MOLDETOX: EVALUATING LANGUAGE MODEL'S STEPWISE FRAGMENT EDITING FOR MOLECULAR DETOXIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used for molecule design and optimization, but systematic evaluation of their capability for toxicity mitigation remains limited. We present MolDeTox, a benchmark for evaluating reference-based molecular detoxification across multiple fragment-level editing tasks. MolDeTox is built on *ToxicityCliff*, a collection of closely related molecular pairs that differ in toxicity for the same endpoint. These examples expose small structural changes associated with reduced toxicity while minimizing unrelated changes to the molecule. We convert these pairs into three QA tasks covering fragment identification, fragment substitution, and final molecule generation. This structure enables step-wise analysis of model behavior across the detoxification process. Our analysis further identifies common failure cases and highlights potential directions for improving performance. We also assess whether the benchmark provides a useful training signal. Fine-tuning a small language model on the training split leads to substantial gains on MolDeTox. The fine-tuned model also improves molecular repair success under an independent oracle-based evaluation setting. Together, these results show that MolDeTox serves not only as an evaluation benchmark, but also as a useful resource for developing and analyzing molecular detoxification systems. We release the evaluation code and dataset.[1]

1 INTRODUCTION

Recent advances in large language models (LLMs) have significantly expanded their capabilities beyond natural language processing, enabling applications in scientific domains such as chemistry (Narayanan et al., 2025; Zhao et al., 2025) and biology (Fallahpour et al., 2025; 2026; Istrate et al., 2025). They have also demonstrated an increasing ability to understand molecular structures and infer physicochemical properties directly from structural representations, making them promising tools for drug discovery.

These advances have extended to molecular optimization, one of the most challenging tasks in drug discovery. Molecular optimization aims to modify molecules toward desired properties while preserving other favorable characteristics, with objectives such as enhancing biological activity, improving selectivity, and reducing toxicity (Yu et al., 2025). Recent studies have shown that LLMs can improve molecular properties, including drug-likeness (QED) and solubility, through prompt-based approaches or instruction tuning (Nguyen & Grover, 2024; Ye et al., 2025). Benchmark datasets have also been introduced to systematically evaluate the molecular optimization capabilities of LLMs (Cai et al., 2025; Li et al., 2025).

However, molecular optimization specifically aimed at toxicity mitigation remains under-explored. Existing studies often consider only a limited set of toxicity endpoints, such as hERG-related cardiotoxicity (Ye et al., 2025). More recently, dedicated benchmarks have begun to emerge. For example, ToxiMol (Lin et al., 2026) evaluates general-purpose multimodal large language models (MLLMs) for molecular toxicity repair across multiple toxicity endpoints. ToxiMol evaluates generated molecules using ToxiEval (Lin et al., 2025), an oracle-based evaluation framework that

[1]Dataset:https://huggingface.co/datasets/MolDeTox/MolDeTox
Code: https://anonymous.4open.science/r/MolDeTox-F2B3

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

combines predicted toxicity with drug-likeness, synthetic accessibility, and structural similarity. While this design enables scalable and open-ended evaluation, ToxiMol is constructed from toxic molecules without paired non-toxic counterparts. As a result, the benchmark evaluates whether a generated molecule satisfies desired endpoint-level criteria, but does not provide a reference structural transformation that can be used to supervise or assess intermediate detoxification steps.

To address this gap, we introduce **MolDeTox**, a large-scale benchmark for molecular detoxification with dedicated training and evaluation splits. At its core, we construct *ToxicityCliff*, a dataset of structurally similar toxic/non-toxic molecular pairs with contrasting toxicity outcomes for the same endpoint. By focusing on highly similar molecular pairs, *ToxicityCliff* captures localized structural changes associated with toxicity mitigation while largely preserving the remaining molecular structure and physicochemical properties. Building on this dataset, we construct the MolDeTox benchmark in a question–answering (QA) format and decompose molecular detoxification into three sub-tasks: (1) identifying toxicity-associated fragments, (2) replacing them with corresponding fragments from the non-toxic counterpart, and (3) generating the final detoxified molecule. Since each task is derived from a paired toxic/non-toxic molecule, MolDeTox provides explicit reference targets for both intermediate structural changes and final molecular outputs. These targets enable reference-based evaluation of whether a model recovers the observed structural transformation associated with the paired non-toxic molecule. Furthermore, the stepwise task design allows us to analyze where models succeed or fail at each stage of the detoxification process.

To further validate the utility of MolDeTox, we fine-tune a small language model using its training split and observe substantial improvements across the benchmark. Moreover, under the same ToxiEval protocol, the MolDeTox-based approaches achieve higher molecular repair performance than the prompt used in ToxiMol. These results demonstrate that the stepwise formulation of MolDeTox provides an effective framework for both learning and evaluating molecular detoxification.

Our contributions can be summarized as follows:

- We construct a novel dataset, *ToxicityCliff*, consisting of structurally similar toxic/non-toxic molecular pairs with contrasting toxicity outcomes, capturing localized structural changes associated with molecular detoxification.
- We propose **MolDeTox**, a QA-based benchmark that decomposes molecular detoxification into interpretable sub-tasks, enabling fine-grained analysis of model behavior and sensitivity to subtle structural differences associated with contrasting toxicity outcomes.
- We demonstrate the utility of the MolDeTox training split, showing substantial improvements in both in-benchmark performance and molecular repair success under an independent evaluation protocol.

2 Related Work

LLMs for molecular optimization. Recent studies have explored the use of large language models (LLMs) for molecular optimization. These approaches leverage LLMs to modify molecular structures toward desired physicochemical, biological, and pharmacological properties (Liu et al., 2023; Ye et al., 2025; Nguyen & Grover, 2024). More recent work has extended this paradigm to multi-property optimization, requiring simultaneous control over multiple molecular objectives (Ye et al., 2025; Dey et al., 2025a;b).

Several benchmarks have also been developed to systematically evaluate the molecular editing and optimization capabilities of LLMs. MolLangBench evaluates language-guided molecular recognition, editing, and generation across diverse molecular properties (Cai et al., 2025), while ChemCoT-Bench evaluates molecular optimization within a broader set of chemical reasoning tasks involving physicochemical properties and biological targets (Li et al., 2025).

LLMs for molecular toxicity As LLMs improve their ability to reason over molecular structures, their capability to predict drug toxicity from structural representations such as SMILES has also advanced. Recent studies (Yang et al., 2025b; Chen et al., 2025) employ prompt-based approaches to predict specific toxicity endpoints, including cardiotoxicity and drug-induced osteotoxicity, directly from SMILES inputs. Furthermore, CoTox (Park et al., 2025) introduces a framework that integrates

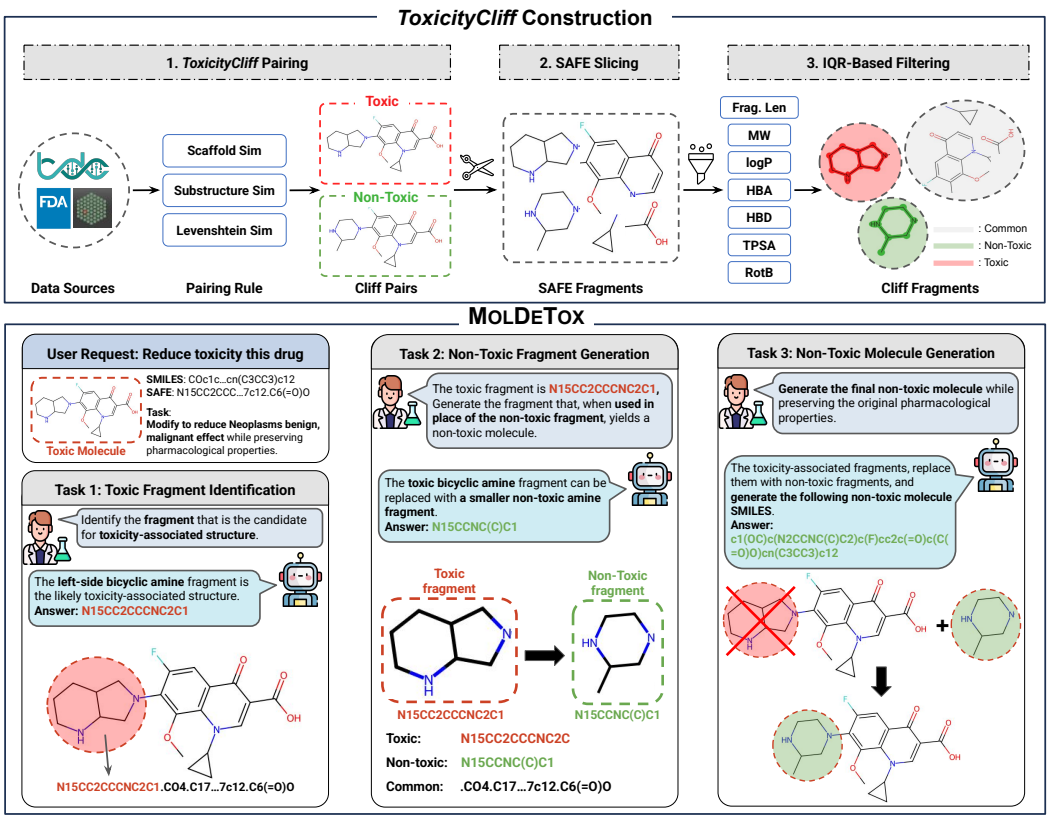


Figure 1: Pipeline of MolDeTox. The upper panel illustrates the *ToxicityCliff* construction pipeline. The lower panel shows how each *ToxicityCliff* pair is converted into three step-wise QA tasks.

both chemical and biological information within a chain-of-thought reasoning paradigm, enabling the prediction of multiple organ-level toxicities.

Benchmarks have also been introduced to evaluate how well LLMs capture molecular toxicity knowledge. ToxReason (Park et al., 2026) presents a benchmark for assessing LLMs' mechanistic toxicity reasoning by requiring models to predict three organ-level toxicities (liver, heart, and kidney) based on adverse outcome pathways(AOPs). Moreover, ToxiMol (Lin et al., 2026) introduces a VQA-style benchmark to evaluate whether multimodal large language models (MLLMs) can repair toxic molecules, covering 11 primary toxicity tasks and 660 representative toxic molecules. Its evaluation determines repair success based on a combination of molecular validity, QED (Bickerton et al., 2012), Synthetic Accessibility, Lipinski's Rule of Five evaluation, Similarity, and a binary safety score predicted by TxGemma-Predict (Wang et al., 2025). In contrast, our **MolDeTox** focuses on structurally similar molecular pairs with opposite toxicity labels collected across diverse toxicity dataset. It formulates molecular detoxification in a reference-based evaluation setting, complementing oracle-based evaluation of open-ended repairs.

3 MolDeTox

MolDeTox is designed to evaluate whether language models can carry out a stepwise molecular detoxification process inspired by practical medicinal chemistry strategies. In medicinal chemistry, toxicity-associated structural features are often identified and subsequently modified or replaced to mitigate toxicity while preserving desirable properties of the original compound (Limban et al., 2018). Following this perspective, MolDeTox decomposes detoxification into three stages: identifying toxicity-relevant fragments, proposing localized non-toxic replacements, and generating a non-toxic analog while preserving as much of the original molecule as possible. This formulation motivates our

dataset construction pipeline, which aims to curate toxic/non-toxic molecule pairs that are globally similar yet locally different in chemically meaningful ways.

3.1 *ToxicityCliff* CONSTRUCTION

MolDeTox is constructed from 10 binary-labeled drug toxicity datasets spanning 49 toxicity endpoints. We collect datasets from multiple sources, including the U.S. Food and Drug Administration (FDA), Therapeutic Data Commons (TDC), and SIDER. From FDA resources, we include DILIst (Thakkar et al., 2020), DICTrank (Qu et al., 2023), and DIRIL (Connor et al., 2024). From TDC, we utilize datasets from the toxicity category, including hERG blockers (Wang et al., 2016), hERG Central (Du et al., 2011), hERG Karim (Karim et al., 2021), Ames Mutagenicity (Xu et al., 2012), Skin Reaction (Alves et al., 2015), Tox21 (Huang et al., 2016), and ClinTox (Gayvert et al., 2016). In addition, we incorporate cytochrome P450 (CYP) inhibition datasets (Veith et al., 2009) from the metabolism category to capture toxicity-related metabolic effects (Kondža et al., 2025). Finally, we include SIDER (Kuhn et al., 2016), which provides drug side-effect information.

To preserve the physicochemical properties of molecules while removing toxic effects, we leverage the concept of activity cliff to the toxicity domain. Activity cliffs, where structurally similar molecules exhibit large differences in biological properties, have been widely studied (Van Tilborg et al., 2022; Kim et al., 2025), and are often explained through matched molecular pair analysis (MMPA), which links small, localized structural changes to significant variations in molecular activity (Kalgutkar, 2019; Yang et al., 2023). Specifically, for each endpoint, we identify pairs of molecules that are structurally highly similar but exhibit opposite toxicity labels, and construct a novel dataset termed *ToxicityCliff*.

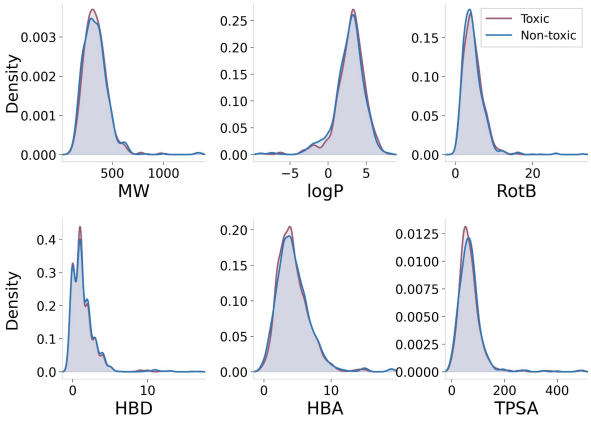


Figure 2: Property distributions of toxic and non-toxic molecules in testset.

To enable minimal molecular edits, we further decompose each molecule into fragment-level representations using Sequential Attachment-based Fragment Embedding(SAFE) (Noutahi et al., 2024). SAFE is a BRICS-based fragment representation that organizes molecules as sequences of chemically meaningful substructures. It offers better interpretability than SMILES and aligning naturally with fragment-based drug discovery. Based on this representation, we categorize fragments within each pair into three groups: common fragments shared by both molecules, toxicity-associated fragments present only in the toxic molecule, and non-toxic fragments only in the non-toxic counterpart. We also compare the frequent toxic fragments with known Structural Alerts in Appendix B.

