Generation of conformational ensembles of small molecules via surrogate model-assisted molecular dynamics

Generation of conformational ensembles of small molecules via surrogate model-assisted molecular dynamics

IOP Publishing Mach. Learn.: Sci. Technol. 5 (2024) 025010 https://doi.org/10.1088/2632-2153/ad3b64

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
24 November 2023

REVISED
6 February 2024

ACCEPTED FOR PUBLICATION
3 April 2024

PUBLISHED
15 April 2024

PAPER

Generation of conformational ensembles of small molecules via surrogate model-assisted molecular dynamics

Juan Viguera Diez[1,2], **Sara Romeo Atance**[1,2], **Ola Engkvist**[1,2] **and Simon Olsson**[1,*]

[1] Department of Computer Science and Engineering, Chalmers University of Technology, Rännvägen 6, 412 58 Göteborg, Sweden
[2] Molecular AI, Discovery Sciences, R&D, AstraZeneca Gothenburg, Pepparedsleden 1, 431 50 Mölndal, Sweden
[*] Author to whom any correspondence should be addressed.

E-mail: simonols@chalmers.se

Keywords: generative models, Boltzmann distribution, molecular conformation generation, molecular dynamics, property prediction, equilibrium sampling

Supplementary material for this article is available online

Abstract

The accurate prediction of thermodynamic properties is crucial in various fields such as drug discovery and materials design. This task relies on sampling from the underlying Boltzmann distribution, which is challenging using conventional approaches such as simulations. In this work, we introduce surrogate model-assisted molecular dynamics (SMA-MD), a new procedure to sample the equilibrium ensemble of molecules. First, SMA-MD leverages deep generative models to enhance the sampling of slow degrees of freedom. Subsequently, the generated ensemble undergoes statistical reweighting, followed by short simulations. Our empirical results show that SMA-MD generates more diverse and lower energy ensembles than conventional MD simulations. Furthermore, we showcase the application of SMA-MD for the computation of thermodynamical properties by estimating implicit solvation free energies.

1. Introduction

Accurately predicting molecular properties is an important task with applications across the sciences. Some prominent examples are drug discovery and material design. Estimating such properties relies on sampling from the underlying Boltzmann distribution. However, generating unbiased and independent samples from the Boltzmann distribution efficiently remains a challenging open problem.

Currently, molecular dynamics (MD) [1] and Markov chain Monte Carlo [2] are the key techniques to draw samples from the Boltzmann distribution. While these techniques asymptotically generate samples from the Boltzmann distribution, many simulation steps are often needed to generate just one independent sample. This problem is particularly prescient for high-dimensional and meta-stable molecular systems. Despite their limitations, these techniques are widely used, especially in combination with enhanced sampling methods [3], which offer different strategies to speed up the generation of independent samples. Important enhanced sampling methods include replica-based approaches [4, 5], flooding [6], meta-dynamics [7], and umbrella sampling [8].

With the advent of deep generative models (DGMs) [9–13], a family of new methods to generate unbiased one-shot equilibrium samples of the Boltzmann distribution were proposed under the name of *Boltzmann Generators* (BG) [14–17]. These methods approximate the Boltzmann distribution of a molecular system with a DGM which allows efficient sampling and exact likelihood evaluation, commonly Normalizing Flows [11]. Efficient sampling allows to side-step iterative simulation methods, and exact likelihood evaluation allows to recover unbiased samples through importance sampling or importance weighing [14]. Previous work has successfully used BGs to sample from large molecules such as proteins [14] or solids [18]. However, these approaches currently do not generalize to different molecular systems, and designing models that are transferable across different molecules remains a challenging task. Another family of recent related

methods, such as implicit transfer operator learning (ITO) [19] or Timewarp [20], tackles the sampling problem by modeling the generative process in MD simulations, yet on much longer time-scales. Nevertheless, currently, these approaches similarly suffer from limited transferability.

Other approaches focus on enumerating the local minima of a potential energy function, so-called conformers. Different architectures have been proposed, such as CGVAE [21], GeoMol [22], or GeoDiff [23]. These methods are transferable across different molecular systems, however, as the generated states represent local minima of potential energy, they are unable to capture entropic effects due to thermal fluctuations, which makes them unsuitable for computing many molecular properties.

Large-scale conformational rearrangements in molecules can be represented by changes in torsion angles (dihedrals), which correspond to rotations occurring around flexible bonds. For example, the structures of biomolecules such as proteins or RNA are compared and analyzed in terms of such angles, using the Ramachandran plot and probabilistic models of local structure [24, 25] as prominent examples. Transitions between different conformations usually account for the slowest processes in simulations. Therefore, generating representative ensembles of torsions is time-consuming and challenging. For this reason, models focusing on torsion angles are useful means to conformational sampling. Recent work encoding conformations in small molecules using torsions include GeoMol [22], Torsional Diffusion [26], Tora3D [27], and VonMisesNet [28]. Apart from capturing major conformational changes, torsion angles are attractive as they reduce the dimensionality of conformational space and are intrinsically invariant to rigid body symmetries. However, even if some of these methods can generate equilibrium samples, they still cannot model stochastic fluctuations in the local structure, and their evaluation as surrogates of the Boltzmann distribution is limited.

