UniSim: A Unified Simulator for Time-Coarsened Dynamics of Biomolecules

Ziyang Yu, Wenbing Huang, Yang Liu

清華大學 TSINGHUA UNIVERSITY 1911

RENMIN UNIVERSITY OF CHINA 1937 中国人民大学

Background & Motivation

■ **Molecular Dynamics (MD)** simulations are essential in various fields

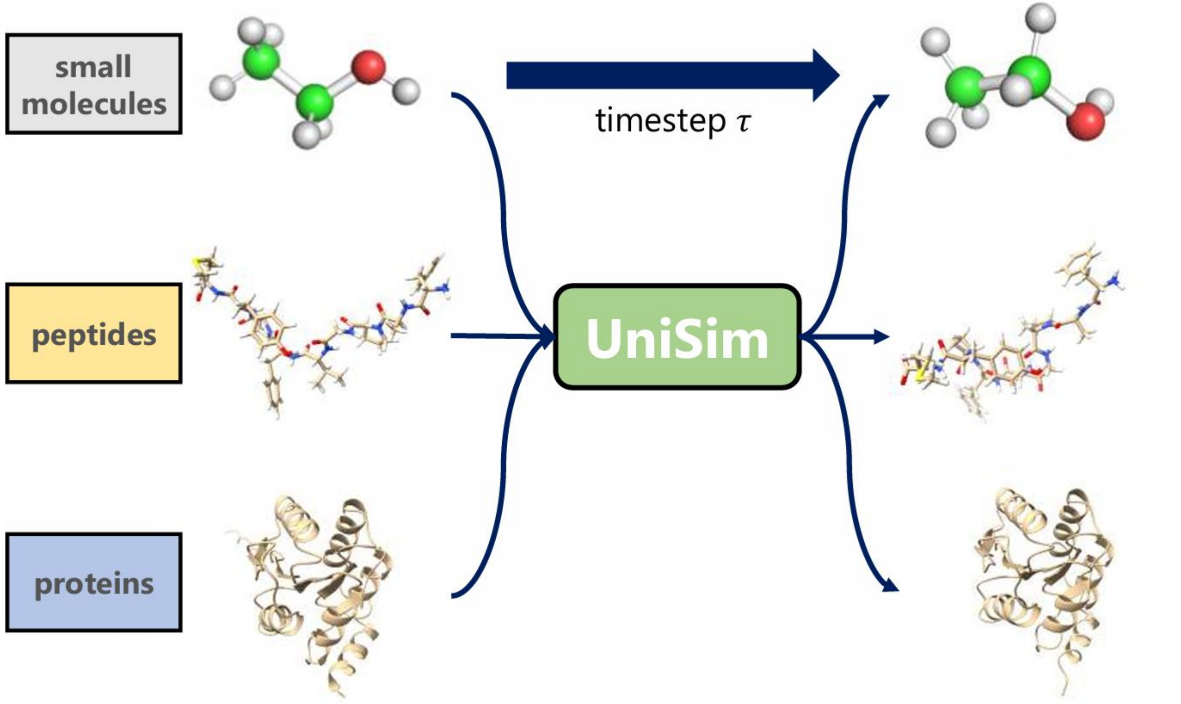
However, current MD methods still struggle with:

- **Traditional Software: Efficiency**
  - Small integration timestep Δt (10^{-15}s) $\longleftrightarrow$ vital biological processes (10^{-3}s)
- **Deep Learning: Transferability**
  - Mostly restricted to a single molecular domain
  - Unable to simulate in different chemical environments