To improve data quality, we apply an interquartile range (IQR)-based filtering strategy (Dash et al., 2023) to remove ambiguous instances. For each toxicity cliff pair, we compute the absolute differences between the physicochemical properties of the toxic and non-toxic molecules, and apply IQR filtering to exclude pairs with excessively large property deviations. The considered properties include six descriptors derived from *Lipinski's Rule* (Lipinski et al., 1997) and *Veber's Rule* (Veber et al., 2002), capturing key physicochemical attributes relevant to molecular behavior. As a result, the property distributions of toxic and non-toxic molecules are highly similar, as shown in Figure 2, indicating that the physicochemical properties are well preserved within each pair. In addition, we apply IQR-based filtering to remove pairs with unusually large fragment counts. The construction process is illustrated in Figure 1, and detailed filtering procedures are provided in Appendix A.

For each endpoint, we applied a 9:1 scaffold-based split based on the scaffold of the toxic molecule, while endpoints with insufficient data were assigned entirely to the test set. Detailed dataset statistics, including the number of training and test samples, are reported in Table A.

3.2 MolDeTox Construction

A key objective of MolDeTox is to formulate molecular detoxification as a structured, stepwise process rather than merely evaluating the final generation of a non-toxic molecule. Our benchmark explicitly tests whether it can (i) identify the toxicity-relevant fragment, (ii) propose a localized non-toxic replacement, and (iii) integrate such edits into a full non-toxic molecule. This step-by-step formulation enables fine-grained analysis of where detoxification succeeds or fails, distinguishing performance on intermediate fragment-level decisions from final molecule generation. Examples for each task can be found in Figure 1.

We denote a toxic/non-toxic molecular pair as (M_t, M_{nt}), and define the corresponding SAFE-based fragment sets as the shared fragment set $\mathcal{F}_{\text{common}}$, the toxic-only fragment set $\mathcal{F}_{t\text{-only}}$, and the non-toxic-only fragment set $\mathcal{F}_{nt\text{-only}}$, as provided in *ToxicityCliff*. Using these molecule- and fragment-level correspondences, we construct three benchmark tasks that capture distinct stages of stepwise molecular detoxification.

Task 1: Toxic Fragment Identification Given a toxic molecule M_t represented by its SMILES and the corresponding SAFE fragment set (connected by '.'), the objective is to identify the toxic-only fragment set $\mathcal{F}_{t\text{-only}}$. This task evaluates whether a model can recognize toxicity-relevant substructures by jointly considering the molecular structure and its fragment-level representation.

Task 2: Non-Toxic Fragment Generation The input is the same as in Task 1, with the addition of the toxic-only fragment set $\mathcal{F}_{t\text{-only}}$. The objective is to generate the non-toxic-only SAFE fragment set $\mathcal{F}_{nt\text{-only}}$. This task evaluates whether a model can recover the fragment-level replacement observed in the paired non-toxic molecule after the toxicity-associated fragment has been identified.

Task 3: Non-Toxic Molecule Generation The input consists only of the toxic molecule M_t. The objective is to generate the full non-toxic molecule M_{nt}, requiring end-to-end detoxification at the whole-molecule level. This task evaluates whether a model can reconstruct the paired non-toxic molecule from the toxic input, providing an end-to-end measure of reference-based editing.

All three tasks are categorized into single- and multi-fragment settings depending on the number of fragments that need to be modified. Detailed descriptions of this setup can be found in Appendix C. By decomposing the task, MolDeTox offers a granular perspective on a model's capabilities. This allows for a fine-grained analysis of whether the generative failure stems from incorrect toxicity identification or a lack of chemical reasoning during the replacement and integration stages. Moreover, each instance is accompanied by its toxicity endpoint description, providing context that allows the model to adapt to the specific toxicity setting. Endpoint descriptions are provided in Appendix F.

4 Experiments and Results

4.1 Experimental Setup

We evaluate model performance on the proposed benchmark across three tasks, each under single- and multi-fragment settings. For all experiments, we set the sampling temperature to 0.7 and run each model three times, reporting the average performance. The corresponding standard deviations are provided in Appendix E. All experiments were conducted using 8 NVIDIA RTX PRO 6000 Blackwell GPUs. We use a consistent toxicity-aware prompting protocol across models, with task-specific instructions for each evaluation setting. Full prompt templates and their components are provided in Appendix Tables P–T.

Evaluated Models We evaluate a diverse set of large language models (LLMs), including both closed-source and open-source models. Closed-source models include GPT-4o (Hurst et al., 2024), GPT-5.2 (OpenAI, 2025), and Gemini 3.1 Flash-Lite (Google DeepMind, 2025), while open-source models include Qwen3 models at multiple scales (4B, 8B, 30B-A3B, and 235B-A22B) (Yang et al., 2025a), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Gemma-3-27B (Team et al., 2025), DeepSeek-Llama-70B (Guo et al., 2025). For the Qwen3-4B-Instruct + *w/ Image* setting, we use the corresponding Qwen3-VL-Instruct model (Bai et al., 2025).

Table 1: Main results on Task 1 and Task 2 of MolDeTox. We compare model performance under single- and multi-fragment settings. Best results are in **bold**, and second-best results are underlined.

Model	Task 1: Toxic Frag. ID			Task 2: NonToxic Frag. Gen.				
	Single	Multi		Single		Multi		
	Acc.(%)	Acc.(%)	F1-Score	Acc.(%)	Lev. Dist	Acc.(%)	F1-Score	Lev. Dist
GPT-4o	36.12	1.37	0.3509	5.73	4.77	0.23	0.0161	4.76
GPT-5.2	39.23	4.56	0.4184	8.11	3.86	0.29	0.0283	4.14
+ w/ Image	38.59	5.27	0.4300	8.54	3.81	0.23	0.0277	4.15
+ 4-Shot	53.80	13.31	0.5461	21.26	**3.56**	4.82	0.1273	**4.01**
Gemini 3.1 Flash-Lite	13.48	1.12	0.0872	3.22	8.40	0.00	0.0078	7.05
Qwen3-235B-A22B-Inst.	35.25	3.44	0.3768	6.33	4.34	0.10	0.0096	5.62
DeepSeek-Llama-70B	34.96	5.40	0.4345	3.36	4.84	0.10	0.0107	4.89
Qwen3-30B-A3B-Inst.	33.01	4.70	0.4289	1.52	5.17	0.00	0.0014	5.82
Gemma-3-27B	40.93	6.11	0.4625	2.18	6.78	0.00	0.0016	8.33
Qwen3-8B	29.66	9.68	0.5261	1.94	4.39	0.19	0.0031	4.72
Llama-3.1-8B-Inst.	16.59	5.69	0.4580	0.56	29.90	0.00	0.0016	47.22
Qwen3-4B-Inst.	28.20	4.21	0.4108	0.85	4.88	0.00	0.0015	5.72
+ w/ Image	27.56	4.36	0.5377	0.14	5.60	0.00	0.0006	7.35
+ 4-Shot	35.99	7.39	0.5010	7.96	4.58	0.53	0.0302	5.24
Qwen3-4B-Inst. SFT								
+ All Tasks	**61.57**	**26.07**	**0.6870**	**28.88**	3.65	**4.84**	**0.1413**	4.71

Variant Settings To explore effective strategies for molecular detoxification, we introduce several variants across the tasks. First, we apply in-context learning (ICL) by providing four examples (4-shot) retrieved from the train set of the same dataset endpoint based on structural similarity to the query molecule (Xian et al., 2025). Examples containing the query molecule or its corresponding answer molecule are excluded from retrieval. In Task 3, we additionally investigate a step-by-step reasoning approach (w/ CoT), where the model first performs toxic fragment identification (Task 1) and replacement (Task 2) before generating the final non-toxic molecule. The corresponding prompt example is provided in Table L.

Finally, we evaluate two molecule generation formats: direct SMILES generation and SAFE-based generation. In SAFE-based Generation, the model generates SAFE strings, which are subsequently decoded into SMILES using a SAFE-to-SMILES decoding function[2].

4.2 Evaluation Metrics

We use task-specific metrics reflecting the objective of each detoxification stage. For Task 1, we report accuracy and F1-Score to evaluate exact and partial identification of toxicity-associated fragments. For Task 2, accuracy, F1-Score, and Levenshtein distance evaluate the correctness and similarity of generated replacement fragments. For Task 3, we evaluate molecular generation using exact-match accuracy (Acc.), BLEU-1, Morgan fingerprint similarity, and molecular validity, following standard molecular generation evaluation protocols (Yang et al., 2025c). We additionally introduce the Property Retention Score (PRS) to quantify whether detoxification preserves the physicochemical properties of the original molecule. Inspired by QED, PRS aggregates properties derived from *Lipinski's Rule* (Lipinski et al., 1997) and *Veber's Rule* (Veber et al., 2002) and measures their change after detoxification. PRS ranges from 0 to 1, with higher values indicating better property preservation. We also report Copy Rate (Copy), defined as the proportion of generated molecules that are identical to the input toxic molecule. A high Copy Rate indicates cases where the model fails to perform any structural modification. Full definitions are provided in Appendix D.

4.3 Performance on MolDeTox across Models

Tables 1 and 2 show a consistent increase in difficulty from Task 1 to Task 3 across models. Performance decreases as the tasks provide less explicit guidance and require greater generative flexibility. Among the prompting strategies, in-context learning yields the most consistent gains, with 4-shot prompting improving performance across all three tasks. For Task 3, we examine CoT prompting and molecule generation format. CoT prompting provides inconsistent gains, particularly in the multi-fragment setting. In contrast, SAFE generation improves average molecular validity over direct SMILES generation by 8.9% and 12.4% in the single- and multi-fragment settings, respectively.

[2] https://github.com/datamol-io/safe

Table 2: Comparison of inference and training strategies on Task 3 of MolDeTox. We compare model performance under single- and multi-fragment settings across SMILES and SAFE generation. Best results are in **bold**, and second-best results are underlined.

	Single Fragment						Multi Fragment					
Model	**Acc.(%)**	**BLEU1**	**Morgan**	**Validity**	**PRS**	**Copy(%)**	**Acc.(%)**	**BLEU1**	**Morgan**	**Validity**	**PRS**	**Copy(%)**
SMILES Generation												
GPT-5.2	**3.54**	0.9455	0.5697	**0.9272**	0.8776	9.2	0.19	0.9101	0.4567	**0.9234**	0.8737	8.6
DeepSeek-Llama-70B	2.40	0.9397	0.3768	0.7020	0.6578	1.3	**0.45**	0.9113	0.3150	0.7064	0.6555	1.3
Qwen3-8B	0.43	**0.9599**	0.5983	0.8474	0.8862	62.7	0.36	**0.9469**	0.4373	0.8579	**0.8773**	56.3
Qwen3-4B-Inst.	0.14	0.8106	**0.6535**	0.9163	**0.9220**	86.3	0.00	0.8728	**0.4631**	0.8450	0.8592	71.8
SAFE Generation												
GPT-4o	1.82	0.8550	0.5459	0.9149	0.8741	17.2	0.13	0.8257	0.4298	0.9167	0.8619	14.6
GPT-5.2	3.82	0.8829	0.5933	0.9328	0.8960	28.0	0.16	0.8685	0.4883	0.9486	0.8952	29.5
+ w/ Image	3.73	0.8658	0.5762	0.9133	0.8960	22.8	0.29	0.8546	0.4725	0.9317	0.8952	22.4
+ CoT	3.68	0.8855	0.5880	0.9373	0.8947	20.3	0.26	0.8524	0.4686	0.9348	0.8944	21.7
+ 4-Shot	15.99	0.9140	0.6586	0.9682	0.9218	15.8	3.17	0.8635	0.4979	0.9533	0.8973	18.1
Gemini 3.1 Flash-Lite	0.45	0.6150	0.1554	0.8857	0.7842	0.5	0.26	0.7112	0.2475	0.9313	0.8138	1.3
Qwen3-235B-A22B-Inst.	1.52	0.8871	0.5778	0.9555	0.9157	30.6	0.10	0.8536	0.4794	0.9510	0.9111	35.5
DeepSeek-Llama-70B	2.37	0.7368	0.4170	0.7999	0.7459	0.4	0.39	0.7172	0.3410	0.7967	0.7394	0.9
Qwen3-30B-A3B-Inst.	0.74	0.9326	0.6872	0.9830	0.9755	83.8	0.00	0.8976	0.5401	0.9861	0.9789	88.4
Gemma-3-27B	1.65	0.8587	0.6091	0.9415	0.8934	34.8	0.06	0.8126	0.4660	0.9328	0.8711	32.0
Qwen3-8B	0.32	0.9282	0.6760	0.9834	0.9673	83.1	0.16	0.8987	0.5403	0.9887	0.9770	87.8
Llama-3.1-8B-Inst.	0.00	0.2917	0.0512	0.5266	0.6482	18.7	0.00	0.2889	0.0524	0.5349	0.6439	16.7
Qwen3-4B-Inst.	0.04	0.9402	0.6949	0.9905	**0.9852**	93.1	0.00	0.8890	0.5345	0.9784	0.9739	90.5
+ w/ Image	0.00	0.9335	0.6883	0.9832	0.9769	92.8	0.05	0.8916	0.5354	0.9821	0.9782	91.3
+ CoT	1.20	0.7498	0.4643	0.8105	0.7802	22.8	0.10	0.7506	0.3867	0.8419	0.7973	25.5
+ 4-Shot	3.47	**0.9422**	**0.7064**	**0.9942**	0.9812	89.1	0.53	**0.9041**	**0.5424**	**0.9991**	**0.9859**	92.5
Qwen3-4B-Inst. SFT												
+ Only Task 3	12.65	0.8985	0.6228	0.9586	0.9107	12.2	3.39	0.7997	0.3954	0.9184	0.8274	1.1
+ Only Task 3 w/ CoT	12.34	0.8808	0.6106	0.9385	0.8924	0.6	1.29	0.7129	0.3012	0.8399	0.7410	0.4
+ All Tasks	19.09	0.9105	0.6504	0.9643	0.9148	2.7	2.29	0.8102	0.3894	0.9303	0.8385	1.5
+ All Tasks w/ CoT	**19.58**	0.8956	0.6405	0.9487	0.8918	0.9	**9.62**	0.7834	0.4062	0.8935	0.7963	0.4

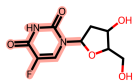
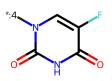
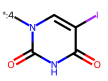
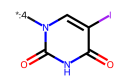
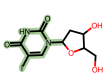
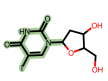
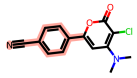
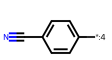
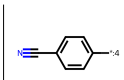
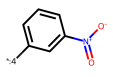
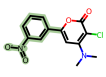
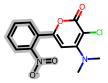
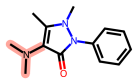
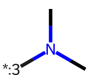
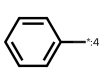
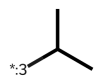
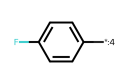
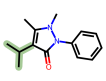
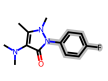
We further identify input copying as a prominent failure mode in Task 3. Similar behavior has been reported in prior LLM-based molecular optimization studies (Ye et al., 2025; Guo et al., 2026), where models reproduce the input molecule without performing the requested modification. In particular, Qwen3 models exhibit high Copy Rates, exceeding 90% for Qwen3-4B-Instruct under SAFE generation. This behavior can lead to high similarity, validity, and property-retention scores despite low exact-match accuracy. Simply reproducing the input preserves a valid molecular structure and its original properties without performing detoxification. Interestingly, this tendency is substantially reduced under the CoT setting, suggesting that explicit intermediate reasoning may help mitigate input copying.

