Kent Academic Repository

Cook, Frederika and Masala, Giovanni Luca (2026) *Trustworthy Molecular AI: Multi-Dimensional Validation of LLM Generated Chemical Descriptions*. In: Paszynski, Maciej and Barnard, Amanda S. and Zhang, Yongjie Jessica, eds. Computational Science – ICCS 2026 Workshops. 26th International Conference, ICCS 2026, Proceedings Part I. Lecture Notes in Computer Science . pp. 337-345. Springer ISBN 978-3-032-29911-6.

Downloaded from

https://kar.kent.ac.uk/114278/ The University of Kent's Academic Repository KAR

The version of record is available from

https://doi.org/10.1007/978-3-032-29912-3_26

This document version

Author's Accepted Manuscript

DOI for this version

Licence for this version

CC BY (Attribution)

Additional information

For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Versions of research works

Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in ***Title of Journal*** , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries

If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies).

Trustworthy Molecular AI: Multi-Dimensional Validation of LLM Generated Chemical Descriptions

Frederika Cook[1 [0009−0007−3151−3887]] and Giovanni Luca Masala[1 [0000−0001−6734−9424]]

School of Computing, University of Kent, UK
frederikacook@gmail.com
g.masala@kent.ac.uk

Abstract. Large Language Models (LLMs) offer unprecedented opportunities for molecular property prediction, yet exhibit concerning capabilities for generating plausible-sounding but factually incorrect chemical descriptions. Whilst traditional evaluation metrics measure linguistic similarity, they cannot detect chemically impossible claims. Domain benchmarks focus on entity-level verification rather than systematically validating whether claimed structures are computationally possible.
We propose a multi-dimensional validation framework, integrating grammatical fluency and lexical validity checking, `RDKit`-based computational chemistry validation, and DrugBank verification. Evaluating three architecturally distinct LLMs (MolT5, LLaMA, TxGemma) across multiple decoding strategies on 451 antibiotic compounds, we demonstrate that multi-dimensional validation identifies distinct error types invisible to individual metrics. Our interactive dashboard enables inspection at multiple granularities, transforming validation from post-hoc comparison into evidence-based analysis.
Our findings reveal systematic blind spots: perfect grammar accompanies fundamental chemical errors, aggregate metrics mask substantial performance differences, and domain-specific pre-training fails to transfer across chemistry tasks. This work sets foundations for trustworthy AI in chemistry, providing quality assurance infrastructure necessary for deploying LLMs in high-stakes applications that carry material consequences.

Keywords: LLMs · drug discovery · AI hallucinations · metrics

1 Introduction

Drug discovery costs exceed $2.8 billion and require 10-16 years per drug [3], meaning many rare disease treatments never reach patients. Large Language Models (LLMs) offer unprecedented opportunities for molecular property prediction by learning complex patterns from chemical structures [11]. However, LLMs generate plausible-sounding but factually incorrect content termed "hallucinations" [2]. In chemistry, wrong functional groups lead to failed syntheses, and incorrect toxicity predictions risk patient safety. For example, when an LLM generates "This compound's β-lactam ring provides broad-spectrum antibiotic

activity," a critical question remains unaddressed - does this molecule actually possess a β-lactam ring, or is the claim a confident fabrication?

Traditional metrics, such as BLEU [13], ROUGE [9], METEOR [1] and BERTScore [19], address only partial validation dimensions: measuring linguistic similarity against a single, static reference. They cannot detect chemically impossible claims. Chemical fingerprint similarity validates structures but ignores linguistic quality. Lexical matching against DrugBank [18] confirms only what the database covers, leaving unverified claims - whether accurate or hallucinated - undetected. As the molecular foundation of medicine, chemistry is a high-stakes domain where errors carry real-world consequences. Medical AI has addressed this through multi-stage validation frameworks [14, 16], recognising that high-stakes domains require validation assessing multiple independent correctness dimensions.

