Microsoft

Physical Consistency Bridges Heterogeneous Data in Molecular Multi-Task Learning

Chang Liu, Microsoft Research AI for Science,
on behalf of the authors
changliu@microsoft.com

Motivation

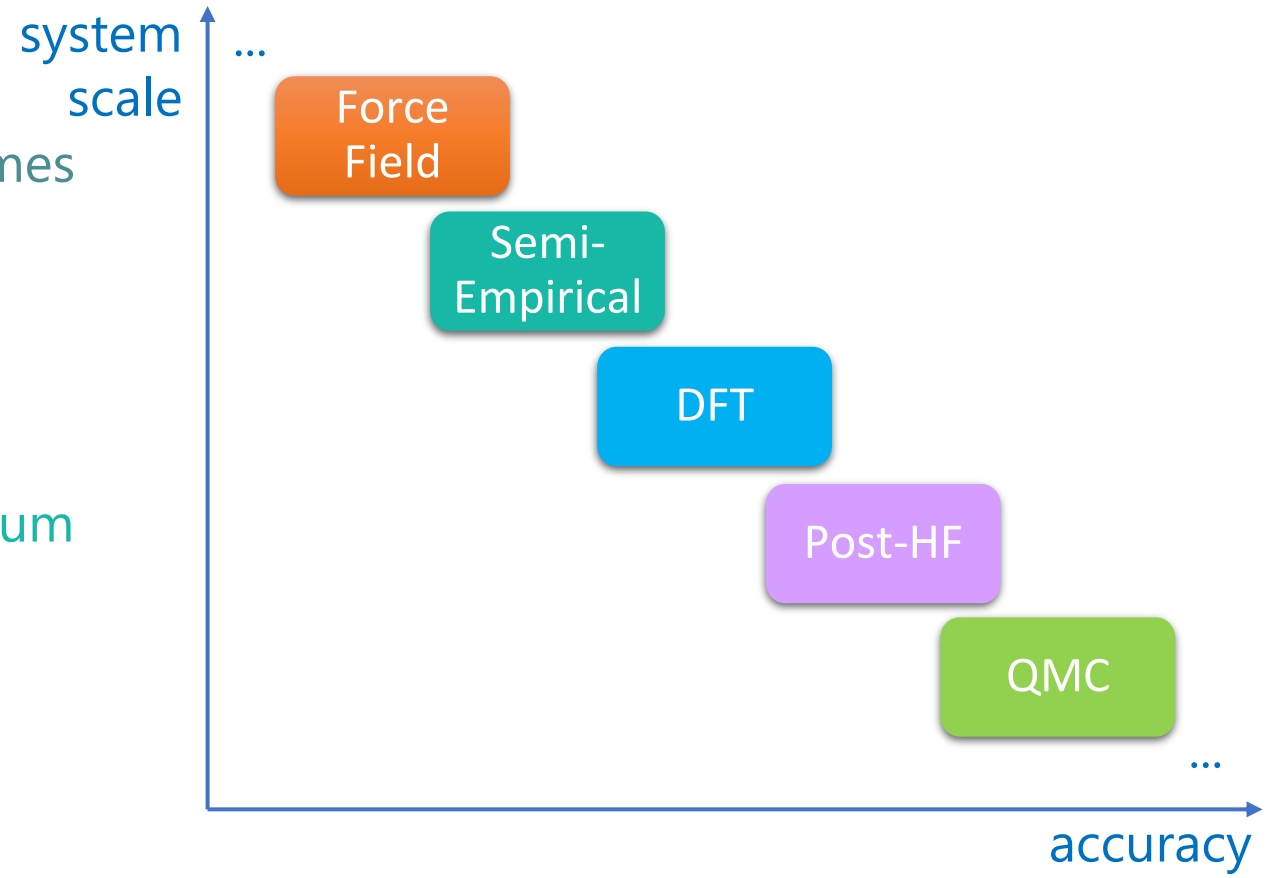
Data Heterogeneity in Molecular Science

- Different levels of accuracy:
  - Some tasks cost more to generate data.
    - E.g., equilibrium structure costs multiple times more than energy does.

