Atom-Motif Contrastive Transformer for Molecular Property Prediction

WENTAO YU, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
SHUO CHEN, School of Intelligence Science and Technology, Nanjing University, Suzhou, China and RIKEN Center for Advanced Intelligence Project, Chuo-ku, Japan
CHEN GONG, School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai, China
BO HAN, Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China and RIKEN Center for Advanced Intelligence Project, Chuo-ku, Japan
GANG NIU, RIKEN Center for Advanced Intelligence Project, Chuo-ku, Japan
MASASHI SUGIYAMA, RIKEN Center for Advanced Intelligence Project, Chuo-ku, Japan and Complexity Science and Engineering, The University of Tokyo, Bunkyo, Japan

Recently, Graph Transformer (GT) models have been widely used in the task of Molecular Property Prediction (MPP) due to their high reliability in characterizing the latent relationship among graph nodes (i.e., the atoms in a molecule). However, most existing GT-based methods usually explore the basic interactions between pairwise atoms, and thus they fail to consider the important interactions among critical motifs (e.g., functional groups consisted of several atoms) of molecules. As motifs in a molecule are significant patterns that are of great importance for determining molecular properties (e.g., toxicity and solubility), overlooking motif interactions inevitably hinders the effectiveness of MPP. To address this issue, we propose a novel Atom-Motif Contrastive Transformer (AMCT), which not only explores the atom-level interactions but also considers the motif-level interactions. Since the representations of atoms and motifs for a given molecule are actually two different views of the same instance, they are naturally aligned to generate the self-supervisory signals for model training. Meanwhile, the same motif can exist in different molecules, and hence we also employ the contrastive loss to maximize the representation agreement of identical motifs across different molecules. Finally, in order to clearly identify the motifs that are critical in deciding the properties of each molecule, we

S. Chen is supported by National Natural Science Fund of China (No. 62506155), Provincial Natural Science Fund of Jiangsu (No. BK20251985), and Suzhou Municipal Leading Talents Fund (2025). C. Gong is supported by NSF of China (Nos. 62336003, 12371510). B. Han is supported by NSFC General Program (No. 62376235) and RGC General Research Fund (No. 12200725).

Authors' Contact Information: Wentao Yu, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China; e-mail: wentao.yu@njust.edu.cn; Shuo Chen (corresponding author), School of Intelligence Science and Technology, Nanjing University, Suzhou, China and RIKEN Center for Advanced Intelligence Project, Chuo-ku, Japan; e-mail: shuo.chen@nju.edu.cn; Chen Gong(corresponding author), School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai, China; e-mail: chen.gong@sjtu.edu.cn; Bo Han, Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China and RIKEN Center for Advanced Intelligence Project, Chuo-ku, Japan; e-mail: bhanml@comp.hkbu.edu.hk; Gang Niu, RIKEN Center for Advanced Intelligence Project, Chuo-ku, Japan; e-mail: gang.niu.ml@gmail.com; Masashi Sugiyama, RIKEN Center for Advanced Intelligence Project, Chuo-ku, Japan and Complexity Science and Engineering, The University of Tokyo, Bunkyo, Japan; e-mail: sugi@k.u-tokyo.ac.jp.

ACM 2157-6912/2026/2-ART35
https://doi.org/10.1145/3787204

further construct a property-aware attention mechanism into our learning framework. Our proposed AMCT is extensively evaluated on 10 popular benchmark datasets, and both quantitative and qualitative results firmly demonstrate its effectiveness when compared with the state-of-the-art methods.

CCS Concepts: • **Applied computing** → **Chemistry**; • **Computing methodologies** → *Supervised learning by classification*;

Additional Key Words and Phrases: Molecular property prediction, contrastive learning, graph transformer

ACM Reference format:
Wentao Yu, Shuo Chen, Chen Gong, Bo Han, Gang Niu, and Masashi Sugiyama. 2026. Atom-Motif Contrastive Transformer for Molecular Property Prediction. *ACM Trans. Intell. Syst. Technol.* 17, 2, Article 35 (February 2026), 20 pages.
https://doi.org/10.1145/3787204

1 Introduction

Molecular Property Prediction (MPP) [1] is a fundamental and critical task in many areas of basic science, such as chemistry, biomedicine, and pharmacology. The primary objective of MPP is to accurately predict the potential properties (e.g., toxicity and solubility) of a given molecule, and this prediction task is particularly challenging due to the intricate nature of interactions within molecules. As shown in Figure 1, depending on the specific prediction objective, MPP can be formulated as a regular classification task, where the goal is to assign a molecule to its predefined classes (e.g., toxic and non-toxic), or it can also be viewed as a regression task, aiming to predict continuous values (e.g., solubility) associated with the molecular properties.

In recent years, **Graph Transformer (GT)** models have been widely used for the MPP task due to their high reliability in characterizing the latent relationship among atoms [2–6]. However, most existing GT-based methods usually explore the basic interactions between pairwise atoms, so they fail to consider the important interactions among critical motifs (e.g., functional groups consisted of several atoms) of molecules. Consequently, these methods may fail to recognize critical patterns hidden in motifs and thus they are limited in their abilities to represent molecules. This is attributed to the fact that motifs represent meaningful subgraph patterns that consistently emerge with notable frequencies [7] and play an even more fundamental role in determining molecular properties when compared with atoms [5, 8, 9]. Furthermore, atom-level interactions are not always sufficient to predict molecular properties, and motif-level interactions are sometimes more important because they provide the detailed structural information of a molecule. For example, if we only consider atom-level interactions, phenol and cyclohexanol will have similar chemical properties, because they have the same molecular graph as shown in Figure 2. However, phenol is actually much more acidic than cyclohexanol. This is because the hydroxyl group in phenol interacts with the π electron cloud on the benzene ring [10], making it easier for phenol to release protons than cyclohexanol (we provide more details and present the motif interaction in phenol in the Supplementary Material 1).

In light of the above-mentioned considerations, we propose a novel **Atom-Motif Contrastive Transformer (AMCT)**, which not only explores the atom-level (*a.k.a.* low-level) interactions but also considers the motif-level (*a.k.a.* high-level) interactions. To incorporate two levels of interactions and enhance the molecular representation capability, we propose to build the atom-motif contrastive learning (as shown in Figure 3). First, considering that the representations of atoms and motifs for a given molecule are actually two different views of the same instance, they are naturally aligned during the training process. As such, they can jointly provide the self-supervisory signals and thereby improve the reliability of the learned molecular representation. Second, it is

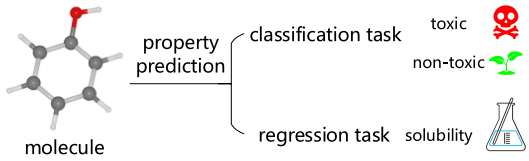


Fig. 1. MPP can be formulated as a classification task or a regression task, depending on the specific prediction objective.

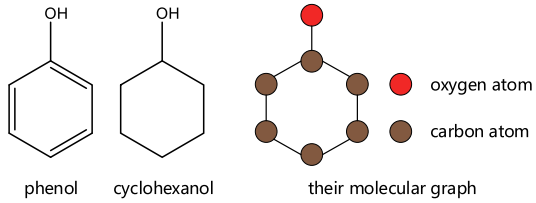


Fig. 2. Molecular formulae of phenol and cyclohexanol, as well as their corresponding molecular graph. As the hydrogen atom is neglected in the MPP task, phenol and cyclohexanol have the same molecular graph.

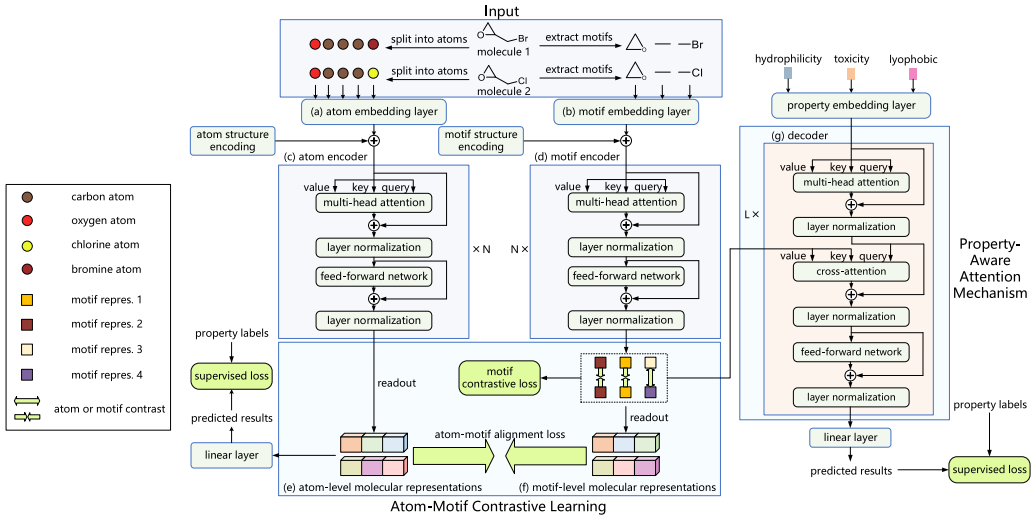


Fig. 3. The framework of our proposed AMCT. Given molecular inputs, atoms and motifs are encoded via embedding layer and encoder layer, respectively. The resulting atom-level and motif-level representations are then processed by a decoder and a linear layer to make predictions.

revealed that identical motifs across different molecules usually have similar chemical properties [11, 12]. For example, carbon rings with NO_2 or NH_2 functional groups tend to be mutagenic [13, 14]. It means that identical motifs should have consistent representations across all molecules. Therefore, we employ another contrastive loss to maximize the representation agreement of identical motifs across different molecules and thus obtain distinguishable motif representations.

Moreover, in order to clearly identify the motifs that are critical in deciding the properties of each molecule, we further construct a property-aware attention mechanism (see the right panel of Figure 3) by using the cross-attention module. Specifically, the cross-attention module calculates the cross-attention weights between the molecular property embeddings and the motif representations.

