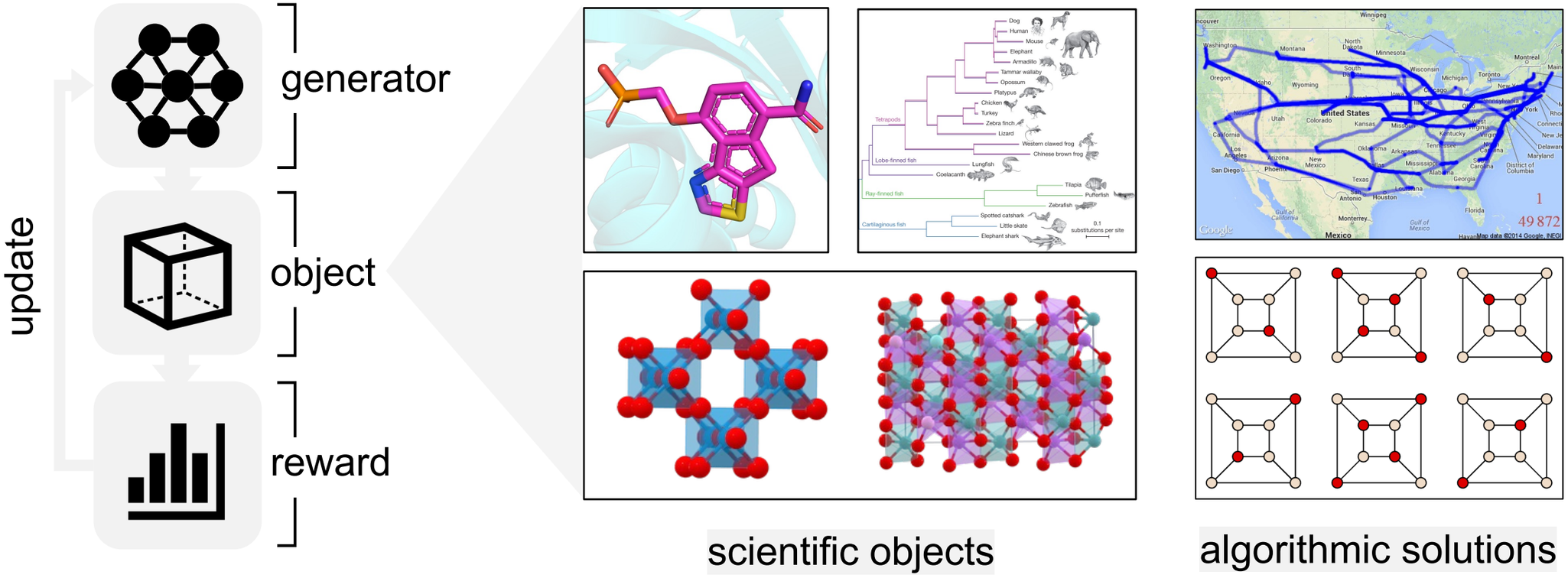

# Pessimistic Backward Policy for GFlowNets

Hyosoon Jang[1] Yunhui Jang[1] Minsu Kim[2] Jinkyoo Park[2] Sungsoo Ahn[1]

[1]POSTECH [2]KAIST


# Function-driven Learning

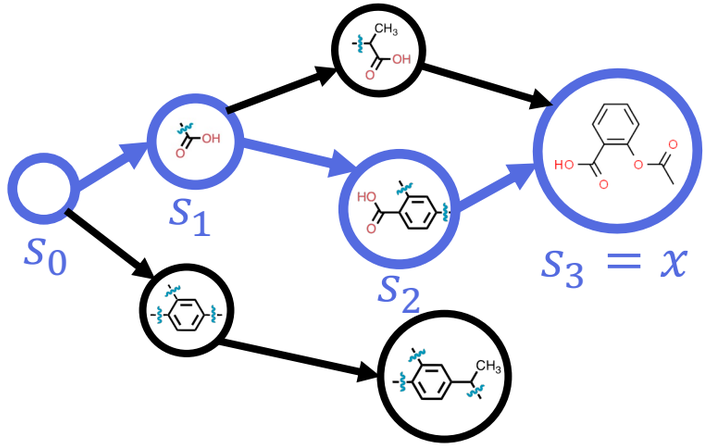
- Given a reward function, we aim to obtain diverse high-reward objects



