Molecule Generation with Fragment Retrieval Augmentation

Seul Lee[1*], Karsten Kreis[2], Srimukh Prasad Veccham[2], Meng Liu[2],
Danny Reidenbach[2], Saee Paliwal[2], Arash Vahdat[2†], Weili Nie[2†]

[1] KAIST, [2] NVIDIA
[*] Work done during an internship at NVIDIA
[†] Equal advising

Motivation

- **Fragment-based drug discovery (FBDD)** has been considered as an effective approach to explore the chemical space.

Motivation

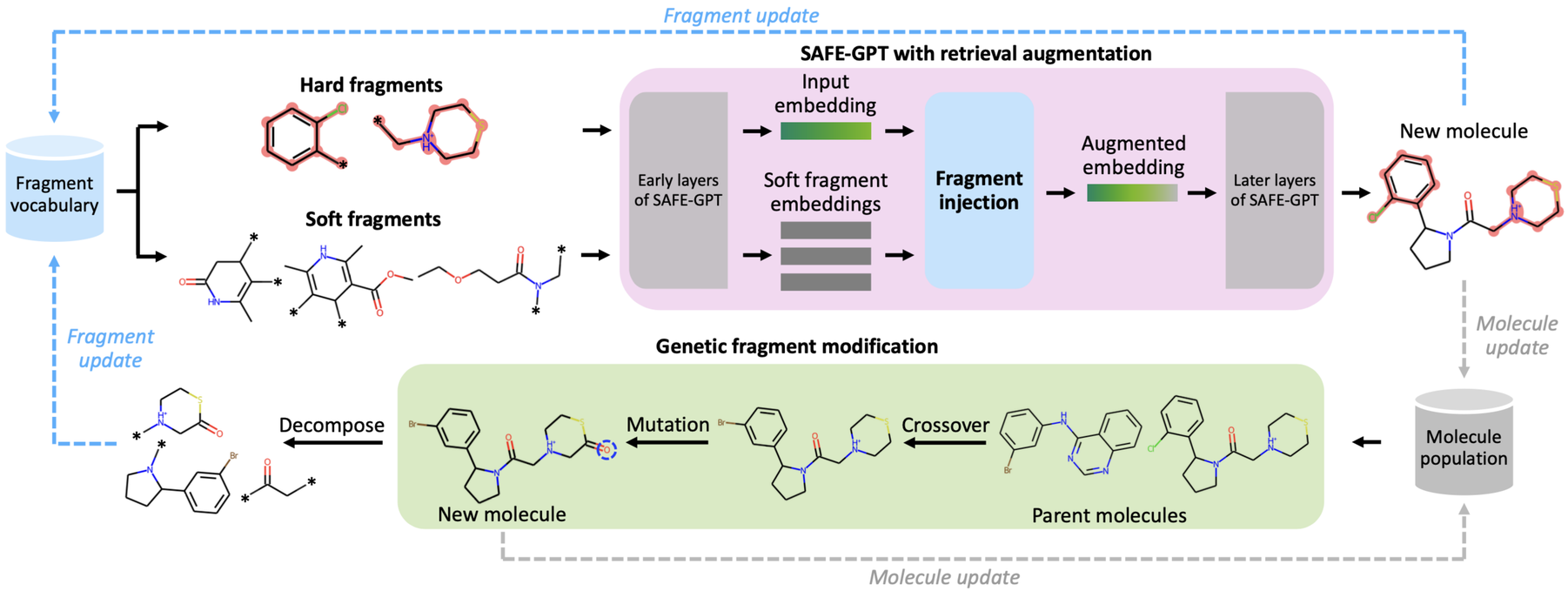
- **Fragment-based drug discovery (FBDD)** has been considered as an effective approach to explore the chemical space.
- Generative models have been adopted in the field of FBDD to accelerate the process.
  - Many fragment-based molecule generation methods show limited exploration as they only reassemble or slightly modify the given fragments.