4.4 Supervised Fine-Tuning on MolDeTox

To assess whether the MolDeTox training split provides useful supervision for molecular detoxification, we fine-tune Qwen3-4B-Instruct either only on the final molecule generation task (*Only Task 3*) or jointly on all three MolDeTox tasks (*All Tasks*). This setup allows us to examine whether intermediate fragment-level supervision benefits final molecule generation. In final molecule generation, we additionally consider CoT-based variants in which the model generates intermediate detoxification steps, including fragment identification and replacement, before producing the final molecule. For all training configurations, single- and multi-fragment samples are balanced at a 1:1 ratio. We use LoRA with rank $r = 32$, $\alpha = 64$, and a dropout rate of 0.05, and train for three epochs with a learning rate of 3×10^{-4}. All models are trained using 8 NVIDIA RTX PRO 6000 Blackwell GPUs.

Fine-tuning on *All Tasks* results in overall stronger performance on Tasks 1 and 2 than the other evaluated models. The benefit of the MolDeTox training set is also evident on Task 3. Both *Only Task 3* and *All Tasks* substantially improve exact-match accuracy over the base model. In particular, training on *All Tasks* further improves single-fragment accuracy from 12.65% with *Only Task 3* to 19.09%. For Task 3, we additionally evaluate CoT-based training, where *All Tasks w/ CoT* shows a pronounced improvement in the more challenging multi-fragment setting, achieving 9.62% accuracy. Fine-tuning also substantially reduces the Copy Rate, and CoT-based training nearly eliminates input copying. Together, these results suggest that supervision from MolDeTox not only improves prediction accuracy but also mitigates a major failure mode in molecular detoxification, where models simply reproduce the input molecule without modification.

Table 3: Representative examples of step-wise detoxification trajectories after fine-tuning on *All Tasks w/ CoT*. Circles and crosses indicate successful and failed predictions for each task, respectively.

Cases	Toxic Molecule	Task 1		Task 2		Task 3	
		Ground Truth	Prediction	Ground Truth	Prediction	Ground Truth	Prediction
Task 1 ○ Task 2 ○ Task 3 ○							
Task 1 ○ Task 2 × Task 3 ×							
Task 1 × Task 2 × Task 3 ×							

5 Step-wise Analysis of Detoxification

A central motivation for MolDeTox is that final-molecule evaluation alone cannot reveal which intermediate structural decisions lead to success or failure. To identify the major bottlenecks in molecular detoxification, we analyze the dependency of prediction outcomes across the decomposed tasks of MolDeTox. Unlike ToxiMol benchmark, MolDeTox is built from paired toxic/non-toxic molecules, allowing intermediate fragment-level targets to be derived from their structural differences. This design enables failure attribution across the detoxification process, revealing where errors emerge and propagate toward the final molecular output.

Figure 3 shows a strong dependency between the three stages of detoxification for Qwen3-4B-Instruct fine-tuned on *All Tasks w/ CoT*. Of the test molecules, 44.5% are correctly localized in Task 1, but only 14.3% remain correct through Tasks 2 and 3. Thus, most molecules with a correct Task 1 prediction are lost at the replacement stage, while no molecule recovers from an incorrect Task 1 prediction at a later stage. Once the correct replacement is identified in Task 2, the corresponding Task 3 reconstruction is consistently successful. These results identify replacement inference as the principal bottleneck following localization and show that errors introduced at either intermediate stage propagate directly to recovery of the paired non-toxic reference molecule.

We further examine representative examples to understand how these failure patterns manifest in actual molecular transformations. Table 3 presents predictions from the same model. The examples reveal that unsuccessful detoxification can originate from different stages of the process. In some cases, the model correctly identifies the toxicity-relevant fragment but fails to generate the corresponding non-toxic replacement, while in others, the error occurs during fragment identification. By exposing these intermediate outcomes, MolDeTox enables a more fine-grained analysis of model behavior than end-to-end evaluation alone, allowing failures in fragment identification, replacement, and final molecule generation to be examined separately.

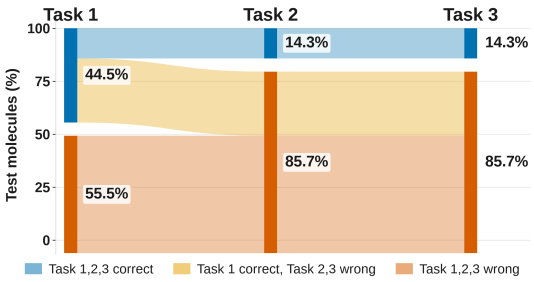


Figure 3: Task-wise success dependency across the three detoxification tasks.

6 Oracle-based Evaluation

We evaluate whether MolDeTox supervision generalizes to molecular detoxification beyond recovering the paired non-toxic molecule. While MolDeTox evaluates recovery of a specific detoxified counterpart, valid detoxification may admit alternative non-toxic molecules that differ from the paired target. We therefore conduct an additional evaluation using the ToxiMol benchmark set and the Safety

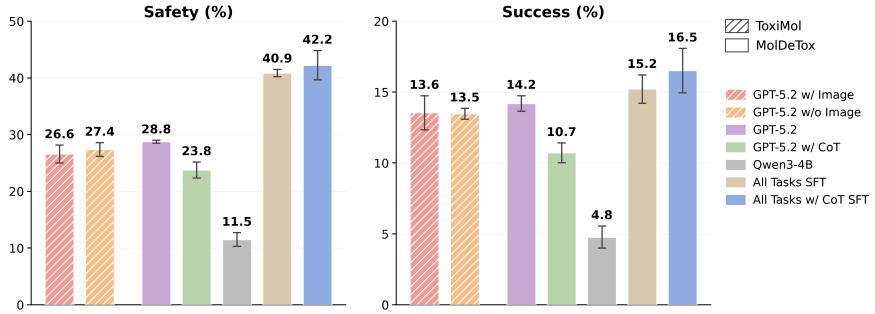


Figure 4: Oracle-based evaluation on the filtered ToxiMol benchmark set of 460 toxic molecules. We evaluate two prompting-based settings, as well as models fine-tuned on *All Tasks* and *All Tasks w/ CoT*, using the Safety and Success metrics from ToxiEval.

and Success metrics defined by ToxiEval (Lin et al., 2025). Safety measures whether the generated molecule satisfies the toxicity criterion determined by TxGemma-9B-Predict (Wang et al., 2025), while Success additionally requires satisfaction of all ToxiEval criteria, including QED, synthetic accessibility, Lipinski's Rule of Five, and structural similarity. To avoid training leakage, we remove ToxiMol instances whose toxic molecules appear in the MolDeTox training set. We additionally exclude instances where the ToxiEval safety oracle classifies the original toxic molecule as non-toxic. This filtering step ensures consistency between the benchmark labels and the oracle-based evaluation. After these filtering steps, 460 toxic molecules remain for evaluation.

For a controlled comparison, both the ToxiMol and MolDeTox settings use the same GPT-5.2 model. All results are reported as the mean over three independent runs, with standard deviations shown in Figure 4. While the prompting-based settings show comparable performance, supervised fine-tuning on the MolDeTox training set consistently improves both Safety and Success. In particular, fine-tuning on *All Tasks*, especially with CoT supervision, achieves the strongest overall performance. These results indicate that MolDeTox supervision transfers beyond paired-target recovery and remains effective under an independent oracle-based evaluation.

7 Conclusion

We introduce MolDeTox, a reference-based benchmark for evaluating fragment-level structural editing across molecular detoxification. Our experiments show that language models remain challenged by subtle structural differences associated with contrasting toxicity outcomes. We further quantify input copying as a prominent failure mode, consistent with prior LLM-based molecular optimization studies. We also find that fragment-based SAFE generation produces more valid molecules than direct SMILES generation, suggesting that generating molecules through fragment composition can offer an advantage over directly generating the full molecular string. Overall, the MolDeTox training set provides effective supervision across the detoxification process, substantially improving performance across all three tasks. Beyond reference-based evaluation, improvements under the oracle-based setting provide additional evidence that MolDeTox supervision can transfer to a broader molecular detoxification setting.

8 Limitations

MolDeTox adopts a reference-based evaluation setting, which may not fully capture valid detoxification solutions that differ from the paired non-toxic reference. The intermediate fragment-level targets are derived from structural differences within *ToxicityCliff* pairs and should not be interpreted as causal toxicity determinants. MolDeTox also considers each toxicity endpoint independently and does not explicitly capture interactions among multiple concurrent toxicity liabilities. Finally, both MolDeTox and our additional oracle-based evaluation rely on existing toxicity labels or computational predictions rather than prospective experimental validation. Accordingly, the actual toxicity reduction of generated molecules requires confirmation through appropriate biological assays.

AI USE STATEMENT

Generative AI tools were used to assist with literature search and related-work discovery, manuscript editing, and code development. All AI-assisted content, including references and code, was reviewed and verified by the authors. The authors take full responsibility for the content of the paper.

ETHICS STATEMENT

We acknowledge that research on molecular toxicity and structural modification may present dual-use risks, as such knowledge could potentially be misused to guide the design of harmful compounds. MolDeTox is designed specifically for toxicity mitigation and does not provide experimental procedures for synthesizing or increasing the toxicity of compounds. We encourage its use for toxicity reduction, safety assessment, and responsible molecular design.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the MolDeTox dataset construction, filtering criteria, and data splits in Section 3 and Appendix A. The definitions of the three benchmark tasks and evaluation metrics are provided in Section 4.2 and Appendix D. Experimental settings, including prompts, model configurations, and supervised fine-tuning hyperparameters, are described in Section 4 and Appendix G. All reported experiments are conducted over three independent runs where applicable, with the corresponding mean and standard deviation reported in Section 4.3 and Appendix E. The dataset, prompts, and code required to reproduce the benchmark and experiments will be provided in the supplementary materials and an anonymous repository.

REFERENCES

Vinicius M Alves, Eugene Muratov, Denis Fourches, Judy Strickland, Nicole Kleinstreuer, Carolina H Andrade, and Alexander Tropsha. Predicting chemically-induced skin reactions. part i: Qsar models of skin sensitization and their application to identify potentially hazardous compounds. *Toxicology and applied pharmacology*, 284(2):262–272, 2015.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.

Feiyang Cai, Jiahui Bai, Tao Tang, Guijuan He, Joshua Luo, Tianyu Zhu, Srikanth Pilla, Gang Li, Ling Liu, and Feng Luo. Mollangbench: A comprehensive benchmark for language-prompted molecular structure recognition, editing, and generation. *arXiv preprint arXiv:2505.15054*, 2025.

Yi-Qi Chen, Tao Yu, Zheng-Qi Song, Chen-Yu Wang, Jiang-Tao Luo, Yong Xiao, Heng Qiu, Qing-Qing Wang, and Hai-Ming Jin. Application of large language models in drug-induced osteotoxicity prediction. *Journal of Chemical Information and Modeling*, 65(7):3370–3379, 2025.

Skylar Connor, Ting Li, Yanyan Qu, Ruth A Roberts, and Weida Tong. Generation of a drug-induced renal injury list to facilitate the development of new approach methodologies for nephrotoxicity. *Drug discovery today*, 29(4):103938, 2024.

Ch Sanjeev Kumar Dash, Ajit Kumar Behera, Satchidananda Dehuri, and Ashish Ghosh. An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal*, 6:100164, 2023.

Vishal Dey, Xiao Hu, and Xia Ning. GeLLM³O: Generalizing large language models for multi-property molecule optimization. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25192–25221, Vienna, Austria, July

2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1225. URL https://aclanthology.org/2025.acl-long.1225/.

Vishal Dey, Xiao Hu, and Xia Ning. Large language models for controllable multi-property multi-objective molecule optimization, 2025b. URL https://arxiv.org/abs/2505.23987.

Fang Du, Haibo Yu, Beiyan Zou, Joseph Babcock, Shunyou Long, and Min Li. hergcentral: a large database to store, retrieve, and analyze compound-human ether-a-go-go related gene channel interactions to facilitate cardiotoxicity assessment in drug development. *Assay and drug development technologies*, 9(6):580–588, 2011.

Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimer, Arnav Shah, Haonan Duan, Omar Ibrahim, Hani Goodarzi, Chris J Maddison, et al. Bioreason: Incentivizing multimodal biological reasoning within a dna-llm model. *arXiv preprint arXiv:2505.23579*, 2025.

Adibvafa Fallahpour, Arman Seyed-Ahmadi, Parsa Idehpour, Omar Ibrahim, Purav Gupta, Jack Naimer, Kevin Zhu, Arnav Shah, Shihao Ma, Abhinav Adduri, et al. Bioreason-pro: Advancing protein function prediction with multimodal biological reasoning. *bioRxiv*, pp. 2026–03, 2026.

Kaitlyn M Gayvert, Neel S Madhukar, and Olivier Elemento. A data-driven approach to predicting successes and failures of clinical trials. *Cell chemical biology*, 23(10):1294–1301, 2016.

Google DeepMind. Gemini 3.1 flash lite model card, 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Flash-Lite-Model-Card.pdf.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yan Guo, Menglan Luo, Wenbo Zhang, Peidong Liu, Jin Liu, Shudong Huang, Jiancheng Lv, Bowen Ke, and Xianggen Liu. Few-shot molecular property optimization via a domain-specialized large language model. *Chemical Science*, 17(10):4928–4941, 2026.

Ruili Huang, Menghang Xia, Dac-Trung Nguyen, Tongan Zhao, Srilatha Sakamuru, Jinghua Zhao, Sampada A Shahane, Anna Rossoshek, and Anton Simeonov. Tox21challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Frontiers in Environmental Science*, 3:85, 2016.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Ana-Maria Istrate, Fausto Milletari, Fabrizio Castrotorres, Jakub M Tomczak, Michaela Torkar, Donghui Li, and Theofanis Karaletsos. rbio1-training scientific reasoning llms with biological world models as soft verifiers. *bioRxiv*, pp. 2025–08, 2025.

Amit S Kalgutkar. Designing around structural alerts in drug discovery. *Journal of medicinal chemistry*, 63(12):6276–6302, 2019.

Abdul Karim, Matthew Lee, Thomas Balle, and Abdul Sattar. Cardiotox net: a robust predictor for herg channel blockade based on deep learning meta-feature ensembles. *Journal of cheminformatics*, 13(1):60, 2021.

Hajung Kim, Jueon Park, Junseok Choe, Sheunheun Baek, Hyeon Hwang, and Jaewoo Kang. Graphcliff: Short-long range gating for subtle differences but critical changes. *arXiv preprint arXiv:2511.03170*, 2025.

Martin Kondža, Josipa Bukić, Ivan Ćavar, and Biljana Tubić. Targeted but troubling: Cyp450 inhibition by kinase and parp inhibitors and its clinical implications. *Drugs and drug candidates*, 4 (2):24, 2025.

