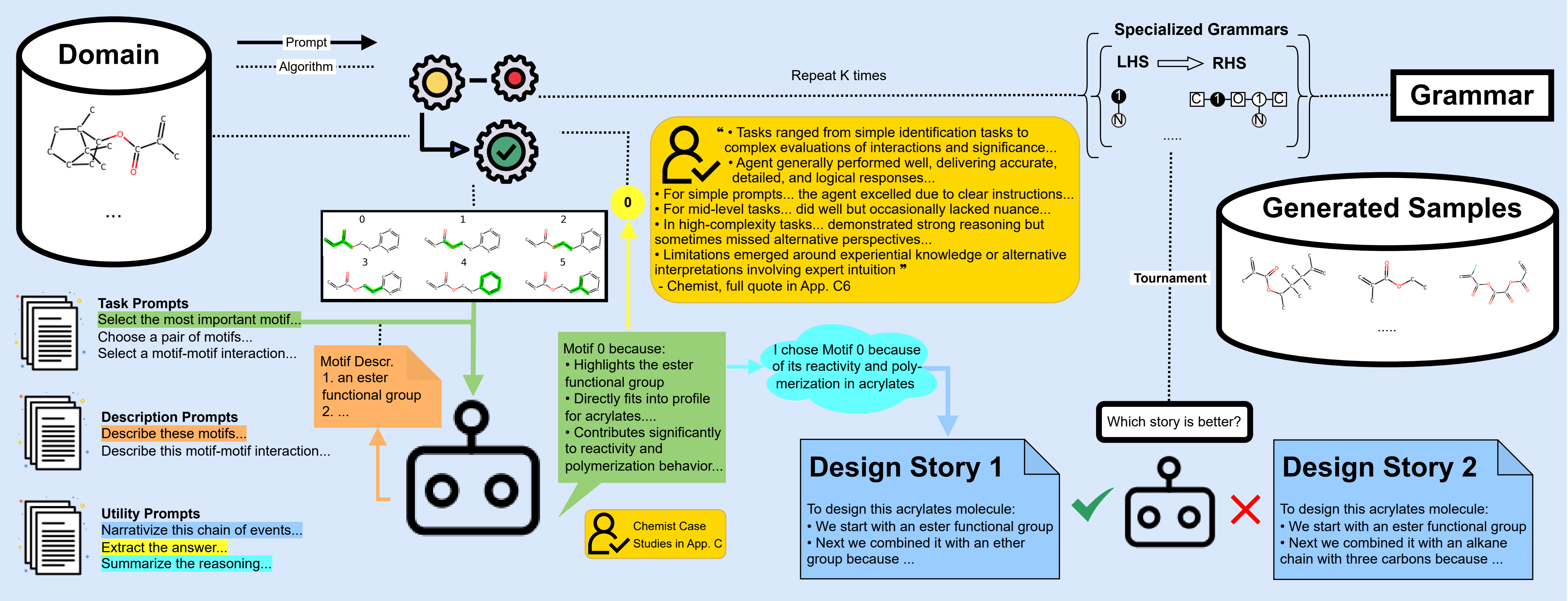
Foundation Molecular Grammar: Multi-Modal Foundation Models Induce Interpretable Molecular Graph Languages

Michael Sun[1*], Gang Liu[1], Weize Yuan[3], Wojciech Matusik[1], Jie Chen[4]

[1]MIT CSAIL, [2]Notre Dame, [3]MIT Chemistry, [4]MIT-IBM Watson AI Lab, [*]Website: michaelsun.tech

LLM accelerates molecular discovery by acting as experts in decomposing molecules, designing design languages.



Task Prompts: Select the most important motif... Choose a pair of motifs... Select a motif-motif interaction...

Description Prompts: Describe these motifs... Describe this motif-motif interaction...

Utility Prompts: Narrativize this chain of events... Extract the answer... Summarize the reasoning...

Motif Descr. 1. an ester functional group 2. ...

Motif 0 because:
- Highlights the ester functional group
- Directly fits into profile for acrylates....
- Contributes significantly to reactivity and polymerization behavior...

