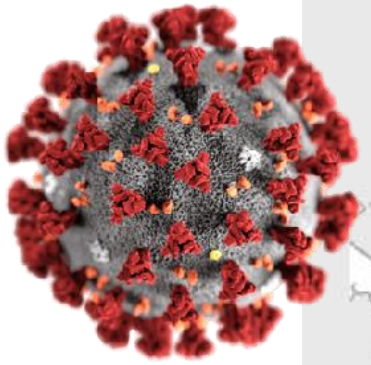
Multi-objective Drug Molecule Generation via MARS

Lei Li

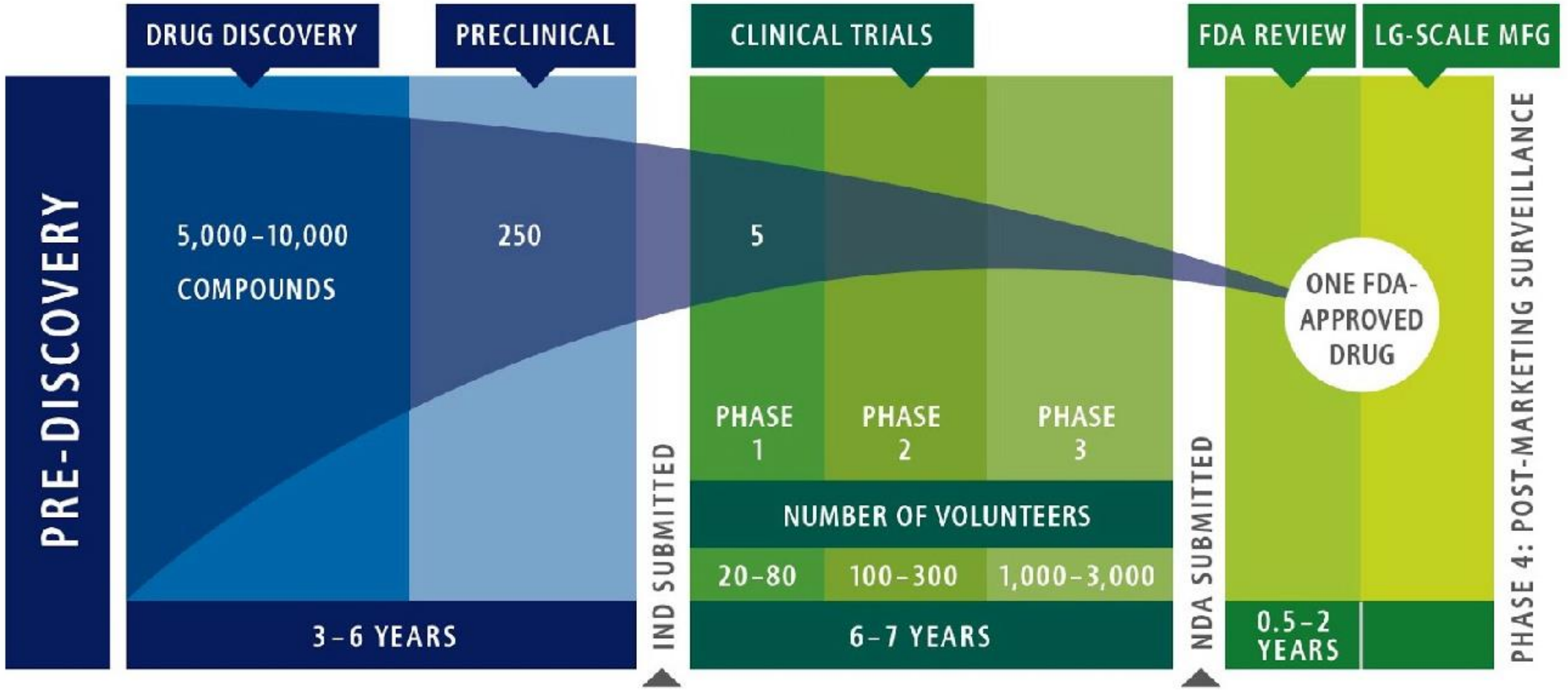
Language Technologies Institute

Carnegie Mellon University
School of Computer Science



- In 2020, Covid-19 pandemic rapidly spread over the world.
- More than 620 million confirmed cases including over 6.5 million deaths.
- Need for acceleration in pharmaceutical research.

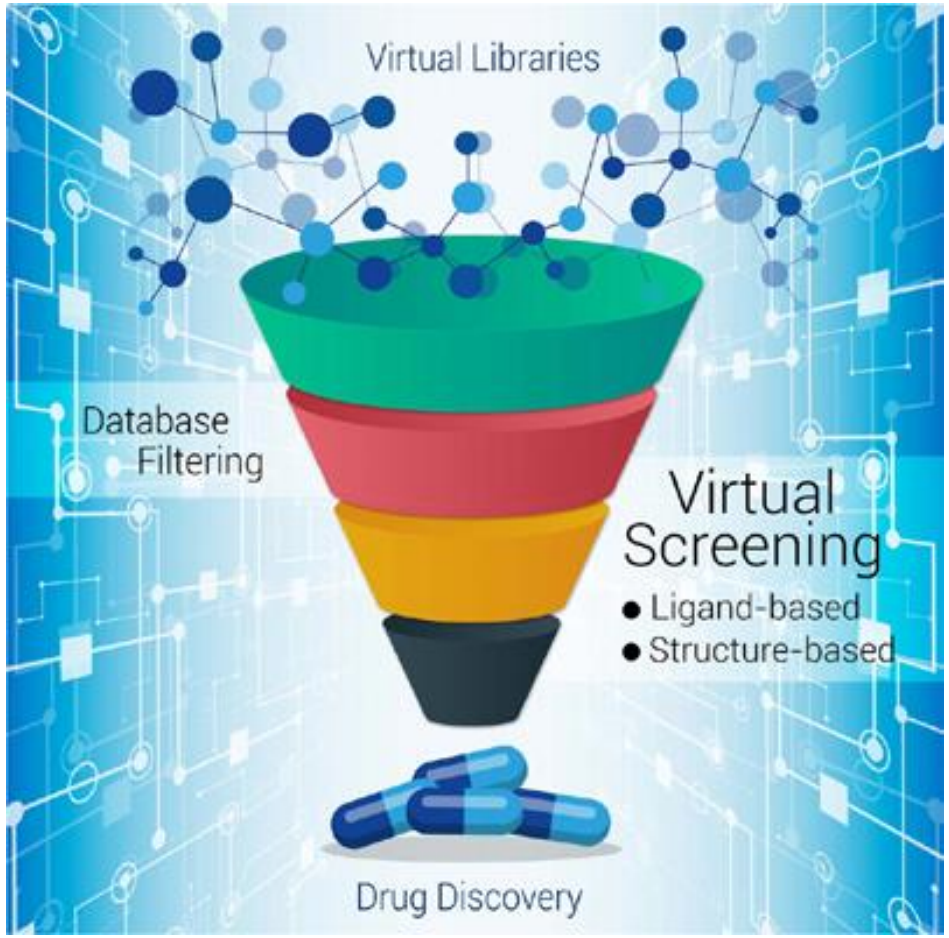
Drug Discovery and Development



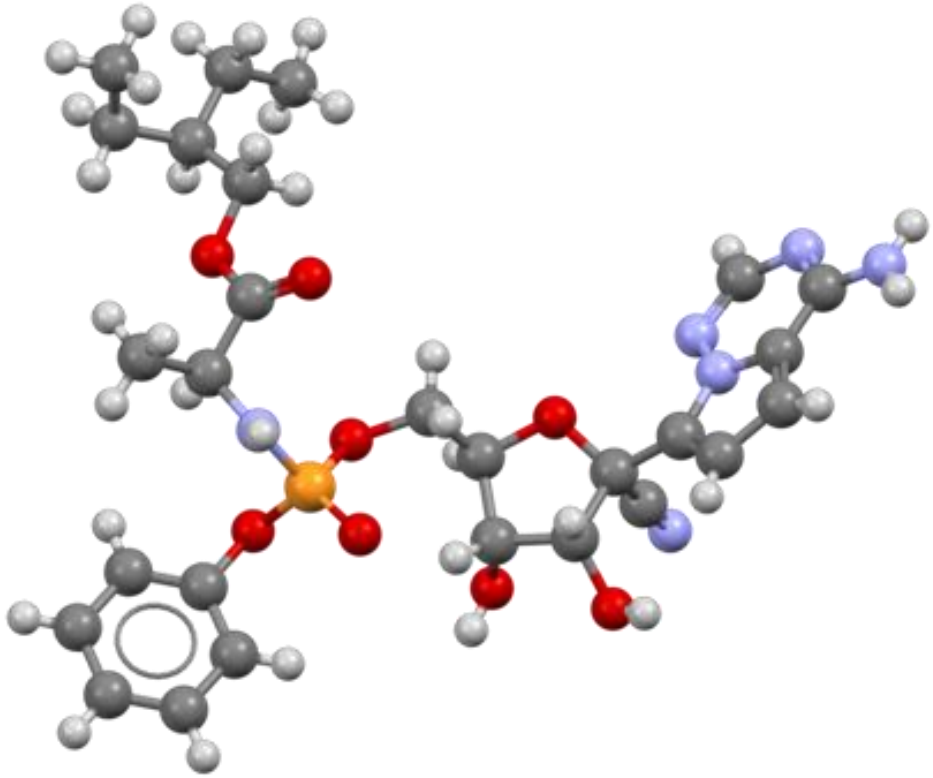
- Drug discovery and development is a long and risky process.
- The successful launch of a new drug cost $1.3 billion on average and more than 10 years.

