LAUGHS: An LLM-compatible Molecular String Representation

Anonymous Authors[1]

Abstract

Large language models (LLMs) are increasingly applied to chemistry, yet their performance depends strongly on how molecules are represented as text. IUPAC names become syntactically unwieldy for complex structures, while graph-serialized strings disperse chemically meaningful moieties across the sequence. Here, we present LAUGHS, an LLM-compatible molecular string representation that decomposes a molecule into named moieties, hierarchically organizes them into a tree structure, and linearizes the result into a natural-language-like string. Tokenization analysis reveals that LAUGHS units align near-perfectly with tokenizer spans, suggesting strong compatibility with LLMs. On the property explanation task, LAUGHS matches IUPAC-level performance across all metrics; on site-specific editing, it substantially outperforms all baselines with a 91.4% exact match rate among valid outputs. Together, our results suggest that semantic mismatch between molecular representations and natural language syntax is a key bottleneck for LLMs in chemistry, and that LAUGHS offers an effective way to address it.

1. Introduction

Pretrained language models have advanced a wide range of natural language processing capabilities, including language understanding, representation learning (Devlin et al., 2019; Wang et al., 2018; 2019; Liu et al., 2019), and text generation (Radford et al., 2019; Brown et al., 2020). More recent instruction-following LLMs (Ouyang et al., 2022; Achiam et al., 2023; Comanici et al., 2025; Yang et al., 2025; Guo et al., 2025) now serve as general-purpose interfaces for question answering (Yue, 2025), summarization (Zhang et al., 2025), translation (Gain et al., 2026), and code generation (Li et al., 2022; Roziere et al., 2023; Jiang et al., 2026). Chemistry is highly compatible with language model architectures because molecules are routinely represented in textual form. Consequently, these architectures have been broadly adopted for molecular property prediction (Ross et al., 2022), reaction outcome prediction (Schwaller et al., 2019), retrosynthesis planning (Schwaller et al., 2020), and de novo molecular generation and optimization. (Bagal et al., 2021; He et al., 2022)

Table 1. Comparison syntactic features of representations. NL: natural language; IUPAC: IUPAC nomenclature; SMILES: Simplified Molecular Input Line Entry System (Weininger, 1988); Spaced units: space-delimited semantic units; Head-modifier syntax: phrase-level composition via explicit markers; Locant: position-specific substitution; Combinatorial: structural assembly via the arrangement of discrete building blocks. ✓/✗ indicate presence/absence.

	Spaced units	Head -modifier	Locant	Combi -natorial
NL	✓	✓	✗	✗
IUPAC	✗	✗	✓	✗
SMILES	✗	✗	✗	✓
LAUGHS (ours)	✓	✓	✓	✓

LLMs receive chemical information about molecules through their textual representations. However, the most widely used molecular representations were developed primarily for systematic nomenclature or machine learning pipelines, and thus diverge considerably from the structural characteristics of natural language. Table 2 summarizes the key properties of natural language with representative molecular representations. IUPAC nomenclature provides the standard system for molecular names (Favre & Powell, 2013), but its detailed rules can make names of complex molecules difficult to understand (Fendos, 2021). SMILES (Weininger, 1988) remains the predominant molecular representation in cheminformatics owing to its compactness and pervasive integration into data pipelines. However, while its traversal syntax surfaced encodes atom- and bond-level information, higher-level chemical abstractions, such as scaffolds and attachment sites, are left implicit and must be inferred from that syntax. Although alternative encodings show that redesigning molecular encodings can improve validity (Krenn et al., 2019; 2022), tokenization consistency (Li & Fourches, 2021), and fragment granularity (Cheng

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the AI for Science workshop (ICML 2026).

et al., 2023), they do not directly address high-level abstractions in a format that language models can effectively leverage.

In this study, we present LAUGHS, an LLM-compatible molecular string representation designed to expose chemical groups. LAUGHS is constructed in three stages: decomposing the molecule into named chemical moieties, organizing them into a root-substituent hierarchy, and linearizing that tree using natural-language-like connectors and delimiter. We analyze how LAUGHS compatible with LLM tokenizers. For downstream tasks, we evaluate structure-grounded property explanation and localized structure editing.

2. Related Work

Prior work on molecular string representations has largely focused on encodings more amenable to deep learning models for property prediction(Jablonka et al., 2024) and molecular generation(Bagal et al., 2021; Chilingaryan et al., 2024; Özçelik et al., 2024). Beyond SMILES, SELFIES (Krenn et al., 2019) guarantees the validity of generated molecules through its robust syntax, and SMILES Pair Encoding (Li & Fourches, 2021) uses data-driven substructure vocabularies. Another line of work has explored fragment-level representations in which functional groups and substructures serve as building blocks for molecular generation and design (Wu et al., 2024), with related extensions marked encoding attachment patterns (Noutahi et al., 2024) and chirality (Mastrolorito et al., 2025).

Existing string encodings were carried over as inputs to language models in chemistry, spanning conventional molecular modeling tasks and agentic workflows. Early molecular language modeling centered on Transformers that were pretrained on SMILES and applied to property prediction (Ross et al., 2022) and sequence-to-sequence tasks such as reaction prediction, retrosynthesis, and molecular optimization (Irwin et al., 2022). Text2Mol and MolT5 (Edwards et al., 2021; 2022) marked an important transition by connecting molecular strings with natural language through bidirectional translation, enabling both molecule-to-text description and text-to-molecule generation. At the scale of LLMs, later work extended this framing by recasting domain tasks as instructions, including Mol-Instructions (Fang et al., 2023), a biomolecular instruction dataset, and ChemLLM (Zhang et al., 2024), an LLM for chemistry with instruction data, benchmarking, and dialogue interaction. Meanwhile, agentic systems leverage language models within tool-using workflows (M. Bran et al., 2024; Zou et al., 2025; Kim et al., 2025). Across this progression, molecular string encodings have been readily adopted by language models, but their suitability for the way LLMs process and reason over chemical information has often been overlooked.