As a result, we can identify influential motifs based on the cross-attention weights. To the best of our knowledge, this is the first attempt to apply a property-aware attention mechanism in the GT model. Thanks to the effective atom-motif contrastive learning and the additional property-aware attention mechanism, the proposed AMCT can successfully extract informative features from molecules, therefore improving the performance in various MPP tasks. We compare our AMCT with 11 existing methods on 10 popular benchmark datasets, and both quantitative and qualitative results clearly demonstrate the superiority of our proposed method. To sum up, the contributions of our work are as follows:

- We propose a new AMCT to simultaneously explore the atom-level interactions and the motif-level interactions, so that we can successfully recognize critical patterns hidden in motifs and improve the reliability of MPP.
- We are the first to construct a property-aware attention mechanism in the GT model to identify the motifs that are critical in deciding the properties of each molecule.
- The experimental results on 10 popular datasets firmly demonstrate the superiority of our proposed AMCT when compared with the state-of-the-art methods.

2 Related Work

This section reviews the typical works related to this article, including MPP, GT model, and motif learning techniques.

2.1 MPP

In the past decades, numerous methods have been proposed for the MPP task. Initially, these methods mainly depend on hand-crafted features such as molecular descriptors and molecular fingerprints. However, these approaches may suffer from limited scalability and flexibility [15, 16]. To address these challenges, deep learning-based approaches, such as **Graph Neural Networks (GNNs)** and GT models, have been introduced to generate highly expressive molecular representations. For instance, Zhang et al. [8] pre-trained a GNN with a motif-based generation task. Although GNN-based approaches have shown promising results, they still have some limitations inherited from GNNs [17–26], such as over-smoothing, over-squashing, and limited expressiveness. To deal with these deficiencies, more GT-based MPP methods have been proposed. For example, Ying et al. [3] introduced the structural information from graphs into the classical transformer [27] and achieved promising results. Despite the noticeable achievements of GT-based MPP methods in recent years, most of them ignore the critical interactions among motifs. Therefore, in this article, we explore the atom-level and motif-level interactions, simultaneously.

2.2 GT

In contrast to the message-passing mechanism in a GNN [28, 29], which primarily aggregates local neighborhood information, a GT model possesses the ability to capture interactions between each pair of nodes through the self-attention mechanism. As an early attempt to generalize the transformer to the graph case, Dwivedi and Bresson [30] employed Laplacian eigenvectors as positional encodings of nodes and computed the corresponding attention weights based on the neighborhood connectivity of each node. To enhance the representation ability of GT models in a self-supervised manner, Zhang et al. [31] proposed to directly contrast two different graph views of the same instance, which shares the similar motivation with the typical works in self-supervised learning. Due to the success of contrastive learning in visual tasks [32], lots of graph contrastive learning methods have been proposed. Contrastive learning has also been introduced into the MPP task. For example, Fang et al. [33] contrasted molecular graphs generated by the knowledge-guided

graph augmentations. To incorporate two levels of interactions, we explore the self-supervision by our proposed atom-motif contrast in this article.

2.3 Motif Learning

Since motifs are significant subgraph patterns that appear consistently with remarkable frequencies [7], they usually play a vital role in graph learning. Yu and Gao [34] constructed a heterogeneous motif graph containing both motif nodes and molecular nodes. They then used the message-passing mechanism to learn the heterogeneous motif graph. Similarly, Wu et al. [5] proposed a heterogeneous molecular GT called Molformer integrating both the atom-level and motif-level nodes. Nevertheless, Molformer highly depends on the exact 3D coordinates of each molecule, which is usually quite hard to acquire. Furthermore, feeding two types of nodes into the same self-attention module poses a practical challenge in distinguishing interactions between different node types. In comparison, we separate atom-level and motif-level interactions by using the atom encoder and motif encoder, respectively. Meanwhile, we build the self-supervision of atom-motif contrast, so that we can conduct the representation learning without using any 3D coordinate information. Consequently, our proposed method actually has remarkable distinctions and superiorities when compared with Molformer.

3 Our Proposed Method

In this section, we will discuss the main technical details of our proposed AMCT, including atom and motif encoding, property-aware decoding, and our proposed loss functions.

3.1 Pipeline of Our Proposed Method

The framework of AMCT is shown in Figure 3. Given a batch of molecules, they are first split into a set of atoms and segmented into a set of motifs, respectively. Second, the atom embedding layer (Figure 3(a)) and the motif embedding layer (Figure 3(b)) generate atom embeddings and motif embeddings, respectively. Third, the atom encoder (Figure 3(c)) and the motif encoder (Figure 3(d)) are used to obtain atom-level and motif-level molecular representations (Figure 3(e) and (f)), respectively. Finally, the decoder (Figure 3(g)) and a linear layer are invoked to obtain the predicted results, where the atom-motif alignment loss, motif contrastive loss, and supervised losses are utilized for model training.

3.2 Atom Encoding

In the process of atom encoding, we first obtain atom embeddings. Then, we use the atom encoder to extract atom-level interactions. Finally, atom-level molecular representations are obtained after the readout operation.

3.2.1 Atom Embedding and Structure Encoding. Here we adopt the atom embedding method provided by the Open Graph Benchmark [35], which is widely acknowledged in the field of MPP. After atom embedding, in a molecule with n atoms, the ith atom is represented by a feature vector $\mathbf{x}_i \in \mathbb{R}^d$, where $i = 1, \ldots, n$ and d is the dimensionality of $\mathbf{x}_i$. Meanwhile, we use the degree centrality to encode the structural information among atoms, namely the connective relationship of atoms in the molecule. Therefore, they can be represented as $\mathbf{b}_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$. Since the degree centrality is applied to each atom, we simply add it to the atom embedding as in Graphormer [3]. Finally, given a molecule with n atoms, the combination of atom embedding and structural information for the ith atom is represented as $\mathbf{h}_i = \mathbf{x}_i + \mathbf{b}_i$ $(i = 1, \ldots, n)$, where $\mathbf{h}_i$ will be fed into the atom encoder.

3.2.2 Atom Encoder and Readout. The atom encoder (see Figure 3(c)) aims to capture the interactions among atoms. Given an input sequence of $\mathbf{h}_1, \ldots, \mathbf{h}_n$, we can represent them as a matrix $\mathbf{H} \in \mathbb{R}^{n \times d}$ with each row corresponding to an $\mathbf{h}_i$ $(i = 1, \ldots, n)$. The computation of **Multi-Head Attention (MHA)** in the atom encoder can be expressed as:

$$\begin{cases} \text{MultiHead}^{\text{atom}}(\mathbf{H}) = \text{Con}(\text{head}_1^{\text{atom}}, \ldots, \text{head}_h^{\text{atom}})\mathbf{W}^{\mathbf{O}}, \\ \text{head}_i^{\text{atom}} = \text{Attention}(\mathbf{H}\mathbf{W}_i^{\mathbf{Q}}, \mathbf{H}\mathbf{W}_i^{\mathbf{K}}, \mathbf{H}\mathbf{W}_i^{\mathbf{V}}), \\ \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\mathbf{Q}\mathbf{K}^\top / \sqrt{d_k}\right)\mathbf{V}, \end{cases} \tag{1}$$

where $\mathbf{W}_i^{\mathbf{Q}}, \mathbf{W}_i^{\mathbf{K}}, \mathbf{W}_i^{\mathbf{V}} \in \mathbb{R}^{d \times d_k}$, and $\mathbf{W}^{\mathbf{O}} \in \mathbb{R}^{hd_k \times d}$ are projection matrices. The notations $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the matrices of query, key, and value, respectively. Here h is the number of attention heads, and $d_k = d/h$ denotes the number of columns of $\mathbf{W}_i^{\mathbf{Q}}$, $\mathbf{W}_i^{\mathbf{K}}$, and $\mathbf{W}_i^{\mathbf{V}}$. Moreover, "Con" denotes the concatenation operation performed along the column. By sequentially passing through N layers, we acquire the atom representations. After that, we use a readout operation (namely, summation function) that aggregates the individual atom representations to generate the entire molecular graph representation (i.e., the atom-level molecular representation $\mathbf{h}^{\text{readout}} \in \mathbb{R}^d$).

3.3 Motif Encoding

Atom interactions successfully capture the low-level details, yet they ignore the high-level structural information among different molecules, so they are not sufficient to predict molecular properties in some cases (e.g., the example in Figure 2). Therefore, motif-level interactions are naturally introduced in our proposed AMCT. In the process of motif encoding, we first extract motifs and then obtain motif embeddings. After that, we use the motif encoder to extract motif-level interactions. Finally, motif-level molecular representations are obtained after the readout operation. Meanwhile, we use the atom-motif alignment loss and the motif contrastive loss to generate the self-supervisory signals.

3.3.1 Motif Extraction. Before describing the motif extraction process, let us first define the molecular graph formally. A molecular graph can be represented as $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V}$ is the node set containing all atoms, and $\mathcal{E}$ is the edge set describing the connectivity among the nodes (i.e., chemical bonds). By performing motif extraction, the molecular graph $\mathcal{G}$ can be decomposed into m motifs, denoted as $\mathcal{M}_1, \ldots, \mathcal{M}_m$. Each motif $\mathcal{M}_i = \langle \mathcal{S}_i, \mathcal{C}_i \rangle$ $(i = 1, \ldots, m)$ represents a subgraph defined by the atoms in $\mathcal{S}_i$ and the chemical bonds in $\mathcal{C}_i$. The combination of all motifs covers the entirety of the molecular graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V} = \bigcup_{i=1}^m \mathcal{S}_i$ and $\mathcal{E} = \bigcup_{i=1}^m \mathcal{C}_i$. In this article, we employ a dynamic and adaptive motif extraction method (i.e., MotifPiece [36]) to extract motifs. After motif extraction, a motif vocabulary is constructed by preprocessing all molecules in the dataset. For example, Figure 4 is the visualization of a subset of the motif vocabulary extracted from **Human Immunodeficiency Virus (*HIV*)** dataset. More details of the motif extraction process are presented in the Supplementary Material 2.