Motivation

Data Heterogeneity in Molecular Science

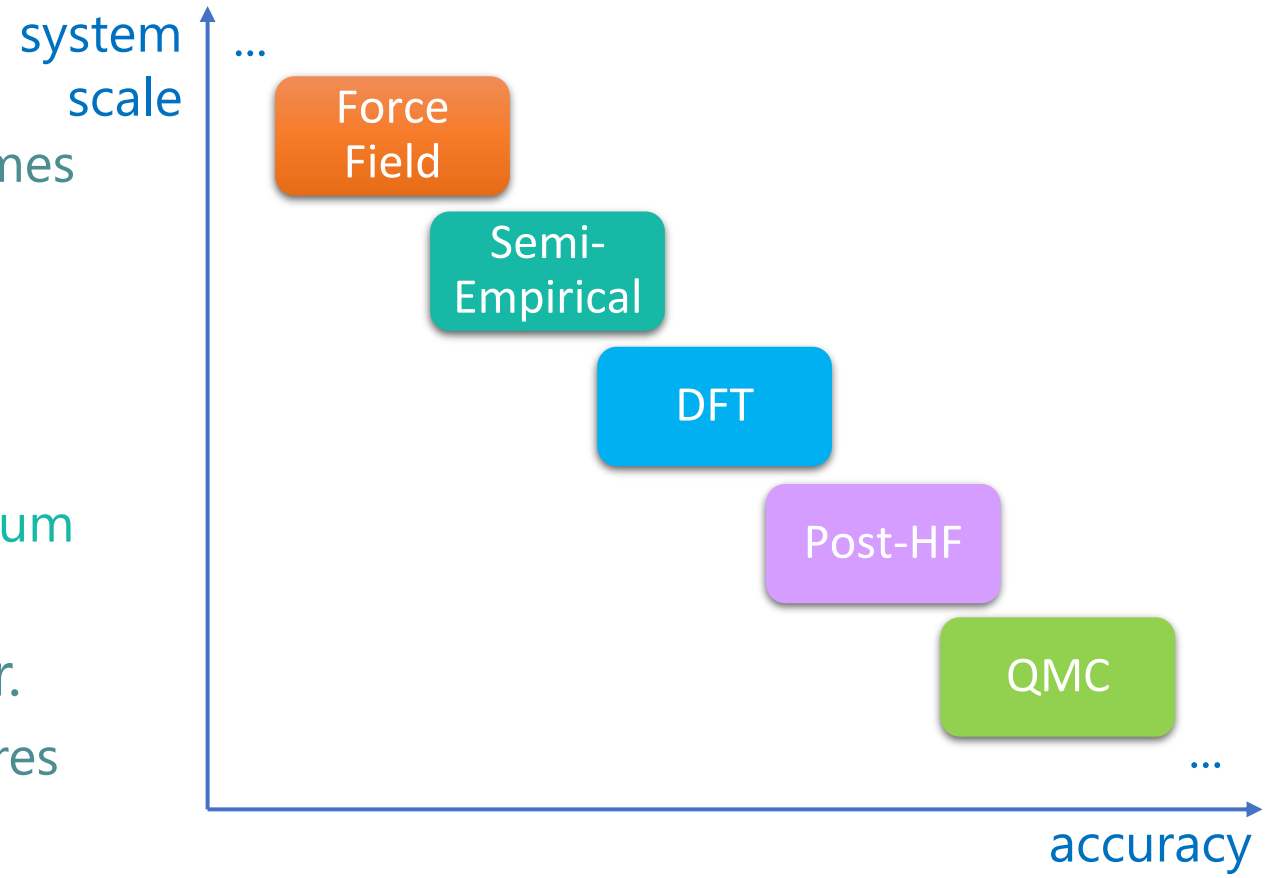
- Different levels of accuracy:
  - Some tasks cost more to generate data.
    - E.g., equilibrium structure costs multiple times more than energy does.
  - Accuracy-efficiency trade-off of data-generation methods.
    - E.g., PubChemQC B3LYP/6-31G*//PM6 generates energy in DFT level, but equilibrium structure in semi-empirical level.