Motivation

- **Fragment-based drug discovery (FBDD)** has been considered as an effective approach to explore the chemical space.
- Generative models have been adopted in the field of FBDD to accelerate the process.
  - Many fragment-based molecule generation methods show limited exploration as they only reassemble or slightly modify the given fragments.
- FBDD + RAG → **Fragment Retrieval-Augmented Generation (f-RAG)**.
  - f-RAG augments the pre-trained molecular language model SAFE-GPT with two types of retrieved fragments: **hard fragments and soft fragments**.

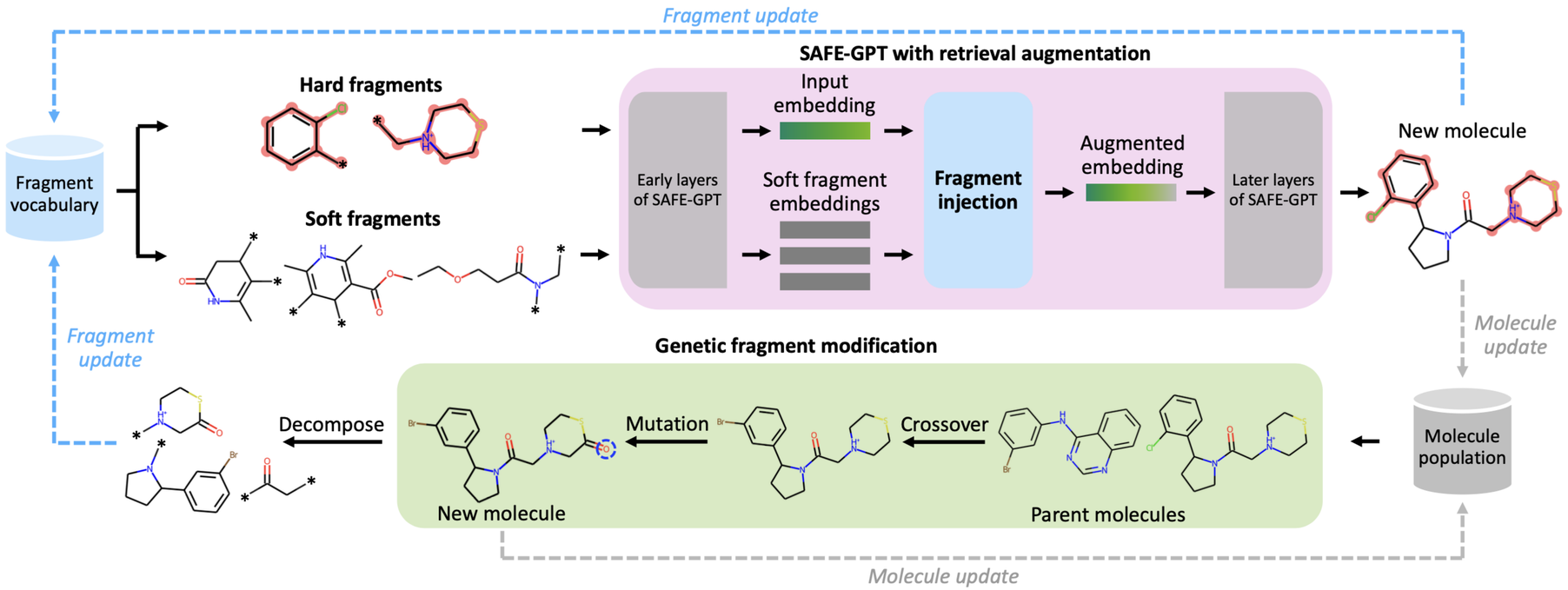
Methodology

- Construct a fragment vocabulary.
  - Decompose known molecules from the existing library into fragments and scoring the fragments.



