WELCOME TO THE

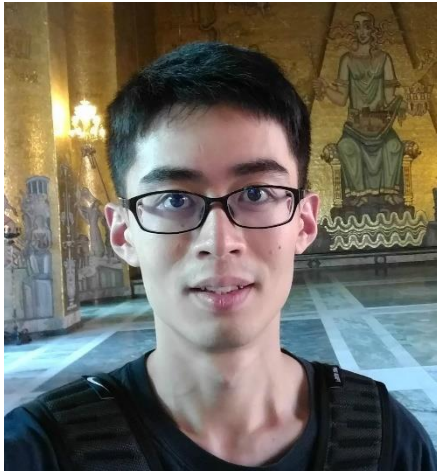
Molecular Team Lecture Series

In this lecture series, MAI LAB Molecular Team will introduce various molecular generation tasks

TODAY’S LECTURE

Multi-Objective Molecule Generation using Interpretable Substructures

Researchers



Wengong Jin
Postdoctoral at
Broad Institute of
MIT and Harvard

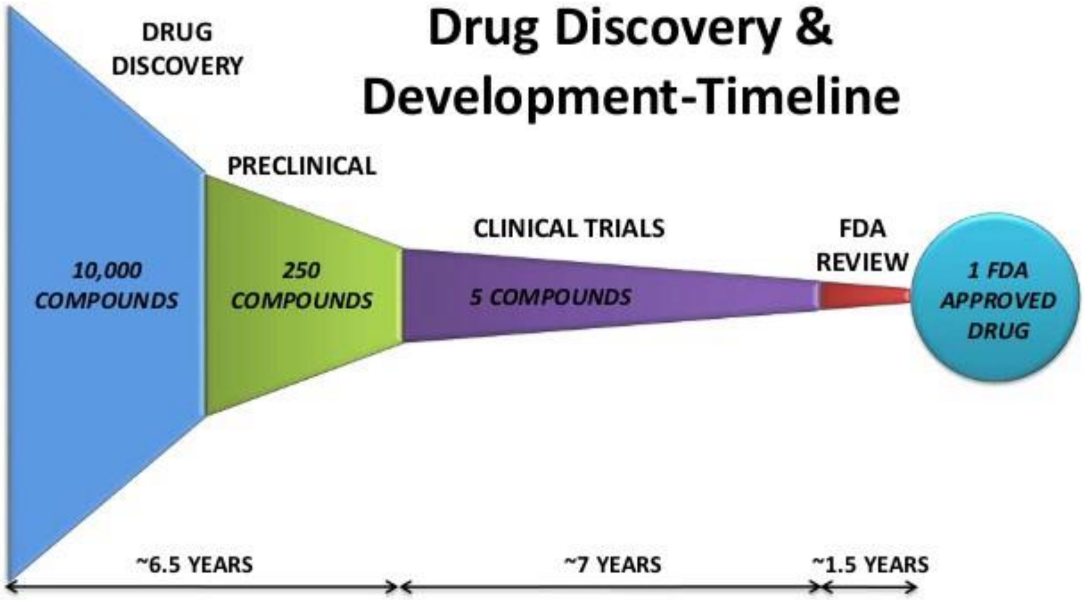
Regina Barzilay
MIT EECS
Professor

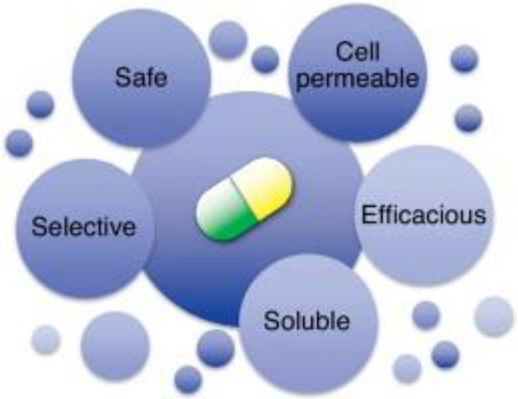
Tommi Jaakkola
MIT CSAIL
Professor

1 Introduction

Background

- ✓ Drug discovery: finding molecules with desired chemical properties
- ✓ Drug needs to satisfy multiple objectives

