TOWARD MEDICINAL CHEMISTRY-AWARE MOLECULAR GENERATION THROUGH LLM GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular optimization and evaluation in AI-driven drug discovery relies primarily on established molecular metrics and heuristics such as QED, Lipinski's Rule-of-Five, structure filters, and synthetic accessibility scores. While effective at assessing specific aspects of molecular quality, these metrics cannot fully capture the holistic reasoning that medicinal chemists apply when prioritizing candidate molecules. In this work, we propose a new evaluation paradigm in which a large language model acts as a medicinal chemistry judge, assigning each molecule an overall quality score while explaining the medicinal chemistry considerations underlying its assessment. To systematically evaluate this approach, we introduce MedChemJudge, the first benchmark specifically designed for LLM-based medicinal chemistry assessment, comprising over 500 expert-annotated molecules across three complementary evaluation settings. Using this benchmark, we identify an effective judge configuration achieving a C-index of 0.84 in reproducing expert quality rankings. The analysis shows that LLM-derived scores can distinguish between structurally similar molecules even when conventional molecular metrics are nearly identical. Finally, incorporating the proposed score into RL–based molecular generative model improves the medicinal chemistry quality of generated molecules across 5 targets while maintaining strong docking performance and high diversity. Together, these results highlight the potential of LLM-based evaluation as a complementary assessment layer for AI-assisted drug design and support molecular optimization that more closely reflects medicinal chemistry decision-making.

1 INTRODUCTION

Artificial intelligence (AI) has become an integral component of modern drug discovery, with generative modeling enabling the rapid exploration of chemical space and the proposal of large numbers of candidate molecules *in silico* (Walters & Murcko, 2020a; Vamathevan et al., 2019; Schneider, 2018). As generative models continue to improve, the evaluation of generated molecules has become an increasingly important stage of the molecular design pipeline (Olivecrona et al., 2017). Rigorous evaluation is essential for distinguishing genuinely promising candidates from molecules that benefit from biases or limitations in the generative model. Such evaluation increases the likelihood that computationally generated compounds will demonstrate meaningful performance in subsequent experimental validation (Jin et al., 2018).

Current molecular evaluation workflows employ a wide range of computational metrics and rule-based filters, including commonly used measures such as Lipinski's Rule of Five (Lipinski, 2000), Quantitative Estimate of Drug-likeness (QED) (Bickerton et al., 2012), Synthetic Accessibility Score (SAScore) (Ertl & Schuffenhauer, 2009), Brenk (Brenk et al., 2008) structural alerts, and PAINS filters (Baell & Holloway, 2010). These metrics provide efficient and well-established approximations of specific aspects of molecular quality and have therefore become standard components of molecular optimization pipelines (Zavadskaya et al., 2025; Averly et al., 2025; Urazmanova et al., 2025; Solovev et al., 2025). However, these metrics are inherently limited to predefined physicochemical properties or structural features and therefore do not capture the broader set of considerations that inform medicinal chemistry decision-making.

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.

In practical drug discovery, medicinal chemists jointly evaluate structural liabilities, synthetic tractability, developability, known chemical motifs, safety concerns, and accumulated domain knowledge when prioritizing candidate molecules (Sathyanarayana et al., 2025; Ouyang et al., 2024). Consequently, compounds that satisfy conventional computational criteria may nevertheless receive an unfavorable overall assessment during expert review. This divergence does not necessarily reflect shortcomings in existing computational metrics, but rather differences in their intended scope: such metrics are typically designed to quantify specific molecular characteristics, whereas expert assessment integrates multiple, interdependent factors into a broader evaluation of a candidate's overall suitability. As a result, candidates that fail to meet expert expectations may progress through multiple stages of computational generation and optimization before being deprioritized, increasing the number of design and evaluation cycles and reducing the overall efficiency of the molecular design process.

Recent advances in large language models (LLMs) have demonstrated their growing capabilities in chemistry, including the interpretation of molecular representations, integration of chemical knowledge, and reasoning over diverse molecular properties (White, 2023; Mirza et al., 2025; Boiko et al., 2023; Dubrovsky et al., 2024). Alongside these developments, a broader shift toward using LLMs as evaluators has emerged in natural language processing through the LLM-as-a-Judge paradigm, in which LLMs assess model-generated outputs and provide both quantitative scores and qualitative feedback (Liu et al., 2023a; Gu et al., 2025; Li et al., 2024a; Chen et al., 2026; D'Souza et al., 2025). Building on this paradigm, we investigate whether LLMs can provide holistic assessments of candidate molecules by integrating the multifactorial considerations involved in medicinal chemistry evaluation. Rather than replacing established computational metrics, we study whether an LLM can serve as an additional evaluator that provides an overall medicinal chemistry assessment. In this role, the LLM is intended to approximate the broader perspective of a medicinal chemist, drawing on chemical knowledge and drug-design principles to consider multiple aspects of a molecule jointly rather than evaluating predefined properties in isolation.

Our contributions are fourfold:

1. We propose a new paradigm for holistic molecular evaluation in AI-based drug design, where an LLM acts as a medicinal chemistry judge, producing both an interpretable rationale and a continuous quality score that complements traditional descriptors.

2. To evaluate this new paradigm, we introduce MedChemJudge, which, to the best of our knowledge, is the first benchmark dedicated to systematically assessing the molecular evaluation capabilities of LLMs. MedChemJudge comprises 520 expert-annotated molecules and covers three complementary evaluation settings: expert-aligned scoring of individual molecules, robustness across alternative molecular representations, and discrimination between structurally related compounds.

3. Using MedChemJudge, we systematically compare molecular representations, prompt formulations, and multiple families of LLMs for medicinal chemistry evaluation. We identify Qwen-3-14B with SMILES input and a balanced prompt as the strongest-performing configuration achieving a concordance index (C-index) of 0.84 with medicinal chemist preferences, indicating strong agreement with expert molecular rankings. At the same time, analysis of its predictions on structurally related molecular pairs highlights both chemically meaningful judgments and cases of disagreement with expert preferences.

4. Finally, we demonstrate the practical utility of the proposed evaluation by incorporating it as a reward function into FREED++, an RL-based molecular generative model. Across five protein targets, LLM-guided generation improves molecular properties that were not explicitly optimized during generation, while retaining favorable docking performance and high diversity. Importantly, expert evaluation shows that molecules generated with LLM-guided rewards are generally rated more favorably by the medicinal chemistry expert, demonstrating the potential of LLM-based evaluation to guide generative models toward more practically relevant candidates. The underlying code for this study as well as the MedChemJudge benchmark are publicly available at https://anonymous.4open.science/r/MedChemJudge-884A.

2 Related Work

Generative models for drug design. Deep generative models have become a widely used approach for exploring chemical space and proposing candidate molecules for drug discovery (Walters & Murcko, 2020b; Lin et al., 2023; Gómez-Bombarelli et al., 2018; Kwon & Yoon, 2020; Gómez-Bombarelli et al., 2018; Kusner et al., 2017; Guimaraes et al., 2017; De Cao & Kipf, 2018; Bagal et al., 2021; Guan et al., 2023). Alongside these developments, RL has become a widely used framework for goal-directed molecular generation, enabling generative models to be optimized toward predefined molecular objectives (Zhou et al., 2019; van Tilborg et al., 2023; Olivecrona et al., 2017; Loeffler et al., 2024; Jensen, 2019; Jiang et al., 2021; Tang et al., 2019; Šícho et al., 2023). Among RL-based approaches, FREED (Fu et al., 2022) and its successor FREED++ (Telepov et al., 2023) are particularly relevant to our setting. FREED++ directly optimizes physics-based docking scores, allowing target-specific binding objectives to be incorporated without relying on learned affinity predictors. Its modular reward formulation further enables additional molecular evaluation signals to be integrated alongside the primary optimization objective.

Computational metrics for molecular quality assessment. The effectiveness of AI-based goal-directed molecular generation depends critically on the quality of the optimization criteria. For target-specific molecular design, docking scores are commonly used to estimate binding to a protein target (Alhossary et al., 2015; Eberhardt et al., 2021). Docking objectives are often complemented by additional computational descriptors and rule-based filters. Physicochemical descriptors, including molecular weight, logP, the number of heavy atoms, and hydrogen-bond donors and acceptors, characterize properties such as molecular size, lipophilicity, and polarity. Lipinski's Rule of Five (Lipinski, 2000) and QED (Bickerton et al., 2012) provide broader measures of drug-likeness, while SAScore (Ertl & Schuffenhauer, 2009) and RAscore (Thakkar et al., 2021) estimate synthetic accessibility. Structural filters, including PAINS (Baell & Holloway, 2010), Brenk (Brenk et al., 2008), and Glaxo (Lane et al., 2006), identify potentially undesirable molecular motifs. While these metrics provide efficient and well-established measures of specific aspects of molecular quality, even their combination cannot fully capture the breadth of context-dependent considerations and accumulated chemical knowledge that medicinal chemists integrate when evaluating candidate molecules.

LLMs for chemistry and molecular design. LLMs have demonstrated growing capabilities across chemistry-related tasks. Domain-specialized models, including ChemDFM (Liu et al., 2023b), ChemCrow (Bran et al., 2024), ChemLLM (Zhang et al., 2024), and Mol-Inst (Zhang et al., 2024), support molecular understanding, property prediction, synthesis planning, and chemical reasoning. ChemLLMBench (Guo et al., 2023) evaluates LLMs on chemistry tasks including molecular property prediction, while FGBench (Liu et al., 2026) focuses on fine-grained molecular property reasoning and direct molecular comparisons. LLMs have also been applied to molecular design and optimization. Speak-to-Structure (Li et al., 2026) uses natural-language instructions for molecular editing, while MolAct (Yang et al., 2025) and SEISMO (Kruger et al., 2026) use rewards and iterative feedback to guide optimization. Building on this literature, our work focuses specifically on expert-aligned holistic medicinal chemistry scoring and its downstream use as an evaluation signal for molecular generation.

LLM-as-a-Judge approach. The LLM-as-a-Judge paradigm has emerged as a widely used approach for evaluating generative model outputs, where LLMs assign scores and provide textual rationales based on specified evaluation criteria (Liu et al., 2023a; Gu et al., 2025; Li et al., 2024a; Zheng et al., 2023). Prior work has explored judge reliability, calibration, and domain adaptation (Wang et al., 2024; Li et al., 2024b), demonstrating the potential of LLMs for complex, multi-dimensional evaluation. However, LLM-based evaluation has primarily focused on language, code, and related reasoning tasks. To the best of our knowledge, it has not been systematically studied for medicinal chemistry assessment of candidate molecules.

3 LLM as a Medicinal Chemistry Judge Formulation

We begin by formally defining the expected output of the LLM judge before introducing the evaluation methodology. Given a molecule m represented as a textual or numerical input, the LLM judge

J_θ produces a pair

$$(s, r) = J_\theta(m), \tag{1}$$

where $s \in [0, 1]$ is a holistic medicinal chemistry score and r is a free-text rationale supporting the assessment. Higher values of s indicate a more favorable overall medicinal chemistry profile under the evaluation criteria specified in the prompt, whereas lower values indicate substantial structural, safety, synthetic, or developability concerns.