In this work, we present surrogate model-assisted molecular dynamics (SMA-MD), a method for generating equilibrium ensembles of molecules. In SMA-MD, generative models are used to sample a diverse ensemble of initial conditions for short molecular simulations. SMA-MD follows a two-step procedure: first, a generative model mixes efficiently across degrees of freedom which exchange slowly in molecular simulations. Second, we reweight samples against the Boltzmann distribution and run short MD simulations to equilibrate the local structure and ensure sampling statistics are unbiased with respect to the target Boltzmann distribution. In this manner, SMA-MD is able to capture entropic effects occurring in all degrees of freedom in a molecule, which are critical for the computation of thermodynamic quantities such as free energy differences. We implement SMA-MD using torsional surrogate models and restrict ourselves to working with small non-cyclic molecules. We probe our method by measuring geometric and thermodynamical (potential and implicit solvation free energies) properties and comparing them to the baseline of classic MD simulations, used as the source of training data. We empirically show that our method can generate diverse and physically realistic ensembles. Equilibrium ensembles generated with SMA-MD present higher conformational coverage and lower average energy than those obtained with conventional MD simulations of similar runtime, closely matching long replica exchange (RE) simulations.

Our main contributions are:

- Introducing SMA-MD: a new approach that combines generative models for slow degrees of freedom with statistical reweighting and short simulations to produce equilibrium conformational ensembles for molecules.
- Evaluation of our method by comparison with ensembles generated by MD and RE simulations.
- Generating a new dataset: consisting of MD simulations of 12530 k non-cyclic small molecules that we use as training data, along with data splits for benchmarking.
- Showcasing a downstream application: estimating relevant observables such as geometrical quantities and free energies of solvation.

2. Methods

2.1. Sampling and molecular properties

Experimental observables (molecular properties) often correspond to averages over the ensemble of 3D arrangements (conformations, $\mathbf{x}$) molecules can adopt, which follow the Boltzmann distribution,

$$\mu(\mathbf{x}) = \mathcal{Z}^{-1} \exp\left(-\beta U(\mathbf{x})\right), \text{with } \mathcal{Z} = \int \mathrm{d}\mathbf{x} \exp\left(-\beta U(\mathbf{x})\right), \tag{1}$$

where $U(\mathbf{x})$ is the potential energy and β is the inverse temperature. We conveniently define the reduced potential $u(x) = \beta U(\mathbf{x})$. Given independent and identically distributed (i.i.d.) conformations sampled from the Boltzmann distribution, thermodynamical quantities can be computed with the *Monte Carlo estimator*:

$$O = \mathbb{E}_{\mathbf{x}\sim\mu(\mathbf{x})}\left[o\left(\mathbf{x}\right)\right] \approx \frac{1}{N}\sum_{i=1}^{N} o\left(\mathbf{x}_i\right), \;\; \mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mu\left(\mathbf{x}\right), \tag{2}$$

where $o(\mathbf{x})$ computes the microscopic contribution of a property for a conformation. Generating i.i.d. samples from the Boltzmann distribution is challenging. However, if a surrogate model $p(\mathbf{x})$ overlaps with $\mu(\mathbf{x})$, unbiased Boltzmann-distributed samples can be generated by reweighting the surrogate distribution [14]. This reweighing or resampling, can be achieved with various algorithms, but the simplest one is known as importance sampling, in which samples from $p(\mathbf{x})$ are assigned statistical weight $w = e^{-u(\mathbf{x})}/p(\mathbf{x})$.

2.2. Boltzmann surrogate model

A Boltzmann surrogate is a generative model trained to generate samples from the Boltzmann distribution. In this work, we consider models that generalize across different molecular systems. We use a torsional generative model and therefore, we structure the generation process in the following two sub-steps.

2.2.1. Local structure generation

We define the local structure of a non-terminal atom as the relative geometry of the atoms connected to it. The local structure of a given atom can be specified as a set of internal coordinates (distances, angles, and dihedral angles). In this work, we restrict ourselves to molecules that do not have rings as cyclic molecules would require special considerations. Local structures around non-terminal atoms are very constrained and highly dependent on the hybridization of the central atom. For this reason, we use a simple method for generating the local structure of molecules:

- Distances and angles are set to the equilibrium force field parameters.
- Dihedral angles are chosen based on the hybridization of the central atom. For example, if an atom is sp^2 hybridized, its local structure will be planar but if it is sp^3 hybridized, it will follow a tetrahedron shape. We provide further details in supplementary material.

2.2.2. Rotatable bond generation: Torsional Diffusion

In this work, following conventions from previous contributions in the context of small molecules [22, 26], we consider a bond to be rotatable if it connects two non-terminal atoms in the chemical graph. In this step, we generate the remaining degrees of freedom, the torsion angles of rotatable bonds, using a DGM to model potentially complex multi-modal distributions. We choose this DGM to be a diffusion model [12, 13, 29]. Torsion angles lie on the circle and therefore the set of torsions of rotatable bonds within a molecule lies on a hyper-torus. Previous work adapted the formalism of diffusion models to operate on this Riemannian manifold [30]. Based on this work, Torsional Diffusion [26], a diffusion model tailored for modeling torsions, was proposed. One of the main innovations behind Torsional Diffusion is the use of 3D-aware torsional updates which are invariant to the choice of reference atoms. Moreover, a new model architecture exploiting symmetries around rotatable bonds, the pseudo-torque layer, was introduced. Because of these beneficial characteristics and excellent performance in similar tasks, in this work, we use Torsional Diffusion as a model for torsion angles around rotatable bonds.

