Molecular Toxicity Prediction with SMILES Representations

Christopher Rohlicek
christopher_rohlicek@brown.edu

Cameron Webster
cameron_webster@brown.edu

Jillian Green
jillian_green@brown.edu

Akshay Shah
akshay_shah@brown.edu

Abstract

Many of the recent advances in machine learning research which have garnered the most public attention are in the domain of natural language processing. However, these new sequence-learning methods can be applied just as well to problems in chemistry. Here, we explore three approaches to the chemistry problem of predicting a compound's toxicity from its expression in SMILES format (1). The approaches we consider are the application of an InceptionResNet to 2-dimensional image representations of our set of molecules, a bidirectional LSTM applied to encodings of our SMILES sequences created by applying a pre-trained BERT-based model that is tailored to SMILES data, and finally a model that combines the feature embeddings of these two models. Measuring model performance by the AUROC achieved by the trained classifier, our InceptionResNet-based CNN model reached a peak validation score of 0.787, our LSTM reached a peak validation score of 0.691, and our composite model reached a peak validation score of 0.942.

1 Introduction

Chemistry is a field that has seen a surge in research motivated by some of the latest techniques in machine learning for applications in the realm of natural language processing. These techniques – such as transformers, attention-based encoding schemes, and powerful recurrent neural network architectures – have gained a great deal of attention for their results in language tasks, but their impact on chemistry is equally groundbreaking. Because of these techniques' fundamental nature as sequence-learning strategies, their application to chemistry is a natural consequence of the treatment of molecular compounds as sequences of atoms and bonds.

In this project we investigate methods of processing molecular sequence data as represented in the SMILES format (1) for the prediction of general toxicity, as indicated by the presence of the SR-P53 protein, a compound shown to indicate the presence of cancerous cells (2). Our dataset comes from the National Institutes of Health (3) and we split it into training, validation, and test sets used across our different models.

The methods that guided our experimentation lie both in the sequence and image domains, applying RNNs and LSTMs to SMILES sequences and CNNs to image representations of our SMILES sequence. These two approaches are main themes in the literature on this type of problem because a SMILES string can reasonably be treated both as a sequence of symbols from a chemical alphabet and as a blueprint for a 2-dimensional image of a compound. Following from these two approaches we implemented both a CNN-based and LSTM-based model, and a third model which combines the two feature representations from the constituent models, which is an approach we believe to be a novelty.

2 Related work

The body of research on problems in molecular property prediction is varied and expanding rapidly. However, some of the main methods that served as points of departure for our experiments were the use of CNNs to learn molecular features from 2-dimensional representations (4; 5), the use of RNNs and LSTMs to learn features from the sequence form of the SMILES representation (6; 7), and using variants of BERT to learn bidirectional context-based embeddings of SMILES sequences (8; 9).

The CNN- and LSTM-based approaches call mostly on the established practices relating to the use of those families of network architectures, requiring little change that is specific to the domain of chemistry. An especially well-suited CNN architecture that appears in the literature and in (4) specifically is the InceptionNet which is defined by its blocks comprising groups of differently-sized filters that are capable of detecting patterns at a variety of sizes. This is useful in the case of processing images of molecules because the features of importance will be represented as different clusters of atoms arranged in recurring patterns that will inevitably be of varying size.

The application of LSTMs and BERT models to SMILES sequences borrows much of the same motivation that guides their application in language-related problems. To tailor BERT models to SMILES data, it is a common practice to train a given encoder on a large composite SMILES dataset in an unsupervised way that finds an optimal encoding strategy through a Masked Language Modeling (MLM) approach. In our experiments we used one such encoder, a derivative of RoBERTa (10) called ChemBERTa, which was trained on a large collection of different unlabeled SMILES data (8).

3 Dataset and Features

The raw dataset consists of two text files from The National Institutes of Health (3), one used for training and validation examples and one for test examples. Each file contains two properties: one column containing the SMILES string representation of molecules and one containing the binary target labels that indicate toxicity. The raw table for training data was then split, keeping 80% of observations for training and 20% of observations for validation. This left a total of 5792 training observations, 1402 for validation, and 268 for testing.

From here, two diverging sequences of preprocessing steps took place. For the CNN input, the sequences were mapped to 300 by 300 RGB images of the molecules using the open-source RDKit library's drawing tool. No normalization nor augmentation was applied to the images since the pixel values were already within the same 0-255 range and the image generation was consistent across all molecules in terms of visual representations of SMILES tokens so augmentation was not deemed to be necessary.