Recent studies have shown that LLM performance in chemistry depends on how a molecule is presented to the model. In property prediction, accuracy varies across molecular representations (e.g. SMILES, DeepSMILES, SELFIES, InChI, and IUPAC names) (Baker et al., 2025). Consistency analyses likewise show that the same molecule can elicit different answers when written as SMILES or as an IUPAC name (Yan et al., 2025). In generative tasks, this sensitivity extends to validity and correction, where format choice and repair strategies affect whether generated strings correspond to valid structures (Tao et al., 2025). Taken together, these studies indicate that molecular encodings do more than specify its identity; they also shape how models behave on chemical tasks, underscoring the need for representations tailored to LLMs.

3. Method

3.1. LAUGHS Representation

Figure 1 illustrates these three stages of LAUGHS encoding for donepezil and highlights how the LAUGHS preserves the decomposition into a root moiety and its substituents. We implement this pipeline as an encoder-decoder pair. Vocabulary statistics are summarized in Appendix A.3.Round-trip reconstruction accuracy of the LAUGHS encoder and decoder is documented in Appendix A.4.

LAUGHS (**LA**nguage for molec**U**lar **G**roup-aware **H**ierarchical **S**trings) represents molecules as natural-language phrases over chemical groups, where the fundamental unit of description is a named moiety and molecular structure is expressed through modifier-clause syntax. Its grammar is grounded in two design principles that mirror the structure of natural language: named chemical groups serve as the basic units of description, and the representation adopts a phrase structure analogous to natural language–comprising a head, modifier clauses, connectors, and delimiters. LAUGHS is constructed from a molecular graph through three stages: fragmentation, tree construction, and linearization.

3.1.1. Fragmentation

The first stage decomposes the input molecule into moieties. We define a vocabulary of frequently occurring named rings, functional groups, and carbon chains as the basic units, and decompose input molecules into these units. All decompositions occur exclusively at single bonds. The detailed matching and bond-cutting rules are given in Appendix A.1.

3.1.2. Hierarchical Moiety Tree Construction

The resulting moieties are then organized into a rooted hierarchy. A priority scheme selects the global root moiety (the core) and orders the attached moieties (substituents)

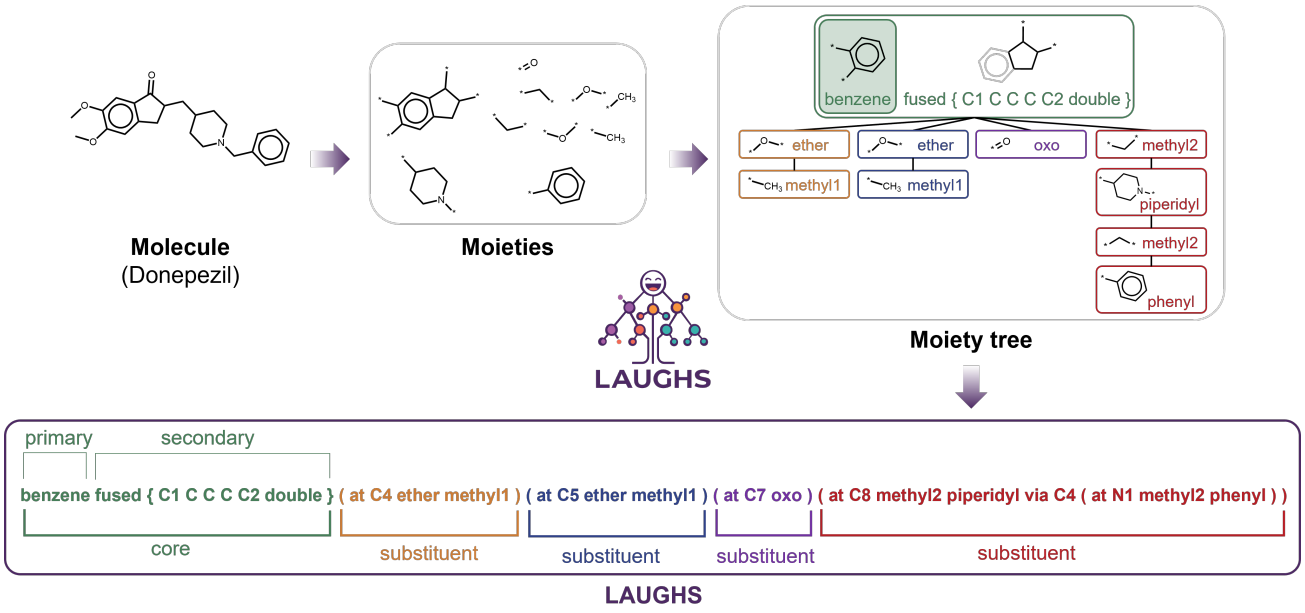

Figure 1. LAUGHS construction example for donepezil. The molecule is first decomposed into a root moiety and its substituents, organized into a rooted moiety tree, and finally rendered as a linear LAUGHS string with named groups and locants. The color coding highlights the correspondence between tree nodes and their spans in the final LAUGHS string.

across the molecular graph. A specialized rule set handles parent–child assignment and locant ordering within complex ring systems. Together, these rules define the moiety tree with one node per moiety and one edge for each parent-child attachment relation. Each edge records the information required to realize that relation in the final LAUGHS string, including attachment-site locants and, when necessary, bond orders. The full hierarchy rules are detailed in Appendix A.2.

3.1.3. Linearization

Once the rooted moiety tree has been defined, the encoder converts it into a LAUGHS string. The tree is traversed via depth-first search, and based on the node and edge information, units are sorted and converting into a modifier-clause string. Its units include parent-form moieties, substituent names, locants, attachment markers such as `at` and `via`, and complex-ring topology markers such as `fused`, `spiro`, and `bridged`. Most units are whitespace-delimited, whereas parentheses and braces serve as dedicated structural tokens. Moieties that share the same name but differ in bonding valence are distinguished by a numerical suffix (e.g., a primary amine is denoted `amino1` and a tertiary amine `amino3`).

