Rethinking Molecular Text Representations for LLMs: An Empirical Study

Arun Raja *
University of Oxford

Garrett M. Morris
University of Oxford

Kian Ming A. Chai
DSO National Laboratories

Abstract

Large language models (LLMs) are increasingly used for molecular tasks, but it remains unclear *which molecular representation to use*. We present a systematic benchmark, MolRepBench, to evaluate LLM molecular competence across nine representations and eight chemical tasks. We benchmark 16 LLMs across five model families, including reasoning and non-reasoning variants, chemistry-specialized LLMs, and closed frontier models. Performance is strongly representation-dependent and no single representation wins across tasks, though CML is the best, followed by MolJSON, InChI and canonical SMILES. Explicit structured text representations (CML and MolJSON) dominate structural tasks; IUPAC dominates semantic tasks, winning molecule retrieval for all 16 LLMs; and SMILES variants are rarely optimal despite their prevalence in pretraining. Chemistry-specialized models perform well with SMILES at the cost of large degradations with structured text representations, suggesting SMILES-only evaluation rewards specialization that does not generalize. Using LLM-as-a-judge, we find that IUPAC produces the highest fraction of correct molecule generations. A mechanistic study via tokenization audits, linear probes, and attention shows that representations are encoded differently inside the model; for example, structured representations require more attention across the molecular span. Our results argue against representation-invariant evaluation and motivate task-aware representation routing for LLM-based chemistry.

1 Introduction

When using large language models (LLMs) for chemistry, molecules can be expressed in various text-based formats. However, the representation used for a task is mostly chosen by convention rather than empirical evidence. A medicinal chemist querying an LLM might give a SMILES string, an IUPAC name, or a SELFIES string—yet we have no systematic understanding of which representation best supports different tasks.

Prior work on molecular LLMs focuses either on training models from scratch on molecular strings (Edwards et al., 2022; Liu et al., 2023) or on fine-tuning foundation models for specific tasks (Zeng et al., 2022; Zhao et al., 2023). Most such works adopt SMILES (Weininger, 1988) as the de facto representation, given its prevalence in chemistry literature and applications. However, there are molecular text representations for different purposes.

Chemistry has diverse machine-readable representations, each different in compactness, invertibility, and application (David et al., 2020; Wigh et al., 2022; Raghunathan and Priyakumar, 2022). An early textual representation was the Wiswesser Line Notation (WLN, Wiswesser 1954). It was designed to be punched onto cards and processed by tabulating machines, and it was the dominant machine-readable representation in the industry.

Rigaudy and Klesney (1979) introduced a nomenclature for the International Union of Pure and Applied Chemistry (IUPAC). IUPAC maintains rules to name any molecule in natural language. It is more human-readable than other text notations. However, the rules are elaborate with edge cases, limiting scalability. In contrast to WLN and IUPAC, Simplified Molecular Input Line Entry System (SMILES, Weininger 1988) is a context-free grammar and can be algorithmically derived. It was created for modern computer software to process and store chemical structures efficiently. Its canonical variant, canonical SMILES, ensured uniqueness for each molecule, so it could be used as a database key. The advent of generative AI in chemistry introduced DeepSMILES (O'Boyle and Dalke, 2018) and SELFIES (Krenn et al., 2020), designed to reduce invalid generations from probabilistic models.

To store, access, and exchange descriptions of molecules on the internet, Extensible Markup Lan-

*arun.raja@dtc.ox.ac.uk

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 59565–59589
October 24-29, 2026 ©2026 Association for Computational Linguistics

guage and JavaScript Object Notation have become popular. Chemical Markup Language (CML, Murray-Rust and Rzepa 2000, 2003) is the first such; MolJSON (Runcie et al., 2026) is the latest (see Section A). They are called **structured text representations** because they explicitly store graph information such as atoms and bond connectivity.

Therefore, there are dominant molecular representations for different purposes and technology. Presently, LLM is a prevalent technology. It has its own affordances (tokenization, in-context learning, alignment with natural language) and failure modes (for example, hallucination). *Which representation should chemistry use for an LLM?*

We address the question via a thorough, systematic study of various molecular text representations for use with LLMs. Our contributions are:

1. **MolRepBench**, a benchmark of eight tasks covering comprehension and generation, each prompted with nine representation formats.
2. **Evaluation across five model families**, spanning non-reasoning and reasoning-capable models, domain-specialized models and closed frontier models. We use **LLM-as-a-judge** to qualitatively evaluate the molecules generated with different representations.
3. **A mechanistic interpretability study** in which we investigate how an LLM 'views' the various molecular representations.

Our findings conclude that current LLMs do not yet possess robust, representation-invariant molecular capabilities. Thus, molecular representation is a key factor when using LLMs for chemistry.

2 Related Work

The effect of representations on LLM performance has been explored in general. For instance, Yauney and Mimno (2021) show that the difficulty of a dataset is related to the alignment between the input representation and the labeling. Al-shaibani and Ahmad (2023) suggest a consonant-based compact representation for English using less computational resources yet achieving performance comparable to the standard text representation. Our work uses both compact and verbose representations to investigate how verbosity impacts an LLM's performance across different chemistry tasks.

A growing body of work applies **large language models (LLMs) to chemistry**. Domain-specific models such as MolT5 (Edwards et al., 2022), Text+Chem T5 (Christofidellis et al., 2023), Galactica (Taylor et al., 2022) and nach0 (Livne et al., 2024) jointly model SMILES and natural language for molecule captioning, caption-to-molecule generation, and property prediction. General-purpose LLMs have also been evaluated on chemistry tasks through prompting and in-context learning (Jablonka et al., 2024; Li et al., 2024), and recent benchmarks such as ChemLLMBench (Guo et al., 2023) and ChemBench (Mirza et al., 2024) measure performance across property prediction, name conversion, and structural reasoning. However, the dominant representation across this literature is SMILES; there are only occasional comparisons to SELFIES or to text-encoded molecular graphs.

A **prior work comparing representations for molecules** by Baker et al. (2025) studies canonical SMILES, DeepSMILES, SELFIES, InChI and IUPAC names for molecular property prediction. Our work builds upon this in *three* aspects (see Table 1). First, the prior work has not explored structured text representations, which may be important given that graph neural networks (GNNs) outperform LLMs in molecular property prediction (Zhong et al., 2024; Gupta et al., 2025; Jacobs et al., 2026). Our work evaluates CML and MolJSON, which textually encode the molecular graph.

Given that LLMs are known to be less competent than GNNs for property *prediction*, comparing text representations on predicting properties, such as solubility, toxicity, and lipophilicity, may offer limited insight. These biochemical properties are measured using wet-lab assays, and their reliable predictions require chemical or physical inductive biases. Therefore, we instead frame such tasks as molecular property *estimation*, where the LLMs are to predict the number of hydrogen bond donors and acceptors, Wildman-Crippen LogP value and topological polar surface area or TPSA for molecules. These properties are computable from the molecular graph (using RDKit, Landrum 2016) without wet-lab experiments. This is the second aspect.

The third aspect is data contamination. Baker et al. (2025) uses datasets from the widely used MoleculeNet (Wu et al., 2018). These datasets are likely in the pretraining corpus (Sainz et al., 2023; Cheng et al., 2025; Xu et al., 2025; Busch et al., 2026). By contrast, our property labels are generated using RDKit, so are unavailable elsewhere.

Aspect	Prior work (Baker et al., 2025)	Ours
Representations	Canonical SMILES, DeepSMILES, SELFIES, InChI, IUPAC. *No* structured text respresentations.	In addition to Baker et al.'s, isomeric SMILES, randomized SMILES, **CML** and **MolJSON**. The last two are structured text representations.
Task framing	Molecular *property prediction* (e.g., solubility, toxicity, lipophilicity): wet-lab properties requiring chemical/physical inductive biases that LLMs lack	Molecular *property estimation* (e.g., HBD, HBA, Wildman–Crippen LogP, TPSA): properties computable from molecular graphs without wet-lab assays
Label in Data	MoleculeNet datasets, likely present in LLM pretraining corpora	Labels generated with RDKit, unavailable elsewhere, mitigating contamination

Table 1: Comparison of MolRepBench with the prior representation-comparison study by Baker et al. (2025).

3 Benchmark Design

We evaluate **nine molecular representations** (Table 2 and Appendix J). Each representation encodes the same molecule but differs in syntactic structure and implicit assumptions. We decided on these nine based on their prevalent use in modern cheminformatics tools and deep learning methods. SMILES is the de facto standard and is widely represented in LLM pretraining corpora. We include three variants: canonical SMILES determines the atom ordering algorithmically; isomeric SMILES adds stereochemical annotations; and randomized SMILES removes canonical ordering by generating a new traversal each time.

Next, we consider two representations developed for generative models. First, DeepSMILES is a derivative of SMILES that eliminates the two grammatical features most prone to autoregressive failure: matched parentheses and paired ring-closure digits. Second, SELFIES guarantees that every string decodes into a valid molecule. We evaluate whether such syntactic fixes help LLMs.

We also study two notations from IUPAC, namely IUPAC name and International Chemical Identifier (InChI, McNaught 2006). IUPAC names are linguistically compositional and human-readable, so the LLM can parse it as natural language rather than as symbolic code. They are also ubiquitous in chemistry texts and chemistry pretraining corpora (Baker et al., 2025). InChI is a compact string with detailed information, such as the chemical formula, atom connections, charge and stereochemistry, organized in layers and sublayers. InChI strings are also canonical.

Lastly, we have CML and MolJSON as structured representations with adjacency information.

Our benchmark, **MolRepBench**, has 8 molecular reasoning and generation tasks, spanning low-level structural parsing, chemically grounded semantic recognition, retrieval, and caption-to-molecule generation. [1] It draws molecular data from ChEBI-20 (Edwards et al., 2022) and ZINC250K (Irwin et al., 2012). The tasks are summarised below (and in Table 3; details in Sections C and D).

Atom Counting tests whether the LLM can accurately parse molecular strings and count atoms of specified elements, probing low-level structural comprehension.

Functional Group Identification evaluates the LLM's ability to recognize chemically meaningful substructures (aldehyde, ester, halide, primary amine, sulfonamide).

Molecular Property Estimation assesses the prediction of physicochemical properties—LogP, TPSA, hydrogen-bond donors (HBD) and hydrogen-bond acceptors (HBA)—directly from the representations, testing quantitative chemical reasoning.

Molecule Retrieval measures alignment between natural language descriptions and molecular structures using constructed distractors (similar scaffold, similar weight, and random ones).

Isomer Discrimination probes whether the LLM can distinguish molecules that differ subtly in atom ordering or stereochemistry. This tests fine-grained structural sensitivity.

Tautomer Recognition evaluates the LLM's understanding of dynamic structural equivalence, where two molecules are the same under relocation of a hydrogen atom or bond rearrangement.

Protonation State Recognition tests whether the LLM can identify molecules that are chemically identical but differ in protonation, assessing awareness of charge-state equivalence.

[1] The code for MolRepBench is available at https://github.com/arunraja-hub/molrepbench.

Representation	Year	Description	Example for L-Alanine ($C_3H_7NO_2$)
Canonical SMILES	1988	Short ASCII strings with canonical ordering	`CC(N)C(=O)O`
Isomeric SMILES	1988	SMILES with isotopic and chiral specifications	`C[C@@H](N)C(=O)O`
Randomized SMILES	2017	Non-canonical atom ordering, generated fresh per call using RDKit	`C(C(=O)O)(N)C`
DeepSMILES	2018	Removes matching parentheses and ring-closure digits	`CCN)C=O)O`
IUPAC name	1979	Natural language description that follows IUPAC nomenclature	`(2S)-2-aminopropanoic acid`
SELFIES	2020	Self-referencing embedded strings	`[C][C@@H1][Branch1]...[=O][O][N]`
CML	2000	XML-based explicit graph	`<molecule><atomArray>... </atomArray><bondArray>... </bondArray></molecule>`
InChI	2005	Layered canonical identifier by IUPAC	`InChI=1S/C3H7NO2/c1-2(4)...`
MolJSON	2026	Explicit graph representation in JSON	`{"atoms": [...],"bonds": [...]}`

Table 2: The nine molecular representations used in this work, with the year of introduction. For the L-Alanine example, SELFIES, CML, InChI, and MolJSON are abbreviated. Complete forms are provided in Section J.

Caption-to-Molecule Generation examines the LLM's generative capability to translate natural language descriptions into valid molecular strings, evaluating both syntactic validity and semantic fidelity.

We employ **various metrics** (Table 3) to evaluate across our diverse tasks. Atom counting uses exact-match accuracy. Functional group identification uses macro-F1 for the 5 binary labels, one for each functional group. In property estimation, logP and TPSA are regression tasks using Spearman correlation ρ, while hydrogen-bond donor/acceptor counts use exact-match accuracy. Molecule retrieval is a multiple-choice task scored by top-1 accuracy. Isomer discrimination, tautomer recognition, and protonation-state recognition are binary classification tasks using accuracy. For caption-to-molecule generation, we report validity rate, exact-match accuracy, Tanimoto similarities (Tanimoto, 1957) on Morgan (Rogers and Hahn, 2010), MACCS (Durant et al., 2002) and RDKit fingerprints (Landrum, 2016), and Fréchet ChemNet Distance (Preuer et al., 2018).

To further evaluate the caption-to-molecule generation task, we also use a closed frontier LLM, Gemini 3 Flash (Google DeepMind, 2025), as an **LLM-as-a-judge** to evaluate the generations. This provides qualitative insights into how each model and representation leads to various failure modes, such as hallucinations.

4 Experiments

We use 16 large language models (LLMs) spanning five families: three open-weight model families covering a range of sizes and reasoning capabilities, three LLMs specialized for chemistry (and their corresponding base models), and two closed frontier models (Table 4; details in Section B).

Tasks	Sources	Metrics
Comprehension		
Atom Counting	ChEBI-20	Exact match
Group identification	ChEBI-20	Macro-F1
Property estimation	ChEBI-20	ρ, Exact match
Molecule Retrieval	ChEBI-20	Top-1
Discrimination		
Isomer Discrimination	ChEBI-20	Accuracy
Tautomer Recognition	ZINC250K	Accuracy
Protonation State	ZINC250K	Accuracy
Generation		
Caption-to-Molecule	ChEBI-20	Validity rate, Tanimoto sim., Exact match, FCD

Table 3: Overview of eight benchmark tasks developed to study how different text representations affect the performance of LLMs in chemistry.

We evaluate the 16 models with 9 molecular representations across the 8 tasks. For each model-task configuration, we identify the representation with the highest mean score and then apply a paired bootstrap test (Koehn, 2004) with 10,000 resamples between the per-instance scores of every other representation and the top scorer.

5 Results

We observe a consistent pattern: *model performance is not representation-invariant*. The same molecule, expressed in different molecular representations, can lead to substantially different outcomes. This suggests that current LLMs' chemical competence is significantly influenced by the

Family	#Param	Model	R
Qwen3	4B	Qwen3-4B	✓
		Qwen3-4B	✗
	30B	Qwen3-30B-A3B	✓
		Qwen3-30B-A3B	✗
Phi-4	14B	Phi-4	✓
		Phi-4-Reasoning	✓
		Phi-4-Reasoning-Plus	✓
OLMo	32B	OLMo-3.1-32B-Instruct	✗
		OLMo-3.1-32B-Think	✓
Specialized	14B	ChemDFM-v2.0-14B	✗
		ChemDFM-R-14B	✓
	24B	Ether0	✓
Base models of specialized	14B	Qwen2.5-14B	✗
	24B	Mistral-Small-24B	✓
Closed frontier	Unknown	GPT-5.4-Mini	✓
		Claude-Haiku-4.5	✓

Table 4: All LLMs evaluated across five model families. The last column indicates if reasoning capability is present (✓) or absent (✗). Qwen3 supports both modes whereas other models only support one. We also include the base model of the ChemDFMs (resp. Ether0), which is Qwen2.5-14B (resp. Mistral-Small-24B). Further details are provided in Section B.

molecular text representation used. In addition, we notice that molecular text representations' embeddings from different LLMs across the nine representations show clear clustering amongst the different representations, indicating that the models fail to build representation-invariant internal encodings. We have provided the t-SNE plots of these embeddings in Appendix M.

5.1 Performance depends on representation

Our central finding is that no single molecular representation is uniformly optimal across tasks. Table 5 summarizes the number of wins per representation across models for each task (detailed tables in Section K). In aggregate, CML achieves the top score in 81 out of all 128 model-task configurations, more than any other representation. MolJSON follows with 75, InChI with 72, and then canonical SMILES with 66.

This performance ranking is surprising. Representations predominant in LLM pretraining, such as the SMILES variants, are not the best. Instead CML and MolJSON are the top two. MolJSON is introduced recently (April 2026), so highly unlikely to be exposed to the LLMs we use (all released before April 2026). This contributes evidence that an LLM's performance on molecular tasks is not limited by its familiarity with a certain representation but by its ability to parse the implicit structure.

In atom counting (Table 6), the SMILES variants achieve high accuracies of 0.80 to 0.90. Even so, InChI, CML and MolJSON surpass them: InChI wins with 12 LLMs, followed by CML and MolJSON at 9 each. This could be due to the explicit mention of atom counts in InChI's formula layer (e.g., C9H8O4), whereas CML and MolJSON only mention each atom explicitly. SELFIES is the worst, losing with all 16 LLMs.

Functional-group identification is led by CML, followed by both IUPAC and MolJSON in 4 models (Tables 5 and 7). Two patterns arise here: CML supports graph reasoning, which is helpful in finding functional groups that are substructures in the molecular graph, while IUPAC exposes the functional groups directly through its functional group morphemes (e.g. carboxylic acid, ester, amide).