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079, 2016.

Hao Li, He Cao, Bin Feng, Yanjun Shao, Xiangru Tang, Zhiyuan Yan, Li Yuan, Yonghong Tian, and Yu Li. Beyond chemical qa: Evaluating llm's chemical reasoning with modular chemical operations. *arXiv preprint arXiv:2505.21318*, 2025.

Carmen Limban, Diana C Nuţă, Cornel Chiriţă, Simona Negreș, Andreea L Arsene, Marina Goumenou, Spyros P Karakitsios, Aristidis M Tsatsakis, and Dimosthenis A Sarigiannis. The use of structural alerts to avoid the toxicity of pharmaceuticals. *Toxicology reports*, 5:943–953, 2018.

Fei Lin, Tengchao Zhang, Ziyang Gong, and Fei-Yue Wang. Toxieval-zkp: A structure-private verification framework for molecular toxicity repair tasks. *IFAC-PapersOnLine*, 59(34):36–41, 2025.

Fei Lin, Ziyang Gong, Cong Wang, Tengchao Zhang, Yonglin Tian, Yining Jiang, Ji Dai, Chao Guo, Xiaotong Yu, Xue Yang, et al. Breaking bad molecules: are mllms ready for structure-level molecular detoxification? In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 11399–11410, 2026.

Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1-3):3–25, 1997.

Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. Chatgpt-powered conversational drug editing using retrieval and domain feedback, 2023. URL https://arxiv.org/abs/2305.18090.

Siddharth M Narayanan, James D Braza, Ryan-Rhys Griffiths, Albert Bou, Geemi Wellawatte, Mayk Caldas Ramos, Ludovico Mitchener, Samuel G Rodriques, and Andrew D White. Training a scientific reasoning model for chemistry. *arXiv preprint arXiv:2506.17238*, 2025.

Tung Nguyen and Aditya Grover. Lico: Large language models for in-context molecular optimization. *arXiv preprint arXiv:2406.18851*, 2024.

Emmanuel Noutahi, Cristian Gabellini, Michael Craig, Jonathan SC Lim, and Prudencio Tossou. Gotta be safe: a new framework for molecular design. *Digital Discovery*, 3(4):796–804, 2024.

OpenAI. Update to gpt-5 system card: Gpt-5.2, December 2025. URL https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf.

Jueon Park, Yein Park, Minju Song, Soyon Park, Donghyeon Lee, Seungheun Baek, and Jaewoo Kang. Cotox: Chain-of-thought-based molecular toxicity reasoning and prediction. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 4002–4007. IEEE, 2025.

Jueon Park, Wonjune Jang, Chanhwi Kim, Yein Park, and Jaewoo Kang. Toxreason: A benchmark for mechanistic chemical toxicity reasoning via adverse outcome pathway. *arXiv preprint arXiv:2604.06264*, 2026.

Yanyan Qu, Ting Li, Zhichao Liu, Dongying Li, and Weida Tong. Dictrank: The largest reference list of 1318 human drugs ranked by risk of drug-induced cardiotoxicity using fda labeling. *Drug Discovery Today*, 28(11):103770, 2023.

Iurii Sushko, Elena Salmina, Vladimir A Potemkin, Gennadiy Poda, and Igor V Tetko. Toxalerts: a web server of structural alerts for toxic chemicals and compounds with potential adverse reactions, 2012.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Shraddha Thakkar, Ting Li, Zhichao Liu, Leihong Wu, Ruth Roberts, and Weida Tong. Drug-induced liver injury severity and toxicity (dilist): binary classification of 1279 drugs by human hepatotoxicity. *Drug discovery today*, 25(1):201–208, 2020.

Derek Van Tilborg, Alisa Alenicheva, and Francesca Grisoni. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of chemical information and modeling*, 62(23): 5938–5951, 2022.

Daniel F Veber, Stephen R Johnson, Hung-Yuan Cheng, Brian R Smith, Keith W Ward, and Kenneth D Kopple. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of medicinal chemistry*, 45(12):2615–2623, 2002.

Henrike Veith, Noel Southall, Ruili Huang, Tim James, Darren Fayne, Natalia Artemenko, Min Shen, James Inglese, Christopher P Austin, David G Lloyd, et al. Comprehensive characterization of cytochrome p450 isozyme selectivity across chemical libraries. *Nature biotechnology*, 27(11): 1050–1055, 2009.

Eric Wang, Samuel Schmidgall, Paul F. Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. Txgemma: Efficient and agentic llms for therapeutics. 2025.

Shuangquan Wang, Huiyong Sun, Hui Liu, Dan Li, Youyong Li, and Tingjun Hou. Admet evaluation in drug discovery. 16. predicting herg blockers by combining multiple pharmacophores and machine learning approaches. *Molecular pharmaceutics*, 13(8):2855–2866, 2016.

Ziting Xian, Jiawei Gu, Lingbo Li, and Shangsong Liang. Molrag: unlocking the power of large language models for molecular property prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15513–15531, 2025.

Congying Xu, Feixiong Cheng, Lei Chen, Zheng Du, Weihua Li, Guixia Liu, Philip W Lee, and Yun Tang. In silico prediction of chemical ames mutagenicity. *Journal of chemical information and modeling*, 52(11):2840–2847, 2012.

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

Hengzheng Yang, Jian Xiu, Weiqi Yan, Kaifeng Liu, Huizi Cui, Zhibang Wang, Qizheng He, Yilin Gao, and Weiwei Han. Large language models as tools for molecular toxicity prediction: Ai insights into cardiotoxicity. *Journal of Chemical Information and Modeling*, 65(5):2268–2282, 2025b.

Zaifei Yang, Hong Chang, Ruibing Hou, Shiguang Shan, and Xilin Chen. Knowmol: Advancing molecular large language models with multi-level chemical knowledge. *arXiv preprint arXiv:2510.19484*, 2025c.

Ziyi Yang, Shaohua Shi, Li Fu, Aiping Lu, Tingjun Hou, and Dongsheng Cao. Matched molecular pair analysis in drug discovery: methods and recent applications. *Journal of Medicinal Chemistry*, 66(7):4361–4377, 2023.

Geyan Ye, Xibao Cai, Houtim Lai, Xing Wang, Junhong Huang, Longyue Wang, Wei Liu, and Xiangxiang Zeng. Drugassist: A large language model for molecule optimization. *Briefings in Bioinformatics*, 26(1):bbae693, 2025.

Jiajun Yu, Yizhen Zheng, Huan Yee Koh, Shirui Pan, Tianyue Wang, and Haishuai Wang. Collaborative expert llms guided multi-objective molecular optimization. *arXiv preprint arXiv:2503.03503*, 2025.

Zihan Zhao, Bo Chen, Ziping Wan, Lu Chen, Xuanze Lin, Shiyang Yu, Situo Zhang, Da Ma, Zichen Zhu, Danyang Zhang, et al. Chemdfm-r: A chemical reasoning llm enhanced with atomized chemical knowledge. *arXiv preprint arXiv:2507.21990*, 2025.

A *ToxicityCliff* CONSTRUCTION DETAILS

Table A: Statistics of ***ToxicityCliff***, including the number of endpoints, molecules, toxicity cliff pairs, and resulting split sizes across datasets.

Dataset	DILIst	DICTrank	DIRIL	hERG	AMES	Skin Reaction	Tox21	ClinTox	CYP Inh.	SIDER	Total
Endpoint N	1	1	1	1	1	1	12	1	5	25	49
Molecule N	97	27	6	4,303	1,133	56	1,635	30	3,355	212	10,854
***ToxicityCliff* N**	87	20	3	5,355	6,444	109	30,236	15	8,502	1,829	52,600
Train	78	18	0	4,831	6,325	106	29,947	14	7,807	1,498	50,624
Test	9	2	3	524	119	3	289	1	695	331	1,976

To construct MolDeTox, we design a multi-stage pairing pipeline that identifies toxic/non-toxic molecule pairs that differ in toxicity label while remaining highly similar in overall structure and molecular characteristics. Our goal is to frame detoxification as a *localized molecular editing problem*, where toxicity changes can be attributed to a small number of fragment-level modifications rather than a complete molecular redesign.

Let M_t and M_{nt} denote a toxic molecule and a non-toxic molecule, respectively. The final toxicity cliff pairs are constructed through the following steps.

Step 1: Toxic/Non-Toxic candidate pairing For each toxicity dataset, we first split molecules according to the endpoint label and canonicalize all SMILES strings. We then construct candidate toxic/non-toxic pairs using stringent whole-molecule similarity criteria, as in MoleculeACE (Van Tilborg et al., 2022). A pair (M_t, M_{nt}) is retained if it satisfies *at least one* of the following:

- Bemis–Murcko scaffold-based ECFP4 Tanimoto similarity ≥ 0.9,
- Full-SMILES ECFP4 Tanimoto similarity ≥ 0.9,
- Normalized SMILES Levenshtein similarity ≥ 0.9

These criteria ensure that paired molecules are globally similar enough to support meaningful toxicity-aware editing.

Step 2: SAFE conversion To analyze local structural differences, we convert each paired SMILES into its SAFE representation. SAFE expresses a molecule as a dot-separated sequence of fragments, allowing explicit comparison of shared and differing substructures between toxic and non-toxic molecules.

Formally, let

$$S_t = \mathrm{SAFE}(M_t), \qquad S_{nt} = \mathrm{SAFE}(M_{nt})$$

where each SAFE string is decomposed into fragment sets by splitting on the dot separator:

$$\mathcal{F}_t = \mathrm{split}(S_t, \texttt{'.'}), \qquad \mathcal{F}_{nt} = \mathrm{split}(S_{nt}, \texttt{'.'})$$

Step 3: SAFE fragment comparison Using the fragment sets $\mathcal{F}_t$ and $\mathcal{F}_{nt}$, we define three fragment groups:

$$\mathcal{F}_{\text{common}} = \mathcal{F}_t \cap \mathcal{F}_{nt},$$
$$\mathcal{F}_{t\text{-only}} = \mathcal{F}_t \setminus \mathcal{F}_{nt},$$
$$\mathcal{F}_{nt\text{-only}} = \mathcal{F}_{nt} \setminus \mathcal{F}_t.$$

We also record their cardinalities:

$$n_{\text{common}} = |\mathcal{F}_{\text{common}}|, \quad n_{t\text{-only}} = |\mathcal{F}_{t\text{-only}}|, \quad n_{nt\text{-only}} = |\mathcal{F}_{nt\text{-only}}|$$

This representation provides a fragment-level view of which substructures are preserved and which are edited between the toxic and non-toxic molecules.

Step 4: SAFE-based fragment filtering We next filter candidate pairs to retain only those that reflect small and interpretable fragment-level edits. Concretely, we apply the following rules:

- **Shared core constraint:**
$$n_{\text{common}} \neq 0$$
This ensures that the toxic and non-toxic molecules preserve at least one common fragment.
- **Non-trivial edit constraint:** pairs with
$$n_{t\text{-only}} = 0 \quad \text{and} \quad n_{nt\text{-only}} = 0$$
are removed, since they do not differ at the fragment level.
- **Fragment length outlier filtering:** we remove pairs if any fragment in $\mathcal{F}_{t\text{-only}}$ or $\mathcal{F}_{nt\text{-only}}$ has SAFE length above an outlier threshold determined from the empirical fragment-length distribution using an interquartile-range (IQR)-based rule. In our benchmark, this corresponds to removing pairs containing toxic-only or non-toxic-only fragments with SAFE length ≥ 28.
- **Fragment count outlier filtering:** we restrict the number of edited fragments using thresholds derived from the empirical distributions of $n_{t\text{-only}}$ and $n_{nt\text{-only}}$ under the same IQR-based outlier rule. In practice, we retain only pairs satisfying
$$n_{t\text{-only}} \leq 4 \qquad \text{and} \qquad n_{nt\text{-only}} \leq 4$$
This removes cases requiring excessively many fragment edits.

Together, these rules ensure that the remaining pairs reflect localized, minimal, and interpretable fragment-level differences.

Step 5: Molecular property filtering To ensure that paired molecules differ mainly in toxicity rather than in broad physicochemical characteristics, we compute RDKit descriptors for both molecules:
$$\{\text{MW}, \text{logP}, \text{TPSA}, \text{HBD}, \text{HBA}, \text{RotB}\}$$

For each descriptor p, we compute the absolute difference
$$\Delta p = |p(M_t) - p(M_{nt})|$$

We then remove pairs that exhibit outlier-level changes in any descriptor using an interquartile-range (IQR)-based filtering procedure over the empirical distribution of descriptor differences. That is, if a pair is an outlier in at least one molecular property difference, it is discarded. This step reduces unrealistic or pharmacologically mismatched pairs and preserves drug-like similarity beyond fragment-level overlap.

Step 6: Final toxicity cliff pairs The remaining pairs form the final **toxicity cliff pairs** used in MolDeTox. Each retained pair is characterized by:

- **high global similarity** at the whole-molecule level,
- **localized and interpretable fragment differences** under SAFE decomposition,
- **limited physicochemical deviation** across key molecular properties

As a result, MolDeTox focuses on toxic/non-toxic pairs where detoxification can be interpreted as a small number of meaningful fragment edits, making the benchmark well-suited for evaluating whether models can identify toxicity-relevant fragments, propose minimal non-toxic edits, and generate full non-toxic analogs while preserving the key characteristics of the original molecule.

