This repo implements macrocyclization of linear molecules to generate macrocycles with chemical diversity and structural novelty.

Setup

Install Macformer from the .yaml file

conda env create -f Macformer\_env.yaml

conda activate Macformer

Quick Start

1. Data processing

The acyclic-macrocyclic SMILES pairs extracted from ChEMBL and ZINC database, respectively, can be found in the data/ folder. Or researcher can process their own macrocyclic compounds from scratch using scripts in the utils/ folder.

fragmentation.py: generate unique acyclic-macrocyclic SMILES pairs

data\_split.py: split the acyclic-macrocyclic SMILES pairs into train, validation, and test datasets with the ratio of 8:1:1

data\_augmentation.py: implement substructure-aligned data augmentation

2. Input files generation

The preprocessing.sh script will generate following input files necessary to train the model.

\*.train.pt : serialized PyTorch file containing training data

\*.valid.pt: serialized PyTorch file containing validation data

\*.vocab.pt: serialized PyTorch file containing vocabulary data

3. Model training

Run the training.sh script to start model training.

The saved checkpoints can be averaged by running the average\_models.sh script.

4. Model evaluation

Run the testing\_beam\_search.sh script to obtain predicted molecules.

The utils/model\_evaluation.py script can be used to calculate the evaluation metrics, including recovery, validity, uniqueness, novelty, and macrocyclization.

5. Pre-trained models

The models pretrained with ChEMBL dataset can be found in the models/ folder.