No existing framework systematically assesses all quality dimensions needed for trustworthy chemistry text generation. This work develops a proof-of-concept multi-dimensional validation framework contrasting reference-based metrics alongside three independent layers: grammatical fluency assessment, lexical validity checking, and computational chemistry validation using `RDKit` [7]. Three architecturally distinct LLMs (MolT5 [4], LLaMA [6], TxGemma [17]), each evaluated across multiple decoding strategies, provide a diverse testbed for examining how architectural and generative choices shape hallucination patterns in chemical description generation.

2 Related Work

Reference-free approaches such as SelfCheckGPT [10] detect linguistic inconsistencies rather than domain errors: LLMs can consistently generate chemically impossible claims whilst passing consistency checks. RAGAS [5], a retrieval-augmented evaluation framework, inherits evaluator biases and exhibits poor calibration for domain-specific correctness. Mol-Hallu [8] applies semantic entailment scoring to molecular descriptions, whilst HalluMat [15] employs multi-source retrieval and contradiction detection for materials science content, but neither approach operates on drug-like molecules, nor cross-validates generated claims against computationally derived molecular properties. Our framework begins to address this gap by systematically combining established techniques — grammar analysis, DrugBank lexical matching, and RDKit-computed structural validation — into a multi-layer diagnostic pipeline, exposing error patterns invisible to any single existing approach.

3 Methods

System Architecture, Data and Models. The framework evaluates LLM-generated molecular descriptions across three independent quality dimensions, each producing diagnostic scores: Layer 1 verifies claimed chemical structures against `RDKit` molecular representations; Layer 2 assesses grammatical and syntactic correctness; and Layer 3 checks terminology against DrugBank vocabularies. Together, these enable categorisation of error types invisible to any single metric.

`SMILES` strings and comprehensive metadata (functional groups, mechanisms of action, physicochemical properties) are extracted from DrugBank 5.1, providing

ground-truth annotations for 451 antibacterial/antimicrobial compounds. Three architecturally distinct LLMs are evaluated: MolT5 (encoder-decoder, ∼220M parameters), specialised in molecule-text translation; LLaMA3.1 (decoder-only, 8B parameters), a general-purpose model; and TxGemma (decoder-only, 9B parameters), optimised for therapeutic applications.

To examine how generation parameters influence hallucination patterns, we systematically varied decoding strategies and prompt formulations across multiple configurations. For beam search, we tested widths of 1, 4, 5, 6, and 8 beams with 1 or 2 groups. For stochastic sampling, temperature $\in [0.1, 1.3]$ (step 0.2), top-$k \in [40, 100]$ (step 20), and top-$p \in [0.7, 0.9, 0.95]$. Nine prompt styles were tested: Prompts A-D employed role-based domain expertise, whilst Prompts E-I targeted task-specific predictions covering LogP estimation, structural optimisation, ADMET (absorption, distribution, metabolism, excretion, and toxicity) profiling, resistance mechanisms, and drug repurposing. This yielded 54 model-strategy-prompt combinations evaluated across all 451 SMILES.

Multi-Dimensional Validation Framework

Layer 1: `RDKit` Computational Chemical Validation. Chemical validation verifies claimed molecular structures against computational representations from `RDKit`. Structural claims are extracted using custom regex patterns, with `SciSpacy` [12] providing named entity recognition for chemical entities in the generated text. Extracted claims are then compared against molecular features computed directly from `SMILES` representations via `RDKit`'s Descriptors and `SMARTS` pattern matching. The pipeline proceeds in four steps: (1) parse the `SMILES` string into a molecule object; (2) extract chemical claims from the generated text (e.g., "contains hydroxyl group"); (3) compute actual molecular features via `RDKit` `SMARTS` pattern matching and descriptor calculation; and (4) classify each claim as verified, hallucinated, or omitted.