3.3.2 Motif Embedding and Structure Encoding. Here we use neural embedding vectors [37] to represent each motif. This is analogous to using word embeddings to represent words in natural language processing tasks. After motif embedding, in a molecule with m motifs, the ith motif is represented by a feature vector $\mathbf{t}_i \in \mathbb{R}^d$, where $i = 1, \ldots, m$. Meanwhile, we use the degree centrality to encode the structural information among motifs, namely the connective relationship of motifs in the molecule. Therefore, they can be represented as $\mathbf{e}_i \in \mathbb{R}^d$ for $i = 1, \ldots, m$. Finally, given a molecule with m motifs, the combination of motif embedding and structural information for the ith motif is represented as $\mathbf{z}_i = \mathbf{t}_i + \mathbf{e}_i$ $(i = 1, \ldots, m)$, where $\mathbf{z}_i$ will be fed into the motif encoder.

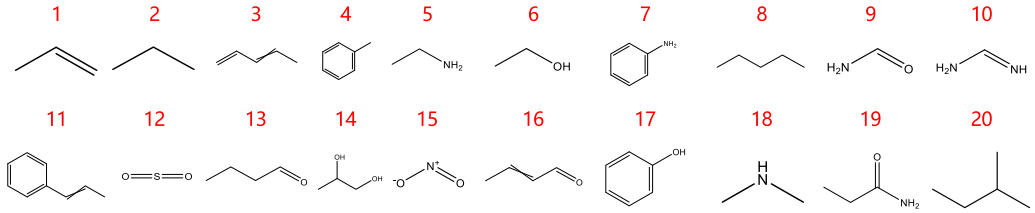


Fig. 4. The visualization of a subset of motif vocabulary extracted from *HIV* dataset by using MotifPiece [36]. Motifs are arranged in descending order based on their frequency of occurrence.

3.3.3 Motif Encoder and Readout. The motif encoder (see Figure 3(d)) is designed to capture the interactions among motifs. Given an input sequence of $\mathbf{z}_1, \ldots, \mathbf{z}_m$, we can represent them as a matrix $\mathbf{Z} \in \mathbb{R}^{m \times d}$ with each row corresponding to a $\mathbf{z}_i$ ($i = 1, \ldots, m$). By sequentially passing through N layers, we acquire the motif representations denoted as $\mathbf{Z}^{\text{motif}} \in \mathbb{R}^{m \times d}$, where each row denotes the representation of each motif. After the summation function for the readout operation, we obtain the motif-level molecular representation $\mathbf{z}^{\text{readout}} \in \mathbb{R}^d$ from $\mathbf{Z}^{\text{motif}}$.

3.3.4 Atom-Motif Alignment Loss. To explore the additional new supervisory information provided by the motifs, we consider the similarity relationship between the atom-level and motif-level molecular representations. Since the representations of atoms and motifs for a given molecule are actually two different views of the same instance, they are naturally aligned to generate the self-supervisory signals for model training. Here, in a batch of q molecules, the atom-motif alignment loss can be calculated as:

$$\mathcal{L}_{\text{align}} = \frac{1}{q} \sum_{i=1}^{q} T^2 \mathcal{L}_{\text{KL}} \left(\sigma\left(\frac{\mathbf{h}_i^{\text{readout}}}{T} \right), \sigma\left(\frac{\mathbf{z}_i^{\text{readout}}}{T} \right) \right), \tag{2}$$

where $\mathcal{L}_{\text{KL}}$ denotes the **Kullback-Leibler (KL)** divergence loss, σ denotes the softmax layer, and T is a hyperparameter controlling the softening effect. We set T to 4 throughout this article, which is a common setting in existing methods [38]. In Equation (2), $\mathbf{h}_i^{\text{readout}}$ and $\mathbf{z}_i^{\text{readout}}$ denote the atom-level and motif-level molecular representations of the ith molecule, respectively. By optimizing $\mathcal{L}_{\text{align}}$, AMCT can further enhance the molecular representation capability.

3.3.5 Motif Contrastive Loss. Since atom-motif alignment loss is performed intra-molecule and it constrains the consistency between atoms and motifs in the same molecule, we naturally seek to perform inter-molecule contrast and investigate the consistency across different molecules. Given that identical motifs across different molecules exhibit similar chemical properties [11, 12], it is expected that they should have similar representations across all molecules. To achieve this, we propose the motif contrastive loss, which maximizes the representation agreement of identical motifs across different molecules. Meanwhile, the representations of motifs belonging to different classes are pulled away. Specifically, the proposed motif contrastive loss can be defined as:

$$\mathcal{L}_{\text{contrastive}}^{\text{motif}}(\mathbf{Z}_{i:}^{\text{motif}}) = \log \frac{\sum_{k=1}^{l} \mathbb{1}_{[y_i = y_k]} e^{\langle \mathbf{Z}_{i:}^{\text{motif}}, \mathbf{Z}_{k:}^{\text{motif}} \rangle}}{\sum_{j=1}^{l} e^{\left\langle \mathbf{Z}_{i:}^{\text{motif}}, \mathbf{Z}_{j:}^{\text{motif}} \right\rangle}}, \tag{3}$$

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{l} \sum_{i=1}^{l} \mathcal{L}_{\text{contrastive}}^{\text{motif}}(\mathbf{Z}_{i:}^{\text{motif}}), \tag{4}$$

where l is the number of motifs in a batch of q molecules, $\mathbf{Z}_{i:}^{\text{motif}}$, $\mathbf{Z}_{j:}^{\text{motif}}$, and $\mathbf{Z}_{k:}^{\text{motif}}$ denote the ith, jth, and kth rows of $\mathbf{Z}^{\text{motif}}$, respectively. In Equation (3), the symbol $\langle \cdot \rangle$ expresses the inner product,

and $\mathbb{1}_{[\cdot]}$ is an indicator function which equals 1 if the argument inside the bracket holds, and 0 otherwise. In Equation (3), y_i and y_k are the motif labels of $\mathbf{Z}^{\text{motif}}_{i:}$ and $\mathbf{Z}^{\text{motif}}_{k:}$, respectively. In the proposed motif contrastive loss, the motifs belonging to identical/different classes are regarded as positive/negative pairs. By optimizing $\mathcal{L}_{\text{contrastive}}$, our proposed AMCT not only ensures the consistency in motif representations, but also improves the discriminating ability of learned motif representations.

3.4 Property-Aware Decoding

A good decoding process is also important to obtain the reliable representation. However, most existing GT models equally decode all properties without considering their importance. In contrast, we propose the property-aware decoding. Specifically, we first obtain property embeddings and then use the decoder to extract property-aware molecular representations. Finally, predicted results are obtained after the linear projection.

3.4.1 Property Embedding. Similar to motif embeddings, property embeddings are obtained via neural embedding vectors [37]. This process is analogous to using word embeddings to represent words in natural language processing tasks. Specifically, the property embeddings come from a learnable property embedding layer. Therefore, the property embeddings can be denoted as $\mathbf{P} \in \mathbb{R}^{c\times d}$, where c is the number of molecular property categories, and d is the number of dimensions.

3.4.2 Decoder and Linear Layer. The decoder (see Figure 3(g)) is designed to extract property-aware molecular representations. Given an input sequence of $\mathbf{p}_1, \ldots, \mathbf{p}_c$, we can represent them as a matrix $\mathbf{P} \in \mathbb{R}^{c\times d}$ with each row corresponding to a $\mathbf{p}_i$ ($i = 1, \ldots, c$). The matrix $\mathbf{P}$ is first fed into the MHA, by which we obtain $\mathbf{P}^{(1)} \in \mathbb{R}^{c\times d}$. Next, $\mathbf{P} + \mathbf{P}^{(1)}$ is operated by a layer normalization operation, by which we obtain $\mathbf{P}^{(2)}$. To identify the motifs that are critical in deciding the properties of each molecule, we construct a property-aware attention mechanism. Specifically, we employ a cross-attention module, which uses property embeddings $\mathbf{P}^{(2)}$ as queries and motif representations $\mathbf{Z}^{\text{motif}}$ as keys and values. To summarize, this process is computed as:

$$\begin{cases} \text{PAware}(\mathbf{Z}^{\text{motif}}, \mathbf{P}^{(2)}) = \text{Con}(\text{head}_1^{\text{cross}}, \ldots, \text{head}_h^{\text{cross}})\mathbf{W}^{\mathbf{O}}, \\ \text{head}_i^{\text{cross}} = \text{croatt}(\mathbf{P}^{(2)}\mathbf{W}_i^{\mathbf{Q}}, \mathbf{Z}^{\text{motif}}\mathbf{W}_i^{\mathbf{K}}, \mathbf{Z}^{\text{motif}}\mathbf{W}_i^{\mathbf{V}}), \\ \text{croatt}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{crossweight}(\mathbf{Q}, \mathbf{K})\mathbf{V}, \\ \text{crossweight}(\mathbf{Q}, \mathbf{K}) = \text{softmax}\left(\mathbf{Q}\mathbf{K}^\top/\sqrt{d_k}\right), \end{cases} \tag{5}$$

where "PAware" denotes the property-aware attention mechanism, "croatt" denotes the computation of the cross-attention module, and "crossweight" denotes the computation of cross-attention weights. The matrix of cross-attention weights is denoted as $\mathbf{A} \in \mathbb{R}^{c\times m}$, which indicates the strength of interactions between property embeddings and motif representations. Therefore, a motif with larger cross-attention weight is considered to have a greater contribution to the molecular property. After training, we normalize the elements of $\mathbf{A}$ to the range $[0, 1]$. We then identify the motifs that are critical in deciding the properties of each molecule by considering the cross-attention weights that exceed the threshold α, where $\alpha \in [0, 1]$ is a hyperparameter. By sequentially passing through L layers, we acquire the property-aware molecular representations. After the linear projection, the predicted results $\mathbf{o} \in \mathbb{R}^c$ are finally obtained. Note that our proposed property-aware attention mechanism is useful for both classification task and regression task. For classification tasks, our property-aware attention mechanism can identify the motifs in the specific molecule (e.g., ethanol) that are critical in deciding the toxicity property. For regression tasks, our proposed mechanism can identify the motifs in the specific molecule (e.g., acetic acid) that are critical in

deciding the solubility property. Specifically, the visualization results of the motifs identified by the property-aware attention mechanism are presented in Section 4.3.2.

