MARS:

MARKOV MOLECULAR SAMPLING

FOR
MULTI-OBJECTIVE DRUG DISCOVERY

Xie, Yutong, Chence Shi, Hao Zhou, Yuwei Yang,
Weinan Zhang, Yong Yu, and Lei Li.

Presented by: Samira Mali

Monte Carlo importance sampling

- sampling distributions

- target distribution

- Uniform energy model
- Non- uniform energy model

- Annealing
- Metropolis- Hasting

Why sampling from a distribution p(x) is hard?

Liu, Jun S. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

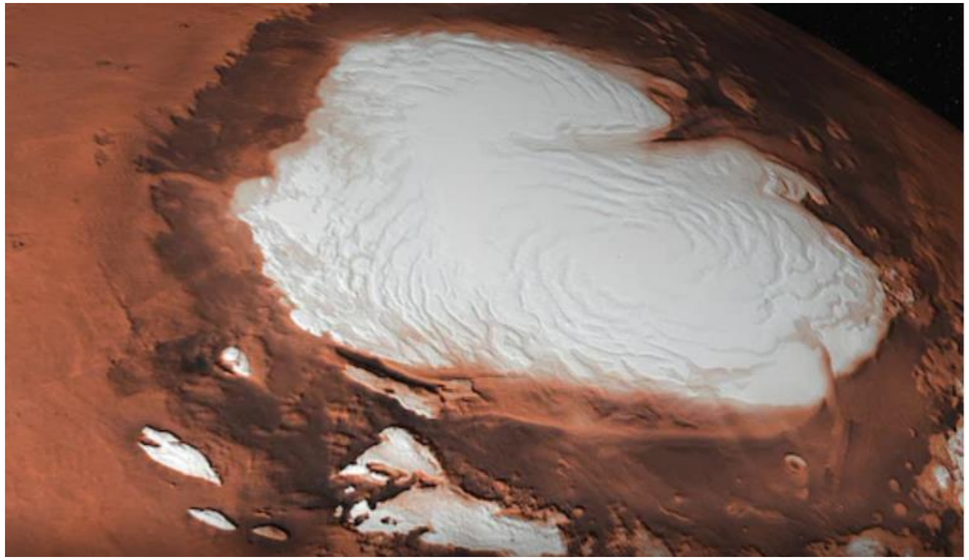
No water on MARS?

Single properties:

Druglikeness (QED)
octanol-water partition coefficient (logP)

Solubility in both water and fat.

The value is greater than one if a substance is more soluble in fat-like solvents, and less than one if it is more soluble in water.



- **C1**: It should satisfy multiple properties with high scores;

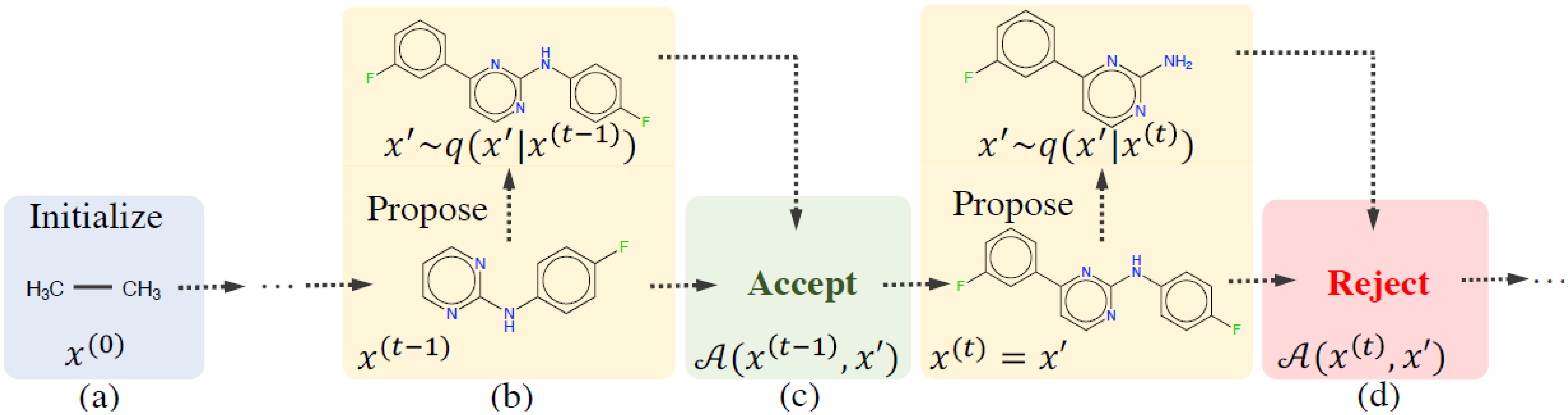
  C2: It should produce novel and diverse molecules;