In silico Drug Discovery

- Using computers to predict drug activities.
- Faster and cheaper.
- But drug candidates are either existing molecules or manually designed ones.
- Larger virtual library?



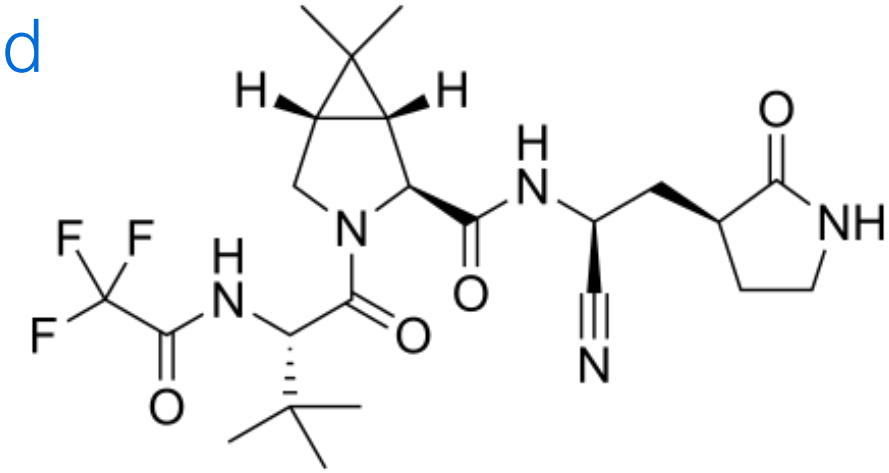
Discovering Treatment for COVID-19



Remdesivir

Originally for hepatitis C

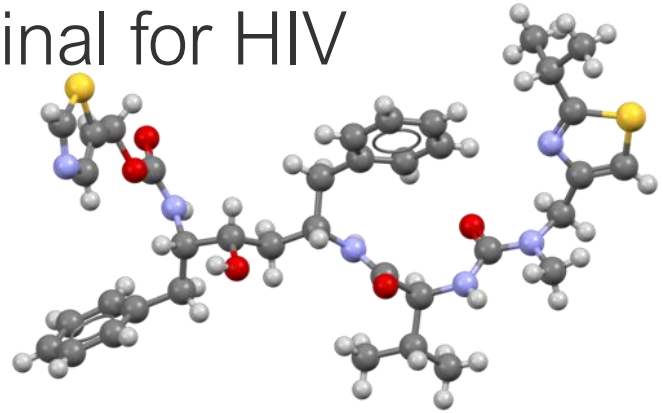
Paxlovid



Nirmatrelvir

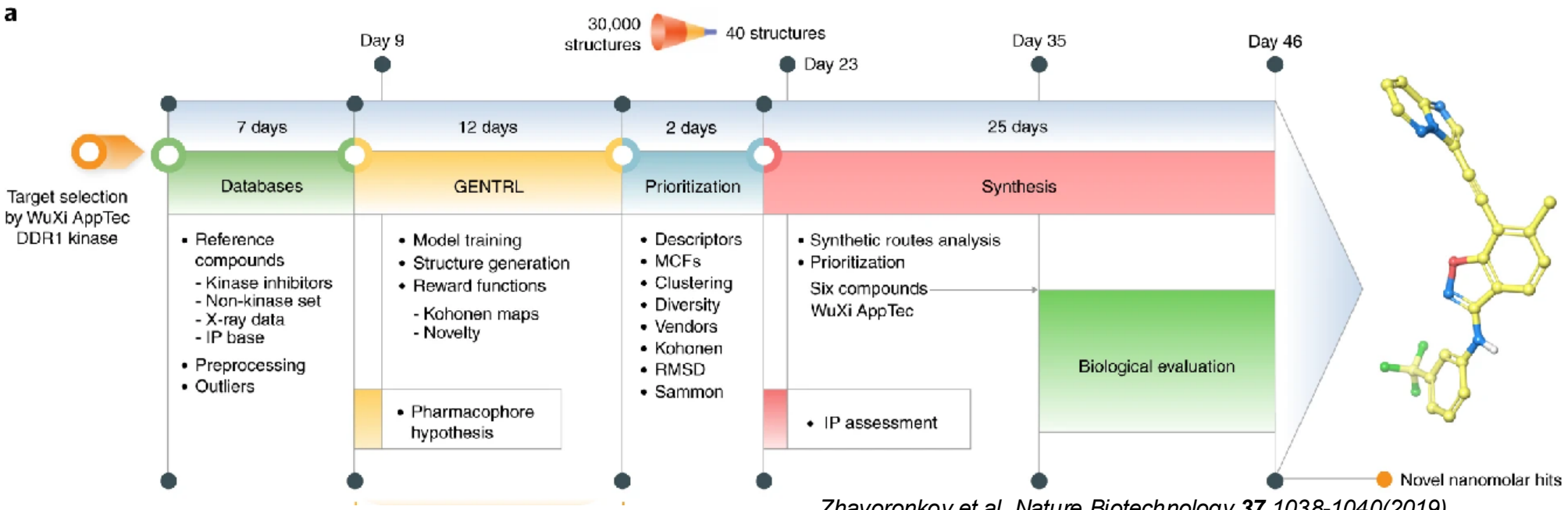
by modifying prior protease inhibitor

Ritonavir: original for HIV



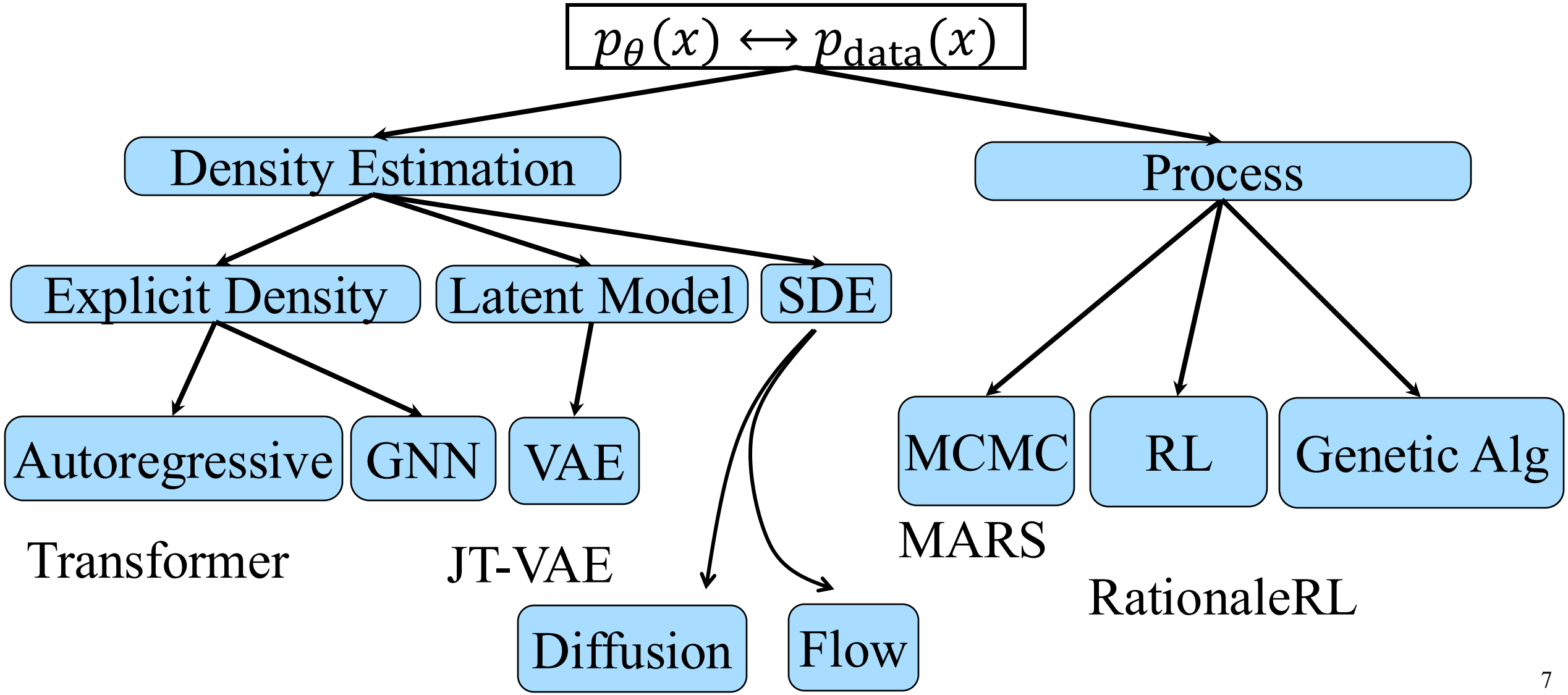
AI Powered Drug Discovery

- AI can generate drug-like molecules with desired properties.
- Using deep generative model (GENTRL) to design novel and biological active drug candidates in 46 days.

