MTGIB-UNet: A Multi-Task Graph Information Bottleneck and Uncertainty Weighted Network for ADMET Prediction

Xuqiang Li[1,2,#], **Wenjie Du**[1,2,#], **Jun Xia**[3], **Jianmin Wang**[4], **Xiaoqi Wang**[5], **Yang Yang**[6,*], **Yang Wang**[1,2,*]
[1]University of Science and Technology of China (USTC), Hefei, China
[2]Suzhou Institute for Advanced Research, USTC, Suzhou, China
[3]Zhejiang University
[4]Yonsei University
[5]Northwestern Polytechnical University
[6]Shanghai Jiao Tong University
{xuqiangli, duwenjie}@mail.ustc.edu.cn, junxia@zju.edu.cn, jmwang113@hotmail.com, xqw@nwpu.edu.cn, yangyang@cs.sjtu.edu.cn, angyan@ustc.edu.cn

Abstract

Accurate prediction of ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties is crucial in drug development, as these properties directly impact a drug's efficacy and safety. However, existing multi-task learning models often face challenges related to noise interference and task conflicts when dealing with complex molecular structures. To address these issues, we propose a novel multi-task Graph Neural Network (GNN) model, **MTGIB-UNet**. The model begins by encoding molecular graphs to capture intricate molecular structure information. Subsequently, based on the Graph Information Bottleneck (GIB) principle, the model compresses the information flow by extracting subgraphs, retaining task-relevant features while removing noise for each task. These embeddings are then fused through a gated network that dynamically adjusts the contribution weights of auxiliary tasks to the primary task. Specifically, an uncertainty weighting (UW) strategy is applied, with additional emphasis placed on the primary task, allowing dynamic adjustment of task weights while strengthening the influence of the primary task on model training. Experiments on standard ADMET datasets demonstrate that our model outperforms existing methods. Additionally, the model shows good interpretability by identifying key molecular substructures related to specific ADMET endpoints.

1 Introduction

Computer-aided drug design (CADD) has emerged as a key area of focus at the intersection of artificial intelligence and scientific domains [Rentzsch *et al.*, 2019; Zhao *et al.*, 2020; Yang and Du, 2022]. In the costly drug discovery process, approximately half of clinical trial failures are attributed to an insufficient understanding of the absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of candidate drugs, which are crucial for their efficacy and safety as therapeutic agents [Norinder and Bergström, 2006; Feinberg *et al.*, 2020]. Poor pharmacokinetic properties and unacceptable potential characteristics, such as toxicity, pose significant risks to human health and the environment and are primary reasons for the exclusion of candidate drugs. In this context, ADMET prediction lays the foundation for drug-candidate selection, thereby supporting advancements in CADD [Du *et al.*, 2023a].

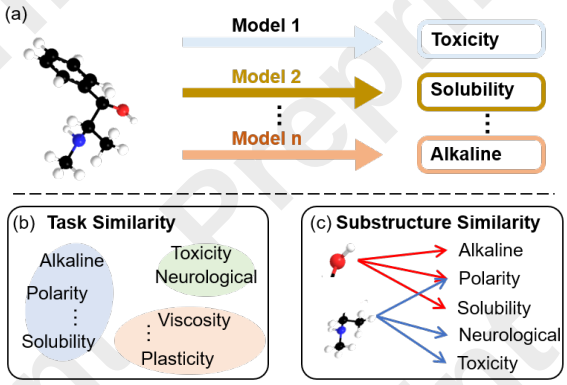

Figure 1: Illustrations of different notions. (a) single task; (b) task similarity based grouping; (c) substructure similarity based grouping. Best viewed in color.

Today, machine learning (ML) methods are widely employed for this purpose. Typically, such methods involve capturing compound structures by molecular descriptors such as SMILES strings, molecular topological graphs, molecular fingerprints, etc., and then analyzing them using elaborately designed algorithms in classification and regression tasks to explore potential quantitative structure-activity relationships (QSAR) [Fang *et al.*, 2024b; Du *et al.*, 2023b]. For example, CORAL utilized Monte Carlo models to encode

SMILES for predicting pIC50 values [Azimi *et al.*, 2023; Du *et al.*, 2024]. HRGCN+ utilized molecular graphs as the input to capture the intricate QSAR [Wu *et al.*, 2021]. However, when applied to multi-tasks scenarios, their efficacy diminishes significantly. One primary reason for this decline is the pervasive issue of optimization in multi-task prediction datasets [Zhang and Yang, 2021], manifesting in two major aspects.

Intrinsic connections in multi-tasks. In the realm of molecular field, many predictive models focus solely on individual molecular properties (Figure 1(a)), neglecting the interdependencies among these properties [Wang *et al.*, 2024; Xu *et al.*, 2025]. In reality, there are complex interactions between ADMET properties, while single-task models are limited in capturing [Kim *et al.*, 2024]. For instance, studies have shown that a drug's solubility is closely related to its lipophilicity, which in turn affects the drug's ability to permeate cell membranes [Gleeson *et al.*, 2011; Du *et al.*, 2023c]. This relationship directly influences the drug's absorption characteristics and subsequently impacts its distribution, metabolism, and excretion [Swanson *et al.*, 2023] (Figure 1 (b)). Although multitask learning has been effectively implemented in this field, previous studies ignore the potential conflicts between tasks during training, which may induce negative task interference and result in suboptimal model performance [Xu *et al.*, 2017; Guo *et al.*, 2018]. Dynamic weighting strategies appear to be a potential approach to mitigate optimization discrepancies between tasks [Sener and Koltun, 2018; Vandenhende *et al.*, 2021], but optimizing the weights for different tasks requires consideration of the true physical meaning of chemical tasks to be applicable for ADMET prediction.

Intrinsic connections in molecular substructure. Molecular representation plays a crucial role in determining the predictive performance of machine learning models [Gao *et al.*, 2024a]. A single molecule possesses multiple physicochemical properties, each of which is often influenced by distinct substructural fragments. For instance, the -NH2 group is frequently associated with toxicity, while the -COOH group exhibits hydrophilicity. These substructures hold significant chemical relevance, suggesting that identifying substructures unique to specific properties could enhance model interpretability and performance [Fang *et al.*, 2024a; Du *et al.*, 2025] (Figure 1 (c)). Furthermore, the similarity between different substructures may also provide insights into the underlying commonalities among related tasks, such as -OH is both alkaline and has a certain hydrophilicity.

