Key Substructure Learning with Chemical Intuition for Material Property Prediction

Peiliang Zhang[1], Jingling Yuan[1]*, Lin Li[1], Wen Luo[1], Jiwei Hu[2], Xin Li[3, 4]

1 Wuhan University of Technology

2 Wuhan Fiberhome Technical Services Co.,Ltd

3 University of Science and Technology of China

4 iFLYTEK Co., Ltd

2024.07.03

Background & Motivation

Background

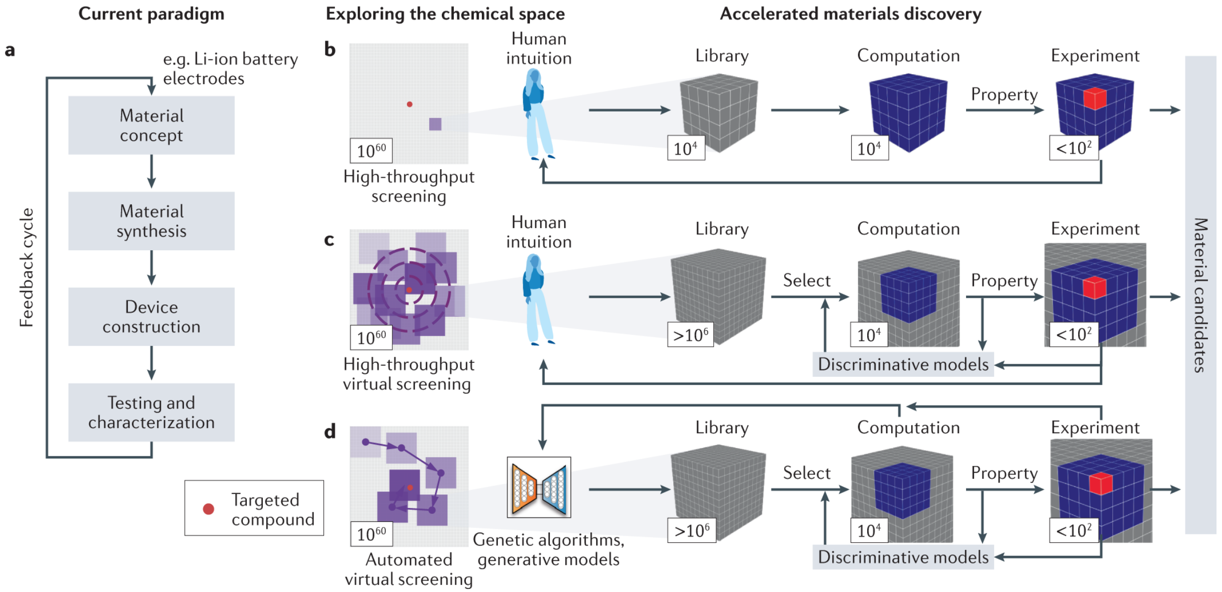

Fig 1. Traditional and accelerated approaches to materials discovery [1].

Motivation

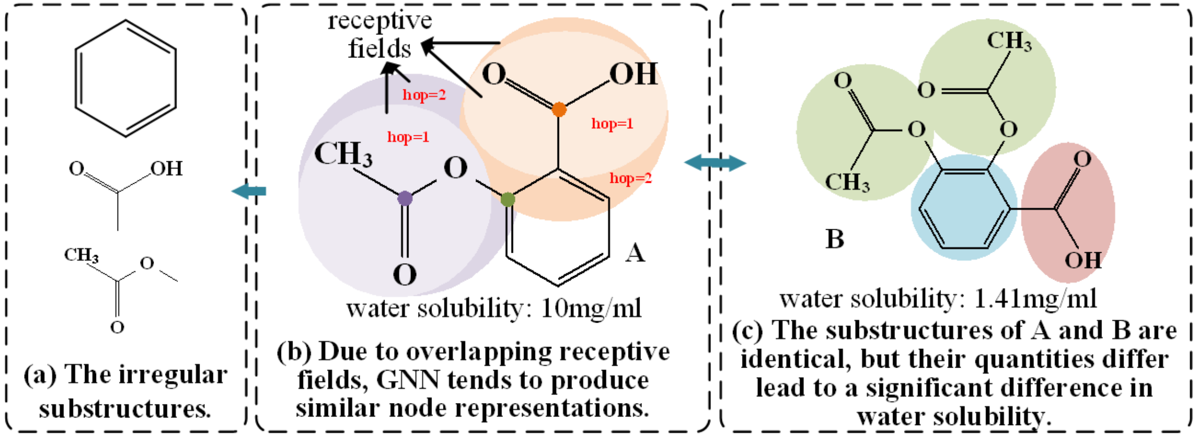

Fig 2. The motivation statement for KSCI.

