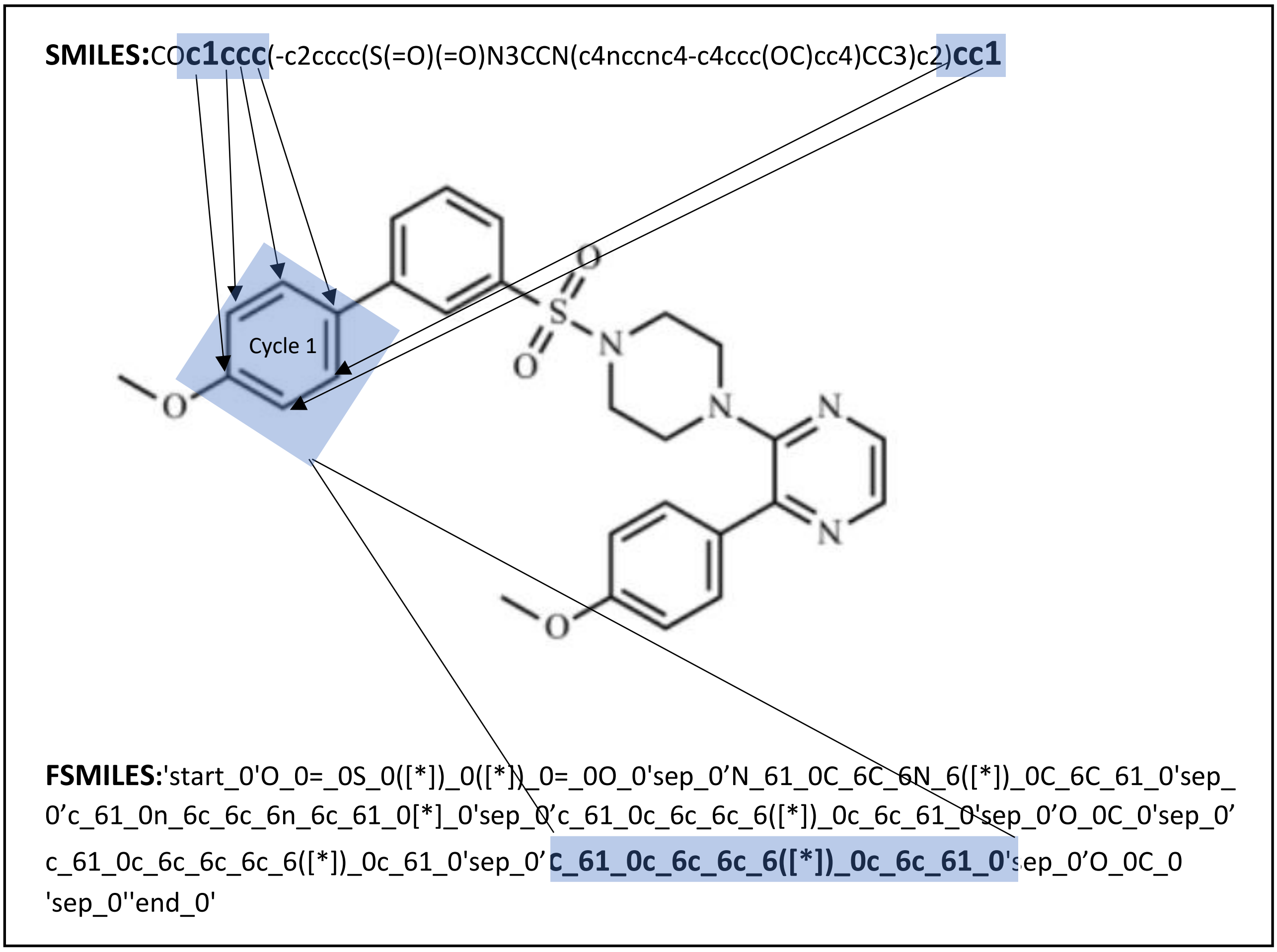
FSMILES VS SMILES

Youcef Bagdad



SMILES:CO**c1ccc**(-c2cccc(S(=O)(=O)N3CCN(c4nccnc4-c4ccc(OC)cc4)CC3)c2)**cc1**

FSMILES:'start_0'O_0=_0S_0([*])_0([*])_0=_0O_0'sep_0’N_61_0C_6C_6N_6([*])_0C_6C_61_0'sep_0’c_61_0n_6c_6c_6n_6c_61_0[*]_0'sep_0’c_61_0c_6c_6c_6([*])_0c_6c_61_0'sep_0’O_0C_0'sep_0’c_61_0c_6c_6c_6c_6([*])_0c_61_0'sep_0’**c_61_0c_6c_6c_6([*])_0c_6c_61_0**'sep_0’O_0C_0'sep_0''end_0'

Figure 1 : Comparison between SMILES and FSMILES representations in the succecivity of atom cycles. The parts of the SMILES and FSMILES representations belonging to cycle 1 are highlighted in the same color as this cycle (in blue).