In response to the limitations of existing models, we propose the **M**ulti-**T**ask **G**raph **I**nformation **B**ottleneck and **U**ncertainty Weighted **NET**work (MTGIB-UNet) for ADMET prediction, a novel framework that carefully considers inter-task and inter-subgraph correlations to enhance the performance of the primary task. This framework significantly improves the accuracy of ADMET predictions. As illustrated in Figure 2, MTGIB-UNet begins by grouping tasks based on their initial similarities. Following this, an improved Graph Information Bottleneck (GIB) theory is employed to capture the core substructures relevant to different tasks, which are then used to further refine the task groupings. Based on these criteria, a select number of auxiliary tasks are chosen to enhance the learning of the primary task. Finally, an uncertainty-weighting method is employed to balance the contributions of the primary and auxiliary tasks, leading to precise predictive outcomes. Our contributions in this work are summarized as follows:

- **Introduction of MTGIB-UNet:** We present the Multi-Task Graph Information Bottleneck and Uncertainty Weighted Network (MTGIB-UNet), a pioneering framework designed to capture inter-task and inter-subgraph correlations, enabling more accurate predictions across multiple ADMET properties.
- **Auxiliary Task Grouping:** Our approach integrates GIB theory into the model to effectively filter out irrelevant information. By combining this with task similarity, we optimize the selection of auxiliary tasks, and UW strategy is introduced to keep balance between tasks, thereby enhancing the overall performance of the model.
- **Validation through Extensive Experiments:** We conducted comprehensive experiments to validate the effectiveness of our proposed model on real-world ADMET datasets, demonstrating significant improvements over existing state-of-the-art methods.

2 Related Work

2.1 ADMET Property Prediction

In recent years, the rapid development of deep learning has significantly advanced ADMET property prediction. Initially, researchers employed various machine learning (ML) algorithms to predict ADMET properties [Kim *et al.*, 2024]. With the success of deep learning across various fields, deep learning models have shown strong potential in ADMET research [Tao and Abe, 2025]. For example, [Yang *et al.*, 2019] applied a GCNN model to a dataset from Amgen Inc., significantly improving the prediction accuracy for multiple endpoints. [De Carlo *et al.*, 2024] predict ADMET properties from molecule SMILES using a bottom-up approach with attention-based graph neural networks. These methods often fail to capture the complex interdependencies between ADMET properties, which is crucial for accurate prediction, we propose an improved method that integrates task weighting and grouping strategies in multitask learning to better capture these connections and enhance prediction accuracy.

2.2 Multitask Learning

Multitask learning (MTL) leverages knowledge from related tasks to improve performance, optimizing weights through a combined loss function of multiple tasks. While this can enhance predictions, training different tasks on a shared model can increase optimization challenges and risk negative transfer [Caruana, 1997]. Task weighting addresses this by assigning different weights to task losses, suppressing noise and promoting beneficial training signals. Homoscedastic uncertainty [Kendall *et al.*, 2018] and Dynamic Weight Averaging (DWA) [Liu *et al.*, 2019] are common methods to balance task losses and learning speeds. Task grouping can further mitigate these risks by balancing task losses and clustering tasks into subsets. Techniques such as differentiable

pruning [Gao *et al.*, 2024b] and status theory with maximum flow [Du *et al.*, 2023a] have been developed to optimize task grouping and model weights simultaneously. However, these methods still face limitations in addressing task conflicts. We build upon the GIB theory with an uncertainty weighting strategy to better manage inter-task dynamics and minimize negative impact.

2.3 Preliminaries

Problem Definition

Given a set of molecular compounds, the goal is to predict multiple ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties. These properties correspond to K endpoint tasks $T = \{t_1, t_2, \ldots, t_K\}$, where each task represents a specific ADMET property. These include both classification (binary) and regression (continuous) tasks.

Definition 1. *(IB-Graph)* Molecules are represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes (atoms) and $\mathcal{E}$ is the set of edges (bonds).

The GIB method [Yu *et al.*, 2021] identifies and extracts the most relevant substructures for a given task $\mathbf{Y}$. By compressing the graph $\mathcal{G}$ to obtaion the core subgraph $\mathcal{G}_{\text{IB}}$, which is determined by:

$$\mathcal{G}_{\text{IB}} = \arg\min_{\mathcal{G}_{\text{IB}}} - I(\mathbf{Y}; \mathcal{G}_{\text{IB}}) + \beta I(\mathcal{G}; \mathcal{G}_{\text{IB}}), \tag{1}$$

where $I(\cdot, \cdot)$ denotes the mutual information, and β is a Lagrangian multiplier that controls the trade-off between the prediction and compression.

Definition 2. *(UW)* The Uncertainty Weighting (UW) strategy [Kendall *et al.*, 2018] dynamically adjusts task weights in multi-task learning based on each task's uncertainty. This strategy utilizes a learnable parameter σ_k^2 for each task, modifying the loss calculation as follows:

$$\mathcal{L}_{\text{uncertainty}} = \sum_{k=1}^{K} \left(\frac{1}{2\sigma_k^2} \mathcal{L}_k + \delta_k \right), \tag{2}$$

where σ_k^2 adjusts the weight of each task's loss $\mathcal{L}_k$. The regularization term $\delta_k = \log\left(1 + \sigma_k^2\right)$ ensures stability by preventing the assignment of negative weights.

3 Methodology

The architecture of our model is illustrated in Figure 2. The model consists of five components: the Molecular Graph Encoder, Task-oriented Subgraph Extraction, Auxiliary Task Grouping, Task-Centered Gating, and Multi-task Prediction. The Molecular Graph Encoder processes the molecular graphs to obtain atomic embeddings using a three-layer GCN. The Task-oriented Subgraph Extraction module applies the GIB principle to extract task-specific subgraphs. Next, these subgraphs are then utilized in the Auxiliary Task Grouping module to identify and group tasks with similar subgraph features. The Task-Centered Gating module optimizes the contribution of the grouped auxiliary tasks to the primary task. Finally, the Multi-task Prediction module generates predictions for each task, integrating an uncertainty weighting strategy to dynamically adjust the task weights. Detailed explanation of each component is following:

3.1 Molecular Graph Encoder Module

Each node v_i is encoded as a 40-dimensional feature vector $\mathbf{x}_i \in \mathbb{R}^d$, where d is the dimensionality. The set of all node feature vectors constitutes the feature matrix $\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times d}$, where each row corresponds to a node's feature vector.The specific details of the node and edge features are provided in Appendix A.

The feature matrix $\mathbf{X}$ is then passed through three layers of GCN [Kipf and Welling, 2017] to update information:

$$\mathbf{H} = \texttt{GCN}(\mathbf{X}). \tag{3}$$

Here, $\mathbf{H}$ represents the updated graph.

3.2 Task-oriented Subgraph Extraction Module

We introduce the GIB theory to extract the task k's core substructure $\mathcal{G}_k$, as defined in Equation (1). Specifically, We optimize the following two items to obtain $\mathcal{G}_k$:

Minimizing: $-I\left(\mathbf{Y}; \mathcal{G}_k\right)$ We consider $P_\theta\left(\mathbf{Y} \mid \mathcal{G}_k\right)$ as the variational estimation of $P\left(\mathbf{Y} \mid \mathcal{G}_k\right)$. Thus, we derive:

$$\begin{aligned} I\left(\mathbf{Y}; \mathcal{G}_k\right) &\geq \mathbb{E}_{(\mathbf{Y}, \mathcal{G}_k)} \log \left[\frac{P_\theta\left(\mathbf{Y} \mid \mathcal{G}_k\right)}{P(\mathbf{Y})} \right] \\ &= \mathbb{E}_{(\mathbf{Y}, \mathcal{G}_k)} \log \left[P_\theta\left(\mathbf{Y} \mid \mathcal{G}_k\right) \right] + H(\mathbf{Y}) := \mathcal{L}_{pre}, \end{aligned} \tag{4}$$

where $H(\mathbf{Y})$ is a constant, and thus omitted in the model optimization process. The proof is given in Appendix B.