The pipeline examines functional groups via `SMARTS` substructure matching with fuzzy string matching (Levenshtein distance threshold of 3 to accommodate terminological variation), ring systems via exact size matching, molecular properties (molecular weight, LogP, H-bond donors/acceptors with tolerance-based thresholds), and peptide sequences. 21 functional group patterns are maintained, with precision and recall calculated per group and hallucinations categorised by type (functional group, ring, or property) for diagnostic error profiling.

Layer 2: Grammar and Linguistic Quality. Linguistic validation assesses grammatical correctness, syntactic validity, and generation completeness using `LanguageTool`'s rule-based checker and `spaCy` for syntactic parsing. All scores normalise to $[0, 1]$; penalty thresholds are heuristic and configurable.

A tolerance threshold $\tau = 0.05$ defines the error rate at which the grammar score reaches zero, accommodating `LanguageTool`'s known false positive rate on scientific text containing domain-specific chemical terminology:

$$\text{error_rate} = \frac{\text{grammar_errors}}{\text{word_count}}, \qquad \text{score}_{\text{grammar}} = \max\left(0,\, 1 - \frac{\text{error_rate}}{\tau}\right)$$

Fragment detection identifies sentences lacking verbal predicates and exceeding five tokens, penalised as:

$$\text{penalty}_{\text{fragment}} = \min\left(\frac{\text{fragment_count}}{\text{sentence_count}} \times 0.5,\ 0.4\right)$$

Lexical validity penalises out-of-vocabulary words and degenerate tokens equally, with chemistry-specific terms exempt from OOV penalties.

Generation completeness flags outputs not ending with valid punctuation as potentially truncated, incurring a heuristic penalty of 0.15. Token utilisation relative to `max_tokens` is recorded as a diagnostic signal, distinguishing concise outputs from those exhausting the generation budget.

Layer 3: DrugBank Lexical Verification. DrugBank lexical verification cross-references NER-tagged terms in generated descriptions against reference terms extracted from DrugBank text fields, including drug classifications, mechanisms of action, pharmacodynamics, and therapeutic categories. `SciSpacy` identifies chemical, drug, and medical entities in both the generated text and DrugBank fields; matched terms are determined by set intersection. Terms are classified as: (1) verified (NER-tagged and present in DrugBank), (2) unmatched (NER-tagged but absent, flagged as potential hallucinations), or (3) untagged (general vocabulary not recognised as chemical entities). The verification rate is:

$$\text{verification_rate} = \frac{|\text{matched_terms}|}{|\text{NER_tagged_terms}|} \times 100$$

4 Results

Results are presented via an interactive dashboard and organised across three comparison strategies: benchmark validation, intra-model consistency, and cross-model architectural analysis.

Fig. 1. Dashboard overview: radar plot (left) and dataset summary (right).

Multi-Dimensional Validation Reveals Complementary Errors. Across six metrics (BLEU, ROUGE, METEOR, BERTScore F1, semantic similarity via MPNet, and chemical similarity via ChemBERTa), all three LLMs exhibit broadly similar aggregate profiles (Fig. 1), though MolT5 outperforms decoder-only models on embedding-based measures. Linguistic metrics show weaker correlations with embedding measures (0.24–0.64 across configurations). This apparent parity masks critical differences in chemical validity: none of these metrics provide information about grammatical correctness, lexical validity, or factual accuracy, requiring the dedicated validation layers described in Section 3.

