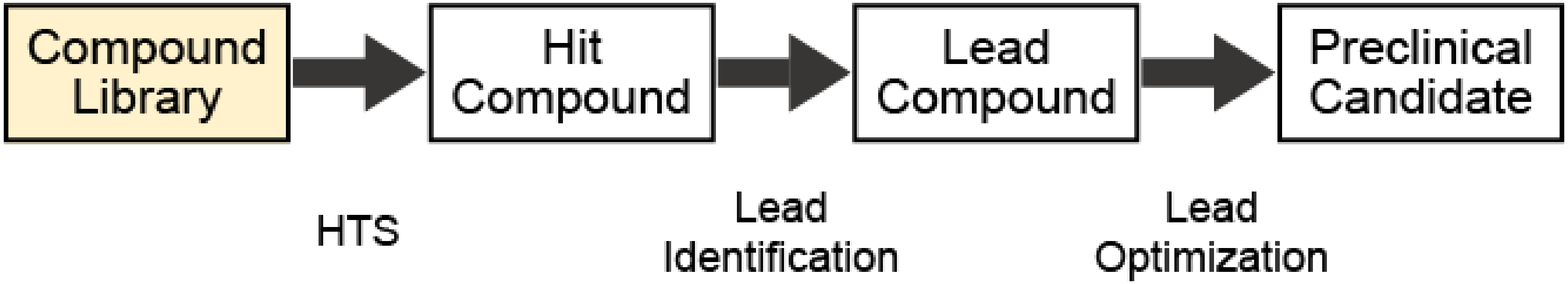
Deep Generative Model for De Novo Drug Design

2019/11/14

M2 Koki Sasamoto

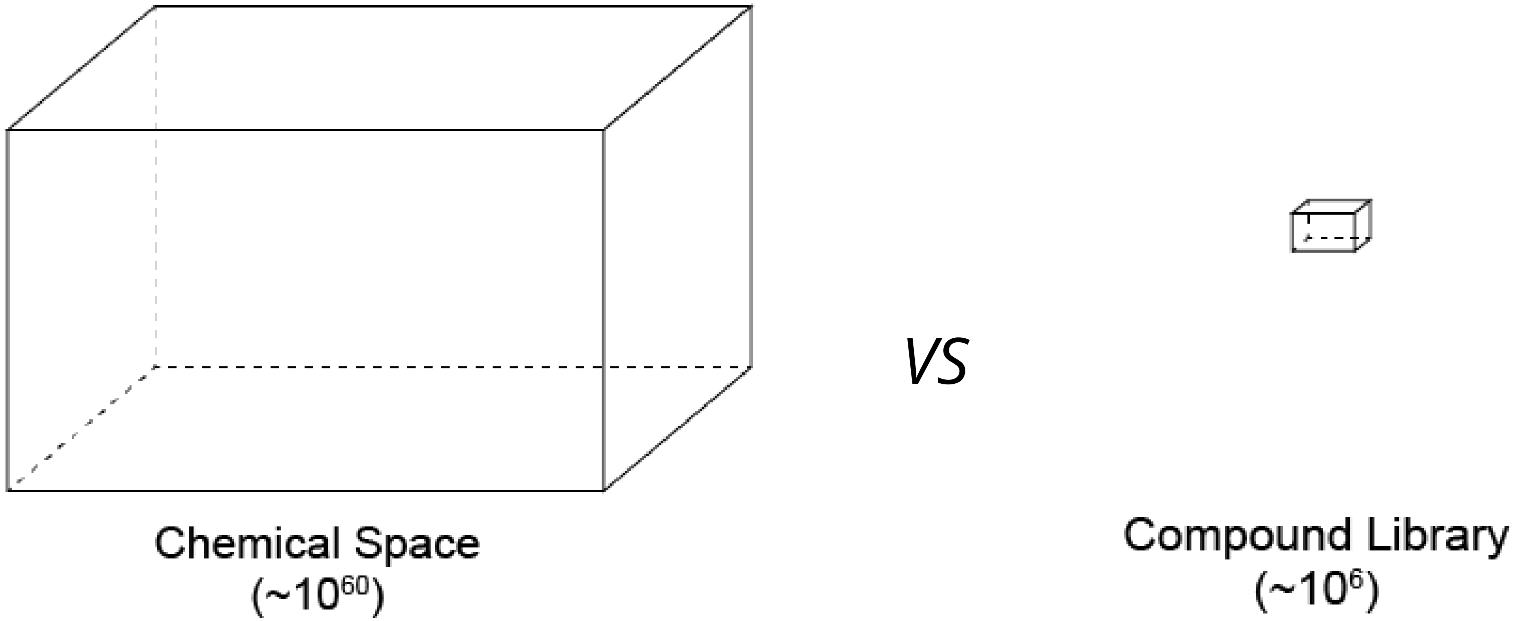
Drug discovery

- Flow chart of drug discovery process



- Compound library is used to screen hit compound.
- Good hit compound reduces time and money.
- **Diverse** and **high-quality** compound library is **needed**.

Chemical space



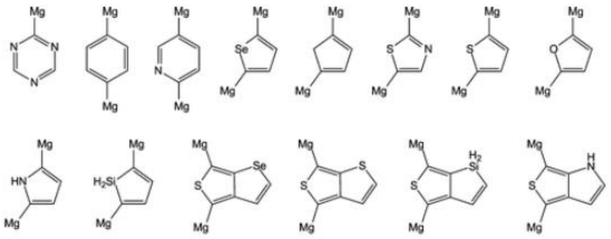
VS

- Chemical space is vast, and **only a tiny fraction was collected** as compound libraries.

J. Comput.- Aided Mol. Des. **2013**, *27*, 675−679.

Construction of virtual library

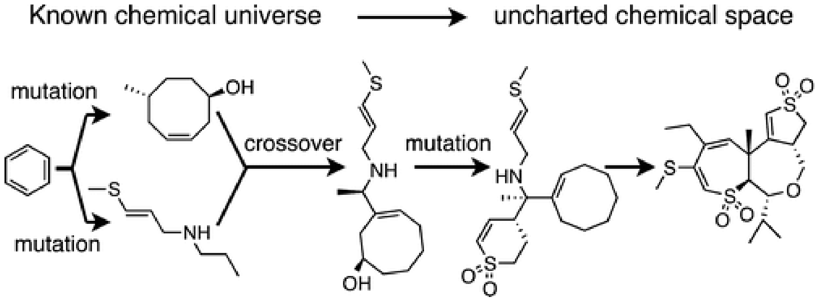
Building block



combination

J. Phys. Chem. Lett., **2011**, *2*, 2241-2251.

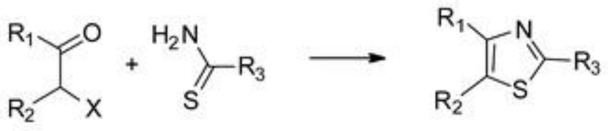
Genetic algorithm



J. Am. Chem. Soc., **2013**, *135*, 7296-7303.

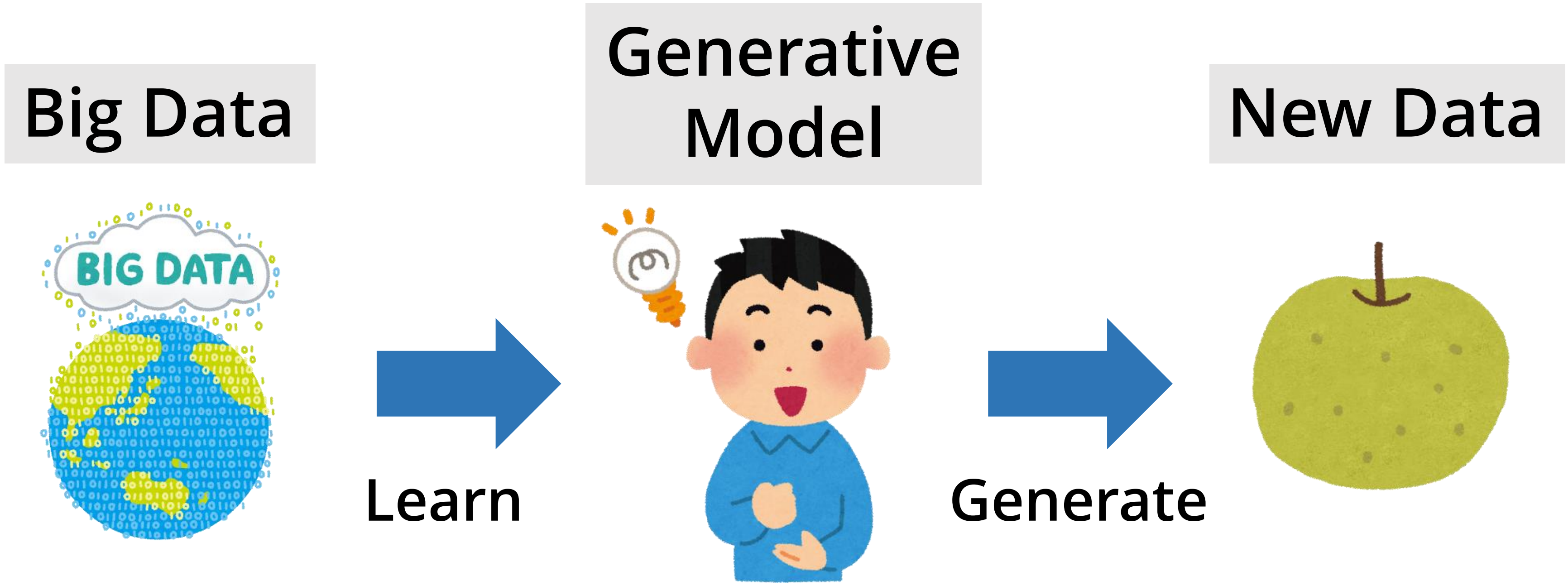
Reaction-based rule

```
[#6:6]-[C;R0:1](=[OD1])-[CH1;R0:5](-[#6:7])-[*;#17,#35,#53].[NH2:2]-[C:3]=[SD1:4]>>
[c:1]2(-[#6:6]):[n:2]:[c:3]:[s:4][c:5]([#6:7]):2
```



J. Chem. Inf. Model, **2011**, *51*, 3093-3098.

Deep generative model



Big Data

Learn

Generative Model

Generate

New Data

・**Generative model generates realistic data** from feature of data.

→ Drug-like molecules can be generated by generative model learning features of biologically-active compounds.

Contents

1. Deep learning methods in drug design

   - RNN

   - RNN with RL (ReLeaSE)

   - VAE

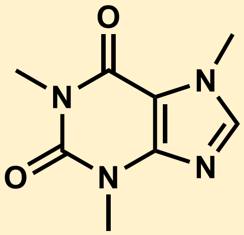
   - Graph / GAN (MolGAN)

2. Application in drug discovery (GENTRL)

3. Summary

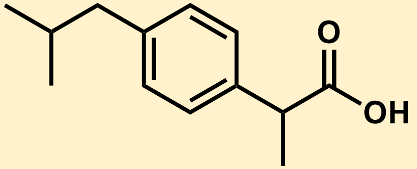
SMILES

- **Examples of SMILES representation**



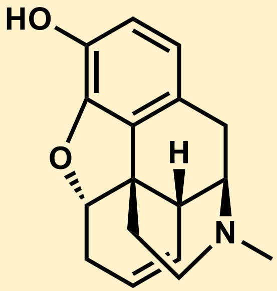
Caffeine

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

M. H. S. Segler *et al.*, *ACS Cent. Sci.*, **2018**, *4*, 120-131.

De novo drug design by RNN

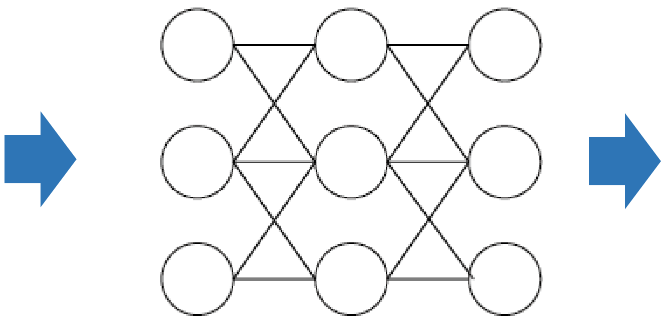
- Generation of sentence

learn English grammar

RNN

input

Chemistry is ...



output

O Chemistry is “important”

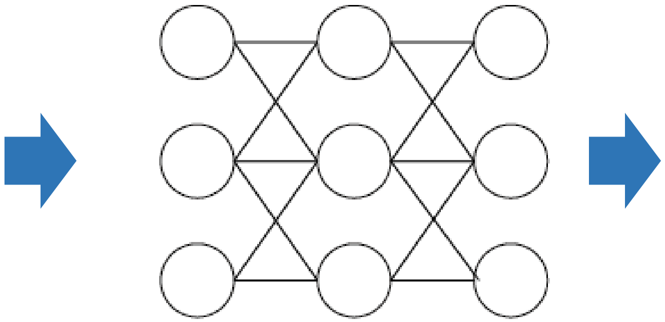
O Chemistry is “fascinating”