  C3: Its generation process does not rely on either expert annotated or wet experimental data collected from a biochemistry lab

Framework of MARS

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

$$\mathcal{A}(x, x') = \min\left\{1, \frac{\pi^\alpha(x')q(x|x')}{\pi^\alpha(x)q(x'|x)}\right\}$$

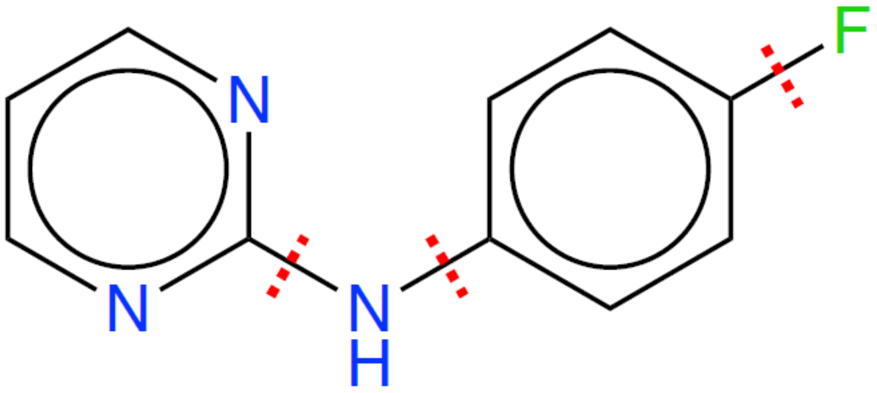


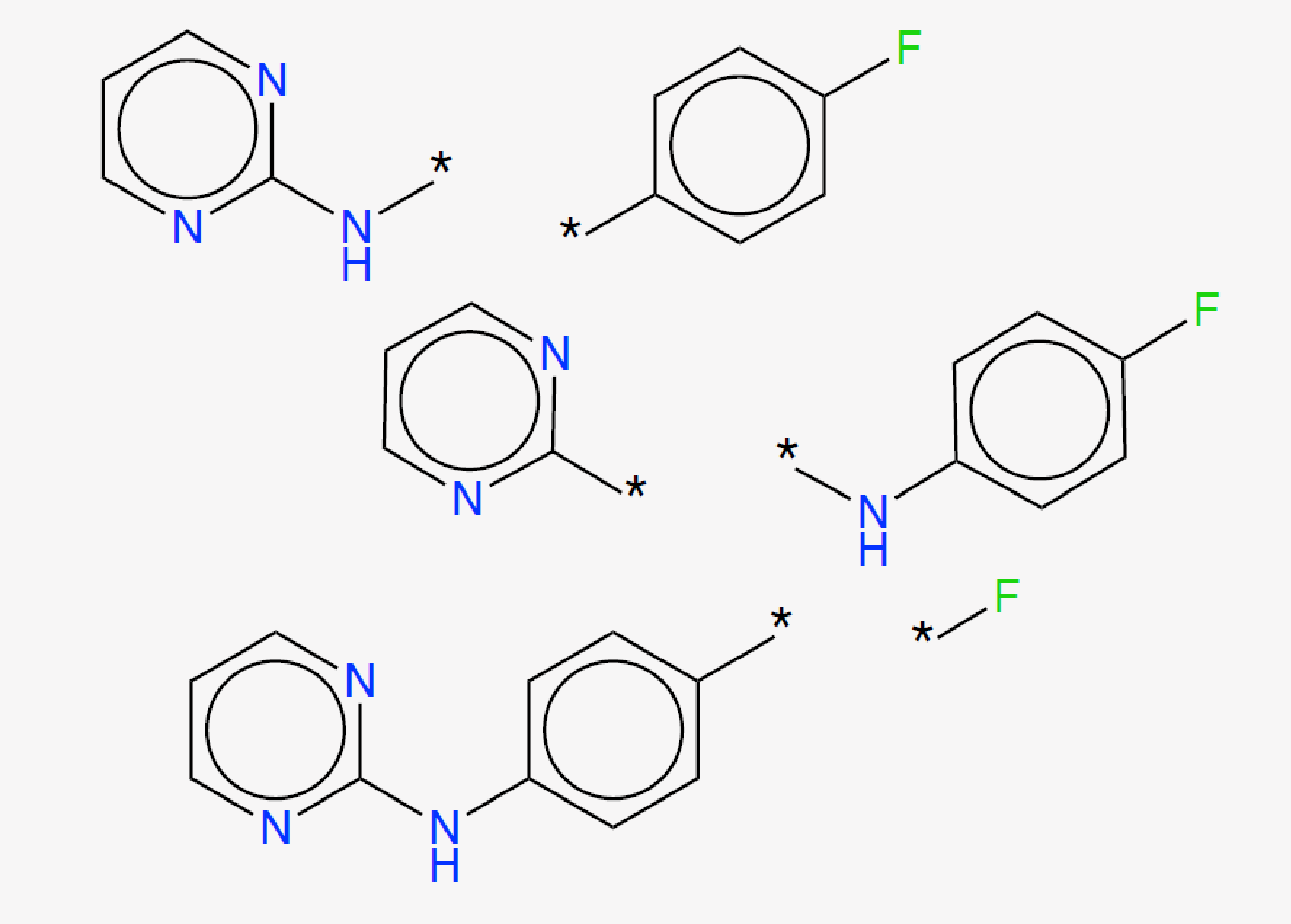
$$q(x'|x) = \frac{1}{2} \cdot p_{\text{add}}(x, u) \cdot p_{\text{frag}}(x, u, k)$$

