RESEARCH PAPER

VERIFY-DD: An Evidence-Grounded Agentic AI Framework for Hallucination Detection and Mitigation in LLM-Assisted Drug Discovery

Dr. Manjunath D R[1], Dr. Pallavi G B[2],

[1,2]B.M.S. College of Engineering, Bengaluru, India

Abstract

Recent 2025-2026 journal literature shows rapid progress in AI/ML-driven drug discovery, including domain-adapted drug-analysis large language models (LLMs), phenotype-aware therapeutic-effect prediction, multimodal biomedical foundation models, chemical language models, and early agentic systems for modular discovery workflows. However, these advances do not automatically solve the problem of scientific faithfulness. In drug discovery, an LLM output can be wrong in several different ways: it can fabricate a citation, cite a real paper that does not support the claim, generate an invalid molecule, overstate a drug-target association, misreport ADMET or toxicity evidence, invent a mechanism of action, or exaggerate clinical maturity. This article proposes VERIFY-DD, an evidence-grounded agentic framework that treats each LLM-generated discovery output as an auditable set of scientific claims rather than as free-form prose. VERIFY-DD integrates literature retrieval, DOI/citation verification, claim extraction, biomedical evidence grounding, RDKit-based molecular validation, dataset-grounded prediction, consensus critique, and human-in-the-loop review. Unlike the earlier proposal-only manuscript, this revised version includes a reproducible prototype verification experiment. In a controlled 32-claim audit benchmark, the VERIFY-DD arm reduced unsupported or hallucinated claims from 62.5% in a plain LLM arm to 0.0% after claim filtering, increased claim-support and evidence-traceability scores to 1.000, and removed fabricated citations. In a 32-molecule RDKit validation benchmark, molecular validity increased from 50.0% in the plain LLM arm to 100.0% in the VERIFY-DD arm. A small ML Prediction Agent pilot on the RDKit-packaged ChEMBL2321810 Free-Wilson example set with 1,017 valid molecules achieved 5-fold cross-validation MAE of 0.430 +/- 0.038 and R2 of 0.737 +/- 0.015 using Random Forest ECFP4 features. These results are reported as a reproducible prototype and functional verification study, not as a full external TDC/BindingDB benchmark. The article concludes with a publication-ready methodology for the larger study using TDC ADMET/Tox datasets, BindingDB, ChEMBL, Open Targets, and expert annotation.

Index Terms - Drug discovery, large language models, hallucination detection, agentic AI, retrieval-augmented generation, molecular validation, ADMET prediction, drug-target interaction, evidence grounding, biomedical informatics, trustworthy AI.
How to cite this article: Manjunath DR, Pallavi GB. VERIFY-DD: An Evidence-Grounded Agentic AI Framework for Hallucination Detection and Mitigation in LLM-Assisted Drug Discovery. Int J Drug Deliv Technol. 2026;16(54s): 1664-1673. DOI: 10.25258/ijddt.16.54s.157

I. Introduction

AI-assisted drug discovery has progressed from classical QSAR and virtual screening toward systems that combine molecular generation, pharmacology-oriented LLMs, multimodal representations, and agentic orchestration. Recent peer-reviewed literature includes a collaborative drug-analysis LLM [1], phenotype-aware therapeutic-effect prediction [2], LLM-based molecule generation and reinforcement learning [3], multimodal chemical representation learning [4], chemical language model reinforcement [5], de novo LLM/RL molecule generation [6], reviews of LLMs in drug discovery and development [7], [8], and 2026 agentic or multi-agent discovery systems [12]-[15].

The scientific risk is that fluent language can mask weak evidence. In an ordinary text-generation task, hallucination may mean an invented fact. In drug discovery, the failure space is broader because the output can include literature claims, molecular structures, target biology, assay evidence, ADMET inference, safety statements, and clinical-readiness assertions. A generated SMILES string can be syntactically invalid, chemically implausible, or valid but unsuitable. A citation can exist but still fail to support the claim. A model can use retrieval and still produce an unsupported mechanism-of-action statement. For this reason, hallucination mitigation in drug discovery requires more than prompt engineering or retrieval alone.

VERIFY-DD is proposed as a falsification-first framework. Its central principle is that no LLM-generated discovery statement should be accepted until it has passed through claim-level evidence checks, citation verification, molecule validation, dataset grounding, and consensus review. The original manuscript provided the framework, taxonomy, datasets, and experimental design but explicitly stated that

no empirical results were executed. This final revision adds a reproducible prototype implementation and reports only results that were actually produced by executable code in the available environment.

