Lecture04.
Molecular design using deep generative model

HITS 임 재 창

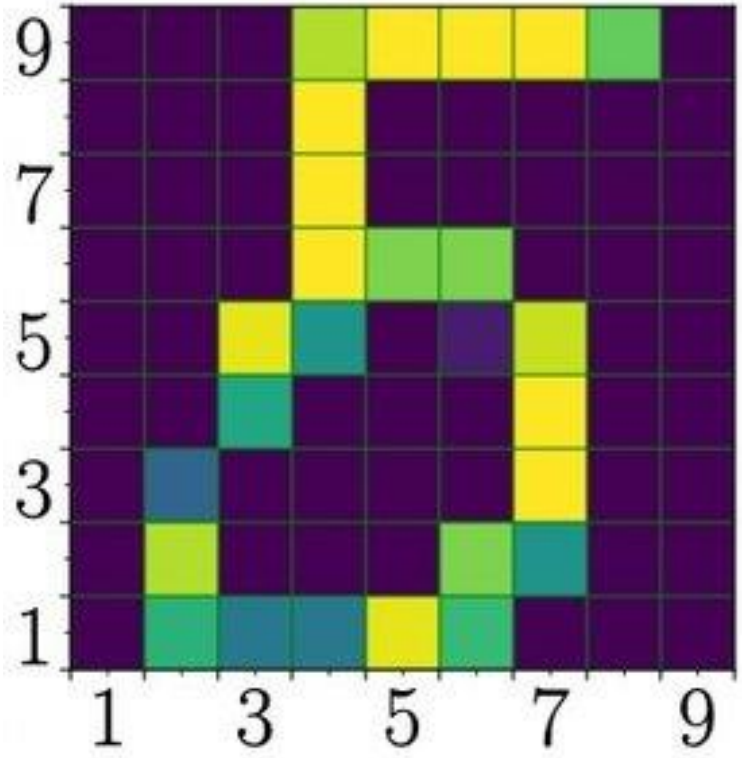
목차

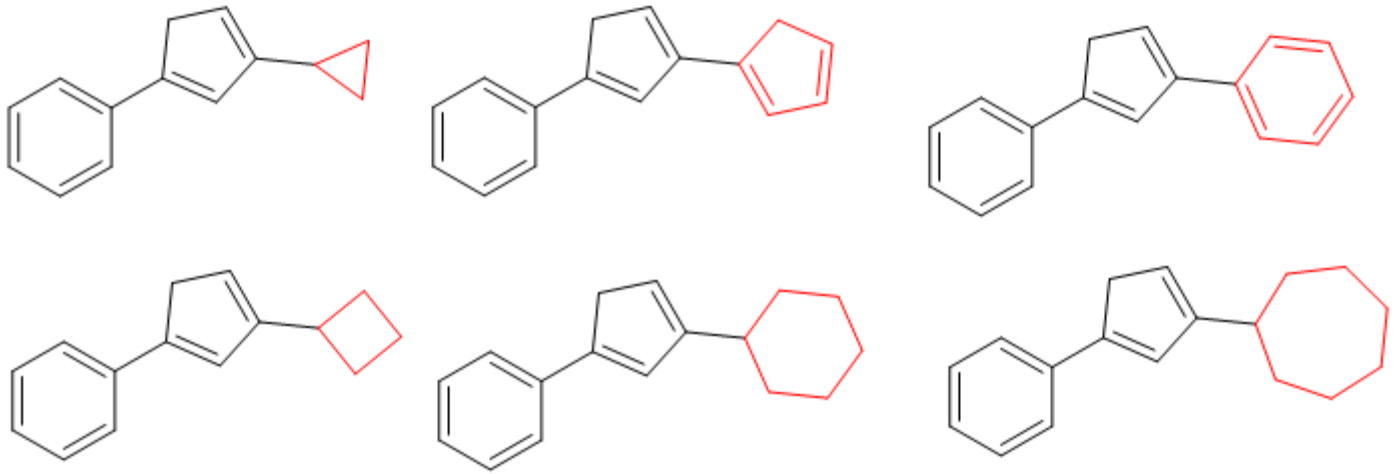
- Molecular design with language model
- Molecular design with variational autoencoder
- Molecular design with generative adversarial network (+adversarially regularized autoencoder)
- Molecular graph generative model
- Scaffold based vs de novo design in deep learning based molecular design

Molecular generative model

- Generative model이란? Data distribution, P(x) 을 학습하는 것

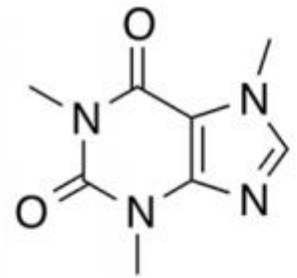
- Data distribution, P(x)란?





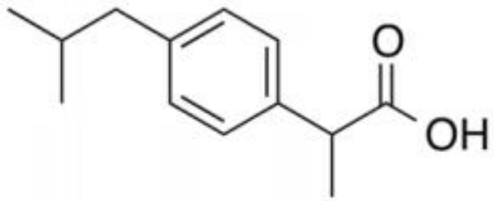
Molecular design using language model

- 주어진 smiles string piece로 부터 다음 character의 확률을 예측함



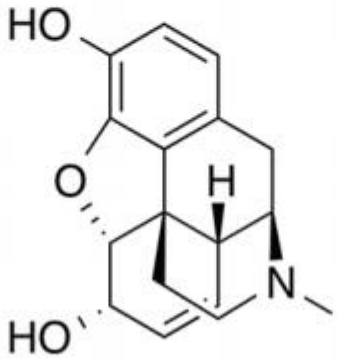
Caffeine

CN1c2ncn(C)c2C(=O)N(C)C1=O



Ibuprofen

CC(C)Cc1ccc(cc1)C(C)C(O)=O



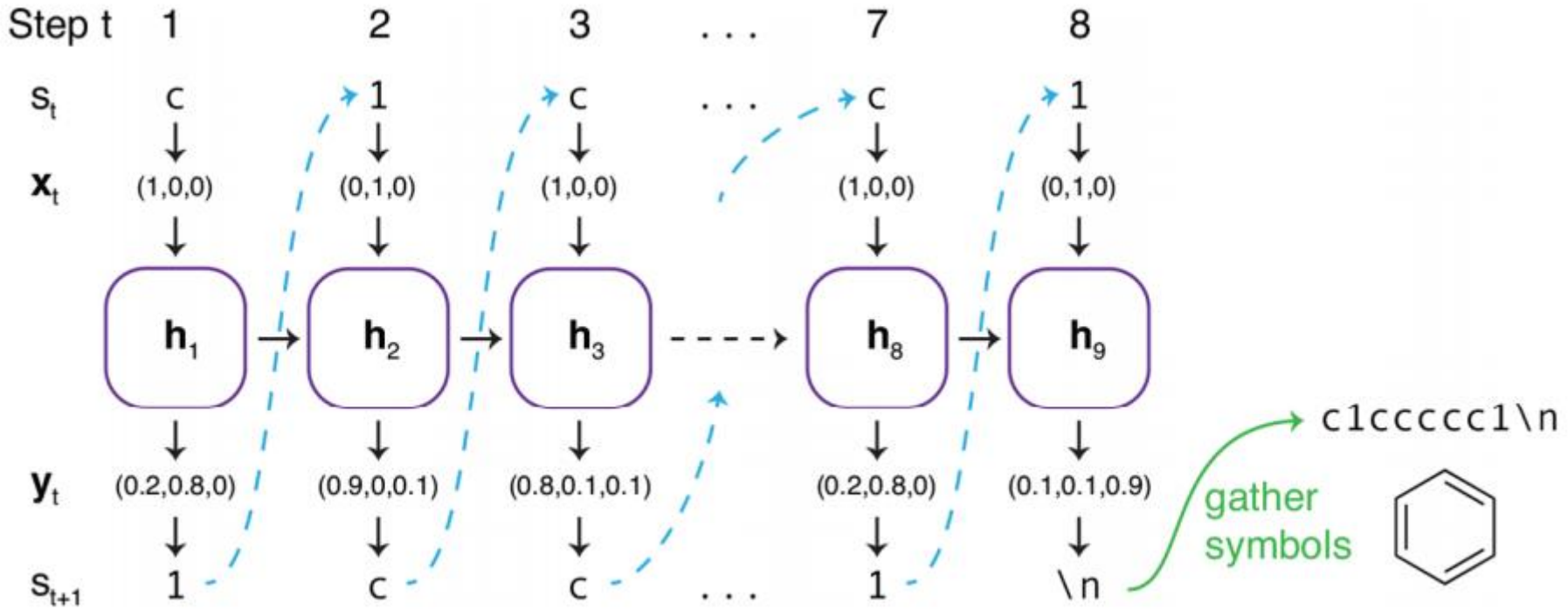
Morphine

[H][C@]12C=C[C@H](O)[C@@H]3Oc4c5c(C[C@H]1N(C)CC[C@@]235)ccc4O

ACS Cent. Sci. 2018, 4, 1, 120–131

Molecular design using language model

- 주어진 smiles string piece로 부터 다음 character의 확률을 예측함



ACS Cent. Sci. 2018, 4, 1, 120–131

생성모델 평가기준

- Validity: the ratio of the number of valid molecules to the number of generated samples. The validity was checked by using RDKit.[39]
- Uniqueness: the ratio of the number of unrepeated molecules to the number of valid molecules.
- Novelty: the ratio of the number of molecules which are not included in the training set to the number of unique molecules.
- Novel/sample: the ratio of the number of valid, unique, and novel molecules to the total number of generated samples.
- Diversity: $\left(1.0 - \frac{1}{N}\sum_{i,j>i} \text{similarity}_{ij}\right)$ for all N molecule pairs $(i, j > i)$ in the test set. The similarity between two molecules was computed with Tanimoto similarity[40] between their Morgan fingerprints[41] with radius of 4 and 2048 bits.

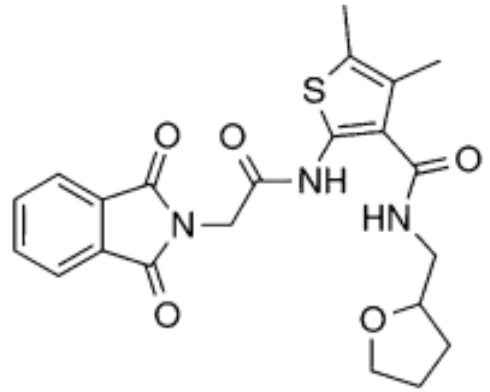
J. Chem. Inf. Model. 2020, 60, 1, 29–36

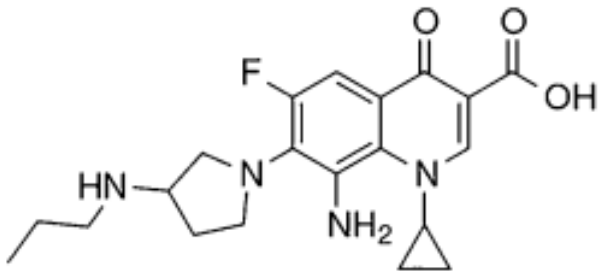
Results

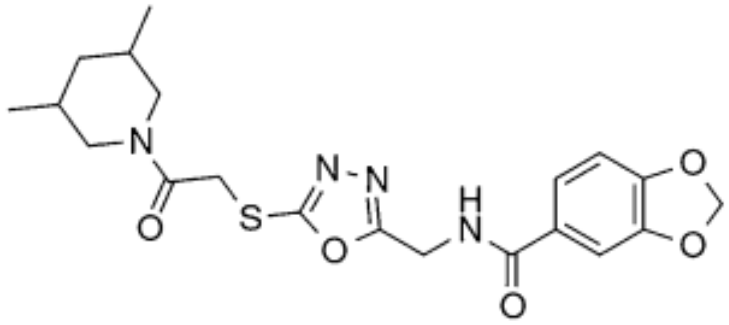
Batch	Generated Example	valid
0	Oc.BK5i%ur+7oAFc7L3T=F8B5e=n)CS6RCTAR((OVCp1CApb)	no
1000	OF=CCC2OCCCC)C2)C1CNC2CCCCCCCCCCCCCCCCCCCCCCC	no
2000	O=C(N)C(=O)N(c1occc1OC)c2ccccc2OC	yes
3000	O=C1C=2N(c3cc(ccc3OC2CCC1)CCCc4cn(c5c(Cl)cccc54)C)C	yes