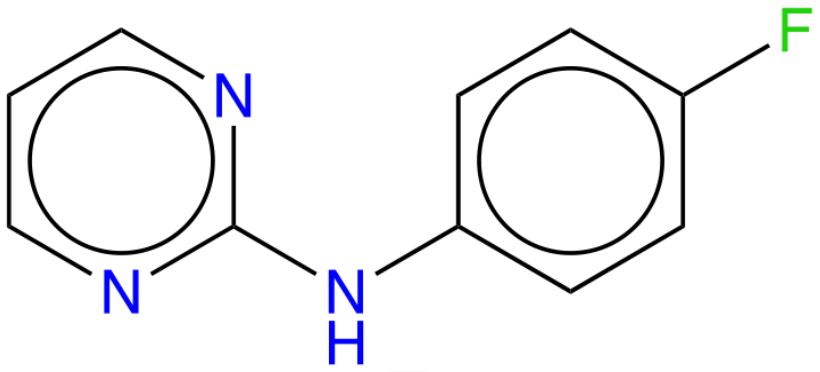


*Zhavoronkov et al. Nature Biotechnology **37**,1038-1040(2019)*

Generative Models for Molecules

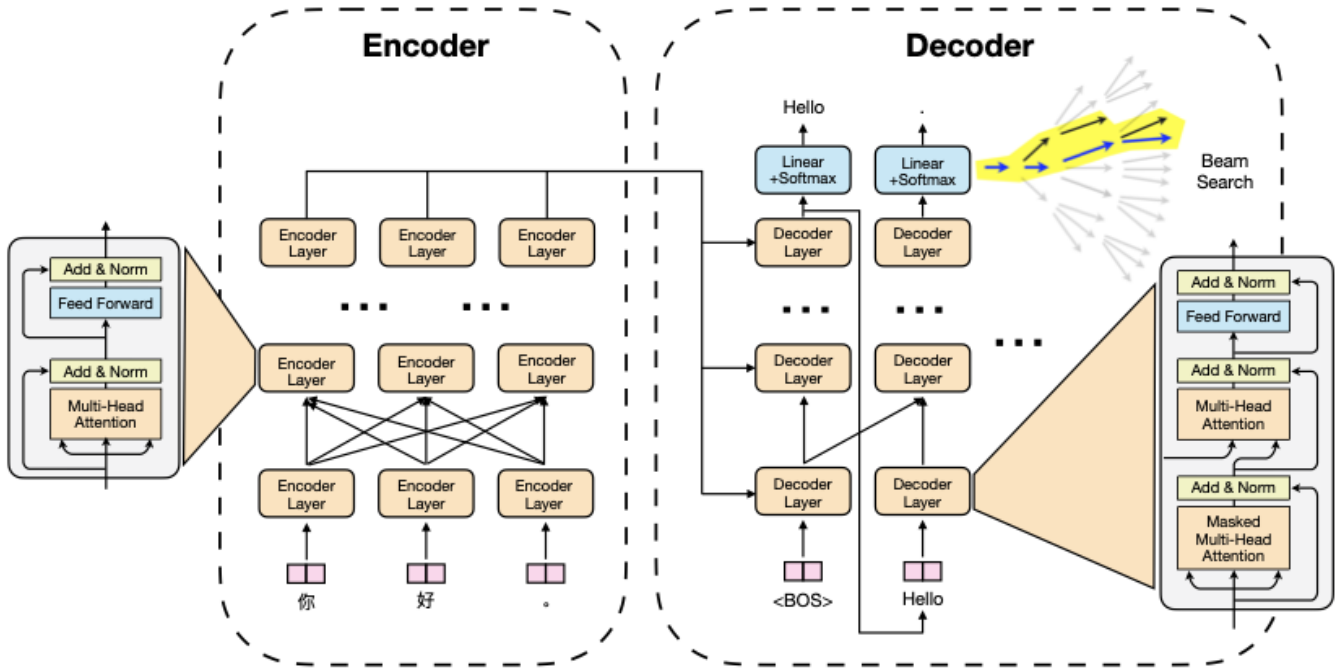


Perspective 1: Molecule Sequence Generation



C10N3FH8

Fc2ccc(Nc1ncccn1)cc2

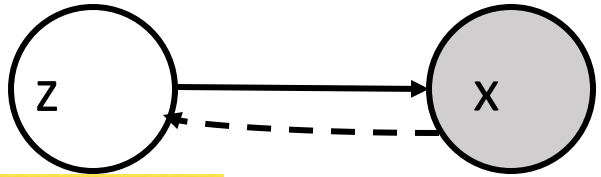


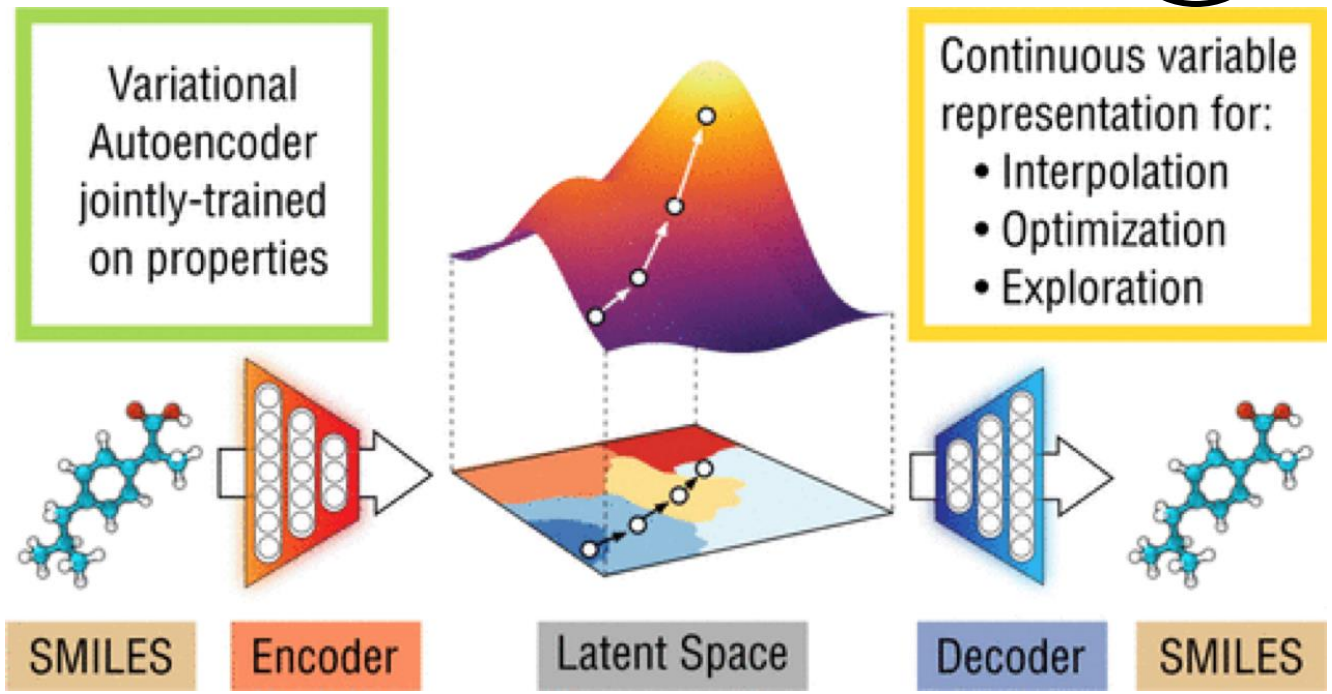
Transformer

$$p(x) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)\cdots$$

Generation from Latent Space

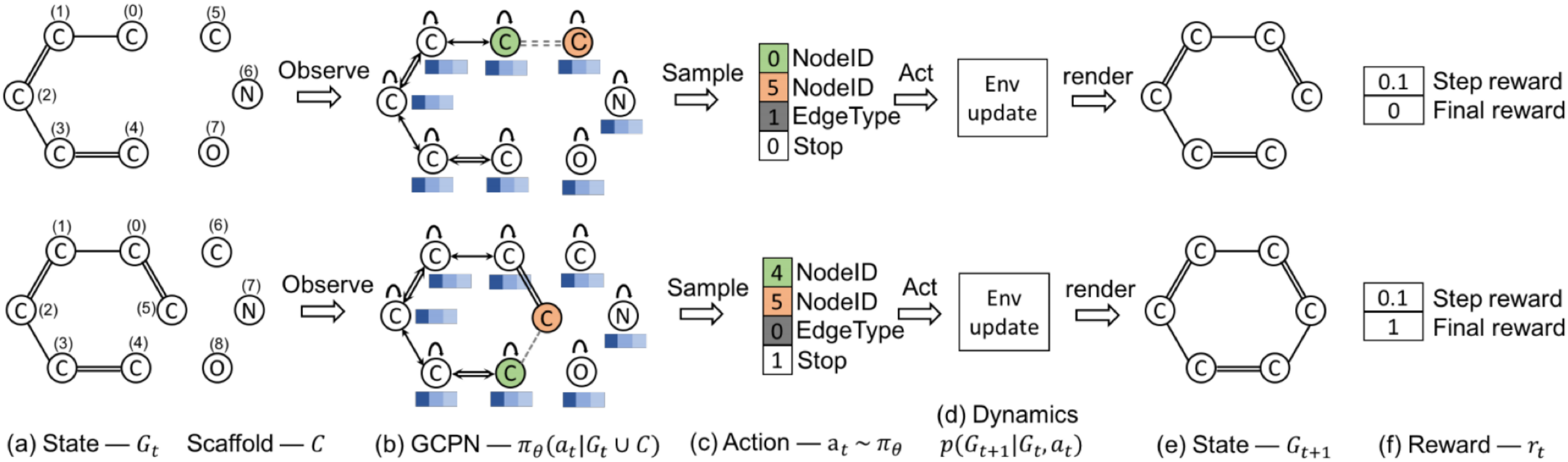
$$p(x) = \int p(z)p(x|z)dz$$





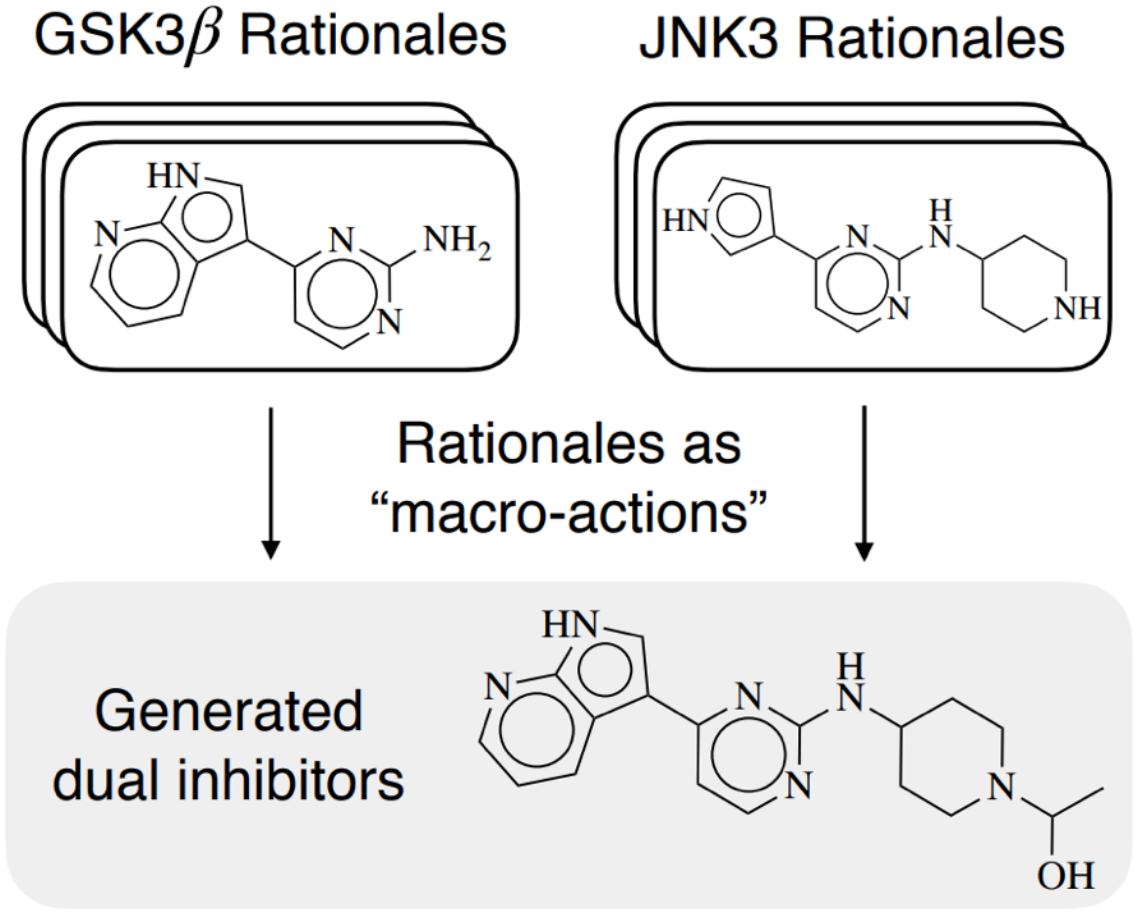
