

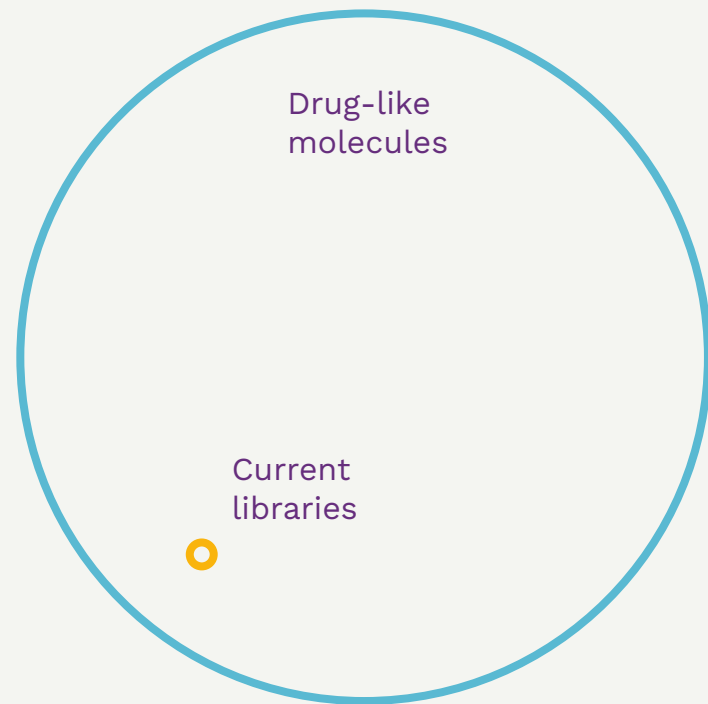
RGFN: Synthesizable Molecular Generation Using GFlowNets

Michał Koziarski*, **Andrei Rekes***, Dmytro Shevchuk*, Almer van der Sloot, Piotr Gaiński, Yoshua Bengio, Cheng-Hao Liu, Mike Tyers, Robert A. Batey

Why generative models?

The size of chemical space of drug-like molecules is often estimated at 10^{60} .

Generative models offer a promise of being able to explore that vast space **without** being limited to existing libraries.



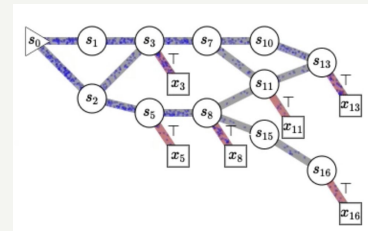
Why GFlowNets?

Generative Flow Networks (GFlowNets) are a relatively new family of generative models.

Goal: generating **high reward**, **diverse** samples in an **amortized** manner. All crucial in drug discovery!

