August 3–7, 2025
KDD2025

# Electron-Informed Coarse-Graining Molecular Representation Learning for Real-World Molecular Physics

: Decomposition-Based Molecular Representation Learning for Predicting Molecular Properties in Real-World Chemical Experiments[1,2]

Gyoung S. Na[a] and **Chanyoung Park**[b]
[a]Korea Research Institute of Chemical Technology (KRICT)
[b]Korea Advanced Institute of Science and Technology (KAIST)
*ngs0@krict.re.kr, cy.park@kaist.ac.kr*

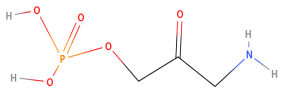
**Paper and source code:**
[1] *https://doi.org/10.1145/3690624.3709270*
[2] *https://github.com/ngs00/hedmol*

KRICT KAIST

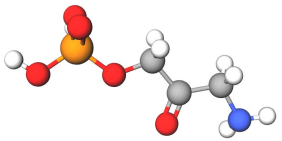
# Three Common Molecular Descriptors for Machine Learning

Characteristics of Three Essential Molecular Descriptors in Electron- and Atom-Level Information

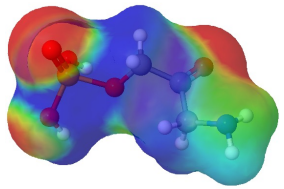
**| Commonly Used Molecular Descriptors used in Machine Learning for Chemistry**



Atom-level 2D molecular Structure (e.g., SMILES)



Atom-level 3D molecular structure



Electron-level molecular descriptor

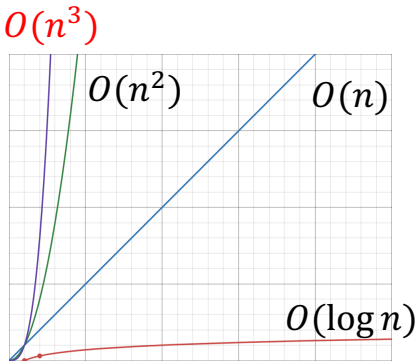
Measurement cost

Low (Least Informative)

High (Most informative)

Electron-level information is an essential feature describing the nature of molecules.

However, acquiring electron-level information is expensive and sometimes infeasible in real-world chemical applications.




Time complexity of standard quantum mechanical calculation methods

# Limitations of Existing Methods

Basic Assumptions on Molecular Representation Learning and Its Limitations

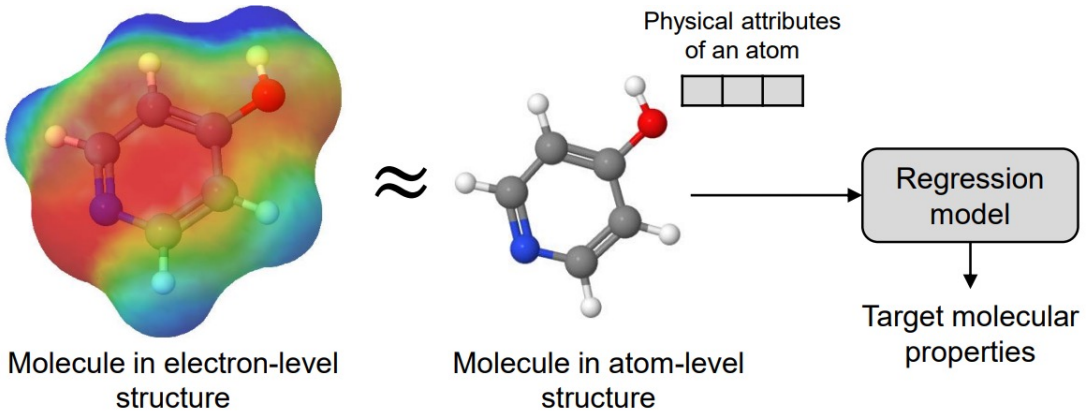
## Pros and Cons of Existing Methods

1. **(Atom-level)** 2D GNNs with 2D atom-level molecular structure [1, 2]
   - Pros: Low cost, Practicality, Generality
   - Cons: Low accuracy, Interpretability

2. **(Atom-level)** 3D GNNs with public quantum mechanical datasets [3, 4, 5]
   - Pros: High accuracy
   - Cons: High cost

3. **(Electron-level)** Domain-specific neural networks with quantum mechanical calculations [6, 7]
   - Pros: High accuracy, Interpretability
   - Cons: High cost, Generality

## Basic Assumptions on Atom-level Methods (1 & 2)

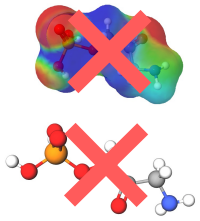



- **Assumption:** Atom-level molecular structure is sufficient to embed the electron-level information
- However, electrons have uncertainty → Describing electron-level information using only atom-level descriptors is not accurate and chemically valid

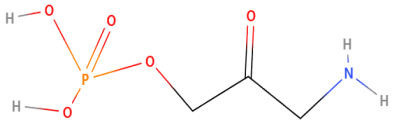
# Molecular Representation Learning for Real-World Applications

Molecular Representation Learning Methods for Complex and Large-Scale Molecules in Real-World Chemical Applications

- **Our work:** Molecular representation learning method that can estimate electron-level information from atom-level 2D molecular structures *without expensive quantum mechanical calculations*



No quantum mechanical calculations



2D molecular structure (SMILES)

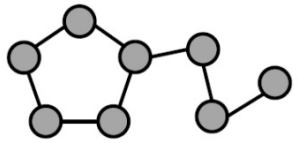
Machine Learning Model

Target molecular property (e.g., Lipophilicity, Aqueous solubility, Toxicity)

- We reformulated the molecular property prediction problem based on decomposed substructures.