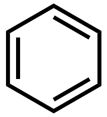
X Chemistry is “potato”

X Chemistry is “runs”

- Generation of chemical structure

c
(SMILES)



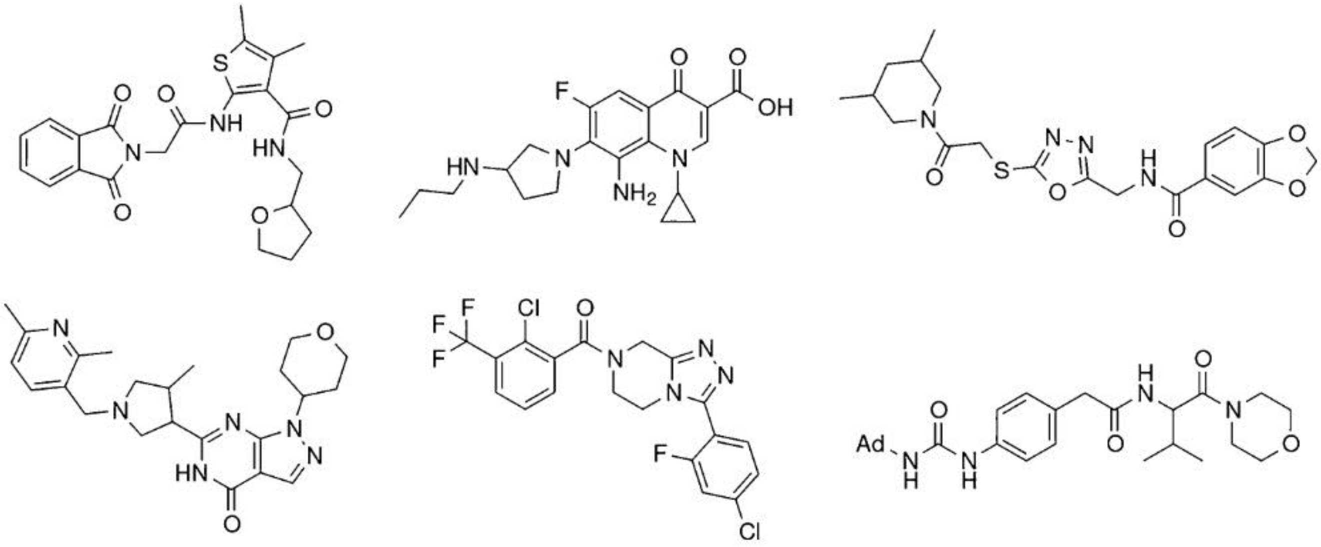


c1ccccc1

M. H. S. Segler *et al.*, *ACS Cent. Sci.*, **2018**, *4*, 120-131.

De novo drug design by RNN

Examples of generated novel molecules

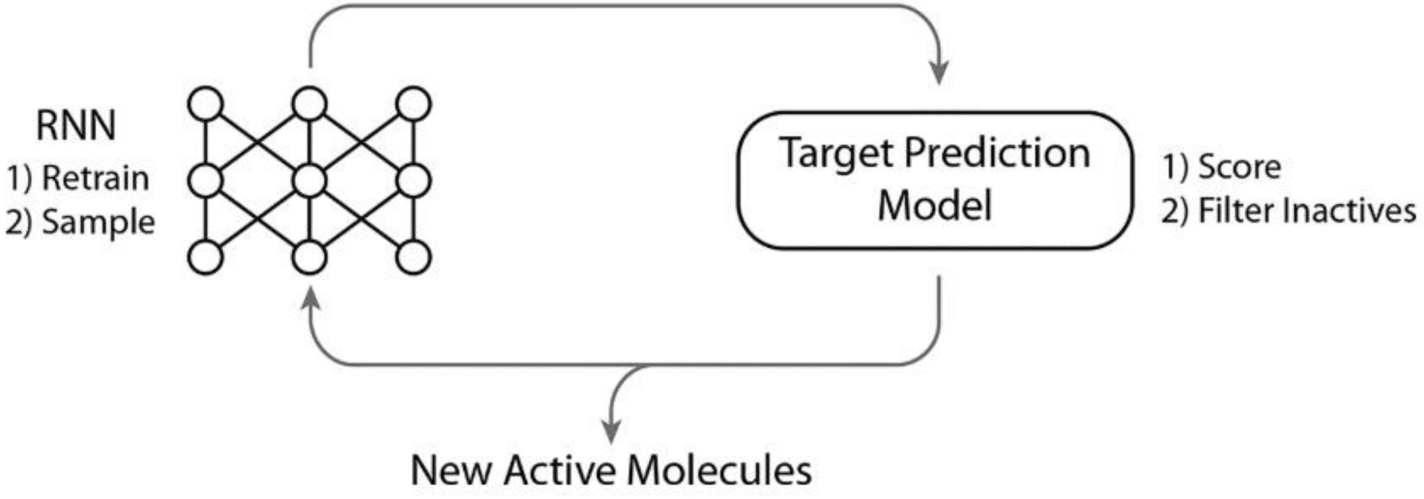


- **976327** molecules were **generated**.
- **847955** molecules were **novel**.
- **75%** of new molecules were **highly scored** ("core" or "backup") by AstraZeneca filter.

M. H. S. Segler *et al.*, *ACS Cent. Sci.*, **2018**, *4*, 120-131.

De novo design cycle

▶ Scheme



"**Synthesis**" … molecule generation

"**Virtual Assay**" … best molecule selection by machine learning

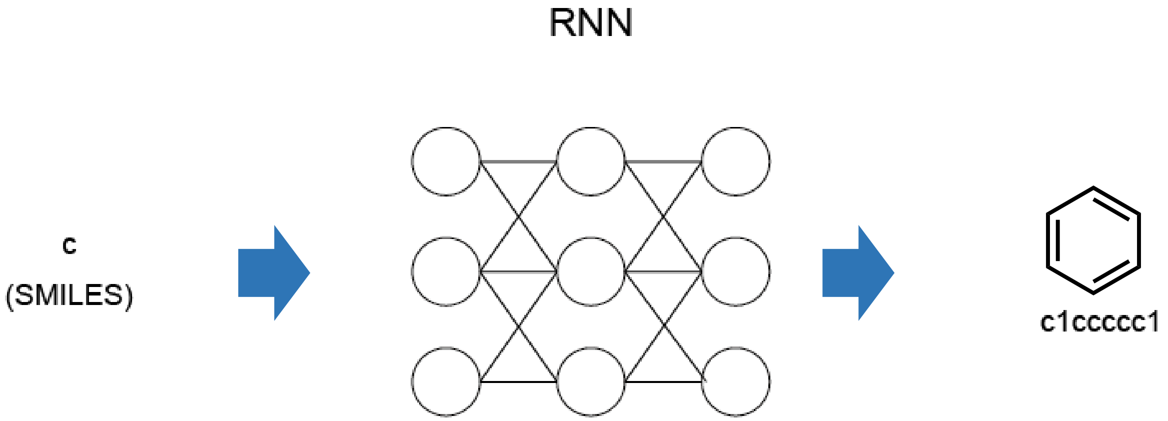
"**Design**" … retraining RNN model by best molecules

⬇

6% of known active molecules were **re-generated.**

M. H. S. Segler *et al.*, *ACS Cent. Sci.*, **2018**, *4*, 120-131.

De novo drug design by RNN



Data **541555** bioactive molecules

Fine-tune **25** molecules with known agonistic activity
on **RXR** (retinoid X receptor) and/or
PPAR (peroxisome proliferator-activated receptor)

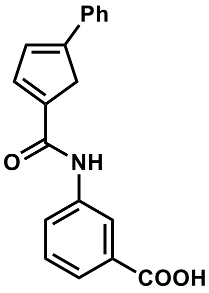
Result **1000** molecules (90% were valid and novel)

5 molecules were **synthesized** and **tested** in vitro.

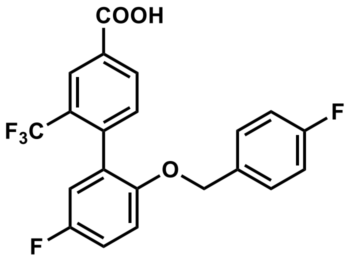
G. Schneider *et al*., *Mol. Inf*., **2018**, *37*, 1700153.

De novo drug design by RNN

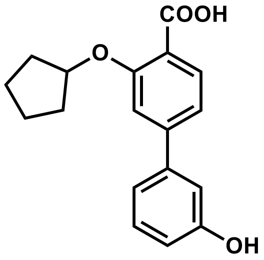
- **Synthesized novel molecules and these bioactivity**



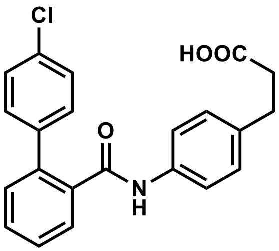
1



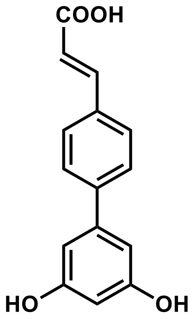
2



3

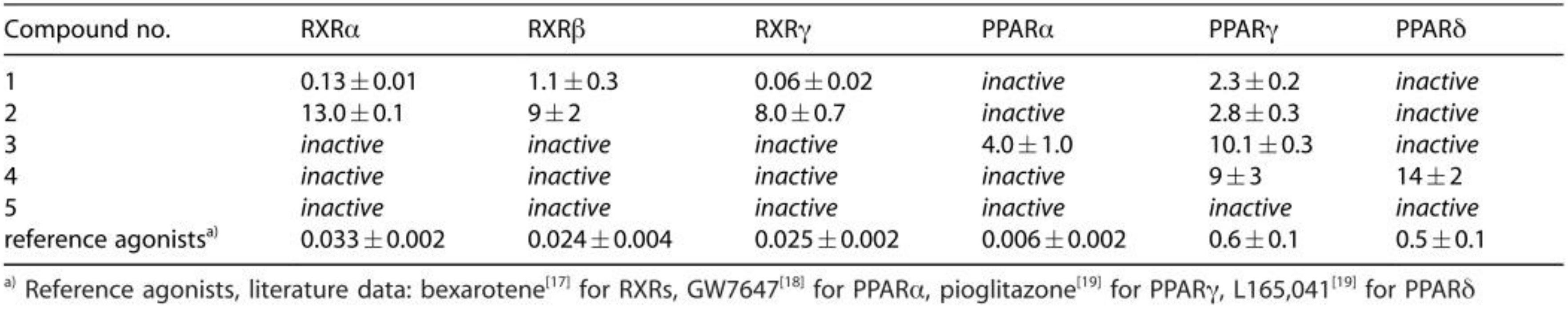


4



5

Bioactivity (EC$_{50}$ / uM)



Compound no.	RXRα	RXRβ	RXRγ	PPARα	PPARγ	PPARδ
1	0.13 ± 0.01	1.1 ± 0.3	0.06 ± 0.02	*inactive*	2.3 ± 0.2	*inactive*
2	13.0 ± 0.1	9 ± 2	8.0 ± 0.7	*inactive*	2.8 ± 0.3	*inactive*
3	*inactive*	*inactive*	*inactive*	4.0 ± 1.0	10.1 ± 0.3	*inactive*
4	*inactive*	*inactive*	*inactive*	*inactive*	9 ± 3	14 ± 2
5	*inactive*	*inactive*	*inactive*	*inactive*	*inactive*	*inactive*
reference agonists[a)]	0.033 ± 0.002	0.024 ± 0.004	0.025 ± 0.002	0.006 ± 0.002	0.6 ± 0.1	0.5 ± 0.1

[a)] Reference agonists, literature data: bexarotene[17] for RXRs, GW7647[18] for PPARα, pioglitazone[19] for PPARγ, L165,041[19] for PPARδ

G. Schneider *et al*., *Mol. Inf*., **2018**, *37*, 1700153.

De novo drug design by RNN

Pros

- Diverse set of molecules could be generated.
- Generated molecules had drug-like properties.