Motivation

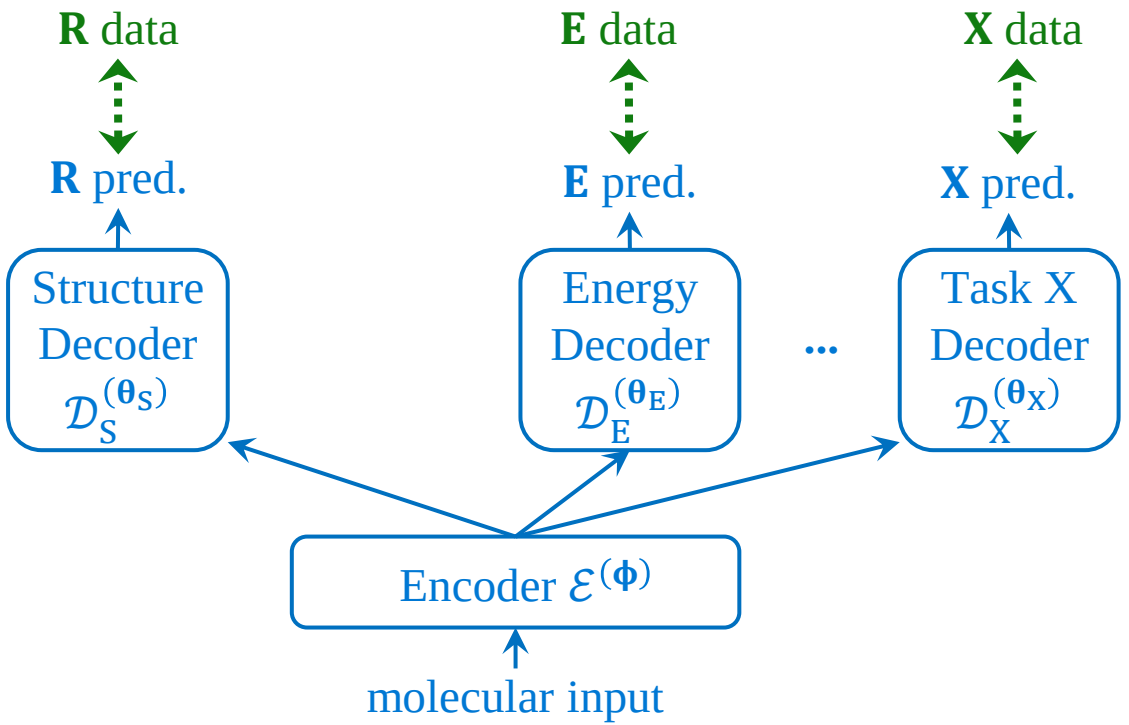
Data Heterogeneity in Molecular Science

- 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.