3.2. Tokenizer Alignment

We analyzed tokenizer alignment to provide a quantitative measure of how compatible the representation is with LLMs. To distinguish tokenizer-defined units from the lexical units defined by the representation itself, we use the term "token" to refer specifically to the former. We report two tokenization statistics: token purity(Purity) and tokenization consistency (Consistency). Purity measures the fraction of tokens that correspond to exactly one representation unit. Consistency measures the fraction of unit occurrences that follow the most common tokenization pattern. Tokenizer Alignment (Alignment) is calculated by Purity $\times$ Consistency and reflects unit-token one-to-one mapping. Formal definitions are provided in Appendix D. Tokenization was performed using OpenAI's `tiktoken` library, specifically the `o200k_base` encoding used for GPT-4.1-mini. We analyze tokenizer alignment on 14,307 molecules derived from the BACE, BBBP, HIV, and Tox21 benchmarks in MoleculeNet (Wu et al., 2018).

3.3. Evaluation

3.3.1. Property Explanation

The property explanation task evaluates whether the molecular representation gives LLMs enough structural evidence to produce a chemically meaningful account of a molecular property. On 500 BACE molecules from MoleculeNet (Wu et al., 2018), GPT-4.1-mini was employed to generate property-oriented explanations from four input representations: IUPAC names (abbreviated as IUPAC), SMILES, SELFIES, and LAUGHS.

The outputs are then assessed with an LLM-as-a-judge setup (Zheng et al., 2023): GPT-5.4-mini serves as the judge un-

der the G-Eval framework (Liu et al., 2023) implemented through the DeepEval package. For Plausibility, Groundedness, Specificity, and Consistency, we use separate G-Eval rubrics and report their arithmetic mean as the overall explanation score (Overall). The generation prompt is given in Appendix E, and the judging criteria and prompt template are summarized in Appendix F.

3.3.2. Structure Editing

We constructed an editing-instruction dataset from 500 seed molecules sampled from MOSES (Polykovskiy et al., 2020). The dataset covers four molecular representations (IUPAC, SMILES, SELFIES, and LAUGHS) and three edit types (elimination, substitution, and swap). For each seed molecule and edit type, we generated five instruction variants, yielding 7,500 instructions in total.

Each benchmark instance consists of an original molecule, a natural-language edit instruction, and a reference edited molecule, ensuring that the intended structural change is matched across representations. We evaluated GPT-4.1-mini on localized structure editing with this dataset, and scored Validity (the fraction of outputs that decode to valid molecules) and Exact Match (the fraction of outputs that structurally match the reference edited molecule) as a representative metrics.

4. Results

4.1. Tokenizer Alignment

Tokenizer alignment measures how well tokenizer-recognized spans align with the semantic units of a representation, indicating its suitability for LLM processing. As shown in Figure 2, LAUGHS attains near-perfect scores of 0.999 across all metrics, demonstrating that its grammatical structure provides LLMs with highly stable and consistent input.

Tokenizer alignment reflects the grammatical characteristics of a representation along two distinct dimensions. IUPAC achieved strong token purity (0.753), but exhibited the lowest tokenization consistency (0.767), indicating that the prefix- and suffix-based string transformation rules governing IUPAC names give rise to diverse tokenization patterns. Conversely, SELFIES reached perfect tokenization consistency (1.000), but its token purity drops to 0.619, suggesting that stable segmentation alone does not guarantee chemically meaningful token boundaries. SMILES underperformed in both Purity (0.616) and Consistency (0.778), which can be attributed to its graph-traversal linearization scheme, which disperses chemical information across heterogeneous string patterns.

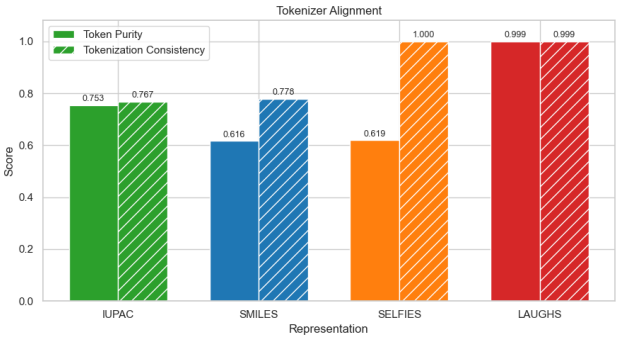

Figure 2. Tokenizer alignment scores across molecular representations. Token Purity (solid) and Tokenization Consistency (hatched) are reported for IUPAC, SMILES, SELFIES, and LAUGHS. LAUGHS achieves near-perfect scores on both metrics (0.999), whereas baseline representations show diverging patterns across the two metrics: SELFIES reaches perfect Consistency (1.000) but low Token Purity (0.619), and IUPAC shows the reverse trend with higher Token Purity (0.753) but lower Consistency (0.767). SMILES underperforms on both metrics.

4.2. Property Explanation

Property explanation tests whether a molecular representation gives an LLM enough structural evidence to produce a chemically meaningful account of a molecular property. In Table 2, IUPAC achieves highest score in all property explanation metrics, followed by LAUGHS. The leading performance of IUPAC may reflect both its chemically systematic structure and its likely prevalence in pretraining corpora. Notably, LAUGHS surpasses SMILES and SELFIES in this zero-shot setting, despite the complete LAUGHS syntax not being part of standard molecular representations.

Among our criteria, Groundedness is particularly revealing, as it assesses whether an explanation anchors to structural evidence such as functional groups and attachment positions. IUPAC and LAUGHS show the closest agreement on this metric (0.019), outperforming traversal-oriented strings. Such alignment suggests that LAUGHS's hierarchical and moiety-based form offers stronger support for structure-grounded answering. Furthermore, this advantage likely arises because LAUGHS leverages familiar chemical terms, which partially parallels the system in IUPAC nomenclature.

