Going beyond Data Limitation with Physical Laws

Chang Liu
Zhongguancun Academy

北京中关村学院 x 中关村人工智能研究院
Zhongguancun Academy Zhongguancun Institute of Artificial Intelligence

Computational Methods for Molecular Science

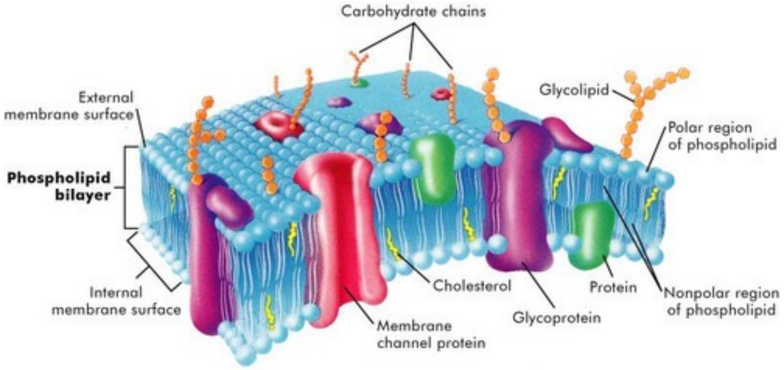
Molecular science tasks

biomolecule understanding

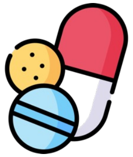
drug design

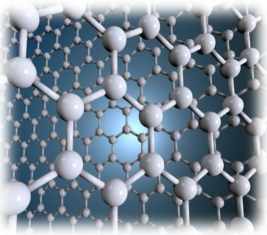
material/catalyst discovery

...


Carbohydrate chains
Glycolipid
External membrane surface
Polar region of phospholipid
Phospholipid bilayer
Internal membrane surface
Cholesterol
Protein
Nonpolar region of phospholipid
Membrane channel protein
Glycoprotein

Modified from Thibodeau GA, Patton KT: *Anatomy & physiology*, ed 6, St Louis, 2007, Mosby.





Computational Methods for Molecular Science

Molecular science tasks | **Macroscropic properties** ← **Thermodynamic ensemble** ← **Microscopic properties** ← **Electronic structure**

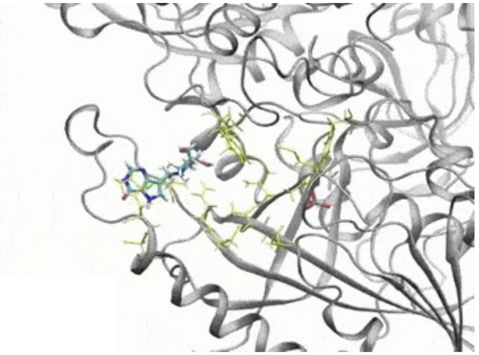
Molecular science tasks

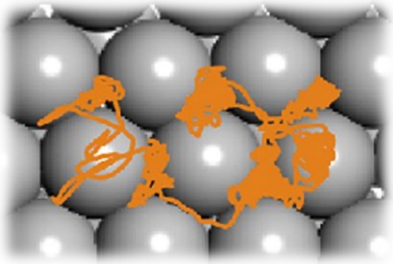
- biomolecule understanding
- drug design
- material/catalyst discovery
- …

Macroscropic properties

- binding affinity
- stability
- activation energy
- conductivity

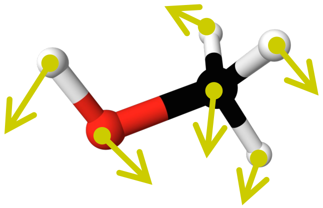
Thermodynamic ensemble





Microscopic properties

- E
- ϵ_Δ
- $\mathbf{f}$



Electronic structure

- $\psi(\mathbf{r}_1,\cdots,\mathbf{r}_N)$
- $\{\phi_i(\mathbf{r}_j)\}$
- $\rho(\mathbf{r})$

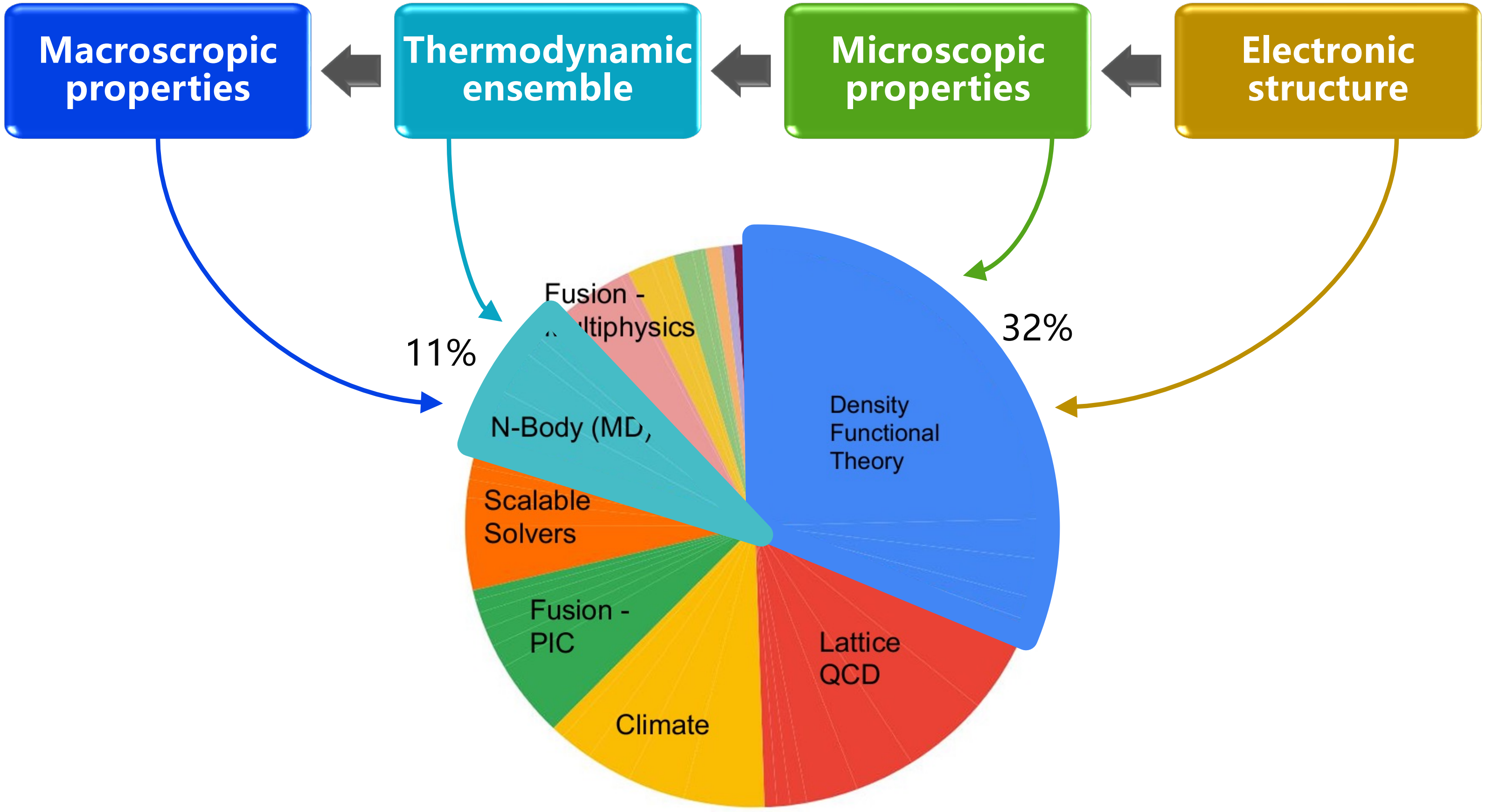
Computational Methods for Molecular Science