2.2.3. Training

Since the local structure generation module has no learnable parameters, in this section we solely elaborate on how we train the rotatable bonds model. We train Torsional Diffusion models against a set of target conformations with torsions $\tau_0 \sim p_0$. During training, noise with intensity t is added, resulting in noisy torsions τ_t. Model optimization is performed by minimizing the denoising score matching loss [13, 31],

$$J_{DSM}\left(\theta\right) = \mathbb{E}_t\left[\lambda\left(t\right)\mathbb{E}_{\boldsymbol{\tau}_0\sim p_0, \boldsymbol{\tau}_t\sim p_{t|0}(\cdot|\,\boldsymbol{\tau}_0)}\left[||s\left(\boldsymbol{\tau}_t, t\right) - \nabla_{\boldsymbol{\tau}_t}\log p_{t|0}\left(\boldsymbol{\tau}_t|\boldsymbol{\tau}_0\right)||^2\right]\right], \tag{3}$$

where noise level t is uniformly sampled, $\lambda(t) = 1/\mathbb{E}_{\boldsymbol{\tau}\sim p_{t|0}(\cdot|0)}\left[||\nabla_{\boldsymbol{\tau}_t}\log p_{t|0}(\boldsymbol{\tau}|\mathbf{0})||^2\right]$, $s(\boldsymbol{\tau}_t, t)$ is the neural network prediction for the score and $p_{t|0}(\boldsymbol{\tau}_t|\boldsymbol{\tau}_0)$ is the perturbation kernel. More details about training can be found in supplementary material.

Please note that, if we directly trained a model using molecular conformations obtained from simulations (or other methods), we would create a distribution shift between training and inference. Therefore, to compensate for this, during training, we substitute the local structures from target conformations with the ones generated by our local structure model.

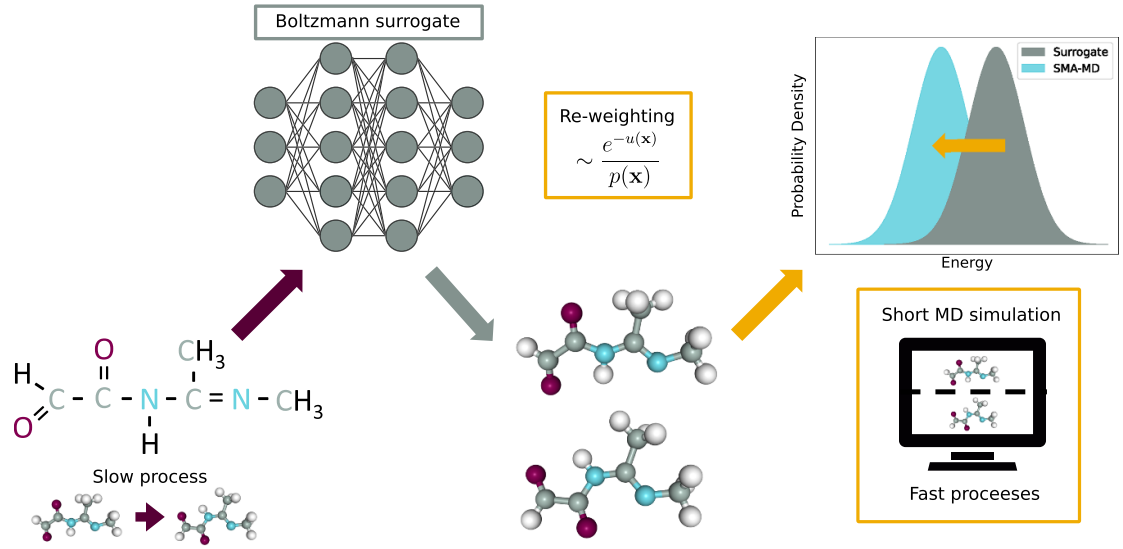

Figure 1. The SMA-MD workflow: starting from the chemical graph describing a molecule, SMA-MD generates its conformational equilibrium ensemble. First, we generate conformations using a generative model trained to generate samples from the Boltzmann distribution. Then, we reweight against the Boltzmann distribution and run short MD simulations. The generated ensembles can then be used for computing downstream properties of interest.

2.3. SMA-MD

In this work, we present SMA-MD as an approach to efficiently sample from the Boltzmann distribution. The overall workflow of SMA-MD is summarized in figure 1 and consists of two main steps. In the first step, we use a generative model trained to emulate the Boltzmann distribution of molecules (the Boltzmann surrogate) to generate conformations. These conformations are constructed in two sub-steps: starting by generating local geometries with a deterministic algorithm and then sampling the torsion angles using a DGM to mix across the molecules' 'slow degrees of freedom'. Next, in the second step, we use short parallel MD simulations to thermalize and mix the fast degrees of freedom. Combining simulations with an exact reweighting scheme enables us to generate unbiased samples from the Boltzmann distribution of the molecule. These ensembles can then be used to compute thermodynamic properties of interest.

2.3.1. Sampling from the surrogate model

We first generate the degrees of freedom corresponding to the local structure and then we sample the torsions, $\boldsymbol{\tau}$. The torsions are sampled through integration of the probability flow (neural) ODE (ordinary differential equation), corresponding to the Torsional Diffusion model, which further enables exact reweighting through sample likelihood calculation. We denote $p_0(\boldsymbol{\tau})$ the neural ODE sampler likelihood in torsional space. However, the Boltzmann distribution is generally specified as a function of the 3D coordinates of the atoms in a molecule, $\mathbf{x}$. Therefore, to allow for compatibility with the Boltzmann measure, we need to express $p_0(\boldsymbol{\tau})$ in Euclidean space instead. As all the generated geometric quantities correspond to internal coordinates, and the local structure is generated deterministically, the Euclidean likelihood can be computed as

$$p(\mathbf{x}) = p_0(\boldsymbol{\tau}) / |\det(J_{\text{int}\to\text{euc}}(\mathbf{x}))|, \tag{4}$$

where $J_{\text{int}\to\text{euc}}(\mathbf{x})$ is the Jacobian of the transformation from internal to Euclidean coordinates. For details about the sampling procedure, see supplementary material.