$y = g(G)$

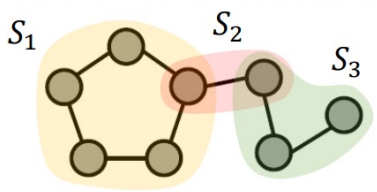
Original problem



Problem reformulation

$y = (g \circ h)(\varepsilon_1, \varepsilon_2 \ldots, \varepsilon_K)$

Reformulated problem

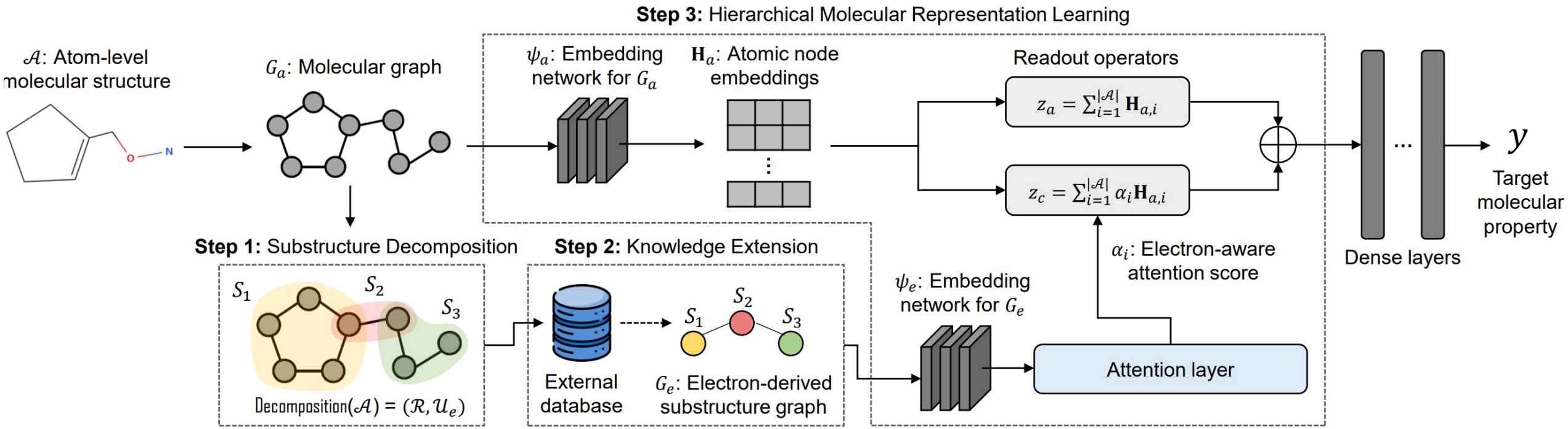



- $G$ : Atom-level 2D molecular structure (e.g., SMILES)
- $\varepsilon_k$: Electron-level information of the $k$-th substructure
- $h$ : Learnable embedding function for $\varepsilon_k$

# HEDMoL: Hierarchical Electron-Derived Molecular Learning

Electron-Informed Coarse-Graining Molecular Representation Learning for Predicting Experimentally-Measured Molecular Properties

## The Overall Representation Learning and Prediction Processes of HEDMoL

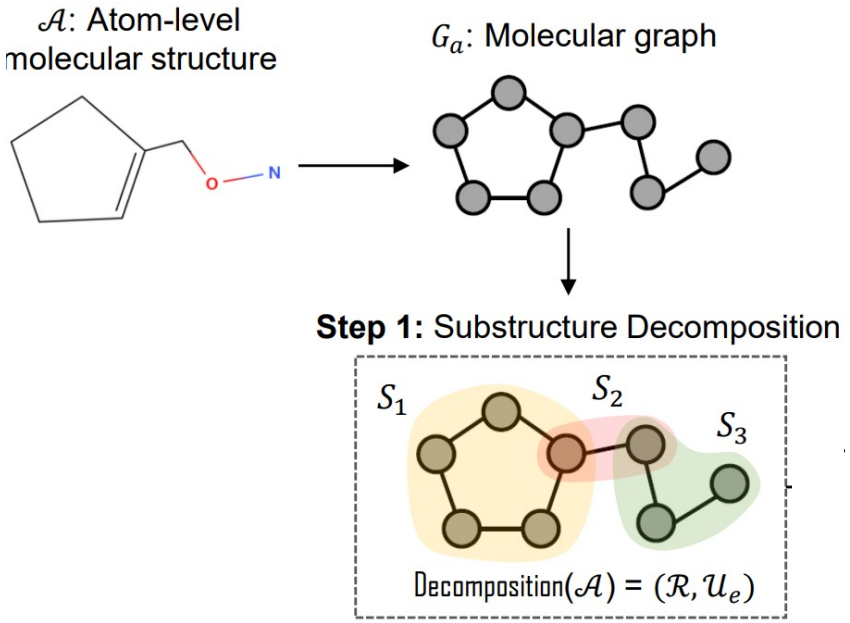



1. The input atom-level 2D molecular structure is decomposed into atom-level substructures $S_1, S_2, \ldots, S_K$.
2. Electron-level information in readily accessible public databases (e.g., QM9) [8] is assigned to $S_1, S_2, \ldots, S_K$.
3. Joint representation learning is conducted on the input molecular graph $G_a$ and electron-informed subgraphs $G_1, G_2, \ldots, G_K$.