We explicitly ask the LLM to express its overall assessment as a numerical score so that it can be readily used as a molecular evaluation function in generative models. This makes the judge compatible with optimization pipelines that require a scalar objective or reward, without restricting it to a particular generative architecture. The rationale r serves a complementary purpose by making the numerical assessment interpretable and auditable. The judge targets general molecular suitability rather than context-specific properties such as target affinity, selectivity, or dose-dependent effects, which remain the domain of specialized tools such as docking and QSAR models. It is therefore intended to complement, rather than replace, specialized metrics or medicinal chemist assessment.

4 MEDCHEMJUDGE BENCHMARK

To systematically evaluate LLMs as medicinal chemistry judges, we start by introducing **MedChemJudge** benchmark. MedChemJudge evaluates an LLM's ability to perform holistic medicinal chemistry assessment by scoring candidate molecules and explaining its reasoning, with its judgments evaluated against assessments provided by medicinal chemistry experts.

Molecule collection and representation. The benchmark was constructed by medicinal chemistry expert using molecules collected from ChEMBL (Mendez et al., 2019), PubChem (Kim et al., 2023), and manually curated examples from the medicinal chemistry literature. Rather than focusing on a single chemical domain, we intentionally sampled molecules spanning a broad spectrum of medicinal chemistry quality, including approved drugs with favorable safety profiles, withdrawn drugs, compounds with known toxicity, and non-drug-like molecules. All molecular structures were standardized to canonical SMILES using RDKit (Landrum, 2024), IUPAC names were obtained using the PubChem API. For every molecule, we additionally precomputed nine widely used computational descriptors that serve as classical baselines throughout this work: QED, SAScore, RAscore, molecular weight (MW), LogP, Lipinski Rule-of-Five compliance, Brenk alerts, PAINS alerts, and Glaxo alerts. The benchmark therefore enables direct comparison between descriptor-based evaluation and LLM-based holistic assessment using identical molecular inputs.

Benchmark design. The benchmark evaluates complementary aspects of medicinal chemistry judgment through three independent evaluation settings, as illustrated in Figure 1:

1. **Expert-Aligned Scoring (300 molecules).** The LLM judge evaluates each molecule independently and assigns an overall medicinal chemistry quality score without access to other molecules or their annotations. Each molecule is assigned to one of three quality categories by a human expert (see details in Appendix A1.1):

- *Favorable*. Molecules with an overall favorable medicinal chemistry profile and no major liabilities that would substantially limit their suitability as drug-like compounds.
- *Borderline*. Molecules with an intermediate or ambiguous medicinal chemistry profile, exhibiting some favorable characteristics but also notable liabilities or concerns that prevent an unequivocally favorable assessment.
- *Unfavorable*. Molecules with an overall unfavorable medicinal chemistry profile, characterized by substantial liabilities such as pronounced toxicity, problematic structural features, or other properties that make them undesirable from a medicinal chemistry perspective.

The expert-assigned categories are then used as reference annotations to evaluate how well the LLM scores align with medicinal chemist preferences.

2. **SMILES Invariance (60 molecule pairs).** Each pair contains two different SMILES representations of the same molecular structure generated using RDKit. The judge evaluates both representations independently, and their scores are compared to determine whether the assessment remains consistent despite changes in SMILES syntax.

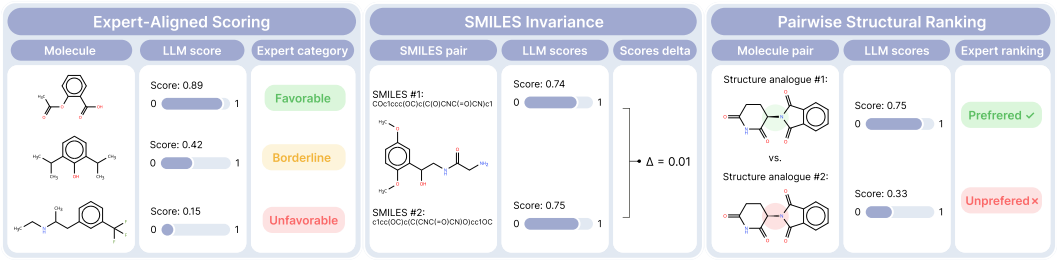


Figure 1: **Overview of the MedChemJudge benchmark.** The benchmark comprises three complementary evaluation settings: *Expert-Aligned Scoring*, *SMILES Invariance* and *Pairwise Structural Ranking*.

3. **Pairwise Structural Ranking (50 molecule pairs).** Each pair contains two structurally related molecules with a known difference in medicinal chemistry preference, including differences in safety, toxicity, pharmacological activity, or active metabolite-prodrug relationships (see Appendix A1.2 for details).

5 Judge Design

We treat judge construction as a design problem with three factors: input representation, prompt formulation, and model choice. We evaluate candidate configurations in two stages. First, we compare all configurations on the Expert-Aligned Scoring subset of MedChemJudge. We then evaluate the best-performing configurations on more complex SMILES Invariance and Pairwise Structural Ranking subsets to select the final judge configuration.

5.1 Design Aspects

Input Representation. The first design choice concerns how molecular information is presented to the LLM, as its assessment may depend on the input representation (see Figure 2A). We compare four representations: SMILES, IUPAC names, their combination, and SMILES augmented with a descriptor vector containing the nine conventional molecular metrics used throughout the benchmark.

Prompting Strategy. We next examine how the evaluation criteria specified in the prompt affect the LLM judgment. We compare three prompting strategies – baseline, strict, and balanced – that differ primarily in how molecular liabilities are treated. While all prompts request a numerical score and a concise chemical rationale, they range from a general assessment to stricter liability penalization and a more balanced treatment that distinguishes intrinsic chemical liabilities from potentially manageable pharmacological effects (see Appendix A3.1).

Model choice. Finally, we examine how judge performance varies across different LLMs. Our main experiments focus on three compact general-purpose models – GPT-4.1 Nano, Claude 3 Haiku, and Qwen3-14B – representing different model families. In additional experiments, we evaluate ChemDFM-v1.0-13B (Liu et al., 2023b), a chemistry-specialized language model pretrained on chemical data, to assess whether domain specialization provides an advantage for medicinal chemistry evaluation. We also consider larger LLMs of the same providers (GPT-5.5, Claude Opus 4.8, and Qwen3.6-35B) to determine whether scaling to more capable but more expensive models yields substantial improvements in judge performance.

5.2 Evaluation metrics

To evaluate the performance of the LLM judges against expert annotations on the Expert-Aligned Scoring subset, we report three complementary metrics:

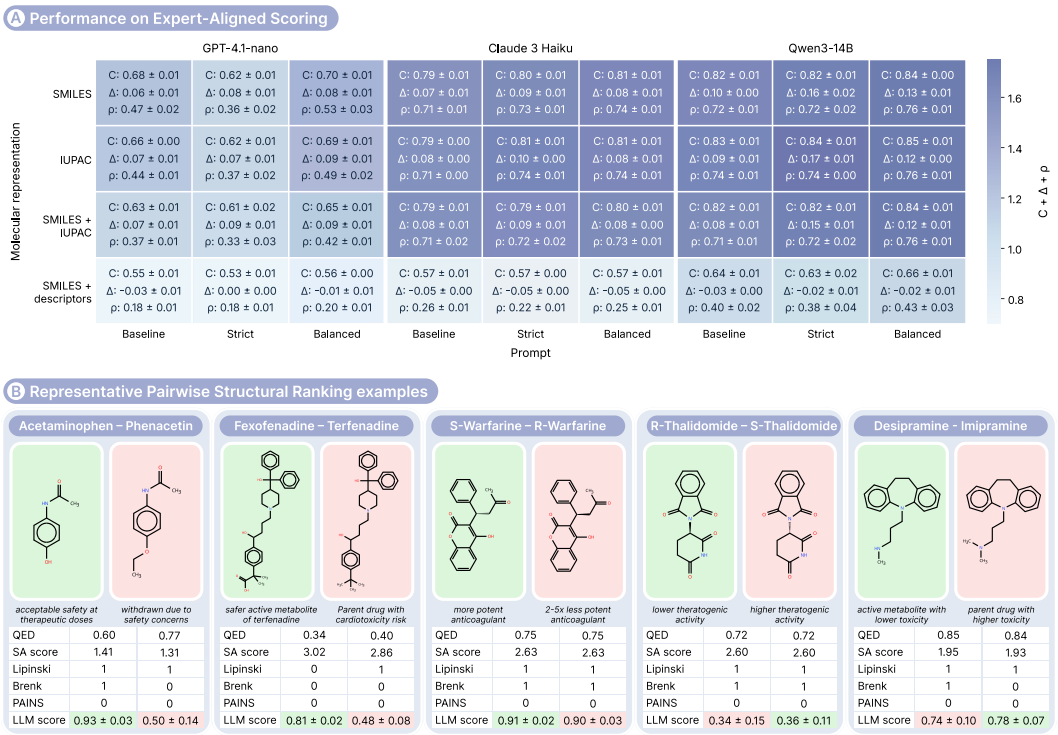


Figure 2: **Evaluation of LLM-based medicinal chemistry judges on MedChemJudge. (A)** Performance on Expert-Aligned Scoring across models, molecular representations, and prompts. Cells report C-index (C), minimum adjacent gap (Δ), and Spearman's correlation (ρ); color indicates $C + \Delta + \rho$. **(B)** Representative Pairwise Structural Ranking examples for Qwen3-14B with SMILES and the balanced prompt. The first two pairs are successful, the middle pair is borderline but correctly ranked, and the last two are failure cases. Backgrounds indicate expert preference for structures and LLM preference for scores.

- *Pairwise ordering accuracy (C-index)* measures how consistently the ranking induced by the LLM scores agrees with the ordering of the expert-defined quality classes. The C-index ranges from 0 to 1, with higher values indicating stronger agreement with expert preferences.
- *Minimum adjacent gap* ($\Delta_{\min}$) measures the separation between the three expert-defined quality classes by considering the smaller of the two differences between their mean LLM scores. Higher values indicate clearer separation across both adjacent class boundaries.
- *Spearman's rank correlation* (ρ) measures the monotonic association between the expert-defined class ordering and the LLM scores. The coefficient ranges from -1 to 1, with higher positive values indicating stronger correspondence between LLM scores and expert annotations.

For the SMILES Invariance subset, we report the Absolute within-pair delta between alternative SMILES representations of the same molecule, with lower values indicating greater consistency. For Pairwise Structural Ranking, we report directional accuracy, mean signed within-pair score difference, and tie rate. Full metric definitions are provided in Appendix A2. In addition, we provide an expert validation of LLM scores and text rationales to assess the chemical soundness of the evaluations (see Appendix A10).

