Masood, Muhammad Arslan; Kaski, Samuel; Ceulemans, Hugo; Herman, Dorota; Heinonen, Markus

Balancing Imbalanced Toxicity Models : Using MolBERT with Focal Loss

Published in:
AI in Drug Discovery - 1st International Workshop, AIDD 2024, Held in Conjunction with ICANN 2024, Proceedings

DOI:
10.1007/978-3-031-72381-0_8

Published: 01/01/2025

Document Version
Publisher's PDF, also known as Version of record

Balancing Imbalanced Toxicity Models: Using MolBERT with Focal Loss

Muhammad Arslan Masood[1,2(✉)], Samuel Kaski[2,3], Hugo Ceulemans[1], Dorota Herman[1], and Markus Heinonen[2]

[1] Drug Discovery Data Sciences, Janssen Pharmaceutica NV, Turnhoutseweg 30, 2340 Beerse, Belgium
[2] Department of Computer Science, Aalto University, Espoo, Finland
arslan.masood@aalto.fi
[3] Department of Computer Science, University of Manchester, Manchester, UK
https://www.aalto.fi/en/department-of-computer-science

Abstract. Drug-induced liver injury (DILI) presents a multifaceted challenge, influenced by interconnected biological mechanisms. Current DILI datasets are characterized by small sizes and high imbalance, posing difficulties in learning robust representations and accurate modeling. To address these challenges, we trained a multi-modal multi-task model integrating preclinical histopathologies, biochemistry (blood markers), and clinical DILI-related adverse drug reactions (ADRs). Leveraging pre-trained BERT models, we extracted representations covering a broad chemical space, facilitating robust learning in both frozen and fine-tuned settings. To address imbalanced data, we explored weighted Binary Cross-Entropy (w-BCE) and weighted Focal Loss (w-FL) . Our results demonstrate that the frozen BERT model consistently enhances performance across all metrics and modalities with weighted loss functions compared to their non-weighted counterparts. However, the efficacy of fine-tuning BERT varies across modalities, yielding inconclusive results. In summary, the incorporation of BERT features with weighted loss functions demonstrates advantages, while the efficacy of fine-tuning remains uncertain.

Keywords: Toxicity · DILI · BERT · Focal loss

1 Introduction and Background

Thalidomide, the tragedy of birth defects led the foundation of systematic testing of drugs safety prior to marketing (Kim and Scialli, 2011). Pharmacovigilance efforts start with in-vitro and in-vivo studies during the drug development stage, continue through clinical trial and post-marketing surveillance.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-72381-0_8.

D.-A. Clevert et al. (Eds.): AIDD 2024, LNCS 14894, pp. 82–97, 2025.
https://doi.org/10.1007/978-3-031-72381-0_8

The liver, as the primary organ affected by xenobiotics, plays a crucial role in drug metabolism (Stanley, 2017). Drug-induced liver injury (DILI) stands as a significant cause of late-stage drug failure and post-marketing drug withdrawal (Watkins, 2011; Parasrampuria et al., 2018). Toxic compounds can be categorized into intrinsic toxins, whose toxicity can be modeled based on chemical information, and idiosyncratic toxins, which pose challenges in both preclinical and clinical modeling due to their unpredictable effects influenced by genetic variations (Lancaster et al., 2015; Parasrampuria et al., 2018). Over the years, several methods have been developed to model DILI using molecular structure and various fingerprints (Cruz-Monteagudo et al., 2008; Chen et al., 2013b; Xu et al., 2015; Ai et al., 2018; Wang et al., 2019; Asilar et al., 2020). Combining other modalities with molecular features, such as transcriptomics (Wang et al., 2019a), physicochemical properties (Ekins et al., 2010; Chen et al., 2013a), and selected in-vitro assays (Williams et al., 2020), has been shown to provide robust DILI models. During the drug design process, toxicity assessment spans multiple stages, encompassing in-vitro assays, preclinical investigations, and clinical trials. Toxicity presents across diverse endpoints and species, thus prompting a multitask approach for data integration and cross-modality learning. This strategy has demonstrated promise in extracting toxicity patterns by jointly considering various dose administration methods, endpoints, and species, particularly in acute toxicity modeling (Sosnin et al., 2019; Jain et al., 2021). Moreover, extending this approach to incorporate joint learning from in-vitro, in-vivo, and clinical data has improved balanced accuracy (as defined in Eq. 7) of the ClinTox dataset. (Sharma et al., 2023).

Class imbalance is a prevalent issue in toxicity datasets, where negative instances vastly outnumber positive ones. This disparity makes machine learning models inaccurate, as classifiers trained on imbalanced data tend to prioritize the majority class, leading to ineffective performance on the minority class (Rawat and Mishra, 2022). To address this, various strategies are employed, including resampling techniques like oversampling and undersampling. Oversampling methods such as Synthetic Minority Oversampling Technique (SMOTE) artificially increase the number of minority instances (Chawla et al., 2002), while undersampling involves reducing the number of majority instances (Laveti et al., 2021; Lee and Seo, 2022). However, both approaches have drawbacks; undersampling may lead to loss of valuable data, while oversampling can be computationally intensive (Rawat and Mishra, 2022). Cost-Sensitive Learning (CSL) can also be used as this method assigns higher costs to samples from the minority class (Elkan, 2001; López et al., 2012). Unlike resampling techniques, CSL maintains the original data distribution while enhancing computational efficiency. CSL, coupled with traditional machine learning algorithms such as Random Forest (RF) and Support Vector Machine (SVM), has been used for drug discovery application, including compound activity estimation (Alashwal and Lucman, 2020), CYP450 modeling (Eitrich et al., 2007), and Drug-Induced Liver Injury (DILI) modeling (Moein et al., 2023), demonstrating improvements in some cases. In the realm of deep learning, binary-cross-entropy loss serves as a

