Generative Modeling of Molecular Dynamics Trajectories

Bowen Jing*, Hannes Stark*, Tommi Jaakkola, Bonnie Berger

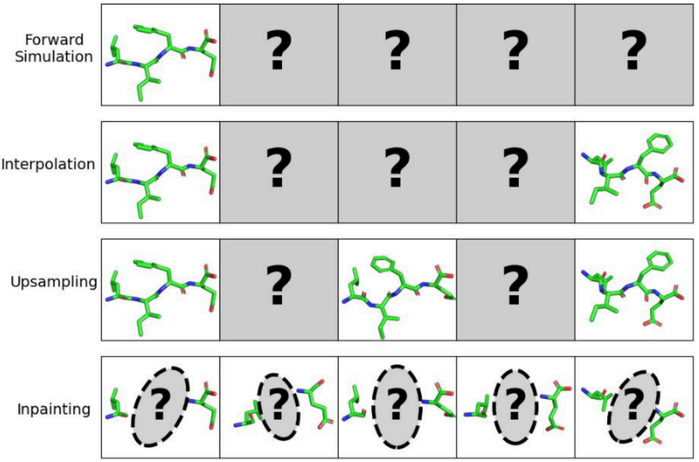
Massachusetts Institute of Technology

TL;DR

Simulating molecular dynamics is a classical method to computationally obtain insights about molecules or to design new ones.

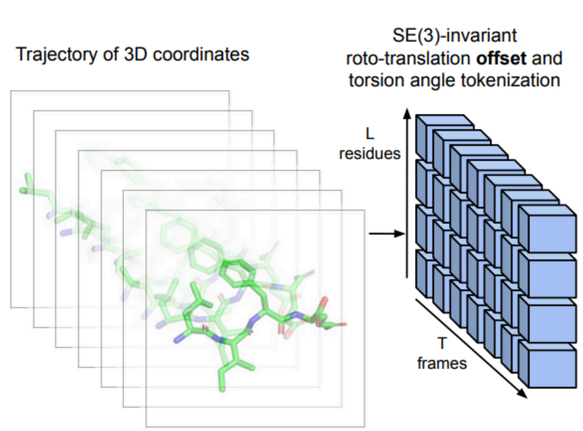
1. Existing generative models for MD sample trajectories frame-by-frame.
2. We sample all frames jointly, similar to video diffusion models.
3. This unlocks new application capabilities next to standard *forward simulation: interpolation, upsampling, and inpainting.*
4. We use a scalable interpolant transformer to achieve all these tasks.
5. We provide experiments generalizing to unseen molecules on tetrapeptides and protein monomers.

MDGen tasks



1. *Forward simulation* —given the initial frame of a trajectory, we sample a potential time evolution of the molecular system.
2. *Interpolation*—given the frames at the two endpoints of a trajectory, we sample a plausible path connecting the two. In chemistry, this is known as transition path sampling and is important for studying reactions and conformational transitions.
3. *Upsampling*—given a trajectory with timestep Δt between frames, we upsample the "framerate" by a factor of M to obtain a trajectory with timestep $\Delta t/M$. This infers fast motions from trajectories saved at less frequent intervals.
4. Inpainting—given part of a molecule and its trajectory, we generate the rest of the molecule (and its time evolution) to be consistent with the known part of the trajectory. This ability could be applied to design molecules to scaffold desired dynamics.

MDGen tokenization and flow model



We tokenize a trajectory as SE(3)-invariant roto-translation offsets from one or more key-frames, which are the frames we condition on for the corresponding taks. Additionally, we have 7 torsion angles per residue.

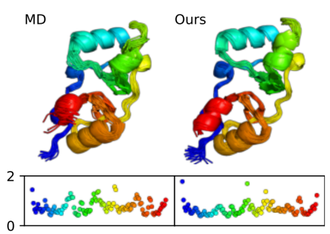
$$\boldsymbol{x}_t^l = ((R, \mathbf{t}), (\psi, \phi, \omega, \chi_1 \dots \chi_4)), \quad \boldsymbol{x} \in \left(\left[SE(3) \times \mathbb{T}^7 \right]^L \right)^T$$

Thus, with L residues and T frames in an MD trajectory, we obtain an array of L x T SE(3) invariant tokens. We generate such tokens using a scalable interpolant transformer based on stochastic interpolants and flow matching.