2.3.2. Reweighting and MD fine-tuning

After generating molecular conformations with the surrogate model, we post-process them by taking two extra steps which we have observed to be critical to generate physically realistic structures.

2.3.2.1. Reweighting according to Boltzmann weights

Given a set of sampled conformers, we compute per-sample weights with

$$w(\mathbf{x}) = \frac{e^{-u(\mathbf{x})}}{p(\mathbf{x})}. \tag{5}$$

We note that reweighting only guarantees improvement in the case of complete domain coverage, which is not fulfilled in general. However, we observe that the surrogate model tends to broadly cover the domain, even generating high-energy metastable conformations. One important role of the reweighting step is ruling out these states as we illustrate in supplementary material.

2.3.2.2. MD fine-tuning: short parallel simulations

Generally, we expect reweighting to align the generated ensemble more closely to the target Boltzmann distribution. However, the ensemble still misses two fundamental ingredients in our implementation: stochastic fluctuations in the local structure and the coupling of these fluctuations with the rotatable bonds. We hypothesize that these two features can be recovered in short simulations. Therefore, we run parallel simulations on the reweighted ensemble using the REFORM library extending OpenMM [32, 33]. For the experiments performed in this work, we run 1 *ns* simulations per sample.

2.4. The *MDQM9-nc* dataset

We generated approximately Boltzmann-distributed samples using MD by simulating 12 530 non-cyclic molecules from the Quantum Machine 9 (QM9) dataset [34] in vacuum and room temperature using the GAFF force field [35]. Thus, we call our dataset *MDQM9-nc*. We carried out these simulations using the `openmmforcefields` [36] and OpenMM packages [33]. All initial conditions were generated by energy minimizing the QM9 geometry in the corresponding GAFF force field. We sampled different molecules proportionally to their number of heavy atoms with a median sampling time of 36.5 ns. Moreover, for 100 molecules from the test set (10%), we run longer 100 ns RE simulations. These long RE simulations are used as ground truth in our experiments. The *MDQM9-nc* dataset is available at https://github.com/olsson-group/mdqm9-nc-loaders. We provide further details about the dataset generation and training, validation, and test splits in supplementary material.

3. Results and discussion

3.1. The impact of the different components of SMA-MD for sampling equilibrium conformations of molecular systems

We showcase the contribution of the different components of SMA-MD by analyzing state populations and potential energies for a molecule in the test set at the two different sampling stages (sampling from the surrogate model and applying post-processing). To find the slowest transitions between metastable states, we use time-lagged independent component analysis (TICA) [37], a linear dimensionality reduction technique that identifies the linear combinations of molecular features that maximizes the autocorrelation. In figure 2(a) we observe that the Boltzmann surrogate captures the relevant states, potentially transitioning to each other in slow processes. Moreover, reweighting and running short simulations adjust the populations and reduce the energy of the ensemble, matching ground-truth long RE simulations figure 2(b).

3.2. SMA-MD generates similar local structures to MD

As the marginal distributions of the degrees of freedom in the local structure are often unimodal, in order to compare local structures generated by different methods, we compute the mean absolute error (MAE) of the estimated mean and standard deviation of these distributions. In table 1 we observe that local structures generated by SMA-MD and MD exhibit remarkable similarity.

3.3. SMA-MD outperforms MD in conformer generation

As previously introduced, a conformer is a local minimum in the molecular energy landscape. Therefore, they are also local maxima of the density landscape. Given an ensemble, we can extract its conformers by finding these local minima. Following this logic, we retrieve conformers from ensembles generated by SMA-MD and MD.

To obtain a set of conformers from an ensemble, we first find modes in the marginal distributions of torsions. Then, we check how many of the possible combinations of modes in the marginal distributions appear in the ensemble. Next, we refine the set of extracted conformers eliminating duplicates corresponding to atom permutations or global inversions of the geometry. Finally, we check that all conformers satisfy a dissimilarity threshold in their relative root mean square deviation (RMSD). We provide additional details about this procedure in supplementary material.

In table 2 we show the average minimum RMSD (AMR) and coverage (COV) of our method in precision and recall modes, see supplementary material for precise definitions. We report remarkable agreement of SMA-MD with RE, clearly outperforming the training-set-like MD trajectories (MD (15 min)). Because the runtime of SMA-MD (using a single GPU) is longer (20 min) than that of the training set simulations

