nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00654-0

Knowledge graph-enhanced molecular contrastive learning with functional prompt

In the format provided by the authors and unedited

Supplementary Information

Related work

Knowledge graph

A Knowledge Graph (KG) is a powerful tool for representing graph-structured knowledge that includes multiple entities and semantic relations. It has been used extensively in research and industry, particularly in association with search engines, recommendation systems, bioinformatics, large-scale data analysis, and natural language understanding [1].

From the perspective of the Semantic Web, a KG is typically composed of statements in the form of Resource Description Framework (RDF) triples [2, 3]. Each RDF triple is denoted as (s, p, o), where s represents a subject entity, p represents the predicate, and o can either be an entity or a datatype value [4]. If o is an entity, then this RDF triple is a relational fact, and p represents the relation between these two entities (also known as an object property). On the other hand, p represents a data property if o is a literal with some data type, such as a string, date, or number. In this case, the triple represents a literal, such as an entity attribute, textual definition, or comment.

KGs are often accompanied by an ontology, referred to as a schema, that uses languages from the Semantic Web community such as RDF Schema (RDFS) and Web Ontology Language (OWL) for richer semantics and higher quality [5, 6]. They typically define hierarchical classes, properties (i.e., the terms used as relations), concept and relation hierarchies, constraints (e.g., relation domain and range, and class disjointness), and logical expressions (e.g., relation composition). These languages have defined a number of built-in vocabularies for representing this knowledge, such as *rdf:subClassOf*, *rdf:type*, and *owl:disjointWith*. An ontology can be represented as RDF triples using these vocabularies. For example, containment between two classes can be expressed with *rdfs:subClassOf*, while *rdf:type* is used to describe the membership between an instance and a class. KGs equipped with schemas and constraints can support symbolic reasoning, such as consistency checking for change [7], which can find logical violations and entailment reasoning that infers hidden knowledge according to defined logics.

Previous attempts have built KGs from public chemical databases and scientific literature to extract associations between chemicals and diseases or drug pairs [8, 9]. In recent research, KCL [10] constructed a KG to describe the relations between elements and their corresponding chemical attributes. However, due to the simple relations within that KG, it cannot accommodate comprehensive and well-organized basic chemical knowledge, which hinders the model's exploration of knowledge to a certain extent.

Graph-based molecular representation learning

With the rise of deep learning, conventional supervised approaches have proven effective for molecular property prediction tasks. However, early works that utilized recurrent neural networks (RNNs) to encode SMILES strings into molecular representations suffered from a loss of topological information, as similar molecules could be encoded into distinct SMILES strings. To address this, recent attempts have turned to graph neural networks (GNNs) applied to molecular graphs to better capture the structure of molecules [11, 12, 13, 14, 15, 16, 17].

Duvenaud et al.[18] were the first to apply convolutional networks to encode molecular graphs into neural fingerprints. This approach has since been extended in various ways, with most methods focusing on atom-based message passing[19, 20, 21]. To better incorporate bond information, Coley et al.[11] concatenated features of an atom and its neighboring bonds to create atom-bond feature vectors. MPNN[12] was proposed to utilize features of both atoms and bonds. While some studies have attempted to leverage edge information using modules like the edge attention mechanism [13] and edge memory module [14], these models still primarily build on node-based MPNNs. Recently, DMPNN [15], CMPNN [16], and CoMPT [17] have been introduced as alternatives that enhance message interactions between bonds and atoms during message passing. However, these methods face challenges such as scarcity of labeled

data, as well as difficulties in generalizing to cases not seen during the training phase [22, 23, 24], greatly hindering their practical applicability.

Self-supervised learning for molecular graphs

To address these issues, researchers have turned to self-supervised learning (SSL), where a model is pre-trained on large, unlabeled molecular datasets and then transferred to downstream tasks with limited labels [25, 22]. SSL methods can be classified into two categories based on their design patterns: predictive methods and contrastive methods [26].

Predictive methods construct prediction tasks by exploiting the intrinsic properties of data. N-GRAM [25] applies the idea of N-gram in NLP to molecular graphs by predicting the atoms' attributes to create atoms embeddings. Hu et al. [22] predict contextual graph structures and masked node/edge attributes based on a masked molecule to capture atom features by learning the regularities of attributes distributed. Instead of predicting the atom type, GROVER [23] randomly masks a local subgraph of the target node/edge and predicts this contextual property from node embeddings. MGSSL [27] proposes a motif generation task, which aims to exploit the data distribution of graph motifs (significant subgraph patterns that occur frequently and contain semantic information). GEM [28] utilizes the molecular geometry described by bond angles and bond length in pretext tasks.

Contrastive methods construct pairwise augmented data, which aims to force positive pairs to be closer and push negative pairs apart. One key component is to generate informative and diverse views from each data instance. MoCL [24] proposes a bioisosteres substitution method, which produces a new molecule with similar physical or chemical properties as the original one. GeoGCL [29] captures geometry-aware structural information by contrasting the 2D-3D geometric views. MolCLR [30] introduces molecule graph augmentation strategies to generate contrastive pairs, namely atom masking, bond deletion, and subgraph removal. GraphMVP [31] performs SSL by leveraging the correspondence and consistency between 2D topological structures and 3D geometric views.

