Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Agentic reinforcement learning empowers next-generation chemical language models for molecular design and synthesis

Hao Li

lihao1984@pku.edu.cn

Peking University https://orcid.org/0000-0002-3200-0270

He Cao

International Digital Economy Academy (IDEA)

Shenyao Peng

Yuanpei College, Peking University

Zijing Liu

International Digital Economy Academy (IDEA) https://orcid.org/0000-0002-0189-7409

Bin Feng

International Digital Economy Academy (IDEA)

Yu Wang

School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School

Zhiyuan Yan

Peking University

Liuzhenghao Lv

Peking University Shenzhen Graduate School

Yang Zhang

National University of Singapore https://orcid.org/0000-0002-2739-1916

Yonghong Tian

Peking University https://orcid.org/0000-0002-2978-5935

Yu Li

International Digital Economy Academy https://orcid.org/0000-0003-1865-8276

Li Yuan

Peking University https://orcid.org/0000-0002-2120-5588

Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9959783/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Agentic reinforcement learning empowers next-generation chemical language models for molecular design and synthesis

Hao Li[1,2,3*], **He Cao**[2*], **Shenyao Peng**[1,3*], **Zijing Liu**[2], **Bin Feng**[2], **Yu Wang**[1,3], **Zhiyuan Yan**[1,3], **Zhenghao Lvliu**[1,3], **Yang Zhang**[4], **Yonghong Tian**[1,3†], **Yu Li**[2†], **and Li Yuan**[1,3†]

[1] *School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School, Shenzhen, 518055, China*
[2] *International Digital Economy Academy (IDEA)*
[3] *Peng Cheng Laboratory, Shenzhen, 518000, China*
[4] *National University of Singapore, Singapore*
[*] These authors contributed equally to this work
[†] Corresponding authors: yuanli-ece@pku.edu.cn, liyu@idea.edu.cn, yhtian@pku.edu.cn

ABSTRACT

Language models are revolutionizing the biochemistry domain, assisting scientists in drug design and chemical synthesis with high efficiency. Yet current approaches struggle between small language models prone to hallucination and limited knowledge retention, and large cloud-based language models plagued by privacy risks and high inference costs. To bridge this gap, we introduce ChemCRAFT, a novel framework leveraging agentic reinforcement learning to decouple chemical reasoning from knowledge storage. Instead of forcing the model to memorize vast chemical data, our approach empowers the language model to interact with a sandbox for precise information retrieval. This externalization of knowledge allows a locally deployable small model to achieve superior performance while being highly token-efficient compared to traditional multi-agent systems. To enable small language models for agent-calling ability, we build an agentic trajectory construction pipeline and a comprehensive chemical-agent sandbox. Based on sandbox interactions, we constructed ChemToolDataset, the first large-scale chemical tool trajectory dataset. Simultaneously, we propose SMILES-GRPO to build a dense chemical reward function, promoting the model's ability to call chemical agents. Evaluations across diverse aspects of drug design show that ChemCRAFT outperforms current cloud-based LLMs in molecular structure analysis, molecular optimization, and synthesis pathway prediction, demonstrating that scientific reasoning is not solely an emergent ability of model scale, but a learnable policy of tool orchestration. Furthermore, assessments in real-world drug optimization scenarios, including fentanyl analog optimization and pro-inflammatory GSTO1 inhibitors optimization, reveal that ChemCRAFT generates optimization trajectories mirroring human experts, confirming its reliability as a scientific assistant. This work establishes a cost-effective and privacy-preserving paradigm for AI-aided chemistry, opening new avenues for accelerating molecular discovery with locally deployable agents. Code available at https://github.com/HowardLi1984/ChemCraft.

1 Introduction

Chemical Language Models (CLMs) have emerged as transformative tools in accelerating drug discovery and materials science[1–4], demonstrating remarkable potential in tasks ranging from molecular property prediction[5–7] and de novo design[8–11] to retrosynthesis planning[12–14] and reaction condition recommendation[15–17]. Current research has evolved from simple generation into two distinct technical trajectories[18]: The *optimization-centric paradigm* leveraging parameter-space manipulation to internalize scientific knowledge; The *system-centric paradigm*, employing multi-LLM collaboration and external tools to address complex drug discovery problems.

The *optimization-centric paradigm* primarily utilizes supervised finetuning or continue pretraining of models on chemical corpora[19–23]. While these models achieve competitive performance on in-distribution metrics[20,24,25], they fundamentally learn a shallow mapping from chemical inputs directly to outputs rather than multi-step reasoning[26–28]. This approach bypasses the analytical processes such as structural analysis and logical verification that are essential for scientific discovery[29–31], leading to catastrophic forgetting and limited adaptability across disparate chemical tasks[32–34]. To resolve this shallow mapping, reinforcement learning is introduced to enhance logical reasoning capabilities[35–38]. However, the reasoning process in these models is frequently dominated by token-intensive, low-level tasks such as valency checks and structural parsing[39,40]. This

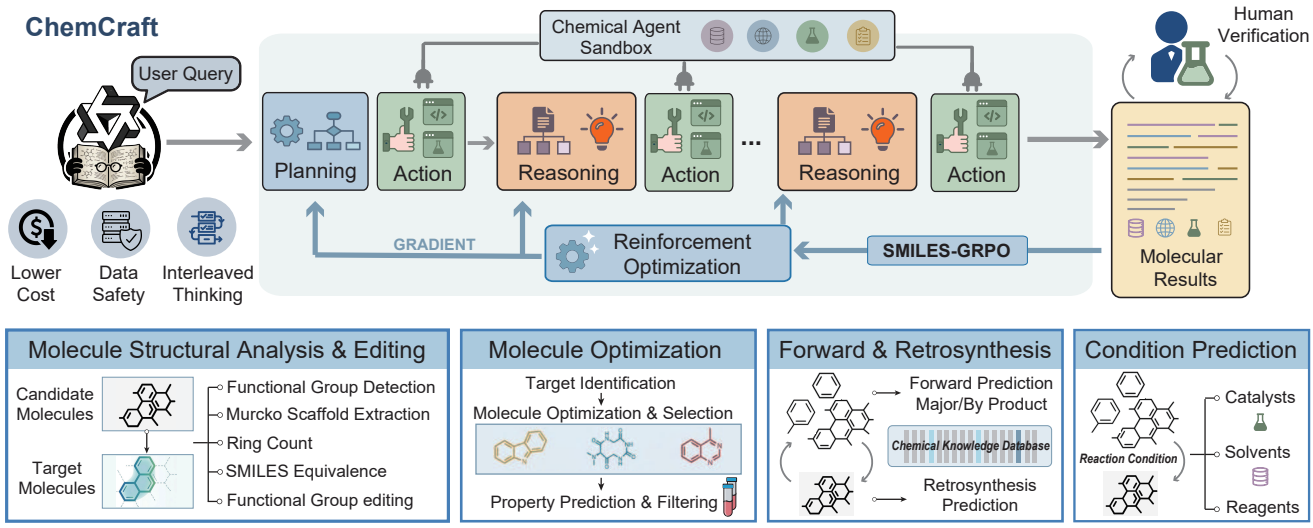


Figure 1. The ChemCRAFT framework. Top: ChemCRAFT comprises a interleaved reasoning process through planning, reasoning, and action steps. By interfacing with a *Chemical Agent Sandbox*, the framework offloads the burden of chemical knowledge retrieval to chemical tools, enabling the model to prioritize multi-step reasoning while mitigating catastrophic forgetting. **Bottom:** ChemCRAFT covers chemical discovery and design tasks including molecule analysis, manipulation, reaction forward prediction, retrosynthesis, and reaction condition prediction.

strategy is inefficient as it diverts the limited context window of the model away from high-level chemical derivation toward basic syntactic validation[41,42].

Alternatively, the *system-centric paradigm* designs multi-agent systems that integrate chemical tools with LLM-APIs through prompt-based guidance to solve scientific problems[43–46]. This paradigm is training-free and allows for rapid deployment, yet it introduces significant bottlenecks regarding privacy risks and high operational costs associated with commercial APIs[47–49]. More critically, these systems remain static because they lack the mechanism for gradient backpropagation[50]. Without the ability to update parameters based on scientific observations or tool feedback, the system cannot evolve its underlying policy or automatically optimize parameters to improve performance in complex chemical domains.