# HEDMoL: Hierarchical Electron-Derived Molecular Learning

Electron-Informed Coarse-Graining Molecular Representation Learning for Predicting Experimentally-Measured Molecular Properties

## The Overall Representation Learning and Prediction Processes of HEDMoL

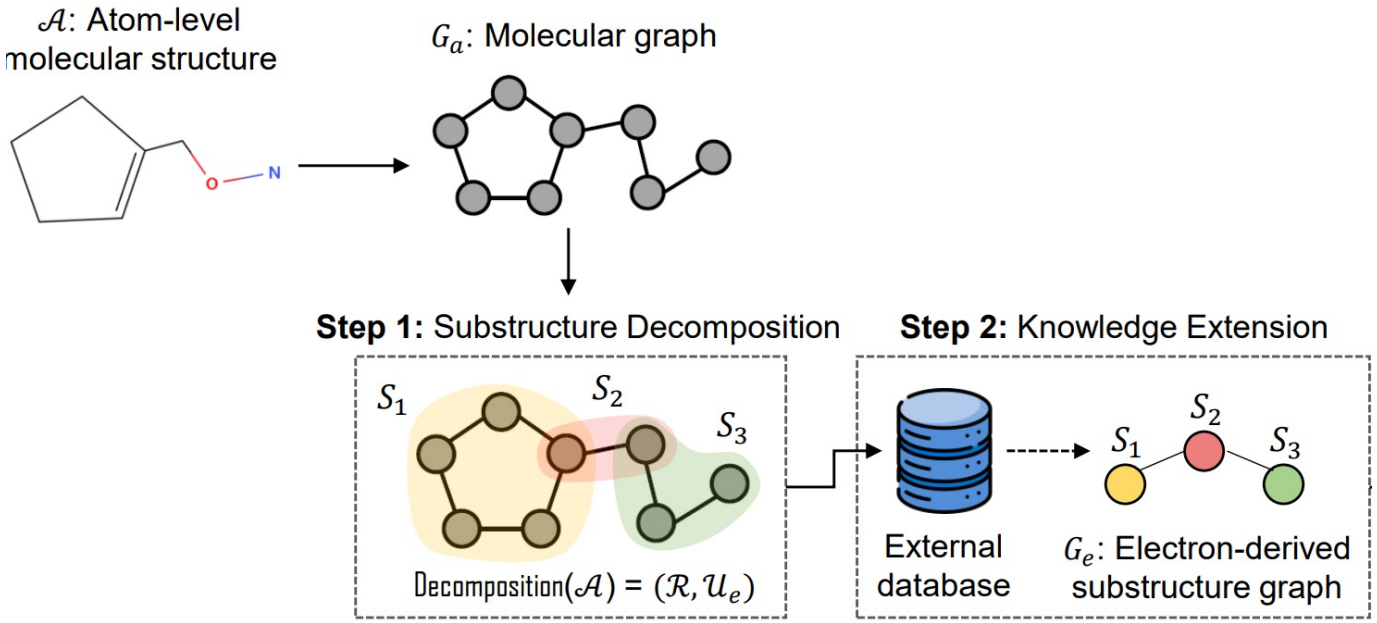



1 The input atom-level 2D molecular structure is decomposed into atom-level substructures $S_1, S_2, \ldots, S_K$.

# HEDMoL: Hierarchical Electron-Derived Molecular Learning

Electron-Informed Coarse-Graining Molecular Representation Learning for Predicting Experimentally-Measured Molecular Properties

## The Overall Representation Learning and Prediction Processes of HEDMoL

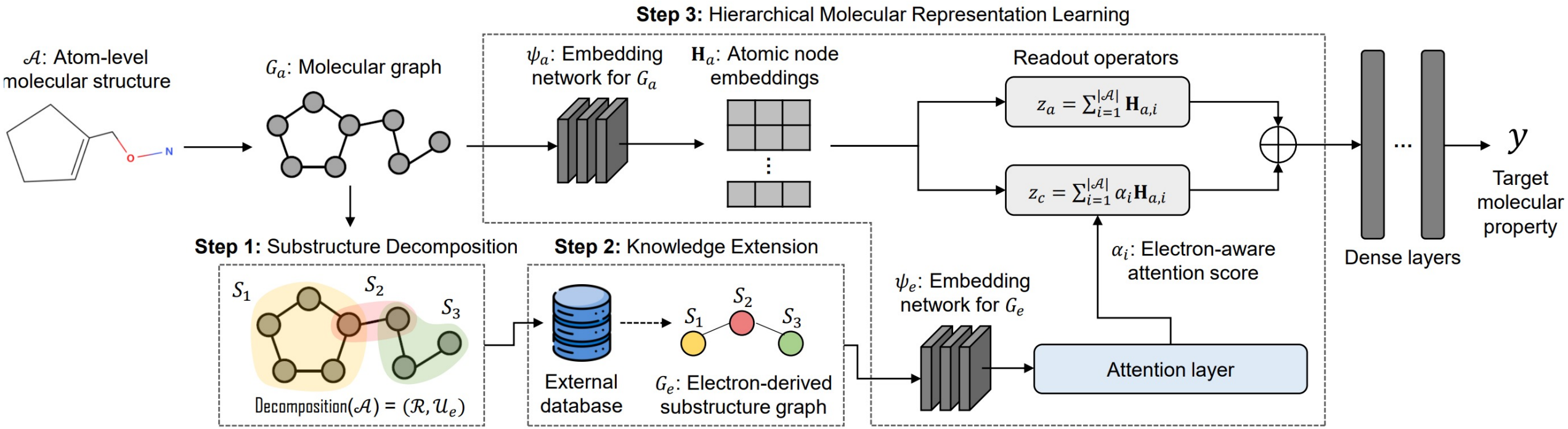



1. The input atom-level 2D molecular structure is decomposed into atom-level substructures $S_1, S_2, \ldots, S_K$.
2. Electron-level information in readily accessible public databases (e.g., QM9) [8] is assigned to $S_1, S_2, \ldots, S_K$.

# HEDMoL: Hierarchical Electron-Derived Molecular Learning

Electron-Informed Coarse-Graining Molecular Representation Learning for Predicting Experimentally-Measured Molecular Properties

## The Overall Representation Learning and Prediction Processes of HEDMoL




1. The input atom-level 2D molecular structure is decomposed into atom-level substructures $S_1, S_2, \ldots, S_K$.
2. Electron-level information in readily accessible public databases (e.g., QM9) [8] is assigned to $S_1, S_2, \ldots, S_K$.
3. Joint representation learning is conducted on the input molecular graph $G_a$ and electron-informed subgraphs $G_1, G_2, \ldots, G_K$.