Prompts for pre-trained models

Prompt Learning has emerged as a new paradigm in NLP, achieving outstanding performance on various tasks [32]. In most SSL models, the pre-training phase uses pretext tasks that differ from the downstream task objectives. To bridge this gap, prompts are introduced to stimulate pre-trained model knowledge. Originally, prompts were manually designed discrete words [33, 34], requiring domain expertise and validation [35]. To facilitate labor-intensive prompt engineering, automatic searches for discrete prompts have been explored [36, 37], although sub-optimal. Recent works have eliminated discrete spaces by using continuous prompts, appending learnable continuous embeddings as prompt templates to input sequences [38, 39, 40].

Details of KANO

Statistics of ElementKG

The statistics for ElementKG are presented in Supplementary Table 1, which includes the counts and examples of axioms, classes, object properties, data properties, and entities.

Input features for nodes and edges in original and augmented molecular graphs

We use RDKit to extract atom and bond features as the input of the graph encoder. Supplementary Table 2 shows the specific chemical knowledge that these features take into account.

Augmented molecular graphs, unlike original molecular graphs, contain not only atoms and bonds, but also element entities and their relations. To illustrate, consider the augmented molecular graph shown

Ontology Metrics	Counts	Examples
Axiom	52140	*rdfs:subClassOf*, *owl:disjointWith*
Class	107	
Class (element)	27	*Metals*, *AlkaliMetals*, *ReactiveNonmetals*
Class (functional group)	80	*Ester*, *CarboxylicAcid*, *Cyanates*
Object property	94	*inRadiusGroup1*, *hasStateGas*, *isPartOf*
Data property	20	*hasAtomic*, *hasHeat*, *hasBondType*
Entity	201	
Entity (element)	118	*C*, *N*, *O*
Entity (functional group)	83	*Carboxyl*, *Cyanate*, *Borono*

Supplementary Table 1: The statistics of ElementKG.

Feature type	Features	Description	Size
atom	atom type	type of atom (e.g., C, N, O), by atomic number	110
	degree	number of bonds the atom is involved in	6
	formal charge	integer electronic charge assigned to atom	5
	chirality tag	unspecified, tetrahedral CW/CCW, or other	5
	number of hydrogens	number of bonded hydrogen atoms	5
	hybridization	sp, sp^2, sp^3, sp^3d, or sp^3d^2	5
	aromaticity	whether the atom is part of an aromatic system	1
	atomic mass	mass of the atom, divided by 100	1
bond	bond type	single, double, triple or aromatic	4
	conjugated	whether the bond is conjugated	1
	in ring	whether the bond is part of a ring	1
	stereo	none, any, E/Z or cis/trans	6

Supplementary Table 2: Feature extraction for atoms and bonds.

in Figure 1(b). We identify the element types present in the graph, retrieve the corresponding element entities and their relations from ElementKG, and use these to construct an element relation subgraph consisting of element entities as nodes and relations as edges. We use the embeddings of the corresponding element entities from ElementKG as input features for the element nodes. To represent the multiple relations between two element entities, we apply only one edge and implement mean pooling on the embeddings of the relations, using the resulting representation as input for this edge. Next, we link the element nodes to their corresponding atom nodes to connect the element relation subgraph and the original molecular graph. The input features of atoms and bonds in the original molecular graph follow Supplementary Table 2. The input features of edges connecting element nodes and corresponding atom nodes are randomly initialized with different embeddings according to the element type.

Details of molecular datasets

Supplementary Table 3 provides an overview of the benchmark datasets used in our work, including information on task type, data size, split type, and evaluation metrics. These datasets, taken from MoleculeNet [41], cover a diverse range of molecular properties, including physiology (BBBP, Tox21, ToxCast, SIDER, ClinTox), biophysics (BACE, MUV, HIV), physical chemistry (ESOL, FreeSolv, Lipophilicity), and quantum mechanics (QM7, QM8, QM9). Physiology focuses on macroscopic life systems, while biophysics uses physical methods to study biological phenomena. Physical chemistry analyzes the principles and laws of the chemical behavior of material systems from the perspective of physics, and quantum mechanics de-

scribes the most microscopic electronic properties.

Task Type	Metric	Category	Dataset	# Tasks	# Compounds	Split
Classification	ROC-AUC	Physiology	BBBP	1	2,039	scaffold split
			Tox21	12	7,831	scaffold split
			ToxCast	617	8,575	scaffold split
			SIDER	27	1,427	scaffold split
			ClinTox	2	1,478	scaffold split
		Biophysics	BACE	1	1,513	scaffold split
			MUV	17	93087	scaffold split
			HIV	1	41127	scaffold split
Regression	RMSE	Physical chemistry	ESOL	1	1,128	scaffold split
			FreeSolv	1	642	scaffold split
			Lipophilicity	1	4,200	scaffold split
	MAE	Quantum mechanics	QM7	1	6,830	scaffold split
			QM8	12	21,786	scaffold split
			QM9	3	133,885	random split

Supplementary Table 3: Summary of all the benchmarks for molecular property predictions used in this work. Our work utilizes 8 graph classification datasets and 6 graph regression datasets, with the number of binary prediction tasks for each dataset indicated in #Tasks.

Molecular classification datasets:

- *BBBP* [42] contains records of whether a compound carries the permeability property of penetrating the blood-brain barrier.
- *Tox21* [43] is a public database measuring the toxicity of compounds, which has been used in the 2014 Tox21 Data Challenge.
- *ToxCast* [44] contains multiple toxicity labels over thousands of compounds by running high-throughput screening tests on thousands of chemicals.
- *SIDER* [45] records marketed drugs along with their adverse drug reactions, also known as the Side Effect Resource.
- *ClinTox* [46] compares drugs approved through FDA and drugs eliminated due to toxicity during clinical trials.
- *BACE* [47] is collected for recording compounds that could act as the inhibitors of human β-secretase 1 (BACE-1) in the past few years.
- *MUV* [48] is a subset of PubChem BioAssay by applying a refined nearest neighbor analysis, designed for validation of virtual screening techniques.
- *HIV* [49] contains experimentally measured abilities of over 40,000 molecules to inhibit HIV replication.

