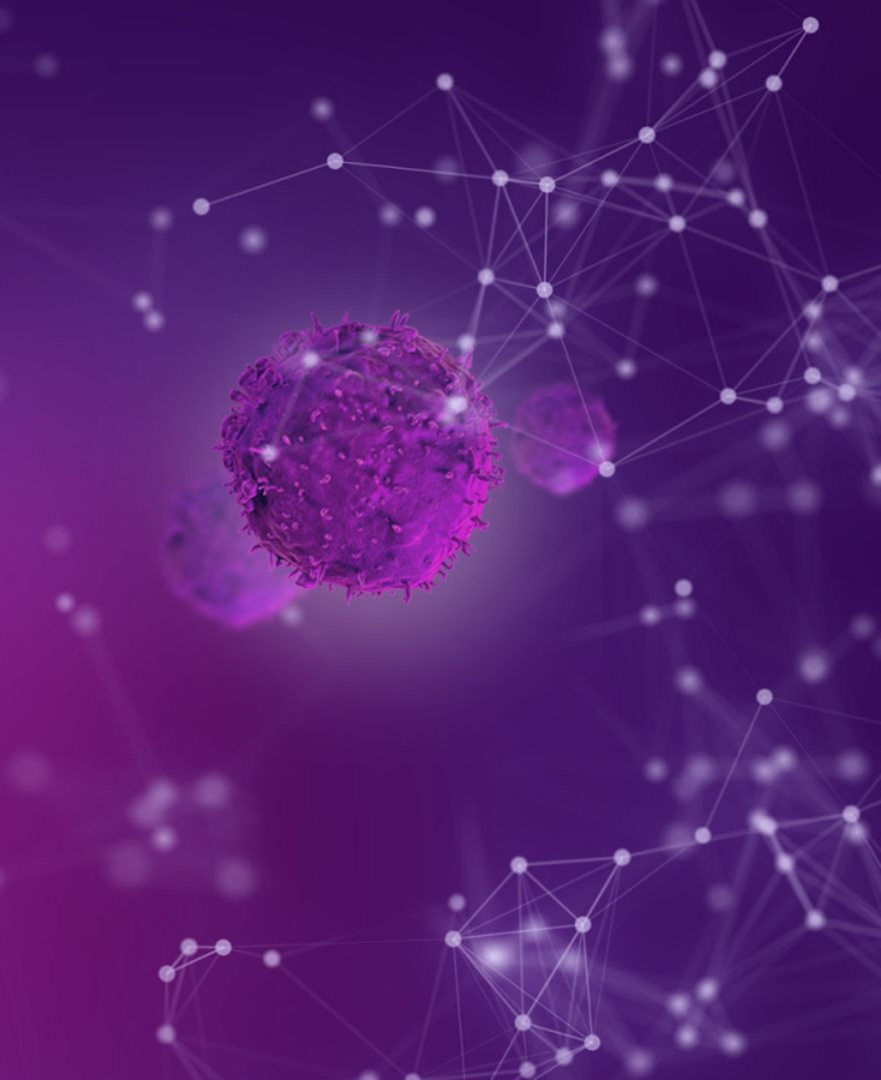
ardigen

Artificial Intelligence & Bioinformatics
for Precision Medicine

group of machine
gmum
learning research

25 November 2019



Agenda

1. Goals and Obstacles in Generating Novel Compounds
   a. Chemical perspective
   b. Machine learning perspective
2. Overview of Generative Models in Chemistry
   a. Graph-based
   b. SMILES-based
3. Coding...

ardigen

gmum group of machine learning research

Goals and Obstacles in Generating Novel Compounds

Drug Design Point of View

Goals of Computer-Aided Drug Design

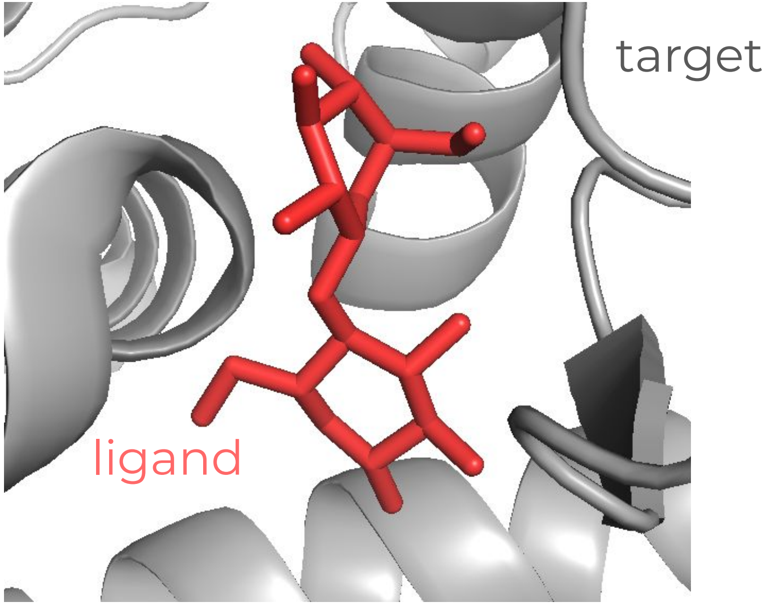
The chemical space of pharmacologically active compounds is estimated to be in the order of 10^{60}.