B Comparative analysis of Structural Alerts

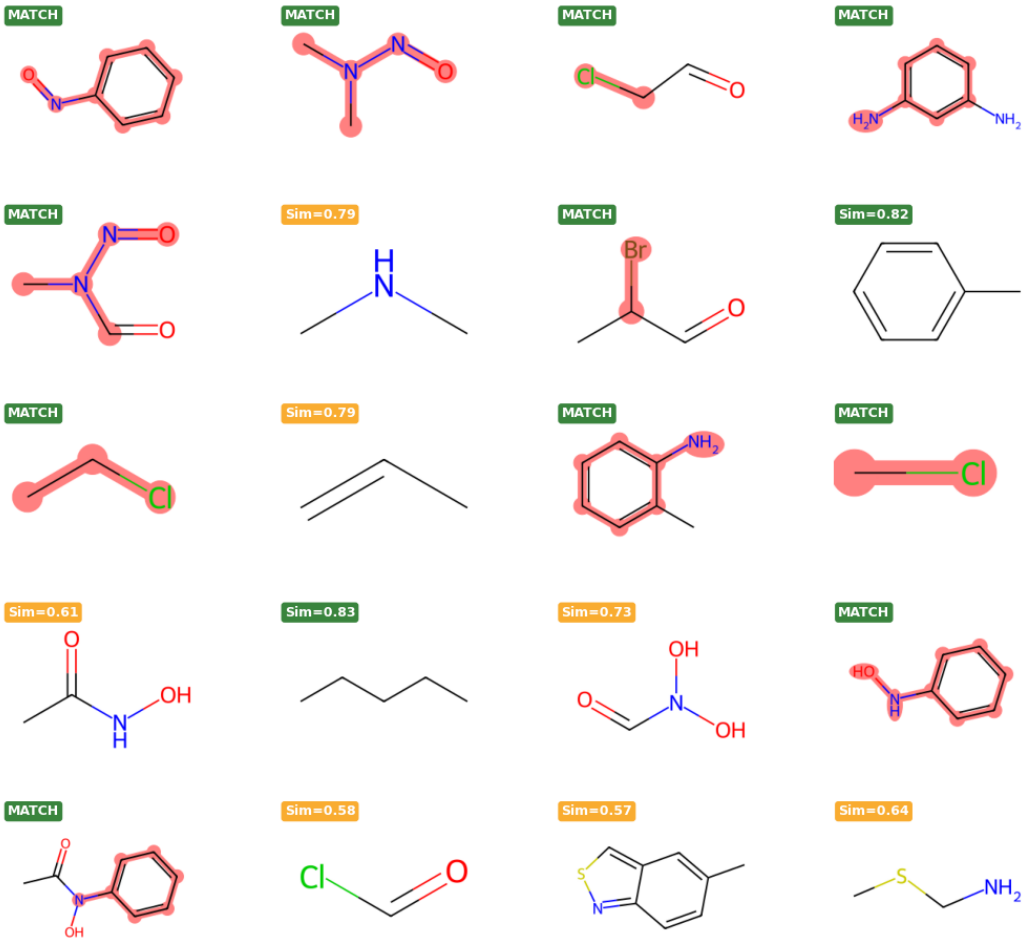


Figure 5: Structural alert overlap for the top-20 Ames Mutagenicity toxicity-associated fragments.

To assess the reliability of toxic fragments extracted from *ToxicityCliff*, we further compare them against known structural alerts (SAs) collected in ToxAlerts (Sushko et al., 2012). ToxAlerts is a curated repository of toxicity-related SAs collected from expert knowledge and literature sources for various toxicity endpoints. We select the most closely aligned endpoints, namely Ames Mutagenicity and Skin Reaction, and compare the top-20 most frequent toxicity-associated fragments from each dataset against the corresponding endpoint-specific SAs.

For each extracted fragment, we first determine whether it directly matches a known SA for the corresponding endpoint. For fragments without an exact match, we compute the maximum Tanimoto similarity between the fragment and the full set of SAs using 1024-bit ECFP4 fingerprints. In Figures 5 and 6, green labels indicate either exact SA matches or high structural similarity scores of at least 0.8, while yellow `Sim=` labels indicate moderate structural similarity scores of at least 0.5.

A substantial portion of the extracted fragments directly matches known SAs, while many of the remaining fragments still exhibit moderate structural similarity to documented alerts. These results demonstrate that the toxic fragments extracted by MolDeTox are strongly aligned with established toxicity knowledge and capture meaningful toxicity-relevant substructures. This supports the reliability of our data-driven fragment extraction procedure and suggests that the toxic-only fragments derived from *ToxicityCliff* provide chemically grounded targets for toxicity-aware molecular editing.

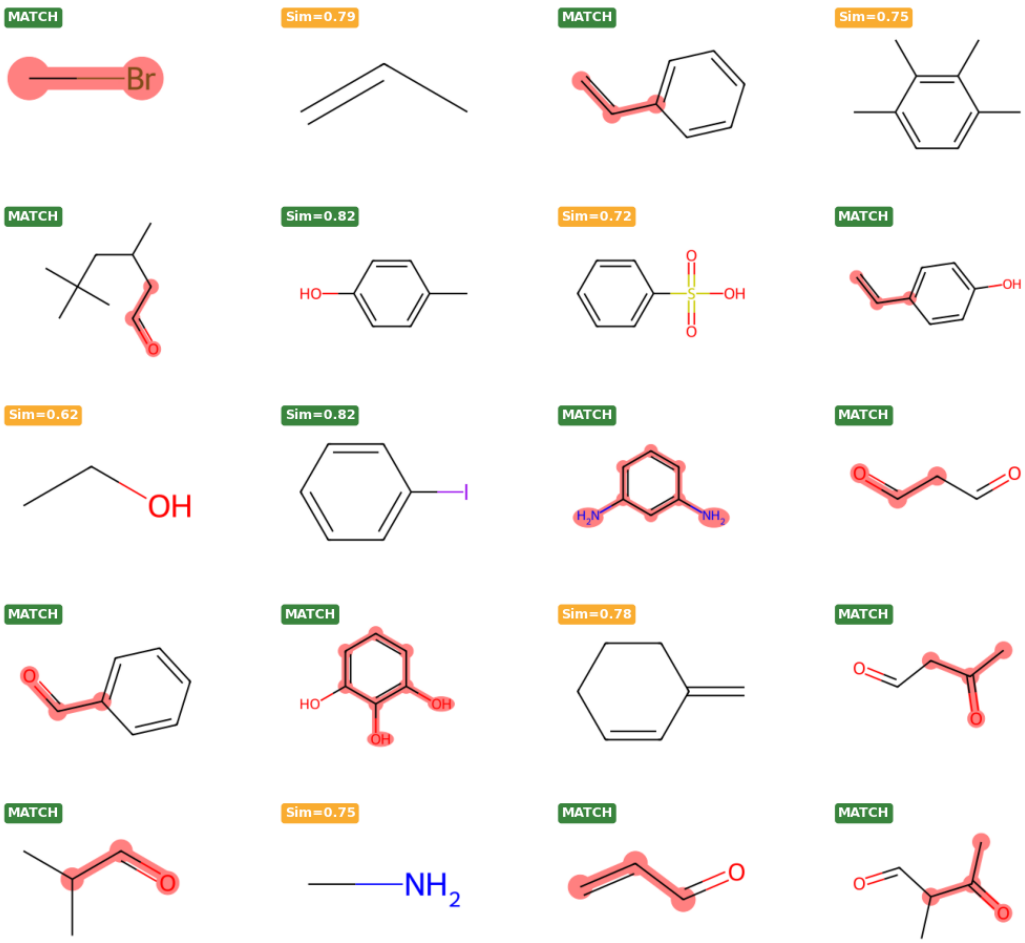


Figure 6: Structural alert overlap for the top-20 Skin Reaction toxicity-associated fragments.

C MolDeTox Construction Details

C.1 Task Construction Details

MolDeTox instances are constructed from toxicity cliff pairs (M_t, M_{nt}) and their corresponding SAFE fragment decompositions. From each pair, we derive the shared fragment set $\mathcal{F}_{\text{common}}$, the toxic-only fragment set $\mathcal{F}_{t\text{-only}}$, and the non-toxic-only fragment set $\mathcal{F}_{nt\text{-only}}$, which are then used to instantiate the three benchmark tasks.

Task 1: Toxic Fragment Identification For Task 1, the input is the toxic molecule M_t, and the target output is the toxic-only SAFE fragment set $\mathcal{F}_{t\text{-only}}$.

Task 2: Non-Toxic Fragment Generation For Task 2, the input consists of the toxic molecule M_t together with the toxic-only fragment set $\mathcal{F}_{t\text{-only}}$, and the target output is the non-toxic-only SAFE fragment set $\mathcal{F}_{nt\text{-only}}$.

Task 3: Non-Toxic Molecule Generation For Task 3, the input is the toxic molecule M_t, and the target output is the full non-toxic molecule M_{nt}.

Table B: Benchmark instance statistics of MolDeTox across tasks and fragment settings.

Task	Train		Test		Total
	Single	Multi	Single	Multi	
Task 1: Toxic Fragment Identification	12,213	38,411	1,026	950	52,600
Task 2: Non-Toxic Fragment Generation	9,646	40,978	943	1,033	52,600
Task 3: Non-Toxic Molecule Generation	9,646	40,978	943	1,033	52,600
Total benchmark instances	31,505	120,367	2,912	3,016	157,800

Table C: Source-level benchmark instance statistics of MolDeTox. We report endpoint counts, molecule counts, and the number of benchmark QA instances in the train and test splits. Raw *ToxicityCliff* pair statistics are reported in Table A of the main paper.

Source	Endpoint N	Molecule N	Train QA N	Test QA N	Total QA N
AMES	1	1,133	18,975	357	19,332
ClinTox	1	30	42	3	45
DICTrank	1	27	54	6	60
DILIst	1	97	234	27	261
DIRIL	1	6	0	9	9
hERG	1	4,303	14,493	1,572	16,065
Metabolism	5	3,355	23,421	2,085	25,506
SIDER	25	212	4,494	993	5,487
Skin Reaction	1	56	318	9	327
Tox21	12	1,635	89,841	867	90,708
Total	**49**	**10,854**	**151,872**	**5,928**	**157,800**

Single- and Multi-fragment settings We divide benchmark instances into single- and multi-fragment settings based on the number of fragment tokens in the SAFE labels. Let N_t and N_{nt} denote the numbers of toxic-only and non-toxic-only fragments, respectively.

For Task 1, an instance is single-fragment if $N_t = 1$, and multi-fragment if $N_t \geq 2$. For Task 2 and Task 3, an instance is single-fragment if $N_t = 1$ and $N_{nt} = 1$, and multi-fragment if $N_t \geq 2$ or $N_{nt} \geq 2$. This partition is applied consistently to all QA variants built from the same molecule-pair rows.

C.2 Benchmark Statistics

The toxicity cliff pair statistics and endpoint-level split results are summarized in the main paper in Table A. In total, MolDeTox is constructed from 52,600 toxicity cliff pairs. Each pair is converted into one QA instance for each of the three tasks, resulting in 52,600 instances per task and 157,800 benchmark instances in total.

Table B reports the train/test and single-/multi-fragment distributions for each task. MolDeTox contains 151,872 training instances and 5,928 test instances, with multi-fragment cases accounting for the majority of the benchmark. Additional endpoint level statistics are provided in Tables C.

D Evaluation Metric Details

D.1 Task 1 Metrics (Toxic Fragment Identification)

Task 1 evaluates whether the model correctly identifies the toxicity-associated fragment(s). Since SAFE represents multiple fragments as dot-separated strings, we split both the prediction and reference at the dot ('.') and compare the resulting fragments irrespective of their order. Results are

reported separately for single-fragment and multi-fragment examples according to the number of reference fragments.

Single / Multi Acc. (%) A prediction is considered correct when it contains exactly the same fragment(s) as the reference, regardless of their order. For multi-fragment examples, all reference fragments must be recovered without additional or missing fragments. Accuracy is the percentage of examples satisfying this exact-match criterion.

Multi F1-Score Exact match requires the entire set of fragments to be correct and therefore does not provide partial credit when only some fragments are successfully identified. For multi-fragment settings, we additionally compute fragment-level Precision, Recall, and F1-Score based on the overlap between predicted and reference fragments:

$$\text{Precision} = \frac{N_{\text{match}}}{N_{\text{pred}}}, \qquad \text{Recall} = \frac{N_{\text{match}}}{N_{\text{gold}}},$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

where N_{match} is the number of matched fragments, and N_{pred} and N_{gold} are the numbers of predicted and reference fragments, respectively. We report the average F1-Score over the multi-fragment subset.

D.2 Task 2 Metrics (Non-Toxic Fragment Generation)

Task 2 evaluates whether the model generates the correct replacement fragment(s). As in Task 1, the predicted and reference SAFE strings are split at the dot ('.') into individual fragments and compared irrespective of their order. Results are reported separately for single-fragment and multi-fragment examples.

Single / Multi Acc. (%) We use the same fragment-level exact-match criterion as in Task 1. A prediction is considered correct only when the generated fragments exactly match the reference fragments, with no additional or missing fragments. Accuracy is reported separately for the single- and multi-fragment subsets.

Single / Multi Lev. Dist. In addition to exact match, we report fragment-level Levenshtein distance to measure the similarity of non-exact predictions to the reference fragments. Rather than computing the edit distance over the entire SAFE string, we compare individual fragments. For each predicted fragment p_j, we compute its Levenshtein distance to every reference fragment and retain the minimum distance:

$$d(p_j) = \min_{g_i \in G} \text{Lev}(p_j, g_i),$$

where G denotes the collection of reference fragments. The fragment-level Levenshtein distance for an example is then computed as

$$\text{LevFrag} = \frac{1}{N_{\text{pred}}} \sum_{j=1}^{N_{\text{pred}}} d(p_j),$$

where N_{pred} is the number of predicted fragments. We report the average distance over the corresponding single- or multi-fragment subset, with lower values indicating closer agreement with the reference fragments.

Multi F1-Score For multi-fragment examples, we additionally report the fragment-level F1-Score defined in Task 1, providing partial credit when only a subset of the reference replacement fragments is correctly generated.

D.3 Task 3 Metrics (Non-Toxic Molecule Generation)

Task 3 compares a predicted molecule string $\hat{m}$ with the gold non-toxic target molecule string m_{nt}. The code first attempts RDKit parsing and canonicalization, and then evaluates exact-match accuracy, string similarity, structural similarity, chemical validity, and a property-based score.

Acc. (%) The code computes molecule-level exact match by checking whether the canonical SMILES strings of the predicted molecule and the gold non-toxic molecule are identical:

$$\mathrm{EM}_{\mathrm{mol}}(\hat{m}, m_{nt}) = \begin{cases} 1 & \text{if } \mathrm{canon}(\hat{m}) = \mathrm{canon}(m_{nt}), \\ 0 & \text{otherwise} \end{cases}$$

The reported Acc.(%) is obtained by averaging this exact-match indicator over the relevant subset and converting it to a percentage:

$$\mathrm{Acc}_{\mathrm{mol}}(\%) = 100 \cdot \mathbb{E}[\mathrm{EM}_{\mathrm{mol}}]$$

Thus, Task 3 accuracy is the percentage version of exact-match correctness after canonical SMILES normalization.

BLEU1 We compute unigram BLEU between the predicted molecular string and the target molecular string.

Morgan FTS To evaluate structural similarity beyond string overlap, we compute Tanimoto similarity using Morgan fingerprints derived from canonical SMILES.

Validity Validity is the fraction of predictions that can be parsed as chemically valid molecules by RDKit. For direct SMILES generation, the model output is evaluated directly as a SMILES string. For SAFE generation, the model output is first converted into SMILES using the deterministic SAFE-to-SMILES conversion function provided by the original SAFE implementation, and the resulting SMILES string is then parsed by RDKit. If SAFE-to-SMILES conversion fails, the prediction is counted as invalid and assigned a validity score of 0. Thus, the SAFE-to-SMILES conversion step does not mask invalid generations or introduce an additional learned decoding model; it only reconstructs SMILES from generated SAFE strings, with failed reconstructions explicitly penalized.

PRS PRS is a property-retention score that measures how well a generated non-toxic molecule preserves the drug-relevant physicochemical profile of the original toxic molecule. It is computed from six molecular properties: molecular weight (MW), logP, hydrogen-bond acceptors (HBA), hydrogen-bond donors (HBD), polar surface area (PSA), and rotatable bonds (RotB). These properties cover the main factors used in *Lipinski's Rule* (Lipinski et al., 1997) and *Veber's Rule* (Veber et al., 2002).