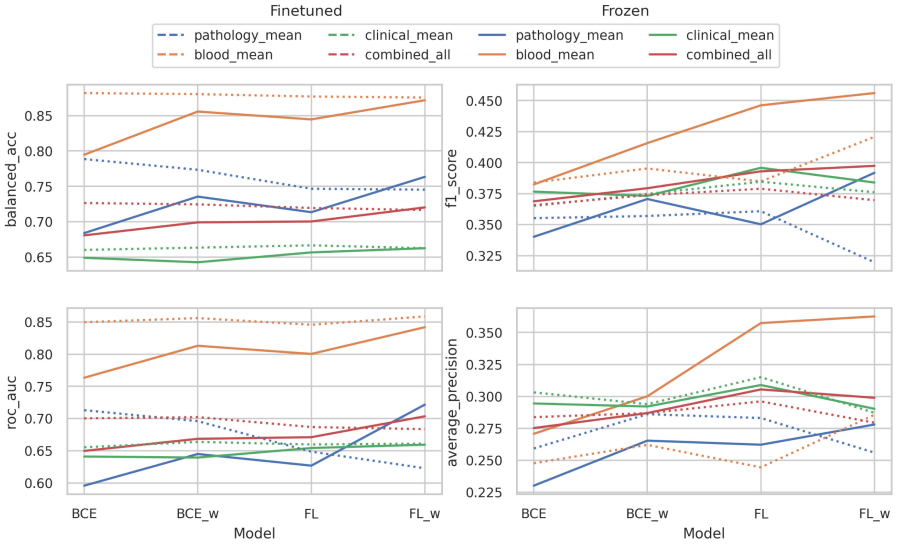

Fig. 1. Mean performance across all tasks was evaluated using multiple metrics. Solid lines represent Frozen-BERT, while dotted lines indicate fine-tuned BERT. Performance improved from BCE to weighted-BCE, Focal loss, and weighted Focal loss for the Frozen-BERT model. However, this trend is inconsistent for fine-tuned BERT. In certain tasks, such as pathological ones, the fine-tuned model outperformed Frozen-BERT. In other cases, negative transfer is observed.

common choice for training binary classification models, often augmented with a weighting factor to elevate the cost of positive instances, thus ensuring a balanced contribution from both classes in the overall loss function. Focal loss represents a refinement of BCE loss, introducing a modulating factor that aids in distinguishing between easy and difficult examples naturally favours minority class Lin et al. (2018).

One method for representing three-dimensional chemical structures as text strings is the Simplified Molecular Input Line Entry System (SMILES), which employs a defined set of ordered rules and specific syntax Weininger (1988). The chemical characteristics of a compound $\mathbf{x}_c$ can be described through various modalities. Encoding schemes like Mold, PaDel, RDF, ECFC, and Marvin molecular descriptors have been developed to capture molecular structural properties. Despite their individual successes, there's no universal encoding scheme or algorithm to *rule them all* (Gao et al., 2020). In drug development, small-scale datasets often fail to adequately represent the vast chemical space, leading to models trained on handcrafted features that struggle to generalize to unseen chemical spaces (Moein et al., 2023). To address this limitation, researchers leverage representations derived from large amounts of unlabeled data (Harnik and Milo, 2024). Various models such as Variational Autoencoders (VAEs) (Kingma and Welling, 2013), Normalizing Flows (NFs) (Rezende and Mohamed, 2015),

and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) aim to uncover low-dimensional latent representations $\phi(\mathbf{x}) \in \mathbb{R}^d$ of complex, high-dimensional objects $\mathbf{x} \in \mathbb{R}^D$, where $d \ll D$ (Ruthotto and Haber, 2021). These models transform data into a vectorized space, generating concise and well-structured representations that encompass broader chemical space (Li et al., 2022). To facilitate learning of underlying chemistry, various pretext tasks are carefully designed, including input translation between modalities (Winter et al., 2019; Yang et al., 2019), input reconstruction (Wang et al., 2019b; Li and Fourches, 2020; Maziarka et al., 2020), and recovering masked or corrupted input (Liu et al., 2023).

In recent years, various transformer-based models have been applied to molecular representation learning (Chithrananda et al., 2020; Li and Jiang, 2021; Ahmad et al., 2022; Irwin et al., 2022) with many studies opted for transformer based BERT architecture (Li and Jiang, 2021; Liu et al., 2023; Shermukhamedov et al., 2023). BERT (Bidirectional Encoder Representations from Transformers) is pre-trained on large text corpora using two objectives, defined by Devlin et al. (2019) as the "masked language model" (MLM) and "next sentence prediction" (NSP) task. During pre-training, BERT learns bidirectional contextual embeddings for each token, capturing nuanced word meanings within the sentence context. Utilizing the Transformer architecture, BERT employs self-attention mechanisms to dynamically weigh word importance. By fine-tuning on task-specific labeled data, BERT adapts its learned representations to various downstream natural language processing tasks, achieving state-of-the-art performance. BERT can learn molecular representations by treating molecular structures as token sequences. Pre-training BERT on large molecular datasets with appropriate objectives, such as incorporating physicochemical properties or molecule relationships, enables it to learn robust chemical representations (Fabian et al., 2020). Fine-tuning the pre-trained BERT model on small task-specific labeled data can provide improved performance in some drug discovery applications (Liu et al., 2023).