I chose Motif 0 because of its reactivity and polymerization in acrylates

“ • Tasks ranged from simple identification tasks to complex evaluations of interactions and significance...
• Agent generally performed well, delivering accurate, detailed, and logical responses...
• For simple prompts... the agent excelled due to clear instructions...
• For mid-level tasks... did well but occasionally lacked nuance...
• In high-complexity tasks... demonstrated strong reasoning but sometimes missed alternative perspectives...
• Limitations emerged around experiential knowledge or alternative interpretations involving expert intuition ”
- Chemist, full quote in App. C6

Design Story 1

To design this acrylates molecule:
- We start with an ester functional group
- Next we combined it with an ether group because ...

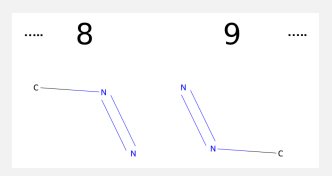
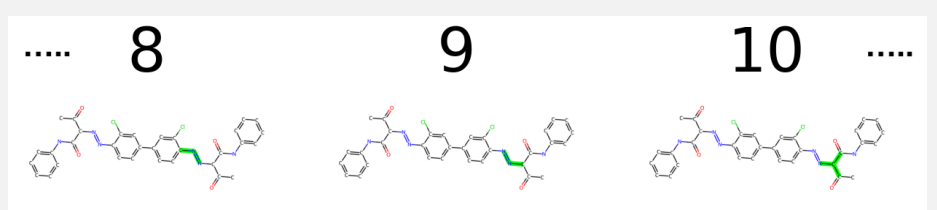
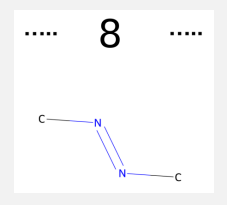
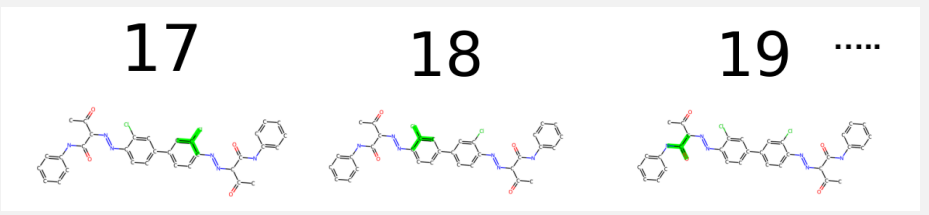
Design Story 2

To design this acrylates molecule:
- We start with an ester functional group
- Next we combined it with an alkane chain with three carbons because ...

Table 7. We tally LLM judge decisions against the human decision (Gold) across 3 repeated calls per molecule. For 6 molecules (columns with answer in [brackets]), the expert found both designs equally reasonable, so we exclude those from the total.

	Isocyanates		Acrylates		HOPV		PTC		Total
Gold	*A*	*B*	*A*	*B*	*A*	*B*	*A*	*B*	
1	*BAAB*[*B*]	*BBBB*[*B*]	*AABAB*	*BBABB*	*BBAAA*	*BBBBA*	[*B*]*BABA*	[*A*]*BBBA*	25/36
2	*BAAB*[*B*]	*BBBA*[*A*]	*AAAAA*	*BBBBA*	*BBABA*	*BBBBA*	[*B*]*BABA*	[*B*]*BBAB*	25/36
3	*BAAB*[*B*]	*ABBA*[*B*]	*AAAAB*	*BBBBB*	*BAAAB*	*BBBBB*	[*B*]*BABA*	[*A*]*ABBB*	26/36
Score	6/12	9/12	12/15	13/15	9/15	13/15	6/12	9/12	77/108 = 71%

LLM-as-a-judge chooses which decomp. was better. **Expert agrees with LLM on 71% of cases.** Using LLM judge, we host Swiss-style tournament to rank decomps. The top *k* tournament-winning decomps craft a **generative molecular grammar.**