All three models generated descriptions for acetic acid (DB01060), and achieved grammar scores above 0.78 yet qualitatively distinct failure modes are revealed (Fig. 2):

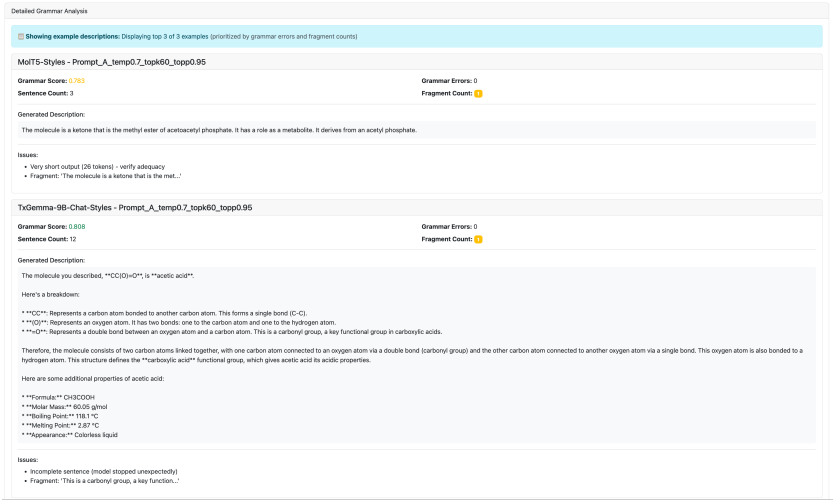

Fig. 2. MolT5 and TxGemma grammar analysis for acetic acid across aggregated configurations.

- **MolT5**: "The molecule is a ketone that is the methyl ester of acetoacetyl..." (score: 0.783) - grammatically acceptable but chemically inaccurate.
- **LLaMA**: "A simple molecule!" yet identified the compound as *formaldehyde* (score: 1.000) - perfect grammar, completely incorrect chemistry.
- **TxGemma**: Correctly identifies acetic acid with extensive structural detail despite the lowest grammar score (0.808).

LLaMA demonstrates seemingly perfect grammar whilst misidentifying the molecule entirely; TxGemma scores lower yet provides correct identification. Scores are heuristic and best interpreted as relative diagnostic signals rather than absolute quality measures, but the contrast demonstrates that linguistic fluency and chemical accuracy must be evaluated independently.

Context-Dependent Error Detection Through Lexical Validation. Despite consistent aggregate performance (mean lexical score: 0.833, OOV: 23.7%),

substantial configuration-dependent variation emerges (Fig. 3). Lexical scores span approximately 0.4–1.0 and OOV rates vary considerably across configurations, demonstrating that decoding strategy and prompt formulation systematically influence vocabulary selection in ways invisible to aggregate metrics. The dashboard enables per-configuration inspection to identify which hyperparameter combinations drive terminological inconsistency.

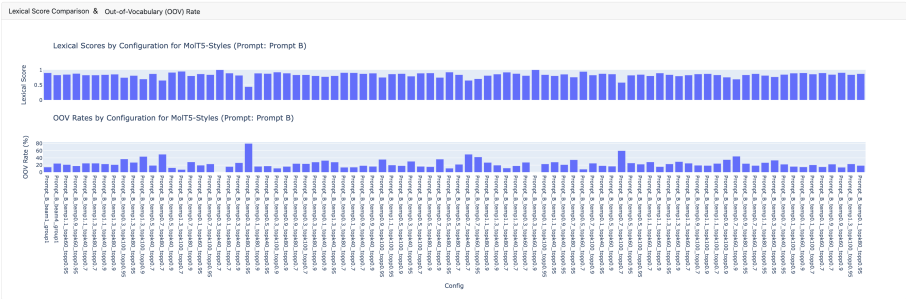

Fig. 3. Lexical validity and OOV rate distributions for MolT5 across configurations.

Configuration Effects on Linguistic Similarity. BERTScore F1 across nine prompts reveal systematic architectural differences under sampling (Fig. 4).