Automatic chemical design using a data-driven continuous representation of molecules. Gomez-Bombarelli et al. ACS Central Science 2016.

Reinforcement-Learning based Generation



Graph convolutional policy network for goal-directed molecular graph generation. You et al, NeurIPS 2018

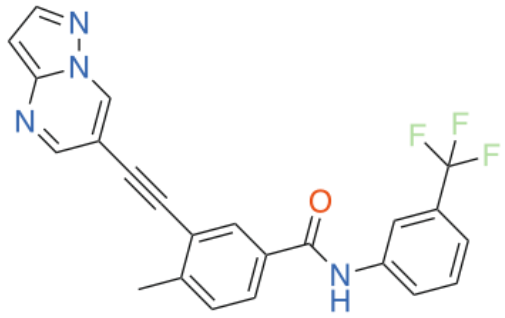
RationaleRL



Multi-objective Molecule Generation using Interpretable Substructures. Jin et al, ICML 2020.

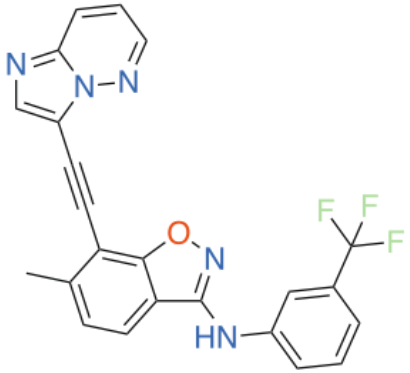
Challenges

- Current generative model can only find molecules with high similarities comparing to existing drugs/potential drug candidates.