3.5 Model Training

Since the supervised loss depends on the specific prediction objective, we adopt the cross-entropy loss for classification tasks, while we use the squared error loss for regression tasks. Given a batch of q molecules, we employ two supervised loss functions, where one for the matrix of predicted results $\mathbf{O} \in \mathbb{R}^{q\times c}$, and the other for atom-level molecular representations $\mathbf{H}^{\text{readout}}$. Since $\mathbf{H}^{\text{readout}} \in \mathbb{R}^{q\times d}$, we use a linear layer to obtain $\mathbf{H}^{\text{linear}} \in \mathbb{R}^{q\times c}$. The overall loss of our proposed AMCT is shown in Equation (6), which assembles the supervised loss $\mathcal{L}_{\text{sup}}(\mathbf{O})$, the supervised loss $\mathcal{L}_{\text{sup}}(\mathbf{H}^{\text{linear}})$, the atom-motif alignment loss $\mathcal{L}_{\text{align}}$, and the motif contrastive loss $\mathcal{L}_{\text{contrastive}}$, namely:

$$\mathcal{L} = \mathcal{L}_{\text{sup}}(\mathbf{O}) + \mathcal{L}_{\text{sup}}(\mathbf{H}^{\text{linear}}) + \lambda_{\text{a}}\mathcal{L}_{\text{align}} + \lambda_{\text{b}}\mathcal{L}_{\text{contrastive}}, \tag{6}$$

where $\lambda_{\text{a}} > 0$ and $\lambda_{\text{b}} > 0$ are hyperparameters adjusting the impacts of $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{contrastive}}$, respectively. In Equation (6), if we are given a classification task, then the two loss functions are:

$$\mathcal{L}_{\text{sup}}(\mathbf{O}) = -\frac{1}{qc}\sum\nolimits_{i=1}^{q}\sum\nolimits_{j=1}^{c}\mathbf{Y}_{ij}\ln\mathbf{O}_{ij}, \tag{7}$$

and

$$\mathcal{L}_{\text{sup}}(\mathbf{H}^{\text{linear}}) = -\frac{1}{qc}\sum\nolimits_{i=1}^{q}\sum\nolimits_{j=1}^{c}\mathbf{Y}_{ij}\ln\mathbf{H}^{\text{linear}}_{ij}, \tag{8}$$

respectively. If we are given a regression task, then they are:

$$\mathcal{L}_{\text{sup}}(\mathbf{O}) = \frac{1}{qc}\sum\nolimits_{i=1}^{q}\sum\nolimits_{j=1}^{c}(\mathbf{Y}_{ij} - \mathbf{O}_{ij})^2, \tag{9}$$

and

$$\mathcal{L}_{\text{sup}}(\mathbf{H}^{\text{linear}}) = \frac{1}{qc}\sum\nolimits_{i=1}^{q}\sum\nolimits_{j=1}^{c}(\mathbf{Y}_{ij} - \mathbf{H}^{\text{linear}}_{ij})^2, \tag{10}$$

respectively, where $\mathbf{Y} \in \mathbb{R}^{q\times c}$ is the ground truth matrix.

Remark. Here we discuss the efficiency and scalability of our proposed AMCT. On one hand, although our proposed AMCT contains several components (e.g., motif extraction and atom-motif contrastive learning), they can be efficiently implemented, as revealed by the efficiency analysis in Section 4.5. On the other hand, we emphasize that all components in our proposed AMCT are essential and non-negligible, which will also be verified by our comprehensive ablation studies presented in Section 4.3.1. Consequently, our proposed AMCT is easy to deploy in real-world applications, such as molecular design and drug discovery. Furthermore, AMCT can be easily scaled to multi-property tasks owing to its high efficiency.

4 Experiments

In this section, we conduct comprehensive experiments to evaluate the effectiveness and efficiency of our proposed AMCT. First, to ensure reproducibility and facilitate fair comparisons, we present comprehensive details of experimental settings, including implementation details, dataset statistics, and the descriptions of baseline methods. Second, to evaluate the effectiveness of our proposed AMCT, we compare AMCT with baseline methods on both classification tasks and regression tasks across 10 benchmark datasets. Third, to verify the effectiveness of each key component in our proposed AMCT, we perform ablation studies and provide visualization results, respectively. Fourth, to evaluate whether the performance of our proposed AMCT is sensitive to different values

of hyperparameters, we conduct parametric sensitivity analyses. Finally, to evaluate the efficiency of our proposed AMCT, we compare the inference time of AMCT with baseline methods.

4.1 Details of Experimental Settings

In this subsection, we provide comprehensive details of experimental settings to ensure reproducibility and facilitate fair comparisons. We first describe the implementation details of our proposed AMCT. Then, we introduce 10 benchmark datasets used in our experiments, along with their statistical properties. Finally, we introduce the compared baseline methods.

4.1.1 Implementation Details. To improve the usability of our proposed AMCT, we first build a unified pre-trained model on a large-scale dataset and then fine-tune it on other datasets. Specifically, we pre-train our model on *ZINC* dataset [39], which is a large-scale dataset consisting of about 250,000 molecules. After that, we fine-tune the pre-trained model on classification task and regression task, respectively. Furthermore, all experiments in this article are conducted on a Linux server with a 2.90 GHz Intel Xeon Gold 6326 CPU with 64 GB of RAM and an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. Our proposed method is implemented via Python 3.8.8, PyTorch 1.12.0, and PyTorch Geometric 2.1.0 with Adam optimizer. We use RDKit [40] to preprocess the Simplified Molecular-Input Line-Entry System [41] strings from various datasets.

4.1.2 Datasets. To validate the effectiveness of our proposed AMCT, we perform extensive experiments on 10 widely used benchmark datasets (i.e., *HIV*, *ToxCast*, *Tox21*, *BBBP*, *BACE*, *SIDER*, *ESOL*, *FreeSolv*, *Lipo*, and *PDBbind* [42]). These datasets include both classification task and regression task. The statistical information of the above benchmark datasets is described in Table 1. These benchmark datasets cover a wide range of domains, including biophysics, physiology, and physical chemistry. We utilize these datasets to predict various properties, such as (1) *HIV* replication inhibition in *HIV* dataset, (2) toxicological properties of 617 types in *ToxCast* dataset, (3) toxicity measurements such as nuclear receptors and stress response in *Tox21* dataset, (4) blood-brain barrier permeability in *BBBP* dataset, (5) inhibition to human β-secretase 1 in *BACE* dataset, (6) drug side effects of 27 system organ classes in *SIDER* dataset, (7) solubility in *ESOL* dataset, (8) hydration free energies in *FreeSolv* dataset, (9) lipophilic properties in *Lipo* dataset, and (10) protein-ligand binding affinity in *PDBbind* dataset. For all 10 datasets, we employ the scaffold splitting procedure [43], which is widely acknowledged in the field of MPP. For all experiments, we use the **Area under the ROC Curve (AUC)** as the evaluation metric (higher values are better) for classification tasks, and use the **Root Mean Square Error (RMSE)** as the evaluation metric (lower values are better) for regression tasks. To ensure a fair comparison, we calculate the mean AUC, mean RMSE, and the corresponding standard deviation over 10 independent runs for each method on each dataset. In addition, the paired t-test with significance level 0.05 is employed to statistically compare the results of various methods.

4.1.3 Baseline Methods. We compare our proposed AMCT with the following baseline methods, which can be categorized into three groups. The first group is composed of TF-Robust [44], which takes molecular fingerprints as the input. The second group consists of six GNN-based methods, namely Weave [45], GCN [46], SchNet [47], MGCN [48], MGSSL [8], and GREA-GCN [49]. The third group contains four GT-based methods, i.e., MAT [50], R-MAT [51], Graphormer [3], and Molformer [5].

4.2 Experimental Results on Classification Tasks and Regression Tasks

To evaluate the effectiveness of our proposed AMCT, we compare AMCT with baseline methods on both classification tasks and regression tasks across 10 benchmark datasets. Tables 2 and 3 show the

Table 1. Statistical Information of 10 Widely Used Benchmark Datasets

Dataset	# Graphs	# Average Nodes	# Max Nodes	# Average Edges	# Max Edges	# Classes	Task
HIV	41,127	25.5	222	54.9	502	1	Classification
ToxCast	8,576	18.8	124	38.5	268	617	Classification
Tox21	7,831	18.6	132	38.6	290	12	Classification
BBBP	2,039	24.1	132	51.9	290	1	Classification
BACE	1,513	34.1	97	73.7	202	1	Classification
SIDER	1,427	33.6	492	70.7	1,010	27	Classification
ESOL	1,128	13.3	55	27.4	124	-	Regression
FreeSolv	642	8.7	24	16.8	50	-	Regression
Lipo	4,200	27.0	115	59.0	236	-	Regression
PDBbind	19,310	33.0	224	35.0	247	-	Regression

Table 2. Experimental Results (i.e., AUC ↑) of Various Methods on Six Datasets for Classification Tasks

