CAAI Transactions on Intelligence Technology

WILEY

ORIGINAL RESEARCH OPEN ACCESS

Addressing Long-Tailed Drug–Drug Interactions Through Minimisation of Predictive Uncertainty and Loss Sharpness

Chao Liu[1] | Xizhao Wang[1] | Farhad Pourpanah[2] | Sam Kwong[3]

[1]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China | [2]Department of Electrical and Computer Engineering, Queens University, Kingston, Ontario, Canada | [3]Department of Computing and Decision Sciences, Lingnan University, Hong Kong, China

Correspondence: Xizhao Wang (xzwang@szu.edu.cn)

Received: 23 August 2025 | **Revised:** 16 June 2026 | **Accepted:** 8 July 2026

Keywords: drug–drug interaction prediction | dynamic reweighting | long-tailed distribution | predictive uncertainty | sharpness-aware minimisation

ABSTRACT

Drug–Drug Interaction (DDI) prediction is critical for ensuring patient safety, particularly under long-tailed distributions, where frequent (head) interactions dominate whereas rare (tail) interactions remain underrepresented. Conventional loss functions such as cross-entropy often tend to overfit head classes while they underperform on rare but clinically important classes. Recent studies have shown that sharpness-aware minimisation (SAM) improves generalisation by encouraging solutions that lie in flatter regions of the loss landscape. Given that the loss function plays a fundamental role in model training and that uncertainty estimation is crucial for handling ambiguous or difficult samples, we propose the uncertainty and SAM (USAM) loss for improving DDI prediction under long-tail scenarios. The USAM loss integrates predictive uncertainty with SAM via two key components: (i) a dynamic reweighting strategy based on uncertainty, which leverages prediction entropy to emphasise difficult samples and (ii) a class-balanced SAM regularisation term, which encourages flatter minima and enhances generalisation across imbalanced classes. Extensive experiments on four long-tailed DDI benchmarks demonstrate that the USAM loss consistently outperforms existing methods in terms of both F1 score and recall. These results highlight the effectiveness of incorporating uncertainty modelling with sharpness-aware optimisation in addressing long-tailed problems.

1 | Introduction

In recent years, the rapid development of artificial intelligence (AI) has provided new technological pathways for clinical decision support and demonstrated promising applications across multiple medical scenarios. For example, in the noninvasive diagnosis of clinically significant portal hypertension, AI-based methods have significantly improved diagnostic efficiency and decision-making quality [1]. In the treatment of complex diseases such as diabetes and cancer, polypharmacy has become an important and promising strategy [2]. However, although it can improve therapeutic efficacy and reduce drug resistance, it also increases the risk of adverse drug reactions, which in severe cases may lead to life-threatening events. Against this background, DDI has emerged as a key factor affecting the safety of combination therapy. AI-driven DDI prediction methods aim to identify potential interaction risks and thereby reduce the occurrence of adverse drug reactions. Therefore, enhancing the performance of DDI prediction models is critical for ensuring patient safety and supporting the formulation of rational treatment plans [3].

In clinical practice, misclassification costs are highly asymmetric. Head classes typically correspond to common and mild

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2026 The Author(s). *CAAI Transactions on Intelligence Technology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.

interactions (e.g., mild gastrointestinal discomfort), whereas tail classes often represent rare but severe or even fatal interactions (e.g., severe bleeding events involving warfarin). Missing tail-class interactions (false negatives) can thus pose life-threatening risks. Consequently, prioritising improved recall for tail classes while maintaining overall reliability is a critical clinical requirement. By reducing false negatives, improved tail-class recall can better support clinical decision-making by mitigating the risk of missing high-severity DDIs, thereby facilitating safer prescribing in patients receiving complex polypharmacy treatments.

As more new drugs are introduced to the market, their interactions with other drugs remain unclear. Traditional DDI prediction methods rely on in vitro experiments and clinical validation. While these methods offer high accuracy, they are costly and inefficient, making it difficult to meet the practical demands of massive drug combination scenarios [4]. In recent years, with the continuous accumulation of biomedical data and the rapid development of machine learning technologies, data-driven DDI prediction methods have emerged. These methods typically integrate multi-source data such as SMILES molecular structures, target protein information, knowledge graphs and drug properties, using deep neural networks to model potential drug interactions, achieving promising results [5].

Although existing methods have achieved significant results on multiple datasets, most overlook the common long-tailed distribution in DDI data [6, 7]. As shown in Figure 1, in the DDI-DB171 [5] dataset, a few high-frequency interaction types (head classes) account for the majority of samples, whereas many low-frequency interaction types (tail classes) are scarce. Because the training process minimises overall loss, models tend to focus on head classes with abundant training samples, leading to poor performance on tail classes [5, 8]. This class imbalance significantly limits model reliability and generalisation in real clinical settings.

Furthermore, models tend to output low confidence for tail-class samples and ambiguous class boundaries, resulting in high predictive uncertainty, reflected in the large entropy of the predicted probability distribution [9]. Several studies have attempted to address these challenges through loss function design. Class-balanced (CB) loss [10] reweights cross-entropy (CE) loss based on class frequency, alleviating imbalance but ignoring sample difficulty; under extreme skew, such frequency-driven reweighting may degrade calibration performance because of overemphasis on tail classes. Focal loss (FL) [11] focuses training on hard samples via confidence-based modulation, but it is sensitive to noise and may shift attention towards unstable examples in the presence of label ambiguity. Tail focal loss (TFL) [5] further emphasises rare DDI types, but it only considers the ground-truth class and ignores the full prediction distribution, limiting inter-class modelling. Overall, under severe class imbalance, CE loss is dominated by head classes and yields poor tail-class recall, FL is vulnerable to noisy hard samples and CB loss may harm calibration because of aggressive reweighting. These limitations hinder effective tail-class learning.

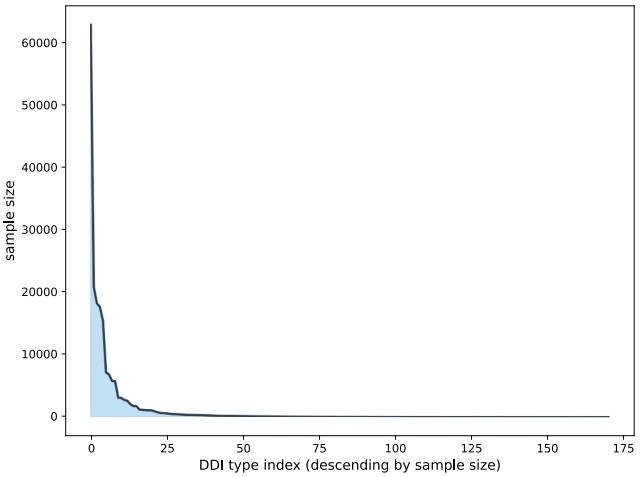


FIGURE 1 | DDI types follow a long-tail distribution in the DDI-DB171 dataset.

To enhance generalisation, sharpness-aware minimisation (SAM) [12] has recently been proposed to simultaneously minimise the training loss and the sharpness of the loss landscape. By guiding model parameters towards flatter local minima, SAM emphasises neighbourhood loss stability, thereby reducing overfitting and improving performance on complex or long-tailed distributions [13]. To illustrate this effect, Figure 2 presents a 2D contour visualisation comparing the loss landscapes of the standard CE loss and the proposed USAM loss. As shown in the figure, although the standard CE loss tends to converge to a sharp minimum characterised by dense contours and steep gradients, USAM loss yields a flatter loss landscape, resulting in a broader and more stable optimum.

In this work, we propose USAM loss, a unified framework for long-tailed DDI prediction that leverages predictive uncertainty and sharpness-aware soptimisation to improve learning under class imbalance. The key contributions of this work are as follows:

1. We propose USAM loss, a unified framework for long-tailed DDI prediction that jointly incorporates uncertainty-aware weighting and class-balanced sharpness-aware optimisation.
2. We design an entropy-based uncertainty weighting mechanism that captures predictive uncertainty over the full output distribution and adaptively emphasises difficult samples.
3. We perform extensive ablation studies and comparisons with representative long-tail learning methods, including FL, TFL and their SAM-based variants, to quantify the contribution of each component.
4. Experimental results on benchmark DDI datasets demonstrate that USAM loss achieves state-of-the-art performance

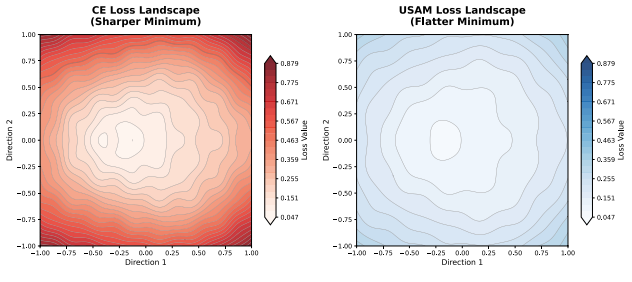


FIGURE 2 | Loss landscape contour: CE loss versus USAM loss.

while maintaining balanced performance across head, medium and tail classes.