$$q(x'|x) = \frac{1}{2} \cdot p_{\text{del}}(x, b)$$

Fragments

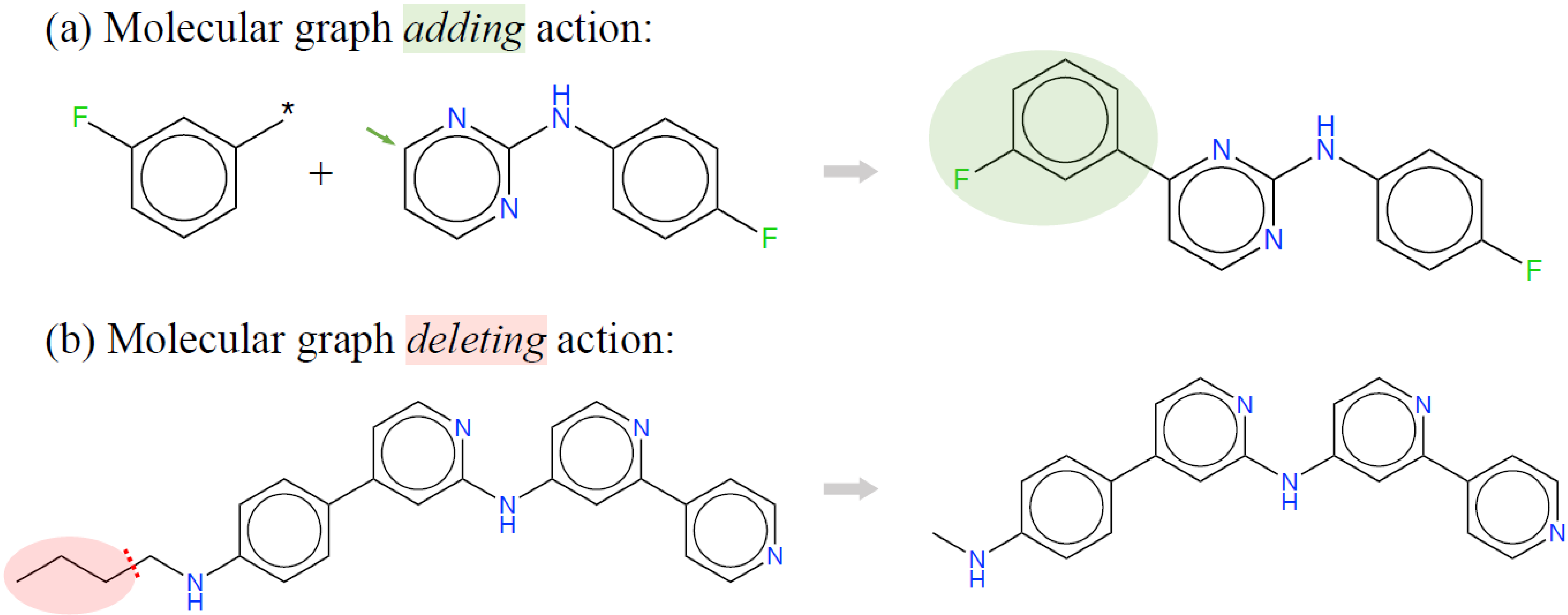
Fragments in molecules





Fragment vocabulary

Molecular graph editing actions



(a) Molecular graph *adding* action:

(b) Molecular graph *deleting* action:

Algorithm 1: MARS

```
Set $N$ initial molecules $\{x_i^{(0)}\}_{i=1}^N$ and initialize the molecular graph 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
        Compute probability distributions $(p_{\text{add}}, p_{\text{frag}}, p_{\text{del}}) = \mathcal{M}_\theta(x_i^{(t-1)})$ as Equations 7-9
        Sample a candidate molecule $x'$ from the proposal distribution $q(x' \mid x_i^{(t-1)})$ defined with
         probability distributions $p_{\text{add}}, p_{\text{frag}}, p_{\text{del}}$ as Equations 3-4
        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 $x_i^{(t)} = x_i^{(t-1)}$
        if The candidate improves the objectives, i.e. $\pi(x') > \pi(x_i^{(t-1)})$ then
            Adding the editing record $(x_i^{(t-1)}, x')$ into the dataset $\mathcal{D}$
    $\theta^{new} \longleftarrow \arg\max \log M_\theta(\mathcal{D})$
```

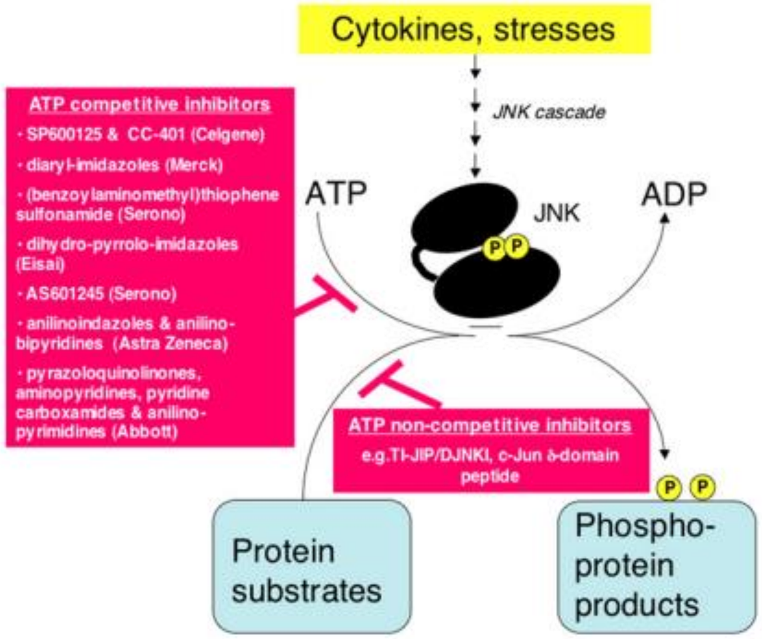
$$q(x'|x) = \frac{1}{2} \cdot p_{\text{add}}(x, u) \cdot p_{\text{frag}}(x, u, k) \tag{3}$$

$$q(x'|x) = \frac{1}{2} \cdot p_{\text{del}}(x, b) \tag{4}$$

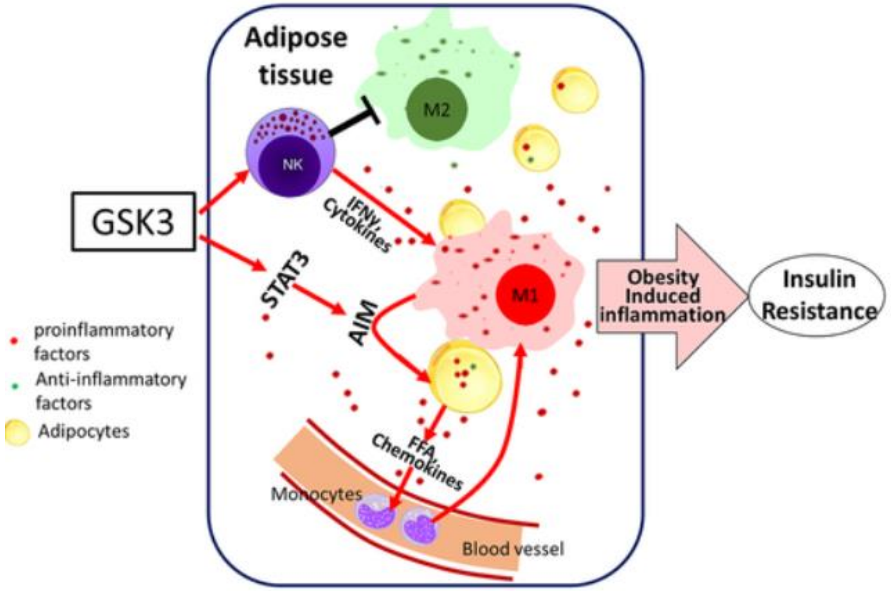
$$p_{\text{add}}(x) = \text{Softmax}(\{\text{MLP}_{\text{node}}(\boldsymbol{h}_u^{\text{node}})\}_{u=1}^n) \in [0, 1]^n \tag{7}$$

