GeoTMI: Predicting Quantum Chemical Property with Easy-to-Obtain Geometry via Positional Denoising

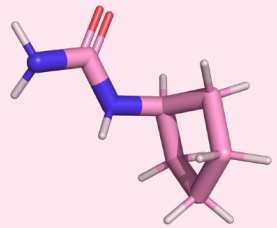
Hyeonsu Kim, Jeheon Woo, Seonghwan Kim, Seokhyun Moon, Jun Hyeong Kim, Woo Youn Kim*

Intelligent Chemistry Lab
Department of Chemistry, KAIST

Contents

Overview

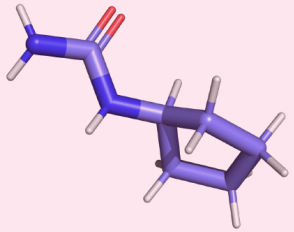
Quantum chemical calculations



$\tilde{X}$

Easy-to-obtain geometry

Geometry optimization



X

High-level geometry

Property calculation

Y

- HOMO-LUMO gap
- Energy
- Barrier height
- Relaxed energy
- ...

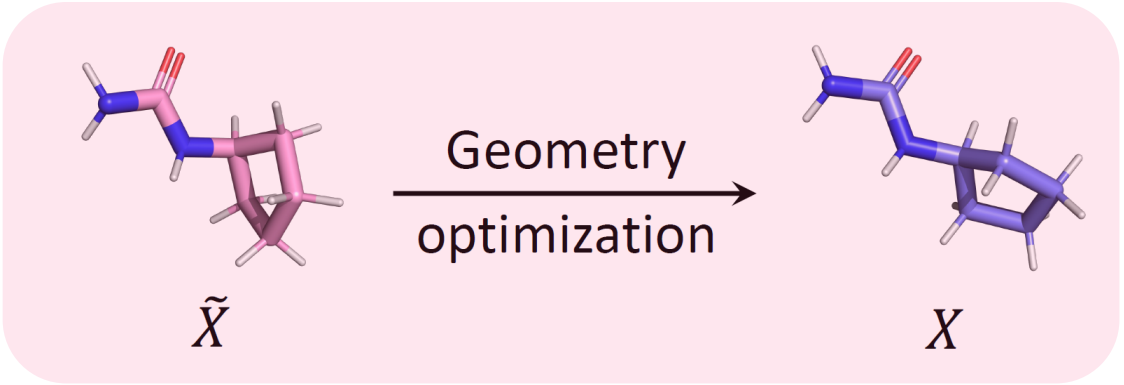
Quantum chemical properties

Schrödinger equation

$H\psi = E\psi$

Overview

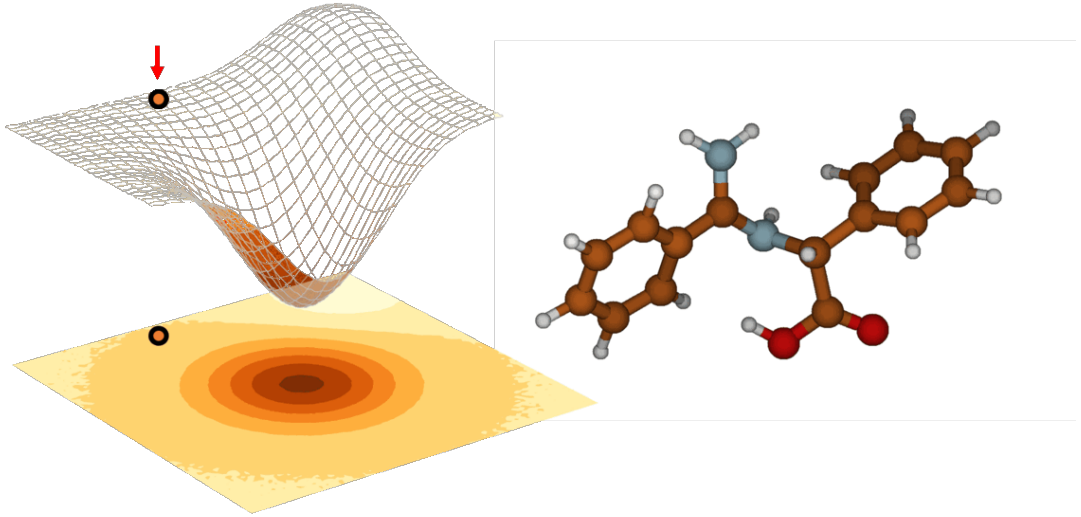
Geometry optimization to obtain X from $\tilde{X}$



