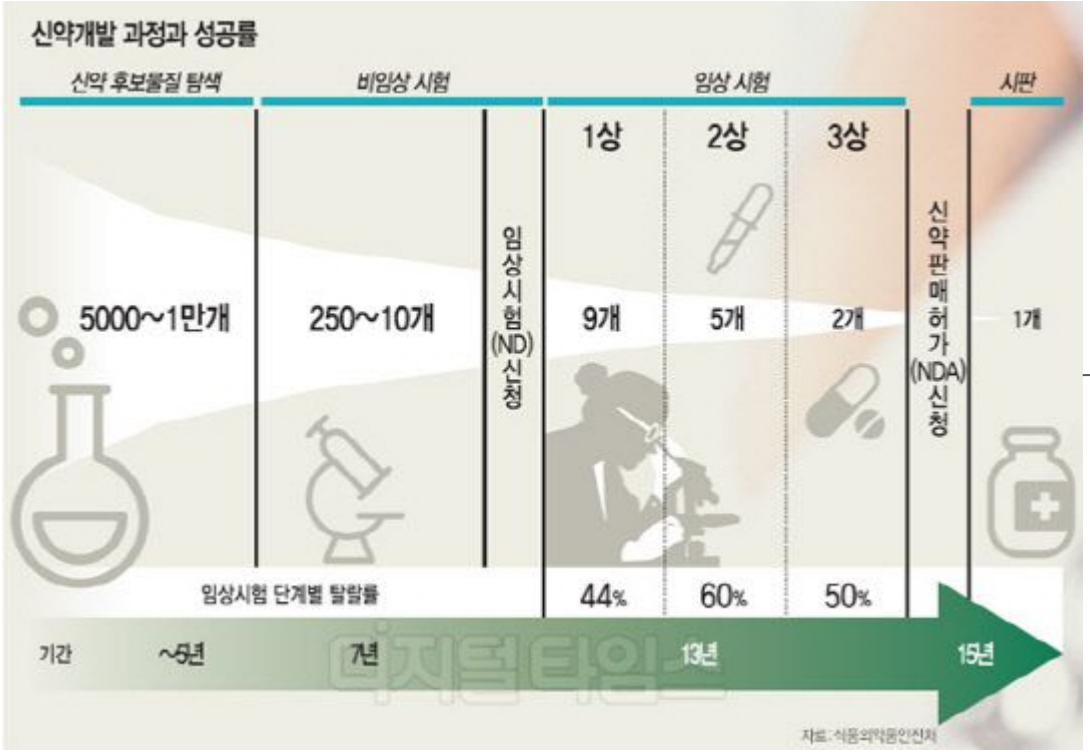
Automating chemical Design a Data-Driven Continuous Representation of Molecules

