molecules

MDPI

Article

SG-ATT: A Sequence Graph Cross-Attention Representation Architecture for Molecular Property Prediction

Yajie Hao [1], Xing Chen [1], Ailu Fei [1], Qifeng Jia [1], Yu Chen [1], Jinsong Shao [1], Sanjeevi Pandiyan [1] and Li Wang [1,2,*]

[1] School of Information Science and Technology, Nantong University, Nantong 226001, China; 2110310030@stmail.ntu.edu.cn (Y.H.); 2230310036@stmail.ntu.edu.cn (X.C.); 2230310037@stmail.ntu.edu.cn (A.F.); 2230310054@stmail.ntu.edu.cn (Q.J.); 2110310019@stmail.ntu.edu.cn (Y.C.); sylershao@gmail.com (J.S.); gpsanjeevi@gmail.com (S.P.)

[2] Research Center for Intelligent Information Technology, Nantong University, Nantong 226001, China

* Correspondence: wangli@ntu.edu.cn

Abstract: Existing formats based on the simplified molecular input line entry system (SMILES) encoding and molecular graph structure are designed to encode the complete semantic and structural information of molecules. However, the physicochemical properties of molecules are complex, and a single encoding of molecular features from SMILES sequences or molecular graph structures cannot adequately represent molecular information. Aiming to address this problem, this study proposes a sequence graph cross-attention (SG-ATT) representation architecture for a molecular property prediction model to efficiently use domain knowledge to enhance molecular graph feature encoding and combine the features of molecular SMILES sequences. The SG-ATT fuses the two-dimensional molecular features so that the current model input molecular information contains molecular structure information and semantic information. The SG-ATT was tested on nine molecular property prediction tasks. Among them, the biggest SG-ATT model performance improvement was 4.5% on the BACE dataset, and the average model performance improvement was 1.83% on the full dataset. Additionally, specific model interpretability studies were conducted to showcase the performance of the SG-ATT model on different datasets. In-depth analysis was provided through case studies of in vitro validation. Finally, network tools for molecular property prediction were developed for the use of researchers.

Keywords: molecular graph; knowledge graph; cross-attention; molecular property prediction

check for updates

Citation: Hao, Y.; Chen, X.; Fei, A.; Jia, Q.; Chen, Y.; Shao, J.; Pandiyan, S.; Wang, L. SG-ATT: A Sequence Graph Cross-Attention Representation Architecture for Molecular Property Prediction. *Molecules* **2024**, *29*, 492. https://doi.org/10.3390/molecules29020492

Academic Editor: Francisco Torrens

Received: 1 December 2023
Revised: 6 January 2024
Accepted: 14 January 2024
Published: 19 January 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Currently, well-designed molecular representations, such as molecular fingerprints or molecular descriptors, can achieve good results for specific tasks. However, most molecular characterizations based on feature engineering are optimized for specific tasks and lack generality [1–3]. With the development of deep neural networks, various molecular representation models have freed researchers from complex and time-consuming feature engineering [4,5]. Molecular characterization learning has also officially entered the era of deep learning, e.g., in compound property prediction, drug screening of small molecules, drug interactions, and more foreseeable applications [6–8].

Molecular property prediction methods based on molecular graphs and SMILES sequence-based molecular properties have rapidly developed [9–11]. To address the problem of inadequate representation of molecular features in a single dimension [12], Pang et al. proposed a multidimensional molecular feature encoding model in 2021. They used the molecular SMILES sequence encoding as input and obtained the 1D and 2D feature vectors of molecules using the transformer model and message-passing neural networks (MPNN) model, respectively [13–15]. In addition, basic chemical domain knowledge is an important prior knowledge in molecular information, and work has been implemented to utilize domain prior chemical knowledge for multimodal learning [16]. Combining the above ideas, Fang et al. proposed the knowledge-enhanced contrastive learning (KCL) framework to construct a chemical element knowledge graph (CKG) based on the periodic table of elements to expand the molecular graph by expanding relationships between elements and their basic chemical properties on the original graph. The expanded molecular graph contains the topology of the molecule and basic chemical domain knowledge of the element [17]. Hasebe's work proposes a knowledge-embedded message-passing neural network [18]. This network can combine molecular properties and knowledge annotations on molecular graphs as supervision to learn molecular properties and structures. Incorporating knowledge embedding into the model for multimodal learning of molecular attributes and manually annotated knowledge makes the graph neural networks (GNNs) more generalizable and physically consistent. The debut of ChatGPT has attracted great attention from the natural language processing (NLP) community [19]. Large language models have proved their ability to perform various NLP tasks. With the success of language processing models based on transformer architecture in many fields, related researchers have also attempted to introduce it and pre-train with fine-tuning models for molecular characterization [20]. The topological information of the molecule is expressed by aggregating information according to the presence or absence of edges between atoms when aggregating information after self-attention has been computed during pre-training.

In current molecular property prediction tasks, molecular representations are usually represented using only molecular graphs or SMILES sequence encoding. However, neither type of input can express sufficiently rich molecular semantic and structural information on its own. Because SMILES sequences alone encode a complex molecular spatial structure compressed into a single linear sequence, they cannot encode the spatial structure of the molecule itself. The molecular graph-based approach heavily depends on the amount of data, and it performs worse than the molecular descriptor-based approach when the dataset is small. In addition, GNNs are prone to over-smoothing problems, so the number of GNN layers is usually only two to four, which limits their feature extraction capability. In addition, the method overemphasizes the importance of GNNs for structure perception and ignores the generalization of the graph structure.

Obtaining models with sufficient prediction accuracy requires large datasets for pre-training. It is difficult to achieve training goals using small datasets. One solution is to use transfer learning, which can improve the prediction accuracy for tasks with small datasets [21]. However, to apply transfer learning, another dataset with molecular properties related to the target properties must be available. Therefore, to address the universality of prediction models across different datasets, multimodal learning using prior knowledge of physics and chemistry has improved the prediction model performance by incorporating knowledge into deep learning. For example, existing molecular representation models usually encode atoms in a molecular graph as individuals who can interact only in the presence of chemical bonds but do not consider the correlations between atoms, which qualitatively exist between unconnected atoms in the molecular graph. If a knowledge graph is constructed based on the relationships between elements using domain knowledge, then the original molecular graph can be enhanced through knowledge-guided expansion. This expansion helps establish connections between atoms that share common properties but are not directly connected via chemical bonds.

Combining these two points, the multidimensional feature encoding that combines 1D and 2D features complements the structural information of molecules implicitly in SMILES and reduces the requirement of relevant molecular representation models for GNN to extract implicit chemical information, which can further improve the accuracy of prediction tasks. In addition, if domain knowledge is applied to the prediction of molecular attributes, it is helpful for the model to capture the interaction between atoms and the characteristics of the molecular substructure.

To solve these problems, multidimensional molecular feature encoding is explored and combined with a molecular knowledge graph (MKG), and the SG-ATT model is proposed. Thus, the SG-ATT can learn molecular features from 1D semantic and 2D topological structure perspectives. In the process of learning molecular graph features, the proposed MKG incorporates features at both the atom and functional group levels, thereby enhancing the representation capability of molecular graphs. After feature encoding, a cross-attention method is used to extract molecular features adaptively based on different weighted molecular features, and finally, a composite molecular representation is obtained.

The main contributions of SG-ATT are as follows:

(1) Fusing 1D and 2D molecular features encoding to learn molecular representations from a multidimensional view.
(2) Considering the generalization ability of the model for small datasets, the molecular graph is enhanced using MKG guidance, which allows the model to learn richer molecular information.
(3) Adopting a cross-attention mechanism to improve the learning ability of the representation of molecular.

2. Results

2.1. Dataset and Setup

To benchmark the SG-ATT model performance, nine widely used molecular property-related datasets were selected from the MoleculeNet dataset, covering physical chemistry (ESOL, FreeSolv), Quantum mechanics (QM7), biophysics (BACE, HIV), and physiology (BBBP, Tox21, ToxCast, SIDER, and ClinTox) [22]. The number of tasks in these datasets is 1 to 617, and the size of the datasets is 642 to 41,127. Table A1 in Appendix C summarizes the details of the datasets.