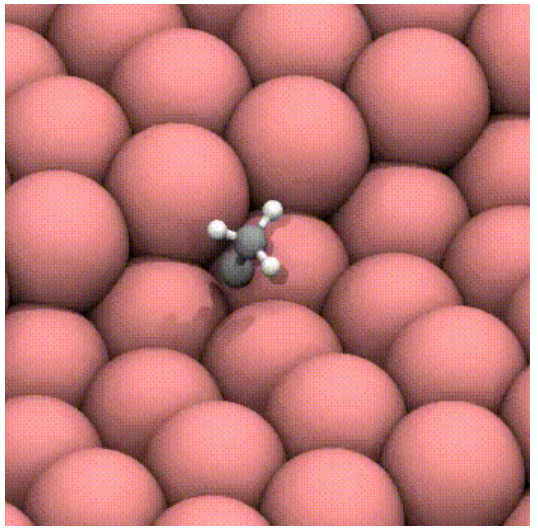
Property calculation → Y

- HOMO-LUMO gap
- Energy
- Barrier height
- Relaxed energy
- …

To obtain X, a geometry optimization process with a starting geometry $\tilde{X}$ should be preceded, which is also based on quantum calculation



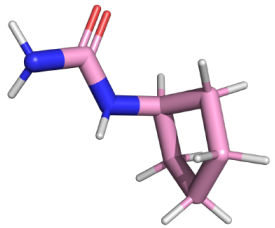
optimized to equilibrium molecular geometry



optimized to adsorption geometry

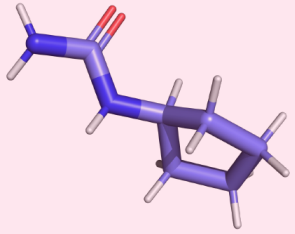
Overview

Calculation of quantum chemical properties (Y) from X



$\tilde{X}$

Geometry optimization



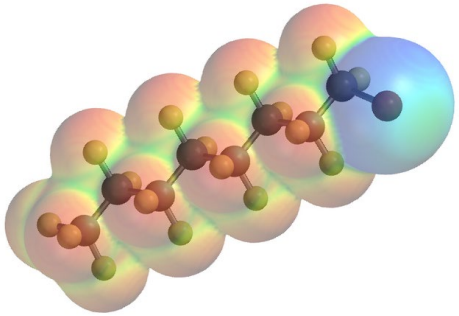
X

Property calculation

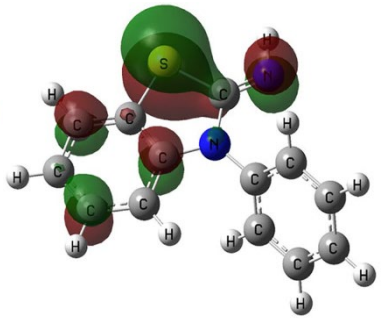
Y

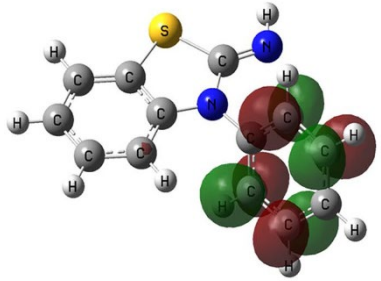
- HOMO-LUMO gap
- Energy
- Barrier height
- Relaxed energy
- ...

Molecular properties
(e.g., QM9, PCQM)