LAUGHS outperforms SMILES and SELFIES across all explanation metrics. Its Overall score surpasses these baselines by 0.095 and 0.066, respectively, which are substantial margins for this evaluation setting. When compared to SELFIES, the advantage is most notable in Specificity, where LAUGHS exceeds it by 0.130, indicating that LAUGHS generates molecule-specific responses rather than generic property descriptions. Relative to SMILES, the gains are distributed across Plausibility, Specificity, and Consistency, all three exceeding a margin of 0.10 for LAUGHS. Based on these results, the explanatory advantage of LAUGHS stems

Table 2. Comparison of molecular representations across property explanation and structure editing. Task-specific metrics are defined in the method section. Best values are shown in bold and second-best values are underlined.

Task	Metric	IUPAC	SMILES	SELFIES	LAUGHS
Property Explanation	Plausibility	**0.721**	0.574	0.633	<u>0.679</u>
	Groundedness	**0.603**	0.516	0.522	<u>0.584</u>
	Specificity	**0.789**	0.644	0.617	<u>0.747</u>
	Consistency	**0.784**	0.637	0.715	<u>0.742</u>
Structure Editing	Validity	0.906	0.832	**0.976**	<u>0.975</u>
	Exact Match	<u>0.811</u>	0.699	0.227	**0.914**

from pairing chemical named vocabulary with explicit structural relations.

4.3. Structure Editing

Structure editing evaluates whether LLMs can identify a target substructure and apply local modifications while preserving the rest of the molecule. Within Table 2, LAUGHS outperforms exact match rate (0.914), followed by IUPAC, SMILES, and SELFIES. This result suggests that LAUGHS generates more LLM-tractable tokens than IUPAC, a finding corroborated by its superior tokenizer alignment scores. SELFIES scored considerably lower than even SMILES — another atom-level representation — which can be attributed to the fact that its atomic units simultaneously carry grammatical roles, making it difficult to correctly reassign these functional units through local edits alone.

In terms of Validity, SELFIES, which was designed for syntactic robustness, attains the highest score (0.976), while LAUGHS reaches a nearly identical level (0.975). By contrast, IUPAC and SMILES exhibit lower validity despite their widespread use. Although SELFIES was designed to operate exclusively with syntactically valid units, imperfect alignment between those units and LLM tokens can still produce unsupported expressions. Nevertheless, the high validity scores of both SELFIES and LAUGHS demonstrate that principled representation design can improve the syntactic validity of LLM-generated outputs. F1-scores follow the same ordering (Appendix G).

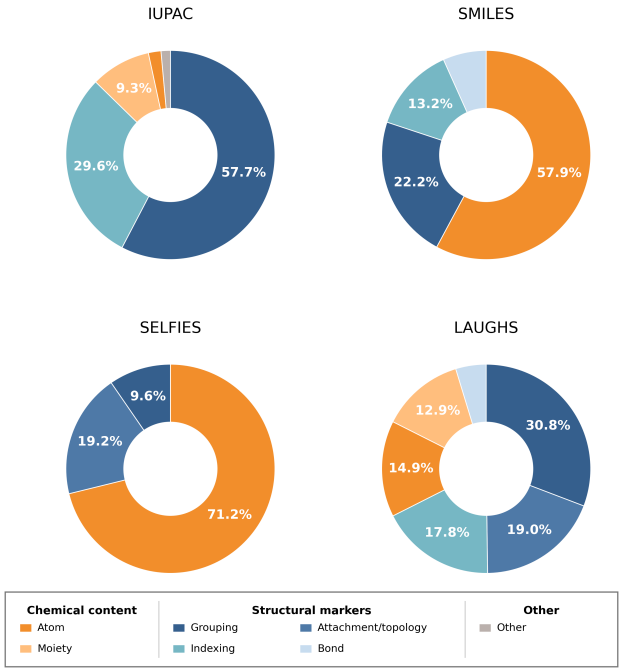

Figure 3. Comparative functional composition of high-frequency units. Distributions across seven roles reveal the occurrence patterns for the top 50 units in each representation. LAUGHS spans a diverse repertoire of moiety and structural markers, whereas baselines skew toward restrictive formal roles (e.g., atom tokens in SELFIES, grouping markers in IUPAC).

5. Discussion

5.1. How representation syntax governs LLM performance

Our results show a consistent pattern across all evaluation tasks: representations that more closely mirror natural language syntax elicit stronger LLM performance. This pattern cannot be explained by chemical information alone. SMILES and SELFIES encode complete molecular graphs, yet they lag behind LAUGHS and IUPAC on explanation and editing alike. Conversely, IUPAC, the most chemically systematic representation, leads on property explanation but falls substantially behind LAUGHS on editing task. The factor that best accounts for performance differences across all tasks is how well each representation's syntactic structure aligns with the statistical regularities of natural language that LLMs internalize during pretraining.

The distribution of unit types across representations explains the observed differences in LLM performance. Figure 3 categorizes the 50 most frequent units in each representation into seven functional roles. LAUGHS has the highest proportion of moiety tokens, followed by IUPAC, while IUPAC ranks first and LAUGHS second in indexing tokens. Representations rich in moiety and indexing information — namely IUPAC and LAUGHS — tend to elicit explanations that more specifically characterize substituent identity and

attachment position. By contrast, SMILES and SELFIES convey chemical information only at the atomic level, which limits the recognizable structural features available to the model and consequently reduces Specificity while increasing the rate of implausible attributions. Grouping units serve to encode representational grammar; the high proportion in IUPAC reflects its rule-dense locant syntax. Paradoxically, this highly systematic framework also contributes to the difficulty of generating exact molecular strings, a challenge that has been similarly documented among human learners (Bodé et al., 2016; Fendos, 2021; Taskin & Bernholt, 2014).

5.2. Natural language structure as a design principle for LLM-facing representations

The consistent advantage of LAUGHS across evaluation tasks can be attributed to three structural features that collectively make it resemble natural language more closely than any of the baseline representations.

