Challenge 2: Molecule Generation

Giovanni Filomeno

Artificial Intelligence in Life Science

JOHANNES KEPLER UNIVERSITY LINZ

Problem overview

This challenge focuses on developing machine learning techniques to generate realistic and novel molecular structures. The primary objective is to minimize the Frechet ChemNet Distance (FCD), ensuring high quality while maintaining high scores for novelty, validity, and uniqueness.

Model Architecture

- **Input:** Raw character-level Smiles with special tokens
- **Embedding:** 128-dim lookup table
- **Back-bone:** 2 layer LSTM, 512 hidden units, dropout=0.3
- **Output Head:** Linear → Softmax over vocabulary
- **Sampling:** Multinomial sampling with adjustable temperature to control diversity of generated molecules

Training Strategy & Tuning

Dataset:

- SMILES encoded as token sequences
- Train-validation split: 95%-5%, random stratification
- Batch size: 512

Training Setup:

- Optimizer: AdamW, learning rate = 1e-3
- Loss function: Cross-Entropy (ignoring PAD tokens)
- Scheduler: Cosine Annealing Warm Restarts
- Early Stopping: Patience of 5 epochs based on validation loss

Results

#	FCD	Novelty	Uniqueness	Validity	Description	Upload Date
1	0.418	0.961	1.0	1.0	Try_before_christmas_colab_3	May 25, 2025, 4:53 p.m.

< 0.8

high