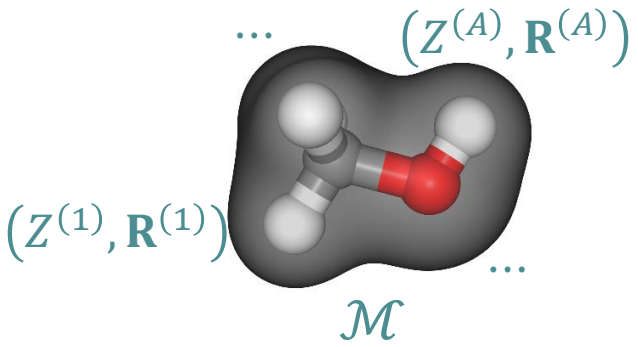
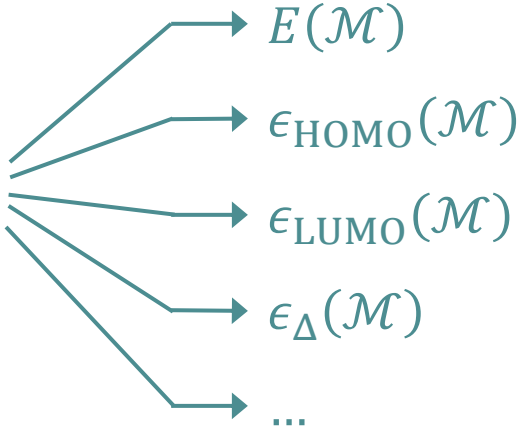
Microsoft

Self-Consistency Training for Density-Functional-Theory Hamiltonian Prediction

Chang Liu
on behalf of the team
changliu@microsoft.com

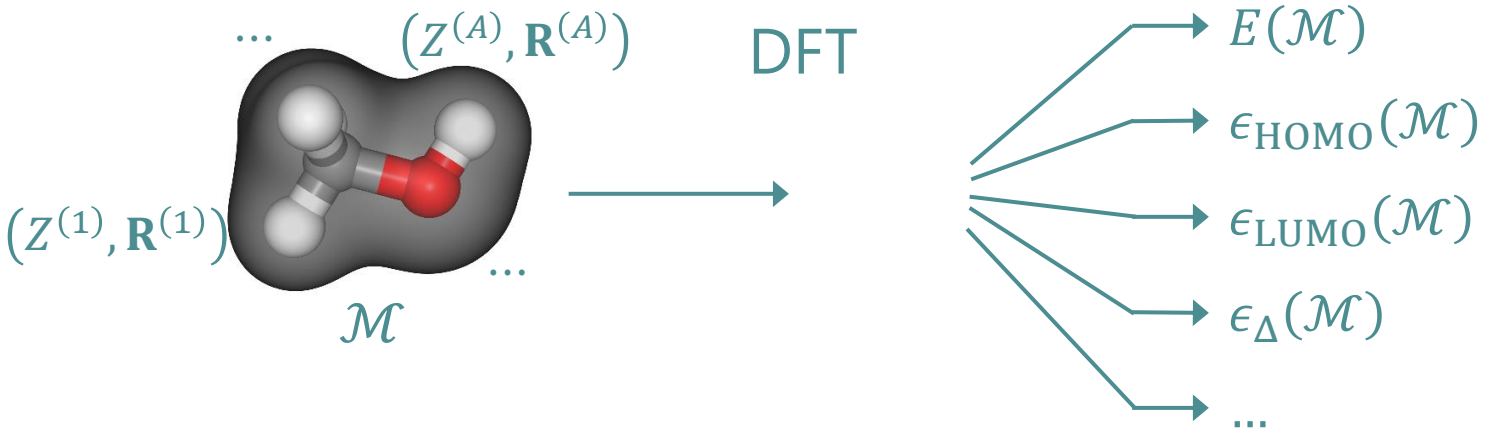
Hamiltonian Prediction





- Molecular properties: interaction among electrons and atomic nuclei.

Hamiltonian Prediction



- Molecular properties: interaction among electrons and atomic nuclei.
- DFT: solve electronic structure hence properties.

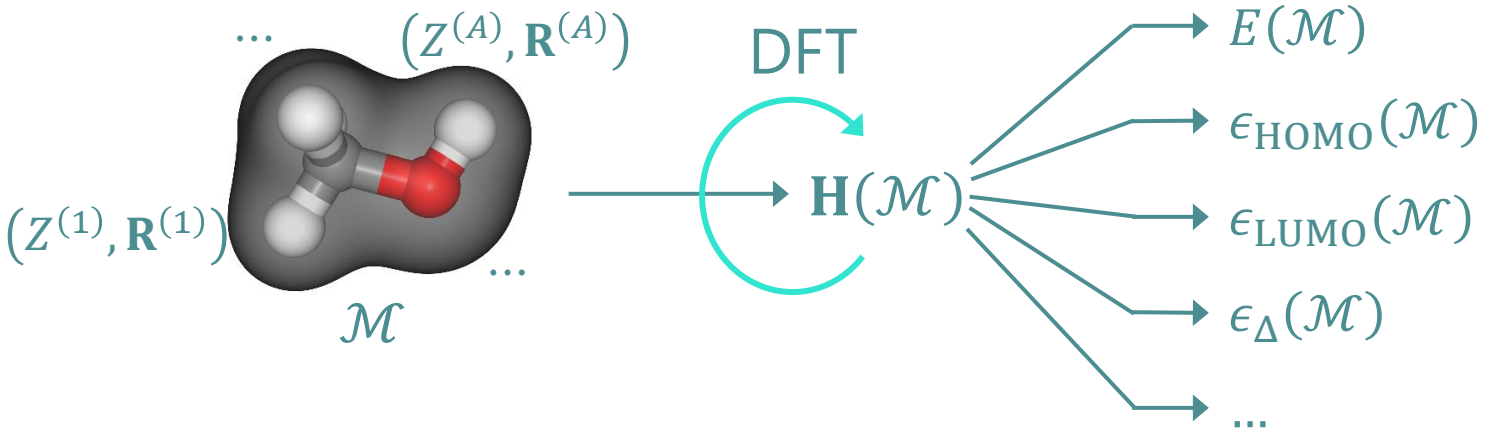
Hamiltonian Prediction

- Molecular properties: interaction among electrons and atomic nuclei.
- DFT: solve electronic structure hence properties.
- Hamiltonian: raw DFT solution, derive all properties.