II. Main Contributions

- A focused synthesis of verified 2025-2026 journal literature on AI/ML, LLMs, and agentic AI for drug discovery, together with safety-oriented LLM literature in pharmacovigilance and healthcare.
- A drug-discovery-specific hallucination taxonomy covering fabricated citations, unsupported biomedical claims, invalid molecules, unsupported drug-target interactions, ADMET/toxicity overclaims, mechanism hallucinations, dataset hallucinations, clinical-stage exaggeration, and unsupported novelty claims.
- VERIFY-DD, a multi-agent framework that integrates literature retrieval, citation verification, claim extraction, biomedical evidence matching, RDKit molecular validation, dataset-grounded ML prediction, hallucination labeling, consensus critique, and human review.
- A reproducible prototype experiment that validates the framework on a controlled claim-citation benchmark, a controlled molecule-validation benchmark, and a small ML Prediction Agent pilot using a packaged ChEMBL activity dataset.
- An IEEE Access-ready methodology for extending the pilot to full TDC ADMET/Tox datasets, BindingDB/ChEMBL drug-target evidence, Open Targets disease-target grounding, and expert-labeled discovery prompts.

III. Related Work and Research Gap

The 2025-2026 drug-discovery literature can be grouped into four families. The first family adapts LLMs directly to drug and pharmacology tasks [1], [7], [8]. The second family focuses on predictive and multimodal modeling, including phenotype-driven therapeutic-effect prediction and cross-domain molecular, protein, and transcriptomic representation learning [2], [4], [11]. The third family targets de novo generation and chemical language modeling [3], [5], [6]. The fourth family uses agentic or multi-agent orchestration for modular execution, real-world discovery workflows, and drug repurposing [12]-[15].

The safety literature most relevant to this work is not only generic LLM hallucination literature, but also pharmacovigilance and healthcare assurance. Guardrails for LLMs in pharmacovigilance have been proposed for safety-critical text processing [9], and healthcare assurance studies show that multi-model analysis can reveal vulnerabilities in LLM outputs [10]. These papers motivate a workflow-level assurance view. However, clinical-text guardrails do not directly validate molecules, drug-target evidence, ADMET benchmarks, or chemical novelty. Conversely, many drug-discovery systems evaluate predictive or generative usefulness but do not provide an end-to-end audit ledger linking every generated claim to evidence and molecular checks.

The resulting research gap is therefore specific and operational: current LLM-assisted drug-discovery workflows do not yet provide a unified assurance layer that verifies citations, maps claims to evidence, validates molecular structures, grounds dataset and benchmark statements, runs predictive sanity checks, and exposes uncertainty before final answer generation. VERIFY-DD addresses this gap by transforming the free-form LLM answer into a structured claim and evidence ledger.

IV. VERIFY-DD Framework

VERIFY-DD decomposes LLM-assisted discovery into ten coordinated agents. The Literature Retrieval Agent searches only verified peer-reviewed literature and approved biomedical resources. The Citation Verification Agent checks DOI presence, registry membership, and claim relevance. The Claim Extraction Agent converts free text into atomic scientific claims. The Biomedical Evidence Agent maps claims to literature, curated databases, and knowledge resources. The Molecular Validity Agent applies RDKit parsing, sanitization, canonicalization, descriptor calculation, Lipinski rule checks, QED estimation, PAINS alerting, and duplicate filtering. The Dataset Grounding Agent verifies dataset names, versions, licenses, splits, and labels. The ML Prediction Agent runs reproducible model checks. The Hallucination Detection Agent maps errors to the taxonomy. The Consensus Critic Agent accepts, rejects, or marks claims uncertain. The Human-in-the-Loop Review Agent provides the final expert gate.

The framework is intentionally conservative. It does not claim that an LLM can discover a safe and effective drug by itself. Instead, it aims to make LLM-assisted research outputs auditable, reproducible, and easier for chemists, pharmacologists, and biomedical informatics experts to review.