# Generative Flow Networks (GFlowNets)

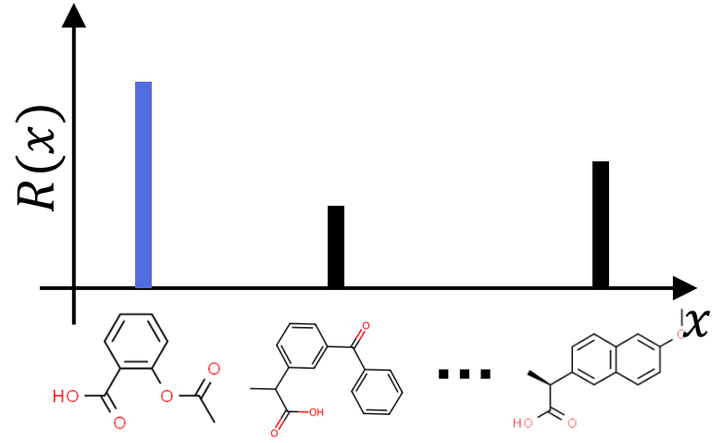
- GFlowNets construct objects with a forward policy $P_F$:

$$x \sim \prod_t P_F(s_{t+1}|s_t)$$




- **compositionality in the generation**:
  inducing an object $x$ with a trajectory of states $\tau = (s_0, \ldots, s_T = x)$

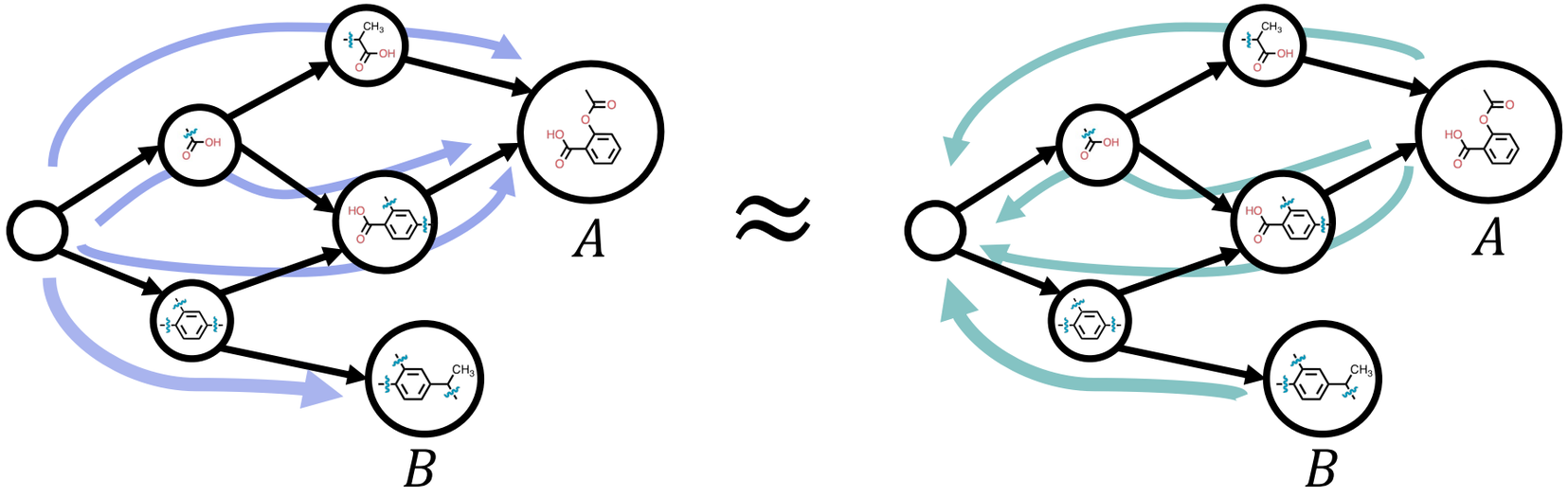
$$P_F(x) \propto R(x)$$




- **diversity in the generation**:
  sampling with a probability proportional to the reward $R(x)$