The remainder of this paper is organised as follows: Section 2 reviews related work on DDI prediction methods, long-tailed learning strategies and the theory of SAM. Section 3 formalises the task. Section 4 presents the design motivation and modelling details of the proposed USAM loss. Section 5 provides experimental settings and a comprehensive analysis of the results. Section 6 concludes the paper and discusses future research directions.

2 | Related Works

This section first reviews representative methods for DDI prediction, then discusses typical loss function designs in the field of long-tailed learning and finally summarises the theory of SAM.

2.1 | Drug–Drug Interaction Prediction

Existing DDI prediction methods broadly fall into two categories: similarity-based and graph-based approaches [5]. Similarity-based methods rely on the fundamental assumption that similar drugs tend to produce similar DDIs. DDIMDL [14] employs deep neural networks to extract diverse modal features of drug molecules and subsequently fuses the outputs of various models through voting. MDDI-SCL [6] enhances drug pair representations using supervised contrastive learning combined with multi-scale feature fusion. MDF-SA-DDI [7] combines autoencoders and Transformer [15] self-attention mechanisms to improve similarity learning. SSF-DDI [16] jointly models molecular sequences and substructure features to boost prediction performance on unknown drugs. SRR-DDI [17] refines substructure features through self-attention mechanisms and incorporates drug similarity information to improve generalisation. These methods excel at integrating chemical structures and other features, but simply hashing molecular structures into binary vectors may lose important spatial and semantic information in SMILES representations.

Graph-based methods focus on leveraging knowledge graphs (KGs) and chemical structural spatial information to learn drug embeddings through neighbourhood aggregation. This helps reveal latent relationships. For example, KGNN [18] captures high-order topology and semantic information via multi-hop neighbourhood aggregation. LaGAT [19] designs a link-aware graph attention mechanism to generate semantically relevant neighbour selection paths for specific drug pairs. DSN-DDI [20] learns local and global drug representations simultaneously, balancing features of single drugs and drug pairs. MDNN [21] introduces a dual-path structure that incorporates drug KG paths and heterogeneous feature paths. MUFFIN [22] models molecular structures and KG features jointly, mitigating annotation scarcity through multi-scale feature fusion. HMGRL [23] integrates explicit and implicit multi-relation associations in heterogeneous graphs to achieve multi-view fusion. RaGSECo [24] fuses multi-relation graphs with drug similarity graphs and performs collaborative contrastive learning from both interaction and similarity perspectives. MeTDDI [25] achieves interpretable prediction using local-global self-attention and motif modelling. KnowDDI [26] uncovers structural mechanisms of DDI through knowledge subgraph learning. Mol-DDI [27] relies on the 'structure determines function' hypothesis and achieves high-precision prediction using only molecular geometric structure information. DrugDAGT [28] combines bond-level and atom-level dual attention mechanisms with graph contrastive learning to improve interpretability for multiple DDI types. Du et al. [29] introduce a searchable mechanism for subgraph selection and encoding spaces to enable structured modelling tailored to specific data. DMFDDI [30] fuses molecular graphs, DDI networks and biochemical similarity features. It enhances prediction performance via multi-modal attention mechanisms. Additionally, DFPDDI [31] incorporates edge-aware enhanced contrastive learning and mutual information constraints to better distinguish different types of drug relationships.

Although these methods achieve significant advances in overall performance, most still assume balanced data distributions. This assumption limits their effectiveness in handling long-tail distribution challenges commonly seen in DDI prediction. To address this, Zheng et al. [5] propose the TFMD framework. It integrates multi-modal information and tail-enhancing strategies to significantly improve prediction performance on long-tail datasets. Overall, designing models that balance both overall performance and tail-class prediction capability remains an important future research direction in DDI prediction.

2.2 | Advances in Long-Tailed Learning

Long-tailed distribution is a common problem in deep learning. Existing methods mainly focus on three aspects: loss function design, feature and data enhancement and fine-tuning strategies. For loss functions, weighted cross entropy (WCE) loss and Meta-Softmax (BALMS) loss [32] adjust weights or logits based on class frequency. Label-Distribution-Aware Margin (LDAM) loss [33] aims to enlarge the decision boundary for minority classes. FL [11] focuses on giving more weight to hard-to-classify samples. For DDI tasks, TFL [5] builds on FL and increases the gradient for tail classes to reduce gradient vanishing. In addition, GCL [34] reduces Softmax saturation using Gaussian noise and CB [10] loss balances class weights based on the number of effective samples. Recently, Chen et al. [35] propose uncertainty calibration based on knowledge transfer, which adjusts the prediction distribution using head class statistics and helps reduce overconfidence in head classes. For feature and data enhancement, H2T [36] transfers the semantics of head classes to tail classes to correct the decision boundary. SHIKE [37] uses an expert mixture structure to increase feature diversity. Context-based oversampling [38] and image mixing based on optimal transport [39] increase the diversity and semantic quality of minority samples. In fine-tuning strategies, LIFT [40] shows that light fine-tuning can reduce class inconsistency. GNM-PT [41] finds flatter loss areas through Gaussian neighbourhood minimisation, which improves model generalisation. LPT [42] introduces trainable prompts into a frozen pre-trained model and uses a two-stage training strategy to learn general features and class-specific features separately, to adapt to long-tail data distributions.

Different from these methods, this paper proposes a loss function that combines normalised entropy and class weights based

on the effective number of samples. The method models prediction uncertainty and increases the margin between the true class and other classes, which improves the performance of tail class classification.

2.3 | Applications of Sharpness-Aware Minimisation

In recent years, sharpness-aware minimisation (SAM) has attracted increasing attention as an effective optimisation strategy to enhance the generalisation ability of deep learning models. Proposed by Foret et al. [12], SAM introduces perturbations within the neighbourhood of model parameters to guide the model towards flatter minima, thereby reducing the risk of overfitting and improving robustness to data perturbations.

With the advancement of research, SAM is gradually applied to long-tailed classification tasks. ImbSAM [13] applies SAM optimisation exclusively to tail classes to enhance their generalisation. CC-SAM [43] performs SAM optimisation independently for each class, with a class-specific perturbation radius that increases from head to tail classes, thereby balancing generalisation across classes. Focal-SAM [44] integrates the hard-sample modelling mechanism of FL with SAM, effectively improving the model's ability to learn from tail classes and ambiguous boundary samples. SSE-SAM [45] adopts a staged optimisation strategy, focusing on head-class samples during the early training phase and gradually shifting towards tail classes in later stages, aiming to achieve more balanced learning across different classes.

In addition, the application of SAM extends to other tasks. Wang et al. propose the U-MedSAM [46] model, which integrates an uncertainty-aware loss function with the SAM optimiser and achieves excellent performance in medical image segmentation. Ma et al. [47] apply SAM optimisation to fault diagnosis tasks, significantly improving model generalisation under noisy labels.

These studies suggest that combining uncertainty modelling with SAM optimisation effectively improves model generalisation. Different from the aforementioned works, our method simultaneously considers predictive uncertainty and class-based SAM optimisation in long-tailed data scenarios. Experimental results demonstrate that the proposed method significantly enhances the generalisation performance of DDI prediction models. Table 1 presents a comparison of representative SAM-based optimisation methods.

3 | Problem Formulation and Preliminaries

In this section, we will provide a formulated description of DDI and introduce some of the preliminary concepts used in this paper.

3.1 | Task Definition of DDI Prediction

Given: (1) A drug set $\mathcal{D} = \{d_1, d_2, ..., d_M\}$ containing M unique drugs; (2) A feature vector $x_m \in \mathbb{R}^d$ for each drug d_m and (3) A training dataset of known DDI types structured as

$$\mathcal{S} = \{(e_i, y_i)\}_{i=1}^{|\mathcal{S}|}, \tag{1}$$

where $e_i = (d_m, d_n)$ represents the i-th interacting drug pair ($m < n$ assuming undirected interactions) and $y_i \in \{1, 2, ..., K\}$ denotes its corresponding interaction type among K specific classes.