To resolve these challenges, we introduce ChemCRAFT, a locally deployable chemical agent framework that trains compact language models (7B) to reason with chemistry agents. As shown in Figure 1, inspired by human scientific practice, where chemists combine internal chemical principles with literature search, computational simulation, and database retrieval, ChemCRAFT uses the language model as a central scientific reasoner and delegates verifiable domain operations to a Chemical Agent Sandbox comprising computational, predictive, and retrieval-based tools[41,51,52]. This design bridges the two existing paradigms: it avoids the catastrophic forgetting and shallow input-output mapping of knowledge-injection training, while overcoming the static, prompt-only nature of tool-wrapped commercial agents by making tool orchestration a trainable local policy. To enable this policy, we construct ChemToolDataset, the first large-scale chemical tool-integrated trajectory dataset, where reasoning steps are interleaved with tool calls and sandbox observations. ChemCRAFT is then optimized through a two-stage training paradigm: cold-start training establishes the basic Planning→Action→Reasoning behavior, and SMILES-GRPO further improves tool-use strategies with chemistry-aware rewards covering SMILES validity, molecular understanding, editing fidelity, property improvement, and optimization success. Through this agentic training paradigm, chemical expertise is acquired from validated tool interaction while preserving the backbone model's general reasoning prior.

To rigorously validate the application potential of our framework across diverse chemical domains, we employ ChemCoT-Bench[53], a comprehensive evaluation suite designed to assess the multi-step reasoning and tool-use capabilities of chemical agents. The evaluation covers 22 critical tasks, ranging from fundamental molecular structure understanding and editing to molecular property optimization and reaction-related tasks, which provides a holistic view of an agent's proficiency. Extensive evaluations reveal the effectiveness and performance of ChemCRAFT among open-source models. Remarkably, despite its compact parameter size, it demonstrates problem-solving capabilities that not only significantly surpass the baselines of similar scales but also rival, and in specific reasoning-intensive tasks, exceed those of leading commercial LLM APIs. Beyond standardized metrics, we further assess the model's utility in realistic drug discovery scenario, focusing on the structural refinement and optimization of fentanyl analogs and pro-inflammatory GSTO1 inhibitors. ChemCRAFT successfully navigate the chemical space to propose a viable optimization strategy. Notably, the generated synthesis pathway and structural modifications exhibit high fidelity to the verified protocols reported in the Journal of Medicinal Chemistry. This alignment serves as strong proof of

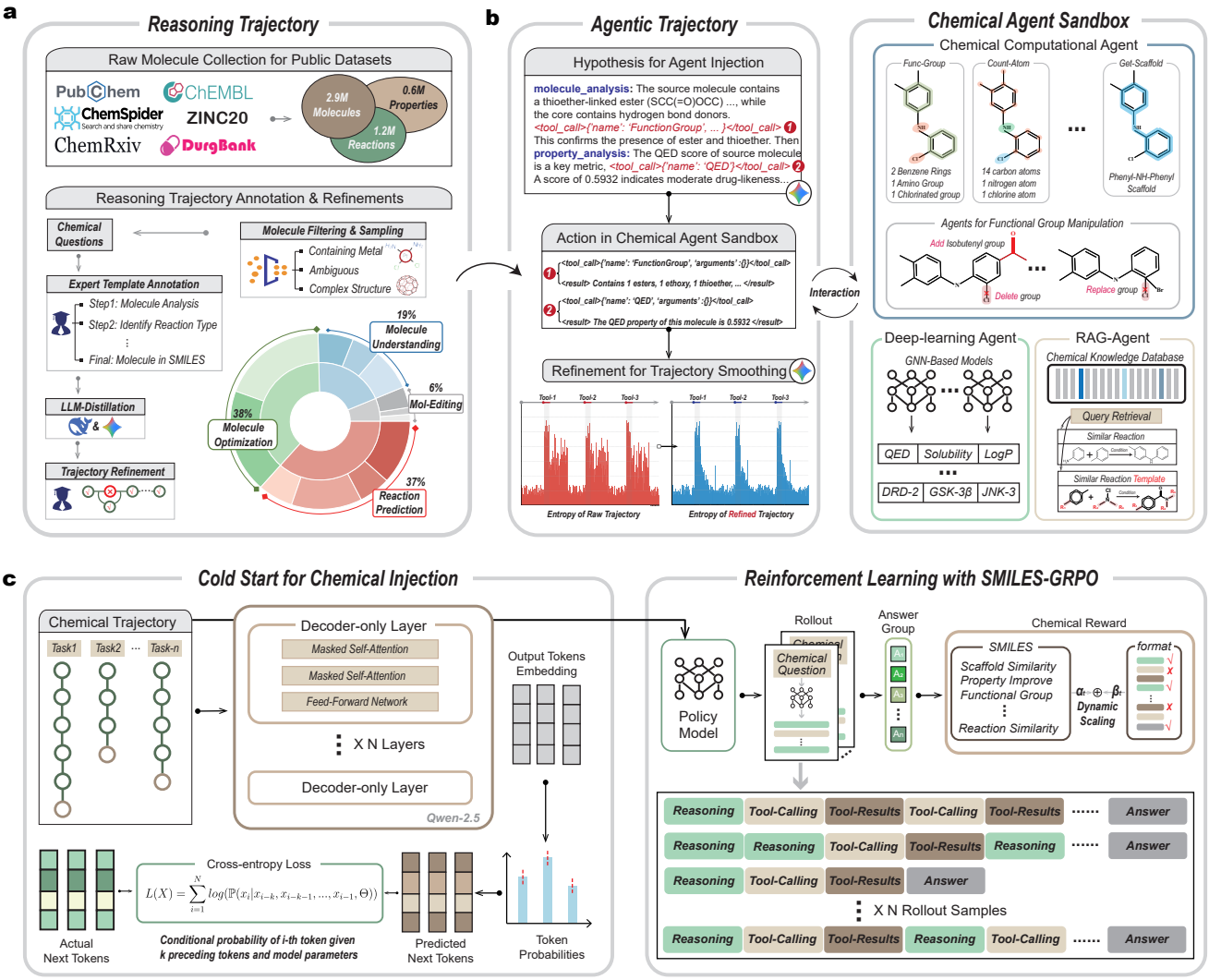


Figure 2. Overview of the Data-curation and Training Pipeline. **a** The annotation and refinement pipelines of the pure reasoning trajectory. **b** The agentic reasoning trajectory dataset is built based on the reasoning trajectory with the interaction of the Chemical Agent Sandbox. **c** The two-stage training scheme empowers ChemCRAFT to leverage diverse tools for addressing complex chemical tasks. The first stage involves cold-start training using token-level loss to enable the model to memorize the tool invocation format. The second stage employs SMILES-GRPO to enhance the model's agent-calling ability.

concept, demonstrating that our data-centric, tool-integrated paradigm can effectively replicate the decision-making process of human medicinal chemists in complex, open-ended research tasks.

2 Results

2.1 Framework Construction

To operationalize the Cognitive Decoupling paradigm, we established a systematic pipeline for data curation and model training, as illustrated in Figure 2a and 2b. This pipeline transforms raw chemical data into high-quality chemical reasoning trajectories with fine-grained interaction with our chemical agent sandbox. Furthermore, to subsequently align the model's capability with the high-quality chemical reasoning and agent-interaction, Figure 2c introduces our two-stage training process involving the cold-start supervised finetuning and the chemical reinforcement learning.

Task Formalization: Mapping Chemical Challenges with ChemCoTBench. Our data curation begins with the aggregation of raw molecular structures and reaction data from public repositories, including PubChem, ChEMBL, and ZINC. However, simply feeding models with raw data is insufficient for cultivating scientific intelligence, as current evaluations often reduce chemistry to simple factual recall or property prediction, neglecting the step-by-step operational reasoning required for real-

world discovery. To bridge this gap and provide a rigorous grounding for our learning objectives, we align our framework with ChemCoTBench. Unlike traditional benchmarks, ChemCoTBench systematically decomposes the chemical discovery process into 4 major tasks and 22 subtasks, establishing a direct mapping between fundamental reasoning and high-stakes practical applications. Specifically, it connects foundational Molecule Understanding and Editing to downstream challenges: Molecule Optimization serves as a proxy for de novo drug design (e.g., optimizing binding affinity for targets like DRD2 or GSK3-β), while Reaction Prediction (including forward prediction and retrosynthesis) underpins automated chemical synthesis planning. By decomposing these complex challenges into explicit sequences of modular chemical operations (e.g., addition, deletion, substitution), ChemCoTBench defines the fundamental logic for our work, converting abstract chemical challenges into actionable, traceable reasoning steps that mirror the rigorous workflows of professional chemists.

Innovating Data Construction: Tool-Integrated Trajectories. Traditionally, constructing reasoning datasets relies on distilling traces from powerful closed-source LLMs by prompting them to mimic expert thought patterns. However, this static textual approach fails to capture the interactive, hypothesis-driven nature of the modern scientific workflow. To transcend this limitation, we propose a novel tool-integrated reasoning trajectory construction method, as visualized in Figure 2a and 2b.

First, to ground the model's reasoning in chemical calculation tools, we engineer the Chemical Agent Sandbox, a high-performance execution environment. As shown in Figure 2b, our Chemical Agent Sandbox includes essential chemical computations (RDKit structure parsing, property prediction, reaction retrieval, etc) that are refactored into scalable microservices. This sandbox encapsulates domain complexity, allowing the LLM to drive professional-grade chemical calculations via standardized API interfaces. This design effectively decouples the reasoning from chemical knowledge and calculations, preventing the model from struggling with error-prone internal arithmetic and allowing it to focus on high-level strategy.