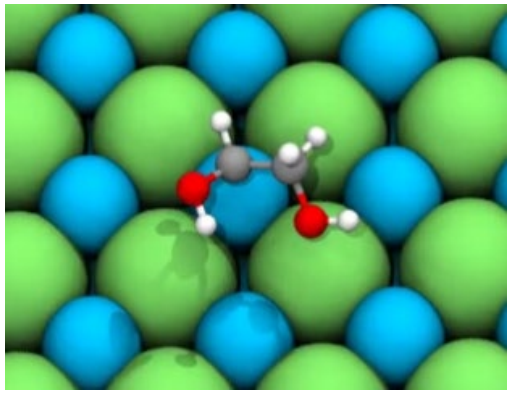
electron density



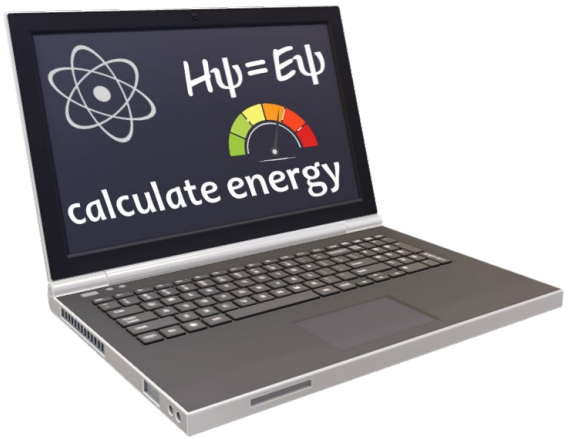


orbitals

Energies in surface catalyst systems
(e.g., OC20, OC22)

