Uncertainty-Guided Adaptive Local Adversarial Perturbation for Reliable Molecular Property Prediction

Xuxin Zhao , **Wen Zhang*** and **Shichao Liu***

College of Informatics, Huazhong Agricultural University, Wuhan 430070, China

xuxin7286@webmail.hzau.edu.cn, {zhangwen, scliu}@mail.hzau.edu.cn

Abstract

Molecular property prediction is a fundamental task with wide-ranging applications in chemistry, materials science, and drug discovery. Beyond predictive accuracy, the reliability of model predictions, particularly their uncertainty awareness, is critical for high-stakes downstream applications. However, most existing methods predominantly optimize performance metrics while overlooking prediction reliability, often producing results that are both inaccurate and overconfident. To address this challenge, we propose an uncertainty-guided framework that integrates adaptive local adversarial perturbation with evidential deep learning for reliable molecular property prediction. The proposed method identifies sensitive molecular regions through gradient-based sensitivity analysis and applies locally adaptive perturbations whose magnitudes are dynamically modulated by predictive uncertainty. This approach explicitly challenges the model to preserve predictive stability under structure-aware perturbations, thereby enhancing its sensitivity to critical molecular substructures. Furthermore, we introduce an evidential calibration strategy that aligns uncertainty with the prediction discrepancy between clean and adversarial molecular views, enabling the learned evidence to faithfully reflect prediction stability. Extensive experiments on benchmark molecular property prediction datasets demonstrate that our approach achieves state-of-the-art predictive performance while providing more reliable uncertainty estimates than widely used uncertainty quantification methods. These results highlight the potential of the proposed framework as a reliable tool for risk-sensitive virtual screening and molecular decision-making. Our code is available at https://github.com/Ubehind/ugap.

*Corresponding author

1 Introduction

Molecular property prediction is essential for modern drug discovery and materials design. With the rapid development of graph neural networks (GNNs), molecular structures are commonly modeled as graphs, where atoms correspond to nodes and chemical bonds to edges, enabling expressive learning of structure–property relationships [Kang *et al.*, 2024; Xiong *et al.*, 2026]. These models have achieved impressive performance on standard benchmarks. However, despite their empirical success, existing molecular property predictors often suffer from limited generalization ability [Zhang *et al.*, 2024; Han *et al.*, 2023]. In real-world applications, chemical space is vast and sparsely sampled, requiring models to extrapolate to novel compounds or molecular classes that are underrepresented in training data. Under such conditions, prediction errors become difficult to detect, and models tend to produce overconfident outputs even when their predictions are unreliable [Yu *et al.*, 2022]. This mismatch between predictive confidence and actual correctness raises a critical question for practical deployment: when can a model's prediction be trusted? Addressing this question is as important as improving predictive accuracy, particularly in high-stakes scenarios such as virtual screening and early-stage drug discovery.

Adversarial training has been widely explored to improve robustness and generalization in machine learning. Early work in computer vision and natural language processing showed that training on adversarially perturbed samples could enhance model stability under distributional, syntactic, or semantic variations [Xie *et al.*, 2020; Li and Qiu, 2021]. More recently, adversarial training was adapted to graph-structured data [Kong *et al.*, 2022], where perturbations were applied to node features, edges, or latent representations. In molecular property prediction, adversarial perturbations were typically applied to atomic embeddings or learned representations [Jin *et al.*, 2025]. These methods were shown to help GNNs generalize to out-of-distribution samples by encouraging consistent predictions under input variations. However, existing approaches often rely on uniform perturbation strategies, which overlook the heterogeneous importance of different molecular regions and may introduce excessive perturbations to fragile representations. Moreover, multi-step iterative perturbation methods incur substantial computational overhead, limiting their scalability for large molecular datasets.

In parallel, uncertainty quantification has received growing attention as a way to assess the reliability of model predictions. Bayesian neural networks, Monte Carlo dropout, and deep ensembles were proposed to estimate predictive uncertainty by approximating posterior distributions over model parameters [Jospin *et al.*, 2022; Lakshminarayanan *et al.*, 2017; Gal and Ghahramani, 2016]. Despite their effectiveness, these approaches are often computationally expensive and typically require multiple inference passes. Evidential deep learning later emerged as a deterministic alternative that models uncertainty by predicting parameters of evidential distributions [Sensoy *et al.*, 2018; Amini *et al.*, 2020], enabling uncertainty estimation in a single forward pass. This paradigm was successfully applied to regression and classification tasks, including molecular property prediction [Soleimany *et al.*, 2021]. However, existing evidential methods typically rely on implicit assumptions linking confidence to reliability and lack explicit mechanisms to calibrate evidence against prediction stability under perturbations.

To address these challenges, we propose UGAP, an uncertainty-guided adaptive perturbation framework for reliable molecular property prediction. UGAP integrates adversarial training with evidential learning to couple predictive performance with uncertainty estimation. Motivated by the observation that molecular properties are often governed by specific functional groups rather than the entire molecular structure, UGAP identifies salient molecular regions via gradient-based sensitivity analysis and selectively perturbs influential atoms. Furthermore, the perturbation intensity is adaptively modulated by predictive uncertainty: stronger perturbations are applied to high-confidence samples to probe representational robustness, while milder perturbations are used for low-confidence samples whose feature representations are not yet well established. In addition, UGAP introduces a stability-aware evidence calibration strategy that regularizes evidential outputs using discrepancies between clean and perturbed predictions, thereby aligning evidential magnitude with predictive stability. Extensive experiments demonstrate that UGAP consistently outperforms existing baselines in both predictive accuracy and uncertainty reliability, offering an effective pathway toward trustworthy molecular property prediction.