The task is to learn a prediction function f_θ, such that for any drug pair $e_i = (d_m, d_n)$, the model outputs a probability distribution over the K interaction types:

$$f_\theta(x_m, x_n) = \mathbf{p}_i = [p_{i,1}, p_{i,2}, ..., p_{i,K}], \quad \text{s.t.} \quad \sum_{k=1}^{K} p_{i,k} = 1. \tag{2}$$

The objective is to minimise the discrepancy between the predicted distribution $\mathbf{p}_i$ and the ground-truth label y_i. Across a training batch of N samples, this is typically optimised using the multi-class CE loss:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(y_i = k) \log p_{i,k} = -\frac{1}{N}\sum_{i=1}^{N} \log p_{i,y_i}, \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function. The term p_{i,y_i} precisely corresponds to the predicted probability of the i-th sample for its true label y_i. This basic formulation forms the foundational supervision signal, which will be further extended in Section 4 to address long-tailed distributions and prediction uncertainty.

TABLE 1 | Comparison of representative SAM-based optimisation methods.

Method	Hard-sample	Class handling	Hyperparameters	Cost	Limitations
ImbSAM	Tail emphasis	Class-aware SAM	Sampling ratio	High (SAM)	Tail-focused; may weaken global class consistency.
CC-SAM	None	Class-conditioned radius	Class-specific radius	High (SAM)	Requires class-dependent tuning; higher training complexity.
Focal-SAM	$(1 - p_t)^\gamma$	None	γ	High (SAM)	Focuses on confidence; ignores inter-class ambiguity.
SSE-SAM	Stage-wise	Head/tail scheduling	Schedule design	High (SAM)	Relies on manual schedules; limited flexibility.
USAM (ours)	Entropy-based	Class-balanced SAM	α, β	High (SAM)	Uses fixed class weights from data priors; retains SAM overhead.

3.2 | Preliminaries

3.2.1 | Loss Functions

As shown in Equation (4), FL [11] introduces a modulating factor $\left(1 - p_{i,y_i}\right)^{\gamma}$ to extend the conventional CE loss, where p_{i,y_i} denotes the model's predicted probability of the i-th sample for its ground-truth class y_i. This formulation effectively reduces the loss contribution of easily classifiable examples (when p_{i,y_i} is close to 1) while increasing emphasis on hard examples (when p_{i,y_i} is close to 0). By dynamically adjusting weights based on classification confidence, the method demonstrates superior performance in scenarios with extreme class imbalance, particularly in tasks like object detection and multi-label classification where foreground-background imbalance is prevalent.

$$\mathcal{L}_{\mathrm{FL}} = -\frac{1}{N}\sum_{i=1}^{N}\left(1 - p_{i,y_i}\right)^{\gamma}\log p_{i,y_i}. \tag{4}$$

As shown in Equation (5), CB loss [10] adopts a class-level weighting strategy based on the 'effective number of samples theory. The method assigns class-specific weights through the formulation $w_c = \frac{1-\beta}{1-\beta^{n_{y_i}}}$, where n_{y_i} denotes the number of samples in class y_i and β serves as a hyperparameter controlling the balance intensity. This weighting mechanism can be combined with loss functions like CE or FL to further mitigate the underfitting issue of minority classes in imbalanced datasets.

$$\mathcal{L}_{\mathrm{CB}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1-\beta}{1-\beta^{n_{y_i}}}\log p_{i,y_i}. \tag{5}$$

BS loss [32] addresses the class imbalance problem from a probabilistic modelling perspective by correcting the normalisation bias inherent in conventional Softmax under imbalanced distributions. It incorporates class frequency priors directly into the denominator of the Softmax function, as formulated in Equation (6), thereby rebalancing the contribution of each class during probability calibration. This probabilistic reweighting strategy effectively mitigates the prediction bias towards majority classes while preserving the discriminative power of the learn representations.

$$\mathcal{L}_{\mathrm{BS}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log p_{i,y_i} + \log\frac{1}{\sum_{k=1}^{K} n_k}\right]. \tag{6}$$

This method does not require weighting or restructuring the loss function, but rather naturally adapts to long-tailed distributions by adjusting the normalisation process of logits, demonstrating strong performance particularly when class distributions are consistent between the training and test sets.

TFL [5] is designed to more effectively model tail classes in DDI prediction tasks. The loss function is defined as follows:

$$\mathcal{L}_{\mathrm{TFL}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\left(1 - p_{i,y_i}\right)^{\gamma}\log p_{i,y_i} + \log p_{i,y_i}^{\beta F_{T_S}(y_i)}\right], \tag{7}$$

where:

$$F_{T_S}(y_i) = \begin{cases} 1, & \text{if } P_{y_i} > T_S \\ 0, & \text{otherwise.} \end{cases}, \quad P_{y_i} = \frac{\sum_{m=0}^{M} n_m}{\sum_{k=0}^{N} n_k}. \tag{8}$$

This function augments the original FL with an additional loss term $\log p_{i,y_i}^{\beta F_{T_S}(y_i)}$ to counteract gradient reduction for tail samples in FL, thereby preventing vanishing gradients. Where p_{i,y_i} denotes the predicted probability of the ground-truth label, $F_{T_S}(y_i)$ is a threshold function determining whether the input class belongs to the tail classes and β serves as a hyperparameter controlling the penalty intensity applied to tail classes. The threshold function $F_{T_S}(y_i)$ returns 1 if the corresponding class y_i is a tail class, otherwise 0.

3.2.2 | Sharpness-Aware Minimisation

SAM is an optimisation method designed to improve the generalisation ability of models. Its core idea is to minimise the loss at the current parameter point and consider the worst-case loss within the neighbourhood of the parameters. This guides the model to converge to a flatter optimum, thereby enhancing its robustness against input perturbations and overfitting.

Specifically, SAM defines the flatness of the loss function $\mathcal{L}(\theta)$ at the network parameters θ as follows:

$$\mathcal{L}_{\mathrm{sam}}(\theta) := \max_{\|\epsilon\|_2 \le \rho}\left[\mathcal{L}(\theta + \epsilon) - \mathcal{L}(\theta)\right], \tag{9}$$

where ϵ denotes the perturbation vector added to the parameters θ and ρ is a hyperparameter that controls the radius of the perturbation. The optimisation objective of SAM is to minimise $\mathcal{L}_{\mathrm{sam}}(\theta)$, which involves searching for the direction of perturbation within the ρ-neighbourhood that leads to the maximum loss. This encourages the model to maintain consistently low loss within the local neighbourhood, thereby enhancing its generalisation ability to unseen samples.

4 | Our Proposed Loss: USAM

This section systematically introduces the proposed USAM loss function in three aspects: the definition of the loss function, the optimisation strategy and its application to the DDI prediction task. The method integrates predictive uncertainty and parameter perturbation information to enhance the model's generalisation ability in complex tasks involving long-tailed data.

4.1 | Definition of the USAM Loss Function

We design the weighting strategy of the USAM loss at both the individual sample level and the class level. Specifically, we use the model's predicted entropy for each sample to quantify its classification difficulty and apply dynamic weighting accordingly. A higher predicted entropy indicates greater uncertainty for the sample and thus warrants a larger loss weight. Additionally, if the sample belongs to a tail class, it should be assigned a higher loss weight. The definition of the USAM loss is as follows:

$$\mathcal{L}_u(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left(1 + H_i + \frac{1-\beta}{1-\beta^{n_{y_i}}}\right)\log p_{i,y_i}, \tag{10}$$

where N denotes the number of samples in the batch, p_{i,y_i} represents the predicted probability of the i-th sample for its true label y_i, n_{y_i} stands for the number of samples belonging to class y_i, θ is the model parameter and H_i is the normalised entropy of the predicted distribution for the sample, defined as follows:

$$H_i = -\frac{\sum_{k=1}^{K} p_{ik} \log p_{ik}}{\log K}, \tag{11}$$

where K denotes the total number of classes. The loss weights in Equation (10) consist of three parts: (1) Baseline ('1'): It provides standard Cross Entropy supervision and keeps the gradients stable for all samples. (2) Sample-level (H_i): It gives more weight to hard samples with high prediction uncertainty. (3) Class-level $\left(\frac{1-\beta}{1-\beta^{n_{y_i}}}\right)$: It adjusts for the long-tailed distribution by increasing the weight of minority classes.

Unlike a multiplicative approach, a single small factor (e.g., when $H_i \to 0$ for easy samples) does not make the whole loss close to zero. Furthermore, this additive structure successfully separates sample difficulty and class imbalance. For example, a hard head sample gets extra weight from H_i, but its total loss is limited because its class weight is near zero. This prevents the model from ignoring hard head samples or paying too much attention to head classes. In contrast, a hard tail sample gets large weights from both parts. As shown in Figure 3, the average normalised predictive entropy of tail classes is always higher than that of head classes during the whole training process, which proves this design is reasonable and effective.

