Physical Consistency Bridges Heterogeneous Data in Molecular Multi-Task Learning

Yuxuan Ren, Dihan Zheng, Chang Liu, Peiran Jin, Yu Shi, Lin Huang, Jiyan He, Shengjie Luo, Tao Qin, Tie-Yan Liu

Microsoft Research AI for Science ✉ changliu@microsoft.com

Microsoft

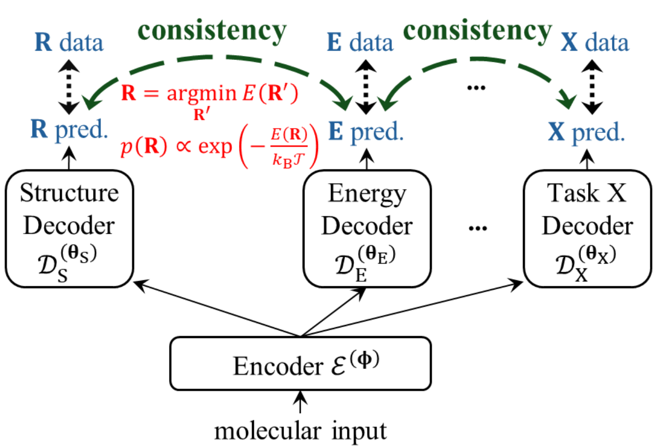
MOTIVATION

Data Heterogeneity in Molecular Science:

- Different levels of accuracy:
  - Some tasks cost more to generate data.
    - E.g., equilibrium structure costs multiple times more than energy does.
  - Accuracy-efficiency trade-off of data-generation methods.
    - E.g., PubChemQC B3LYP/6-31G*//PM6 generates energy in DFT level, but equilibrium structure in semi-empirical level.
- Tasks cannot directly benefit each other.
  - E.g., force labels on off-equilibrium structures cannot yet directly improve equilibrium structure.

GENERAL IDEA

Multi-task Learning with Physical Consistency:



PHYSICAL CONSISTENCY TRAINING

Optimality Consistency

Equilibrium structure is the argmin of energy:

$\mathbf{R}^\star(\mathcal{G}) = \operatorname{argmin}_{\mathbf{R}} E_{\mathcal{G}}(\mathbf{R})$ ➔➔

$$\min_{\theta} \mathbb{E}_{\boldsymbol{\eta}} \max\{0, E_{\phi,\mathcal{G}}(\mathbf{R}_\theta^\star(\mathcal{G})) - E_{\phi,\mathcal{G}}(\mathbf{R}_\theta^\star(\mathcal{G}) + \boldsymbol{\eta})\}.$$

- Gradient-norm loss $\|\nabla E_{\phi,\mathcal{G}}(\mathbf{R}_\theta^\star(\mathcal{G}))\|^2$ or just $E_{\phi,\mathcal{G}}(\mathbf{R}_\theta^\star(\mathcal{G}))$ as a loss are unstable.
- Only structure-related parameters θ are optimized.

Specification for Diffusion Model

To obtain $\mathbf{R}_\theta^\star(\mathcal{G})$:

- Using reverse process is prohibitively costly for optimization.
- Leveraging the denoising formulation: $\mathbf{D}_{\theta,\mathcal{G},t}(\mathbf{R}_t)$ targets $\mathbb{E}_{|\mathcal{G}}[\mathbf{R}_0 | \mathbf{R}_t]$.
- **Symmetry breaking:** Taking $t < T$ but close to T.
- $\min_{\theta} \mathbb{E}_{\boldsymbol{\eta}} \max\left\{0,\ E_{\phi,\mathcal{G}}\left(\mathbf{D}_{\theta,\mathcal{G},t}(\boldsymbol{\epsilon})\right) - E_{\phi,\mathcal{G}}\left(\mathbf{D}_{\theta,\mathcal{G},t}(\boldsymbol{\epsilon}) + \boldsymbol{\eta}\right)\right\}.$

Score Consistency

Equilibrium structure is a sample from the thermodynamic distribution at low temperature:

$\mathbf{R}^\star(\mathcal{G}) \sim p_{\mathcal{G}}(\mathbf{R}) \propto \exp\left(-\frac{E_{\mathcal{G}}(\mathbf{R})}{k_B \mathcal{T}}\right)$ ➔➔

$$\min_{\theta} \mathbb{E}_{\mathbf{R}} \left\| \nabla \log p_{\theta,\mathcal{G}}(\mathbf{R}) + \frac{\nabla E_{\phi,\mathcal{G}}(\mathbf{R})}{k_B \mathcal{T}} \right\|^2.$$