Lexically stable unit identities. LAUGHS assigns each structural unit a fixed lexical identity regardless of molecular context: a terminal methyl group is consistently represented as `methyl1`, and a quaternary carbon center as `methyl4`. Its syntax provides a stable token-to-meaning association, which stands in contrast to SMILES and SELFIES, where the structural role of an atom token is entirely determined by its position in the traversal sequence. The high Specificity and Groundedness scores of LAUGHS in property explanation, and its superior exact-match rate in structure editing, reflect the model's ability to identify and reason about stable unit identities without requiring contextual inference.

Explicit modifier-clause syntax for branching. LAUGHS encodes molecular branching through a post-modification syntax in which connectors such as `at` and `via` explicitly specify attachment sites, in the same way that prepositional phrases in natural language specify the relationship between modifier and its head. In contrast, SMILES parentheses and SELFIES branch tokens encode topological branching without lexical content, requiring the model to reconstruct attachment context from positional reasoning. This distinction accounts for the editing advantage of LAUGHS: targeted structural modifications correspond to local clause substitutions in LAUGHS, whereas context-dependent representations require globally consistent re-traversal of the molecular string.

Whitespace delimitation and tokenizer alignment. LLM tokenizers are optimized to segment whitespace-separated strings into stable and reusable units. This design directly addresses a known tokenization bottleneck in chemistry LLMs: general-domain tokenizers tend to fragment SMILES into semantically uninformative subtokens, whereas chemistry-specific molecular tokenizers can still suffer from incomplete coverage of the SMILES/OpenSMILES space (Kalamkar et al., 2025; Wadell et al., 2025).The whitespace-delimited architecture of LAUGHS produces near-perfect tokenizer alignment (0.999), as each LAUGHS unit maps onto tokenizer spans in the same manner as natural language words. SMILES, without delimiters, yields fragmented and inconsistent tokenization. IUPAC, though lexically familiar, compresses complex structural information into long compound tokens that segment unpredictably. SELFIES achieves consistent tokenization at the symbol level but only by sacrificing meaningful unit boundaries. LAUGHS achieves high scores on both Purity and Consistency simultaneously, because whitespace delimitation directly instantiates the segmentation assumptions underlying the tokenizer.

These three properties suggest that syntactic alignment with natural language conventions may be a productive design consideration for LLM-compatible molecular representations, independent of chemical information completeness. Chemical completeness and compact encodings are essential for cheminformatics pipelines, but do not necessarily translate to stronger LLM performance. LAUGHS offers one instantiation of how these objectives can be reconciled.

5.3. Limitations

We introduced a molecular format that marks structural features of molecules, but several limitations remain. First, our tokenizer analysis focused on a single setting. Although the ability of language models to understand a representation may vary with the tokenizer, we restricted our analysis to the tokenizer used for GPT-4.1-mini. Demonstrating the generality of the representation will require benchmarking across multiple tokenizers. Second, because our evaluation used pretrained LLMs, the composition of the pretraining data is unknown. This may provide prior knowledge for widely used representations such as IUPAC and SMILES, while disadvantaging more recent formats such as SELFIES and LAUGHS. Third, the current implementation of LAUGHS does not cover stereochemistry. This limitation will be addressed in future work on chemical reaction tasks.

6. Conclusion

We present LAUGHS, a human-readable hierarchical molecular string representation designed to decompose molecules into named moieties, hierarchically construct a tree structure, and describe it into a natural language-like text format. Across property explanation, localized structure editing, and tokenizer compatibility analysis, our results demonstrate that surfacing chemically meaningful units and utilizing delimiter enhance language models' ability to interpret and manipulate molecular structure. In particular, LAUGHS matches explanation quality close to IUPAC while achiev-

ing the strongest editing performance overall. Moreover, it attains near-perfect compatibility between tokens and their chemical units.

This work redefines molecular representation as a strategic interface design problem. Aligning chemical abstractions with tokenization scheme provides a robust foundation for LLM-driven chemistry, ensuring that representations act not merely as storage formats but as communication channels between human knowledge and model processing.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Bagal, V., Aggarwal, R., Vinod, P., and Priyakumar, U. D. Molgpt: molecular generation using a transformer-decoder model. *Journal of chemical information and modeling*, 62(9):2064–2076, 2021.

Baker, G. A., Sanz-Guerrero, M., and von der Wense, K. Molecular string representation preferences in pre-trained llms: A comparative study in zero- & few-shot molecular property prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1071–1085, 2025. doi: 10.18653/v1/2025.emnlp-main.56.

Bodé, N. E., Caron, J., and Flynn, A. B. Evaluating students' learning gains and experiences from using http://nomenclature101.com. *Chemistry Education Research and Practice*, 17(4):1156–1173, 2016.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Cheng, A. H., Cai, A., Miret, S., Malkomes, G., Phielipp, M., and Aspuru-Guzik, A. Group selfies: a robust fragment-based molecular string representation. *Digital Discovery*, 2(3):748–758, 2023.

Chilingaryan, G., Tamoyan, H., Tevosyan, A., Babayan, N., Hambardzumyan, K., Navoyan, Z., Aghajanyan, A., Khachatrian, H., and Khondkaryan, L. Bartsmiles: Generative masked language models for molecular representations. *Journal of Chemical Information and Modeling*, 64(15):5832–5843, 2024.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

Edwards, C., Zhai, C., and Ji, H. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 595–607, 2021.

Edwards, C., Lai, T., Ros, K., Honke, G., Cho, K., and Ji, H. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 375–413, 2022.

Fang, Y., Liang, X., Zhang, N., Liu, K., Huang, R., Chen, Z., Fan, X., and Chen, H. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*, 2023.

Favre, H. A. and Powell, W. H. *Nomenclature of organic chemistry: IUPAC recommendations and preferred names 2013*. Royal Society of Chemistry, 2013.

Fendos, J. Combining jigsaws, rule-based learning, and retrieval practice improves IUPAC nomenclature competence. *Journal of Chemical Education*, 98(5):1503–1517, 2021.