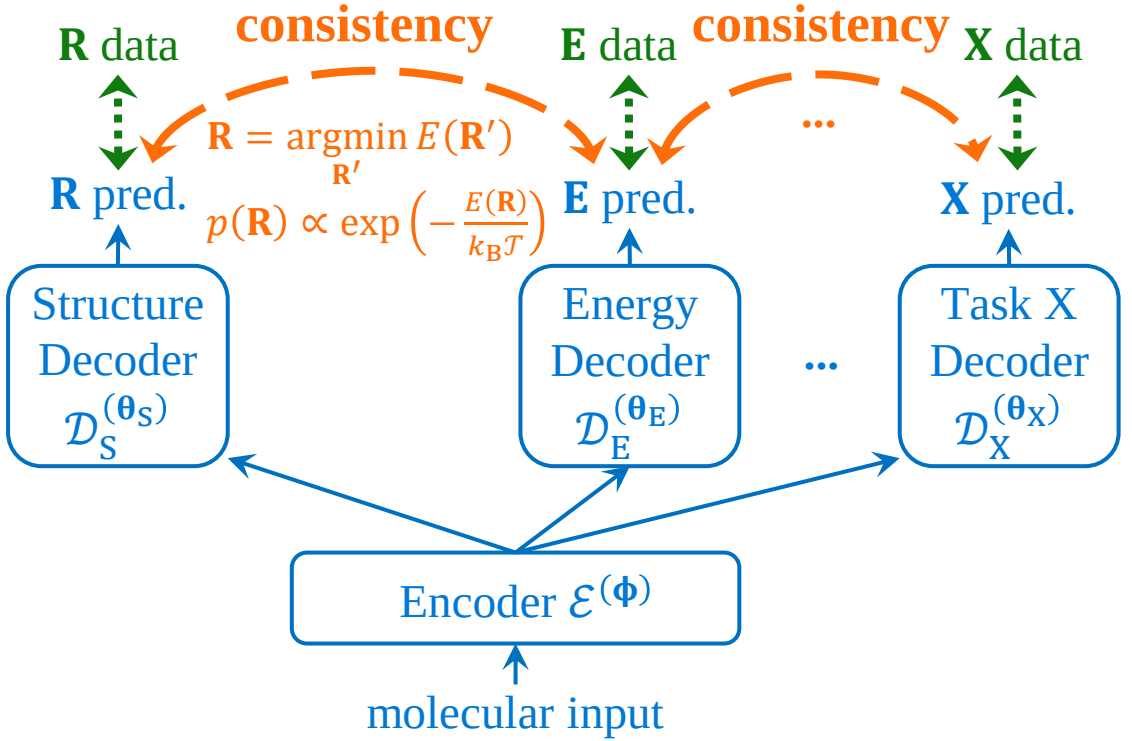
General Idea

- Basic approach: multi-task learning.
  - Connection in the input end: shared input processor (encoder).



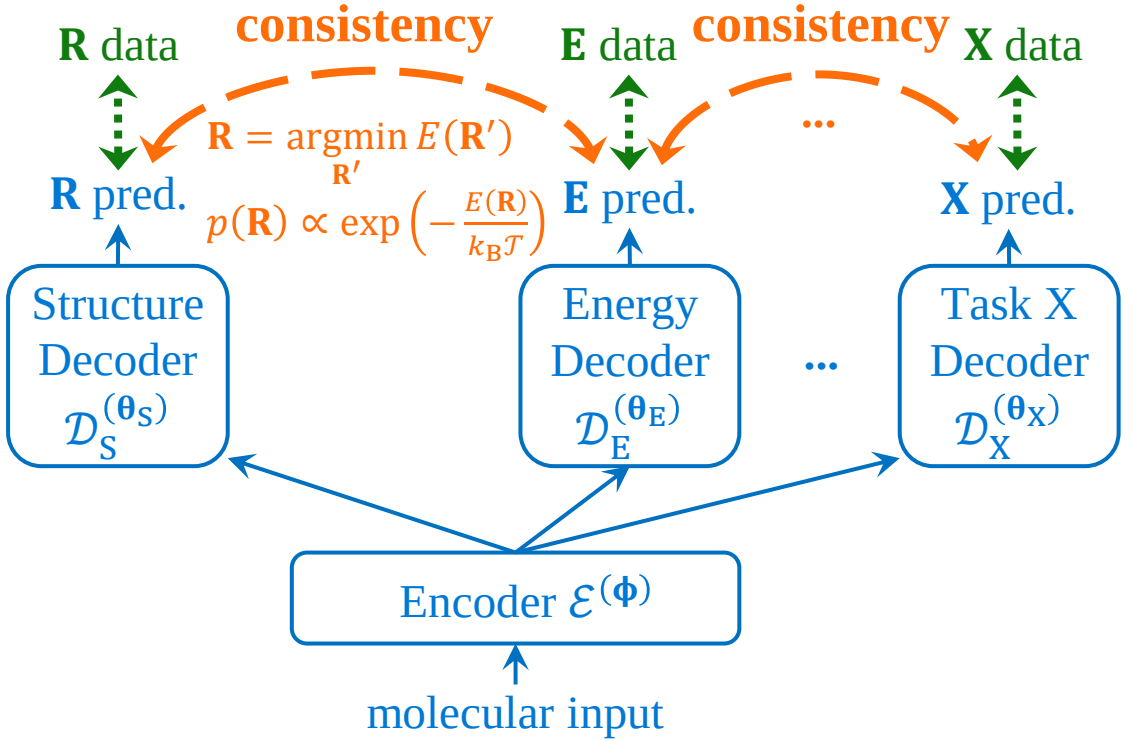
General Idea

- Basic approach: multi-task learning.
  - Connection in the input end:
    shared input processor (encoder).
- Proposed approach: **scientific consistency**.
  - Scientific tasks originate in fundamental scientific laws, which explicitly connect them.
  - Connection in the output end:
    direct information exchange among tasks:



General Idea

- Basic approach: multi-task learning.
  - Connection in the input end: shared input processor (encoder).
- Proposed approach: **scientific consistency**.
  - Scientific tasks originate in fundamental scientific laws, which explicitly connect them.
  - Connection in the output end: direct information exchange among tasks:
    - Information of a higher level of accuracy can flow from one task (e.g., energy) to another (e.g., equilibrium structure).
    - Data of a related task (e.g., force) can directly improve the performance of the concerned task (e.g., equilibrium structure).



General Method

- Equilibrium structure is the argmin of energy:
  $\mathbf{R}^\star(\mathcal{G}) = \underset{\mathbf{R}}{\operatorname{argmin}}\, E_{\mathcal{G}}(\mathbf{R})$.
- Equilibrium structure is a sample from the thermodynamic distribution defined by the energy at low temperature:
  $\mathbf{R}^\star(\mathcal{G}) \sim p_{\mathcal{G}}(\mathbf{R}) \propto \exp\left(-\frac{E_{\mathcal{G}}(\mathbf{R})}{k_B \mathcal{T}}\right)$.
- Force is the gradient of energy:
  Force labels on off-equilibrium structures
  → better energy landscape
  → better equilibrium structure.

Specification for Diffusion Model

- Diffusion model for structure generation:

Target distribution $p_{\mathcal{G}}(\mathbf{R})$

$$\mathrm{d}\mathbf{R}_t = -\frac{\beta_t}{2}\mathbf{R}_t\,\mathrm{d}t + \sqrt{\beta_t}\,\mathrm{d}\mathbf{W}_t$$

$p_T(\mathbf{R})$ Simple distribution

$$\mathrm{d}\mathbf{R}_{\bar{t}} = \frac{\beta_{T-\bar{t}}}{2}\mathbf{R}_{\bar{t}}\,\mathrm{d}\bar{t} + \beta_{T-\bar{t}}\nabla \log p_{\mathcal{G},T-\bar{t}}(\mathbf{R}_{\bar{t}})\,\mathrm{d}\bar{t} + \sqrt{\beta_{T-\bar{t}}}\,\mathrm{d}\mathbf{W}_{\bar{t}}$$

Specification for Diffusion Model

- Diffusion model for structure generation:

Target distribution $p_{\mathcal{G}}(\mathbf{R})$

$$\mathrm{d}\mathbf{R}_t = -\frac{\beta_t}{2}\mathbf{R}_t\,\mathrm{d}t + \sqrt{\beta_t}\,\mathrm{d}\mathbf{W}_t$$

$p_T(\mathbf{R})$ Simple distribution

$$\mathrm{d}\mathbf{R}_{\bar{t}} = \frac{\beta_{T-\bar{t}}}{2}\mathbf{R}_{\bar{t}}\,\mathrm{d}\bar{t} + \beta_{T-\bar{t}}\nabla \log p_{\mathcal{G},T-\bar{t}}(\mathbf{R}_{\bar{t}})\,\mathrm{d}\bar{t} + \sqrt{\beta_{T-\bar{t}}}\,\mathrm{d}\mathbf{W}_{\bar{t}}$$

$$\approx \mathbf{s}_{\theta,\mathcal{G},T-\bar{t}}(\mathbf{R}_{\bar{t}})$$

Denoising score matching:

$$\mathbb{E}_{p_0(\mathbf{R}_0)}\mathbb{E}_{p(\mathbf{R}_t|\mathbf{R}_0)}\left\|\mathbf{s}_{\theta,\mathcal{G},t}(\mathbf{R}_t) - \nabla_{\mathbf{R}_t}\log p(\mathbf{R}_t|\mathbf{R}_0)\right\|^2$$

$$= \mathbb{E}_{p_0(\mathbf{R}_0)}\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\mathbf{s}_{\theta,\mathcal{G},t}\left(\sqrt{\bar{\alpha}_t}\,\mathbf{R}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}\right) + \frac{\boldsymbol{\epsilon}}{\sqrt{1-\bar{\alpha}_t}}\right\|^2,$$

$$\bar{\alpha}_t \coloneqq \exp\left(-\int_0^t \beta_s\,\mathrm{d}s\right).$$

Specification for Diffusion Model

- Diffusion model for structure generation:

Target distribution $p_{\mathcal{G}}(\mathbf{R})$

$$\mathrm{d}\mathbf{R}_t = -\frac{\beta_t}{2}\mathbf{R}_t\,\mathrm{d}t + \sqrt{\beta_t}\,\mathrm{d}\mathbf{W}_t$$

$p_T(\mathbf{R})$ Simple distribution

$$\mathrm{d}\mathbf{R}_{\bar{t}} = \frac{\beta_{T-\bar{t}}}{2}\mathbf{R}_{\bar{t}}\,\mathrm{d}\bar{t} + \beta_{T-\bar{t}}\nabla\log p_{\mathcal{G},T-\bar{t}}(\mathbf{R}_{\bar{t}})\,\mathrm{d}\bar{t} + \sqrt{\beta_{T-\bar{t}}}\,\mathrm{d}\mathbf{W}_{\bar{t}}$$

$$\mathbf{D}_{\theta,\mathcal{G},t}(\mathbf{R}_t) \coloneqq \frac{\mathbf{R}_t + (1-\bar{\alpha}_t)\,\mathbf{s}_{\theta,\mathcal{G},t}(\mathbf{R}_t)}{\sqrt{\bar{\alpha}_t}} \quad \Leftarrow \quad \approx \mathbf{s}_{\theta,\mathcal{G},T-\bar{t}}(\mathbf{R}_{\bar{t}})$$

Denoising loss:

$$\mathbb{E}_{p_0(\mathbf{R}_0)}\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\mathbf{D}_{\theta,\mathcal{G},t}\left(\sqrt{\bar{\alpha}_t}\,\mathbf{R}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}\right) - \mathbf{R}_0\right\|^2.$$

Denoising score matching:

$$\mathbb{E}_{p_0(\mathbf{R}_0)}\mathbb{E}_{p(\mathbf{R}_t|\mathbf{R}_0)}\left\|\mathbf{s}_{\theta,\mathcal{G},t}(\mathbf{R}_t) - \nabla_{\mathbf{R}_t}\log p(\mathbf{R}_t|\mathbf{R}_0)\right\|^2$$

$$= \mathbb{E}_{p_0(\mathbf{R}_0)}\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\mathbf{s}_{\theta,\mathcal{G},t}\left(\sqrt{\bar{\alpha}_t}\,\mathbf{R}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}\right) + \frac{\boldsymbol{\epsilon}}{\sqrt{1-\bar{\alpha}_t}}\right\|^2,$$

$$\bar{\alpha}_t \coloneqq \exp\left(-\int_0^t \beta_s\,\mathrm{d}s\right).$$

Specification for Diffusion Model

- Equilibrium structure is the argmin of energy:
  $\mathbf{R}^{\star}(\mathcal{G}) = \underset{\mathbf{R}}{\operatorname{argmin}} E_{\mathcal{G}}(\mathbf{R})$.
  ➔ $\min_{\theta} \mathbb{E}_{\boldsymbol{\eta}} \max\{0,\ E_{\phi,\mathcal{G}}(\mathbf{R}^{\star}_{\theta}(\mathcal{G})) - E_{\phi,\mathcal{G}}(\mathbf{R}^{\star}_{\theta}(\mathcal{G}) + \boldsymbol{\eta})\}$.
  - Gradient-norm loss $\|\nabla E_{\phi,\mathcal{G}}(\mathbf{R}^{\star}_{\theta}(\mathcal{G}))\|^2$ or just $E_{\phi,\mathcal{G}}(\mathbf{R}^{\star}_{\theta}(\mathcal{G}))$ as a loss are unstable.
  - Only structure-related parameters θ are optimized.

Optimality Consistency

Specification for Diffusion Model

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

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

  - Using reverse process is unaffordably costly for optimization.

