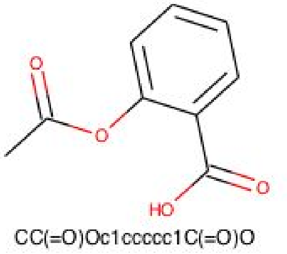
CognitiveChem

Drug Design with RNNs and Molecular Embeddings

Drug Discovery

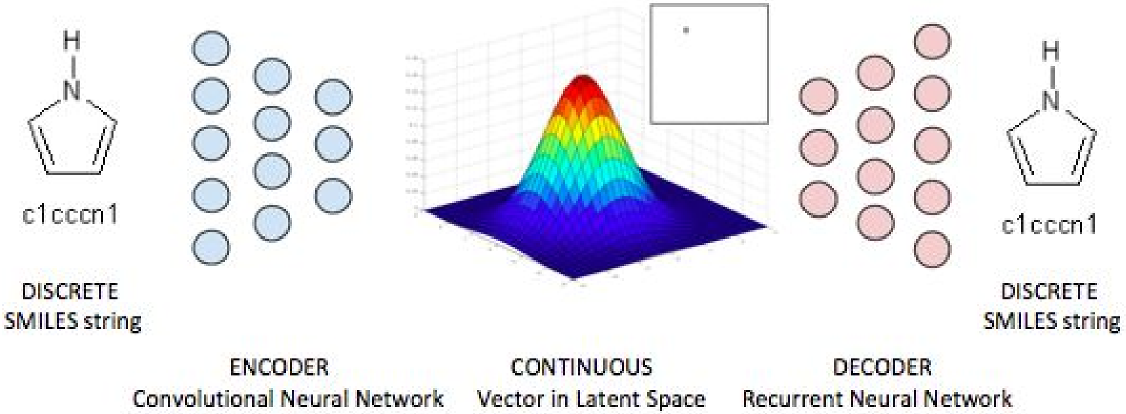
- Finding novel molecules with desirable properties is critical to chemical and pharmaceutical engineering
- Optimization in molecular space is challenging
- We implemented two papers on the topic
  - Rafael Gomez-Bombarelli et al. “Automatic chemical design using a data-driven continuous representation of molecules”
  - Segler et al. “Generating Focussed Molecule Libraries for Drug Discovery with Recurrent Neural Networks”

- Dataset from ChEMBL
  - 1.6 million molecules in SMILES format
- Simplified Molecular-Input Line-Entry System
  - Describes graph structure of molecule
  - Side chains denoted with parentheses
  - Cycles connected by integers



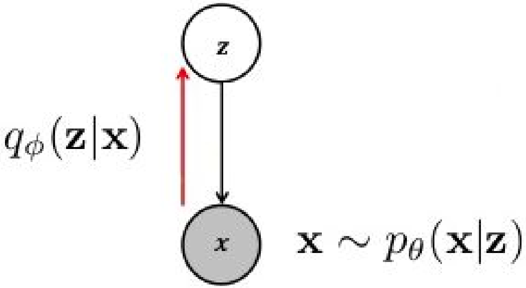
- RDKit for preprocessing
  - Verifying SMILES represent valid molecules
  - Determining special properties (drug-like, etc.)
  - Analysis and visualization

VAE to learn a continuous representation



Model structure

- Encoder : 3 convolutional layers
- Decoder : 3 recurrent layers with 501 GRU cells each
- Latent space : diagonal Gaussian with 292 dimensions

