Molecular Contrastive Learning with Chemical Element Knowledge Graph

Yin Fang[1,2,3*], Qiang Zhang[1,2,3*], Haihong Yang[1,2,3], Xiang Zhuang[1,2,3], Shumin Deng[1,3], Wen Zhang[4], Ming Qin[1,2,3], Zhuo Chen[1,2,3], Xiaohui Fan[5,6,7], Huajun Chen[1,2,3†]

[1] College of Computer Science and Technology, Zhejiang University
[2] ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University
[3] AZFT Joint Lab for Knowledge Engine [4] School of Software Technology, Zhejiang University
[5] Pharmaceutical Informatics Institute, College of Pharmaceutical Sciences, Zhejiang University
[6] Innovation Center in Zhejiang University, State Key Laboratory of Component-Based Chinese Medicine
[7] Westlake Laboratory of Life Sciences and Biomedicine
{fangyin, qiang.zhang.cs, haihong825, zhuangxiang, 231sm, wenzhang2015, qinandming, zhuo.chen, fanxh, huajunsir}@zju.edu.cn

Code：https://github.com/ZJU-Fangyin/KCL

（AAAI-2022）

gesis Leibniz-Institut für Sozialwissenschaften

Reported by Rongrong Ma

1. Introduction

2. Approach

3. Experiments

Introduction

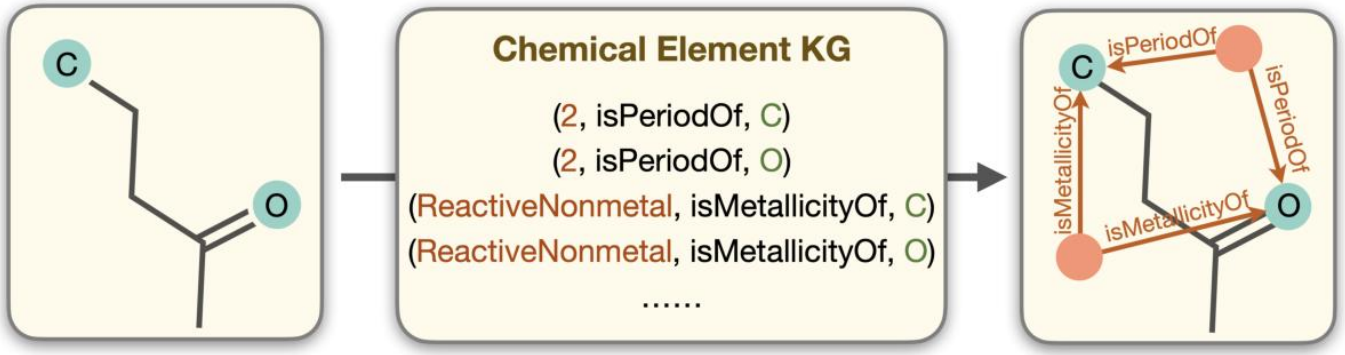


Figure 1: Chemical Element KG builds associations between atoms that are not directly connected by bonds but related in fundamental chemical attributes, as denoted by red arrows.

Prior works fail to incorporate fundamental domain knowledge into graph semantics and thus ignore the correlations between atoms that have common attributes but are not directly connected by bonds.