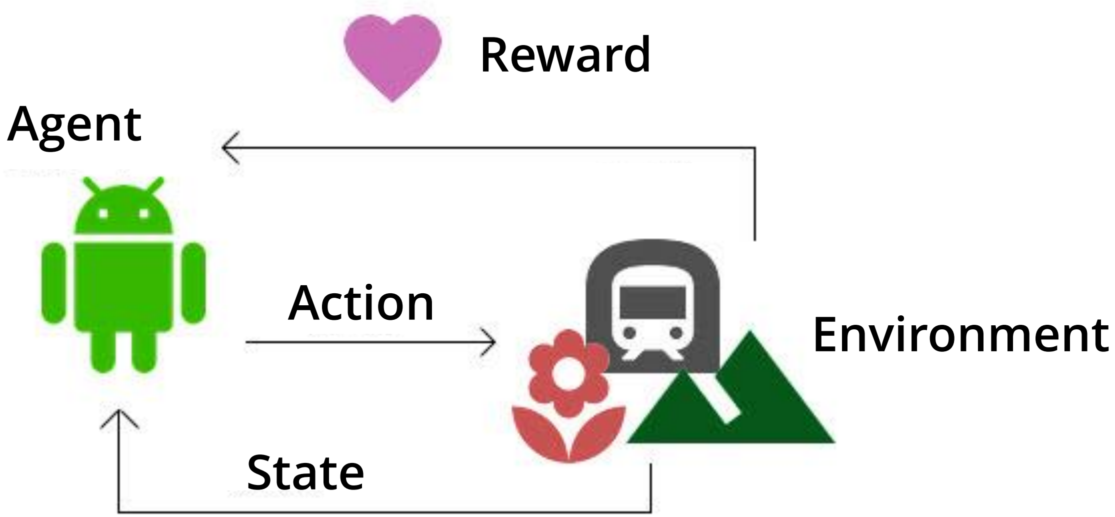
Cons

- Chemical space was restricted by training set.
- Properties of generated molecules couldn’t be controlled.

➡ **Reinforcement Leaning (RL)**

Reinforcement Learning (RL)

Scheme of Reinforcement Learning

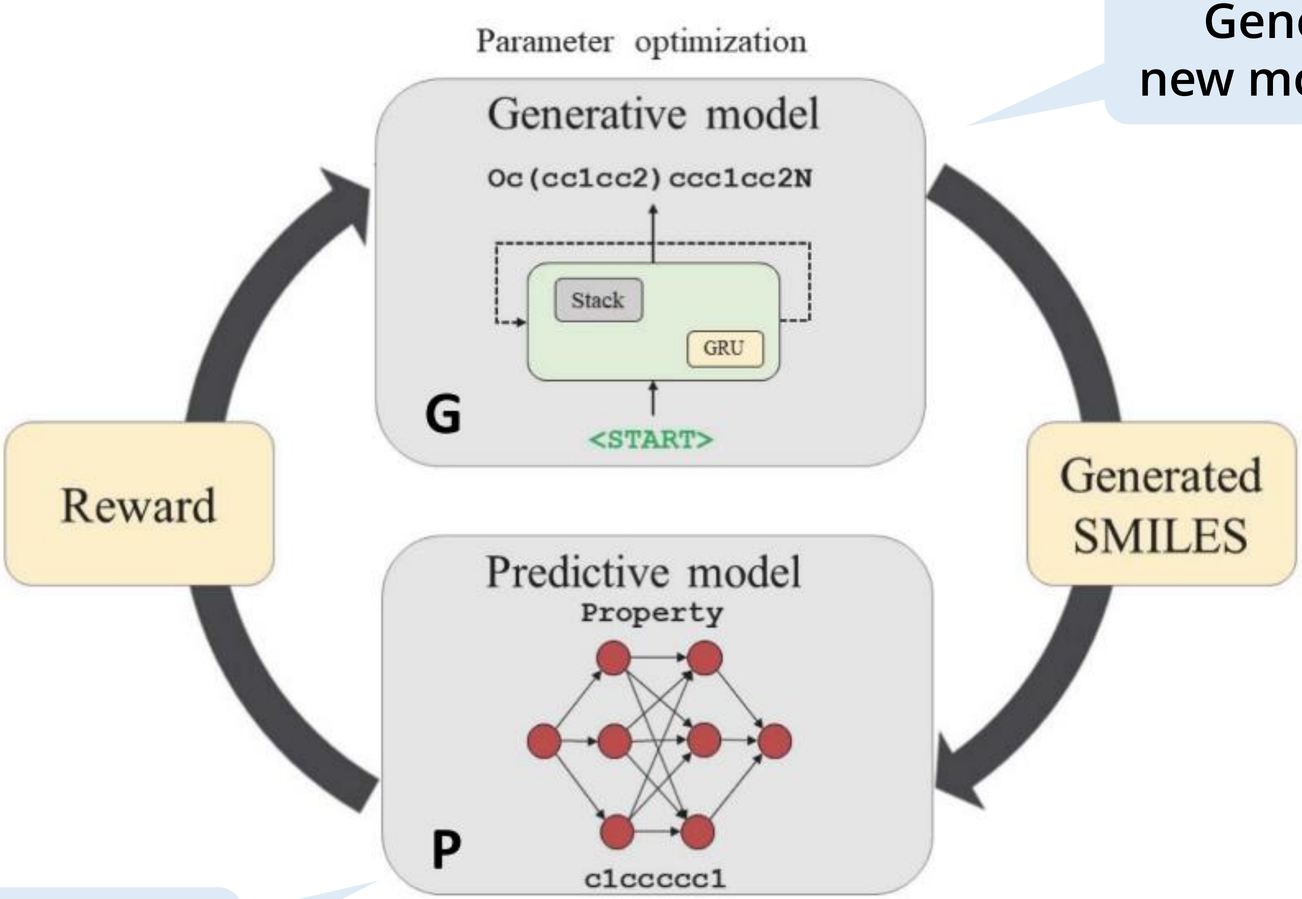


Application

AlphaGo

De novo drug design by RL

- **Scheme of “ReLeaSE”**



Generate new molecules

Predict properties

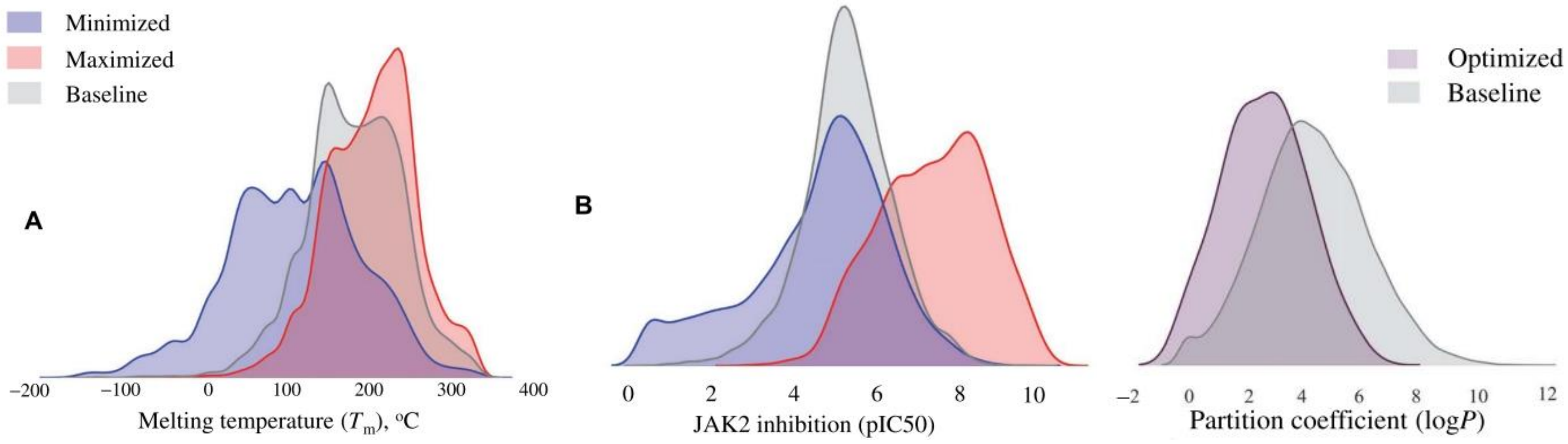
Popova *et al.*, *Sci. Adv.*, **2018**, *4*, No. eaap7885.

De novo drug design by RL

Target Properties

- **Tm** (Melting point)
- **logP** (n-octanol / water partition coefficient)
- **pIC$_{50}$** for **JAK2** (janus protein kinase 2)

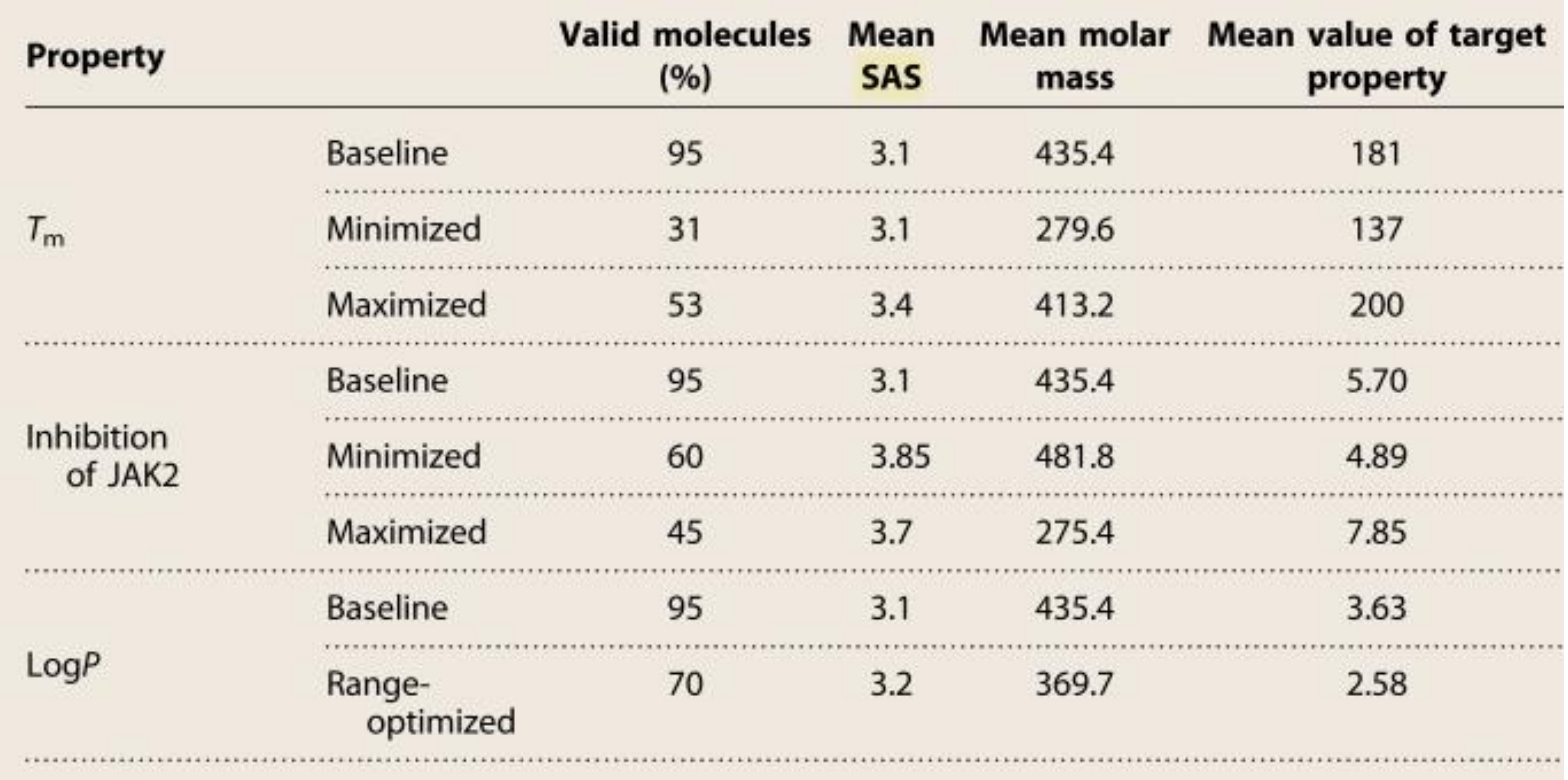
Distribution of predicted properties



Popova *et al*., *Sci. Adv*., **2018**, *4*, No. eaap7885.

