Computational Prediction of Molecular Toxicity

S. Jahangiri

Summary

Prediction of molecular toxicity with computational models is essential to reduce the cost of developing new industrial chemicals and new drugs. The purpose of this project is to use advanced computational methods to predict molecular toxicity based on easily-accessible structural information of molecules such as SMILES strings. The methods used here are machine learning models based on convolutional neural network (CNN), singular value decomposition (SVD) and support vector machine (SVM). The SMILES strings were converted to a two-dimensional matrix that was used as input to the CNN. In order to detect those parts of the molecules that have significant contributions to the molecular toxicity, the smiles strings were tokenized with a model inspired by the n-gram model used in computational linguistics. The important features detected in this way were then used to classify the molecules by applying a SVM. In order to use all the available toxicity data with respect to all targets for predicting the output of one specific target, the SVD method was applied to the matrix obtained by using all target results. Overall, the models implemented here have a fair performance on the test datasets and the analysis performed on their results provides guidelines for further improvements.

1. Dataset

The dataset investigated here contains 8014 molecules with toxicity information for 12 targets. About 17.3 % of the target data is unknown and about 7.5 % of the molecules in the know dataset are toxic. The latter indicates that the target data is unbalanced. The maximum length of the SMILES strings in this dataset is 342.

2. Models

2.1. *Convolutional Neural Network*

The CNN architecture contains two convolutional and two fully connected layers. Pooling layers were also added to each convolution layer and dropout layers were added to all convolution and dense layers in order to avoid overfitting by turning off some neocons randomly. The one-dimensional SMILES strings were converted to two-dimensional matrices by assigning a vector of length 27 to each symbol in the string. Each component of the vector is assigned to a special character and the presence of that character turns the assigned component to 1 while all other components are set to 0. A maximum length of 400 characters were assigned to the SMILES strings and for those vectors that have indices larger than the length of the SMILES string, all components were set to 0. In this way, each SMILES string is converted to a matrix of 400 x 27. These matrices were then used as input to the CNN model. The loss of the model for both the training and test sets is plotted for 10 epochs in Figure 1. The changes in the loss values indicate that the performance of the model is acceptable although running the model for a larger number of epochs is necessary for obtaining more accurate results.

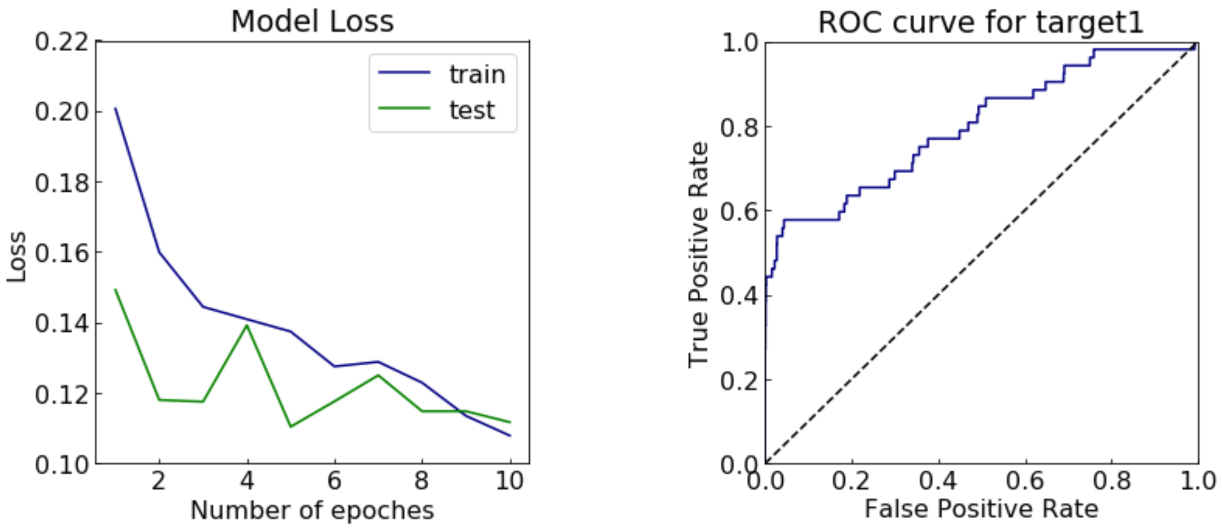


Figure 1. The model loss as a function of number of epochs obtained for the CNN model (left). The ROC curve of the CNN model for target1 (right).