Method

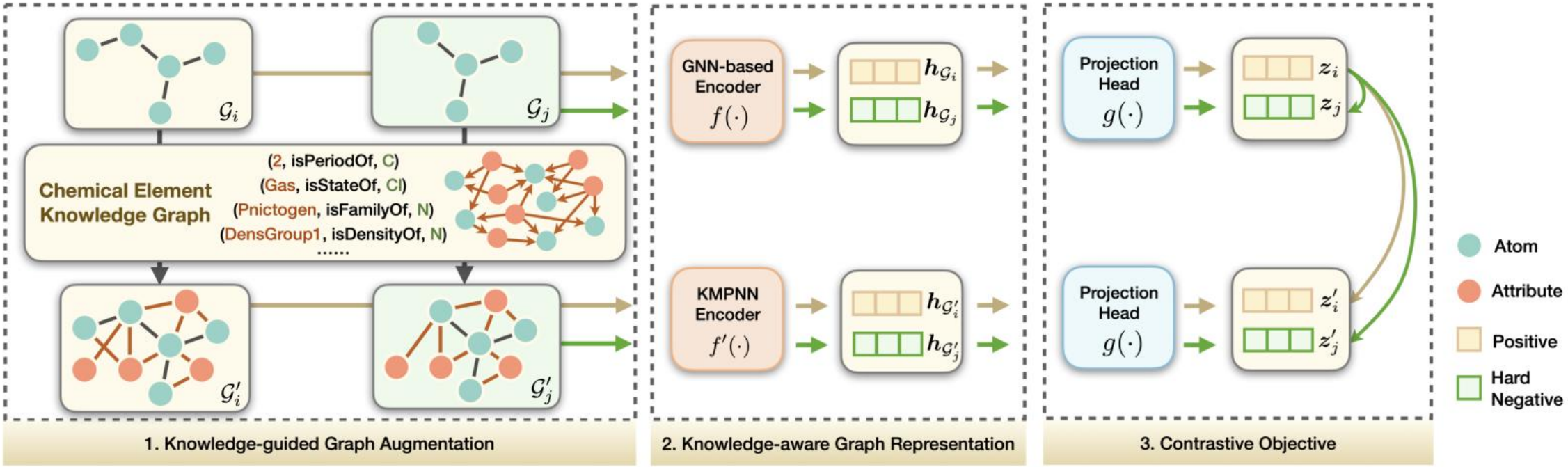


Figure 2: An illustrative example for KCL. We ignore edge directions in four molecular graphs due to space limitation (the direction of an edge between an attribute and an atom is from the former to the latter, while an edge between atoms is bidirectional). Module 1: Knowledge-guided graph augmentation converts the original molecular graph $\mathcal{G}$ into the augmented molecular graph $\mathcal{G}'$ based on Chemical Element KG. Module 2: Knowledge-aware graph representation captures representations from two graph views separately. Module 3: Contrastive objective trains the encoders and the projection head to maximize agreement between positives and disagreement between hard negatives (e.g., $\mathcal{G}_j$ act as the hard negative of $\mathcal{G}_i$) via a contrastive loss.

Method

	Chemical Element KG
Elements	118
Attributes	107
Entities	225
Relation Types	17
KG Triples	1643

Table 1: The statistics of Chemical Element KG.

triples in the form of (Gas, isStateOf, Cl)

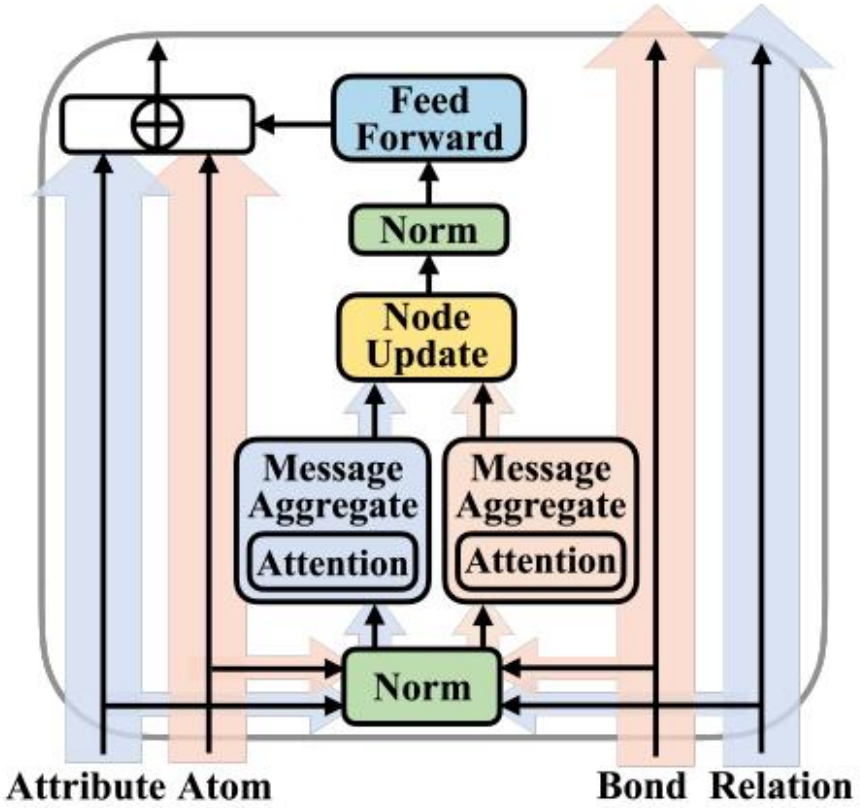


Figure 5: Architecture of KMPNN.

