De novo Design

De novo Drug Design: Generating

WorkFlow

Data collection

- ASK1 inhibitor data
- Chembl data
- Collected by PubChem and Thomson-Pharma

Data preprocessing

- Remove duplicate compounds
- Remove Isomeric from SMILES Structure

Molecular Generator

- GAN
- Conditional GAN
- Generative RNN
- Reinforcement Learning
- Variational AutoEncoder

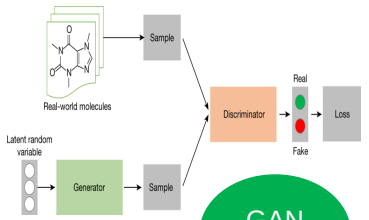
Result

- Canonize SMILES
- Tanimoto Similarity
- T-SNE

QSAR

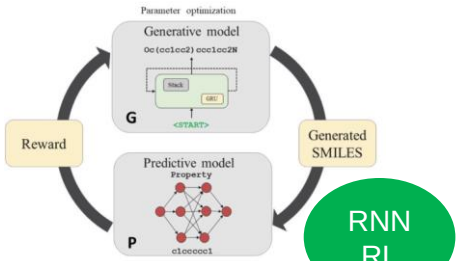
- Predict biological Activity
- Random Forest
- CNN

Molecular Generator



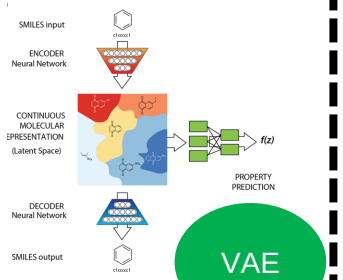
GAN
cGAN

- (Conditional) Generative Adversarial Network
- Divide logP into 5 segments to create a new molecule



RNN
RL

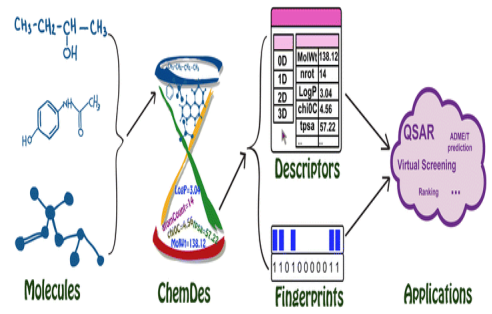
- Reinforcement learning
- Generative model: Stack GRU (Generative RNN)
- Prediction model: Random forest



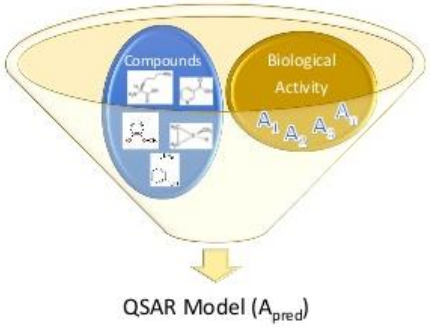
VAE

- Variational AutoEncoder
- Encoder: 1D CNN
- Decoder: GRU(RNN)

De novo Drug Design: QSAR



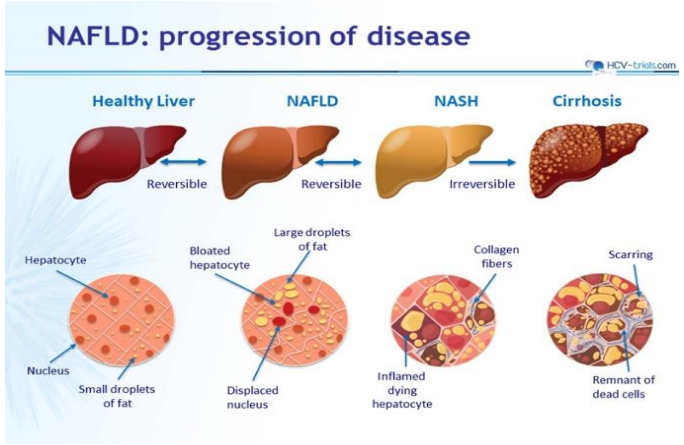
- Convert IC50 to pIC50
- Generating Descriptor & Fingerprint (ECFP4, MACCS Fingerprint)



- RandomForest, Chemception (CNN)
- Regression Model
- Classification Model

(Active Compound pIC50 >=7)

1. Data



1)Target Data (NASH)

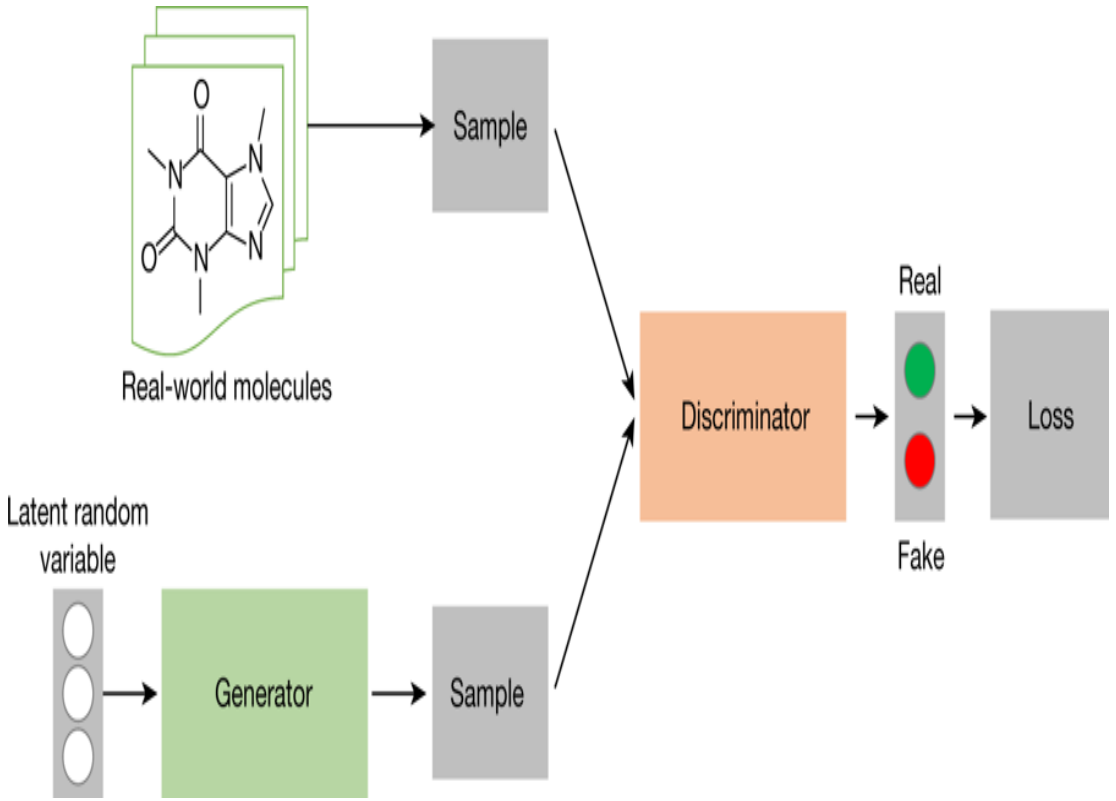
-Finetuning Data
-Collected by PubChem and Thomson-Pharma
-ASK1 Inhibitor(624 case)