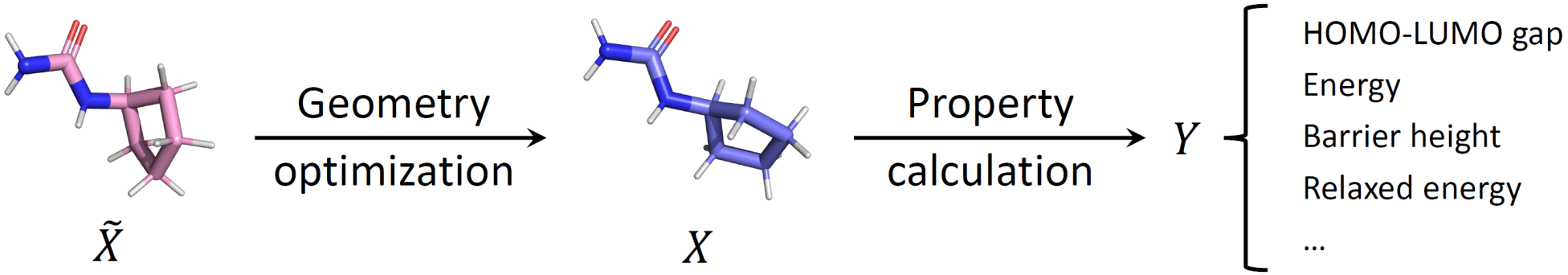


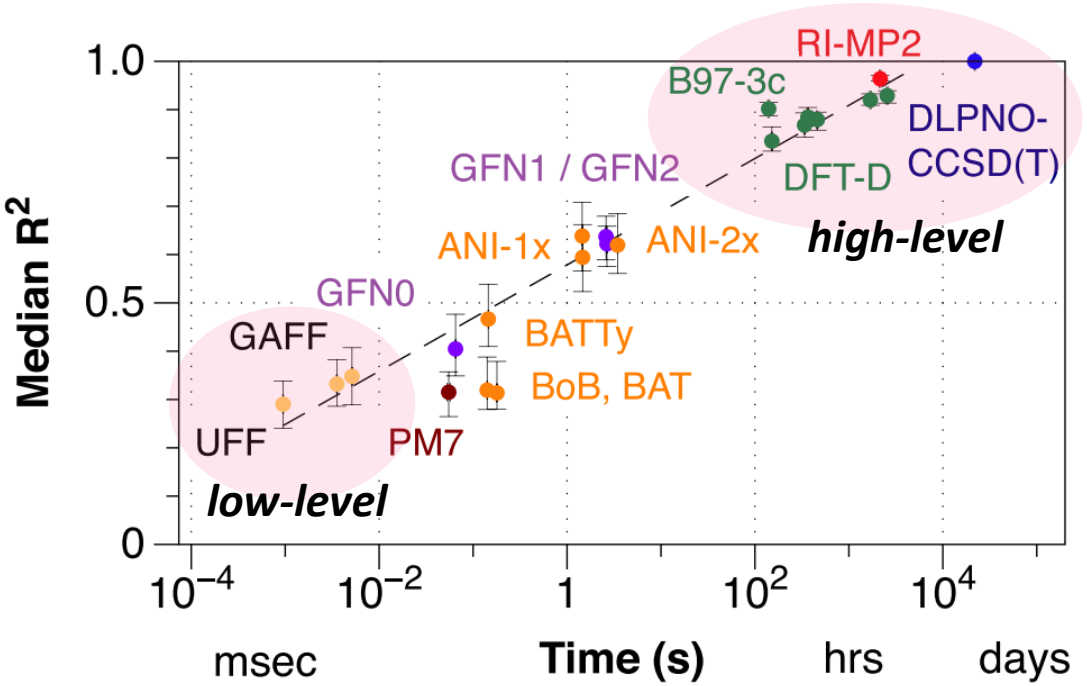
adsorption on surface



Overview

Geometry optimization to obtain X

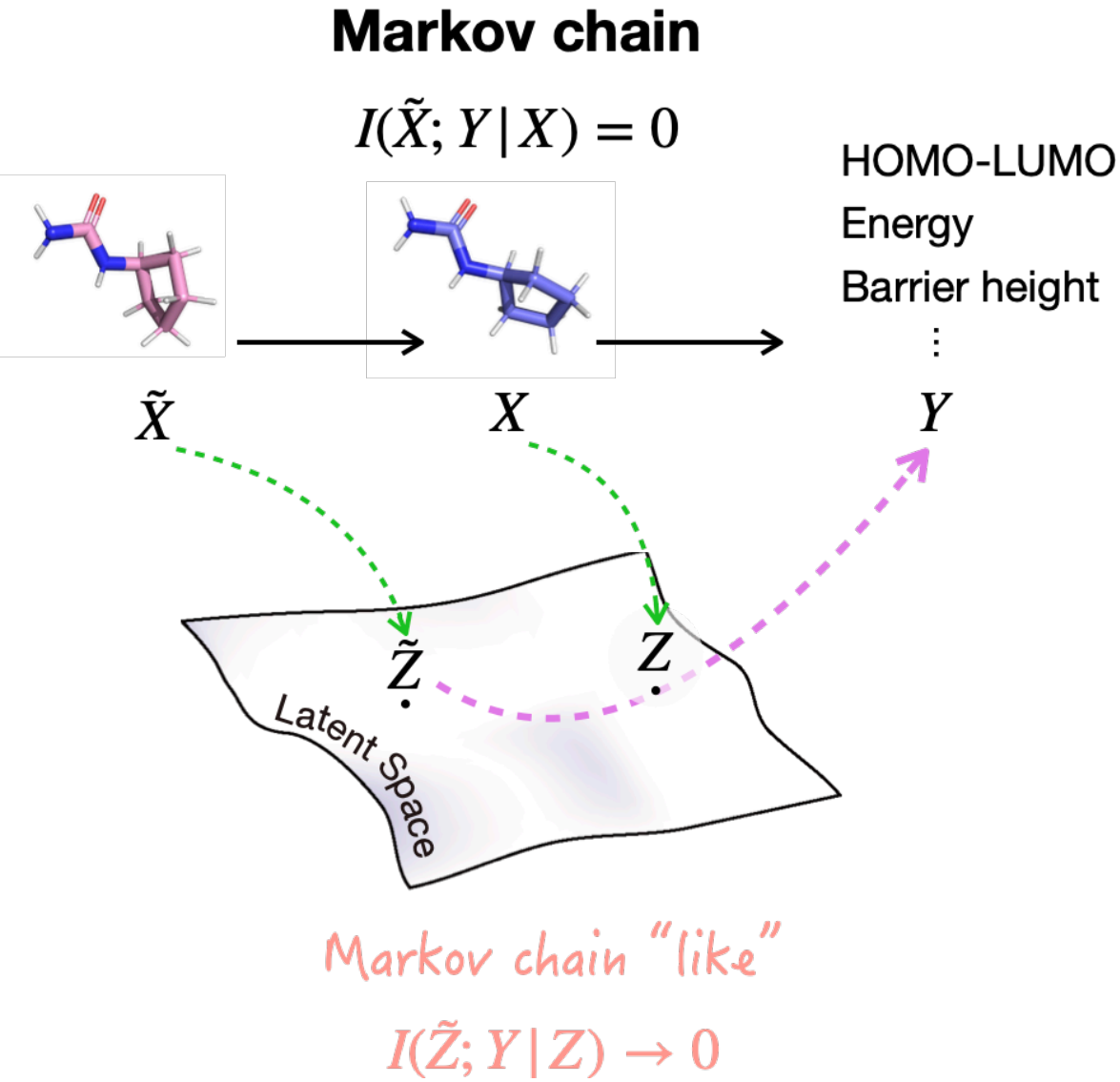




- ✓ Most 3D GNN studies have achieved great success in many quantum chemical property (Y) prediction tasks using X.
- ✓ The usage of X is **infeasible in real-world applications.**
- ✓ To tackle this, we propose a novel training framework for 3D GNNs, **GeoTMI**, which aims to predict high-level quantum properties from an easy-to-obtain low-level geometry ($\tilde{X}$).
- ✓ We also present a theoretical basis for fully exploiting such easy-to-obtain geometries to predict accurate target properties.

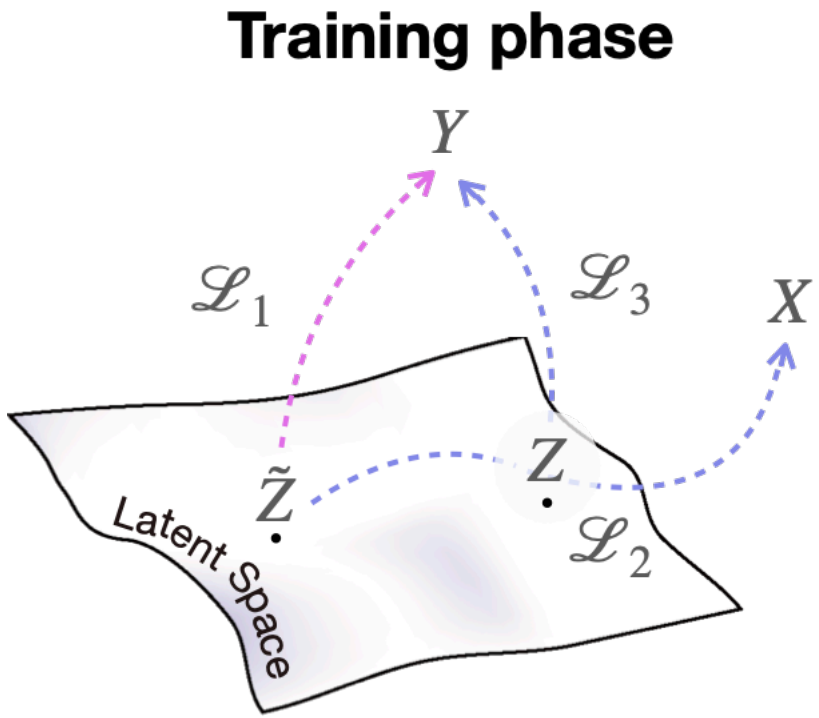
Folmsbee, D., & Hutchison, G. (2021). *International Journal of Quantum Chemistry*, *121*(1), e26381.

Introducing GeoTMI



- ✓ The relation between data $(\tilde{X}, X, Y)$ can be assumed as a **Markov chain**, and thus the conditional mutual information $I(\tilde{X}; Y|X)$ must be zero.
- ✓ The goal of GeoTMI is to obtain a proper representation $\tilde{Z}$ in predicting Y, by aligning it to Z with more enriching information for Y.
- ✓ It can be achieved by minimizing $I(\tilde{Z}; Y|Z)$, which is a physical inductive bias derived from the **Markov chain** relation.

Introducing GeoTMI



Intractable Maximize Minimize

$$I(\tilde{Z};Z;Y) = I(\tilde{Z};Y) - I(\tilde{Z};Y|Z)$$

$$\geq -(\mathscr{L}_1 + \mathscr{L}_2 + \mathscr{L}_3)$$

Tractable Surrogate LB

- ✓ By maximizing three-term mutual information $I(\tilde{Z};Z;Y)$, we can account for the **physical inductive bias** while **accurately predicting Y**.
- ✓ However, because the $I(\tilde{Z};Z;Y)$ is intractable, we adopted a tractable lower bound (LB) of it.
- ✓ We propose a tractable surrogate LB:

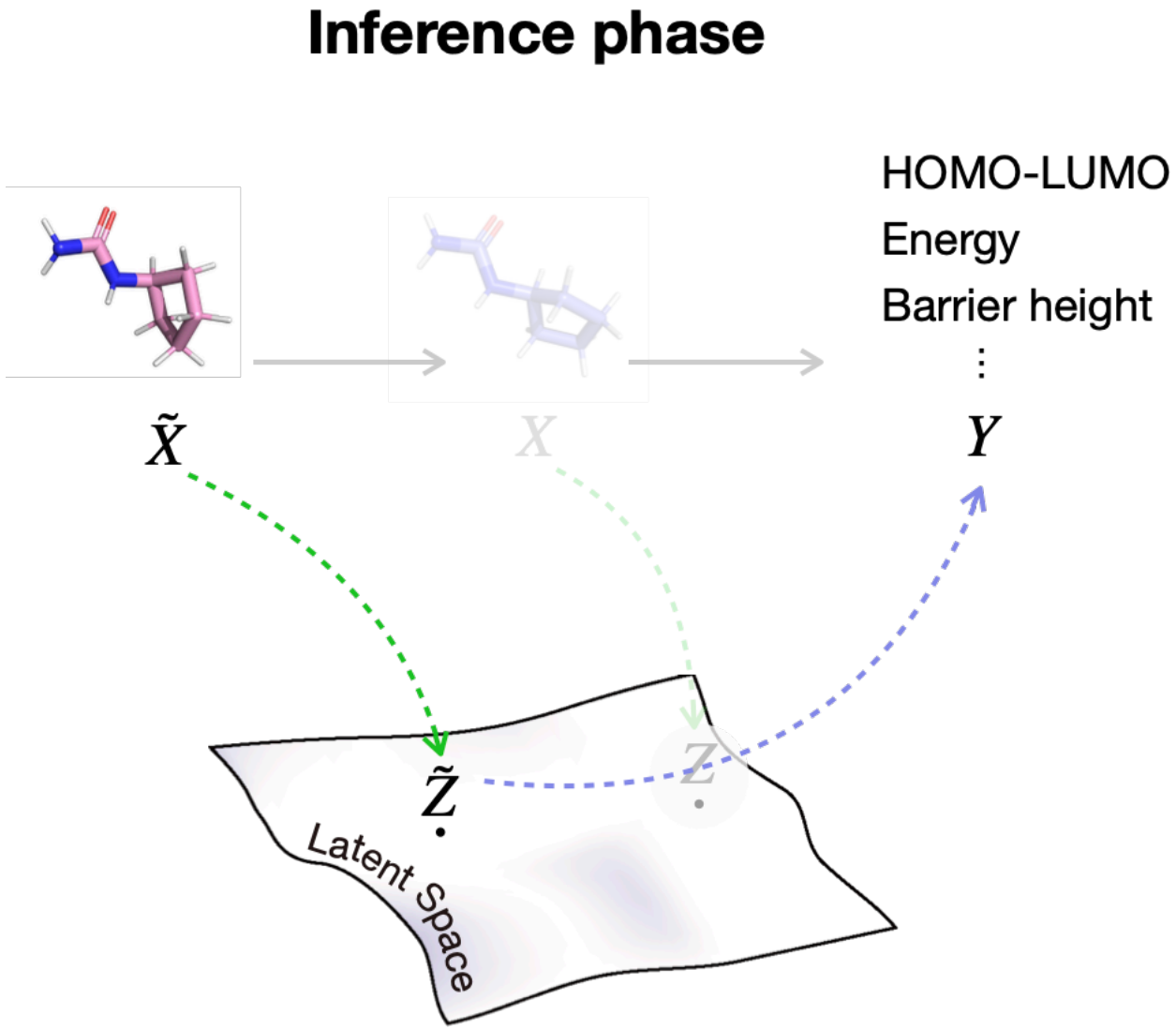
$$\mathscr{L}_1 = H(Y|\tilde{Z})$$

$$\mathscr{L}_2 = H(X|\tilde{Z})$$

$$\mathscr{L}_3 = H(Y|Z)$$

*We assumed the encoder is invertible.

Introducing GeoTMI



- ✓ While we utilized X in training phase, we circumvent it in inference phase.
- ✓ In summary, GeoTMI is a **model agnostic training framework**, which maximizes "**Three-term Mutual Information**" (TMI) using surrogate loss.