# Training of GFlowNets

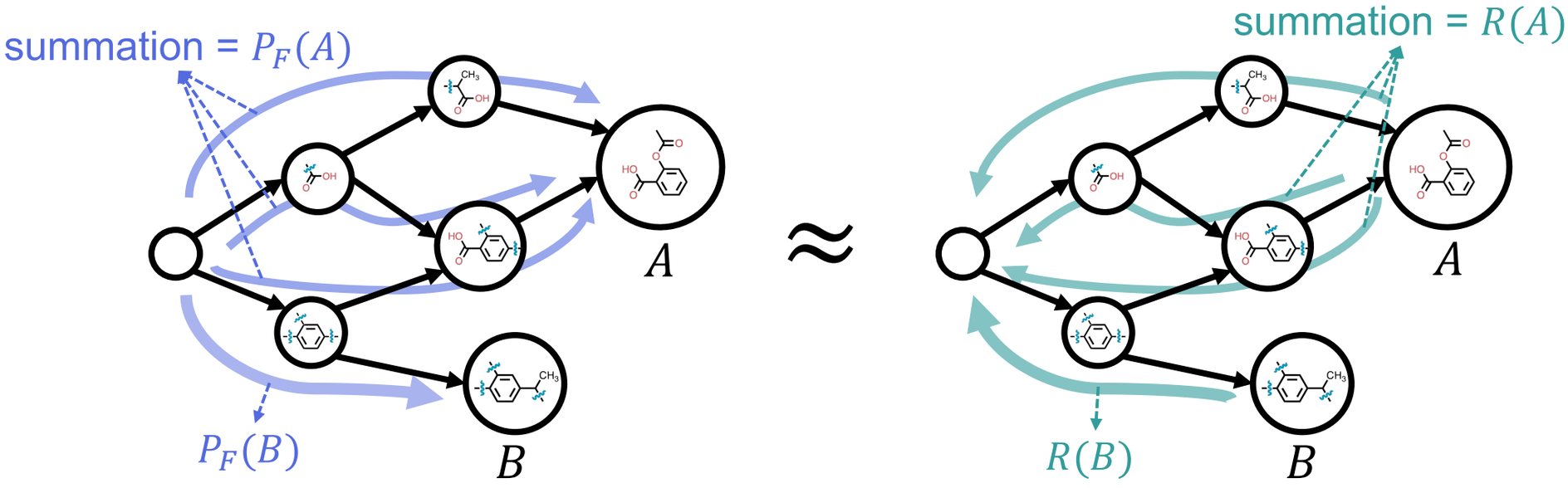
- GFlowNets align the forward policy with a backward policy over trajectories
  - similar to aligning diffusion with reverse diffusion or encoder with decoder