We selected predictive entropy as the uncertainty measure based on both computational and task-specific considerations. First, alternative uncertainty measures, such as predictive variance and mutual information, are typically estimated using Bayesian neural networks or Monte Carlo dropout, which require multiple forward passes per sample. Because the SAM optimiser already requires two forward-backwards passes per optimisation step, combining SAM with these approaches would introduce substantial computational overhead. In contrast, predictive entropy can be computed directly from a single forward pass, making it highly computationally efficient. Second, DDI prediction is a large-scale multi-class classification task involving up to 171 interaction types. Predictive entropy captures the uncertainty over the entire output distribution and therefore reflects the model's confusion among multiple interaction types. This property makes it particularly suitable for the complex DDI setting.

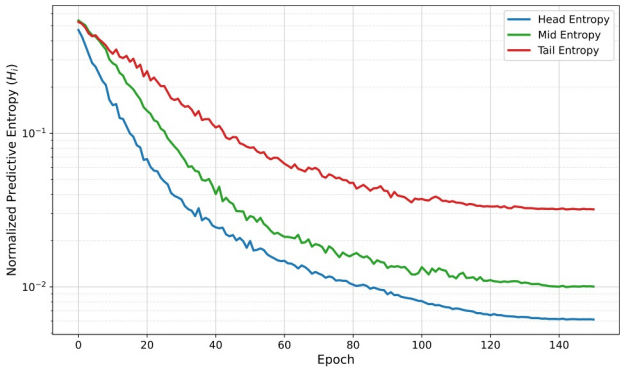


FIGURE 3 | Normalised predictive entropy across epochs under standard cross-entropy loss on the DDI-DB171 test set.

4.1.1 | *SAM Regularisation*

To further improve the generalisation performance of the model and reduce the risk of converging to sharp local minima, we introduce SAM as a regularisation term. The key idea of SAM is to optimise the worst-case loss within the neighbourhood of the model parameters, thereby guiding the model towards flatter minima and enhancing its generalisation ability. Its mathematical formulation can be expressed as a bilevel optimisation problem:

$$\mathcal{L}_{\text{sam}}(\theta) = \max_{\|\epsilon\|_2 \le \rho} [\mathcal{L}_u(\theta + \epsilon) - \mathcal{L}_u(\theta)]. \tag{12}$$

The inner optimization aims to find a perturbation vector ϵ within an L_2 ball centred at the parameter θ with radius ρ, such that the loss increment $\mathcal{L}_u(\theta + \epsilon) - \mathcal{L}_u(\theta)$ is maximised. The outer optimisation then adjusts the parameter θ to minimise this maximum loss increment.

When θ lies at a sharp minimum, small perturbations ϵ can cause a significant increase in the loss (i.e., a large max[·] value). In contrast, flat minima are less sensitive to perturbations (i.e., a small max[·] value). By minimising the maximum loss increment, the parameters are guided towards flatter regions, thereby enhancing the model's robustness to both parameter and data perturbations.

Therefore, the final USAM loss is defined as follows:

$$\mathcal{L}(\theta) = \mathcal{L}_u(\theta) + \alpha \mathcal{L}_{\text{sam}}(\theta), \tag{13}$$

where α is a tuning parameter used to balance the weights between the training loss and the SAM regularisation term. By minimising $\mathcal{L}(\theta)$, the model not only fits well on the training data (corresponding to $\mathcal{L}_u(\theta)$) but also tends to lie in a flatter parameter region (corresponding to $\mathcal{L}_{\text{sam}}(\theta)$), thereby enhancing its generalisation ability and robustness.

4.2 | Optimization Strategy of USAM Loss

In each iteration, the USAM loss addresses the min–max optimisation problem in two sequential steps. First, it approximates the inner maximisation problem via a first-order Taylor expansion around the current parameter θ. An adversarial perturbation is generated along the gradient direction to maximise the loss within a bounded neighbourhood:

$$\hat{\epsilon}(\theta) \approx \rho \frac{\nabla_\theta \mathcal{L}_u(\theta)}{\|\nabla_\theta \mathcal{L}_u(\theta)\|_2}. \tag{14}$$

Second, USAM loss updates the model parameters by minimising the sharpness-adjusted joint loss. The corresponding gradient is computed as follows:

$$\begin{aligned} \nabla_\theta \mathcal{L}(\theta) &\approx \nabla_\theta (\mathcal{L}_u(\theta) + \alpha[\mathcal{L}_u(\theta + \hat{\epsilon}) - \mathcal{L}_u(\theta)]) \\ &\approx (1 - \alpha)\nabla_\theta \mathcal{L}_u(\theta) + \alpha \nabla_\theta \mathcal{L}_u(\theta + \hat{\epsilon}). \end{aligned} \tag{15}$$

This two-step formulation seamlessly unifies predictive uncertainty and sharpness-aware optimisation. Specifically, the adversarial perturbation $\hat{\epsilon}(\theta)$ explicitly probes the local curvature, steering the model away from sharp, error-prone regions. Subsequently, the ensemble gradient in Equation (15) balances standard empirical risk minimisation with sharpness regularisation, weighted by α. By penalising steep loss fluctuations within the parameter neighbourhood, USAM loss effectively guides the optimisation towards flatter minima, thereby bolstering generalisation and robustness in complex scenarios.

4.3 | USAM Loss for the DDI Prediction Task

As shown in Figure 4, the proposed framework comprises two core modules: a drug representation learning module and a DDI prediction module. The drug representation learning module leverages four modality-specific features, including SMILES strings, molecular graphs, targets and enzymatic pathways, to generate drug embedding vectors. Building on these representations, the DDI prediction module concatenates drug pairs at the feature level and feeds them into a multi-layer perceptron (MLP) classifier to predict interaction types.

As shown in Figure 4, to fairly evaluate the performance of the proposed USAM loss, we adopted the same drug representation learning module as used in TFMD [5]. TFMD introduces a modality enhancement mechanism that enhances the representation power of primary modalities by fusing auxiliary modalities. Specifically, the model utilises SMILES information to enhance the representation of the drug graph structure, whereas employing enzyme information to enhance the representation of targets. Because Graph Neural Networks (GNNs) struggle to capture three-dimensional structural features, such as chirality, inherent in SMILES strings, SMILES is employed as a supplement for spatial information to augment the spatial expressive capability of the drug graph. Concurrently, the enzyme modality facilitates the modelling of specific interactions between enzymes and targets, thereby enriching the biological semantics of target representations. Furthermore, TFMD utilises max pooling to extract core features from each modality and subsequently fuses these multimodal features. Specifically, for a given drug d_i, TFMD encodes its four modalities to obtain their respective representation vectors: $\mathbf{s}_i$ (SMILES), $\mathbf{G}_i$ (Graph structure), $\mathbf{T}_i$ (Target) and $\mathbf{e}_i$ (Enzyme). These four representations are then fused via a multi-layer neural network to derive the final representation x_i of d_i, which is subsequently used for DDI type prediction.

The final classifier adopts a fully connected neural network architecture, which takes the enhanced multi-dimensional feature vector as input and maps it from the feature space to the label space through a nonlinear activation function.

The algorithm procedure of USAM loss for the DDI prediction task is shown in Algorithm 1.

ALGORITHM 1 | USAM loss for DDI prediction task.

Require: Training set $\mathcal{S} = \{(e_i, y_i)\}_{i=1}^{|\mathcal{S}|}$, where pair $e_i = (d_m, d_n)$ corresponds to features (x_m, x_n); hyperparameter α; perturbation radius ρ; learning rate η

Ensure: The model parameter θ trained with USAM loss

1: Initialise the model parameter θ_0 and train step $t = 0$
2: **while** not converged **do**
3: Sample batch $B = \{(e_i, y_i)\}_{i=1}^{b}$
4: Compute base loss $\mathcal{L}_u(\theta)$
5: Compute $\nabla_\theta \mathcal{L}_u(\theta)$ and $\hat{\epsilon}(\theta)$ according to Equation (14)
6: Perturb θ with $\hat{\epsilon}(\theta)$, and compute gradient $g_1 = \nabla_\theta \mathcal{L}_u(\theta)|_{\theta+\hat{\epsilon}}$
7: Compute $g_2 = (1-\alpha)\nabla_\theta \mathcal{L}_u(\theta)|_\theta$
8: Update weights: $\theta_{t+1} \leftarrow \theta_t - \eta(\alpha g_1 + g_2)$
9: $t = t + 1$
10: **end while**

5 | Experiments

In this section, we will provide a detailed description of the experimental setup and thoroughly analyse the experimental results.