In property estimation (Tables 8 to 11), SMILES variants and IUPAC dominate in LogP and TPSA regression while being average in hydrogen bond donor (HBD) and acceptor (HBA) counts. For HBA and HBD counts, structured representations dominate. With Qwen3-4B, CML is twice as good as SMILES, increasing accuracy to 0.79 from 0.30; with GPT-5.4-Mini, to 0.98 from 0.81. MolJSON, while generally trailing CML, is better than the other representations in the HBA and HBD counts. The results reinforce the intuition that the structured text representations are apt at tracking the various atoms and bonds while SMILES and IUPAC are apt for the regression tasks.

Structured text representations, CML and MolJSON, dominate isomer discrimination and tautomer recognition (Tables 5, 13 and 20). These tasks require graph understanding and fine-grained structural disambiguation, which are both facilitated by CML and MolJSON. Canonical and isomeric SMILES are the top two representations for protonation-state recognition: SMILES variants present charges as visible characters, for example `CC(=O)O` (acid) versus `CC(=O)[O-]` (anion). IUPAC and SELFIES are the worst and collapse to random performance with many LLMs. These two do not explicitly mention the charge: IUPAC names are neutral by definition, and SELFIES grammar enforces valence constraints by construction.

However, IUPAC shines in molecule retrieval (Table 12): it is the best representation for all the 16 LLMs. Canonical SMILES and isomeric SMILES

Representation	Atom Count.	Func. Groups	Property Estimation				Retrieval	Isomer Disc.	Cap.-to-Mol.	Tautomer	Protonation	**Total**
			logP	TPSA	HBD	HBA						
Canonical SMILES	0	3	7	7	3	5	13	0	9	7	12	66
Isomeric SMILES	0	2	10	7	2	6	11	0	8	5	12	63
Randomized SMILES	0	2	5	5	3	4	7	4	7	0	1	38
DeepSMILES	0	1	3	3	0	4	8	0	0	3	8	30
IUPAC	0	4	16	9	3	3	16	0	9	3	1	64
SELFIES	0	0	2	3	1	0	0	0	3	3	1	13
CML	9	9	2	6	13	8	1	14	0	13	6	81
InChI	12	0	9	8	1	15	9	1	0	9	8	72
MolJSON	9	4	3	7	0	13	3	13	4	12	7	75

Table 5: Number of model configurations where each representation is among the best (paired bootstrap, 95% confidence interval). Ties award a point to every indistinguishable representation.

trail closely behind, with 13 and 11 LLMs, respectively. IUPAC's dominant performance highlights its morphemic structure, which is directly related to the natural language descriptions of molecules present in molecule retrieval questions. Also, IUPAC performs the best in caption-to-molecule generation alongside the SMILES variants (see also Section 5.3).

Therefore, the optimal representation depends on the task. Counting and structural-disambiguation tasks benefit from explicit, addressable formats (CML, MolJSON, InChI), while description-matching tasks favour IUPAC.

5.2 Chemistry-Specialized Post-Training limits robustness across representations

To understand the capabilities of chemistry-specialized models and isolate the effect of domain post-training from the base model's general capability, we compare the specialized models with their respective base models. We find that chemistry-specialized models do not necessarily perform better than their base models across representations, which surfaces the need for reporting performance across multiple representations.

ChemDFM-R-14B with canonical SMILES is the best in the entire study for caption-to-molecule generation in terms of exact match (Table 15), validity rate (Table 14), and FCD (Table 19). On SMILES representations, the ChemDFMs improve across tasks over its base model, Qwen2.5-14B: on molecule retrieval, accuracy rises from 0.27 to 0.72 (resp. 0.80) for ChemDFM-v2.0 (resp. ChemDFM-R); on tautomer recognition, from 0.26 to 0.88 (resp. 0.92); and on protonation state recognition, from 0.28 to 0.88 (resp. 0.95).

However, these improvements are not uniform across representations. The ChemDFMs score near zero with CML and InChI for molecule generation. This suggests that the fine-tuning data, likely dominated by SMILES, strengthens SMILES-based chemical reasoning at the cost of generalization to markup and line-notation formats. Compared to Qwen2.5-14B in property estimation, ChemDFM-v2.0 is worse with MolJSON, CML and especially InChI. In atom counting, ChemDFM-v2.0 is worse with IUPAC, SELFIES, and especially InChI.

Hence, fine-tuning Qwen2.5-14B seems to erode structured-format understanding for a boost in SMILES fluency, and the poor performance is most visible on tasks where the model needs to parse the representations rather than just pattern-match. This may be attributed to the *catastrophic forgetting* when LLMs are fine-tuned or post-trained on tasks with a shift in distribution, making them unable to retain past knowledge or capability (Kirkpatrick et al., 2017; Liao et al., 2025).

Similarly, Ether0-24B performs worse than its base model (Mistral-Small-24B) on essentially every task in our benchmark. Details in Section F.

Therefore, evaluating chemistry LLMs on SMILES-only tasks systematically rewards specialization that does not generalize. Reporting performance across multiple representations, as we do here, is necessary to surface trade-offs.

5.3 Caption-to-molecule differentiates representations best

Caption-to-molecule generation (Tables 14–19) requires constructing an exact molecular graph from a natural language description. Under the paired bootstrap analysis (Table 5), canonical SMILES and IUPAC jointly lead (9 points), followed by isomeric SMILES (8) and randomized SMILES (7). Three key trends emerge from this task.

First, validity does not imply structural correctness. ChemDFM-R-14B produces 97.6% valid molecules with canonical SMILES, yet its Morgan Tanimoto similarity is 0.68 (Table 16). Similarly, GPT-5.4-Mini produces 91.6% valid molecules

with canonical SMILES, yet its Morgan Tanimoto similarity with canonical SMILES is 0.61.

Next, robustness across representations distinguishes frontier from specialized models. ChemDFM-R-14B's good results with canonical SMILES come at the cost of complete failure with CML, MolJSON and InChI, where it produces no valid generations. GPT-5.4-Mini's canonical SMILES scores are slightly lower compared to ChemDFM-R-14B but the model produces non-trivial valid generations for structured text representations (Morgan Tanimoto similarity for CML is 0.24; for MolJSON, 0.30). Hence, frontier models are more robust in generation, while ChemDFM-R's good results are limited to SMILES.

Third, IUPAC's natural language and morphemic structure is particularly relevant to molecule generation. As mentioned previously, the ChemDFMs achieve the best across the generation metrics when using canonical SMILES. However, ChemDFMs' performance with other representations drops severely. This could be attributed to ChemDFMs being post-trained heavily on chemical text that contains a large amount of SMILES strings. Meanwhile, across different generation metrics and non-ChemDFM models, IUPAC gives the best performance. This is due to ChEBI-20's captions referencing systematic-name fragments that are directly relevant to the IUPAC nomenclature.

5.4 LLM-as-a-judge reveals major sources of error in generation across representations

While validity rate, exact match, and the Tanimoto similarity metrics give an understanding of the quality and correctness of the generations, we also seek to know exactly where each model goes wrong with different representations. We use Gemini 3 Flash to judge the generations for 10 captions from ChEBI-20, for all models and representations. The LLM-as-a-judge categorizes errors into five types: stereochemistry, missing substituents, wrong scaffold, hallucination and/or syntax error (Figure 1). It scores the generations against representation faithfulness, chemical reasoning, and overall quality, each on a 1–to–5 scale (Figure 2). Despite the small sample size, we can glean some insights into the errors for each representation.

IUPAC leads the performance with 50% correct generations, while SELFIES and DeepSMILES are the worst. Canonical SMILES has the highest representation faithfulness but relatively low overall quality; it has a significant issue with wrong scaffolds and chemical reasoning, despite SMILES being most represented in chemical training data. When prompted to generate in CML, DeepSMILES or SELFIES, LLMs score low on representation faithfulness, with syntax being the major error. IUPAC is the only representation where both reasoning and faithfulness scores are high, so we should prioritise IUPAC names for molecule generation. MolJSON has a higher representation faithfulness score while it closely trails IUPAC in terms of chemical reasoning and overall quality.

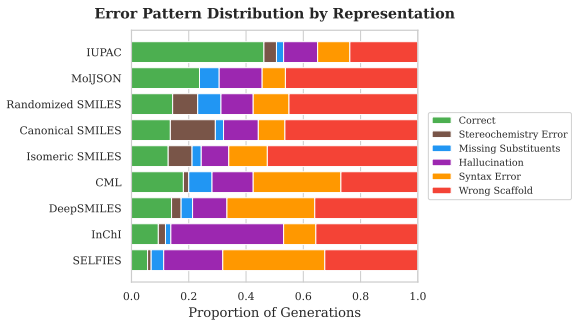


Figure 1: Proportion of error modes in molecule generation across various representations as determined by Gemini 3 Flash.

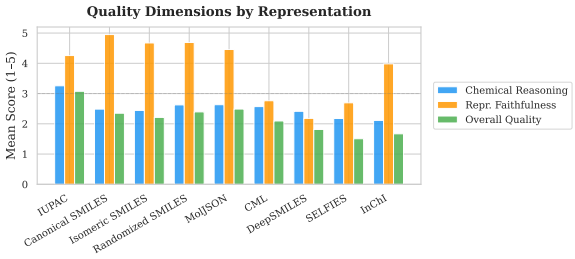


Figure 2: Gemini 3 Flash's scoring (from 1 to 5) across representations in the molecule generation task, based on faithfulness, chemical reasoning, and overall quality.

While the hallucination errors are less frequent compared to other error types, they remain a major issue with current LLMs (Rawte et al., 2023). For instance, ChemDFM-R-14B just generates a string of multiple 'C's, and OLMo-3.1-32B-Think, when asked for a CML output, outputs a malformed SMILES string. On the other hand, Ether0 wrongly identifies pneumocandin, an antifungal drug (Denning, 1997), as a chemical weapon and refuses to proceed with the response. Given that hallucination errors are diverse, using LLM-as-a-judge in addition to rigid metrics helps to analyse the qualitative nature of these errors.

Finally, reasoning LLMs reduce hallucination

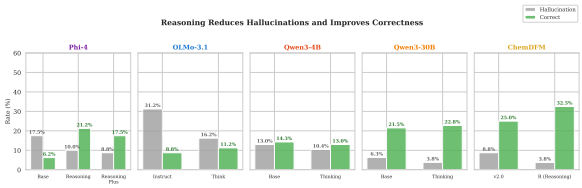


Figure 3: Reasoning models reduce hallucination.

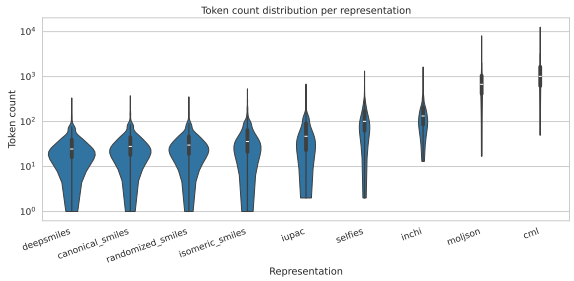


Figure 4: Token count distribution across 9 representations using Qwen3's tokenizer.

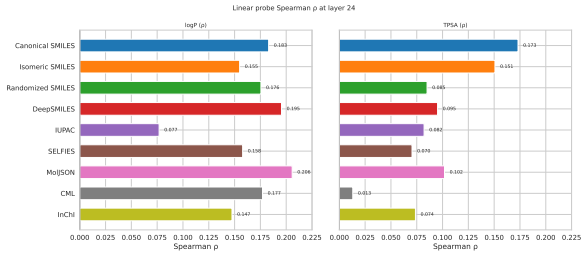


Figure 5: Linear probing performance of Qwen3-4B across all representations for molecular weight, log P, and TPSA prediction

(Figure 3). For OLMo, the reasoning variant halves the hallucination rate to 16.2% from the non-reasoning variant's 31.2%; and for ChemDFM, the reasoning variant halves the hallucination rate to 3.8% from the non-reasoning variant's 8.8%. Moreover, reasoning LLMs also lead to more correct responses generally. The observations carry over to Phi-4 and Qwen3 models, though the improvements are not as pronounced.

6 Mechanistic Analysis

To explain the representation-dependence in performance in Section 5, we conduct mechanistic analysis using the Qwen3-4B (36 transformer layers and 151,669 token vocabulary). We use this model as it has open weights, it is small, and all its parameters are active. We conduct (1) a tokenization audit that examines how each representation is segmented, (2) linear probing of hidden states to measure what chemical information the model encodes internally, and (3) attention analysis to locate where the model attends for each representation.

6.1 Tokenization Audit

We tokenize 1,000 molecules from ChEBI-20 in all nine representations and measure the total token count of the actual molecular representation and the length of reasoning tokens.

Representations vary by an order of magnitude in token cost. DeepSMILES is the most compact (median 25 tokens), followed by canonical SMILES (28), randomized SMILES (30), and isomeric SMILES (36). IUPAC names require 47 tokens on average. SELFIES is 3.6 times longer than canonical SMILES (median 101). InChI, MolJSON and CML are the longest. Such representation lengths are costly given fixed context windows and quadratic attention scaling. However, the reasoning token length presents a trade-off (Figure 6): explicit, verbose representations like CML and MolJSON result in shorter reasoning token length.

6.2 Linear Probing of Hidden States

Using the same 1,000 molecules from Section 6.1, we use linear probes to see if the hidden embeddings can already estimate molecular properties. This can elicit if the representation already provides sufficient signals in the intermediate layers before the final predictions. We train linear probes (ridge regression for continuous targets) on the hidden states of each layer to estimate three molecular properties: molecular weight (MW), LogP and TPSA. We use the intermediate embeddings (the 24th layer) as inputs to train the probes.

The SMILES variants have some of the best performances. This is expected, given that they are likely present during pre-training. However, MolJSON, which is created after Qwen3-4B was released, results in the best performances. CML, though an explicit graph-based representation like MolJSON, underperforms to varying degrees on all probing tasks, indicating that the internal representations from the XML format are not sufficiently informative. However, the performance of Qwen3-4B when prompted on logP and TPSA estimation leads to much higher Spearman correlation across all representations (Tables 9 and 8). This disconnect between benchmark performance and linear probe performance suggests two points: (1) LLMs can better estimate molecular properties through

the generation process during inference instead of storing them as linearly accessible features in its hidden states (Belinkov, 2022; Gao et al., 2024), and (2) the representation choice affects the LLM's ability to leverage encoded knowledge during generation as opposed to just impacting the richness of the encodings.

6.3 Attention analysis

We extract attention weights from all 36 layers and 32 heads for 50 selected molecules across all representations. We measure two quantities: (i) *last token to molecule attention* (Figure 7), which is the fraction of the final prediction token's attention towards the molecule tokens, and (ii) *within-molecule attention* (Figure 8), which is the self-attention amongst the molecule's tokens. We perform this analysis from layers 17 to 24. We chose these layers as they form the intermediate layers and exhibit reasonable performance during linear probing.

CML and MolJSON have the largest last token to molecule attention, nearly 5 to 10 times that of canonical SMILES. This is consistent with the long length of the explicit graph representations: the model must look further back to gather molecular information. By contrast, SMILES variants have 20 times larger within-molecule attention compared to MolJSON and CML, suggesting that compact representations encourage richer token-to-token interaction within the molecular string.

7 Conclusion

We introduced MolRepBench, a systematic evaluation of nine molecular text representations across eight tasks and 16 LLMs spanning open-weight, reasoning-tuned, domain-specialized, and closed frontier models. Our results show that **LLM-based chemical reasoning is not representation-invariant**. The same molecule, expressed in different molecular representations, can yield significantly different outcomes. CML and MolJSON dominate tasks that require understanding of molecular graphs: atom counting, isomer discrimination, and tautomer recognition. IUPAC names dominate semantically grounded tasks such as molecule retrieval and caption-to-molecule generation, where their morphemic structure highlights key chemical units. Despite SMILES prevalence in pretraining corpora, they are rarely optimal.

Our findings suggest to adopt the following: (1) benchmarks should report performances across multiple molecular representations, (2) an LLM-judge should be used to qualitatively analyse the errors (in LLM outputs) beyond quantitative metrics such as validity rate and Tanimoto similarities, and (3) molecular applications should select the molecular representation based on the task.

Limitations

Our goal is to understand how LLMs use the various text representations to solve various tasks, and we have strived to cover a wide number of chemistry tasks and representations relevant to current cheminformatics tools and LLMs. While more chemistry tasks can be added to enhance coverage, our benchmark have already led to clear and actionable insights that could enable robust use of LLMs for chemistry. The study have solely focused on the use of text representations with LLMs and have not explored multi-modal models, which are becoming more common. Future work can understand how multi-modal models perform when given 1D, 2D and 3D representations of molecules.

Ethics Statement

Our work does not involve any human subjects. We use publicly available datasets, open-source LLMs, and openly accessible frontier LLMs. We intend to understand the effect of molecular text representations with LLMs, and there is potential for dual use. Our findings could aid medicinal chemists working on drug discovery or nefarious users trying to create harmful molecules.

Acknowledgements

A.R.'s PhD program is supported by the Agency for Science, Technology, and Research and the SABS R3 CDT program via the Engineering and Physical Sciences Research Council. This work was done during A.R.'s internship at DSO National Laboratories as part of the DSO-AISG Incentive Award. We would like to thank DSO and AI Singapore for the computational resources, which played a significant role in this research. We would also like to thank Dr Hongtao Zhao, Dr Christian Tyrchan, Dr Eva Nittinger, Prof. Charlotte M. Deane, and Prof. Michael M. Bronstein for their advice.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen,

Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and Guoqing Zheng. 2025. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Maged S. Al-shaibani and Irfan Ahmad. 2023. Consonant is all you need: a compact representation of English text for efficient NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11578–11588, Singapore. Association for Computational Linguistics.