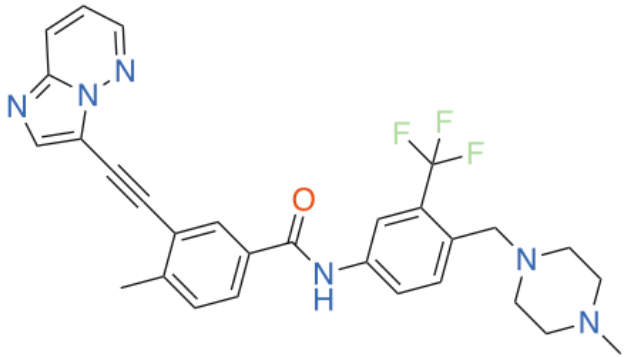
Gao et al.
Compound 7r
6 nM

Previously Reported Drug Candidate



Zhavoronkov et al.
Compound 1
10 nM

AI Designed Drug Candidate



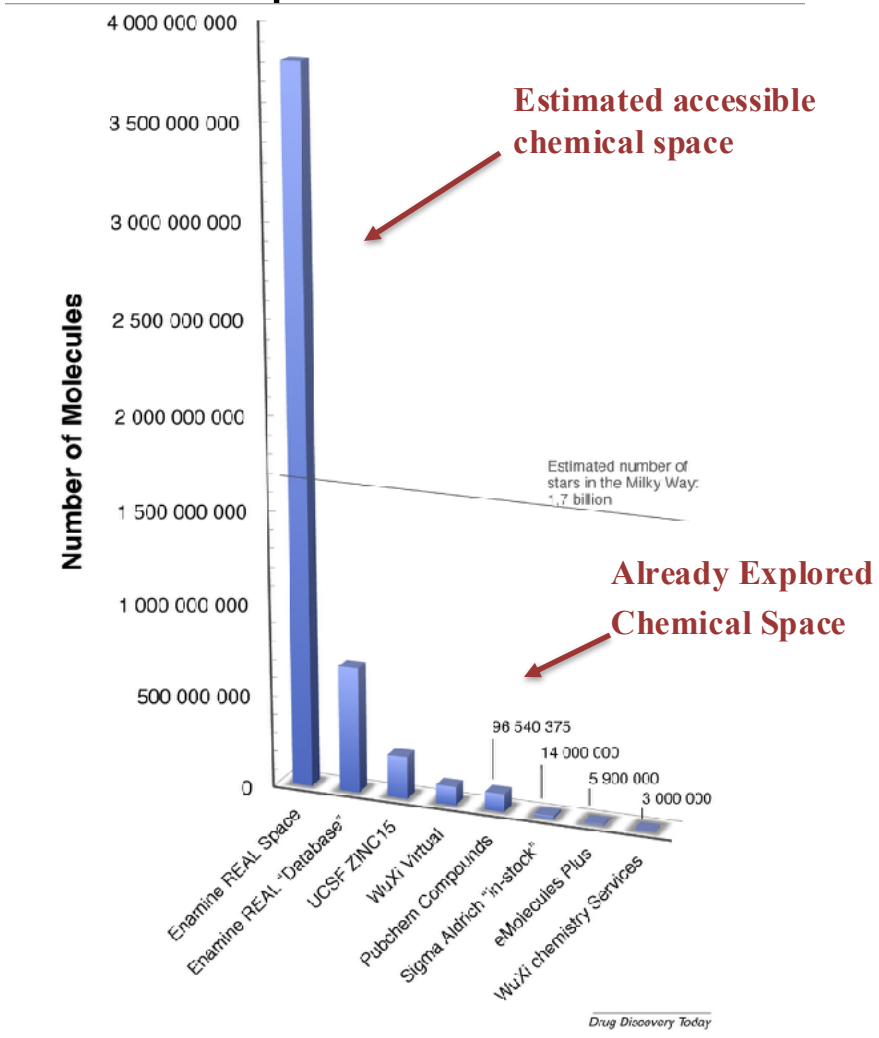
Ponatinib
9 nM

FDA Approved Drug

Walters et al. Nature Biotechnology ***38****, 143–145(2020)*

Practical Drug Discovery

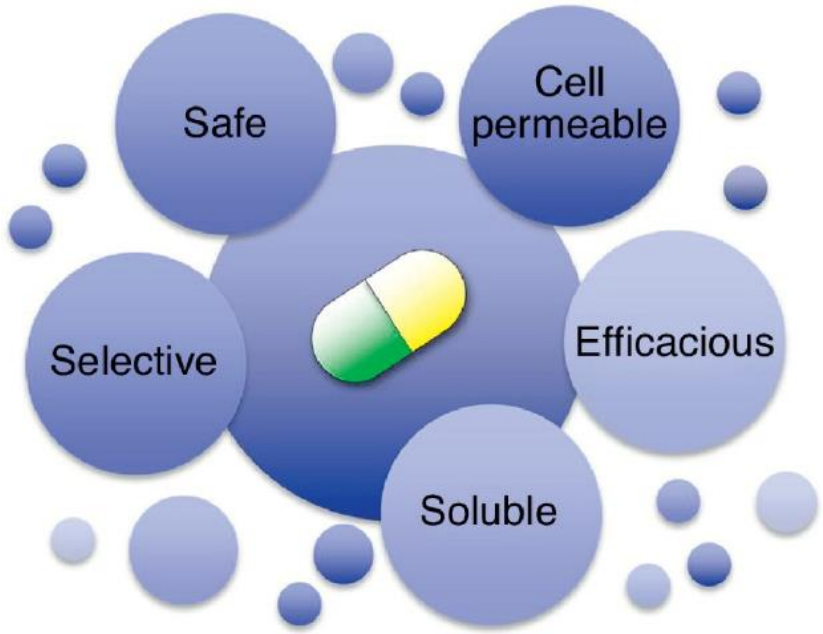
- Broader exploration of chemical space.



- Chemical space is huge but under-explored.
- About 85% drug candidates fail at clinical trials.
- Finding diverse drug candidates can improve the success rate in clinical trials.

Hoffmann et al. Drug Discovery Today ***24****(5), 1148-1156(2019)*

Multi-Objective Optimization for Drug Discovery

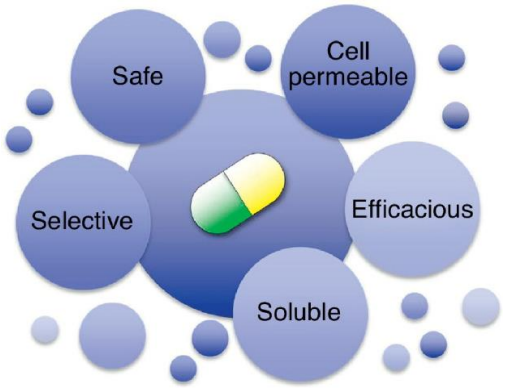


- Successful drugs need to meet multiple requirements, e.g. bioactivity, safety and etc.
- In drug discovery, these properties should also be optimized.

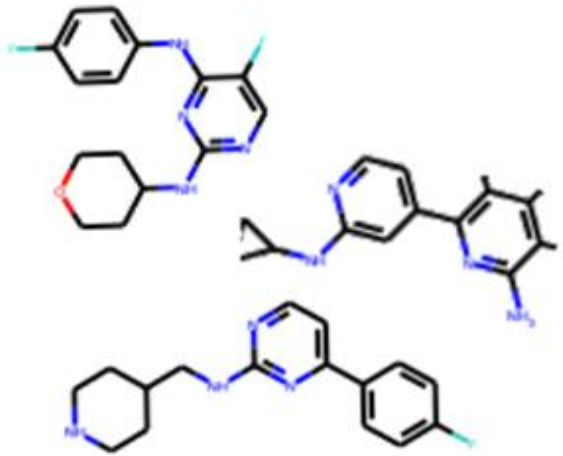
*Nicolaou et al. Drug Discovery Today: Technologies **10**(3), e427-e435(2019)*

Goal: Finding Diverse Molecules Satisfying Multiple Objectives

Satisfy multiple properties with high scores



Produce diverse and novel molecules



Does not rely on lab measured data

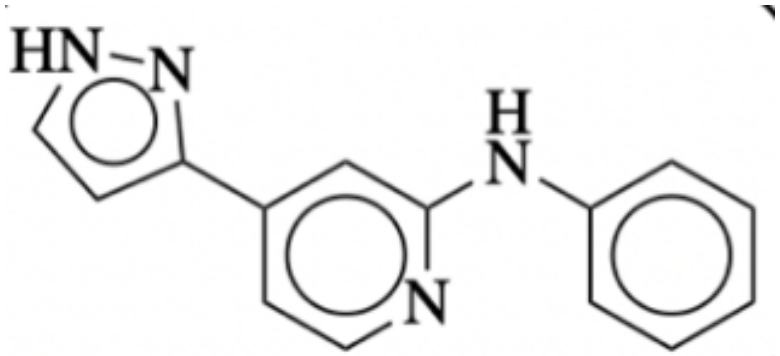
MARS: Markov Molecule Sampling

Yutong Xie

- MCMC sampling from

$$\pi(x) = \underbrace{s_1(x) \circ s_2(x) \circ s_3(x) \circ \cdots \circ s_K(x)}_{\textbf{desired properties}}$$

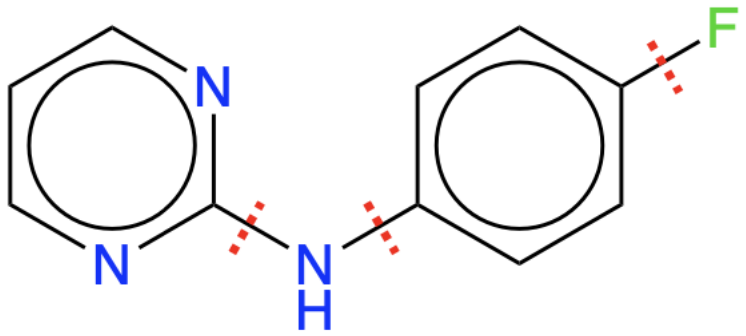
- $s_k(x)$ is scoring function
  - QED: drug-likeness
  - SA: possibility to synthesize