北京中关村学院 x 中关村人工智能研究院

Zhongguancun Academy Zhongguancun Institute of Artificial Intelligence

Molecular science tasks

- biomolecule understanding
- drug design
- material/catalyst discovery
- …



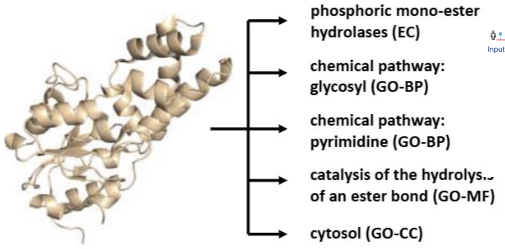
Macroscropic properties ← Thermodynamic ensemble ← Microscopic properties ← Electronic structure

AI for Scientific Computation

Macroscopic properties

Macroscopic property prediction:

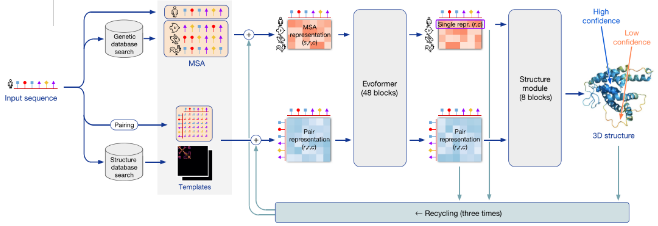
scalability

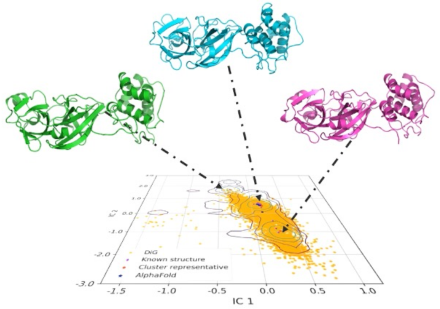

phosphoric mono-ester hydrolases (EC)
chemical pathway: glycosyl (GO-BP)
chemical pathway: pyrimidine (GO-BP)
catalysis of the hydrolys.. of an ester bond (GO-MF)
cytosol (GO-CC)

Thermodynamic ensemble

Structure prediction and sampling:

accuracy & scalability

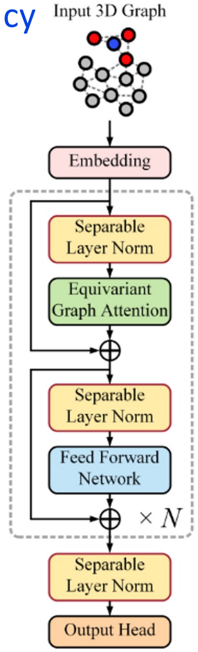

Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)



Microscopic properties

ML force field:

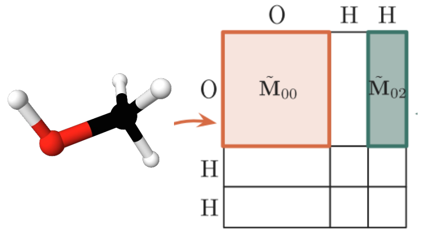
accuracy



Electronic structure

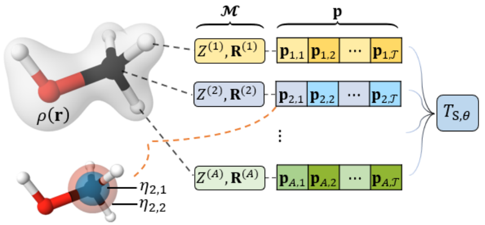
Hamiltonian prediction:

scalability

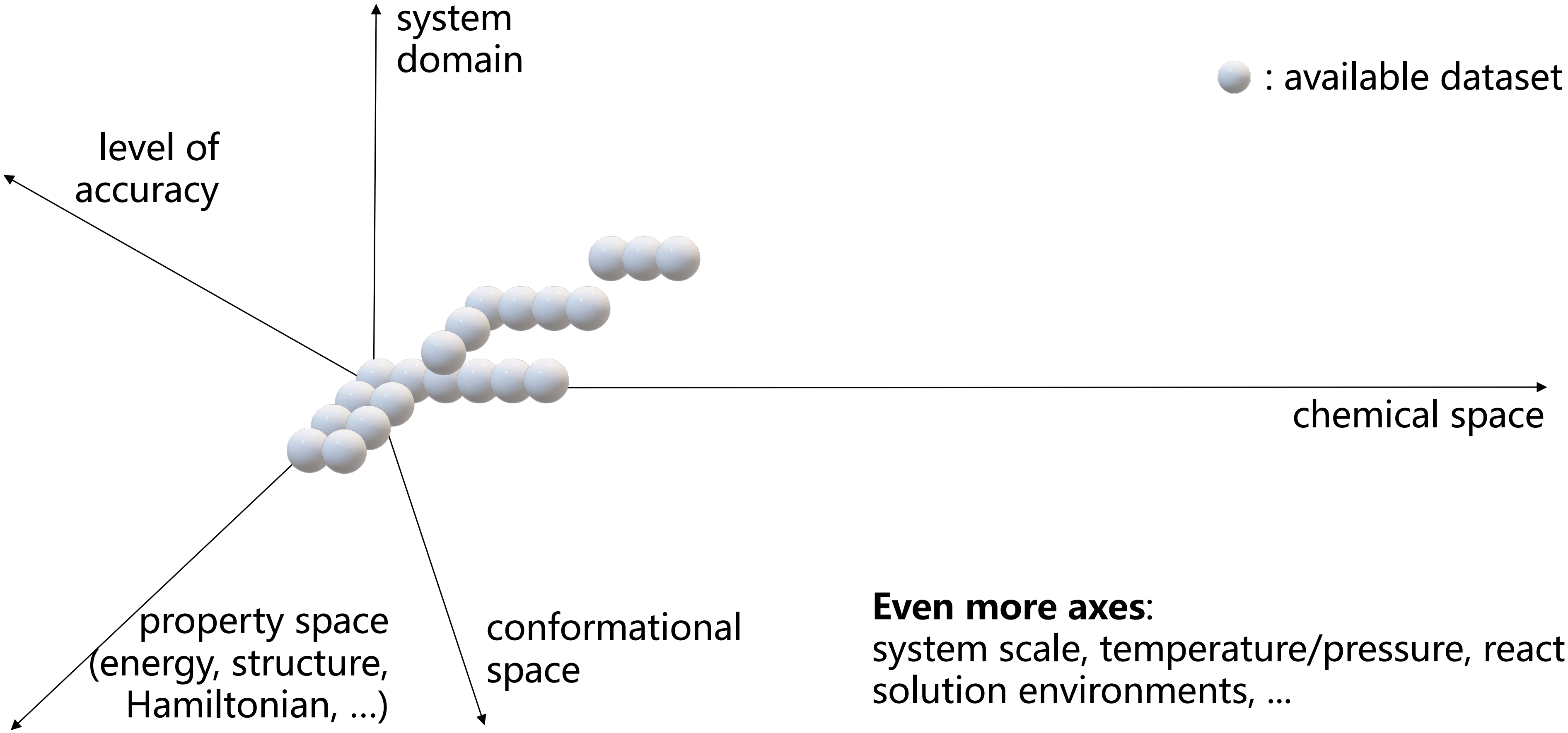


Learned functional:

accuracy


$\mathcal{M}$
p
$\rho(\mathbf{r})$
$Z^{(1)}, \mathbf{R}^{(1)}$
$Z^{(2)}, \mathbf{R}^{(2)}$
$Z^{(A)}, \mathbf{R}^{(A)}$
$\mathbf{p}_{1,1}$ $\mathbf{p}_{1,2}$ ⋯ $\mathbf{p}_{1,\mathcal{J}}$
$\mathbf{p}_{2,1}$ $\mathbf{p}_{2,2}$ ⋯ $\mathbf{p}_{2,\mathcal{J}}$
$\mathbf{p}_{A,1}$ $\mathbf{p}_{A,2}$ ⋯ $\mathbf{p}_{A,\mathcal{J}}$
$\eta_{2,1}$
$\eta_{2,2}$
$T_{S,\theta}$