2.2. Benchmarking Methods

The SG-ATT model was compared with the following baseline methods to test its performance on downstream tasks. The graph convolutional network (GCN) and model Weave are two types of graph convolutional methods [23,24]. The Message-Passing Neural Network (MPNN) considers the edge features and strengthens the message interactions between bonds and atoms during message passing [15]. Directed Message Passing Neural Networks (DMPNNs) and Communicative Message-Passing Neural Networks (CMPNNs) are improvements to MPNN [25,26]. The Communicative Message Passing Transformer (CoMPT) considers edge features and enhances the message interaction between bonds and atoms in the message-passing process [27]. Fusion network of molecular fingerprints and molecular graphs for molecular property prediction (FP-GNN) [28].

(1) GCN is a convolutional method that focuses on learning the relationship with the nearest neighbor node [23].
(2) Weave transformed feature vectors using pair features with distant atoms in addition to atom features focusing only on atoms [24]. MPNN utilizes features from nodes and edges and then summarizes them into a framework [15].
(3) DMPNN treated the molecular graph as an edge-oriented directed structure, avoiding information redundancy during iterations [25].
(4) CMPNN improves molecular embedding by enhancing message interaction between nodes and edges through communication kernel [26].
(5) CoMPT invokes a communicative message-passing paradigm based on Transformer [27].
(6) FP-GNN combines and simultaneously learns information from molecular graphs and fingerprints for molecular property prediction [28].

2.3. Experimental Setup

Using the Adam optimizer with an initial learning rate of 0.0001 and a batch size set to 64, these configurations are implemented through PyTorch (version 1.12.0) and the Deep Graph Library [29,30]. For the graph encoder, the dimension size of the feature vector is set to 128. In the sequence encoder, the number of attention heads in the transformer is set to 8, and the final feature length is set to 128. To prevent overfitting, an early stopping mechanism is employed to ensure SG-ATT runs for a maximum of 100 epochs.

Hyperparameter selection is performed using a random search approach. Additionally, for each experimental run, the average of results from 10 random seeds is taken. All code development is conducted on a GPU (NVIDIA GeForce 3080ti, NVIDIA, Santa Clara, CA, USA) utilizing an Ubuntu server.

2.4. Experimental Results and Analysis

In this section, we tested whether the proposed SG-ATT model approach outperformed the benchmark approach. Table 1 provides the following observations. Firstly, SG-ATT consistently achieved the best performance across all regression datasets, the best results on three classification datasets, and the second-best result on one dataset. The significant overall dataset performance demonstrates the effectiveness of SG-ATT in molecular property prediction tasks. Furthermore, SG-ATT exhibited a relative improvement of 2.9% over the state-of-the-art (SOTA) baseline in the small dataset FreeSolv, which comprises only 642 labeled molecules. This result confirms the advantage of SG-ATT in small datasets for several reasons. Firstly, the model leverages functional group-level knowledge, amplifying its capacity to capture distinctive features of hydrophilic functional groups such as hydroxyl (-OH) and amino (-NH_2), as well as hydrophobic functional groups like alkyl (-CH_3). These functional groups often exert a crucial influence on a molecule's solubility in various solvents. Secondly, the SG-ATT model seamlessly integrates multidimensional features and chemical element knowledge, empowering it not only to glean more nuanced molecular representations from limited datasets but also to enhance its generalization prowess across a spectrum of datasets.

To further validate the model's performance, we treated each task on the Tox21, ToxCast, and SIDER datasets as a separate attribute and then compared ROC-AUC among tasks within each dataset. Figure 1 illustrates the comparison of ROC-AUC for nine attribute tasks on each dataset using different models. Our model achieved state-of-the-art performance in 20 out of the total 27 attribute tasks.

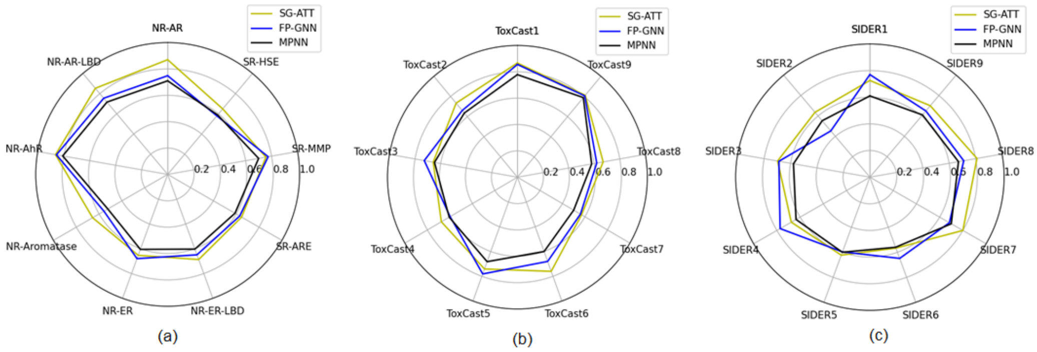


Figure 1. The test results of attribute tasks on multitask datasets. (**a**) represents the testing results on the Tox21 dataset; (**b**) represents the testing results on the ToxCast dataset; (**c**) represents the testing results on the SIDER dataset.

Table 1. Comparison of data from molecular attribute prediction models.

Metric	ROC-AUC				RMSE		MAE
Dataset	**BBBP**	**BACE**	**HIV**	**ClinxTox**	**ESOL**	**FreeSolv**	**QM7**
Weave [2016]	0.837	0.791	—	0.823	1.158	2.398	—
GCN [2017]	0.877	0.854	0.740	0.807	1.068	2.900	—
MPNN [2017]	0.913	0.815	0.770	0.879	1.167	2.185	111.4
DMPNN [2019]	0.919	0.852	0.771	0.897	0.980	2.177	103.5
CMPNN [2020]	0.927	0.869	0.782	0.902	0.798	0.956	75.1
CoMPT [2021]	0.938	0.871	—	**0.934**	0.774	1.855	65.3

Table 1. *Cont.*

Metric	ROC-AUC				RMSE		MAE
Dataset	**BBBP**	**BACE**	**HIV**	**ClinxTox**	**ESOL**	**FreeSolv**	**QM7**
FP-GNN [2022]	0.916	0.860	0.824	0.840	0.675	0.905	—
SG-ATT	**0.943**	**0.910**	**0.827**	0.920	**0.646**	**0.879**	**64.1**

Bold cells indicate the best results achieved, underlined cells indicate the second-best results achieved, and "—" indicates that the model did not undergo corresponding experimental testing.

2.5. Ablation Experiments

The ablation study was conducted to verify the importance of multidimensional features in the SG-ATT model on model performance. The experiments were repeated five times on each encoder to obtain stable results. Table 2 shows the experimental results of SG-ATT.

Table 2. Results of SG-ATT ablation experiments.

Metric	ROC-AUC				RMSE		MAE
Dataset	**BBBP**	**BACE**	**HIV**	**ClinxTox**	**ESOL**	**FreeSolv**	**QM7**
SG-ATT	**0.943**	**0.910**	**0.827**	**0.920**	**0.646**	**0.879**	**64.1**
-w/o Graph	0.883	0.864	0.757	0.853	0.858	1.126	68.7
-w/o Sequnce	0.922	0.879	0.788	0.910	0.692	0.904	67.3

Bold cells indicate the best results achieved.

The bold cells indicate the best results obtained through SG-ATT, and the underlined cells indicate the second-best results obtained via SG-ATT. w/o Graph indicates the removal of the graph feature module in the SG-ATT model, and w/o Sequence indicates the removal of the sequence feature module in the SG-ATT model. The experimental results in Table 2 show that the exclusion of either component (Transformer and AMPNN) can easily degrade the model performance. Excluding the AMPNN module causes more obvious model performance degradation, which proves the effectiveness of AMPNN in improving SG-ATT performance. The SG-ATT experiments obtain the worst results if both components are simultaneously excluded. The ablation experiments confirm that the SG-ATT model can learn richer molecular representations and improve the prediction performance of downstream tasks.

2.6. Exploration of Model Interpretability

Figure 2 depicts the prediction results of the four models on the BBBP test set, with red dots representing instances predicted with a label of 0 and blue dots representing instances predicted with a label of 1. When predictions are entirely accurate, there are a total of 42 instances with a label of 0 in the BBBP test set, represented by red dots. However, in the prediction plot of GCN, only 18 instances are predicted to be 0, whereas in SG-ATT, 30 instances are predicted to be 0. This highlights that the SG-ATT model, through the fusion of multi-dimensional features and prior knowledge, can capture rich features of different molecular types. Consequently, compared to other baseline models, the SG-ATT model yields more accurate prediction results.