Specification for Diffusion Model

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

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

  - Using reverse process is unaffordably costly for optimization.
  - Leveraging the denoising formulation: $\mathbf{D}_{\theta,\mathcal{G},t}(\mathbf{R}_t)$ targets $\mathbb{E}_{|\mathcal{G}}[\mathbf{R}_0|\mathbf{R}_t]$

    ➔ $\mathbf{D}_{\theta,\mathcal{G},T}(\boldsymbol{\epsilon})$ targets $\mathbb{E}_{|\mathcal{G}}[\mathbf{R}_0]$ =?= $\mathbf{R}^{\star}(\mathcal{G})$?

    - The target distribution should be rotationally invariant:

      $p_{\mathcal{G}}(\mathbf{R}_0) = p_{\mathcal{G}}(\mathbf{Q}\,\mathbf{R}_0)$ ➔ $\mathbb{E}_{|\mathcal{G}}[\mathbf{R}_0] = \mathbb{E}_{\mathrm{Unif}(\mathbf{Q})}[\mathbf{Q}\,\mathbf{R}^{\star}(\mathcal{G})] = 0.$

Specification for Diffusion Model

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

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

  - Using reverse process is unaffordably costly for optimization.
  - Leveraging the denoising formulation: $\mathbf{D}_{\theta,\mathcal{G},t}(\mathbf{R}_t)$ targets $\mathbb{E}_{|\mathcal{G}}[\mathbf{R}_0|\mathbf{R}_t]$

    ➔ $\mathbf{D}_{\theta,\mathcal{G},T}(\boldsymbol{\epsilon})$ targets $\mathbb{E}_{|\mathcal{G}}[\mathbf{R}_0]$ =?= $\mathbf{R}^{\star}(\mathcal{G})$?

    - The target distribution should be rotationally invariant:

      $p_{\mathcal{G}}(\mathbf{R}_0) = p_{\mathcal{G}}(\mathbf{Q}\,\mathbf{R}_0)$ ➔ $\mathbb{E}_{|\mathcal{G}}[\mathbf{R}_0] = \mathbb{E}_{\mathrm{Unif}(\mathbf{Q})}[\mathbf{Q}\,\mathbf{R}^{\star}(\mathcal{G})] = 0$.
    - Taking $\tau < T$ but close to T: $\mathbf{D}_{\theta,\mathcal{G},\tau}(\boldsymbol{\epsilon})$ targets $\mathbb{E}_{|\mathcal{G}}[\mathbf{R}_0|\mathbf{R}_\tau = \boldsymbol{\epsilon}]$,

      $p_{\mathcal{G}}(\mathbf{R}_0|\mathbf{R}_\tau) \propto p_{\mathcal{G}}(\mathbf{R}_0)\mathcal{N}\left(\mathbf{R}_0 \mid \mathbf{R}_\tau/\sqrt{\bar{\alpha}_\tau}, (1/\bar{\alpha}_\tau - 1)\mathbf{I}\right)$ assigns a larger probability along the orientation of $\mathbf{R}_\tau$: **symmetry breaking**!

➔ $\min_{\theta} \mathbb{E}_{\boldsymbol{\eta}} \max\left\{0,\ E_{\boldsymbol{\phi},\mathcal{G}}\left(\mathbf{D}_{\theta,\mathcal{G},\tau}(\boldsymbol{\epsilon})\right) - E_{\boldsymbol{\phi},\mathcal{G}}(\mathbf{D}_{\theta,\mathcal{G},\tau}(\boldsymbol{\epsilon}) + \boldsymbol{\eta})\right\}$.

Specification for Diffusion Model

- 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 unaffordably costly for optimization.

Specification for Diffusion Model

- 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 unaffordably costly for optimization.
  - $\mathbf{s}_{\theta,\mathcal{G},t=0}(\mathbf{R})$ targets $\nabla \log p_{\theta,\mathcal{G}}(\mathbf{R})$.
  - But $\mathbf{s}_{\theta,\mathcal{G},t}(\mathbf{R}_t) = \frac{\sqrt{\bar{\alpha}_t}\,\mathbf{D}_{\theta,\mathcal{G},t}(\mathbf{R}_t) - \mathbf{R}_t}{1-\bar{\alpha}_t}$: 0/0 near $t = 0$.