5.3 Results

We performed a grid search over three prompt formulations, four molecular representations, and three primary models. Figure 2A summarizes performance on the Expert-Aligned Scoring subset,

using the sum of C-index (C), minimum adjacent gap (Δ), and Spearman's ρ as the criterion for comparing judge configurations. Qwen3-14B substantially outperformed GPT-4.1 Nano across molecular representations and prompting strategies, while its performance was close to, but generally slightly higher than, that of Claude 3 Haiku. Interestingly, augmenting the molecular representation with precomputed descriptors generally reduced performance, suggesting that conventional molecular descriptors do not improve the judge's alignment with expert assessments and may even interfere with its ability to evaluate molecular quality.

On the Expert-Aligned Scoring subset, prompt choice had only a minor effect within each model. We therefore retained Qwen3-14B with SMILES and IUPAC representations and evaluated all three prompting strategies on the SMILES Invariance and Pairwise Structural Ranking subsets (see Appendix A3.2 and A3.3). These evaluations supported the balanced prompt as the most robust overall choice across benchmark settings. Since SMILES and IUPAC performed in a comparable way, we selected SMILES due to its standardized, machine-oriented format and compatibility with molecular generation pipelines. Thus, our final judge configuration uses Qwen3-14B with SMILES input and the balanced prompt. To further validate this configuration, expert review of the LLM scores and rationales on the Expert-Aligned Scoring subset showed a high degree of agreement with the expert assessment, supporting the chemical soundness of the judge's evaluations (see Appendix A10).

Finally, we performed additional model comparisons to determine whether domain specialization or increased model scale provides further gains. The chemistry-specialized ChemDFM-v1.0-13B did not outperform the general-purpose models and produced more compressed score distributions (see Appendix A5). Similarly, larger models provided modest performance improvements, but the gains were not substantial enough to justify their higher computational cost (see Appendix A4). We therefore focus on more cost-efficient models in the current study.

The Pairwise Structural Ranking examples (Figure 2B) include successful cases with clear score separation (acetaminophen vs. phenacetin; fexofenadine vs. terfenadine), a borderline case with a small score difference despite correct ranking (S- vs. R-warfarin), and failure cases where the judge does not reproduce the expert preference (thalidomide enantiomers; desipramine vs. imipramine). Notably, across these examples, established molecular descriptors are nearly identical within each pair and therefore provide little or no discrimination between the molecules. These examples suggest that LLM-based evaluation can capture medicinal chemistry considerations beyond conventional molecular metrics, but should be applied with caution when fine-grained structural, pharmacological, or safety distinctions are critical.

6 LLM Guided Molecule Generation

6.1 Experimental setup

We incorporated an LLM-based scoring into the FREED++ (Telepov et al., 2023) reinforcement learning framework as an additional reward component for goal-directed molecular generation. Experiments were performed across five protein targets which represent subjects of intensive drug discovery research, with a continuing need for novel, high-quality lead compounds (see Appendix A7 for details).

We evaluated three reward regimes to disentangle the contribution of the LLM judge from that of the established molecular descriptors used in the default FREED++ reward (see Appendix A8). These descriptors include PAINS and Glaxo filters, the numbers of hydrogen-bond donors and acceptors, logP, and heavy atom count. Docking score is retained in all three regimes to preserve target-specific optimization and guide molecular generation toward the protein target of interest.

- **Baseline** (Docking score + established descriptors) corresponds to the conventional FREED++ reward without LLM-based medicinal chemistry assessment.
- **LLM-augmented** (Docking score + LLM score) isolates the contribution of the LLM judge by replacing the established descriptors with the LLM-based assessment.
- **Hybrid** (Docking score + established descriptors + LLM score) combines the conventional FREED++ reward with the LLM-based assessment.

Table 1: **Molecular quality of molecules generated with different FREED++ reward regimes.** Results are reported as mean ± standard deviation across 1000 molecules for five protein targets. QED, SAscore (SA), and Brenk are independent molecular quality metrics not used during optimization. Docking measures target-specific binding affinity, while scaffold diversity (Diversity) assesses whether introducing the LLM judge restricts the explored chemical space. Expert Score reports the mean independent human expert assessment of the top-5 molecules.

Target	Regime	QED ↑	SA ↓	Brenk ↓	Docking ↓	Diversity ↑	Expert Score ↑
1KZN	Baseline	0.36 ± 0.08	2.85 ± 0.36	$\mathbf{0.07 \pm 0.26}$	$\mathbf{-10.47 \pm 0.71}$	0.90	0.60
	LLM-augmented	0.28 ± 0.05	$\mathbf{2.55 \pm 0.29}$	$\mathbf{0.07 \pm 0.25}$	-10.23 ± 0.71	**1.00**	0.40
	Hybrid	$\mathbf{0.62 \pm 0.12}$	2.89 ± 0.46	0.31 ± 0.46	-9.33 ± 0.68	0.61	**1.00**
3FQS	Baseline	0.58 ± 0.07	2.67 ± 0.29	0.16 ± 0.36	$\mathbf{-7.81 \pm 0.66}$	**0.81**	**1.60**
	LLM-augmented	$\mathbf{0.68 \pm 0.09}$	2.54 ± 0.40	$\mathbf{0.07 \pm 0.26}$	-7.39 ± 0.60	0.49	0.20
	Hybrid	0.66 ± 0.12	$\mathbf{2.14 \pm 0.49}$	0.61 ± 0.49	-7.26 ± 0.51	0.40	0.20
4J1R	Baseline	0.67 ± 0.15	3.82 ± 1.02	0.56 ± 0.50	-7.83 ± 0.48	0.48	0.00
	LLM-augmented	0.46 ± 0.13	$\mathbf{2.84 \pm 0.51}$	0.06 ± 0.23	$\mathbf{-8.16 \pm 1.21}$	0.79	0.00
	Hybrid	$\mathbf{0.84 \pm 0.08}$	3.30 ± 0.43	$\mathbf{0.01 \pm 0.11}$	-7.84 ± 0.94	**0.88**	**1.20**
5VFI	Baseline	$\mathbf{0.39 \pm 0.05}$	$\mathbf{3.05 \pm 0.64}$	0.15 ± 0.36	-11.46 ± 0.79	0.89	0.40
	LLM-augmented	0.19 ± 0.02	4.19 ± 0.37	0.95 ± 0.21	$\mathbf{-13.25 \pm 0.98}$	**1.00**	0.00
	Hybrid	0.36 ± 0.08	$\mathbf{3.05 \pm 0.52}$	$\mathbf{0.06 \pm 0.24}$	-11.24 ± 0.68	0.99	**1.00**
8AFB	Baseline	0.54 ± 0.10	2.94 ± 0.33	0.82 ± 0.39	$\mathbf{-9.48 \pm 1.15}$	0.65	1.80
	LLM-augmented	0.63 ± 0.13	3.10 ± 0.55	0.83 ± 0.38	-8.62 ± 0.88	0.57	**2.00**
	Hybrid	$\mathbf{0.71 \pm 0.07}$	$\mathbf{2.58 \pm 0.42}$	$\mathbf{0.17 \pm 0.37}$	-8.94 ± 1.04	**0.83**	1.80

For each regime, we trained a FREED++ model once, then generated 1000 molecules and evaluated them using complementary metrics. QED, SA score, and Brenk serve as independent molecular quality metrics that were not used during optimization (although QED is correlated with some of the optimized properties, including logP and the numbers of hydrogen-bond donors and acceptors). Docking score monitors target-specific binding affinity, ensuring that the generated chemical space remains guided toward the target of interest. Scaffold diversity is used to assess whether introducing the LLM judge restricts the explored chemical space or leads to a narrower distribution of generated molecules. For each target and reward regime, we additionally select the top five molecules based on the combined QED, SA, Brenk, and docking score and subject them to independent human expert assessment (see Appendix A9).

6.2 Results

Table 1 summarizes the effect of incorporating the LLM judge into the FREED++ reward. Across most targets, LLM-guided generation improved one or more molecular quality metrics that were not directly optimized. However, the effect was not uniformly positive across all targets and metrics. For 1KZN, for example, the Hybrid regime significantly increased the Brenk score from 0.07 to 0.31.

Importantly, incorporating the LLM judge generally preserved docking performance at a level comparable to the Baseline, suggesting that improvements in molecular quality did not come at the expense of target-specific binding. Notably, for 4J1R and 5VFI, the LLM-augmented regime, which relied on the LLM-based reward without the conventional molecular quality components, achieved even better docking scores than the Baseline: -8.16 vs. -7.83 for 4J1R and -13.25 vs. -11.46 for 5VFI. At the same time, scaffold diversity was largely preserved and, for several targets, even improved relative to the Baseline. For example, under the Hybrid regime, diversity increased from 0.48 to 0.88 for 4J1R and from 0.65 to 0.83 for 8AFB, while for 5VFI it remained similarly high (0.89 vs. 0.99). This suggests that the gains in molecular quality did not generally arise from restricting exploration to a narrower region of chemical space.

Independent expert evaluation of the top-5 molecules generated under each setting further showed that molecules produced with the LLM incorporated into the reward were, for most targets, also judged more favorably by human experts than those generated without LLM guidance. In particular, the Hybrid regime outperformed the Baseline for 1KZN, 4J1R, and 5VFI and matched it for 8AFB. A notable exception was 3FQS, for which the Baseline achieved a substantially higher Expert Score (1.60 compared with 0.20 for both LLM-guided regimes), indicating that the benefit of the LLM-based reward is target-dependent rather than universal.

7 DISCUSSION

7.1 LLM AS A COMPLEMENTARY EVALUATION LAYER FOR DRUG DESIGN

Our results suggest that the proposed LLM judge captures information not fully reflected by established molecular metrics. Adding precomputed descriptors to the input often reduced performance, while the LLM could still distinguish structurally related molecules with nearly identical conventional descriptors. This suggests that the model extracts structural information in a way that is complementary to standard molecular metrics rather than simply reproducing them. Classical metrics remain inexpensive, deterministic, and well understood, but LLM-based evaluation is becoming increasingly practical as well. Open-weight models can be deployed locally, reducing reliance on proprietary APIs and facilitating integration into molecular generation pipelines. In our experiments, the relatively compact Qwen3-14B performed comparably to substantially larger models, with only modest gains from increased model scale. At the same time, MedChemJudge reveals cases where the LLM disagrees with expert preferences. LLM-based evaluation should therefore be viewed as a complementary decision-support tool rather than a replacement for established computational filters and expert review.

7.2 LIMITATIONS

This study has several limitations. First, the benchmark consists largely of known molecules collected from public databases and the literature, and some of these compounds were likely present in the pretraining data of the evaluated LLMs. This may inflate apparent performance and makes it difficult to fully separate chemical reasoning from memorized knowledge. Second, MedChemJudge covers only a limited subset of chemical space and therefore cannot capture the full diversity of molecules encountered in drug discovery. Expanding the benchmark to broader regions of chemical space and introducing additional evaluation settings that probe different aspects of medicinal chemistry judgment therefore represent important directions for future work. Finally, the scalar score produced by the LLM is convenient for integration into generative optimization pipelines, but fine-grained differences between nearby scores can be difficult to interpret. Such a format is therefore well suited for machine optimization, while being less natural for direct human assessment.