Molecular regression datasets:

- *ESOL* [50] is a small dataset documenting the solubility of compounds .
- *FreeSolv* [51] is selected from the Free Solvation Database, which contains the hydration free energy of small molecules in water from both experiments and alchemical free energy calculations.

- *Lipophilicity* [52] is collected from the ChEMBL database [52], containing the experimental results of the octanol or water partition coefficient, which reflects the solubility of the molecules.
- *QM7* [53] is a subset of GDB-13, which provides information about the spatial structures of the molecules. It records various electronic properties that are stable and synthetically obtainable, such as HOMO and LUMO determined by *ab-initio* density function theory (DFT), and atomization energy.
- *QM8* [54] uses a variety of quantum mechanics methods to calculate the electronic spectrum and excited state energy of small molecules.
- *QM9*[1] [55] provides multiple data information on geometry, energy, electronic and thermodynamic properties of small molecules calculated by DFT.

As shown in Supplementary Table 3, we apply the scaffold splitting [56] on all benchmarks except QM9. Scaffold splitting splits the molecules with distinct two-dimensional structural frameworks into different subsets, providing a more challenging but practical setting since the test molecule can be structurally different from the training set. For QM9, random splitting was implemented following the settings in related works [25, 30]. We used RMSE and MAE as evaluation metrics for regression tasks.

Baselines

To demonstrate the effectiveness of KANO, we comprehensively evaluate its performance in comparison with supervised and self-supervised learning baselines.

Supervised learning baselines

We compare KANO with two widely used graph neural networks, GCN [57] and GIN [58], and two variants of message passing neural networks, MPNN [12], and DMPNN [15] and CMPNN [16], which are specifically designed for molecules.

Self-supervised learning baselines

For predictive-based self-supervised learning, we include the following pre-training models: N-GRAM [25], which assembles node embeddings in short walks and uses Random Forest or XGBoost to predict molecular properties; Hu et al.[22] and GROVER[23], which incorporate both node-level and graph-level knowledge in pretext tasks; MGSSL [27], which uses a motif generation pretext task; and GEM [28], which proposes several dedicated geometry-level self-supervised learning strategies to learn molecular geometry knowledge. For contrastive-based models, we include GraphMVP [31], which incorporates 3D geometric information into graph self-supervised learning, and MolCLR [30], which applies general graph augmentation methods to molecules. It is worth noting that to make the comparison fairer, we replaced the GCN and GIN backbones used in the original MolCLR paper with the CMPNN backbone. As a result, we included an additional baseline, denoted as $\text{MolCLR}_{\text{CMPNN}}$, for comprehensive comparison with our method.

Baseline experimental setup

In our study, we compared KANO with 12 baseline methods, including GCN [57], GIN [58], N-GRAM [25], Hu et al.[22], MGSSL [27], GraphMVP [31], MPNN [12], DMPNN [15], CMPNN [16], GEM [28], GROVER[23], and MolCLR [30]. The results of GCN, GIN, N-GRAM, and Hu et al. were taken from the paper of MolCLR, while the results of MGSSL and GraphMVP were obtained from the original text of these two articles. We reproduced the results of MPNN, DMPNN, CMPNN, GEM, GROVER, and MolCLR based on their respective

[1]Since the ranges of different targets in QM9 vary largely, we only take the “homo”, “lumo”, and “gap” targets to calculate the average MAE in our experiments.

GitHub code repositories. To ensure a fair comparison, we followed the experimental setup of previous works, which involved conducting experiments using three different splits of train/val/test data. This experimental setup was designed to evaluate the robustness of the models and minimize the impact of potential variations in model performance caused by different data splits.

Addtional experimental results

KANO achieves robust performance on identical data splits. Although we followed the same experimental settings as previous work, it is possible that different data splits may result in different performance outcomes. To ensure fairer comparison, we conducted additional experiments on three exact random splits for QM9 and three exact scaffold splits for all other downstream tasks, as detailed in Supplementary Table 4 and 5. The results on identical data splits show that KANO consistently achieves high performance, demonstrating its robustness and superiority over other methods.

Category	**Physiology**					**Biophysics**		
Dataset	BBBP	Tox21	ToxCast	SIDER	ClinTox	BACE	MUV	HIV
# Molecules	2039	7831	8575	1427	1478	1513	93807	41127
# Tasks	1	12	617	27	2	1	17	1
MPNN	86.5±2.6	82.5±2.6	67.2±2.3	57.4±2.0	85.6±1.1	82.8±4.0	77.7±4.3	79.4±4.0
DMPNN	85.3±3.5	81.9±2.2	67.3±2.1	58.5±1.7	86.6±2.9	82.9±2.8	80.9±2.8	80.8±1.7
CMPNN	91.3±4.0	81.7±2.7	69.2±0.6	64.1±0.8	89.6±2.1	91.0±3.2	80.6±2.4	81.8±1.6
GEM	88.1±7.8	80.9±0.6	71.0±0.2	61.4±0.9	86.5±1.3	89.8±4.0	76.2±5.1	80.5±3.7
GROVER	84.1±4.1	80.0±2.0	51.1±0.5	58.2±5.0	70.7±7.2	83.3±9.0	70.4±1.5	67.5±1.5
$MolCLR_{GCN}$	73.2±0.6	73.5±1.9	68.7±2.0	61.0±2.8	85.9±2.3	82.6±5.6	73.5±1.3	76.8±2.4
$MolCLR_{GIN}$	73.5±0.4	76.7±2.1	65.2±1.8	60.7±5.7	90.4±1.7	83.5±1.8	75.5±1.8	77.6±3.2
$MolCLR_{CMPNN}$	72.4±0.7	78.4±2.6	69.1±1.2	59.7±3.4	88.0±4.0	85.0±2.4	74.5±2.1	77.8±5.5
KANO	**96.0±1.6**	**83.7±1.3**	**73.2±1.6**	**65.2±0.8**	**94.4±0.3**	**93.1±2.1**	**83.7±2.3**	**85.1±2.2**