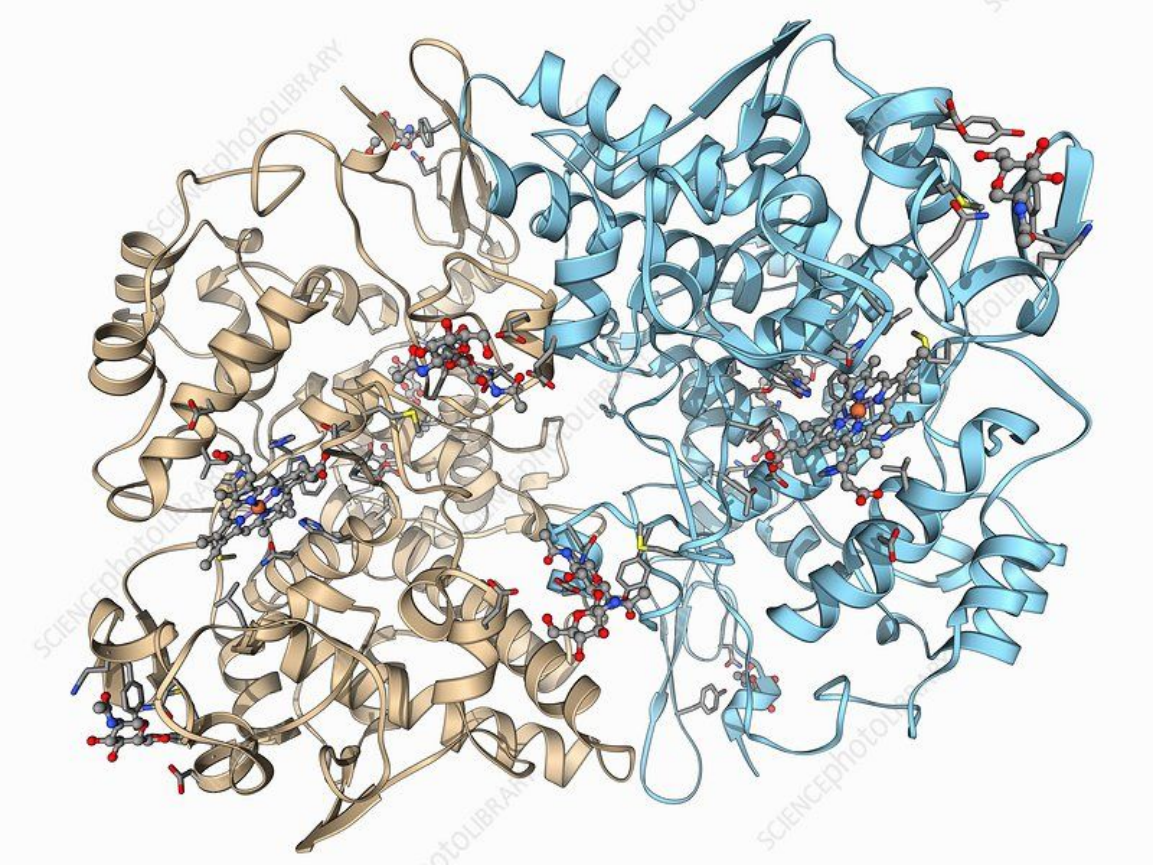
김 준 태

Drug Discovery

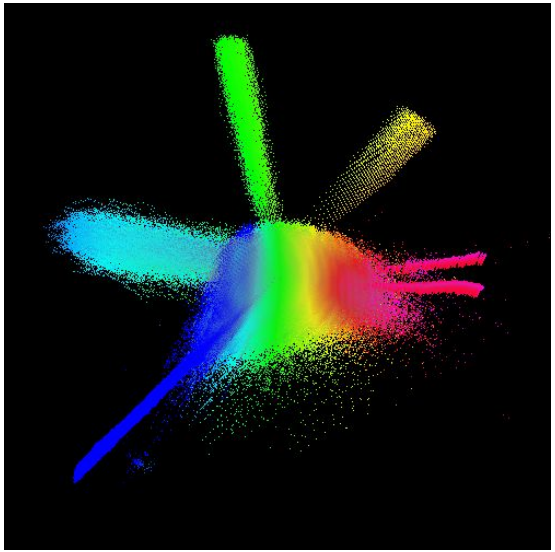


최소 **1300억원**

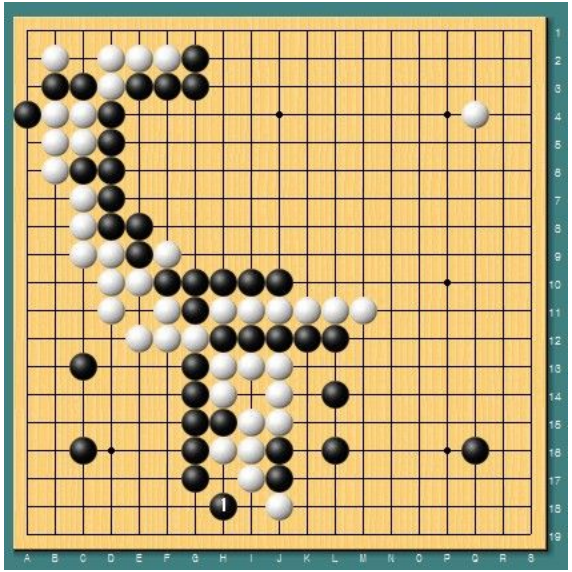
Drug Discovery



Chemical space



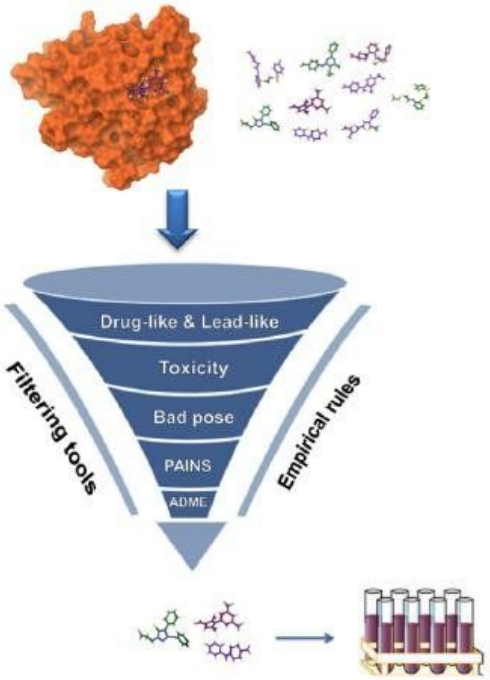
$>$



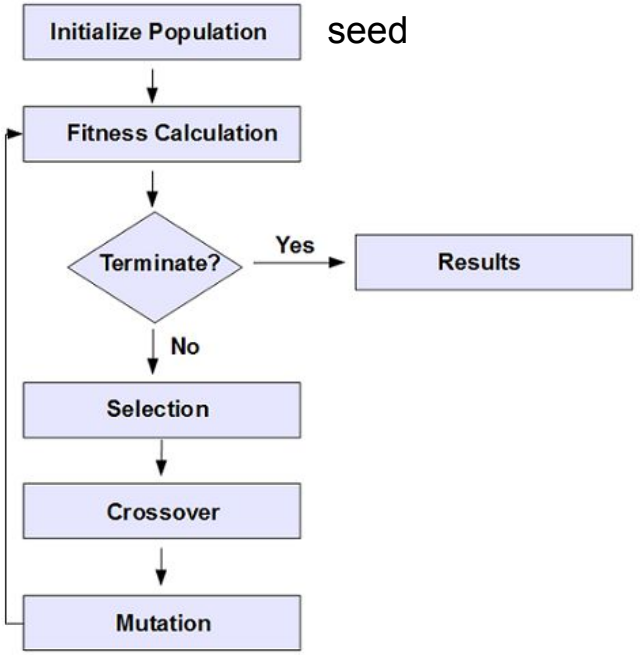
10^{60}

2^{170}

Research Background

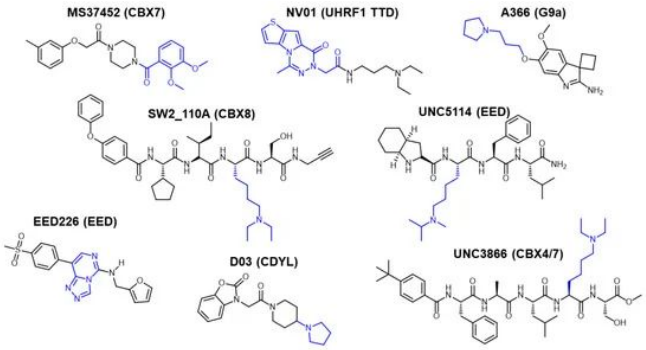


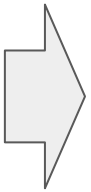
virtual screening

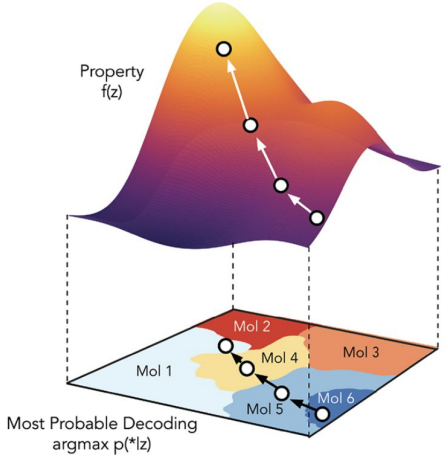