# Energy-Based Physical Consistency Regularization

Training Objective of HEDMoL With Energy-Based Physical Consistency Regularization for Robust Molecular Representation Learning

## Energy-Based Physical Consistency

$$E_{p,k} + \epsilon_k = \hat{E}_{a,k} = \hat{E}_{e,k}, \forall k = 1, 2, ..., |\mathcal{R}|$$

$E_{p,k}$: Ground-truth physical energy of $S_k$ with uncertainty $\epsilon_k$ (Retrieved from DB)

$\hat{E}_{a,k}$: Predicted energy from atom-level graph embedding of $S_k$

$\hat{E}_{e,k}$: Predicted energy from electron-informed embedding of $S_k$

- This constraint enforces the graph embeddings of $G_a$ and $G_e$ to represent the same potential energy, which is essential information in describing molecules.

## Loss Function of the Training Process

$$L = \underbrace{\sum_{n=1}^{|\mathcal{D}|} L_p(y_n, f_d(\mathbf{z}_n))}_{\text{Prediction error}} + \lambda \underbrace{\left( \sum_{n=1}^{|\mathcal{D}|} \Omega_{a,n} + \Omega_{e,n} \right)}_{\text{Regularization terms}}$$

$$\Omega_{a,n} = \sum_{k=1}^{|\mathcal{R}_n|} \max\{|E_{p,k} - f_e(g_e(S_k))| - \alpha, 0\}$$

$$\Omega_{e,n} = \sum_{k=1}^{|\mathcal{R}_n|} \max\{|E_{p,k} - f_e(\mathbf{H}_{e,k})| - \alpha, 0\}$$

$\hat{E}_{a,k}$ $\hat{E}_{e,k}$

$f_e$: Learnable energy function

$g_e$: GNN encoder

$g_e(S_k)$: Atom-level graph embedding of $S_k$

$\mathbf{H}_{e,k}$: Electron-informed embedding of $S_k$

$\alpha$: Margin

# Experiment Settings

Experimental Evaluations on Real-World Benchmark Molecular Datasets

## Benchmark Molecular Datasets for Experimental Evaluations

- We conduct experimental evaluations on eight benchmark molecular datasets.
- We focus on experimentally-collected data in real-world chemical experiments.

| Application Category | Dataset | Target Molecular Property | # of Instances |
|---|---|---|---|
| Physicochemistry | Lipop [32] | Lipophilicity | 4,200 |
| | ESOL [12] | Aqueous solubility | 1,128 |
| | ADMET [61] | Aqueous solubility | 4,801 |
| Toxicity | IGC50 [59] | Tetrahymena pyriformis toxicity | 1,791 |
| | LC50 [59] | Fathead minnow toxicity | 822 |
| | LD50 [59] | Oral rat toxicity | 7,412 |
| Pharmacokinetics | LMC-H [32] | Microsomal clearance in human | 5,347 |
| | LMC-R [32] | Microsomal clearance in rat | 2,165 |

## Competitor Methods in Three Different Approaches

1. Tree-based methods: XGB-Mor, XGB-FC, XGB-MK[1]
2. 3D GNNs: SchNet, DimeNet, PhysChem, M3GNet, FAENet[2]
3. 2D GNNs: GATv2, GIN, EGC, MPNN, UniMP, FiLM, MEGNet, DMPNN, AttFP

[1]We used Morgan, functional-class, and MACCS Key fingerprints.
[2]We employed force-field-based semi-empirical method to generate 3D structures in feasible time.

## Implementation Notes

- GNN for $G_a$: Efficient graph convolution (EGC) [9] or SchNet [10]
- GNN for $G_e$: Graph isomorphism network (GIN) [11]
- Graph decomposition: Junction tree algorithm [12]

| Dataset | $\psi_a$ | $\psi_e$ |
|---|---|---|
| Lipop | EGC | GIN |
| ESOL | EGC | GIN |
| ADMET | EGC | GIN |
| IGC50 | EGC | GIN |
| LC50 | SchNet | GIN |
| LD50 | SchNet | GIN |
| LMC-H | SchNet | GIN |
| LMC-R | SchNet | GIN |

Source code: https://github.com/ngs00/hedmol



Gyoung S. Na and Chanyoung Park, ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2025.

# Prediction Accuracy on Experimental Molecular Datasets

Characteristics of Three Essential Molecular Descriptors in Electron- and Atom-Level Information

- HEDMoL achieves state-of-the-art accuracy in predicting molecular properties on the eight benchmark molecular datasets (Metric: $R^2$).
- Although AttFP was designed for drug-like molecules, HEDMoL outperforms AttFP in the physicochemistry datasets (Lipop, ESOL, ADMET).
- Accuracy improvements are remarkable on the LC50, LD50, LMC-H, and LMC-R datasets (Large molecules).

