Assessment Report **Lu Zhu**
InVivo AI zhu.lu@hotmail.com

Problem Statement

Build classifiers that predict each of the target values from the given the following dataset. The first column contains the SMILES representation of the molecules and the subsequent columns contain binary information about the molecules (ex. toxic /non-toxic).

Summary

In this report, I established prediction models for chemical toxicity prediction. Hidden features are derived from the SMILES strings, which represent the chemical structures, instead of learning from crafted features. To discover hidden representations for the SMILES strings, the convolutional neural networks (CNNs) were used. To improve the accuracy for a given target, the activity of compounds for the other targets were also used as features for the toxicity prediction. The source code is available at `https://github.com/zhu0619/Toxicity`.

Investigation

Representation of chemical compound

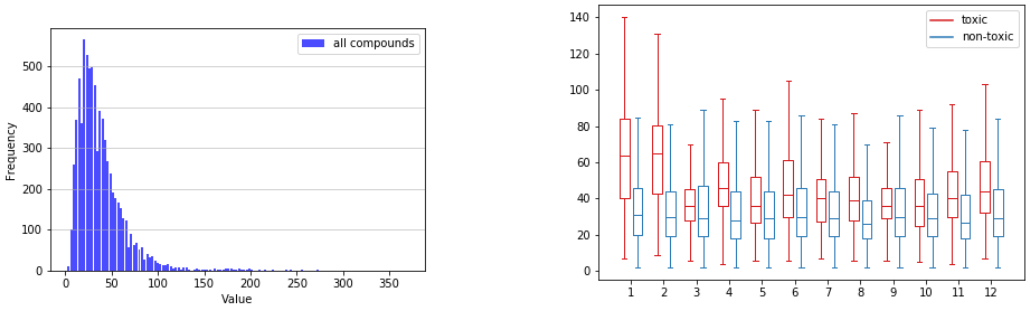
Among the many types of chemical information of chemical compounds (such as weight, molecular formula, rings, atoms, SMILES and InChI), SMILES represents the chemical structure as a line of ASCII characters. Atoms (e.g., carbon, nitrogen, and oxygen), bonds (e.g., single, double, and triple bonds), rings (e.g., open ring, close ring, and ring number), aromaticity, and branching can be represented with SMILES. As a one-dimensional representation for the chemical structure, SMILES can be converted into a two- or three-dimensional chemical structure, which means that it contains sufficient structure information to derive higher dimensions [2].

Canonicalization

There are usually a large number of valid generic SMILES which represent a given structure. However, one chemical compound can be written in different SMILES structures. For example, `C1=COC=C1` can be also written in `c1cocc1`. A canonicalization algorithm exists to generate one special generic SMILES among all valid possibilities; this special one is known as the "unique SMILES". SMILES written with isotopic and chiral specifications are collectively known as "isomeric SMILES" [1]. The isotopic and the chiral specifications are potentially important for drug toxicity. therefore, in this project, I choose "isomeric SMILES" format to canonicalize the given SMILES from dataset. The isomeric SMILES can be requested by "pubchempy" API `https://pubchempy.readthedocs.io/en/latest/api.html`. 84 compounds which the isomeric SMILES are unavailable from "pubchempy" were excluded for the analysis.

SMILES length of chemical compounds

In general, SMILES strings have variable lengths depending on the complexity of the chemical structure. To effectively learn the hidden representations for SMILES, the inputs are limited to a specified maximum length, λ. Zero values are pre-padded if a sequence is shorter than



(a) Histogram of SMILES for target1

(b) Length distribution for each target.

Figure 1: SMILES length distribution. The median length is 32, the average length is 38.60, maximum length is 370. Majority of the SMILES length is less than 100.

the maximum length. Otherwise, the sequence is truncated to the maximum length. All input lengths are fixed to the maximum length of λ after the truncated or padded processes [2].

As shown in Figure 1a, for most of the compounds, the SMILES length is less than 100 characters. Figure 1b demonstrates that the SMILES length for both toxic and non-toxic compounds, while the length of toxic compounds is larger than the non-toxic compounds in general. Therefore, I limit the SMILES to a maximum length, λ, as 100. The input x is encoded into a $\lambda \times |X|$-dimensional binary matrix according to the encoding scheme and maximum length. $|X|$ is the number of characters appeared in the all SMILES. The input collection is represented by a $N_t \times \lambda \times |X|$- dimensional dimensional binary tensor.

Evaluation metrics

In the case of toxicity prediction, we should avoid misprediction of the toxic compounds as non-toxic compounds as possible. That means to maximize the recall of toxic compounds. At the same time, we should also maximize the precision of non-toxic compounds.