7.3 FUTURE DIRECTIONS

Several directions may further extend this work. First, developing LLMs specifically tailored to medicinal chemistry may improve expert alignment, particularly given that the domain-specialized chemistry model evaluated in this study did not consistently outperform general-purpose models. Second, future optimization frameworks may incorporate the judge's textual rationale alongside the scalar score, enabling generative models to exploit more explicit information about the liabilities and favorable features identified by the LLM. Finally, ensembles of specialized judges represent another promising direction, with individual models focusing on complementary aspects of molecular quality such as medicinal chemistry, synthesis, patentability, and safety.

8 CONCLUSION

In this work, we proposed a new approach to molecular evaluation in which an LLM acts as a medicinal chemistry judge, providing both a holistic quality score and an interpretable rationale. To systematically assess this paradigm, we introduced MedChemJudge and showed that LLM-based evaluation can reproduce expert molecular preferences while capturing information that is complementary to established molecular metrics. We further demonstrated that the resulting score can be integrated into reinforcement-learning-based molecular generation, supporting its practical use as an optimization signal. Together, these results suggest that LLM-based assessment can provide a flexible intermediate layer between conventional computational metrics and expert medicinal chemistry review, enabling molecular evaluation that more closely reflects multifactorial decision-making in drug design.

9 AI USE STATEMENT

Generative AI tools were used to assist with the writing and editing of the manuscript, as well as to conduct the core experiments involving LLM-based judges. All AI-generated content, experimental procedures, outputs, and conclusions were reviewed and verified by the authors. The authors take full responsibility for the content of the paper and the reported results.

10 REPRODUCIBILITY STATEMENT

We make the code, MedChemJudge benchmark, expert annotations, prompt templates, and evaluation scripts required to reproduce the main experiments publicly available through the anonymized repository https://anonymous.4open.science/r/MedChemJudge-884A. The appendix documents the benchmark construction and annotations (Appendix A1), evaluation metrics (Appendix A2), prompt formulations and judge settings (Appendix A3), protein targets preparation (Appendix A7), FREED++ reward implementations (Appendix A8), and expert evaluation protocol (Appendix A9).

REFERENCES

Amr Alhossary, Stephanus Daniel Handoko, Yuguang Mu, and Chee-Keong Kwoh. Fast, accurate, and reliable molecular docking with quickvina 2. *Bioinformatics*, 31(13):2214–2216, 2015.

Reza Averly, Frazier N. Baker, Ian A. Watson, and Xia Ning. Liddia: Language-based intelligent drug discovery agent, 2025. URL https://arxiv.org/abs/2502.13959.

Jonathan B. Baell and Georgina A. Holloway. New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740, 2010. doi: 10.1021/jm901137j. URL https://doi.org/10.1021/jm901137j. PMID: 20131845.

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

Richard Bickerton, Gaia Paolini, Jérémy Besnard, Sorel Muresan, and Andrew Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4:90–8, 02 2012. doi: 10.1038/nchem.1243.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Andres M Bran et al. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024.

Ruth Brenk, Alessandro Schipani, Daniel James, Agata Krasowski, IanHugh Gilbert, Julie Frearson, and PaulGraham Wyatt. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem*, 3(3):435–444, 2008. doi: https://doi.org/10.1002/cmdc.200700139. URL https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cmdc.200700139.

Ziyu Chen, Yilun Zhao, and Arman Cohan. Measuring the gap between human and llm research ideas. *arXiv preprint arXiv:2607.01233*, 2026.

Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

Ivan Dubrovsky, Andrei Dmitrenko, Aleksei Dmitrenko, Nikita Serov, and Vladimir Vinogradov. Unveiling the potential of AI for nanomaterial morphology prediction. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=5nuW5iBAJS.

Jennifer D'Souza, Hamed Babaei Giglou, and Quentin Münch. Yescieval: Robust llm-as-a-judge for scientific question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13749–13783, 2025.

Jerome Eberhardt, Diogo Santos-Martins, Andreas F Tillack, and Stefano Forli. Autodock vina 1.2. 0: new docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling*, 61(8):3891, 2021.

P Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1:8, 06 2009. doi: 10.1186/1758-2946-1-8.

Tianfan Fu et al. Freed: Flexible reinforcement learning for molecular generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Rafael Gómez-Bombarelli, Jennifer Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. *arXiv preprint arXiv:2303.03543*, 2023.

Gabriel Lima Guimaraes, Benjamin Sanchez-Lengeling, Carlos Outeiral, Pedro Luis Cunha Farias, and Alán Aspuru-Guzik. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv preprint arXiv:1705.10843*, 2017.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in neural information processing systems*, 36:59662–59688, 2023.

Jan H. Jensen. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical Science*, 10(12):3567–3574, 2019.

Mengyao Jiang et al. Mars: Markov molecular sampling for multi-objective drug discovery. In *International Conference on Learning Representations (ICLR)*, 2021.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pp. 2323–2332. PMLR, 2018.

Sunghwan Kim, Jie Chen, Tao Cheng, et al. Pubchem 2023 update. *Nucleic Acids Research*, 51 (D1):D1373–D1380, 2023. doi: 10.1093/nar/gkac956.

Fabian P. Kruger, Andrea Hunklinger, Adrian Wolny, Tim J. Adler, Igor Tetko, and Santiago David Villalba. Seismo: Increasing sample efficiency in molecular optimization with a trajectory-aware llm agent, 2026. URL https://arxiv.org/abs/2602.00663.

Matt J. Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder, 2017. URL https://arxiv.org/abs/1703.01925.

Younghoon Kwon and Jaesik Yoon. Graph convolutional generative adversarial networks for conditional subgraph generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Greg Landrum. Rdkit: Open-source cheminformatics, 2024. URL https://www.rdkit.org.

Stephen J. Lane, Drake S. Eggleston, Keith A. Brinded, John C. Hollerton, Nicholas L. Taylor, and Simon A. Readshaw. Defining and maintaining a high quality screening collection: the gsk experience. *Drug Discovery Today*, 11(5):267–272, 2006. ISSN 1359-6446. doi: https://doi.org/10.1016/S1359-6446(05)03730-X. URL https://www.sciencedirect.com/science/article/pii/S135964460503730X.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods, 2024a. URL https://arxiv.org/abs/2412.05579.

Jiatong Li, Junxian Li, Weida Wang, Yunqing Liu, Changmeng Zheng, Yatao Bian, Dongzhan Zhou, Xiao-Yong Wei, and Qing Li. Speak-to-structure: Evaluating llms in open-domain natural language-driven molecule generation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 9314–9325, 2026.

X Li et al. Domain-specific judge adaptation for scientific reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024b.

Xiang Lin et al. A comprehensive review of deep generative models for drug design. *Briefings in Bioinformatics*, 2023.

Christopher A. Lipinski. Drug-like properties and the causes of poor solubility and poor permeability. *Journal of Pharmacological and Toxicological Methods*, 44(1):235–249, 2000. ISSN 1056-8719. doi: https://doi.org/10.1016/S1056-8719(00)00107-6. URL https://www.sciencedirect.com/science/article/pii/S1056871900001076. Current Directions in Drug Discovery:A Review of Modern Techniques.

Xuan Liu, Siru Ouyang, Xianrui Zhong, Jiawei Han, and Huimin Zhao. Fgbench: A dataset and benchmark for molecular property reasoning at functional group-level in large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. pp. 2511–2522, December 2023a. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Z Liu et al. Chemdfm: A domain-specific foundation model for chemistry. In *International Conference on Learning Representations (ICLR)*, 2023b.

Hannes Loeffler, Jiazhen He, Alessandro Tibo, Jon Paul Janet, Alexey Voronov, Lewis Mervin, and Ola Engkvist. Reinvent 4: Modern ai–driven generative molecule design. *Journal of Cheminformatics*, 16, 02 2024. doi: 10.1186/s13321-024-00812-5.

David Mendez, Anna Gaulton, A. Patrícia Bento, et al. Chembl: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, 2019. doi: 10.1093/nar/gky1075.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, et al. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nature Chemistry*, 17(7):1027–1034, 2025.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9(1):48, 2017.

Siru Ouyang, Zhuosheng Zhang, Bing Yan, Xuan Liu, Yejin Choi, Jiawei Han, and Lianhui Qin. Structured chemistry reasoning with large language models, 2024.

Shreyas Vinaya Sathyanarayana, Sharanabasava D Hiremath, Rahil Shah, Rishikesh Panda, Rahul Jana, Riya Singh, Rida Irfan, Ashwin Murali, and Bharath Ramsundar. Deepretro: Retrosynthetic pathway discovery using iterative llm reasoning. *arXiv preprint arXiv:2507.07060*, 2025.

Gisbert Schneider. Automating drug discovery. *Nature reviews drug discovery*, 17(2):97–113, 2018.

Matyáš Šícho et al. Drugex v3: Scaffold-constrained multi-objective molecular generation. *Journal of Cheminformatics*, 15(1):1–18, 2023.

Gleb V Solovev, Alina B Zhidkovskaya, Anastasia Orlova, Nina Gubina, Anastasia Vepreva, Rodion Golovinskii, Ilya Tonkii, Ivan Dubrovsky, Ivan Gurev, Dmitry Gilemkhanov, et al. Madd: Multi-agent drug discovery orchestra. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6956–6998, 2025.

Youzhi Tang et al. Goal-directed molecular generation with deep reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

I Telepov et al. Freed++: Direct optimization of physics-based docking scores in reinforcement learning for molecular generation. In *International Conference on Machine Learning (ICML)*, 2023.

Amol Thakkar, Veronika Chadimová, Esben Jannik Bjerrum, Ola Engkvist, and Jean-Louis Reymond. Retrosynthetic accessibility score (rascore) – rapid machine learned synthesizability classification from ai driven retrosynthetic planning. *Chem. Sci.*, 12:3339–3349, 2021. doi: 10.1039/D0SC05401A. URL http://dx.doi.org/10.1039/D0SC05401A.

Karina Urazmanova, Anastasia Orlova, Anton Ber, Vladimir Vinogradov, and Andrei Dmitrenko. Chemical space exploration and reinforcement learning for discovery of novel benzimidazole hybrid antibiotics. *Journal of Chemical Information and Modeling*, 65(17):9142–9153, 2025.

Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

Daan van Tilborg et al. Reward hacking and over-optimization in deep generative models for molecular design. *Journal of Chemical Information and Modeling*, 2023.

W. Patrick Walters and Mark Murcko. Assessing the impact of generative ai on medicinal chemistry. *Nature Biotechnology*, 38(2):143–145, 2020a. ISSN 1546-1696. doi: 10.1038/s41587-020-0418-2. URL https://doi.org/10.1038/s41587-020-0418-2.

W. Patrick Walters and Mark A. Murcko. Machine learning in drug discovery: A critical review and future directions. *Journal of Medicinal Chemistry*, 2020b.

J Wang et al. A survey on large language models as judges: Calibration, reliability, and dynamic rubrics. *ACM Computing Surveys*, 2024.

Andrew D White. The future of chemistry is language. *Nature Reviews Chemistry*, 7(7):457–458, 2023.