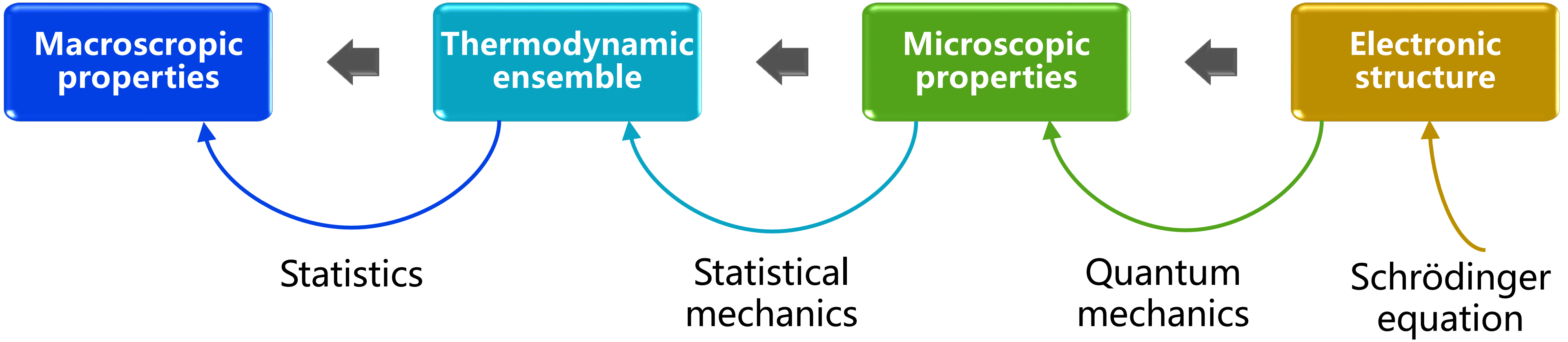
The Curse: Data Scarcity



● : available dataset

Even more axes:
system scale, temperature/pressure, reaction, solution environments, ...

The Blessing: Physical Laws!



Macroscropic properties

Thermodynamic ensemble

Microscopic properties

Electronic structure

Statistics

Statistical mechanics

Quantum mechanics

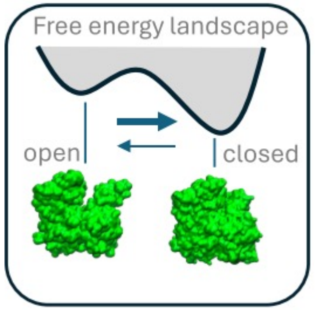
Schrödinger equation

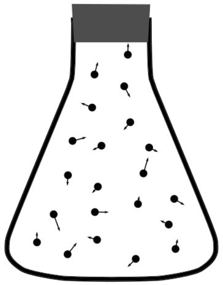
$$\Delta G_{21} = \log\frac{p_1}{p_2}$$

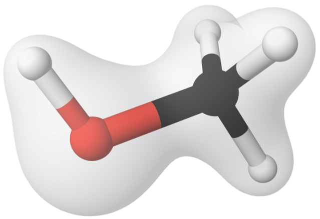
$$p(\mathbf{R}) \propto \exp\left(-\frac{E(\mathbf{R})}{kT}\right)$$

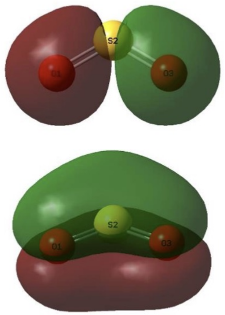
$$E(\mathbf{R}) = E[\psi_{\mathbf{R}}]$$

$$\hat{H}_{\mathbf{R}}\psi = E\psi$$

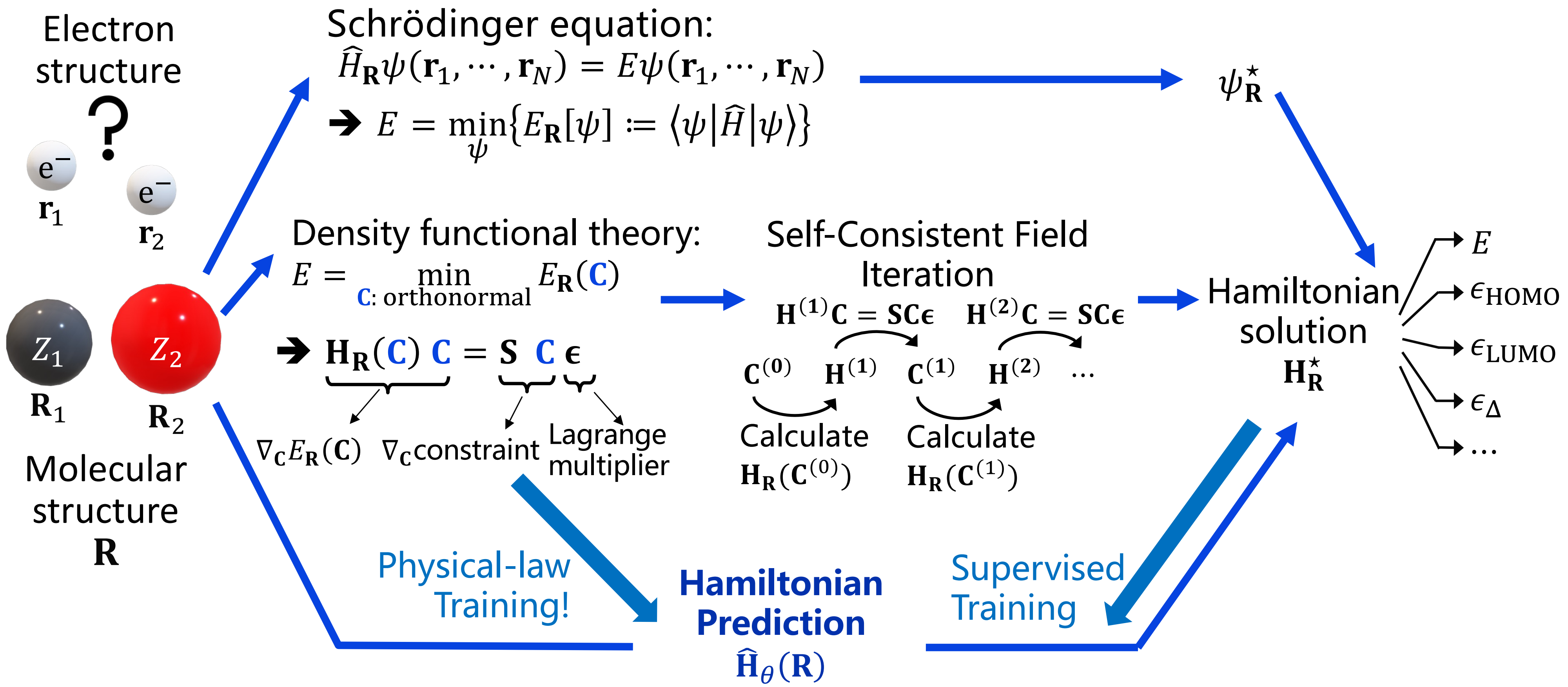








Electronic Structure



Electron structure

?

e^- $\mathbf{r}_1$ e^- $\mathbf{r}_2$

Z_1 Z_2 $\mathbf{R}_1$ $\mathbf{R}_2$

Molecular structure $\mathbf{R}$

Schrödinger equation:

$$\hat{H}_{\mathbf{R}}\psi(\mathbf{r}_1,\cdots,\mathbf{r}_N) = E\psi(\mathbf{r}_1,\cdots,\mathbf{r}_N)$$

$$\rightarrow E = \min_{\psi}\{E_{\mathbf{R}}[\psi] \coloneqq \langle\psi|\hat{H}|\psi\rangle\}$$

