MolMiner: Transformer architecture for fragment-based autoregressive generation of molecular stories

Raul Ortega-Ochoa[1], Tejs Vegge[1], Jes Frellsen[1]
[1]Technical University of Denmark, Kgs. Lyngby, Denmark

rauoc@dtu.dk

Deep generative models for molecular discovery [1] have become a very popular choice in new high-throughput screening paradigms [2,3]. These models have been developed inheriting from the advances in natural language processing and computer vision, achieving ever greater results. However, generative molecular modelling has unique challenges that are often overlooked. Chemical validity, interpretability of the generation process and flexibility to variable molecular sizes are among some of the remaining challenges for generative models in computational materials design. In this work, we propose an autoregressive approach that decomposes molecular generation into a sequence of discrete and interpretable steps using molecular fragments as units, a 'molecular story'. Enforcing chemical rules in the stories guarantees the chemical validity of the generated molecules, the discrete sequential steps of a molecular story makes the process transparent improving interpretability, and the autoregressive nature of the approach allows the size of the molecule to be a decision of the model. We demonstrate the validity of the approach in a multi-target inverse design of electroactive organic compounds, focusing on the target properties of solubility, redox potential, and synthetic accessibility. Our results show that the model can effectively bias the generation distribution according to the prompted multi-target objective.

Figures

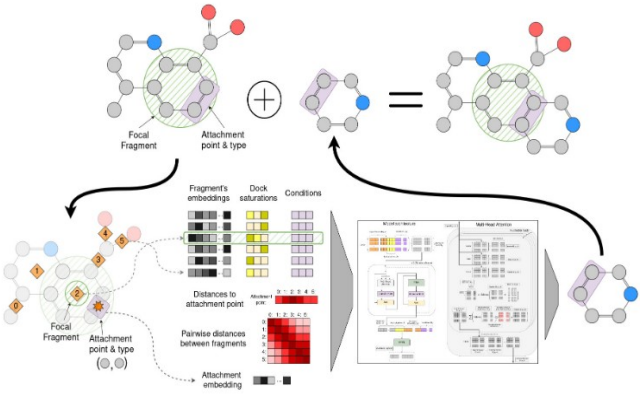

Figure 1. Integration of Molminer in the creation of a molecular story step. The model takes the molecule to be grown at a focal fragment through an attachment point and predicts which fragment at which attachment configuration should be added, resulting in a step of a ‘molecular story’

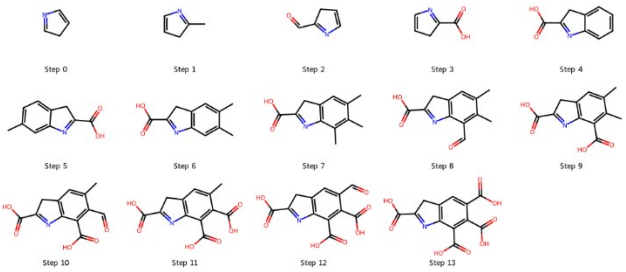

Figure 2. An example of a story generated in the creation of a compound.

References

[1] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik, *Science*, **361**(6400) (2018) 360–365.

[2] Julia Westermayr, Joe Gilkes, Rhyan Barrett, and Reinhard J. Maurer, *Nature Computational Science*, **3**(2) (2023) 139–148.

[3] Raul Ortega Ochoa, Bardi Benediktsson, Renata Sechi, Peter Bjørn Jørgensen, and Arghya Bhowmik, *J. Mater. Chem. A*, **11** (2023) 26551–26561.