- The topological information for **irregular substructures**
- The **importance of substructures** to molecular properties

Related Work

GNN-based	Consist of multilayer GNNs	Over-smoothing problem[2]	**The substructure topological difference**	
Transformer -based	Handling fixed-size substructures	Lack of adaptability to structures	**The irregularity of substructures**	

Key Substructure Learning with Chemical Intuition (KSCI)

[1] Yao Z, et al. Machine learning for a sustainable energy future. Nature Reviews Materials, 2023.
[2] Jaiswal A, et al. Graph ladling: Shockingly simple parallel GNN training without intermediate communication. ICML, 2023.

Methodology

Overview

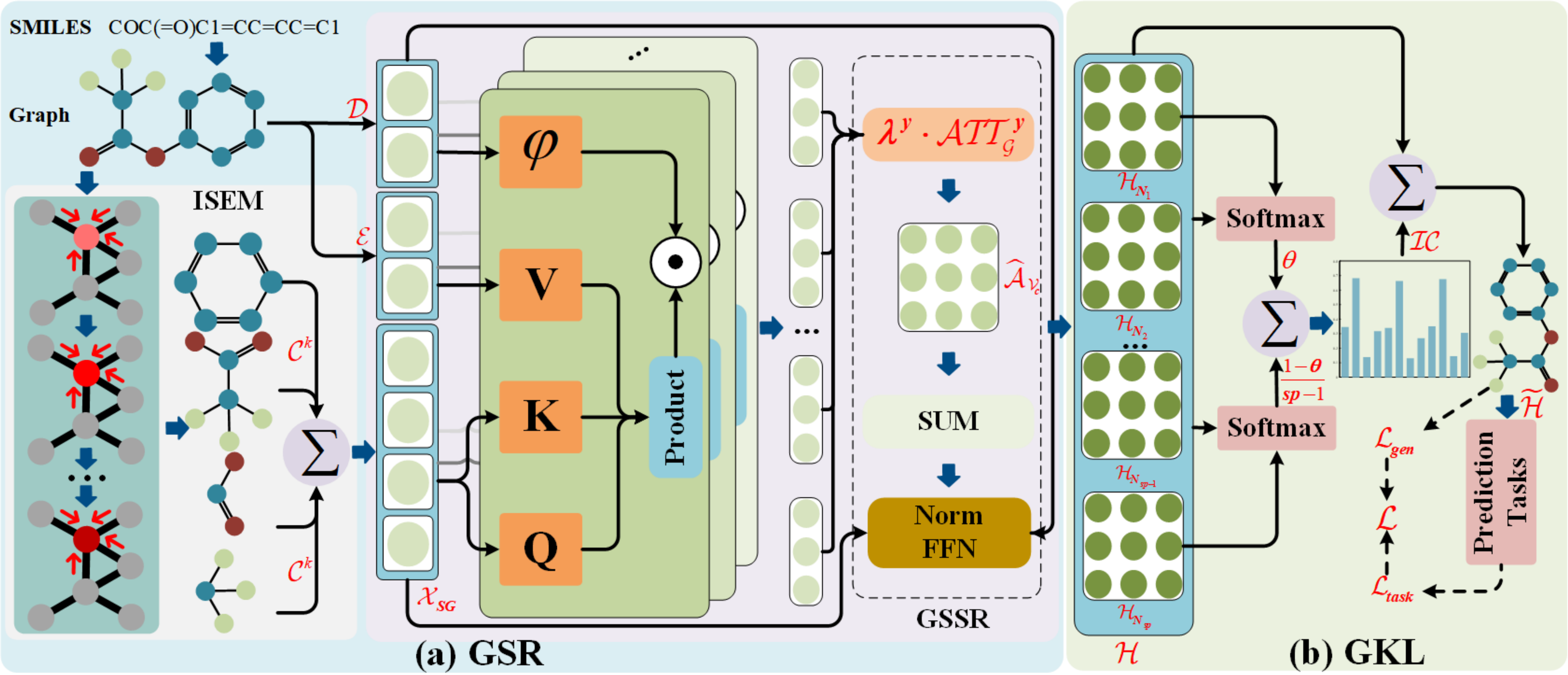

Fig 3. The structure of the KSCI.