Method

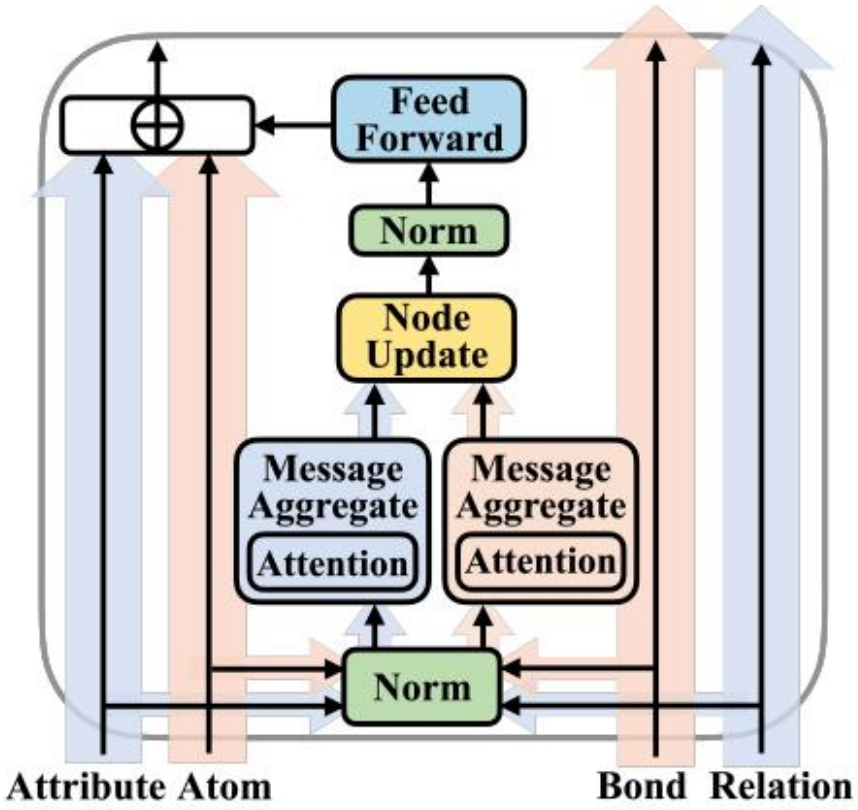


Figure 5: Architecture of KMPNN.

$$\alpha_{uv} = \frac{\exp\left(\text{LeakyReLU}\left(\boldsymbol{a}^T\left[\boldsymbol{W}\boldsymbol{h}_u || \boldsymbol{W}\boldsymbol{h}_v\right]\right)\right)}{\sum_{k\in\mathcal{N}_u}\exp\left(\text{LeakyReLU}\left(\boldsymbol{a}^T\left[\boldsymbol{W}\boldsymbol{h}_u || \boldsymbol{W}\boldsymbol{h}_k\right]\right)\right)}, \quad (1)$$

