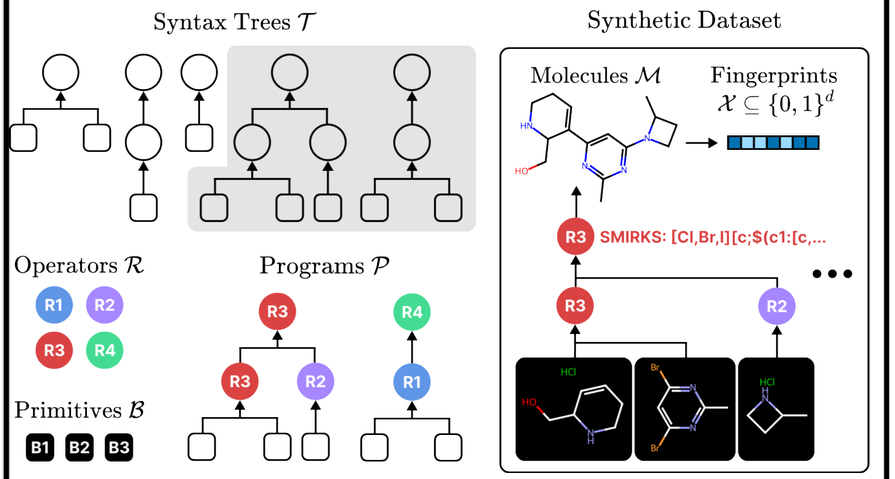
Procedural Synthesis of Synthesizable Molecules

Michael Sun[1], Alston Lo[1], Minghao Guo[1], Jie Chen[3], Connor Coley[2], Wojciech Matusik[1]

[1]MIT CSAIL, [2]MIT Chemical Engineering, [3]MIT-IBM Watson AI Lab

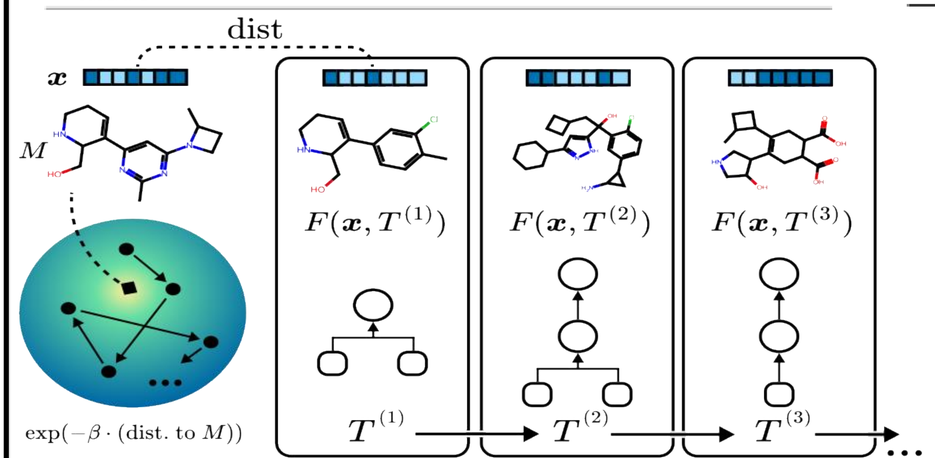
Preliminaries



Algorithm 1 Construction of training dataset.

Require: A synthetic dataset $\mathcal{D}_0 \subseteq \mathcal{P} \times \mathcal{B}^*$ of programs (Section 4.1.1).