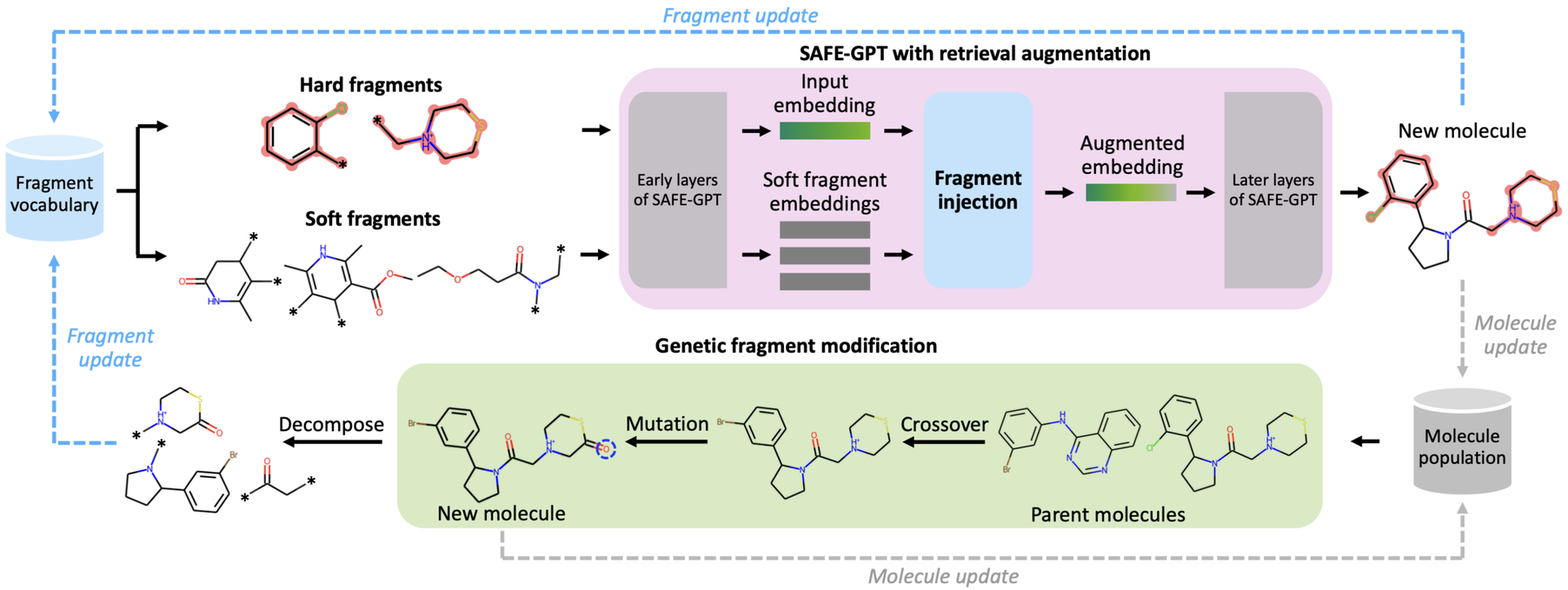
Methodology

- Construct a fragment vocabulary.
  - Decompose known molecules from the existing library into fragments and scoring the fragments.
- f-RAG retrieves fragments that will be explicitly included in the new molecule (i.e., **hard fragments**).
  - Hard fragments serve as the input context to the molecular language model that predicts the remaining fragments.
- f-RAG retrieves fragments that will not be part of the generated molecule but provide guidance (i.e., **soft fragments**).
  - The soft fragment embeddings are fused with the hard fragment embeddings through a lightweight **fragment injection module** in the middle of SAFE-GPT.



Methodology

- f-RAG updates the fragment vocabulary with generated fragments via an iterative refinement process which is further enhanced with **post-hoc genetic fragment modification**.



Experiments: PMO Benchmark

- f-RAG outperformed the previous methods in the PMO goal-directed hit generation benchmark.
- f-RAG achieved improved trade-offs between optimization performance, diversity, novelty, and synthesizability.

