Neural Graph Matching Improves Retrieval Augmented Generation in Molecular Machine Learning

Runzhong Wang*, Rui-Xi Wang*, Mrunali Manjrekar, Connor W. Coley

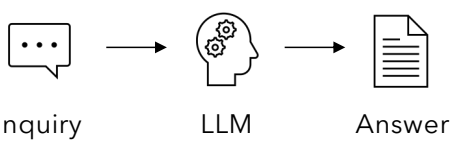
ICML 2025

https://github.com/coleygroup/ms-pred

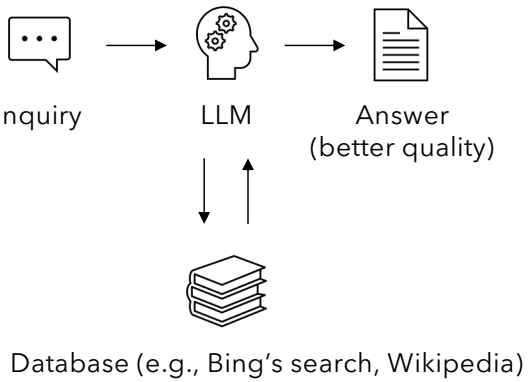
ChemE MIT Chemical Engineering MIT EECS

Retrieval Augmented Generation (RAG)

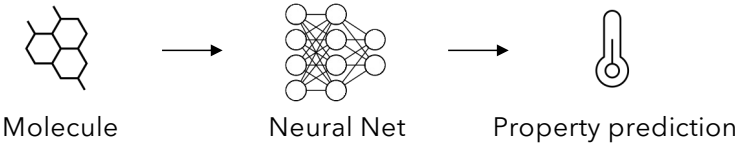
Text generation (LLMs)



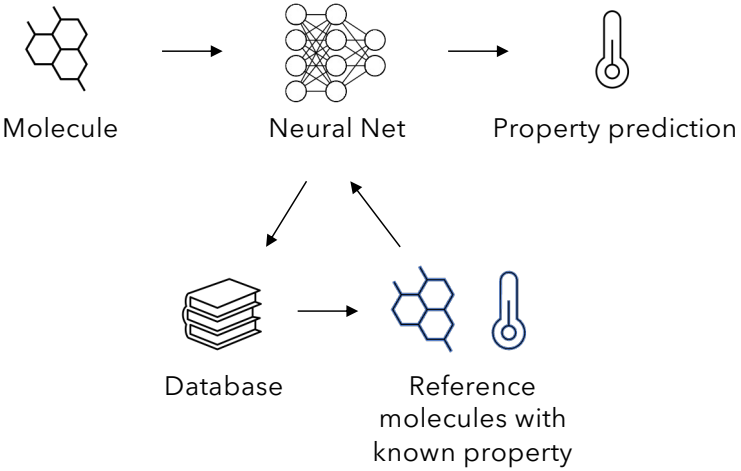
with RAG:



Molecular machine learning
(focus of this paper)

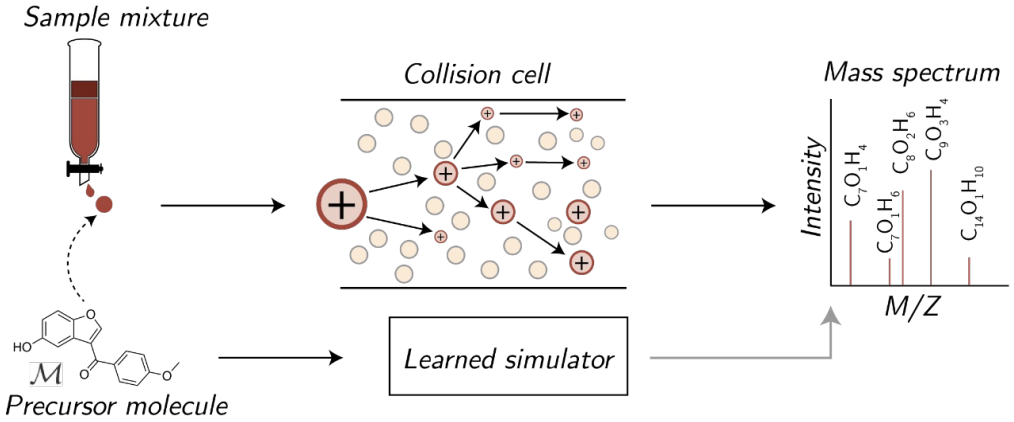


with RAG:

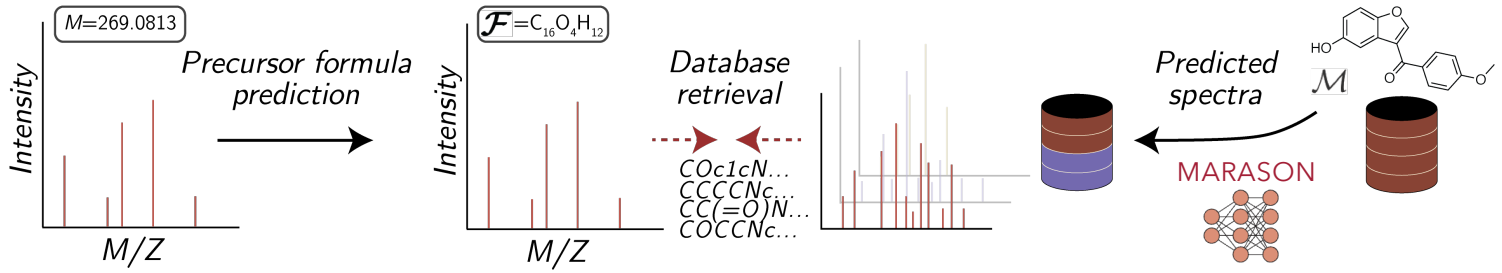


We took mass spectra simulation as the case study

Task: mimic what's happening in the analytical tool, mass spectrometer (MS)

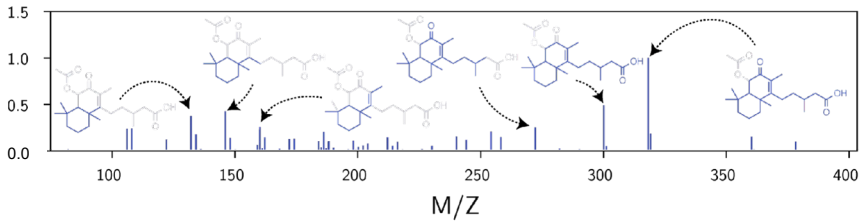


Scientific value: MS simulation models could help structural elucidation

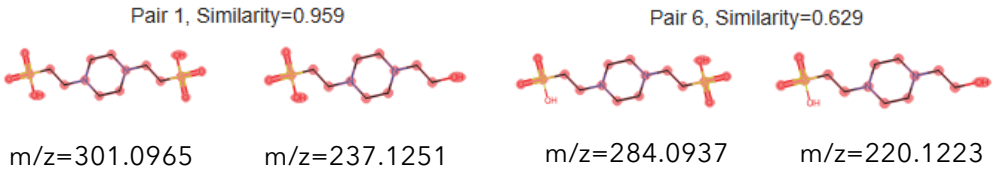


Key insight: matching reference and target molecuels

In mass spectrometry, peaks are attributed to molecular fragments

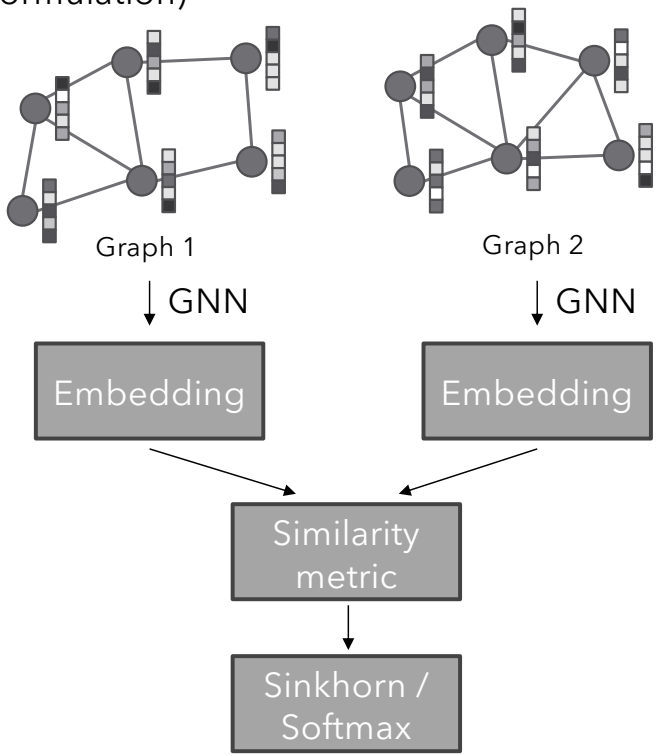


The reference structure might have fragments with distinct masses, and the mapping is not obvious until the molecular graphs are matched