2 Materials and Method

2.1 Datasets

The preclinical study integrates liver histopathology endpoints from the TG-Gates dataset (Igarashi et al., 2015), covering 170 compounds administered to rats across varying concentrations and exposure conditions, later expanded to 430 compounds with re-annotated `INHAND` labels (Moein et al., 2023). Out of 55 liver endpoints, we focus on 12 for this study. We extend preclinical tasks by incorporating selected blood markers (ALP, AST, ALT, GTP, TC, TG, TBIL, DBIL) from biochemistry database provided by TG-GATES converting both histopathological and bloodmarker labels into binary labels using expert-derived thresholds. Additionally, we enrich preclinical data with DILI related adverse drug reactions (ADRs) extracted from the SIDER dataset , comprising 6060 ADRs associated with 1430 drugs (Kuhn et al., 2016). Further details regarding

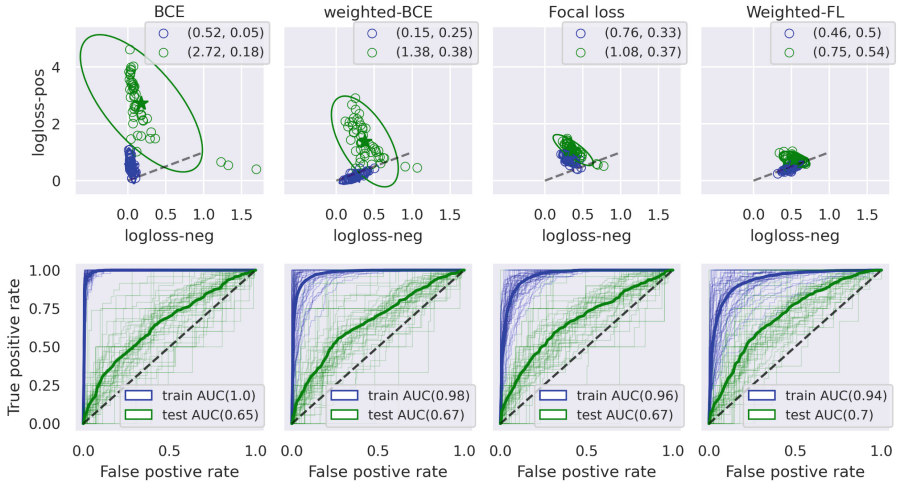

Fig. 2. Results from Frozen-BERT model. **Top row**: The log-loss analysis of positive and negative data points is depicted in this plot, where blue and green colors represent the training and testing data, respectively. Task-wise means are represented by small blobs, while ellipses indicate the 95% confidence interval. This visualization revealed that the network tends to be biased towards the majority class (negatives), leading to significantly lower log-loss for negatives, particularly evident with Binary Cross-Entropy (BCE). Transitioning to weighted Binary Cross-Entropy (BCE-W), the model is forced to equally prioritize both negatives and positives, resulting in a decrease in log-loss for positives. Focal Loss naturally emphasizes on hard examples, which, in this context, are positive examples. Weighted Focal Loss further supported the model by applying additional weighting to positive examples, as a result further reduced in logloss of positive instances. **Bottom row**: This plot presents the ROC-AUC for the train and test sets. Light lines represent task-wise ROC-AUC, while the thick line represents the mean ROC-AUC across all tasks. Weighted Focal Loss provided the highest validation ROC-AUC

task selection, binarization, and distributions are available in the supplementary material.

2.2 Loss Functions

We consider a modeling problem from molecules $\mathbf{x}$ to binary toxicity profiles $\mathbf{y} \in \{0,1\}^P$ of $P = 50$ endpoints from a dataset $D = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ of size $N \approx 2000$. We assume a function $f(\mathbf{x};\theta) \in [0,1]^P$ that outputs separate probabilities for endpoints, and we use a shorthand $f_{np} = f(\mathbf{x}_n;\theta)_p$.

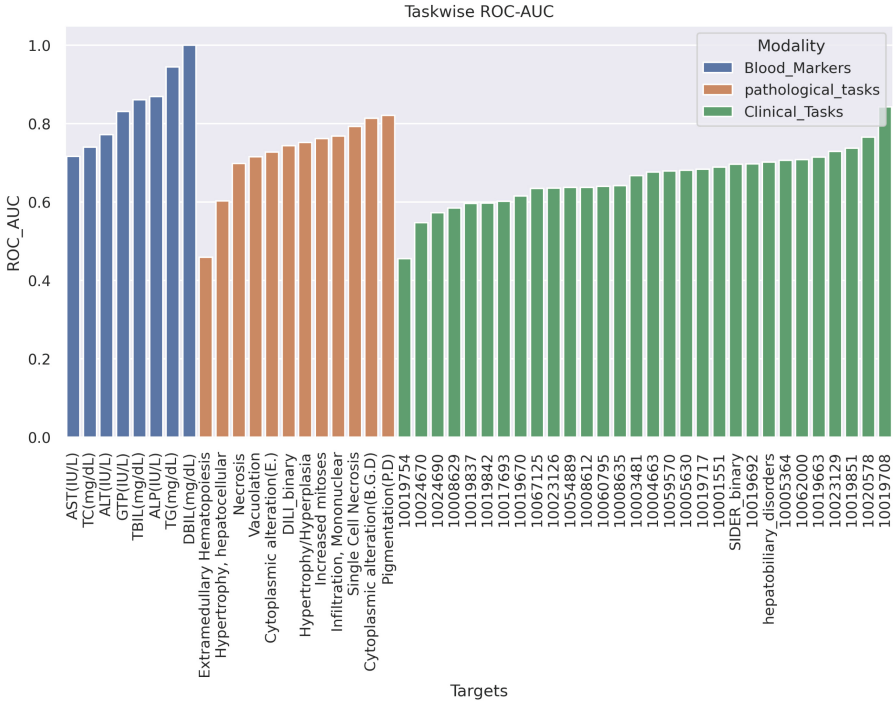

Fig. 3. This plot shows task-wise performance, with colors representing different modalities and tasks organized from lowest to highest ROC score. This plot also shows that clinical tasks were the most challenging to model. Additionally, the model failed to learn two tasks: Extramedullary (pathological) and 100197554 (clinical)