Anthropic. 2025. Introducing Claude Haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5. Accessed: 2026-05-13.

George Arthur Baker, Mario Sanz-Guerrero, and Katharina von der Wense. 2025. Molecular string representation preferences in pretrained llms: A comparative study in zero- & few-shot molecular property prediction. In *Conference on Empirical Methods in Natural Language Processing*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Guy W. Bemis and Mark A. Murcko. 1996. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39 15:2887–93.

Melanie Burger. 2015. Chemdoodle web components: Html5 toolkit for chemical graphics, interfaces, and informatics. *Journal of cheminformatics*, 7:35.

Matthias Busch, Marius Tacke, Sviatlana V. Lamaka, Mikhail L. Zheludkevich, Christian J. Cyron, Christian Feiler, and Roland C. Aydin. 2026. In-context molecular property prediction with llms: A blinding study on memorization and knowledge conflicts.

Yuxing Cheng, Yi Chang, and Yuan Wu. 2025. A survey on data contamination for large language models.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6140–6157. PMLR.

Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. 2020. Molecular representations in ai-driven drug discovery: a review and practical guide. *Journal of Cheminformatics*, 12.

Mark Davies, Michał Nowotka, George Papadatos, Nathan Dedman, Anna Gaulton, Francis Atkinson, Louisa Bellis, and John P. Overington. 2015. Chembl web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Research*, 43(W1):W612–W620.

D W Denning. 1997. Echinocandins and pneumocandins–a new antifungal class with a novel mode of action. *Journal of Antimicrobial Chemotherapy*, 40(5):611–614.

Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. 2002. Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42 6:1273–80.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Muhan Gao, TaiMing Lu, Kuai Yu, Adam Byerly, and Daniel Khashabi. 2024. Insights into LLM long-context failures: When transformers know but don't tell. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7611–7625, Miami, Florida, USA. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 3 flash model card. Technical report, Google DeepMind. Model released December 17, 2025.

Taicheng Guo, kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In *Advances in Neural Information Processing Systems*, volume 36, pages 59662–59688. Curran Associates, Inc.

Sonakshi Gupta, Akhlak Mahmood, Shivank Shukla, and Rampi Ramprasad. 2025. Benchmarking large language models for polymer property predictions. *Macromolecular Rapid Communications*.

Marcus D. Hanwell, Wibe A. de Jong, and Christopher J. Harris. 2017. Open chemistry: Restful web apis, json, nwchem and the modern web application. *Journal of Cheminformatics*, 9.

John J. Irwin, T. Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. 2012. Zinc: A free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52:1757 – 1768.

Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. 2024. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6:161 – 169.

Ryan Jacobs, Maciej P. Polak, Lane E. Schultz, Hamed Mahdavi, Vasant Honavar, and Dane Morgan. 2026. Regression with large language models for materials and molecular property prediction.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. 2020. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Greg Landrum. 2016. Rdkit: Open-source cheminformatics software.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. 2024. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6071–6083.

Chonghua Liao, Ruobing Xie, Xingwu Sun, Haowen Sun, and Zhanhui Kang. 2025. Exploring forgetting in large language model pre-training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2112–2127, Vienna, Austria. Association for Computational Linguistics.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. 2023. Multi-modal molecule structure-text model for text-based retrieval and editing. *Nature Machine Intelligence*.

Micha Livne, Zulfat Miftahutdinov, Elena Tutubalina, Maksim Kuznetsov, Daniil Polykovskiy, Annika Brundyn, Aastha Jhunjhunwala, Anthony Costa, Alex Aliper, Alán Aspuru-Guzik, and Alex Zhavoronkov. 2024. nach0: multimodal natural and chemical languages foundation model. *Chemical Science*, 15(22):8380–8389.

Alan McNaught. 2006. The iupac international chemical identifier: Inchl-a new standard for molecular informatics. *Chemistry international*, 28:12–14.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, Amir Mohammad Elahi, Mehrdad Asgari, Juliane Eberhardt, Hani M. Elbeheiry, María Victoria Gil, Maximilian Greiner, Caroline T. Holick, Christina Glaubitz, Tim Hoffmann, Abdelrahman Ibrahim, Lea C. Klepsch, Yannik Köster, Fabian Alexander Kreth, Jakob Meyer, Santiago Miret, Jan Matthias Peschel, Michael Ringleb, Nicole Roesner, Johanna Schreiber, Ulrich S. Schubert, Leanne M. Stafast, Dinga Wonanke, Michael Pieler, Philippe Schwaller, and Kevin Maik Jablonka. 2024. Are large language models superhuman chemists?

Mistral AI. 2025. mistralai/Mistral-Small-24B-Instruct-2501. `https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501`. Accessed: 2025-04-22.

Peter Murray-Rust and Henry Rzepa. 2000. Chemical markup, xml, and the worldwide web. 1. basic principles. *Journal of Chemical Information and Computer Sciences*, 39.

Peter Murray-Rust and Henry Rzepa. 2003. Chemical markup, xml, and the world wide web. 4. cml schema. *Journal of chemical information and computer sciences*, 43:757–72.

Siddharth M. Narayanan, James D. Braza, Ryan-Rhys Griffiths, Albert Bou, Geemi Wellawatte, Mayk Caldas Ramos, Ludovico Mitchener, Samuel G. Rodriques, and Andrew D. White. 2025. Training a scientific reasoning model for chemistry.

Noel M. O'Boyle and Andrew Dalke. 2018. Deepsmiles: An adaptation of smiles for use in machine-learning of chemical structures. *ChemRxiv*.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang,

Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2026. Olmo 3.

OpenAI. 2026. Introducing GPT-5.4. Released March 5, 2026.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. 2018. Fréchet chemnet distance: A metric for generative models for molecules in drug discovery.

Shampa Raghunathan and U. Deva Priyakumar. 2022. Molecular representations for machine learning applications in chemistry. *International Journal of Quantum Chemistry*, 122(7):e26870.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Jean Rigaudy and S. P. Klesney. 1979. Nomenclature of organic chemistry: sections a, b, c, d, e, f and h - 1979 ed.

David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50 5:742–54.

Nicholas T. Runcie, Fergus Imrie, and Charlotte M. Deane. 2026. Molecular representations for large language models.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Matthew Swain. 2018. CommonChem: A data format for chemical information.

Taffee T Tanimoto. 1957. Ibm internal report. *Nov*, 17:1957.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science.

Y. Wang, J. Xiao, T. Suzek, J. Zhang, J. Wang, and S. Bryant. 2009. Pubchem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*, 37(Web-Server-Issue):623–633.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36.

Daniel S. Wigh, Jonathan M. Goodman, and Alexei A. Lapkin. 2022. A review of molecular representation in the age of machine learning. *WIREs Computational Molecular Science*, 12(5):e1603.

William J. Wiswesser. 1954. *A Line-Formula Chemical Notation*. Crowell, New York.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: A benchmark for molecular machine learning.

Xinnuo Xu, Rachel Lawrence, Kshitij Dubey, Atharva Pandey, Risa Ueno, Fabian Falck, Aditya V. Nori, Rahul Sharma, Amit Sharma, and Javier Gonzalez. 2025. Re-imagine: Symbolic benchmark synthesis for reasoning evaluation.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Gregory Yauney and David Mimno. 2021. Comparing text representations: A theory-driven approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5527–5539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Botao Yu, Frazier N. Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2022. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature Communications*.

Haiteng Zhao, Shengchao Liu, Chang Ma, Hannan Xu, Jie Fu, Zhihong Deng, Lingpeng Kong, and Qi Liu. 2023. GIMLET: A unified graph-text model for instruction-based molecule zero-shot learning. *Advances in Neural Information Processing Systems*.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, Kai Yu, and Xin Chen. 2024. Developing ChemDFM as a large language foundation model for chemistry. *Cell Reports Physical Science*. ArXiv:2401.14818.

Zihan Zhao et al. 2025. ChemDFM-R: A chemical reasoning LLM enhanced with atomized chemical knowledge. *arXiv preprint arXiv:2507.21990*.

Zhiqiang Zhong, Kuangyu Zhou, and Davide Mottin. 2024. Benchmarking large language models for molecule prediction tasks. *ArXiv*, abs/2403.05075.

A Structured molecular representations

Structured representations of molecules are understudied in machine learning. JavaScript Object Notation (JSON) and Extensible Markup Language (XML) have been widely used for storing hierarchically structured data, providing schema-validated, human-readable, and programming language-agnostic storage formats. They have become widely used standards for data interchange on the internet and between scientific tools. In chemistry, JSON and XML formats have been developed for various purposes. Chemical Markup Language or CML (Murray-Rust and Rzepa, 2000) was the first domain-specific application of XML, encoding a molecule as nested elements, namely, `atomArray`, `bondArray`, and per-atom or per-bond attributes. This XML-based format makes every chemical component explicit and machine-parseable, so that chemical data could be represented or stored properly on web platforms.

The cheminformatics community has shifted towards JSON for the same data interchange role, motivated by lighter parsers and native compatibility with web services and document stores. Chemical JSON (Hanwell et al., 2017), built based on CML, is the native representation of Avogadro 2. It stores a molecule as property arrays that map directly onto in-memory data structures and can be stored efficiently in binary JSON for database storage. Toolkit-oriented schemas such as CommonChem's JSON (Swain, 2018) optimise for representation conversions between different cheminformatics libraries, while the ChemDoodle JSON representation (Burger, 2015) is meant for rendering molecular graphics on web browsers. Public databases like PubChem (Wang et al., 2009) and ChEMBL (Davies et al., 2015) also expose molecules as JSON, though their schemas are optimized for retrieving data such as property tables and assay records. Recently, Runcie et al. (2026) introduced another JSON representation for LLMs, MolJSON, and we investigate this in our benchmarking study.

B Model families

The open-weight models are served locally via vLLM (Kwon et al., 2023) on H200 GPUs, while closed models are accessed via their respective API.

Qwen3 (by Alibaba). The Qwen3 series (Yang et al., 2025) supports a switchable reasoning mode: `enable_reasoning=True` activates chain-of-thought reasoning within `<think>` tags; `enable_reasoning=False` gives direct responses. We evaluate Qwen3-4B (dense) and Qwen3-30B-A3B (Mixture-of-Experts model, 3B active parameters) under *both* reasoning conditions, providing a clean within-model reasoning ablation.

Phi-4 (by Microsoft). The Phi-4 family (Abdin et al., 2024, 2025) has three 14B models at different levels of reasoning: Phi-4, Phi-4-Reasoning (RL post-trained for reasoning), and Phi-4-Reasoning-Plus (stronger RL).

OLMo-3 (by AI2). The OLMo-3 series (Olmo et al., 2026) provides fully open-weight 32B models. We include instruction-tuned OLMo-3.1-32B-Instruct and the reasoning variant OLMo-3.1-32B-Think.

Domain-specialized models. ChemDFM or Dialogue Foundation Model for Chemistry (Zhao et al., 2024) is a chemistry-specialist LLM family which consists of ChemDFM-v2.0-14B and ChemDFM-R-14B (Zhao et al., 2025). These models were obtained by finetuning Qwen2.5-14B (Yang et al., 2024) on chemical literature and molecular data. In addition, we also benchmark reasoning-capable Ether0 (Narayanan et al., 2025) which is based on Mistral-Small-24B-Instruct-2501 (Mistral AI, 2025). We also benchmark the base LLMs of these domain-specialized models, namely Qwen2.5-14B and Mistral-Small-24B. Benchmarking such chemistry-specialized LLMs allows us to isolate the effect of domain-specialized LLMs from general LLMs.

Closed frontier models. GPT-5.4-Mini (OpenAI, 2026) and Claude-Haiku-4.5 (Anthropic, 2025) have been selected as closed-source frontier models for benchmarking. They were selected as they were two of the top mid-tier models by big players in the industry, namely, OpenAI and Anthropic. Their input token costs between \$0.75 to \$1.00 per 1M tokens based on which we could make a fair comparison and these versions allowed us to keep our API costs within the limited budget we were operating with.

C Source datasets

ChEBI-20 pairs about 33,000 molecules with expert-written natural language descriptions and provides standard train/validation/test splits (Edwards et al., 2022). We use the test split as the source of molecules for our tasks. From the ChEBI-20 test split ($\sim$3.3K molecules), we select a stratified subset of 200 molecules for the comprehension benchmarks. We apply a complexity cut-off before sampling: molecules must have a molecular weight $\geq$ 300 Da and at least 2 rings. This cut-off excludes trivially simple molecules (e.g. small acyclic fragments, single amino acids) where all representations may succeed equally, ensuring that the benchmark discriminates between representations on structurally meaningful cases.

For comprehension tasks (atom counting, functional group identification, property estimation), we use 200 molecules. For retrieval we use 78 molecules (each appearing as the correct answer once). Each retrieval question needs three distractor types (scaffold match, similar molecular weight, random). Not every molecule in the comprehension subset has another molecule in the test set that shares its Murcko scaffold (Bemis and Murcko, 1996). Therefore, we ended up with a small dataset of 78 questions for this task, but it was sufficient to evaluate all models and draw clear insights. For isomer discrimination we construct 500 pairs. For caption-to-molecule generation, we use 200 molecules from the full test set.

ZINC250K is a curated, drug-like subset of the ZINC database (Irwin et al., 2012) containing 250,000 commercially available molecules. Unlike ChEBI-20, ZINC250K lacks natural language descriptions but provides a much larger pool of structurally diverse molecules. We use it to construct paired discrimination datasets (tautomer recognition and protonation state recognition) that require *generating* chemically related variants of existing molecules as multiple choices for the questions. The large, diverse pool of drug-like molecules in ZINC250K helps with the generation of these variants.

D Tasks

Atom Counting. Given a molecular string, the LLM is to count the number of atoms of a specified element (C, N, O, S, F, or Cl). The ground truth is computed from RDKit atom iteration. We measure performance by exact match accuracy.

Functional Group Identification. For each molecule, we prompt the LLM to identify the presence of different functional groups: aldehyde, ester, halide, primary amine, and sulfonamide. This is a binary classification task, where the LLM indicates the presence of each functional group with a *yes* or a *no*. The ground truth is detected via RDKit SMARTS matching, and the performance is measured by macro-averaged F1 across all 5 functional groups.

Molecular Property Estimation. The LLM is prompted to predict four different RDKit-computed properties when given each of the 9 representations separately. The properties are Wildman-Crippen LogP value, Topological Polar Surface Area and H-bond donor and acceptor counts, evaluated by exact match accuracy.

Molecule Retrieval. Given a ChEBI natural language description, identify the correct molecule from four candidates. This is a multiple-choice question task. The three distractors are selected based on the following factors: one shares the Murcko scaffold (same core, different substituents) (Bemis and Murcko, 1996), one has a similar molecular weight ($\pm$10%), and one is random. The metric used is top-1 accuracy.

Isomer Discrimination. The LLM, when given two molecular strings, has to determine whether they encode the same molecule or not, which is a binary classification task. We construct 500 pairs split across three types: positive pairs (same molecule but different atom orderings), stereoisomer pairs (molecules differing only in R/S or E/Z configuration), and substitution pairs (one atom changed). The stereoisomer pairs specifically probe whether representations that encode stereocenters help models detect chirality differences.

Tautomer Recognition. Here the task is to determine whether two given molecular strings are tautomeric forms of the same molecule. This is a binary classification task. Two molecules are tautomers if and only if they map to the same RDKit canonical tautomer. It contains 250 molecule pairs (125 positive: confirmed tautomeric forms; 125 negative: chemically confusable non-tautomers). Positive pairs span four tautomer classes: keto/enol, amide/imidic acid, heterocyclic, and nitroso/oxime, identified via RDKit's `TautomerEnumerator`. Negative pairs are

unrelated molecules that do not share a canonical tautomer.

Protonation State Recognition. Given two molecular strings from ZINC250K, the LLM has to determine whether they represent the same molecule at different protonation states, which is a binary classification task. It contains 210 pairs total. Two molecules are considered protonation variants if and only if their RDKit-neutralized canonical SMILES are identical. Negatives include molecules with similar charge patterns but genuinely different structures. This task contains 210 molecule pairs (105 positive pairs, which are same molecule at different protonation states and 105 negative pairs). Positive pairs cover carboxylic acid/carboxylate, amine/ammonium, phenol/phenolate, and zwitterion variants generated via RDKit's `MolStandardize` module (Landrum, 2016). Ground truth uses RDKit's `Uncharger`-based neutralization: two molecules are protonation variants if and only if they share the same neutral canonical SMILES.

Caption-to-Molecule Generation. This task involves generating the molecular string in the target representation when given a natural language molecular description or caption from the ChEBI-20 dataset. We use the following metrics for evaluation: validity rate (fraction of the generated molecular strings that are parseable to a valid RDKit molecule), exact match (canonical SMILES of generated molecule matches ground truth), Tanimoto similarity (2048-bit Morgan fingerprint with radius 2), and Fréchet ChemNet Distance or FCD (Preuer et al., 2018).

E Prompt templates

This section lists the exact prompt templates used for each benchmark task. Placeholders are shown in `{curly braces}`. All models receive the same task prompt; the only variation across conditions is whether the model's reasoning/reasoning mode is enabled (controlled via chat template parameters or system messages, not via the prompt itself). For the generation benchmark, we use the structured output capability supported by vLLM (Section E.6).

