Molecule classification using 2D Convolutional Neural Network and Ensembles Methods

Istvan Lakatos[a], Andras Hajdu[a], Balazs Harangi[a]

[a]University of Debrecen, Faculty of Informatics
POB 400, 4002 Debrecen, Hungary

Abstract

In the field of pharmaceutical industry there are many researches, developments and manufacturing which have aims to find new drugs or molecules[4]. One of the first steps in the drug discovery process is to test already well known compounds to see their effects and reactions. The question arises as to whether this expensive and time-consuming procedure can be simplified and enhanced by using computer simulations or estimates to perform better filtering than classical high-throughput screening (HTS) [2].

In this paper, we investigate how we can apply image classification based approaches like convolutional neural networks (CNNs) to predict toxicity or activity of different unknown molecules based on their structural information and their known components. One of the most cited and used dataset [1, 3, 6] related to molecule toxicity prediction is the Tox21 Data Challenge Molecule Database [5]. Since the Tox21 dataset contains the structural information of compounds in the commonly used SMILES strings, so we needed to develop a method to generate images and classify them based on this format. We applied a 3D molecule visualization tool to generate 2D mapping of the 3D model like six faces of an object then trained our CNN to classify each image separately. The final label of the molecule is derived from an ensemble of these six class labels.

Our experimental results show that the using of CNNs to molecule toxicity prediction problems is a promising approach. For the ensembles of class labels of the 6 faces, we have evaluated both majority voting and probability sum as an aggregation function and got ROC-AUC values of 0.7733 and 0.8089 respectively.

Keywords: molecule classification, convolutional neural network

Acknowledgements. Research was supported in part by the Janos Bolyai Research Scholarship of the Hungarian Academy of Sciences and the projects EFOP-3.6.2-16-2017-00015 supported by the European Union, co-finances by the European Social Fund.

References

[1] Ahmed Abdelaziz et al. "Consensus Modeling for HTS Assays Using In silico Descriptors Calculates the Best Balanced Accuracy in Tox21 Challenge". In: *Frontiers in Environmental Science* 4 (2016), p. 2. DOI: 10.3389/fenvs.2016.00002.

[2] Jayme L Dahlin and Michael A Walters. "The essential roles of chemistry in high-throughput screening triage". In: *Future Med Chem* 6 (July 2014), pp. 1265–190. DOI: 10.4155/fmc.14.60.

[3] Andreas Mayr et al. "DeepTox: Toxicity Prediction using Deep Learning". In: *Frontiers in Environmental Science* 3 (2016), p. 80. DOI: 10.3389/fenvs.2015.00080.

[4] Steven Paul et al. "How to Improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge". In: *Nature reviews. Drug discovery* 9 (Feb. 2010), pp. 203–14. DOI: 10.1038/nrd3078.

[5] *Tox21 Data Challenge 2014*. URL: https://tripod.nih.gov/tox21/challenge/index.jsp.

[6] Zhenqin Wu et al. "MoleculeNet: A Benchmark for Molecular Machine Learning". In: *Chemical Science* 9 (Mar. 2017). DOI: 10.1039/C7SC02664A.