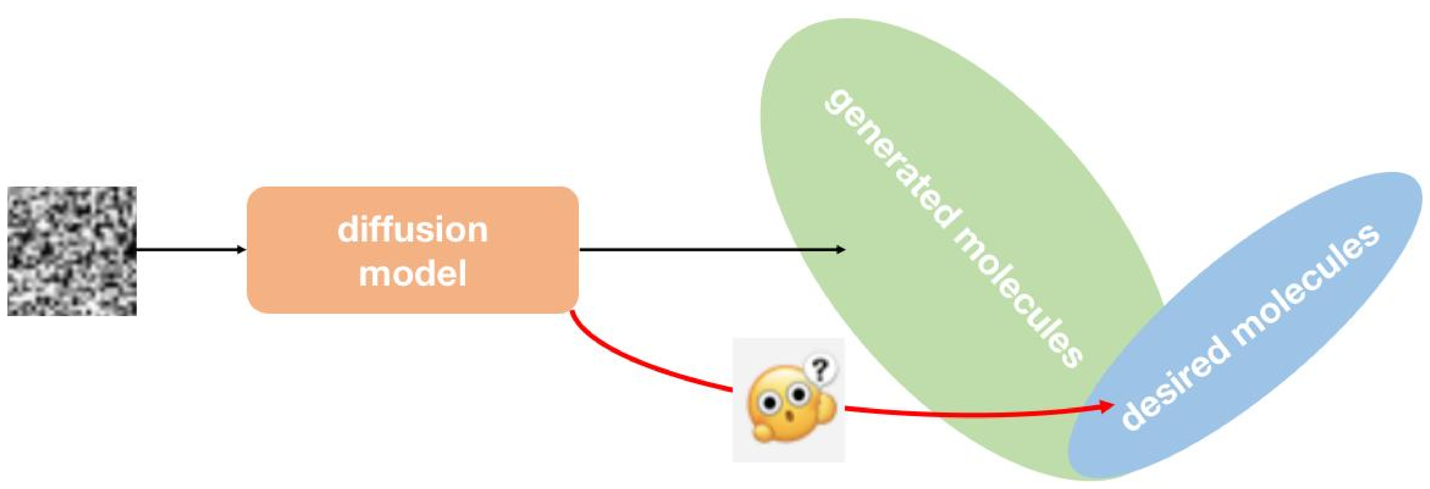
A PROPERTY-GUIDED DIFFUSION MODEL FOR GENERATING MOLECULAR GRAPHS

Changsheng Ma, Taicheng Guo†, Qiang Yang*, Xiuying Chen*, Xin Gao*, Shangsong Liang*, Nitesh Chawla†, Xiangliang Zhang†*