- Proper calculation of $\log p_{\theta,\mathcal{G}}(\mathbf{R})$ (solving ODE) is prohibitively costly for optimization.
- $\mathbf{s}_{\theta,\mathcal{G},t=0}(\mathbf{R})$ targets $\nabla \log p_{\theta,\mathcal{G}}(\mathbf{R})$.
- $\mathbf{s}_{\theta,\mathcal{G},t}(\mathbf{R}_t) = \frac{\sqrt{\bar{\alpha}_t}\, \mathbf{D}_{\theta,\mathcal{G},t}(\mathbf{R}_t) - \mathbf{R}_t}{1-\bar{\alpha}_t}$: 0/0 near t=0.
- Taking $t > 0$ but close to 0:

$$➔ \min_{\theta} \mathbb{E}_{p_t(\mathbf{R})} \left\| \frac{\sqrt{\bar{\alpha}_t}\, \mathbf{D}_{\theta,\mathcal{G},t}(\mathbf{R}) - \mathbf{R}}{1-\bar{\alpha}_t} + \frac{\nabla E_{\phi,\mathcal{G}}(\mathbf{R})}{k_B \mathcal{T}} \right\|^2.$$

- Does not contradict with the optimality consistency loss: one is near T, one is near 0.

EXPERIMENTS

Train on low-accuracy structures, test RMSD (↓) vs. high-accuracy structures

Training Data	Test Set / Generated by / Struct. Stat.	PCQ Denoising Mean	PCQ Denoising Min	PCQ DDIM Mean	PCQ DDIM Min	QM9 Denoising Mean	QM9 Denoising Min	QM9 DDIM Mean	QM9 DDIM Min
PM6	Multi-Task	1.189	0.655	1.041	0.361	0.928	0.545	0.669	0.197
	Consistency	**1.158**	**0.645**	**1.007**	**0.346**	**0.848**	**0.490**	**0.650**	**0.194**
PM6 & SPICE force	Multi-Task	1.161	0.631	1.047	0.373	0.876	0.486	0.670	0.207
	Consistency	**1.147**	**0.590**	**1.013**	**0.345**	**0.842**	**0.485**	**0.644**	**0.194**
PM6 & subset force	Multi-Task	1.199	0.672	1.027	0.365	0.914	0.545	0.648	0.193
	Consistency	**1.113**	**0.629**	**1.019**	**0.351**	**0.836**	**0.488**	**0.646**	**0.192**

With finetuning

(Pre-)Training Data	Test Set / Generated by / Struct. Stat.	PCQ Denoising Mean	PCQ Denoising Min	PCQ DDIM Mean	PCQ DDIM Min	QM9 Denoising Mean	QM9 Denoising Min	QM9 DDIM Mean	QM9 DDIM Min
PM6	Multi-Task	1.158	0.614	0.921	0.220	0.889	0.467	0.501	0.090
	Consistency	**1.152**	**0.610**	**0.918**	**0.218**	**0.835**	**0.420**	**0.493**	**0.076**
PM6 & SPICE force	Multi-Task	1.161	0.618	0.930	0.219	0.855	0.444	0.505	0.081
	Consistency	**1.132**	**0.581**	**0.916**	**0.215**	**0.832**	**0.418**	**0.492**	**0.073**
PM6 & subset force	Multi-Task	1.143	0.603	0.927	0.224	0.855	0.441	0.497	0.080
	Consistency	**1.099**	**0.542**	**0.914**	**0.215**	**0.822**	**0.419**	**0.490**	**0.076**

Analysis

$$\mathrm{EGap} := \frac{E_{\phi,\mathcal{G}}\left(\mathbf{R}^\star_{\mathrm{pred},\theta}(\mathcal{G})\right) - E_{\phi,\mathcal{G}}(\mathbf{R}^\star(\mathcal{G}))}{|E_{\phi,\mathcal{G}}(\mathbf{R}^\star(\mathcal{G}))|}$$

Train Set	PM6	PM6 with SPICE force	PM6 with PM6 subset force
Multi-Task	0.1278	0.0546	0.1163
Consistency	**0.1172**	**0.0306**	**0.1013**