To validate whether the molecular embeddings learned through the SG-ATT model align with real-world knowledge, experiments were conducted for the blood–brain barrier permeability task using randomly selected eight molecules with different scaffolds. Different molecules are depicted in Figure 3, where distinct molecular scaffolds exhibit both similar and dissimilar cases. As shown in Figure 4, the trained model was employed to obtain embedding vectors for the eight molecules. The cosine similarity between different vectors was computed to assess the features learned through the model concerning molecular characteristics for various scaffolds. The specific SMILES representations for each molecular graph are provided in Appendix D.

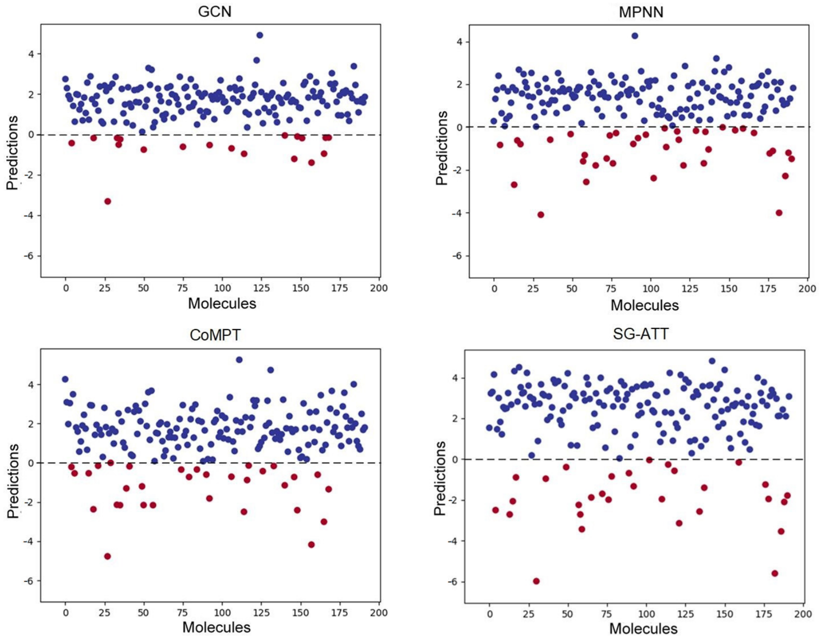


Figure 2. BBBP dataset model test results.

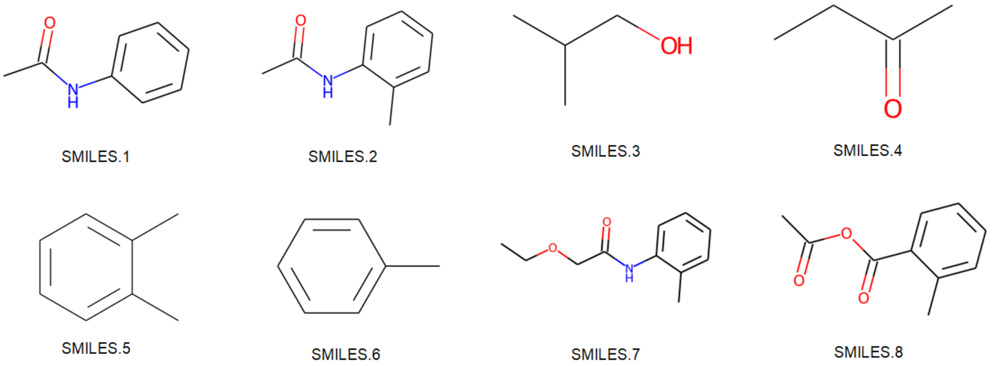


Figure 3. Molecules with different scaffolds.

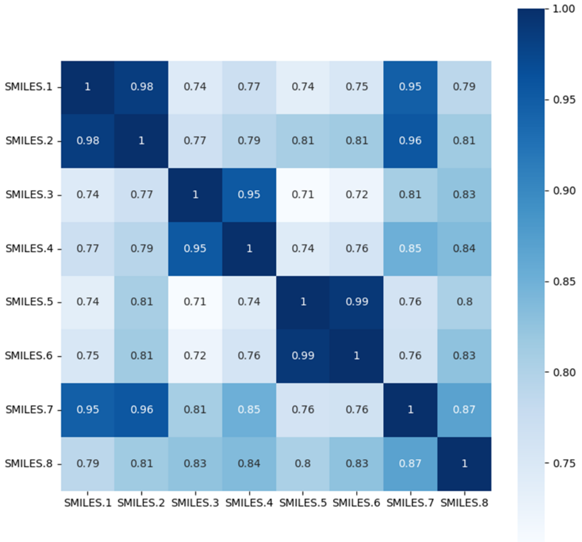


Figure 4. Similarity heatmap of molecules with different scaffolds.

T-distributed random neighbor embedding (t-SNE) is a machine learning algorithm used for data dimensionality reduction, visualizing high-dimensional data, and providing an intuitive understanding of the data distribution. To further investigate the effectiveness of SG-ATT in learning molecular features, the dimensionality of the embedding vectors learned through the SG-ATT model was reduced to 2D using the t-SNE visualization method. As shown in Figure 5, it can be observed that the molecular embedding vectors obtained through SG-ATT can clearly distinguish active molecules from inactive ones with higher classification accuracy. Additionally, the embeddings of different active molecules exhibit a more clustered pattern. Therefore, SG-ATT can learn differentiated feature representations among different active molecules and capture similar features among molecules of the same class, which is crucial for its outstanding performance in molecular property prediction tasks.

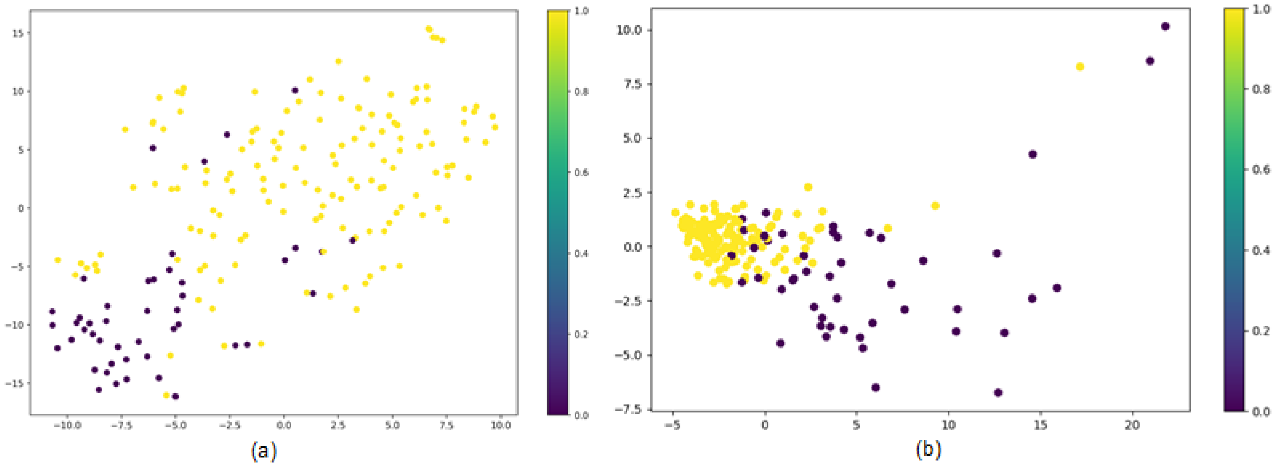


Figure 5. t-SNE feature dimension reduction. (**a**) represents the results of the MPNN model trained on the BBBP dataset, while (**b**) represents the results of the SG-ATT model trained on the same dataset.

2.7. Case Analysis

Chronic hepatitis B caused by hepatitis B virus (HBV) is a kind of serious disease. HBV inhibitory drugs are developed from a variety of compounds, and to further explore whether anti-HBV compounds are potentially promising for research, it is necessary to validate their anti-HBV capacity (IC_{50}) and HepG2 hepatocytotoxicity (CC_{50}). We used compounds screened with the fields of target name and target tissue from the CHEMBL database as training data and set the HBV (IC_{50}) concentration and hepatotoxicity concentration (CC_{50}) thresholds at 20 µM (the smaller the IC_{50} value; the more specific the performance of the antibody) and 100 µM (higher CC_{50} concentration indicates lower toxicity), respectively, to convert the compound activity data against the target into binary data for prediction [31]. To validate the reliability of the SG-ATT model, we also used in vitro validation to test the model on IC_{50} and CC_{50} to predict the properties of the compounds. Throughout the in vitro validation, we tested the activity data of 100 compounds for IC_{50} and CC_{50}. Table 3 shows the final test results of CC_{50} and IC_{50} on CHEMBL data and in vitro validation data.