Building upon this infrastructure, we formulate the agent-interacted reasoning trajectories via a Hypothesis-Action-Refinement pipeline (Figure 2b). Traditional methods[54,55] often naively prepend tool interactions to reasoning traces, leading to two critical issues: (1) tool calls appearing at positions that defy natural scientific logic, and (2) feedback failing to inform subsequent reasoning. Such disjointed trajectories exhibit significant semantic turbulence, as evidenced by the irregular probability distributions in the entropy of raw trajectory analysis (Figure 2b).

To resolve these discrepancies, our pipeline decouples the construction process into three strategic phases. In the Hypothesis step, a teacher model identifies critical information gaps within the original non-agent trajectory and determines the optimal insertion points for tool calls to ensure logical consistency. During the Action step, the Chemical Agent Sandbox executes the identified operations to generate precise interaction results. Finally, in the Refinement step, the entire interaction history is reinjected into the teacher model to rewrite the reasoning process. This reflective mechanism smooths the semantic transitions between textual reasoning and tool outputs. As illustrated by the entropy analysis in Figure 2b, this process effectively converts rigid API logs into fluid, expert-level scientific narratives, where the agent interprets evidence and adjusts its strategy in a manner consistent with human domain experts.

The Two-Stage Training Paradigm. Leveraging these curated trajectories, we employed a progressive training strategy in Figure 2c to instill both chemical knowledge and agentic behavior. **Stage 1: Cold-Start SFT:** We first subjected Qwen2.5-7B as the base model to supervised fine-tuning using the synthesized agentic trajectories. This Cold Start phase is critical for initializing the model's understanding of chemical syntax and establishing the basic Planning→Action→Reasoning behavioral pattern, providing a stable policy foundation for subsequent optimization. **Stage 2: Reinforcement Learning with Chemical-Aware Rewards:** To transcend mere imitation and achieve robust problem-solving, we advanced to a reinforcement learning framework utilizing Group Relative Policy Optimization (GRPO)[56]. Departing from generic scalar feedback mechanisms, we engineered a multidimensional chemical reward function designed to evaluate the scientific validity of the entire reasoning chain; this signal rigorously integrates structural validity via exact SMILES matching and scaffold similarity, functional fidelity through the alignment of functional groups and reaction templates, and optimization success based on the magnitude of property improvement (e.g., ΔLogP, ΔQED). By optimizing against these granular chemical metrics rather than simple text overlap, the model learns to strategize its tool usage and reasoning path to maximize scientific accuracy.

2.2 Experimental Analysis on Chemical Domains

Our goal is to evaluate ChemCRAFT's utility in chemical research from fundamental structural parsing to the intricate synthesis planning. Thus, by utilizing the ChemCoTBench framework, we subject our ChemCRAFT models to a rigorous assessment across fundamental molecular understanding, property-guided optimization, and reaction prediction, benchmarking them against premier reasoning-enhanced LLMs and domain-specific baselines. Our empirical analysis reveals that by integrating precise interactions with the chemical agent sandbox, a smaller agentic LLM can effectively decouple logical reasoning from factual retrieval. This integration mitigates knowledge-induced hallucinations and prevents the degradation of reasoning capabilities typically observed when models struggle with internalizing vast chemical datasets, achieving expert-level performance in rigorous scientific tasks.

Models	Func-Group		Scaffold		SMILES	Molecule-Edit		
	FG↓	Ring↓	Murcko↑	Ring-sys↑	Eq.↑	Add↑	Delete↑	Sub↑
Gemini-2.5-Pro-think	0.11	0.60	0.51	87.5	82	**100.0**	85.0	81.7
Claude3.7-sonnet-think	0.21	1.60	0.40	80.0	84	85.0	80.0	**83.4**
DeepSeek-R1	0.27	1.55	0.34	45.0	65	70.0	70.0	68.3
O3-Mini	0.13	0.60	0.39	75.0	78	65.0	55.0	80.0
DeepSeek-V3	0.15	1.50	0.24	76.7	77	70.0	75.0	76.7
Qwen3-235B-A22B	0.42	1.00	0.34	82.5	75	40.0	75.0	66.7
Qwen2.5-32B-Instruct	0.36	0.65	0.12	53.3	62	50.0	50.0	48.3
ChemCrow	0.33	0.55	0.31	52.0	59	42.8	45.4	49.0
SciToolAgent	0.91	1.75	0.14	27.3	89	0.0	50.0	80.0
ChemCRAFT	**0.06**	**0.15**	0.26	**100.0**	**97**	90.0	**100.0**	73.0

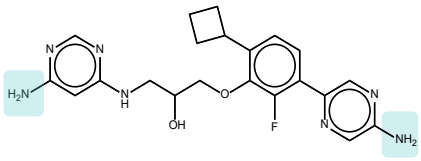
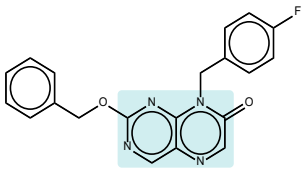
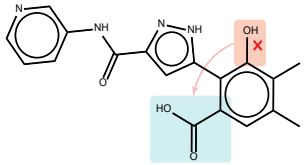
Function-Group-Detection	**Ring-System-Detection**	**Molecule-Edit-Substitute**
		
Giving you an input molecule, help me count the **amino group** number in the molecule.	Determine whether the ring system is in the target molecule.	Help me delete the **hydroxyl** group and add **carboxyl** group.

Table 1. Performance Comparison for Molecule Structural Understanding and Editing. Results demonstrate that ChemCRAFT consistently outperforms specialized biochemical agents (ChemCrow and SciToolAgent) and general LLMs (Gemini, Claude, etc) across most subtasks. Selected case studies are provided to demonstrate the model's performance on key representative tasks.

Molecular Structural Understanding & Editing. Fundamentally, an AI chemist must accurately perceive and manipulate molecular structures before it can engage in complex design. To rigorously evaluate this, we assessed models on Molecule Understanding, which demands the precise recognition of functional groups (critical for reactivity) and ring systems (the stable building blocks of drug scaffolds). We quantified performance using Mean Absolute Error (MAE) for counting tasks and Tanimoto similarity for structural scaffold extraction. As shown in Table 1, commercial general LLMs often suffer chemical hallucination, misinterpreting the linearized SMILES syntax and failing to map 1D strings to 2D topologies. In contrast, ChemCRAFT achieves near-perfect precision. For example, in the Function-Group-Detection task (Table 1, bottom-left), where the model detects the functional groups in a molecule architecture, ChemCRAFT achieves an MAE of 0.06, significantly outperforming general LLMs like Gemini-2.5-Pro (MAE 0.11). This perceptual acuity extends to hierarchical structural reasoning; in the Ring-System-Detection task (Table 1, bottom-middle), which requires determining the existence of specific fused rings, our model attains 100% accuracy compared to 87.5% for Gemini-2.5-Pro[57].

Crucially, these understanding capabilities can be translated into actionable Molecule Editing capabilities (Add, Delete, Substitute, etc) that serve as the fundamental operation of chemical design. In the Molecule-Edit-Substitute visualization (Table 1, bottom-right), the challenge involves precisely targeting a functional group for removal and attaching a functional group at the exact locus without disrupting the surrounding scaffold. ChemCRAFT leverages its agent sandbox to validate valency and connectivity, achieving a 100.0% success rate in deletion tasks and a 73% success rate in substitution tasks. These results show that ChemCRAFT goes beyond simple text generation by grounding its editing process in chemical constraints, allowing for rigorous and valid molecular transformations.

Navigating the Optimization Landscape. Moving beyond static editing, we challenged the models with the open-ended task of Molecule Optimization—a proxy for de novo drug design where the goal is to modify a lead compound to maximize specific properties while retaining structural validity. We evaluated performance across two tiers of difficulty: (1) Physicochemical Properties (LogP, QED, Solubility), which follow defined heuristic rules, and (2) Protein-Ligand Binding Affinities (DRD2, GSK3-β, JNK3), which require inferring complex bio-interactions. Performance was quantified by the Property Improvement score (Δ), measuring the net gain over the starting molecule.