5.1 | Datasets and Evaluation Metrics

5.1.1 | *Datasets*

Our experiments were conducted on four benchmark DDI datasets adopted from TFMD [5]: DDI-DB110, DDI-DB171, DDIMDL and MUFFIN. Among them, DDI-DB110 is a subset of DDI-DB171, obtained by ranking DDI types according to their

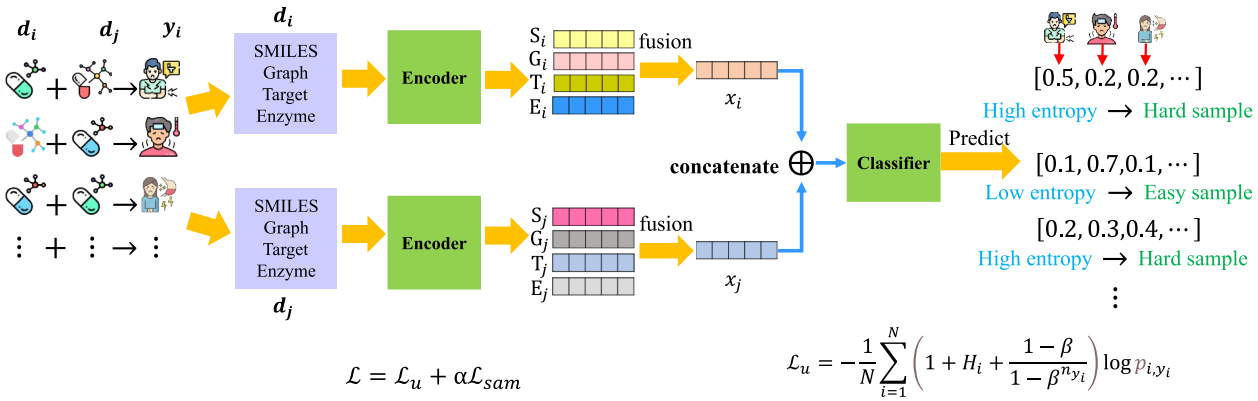


FIGURE 4 | USAM loss for the DDI prediction task.

sample counts and retaining the 110 most frequent classes. In contrast, DDI-DB171 contains all 171 interaction types, including many classes with very limited samples, resulting in a more severe long-tailed distribution.

DDIMDL and MUFFIN are widely used benchmark datasets for DDI prediction and are included to evaluate the generalisability of different methods across datasets with varying levels of class imbalance. To quantitatively characterise the degree of class imbalance, we report the class imbalance ratio (CIR), normalised entropy (H_{norm}) and Gini coefficient for each dataset. Lower values of H_{norm} and higher Gini coefficients correspond to more severe class imbalance. As shown in Table 2, DDI-DB171 exhibits the most severe long-tailed characteristic ($H_{\text{norm}} = 0.5292$, Gini $= 0.9172$), whereas DDI-DB110, DDIMDL and MUFFIN represent relatively less extreme but still highly imbalanced settings. Detailed statistics of all datasets are summarised in Table 2.

5.1.2 | *Evaluation Metrics*

We evaluate the performance of the USAM loss on long-tailed tasks using six standard evaluation metrics, namely Accuracy, Precision, Recall, F1-score, AUC and AUPR.

5.2 | Baselines and Experiments Setup

5.2.1 | *Baseline*

5.2.1.1 | *DDI Baselines.* Our comparative analysis includes one conventional statistical method, Support Vector Machine (SVM), alongside six state-of-the-art (SOTA) models recognised for DDI prediction: DDIMDL [14], MUFFIN [22], MDF-SA-DDI [7], MDDI-SCL [6] and TFMD [5]. DDIMDL operates as a multimodal framework, integrating diverse drug data sources through a straightforward late fusion strategy. MUFFIN employs CNN-based feature extraction for its bimodal fusion approach. MDF-SA-DDI leverages an auto-encoder combined with multi-head attention to integrate DDI vector representations. MDDI-SCL introduces a contrastive learning paradigm specifically designed for DDI classification tasks. Notably, MDF-SA-DDI and MDDI-SCL incorporate FL to mitigate challenges posed by imbalanced class distributions. Several benchmarks, including MDF-SA-DDI, MDDI-SCL and TFMD, have demonstrated superior performance compared to alternative approaches like KGNN.

5.2.1.2 | *Loss Function Baselines.* For loss functions, we evaluate standard CE alongside five established techniques prominent in computer vision for addressing long-tailed data distributions: WCE, FL [11], CB [10], BS [32], LDAM [33] and TFL [5] loss.

5.2.2 | *Implementation Details*

USAM loss operates on the PyTorch framework, with all experiments executed on Tesla V100 GPUs for training and testing. To ensure fair comparison with baseline methods in DDI prediction, all encoder and classifier parameters adopt configurations from TFMD. Specifically, for SMILES chemical notations, a pre-trained Transformer encoder captures long-range dependencies in molecular sequences through self-attention mechanisms, generating feature vectors with chemical bond connectivity data; for molecular graph structures, GNNs perform topological encoding where message-passing mechanisms extract spatial configuration features between atoms; regarding target/enzyme features, binary vector intersection operations compute drug-pair Jaccard similarity.

The model uses a learning rate of 0.003 across 150 epochs with a batch size of 2048. The search range for α is {0.1, 0.25, 0.5, 0.75, 1.0} and ρ is set to 0.05. The weight β for the class-balanced loss was searched over {0.99, 0.999, 0.9999} and set to 0.999 based on experimental results.

5.3 | Main Results

Based on the results in Table 3, USAM loss outperforms existing methods across all four datasets, demonstrating a significant overall advantage. On the DDI-DB110 dataset, USAM loss achieves an F1 score of 0.9203, corresponding to an absolute improvement of 1.75% points over the best-performing baseline, TFMD ($F1 = 0.9028$). On the more challenging DDI-DB171 dataset, USAM loss attains an F1 score of 0.8580, exceeding TFMD ($F1 = 0.7995$) by 5.85% points, highlighting its robust generalisation capability under imbalanced scenarios. On the DDIMDL dataset, USAM loss achieves an F1 score of 0.8913, representing an absolute improvement of 4.27% points compared to TFMD ($F1 = 0.8486$). On the MUFFIN dataset, USAM loss reaches an F1 score of 0.9618, surpassing TFMD ($F1 = 0.9458$) by 1.60% points. In contrast, other methods, although showing strong performance on certain metrics, often struggle to maintain overall effectiveness. Overall, USAM loss effectively mitigates the challenges posed by class imbalance and significantly enhances the model's ability to distinguish between highly similar samples, achieving leading performance across multiple key metrics.

5.4 | Discussions

5.4.1 | *Comparison of Different Loss Functions*

From the results in Table 4, it can be observed that USAM loss outperforms the compared methods on most metrics across both datasets. For example, on DDI-DB110, USAM loss achieves an accuracy of 0.9536 and an F1 score of 0.9222, both

TABLE 2 | The Statistics and division of datasets.

Dataset	DDI	Class	Drug	Feature	CIR	H_{norm}	Gini
DDIMDL	37,243	65	569	GSTEP	3270	0.5635	0.8575
MUFFIN	172,426	81	1569	GS	5243	0.5698	0.8652
DDI-DB110	198,631	110	1178	GSTE	3304.26	0.5751	0.8759
DDI-DB171	199,052	171	1178	GSTE	31,390.5	0.5292	0.9172

TABLE 3 | Evaluation results of different methods.

Datasets	Metrics	Methods						
		SVM	DDIMDL	MUFFIN	MDF-SA-DDI	MDDI-SCL	TFMD	USAM
DDI-DB110	Acc	0.336	0.5513	0.8661	0.8415	0.9201	0.9439	**0.9536**
	Pre	0.8042	0.5418	0.7246	0.8512	0.8696	0.9287	**0.9308**
	Rec	0.0213	0.2463	0.5674	0.7124	0.8784	0.8919	**0.9106**
	F1	0.0955	0.3031	0.6109	0.7627	0.8690	0.9028	**0.9203**
	AUC	0.8080	0.9801	0.9945	0.9771	**0.9976**	0.9970	0.9973
	AUPR	0.0835	0.5239	0.7987	0.8246	0.9199	0.9366	**0.9428**
DDI-DB171	Acc	0.3463	0.5606	0.8807	0.8903	0.5843	0.9410	**0.9538**
	Pre	0.8708	0.3708	0.5548	0.9115	0.7123	0.9205	**0.9298**
	Rec	0.0202	0.1759	0.5307	0.5663	0.6206	0.7806	**0.7968**
	F1	0.0476	0.2174	0.4833	0.6195	0.5860	0.7995	**0.8580**
	AUC	0.7381	0.9664	0.9844	0.9913	0.9446	0.9959	**0.9960**
	AUPR	0.0938	0.4599	0.6617	0.8208	0.6665	0.8941	**0.9032**
DDIMDL	Acc	0.3761	0.8173	0.8258	0.9109	0.9019	0.9190	**0.9301**
	Pre	0.8496	0.7935	0.6111	0.8812	0.8333	0.8933	**0.9062**
	Rec	0.0305	0.5919	0.5310	0.8460	0.8216	0.8659	**0.8768**
	F1	0.0912	0.6596	0.5498	0.8263	0.8205	0.8486	**0.8913**
	AUC	0.7921	0.9864	0.9954	0.9756	**0.9974**	0.9906	0.9932
	AUPR	0.0699	0.7511	0.7305	0.8967	0.8831	0.9013	**0.9078**
MUFFIN	Acc	0.2808	0.8583	0.9360	0.9408	0.9553	0.9728	**0.9826**
	Pre	0.8146	0.8595	0.8970	0.9506	0.9414	0.9648	**0.9742**
	Rec	0.0177	0.7305	0.8035	0.9071	0.9411	0.9455	**0.9498**
	F1	0.0907	0.7384	0.8351	0.9353	0.9391	0.9458	**0.9618**
	AUC	0.6921	0.9969	**0.9989**	0.9982	0.9976	0.9949	0.9952
	AUPR	0.0485	0.8729	0.9296	0.9584	0.9651	0.9666	**0.9702**