2 Method

Each molecule is represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where our goal is to learn a mapping $f : \mathcal{G} \rightarrow (\hat{y}, u)$ that produces accurate predictions $\hat{y}$ and reliable uncertainty estimates u. To achieve this, we propose UGAP, a framework integrating evidential learning with adaptive perturbations to optimize both generalization and uncertainty fidelity. As illustrated in Figure 1, raw atomic features are first converted into embeddings. UGAP then applies perturbations to salient atoms with intensities dynamically modulated by uncertainty. Finally, a stability-aware calibration module refines uncertainty by regularizing the discrepancy between clean and perturbed predictions. This mechanism aligns evidence magnitude with structural robustness, establishing a reliable safeguard for molecular property prediction.

2.1 Evidential Learning for Regression and Classification

To quantify predictive uncertainty, we adopt Evidential Deep Learning (EDL). Unlike standard networks producing point estimates, EDL outputs the hyperparameters of a prior distribution over the target parameters.

Regression Task

For continuous property prediction, the target y is modeled as a Gaussian distribution with a Normal-Inverse-Gamma (NIG) prior $(\mu, \sigma^2) \sim \text{NIG}(\gamma, v, \alpha, \beta)$ [Amini *et al.*, 2020]. The predicted mean and epistemic uncertainty u are derived from the estimated hyperparameters:

$$\mu = \gamma, \quad u = \frac{\beta}{v(\alpha - 1)} \tag{1}$$

The model was optimized using the negative log-likelihood $\mathcal{L}_{NLL}$:

$$\begin{aligned} \mathcal{L}_{NLL} =& \frac{1}{2}\log\left(\frac{\pi}{v}\right) - \alpha\log(\Omega) + \log\left(\frac{\Gamma(\alpha)}{\Gamma(\alpha + 1/2)}\right) \\ &+ \left(\alpha + \frac{1}{2}\right)\log\left((y-\gamma)^2 v + \Omega\right) \end{aligned} \tag{2}$$

where $\Omega = 2\beta(1 + v)$. $\mathcal{L}_{NLL}$ aims to fit the ground truth by maximizing the marginalized likelihood of the observed data. Furthermore, an auxiliary regularizer $\mathcal{L}_{REG}$ was utilized to penalize evidence associated with inaccurate predictions:

$$\mathcal{L}_{REG} = |y - \gamma| \cdot (2v + \alpha) \tag{3}$$

The total regression loss was formulated as:

$$\mathcal{L}_{reg} = \mathcal{L}_{NLL} + \lambda\mathcal{L}_{REG} \tag{4}$$

Classification Task

In classification tasks, class probabilities follow a Dirichlet distribution $\text{Dir}(\mathbf{p}|\boldsymbol{\alpha})$ [Sensoy *et al.*, 2018]. Evidence $e_k \geq 0$ for each class defines the concentration parameters $\alpha_k = e_k + 1$ and total evidence $S = \sum_{k=1}^{K} \alpha_k$. The predicted probability $\hat{p}_k$ and uncertainty u are:

$$\hat{p}_k = \frac{\alpha_k}{S}, \quad u = \frac{K}{S} \tag{5}$$

Training was supervised by the expected cross-entropy loss $\mathcal{L}_{ce}$:

$$\begin{aligned} \mathcal{L}_{ce} &= \int \left[\sum_{k=1}^{K} -y_k \log p_k\right] \text{Dir}(\mathbf{p}|\boldsymbol{\alpha}) d\mathbf{p} \\ &= \sum_{k=1}^{K} y_k \left(\psi(S) - \psi(\alpha_k)\right) \end{aligned} \tag{6}$$

This loss minimizes the prediction error by aligning the multinomial opinions with the categorical ground truth labels. A KL divergence term $\mathcal{L}_{KL}$ was incorporated to penalize misleading evidence:

$$\mathcal{L}_{KL} = \text{KL}\left[\text{Dir}(\mathbf{p}|\tilde{\alpha}) \parallel \text{Dir}(\mathbf{p}|\mathbf{1})\right] \tag{7}$$

where $\tilde{\alpha}_k = y_k + (1 - y_k)\alpha_k$. The final classification loss was integrated as:

$$\mathcal{L}_{cls} = \mathcal{L}_{ce} + \lambda\mathcal{L}_{KL} \tag{8}$$

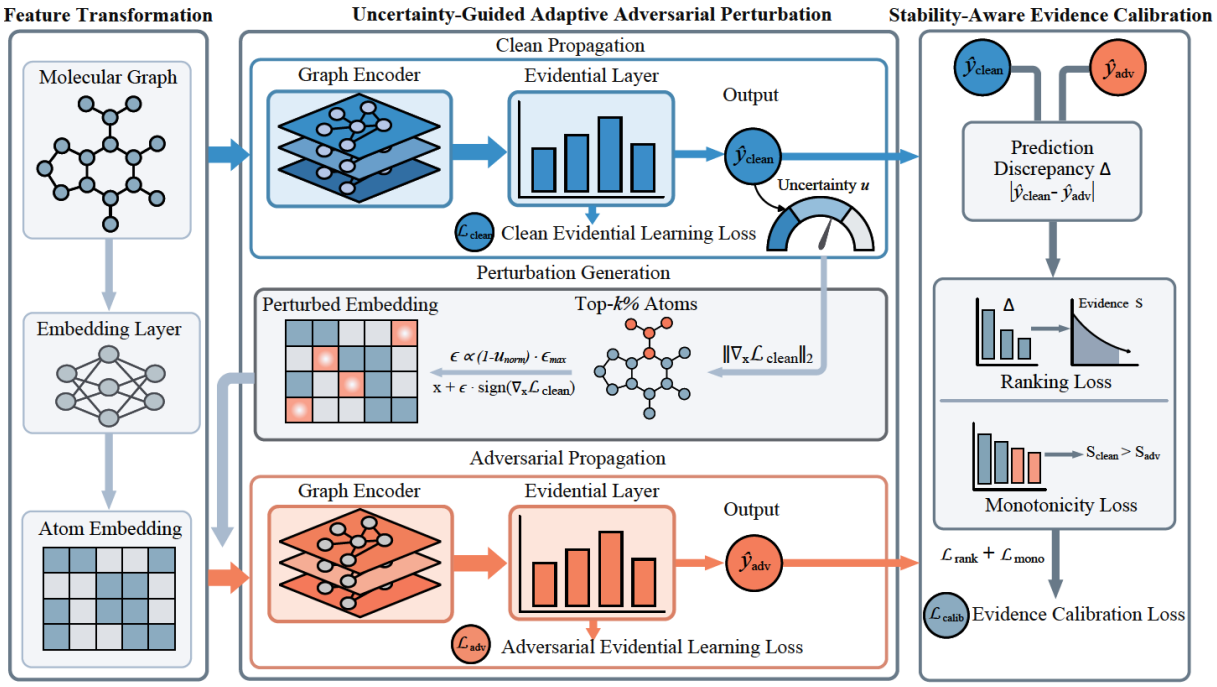

Figure 1: Overview of the UGAP framework. Raw atomic features are first converted into embeddings. UGAP then applies perturbations to salient atoms, with intensities modulated by uncertainty. Finally, a stability-aware calibration module refines uncertainty via a ranking loss to suppress evidence for high-discrepancy samples and a monotonicity loss to ensure perturbed evidence is lower than clean evidence.

2.2 Uncertainty-Guided Adaptive Adversarial Perturbation

Clean Propagation

A clean propagation path was first executed to perform evidential learning. Each molecular graph was initially processed through an embedding layer to convert raw atomic features into atomic embeddings $\mathbf{x}_i$. These embeddings were then passed through the graph encoder to obtain representations for computing the evidential learning loss $\mathcal{L}_{\text{clean}}$, which is defined as:

$$\mathcal{L}_{\text{clean}} = \begin{cases} \mathcal{L}_{\text{reg}}, & \text{for regression tasks} \\ \mathcal{L}_{\text{cls}}, & \text{for classification tasks} \end{cases} \tag{9}$$

To identify structural components where the model exhibits high sensitivity, we examined the gradient of $\mathcal{L}_{\text{clean}}$ with respect to each initial embedding $\mathbf{x}_i$. Atoms that play a decisive role in predictions typically exhibit larger gradients. We quantified atomic saliency s_i by computing the L_2 norms of these gradients:

$$s_i = \|\nabla_{\mathbf{x}_i} \mathcal{L}_{\text{clean}}\|_2 \tag{10}$$

The computed saliency scores provided the basis for targeted perturbations in the subsequent adversarial path.

Perturbation Generation and Adversarial Propagation

Atoms were ranked by saliency scores, and a top-k (recommended $k \in [0.2, 0.6]$) ratio was selected as perturbation targets. An adversarial perturbation mechanism was designed to generate perturbed representations $\mathbf{x}'_i$. The perturbation radius ϵ was dynamically adjusted based on normalized batch uncertainty $\bar{u}$ to calculate the adaptive intensity:

$$\epsilon = (1 - \bar{u}) \cdot \epsilon_{\max} \tag{11}$$

With $\epsilon_{\max} \in [0.06, 0.1]$ recommended, this modulation assigns larger perturbations to high-confidence samples to probe robustness, while applying milder noise to uncertain ones to maintain stability. Following this, adversarial perturbations were applied to generate the perturbed representations:

$$\mathbf{x}'_i = \mathbf{x}_i + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}_i} \mathcal{L}_{\text{clean}}) \tag{12}$$

The resulting $\mathbf{x}'_i$ were then passed through an adversarial propagation path to calculate the adversarial evidential loss $\mathcal{L}_{\text{adv}}$, defined analogously to $\mathcal{L}_{\text{clean}}$. This dual-loss structure encourages the model to learn robust representations from localized structural variations.

2.3 Stability-Aware Evidence Calibration

We developed a calibration scheme to align predictive evidence with structural robustness. In this framework, the prediction shift $\Delta\hat{y}$ serves as a metric for structural sensitivity. For regression tasks, $\Delta\hat{y}$ is defined as the absolute difference $|\hat{y}_{\text{clean}} - \hat{y}_{\text{adv}}|$, while for classification, it is calculated as the Kullback-Leibler divergence between the two predictive probability distributions. Based on this metric, the stability-aware ranking loss $\mathcal{L}_{\text{rank}}$ was employed to ensure that samples exhibiting larger prediction shifts are assigned lower evidence S. Following the parameterization in [Sensoy *et al.*, 2018] and [Amini *et al.*, 2020], the total evidence S is quantified as

$S = \sum_{k=1}^{K} \alpha_k$ for classification and $S = 2\alpha + \nu$ for regression:

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \ln\left(1 + \exp\left(\log \frac{S_i}{S_j}\right)\right), \tag{13}$$

where $\mathcal{P} = \{(i,j) \mid \phi_{i,j} \geq \text{percentile}(\Phi, 50)\}$ denotes the set of sample pairs whose shift differences $\phi_{i,j} = \Delta\hat{y}_i - \Delta\hat{y}_j$ rank in the top 50%. We restricted the calibration to the top 50% of pairs because forcing the model to distinguish between pairs with nearly identical stability yields little useful information and creates minor gradients that dilute stronger signals. By filtering for these clear differences, the ranking objective provided a more effective signal for evidence calibration.

Furthermore, based on the principle of epistemic consistency, we implemented the monotonicity constraint $\mathcal{L}_{\text{mono}} = \max(0, S_{\text{adv}} - S_{\text{clean}})$ to ensure that evidence from perturbed passes remains below that of clean passes. In this term, S_{clean} was detached from the computational graph, forcing the gradient to exclusively penalize overconfident S_{adv} values. The total calibration loss $\mathcal{L}_{\text{calib}}$ was then defined as $\mathcal{L}_{\text{calib}} = \mathcal{L}_{\text{rank}} + \mathcal{L}_{\text{mono}}$.