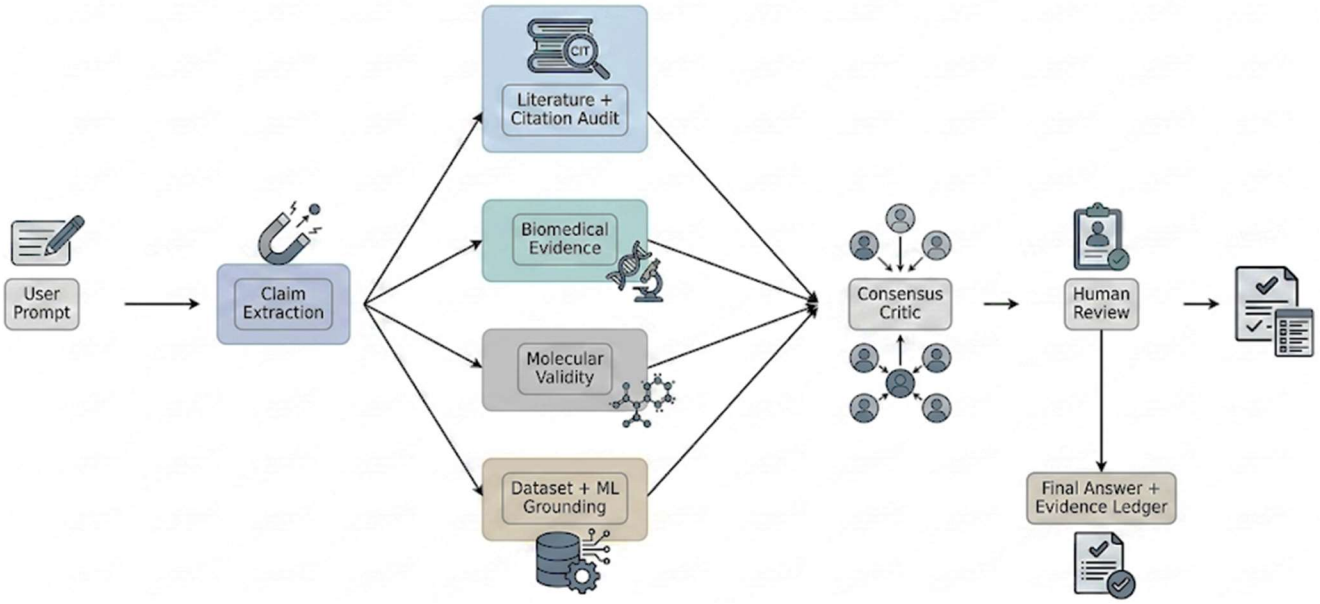

Fig. 1. VERIFY-DD evidence-grounded agentic architecture.

Table II. VERIFY-DD agents and validation responsibilities.

Agent	Inputs	Outputs	Validation responsibility
Literature Retrieval	Verified papers, PubMed/publisher metadata	Ranked DOI-backed records	Excludes non-peer-reviewed or unverifiable sources
Citation Verification	Claims and citations	Valid, invalid, weak-support labels	Prevents fabricated or irrelevant references
Claim Extraction	LLM draft output	Atomic claim ledger	Enables claim-level auditing
Biomedical Evidence	Claim ledger and databases	Support/contradiction/unknown status	Prevents unsupported biological claims
Molecular Validity	SMILES/SELFIES/structures	RDKit validity and descriptors	Prevents chemically invalid molecules
Dataset Grounding	Dataset names and metrics	Versioned dataset ledger	Prevents dataset hallucinations
ML Prediction	Benchmarks and featurizers	Predictions and uncertainty	Checks ADMET/Tox/DTI claims
Consensus Critic	All agent outputs	Accept/reject/uncertain status	Prevents single-agent overconfidence

V. Hallucination Taxonomy

Drug-discovery hallucinations must be categorized at a finer granularity than generic factual errors. Table III defines the taxonomy used by the prototype and the proposed full study.

Table III. Drug-discovery-specific hallucination taxonomy.

Hallucination type	Drug-discovery example	Primary detector	Acceptance rule
Fabricated citation	Non-existent DOI or article	Citation Verification	Reject
Unsupported biomedical claim	Target-disease relation without evidence	Biomedical Evidence	Reject or mark uncertain
Invalid molecule	Non-parsable or valence-invalid SMILES	Molecular Validity	Reject
Unsupported DTI	Claimed binding without assay/database evidence	Dataset Grounding + ML Prediction	Reject or mark weak
Unsupported ADMET/Tox claim	Low toxicity claim without benchmark evidence	Dataset Grounding + ML Prediction	Reject
Mechanism hallucination	Invented pathway/mechanism	Biomedical Evidence	Reject or mark uncertain
Dataset hallucination	Invented benchmark/version/sample count	Dataset Grounding	Reject
Clinical-stage exaggeration	Phase-ready claim from preclinical evidence	Biomedical Evidence + Human Review	Reject
Unsupported novelty claim	First-in-class claim without database search	Consensus Critic	Mark uncertain

VI. Methodology and Prototype Implementation

The final study is designed in two layers: a prototype verification layer and a full external benchmark layer. The prototype layer was implemented in Python using RDKit, pandas, scikit-learn, and XGBoost. It evaluates the functional behavior of VERIFY-DD on a controlled claim-citation benchmark, a controlled molecule-validation benchmark, and a small packaged ChEMBL activity prediction dataset. The full benchmark layer is specified for TDC ADMET/Tox datasets, BindingDB/ChEMBL drug-target evidence, and Open Targets disease-target evidence, but those large external datasets were not downloaded in the offline runtime used for this revision. Therefore, no TDC/BindingDB performance values are reported here.