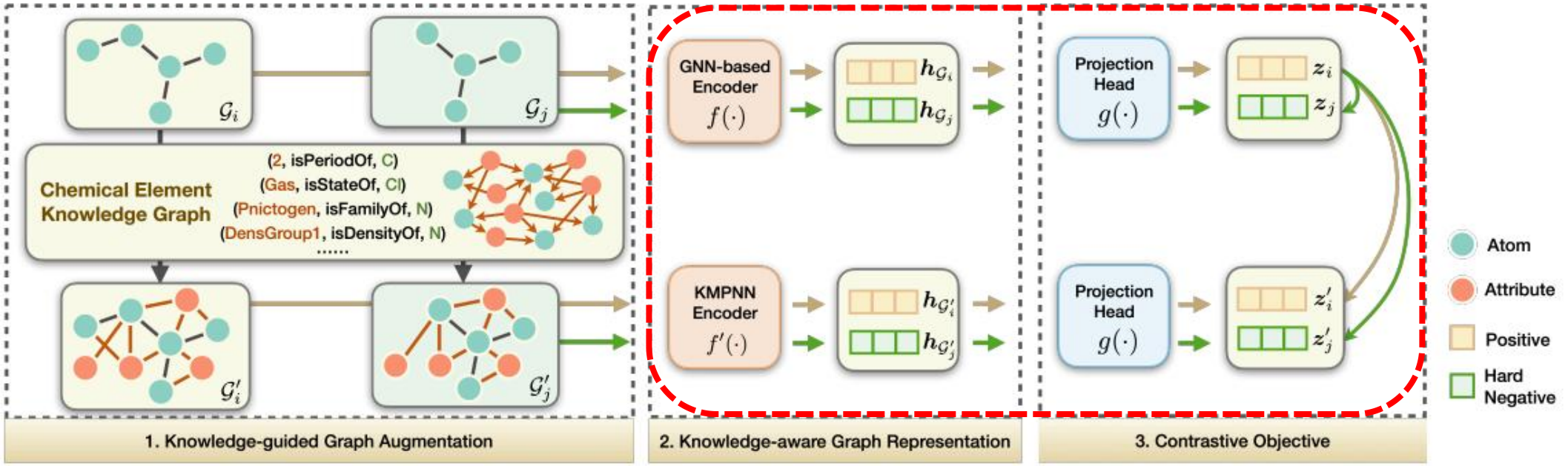
$$\text{MSG}_0 = \alpha_{uv}\boldsymbol{W}_0\boldsymbol{h}^{k-1}(e_{uv}) \cdot \boldsymbol{h}^{k-1}(u), \quad (2)$$

$$\text{MSG}_1 = \beta_{uv}\boldsymbol{W}_1\boldsymbol{h}^{k-1}(e_{uv}) \cdot \boldsymbol{h}^{k-1}(u), \quad (3)$$

$$\boldsymbol{m}^k(v) = \sum_{k\in\mathcal{N}_u}\text{MSG}(\boldsymbol{h}^{k-1}(e_{uv}), \boldsymbol{h}^{k-1}(u)), \quad (4)$$

$$\boldsymbol{h}^k(v) = \text{GRU}(\boldsymbol{h}^{k-1}(v), \boldsymbol{m}^k(v)), \quad (5)$$

Method



Morgan Fingerprints

$$\boldsymbol{h}_{\mathcal{G}'} = \text{Set2set}(\boldsymbol{h}^K(v)), \tag{6}$$

$$s(\boldsymbol{e}_1, \boldsymbol{e}_2) = \frac{N_{12}}{N_1 + N_2 - N_{12}}, \tag{7}$$

$$\ell_i = -\log \frac{e^{\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}'_i)/\tau}}{\sum_{j=1}^{N} \left(e^{\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}'_j)/\tau} + e^{\text{sim}(\boldsymbol{z}'_i, \boldsymbol{z}_j)/\tau} \right)}, \tag{8}$$

$$f(\mathbf{h}, \mathbf{r}, \mathbf{t}) = \|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\|, \tag{9}$$

$$\boldsymbol{h}^k(v) = \text{U}(\boldsymbol{h}^{k-1}(v), \text{Agg}(\{\boldsymbol{h}^{k-1}(u)), \forall u \in \mathcal{N}_v\})), \tag{10}$$

Experiments

Type	Category	Dataset	# Tasks	# Compounds	Split	Metric
Classification	Biophysics	BBBP	1	2039	Scaffold	ROC-AUC
	Physiology	SIDER	27	1427	Random	ROC-AUC
		ClinTox	2	1478	Random	ROC-AUC
		BACE	1	1513	Scaffold	ROC-AUC
		Tox21	12	7831	Random	ROC-AUC
		ToxCast	617	8575	Random	ROC-AUC
Regression	Physical chemistry	FreeSolv	1	642	Random	RMSE
		ESOL	1	1128	Random	RMSE

Table 5: Dataset information.