Minimizing: $I(\mathcal{G}_k; \mathcal{G})$ Following the information-bottleneck principle, we encourage the encoder to discard superfluous structures by perturbing node embeddings. For each node i with hidden state $\mathbf{H}_i$ we first obtain a retention logit:

$$s_i = \text{MLP}(\mathbf{H}_i), \quad p_i = \sigma(s_i), \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid. A Bernoulli mask $\lambda_i \sim \text{Bernoulli}(p_i)$ is then relaxed with the Gumbel-Sigmoid trick [Maddison *et al.*, 2017] to keep gradients:

$$\lambda_i = \sigma\Big(\frac{1}{t} \log \tfrac{p_i}{1-p_i} + \log \tfrac{u}{1-u}\Big), \qquad u \sim \mathcal{U}(0,1),\ t > 0. \tag{6}$$

The resulting perturbed embedding is

$$\mathbf{T}_i = \lambda_i \mathbf{H}_i + (1 - \lambda_i)\varepsilon, \qquad \varepsilon \sim \mathcal{N}(\mu_{\mathbf{H}}, \sigma_{\mathbf{H}}^2), \tag{7}$$

so uninformative nodes (small p_i) are replaced by Gaussian noise. This operation yields a pruned subgraph $\mathcal{G}_k$ that retains the salient information of $\mathcal{G}$.

Finally, the mutual information $I(\mathcal{G}_k; \mathcal{G})$ admits the following upper bound, a detailed proof is in Appendix B.2:

$$\begin{aligned} I(\mathcal{G}_k; \mathcal{G}) &\leq \mathbb{E}_{\mathcal{G}}[-\frac{1}{2} \log A + \frac{1}{2N} A + \frac{1}{2N} B^2] \\ &:= \mathcal{L}_{\text{MI}}(\mathcal{G}_k, \mathcal{G}), \end{aligned} \tag{8}$$

where $A = \sum_{j=1}^{N} (1 - \lambda_j)^2$ and $B = \frac{\sum_{j=1}^{N} \lambda_j (\mathbf{H} - \mu_{\mathbf{H}})}{\sigma_{\mathbf{H}}}$.

Therefore, the following objective function must be optimized:

$$\mathcal{L} = \mathcal{L}_{\text{pre}} + \beta \mathcal{L}_{\text{MI}}, \tag{9}$$

where $\mathcal{L}_{\text{pre}}$ is the cross-entropy (classification) or MSE (regression) loss, and $\mathcal{L}_{\text{MI}}$ is the Kullback-Leibler (KL) divergence between the extracted core subgraphs and random

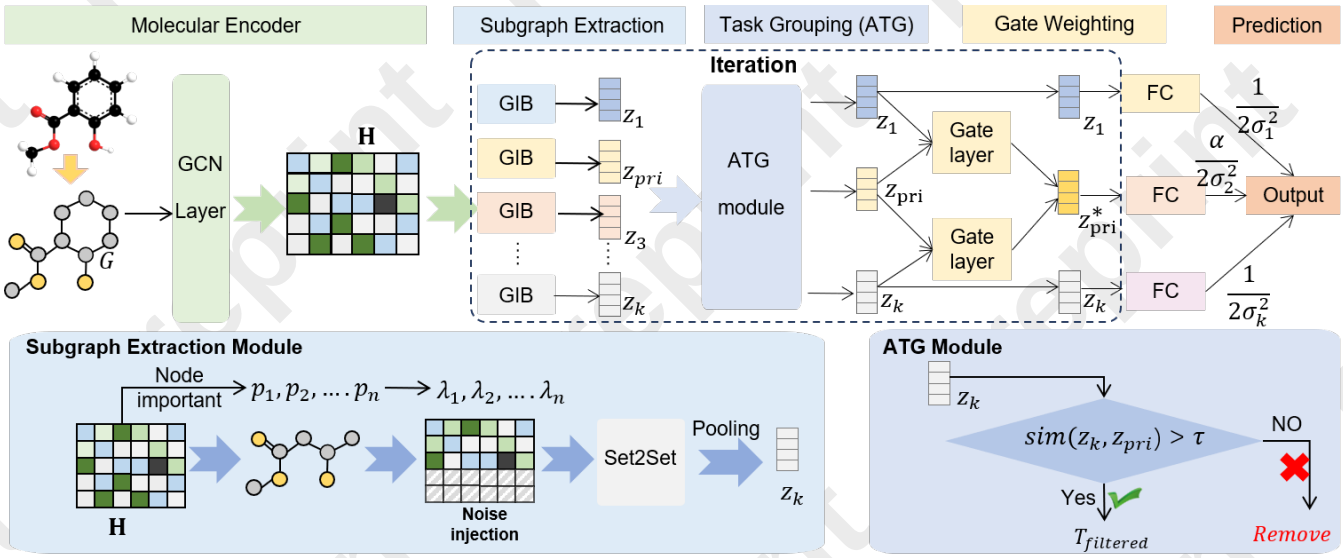

Figure 2: Overview of our proposed multi-task learning model. First, the **Molecular Graph Encoder** uses a three-layer Graph Convolutional Network (GCN) to encode molecular structures into atomic embeddings. Next, the **Subgraph Extraction** module applies the GIB to extract subgraphs specific to each task. These subgraphs are then used by the **Auxiliary Task Grouping** module to identify and group tasks with similar subgraph features, facilitating knowledge transfer. The Task-Centered **Gating module** optimizes the contribution of these grouped auxiliary tasks to the primary task through gating mechanisms. Finally, the Multi-task **Prediction module** generates predictions for each task, incorporating an uncertainty weighting strategy to dynamically adjust task weights, thereby enhancing the overall predictive performance. Best viewed in color.

noise graphs.The parameter β acts as a factor to balance the contributions of these two loss components. Finally, the subgraph $\mathcal{G}_k$ is pooled by the `Set2Set` network [Vinyals *et al.*, 2016], yielding the representation vector z_k:

$$z_k = \texttt{Set2Set}(\mathcal{G}_k). \tag{10}$$

This vector serves as a compact representation that encodes the essential information of the subgraph.

3.3 Auxiliary Task Grouping Module (ATG)

Task Similarity Calculation

First, we calculate the similarity between each auxiliary task and the primary task by comparing their core subgraph representations z_{pri} and z_k. The similarity score S_k could be calculated as:

$$\text{S}_k = \text{sim}(z_k, z_{\text{pri}}). \tag{11}$$

Here, $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity. A higher S_k indicates a greater similarity.