Find a novel active compound

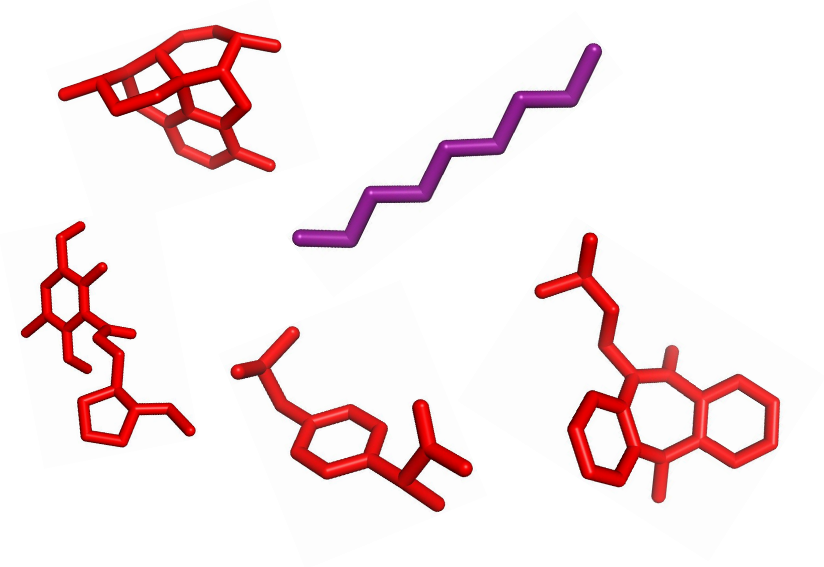
1. *de novo* molecule generation
2. generating from a given core (optimization)