KSCI is driven by chemical intuition[1] and mainly consists of two modules: Graph Self-attention-based Irregular Substructure Representation (GSR), as shown in Fig 3(a) and Gain-driven Key Substructure Learning (GKL) , as shown in Fig 3(b) .

[1] Choung O, et al. Extracting medicinal chemistry intuition via preference machine learning. Nature Communications, 2023.

Methodology

Graph Self-attention-based Irregular Substructure Representation (GSR)

Chemical Intuition: Molecular substructures are usually irregular in size, shape, and atoms. Different substructures also exhibit distinct topological structures.

➢ Irregular Substructure Extraction Module (ISEM)

ISEM utilizes subgraph feature extraction to calculate the weighted sum of substructure coefficient and substructure features with different hops to generate irregular molecular substructure embedding.

- The irregular substructure representation:

$$\mathcal{X}_{SG}(\mathcal{V}_c,\mathcal{G}) = \sum_{l=1}^{L_G}\sum_{k=1}^{K}\sum_{\mathcal{V}_s\in\mathcal{V}_c^k} \mathcal{C}^k \cdot GNN_l^k(\mathcal{V}_s)$$

- The substructure coefficient:

$$\mathcal{C}^k = \frac{IE_{\mathcal{V}_c^k}}{\sum_{\mathcal{V}_s\in\mathcal{V}_c^k} IE_{\mathcal{V}_s^k}}$$

- The node feature entropy weighting:

$$IE_{\mathcal{V}_c^k} = \frac{exp(\mathcal{Z}(\mathcal{V}_c))\cdot\log(exp(\mathcal{Z}(\mathcal{V}_c)))}{\sum_{\mathcal{V}_s\in\mathcal{V}_c^{w/o}} exp(\mathcal{Z}(\mathcal{V}_s))\cdot\log(exp(\mathcal{Z}(\mathcal{V}_s)))}$$

➢ Graph Self-attention-based Substructure Representation (GSSR)

Considering the topological features of material molecular structure and the limitations of GNNs, we design graph self-attention for extracting feature information of molecules. The graph self-attention extracts the molecular feature using the position encoding linear function φ(∗) and the graph kernel function Φ(∗).

- For a molecular atom $\mathcal{V}_c$, we generate its absolute position encoding embedding by $\varphi\left(\mathcal{D}_{\mathcal{V}_c}\right) = \mathcal{P}_{ae}(\mathcal{D}_{\mathcal{V}_c}) = [\mathcal{P}^1,\mathcal{P}^2,\ldots,\mathcal{P}^{RS}]$.
- Based on the absolute position of atoms and the substructure topological information, we design graph self-attention:

$$\mathcal{ATT}_{\mathcal{G}} = \Phi(Sub(\mathcal{V}_c))\cdot\varphi\left(\mathcal{D}_{\mathcal{V}_c}\right) = \sum_{\mathcal{V}_m\in\mathcal{V}} \frac{\delta_{\mathcal{G}}(\mathcal{V}_c,\mathcal{V}_m)}{\sum_{\mathcal{V}_w\in\mathcal{V}}\delta_{\mathcal{G}}(\mathcal{V}_c,\mathcal{V}_w)}\cdot\gamma(\mathcal{E}_c)\cdot\varphi\left(\mathcal{D}_{\mathcal{V}_c}\right)$$

Learning from the multi-head self-attention of Transformer encoder, the aggregation computation of $\mathcal{ATT}_{\mathcal{G}}$ is:

$$\widehat{\mathcal{A}}_{\mathcal{V}_c} = \sum_{y=1}^{Y}\lambda^y\cdot\mathcal{ATT}_{\mathcal{G}}^y(\mathcal{V}_c) = \sum_{y=1}^{Y}\frac{exp(\mathcal{ATT}_{\mathcal{G}}^y(\mathcal{V}_c))}{\sum_{t=1}^{Y}\mathcal{O}_{[t\neq y]}\cdot exp(\mathcal{ATT}_{\mathcal{G}}^t(\mathcal{V}_c))}\cdot\mathcal{ATT}_{\mathcal{G}}^y(\mathcal{V}_c)$$