* King Abdullah University of Science and Technology, Jeddah, Saudi Arabia

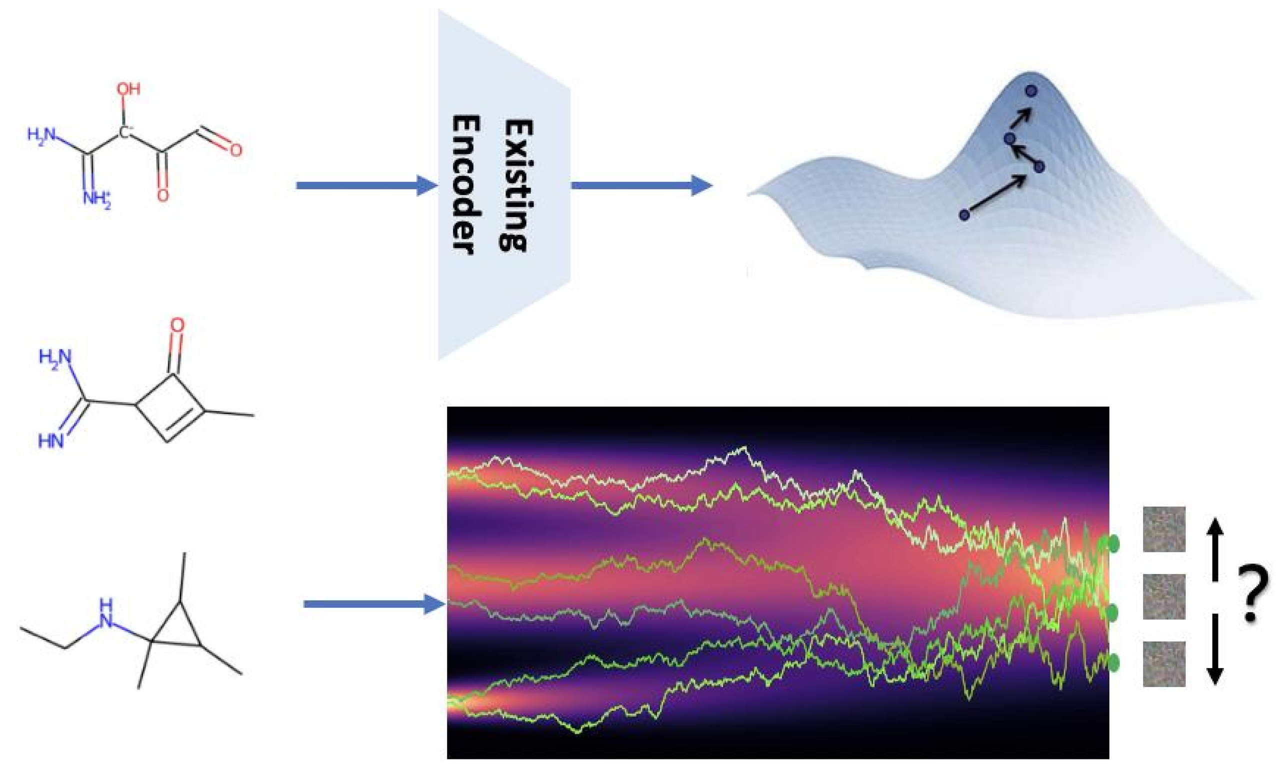
* MBZUAI, UAE † University of Notre Dame, IN, US

Objectives



Generate molecules exhibiting specific characteristics (properties) while maintaining optimal generative efficacy

Issues



The latent space of existing generators (upper) possesses semantics, allowing for Bayesian optimization to find desired molecules, whereas the diffusion model cannot, because its latent space (lower) has no semantics.

Methods

Theoretical Foundation

initial reverse process

$$p_\theta\left(G_{t-1} \mid G_t\right) = \mathcal{N}\left(\mu_{\theta_t}, \sigma_{\theta_t}^2 \mathbf{I}\right).$$