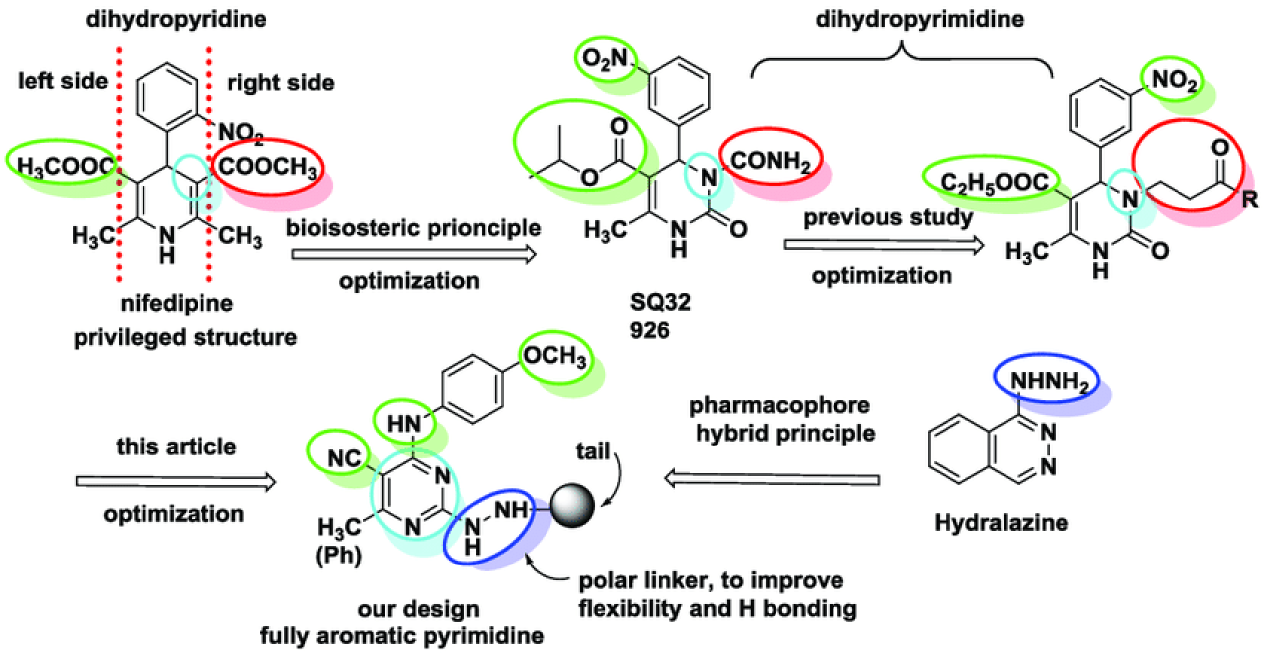




Goal

- Learn to generate sample molecules in the intersection of multiple property constraints
- Multi-property optimization is challenging
- In this paper, composing molecules from a vocabulary of substructures
- Molecular rationales are identified from molecules as substructures

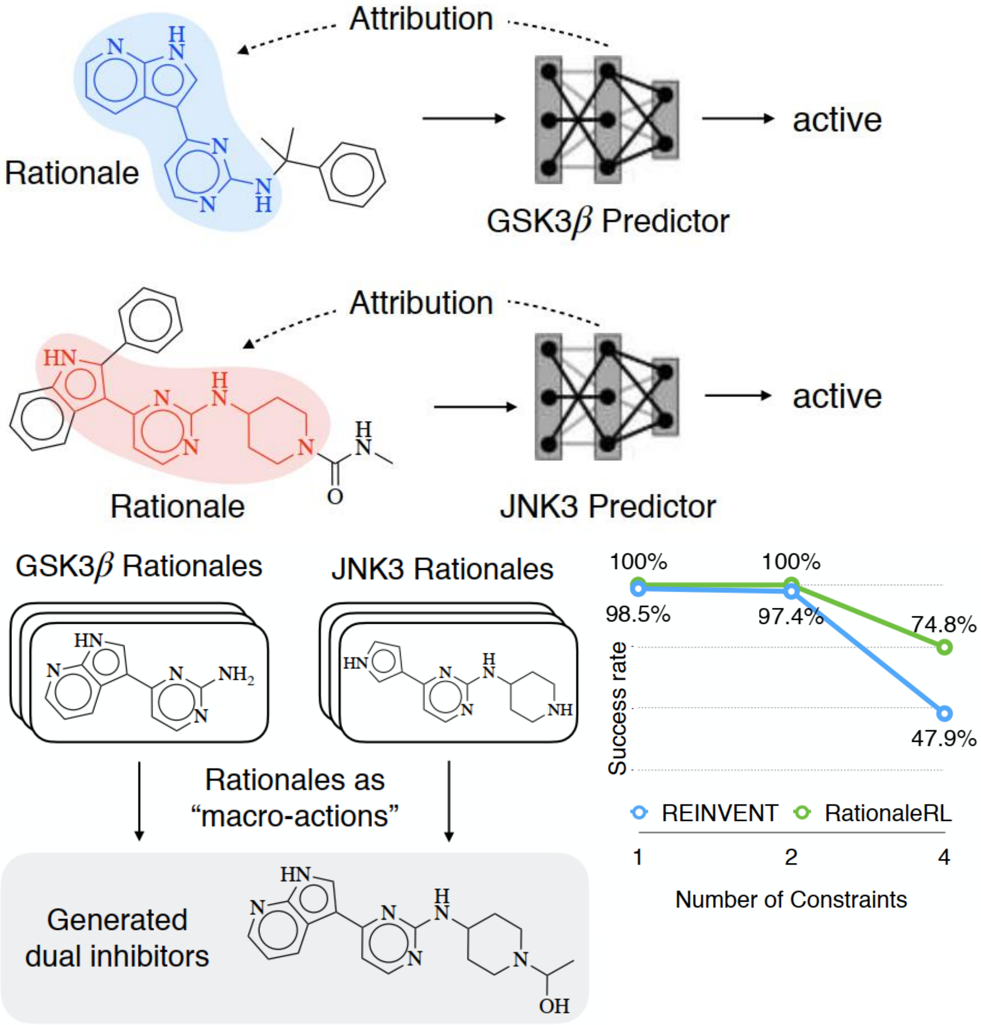
$$P(\mathcal{G}) = \sum_{\mathcal{S}} P(\mathcal{G}|\mathcal{S})P(\mathcal{S})$$



Prior Works

- **Generation Methods for Molecule Design**
  - **JT-VAE (Jin et al., ICML 2018)**
    - Generate molecular graphs in two phases
    - 1) generating a tree-structured scaffold over chemical substructures
    - 2) combining them into a molecule with a graph MPNN
    - This model contains auxiliary property predictors over the VAE latent space
  - **REINVENT (Olivecrona et al., *JChem* 2017)**
    - from AstraZeneca R&D center
    - RL model generating molecules based on their SMILES strings
    - Model is pre-trained over one million molecules and then fine-tuned under property reward
  - **GCPN (You et al., NIPS 2018)**
    - from Stanford University (Jure Leskovec)
    - GCN based model for goal-directed graph generation through RL
    - RL model is trained to optimize domain-specific rewards and adversarial loss
    - They use GAN to help generate realistic molecules

Overview


Attribution
Rationale
GSK3β Predictor
active
Attribution
Rationale
JNK3 Predictor
active
GSK3β Rationales
JNK3 Rationales
Rationales as
"macro-actions"
Generated
dual inhibitors
100%
100%
98.5%
97.4%
74.8%
47.9%
Success rate
REINVENT
RationaleRL
1
2
4
Number of Constraints