Hamiltonian Prediction

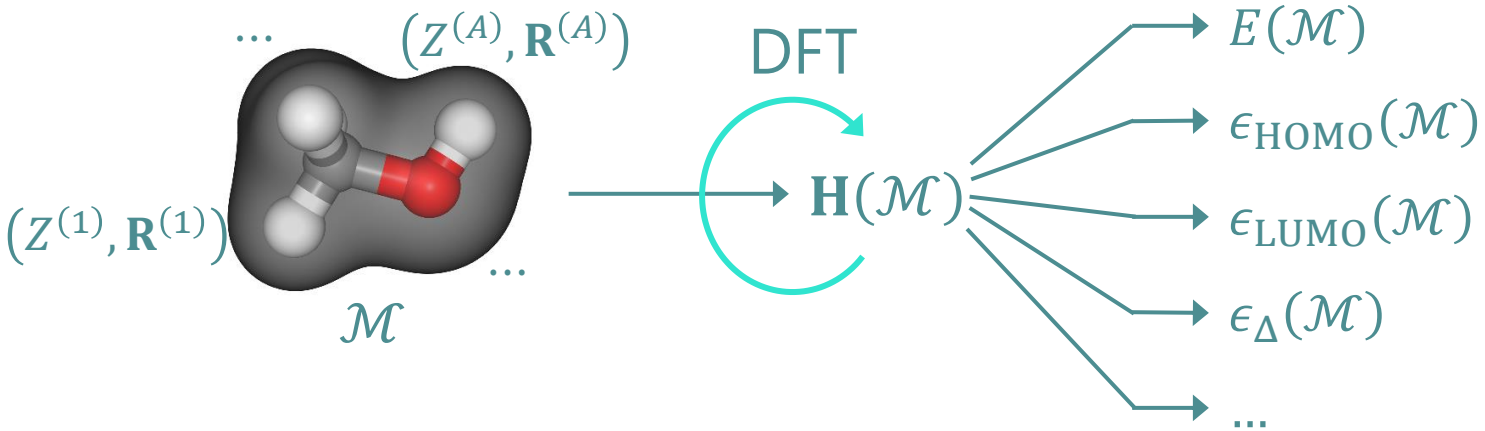
- Molecular properties: interaction among electrons and atomic nuclei.
- DFT: solve electronic structure hence properties.
- Hamiltonian: raw DFT solution, derive all properties.
- Hamiltonian prediction:
  "root property" prediction, provide all properties that DFT can.

Hamiltonian Prediction



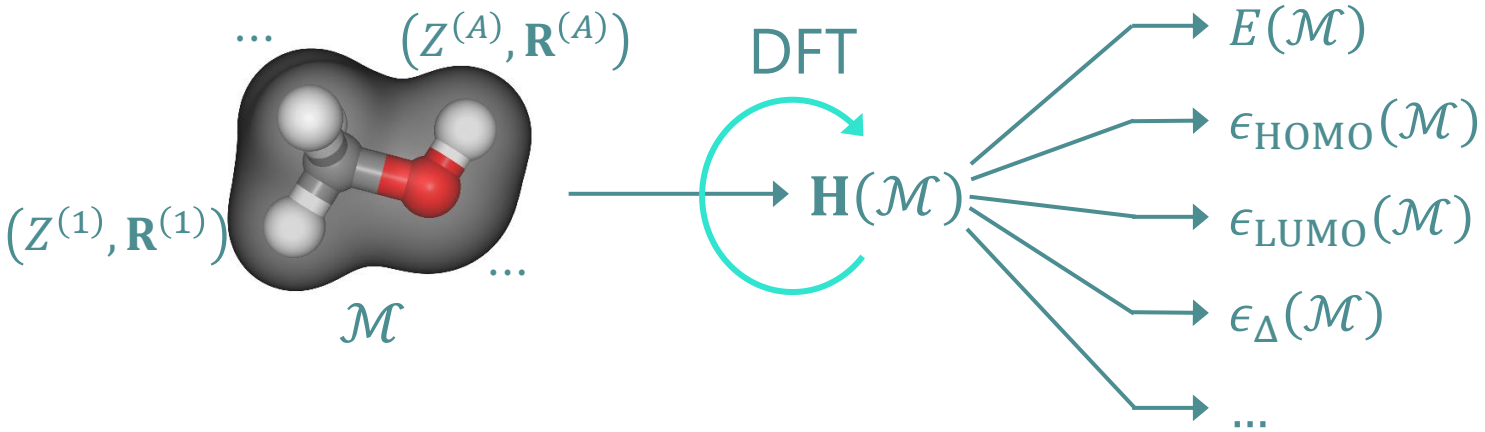
- Hamiltonian prediction has a self-consistency principle: **Training without label!**
  - Distinction from common property prediction: data-free training / self-improvement.
  - Compensating data scarcity with scientific laws.

Hamiltonian Prediction



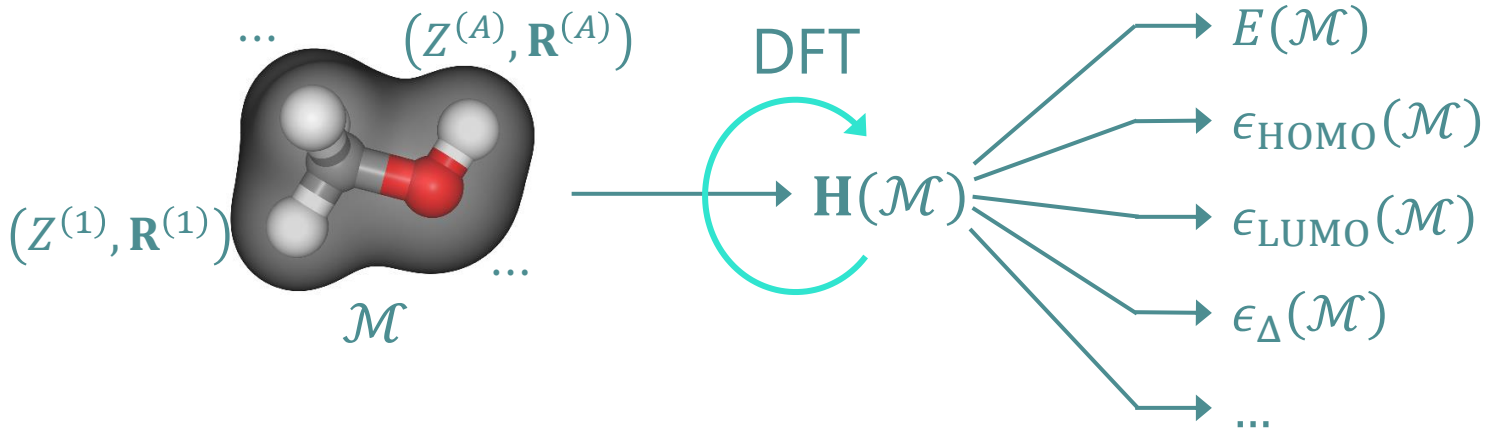
- Hamiltonian prediction has a self-consistency principle: **Training without label!**
  - Distinction from common property prediction: data-free training / self-improvement.
  - Compensating data scarcity with scientific laws.
- Unique benefits:
  - Exact generalization to arbitrary workload beyond labeled data (also for other properties).