Profile of a Perfect Compound

bioactivity

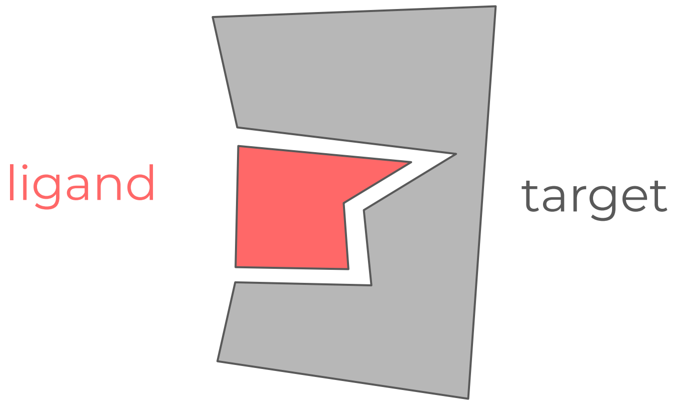


drug-likeness

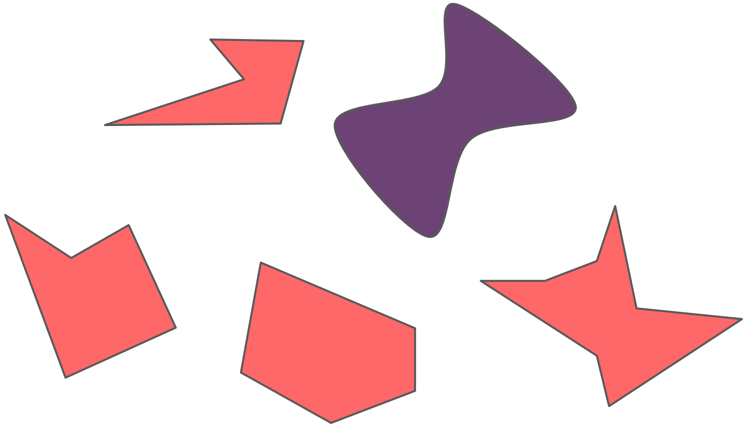


Profile of a Perfect Compound

bioactivity

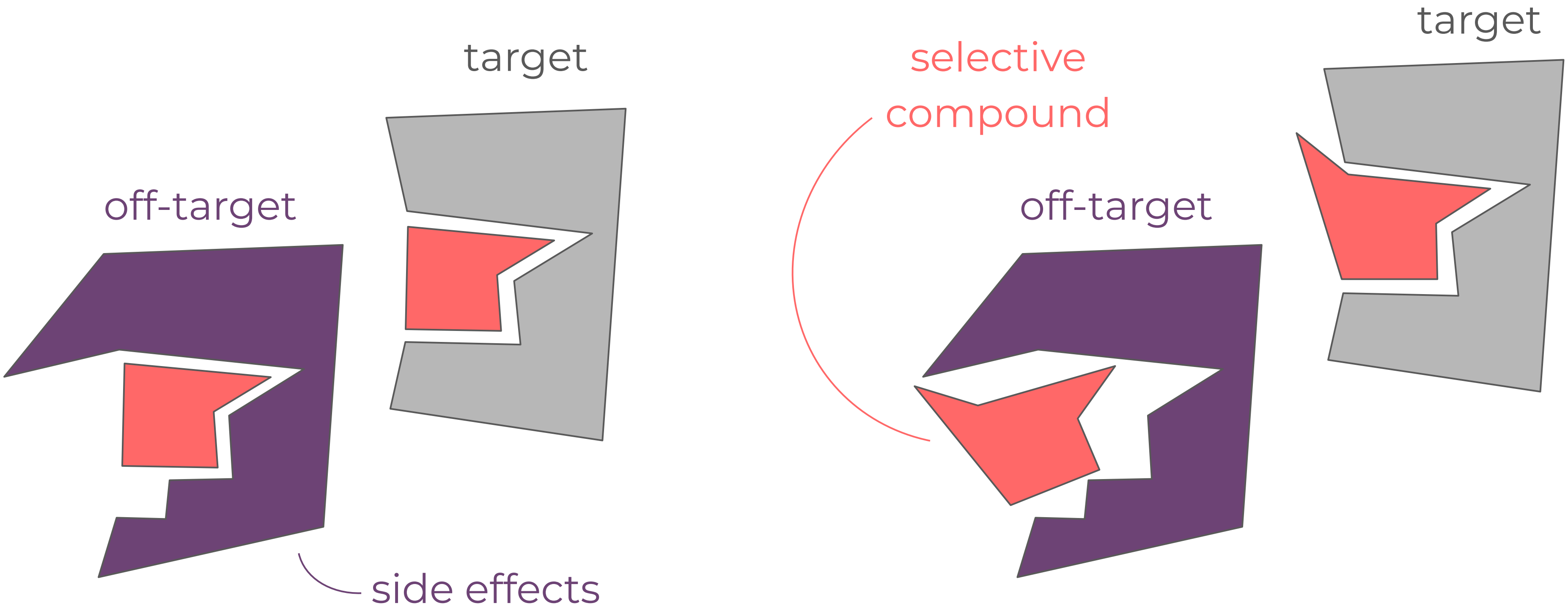


drug-likeness



Issues to Overcome

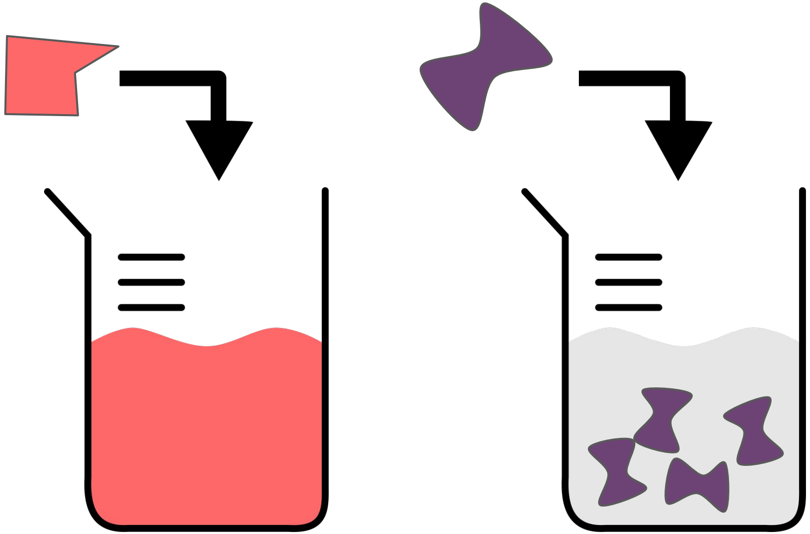
selectivity



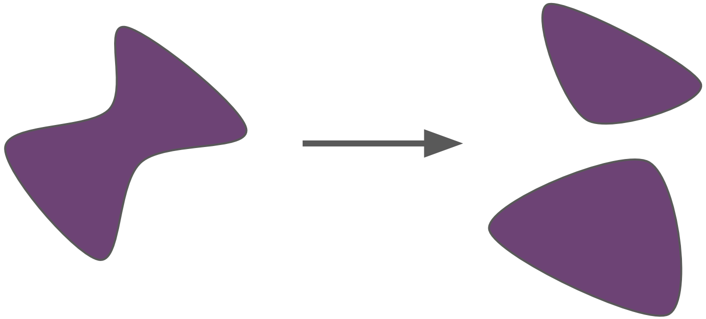
Issues to Overcome

ADME properties

solubility
(**d**istribution)



stability
(**m**etabolism)