Oracle	f-RAG (ours)	Genetic GFN	Mol GA	REINVENT	Graph GA
albuterol_similarity	**0.977** ± 0.002	0.949 ± 0.010	0.896 ± 0.035	0.882 ± 0.006	0.838 ± 0.016
amlodipine_mpo	0.749 ± 0.019	**0.761** ± 0.019	0.688 ± 0.039	0.635 ± 0.035	0.661 ± 0.020
celecoxib_rediscovery	0.778 ± 0.007	**0.802** ± 0.029	0.567 ± 0.083	0.713 ± 0.067	0.630 ± 0.097
deco_hop	**0.936** ± 0.011	0.733 ± 0.109	0.649 ± 0.025	0.666 ± 0.044	0.619 ± 0.004
drd2	**0.992** ± 0.000	0.974 ± 0.006	0.936 ± 0.016	0.945 ± 0.007	0.964 ± 0.012
fexofenadine_mpo	**0.856** ± 0.016	**0.856** ± 0.039	0.825 ± 0.019	0.784 ± 0.006	0.760 ± 0.011
gsk3b	**0.969** ± 0.003	0.881 ± 0.042	0.843 ± 0.039	0.865 ± 0.043	0.788 ± 0.070
isomers_c7h8n2o2	0.955 ± 0.008	**0.969** ± 0.003	0.878 ± 0.026	0.852 ± 0.036	0.862 ± 0.065
isomers_c9h10n2o2pf2cl	0.850 ± 0.005	**0.897** ± 0.007	0.865 ± 0.012	0.642 ± 0.054	0.719 ± 0.047
jnk3	**0.904** ± 0.004	0.764 ± 0.069	0.702 ± 0.123	0.783 ± 0.023	0.553 ± 0.136
median1	0.340 ± 0.007	**0.379** ± 0.010	0.257 ± 0.009	0.356 ± 0.009	0.294 ± 0.021
median2	**0.323** ± 0.005	0.294 ± 0.007	0.301 ± 0.021	0.276 ± 0.008	0.273 ± 0.009
mestranol_similarity	0.671 ± 0.021	**0.708** ± 0.057	0.591 ± 0.053	0.618 ± 0.048	0.579 ± 0.022
osimertinib_mpo	**0.866** ± 0.009	0.860 ± 0.008	0.844 ± 0.015	0.837 ± 0.009	0.831 ± 0.005
perindopril_mpo	**0.681** ± 0.017	0.595 ± 0.014	0.547 ± 0.022	0.537 ± 0.016	0.538 ± 0.009
qed	0.939 ± 0.001	**0.942** ± 0.000	0.941 ± 0.001	0.941 ± 0.000	0.940 ± 0.000
ranolazine_mpo	**0.820** ± 0.016	0.819 ± 0.018	0.804 ± 0.011	0.760 ± 0.009	0.728 ± 0.012
scaffold_hop	0.576 ± 0.014	**0.615** ± 0.100	0.527 ± 0.025	0.560 ± 0.019	0.517 ± 0.007
sitagliptin_mpo	0.601 ± 0.011	**0.634** ± 0.039	0.582 ± 0.040	0.021 ± 0.003	0.433 ± 0.075
thiothixene_rediscovery	**0.584** ± 0.009	0.583 ± 0.034	0.519 ± 0.041	0.534 ± 0.013	0.479 ± 0.025
troglitazone_rediscovery	0.448 ± 0.017	**0.511** ± 0.054	0.427 ± 0.031	0.441 ± 0.032	0.390 ± 0.016
valsartan_smarts	**0.627** ± 0.058	0.135 ± 0.271	0.000 ± 0.000	0.178 ± 0.358	0.000 ± 0.000
zaleplon_mpo	0.486 ± 0.004	**0.552** ± 0.033	0.519 ± 0.029	0.358 ± 0.062	0.346 ± 0.032
Sum	**16.928**	16.213	14.708	14.196	13.751