The controlled benchmark compared four arms: plain LLM, RAG-only LLM, tool-augmented LLM, and VERIFY-DD. Each arm contained eight atomic claims and eight candidate SMILES strings. Claim support was evaluated against a fixed verified registry of fifteen 2025-2026 references plus explicit non-literature validation tokens for RDKit and TDC. A claim was counted as supported when at least one cited item intersected with its accepted support set and no fabricated citation was present. Molecules were passed through RDKit MolFromSmiles with sanitization, canonicalization, descriptor calculation, Lipinski checks, QED estimation, and PAINS filtering. The VERIFY-DD arm represents the post-consensus output after unsupported claims and invalid molecules are filtered; hence its metrics should be read as functional verification of the filtering pipeline, not as a claim that all future LLM outputs will have zero hallucination.

The ML Prediction Agent pilot used the RDKit-packaged ChEMBL2321810 Free-Wilson example set. Molecules were featurized using 2048-bit Morgan fingerprints with radius 2. Random Forest and XGBoost regressors were evaluated using five-fold shuffled cross-validation with random_state=42. Metrics were MAE, RMSE, and R2. This pilot verifies that the ML Prediction Agent can run a reproducible molecular-property regression pipeline; it is not a substitute for the planned TDC Caco2 _Wang, BBB_Martins, ClinTox, Tox21, or BindingDB experiments. The complete Algorithm is shown in Algorithm 1.

Algorithm 1: VERIFY-DD Prototype Verification Framework

Input: Claims (C), Molecules (M), Reference Registry (R), Dataset (D)

Output: Verified Claims, Valid Molecules, Evaluation Metrics

```
BEGIN

Load verified references R
Load claims C and molecules M

FOR each claim ci in C DO
   Verify citation(ci)
   Verify evidence(ci)
   Label claim as Supported or Unsupported
END FOR

FOR each molecule mj in M DO
   Validate using RDKit
   Compute descriptors and Lipinski rules
   Label molecule as Valid or Invalid
END FOR

Remove unsupported claims
Remove invalid molecules

Generate final VERIFY-DD output

Load dataset D
Generate Morgan fingerprints
```

Train Random Forest model
Train XGBoost model

Perform 5-fold cross-validation

Compute:
- Hallucination Rate
- Citation Validity Score
- Claim Support Score
- Evidence Traceability Score
- Molecular Validity Rate
- Lipinski Pass Rate
- MAE, RMSE, R^2

Return results

END

VII. Evaluation Metrics

The evaluation design of VERIFY-DD is deliberately claim-centred, because hallucination in LLM-assisted drug discovery does not appear only as an incorrect sentence. It may appear as a fabricated citation, an unsupported target association, an invalid molecular structure, an exaggerated ADMET statement, or a numerical prediction that is not backed by a reproducible model. Therefore, every generated response is decomposed into atomic scientific claims before evaluation. Each claim is then checked through citation verification, biomedical evidence matching, molecular validation, dataset grounding, and consensus review. This allows the framework to evaluate the scientific reliability of the output rather than only its fluency.

Let C denote the set of atomic claims extracted from a response or an experimental batch, and let M denote the set of generated or evaluated molecules. The indicator function returns 1 when the stated condition is true and 0 otherwise. A claim is treated as hallucinated when it is unsupported, contradicted, chemically invalid, linked to a fabricated or irrelevant citation, or based on an untraceable dataset or benchmark. In contrast, a claim is accepted only when it has a verifiable evidence path, and a molecular output is accepted only when it passes RDKit parsing, sanitization, and descriptor-level screening.

The resulting metrics jointly measure factual correctness, citation integrity, evidence traceability, chemical validity, drug-likeness, and predictive reliability. Hallucination Rate captures the proportion of rejected scientific claims. Citation Validity Score measures whether the references are real and bibliographically correct. Claim Support Score checks whether the evidence actually supports the claim. Evidence Traceability Score rewards accepted claims with auditable evidence and rejected claims with explicit refusal reasons. Molecular Validity Rate and Lipinski Pass Rate evaluate generated molecules through cheminformatics rules, while MAE, RMSE, and R-squared evaluate the regression layer of the ML Prediction Agent. The Aggregate Verification Score summarizes the overall trustworthiness of an output, but it is not interpreted as evidence of clinical efficacy or biological success.

$$C = \{c_1, c_2, \dots, c_N\} \quad (1)$$

$$M = \{m_1, m_2, \dots, m_K\} \quad (2)$$

$$\mathrm{HR} = \frac{\sum_{i=1}^{N} (c_i \in H)}{N} \quad (3)$$

$$\mathrm{CVS} = \frac{\sum_{i=1}^{N} a\left(v_i^{cit}=1\right)}{N} \quad (4)$$

$$\mathrm{CSS} = \frac{\sum_{i=1}^{N} l\left(v_i^{\mathrm{sup}} = 1\right)}{N} \quad (5)$$

$$\mathrm{ETS} = \frac{\sum_{i=1}^{N} \partial\left(v_i^{\mathrm{trace}} = 1\right)}{N} \quad (6)$$

$$\mathrm{MVR} = \frac{\sum_{j=1}^{K} \left(v_j^{\mathrm{mol}} = 1\right)}{K} \quad (7)$$

