INVESTIGATING SUBSTRUCTURE VOCABULARY IN BLOCK-BASED MOLECULAR GENERATION

Tian Xiang Cheng, Dr Shen Bingquan
Raffles Institution (Junior College), 1 Raffles Institution Ln, Singapore 575954
DSO National Laboratories, 12 Science Park Dr, Singapore 118225

ABSTRACT

Recent *de novo* molecular design methods make use of atom-based, fragment-based or reaction-based representations as building blocks. Fragment-based approaches constrain generation of new molecules to known substructures. The chosen substructures thus greatly affect the range of chemical space explored (diversity of molecules) and effectiveness of traversal (optimal solutions through considering fewer potential molecules). We seek to understand the impacts of substructure vocabulary on training and generating molecules.

BACKGROUND AND MOTIVATION

The occurrence and risk of infectious disease has been rising with the rapid changes in climate and demography (Baker et al., 2021), necessitating a faster and more efficient drug discovery process to combat diseases. Yet drug discovery has traditionally been a costly and time-consuming process. As such, the field has sought to automate the process using computational methods, primarily a) Simulation, b) Virtual screening and c) *De novo* drug design (Meyers et al., 2021), each with their respective pros and cons. Our focus is on *de novo* drug design, which has recently seen much success. *De novo* drug design is based on optimisation, evolutionary strategies and deep generative models. It can thus leverage on the recent advances in reinforcement learning and generative models such as Generative Adversarial Networks and Variational Autoencoders. Deep learning has achieved a human-level accuracy in computer vision and can generate realistic text and images (Chai et al., 2021). The maturity in the field of deep learning thus makes it a very promising direction to work towards in drug design.

Traditionally, deep generative models involved atom-based generation, however the model must generate chemically invalid intermediaries and delay validation until a complete graph is generated. This also creates problems in generating ring structures. As such, newer techniques make use of fragment-based techniques, where chemically valid molecular substructures are joined together. (We will use the terms *fragment* and *substructure* interchangeably.) Substructures refer to one or more atoms bonded together, with incomplete bonds or attachment points. These attachment points can then bond to other molecules. This allows more complicated substructures to be included, which retains more complicated chemical properties unique to the substructures. Furthermore, this allows larger and more complex molecules to be generated, exploring more of the chemical space (Meyers et al., 2021).

Currently, there are limited or no standardisations as to what type and number of substructures are used and how they are obtained. Different methods use differing heuristics-based substructure sampling methods. One example would be a simple method proposed by Jin et. al (2020): For each molecule in a dataset, bonds attached to a ring are split. The substructures are then evaluated for how many times they occur, if they occur more than 100 times, they are retained, otherwise they are split down further. Though simple and efficient, the method is largely heuristics, and the number 100 is arbitrary.

We seek to highlight the importance of types of substructure and consequently encourage standardisation of substructure sampling methods or constructing substructure datasets.

HYPOTHESIS

We hypothesise that increasing the size and number of substructures will improve the diversity and drug-likeness of the generated molecules.

METHODS

We chose the Proximal Policy Optimisation[1] (PPO) method of molecular generation as a standard method. It is a tested and proven reinforcement learning algorithm, chosen for its

[1] https://github.com/GFNOrg/gflownet/blob/master/mols/ppo.py

ease of use and good (or acceptable) performance. We will also be using Python for its flexibility, ease in handling large datasets and maturity in machine learning applications. We will primarily use the RDKit[2] library, an open-source cheminformatics and machine learning library, for handling molecules, such as printing and fingerprinting.

The original dataset of substructures provided by the authors of the PPO method consists of 105 molecular substructures[3], of relatively small molecular size. They have an average molecular weight of 74.3 g/mol. We call this fragment set A.