Background & Motivation

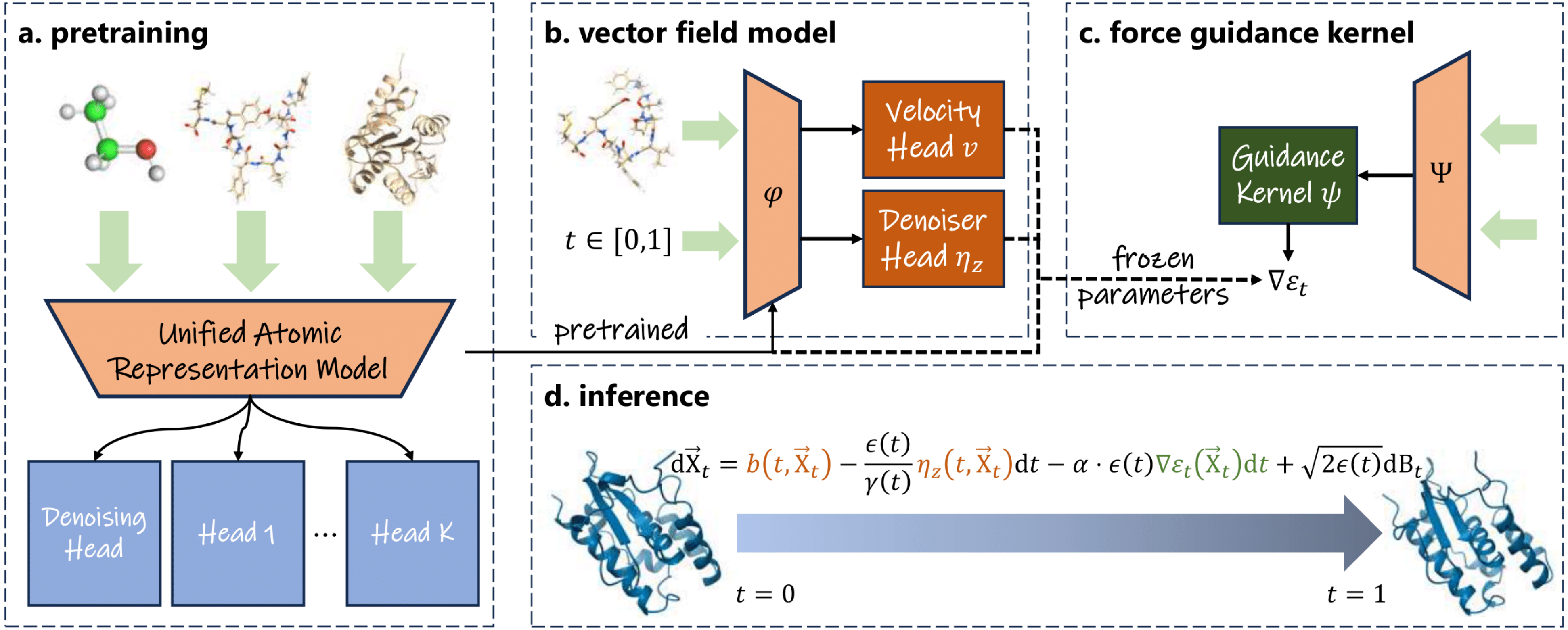
- A better solution requires:
- **Great Efficiency:** time-coarsened dynamics
- **Unified Simulation:** one model for multiple domains
- **Adaptability:** simulations in different environments

Learns the push forward from $\mathbf{X}_t$ to $\mathbf{X}_{t+\tau}$, where $\tau \gg \Delta t$.



Overview

- Unified Representation Model: leverages the cross-domain knowledge from pretraining
- Vector Field Model: follows time-coarsened dynamics using stochastic interpolants
- Force Guidance Kernel: helps adapt to different chemical environments



$$d\vec{X}_t = b(t,\vec{X}_t) - \frac{\epsilon(t)}{\gamma(t)}\eta_z(t,\vec{X}_t)dt - \alpha \cdot \epsilon(t)\nabla\varepsilon_t(\vec{X}_t)dt + \sqrt{2\epsilon(t)}dB_t$$

Challenges & Solutions

■ How to deal with the **scale discrepancy** between molecular systems?