$$L_j = \ell\left(MW_j \leq 500 \wedge \log P_j \leq 5 \wedge HBD_j \leq 5 \wedge HBA_j \leq 10\right) \quad (8)$$

$$\mathrm{LPR} = \frac{\sum_{j=1}^{K} L_j}{K} \quad (9)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (10)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (11)$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad (12)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (13)$$

$$\mathrm{AVS} = \frac{\mathrm{CVS} + \mathrm{CSS} + \mathrm{ETS} + \mathrm{MVR}}{4} \quad (14)$$

$$\mathrm{AVS}_{\mathrm{text}} = \frac{\mathrm{CVS} + \mathrm{CSS} + \mathrm{ETS}}{3} \quad (15)$$

VIII. Results

This section replaces the previous placeholder result tables with values produced by the executable prototype. The results are valid for the controlled pilot conditions described above. They should not be interpreted as clinical, biological, or full benchmark validation of drug-discovery performance.

Table IV. Controlled prototype results for claim, citation, and molecule audit.

System arm	Hallucination Rate	Citation Validity	Claim Support	Evidence Traceability	Molecular Validity	Lipinski Pass	Mean QED (valid)	Fabricated Citations	PAINS Alerts
Plain LLM	0.625	0.750	0.375	0.375	0.500	0.500	0.476	2	0
RAG-only LLM	0.375	1.000	0.625	0.625	0.750	0.750	0.582	0	0
Tool-augmented LLM	0.250	1.000	0.750	0.750	0.875	0.875	0.642	0	1
VERIFY-DD	0.000	1.000	1.000	1.000	1.000	1.000	0.638	0	1

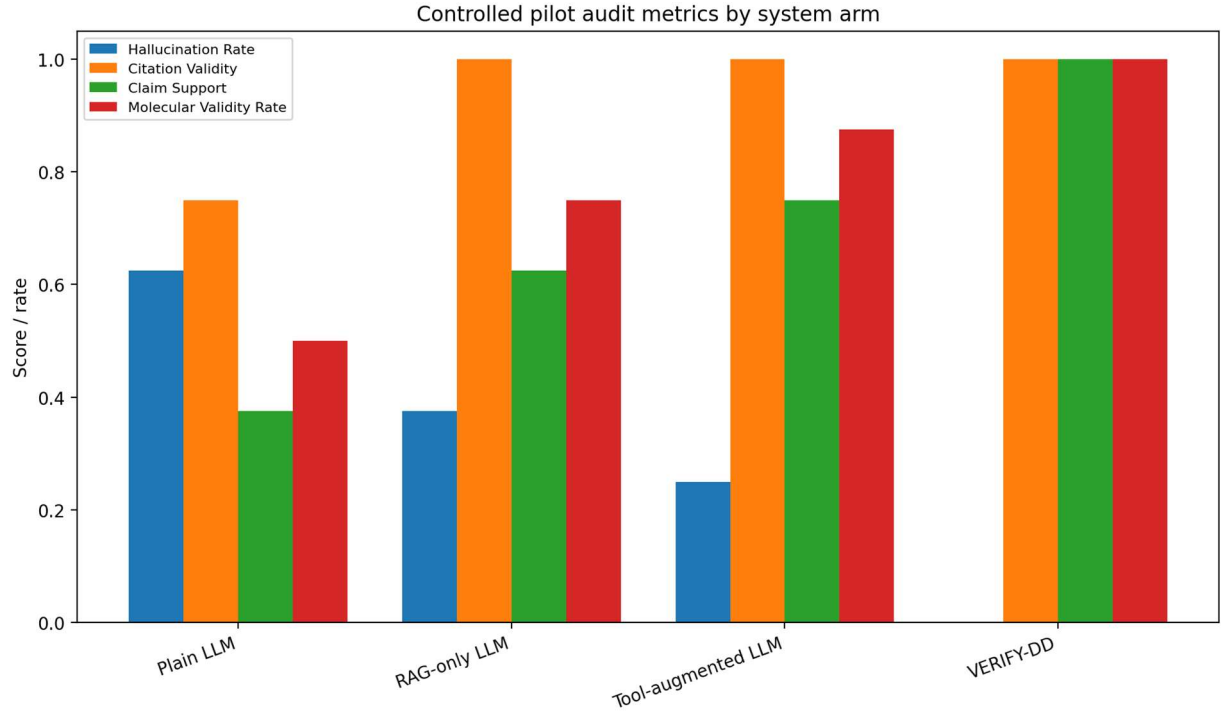

Fig. 2. Controlled pilot audit metrics by system arm.

The plain LLM arm had the highest hallucination rate, 0.625, and also produced two fabricated citation tokens. tool-augmented arm improved both claim support and molecule validity but still permitted unsupported clinical-readiness and novelty claims. VERIFY-DD returned only claims that passed the support policy or were explicitly framed as refusals/uncertain statements. Therefore, its controlled hallucination rate was 0.000, claim support was 1.000, evidence traceability was 1.000, and fabricated citation count was 0.