Genetic Algorithm

Contribution

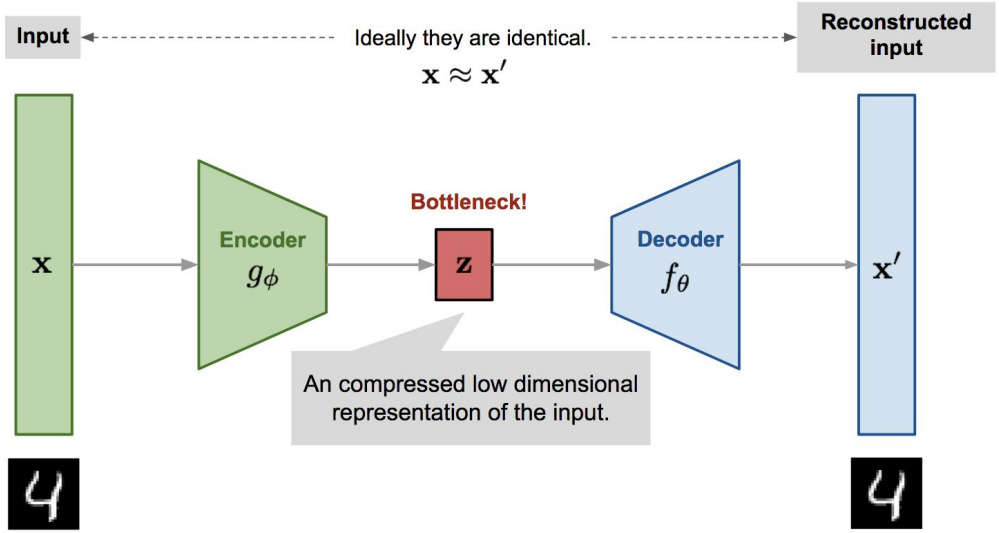


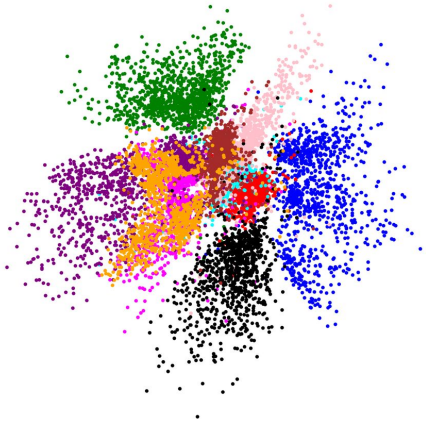




- 연속적이지 않은 데이터들을 **연속적인 latent space**로 매핑하자
  - **Hand craft rule**이 필요없다
  - Gradient를 이용하여 **chemical space**를 **search할 수 있음**
  - 상대적으로 **적은 데이터**를 이용하더라도 **큰 chemical space**를 만들수있다.

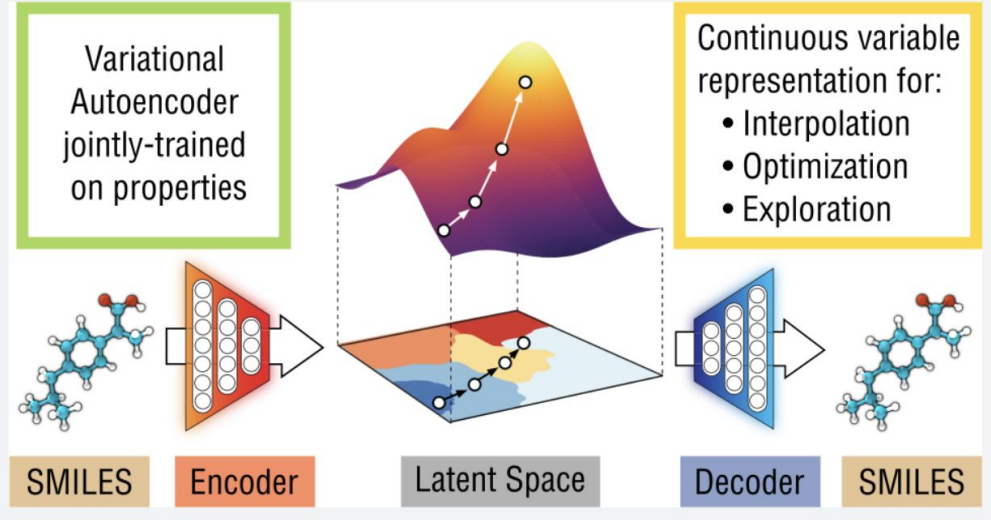
Contribution





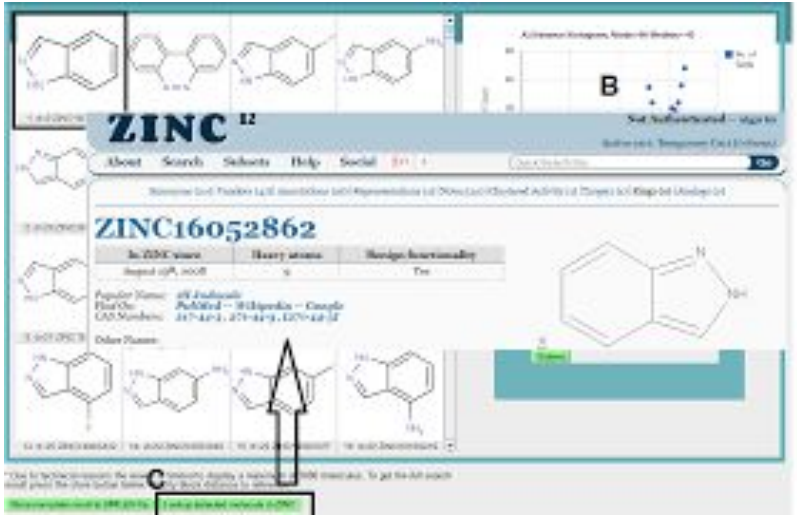
연속적이지 않은 숫자 이미지 사이 사이들을 **encoding**을 통해 **latent space**로 만들어 **연속적** 이게 만들었다.