E.1 Atom Counting

Prompt - Atom Counting

```
How many {element} atoms are in the following
molecule?
Molecule: {molecule_str}

Please reason step by step, and put
your final answer within \boxed{}.

Example format: After reasoning, conclude
with: The answer is \boxed{42}
```

Variables: `{element}` ∈ {C, N, O, S, F, Cl}; `{molecule_str}`: molecule in the target representation.

E.2 Functional Group Identification

Prompt - Functional Group Identification

```
Does the following molecule contain a
{functional_group}?
Molecule: {molecule_str}

Please reason step by step, and put
your final answer (Yes or No) within \boxed{}.

Example format: After reasoning, conclude
with: The answer is \boxed{Yes} or \boxed{No}
```

Variables: `{functional_group}` ∈ {Primary amine, Ester, Aldehyde, Sulfonamide, Halide}; `{molecule_str}`: molecule in the target representation.

E.3 Property estimation

Prompt - Property estimation

```
Estimate the {property_display} of the
following molecule.
Molecule: {molecule_str}

Please reason step by step, and put
your final answer (a number) within \boxed{}.

Example format: After reasoning, conclude
with: The answer is \boxed{3.5}
```

Variables: `{property_display}` ∈ {LogP (partition coefficient), Topological Polar Surface Area (TPSA), number of hydrogen bond donors, number of hydrogen bond acceptors}; `{molecule_str}`: molecule in the target representation.

E.4 Molecule Retrieval

Prompt - Molecule Retrieval

```
Which of the following molecules matches this
description?
Description: {description}
A: {molecule_A}
B: {molecule_B}
C: {molecule_C}
D: {molecule_D}
```

```
Please reason step by step, and put
your final answer (only the choice letter)
within \boxed{}.

Example format: After reasoning, conclude
with: The answer is \boxed{A}
```

Variables: `{description}`: natural language molecule description from ChEBI-20; `{molecule_A}`-`{molecule_D}`: four candidate molecules in the target representation (one correct, three distractors).

E.5 Isomer Discrimination

Prompt - Isomer Discrimination

```
Do the following two molecular representations
refer to the same molecule?
Molecule 1: {molecule_1}
Molecule 2: {molecule_2}

Please reason step by step, and put
your final answer (Yes or No) within \boxed{}.

Example format: After reasoning, conclude
with: The answer is \boxed{Yes} or \boxed{No}
```

Variables: Pairs are drawn from three categories: identical molecules (positive), constitutional isomers (negative-substitution), and stereoisomers (negative-stereoisomer).

E.6 Caption-to-Molecule Generation

Prompt - Caption-to-Molecule Generation

```
Given a description of a molecule, generate
the corresponding {rep_name} string.

IMPORTANT: Don't restate the full description
- get straight to generating the molecule.
Keep any reasoning brief and concise. Output
tokens are limited, so provide your final
answer as quickly as possible.

Description: {description}

Provide your final answer in the format:
Final answer: <molecule_string>
```

Prompt - MolJSON instruction

```
Your final output must be ONLY a valid JSON
object with "atoms" and "bonds" arrays.

Format:
{"atoms": [{"id": "C1", "element": "C"},
...],
"bonds": [{"source": "C1", "target": "C2",
"order": 1.0}, ...],
"charges": null, "aromatic_n_h": null}

Keep the reasoning concise and short.
Don't ramble, be decisive. As soon as
possible, output the JSON object.
```

We also provide LLMs the schema for MolJSON provided by Runcie et al. (2026) using vLLM's `StructuredOutputsParams` for constraining the model's output to a valid format (Kwon et al., 2023).

Prompt - JSON formatting for non-MolJSON

```
Your final output must be a JSON object in
this format:
{"molecule": "..."}

Where the value is the complete {rep_name}
string. Keep the reasoning concise and
short. Don't ramble, be decisive. As soon as
possible, output the JSON object.
```

E.7 Tautomer Recognition

Prompt - Tautomer Recognition

```
Are the following two molecules tautomeric
forms of the same molecule?
Molecule 1: {molecule_1}
Molecule 2: {molecule_2}

Please reason step by step, and put
your final answer (Yes or No) within \boxed{}.

Example format: After reasoning, conclude
with: The answer is \boxed{Yes} or \boxed{No}
```

E.8 Protonation State Recognition

Prompt - Protonation State Recognition

```
Are the following two molecules different
protonation states of the same molecule?
Molecule 1: {molecule_1}
Molecule 2: {molecule_2}

Please reason step by step, and put
your final answer (Yes or No) within \boxed{}.

Example format: After reasoning, conclude
with: The answer is \boxed{Yes} or \boxed{No}
```

F Ether0's performance across representations

Ether0-24B is post-trained from Mistral-Small-24B-Instruct-2501 for chemistry, yet Mistral-Small-24B outperforms Ether0 on essentially every task in our benchmark. On property estimation, Ether0 scores 0.102 in accuracy with canonical SMILES versus Mistral's 0.444. On molecule retrieval, 0.269 versus 0.513. On atom counting, 0.195 versus 0.360. On tautomer and protonation recognition Ether0 is at or below the 0.50 random baseline on every representation, while Mistral exceeds 0.95 accuracy on canonical SMILES, isomeric SMILES, MolJSON, and InChI for tautomer; and 0.99 for

protonation. Most strikingly, Ether0 fails caption-to-molecule generation across the board: validity rates are 0.00 to 0.20 (Table 14) and exact match is uniformly 0.00 (Table 15). This underperformance by Ether0 is surprising as its post-training involves tasks relevant to ours, such as IUPAC name to SMILES translation, molecular captioning with molecules from Yu et al. (2024), and functional group suggestion (Narayanan et al., 2025). The potential reasons could be that Ether0's post-training might have eroded its general molecular reasoning, which is exhibited somewhat by its base model, and it is struggling to generalize beyond its training distribution as noted by Narayanan et al. (2025).

G Reasoning is not consistently beneficial

The model families we have selected help to analyse the effect of reasoning mode from three different angles: (1) between the reasoning and non-reasoning modes of the Qwen3 models, (2) the Phi-4 models, which are all reasoning models but with progressively more reasoning supervision and, (3) the OLMo-3.1 models, which have instruction-tuned and reasoning variants.

The Qwen3 differences between their reasoning and non-reasoning variants are quite small. Qwen3-30B atom counting on canonical SMILES moves from 0.855 (non-reasoning) to 0.805 (reasoning), within confidence intervals. OLMo3.1-32B-Think vs. Instruct is largely positive on generation (MACCS tanimoto similarity with IUPAC improves from 0.312 to 0.511) but regresses on tautomer recognition across every SMILES variant (for example, with canonical SMILES it drops from 0.744 to 0.620), suggesting reasoning-style training can push the model away from competent recognition performance.

The Phi-4 family's reasoning progression also reveals some inconsistent patterns. From Phi-4 to Phi-4-Reasoning, atom counting with canonical SMILES jumps from 0.425 to 0.715 in accuracy. In TPSA property estimation, with MolJSON and InChI we see significant progress but the improvements are more stagnant with the SMILES variants. In HBD counting, there is improvement from Phi-4 to Phi-4 Reasoning across all representations but from Phi-4-Reasoning to Reasoning-Plus, there are significant drops which may point to the heavier post-training resulting in catastrophic forgetting (Kirkpatrick et al., 2017; Liao et al., 2025). However, across the three models from Phi-4 to Phi-4-Reasoning Plus, on caption-to-molecule generation with canonical SMILES, the validity rate drops from 0.712 to 0.324 to 0.356 (Table 14) while exact match rises incrementally from 0.016 to 0.048 to 0.056 (Table 15).

Therefore, across model families, enabling reasoning mode does not uniformly improve performance. In several cases, reasoning variants perform similarly to or worse than their non-reasoning counterparts. However, in the case of molecule generation, reasoning models help cut down on hallucination (Figure 3) as we investigate in Section 5.4.

H Tokenization audit diagrams

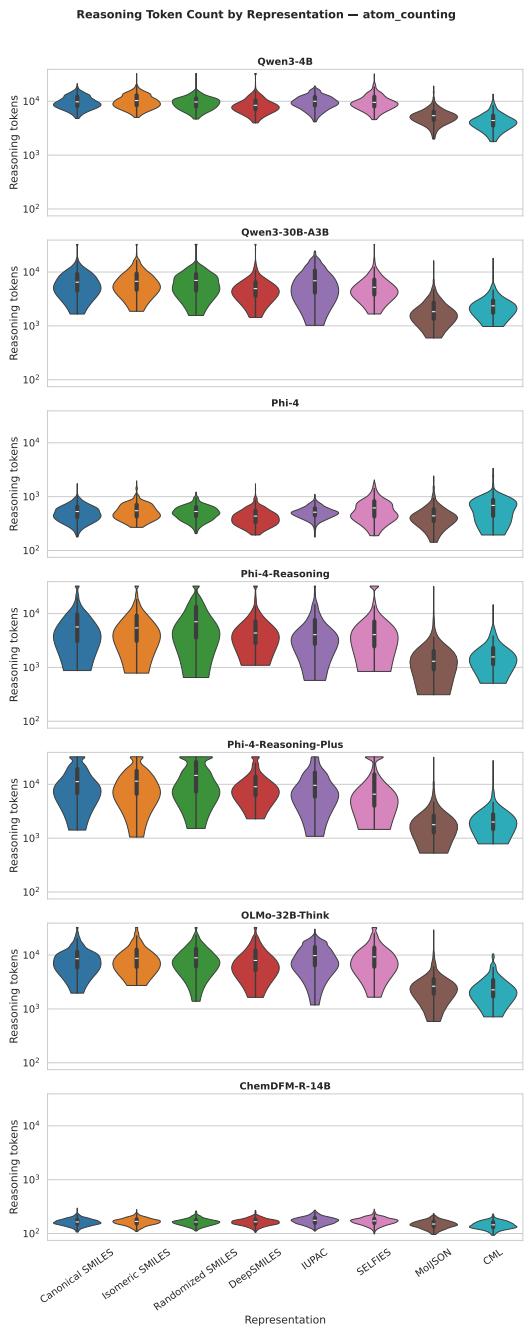


Figure 6: Reasoning token lengths for atom counting task

I Attention analysis

We extract attention weights from all 36 layers and 32 heads for 50 selected molecules across all representations. We measure two quantities: (i) *last token to molecule attention*, which is the fraction of the final prediction token's attention towards the molecule tokens, and (ii) *within-molecule attention*, which is the self-attention amongst the molecule's

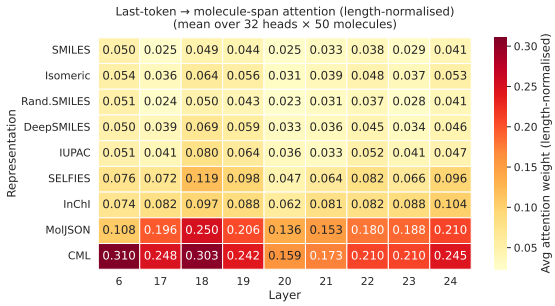


Figure 7: Attention between the last token and the molecule (length-normalised)

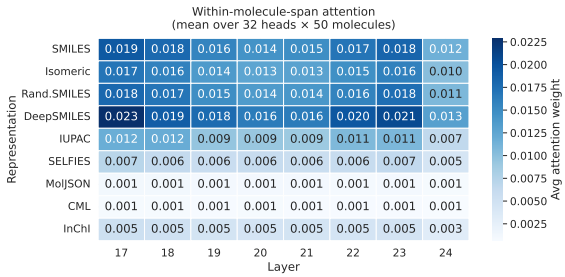


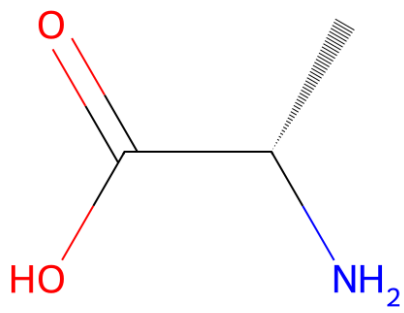
Figure 8: Attention within the molecule

tokens. We perform this analysis from layers 17 to 24. We chose this range of layers as they form the intermediate layers and exhibit reasonable performance during linear probing.

CML and MolJSON have the largest last token to molecule attention, nearly 5 to 10 times that of canonical SMILES. This is consistent with the long length of the explicit graph representations: the model must look back further to gather molecular information compared to other representations. By contrast, SMILES variants have 20 times larger within molecule attention compared to MolJSON and CML, suggesting that compact representations encourage richer token-to-token interaction within the molecular string.

J Molecular text representations

This section lists the nine molecular representations used in this work, illustrated for L-Alanine ($C_3H_7NO_2$).



Canonical SMILES

```
CC(N)C(=O)O
```

Isomeric SMILES

```
C[C@H](N)C(=O)O
```

Randomized SMILES

```
C(C(=O)O)(N)C
```

DeepSMILES

```
CCN)C=O)O
```

IUPAC

```
(2S)-2-aminopropanoic acid
```

SELFIES

```
[C][C][Branch1][C][N][C][=Branch1][C][=O][O]
```

MolJSON


```
{"atoms": [
  {"id":"C1","element":"C"}, {"id":"C2","element":"C"}, {"id":"N1","element":"N"}, {"id":"C3","element":"C"}, {"id":"O1","element":"O"}, {"id":"O2","
      element":"O"}
],
"bonds": [
  {"source":"C1","target":"C2","order":1.0}, {"source":"C2","target":"N1","order":1.0}, {"source":"C2","target":"C3","order":1.0}, {"source":"C3","
      target":"O1","order":2.0}, {"source":"C3","target":"O2","order":1.0}
]}
```


CML


```
<molecule spinMultiplicity="2">
 <atomArray>
  <atom id="a1" elementType="C" hydrogenCount="3"/>
  <atom id="a2" elementType="C" hydrogenCount="1"/>
  <atom id="a3" elementType="N" hydrogenCount="2"/>
  <atom id="a4" elementType="C" hydrogenCount="0"/>
  <atom id="a5" elementType="O" hydrogenCount="0"/>
  <atom id="a6" elementType="O" hydrogenCount="1"/>
 </atomArray>
 <bondArray>
  <bond atomRefs2="a1 a2" order="1"/>
  <bond atomRefs2="a2 a3" order="1"/>
  <bond atomRefs2="a2 a4" order="1"/>
  <bond atomRefs2="a4 a5" order="2"/>
  <bond atomRefs2="a4 a6" order="1"/>
 </bondArray>
</molecule>
```


InChI

```
InChI=1S/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6)
```

K Tables of results for all tasks in MolRepBench

This section contains the full results in Tables 6 to 21.