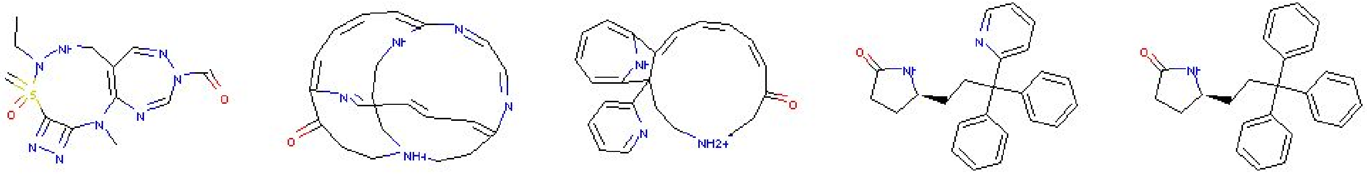


$q_\phi(\mathbf{z}|\mathbf{x})$
z
x
$\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{z})$

$$L(x;\theta;\phi) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x)||p(z))$$
$$\leq \log p(x)$$

Results for the VAE approach

- Reconstruction accuracy
  - 99% character-level
  - 70% molecule-level
- VAE objective
  - Latent space fits the Gaussian prior well
  - Allows interpolation between molecules in the latent space
- Future work
  - Regression model with latent space as input
  - Perform gradient ascent w.r.t input to find the best regions of the latent space

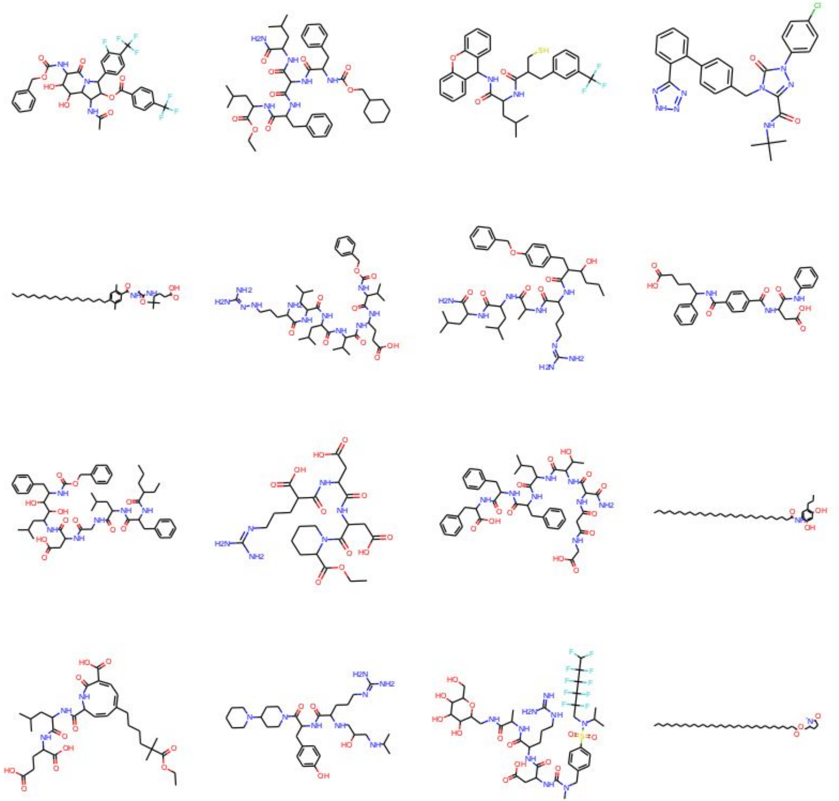


Novel molecule generation with LSTM model

- Model Structure
    - 3 stacked LSTM layers, 1024 neurons each
    - 64-timestep sequences from SMILES
    - Roughly 21 million parameters
- Batches of 128, drawn uniformly
- 100 molecules generated every 500 batches to check progress
    - Generation stochasticity controlled by temperature parameter

- Trained on 80% of full dataset
    - 1.3 million molecules
    - Roughly 70 million training sequences
    - Results
        - 75.9% valid
        - 40.4% drug-like
- Fine-tuned on lead-like molecules
    - 3% of full dataset
    - Roughly 100K sequences
    - Results
        - 93.5% valid
        - 92.8% drug-like

Full Dataset



Lead-like