Auxiliary Task Selection

Tasks are ranked in descending order by their similarity scores S_k. A threshold $\tau = 0.7$ is applied to obtain a filtered set $\mathcal{T}_{\text{filtered}}$, which includes tasks with scores above the threshold:

$$\mathcal{T}_{\text{filtered}} = \{z_k \mid \text{S}_k \geq \tau\}. \tag{12}$$

To ensure an adequate number of auxiliary tasks, a minimum $m = 2$ is enforced. The final number of selected tasks N_{selected} is the maximum of m and the size of the filtered set:

$$N_{\text{selected}} = \max(m, |\mathcal{T}_{\text{filtered}}|). \tag{13}$$

The primary task representation z_{pri} and the selected auxiliary task representations z_k (where $k \in \mathcal{T}_{\text{filtered}}$) are then passed to the next module.

3.4 Task-Centered Gating Module

The task-centric module aims to optimize the contribution of each auxiliary task to the primary task. Each gating network in the module consists of a multi-layer feed-forward neural network and a sigmoid activation function. The input of each network is z_k of the auxiliary task, and it outputs two scalar weights: (1) θ_k, for the primary task and (2) $1 - \theta_k$, for the auxiliary task.

$$\theta_k = \texttt{Sigmoid}(\mathbf{W}_{\text{gate}} \cdot z_k), \quad k \in \mathcal{T}_{\text{filtered}} \tag{14}$$

The final embedding for the primary task is obtained by summing the weighted embeddings from all auxiliary tasks:

$$\hat{z}_{\text{pri}} = \sum_{k \in \mathcal{T}_{\text{filtered}}} (\theta_k z_k + (1 - \theta_k) z_{pri}). \tag{15}$$

This combined embedding $\hat{z}_{\text{pri}}$ is then passed through fully connected layers to predict the primary task. The embeddings of the other auxiliary tasks z_k remain unchanged.

3.5 Multi-task Prediction Module

Each task's embedding z_k is processed through a task-specific three-layer fully-connected network to generate the corresponding predictions:

$$y_{pri} = \text{FC}_{pri}(\hat{z}_{\text{pri}}), \quad y_k = \text{FC}_k(z_k). \tag{16}$$

This formula captures the process of generating the final predicted value for each task. These predicted values are then used to calculate the task-specific losses.

Uncertainty Weighting
Considering potential optimizing conflicts in multi-task training, an uncertainty weighting strategy is applied, as described in Definition 2. This strategy adjusts the weight of each task dynamically, reflecting the uncertainty associated with each task.

Loss Function
The total loss $\mathcal{L}_{\text{total}}$ for multi-task learning is calculated by aggregating the losses from all tasks. The loss $\mathcal{L}$ for each task is defined as shown in Equation (9). After introducing the uncertainty weighting strategy, the total loss function is expressed as:

$$\mathcal{L}_{\text{total}} = \left(\frac{\alpha}{2\sigma_{\text{pri}}^2}\right)\mathcal{L}^{\text{pri}} + \sum_{k\in\mathcal{T}_{\text{filtered}}}\left(\frac{1}{2\sigma_k^2}\mathcal{L}^k\right) + \sum_{k\in\{\text{pri}\}\cup\mathcal{T}_{\text{filtered}}}\delta_k, \tag{17}$$

where α scales the primary-task loss, σ_k^2 are learnable uncertainty parameters that adjust the weight of each task's loss based on its uncertainty, and $\delta_k = \log\left(1+\sigma_k^2\right)$ acts as a regularization term to maintain the stability of the learning process. $\mathcal{L}^{\text{pri}}$ and $\mathcal{L}^k$ represent the losses for the primary task and auxiliary tasks, respectively.

4 Experiment

4.1 Experimental Setup

Datasets. In this study, we utilized the dataset constructed by [Du *et al.*, 2023a], consisting of 24 endpoints, including 18 classification tasks and 6 regression tasks, collected from 8 published studies related to ADMET properties. It comprises 43,291 drug-like compounds, with 28,153 compounds involved in classification tasks and 16,545 in regression tasks. Depending on the specific tasks, each molecular may be associated with one or more endpoint labels [Yang *et al.*, 2018; Wang *et al.*, 2020; Delaney, 2004].
Evaluation Metrics. For the regression tasks, including Caco-2 permeability, PPB, LD50, IGC50, ESOL, and logD7.4, the coefficient of determination (R^2) is used as the evaluation metric, covering a total of 6 tasks. For the remaining 18 classification tasks, including HIA, OB, P-gp inhibitor, P-gp substrates, BBB, CYP1A2 inhibitor, and others, the area under the receiver operating characteristic curve (ROC-AUC) is used as the evaluation metric.
Baselines. We compare our model with several state-of-the-art models including: Uni-Mol [Zhou *et al.*, 2023], ChemBFN [Tao and Abe, 2025], and MTGL [Du *et al.*, 2023a], MT-GCN [Montanari *et al.*, 2019] and MGA [Peng *et al.*, 2020], along with single-task variants like ST-GCN and ST-MGA [Montanari *et al.*, 2019; Peng *et al.*, 2020]. Detailed descriptions are in Appendix C.

Implementation Details
In the model implementation, the input feature dimension is set to 40, and the core computations are performed via three GCN layers of dimensions 64, 128, and 128.

During training, the Adam optimizer was employed with a learning rate of 0.001 and a weight decay of 10^{-5}. The batch size was set to 128, and the model was trained for 200 epochs. Additionally, the UW coefficient for the primary task, α, was set to 1.2. All experiments were conducted on a Tesla V100-PCIE-16GB GPU. Each experiment was repeated eight times, with the mean and variance reported.

4.2 Model Performance

Table 1 presents the quantitative results for ADMET property prediction using various baseline models. MTGIB-UNet consistently outperformed other models, achieving superior results in both classification and regression tasks. Specifically, MTGIB-UNet demonstrated significant improvements across 14 classification tasks and 5 regression tasks, surpassing the second-best model by 0.79 % and 0.85 %, respectively. Overall, MTGIB-UNet secured the first position in 19 out of 24 tasks and the second position in the remaining 5 tasks.

In particular, for the BBB endpoint, MTGIB-UNet improved the ROC-AUC score from 0.973 to 0.979, representing a 0.62 % increase compared to MTGL. For the IGC50 endpoint, the ROC-AUC score was elevated from 0.819 to 0.824, marking a 0.61 % improvement. When considering all endpoints, MTGIB-UNet achieved an overall performance enhancement of 1.03% compared to MTGL. Notably, when compared to single-task models like ST-GCN and ST-MGA, MTGIB-UNet demonstrated an average improvement of 7.32 % to 4.28 %, underscoring its superiority within a multi-task learning framework.

4.3 Ablation Studies

We conduct a series of ablation studies to evaluate the contribution of various components in our proposed model, specifically the Task-oriented Subgraph Extraction Module, Uncertainty Weighting Analysis, and Auxiliary Task Grouping Module. The primary results of these studies are summarized in Table 2, more detailed results provided in Appendix D.

Impact of the Task-oriented Subgraph Extraction Module
This module constitutes a critical component of our model. To assess the impact of accurate subgraph extraction on the target prediction properties, we conducted experiments by excluding the GIB module from the model. The results, presented in Table 2, reveal a notable degradation in model performance across both classification and regression tasks when this module is removed.