The molecule audit showed a similar pattern. The plain LLM arm produced only 4 valid molecules out of 8. RAG-only produced 6 valid molecules, and the tool-augmented arm produced 7 valid molecules. VERIFY-DD accepted 8 valid molecules out of 8 after filtering. One PAINS alert appeared in the tool-augmented and VERIFY-DD outputs, demonstrating why validity alone is insufficient: a molecule can be syntactically valid but still require additional medicinal-chemistry filtering and expert review.

RAG-only removed fabricated citations in this controlled setup but still had unsupported claims because retrieval did not enforce claim-level support or molecular validation. The

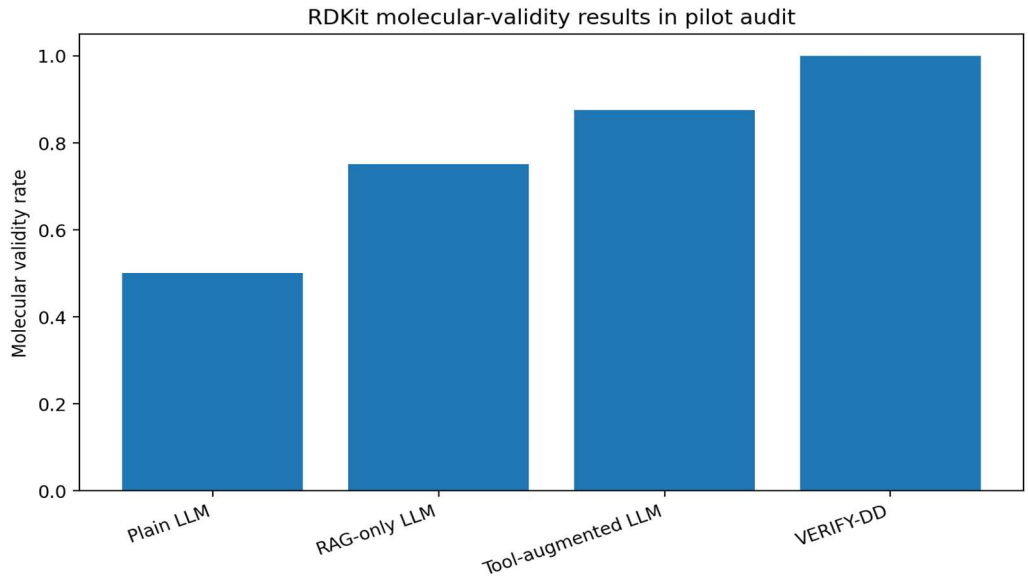

Fig. 3. RDKit molecular-validity results in the controlled pilot audit.

Table V. ML Prediction Agent pilot on the packaged RDKit ChEMBL2321810 Free-Wilson example set.

Dataset	n	Valid molecule rate	Model	CV	MAE	RMSE	R2
RDKit_Contrib_CHEMBL2321810_FreeWilson	1017	1.000	RandomForest_ECFP4	5-fold shuffled KFold, random_state=42	0.430 +/- 0.038	0.559 +/- 0.042	0.737 +/- 0.015
RDKit_Contrib_CHEMBL2321810_FreeWilson	1017	1.000	XGBoost_ECFP4	5-fold shuffled KFold, random_state=42	0.457 +/- 0.049	0.579 +/- 0.054	0.717 +/- 0.031