2.4 Training Objective

To simultaneously account for generalization capability and uncertainty fidelity, the framework was optimized using a multi-objective loss function. The total loss $\mathcal{L}_{\text{total}}$ integrates the evidential learning losses from both clean and adversarial propagation paths along with the calibration loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{clean}} + \lambda_1 \mathcal{L}_{\text{adv}} + \lambda_2 \mathcal{L}_{\text{calib}}, \tag{14}$$

where λ_1 and λ_2 are hyperparameters that coordinate adversarial training and evidence calibration. This joint objective ensures that the model maintains robust generalization across molecular structures while producing uncertainty estimates that faithfully reflect structural sensitivity.

3 Experiments

3.1 Experimental Setup

Datasets

We evaluated the framework on eight datasets from MoleculeNet [Wu *et al.*, 2018]. The classification tasks include BBBP, ClinTox, BACE, Tox21, and SIDER, where ROC-AUC is employed as the primary metric. The regression tasks comprise FreeSolv, ESOL, and Lipophilicity, with RMSE used for evaluation. To rigorously assess the model's generalization ability, we adopted a strict scaffold splitting strategy with a ratio of 8:1:1 for training, validation, and testing.

Baselines

We conducted extensive experiments to investigate the performance of the proposed framework on molecular property prediction tasks. To ensure a comprehensive evaluation, we compared our approach against several groups of baselines, focusing on their predictive accuracy and reliability in chemical contexts.

Molecular property prediction models: This category includes fundamental GNN backbones such as GIN [Xu *et al.*, 2018], GPS [Rampášek *et al.*, 2022], and CMPNN [Song *et al.*, 2020]. We integrated our framework into these encoders to validate its versatility in improving molecular representations. Additionally, we considered state-of-the-art pre-trained models, including GraphMVP [Liu *et al.*, 2021], GROVER [Rong *et al.*, 2020], MolCLR [Wang *et al.*, 2022], ImageMol [Zeng *et al.*, 2022], Uni-Mol [Zhou *et al.*, 2023], MotiL [Liu *et al.*, 2025], and SCAGE [Qiao *et al.*, 2025], which serve as strong competitive benchmarks for molecular property prediction.

Adversarial perturbation methods: To evaluate the impact of different perturbation schemes on molecular property prediction, we compared our strategy against classic adversarial training techniques, including FGSM [Goodfellow *et al.*, 2014] and PGD [Madry *et al.*, 2017], as well as graph-specific data augmentation methods such as FLAG [Kong *et al.*, 2022] and CAP [Xue *et al.*, 2021]. These comparisons were performed to evaluate the effectiveness of our uncertainty-guided perturbation in enhancing the generalization performance of molecular property predictors.

Predictive reliability baselines: Given the importance of confidence estimation in drug discovery, we assessed our framework's ability to provide reliable uncertainty. In addition to vanilla EDL [Sensoy *et al.*, 2018; Amini *et al.*, 2020], we included representative uncertainty quantification methods, namely Deep Ensembles [Lakshminarayanan *et al.*, 2017] and MC-Dropout [Gal and Ghahramani, 2016]. These were employed to establish benchmarks for uncertainty estimates, ensuring that the model's confidence aligns with its actual predictive performance on complex chemical structures.

3.2 Performance on Molecular Property Prediction

The results in Table 1 demonstrate that the proposed UGAP framework consistently achieved superior performance across both classification and regression benchmarks. In classification tasks, our method exhibited significant advantages, particularly on the BBBP, ClinTox, and BACE datasets, where it reached the highest ROC-AUC scores of 75.9, 96.2, and 88.2, respectively. Compared to competitive pre-trained baselines such as Uni-Mol and SCAGE, the UGAP+CMPNN model observed a notable margin of improvement. For instance, on the ClinTox dataset, the model yielded a ROC-AUC of 96.2, significantly outperforming the vanilla CMPNN backbone (90.5) and even the second-best model, SCAGE (93.7). Notably, these results indicate that our framework can achieve performance comparable to models pre-trained on large-scale datasets.

In regression tasks, the UGAP framework maintained its superiority by achieving the lowest RMSE on FreeSolv (1.61) and ESOL (0.80). Specifically, on the FreeSolv dataset, our method improved the baseline CMPNN from 1.92 to 1.61, outperforming Uni-Mol's score of 1.72. For the ESOL task, UGAP+CMPNN attained an RMSE of 0.80, which surpassed advanced pre-training methods like MotiL (0.82) and GROVER (0.88). Moreover, the backbone-agnostic nature was validated as UGAP+GIN and UGAP+GPS also showed consistent gains over their respective vanilla versions. Overall, these results demonstrate that the proposed framework