Model	Reasoning	Canonical SMILES	Isomeric SMILES	Randomized SMILES	DeepSMILES	IUPAC	SELFIES	MolJSON	CML	InChI
Qwen3-4B	✓	0.740$_{\pm 0.031}$	0.715$_{\pm 0.032}$	0.690$_{\pm 0.033}$	0.700$_{\pm 0.032}$	0.710$_{\pm 0.032}$	0.365$_{\pm 0.034}$	0.800$_{\pm 0.028}$	**0.995$_{\pm 0.005}$**	0.955$_{\pm 0.015}$
Qwen3-4B	×	0.740$_{\pm 0.031}$	0.795$_{\pm 0.029}$	0.660$_{\pm 0.034}$	0.695$_{\pm 0.033}$	0.700$_{\pm 0.032}$	0.375$_{\pm 0.034}$	0.830$_{\pm 0.027}$	**0.995$_{\pm 0.005}$**	0.945$_{\pm 0.016}$
Qwen3-30B	✓	0.805$_{\pm 0.028}$	**0.820$_{\pm 0.027}$**	0.735$_{\pm 0.031}$	0.885$_{\pm 0.023}$	0.780$_{\pm 0.029}$	0.410$_{\pm 0.035}$	0.990$_{\pm 0.007}$	**0.995$_{\pm 0.005}$**	0.995$_{\pm 0.005}$
Qwen3-30B	×	**0.855$_{\pm 0.025}$**	0.810$_{\pm 0.028}$	0.740$_{\pm 0.031}$	0.850$_{\pm 0.025}$	0.785$_{\pm 0.029}$	0.425$_{\pm 0.035}$	0.990$_{\pm 0.007}$	**0.995$_{\pm 0.005}$**	0.995$_{\pm 0.005}$
Phi-4	✓	0.425$_{\pm 0.035}$	0.445$_{\pm 0.035}$	0.395$_{\pm 0.035}$	0.395$_{\pm 0.035}$	0.515$_{\pm 0.035}$	0.255$_{\pm 0.031}$	0.925$_{\pm 0.019}$	0.720$_{\pm 0.032}$	**1.000**
Phi-4-Reasoning	✓	0.715$_{\pm 0.032}$	0.750$_{\pm 0.031}$	0.635$_{\pm 0.034}$	0.735$_{\pm 0.031}$	0.635$_{\pm 0.034}$	0.365$_{\pm 0.034}$	0.970$_{\pm 0.012}$	0.965$_{\pm 0.013}$	**1.000**
Phi-4-Reasoning-Plus	✓	0.705$_{\pm 0.032}$	0.710$_{\pm 0.032}$	0.620$_{\pm 0.034}$	0.790$_{\pm 0.029}$	0.695$_{\pm 0.033}$	0.375$_{\pm 0.034}$	0.975$_{\pm 0.011}$	0.990$_{\pm 0.007}$	**1.000**
Qwen2.5-14B	×	0.195$_{\pm 0.028}$	0.180$_{\pm 0.027}$	0.175$_{\pm 0.027}$	0.185$_{\pm 0.028}$	0.225$_{\pm 0.030}$	0.135$_{\pm 0.024}$	0.750$_{\pm 0.031}$	0.290$_{\pm 0.032}$	0.895$_{\pm 0.022}$
ChemDFM-v2.0-14B	×	0.125$_{\pm 0.023}$	0.145$_{\pm 0.025}$	0.170$_{\pm 0.027}$	0.085$_{\pm 0.020}$	0.110$_{\pm 0.022}$	0.030$_{\pm 0.012}$	0.715$_{\pm 0.032}$	0.300$_{\pm 0.032}$	0.275$_{\pm 0.032}$
ChemDFM-R-14B	✓	0.260$_{\pm 0.031}$	0.265$_{\pm 0.031}$	0.210$_{\pm 0.029}$	0.140$_{\pm 0.025}$	0.190$_{\pm 0.028}$	0.095$_{\pm 0.021}$	0.840$_{\pm 0.026}$	0.215$_{\pm 0.029}$	0.905$_{\pm 0.021}$
Mistral-Small-24B-Instruct-2501	×	0.360$_{\pm 0.034}$	0.330$_{\pm 0.033}$	0.330$_{\pm 0.033}$	0.285$_{\pm 0.032}$	0.560$_{\pm 0.035}$	0.280$_{\pm 0.032}$	0.995$_{\pm 0.005}$	0.765$_{\pm 0.030}$	**1.000**
Ether0-24B	✓	0.195$_{\pm 0.028}$	0.175$_{\pm 0.027}$	0.160$_{\pm 0.026}$	0.060$_{\pm 0.017}$	0.090$_{\pm 0.020}$	0.125$_{\pm 0.023}$	0.280$_{\pm 0.032}$	0.085$_{\pm 0.020}$	0.200$_{\pm 0.028}$
OLMo-3.1-32B-Instruct	×	0.590$_{\pm 0.035}$	0.595$_{\pm 0.035}$	0.525$_{\pm 0.035}$	0.505$_{\pm 0.035}$	0.460$_{\pm 0.035}$	0.330$_{\pm 0.033}$	0.960$_{\pm 0.014}$	0.930$_{\pm 0.018}$	0.960$_{\pm 0.014}$
OLMo-3.1-32B-Think	✓	0.725$_{\pm 0.032}$	0.785$_{\pm 0.029}$	0.700$_{\pm 0.032}$	0.710$_{\pm 0.032}$	0.655$_{\pm 0.034}$	0.375$_{\pm 0.034}$	0.950$_{\pm 0.015}$	0.990$_{\pm 0.007}$	**1.000**
GPT-5.4-Mini	✓	0.815$_{\pm 0.028}$	0.805$_{\pm 0.028}$	**0.825$_{\pm 0.027}$**	**0.900$_{\pm 0.021}$**	**0.910$_{\pm 0.020}$**	0.405$_{\pm 0.035}$	**1.000**	**0.995$_{\pm 0.005}$**	**1.000**
Claude-Haiku-4.5	✓	0.765$_{\pm 0.030}$	0.790$_{\pm 0.029}$	0.675$_{\pm 0.033}$	0.890$_{\pm 0.022}$	0.755$_{\pm 0.030}$	**0.525$_{\pm 0.035}$**	0.995$_{\pm 0.005}$	0.985$_{\pm 0.009}$	**1.000**

Table 6: Atom Counting (accuracy)

Model	Reasoning	Canonical SMILES	Isomeric SMILES	Randomized SMILES	DeepSMILES	IUPAC	SELFIES	MolJSON	CML	InChI
Qwen3-4B	✓	0.583$_{\pm 0.054}$	0.676$_{\pm 0.066}$	0.530$_{\pm 0.048}$	0.579$_{\pm 0.022}$	0.719$_{\pm 0.082}$	0.262$_{\pm 0.020}$	0.680$_{\pm 0.043}$	0.765$_{\pm 0.078}$	0.348$_{\pm 0.023}$
Qwen3-4B	×	0.591$_{\pm 0.037}$	0.596$_{\pm 0.033}$	0.545$_{\pm 0.034}$	0.566$_{\pm 0.030}$	0.729$_{\pm 0.078}$	0.322$_{\pm 0.053}$	0.647$_{\pm 0.047}$	0.847$_{\pm 0.075}$	0.338$_{\pm 0.025}$
Qwen3-30B	✓	0.612$_{\pm 0.037}$	0.837$_{\pm 0.074}$	0.590$_{\pm 0.033}$	0.587$_{\pm 0.033}$	0.720$_{\pm 0.063}$	0.293$_{\pm 0.020}$	0.741$_{\pm 0.034}$	0.827$_{\pm 0.058}$	0.401$_{\pm 0.025}$
Qwen3-30B	×	0.806$_{\pm 0.080}$	0.833$_{\pm 0.075}$	0.684$_{\pm 0.064}$	0.597$_{\pm 0.028}$	0.743$_{\pm 0.062}$	0.343$_{\pm 0.053}$	0.730$_{\pm 0.029}$	**0.854$_{\pm 0.066}$**	0.418$_{\pm 0.027}$
Phi-4	✓	0.347$_{\pm 0.032}$	0.448$_{\pm 0.058}$	0.332$_{\pm 0.029}$	0.313$_{\pm 0.033}$	0.665$_{\pm 0.072}$	0.276$_{\pm 0.028}$	0.470$_{\pm 0.034}$	0.479$_{\pm 0.035}$	0.294$_{\pm 0.022}$
Phi-4-Reasoning	✓	0.517$_{\pm 0.039}$	0.578$_{\pm 0.043}$	0.510$_{\pm 0.031}$	0.488$_{\pm 0.030}$	0.617$_{\pm 0.066}$	0.351$_{\pm 0.053}$	0.614$_{\pm 0.036}$	0.757$_{\pm 0.082}$	0.400$_{\pm 0.033}$
Phi-4-Reasoning-Plus	✓	0.642$_{\pm 0.041}$	0.604$_{\pm 0.043}$	0.522$_{\pm 0.044}$	0.525$_{\pm 0.029}$	0.711$_{\pm 0.073}$	0.355$_{\pm 0.050}$	0.710$_{\pm 0.032}$	0.807$_{\pm 0.082}$	0.407$_{\pm 0.036}$
Qwen2.5-14B	×	0.205$_{\pm 0.016}$	0.241$_{\pm 0.018}$	0.225$_{\pm 0.015}$	0.230$_{\pm 0.021}$	0.394$_{\pm 0.040}$	0.232$_{\pm 0.016}$	0.241$_{\pm 0.024}$	0.289$_{\pm 0.021}$	0.214$_{\pm 0.018}$
ChemDFM-v2.0-14B	×	0.328$_{\pm 0.023}$	0.387$_{\pm 0.028}$	0.298$_{\pm 0.020}$	0.363$_{\pm 0.031}$	0.367$_{\pm 0.029}$	0.168$_{\pm 0.018}$	0.269$_{\pm 0.024}$	0.290$_{\pm 0.021}$	0.170$_{\pm 0.024}$
ChemDFM-R-14B	✓	0.561$_{\pm 0.044}$	0.576$_{\pm 0.040}$	0.579$_{\pm 0.046}$	0.467$_{\pm 0.040}$	0.504$_{\pm 0.039}$	0.320$_{\pm 0.034}$	0.196$_{\pm 0.028}$	0.277$_{\pm 0.028}$	0.143$_{\pm 0.022}$
Mistral-Small-24B-Instruct-2501	×	0.335$_{\pm 0.039}$	0.463$_{\pm 0.075}$	0.350$_{\pm 0.036}$	0.376$_{\pm 0.032}$	0.680$_{\pm 0.067}$	0.310$_{\pm 0.056}$	0.331$_{\pm 0.039}$	0.408$_{\pm 0.036}$	0.300$_{\pm 0.027}$
Ether0-24B	✓	0.393$_{\pm 0.042}$	0.382$_{\pm 0.033}$	0.421$_{\pm 0.041}$	0.402$_{\pm 0.028}$	0.326$_{\pm 0.026}$	0.227$_{\pm 0.021}$	0.203$_{\pm 0.019}$	0.220$_{\pm 0.023}$	0.177$_{\pm 0.016}$
OLMo-3.1-32B-Instruct	×	0.403$_{\pm 0.049}$	0.406$_{\pm 0.039}$	0.296$_{\pm 0.038}$	0.203$_{\pm 0.021}$	0.757$_{\pm 0.078}$	0.195$_{\pm 0.018}$	0.478$_{\pm 0.038}$	0.597$_{\pm 0.078}$	0.211$_{\pm 0.024}$
OLMo-3.1-32B-Think	✓	0.526$_{\pm 0.037}$	0.463$_{\pm 0.037}$	0.445$_{\pm 0.055}$	0.143$_{\pm 0.019}$	0.729$_{\pm 0.078}$	0.192$_{\pm 0.019}$	0.686$_{\pm 0.064}$	0.782$_{\pm 0.085}$	0.256$_{\pm 0.019}$
GPT-5.4-Mini	✓	**0.849$_{\pm 0.074}$**	**0.844$_{\pm 0.074}$**	**0.738$_{\pm 0.056}$**	**0.637$_{\pm 0.035}$**	0.739$_{\pm 0.061}$	**0.519$_{\pm 0.036}$**	**0.797$_{\pm 0.062}$**	0.724$_{\pm 0.048}$	**0.622$_{\pm 0.055}$**
Claude-Haiku-4.5	✓	0.818$_{\pm 0.072}$	0.813$_{\pm 0.073}$	0.608$_{\pm 0.027}$	0.493$_{\pm 0.032}$	**0.771$_{\pm 0.073}$**	0.410$_{\pm 0.028}$	0.743$_{\pm 0.058}$	0.849$_{\pm 0.074}$	0.463$_{\pm 0.028}$

Table 7: Functional Groups (Macro F1)

Model	Reasoning	Canonical SMILES	Isomeric SMILES	Randomized SMILES	DeepSMILES	IUPAC	SELFIES	MolJSON	CML	InChI
Qwen3-4B	✓	0.616$_{\pm 0.051}$	0.629$_{\pm 0.052}$	0.547$_{\pm 0.058}$	0.369$_{\pm 0.071}$	0.709$_{\pm 0.041}$	0.427$_{\pm 0.063}$	0.589$_{\pm 0.053}$	0.570$_{\pm 0.057}$	0.740$_{\pm 0.039}$
Qwen3-4B	×	0.677$_{\pm 0.046}$	0.643$_{\pm 0.050}$	0.559$_{\pm 0.054}$	0.320$_{\pm 0.073}$	0.727$_{\pm 0.048}$	0.281$_{\pm 0.071}$	0.561$_{\pm 0.059}$	0.496$_{\pm 0.066}$	0.742$_{\pm 0.033}$
Qwen3-30B	✓	0.710$_{\pm 0.046}$	0.737$_{\pm 0.040}$	0.713$_{\pm 0.040}$	0.583$_{\pm 0.058}$	0.720$_{\pm 0.042}$	0.534$_{\pm 0.056}$	0.682$_{\pm 0.047}$	0.632$_{\pm 0.053}$	0.743$_{\pm 0.040}$
Qwen3-30B	×	0.698$_{\pm 0.045}$	0.723$_{\pm 0.042}$	0.669$_{\pm 0.044}$	0.643$_{\pm 0.049}$	0.697$_{\pm 0.047}$	0.513$_{\pm 0.058}$	0.654$_{\pm 0.050}$	0.674$_{\pm 0.044}$	0.732$_{\pm 0.042}$
Phi-4	✓	0.661$_{\pm 0.050}$	0.705$_{\pm 0.043}$	0.660$_{\pm 0.048}$	0.412$_{\pm 0.070}$	0.760$_{\pm 0.037}$	0.520$_{\pm 0.055}$	0.613$_{\pm 0.050}$	0.462$_{\pm 0.059}$	0.752$_{\pm 0.038}$
Phi-4-Reasoning	✓	0.757$_{\pm 0.033}$	0.757$_{\pm 0.039}$	0.725$_{\pm 0.041}$	0.574$_{\pm 0.056}$	0.791$_{\pm 0.033}$	0.464$_{\pm 0.060}$	0.669$_{\pm 0.049}$	0.740$_{\pm 0.038}$	0.800$_{\pm 0.029}$
Phi-4-Reasoning-Plus	✓	0.661$_{\pm 0.048}$	0.741$_{\pm 0.042}$	0.756$_{\pm 0.039}$	0.573$_{\pm 0.054}$	0.821$_{\pm 0.029}$	0.637$_{\pm 0.048}$	0.718$_{\pm 0.048}$	0.687$_{\pm 0.052}$	0.766$_{\pm 0.034}$
Qwen2.5-14B	×	0.215$_{\pm 0.078}$	0.179$_{\pm 0.076}$	0.152$_{\pm 0.071}$	0.184$_{\pm 0.073}$	0.454$_{\pm 0.061}$	0.312$_{\pm 0.072}$	0.217$_{\pm 0.069}$	0.344$_{\pm 0.064}$	0.284$_{\pm 0.071}$
ChemDFM-v2.0-14B	×	0.448$_{\pm 0.069}$	0.193$_{\pm 0.072}$	0.337$_{\pm 0.077}$	0.039$_{\pm 0.091}$	0.490$_{\pm 0.065}$	0.132$_{\pm 0.074}$	0.114$_{\pm 0.083}$	-0.041$_{\pm 0.088}$	-0.094$_{\pm 0.114}$
ChemDFM-R-14B	✓	0.838$_{\pm 0.026}$	0.805$_{\pm 0.026}$	0.835$_{\pm 0.026}$	0.737$_{\pm 0.035}$	0.814$_{\pm 0.032}$	0.345$_{\pm 0.066}$	0.186$_{\pm 0.068}$	0.268$_{\pm 0.061}$	0.348$_{\pm 0.066}$
Mistral-Small-24B-Instruct-2501	×	0.651$_{\pm 0.049}$	0.704$_{\pm 0.042}$	0.616$_{\pm 0.051}$	0.465$_{\pm 0.065}$	0.766$_{\pm 0.036}$	0.371$_{\pm 0.064}$	0.618$_{\pm 0.052}$	0.521$_{\pm 0.068}$	0.655$_{\pm 0.051}$
Ether0-24B	✓	0.078$_{\pm 0.065}$	0.166$_{\pm 0.073}$	0.190$_{\pm 0.063}$	0.054$_{\pm 0.074}$	0.170$_{\pm 0.064}$	0.124$_{\pm 0.075}$	-0.002$_{\pm 0.069}$	-0.030$_{\pm 0.069}$	0.045$_{\pm 0.077}$
OLMo-3.1-32B-Instruct	×	0.517$_{\pm 0.058}$	0.452$_{\pm 0.065}$	0.582$_{\pm 0.054}$	0.332$_{\pm 0.075}$	0.703$_{\pm 0.045}$	0.218$_{\pm 0.081}$	0.481$_{\pm 0.063}$	0.414$_{\pm 0.067}$	0.672$_{\pm 0.047}$
OLMo-3.1-32B-Think	✓	0.400$_{\pm 0.068}$	0.552$_{\pm 0.053}$	0.527$_{\pm 0.056}$	0.416$_{\pm 0.066}$	0.710$_{\pm 0.044}$	0.350$_{\pm 0.067}$	0.631$_{\pm 0.049}$	0.637$_{\pm 0.052}$	0.718$_{\pm 0.043}$
GPT-5.4-Mini	✓	**0.878$_{\pm 0.024}$**	**0.866$_{\pm 0.025}$**	**0.862$_{\pm 0.027}$**	**0.840$_{\pm 0.028}$**	**0.872$_{\pm 0.030}$**	**0.816$_{\pm 0.032}$**	**0.875$_{\pm 0.022}$**	**0.862$_{\pm 0.015}$**	**0.900$_{\pm 0.021}$**
Claude-Haiku-4.5	✓	0.780$_{\pm 0.032}$	0.758$_{\pm 0.040}$	0.752$_{\pm 0.034}$	0.696$_{\pm 0.043}$	0.786$_{\pm 0.039}$	0.720$_{\pm 0.040}$	0.688$_{\pm 0.043}$	0.735$_{\pm 0.039}$	0.778$_{\pm 0.036}$

Table 8: Property Estimation - Log P (ρ)