Gain, B., Bandyopadhayay, D., Ekbal, A., and Singh, T. N. Bridging the linguistic divide: a survey on leveraging large language models for machine translation. *Language Resources and Evaluation*, 60(2):40, 2026.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

He, J., Nittinger, E., Tyrchan, C., Czechtizky, W., Patronov, A., Bjerrum, E. J., and Engkvist, O. Transformer-based molecular optimization beyond matched molecular pairs. *Journal of cheminformatics*, 14(1):18, 2022.

Irwin, R., Dimitriadis, S., He, J., and Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, 2022.

Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A., and Smit, B. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, 2024.

Jiang, J., Wang, F., Shen, J., Kim, S., and Kim, S. A survey on large language models for code generation. *ACM Transactions on Software Engineering and Methodology*, 35(2):1–72, 2026.

Kalamkar, P., Letcher, N., Chami, M., Lad, S., Mohanty, S., and Pendse, P. The tokenization bottleneck: How vocabulary extension improves chemistry representation learning in pretrained language models, 2025.

Kim, H., Jang, Y., and Ahn, S. Mt-mol: Multi agent system with tool-based reasoning for molecular optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11544–11573, 2025. doi: 10.18653/v1/2025.findings-emnlp.619.

Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. Selfies: a robust representation of semantically constrained graphs with an example application in chemistry. *arXiv preprint arXiv:1905.13741*, 1(3), 2019.

Krenn, M., Ai, Q., Barthel, S., Carson, N., Frei, A., Frey, N. C., Friederich, P., Gaudin, T., Gayle, A. A., Jablonka, K. M., et al. Selfies and the future of molecular string representations. *Patterns*, 3(10), 2022.

Li, X. and Fourches, D. Smiles pair encoding: a data-driven substructure tokenization algorithm for deep learning. *Journal of chemical information and modeling*, 61(4): 1560–1569, 2021.

Li, Y., Choi, D., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Eccles, T., Keeling, J., Gimeno, F., Dal Lago, A., et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023.

M. Bran, A., Cox, S., Schilter, O., Baldassari, C., White, A. D., and Schwaller, P. Augmenting large language models with chemistry tools. *Nature machine intelligence*, 6(5):525–535, 2024.

Mastrolorito, F., Ciriaco, F., Togo, M. V., Gambacorta, N., Trisciuzzi, D., Altomare, C. D., Amoroso, N., Grisoni, F., and Nicolotti, O. fragsmiles as a chemical string notation for advanced fragment and chirality representation. *Communications Chemistry*, 8(1):26, 2025.

Noutahi, E., Gabellini, C., Craig, M., Lim, J. S., and Tossou, P. Gotta be safe: a new framework for molecular design. *Digital Discovery*, 3(4):796–804, 2024.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Özçelik, R., de Ruiter, S., Criscuolo, E., and Grisoni, F. Chemical language modeling with structured state space sequence models. *Nature Communications*, 15(1):6176, 2024.

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., et al. Molecular sets (moses): a benchmarking platform for molecular generation models. *Frontiers in pharmacology*, 11:565644, 2020.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., and Das, P. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., Adi, Y., Liu, J., Sauvestre, R., Remez, T., et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

Schwaller, P., Petraglia, R., Zullo, V., Nair, V. H., Haeuselmann, R. A., Pisoni, R., Bekas, C., Iuliano, A., and Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical science*, 11(12):3316–3325, 2020.

Tao, W., Tang, J., Chan, A., Hooi, B., Bi, B., Peng, N., Liu, Y., and Wang, Y. How to make large language models generate 100% valid molecules? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26565–26580, 2025.

Taskin, V. and Bernholt, S. Students' understanding of chemical formulae: A review of empirical research. *International Journal of Science Education*, 36(1):157–185, 2014.

Wadell, A., Bhutani, A., and Viswanathan, V. Tokenization for molecular foundation models, 2025.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pp. 353–355, 2018.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Wu, J.-N., Wang, T., Chen, Y., Tang, L.-J., Wu, H.-L., and Yu, R.-Q. t-smiles: a fragment-based molecular representation framework for de novo ligand design. *Nature Communications*, 15(1):4993, 2024.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Yan, B., Chen, A., and Cho, K. Inconsistency of llms in molecular representations. *Digital Discovery*, 4(10): 2876–2892, 2025.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yue, M. A survey of large language model agents for question answering. *arXiv preprint arXiv:2503.19213*, 2025.

Zhang, D., Liu, W., Tan, Q., Chen, J., Yan, H., Yan, Y., Li, J., Huang, W., Yue, X., Ouyang, W., et al. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*, 2024.

Zhang, H., Yu, P. S., and Zhang, J. A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, 57(11):1–41, 2025.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zou, Y., Cheng, A. H., Aldossary, A., Bai, J., Leong, S. X., Campos-Gonzalez-Angulo, J. A., Choi, C., Ser, C. T., Tom, G., Wang, A., Zhang, Z., Yakavets, I., Hao, H., Crebolder, C., Bernales, V., and Aspuru-Guzik, A. El agente: An autonomous agent for quantum chemistry. *Matter*, 8(7):102263, 2025.

A. LAUGHS Construction Details

A.1. Fragmentation Rules

The encoder begins by decomposing the input molecular graph into moieties. Named rings and functional groups are matched with predefined SMARTS patterns, and the remaining non-ring sp^3 carbons are assigned to longest-path alkyl-chain motifs. Fragmentation is selective rather than bond-exhaustive: ring bonds are never cut, and cuts are introduced only when they are needed to separate distinct matched motifs or ring-system boundaries. The inter-moiety bond order is preserved at each cut site and later reused during encoding and decoding.

A.2. Hierarchy Rules

Hierarchy assignment uses two different rule sets. At the whole-molecule level, fragment-to-fragment priority first orients graph edges by connection count, heavy-atom count, and sorted atomic numbers. Root selection for the moiety tree then uses a separate criterion based on unique child-type count, heavy-atom count, atomic numbers, and neighbor-tree priority. Inside a complex ring fragment, parent-child assignment among constituent rings follows an eight-tier rule based on bridgehead count, port count, heavy-atom count, sequential atomic numbers, positional atomic numbers, double-bond count, external-fragment priority, and substituent-tree priority. Locant assignment within each ring further prioritizes bridgehead status and bridgehead traversal order before heteroatom, parent attachment, and substituent priority.