Table 3. Test results on CHEMBL screening data and in vitro validation data.

Dataset	ROC-AUC
CHEMBL_{CC50}	0.817
in vitro validation_{CC50}	0.704
CHEMBL_{IC50}	0.783
in vitro validation_{IC50}	0.696

We counted the model prediction results and listed the top 10 compounds ranked using µM prediction size. The first represents the largest µM, and the 10th represents the smallest μM. Table 4 shows the IC_{50} prediction results and in vitro validation results obtained through the SG-ATT model for compounds with specified thresholds, and Table 5 shows the CC_{50} prediction results and in vitro validation results of the model for compounds with specified thresholds. The tests of IC_{50} and CC_{50} for compounds not present in the training data show that the model can predict different physicochemical properties of compounds as well as generalizability to different datasets.

Table 4. Top 10 compounds for predicting IC_{50}.

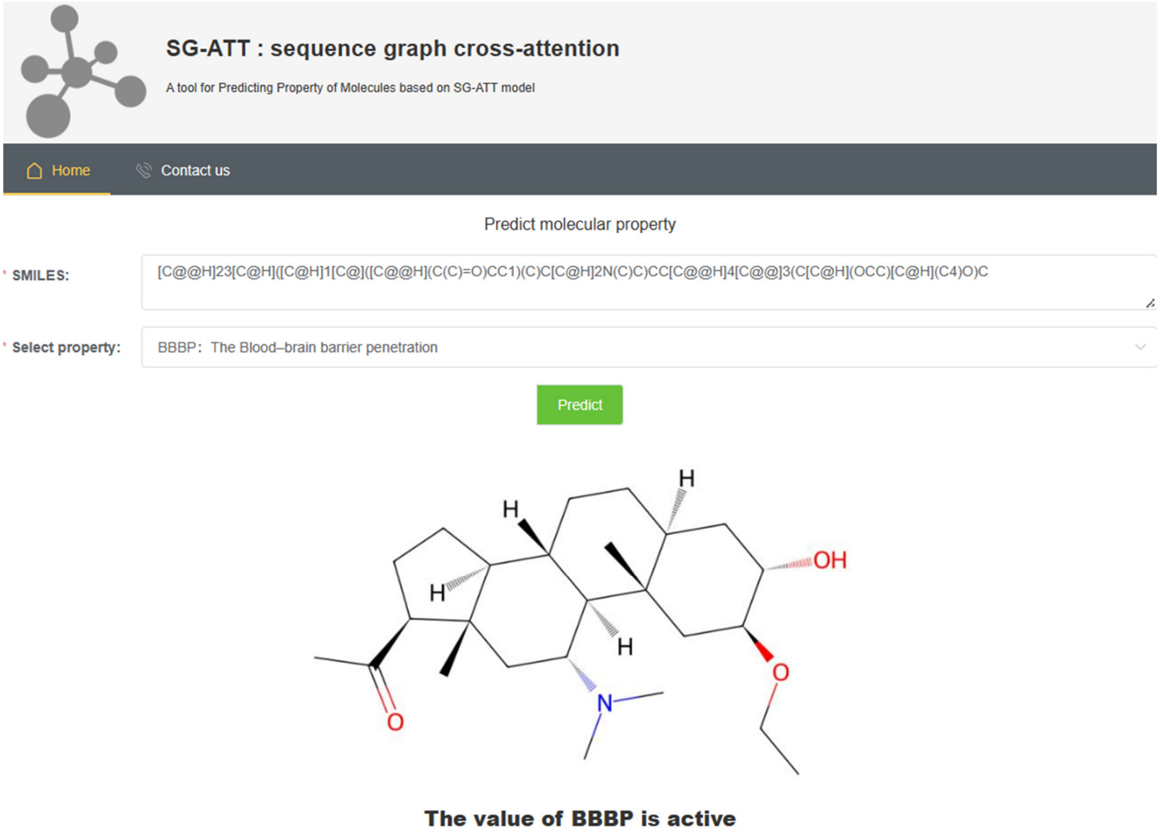
Rank	SMILES	In Vitro Validation (μM)
1	O=C(N[C@H](CO)CC1=CC=CC=C1)/C(NC(C2=C(F)C=CC=C2)=O)=C\C3=CC=CC=C3	>20
2	O=C(N[C@H](CO)CC1=CC=CC=C1)/C(NC(C2=CC=CC=C2)=O)=C\C3=CC=CC=C3Br	>20
3	O=C(N[C@H](CO)CC1=CC=CC=C1)/C(NC(C2=CC=CC=C2)=O)=C\C3=CC=CC=C3Cl	>20
4	Br/C(C1=CC=CC=C1)=C(NC(C2=CC=CC=C2)=O)/C(N[C@H](CO)CC3=CC=CC=C3)=O	3.86
5	Br/C(C1=CC=C(Br)C=C1)=C(NC(C2=CC=CC=C2)=O)/C(N[C@@H](CC3=CC=CC=C3)CO)=O	4.45
6	O=C(N[C@H](CO)CC1=CC=CC=C1)/C(NC(C2=CC=CC=C2)=O)=C\C3=CC=CC(C)=C3	>20
7	O=C(N1CCCCC1)[C@@H](NC(C2=CC=CC=C2)=O)CC3=CC=CC=C3	>20
8	Br/C(C1=CC=CC=C1OC)=C(NC(C2=CC=C([N+]([O-])=O)C=C2)=O)/C(N[C@@H](CO)C(C)C)=O	2.06
9	Br/C(C1=CC=CC=C1Cl)=C(NC(C2=CC=CC=C2)=O)/C(N[C@@H](CC3=CC=CC=C3)CO)=O	1.88
10	O=C1C2=CC(OCCN(CC)CC)=CC=C2N=C(SCC3=C(OC)C=CC=C3)N1CC4=CC=CS4	1.93

Table 5. Top 10 compounds for predicting CC_{50}.

Rank	SMILES	In Vitro Validation (μM)
1	Br/C(C1=CC=C(O)C=C1)=C(NC(C2=CC=CC=C2)=O)/C(N[C@H](CO)CC3=CC=CC=C3)=O	>100
2	O=C(N[C@H](CO)CC1=CC=CC=C1)/C(NC(C2=C(F)C=CC=C2)=O)=C\C3=CC=CC=C3	>100
3	O=C(N[C@H](CO)CC1=CC=CC=C1)/C(NC(C2=CC=CC=C2)=O)=C\C3=CC=CC=C3Cl	>100
4	Br/C(C1=CC=CC=C1Br)=C(NC(C2=CC=CC=C2)=O)/C(N[C@@H](CC3=CC=CC=C3)CO)=O	>100
5	O=C(N[C@H](CO)CC1=CC=CC=C1)/C(NC(C2=CC=CC=C2)=O)=C\C3=CC=CC(C)=C3	>100
6	O=C(N[C@H](CO)CC1=CC=CC=C1)/C(NC(C2=CC=CC=C2)=O)=C\C3=CC=C(Br)C=C3	72.66
7	Br/C(C1=CC=C(C#N)C=C1)=C(NC(C2=CC=CC=C2)=O)/C(N[C@@H](CC3=CC=CC=C3)CO)=O	>100
8	O=C(N/C(C(N1CCCCC1)=O)=C(Br)/C2=NC=CS2)C(C=C3)=CC=C3F	>100
9	O=C(NC(CO)C(C)C)[C@@H](NC(C1=CC=CC=C1)=O)CC2=CC=C(OCCN(C)C)C=C2	89.10
10	O=C(N[C@H](CC(C)C)CO)/C(NC(C1=CC=C([N+]([O-])=O)C=C1)=O)=C(Br)\C2=CC=CC=C2OC	86.61

2.8. Web Server Implementation

To facilitate researchers in implementing the proposed SG-ATT model, an easy-to-use web server was built on http://vectorspaceai.cn/SG-ATT (accessed on 12 June 2023). The server interface is shown in Figure 6. To obtain the desired results, the user only needs to upload the SMILES string of the compound of interest and select the desired predicted molecular properties. Finally, the service returns the molecular graph of the corresponding compound along with the predicted results of the desired property.