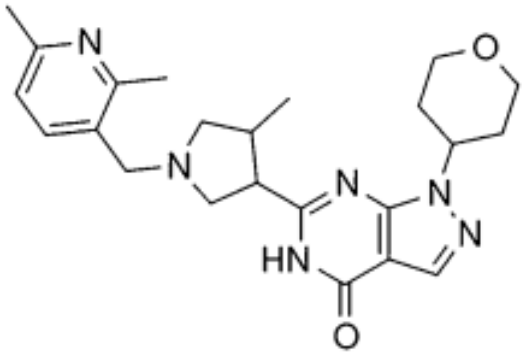
- 최종적으로 976,327번 생성시도, 이중 97.7% valid
- Training set과 겹치는 분자 제거 후: 864,880
- 중복분자 제거 후: 847,955

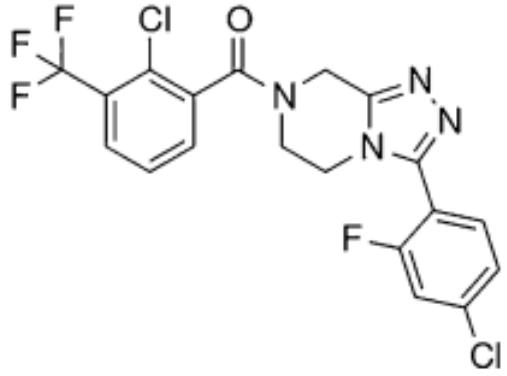
Results

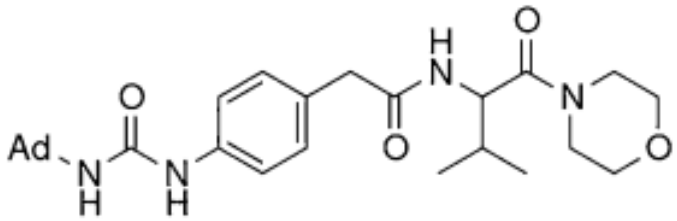

S
O
O
NH
HN
O
N
O
O


F
O
O
OH
HN
N
N
NH2

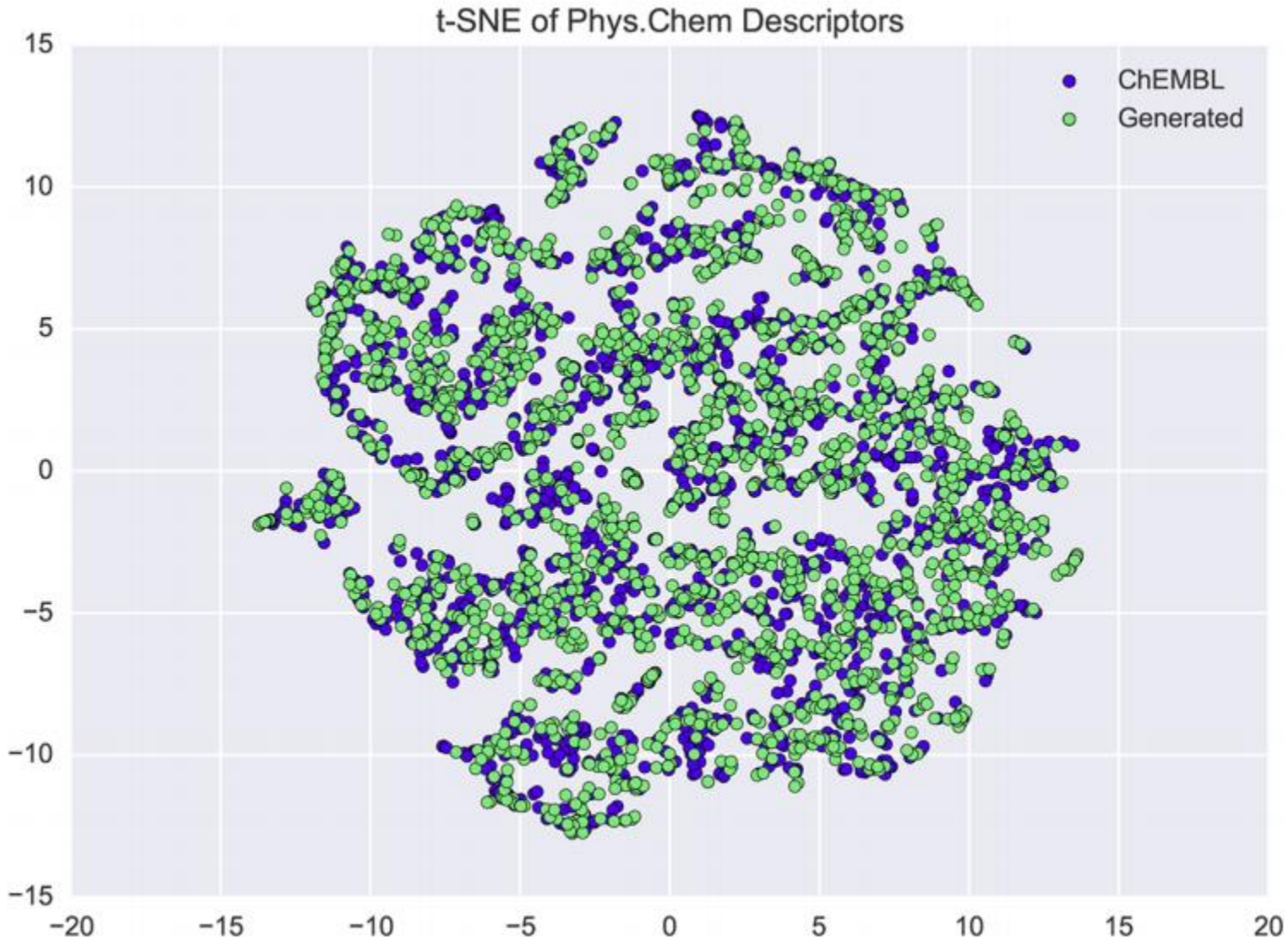

N
O
S
N
N
O
H
N
O
O
O


N
O
N
N
N
HN
N
O


F
F
F
Cl
O
N
N
N
F
Cl


Ad
N
H
O
N
H
H
N
O
O
N
O

Results


t-SNE of Phys.Chem Descriptors
ChEMBL
Generated
15
10
5
0
−5
−10
−15
−20
−15
−10
−5
0
5
10
15
20

Molecular design using language model의 장단점

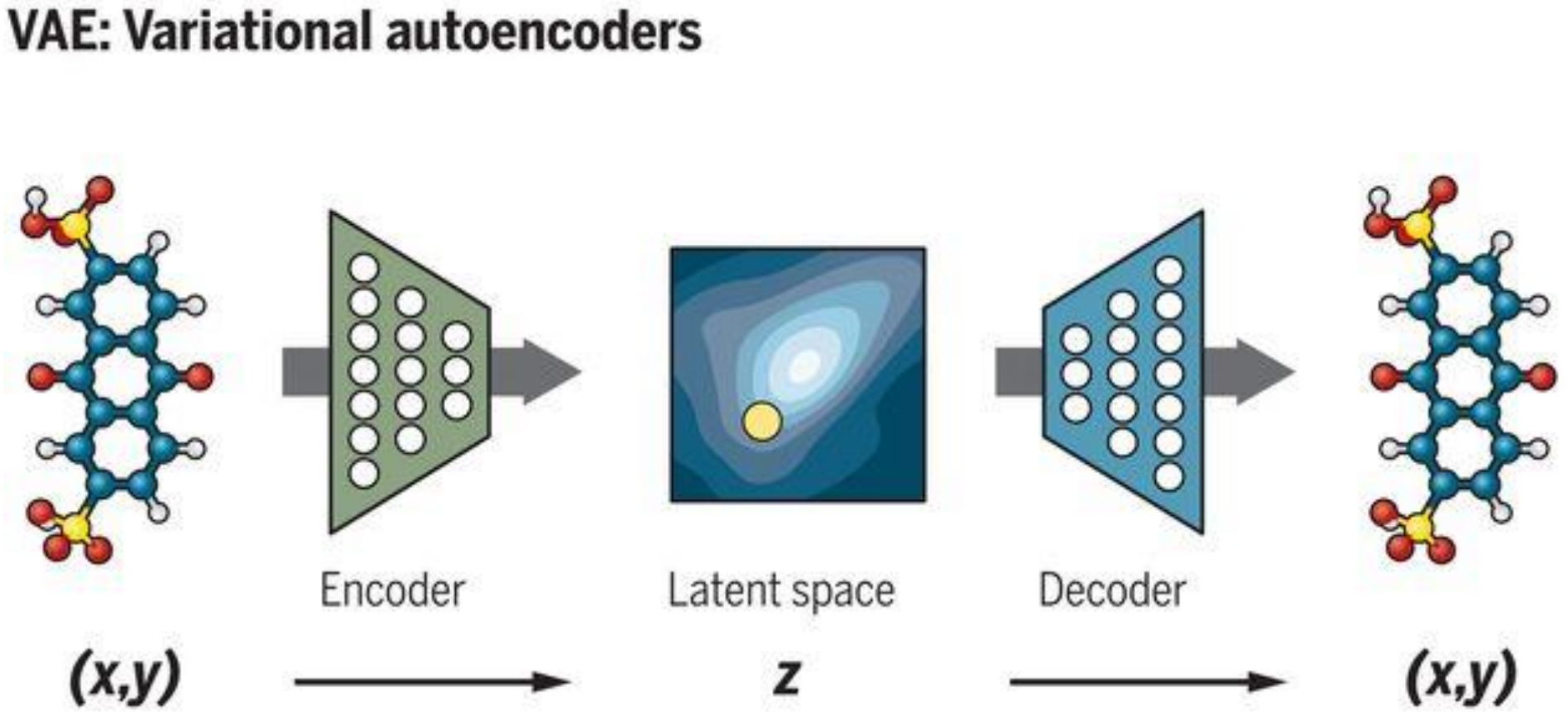
장점

- 구현하기 쉬움 (library가 잘 구축 되어있음)
- 학습이 쉬움

단점

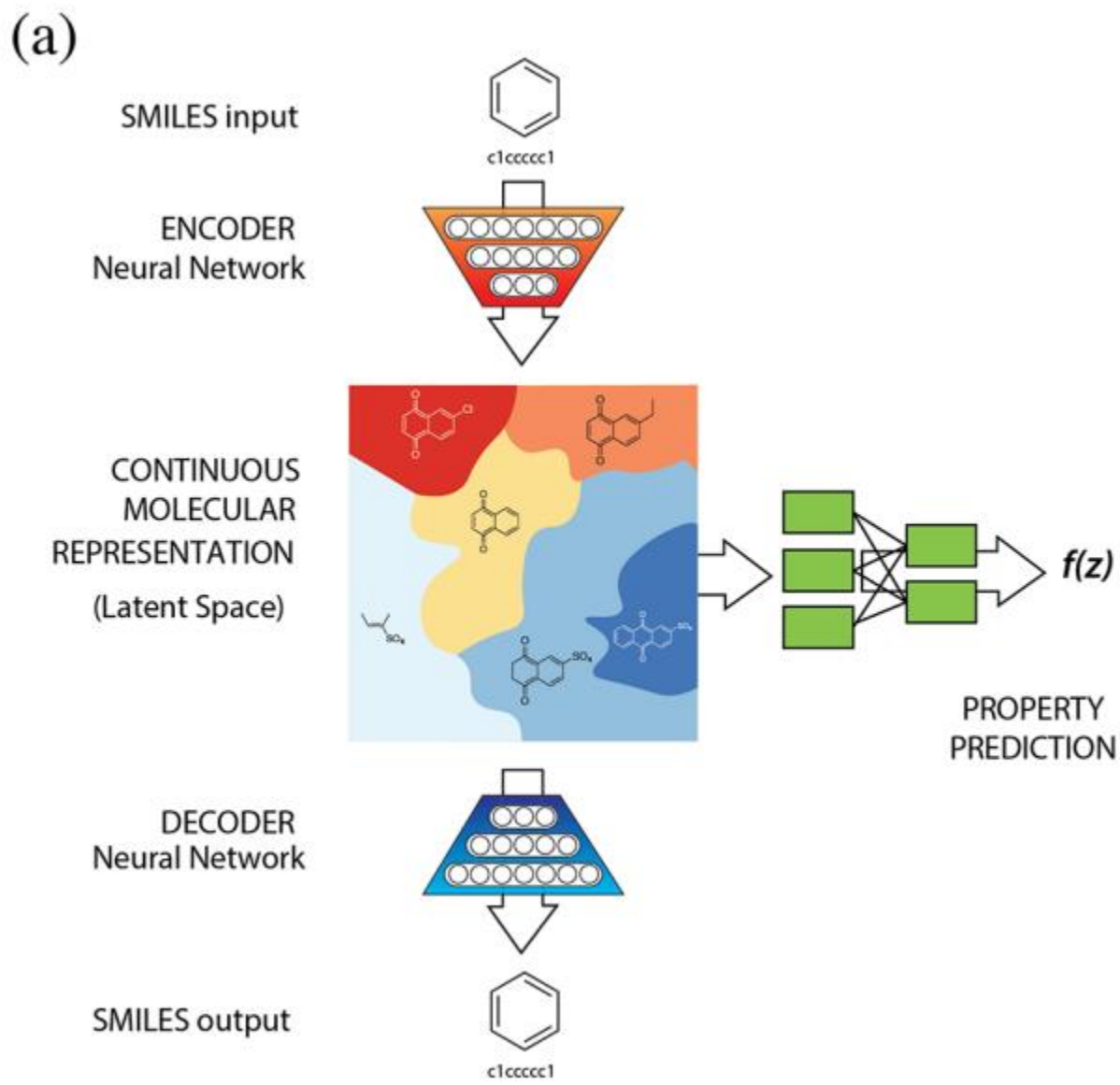
- Latent space 분석이 불가능 (latent vector modification이 안됨)