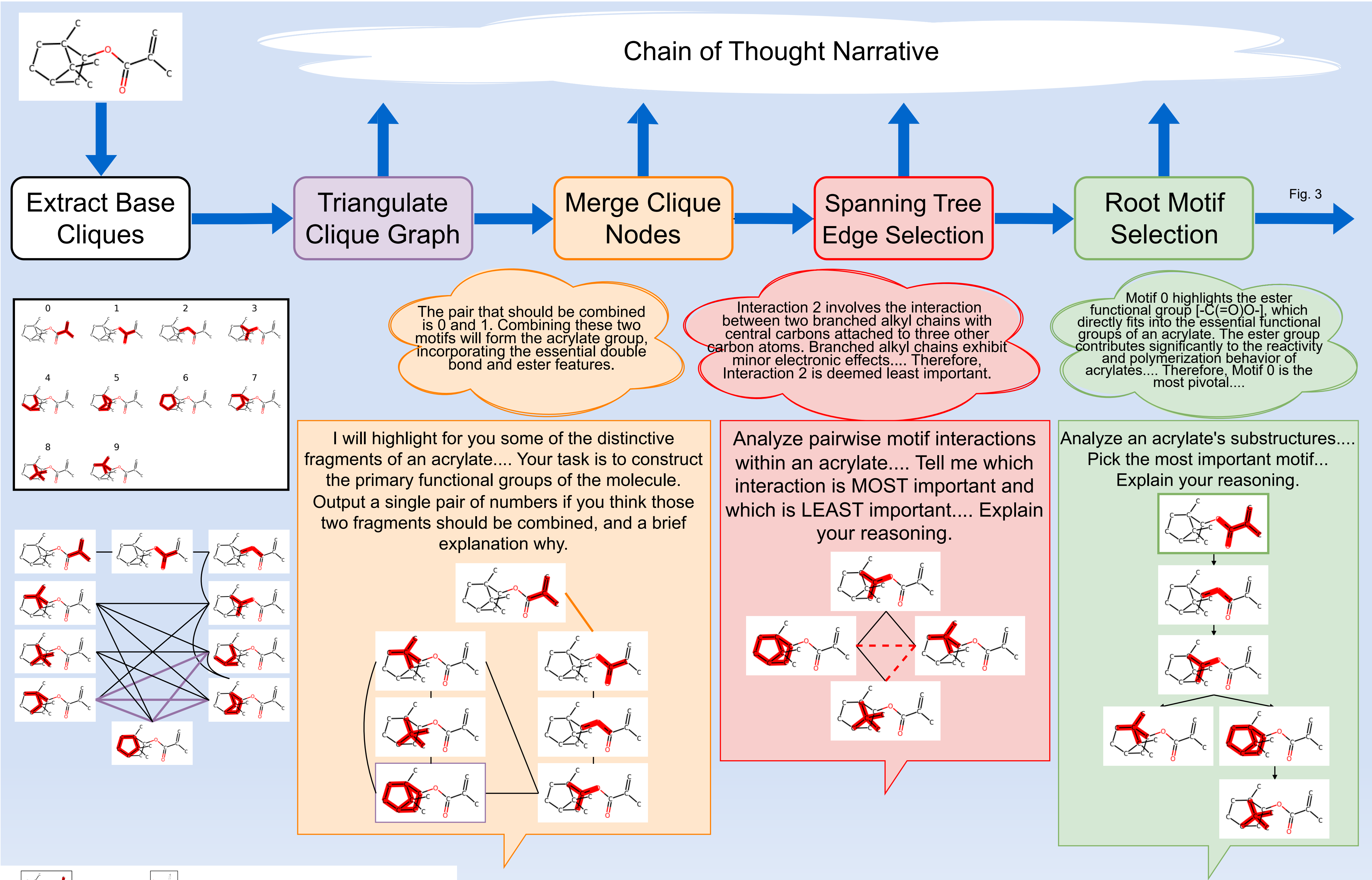
Prompt	Response	Expert Evaluation
...I want you to explain, concisely, what each numbered motif is ... 	.. **Motif 8:** Di-substituted carbon chain with a nitrile group (N=C) on one end and another nitrile (N≡C) in the middle. **Motif 9:** ...	... Each motif was correctly categorized with its chemical identity and described in terms of its structure and connectivity, reflecting a solid understanding of organic chemistry principles ...
...I want you to tell me if any two of them should be combined to-gether to form a more meaningful substructure ... 	**Pair: 8 and 9** **Explanation:** ... Combining these two fragments could provide a more comprehensive understanding of the arrangement and connectivity of nitrile groups along the carbon chain, which is critical in analyzing the structure and reactivity of the molecule ...	Medium, PTC molecules are more complex, so it requires considering combinations of more than one pair of motifs ...
...I want you to explain, concisely, what each numbered motif is ... 	... *Motif 8**: Butanedinitrile - A nitrile with a CN group at each end of a butane backbone. ...	...
...	...	...
... I want you to pick only ONE of these as the root motif most essential to its chemical profile ... 	... From the motifs listed, **Motif 17 (Chloroacetylene)** and **Motif 18 (1,1 Dichloroethane)** contain halide groups ... Given the need to select one motif, **Motif 18** (1,1 Dichloroethane) stands out slightly more due to the presence of two chlorine atoms ...	... GPT choose Motif 18 as the most essential root motif, which is correct. This is because 1,1 Dichloroethane obtain higher significancy due to the presence of two chlorine atoms ...