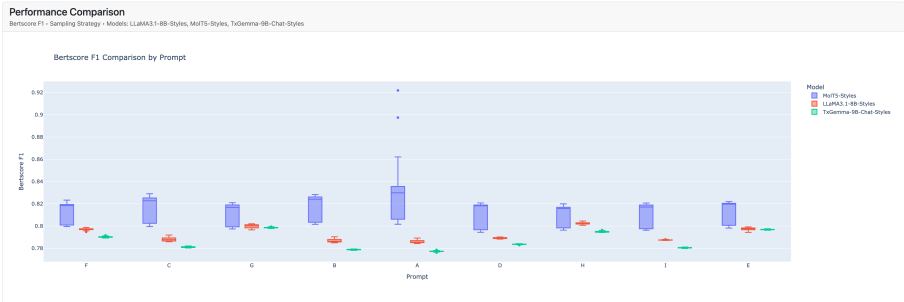

Fig. 4. BERTScore F1 comparison across prompt styles for sampling strategies.

MolT5 exhibits substantial variance under Prompt A (IQR: 0.80-0.86) relative to all other prompts (IQR: 0.01-0.02), likely reflecting baseline bias: the MolT5 beam-1/group-1 baseline was itself generated under Prompt A, inflating scores for that configuration and artificially compressing variance elsewhere. Decoder-only architectures (LLaMA and TxGemma) produce consistently compressed distributions across all prompts, consistent with greater architectural distance from the MolT5 reference.

Task-Specific Validation: LogP Prediction Performance. LogP values were extracted from 37,884 generated text outputs via regex and compared against RDKit-calculated ground truth (Fig. 5). TxGemma produces the highest coverage (35,687 outputs containing a LogP value, 94.2%) though with substantial

error (MAE: 2.564, RMSE: 3.429), reflecting a systematic tendency to generate plausible-sounding but imprecise property values. LLaMA achieves comparable coverage (21,836 outputs, 57.6%, MAE: 2.675, RMSE: 3.592) without chemistry-specific training. Although MolT5 was trained for molecule-text translation, it failed to mention LogP values in 99.9% of prompted descriptions (36 outputs, 0.1%, MAE: 4.166, RMSE: 5.552). This counterintuitive result demonstrates that domain-specific pre-training on molecular structure does not necessarily transfer to property prediction tasks.

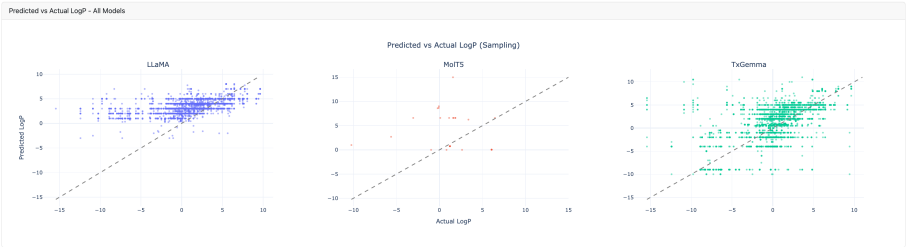

Fig. 5. Predicted versus actual LogP values for all three LLMs (sampling strategy).

5 Discussion and Conclusion

Single-metric frameworks miss critical failures: LLaMA achieved perfect grammar (1.000) whilst misidentifying acetic acid as formaldehyde, an error invisible to linguistic evaluation, whilst TxGemma's accurate identification received lower grammar scores. Trustworthy chemistry AI requires validation across orthogonal dimensions - linguistic fluency, lexical appropriateness, structural validity, and factual accuracy - none of which is redundant with the others.

LogP prediction challenges domain-specific pre-training assumptions, as demonstrated by MolT5 poor performance, and outperformed by the general-purpose LLaMA and therapeutic TxGemma models. This demonstrates that structural pre-training does not transfer to property prediction tasks, and that model selection should be task-specific rather than architecture-driven.

Mol-Hallu and HalluMat have advanced entity-level hallucination detection, but address only one dimension among several. Our contribution is architectural, offering an alternative perspective. By systematically combining established validation dimensions into a multilayer diagnostic pipeline, our framework prioritises transparency and auditability, revealing where and how LLMs fail in high-stakes chemical applications. The dashboard shifts LLM selection from aggregate benchmark performance to per-configuration correctness profiling, enabling deployment decisions grounded in application-specific requirements.