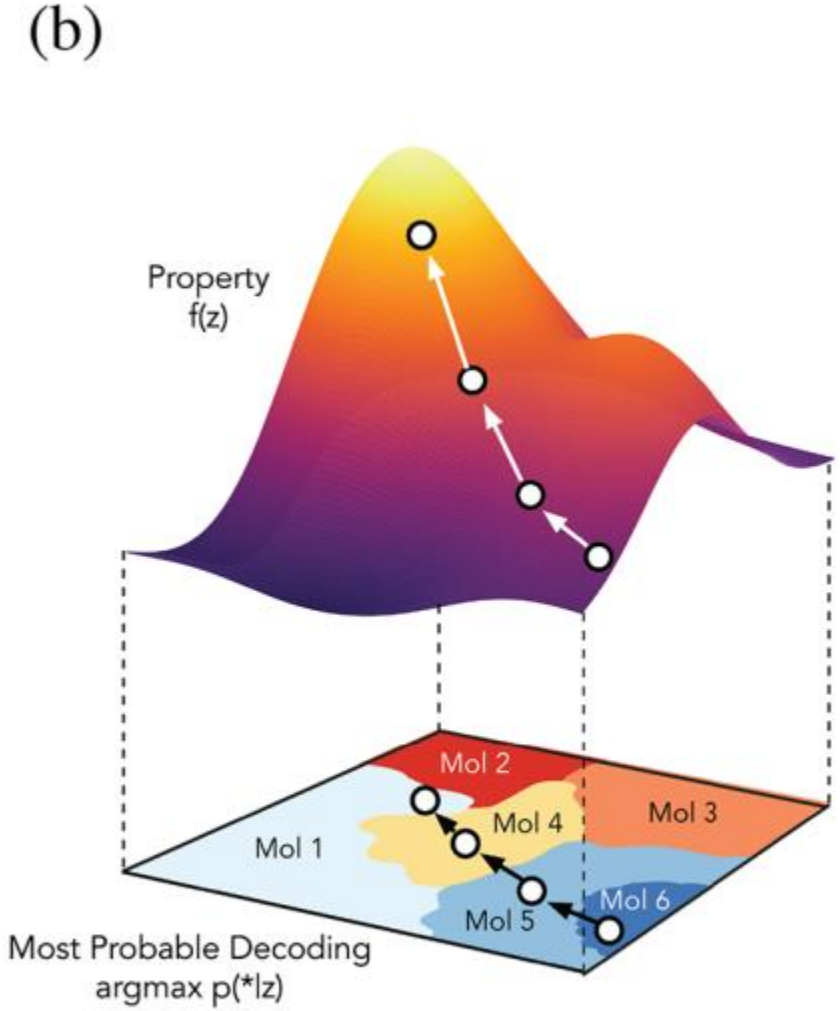
Molecular design using variational autoencoder



Science 27 Jul 2018: Vol. 361, Issue 6400,
pp. 360-365 DOI: 10.1126/science.aat2663

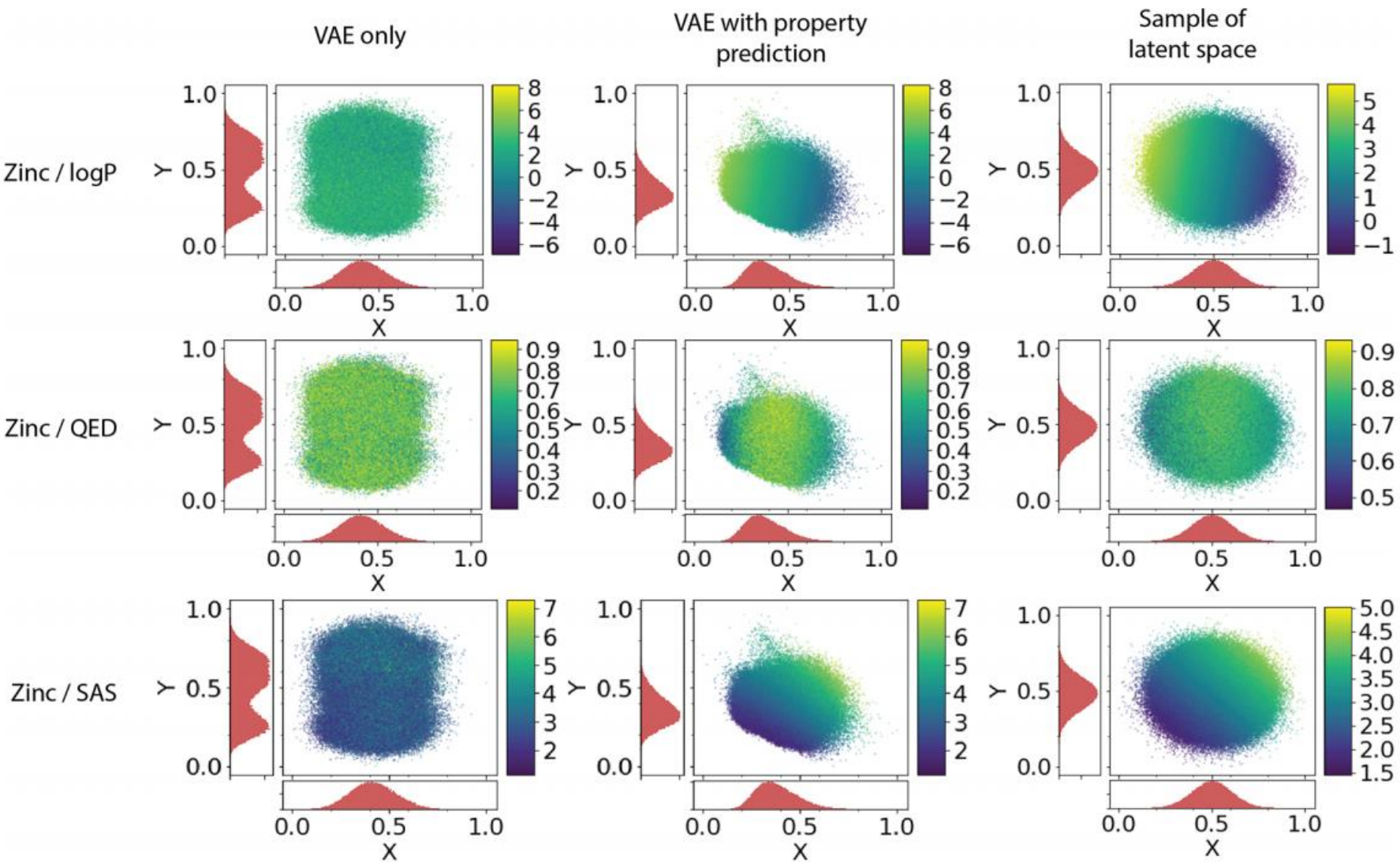
Molecular design using variational autoencoder