Supplementary Table 4: Test performance of different models on eight classification benchmarks of physiology and biophysics. The mean and standard deviation of test ROC-AUC (%) on the three exact *scaffold splits* are reported.
*The best performing results are marked in bold.

Category	**Physical chemistry**			**Quantum mechanics**		
Dataset	ESOL	FreeSolv	Lipophilicity	QM7	QM8	QM9
# Molecules	1128	642	4200	7160	21786	133885
# Tasks	1	1	1	1	12	3
MPNN	1.297±0.167	1.621±0.419	0.729±0.019	110.572±5.675	0.0149±0.0003	0.00553±0.00004
DMPNN	1.078±0.111	1.673±0.402	0.752±0.043	102.544±3.840	0.0154±0.0003	0.00540±0.00004
CMPNN	0.767±0.092	1.572±0.557	0.629±0.024	73.120±3.482	0.0152±0.0004	0.00365±0.00004
GEM	0.770±0.025	1.644±0.195	0.649±0.136	57.493±2.541	0.0166±0.0002	0.00521±0.00001
GROVER	1.475±0.440	3.235±1.394	0.974±0.097	94.291±7.869	0.0177±0.0022	0.00688±0.00062
$MolCLR_{GCN}$	1.102±0.021	2.241±0.259	0.819±0.003	90.216±3.486	0.0189±0.0002	0.00480±0.00002
$MolCLR_{GIN}$	1.091±0.035	2.017±0.154	0.824±0.002	87.125±5.187	0.0175±0.0004	0.00473±0.00002
$MolCLR_{CMPNN}$	0.911±0.082	2.021±0.133	0.875±0.003	89.823±6.334	0.0179±0.0009	0.00475±0.00001
KANO	**0.670±0.019**	**1.142±0.258**	**0.566±0.007**	**56.405±2.780**	**0.0123±0.0001**	**0.00320±0.00001**

Supplementary Table 5: Test performance of different models on six regression benchmarks of physical chemistry and quantum mechanics. The mean and standard deviation of test RMSE (for ESOL, FreeSolv, Lipophilicity) or MAE (for QM7, QM8, QM9) on the three exact *scaffold splits* (ESOL, FreeSolv, Lipophilicity, QM7, QM8) and *random splits* (QM9) are reported.
*The best performing results are marked in bold.

KANO demonstrates high performance on cluster splitting datasets. While scaffold splitting is a commonly used technique for dividing molecular datasets, some scaffolds can be similar yet unique. To ensure

that the train/validation/test sets have diverse structural differences, we propose cluster splitting. In this approach, molecules are clustered based on their molecular Morgan Fingerprints using the Butina algorithm [59], with a distance threshold of 0.4. The resulting groups are then balanced in size to avoid putting only the smallest clusters in the testing set.

Supplementary Table 6 and 7 present a comparison of the performance of KANO with other baseline methods on cluster splitting. The mean and standard deviation of the results were reported. Our results show that KANO consistently outperforms other baselines on cluster splitting datasets.

Category	**Physiology**					**Biophysics**		
Dataset	BBBP	Tox21	ToxCast	SIDER	ClinTox	BACE	MUV	HIV
# Molecules	2039	7831	8575	1427	1478	1513	93807	41127
# Tasks	1	12	617	27	2	1	17	1
MPNN	89.1±4.7	83.9±1.5	69.8±1.3	62.5±2.2	92.3±1.9	77.2±6.7	82.3±2.0	80.6±0.3
DMPNN	87.5±6.2	84.1±1.4	69.9±1.4	62.7±1.6	91.2±2.8	80.3±5.3	79.0±0.9	81.4±1.4
CMPNN	92.2±5.8	83.4±1.2	71.7±2.0	62.8±1.2	89.8±3.8	85.5±3.2	82.6±0.4	79.1±1.4
GEM	86.3±10.1	83.4±0.6	72.3±0.6	61.6±2.8	89.4±4.1	82.3±8.0	78.7±0.9	80.7±3.5
GROVER	88.7±3.6	80.8±2.4	59.2±3.7	60.1±2.3	79.6±1.3	79.3±8.3	72.2±0.5	68.6±4.6
$MolCLR_{GCN}$	75.1±0.9	80.9±1.3	70.1±1.5	60.2±2.1	84.1±5.6	81.5±3.1	72.1±1.6	72.7±1.9
$MolCLR_{GIN}$	74.6±1.7	81.5±0.8	69.6±0.8	62.1±1.2	82.5±5.7	79.6±2.7	73.2±0.8	76.1±2.1
$MolCLR_{MPNN}$	73.9±2.1	80.2±1.0	71.0±1.1	59.6±1.0	83.3±6.7	80.4±5.9	71.3±1.4	74.5±1.7
KANO	**96.6±2.2**	**85.8±0.9**	**74.7±1.5**	**64.5±1.4**	**93.8±1.8**	**87.9±1.9**	**83.7±2.3**	**82.1±0.5**