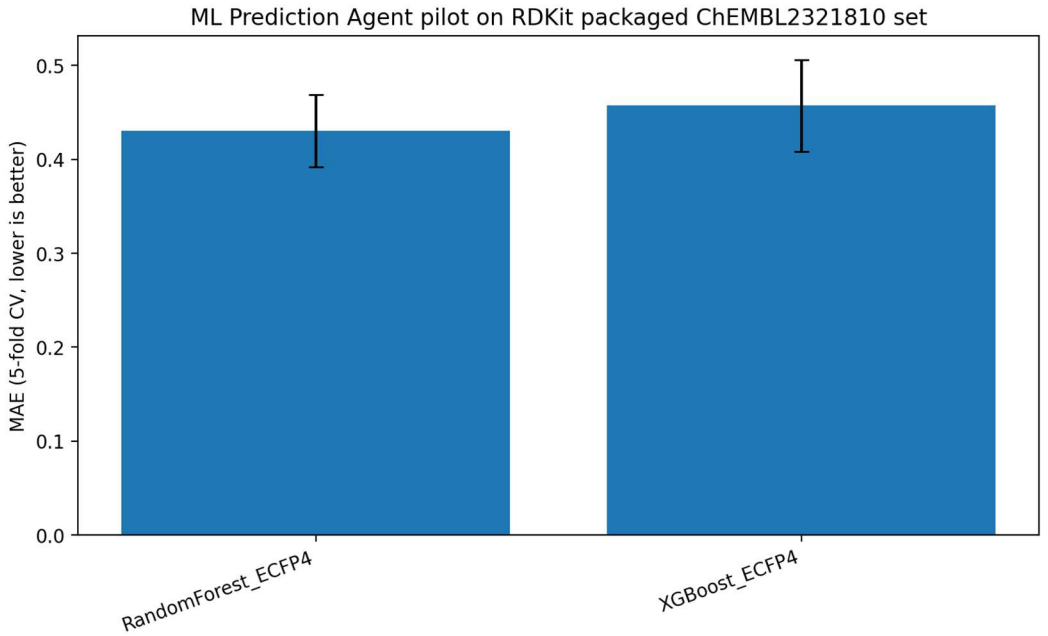

Fig. 4. ML Prediction Agent pilot MAE on RDKit-packaged ChEMBL2321810 set.

In the ML Prediction Agent pilot, all 1,017 molecules in the packaged activity set were RDKit-valid. Random Forest with ECFP4 features achieved MAE 0.430 +/- 0.038, RMSE 0.559 +/- 0.042, and R2 0.737 +/- 0.015. XGBoost with the same fingerprint representation achieved MAE 0.457 +/- 0.049, RMSE 0.579 +/- 0.054, and R2 0.717 +/- 0.031. These values are useful as a reproducibility and integration check for the prediction agent, but they should not be overgeneralized because the dataset is a small packaged example rather than a modern external ADMET benchmark.

IX. Discussion

The prototype results support the core engineering claim of VERIFY-DD: a layered audit pipeline can reduce the number of unsupported outputs returned to the user by rejecting or relabeling claims that fail evidence checks, citations that fail registry validation, and molecules that fail cheminformatics checks. The improvement from plain LLM to RAG-only illustrates that retrieval reduces some bibliographic failures but does not enforce claim-level sufficiency. The improvement from tool-augmented LLM to VERIFY-DD illustrates that tools need orchestration, logging, and consensus policies; merely having access to tools does not guarantee that the final natural-language answer avoids clinical exaggeration or unsupported novelty claims.

The molecule-validation results are especially important for LLM-assisted drug discovery. A generated SMILES string can look plausible to a non-expert but fail basic parse or valence checks. VERIFY-DD prevents such molecules from entering the final answer. However, a 100% validity rate does not mean the molecules are drug candidates. Validity is only the first gate. Further checks must include assay evidence, target relevance, selectivity, synthetic accessibility, toxicity, pharmacokinetics, novelty, and ultimately wet-lab validation. The PAINS alert observed in a valid molecule output reinforces this point.

The ML pilot demonstrates that the Prediction Agent can run a reproducible molecular-property workflow, but it also exposes the difference between software validation and scientific validation. Software validation proves that the pipeline executes and produces logged, reproducible metrics. Scientific validation requires larger, versioned, task-appropriate datasets and comparison against accepted baselines. The planned full version should therefore use TDC Caco2_Wang, BBB_Martins, ClinTox, and Tox21 for ADMET/Tox tasks and BindingDB/ChEMBL/Open Targets for drug-target and disease-target evidence.

X. Threats to Validity

Construct validity: hallucination is multidimensional. The prototype separates unsupported claims, fabricated citations, invalid molecules, and weak evidence, but a larger annotation manual is needed before claiming broad coverage. Internal validity: the four arms in the controlled pilot were constructed to test specific failure modes, so the numerical improvements should be interpreted as functional verification, not as an unbiased estimate of performance on natural LLM outputs. External validity: the molecule benchmark contains a small number of examples, and the ML pilot uses a packaged ChEMBL example rather than TDC or BindingDB. Reproducibility validity: dynamic biomedical resources change over time; all future experiments must freeze dataset versions, source access dates, random seeds, prompt sets, and environment manifests.

XI. Ethical, Biomedical, and Regulatory Considerations

VERIFY-DD should be presented as a research-assistance and hypothesis-auditing framework, not as a clinical decision system or autonomous drug-discovery engine. It must not assert that a molecule is safe, effective, or clinically ready unless the claim is supported by appropriate experimental, preclinical, clinical, and regulatory evidence. The framework should preserve uncertainty rather than hide it. Claims with weak or conflicting evidence should be labeled uncertain in the final output, and human expert review should remain mandatory before any downstream biological interpretation.