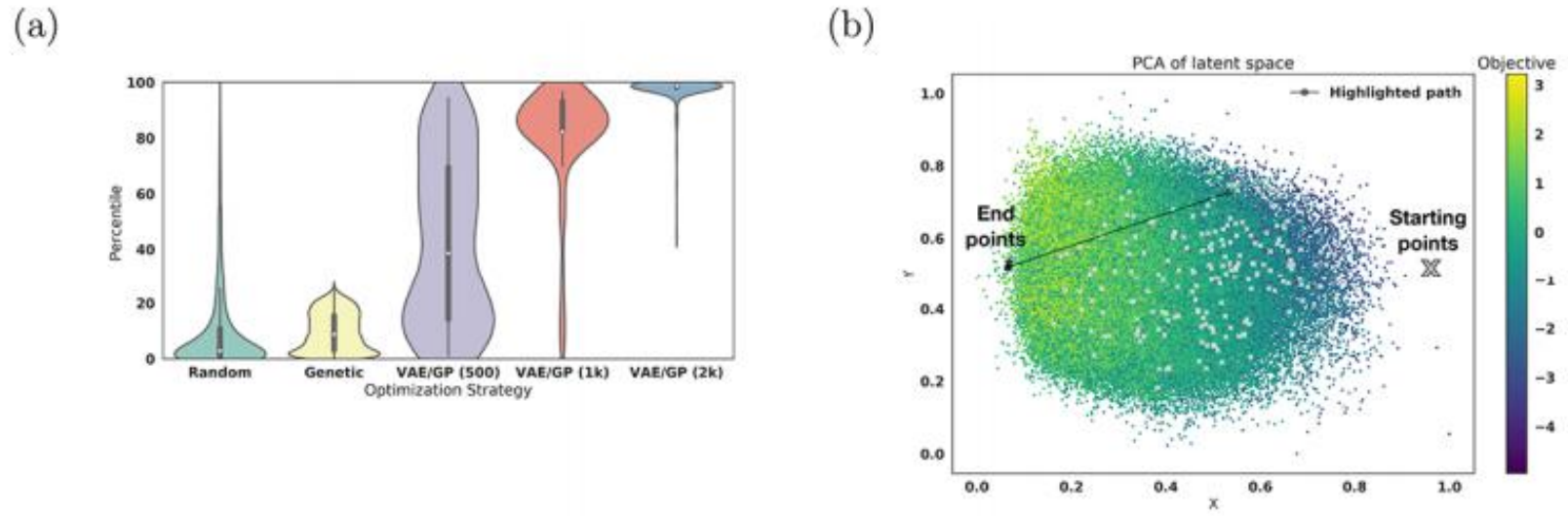
ACS Cent. Sci. 2018, 4, 2, 268–276

Results



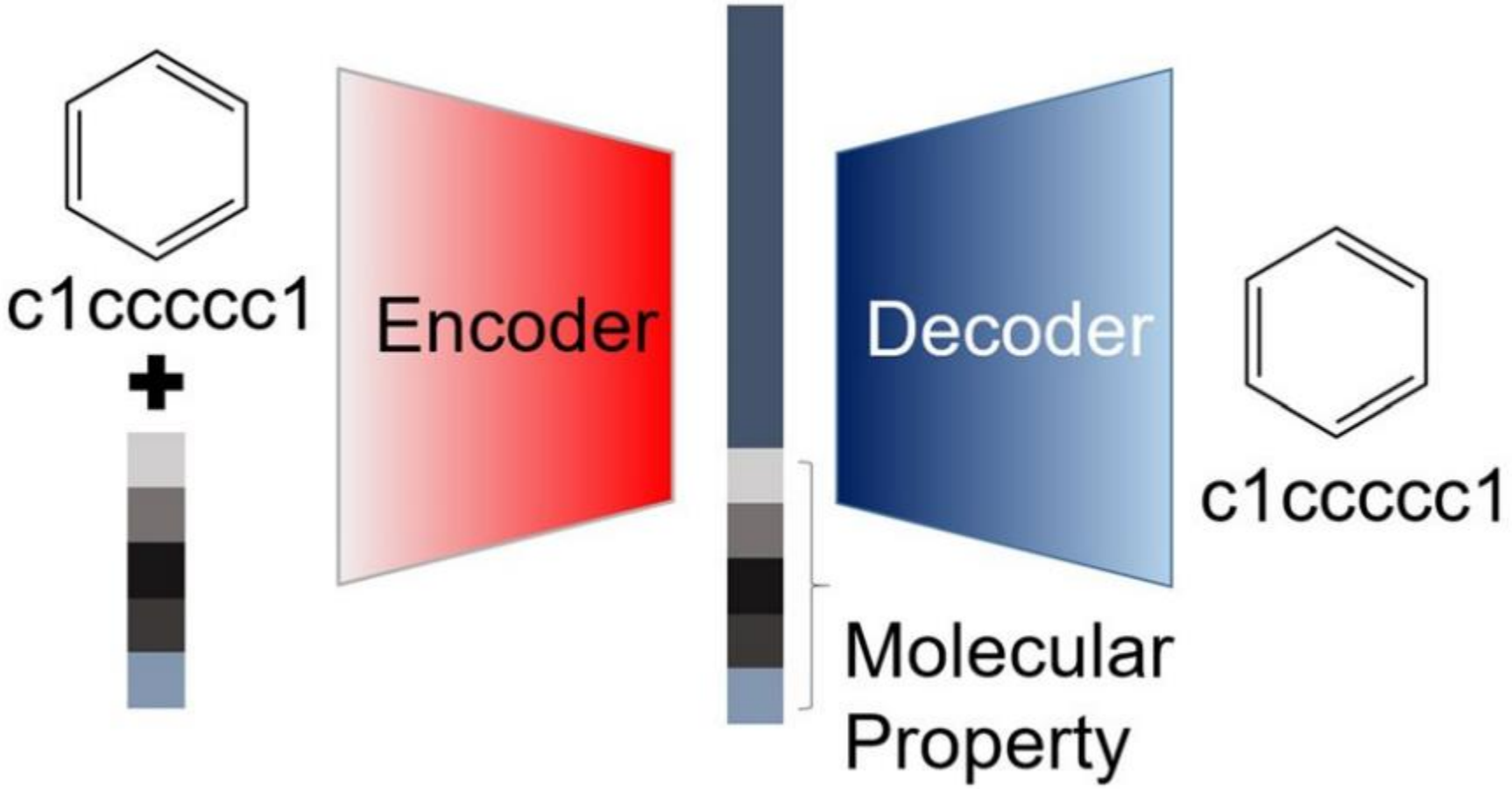
ACS Cent. Sci. 2018, 4, 2, 268–276

Results



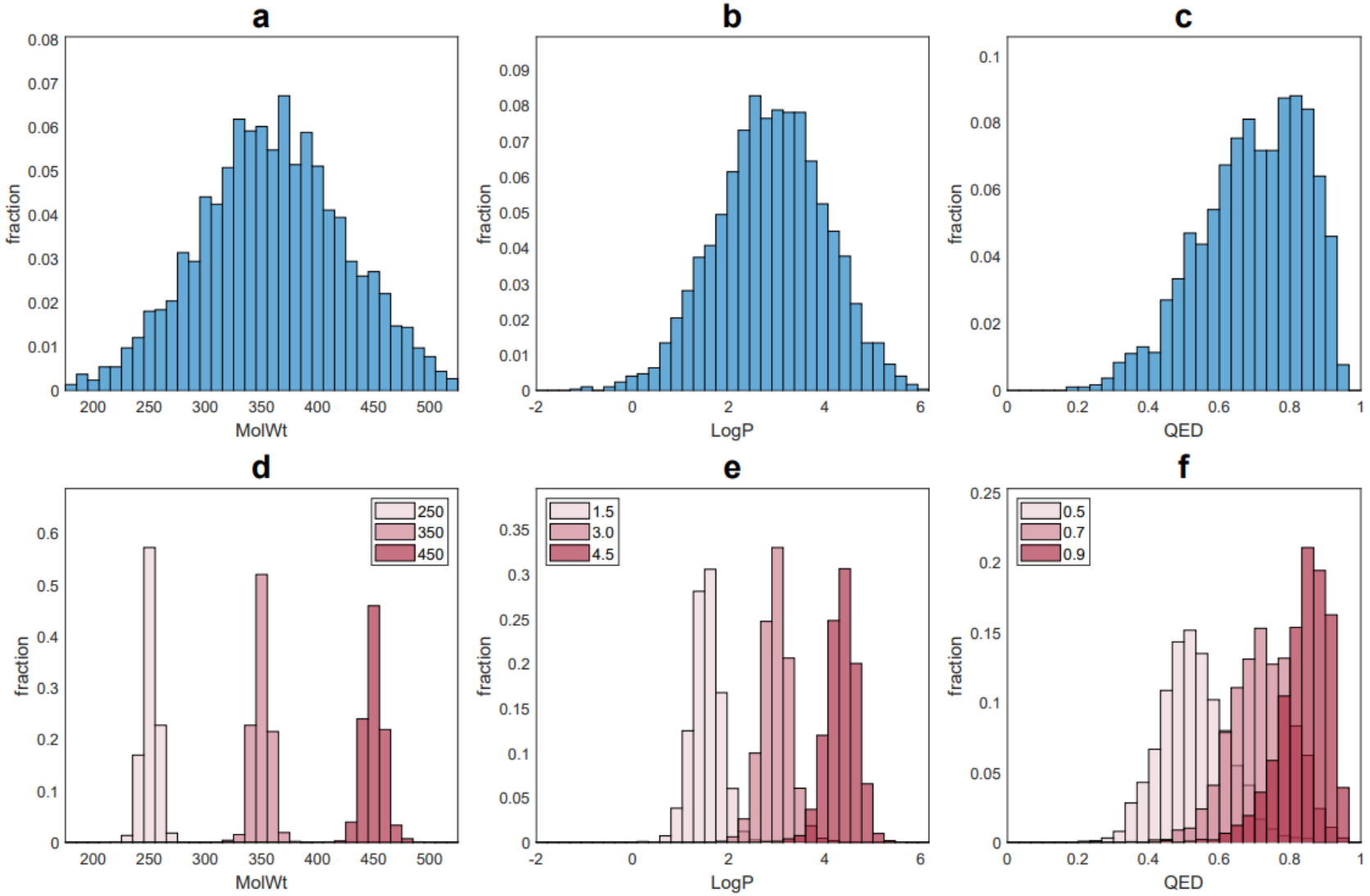
ACS Cent. Sci. 2018, 4, 2, 268–276

Conditional variational autoencoder



Journal of Cheminformatics volume 10, Article number: 31 (2018)

Results



J. Chem. Inf. Model. 2019, 59, 1, 43–52

Molecular design using variational autoencoder의 장단점

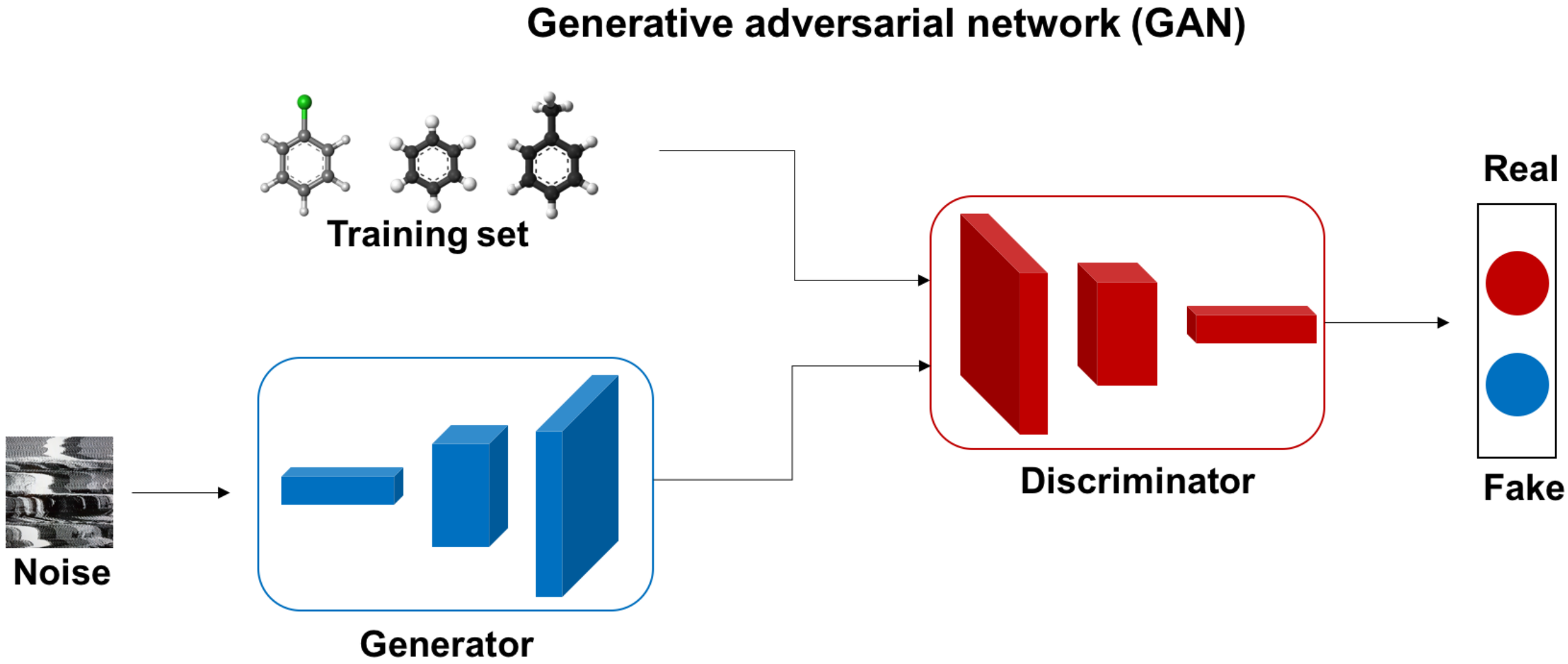
장점

- 구현하기가 상대적으로 수월함
- 난이도 대비 상대적으로 우수한 결과를 보여줌

단점

- Prior assumption이 큰 restriction으로 작용함
- Language modeling (smiles)에서 최적의 모델은 아님

Molecular design with generative adversarial network



https://github.com/chennnnnyize/Generative-Molecules

Generative adversarial network

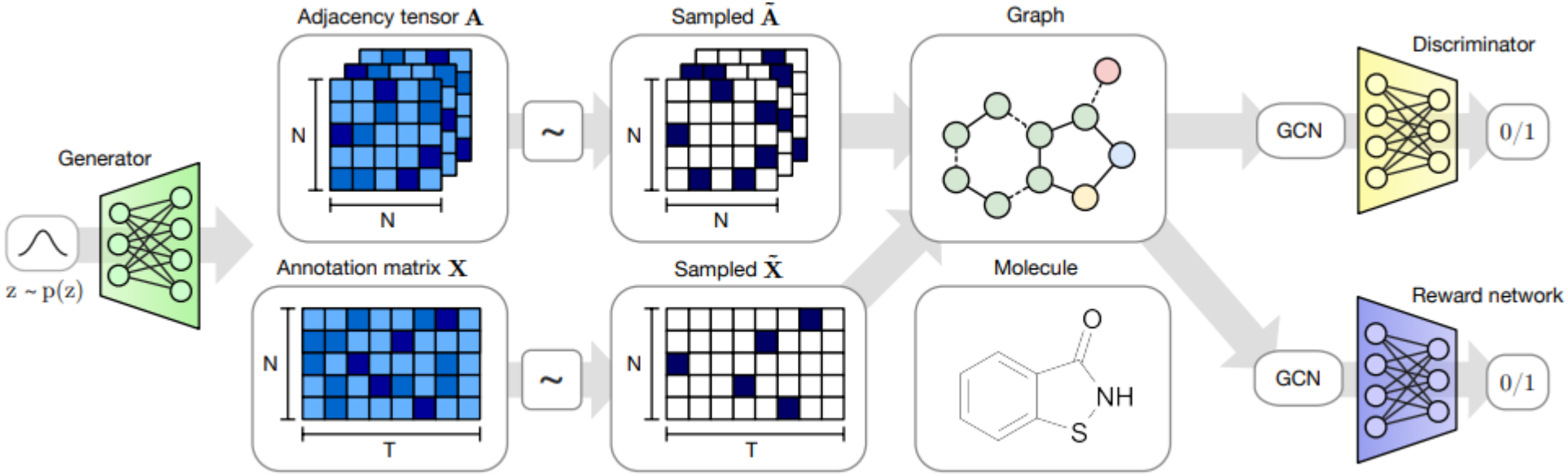
- 이미지 생성에서 최고의 성능을 보이는 GAN, 과연 분자생성에서도 최고의 모델 구조일까?

- GAN은 continuous한 object 생성에 좋은 성능을 보여줌. Language, 분자와 같이 discrete한 object를 생성하기 위해서는 적합하지 않음

- 때문에 분자생성연구에서 GAN은 많이 사용되지 않으며, 사용되는 경우 주로 작은 분자 생성에 사용됨

- RL과 같은 기법과 GAN을 결합하여, 이러한 한계를 극복하려는 시도들이 있음.