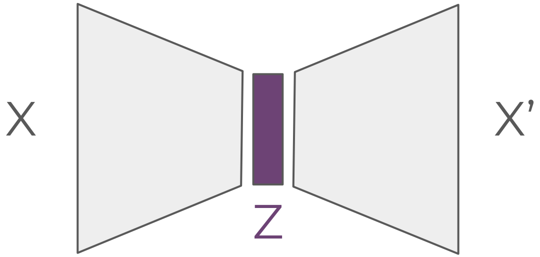


Machine Learning Point of View

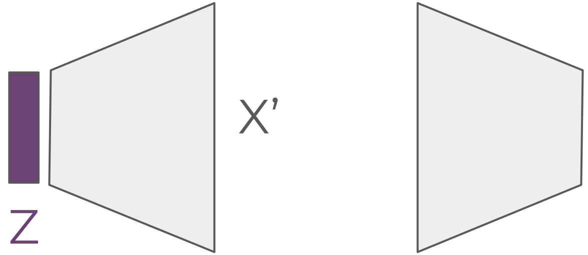
Generative Models 101

Typical deep generative architectures

Autoencoders



Generative Adversarial Networks



Recurrent Neural Networks

$$X'_t \xrightarrow{\text{supervision}} X'_{t+1}$$

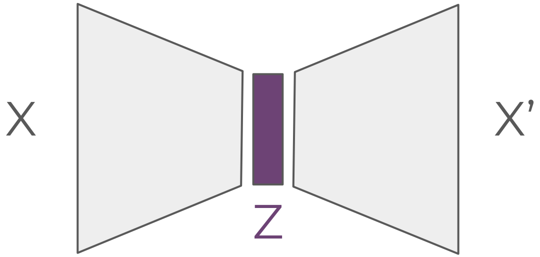
Reinforcement Learning

$$X'_t \xrightarrow[\text{reward}]{\text{action}} X'_{t+1}$$

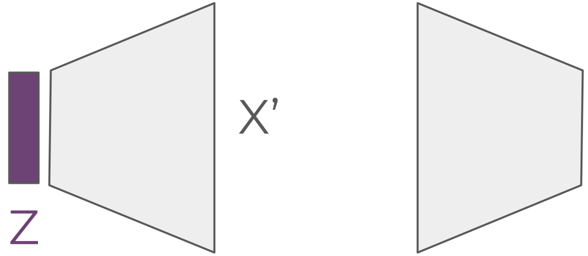
Generative Models 101

Typical deep generative architectures

Autoencoders



Generative Adversarial Networks



Recurrent Neural Networks

$$X'_t \xrightarrow{\text{supervision}} X'_{t+1}$$

Reinforcement Learning

$$X'_t \xrightarrow[\text{reward}]{\text{action}} X'_{t+1}$$

Problems with Chemical Representations

1. Fingerprints

[011001101001000]

Probably easy to generate, but what molecule is that?

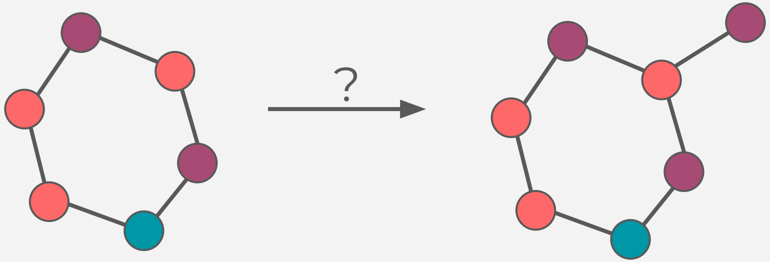
2. SMILES

C1=CC(=C(C=C1CCN)O)O

context-free grammar

3. Graphs

Graphs are intuitive but discrete.



Only some connected substructures make sense in chemistry.

Constraints

optimization vs filtering

1. Optimize properties during generation
   a. Bayesian optimization (AE, GAN)
   b. Conditional generation (GAN, AE, RNN)
   c. Reward-driven generation (RF)

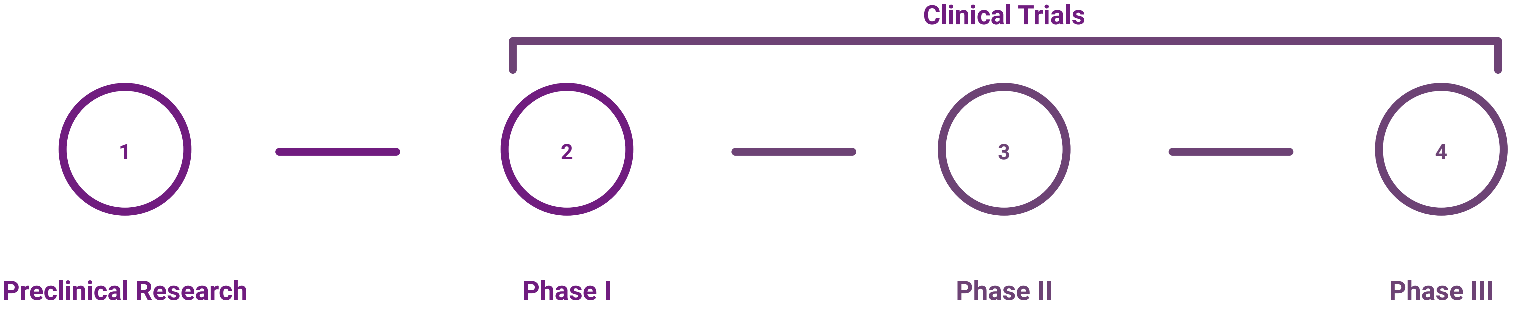
2. Filter out compounds after generation
   a. Costly simulations, e.g. docking
   b. Expertise of medchems

Evaluation of Generated Compounds

validity = (# valid compounds) / (# generated)

uniqueness = (# unique, valid compounds) / (# valid)

novelty = (# unique compounds not in the training set) / (# unique)