2)Pretrain Data

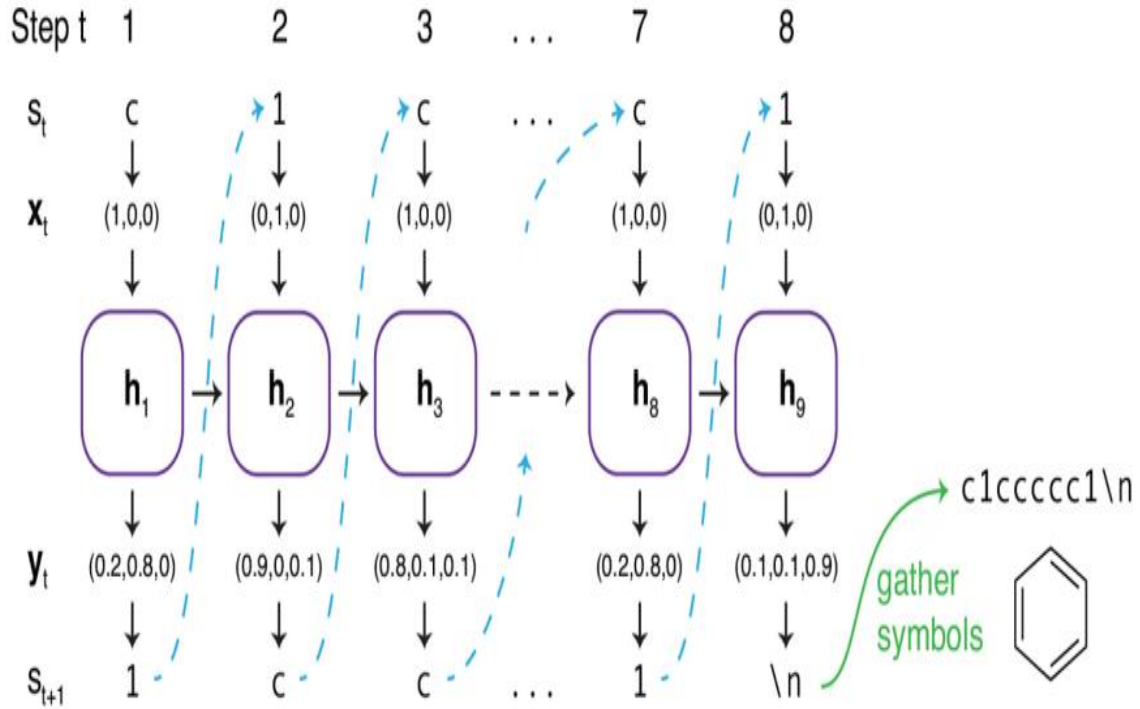
-Enrichment & pretrain Data
-ChEMBL: RNN, Reinforcement Learning
-ZINC: Variational Autoencoder