Methods	Classification (ROC-AUC% ↑)					Regression (RMSE ↓)		
	BBBP	ClinTox	BACE	Tox21	SIDER	FreeSolv	ESOL	Lipo
GIN	65.8 ± 1.1	85.0 ± 1.4	77.1 ± 2.1	70.2 ± 1.8	57.3 ± 1.6	2.71 ± 0.13	1.45 ± 0.16	0.87 ± 0.06
GPS	64.8 ± 1.6	87.2 ± 1.9	78.0 ± 2.5	70.5 ± 1.4	60.8 ± 2.1	2.65 ± 0.12	1.40 ± 0.12	0.90 ± 0.07
CMPNN	71.3 ± 1.1	90.5 ± 0.9	81.3 ± 1.0	72.0 ± 0.8	61.6 ± 1.0	2.15 ± 0.10	1.19 ± 0.07	0.85 ± 0.06
GraphMVP	72.4 ± 1.0	87.1 ± 0.8	81.2 ± 1.2	75.9 ± 1.0	63.9 ± 1.1	2.15 ± 0.10	1.03 ± 0.09	0.75 ± 0.05
GROVER	69.5 ± 0.8	86.2 ± 0.7	81.0 ± 0.8	73.5 ± 1.2	65.4 ± 1.2	2.20 ± 0.08	0.88 ± 0.10	0.82 ± 0.05
MolCLR	73.0 ± 0.7	90.4 ± 0.7	83.5 ± 1.0	75.7 ± 0.6	60.7 ± 0.8	2.79 ± 0.09	0.84 ± 0.08	0.74 ± 0.06
ImageMol	70.9 ± 1.2	89.1 ± 1.0	83.9 ± 1.1	73.3 ± 0.8	66.0 ± 0.9	2.81 ± 0.11	1.07 ± 0.09	0.74 ± 0.05
Uni-Mol	72.8 ± 0.6	92.9 ± 0.9	85.7 ± 0.9	$\mathbf{77.9 \pm 0.6}$	65.1 ± 0.7	1.72 ± 0.06	0.91 ± 0.07	0.69 ± 0.04
MotiL	73.2 ± 1.0	92.8 ± 0.8	86.4 ± 0.7	74.9 ± 0.7	66.4 ± 0.5	1.97 ± 0.07	0.82 ± 0.06	0.69 ± 0.04
SCAGE	73.5 ± 0.7	93.7 ± 1.0	85.7 ± 0.6	76.0 ± 0.6	66.1 ± 0.5	1.80 ± 0.08	0.83 ± 0.04	$\mathbf{0.65 \pm 0.02}$
UGAP+GIN	73.3 ± 0.7	93.3 ± 0.8	85.6 ± 0.9	74.6 ± 0.8	62.6 ± 0.8	1.75 ± 0.09	0.92 ± 0.09	0.81 ± 0.05
UGAP+GPS	72.0 ± 0.8	92.2 ± 0.9	85.2 ± 0.7	74.9 ± 0.7	63.4 ± 0.7	1.90 ± 0.10	1.05 ± 0.07	0.84 ± 0.05
UGAP+CMPNN	$\mathbf{75.9 \pm 0.7}$	$\mathbf{96.2 \pm 0.6}$	$\mathbf{88.2 \pm 0.5}$	76.7 ± 0.6	$\mathbf{67.0 \pm 0.5}$	$\mathbf{1.61 \pm 0.05}$	$\mathbf{0.80 \pm 0.04}$	0.67 ± 0.03

Table 1: The average RMSE and ROC-AUC results (with standard deviation) on MoleculeNet's classification and regression test sets under scaffold splitting. The best results are highlighted in bold, with the second-best results underlined.

effectively enhances the predictive performance and generalization of GNNs on molecular property tasks.

3.3 Ablation Study

We conducted ablation experiments on eight benchmark datasets to evaluate the contribution of each UGAP component. We compared UGAP against various perturbation baselines and assessed the significance of its core loss functions.

Effectiveness of Perturbation Strategies

The results in Table 2 demonstrated that gradient-based perturbation strategies consistently outperformed the random noise baseline. Compared to global methods like PGD and FLAG, UGAP w/o Adaptive, which applied perturbations only to high-gradient local regions, achieved superior performance across all datasets. For instance, on ClinTox, it reached an ROC-AUC of 95.2, marking a significant improvement over FLAG's 92.1. This suggests that targeting chemically influential atoms is more effective for regularizing molecular representations than uniform perturbations. Furthermore, the full UGAP model, which incorporated uncertainty-guided intensity, attained the best results in most cases, such as an RMSE of 1.61 on FreeSolv. This indicated that adaptively modulating perturbation strength based on model confidence further enhances the model's predictive performance.

Impact of Loss Components

We investigated the necessity of the adversarial loss ($\mathcal{L}_{\text{adv}}$) and the calibration loss ($\mathcal{L}_{\text{calib}}$). The performance of UGAP w/o $\mathcal{L}_{\text{adv}}$ dropped sharply to levels near the vanilla model, which verified that adversarial training is the primary driver for boosting predictive accuracy. In contrast, UGAP w/o L_{calib} maintained high accuracy and showed slightly better results on specific datasets like BBBP and Lipo. However, the full UGAP model achieved more consistent performance across the majority of tasks, such as Tox21 (76.7) and ESOL (0.80). This suggested that while $\mathcal{L}_{\text{calib}}$ is primarily designed to enhance uncertainty reliability, it tends to provide a stabilizing regularization effect. The results indicated that incorporating calibration mechanisms does not significantly compromise, and in many cases remains compatible with, the model's overall predictive power.

3.4 Uncertainty-Informed Predictive Performance

To evaluate uncertainty reliability and its relationship with model performance, we conducted an analysis across confidence intervals on both classification (BACE, BBBP) and regression (ESOL, FreeSolv) tasks. As Figure 2 illustrates, we ranked test samples by predicted confidence and evaluated predictive metrics within each interval.

In classification tasks, the ROC-AUC of our proposed strategy (red bars) consistently increased as confidence rose, outperforming vanilla Evidence, Ensemble, and Dropout methods, particularly in high-confidence regions. For instance, on

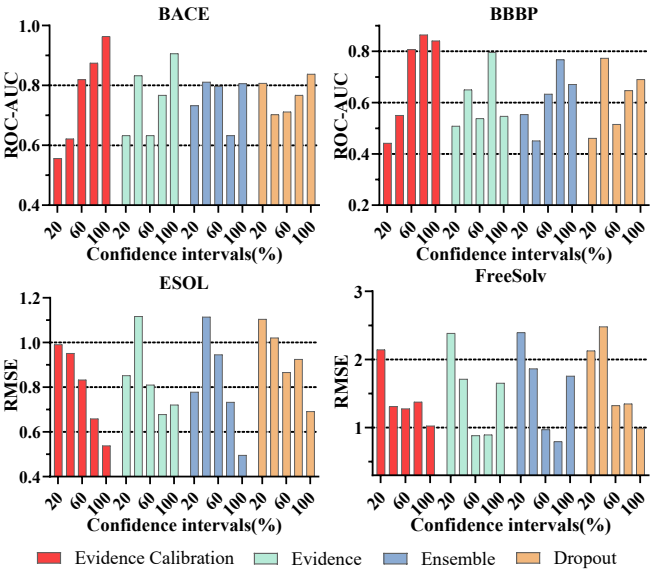

Figure 2: Performance comparison across confidence intervals. Evidence Calibration consistently outperforms baseline methods in high-confidence regions.

