Drug Discovery: Variational Autoencoder Techniques for Molecule Generation

Andrew Jacobson jonaj2@illinois.edu
Dixon Liang dixonl2@illinois.edu
John Judge jmjudge2@illinois.edu
Megan Masanz mjneuman@illinois.edu

Baseline Background & Literature Review

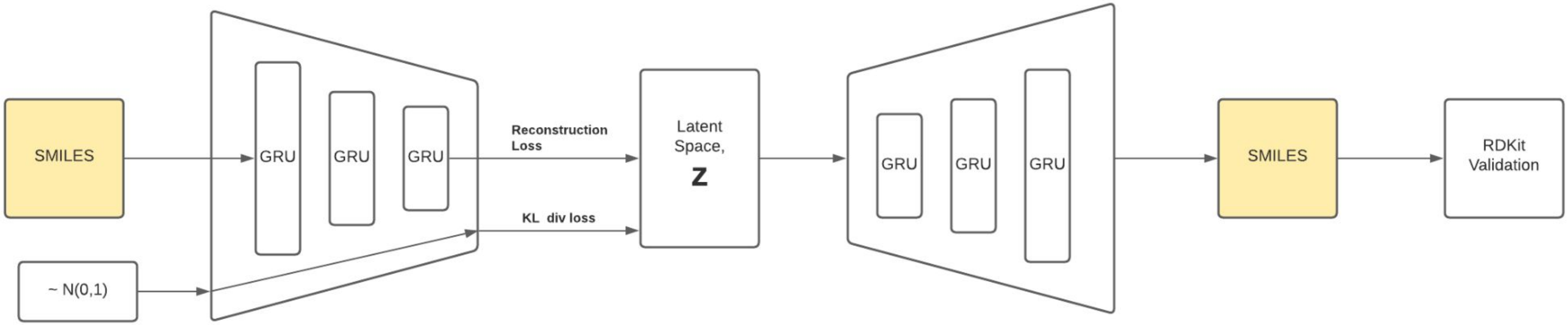
Motivation

- Search Space, Drug-Like, Synthesizable
  - Test validity against RDKit
- Baseline Model
  - Character Based Chemical VAE
  - Aspuru-Guzik
  - “Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules”

Improvements for Validity of Molecules

- KL Cost Annealing
  - Regularization hyperparameter during training process
- Teacher Forcing
  - Molecular Sets (MOSES) Implementation of VAE
- Self-Referencing Embedded Strings (SELFIES) vs Simplified Molecular INput Line Entry System (SMILES)
  - Adjusting the input with constraints

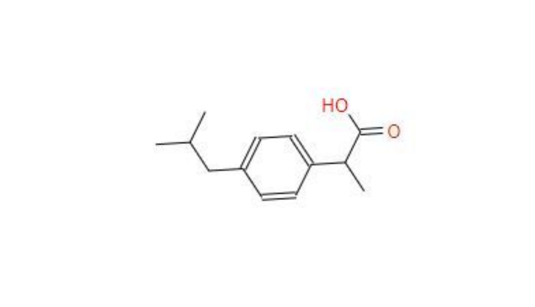
Baseline VAE Architecture

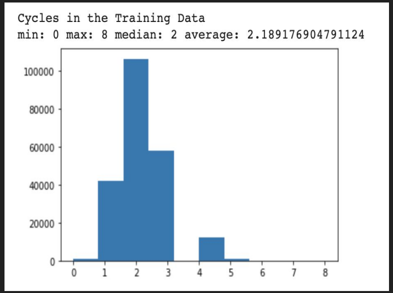


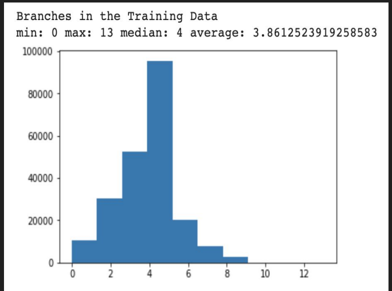
- Encoder
  - Three 1D Convolutions of size [9, 9, 10] and associated kernels of [9, 9, 11]
  - Dropout Layer
  - ReLU
- Decoder
  - Three layers of GRU
- Loss Function
  - Maximization of log-likelihood of input distribution given latent distribution