2. Molecular Generator: GAN, RNN

<GAN Structure>



<Generative RNN Structure>

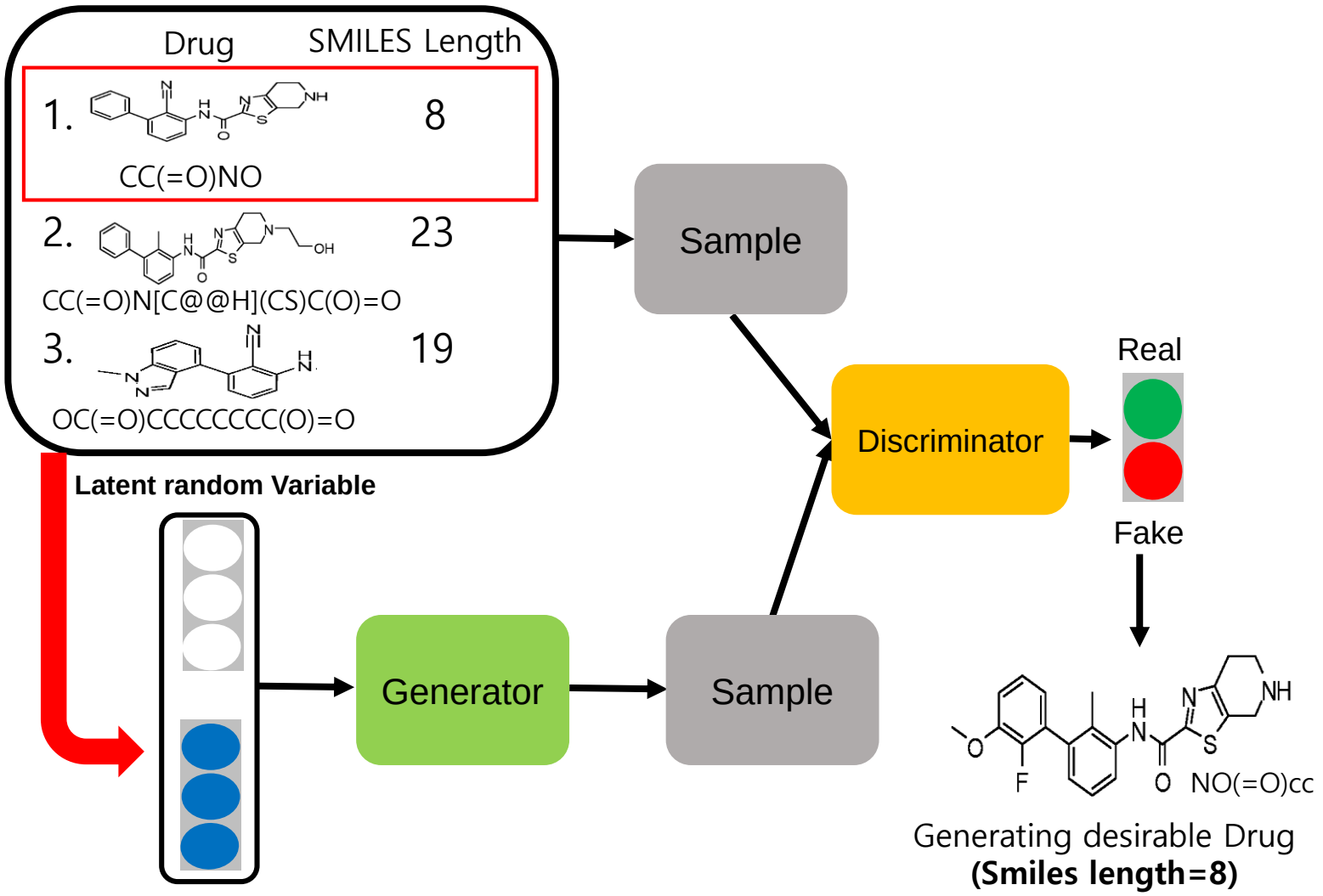


Ref.

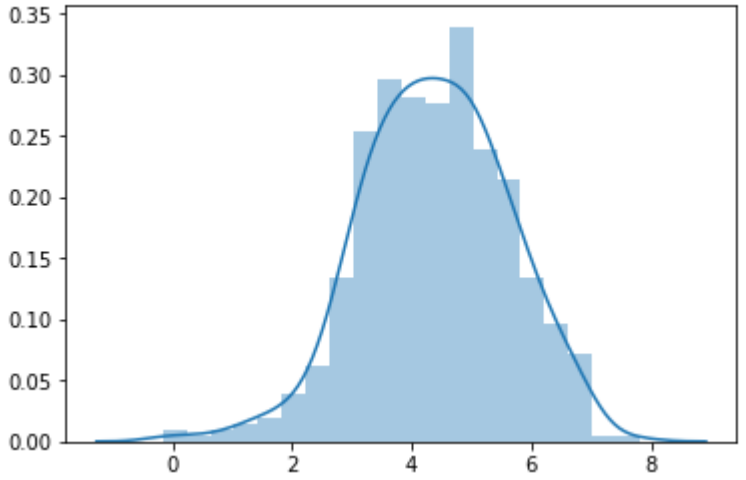
1.Generative Recurrent Networks for De Novo Drug Design, **Gupta A** (2017)
2.SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient, **Lantao Yu** (2017)
3.Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks, **Marwin H. S. Segler**(2018)
4.Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models, **Gabriel Lima Guimaraes** (2017)