Binary-Cross-Entropy Loss (BCE). The binary cross-entropy (BCE) training loss is appropriate for this problem

$$\mathcal{L}_{\text{BCE}} = \sum_{p=1}^{P} \sum_{n=1}^{N} y_{np} \log \sigma(f_{np}) + (1 - y_{np}) \log(1 - \sigma(f_{np})) \tag{1}$$

Weighted-BCE. The toxicity datasets are generally zero-inflated with negatives being much more common, however, the BCE treats each observation as equally important, and will lead the model to focus more on negatives. We can tackle this positive-negative imbalance by overweighting the positive datapoints within each endpoint,

$$\mathcal{L}^{\text{w}}_{\text{BCE}} = \sum_{n=1}^{N} \sum_{p=1}^{P} w_p^+ y_{np} \log \sigma(f_{np}) + (1 - y_{np}) \log \left(1 - \sigma(f_{np})\right) \tag{2}$$

where $w_p^+ = N_{p-}/N_{p+} \in \mathbb{R}^+$ is the inverse ratio of positives N_{p+} to negatives N_{p-} in endpoint p. Here we only upscale the positives while leaving negatives

Table 1. The distribution of positive and negative samples across each modality. For visualization purpose, a molecule is classified as positive if it is active in any task, and negative if it is inactive in all tasks. The total represents the number of unique SMILES in the complete dataset.

	positive	negative	Modality Sum
Pathologies	176	234	410
Biochemistry (blood markers)	124	286	410
Clinical	749	470	1219
Total	1049	990	1554*

intact. This loss ensures that within-class positive and negative observations have equal mass. Further, we can try to find the balancing, by selecting optimal α through cross validation

$$w_p^+ = \alpha \frac{N_{p-}}{N_{p+}} + (1-\alpha)1 \tag{3}$$

where $\alpha \in [0, 1]$ denotes the positive balancing.

Focal Loss. In scenarios of significant class imbalance, mere weighting can be insufficient, as it fails to discriminate between easy and challenging examples, thereby risking the overwhelming of gradients by the dominant class. A remedy for this issue is focal loss, initially devised for object detection within images (Lin et al., 2018). This approach incorporates a modulating parameter alongside cross-entropy loss, thereby decrease the influence of accurately classified examples and consequently mitigating their overall impact. This modulating factor can be adopted and integrated into our binary cross-entropy loss framework.

$$\mathcal{L}_{\text{FL}} = \sum_{n=1}^{N} \sum_{p=1}^{P} (1-\sigma(f_{np}))^{\gamma}\, y_{np} \log \sigma(f_{np}) + \sigma(f_{np})^{\gamma} (1-y_{np}) \log\left(1-\sigma(f_{np})\right) \tag{4}$$

Weighted Focal Loss. Focal loss can also be assisted by incorporating positive weighting as described earlier.

$$\mathcal{L}_{\text{FL}}^{\text{w}} = \sum_{n=1}^{N} \sum_{p=1}^{P} w_p^+ (1-\sigma(f_{np}))^{\gamma}\, y_{np} \log \sigma(f_{np}) + \sigma(f_{np})^{\gamma} (1-y_{np}) \log\left(1-\sigma(f_{np})\right) \tag{5}$$

2.3 Models

Baseline. We are using Random Forest as our baseline. Random Forest is a robust baseline as it combines decision trees through ensemble learning, reducing overfitting and providing reliable results. Additionally, Random Forest maintains interpretability and scales efficiently for large datasets. To optimize performance, we conducted individual task-specific hyperparameter searches and presented the mean results across all tasks in Table 2. The hyperparameter search space details are provided in supplementary Table 2.

MolBERT. The MolBERT model Fabian et al. (2020), an adaptation of the BERT architecture Devlin et al. (2019), consists of 12 attention heads, 12 layers, and a 768-dimensional hidden layer, containing 85 million parameters. It is primarily optimized for the masked token estimation, employing cross-entropy loss. Additionally, it incorporates physicochemical properties computed via RDKit as an auxiliary task, with optimization achieved through mean squared error. The final loss function is determined by the arithmetic mean of all individual task losses. This model is pretrained for 100 epochs using the Adam optimizer.

MLP Head. This MLP head consists of an input-hidden-output layers, where $\mathbf{x}_0$ is initialized as the input features $\mathbf{x}$, which can be either BERT features or ECFP . We utilize dropout for regularization, batch normalization for training stability, and the rectified linear unit (ReLU) activation function as the default activation. Additionally, the network incorporates a skip connection, merging the input and output of the hidden layer, enhancing information flow. Finally, the output layer generates logits, which can be transformed into probabilities by passing through a sigmoidal activation function.

$$\begin{aligned}
\mathbf{x}_0 &= \mathbf{x} \quad \texttt{BERT features or ECFP} \\
\mathbf{x}_\ell &= \text{Dropout}(\text{ReLU}(\text{BatchNorm}(W_\ell \mathbf{x}_0 + \mathbf{b}_\ell))) \\
\tilde{\mathbf{x}}_{\ell+1} &= \text{BatchNorm}(W_{\ell+1}\mathbf{x}_\ell + \mathbf{b}_{\ell+1}) \\
\mathbf{x}_{\ell+1} &= \text{Dropout}(\text{ReLU}(\mathbf{x}_\ell + \tilde{\mathbf{x}}_{\ell+1})) \\
x_{out} &= W_{\ell+2}\mathbf{x}_{\ell+1} + \mathbf{b}_{\ell+1}
\end{aligned} \tag{6}$$

The hyper-parameters of this model are given in Table 1 in supplementary.