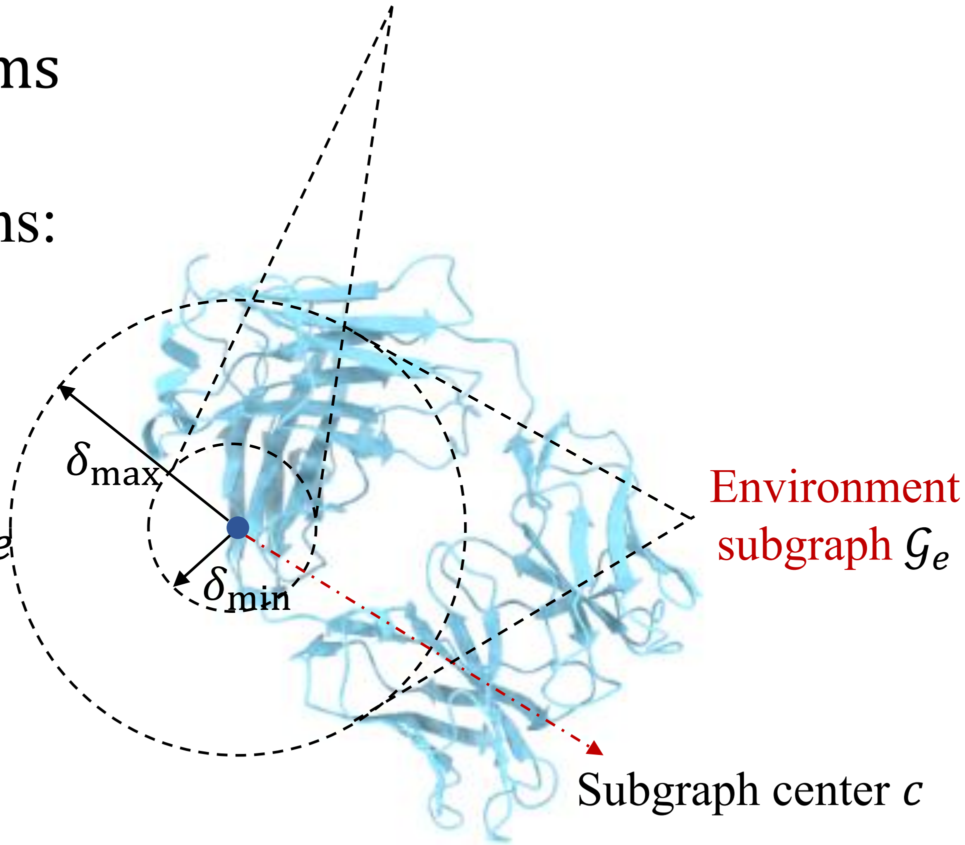
Gradient-Environment Subgraph

- For each macromolecule $\mathcal{G}$ with more than 1,000 atoms
- Randomly select an atom c from the molecule
- Given $\delta_{\min} < \delta_{\max}$, define the following two subgraphs:

$$\mathcal{G}_g = \left\{ j \middle| j \in \mathcal{G}, \left\| x_j - x_c \right\|_2 < \delta_{\min} \right\},$$

$$\mathcal{G}_e = \left\{ j \middle| j \in \mathcal{G}, \left\| x_j - x_c \right\|_2 < \delta_{\max} \right\},$$

- $\mathcal{G}_g$ will serve as the input in place of $\mathcal{G}$, and atoms in $\mathcal{G}_e$ will participate in the calculation of training objectives


Gradient subgraph $\mathcal{G}_g$
$\delta_{\max}$
$\delta_{\min}$
Environment subgraph $\mathcal{G}_e$
Subgraph center c

Challenges & Solutions

- How to identify **specific substructures** (*e.g.*, α-carbon in amino acids) under the premise of unified representation?

Atom Embedding Expansion

- Use the periodic table as the basic vocabulary $\boldsymbol{A}_b \in \mathbb{R}^{A\times H}$
- Predefine the expanded dimension D and initialize the expanded vocabulary $\boldsymbol{A}_e \in \mathbb{R}^{A\times D\times H}$
- For atom i of the molecular graph $\mathcal{G}$, calculate the expanded weight vector:

$$\boldsymbol{n}_i = \sum_{j\in\mathcal{N}_i} \mathrm{rbf}(d_{ij}) \odot \boldsymbol{A}_b[j] \in \mathbb{R}^H,$$

$$\boldsymbol{w}_i = \mathrm{softmax}\big(\mathrm{lin}(\boldsymbol{A}_b[i], \boldsymbol{n}_i)\big) \in [0,1]^D,$$

