Thesis Proposal: [provisional title] Retrosynthesis Development and Implementation in Molecular Generators

Julius Cathalina

September 27, 2021

[2] - De novo drug design: generate novel molecular structures from building blocks without a priori relationships - Use deep reinforcement learning techniques - Create better leads for drug candidates, current pipeline very expensive and (relatively) inefficient (5 in 5000) - Techniques like HTS have helped a lot but... - This samples a very small sample of the large chemical space (106̂0 order) - selection, design, synth very hard - CADD can be improved... - as has been shown by methods such as QSAR and pharmacophore modelling - data is increasing and is becoming cheaper to collect on compounds - but when it comes to pharmaceuticals, the therapeutic targets are not trivial (complex signalling pathways, toxicity etc.) - DNDD to design chemical entities that fit a certain constraint... - We can explore more of the chemical space using this technique - however, synthetic accessibility is a challenge with the novel compounds - we have different ways to approach DNDD... - structure based, ligand based, EA... - constraints can be e.g.: solubility (range), toxicity, chemical groups etc. - DRL approaches in DNDD usually are comprised of a generator and an RL agent that interacts with the molecules to modify them to obtain desirable properties. - The molecules can be represented in a variety of ways, such as graphs, or SMILES (char sequences) and this of course has to be compatible with the type of network that we will use. - we train the NN On tokens of pre-existing data (e.g. known bio-active compounds from ChEMBL) for specific biological targets. - After learning the probability distribution of tokens on the pre-existing data, we can now use this knowledge (during molecule generation) which token is optimal next based on the vocab, and the sequence generated previously - Many

examples of DRL being used for DNDD have been published, backed by architectures such as RNNs, CNNs, GANs etc. - TODO: REFER TO SOME OF THESE OTHER EXAMPLES, AND MENTION THAT OURS IS BASED ON DRUGEX V2 - We must also take into account that we need to determine evaluation criteria to analyse our results. - Namely, there does not seem to be a standard approach to evaluate these generated molecular compounds... - What are the important things that we should test then..? - We must look at the diversity of the generated compounds, e.g. are they far enough from the mean of the training distribution? Are all the generated sequences similar? (check using levenshtein distance if SMILES, or Tanimoto/Dice if using fingerprints). - How do we select "good" evaluation threshold for these new generated drugs? - we could apply ADMET and QSAR approaches prior to synthesis and in vitro testing to assess the relevance of the designed molecules (Muratov et al.) to attempt to be as objective in our assessment as possible. - The crux of this thesis, aside from experimenting with the way in which the molecules are generated, is the incorporation of methods that ensure that the synthetic feasibility of the generated molecules are prioritized. we can use things such as SA-score, and then compare these results by using RA-score developed by AstraZeneca. - Certain approaches already implement the ability to generate realistic compounds at the level of the generation process. "The SPROUT algorithm assigns a different penalty to each fragment WHILE assembling the compounds based on a db of fragments with known complexity." [2] Maybe we can apply the same logic here in drugex v2? - - We can attempt to introduce the synthetic feasibility at the level of molecule generation and then compare the results produced by something such as AiZynthFinder and use that as a benchmark to see if the molecules that we generate improve in how synthesizable they are without hindering the other desirable properties. - By using multiple retrosynthesis planners, we can get a less biased result pool to validate these results with. E.g. we can make use of aizynthfinder, retro* and others in combination and assess the differences in results given by these approaches. - We may use similar approaches to an older algorithm called SPROUT to do the synth check at generation level [1]. However, as the authors themselves mention, the limitation here is that the databases used to penalize the complexity of certain structures, are not complete. They propose that using much larger databases is a possible way to ameliorate the problem, but we need to investigate if this is truly a smart approach. This is not a trivial decision, as this algorithm was published about 15 years ago and the current databases and the amount of data has grown dramatically since then. - Additionally, the paper describing the SPROUT algorithm also mentions the idea that we want to implement: They had access to a rule-based expert system called CAESA that used retrosynthetic analysis to estimate synthetic accessibility of a generated compound... - However, they also mentioned that it is too slow for the sheer amount of generated compounds, but would be useful to prune early in the structure generation stage, and avoid combinatorial explosion (were it tractable) - RA Score can serve the same purpose as this, and it is an idea to incorporate this early on in DrugEx's structure generation loop to steer the novel molecules into a direction that avoids synthetically infeasible products - This does come with the uncertainty of RA score's performance, but we can subsequently benchmark this as mentioned above by scoring a sample of the generated structures using multiple retrosynthetic engines. - The authors of SPROUT noted that ”many cases of synthetic complexity were caused by the presence of uncommon substitution patterns in rings and chains rather than from the presence of more obvious complex features such as stereocenters”. [1]. - We will be simply focusing on the insight generated by the SPROUT paper, and other ideas such as a new much larger version of a complexity database such as the one used in said paper, will be left to future research. - The scope of this paper will thus revolve around finding the correct implementation of RA Score within DrugEx, and the tuning thereof, i.e. where and how do we prune the structure that is being generated? Is pruning our only option or can we steer the generation itself (e.g. manipulate the distribution of likely next SMILES char). - Next to testing samples for their feasibility using various retrosynthesis engines, we will also compare the usefulness to much simpler measures of synthetic accessibility such as the SA score to see if the impact of this method is sufficient to consider further investigation.

- The tools I will be using are as follows... - All code will be written in Python ¿= 3.7 - I will be using pytorch for the construction of any neural networks - I will be using jupyter notebooks for experimentation and in-code example-based explanations... - Any code that is eventually used in the main pipeline for the generation of the molecules, will be extracted as a module and added as source code - I will be using Git and Azure DevOps for version control and planning (eventually automatically building, testing and deploying the model if time allows, but this is not a priority). - I will adhere to 2 week sprints and have a presentable demo every meeting. - Experiments for the thesis are expected to be completed in late November.

- Deliverables include... - A modified version of DrugEx that can generate novel molecular structures with emphasis on synthetic feasibility at the structure-generation level. - Several examples of generated molecules and analysis of their complexity relative to the configuration used to generate them (analysis done through e.g. Retro* or AiZynthFinder, manual inspection with Anthe etc.) - Documentation on how to use the model and how to eventually expand on the code (for future students, or myself!) - A thesis (30 pages, excl. refs). - All code & data used for the experiments, ideally tested where necessary and deployable - A simple interface for quickly being able to reproduce and validate all experiments - reports generated from experiments

References

[1] Krisztina Boda and A Peter Johnson. Molecular complexity analysis of de novo designed ligands. *Journal of medicinal chemistry*, 49(20):5869–5879, 2006.

[2] Varnavas D Mouchlis, Antreas Afantitis, Angela Serra, Michele Fratello, Anastasios G Papadiamantis, Vassilis Aidinis, Iseult Lynch, Dario Greco, and Georgia Melagraki. Advances in de novo drug design: from conventional to machine learning methods. *International journal of molecular sciences*, 22(4):1676, 2021.