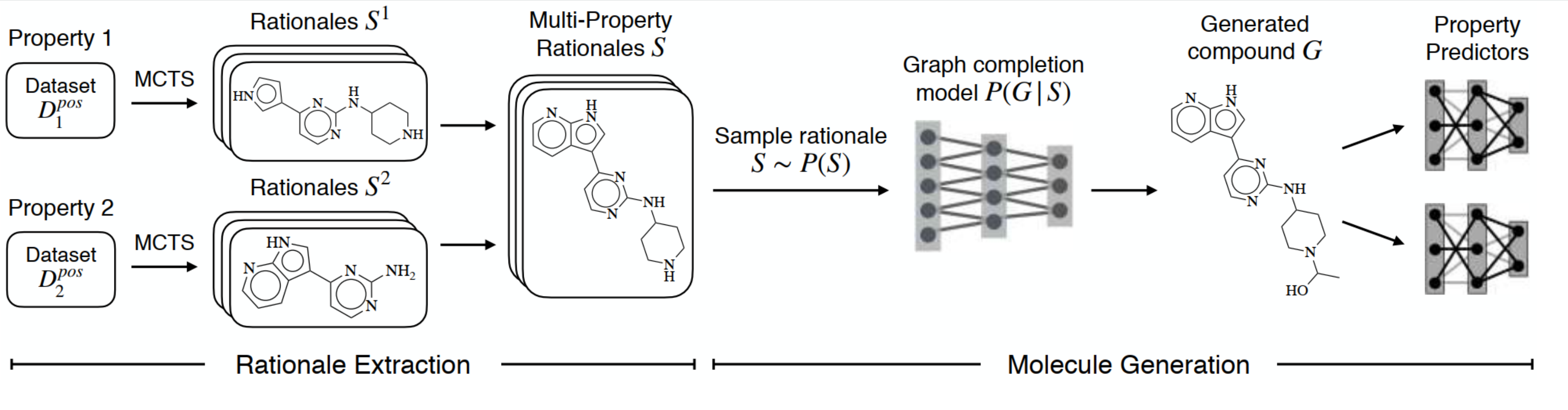
MT

2

Methods

This Paper

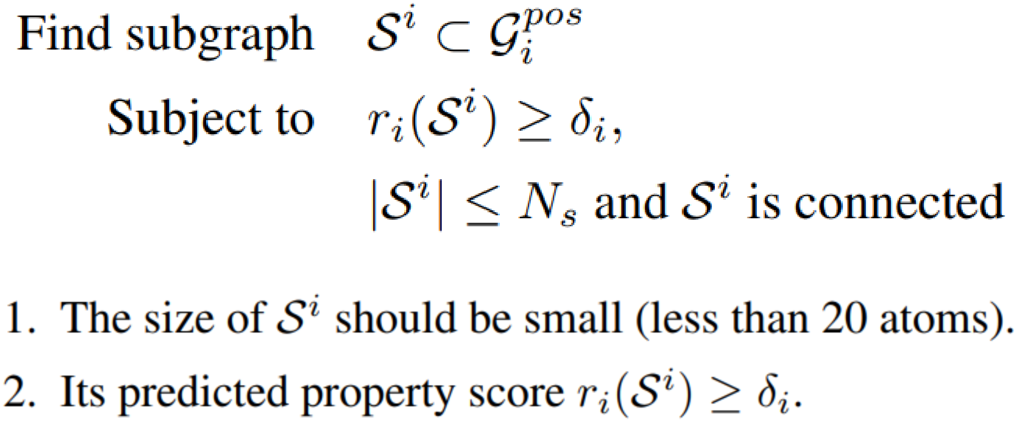
- Construct/extract rationales for each individual property by using MCTS
- Combine each individual property rationales as multi-property rationales
- Learns a graph completion model P(G|S) and rationale distribution P(S)
- Completes a full molecule G given a rationale S



Algorithm Process (1-1)

✓ This model generates molecules by first sampling a rationale *S* from the vocabulary.

✓ **Rationale extraction process**



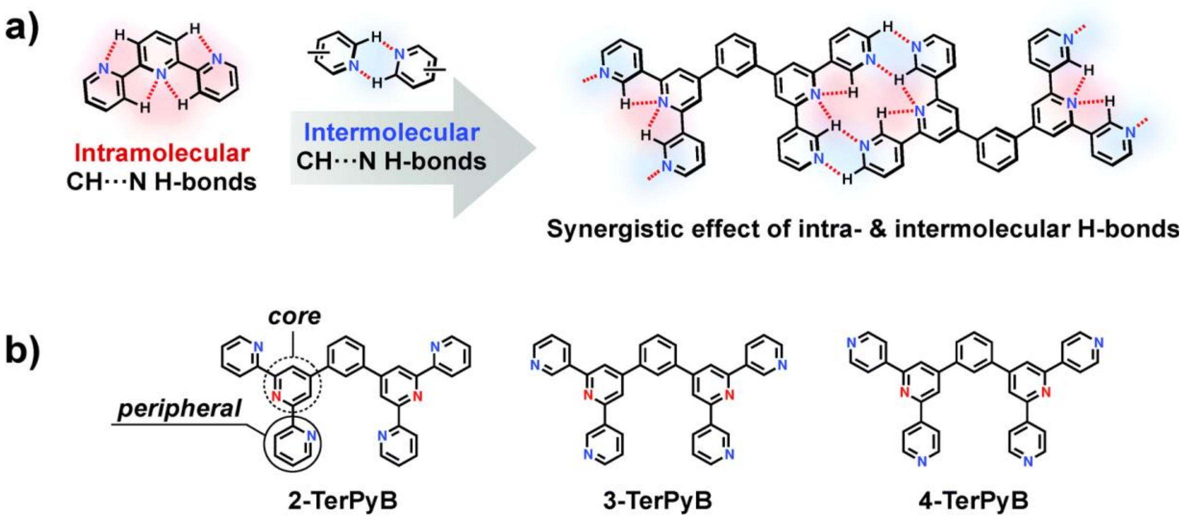
$$\begin{aligned} \text{Find subgraph} \quad & \mathcal{S}^i \subset \mathcal{G}_i^{pos} \\ \text{Subject to} \quad & r_i(\mathcal{S}^i) \geq \delta_i, \\ & |\mathcal{S}^i| \leq N_s \text{ and } \mathcal{S}^i \text{ is connected} \end{aligned}$$