For the sequence model, the SMILES strings were run through a pre-trained BERT tokenizer designed specifically to handle SMILES sequences that also padded the inputs to the length of the longest sequence in the training set: 282 tokens. These padding tokens were subsequently masked before input to the base layers of the sequence model. Care was taken to preserve the observation pairings of image representations and their corresponding tokenized representations by developing a table that mapped sequences to image file names and target values. Figure 1 is an example of one image, sequence, and target observation.

4 Methods

4.1 Image Model

Our model uses transfer learning from InceptionResNetV2 with ImageNet weights as a base and a dense network as the head. The head model includes a GlobalAveragePooling2D layer and a dense layer with softmax activation. The loss function is categorical crossentropy since the network is performing binary classification. Adam optimizer was chosen with default learning rate 0.01. Finally, the model is trained over three epochs and AUROC is the metric tracked. In all of these binary classification tasks we choose AUROC as our metric of interest because it gives us the most information about the quality of classification being made by our model. This is common practice in

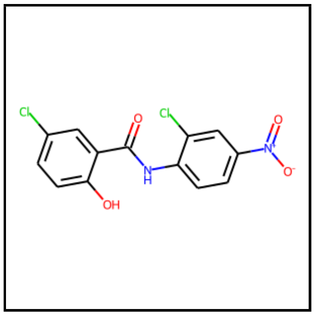

Figure 1: The training example corresponding to the SMILES sequence O=C(Nc1ccc([N+](=O)[O-])cc1Cl)c1cc(Cl)ccc1O (5-chloro-N-(2-chloro-4-nitrophenyl)-2-hydroxybenzamide) of the positive toxicity target class

molecular property prediction problems also due to the issue of heavily imbalanced datasets, which can make accuracy a less informative metric that by being heavily biased based towards the majority class.

4.2 Sequence Model

Our sequence model consists of three bidirectional LSTM layers with 25% dropout in between. We included bidirectional layers to allow the model full access to forward and backward information regarding a sequence at each time step. Following these recurrent layers was a one-dimensional convolutional layer with kernel size three and a subsequent global average pooling layer before the softmax-activated final dense layer. The final dense layer had a sigmoid activation and the pooling layer used the Glorot uniform kernel initializer. An important architectural change we implemented was the shift from an embedded one-hot representation of the features to a RoBERTa based tokenizer of the input sequences, ChemBERTa. The model was fit using 30 epochs and a batch size of 128. We included an early stopping callback with patience of four monitoring the validation AUROC and ReduceLROnPlateau as a learning rate scheduler.

4.3 Composite Model

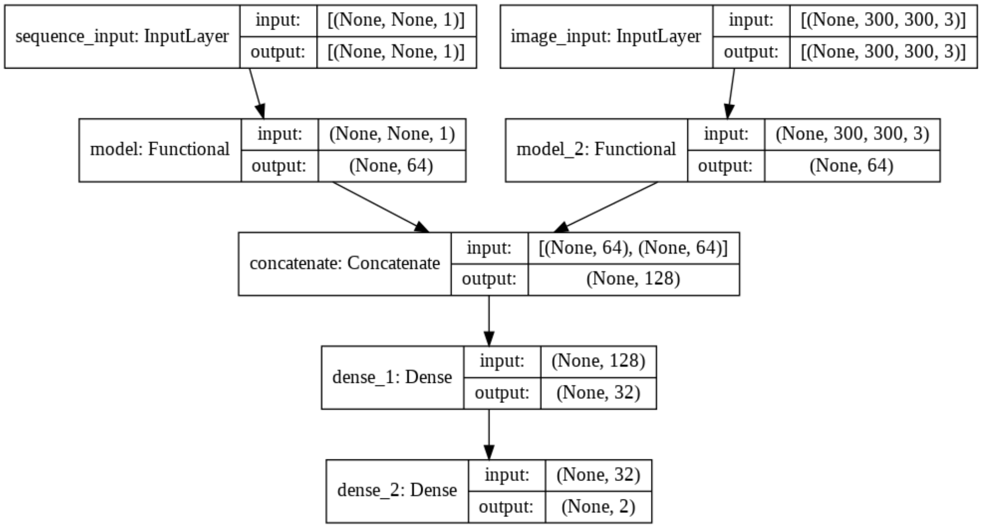

Figure 2: Diagram of composite model

The final step in our process was to create a composite model, shown in Figure 2, that combines the sequence and image feature embeddings by concatenating them and feeding them through two dense layers. This model takes in two separate inputs, the tokenized SMILES sequences and the

2-dimensional image representations of the SMILES compounds, and is still a binary classifier of SR-P53. To combine the two models, we used a global average pooling layer to achieve matching shapes and then stripped off the original models' final dense layers. In addition to concatenating the models there are two more dense layers to reduce the final output to shape (None, 1) for binary classification. Layer 'dense1' has relu activation function and l2 regularization to help normalize the model concatenation process by ensuring the weights of one model do not overshadow the other.