$$\Rightarrow \quad P_F(x) \propto R(x)$$

# Training of GFlowNets

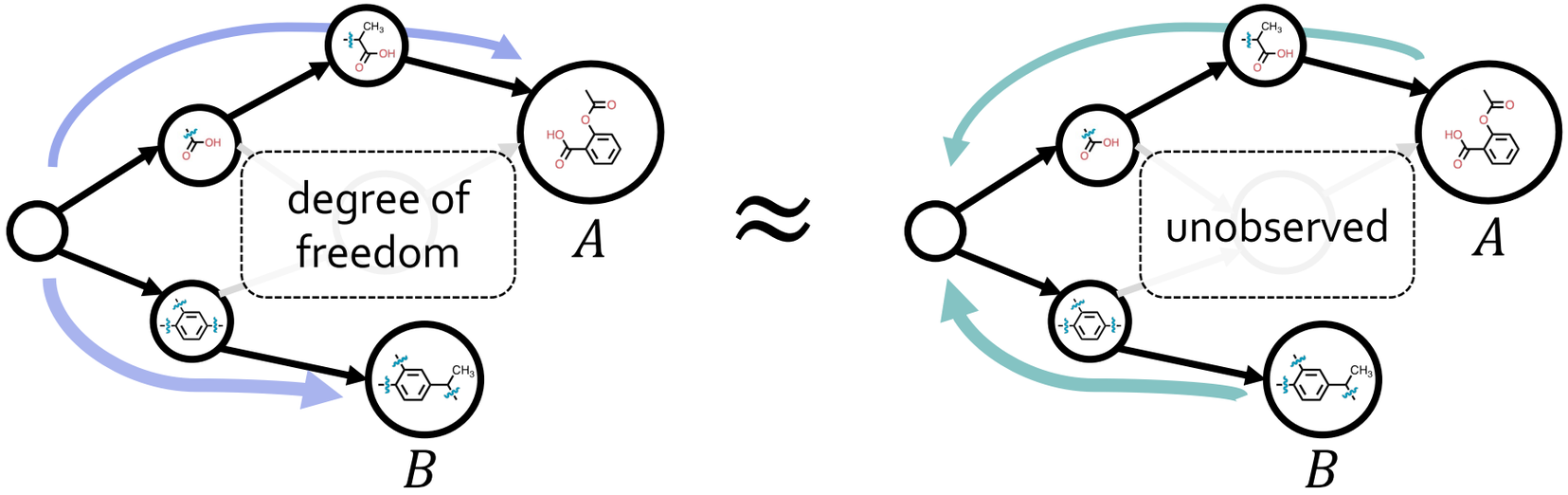
- The forward policy captures the target densities over trajectories
  that decompose the object's reward through the backward policy




$$\Rightarrow \quad R(A) > R(B) \quad \text{and} \quad P_F(A) > P_F(B)$$

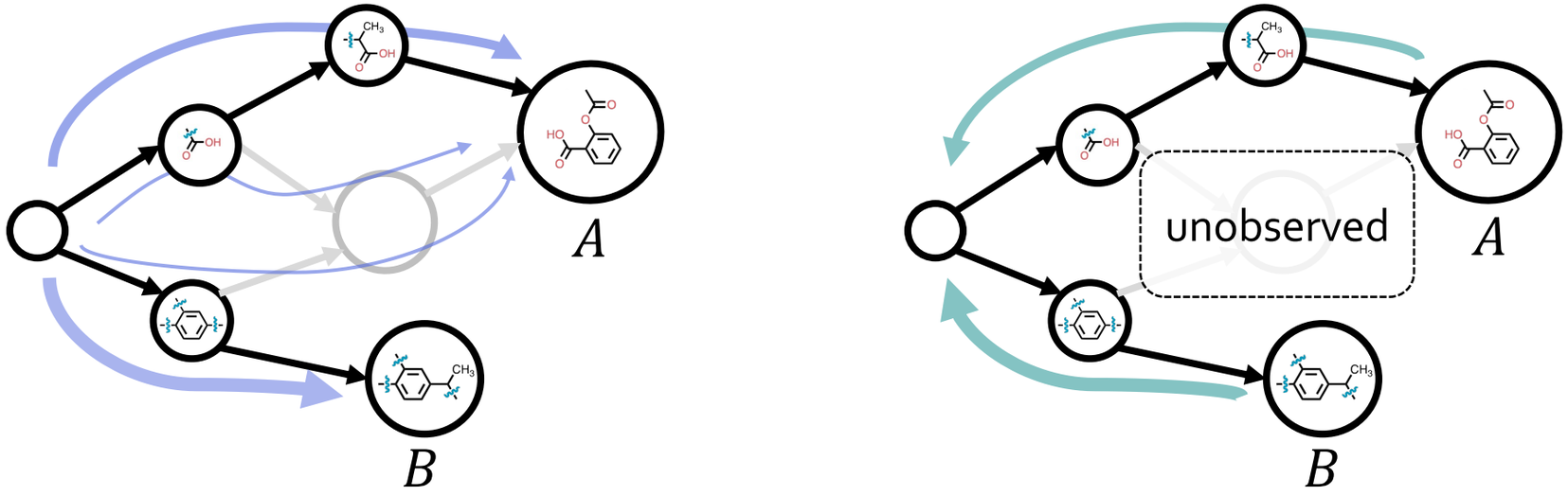
# Motivation: Under-determined Probability