1. The size of $\mathcal{S}^i$ should be small (less than 20 atoms).
2. Its predicted property score $r_i(\mathcal{S}^i) \geq \delta_i$.

Removing Peripheral Bond

Algorithm Process (1-2)

- **Rationale search problem can be solved by MCTS**
  - ✓ **Root:** positive molecules
  - ✓ **State:** subgraph
  - ✓ **Action:** bond deletions (one peripheral non-aromatic bond or one peripheral ring from each state)

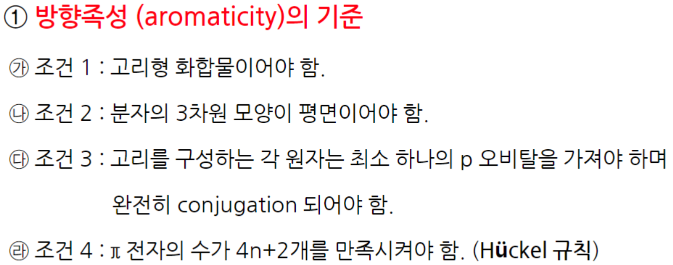


Synergistic effect of intra- & intermolecular H-bonds

2-TerPyB

3-TerPyB

4-TerPyB



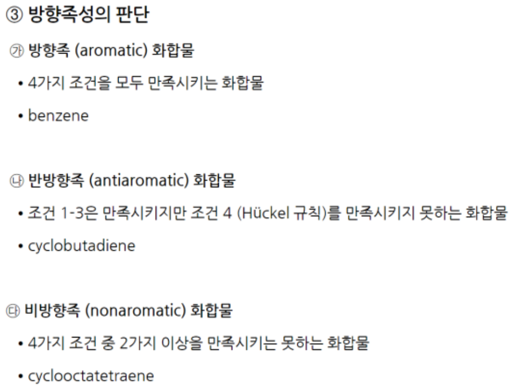
① 방향족성 (aromaticity)의 기준

㉮ 조건 1 : 고리형 화합물이어야 함.

㉯ 조건 2 : 분자의 3차원 모양이 평면이어야 함.

㉰ 조건 3 : 고리를 구성하는 각 원자는 최소 하나의 p 오비탈을 가져야 하며 완전히 conjugation 되어야 함.

㉱ 조건 4 : π 전자의 수가 4n+2개를 만족시켜야 함. (Hückel 규칙)



③ 방향족성의 판단

㉮ 방향족 (aromatic) 화합물

- 4가지 조건을 모두 만족시키는 화합물
- benzene

㉯ 반방향족 (antiaromatic) 화합물

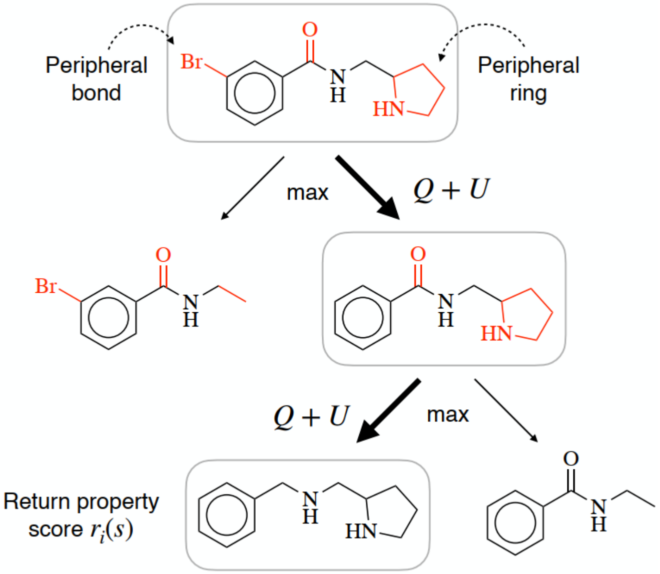
- 조건 1-3은 만족시키지만 조건 4 (Hückel 규칙)를 만족시키지 못하는 화합물
- cyclobutadiene

㉰ 비방향족 (nonaromatic) 화합물

- 4가지 조건 중 2가지 이상을 만족시키는 못하는 화합물
- cyclooctatetraene

Algorithm Process (1-3)

- **MCTS for molecules**
  - ✓ MCTS = RL + Search
  - ✓ peripheral bonds and rings are highlighted in red
  - ✓ In forward pass, the model deletes a peripheral bond
  - ✓ In backward pass, the model updates the statistics



