ECO-AI HACKATHON TASK 1

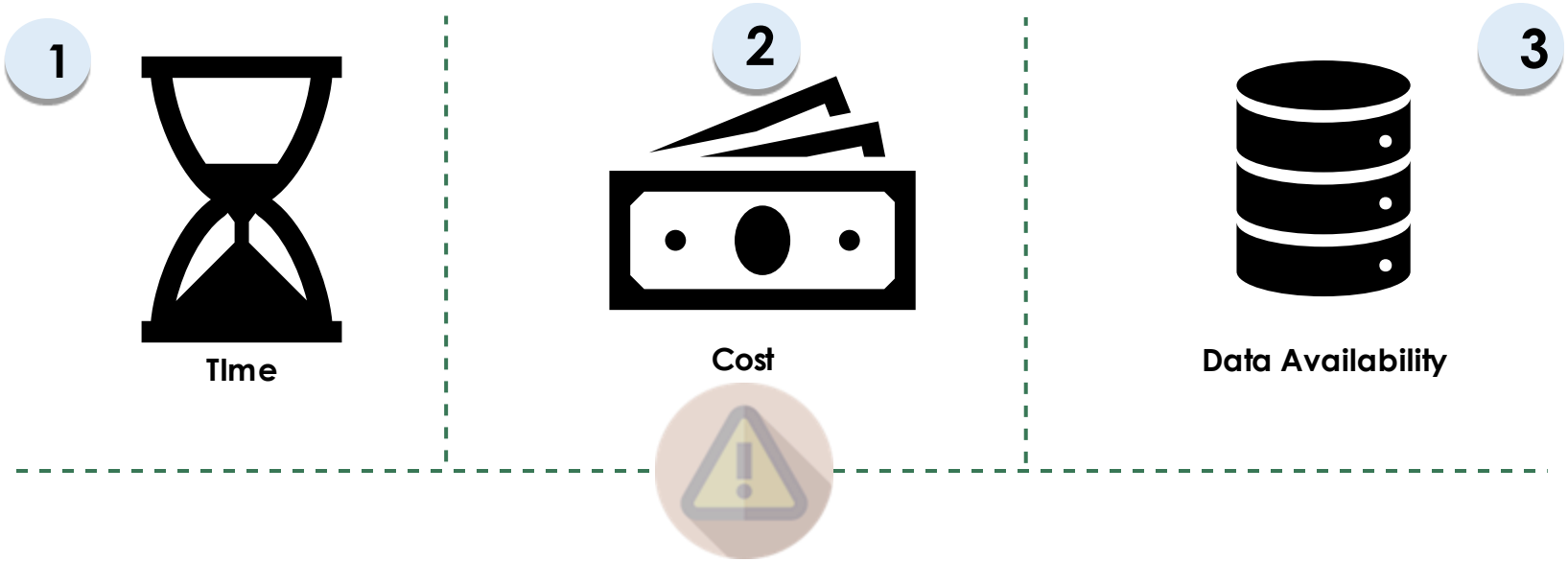
Auto-Encoders for molecules representation and property prediction

AUTOBOTS TEAM
Abu Abu, Thomas Bernet, Rohit Murali, Minghui Ye, Amin Zarei

INTRODUCTION

- Developments in molecule representation, property prediction, and descriptor-based molecular generation enhance CO2 solvent screening and design with machine learning.
- Refined representation aids in estimating CO2 absorption capacity in regression settings, particularly useful with scarce training data.

PROBLEM STATEMENT



- It is time consuming and costly to synthesize chemicals (energy, solvents, materials)
- Limited experimental data availability poses challenges, leading to exploration of alternative strategies like using molecule databases and combinatorial-based approaches for pre-training and transfer learning.
- Software for chemical generation can improve learning efficacy with limited datasets.

APPROACH: PROCESS DESCRIPTION

- Train VAEs with a dataset of 20,938 amines from ZINC and QM9 databases for molecular design.
- Objective: Develop encoder-decoder models with continuous representation akin to architecture in Figure 1.
- Carry out tasks like molecule reconstruction and predicting CO2 absorption capacity descriptors.