(↑) Example run. GPT-4o reasons and chooses between presented options. Experts validate the traces.

(→) Tallying turn-by-turn expert agreement of prompt-response pairs

Dataset		Easy	Medium	Hard
Small Dataset	Correct	6	3	0
	Partial	0	1	0
Real-World Dataset	Correct	5	2	2
	Partial	0	1	0

GPT-4o can play a core role within an algorithm to provide interpretability and accuracy in decomposing molecules.



The pair that should be combined is 0 and 1. Combining these two motifs will form the acrylate group, incorporating the essential double bond and ester features.

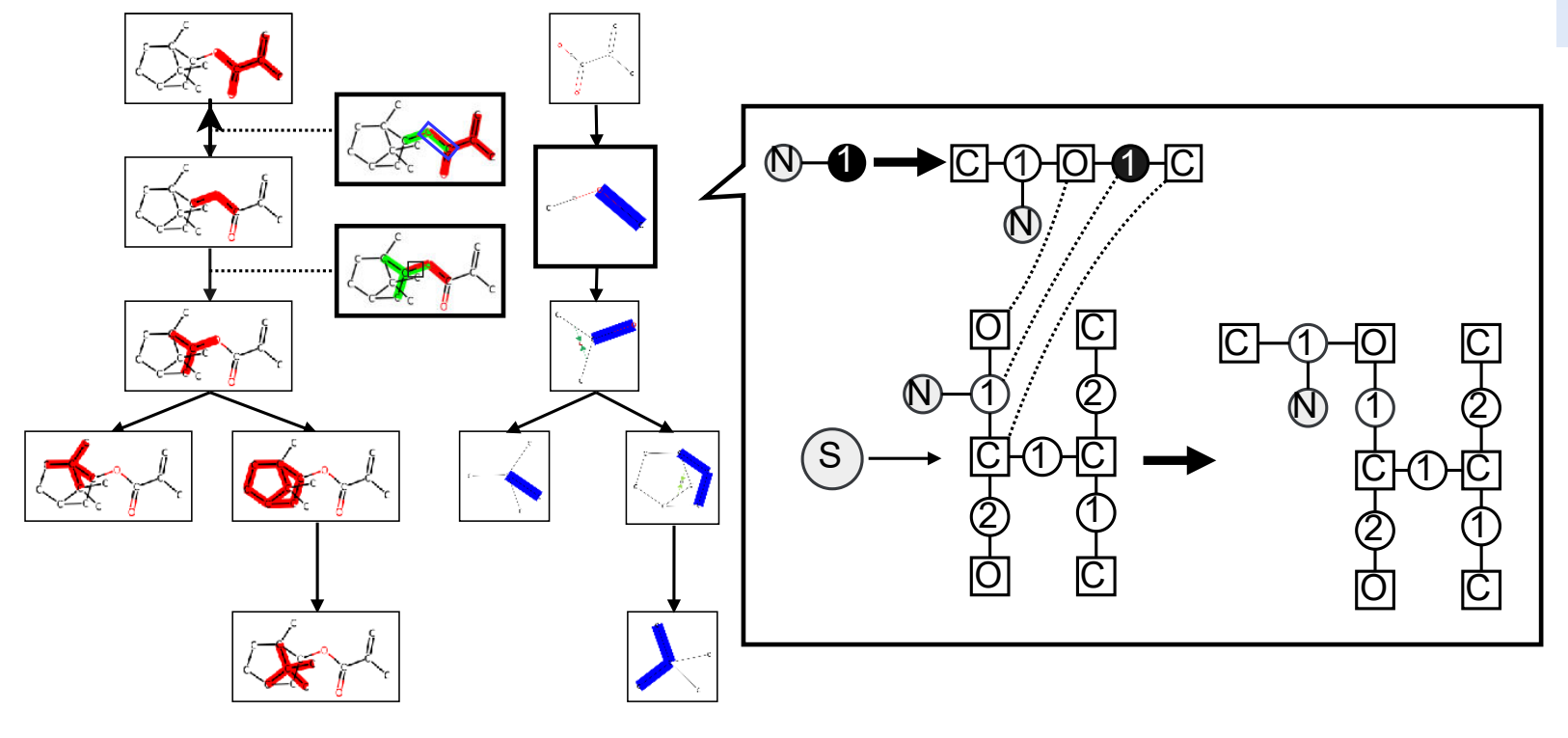
Interaction 2 involves the interaction between two branched alkyl chains with central carbons attached to three other carbon atoms. Branched alkyl chains exhibit minor electronic effects.... Therefore, Interaction 2 is deemed least important.