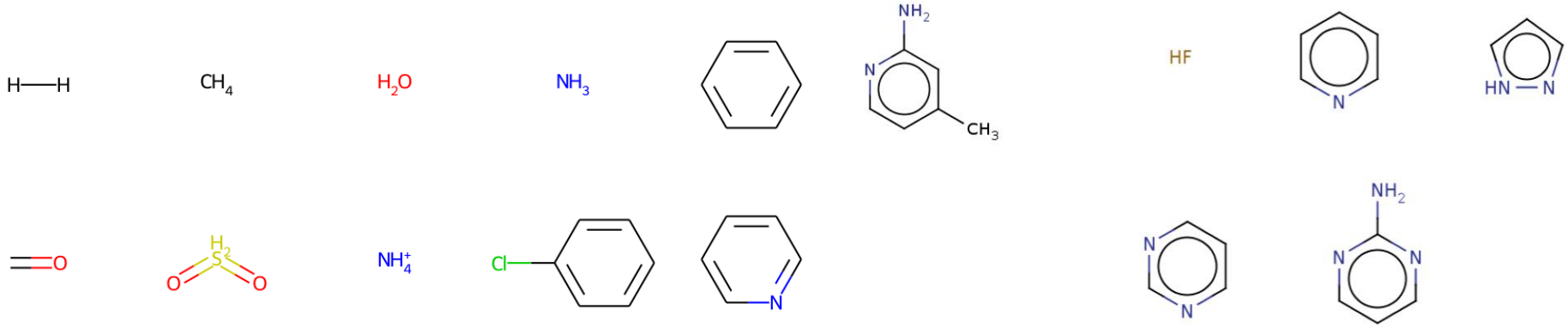
- $\pi(x)$ is the unnormalized distribution over molecule space.

Xie et al. MARS: Markov Molecular Sampling for Multi-objective Drug Discovery. ICLR 2021.

Mining Molecular Fragments

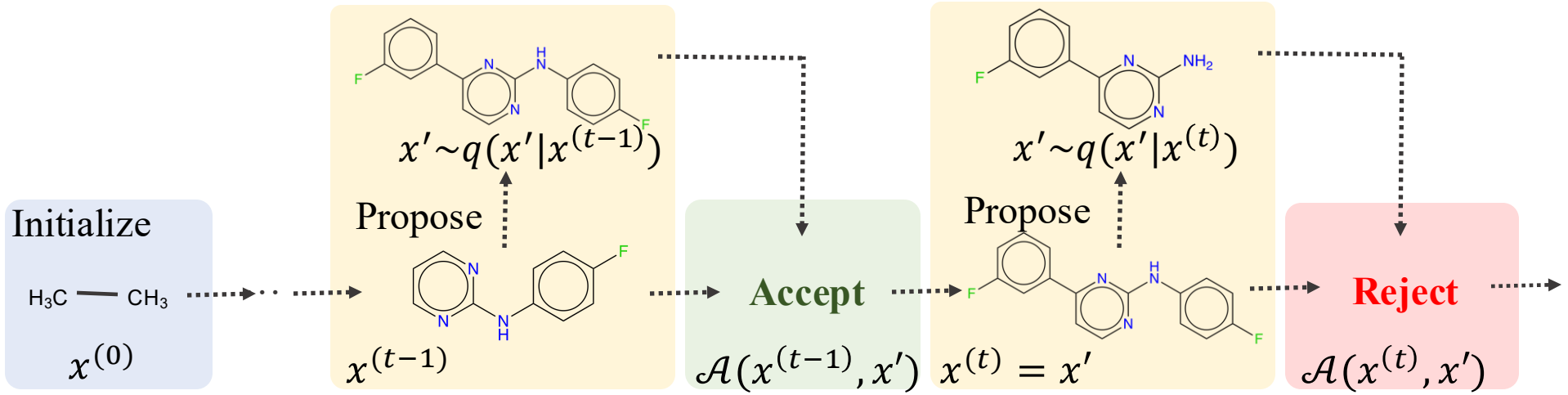


ChEMBL: 2.3M compounds
Breaking single bond and Mining 1000 most frequent fragments with no more than 10 heavy atoms



866	9.6%	c1ccncn1	2	25.4%	F	2	18.2%	F
29	8.8%	c1ccncc1	289	1.9%	C1=NCCN1	29	3.2%	c1ccncc1
380	2.5%	c1cn[nH]c1	694	1.7%	Nc1cc(C)ccn1	380	3.1%	c1cn[nH]c1
289	2.4%	C1=NCCN1	754	1.4%	NC1CCC1	866	2.8%	c1ccncn1
604	2.2%	C=C1SC(=O)NC1=O	686	1.0%	NC1CCOCC1	662	2.4%	Nc1ncccn1

MARS: Iterative Graph Editing

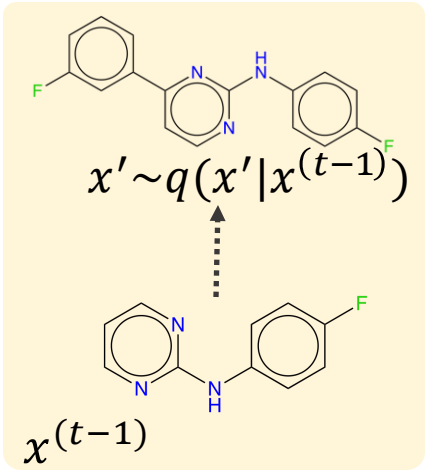


- Start from initial molecule (e.g. C2H6)
- For each step, propose a new molecule x' by modifying existing one, $x' \sim q(x'|x^{(t-1)})$
- Accept wrt ratio: $$\mathcal{A}(x, x') = min\left\{1, \frac{\pi^{\alpha}(x')q(x|x')}{\pi^{\alpha}(x)q(x'|x)}\right\}$$

Adaptive Proposal Learning in MARS

- MCMC sampling

$$\pi(x) = s_1(x) \circ s_2(x) \circ s_3(x) \circ \cdots \circ s_K(x)$$



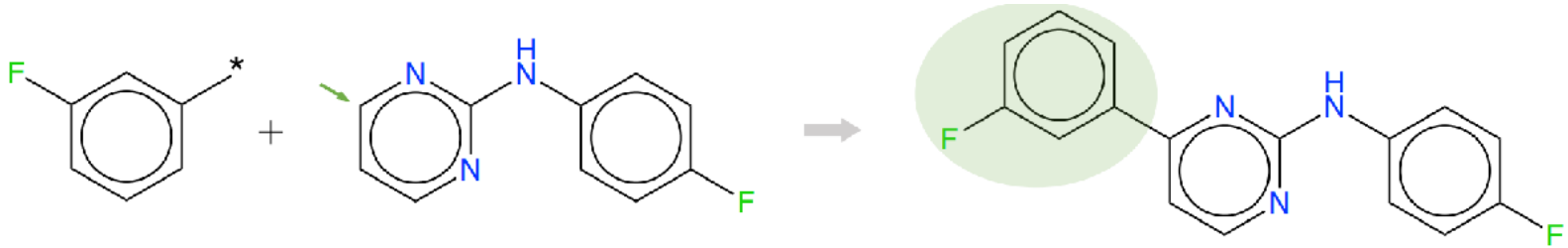
- Markov-chain with annealing scheme to find optimal samples

- Adaptive MCMC proposal: MPNN learned on sampled molecular graphs

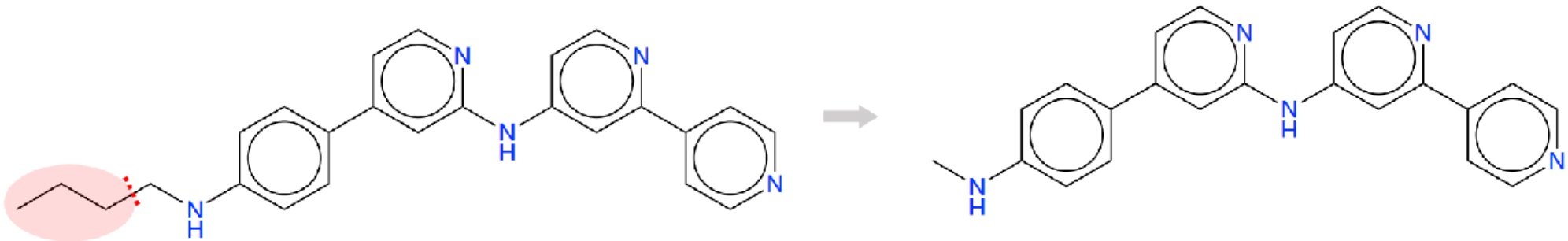
Xie et al. MARS: Markov Molecular Sampling for Multi-objective Drug Discovery. ICLR 2021.

Molecular Graph Editing

- Adding fragment



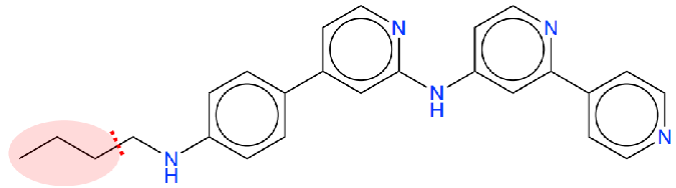
- Deleting fragment



Xie et al. MARS: Markov Molecular Sampling for Multi-objective Drug Discovery. ICLR 2021.