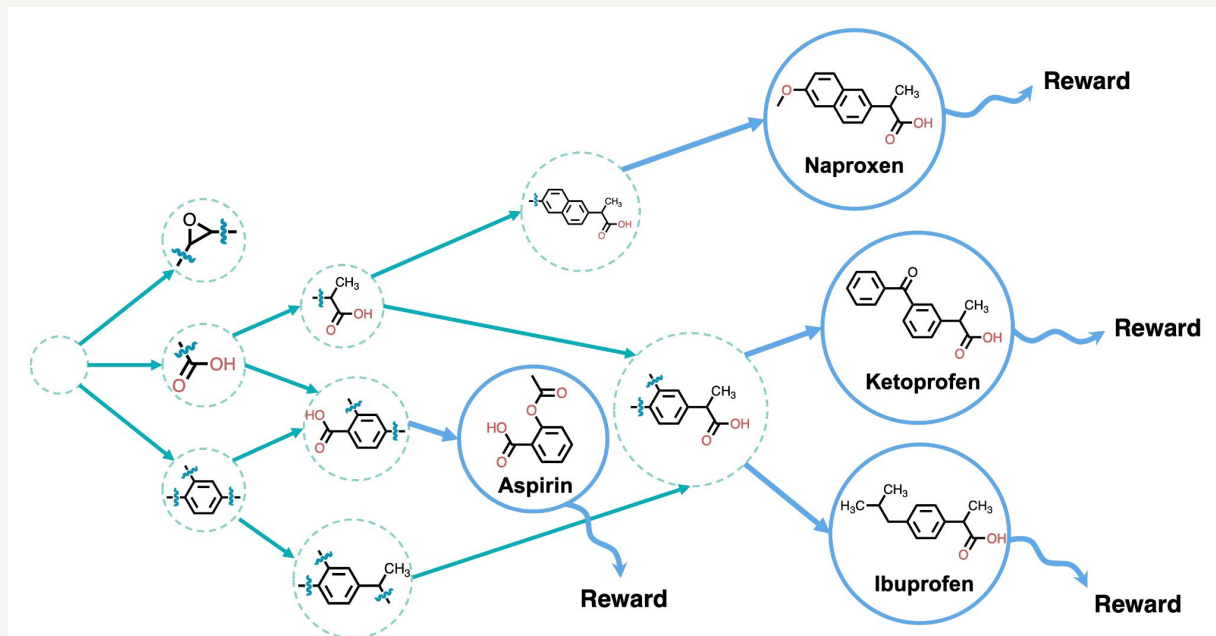
Shortcomings of the existing methods:

- MCMC - lack of amortization,
- RL - mean-seeking behaviour; mode collapse



How to do it? On high level: ensure that the probability of generating a sample is proportional to its reward: $p(\mathbf{x}) \sim R(\mathbf{x})$. This can be done by training a sampling policy $\pi(\mathbf{x})$ (a machine learning model).

GFlowNets for Molecule Design



Key ingredients of GFlowNets:

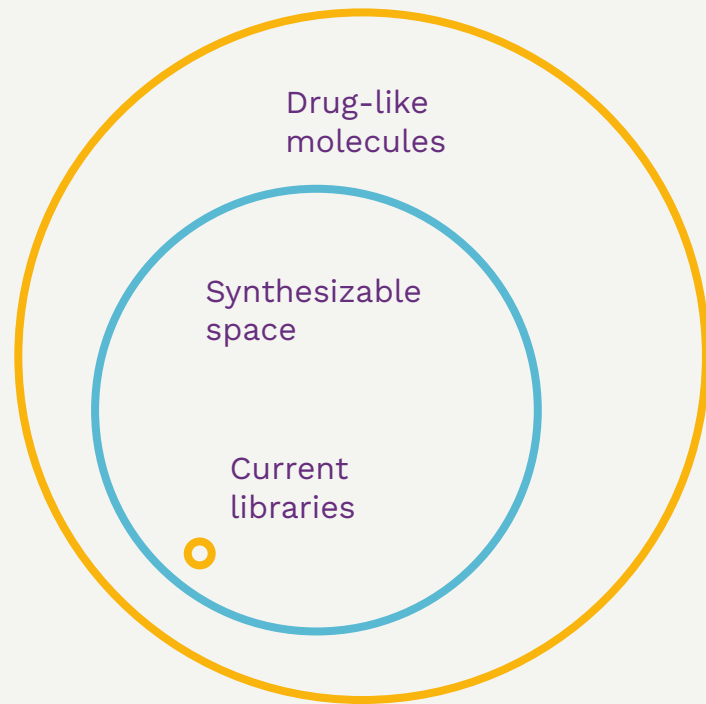
- **State** = current molecule
- **Action space** = fragments to add
- **Reward** function = property of interest

How do we ensure molecules are synthesizable?