As illustrated in the density plots of Figure 3a, general LLMs (e.g., Qwen-2.5-7B[58], Deepseek-R1) often struggle to escape local minima, producing distributions centered near zero ($\Delta \approx 0$). In contrast, ChemCRAFT (red distribution) consistently shifts the optimization trajectory toward positive gains. For instance, in Solubility optimization, ChemCRAFT achieves a

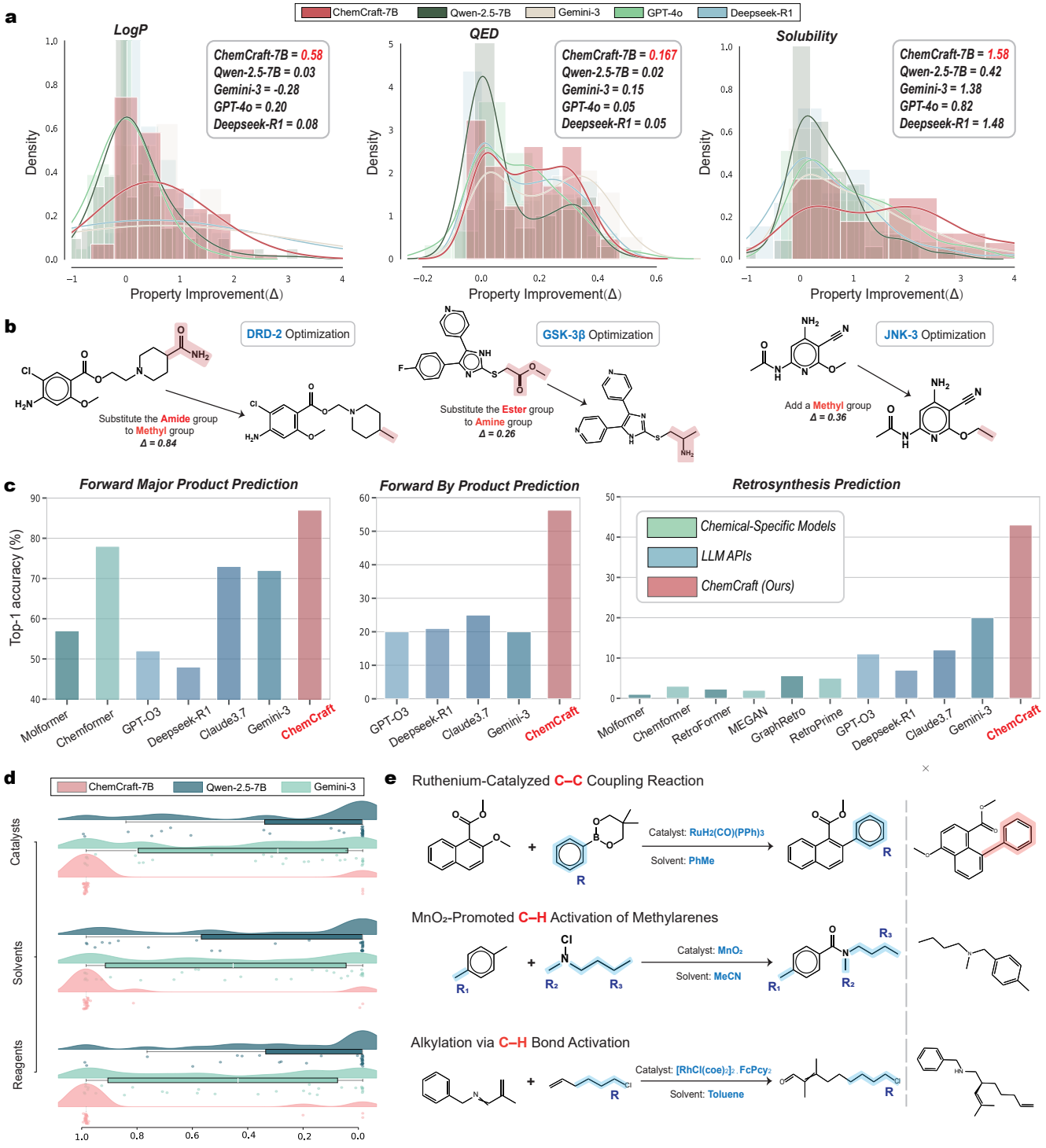


Figure 3. Experimental Analysis for Molecular Optimization and Reaction Predictions. **a** Distribution analysis for property improvements on LogP, QED, and Solubility. **b** Optimized molecule visualizations for protein-activated properties, including DRD2, JNK3, GSK-3β. **c** Comparison between ChemCRAFT, chemical-specific models, and LLM-APIs on forward major / by product prediction and retrosynthesis prediction. **d** Similarity distribution for reaction condition recommendations. **e** Reaction prediction visualizations on complex reaction types.

mean improvement of $\Delta = 1.58$, nearly quadrupling the performance of the baseline Qwen model ($\Delta = 0.42$) and surpassing commercial reasoning models like Gemini-2.5-Pro ($\Delta = 1.38$).

This statistical success translates into chemically meaningful design strategies, as visualized in the drug-target case studies in Figure 3b. Here, the model effectively deploys the modular "Add/Delete/Substitute" operations validated in our Editing tasks to solve complex biological constraints. In the GSK-3β case (middle), the model strategically substitutes an ester group with an amine ($\Delta = 0.26$). This is a non-trivial bioisosteric replacement that significantly alters the hydrogen bond donor/acceptor profile to fit the binding pocket, demonstrating that the model is not merely randomizing atoms but applying medicinal chemistry logic. In the DRD-2 case (left), it executes a precise scaffold-hopping maneuver, replacing a flexible amide linker with a methyl group ($\Delta = 0.84$) to potentially lock the conformation. Crucially, these optimizations are achieved with high sample efficiency. Unlike traditional genetic algorithms that may require thousands of oracle calls (property evaluations), ChemCRAFT reaches these high-value candidates within minimal interaction steps. This efficiency suggests that our agent does not rely on brute-force search, but rather enables a Hypothesis-Action-Verification loop that is practical for real-world campaigns where wet-lab feedback is costly.

Predicting Complex Reaction Outcomes. To evaluate chemical intuition at the systemic level, we assessed models on three pillars of reaction prediction: (1) Forward Prediction, encompassing both Major Products and, crucially, By-Products (vital for assessing impurity profiles and purification risks); (2) Retrosynthesis (w/o reaction type), testing the ability to deconstruct targets into available precursors; and (3) Condition Recommendation, requiring precise suggestions for reagents, solvents, and catalysts. We benchmarked against domain-specialized models (e.g., Chemformer[59], GraphRetro[60], RetroPrime[61]) and general reasoning LLMs (e.g., GPT-o3, DeepSeek-R1[56]).

As shown in Figure 3c, while specialized models like Chemformer achieve competitive accuracy on Major Product prediction due to extensive pre-training on USPTO data, they significantly underperform on By-Product Prediction and Retrosynthesis on out-of-distribution samples. We attribute this to architectural rigidity: sequence-based and graph-based expert models are typically optimized for single-outcome generation and often rely on precise atom-mapping. This dependency limits their generalization to noisy or unmapped real-world inputs. In contrast, ChemCRAFT achieves a top-1 accuracy of over 50% in by-product prediction and dominates the retrosynthesis task (40% + vs < 20% for baselines). Furthermore, in Condition Recommendation (Figure 3d), the violin plots reveal that while general LLMs suffer from recommending convincing reaction conditions (distributions centered near 0.0), our ChemCRAFT produces highly concentrated predictions aligned with ground truth, particularly for complex catalysts.

The model's superior grasp of reactivity is further evidenced in the case studies (Figure 3e). We selected challenging modern reactions involving non-polar bond activation (C-C, C-H, C-F), transformations that are underrepresented in standard patent data (USPTO) and typically confound expert models. In the Ruthenium-catalyzed C-C coupling case[62] (top), Gemini-2.5-Pro misses the competitive advantage of C-O activation over remote C-H activation. In the MnO_2 oxidation case[63] (middle), it underestimates the oxidation depth, predicting an amide rather than the correct carbonyl product. In the Cobalt-catalyzed C-F activation[64] (bottom), the baseline fails to recognize that the specific additives are designed to suppress the benzyne pathway, leading to a regio-isomer error. ChemCRAFT correctly predicts these outcomes not by mere intuition, but by evidence-based reasoning: its sandbox retrieves analogous reaction templates from the library, allowing the agent to ground its mechanism in empirical precedent rather than hallucinating plausible but incorrect pathways.

2.3 Algorithmic Validation: Efficiency, Generalization, and the Power of Agentic RL

We posit that the superior performance of ChemCRAFT transcends mere domain data exposure, stemming instead from a cognitive decoupling architecture refined via reinforcement learning. To rigorously verify this hypothesis, we conducted a comprehensive ablation analysis evaluating training dynamics, generalization capabilities, and inference efficiency, as summarized in Figure 4.

The Necessity of Agentic Reinforcement Learning. The limitations of standard SFT become evident when analyzing the training trajectories in Figure 4b. Models trained via "Raw SFT" without tool integration plateau early, as parameters alone are insufficient for precise scientific computation. While introducing tool-use trajectories ("Tool-Augmented SFT") resolves basic syntax errors and boosts performance, these models effectively learn the *format* of tool usage rather than the *strategy*. This is where the necessity of our agentic RL paradigm becomes undeniable. As illustrated in the right panel of Figure 4b, while the SFT baseline stagnates, the RL-driven training curve continues to climb, propelling the model toward a higher performance ceiling. This gain is mechanistically explained by Figure 4c: the RL-trained model exhibits a statistically significant increase in tool-calling frequency for complex subtasks like molecule structure analysis and reaction-related tasks. Unlike the SFT model, which remains a passive tool-user, the RL agent actively queries the sandbox to verify hypotheses, demonstrating a self-correction behavior where suboptimal initial actions are autonomously refined based on feedback.