$\psi^{\star}_{\mathbf{R}}$

Density functional theory:

$$E = \min_{\mathbf{C}:\ \text{orthonormal}} E_{\mathbf{R}}(\mathbf{C})$$

$$\rightarrow \underbrace{\mathbf{H}_{\mathbf{R}}(\mathbf{C})\ \mathbf{C}}_{\nabla_{\mathbf{C}} E_{\mathbf{R}}(\mathbf{C})} = \underbrace{\mathbf{S}\ \mathbf{C}}_{\nabla_{\mathbf{C}}\text{constraint}}\ \underbrace{\boldsymbol{\epsilon}}_{\text{Lagrange multiplier}}$$

Self-Consistent Field Iteration

$\mathbf{H}^{(1)}\mathbf{C} = \mathbf{S}\mathbf{C}\boldsymbol{\epsilon}$ $\mathbf{H}^{(2)}\mathbf{C} = \mathbf{S}\mathbf{C}\boldsymbol{\epsilon}$

$\mathbf{C}^{(0)}$ $\mathbf{H}^{(1)}$ $\mathbf{C}^{(1)}$ $\mathbf{H}^{(2)}$ $\cdots$

Calculate $\mathbf{H}_{\mathbf{R}}(\mathbf{C}^{(0)})$ Calculate $\mathbf{H}_{\mathbf{R}}(\mathbf{C}^{(1)})$

Hamiltonian solution $\mathbf{H}^{\star}_{\mathbf{R}}$

E

ϵ_{HOMO}

ϵ_{LUMO}

ϵ_{Δ}

$\dots$

Physical-law Training!

Hamiltonian Prediction $\hat{\mathbf{H}}_{\theta}(\mathbf{R})$

Supervised Training

Zhang et al. Self-Consistency Training for Density-Functional-Theory Hamiltonian Prediction. *ICML*, 2024.

Physical-Law Training for Electronic Structure

- Self-consistency training for Hamiltonian prediction:

$$L_{\mathrm{self\_con}}(\theta) = \left\| \hat{\mathbf{H}}_\theta(\mathbf{R}) - \mathbf{H}_\mathbf{R}\left(\mathbf{C}_\mathbf{R}\left(\hat{\mathbf{H}}_\theta(\mathbf{R})\right)\right) \right\|_\mathrm{F}^2 .$$

  - **Label-free**: distinction from predicting other properties.
  - Not just a regularization: it fully **determines** the solution.

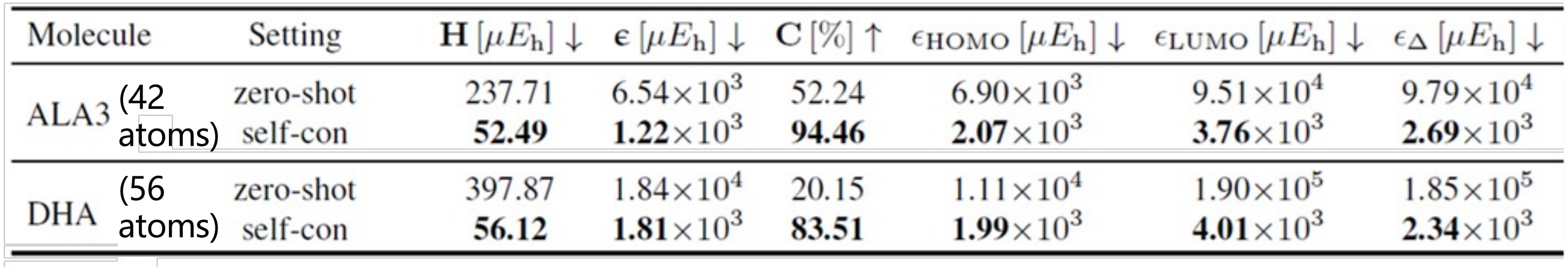
Physical-Law Training for Electronic Structure

- Generalization beyond available data

Largest labeled dataset (≤ 31 atoms) + unlabeled molecules → test on larger molecules

Hamiltonian

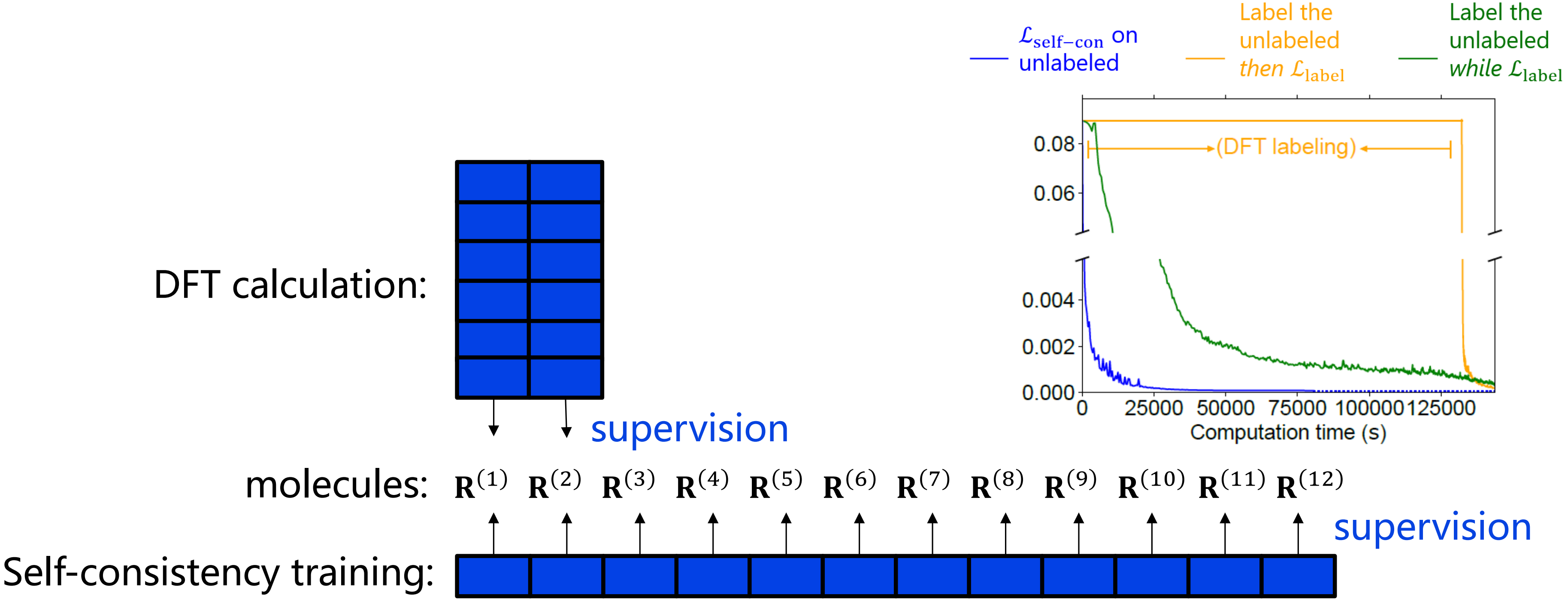
Derived properties



Molecule	Setting	$\mathbf{H}$ [μE_h] ↓	ϵ [μE_h] ↓	$\mathbf{C}$ [%] ↑	ϵ_HOMO [μE_h] ↓	ϵ_LUMO [μE_h] ↓	ϵ_Δ [μE_h] ↓
ALA3 (42 atoms)	zero-shot	237.71	6.54×10^3	52.24	6.90×10^3	9.51×10^4	9.79×10^4
	self-con	**52.49**	$\mathbf{1.22}\times10^3$	**94.46**	$\mathbf{2.07}\times10^3$	$\mathbf{3.76}\times10^3$	$\mathbf{2.69}\times10^3$
DHA (56 atoms)	zero-shot	397.87	1.84×10^4	20.15	1.11×10^4	1.90×10^5	1.85×10^5
	self-con	**56.12**	$\mathbf{1.81}\times10^3$	**83.51**	$\mathbf{1.99}\times10^3$	$\mathbf{4.01}\times10^3$	$\mathbf{2.34}\times10^3$

Physical-Law Training for Electronic Structure

- Amortization effect: more efficient than running DFT to generate labels.