ardigen

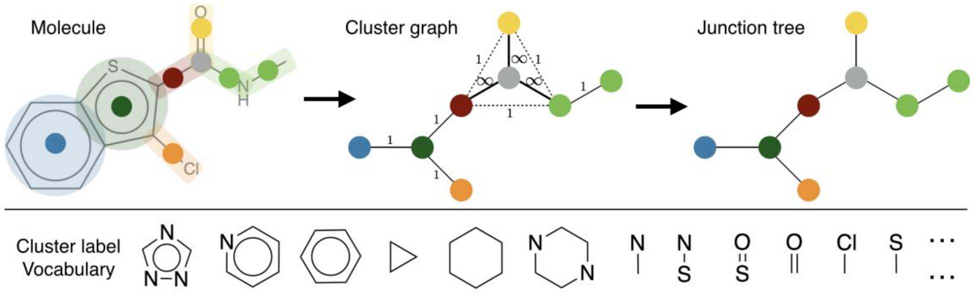
group of machine
gmum
learning research

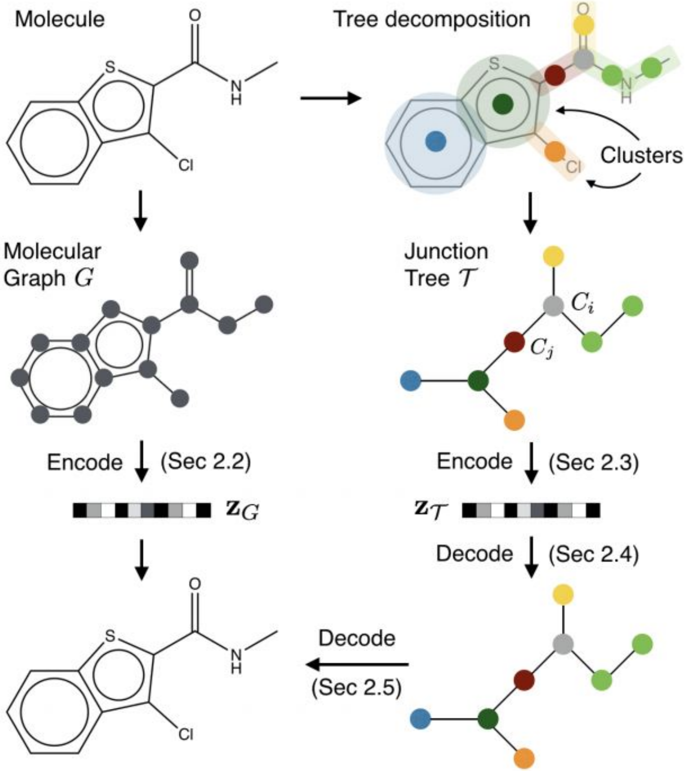
Overview of Generative Models in Chemistry

Graph-Based Models

JT-VAE

- VAE-based model
- constructs a continuous latent space of molecules
- encodes and decodes molecules using junction trees

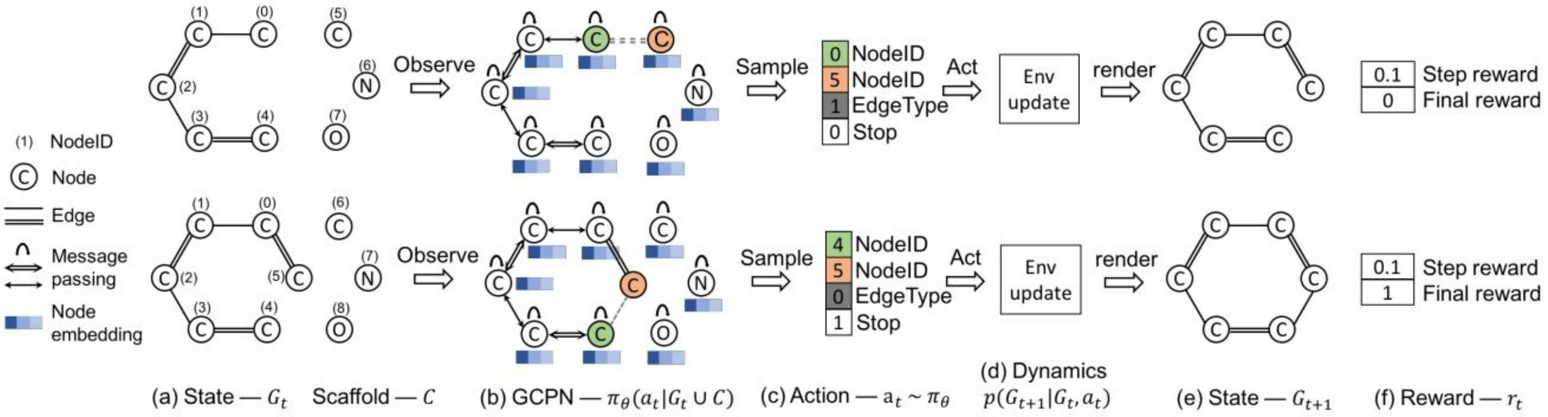




Jin, W., Barzilay, R., & Jaakkola, T. (2018). Junction tree variational autoencoder for molecular graph generation.