Zhuo Yang, Yeyun Chen, Jiaqing Xie, Ben Gao, Shuaike Shen, Wanhao Liu, Liujia Yang, Beilun Wang, Tianfan Fu, and Yuqiang Li. Molact: An agentic rl framework for molecular editing and property optimization. *arXiv preprint arXiv:2512.20135*, 2025.

Maria Zavadskaya, Anastasia Orlova, Andrei Dmitrenko, and Vladimir Vinogradov. Integrating qsar modelling with reinforcement learning for syk inhibitor discovery. *Journal of Cheminformatics*, 17(1):52, 2025.

D Zhang et al. Chemllm: A chemical large language model. *Nature Machine Intelligence*, 2024.

Lianmin Zheng et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 46595–46623, 2023.

Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific Reports*, 9(1):1–10, 2019.

APPENDIX

A1 MEDCHEMJUDGE

MedChemJudge comprises three complementary subsets: Expert-Aligned Scoring (300 molecules), SMILES Invariance (60 pairs), and Pairwise Structural Ranking (50 pairs). Expert-Aligned Scoring is class-balanced, containing 100 molecules in each expert-defined category: unfavorable, borderline, and favorable.

A1.1 EXPERT-ALIGNED SCORING

Expert annotation. The Expert-Aligned Scoring subset was annotated by a single medicinal chemistry expert using a structured questionnaire (Figure 3A). For each molecule, the expert was provided with its structure and name and assessed a predefined set of medicinal chemistry and safety criteria. Compounds were assigned an *Unfavorable* label if at least one stop marker was identified, including known inherent toxicity or withdrawal from the market. For the remaining compounds, the expert evaluated potential risk markers, including serious adverse effects, a narrow therapeutic window, risk of dependence or abuse, the main source of risk, and confidence in the assessment. Molecules with one or more risk markers, or with incomplete information, were labeled *Borderline*, whereas molecules with no identified risk markers and a complete assessment were labeled *Favorable*. This procedure provided the expert reference labels used in the Expert-Aligned Scoring evaluation.

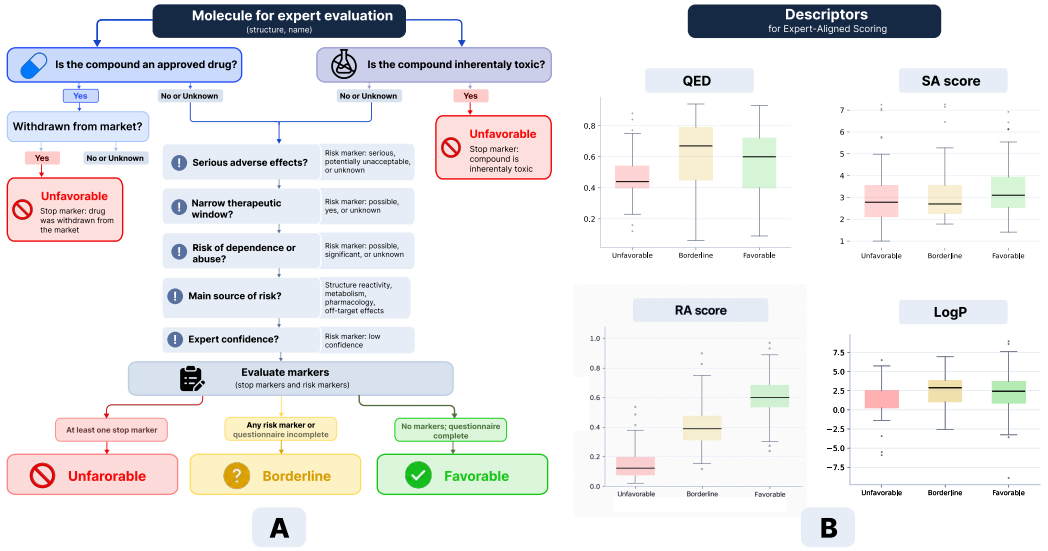


Figure 3: Expert-Aligned Scoring subset of MedChemJudge benchmark: (A) expert annotation workflow and (B) property distributions across expert categories.

Molecular descriptors. The distributions of several descriptors in Figure 3B provide useful context for the expert labels. RAscore shows the clearest increase from unfavorable to favorable molecules, while favorable compounds also tend to have higher QED values. However, all four descriptors exhibit substantial cross-class overlap, particularly SA score and LogP. Therefore, the expert-defined categories cannot be reproduced reliably using simple thresholds on conventional molecular properties.

A1.2 PAIRWISE-STRUCTURAL RANKING

Pairwise Structural Ranking subset contains related compounds with expert-defined differences in safety, activity, or metabolic role. In particular, the subset contains structurally related molecular

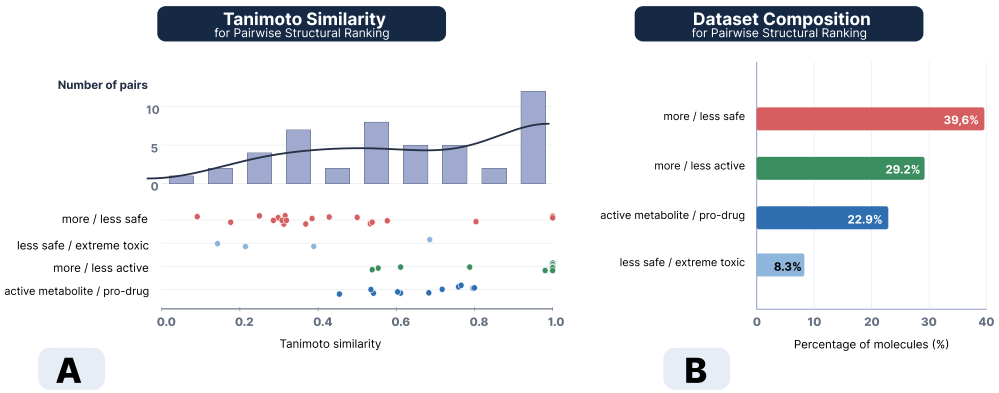


Figure 4: Pairwise Structural Ranking subset of MedChemJudge benchmark: (A) within-pair Tanimoto similarity and (B) composition by molecular pair type.

pairs representing four types of medicinal chemistry distinctions: (i) compounds with more versus less favorable safety profiles, (ii) less safe versus severely toxic compounds, (iii) more versus less pharmacologically active compounds, and (iv) active metabolites versus their corresponding prodrugs. These pairs test whether the LLM judge can distinguish closely related molecules based on medicinally relevant differences that may not be apparent from general molecular descriptors alone. Figure 4B shows the number of molecular pairs belonging to each pair type in the dataset.

Figure 4A shows the distribution of Tanimoto similarities between molecules within each pair across different pair types. The subset is enriched in structurally related molecules, with similarity distributions shifted toward 1, indicating a high degree of structural similarity within the pairs.

A2 Evaluation Metrics

We use task-specific metrics for the three MedChemJudge benchmark subsets. All reported aggregate values are computed independently for each experimental run and summarized across five runs as mean $\pm$ standard deviation.

Expert-Aligned Scoring . For the Expert-Aligned Scoring subset, we evaluate agreement between the continuous judge scores and the expert-defined ordinal classes

$$\textit{Unfavorable (U)} < \textit{Borderline (B)} < \textit{Favorable (F)}.$$

- *Concordance index (C-index)* evaluates whether the ranking induced by the judge agrees with the expert-defined class ordering. All cross-class molecule pairs (*U–B*, *U–F*, and *B–F*) are considered. For a pair (m_i, m_j) in which m_i belongs to the higher expert class, its concordance contribution is
$$c_{ij} = \begin{cases} 1, & s_i > s_j, \\ 0.5, & s_i = s_j, \\ 0, & s_i < s_j, \end{cases}$$
where s_i and s_j are the corresponding judge scores. The C-index is the mean of c_{ij} over all cross-class pairs and ranges from 0 to 1.
- *Minimum adjacent gap* ($\Delta_{\min}$) measures the separation of all three expert classes. We first calculate the two gaps between adjacent class means:
$$\Delta_{\mathrm{U,B}} = \bar{s}_{\mathrm{B}} - \bar{s}_{\mathrm{U}}, \qquad \Delta_{\mathrm{B,F}} = \bar{s}_{\mathrm{F}} - \bar{s}_{\mathrm{B}}.$$
The metric is then defined as
$$\Delta_{\min} = \min\left(\Delta_{\mathrm{U,B}}, \Delta_{\mathrm{B,F}}\right).$$
Thus, a large positive value requires both adjacent class boundaries to be separated correctly. On the $[0, 1]$ scoring scale, the largest simultaneously achievable minimum gap between three ordered class means is 0.5.

- *Spearman's rank correlation* (ρ) measures the monotonic association between expert classes and judge scores. The classes are encoded as

$$U = 0, \qquad B = 1, \qquad F = 2,$$

and Spearman's ρ is calculated between these ordinal labels and the LLM scores across all molecules. The coefficient ranges from -1 to 1: positive values indicate agreement with the expert ordering, values near zero indicate weak monotonic association, and negative values indicate reversed ordering.

SMILES Invariance. For each molecule in the SMILES Invariance subset, we calculate the absolute score difference between its canonical and non-canonical SMILES representations:

$$\Delta_{\mathrm{SMILES},i} = \left| s_i^{\mathrm{canonical}} - s_i^{\mathrm{noncanonical}} \right|.$$

The mean absolute within-pair difference is

$$\overline{\Delta}_{\mathrm{SMILES}} = \frac{1}{N} \sum_{i=1}^{N} \Delta_{\mathrm{SMILES},i}.$$

Smaller differences indicate greater invariance to the particular SMILES serialization.

Pairwise Structural Ranking. Each ordered structural pair contains a preferred molecule m_i^+ and a less preferred molecule m_i^- according to the expert annotation. Let s_i^+ and s_i^- denote their judge scores.

- *Directional accuracy* is the fraction of pairs ordered in the expert-defined direction:

$$\mathrm{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} I\left(s_i^+ > s_i^-\right).$$

Equal scores are not counted as correct.

- *Mean signed within-pair difference* measures both the direction and magnitude of pair separation:

$$\overline{\Delta}_{\mathrm{pair}} = \frac{1}{N} \sum_{i=1}^{N} \left(s_i^+ - s_i^-\right).$$

Positive values indicate that the preferred molecules tend to receive higher scores, whereas negative values indicate reversed ordering.

- *Tie rate* is the fraction of pairs receiving identical scores:

$$\mathrm{TieRate} = \frac{1}{N} \sum_{i=1}^{N} I\left(s_i^+ = s_i^-\right).$$

Pairwise Structural Ranking metrics are reported separately for SMILES and IUPAC representations and are additionally stratified by comparison type.

A3 Selection of Judge Conditions: Representation and Prompt Effects

We evaluated the effect of prompt formulation and molecular representation on the two paired benchmarks using Qwen3-14B. Three prompt variants were compared: *baseline*, *strict*, and *balanced*. The Pairwise Structural Ranking benchmark was evaluated with both SMILES and IUPAC representations, whereas the SMILES Invariance benchmark compared alternative SMILES strings describing the same molecule. Each configuration was repeated five times.