Our composite model parameters include an Adam optimizer with learning rate 0.005, binary crossentropy loss, and an AUROC evaluation metric. These parameters are in line with the individual sequence and image models. We included a learning rate scheduler 'ReduceLROnPlateau' monitoring validation AUROC. Early stopping was also added with a patience of four and validation AUROC monitor. The model was fit with 10 epochs and a batch size of 32.

5 Experiments/Results/Discussion

As we mentioned above, our sequence model achieved a peak validation AUROC of 0.691, and we show the model's training graphs in figure 3. However, this was not our best model so we continue with our other two approaches.

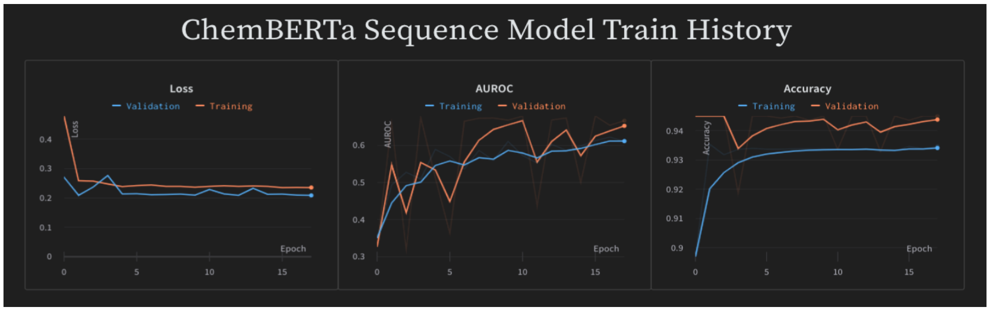

Figure 3: Sequence model training history

Our InceptionResNet-based model was able to achieve very promising results in relatively few epochs, training for only three, due to the advantage of initializing to the pretrained ImageNet weights. As shown in the training graphs in figure 4, this model achieved a peak validation AUROC of 0.787 and a test AUROC of 0.847:

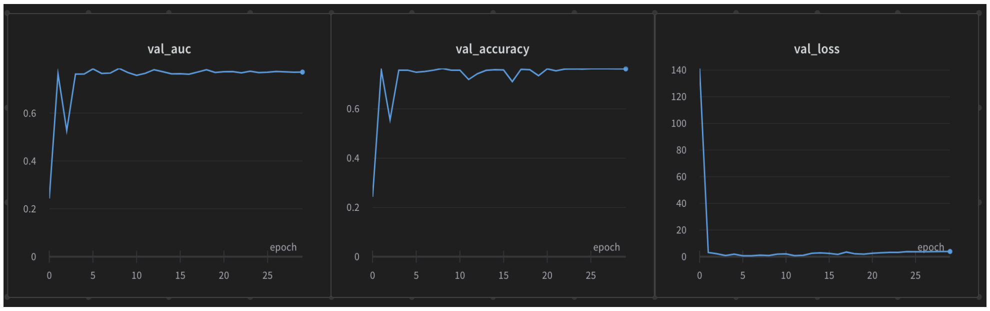

Figure 4: Image model training history

Finally, we trained the composite model for 10 epochs and it achieved a peak validation AUROC of 0.942 (see figure 5).

To properly evaluate the model, we passed the previously untouched test dataset into the model, yielding an AUROC score of 0.897. While these results support our hypothesis of the superior performance of a model that utilizes both feature extraction methods, this result came alongside many irregularities in our data pipeline. Given more time, we would need to verify these results more thoroughly.

Because our composite model's results are in need of further verification, our InceptionNet

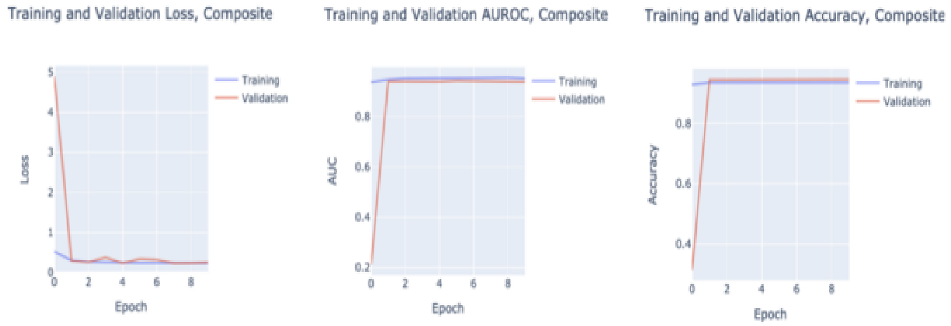

Figure 5: composite model training history

model is our most highly performing and reliable classifier so it is the focus of our final analysis. Below we see that model's ROC curve (figure 6) and a confusion matrix for a threshold probability of 0.99 (figure 7).