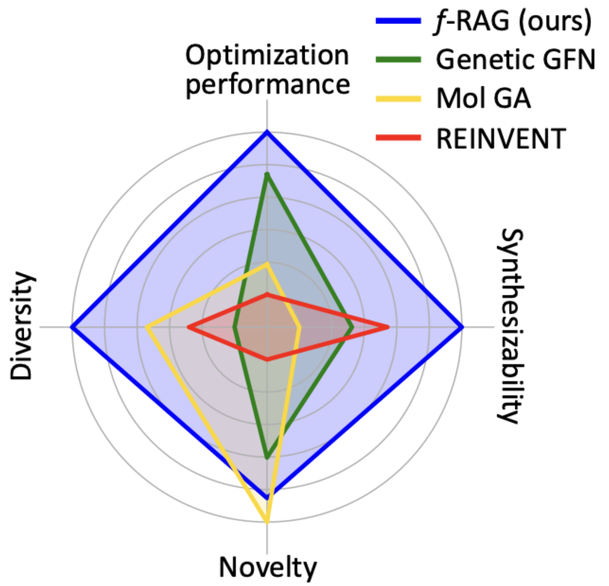


Figure 1: **A radar plot of target properties.** f-RAG strikes better balance among optimization performance, diversity, novelty, and synthesizability than the state-of-the-art techniques on the PMO benchmark [10].

Experiments: Constrained Docking Score Optimization

- f-RAG outperformed the previous methods in docking score (DS) optimization under QED, SA, and novelty constraints.
  - (the maximum similarity with the training molecules) < 0.4
  - DS < (the median DS of known active molecules)
  - QED > 0.5
  - SA < 5
- With the dynamic update, f-RAG can discover molecules that have higher DS than the top molecule in the training set.

Method	Target protein				
	parp1	fa7	5ht1b	braf	jak2
JT-VAE [16]	-9.482 ± 0.132	-7.683 ± 0.048	-9.382 ± 0.332	-9.079 ± 0.069	-8.885 ± 0.026
REINVENT [35]	-8.702 ± 0.523	-7.205 ± 0.264	-8.770 ± 0.316	-8.392 ± 0.400	-8.165 ± 0.277
Graph GA [14]	-10.949 ± 0.532	-7.365 ± 0.326	-10.422 ± 0.670	-10.789 ± 0.341	-10.167 ± 0.576
MORLD [15]	-7.532 ± 0.260	-6.263 ± 0.165	-7.869 ± 0.650	-8.040 ± 0.337	-7.816 ± 0.133
HierVAE [17]	-9.487 ± 0.278	-6.812 ± 0.274	-8.081 ± 0.252	-8.978 ± 0.525	-8.285 ± 0.370
GA+D [32]	-8.365 ± 0.201	-6.539 ± 0.297	-8.567 ± 0.177	-9.371 ± 0.728	-8.610 ± 0.104
MARS [45]	-9.716 ± 0.082	-7.839 ± 0.018	-9.804 ± 0.073	-9.569 ± 0.078	-9.150 ± 0.114
GEGL [1]	-9.329 ± 0.170	-7.470 ± 0.013	-9.086 ± 0.067	-9.073 ± 0.047	-8.601 ± 0.038
RationaleRL [18]	-10.663 ± 0.086	-8.129 ± 0.048	-9.005 ± 0.155	*No hit found*	-9.398 ± 0.076
FREED [46]	-10.579 ± 0.104	-8.378 ± 0.044	-10.714 ± 0.183	-10.561 ± 0.080	-9.735 ± 0.022
PS-VAE [20]	-9.978 ± 0.091	-8.028 ± 0.050	-9.887 ± 0.115	-9.637 ± 0.049	-9.464 ± 0.129
MOOD [24]	-10.865 ± 0.113	-8.160 ± 0.071	-11.145 ± 0.042	-11.063 ± 0.034	-10.147 ± 0.060
RetMol [42]	-8.590 ± 0.475	-5.448 ± 0.688	-6.980 ± 0.740	-8.811 ± 0.574	-7.133 ± 0.242
GEAM [25]	-12.891 ± 0.158	-9.890 ± 0.116	-12.374 ± 0.036	-12.342 ± 0.095	-11.816 ± 0.067
f-RAG (ours)	**-12.945** ± 0.053	**-9.899** ± 0.205	**-12.670** ± 0.144	**-12.390** ± 0.046	**-11.842** ± 0.316