Molecular design with generative adversarial network



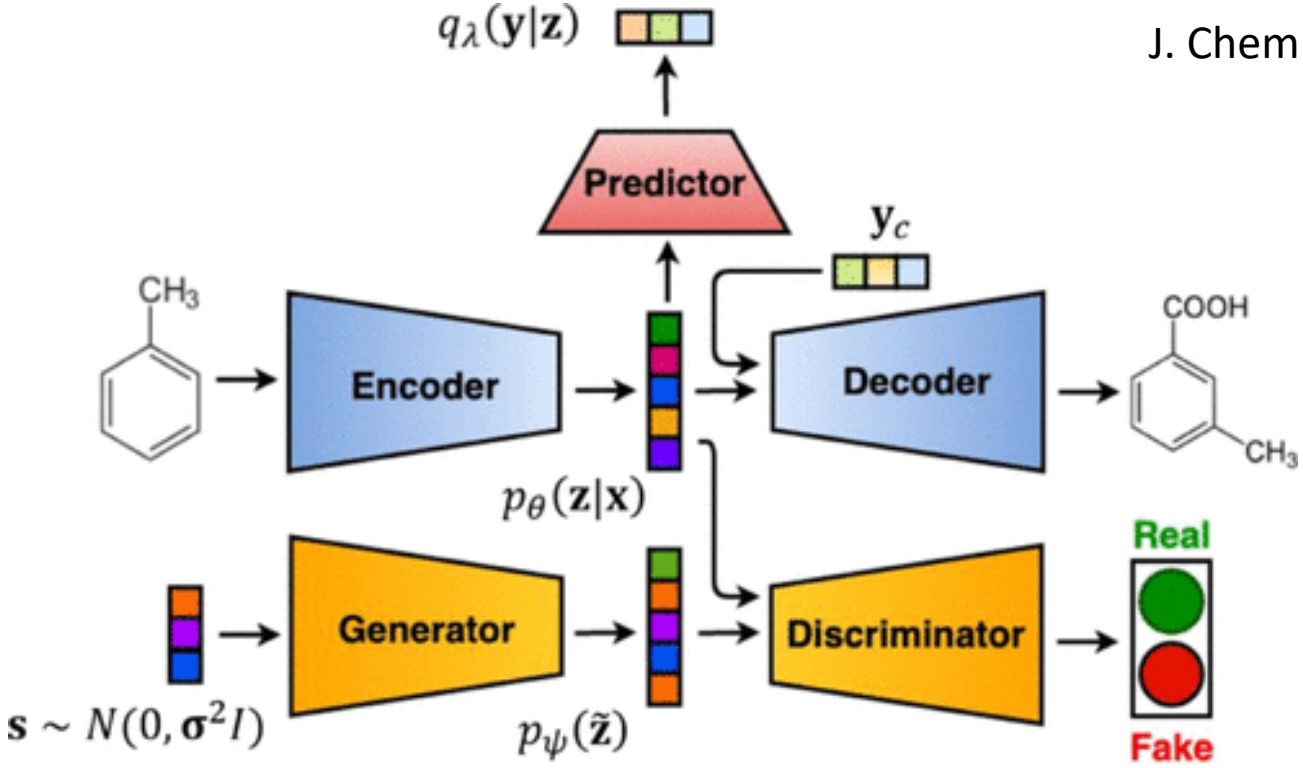
Algorithm	Valid	Unique	Novel
CharacterVAE	10.3	67.5	90.0
GrammarVAE	60.2	9.3	80.9
GraphVAE	55.7	**76.0**	61.6
GraphVAE/imp	56.2	42.0	75.8
GraphVAE NoGM	81.0	24.1	61.0
MolGAN	**98.1**	10.4	**94.2**

arXiv:1805.11973

Adversarially regularized autoencoder (ARAE)

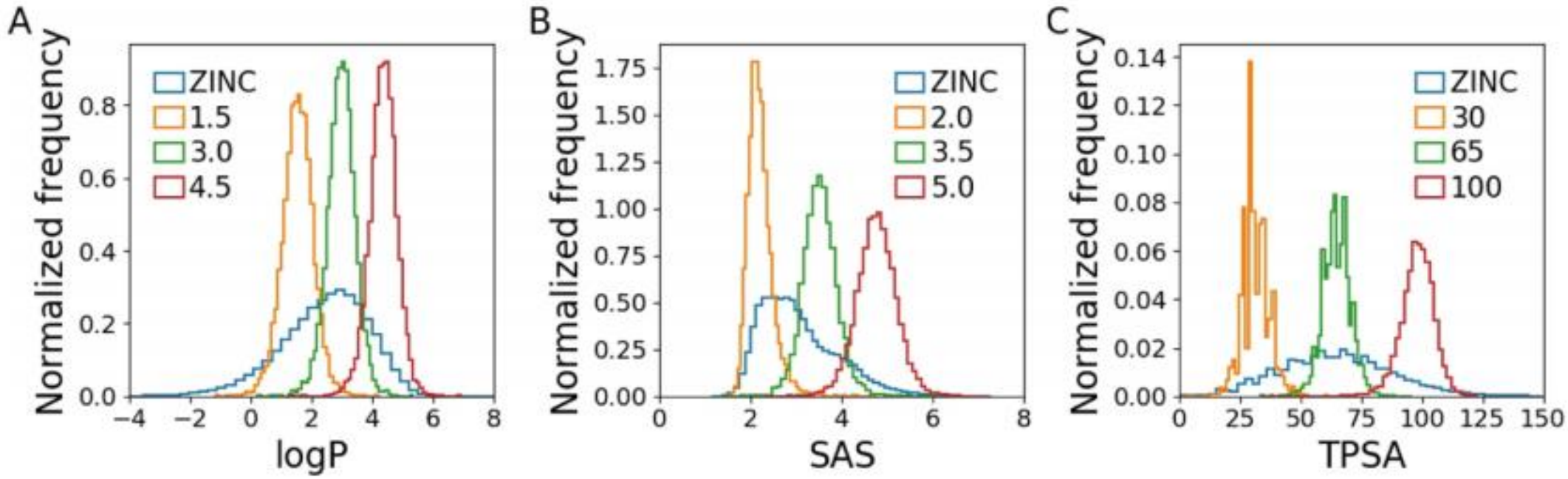
VAE와 GAN의 장점을 취합하여 만든 모델

- VAE의 prior를 GAN의 adversarial training으로 대체하여 VAE의 단점 보완
- Adversarial training을 latent space (continuous)에 적용하여 GAN의 단점 보완

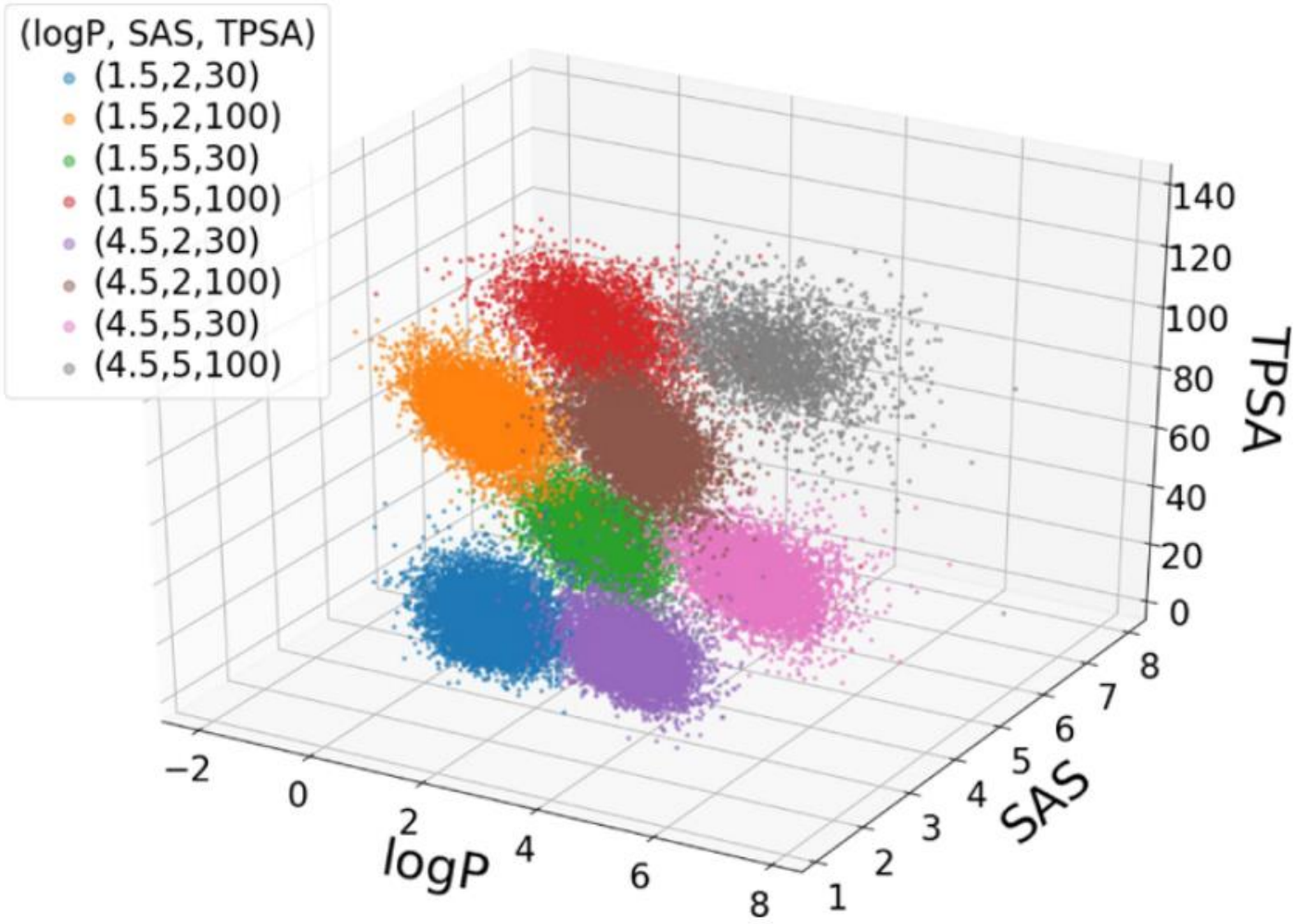


J. Chem. Inf. Model. 2020, 60, 1, 29–36

Results


A
Normalized frequency
ZINC
1.5
3.0
4.5
logP
B
Normalized frequency
ZINC
2.0
3.5
5.0
SAS
C
Normalized frequency
ZINC
30
65
100
TPSA

Results


(logP, SAS, TPSA)
(1.5,2,30)
(1.5,2,100)
(1.5,5,30)
(1.5,5,100)
(4.5,2,30)
(4.5,2,100)
(4.5,5,30)
(4.5,5,100)
logP
SAS
TPSA

