Jennifer HEMMERICH

Thesis Supervisor: Gerhard ECKER, Dept. of Pharmaceutical Chemistry, University of Vienna.

Associated MolTag student, 3rd Funding Period

Defense: 26.06.2020 (online)

Thesis title: Deep learning for toxicity prediction.

For many years, machine learning has been an integral part of in silico toxicology. Despite a plethora of research efforts, only a very limited number of models are available to predict complex apical toxicity endpoints such as organ toxicities. The Merck Kaggle Competition and the Tox21 challenge highlighted the suitability of deep learning for bioactivity and toxicity predictions. Thus, deep learning could give rise to novel possibilities in in silico toxicology. However, toxicological datasets are often small and imbalanced. While small datasets might not contain enough data for deep learning to be superior to traditional machine learning, imbalanced datasets can lead to models ignoring the minority classes. Thus, to utilize deep learning in in silico toxicology, it is essential to gather novel datasets and to address imbalances in them.

This thesis aims to, firstly, gain insights into datasets by rigorous data analysis, and secondly, to introduce conformational oversampling (COVER) as a new method to balance datasets by using 3D conformations.

To generate novel datasets, we introduce a chemical structure standardisation workflow which facilitates merging of datasets from different sources and enables the user to detect duplicates via the generated InChIKey. In our first study this workflow was used to support an experimental study which investigated drug uptake by SLC transporters. By comparing the screening compound library with the drug chemical space represented by DrugBank, we showed that the tested compounds are representative of this space and do not show any bias towards specific physicochemical properties. Thus, SLC related drug uptake was independent from physicochemical properties and is therefore most likely the rule and not the exception. In our second study we compiled a dataset for mitochondrial toxicity using our workflow. Through data analysis we identified important physicochemical properties for mitochondrial toxicity. Subsequently we trained different types of machine learning and deep learning models using our novel dataset. Using known and novel structural alerts we gained insights into possible mechanisms of toxicity for positively predicted molecules.

The last two studies shift the focus from gathering novel datasets towards training of neural networks with imbalanced data. In the first study we introduced conformational oversampling (COVER), which uses 3D conformations to balance a dataset. We demonstrated that the 3D information of molecules is sufficient to balance datasets by oversampling. Applying COVER to the Tox21 dataset greatly increased the model predictivity and was comparable to oversampling with SMOTE. In our last study, we explored the application of COVER from 3D descriptor-based learning to image-based learning. Both studies showed that COVER can be applied successfully to different training inputs for neural networks.

Conclusively, this thesis highlights the importance of data analysis, by, firstly, supporting an experimental study and, secondly, increasing the insight into machine learning predictions. In addition, we could show that COVER is a viable method to oversample datasets utilizing 3D descriptors and images.