Methods	*ToxCast*	*Tox21*	*BBBP*	*BACE*	*SIDER*	*HIV*
TF-Robust [44]	0.585±0.031 ✓	0.698±0.012 ✓	0.860±0.087 ✓	0.824±0.022 ✓	0.607±0.033 ✓	0.634±0.036 ✓
Weave [45]	0.678±0.024 ✓	0.741±0.044 ✓	0.837±0.065 ✓	0.791±0.008 ✓	0.543±0.034 ✓	0.673±0.049 ✓
GCN [46]	0.650±0.025 ✓	0.772±0.041 ✓	0.877±0.036 ✓	0.854±0.011 ✓	0.593±0.035 ✓	0.760±0.012 ✓
SchNet [47]	0.679±0.021 ✓	0.767±0.025 ✓	0.847±0.024 ✓	0.750±0.033 ✓	0.545±0.038 ✓	0.702±0.034 ✓
MGCN [48]	0.663±0.009 ✓	0.707±0.016 ✓	0.850±0.064 ✓	0.734±0.030 ✓	0.552±0.018 ✓	0.738±0.016 ✓
MGSSL [8]	0.638±0.003 ✓	0.764±0.004 ✓	0.705±0.011 ✓	0.797±0.008 ✓	0.605±0.007 ✓	0.795±0.011 ✓
GREA-GCN [49]	0.658±0.006 ✓	0.785±0.008 ✓	0.679±0.018 ✓	0.732±0.035 ✓	0.569±0.023 ✓	0.762±0.019 ✓
MAT [50]	0.678±0.009 ✓	0.785±0.011 ✓	0.737±0.009 ✓	0.846±0.025 ✓	0.597±0.012 ✓	0.772±0.028 ✓
R-MAT [51]	0.685±0.009 ✓	0.791±0.009 ✓	0.745±0.010 ✓	0.858±0.041 ✓	0.600±0.013 ✓	0.786±0.025 ✓
Graphormer [3]	0.703±0.010 ✓	0.790±0.011 ✓	0.917±0.010 ✓	0.860±0.013 ✓	0.616±0.010 ✓	0.805±0.005 ✓
Molformer [5]	0.691±0.012 ✓	0.783±0.012 ✓	0.918±0.008 ✓	0.881±0.014 ✓	0.605±0.011 ✓	0.804±0.010 ✓
AMCT (Ours)	**0.735±0.011**	**0.816±0.013**	**0.937±0.014**	**0.922±0.010**	**0.641±0.011**	**0.827±0.013**

The best record in each column is bolded. The "✓" denotes that our AMCT is significantly better than existing methods revealed by paired t-test with significance level 0.05.

AUCs and RMSEs of different methods on 10 datasets, respectively, where the best record on each dataset has been highlighted in bold. According to Tables 2 and 3, we can find that our proposed AMCT achieves the best performance among the compared methods on all datasets. For example, on *FreeSolv* and *PDBbind*, AMCT outperforms the second-best method (i.e., Molformer) by a margin of 8.3% and 5.2%, respectively. Although Molformer also characterizes motif-level interactions, it overlooks the agreement between atom-level and motif-level molecular representations. Since our proposed AMCT enriches self-supervisory signals by contrasting atom-level representations and motif-level representations, it recognizes critical patterns hidden in motifs. Consequently, our proposed AMCT significantly surpasses other GT-based methods.

4.3 Verification of Our Proposed AMCT

To verify the effectiveness of each key component in our proposed AMCT, we first carry out ablation studies. Second, we provide visualization results to investigate whether the identified motifs by AMCT contribute to MPP.

4.3.1 Ablation Study. To evaluate the effectiveness of each key component in our proposed AMCT, here we perform ablation studies from three aspects. First, we perform ablation studies on our proposed losses and property-aware attention mechanism. Our proposed AMCT employs atom-motif alignment loss and motif contrastive loss to enrich the self-supervisory signals. In addition, we use the property-aware attention mechanism to improve the reliability of MPP. To shed

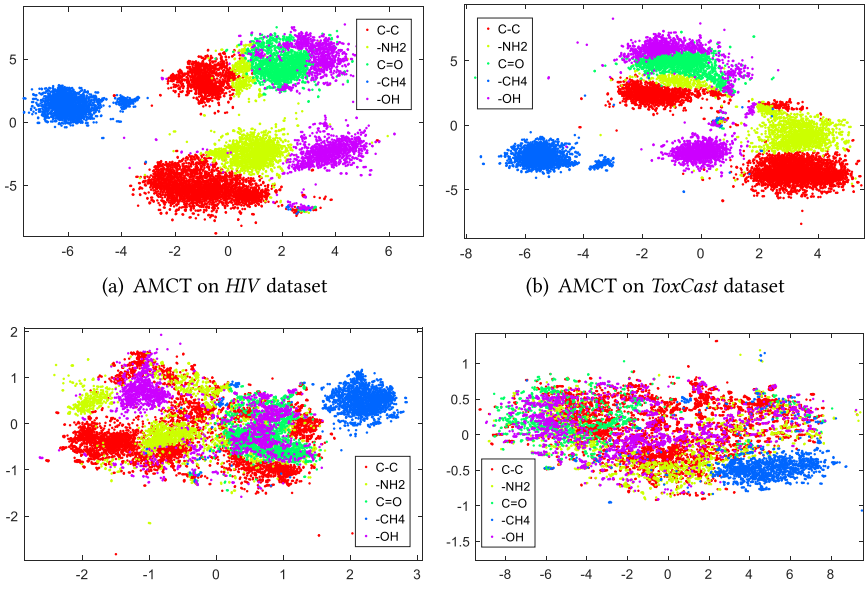
Table 3. Experimental Results (i.e., RMSE ↓) of Various Methods on Four Datasets for Regression Tasks

Methods	*ESOL*	*FreeSolv*	*Lipo*	*PDBbind*
TF-Robust [44]	1.722±0.038 ✓	4.122±0.066 ✓	0.909±0.039 ✓	1.512±0.027 ✓
Weave [45]	1.158±0.055 ✓	2.398±0.046 ✓	0.813±0.060 ✓	1.504±0.030 ✓
GCN [46]	1.068±0.050 ✓	2.870±0.140 ✓	0.851±0.082 ✓	1.413±0.025 ✓
SchNet [47]	1.045±0.064 ✓	3.221±0.764 ✓	0.912±0.103 ✓	1.430±0.058 ✓
MGCN [48]	1.266±0.147 ✓	3.352±0.014 ✓	1.111±0.044 ✓	1.441±0.046 ✓
MGSSL [8]	1.179±0.008 ✓	2.936±0.071 ✓	1.106±0.077 ✓	1.498±0.033 ✓
GREA-GCN [49]	0.897±0.036 ✓	1.829±0.368 ✓	0.786±0.036 ✓	1.406±0.029 ✓
MAT [50]	0.838±0.012 ✓	1.744±0.425 ✓	0.708±0.017 ✓	1.447±0.035 ✓
R-MAT [51]	0.833±0.015 ✓	1.912±0.364 ✓	0.685±0.029 ×	1.432±0.038 ✓
Graphormer [3]	0.829±0.014 ✓	2.210±0.014 ✓	0.901±0.012 ✓	1.419±0.031 ✓
Molformer [5]	0.848±0.013 ✓	1.145±0.017 ✓	0.740±0.012 ✓	1.303±0.036 ✓
AMCT (Ours)	**0.801±0.013**	**1.062±0.014**	**0.672±0.011**	**1.251±0.015**

The best record in each column is bolded. The "✓" ("×") denotes that our AMCT is significantly better (worse) than existing methods revealed by paired t-test with significance level 0.05.

light on the contributions of these components, we report the experimental results of AMCT when each of these components is removed on 10 datasets, which are shown in Table 4. For simplicity, "AMCT (w/o ALoss)," "AMCT (w/o CLoss)," and "AMCT (w/o PAware)" denote the reduced models by removing $\mathcal{L}_{\text{align}}$, $\mathcal{L}_{\text{contrastive}}$, and property-aware attention mechanism, respectively. It can also be observed that the performance decreases when either $\mathcal{L}_{\text{align}}$ or $\mathcal{L}_{\text{contrastive}}$ is removed, showing that both loss functions contribute significantly to satisfactory performance. In particular, AMCT is able to significantly improve the performance by using motif contrastive loss. For example, AUC can be decreased by more than 2% on *BACE* dataset without $\mathcal{L}_{\text{contrastive}}$. Meanwhile, we select five types of motifs and then visualize their motif representations obtained from AMCT and "AMCT (w/o CLoss)" on *HIV* and *ToxCast* datasets via using t-SNE method [52], respectively. As shown in Figure 5, the 2D projections of motif representations obtained from AMCT (see Figure 5(a) and (b)) show more compact clusters when compared with "AMCT (w/o CLoss)" (see Figure 5(c) and (d)). Therefore, it demonstrates that our proposed motif contrastive loss is beneficial for promising performance. Moreover, the performance decreases when removing the property-aware attention mechanism, which validates that it can help obtain reliable molecular representations. Second, we carry out an ablation study on motif extraction. Specifically, "AMCT (w/o motif)" refers to the reduced model that utilizes only atom-level information. In Table 4, we can find that the performance of "AMCT (w/o motif)" consistently degrades on 10 datasets, demonstrating the significance of the proposed motif-level interactions. Third, we employ the concatenation operation to replace $\mathcal{L}_{\text{align}}$. This variant method is termed as "AMCT (Concat.)." Experimental results in Table 4 demonstrate that the performance of "AMCT (Concat.)" is lower than "AMCT." This means that atom-motif alignment loss is obviously superior to the direct concatenation operation.

4.3.2 Visualization Results. To evaluate whether the identified motifs by our proposed AMCT are critical in deciding the properties of each molecule by using cross-attention weights, we visualize the motifs identified by our proposed AMCT (the threshold α is set to 0.5) in *HIV*, *ToxCast*, and *Tox21* datasets, respectively. Meanwhile, we also visualize atom attention weights learned from Graphormer. It is noteworthy that molecules are all selected from the test set of datasets. The



(a) AMCT on *HIV* dataset (b) AMCT on *ToxCast* dataset

(c) AMCT (w/o CLoss) on *HIV* dataset (d) AMCT (w/o CLoss) on *ToxCast* dataset

Fig. 5. The t-SNE visualizations of motif representations obtained from different models on two datasets.