inject desired properties y

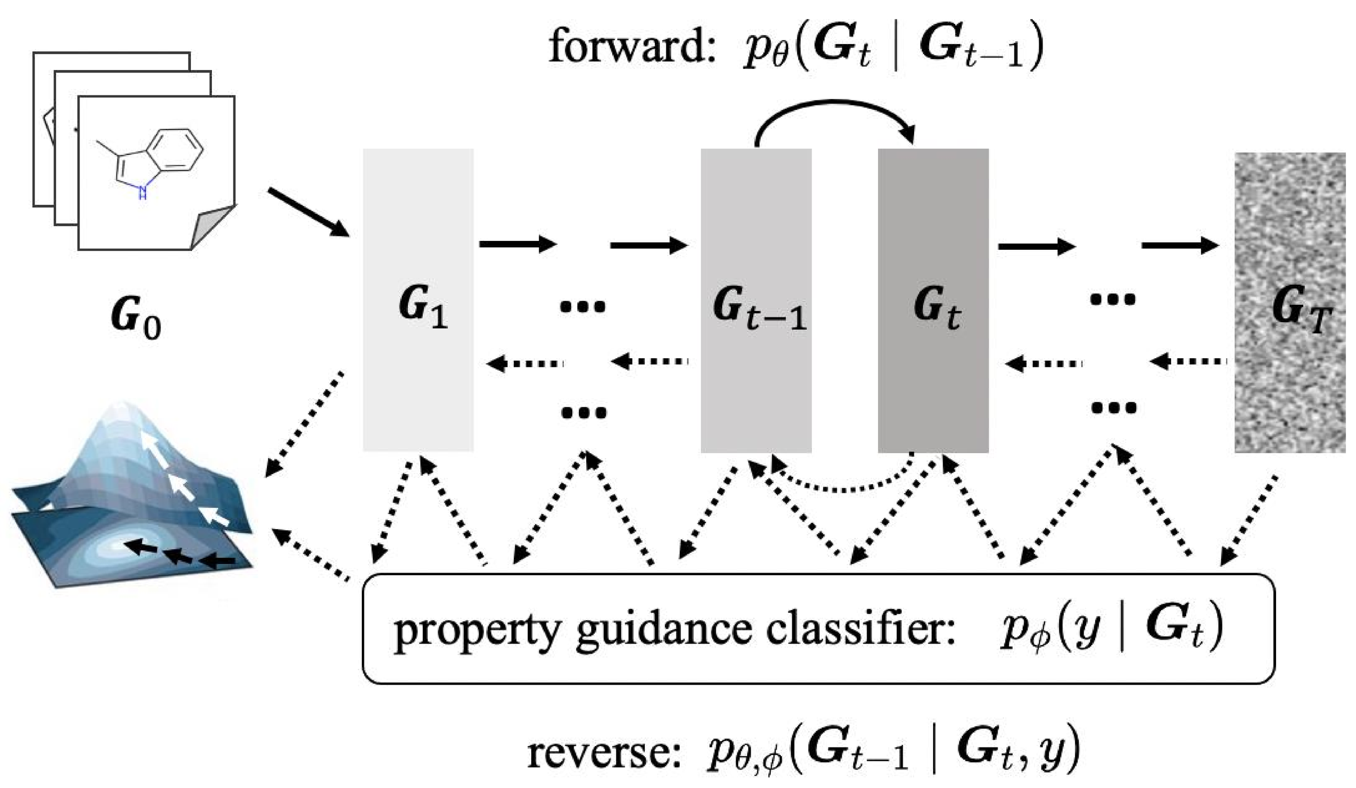
$$p_{\theta,\phi}\left(G_{t-1} \mid G_t, y\right) = Z p_\theta\left(G_{t-1} \mid G_t\right) p_\phi\left(y \mid G_{t-1}\right)$$

post-injection reverse process [1]

$$p_\theta\left(G_{t-1} \mid G_t\right) p_\phi\left(y \mid G_{t-1}\right) = \mathcal{N}(\mu_{\theta_t} + \sigma_{\theta_t}^2 \mathbf{I} g, \sigma_{\theta_t}^2)$$

$$g = \nabla_{G_{t-1}} \log p_\phi\left(y \mid G_{t-1}\right).$$

Model Architecture



Results

Generation Performance

Table 1: Generation performance on QM9. Results are the means and standard deviations of three independent runs.