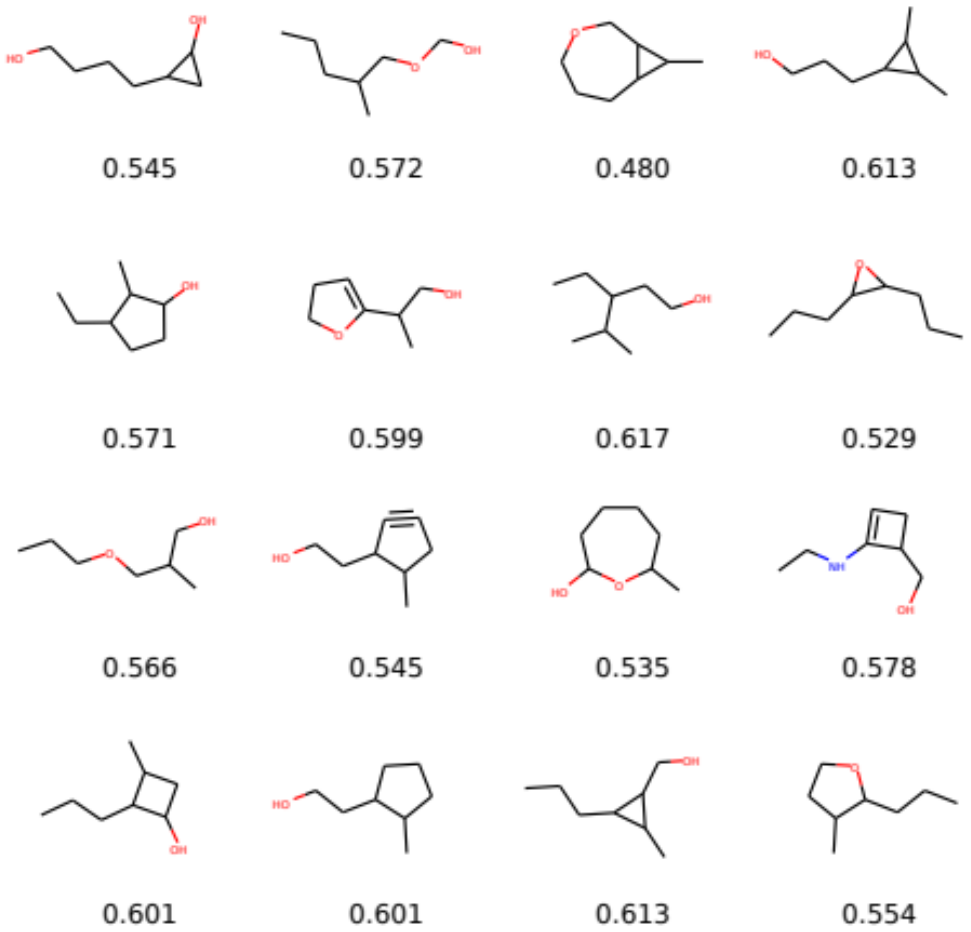
Molecular design using GAN의 장단점

장점

- Prior를 가정하지 않기 때문에 restriction으로 인한 부작용이 없음

단점

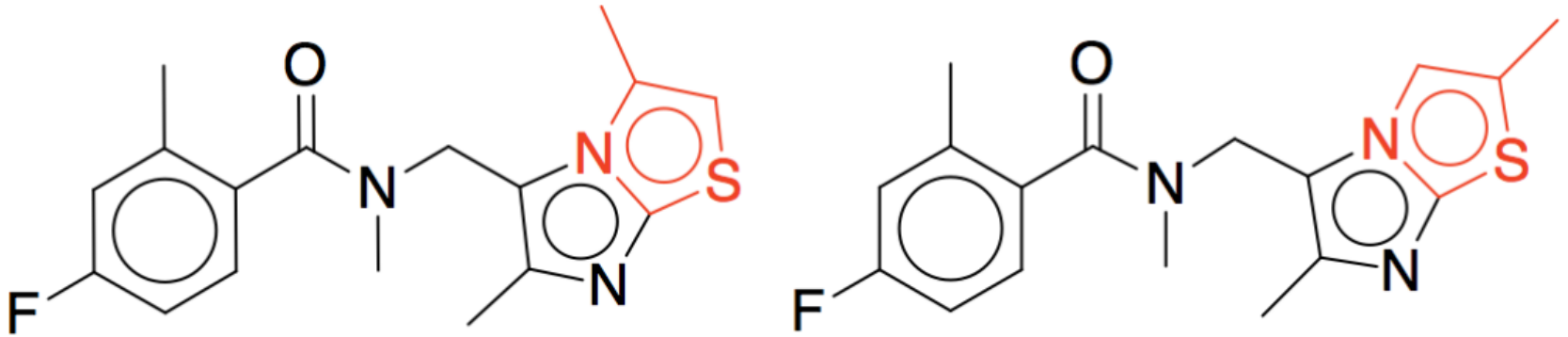
- Discrete한 object에서 좋은 성능을 내지 못함 (작은 분자들만 적용가능)
- 추가적인 module (RL)등이 필요함



Graph vs smiles

Smiles의 문제점 및 graph의 장점

- 유사한 분자가 매우 다른 smiles로 표현됨 (학습에 어려움이 가중됨)
- Graph가 smiles보다 분자를 표현할 수 있는 보다 자연스러운 representation



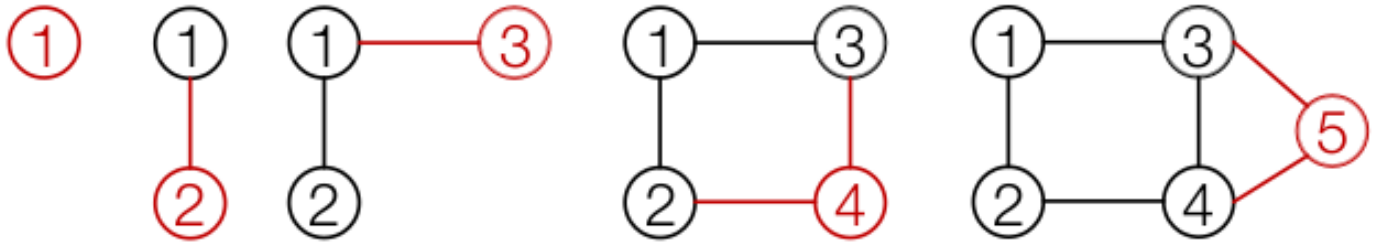
Cc1cn2c(CN(C)C(=O)c3ccc(F)cc3C)c(C)nc2s1

Cc1cc(F)ccc1C(=O)N(C)Cc1c(C)nc2scc(C)n12

Molecular graph generative model

- 분자는 2D, smiles (=sequence)는 1D
- 2D 그래프를 어떻게 생성할 것인가?

- **Node-level**: At each step, a new node is added

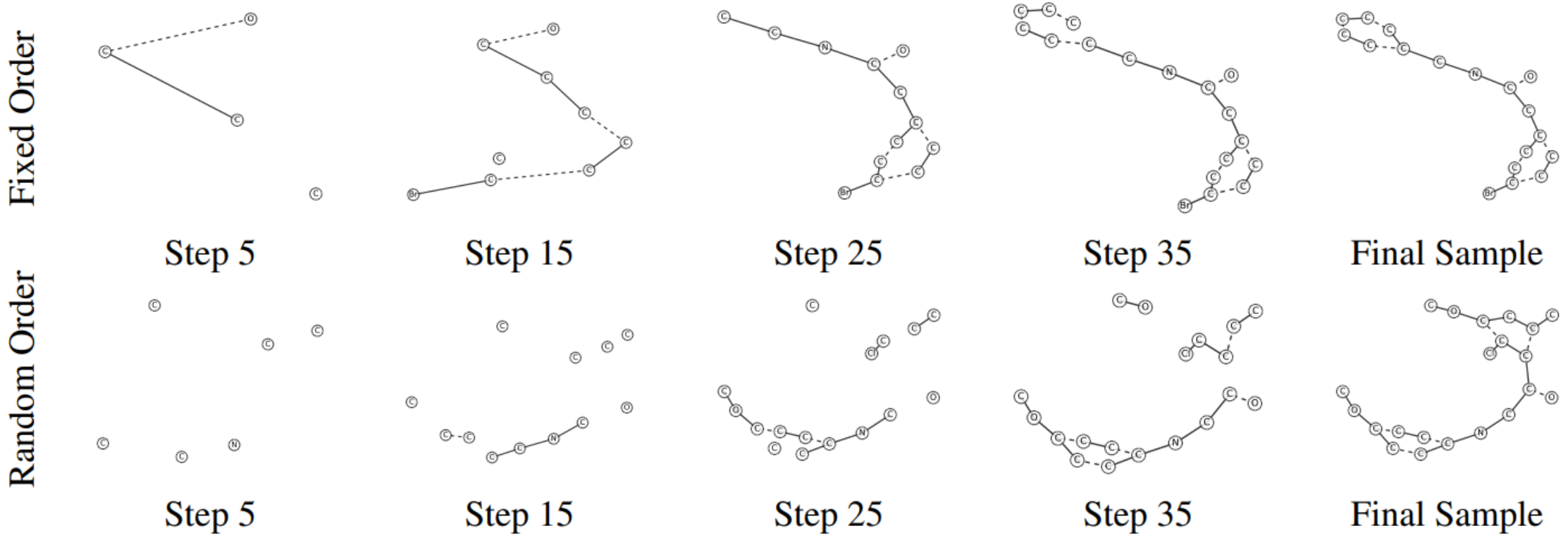


$S^{\pi} = (S_1^{\pi} , S_2^{\pi} , S_3^{\pi} , S_4^{\pi} , S_5^{\pi})$

"Add node 1" ... "Add node 5"

https://snap-stanford.github.io/cs224w-notes/machine-learning-with-networks/graph-generative-models

Molecular graph generative model



Results

Model	% valid	% novel	% valid and novel
SMILES VAE	0.804 ± 0.016	0.986 ± 0.000	0.793 ± 0.016
SMILES GRU1	0.886 ± 0.002	0.984 ± 0.000	0.872 ± 0.002
SMILES GRU2	0.932 ± 0.002	0.965 ± 0.001	0.899 ± 0.002
SMILES LSTM	0.935 ± 0.006	0.975 ± 0.001	0.912 ± 0.006
MolMP ($\alpha = 1.0$)	0.952 ± 0.002	0.98 ± 0.001	0.933 ± 0.001
MolMP ($\alpha = 0.8$)	0.962 ± 0.002	0.984 ± 0.001	0.946 ± 0.001
MolMP ($\alpha = 0.6$)	0.963 ± 0.001	*0.988 ± 0.001***	0.951 ± 0.001
MolRNN ($\alpha = 1.0$)	0.967 ± 0.001	0.959 ± 0.000	0.928 ± 0.001
MolRNN ($\alpha = 0.8$)	*0.970 ± 0.001*	0.976 ± 0.001	0.947 ± 0.001
MolRNN ($\alpha = 0.6$)	*0.970 ± 0.001*	0.985 ± 0.000	*0.955 ± 0.001****

Journal of Cheminformatics volume 10, Article number: 33 (2018)

Molecular graph generative model의 장단점

장점

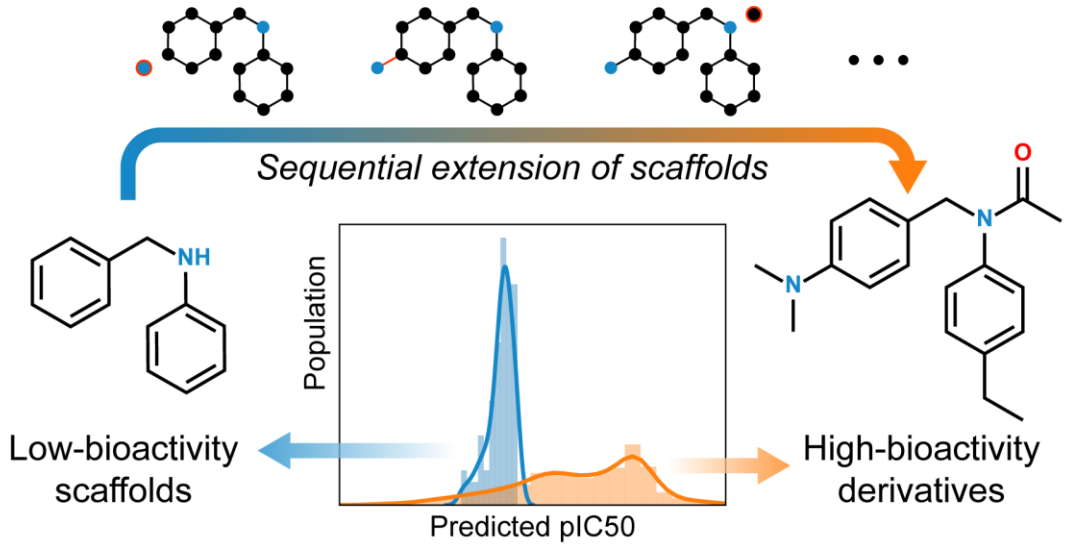
- 이론적 및 직관적으로 분자를 가장 잘 표현할 수 있는 representation으로서, 모델학습과정을 효율적으로 만들어줄 수 있음

단점

- 학습과정 및 모델 구현이 복잡함. (GPU 최적화가 어려움)
- 몇몇 task에 대해서 (약간의) 좋은 성능을 보여주지만 이론적 당위성과 별개로, smiles에 비해 뛰어난 성능을 보여주는 결과들이 없음. (smiles에서는 안되는 것이 graph로 하면 되는 경우 보고된 바 없음)

