Balancing Imbalanced Toxicity Model: Using MolBERT with Focal Loss

Muhammad Arslan Masood

19.09.2024

Drug Induced Liver Injury (DILI)

- 9 out of 10 drugs fail in clinical trials
- 50% of failures stem from unexpected toxicity
- Drug-induced liver injury (DILI) is a major culprit behind late-stage drug failures

Modeling DILI

- Challenges
  - Limited input space (limited chemistry)
  - Limited output space (limited targets)
  - Highly imbalanced dataset
- Proposed solution
  - Leverage pretraining to learn robust molecular representations
  - Incorporation of other (Hematology and clinical) modalities
  - Leverage weighted loss to tackle imbalance
  - Provide biological context in pretraining

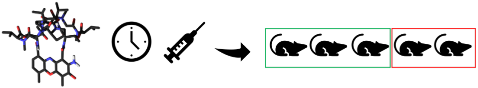
Preclinical Liver Histopathology Tasks

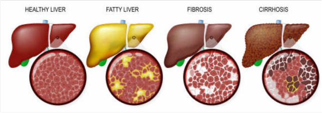
$$\mathcal{D}_{\text{invivo}} = \{(\mathsf{x}_i^{\text{invivo}}, \mathsf{y}_i)\}_{i=1}^{N}$$

$$\mathsf{y}_{i=(n,t,d,p)} \in \{y_i\}_{k=1}^{K}$$

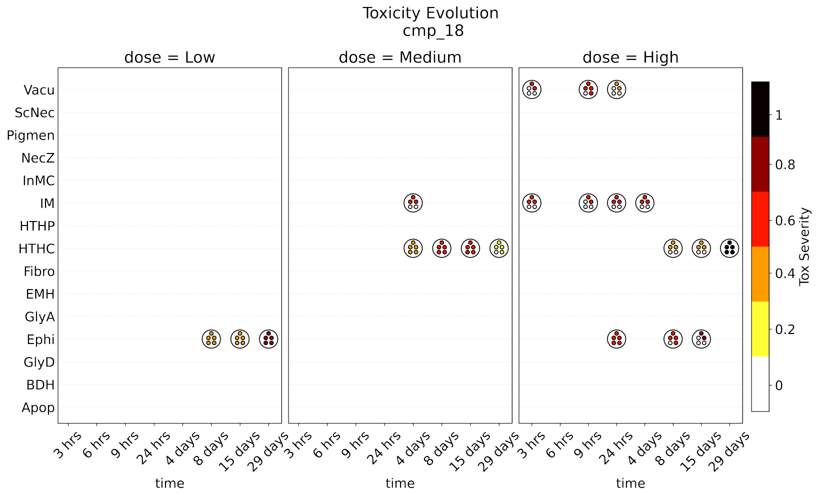
$$y_i \in \{s_0, s_1, \ldots, s_5\}$$

Compounds $n = 1, \ldots, N$
Timepoints $t \in \{t_1, \ldots, t_8 \in \mathbb{R}\}$
Doses $d \in \{d_1, d_2, d_3 \in \mathbb{R}\}$
Histopathological endpoints $p = 1, \ldots, P$
Multiple animal replicates $k = 1, \ldots, K$
Severity levels $y_i \in \{s_0, s_1, \ldots, s_5\}$





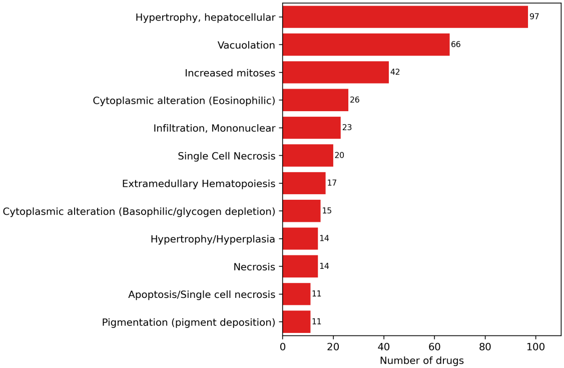
Preclinical Liver Histopathology Tasks



Binary Label assignment

- Pooled over dose and time and only consider the 'nominal' toxicity y_{np}.

$$y_{np} = \begin{cases} 1, & \text{if } \exists \text{ a combination } (d, t) \text{ such that } \sum_k \mathbb{I}(y_{ndtpk} \neq s_0) \geq 2 \\ 0, & \text{otherwise} \end{cases}$$



Expanding Output Space

Preclinical Hematology Tasks

- TG-GATES Hematology

Description	Thresholds
Alkaline Phosphatase (ALP)	1.5
Aspartate Aminotransferase (AST)	2
Alanine Aminotransferase (ALT)	2
Gamma-Glutamyl Transferase (GTP)	3
Total Cholesterol (TC)	1.5
Triglycerides (TG)	3
Total Bilirubin (TBIL)	dependent on $y_{control}$
Direct Bilirubin (DBIL)	dependent on $y_{control}$