Method	% VwoC ↑	NSPDK ↓	FCD ↓	% Validity ↑	% Uniqueness ↑	% Novelty ↑	% V.U.N ↑
GraphAF	67	0.020 ± 0.003	5.268 ± 0.403	**100.00**	94.51	88.83	83.95
GraphDF	82.67	0.063 ± 0.001	10.816 ± 0.020	**100.00**	97.62	**98.10**	95.77
MoFlow	91.36 ± 1.23	0.017 ± 0.003	4.467 ± 0.595	$\mathbf{100.00 \pm 0.00}$	98.65 ± 0.57	94.72 ± 0.77	93.44 ± 0.44
EDP-GNN	47.52 ± 3.60	0.005 ± 0.001	2.680 ± 0.221	$\mathbf{100.00 \pm 0.00}$	$\mathbf{99.25 \pm 0.05}$	86.58 ± 1.85	85.93 ± 0.09
GraphEBM	8.22 ± 2.24	0.030 ± 0.004	6.143 ± 0.411	$\mathbf{100.00 \pm 0.00}$	97.90 ± 0.14	97.01 ± 0.17	94.97 ± 0.02
GDSS	95.72 ± 1.94	0.003 ± 0.000	2.900 ± 0.282	$\mathbf{100.00 \pm 0.00}$	98.46 ± 0.61	86.27 ± 2.29	84.94 ± 1.40
Ours	$\mathbf{96.98 \pm 1.23}$	$\mathbf{0.002 \pm 0.000}$	$\mathbf{2.204 \pm 0.065}$	$\mathbf{100.00 \pm 0.00}$	98.52 ± 0.15	97.23 ± 1.05	$\mathbf{95.79 \pm 0.16}$

Single/Multi-Property-Guided Performance

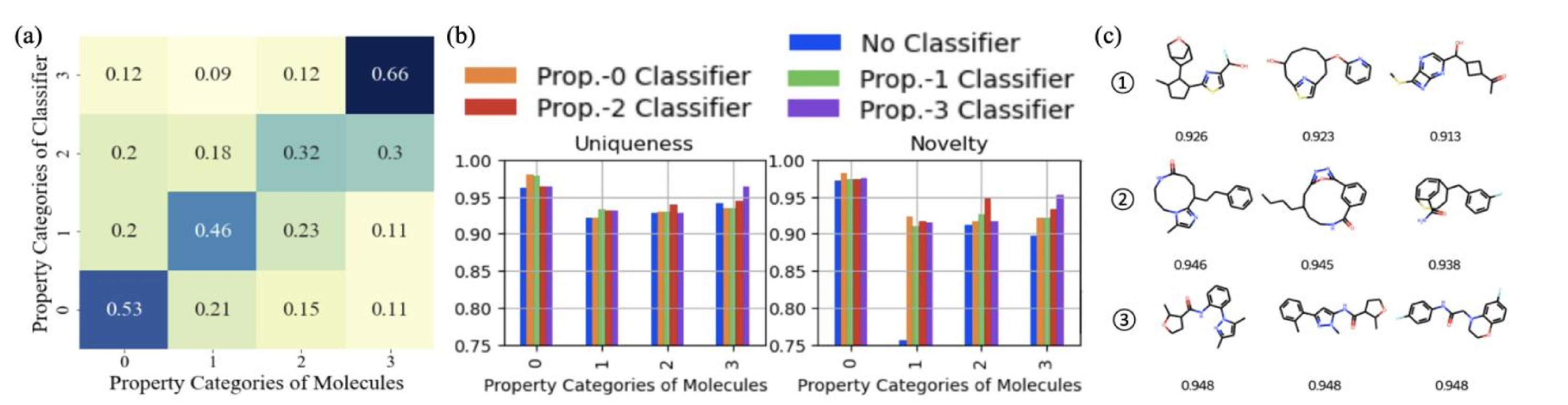


Fig. 3: Property-Guided performance. (a): The proportion of molecules in each category relative to all molecules generated under the guidance of different classifiers. (b): The uniqueness and novelty value of generated molecules within each individual class under the guidance of different classifiers. (c): Top 3 molecules with QED values out of 10,000 randomly generated molecules by different models: ① Without guidance, ② QED-guided, ③ QED & Ring-guided.

Conclusions

Acknowledging the constraints of diffusion-based models in generating molecules with specific properties, we addressed the issue by:

1. Integrating a time-dependent classifier to guide the sampling process toward desired properties.

2. Extending to multi-property-guided molecular generation, enabling the concurrent satisfaction of multiple properties.

[1] Dhariwal et al. "Diffusion models beat gans on image synthesis", NeurIPS (2021).