Model	Reasoning	Canonical SMILES	Isomeric SMILES	Randomized SMILES	DeepSMILES	IUPAC	SELFIES	MolJSON	CML	InChI
Qwen3-4B	✓	0.642$_{\pm 0.051}$	0.719$_{\pm 0.043}$	0.548$_{\pm 0.060}$	0.416$_{\pm 0.067}$	0.404$_{\pm 0.066}$	0.586$_{\pm 0.056}$	0.396$_{\pm 0.062}$	0.496$_{\pm 0.064}$	0.489$_{\pm 0.060}$
Qwen3-4B	×	0.626$_{\pm 0.051}$	0.673$_{\pm 0.051}$	0.632$_{\pm 0.050}$	0.378$_{\pm 0.065}$	0.546$_{\pm 0.053}$	0.620$_{\pm 0.056}$	0.418$_{\pm 0.061}$	0.505$_{\pm 0.063}$	0.527$_{\pm 0.048}$
Qwen3-30B	✓	0.914$_{\pm 0.036}$	0.959$_{\pm 0.008}$	0.861$_{\pm 0.046}$	0.820$_{\pm 0.047}$	0.954$_{\pm 0.008}$	0.830$_{\pm 0.054}$	0.972$_{\pm 0.007}$	0.975$_{\pm 0.005}$	0.973$_{\pm 0.006}$
Qwen3-30B	×	0.936$_{\pm 0.026}$	0.967$_{\pm 0.005}$	0.927$_{\pm 0.028}$	0.798$_{\pm 0.052}$	0.928$_{\pm 0.028}$	0.739$_{\pm 0.068}$	0.975$_{\pm 0.005}$	0.968$_{\pm 0.006}$	0.976$_{\pm 0.005}$
Phi-4	✓	0.449$_{\pm 0.066}$	0.399$_{\pm 0.072}$	0.393$_{\pm 0.065}$	0.390$_{\pm 0.067}$	0.516$_{\pm 0.067}$	0.195$_{\pm 0.066}$	0.404$_{\pm 0.064}$	0.585$_{\pm 0.056}$	0.369$_{\pm 0.063}$
Phi-4-Reasoning	✓	0.887$_{\pm 0.031}$	0.891$_{\pm 0.033}$	0.820$_{\pm 0.050}$	0.791$_{\pm 0.061}$	0.855$_{\pm 0.043}$	0.668$_{\pm 0.065}$	0.756$_{\pm 0.056}$	0.743$_{\pm 0.051}$	0.866$_{\pm 0.032}$
Phi-4-Reasoning-Plus	✓	0.660$_{\pm 0.070}$	0.587$_{\pm 0.080}$	0.545$_{\pm 0.080}$	0.723$_{\pm 0.061}$	0.752$_{\pm 0.052}$	0.708$_{\pm 0.059}$	0.628$_{\pm 0.064}$	0.736$_{\pm 0.050}$	0.671$_{\pm 0.055}$
Qwen2.5-14B	×	0.243$_{\pm 0.071}$	0.145$_{\pm 0.073}$	0.214$_{\pm 0.071}$	0.169$_{\pm 0.072}$	0.231$_{\pm 0.072}$	0.190$_{\pm 0.070}$	0.158$_{\pm 0.075}$	0.207$_{\pm 0.072}$	0.223$_{\pm 0.072}$
ChemDFM-v2.0-14B	×	0.024$_{\pm 0.088}$	-0.134$_{\pm 0.080}$	-0.032$_{\pm 0.081}$	0.457$_{\pm 0.090}$	0.331$_{\pm 0.084}$	0.399$_{\pm 0.070}$	0.112$_{\pm 0.079}$	0.022$_{\pm 0.082}$	-0.025$_{\pm 0.110}$
ChemDFM-R-14B	✓	0.972$_{\pm 0.005}$	0.972$_{\pm 0.004}$	0.964$_{\pm 0.008}$	0.894$_{\pm 0.016}$	0.971$_{\pm 0.005}$	0.687$_{\pm 0.049}$	0.547$_{\pm 0.054}$	0.645$_{\pm 0.037}$	0.650$_{\pm 0.053}$
Mistral-Small-24B-Instruct-2501	×	0.730$_{\pm 0.042}$	0.683$_{\pm 0.046}$	0.746$_{\pm 0.038}$	0.764$_{\pm 0.040}$	0.755$_{\pm 0.039}$	0.717$_{\pm 0.043}$	0.759$_{\pm 0.038}$	0.704$_{\pm 0.044}$	0.800$_{\pm 0.034}$
Ether0-24B	✓	0.111$_{\pm 0.070}$	0.094$_{\pm 0.071}$	0.035$_{\pm 0.070}$	0.156$_{\pm 0.073}$	-0.148$_{\pm 0.073}$	-0.015$_{\pm 0.071}$	0.117$_{\pm 0.071}$	0.065$_{\pm 0.070}$	0.029$_{\pm 0.073}$
OLMo-3.1-32B-Instruct	×	0.493$_{\pm 0.059}$	0.475$_{\pm 0.068}$	0.491$_{\pm 0.061}$	0.333$_{\pm 0.072}$	0.513$_{\pm 0.061}$	0.141$_{\pm 0.080}$	0.539$_{\pm 0.056}$	0.453$_{\pm 0.070}$	0.435$_{\pm 0.070}$
OLMo-3.1-32B-Think	✓	0.574$_{\pm 0.061}$	0.737$_{\pm 0.043}$	0.632$_{\pm 0.054}$	0.393$_{\pm 0.071}$	0.752$_{\pm 0.045}$	0.603$_{\pm 0.061}$	0.647$_{\pm 0.047}$	0.674$_{\pm 0.045}$	0.715$_{\pm 0.045}$
GPT-5.4-Mini	✓	**0.997$_{\pm 0.001}$**	**0.994$_{\pm 0.002}$**	**0.995$_{\pm 0.001}$**	**0.970$_{\pm 0.006}$**	**0.994$_{\pm 0.002}$**	**0.956$_{\pm 0.009}$**	**0.994$_{\pm 0.002}$**	**0.996$_{\pm 0.001}$**	**0.992$_{\pm 0.002}$**
Claude-Haiku-4.5	✓	0.968$_{\pm 0.005}$	0.964$_{\pm 0.006}$	0.942$_{\pm 0.023}$	0.932$_{\pm 0.011}$	0.962$_{\pm 0.007}$	0.950$_{\pm 0.009}$	0.964$_{\pm 0.007}$	0.973$_{\pm 0.006}$	0.949$_{\pm 0.024}$

Table 9: Property Estimation - TPSA (ρ)

Model	Reasoning	Canonical SMILES	Isomeric SMILES	Randomized SMILES	DeepSMILES	IUPAC	SELFIES	MolJSON	CML	InChI
Qwen3-4B	✓	0.300±0.033	0.315±0.034	0.250±0.031	0.185±0.027	0.585±0.035	0.165±0.026	0.520±0.035	0.785±0.030	0.230±0.030
Qwen3-4B	×	0.290±0.032	0.285±0.033	0.310±0.033	0.205±0.029	0.550±0.034	0.140±0.023	0.505±0.036	0.775±0.030	0.230±0.030
Qwen3-30B	✓	0.435±0.036	0.495±0.034	0.475±0.036	0.265±0.032	0.615±0.034	0.130±0.023	0.620±0.035	0.760±0.031	0.380±0.035
Qwen3-30B	×	0.450±0.036	0.445±0.035	0.410±0.036	0.310±0.031	0.590±0.034	0.105±0.022	0.685±0.033	0.775±0.030	0.410±0.035
Phi-4	✓	0.340±0.035	0.335±0.034	0.286±0.033	0.190±0.028	0.515±0.036	0.215±0.029	0.120±0.022	0.730±0.032	0.215±0.029
Phi-4-Reasoning	✓	0.350±0.034	0.365±0.034	0.330±0.034	0.235±0.030	0.583±0.037	0.111±0.022	0.410±0.034	0.780±0.030	0.410±0.035
Phi-4-Reasoning-Plus	✓	0.370±0.036	0.360±0.034	0.330±0.032	0.261±0.032	0.485±0.036	0.105±0.021	0.550±0.036	0.760±0.030	0.420±0.034
Qwen2.5-14B	×	0.168±0.026	0.200±0.028	0.136±0.024	0.182±0.028	0.286±0.032	0.078±0.018	0.097±0.021	0.178±0.027	0.135±0.025
ChemDFM-v2.0-14B	×	0.287±0.034	0.204±0.030	0.230±0.033	0.123±0.027	0.310±0.034	0.123±0.026	0.050±0.016	0.254±0.033	0.044±0.019
ChemDFM-R-14B	✓	0.465±0.035	0.460±0.035	0.500±0.035	0.290±0.033	0.465±0.035	0.145±0.025	0.130±0.024	0.140±0.024	0.175±0.026
Mistral-Small-24B-Instruct-2501	×	0.255±0.031	0.295±0.033	0.230±0.030	0.140±0.025	0.475±0.035	0.125±0.023	0.150±0.025	0.710±0.033	0.140±0.025
Ether0-24B	✓	0.172±0.027	0.155±0.027	0.135±0.024	0.046±0.016	0.096±0.021	0.115±0.022	0.115±0.021	0.121±0.024	0.155±0.026
OLMo-3.1-32B-Instruct	×	0.137±0.025	0.182±0.028	0.109±0.023	0.112±0.023	0.500±0.038	0.117±0.024	0.246±0.032	0.708±0.033	0.127±0.026
OLMo-3.1-32B-Think	✓	0.150±0.025	0.195±0.029	0.155±0.025	0.170±0.028	0.550±0.036	0.085±0.019	0.410±0.036	0.785±0.030	0.180±0.027
GPT-5.4-Mini	✓	**0.810±0.028**	**0.815±0.028**	**0.765±0.031**	**0.535±0.036**	**0.730±0.031**	0.195±0.028	**0.795±0.029**	**0.980±0.006**	**0.682±0.031**
Claude-Haiku-4.5	✓	0.620±0.034	0.640±0.034	0.555±0.036	0.392±0.035	0.665±0.033	**0.295±0.033**	0.575±0.034	0.920±0.020	0.455±0.035

Table 10: Property Estimation - HBD (Exact Match)

Model	Reasoning	Canonical SMILES	Isomeric SMILES	Randomized SMILES	DeepSMILES	IUPAC	SELFIES	MolJSON	CML	InChI
Qwen3-4B	✓	0.385±0.034	0.395±0.034	0.280±0.032	0.325±0.032	0.365±0.034	0.210±0.029	0.330±0.033	0.485±0.034	0.465±0.035
Qwen3-4B	×	0.325±0.033	0.380±0.034	0.275±0.032	0.340±0.032	0.355±0.034	0.235±0.030	0.335±0.033	0.475±0.035	0.445±0.035
Qwen3-30B	✓	0.425±0.035	0.455±0.036	0.385±0.035	0.450±0.035	0.420±0.035	0.240±0.030	0.500±0.035	0.475±0.034	0.465±0.036
Qwen3-30B	×	0.450±0.036	0.420±0.035	0.410±0.035	0.455±0.035	0.430±0.035	0.230±0.030	0.515±0.034	0.460±0.034	0.485±0.035
Phi-4	✓	0.255±0.031	0.210±0.029	0.260±0.031	0.165±0.026	0.270±0.031	0.100±0.022	0.440±0.033	0.390±0.034	0.365±0.034
Phi-4-Reasoning	✓	0.415±0.035	0.435±0.034	0.430±0.035	0.430±0.034	0.345±0.033	0.187±0.027	0.470±0.035	0.445±0.034	0.485±0.035
Phi-4-Reasoning-Plus	✓	0.430±0.035	0.420±0.034	0.455±0.036	0.455±0.035	0.390±0.034	0.220±0.029	0.465±0.035	0.420±0.034	0.410±0.034
Qwen2.5-14B	×	0.118±0.023	0.145±0.025	0.077±0.020	0.074±0.019	0.076±0.019	0.086±0.020	0.310±0.032	0.146±0.026	0.254±0.030
ChemDFM-v2.0-14B	×	0.130±0.025	0.050±0.016	0.093±0.022	0.037±0.016	0.075±0.019	0.089±0.022	0.250±0.032	0.141±0.026	0.139±0.034
ChemDFM-R-14B	✓	0.145±0.024	0.115±0.023	0.155±0.026	0.165±0.027	0.150±0.026	0.100±0.021	0.210±0.028	0.165±0.026	0.430±0.034
Mistral-Small-24B-Instruct-2501	×	0.140±0.024	0.205±0.028	0.195±0.027	0.141±0.025	0.250±0.030	0.085±0.019	0.395±0.034	0.315±0.033	0.405±0.035
Ether0-24B	✓	0.045±0.015	0.045±0.015	0.060±0.016	0.080±0.019	0.081±0.019	0.030±0.012	0.055±0.016	0.026±0.011	0.055±0.016
OLMo-3.1-32B-Instruct	×	0.339±0.036	0.309±0.035	0.341±0.036	0.303±0.033	0.301±0.034	0.168±0.026	0.434±0.037	0.304±0.034	0.361±0.035
OLMo-3.1-32B-Think	✓	0.440±0.035	0.435±0.035	0.440±0.035	0.410±0.035	0.390±0.035	0.200±0.029	0.465±0.035	0.470±0.035	0.460±0.035
GPT-5.4-Mini	✓	**0.590±0.035**	**0.570±0.036**	**0.540±0.035**	**0.600±0.034**	**0.640±0.034**	0.225±0.029	**0.680±0.032**	**0.606±0.023**	**0.645±0.033**
Claude-Haiku-4.5	✓	0.462±0.035	0.455±0.035	0.425±0.035	0.370±0.034	0.440±0.037	**0.335±0.033**	0.480±0.036	0.495±0.035	0.435±0.036

Table 11: Property Estimation - HBA (Exact-Match)

Model	Reasoning	Canonical SMILES	Isomeric SMILES	Randomized SMILES	DeepSMILES	IUPAC	SELFIES	MolJSON	CML	InChI
Qwen3-4B	✓	0.821±0.044	0.769±0.048	0.808±0.045	0.808±0.045	0.885±0.036	0.615±0.055	0.795±0.046	0.282±0.051	0.795±0.046
Qwen3-4B	×	0.795±0.046	0.859±0.040	0.705±0.052	0.667±0.054	0.910±0.033	0.615±0.055	0.808±0.045	0.295±0.052	0.769±0.048
Qwen3-30B	✓	0.885±0.036	0.910±0.033	0.833±0.042	0.872±0.038	0.910±0.033	0.692±0.053	0.795±0.046	0.244±0.049	0.897±0.035
Qwen3-30B	×	0.897±0.035	0.897±0.035	0.821±0.044	0.833±0.042	0.936±0.028	0.679±0.053	0.795±0.046	0.295±0.052	0.859±0.040
Phi-4	✓	0.603±0.056	0.641±0.055	0.564±0.057	0.577±0.056	0.705±0.052	0.385±0.055	0.462±0.057	0.205±0.046	0.551±0.057
Phi-4-Reasoning	✓	0.808±0.045	0.859±0.040	0.769±0.048	0.756±0.049	0.872±0.038	0.641±0.055	0.744±0.050	0.269±0.051	0.795±0.046
Phi-4-Reasoning-Plus	✓	0.846±0.041	0.808±0.045	0.859±0.040	0.821±0.044	0.923±0.030	0.590±0.056	0.795±0.046	0.231±0.048	0.846±0.041
Qwen2.5-14B	×	0.167±0.042	0.269±0.051	0.269±0.051	0.154±0.041	0.269±0.051	0.115±0.036	0.128±0.038	0.141±0.040	0.256±0.050
ChemDFM-v2.0-14B	×	0.667±0.054	0.718±0.051	0.654±0.054	0.449±0.057	0.692±0.053	0.410±0.056	0.295±0.052	0.192±0.045	0.359±0.055
ChemDFM-R-14B	✓	0.603±0.056	0.577±0.056	0.628±0.055	0.551±0.057	0.795±0.046	0.538±0.057	0.449±0.057	0.141±0.040	0.410±0.056
Mistral-Small-24B-Instruct-2501	×	0.513±0.057	0.564±0.057	0.526±0.057	0.423±0.056	0.628±0.055	0.372±0.055	0.436±0.057	0.462±0.057	0.538±0.057
Ether0-24B	✓	0.269±0.051	0.231±0.048	0.231±0.048	0.244±0.049	0.282±0.051	0.231±0.048	0.218±0.047	0.077±0.030	0.397±0.056
OLMo-3.1-32B-Instruct	×	0.628±0.055	0.615±0.055	0.603±0.056	0.487±0.057	0.769±0.048	0.474±0.057	0.449±0.057	0.205±0.046	0.590±0.056
OLMo-3.1-32B-Think	✓	0.744±0.050	0.782±0.047	0.731±0.051	0.705±0.052	0.910±0.033	0.462±0.057	0.615±0.055	0.192±0.045	0.769±0.048
GPT-5.4-Mini	✓	**0.949±0.025**	**1.000**	**0.936±0.028**	**0.949±0.025**	**0.962±0.022**	0.718±0.051	**0.962±0.022**	**0.923±0.030**	**0.949±0.025**
Claude-Haiku-4.5	✓	0.885±0.036	0.936±0.028	0.846±0.041	0.910±0.033	0.949±0.025	**0.846±0.041**	0.833±0.042	0.821±0.044	0.833±0.042

Table 12: Molecule Retrieval (accuracy)