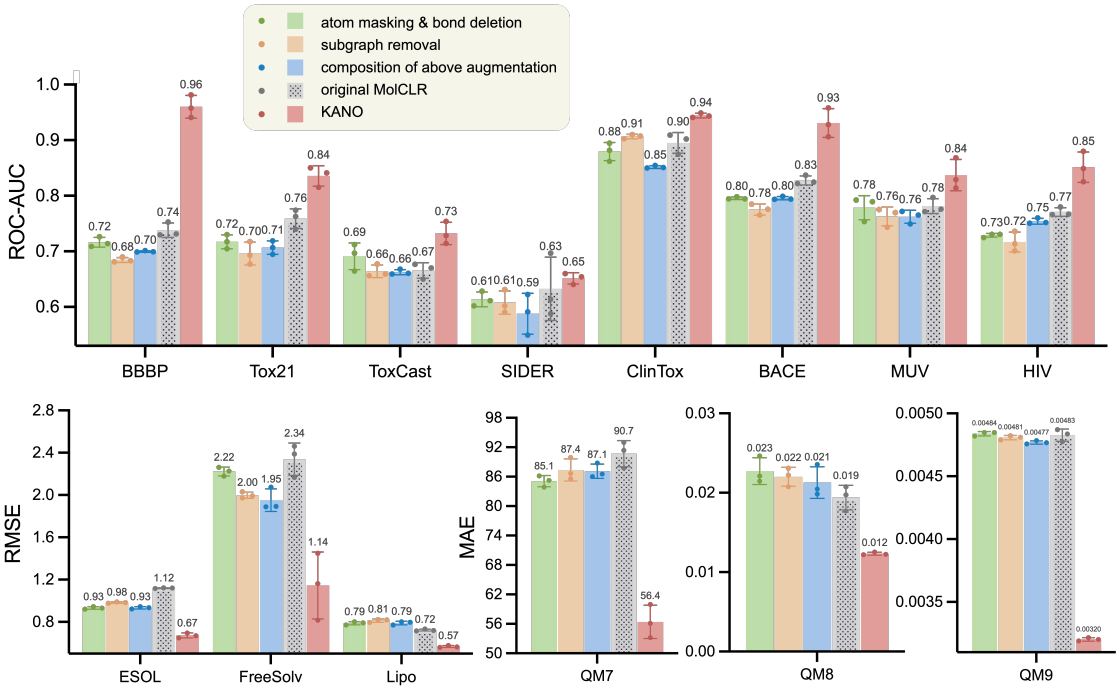
Supplementary Table 6: Test performance of different models on eight classification benchmarks of physiology and biophysics. The mean and standard deviation of test ROC-AUC (%) on the three exact *cluster splits* are reported.
*The best performing results are marked in bold.

Category	**Physical chemistry**			**Quantum mechanics**	
Dataset	ESOL	FreeSolv	Lipophilicity	QM7	QM8
# Molecules	1128	642	4200	7160	21786
# Tasks	1	1	1	1	12
MPNN	0.938±0.085	1.272±0.085	0.728±0.024	82.772±2.225	0.0143±0.0005
DMPNN	0.877±0.082	1.306±0.041	0.712±0.015	77.260±1.055	0.0147±0.0004
CMPNN	0.655±0.049	0.976±0.145	0.629±0.013	60.799±1.328	0.0115±0.0002
GEM	0.659±0.028	0.887±0.196	0.610±0.060	56.465±2.478	0.0126±0.0002
GROVER	0.947±0.204	1.842±0.296	0.841±0.029	112.918±3.121	0.0181±0.0046
$MolCLR_{GCN}$	0.733±0.102	1.150±0.069	0.838±0.014	73.128±1.592	0.0140±0.0002
$MolCLR_{GIN}$	0.706±0.045	1.148±0.076	0.832±0.021	72.154±2.155	0.0144±0.0003
$MolCLR_{MPNN}$	0.719±.069	1.152±0.093	0.843±0.013	68.251±1.645	0.0135±0.0002
KANO	**0.588±0.036**	**0.767±0.139**	**0.589±0.011**	**54.413±0.639**	**0.0099±0.0002**

Supplementary Table 7: Test performance of different models on five regression benchmarks of physical chemistry and quantum mechanics. The mean and standard deviation of test RMSE (for ESOL, FreeSolv, Lipophilicity) or MAE (for QM7, QM8, QM9) on the three exact *cluster splits* are reported.
*The best performing results are marked in bold.