Motif 0 highlights the ester functional group [-C(=O)O-], which directly fits into the essential functional groups of an acrylate. The ester group contributes significantly to the reactivity and polymerization behavior of acrylates.... Therefore, Motif 0 is the most pivotal....

I will highlight for you some of the distinctive fragments of an acrylate.... Your task is to construct the primary functional groups of the molecule. Output a single pair of numbers if you think those two fragments should be combined, and a brief explanation why.

Analyze pairwise motif interactions within an acrylate.... Tell me which interaction is MOST important and which is LEAST important.... Explain your reasoning.

Analyze an acrylate's substructures.... Pick the most important motif... Explain your reasoning.



(←) Example conversion from clique tree to production rules; (Left) Each node of tree contains a substructure (red), edges mean shared bonds between substructures; (Right-top) Rule extracted from second clique of the tree, with a non-terminal hyperedge for the LHS and the clique’s substructure being the RHS; (Right-bottom) example of rule application, dashed connections are matching bonds & atoms

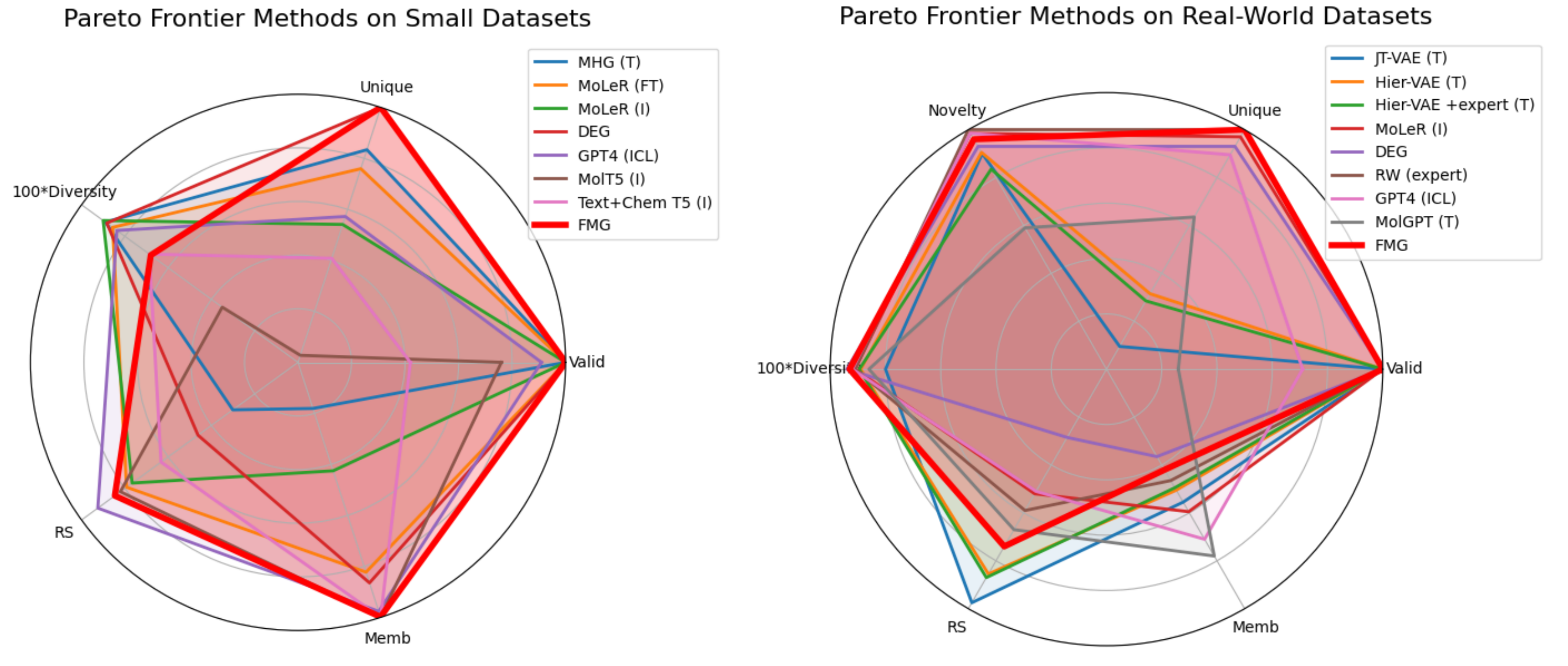
Table 1. Results on Small Datasets Isocyanates (11 examples), Acrylates (32) and Chain Extenders (11). Best rounded result(s) for each metric **bolded**. (T): Training from scratch; (FT): Fine-tuning, (I): Inference, using posterior interpolation. Since T5 (I) methods struggle to generate sufficient valid and unique samples, we exclude them from ranking.