Table 4. Ablation Study on 10 Datasets

Task Type	Dataset	AMCT (w/o ALoss)	AMCT (w/o CLoss)	AMCT (w/o PAware)	AMCT (w/o Motif)	AMCT (Concat.)	AMCT
Classification (AUC ↑)	*ToxCast*	0.692±0.022✓	0.689±0.018✓	0.681±0.020✓	0.633±0.036✓	0.664±0.019✓	**0.735±0.011**
	Tox21	0.785±0.016✓	0.782±0.018✓	0.781±0.025✓	0.738±0.029✓	0.768±0.023✓	**0.816±0.013**
	BBBP	0.904±0.017✓	0.902±0.020✓	0.861±0.016✓	0.741±0.048✓	0.783±0.023✓	**0.937±0.014**
	BACE	0.879±0.015✓	0.870±0.023✓	0.854±0.017✓	0.748±0.037✓	0.775±0.024✓	**0.922±0.010**
	SIDER	0.602±0.017✓	0.605±0.018✓	0.587±0.022✓	0.511±0.035✓	0.539±0.023✓	**0.641±0.011**
	HIV	0.790±0.016✓	0.788±0.020✓	0.771±0.022✓	0.722±0.042✓	0.782±0.026✓	**0.827±0.013**
Regression (RMSE ↓)	*ESOL*	1.229±0.047✓	1.153±0.060✓	1.198±0.045✓	1.589±0.044✓	1.338±0.048✓	**0.801±0.013**
	FreeSolv	1.494±0.056✓	1.525±0.042✓	1.685±0.059✓	2.035±0.051✓	1.847±0.066✓	**1.062±0.014**
	Lipo	1.218±0.032✓	1.337±0.036✓	1.543±0.039✓	1.754±0.058✓	1.496±0.057✓	**0.672±0.011**
	PDBbind	1.511±0.058✓	1.493±0.037✓	1.458±0.033✓	1.946±0.058✓	1.584±0.048✓	**1.251±0.015**

The best record in each row is bolded. The "✓" denotes that our AMCT is significantly better than reduced models revealed by paired t-test with significance level 0.05.

bluer the circles in the visualization results are, the larger the attention weights are. The additional visualization results are provided in the Supplementary Material 3.

HIV Dataset. Figure 6 shows the visualizations in *HIV* dataset. The task of *HIV* dataset is to predict whether a molecule will inhibit HIV replication or not. In the third example in Figure 6, we can observe that the fourth motif (namely, imidazole) has the highest cross-attention weight in motifs identified by AMCT. This makes sense because imidazole derivatives can easily bind to a variety of enzymes, proteins, and receptors in biological systems through various non-covalent interactions [53]. As a result, imidazole derivatives have potent biological activities, and several imidazole-containing compounds have been approved for the treatment of HIV infection or have been deployed in various stages of clinical trials [54]. In other words, the identified motif (namely,

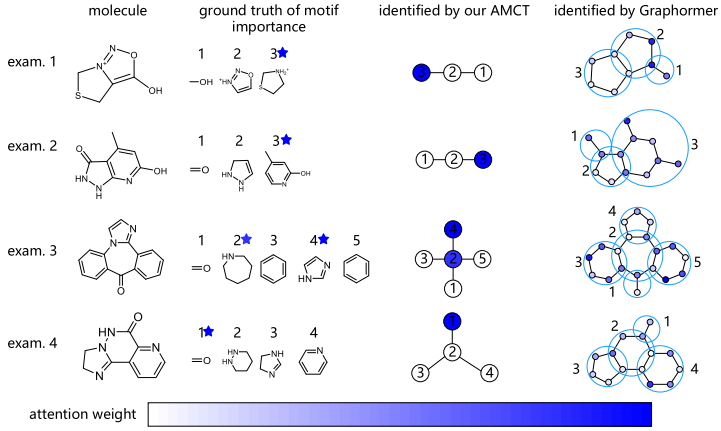


Fig. 6. Visualization of identified motifs of four molecules in *HIV* dataset. Their IDs in the dataset are 971, 3,134, 8,910, and 9,427, respectively. The blue stars in the second column indicate the motifs that are considered important in chemistry for deciding molecular properties. The blue hollow circles in the last column contain atoms in the corresponding motifs, and the numbers around them are their corresponding motif IDs.

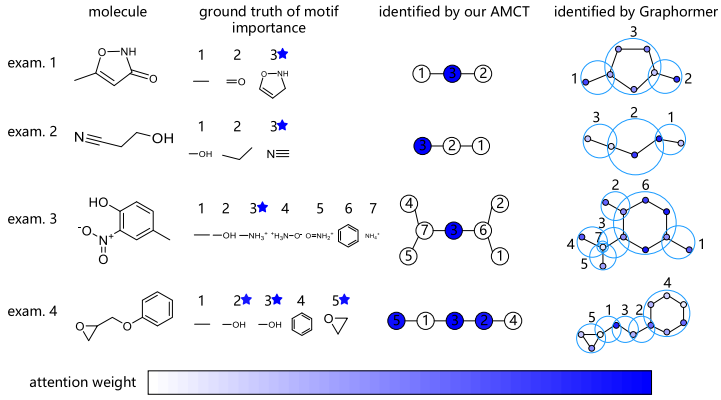


Fig. 7. Visualization of identified motifs of four molecules in *ToxCast* dataset. Their IDs in the dataset are 5, 631, 1,106, and 1,283, respectively. The blue stars in the second column indicate motifs that are considered important in chemistry for deciding molecular properties. The blue hollow circles in the last column contain atoms in the corresponding motifs, and the numbers around them are their corresponding motif IDs.

imidazole) is capable of inhibiting HIV replication. Consequently, this example firmly validates that our proposed property-aware attention mechanism is able to identify the motifs that are critical in deciding the properties of molecules. However, the attention weights of atoms in imidazole obtained from Graphormer are small, which means that Graphormer does not really identify motifs that are critical in deciding molecular properties. This is because Graphormer only considers atom-level interactions but neglects motif-level interactions. In contrast, owing to our proposed property-aware attention mechanism, identified motifs are critical to determining the properties of molecules.

ToxCast Dataset. Figure 7 shows the visualizations in *ToxCast* dataset. The task of *ToxCast* dataset is to predict whether a molecule is toxic or not. In the second example in Figure 7, we can observe that "N≡" (namely, cyanide) has the highest cross-attention weight in motifs identified by AMCT,

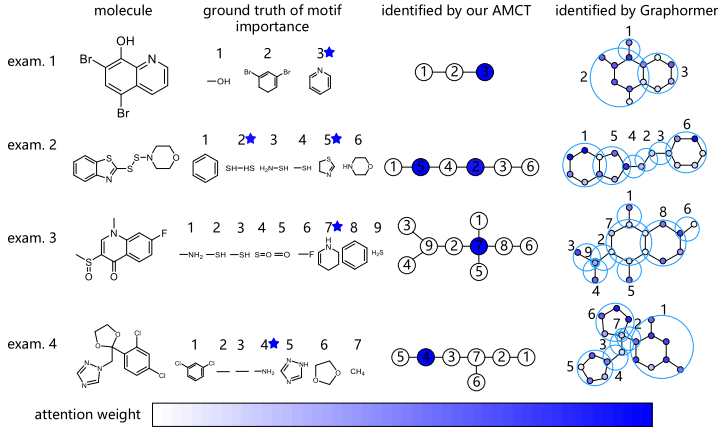
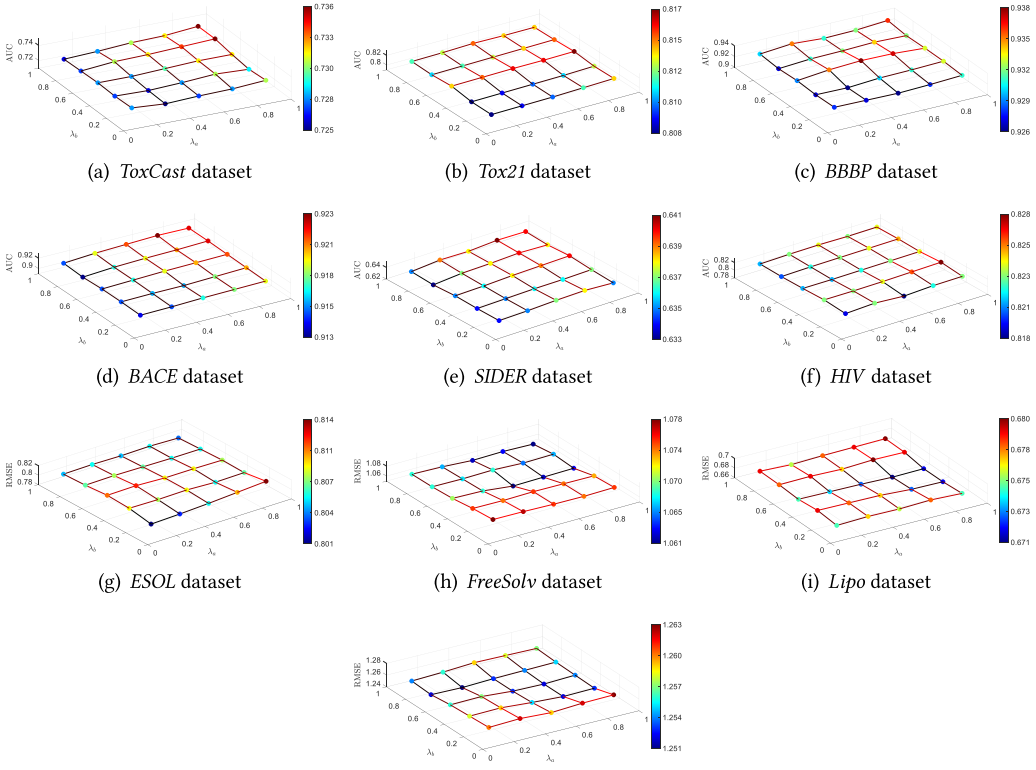


Fig. 8. Visualization of identified motifs of four molecules in *Tox21* dataset. Their IDs in the dataset are 890, 1,264, 1,386, and 2,643, respectively. The blue stars in the second column indicate motifs that are considered important in chemistry for deciding molecular properties. The blue hollow circles in the last column contain atoms in the corresponding motifs, and the numbers around them are their corresponding motif IDs.