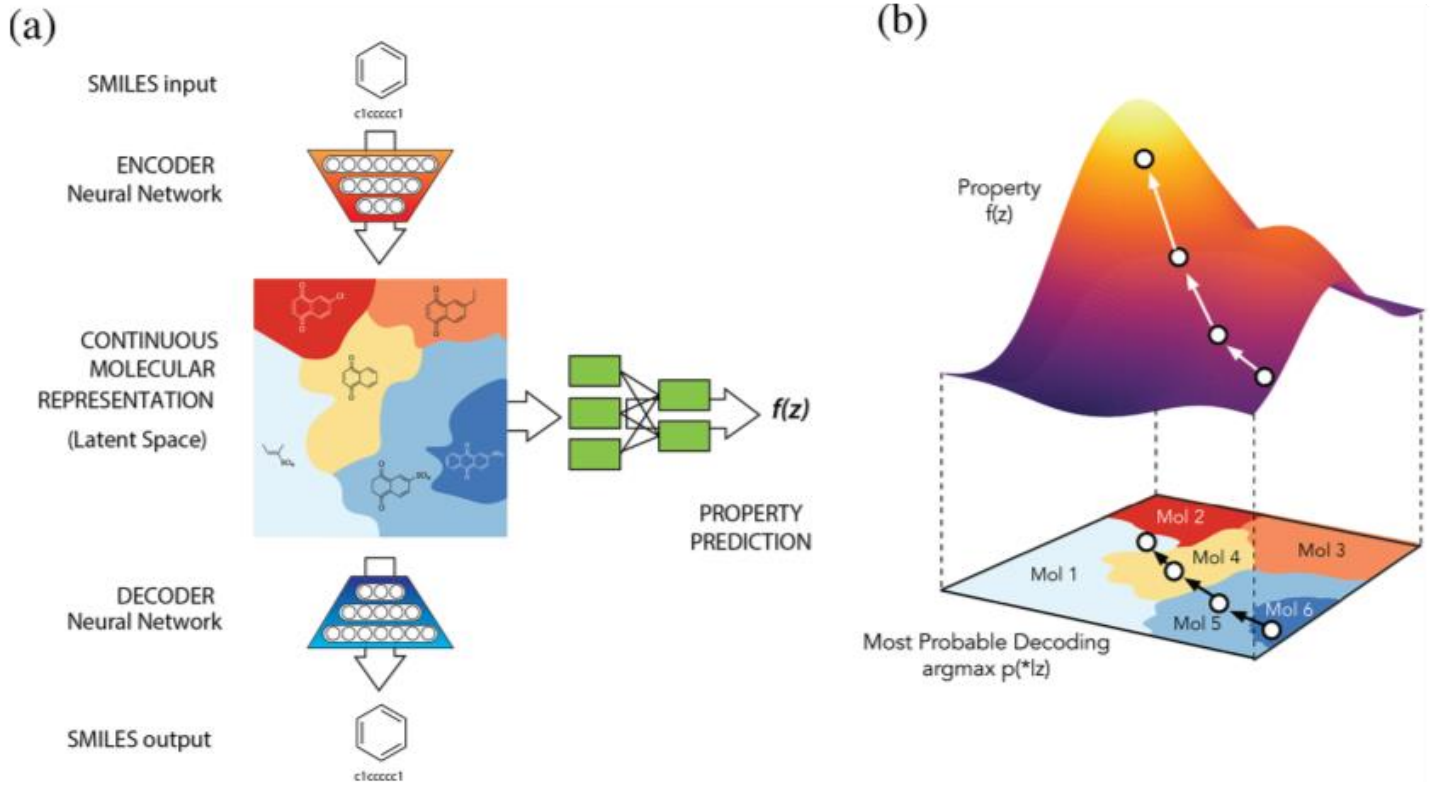


Fig. 1: (a) A diagram of the autoencoder used for molecular design, including the joint property prediction model (b) Gradient-based optimization in continuous latent space. [Gómez-Bombarelli et al., 2018]

DATASETS

1

zinc_small.txt (723.95 kB)

This preview is truncated due to the large file size. Create a Notebook or download this file to see the full content. **Download**

```
idx,smiles
1,N[C@H](Cc1ccc(Cl)cc1)C(=O)O
2,N[C@@H](Cc1ccc(O)c(I)c1)C(=O)O
3,N[C@H](Cc1ccccc1)C(=O)O
4,N[C@H](Cc1ccc(O)cc1)C(=O)O
5,CC(=O)NCCc1c(Cc2ccccc2)[nH]c2ccccc12
6,N[C@@H](Cc1ccc(Br)cc1)C(=O)O
7,N[C@H](Cc1ccc(Br)cc1)C(=O)O
8,CC(C)[C@H](N)C(=O)Nc1ccc2ccccc2c1
9,N[C@H](C(=O)O)[C@H](O)c1ccccc1
```

20938

unique values

2

0SelectedSMILES_QM9.csv (3.07 MB)

Detail | Compact | Column — 2 of 2 columns

# idx	smiles
1	C
2	N
3	O
4	C#C
5	C#N
6	C=O
7	CC

132040

unique values

RESULTS

Dataset 1 – Zinc_small	Validity	Diversity	Reconstruction Accuracy
SMILES	17.1%	4.6%	84.9%
SELFIES	72.1%	37.1%	85.2%

Dataset 1 – QM9 Dataset	Validity	Diversity	Reconstruction Accuracy
SMILES	20.7%	3.0%	74.8%
SELFIES	89%	25%	90%

Challenges: Sampled molecules from latent space

- 'CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
  CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC’
- 'N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\
  N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N
  \\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\
  \N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\
  N\\N\\N\\N\\N\\N\\N\\N\\N\\N\\N'
- 'Br/Br’,
- 'NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
  NNNN'

Opportunities for Future Work

- Predict CO2 absorption capacity descriptors of regenerated molecules
- Compare other VAE approaches such as CGVAE, Grammar VAE