KG-guided contrastive learning outperforms universal graph augmentation strategies. To investigate the impact of various graph augmentation methods on model performance, we compare our KG-guided contrastive learning approach to the MolCLR-type graph augmentation strategy with CMPNN as the backbone ($MolCLR_{CMPNN}$), which serves as an important baseline. Supplementary Fig. 1 provides an intuitive comparison of the performance of different architectures on 14 downstream tasks. Our results show that KG-guided contrastive learning outperforms MolCLR-type augmentations with the CMPNN backbone, indi-

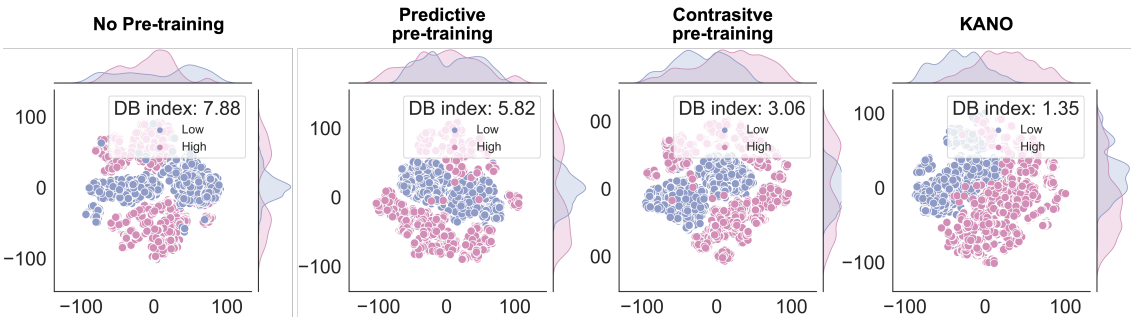
cating that our approach is superior to universal graph augmentation strategies in improving model performance. Furthermore, we confirm that CMPNN is not the primary factor contributing to the improvement in model performance, which further supports the effectiveness of our KG-guided approach.



Supplementary Fig. 1: Comparison of model performance with different graph augmentation strategies. The first three bars represent the performance of MolCLR-type graph augmentation strategy with CMPNN as the backbone. The green, yellow, and blue bars correspond to the three sub-strategies of MolCLR-type augmentation, including atom masking and bond deletion, subgraph removal, and the combination of the two, respectively. The gray bar indicates the best performance achieved by the original MolCLR, and the red bar shows the results of KANO. The results are reported as mean values +/- SD on three independent runs. The error bars represent the SD, while the dots represent three individual data points.

KG-guided contrastive learning produces semantically meaningful representations. To intuitively observe whether the molecular representations derived from our pre-trained model hint at chemistry knowledge, we analyze the representations learned by pre-trained KANO by mapping them to the two-dimensional space with t-SNE algorithm [60]. We select 9,124 molecules from the QM9 dataset and divided them into two groups based on their Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO) obtained from density functional theory (DFT): 5,036 molecules with high HOMO-LUMO gaps and 4,088 molecules with low HOMO-LUMO gaps. The HOMO and LUMO energies play a key role in determining the molecular properties, such as redox ability, optical properties, and chemical reactivity, and the HOMO-LUMO gap represents the energy difference between the HOMO and LUMO, which is an indicator of molecular stability.

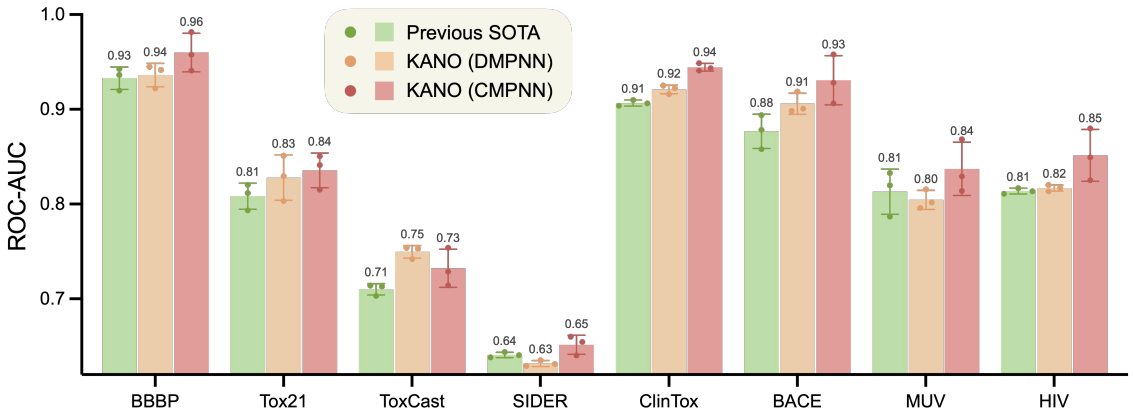
We evaluate the ability of different models to distinguish between these two groups of molecules, as shown in Supplementary Figure 2. The models include a graph encoder trained from scratch ("No pre-training"), the representative predictive-based pre-trained model GROVER [23] ("Predictive pre-training"), and contrastive-based pre-trained model with universal graph augmentation strategy $\text{MolCLR}_{\text{CMPNN}}$ ("Contrastive pre-training") [30]. We calculated the Davids Bouldin (DB) index to measure the separation of the clusters, with a lower DB index indicating better cluster separation. Visually, all three models show a separation trend of the two clusters, but our model outperforms the others, as supported by the lower DB index. These results confirm that KG-fused KANO effectively incorporates basic element knowledge



Supplementary Fig. 2: A molecular representation visualization of different methods. A molecular representation visualization comparing different methods. The high cluster contains molecules with high HOMO-LUMO gaps (pink), whereas the low cluster contains those with low HOMO-LUMO gaps (blue). A lower DB index means that the clustering has a more appropriate separation.

into the contrastive learning framework, producing semantically meaningful molecular representations that outperform predictive pre-training models in terms of computational cost. Furthermore, our model maintains semantic coherence within molecules, resulting in better performance than the contrastive pre-training model.