As a prediction result, every compound is assigned with a probability values for each target. In practice, I'd rather not use AUC of ROC as metrics for tuning the parameters due to its weakness on the imbalanced dataset. Also, I observed during the experiments, an optimal threshold value based on the ROC curve often favors the majority class (non-toxic), which gives bad recall and precision for the minority class(toxic) In this project, I use 0.5 as a fixed threshold, and evaluate the models with precision, f-beta score (beta: 1.5, with favor to recall) which reflects the overall precision and recall, and recall of the toxic class.

Evaluation and analysis

Imbalance dataset and weighted CNNs

Figure 2 shows that the dataset is highly imbalanced for each target. The predictor can be overwhelmed by the majority class. In that case, a high overall accuracy often comes with low precision and low recall for the monority class. In this case, a predictor can achieve high overall accuracy by predicting mainly 'non-toxic'. The recall and precision for the 'toxic' class could be very low. The typical solutions for learning form imbalance dataset is to

1. Re-sampling; over-sampling the minority class; under-sampling the majority class.
2. Class weight to enhance the minority class.

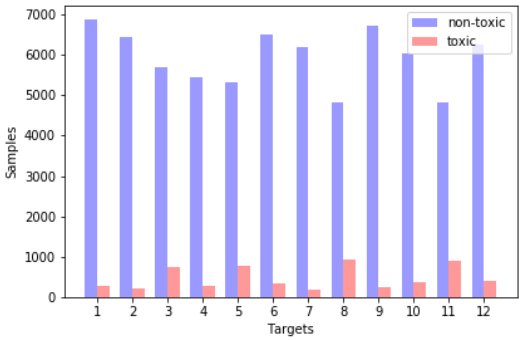


Figure 2: Number of samples in non-toxic and toxic classes for 12 targets.

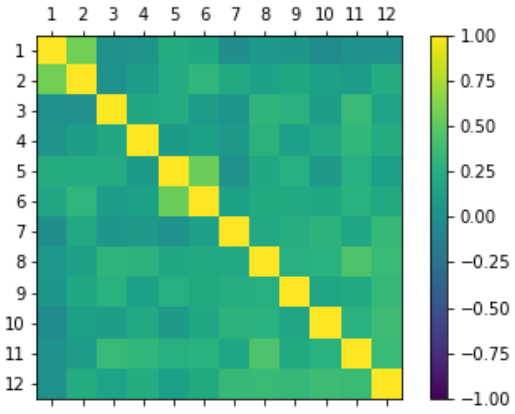


Figure 3: Activity correlation between targets.

To address the effect of these two approaches, I performed 5-fold cross-validation on the same datasets.

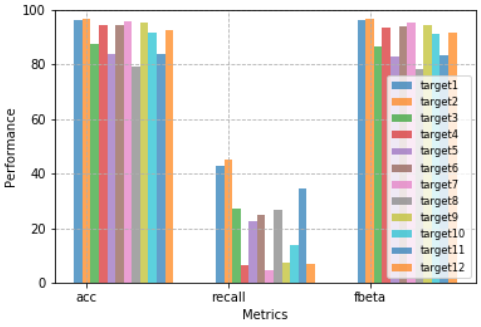
As we can see in Figure 4a, although the re-sampling strategy can reach a higher accuracy and f beta score, the recall for the toxic class is much worse than the class weighted CNN model. The CNN model using class weighted can achieve better recall for predicting the toxic chemical compounds.

The toxicity correlation of the 12 targets

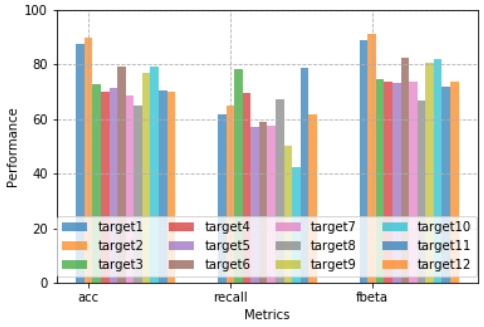
To investigate whether the information about other targets can improve the accuracy for a given target, I performed a correlation analysis of the toxicities of 12 targets. From the correlation matrix and the plot (Figure 3), we observe the strong correlation between target1 and target2 (0.57), target5 and target6 (0.54), and target8 and target11 (0.44). For the rest pairs of targets, the correlations are lower than 0.4. Nevertheless, the activity of a compound on the other targets would be helpful for toxicity prediction.