2. Molecular Generator: Conditional GAN

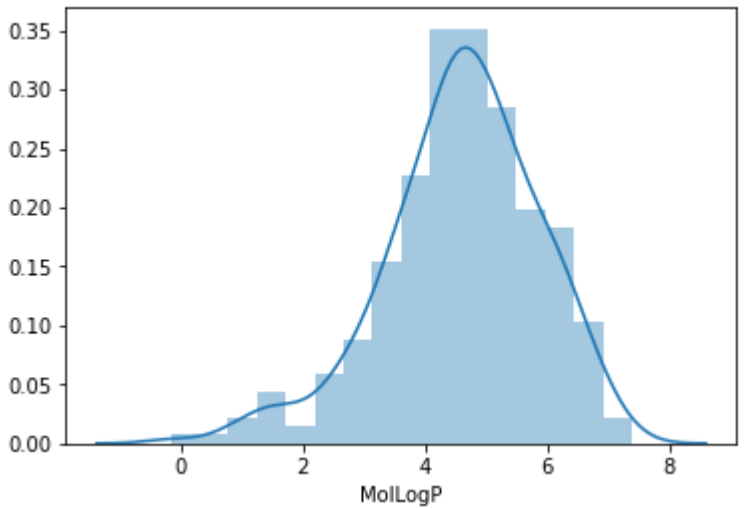
<Conditional GAN Structure>



<ASK1 Inhibitor의 logP의 분포 >

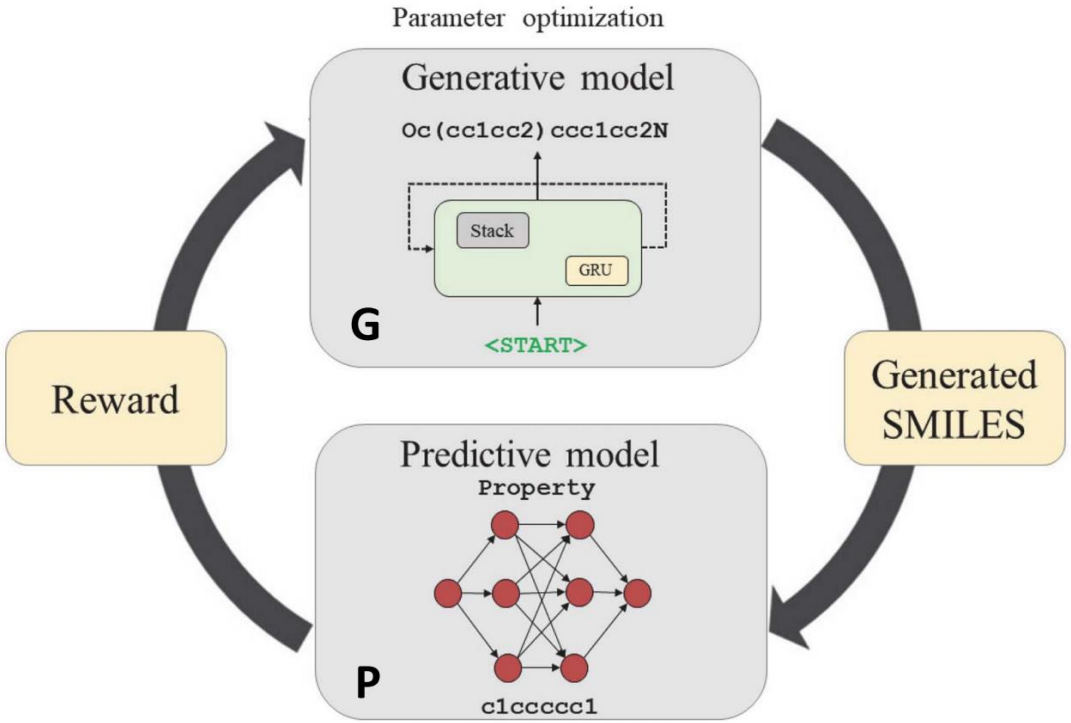


<cGAN으로 생성한 logP의 분포 >

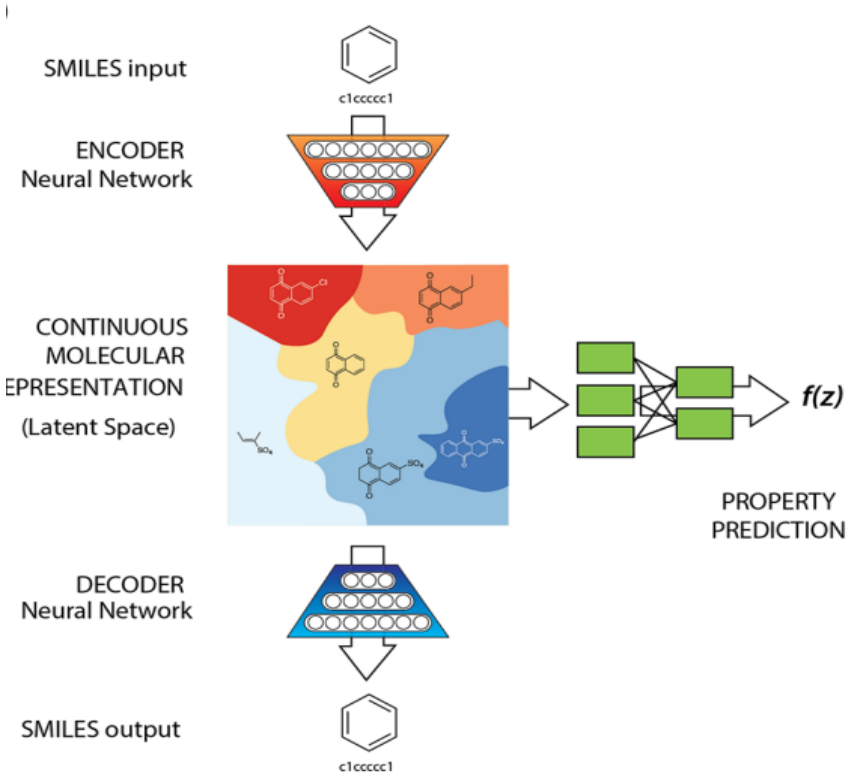


2. Molecular Generator: RL, VAE

<Reinforcement Learning Cycle>



<Molecular VAE>



Ref.

1)Deep reinforcement learning for de novo drug design, **Mariya Popova** (2018)

2)Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules, **Rafael Gómez-Bombarelli** (2018)

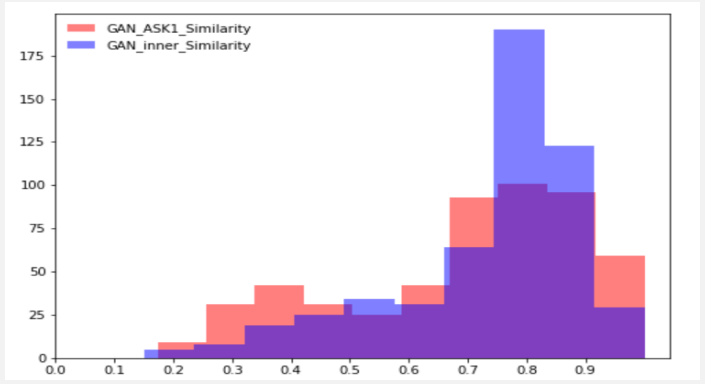
3)Molecular generative model based on conditional variational autoencoder for de novo molecular design, **Jaechang Lim**(2018)