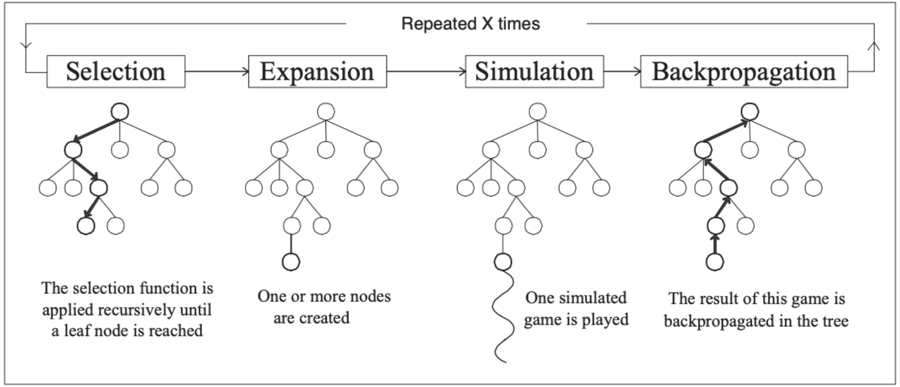
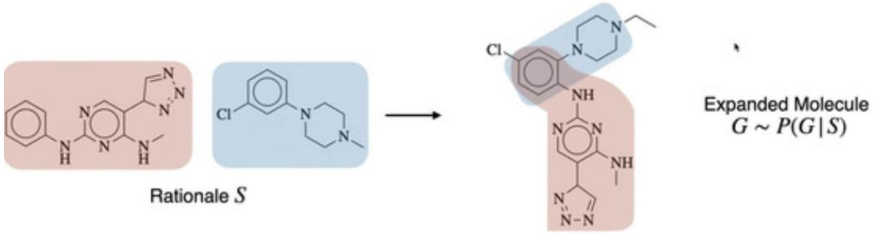

Repeated X times
Selection
Expansion
Simulation
Backpropagation
The selection function is applied recursively until a leaf node is reached
One or more nodes are created
One simulated game is played
The result of this game is backpropagated in the tree

Figure 1: Outline of a Monte-Carlo Tree Search.

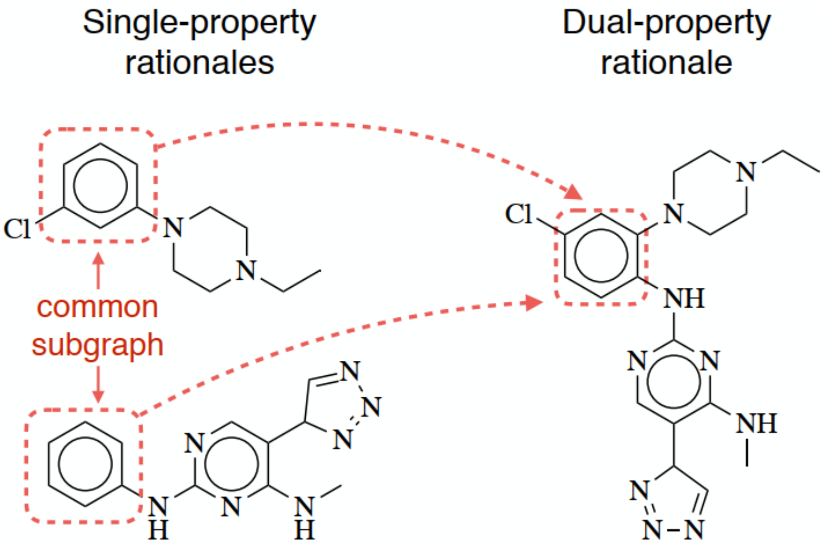
Let’s Dig Deeper

- Rationales are “partial” molecules
- We need to complete them into a full molecule
- Learn a molecule completion model P(G|S) to connect the rationales
- We model P(G|S) as an autoregressive process
- We use a simple atom-by-atom molecule completion model
- In each step, we add an atom to the current molecule, and predict its associated bonds



Algorithm Process (1-4)

- **Multi-property rationale construction**
  - ✓ Given two single-property rationales, find their maximum common substructure
  - ✓ Then, superposing two rationales so that their MCS coinicdes



$$\forall i : r_i(\mathcal{S}^{[M]}) \geq \delta_i, i = 1, \cdots, M$$

$$C_{\mathcal{S}}^{M} = \bigcup_{(\mathcal{S}^1, \cdots, \mathcal{S}^M)} \text{MERGE}(\mathcal{S}^1, \cdots, \mathcal{S}^M)$$

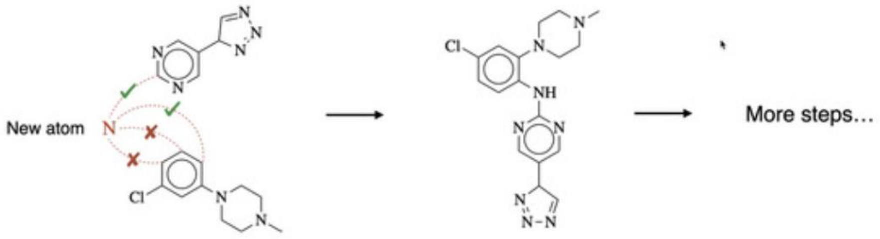
$$V_{\mathcal{S}}^{[M]} = \{\mathcal{S} \in C_{\mathcal{S}}^{M} \mid r_i(\mathcal{S}^{[M]}) \geq \delta_i, \forall i\}$$

Algorithm Process (2)

- **Graph Completion**
  - ✓ VAE, which completes a full molecule *G* given a rationale *S*
  - ✓ Encoder: message passing network for atom representation
  - ✓ Decoder: generates molecule *G* by BFS, must include subgraph

$$P(\mathcal{G}|\mathcal{S}) = \int_{\boldsymbol{z}} P(\mathcal{G}|\mathcal{S}, \boldsymbol{z})P(\boldsymbol{z})d\boldsymbol{z}$$

$$\{\boldsymbol{h}_v\} = \mathrm{MPN}_e\left(\mathcal{G}, \{\boldsymbol{e}(a_u)\}, \{\boldsymbol{e}(b_{uv})\}\right)$$



1. Predict whether there will be a new atom attached to v_t:

$$\boldsymbol{p}_t = \mathrm{sigmoid}(\mathrm{MLP}(\boldsymbol{h}_{v_t}^{(t)}, \boldsymbol{h}_{\mathcal{G}_t}, \boldsymbol{z}_{\mathcal{G}})) \qquad (13)$$

   where $\mathrm{MLP}(\cdot, \cdot, \cdot)$ is a ReLU network whose input is a concatenation of multiple vectors.