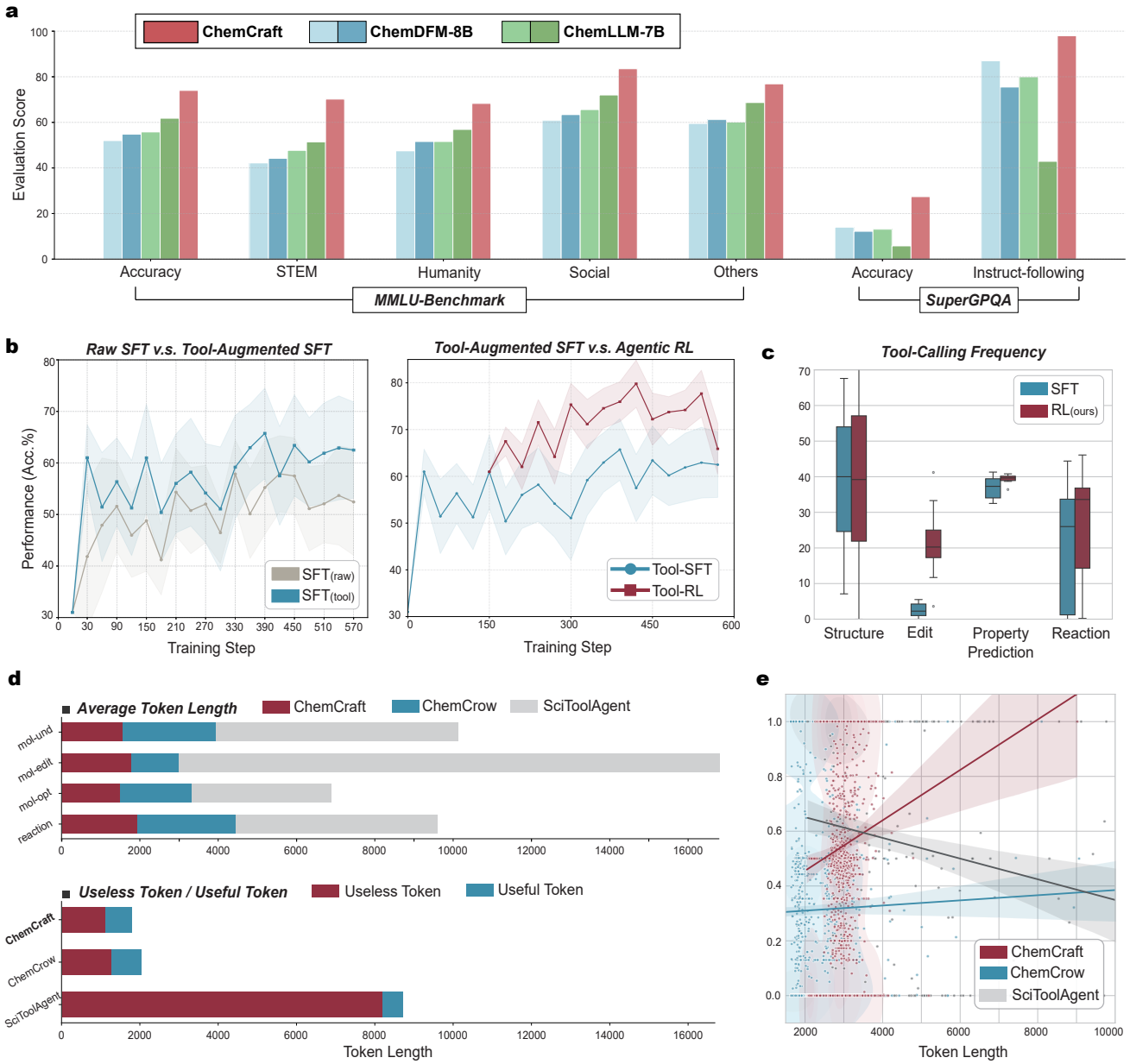


Figure 4. **a** Our ChemCRAFT outperforms other chemical language models on general reasoning domains, showing our training paradigm effectively preserves the foundational reasoning capabilities and knowledge of LLM backbone. **b** Performance curves analysis for non-tool-calling SFT, tool-augmented SFT, and Agentic-RL. **c** Tool-calling frequency analysis between the SFT and Agentic-RL. The RL training paradigm boost the model's interaction with the Chemical Sandbox. **d** ChemCRAFT has low inference cost and a more useful token length compared to other chemical multi-agent systems (ChemCrow, SciToolAgent). **e** Performance and token-length correlation analysis shows that ChemCRAFT's reasoning accuracy positively correlates with token cost; other baselines show redundant tokens with no performance gain.

Defying Catastrophic Forgetting. A pervasive risk in training specialized scientific models is the degradation of general reasoning capabilities, known as catastrophic forgetting. Figure 4a demonstrates that ChemCRAFT effectively circumvents this trade-off. While domain-adapted models like ChemLLM often compromise general capabilities due to rigid knowledge injection, ChemCRAFT preserves robust performance across general benchmarks. It achieves strong results on MMLU (STEM, Humanities, and Social Sciences) and SuperGPQA, significantly outperforming baselines. We attribute this robustness to the nature of our agentic training paradigm: rather than forcing the model to memorize static chemical knowledge or specific response templates via standard SFT, our approach instills a universal problem-solving methodology—analyzing the query, formulating a step-by-step plan, and orchestrating appropriate tools. This ensures that the model preserves its intrinsic cognitive structures for logic and analysis, allowing specialized chemical proficiency to coexist with, rather than overwrite, general

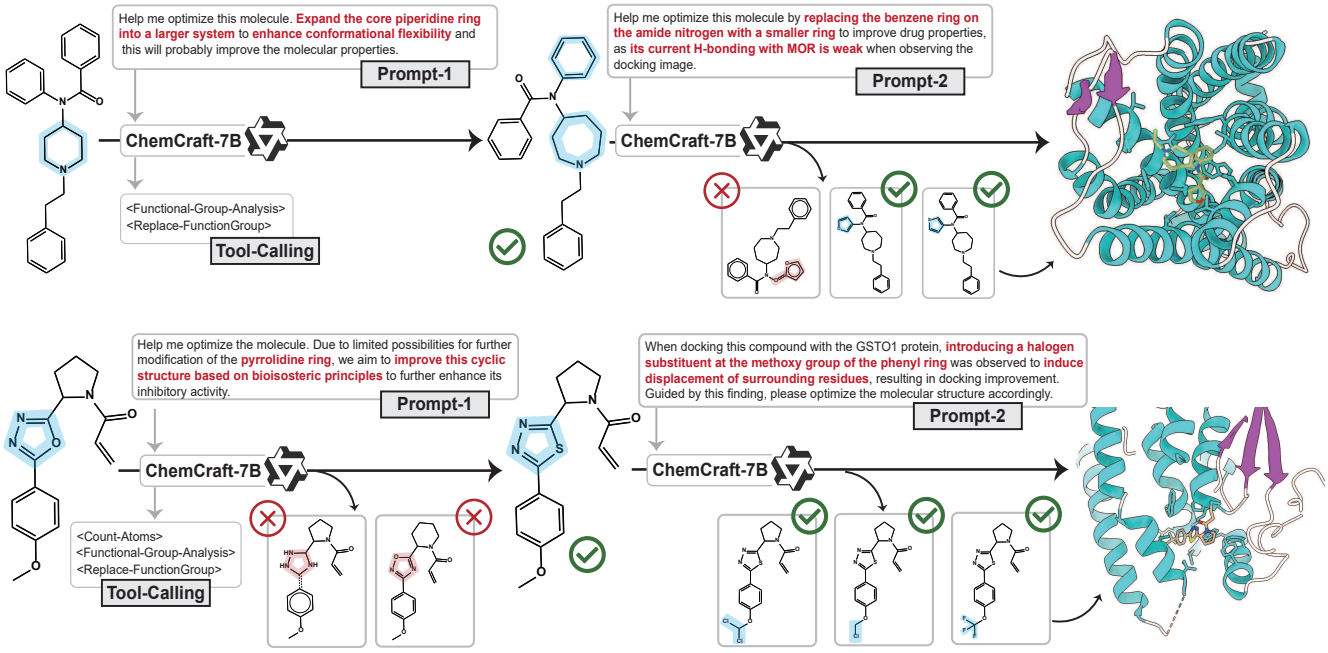


Figure 5. ChemCRAFT's assessment on real-world drug optimization. We demonstrate the robust optimization capabilities of ChemCRAFT on two prominent drug scaffolds: (a) Fentanyl analogues and (b) pro-inflammatory GSTO1 inhibitors, both recently featured in *the Journal of Medicinal Chemistry*. By prompting the model with the original medicinal chemists' rationales, ChemCRAFT successfully replicates the reported optimization trajectories. The docking poses within the target protein binding sites further validate the structural integrity of the optimized molecules from ChemCRAFT.

intelligence.