3. Result (Tanimoto Similarity,T-SNE)

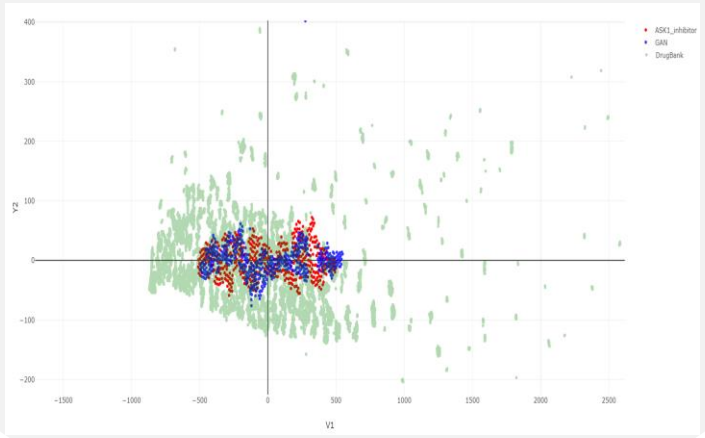
GAN

- ASK1 데이터를 3분할하여 GAN으로 528개 Compound 생성

(**A**)Tanimoto Simlarity



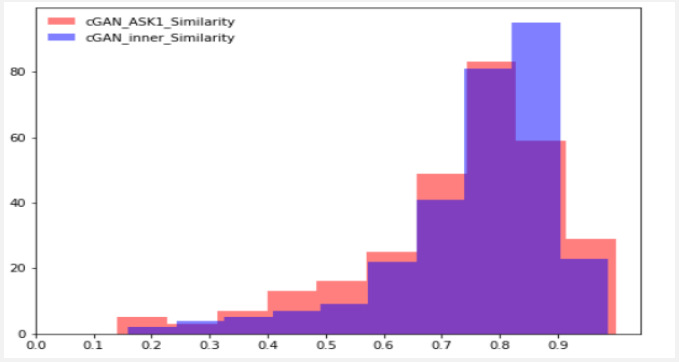
(B)T-SNE



cGAN

-ASK1 데이터를 logP 기준으로 5구간을 나누어 각각 생성
-cGAN으로 287개 compound 생성

(**A**)Tanimoto Simlarity

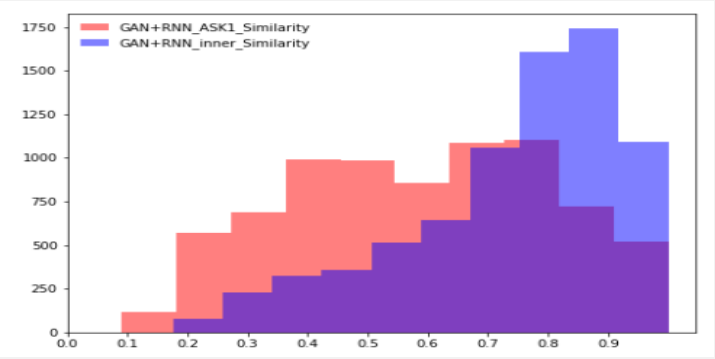


(B)T-SNE

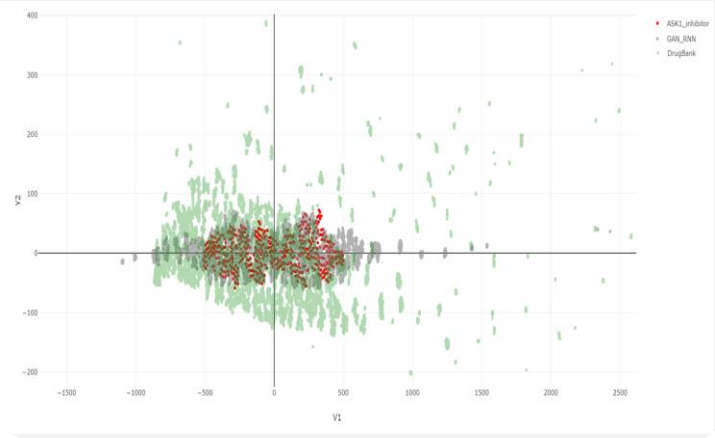
GAN+RNN

-Input Data: ASK1 Inhibitor + GAN으로 생성한 데이터
-RNN으로 7,638개 compound 생성

(**A**)Tanimoto Simlarity



(B)T-SNE



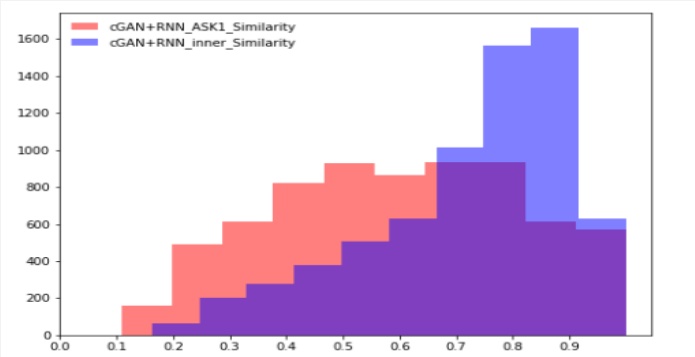
3. Result (Tanimoto Similarity,T-SNE)

cGAN+RNN

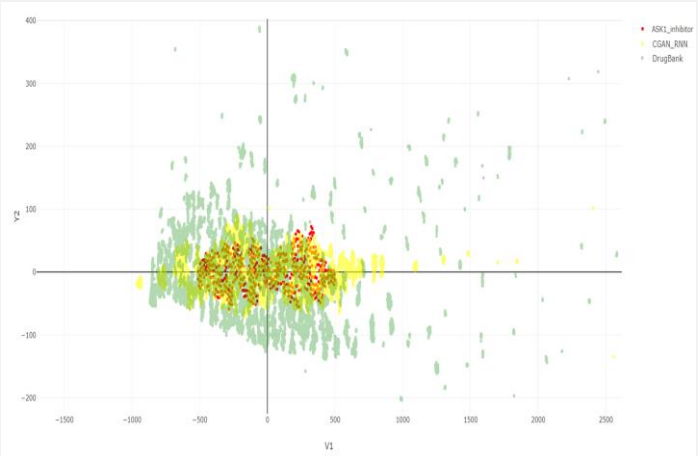
-Input Data: ASK1 Inhibitor + cGAN으로 생성한 데이터