2.4 Feature Extraction

ECFP Fingerprints. ECFP or Extended-Connectivity Fingerprints (Rogers and Hahn, 2010), is a method used in cheminformatics to represent molecular structures as binary fingerprints, capturing structural information by encoding the presence or absence of substructural features within a specified radius around each atom. Through iterative traversal of the molecular structure, unique substructural fragments are identified and hashed into a fixed-length bit vector, generating a binary fingerprint where each bit indicates the presence or absence

of a specific substructural fragment. We encoded each molecule into fix 1024 dimensional binary vector by using radius 6. We have compared ECFP fingerprints with BERT features explained below.

BERT Features. We encoded preclinical and clinical SMILES into continuous features, utilizing a large transformer model MolBERT, pretrained on 1.6 million SMILES via masking, alongside physicochemical properties (Fabian et al., 2020). Extracting a pooled output of dimension 764 from the pretrained model, we employed these features to train an MLP head. This strategy allowed us to leverage a significant volume of unlabeled data,and encapsulated the contextual information of larger chemical space.

2.5 Evaluation

Here, we briefly sketch the evaluation metrics used in model selection and to report final results.

Balanced Accuracy. Given the imbalance between positive and negative instances, using accuracy as a performance metric becomes inadequate. Therefore, we chose balanced accuracy, which represents the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate). We compute the balanced accuracy at varying thresholds for each task and select the threshold ($\tau_p^{\max}$) that yields the highest balanced accuracy.

$$\begin{aligned} \mathrm{BA}(\tau_p) &= \frac{1}{2}\left(\mathrm{Sensitivity}(\tau_p) + \mathrm{Specificity}(\tau_p)\right) \\ \tau_p^{\max} &= \arg\max_{\tau_p} \mathrm{BA}(\tau_p) \\ \mathrm{BA} &= \frac{1}{P}\sum_{p=1}^{p} \mathrm{BA}(\tau_p^{\max})_p \end{aligned} \tag{7}$$

ROC AUC. The ROC curve, generated by plotting true positive rates (TPR) against false positive rates (FPR) at various thresholds(τ_p), illustrates the trade-off in model performance. The area under this curve (ROC AUC) condenses the curve's information into a single value, ranging between 0.5 (no discrimination) and 1.0 (ideal discrimination).

AUPR. The ROC-AUC curve can yield overly optimistic results with highly imbalanced datasets, thus we used Precision-Recall (PR) curves (Davis and Goadrich, 2006; Forman and Scholz, 2010). The Average Precision (AP) score provides a summary of a precision-recall curve by calculating the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight (Zhu, 2004):

$$\mathrm{AP} = \sum_n (R_n - R_{n-1}) P_n \tag{8}$$

where P_n and R_n denote the precision and recall at the n-th threshold, respectively. We selected the optimal hyperparmeters based on AP-score

F1-score. This metric combines the precision and recall using the harmonic mean. To select the optimal threshold, we followed the similar procedure to balance accuracy

$$\text{F1 score} = 2 \times \frac{\text{Precision}(\tau_p) \times \text{Recall}(\tau_p)}{\text{Precision}(\tau_p) + \text{Recall}(\tau_p)} \tag{9}$$

Log-Loss. To compute the loss of positive and negative instances for each task, we use the following equations:

$$\begin{aligned} \mathcal{L}^p_{\text{pos}} &= \frac{1}{N_{\text{pos}}} \sum_{n=1}^{N} (y_{np} \log \sigma(f_{np})) \\ \mathcal{L}^p_{\text{neg}} &= \frac{1}{N_{\text{neg}}} \sum_{n=1}^{N} ((1 - y_{np}) \log(1 - \sigma(f_{np}))) \end{aligned} \tag{10}$$

3 Results and Discussions

Table 2. Comparison of different loss functions with ECFP and BERT features. We also showed the effect of BERT fine-tuning

Model	Loss type				Features		Finetuning	Metrics			
	BCE	BCE^w	FL	FL^w	ECFP	BERT		BA	F1	ROC	AP
RF	-	-	-	-	-	-	-	0.67 ± 0.002	0.36 ± 0.003	0.65 ± 0.004	0.27 ± 0.003
MT	✓	-	-	-	✓	-	-	0.67 ± 0.004	0.34 ± 0.001	0.62 ± 0.003	0.26 ± 0.002
	-	✓	-	-	✓	-	-	0.66 ± 0.003	0.34 ± 0.004	0.63 ± 0.002	0.26 ± 0.001
	-	-	✓	-	✓	-	-	0.67 ± 0.004	0.37 ± 0.002	0.64 ± 0.003	0.28 ± 0.004
	-	-	-	✓	✓	-	-	0.68 ± 0.001	0.35 ± 0.003	0.65 ± 0.002	0.26 ± 0.001
	✓	-	-	-	-	✓	-	0.68 ± 0.003	0.37 ± 0.004	0.65 ± 0.001	0.28 ± 0.003
	-	✓	-	-	-	✓	-	0.70 ± 0.002	0.38 ± 0.001	0.67 ± 0.003	0.29 ± 0.002
	-	-	✓	-	-	✓	-	0.70 ± 0.001	0.39 ± 0.003	0.67 ± 0.004	0.31 ± 0.001
	-	-	-	✓	-	✓	-	0.72 ± 0.004	0.40 ± 0.002	0.70 ± 0.003	0.30 ± 0.001
	✓	-	-	-	-	✓	✓	0.73 ± 0.001	0.37 ± 0.002	0.70 ± 0.003	0.28 ± 0.004
	-	✓	-	-	-	✓	✓	0.72 ± 0.004	0.37 ± 0.001	0.70 ± 0.002	0.29 ± 0.003
	-	-	✓	-	-	✓	✓	0.72 ± 0.003	0.38 ± 0.004	0.69 ± 0.001	0.30 ± 0.002
	-	-	-	✓	-	✓	✓	0.72 ± 0.002	0.37 ± 0.003	0.68 ± 0.002	0.28 ± 0.001

We observed a significant performance gain when using BERT features compared to ECFP, as highlighted in Table 2. Figure 1 compares weighted and non-weighted loss functions for both frozen and fine-tuned BERT models. For the frozen BERT model, we observed a consistent performance improvement in balanced accuracy, F1-score, ROC-AUC, and average precision across all modalities

when transitioning from Binary Cross-Entropy (BCE) to weighted Binary Cross-Entropy (BCE-w), Focal Loss (FL), and weighted Focal Loss (FL-w). However, this trend was not consistent in the fine-tuned BERT model. Fine-tuning BERT improved performance in some modalities but decreased it in others. In conclusion, BERT features outperformed ECFP, weighted loss functions were superior to unweighted ones, and the effectiveness of fine-tuning remained inconclusive.