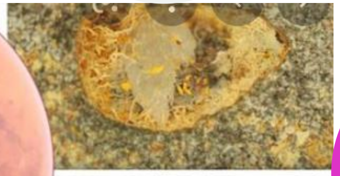
$$p_{\text{frag}}(x, u) = \text{Softmax}(\text{MLP}'_{\text{node}}(\boldsymbol{h}_u^{\text{node}})) \in [0, 1]^{|V|} \tag{8}$$

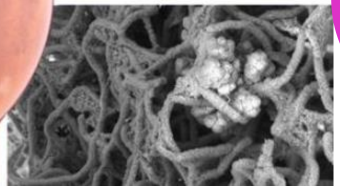
$$p_{\text{del}}(x) = \text{Softmax}(\{\text{MLP}_{\text{edge}}(\boldsymbol{h}_b^{\text{edge}})\}_{b=1}^{2m}) \in [0, 1]^{2m} \tag{9}$$



Cytokines, stresses
JNK cascade
ATP
JNK
ADP
P P
ATP competitive inhibitors
· SP600125 & CC-401 (Celgene)
· diaryl-imidazoles (Merck)
· (benzoylaminomethyl)thiophene sulfonamide (Serono)
· dihydro-pyrrolo-imidazoles (Eisai)
· AS601245 (Serono)
· anilinoindazoles & anilino-bipyridines (Astra Zeneca)
· pyrazoloquinolinones, aminopyridines, pyridine carboxamides & anilino-pyrimidines (Abbott)
ATP non-competitive inhibitors
e.g.TI-JIP/DJNKI, c-Jun δ-domain peptide
P P
Protein substrates
Phospho-protein products







Alzheimer disease in MARS?

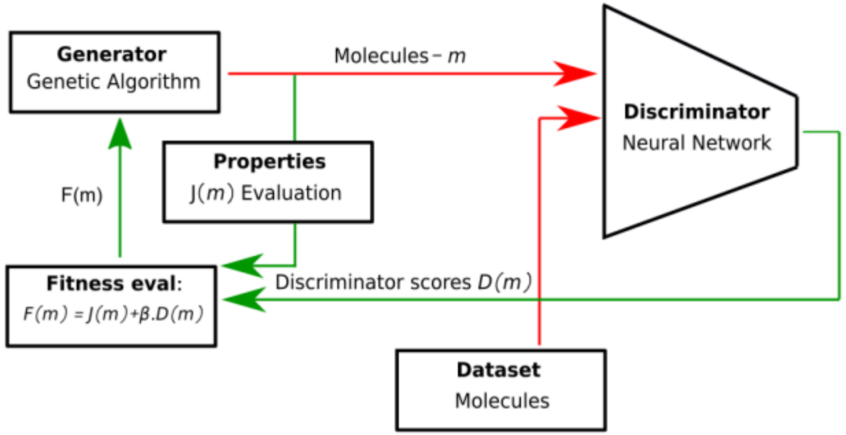
--Experiment Setup

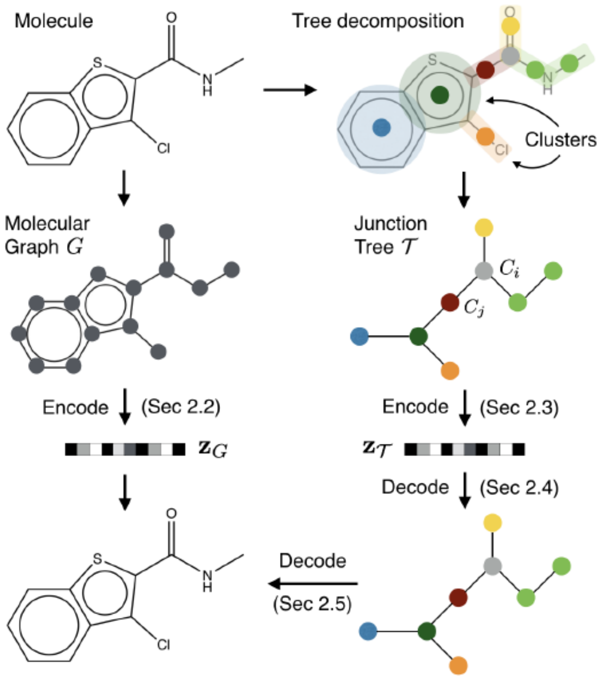
- Biological objectives.