Usefulness of the activity on the other targets

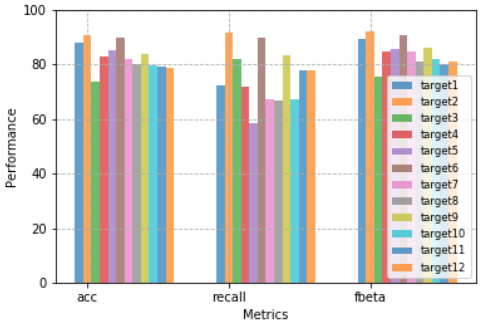
To take advantage of the activity information on the other targets, I include multiplying the binary vectors 11×1 by the number of maximum length size to $11 \times \lambda$. Therefore, the input



(a) Re-sampling



(b) Weighted class



(c) Weighted class + target features

Figure 4: Evaluation results on the test dataset. The performances are the average values from 5-fold cross validation using threshold of 0.5. The 'acc' represents the overall accuracy of prediction. The 'fbeta' indicates the overall weighted average F beta. The 'recall' is the recall for the minor class ('toxic' class).

collection is represented by an $N \times \lambda \times (|X| + 11)$- dimensional binary tensor.

5-fold cross-validation was performed for the CNN model with both class weight and the activities of the other targets. Figure 4c shows the final performance of the test dataset. Obviously, the overall accuracy the f-beta score, and the recall of toxic class can be improved by adding the activities of the other targets to baseline class weighted CNN model.

Appendix

1. The detailed training processes of each model can be found in `*_targets.loss.cv.sub.png` and `*_targets.acc.cv.sub.png`.
2. The pre-built models are exported to `*.h5` files.
3. The predicted toxicity of the compounds which were previously unknown can be found the target*prediction.csv.(Default threshold is 0.5.)

References

[1] Daylight Headquarters. Smiles - a simplified chemical language.