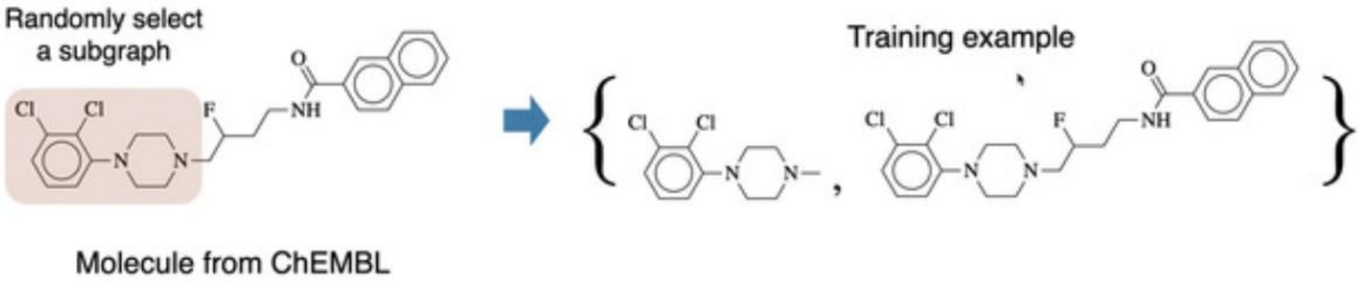
2. If $\boldsymbol{p}_t < 0.5$, discard v_t and move on to the next node in $\mathcal{Q}$. Stop generation if $\mathcal{Q}$ is empty. Otherwise, create a new atom u_t and predict its atom type:

$$\boldsymbol{p}_{u_t} = \mathrm{softmax}(\mathrm{MLP}(\boldsymbol{h}_{v_t}^{(t)}, \boldsymbol{h}_{\mathcal{G}_t}, \boldsymbol{z}_{\mathcal{G}})) \qquad (14)$$

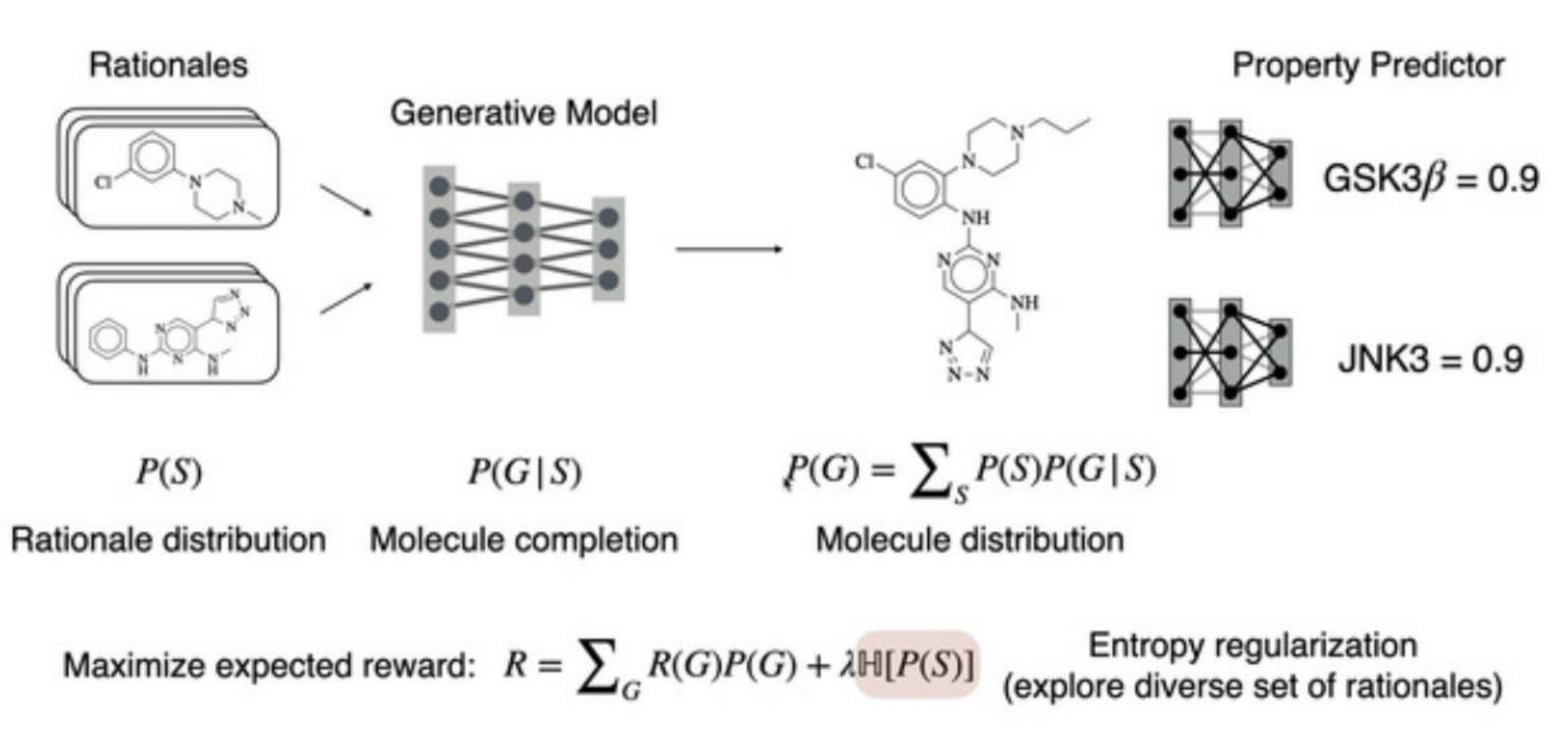
3. Predict the bond type between u_t and other frontier nodes in $\mathcal{Q} = \{q_1, \cdots, q_n\}$ $(q_1 = v_t)$. Since atoms are generated in breadth-first order, there are no bonds between u_t and atoms not in $\mathcal{Q}$.