$y_{ntdpk} \in \mathbb{R}$

$\mathbb{R} \rightarrow \{0, 1\}$

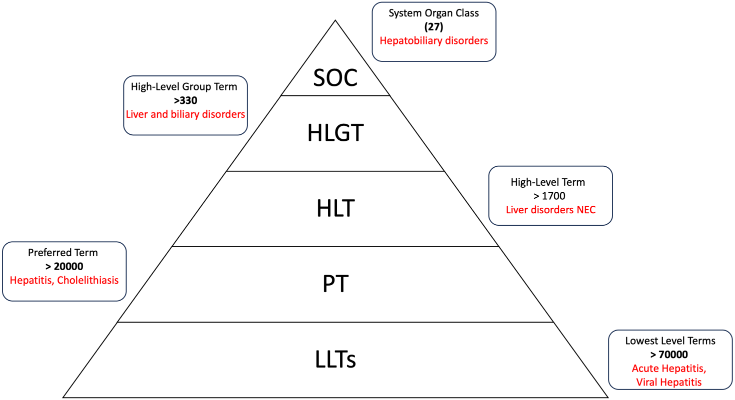
$y_{\text{control}} = \frac{1}{K} \sum_{k=1}^{K} y_{ntdpk} \mid d \in \{0\}$

$y_{\text{obs}} = y_{ntdpk} \mid d \notin \{0\}$

$$y'_{ndtpk} = \begin{cases} 1, & \text{if } y_{\text{obs}} > \text{th} \times y_{\text{control}} \\ 0, & \text{otherwise} \end{cases}$$

Clinical Tasks

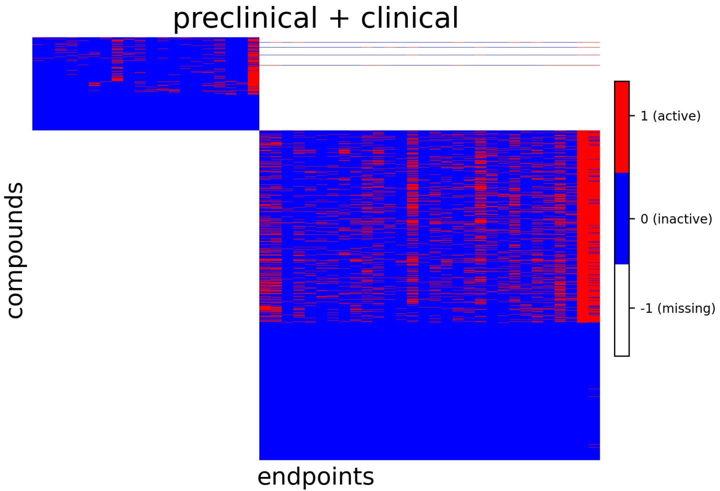
- SIDER dataset (1430 drugs, 6060 ADRs)[1]



[1] https://www.meddra.org/how-to-use/basics/hierarchy

Preclinical + clinical output space

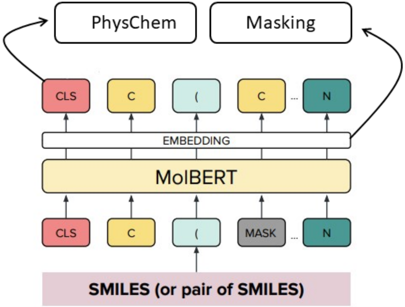
- Preclinical: drugs 410, task 20, 7.38% active
- Clinical: drugs 1219, task 28+2, 13.7% active



Expanding Input space

Molecular Representation learning

- Unsupervised pretraining
  - MolBERT [2]
  - GuacaMol benchmark dataset (1.26m compounds)
  - 12 attention heads, 12 layers, 768 dimensional hidden layer, $\approx$ 85M parameters

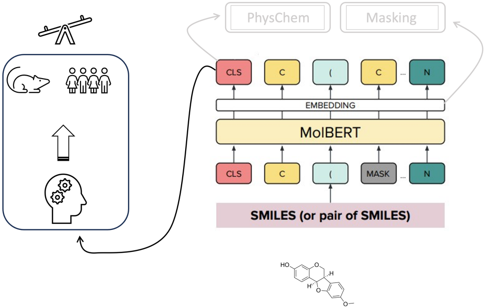


[2] Fabian et. al. Molecular representation learning with language models and domain-relevant auxiliary tasks

Molecular Representation learning

Given a set of SMILES and invivo labels
$\mathcal{D}_{\text{invivo}} = \{(s_n^{\text{invivo}}, \mathbf{y}_n)\}_{n=1}^{N}$

$$\mathbf{h}_n = \mathrm{MolBERT}(s_n^{\text{invivo}}; \Theta^*_{\text{pretrain}}), \quad \mathbf{h}_n \in \mathbb{R}^d$$
$$\hat{f}_{np} = \mathrm{head}_{\text{invivo}}(\mathbf{h}_n; \Theta_{\text{MLP}})$$