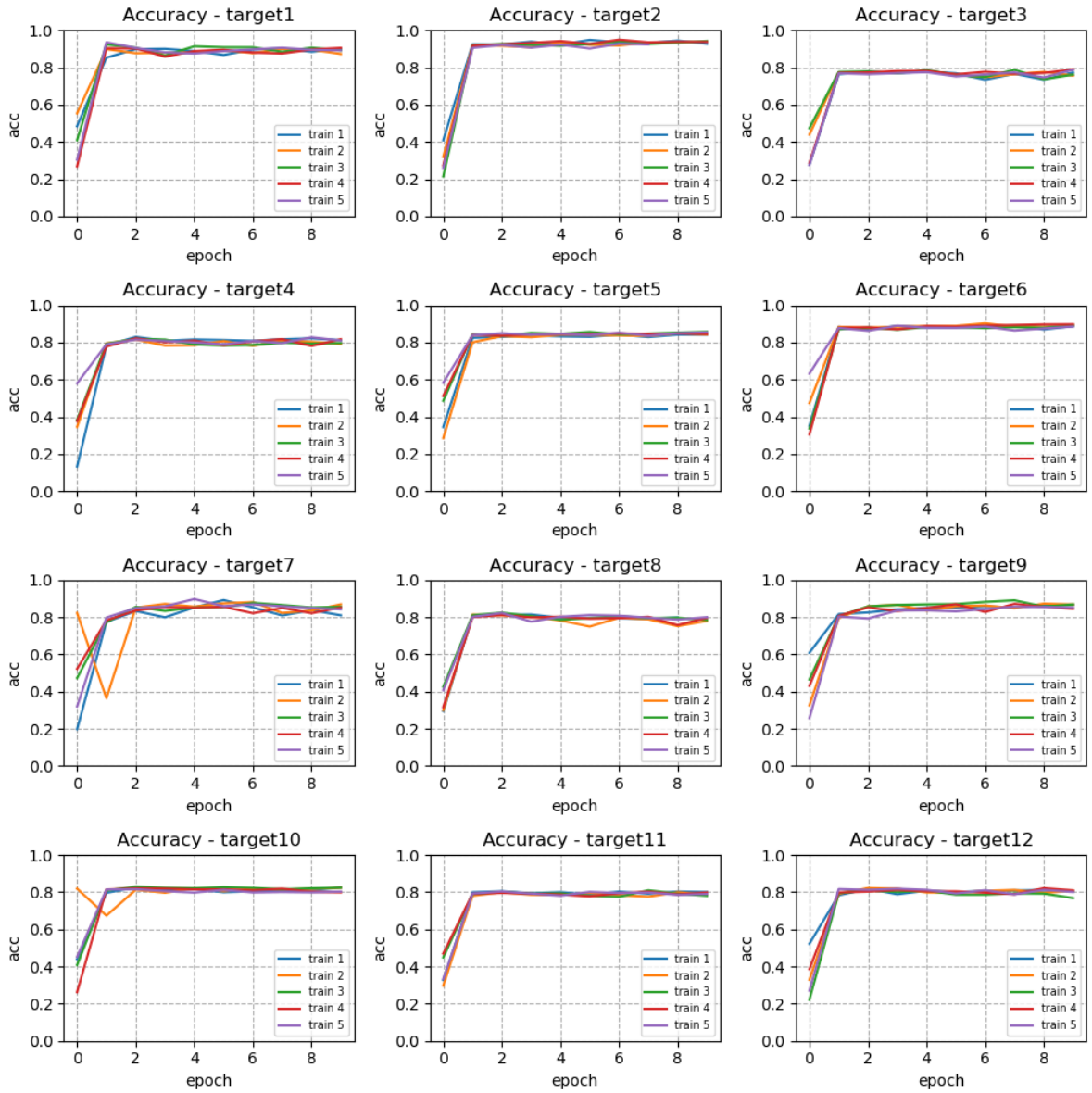


Figure 5: weighted.target.12targets.loss

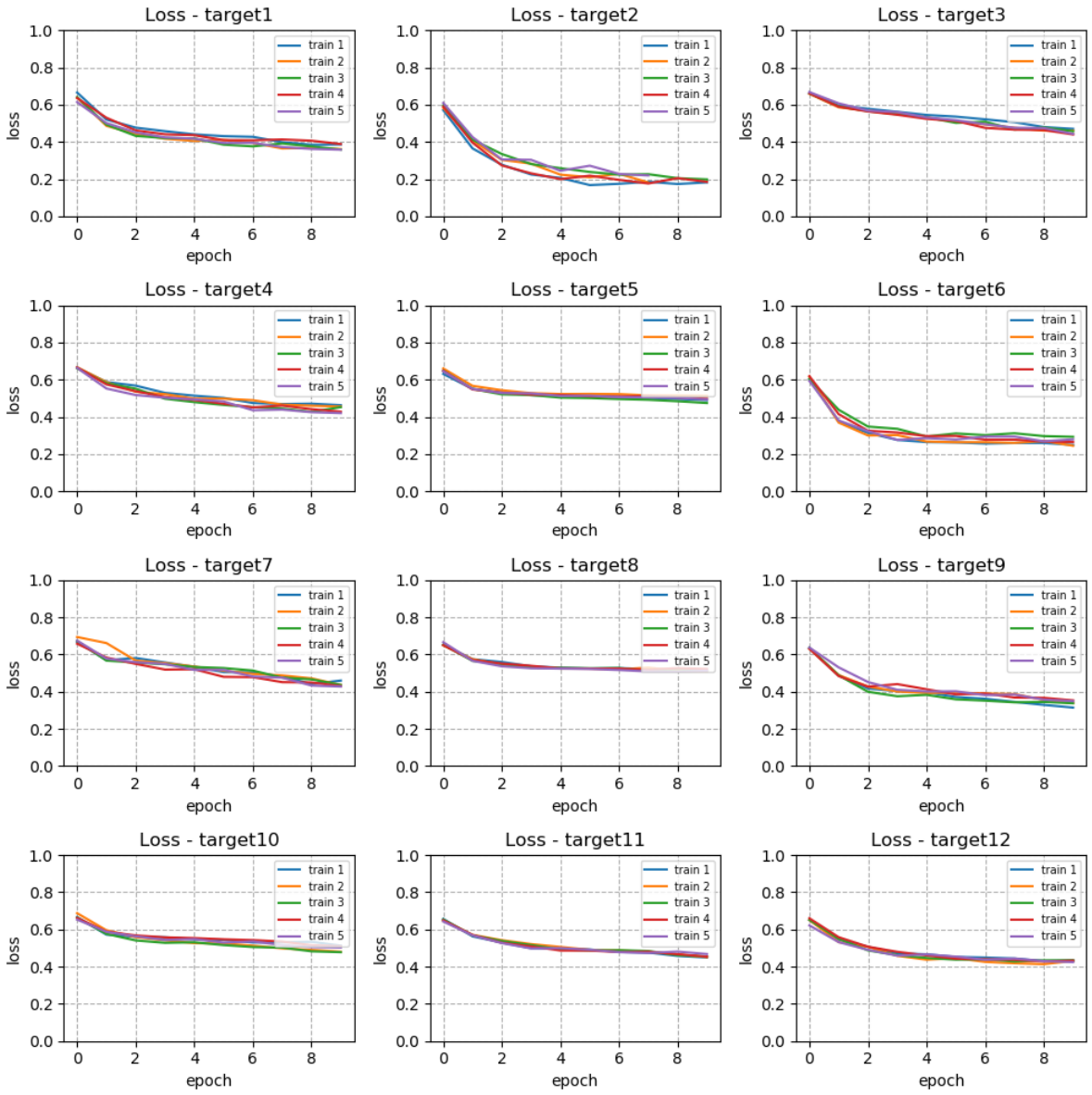


Figure 6: weighted.target.12targets.loss

[2] Sunyoung Kwon and Sungroh Yoon. DeepCCI. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics - ACM-BCB '17*, 2017.