Specification for Diffusion Model

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

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

  Score Consistency

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

Experiments

- Zero-shot:

Trained on PubChemQC B3LYP/6-31G*//PM6:

Test Set	PCQ				QM9			
Generated by	Denoising		DDIM		Denoising		DDIM	
Struct. Stat.	Mean	Min	Mean	Min	Mean	Min	Mean	Min
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**

"Free lunch" redeemed from scientific laws!

Trained on PubChemQC B3LYP/6-31G*//PM6 **+ additional force data**:

Additional Training Data	Test Set	PCQ				QM9			
	Generated by	Denoising		DDIM		Denoising		DDIM	
	Struct. Stat.	Mean	Min	Mean	Min	Mean	Min	Mean	Min
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**

Experiments

- With finetuning:

Trained on PubChemQC B3LYP/6-31G*//PM6:

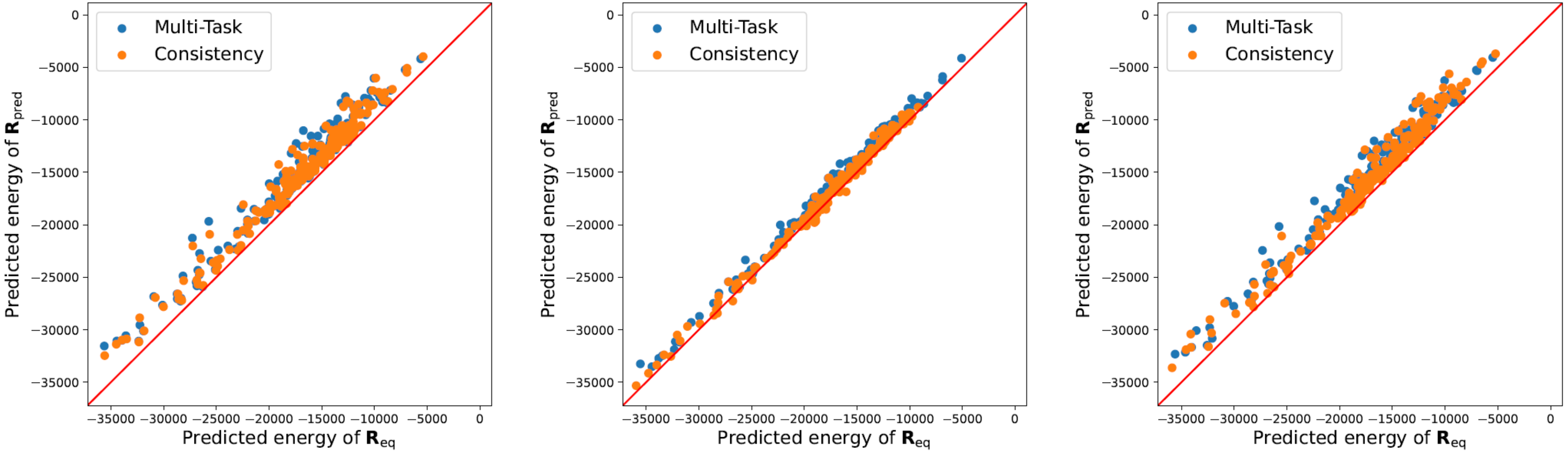
Test Set	PCQ				QM9			
Generated by	Denoising		DDIM		Denoising		DDIM	
Struct. Stat.	Mean	Min	Mean	Min	Mean	Min	Mean	Min
Multi-Task	1.158	0.614	0.921	0.220	0.889	0.467	0.501	0.090
Consistency	**1.152**	**0.610**	**0.918**	**0.218**	**0.835**	**0.420**	**0.493**	**0.076**

Trained on PubChemQC B3LYP/6-31G*//PM6 **+ additional force data**:

Additional Training Data	Test Set	PCQ				QM9			
	Generated by	Denoising		DDIM		Denoising		DDIM	
	Struct. Stat.	Mean	Min	Mean	Min	Mean	Min	Mean	Min
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**

Experiments

- Analysis



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

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**

Microsoft

Thank You