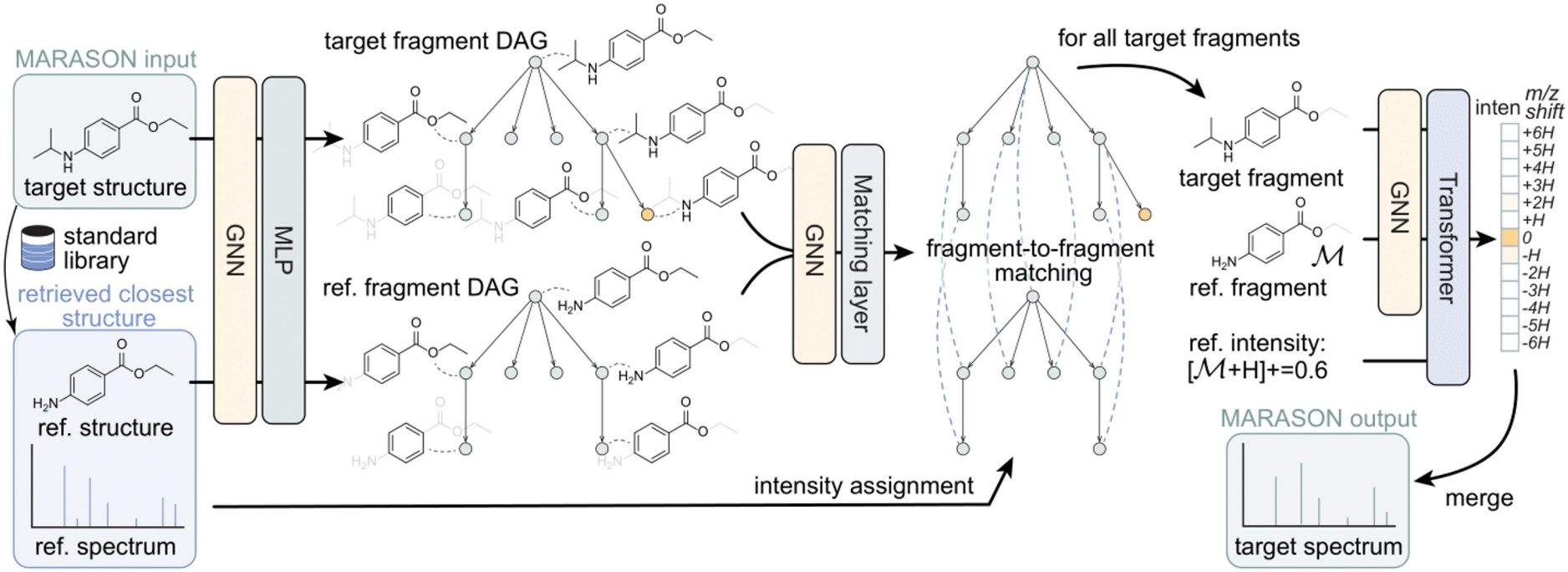
	No RAG	Concat RAG	Matching RAG
Cosine sim.	0.739	0.737 (-0.3%)	0.757 (+2.4%)

Neural graph matching (linear matching formulation)



Wang et al. *TPAMI*. 2023

MARASON: RAG with graph matching for MS/MS simulation



MARASON

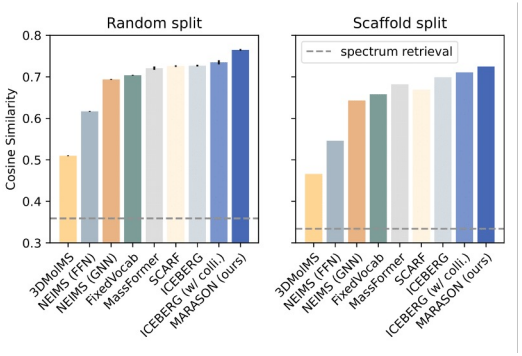
- RAG query by Tanimoto similarity on the training dataset (could further extend)
- Once target and reference fragments are generated, MARASON uses a nested neural graph matching layer to match them

MARASON shows state-of-the-art accuracy on benchmarks

Retrieval accuracy, NIST'20:

Accuracy @ Top-k	1	2	3	4	5	8	10
Random	0.026±0.001	0.052±0.001	0.076±0.002	0.098±0.001	0.120±0.001	0.189±0.003	0.233±0.004
3DMolMS (Hong et al., 2023)	0.055±0.003	0.105±0.000	0.146±0.005	0.185±0.007	0.225±0.009	0.332±0.005	0.394±0.008
FixedVocab (Murphy et al., 2023)	0.172±0.004	0.304±0.004	0.399±0.002	0.466±0.007	0.522±0.012	0.638±0.009	0.688±0.006
NEIMS (FFN) (Wei et al., 2019)	0.105±0.003	0.243±0.012	0.324±0.013	0.387±0.011	0.440±0.014	0.549±0.010	0.607±0.005
NEIMS (GNN) (Zhu et al., 2020)	0.175±0.005	0.305±0.003	0.398±0.002	0.462±0.004	0.515±0.005	0.632±0.007	0.687±0.005
MassFormer (Young et al., 2024a)	0.191±0.008	0.328±0.006	0.422±0.004	0.491±0.002	0.550±0.005	0.662±0.005	0.716±0.003
SCARF (Goldman et al., 2023)	0.187±0.008	0.321±0.006	0.417±0.007	0.486±0.008	0.541±0.009	0.652±0.008	0.708±0.009
ICEBERG (Goldman et al., 2024)	0.189±0.012	0.375±0.005	0.489±0.007	0.567±0.005	0.623±0.004	0.725±0.003	0.770±0.002
ICEBERG (w/ collion energy)	0.202±0.009	0.399±0.008	0.513±0.008	0.585±0.008	0.639±0.010	0.749±0.006	0.793±0.007
MARASON (ours)	**0.278±0.002**	**0.455±0.004**	**0.562±0.009**	**0.636±0.006**	**0.685±0.004**	**0.784±0.002**	**0.827±0.004**