Algorithm Process (3)

- **Pre-training Molecule Completion**
  - ✓ Molecule completion model can be trained w/o “property” predictors
  - ✓ Pre-train molecule completion on a large dataset (ChEMBL)



Putting Everything Together



$P(S)$

Rationale distribution

$P(G \mid S)$

Molecule completion

$P(G) = \sum_S P(S)P(G \mid S)$

Molecule distribution

Maximize expected reward: $R = \sum_G R(G)P(G) + \lambda \mathbb{H}[P(S)]$

Entropy regularization
(explore diverse set of rationales)

MT

3

Results

Evaluation Metric

1) Success rate:

✓ How often do generated molecules satisfy all the property constraints?

✓ Following REINVENT, we use property predictors to compute this metric

2) Diversity:

✓ Average pairwise molecule distance

$$\mathrm{Diversity} = 1 - \frac{2}{n(n-1)} \sum\nolimits_{X,Y} \mathrm{sim}(X, Y)$$

3) Novelty:

✓ We don't want to rediscover existing drugs known to satisfy all the constraints

$$\mathrm{Novelty} = \frac{1}{n} \sum\nolimits_{\mathcal{G}} \mathbf{1}\left[\mathrm{sim}(\mathcal{G}, \mathcal{G}_{\mathrm{SNN}}) < 0.4\right]$$