Objective



- Molecule이 input으로 들어가고 output도 molecule로 나오는 VAE를 학습
- VAE를 학습하면 latent space가 생성된다.
- 학습된 latent space에서 gradient를 이용하여 원하는 molecule을 생성하자

Dataset

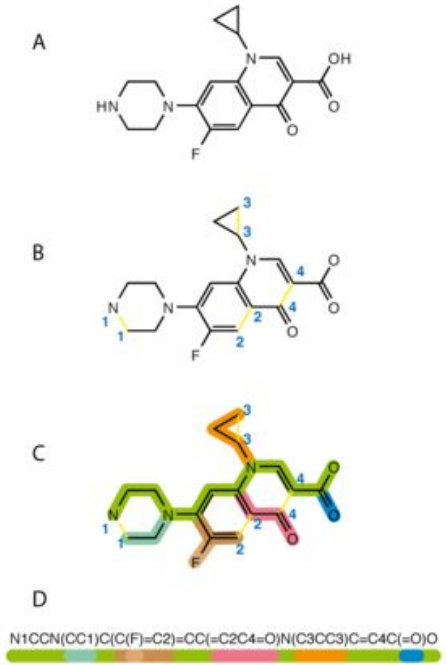


- 상업적으로 이용가능한 분자 정보 있음
- 어느정도 화학적 요소가 만족되는 분자들
- 250,000개의 데이터를 샘플링해서 사용

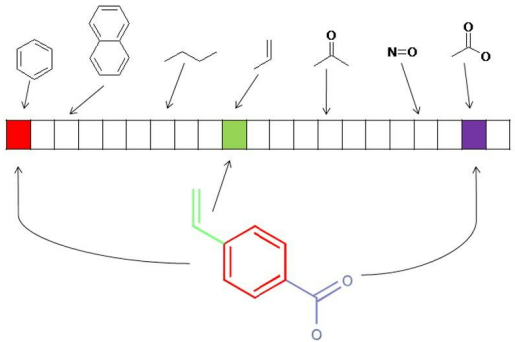