Experiments

Task	Classification (ROC-AUC)						Regression (RMSE)	
Dataset	BBBP	Tox21	ToxCast	SIDER	ClinTox	BACE	ESOL	FreeSolv
#Molecules	2039	7831	8575	1427	1478	1513	1128	642
#Tasks	1	12	617	27	2	1	1	1
GCN (Kipf and Welling 2017)	0.877	0.772	0.650	0.638	0.807	0.854	1.068	2.900
Weave (Kearnes et al. 2016)	0.837	0.741	0.678	0.621	0.823	0.791	1.158	2.398
MPNN (Gilmer et al. 2017)	0.913	0.808	0.691	0.641	0.879	0.815	1.167	2.185
DMPNN (Yang et al. 2019)	0.919	0.826	0.718	0.632	0.897	0.852	0.980	2.177
CMPNN (Song et al. 2020)	0.927	0.806	0.738	0.636	0.902	0.869	0.798	0.956
CoMPT (Chen et al. 2021)	0.938	0.809	0.740	0.634	0.934	0.871	0.774	1.855
N-GRAM (Liu et al. 2019)	0.912	0.769	-	0.632	0.870	0.876	1.100	2.512
Hu et al. (Hu et al. 2020)	0.915	0.811	0.714	0.614	0.762	0.851	-	-
GROVER (Rong et al. 2020)	0.940	0.831	0.737	0.658	0.944	0.894	0.831	1.544
KCL(GCN)	0.956	0.856	**0.757**	0.666	0.945	**0.934**	**0.582**	0.854
KCL(KMPNN)	**0.961**	**0.859**	0.740	**0.671**	**0.958**	0.924	0.732	**0.795**

Table 2: The property prediction performance (lower is better for regression) of KCL under the fine-tune protocol, compared with supervised learning (first group) and pre-training methods (second group) baselines on 8 datasets.

Experiments

Dataset	BBBP	Tox21	ToxCast	SIDER	ClinTox	BACE
Node	0.843	0.728	0.633	0.577	0.635	0.746
Edge	0.833	0.715	0.619	0.605	0.630	0.657
Subgraph	0.815	0.727	0.625	0.583	0.603	0.629
Attribute	0.826	0.726	0.623	0.621	0.671	0.796
InfoGraph	0.611	0.615	0.562	0.502	0.458	0.594
MICRO	0.830	0.718	0.595	0.573	0.735	0.708
GraphCL	0.697	0.739	0.624	0.605	0.760	0.755
JOAO	0.714	0.750	0.632	0.605	0.813	0.773
MoCL	0.905	0.768	0.653	0.628	0.750	0.845
KCL(G)	**0.929**	0.821	0.696	0.620	**0.909**	**0.902**
KCL(K)	0.927	**0.825**	**0.709**	**0.659**	0.898	0.860

Table 3: The performance of KCL under the linear protocol on 6 datasets, compared with contrastive learning baselines. The metric is ROC-AUC.

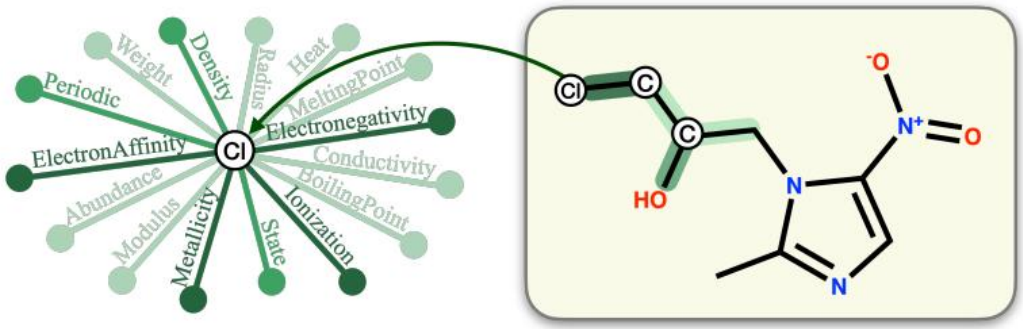


Figure 4: An attention visualization example of different types of neighbors (attributes and atoms) in the BBBP dataset. The attention weights assigned for bonds connected to the two C atoms are visualized on the right. The darker the color, the higher the attention.