Cosine similarity, NIST'20:



Retrieval accuracy, MassSpecGym:

Accuracy @ Top-k	1	5	20
Precursor m/z	0.0209 (0.0166-0.0259)	0.0852 (0.0765-0.0953)	0.2265 (0.2126-0.2401)
NEIMS (FFN) (Wei et al., 2019)	0.0762 (0.0677-0.0854)	0.2270 (0.2132-0.2412)	0.4412 (0.4251-0.4575)
NEIMS (GNN) (Zhu et al., 2020)	0.0363 (0.0305-0.0429)	0.1355 (0.1246-0.1468)	0.3377 (0.3226-0.3537)
FraGNNet (Young et al., 2024b)	0.3193 (0.3040-0.3350)	0.6320 (0.6164-0.6476)	0.8270 (0.8145-0.8393)
MARASON (ours)	**0.3403 (0.3286-0.3520)**	**0.6404 (0.6277-0.6519)**	**0.8539 (0.8448-0.8624)**

Ablation study, NIST'20:

Base model	RAG strategy	Match layer	Cosine sim.
MARASON (shared GNN)	No RAG	-	0.739
	Concat	-	0.737 (-0.3%)
	Hungarian	-	0.746 (+0.9%)
	RRWM	-	0.742 (+0.4%)
	NGM	Sinkhorn	0.749 (+1.4%)
	NGM	Softmax	0.753 (+1.9%)
MARASON (not shared GNN)	NGM	Sinkhorn	0.753 (+1.9%)
	NGM	Softmax	**0.757 (+2.4%)**

NIST'20: NIST standard reference database. National Instituteof Standards and Technology, 2020.
MassSpecGym: A benchmark for the discovery and identification of molecules. NeurIPS 2024.

Thank you!

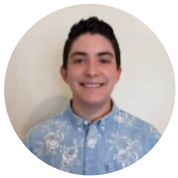
MIT Coley Group MS/MS Team


Prof. Connor Coley


Dr. Runzhong Wang


Rui-Xi (Ray) Wang


Mrunali Manjrekar

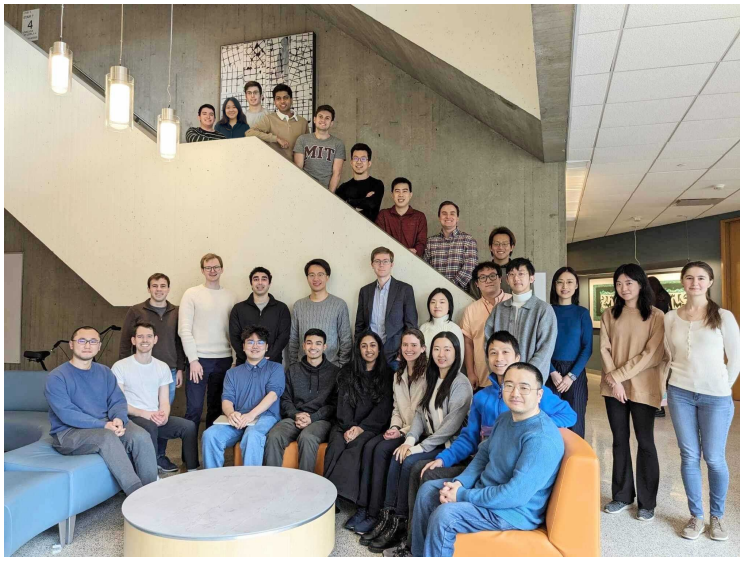

Joules Provenzano


Montgomery Bohde


Magdalena Lederbauer


Dr. Samuel Goldman
(now at MPM)

MIT Coley Group



Code https://github.com/coleygroup/ms-pred

Funding Support

MLPDS
Machine Learning for Pharmaceutical Discovery and Synthesis

DSO NATIONAL LABORATORIES

UROP
Undergraduate Research Opportunities Program