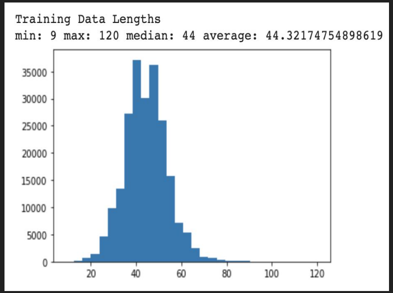
Data

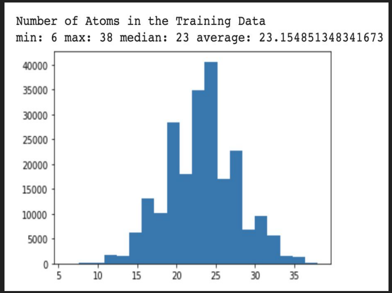
- ZINC15 Dataset
  - Sourced from the deepchem python library
  - 250k-molecule dataset of “lead-like compounds”
  - Used subset of 10k for initial testing
- SMILES Representation
  - CC(C)CC1=CC=C(C=C1)C(C)C(=O)O
- Conversion to SELFIES from SMILES





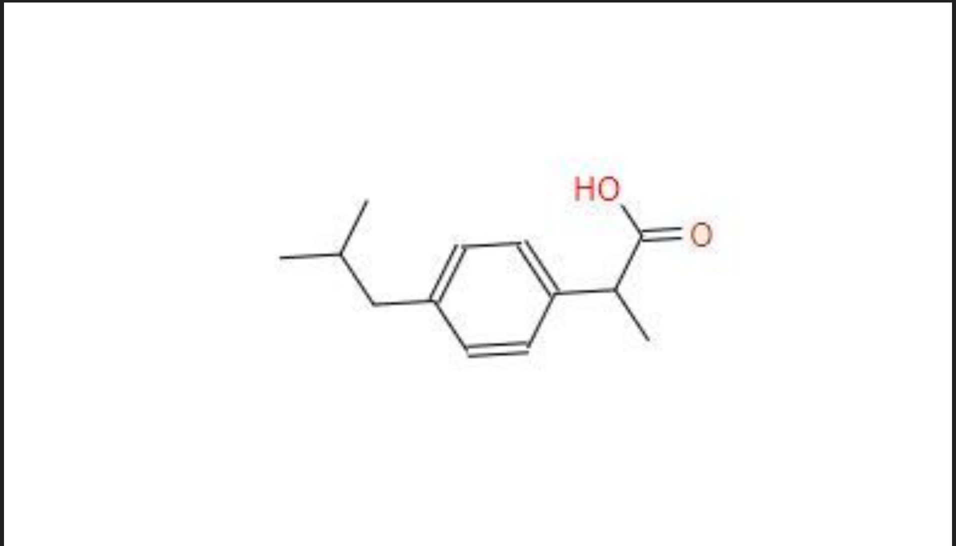






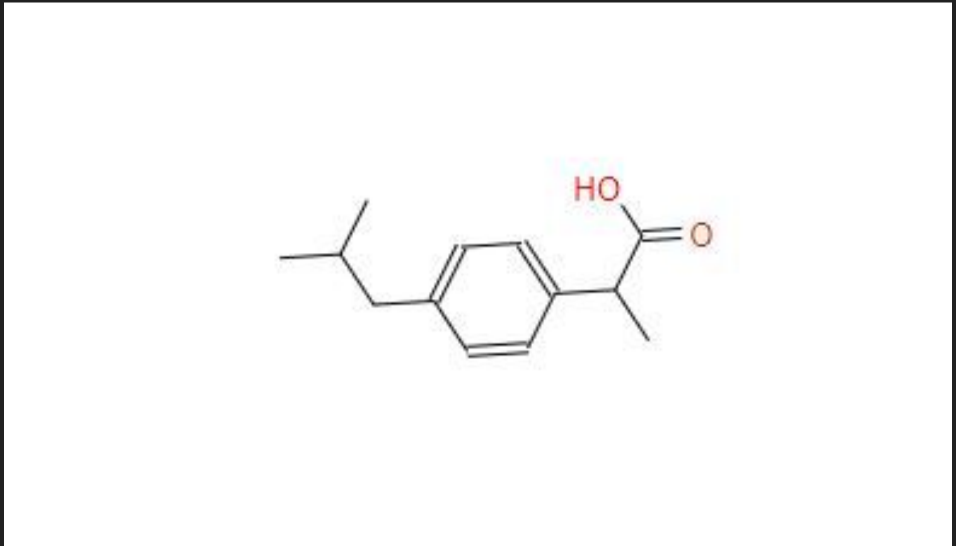
SMILES

CC(C)CC1=CC=C(C=C1)C(C)C(=O)O



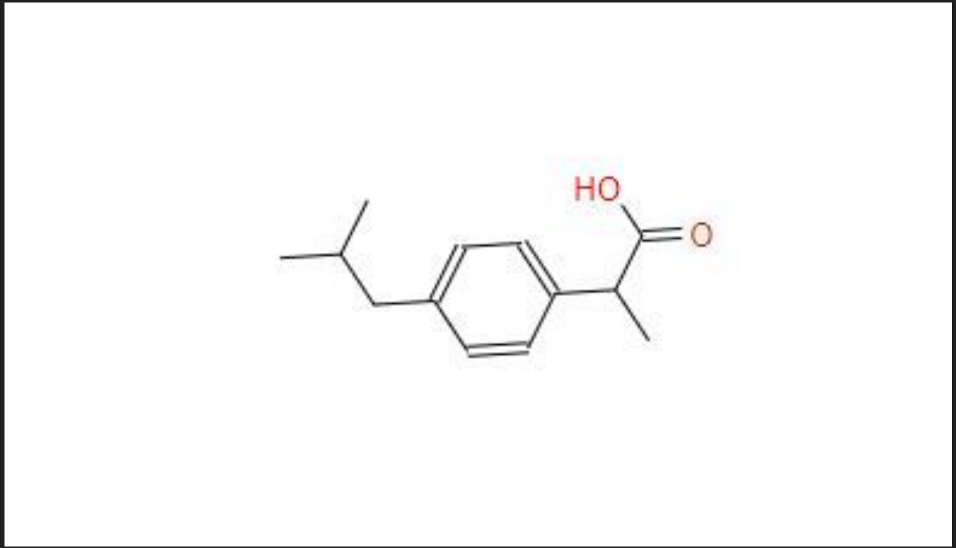
SMILES

CC(C)CC1=CC=C(C=C1)C(C)C(=O)O



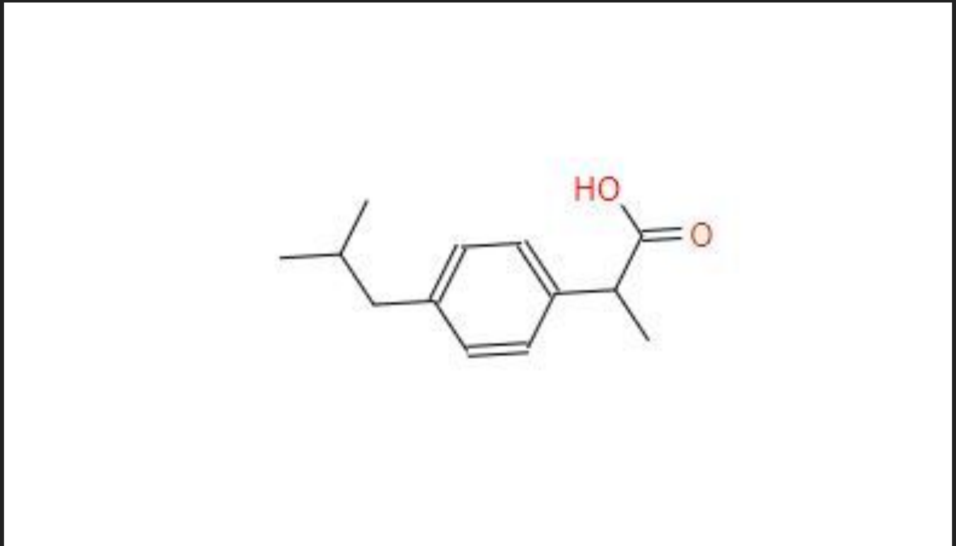
SMILES