- The expanded embedding of atom i is given by:

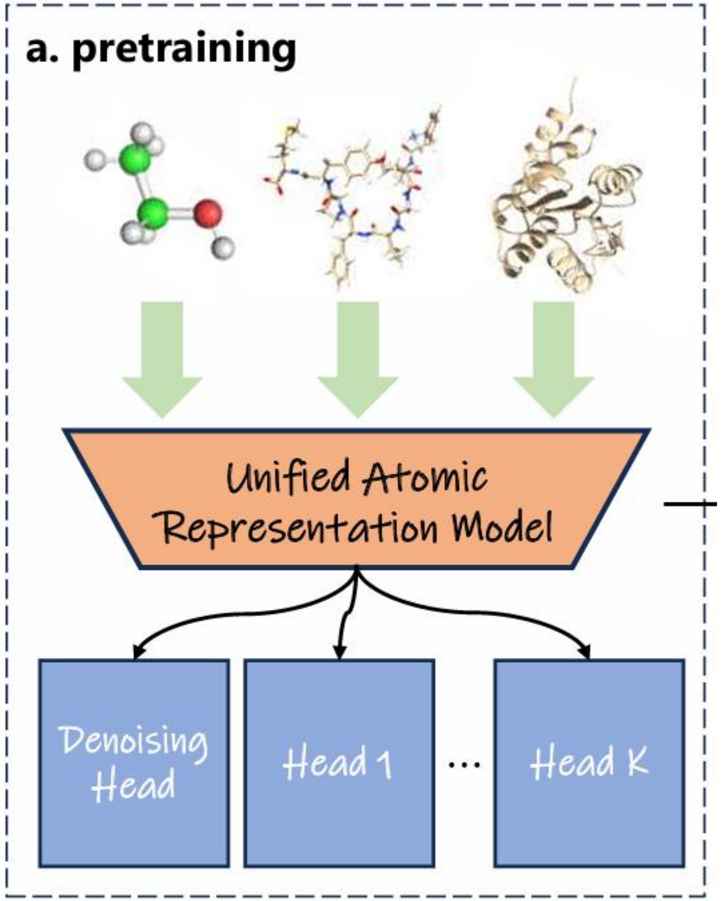
$$\boldsymbol{z}_i = \mathrm{lin}\big(\boldsymbol{A}_b[i], w_i^\top \boldsymbol{A}_e[i], \boldsymbol{n}_i\big) \in \mathbb{R}^H.$$

Challenges & Solutions

- How to deal with **inconsistent force labels** caused by using different force field parameters?
- How to deal with the mixture of equilibrium and off-equilibrium conformations?

Unified Multi-Head Pretraining

- For different states
  - **Equilibrium:** denoising pretraining
  - **Off-equilibrium:** pretraining with force labels
- For different force field parameters
  - **Multi-Head:** Use *K* output heads corresponding to *K* different force fields



Challenges & Solutions

- How to perform MD simulation in different **chemical environments** (*e.g.*, solvation)?
- **Notice:** the potential $\varepsilon(\cdot)$ is a good reflection of the chemical environment.

Force Guidance Kernel

- *Stochastic Interpolant*

$$\begin{tikzcd} \mathbf{X}_t = t\mathbf{X}_1 + (1-t)\mathbf{X}_0 + \sqrt{t(1-t)}\sigma_s\mathbf{Z}, \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \arrow[r, leftrightarrow, "\gamma(t) = \sqrt{t(1-t)}\sigma_s"] & \mathrm{d}\mathbf{X}_t = b(t, \mathbf{X}_t)\mathrm{d}t - \frac{\epsilon(t)}{\gamma(t)}\eta_z(t, \mathbf{X}_t)\mathrm{d}t + \sqrt{2\epsilon(t)}\mathrm{d}\mathbf{B}_t \end{tikzcd}$$