Hamiltonian Prediction



- Hamiltonian prediction has a self-consistency principle: **Training without label!**
  - Distinction from common property prediction: data-free training / self-improvement.
  - Compensating data scarcity with scientific laws.
- Unique benefits:
  - Exact generalization to arbitrary workload beyond labeled data (also for other properties).
  - Amortization of DFT calculation: more efficient than running DFT to generate labels.

Hamiltonian Prediction



- Hamiltonian prediction has a self-consistency principle: **Training without label!**
  - Distinction from common property prediction: data-free training / self-improvement.
  - Compensating data scarcity with scientific laws.
- Unique benefits:
  - Exact generalization to arbitrary workload beyond labeled data (also for other properties).
  - Amortization of DFT calculation: more efficient than running DFT to generate labels.
  - Extending applicable scale of Hamiltonian prediction.

Background: DFT Formulation

- Describe the N-electron state by orbitals $\{\phi_i(\mathbf{r})\}_{i=1}^{N}$.

Background: DFT Formulation

- Describe the N-electron state by orbitals $\{\phi_i(\mathbf{r})\}_{i=1}^{N}$.
- Expressed as coefficients $\mathbf{C}$ under a basis set: $\phi_i(\mathbf{r}) = \sum_{\alpha=1}^{B} \mathbf{C}_{\alpha i}\, \eta_{\mathcal{M},\alpha}(\mathbf{r})$.

Background: DFT Formulation

- Describe the N-electron state by orbitals $\{\phi_i(\mathbf{r})\}_{i=1}^{N}$.
- Expressed as coefficients $\mathbf{C}$ under a basis set: $\phi_i(\mathbf{r}) = \sum_{\alpha=1}^{B} \mathbf{C}_{\alpha i}\, \eta_{\mathcal{M},\alpha}(\mathbf{r})$.
- Solve for $\mathbf{C}$ by minimizing $E_{\mathcal{M}}(\mathbf{C}\mathbf{C}^{\top})$, s.t. $\mathbf{C}^{\top}\mathbf{S}_{\mathcal{M}}\,\mathbf{C} = \mathbf{I}$.
  - $E_{\mathcal{M}}(\mathbf{C}\mathbf{C}^{\top})$ is a known, explicit function (given an approximate XC functional).
  - $(\mathbf{S}_{\mathcal{M}})_{\alpha\beta} \coloneqq \langle \eta_{\mathcal{M},\alpha} | \eta_{\mathcal{M},\beta} \rangle$.

Background: DFT Formulation

- Describe the N-electron state by orbitals $\{\phi_i(\mathbf{r})\}_{i=1}^{N}$.
- Expressed as coefficients $\mathbf{C}$ under a basis set: $\phi_i(\mathbf{r}) = \sum_{\alpha=1}^{B} \mathbf{C}_{\alpha i}\, \eta_{\mathcal{M},\alpha}(\mathbf{r})$.
- Solve for $\mathbf{C}$ by minimizing $E_{\mathcal{M}}(\mathbf{C}\mathbf{C}^{\top})$, s.t. $\mathbf{C}^{\top}\mathbf{S}_{\mathcal{M}}\,\mathbf{C} = \mathbf{I}$.
  - $E_{\mathcal{M}}(\mathbf{C}\mathbf{C}^{\top})$ is a known, explicit function (given an approximate XC functional).
  - $(\mathbf{S}_{\mathcal{M}})_{\alpha\beta} \coloneqq \langle \eta_{\mathcal{M},\alpha} | \eta_{\mathcal{M},\beta} \rangle$.
- Solve the optimization problem: $\nabla_{\mathbf{C}}\left[E_{\mathcal{M}}(\mathbf{C}\mathbf{C}^{\top}) - \mathrm{tr}\left((\mathbf{C}^{\top}\mathbf{S}_{\mathcal{M}}\mathbf{C} - \mathbf{I})\boldsymbol{\epsilon}\right)\right] = 0$

Background: DFT Formulation

- Describe the N-electron state by orbitals $\{\phi_i(\mathbf{r})\}_{i=1}^{N}$.
- Expressed as coefficients $\mathbf{C}$ under a basis set: $\phi_i(\mathbf{r}) = \sum_{\alpha=1}^{B} \mathbf{C}_{\alpha i}\, \eta_{\mathcal{M},\alpha}(\mathbf{r})$.
- Solve for $\mathbf{C}$ by minimizing $E_{\mathcal{M}}(\mathbf{C}\mathbf{C}^{\top})$, s.t. $\mathbf{C}^{\top}\mathbf{S}_{\mathcal{M}}\,\mathbf{C} = \mathbf{I}$.
  - $E_{\mathcal{M}}(\mathbf{C}\mathbf{C}^{\top})$ is a known, explicit function (given an approximate XC functional).
  - $(\mathbf{S}_{\mathcal{M}})_{\alpha\beta} \coloneqq \langle \eta_{\mathcal{M},\alpha} | \eta_{\mathcal{M},\beta} \rangle$.
- Solve the optimization problem: $\nabla_{\mathbf{C}}\left[E_{\mathcal{M}}(\mathbf{C}\mathbf{C}^{\top}) - \mathrm{tr}\left((\mathbf{C}^{\top}\mathbf{S}_{\mathcal{M}}\mathbf{C} - \mathbf{I})\boldsymbol{\epsilon}\right)\right] = 0$

$$\rightarrow \underbrace{\mathbf{H}_{\mathcal{M}}(\mathbf{C})}_{\coloneqq \nabla E_{\mathcal{M}}(\cdot)|_{\mathbf{C}\mathbf{C}^{\top}}}\mathbf{C} = \mathbf{S}_{\mathcal{M}}\,\mathbf{C}\,\boldsymbol{\epsilon}.$$

Kohn-Sham equation

DFT Calculation → Self-Consistency Training

Kohn-Sham equation

$$\mathbf{H}_{\mathcal{M}}(\mathbf{C})\,\mathbf{C} = \mathbf{S}_{\mathcal{M}}\,\mathbf{C}\,\boldsymbol{\epsilon}$$

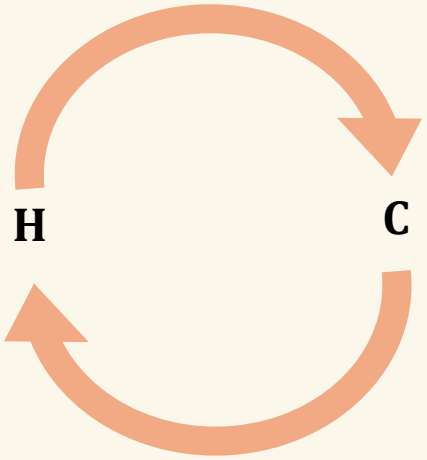
DFT Calculation → Self-Consistency Training