Methodology

Gain-driven Key Substructure Learning (GKL)

Fundamental Theories: (1) **the main efficacy of molecules is usually determined by one or several functional groups (i.e., substructures)** of the material molecules [1]. (2) **the important functional groups can dictate molecular unique properties**, while the underlying functional groups usually determine molecular fundamental properties [2] .

- Considering the chemical above foundational theories, we design a Cross Probability Function (CPF) based on the network structure to simulate the effect of functional group synergy on molecular properties:

$$\mathcal{R}_{\mathcal{V}_{ci}}^{h}=\sum_{i=1}^{Sp}(\underbrace{\theta \cdot \sigma(\mathcal{H}_{\mathcal{V}_{ci}}^{h},\mathcal{H}^{h})}_{I}+\underbrace{(1-\theta)\cdot \sum_{j=1,j\neq i}^{Sp}\sigma(\mathcal{H}_{\mathcal{V}_{cj}}^{h},\mathcal{H}^{h})/(Sp-1)}_{U})$$

I simulates the determinacy of important substructures on molecular properties, and U calculates the effect of other substructures on the molecular underlying properties.

- GKL quantifies the importance of each substructure based on substructure gain in material molecule . The computation of the finish representation $\widetilde{\mathcal{H}}_h$ for material molecule $\mathcal{D}_h$ is carried out as follows:

$$\widetilde{\mathcal{H}}_h=\sum_{s_p=0}^{S_P-1}\frac{\mathcal{D}_h^{Sub}(0,s_p)\cdot IC_h(s_p,0)}{\rho}$$

Prediction Tasks and Model Optimization

The generation loss $\mathcal{L}_{gen}$ in KSCI is defined as the reconstruction error between input features $\mathcal{D}_h$ and the generated graph-level embedding $\mathcal{H}_h$. The property prediction loss functions $\mathcal{L}_{task}$ is defined as the Mean Absolute Error Loss.

$$\mathcal{L}=\alpha\cdot\mathcal{L}_{gen}+\beta\cdot\mathcal{L}_{task}=\alpha\cdot\left\|\mathcal{D}_h-\mathcal{H}_h\right\|_2+\beta\cdot\frac{1}{|\mathcal{A}|}\sum_{a=1}^{A}\left|\mathcal{R}eal_a-\mathcal{R}eg_a\right|$$

[1] Mokaya M, et al. Testing the limits of smiles-based de novo molecular generation with curriculum and deep reinforcement learning. Nature Machine Intelligence, 2023.
[2] Wang H, et al. Scientific discovery in the age of artificial intelligence, 2023.

Experiments & Results

Datasets and Experimental Setup

- **Dataset**
  - Four material dataset: QMOF, C2DB, Materials Project, and hMOF
  - training set: validation set: test set = 6: 2: 2
- **Property Prediction**
  - Predictive Properties: band gap, work function, formation and adsorption
  - Evaluation Indicator: Mean Absolute Error (MAE)
- **Experimental Setup**
  - Training: two NVIDIA GeForce RTX 4090 24G computing graphics cards and Intel(R) Core(TM) i9-12900KF
  - Optimizer: Adam
  - Parameter Settings: learning from{0.01, 0.005, 0.001, 0.0005, 0.0001}, θ is 0.6, iterations from {1, 2, 3, 5, 10, 15, 20}, batch size is 128, dropout rate is 0.1, Graph self-attention heads and neighbor aggregation hops are 2, and encoder layers are 6.
- **Comparison Models**
  - Material Domain Models: CGCNN, SchNet, MEGNet, MPNN, MoFormer, GATGNN, GMPNN, BNM-CDGNN
  - Generalized Molecular Representation Models: MR-GNN, SAN, SAT

Table 1. The statistics of experimental datasets.

