mta ttk

Prof. GYÖRGY M. KESERŰ
scientific advisor
MEDICINAL CHEMISTRY RESERACH GROUP
RESEARCH CENTRE FOR NATURAL SCIENCES
HUNGARIAN ACADEMY OF SCIENCES

H-1117 BUDAPEST, MAGYAR TUDÓSOK KRT. 2.
ADDRESS: H-1519 BUDAPEST, P.O.BOX 286.
PHONE: +361-382-6821
E-MAIL: KESERU.GYORGY@TTK.MTA.HU
WWW.TTK.MTA.HU

August 25, 2016

External examination report on the PhD thesis by

Sabina Podlewska

Development of machine learning-based tools for computer-aided drug design

Sabina Podlewska's thesis focuses to a relatively new field of drug design related to the development and application of machine learning techniques for virtual screening. The computational efficiency and different performance metrics validate machine learning tools in drug discovery applications from which virtual screening of large compound databases is of utmost importance. Given the demonstrated success of virtual screening in the identification of new chemotypes for therapeutically relevant protein targets these approaches represent a cost-efficient alternative to resource intensive physical screening campaigns. Based on these considerations I acknowledge the relevance of the topic of the thesis. Results achieved during the PhD work could contribute significantly to this emerging area of virtual screening technologies.

The thesis is based on 7 research papers that are all connected to the development and application of machine learning tools for drug discovery problems. In addition the candidate provided a 84 page long summary of the thesis. The summary starts with a short introduction and followed by a chapter with the aim of the program and the list of the underlying publications. The main part of the summary describes the results reported in these publications. Carefully examining the publications I identified the major achievements of the author as follows:

1. The author investigated several parameters of machine learning methods that impact their performance in virtual screening settings. These studies included investigations on the influence of the size and composition of inactive subsets and also the hyperparameters of support vector machines. The coverage of the chemical space was identified as a key parameter in the performance of machine learning tools. In addition, it was found that the 1:9-1:10 inactive ratio provided the best separation of actives from inactives. Finally, the candidate pointed out the importance of optimized SVM parameters that should be optimized through Bayesian optimization to achieve superior performance in virtual screening (papers 1, 2 and 3).
2. The author tested the performance of extremely randomized machine learning tools in activity predictions. These investigations included Extreme Entropy Machine and Extremely Randomized Trees methods and revealed that these techniques perform somewhat better than conventional machine learning tools. More importantly she demonstrated that they are computationally less expensive and easier to optimize

that makes these tools promising in the virtual screening of large databases (paper 4).

3. The author developed a new methodology for the treatment of data inconsistencies often detected in drug discovery datasets. To achieve this goal she modified the SVM algorithm introducing the class-weighting and variance weighting techniques. The study showed that incorporating the uncertainty of biological experiments might improve the performance of machine learning based classifications, specifically the balanced accuracy of selecting actives (paper 5).
4. The author developed a highly efficient multi-step protocol for the automatic evaluation of docking results. The developed methodology is a machine learning technology that uses structural interaction fingerprints and spectrophores for the post-processing of results obtained in docking-based virtual screening. The protocol has been tested to identify serotonin 5HT6 and 5HT7 ligands. The new evaluation protocol improved the discrimination of active and inactive compounds significantly and the author achieved almost perfect classification (paper 6).
5. The candidate developed a fingerprint-based machine learning application for the identification of new 5HT6 ligands. It has been shown that a fingerprint based consensus approach with the sequential minimal optimization method enriched the actives in the screened database significantly. 23 compounds from the hit list of the subsequent docking screen were ordered and two of them showed submicromolar affinity and promising selectivity towards 5HT6 receptors. Interestingly, both compounds are non-basic and represent novel chemotypes (paper 7).

The publications follow a logical progression of ideas from the development and optimization of machine learning tools to their several relevant applications including the treatment of data inconsistencies, the development of an automated evaluation protocol for docking and a real life virtual screening application. All chapters are published in respected, peer-reviewed journals. Investigating the details of these works I conclude that all of the five achievements are based on new and true scientific results and make a strong basis of the PhD dissertation. In depth analysis of the candidate's contribution to these results revealed that Sabina Podlewska contributed more than 50% to 4 papers, gave significant contribution (larger than 25%) to the other two. Based on this evaluation I conclude that the work presented in the present thesis is the major contribution of the candidate.