Methods	Classification (ROC-AUC% ↑)					Regression (RMSE ↓)		
	BBBP	ClinTox	BACE	Tox21	SIDER	FreeSolv	ESOL	Lipo
Vanilla (CMPNN)	71.3 ± 1.1	90.5 ± 0.9	81.3 ± 1.0	72.0 ± 0.8	61.6 ± 1.0	2.15 ± 0.10	1.19 ± 0.07	0.85 ± 0.06
Comparison of Perturbation Strategies								
Random	71.8 ± 1.0	90.9 ± 0.7	81.9 ± 1.2	72.2 ± 0.9	61.8 ± 0.8	2.13 ± 0.08	1.15 ± 0.09	0.85 ± 0.05
FGSM	72.5 ± 0.9	91.8 ± 0.8	82.8 ± 1.0	73.1 ± 0.7	62.3 ± 1.0	2.07 ± 0.09	1.09 ± 0.08	0.83 ± 0.06
PGD	73.4 ± 0.5	92.7 ± 0.7	85.0 ± 0.5	74.5 ± 0.8	63.8 ± 0.6	1.93 ± 0.07	0.92 ± 0.04	0.79 ± 0.05
FLAG	73.4 ± 0.8	92.1 ± 0.4	85.5 ± 0.8	74.3 ± 0.7	64.0 ± 0.5	1.87 ± 0.05	0.90 ± 0.06	0.76 ± 0.03
CAP	73.1 ± 0.9	93.2 ± 0.4	86.4 ± 0.7	75.0 ± 0.7	64.5 ± 0.6	1.80 ± 0.06	0.86 ± 0.06	0.75 ± 0.04
UGAP w/o Local	73.7 ± 0.7	93.0 ± 0.5	85.1 ± 0.6	74.8 ± 0.7	64.6 ± 0.4	1.75 ± 0.07	0.88 ± 0.05	0.70 ± 0.02
UGAP w/o Adaptive	75.2 ± 0.8	95.2 ± 0.4	87.4 ± 0.6	75.7 ± 0.6	66.2 ± 0.5	1.68 ± 0.05	0.84 ± 0.06	0.68 ± 0.03
Ablation on Loss Functions								
UGAP w/o L_{adv}	71.5 ± 0.7	90.8 ± 0.6	81.5 ± 0.7	72.4 ± 0.6	62.7 ± 0.7	1.90 ± 0.06	1.15 ± 0.07	0.80 ± 0.04
UGAP w/o L_{calib}	$\mathbf{76.0 \pm 0.8}$	95.6 ± 0.5	87.9 ± 0.5	75.6 ± 0.5	67.8 ± 0.4	1.70 ± 0.09	0.82 ± 0.05	$\mathbf{0.66 \pm 0.02}$
UGAP (Full)	75.9 ± 0.7	$\mathbf{96.2 \pm 0.6}$	$\mathbf{88.2 \pm 0.5}$	$\mathbf{76.7 \pm 0.6}$	$\mathbf{67.0 \pm 0.5}$	$\mathbf{1.61 \pm 0.05}$	$\mathbf{0.80 \pm 0.04}$	0.67 ± 0.03

Table 2: Performance comparison across perturbation strategies and ablation studies of UGAP on MoleculeNet's classification and regression test sets under scaffold splitting. The best results are highlighted in bold.

the BACE dataset, our method achieved an ROC-AUC near 1.0 in the top confidence tier, whereas baseline methods exhibited a performance gap or fluctuations.

Similarly, in regression tasks, the RMSE of our method showed a clear downward trend as confidence increased, indicating that samples with higher confidence are indeed associated with lower predictive errors. Compared to the erratic behavior of MC-Dropout and sub-optimal performance of standard Ensembles, our method provided a more monotonic and dependable relationship between confidence and accuracy. These results validate that our framework enhances predictive performance and ensures that model confidence is a reliable indicator of actual performance, which is crucial for decision-making in drug discovery.

3.5 Mechanism of Evidence Calibration

To elucidate how the perturbation informs uncertainty estimation, we analyzed the relationship between predictive discrepancy Δ and evidence magnitudes on the BACE and ESOL datasets. Intuitively, a large Δ indicated that the model's prediction was sensitive to structural perturbations, which should ideally correspond to lower evidence (higher uncertainty). As illustrated in Figure 3, we categorized samples into Low, Mid, and High Δ regions and observed the distribution of evidence values. In the vanilla evidential model (green boxes), although there was a slight downward trend in evidence as Δ increased, the distinction between different discrepancy regions was relatively blurred. This suggested that without explicit alignment, standard evidential learning struggles to fully capture the structural instability of molecules. By incorporating our evidential calibration strategy, the UGAP model (red boxes) demonstrated a much more pronounced and monotonic decrease in evidence values as Δ rose. Specifically, in high-discrepancy regions, the evidence magnitudes were significantly suppressed compared to the vanilla version. This clear correlation confirmed that our framework effectively utilized the predictive discrepancy induced by perturbations to regularize evidence, thereby ensuring that the model's confidence was grounded in its structural robustness.

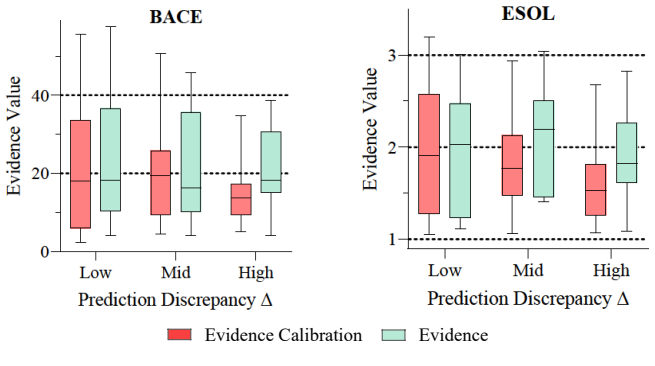

Figure 3: Distribution of evidence values across different prediction discrepancy (Δ) regions.

