

Introduction

The ICH M7 guideline recommends an initial in silico assessment of pharmaceutical impurities using two complementary (Q)SAR methodologies to predict bacterial mutagenicity.¹ It also suggests expert review if warranted (i.e., cases where model predictions are discordant, inconclusive, or outside the applicability domain). This study investigates whether a large language model (LLM) can assist in generating expert-level justifications for mutagenicity predictions based on in silico outputs.

Methodology

1

A dataset of 684 chemicals with known experimental results was analyzed using CASE Ultra v1.9.2.6, along with the GT1_BMUT v1.9.2.4 and GT_EXPERT v1.9.2.4 models. Outputs were further analyzed with the Konsolidator tool (v7.0), which uses structural alerts in the query compound to identify structural analogs and generate a final call: POSITIVE, NEGATIVE, or NO CALL.

2

Compounds were subsequently categorized

- █ predicted correctly
- █ predicted incorrectly
- █ no call

3

From this dataset, 75 compounds (25 per category) were randomly selected. Each Konsolidator report was submitted to ChatGPT v4o for interpretation under a standardized prompt:

“Please summarize the results for this chemical according to the ICH M7 guideline in the form of an executive summary. Be sure to indicate whether the chemical is positive or negative for bacterial mutagenicity, the ICH M7 classification, and the rationale used to support the conclusion.”

4

ChatGPT was configured without memory and allowed to browse the web. The same 75 compounds were also independently evaluated by a human expert under conditions of a blind experiment.

5

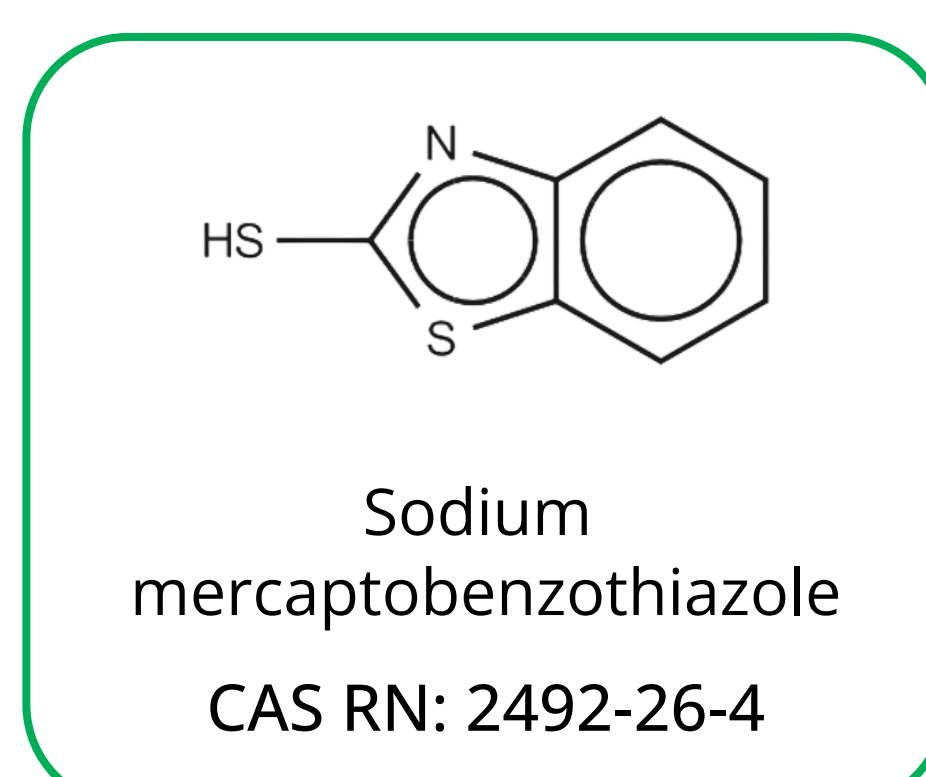
The results produced by ChatGPT and the human expert were compared with experimental outcomes for each compound to evaluate performance.

Results

The human expert achieved the highest prediction accuracy for bacterial mutagenicity (80%) and 100% coverage. ChatGPT’s performance was notable, with 61% accuracy and 91% coverage. ChatGPT also demonstrated a marked ability to correctly identify negative compounds with a specificity of 79%. In contrast, its ability to correctly identify positive compounds was low (sensitivity of 38%).