DFT calculation:

supervision

molecules: $\mathbf{R}^{(1)}$ $\mathbf{R}^{(2)}$ $\mathbf{R}^{(3)}$ $\mathbf{R}^{(4)}$ $\mathbf{R}^{(5)}$ $\mathbf{R}^{(6)}$ $\mathbf{R}^{(7)}$ $\mathbf{R}^{(8)}$ $\mathbf{R}^{(9)}$ $\mathbf{R}^{(10)}$ $\mathbf{R}^{(11)}$ $\mathbf{R}^{(12)}$

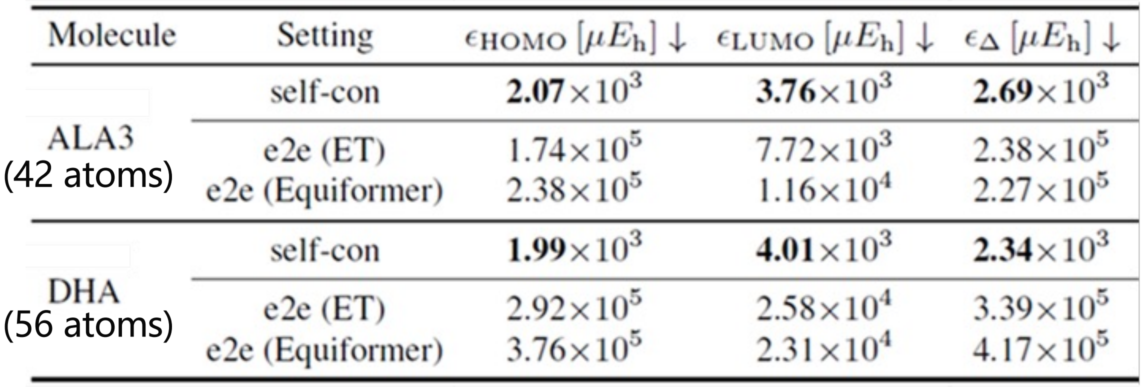
supervision

Self-consistency training:

Physical Law for Microscopic Properties

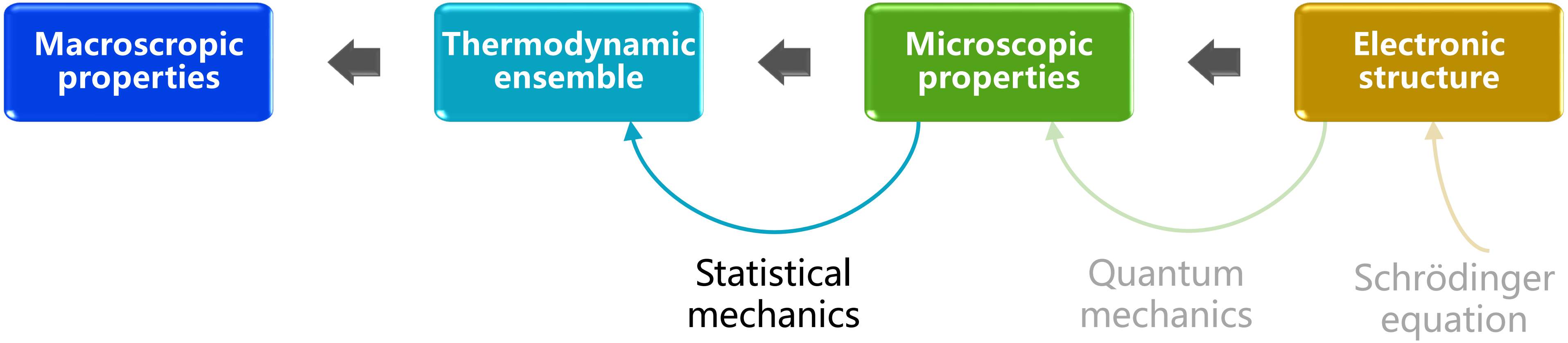
- Passing physical-law information to microscopic properties

labeled dataset (≤ 31 atoms) + unlabeled molecules → test on larger molecules



Molecule	Setting	ϵ_{HOMO} $[\mu E_{\text{h}}] \downarrow$	ϵ_{LUMO} $[\mu E_{\text{h}}] \downarrow$	ϵ_{Δ} $[\mu E_{\text{h}}] \downarrow$
ALA3 (42 atoms)	self-con	$\mathbf{2.07} \times 10^3$	$\mathbf{3.76} \times 10^3$	$\mathbf{2.69} \times 10^3$
	e2e (ET)	1.74×10^5	7.72×10^3	2.38×10^5
	e2e (Equiformer)	2.38×10^5	1.16×10^4	2.27×10^5
DHA (56 atoms)	self-con	$\mathbf{1.99} \times 10^3$	$\mathbf{4.01} \times 10^3$	$\mathbf{2.34} \times 10^3$
	e2e (ET)	2.92×10^5	2.58×10^4	3.39×10^5
	e2e (Equiformer)	3.76×10^5	2.31×10^4	4.17×10^5

The Blessing: Physical Laws!



Macroscropic properties

Thermodynamic ensemble

Microscopic properties

Electronic structure

Statistical mechanics

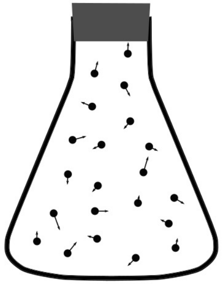
Quantum mechanics

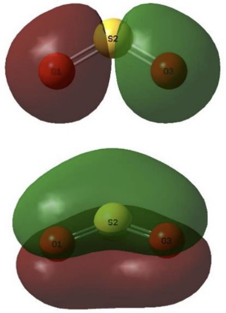
Schrödinger equation

$$p(\mathbf{R}) \propto \exp\left(-\frac{E(\mathbf{R})}{kT}\right)$$

$$E(\mathbf{R}) = E[\psi_{\mathbf{R}}]$$

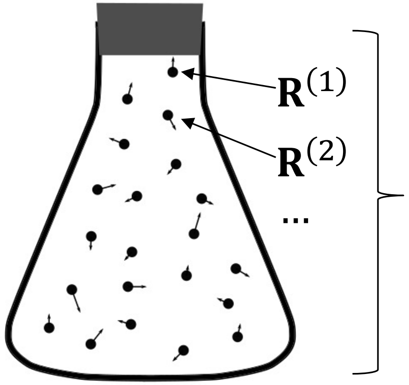
$$\hat{H}_{\mathbf{R}}\psi = E\psi$$



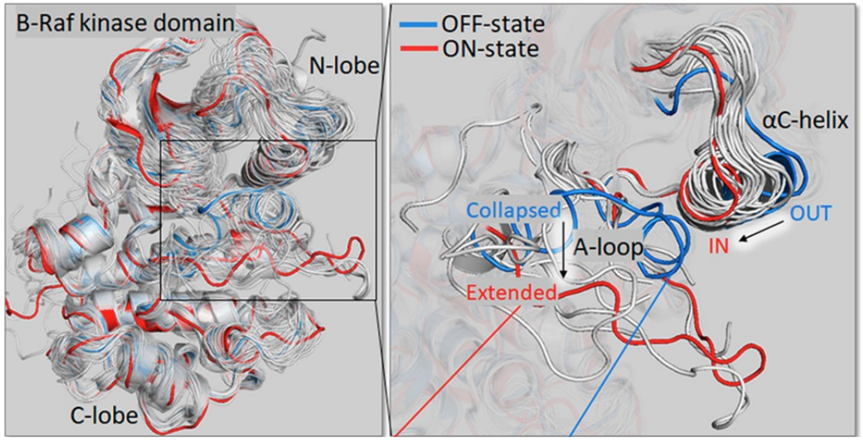


Thermodynamic Ensemble

- A molecule exists in real world with structures $\mathbf{R}$ following a distribution.