CC(C)CC1=CC=C(C=C1)C(C)C(=O)O



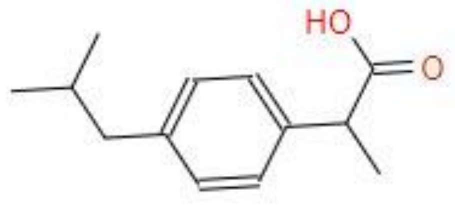
SMILES

CC(C)CC1=CC=C(C=C1)C(C)C(=O)O



SMILES

CC(C)CC1=CC=C(C=C1)C(C)C(=O)O



Training Environments

- Azure ML
  - Description: STANDARD_NC12
  - Metrics/Testing Environment
- Google Colab
  - GPU Enabled
  - Collaborative Environment

Approach - Phase 1

- Baseline Variational Autoencoder (VAE) Model
  - CPU w/ sample dataset
  - GPU w/ dataset of 250K
  - Hyperparameter tuning
    - Epochs
    - Learning Rate
    - Dropout

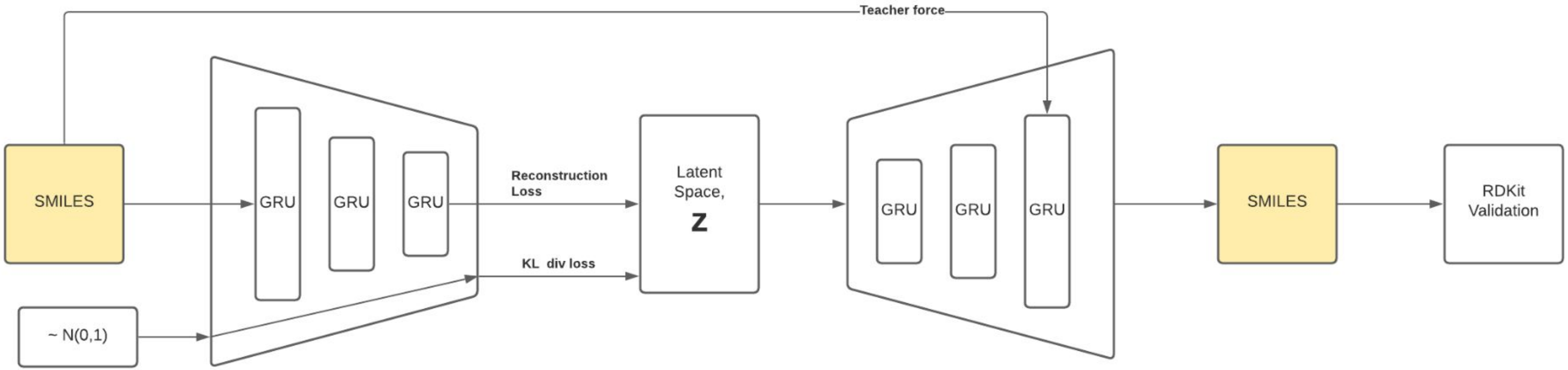
Approach - Phase 2

- VAE with Kullback-Leibler (KL) Cost Annealing
  - Theory:
    - Loss is comprised of reconstruction loss as well as a regularization term
    - Applying a weight to the KL Divergence so it starts at 0 and gradually increase
    - Early training emphasizes reconstruction loss
    - Later training emphasized KL divergence loss
  - Implementation: removing cost annealing from the base model implementation resulting in degradation regarding the number of valid molecules generated

Approach - Phase 3

- VAE with Teacher Forcing
  - Deepchem - initial attempt to update their VAE implementation with teacher forcing.
  - Moses - implemented teacher forcing by default