For each molecule m, we compute a QED-inspired aggregated property score:

$$S(m) = \sum_i w_i \, d_i(m), \quad i \in \{\mathrm{MW}, \mathrm{logP}, \mathrm{HBA}, \mathrm{HBD}, \mathrm{PSA}, \mathrm{RotB}\}$$

where $d_i(m)$ denotes the desirability function for the i-th physicochemical property, and w_i is its corresponding weight. For a toxic molecule m_t and a generated molecule $\hat{m}$, we then measure the absolute difference between their aggregated property scores:

$$x = |S(m_t) - S(\hat{m})|$$

This difference is converted into the final PRS using exponential decay:

$$\mathrm{PRS}(\hat{m}, m_t) = \exp(-x)$$

The reported PRS is the average over the evaluation set.

Table D: Main results on Task 1 and Task 2 of MolDeTox (mean ± std over three runs). Best results are in **bold**, and second-best results are underlined.

Model	Task 1: Toxic Frag. ID			Task 2: NonToxic Frag. Gen.				
	Single	Multi		Single		Multi		
	Acc.(%)	Acc.(%)	F1	Acc.(%)	Lev. Dist	Acc.(%)	F1	Lev. Dist
GPT-4o	36.12 ± 0.90	1.37 ± 0.22	0.3509 ± 0.0063	5.73 ± 0.80	4.77 ± 0.34	0.23 ± 0.15	0.0161 ± 0.0048	4.76 ± 0.32
GPT-5.2	39.23 ± 0.21	4.56 ± 0.37	0.4184 ± 0.0006	8.11 ± 0.28	3.86 ± 0.08	0.29 ± 0.10	0.0283 ± 0.0012	4.14 ± 0.03
+ w/ Image	38.59 ± 0.17	5.27 ± 0.32	0.4300 ± 0.0016	8.54 ± 0.11	3.81 ± 0.03	0.23 ± 0.11	0.0277 ± 0.0006	4.15 ± 0.04
+ 4-Shot	53.80 ± 0.60	13.31 ± 0.64	0.5461 ± 0.0026	21.26 ± 0.50	**3.56 ± 0.03**	4.82 ± 0.18	0.1273 ± 0.0017	**4.01 ± 0.07**
Gemini 3.1 Flash-Lite	13.48 ± 9.68	1.12 ± 0.44	0.0872 ± 0.0445	3.22 ± 2.04	8.40 ± 1.92	0.00 ± 0.00	0.0078 ± 0.0065	7.05 ± 0.95
Qwen3-235B-A22B-Inst.	35.25 ± 0.57	3.44 ± 0.22	0.3768 ± 0.0031	6.33 ± 0.32	4.34 ± 0.04	0.10 ± 0.10	0.0096 ± 0.0020	5.62 ± 0.07
DeepSeek-Llama-70B	34.96 ± 1.14	5.40 ± 0.16	0.4345 ± 0.0071	3.36 ± 0.16	4.84 ± 0.05	0.10 ± 0.10	0.0107 ± 0.0012	4.89 ± 0.05
Qwen3-30B-A3B-Inst.	33.01 ± 0.69	4.70 ± 0.06	0.4289 ± 0.0015	1.52 ± 0.16	5.17 ± 0.06	0.00 ± 0.00	0.0014 ± 0.0003	5.82 ± 0.01
Gemma-3-27B	40.93 ± 0.41	6.11 ± 0.37	0.4625 ± 0.0015	2.18 ± 0.27	6.78 ± 0.07	0.00 ± 0.00	0.0016 ± 0.0007	8.33 ± 0.13
Qwen3-8B	29.66 ± 0.74	9.68 ± 0.28	0.5261 ± 0.0036	1.94 ± 0.24	4.39 ± 0.05	0.19 ± 0.00	0.0031 ± 0.0003	4.72 ± 0.03
Llama-3.1-8B-Inst.	16.59 ± 4.50	5.69 ± 2.62	0.4580 ± 0.0268	0.56 ± 0.27	29.90 ± 13.80	0.00 ± 0.00	0.0016 ± 0.0014	47.22 ± 10.03
Qwen3-4B-Inst.	28.20 ± 0.15	4.21 ± 0.11	0.4108 ± 0.0033	0.85 ± 0.11	4.88 ± 0.03	0.00 ± 0.00	0.0015 ± 0.0010	5.72 ± 0.02
+ w/ Image	27.56 ± 0.35	4.36 ± 0.32	0.5377 ± 0.0018	0.14 ± 0.06	5.60 ± 0.05	0.00 ± 0.00	0.0006 ± 0.0003	7.35 ± 3.06
+ 4-Shot	35.99 ± 1.21	7.39 ± 0.15	0.5010 ± 0.0012	7.96 ± 0.55	4.58 ± 0.09	0.53 ± 0.00	0.0302 ± 0.0008	5.24 ± 0.02
Qwen3-4B-Inst. SFT								
+ All Tasks	**61.57 ± 0.15**	**26.07 ± 0.49**	**0.6870 ± 0.0024**	**28.88 ± 2.10**	3.65 ± 0.09	**4.84 ± 0.70**	**0.1413 ± 0.0065**	4.71 ± 0.24

Copy Rate (%) By construction, the toxic and non-toxic molecules of a toxicity cliff are structurally very similar. A prediction that simply returns the input molecule unchanged is therefore highly similar to the gold non-toxic molecule as well, and scores well on the similarity-based metrics while being incorrect. We report the Copy Rate to detect this behavior: the fraction of predictions that are identical to the toxic input molecule m_t after canonicalization,

$$\mathrm{Copy}(\hat{m}, m_t) = \begin{cases} 1 & \text{if } \mathrm{canon}(\hat{m}) = \mathrm{canon}(m_t), \\ 0 & \text{otherwise} \end{cases} \qquad \mathrm{Copy}(\%) = 100 \cdot \mathbb{E}[\mathrm{Copy}]$$

As in Validity, SAFE predictions are converted to SMILES before comparison, and predictions that fail to parse are counted as non-copies.

E Main Results with Standard Deviations

To assess the stability of the main results, we repeat each experiment three times and report the mean and standard deviation in Tables D, E and F. Unless otherwise noted, all results are reported as mean ± standard deviation over three independent runs.

Table E: Task 3 of MolDeTox, accuracy and similarity to the target molecule (mean ± std over three runs). Best results are in **bold**, and second-best results are underlined.

	Single Fragment			Multi Fragment		
Model	**Acc.(%)**	**BLEU1**	**Morgan**	**Acc.(%)**	**BLEU1**	**Morgan**
SMILES Generation						
GPT-5.2	**3.54 ± 0.37**	0.9455 ± 0.0016	0.5697 ± 0.0019	0.19 ± 0.17	0.9101 ± 0.0010	0.4567 ± 0.0043
DeepSeek-Llama-70B	2.40 ± 0.52	0.9397 ± 0.0024	0.3768 ± 0.0059	**0.45 ± 0.06**	0.9113 ± 0.0025	0.3150 ± 0.0047
Qwen3-8B	0.43 ± 0.74	**0.9599 ± 0.0137**	0.5983 ± 0.0004	0.36 ± 0.62	**0.9469 ± 0.0265**	0.4373 ± 0.0346
Qwen3-4B-Inst.	0.14 ± 0.41	0.8106 ± 0.0011	**0.6535 ± 0.0020**	0.00 ± 0.00	0.8728 ± 0.0444	**0.4631 ± 0.0111**
SAFE Generation						
GPT-4o	1.82 ± 0.34	0.8550 ± 0.0139	0.5459 ± 0.0102	0.13 ± 0.06	0.8257 ± 0.0050	0.4298 ± 0.0057
GPT-5.2	3.82 ± 0.12	0.8829 ± 0.0062	0.5933 ± 0.0064	0.16 ± 0.15	0.8685 ± 0.0058	0.4883 ± 0.0035
+ w/ Image	3.73 ± 0.23	0.8658 ± 0.0022	0.5762 ± 0.0024	0.29 ± 0.28	0.8546 ± 0.0098	0.4725 ± 0.0016
+ CoT	3.68 ± 0.38	0.8855 ± 0.0063	0.5880 ± 0.0057	0.26 ± 0.06	0.8524 ± 0.0030	0.4686 ± 0.0023
+ 4-Shot	15.99 ± 0.41	0.9140 ± 0.0056	0.6586 ± 0.0027	3.17 ± 0.22	0.8635 ± 0.0024	0.4979 ± 0.0011
Gemini 3.1 Flash-Lite	0.45 ± 0.37	0.6150 ± 0.0284	0.1554 ± 0.0438	0.26 ± 0.20	0.7112 ± 0.0349	0.2475 ± 0.0615
Qwen3-235B-A22B-Inst.	1.52 ± 0.06	0.8871 ± 0.0024	0.5778 ± 0.0044	0.10 ± 0.10	0.8536 ± 0.0017	0.4794 ± 0.0010
DeepSeek-Llama-70B	2.37 ± 0.48	0.7368 ± 0.0110	0.4170 ± 0.0059	0.39 ± 0.10	0.7172 ± 0.0081	0.3410 ± 0.0057
Qwen3-30B-A3B-Inst.	0.74 ± 0.11	0.9326 ± 0.0008	0.6872 ± 0.0011	0.00 ± 0.00	0.8976 ± 0.0014	0.5401 ± 0.0010
Gemma-3-27B	1.65 ± 0.06	0.8587 ± 0.0020	0.6091 ± 0.0018	0.06 ± 0.05	0.8126 ± 0.0029	0.4660 ± 0.0005
Qwen3-8B	0.32 ± 0.11	0.9282 ± 0.0023	0.6760 ± 0.0011	0.16 ± 0.06	0.8987 ± 0.0026	0.5403 ± 0.0018
Llama-3.1-8B-Inst.	0.00 ± 0.00	0.2917 ± 0.2510	0.0512 ± 0.0439	0.00 ± 0.00	0.2889 ± 0.2500	0.0524 ± 0.0452
Qwen3-4B-Inst.	0.04 ± 0.06	0.9402 ± 0.0035	0.6949 ± 0.0022	0.00 ± 0.00	0.8890 ± 0.0031	0.5345 ± 0.0020
+ w/ Image	0.00 ± 0.00	0.9335 ± 0.0012	0.6883 ± 0.0008	0.05 ± 0.06	0.8916 ± 0.0090	0.5354 ± 0.0281
+ CoT	1.20 ± 0.32	0.7498 ± 0.0123	0.4643 ± 0.0063	0.10 ± 0.10	0.7506 ± 0.0095	0.3867 ± 0.0076
+ 4-Shot	3.47 ± 0.26	**0.9422 ± 0.0018**	**0.7064 ± 0.0018**	0.53 ± 0.00	**0.9041 ± 0.0005**	**0.5424 ± 0.0007**
Qwen3-4B-Inst. SFT						
+ Only Task 3	12.65 ± 0.95	0.8985 ± 0.0005	0.6228 ± 0.0059	3.39 ± 0.84	0.7997 ± 0.0157	0.3954 ± 0.0104
+ Only Task 3 w/ CoT	12.34 ± 0.58	0.8808 ± 0.0029	0.6106 ± 0.0005	1.29 ± 0.15	0.7129 ± 0.0042	0.3012 ± 0.0044
+ All Tasks	19.09 ± 0.83	0.9105 ± 0.0031	0.6504 ± 0.0040	2.29 ± 0.28	0.8102 ± 0.0073	0.3894 ± 0.0040
+ All Tasks w/ CoT	**19.58 ± 0.52**	0.8956 ± 0.0075	0.6405 ± 0.0048	**9.62 ± 0.64**	0.7834 ± 0.0021	0.4062 ± 0.0048

Table F: Task 3 of MolDeTox, validity, property retention and copying (mean ± std over three runs). Best and second-best of the remaining columns are in **bold** and underlined.

	Single Fragment			Multi Fragment		
Model	**Validity**	**PRS**	**Copy(%)**	**Validity**	**PRS**	**Copy(%)**
SMILES Generation						
GPT-5.2	**0.9272 ± 0.0012**	0.8776 ± 0.0077	9.22 ± 0.41	**0.9234 ± 0.0078**	0.8737 ± 0.0074	8.64 ± 0.76
DeepSeek-Llama-70B	0.7020 ± 0.0094	0.6578 ± 0.0136	1.30 ± 0.21	0.7064 ± 0.0028	0.6555 ± 0.0064	1.34 ± 0.30
Qwen3-8B	0.8474 ± 0.0507	0.8862 ± 0.0088	62.67 ± 0.67	0.8579 ± 0.0425	**0.8773 ± 0.0048**	56.29 ± 0.81
Qwen3-4B-Inst.	0.9163 ± 0.0034	**0.9220 ± 0.0018**	86.30 ± 0.82	0.8450 ± 0.0201	0.8592 ± 0.0064	71.79 ± 1.51
SAFE Generation						
GPT-4o	0.9149 ± 0.0139	0.8741 ± 0.0148	17.18 ± 0.95	0.9167 ± 0.0039	0.8619 ± 0.0041	14.63 ± 0.06
GPT-5.2	0.9328 ± 0.0066	0.8960 ± 0.0120	27.96 ± 1.82	0.9486 ± 0.0061	0.8952 ± 0.0118	29.46 ± 0.53
+ w/ Image	0.9133 ± 0.0023	0.8960 ± 0.0120	22.84 ± 0.20	0.9317 ± 0.0103	0.8952 ± 0.0118	22.35 ± 0.11
+ CoT	0.9373 ± 0.0070	0.8947 ± 0.0087	20.28 ± 1.09	0.9348 ± 0.0030	0.8944 ± 0.0036	21.70 ± 1.36
+ 4-Shot	0.9682 ± 0.0074	0.9218 ± 0.0081	17.22 ± 3.48	0.9533 ± 0.0027	0.8973 ± 0.0030	18.37 ± 0.90
Gemini 3.1 Flash-Lite	0.8857 ± 0.0076	0.7842 ± 0.0669	0.49 ± 0.43	0.9313 ± 0.0034	0.8138 ± 0.0217	1.34 ± 0.81
Qwen3-235B-A22B-Inst.	0.9555 ± 0.0021	0.9157 ± 0.0021	30.64 ± 1.59	0.9510 ± 0.0024	0.9111 ± 0.0019	35.46 ± 0.96
DeepSeek-Llama-70B	0.7999 ± 0.0143	0.7459 ± 0.0154	0.40 ± 0.27	0.7967 ± 0.0098	0.7394 ± 0.0081	0.87 ± 0.25
Qwen3-30B-A3B-Inst.	0.9830 ± 0.0011	0.9755 ± 0.0024	83.85 ± 0.70	0.9861 ± 0.0015	0.9789 ± 0.0015	88.44 ± 0.65
Gemma-3-27B	0.9415 ± 0.0031	0.8934 ± 0.0055	34.80 ± 0.52	0.9328 ± 0.0025	0.8711 ± 0.0034	31.99 ± 0.47
Qwen3-8B	0.9834 ± 0.0006	0.9673 ± 0.0024	83.09 ± 0.41	0.9887 ± 0.0015	0.9770 ± 0.0018	87.77 ± 0.49
Llama-3.1-8B-Inst.	0.5266 ± 0.4528	0.6482 ± 0.1217	18.66 ± 23.61	0.5349 ± 0.4631	0.6439 ± 0.1133	16.74 ± 21.36
Qwen3-4B-Inst.	0.9905 ± 0.0038	**0.9852 ± 0.0029**	93.14 ± 0.27	0.9784 ± 0.0034	0.9739 ± 0.0034	90.49 ± 0.79
+ w/ Image	0.9832 ± 0.0021	0.9769 ± 0.0020	92.82 ± 0.41	0.9821 ± 0.0009	0.9782 ± 0.0016	91.28 ± 0.78
+ CoT	0.8105 ± 0.0138	0.7802 ± 0.0110	22.79 ± 0.56	0.8419 ± 0.0119	0.7973 ± 0.0120	25.48 ± 1.21
+ 4-Shot	**0.9942 ± 0.0013**	0.9812 ± 0.0015	90.30 ± 1.73	**0.9991 ± 0.0005**	**0.9859 ± 0.0004**	93.08 ± 1.39
Qwen3-4B-Inst. SFT						
+ Only Task 3	0.9586 ± 0.0018	0.9107 ± 0.0027	12.20 ± 0.78	0.9184 ± 0.0171	0.8274 ± 0.0152	1.14 ± 0.48
+ Only Task 3 w/ CoT	0.9385 ± 0.0053	0.8924 ± 0.0021	0.63 ± 0.34	0.8399 ± 0.0048	0.7410 ± 0.0053	0.36 ± 0.31
+ All Tasks	0.9643 ± 0.0012	0.9148 ± 0.0013	2.74 ± 0.21	0.9303 ± 0.0042	0.8385 ± 0.0043	1.54 ± 0.20
+ All Tasks w/ CoT	0.9487 ± 0.0093	0.8918 ± 0.0070	0.85 ± 0.08	0.8935 ± 0.0039	0.7963 ± 0.0049	0.39 ± 0.14