Kohn-Sham equation

$$\mathbf{H}_{\mathcal{M}}(\mathbf{C})\,\mathbf{C} = \mathbf{S}_{\mathcal{M}}\,\mathbf{C}\,\boldsymbol{\epsilon}$$

$\Longleftrightarrow$

$\mathbf{C}_{\mathcal{M}}(\mathbf{H})$:
solution to $\mathbf{H}\,\mathbf{C} = \mathbf{S}_{\mathcal{M}}\,\mathbf{C}\,\boldsymbol{\epsilon}$

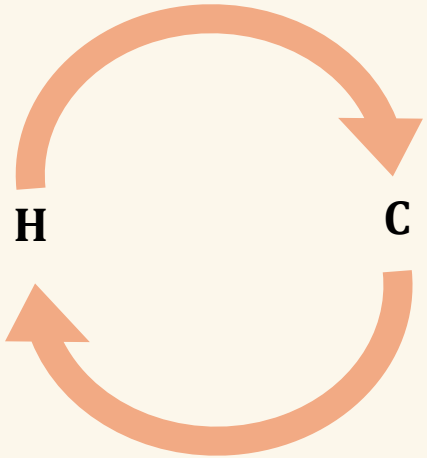


DFT Calculation → Self-Consistency Training

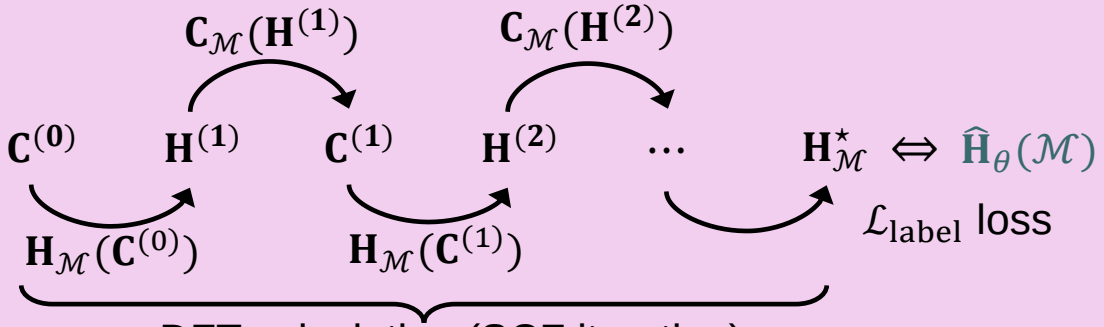
Kohn-Sham equation

$$\mathbf{H}_{\mathcal{M}}(\mathbf{C})\,\mathbf{C} = \mathbf{S}_{\mathcal{M}}\,\mathbf{C}\,\boldsymbol{\epsilon}$$

$\mathbf{C}_{\mathcal{M}}(\mathbf{H})$:
solution to $\mathbf{H}\,\mathbf{C} = \mathbf{S}_{\mathcal{M}}\,\mathbf{C}\,\boldsymbol{\epsilon}$



Supervised training

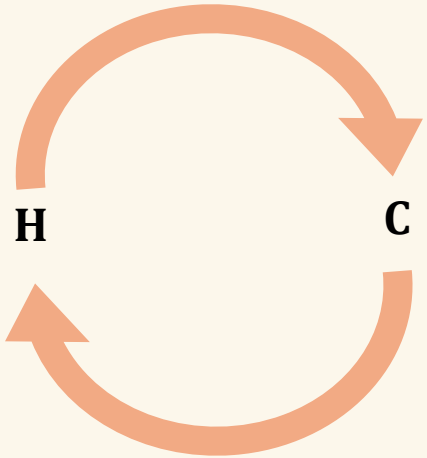


DFT Calculation → Self-Consistency Training

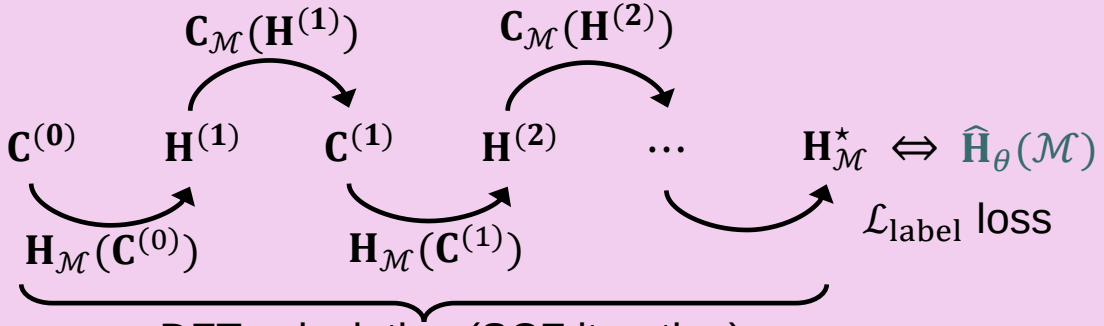
Kohn-Sham equation

$$\mathbf{H}_{\mathcal{M}}(\mathbf{C})\,\mathbf{C} = \mathbf{S}_{\mathcal{M}}\,\mathbf{C}\,\boldsymbol{\epsilon}$$

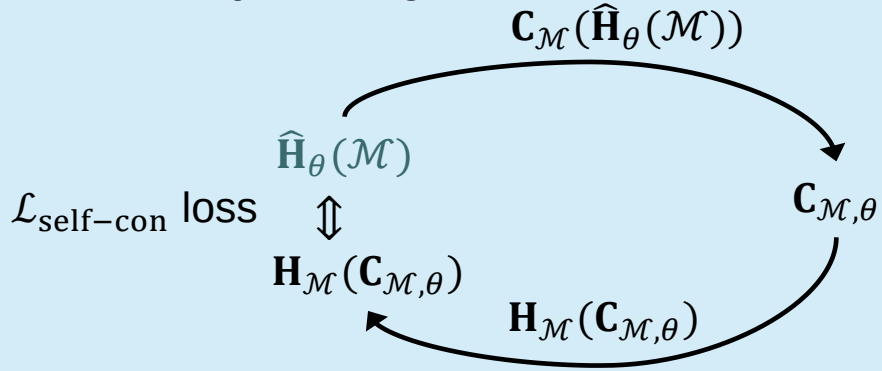
$\mathbf{C}_{\mathcal{M}}(\mathbf{H})$:
solution to $\mathbf{H}\,\mathbf{C} = \mathbf{S}_{\mathcal{M}}\,\mathbf{C}\,\boldsymbol{\epsilon}$