Dataset	#Molecular	Property	Unit	#Training	#Validation	#Test
QMOF	18,633	Band Map	eV	11,180	3,727	3,726
C2DB	3,316	Work Function (WF)	eV	1,990	663	663
MP	32,665	Formation Band Gap	eV/atom eV	19,599	6,533	6,533
hMOF	113,665	CO_2 Adsorption CH_4 Adsorption	mol/kg mol/kg	68,199	22,733	22,733

Experiments & Results

Gain-driven Key Substructure Learning (GKL)

➢ Overall Prediction Performance of KSCI

KSCI achieves the optimal performance in both QMOF, C2DB, and Materials Project datasets, which indicates the effectiveness of KSCI in molecular representation and property prediction.

➢ Performance Comparison of Model Types

There are differences between domain models and general models in property prediction performance, which is caused by the capture of critical properties in molecular structures. The comprehensive performance of the most GNNs-based models (CGCNN, MPNN, GATGNN and BNM-CDGNN) outperforms the Transformer-based models in both datasets. This is mainly attributed to the representation of molecular topology by GNNs.

➢ Performance Analysis of Different Properties

In the different property predictions, KSCI and BNM-CDGNN achieve satisfactory prediction performance. Nevertheless, the prediction performance of BNM-CDGNN is suboptimal in datasets with more elemental categories, such as QMOF (79 elements) and C2DB (60 elements).

Table 2. The results of property prediction by various models. (The bold values indicate the best results and the underline denotes the second-best results.)

Type	Model	Backbone	QMOF	C2DB	Materials Project	
			Band Gap	WF	Formation	Band Gap
Domain	CGCNN [23]	GNN	0.283	0.226	0.054	0.283
	SchNet [16]	CNN	0.316	0.231	0.063	0.307
	MEGNet [2]	GNN	0.281	0.224	0.053	0.291
	MoFormer [1]	GNN&T	0.401	0.226	0.059	0.289
	MPNN [25]	GNN	0.305	0.229	0.052	0.289
	GATGNN [9]	GNN	0.303	0.218	0.051	0.296
	GMPNN [14]	GNN	0.299	0.223	0.047	<u>0.278</u>
	BNM-CDGNN [10]	GNN	<u>0.269</u>	<u>0.201</u>	**0.043**	0.279
General	MR-GNN [24]	GNN	0.312	0.228	0.059	0.281
	SAN [8]	T	0.312	0.224	0.048	0.281
	SAT [3]	GT	0.301	0.216	0.045	0.279
	KSCI	GT	**0.268**	**0.189**	**0.043**	**0.270**

Type	Model	Backbone	hMOF			
			CO_2 Adsorption		CH_4 Adsorption	
			0.5 bar	2.5 bar	0.5 bar	2.5 bar
Domain	CGCNN [23]	GNN	0.416	0.862	0.145	0.362
	SchNet [16]	CNN	0.446	0.924	0.150	0.375
	MEGNet [2]	GNN	0.428	0.856	0.147	0.359
	MoFormer [1]	GNN&T	0.569	1.25	0.196	0.403
	MPNN [25]	GNN	0.398	0.727	0.136	0.342
	GATGNN [9]	GNN	0.379	0.741	0.135	0.332
	GMPNN [14]	GNN	0.350	0.755	0.137	0.302
	BNM-CDGNN [10]	GNN	**0.336**	<u>0.683</u>	<u>0.132</u>	<u>0.286</u>
General	MR-GNN [24]	GNN	0.381	0.841	0.139	0.336
	SAN [8]	T	0.371	0.902	0.146	0.323
	SAT [3]	GT	0.353	0.819	0.144	0.303
	KSCI	GT	<u>0.339</u>	**0.662**	**0.131**	**0.269**

