CSCE 670 Project Summary: Molecule Recommendation

Yuncheng Yu
yyc@tamu.edu

Pragati Naikare
pragatinaikare311@tamu.edu

Jacob Helwig
jacob.a.helwig@tamu.edu

Ya-Ru Yang
yaruyang@tamu.edu

Our project aims to leverage users' past research experiences to recommend innovative synthesized molecules. These recommendations will serve as potential subjects of investigation for researchers in biomedical, drug, and chemistry-related fields. To achieve this, we have developed a powerful molecule recommendation pipeline, tailored to recommend molecules to researchers based on their past interactions. Our contributions are as follows:

1. **MolRec Dataset**. The MolRec dataset describes researcher-molecule interactions and contains 5.4K ratings spanning 1.2K authors and over 100 chemical compounds [Hastings et al., 2016].
2. **A 2 stage recommendation pipeline**. Given a researcher's historical molecule interactions, our recommendation pipeline curates a set of molecules from the MolRec data, and furthermore generates novel molecules which are aligned with the researcher's interests.

   Stage I In the first stage, we use state-of-the-art methods for recommendations from the MolRec dataset based on collaborative filtering and chemical semantic similarity [Barros et al., 2021]. Our results show that a hybrid of alternating least squares and chemical semantic similarity based on the molecule's ChEBI ontology give the best performance on a held-out test set.

   Stage II In the second stage, we train a Junction Tree Variational Autoencoder [Jin et al., 2018] on the Zinc250K dataset [Sterling and Irwin, 2015] of 250K drug-like compounds. We are then able to condition generations from the VAE on the molecules recommended in the first stage to generate novel molecules beyond those in the MolRec data based on the researcher's interests.

References

Janna Hastings, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. Chebi in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*, 44(D1): D1214–D1219, 2016.

Marcia Barros, Andre Moitinho, and Francisco M Couto. Hybrid semantic recommender system for chemical compounds in large-scale datasets. *Journal of cheminformatics*, 13:1–18, 2021.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR, 2018.

Teague Sterling and John J Irwin. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11):2324–2337, 2015.