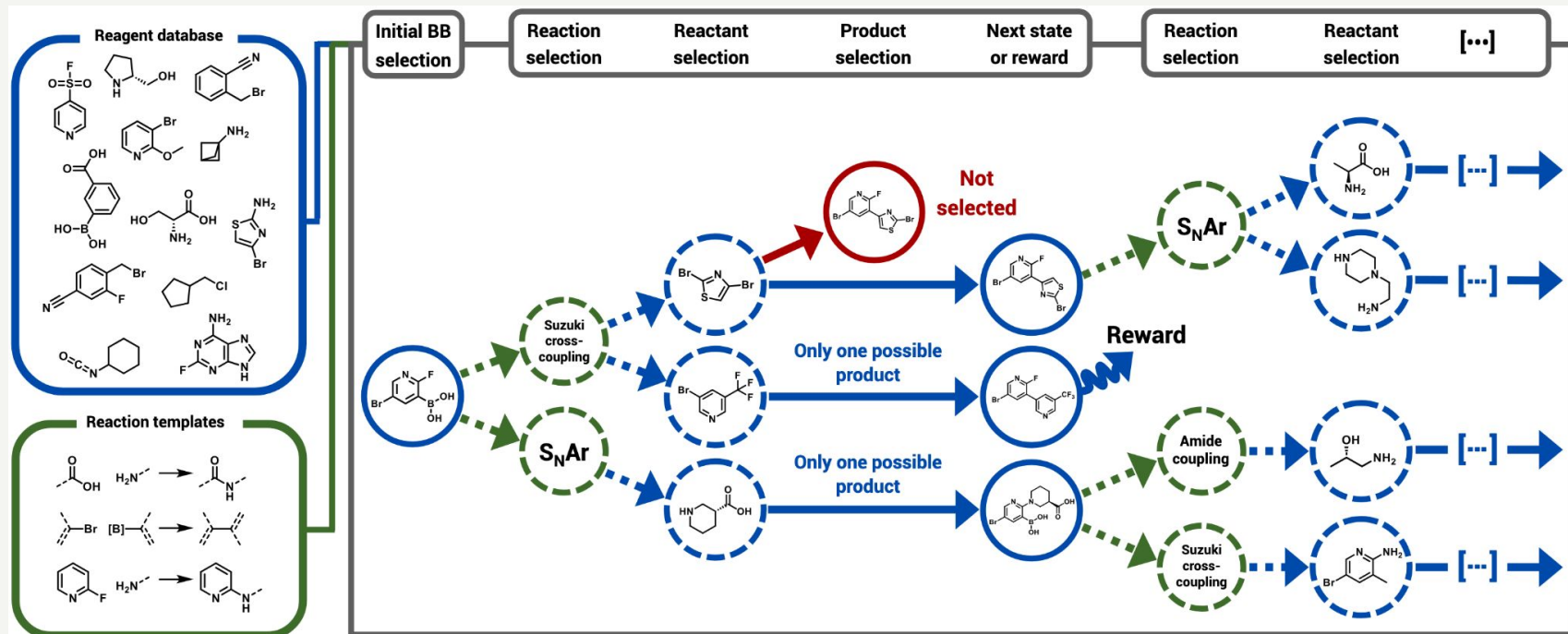
Constraint Trade-Off

The goal: constrain the searchable space to highly synthesizable compounds.

(while increasing the search space size as much as possible!)