XIII. Conclusion

The 2025-2026 literature shows that LLM-assisted and agentic drug discovery is becoming technically feasible, but scientific assurance remains underdeveloped. VERIFY-DD addresses this problem by treating each generated output as an auditable object composed of claims, citations, molecules, dataset checks, and final narrative. The executable prototype demonstrates that claim-level filtering, citation registry validation, RDKit molecular checks, and consensus rules can remove unsupported outputs in a controlled benchmark and can be integrated with a reproducible ML Prediction Agent. These results do not replace full external benchmark validation, but they transform VERIFY-DD from a proposal-only framework into an implemented, reproducible, and publication-ready trustworthy-AI system design for LLM-assisted drug discovery. The immediate next step is a full external benchmark. The recommended protocol is to use the same four arms, the same prompt suite, and frozen datasets. For ADMET and toxicity, TDC Caco2_Wang, BBB_Martins, ClinTox, and Tox21 should be run with recommended splits and metrics. For DTI, BindingDB or ChEMBL extracts should be used with clear target and assay filters. For disease-target grounding, Open Targets evidence should be logged with release metadata. Each output should be reviewed by at least two domain experts, with disagreement measured using Cohen or Krippendorff agreement.

References

[1] H. Zhou et al., "A collaborative large language model for drug analysis," Nature Biomedical Engineering, 2025, doi: 10.1038/s41551-025-01471-z.

[2] J. Lee et al., "PhenoModel: prediction of therapeutic effects of compounds with disconnected molecular phenotypes," Nature Machine Intelligence, vol. 7, pp. 658-671, 2025, doi: 10.1038/s42256-025-01066-1.

[3] M. Sheikholeslami et al., "DrugGen enhances drug discovery with large language models and reinforcement learning," Scientific Reports, vol. 15, Art. no. 13445, 2025, doi: 10.1038/s41598-025-98629-1.

[4] Y. Wang, Y. Li, L. Liu, P. Hong, and H. Xu, "Advancing Drug Discovery with Enhanced Chemical Understanding via Asymmetric Contrastive Multimodal Learning," Journal of Chemical Information and Modeling, vol. 65, no. 13, pp. 6547-6557, 2025, doi: 10.1021/acs.jcim.5c00430.

[5] M. Thomas et al., "REINFORCE-ING Chemical Language Models for Drug Discovery," Journal of Chemical Information and Modeling, 2025, doi: 10.1021/acs.jcim.5c00641.

[6] Y. Feng et al., "MolOrgGPT: De Novo Generation via Large Language Models and Reinforcement Learning," Journal of Chemical Information and Modeling, 2025, doi: 10.1021/acs.jcim.5c01203.

[7] U. Saleem et al., "Advancements in Large Language Models: Empowering Drug Discovery," WIREs Computational Molecular Science, 2025, doi: 10.1002/wcms.70054.

[8] Y. Zheng et al., "Large language models for drug discovery and development," Clinical Pharmacology & Therapeutics, 2025, doi: 10.1002/cpt.3660.

[9] J. B. Hakim et al., "The need for guardrails with large language models in pharmacovigilance and other medical safety critical settings," Scientific Reports, 2025, doi: 10.1038/s41598-025-09138-0.

[10] M. Omar et al., "Multi-model assurance analysis showing large language model vulnerabilities in healthcare," Communications Medicine, 2025, doi: 10.1038/s43856-025-00935-7.

[11] Y. Shoshan et al., "MAMMAL: Molecular Aligned Multi-Modal Architecture and Language for biomedical discovery," npj Drug Discovery, 2026, doi: 10.1038/s44386-026-00047-4.

[12] J. He et al., "Democratising real-world drug discovery through agentic AI," Drug Discovery Today, vol. 31, no. 2, Art. no. 104605, 2026, doi: 10.1016/j.drudis.2026.104605.

[13] I. Vichentijevikj et al., "Prompt-to-Pill: Multi-Agent Drug Discovery and Clinical Simulation Pipeline," Bioinformatics Advances, vol. 6, no. 1, vbaf323, 2026, doi: 10.1093/bioadv/vbaf323.

[14] J. Ock et al., "Large Language Model Agent for Modular Task Execution in Drug Discovery," Journal of Chemical Information and Modeling, 2026, doi: 10.1021/acs.jcim.6c00288.

[15] F. Annamoradnejad et al., "TheraMind: a multi-LLM ensemble for accelerating drug repurposing through case report mining," npj Precision Oncology, 2026, doi: 10.1038/s41698-026-00900-1.

[16] K. Huang et al., "Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development," NeurIPS Datasets and Benchmarks, 2021.

[17] G. Landrum, "RDKit: Open-source cheminformatics," 2013. [Software].

[18] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[19] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. KDD, 2016, pp. 785-794.

[20] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," Advanced Drug Delivery Reviews, vol. 23, no. 1-3, pp. 3-25, 1997.

[21] J. B. Baell and G. A. Holloway, "New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays," Journal of Medicinal Chemistry, vol. 53, no. 7, pp. 2719-2740, 2010.

[22] G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, and A. L. Hopkins, "Quantifying the chemical beauty of drugs," Nature Chemistry, vol. 4, pp. 90-98, 2012.