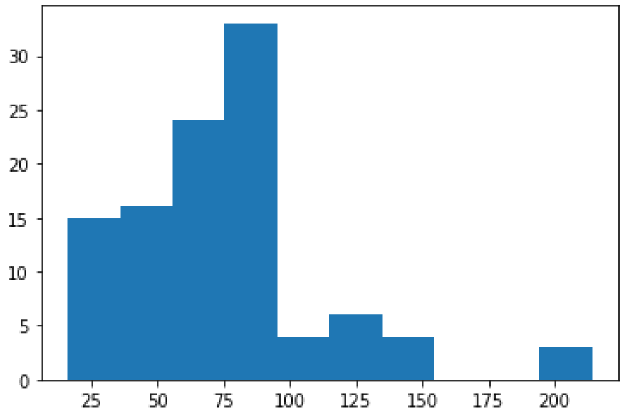


Figure 1: Histogram of fragment set A molecular weights

The new dataset of substructures is obtained from the previously mentioned paper (Jin et al., 2020). This dataset is chosen for its simple heuristics and as it is used for generation of similar molecules – drug-like small organic molecules. The substructure construction method was described earlier, and the initial dataset of molecules used was the ChEMBL database. As quoted from the website[4], "ChEMBL is a manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs." This precisely matches our objective in molecular generation.

However, the substructures generated by the method are expressed in a different format as to what is readable by the PPO algorithm. As such, a conversion process was written by us.

[2] https://www.rdkit.org/
[3] https://github.com/GFNOrg/gflownet/blob/master/mols/data/blocks_PDB_105.json
[4] https://www.ebi.ac.uk/chembl/

The new dataset of substructures from Jin et al. (2020) and converted to a readable format consists of 5030 molecular substructures, of comparable but larger molecular size. They have an average molecular weight of 97.5 g/mol. We call this fragment set B.

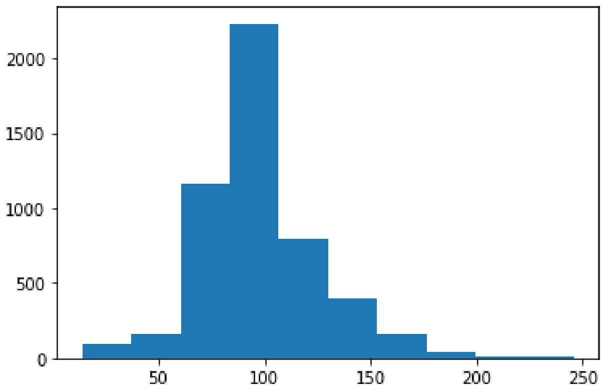
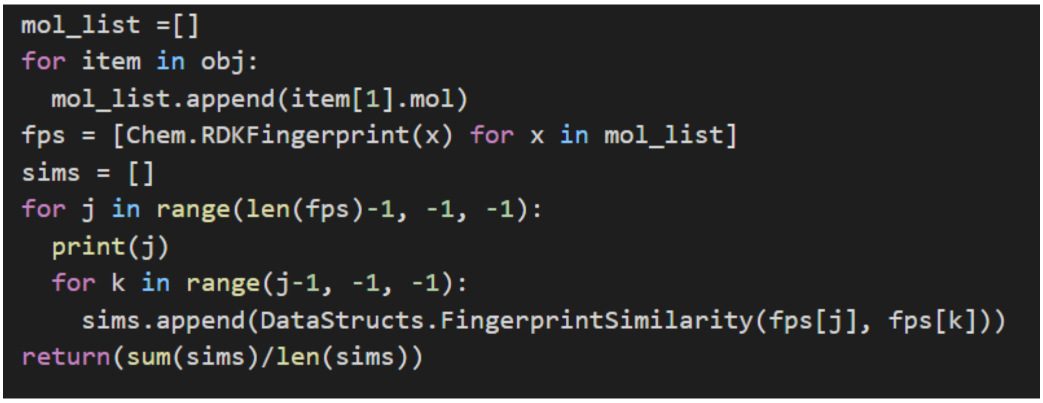


Figure 2: Histogram of fragment set B molecular weights

PPO is trained with both sets for 1000 iterations each and optimised for octanol–water partition coefficient (LogP) as a dummy desired drug-property. Generated molecules are then assessed for their LogP value and similarity.