QM9 Dataset

- 9개 이하의 atom으로 이루어진 분자 데이터셋
- 108,000개의 분자 데이터

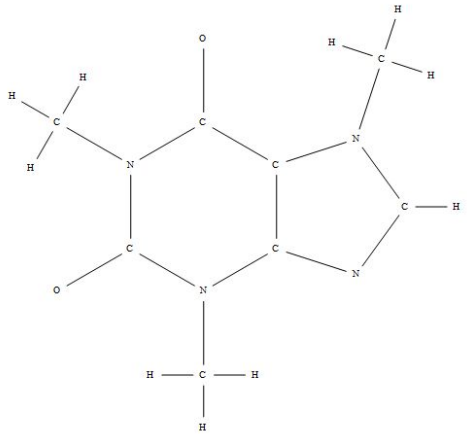
Data Description



smiles

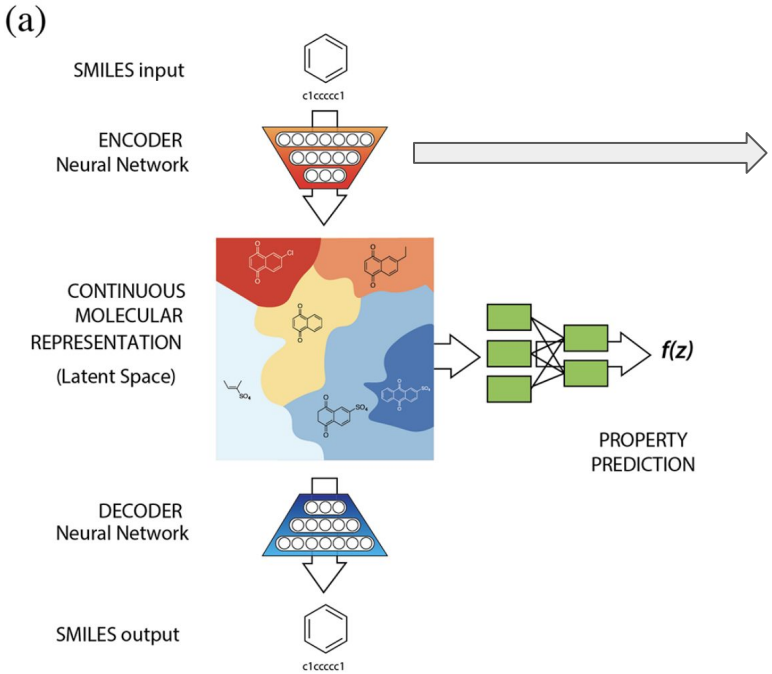


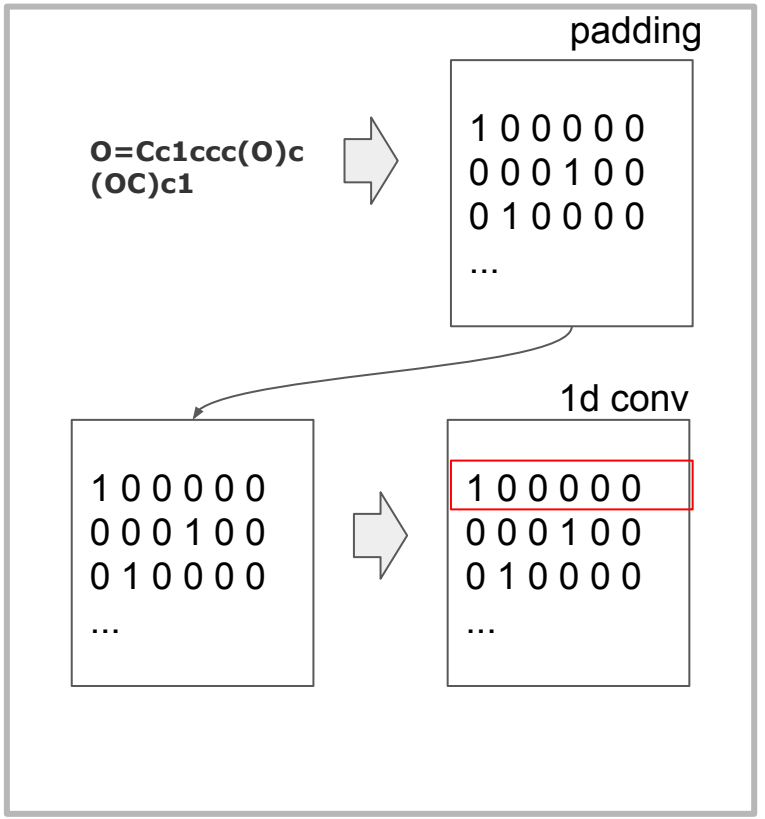
fingerprint



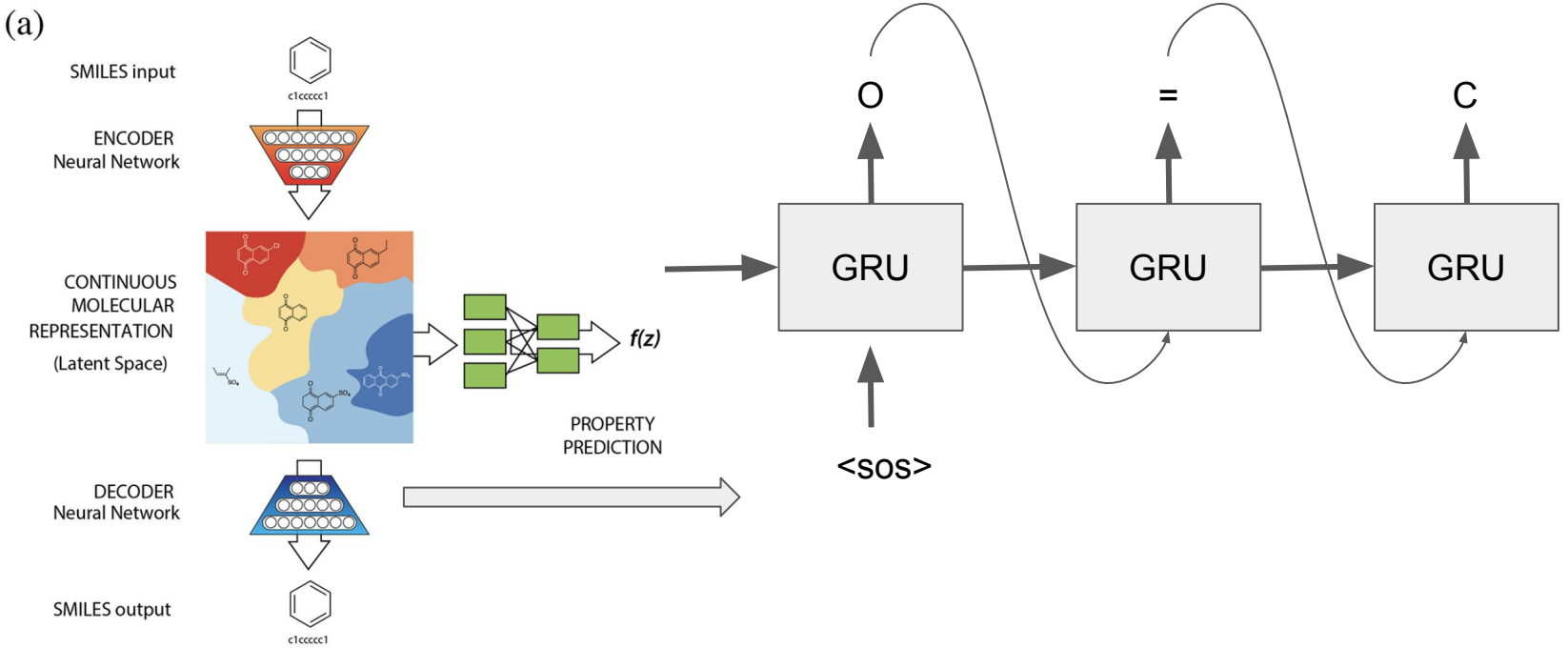
graph

Model (Encoder)