- Unobserved trajectories under-determine the sampling probability
  - a degree of freedom in the forward probability over unobserved trajectories



# Motivation: Under-determined Probability

- This can cause the forward policy to favor low-reward objects
  - due to the degree of freedom lower than unobserved target densities

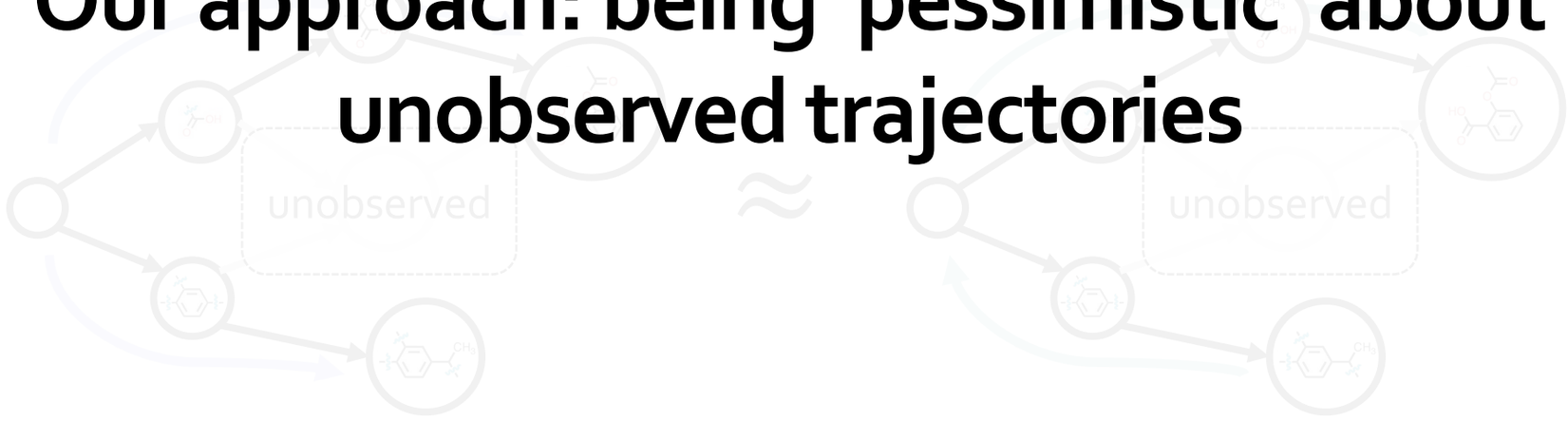



$$\Rightarrow \quad R(A) > R(B) \quad \text{but} \quad P_F(A) < P_F(B)$$

# Motivation: Under-determined Probability

- Unobserved trajectories **under-determine** the sampling probability
  - even make GFlowNets to favor low-reward objects