Model	Reasoning	Canonical SMILES	Isomeric SMILES	Randomized SMILES	DeepSMILES	IUPAC	SELFIES	MolJSON	CML	InChI
Qwen3-4B	✓	0.252±0.019	0.266±0.020	0.602±0.022	0.268±0.020	0.253±0.019	0.260±0.020	0.648±0.021	0.658±0.021	0.250±0.019
Qwen3-4B	×	0.262±0.020	0.262±0.020	0.610±0.022	0.276±0.020	0.251±0.019	0.276±0.020	0.638±0.022	0.660±0.021	0.252±0.019
Qwen3-30B	✓	0.250±0.019	0.258±0.020	0.612±0.022	0.260±0.020	0.251±0.019	0.256±0.020	0.632±0.022	0.640±0.021	0.250±0.019
Qwen3-30B	×	0.252±0.019	0.252±0.019	0.608±0.022	0.250±0.019	0.251±0.019	0.258±0.020	0.640±0.021	0.640±0.021	0.250±0.019
Phi-4	✓	0.252±0.019	**0.272±0.020**	0.462±0.022	0.272±0.020	0.248±0.019	0.290±0.020	0.620±0.022	0.604±0.022	0.254±0.019
Phi-4-Reasoning	✓	0.260±0.020	0.266±0.020	0.566±0.022	0.270±0.020	0.253±0.019	0.254±0.019	0.690±0.021	**0.712±0.020**	0.250±0.019
Phi-4-Reasoning-Plus	✓	0.258±0.020	0.264±0.020	0.560±0.022	0.266±0.020	0.257±0.020	0.262±0.020	0.660±0.021	0.652±0.021	0.250±0.019
Qwen2.5-14B	×	0.066±0.011	0.064±0.011	0.112±0.014	0.082±0.012	0.078±0.012	0.094±0.013	0.178±0.017	0.244±0.019	0.046±0.009
ChemDFM-v2.0-14B	×	**0.268±0.020**	0.244±0.019	0.372±0.022	0.306±0.021	0.257±0.020	0.296±0.020	0.564±0.022	0.572±0.022	0.248±0.019
ChemDFM-R-14B	✓	0.260±0.020	0.256±0.020	0.372±0.022	0.278±0.020	0.253±0.019	**0.330±0.021**	0.590±0.022	0.598±0.022	0.264±0.020
Mistral-Small-24B-Instruct-2501	×	0.262±0.020	0.260±0.020	0.530±0.022	0.272±0.020	0.251±0.019	0.278±0.020	0.640±0.021	0.690±0.021	0.250±0.019
Ether0-24B	✓	0.162±0.016	0.160±0.016	0.168±0.017	0.194±0.018	0.240±0.019	0.240±0.019	0.268±0.020	0.220±0.019	**0.300±0.021**
OLMo-3.1-32B-Instruct	×	0.236±0.019	0.262±0.020	0.624±0.022	0.322±0.021	0.244±0.019	0.320±0.021	0.584±0.022	0.614±0.022	0.248±0.019
OLMo-3.1-32B-Think	✓	0.258±0.020	0.256±0.020	**0.714±0.020**	**0.406±0.022**	**0.259±0.020**	0.310±0.021	**0.708±0.020**	0.696±0.021	0.254±0.019
GPT-5.4-Mini	✓	0.250±0.019	0.252±0.019	0.348±0.021	0.252±0.019	0.251±0.019	0.256±0.020	0.426±0.022	0.440±0.022	0.258±0.020
Claude-Haiku-4.5	✓	0.252±0.019	0.252±0.019	0.514±0.022	0.252±0.019	0.251±0.019	0.252±0.019	0.612±0.022	0.570±0.022	0.250±0.019

Table 13: Isomer Discrimination (accuracy)

Model	Reasoning	Canonical SMILES	Isomeric SMILES	Randomized SMILES	DeepSMILES	IUPAC	SELFIES	MolJSON	CML	InChI
Qwen3-4B	✓	0.624±0.031	0.584±0.031	0.652±0.030	0.016±0.008	0.600±0.031	**0.884±0.020**	0.708±0.029	0.336±0.030	0.004±0.004
Qwen3-4B	×	0.592±0.031	0.552±0.032	0.596±0.031	0.016±0.008	0.588±0.031	0.876±0.021	0.748±0.028	0.344±0.030	0.004±0.004
Qwen3-30B	✓	0.712±0.029	0.692±0.029	0.696±0.029	0.008±0.006	0.644±0.030	0.532±0.032	0.560±0.031	0.288±0.029	0.032±0.011
Qwen3-30B	×	0.700±0.029	0.624±0.031	0.700±0.029	0.000	0.660±0.030	0.512±0.032	0.612±0.031	0.276±0.028	0.036±0.012
Phi-4	✓	0.712±0.029	0.712±0.029	0.716±0.029	0.008±0.006	0.484±0.032	0.080±0.017	0.056±0.015	0.004±0.004	0.020±0.009
Phi-4-Reasoning	✓	0.324±0.030	0.240±0.027	0.296±0.029	0.020±0.009	0.452±0.032	0.380±0.031	0.476±0.032	0.112±0.020	0.016±0.008
Phi-4-Reasoning-Plus	✓	0.356±0.030	0.248±0.027	0.312±0.029	0.028±0.010	0.464±0.032	0.424±0.031	0.500±0.032	0.112±0.020	0.016±0.008
Qwen2.5-14B	×	0.664±0.030	0.624±0.031	0.560±0.031	0.008±0.006	0.436±0.031	0.372±0.031	0.004±0.004	0.068±0.016	0.020±0.009
ChemDFM-v2.0-14B	×	0.696±0.029	0.680±0.030	0.708±0.029	0.020±0.009	0.148±0.023	0.252±0.028	0.112±0.020	0.000	0.000
ChemDFM-R-14B	✓	**0.976±0.010**	**0.980±0.009**	**0.964±0.012**	0.024±0.010	0.148±0.023	0.308±0.029	0.228±0.027	0.000	0.000
Mistral-Small-24B-Instruct-2501	×	0.604±0.031	0.652±0.030	0.644±0.030	0.016±0.008	0.524±0.032	0.208±0.026	0.016±0.008	0.120±0.021	0.028±0.010
Ether0-24B	✓	0.004±0.004	0.000	0.008±0.006	0.000	0.000	0.200±0.025	0.000	0.000	0.000
OLMo-3.1-32B-Instruct	×	0.312±0.029	0.236±0.027	0.272±0.028	0.020±0.009	0.372±0.031	0.216±0.026	0.248±0.027	0.000	0.004±0.004
OLMo-3.1-32B-Think	✓	0.384±0.031	0.348±0.030	0.352±0.030	0.016±0.008	0.576±0.031	0.700±0.029	**0.836±0.023**	0.000	0.004±0.004
GPT-5.4-Mini	✓	0.916±0.018	0.932±0.016	0.928±0.016	**0.288±0.029**	0.348±0.030	0.732±0.028	0.416±0.031	**0.436±0.031**	**0.104±0.019**
Claude-Haiku-4.5	✓	0.832±0.024	0.828±0.024	0.812±0.025	0.028±0.010	**0.668±0.030**	0.648±0.030	0.372±0.031	0.368±0.031	0.048±0.014

Table 14: Caption-to-Molecule (Validity Rate)

Model	Reasoning	Canonical SMILES	Isomeric SMILES	Randomized SMILES	DeepSMILES	IUPAC	SELFIES	MolJSON	CML	InChI
Qwen3-4B	✓	0.052±0.014	0.036±0.012	0.044±0.013	0.008±0.006	0.292±0.029	0.000	0.040±0.012	0.000	0.000
Qwen3-4B	×	0.052±0.014	0.040±0.012	0.036±0.012	0.008±0.006	0.296±0.029	0.000	0.052±0.014	0.000	0.000
Qwen3-30B	✓	0.056±0.015	0.080±0.017	0.056±0.015	0.008±0.006	**0.348±0.030**	0.008±0.006	0.052±0.014	0.012±0.007	0.008±0.006
Qwen3-30B	×	0.080±0.017	0.056±0.015	0.076±0.017	0.000	0.340±0.030	0.000	0.068±0.016	0.004±0.004	0.008±0.006
Phi-4	✓	0.016±0.008	0.020±0.009	0.016±0.008	0.004±0.004	0.184±0.025	0.000	0.012±0.007	0.000	0.012±0.007
Phi-4-Reasoning	✓	0.048±0.014	0.032±0.011	0.044±0.013	0.000	0.248±0.027	0.004±0.004	0.056±0.015	0.016±0.008	0.004±0.004
Phi-4-Reasoning-Plus	✓	0.056±0.015	0.044±0.013	0.040±0.012	0.000	0.256±0.028	0.004±0.004	0.056±0.015	0.024±0.010	0.004±0.004
Qwen2.5-14B	×	0.000	0.000	0.000	0.004±0.004	0.116±0.020	0.004±0.004	0.004±0.004	0.004±0.004	0.008±0.006
ChemDFM-v2.0-14B	×	0.240±0.027	0.264±0.028	0.244±0.027	0.004±0.004	0.028±0.010	0.000	0.012±0.007	0.000	0.000
ChemDFM-R-14B	✓	**0.376±0.031**	**0.384±0.031**	**0.392±0.031**	0.016±0.008	0.056±0.015	0.004±0.004	0.008±0.006	0.000	0.000
Mistral-Small-24B-Instruct-2501	×	0.008±0.006	0.016±0.008	0.020±0.009	0.008±0.006	0.196±0.025	0.004±0.004	0.004±0.004	0.004±0.004	0.016±0.008
Ether0-24B	✓	0.000	0.000	0.000	0.000	0.000	0.004±0.004	0.000	0.000	0.000
OLMo-3.1-32B-Instruct	×	0.000	0.008±0.006	0.008±0.006	0.004±0.004	0.132±0.021	0.004±0.004	0.008±0.006	0.000	0.004±0.004
OLMo-3.1-32B-Think	✓	0.016±0.008	0.016±0.008	0.016±0.008	0.008±0.006	0.260±0.028	0.004±0.004	0.052±0.014	0.000	0.000
GPT-5.4-Mini	✓	0.304±0.029	0.300±0.029	0.288±0.029	**0.124±0.021**	0.260±0.028	**0.020±0.009**	**0.204±0.026**	**0.156±0.023**	**0.084±0.018**
Claude-Haiku-4.5	✓	0.160±0.023	0.176±0.024	0.148±0.023	0.012±0.007	0.332±0.030	0.012±0.007	0.112±0.020	0.128±0.021	0.032±0.011

Table 15: Caption-to-Molecule (Exact Match)

Model	Reasoning	Canonical SMILES	Isomeric SMILES	Randomized SMILES	DeepSMILES	IUPAC	SELFIES	MolJSON	CML	InChI
Qwen3-4B	✓	0.203±0.017	0.193±0.017	0.212±0.017	0.013±0.007	0.439±0.027	0.000	0.154±0.017	0.000±0.000	0.000±0.000
Qwen3-4B	×	0.200±0.018	0.178±0.016	0.192±0.017	0.013±0.007	0.428±0.027	0.000	0.166±0.017	0.000±0.000	0.000±0.000
Qwen3-30B	✓	0.275±0.018	0.269±0.019	0.260±0.018	0.008±0.006	0.494±0.027	0.047±0.007	0.180±0.020	0.029±0.008	0.014±0.006
Qwen3-30B	×	0.269±0.019	0.233±0.018	0.286±0.019	0.000	0.493±0.027	0.037±0.005	0.213±0.021	0.013±0.005	0.014±0.006
Phi-4	✓	0.208±0.013	0.198±0.013	0.200±0.013	0.005±0.004	0.315±0.025	0.004±0.002	0.018±0.007	0.000	0.017±0.008
Phi-4-Reasoning	✓	0.146±0.017	0.097±0.015	0.126±0.016	0.006±0.004	0.341±0.027	0.021±0.005	0.153±0.018	0.031±0.009	0.010±0.005
Phi-4-Reasoning-Plus	✓	0.158±0.019	0.112±0.017	0.126±0.016	0.003±0.002	0.353±0.027	0.030±0.006	0.174±0.019	0.036±0.010	0.006±0.004
Qwen2.5-14B	×	0.126±0.008	0.123±0.009	0.108±0.009	0.004±0.004	0.244±0.022	0.019±0.006	0.004±0.004	0.010±0.004	0.009±0.006
ChemDFM-v2.0-14B	×	0.458±0.025	0.453±0.025	0.452±0.025	0.011±0.006	0.076±0.013	0.001±0.001	0.023±0.007	0.000	0.000
ChemDFM-R-14B	✓	**0.677±0.019**	**0.679±0.019**	**0.682±0.020**	0.018±0.008	0.090±0.016	0.005±0.004	0.033±0.007	0.000	0.000
Mistral-Small-24B-Instruct-2501	×	0.190±0.014	0.194±0.014	0.194±0.014	0.011±0.006	0.335±0.025	0.029±0.008	0.006±0.004	0.014±0.005	0.019±0.008
Ether0-24B	✓	0.000	0.000	0.000	0.000	0.000	0.004±0.004	0.000	0.000	0.000
OLMo-3.1-32B-Instruct	×	0.043±0.005	0.045±0.008	0.053±0.008	0.006±0.004	0.240±0.023	0.008±0.004	0.024±0.006	0.000	0.004±0.004
OLMo-3.1-32B-Think	✓	0.084±0.012	0.080±0.012	0.079±0.011	0.011±0.006	0.393±0.026	0.011±0.005	0.157±0.017	0.000	0.004±0.004
GPT-5.4-Mini	✓	0.605±0.022	0.609±0.021	0.598±0.021	**0.188±0.023**	0.304±0.028	**0.103±0.010**	**0.295±0.026**	**0.239±0.023**	**0.092±0.018**
Claude-Haiku-4.5	✓	0.464±0.022	0.465±0.022	0.421±0.022	0.020±0.009	**0.499±0.027**	0.080±0.009	0.238±0.024	0.195±0.022	0.044±0.013

Table 16: Caption-to-Molecule (Morgan Tanimoto Similarity)

Model	Reasoning	Canonical SMILES	Isomeric SMILES	Randomized SMILES	DeepSMILES	IUPAC	SELFIES	MolJSON	CML	InChI
Qwen3-4B	✓	0.389±0.022	0.371±0.022	0.410±0.022	0.014±0.007	0.546±0.029	0.002±0.001	0.331±0.020	0.013±0.003	0.001±0.001
Qwen3-4B	×	0.383±0.023	0.363±0.023	0.387±0.023	0.015±0.007	0.537±0.029	0.002±0.001	0.350±0.020	0.021±0.004	0.001±0.001
Qwen3-30B	✓	0.518±0.023	0.503±0.023	0.497±0.023	0.008±0.006	0.603±0.029	0.105±0.011	0.320±0.023	0.078±0.012	0.020±0.008
Qwen3-30B	×	0.514±0.023	0.456±0.024	0.518±0.024	0.000	0.605±0.028	0.101±0.010	**0.362±0.024**	0.048±0.008	0.025±0.009
Phi-4	✓	0.484±0.021	0.475±0.021	0.470±0.021	0.006±0.004	0.421±0.028	0.010±0.004	0.033±0.009	0.000	0.019±0.008
Phi-4-Reasoning	✓	0.239±0.023	0.165±0.020	0.220±0.022	0.008±0.005	0.415±0.030	0.045±0.007	0.285±0.023	0.058±0.012	0.013±0.007
Phi-4-Reasoning-Plus	✓	0.256±0.023	0.183±0.021	0.228±0.023	0.005±0.003	0.428±0.030	0.053±0.008	0.319±0.024	0.058±0.013	0.009±0.005
Qwen2.5-14B	×	0.333±0.017	0.307±0.017	0.270±0.017	0.004±0.004	0.345±0.026	0.033±0.007	0.004±0.004	0.023±0.007	0.010±0.006
ChemDFM-v2.0-14B	×	0.625±0.027	0.609±0.027	0.631±0.027	0.015±0.007	0.121±0.019	0.001±0.001	0.050±0.010	0.000	0.000
ChemDFM-R-14B	✓	**0.892±0.012**	**0.891±0.011**	**0.883±0.013**	0.021±0.009	0.128±0.020	0.005±0.004	0.100±0.013	0.000	0.000
Mistral-Small-24B-Instruct-2501	×	0.409±0.022	0.435±0.022	0.428±0.022	0.013±0.007	0.454±0.028	0.057±0.010	0.009±0.005	0.040±0.008	0.025±0.009
Ether0-24B	✓	0.000±0.000	0.000	0.000±0.000	0.000	0.000	0.004±0.004	0.000	0.000	0.000
OLMo-3.1-32B-Instruct	×	0.140±0.014	0.118±0.015	0.133±0.015	0.008±0.005	0.312±0.027	0.019±0.005	0.082±0.012	0.000	0.004±0.004
OLMo-3.1-32B-Think	✓	0.199±0.018	0.186±0.018	0.188±0.018	0.014±0.007	0.511±0.029	0.028±0.007	0.324±0.020	0.000	0.003±0.003
GPT-5.4-Mini	✓	0.839±0.018	0.852±0.016	0.845±0.016	**0.251±0.026**	0.339±0.029	**0.249±0.015**	0.360±0.029	**0.367±0.028**	**0.101±0.019**
Claude-Haiku-4.5	✓	0.722±0.022	0.715±0.022	0.689±0.023	0.022±0.009	**0.617±0.028**	0.186±0.013	0.326±0.028	0.296±0.026	0.047±0.013

Table 17: Caption-to-Molecule (MACCS Tanimoto Similarity)