Specific comments and questions:

Paper 1:
Inactive subset generation was investigated by three different databases including ZINC, MDDR and DUD. These databases, however, are very different in size and composition. ZINC contains all the commercially available compounds, MDDR is a database of compounds with measured biological activity on particular targets and DUD contains compounds similar to

bioactive compounds. The authors kept the number of inactive molecules constant for selecting inactives from these databases. Selecting true inactives from MDDR resulted in zero probability of actives for this set. In the case of ZINC and DUD the compounds assumed to be inactive that gives larger probability of actives as compared to the MDDR-based set of inactives. Furthermore, ZINC and DUD are different in size and contain actives with different probability (random for ZINC and higher for decoys of DUD). Selecting the same number of compounds from ZINC and DUD therefore should yield sets with different probability of actives. The best performance was obtained when inactives were randomly picked from ZINC. This is not unexpected since the ZINC-base set has the lowest probability of actives and the enrichment of actives in the inactive subset is smaller with random selection as compared to diverse selection that covers all the chemotypes. I think it would be reasonable to perform this analysis on publically available HTS datasets that typically contain high number experimental inactives. In the case of success the methodology would be useful for selecting compounds in sequential or iterative screening paradigms.

Paper 2:
The author tested 5 different machine learning method with CDK and MACCS fingerprints. Interestingly, in the case of SMO CDK outperforms MACCS but this is not the case for the other methods (Fig. 1). What is the special feature of the SMO method that could explain this? It was found that Naïve Bayesian was insensitive to active to inactive ratio. How the author could rationalize this? Table 1 shows the evaluated parameters with the five machine learning methods and the two fingerprints as applied for 5HT1A, HIV PR and Metalloproteinase. Largest differences were obtained with training sets containing the highest and lowest number of inactives. The two most significant differences were found with SMO/CDK and RF/CDK models. What is the special feature of these methods that could explain this? Furthermore, machine learning methods provided larger differences in performance with CDK as compared to MACCS. How can we rationalize this? Data in this table suggest that machine learning methods have more significant impact on the performance than the fingerprints and the target. I think this is a critical question that would be interesting to investigate on further examples.

Paper 3:
The author concluded that when a model of strong predictive power is needed in a relatively short time, then, the random search should be used; and when the strongest model is desired and the computational time is not a limitation, the Bayesian approach should be used. However, from Fig. 2 it seems that Bayesian optimization outperforms random optimization and furthermore both are fast enough. Is it reasonable to use random optimization at all? When applying extremely randomized methods (EEM and ET) the improvement in BAC was very much limited and obviously non-significant. Except form their computational efficiency what is the advantage that could validate these methods?

Paper 4:
I see the application of extremely randomized machine learning tools for compound selection interesting since these methods have been scarcely tested in this paradigm. Here, however, ET and RF probabilities seem to be virtually the same. There is more difference between EEM and SVM but again, improvement in BAC is less than 1%. In my opinion these results could not validate the application of extremely randomized tools, however, further studies might specific advantages.

Paper 5:
I think that the treatment of data inconsistency is one of the key points when compiling training sets for predictive CADD tools, especially for ligand based approaches. This paper can be considered as an important contribution to the field, however achieving more significant improvement in BAC needs more research. Here ΔBAC is around 10% in the best.

Paper 6:
In my opinion the most important result of this PhD thesis is the development of the multi-step protocol for the evaluation of docking results. The results are impressive and suggest that the machine learning based method with SIFts and Spectrophores should be useful as a general protocol. Did the author test the protocol on other targets?

Paper 7:
The author described a fingerprint based machine learning protocol for the identification of new 5HT6 ligands. In this case virtual screening has been started by the machine learning approach and final compounds were selected by docking. Having the automated evaluation protocol in hand why the author did not use this methodology for the selection of compounds?

In summary, Sabina Podlewska investigated a number of machine learning approaches in virtual screening settings. First she examined the impact of different parameters such as the size and composition of inactives and the optimization of SVM parameters on the performance of these methods. Next she compared the performance of extremely randomized methods to conventional techniques. Then she applied these tools to relevant drug design challenges from the treatment of data inconsistencies, through the development of an automated evaluation protocol for docking results, to the fingerprint based identification of new 5HT6 chemotypes. Based on these results I strongly support the acceptance of this thesis and warmly suggest the Committee issuing the PhD degree to Sabina Podlewska.

Given the Jagellonian University rules of granting honours for PhD thesis I also investigated the PhD exam mark (the candidate passed the exam with excellent degree - level 5), the summ of PhD papers impact factors (the candidate IF is 24.635), the time of PhD studies (the candidate completed the thesis within 4 years). The excellent work reported in these papers together with fulfilling these criteria nominates Sabina Podlewska for the honour and therefore I recommend the thesis for distinctions for the Committee.

Prof. György M Keserű PhD DSc

(RCNS, Hungary)