-RNN으로 6,934개 compound 생성

(**A**)Tanimoto Simlarity



(B)T-SNE



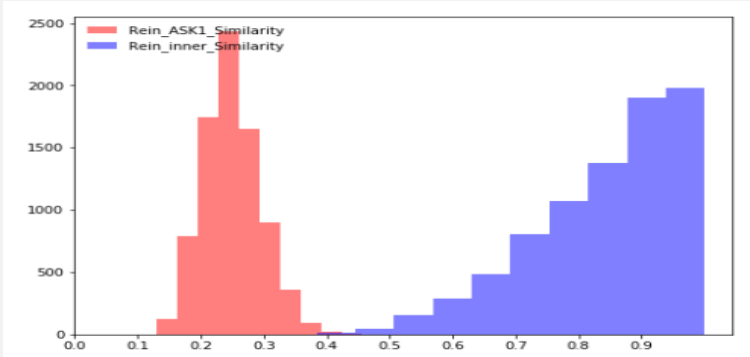
Reinforcement Learning

-Input Data: 1)Generative model: Chembl Database 22

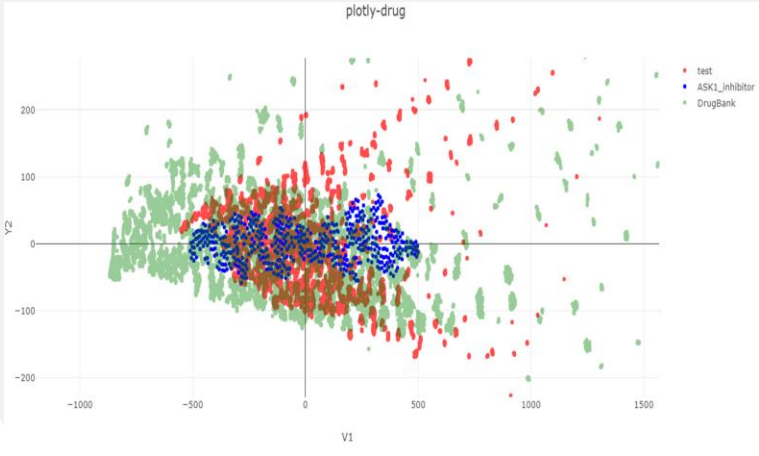
2)prediction model: ASK1 inhibitor

-강화학습으로 18,694개 Compound 생성

(**A**)Tanimoto Simlarity



(B)T-SNE

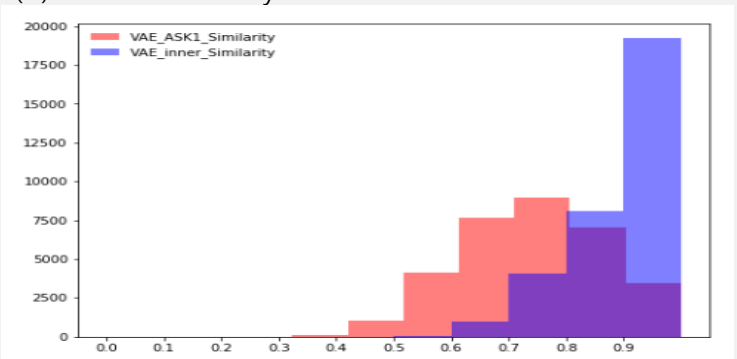


VAE

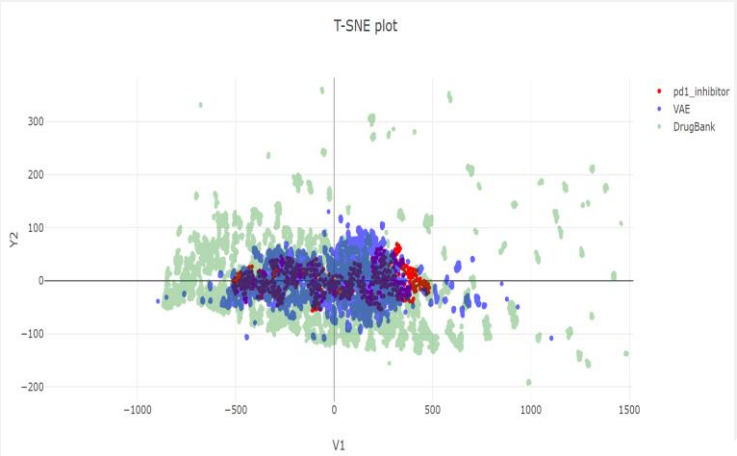
-Input Data: Pretrained Data: ZINC Database, Target Data: ASK1 inhibitor