Note: The bold values indicate the best performance results.

exceeding those of other methods. On the more challenging DDI-DB171, USAM loss still maintains the highest accuracy (0.9526) and F1 score (0.8573), demonstrating its stable generalisation ability in complex and imbalanced data environments. In contrast, methods such as WCE, although showing strong performance in terms of recall, often sacrifice precision, leading to an overall performance decline. USAM loss, however, achieves a better balance between precision and recall while remaining competitive on metrics such as AUPR. These results indicate that USAM loss effectively alleviates the class imbalance problem, enhances the model's ability to distinguish samples that are difficult to classify and thus significantly outperforms traditional loss functions in overall performance.

5.4.2 | *Ablation Studies and Baseline Comparison*

This section conducts a comprehensive analysis of the impact of each component of the USAM loss on the head, middle and tail classes, as well as on the overall Macro-F1 score in the DDI prediction task. Specifically, the loss function in Equation (10) can be decomposed into three parts: the CE loss term $\mathcal{L}_{\mathrm{CE}}$, the uncertainty-aware term $\mathcal{L}_{\mathrm{un}}$ and the class-balanced loss term $\mathcal{L}_{\mathrm{CB}}$.

$$\mathcal{L}_u(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left(\underbrace{\log p_{i,y_i}}_{\mathcal{L}_{\mathrm{CE}}} + \underbrace{H_i \log p_{i,y_i}}_{\mathcal{L}_{\mathrm{un}}} + \underbrace{\frac{1-\beta}{1-\beta^{n_{y_i}}}\log p_{i,y_i}}_{\mathcal{L}_{\mathrm{CB}}}\right). \quad (16)$$

Therefore, the total loss in Equation (13) can be expressed as the sum of the above three loss terms and the SAM regularisation term $\mathcal{L}_{\mathrm{sam}}(\theta)$.

As shown in Table 5, significant differences in model performance across different class distributions are observed when USAM loss components are selectively removed or retained. First, removing the CE loss term causes a substantial drop in the overall Macro-F1 score, indicating that CE remains critical for maintaining the model's fundamental discriminative ability and learning main class boundaries.

Second, adding the uncertainty-aware term ($\mathcal{L}_{\mathrm{un}}$) to the CE baseline significantly improves the tail class performance. As observed in Table 5, using only $\mathcal{L}_{\mathrm{un}}$ achieves an F1 score of 0.8351 on DDI-DB110, which is lower than the Focal Loss (FL) baseline ($F1 = 0.8990$ in Table 4). Nevertheless, we specifically

TABLE 4 | Results of different loss functions.

Datasets	Metrics	Loss									
		CE	WCE	FL	CB	BS	LDAM	TFL	FL + SAM	LDAM + SAM	TFL + SAM	USAM
DDI-DB110	Acc	0.9473	0.8762	0.9436	0.9205	0.9466	0.9456	0.9439	0.9461	0.9482	0.9495	**0.9536**
	Pre	0.9202	0.7957	0.9372	0.8669	0.8480	0.9422	0.9287	0.9381	**0.9451**	0.9385	0.9338
	Rec	0.8835	**0.9318**	0.8841	0.9181	0.9255	0.8734	0.8919	0.8905	0.8821	0.9214	0.9108
	F1	0.8998	0.8458	0.8990	0.8838	0.8760	0.8901	0.9028	0.9052	0.8955	0.9148	**0.9222**
	AUC	0.9970	0.9962	0.9968	0.9961	0.9969	0.9918	0.9970	0.9970	0.9935	0.9971	**0.9972**
	AUPR	0.9343	0.9349	0.9332	**0.9406**	0.9342	0.8842	0.9366	0.9354	0.8912	0.9391	0.9401
DDI-DB171	Acc	0.9353	0.6604	0.9311	0.3535	0.9316	0.9396	0.9410	0.9358	0.9432	0.9478	**0.9526**
	Pre	0.9150	0.7061	0.9214	0.4467	0.6249	0.8996	0.9205	0.9225	0.9054	**0.9315**	0.9286
	Rec	0.4752	**0.8574**	0.7028	0.7302	0.8143	0.7156	0.7806	0.7254	0.7365	0.8215	0.7962
	F1	0.4886	0.7534	0.7334	0.4448	0.7693	0.7434	0.7995	0.7512	0.7618	0.8265	**0.8573**
	AUC	0.9798	0.9818	0.9901	0.9411	0.9882	0.9719	0.9959	0.9925	0.9785	0.9964	**0.9967**
	AUPR	0.7215	0.8367	0.8458	0.7058	0.8215	0.6724	0.8941	0.8562	0.6985	**0.8975**	0.8958

Note: The bold values indicate the best performance results.

adopt prediction entropy as the uncertainty measure. The primary reason is that FL mainly relies on the probability of the target class, whereas entropy captures the uncertainty of the entire prediction distribution. In a challenging DDI task with numerous relations (e.g., 110 classes), the probability distribution over nontarget classes provides useful information regarding model confusion, which helps identify difficult samples more comprehensively.

Third, integrating the class balancing loss term ($\mathcal{L}_{CB}$) further elevates tail class performance. As shown in Table 4, although LDAM loss alone ($F1 = 0.8901$) performs slightly better than CB loss alone ($F1 = 0.8838$), our framework selects CB loss. CB loss operates as a class-dependent weight directly in the loss function and does not modify the model outputs. In contrast, LDAM loss alters the logit distribution by introducing class-dependent margins. Because USAM loss already includes a dynamic uncertainty weighting mechanism, utilising CB loss provides a more consistent way to handle class imbalance while keeping the overall optimisation framework simple.

Finally, introducing the SAM regularisation term achieves the optimal overall performance. To investigate whether SAM works well with existing long-tail learning methods, we evaluated FL + SAM, LDAM + SAM and TFL + SAM (Table 4). The results demonstrate that SAM consistently improves the performance of these methods (e.g., TFL + SAM achieves a high Recall of 0.9214). However, USAM loss still obtains the best overall performance (Accuracy = 0.9536, $F1 = 0.9222$, AUC = 0.9972 on DDI-DB110) and improves the tail-class F1 score to 0.8908 (Table 5). We believe this advantage is related to the different optimisation behaviours. LDAM loss modifies the decision boundary through class-dependent margins and TFL uses nonlinear reweighting. In contrast, USAM loss uses a simple additive loss design (CE + entropy + CB) without explicitly changing the logit space. This formulation synergises more effectively with SAM, allowing it to search for flat minima with greater flexibility and thereby leading to superior generalisation performance.

5.4.3 | *Synergistic Mechanism of USAM Loss*

The decoupled ablation results in Table 5 indicate that class-balanced SAM and uncertainty weighting address different aspects of the learning process. Class-balanced SAM primarily alleviates macro-level class imbalance, whereas uncertainty weighting emphasises difficult and ambiguous samples at the micro level. Their combination enables the model to simultaneously balance class distributions and focus on confusing boundary samples. Meanwhile, assigning larger weights to hard samples may increase the model's sensitivity to noisy or ambiguous instances. By optimising the worst-case loss within a local parameter neighbourhood, SAM acts as a regulariser that reduces this sensitivity and encourages convergence towards flatter minima. As a result, the proposed USAM loss improves generalisation while effectively exploiting informative hard samples.

5.4.4 | *Discussion on the Head–Tail Tradeoff*

In conventional long-tailed learning, improving tail-class performance often comes at the expense of head-class performance.

TABLE 5 | Results of different classes (Macro-F1).