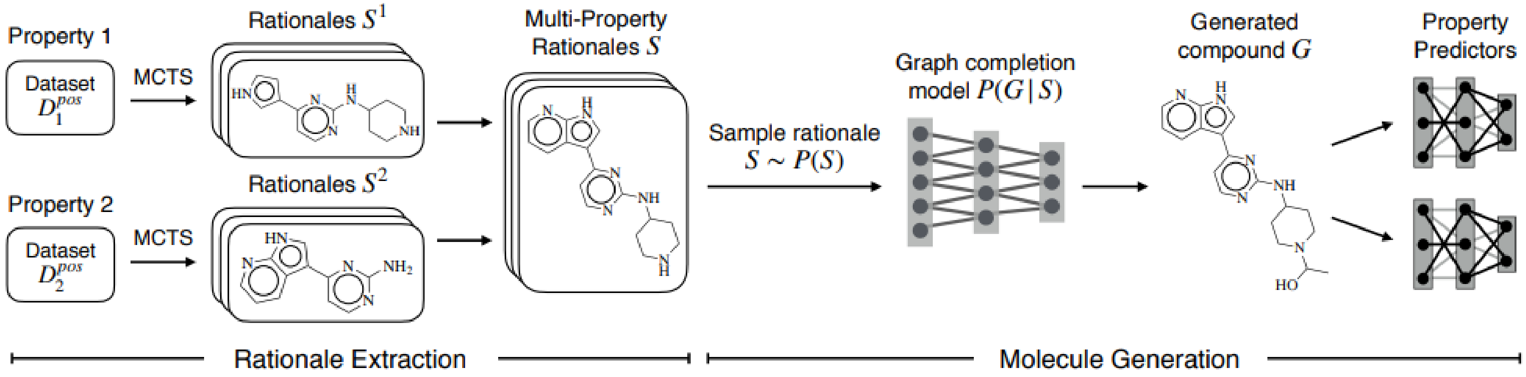
GSK3: Inhibition against glycogen synthase kinase-3.

JNK3: Inhibition against c-Jun N-terminal kinase-3.

- Non-biological objectives
- Multi-objective generation setting







Baselines

- **GCPN** (You et al., 2018) --> Paper 1, presented by Bernard in this class
- **JT-VAE** (Jin et al., 2018)
- **RationaleRL** (Jin et al., 2020)
- **GA+D** (Nigam et al., 2020)

Evaluation metrics

- Success rate (SR):

  Percentage of generated molecules that are evaluated as positive on all given objectives

- Novelty (Nov):

  Percentage of generated molecules with similarity less than 0.4

compared to the nearest neighbor xSNN in the training set

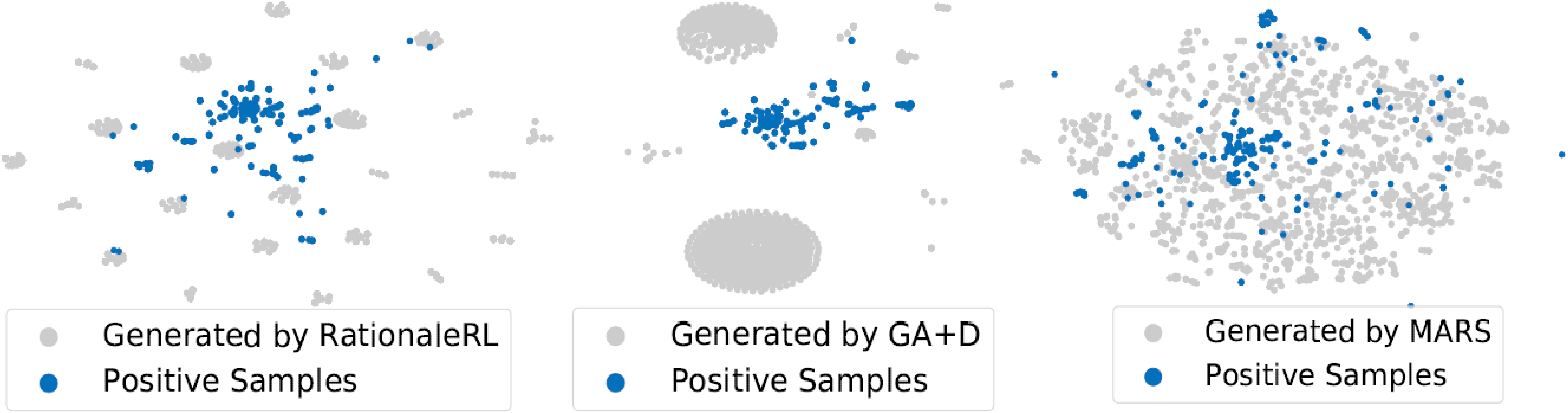
- Diversity (Div):

  Measures the diversity of generated molecules,

- PM:

  Product of the above three metrics

Results



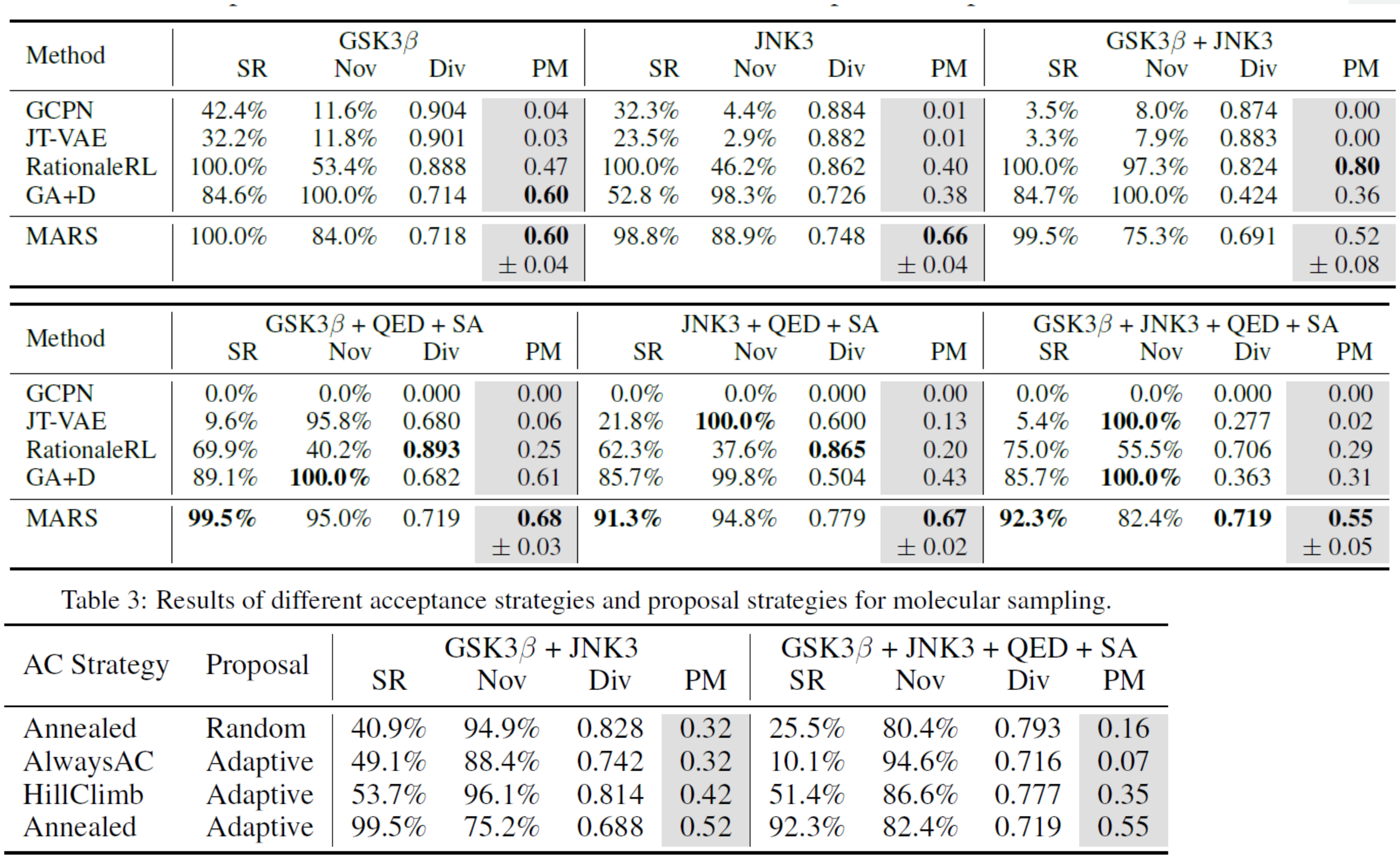
(a) RationaleRL

(b) GA+D

(c) MARS

Figure 3: t-SNE visualization of generated molecules (gray) and positive molecules in the training set (blue).