To analyze the impact of different loss functions, we computed the log-loss for positive and negative instances generated from Frozen-BERT, as shown in Fig. 2. Task-wise means, calculated using Eq. 10, are represented by small blobs with ellipses indicating the 95% confidence interval across all tasks. The visualization highlighted a network bias towards the majority class (negatives), resulting in elevated log-loss for positive instances, particularly with Binary Cross-Entropy (BCE). Transitioning to weighted Binary Cross-Entropy (BCE-W) prompted the model to equally prioritize both positives and negatives, decreasing log-loss for positives compared to BCE. Moreover, Focal Loss naturally emphasizes on hard examples, in this case, positive instances, lead to significantly lower log-loss for positives. Weighted Focal Loss further supported this by assigning additional weight to positive examples, consequently reducing the log-loss of positive instances even further. Further, in our experience frozen-BERT model provided the highest ROC-AUC with weighted Focal loss as depicted in the bottom row of the Fig. 2.

We computed a task-wise performance, as depicted in Fig. 3. Our models achieved the highest ROC-AUC for biochemistry related tasks and lowest for the clinical tasks the highest ROC scores. Interestingly, the model encounters difficulty in learning two specific tasks: Extramedullary from the pathological category and 100197554 from the clinical category.

4 Supplementary Information

The related supplementary information can be found on project GitHub repository https://github.com/Arslan-Masood/Tox_balance

Acknowledgments. The authors acknowledge financial support from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 956832, "Advanced Machine learning for Innovative Drug Discovery" (AIDD).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

Ahmad, W., Simon, E., Chithrananda, S., Grand, G. and Ramsundar, B.: ChemBERTa-2: Towards chemical foundation models. arXiv:2209.01712 (2022)

Ai, H., et al.: Predicting drug-induced liver injury using ensemble learning methods and molecular fingerprints. Toxicol. Sci. **165**(1), 100–107 (2018). ISSN 1096-6080, 1096-0929. https://doi.org/10.1093/toxsci/kfy121, https://academic.oup.com/toxsci/article/165/1/100/5000032

Alashwal, H., Lucman, J.: Utilizing cost-sensitive machine learning classifiers to identify compounds that inhibit Alzheimer's APP translation. In: Proceedings of the 2020 4th International Conference on Cloud and Big Data Computing, pp. 113–117, Virtual United Kingdom. ACM (2020). ISBN 978-1-4503-7538-2. https://doi.org/10.1145/3416921.3416931, https://dl.acm.org/doi/10.1145/3416921.3416931

Asilar, E., Hemmerich, J., Ecker, G.F.: Image based liver toxicity prediction. J. Chem. Inform. Model. **60**(3), 1111–1121 (2020). ISSN 1549-9596, 1549-960X. https://doi.org/10.1021/acs.jcim.9b00713, https://pubs.acs.org/doi/10.1021/acs.jcim.9b00713

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. **16**, 321–357 (2002). ISSN 1076-9757. https://doi.org/10.1613/jair.953, https://www.jair.org/index.php/jair/article/view/10302

Chen, M., Borlak, J., Tong, W.: High lipophilicity and high daily dose of oral medications are associated with significant risk for drug-induced liver injury. Hepatology, **58**(1), 388–396 (2013). ISSN 02709139. https://doi.org/10.1002/hep.26208, https://onlinelibrary.wiley.com/doi/10.1002/hep.26208

Chen, M., et al. Quantitative structure-activity relationship models for predicting drug-induced liver injury based on FDA-approved drug labeling annotation and using a large collection of drugs. Toxicol. Sci. **136**(1), 242–249 (2013). ISSN 1096-6080, 1096-0929. https://doi.org/10.1093/toxsci/kft189, https://academic.oup.com/toxsci/article-lookup/doi/10.1093/toxsci/kft189

Chithrananda, S., Grand, G. and Ramsundar, B.: ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885 (2020)

Cruz-Monteagudo, M., Cordeiro, M.N.D., Borges, F.: Computational chemistry approach for the early detection of drug-induced idiosyncratic liver toxicity: early Detection of Drug-Induced Idiosyncratic Liver Toxicity. Jo. Comput. Chem. **29**(4), 533–549 (2008.) ISSN 01928651. https://doi.org/10.1002/jcc.20812, https://onlinelibrary.wiley.com/doi/10.1002/jcc.20812

Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning, ICML 2006, pp. 233–240, New York, NY, USA. Association for Computing Machinery (2006). ISBN 978-1-59593-383-6. https://doi.org/10.1145/1143844.1143874, https://doi.org/10.1145/1143844.1143874

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2019)

Eitrich, T., Kless, A., Druska, C., Meyer, W., Grotendorst, J.: Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques. J. Chem. Inform. Model. **47**(1), 92–103 (2007). ISSN 1549-9596. https://doi.org/10.1021/ci6002619, https://doi.org/10.1021/ci6002619. Publisher: American Chemical Society