Result Table

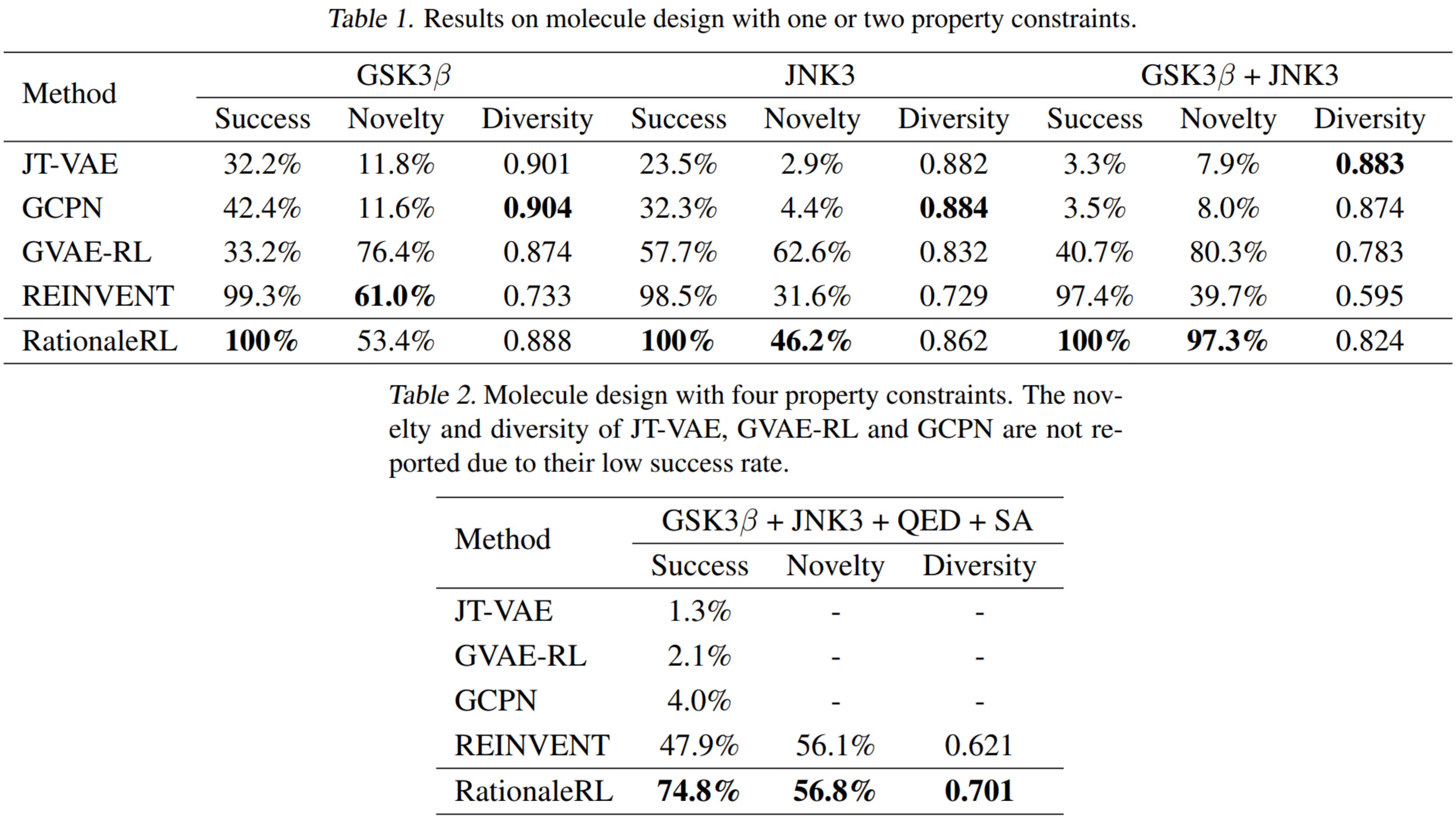


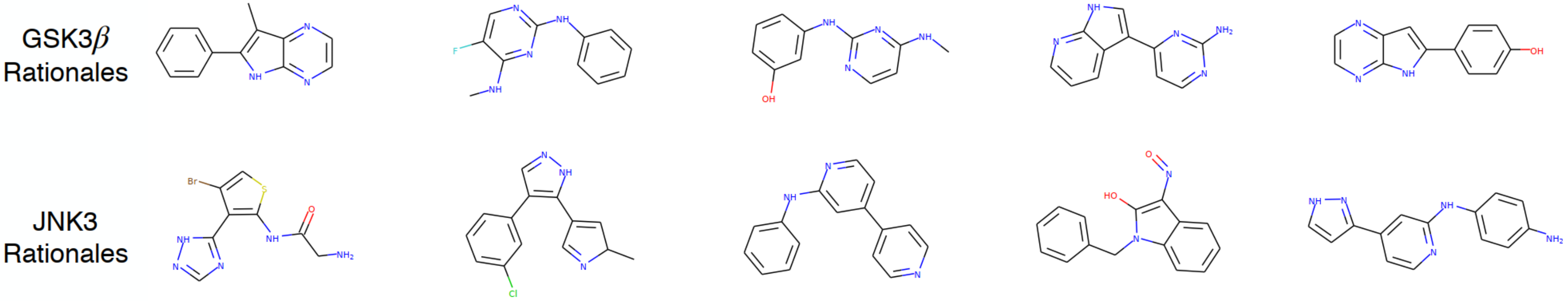
Table 1. Results on molecule design with one or two property constraints.

Method	GSK3β			JNK3			GSK3β + JNK3		
	Success	Novelty	Diversity	Success	Novelty	Diversity	Success	Novelty	Diversity
JT-VAE	32.2%	11.8%	0.901	23.5%	2.9%	0.882	3.3%	7.9%	**0.883**
GCPN	42.4%	11.6%	**0.904**	32.3%	4.4%	**0.884**	3.5%	8.0%	0.874
GVAE-RL	33.2%	76.4%	0.874	57.7%	62.6%	0.832	40.7%	80.3%	0.783
REINVENT	99.3%	**61.0%**	0.733	98.5%	31.6%	0.729	97.4%	39.7%	0.595
RationaleRL	**100%**	53.4%	0.888	**100%**	**46.2%**	0.862	**100%**	**97.3%**	0.824

Table 2. Molecule design with four property constraints. The novelty and diversity of JT-VAE, GVAE-RL and GCPN are not reported due to their low success rate.

Method	GSK3β + JNK3 + QED + SA		
	Success	Novelty	Diversity
JT-VAE	1.3%	-	-
GVAE-RL	2.1%	-	-
GCPN	4.0%	-	-
REINVENT	47.9%	56.1%	0.621
RationaleRL	**74.8%**	**56.8%**	**0.701**

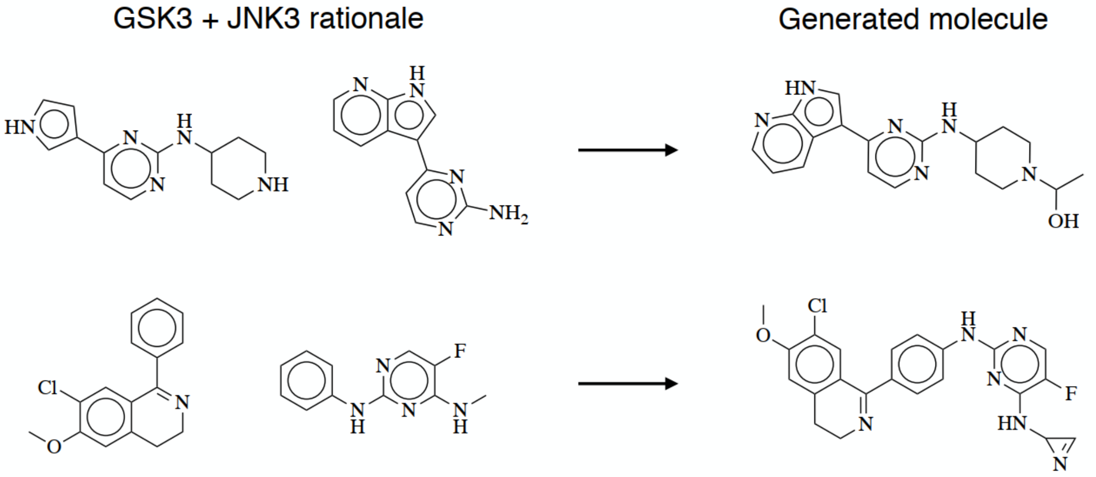
Examples



GSK3β Rationales

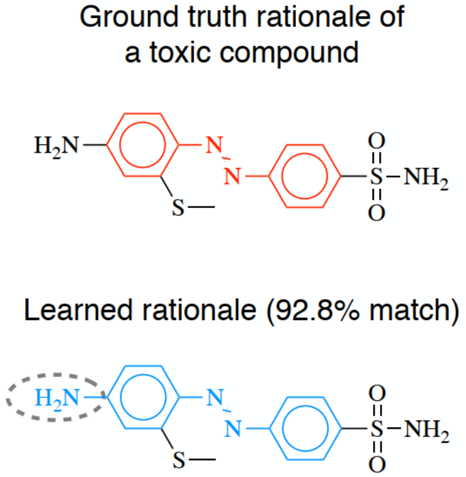
JNK3 Rationales

Figure 6. Sample rationales of GSK3β (top) and JNK3 (bottom).



GSK3 + JNK3 rationale

Generated molecule



Ground truth rationale of a toxic compound

Learned rationale (92.8% match)

Summary

- **Molecular graph generation is particularly challenging due to multiple constraints**

- **In this paper, authors propose hierarchical RL based on rationales**

- **Rationales are extracted by MCTS and then combined to be formed full molecules by graph VAE**

- **Limitation:** instead of atom-by-atom generation, once can use motif substructures mechanisms. (Hierarchical Generation of Molecular Graph using Structural Motif Jin et al., ICML 2020)

Thank you