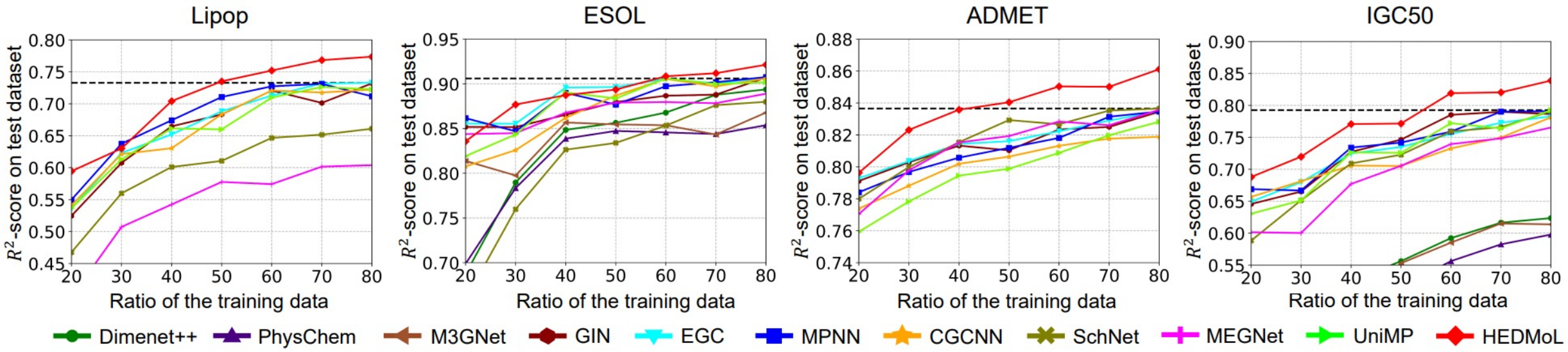
| Input Type | Method | Lipop | ESOL | ADMET | IGC50 | LC50 | LD50 | LMC-H | LMC-R |
|---|---|---|---|---|---|---|---|---|---|
| Molecular Fingerprint | XGB-Mor [42] | 0.531 (0.024) | 0.659 (0.045) | 0.717 (0.021) | 0.621 (0.040) | 0.390 (0.133) | 0.497 (0.016) | 0.505 (0.018) | **0.617 (0.058)** |
| | XGB-FC [42] | 0.578 (0.018) | 0.686 (0.052) | 0.720 (0.009) | 0.628 (0.023) | 0.501 (0.052) | 0.519 (0.025) | 0.503 (0.007) | **0.612 (0.015)** |
| | XGB-MK [49] | 0.542 (0.041) | 0.764 (0.047) | 0.761 (0.020) | 0.680 (0.037) | 0.486 (0.112) | 0.526 (0.021) | 0.471 (0.019) | 0.591 (0.033) |
| 3D Graph | SchNet [45] | 0.667 (0.021) | 0.881 (0.026) | 0.834 (0.012) | 0.765 (0.034) | 0.467 (0.025) | 0.527 (0.062) | 0.456 (0.024) | 0.573 (0.043) |
| | DimeNet [17] | N/R | 0.878 (0.025) | N/R | 0.779 (0.019) | N/A | 0.541 (0.045) | 0.352 (0.101) | N/A |
| | PhysChem [67] | 0.694 (0.024) | 0.848 (0.032) | N/A | 0.814 (0.017) | N/A | 0.511 (0.053) | N/A | N/A |
| | M3GNet [6] | N/A | 0.857 (0.025) | N/A | 0.697 (0.029) | N/A | 0.531 (0.034) | N/A | N/A |
| | FAENet [14] | 0.670 (0.036) | 0.869 (0.013) | 0.788 (0.020) | 0.708 (0.015) | 0.528 (0.094) | 0.474 (0.020) | 0.437 (0.025) | 0.528 (0.035) |
| 2D Graph | GATv2 [3] | 0.677 (0.053) | 0.891 (0.020) | 0.828 (0.014) | 0.795 (0.013) | 0.502 (0.063) | 0.498 (0.030) | 0.424 (0.027) | 0.560 (0.036) |
| | GIN [64] | 0.702 (0.031) | 0.897 (0.022) | 0.833 (0.017) | 0.799 (0.021) | 0.543 (0.080) | 0.515 (0.044) | 0.443 (0.027) | 0.568 (0.020) |
| | EGC [52] | 0.708 (0.043) | 0.896 (0.017) | 0.838 (0.012) | 0.808 (0.029) | 0.575 (0.045) | 0.497 (0.034) | 0.441 (0.023) | 0.566 (0.017) |
| | MPNN [19] | 0.711 (0.022) | 0.894 (0.023) | 0.830 (0.014) | 0.797 (0.018) | 0.532 (0.064) | 0.469 (0.040) | 0.449 (0.057) | 0.564 (0.031) |
| | UniMP [47] | 0.702 (0.030) | 0.886 (0.025) | 0.833 (0.014) | 0.793 (0.027) | 0.504 (0.031) | 0.470 (0.025) | 0.422 (0.061) | 0.579 (0.036) |
| | FiLM [2] | 0.703 (0.048) | 0.894 (0.031) | 0.836 (0.014) | 0.783 (0.046) | 0.526 (0.042) | 0.475 (0.032) | 0.421 (0.050) | 0.568 (0.032) |
| | MEGNet [7] | 0.604 (0.023) | 0.889 (0.027) | 0.826 (0.038) | 0.754 (0.026) | 0.574 (0.122) | 0.505 (0.027) | 0.422 (0.032) | 0.607 (0.041) |
| | DMPNN [66] | 0.716 (0.037) | 0.879 (0.013) | 0.820 (0.018) | 0.787 (0.008) | 0.566 (0.098) | 0.521 (0.011) | 0.494 (0.011) | 0.605 (0.043) |
| | AttFP [62] | 0.710 (0.021) | **0.909 (0.018)** | 0.841 (0.017) | 0.807 (0.013) | **0.642 (0.079)** | 0.513 (0.016) | 0.456 (0.031) | 0.588 (0.032) |
| | HEDMoL | **0.759 (0.043)** | **0.914 (0.016)** | **0.865 (0.014)** | **0.840 (0.010)** | **0.663 (0.053)** | **0.572 (0.035)** | **0.551 (0.008)** | **0.639 (0.035)** |