Ekins, S., Williams, A.J., Xu, J.J.: A predictive ligand-based bayesian model for human drug-induced liver injury. Drug Metab. Dispos. **38**(12), 2302–2308 (2010). ISSN 0090-9556, 1521-009X. https://doi.org/10.1124/dmd.110.035113, http://dmd.aspetjournals.org/lookup/doi/10.1124/dmd.110.035113

Elkan, C.: The foundations of cost-sensitive learning. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI 2001, pp. 973–978, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.(2001). ISBN 978-1-55860-812-2

Fabian, B., et al.: Molecular representation learning with language models and domain-relevant auxiliary tasks. arXiv:2011.13230 (2020)

Forman, G., Scholz, M.: Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. ACM SIGKDD Explor. Newslett. **12**(1), 49–57 (2010). ISSN 1931-0145, 1931-0153. https://doi.org/10.1145/1882471.1882479, https://dl.acm.org/doi/10.1145/1882471.1882479

Gao, K., Nguyen, D.D., Sresht, V., Mathiowetz, A.M., Tu, M., Wei, G.W.: Are 2D fingerprints still valuable for drug discovery? Phys. Chem. Chem. Phys. **22**(16), 8373–8390 (2020). ISSN 1463-9076, 1463-9084. https://doi.org/10.1039/D0CP00305K, http://xlink.rsc.org/?DOI=D0CP00305K

Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27 (2014)

Harnik, Y., Milo, A.: A focus on molecular representation learning for the prediction of chemical properties. Chem. Sci. **15**(14), 5052–5055 (2024). ISSN 2041-6520, 2041-6539. https://doi.org/10.1039/D4SC90043J, https://xlink.rsc.org/?DOI=D4SC90043J

Igarashi, Y., et al.: Open TG-GATEs: a large-scale toxicogenomics database. Nucleic Acids Res. **43**(D1), D921–D927 (2015). ISSN 1362-4962, 0305-1048. https://doi.org/10.1093/nar/gku955, https://academic.oup.com/nar/article/43/D1/D921/2439524

Irwin, R., Dimitriadis, S., He, J., Bjerrum, E.J.: Chemformer: a pre-trained transformer for computational chemistry. Mach. Learn. Sci. Technol. **3**(1), 015022 (2022). ISSN 2632-2153. https://doi.org/10.1088/2632-2153/ac3ffb, https://dx.doi.org/10.1088/2632-2153/ac3ffb. Publisher: IOP Publishing

Jain, S., et al.: Large-scale modeling of multispecies acute toxicity end points using consensus of multitask deep learning methods. J. Chem. Inform. Model. **61**(2), 653–663 (2021). ISSN 1549-9596, 1549-960X. https://doi.org/10.1021/acs.jcim.0c01164, https://pubs.acs.org/doi/10.1021/acs.jcim.0c01164

Kim, J.H., Scialli, A.R.: Thalidomide: the tragedy of birth defects and the effective treatment of disease. Toxicol. Sci. **122**(1), 1–6 (2011). ISSN 1096-6080, 1096-0929. https://doi.org/10.1093/toxsci/kfr088, https://academic.oup.com/toxsci/article/1672454/Thalidomide:

Kingma D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

Kuhn, M., Letunic, I., Jensen, L.J., Bork, P.: The SIDER database of drugs and side effects. Nucleic Acids Res. **44**(D1), D1075–D1079 (2016) ISSN 0305-1048, 1362-4962. https://doi.org/10.1093/nar/gkv1075, https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1075

Lancaster, E.M., Hiatt, J.R., Zarrinpar, A.: Acetaminophen hepatotoxicity: an updated review. Arch. Toxicol. **89**, 193–199 (2014). https://doi.org/10.1007/s00204-014-1432-2

Laveti, R.N., Mane, A.A., Pal, S.N.: Dynamic stacked ensemble with entropy based undersampling for the detection of fraudulent transactions. In: 2021 6th International Conference for Convergence in Technology (I2CT), pp. 1–7, Maharashtra,

India. IEEE (2021). ISBN 978-1-72818-876-8. https://doi.org/10.1109/I2CT51068.2021.9417896, https://ieeexplore.ieee.org/document/9417896/

Lee, W., Seo, K.: Downsampling for binary classification with a highly imbalanced dataset using active learning. Big Data Res. **28**, 100314 (2022). ISSN 22145796. https://doi.org/10.1016/j.bdr.2022.100314, https://linkinghub.elsevier.com/retrieve/pii/S2214579622000089

Li, J., Jiang, X.: Mol-BERT: an effective molecular representation with BERT for molecular property prediction. Wireless Commun. Mob. Comput. **2021**, 1–7 (2021). ISSN 1530-8677, 1530-8669. https://doi.org/10.1155/2021/7181815, https://www.hindawi.com/journals/wcmc/2021/7181815/

Li, X., Fourches, D.: Inductive transfer learning for molecular activity prediction: *Next-Gen QSAR Models with MolPMoFiT*. J. Cheminform. **12**(1), 1–15 (2020). https://doi.org/10.1186/s13321-020-00430-x

Li, Z., Jiang, M., Wang, S., Zhang, S.: EEP learning methods for molecular representation and property prediction. Drug Discov. Today **27**(12), 103373 (2022). ISSN 1878-5832. https://doi.org/10.1016/j.drudis.2022.103373

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. arXiv:1708.02002 (2018)

Liu, Y., Zhang, R., Li, T., Jiang, J., Ma, J., Wang, P.: MolRoPE-BERT: an enhanced molecular representation with rotary position embedding for molecular property prediction. J. Mol. Graph. Model. **118**, 8344 (2023) ISSN 1093-3263. https://doi.org/10.1016/j.jmgm.2022.108344, https://www.sciencedirect.com/science/article/pii/S1093326322002236

López, V., Fernández, A., Moreno-Torres, J.G., Herrera, F.: Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. Expert Systems with Applications, **39**(7), 6585–6608 (2012). ISSN 09574174. https://doi.org/10.1016/j.eswa.2011.12.043, https://linkinghub.elsevier.com/retrieve/pii/S0957417411017143