GCPN

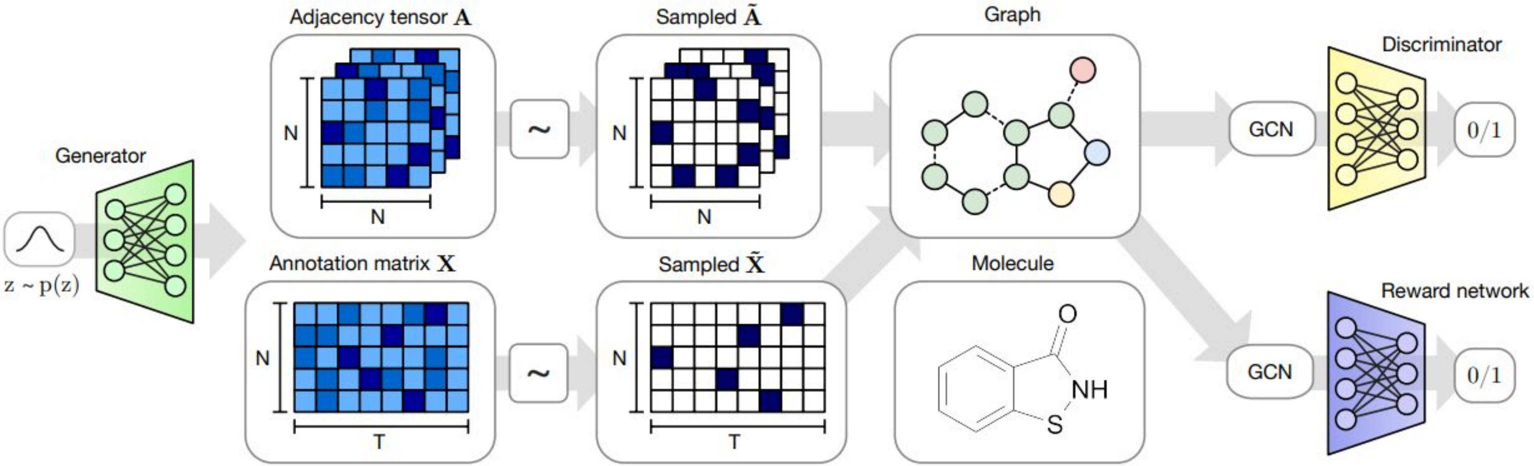
- RL-based model
- operates on graphs by adding edges bond by bond
- GAN holds the drug-like distribution



You, J., Liu, B., Ying, Z., Pande, V., & Leskovec, J. (2018). Graph convolutional policy network for goal-directed molecular graph generation.

MolGAN

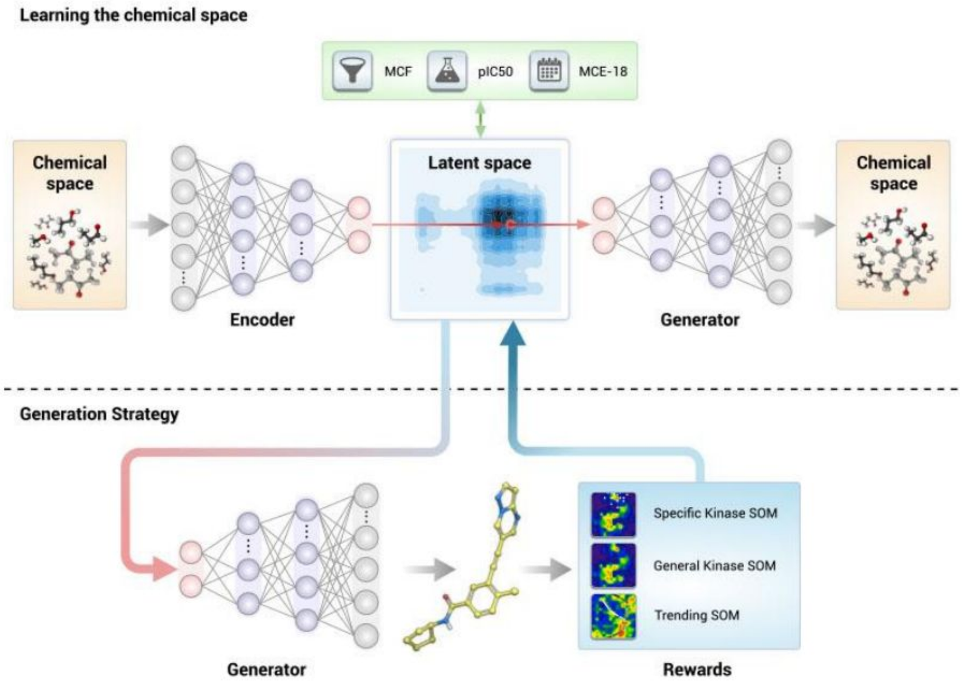
- GAN-based model
- generates discrete adjacency matrices
- discrete reparametrization - Gumbel softmax


Adjacency tensor $\mathbf{A}$
Sampled $\tilde{\mathbf{A}}$
Graph
Discriminator
N
N
~
N
N
GCN
0/1
Generator
$z \sim p(z)$
Annotation matrix $\mathbf{X}$
Sampled $\tilde{\mathbf{X}}$
Molecule
Reward network
N
T
~
N
T
O
NH
S
GCN
0/1

De Cao, N., & Kipf, T. (2018). MolGAN: An implicit generative model for small molecular graphs.

GENTRL

- VAE-based with RL component
- inhibitors of DDR1 discovered in 21 days
- designed, synthesized, and validated in less than 2 months
- SOMs were used to calculate rewards (Kohonen 1997)