Method	Valid (Avg.)	Unique			Div.			RS			Memb.		
Train Data	100%	100%	100%	100%	0.61	0.67	0.80	100%	100%	100%	100%	100%	100%
JT-VAE (T)	100%	5.8%	0.5%	2.3%	0.72	0.29	0.62	5.5%	4.9%	2.2%	66.5%	48.64%	79.6%
Hier-VAE (T)	100%	99.6%	99.7%	99.8%	0.83	0.83	0.83	1.85%	3.04%	2.69%	0.05%	0.82%	43.6%
MHG (T)	100%	75.9%	86.8%	87.4%	0.88	0.89	0.90	2.97%	36.8%	50.6%	12.1%	0.93%	41.2%
MoLeR (FT)	100%	87.1%	40.7%	**100%**	0.86	0.80	0.91	69.2%	**97.7%**	70.7%	77.3%	72.2%	97.3%
MoLeR (I)	100%	65.7%	45.4%	51.1%	**0.90**	**0.90**	0.90	61.3%	76.2%	92.3%	0.08%	32.0%	95.5%
STONED	100%	**100%**	99.8%	99.8%	0.85	0.84	**0.93**	5.63%	11.2%	6.78%	79.8%	47.9%	61.0%
DEG	100%	**100%**	**100%**	**100%**	0.86	0.86	**0.93**	27.2%	43.9%	67.5%	96.3%	69.6%	93.5%
GPT4 (ICL)	91%	73.0%	35.1%	63.5%	0.86	0.78	0.87	**84.4%**	95.0%	98.0%	93.7%	99.7%	99.5%
MolT5 (I)	76%	0.9%	0.3%	7.1%	0.09	0.21	0.75	98.1%	99.6%	48.5%	99.8%	100%	100%
Text+Chem T5 (I)	42%	26.2%	46.4%	49.8%	0.55	0.71	0.80	87.6%	58.3%	43.9%	100%	100%	100%
FMG	100%	**100%**	**100%**	**100%**	0.73	0.46	0.85	61.7%	93.0%	**99.1%**	**99.6%**	**100%**	**99.8%**

Table 2. Results on Real-World Datasets HOPV (316 examples) and PTC (348). Same protocol as Table 1. Since VAE (T) methods struggle to generate sufficient unique samples, we exclude them from ranking.