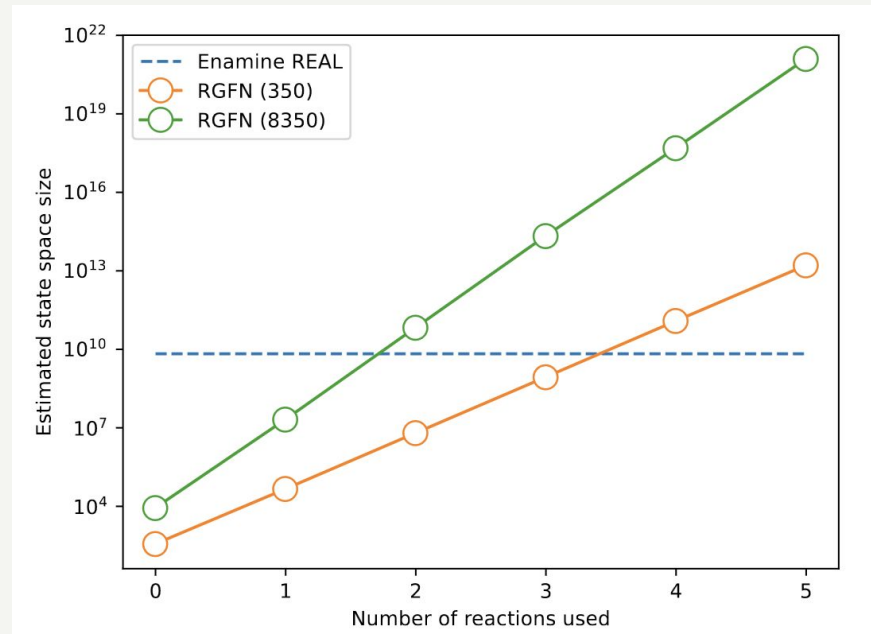
Reaction-GFlowNet



Chemical language

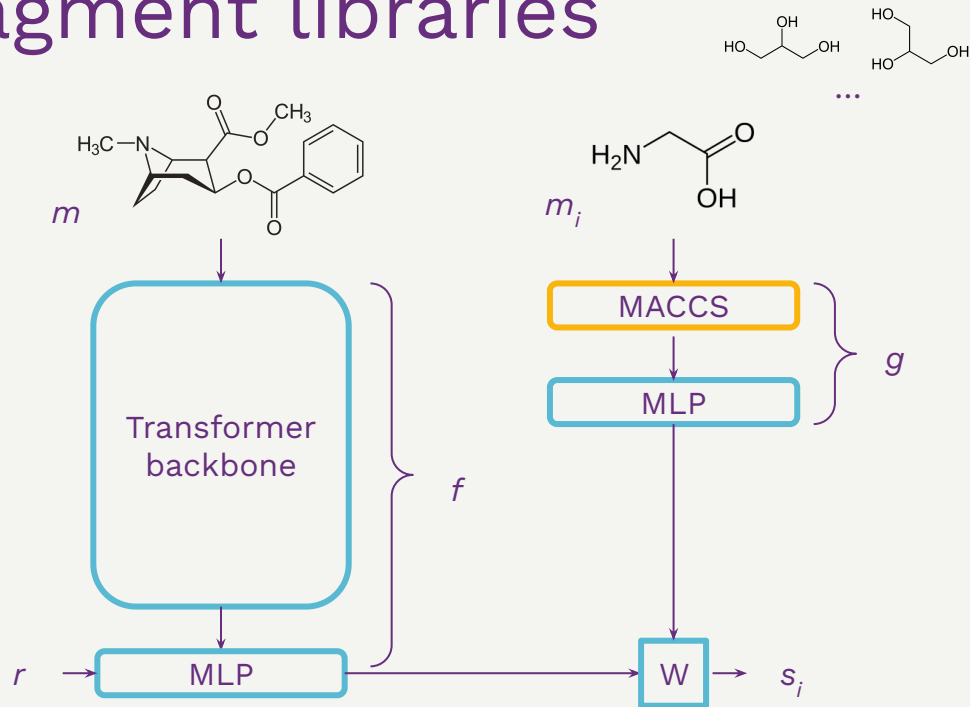
We select a total of **350** affordable reagents ($\leq \$200/\text{g}$) **and 17** high-yield reactions.

Combinatorial space generated by this approach with depth 4 is **larger than most compound libraries.**