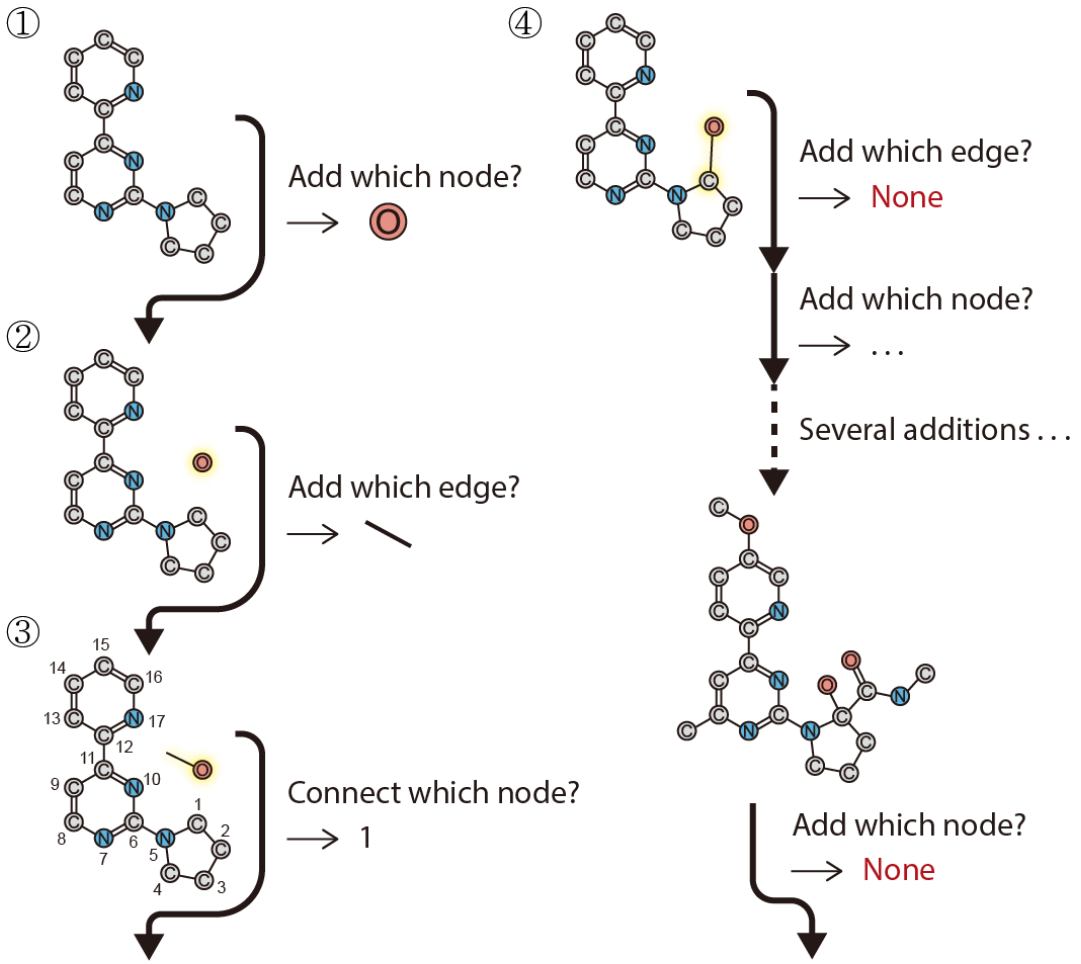
Scaffold based vs de novo design

- 이상적으로 de novo design이 가장 좋은 접근법
- 대부분의 현재 deep learning기반 분자 생성모델들은 de novo design에 focusing되어 있음.
- 하지만 광대한 chemical space를 scratch에서 탐색하는 것은 현실적이지 않음
- 실제로는 활성이 검증된 scaffold들로 출발하는 경우가 더 많음
- 이러한 현실을 반영하는 deep learning model이 필요함

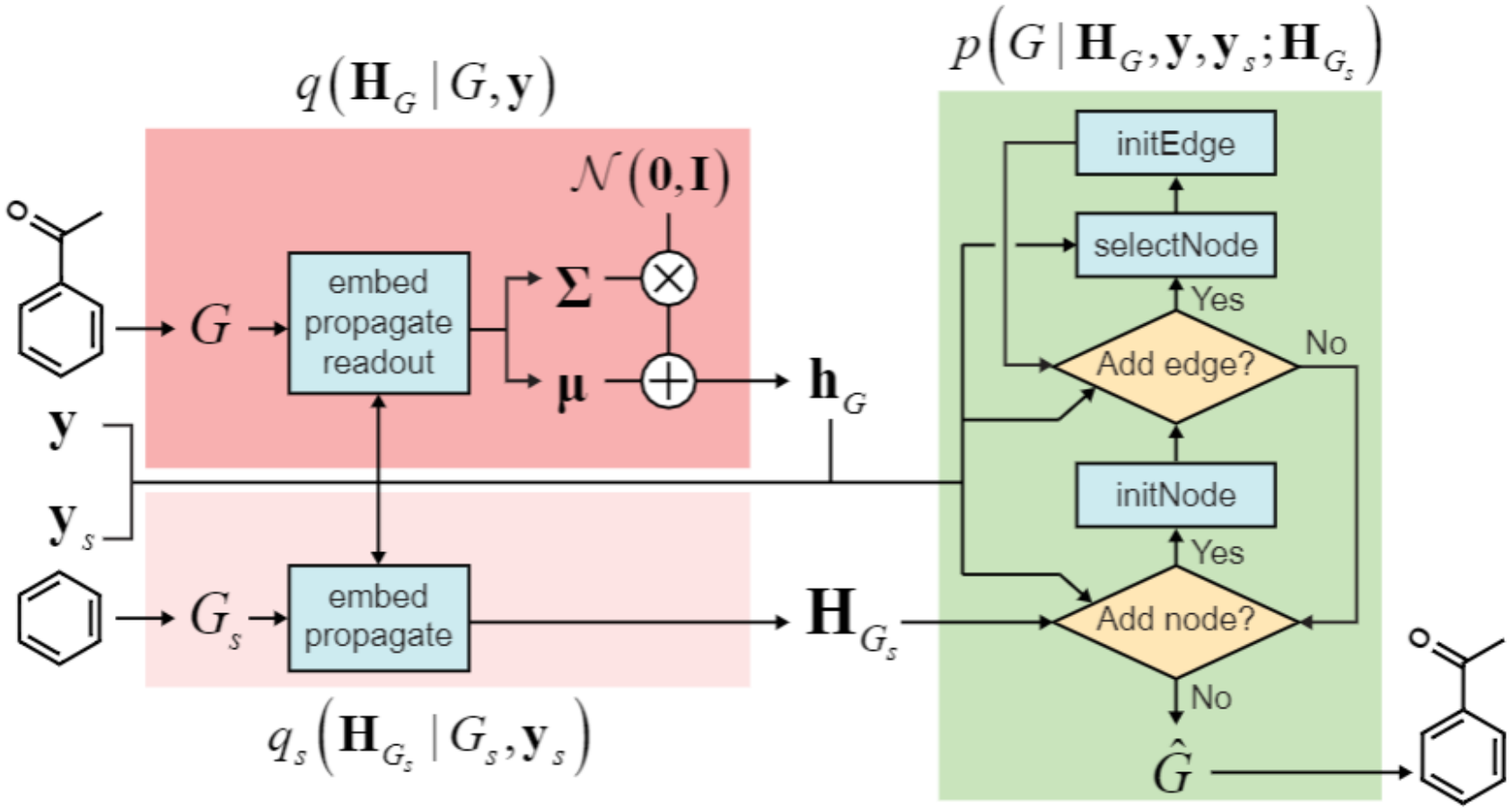


Scaffold based graph generative model

- 주어진 scaffold에 atom과 bond를 추가하는 방식으로 새로운 분자를 디자인



Scaffold based graph generative model



Scaffold based graph generative model

Algorithm 1 Scaffold-based graph generation

Inputs: G, S, $\mathbf{y}$, $\mathbf{y}_S$ ▷ Whole/scaffold graphs and properties

$G_0 \leftarrow S$

$\tilde{\mathbf{y}} \leftarrow \text{concat}\left(\mathbf{y}, \mathbf{y}_S\right)$

if $G \neq (\emptyset, \emptyset)$ **then** ▷ Learning phase

  $\left(\mathbf{H}_{V(G)}, \mathbf{H}_{E(G)}\right) \leftarrow \text{embed}(G)$

  $\mathbf{H}_{V(G)} \leftarrow \text{propagate}^{(k)}\left(\mathbf{H}_{V(G)}, \mathbf{H}_{E(G)}, \tilde{\mathbf{y}}\right)$

  $\mathbf{z} \sim \text{reparam} \circ \text{readout}\left(\mathbf{H}_{V(G)}\right)$ ▷ Vector representation of the target graph

else

  $\mathbf{z} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right)$ ▷ Generation phase

end if

$\tilde{\mathbf{z}} \leftarrow \text{concat}\left(\mathbf{z}, \tilde{\mathbf{y}}\right)$

$\left(\mathbf{H}_{V(G_0)}, \mathbf{H}_{E(G_0)}\right) \leftarrow \text{embed}\left(G_0\right)$ ▷ Node and edge feature vectors

$\mathbf{H}_{V(G_0)} \leftarrow \text{propagate}^{(k)}\left(\mathbf{H}_{V(G_0)}, \mathbf{H}_{E(G_0)}, \tilde{\mathbf{y}}\right)$ ▷ Initial update of the scaffold nodes

$t \leftarrow 1$ ▷ Node addition counter

$v_t \sim \text{Cat} \circ \text{addNode}\left(\mathbf{H}_{V(G_{t-1})}, \mathbf{H}_{E(G_{t-1})}, \tilde{\mathbf{z}}\right)$ ▷ Sample a node type or STOP

while $v_t \neq \text{STOP}$ **do**

  $V(G_t) \leftarrow V(G_{t-1}) \cup \{v_t\}$ ▷ Add the new node

  $\mathbf{H}_{V(G_t)} \leftarrow \mathbf{H}_{V(G_{t-1})} \cup \left\{\text{initNode}\left(v_t, \mathbf{H}_{V(G_{t-1})}\right)\right\}$ ▷ Initialize and add a new node vector

Scaffold based graph generative model

$\mathbf{H}_{V(G_t)} \leftarrow \mathbf{H}_{V(G_{t-1})} \cup \{\mathsf{initNode}\,(v_t, \mathbf{H}_{V(G_{t-1})})\}$ ▷ Initialize and add a new node vector

$E_{t,0} \leftarrow E(G_{t-1})$; $\mathbf{H}_{E_{t,0}} \leftarrow \mathbf{H}_{E(G_{t-1})}$ ▷ Prepare edge additions

$i \leftarrow 1$ ▷ Edge addition counter

$e_{t,i} \sim \mathrm{Cat} \circ \mathsf{addEdge}\,(\mathbf{H}_{V(G_t)}, \mathbf{H}_{E_{t,i-1}}, \tilde{\mathbf{z}})$ ▷ Sample an edge type or STOP

while $e_{t,i} \neq \mathsf{STOP}$ **do**

  $v_{t,i} \sim \mathrm{Cat} \circ \mathsf{selectNode}\,(\mathbf{H}_{V(G_t)}, \mathbf{H}_{E_{t,i-1}}, \tilde{\mathbf{z}})$ ▷ Sample a node to connect

  $E_{t,i} \leftarrow E_{t,i-1} \cup \{(v_t, v_{t,i})\}$ ▷ Add the new edge (with type $e_{t,i}$)

  $\mathbf{H}_{E_{t,i}} \leftarrow \mathbf{H}_{E_{t,i-1}} \cup \{\mathsf{initEdge}\,(e_{t,i}, \mathbf{H}_{V(G_t)})\}$ ▷ Initialize and add a new edge vector