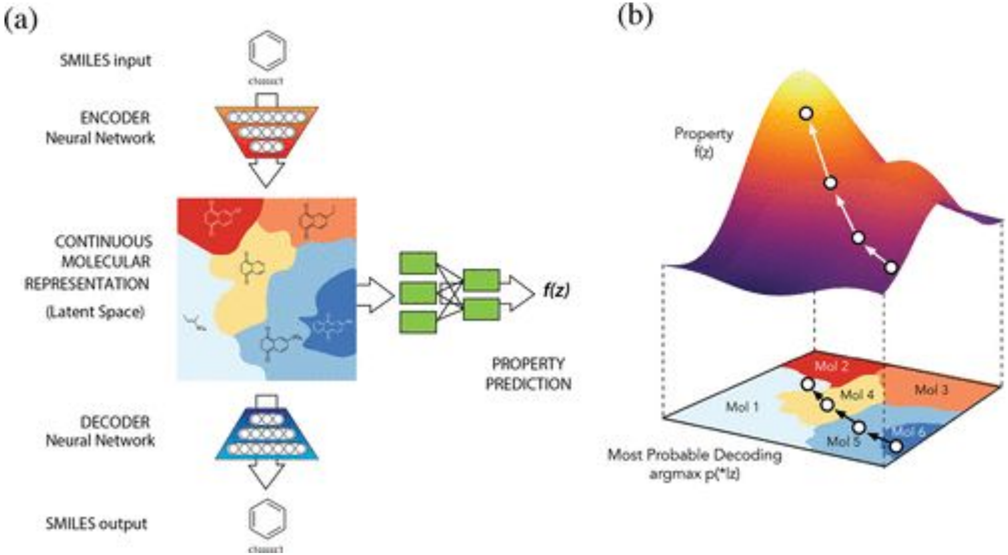


Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., ... & Volkov, Y. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors.

SMILES-Based Models

Character VAE

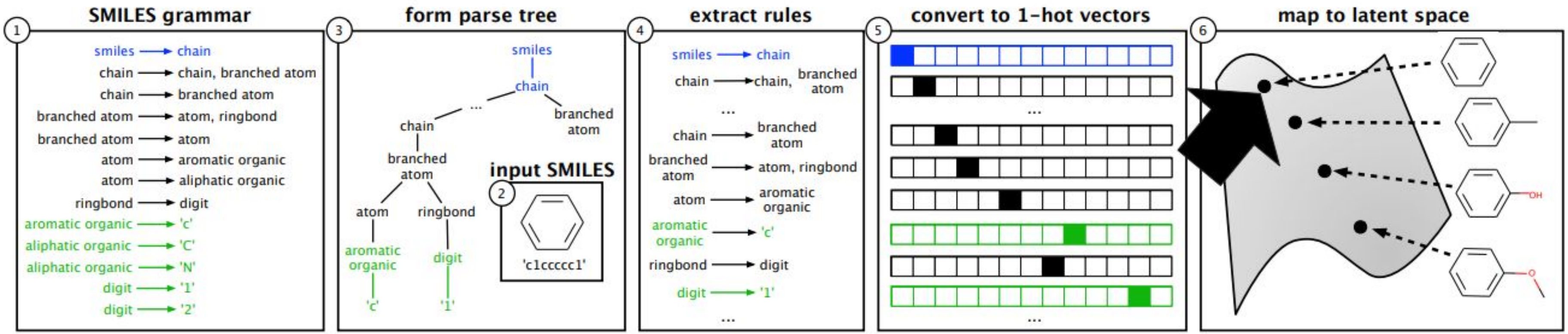
- VAE-based model



Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., ... & Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules.

Grammar VAE

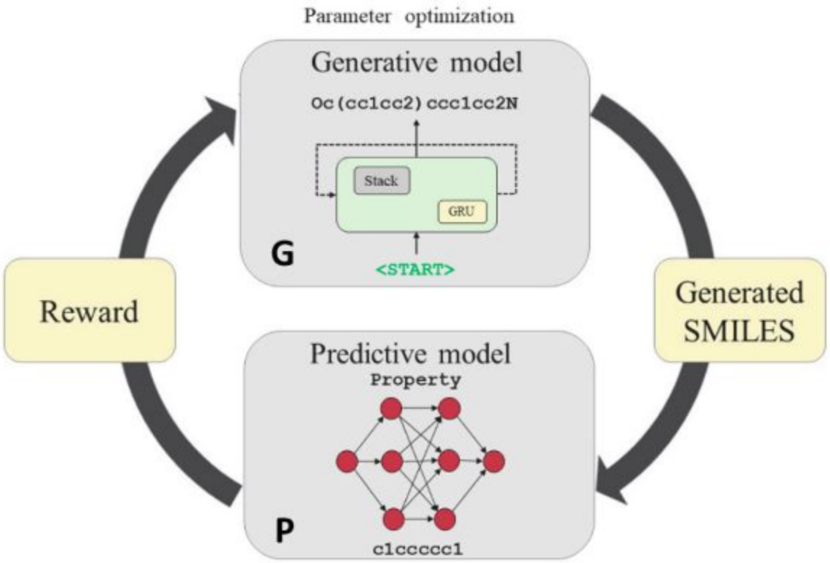
- VAE-based model
- encodes and decodes production rules of the SMILES grammar



Kusner, M. J., Paige, B., & Hernández-Lobato, J. M. (2017, August). Grammar variational autoencoder.