Large molecules

# Prediction Accuracy on Small Training Datasets

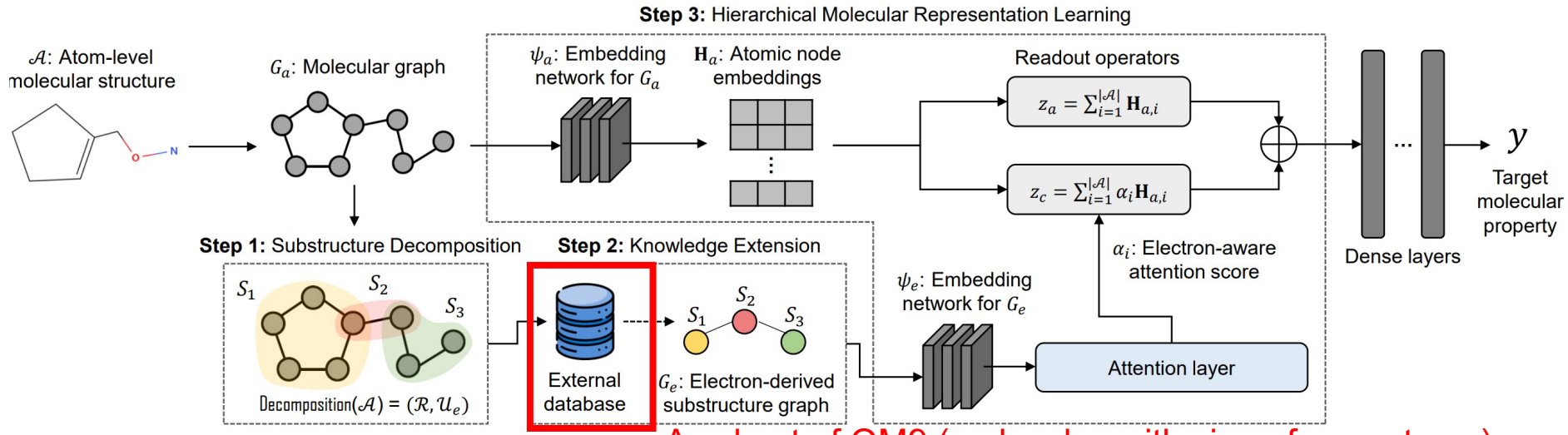
Prediction Accuracy of HEDMoL for Different Volumes of Training Datasets

- Lack of training data is one of the main challenges of machine learning in chemical applications [13].
- We measure the prediction accuracy for different training data sizes on the benchmark molecular datasets.
- **HEDMoL is relatively robust to the size of the training data and achieves state-of-the-art accuracy for most settings.**



# Prediction Accuracy for Different External Calculation Databases

Experimental Evaluations on Prediction Accuracy of HEDMoL for Different Volumes and Diversities of External Calculation Databases



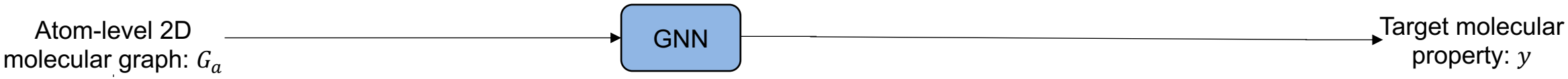

A subset of QM9 (molecules with six or fewer atoms)

- The quality of the external calculation database in the knowledge extension step is a crucial factor in the prediction accuracy of HEDMoL.
- We construct the external database from the QM9 dataset by extracting small molecules composed of six or fewer atoms.
- **HEDMoL is robust to the size of the external database.**

| Dataset | $C = 3$ | $C = 4$ | $C = 5$ | $C = 6$ | $C = 7$ | $C = 8$ |
|---|---|---|---|---|---|---|
| Lipop | 0.732 (0.037) | 0.723 (0.053) | 0.735 (0.047) | 0.736 (0.030) | 0.738 (0.040) | 0.736 (0.037) |
| ESOL | 0.914 (0.018) | 0.911 (0.016) | 0.915 (0.015) | 0.915 (0.016) | 0.916 (0.015) | 0.914 (0.019) |
| ADMET | 0.867 (0.007) | 0.862 (0.015) | 0.868 (0.012) | 0.861 (0.010) | 0.867 (0.012) | 0.865 (0.014) |
| IGC50 | 0.829 (0.017) | 0.833 (0.012) | 0.828 (0.016) | 0.835 (0.017) | 0.825 (0.018) | 0.831 (0.016) |

# Ablation Studies

Ablation Studies to Evaluate Effectiveness of Each Module in HEDMoL

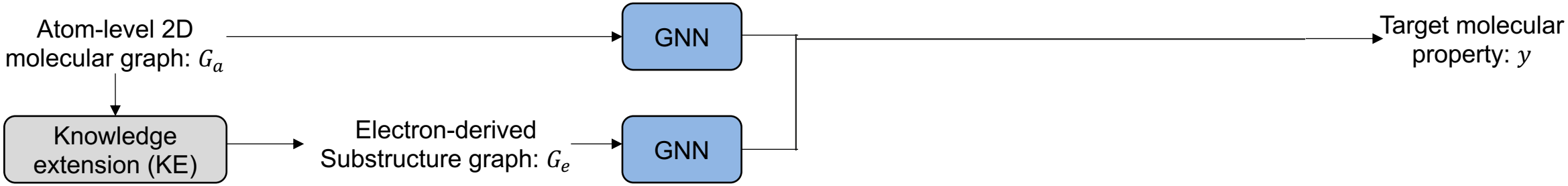



- EGC and GIN: Baseline 2D GNNs that predict $y$ from $G_a$.
- **EGC and GIN show significantly low prediction accuracy because they are solely based on the atom-level 2D structures.**

| Dataset | EGC | GIN | KE | KE+HRL | KE+PCR | HEDMoL |
|---|---|---|---|---|---|---|
| Lipop | 0.708 (0.043) | 0.702 (0.031) | 0.722 (0.051) | 0.726 (0.049) | 0.731 (0.044) | 0.759 (0.043) |
| ESOL | 0.896 (0.017) | 0.897 (0.022) | 0.913 (0.015) | 0.913 (0.015) | 0.915 (0.017) | 0.914 (0.016) |
| ADMET | 0.838 (0.012) | 0.833 (0.017) | 0.862 (0.014) | 0.864 (0.015) | 0.860 (0.013) | 0.865 (0.014) |
| IGC50 | 0.808 (0.029) | 0.799 (0.021) | 0.827 (0.018) | 0.833 (0.017) | 0.826 (0.014) | 0.840 (0.010) |
| LC50 | 0.575 (0.045) | 0.543 (0.080) | 0.631 (0.091) | 0.640 (0.084) | 0.603 (0.061) | 0.663 (0.053) |