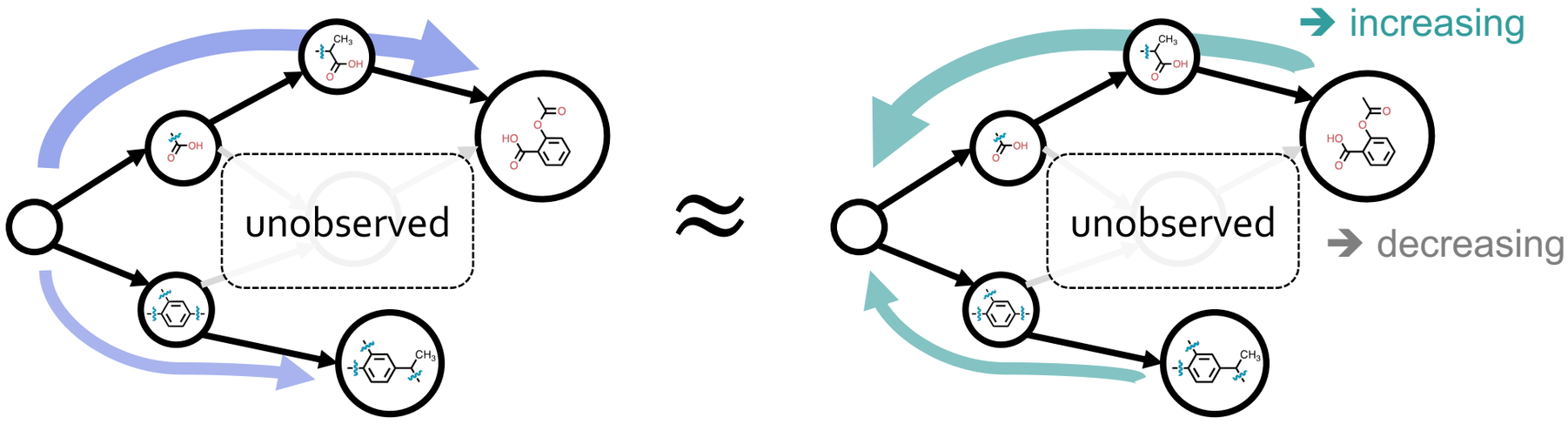
## Our approach: being 'pessimistic' about unobserved trajectories




$\Rightarrow$ $P(x) \overset{?}{\propto} R(x)$

# Our: Pessimistic Backward Policy

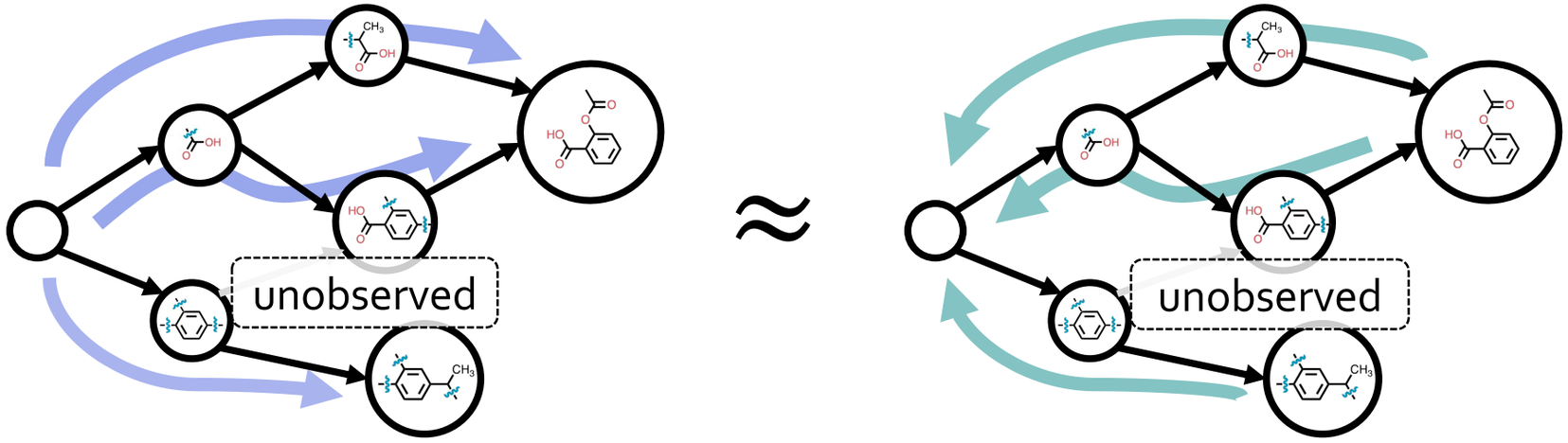
- For an observed object, our backward policy minimizes and increases the unobserved and observed target densities, respectively.