Model	Reasoning	Canonical SMILES	Isomeric SMILES	Randomized SMILES	DeepSMILES	IUPAC	SELFIES	MolJSON	CML	InChI
Qwen3-4B	✓	0.257±0.019	0.248±0.019	0.259±0.019	0.014±0.007	0.475±0.028	0.004±0.004	0.160±0.018	0.000	0.001±0.001
Qwen3-4B	×	0.248±0.019	0.233±0.019	0.242±0.019	0.014±0.007	0.458±0.028	0.004±0.004	0.182±0.019	0.000	0.001±0.001
Qwen3-30B	✓	0.356±0.020	0.347±0.020	0.330±0.019	0.008±0.006	**0.532±0.028**	0.041±0.008	0.195±0.021	0.040±0.010	0.019±0.008
Qwen3-30B	×	0.350±0.020	0.309±0.020	0.365±0.021	0.000	0.527±0.028	0.034±0.006	0.226±0.022	0.021±0.007	0.025±0.009
Phi-4	✓	0.307±0.017	0.289±0.016	0.286±0.016	0.006±0.004	0.357±0.027	0.003±0.001	0.020±0.007	0.000±0.000	0.017±0.008
Phi-4-Reasoning	✓	0.173±0.019	0.112±0.016	0.158±0.019	0.007±0.005	0.362±0.028	0.015±0.005	0.180±0.019	0.034±0.010	0.012±0.006
Phi-4-Reasoning-Plus	✓	0.188±0.020	0.137±0.018	0.154±0.018	0.004±0.003	0.378±0.028	0.018±0.005	0.207±0.021	0.040±0.011	0.007±0.005
Qwen2.5-14B	×	0.180±0.012	0.185±0.013	0.149±0.011	0.004±0.004	0.274±0.024	0.014±0.005	0.004±0.004	0.011±0.005	0.009±0.006
ChemDFM-v2.0-14B	×	0.522±0.026	0.511±0.026	0.525±0.026	0.013±0.007	0.093±0.016	0.000	0.026±0.008	0.000	0.000
ChemDFM-R-14B	✓	**0.762±0.017**	**0.762±0.018**	**0.759±0.018**	0.019±0.008	0.106±0.018	0.004±0.004	0.036±0.008	0.000	0.000
Mistral-Small-24B-Instruct-2501	×	0.250±0.017	0.285±0.018	0.261±0.017	0.011±0.006	0.368±0.026	0.032±0.008	0.006±0.004	0.024±0.007	0.024±0.009
Ether0-24B	✓	0.000±0.000	0.000	0.000±0.000	0.000	0.000	0.004±0.004	0.000	0.000	0.000
OLMo-3.1-32B-Instruct	×	0.079±0.009	0.074±0.011	0.081±0.011	0.009±0.005	0.264±0.025	0.007±0.004	0.035±0.007	0.000	0.004±0.004
OLMo-3.1-32B-Think	✓	0.124±0.014	0.112±0.014	0.120±0.014	0.012±0.006	0.426±0.027	0.008±0.004	0.157±0.018	0.000	0.004±0.004
GPT-5.4-Mini	✓	0.696±0.021	0.707±0.020	0.697±0.020	**0.204±0.023**	0.313±0.028	**0.101±0.011**	**0.314±0.027**	**0.247±0.024**	**0.094±0.018**
Claude-Haiku-4.5	✓	0.545±0.022	0.560±0.023	0.505±0.022	0.021±0.009	0.530±0.028	0.079±0.009	0.257±0.025	0.212±0.023	0.048±0.013

Table 18: Caption-to-Molecule (RDKit Tanimoto Similarity)

Model	Reasoning	Canonical SMILES	Isomeric SMILES	Randomized SMILES	DeepSMILES	IUPAC	SELFIES	MolJSON	CML	InChI
Qwen3-4B	✓	12.069	12.286	11.798	51.767	6.248	41.612	19.528	38.580	N/A
Qwen3-4B	×	12.454	12.186	12.784	50.127	6.969	41.610	18.420	36.003	N/A
Qwen3-30B	✓	8.141	8.058	8.657	56.978	**5.227**	28.994	17.361	23.819	35.137
Qwen3-30B	×	8.650	8.467	8.227	N/A	5.883	29.035	16.145	26.109	32.779
Phi-4	✓	8.466	9.017	8.563	61.488	8.496	33.466	29.659	N/A	42.122
Phi-4-Reasoning	✓	13.159	15.857	14.433	39.255	8.074	29.703	19.044	26.084	56.278
Phi-4-Reasoning-Plus	✓	12.900	14.178	13.588	37.846	8.616	30.986	16.793	27.363	46.232
Qwen2.5-14B	×	12.590	13.450	13.624	39.922	9.832	38.124	N/A	32.164	36.180
ChemDFM-v2.0-14B	×	5.437	5.736	10.826	49.030	17.443	41.865	26.946	N/A	N/A
ChemDFM-R-14B	✓	**3.544**	10.992	**3.666**	43.437	21.877	41.883	23.124	N/A	N/A
Mistral-Small-24B-Instruct-2501	×	10.729	9.950	10.412	45.993	7.762	32.571	39.151	27.853	35.916
Ether0-24B	✓	N/A	N/A	39.400	N/A	N/A	41.344	N/A	N/A	N/A
OLMo-3.1-32B-Instruct	×	17.039	18.432	17.412	38.455	11.165	34.303	23.693	N/A	N/A
OLMo-3.1-32B-Think	✓	15.712	15.881	15.173	41.208	6.909	35.765	18.712	N/A	N/A
GPT-5.4-Mini	✓	4.077	**3.969**	3.994	**18.590**	10.428	**21.880**	**12.969**	**14.820**	**27.641**
Claude-Haiku-4.5	✓	5.277	5.460	5.901	40.766	5.527	21.898	15.165	15.280	33.615

Table 19: Caption-to-Molecule (FCD) (For some models/representations, FCD is not computable due to the absence of valid generations so marked as N/A.)

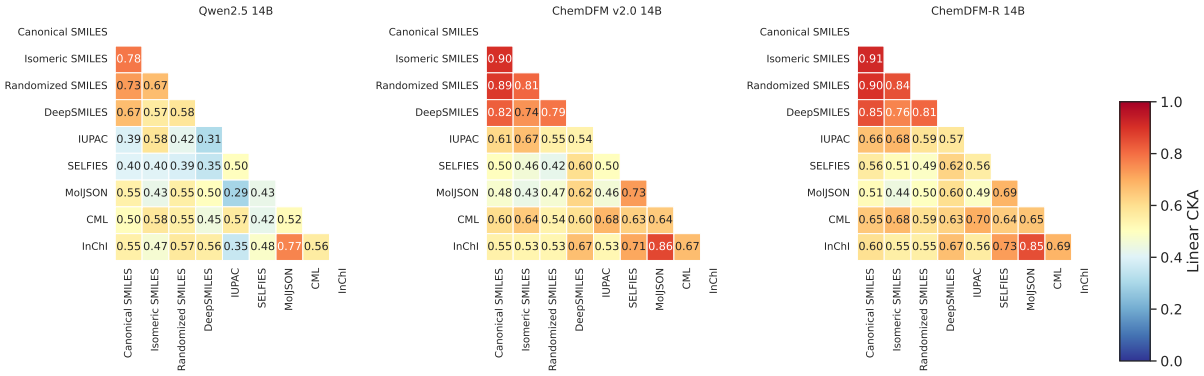
Model	Reasoning	Canonical SMILES	Isomeric SMILES	Randomized SMILES	DeepSMILES	IUPAC	SELFIES	MolJSON	CML	InChI
Qwen3-4B	✓	0.808±0.025	0.792±0.026	0.668±0.030	0.744±0.028	0.498±0.032	0.500±0.032	0.796±0.026	0.808±0.025	0.684±0.029
Qwen3-4B	×	0.796±0.026	0.776±0.026	0.640±0.030	0.724±0.028	0.498±0.032	0.504±0.032	0.780±0.026	0.824±0.024	0.704±0.029
Qwen3-30B	✓	0.832±0.024	0.780±0.026	0.740±0.028	0.828±0.024	0.498±0.032	0.488±0.032	0.844±0.023	0.836±0.023	0.804±0.025
Qwen3-30B	×	0.812±0.025	0.820±0.024	0.708±0.029	0.864±0.022	0.498±0.032	0.492±0.032	0.864±0.022	0.848±0.023	0.824±0.024
Phi-4	✓	0.872±0.021	0.864±0.022	0.820±0.024	0.856±0.022	0.498±0.032	0.440±0.031	**0.948±0.014**	0.888±0.020	0.912±0.018
Phi-4-Reasoning	✓	0.868±0.021	0.860±0.022	0.820±0.024	0.868±0.021	0.502±0.032	0.516±0.032	0.884±0.020	0.908±0.018	0.928±0.016
Phi-4-Reasoning-Plus	✓	0.892±0.020	0.892±0.020	0.808±0.025	0.864±0.022	0.502±0.032	0.516±0.032	0.900±0.019	0.892±0.020	**0.932±0.016**
Qwen2.5-14B	×	0.100±0.019	0.140±0.022	0.128±0.021	0.144±0.022	0.256±0.028	0.188±0.025	0.168±0.024	0.192±0.025	0.128±0.021
ChemDFM-v2.0-14B	×	0.656±0.030	0.628±0.031	0.628±0.031	0.640±0.030	0.880±0.021	**0.828±0.024**	0.672±0.030	0.348±0.030	0.736±0.028
ChemDFM-R-14B	✓	0.532±0.032	0.532±0.032	0.580±0.031	0.528±0.032	**0.916±0.018**	0.724±0.028	0.552±0.032	0.220±0.026	0.604±0.031
Mistral-Small-24B-Instruct-2501	×	0.740±0.028	0.708±0.029	0.632±0.031	0.668±0.030	0.556±0.031	0.692±0.029	0.764±0.027	0.792±0.026	0.804±0.025
Ether0-24B	✓	0.064±0.016	0.076±0.017	0.048±0.014	0.080±0.017	0.185±0.025	0.280±0.028	0.144±0.022	0.280±0.028	0.260±0.028
OLMo-3.1-32B-Instruct	×	0.744±0.028	0.748±0.028	0.580±0.031	0.708±0.029	0.490±0.032	0.496±0.032	0.832±0.024	0.872±0.021	0.864±0.022
OLMo-3.1-32B-Think	✓	0.620±0.031	0.592±0.031	0.516±0.032	0.548±0.032	0.498±0.032	0.500±0.032	0.680±0.030	0.696±0.029	0.540±0.032
GPT-5.4-Mini	✓	**0.904±0.019**	**0.912±0.018**	**0.844±0.023**	**0.900±0.019**	0.526±0.032	0.508±0.032	0.884±0.020	**0.912±0.018**	0.924±0.017
Claude-Haiku-4.5	✓	0.796±0.026	0.812±0.025	0.760±0.027	0.764±0.027	0.498±0.032	0.500±0.032	0.912±0.018	0.892±0.020	0.860±0.022

Table 20: Tautomer Recognition (accuracy)

Model	Reasoning	Canonical SMILES	Isomeric SMILES	Randomized SMILES	DeepSMILES	IUPAC	SELFIES	MolJSON	CML	InChI
Qwen3-4B	✓	0.990±0.007	0.990±0.007	0.824±0.026	0.971±0.012	0.495±0.035	0.500±0.035	0.957±0.014	0.881±0.022	0.843±0.025
Qwen3-4B	×	**0.995±0.005**	0.981±0.009	0.819±0.027	0.976±0.011	0.495±0.035	0.500±0.035	0.967±0.012	0.929±0.018	0.848±0.025
Qwen3-30B	✓	0.986±0.008	**0.995±0.005**	0.810±0.027	0.967±0.012	0.548±0.035	0.510±0.035	0.990±0.007	**0.990±0.007**	0.781±0.029
Qwen3-30B	×	0.976±0.011	0.986±0.008	0.833±0.026	0.981±0.009	0.524±0.035	0.510±0.035	0.990±0.007	0.986±0.008	0.776±0.029
Phi-4	✓	0.981±0.009	0.981±0.009	0.895±0.021	0.943±0.016	0.514±0.035	0.529±0.035	0.981±0.009	0.724±0.031	**1.000**
Phi-4-Reasoning	✓	0.986±0.008	0.981±0.009	0.848±0.025	0.938±0.017	0.500±0.035	0.505±0.035	0.990±0.007	0.952±0.015	**1.000**
Phi-4-Reasoning-Plus	✓	0.971±0.012	**0.995±0.005**	0.838±0.025	0.976±0.011	0.538±0.035	0.510±0.035	0.986±0.008	0.957±0.014	0.995±0.005
Qwen2.5-14B	×	0.229±0.029	0.229±0.029	0.157±0.025	0.229±0.029	0.279±0.031	0.205±0.028	0.200±0.028	0.252±0.030	0.190±0.027
ChemDFM-v2.0-14B	×	0.876±0.023	0.848±0.025	0.857±0.024	0.710±0.031	**0.611±0.034**	**0.633±0.033**	0.776±0.029	0.390±0.034	0.833±0.026
ChemDFM-R-14B	✓	0.952±0.015	0.914±0.019	0.752±0.030	0.857±0.024	0.562±0.034	0.548±0.034	0.724±0.031	0.371±0.033	0.752±0.030
Mistral-Small-24B-Instruct-2501	×	0.990±0.007	**0.995±0.005**	0.929±0.018	0.962±0.013	0.500±0.035	0.538±0.034	0.948±0.015	0.933±0.017	0.962±0.013
Ether0-24B	✓	0.376±0.034	0.338±0.033	0.252±0.030	0.343±0.033	0.264±0.031	0.471±0.035	0.395±0.034	0.410±0.034	0.595±0.034
OLMo-3.1-32B-Instruct	×	0.895±0.021	0.862±0.024	0.786±0.028	0.876±0.023	0.481±0.035	0.519±0.035	0.929±0.018	0.938±0.017	0.905±0.020
OLMo-3.1-32B-Think	✓	0.895±0.021	0.905±0.020	0.600±0.034	0.933±0.017	0.510±0.035	0.495±0.035	0.895±0.021	0.952±0.015	0.690±0.032
GPT-5.4-Mini	✓	**0.995±0.005**	**0.995±0.005**	**0.967±0.012**	**1.000**	0.510±0.035	0.500±0.035	**0.995±0.005**	**0.990±0.007**	**1.000**
Claude-Haiku-4.5	✓	**0.995±0.005**	0.986±0.008	0.943±0.016	0.990±0.007	0.495±0.035	0.500±0.035	0.990±0.007	0.871±0.023	**1.000**

Table 21: Protonation State Recognition (accuracy)

L CKA similarity between Qwen2.5 and ChemDFMs across representations



(a) Qwen2.5-14B (base) (b) ChemDFM-v2.0-14B (c) ChemDFM-R-14B

Figure 9: Linear CKA similarity between representation pairs of chemistry-specialized models, ChemDFM-v2.0 and ChemDFM-R, and their base model Qwen2.5-14B. Higher values (darker red) indicate the model encodes two representations more similarly.

M t-SNE plots of molecular text representations across different LLMs

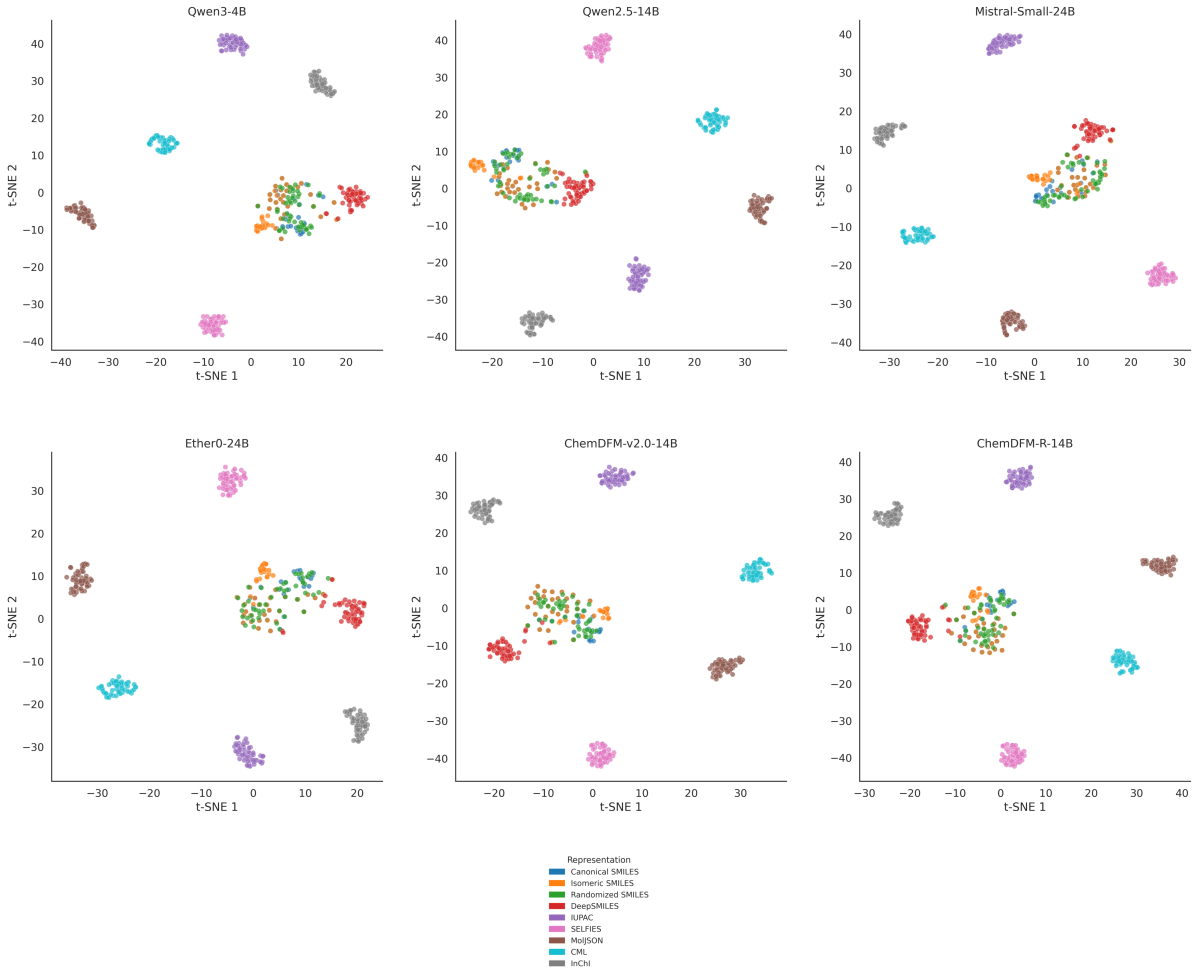


Figure 10: t-SNE plots of the various text representations of 10 molecules with different LLMs