Expanding the possibilities of KANO with pluggable graph encoders. Our graph encoder is pluggable, meaning it can be replaced with any model that considers edge information during message passing. To investigate the impact of different message passing architectures, we replaced the original CMPNN backbone [16] with DMPNN [12]. Supplementary Figure 3 shows the test ROC-AUC (averaged over three runs) for previous SOTA, KANO with CMPNN, and KANO with DMPNN across eight molecular property prediction classification benchmarks. KANO with DMPNN achieves competitive performance in most cases, and outperforms previous SOTA in many cases. This suggests that different graph encoder architectures are not a critical factor affecting model performance, as long as edge information is reasonably considered in the message passing process.



Supplementary Fig. 3: Performance of KANO with different message passing architectures for classification tasks. The green, yellow, and red bars represent the performance of the previous SOTA model, KANO with DMPNN as the graph encoder, and KANO with CMPNN as the graph encoder, respectively. The results are reported as mean values +/- SD on three independent runs. The error bars represent the SD, while the dots represent three individual data points.

References

[1] Hogan, A. *et al. Knowledge Graphs*. *ACM Computing Surveys (CSUR)* **54**, 1–37 (Morgan & Claypool Publishers, 2021).

[2] Domingue, J. & Fensel, D. *Handbook of Semantic Web Technologies* (Springer, 2011).

[3] Pan, J. Z. Resource description framework. *Handbook on Ontologies* 71–90 (Springer, 2009).

[4] Chen, J. *et al.* Low-resource learning with knowledge graphs: A comprehensive survey. Preprint at https://arxiv.org/abs/2112.10006 (2021).

[5] Horrocks, I. Ontologies and the semantic web. *Commun. ACM* **51**, 58–67 (2008).

[6] Baader, F., Horrocks, I. & Sattler, U. Description logics as ontology languages for the semantic web. *Mechanizing Mathematical Reasoning, Essays in Honor of Jörg H. Siekmann on the Occasion of His 60th Birthday* 228–248 (Springer, 2005).

[7] Flouris, G., Huang, Z., Pan, J. Z., Plexousakis, D. & Wache, H. Inconsistencies, negations and changes in ontologies. In *AAAI Conference on Artificial Intelligence* **21**, 1295–1300 (AAAI Press, 2006).

[8] Delmas, M. *et al.* Building a knowledge graph from public databases and scientific literature to extract associations between chemicals and diseases. *Bioinformatics* **37**, 3896–3904 (2021).

[9] Lin, X., Quan, Z., Wang, Z., Ma, T. & Zeng, X. KGNN: knowledge graph neural network for drug-drug interaction prediction. In *International Joint Conference on Artificial Intelligence* 2739–2745 (ijcai.org, 2020).

[10] Fang, Y. *et al.* Molecular contrastive learning with chemical element knowledge graph. In *AAAI Conference on Artificial Intelligence* **36**, 3968–3976 (AAAI Press, 2022).

[11] Coley, C. W., Barzilay, R., Green, W. H., Jaakkola, T. S. & Jensen, K. F. Convolutional embedding of attributed molecular graphs for physical property prediction. *Journal of chemical information and modeling* **57**, 1757–1772 (2017).

[12] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In *34th International Conference on Machine Learning*, **70**, 1263–1272 (PMLR, 2017).

[13] Shang, C. *et al.* Edge attention-based multi-relational graph convolutional networks. Preprint at https://arxiv.org/abs/1802.04944 (2018).

[14] Withnall, M., Lindelöf, E., Engkvist, O. & Chen, H. Building attention and edge message passing neural networks for bioactivity and physical-chemical property prediction. *J. Cheminformatics* **12**, 1 (2020).

[15] Yang, K. *et al.* Are learned molecular representations ready for prime time? Preprint at http://arxiv.org/abs/1904.01561 (2019).

[16] Song, Y. *et al.* Communicative representation learning on attributed molecular graphs. In *International Joint Conference on Artificial Intelligence*, 2831–2838 (ijcai.org, 2020).

[17] Chen, J., Zheng, S., Song, Y., Rao, J. & Yang, Y. Learning attributed graph representations with communicative message passing transformer. Preprint at https://arxiv.org/abs/2107.08773 (2021).

[18] Duvenaud, D. *et al.* Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems*, 2224–2232 (2015).

[19] Kearnes, S., McCloskey, K., Berndl, M., Pande, V. S. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.* **30**, 595–608 (2016).

[20] Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nature communications* **8**, 1–8 (2017).

[21] Schütt, K. *et al.* Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, 991–1001 (2017).

[22] Hu, W. *et al.* Strategies for pre-training graph neural networks. In *8th International Conference on Learning Representations* (OpenReview.net, 2020).

[23] Rong, Y. *et al.* Self-supervised graph transformer on large-scale molecular data. In *Advances in Neural Information Processing Systems* **33**, 12559–12571 (2020).

[24] Sun, M., Xing, J., Wang, H., Chen, B. & Zhou, J. Mocl: Contrastive learning on molecular graphs with multi-level domain knowledge. In *The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 3585–3594 (2021).

[25] Liu, S., Demirel, M. F. & Liang, Y. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. In *Advances in Neural Information Processing Systems* **32** (2019).

[26] Fang, Y., Zhang, Q., Chen, Z., Fan, X. & Chen, H. Knowledge-informed molecular learning: A survey on paradigm transfer. Preprint at https://arxiv.org/abs/2202.10587 (2022).

[27] Zhang, Z., Liu, Q., Wang, H., Lu, C. & Lee, C. Motif-based graph self-supervised learning for molecular property prediction. In *Advances in Neural Information Processing Systems* **34**, 15870–15882 (2021).