$$\Rightarrow \quad R(A) > R(B) \quad \text{and} \quad P_F(A) > P_F(B)$$

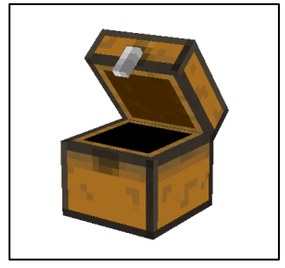
# Our: Pessimistic Backward Policy

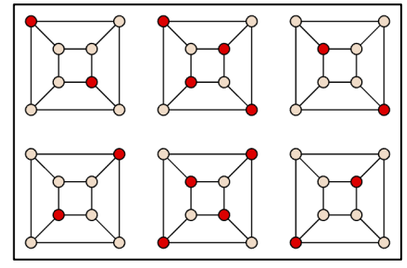
- In each training round, we adapt our backward policy for the observed trajectories

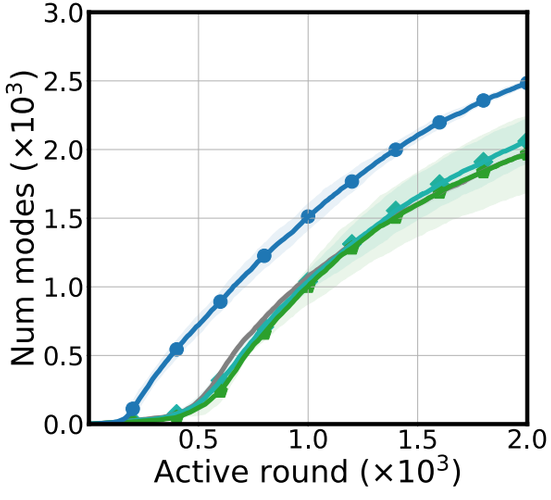


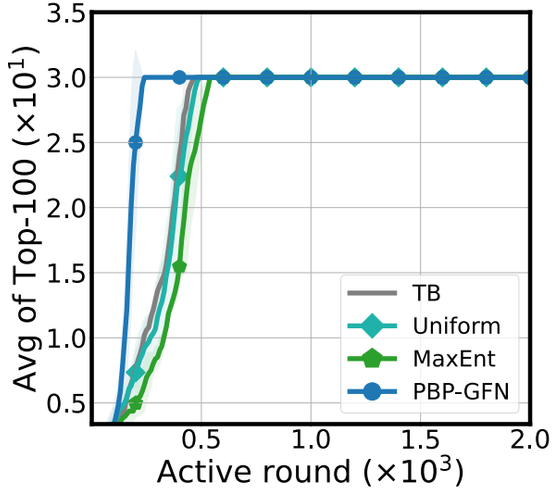

$$\Rightarrow \quad R(A) > R(B) \quad \text{and} \quad P_F(A) > P_F(B)$$

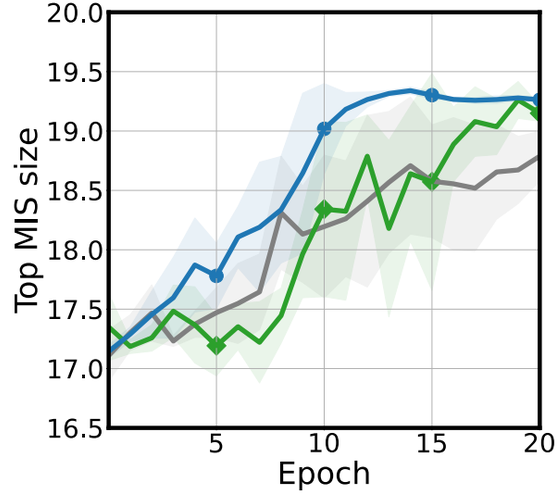
# Experiments: Bag and Maximum Independent Set