The Efficiency of a Unified Agent Architecture. Finally, we address the computational feasibility of AI chemists. Existing Multi-Agent Systems (e.g., ChemCrow[43], SciToolAgent[65]) rely on massive commercial models and extensive inter-agent dialogue (e.g., hand-offs between a "Planner," "Critic," and "Executor"), resulting in prohibitive token costs. ChemCRAFT internalizes these roles into a single, streamlined reasoning stream. The impact of this architectural shift is quantified in Figure 4d: ChemCRAFT reduces the average token length by approximately 65% compared to SciToolAgent while maintaining a higher density of "Useful Tokens". Figure 4e presents a normalized correlation analysis, where task metrics are mapped to a $[0, 1]$ scale to visualize the return on inference compute. The trend line illustrates an ideal reasoning efficiency, where increased token expenditure (longer CoT) linearly translates to performance gains. ChemCRAFT closely adheres to this trajectory, demonstrating effective scaling of test-time compute. In contrast, multi-agent frameworks like SciToolAgent and ChemCrow diverge significantly, suffering from excessive token overhead with diminishing performance returns. This places ChemCRAFT in the optimal quadrant of high performance and low inference cost. The efficiency breakthrough resolves the data privacy and latency bottlenecks inherent to cloud-based solutions, making it feasible to deploy expert-level AI chemists on local hardware.

2.4 Assessment on Real-World Drug Optimization Scenarios

To demonstrate the utility of ChemCRAFT in a realistic pharmaceutical setting, we conducted a blind case study focused on the structural optimization of key drugs whose optimization strategies have been reported in the Journal of Medicinal Chemistry. These strategies, proposed and validated by expert chemists, serve as benchmarks for validating the outputs generated by our ChemCRAFT. As shown in Figure 5, ChemCRAFT successfully identifies optimization paths consistent with those validated by medicinal chemists. The high-quality docking results of these refined molecules demonstrate that ChemCRAFT can serve as a dependable co-pilot in practical pharmaceutical design, producing both chemically rational and biologically plausible leads.

Fentanyl Analogues Optimization. Fentanyl is a potent synthetic opioid with highly effective analgesic properties, yet its safety profile is problematic due to a high risk of causing respiratory depression and fatal overdose. The structural optimization of fentanyl aims to design analogs with improved therapeutic indices (reduced toxicity and addiction potential), while preserving their analgesic potency. As shown in the top case of Figure 5 , the optimization process contains two stages:

(1) Stage-1: Conformation Flexibility Improvement. As reported in the related journal[66], the core piperidine ring imposes significant conformational constraints on the seed molecule. To address this, we initially targeted this scaffold for refinement.

By prompting ChemCRAFT with the medicinal chemist's rationale, the model successfully sampled the optimized molecule in a single inference step.

(2) Stage-2: Amide Nitrogen Optimization. As reported in the related journal[66], replacing the benzene ring on the amide nitrogen with a smaller ring to improve drug properties, as its current H-bonding with MOR is weak when observing the docking image. Upon receiving the strategic prompt, ChemCRAFT produced multiple distinct five-membered ring analogs via iterative sampling. Each generated structure (Furan, Thiophene) mirrors the expert-designed molecules from the study, validating ChemCRAFT's ability to serve as an effective and creative assistant in drug discovery.

GSTO1 Inhibitor Optimization. GSTO1-specific covalent inhibitors serve as pivotal probes and potential therapeutics for modulating NLRP3-driven inflammatory responses, yet their clinical translation is often hindered by excessive electrophilicity, leading to off-target toxicity and poor metabolic stability. The structural optimization of these inhibitors aims to discover novel chemotypes with improved selectivity and drug-like properties, while maintaining potent target engagement with the GSTO1 active site. As shown in bottom case of Figure 5, the optimization process contains two stages:

(1) Stage-1: Cyclic structure Improvement. As reported in the related journal[67], the pyrrolidine ring structure has limited modification possibilities. By prompting ChemCRAFT with the medicinal chemist's rationale, ChemCRAFT successfully sampled the optimized molecule after multiple inference steps.

(2) Stage-2: Docking Optimization. As reported in the related journal articles[67], when docking this compound with the GSTO1 protein, introducing a halogen substituent at the methoxy group of the phenyl ring was observed to induce displacement of surrounding residues, resulting in docking improvement. Based on this chemist's rationale, ChemCRAFT produced multiple optimized candidates with a halogen substituent. The docking scene further demonstrates the ChemCRAFT's optimization effectiveness.

Human-AI Synergy in Practical Drug Design. It is crucial to contextualize the role of ChemCRAFT within the broader paradigm of AI-aided drug discovery. As illustrated by the prompts in Figure 5, the initial strategic hypotheses—such as proposing a bioisosteric replacement or incorporating a halogen substituent—are guided by the intuition of human medicinal chemists. We emphasize that ChemCRAFT is designed not to autonomously replace these high-level strategic leaps, but rather to function as an expert-level collaborative copilot. Traditional LLMs often struggle to translate abstract human directives into valid structural modifications without suffering from chemical hallucination. In contrast, ChemCRAFT excels at flawlessly executing human-guided intents: it autonomously navigates the granular chemical space, orchestrates the sandbox to strictly validate valency and topological constraints, and efficiently screens for optimal binding poses. This human-in-the-loop synergy ensures that the molecular outputs are not merely computational artifacts, but synthetically viable candidates deeply grounded in practical medicinal chemistry.

3 Discussion

ChemCRAFT introduces a cognitive decoupling paradigm that fundamentally shifts the role of language models in chemistry from knowledge containers to the scientific reasoner. By separating high-level reasoning from deterministic chemical calculations via an agent sandbox and optimizing this process through reinforcement learning, we demonstrate that a compact, locally deployable 7B model can surpass the performance of massive, cloud-based LLMs in complex chemical tasks. This framework effectively resolves the deployment trilemma of performance, cost, and privacy, providing a secure and scalable solution for sensitive biochemical research. However, while ChemCRAFT establishes a robust chemical assistant, its current implementation faces specific structural limitations and potential unintended risks that warrant further exploration. Thus, we discuss the boundaries of our current approach and propose strategic paths for mitigation.

The future evolution of ChemCRAFT is based on three critical dimensions of scaling. **(1).** Tool-level Scaling. Tool-integrated scaling is essential to expand the chemical sandbox beyond basic informatics toward more advanced computational and search-based microservices, enabling the model to orchestrate more complex, multi-step scientific workflows. **(2).** Data-level Scaling. Data and policy scaling must be optimized to enhance the efficiency of reinforcement learning; this includes exploring curriculum-based sampling to manage task difficulty and investigating the underlying interactions—both synergistic and antagonistic—between disparate chemical reasoning objectives. **(3).** Modality-level Scaling. Modality scaling represents a vital frontier for expanding the model's domain. While textual representations are effective for small molecules, transitioning toward multimodal backbones that incorporate 3D protein structures and visual coordinate data is necessary to address high-stakes challenges in protein-ligand design and large-molecule therapeutics. Addressing these frontiers will be central to developing a more universal, secure, and physically-grounded chemical intelligence.

4 Methods

We present a unified framework designed to empower small language models with autonomous reasoning and operational capabilities in the chemical domain. As illustrated in Figure 2, our methodology is orchestrated across three interconnected pillars. First, to overcome the limitations of static knowledge storage, we construct a Chemical Agent Sandbox, an interactive environment that externalizes domain expertise through computational[68], deep learning, and retrieval-based[69,70] tools. Second, to bridge the gap between abstract chemical knowledge and actionable problem-solving, we introduce a pipeline for constructing High-Quality Reasoning Trajectories, incorporating tool-integrated narratives and reflective refinement. Finally, we implement a Two-Stage Training Paradigm, which progressively transitions the model from supervised behavioral cloning to robust reinforcement learning, optimizing for scientific validity and rigorous chemical constraints via the SMILES-GRPO mechanism.

4.1 Construction of the Chemical Agent Sandbox

To endow the language model with comprehensive domain expertise, we established a Chemical Agent Sandbox, an interactive environment designed to decouple chemical reasoning from static knowledge storage. This framework covers the full spectrum of the chemical discovery pipeline, ranging from molecular structure analysis and de novo design to property optimization and retrosynthetic planning. To achieve robust performance across these diverse tasks, we encapsulate domain-specific tools into three distinct categories of agents:

- **Computational Software Agents:** Leveraging robust cheminformatics libraries, primarily RDKit and RDChiral, these agents provide deterministic capabilities for molecular graph manipulation. They are responsible for verifying topological validity, analyzing stereochemistry, and executing precise structural editing, ensuring that the model's generated molecules adhere to strict chemical rules.