Supervised training



Self-Consistency training



$$\mathcal{L}_{\text{self-con}}(\theta;\mathcal{D}) \coloneqq \frac{1}{|\mathcal{D}|}\sum_{\mathcal{M}\sim\mathcal{D}} \left\| \widehat{\mathbf{H}}_{\theta}(\mathcal{M}) - \mathbf{H}_{\mathcal{M}}\left(\mathbf{C}_{\mathcal{M}}\left(\widehat{\mathbf{H}}_{\theta}(\mathcal{M})\right)\right)\right\|_{\mathrm{F}}^{2}.$$

Self-Consistency Training

- Self-consistency loss:

$$\mathcal{L}_{\text{self-con}}(\theta;\mathcal{D}) \coloneqq \frac{1}{|\mathcal{D}|}\sum_{\mathcal{M}\sim\mathcal{D}} \left\| \widehat{\mathbf{H}}_\theta(\mathcal{M}) - \mathbf{H}_{\mathcal{M}}\left(\mathbf{C}_{\mathcal{M}}\left(\widehat{\mathbf{H}}_\theta(\mathcal{M})\right)\right)\right\|_{\mathrm{F}}^2.$$

  - Not just a regularization: it **determines** the DFT solution (label).

Self-Consistency Training

- Self-consistency loss:

$$\mathcal{L}_{\text{self-con}}(\theta;\mathcal{D}) \coloneqq \frac{1}{|\mathcal{D}|}\sum_{\mathcal{M}\sim\mathcal{D}} \left\| \widehat{\mathbf{H}}_\theta(\mathcal{M}) - \mathbf{H}_{\mathcal{M}}\left(\mathbf{C}_{\mathcal{M}}\left(\widehat{\mathbf{H}}_\theta(\mathcal{M})\right)\right)\right\|_{\mathrm{F}}^2.$$

  - Not just a regularization: it **determines** the DFT solution (label).
  - Minimizing the gap **unnecessarily** drives $\widehat{\mathbf{H}}_\theta(\mathcal{M})$ towards $\mathbf{H}_{\mathcal{M}}\left(\mathbf{C}_{\mathcal{M}}\left(\widehat{\mathbf{H}}_\theta(\mathcal{M})\right)\right)$:
    - The latter may even be farther from the solution, in which case both are driven to the solution.
    - Should not apply stop-gradient to the latter.

Self-Consistency Training

- Self-consistency loss:

$$\mathcal{L}_{\text{self-con}}(\theta;\mathcal{D}) \coloneqq \frac{1}{|\mathcal{D}|}\sum_{\mathcal{M}\sim\mathcal{D}} \left\| \widehat{\mathbf{H}}_\theta(\mathcal{M}) - \mathbf{H}_{\mathcal{M}}\left(\mathbf{C}_{\mathcal{M}}\left(\widehat{\mathbf{H}}_\theta(\mathcal{M})\right)\right)\right\|_{\mathrm{F}}^2.$$

  - Not just a regularization: it **determines** the DFT solution (label).
  - Minimizing the gap **unnecessarily** drives $\widehat{\mathbf{H}}_\theta(\mathcal{M})$ towards $\mathbf{H}_{\mathcal{M}}\left(\mathbf{C}_{\mathcal{M}}\left(\widehat{\mathbf{H}}_\theta(\mathcal{M})\right)\right)$:
    - The latter may even be farther from the solution, in which case both are driven to the solution.
    - Should not apply stop-gradient to the latter.
  - Numerically stable implementation of differentiation through eigensolver.
  - GPU implementation of Hamiltonian construction $\mathbf{H}_{\mathcal{M}}(\mathbf{C})$.

Unique Benefits

1. Generalization beyond labeled data: $\mathcal{L}_{\text{label}}\left(\theta; \underbrace{\overline{\mathcal{D}^{(1)}}}_{\text{limited labeled dataset}}\right) + \lambda\, \mathcal{L}_{\text{self-con}}\left(\theta; \underbrace{\mathcal{D}^{(2)}}_{\text{unlimited unlabeled dataset}}\right)$.

Unique Benefits

1. Generalization beyond labeled data: $\mathcal{L}_{\text{label}}\left(\theta; \overline{\mathcal{D}^{(1)}}\right) + \lambda\, \mathcal{L}_{\text{self-con}}\left(\theta; \mathcal{D}^{(2)}\right)$.
   limited labeled dataset — unlimited unlabeled dataset
2. Amortization effect: efficiency over DFT labeling.

Cost of one iteration: 

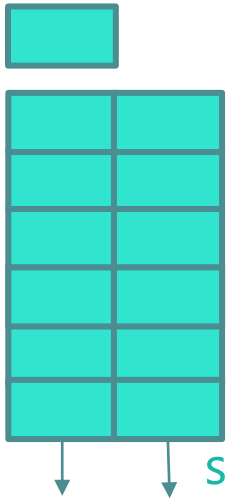
Training molecules: $\mathcal{M}^{(1)}\mathcal{M}^{(2)}\mathcal{M}^{(3)}\mathcal{M}^{(4)}\mathcal{M}^{(5)}\mathcal{M}^{(6)}\mathcal{M}^{(7)}\mathcal{M}^{(8)}\mathcal{M}^{(9)}\mathcal{M}^{(10)}\mathcal{M}^{(11)}\mathcal{M}^{(12)}$

Unique Benefits

1. Generalization beyond labeled data: $\mathcal{L}_{\text{label}}\left(\theta; \underbrace{\overline{\mathcal{D}^{(1)}}}_{\text{limited labeled dataset}}\right) + \lambda\, \mathcal{L}_{\text{self-con}}\left(\theta; \underbrace{\mathcal{D}^{(2)}}_{\text{unlimited unlabeled dataset}}\right)$.
2. Amortization effect: efficiency over DFT labeling.