SG-ATT : sequence graph cross-attention

A tool for Predicting Property of Molecules based on SG-ATT model

Predict molecular property

* SMILES: [C@@H]23[C@H]([C@H]1[C@]([C@@H](C(C)=O)CC1)(C)C[C@H]2N(C)C)CC[C@@H]4[C@@]3(C[C@H](OCC)[C@H](C4)O)C

* Select property: BBBP：The Blood–brain barrier penetration

The value of BBBP is active

Figure 6. Screen capture of the web tool driven by SG-ATT.

3. Materials and Methods

Molecular property prediction can be considered as classification tasks and regression tasks to determine whether a molecule has certain activities. In the SG-ATT approach, a molecule is represented by a sequence of SMILES, which consists of a series of chemical atoms and chemical bonds. The sequence of SMILES is denoted as $S_l = \{s_1, s_2, \cdots s_i\}$, where i is the length of S_l.

In this section, the general framework of the SG-ATT model is first introduced; then, the specific methods of each module are described. The structure of the SG-ATT model is shown in Figure 7. The input to SG-ATT is a sequence of one-dimensional molecular SMILES. In part I, the SMILES sequence is converted into a molecular graph using the RDKit toolkit; then, the original molecular graph is enhanced through MKG and used as the input to the graph encoder. For the graph encoder, atomic feature vectors of the molecules are generated by combining feature attention through the atom-aware message-passing neural network (AMPNN). In part II, the SMILES sequence is directly encoded using the frequent consecutive subsequence (FCS) algorithm to generate sequence data as input to the sequence encoder [32]. The sequence encoder utilizes the transformer architecture to encode the sequence data directly and generate a feature vector that represents the one-dimensional sequence of molecules. In part III, 2D molecular graph feature vectors and 1D sequence feature vectors are fed into a cross-attention block. The feature mixing part generates a high-dimensional vector on which a decoder is implemented for downstream molecular property prediction tasks.

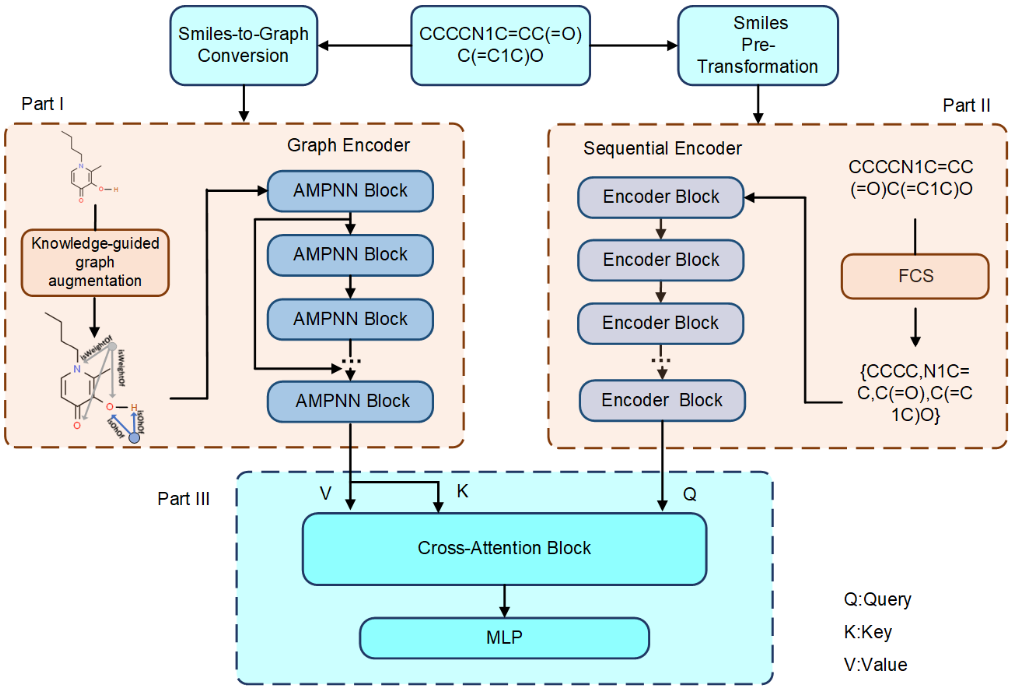


Figure 7. SG-ATT model framework.

3.1. Feature Extraction

SMILES is converted to undirected graphs using the RDKit toolkit, denoted as $G = \{V, E\}$ [33], where V is a set of n nodes, and E is a series of edges. e_{ij} is the edge $(i,\ j)$, representing the initial feature. Let *h(j)* be the node hiding state and $h(e_{ij})$ be the edge hiding state. Learning the graph encodes, $f = G \rightarrow R^{d}$, which encodes the input graph into a vector representation with no labels.

The SG-ATT model applies a data-driven sequence pattern mining algorithm called the FCS algorithm. This algorithm can progressively decompose the SMILES sequence of a molecule into smaller subsequences. Molecular properties are often determined using specific substructures, and FCS decomposes molecular SMILES sequences into medium-sized substructures that are more likely to provide clear indications. The specific representation is shown in Equation (1):

$$FCS(s_l) = D_i = \{d_1, d_2 \dots d_q | d_k \in C\} \tag{1}$$

where C is the FCS algorithm vocabulary, s is the molecular SMILES code, and D is the s FCS subsequence.

3.2. Sequence Encoding: Transformer

The Transformer encoder relies on an attention mechanism to compute contextual features, which is significantly different from the Recurrent Neural Network and Convolutional Neural Networks [34,35]. The Transformer encoder is suitable for encoding sequential information and has been widely used in NLP. The multiheaded attention mechanism enables the Transformer encoder to learn the features of different subsequences in a sequence. When processing sequence information, the Transformer encoder can associate different positions of the sequence to obtain embeddings that contain contextual information.

The FCS-encoded sequence is input to the Transformer encoder for sequence encoding to generate a feature vector that contains 1D sequence-structure information $T_{transformer}$. The specific process is shown in Equations (2) and (3).

$$D = FCS(s) \tag{2}$$

and

$$T_1 = Transformer(D). \tag{3}$$

3.3. Knowledge Graph Enhancement

Previous work has been conducted to construct knowledge graphs (KGs) from public chemical databases and scientific literature to extract associations between chemicals and disease or drug pairs [36–38]. Considering previous research, it is known that microscopic connections exist even between atoms in a molecule that are directly connected through chemical bonds [17]. Based on this approach, MKG creates triples based on the chemical elements in the periodic table of the chemical elements and their properties, with each triple containing five properties: periodicity, metallicity, group, radius, and weight. Moreover, functional groups in a molecule often play a crucial role in determining its properties. Recognizing this importance, MKG has incorporated information on commonly found functional groups in molecules. When it is found that a certain substructure exists in the molecule, a virtual node connecting the atoms is established among the substructure according to the guidance of MKG to capture the information of the substructure. Specifically, as shown in Figure 8, a connection is established between an atom in a molecule extracted in the MKG, and a 1-hop neighbor property based on that atom, and a triple is added as an edge. Since some element attributes are continuous, it is difficult for MKG to model their associations. Therefore, continuous element attributes are sampled into discrete grouping labels (e.g., radius group 1 and radius group 2). For example, at the atomic level, a node "Weight2" and an edge from "Weight2 (divide into the second group based on weight.)" to "O (oxygen)" are added to the original molecular graph based on the triple (Weight2, isweightOf, O). Similarly, at the functional group level, two edges from "OH (oxhydryl)" to "O (oxygen)" and "H (Hydrogenium)" are added, respectively. Each edge between a property and an atom is oriented from the former to the latter, while the edges between atoms are bidirectional. Similarly, when a molecular fragment is found in the molecule, a connection is established in the substructure through the obtained atomic number. Then, an augmented molecular graph is obtained, which preserves the original molecular structure and introduces virtual nodes to capture long-distance atomic interactions and molecular substructure information. The augmented molecular graph G' considers the microscopic correlations between atoms, containing richer and more complex information.