Stochastic Process

SDE

Vector field model $\mathbf{X}_0 \xrightarrow{b, \eta_z} \mathbf{X}_t \sim q_t(\cdot) \longrightarrow \mathbf{X}_1$

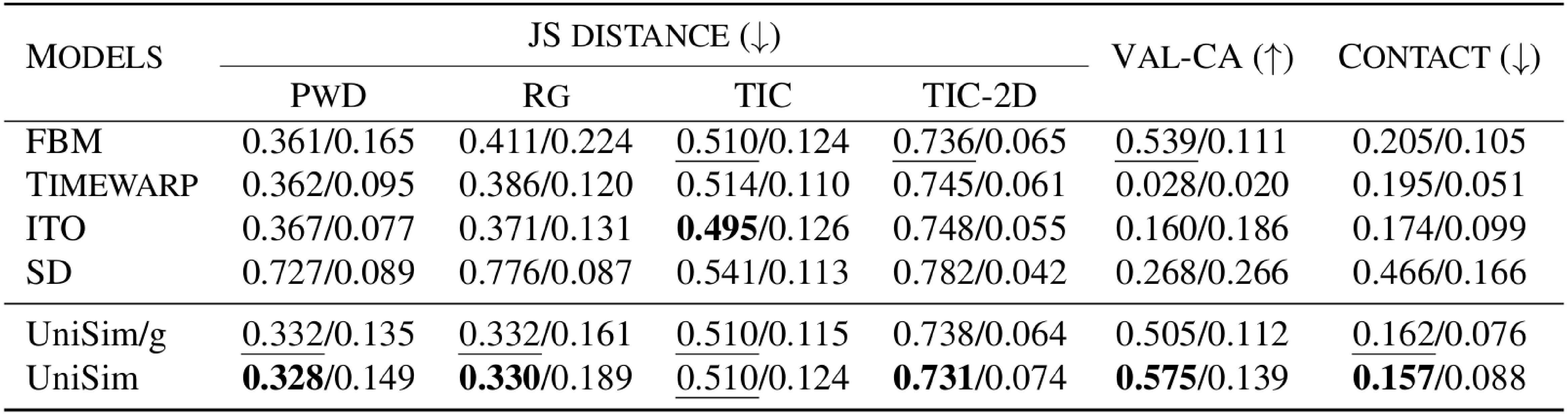
Force Guidance Kernel $\mathbf{X}_0 \xrightarrow{b', \eta_z'} \mathbf{X}_t' \sim p_t(\cdot) \longrightarrow \mathbf{X}_1'$

- We prove that, if $b' = b, \eta_z' = \eta_z + \alpha\gamma(t)\nabla\varepsilon_t$, then $p_t \propto q_t \exp(-\alpha\varepsilon_t)$ under some assumptions, where ε_t is called the intermediate potential that satisfies $\varepsilon_0 = \varepsilon_1 = \varepsilon$.

✓ *Parameters of the vector field model are frozen => The force guidance kernel is pluggable!*