A.3. Surface Vocabulary

The current LAUGHS surface vocabulary comprises 33 structural tokens, 224 motif-name tokens, and 74 locant tokens, for 331 distinct tokens overall. The structural inventory includes keywords, brackets, bond-order markers, fold markers, ring-atom symbols, and bond-range markers. The motif inventory is the union of named fragment tokens together with their parent-form and substituent-form variants. This controlled vocabulary is used to compose the final LAUGHS string and supports paired decoding back to SMILES.

A.4. Round-Trip Reconstruction

For debugging, we evaluated encoder-decoder round-trip reconstruction on 44,671 molecules from the BACE, BBBP, HIV, and Tox21 MoleculeNet benchmarks (Wu et al., 2018) after filtering invalid molecules, ionic species, and organometallics and removing chirality. Round-trip accuracy was 100%.

B. Tokenizer-Facing Examples

Representative LAUGHS strings expose recurring lexical items such as `benzene`, `phenyl`, `hydroxy`, `amide`, `oxo`, `chloro`, `ring`, and `fused`. In contrast, SMILES expresses the same chemical content through character-level fragments and punctuation. We view this difference as the main reason LAUGHS is a promising interface for generic LLM tokenization: the representation reuses stable surface words rather than requiring the model to infer higher-level concepts from traversal syntax alone. In the current study, tokenization is measured with OpenAI's `tiktoken` library using the model-specific encoding for GPT-4.1-mini.

A token denotes a span returned by the tokenizer, whereas a unit denotes a representation-defined lexical item. In LAUGHS, units are largely whitespace-delimited lexical items, with structural punctuation such as parentheses and braces treated as separate tokens. The tokenization results suggest two practical design rules for LLM-interpretable representations. First, a chemically meaningful unit should map to one token whenever possible. Second, the same unit should follow the same tokenization pattern whenever it recurs. The token-unit alignment rate and tokenization consistency reported in the main text make these desiderata explicit and provide a compact way to compare alternative molecular serializations.

C. Tokenizer Alignment Details

C.1. Tokenizer Corpus Construction

The tokenizer alignment corpus is obtained from the BACE, BBBP, Tox21, and balanced HIV MoleculeNet subsets (18,043 rows total). We retain molecules with non-empty IUPAC, SELFIES, and LAUGHS strings, exclude invalid, multi-component, charged, and organometallic molecules, remove stereochemistry, and deduplicate by canonical achiral

Table 3. Summary statistics for the representation-native unit vocabularies on the benchmark-derived aligned four-representation subset described in Appendix C.1. Top-10 share denotes the fraction of all unit occurrences accounted for by the ten most frequent units in that representation.

Representation	Distinct units	Units seen once	Top-10 share	Mean units/string
IUPAC	5,346	2,600	0.666	29.0
SMILES	80	6	0.910	37.6
SELFIES	89	11	0.897	36.6
LAUGHS	386	56	0.672	44.4

SMILES. This procedure yields 14,307 aligned molecules for tokenizer analysis.

C.2. Unit-Type Distribution Analysis

As a descriptive complement to the main tokenizer benchmark, we also analyze the native unit vocabularies of the four molecular representations on this aligned four-view benchmark subset.

For this analysis, units are defined by representation-specific lexical rules rather than by the tokenizer. In IUPAC, we split the string into alphabetic spans, locant numbers, and punctuation symbols such as hyphens, commas, parentheses, brackets, and slashes. In SMILES, units are lexical SMILES tokens such as atom symbols, aromatic atoms, bond symbols, branch markers, bracket atoms, and ring indices. In SELFIES, units are bracketed SELFIES symbols. In LAUGHS, units are whitespace-delimited lexical items, with ( ) { } treated as separate structural tokens.

For the category-level comparison in Figure 3, we collapse the detailed labels into seven functional roles: atom tokens, moiety tokens, indexing markers, bond markers, attachment/topology markers, grouping markers, and other. Atom tokens directly denote atoms or elements; moiety tokens cover named groups, moieties, and chemically descriptive modifiers; indexing markers cover locants and ring or site indices; bond markers cover explicit bond-order symbols; attachment/topology markers cover units that primarily express attachment or higher-order topology; grouping markers cover delimiters and grouping punctuation; and other collects the remaining formal residues. The figure aggregates occurrence counts within the top 50 most frequent units of each representation under this seven-way grouping.

Table 3 summarizes vocabulary breadth and concentration. IUPAC has by far the largest inventory and the heaviest low-frequency tail, with 5,346 distinct units and 2,600 units seen only once in the benchmark subset. SMILES and SELFIES are much more concentrated, with only 80 and 89 distinct units and top-10 share near 0.90. LAUGHS occupies an intermediate regime: it exposes a much broader reusable vocabulary than SMILES or SELFIES, but remains far less diffuse than IUPAC, with 386 distinct units, 56 units seen only once in the benchmark subset, and top-10 share 0.672. These statistics support the view that IUPAC has the broadest lexical inventory, SMILES and SELFIES rely on much narrower recurring inventories, and LAUGHS lies between them.

D. Formal Definition of the Token-Unit Alignment Rate and Tokenization Consistency

Let r denote a representation of length L_r, and let $I_r = \{1, \dots, L_r\}$ be its position index set. Let

$$\mathcal{T}(r) = (T_1, \dots, T_m),$$
$$\mathcal{U}(r) = (U_1, \dots, U_n)$$

be the ordered partitions of I_r induced by tokenizer spans and by representation-unit spans, respectively.