				DDI-DB110				DDI-DB171			
$\mathcal{L}_{CE}$	$\mathcal{L}_{un}$	$\mathcal{L}_{CB}$	$\mathcal{L}_{sam}(\theta)$	**Head**	**Med**	**Tail**	**All**	**Head**	**Med**	**Tail**	**All**
✓	×	×	×	0.9452	0.9154	0.8676	0.8998	0.8202	0.6800	0.3557	0.4886
×	✓	×	×	0.8805	0.8512	0.8025	0.8351	0.7801	0.6402	0.3528	0.4725
×	×	✓	×	0.8601	0.8402	0.8339	0.8412	0.7205	0.6300	0.3820	0.4816
✓	✓	×	×	0.9460	0.9172	0.8714	0.9023	0.9251	0.8800	0.7934	0.8298
✓	✓	✓	×	0.9492	0.9258	0.8816	0.9106	0.9304	0.8912	0.7985	0.8364
×	×	✓	✓	0.8722	0.8510	0.8451	0.8524	0.7450	0.6601	0.3986	0.5023
✓	✓	✓	✓	0.9555	0.9380	0.8908	0.9203	0.9482	0.9150	0.8201	0.8580

This expected head–tail tradeoff can indeed be observed in our individual ablation studies. For example, as shown in Table 5, when using only the class-balanced loss ($\mathcal{L}_{CB}$) on the DDI-DB110 dataset, the model improves the tail-class F1 score but suffers a substantial reduction in the head-class F1 score, decreasing from 0.9452 (the CE baseline) to 0.8601. Similar trends are observed when other components are applied independently.

Interestingly, this tradeoff becomes much less pronounced within the USAM loss. One possible explanation is that the standard cross-entropy term ($\mathcal{L}_{CE}$) is retained in the final objective, helping preserve the discriminative ability on head classes, whereas the entropy and CB terms provide additional support for uncertain and minority-class samples. In addition, the SAM optimisation strategy may further improve generalisation across different class groups by seeking flatter minima in the optimisation landscape. Together, these factors may help explain why USAM loss achieves balanced improvements on head, medium and tail classes. As a result, the proposed loss improves tail-class performance without substantial degradation on head classes.

5.4.5 | *Parameter Sensitivity Analysis*

We analyse the effect of the SAM regularisation coefficient α on the performance of the DDI prediction model. The experiments set the search range for α as $\{0.1, 0.25, 0.5, 0.75, 1.0\}$ and evaluate the performance of the USAM loss on the DDI-DB171 dataset using the F1 score. As shown in Figure 5, the model achieves the best performance when α is set to 0.25.

When α is too small (e.g., 0.1), the sharpness-aware regularisation becomes insufficient. In the long-tailed DDI-DB171 dataset, many tail classes contain only a few samples. Without a sufficiently large α to smooth the loss landscape, the model tends to memorise these rare samples, leading to overfitting and convergence to sharp local minima. During testing, even minor feature deviations may cause the model to misclassify these tail instances, resulting in a significant drop in the F1 score.

In contrast, when α becomes too large (e.g., $\alpha \geq 0.5$), the performance gradually declines. A large α imposes excessively strong smoothness constraints during optimisation, which may oversmooth the decision boundaries and reduce the model's ability to capture discriminative features. This can lead to underfitting, particularly for head classes with more complex feature distributions.

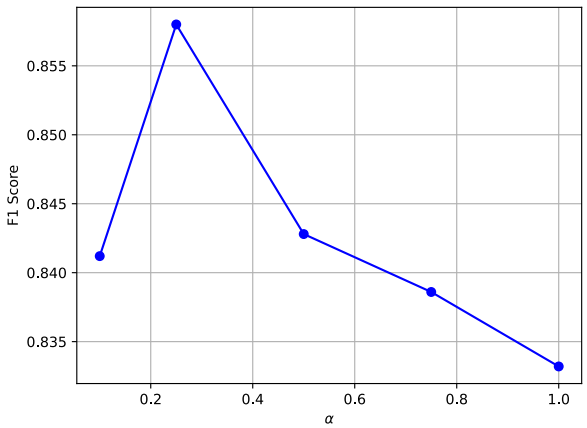


FIGURE 5 | F1 scores of different α values in the DDI-DB171 dataset.

6 | Conclusion

To address performance degradation in DDI prediction caused by long-tailed data, this work proposes the USAM loss function. USAM loss tackles this by dynamically adjusting training focus: it measures the uncertainty (entropy) of each sample's prediction to identify instances that are difficult to classify and assigns them higher weights within the CE loss. Crucially, USAM loss also incorporates a sharpness-aware minimisation strategy specifically adapted for class imbalance. This combination guides the model towards flatter, more robust regions in the loss landscape, significantly improving generalisation, particularly for the challenging tail classes. Experimental results confirm USAM loss effectiveness in enhancing DDI prediction model performance on long-tailed distributions.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grants 62376161 and U24A20322) and the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Reference Number: STG5/E-103/24-R).

Funding

This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (Grant STG5/E-103/24-R) and National Natural Science Foundation of China (Grant 62376161, U24A20322).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data will be made available on request.

References

1. Z. Du, L. Yang, H. He, X. Wu, X. Qi, and Y. Zhang, "Artificial Intelligence for the Noninvasive Diagnosis of Clinically Significant Portal Hypertension," *EngMedicine* 2, no. 2 (2025): 100069, https://doi.org/10.1016/j.engmed.2025.100069.

2. Y. Xia, A. Xiong, Z. Zhang, Q. Zou, and F. Cui, "A Comprehensive Review of Deep Learning-Based Approaches for Drug-Drug Interaction Prediction," *Briefings in Functional Genomics* (2025).

3. S. Chen, T. Li, L. Yang, et al., "Artificial Intelligence-Driven Prediction of Multiple Drug Interactions," *Briefings in Bioinformatics* 23, no. 6 (2022): bbac427, https://doi.org/10.1093/bib/bbac427.

4. L. Yue, D. Tian, W. Chen, X. Han, and M. Yin, "Deep Learning for Heterogeneous Medical Data Analysis," *World Wide Web* 23, no. 5 (2020): 2715–2737, https://doi.org/10.1007/s11280-019-00764-z.

5. L. N. Zheng, C. G. Dong, W. E. Zhang, X. Chen, L. Yue, and W. Chen, "Devil in the Tail: A Multi-Modal Framework for Drug-Drug Interaction Prediction in Long Tail Distinction," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, (Association for Computing Machinery, 2024), 3395–3404.

6. S. Lin, W. Chen, G. Chen, S. Zhou, D. Wei, and Y. Xiong, "MDDI-SCL: Predicting Multi-Type Drug-Drug Interactions via Supervised Contrastive Learning," *Journal of Cheminformatics* 14, no. 1 (2022): 81, https://doi.org/10.1186/s13321-022-00659-8.

7. S. Lin, Y. Wang, L. Zhang, et al., "MDF-SA-DDI: Predicting Drug-Drug Interaction Events Based on Multi-Source Drug Fusion, Multi-Source Feature Fusion and Transformer Self-Attention Mechanism," *Briefings in Bioinformatics* 23, no. 1 (2022): bbab421, https://doi.org/10.1093/bib/bbab421.

8. Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 9 (2023): 10795–10816, https://doi.org/10.1109/tpami.2023.3268118.

9. D. Wood, T. Papamarkou, M. Benatan, and R. Allmendinger, "Model-Agnostic Variable Importance for Predictive Uncertainty: An Entropy-Based Approach," *Data Mining and Knowledge Discovery* 38, no. 6 (2024): 4184–4216, https://doi.org/10.1007/s10618-024-01070-7.

10. Y. Cui, M. Jia, T. Lin, Y. Song, and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (IEEE, 2019), 9268–9277.

11. T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision*, (IEEE, 2017), 2980–2988.

12. P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-Aware Minimization for Efficiently Improving Generalisation," in *Proceedings of the International Conference on Learning Representations*, (OpenReview.net, 2021).

13. Y. Zhou, Y. Qu, X. Xu, and H. Shen, "ImbSAM: A Closer Look at Sharpness-Aware Minimization in Class-Imbalanced Recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, 2023): 11345–11355.

14. Y. Deng, X. Xu, Y. Qiu, J. Xia, W. Zhang, and S. Liu, "A Multi-modal Deep Learning Framework for Predicting Drug-Drug Interaction Events," *Bioinformatics* 36, no. 15 (2020): 4316–4322, https://doi.org/10.1093/bioinformatics/btaa501.

15. A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is all You Need," in *Proceedings of the Advances in Neural Information Processing Systems*, (Curran Associates, Inc., 2017).

16. J. Zhu, C. Che, H. Jiang, J. Xu, J. Yin, and Z. Zhong, "SSF-DDI: A Deep Learning Method Utilising Drug Sequence and Substructure Features for Drug-Drug Interaction Prediction," *BMC Bioinformatics* 25, no. 1 (2024): 39, https://doi.org/10.1186/s12859-024-05654-4.