Experiments

■ Compare with time-coarsened dynamics baselines on peptides

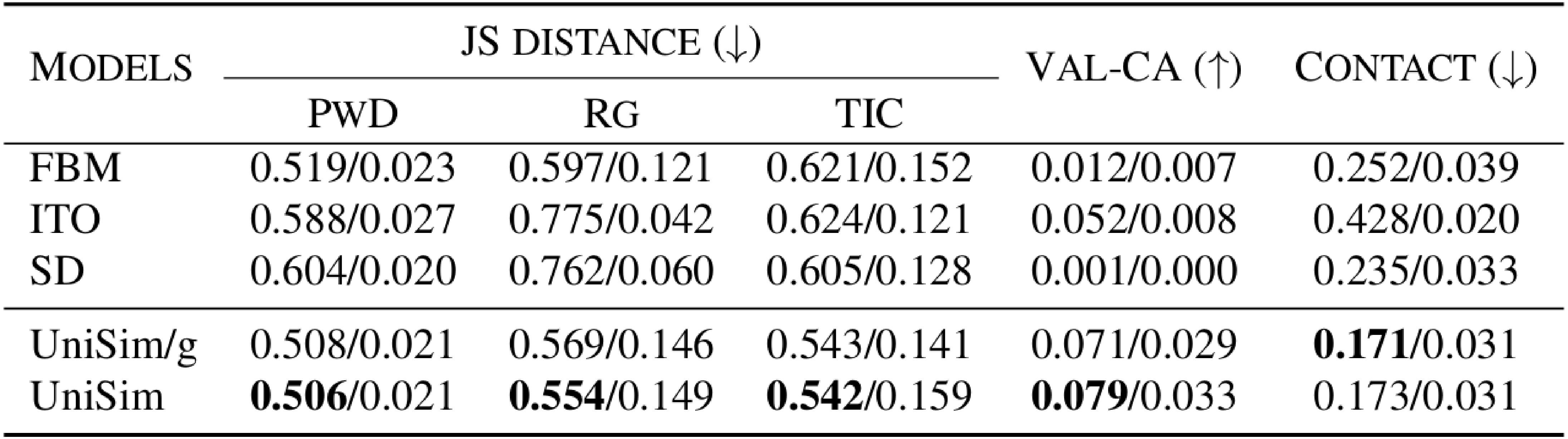


MODELS	JS DISTANCE (↓)				VAL-CA (↑)	CONTACT (↓)
	PWD	RG	TIC	TIC-2D		
FBM	0.361/0.165	0.411/0.224	0.510/0.124	0.736/0.065	0.539/0.111	0.205/0.105
TIMEWARP	0.362/0.095	0.386/0.120	0.514/0.110	0.745/0.061	0.028/0.020	0.195/0.051
ITO	0.367/0.077	0.371/0.131	**0.495**/0.126	0.748/0.055	0.160/0.186	0.174/0.099
SD	0.727/0.089	0.776/0.087	0.541/0.113	0.782/0.042	0.268/0.266	0.466/0.166
UniSim/g	0.332/0.135	0.332/0.161	0.510/0.115	0.738/0.064	0.505/0.112	0.162/0.076
UniSim	**0.328**/0.149	**0.330**/0.189	0.510/0.124	**0.731**/0.074	**0.575**/0.139	**0.157**/0.088

- ➢ All models perform the simulation for each molecular system with 1,000 frames.
- ➢ UniSim/g denotes only using the vector field model for inference, without the force guidance kernel.

Experiments

- **Compare with time-coarsened dynamics baselines on proteins with fine-tuning**