$\mathcal{D} \leftarrow \varnothing$
for each $(P, B) \in \mathcal{D}_0$ **do**
  Turn (P, B) into a fully-filled program $T \in \partial\mathcal{P}$ whose root is attributed with FP(P, B).
  for each $\Lambda \in 2^T$ containing the root and closed under parent$(\cdot)$ **do**
    Frontier$(\Lambda) \leftarrow \{i \in T \mid i \notin \Lambda$ and parent$(i) \in \Lambda\}$
    Populate node features $\boldsymbol{H}$ and labels $\boldsymbol{Y}$ based on P and B (App. F)
    for $i \in T - \Lambda$ **do**
      Mask the feature in $\boldsymbol{H}$ corresponding to node i
    for $i \in T -$ Frontier(Λ) **do**
      Mask the label in $\boldsymbol{Y}$ corresponding to node i
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(T, \boldsymbol{H}, \boldsymbol{Y})\}$
return $\mathcal{D}$



Our Metropolis-Hastings algorithm iteratively refines the syntax tree skeleton towards the stationary distribution which is proportional to the inverse distance to our target molecule M.

Method

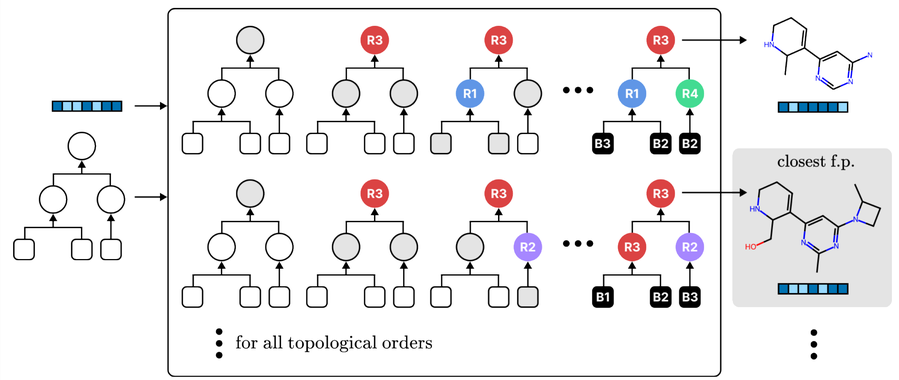
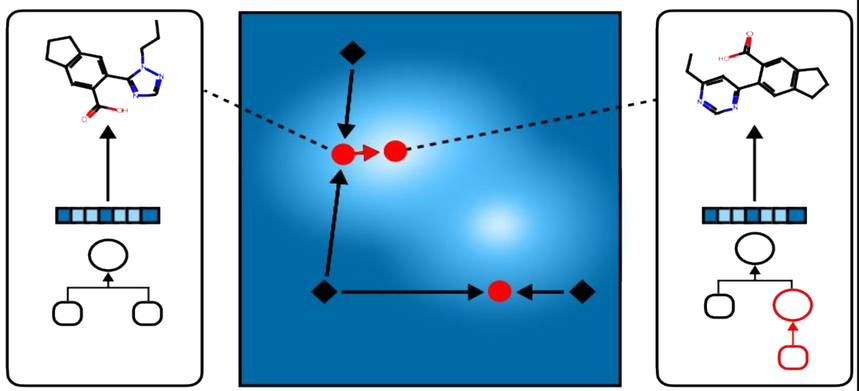


Illustration of our decoding scheme F: (Left) The input is a Morgan fingerprint x and syntax skeleton T; (Middle) Decode once for every topological ordering of the tree, tracking all partial programs with a stack; (Right) Execute all decoded programs, then returning the closest analog which minimizes distance to x.

# Rxns	$\lvert\hat{\mathcal{T}}_{k\setminus k-1}\rvert$
6	298
7	243
8	112
9	63
10	42
11	22
12	11
13	4
14	2

# Rxns	$\lvert\hat{\mathcal{T}}_{k\setminus k-1}\rvert$ ($\lvert\mathcal{T}_{k\setminus k-1}\rvert$)	# Topo. Orders (Max, Mean, Std)$_{\hat{\mathcal{T}}_{k\setminus k-1}}$
1	2 (2)	2, 1.5, 0.5
2	6 (6)	8, 4.17, 2.79
3	22 (22)	80, 19.59, 20.55
4	83 (90)	896, 152.02, 215.53
5	209 (394)	19200, 2506.25, 3705.77