-VAE로 32,464개 compound 생성

(**A**)Tanimoto Simlarity

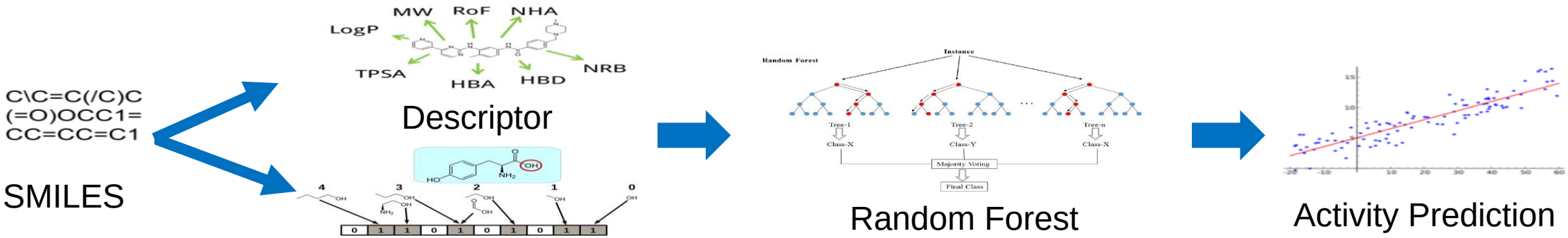


(B)T-SNE

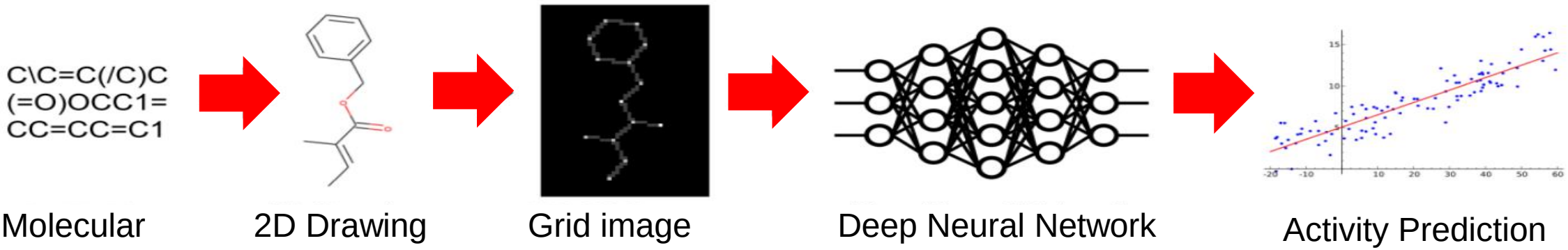


4. QSAR Model

<Random Forest QSAR>



<Chemception>

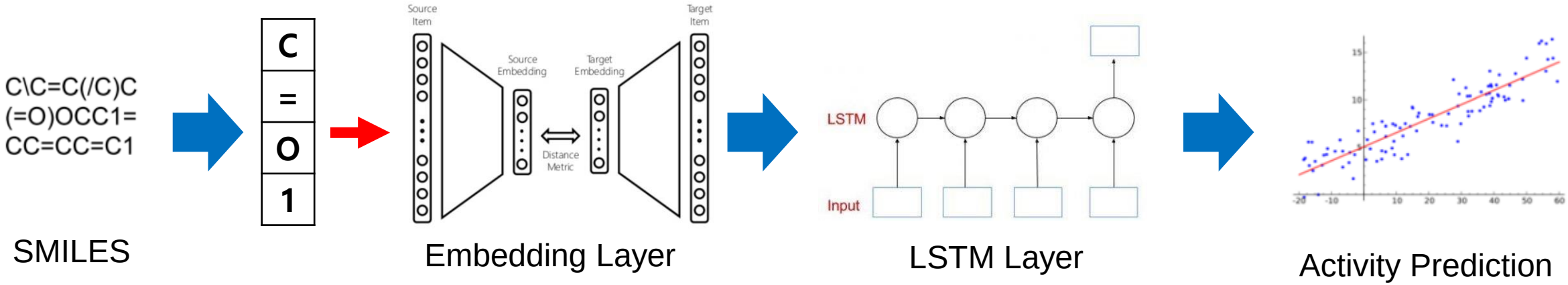


Ref.

1) Chemception: A Deep Neural Network with Minimal Chemistry Knowledge Matches the Performance of Expert-developed QSAR/QSPR Models, **Garrett B. Goh**(2017)
2) How Much Chemistry Does a Deep Neural Network Need to Know to Make Accurate Predictions?, **Garrett B. Goh**,(2017)
3) In silico modelling of permeation enhancement potency in Caco-2 monolayers based on molecular descriptors and random forest, **Søren H.Welling**(2015)

4. Embedded RNN QSAR

<Embedded RNN QSAR>



QSAR Model	Descriptor 사용	Fingerprint 사용	Regression model R2 Score
Embedded RNN	X	X	0.852
Random Forest	O	O	0.848
Chemception(CNN)	X	X	0.840

4. Result

<Random Forest>

Model	The number of Generating	Active(pIC50>=7)	Ratio of Active
GAN	405	46	11.36%
CGAN	218	14	6.42%
GAN+RNN	6769	347	5.13%
CGAN+RNN	6011	282	4.69%

<Chemception>

Model	The number of Generating	Active(pIC50>=7)	Ratio of Active
GAN	405	93	22.96%
CGAN	218	26	11.93%
GAN+RNN	6769	905	13.37%
CGAN+RNN	6011	710	11.81%

<Embedded RNN>

Model	The number of Generating	Active(pIC50>=7)	Ratio of Active
GAN	405	70	17.28%
CGAN	218	23	10.55%
GAN+RNN	6769	532	7.86%
CGAN+RNN	6011	428	7.12%

뒷부분 참조

3-1. Result

GAN

- ASK1 데이터를 3분할하여 GAN으로 543개 Compound 생성

1)Valid SMILES: 543 → 543
2)Canonical SMILES로 변환 후 중복 제거 : 543 → 528개
3)ASk1_inhibitor와 생성한 Compound의 Tanimoto Similitity 0.85 이하 비율 : 0.76

cGAN

- ASK1 데이터를 logP 기준으로 5구간을 나누어 각각 생성
- 294개 compound 생성

1)Valid SMILES: 294(43,57,67,58,69)→ 294개
2)Canonize SMILES 변환 후 중복 SMILES 제거: 294→ 287개
3)각 생성한 Compound들의 logP 평균: 2.66, 4.18, 4.46, 5.16, 5.85
4)ASk1_inhibitor와 생성한 Compound의 Tanimoto Similarity 0.85 이하 비율: 0.75