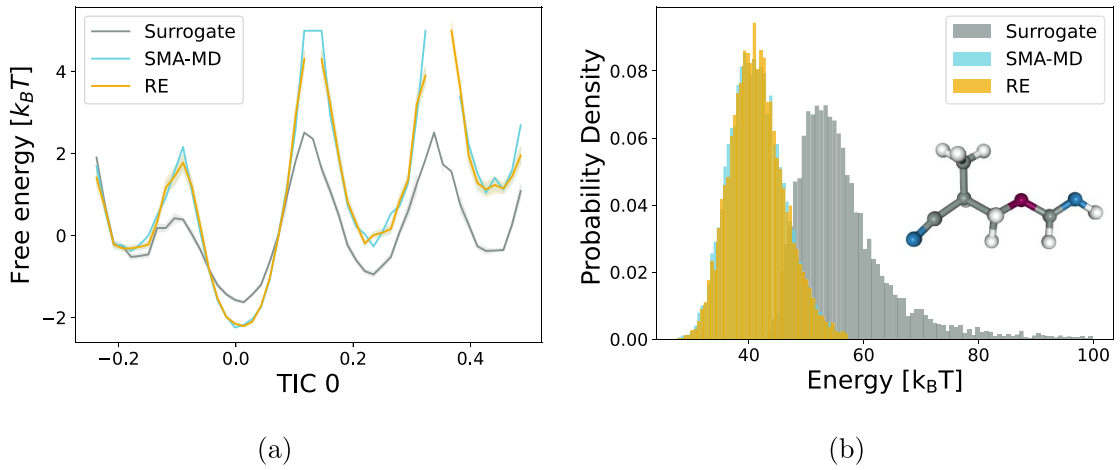

Figure 2. Free energy of slowest TICA component (a) and probability density of ensemble energies (b). The shaded areas in (a) correspond to the standard deviation of ten-fold cross-validation. The Boltzmann surrogate captures the relevant states, potentially transitioning to each other in slow processes. Moreover, reweighting and running short simulations adjust the populations and reduce the energy of the ensemble, matching ground-truth long RE simulations.

Table 1. Average relative difference of distribution parameters in the local structure between generated and MD for all molecules in the test set. Errors are not shown for being smaller than the last digit.

Distances		Angles		Dihedrals	
Average (%)	Std (%)	Average (%)	Std (%)	Average (%)	Std (%)
0.04	2.6	0.13	1.3	0.14	1.4

Table 2. Average minimum root square deviation (AMR) and coverage (COV) of the ensembles generated by SMA-MD and MD simulations of different runtimes against replica exchange. For COV, the threshold is set to $\delta = 0.75$ Å. Bold values indicate best results.

	MD (15 min)		SMA-MD (20 min)		MD (20 min)	
	Precision	Recall	Precision	Recall	Precision	Recall
AMR (Å,↓)	0.14 ± 0.02	0.24 ± 0.03	$\mathbf{0.08 \pm 0.01}$	$\mathbf{0.10 \pm 0.02}$	0.13 ± 0.01	0.21 ± 0.02
COV (↑)	0.95 ± 0.02	0.87 ± 0.02	$\mathbf{0.98 \pm 0.01}$	$\mathbf{0.96 \pm 0.01}$	0.97 ± 0.01	0.90 ± 0.02

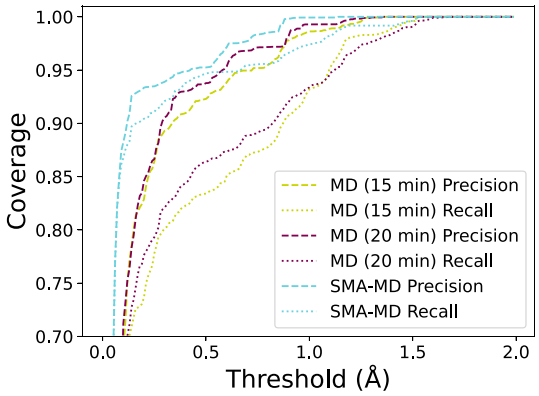

Figure 3. Coverage (precision and recall) as a function of threshold.

(15 min), see supplementary material for details, we further challenged SMA-MD by comparing to longer MD simulations of comparable runtime. We observe that SMA-MD still clearly outperforms these longer simulations (MD (20 min)) by recovering 6% more conformers in the ground-truth ensemble. In figure 3 we compare the Coverage for different threshold values. Here we observe that SMA-MD achieves the best Coverage among the three methods independently of the choice of threshold and the improvement margin becomes greater as we reduce the threshold.

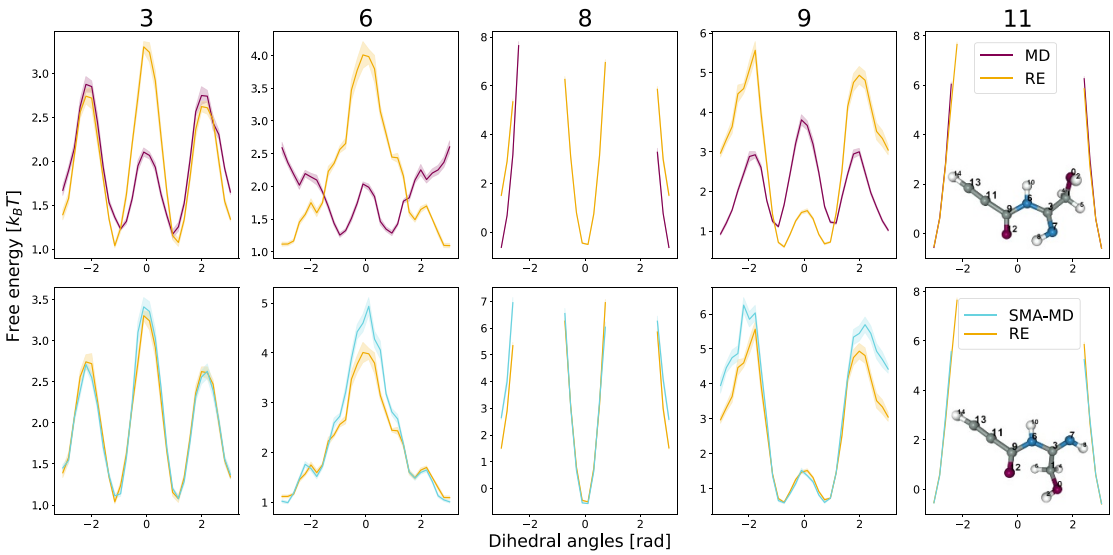

Figure 4. Comparison of dihedral angles of rotatable bonds generated by molecular dynamics simulations of similar runtime to SMA-MD (MD, top) and our method (SMA-MD, bottom) vs replica exchange (RE). The shaded areas correspond to the standard deviation of ten-fold cross-validation. While MD shows different marginal distributions, our method recovers the distributions generated with RE. Torsions of atoms 13 and 14 are omitted for clarity of presentation, having similar marginal distributions under the three methods.