For a token span T, define

$$\nu_r(T) := \sum_{U \in \mathcal{U}(r)} \mathbf{1}(T \cap U \neq \varnothing),$$

the number of representation units overlapped by that token. The token-unit alignment rate is

$$A(r) := \frac{1}{|\mathcal{T}(r)|} \sum_{T \in \mathcal{T}(r)} \mathbf{1}\big(\nu_r(T) = 1\big).$$

Equivalently, $A(r)$ is the fraction of tokens that correspond to exactly one representation unit.

To characterize whether the same unit type is segmented consistently across a corpus, let $\lambda_r(U)$ denote the unit type of $U \in \mathcal{U}(r)$, and let $\pi_r(U)$ denote the ordered tokenization pattern induced on that unit. For a collection $\mathcal{R}$, define

$$N_u(p) := \sum_{r \in \mathcal{R}} \sum_{U \in \mathcal{U}(r)} \mathbf{1}\big(\lambda_r(U) = u,\ \pi_r(U) = p\big).$$

Then let

$$N_u = \sum_p N_u(p),$$
$$M_u = \max_p N_u(p),$$

and define corpus-level tokenization consistency as

$$C(\mathcal{R}) = \frac{\sum_u M_u}{\sum_u N_u}.$$

This is the fraction of representation units that follow the most frequent tokenization pattern for their unit type.

E. Property-Explanation Prompt

Property explanations were generated with a shared instruction template across representations. The system prompt was:

```
You are an expert chemist. Your task is to explain the structural basis for a given
↪  molecular property.

You will receive a molecule in a text representation, along with a property name and its
↪  observed value. Explain why this molecule exhibits that property value based on its
↪  structural features.

IMPORTANT: Do NOT mention the name or type of the molecular representation you are reading
↪  (e.g., do not say "in SMILES notation" or "the LAUGHS representation shows"). Simply
↪  describe the molecular structure and its relationship to the property directly.

Rules:
1. Return a single JSON object with one key: "explanation".
2. Identify at least 2 specific substructures or physicochemical features that contribute
↪  to the observed property value.
3. For each feature, explain the mechanistic link: how does this structural element
↪  promote or hinder the property?
4. Be concrete: name the specific functional groups, ring systems, or atom patterns you
↪  observe.
5. End with one sentence noting limitations of structure-only reasoning.
```

The user template was identical across representations except for the molecular field (`Molecule: {smiles}` or `Molecule: {laughs}`):

```
Sample ID: {id}
Molecule: {representation}
Property: {property_name}
Observed value: {property_value}

Explain the structural basis for the observed property value of this molecule.

Requirements:
- Do NOT predict or re-state the property value. Focus entirely on structural rationale.
- Do NOT mention the name or format of the representation. Just describe the structure.
- Identify specific substructures (functional groups, ring systems, chains) relevant to
↪  the property.
- For each substructure, explain the mechanism by which it contributes to the observed
↪  value.
- Discuss at least 2 distinct structural features with clear structure-to-property
↪  reasoning.
- Avoid generic statements; every claim must reference a concrete part of this molecule.
```

F. G-Eval Judge Criteria and Prompt

Property explanations were evaluated with four separate G-Eval metrics in DeepEval: Plausibility, Groundedness, Specificity, and Consistency. The Overall score reported in the main text is the arithmetic mean of these four scores rather than a fifth independently judged dimension.

The criterion strings passed to DeepEval were:

1. **Plausibility**: Evaluate whether the explanation makes chemically and structurally plausible claims. Penalize chemically implausible or structurally unjustified claims.
2. **Groundedness**: Evaluate whether the explanation is supported by the provided molecular or reaction description. Reward claims tied to explicit features in the input and penalize unsupported leaps.
3. **Specificity**: Evaluate whether the explanation is concrete and structurally specific rather than generic. Reward references to precise groups, positions, linkers, substituents, or localized features.
4. **Consistency**: Evaluate whether the explanation is internally coherent and whether its claims support one consistent structural interpretation. Penalize contradictions or mismatched causal claims.

DeepEval's default G-Eval template constructs the judge prompt in two stages. For each criterion, it first asks the judge model to generate 3-4 evaluation steps from the criterion and the evaluation parameters (`input` and `actual_output`). It then asks the judge to score the test case using those generated steps and to return a JSON object with an integer score from 0 to 10 together with a brief reason. The default template is:

Here {`parameters`} denotes DeepEval's textual placeholder for the fields supplied to the metric. In our setup, it corresponds to `input` and `actual_output`.

```
Step-generation prompt:
Given an evaluation criteria which outlines how you should judge the {parameters},
↪  generate 3-4 concise evaluation steps based on the criteria below.
Evaluation Criteria: {criteria}
Return JSON with the "steps" key as a list of strings.

Scoring prompt:
You are an evaluator.
Given the evaluation steps, assess the response below and return JSON with two fields:
- "score": an integer between 0 and 10
- "reason": a brief explanation grounded in the evaluation steps
Evaluation Steps: {evaluation_steps}
Test Case: {test_case_content}
Parameters: {parameters}
```

G. Representative Edit Operations

The editing benchmark uses functional-group-level instructions derived from MOSES seed molecules (Polykovskiy et al., 2020). For each seed, three edit families are instantiated:

- **elimination**: remove a designated functional group from the molecule,
- **substitution**: replace one functional group with another at the same local position,
- **swap**: exchange one functional group identity for another specified alternative.

Five variants are generated for each edit family, yielding 15 instructions per seed and 7,500 editing instructions in total across 500 seeds. These operations are not a full natural-language design benchmark, but they illustrate the practical advantage of LAUGHS for editing: the editable object is already surfaced as a named root moiety, group, or site.

In the main benchmark, **Validity** denotes the fraction of generated outputs that correspond to valid molecules, and **Exact Match** denotes the fraction of outputs that follow the requested edit exactly. We additionally report F1-score as a

supplementary fidelity measure, using the standard definition $F_1 = 2PR/(P + R)$, where P and R denote precision and recall. The resulting F1-scores are 0.896 for IUPAC, 0.823 for SMILES, 0.370 for SELFIES, and 0.955 for LAUGHS, following the same ordering as the main-text metrics.