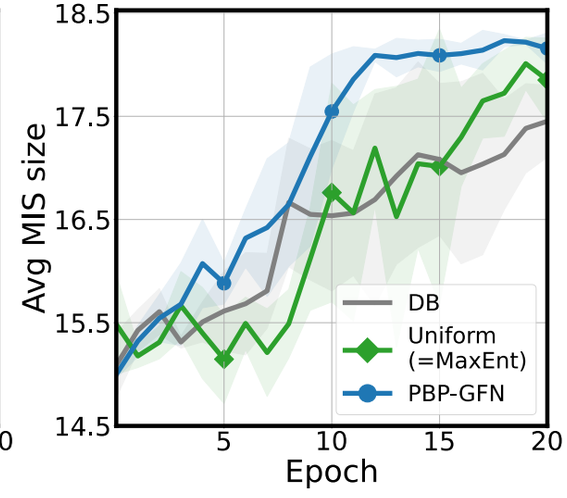




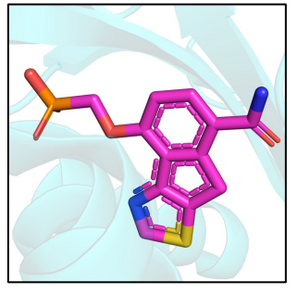


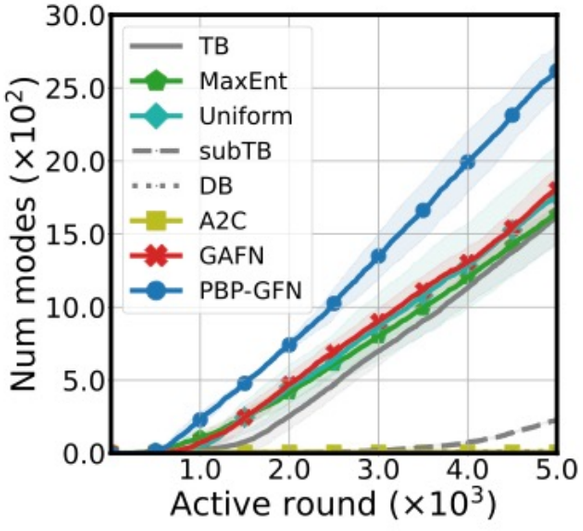



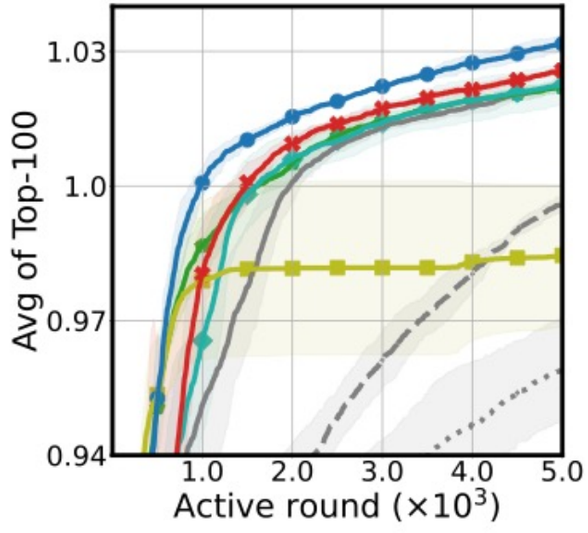



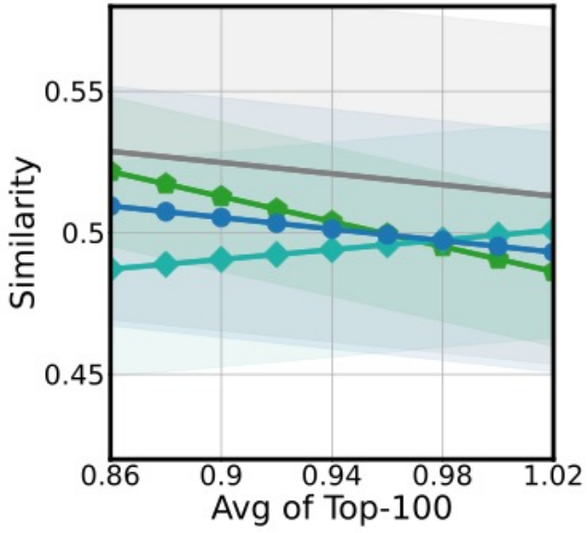



# Experiments: Molecule Generation





# Experiments: RNA sequence