# Ablation Studies

Ablation Studies to Evaluate Effectiveness of Each Module in HEDMoL

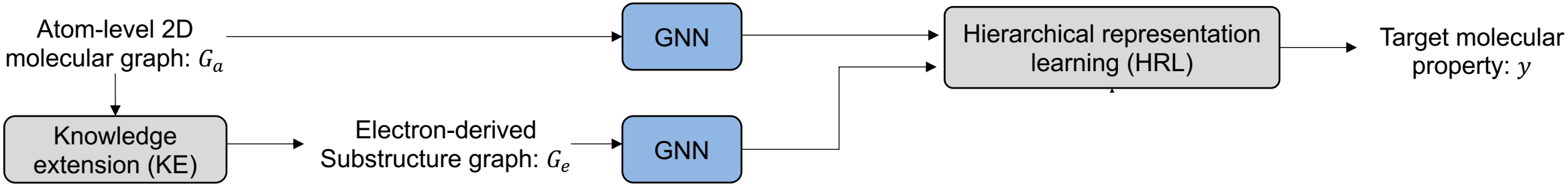



- KE: Predicts $y$ from the molecular embeddings of both branches (atom-level and electron-level)
- **KE shows consistent accuracy improvements for all benchmark datasets.**

| Dataset | EGC | GIN | KE | KE+HRL | KE+PCR | HEDMoL |
|---|---|---|---|---|---|---|
| Lipop | 0.708 (0.043) | 0.702 (0.031) | 0.722 (0.051) | 0.726 (0.049) | 0.731 (0.044) | 0.759 (0.043) |
| ESOL | 0.896 (0.017) | 0.897 (0.022) | 0.913 (0.015) | 0.913 (0.015) | 0.915 (0.017) | 0.914 (0.016) |
| ADMET | 0.838 (0.012) | 0.833 (0.017) | 0.862 (0.014) | 0.864 (0.015) | 0.860 (0.013) | 0.865 (0.014) |
| IGC50 | 0.808 (0.029) | 0.799 (0.021) | 0.827 (0.018) | 0.833 (0.017) | 0.826 (0.014) | 0.840 (0.010) |
| LC50 | 0.575 (0.045) | 0.543 (0.080) | 0.631 (0.091) | 0.640 (0.084) | 0.603 (0.061) | 0.663 (0.053) |

# Ablation Studies

Ablation Studies to Evaluate Effectiveness of Each Module in HEDMoL

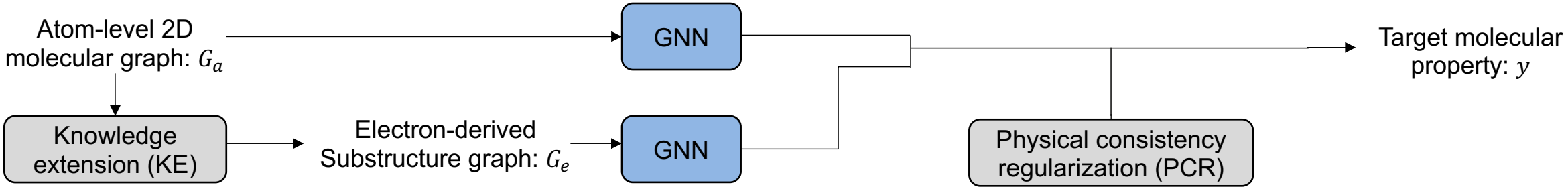



- KE+HRL: Employs KE and HRL in molecular representation learning.
- **KE+HRL shows marginal accuracy improvements over KE.**

| Dataset | EGC | GIN | KE | KE+HRL | KE+PCR | HEDMoL |
|---|---|---|---|---|---|---|
| Lipop | 0.708 (0.043) | 0.702 (0.031) | 0.722 (0.051) | 0.726 (0.049) | 0.731 (0.044) | 0.759 (0.043) |
| ESOL | 0.896 (0.017) | 0.897 (0.022) | 0.913 (0.015) | 0.913 (0.015) | 0.915 (0.017) | 0.914 (0.016) |
| ADMET | 0.838 (0.012) | 0.833 (0.017) | 0.862 (0.014) | 0.864 (0.015) | 0.860 (0.013) | 0.865 (0.014) |
| IGC50 | 0.808 (0.029) | 0.799 (0.021) | 0.827 (0.018) | 0.833 (0.017) | 0.826 (0.014) | 0.840 (0.010) |
| LC50 | 0.575 (0.045) | 0.543 (0.080) | 0.631 (0.091) | 0.640 (0.084) | 0.603 (0.061) | 0.663 (0.053) |

# Ablation Studies

Ablation Studies to Evaluate Effectiveness of Each Module in HEDMoL

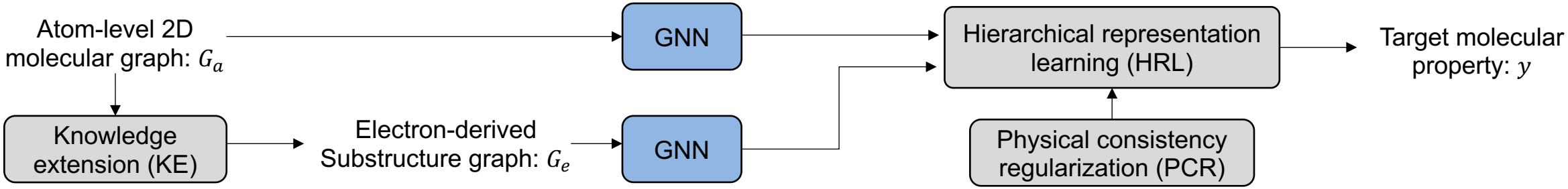



- KE+PCR: Utilizes PCR in molecular representation learning processes of individual EGC and GIN.
- **The performance turns out to be similar to that of KE+HRL.**

| Dataset | EGC | GIN | KE | KE+HRL | KE+PCR | HEDMoL |
|---|---|---|---|---|---|---|
| Lipop | 0.708 (0.043) | 0.702 (0.031) | 0.722 (0.051) | 0.726 (0.049) | 0.731 (0.044) | 0.759 (0.043) |
| ESOL | 0.896 (0.017) | 0.897 (0.022) | 0.913 (0.015) | 0.913 (0.015) | 0.915 (0.017) | 0.914 (0.016) |
| ADMET | 0.838 (0.012) | 0.833 (0.017) | 0.862 (0.014) | 0.864 (0.015) | 0.860 (0.013) | 0.865 (0.014) |
| IGC50 | 0.808 (0.029) | 0.799 (0.021) | 0.827 (0.018) | 0.833 (0.017) | 0.826 (0.014) | 0.840 (0.010) |
| LC50 | 0.575 (0.045) | 0.543 (0.080) | 0.631 (0.091) | 0.640 (0.084) | 0.603 (0.061) | 0.663 (0.053) |