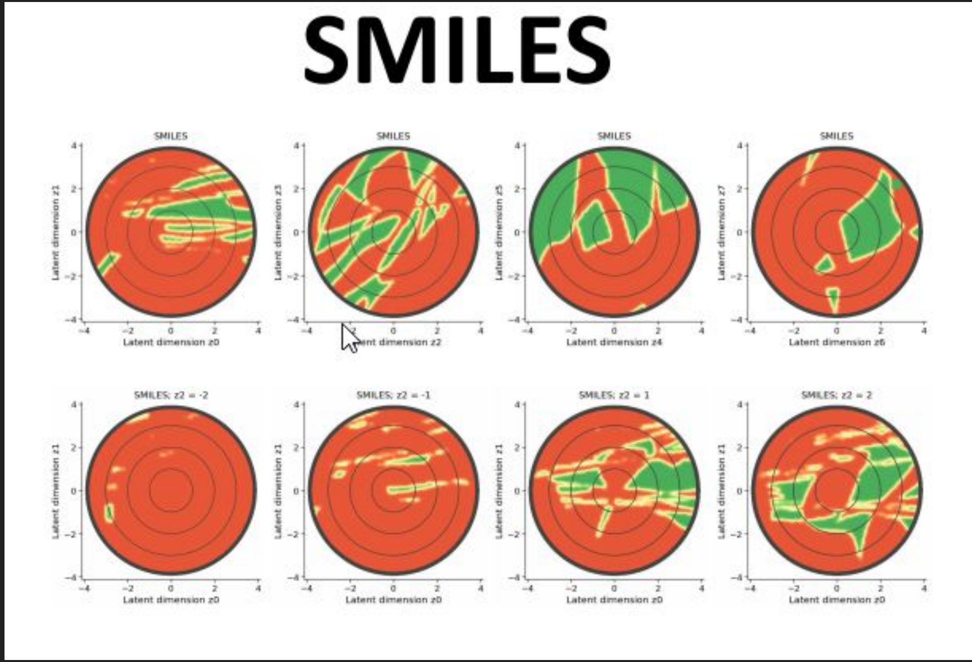
SMILES, with teacher forcing



Approach - Phase 4

- VAE using SELFIES instead of SMILES

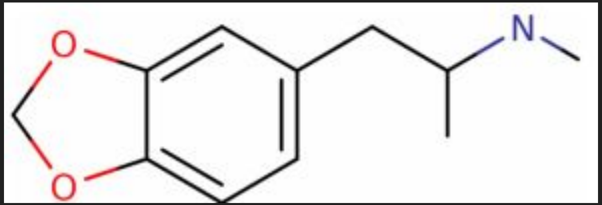
SELFIES: Motivation



Krenn, Mario, Florian Häse, Akshat Kumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. "Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation." Machine Learning: Science and Technology 1, no. 4 (November 2020): 045024.

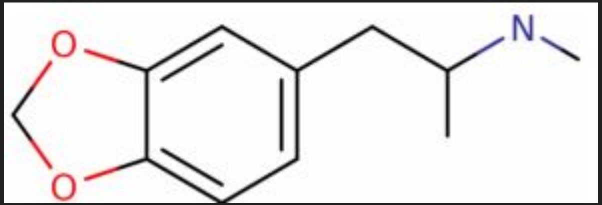
SMILES vs SELFIES

Example: MDMA



SMILES vs SELFIES

Example: MDMA



SMILES:

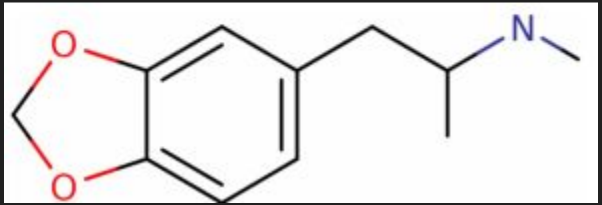
CNC(C)CC1=CC=C2C(=C1)OCO2

SELFIES:

[C][N][C][Branch1_1][C][C][C][C][=C][C][=C][C][Branch1_2][Ring2][=C][Ring1][Branch1_2][O][C][O][Ring1][Branch1_2]