Experiments

- **Task 1. Molecular property prediction ($QM9_M$)**
- **Task 2. IS2RE prediction (OC20)**
- **Task 3. Reaction property prediction (Reaction barrier height)***

Task	$\tilde{\mathbf{x}}$	$\mathbf{x}$
$QM9_M$	MMFF-optimized geometries	DFT-optimized geometries
OC20	Initial structures (IS)	Relaxed structures (RS)
Reaction barrier height	Reactant and product geometries	Reactant and transition state geometries

✓ Demonstrate the effectiveness of the GeoTMI in various applicable tasks.

*See Appendix

Experiments

- **Molecular property prediction ($QM9_M$)**

Table 1: MAEs for QM9's properties. The best performance among the models that do not use X in the inference (Infer.) process is shown in bold. The values of Transformer-M using X were borrowed from Luo et al. [17]. The performance of GeoTMI integrated with SchNet and DimeNet++ is provided in Appendix B.3.

Methods	Input type (Train / Infer.)	U_0 (meV)	μ (D)	α (Bohr^3)	ϵ_{HOMO} (meV)	ϵ_{LUMO} (meV)	GAP (meV)	R^2 (Bohr^2)	C_v ($\frac{cal}{mol \cdot K}$)	ZPVE (meV)
Transformer-M [17]	X/X	14.8	-	-	26.5	23.8	-	-	-	-
EGNN	X/X	12.9	0.0350	0.0759	31.2	26.6	51.1	0.130	0.0336	1.59
Transformer-M	2D / 2D	38.2	0.309	0.171	53.6	52.5	77.1	11.4	0.0669	4.79
Transformer-M	2D, X / 2D	43.9	0.245	0.160	48.7	46.3	68.4	10.3	0.0683	3.85
EGNN	$\tilde{X}/\tilde{X}$	17.4	0.133	0.125	38.4	34.4	58.0	5.60	0.0445	1.97
EGNN + GeoTMI	$X, \tilde{X}/\tilde{X}$	**14.5**	**0.100**	**0.105**	**35.7**	**31.2**	**53.2**	**4.08**	**0.0407**	**1.76**
Improvements by GeoTMI (%)		16.7	24.8	16.0	7.03	9.30	8.28	27.1	8.54	10.7
Improvements in SchNet (%)		0.00	33.2	18.1	15.3	16.5	13.3	22.6	4.87	7.39
Improvements in DimeNet++ (%)		17.2	9.29	11.4	6.72	7.12	1.78	9.11	9.52	9.52

✓ GeoTMI achieved consistent performance improvements across all properties and models using $\widetilde{X}$, demonstrating effectiveness of GeoTMI.

Experiments

- **Molecular property prediction ($QM9_M$)**

Table 1: MAEs for QM9's properties. The best performance among the models that do not use X in the inference (Infer.) process is shown in bold. The values of Transformer-M using X were borrowed from Luo et al. [17]. The performance of GeoTMI integrated with SchNet and DimeNet++ is provided in Appendix B.3.

Methods	Input type (Train / Infer.)	U_0 (meV)	μ (D)	α (Bohr^3)	ϵ_{HOMO} (meV)	ϵ_{LUMO} (meV)	GAP (meV)	R^2 (Bohr^2)	C_v ($\frac{cal}{mol \cdot K}$)	ZPVE (meV)
Transformer-M [17]	X/X	14.8	-	-	26.5	23.8	-	-	-	-
EGNN	X/X	12.9	0.0350	0.0759	31.2	26.6	51.1	0.130	0.0336	1.59
Transformer-M	2D / 2D	38.2	0.309	0.171	53.6	52.5	77.1	11.4	0.0669	4.79
Transformer-M	2D, X / 2D	43.9	0.245	0.160	48.7	46.3	68.4	10.3	0.0683	3.85
EGNN	$\tilde{X}/\tilde{X}$	17.4	0.133	0.125	38.4	34.4	58.0	5.60	0.0445	1.97
EGNN + GeoTMI	$X, \tilde{X}/\tilde{X}$	**14.5**	**0.100**	**0.105**	**35.7**	**31.2**	**53.2**	**4.08**	**0.0407**	**1.76**
Improvements by GeoTMI (%)		16.7	24.8	16.0	7.03	9.30	8.28	27.1	8.54	10.7
Improvements in SchNet (%)		0.00	33.2	18.1	15.3	16.5	13.3	22.6	4.87	7.39
Improvements in DimeNet++ (%)		17.2	9.29	11.4	6.72	7.12	1.78	9.11	9.52	9.52

✓ While both the MMFF geometry and the molecular the 2D graph are easy-to-obtain inputs, the MMFF geometry contains more useful information for learning the relationship between molecules and their quantum chemical properties.