SMILES (Simplified Molecular Input Line Entry System) is an approach for encoding molecular structures into ASCII strings, capturing atomic composition, connectivity, and sometimes the stereochemical configuration of compounds. This technique represents molecules as graphs, where atoms and bonds are symbolized by nodes and edges, respectively, and employs a depth-first search for linear enumeration of atoms. However, this enumeration technique of the hole molecule has limitations, particularly a sometimes discontinuous representation of cycles as illustrated in Figure 1, SMILES section, where it can be seen that atoms of cycle 1 are dispersed into two parts, each at one end of the SMILES string.

In an approach that aims to use these character strings to create attention-based generative models, this limitation could lead to a misinterpretation of molecular structures, notably cycles, thus affecting the accuracy and relevance of the generated molecules.

Indeed, current state-of-the-art generative models, whether based on SMILES or molecular graphs, have found that their models struggle to generate molecules with correct cyclic structures having correct ring size or correct number of cycles (as mentioned in our article of interest by Feng et al. 2024).

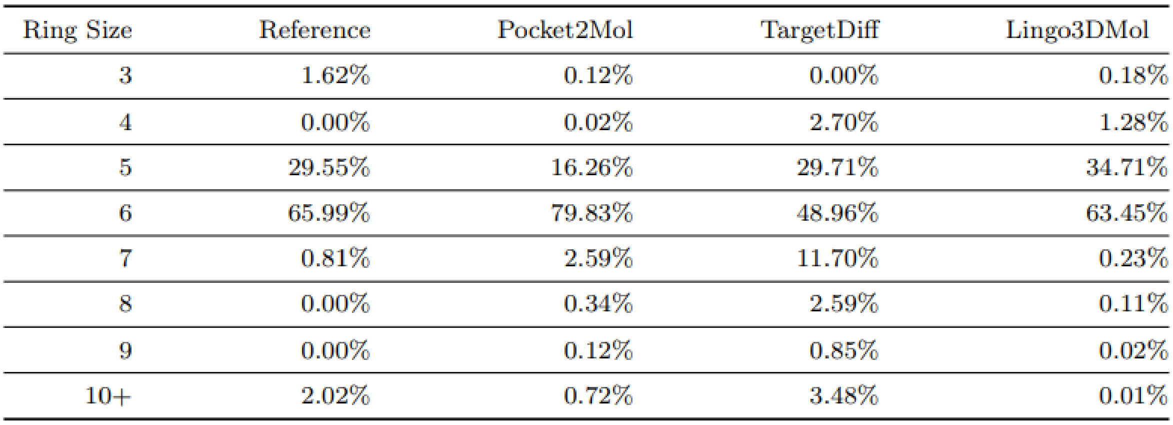
To overcome this limitation of SMILES, Feng *et al.* introduced FSMILES (Fragment SMILES), a new molecular representation that strategically divides molecules into fragments. Each fragment, potentially containing a cycle, is then encoded separately in SMILES format, ensuring a continuous representation of cycles, as demonstrated in Figure 1, FSMILES section, where the cycle 1 is presented continuously, unlike in SMILES. This new representation could improve the accuracy of generative models by making these molecular details easier to capture and interpret by attention models.

FSMILES also enhance traditional SMILES notation by explicitly adding the size of cycles near each cyclic atom within fragments. This provides essential additional details, absent in SMILES. This integration further highlights cycles, thus allowing for potentially more accurate recognition and interpretation of these important structures by attention models.

Therefore, with their ability to provide an integral and detailed representation of molecular structures, especially cycles, FSMILES represents potentially a better option than SMILES for training generative models.

To show the efficacy of the model 'Lingo3DMol' developed in Feng et al. study which employs FSMILES, the following table offers a comparative analysis against two other different models.

Table 1 : Comparison of the ring size distribution in molecules generated by different methods (from Feng *et al.* 2024).



Ring Size	Reference	Pocket2Mol	TargetDiff	Lingo3DMol
3	1.62%	0.12%	0.00%	0.18%
4	0.00%	0.02%	2.70%	1.28%
5	29.55%	16.26%	29.71%	34.71%
6	65.99%	79.83%	48.96%	63.45%
7	0.81%	2.59%	11.70%	0.23%
8	0.00%	0.34%	2.59%	0.11%
9	0.00%	0.12%	0.85%	0.02%
10+	2.02%	0.72%	3.48%	0.01%

Table 1 reveals that the Lingo3DMol model has a lower tendency to generate molecules with a ring size greater than 7 compared to the TargetDiff (Guan *et al.* 2023) and Pocket2Mol (Peng *et al.* 2022) models, aligning with reference statistics. This observation suggests that Lingo3DMol, a model leveraging FSMILES (and other features like global and central coordinates of atoms), demonstrates promising potential in avoiding the generation of molecules with undesirable ring sizes. This further reinforces our confidence in choosing FSMILES for the development of generative models for molecules.

References :

Feng W, Wang L, Lin Z *et al.* Generation of 3D molecules in pockets via a language model. *Nat Mach Intell* 2024;**6**:62–73.

Guan J, Qian WW, Peng X *et al.* 3D Equivariant Diffusion for Target-Aware Molecule Generation and Affinity Prediction. 2023, DOI: 10.48550/arXiv.2303.03543.

Peng X, Luo S, Guan J *et al.* Pocket2Mol: Efficient Molecular Sampling Based on 3D Protein Pockets. 2022, DOI: 10.48550/arXiv.2205.07249.