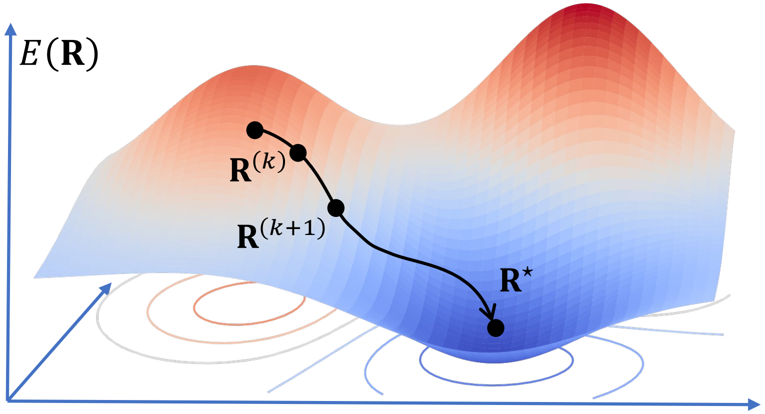


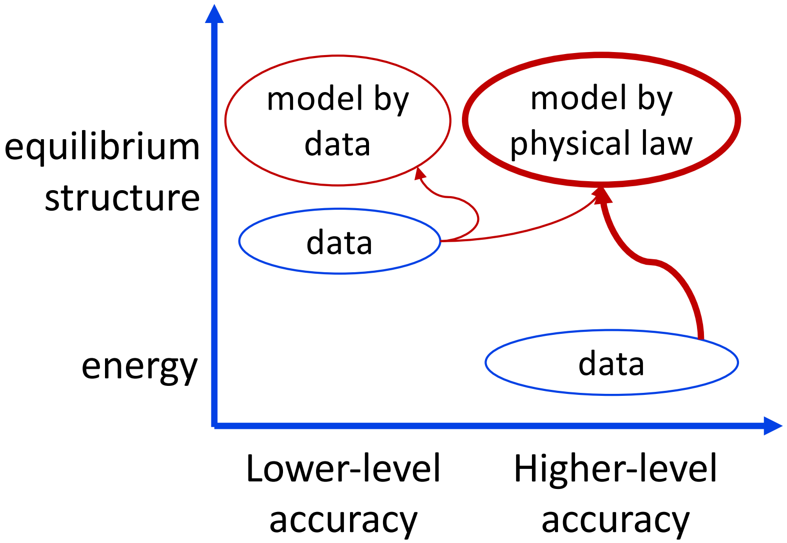
- More detailed description/knowledge



https://en.wikipedia.org/wiki/Ensemble_(mathematical_physics)

Data Heterogeneity: Energy and Equilibrium Structure





Connection between Structure and Energy

- Optimality consistency:

$\mathbf{R}^\star = \operatorname*{argmin}_{\mathbf{R}} E(\mathbf{R})$ ➔ $$\min_{\theta} \mathbb{E} \max\{0,\ E(\mathbf{R}_\theta^\star) - E(\mathbf{R}_\theta^\star + \boldsymbol{\eta})\}.$$

$\mathbf{D}_\theta(\boldsymbol{\epsilon}, t)$ for large $t \approx T$.

- Score consistency:

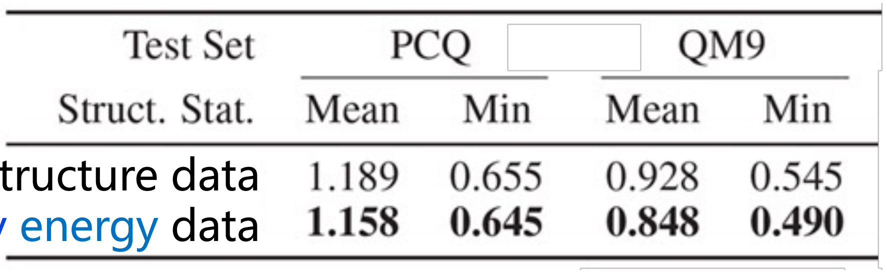
$\mathbf{R}^\star \sim \exp\left(-\frac{E(\mathbf{R})}{k\mathcal{T}}\right)$ for small $\mathcal{T}$ ➔ $$\min_{\theta} \mathbb{E}_{\mathbf{R}} \left\| \nabla \log p_\theta(\mathbf{R}) + \frac{\nabla E(\mathbf{R})}{k\mathcal{T}} \right\|^2.$$

$\frac{\alpha_t\, \mathbf{D}_\theta(\mathbf{R},t) - \mathbf{R}}{\sigma_t^2}$ for small $t \approx 0$

Ren et al. Physical Consistency Bridges Heterogeneous Data in Molecular Multi-Task Learning. *NeurIPS*, 2024.

Accuracy beyond Training Data

Evaluation:
error w.r.t high-accuracy structure



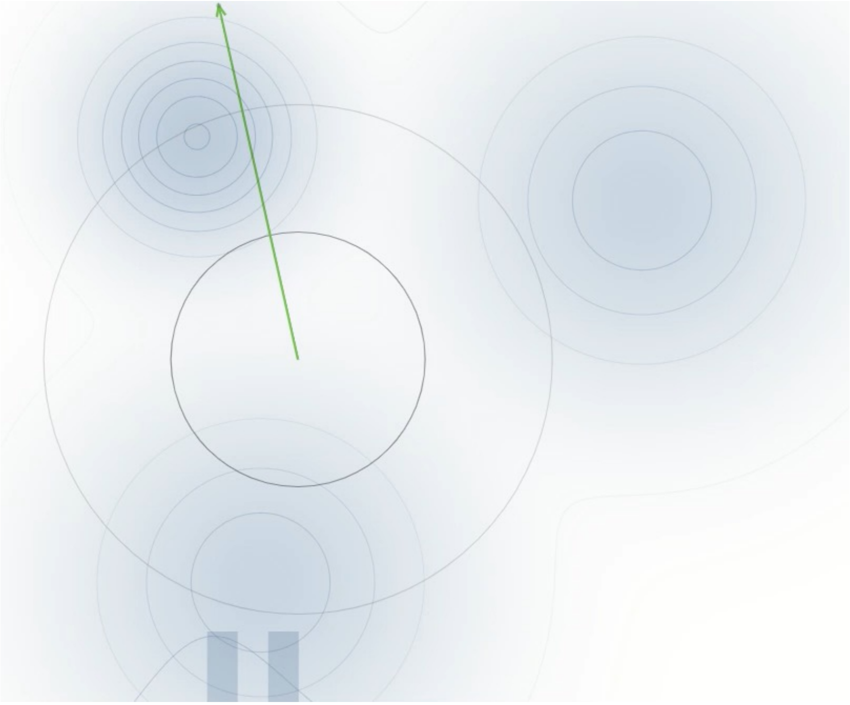
Training:

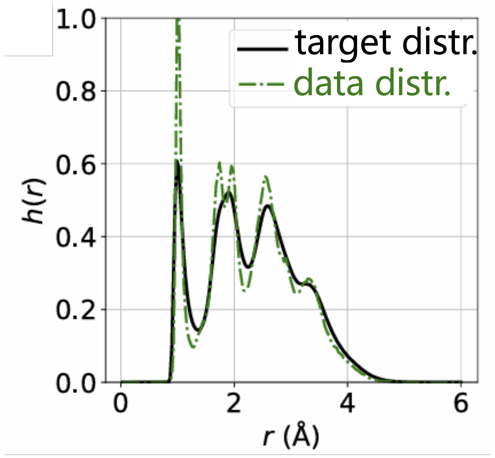
Test Set	PCQ		QM9	
Struct. Stat.	Mean	Min	Mean	Min
low-accuracy structure data	1.189	0.655	0.928	0.545
+high-accuracy energy data	**1.158**	**0.645**	**0.848**	**0.490**

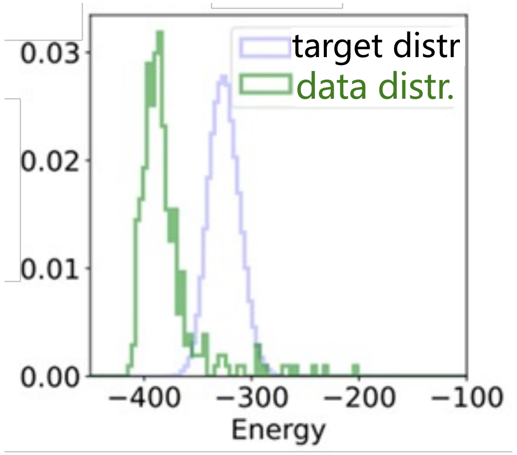
Ren et al. Physical Consistency Bridges Heterogeneous Data in Molecular Multi-Task Learning. *NeurIPS*, 2024.