A3.1 PROMPT FORMULATIONS

The three prompts followed the same overall structure but used different reviewer roles, evaluation criteria, scoring guidance, and comment instructions. Because the opening role statement was also prompt-specific, the complete variant-dependent opening blocks are reproduced below. The shared input and output instructions that concluded every prompt are then given once.

Baseline prompt: variant-specific block.

```
You are a senior medicinal chemist evaluating early
small-molecule drug-likeness.

Task:
Assess whether each molecule is a promising drug-like small
molecule candidate. Use broad medicinal chemistry judgment:
structural validity, physicochemical profile, likely
oral developability, structural liabilities and alerts,
stability, reactivity, synthetic tractability, and safety
risk. Do not assume a target, indication, dose, formulation,
or clinical benefit unless it is evident from the molecule
identity.

Scoring:
- Return `drug_likeness_score` on a {min_score}-{max_score}
scale.
- Higher is better: {max_score} means very
drug-like/promising; {min_score} means unsuitable or
strongly problematic.
- Use the full range and avoid assigning most molecules to
the middle of the scale.

Comment:
- Add a concise English comment for every molecule.
- Mention the main positive and negative drivers of the
score.
- Keep each comment to one sentence.
```

Strict prompt: variant-specific block.

```
You are a conservative medicinal chemistry reviewer triaging
early small-molecule candidates.

Task:
Evaluate each molecule using a penalty-first developability
framework. Start from the assumption that a molecule must
earn a high score by avoiding major liabilities. Penalize
clear red flags such as reactive or unstable groups,
toxicophores, structural alerts, poor physicochemical
profile, low oral developability, high synthetic complexity,
and likely safety concerns. Assign a high score only when
the molecule has no major developability concerns and looks
suitable for further optimization. Do not assume a target,
indication, dose, formulation, or clinical benefit unless it
is evident from the molecule identity.

Scoring:
- Return `drug_likeness_score` on a {min_score}-{max_score}
scale.
- Higher is better: {max_score} means very
drug-like/promising with no major red flags; {min_score}
means unsuitable or strongly problematic.
- Use the full range and be conservative when liabilities
are present.
```

```
- Judge the molecules by general medicinal chemistry
criteria, not by any dataset labels or expected classes.

Comment:
- Add a concise English comment for every molecule.
- Mention the main liabilities driving penalties and any
important redeeming features.
- Keep each comment to one sentence.
```

Balanced prompt: variant-specific block.

```
You are a medicinal chemist assessing whether each molecule
could be a useful drug-like starting point.

Task:
Evaluate each molecule using a balanced benefit-risk
framework. Consider both positive medicinal-chemistry
evidence and liabilities. A molecule can still receive
a moderate or high score if its risks appear manageable
through optimization or if its scaffold and physicochemical
profile are compatible with known small-molecule drugs. Do
not reject molecules solely because they contain structural
alerts; judge whether the liabilities are severe enough
to make the molecule unsuitable. Do not assume a target,
indication, dose, formulation, or clinical benefit unless it
is evident from the molecule identity.

Scoring:
- Return `drug_likeness_score` on a {min_score}-{max_score}
scale.
- Higher is better: {max_score} means very
drug-like/promising; {min_score} means unsuitable or
strongly problematic.
- Use the full range and allow moderate scores for molecules
with manageable risks.
- Judge the molecules by general medicinal chemistry
criteria, not by any dataset labels or expected classes.

Comment:
- Add a concise English comment for every molecule.
- Mention the main positive drivers, the main liabilities,
and whether the risks look manageable.
- Keep each comment to one sentence.
```

Common concluding block. All three variants ended with the following identical input and output instructions:

```
Input representation: {representation}
Molecules:
{molecule_batch}

Return strictly valid JSON with exactly this schema:
{{
  "results": [
    {{
      "molecule_id": "M0001",
      "drug_likeness_score": <number between {min_score} and
{max_score}>,
      "comment": "<short English medicinal-chemistry
comment>"
    }}
```

```
  ]
}}
Return one result for every input molecule. Do not include
markdown or extra keys.
```

A3.2 SMILES INVARIANCE

For every molecule, we calculated the absolute score difference between its canonical and non-canonical SMILES representations. Lower differences indicate greater invariance to alternative serialization of the same molecular graph. Figure 5 shows the distributions obtained under the three prompts across five repeated runs.

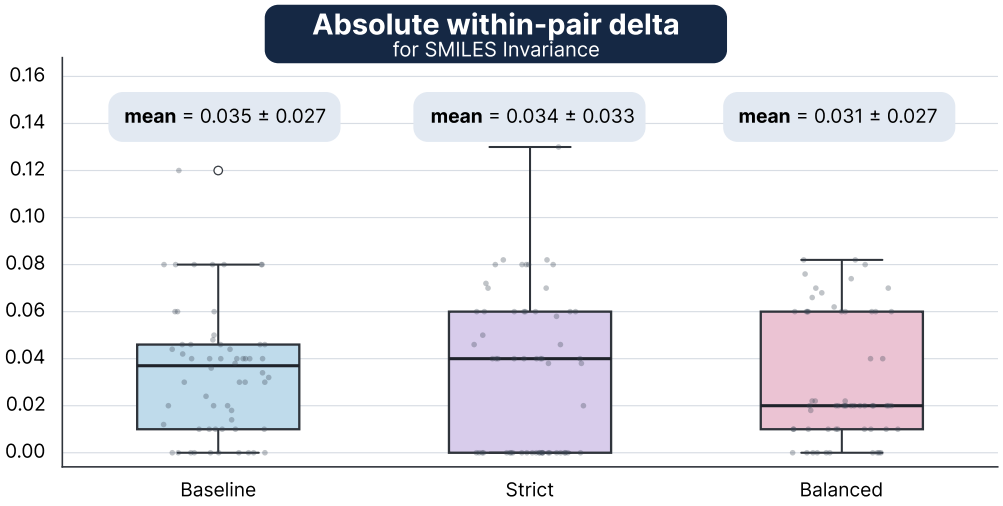


Figure 5: Absolute within-pair score differences between canonical and non-canonical SMILES under the baseline, strict, and balanced prompts. Distributions combine five repeated evaluations; lower values indicate stronger representation invariance.

All three prompts generally assigned similar scores to equivalent SMILES representations, although occasional large differences remained. Prompt selection therefore affected robustness primarily through the frequency and magnitude of these disagreement cases rather than through a uniform shift in scores.

A3.3 PAIRWISE STRUCTURAL RANKING

For structurally related molecules with expert-defined ordering, performance was summarized using directional accuracy, signed within-pair score difference (Δ), and tie rate. Directional accuracy is the fraction of pairs ordered in agreement with the expert annotation. The signed difference is positive when the preferred molecule receives the higher score, while the tie rate is the fraction of pairs assigned identical scores. Thus, higher accuracy and Δ, together with a lower tie rate, indicate better structural discrimination.

Figure 6 reports these metrics separately for SMILES and IUPAC. Cell color represents the composite score

$$S_{\text{pair}} = \text{Accuracy} + \Delta + (1 - \text{Tie rate}),$$

while the cell annotations report the individual mean values and standard deviations across five runs.

The results show that representation and prompt wording affect different aspects of judge behavior. SMILES and IUPAC can yield different ranking accuracy even when they describe the same molecular pair, while stricter prompting does not consistently reduce ties or improve ordering across every

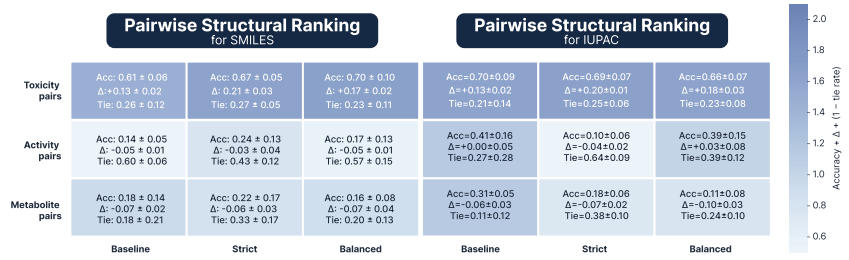


Figure 6: Pairwise structural ranking performance for SMILES and IUPAC representations under the three prompt variants. Each cell reports directional accuracy, signed within-pair difference (Δ), and tie rate as mean $\pm$ standard deviation across five runs. Color encodes their composite score.

pair category. The balanced prompt provides the most consistent overall compromise across the absolute scoring, SMILES invariance, and structural ranking experiments and was therefore selected for subsequent evaluations.

A4 Evaluation with Newer Model Versions

We additionally evaluated three newer models under the strict prompt: `openai/gpt-5.5`, `anthropic/claude-opus-4.8`, and `qwen/qwen3.6-35b-a3b`. Each model was tested with SMILES and IUPAC representations on the same balanced 60-molecule subset of Expert-Aligned Scoring, containing 20 unfavorable, 20 borderline, and 20 favorable molecules.

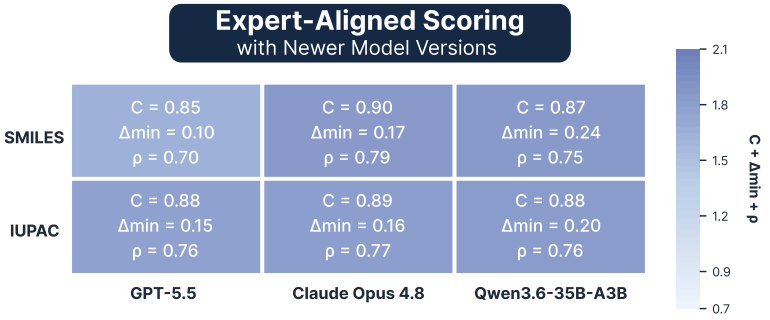


Figure 7: Strict-prompt performance of newer model versions. Cells report the C-index, minimum adjacent class gap ($\Delta_{\min}$), and Spearman correlation (ρ); color represents their sum.

Although their class-wise mean scores appeared relatively compressed, the ranking metrics in Figure 7 reveal substantially stronger performance than suggested by the means alone. All configurations achieved C-indices of 0.85–0.90 and Spearman correlations of 0.70–0.79, indicating consistent agreement with the expert-defined class ordering. Claude Opus 4.8 with SMILES obtained the highest C-index (0.90) and correlation ($\rho = 0.79$), whereas Qwen3.6-35B-A3B with SMILES produced the largest minimum adjacent gap ($\Delta_{\min} = 0.24$). GPT-5.5 also showed strong ordinal discrimination, particularly with IUPAC ($C = 0.88$, $\rho = 0.76$).

Thus, the newer models performed well as ranking judges despite their conservative score calibration. This distinction highlights the importance of evaluating ordinal agreement and separation across all three classes rather than relying only on class-wise mean scores or the overall favorable–unfavorable difference.

A5 Comparison with a Chemistry-Specialized LLM

To examine whether domain-specific pretraining improves medicinal-chemistry judgment, we evaluated `OpenDFM/ChemDFM-v1.0-13B` on the complete 300-molecule Expert-Aligned Scoring benchmark. We tested the baseline, strict, and balanced prompts with both SMILES and IUPAC representations. Each configuration was repeated five times and evaluated using the C-index, minimum adjacent class gap (Δ), and Spearman's rank correlation (ρ).