# Rxns	# Topo. Masks ()$_{\hat{\mathcal{T}}_{k\setminus k-1}}$	# Topo. Masks ()$_{\mathcal{T}_{k\setminus k-1}}$
1	5, 4, 1	5, 4, 1
2	11, 7.67, 2.56	11, 7.67, 2.56
3	26, 14.36, 5.86	26, 14.36, 5.86
4	56, 27.99, 12.47	56, 26.73, 12.78
5	131, 65.07, 26.36	131, 49.74, 27.09
6	287, 165.12, 61.43	287, 92.67, 56.29

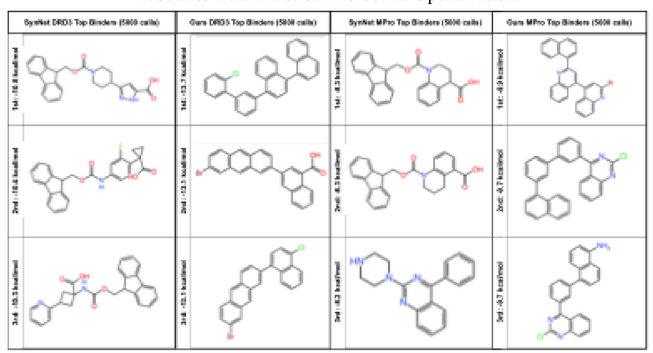


Our genetic algorithm over the joint design space combines the strategies of semantic crossover (→) and syntactical mutation (→) to encourage both global improvement and local exploration.

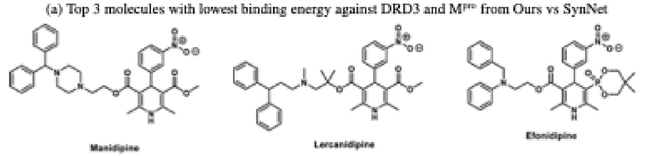
Results

Category	Method	Score ↑ Value	Score ↑ Rank	AUC ↑ Value	AUC ↑ Rank	SA ↓ Value	SA ↓ Rank
Screening	Screening	0.426	20	0.377	20	3.097	8
	MolPAL	0.472	16	0.444	15	3.018	4
String	REINVENT	0.697	2	0.607	2	3.415	9
	REINVENT-SELFIES	0.682	3	0.578	4	3.791	15
	STONED	0.609	8	0.555	6	5.550	24
	smiles_lstm_hc	0.667	5	0.544	8	3.036	5
	smiles_ga	0.548	11	0.503	10	5.422	23
	selfies_lstm_hc	0.539	12	0.431	16	3.743	14
	selfies_vae_bo	0.428	19	0.383	19	3.522	11
	smiles_vae_bo	0.422	21	0.376	21	3.084	7
	selfies_ga	0.48	15	0.337	23	5.709	26
	pasithea	0.338	24	0.326	24	3.66	12
Graph	Graph-GA	**0.701**	1	0.601	3	3.982	17
	GPBO	0.642	6	0.570	5	3.954	16
	DST	0.555	10	0.479	11	4.146	20
	mars	0.507	14	0.47	12	4.232	21
	mimosa	0.538	13	0.463	13	4.3	22
	gflownet	0.461	17	0.419	17	4.05	19
	gflownet_al	0.417	22	0.387	18	4.005	18
	jt_vae_bo	0.388	23	0.371	22	3.5	10
	graph_mcts	0.317	25	0.28	25	3.732	13
	moldqn	0.213	26	0.187	26	5.604	25
Synthesis	SynNet	0.578	9	0.545	7	3.075	6
	DoG-Gen	0.634	7	0.511	9	2.793	2
	DoG-AE	0.460	18	0.450	14	2.857	3
	Ours	0.670	4	**0.608**	1	**2.739**	1

Avg. across GSK3β, JNK3, DRD2, Median1, Median2, Rediscovery, Osimertinib, 6 others
Baselines from Practical Molecular Optimization.



(a) Top 3 molecules with lowest binding energy against DRD3 and M^{Pro} from Ours vs SynNet



(b) Top binders against M^{Pro} from literature, based on consensus docking scores (Ghahremanpour et al., 2020)