Results, cont’d



Method	GSK3β				JNK3				GSK3β + JNK3			
	SR	Nov	Div	PM	SR	Nov	Div	PM	SR	Nov	Div	PM
GCPN	42.4%	11.6%	0.904	0.04	32.3%	4.4%	0.884	0.01	3.5%	8.0%	0.874	0.00
JT-VAE	32.2%	11.8%	0.901	0.03	23.5%	2.9%	0.882	0.01	3.3%	7.9%	0.883	0.00
RationaleRL	100.0%	53.4%	0.888	0.47	100.0%	46.2%	0.862	0.40	100.0%	97.3%	0.824	**0.80**
GA+D	84.6%	100.0%	0.714	**0.60**	52.8 %	98.3%	0.726	0.38	84.7%	100.0%	0.424	0.36
MARS	100.0%	84.0%	0.718	**0.60** ± 0.04	98.8%	88.9%	0.748	**0.66** ± 0.04	99.5%	75.3%	0.691	0.52 ± 0.08

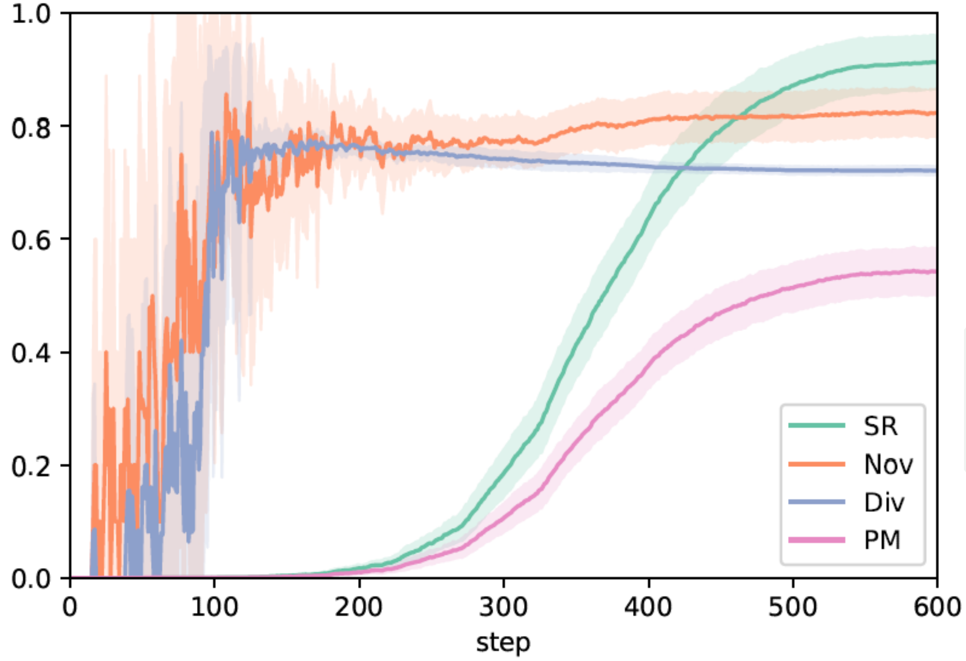
Method	GSK3β + QED + SA				JNK3 + QED + SA				GSK3β + JNK3 + QED + SA			
	SR	Nov	Div	PM	SR	Nov	Div	PM	SR	Nov	Div	PM
GCPN	0.0%	0.0%	0.000	0.00	0.0%	0.0%	0.000	0.00	0.0%	0.0%	0.000	0.00
JT-VAE	9.6%	95.8%	0.680	0.06	21.8%	**100.0%**	0.600	0.13	5.4%	**100.0%**	0.277	0.02
RationaleRL	69.9%	40.2%	**0.893**	0.25	62.3%	37.6%	**0.865**	0.20	75.0%	55.5%	0.706	0.29
GA+D	89.1%	**100.0%**	0.682	0.61	85.7%	99.8%	0.504	0.43	85.7%	**100.0%**	0.363	0.31
MARS	**99.5%**	95.0%	0.719	**0.68** ± 0.03	**91.3%**	94.8%	0.779	**0.67** ± 0.02	**92.3%**	82.4%	**0.719**	**0.55** ± 0.05

Table 3: Results of different acceptance strategies and proposal strategies for molecular sampling.

AC Strategy	Proposal	GSK3β + JNK3				GSK3β + JNK3 + QED + SA			
		SR	Nov	Div	PM	SR	Nov	Div	PM
Annealed	Random	40.9%	94.9%	0.828	0.32	25.5%	80.4%	0.793	0.16
AlwaysAC	Adaptive	49.1%	88.4%	0.742	0.32	10.1%	94.6%	0.716	0.07
HillClimb	Adaptive	53.7%	96.1%	0.814	0.42	51.4%	86.6%	0.777	0.35
Annealed	Adaptive	99.5%	75.2%	0.688	0.52	92.3%	82.4%	0.719	0.55

Conclusion

- MARS includes a trainable proposal to modify chemical graph fragments, which is parameterized by an MPNN.
- Our experiments verify that MARS outperforms prior approaches on five out of six molecule generation tasks.
- It is capable of finding novel and diverse bioactive molecules that are both drug-like and highly synthesizable.


step
SR
Nov
Div
PM