17. D. Niu, L. Xu, S. Pan, L. Xia, and Z. Li, "SRR-DDI: A Drug-Drug Interaction Prediction Model With Substructure Refined Representation Learning Based on Self-Attention Mechanism," *Knowledge-Based Systems* 285 (2024): 111337, https://doi.org/10.1016/j.knosys.2023.111337.

18. X. Lin, Z. Quan, Z. Wang, T. Ma, and X. Zeng, "KGNN: Knowledge Graph Neural Network for Drug-Drug Interaction Prediction," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, (2020), 2739–2745.

19. Y. Hong, P. Luo, S. Jin, and X. Liu, "LaGAT: Link-Aware Graph Attention Network for Drug-Drug Interaction Prediction," *Bioinformatics* 38, no. 24 (2022): 5406–5412, https://doi.org/10.1093/bioinformatics/btac682.

20. Z. Li, S. Zhu, B. Shao, X. Zeng, T. Wang, and T. Liu, "DSN-DDI: An Accurate and Generalised Framework for Drug-Drug Interaction Prediction by Dual-View Representation Learning," *Briefings in Bioinformatics* 24, no. 1 (2023): bbac597, https://doi.org/10.1093/bib/bbac597.

21. T. Lyu, J. Gao, L. Tian, Z. Li, P. Zhang, and J. Zhang, "MDNN: A Multi-modal Deep Neural Network for Predicting Drug-Drug Interaction Events," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, (2021), 3536–3542.

22. Y. Chen, T. Ma, X. Yang, J. Wang, B. Song, and X. Zeng, "MUFFIN: Multi-Scale Feature Fusion for Drug-Drug Interaction Prediction," *Bioinformatics* 37, no. 17 (2021): 2651–2658, https://doi.org/10.1093/bioinformatics/btab169.

23. M. Jiang, G. Liu, Y. Su, W. Jin, and B. Zhao, "Hierarchical Multi-Relational Graph Representation Learning for Large-Scale Prediction of Drug-Drug Interactions," *IEEE Transactions on Big Data* 11, no. 3 (2025): 961–975, https://doi.org/10.1109/tbdata.2025.3536924.

24. M. Jiang, G. Liu, B. Zhao, Y. Su, and W. Jin, "Relation-Aware Graph Structure Embedding With Co-Contrastive Learning for Drug-Drug Interaction Prediction," *Neurocomputing* 572 (2024): 127203, https://doi.org/10.1016/j.neucom.2023.127203.

25. Y. Zhong, G. Li, J. Yang, et al., "Learning Motif-Based Graphs for Drug-Drug Interaction Prediction via Local-Global Self-Attention," *Nature Machine Intelligence* 6, no. 9 (2024): 1094–1105, https://doi.org/10.1038/s42256-024-00888-6.

26. Y. Wang, Z. Yang, and Q. Yao, "Accurate and Interpretable Drug-Drug Interaction Prediction Enabled by Knowledge Subgraph Learning," *Communication and Medicine* 4, no. 1 (2024): 59, https://doi.org/10.1038/s43856-024-00486-y.

27. Z. Jiang, P. Ding, C. Shen, and X. Dai, "Geometric Molecular Graph Representation Learning Model for Drug-Drug Interactions Prediction," *IEEE Journal of Biomedical and Health Informatics* 28, no. 12 (2024): 7623–7632, https://doi.org/10.1109/jbhi.2024.3453956.

28. Y. Chen, J. Wang, Q. Zou, et al., "DrugDAGT: A Dual-Attention Graph Transformer With Contrastive Learning Improves Drug-Drug Interaction Prediction," *BMC Biology* 22, no. 1 (2024): 233, https://doi.org/10.1186/s12915-024-02030-9.

29. H. Du, Q. Yao, J. Zhang, Y. Liu, and Z. Wang, "Customised Subgraph Selection and Encoding for Drug-Drug Interaction Prediction," in

Proceedings of the Advances in Neural Information Processing Systems, (Curran Associates, Inc., 2024), 109582–109608.

30. Y. Gan, W. Liu, G. Xu, C. Yan, and G. Zou, "DMFDDI: Deep Multimodal Fusion for Drug-Drug Interaction Prediction," *Briefings in Bioinformatics* 24, no. 6 (2023): bbad397, https://doi.org/10.1093/bib/bbad397.

31. L. Meng, Y. He, C. Sun, L. Huang, T. Hu, and F. Yang, "Learning Personalised Drug Features and Differentiated Drug-Pair Interaction Information for Drug-Drug Interaction Prediction," *Neural Networks* 181 (2025): 106828, https://doi.org/10.1016/j.neunet.2024.106828.

32. J. Ren, C. Yu, S. Sheng, et al., "Balanced Meta-Softmax for Long-Tailed Visual Recognition," in *Proceedings of the Advances in Neural Information Processing Systems*, (Curran Associates, Inc., 2020), 4175–4186.

33. K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning Imbalanced Datasets With Label-Distribution-Aware Margin Loss," in *Proceedings of the Advances in Neural Information Processing Systems*, (Curran Associates, Inc., 2019).

34. M. Li, Y. Cheung, and Y. Lu, "Long-Tailed Visual Recognition via Gaussian Clouded Logit Adjustment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (IEEE, 2022), 6929–6938.

35. J. Chen and B. Su, "Transfer Knowledge From Head to Tail: Uncertainty Calibration Under Long-Tailed Distribution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (IEEE, 2023), 19978–19987.

36. M. Li, Z. Hu, Y. Lu, W. Lan, Y. Cheung, and H. Huang, "Feature Fusion From Head to Tail for Long-Tailed Visual Recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, (AAAI, 2024), 13581–13589, https://doi.org/10.1609/aaai.v38i12.29262.

37. Y. Jin, M. Li, Y. Lu, Y. Cheung, and H. Wang, "Long-Tailed Visual Recognition via Self-Heterogeneous Integration With Knowledge Excavation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (IEEE, 2023), 23695–23704.

38. S. Park, Y. Hong, B. Heo, S. Yun, and J. Choi, "The Majority Can Help the Minority: Context-Rich Minority Oversampling for Long-Tailed Classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (IEEE, 2022), 6887–6896.

39. J. Gao, H. Zhao, Z. Li, and D. Guo, "Enhancing Minority Classes by Mixing: An Adaptative Optimal Transport Approach for Long-Tailed Classification," in *Proceedings of the Advances in Neural Information Processing Systems*, (Curran Associates, Inc., 2023), 60329–60348.

40. J. Shi, T. Wei, Z. Zhou, J. Shao, X. Han, and Y. Li, "Long-Tail Learning With Foundation Model: Heavy Fine-Tuning Hurts," in *Proceedings of the 41st International Conference on Machine Learning*, (PMLR, 2024), 45014–45039.

41. M. Li, Y. Liu, Y. Lu, Y. Zhang, Y. Cheung, and H. Huang, "Improving Visual Prompt Tuning by Gaussian Neighborhood Minimization for Long-Tailed Visual Recognition," in *Proceedings of the Advances in Neural Information Processing Systems*, (Curran Associates, Inc., 2024), 103985–104009.

42. B. Dong, P. Zhou, S. Yan, and W. Zuo, "LPT: Long-Tailed Prompt Tuning for Image Classification," in *Proceedings of the the Eleventh International Conference on Learning Representations*, (OpenReview.net, 2023).

43. Z. Zhou, L. Li, P. Zhao, P. Heng, and W. Gong, "Class-Conditional Sharpness-Aware Minimization for Deep Long-Tailed Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (IEEE, 2023), 3499–3509.

44. S. Li, Q. Xu, Z. Yang, et al., "Focal-SAM: Focal Sharpness-Aware Minimization for Long-Tailed Classification," in *Proceedings of the Forty-second International Conference on Machine Learning*, (PMLR, 2025).

45. X. Lyu, Q. Xu, Z. Yang, S. Lyu, and Q. Huang, "SSE-SAM: Balancing Head and Tail Classes Gradually Through Stage-Wise Sam," in *Proceedings of the AAAI Conference on Artificial Intelligence*, (AAAI, 2025), 19278–19286, https://doi.org/10.1609/aaai.v39i18.34122.

46. X. Wang, X. Liu, P. Huang, P. Huang, S. Hu, and H. Zhu, "U-MedSAM: Uncertainty-Aware Medsam for Medical Image Segmentation," in *Proceedings of the Medical Image Segmentation Challenge* (Springe, 2024), 206–217.

47. Y. Ma, J. Yang, and R. Yan, "Sharpness-Aware Gradient Alignment for Domain Generalization With Noisy Labels in Intelligent Fault Diagnosis," *IEEE Transactions on Instrumentation and Measurement* 73 (2024): 10–19, https://doi.org/10.1109/tim.2024.3413128.