Problem with Learning Thermodynamic Ensemble

- Data are biased from finite-length simulation







Connection between Distribution and Energy

- Energy function defines the target: $p_0(\mathbf{R}_0) \propto \exp\left(-\frac{E(\mathbf{R}_0)}{k\mathcal{T}}\right)$.
- Diffusion-model learning target:

$$\mathbf{s}_\theta(\mathbf{R}_t, t) \to \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{\underbrace{p_0(\mathbf{R}_0)}_{\approx q(\mathbf{R}_0)\ \text{data distribution}} p(\mathbf{R}_t|\mathbf{R}_0)} \left\| \mathbf{s}_\theta(\mathbf{R}_t, t) - \frac{1}{\alpha_t} \underbrace{\nabla \log p_0(\mathbf{R}_0)}_{-\frac{\nabla E(\mathbf{R}_0)}{k\mathcal{T}}} \right\|^2.$$

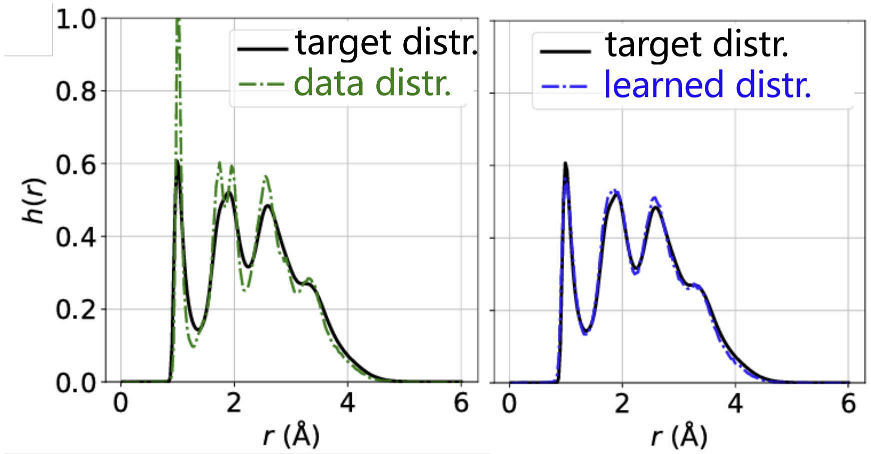
$$\mathbf{s}_{\theta_{\text{debiased}}}(\mathbf{R}_t, t) \to$$

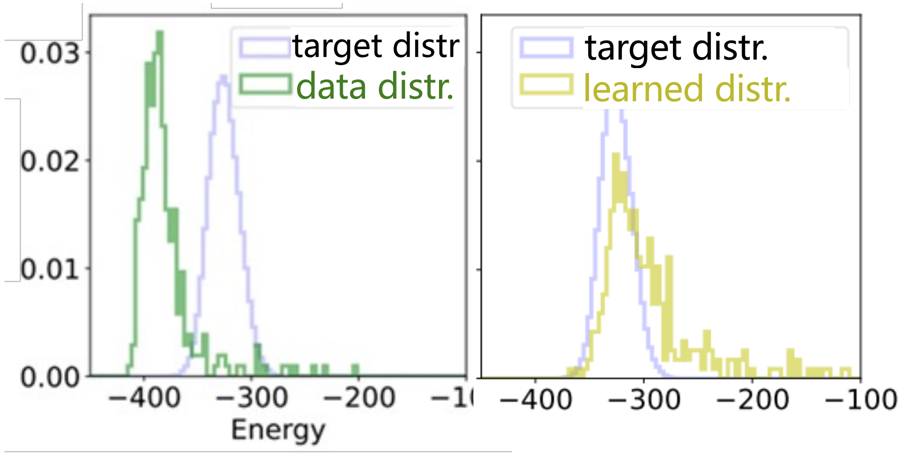
- Energy function corrects data bias:

$$\left\| \mathbf{s}_{\theta_{\text{debiased}}}(\mathbf{R}_t, t) - \nabla \log p_t(\mathbf{R}_t) \right\|^2 \leq \left\| \mathbf{s}_{\theta_{\text{data}}}(\mathbf{R}_t, t) - \nabla \log p_t(\mathbf{R}_t) \right\|^2.$$

Guo et al. Potential Score Matching: Debiasing Molecular Structure Sampling with Potential Energy Guidance. *TMLR*, 2025.

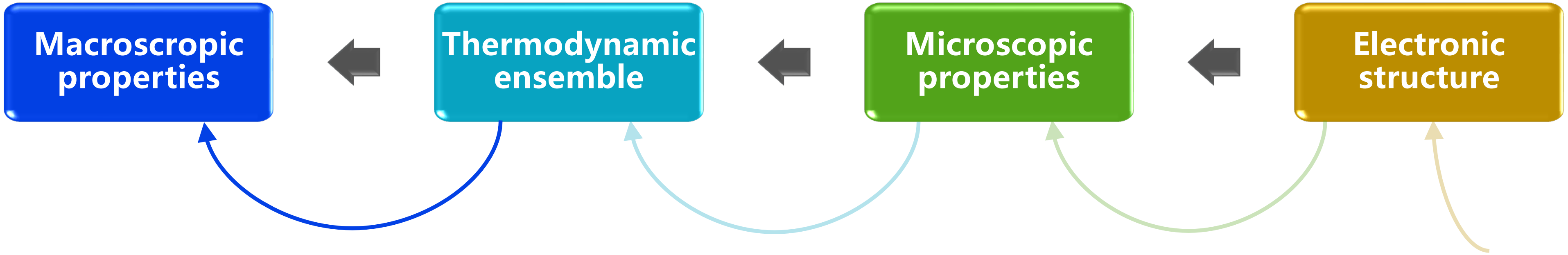
Accuracy beyond Training Data





Guo et al. Potential Score Matching: Debiasing Molecular Structure Sampling with Potential Energy Guidance. *TMLR*, 2025.

The Blessing: Physical Laws!