De novo drug design by RL

- **Analysis of generated molecules**



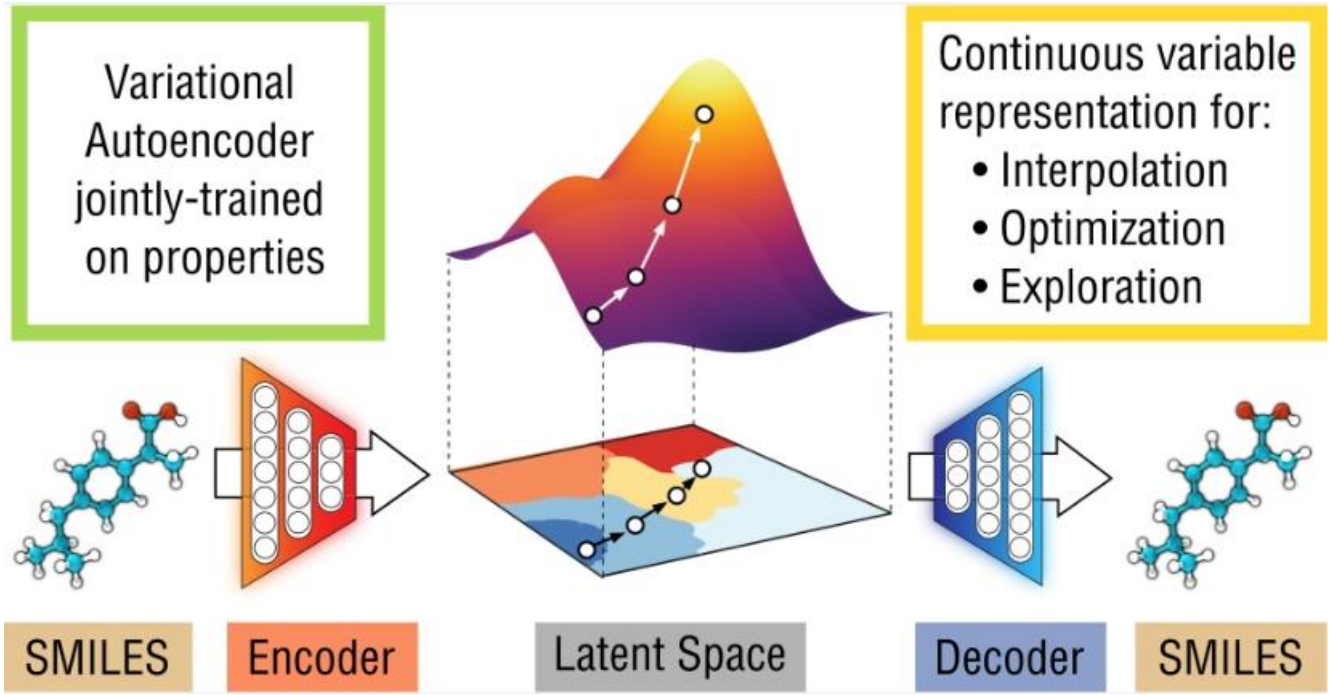
Property		Valid molecules (%)	Mean SAS	Mean molar mass	Mean value of target property
T_m	Baseline	95	3.1	435.4	181
	Minimized	31	3.1	279.6	137
	Maximized	53	3.4	413.2	200
Inhibition of JAK2	Baseline	95	3.1	435.4	5.70
	Minimized	60	3.85	481.8	4.89
	Maximized	45	3.7	275.4	7.85
Log*P*	Baseline	95	3.1	435.4	3.63
	Range-optimized	70	3.2	369.7	2.58

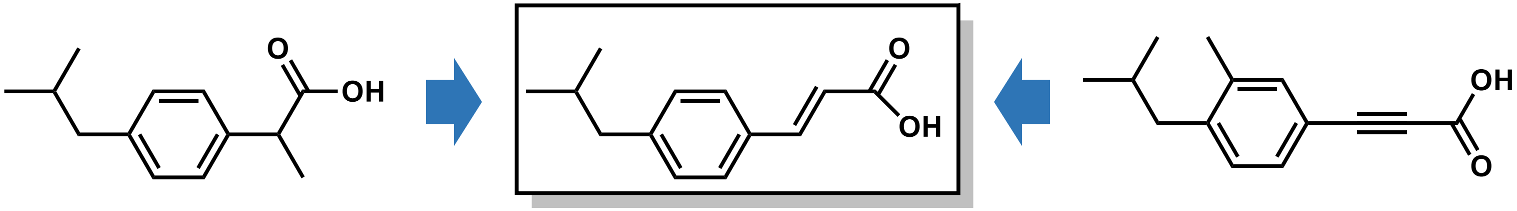
(SAS = synthetic accessibility score)

With reinforcement learning, the proportion of valid molecules was lowered.

Popova *et al*., *Sci. Adv*., **2018**, *4*, No. eaap7885.

De novo drug design by VAE

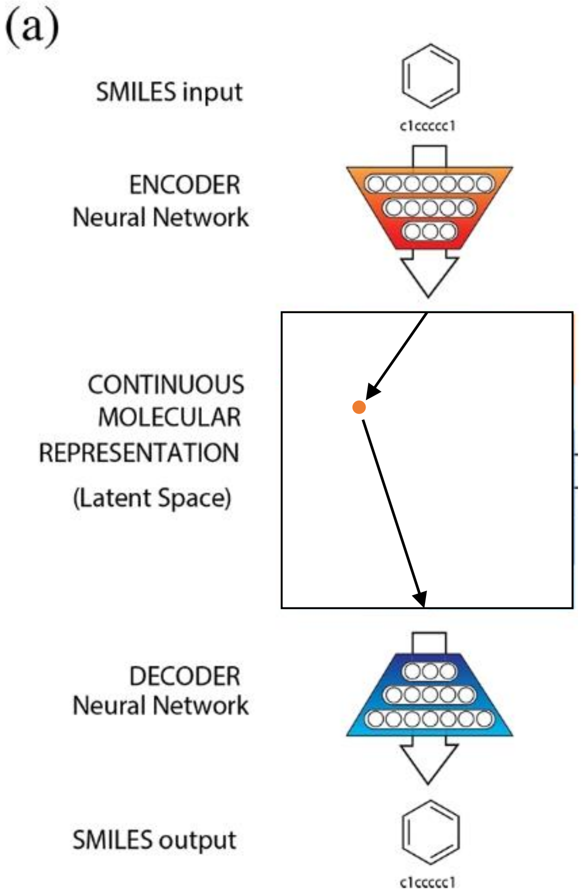

Variational Autoencoder jointly-trained on properties
Continuous variable representation for:
• Interpolation
• Optimization
• Exploration
SMILES
Encoder
Latent Space
Decoder
SMILES

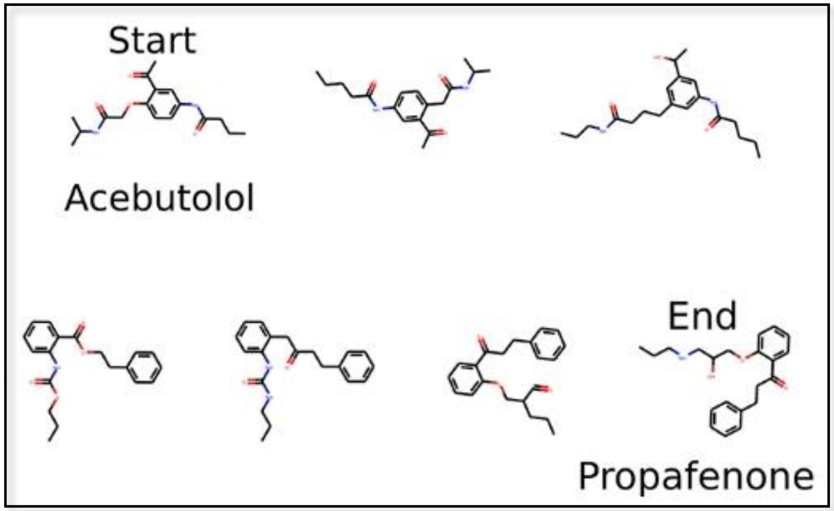


Generated

R. Gómez-Bombarelli *et al.*, *ACS Cent. Sci*, **2018**, *4*, 268-276.

Encoder and Decoder



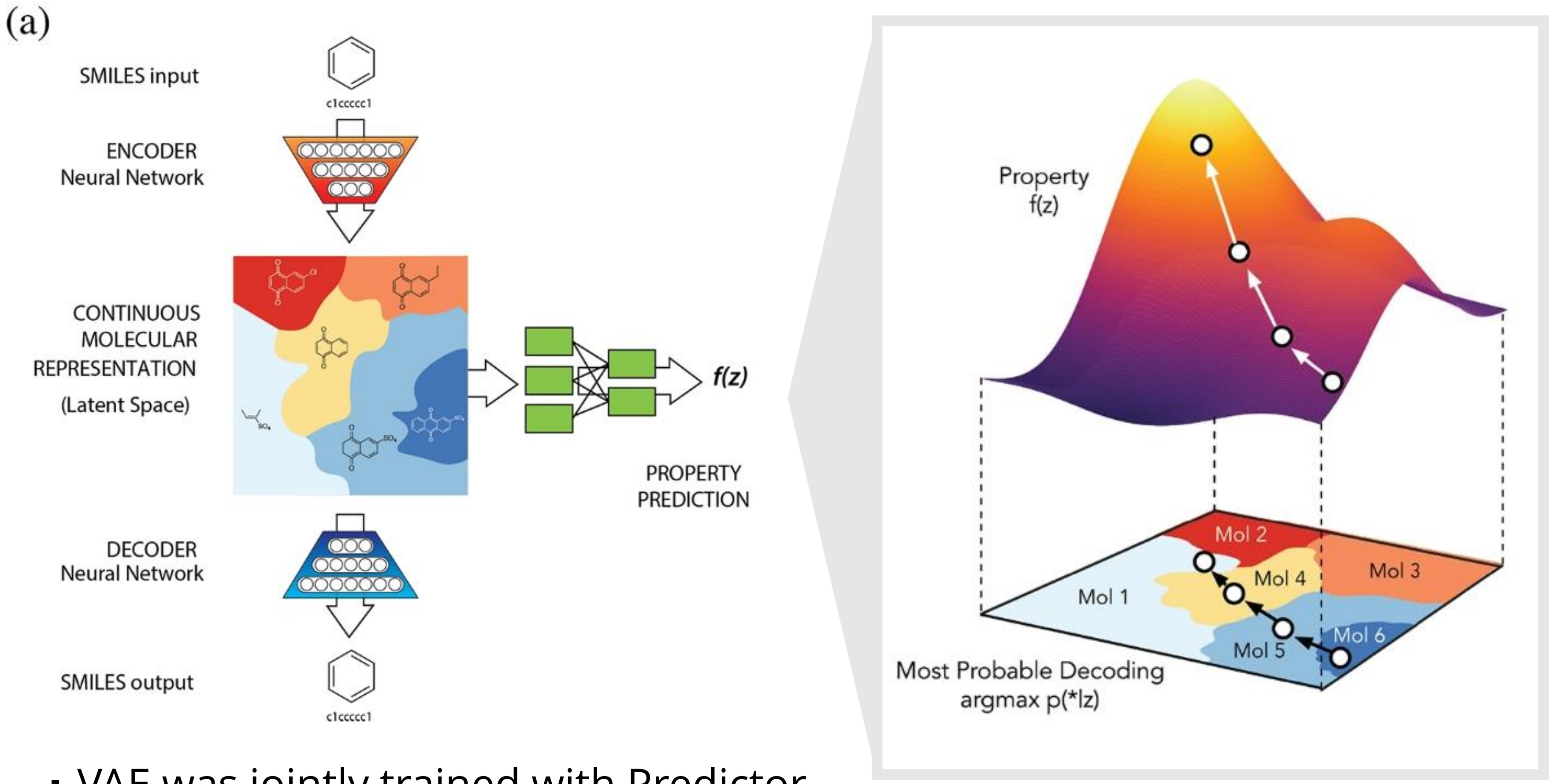


interpolation between two
molecules in latent space

- VAE learns about characteristic feature of a training set.
- **Similar molecules** were mapped **close together** in latent space.

R. Gómez-Bombarelli *et al.*, *ACS Cent. Sci*, **2018**, *4*, 268-276.

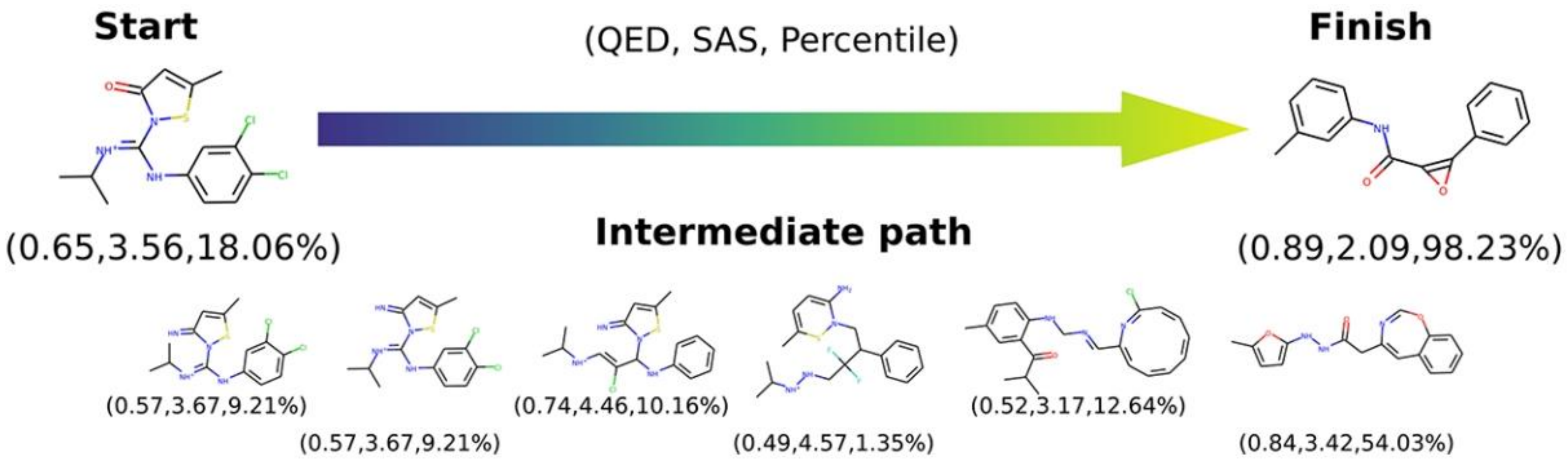
Predictor



- VAE was jointly trained with Predictor.
- 7,500,000 molecules were generated from **250,000** samples.