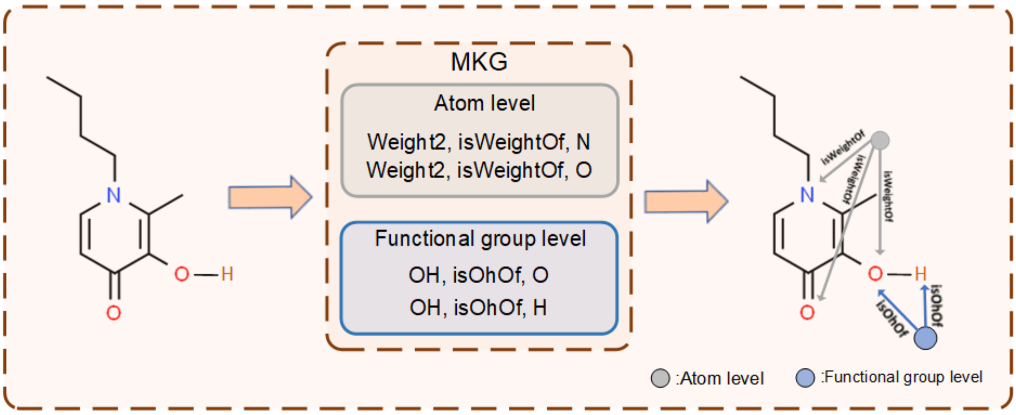


Figure 8. Molecular knowledge graph enhancement process.

3.4. Molecular Graph Encoding: AMPNN

With the inclusion of domain knowledge extracted from MKG in the augmented molecular graph, KMPNN is proposed to enhance the feature encoding of the molecular graph [17]. However, considering the inclusion of different types of molecular information in the augmented molecular graph, AMPNN extends the feature aggregation method of KMPNN. After the message-passing process, it assigns different feature aggregation methods to various types of messages. By employing distinct aggregation methods, the model can better capture the features of each message type, enhancing the richness of information. Figure A1 in Appendix A describes the detailed flow of AMPNN. The key idea behind AMPNN is to use two types of message-passing channels to encode different types of neighboring nodes. During the message-passing process, a feature fusion method based on the attention mechanism is used to better incorporate the correlations between features into the model. The input to the encoder is an augmented molecular graph $G' = \{V, E\}$. Let N_i denote the set of neighbors of node i, and X_j is the initial feature of node j. Then, K rounds of messaging are applied to all nodes. Heterogeneous message passing is enabled using two M_t functions, where M_{t1} is applied to the attributes in the neighbors, and M_{t2} is applied to the neighbors that represent atoms.

In addition, the message-passing process is extended via a self-attentive mechanism, where the attention coefficients are calculated using Equation (4) and normalized through the Softmax function to make the coefficients easily comparable among different nodes [39].

$$\alpha_{uv} = \frac{exp\left(LeakyReLU\left(a^T[Wh_u||Wh_v]\right)\right)}{\sum_{K\in N_v} exp\left(LeakyReLU\left(a^T[Wh_u||Wh_v]\right)\right)}, \tag{4}$$

where $(\cdot)^T$ denotes the transposition, and $||$ denotes the crosstalk operation. a^T is a parameter vector with a dimension of R^{2F}, where F is the dimension of the node feature vector and serves as a parameter for the feedforward neural network. Based on this attention coefficient, we can derive Equation (5):

$$h_i{}' = \sigma\left(\alpha_{ij} W_1 h_{e(ij)}{}^{K-1}\right), \tag{5}$$

where α_{ij} represents the attention coefficient between attributes, and W_1 is the weight matrix for relationships. The message passing function is represented using Equation (6):

$$M_{t1} = h_i{}' \cdot h_i{}^{k-1}. \tag{6}$$

Due to the varying importance of messages passed from different neighboring atoms to the central atom, neighboring atoms adhere to a common process with distinct parameters, as depicted in Equations (7)–(9).

$$\beta_{ij} = \frac{exp\left(LeakyReLU\left(b^T\left[Wh_i||Wh_j\right]\right)\right)}{\sum_{K\in N_i} exp\left(LeakyReLU\left(b^T\left[Wh_i||Wh_j\right]\right)\right)}, \tag{7}$$

$$h_i{}' = \sigma\left(\beta_{ij} W_2 h_{e(ij)}{}^{K-1}\right), \tag{8}$$

and

$$M_{t2} = h_i{}' \cdot h_i{}^{k-1} \tag{9}$$

where β_{ij} represents the attention coefficient between atoms, and W_2 is the weight matrix for bonds. During the message-passing process, aggregation collects messages from neighboring edges, as shown in Equations (10) and (11).

$$e_{ij}^{K-1} = M_t\left(h_{e(ij)}{}^{K-1} \cdot h_i{}^{k-1}\right) \tag{10}$$

and

$$e_i^K = Aggregate\left(e_{ij}^{K-1}\right), \tag{11}$$

where aggregation types include summation and averaging. Furthermore, the node's hidden state is updated using the GRU function, depicted in Equation (12).

$$h^K(v) = GRU\left(h^{K-1}(e_{uv}),\ e_i^K(v)\right), \tag{12}$$

where GRU is a gated recurrent neural network [40]. After K iterations, a readout operator (13) is applied to obtain a graph-level representation of the molecule.

$$T_{kmpnn}^{i} = Set2Set\left(h^{K}(v) \middle| v \in G_{i}'\right), \tag{13}$$

where Set2Set is specifically designed to operate on sets and is more expressive than simply summing the final node states [41].

3.5. Cross-Attention Feature Fusion

The first of the feature vectors is passed into the feature blending section, which blends the features from the 1D sequence feature encoder and 2D graphical feature encoder. This information is combined through a cross-attention mechanism (14):

$$Attention(Q, K, V) = Softmax\left(\frac{QK^{T}}{\sqrt{C/d}}\right) \cdot V, \tag{14}$$

where Q is created from the output of the sequence encoder, and K and V are generated from the output of the graph encoder using the projection functions $f = w^{T}x + b$ (where w and b are weight and bias, respectively). C and d are the embedding dimensions and the number of heads, respectively.

When predicting molecular properties, the multidimensional features of molecules are encoded into a feature decoder. This feature decoder then produces a final predicted label that indicates the molecular properties. Using a decoder consisting of a three-layer feedforward neural network, a strong correlation can be established between input features and output results.

3.6. Loss Function

To build a molecular property prediction model, a two-channel multidimensional molecular feature encoder is constructed. The transformer is applied to the sequence feature encoder, and the enhanced graph is input to the graph feature encoder, AMPNN. The multidimensional feature decoder decodes all embedding vectors from the embedding space. For the classification task, the output of the decoder is a label of 0 or 1, where 1 indicates that the predicted molecule has a certain activity, and 0 indicates that no predicted molecule has a certain activity. For the regression task, the output of the decoder is a logistic value. The SG-ATT model is optimized using the binary cross-entropy loss function; for the regression task, the RMSE loss function is used, where y denotes the true label and x denotes the model prediction, as shown in Equations (15) and (16).

$$Loss_{classification} = -\sum[yln(x) + (1-y)ln(1-x)] \tag{15}$$

and

$$Loss_{regression} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y-x)^{2}}. \tag{16}$$

Backpropagation is propagated from the output layer to the previous layer. With this end-to-end approach, the model is trained with all trainable parameters. The results show that end-to-end training can greatly improve the performance of the model because all trainable parameters accept the gradient of the loss function. In this study, the losses are propagated through a two-channel multidimensional molecular feature encoder and a multidimensional feature decoder.

4. Conclusions

Combining MKG-guided molecular graph enhancement, connections are established between atoms with similar attributes as well as between atoms within the same substructure. This knowledge-enhanced molecular graph is believed to better learn the information transfer between nodes to supplement the shortcomings of insufficient long-distance information transmission and tends to overlook substructure features in the molecular graph. In addition, many existing works focus on the feature learning of molecular substructures, which are obtained by directly slicing the molecular substructure to replace the original molecular graph or by fusing the substructure with the original graph for features. We aim to supplement important substructure information in molecules from a sequence perspective by slicing the molecular SMILES sequence with the FCS algorithm.