3.6 Case Study

To evaluate perturbation sites and uncertainty reliability, we visualized representative molecules from the BBBP dataset (Figure 4). The light-red regions indicate the salient substructures targeted by our adaptive perturbation, while blue dashed circles highlight chemically significant groups, such as the sulfonamide group, hydroxyl group, and thioether linkage. In the BBBP task, these groups are pivotal; for instance, polar groups like sulfonamides and hydroxyls typically hinder passive diffusion across the lipid-rich blood-brain barrier, making them critical focal points for model sensitivity. By visualizing these perturbation regions, we revealed the saliency regions of the model's decision-making process. The high alignment between these regions and known chemical pharmacophores demonstrated that UGAP's learning process possesses physicochemical transparency.

Furthermore, the uncertainty estimates u provided crucial insights into predictive reliability. In cases (a) and (c), UGAP yielded high-confidence predictions with low uncertainty, accurately aligning with true labels. For the complex scaffold

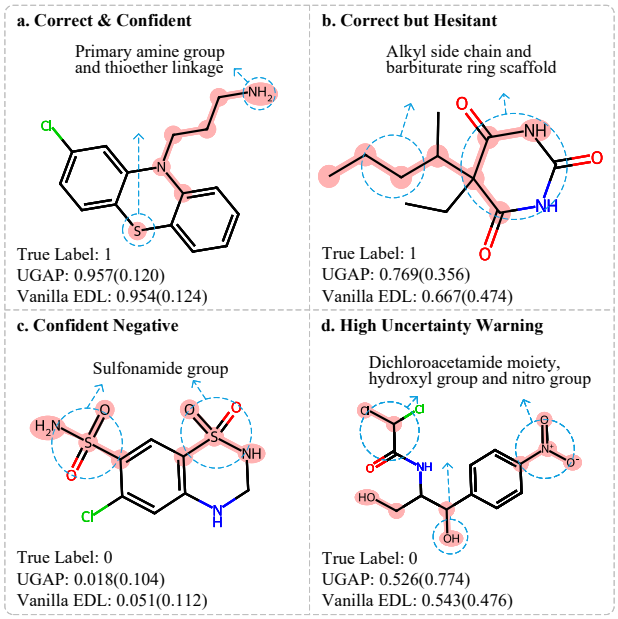

Figure 4: Visualization of Case Studies on the BBBP dataset. The light-red highlighted areas indicate the salient molecular regions identified by UGAP for perturbation, with specific chemical groups further labeled by blue dashed circles. For each molecule, the values outside the parentheses represent the predicted probability of belonging to Label 1, while the values inside the parentheses denote the corresponding uncertainty.

in case (b), UGAP exhibited a moderate uncertainty increase ($u = 0.356$), suggesting a more cautious yet correct inference—notably more confident than the Vanilla EDL, which yielded a higher uncertainty ($u = 0.474$). More importantly, case (d) triggered a significant uncertainty warning ($u = 0.774$). Despite a marginal prediction error, UGAP effectively flagged this risk with substantially higher uncertainty than the baseline ($u = 0.476$). This ability of the model to recognize its own cognitive boundaries through faithful uncertainty estimation demonstrates that UGAP provides a robust safeguard for reliable molecular property prediction.

3.7 Hyperparameter Analysis

We investigated the sensitivity of UGAP to the adversarial loss coefficient λ_1 and the evidential calibration coefficient λ_2 on the BACE and ESOL datasets (Figure 5).

Experimental results on 3D performance surfaces revealed that λ_1 served as a primary driver for predictive accuracy. Within a range centered around 1.0, increasing λ_1 consistently improved the ROC-AUC for BACE and reduced the RMSE for ESOL, which validated the efficacy of our structural-aware perturbations in enhancing generalization. Compared to λ_1, λ_2 showed a marginal impact on accuracy, allowing for uncertainty calibration with minimal loss in predictive performance.

The role of λ_2 was further evaluated for uncertainty quantification. Notably, when $\lambda_2 = 0$, the framework approximates standard evidential learning. For classification, we assessed calibration by calculating the signed calibration bias

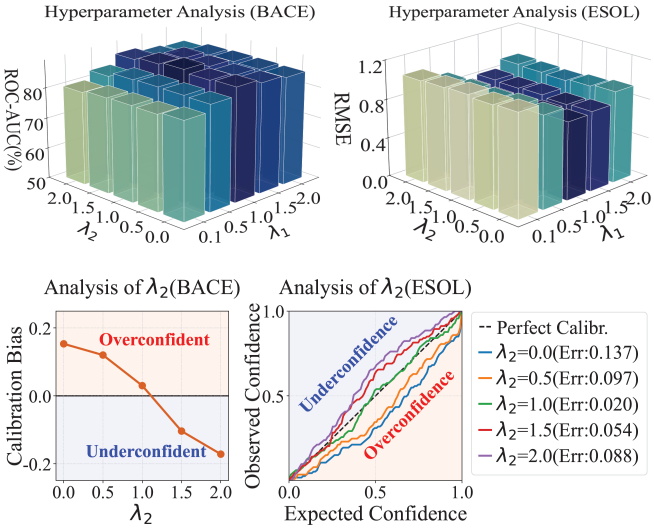

Figure 5: Hyperparameter sensitivity analysis. The top row shows 3D performance surfaces for λ_1 and λ_2. The bottom row illustrates the impact of λ_2 on uncertainty quantification.

(the difference between accuracy and confidence). It was observed that lower λ_2 often led to overconfidence, while excessively high values resulted in underconfidence. For regression, we evaluated calibration by constructing two-sided confidence intervals $[\hat{y} \pm z_p\sigma]$ across a range of theoretical confidence levels $p \in [0, 1]$. We then measured the empirical coverage by calculating the actual proportion of ground-truth values falling within these predicted intervals. As illustrated in the reliability curves, adjusting λ_2 effectively shifted the model between underconfident (above the diagonal) and overconfident (below the diagonal) states. This mechanism helped align observed and expected confidence levels, thereby improving the reliability of uncertainty estimates.