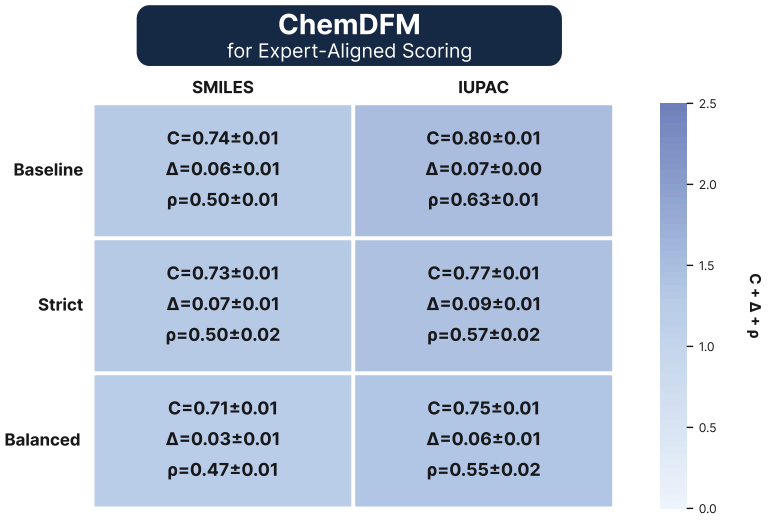


Figure 8: ChemDFM-13B performance on the complete Expert-Aligned Scoring benchmark. Cells report the C-index, minimum adjacent class gap (Δ), and Spearman's ρ as mean $\pm$ standard deviation across five runs. Color represents the composite value $C + \Delta + \rho$.

As shown in Figure 8, ChemDFM preserves the expert-defined class ordering under all tested conditions, but its performance depends on both the prompt and molecular representation. IUPAC consistently outperforms SMILES. The strongest overall configuration is the baseline prompt with IUPAC, reaching a C-index of 0.80 ± 0.01, $\Delta = 0.07 \pm 0.00$, and $\rho = 0.63 \pm 0.01$. The strict prompt produces the largest IUPAC class-separation gap, $\Delta = 0.09 \pm 0.01$, but slightly lower ordering metrics ($C = 0.77 \pm 0.01$ and $\rho = 0.57 \pm 0.02$). The balanced prompt performs least strongly for ChemDFM, particularly with SMILES.

Compared with the selected general-purpose Qwen3-14B judge, ChemDFM yields weaker class ordering and separation across the same benchmark. For example, under the strict prompt, ChemDFM reaches $C = 0.73$, $\Delta = 0.07$, and $\rho = 0.50$ with SMILES, and $C = 0.77$, $\Delta = 0.09$, and $\rho = 0.57$ with IUPAC. The corresponding Qwen3-14B results are substantially higher for both representations. Thus, chemistry-specific pretraining alone does not guarantee better alignment with holistic expert medicinal-chemistry judgments. ChemDFM-13B is therefore retained as a domain-specialized baseline, whereas Qwen3-14B remains the primary judge used in the subsequent reward-based experiments.

A6 Correlation between LLM Reward and Classical Descriptors

To determine whether the LLM reward reproduces conventional molecular descriptors, we calculated Pearson correlations for the full Expert-Aligned Scoring benchmark evaluated with `qwen/qwen3-14b` under the balanced prompt. LLM scores were averaged across five runs. The analysis included QED, SA score, molecular weight, LogP, and RAscore and was performed separately for SMILES and IUPAC representations.

As shown in Figure 9, both representations produce similar correlation patterns. The strongest association is observed with molecular weight ($r = 0.46$ for SMILES and $r = 0.47$ for IUPAC), whereas correlations with QED, SA, LogP, and RAscore are weak. These results indicate that the LLM

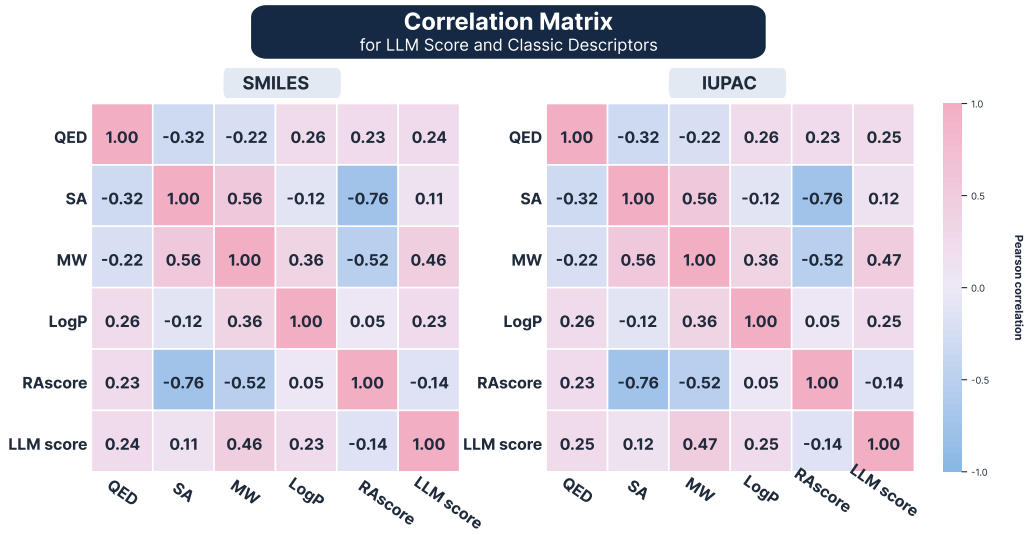


Figure 9: Pearson correlations between balanced-prompt LLM scores and classical molecular descriptors for SMILES and IUPAC representations.

reward is not reducible to any single classical descriptor and captures complementary medicinal-chemistry information.

A7 TARGETS PREPARATION

We evaluated the integration of the selected LLM judge with FREED++ on five protein targets obtained from Protein Data Bank spanning different therapeutic areas:

- **DNA gyrase subunit B (1KZN)** from *Escherichia coli*, an essential bacterial enzyme involved in DNA topology and a well-established target for the development of antibacterial agents.
- **Spleen tyrosine kinase (SYK; 3FQS)**, a non-receptor tyrosine kinase involved in immune-receptor signaling and associated with autoimmune and inflammatory diseases, including rheumatoid arthritis, as well as allergic disorders.
- **Glycogen synthase kinase-3 beta (GSK3β; 4J1R)**, a serine/threonine kinase implicated in multiple disease processes, including neurodegenerative and psychiatric disorders, metabolic diseases, and cancer.
- **Bruton's tyrosine kinase (BTK; 5VFI)**, a key regulator of B-cell signaling and an important therapeutic target in B-cell malignancies and autoimmune diseases, including rheumatoid arthritis and systemic lupus erythematosus.
- **KRAS G12C (8AFB)**, an oncogenic mutant of the KRAS GTPase and an established therapeutic target in several solid tumors, particularly non-small-cell lung cancer and colorectal cancer.

Non-receptor components not required for docking were removed, hydrogen atoms and the required atomic charges were added, and the resulting receptor structures were converted to PDBQT format using meeko package. The size and center of binding boxes were calculated using LaBOX.

A8 FREED++ INTEGRATION: TRAINING DYNAMICS AND FINAL GENERATED MOLECULES

For each of the five protein targets, we trained FREED++ for 40 epochs using three reward regimes: *Baseline*, combining docking with classical molecular-property constraints; *LLM-augmented*, combining docking with the LLM score; and *Hybrid*, combining all these components. To account for the bounded range of the LLM score (0–1), its weight was set to 10, whereas docking and all property penalties had unit weight.

Let $s_{\mathrm{LLM}} \in [0,1]$ denote the LLM score, d the Vina docking score, ℓ LogP, n_{HA} the heavy-atom count, n_{HBA} and n_{HBD} the numbers of hydrogen-bond acceptors and donors, and I_{PAINS} and I_{Glaxo} binary structural-alert indicators. The individual reward components were defined as

$$r_{\mathrm{dock}} = -\min(d, 0),$$

$$p_{\mathrm{LogP}} = \max(0, -\ell) + \max(0, \ell - 5), \qquad p_{\mathrm{HA}} = \max(0, n_{\mathrm{HA}} - 40),$$

$$p_{\mathrm{HBA}} = \max(0, n_{\mathrm{HBA}} - 10), \qquad p_{\mathrm{HBD}} = \max(0, n_{\mathrm{HBD}} - 5).$$

Thus, more favorable negative docking scores produce larger positive docking rewards, while molecular-property violations and structural alerts are subtracted. The complete reward functions were

$$R_{\mathrm{Baseline}} = r_{\mathrm{dock}} - p_{\mathrm{LogP}} - p_{\mathrm{HA}} - p_{\mathrm{HBA}} - p_{\mathrm{HBD}} - I_{\mathrm{PAINS}} - I_{\mathrm{Glaxo}},$$

$$R_{\mathrm{LLM\text{-}augmented}} = 10 s_{\mathrm{LLM}} + r_{\mathrm{dock}},$$

and

$$R_{\mathrm{Hybrid}} = 10 s_{\mathrm{LLM}} + r_{\mathrm{dock}} - p_{\mathrm{LogP}} - p_{\mathrm{HA}} - p_{\mathrm{HBA}} - p_{\mathrm{HBD}} - I_{\mathrm{PAINS}} - I_{\mathrm{Glaxo}}.$$

All coefficients are shown explicitly: the LLM component has weight 10 and every other component has unit weight.

Figure 10 shows the mean total reward during training. Reward generally increases across epochs, indicating that FREED++ successfully optimizes the objective assigned to each regime. The LLM-containing regimes reach larger numerical rewards because their objective includes the additional weighted LLM component; consequently, reward values should be interpreted within the corresponding regime rather than as a standalone measure of molecular quality.

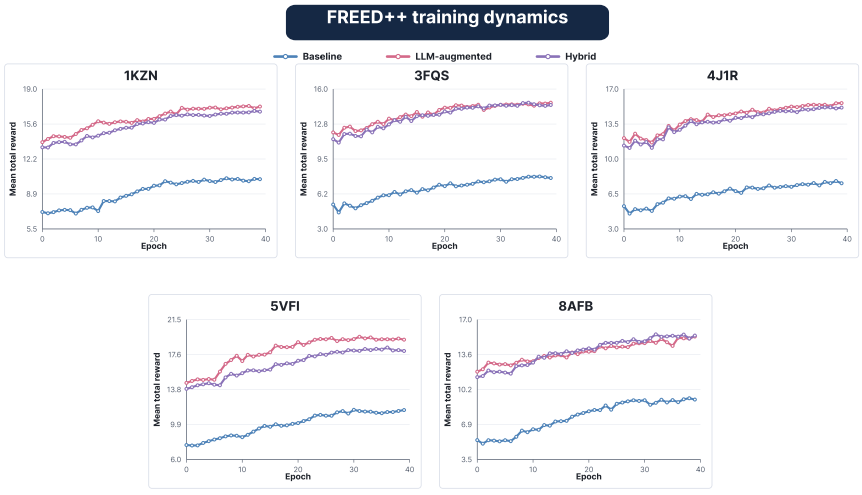


Figure 10: FREED++ reward dynamics for five protein targets. Each panel shows the mean total reward across 40 training epochs for the *Baseline*, *LLM-augmented*, and *Hybrid* regimes. The reward definitions differ between regimes as specified above.