Experiments

- **IS2RE prediction (OC20)**

Table 3: Results on the OC20 IS2RE test set with different methods based on Equiformer architectures [16]. The Equiformer* denotes a model that reduces the number of transformer blocks from 18 to 4 while keeping other hyperparameters the same. The best performance among the Equiformer* models is shown in bold, and its improvement rate is shown in the last row.

	Energy MAE (eV) ↓					EwT (%) ↑				
Methods	ID	OOD Ads	OOD Cat	OOD Both	Average	ID	OOD Ads	OOD Cat	OOD Both	Average
Equiformer + Noisy Nodes [16]	0.417	0.548	0.425	0.474	0.466	7.71	3.70	7.15	4.07	5.66
Equiformer*	0.515	0.651	0.531	0.603	0.575	4.81	2.50	4.45	2.86	3.66
Equiformer* + Noisy Nodes	0.449	0.606	0.460	0.540	0.513	6.47	3.04	5.83	3.52	4.72
Equiformer* + GeoTMI	**0.425**	**0.583**	**0.440**	**0.521**	**0.492**	**7.60**	**3.86**	**6.97**	**4.03**	**5.62**
Improvement (%)	17.6	10.5	17.1	13.7	14.4	58.0	54.4	56.6	40.9	53.8

✓ Both Noisy Nodes and GeoTMI show performance improvements over the baseline Equiformer*, but GeoTMI achieves better performance gains across all metrics.

Conclusion

- ✓ **We propose GeoTMI, a model-agnostic training framework designed to exploit easy-to-obtain geometry for accurate prediction of quantum chemical properties.**

- ✓ **We envision that the GeoTMI becomes a new solution to solve the practical infeasibility of high-cost 3D geometry in many other chemistry fields.**

Appendix. Experiments

- **Reaction property prediction (Reaction barrier height)**

Table 2: MAEs for predicted reaction barrier (kcal/mol). The best performance among the models that do not use X in the inference (Infer.) process is shown in bold.

Methods	Input type (Train / Infer.)	Dataset	
		CCSD(T)-UNI	B97-D3
DimeRxn	X/X	2.38	1.92
D-MPNN	2D / 2D	4.59	4.91
DimeRxn	$\tilde{X}/\tilde{X}$	6.03	7.32
DimeRxn + GeoTMI	$X, \tilde{X}/\tilde{X}$	**3.90**	**4.17**
Improvements by GeoTMI (%)		35.3	43.0

✓ The usefulness of 3D easy-to-obtain geometry with the GeoTMI can be confirmed as in the previous two experiments

Appendix. Experiments

- **IS2RE prediction (OC20)**

Table 3: Results on the OC20 IS2RE test set with different methods based on Equiformer architectures [16]. The Equiformer* denotes a model that reduces the number of transformer blocks from 18 to 4 while keeping other hyperparameters the same. The best performance among the Equiformer* models is shown in bold, and its improvement rate is shown in the last row.

	Energy MAE (eV) ↓					EwT (%) ↑				
Methods	ID	OOD Ads	OOD Cat	OOD Both	Average	ID	OOD Ads	OOD Cat	OOD Both	Average
Equiformer + Noisy Nodes [16]	0.417	0.548	0.425	0.474	0.466	7.71	3.70	7.15	4.07	5.66
Equiformer*	0.515	0.651	0.531	0.603	0.575	4.81	2.50	4.45	2.86	3.66
Equiformer* + Noisy Nodes	0.449	0.606	0.460	0.540	0.513	6.47	3.04	5.83	3.52	4.72
Equiformer* + GeoTMI	**0.425**	**0.583**	**0.440**	**0.521**	**0.492**	**7.60**	**3.86**	**6.97**	**4.03**	**5.62**
Improvement (%)	17.6	10.5	17.1	13.7	14.4	58.0	54.4	56.6	40.9	53.8

✓ The prediction performance is similar to the original model for EwT and even better for OOD Ads.