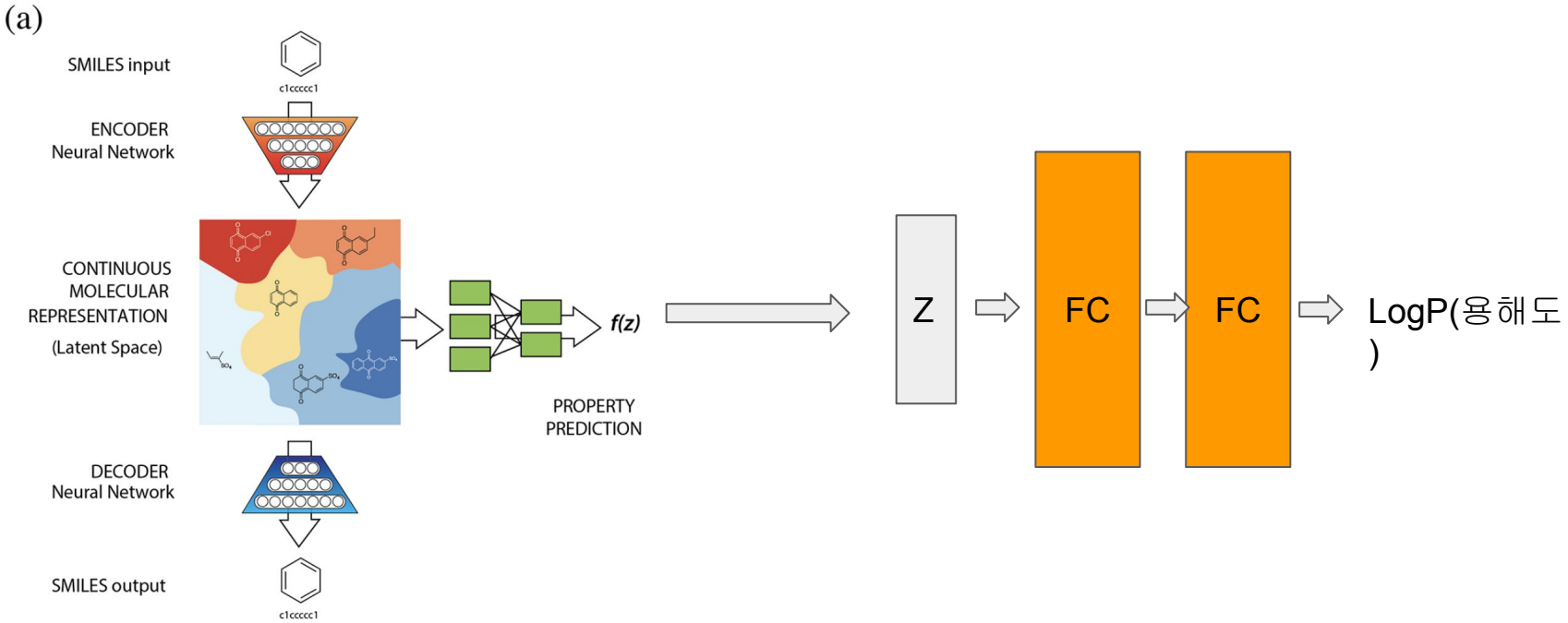




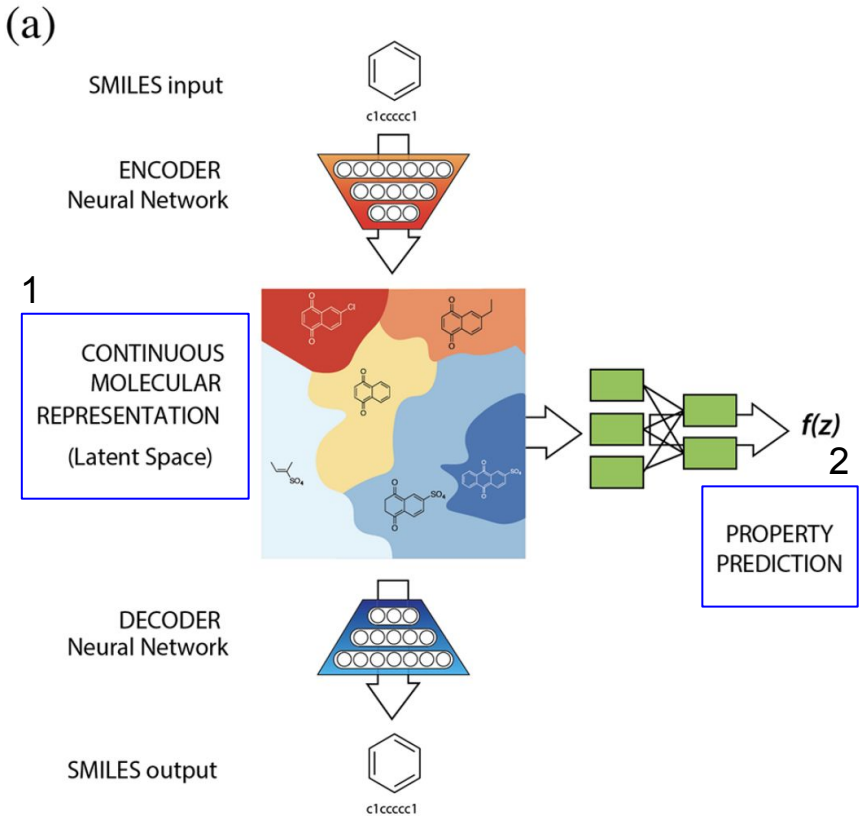
Model (Decoder)

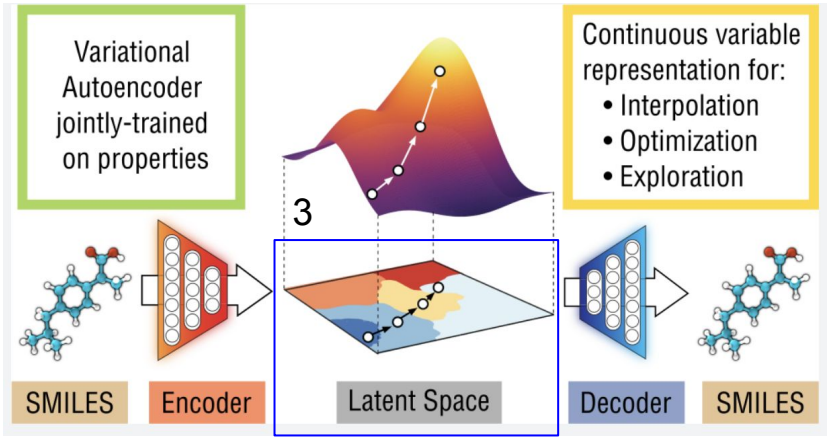


Model (Property Prediction)