- **Deep Learning-based Agents:** Serving as the predictive intuition of the system, these agents interface with pre-trained models and frameworks like PyTDC. By providing rapid quantitative feedback on physicochemical properties (e.g., QED, LogP) and ADMET profiles, they guide the language model in navigating the high-dimensional chemical space during multi-objective optimization tasks.

- **Retrieval-based Agents:** To anchor synthesis planning in empirical reality, these agents maintain a curated repository of reaction equations and templates. Through substructure matching and similarity search, they enable the model to retrieve analogous reaction pathways and validate reaction conditions, promoting mechanistic feasibility in synthesis prediction.

4.2 Construction of High-Quality Reasoning Trajectory

High-quality trajectory construction is pivotal for enabling the model to interact effectively with the chemical sandbox and solve complex domain tasks. We approach this construction in two phases: first by structuring the abstract chemical space into traceable reasoning steps, and subsequently by integrating dynamic tool interactions through a Hypothesis-Action-Refinement protocol.

- **Generating Chemical Reasoning Trajectories**. To transcend simple factual recall, we align our data construction with ChemCoTBench, a framework that decomposes the chemical discovery process into 9 major tasks and 22 subtasks. This structure bridges fundamental capabilities, such as Molecule Understanding and Editing, with high-stakes downstream applications like Molecule Optimization (e.g., binding affinity improvement) and Reaction Prediction (e.g., retrosynthesis planning). By breaking down these complex challenges into explicit sequences of modular chemical operations (e.g., functional group addition or substitution), we convert abstract chemical problems into actionable, step-by-step reasoning scaffolds. This rigorous decomposition ensures the model learns the operational logic required for real-world discovery rather than mere property memorization.

- **Generating Tool-Integrated Reasoning Trajectories**. To capture the interactive nature of modern scientific workflows, we propose a Tool-Integrated Reasoning construction method based on the Hypothesis-Action-Refinement pipeline. First, we leverage the Chemical Agent Sandbox to decouple reasoning from calculation, allowing the LLM to offload error-prone computations (e.g., RDKit parsing, QED evaluation) to external microservices. Given an initial non-agent trajectory τ^0, the **Hypothesis** step asks a teacher model to inspect the current reasoning state, identify missing chemical evidence, and determine both the tool type and insertion position. This step prevents tool calls from being mechanically prepended to the trace and ensures that each call is motivated by an explicit scientific uncertainty. The **Action** step executes the selected operation in the Chemical Agent Sandbox and returns verified observations, such as structural validity, molecular properties, retrieved reaction templates, or predicted synthesis conditions. The **Refinement** step then reinjects the hypothesis, action, and observation into the teacher model, which rewrites the local reasoning segment so that the tool feedback causally supports the subsequent conclusion.

The refinement step is introduced to suppress entropy oscillation at the boundary between natural-language reasoning and tool feedback. For a trajectory $\tau = (y_1, \ldots, y_T)$ and a model distribution $p_t(v) = P_\theta(v \mid x, y_{<t})$ over vocabulary token $v \in \mathcal{V}$, the token-level predictive entropy is:

$$H_t(\tau) = -\sum_{v \in \mathcal{V}} p_t(v) \log p_t(v). \tag{1}$$

When a raw tool log is inserted without semantic smoothing, the model often exhibits abrupt entropy jumps around tool boundaries because the next token distribution switches between free-form chemical reasoning, rigid API syntax, and numerical observations. We visualize on Fig. 2(b) and quantify this instability with the boundary entropy oscillation:

$$\Omega_{\mathcal{B}}(\tau) = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} |H_{b+1}(\tau) - H_b(\tau)|, \tag{2}$$

where $\mathcal{B}$ denotes the set of transition positions between reasoning tokens, tool-call tokens, and observation tokens. A large $\Omega_{\mathcal{B}}$ indicates a non-smooth semantic distribution. In practice, such oscillation can cause unstable token-level supervision, high-variance gradients during SFT, spurious tool-call placement, and weak credit assignment in later reinforcement learning. It may also produce trajectories where the model copies an observation without explaining its chemical implication, leading to hallucinated or under-utilized tool feedback.

To mitigate this problem, the Refinement step rewrites the local context under two constraints: preserving the verified chemical result and reducing the abrupt semantic transition around tool boundaries. Conceptually, the refined trajectory τ^* is selected as:

$$\tau^* = \arg\min_{\tau} \left[\mathcal{L}_{\text{faith}}(\tau, \tau^0, o) + \gamma\, \Omega_{\mathcal{B}}(\tau) \right], \tag{3}$$

where o is the sandbox observation, $\mathcal{L}_{\text{faith}}$ penalizes deviation from the original task answer and verified tool result, and γ controls the smoothness strength. We implement this objective through teacher-model rewriting rather than gradient optimization: the teacher is prompted to preserve the chemical conclusion, explicitly cite the observation that supports it, and rewrite only the surrounding reasoning span. This transforms rigid tool logs into fluid, expert-level scientific narratives, where the agent interprets evidence, validates hypotheses, and dynamically adjusts its strategy, mirroring the cognitive process of a professional chemist.

4.3 Model Training: A Two-Stage Paradigm

To effectively instill both chemical domain knowledge and agentic reasoning capabilities into the model, we employ a progressive two-stage training paradigm. This approach first establishes a stable behavioral policy through supervised learning and subsequently refines the model's problem-solving strategies using reinforcement learning with chemistry-specific feedback.

4.3.1 Stage 1: Cold-Start Supervised Fine-Tuning

The primary objective of the "Cold-Start" phase is to initialize the model's understanding of chemical syntax and establish the fundamental "Think $\rightarrow$ Call Tool $\rightarrow$ Observe" behavioral pattern. We utilize the synthesized tool-integrated trajectories to perform Supervised Fine-Tuning (SFT) on the base model.

We treat the agentic interaction as a sequential generation task. Given an input prompt x and a target trajectory $y = (y_1, y_2, \ldots, y_T)$ consisting of interleaved reasoning thoughts, tool calls, and observations, we optimize the model parameters θ by minimizing the negative log-likelihood of the next token. The loss function $\mathcal{L}_{\text{SFT}}$ is defined as:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{SFT}}} \left[\sum_{t=1}^{T} \log P_\theta(y_t \mid x, y_{<t}) \right] \tag{4}$$

where $\mathcal{D}_{\text{SFT}}$ represents the curated dataset of tool-integrated trajectories, and $y_{<t}$ denotes the history of tokens preceding step t. This phase ensures that the model learns to strictly adhere to the formatting requirements of the *Chemical Agent Sandbox* and mimics the expert-level reasoning logic embedded in the training data, providing a robust policy initialization π_{ref} for the subsequent reinforcement learning stage.

4.3.2 Stage 2: Reinforcement Learning with SMILES-GRPO

While SFT provides a strong foundation, it is limited by the static nature of imitation learning. To transcend mere mimicry and enable the model to explore novel chemical spaces, we advance to a reinforcement learning framework. We adopt Group Relative Policy Optimization (GRPO), which eliminates the need for a value function approximation by normalizing advantages within a group of sampled outputs.

The GRPO Objective. For each chemical query q, we sample a group of G outputs $\{o_1, o_2, \ldots, o_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$. The optimization objective maximizes the surrogate objective while constraining the policy update via the KL-divergence. The GRPO loss function is formulated as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}} \left[\frac{1}{G}\sum_{i=1}^{G} \left(\min\left(r_i \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, r_i \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon \right) \right) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta || \pi_{\text{ref}}) \right) \right] \tag{5}$$

where ε is the clipping parameter, β controls the KL-divergence penalty to prevent policy collapse, and r_i is the advantage term, calculated by normalizing the rewards within the group: $r_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G})}$.