R. Gómez-Bombarelli *et al.*, *ACS Cent. Sci*, **2018**, *4*, 268-276.

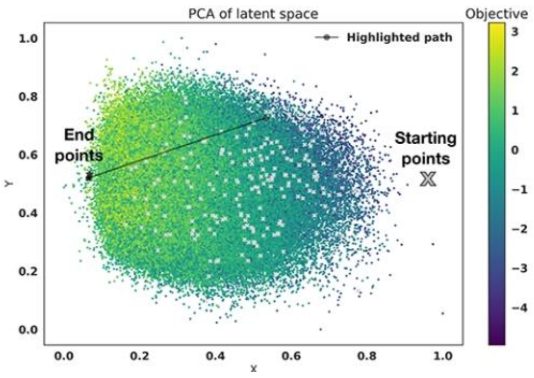
De novo drug design by VAE



- VAE **optimized** (**5 * QED – * SAS**).

  (QED = Qualitative Estimate of Drug-likeness, SAS = Synthetic Accessibility Score)

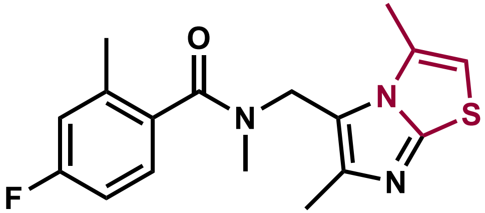
- Molecular optimization was achieved efficiently by gradient-based search.

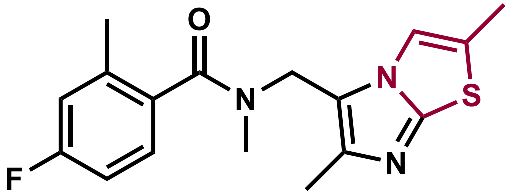


R. Gómez-Bombarelli *et al.*, *ACS Cent. Sci*, **2018**, *4*, 268-276.

Problems of SMILES representation

- **SMILES is not designed to capture molecular similarity.**

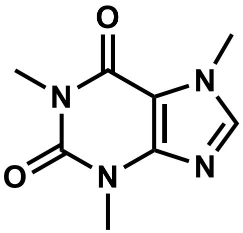




Cc1cn(CN(C)C(=O)c3ccc(F)cc3C)c(C)c(C)nc2s1

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

- **SMILES is not robust to small molecules.**



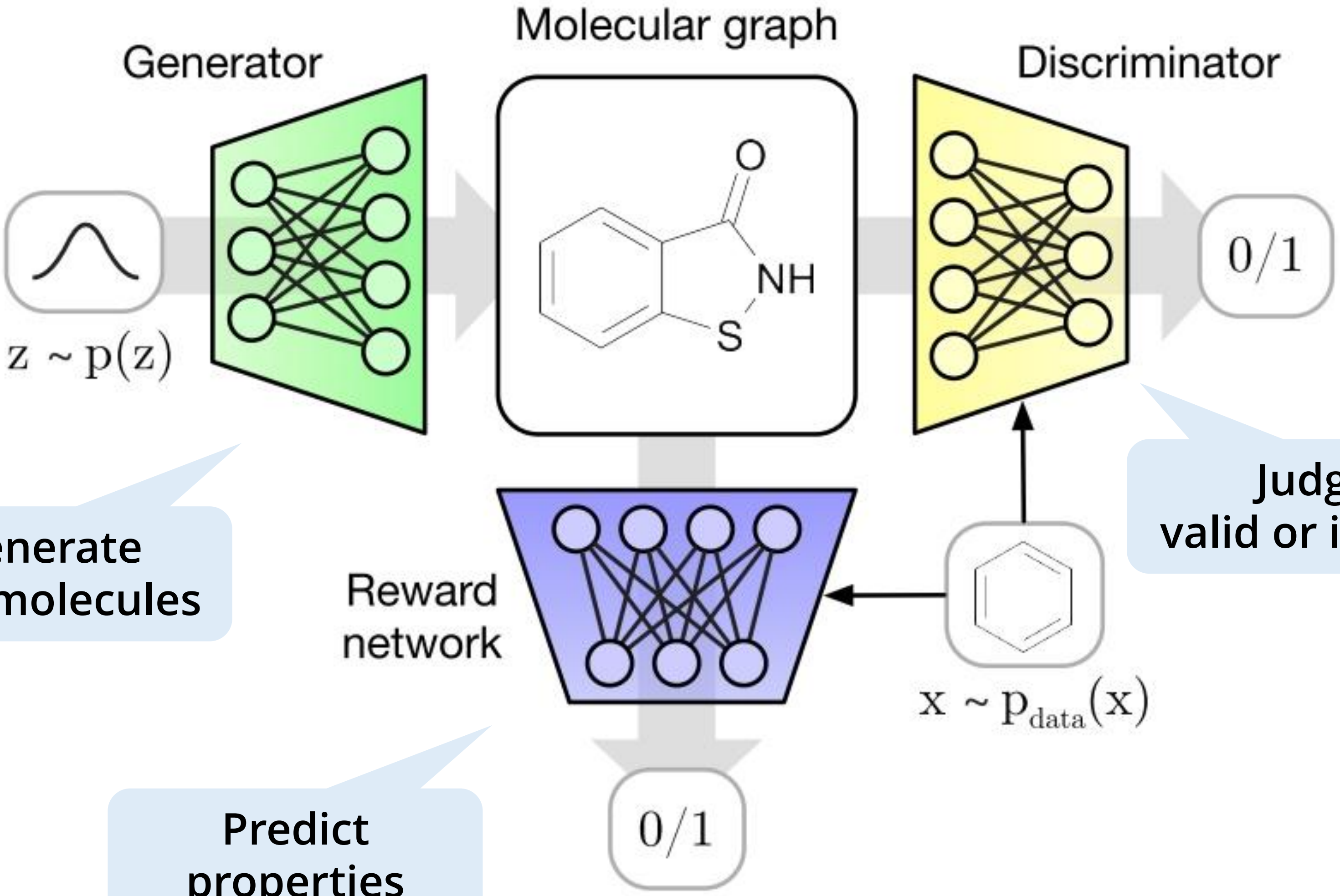
invalid representation

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

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

MolGAN

- Scheme of MolGAN



Generate new molecules

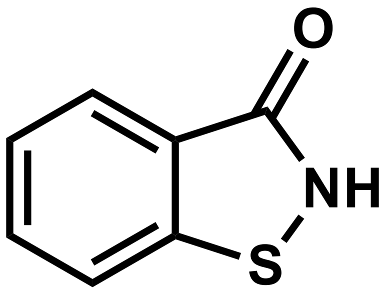
Judge valid or invalid

Predict properties

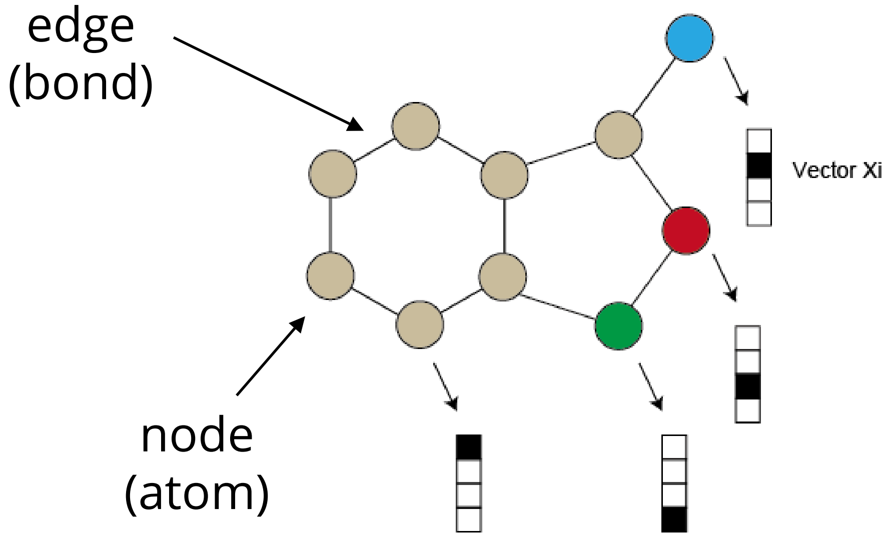
arXiv:1805.11973

Graph representation

Chemical Structure



Molecular Graph

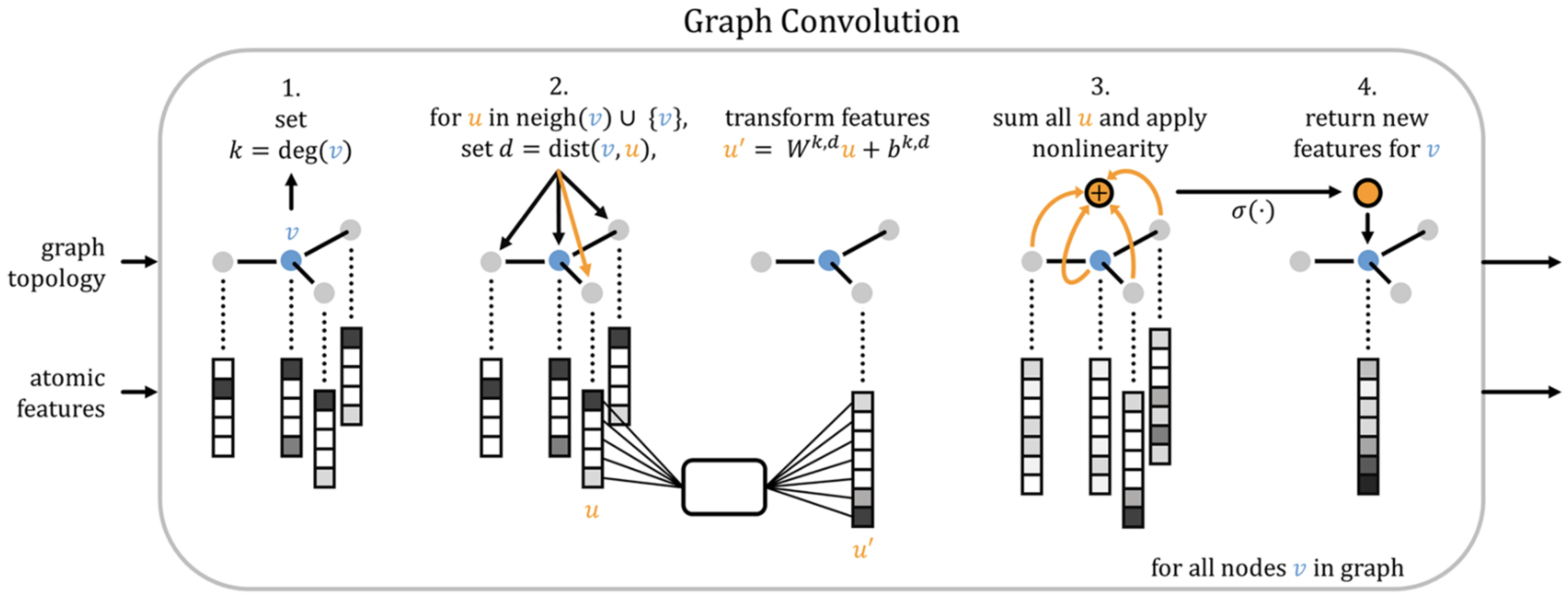

edge (bond)

node (atom)

Vector Xi

- **Graph** … collection of **nodes** and **edges**
- Machine learning model **don't have to learn rules** of molecular representations.