GAN+RNN

-Input Data: ASk1 Inhibitor + GAN으로 생성한 데이터
-RNN으로 10만개의 compound 생성

1)Valid SMILES: 100,000 → 65,629개
2)Canonize SMILES 변환 후 중복 SMILES 제거: 65,629 → 7,638개
3)ASk1 inhibitor와 생성한 Compound의 Tanimoto Similarity 0.85 이하 비율: 0.88

cGAN+RNN

-Input Data: ASk1 Inhibitor + cGAN으로 생성한 데이터
-RNN으로 10만개의 compound 생성

1)Valid SMILES: 100,000 → 67,687개
2)Canonize SMILES 변환 후 중복 SMILES 제거: 65,629 → 6,934개
3)ASk1 inhibitor와 생성한 Compound의 Tanimoto Similarity 0.85 이하 비율: 0.86

3-1. Result

Reinforcement Learning

-Input Data: Generative model: Chembl Database 22, prediction model: ASK1 inhibitor

-강화학습으로 50,000개의 Compound 생성

1)Active Probability가 0.8 이상인 Molecular 50,000개 생성

2)Valid SMILES: 50,000 → 18,691

3)Canonize SMILES 변환 후 중복 SMILES 제거: 18,691 → 18,691

4)ASk1 inhibitor와 생성한 Compound의 Tanimoto Similarity 0.85 이하 비율: 1.00

VAE

-Input Data: Pretrained Data: ZINC Database, Target Data: ASK1 inhibitor

-ASK1 Inhibitor 1개당 100,000개씩 생성

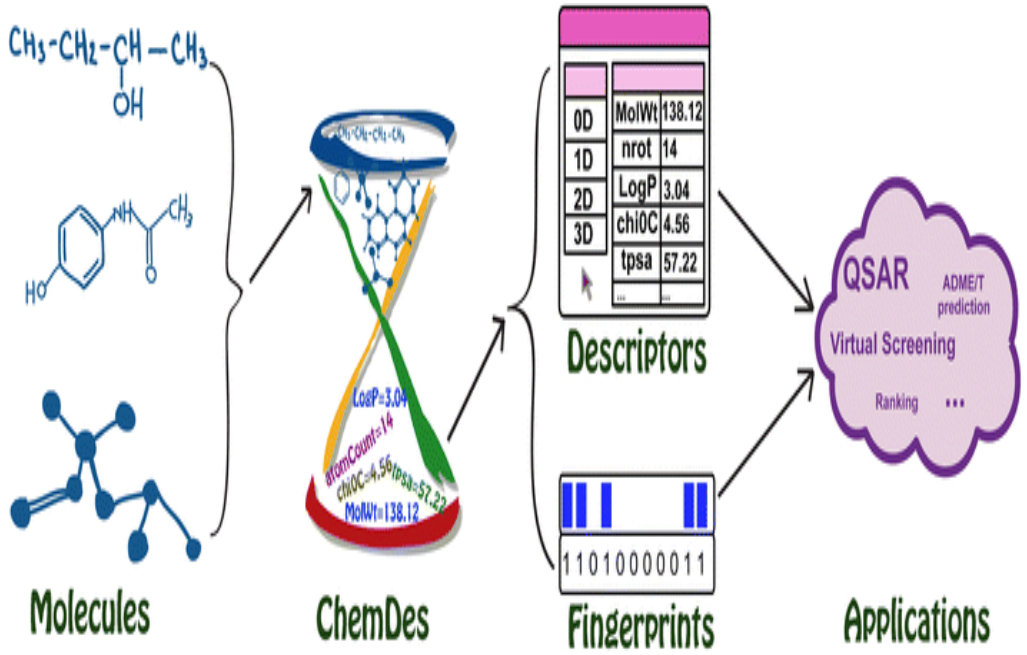
1)Valid SMILES: 36,046 → 32,464

2)Canonize SMILES 변환 후 중복 SMILES 제거: 32,464 → 32,464

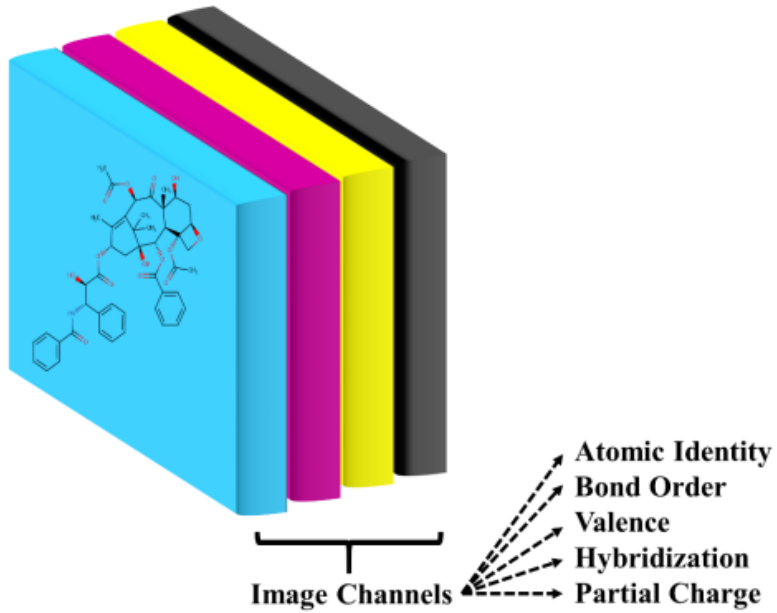
3)ASk1_inhibitor와 생성한 Compound의 tanimoto Similarity 0.85 이하 비율: 0.90

4. QSAR

<Random Forest QSAR>



<Chemception>



Ref.

1) Chemception: A Deep Neural Network with Minimal Chemistry Knowledge Matches the Performance of Expert-developed QSAR/QSPR Models, **Garrett B. Goh**(2017)
2) How Much Chemistry Does a Deep Neural Network Need to Know to Make Accurate Predictions?, **Garrett B. Goh**,(2017)
3) In silico modelling of permeation enhancement potency in Caco-2 monolayers based on molecular descriptors and random forest, **Søren H.Welling**(2015)