This study introduces the SG-ATT model, a multidimensional molecular feature attention encoding based on knowledge graphs designed for predicting molecular properties. By integrating multidimensional features and knowledge graph characteristics, the SG-ATT model's ability to capture molecular information in limited datasets has been enhanced, thereby improving its generalization capabilities across diverse datasets. Additionally, the incorporation of functional group information allows the SG-ATT model to learn structural features that have a significant impact on molecular properties, consequently enhancing the model's predictive accuracy. Through comparative experiments, visual studies, and case analyses, we demonstrated the improved performance of the SG-ATT model. However, existing research does have certain limitations that require attention. On the one hand, current Molecular Knowledge Graphs (MKGs) primarily integrate atomic and functional group features, leaving ample room for further expansion. On the other hand, the one-dimensional sequence encoding module holds the potential for more significant optimization, such as incorporating atom distance matrices to reinforce self-attention mechanisms. In our upcoming work, we are committed to addressing these limitations and refining the model accordingly.

Author Contributions: Conceptualization, Y.H. and L.W.; Methodology, Y.H. and J.S.; Software, Y.H.; Validation, Y.H., X.C., and A.F.; Formal Analysis, Y.H. and X.C.; Investigation, Y.H., Q.J., and Y.C.; Resources, Y.H.; Data Curation, Y.H.; Writing—Original Draft Preparation, Y.H., S.P., and L.W.; Writing—Review and Editing, Y.H., S.P., and L.W.; Visualization, Y.H.; Supervision, Y.H. and L.W.; Project Administration, Y.H. and J.S.; Funding Acquisition, L.W. and S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 81873915), funded by Wang Li and the Foreign Youth Talent Program of the Ministry of Science and Technology, China (No. QN2022014011L), funded by Sanjeevi Pandiyan.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets and codes used in this study are available at GitHub: https://github.com/NTU-MedAI/SG-ATT (accessed on 20 April 2023).

Acknowledgments: We thank our partners who provided all the help during the research process and the team for their great support.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Details of AMPNN

KMPNN builds upon MPNN by considering the establishment of transmission channels for different types of messages, utilizing two distinct message-passing channels. AMPNN, based on KMPNN, considers the creation of distinct message aggregation methods for the transmission of different message types, enabling a more diverse aggregation of various message types. Figure A1 illustrates the architecture of AMPNN. It represents different types of nodes in the molecular graph, including atoms and attributes, through heterogeneous message passing. During the message delivery process, nodes of the same type share message parameters, and different parameters are used to calculate attention coefficients between atoms and attributes, as well as between atoms and other atoms.

Additionally, in the message aggregation stage, different types of messages are assigned different feature aggregation methods.

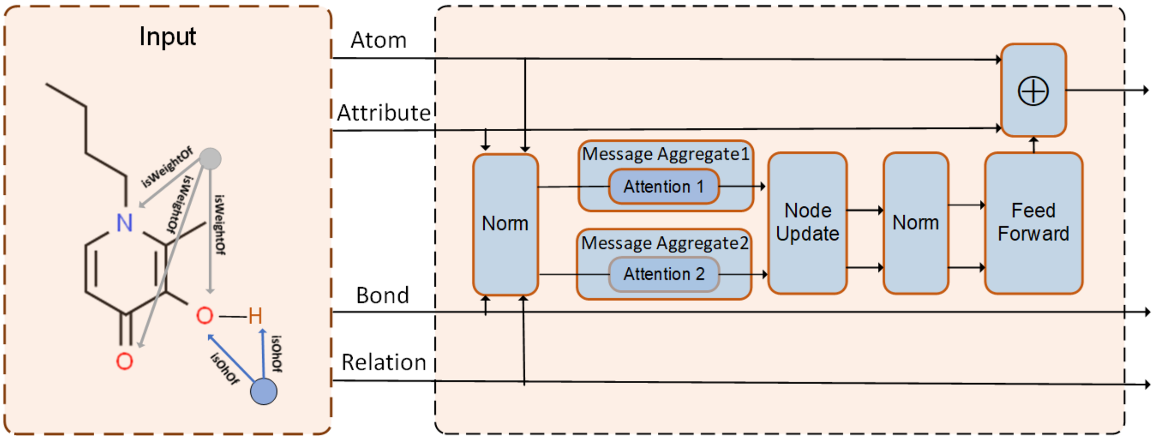


Figure A1. Architecture of AMPNN.

Appendix B. Details of GRU and Transformer

The Gated Recurrent Unit (GRU) is proposed to address the problems of long-term memory and gradient in backpropagation. The GRU combines the current input x^t with the hidden state h^{t-1}, passed from the previous node (this hidden state contains the relevant information of the previous node), and the GRU produces the output y^t of the currently hidden node and the hidden state h^t passed to the next node.

The encoder of the Transformer encodes an input sequence of symbol representations $x = (x_1, \ldots, x_n)$ to a sequence of continuous representations $z = (z_1, \ldots, z_n)$. Given z, the decoder then generates an output sequence $y = (y_1, \ldots, y_n)$ of symbols one element at a time. At each step, the model is auto-regressive, consuming the previously generated symbols as additional input when generating the next. An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed using a compatibility function of the query with the corresponding key. The Transformer follows this overall architecture using stacked self-attention and pointwise, fully connected layers for both the encoder and decoder.

Appendix C. Details of Dataset

Appendix C.1. Dataset Description

Table A1 summarizes the information on the benchmark datasets, including the task type, dataset size, and split type. Detailed information for each dataset is shown below. These datasets play a significant role in the fields of chemistry and pharmaceutical research, holding practical value for tasks such as developing predictive models and conducting drug design.

The BACE dataset consists of drug molecules with diverse structural characteristics, serving as a platform for predicting bioactivity against beta-secretase [42].

The HIV dataset includes compounds along with their viral inhibition activities, providing crucial insights for antiretroviral drug research [43].

The BBBP dataset evaluates blood–brain barrier permeability prediction using molecular features, aiding in an understanding molecule–brain interactions [44].

Table A1. Dataset information.

Type	Category	Dataset	Tasks	Compounds	Split
Classification	Biophysics	BACE	1	1513	Scaffold
		HIV	1	41,127	Scaffold
	Physiology	SIDER	27	1427	Random
		ClinTox	2	1478	Random
		BBBP	1	2039	Scaffold
		Tox21	12	7831	Random
		ToxCast	617	8575	Random
Regression	Physical chemistry	FreeSolv	1	642	Random
		ESOL	1	1128	Random
	Quantum mechanics	QM7	1	7160	Random

The ClinTox dataset encompasses drug compounds and associated toxicity labels, offering valuable insights into drug safety assessment [45].

The Tox21 dataset provides molecular features and covers diverse toxicity activity labels, facilitating toxicity prediction tasks [46].

The ToxCast dataset features various toxicity activity data, offering a comprehensive resource for predictive toxicology [47].

The SIDER dataset includes molecular structures and drug side effect labels, facilitating the analysis of drug safety profiles [48].

The ESOL dataset focuses on predicting the solubility of different molecules, contributing to drug formulation studies [49].

The FreeSolv dataset involves predicting the free energy changes of diverse molecular complexes, a valuable resource for free energy calculations of solvation [50].

QM7: This is a dataset designed for quantum chemistry calculations aimed at studying the electronic structures of small organic molecules and predicting the quantum mechanical properties of molecules [51].

Appendix C.2. Dataset Splitting

As shown in Table A1, molecular skeleton-based splitting and random splitting were applied to all tasks for all datasets [52]. Molecular skeleton-based splitting splits molecules with different 2D structural frameworks into different subsets. This is a more challenging but practical setup because the test molecules may have different structures from the training set. All datasets were divided into training, validation, and test sets at a ratio of 8:1:1. In addition, scaffold splitting was performed for the BACE, BBBP, and HIV datasets, and random splitting was used for the remaining datasets.

Appendix D. Specific SMILES for Each Molecular Graph

Table A2 lists the specific SMILES representations corresponding to each molecule.

Table A2. Specific SMILES.

