Review of the dissertation thesis

Author: **Mgr. Petr Škoda**
Title: **Representation of chemical compounds and its utilization in similarity search**
Opponent: **Martin Modrák, PhD.**

The thesis combines an extensive series of explorations in cheminformatics, related to the task of virtual screening - finding candidate molecules that are worth testing for interactions with a known target molecule within a large molecular database.

Contributions to Virtual Screening Infrastructure

The initial chapters introduce several (presumably) minor contributions of the author to the area: some adjustments to the Molpher software for finding “intermediate” molecules between multiple active molecules, the ViSeT pipelining tool, and the Prankweb server useful for ligand binding site prediction. Here ViSeT appears the weakest as I did not understand in what ways it is different from many other pipelining tools broadly available (Snakemake, Nextflow, Chipster, Galaxy, ...). I therefore suspect that integrating cheminformatics tools with existing pipelining software might have been preferable.

The Prankweb server looks and works nice, albeit it is also one of many similar tools, despite having some very good features - especially the efficient visualisation with LiteMol. Finally, Molpher and its variant for molecular scaffolds is in my opinion a very original approach and an interesting addition to the cheminformatics toolset.

Similarly to ViSeT, the final part, Chapter 7, where a novel benchmarking environment is presented, is in my view a weaker contribution as I would guess improving and extending an existing benchmarking platform would be more beneficial for the community in the long term than creating one more new platform that will likely end up abandoned.

Chapter 7 - as well as the rest of the thesis - relies solely on the area under curve (AUC) metric for evaluation. While this is apparently a standard in the field, I believe it may make the evaluations both fragile and less useful for practical applications. The robustness of the findings could have been better evaluated if more metrics were reported, and the extent to which the results differ between metrics discussed. If drug discovery is the ultimate goal, it also might be sensible to assign utilities to all four screening outcomes (true/false x positive/negative) and see which tools give the highest expected utility for several plausible real-world utility assignments. Finally, the reported AUC is averaged across multiple train/test splits and multiple screening tasks, without reporting any measure of its variability/uncertainty. I suspect that when variability would be reported, a large fraction of the reported differences between methods would prove to

be smaller than the variability and thus unreliable. This is especially relevant as the author acknowledges that choosing different train/test splits of data can greatly influence the observed AUC.

A nice point of Chapter 7 was the systematic search for applications of virtual screening and its benchmarks.

Fingerprints

The most extensive part of the contribution is Chapter 6 describing four novel molecular fingerprints. The design of the fingerprints is driven primarily by engineering considerations and a lot of the components of the methods appear to be more “clever hacks” than “elegant algorithms”, but I consider this a good thing.

Of the four techniques, I am most impressed with the pharmacophore approach which shows the most convincing evaluation results. On the other end, I find the target-oriented generic fingerprint (VectorFp) method least appealing as it appears that the number of tested VectorFp parametrizations is vastly larger than the amount of benchmarking data sets and therefore taking only the “best” VectorFp variant might be overfitting. Despite possible overfitting, the reported performance gains are modest.

Overall, the evaluation of the fingerprints is only mildly convincing. Interestingly a lot of the issues I find with the evaluation methodology for the individual fingerprints is discussed as failures of previous benchmarks in Chapter 7, so at least the author was aware of them. Most notably, baselines for comparison are different for each fingerprint. Also quite often a best version of a method across multiple possibilities is chosen using the same dataset that is then used for comparison to baseline, possibly overfitting and likely overestimating performance. Intriguingly, the author also chose not to include his own fingerprints in the final benchmark in Chapter 7 which would have alleviated some of the issues.

The absence of any quantification of the variability of the reported average AUC makes it harder to judge, but some of the discussions seem to be just chasing fluctuations that might as well be noise (e.g. page 90 “against the targets 548, 600, the aromaticity provides the best AUC performance”) I would advise the candidate to always consider that variability might not be meaningful, especially when the differences are not really big as is the case here. In a similar vein, the fact that different versions of the VectorFp fingerprint are best for different tasks does not necessarily imply that we can - even in principle - choose a good VectorFp version for a given task. We would expect part of the variability to be unpredictable.

Some of the applications of statistics/machine learning in the work appear to no fit the task very well and little justification is given to convince me otherwise. For example the forced binning of continuous features for use with the Naive Bayes classifier indicates that logistic regression or

other approach that works with continuous features seamlessly might have been a more natural approach.

Formal Aspects

A major issue I have is that it is hard for me to distinguish what was the actual scientific contribution of the candidate in various software/methods presented in the thesis. Since many of the papers describing the software/methods do not list him as the first author, I believe it was not 100%. It is therefore unclear which parts of the text describe someone else's work (when it is a necessary introduction to the actual contribution of the candidate) and what is the actual contribution. Explicitly delineating this is in my view important.

I was also slightly confused that sections marked as taken from the author's publications did not mention from which manuscript they were taken, which made it harder when I wanted to access the full publication for context. Personally, I would advise the candidate to make author versions of their publications available on a personal website as is usually allowed by the publisher. Several of them were outside my institutional subscriptions and I had trouble accessing them.

The process of inserting parts of manuscripts into the thesis also left some errors (references to non-existent sections etc.). Some tables - which were likely optimized for space in the original manuscript - appeared crowded or otherwise inefficient in the thesis.

The writing itself is of acceptable quality and mostly easy to follow, albeit with frequent minor grammatical mistakes. Sometimes terms are explained only after they've been used extensively (e.g., AUC is first defined in section 7.2)

I am glad that some of the code underlying the work was easy to access. Unfortunately, for large portions of the thesis, including the whole chapter on fingerprints, the code is not publicly available, making the research less useful to the community and harder to evaluate.

Conclusion

The thesis is wide in scope and documents considerable research effort. While my impression from the text is that most - if not all - of the approaches presented ultimately failed to improve the state of the art noticeably, I believe the underlying scholarship is solid and the lack of breakthrough is due to the inherent difficulty of the problem and not for lack of trying. The topic of the thesis is timely and progress in virtual screening could plausibly have important positive impact on real-world drug discovery and research.

Despite the flaws I mentioned, the thesis contains worthwhile contributions to the development of cheminformatics. The thesis shows that the candidate is capable of independent creative scientific work and I recommend the thesis for defence.