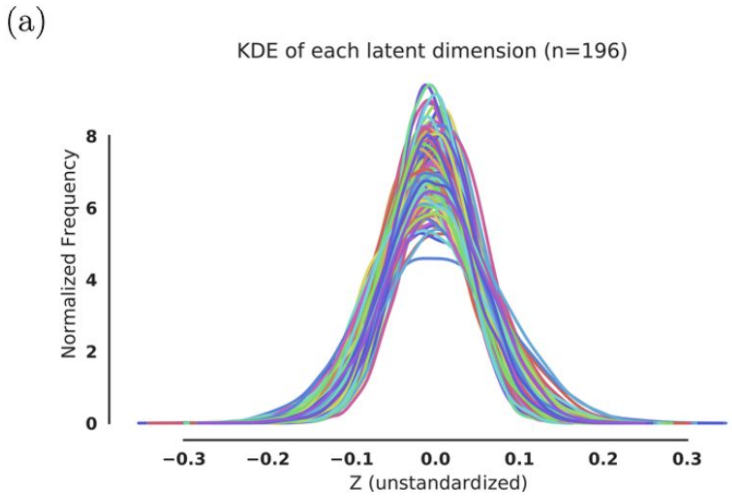


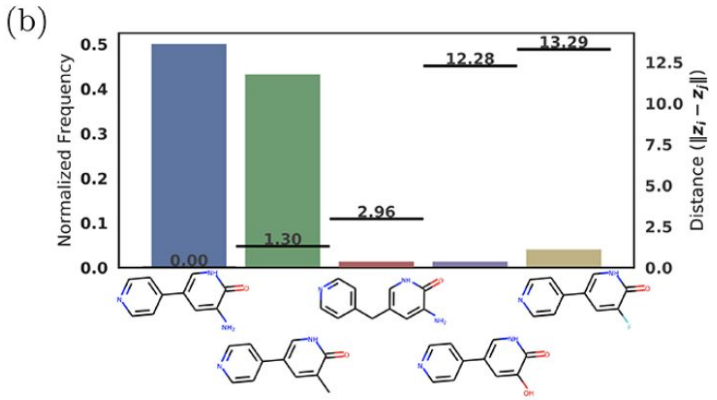
Model


(a)
SMILES input
c1ccccc1
ENCODER
Neural Network
1
CONTINUOUS
MOLECULAR
REPRESENTATION
(Latent Space)
f(z)
2
PROPERTY
PREDICTION
DECODER
Neural Network
SMILES output
c1ccccc1


Variational
Autoencoder
jointly-trained
on properties
Continuous variable
representation for:
• Interpolation
• Optimization
• Exploration
3
SMILES
Encoder
Latent Space
Decoder
SMILES

Mapping to latent space

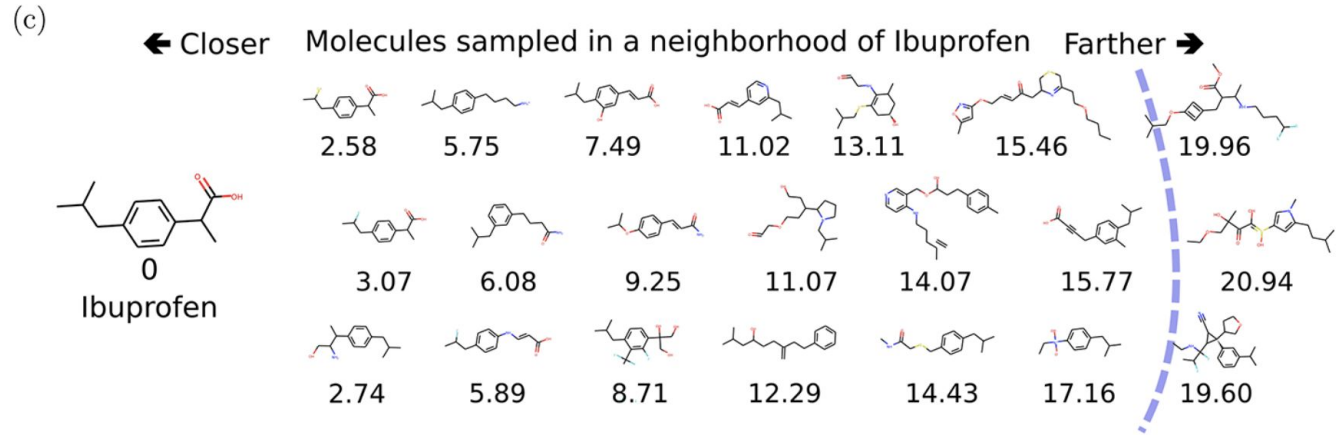


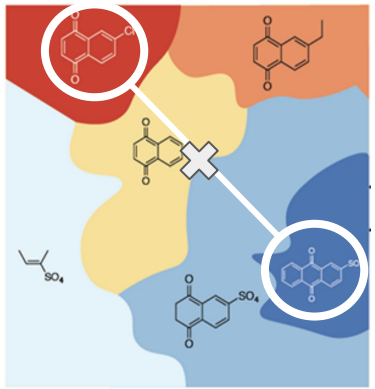


Dataset	ZINC	QM9
Decoding Prob	73.9	79.3

- **새로운 분자**를 생성하기 위해 **Gaussian Noise**를 추가함
- Gaussian Noise를 추가할수록 **Distance가 멀어짐**
- **유사한 분자** 구조일수록 **Distance가 가까움**

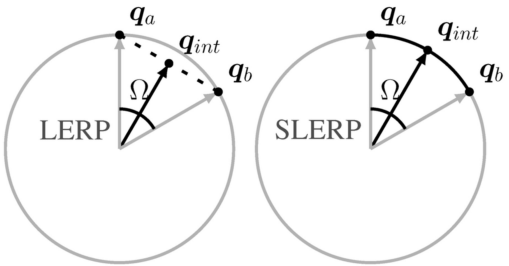
Molecule Distance



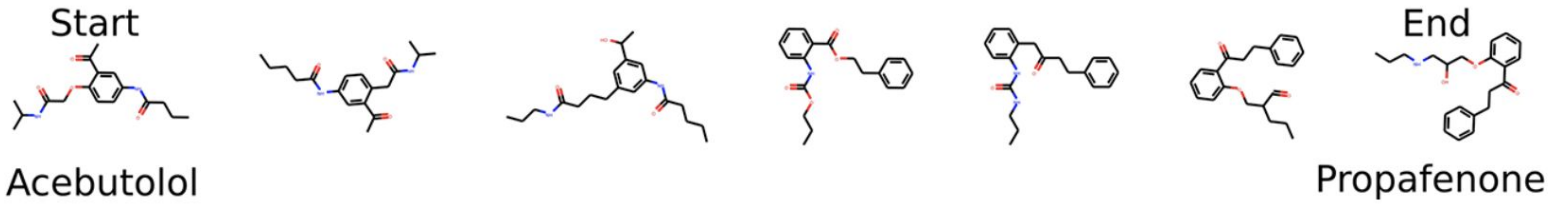


continuous한 representation이기 때문에 **interpolation**도 가능하다

Molecule Distance



- Linear Interpolation은 **너무 sparse**해서 **spherical interpolation** 사용함.