Graph convolution


Graph Convolution
graph topology
atomic features
1. set $k = \deg(v)$
v
2. for u in $\text{neigh}(v) \cup \{v\}$, set $d = \text{dist}(v, u)$,
u
transform features $u' = W^{k,d}u + b^{k,d}$
u'
3. sum all u and apply nonlinearity
$\sigma(\cdot)$
4. return new features for v
for all nodes v in graph

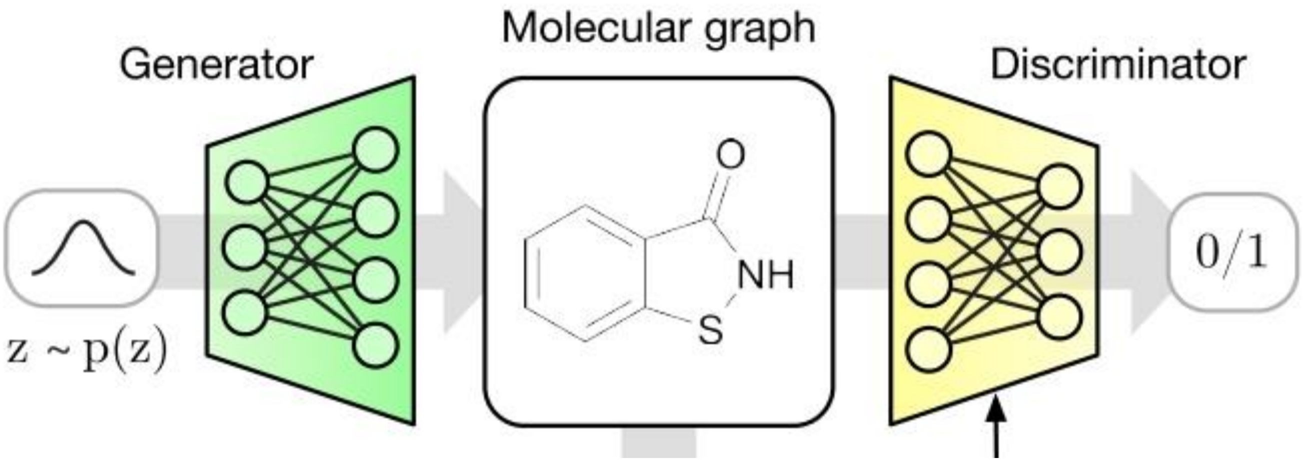
- New vector = self +adjacent vector

→ New vector **includes** the information of **the surrounding environment**.

H. Altae-Tran, *et al.*, *ACS Cent. Sci.* **2017**, *3*, 283-293.

GAN

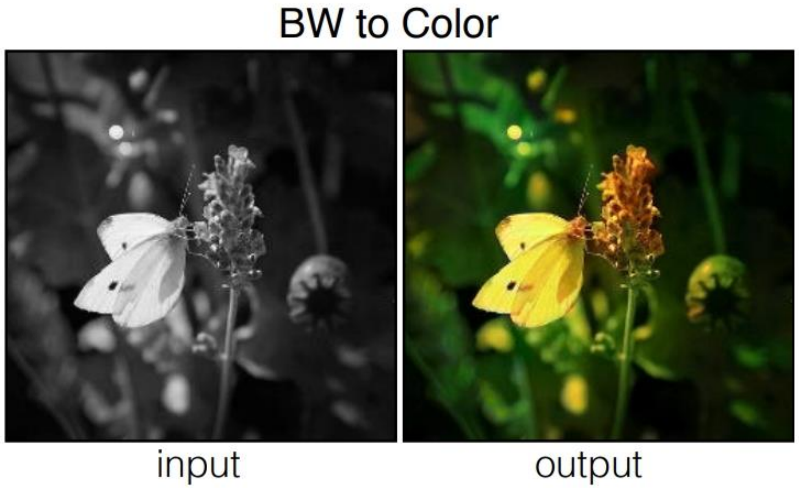
- **Scheme of Generative Adversarial Network (GAN)**



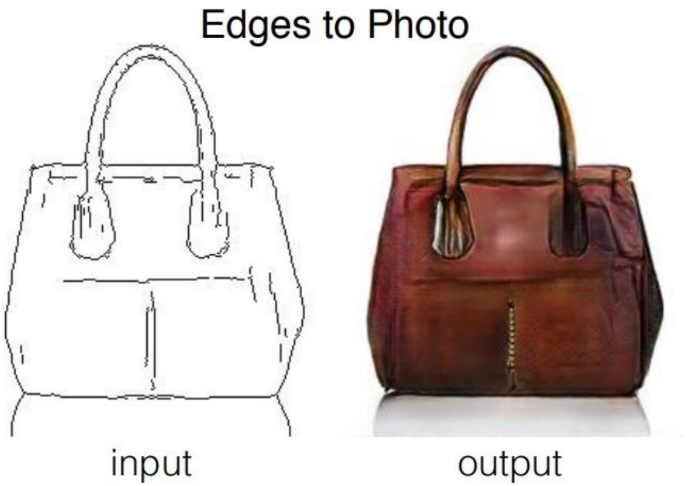
Manufacture of counterfeit money vs Police

GAN

https://arxiv.org/abs/1809.11096



BW to Color

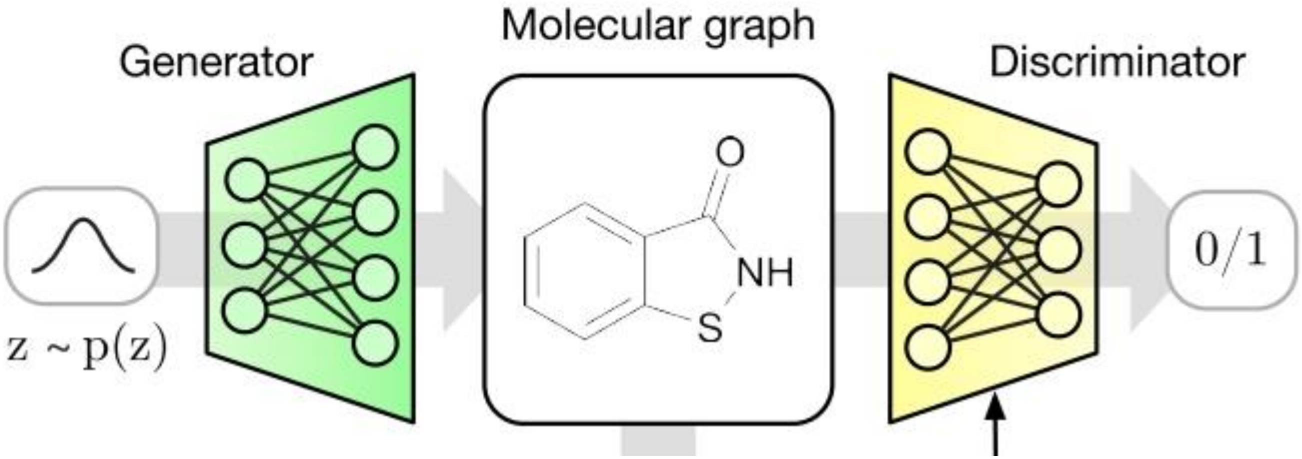


Edges to Photo

https://arxiv.org/abs/1611.07004

GAN

- **Scheme of Generative Adversarial Network (GAN)**



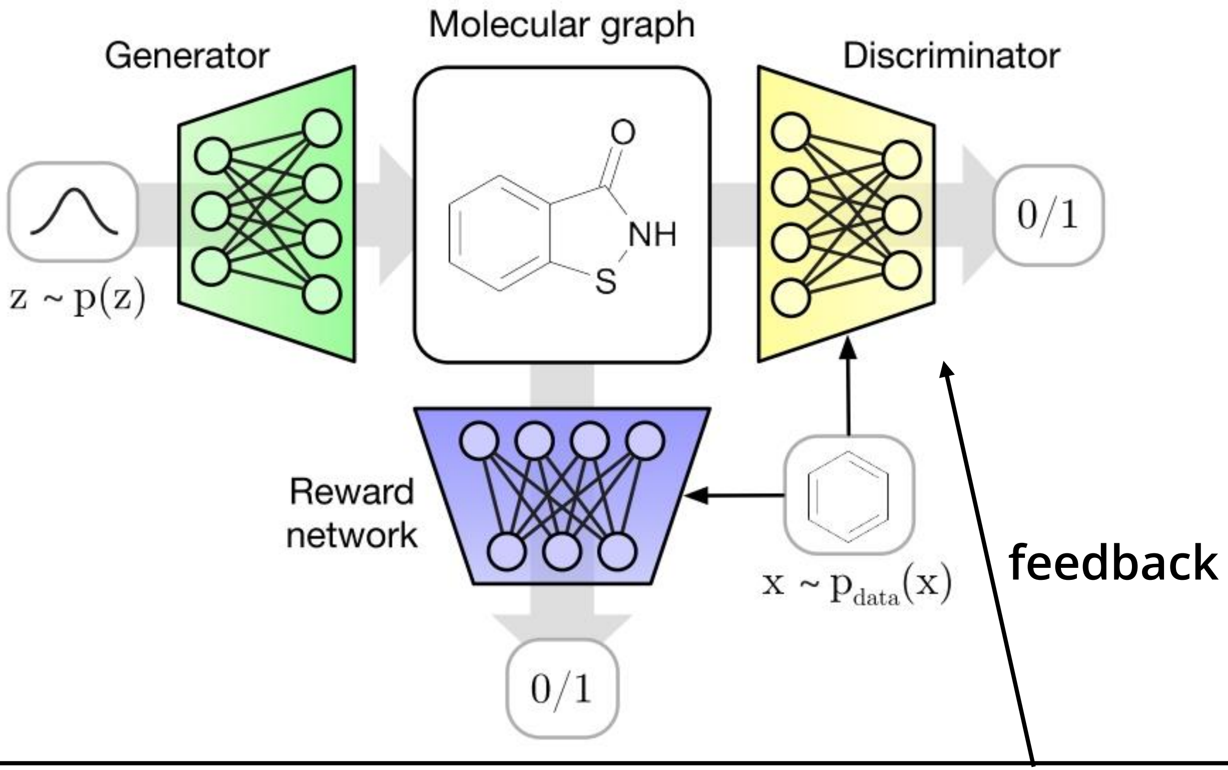
Generator : generate molecules similar to training set

Discriminator : discriminate generated molecules from training set

arXiv:1805.11973

Reward network

- **Scheme of Reinforcement Learning (RL)**



feedback

Reward : Valid, Drug-likeness, Synthesizability, Solubility

+

Generated molecules or Training set ?

arXiv:1805.11973

Performance of MolGAN

Results

Objective	Algorithm	Valid (%)	Unique (%)	Druglikeness	Synthesizability	Solubility
Druglikeness	ORGAN	88.2	69.4	0.52		
	Naive RL	97.1	97.1	0.57		
	MolGAN	**99.9**	2.0	**0.61**		
Synthesizability	ORGAN	96.5	45.9		0.83	
	Naive RL	97.7	13.6		0.83	
	MolGAN	**99.4**	2.1		**0.95**	
Solubility	ORGAN	94.7	54.3			0.55
	Naive RL	92.7	100.0			0.78
	MolGAN	**99.8**	2.3			**0.89**
All	ORGAN	96.1	97.2	0.52	0.71	0.53
	MolGAN	**97.4**	2.4	0.47	**0.84**	**0.65**

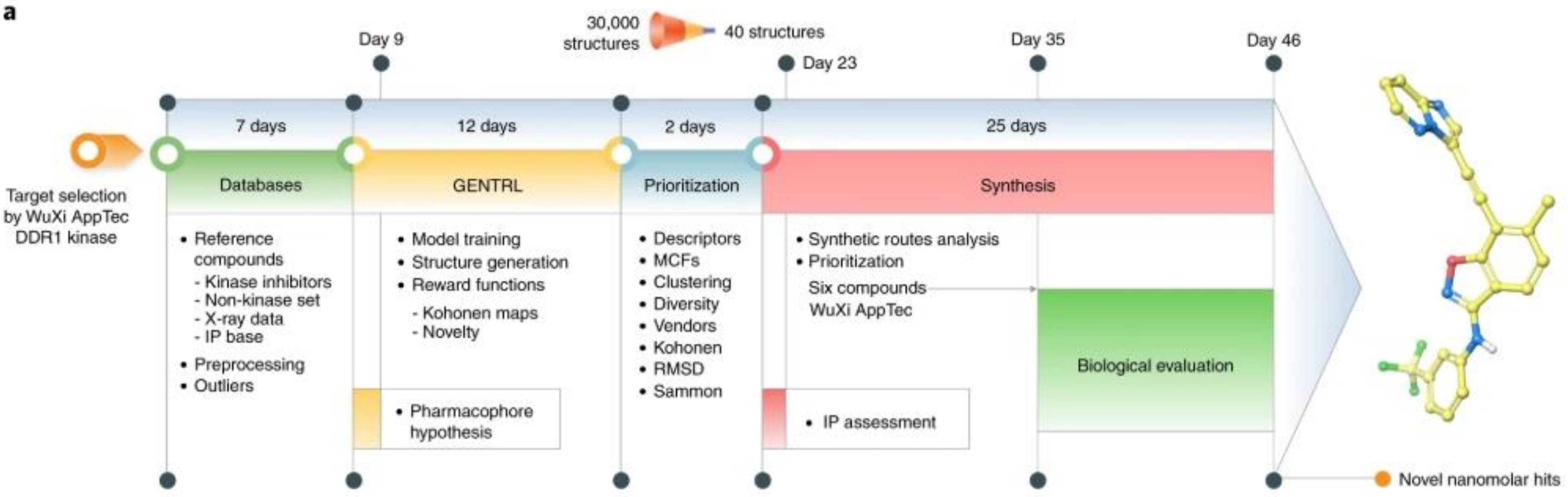
ORGAN (SMILES instead of graph), Naïve RL (without GAN)