	Human Expert	ChatGPT Decision
True Negatives (TN)	38	31
False Positives (FP)	7	8
False Negatives (FN)	8	18
True Positives (TP)	22	11
No Calls	0	7
Accuracy	0.80	0.62
Sensitivity	0.73	0.38
Specificity	0.84	0.79
Coverage	1.00	0.91

The LLM’s handling of “NO CALL” cases were particularly interesting. It provided a conclusive interpretation in 21 out of 25 cases (84%), with an accuracy rate of 64%. All outputs referenced structural alerts and analogs, but the quality and depth of analysis varied across compounds. However, the number of “NO CALL” cases is 7 because ChatGPT failed to provide conclusive POSITIVE/NEGATIVE predictions for 3 compounds that were predicted POSITIVE by (Q)SAR.



Sodium mercaptobenzothiazole, found to be out of domain in both statistical and expert rule-based models, was correctly predicted negative by ChatGPT.

ChatGPT also acknowledged the following, demonstrating its ability to apply reasoning that extended beyond the contents of the ICH M7 guideline, in this case considering information contained within the M7(R2) Q&As³ document:

Carcinogenicity:

- Mixed rodent data with both **negative and equivocal/potentially carcinogenic** findings.
- Classified by IARC as **Group 2A: Probably carcinogenic to humans**.
- However, per ICH M7, **carcinogenicity data are not directly used** for mutagenicity-based classification unless mutagenicity is uncertain.

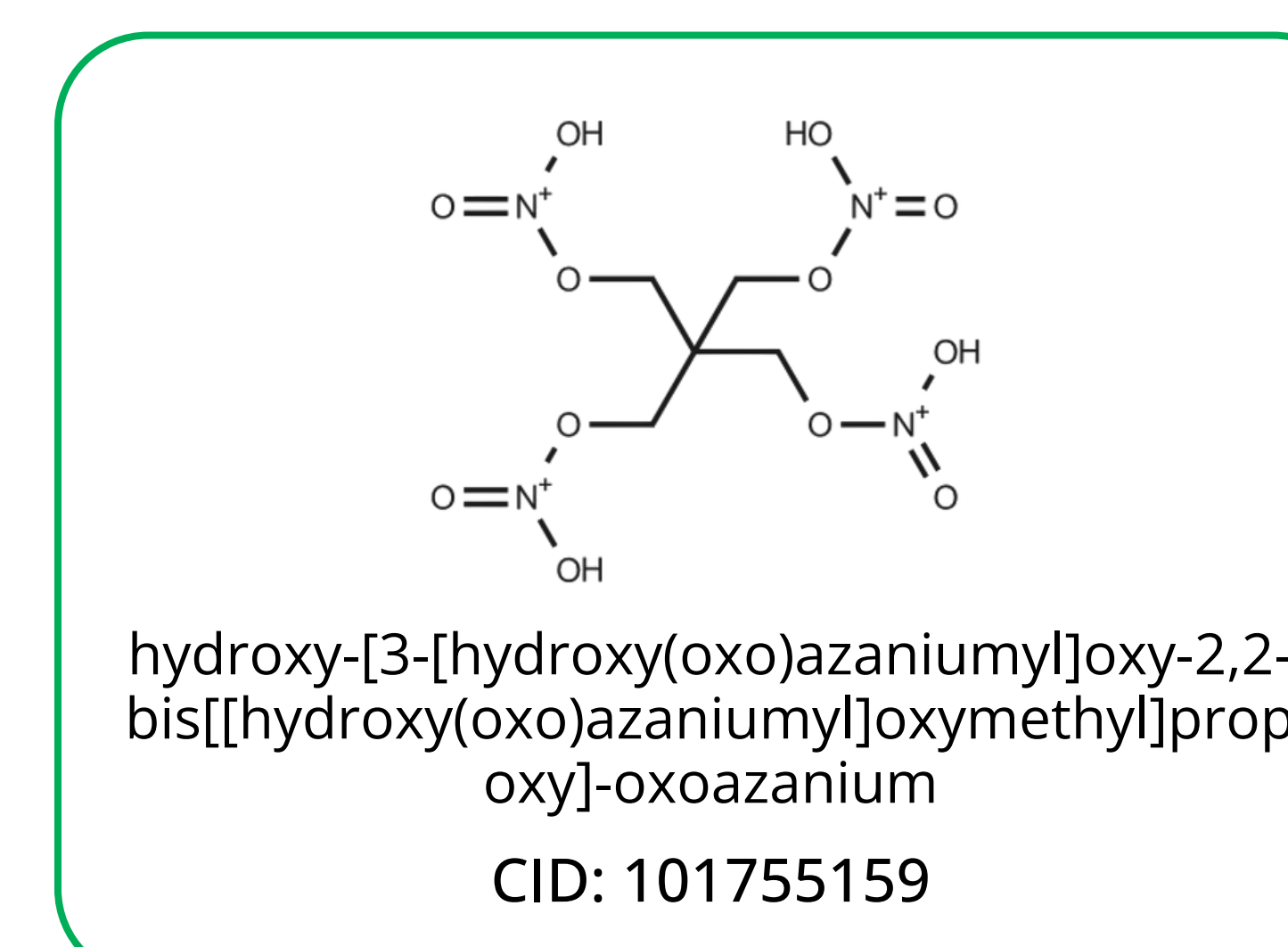
Hydroxy-[3-[hydroxy(oxo)azaniumyl]oxy-2,2-bis[[hydroxy(oxo)azaniumyl]oxymethyl]propoxy]-oxoazanium, negative in the bacterial mutagenicity assay, was found to be out-of-domain in both statistical and expert rule-based models, and zero alerts were observed. ChatGPT included the following in its response:

ICH M7 Classification

- Final ICH M7 Category: Class 3 — Alerting Structure, but Inadequate Data

Recommendation:

- Consider experimental Ames testing to resolve uncertainty.
- Alternatively, quantum chemical calculations or testing of a neutralized form (if ionic) may help reduce uncertainty around the unknown fragments.



AI Hallucinations

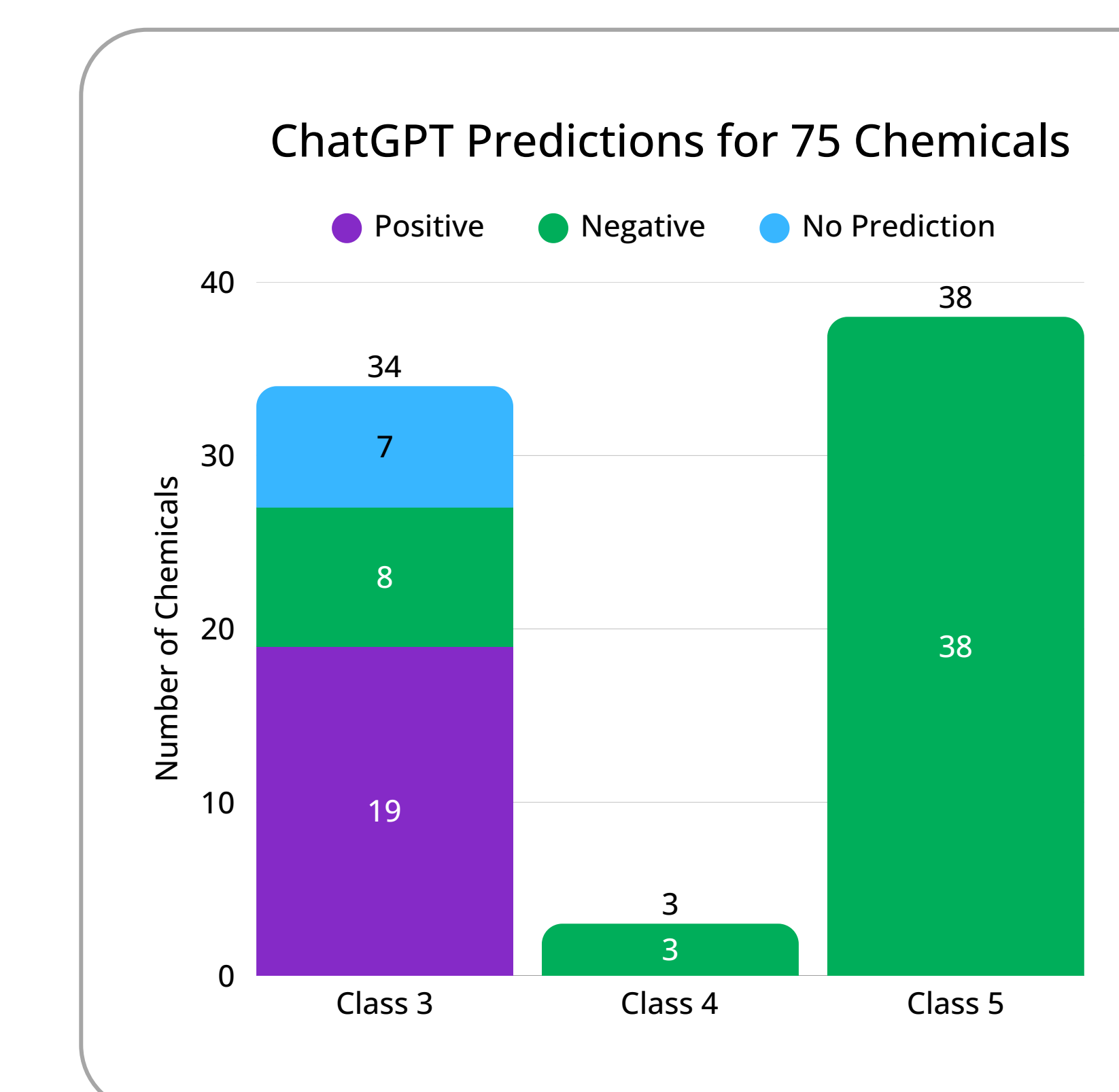
AI hallucinations are plausible but false statements generated by language models, and may persist because early training and evaluation procedures reward guessing over acknowledging uncertainty.²