Method	Valid (Avg.)	Unique		Novelty		Div.		RS		Memb.	
Train Data	100%	100%	100%	N/A	N/A	0.86	0.94	51%	87%	100%	30%
JT-VAE (T)	100%	11%	8%	100%	80%	0.77	0.83	99%	96%	84%	27%
Hier-VAE (T)	100%	43%	20%	96%	85%	0.87	0.91	79%	92%	76%	25%
Hier-VAE +expert (T)	100%	29%	28%	92%	75%	0.86	0.93	84%	90%	82%	17%
MoLeR (FT)	100%	**100%**	99%	**100%**	99%	0.90	0.92	42%	71%	60%	30%
MoLeR (I)	100%	**100%**	94%	**100%**	97%	0.90	0.92	25%	79%	74%	45%
DEG	100%	98%	88%	99%	87%	**0.93**	**0.95**	19%	38%	46%	27%
RW (expert)	100%	**100%**	**100%**	**100%**	**100%**	0.89	0.93	58%	60%	71%	22%
GPT4 (ICL)	71%	95%	84%	99%	98%	0.91	0.93	46%	56%	53%	**89%**
MolGPT (T)	26%	86%	41%	71%	47%	0.84	0.88	43%	**91%**	86%	70%
MolT5 (I)	61%	12%	20%	**100%**	95%	0.41	0.77	8%	**91%**	0%	45%
Text+Chem T5 (I)	48%	81%	95%	67%	91%	0.87	0.91	42%	50%	**88%**	47%
FMG	100%	**100%**	**100%**	**100%**	92%	**0.93**	0.93	**70%**	78%	38%	46%

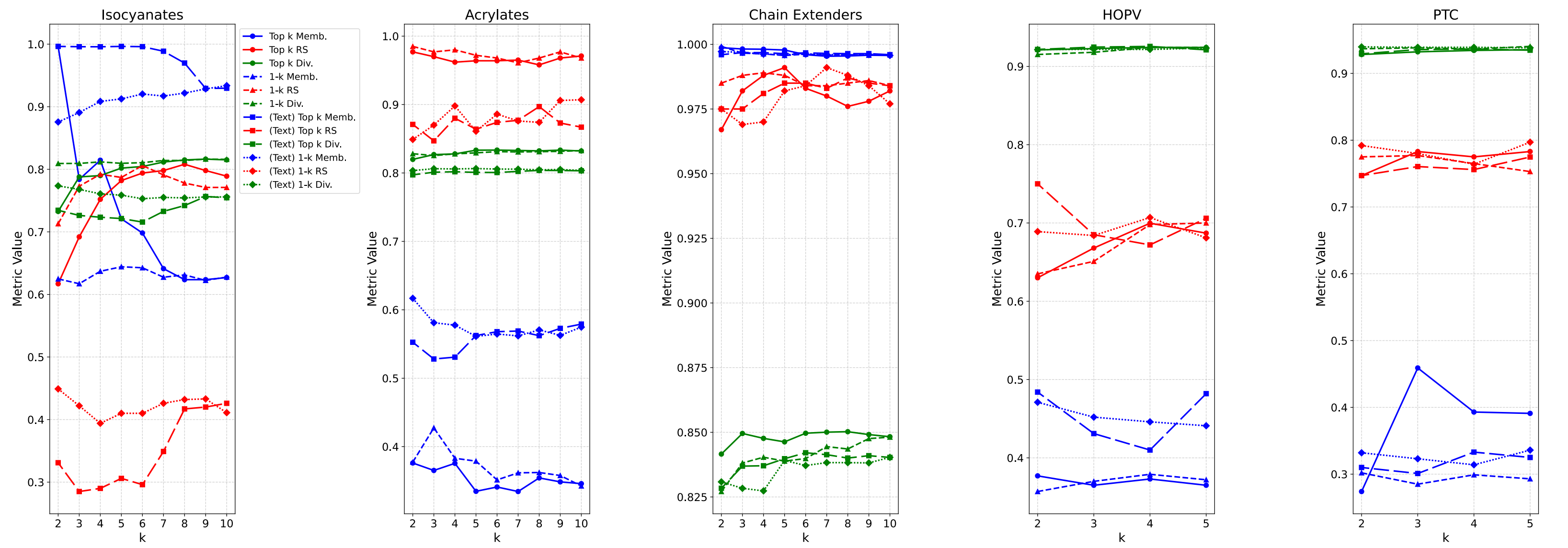
Ablation studies demonstrate versatility of method and key enablers of performance.