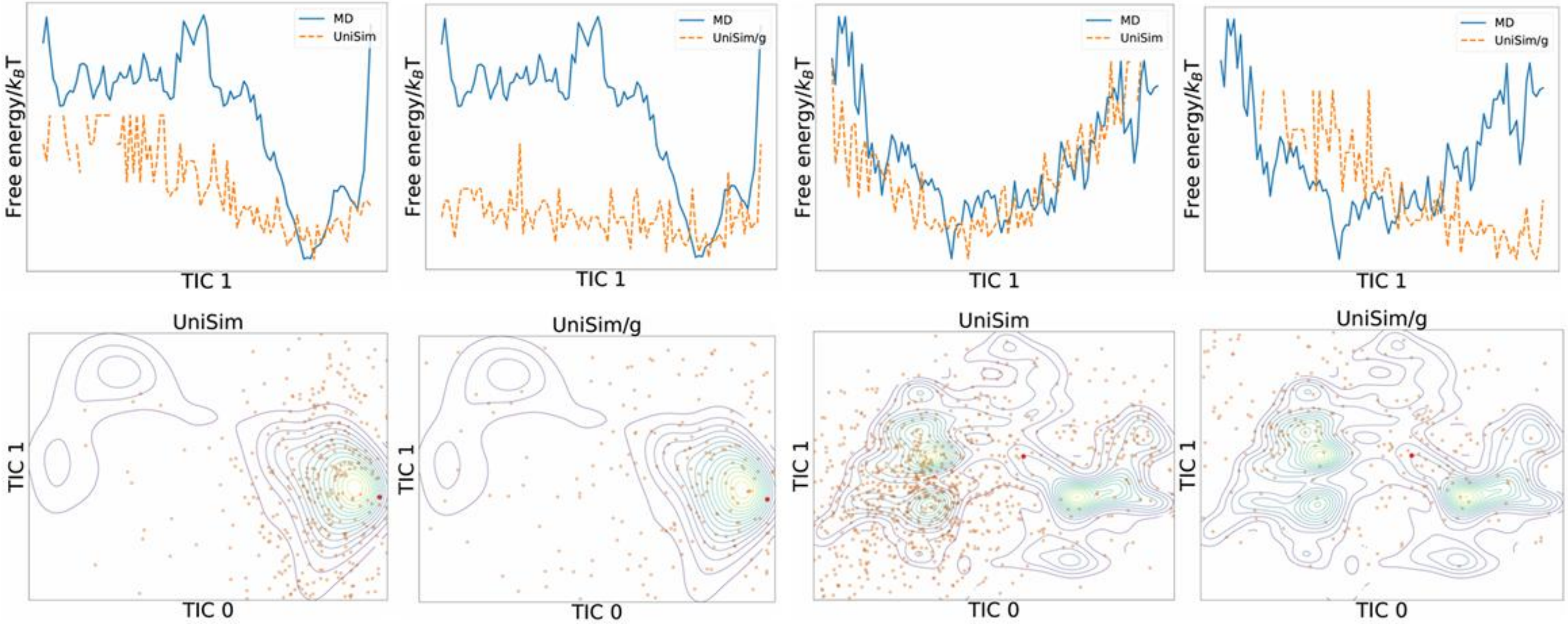


MODELS	JS DISTANCE (↓)			VAL-CA (↑)	CONTACT (↓)
	PWD	RG	TIC		
FBM	0.519/0.023	0.597/0.121	0.621/0.152	0.012/0.007	0.252/0.039
ITO	0.588/0.027	0.775/0.042	0.624/0.121	0.052/0.008	0.428/0.020
SD	0.604/0.020	0.762/0.060	0.605/0.128	0.001/0.000	0.235/0.033
UniSim/g	0.508/0.021	0.569/0.146	0.543/0.141	0.071/0.029	**0.171**/0.031
UniSim	**0.506**/0.021	**0.554**/0.149	**0.542**/0.159	**0.079**/0.033	0.173/0.031

✓ UniSim outperforms other baselines on comprehensive metrics, especially on validity.