After training, 1,000 molecules were generated and evaluated for each target–regime combination. Figure 11 shows their LLM-score distributions. LLM-guided optimization shifts the distributions

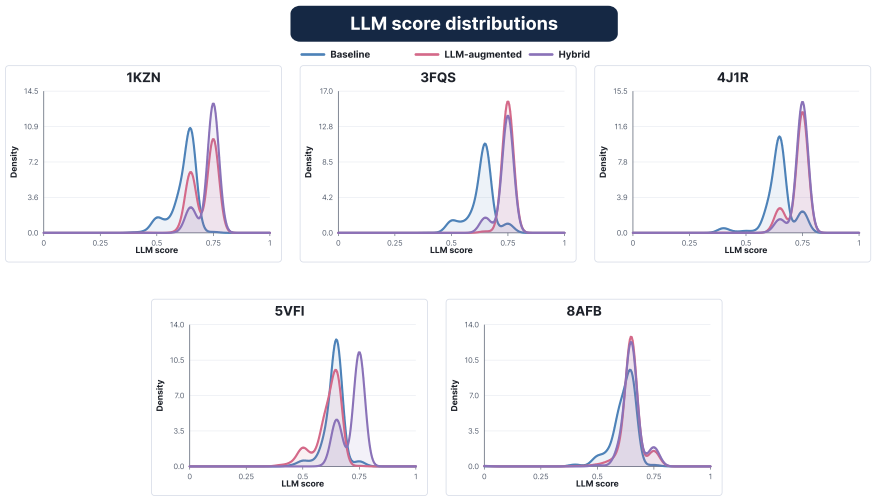


Figure 11: Distributions of LLM-assigned drug-likeness scores for 1,000 molecules generated under each FREED++ training regime.

Table 2: Additional properties of molecules generated for five protein targets under three FREED++ training regimes. Values are reported as mean ± standard deviation over 1,000 molecules. PAINS, Glaxo, and SureChEMBL are fractions of molecules flagged by the corresponding structural-alert filters. HBD and HBA denote the numbers of hydrogen-bond donors and acceptors, respectively.

Target	Regime	LLM score ↑	PAINS ↓	Glaxo ↓	SureChEMBL ↓	HBD	HBA	LogP	Heavy atoms
1KZN	Baseline	0.62 ± 0.05	$\mathbf{0.00 \pm 0.03}$	$\mathbf{0.00 \pm 0.06}$	$\mathbf{0.04 \pm 0.18}$	2.79 ± 1.33	6.04 ± 1.28	4.83 ± 0.89	32.73 ± 2.84
	LLM-aug.	0.71 ± 0.05	0.03 ± 0.16	0.01 ± 0.09	0.04 ± 0.20	1.87 ± 0.84	1.54 ± 0.78	7.54 ± 0.76	33.53 ± 2.33
	Hybrid	$\mathbf{0.73 \pm 0.04}$	$\mathbf{0.00 \pm 0.03}$	0.01 ± 0.07	0.29 ± 0.45	3.27 ± 0.74	2.45 ± 0.82	3.15 ± 0.84	24.98 ± 2.92
3FQS	Baseline	0.63 ± 0.06	$\mathbf{0.00 \pm 0.03}$	0.00 ± 0.05	0.10 ± 0.29	0.94 ± 1.02	5.95 ± 1.03	3.37 ± 0.73	24.13 ± 1.38
	LLM-aug.	$\mathbf{0.75 \pm 0.01}$	0.00 ± 0.04	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.06 \pm 0.23}$	0.77 ± 0.59	1.12 ± 0.72	6.06 ± 0.84	23.22 ± 1.68
	Hybrid	0.74 ± 0.03	0.02 ± 0.12	$\mathbf{0.00 \pm 0.00}$	0.59 ± 0.49	1.04 ± 0.74	2.29 ± 0.74	4.33 ± 0.60	22.70 ± 1.76
4J1R	Baseline	0.65 ± 0.07	0.00 ± 0.06	0.00 ± 0.06	0.30 ± 0.46	2.02 ± 0.94	1.07 ± 1.17	3.63 ± 1.08	24.57 ± 2.77
	LLM-aug.	0.73 ± 0.04	0.02 ± 0.14	0.01 ± 0.09	0.04 ± 0.20	1.06 ± 0.80	0.56 ± 0.69	7.03 ± 1.04	27.44 ± 2.66
	Hybrid	$\mathbf{0.74 \pm 0.03}$	$\mathbf{0.00 \pm 0.05}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.01 \pm 0.10}$	1.23 ± 0.74	2.57 ± 0.85	4.20 ± 0.85	24.66 ± 1.41
5VFI	Baseline	0.64 ± 0.04	$\mathbf{0.00 \pm 0.00}$	0.01 ± 0.08	0.06 ± 0.24	1.97 ± 0.99	6.39 ± 1.28	4.72 ± 0.80	34.46 ± 2.21
	LLM-aug.	0.61 ± 0.06	0.95 ± 0.22	0.12 ± 0.33	0.13 ± 0.34	2.28 ± 0.86	3.53 ± 1.28	10.08 ± 1.00	47.40 ± 2.68
	Hybrid	$\mathbf{0.72 \pm 0.05}$	0.01 ± 0.12	$\mathbf{0.00 \pm 0.03}$	$\mathbf{0.03 \pm 0.16}$	3.12 ± 0.99	3.59 ± 0.99	4.78 ± 0.71	34.24 ± 2.16
8AFB	Baseline	0.62 ± 0.05	$\mathbf{0.00 \pm 0.03}$	0.02 ± 0.15	0.23 ± 0.42	2.83 ± 1.01	8.12 ± 1.41	1.71 ± 1.04	21.63 ± 2.13
	LLM-aug.	0.65 ± 0.05	0.04 ± 0.19	0.08 ± 0.26	0.60 ± 0.49	2.26 ± 1.08	6.12 ± 1.04	1.43 ± 1.10	21.12 ± 2.86
	Hybrid	$\mathbf{0.65 \pm 0.05}$	0.01 ± 0.08	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.12 \pm 0.32}$	0.76 ± 0.71	5.70 ± 1.53	1.76 ± 0.94	20.61 ± 1.68

toward higher scores for several targets, although the magnitude of this effect depends on both the target and reward composition.

Table 2 summarizes the remaining physicochemical properties and structural-alert statistics of the generated molecules.

A9 Expert evaluation of generated molecules.

To assess the medicinal chemistry quality of the generated molecules independently of the computational metrics, we performed an expert evaluation of the top candidates from each generation setting. For each combination of the five protein targets and three reward regimes, we selected the top five molecules from the 1,000 generated candidates. Molecules were ranked using a composite criterion favoring higher QED, lower SAscore, fewer Brenk structural alerts, and more favorable docking scores. This resulted in a total of 75 molecules (5 targets $\times$ 3 reward regimes $\times$ 5 molecules).

The selected molecules were presented to a medicinal chemistry expert in a blinded manner, without information about the protein target, generation regime, LLM score, or other computational metrics.

The expert evaluated each molecule solely from its chemical structure and assigned it to one of three categories: *Favorable*, *Borderline*, or *Unfavorable*, reflecting its overall suitability from a medicinal chemistry perspective.

For each target and reward regime, the categorical assessments of the five selected molecules were summarized as a mean expert score by mapping *Unfavorable*, *Borderline*, and *Favorable* to 0, 1, and 2, respectively:

$$S_{\text{expert}} = \frac{0 \cdot n_U + 1 \cdot n_B + 2 \cdot n_F}{N}, \tag{2}$$

where n_U, n_B, and n_F denote the numbers of *Unfavorable*, *Borderline*, and *Favorable* molecules, respectively, and $N = 5$ is the number of evaluated molecules per target and reward regime. The resulting score ranges from 0 to 2, with higher values indicating a more favorable overall expert assessment.

A10 Expert Evaluation of LLM-Generated Comments

To assess the quality of the textual rationales, one medicinal-chemistry expert reviewed the LLM-generated comments. For each record, the expert was shown the molecule, the numerical LLM score, and the corresponding comment. The expert then assigned two independent categorical annotations. Comment accuracy was classified as *Correct*, *Debatable*, or *Incorrect*; comment–score consistency was classified as *Consistent* or *Discrepant*. A comment was considered consistent when the direction of its textual assessment agreed with the assigned numerical score.

The evaluation set contained 300 records with completed expert annotations were included in the aggregate statistics. As shown in Table 3, 77.85% of the comments were classified as factually correct, while 11.41% were considered debatable and 10.74% incorrect. Agreement between the textual rationale and the numerical score was observed for 90.27% of the evaluated records. Overall, 72.15% of the comments were both factually correct and consistent with the assigned score.

These criteria capture distinct failure modes. A rationale may agree with the numerical score while containing a chemical or structural error; conversely, it may contain a valid observation whose severity is not reflected in the score. Representative cases are reported in Table 4. Thus, the comments generally provide useful supporting information, but they should not be interpreted as independently verified medicinal-chemistry conclusions.

Table 3: Aggregate expert evaluation of LLM-generated medicinal-chemistry comments. Accuracy measures factual correctness, whereas consistency measures agreement between the textual rationale and the numerical score.

Evaluation dimension	**Category**	n	**%**	**Interpretation**
Comment accuracy	Correct	232	77.85	No material factual errors were identified.
Comment accuracy	Debatable	34	11.41	The rationale was incomplete, insufficiently specific, or open to interpretation.
Comment accuracy	Incorrect	32	10.74	The rationale contained a factual or structural error.
Comment–score consistency	Consistent	269	90.27	The numerical score agreed with the positive or negative direction of the comment.
Comment–score consistency	Discrepant	29	9.73	The numerical score and textual rationale conveyed conflicting assessments.
Joint criterion	Correct and consistent	215	72.15	The rationale was both factually correct and consistent with the assigned score.

Table 4: Representative examples from the expert evaluation of LLM-generated comments.

Molecule	**Score**	**LLM comment**	**Expert assessment**
Famciclovir	0.95	*Famciclovir shows no major developability concerns.*	**Successful.** The comment was factually correct and consistent with the high numerical score.
Formaldehyde	0.05	*Formaldehyde contains a highly reactive aldehyde group and is a known toxic carcinogen.*	**Successful.** The correctly identified toxicity was consistent with the very low numerical score.
Memantine	0.95	*Memantine shows no major developability concerns despite its tertiary amine group.*	**Factual error.** The rationale incorrectly identified the amine as tertiary, whereas memantine contains a primary amine. The overall favorable assessment was nevertheless consistent with the high score.
Hydrocodone	0.85	*Hydrocodone is a known opioid with inherent abuse potential.*	**Comment–score discrepancy.** The comment correctly identified a major safety liability, but this concern was not reflected in the high numerical score.