In the inpainging setting, we additionally generate discrete residues types for each inpainted residue => token dimensionality increases by 20. As discrete flow matching framework to achieve this we choose Dirichlet Flow Matching.

Setting	Key frames	Generate	Conditioned on	Token dim.
Forward simulation	g_1	$g_{1\cdots T}, \boldsymbol{\tau}_{1\cdots T}$	$g_1, \boldsymbol{\tau}_1, A$	21
Interpolation	g_1, g_T	$g_{1\cdots T}, \boldsymbol{\tau}_{1\cdots T}$	$g_{1,T}, \boldsymbol{\tau}_{1,T}, A$	28
Upsampling	g_1	$g_{1\cdots T}, \boldsymbol{\tau}_{1\cdots T}$	$g_{1+\{1,2,\cdots\}M}, \boldsymbol{\tau}_{1+\{1,2,\cdots\}M}, A$	21
Inpainting	g_1, g_T	$g_{1\cdots T}, A$	$g_{1\cdots T}^{\text{known}}$	7 (+20)

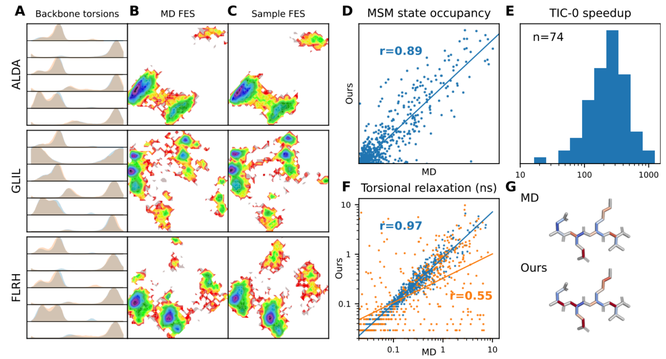
Protein Forward sim.



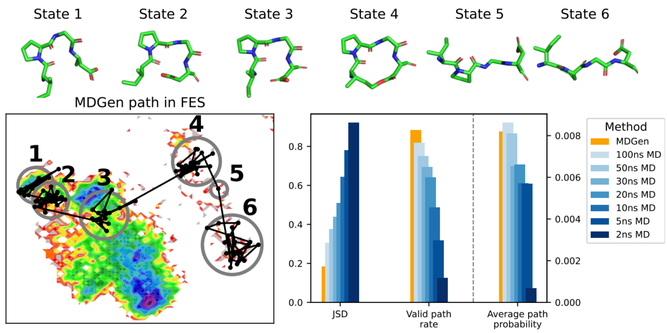
Tetrapeptide forward sim.

C.V.	Ours	10 ns	1 ns	100 ps	100 ns
Torsions (bb)	**.130**	.145	.212	.311	.103
Torsions (sc)	**.093**	.111	.261	.403	.055
Torsions (all)	**.109**	.125	.240	.364	.076
TICA-0	**.230**	.323	.432	.477	.201
TICA-0,1 joint	**.316**	.424	.568	.643	.268
MSM states	**.235**	.363	.493	.527	.208
Runtime	60s	1067s	107s	11s	3h

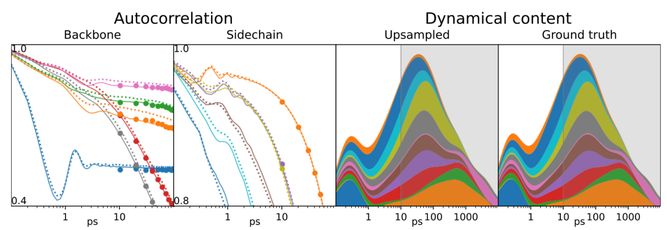
Tetrapeptide forward sim.



Tetrapeptide interpolation



Tetrapeptide upsampling



Tetrapeptide inpainting

Method	High Flux	Random Path
MDGen	**52.1%**	**62.0%**
DynMPNN	17.4%	24.5%
S-MPNN	16.3%	13.5%