Impact of Uncertainty Weighting (UW)
The UW strategy is designed to dynamically adjust task weights during training, prioritizing tasks based on their uncertainty. We conducted an ablation study by setting uniform task weights to 1, as detailed in Table 2. The results clearly demonstrate that the absence of UW leads to decreased model performance, especially on tasks with higher uncertainty. This underscores the importance of adaptive task weighting in improving overall model efficacy.

Impact of the Auxiliary Task Grouping (ATG) Module
The ATG module is a crucial component of our multi-task learning model. To evaluate its impact, we performed an ablation study by removing the ATG strategy and treating auxiliary tasks as independent, without grouping based on their re-

Endpoint	Metric	ST-GCN	ST-MGA	MT-GCN	MGA	MTGL	ChemBFN	Uni-Mol	MTGIB-UNet	Imp(↑)
HIA	ROC-AUC	$0.916_{(0.054)}$	$0.972_{(0.014)}$	$0.899_{(0.057)}$	$0.911_{(0.034)}$	$0.981_{(0.011)}$	$0.944_{(0.010)}$	**0.987**$_{(0.012)}$	$0.985_{(0.006)}$	5.317
OB	ROC-AUC	$0.716_{(0.035)}$	$0.710_{(0.035)}$	$0.728_{(0.031)}$	$0.745_{(0.029)}$	$0.749_{(0.022)}$	$0.753_{(0.024)}$	$0.720_{(0.023)}$	**0.761**$_{(0.003)}$	4.023
P-gp inhibitor	ROC-AUC	$0.916_{(0.012)}$	$0.917_{(0.006)}$	$0.895_{(0.014)}$	$0.901_{(0.010)}$	$0.928_{(0.008)}$	$0.922_{(0.011)}$	$0.894_{(0.009)}$	**0.939**$_{(0.007)}$	3.138
P-gp substrates	ROC-AUC	$0.775_{(0.034)}$	$0.755_{(0.014)}$	$0.733_{(0.044)}$	$0.719_{(0.035)}$	$0.801_{(0.031)}$	$0.774_{(0.034)}$	$0.726_{(0.032)}$	**0.809**$_{(0.015)}$	7.193
BBB	ROC-AUC	$0.956_{(0.008)}$	$0.959_{(0.004)}$	$0.945_{(0.007)}$	$0.956_{(0.010)}$	$0.973_{(0.005)}$	$0.931_{(0.009)}$	$0.950_{(0.007)}$	**0.979**$_{(0.003)}$	2.744
CYP1A2 inhibitor	ROC-AUC	$0.932_{(0.007)}$	$0.931_{(0.013)}$	$0.914_{(0.009)}$	$0.940_{(0.006)}$	$0.952_{(0.005)}$	$0.947_{(0.008)}$	$0.951_{(0.006)}$	**0.959**$_{(0.002)}$	2.223
CYP2C19 inhibitor	ROC-AUC	$0.774_{(0.012)}$	$0.781_{(0.008)}$	$0.775_{(0.011)}$	$0.795_{(0.019)}$	$0.804_{(0.015)}$	$0.780_{(0.020)}$	$0.793_{(0.017)}$	**0.806**$_{(0.004)}$	2.545
CYP2C9 inhibitor	ROC-AUC	$0.746_{(0.016)}$	$0.764_{(0.017)}$	$0.771_{(0.016)}$	$0.798_{(0.019)}$	$0.794_{(0.019)}$	$0.722_{(0.025)}$	$0.733_{(0.021)}$	**0.802**$_{(0.005)}$	5.368
CYP2D6 inhibitor	ROC-AUC	$0.848_{(0.016)}$	$0.841_{(0.022)}$	$0.839_{(0.015)}$	**0.877**$_{(0.017)}$	$0.869_{(0.016)}$	$0.800_{(0.019)}$	$0.860_{(0.017)}$	$0.871_{(0.007)}$	2.747
CYP3A4 inhibitor	ROC-AUC	$0.892_{(0.006)}$	$0.915_{(0.006)}$	$0.865_{(0.007)}$	$0.875_{(0.006)}$	$0.916_{(0.007)}$	$0.878_{(0.014)}$	$0.914_{(0.009)}$	**0.920**$_{(0.006)}$	2.958
Half life	ROC-AUC	$0.725_{(0.011)}$	$0.708_{(0.024)}$	$0.688_{(0.035)}$	$0.707_{(0.017)}$	$0.727_{(0.022)}$	$0.715_{(0.030)}$	$0.709_{(0.026)}$	**0.738**$_{(0.011)}$	3.756
Clearance	ROC-AUC	$0.723_{(0.030)}$	$0.710_{(0.015)}$	$0.686_{(0.031)}$	$0.740_{(0.027)}$	$0.779_{(0.027)}$	$0.768_{(0.032)}$	$0.707_{(0.029)}$	**0.781**$_{(0.017)}$	6.924
Hepatotoxicity	ROC-AUC	$0.653_{(0.040)}$	$0.669_{(0.022)}$	$0.612_{(0.039)}$	$0.713_{(0.053)}$	$0.701_{(0.036)}$	$0.697_{(0.041)}$	$0.660_{(0.037)}$	**0.725**$_{(0.031)}$	7.864
Respiratory toxicity	ROC-AUC	$0.842_{(0.018)}$	**0.872**$_{(0.013)}$	$0.810_{(0.014)}$	$0.828_{(0.021)}$	$0.859_{(0.010)}$	$0.818_{(0.028)}$	$0.865_{(0.015)}$	$0.867_{(0.006)}$	2.969
Cardiotoxicity-1	ROC-AUC	$0.707_{(0.026)}$	$0.703_{(0.020)}$	$0.683_{(0.028)}$	$0.684_{(0.023)}$	$0.740_{(0.023)}$	$0.745_{(0.027)}$	$0.761_{(0.024)}$	**0.765**$_{(0.015)}$	6.610
Cardiotoxicity-5	ROC-AUC	$0.620_{(0.015)}$	$0.637_{(0.010)}$	$0.626_{(0.027)}$	$0.623_{(0.014)}$	$0.641_{(0.014)}$	$0.630_{(0.022)}$	**0.677**$_{(0.019)}$	$0.653_{(0.007)}$	2.627
Cardiotoxicity-10	ROC-AUC	$0.627_{(0.013)}$	$0.611_{(0.015)}$	$0.609_{(0.022)}$	$0.603_{(0.026)}$	$0.654_{(0.010)}$	$0.642_{(0.014)}$	$0.658_{(0.012)}$	**0.663**$_{(0.004)}$	5.381
Cardiotoxicity-30	ROC-AUC	$0.664_{(0.036)}$	$0.653_{(0.036)}$	$0.645_{(0.036)}$	$0.709_{(0.035)}$	$0.723_{(0.029)}$	$0.704_{(0.035)}$	$0.709_{(0.032)}$	**0.728**$_{(0.016)}$	6.012
Caco-2 permeability	R^2	$0.451_{(0.033)}$	$0.519_{(0.014)}$	$0.374_{(0.022)}$	$0.385_{(0.031)}$	$0.523_{(0.025)}$	$0.443_{(0.027)}$	$0.510_{(0.026)}$	**0.527**$_{(0.023)}$	15.101
PPB	R^2	$0.577_{(0.028)}$	$0.585_{(0.004)}$	$0.589_{(0.036)}$	$0.568_{(0.038)}$	$0.626_{(0.029)}$	$0.537_{(0.030)}$	$0.562_{(0.028)}$	**0.637**$_{(0.005)}$	10.262
LD50	R^2	$0.588_{(0.018)}$	$0.617_{(0.018)}$	$0.503_{(0.017)}$	$0.492_{(0.029)}$	$0.635_{(0.015)}$	$0.622_{(0.020)}$	$0.622_{(0.017)}$	**0.640**$_{(0.011)}$	9.831
IGC50	R^2	$0.703_{(0.055)}$	$0.818_{(0.021)}$	$0.618_{(0.027)}$	$0.772_{(0.021)}$	$0.819_{(0.008)}$	$0.797_{(0.014)}$	**0.832**$_{(0.012)}$	$0.824_{(0.003)}$	7.632
ESOL	R^2	$0.814_{(0.030)}$	$0.896_{(0.013)}$	$0.824_{(0.030)}$	$0.866_{(0.020)}$	$0.931_{(0.038)}$	$0.907_{(0.042)}$	$0.920_{(0.039)}$	**0.935**$_{(0.006)}$	6.285
logD7.4	R^2	$0.759_{(0.056)}$	$0.904_{(0.008)}$	$0.770_{(0.019)}$	$0.838_{(0.018)}$	$0.915_{(0.008)}$	$0.863_{(0.025)}$	$0.796_{(0.022)}$	**0.922**$_{(0.007)}$	10.419