3.4. Covering conformations separated by high free-energy barriers

The conformer generation results above, suggest that our generative model covers the local free energy minima of small molecules well and that we may cover the conformational space of small molecules faster than regular MD simulations. We illustrate that this indeed is the case through the following example. We find SMA-MD samples states present in the long RE simulations separated by barriers that are never overcome in MD simulations of comparable runtime to SMA-MD. In figure 4, we show how SMA-MD samples a state (torsion 8) that is not sampled by MD. Overall, the marginal distributions of MD do not match RE, suggesting the simulated ensembles are not fully equilibrated. However, SMA-MD shows remarkable agreement. The difference in mean ensemble energy w.r.t. to RE is $-0.37\ k_BT$ for SMA-MD and $+3.15\ k_BT$ for MD.

3.5. SMA-MD generates more diverse and energetically favorable ensembles than MD

In section 3.1 we have shown that SMA-MD outperforms conventional MD simulations in conformer generation, presenting better precision and recall w.r.t. ground-truth simulations. Moreover, we have illustrated in section 3.4 how SMA-MD is able to sample across high-energy barriers, where MD falls short.

Furthermore, we show that SMA-MD generates samples in a more statistically efficient manner by comparing the ensemble potential energies. We do this by computing the difference in average ensemble energies and the Jensen-Shanon divergence (Div) between energy histograms. We show in figure 5(a) that SMA-MD generates more similar energy averages and distributions to RE than MD, and that the agreement with RE is very high.

We attribute a number of factors to the improved performance of SMA-MD compared to MD. First, SMA-MD combines data aggregation and post-processing. On the one hand, even if simulations used as training data are not fully converged, it is possible to learn relevant structures from similar molecules. On the other, reweighting and short simulations help to equilibrate the populations and compensate for small deviations in the geometries generated by the surrogate. Combining these two elements, SMA-MD shows robustness against training on biased data (non-converged simulations). Second, SMA-MD does not need initial conditions. One important limitation of MD simulations is their sensitivity to the initial coordinates, often obtained from an experimental crystal structure. Indeed, a main success of Markov state modeling [38] is its ability to use simulation data from different initial conditions to make quantitative predictions. In contrast, SMA-MD not only does not require an initial condition but provides a way of initializing simulations with several different representative initial structures to boost convergence. Third, SMA-MD allows for parallelization. MD simulations are intrinsically sequential, however, all the individual components in SMA-MD allow for parallelization, which makes SMA-MD a more suitable method for modern computing hardware.

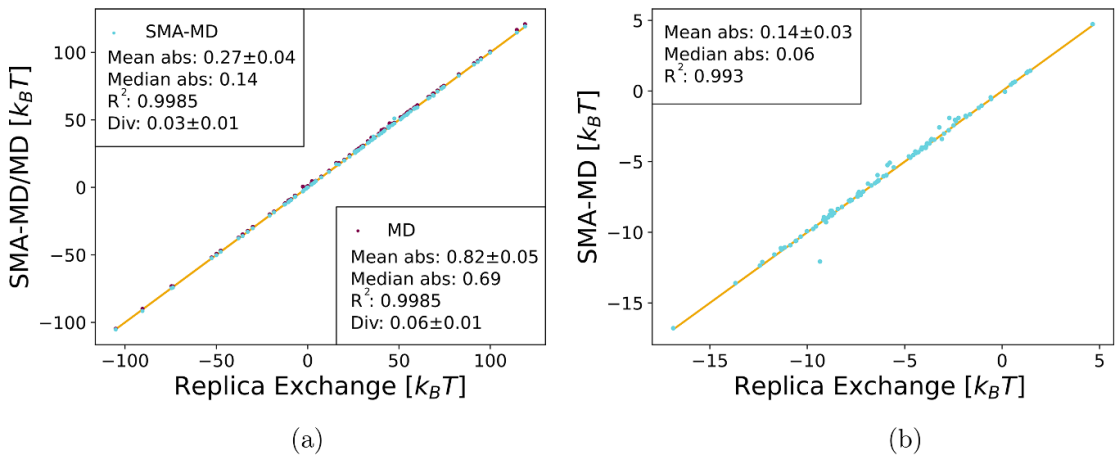

Figure 5. Average potential energy (a) and free energy from vacuum to implicit water (b) of ensembles generated using different methods versus replica exchange. Error bars are smaller than dots. MD here is an MD simulation of comparable runtime to SMA-MD.

3.6. Prediction of molecular properties using SMA-MD: solvation free energy

Finally, after finding great agreement between SMA-MD and RE in the previous analyses, we illustrate how the equilibrium ensembles generated by SMA-MD could be used for downstream tasks by estimating solvation free energies. We use the improved generalized Born model (GB-Neck2) [39] available in OpenMM and experimentally validated with the GAFF force field in previous work [40]. We set an effective number of samples threshold of 100. Details are available in supplementary material. Results in figure 5 (right) show remarkable agreement between the two methods.