4 Conclusion

In this paper, we presented UGAP, an uncertainty-guided adaptive perturbation framework designed for molecular property prediction. By selectively applying perturbations to chemically influential regions of a molecule, our approach improved the model's generalization capabilities across various chemical domains. Extensive experiments demonstrated that UGAP consistently enhanced the predictive performance of diverse GNN backbones. Furthermore, the integrated evidential calibration mechanism improved the reliability of confidence estimates for molecular property predictions. By leveraging prediction discrepancy under perturbation, this mechanism facilitated a better alignment between estimated uncertainty and actual prediction errors. These findings suggest that UGAP offers a practical solution for uncertainty-aware molecular modeling, supporting more informed and trustworthy decision-making in the drug discovery process.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62472191, 62372204); Huazhong Agricultural University Scientific & Technological Self-innovation Foundation and Fundamental Research Funds for the Central Universities (2662024SZ006).

References

[Amini *et al.*, 2020] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14927–14937. Curran Associates, Inc., 2020.

[Gal and Ghahramani, 2016] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[Han *et al.*, 2023] Shen Han, Haitao Fu, Yuyang Wu, Ganglan Zhao, Zhenyu Song, Feng Huang, Zhongfei Zhang, Shichao Liu, and Wen Zhang. Himgnn: a novel hierarchical molecular graph representation learning framework for property prediction. *Briefings in Bioinformatics*, 24(5):bbad305, 2023.

[Jin *et al.*, 2025] Shuting Jin, Xiangrong Liu, Junlin Xu, Sisi Yuan, Hongxing Xiang, Lian Shen, Chunyan Li, Zhangming Niu, and Yinhui Jiang. Adaptive symmetry-based adversarial perturbation augmentation for molecular graph representations with dual-fusion attention information. *Information Fusion*, 120:103062, 2025.

[Jospin *et al.*, 2022] Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.

[Kang *et al.*, 2024] Linjia Kang, Songhua Zhou, Shuyan Fang, and Shichao Liu. Adapting differential molecular representation with hierarchical prompts for multi-label property prediction. *Briefings in Bioinformatics*, 25(5):bbae438, 2024.

[Kong *et al.*, 2022] Kezhi Kong, Guohao Li, Mucong Ding, Zuxuan Wu, Chen Zhu, Bernard Ghanem, Gavin Taylor, and Tom Goldstein. Robust optimization as data augmentation for large-scale graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 60–69, 2022.

[Lakshminarayanan *et al.*, 2017] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[Li and Qiu, 2021] Linyang Li and Xipeng Qiu. Token-aware virtual adversarial training in natural language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8410–8418, 2021.

[Liu *et al.*, 2021] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*, 2021.

[Liu *et al.*, 2025] Ziyang Liu, Chaokun Wang, Shuwen Zheng, Cheng Wu, Hao Feng, Li Xu, Yue Zheng, Liang Rong, and Peng Li. Molecular motif learning as a pre-training objective for molecular property prediction. *Nature Communications*, 2025.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[Qiao *et al.*, 2025] Jianbo Qiao, Junru Jin, Ding Wang, Saisai Teng, Junyu Zhang, Xuetong Yang, Yuhang Liu, Yu Wang, Lizhen Cui, Quan Zou, et al. A self-conformation-aware pre-training framework for molecular property prediction with substructure interpretability. *Nature Communications*, 16(1):4382, 2025.

[Rampášek *et al.*, 2022] Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.

[Rong *et al.*, 2020] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571, 2020.

[Sensoy *et al.*, 2018] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[Soleimany *et al.*, 2021] Ava P Soleimany, Alexander Amini, Samuel Goldman, Daniela Rus, Sangeeta N Bhatia, and Connor W Coley. Evidential deep learning for guided molecular property prediction and discovery. *ACS central science*, 7(8):1356–1367, 2021.

[Song *et al.*, 2020] Ying Song, Shuangjia Zheng, Zhangming Niu, Zhang-Hua Fu, Yutong Lu, and Yuedong Yang. Communicative representation learning on attributed molecular graphs. In *29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence (IJCAI-PRICAI2020)*. International Joint Conferences on Artificial Intelligence Organization, 2020.

[Wang *et al.*, 2022] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive

learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

[Wu *et al.*, 2018] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[Xie *et al.*, 2020] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 819–828, 2020.

[Xiong *et al.*, 2026] Zhankun Xiong, Ziyan Wang, Feng Huang, Minyao Qiu, Shuyan Fang, Liuqing Yang, Xionghui Zhou, Shichao Liu, Ping Zhang, and Wen Zhang. Multi-to-uni modal knowledge transfer pretraining for molecular representation learning. *Nature Communications*, 2026.

[Xu *et al.*, 2018] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[Xue *et al.*, 2021] Haotian Xue, Kaixiong Zhou, Tianlong Chen, Kai Guo, Xia Hu, Yi Chang, and Xin Wang. Cap: Co-adversarial perturbation on weights and features for improving generalization of graph neural networks. *arXiv preprint arXiv:2110.14855*, 2021.

[Yu *et al.*, 2022] Jie Yu, Dingyan Wang, and Mingyue Zheng. Uncertainty quantification: Can we trust artificial intelligence in drug discovery? *Iscience*, 25(8), 2022.

[Zeng *et al.*, 2022] Xiangxiang Zeng, Hongxin Xiang, Linhui Yu, Jianmin Wang, Kenli Li, Ruth Nussinov, and Feixiong Cheng. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence*, 4(11):1004–1016, 2022.

[Zhang *et al.*, 2024] Haohui Zhang, Juntong Wu, Shichao Liu, and Shen Han. A pre-trained multi-representation fusion network for molecular property prediction. *Information Fusion*, 103:102092, 2024.

[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. In *The eleventh international conference on learning representations*, 2023.