonzalez A, Rubanova Y, Veličković P, Kirkpatrick J and Battaglia P 2022 Simple GNN regularisation for 3D molecular property prediction and beyond *Int. Conf. on Learning Representations*
[102] Vincent P, Larochelle H, Bengio Y and Manzagol P-A 2008 Extracting and composing robust features with denoising autoencoders *Proc. 25th Int. Conf. on Machine Learning (ICML '08)* (Association for Computing Machinery) pp 1096–103
[103] Hsu T, Sadigh B, Bertin N, Park C W, Chapman J, Bulatov V and Zhou F 2022 Score-based denoising for atomic structure identification (arXiv:2212.02421)
[104] Hoja J, Medrano Sandonas L, Ernst B G, Vazquez-Mayagoitia A, DiStasio R A and Tkatchenko A 2021 QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules *Sci. Data* **8** 43
[105] Mortazavi M, Brandenburg J G, Maurer R J and Tkatchenko A 2018 Structure and stability of molecular crystals with many-body dispersion-inclusive density functional tight binding *J. Phys. Chem. Lett.* **9** 399–405
[106] Seifert G, Porezag D and Frauenheim T 1996 Calculations of molecules, clusters and solids with a simplified LCAO-DFT-LDA scheme *Int. J. Quantum Chem.* **58** 185–92
[107] Elstner M, Porezag D, Jungnickel G, Elsner J, Haugk M, Frauenheim T, Suhai S and Seifert G 1998 Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties *Phys. Rev.* B **58** 7260–8
[108] Gaus M, Cui Q and Elstner M 2011 DFTB3: extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB) *J. Chem. Theory Comput.* **7** 931–48
[109] Tkatchenko A, DiStasio R A, Car R and Scheffler M 2012 Accurate and efficient method for many-body van der Waals interactions *Phys. Rev. Lett.* **108** 236402
[110] Ambrosetti A, Reilly A M, DiStasio R A and Tkatchenko A 2014 Long-range correlation energy calculated from coupled atomic response functions *J. Chem. Phys.* **140** 18A508
[111] Blum L C and Reymond J-L 2009 970 million druglike small molecules for virtual screening in the chemical Universe database GDB-13 *J. Am. Chem. Soc.* **131** 8732–3
[112] Adamo C and Barone V 1999 Toward reliable density functional methods without adjustable parameters: the PBE0 model *J. Chem. Phys.* **110** 6158–70
[113] Perdew J P, Ernzerhof M and Burke K 1996 Rationale for mixing exact exchange with density functional approximations *J. Chem. Phys.* **105** 9982–5
[114] Blum V, Gehrke R, Hanke F, Havu P, Havu V, Ren X, Reuter K and Scheffler M 2009 *Ab initio* molecular simulations with numeric atom-centered orbitals *Comput. Phys. Commun.* **180** 2175–96
[115] Ren X, Rinke P, Blum V, Wieferink J, Tkatchenko A, Sanfilippo A, Reuter K and Scheffler M 2012 Resolution-of-identity approach to Hartree–Fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions *New J. Phys.* **14** 053020
[116] Sohl-Dickstein J, Weiss E, Maheswaranathan N and Ganguli S 2015 Deep unsupervised learning using nonequilibrium thermodynamics *Proc. 32nd Int. Conf. on Machine Learning (Lille, France, 7–9 July 2015)* (*Proc. Machine Learning Research* vol 37) ed F Bach and D Blei (PMLR) pp 2256–65
[117] Ho J, Jain A and Abbeel P 2020 Denoising diffusion probabilistic models *Advances in Neural Information Processing Systems* vol 33, ed H Larochelle, M Ranzato, R Hadsell, M Balcan and H Lin (Curran Associates, Inc.) pp 6840–51
[118] Song Y, Sohl-Dickstein J, Kingma D P, Kumar A, Ermon S and Poole B 2021 Score-based generative modeling through stochastic differential equations *Int. Conf. on Learning Representations*
[119] Schütt K, Unke O and Gastegger M 2021 Equivariant message passing for the prediction of tensorial properties and molecular spectra *Proc. 38th Int. Conf. on Machine Learning (18–24 July 2021)* (*Proc. Machine Learning Research* vol 139) ed M Meila and T Zhang (PMLR) pp 9377–88

[120] Bishop C M 2006 *Pattern Recognition and Machine Learning* (*Information Science and Statistics*) (Springer)
[121] Schütt K, Kessel P, Gastegger M, Nicoli K, Tkatchenko A and Muller K-R 2018 SchNetPack: a deep learning toolbox for atomistic systems *J. Chem. Theory Comput.* **15** 448–55
[122] Schütt K T, Hessmann S S P, Gebauer N W A, Lederer J and Gastegger M 2023 SchNetPack 2.0: a neural network toolbox for atomistic machine learning *J. Chem. Phys.* **158** 144801
[123] O'Boyle N M, Banck M, James C A, Morley C, Vandermeersch T and Hutchison G R 2011 Open babel: an open chemical toolbox *J. Cheminform.* **3** 33
[124] Larsen A H *et al* 2017 The atomic simulation environment—a Python library for working with atoms *J. Phys.: Condens. Matter* **29** 273002
[125] Weigend F and Ahlrichs R 2005 Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: design and assessment of accuracy *Phys. Chem. Chem. Phys.* **7** 3297–305
[126] Sun Q 2015 Libcint: an efficient general integral library for Gaussian basis functions *J. Comput. Chem.* **36** 1664–71
[127] Sun Q *et al* 2017 PYSCF: the Python-based simulations of chemistry framework *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **8** e1340
[128] Sun Q *et al* 2020 Recent developments in the PYSCF program package *J. Chem. Phys.* **153** 024109
[129] Kahouli K, Hessmann S S P, Mueller K-R, Nakajima S, Gugler S and Gebauer N W A 2024 MoreRed: molecular relaxation by reverse diffusion with time step prediction *Zenodo* https://doi.org/10.5281/zenodo.10927872
[130] Ruddigkeit L, Van Deursen R, Blum L C and Reymond J-L 2012 Enumeration of 166 billion organic small molecules in the chemical Universe database GDB-17 *J. Chem. Inf. Model.* **52** 2864–75
[131] Ramakrishnan R, Dral P O, Rupp M and von Lilienfeld O A 2014 Quantum chemistry structures and properties of 134 kilo molecules *Sci. Data* **1** 140022
[132] Krizhevsky A 2009 Learning multiple layers of features from tiny images (available at: www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf)
[133] Weininger D 1988 SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules *J. Chem. Inf. Comput. Sci.* **28** 31–36
[134] Karras T, Aittala M, Aila T and Laine S 2022 Elucidating the design space of diffusion-based generative models *Advances in Neural Information Processing Systems* ed A H Oh, A Agarwal, D Belgrave and K Cho
[135] Loshchilov I and Hutter F 2019 Decoupled weight decay regularization *Int. Conf. on Learning Representations*