Experiments

■ Transferability to small molecules with the force guidance kernel

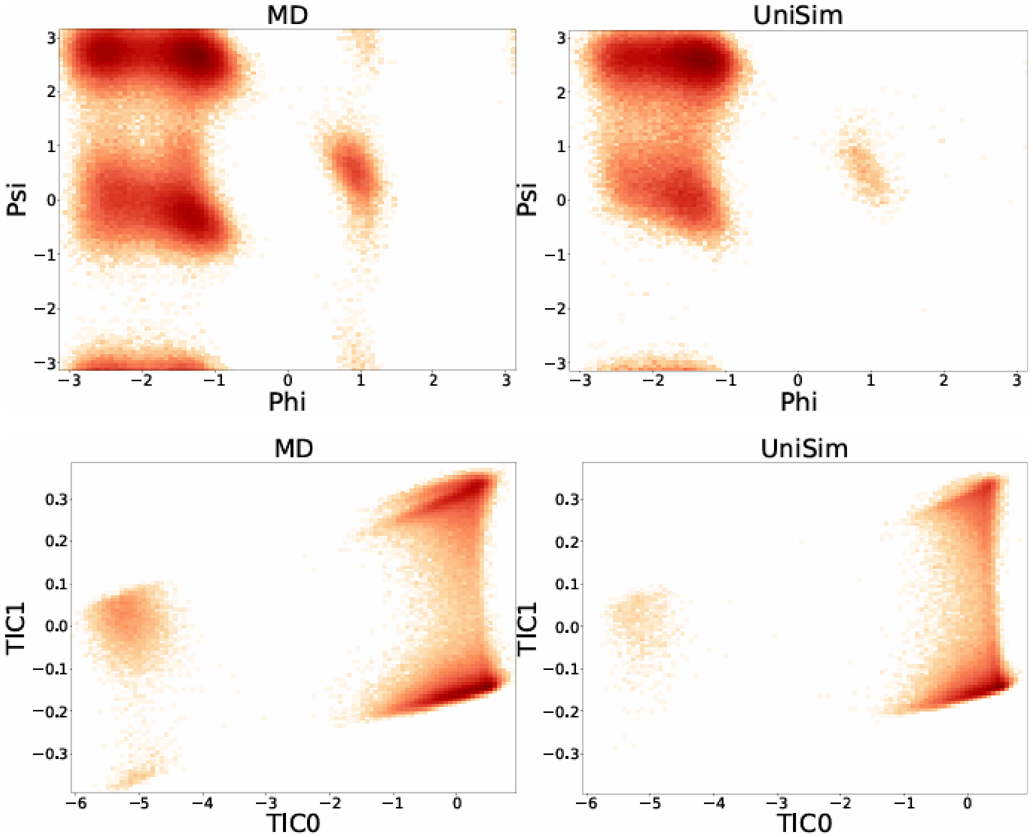


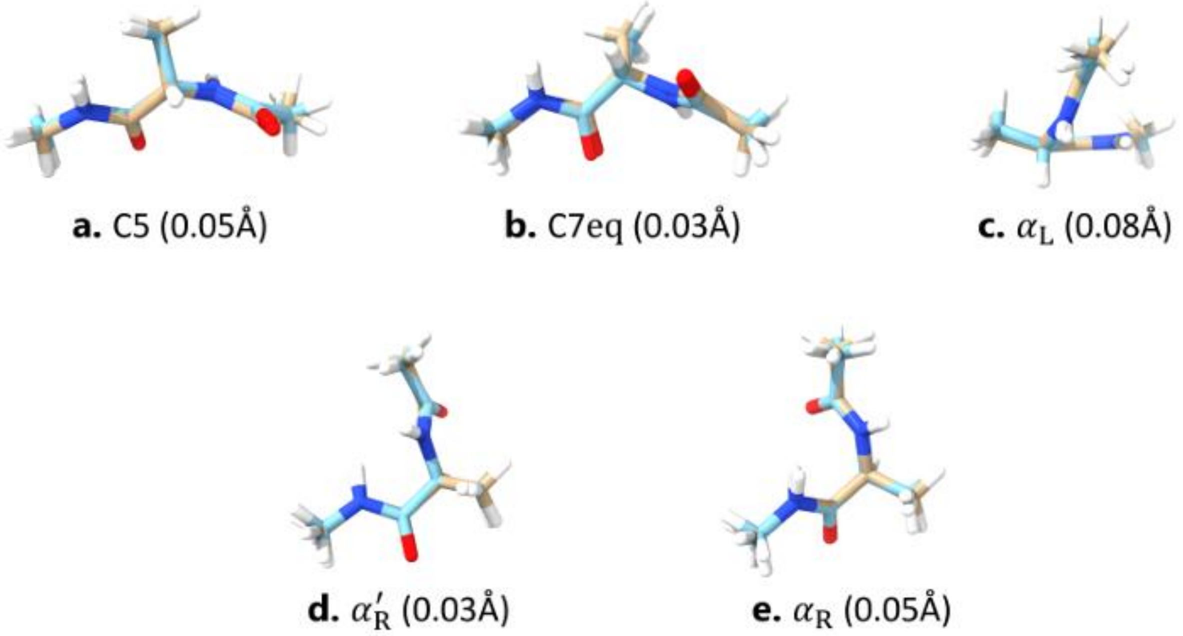
a. Ac-Ala3-NHMe **b.** DHA

✓ The force guidance greatly helps UniSim comprehend the free energy landscape.

Experiments

■ Long-timescale simulations for Alanine-Dipeptide (AD)





yellow: MD blue: UniSim

✓ UniSim robustly reproduces the free energy landscape and successfully explores key metastable states of the alanine-dipeptide system.

Thanks!