Cost of one iteration:

DFT calculation:

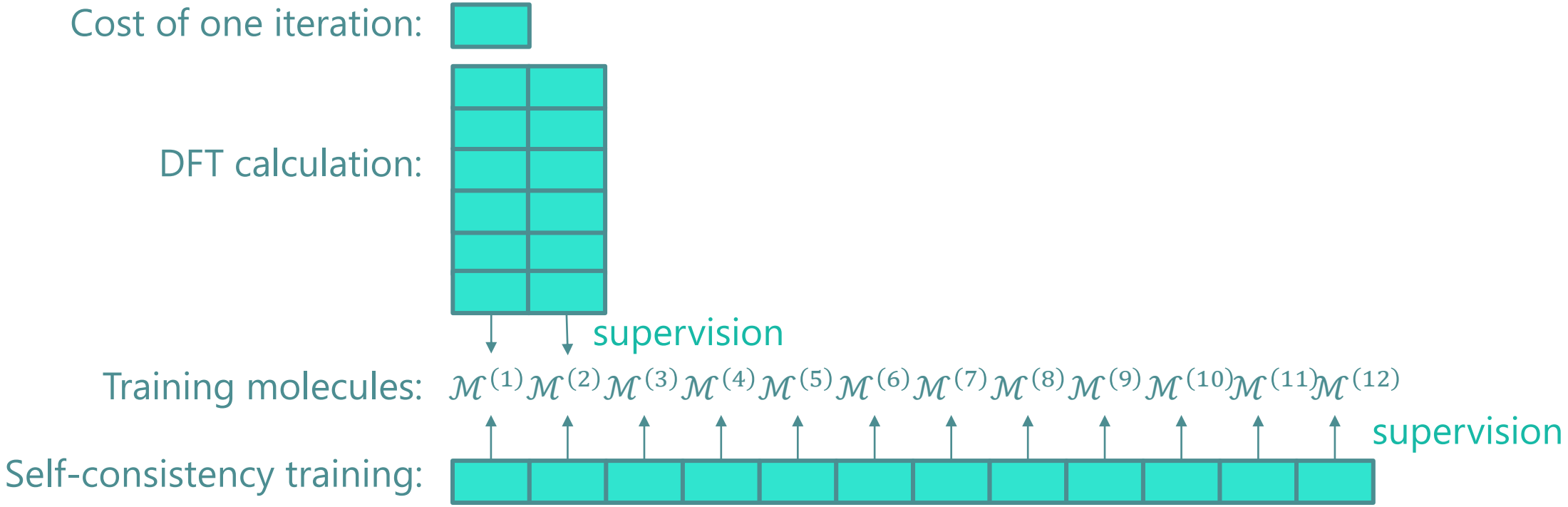


supervision

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

Unique Benefits

1. Generalization beyond labeled data: $\mathcal{L}_{\text{label}}\left(\theta; \overline{\mathcal{D}^{(1)}}\right) + \lambda\, \mathcal{L}_{\text{self-con}}\left(\theta; \mathcal{D}^{(2)}\right)$.
   limited labeled dataset; unlimited unlabeled dataset
2. Amortization effect: efficiency over DFT labeling.



Generalization beyond Labeled Data

- Data-scarce scenario (MD17): 100 labeled + 24900 unlabeled → test.

		Direct prediction	Derived molecular properties					As DFT init.
Molecule	Setting	**H** $[\mu E_\mathrm{h}] \downarrow$	ϵ $[\mu E_\mathrm{h}] \downarrow$	**C** $[\%] \uparrow$	ϵ_HOMO $[\mu E_\mathrm{h}] \downarrow$	ϵ_LUMO $[\mu E_\mathrm{h}] \downarrow$	ϵ_Δ $[\mu E_\mathrm{h}] \downarrow$	SCF Accel. $[\%] \downarrow$
Ethanol	label	160.36	712.54	99.44	911.64	6800.84	6643.11	68.3
	label + self-con	**75.65**	**285.49**	**99.94**	**336.97**	**1203.60**	**1224.86**	**61.5**
Malondi-aldehyde	label	101.19	456.75	99.09	471.92	1093.22	1115.94	69.1
	label + self-con	**86.60**	**280.39**	**99.67**	**274.45**	**279.14**	**324.37**	**62.1**
Uracil	label	88.26	1079.51	95.83	1217.17	12496.1	11850.56	65.8
	label + self-con	**63.82**	**315.40**	**99.58**	**359.98**	**369.67**	**388.30**	**54.5**

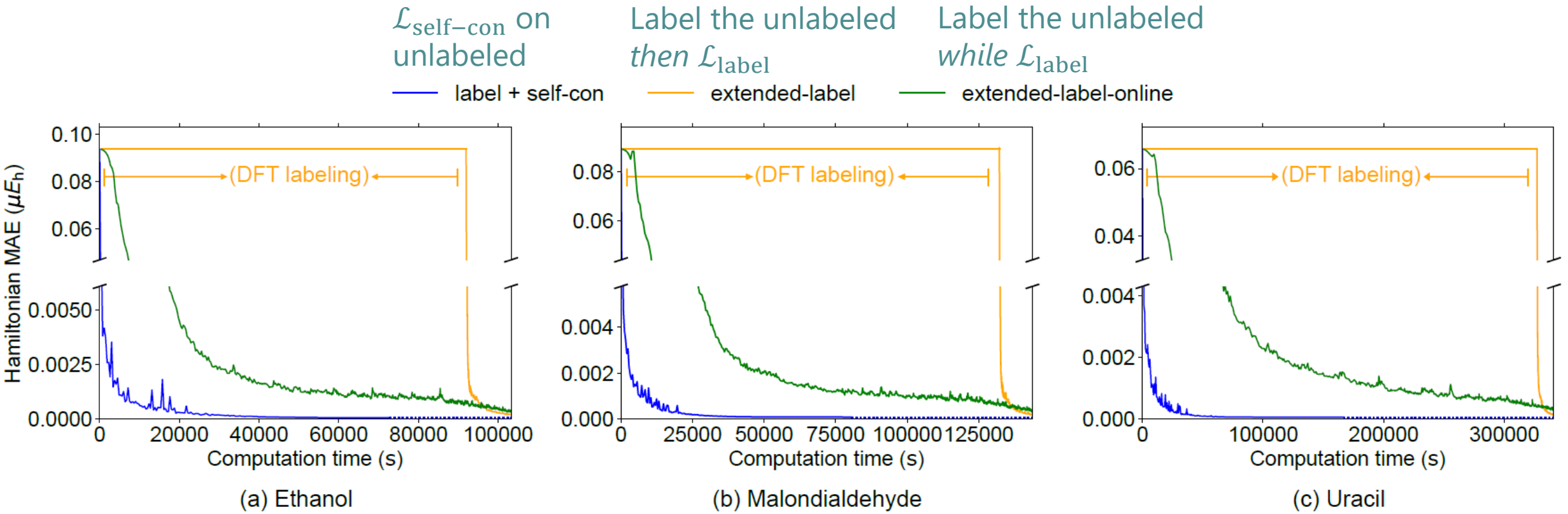
Generalization beyond Labeled Data

- Out-of-distribution (OOD) scenario (QH9):
  labeled small molecules + finetune on unlabeled large molecules → test on large molecules.