Experiments

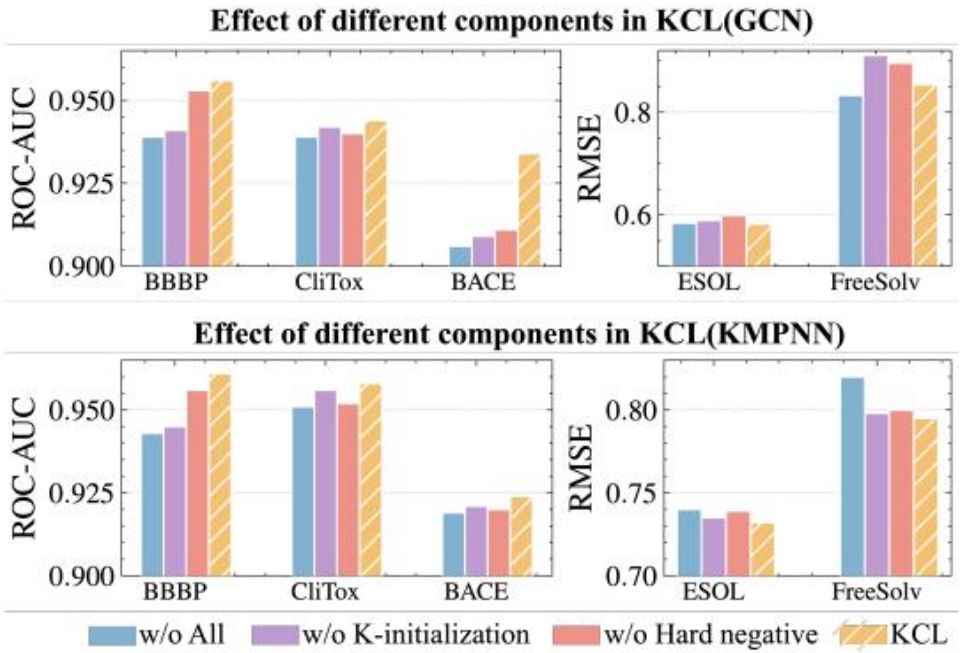


Figure 3: Performance of KCL with different settings under the fine-tune protocol (lower is better for regression).

Task	Classification	Regression
GCN(No contrast)	0.766	1.984
KMPNN(No contrast)	0.806	1.531
KCL(GIN)	0.849	0.718
KCL(GAT)	0.850	0.724
KCL(GCN)	0.852	0.718
KCL(RGCN)	0.831	1.008
KCL(MPNN)	0.833	0.927
KCL(KMPNN)	0.852	0.765

Table 4: Ablation results under the fine-tune protocol. Each value represents the average result of the task, and the underline marks the best in the group.

Experiments

	KCL(GCN)			
Dataset	w/oALL	w/oInit	w/oNS	ALL
BBBP	0.939	0.941	0.953	**0.956**
Tox21	0.846	0.853	0.852	**0.856**
ToxCast	0.750	0.751	0.753	**0.757**
SIDER	0.650	0.663	0.665	**0.666**
CliTox	0.939	0.942	0.940	**0.945**
BACE	0.906	0.909	0.911	**0.934**
ESOL	0.584	0.589	0.599	**0.582**
FreeSolv	**0.833**	0.911	0.896	0.854
Ave(Cls)	0.840	0.843	0.845	**0.852**
Ave(Reg)	**0.709**	0.750	0.748	0.718

Table 9: Ablation results on molecular graphs.

	KCL(KMPNN)			
Dataset	w/oA	w/oInit	w/oNS	ALL
BBBP	0.943	0.945	0.956	**0.961**
Tox21	0.840	0.853	0.856	**0.859**
ToxCast	0.737	0.735	0.739	**0.740**
SIDER	0.650	0.659	0.661	**0.671**
CliTox	0.951	0.956	0.952	**0.958**
BACE	0.919	0.921	0.920	**0.924**
ESOL	0.740	0.735	0.739	**0.732**
FreeSolv	0.820	0.798	0.800	**0.795**
Ave(Cls)	0.840	0.845	0.847	**0.852**
Ave(Reg)	0.780	0.765	0.770	**0.764**

Table 10: Ablation results on the augmented molecular graphs.

Experiments

Task	Classification						Regression	
Dataset	BBBP	Tox21	ToxCast	SIDER	ClinTox	BACE	ESOL	FreeSolv
KCL(GIN)	0.954	0.854	0.748	0.660	**0.945**	0.932	0.580	0.856
KCL(GAT)	0.956	**0.857**	0.750	0.663	0.942	0.930	0.588	0.860
KCL(GCN)	**0.956**	0.856	**0.757**	**0.666**	**0.945**	**0.934**	**0.582**	**0.854**
KCL(R-GCN)	0.936	0.830	0.735	0.637	0.948	0.898	0.780	1.236
KCL(MPNN)	0.940	0.835	0.738	0.640	0.950	0.895	0.743	1.111
KCL(KMPNN)	**0.961**	**0.859**	**0.740**	**0.671**	**0.958**	**0.924**	**0.732**	**0.795**

Table 11: Results comparison with different graph encoders.