(a) Spider plot of pareto-efficient methods in Table 1.

(b) Spider plot of pareto-efficient methods in Table 2.

(↑) Generative coverage, synthesizability, and specificity of FMG exceeds Chemical LMs, VAEs, non-LLM grammar-based methods, and LLM baselines.



(↑) We vary k from 2-10 (small dataset) and 2-5 (real-world dataset) to observe tradeoffs in generation metrics as *k* increases (more top decomps used). We modify FMG to use a text-based encoding (FMG-Text) instead of molecular images.

(↓) For each FMG module, we swap MMFM for a heuristic (-module). Results use $k = 5$.

Method	Novelty			Div.			RS			Memb.		
FMG	99.96	99.87	99.94	0.81	0.83	0.84	78.7	97.2	98.8	64.42	37.88	22.07
FMG (-merge)	99.95	99.88	99.94	0.76	0.83	0.85	39.7	90.3	96.4	93.74	16.40	14.44
FMG (-edge)	99.95	99.87	99.95	0.81	0.83	0.84	66.8	92.7	98.4	58.57	33.83	16.23
FMG (-root)	99.97	99.86	99.94	0.82	0.85	0.86	54.9	87.0	96.2	47.01	22.18	14.84

(↓) We trained FMG on a 1k subset (0.05%) of the refined ZINC dataset used by the MOSES benchmark. Baseline numbers are copied from the MOSES leaderboard.

Model	Valid(↑)	Unique@1k(↑)	Unique@10k(↑)	FCD-TestSF(↓)	SNN-TestSF(↑)	Scaf-TestSF(↑)	IntDiv(↑)	Novelty(↑)
Train	1.00	1.00	1.00	0.48	0.59	0.00	0.86	1.00
HMM	0.08±0.03	0.62±0.12	0.57±0.14	25.43±2.56	0.38±0.01	0.05±0.02	0.85±0.04	**1.00±0.00**
NGram	0.24±0.00	0.97±0.01	0.92±0.00	6.23±0.10	0.50±0.00	0.10±0.01	0.87±0.00	0.97±0.00
Combinatorial	**1.00±0.00**	**1.00±0.00**	0.99±0.00	4.51±0.03	0.44±0.00	0.09±0.00	0.87±0.00	0.99±0.00
CharRNN	0.97±0.03	**1.00±0.00**	**1.00±0.00**	**0.52±0.04**	0.56±0.01	0.11±0.01	0.86±0.00	0.84±0.05
AAE	0.94±0.03	**1.00±0.00**	**1.00±0.00**	1.06±0.24	0.57±0.00	0.08±0.01	0.86±0.00	0.79±0.03
VAE	0.98±0.00	**1.00±0.00**	**1.00±0.00**	0.57±0.03	**0.58±0.00**	0.06±0.01	0.86±0.00	0.69±0.01
JTN-VAE	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**	0.94±0.05	0.52±0.01	0.10±0.01	0.86±0.00	0.91±0.01
LatentGAN	0.90±0.00	**1.00±0.00**	**1.00±0.00**	0.83±0.01	0.51±0.00	0.11±0.01	0.86±0.00	0.95±0.00
FMG (0.05%)	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**	26.30±0.41	0.29±0.00	**0.12±0.00**	**0.90±0.00**	**1.00±0.00**