# Ablation Studies

Ablation Studies to Evaluate Effectiveness of Each Module in HEDMoL

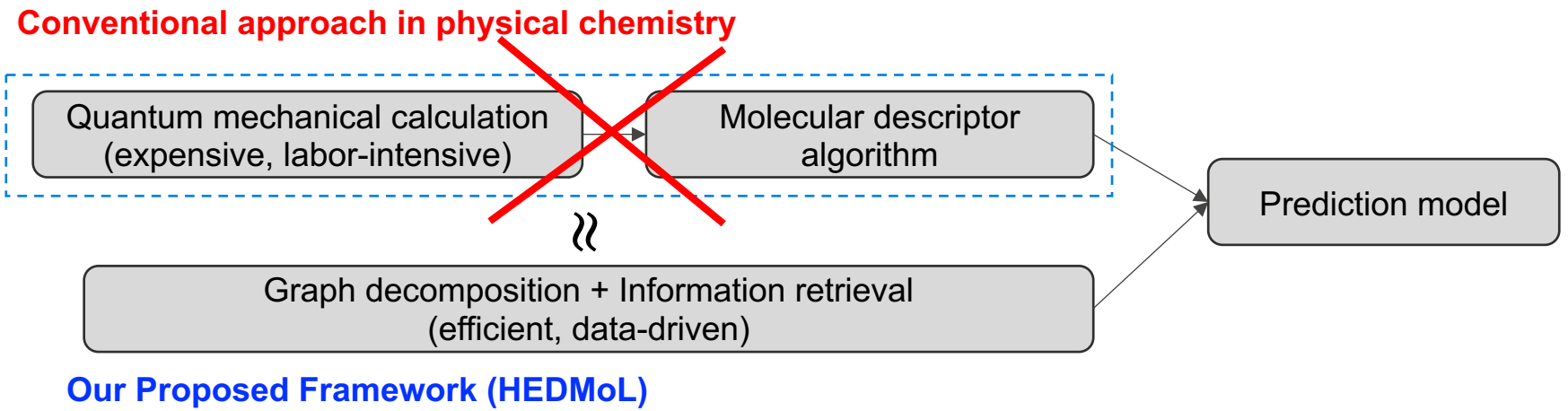



- HEDMoL, which implements all modules, outperformed the baseline methods and ablation models.
- **The evaluations results of the ablation study empirically demonstrated the effectiveness of each module in HEDMoL.**

| Dataset | EGC | GIN | KE | KE+HRL | KE+PCR | HEDMoL |
|---|---|---|---|---|---|---|
| Lipop | 0.708 (0.043) | 0.702 (0.031) | 0.722 (0.051) | 0.726 (0.049) | 0.731 (0.044) | 0.759 (0.043) |
| ESOL | 0.896 (0.017) | 0.897 (0.022) | 0.913 (0.015) | 0.913 (0.015) | 0.915 (0.017) | 0.914 (0.016) |
| ADMET | 0.838 (0.012) | 0.833 (0.017) | 0.862 (0.014) | 0.864 (0.015) | 0.860 (0.013) | 0.865 (0.014) |
| IGC50 | 0.808 (0.029) | 0.799 (0.021) | 0.827 (0.018) | 0.833 (0.017) | 0.826 (0.014) | 0.840 (0.010) |
| LC50 | 0.575 (0.045) | 0.543 (0.080) | 0.631 (0.091) | 0.640 (0.084) | 0.603 (0.061) | 0.663 (0.053) |

# Conclusion

- We propose HEDMoL for learning electron-derived molecular representations of real-world molecules.
- We develop a decomposition-based information retrieval process to generate electron-informed molecular graphs without quantum mechanical calculations.
- HEDMoL replaces an expensive and labor-intensive quantum mechanical calculations with an efficient and data-driven algorithm.



# Conclusion

- Contact: cy.park@kaist.ac.kr
- Lab homepage: https://dsail.kaist.ac.kr/

Thanks for your attention!

# Reference


[1] Chen, C., Ye, W., Zuo, Y., Zheng, C., & Ong, S. P. (2019). Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, *31*(9), 3564-3572.

[2] Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., ... & Zheng, M. (2019). Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, *63*(16), 8749-8760.

[3] Chen, C., & Ong, S. P. (2022). A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, *2*(11), 718-728.

[4] Yang, S., Li, Z., Song, G., & Cai, L. (2021). Deep molecular representation learning via fusing physical and chemical information. *Advances in neural information processing systems*, *34*, 16346-16357.

[5] Duval, A. A., Schmidt, V., Hernández-Garcıa, A., Miret, S., Malliaros, F. D., Bengio, Y., & Rolnick, D. (2023, July). Faenet: Frame averaging equivariant gnn for materials modeling. ICML (pp. 9013-9033). PMLR.

[6] Lee, Y. L., Lee, H., Kim, T., Byun, S., Lee, Y. K., Jang, S., ... & Im, J. (2022). Data-driven enhancement of ZT in SnSe-based thermoelectric systems. *Journal of the American Chemical Society*, *144*(30), 13748-13763.

[7] Roy, D. D., Roy, P., & De, D. (2023). Machine learning and DFT-based combined framework for predicting transmission spectra of quantum-confined bio-molecular nanotube. *Journal of Molecular Modeling*, *29*(11), 338.