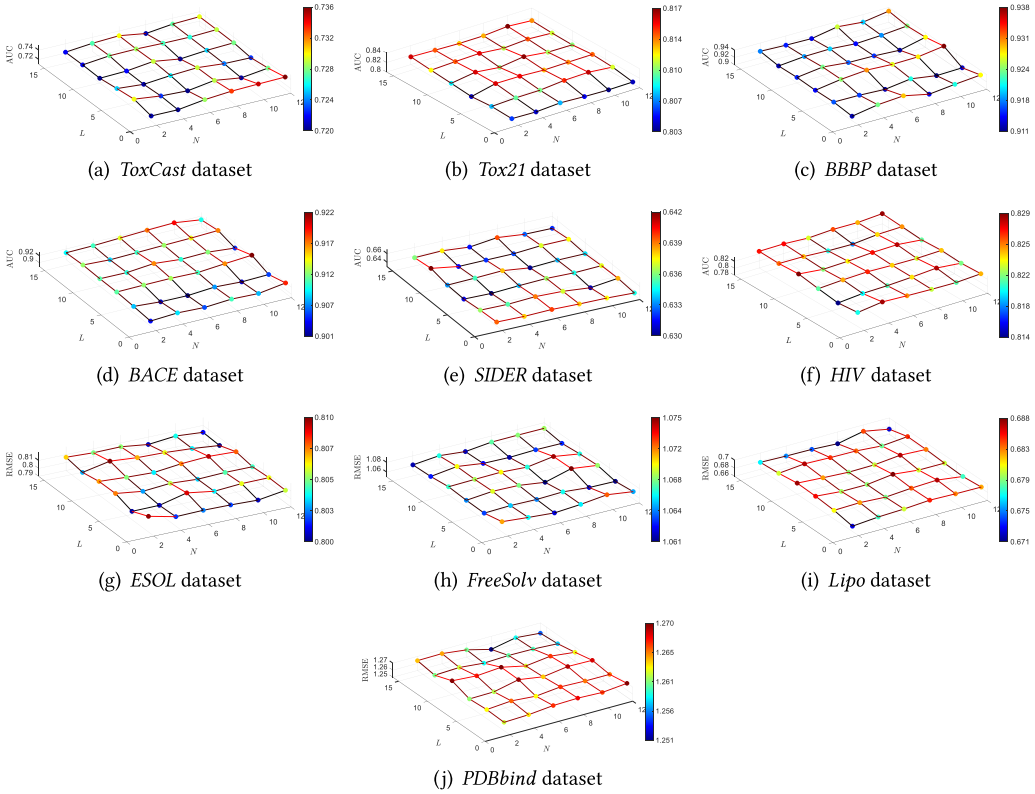
(a) *ToxCast* dataset (b) *Tox21* dataset (c) *BBBP* dataset

(d) *BACE* dataset (e) *SIDER* dataset (f) *HIV* dataset

(g) *ESOL* dataset (h) *FreeSolv* dataset (i) *Lipo* dataset

(j) *PDBbind* dataset

Fig. 9. The sensitivity analyses of hyperparameters λ_a and λ_b on 10 datasets.



(a) *ToxCast* dataset (b) *Tox21* dataset (c) *BBBP* dataset

(d) *BACE* dataset (e) *SIDER* dataset (f) *HIV* dataset

(g) *ESOL* dataset (h) *FreeSolv* dataset (i) *Lipo* dataset

(j) *PDBbind* dataset

Fig. 10. The sensitivity analyses of hyperparameters N and L on 10 datasets.

while other motifs such as "-" have the lowest cross-attention weights. This makes sense because "N≡" can bind to enzymes in cells, interfering with the transfer of oxygen in cellular respiration, leading to symptoms of suffocation [55]. Therefore, molecules with "N≡" are very likely to be toxic. However, in the second example in Figure 7, the attention weight of nitrogen atom learned from Graphormer is small, which means that Graphormer does not really identify motifs that are critical in determining molecular properties. This is because Graphormer only considers atom-level interactions but neglects motif-level interactions. In contrast, owing to our proposed property-aware attention mechanism, identified motifs are critical to deciding the properties of molecules.

Tox21 Dataset. Figure 8 shows the visualizations in *Tox21* dataset. The task of *Tox21* dataset is to predict whether a molecule is toxic or not. In the fourth example in Figure 8, we can observe that "NH_2" has the highest cross-attention weight in motifs identified by AMCT, while other motifs such as benzene ring have the lowest cross-attention weights. This makes sense because carbon rings with "NH_2" functional group tend to be mutagenic [13, 14]. As a result, molecules with "NH_2" are likely to be toxic. Therefore, this example validates that the identified motif is indeed critical in deciding the properties of molecules.

4.4 Sensitivity Analysis

To evaluate whether the performance of our proposed AMCT is sensitive to different values of hyperparameters, here we perform parametric sensitivity analyses. First, since there are two hyperparameters λ_a and λ_b in Equation (6), we report the performance of AMCT under different

Table 5. The Inference Time (Milliseconds) of Different Methods on 10 Datasets

	Classification Tasks						Regression Tasks			
Methods	*ToxCast*	*Tox21*	*BBBP*	*BACE*	*SIDER*	*HIV*	*ESOL*	*FreeSolv*	*Lipo*	*PDBbind*
TF-Robust [44]	2.572	2.593	2.785	2.736	2.793	2.644	2.336	1.787	1.318	3.491
Weave [45]	2.801	2.715	2.191	2.434	2.256	2.345	1.798	1.720	1.703	3.155
GCN [46]	2.298	2.179	2.240	2.401	2.139	2.219	2.270	2.074	2.130	2.987
SchNet [47]	2.653	2.711	2.674	2.425	2.311	2.533	1.724	1.881	1.555	3.170
MGCN [48]	2.696	2.774	2.645	2.734	2.293	2.151	2.493	1.850	2.198	4.913
MGSSL [8]	14.584	13.752	13.638	13.783	15.503	15.687	14.543	10.681	10.375	11.674
GREA-GCN [49]	3.580	3.296	3.143	3.456	3.720	3.708	3.604	3.308	3.154	4.835
MAT [50]	16.780	15.603	17.816	17.392	17.506	16.378	16.340	16.210	15.721	19.182
R-MAT [51]	17.574	18.107	17.686	17.391	16.194	16.800	17.166	16.472	16.816	22.634
Graphormer [3]	12.759	13.812	11.458	10.396	9.675	7.915	9.748	8.109	9.482	9.495
Molformer [5]	23.893	24.310	23.764	23.876	22.625	22.139	21.422	20.468	20.450	22.346
AMCT (Ours)	10.481	8.282	8.320	7.636	7.622	5.463	6.276	4.912	4.107	5.235
Average	9.389	9.178	9.030	8.888	8.720	8.332	8.310	7.456	7.417	9.426

values of them on 10 datasets, which are shown in Figure 9. We can observe that the performance of AMCT is relatively stable under different values of λ_a and λ_b. Second, to evaluate the impact of the number of layers on the performance of AMCT, we carry out a detailed sensitivity analysis on these hyperparameters (i.e., encoder layers N and decoder layers L). As shown in Figure 10, the variations in performance under different values of N and L are small, which confirms that the number of layers have little impact on the performance of our proposed AMCT. Furthermore, ablation studies in Section 4.3.1 validate that the improvement of performance in our proposed AMCT stems from our proposed method (i.e., atom-motif contrastive learning and property-aware attention mechanism).

4.5 Efficiency Analysis

To demonstrate the efficiency of our proposed AMCT, we report the inference time of our proposed AMCT and the compared baseline methods. According to Table 5, the inference time of our proposed method is significantly lower than the average level in most datasets. Moreover, it can be observed that the inference time of our proposed AMCT is competitive with most GT-based methods.

5 Conclusion

In this article, we proposed a novel AMCT for MPP, which simultaneously considers the atom-level and motif-level interactions within the molecule. The main advantage of our method is that it exploits the critical contrast information within atoms and motifs, and thus successfully builds supervisory signals for reliable model training. The proposed contrastive transformer was further integrated with a new property-aware attention mechanism, so that we could finally identify the critical motifs in deciding the molecular properties. Due to the effectiveness of the above contrastive transformer learning as well as the compatibility of the new attention block, our method achieved significantly better results than 11 representative approaches on 10 benchmark datasets.

References

[1] Junyu Lin, Yan Zheng, Xinyue Chen, Yazhou Ren, Xiaorong Pu, and Jing He. 2024. Cross-view contrastive unification guides generative pretraining for molecular property prediction. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1–9.

[2] Yukai Shi, Sen Zhang, Chenxing Zhou, Xiaodan Liang, Xiaojun Yang, and Liang Lin. 2021. GTAE: Graph transformer–based auto-encoders for linguistic-constrained text style transfer. *ACM Transactions on Intelligent Systems and Technology* 12, 3 (2021), 1–16.