[28] Fang, X. *et al.* Geometry-enhanced molecular representation learning for property prediction. *Nat. Mach. Intell.* **4**, 127–134 (2022).

[29] Li, S., Zhou, J., Xu, T., Dou, D. & Xiong, H. Geomgcl: Geometric graph contrastive learning for molecular property prediction. In *AAAI Conference on Artificial Intelligence* **36**, 4541–4549 (AAAI Press, 2022).

[30] Wang, Y., Wang, J., Cao, Z. & Farimani, A. B. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* **4**, 279–287 (2022).

[31] Liu, S. *et al.* Pre-training molecular graph representation with 3d geometry. In *Tenth International Conference on Learning Representations* (OpenReview.net, 2022).

[32] Gu, Y., Han, X., Liu, Z. & Huang, M. PPT: pre-trained prompt tuning for few-shot learning. In *Association for Computational Linguistics*, 8410–8423 (Association for Computational Linguistics, 2022).

[33] Schick, T. & Schütze, H. It's not just size that matters: Small language models are also few-shot learners. In *North American Chapter of the Association for Computational Linguistics*, 2339–2352 (Association for Computational Linguistics, 2021).

[34] Schick, T. & Schütze, H. Exploiting cloze-questions for few-shot text classification and natural language inference. In *European Chapter of the Association for Computational Linguistics*, 255–269 (Association for Computational Linguistics, 2021).

[35] Perez, E., Kiela, D. & Cho, K. True few-shot learning with language models. In *Advances in Neural Information Processing Systems* **34**, 11054–11070 (2021).

[36] Gao, T., Fisch, A. & Chen, D. Making pre-trained language models better few-shot learners. In *Association for Computational Linguistics*, 3816–3830 (Association for Computational Linguistics, 2021).

[37] Shin, T., Razeghi, Y., IV, R. L. L., Wallace, E. & Singh, S. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Empirical Methods in Natural Language Processing*, 4222–4235 (Association for Computational Linguistics, 2020).

[38] Hambardzumyan, K., Khachatrian, H. & May, J. WARP: word-level adversarial reprogramming. In *Association for Computational Linguistics*, 4921–4933 (Association for Computational Linguistics, 2021).

[39] Li, X. L. & Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. In *Association for Computational Linguistics*, 4582–4597 (Association for Computational Linguistics, 2021).

[40] Liu, X. *et al.* GPT understands, too. Preprint at https://arxiv.org/abs/2103.10385 (2021).

[41] Wu, Z. *et al.* Moleculenet: A benchmark for molecular machine learning. *Chemical science* **9**, 513-530 (2018).

[42] Martins, I. F., Teixeira, A. L., Pinheiro, L. & Falcão, A. O. A bayesian approach to *in Silico* blood-brain barrier penetration modeling. *J. Chem. Inf. Model.* **52**, 1686–1697 (2012).

[43] Hartung, T. Toxicology for the twenty-first century. *Nature* **460**, 208–212 (2009).

[44] Richard, A. M. *et al.* Toxcast chemical landscape: paving the road to 21st century toxicology. *Chemical research in toxicology* **29**, 1225–1251 (2016).

[45] Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **44**, 1075–1079 (2016).

[46] Gayvert, K. M., Madhukar, N. S. & Elemento, O. A data-driven approach to predicting successes and failures of clinical trials. *Cell chemical biology* **23**, 1294–1301 (2016).

[47] Subramanian, G., Ramsundar, B., Pande, V. & Denny, R. A. Computational modeling of β-secretase 1 (bace-1) inhibitors using ligand based approaches. *Journal of chemical information and modeling* **56**, 1936–1949 (2016).

[48] Rohrer, S. G. & Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on pubchem bioactivity data. *J. Chem. Inf. Model.* **49**, 169–184 (2009). https://doi.org/10.1021/ci8002649.

[49] IAM Graph Database Repository for Graph Based Pattern Recognition and Machine Learning. In *SSPR/SPR* **5342**, 287–297 (Springer, 2008).

[50] Delaney, J. S. Esol: estimating aqueous solubility directly from molecular structure. *Journal of chemical information and computer sciences* **44**, 1000–1005 (2004).

[51] Mobley, D. L. & Guthrie, J. P. Freesolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design* **28**, 711–720 (2014).

[52] Gaulton, A. *et al.* Chembl: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, 1100–1107 (2012). https://doi.org/10.1093/nar/gkr777.

[53] Blum, L. C. & Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database gdb-13. *Journal of the American Chemical Society* **131**, 8732–8733 (2009).

[54] Ramakrishnan, R., Hartmann, M., Tapavicza, E. & Von Lilienfeld, O. A. Electronic spectra from tddft and machine learning in chemical space. *The Journal of chemical physics* **143**, 084111 (2015).

[55] Ruddigkeit, L., van Deursen, R., Blum, L. C. & Reymond, J. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **52**, 2864–2875 (2012). https://doi.org/10.1021/ci300415d.

[56] Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry* **39**, 2887–2893 (1996).

[57] Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations* (OpenReview.net, 2017).

[58] Xu, K., Hu, W., Leskovec, J. & Jegelka, S. DBLP:conf/iclr/XuHLJ19? In *7th International Conference on Learning Representations* (OpenReview.net, 2019).

[59] Butina, D. Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci.* **39**, 747–750 (1999).

[60] van der Maaten, L. Accelerating t-sne using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).