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

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

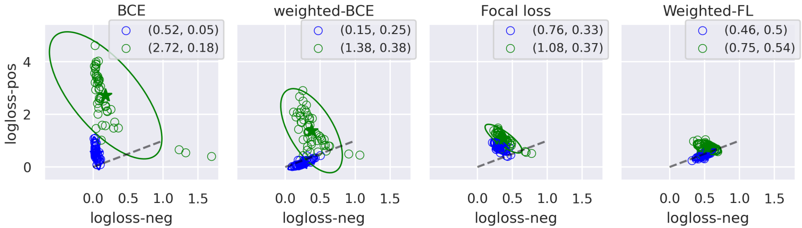
where $\alpha \in [0, 1]$ controls the positive balancing
Focal loss: $\alpha = 0$
Weighted BCE: $\gamma = 0$
BCE: $\alpha = 0$, $\gamma = 0$

Effect of weighting

Taskwise log-loss of positive and negative instances

$$\mathcal{L}_{\text{pos}}^{p} = \frac{1}{N_{\text{pos}}} \sum_{n=1}^{N} \left(y_{np} \log \sigma(f_{np}) \right)$$

$$\mathcal{L}_{\text{neg}}^{p} = \frac{1}{N_{\text{neg}}} \sum_{n=1}^{N} \left((1 - y_{np}) \log(1 - \sigma(f_{np})) \right)$$



Results

- Comparison with baselines

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
	✓	-	-	-	✓	-	-	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
MT	✓	-	-	-	-	✓	-	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

Results

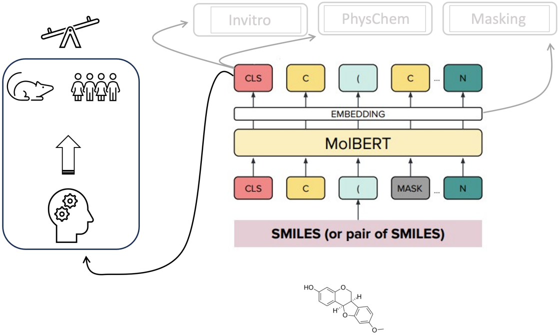
- Weighted losses are better than their non-weighted counterparts
- Focal loss > BCE

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
	✓	-	-	-	✓	-	-	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
MT	✓	-	-	-	-	✓	-	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

Extension of this work

ToxBERT

- MolBERT
  - learns only chemically driven representations
  - lacking in biological knowledge
- ToxBERT
  - Chemically driven rerpesentaions through masking and Physchem
  - Biological interactions through invitro pretraining



MolBERT vs ToxBERT

- invitro-BERT outperforms Random Forest by 27% in Hematology and 32% in Pathology tasks
- invitro-BERT outperforms MolBERT by 29% in Hematology and 10% in Pathology tasks

Model	Features	Loss type	AUPR			
			Hematology	Pathology	Clinical	Combined
RF	ECFP	-	0.37 ± 0.003	0.21 ± 0.002	0.26 ± 0.001	0.27 ± 0.004
MT-MLP	ECFP	BCE	0.33 ± 0.002	0.26 ± 0.003	0.26 ± 0.004	0.26 ± 0.003
	ECFP	BCE^w	0.32 ± 0.004	0.26 ± 0.001	0.25 ± 0.002	0.26 ± 0.003
	ECFP	FL	<u>0.38</u> ± 0.002	<u>0.31</u> ± 0.003	<u>0.26</u> ± 0.004	<u>0.28</u> ± 0.004
	ECFP	FL^w	0.31 ± 0.003	0.28 ± 0.004	0.25 ± 0.002	0.26 ± 0.003
MolBERT	MolBERT	BCE	0.27 ± 0.003	0.23 ± 0.002	0.29 ± 0.004	0.28 ± 0.003
	MolBERT	BCE^w	0.30 ± 0.002	0.27 ± 0.004	0.29 ± 0.001	0.29 ± 0.003
	MolBERT	FL	<u>0.36</u> ± 0.001	0.26 ± 0.002	<u>0.31</u> ± 0.004	<u>0.31</u> ± 0.002
	MolBERT	FL^w	<u>0.36</u> ± 0.003	<u>0.28</u> ± 0.001	0.29 ± 0.004	0.30 ± 0.003
Physchem-BERT	ToxBERT	FL^w	0.46 ± 0.004	0.28 ± 0.003	0.29 ± 0.002	0.31 ± 0.003
invitro-BERT	ToxBERT	FL^w	**0.51** ± 0.002	**0.31** ± 0.003	0.30 ± 0.004	**0.34** ± 0.003

Conclusions

Toxicity is a challenging task to model, following solutions might help

- Expand output space by incorporating closely related modalities
- Expand input space by leveraging unsupervised pretraining
- Leverage auxiliary data to learn better context

Acknowledgment

This project is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 956832, and it is Horizon 2020 Marie Skłodowska-Curie Innovative Training Network - European Industrial Doctorate.