- MolGAN beats other models in terms of optimizing property.
- Unique score of generated molecules was very low.

arXiv:1805.11973

GENTRL

Identification of DDR1 kinase inhibitor by GENTRL

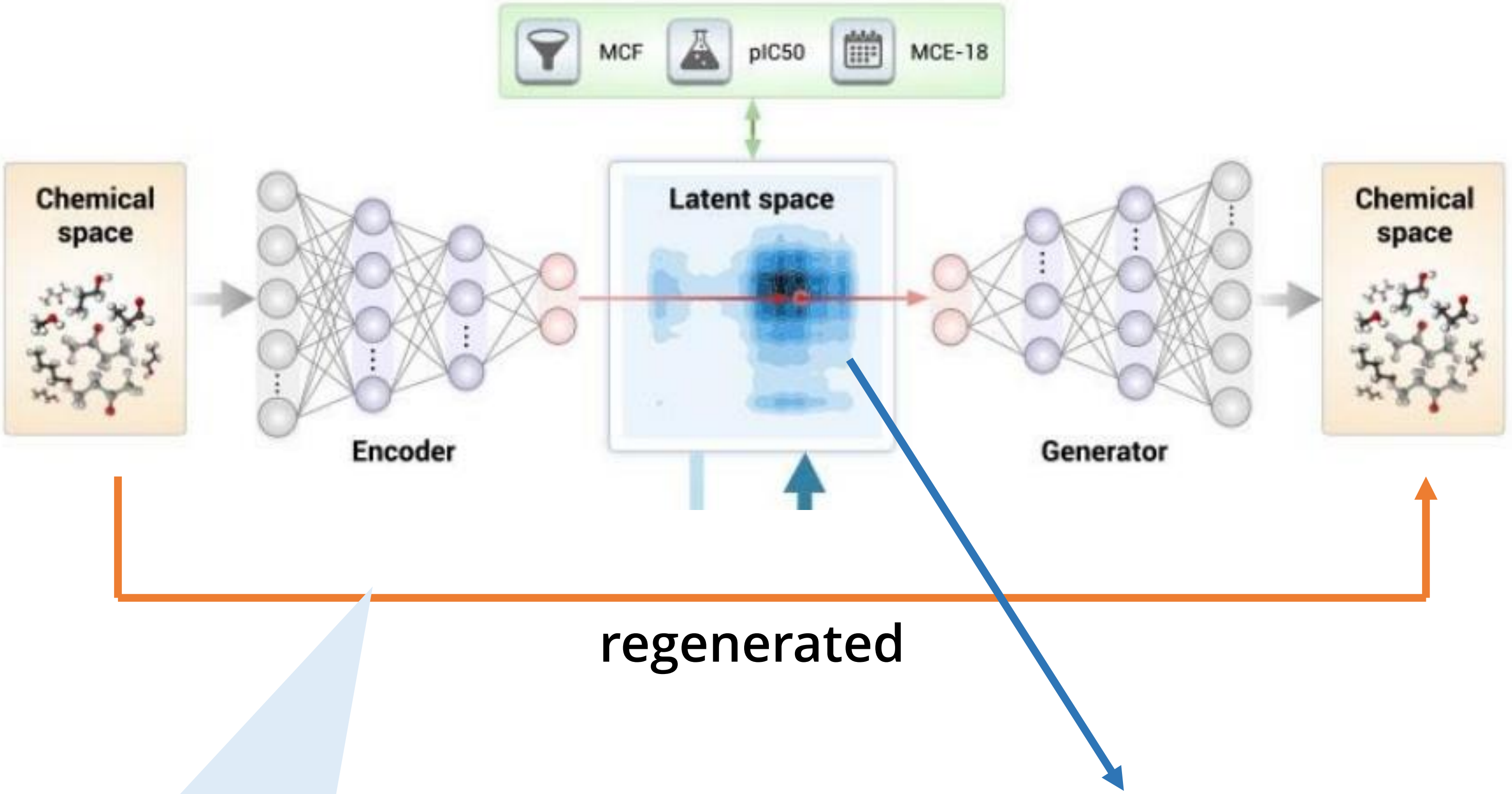


- Day 19 … 30000 molecules were generated by GENTRL.
- Day 23 … 6 molecules were selected by prioritization.
- Day 35 … Synthesis was completed.
- Day 46 … Activities of synthesized molecules were confirmed in cell-based assay.

A. Zhavoronkov *et al.*, *Nat. Biotechnol.*, **2019**, *37*, 1038-1040.

GENTRL

- Creation of chemical space



regenerated

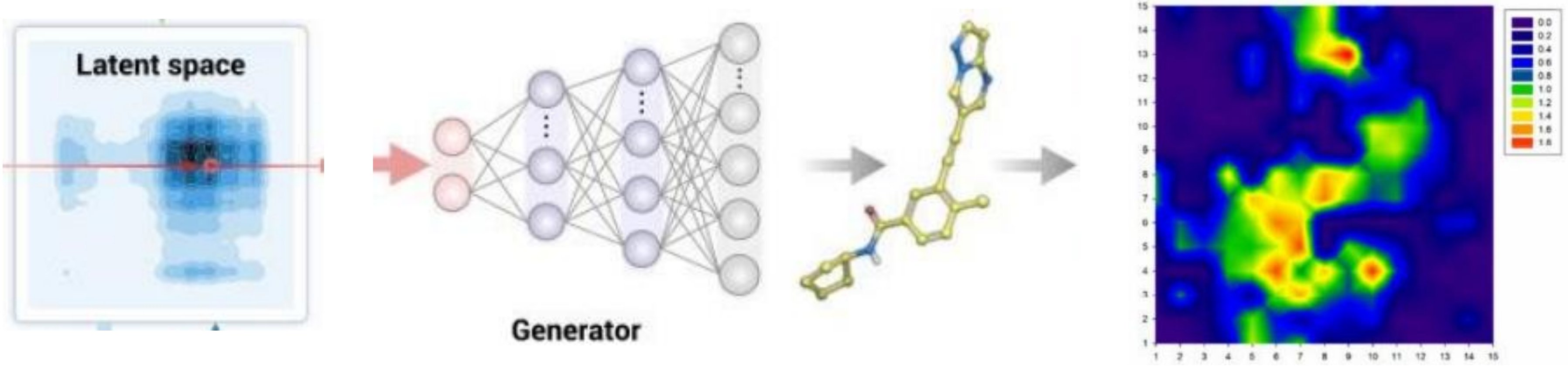
reduction of information

Feature vector

A. Zhavoronkov *et al.*, *Nat. Biotechnol.*, **2019**, *37*, 1038-1040.

GENTRL

- **Molecular generation by Reinforcement Leaning**



SOM
(Self Organizing Map)

Reinforcement Leaning

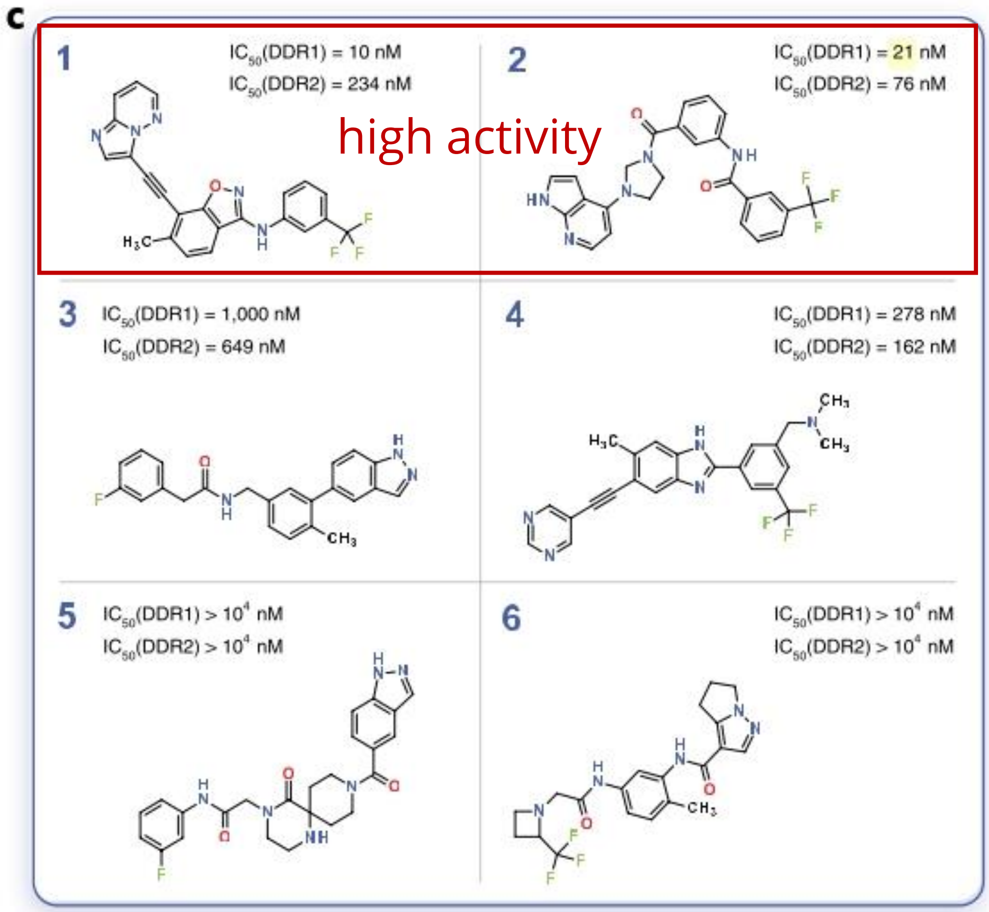
Agent : generator
State : generated molecules
Reward : novelty, kinase inhibition activity, DDR1 inhibition activity

SOM : predict properties of molecules

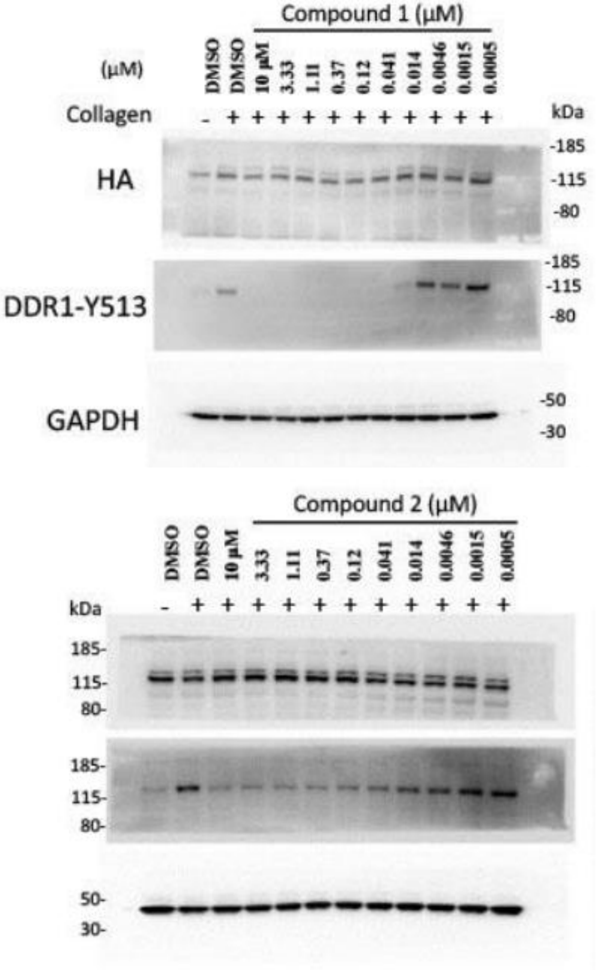
A. Zhavoronkov *et al*., *Nat. Biotechnol*., **2019**, *37*, 1038-1040.

GENTRL

Selected molecules and inhibitory activity in vitro and vivo



high activity



A. Zhavoronkov *et al*., *Nat. Biotechnol*., **2019**, *37*, 1038-1040.

Summary

Chemical Space

- Chemical space is vast (~10^{60}) compared to compound library size (~10^{6}, 10^{8}).
- Generative model can generate 10^{3} ~ 10^{5} drug-like compounds.
- Generative model can control properties of generated molecules by RL.
- The role of generative model is to capture the underlying rules of a data distribution.
- Generative model only reconstruct the training data set.