Name	SMILES
SMILES.1	CC(=O)Nc1ccccc1
SMILES.2	O=C(C)Nc1ccccc1C
SMILES.3	CCC(C)=O
SMILES.4	CC(C)CO
SMILES.5	Cc1ccccc1C
SMILES.6	Cc1ccccc1
SMILES.7	CCOCC(=O)Nc1ccccc1C
SMILES.8	O=C(C)OC(=O)c1ccccc1C

References

1. Kolluri, S.; Lin, J.; Liu, R. Machine Learning and Artificial Intelligence in Pharmaceutical Research and Development: A review. *AAPS J.* **2022**, *24*, 1–10. [CrossRef]
2. Weihua, H.; Bowen, L.; Joseph, G. Strategies for Pre-training Graph Neural Networks. *arXiv* **2019**. [CrossRef]
3. Rong, Y.; Bian, Y.; Xu, T. Self-Supervised Graph Transformer on Large-Scale Molecular Data. *NeurIPS* **2020**, *33*, 12559–12571. [CrossRef]
4. Mater, A.C.; Coot, M.L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2020**, *59*, 2545–2559. [CrossRef] [PubMed]
5. Chuang, K.V.; Gunsalus, L.M.; Keiser, M.J. Learning molecular representations for medicinal chemistry. *J. Med. Chem.* **2020**, *63*, 8705–8722. [CrossRef]
6. Shen, J.; Nicolaou, C.A. Molecular property prediction: Recent trends in the era of artificial intelligence. *Drug Discov. Today Technol.* **2019**, *32*, 29–36. [CrossRef] [PubMed]
7. Mendez, D.; Gaulton, A.; Bento, A.P. ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, 930–940. [CrossRef]
8. Kim, S.; Chen, J.; Cheng, T. PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Res.* **2021**, *49*, 88–395. [CrossRef]
9. David, W. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36. [CrossRef]
10. Arús-Pous, J.; Patronov, A.; Bjerrum, E.J. SMILES-based deep generative scaffold decorator for de-novo drug design. *J. Cheminf.* **2020**, *12*, 1–18. [CrossRef]
11. Shuman, D.I.; Narang, S.K.; Frossard, P. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE SPM.* **2013**, *30*, 83–98. [CrossRef]
12. Zhu, J.; Xia, Y.; Qin, T. Dual-view molecule pre-training. *arXiv* **2021**. [CrossRef]
13. Pang, S.; Zhang, Y.; Song, T. AMDE: A novel attention-mechanism-based multidimensional feature encoder for drug–drug interaction prediction. *Brief. Bioinf.* **2022**, *23*, bbab545. [CrossRef] [PubMed]
14. Vaswani, A.; Shazeer, N.; Parmar, N. Attention is all you need. *arXiv* **2017**. [CrossRef]
15. Justin, G.; Samuel, S.S.; Patrick, R.; Oriol, V.; George, E.D. Neural Message Passing for Quantum Chemistry. *arXiv* **2017**. [CrossRef]
16. Li, R.; Cao, Y.; Zhu, Q.; Bi, G.; Fang, F.; Liu, Y.; Li, Q. How Does Knowledge Graph Embedding Extrapolate to Unseen Data: A Semantic Evidence View. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 5781–5791. [CrossRef]
17. Fang, Y.; Zhang, Q.; Yang, H.; Zhuang, X.; Deng, S.; Zhang, W.; Qin, M.; Chen, Z.; Fan, X.; Chen, H. Molecular Contrastive Learning with Chemical Element Knowledge Graph. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 3968–3976. [CrossRef]
18. Hasebe, T. Knowledge-Embedded Message-Passing Neural Networks: Improving Molecular Property Prediction with Human Knowledge. *ACS Omega* **2021**, *6*, 27955–27967. [CrossRef]
19. Qin, C.; Zhang, A.; Zhang, Z.; Chen, J.; Yasunaga, M.; Yang, D. Is ChatGPT a General-Purpose Natural Language Processing Task Solver? *arXiv* **2023**. [CrossRef]
20. Zhu, W.; Li, Z.; Cai, L. Stepping Back to SMILES Transformers for Fast Molecular Representation Inference. *arXiv* **2023**. [CrossRef]
21. Weiss, K.; Khoshgoftaar, T.M.; Wang, D.A. survey of transfer learning. *J. Big Data* **2016**, *3*, 1–40. [CrossRef]
22. Wu, Z.; Ramsundar, B.; Feinberg, E.N. MoleculeNet: A benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513–530. [CrossRef] [PubMed]
23. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2015**. [CrossRef]
24. Kearnes, S.; McCloskey, K.; Berndl, M. Molecular graph convolutions: Moving beyond fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595–608. [CrossRef] [PubMed]
25. Yang, K.; Swanson, K.; Jin, W. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388. [CrossRef] [PubMed]
26. Song, Y.; Zheng, S.; Niu, Z.; Fu, Z.H.; Lu, Y.; Yang, Y. Communicative Representation Learning on Attributed Molecular Graphs. *IJCAI* **2020**, *2020*, 2831–2838. [CrossRef]
27. Chen, J.; Zheng, S.; Song, Y. Learning attributed graph representations with communicative message passing transformer. *arXiv* **2021**. [CrossRef]
28. Cai, H.; Zhang, H.; Zhao, D. FP-GNN: A versatile deep learning architecture for enhanced molecular property prediction. *Brief Bioinform.* **2022**, *23*, bbac408. [CrossRef]
29. Paszke, A.; Gross, S.; Massa, F. Pytorch: An imperative style, high-performance deep learning library. *arXiv* **2019**. [CrossRef]
30. Wang, M.; Zheng, D.; Ye, Z. Deep graph library: Agraph-centric, highly-performant package for graph neural net. *arXiv* **2019**. [CrossRef]
31. Available online: https://chembl.gitbook.io/chemblinterfacedocumentation/downloads (accessed on 5 March 2023).
32. Huang, K.; Xiao, C.; Glass, L. Moltrans: Molecular interaction Transformer for drug target interaction prediction. *Bioinformatics* **2020**, *37*, 830–836. [CrossRef]
33. Available online: https://www.rdkit.org/docs (accessed on 10 December 2021).
34. Hochreiter, S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int. J. Uncertain. Fuzz.* **1998**, *6*, 107–116. [CrossRef]

35. LeCun, Y.; Bengio, Y. Convolutional networks for images, speech, and time series. *Handb. Brain Theory Neural Netw.* **1995**, *3361*, 1–14.
36. Zeng, X.; Tu, X.; Liu, Y. Toward better drug discovery with knowledge graph. *Curr. Opin. Struct. Biol.* **2022**, *72*, 114–126. [CrossRef] [PubMed]
37. Sun, M.; Xing, J.; Wang, H.; Chen, B.; Zhou, J. Mocl: Contrastive learning on molecular graphs with multi-level domain knowledge. *arXiv* **2021**. [CrossRef]
38. Fang, Y.; Zhang, Q.; Zhang, N.; Chen, Z.; Xiang, Z.; Shao, X.; Fan, X.; Chen, H. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nat. Mach. Intell.* **2023**, *5*, 542–553. [CrossRef]
39. Veličković, P.; Cucurull, G.; Casanova, A. Graph attention networks. *arXiv* **2017**. [CrossRef]
40. Cho, K.; Van Merrienboer, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *Comput. Sci.* **2014**, 103–111. [CrossRef]
41. Vinyals, O.; Bengio, S.; Kudlur, M. Order matters: Sequence to sequence for sets. *arXiv* **2015**. [CrossRef]
42. Subramanian, G.; Ramsundar, B.; Pande, V.; Denny, R.A. Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. *J. Chem. Inf. Model.* **2016**, *56*, 1936–1949. [CrossRef]
43. Available online: http://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data (accessed on 15 April 2022).
44. Martins, I.F.; Teixeira, A.L.; Pinheiro, L. A Bayesian approach to in silico blood-brain barrier penetration modeling. *J. Chem. Inf. Model.* **2012**, *52*, 1686–1697. [CrossRef]
45. Gayvert, K.M.; Madhukar, N.S.; Elemento, O. A data-driven approach to predicting successes and failures of clinical trials. *Cell Chem. Biol.* **2016**, *23*, 1294–1301. [CrossRef] [PubMed]
46. Available online: https://tripod.nih.gov/tox21/challenge/ (accessed on 15 April 2022).
47. Richard, A.M.; Judson, R.S.; Houck, K.A.; Grulke, C.M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.; Martin, M.T.; Wambaugh, J.F.; et al. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.* **2016**, *29*, 1225–1251. [CrossRef] [PubMed]
48. Kuhn, M.; Letunic, I.; Jensen, L.J. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **2016**, *44*, 1075–1079. [CrossRef] [PubMed]
49. Delaney, J.S. ESOL: Estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000–1005. [CrossRef]
50. Mobley, D.L.; Guthrie, J.P. FreeSolv: A database of experimental and calculated hydration free energies, with input files. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 711–720. [CrossRef]
51. Blum, L.C.; Reymond, J. 970 million druglike small molecules for virtual screening in the chemical universe database gdb-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732–8733. [CrossRef]
52. Bemis, G.W.; Murcko, M.A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.