2.2. *SMILES Sequence Feature Extraction*

In order to detect the important molecular features in the SMILES strings a contiguous sequence of n items, also known as n-grams [https://en.wikipedia.org/wiki/N-gram], were extracted from the strings. The n-grams used in this work contain all continues sequences of 2, 3, 4, 5 symbols. For each sequence, the frequency of occurring in all SMILES strings were computed and used to construct a matrix that is similar to the concept of bag-of-words in natural language processing. The mutual information between the frequency of occurrence for each n-gram and a target toxicity was estimated and 5000 n-grams with maximum importance were selected. A SVM model was then applied to classify the SMILES strings to toxic and non-toxic groups using the important n-grams as features to train the SVM classifier.

2.3. *Singular Value Decomposition*

This model is inspired by the recommendation systems in which SVD is used as a collaborator factoring algorithm to predict the interest of a user based on the information available from other users. The algorithm usually searches for users with the same rating patterns and uses this information to make predictions. In this work, SVD is applied to all target information to predict the unknown toxicities. The main difference between this model and the two previously-discussed models is the inclusion of potential correlation between all target data for predicting the molecular toxicities. This model is very useful when toxicity information is available for all targets but prediction is needed for only one specific target.

3. Results

3.1. *Prediction Performance*

The performance of the models implemented here was evaluated by comparing a range of metrics including ROC-AUC, accuracy, recall and f1-test. The ROC-AUC metric is frequently used for similar applications and this metric will be discussed here. In Figure 1, the ROC curve

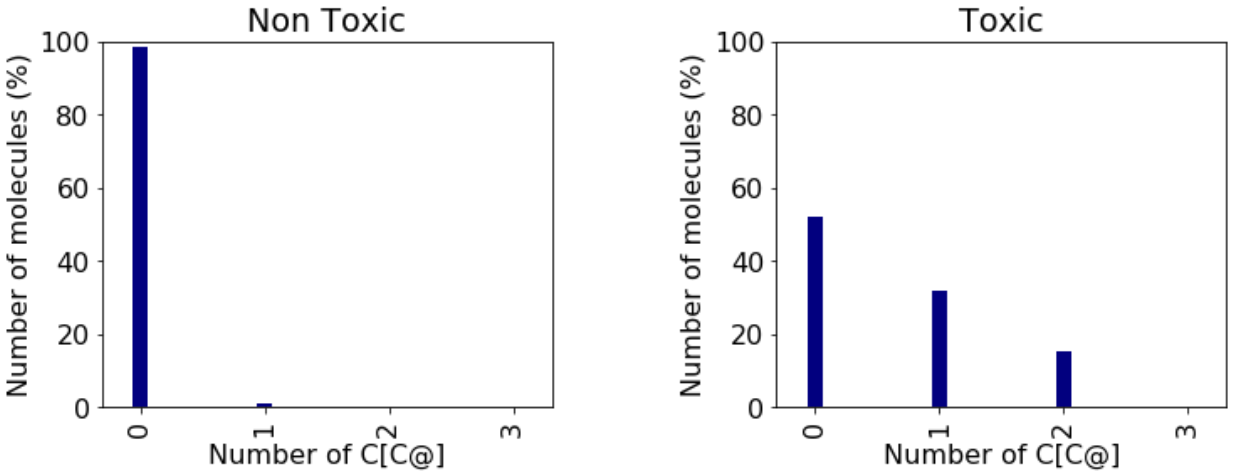
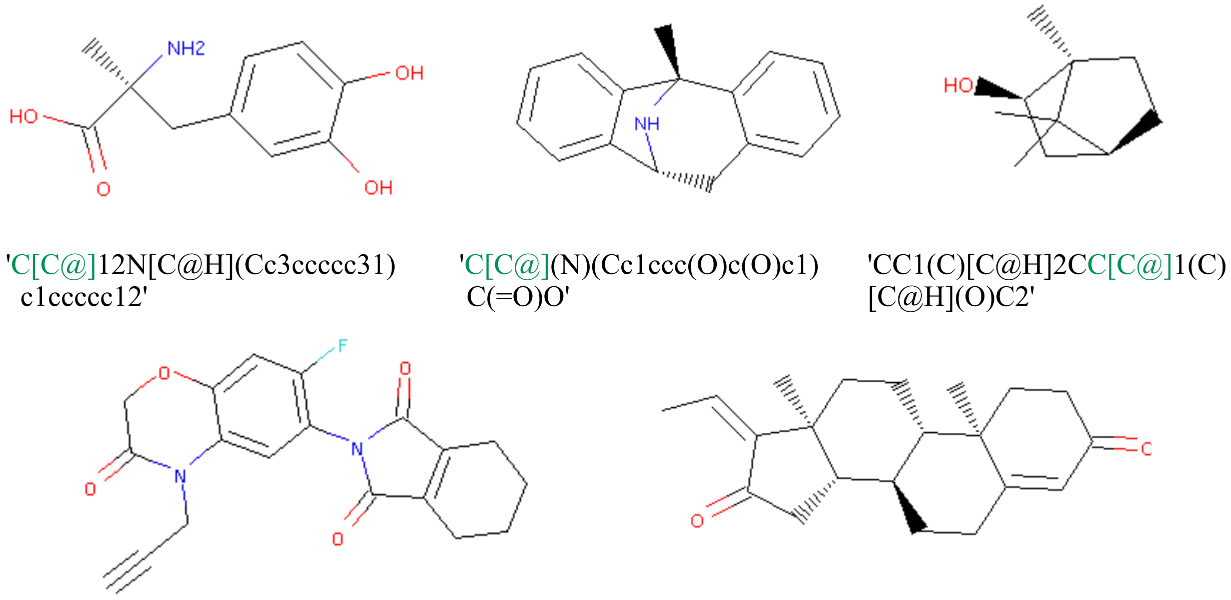