Macroscopic properties

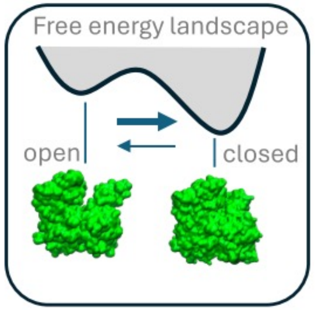
Thermodynamic ensemble

Microscopic properties

Electronic structure

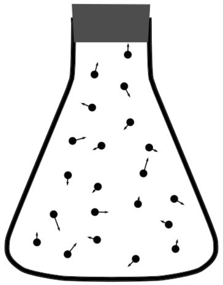
Statistics

$$\Delta G_{21} = \log\frac{p_1}{p_2}$$



Statistical mechanics

$$p(\mathbf{R}) \propto \exp\left(-\frac{E(\mathbf{R})}{kT}\right)$$

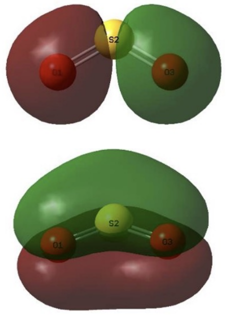


Quantum mechanics

$$E(\mathbf{R}) = E[\psi_{\mathbf{R}}]$$

Schrödinger equation

$$\hat{H}_{\mathbf{R}}\psi = E\psi$$

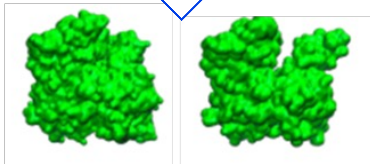


Stability Calculation from Ensemble Model

Thermodynamic ensemble $p_\theta(\mathbf{R})$

Foldedness $x = f(\mathbf{R})$

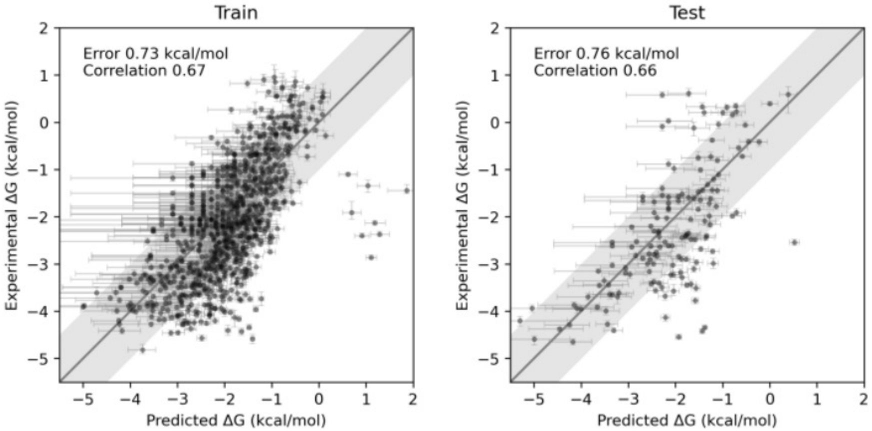
Regularize/ calibrate



p_{folded}, p_{unfolded}

Stability $\log \frac{p_{\text{folded}}}{p_{\text{unfolded}}}$

ΔG calculation



outperforms black-box methods that directly predict ΔG from sequences.

Folding free energy $\Delta G = \log \frac{p_{\text{folded}}}{p_{\text{unfolded}}}$

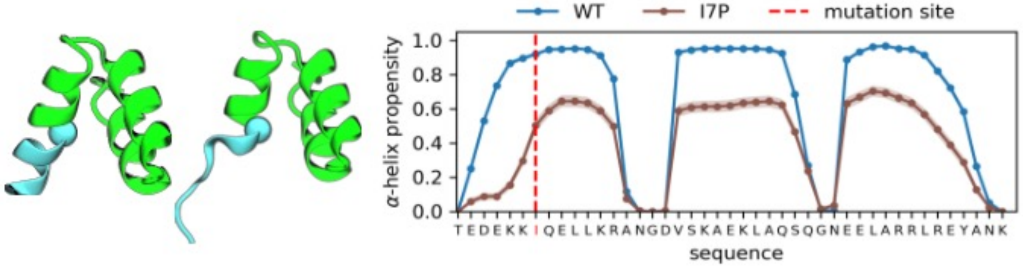
can be experimentally measured!

Lewis et al., Scalable emulation of protein equilibrium ensembles with generative deep learning, *Science*, 2025.

Stability Calculation from Ensemble Model

Thermodynamic ensemble $p_\theta(\mathbf{R})$

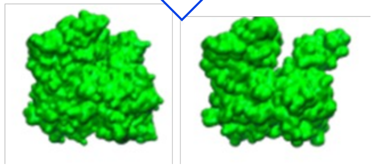
$\Delta\Delta G$ calculation (mutation-caused stability change) with explanation



Wild type HHH_rd1_0335, Mutation I7P

Foldedness $x = f(\mathbf{R})$

Regularize/calibrate



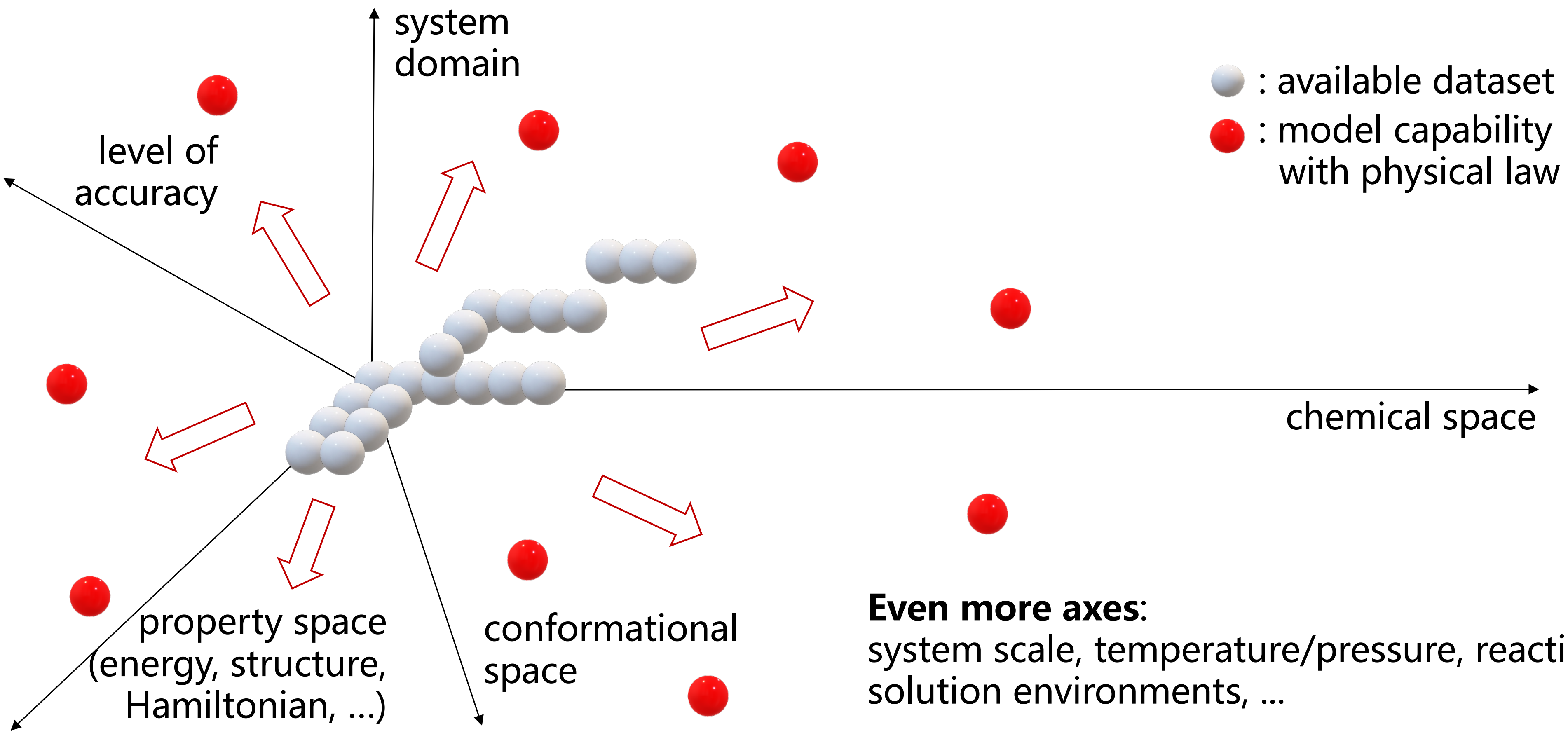
p_{folded}, p_{unfolded}

Stability $\log \frac{p_{\text{folded}}}{p_{\text{unfolded}}}$

Folding free energy $\Delta G = \log \frac{p_{\text{folded}}}{p_{\text{unfolded}}}$ can be experimentally measured!

Lewis et al., Scalable emulation of protein equilibrium ensembles with generative deep learning, *Science*, 2025.

Physical Law: Reasoning in Scientific Computing


system domain
level of accuracy
chemical space
property space (energy, structure, Hamiltonian, ...)
conformational space
○ : available dataset
● : model capability with physical law

Even more axes:
system scale, temperature/pressure, reaction, solution environments, ...

Thank you