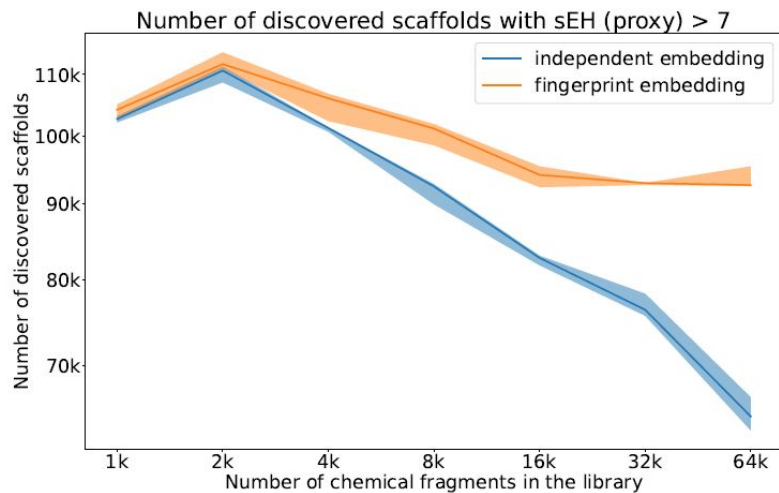
Scaling to large fragment libraries

We use **action embeddings** for chemistry-aware next fragment selection.

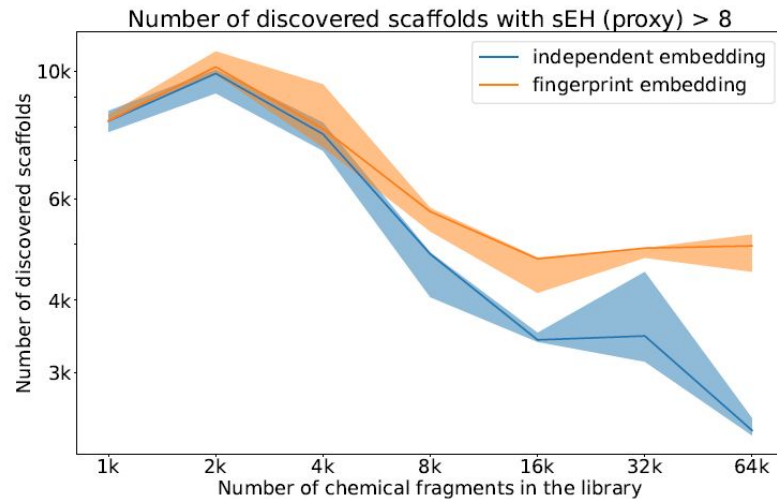


$$p(m_i|m, r) = \sigma^{|\mathcal{M}|}(\mathbf{s})_i, \quad s_i = \phi(W f(m, r))^T g(m_i),$$

Scaling to large fragment libraries



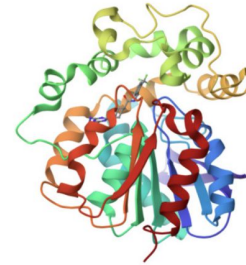
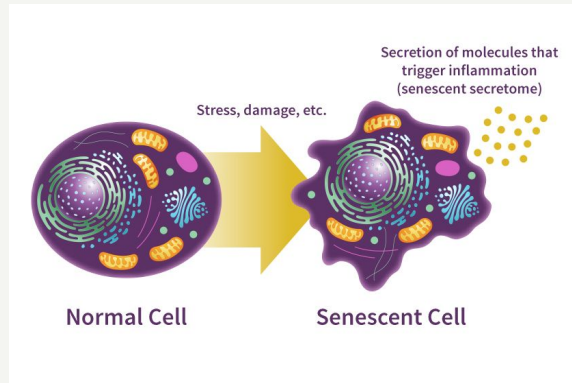
(a) sEH (proxy) > 7



(b) sEH (proxy) > 8

Experimental study

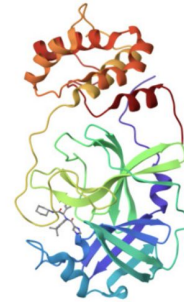
In addition to commonly used sEH_proxy, we consider GPU-accelerated docking for several biologically relevant targets, and challenging activity-based senolytic_proxy.



sEH (PDB ID: 4JNC)



ClpP (PDB ID: 7UVU)