Single Mutation Example

Example: MDMA



SMILES:

CNC(C)CC1=CC=C**N**C(=C1)OCO2

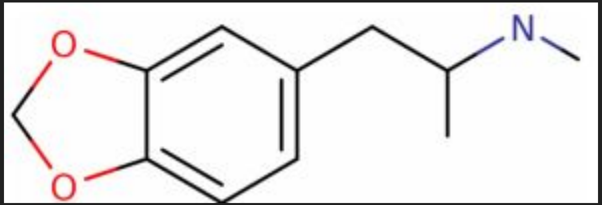
Invalid Syntax

SELFIES:

[C][N][C][Branch1_1][C][C][C][C][=C][C][=C][C][Branch1_2][Ring2][=C][Ring1][Branch1_2][O][C][O][Ring1][Branch1_2]

Single Mutation Example

Example: MDMA

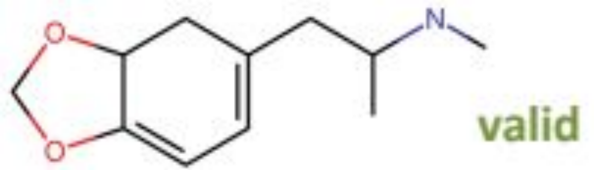


SMILES:

CNC(C)CC1=CC=C**N**C(=C1)OCO2

Invalid Syntax

SELFIES:

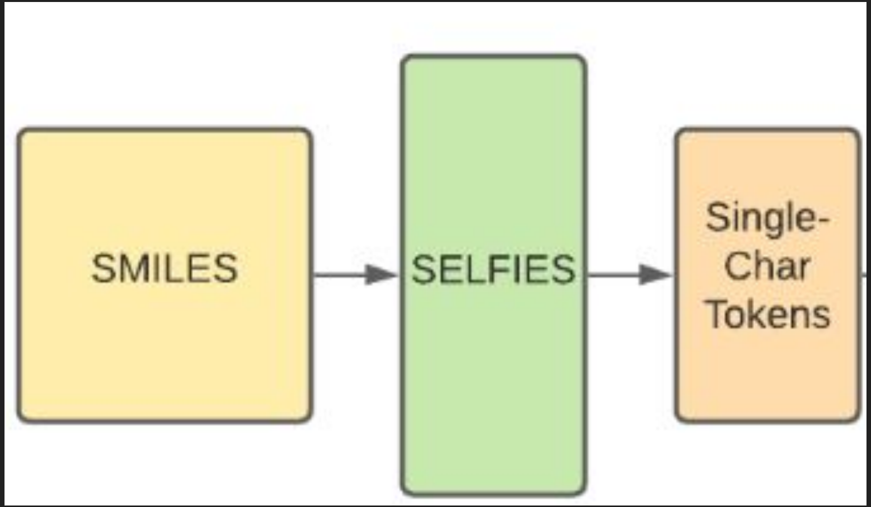


Approach - Phase 4

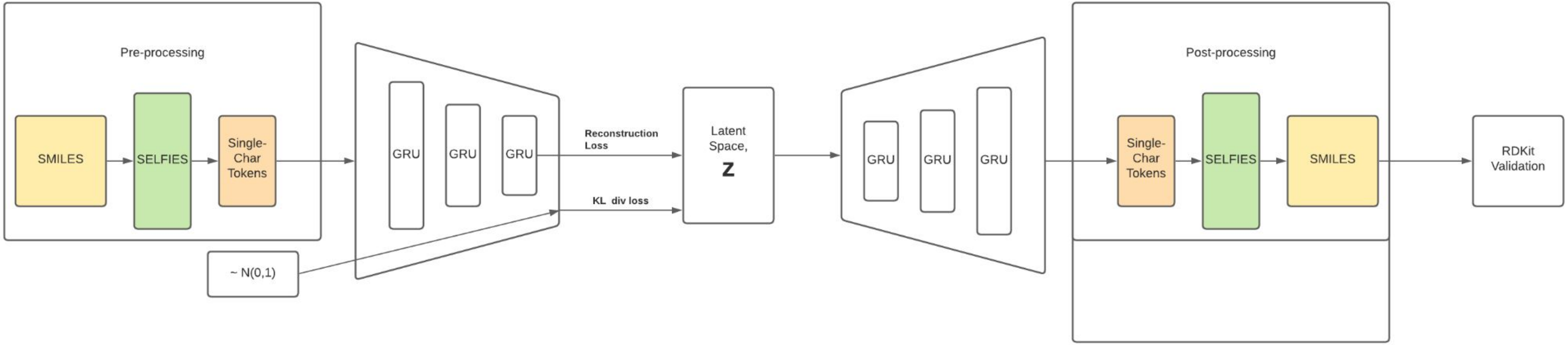
- VAE using SELFIES instead of SMILES

Approach - Phase 4

- VAE using SELFIES instead of SMILES
- Pre-processing
    - Conversion from SMILES to SELFIES
    - Single-character tokenize SELFIES



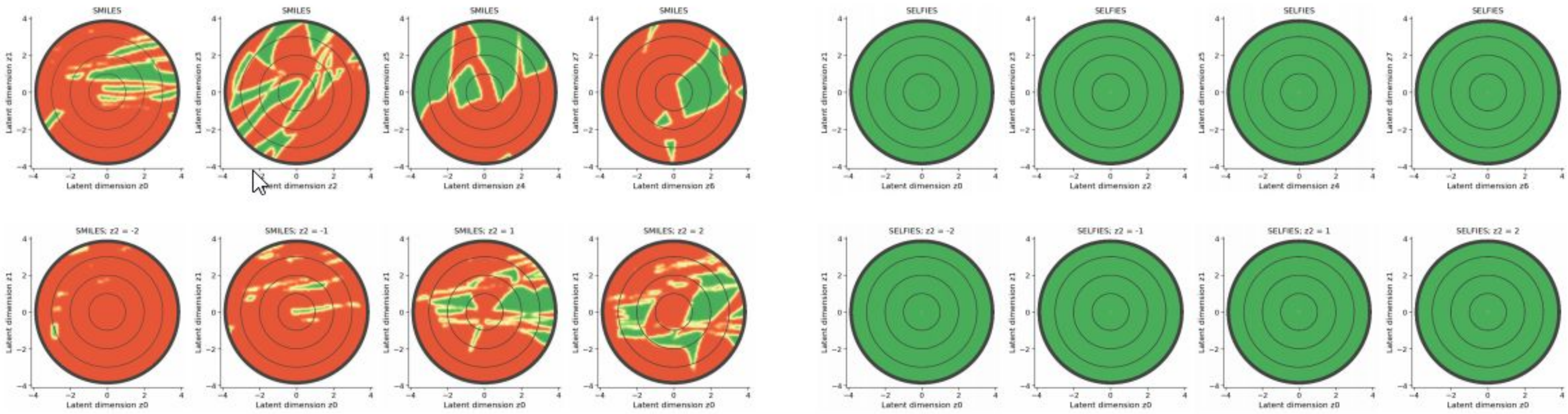
SELFIES, no teacher forcing



Validity of Latent Space in VAE

SMILES

SELFIES



Krenn, Mario, Florian Häse, Akshat Kumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. “Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation.” Machine Learning: Science and Technology 1, no. 4 (November 2020): 045024.

Beyond Syntactic Validity

Other metrics of chemical richness

- QED, SAS
- logP, Lipinski’s Rule of 5

Comparison to Similar Methods

Grammar: G-VAE

- Syntactic validity only

Junction-tree: JTN-VAE

- Restricted sampling space

Results

Baseline VAE	KL Annealing Disabled	Teacher Forcing Enabled	SELFIES instead of SMILES
.2-3%	.04-1%	45-95%	100%

Results are a percentage of syntactically valid molecules. The synthesizability of molecules is arguably more important but harder to determine. https://arxiv.org/pdf/2002.07007.pdf

Conclusion

- VAEs can struggle to learn the rules to a grammar like SMILES
- Teacher Forcing is very helpful for training VAEs
- SELFIES is a superior molecular grammar to SMILES especially for generating syntactically valid molecules.