Transformer-based methods for novel active chemical compounds

During the last 10 years, deep learning has taken off and been successfully applied to many scientific fields that generate knowledge driven on data. Currently, there is an ongoing research on how to apply deep learning models to computer-aided drug design (CADD). Basic CADD problems include *in silico* predicting properties of chemical compounds and generating new molecules with desired properties. Designing new drugs is a long and complicated process that could lasts even 10 years and cost up to 10 billion dollars. However, it can be aided by machine learning algorithms, making the whole procedure quicker and cheaper.

The aim of our project is to provide transformer-based deep learning model for predicting the activity of compounds. Our model will be based on the graph representation of the molecule and will augment the self-attention module of the network with information about the neighbourhood of atoms, the distances between them and the features of bonds that occur between atoms. We will also compare the influence of various atom representations on the performance of graph-based neural networks.

Based on this model we will conduct the graph-to-graph translation of molecules in both supervised and reinforcement learning manner. The resulted model could be helpful in the lead optimization phase of the drug design process, in which researchers analyse a number of good and interesting molecules (named *leads*) but want to further optimize some of their properties.

We will use data from publicly available databases, such as CHEMBL or ZINC as well as the data collected as a result of collaboration of our group with the group of prof. A. Bojarski – we currently have access to a dataset of over 300000 results from docking experiments. The data will need preprocessing to be suitable for our methods. We will mainly focus our attention on predicting and optimizing the physico-chemical features of molecules (such as solubility), metabolic stability and activity towards the given proteins.

Our approach can potentially reduce the cost and time of producing new drugs. It could be used during the virtual screening process to filter out undesirable molecules, which would significantly reduce the expenditure during laboratory tests. The molecules generated by our models could be then synthesised and tested *in-vitro* for possessing the desirable features or could guide medicinal chemists to some new, nonobvious substitutions, which could result in new drugs.

All of these may result in making new drugs cheaper and more accessible.