F TOXICITY ENDPOINT DESCRIPTIONS

Table G: Dataset names, endpoints, and their corresponding descriptions used in MolDeTox.

Dataset Name	Endpoint	Description
hERG Unified	hERG channel blocking activity	The molecule has been evaluated for hERG (human Ether-à-go-go-Related Gene) channel blocking activity under a unified hERG toxicity endpoint that combines data from the hERG, hERG inhibition, and hERG Karim sources. Blockade of the hERG channel can disrupt cardiac repolarization and lead to serious adverse effects, including cardiac arrhythmias and sudden cardiac death.
AMES	Mutagenicity	The molecule is mutagenic, meaning it can cause genetic alterations and DNA damage that may lead to cell death or severe adverse effects.
ClinTox	Clinical toxicity	The molecule has been associated with clinical toxicity, including drugs that have failed clinical trials due to toxicity reasons.
DILIst	Drug-induced liver injury	The molecule is highly likely to cause human liver injury, or actual cases of liver injury have been reported and confirmed.
DICTrank	Cardiotoxicity	The molecule is highly likely to cause human cardiotoxicity, or actual cases of cardiotoxicity have been reported and confirmed.
DIRIL	Renal toxicity	The molecule is highly likely to cause human renal toxicity, or actual cases of renal toxicity have been reported and confirmed.
Skin Reaction	Skin sensitization	The molecule can cause skin sensitization, an immune reaction that leads to allergic contact dermatitis upon repeated exposure.
Tox21	NR-AR	The molecule activates or disrupts the Androgen Receptor (AR) pathway, which regulates male sexual development and function. Disruption of this pathway can affect reproductive development and function.
Tox21	NR-AR-LBD	The molecule binds to the Androgen Receptor Ligand Binding Domain (AR-LBD), affecting androgen signaling pathways. This assay evaluates more direct binding mechanisms compared to the full receptor activity assay.
Tox21	NR-AhR	The molecule activates the Aryl Hydrocarbon Receptor (AhR) pathway, which is involved in xenobiotic metabolism and immune responses. Activation of this receptor can lead to toxic effects such as liver toxicity, carcinogenicity, and immunotoxicity.
Tox21	NR-Aromatase	The molecule inhibits or activates Aromatase, an enzyme essential for estrogen (female hormone) biosynthesis. This assay evaluates whether the chemical can affect aromatase enzyme activity, thereby influencing estrogen levels. Disruption of estrogen balance is important for reproductive health.
Tox21	NR-ER	The molecule activates or disrupts the Estrogen Receptor (ER) pathway, which regulates female sexual development and function. Disruption of this pathway can affect female reproductive development and function, and is associated with conditions such as breast cancer.
Tox21	NR-ER-LBD	The molecule binds to the Estrogen Receptor Ligand Binding Domain (ER-LBD), affecting estrogen signaling pathways. This assay evaluates more direct binding mechanisms compared to the full receptor activity assay.

Continued on next page

Table G continued from previous page

Dataset Name	Endpoint	Description
Tox21	NR-PPAR-gamma	The molecule activates or disrupts the Peroxisome Proliferator-Activated Receptor gamma (PPARγ) pathway, which regulates glucose and lipid metabolism, cell differentiation, and inflammatory responses. This assay evaluates whether the chemical can activate PPAR-gamma, potentially affecting metabolic diseases such as diabetes and obesity.
Tox21	SR-ARE	The molecule activates the Antioxidant Response Element (ARE) pathway, which regulates cellular antioxidant defense mechanisms. This assay evaluates whether the chemical can activate or inhibit the cell's antioxidant defense system in response to oxidative stress.
Tox21	SR-ATAD5	The molecule affects ATAD5 (ATPase family AAA domain-containing protein 5), which plays an important role in DNA damage response and repair. This assay evaluates whether the chemical can affect the ATAD5 pathway, potentially causing DNA damage and genomic instability issues.
Tox21	SR-HSE	The molecule activates the Heat Shock Response Element (HSE) pathway, which responds to cellular stress and protein misfolding. Cells respond to protein denaturation stress (heat, toxic substances, etc.) by inducing the production of heat shock proteins (HSP) to repair damaged proteins. This assay evaluates whether the chemical disrupts the cell's protein quality control system.
Tox21	SR-MMP	The molecule affects Mitochondrial Membrane Potential (MMP), which is an important indicator of mitochondrial functional status. Mitochondria are the cell's energy production factories. This assay evaluates whether the chemical can damage mitochondrial function, potentially causing problems with cellular energy production, which is one of the important mechanisms of cell toxicity.
Tox21	SR-p53	The molecule activates or disrupts the p53 pathway, a critical tumor suppressor pathway involved in cell cycle control and apoptosis. p53 is known as the "guardian of the genome" and responds to DNA damage and cellular stress by inducing cell cycle arrest, DNA repair, and apoptosis (cell death). This assay evaluates whether the chemical can affect the p53 pathway, potentially causing DNA damage, cell death, or cancer development.
Metabolism	CYP1A2_Veith	The molecule inhibits CYP P450 1A2 (Veith et al.). The CYP P450 genes are involved in the formation and breakdown (metabolism) of various molecules and chemicals within cells. Specifically, CYP1A2 localizes to the endoplasmic reticulum and its expression can be induced by polycyclic aromatic hydrocarbons (PAHs), some of which are found in cigarette smoke. It can metabolize PAHs to carcinogenic intermediates and also processes xenobiotics such as caffeine, aflatoxin B1, and acetaminophen. Inhibition can reduce drug metabolism and increase drug-drug interaction risk.
Metabolism	CYP2C19_Veith	The molecule inhibits CYP P450 2C19 (Veith et al.). The CYP P450 genes are essential for the breakdown (metabolism) of various molecules and chemicals within cells. Inhibiting these enzymes can lead to poor metabolism of this drug and co-administered drugs, increasing the risk of drug-drug interactions and adverse effects. CYP2C19 is associated with endoplasmic reticulum functions related to protein processing and transport.

Continued on next page

Table G continued from previous page

Dataset Name	Endpoint	Description
Metabolism	CYP2C9_Veith	The molecule inhibits CYP P450 2C9 (Veith et al.). The CYP P450 genes are involved in the formation and breakdown (metabolism) of various molecules and chemicals within cells. Specifically, CYP2C9 plays a major role in oxidation of both xenobiotic and endogenous compounds. Inhibition can impair metabolic clearance and increase adverse event risk.
Metabolism	CYP2D6_Veith	The molecule inhibits CYP P450 2D6 (Veith et al.). The CYP P450 genes are involved in the formation and breakdown (metabolism) of various molecules and chemicals within cells. CYP2D6 is primarily expressed in the liver and is also highly expressed in regions of the central nervous system, including the substantia nigra. Inhibition can alter metabolic clearance and increase potential toxicity or interaction risk.
Metabolism	CYP3A4_Veith	The molecule inhibits CYP P450 3A4 (Veith et al.). The CYP P450 genes are involved in the formation and breakdown (metabolism) of various molecules and chemicals within cells. CYP3A4 is an important enzyme mainly found in the liver and intestine, and oxidizes many foreign organic molecules (xenobiotics), including toxins and drugs, to support elimination. Inhibition can reduce clearance and increase drug-drug interaction risk.
SIDER	Blood and lymphatic system disorders	The molecule has been associated with blood and lymphatic system disorders, which can include conditions affecting blood cells, clotting mechanisms, or lymphatic circulation. These disorders may manifest as anemia, bleeding disorders, or immune system complications.
SIDER	Cardiac disorders	The molecule has been associated with cardiac disorders, which can include arrhythmias, heart failure, myocardial infarction, or other cardiovascular complications. These conditions can significantly impact heart function and overall cardiovascular health.
SIDER	Congenital, familial and genetic disorders	The molecule has been associated with congenital, familial, and genetic disorders, which may involve birth defects, inherited conditions, or genetic mutations. These disorders can affect development, growth, or long-term health outcomes.
SIDER	Ear and labyrinth disorders	The molecule has been associated with ear and labyrinth disorders, which can include hearing loss, tinnitus, vertigo, or balance problems. These conditions can affect auditory function and spatial orientation.
SIDER	Eye disorders	The molecule has been associated with eye disorders, which can include vision impairment, retinal damage, cataracts, or other ocular complications. These conditions can significantly impact visual function and quality of life.
SIDER	General disorders and administration site conditions	The molecule has been associated with general disorders and administration site conditions, which can include injection site reactions, systemic reactions, or general malaise. These conditions may occur at the site of drug administration or manifest as systemic effects.
SIDER	Hepatobiliary disorders	The molecule has been associated with hepatobiliary disorders, which can include liver damage, hepatitis, cholestasis, or other liver and bile duct complications. These conditions can significantly impact liver function and metabolic processes.

Continued on next page

Table G continued from previous page

Dataset Name	Endpoint	Description
SIDER	Immune system disorders	The molecule has been associated with immune system disorders, which can include autoimmune reactions, hypersensitivity, immunosuppression, or other immune-related complications. These conditions can affect the body's ability to fight infections or maintain immune homeostasis.
SIDER	Infections and infestations	The molecule has been associated with infections and infestations, which may indicate increased susceptibility to infections or direct infectious complications. These conditions can result from immunosuppression or other mechanisms that compromise immune defenses.
SIDER	Injury, poisoning and procedural complications	The molecule has been associated with injury, poisoning, and procedural complications, which can include accidental overdoses, drug interactions, or complications from medical procedures. These conditions may result from improper use, dosage errors, or adverse interactions.
SIDER	Investigations	The molecule has been associated with abnormal laboratory findings or investigations, which can include changes in blood chemistry, liver enzymes, kidney function markers, or other diagnostic parameters. These findings may indicate underlying organ dysfunction or metabolic disturbances.
SIDER	Metabolism and nutrition disorders	The molecule has been associated with metabolism and nutrition disorders, which can include diabetes, electrolyte imbalances, metabolic syndrome, or nutritional deficiencies. These conditions can affect energy metabolism, glucose regulation, or nutrient absorption.
SIDER	Musculoskeletal and connective tissue disorders	The molecule has been associated with musculoskeletal and connective tissue disorders, which can include muscle weakness, joint pain, bone disorders, or connective tissue damage. These conditions can affect mobility, strength, and structural integrity of the musculoskeletal system.
SIDER	Neoplasms benign, malignant and unspecified (incl cysts and polyps)	The molecule has been associated with neoplasms (tumors), including benign, malignant, and unspecified growths, as well as cysts and polyps. These conditions involve abnormal cell growth and may indicate carcinogenic potential or tumor-promoting effects.
SIDER	Nervous system disorders	The molecule has been associated with nervous system disorders, which can include neurotoxicity, seizures, cognitive impairment, or other neurological complications. These conditions can affect brain function, peripheral nerves, or overall neurological health.
SIDER	Pregnancy, puerperium and perinatal conditions	The molecule has been associated with pregnancy, puerperium, and perinatal conditions, which can include complications during pregnancy, childbirth, or the postpartum period. These conditions can affect maternal health, fetal development, or neonatal outcomes.
SIDER	Product issues	The molecule has been associated with product issues, which can include quality problems, contamination, or manufacturing defects. These issues may affect drug safety, efficacy, or stability.
SIDER	Psychiatric disorders	The molecule has been associated with psychiatric disorders, which can include depression, anxiety, psychosis, mood changes, or other mental health complications. These conditions can significantly impact cognitive function, emotional well-being, and behavioral patterns.

Continued on next page

Table G continued from previous page

Dataset Name	Endpoint	Description
SIDER	Renal and urinary disorders	The molecule has been associated with renal and urinary disorders, which can include kidney damage, renal failure, urinary tract complications, or other nephrotoxic effects. These conditions can significantly impact kidney function and fluid-electrolyte balance.
SIDER	Reproductive system and breast disorders	The molecule has been associated with reproductive system and breast disorders, which can include hormonal imbalances, fertility issues, reproductive organ complications, or breast-related conditions. These conditions can affect reproductive health, fertility, or hormonal regulation.
SIDER	Respiratory, thoracic and mediastinal disorders	The molecule has been associated with respiratory, thoracic, and mediastinal disorders, which can include breathing difficulties, lung damage, respiratory infections, or other pulmonary complications. These conditions can significantly impact respiratory function and oxygen exchange.
SIDER	Skin and subcutaneous tissue disorders	The molecule has been associated with skin and subcutaneous tissue disorders, which can include rashes, dermatitis, skin irritation, or other dermatological complications. These conditions can affect skin integrity, appearance, or protective function.
SIDER	Social circumstances	The molecule has been associated with social circumstances, which may indicate impacts on social functioning, relationships, or daily activities. These effects may result from physical or psychological side effects that affect quality of life.
SIDER	Surgical and medical procedures	The molecule has been associated with complications from surgical and medical procedures, which can include adverse reactions during or after medical interventions. These complications may result from drug interactions, procedural risks, or patient-specific factors.
SIDER	Vascular disorders	The molecule has been associated with vascular disorders, which can include blood vessel damage, thrombosis, hypertension, or other circulatory complications. These conditions can affect blood flow, vascular integrity, or cardiovascular function.