Figure 2. The number of C[C@] sequences in the non-toxic and toxic molecules of target1. This particular sequence is absent in majority of the non-toxic molecules while it appears 1 or 2 times in a large portion of the toxic molecules.

optioned from the CNN model applied to target1 is plotted. The AUC value for this target is 0.8 after 10 epochs which is acceptable but can be further improved by increasing the number of epochs. The SMILES strings used here do not contain information about the three-dimensional structure of the molecules which might be important for toxicity predictions. Supplementing the molecular fingerprints with such geometrical information and including extra information about the chemical environment of the atoms, such as number of H atoms connected to them, their hybridization, atomic charges, etc. might further improve the accuracy of the model. The CNN architecture can be also improved by including more layers and further optimizing the network parameters. These modifications can be investigated in future works. The SVM model applied to classify the molecules based on the important SNILES n-grams also has a AUC of 0.8 which is identical to that of the CNN model. This similarity is not surprising as both models use the SMILES strings for their predictions. The accuracy of the SVD model is determined by the root-mean-square error in predicting the activity probabilities for each molecule. The RMSE for the SVD model is about 25 % which is fairly acceptable for this model and means that the SVD probability for the molecular toxicity should be corrected with this uncertainty.

3.2. *Active Sequences*

The SMILES sequence feature extraction model implemented here provides information about the important parts of the molecules that are likely to be responsible for the molecular activity. For instance, among all the sequences investigated here, the C[C@] sequence containing a chiral centre is the longest sequence that was found to be significantly important for target1 activity. In figure 2 the frequency of this sequence in the non-toxic and toxic molecules is plotted. The distribution clearly indicates that a large number of toxic molecules contain this specific sequence while only a very small portion of non-toxic molecules contain C[C@]. Some of the toxic molecules containing the C[C@] sequence and their SMILES strings are presented in Figure 3. Another sequence that was found to be frequent in the toxic molecules is the C4=C sequence which indicates the presence of four rings with at least one double bond in one of the



'C[C@]12N[C@H](Cc3ccccc31)c1ccccc12'

'C[C@](N)(Cc1ccc(O)c(O)c1)C(=O)O'

'CC1(C)[C@H]2CC[C@]1(C)[C@H](O)C2'

'C#CCN1C(=O)COc2cc(F)c(N3C(=O)C4=C(CCCC4)C3=O)cc21'

'C#CCN1C(=O)COc2cc(F)c(N3C(=O)C4=C(CCCC4)C3=O)cc21'

Figure 3. Sample molecules containing the C[C@] and C4=C sequences that are toxic with respect to target1.

rings. The two shortest molecules that contain this sequence are shown in Figure 3. Presence of these two sequences in the toxic molecules indicates the importance of structural features in determining the toxicity of the molecules compared to the presence of specific heteroatoms or functional groups. This finding and the the chemical significance of the C[C@] and C4=C sequences should be investigated in more details.

4. Conclusions and Future Works

In this work, machine learning models based on convolutional neural network, singular value decomposition and support vector machines have been developed to predict the toxicity of molecules based on their SMILES strings. The predictions of these three models can be combined together to obtain unknown molecular toxicities based on majority of votes or any desired ensemble model. The CNN and SVM models have an AUC of ~ 0.8 and the SVD model has a RMSE of ~ 0.25, indicate fair performances of the models. The accuracy of the models can be further improved by tuning the model parameters and increasing the training iterations. In order to detect those parts of the molecules that are highly correlated with the molecular toxicity, a model inspired by the concept of n-grams in computational linguistics was used. The accuracy of this model can be increased by using longer sequences of symbols in the analysis. In order to use the SMILES strings as input for the CNN model, a method was implemented to convert the SMILES strings to two-dimensional matrices. This approach can be also improved by increasing the amount of information included in the expanded matrices. A more robust approach for increasing the accuracy of predictions in such models is to use information about three dimensional structure of the molecules, such as Coulomb matrices, and also include quantum chemical information such as electron density distribution and partial atomic charges.