Philipps Universität Marburg

Faculty of Pharmacy
Department of Pharmaceutical Chemistry
Prof. Dr. Peter Kolb

Tel.: +49 6421 28 25908
Fax: +49 6421 28 28994
E-Mail: peter.kolb@uni-marburg.de
Sec.: Ms. Lydia Hartleben
Tel.: +49 6421 28 21313
E-Mail: lydia.hartleben@uni-marburg.de
Address: Marbacher Weg 6
35032 Marburg
Web: www.kolblab.org

Marburg, June 13, 2018

prof. Zbigniew Madeja, PhD, DSc
Faculty of Biochemistry, Biophysics and Biotechnology
Jagiellonian University
Gronostajowa 7
30-387 Krakow

Review of the doctoral thesis of Mr. Krzysztof Rataj, MSc.

Dear prof. Madeja,
Dear Council Members,

with this letter, it is my pleasure to provide my opinion about the doctoral thesis of Mr. Krzysztof Rataj. His submitted work is entitled "Neighboring Substructures Fingerprint as a New Method of Compound Representation and Its Application in Screening for Novel Selective Compounds for Selected Serotonin Receptors". It deals with the development of a novel kind of fingerprint method in order to assess the similarity between two molecules.

Chemical similarity is always in the eye of the beholder, but when it comes to pharmacology, similarity is defined by which derivatives still bind to a target protein in question. As such, methods that can discriminate between ligands and non-binding molecules are valuable tools in computer-aided drug discovery. Such methods always have to find a balance between speed and accuracy. Speed, because the methods need to be able to process many millions of compounds. Accuracy, because they need to provide the right results. Given these constraints, fingerprint methods, which are based on the two-dimensional structure of a molecule, are very fast while allowing for a decent discrimination between ligands and non-binding molecules, and are thus attractive tools. Krzysztof Rataj has developed a novel way of constructing such fingerprints by taking into account neighboring fragments. The upside of this approach is that it can be applied to various substructure-based fingerprints. He then tests his development on a number of different G protein-coupled receptors, in particular the serotonin receptors 5-HT1B and 5-HT2B.

Mr. Rataj start his thesis with an introduction into the topics of molecular representations in chemoinformatics and what the different methods are at our disposal for virtual screening applications. A particular emphasis is placed on the aspect of selectivity, i.e. the differential affinity of one ligand for

more than one target. This section has textbook-like qualities, in that the explanations are all reading very well and the figures are well-crafted and support the text.

After stating the aim of this thesis, chapter 3 is devoted to the methodology. Mr. Rataj explains how the NSFP, the "Neighboring Substructures Fingerprint" is constructed and applied. He also describes the ways and criteria he used in order to extract the molecular test sets from different sources. Furthermore, to develop the classifiers to predict selectivity of ligands towards 5-HT1B and 5-HT2B, he uses various machine learning techniques, and also explains those briefly. In the end, the most successful method turned out to be Extreme Entropy Machines, EEMs.

However, he does not stop at *post hoc* predictions, but rather goes on to also predict several novel compounds through a combination of fingerprint-based classification and docking. These compounds are then assayed for their selectivity, i.e. the correctness of the prediction. All in all, 29 compounds were acquired and tested, and these are described in the results section. Unfortunately, in the first round only one compound was identified that showed decent binding and a rather amazing selectivity factor of 10'000 for the 5-HT2B. Krzysztof Rataj tried to improve these results by including water molecules during the docking studies, as some of them are quite stable in the existing X-ray structures. Yet, also this increased accuracy did not bear the desired fruit (only a 10-fold selective ligand was found) and Mr. Rataj went on trying to post-filter the results using a structure-interaction fingerprint, a concept which had been developed earlier in prof. Bojarski's lab. A testament to the complexity of the problem of selectivity, also this further modification did not yield the desired ligands.

Mr. Rataj conludes his thesis with chapters on Conclusions and a Discussion. Both parts read very well and show that he is a scientist who critically assesses his results and does not try to oversell them. He also displays the ability to connect his findings to the bigger picture of drug design and selectivity.

It is a particular strength of the thesis that it is supported by informative and clear visual images. These make reading and understanding the thesis quite easy. Especially the flowcharts are a great help, as the research consists of many different strands, and it could be not easy to follow for an outsider. I also liked that Mr. Rataj developed his own nomenclature of the different models and predictors based on sub- and superscripts, as this leads to a good density of information. As I mentioned earlier, the fact that no ligands with the desired properties were identified does not necessarily reflect a deficit on the part of the candidate, but can rather be seen as an effect of the challenging target combination that he picked. Nonetheless, his development, the NSFP, respresents a valuable addition to the tool box of cheminformatics, and the combination of this technology with machine learning in order to construct filters and classifiers is quite novel.

There are a few details which I think could have been solved better or evaluated more critically. As an example, in chapter 4.1, on the retrospective evaluation of the different fingerprint/machine learning combinations, the differences between the different versions are rather small, but are described as clear distinctions. Morover, when describing the criteria for compound selection (page 35) no units are given, and I am not quite sure that the formula $K_i = \frac{IC_{50}}{2}$ holds true at all times. This does not invalidate the results, as it would affect all compounds equally, but is a bit of an imprecision. Lastly, Table 9 and following have a typo in the 4th column. One can see from these criticisms, however, that

they mainly concern typography, and do not touch the substance of the work.

Hence, I would recommend to the Scientific Council of the Faculty of Biochemistry, Biophysics and Biotechnology of the Jagiellonian University to accept the submitted Ph.D. thesis.

Sincerely,

Peter Kolb
Heisenberg Professor