4. Limitations and future work

A major bottleneck of SMA-MD remains the computational cost per sample in comparison to MD. However, as previously discussed, while MD is intrinsically sequential, the SMA-MD framework is fully parallelizable (except for the short individual simulations) and therefore allows for an efficient way to use modern computing hardware, including GPUs. Indeed, this divide-and-conquer strategy is successfully applied in the Markov state modeling community [41–44]. Nevertheless, SMA-MD, in the context presented here does not outperform the state-of-the-art RE in terms of runtime, see supplementary material. This limitation is related to the high cost of sampling from the generative model and the low sample efficiency. Due to continual improvements in the field of DGMs, we believe both of these issues will be resolved in the near future.

Currently, SMA-MD is limited to non-cyclic molecules. This limitation comes from the difficulty of generating realistic ring structures when representing molecules in internal coordinates, e.g. distance, angles, and torsions. This problem is particularly prescient for non-aromatic ring structures, including sugars, which are highly restrained but may undergo concerted slow conformational transitions. While heuristics are available to overcome this problem in e.g. rule-based conformer generations, these methods do not readily allow for the extraction of equilibrium statistics. Therefore, we leave accurate modeling of rings for future work. If modeling non-aromatic rings does not entail significant extra computational cost, we foresee that SMA-MD will be scalable to drug-like molecules. Drug-like molecules in the GEOM-Drugs dataset [45] contain 1.7 aromatic rings and 1.3 non-aromatic rings per molecule. Consequently, the number of rotatable bonds in molecules studied here and those of drug-like molecules, 6.3 and 7.9 respectively, remain comparable.

Finally, we consider training Boltzmann surrogates with large-scale data a promising direction. With data sharing taking an increasing priority, simulation data using similar protocols are deposited in scientific data repositories. Although these simulations may not be fully converged, they still contain valuable information to build general transferable Boltzmann surrogates. Indeed, the data we used to train our model is clearly not converged, however, we can still learn a useful surrogate. Mining this data to train Boltzmann surrogates could increase the transferability and therefore the usefulness of these systems.

5. Conclusions

In this work, we have introduced SMA-MD, an efficient method for generating equilibrium conformational ensembles for molecules. SMA-MD combines a transferable surrogate from the Boltzmann distribution with a statistical reweighting and short simulation post-processing step. The goal is to mix between slow degrees of freedom using the surrogate model and refine the ensemble through reweighting and brief simulations. Here, we implement SMA-MD using torsional generative models and show that this method outperforms conventional MD simulations in diversity and ensemble energy. We showcase the applications of this method in downstream tasks by successfully estimating solvation free energies. Even if the work presented here is limited to small non-cyclic molecules, we believe it motivates further research on solving remaining scientific problems and paves the way towards the adoption of similar methods in practical applications, especially drug-discovery pipelines.

Code availability

Code is available at https://github.com/olsson-group/sma-md.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/olsson-group/mdqm9-nc-loaders.

Acknowledgments

The authors thank Rocío Mercado for discussions and early contributions to this work. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The computations in this work were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

ORCID iDs

Juan Viguera Diez https://orcid.org/0000-0002-0526-1766
Sara Romeo Atance https://orcid.org/0000-0002-0922-2350
Ola Engkvist https://orcid.org/0000-0003-4970-6461
Simon Olsson https://orcid.org/0000-0002-3927-7897

References

[1] Hollingsworth S A and Dror R O 2018 Molecular dynamics simulation for all *Neuron* **99** 1129–43
[2] van Ravenzwaaij D, Cassey P and Brown S D 2018 A simple introduction to Markov Chain Monte–Carlo sampling *Psychon. Bull. Rev.* **25** 143–54
[3] Hénin J, Lelièvre T, Shirts M R, Valsson O and Delemotte L 2022 Enhanced sampling methods for molecular dynamics simulations [article v1.0] *Living J. Comput. Mol. Sci.* **4** 1583
[4] Earl D J and Deem M W 2005 Parallel tempering: theory, applications and new perspectives *Phys. Chem. Chem. Phys.* **7** 3910–6
[5] Pasarkar A P, Bencomo G M, Olsson S and Dieng A B 2023 Vendi sampling for molecular simulations: diversity as a force for faster convergence and better exploration *J. Chem. Phys.* **159** 144108
[6] Grubmüller H 1995 Predicting slow structural transitions in macromolecular systems: conformational flooding *Phys. Rev.* E **52** 2893–906
[7] Laio A and Parrinello M 2002 Escaping free-energy minima *Proc. Natl Acad. Sci.* **99** 12562–6
[8] Torrie G and Valleau J 1977 Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling *J. Comput. Phys.* **23** 187–99
[9] Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial networks (arXiv:1406.2661)
[10] Kingma D P and Welling M 2022 Auto-encoding variational bayes (arXiv:1312.6114)
[11] Papamakarios G, Nalisnick E, Rezende D J, Mohamed S and Lakshminarayanan B 2021 Normalizing flows for probabilistic modeling and inference (arXiv:1912.02762)
[12] Ho J, Jain A and Abbeel P 2020 Denoising diffusion probabilistic models (arXiv:2006.11239)
[13] Song Y, Sohl-Dickstein J, Kingma D P, Kumar A, Ermon S and Poole B 2021 Score-based generative modeling through stochastic differential equations (arXiv:2011.13456)
[14] 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
[15] Köhler J, Krämer A and Noé F 2021 Smooth normalizing flows (arXiv:2110.00351)
[16] Dibak M, Klein L, Krämer A and Noé F 2022 Temperature steerable flows and Boltzmann generators (arXiv:2108.01590)
[17] Wu H, Köhler J and Noé F 2020 Stochastic normalizing flows (arXiv:2002.06707)
[18] Köhler J, Invernizzi M, de Haan P and Noé F 2023 Rigid body flows for sampling molecular crystal structures (arXiv:2301.11355)