Experiments & Results

Ablation Experiments

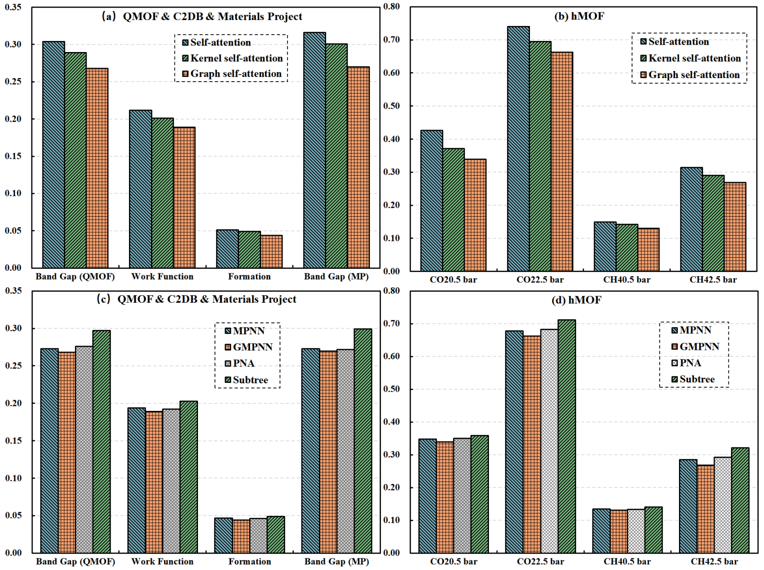

Fig 4. The results of KSCI and its various variants in property predictions.

- ➢ **Effect of Graph Self-attention**
- We replace the graph self-attention in GSSR with either original self-attention or kernel self-attention to verify the effect of different attention computation methods on prediction performance.
- The graph self-attention outperforms the other two methods, while the performance gain achieved by graph self-attention is much higher than that of kernel self-attention.
- ➢ **Effect of Irregular Molecular Substructure.**
- We apply the subtree-based extraction and different subgraph-based extraction methods, such as MPNN, GMPNN, and PNA, to the substructure.
- Subtree-based substructure extraction performs poorly in prediction tasks. MPNN-based substructure coefficient exhibits more flexibility in extracting irregular molecular substructures.

Visual Explanations for KSCI

- ➢ Compared with kernel self-attention, the computation results of KSCI are sparser and better identify the substructures. (Purple and red areas)
- ➢ The graph self-attention highlights the differences between groups centered on A and B, showing that KSCI is equally advantageous in distinguishing similar substructures.

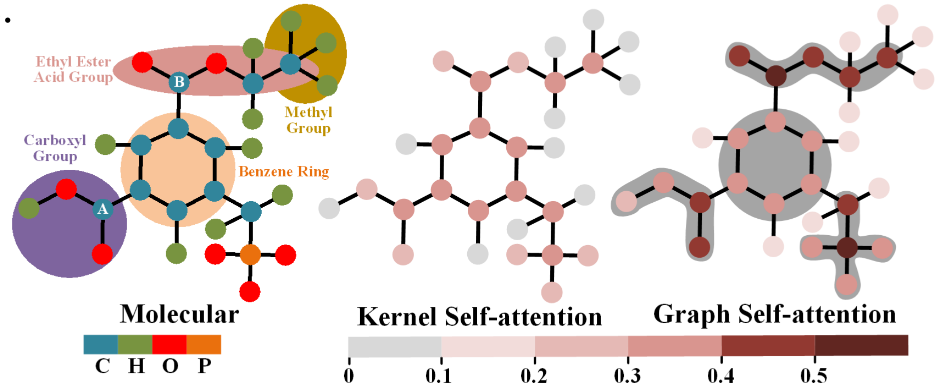

Fig 5. The node features visualization of RUCRUE FSR in different attention.

Conclusion and Acknowledgement

Conclusion

We propose a Key Substructure Representation with Chemical Intuition for Material Property Prediction.

- We propose KSCI for chemical intuition-driven key substructure representation learning to material property prediction, which is promising to accelerate material screening and reduce the research and development cycles for new materials.
- Substructure graph self-attention explicitly encodes topological information about irregular substructures. GKL uncovers the differential gain of sub-structures to molecular representation and highlights the role of key sub structures in molecular property.
- In material property prediction, our proposed KSCI outperforms the state of-the-art model in four real-world material datasets and reduces the Mean Absolute Error from 0.37% to 5.97%.

Acknowledgement

- the National Key Research and Development Program of China (2022YFB2404300)
- the National Natural Science Foundation of China (62276196)
- the Key Research and Development Program of Hubei Province (2021BAA030)
- the Special Project for High-Quality Development of Manufacturing Industry in Hubei Province (2206-420118-89-04-959008)
- the Fundamental Research Funds for the Central Universities (2023vb041)

Thank you for listening