[3] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do transformers really perform badly for graph representation? In *Proceedings of Advances in Neural Information Processing Systems*, 28877–28888.
[4] Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. 2022. Recipe for a general, powerful, scalable graph transformer. In *Proceedings of Advances in Neural Information Processing Systems*, 14501–14515.
[5] Fang Wu, Dragomir Radev, and Stan Z. Li. 2023. Molformer: Motif-based transformer on 3D heterogeneous molecular graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5312–5320.
[6] Adamo Young, Hannes Röst, and Bo Wang. 2024. Tandem mass spectrum prediction for small molecules using graph transformers. *Nature Machine Intelligence* 6, 4 (2024), 404–416.
[7] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. 2002. Network motifs: Simple building blocks of complex networks. *Science* 298, 5594 (2002), 824–827.
[8] Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. 2021. Motif-based graph self-supervised learning for molecular property prediction. In *Proceedings of Advances in Neural Information Processing Systems*, 15870–15882.
[9] Shichang Zhang, Ziniu Hu, Arjun Subramonian, and Yizhou Sun. 2024. Motif-driven contrastive learning of graph representations. *IEEE Transactions on Knowledge and Data Engineering* 36, 8 (2024), 4063–4075.
[10] Miguel Saggu, Nicholas M. Levinson, and Steven G. Boxer. 2011. Direct measurements of electric fields in weak oh$\cdots\pi$ hydrogen bonds. *Journal of the American Chemical Society* 133, 43 (2011), 17414–17419.
[11] Craig J. Hawker and Karen L. Wooley. 2005. The convergence of synthetic organic and polymer chemistries. *Science* 309, 5738 (2005), 1200–1205.
[12] Brian C. Smith. 2018. *Infrared Spectral Interpretation: A Systematic Approach*. CRC Press.
[13] Asim Kumar Debnath, Rosa L. Lopez de Compadre, Gargi Debnath, Alan J. Shusterman, and Corwin Hansch. 1991. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of Medicinal Chemistry* 34, 2 (1991), 786–797.
[14] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. 2020. Parameterized explainer for graph neural network. In *Proceedings of Advances in Neural Information Processing Systems*, 19620–19631.
[15] Philippe Schwaller, Daniel Probst, Alain C. Vaucher, Vishnu H. Nair, David Kreutter, Teodoro Laino, and Jean-Louis Reymond. 2021. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence* 3, 2 (2021), 144–152.
[16] Haoqiang Guo, Sendong Zhao, Haochun Wang, Yanrui Du, and Bing Qin. 2024. MolTailor: Tailoring chemical molecular representation to specific tasks via text prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 18144–18152.
[17] Yaoming Cai, Zijia Zhang, Pedram Ghamisi, Zhihua Cai, Xiaobo Liu, and Yao Ding. 2023. Fully linear graph convolutional networks for semi-supervised and unsupervised classification. *ACM Transactions on Intelligent Systems and Technology* 14, 3 (2023), 1–23.
[18] Jingjing Lin, Zhonglin Ye, Haixing Zhao, and Lusheng Fang. 2022. DeepHGNN: A novel deep hypergraph neural network. *Chinese Journal of Electronics* 31, 5 (2022), 958–968.
[19] Jing Bai, Wentao Yu, Zhu Xiao, Vincent Havyarimana, Amelia C. Regan, Hongbo Jiang, and Licheng Jiao. 2022. Two-stream spatial–temporal graph convolutional networks for driver drowsiness detection. *IEEE Transactions on Cybernetics* 52, 12 (2022), 13821–13833.
[20] Sheng Wan, Jian Yang, and Chen Gong. 2023. Advances of hyperspectral image classification based on graph neural networks. *Acta Electronica Sinica* 51, 6 (2023), 1687–1709.
[21] Wentao Yu, Sheng Wan, Guangyu Li, Jian Yang, and Chen Gong. 2023. Hyperspectral image classification with contrastive graph convolutional network. *IEEE Transactions on Geoscience and Remote Sensing* 61, 1 (2023), 1–15.
[22] Hang Zhou, Wentao Yu, Sheng Wan, Yongxin Tong, Tianlong Gu, and Chen Gong. 2024. Traffic pattern sharing for federated traffic flow prediction with personalization. In *Proceedings of the International Conference on Data Mining*, 1–10.
[23] Hang Zhou, Wentao Yu, Sheng Wan, Yongxin Tong, Tianlong Gu, and Chen Gong. 2025. FedTPS: Traffic pattern sharing for personalized federated traffic flow prediction. *Knowledge and Information Systems* 1, 1 (2025), 1–27.
[24] Wentao Yu, Shuo Chen, Yongxin Tong, Tianlong Gu, and Chen Gong. 2025. Modeling inter-intra heterogeneity for graph federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1–9.
[25] Wentao Yu. 2025. Homophily heterogeneity matters in graph federated learning: A spectrum sharing and complementing perspective. arXiv:2502.13732. Retrieved from https://arxiv.org/abs/2502.13732
[26] Wentao Yu, Chen Gong, Bo Han, Lixin Fan, and Qiang Yang. 2025. Integrating commonality and individuality for graph federated learning: A graph spectrum perspective. Retrieved from https://www.techrxiv.org/doi/full/10.36227/techrxiv.175502607.78991557

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*, 1–11.
[28] Guangyin Jin, Huan Yan, Fuxian Li, Yong Li, and Jincai Huang. 2023. Dual graph convolution architecture search for travel time estimation. *ACM Transactions on Intelligent Systems and Technology* 14, 4 (2023), 1–23.
[29] Xu Chen. 2024. Robust structure-aware graph-based semi-supervised learning: Batch and recursive processing. *ACM Transactions on Intelligent Systems and Technology* 15, 4 (2024), 1–25.
[30] Vijay Prakash Dwivedi and Xavier Bresson. 2021. A generalization of transformer networks to graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence Workshop on Deep Learning on Graphs: Methods and Applications*, 1–8.
[31] Chun-Yang Zhang, Wu-Peng Fang, Hai-Chun Cai, C. L. Philip Chen, and Yue-Na Lin. 2022. Sparse graph transformer with contrastive learning. *IEEE Transactions on Computational Social Systems* 11, 1 (2022), 892–904.
[32] Lingling Xu, Haoran Xie, Zongxi Li, Fu Lee Wang, Weiming Wang, and Qing Li. 2023. Contrastive learning models for sentence representations. *ACM Transactions on Intelligent Systems and Technology* 14, 4 (2023), 1–34.
[33] Yin Fang, Qiang Zhang, Haihong Yang, Xiang Zhuang, Shumin Deng, Wen Zhang, Ming Qin, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2022. Molecular contrastive learning with chemical element knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3968–3976.
[34] Zhaoning Yu and Hongyang Gao. 2022. Molecular representation learning via heterogeneous motif graph neural networks. In *Proceedings of International Conference on Machine Learning*, 25581–25594.
[35] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: Datasets for machine learning on graphs. In *Proceedings of Advances in Neural Information Processing Systems*, 22118–22133.
[36] Zhaoning Yu and Hongyang Gao. 2023. A data-driven approach for effective motif extraction and molecular representation learning. arXiv:2312.15387. Retrieved from https://arxiv.org/abs/2312.15387
[37] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2018. Junction tree variational autoencoder for molecular graph generation. In *Proceedings of International Conference on Machine Learning*, 2323–2332.
[38] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. 2021. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7028–7036.
[39] Teague Sterling and John J. Irwin. 2015. ZINC 15—Ligand discovery for everyone. *Journal of Chemical Information and Modeling* 55, 11 (2015), 2324–2337.
[40] A. Patrícia Bento, Anne Hersey, Eloy Félix, Greg Landrum, Anna Gaulton, Francis Atkinson, Louisa J. Bellis, Marleen De Veij, and Andrew R. Leach. 2020. An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics* 12, 51 (2020), 1–16.
[41] David Weininger, Arthur Weininger, and Joseph L. Weininger. 1989. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences* 29, 2 (1989), 97–101.
[42] Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. 2014. PDB-wide collection of binding data: Current status of the PDBbind database. *Bioinformatics* 31, 3 (2014), 405–412.
[43] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. 2018. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science* 9, 2 (2018), 513–530.
[44] Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, and Vijay Pande. 2015. Massively multitask networks for drug discovery. arXiv:1502.02072. Retrieved from https://arxiv.org/abs/1502.02072
[45] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. 2016. Molecular graph convolutions: Moving beyond fingerprints. *Journal of Computer-Aided Molecular Design* 30 (2016), 595–608.
[46] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of International Conference on Learning Representations*.
[47] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. 2017. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Proceedings of Advances in Neural Information Processing Systems*, 1–11.
[48] Chengqiang Lu, Qi Liu, Chao Wang, Zhenya Huang, Peize Lin, and Lixin He. 2019. Molecular property prediction: A multilevel quantum interactions modeling perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1052–1060.
[49] Gang Liu, Tong Zhao, Jiaxin Xu, Tengfei Luo, and Meng Jiang. 2022. Graph rationalization with environment-based augmentations. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1069–1078.
[50] Łukasz Maziarka, Tomasz Danel, Sławomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanisław Jastrzębski. 2020. Molecule attention transformer. arXiv:2002.08264. Retrieved from https://arxiv.org/abs/2002.08264
[51] Łukasz Maziarka, Dawid Majchrowski, Tomasz Danel, Piotr Gaiński, Jacek Tabor, Igor Podolak, Paweł Morkisz, and Stanisław Jastrzębski. 2024. Relative molecule self-attention transformer. *Journal of Cheminformatics* 16, 3 (2024), 1–14.

[52] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008), 2579–2605.
[53] Amita Verma, Sunil Joshi, and Deepika Singh. 2013. Imidazole: Having versatile biological activities. *Journal of Chemistry* 2013 (2013), 1–12.
[54] Cui Deng, Heng Yan, Jun Wang, Bao-Shan Liu, Kai Liu, and Yu-Min Shi. 2022. The anti-HIV potential of imidazole, oxazole and thiazole hybrids: A mini-review. *Arabian Journal of Chemistry* 15, 11 (2022), 1–16.
[55] Lewis Nelson. 2006. Acute cyanide toxicity: Mechanisms and manifestations. *Journal of Emergency Nursing* 32, 4 (2006), S8–S11.

Received 27 September 2024; revised 10 June 2026; accepted 8 December 2026