	Fine-tune Protocol			Linear Protocol		
	KCL	KMPNN	Abs.Imp.	KCL	KMPNN	Abs.Imp.
BBBP	**0.961**	0.915	+0.046	**0.927**	0.915	+0.012
Tox21	**0.859**	0.804	+0.055	**0.825**	0.804	+0.021
ToxCast	**0.740**	0.725	+0.015	0.709	**0.725**	-0.016
SIDER	**0.671**	0.645	+0.026	**0.659**	0.645	+0.014
ClinTox	**0.958**	0.892	+0.066	**0.898**	0.892	+0.006
BACE	**0.924**	0.856	+0.068	**0.860**	0.856	+0.004
ESOL	**0.736**	0.895	+0.159	**0.736**	0.895	+0.159
FreeSolv	**0.795**	2.167	+1.372	**0.795**	2.167	+1.372
Ave(Cls)	**0.852**	0.806	+0.046	**0.813**	0.806	+0.007
Ave(Reg)	**0.765**	1.531	+0.766	**0.766**	1.531	+ 0.765

Table 12: Results comparison between KCL(KMPNN) and KMPNN without contrastive learning.

Experiments

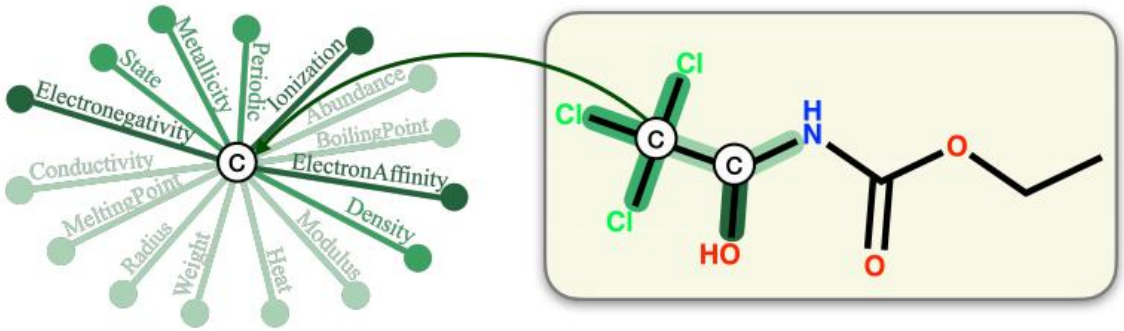


Figure 6: Another example in the BBBP dataset.

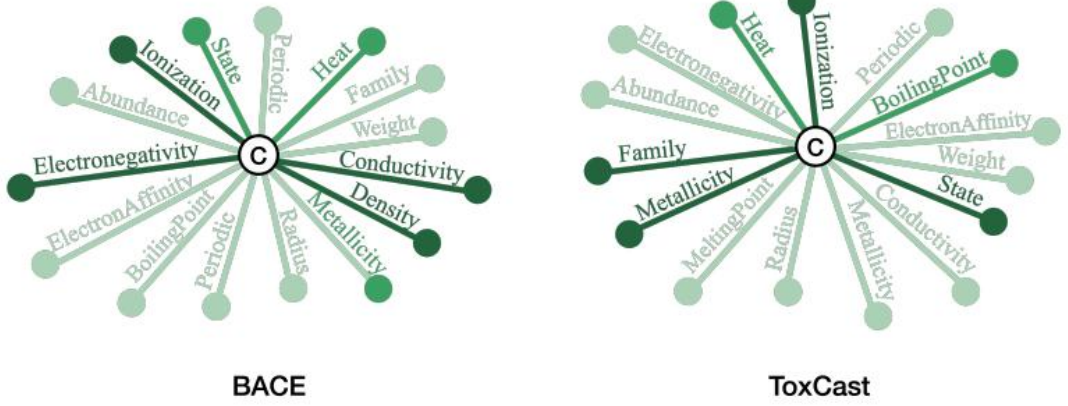


Figure 7: Attention visualization examples of attributes in the BACE and ToxCast datasets.

Thank you !