ReLeaSE

- RNN-based method
- the architecture is augmented with a memory stack
- uses policy gradient to optimize properties

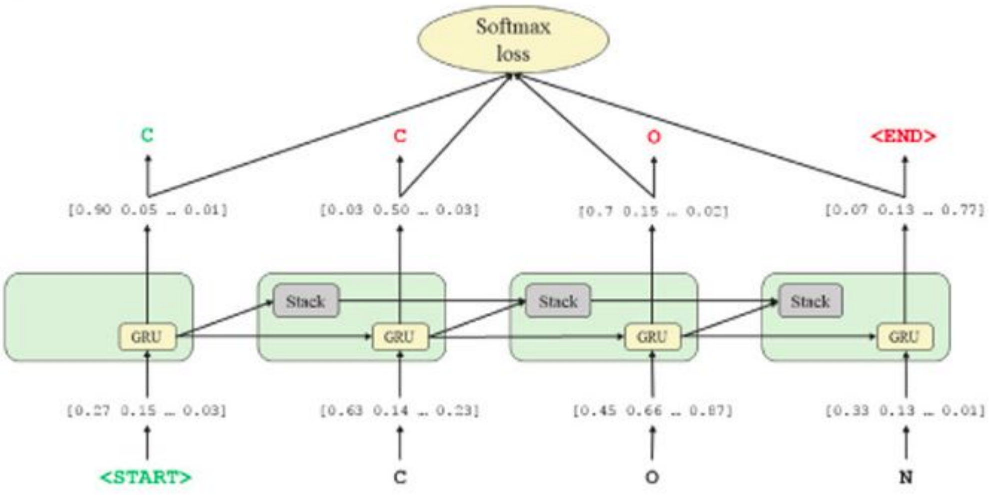


Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design.

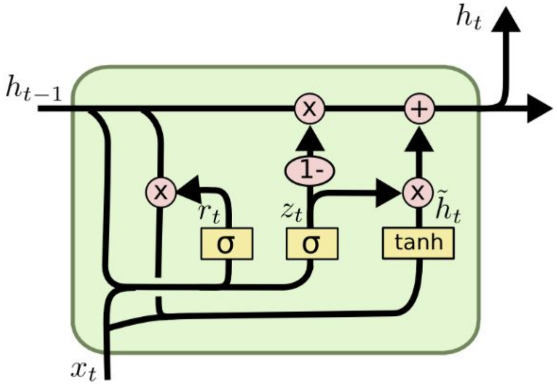
ReLeaSE

Stack-RNN

A stack allows to learn long-range interdependencies. In practice, it is implemented as additional gates realizing PUSH and POP operations.



Gated Recurrent Unit



Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design.

ReLeaSE

Policy Gradient

How to calculate gradients from rewards?

$$\mathrm{f(x)}\ \nabla_\theta \log \mathrm{f(x)} \ = \ \mathrm{f(x)}\ \frac{\nabla_\theta \mathrm{f(x)}}{\mathrm{f(x)}} = \nabla_\theta \mathrm{f(x)}$$

$$\begin{aligned}\nabla_\theta J(\theta) &= \int \nabla_\theta \pi_\theta(\tau) r(\tau) d\tau = \int \pi_\theta(\tau) \nabla_\theta \log \pi_\theta(\tau) r(\tau) d\tau \\ &= E_{\tau \sim \pi_\theta(\tau)}[\nabla_\theta \log \pi_\theta(\tau) r(\tau)]\end{aligned}$$

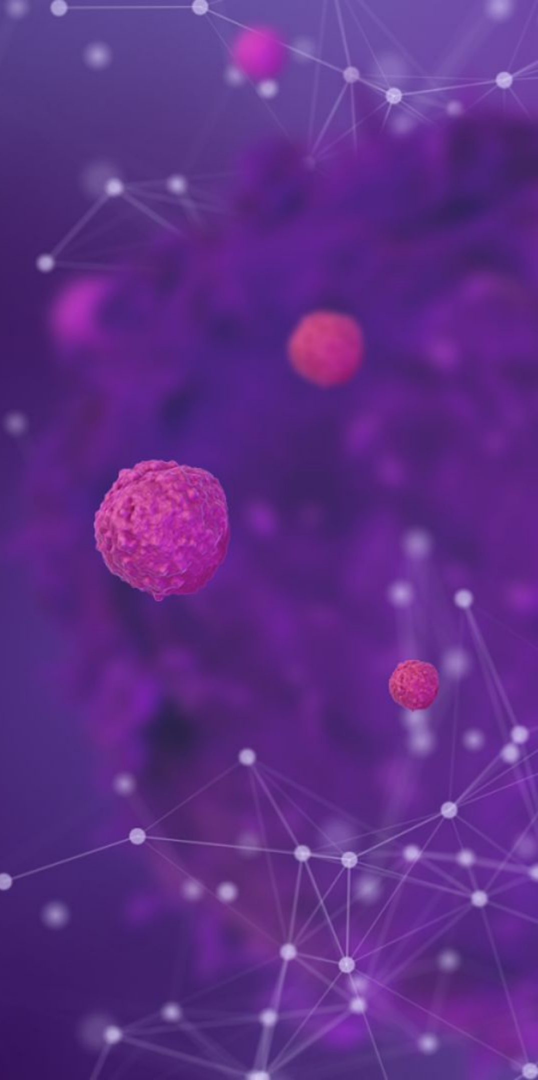
Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design.

ardigen

group of machine
gmum
learning research

Hands-On Drug Discovery Project

tinyurl.com/mlinpl-gen



Contact us

group of machine gmum learning research

ardigen

Artificial Intelligence & Bioinformatics
for Precision Medicine

Our site: gmum.net

Our research topics:

- generative models,
- theoretical understanding of deep learning and optimization,
- natural language processing,
- drug design and cheminformatics,
- unsupervised learning and clustering.

Our site: ardigen.com

Our projects:

- medical imaging
- computer-aided drug design
- single-cell analysis
- more...