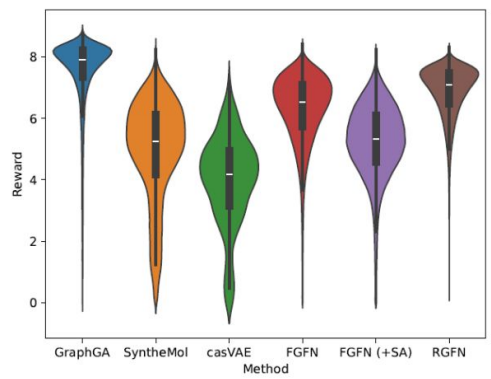


Mpro (PDB ID: 6W63)

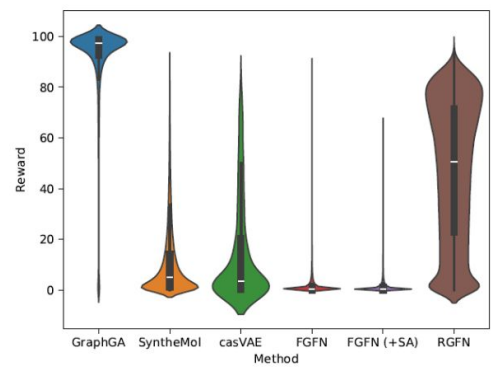


TBLR1 (PDB ID: 5NAF)

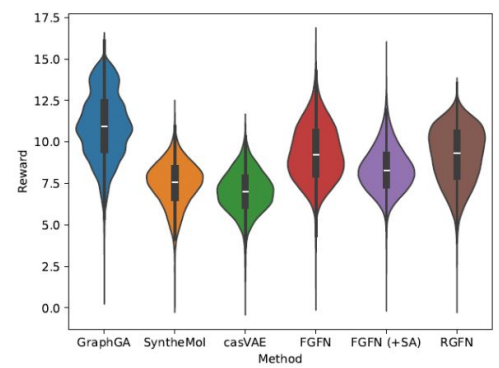
Reward distributions



(a) sEH (proxy)

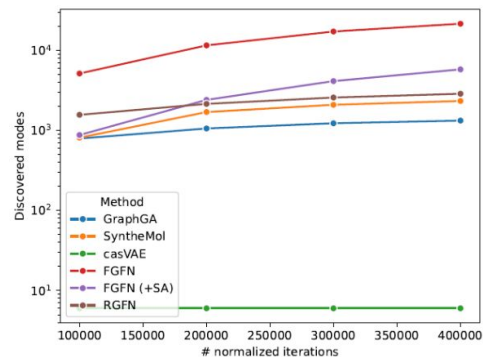


(b) senolytics (proxy)

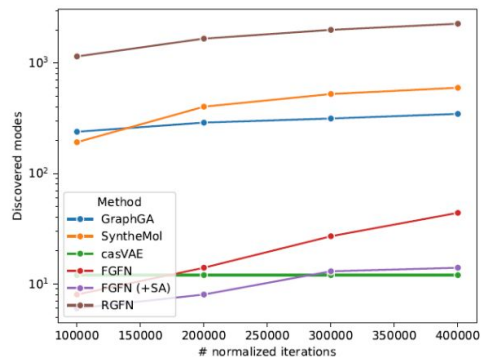


(c) ClpP (docking)

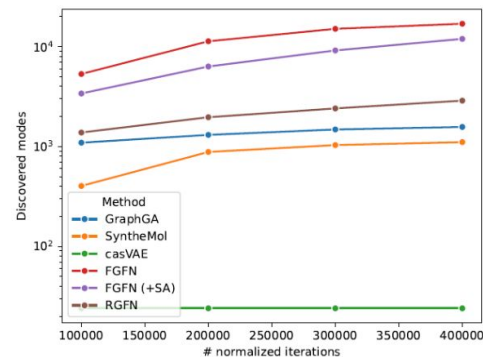
Discovered modes



(a) sEH (proxy)



(b) senolytics (proxy)



(c) ClpP (docking)

Synthesizability

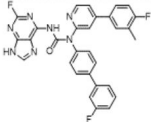
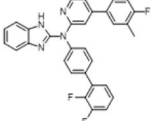
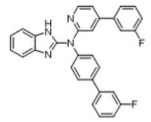
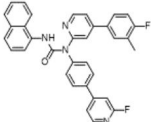
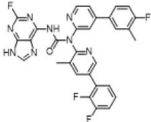
Table 1: Average values of synthesizability-related metrics for top-k modes.