Table 1: Model Performance on ADMET datasets (The best result is in bold, the second-best is underlined. Imp: The improvement relative to the baseline average results. %).

latedness. This configuration effectively represents a single-task strategy. The results, summarized in Table 2, indicate a noticeable decline in performance across all tasks when ATG is removed, highlighting the significance of strategic auxiliary task grouping in enhancing model accuracy.

To further investigate the role of the ATG module, we shuffled the learned groupings from the current model in Appendix D. This additional analysis helps to validate the importance of the ATG module.

	Classification (R^2)		
	PPB	IGC50	ESOL
w/o ATG	$0.624_{(0.014)}$	$0.817_{(0.012)}$	$0.929_{(0.011)}$
w/o GIB	$0.631_{(0.006)}$	$0.812_{(0.016)}$	$0.926_{(0.015)}$
w/o UW	$0.628_{(0.013)}$	$0.818_{(0.007)}$	$0.930_{(0.014)}$
Ours	**0.637**$_{(0.005)}$	**0.824**$_{(0.003)}$	**0.935**$_{(0.006)}$
	Regression (ROC-AUC)		
	P-gp substrates	Cardiotoxicity-1	Cardiotoxicity-5
w/o ATG	$0.792_{(0.009)}$	$0.739_{(0.013)}$	$0.643_{(0.012)}$
w/o GIB	$0.796_{(0.014)}$	$0.736_{(0.014)}$	$0.641_{(0.010)}$
w/o UW	$0.791_{(0.015)}$	$0.740_{(0.010)}$	$0.640_{(0.014)}$
Ours	**0.809**$_{(0.015)}$	**0.765**$_{(0.015)}$	**0.653**$_{(0.007)}$

Table 2: Ablation Study Results.

4.4 Sensitivity Analysis

In this section, we explore the impact of various hyperparameters on the performance of our model.

First, we examine the effect of the parameter β in the GIB layer, which governs the trade-off between preserving task-specific information and reducing redundant information. When β is set to lower values, the model tends to retain more noise, leading to suboptimal performance. On the other hand, higher values of β result in excessive compression of subgraphs, leading to the loss of critical information and, consequently, diminished performance. Optimal performance is achieved at $\beta = 10^{-3}$, where the balance between noise reduction and information preservation is well-maintained, as illustrated in Table 3.

We also conducted a sensitivity analysis on α, which adjusts the weighting of the primary task relative to auxiliary tasks. A higher α increases the model's focus on the primary task, thereby enhancing its influence on overall learning and performance. As α increases, the model's performance initially improves due to this enhanced focus. However, setting α too high leads to diminishing returns, where the model may overly prioritize the primary task at the expense of valuable insights from auxiliary tasks, potentially resulting in a decline in overall performance, as shown in Table 3.

Overall, our sensitivity analysis highlights that carefully chosen values of α and β are critical for optimizing the performance of our model. These parameters significantly influence the model's ability to effectively capture core subgraphs and maintain a balance between different tasks.

4.5 Qualitative Analysis

In this section, we evaluate the interpretability and effectiveness of our model on the primary task Caco-2 permeability and its auxiliary tasks OB and ESOL. We examine the dy-

	Regression (R^2)		Classification (ROC-AUC)	
β	**IGC50**	**logD7.4**	**BBB**	**HIA**
0	$0.782_{(0.010)}$	$0.875_{(0.012)}$	$0.949_{(0.013)}$	$0.956_{(0.009)}$
1e-5	$0.806_{(0.005)}$	$0.902_{(0.008)}$	$0.960_{(0.004)}$	$0.969_{(0.004)}$
1e-4	$0.819_{(0.006)}$	$0.913_{(0.014)}$	$\mathbf{0.979}_{(0.003)}$	$0.980_{(0.015)}$
1e-3	$\mathbf{0.824}_{(0.003)}$	$\mathbf{0.922}_{(0.007)}$	$0.971_{(0.008)}$	$\mathbf{0.985}_{(0.006)}$
1e-1	$0.803_{(0.014)}$	$0.898_{(0.011)}$	$0.967_{(0.008)}$	$0.973_{(0.010)}$
α	**IGC50**	**logD7.4**	**BBB**	**HIA**
1.0	$0.793_{(0.014)}$	$0.907_{(0.012)}$	$0.970_{(0.011)}$	$0.978_{(0.014)}$
1.2	$\mathbf{0.824}_{(0.003)}$	$\mathbf{0.922}_{(0.007)}$	$\mathbf{0.979}_{(0.003)}$	$\mathbf{0.985}_{(0.006)}$
1.5	$0.813_{(0.013)}$	$0.911_{(0.012)}$	$0.973_{(0.010)}$	$0.980_{(0.014)}$
3.0	$0.805_{(0.012)}$	$0.909_{(0.013)}$	$0.968_{(0.011)}$	$0.977_{(0.013)}$
5.0	$0.792_{(0.014)}$	$0.901_{(0.015)}$	$0.962_{(0.012)}$	$0.972_{(0.011)}$

Table 3: Sensitivity analysis for β and α.

namic changes in task weights and total loss over epoch, and analyze the extracted subgraphs for compounds.