[19] Schreiner M, Winther O and Olsson S 2023 Implicit transfer operator learning: multiple time-resolution surrogates for molecular dynamics (arXiv:2305.18046)
[20] Klein L, Foong A Y K, Fjelde T E, Mlodozeniec B, Brockschmidt M, Nowozin S, Noé F and Tomioka R 2023 Timewarp: transferable acceleration of molecular dynamics by learning time-coarsened dynamics (arXiv:2302.01170)
[21] Mansimov E, Mahmood O, Kang S and Cho K 2019 Molecular geometry prediction using a deep generative graph neural network *Sci. Rep.* **9** 20381
[22] Ganea O E, Pattanaik L, Coley C W, Barzilay R, Jensen K F, Green W H and Jaakkola T S 2021 Geomol: torsional geometric generation of molecular 3D conformer ensembles (arXiv:2106.07802)
[23] Xu M, Yu L, Song Y, Shi C, Ermon S and Tang J 2022 Geodiff: a geometric diffusion model for molecular conformation generation (arXiv:2203.02923)
[24] Boomsma W, Mardia K V, Taylor C C, Ferkinghoff-Borg J, Krogh A and Hamelryck T 2008 A generative, probabilistic model of local protein structure *Proc. Natl Acad. Sci. USA* **105** 8932–7
[25] Frellsen J, Moltke I, Thiim M, Mardia K V, Ferkinghoff-Borg J and Hamelryck T 2009 A probabilistic model of RNA conformational space *PLoS Comput. Biol.* **5** e1000406
[26] Jing B, Corso G, Chang J, Barzilay R and Jaakkola T 2023 Torsional Diffusion for molecular conformer generation (arXiv:2206.01729)
[27] Zhang Z *et al* 2023 Tora3d: an autoregressive torsion angle prediction model for molecular 3D conformation generation *J. Cheminf.* **15** 57
[28] Swanson K, Williams J and Jonas E 2023 Von mises mixture distributions for molecular conformation generation (arXiv:2306.07472)
[29] Song Y, Durkan C, Murray I and Ermon S 2021 Maximum likelihood training of score-based diffusion models (arXiv:2101.09258)
[30] Bortoli V D, Mathieu E, Hutchinson M, Thornton J, Teh Y W and Doucet A 2022 Riemannian score-based generative modelling (arXiv:2202.02763)
[31] Hyvärinen A 2005 Estimation of non-normalized statistical models by score matching *J. Mach. Learn. Res.* **6** 695–709
[32] Chen Y Replica exchange for openmm: REFORM (available at: https://github.com/noegroup/reform) (Accessed 17 August 2023)
[33] Eastman P*et al* OpenMM: high performance, customizable molecular simulation (available at: https://openmm.org/)
[34] 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
[35] Wang J, Wolf R M, Caldwell J W, Kollman P A and Case D A 2004 Development and testing of a general amber force field *J. Comput. Chem.* **25** 1157–74
[36] Chodera J *et al* OpenMM force fields: amber and charmm force fields for openmm (available at: https://github.com/openmm/openmmforcefields)
[37] Pérez-Hernández G, Paul F, Giorgino T, Fabritiis G D and Noé F 2013 Identification of slow molecular order parameters for Markov model construction *J. Chem. Phys.* **139** 015102
[38] Prinz J-H, Wu H, Sarich M, Keller B, Senne M, Held M, Chodera J D, Schütte C and Noé F 2011 Markov models of molecular kinetics: generation and validation *J. Chem. Phys.* **134** 174105
[39] Nguyen H, Roe D R and Simmerling C 2013 Improved generalized born solvent model parameters for protein simulations *J. Chem. Theory Comput.* **9** 2020–34
[40] Brieg M, Setzler J, Albert S and Wenzel W 2017 Generalized born implicit solvent models for small molecule hydration free energies *Phys. Chem. Chem. Phys.* **19** 1677–85
[41] Chakrabarti K S *et al* 2022 A litmus test for classifying recognition mechanisms of transiently binding proteins *Nat. Commun.* **13** 3792
[42] Noé F, Schütte C, Vanden-Eijnden E, Reich L and Weikl T R 2009 Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations *Proc. Natl Acad. Sci.* **106** 19011–6
[43] Bowman G R, Bolin E R, Hart K M, Maguire B C and Marqusee S 2015 Discovery of multiple hidden allosteric sites by combining Markov state models and experiments *Proc. Natl Acad. Sci.* **112** 2734–9
[44] Olsson S and Noé F 2019 Dynamic graphical models of molecular kinetics *Proc. Natl Acad. Sci.* **116** 15001–6
[45] Axelrod S and Gómez-Bombarelli R 2022 Geom, energy-annotated molecular conformations for property prediction and molecular generation *Sci. Data* **9** 185
[46] Landrum G RDKit: open-source cheminformatics (available at: www.rdkit.org) (Accessed 17 August 2023)
[47] Kingma D P and Ba J 2017 Adam: a method for stochastic optimization (arXiv:1412.6980)