A recurring issue involved ICH M7 classification. ChatGPT misclassified 8 compounds, labeling them as ICH M7 Class 3 while predicting them as NEGATIVE.

For example, sodium mercaptobenzothiazole was correctly predicted NEGATIVE but assigned Class 3 with rationale using the definition of an ICH M7 Class 5 structure:

“Alerting structure with sufficient data to demonstrate lack of mutagenic potential.”

ChatGPT also assigned 3 compounds as Class 4 using the definition of a Class 5 structure and was not provided with an API or other compound for structural comparison.



ICH M7 Impurity Classifications and Definitions¹

Class	Definition	Proposed action for control (details in Section 7 and 8)
1	Known mutagenic carcinogens	Control at or below compound-specific acceptable limit
2	Known mutagens with unknown carcinogenic potential (bacterial mutagenicity positive*, no rodent carcinogenicity data)	Control at or below acceptable limits (appropriate TTC)
3	Alerting structure, unrelated to the structure of the drug substance; no mutagenicity data	Control at or below acceptable limits (appropriate TTC) or conduct bacterial mutagenicity assay; If non-mutagenic = Class 5; If mutagenic = Class 2
4	Alerting structure, same alert in drug substance or compounds related to the drug substance (e.g., process intermediates) which have been tested and are non-mutagenic	Treat as non-mutagenic impurity
5	No structural alerts, or alerting structure with sufficient data to demonstrate lack of mutagenicity or carcinogenicity	Treat as non-mutagenic impurity

Conclusion

This study highlights the strengths and limitations of generative AI in regulatory toxicology, as well as the role of human experts in this field. While the human expert outperformed the LLM in both accuracy and coverage, the LLM’s ability to process and summarize complex in silico data, particularly in NO CALL cases, was promising.

The study was limited to the use of a single LLM, a fixed prompt, and a manual input process, which may not reflect variability or scalability in real-world applications. Future work should explore automated workflows, prompt engineering, alternative LLMs, and the interpretation of raw (Q)SAR outputs and outputs from other software. Overall, LLMs may serve as valuable tools for generating structured summaries that assist expert workflows, but they should not yet be used as a standalone tool for interpreting in silico results.

References

1. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). M7(R2): Assessment and Control of DNA Reactive (Mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk; Step 4, April 2023.
2. Kalai, A. T.; Nachum, O. Why Language Models Hallucinate. arXiv 2025, arXiv:2509.04664. <https://doi.org/10.48550/arXiv.2509.04664>
3. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. ICH M7(R2) Guideline: Assessment and Control of DNA Reactive (Mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk — Questions and Answers; ICH M7 Implementation Working Group: Geneva, Switzerland, 2022.