As shown in Figures 3 (a) and (b), we examine the core substructures of the compounds clonidine and propylthiouracil across three different tasks. It is evident that the core substructures extracted for the primary task, Caco-2 permeability, bear a closer resemblance to those for the auxiliary task OB. Both tasks are related to important absorption properties, highlighting the role of auxiliary tasks in enhancing the primary task. Interestingly, for the propylthiouracil molecule, the substructures extracted for the primary task, Caco-2 permeability, are more similar to those for the auxiliary task ESOL. The model accurately identifies significant groups like sulfonamide and amide, suggesting that similar substructures can also contribute to enhancing the primary task. Therefore, the enhancement of the primary task is jointly determined by the type of auxiliary tasks and the similarity of the core substructures extracted by GIB.

The model's dynamic weights during training are evident in Figure 3 (c), where the weight of the primary task, Caco-2 permeability, increases rapidly in the early stages, reflecting the model's prioritization of this task. Simultaneously, the auxiliary tasks OB and ESOL receive refined weight adjustments, ensuring balanced multi-task learning. As shown in Figure 3 (d), the total loss decreases consistently over time, with a sharp decline in the first 50 epochs, indicating the model's effective learning and adaptability across tasks.

PPB Classification (R^2)				
Model	**Params (M)**	**Time (min)**	**Memory (G)**	**Performance**
MGA	12.62	11.37	1.27	$0.568_{(0.038)}$
MTGL	0.066	9.25	1.04	$0.626_{(0.029)}$
ChemBFN	54.50	13.65	15.74	$0.537_{(0.030)}$
Uni-Mol	47.61	10.55	7.40	$0.562_{(0.028)}$
Ours	0.77	56.53	1.30	$\mathbf{0.637}_{(0.005)}$

Table 4: Computational resource usage and classification performance on the PPB dataset. **Params (M)** denotes the number of learnable parameters (in millions).

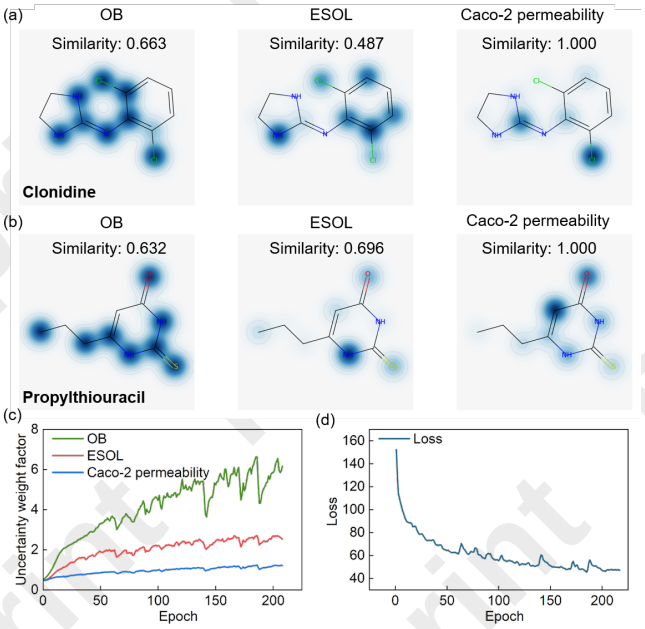

Figure 3: Comparison of different paradigms for Gibbs free energy prediction. (a) Method by concatenation; (b) method by merging; (c) a schematic diagram of the process where acetonitrile (solute) is dissolved in ethanol (solvent); and (d) the illustration of our method. Best viewed in color.

4.6 Resource Usage Analysis

In this section, we evaluate the computational resource usage of our model on the PPB dataset. As shown in Table 4, although the proposed method requires longer training time compared to other baselines, the total number of parameters remains relatively small. Meanwhile, the performance shows a noticeable improvement over other approaches, indicating that our design achieves a favorable balance between model complexity and predictive accuracy.

5 Conclusion and Broader Impact

In this paper, we address the intricate challenge of ADMET property prediction, a pivotal aspect of drug discovery and development. We introduce a novel multi-task GNN model, MTGIB-UNet, that integrates UW and auxiliary learning strategies to enhance predictive accuracy across multiple tasks. Our model functions in two stages: the first stage employs a grouping method based on structural and task similarity, to identify auxiliary task groups for each primary task. The second stage employs a multi-task learning framework incorporating a graph information bottleneck and UW strategies. The graph information bottleneck effectively filters out irrelevant information by constraining the information flow, thereby reducing noise interference, while the UW strategy dynamically adjusts task weights to mitigate potential conflicts between tasks. Our experimental results demonstrate that MTGIB-UNet outperforms existing models in the ADMET prediction domain, underscoring the effectiveness of our integrated approach. Meanwhile, there remains potential for further refinement as in Appendix E.

Acknowledgements

This paper is partially supported by the Project of Stable Support for Youth Team in Basic Research Field, CAS (YSBR005), Natural Science Foundation of China National Major Research Instrument Development Project (No.12227901), the Innovation Program for Quantum Science and Technology (2021ZD0303303), the National Natural Science Foundation of China (22025304, 22033007, 62072427).

Contribution Statement

Xuqiang Li and Wenjie Du are co-frst authors; Yang Wang and Yang Yang are corresponding authors.

References

[Azimi *et al.*, 2023] Atena Azimi, Shahin Ahmadi, Ashwani Kumar, Mahnaz Qomi, and Ali Almasirad. Smiles-based qsar and molecular docking study of oseltamivir derivatives as influenza inhibitors. *Polycyclic Aromatic Compounds*, 43(4):3257–3277, 2023.

[Caruana, 1997] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

[De Carlo *et al.*, 2024] Alessandro De Carlo, Davide Ronchi, Marco Piastra, Elena Maria Tosca, and Paolo Magni. Predicting admet properties from molecule smile: A bottom-up approach using attention-based graph neural networks. *Pharmaceutics*, 16(6):776, 2024.

[Delaney, 2004] John S Delaney. Esol: estimating aqueous solubility directly from molecular structure. *Journal of chemical information and computer sciences*, 44(3):1000–1005, 2004.

[Du *et al.*, 2023a] Bing-Xue Du, Yi Xu, Siu-Ming Yiu, Hui Yu, and Jian-Yu Shi. Mtgl-admet: a novel multi-task graph learning framework for admet prediction enhanced by status-theory and maximum flow. In *International Conference on Research in Computational Molecular Biology*, pages 85–103. Springer, 2023.

[Du *et al.*, 2023b] Wenjie Du, Fenfen Ma, Baicheng Zhang, Jiahui Zhang, Di Wu, Edward Sharman, Jun Jiang, and Yang Wang. Spectroscopy-guided deep learning predicts solid–liquid surface adsorbate properties in unseen solvents. *Journal of the American Chemical Society*, 146(1):811–823, 2023.

[Du *et al.*, 2023c] Wenjie Du, Xiaoting Yang, Di Wu, Fen-Fen Ma, Baicheng Zhang, Chaochao Bao, Yaoyuan Huo, Jun Jiang, Xin Chen, and Yang Wang. Fusing 2d and 3d molecular graphs as unambiguous molecular descriptors for conformational and chiral stereoisomers. *Briefings in Bioinformatics*, 24(1):bbac560, 2023.

[Du *et al.*, 2024] Wenjie Du, Shuai Zhang, Jun Xia Di Wu, Ziyuan Zhao, Junfeng Fang, and Yang Wang. Mmgnn: A molecular merged graph neural network for explainable solvation free energy prediction. 2024.