Setting	**H** $[\mu E_\mathrm{h}] \downarrow$	ϵ $[\mu E_\mathrm{h}] \downarrow$	**C** $[\%] \uparrow$	ϵ_HOMO $[\mu E_\mathrm{h}] \downarrow$	ϵ_LUMO $[\mu E_\mathrm{h}] \downarrow$	ϵ_Δ $[\mu E_\mathrm{h}] \downarrow$	SCF Accel. $[\%] \downarrow$
zero-shot	69.67	403.52	95.72	778.86	12230.49	12203.12	66.3
self-con (all-param)	65.74	375.31	**97.31**	565.50	**1130.55**	**1316.96**	**64.5**
self-con (adapter)	**64.48**	**268.83**	97.12	**449.80**	1220.54	1394.29	65.0

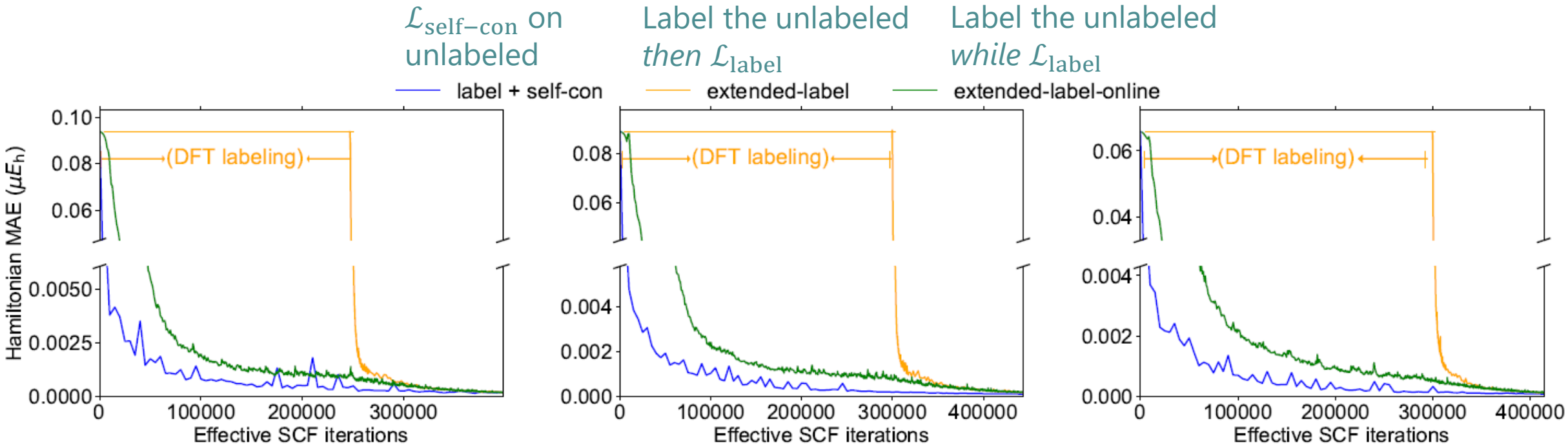
Training Efficiency by Amortization

- Data-scarce scenario (MD17): 100 labeled + 24900 unlabeled → test.



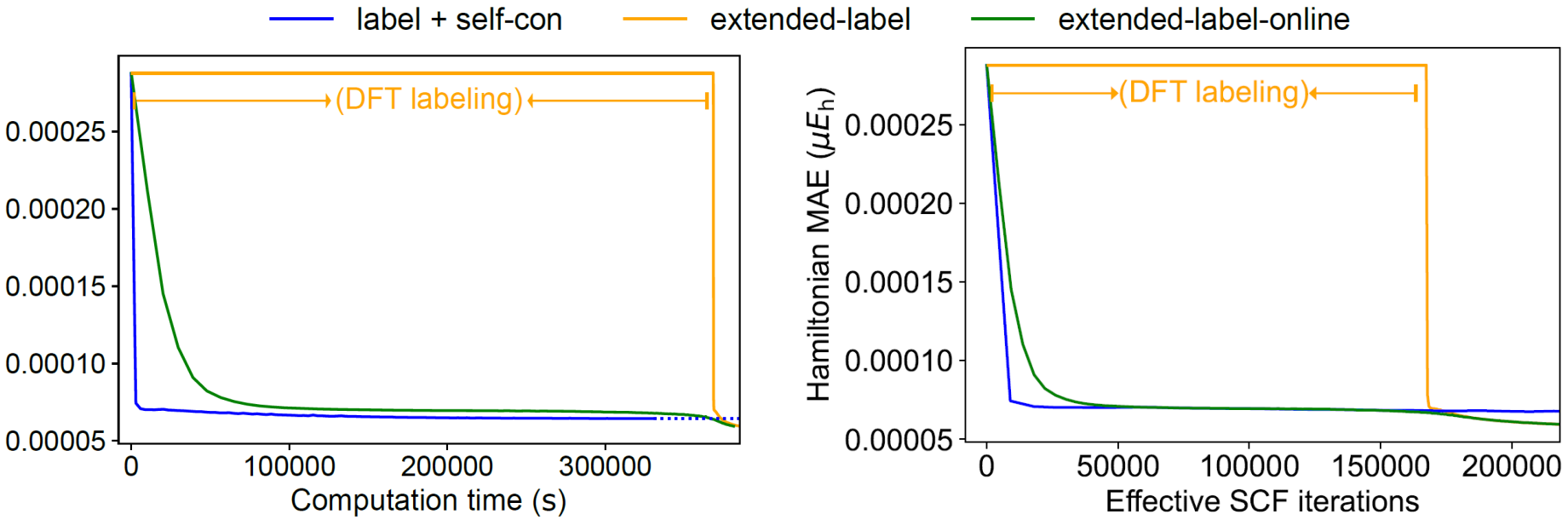
Training Efficiency by Amortization

- Data-scarce scenario (MD17): 100 labeled + 24900 unlabeled → test.