Task	Method	Mol. weight ↓	QED ↑	SAScore ↓	AiZynth ↑
sEH	GraphGA	528.6 ± 42.3	0.21 ± 0.06	3.87 ± 0.24	0.04
	SyntheMol	411.1 ± 66.7	0.57 ± 0.18	<u>2.85 ± 0.55</u>	<u>0.80</u>
	casVAE	<u>421.6 ± 103.4</u>	<u>0.52 ± 0.23</u>	2.41 ± 0.47	0.82
	FGFN	473.4 ± 58.9	0.39 ± 0.13	3.43 ± 0.48	0.14
	FGFN+SA	473.7 ± 62.2	0.36 ± 0.12	3.01 ± 0.50	0.27
	RGFN	495.2 ± 49.6	0.29 ± 0.10	3.09 ± 0.39	0.56
Seno.	GraphGA	485.7 ± 75.6	0.09 ± 0.05	2.92 ± 0.26	0.05
	SyntheMol	<u>441.4 ± 83.5</u>	<u>0.48 ± 0.19</u>	2.77 ± 0.40	0.53
	casVAE	431.5 ± 100.9	0.50 ± 0.19	<u>2.82 ± 0.46</u>	0.65
	FGFN	468.9 ± 47.7	0.42 ± 0.13	3.55 ± 0.52	0.02
	FGFN+SA	451.8 ± 54.5	0.32 ± 0.12	2.83 ± 0.44	0.13
	RGFN	558.7 ± 62.8	0.21 ± 0.09	3.24 ± 0.32	<u>0.58</u>
ClpP	GraphGA	521.0 ± 31.8	0.32 ± 0.07	4.14 ± 0.51	0.00
	SyntheMol	<u>458.2 ± 60.7</u>	<u>0.45 ± 0.16</u>	2.86 ± 0.56	0.56
	casVAE	423.0 ± 61.7	0.47 ± 0.17	2.44 ± 0.41	0.84
	FGFN	548.6 ± 42.9	0.22 ± 0.03	2.94 ± 0.54	0.25
	FGFN+SA	509.2 ± 52.4	0.24 ± 0.04	<u>2.61 ± 0.49</u>	0.33
	RGFN	526.2 ± 37.6	0.23 ± 0.04	2.83 ± 0.22	<u>0.65</u>

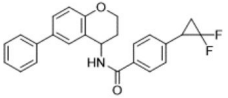
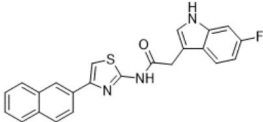
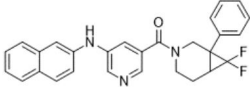
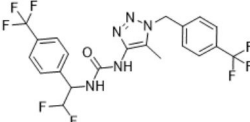
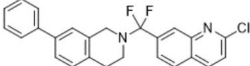
	High reward	Diverse	Synthesizable
GraphGA	✓	✗	✗
SyntheMol	✗	✗	✓
CasVAE	✗	✗	✓
FGFN	✓	✓	✗
FGFN + SA	✓	✓	✗
RGFN	✓	✓	✓

Estimated synthesis cost

RGFN

Position	Ligand	Cost per 0,1 mmol, \$
1		2.07
2		1.90
3		1.76
4		1.37
5		1.80

SyntheMol

Position	Ligand	Cost per 0,1 mmol, \$
1		185.79
2		17.68
3		260.33
4		233.44
5		N/A

Summary

- RGFN guarantees synthesizability out-of-the-box.
- Chemical language used leads to **~100x lower cost of synthesis.**
- Performs well and generates diverse candidates for a set of biologically relevant oracles.
- **Flexible to the choice of fragments and reward functions.**

Links

Paper:



Code:



Join us:



Thank you!