Maziarka, Ł., Danel, T., Mucha, S., Rataj, K., Tabor, J., Jastrzębski, S.: Molecule attention transformer. arXiv:2002.08264 (2020)

Moein, M., et al.: Chemistry-based modeling on phenotype-based drug-induced liver injury annotation: from public to proprietary data. Chem. Res. Toxicol. **36**(8), 1238–1247 (2023). ISSN 0893-228X, 1520-5010. https://doi.org/10.1021/acs.chemrestox.2c00378, https://pubs.acs.org/doi/10.1021/acs.chemrestox.2c00378

Parasrampuria, D.A., Benet, L.Z., Sharma, A.: Why drugs fail in late stages of development: case study analyses from the last decade and recommendations. AAPS J **20**(3), 1–16 (2018). https://doi.org/10.1208/s12248-018-0204-y

Singh Rawat, S., Mishra, A.K.: Review of methods for handling class-imbalanced in classification problems. arXiv:2211.05456 (2022)

Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International Conference on Machine Learning, pp. 1530–1538. PMLR (2015)

Rogers, D., Hahn, M.: Extended-connectivity fingerprints. J. Chem. Inform. Model. **50**(5), 742–754) (2010). ISSN 1549-9596, 1549-960X. https://doi.org/10.1021/ci100050t, https://pubs.acs.org/doi/10.1021/ci100050t

Ruthotto, L., Haber, E.: An introduction to deep generative modeling (2021)

Sharma, B., et al.: Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations. Sci. Rep. **13**(1), 4908 (2023). ISSN 2045-2322. https://doi.org/10.1038/s41598-023-31169-8, https://www.nature.com/articles/s41598-023-31169-8

Shermukhamedov, S., Mamurjonova, D., Probst, M.: Structure to property: chemical element embeddings and a deep learning approach for accurate prediction of chemical properties arXiv:2309.09355 (2023)

Sosnin, S., Karlov, D., Tetko, I.V., Fedorov, M.V.: Comparative study of multitask toxicity modeling on a broad chemical space. J. Chem. Inform. Model. **59**(3), 1062–1072 (2019). ISSN 1549-9596, 1549-960X. https://doi.org/10.1021/acs.jcim.8b00685, https://pubs.acs.org/doi/10.1021/acs.jcim.8b00685

Stanley, L.A.: Chapter 27 - Drug Metabolism. In: Badal, S., Delgoda, R., (eds.) Pharmacognosy, pp. 527–545. Academic Press, Boston (2017). ISBN 978-0-12-802104-0. https://doi.org/10.1016/B978-0-12-802104-0.00027-5, https://www.sciencedirect.com/science/article/pii/B9780128021040000275

Wang, Y., Xiao, Q., Chen, P., Wang, B: In silico prediction of drug-induced liver injury based on ensemble classifier method. Int. J. Mol. Sci. **20**(17), 4106 (2019). ISSN 1422-0067. https://doi.org/10.3390/ijms20174106, https://www.mdpi.com/1422-0067/20/17/4106

Wang, H., Liu, R., Schyman, P., Wallqvist, A.: Deep neural network models for predicting chemically induced liver toxicity endpoints from transcriptomic responses. Front. Pharmacol. **10**, 42 (2019). ISSN 1663-9812. https://doi.org/10.3389/fphar.2019.00042, https://www.frontiersin.org/article/10.3389/fphar.2019.00042/full

Wang, S., Guo, Y., Wang, Y., Sun, H., Huang, J.: SMILES-BERT: large scale unsupervised pre-training for molecular property prediction. In: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pp. 429–436, Niagara Falls NY USA (2019). ACM. ISBN 978-1-4503-6666-3. https://doi.org/10.1145/3307339.3342186, https://dl.acm.org/doi/10.1145/3307339.3342186

Watkins, P.B.: Drug safety sciences and the bottleneck in drug development. Clin. Pharmacol. Ther. **89**(6), 788–790 (2011). ISSN 0009-9236, 1532-6535. https://doi.org/10.1038/clpt.2011.63, https://onlinelibrary.wiley.com/doi/10.1038/clpt.2011.63

Weininger, D.: SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inform. Comput. Sci. **28**(1), 31–36 (1988). ISSN 0095-2338. https://doi.org/10.1021/ci00057a005, https://doi.org/10.1021/ci00057a005. Publisher: American Chemical Society

Williams, D.P., Lazic, S.E., Foster, A.J., Semenova, E., Morgan, P.: Predicting drug-induced liver injury with Bayesian machine learning. Chem. Res. Toxicol **33**(1), 239–248 (2020). ISSN 0893-228X, 1520-5010. https://doi.org/10.1021/acs.chemrestox.9b00264, https://pubs.acs.org/doi/10.1021/acs.chemrestox.9b00264

Winter, R., Montanari, F., Noé, F., Clevert, D.A.: Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chem. Sci. **10**(6), 1692–1701 (2019). ISSN 2041-6520, 2041-6539. https://doi.org/10.1039/C8SC04175J, https://xlink.rsc.org/?DOI=C8SC04175J

Xu, Y., Dai, Z., Chen, F., Gao, S., Pei, J., Lai, L., Deep learning for drug-induced liver injury. J. Chem. Inform. Model. **55**(10), 2085–2093 (2015). ISSN 1549-9596, 1549-960X. https://doi.org/10.1021/acs.jcim.5b00238, https://pubs.acs.org/doi/10.1021/acs.jcim.5b00238

Yang, K., et al.: Analyzing learned molecular representations for property prediction. arXiv:1904.01561 (2019)

Zhu, M.: Recall, precision and average precision. Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, vol. 2, no. 30, p. 6 (2004)