Molecular representation

- SMILES is not robust to small changes or mistakes.
- By using graph representations , generative model don't need to learn complex syntax, but this method is not perfect.
- There is still a need for research on the optimal molecular representation.
  - Junction Tree (arXiv:1802.04364)
  - 3D (arXiv:1810.11347)

Summary

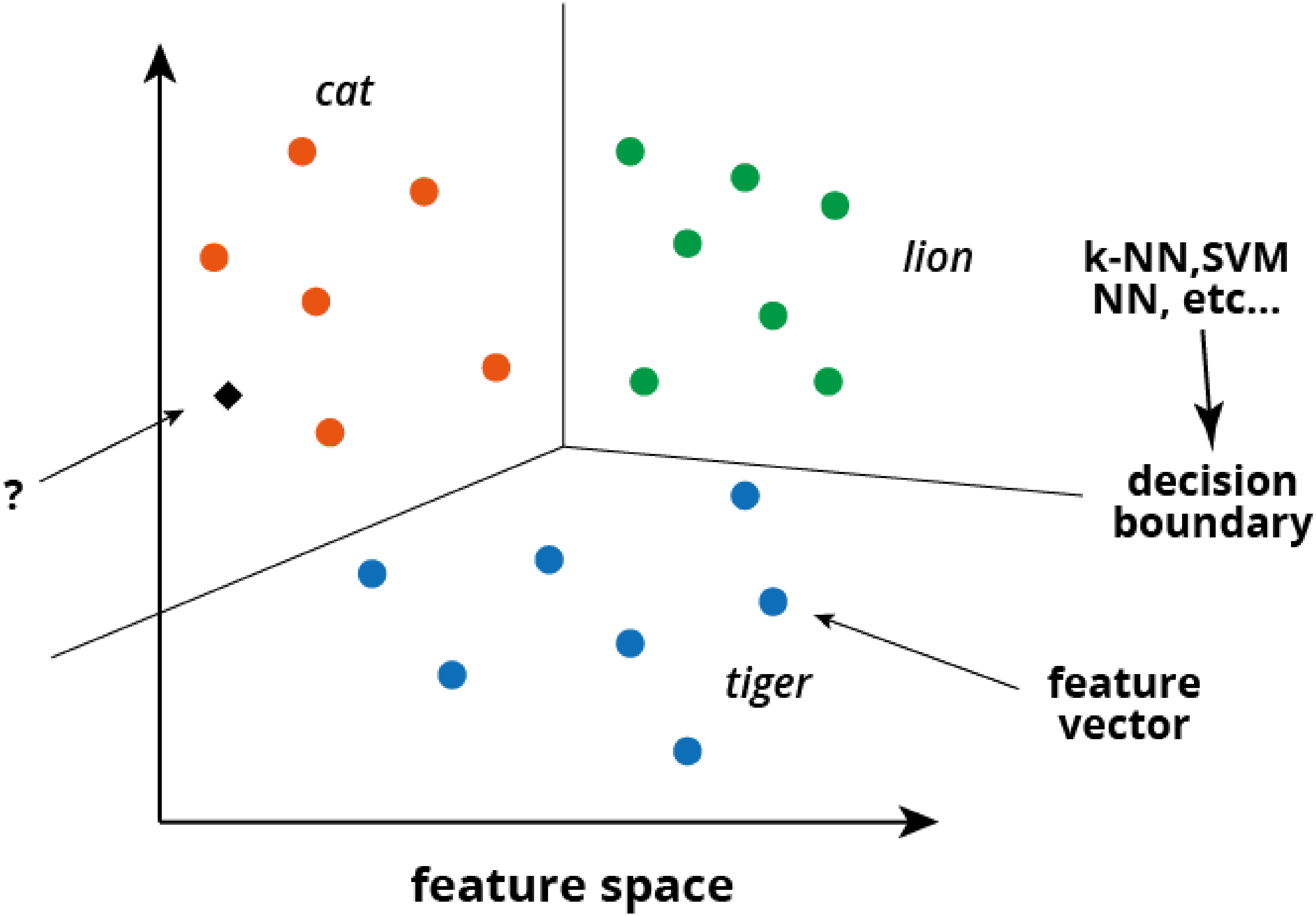
Evaluation of Model

- The performance of each generative model is evaluated by different methods.
  - Number of generated molecules
  - Distribution on 2D map.
  - Properties of generated molecules.
  - Experimental activity.
- Evaluation method of model is needed.
- Several benchmarks are being developed. (J. Chem. Inf. Model, 2019, 59, 1096)

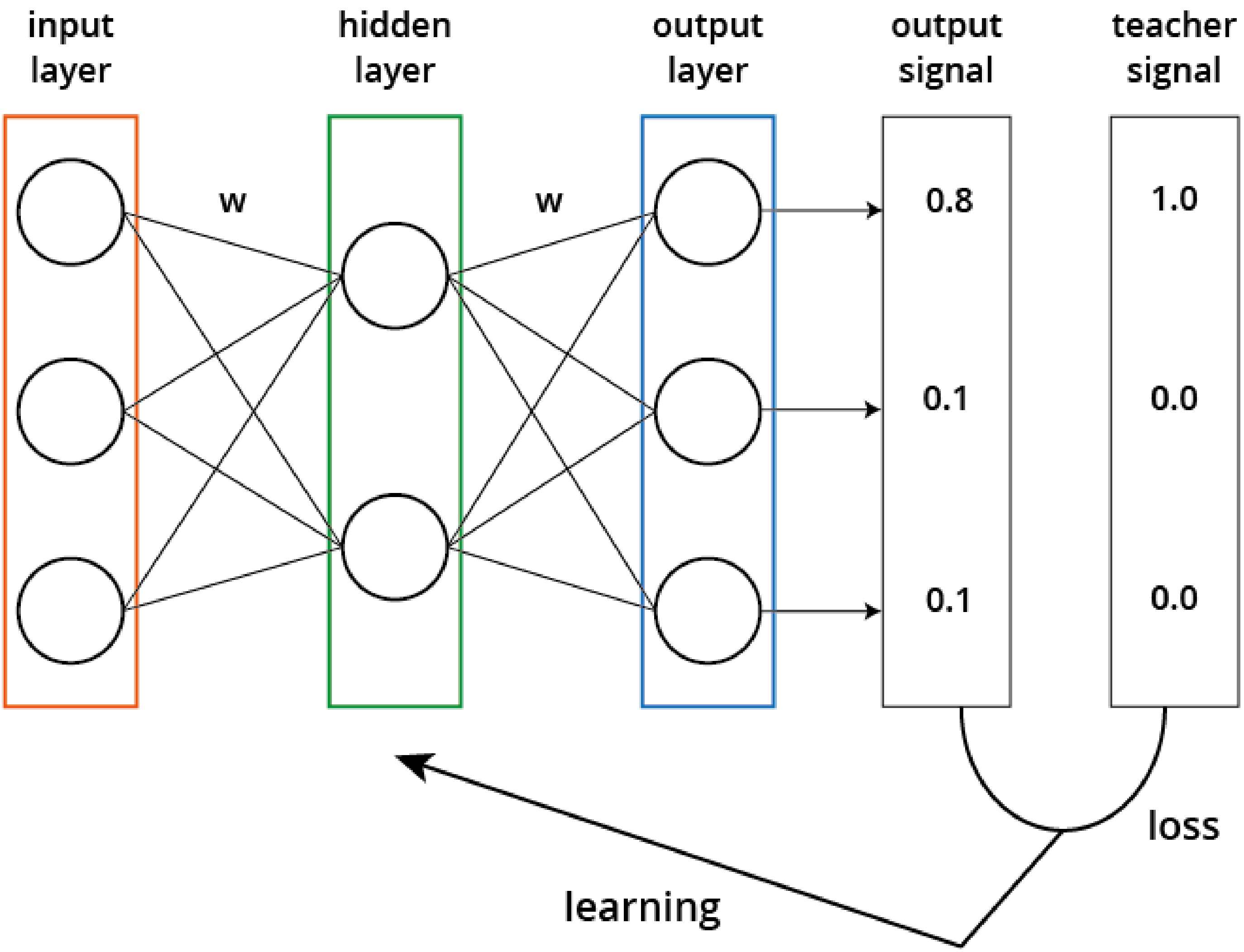
Application in drug discovery

- The generated molecules must be reduced to the number that can be synthesized.
- The generated molecules are necessarily synthesizable.
- SAS (synthetic accessibility score) may prevent generation of molecular diversity.
- Generative model may prove valuable in combination with retrosynthesis AI or virtual screening AI.

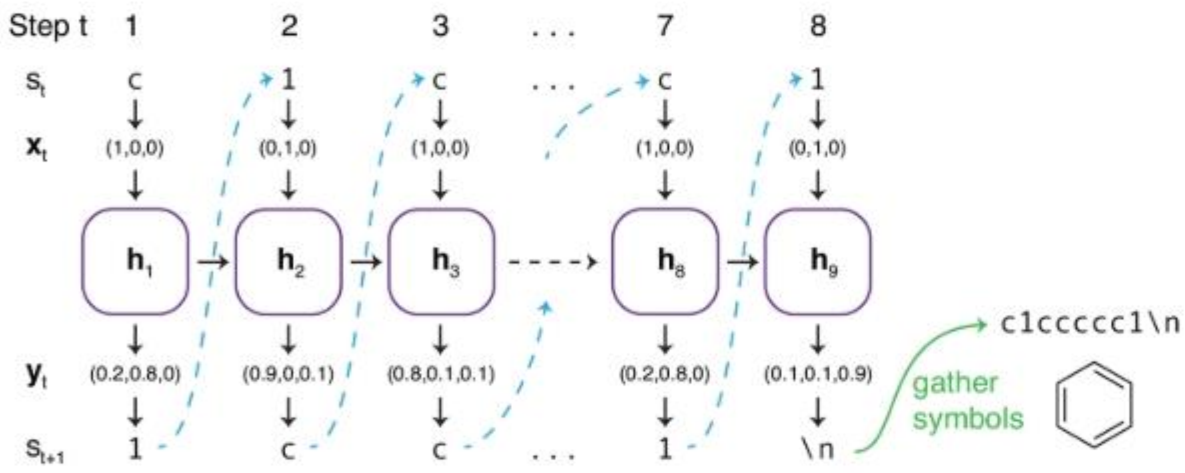
Feature space



NN


input layer
hidden layer
output layer
output signal
teacher signal
w
w
0.8
0.1
0.1
1.0
0.0
0.0
loss
learning

RNN



M. H. S. Segler *et al*., *ACS Cent. Sci*., **2018**, *4*, 120-131.

AZ filter

Class 1 : bland structures

- Fewer than 4 carbon atoms etc.

Class 2 : reactive structures

- Anhydride etc.

Class 3 : frequent hitters

- Nitrophenols etc.

Class 4 : dye-like structures

Class 5 : unlike drug candidates or unsuitable fragments

Class 6 : difficult series or natural compounds

Class 7 : general ugly halogenated structures

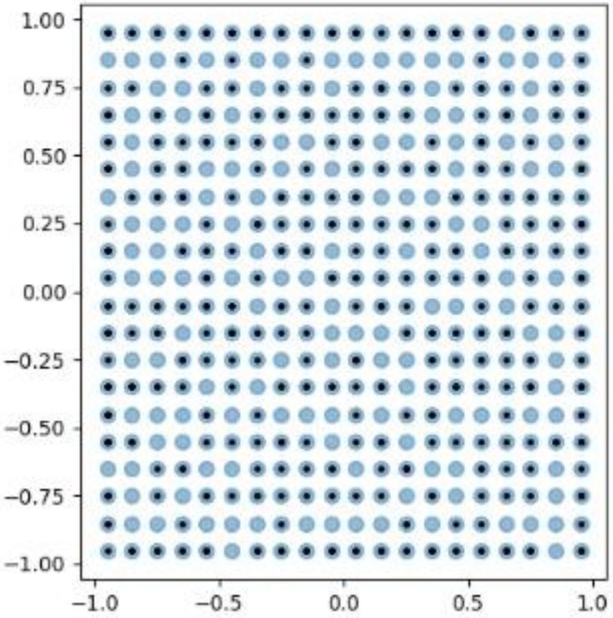
Class 8 : general ugly oxygen

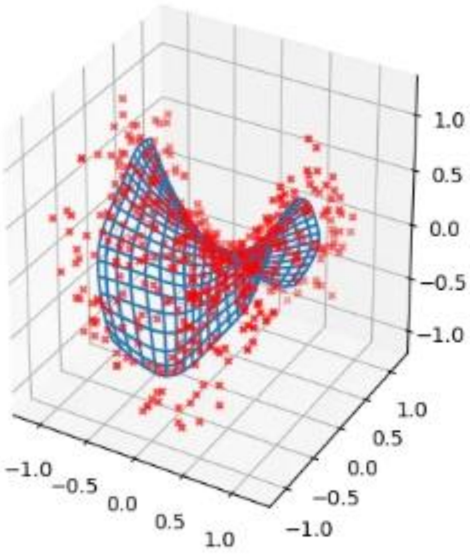
Class 9 : general ugly nitrogen

Class 10 : general ugly sulphur

Nat. Rev. Drug Discovery 2013, 12, 948−962

SOM





https://qiita.com/tohru-iwasaki/items/e51864269767ccc07254