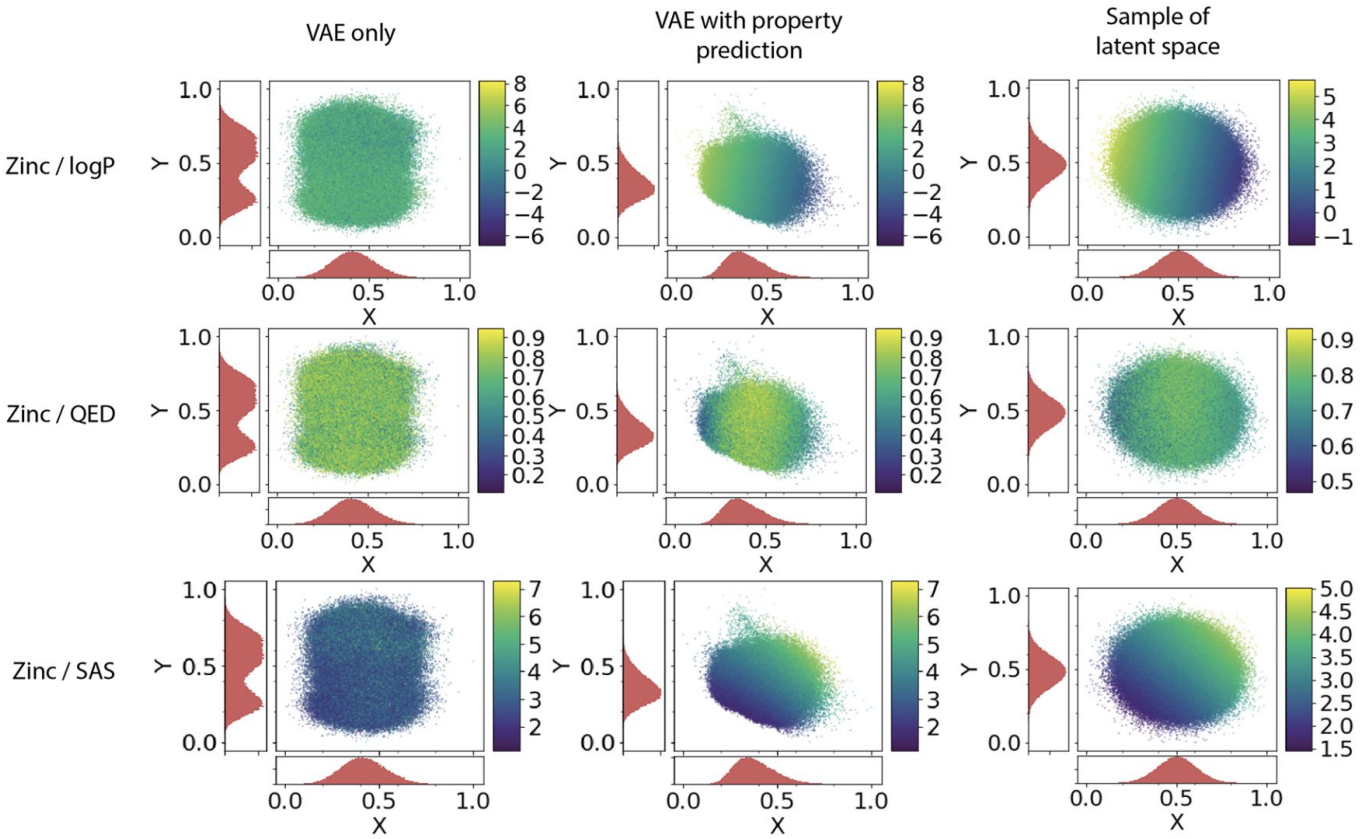


Experiments

Table 1. Comparison of Molecule Generation Results to Original Datasets

source[a]	data set[b]	samples[c]	logP[d]	SAS[e]	QED[f]	% in ZINC[g]	% in emol[h]
Data	ZINC	249k	2.46 (1.43)	3.05 (0.83)	0.73 (0.14)	100	12.9
GA	ZINC	5303	2.84 (1.86)	3.80 (1.01)	0.57 (0.20)	6.5	4.8
VAE	ZINC	8728	2.67 (1.46)	3.18 (0.86)	0.70 (0.14)	5.8	7.0
Data	QM9	134k	0.30 (1.00)	4.25 (0.94)	0.48 (0.07)	0.0	8.6
GA	QM9	5470	0.96 (1.53)	4.47 (1.01)	0.53 (0.13)	0.018	3.8
VAE	QM9	2839	0.30 (0.97)	4.34 (0.98)	0.47 (0.08)	0.0	8.9

Compare with Distribution



Results

Chemical space를 **continuous**하게 만들어 **exploring**할수 있다.

새로운 molecule을 **design**하는데 continuous한 chemical space 사용할 수 있다.

Results

- Molecule을 생성해도 여러 Properties를 만족해야 한다
  - 독성 등등
- 해당 논문의 데이터셋은 target과 관련없는 데이터셋
- valid한 molecule과 새로운 molecule이 나와도 대부분 특허에 걸림
- valid한 molecule과 특허에 걸리지 않는 molecule이 나와도 합성 불가
- valid한 molecule과 특허게 걸리지 않고 합성이 되어도 실험에서 필터링
- 위 조건을 모두 통과하면 후보 물질로 지정
- 하지만 동물실험, 임상실험에서 탈락