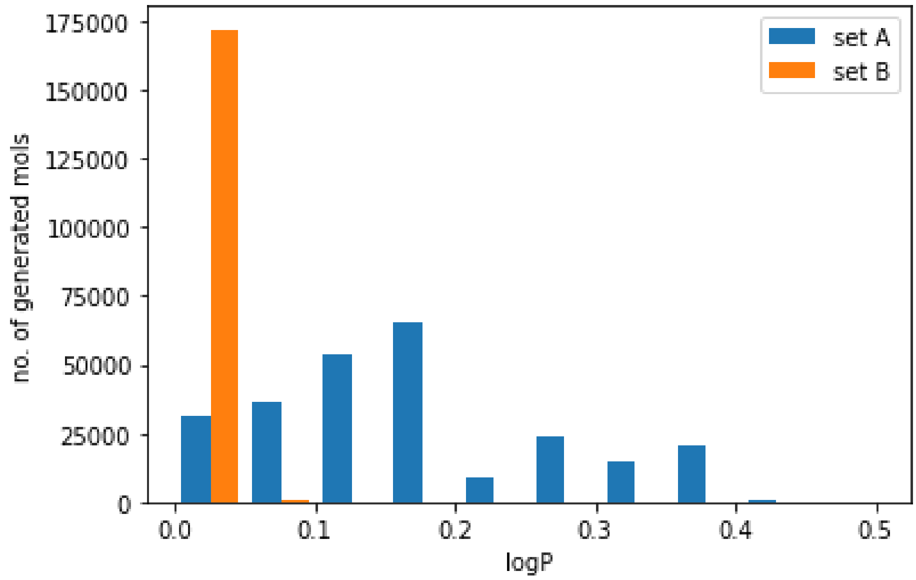
LogP values are generated during the PPO training process for the final molecule. For similarity, first molecular fingerprints are generated using RDKit, then fingerprints are compared between every two molecules once for similarity. Finally, the average similarity is obtained across all comparisons. Below is a part of the process written in Python:



```
mol_list =[]
for item in obj:
  mol_list.append(item[1].mol)
fps = [Chem.RDKFingerprint(x) for x in mol_list]
sims = []
for j in range(len(fps)-1, -1, -1):
  print(j)
  for k in range(j-1, -1, -1):
    sims.append(DataStructs.FingerprintSimilarity(fps[j], fps[k]))
return(sum(sims)/len(sims))
```

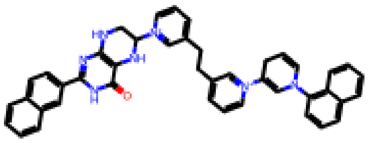
RESULTS AND DISCUSSION

LogP:

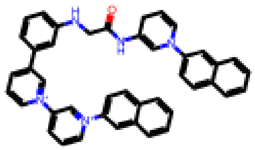


From the diagram, the overall performance of model with fragment set B unfortunately falls off significantly, with generated molecules exhibiting much lower desired property.

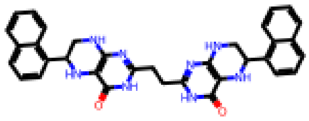
From fragment set A, the top five molecules with highest LogP values are shown here:



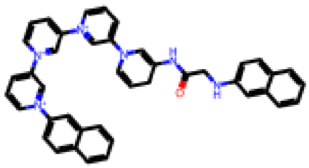
0.505



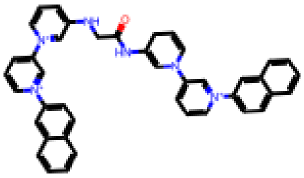
0.487



0.475

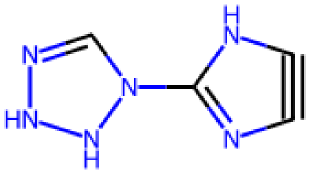


0.466

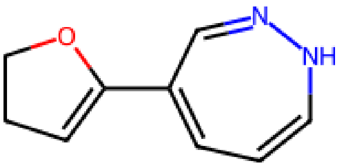


0.465

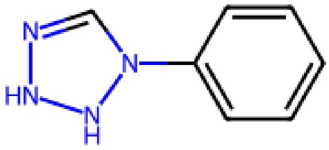
From fragment set B, the top five molecules with highest LogP values are shown here:



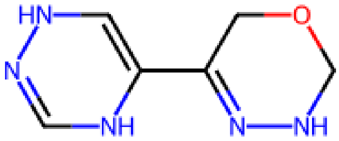
0.317



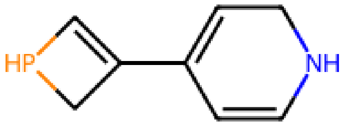
0.236



0.198



0.193



0.191

Similarity:

As described earlier, we look at the average similarity for the 100 generated molecules of highest LogP value.

For fragment set A, the top 100 molecules have a high average similarity of 0.878 out of 1.

For fragment set B, the top 100 molecules have a low average similarity of 0.145 out of 1.

In order to discover new drugs, the higher the diversity (lower the similarity) the more ideal, while retaining high LogP values. Our new fragment set B was able to perform significantly better fragment set A, we believe the larger choice of substructures lead to more choices in generating molecules and introduces additional stochasticity.

One major limitation in our approach may have been running fewer than optimal iterations of training. Originally, the paper recommended 4000 iterations, however due to insufficient compute time and ability, we chose 1000 iterations as an acceptable amount. This may generate molecules with lower LogP values. Another limitation would be that not all substructures in fragment set B are used in training. This is due to different representations of substructures causing chemically invalid structures to be generated. As such these instances are excluded, leading to fewer substructures in fragment set B (though still substantially more than fragment set A).

FUTURE WORK

Our method focused on generating drug-like organic molecules, however fragment-based molecular generation can also be used for polymer generation and other different types of molecules. We would like to explore the types of substructures constructed from these molecules and generation of molecules with them.

We would also like to explore other fragment-based methods and whether they are similarly affected by the substructures used. Finally, we seek to develop our own substructure construction method.

ACKNOWLEDGEMENTS

I would like to thank Dr Shen Bingquan for his constant guidance and support. His patience and expertise greatly aided my research process. I am also grateful for the help and learning opportunities provided by his institution.

This project would not have been possible without the help of my school coordinators for allowing me to conduct research safely and share my findings.

Thank you to my peers and all who have helped me in one way or another.

REFERENCES

[1] Baker, R. E., Mahmud, A. S., Miller, I. F., Rajeev, M., Rasambainarivo, F., Rice, B. L., Takahashi, S., Tatem, A. J., Wagner, C. E., Wang, L.-F., Wesolowski, A., & Metcalf,

C. J. E. (2021, October 13). *Infectious disease in an ERA of global change*. Nature News. Retrieved January 4, 2023, from https://www.nature.com/articles/s41579-021-00639-z

[2] Meyers, J., Fabian, B., & Brown, N. (2021). De novo molecular design and Generative Models. *Drug Discovery Today*, *26*(11), 2707–2715. https://doi.org/10.1016/j.drudis.2021.05.019

[3] Chai, J., Zeng, H., Li, A., & Ngai, E. W. T. (2021). Deep Learning in Computer Vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, *6*, 100134. https://doi.org/10.1016/j.mlwa.2021.100134

[4] Jin, W., Barzilay, R., & Jaakkola, T. (2020, April 18). *Hierarchical generation of molecular graphs using structural motifs*. arXiv.org. Retrieved January 3, 2023, from https://arxiv.org/abs/2002.03230