[Du *et al.*, 2025] Wenjie Du, Shuai Zhang, Zhaohui Cai, Zhiyuan Liu, Junfeng Fang, Jianmin Wang, and Yang Wang. Molecular merged hypergraph neural network for explainable solvation free energy prediction. *Research*, 0(ja), 2025.

[Fang *et al.*, 2024a] Junfeng Fang, Guibin Zhang, Kun Wang, Wenjie Du, Yifan Duan, Yuankai Wu, Roger Zimmermann, Xiaowen Chu, and Yuxuan Liang. On regularization for explaining graph neural networks: An information theory perspective. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[Fang *et al.*, 2024b] Junfeng Fang, Shuai Zhang, Chang Wu, Zhengyi Yang, Zhiyuan Liu, Sihang Li, Kun Wang, Wenjie Du, and Xiang Wang. MolTC: Towards molecular relational modeling in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1943–1958, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[Feinberg *et al.*, 2020] Evan N Feinberg, Elizabeth Joshi, Vijay S Pande, and Alan C Cheng. Improvement in admet prediction with multitask deep featurization. *Journal of medicinal chemistry*, 63(16):8835–8848, 2020.

[Gao *et al.*, 2024a] Bowen Gao, Bo Qiang, Haichuan Tan, Yinjun Jia, Minsi Ren, Minsi Lu, Jingjing Liu, Wei-Ying Ma, and Yanyan Lan. Drugclip: Contrasive protein-molecule representation learning for virtual screening. *Advances in Neural Information Processing Systems*, 36, 2024.

[Gao *et al.*, 2024b] Yuan Gao, Shuguo Jiang, Moran Li, Jin-Gang Yu, and Gui-Song Xia. Dmtg: One-shot differentiable multi-task grouping. In *International Conference on Machine Learning (ICML)*, 2024.

[Gleeson *et al.*, 2011] M Paul Gleeson, Anne Hersey, Dino Montanari, and John Overington. Probing the links between in vitro potency, admet and physicochemical parameters. *Nature reviews Drug discovery*, 10(3):197–208, 2011.

[Guo *et al.*, 2018] Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 270–287, 2018.

[Kendall *et al.*, 2018] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[Kim *et al.*, 2024] Jungwoo Kim, Woojae Chang, Hyunjun Ji, and InSuk Joung. Quantum-informed molecular representation learning enhancing admet property prediction. *Journal of Chemical Information and Modeling*, 64(13):5028–5040, 2024.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional

networks. In *International Conference on Learning Representations*, 2017.

[Liu *et al.*, 2019] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019.

[Maddison *et al.*, 2017] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

[Montanari *et al.*, 2019] Floriane Montanari, Lara Kuhnke, Antonius Ter Laak, and Djork-Arné Clevert. Modeling physico-chemical admet endpoints with multitask graph convolutional networks. *Molecules*, 25(1):44, 2019.

[Norinder and Bergström, 2006] Ulf Norinder and Christel AS Bergström. Prediction of admet properties. *ChemMedChem: Chemistry Enabling Drug Discovery*, 1(9):920–937, 2006.

[Peng *et al.*, 2020] Yuzhong Peng, Yanmei Lin, Xiao-Yuan Jing, Hao Zhang, Yiran Huang, and Guang Sheng Luo. Enhanced graph isomorphism network for molecular admet properties prediction. *Ieee Access*, 8:168344–168360, 2020.

[Rentzsch *et al.*, 2019] Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research*, 47(D1):D886–D894, 2019.

[Sener and Koltun, 2018] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

[Swanson *et al.*, 2023] Kyle Swanson, Parker Walther, Jeremy Leitz, Souhrid Mukherjee, Joseph C Wu, Rabindra V Shivnaraine, and James Zou. Admet-ai: A machine learning admet platform for evaluation of large-scale chemical libraries. *BioRxiv*, pages 2023–12, 2023.

[Tao and Abe, 2025] Nianze Tao and Minori Abe. Bayesian flow network framework for chemistry tasks. *Journal of Chemical Information and Modeling*, 65(3):1178–1187, 2025.

[Vandenhende *et al.*, 2021] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3614–3633, 2021.

[Vinyals *et al.*, 2016] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. In *ICLR (Poster)*, 2016.

[Wang *et al.*, 2020] Xiting Wang, Meng Liu, Lan Zhang, Yun Wang, Yu Li, and Tao Lu. Optimizing pharmacokinetic property prediction based on integrated datasets and a deep learning approach. *Journal of Chemical Information and Modeling*, 60(10):4603–4613, 2020.

[Wang *et al.*, 2024] Jingjing Wang, Zhijiang Yang, Chang Chen, Ge Yao, Xiukun Wan, Shaoheng Bao, Junjie Ding, Liangliang Wang, and Hui Jiang. Mpek: a multitask deep learning framework based on pretrained language models for enzymatic reaction kinetic parameters prediction. *Briefings in Bioinformatics*, 25(5), 2024.

[Wu *et al.*, 2021] Zhenxing Wu, Dejun Jiang, Chang-Yu Hsieh, Guangyong Chen, Ben Liao, Dongsheng Cao, and Tingjun Hou. Hyperbolic relational graph convolution networks plus: a simple but highly efficient qsar-modeling method. *Briefings in Bioinformatics*, 22(5):bbab112, 2021.

[Xu *et al.*, 2017] Yuting Xu, Junshui Ma, Andy Liaw, Robert P Sheridan, and Vladimir Svetnik. Demystifying multitask deep neural networks for quantitative structure–activity relationships. *Journal of chemical information and modeling*, 57(10):2490–2504, 2017.

[Xu *et al.*, 2025] Minghao Xu, Yunteng Geng, Yihang Zhang, Ling Yang, Jian Tang, and Wentao Zhang. GlycanML: A multi-task and multi-structure benchmark for glycan machine learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Yang and Du, 2022] Xiaoting Yang and Wenjie Du. Bond energy assists accurate molecule property prediction. In *Journal of Physics: Conference Series*, volume 2356, page 012047. IOP Publishing, 2022.

[Yang *et al.*, 2018] Ming Yang, Jialei Chen, Liwen Xu, Xiufeng Shi, Xin Zhou, Zhijun Xi, Rui An, and Xinhong Wang. A novel adaptive ensemble classification framework for adme prediction. *RSC advances*, 8(21):11661–11683, 2018.

[Yang *et al.*, 2019] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

[Yu *et al.*, 2021] Junchi Yu, Tingyang Xu, Yu Rong, Yatao Bian, Junzhou Huang, and Ran He. Graph information bottleneck for subgraph recognition. In *International Conference on Learning Representations*, 2021.

[Zhang and Yang, 2021] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609, 2021.

[Zhao *et al.*, 2020] Linlin Zhao, Heather L Ciallella, Lauren M Aleksunes, and Hao Zhu. Advancing computer-aided drug discovery (cadd) by big data and data-driven machine learning modeling. *Drug discovery today*, 25(9):1624–1638, 2020.

[Zhou *et al.*, 2023] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. 2023.