Learning Proposal Network (MPNN)

- Modeling editing actions as node, edge, and graph prediction problems



$$\boldsymbol{h}_b^{\text{edge}} = \text{Concat}(\boldsymbol{h}_v^{\text{node}}, \boldsymbol{h}_w^{\text{node}}) \in \mathbb{R}^{2d}$$

$$\boldsymbol{h}^{\text{graph}} = \text{MaxPooling}(\{\boldsymbol{h}_u^{\text{node}}\}) \in \mathbb{R}^{d}$$

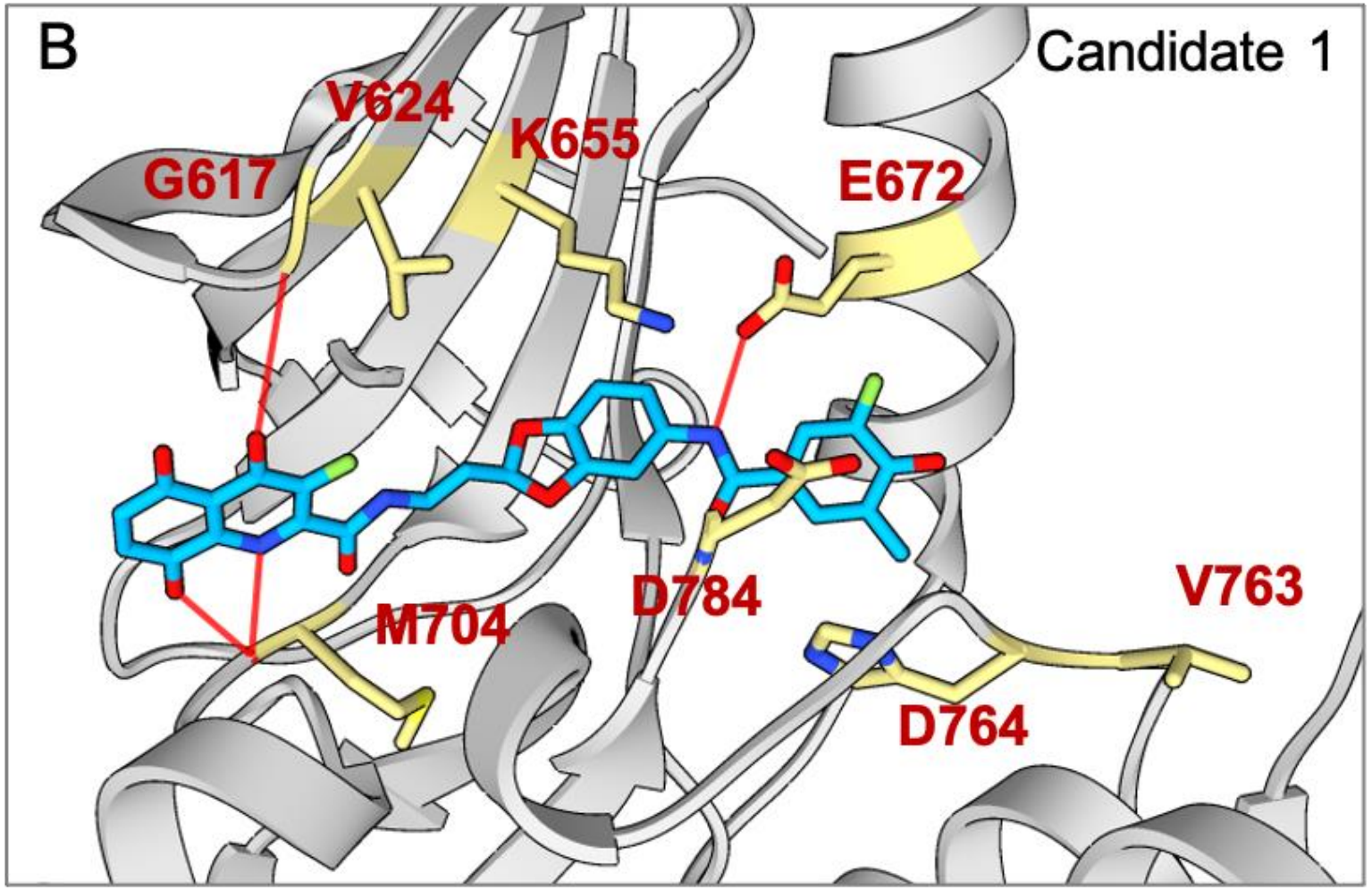
$$p_{\text{add}} = \text{Softmax}(\{\text{MLP}_{\text{node}}(\boldsymbol{h}_u^{\text{node}})\}) \rightarrow \text{Prob. adding}$$

$$p_{\text{frag}} = \text{Softmax}(\text{MLP}_{\text{graph}}(\boldsymbol{h}^{\text{graph}})) \rightarrow \text{Prob. adding}$$

$$p_{\text{del}} = \text{Softmax}(\{\text{MLP}_{\text{edge}}(\boldsymbol{h}_b^{\text{edge}})\}) \rightarrow \text{Prob. deleting}$$

Neural message passing for quantum chemistry, Gilmer et al., ICML 2017

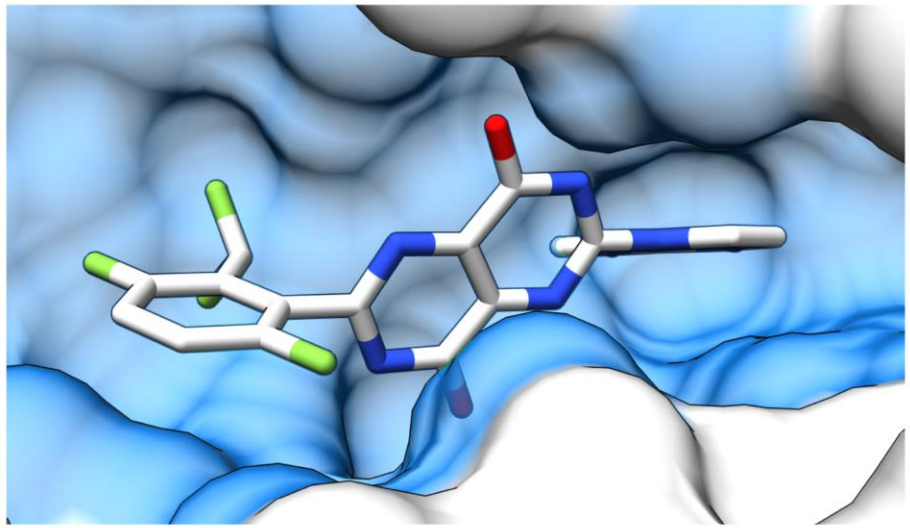
MARS with Docking Score



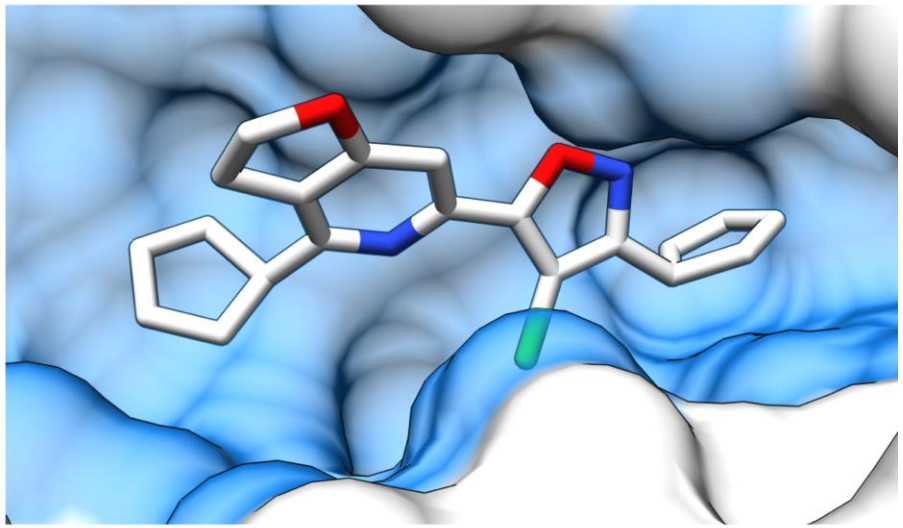
a kinase target

Xie et al. MARS: Markov Molecular Sampling for Multi-objective Drug Discovery. ICLR 2021.