G MolDeTox Prompts

Task 1 SYSTEM PROMPT

You are a molecular toxicity reasoning assistant specialized in SAFE and SMILES representations.

Given:
- A toxic molecule
- Its molecular representation in SAFE and/or SMILES format
- The toxicity endpoint context, when provided

your job is to identify the fragment(s) in the toxic molecule that are most likely associated with toxicity.

Follow these rules carefully:

1. Focus on **toxicity-associated fragment identification**.
- Identify the fragment(s) that are specific to the toxic molecule and are most likely responsible for the toxicity signal.
- If multiple fragments are required, return all of them.
- Preserve the original SAFE fragment format exactly.

2. Output format constraints:
- Return the answer as the **toxic-only SAFE fragment string**.
- If there are multiple fragments, concatenate them as a dot-separated SAFE string.
- Do not paraphrase fragment content or convert it into natural language.

3. Response format:

```
{
  "answer": "..."
}
```

HARD CONSTRAINTS:
- Output ONLY the JSON object.
- Do not include explanations, markdown, or extra text.
- Do not add any extra keys.
- The value of "answer" must be the toxic-only SAFE fragment string exactly.

Table H: Task 1 system prompt for toxic fragment identification.

Task 2 **SYSTEM PROMPT**

You are a molecular toxicity reasoning assistant specialized in SAFE and SMILES representations.

Given:
- A toxic molecule
- Its molecular representation in SAFE and/or SMILES format
- The toxic-only fragment(s) identified from the molecule
- The toxicity endpoint context, when provided

your job is to generate the non-toxic replacement fragment(s) corresponding to the toxic fragment(s).

Follow these rules carefully:

1. Focus on **non-toxic fragment generation**.
- Generate the fragment(s) that can replace the toxic fragment(s) while reducing toxicity.
- If multiple fragments are required, return all of them.
- Preserve the original SAFE fragment format exactly.

2. Output format constraints:
- Return the answer as the **non-toxic-only SAFE fragment string**.
- If there are multiple fragments, concatenate them as a dot-separated SAFE string.
- Do not paraphrase fragment content or convert it into natural language.

3. Response format:

```
{
  "answer": "..."
}
```

HARD CONSTRAINTS:
- Output ONLY the JSON object.
- Do not include explanations, markdown, or extra text.
- Do not add any extra keys.
- The value of "answer" must be the non-toxic-only SAFE fragment string exactly.

Table I: Task 2 system prompt for non-toxic fragment generation.

Task 3 SMILES Generation **SYSTEM PROMPT**

You are a molecular toxicity reasoning assistant specialized in SAFE and SMILES representations.

Given:
- A toxic molecule
- Its molecular representation in SAFE and/or SMILES format
- The toxicity endpoint context, when provided

your job is to generate the final non-toxic molecule as a single SMILES string.

Follow these rules carefully:

1. Focus on **non-toxic molecule generation**.
- Generate a chemically plausible non-toxic molecule.
- Reduce toxicity while preserving the original molecular characteristics as much as possible.
- Return the final molecule, not intermediate fragments.

2. Output format constraints:
- Return the answer as a **single non-toxic molecule SMILES string**.
- Do not return SAFE fragments.
- Do not return multiple candidates.

3. Response format:

```
{
  "answer": "..."
}
```

HARD CONSTRAINTS:
- Output ONLY the JSON object.
- Do not include explanations, markdown, or extra text.
- Do not add any extra keys.
- The value of "answer" must be the final non-toxic molecule SMILES string.

Table J: Task 3 system prompt for direct non-toxic molecule generation.

Task 3 SAFE Generation **SYSTEM PROMPT**

You are a molecular toxicity reasoning assistant specialized in SAFE and SMILES representations.

Given:
- A toxic molecule
- Its molecular representation in SAFE and/or SMILES format
- The toxicity endpoint context, when provided

your job is to generate the final non-toxic molecule in SAFE representation.

Follow these rules carefully:

1. Focus on **non-toxic SAFE generation**.
- Generate the full SAFE representation of the resulting non-toxic molecule.
- Reduce toxicity while preserving the original molecular characteristics as much as possible.
- Return the complete molecule-level SAFE representation, not only edited fragments.

2. Output format constraints:
- Return the answer as the **full non-toxic SAFE string**.
- If multiple fragments are present, concatenate them as a dot-separated SAFE string.
- Do not paraphrase fragment content or convert it into natural language.

3. Response format:

```
{
  "answer": "..."
}
```

HARD CONSTRAINTS:
- Output ONLY the JSON object.
- Do not include explanations, markdown, or extra text.
- Do not add any extra keys.
- The value of "answer" must be the final full non-toxic SAFE string for the whole molecule.

Table K: Task 3 system prompt for full non-toxic SAFE generation.

Task 3 Step-wise CoT SAFE Generation **SYSTEM PROMPT**

You are a molecular toxicity reasoning assistant specialized in SAFE and SMILES representations.

Given:
- A toxic molecule
- Its molecular representation in SAFE and/or SMILES format
- The toxicity endpoint context, when provided

your job is to solve the task through explicit intermediate reasoning steps and generate the final non-toxic molecule in SAFE representation.

Follow these rules carefully:

1. Use **step-wise reasoning**.
- First identify the toxic fragment(s) most likely associated with toxicity.
- Then generate the corresponding non-toxic replacement fragment(s).
- Finally generate the full non-toxic molecule as a SAFE string.

2. Output format constraints:
- Return a single JSON object.
- The JSON must include the final answer and all required intermediate fields.
- Do not omit any required fields.

3. Response format:

```
{
  "answer": "...",
  "step1_only_toxic_safe_fragments": "...",
  "step1_reasoning": "...",
  "step2_only_nontoxic_safe_fragments": "...",
  "step2_reasoning": "...",
  "step3_reasoning": "..."
}
```

HARD CONSTRAINTS:
- Output ONLY the JSON object.
- Do not include markdown or extra text outside the JSON.
- The value of "answer" must be the final full non-toxic SAFE string for the whole molecule.
- All step fields must be included exactly as required.

Table L: Task 3 step-wise chain-of-thought system prompt for full non-toxic SAFE generation.

SAFE Representation EXPLANATION

SAFE (Sequential Attachment-based Fragment Embedding) is a SMILES-compatible string representation that expresses a molecule as a dot-separated sequence of fragments.

How SAFE is constructed:
- **Fragmentation**: A molecule is split into fragments by cutting selected bonds using a slicing algorithm.
- **Slicer**: The default slicer is `brics`, a rule-based method that cuts retrosynthetically relevant bonds to produce chemically meaningful substructures.
- **Attachment Markers**: At each cut site, attachment information is encoded with SMILES-style ring-closure digits (e.g., `1`, `2`, ..., `%10`). Matching digits across fragments indicate where fragments reconnect in the full molecule.
- **Serialization**: The resulting fragments are written as SMILES strings and joined with `.` separators to form a SAFE string.

Important characteristics:
- **Fragment-based representation**: Each token block corresponds to a substructure rather than the entire molecule.
- **Order invariance**: Changing the fragment order does not change the reconstructed molecule.
- **Partial structures**: Individual fragments may look chemically incomplete on their own because they are parts of a larger graph.

Table M: Explanation of the SAFE representation used in MolDeTox.

Common Pair Context PROMPT

Context:
The toxic and non-toxic molecules in this task form a paired example. These paired molecules are structurally very similar and have minimal physicochemical differences, but they differ in toxicity versus non-toxicity for the same endpoint.

Follow this principle carefully:
- Assume that toxicity differences arise from localized structural differences rather than a complete molecular redesign.
- Use this paired setting when identifying toxicity-associated fragments, proposing non-toxic replacement fragments, or generating the final non-toxic molecule.
- Preserve as much of the original molecular structure and characteristics as possible when reasoning about detoxification.

Table N: Common pair-context prompt used across toxic/non-toxic paired tasks in MolDeTox.

Property Preservation PROMPT

Instruction:
When modifying a toxic molecule to make it non-toxic, preserve its original physicochemical and pharmacological properties as much as possible. The goal is to reduce or remove toxicity only for the given endpoint, rather than to redesign the entire molecule or alter unrelated molecular characteristics.

Table O: Property-preservation prompt used in Tasks 2 and 3 of MolDeTox.

Task 1 **QUESTION**

{endpoint_description}
{safe_explanation}
{pair_context}
{preserve_property}

- Toxic molecule (SMILES representation): [toxic_safe_decoded_smiles]
- Toxic molecule (SAFE representation): [toxic_safe]

Task: This toxic molecule belongs to a structurally similar pair that differs only in toxicity for this endpoint. Identify the fragment(s) that are candidates for toxicity-associated structure (the part(s) that drive toxicity for this endpoint) and output them as only_toxic_safe_fragments (dot-separated if multiple).

Output format: a single JSON object with key "answer" and value the only_toxic_safe_fragments string (dot-separated for multiple fragments). Example: {"answer": "frag1.frag2"}

Table P: Task 1 question template for toxic fragment identification. Common context components are abbreviated as placeholders and described separately in the paper.

Task 2 **QUESTION**

{endpoint_description}
{safe_explanation}
{pair_context}
{preserve_property}

- Toxic molecule (SMILES representation): {toxic_safe_decoded_smiles}
- Toxic molecule (SAFE representation): {toxic_safe}

- The fragments that appear only in the toxic molecule (candidates for toxicity-associated structure for this endpoint) are: {only_toxic_safe_fragments}

Task: Output the only_nontoxic_safe_fragments—i.e. the SAFE fragment(s) that, when used in place of the only_toxic_safe_fragments, yield a non-toxic molecule for this endpoint. When modifying the toxic molecule to make it non-toxic, do not change other physicochemical or pharmacological properties; only reduce or remove the drug toxicity for this endpoint.

Output format: a single JSON object with key "answer" and value the only_nontoxic_safe_fragments string (dot-separated for multiple fragments). Example: {"answer": "frag1.frag2"}

Table Q: Task 2 question template for non-toxic fragment generation. Common context components are abbreviated as placeholders and described separately in the paper.

Task 3 SMILES Generation **QUESTION**

{endpoint_description}
{safe_explanation}
{pair_context}
{preserve_property}

- Toxic molecule (SMILES representation): {toxic_safe_decoded_smiles}
- Toxic molecule (SAFE representation): {toxic_safe}

Task: From the toxic molecule above, identify the fragment(s) that are candidates for toxicity-associated structure for this endpoint, then determine the replacement fragment(s) that yield a non-toxic molecule. Output the resulting non-toxic molecule as a single SMILES string (non-toxic_safe_decoded_smiles). When modifying the toxic molecule to make it non-toxic, do not change other physicochemical or pharmacological properties; only reduce or remove the drug toxicity for this endpoint.

Output format: a single JSON object with key "answer" and value the nontoxic_safe_decoded_smiles string. Example: {"answer": "CCO"}

Table R: Task 3 question template for direct non-toxic SMILES generation. Common context components are abbreviated as placeholders and described separately in the paper.

Task 3 SAFE Generation **QUESTION**

{endpoint_description}
{safe_explanation}
{pair_context}
{preserve_property}

- Toxic molecule (SMILES representation): {toxic_safe_decoded_smiles}
- Toxic molecule (SAFE representation): {toxic_safe}

Task: From the toxic molecule above, identify the fragment(s) that are candidates for toxicity-associated structure for this endpoint, then determine the replacement fragment(s) that yield a non-toxic molecule. Output the resulting non-toxic molecule as a single SAFE string (nontoxic_safe). When modifying the toxic molecule to make it non-toxic, do not change other physicochemical or pharmacological properties; only reduce or remove the drug toxicity for this endpoint.

Output format: a single JSON object with key "answer" and value the resulting non-toxic SAFE string. Example: {"answer": "CCO.[*:1]"}

Table S: Task 3 question template for end-to-end non-toxic SAFE generation. Common context components are abbreviated as placeholders and described separately in the paper.

Task 3 Step-wise CoT SAFE Generation **QUESTION**

{endpoint_description}
{safe_explanation}
{pair_context}
{preserve_property}

- Toxic molecule (SMILES representation): {toxic_safe_decoded_smiles}
- Toxic molecule (SAFE representation): {toxic_safe}

Task: Solve the following in ONE call, step by step, using natural-language reasoning.

Step 1 (endpoint-aware toxic fragment identification):
- Identify the fragment(s) most likely responsible for toxicity for this endpoint (dot-separated if multiple).
- In step1_reasoning, identify which fragment is most likely responsible for toxicity for this endpoint and explain *why* the fragment(s) are toxicity-associated for this endpoint, using brief chemical intuition (no need for citations).
- Output the fragment string as step1_only_toxic_safe_fragments.

Step 2 (endpoint-aware non-toxic fragment proposal):
- Using the Step 1 fragment as the part to be replaced, propose non-toxic replacement fragment(s) (dot-separated if multiple) that reduce toxicity for this endpoint while keeping the overall scaffold as similar as possible.
- In step2_reasoning, explain the design intent: what property/alert you are trying to reduce for this endpoint and what you preserve while keeping the overall scaffold as similar as possible.
- Output the fragment string as step2_only_nontoxic_safe_fragments.

Step 3 (construct final non-toxic SAFE):
- Combine Step 1 and Step 2: conceptually remove the toxic fragment and add the proposed non-toxic fragment that reduces toxicity for this endpoint while keeping the overall scaffold as similar as possible.
- In step3_reasoning, describe at a high level how the final molecule changes relative to the toxic molecule.
- Output the final non-toxic molecule as a single full SAFE string under the key "answer".

Important:
- When modifying the toxic molecule to make it non-toxic, do not change other physicochemical or pharmacological properties; only reduce or remove the drug toxicity for this endpoint.
- Your output must be a SINGLE JSON object.
- Do not output any text outside the JSON.
- The fragment fields must be SAFE fragment strings (dot-separated if multiple).

Output format: a single JSON object with the following keys:
- "step1_only_toxic_safe_fragments": string (dot-separated SAFE fragment(s))
- "step1_reasoning": string
- "step2_only_nontoxic_safe_fragments": string (dot-separated SAFE fragment(s))
- "step2_reasoning": string
- "step3_reasoning": string
- "answer": string (the final nontoxic full SAFE string)

Example:

```
{"step1_only_toxic_safe_fragments":"frag1.frag2",
"step1_reasoning":"...",
"step2_only_nontoxic_safe_fragments":"fragA.fragB",
"step2_reasoning":"...",
"step3_reasoning":"...",
"answer":"CCO.[*:1]"}
```

Table T: Task 3 step-wise CoT question template for end-to-end non-toxic SAFE generation. Common context components are abbreviated as placeholders and described separately in the paper.