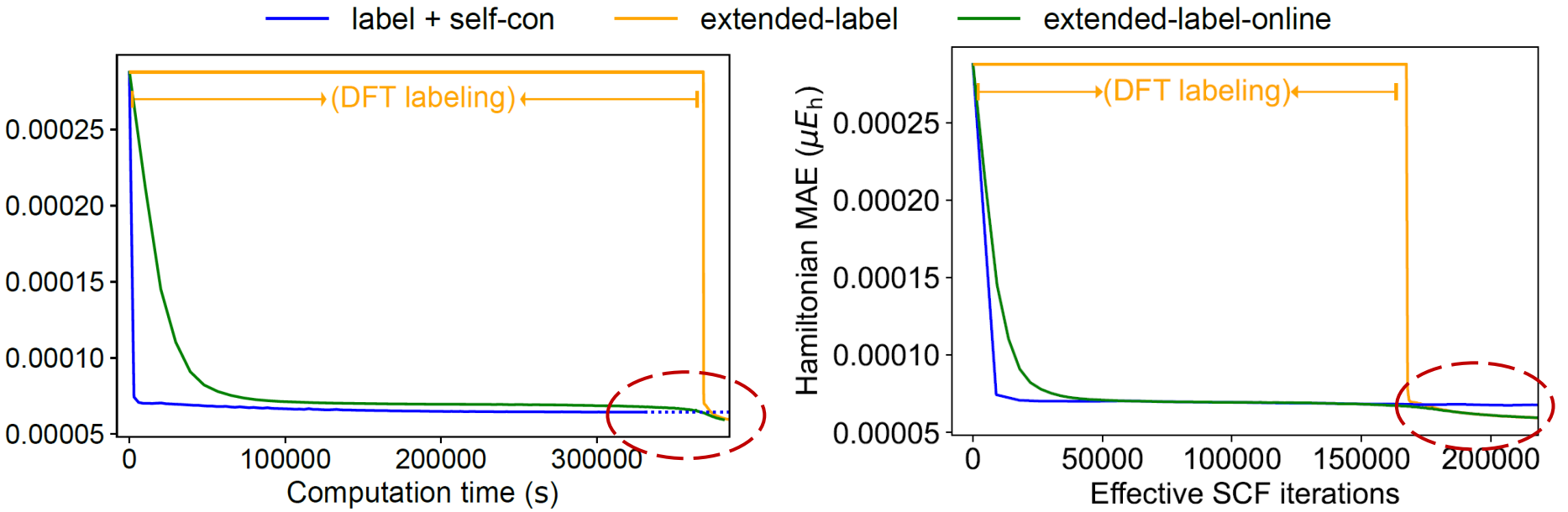
Training Efficiency by Amortization

- OOD scenario:
  labeled small molecules + finetune on unlabeled large molecules → test on large molecules.



Training Efficiency by Amortization

- OOD scenario:
  labeled small molecules + finetune on unlabeled large molecules → test on large molecules.



Final performance: self-consistency training gives better derived molecular properties!

Setting	**H** $[\mu E_h] \downarrow$	ϵ $[\mu E_h] \downarrow$	**C** [%] $\uparrow$	ϵ_{HOMO} $[\mu E_h] \downarrow$	ϵ_{LUMO} $[\mu E_h] \downarrow$	ϵ_Δ $[\mu E_h] \downarrow$	SCF Accel. [%] $\downarrow$
extended-label	**59.67**	330.05	96.63	541.92	6372.12	6445.33	65.2
self-con	64.48	**268.83**	**97.12**	**449.80**	**1220.54**	**1394.29**	**65.0**

Training Efficiency by Amortization

- Direct efficiency comparison with DFT:
  time for solving MD17 structures under the same stopping criteria.

Molecule	criterion $[\mu E_\mathrm{h}]$	$t_\text{self-con}$ [s]	t_DFT [s]
Ethanol	31.0	$\mathbf{4.50\times10^4}$	6.40×10^4
Malondialdehyde	88.9	$\mathbf{4.81\times10^4}$	1.05×10^5
Uracil	177.2	$\mathbf{1.23\times10^5}$	2.15×10^5

Extending Applicable Scale of Hamiltonian Prediction

- Labeled QH9 molecules (≤ 31 atoms) + Finetune on unlabeled larger molecules
  → Test on larger molecules (MD22).

Molecule	Setting	$\mathbf{H}$ $[\mu E_\mathrm{h}]\downarrow$	$\boldsymbol{\epsilon}$ $[\mu E_\mathrm{h}]\downarrow$	$\mathbf{C}$ $[\%]\uparrow$	ϵ_{HOMO} $[\mu E_\mathrm{h}]\downarrow$	ϵ_{LUMO} $[\mu E_\mathrm{h}]\downarrow$	ϵ_Δ $[\mu E_\mathrm{h}]\downarrow$	SCF Accel. $[\%]\downarrow$
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	84.6
	self-con	**52.49**	$\mathbf{1.22}\times10^3$	**94.46**	$\mathbf{2.07}\times10^3$	$\mathbf{3.76}\times10^3$	$\mathbf{2.69}\times10^3$	**64.7**
	e2e (ET)	N/A	N/A	N/A	1.74×10^5	7.72×10^3	2.38×10^5	N/A
	e2e (Equiformer)	N/A	N/A	N/A	2.38×10^5	1.16×10^4	2.27×10^5	N/A
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	170.8
	self-con	**56.12**	$\mathbf{1.81}\times10^3$	**83.51**	$\mathbf{1.99}\times10^3$	$\mathbf{4.01}\times10^3$	$\mathbf{2.34}\times10^3$	**67.0**
	e2e (ET)	N/A	N/A	N/A	2.92×10^5	2.58×10^4	3.39×10^5	N/A
	e2e (Equiformer)	N/A	N/A	N/A	3.76×10^5	2.31×10^4	4.17×10^5	N/A

Extending Applicable Scale of Hamiltonian Prediction

- Labeled QH9 molecules (≤ 31 atoms) + Finetune on unlabeled larger molecules → Test on larger molecules (MD22).
- Outperform end-to-end property predictors: merit of scientific-law supervision!

Molecule	Setting	$\mathbf{H}$ $[\mu E_\mathrm{h}]\downarrow$	$\boldsymbol{\epsilon}$ $[\mu E_\mathrm{h}]\downarrow$	$\mathbf{C}$ $[\%]\uparrow$	ϵ_{HOMO} $[\mu E_\mathrm{h}]\downarrow$	ϵ_{LUMO} $[\mu E_\mathrm{h}]\downarrow$	ϵ_Δ $[\mu E_\mathrm{h}]\downarrow$	SCF Accel. $[\%]\downarrow$
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	84.6
	self-con	**52.49**	$\mathbf{1.22}\times10^3$	**94.46**	$\mathbf{2.07}\times10^3$	$\mathbf{3.76}\times10^3$	$\mathbf{2.69}\times10^3$	**64.7**
	e2e (ET)	N/A	N/A	N/A	1.74×10^5	7.72×10^3	2.38×10^5	N/A
	e2e (Equiformer)	N/A	N/A	N/A	2.38×10^5	1.16×10^4	2.27×10^5	N/A
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	170.8
	self-con	**56.12**	$\mathbf{1.81}\times10^3$	**83.51**	$\mathbf{1.99}\times10^3$	$\mathbf{4.01}\times10^3$	$\mathbf{2.34}\times10^3$	**67.0**
	e2e (ET)	N/A	N/A	N/A	2.92×10^5	2.58×10^4	3.39×10^5	N/A
	e2e (Equiformer)	N/A	N/A	N/A	3.76×10^5	2.31×10^4	4.17×10^5	N/A

Microsoft

Thank You