Integrating 3D Structures



Vina: -12.49 QED: 0.58 SAscore: 0.63



Vina: -12.41 QED: 0.80 SAscore: 0.68

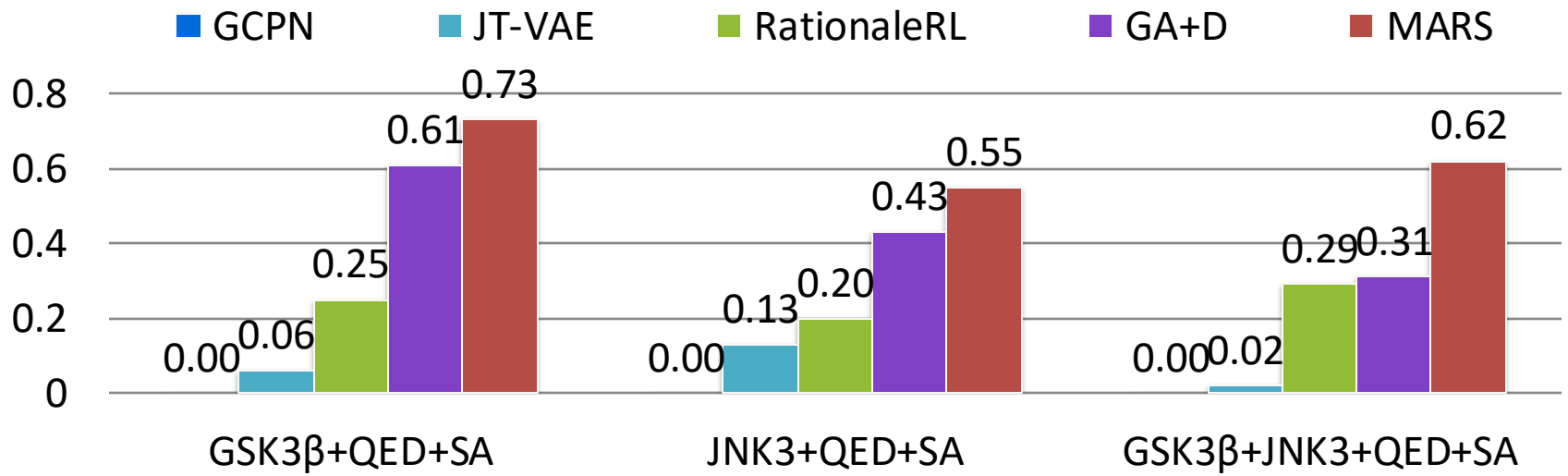
Xie et al. MARS: Markov Molecular Sampling for Multi-objective Drug Discovery. ICLR 2021.

Adaptive Proposal Training

Algorithm 1: MARS.

Initialize molecules $\{x_i^{(0)}\}_{i=1}^{N}$ and the molecular editing model $\mathcal{M}_\theta$;
Create an empty editing model training dataset $\mathcal{D} = \{\}$;
for $t = 1, 2, \ldots$ **do**
  for $i = 1, 2, \ldots, N$ **do**
    Calc probability distributions $(p_{\text{add}}, p_{\text{frag}}, p_{\text{del}}) = \mathcal{M}_\theta(x_i^{(t-1)})$;
    Sample a candidate molecule x' from the proposal distribution $q(x' \mid x_i^{(t-1)})$ defined with probability distributions $p_{\text{add}}, p_{\text{frag}}, p_{\text{del}}$;
    if $u < \mathcal{A}(x_i^{(t-1)}, x')$ *where* $u \sim \mathcal{U}_{[0,1]}$ **then**
      Accept the candidate molecule $x_i^{(t)} = x'$;
    else
      Refuse the candidate molecule $x_i^{(t)} = x^{(t-1)}$;
    end
    if *The candidate improves, i.e.* $\pi(x') > \pi(x_i^{(t-1)})$ **then**
      Adding the editing record $(x_i^{(t-1)}, x')$ into the dataset $\mathcal{D}$;
    end
  end
  Update model $\mathcal{M}_\theta$ with the current dataset $\mathcal{D}$ in a MLE manner;
end

MARS Generates Better Molecules!

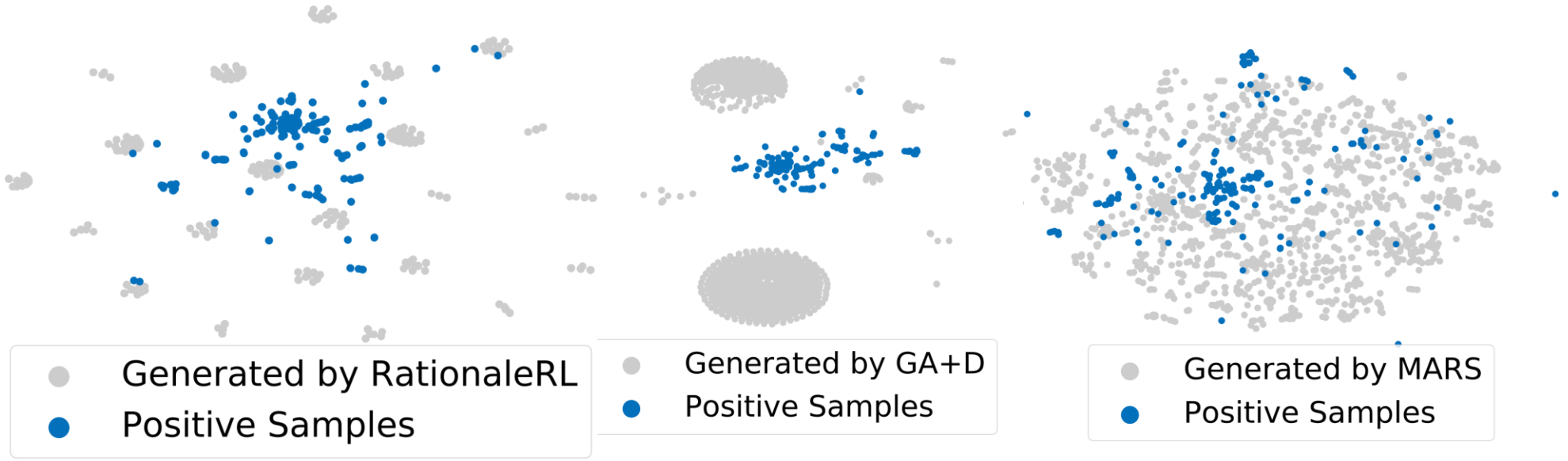


- Generation objectives
  - GSK3β, JNK3: biological inhibition
  - QED: drug-likeness score
  - SA: synthetic accessibility score

- Evaluate w/ Product_score = success_rate * novelty * diversity

Xie et al. MARS: Markov Molecular Sampling for Multi-objective Drug Discovery. ICLR 2021.

MARS explores large chemical space!



Xie et al. MARS: Markov Molecular Sampling for Multi-objective Drug Discovery. ICLR 2021.

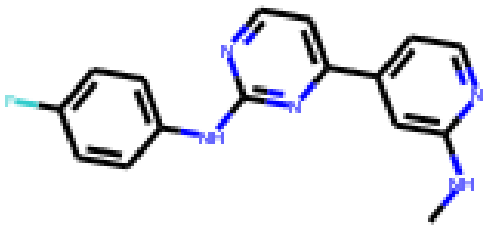
Novel Compounds Found

0.91, 0.85, 0.78, 0.92

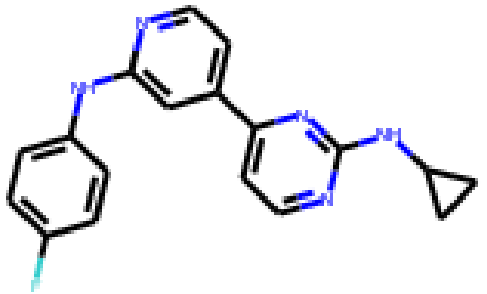
0.90, 0.83, 0.78, 0.93

0.95, 0.75, 0.76, 0.93

0.93, 0.79, 0.75, 0.88



0.89, 0.80, 0.77, 0.88



0.85, 0.87, 0.74, 0.87

Xie et al. MARS: Markov Molecular Sampling for Multi-objective Drug Discovery. ICLR 2021.

Key Takeaways

- Challenges in Drug discovery: multi-objective, novel and diverse, lack of data
  - modeled as a molecule generation problem
- MARS, a simple yet flexible framework for multi-objective
  - Based on MCMC sampling
  - Self-adaptive proposal trained on the fly => no need for data
  - Generates better molecules and explores lager chemical space
- => can discover novel and diverse drug-like molecules
- Challenges remaining:
  - more properties
  - larger molecule, peptide, protein

Reference

Yutong Xie, Chence Shi, Hao Zhou, Yuwei Yang, Weinan Zhang, Yong Yu, Lei Li. MARS: Markov Molecular Sampling for Multi-objective Drug Discovery. ICLR 2021.