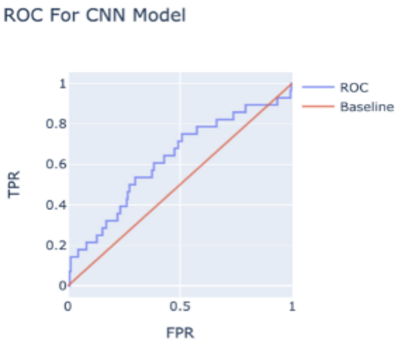

Figure 6: ROC curve for CNN model

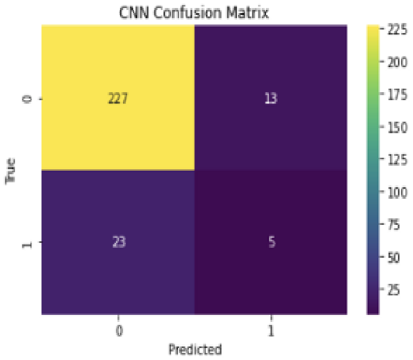

Figure 7: Confusion Matrix for CNN model, threshold of 0.99

6 Conclusion/Future Work

Our experiments here serve as an exploration of some of the state of the art methods in this type of problem and as a proof of concept for a particular approach in combining the different approaches' strengths. The motivation behind our project was to familiarize ourselves with the problem of molecular property prediction, and implement an intuitive extension in the form of our proposed composite model that feeds the feature embeddings from a sequence model and an image model into a multi-layer perceptron. While our composite model gave results that are highly suspect, all other aspects of our endeavour were successful as we implemented a novel approach that we did not find in the literature.

Future work continuing in this line of inquiry would of course further investigate the results that our final model achieved and potential sources of data leakage, which is especially hard to check given the ability for a compound to have multiple different SMILES representations.

A Contributions

All group members contributed equally to the overall project, sharing the research, coding, and writing responsibilities throughout the process.

B System Description

Appendix 1 – The python files for the described experiments are in the GitHub repository: `https://github.com/Molecular-Exploration/toxicity-classification`. The README describes the contents of each python script. The project was also chronicled in a series of blog posts on Medium which can be found at the following addresses: `https://molecularexploration.medium.com/smiles-toxicity-prediction-c674953922bc`, `https://molecularexploration.medium.com/smiles-toxicity-prediction-24e159acc067`, `https://molecularexploration.medium.com/smiles-toxicity-prediction-e75b27863da5`

References

[1] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules." [Online]. Available: https://pubs.acs.org/doi/10.1021/ci00057a005

[2] J.-C. Bourdon, S. Surget, and M. P. Khoury, "Uncovering the role of p53 splice variants in human malignancy: a clinical perspective," *OncoTargets and Therapy*, p. 57, 2013.

[3] "Tox21 data challenge 2014." [Online]. Available: https://tripod.nih.gov/tox21/challenge/data.jsp

[4] G. B. Goh, C. Siegel, A. Vishnu, N. O. Hodas, and N. Baker, "Chemception: A deep neural network with minimal chemistry knowledge matches the performance of expert-developed qsar/qspr models," 2017.

[5] M. Hirohara, Y. Saito, Y. Koda, K. Sato, and Y. Sakakibara, "Convolutional neural network based on SMILES representation of compounds for detecting chemical motif," *BMC Bioinformatics*, vol. 19, no. 19, p. 526, Dec. 2018. [Online]. Available: https://doi.org/10.1186/s12859-018-2523-5

[6] G. B. Goh, N. O. Hodas, C. Siegel, and A. Vishnu, "Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties," 2018.

[7] N. Mriganka and G. Subhasish, "Toxicity detection in drug candidates using simplified molecular-input line-entry system," *International Journal of Computer Applications*, vol. 175, no. 21, p. 1–4, Sep 2020. [Online]. Available: http://dx.doi.org/10.5120/ijca2020920695

[8] S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta: Large-scale self-supervised pretraining for molecular property prediction," 2020.

[9] J. Payne, M. Srouji, D. A. Yap, and V. Kosaraju, "Bert learns (and teaches) chemistry," 2020.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.