This proof-of-concept has limitations. Layer 2 penalty thresholds are heuristic and require empirical calibration against annotated chemical descriptions to replace current manually-set defaults. Upstream errors in `SciSpacy` NER and regex claim extraction propagate into Layers 1 and 3, and the DrugBank vocabulary may penalise valid but uncommon terminology. Implementing these improvements will enable more rigorous evaluation of LLMs in high-stakes chemical applications.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Banerjee, S., et al.: METEOR: An automatic metric for mt evaluation. In: Proc. ACL Workshop on Evaluation Measures for MT. pp. 65–72 (2005)
2. Cossio, M.: A comprehensive taxonomy of hallucinations in large language models (Aug 2025). https://doi.org/10.48550/arXiv.2508.01781
3. DiMasi, J.A., et al.: Innovation in the pharmaceutical industry: New estimates of R&D costs. Journal of Health Economics **47**, 20–33 (May 2016). https://doi.org/10.1016/j.jhealeco.2016.01.012
4. Edwards, C., et al.: Translation between molecules and natural language (Nov 2022). https://doi.org/10.48550/arXiv.2204.11817
5. Es, S., et al.: RAGAs: Automated evaluation of retrieval augmented generation. In: Proc. 18th Conf. European Chapter ACL: System Demonstrations. pp. 150–158 (2024)
6. Grattafiori, A., et al.: The Llama 3 herd of models (Nov 2024). https://doi.org/10.48550/arXiv.2407.21783
7. Landrum, G.: RDKit: Open-source cheminformatics software. http://www.rdkit.org (2006)
8. Li, H., et al.: How to detect and defeat molecular mirage: A metric-driven benchmark for hallucination in LLM-based molecular comprehension (Apr 2025). https://doi.org/10.48550/arXiv.2504.12314
9. Lin, C.: ROUGE: A package for automatic evaluation of summaries. In: Proc. ACL Workshop on Text Summarization. pp. 74–81 (2004)
10. Manakul, P., et al.: SelfCheckGPT: Zero-Resource black-box hallucination detection for generative large language models (Oct 2023). https://doi.org/10.48550/arXiv.2303.08896
11. Mswahili, M., et al.: Transformer-based models for chemical SMILES representation: A comprehensive literature review. Heliyon **10**(20), e39038 (2024)
12. Neumann, M., et al.: ScispaCy: Fast and robust models for biomedical NLP. In: Proc. 18th BioNLP Workshop and Shared Task. pp. 319–327 (2019)
13. Papineni, K., et al.: Bleu: A method for automatic evaluation of machine translation. In: Proc. 40th Annual Meeting ACL. pp. 311–318 (2002)
14. Sujan, M., et al.: Validation framework for the use of AI in healthcare: overview of the new British standard BS30440. BMJ Health Care Informatics **30**(1) (Jun 2023). https://doi.org/10.1136/bmjhci-2023-100749
15. Vangala, B.P., et al.: HalluMat: Detecting hallucinations in LLM-Generated materials science content through multi-stage verification (Dec 2025). https://doi.org/10.48550/arXiv.2512.22396
16. Vasey, B., et al.: Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine **28**(5), 924–933 (May 2022). https://doi.org/10.1038/s41591-022-01772-9
17. Wang, E., et al.: TxGemma: Efficient and agentic LLMs for therapeutics (Apr 2025). https://doi.org/10.48550/arXiv.2504.06196
18. Wishart, D.S., et al.: DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Research **46**(D1), D1074–D1082 (Jan 2018). https://doi.org/10.1093/nar/gkx1037
19. Zhang, T., et al.: BERTScore: Evaluating text generation with BERT (Feb 2020). https://doi.org/10.48550/arXiv.1904.09675