  $i \leftarrow i + 1$

  $e_{t,i} \sim \mathrm{Cat} \circ \mathsf{addEdge}\,(\mathbf{H}_{V(G_t)}, \mathbf{H}_{E_{t,i-1}}, \tilde{\mathbf{z}})$ ▷ Sample a next edge type or STOP

end while

$\mathbf{H}_{E(G_t)} \leftarrow \mathbf{H}_{E_{t,i-1}}$

$E(G_t) \leftarrow E_{t,i-1}$

$G_t \leftarrow (V(G_t), E(G_t))$

$t \leftarrow t + 1$

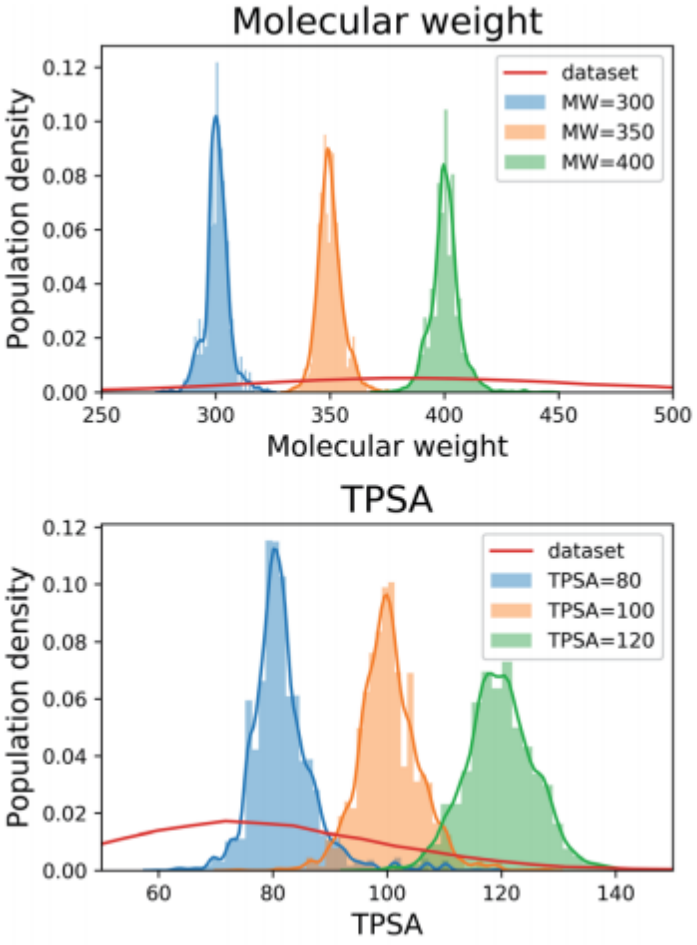
$v_t \sim \mathrm{Cat} \circ \mathsf{addNode}\,(\mathbf{H}_{V(G_{t-1})}, \mathbf{H}_{E(G_{t-1})}, \tilde{\mathbf{z}})$ ▷ Sample a next node type or STOP

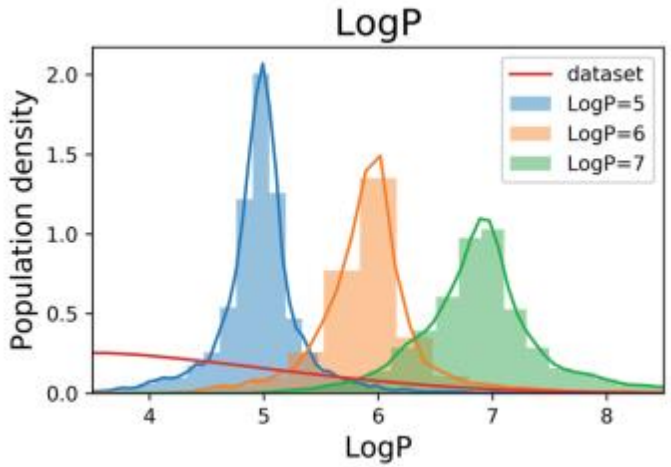
end while

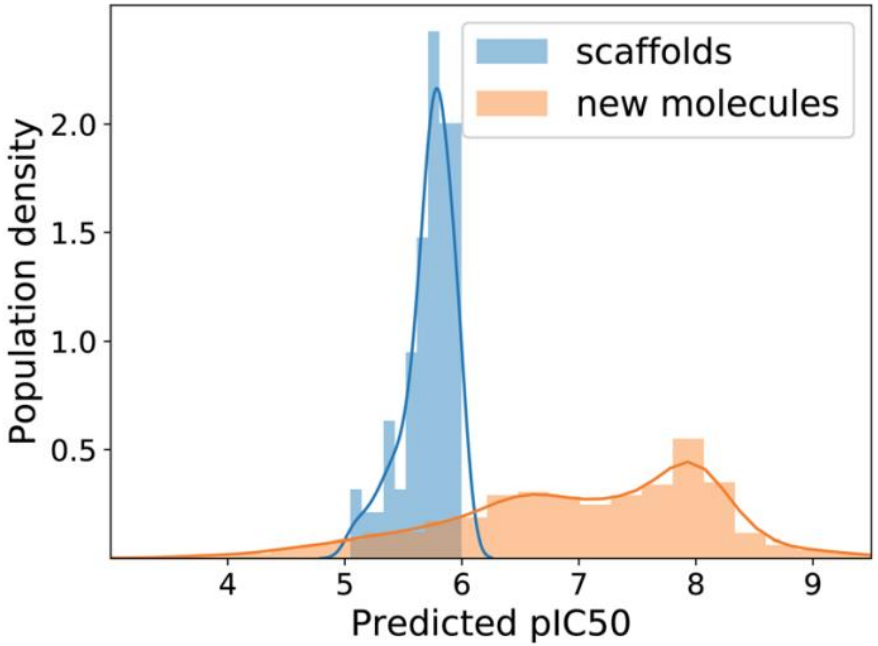
$G_t^* \sim \mathrm{Cat} \circ \mathsf{selectIsomer}\,(G_t, \tilde{\mathbf{z}})$ ▷ Assign the stereoisomerism

return G_t^*

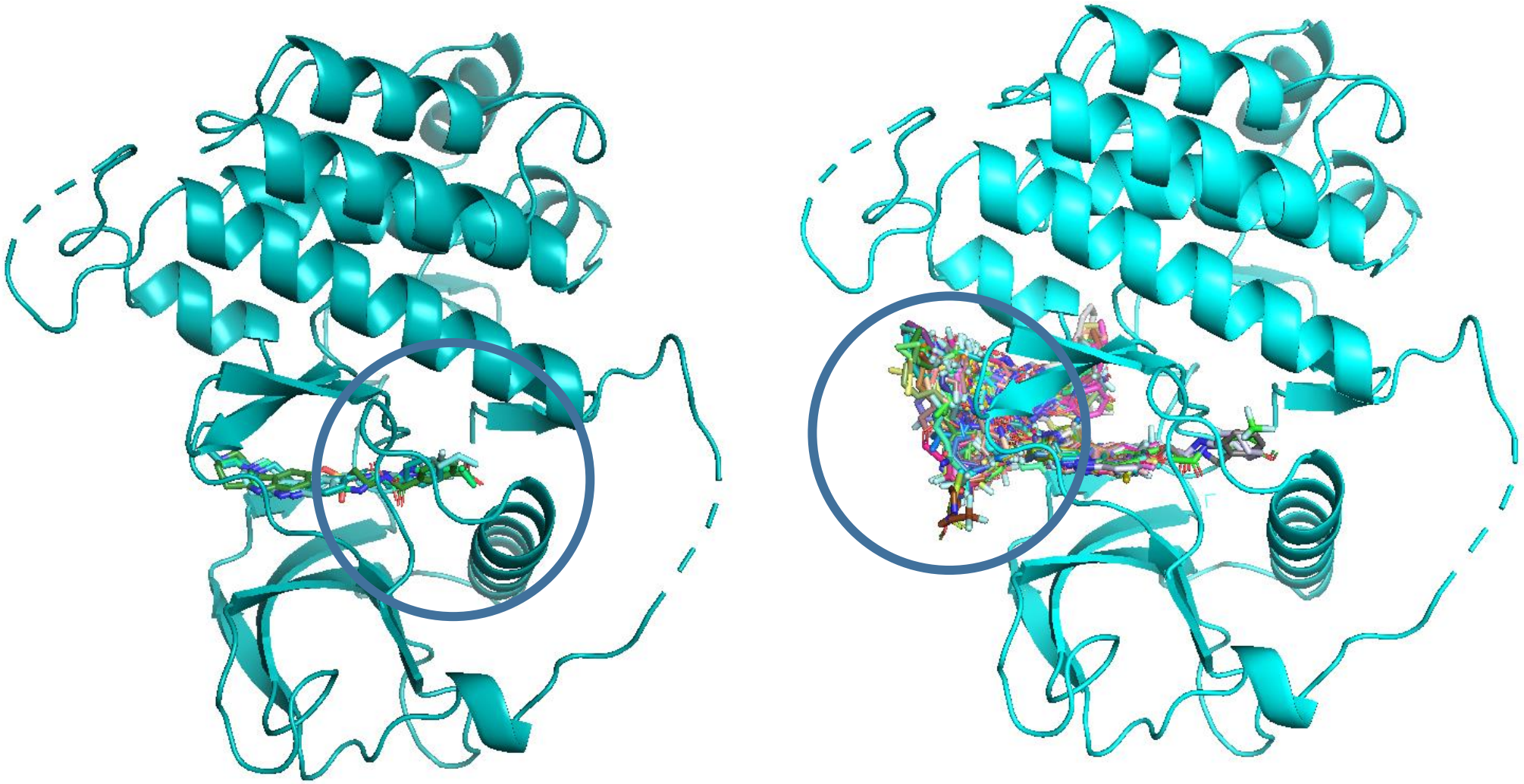
Scaffold based graph generative model



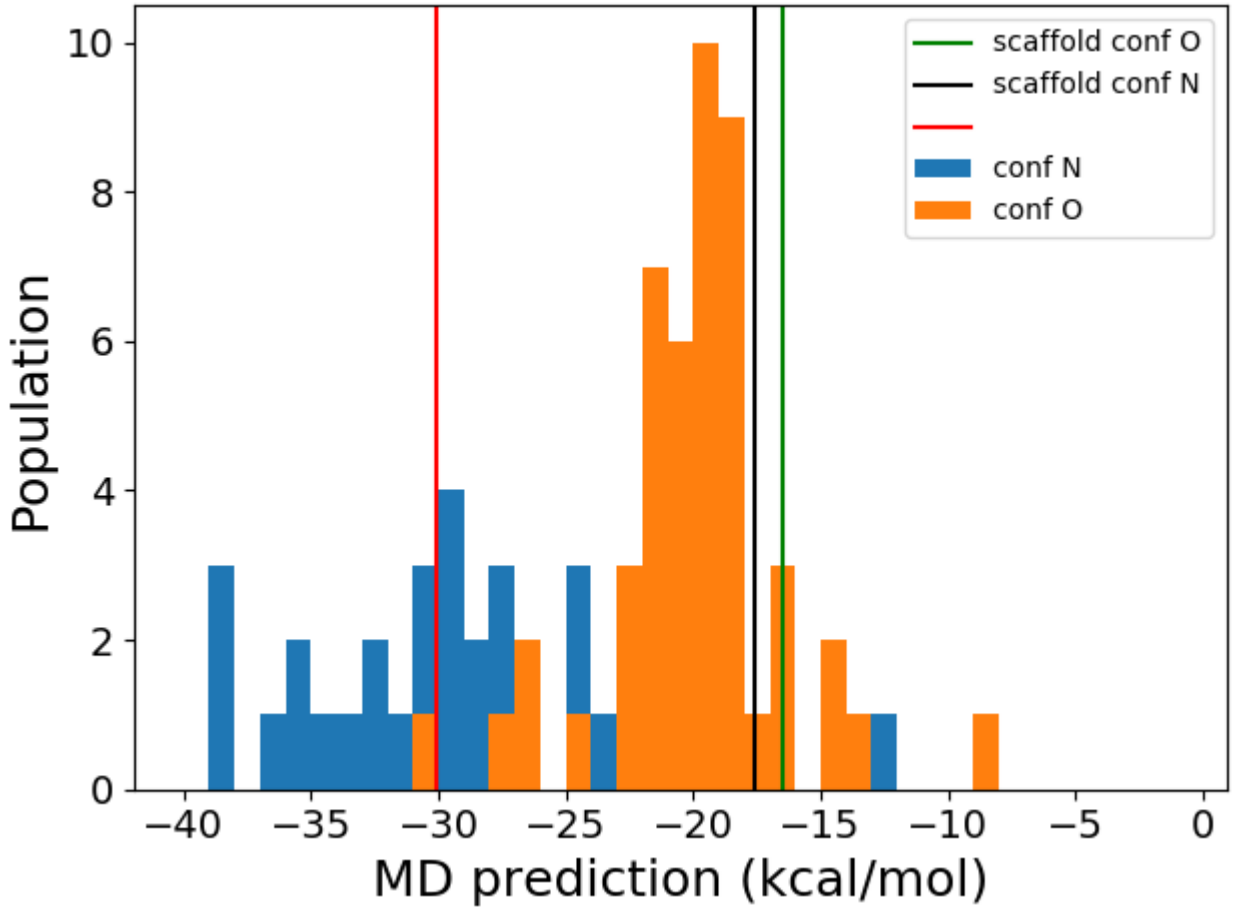




Improve potency of scaffold against XXX target



Improve potency of scaffold against XXX target



MD calculation

- MMPBSA with charmm36 force field
- Reference setting: compare MD results (1ns MD production * 20 trajectory) with experimental results of XXXX and the given scaffold
- 1ns*1 trajectory for the proposed molecules
- Protein structure from PDBID XXX

Thank you