Multidimensional Chemical-Aware Reward. To evaluate generated trajectories beyond surface-form correctness, we design a task-conditioned chemical reward that couples response-format compliance with molecular validity and task-specific chemical objectives. For a query q and generated output o, the reward is defined as:

$$R(q,o) = \lambda_{format}\, R_{format}(o) + \lambda_{\text{val}}\, R_{\text{val}}(o) + \lambda_{\text{task}}\, R_{\tau(q)}(q,o), \tag{6}$$

where R_{format} measures adherence to the required reasoning and tool-calling format, R_{val} is the fraction of generated SMILES strings that are chemically valid under RDKit sanitization, and $R_{\tau(q)}$ denotes the reward associated with the task type $\tau(q)$. All primitive scores are normalized to $[0,1]$, and the non-negative coefficients control their relative importance. For the task-specific term, we instantiate $R_{\tau(q)}$ according to the chemical capability being trained:

$$R_{\tau(q)} = \begin{cases} \alpha_{\text{mae}}\, \exp\left(-\frac{1}{K}\sum_{k=1}^{K} \frac{|\hat{y}_k - y_k|}{\rho_k} \right) + \alpha_{\text{scaf}} \text{Tanimoto}(\text{FP}(\text{BM}(\hat{m})), \text{FP}(\text{BM}(m^*))), & \tau = \text{Understanding, RXN}, \\ \alpha_{\text{fg}}\, \dfrac{|\mathcal{G}^+ \cap \mathcal{G}(\hat{m})| + |\mathcal{G}^- \setminus \mathcal{G}(\hat{m})|}{|\mathcal{G}^+| + |\mathcal{G}^-|}, & \tau = \text{Editing}, \\ \alpha_{\Delta}\, \text{clip}\left(\dfrac{s_q\,[p(\hat{m}) - p(m_0)]}{\eta}, 0, 1 \right) + \alpha_{\text{succ}}\, \mathbb{I}[s_q\,[p(\hat{m}) - p(m_0)] \geq \delta], & \tau = \text{Optimization}. \end{cases} \tag{7}$$

For molecule understanding, $\hat{y}_k$ and y_k are predicted and reference numerical properties, ρ_k is a scale factor used to normalize MAE across properties, and scaffold preservation is measured by the Tanimoto similarity between fingerprints of Bemis–Murcko scaffolds $\text{BM}(\cdot)$. For molecule editing, $\mathcal{G}^+$ and $\mathcal{G}^-$ denote the functional groups that should be introduced and removed, respectively, and $\mathcal{G}(\hat{m})$ is the detected functional-group set in the edited molecule. For molecule optimization, m_0 is the starting molecule, $p(\cdot)$ is the target property, $s_q \in \{+1, -1\}$ specifies whether the property should be maximized or minimized, η normalizes the improvement magnitude, and the success indicator becomes the success rate after averaging over the evaluation set. This design yields chemically grounded feedback for numerical accuracy, scaffold consistency, editing rationality, property improvement, optimization success, and universal SMILES validity.

4.3.3 Details of Hyperparameter and Dataset Split

For the dataset configuration, we applied a systematic 2:8 random split to every subtask within the ChemToolDataset. This strategy allocated 20% of the samples to the cold-start phase and designated the remaining 80% for the subsequent reinforcement learning stage. All training procedures were conducted using the veRL framework (v0.5.0). Hyperparameters included a batch size of 32, a learning rate of 7e-7, and a total training duration of a single epoch.

Data Availability

ChemCoTDataset, the chemical reasoning dataset with well-annotated chain-of-thoughts trajectories, has been deposited in the huggingface repository: ChemCoTDataset. ChemToolDataset, our updated version with the interaction trajectories on Chemical Agent Sandbox, has been deposited in the huggingface repository: ChemCraft-Agent-Trajectory.

Code Availability

All code used in data analysis and preparation of the manuscript, alongside a description of necessary steps for reproducing results, can be found in a GitHub repository ChemCraft Website. The checkpoints of ChemCraft-7B has been deposited in the huggingface repository: ChemCraft-7B.

References

1. Han, Y., Wan, Z., Chen, L., Yu, K. & Chen, X. From generalist to specialist: A survey of large language models for chemistry. In Proceedings of the 31st International Conference on Computational Linguistics, 1106–1123 (2025).
2. Zhang, Q. et al. Scientific large language models: A survey on biological & chemical domains. arXiv preprint arXiv:2401.14656 (2024).
3. Xia, J., Zhu, Y., Du, Y. & Li, S. Z. A systematic survey of chemical pre-trained models. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, 6787–6795 (2023).
4. Lv, L. et al. Navigating chemical-linguistic sharing space with heterogeneous molecular encoding (2025). 2412.20888.
5. Ross, J. et al. Large-scale chemical language representations capture molecular structure and properties. Nat. Mach. Intell. **4**, 1256–1264 (2022).
6. Chithrananda, S., Grand, G. & Ramsundar, B. Chemberta: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885 (2020).
7. Ahmad, W., Simon, E., Chithrananda, S., Grand, G. & Ramsundar, B. Chemberta-2: Towards chemical foundation models. arXiv preprint arXiv:2209.01712 (2022).
8. Grisoni, F. Chemical language models for de novo drug design: Challenges and opportunities. Curr. Opin. Struct. Biol. **79**, 102527 (2023).
9. Bhattacharya, D., Cassady, H. J., Hickner, M. A. & Reinhart, W. F. Large language models as molecular design engines. J. Chem. Inf. Model. **64**, 7086–7096 (2024).
10. Bagal, V., Aggarwal, R., Vinod, P. & Priyakumar, U. D. Molgpt: molecular generation using a transformer-decoder model. J. chemical information modeling **62**, 2064–2076 (2021).
11. Lv, L. et al. Prollama: A protein large language model for multi-task protein language processing. IEEE Transactions on Artif. Intell. (2025).
12. Wan, Y., Hsieh, C.-Y., Liao, B. & Zhang, S. Retroformer: Pushing the limits of end-to-end retrosynthesis transformer. In International Conference on Machine Learning, 22475–22490 (PMLR, 2022).
13. Liu, G., Sun, M., Matusik, W., Jiang, M. & Chen, J. Multimodal large language models for inverse molecular design with retrosynthetic planning. arXiv preprint arXiv:2410.04223 (2024).
14. Ma, Q., Zhou, Y. & Li, J. Automated retrosynthesis planning of macromolecules using large language models and knowledge graphs. Macromol. Rapid Commun. 2500065 (2025).
15. Qian, Y., Li, Z., Tu, Z., Coley, C. & Barzilay, R. Predictive chemistry augmented with text retrieval. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 12731–12745 (2023).
16. Zhang, Y. et al. Large language models to accelerate organic chemistry synthesis. Nat. Mach. Intell. 1–13 (2025).
17. Cao, H. et al. PRESTO: Progressive pretraining enhances synthetic chemistry outcomes. In Al-Onaizan, Y., Bansal, M. & Chen, Y.-N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024, 10197–10224, DOI: 10.18653/v1/2024.findings-emnlp.597 (Association for Computational Linguistics, Miami, Florida, USA, 2024).
18. Zhang, Q. et al. Scientific large language models: A survey on biological & chemical domains. ACM Comput. Surv. **57**, 1–38 (2025).
19. Taylor, R. et al. Galactica: A large language model for science. arXiv preprint arXiv:2211.09085 (2022).
20. Edwards, C. et al. Translation between molecules and natural language. arXiv preprint arXiv:2204.11817 (2022).
21. Zhang, D. et al. Chemllm: A chemical large language model. arXiv preprint arXiv:2402.06852 (2024).
22. Zhao, Z. et al. Chemdfm: A large language foundation model for chemistry. In Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges (2024).
23. Xia, Y. et al. Nature language model: Deciphering the language of nature for scientific discovery. arXiv preprint arXiv:2502.07527 (2025).
24. Pei, Q. et al. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. arXiv preprint arXiv:2310.07276 (2023).
25. Zhuang, X. et al. Instructbiomol: Advancing biomolecule understanding and design following human instructions. arXiv preprint arXiv:2410.07919 (2024).

26. Ding, M. et al. Break the chain: Large language models can be shortcut reasoners. ArXiv **abs/2406.06580** (2024).

27. Li, H. et al. How to detect and defeat molecular mirage: A metric-driven benchmark for hallucination in llm-based molecular comprehension. arXiv preprint arXiv:2504.12314 (2025).

28. Rueda, A. et al. Understanding llm scientific reasoning through promptings and model's explanation on the answers. ArXiv **abs/2505.01482** (2025).

29. Zhang, Y. et al. Exploring the role of large language models in the scientific method: from hypothesis to discovery. npj Artif. Intell. **1**, 14 (2025).

30. Truhn, D., Reis-Filho, J. S. & Kather, J. N. Large language models should be used as scientific reasoning engines, not knowledge databases. Nat. medicine **29**, 2983–2984 (2023).

31. Shojaee, P., Meidani, K., Gupta, S., Farimani, A. B. & Reddy, C. K. Llm-sr: Scientific equation discovery via programming with large language models. arXiv preprint arXiv:2404.18400 (2024).

32. Luo, Y. et al. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. IEEE Transactions on Audio, Speech Lang. Process. (2025).

33. Liu, C. et al. More than catastrophic forgetting: Integrating general capabilities for domain-specific llms. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 7531–7548 (2024).

34. Zheng, J., Qiu, S., Shi, C. & Ma, Q. Towards lifelong learning of large language models: A survey. ACM Comput. Surv. **57**, 1–35 (2025).

35. Zhao, G. et al. Molreasoner: Toward effective and interpretable reasoning for molecular llms. arXiv preprint arXiv:2508.02066 (2025).

36. Zhuang, J. et al. Reasoning-enhanced large language models for molecular property prediction. arXiv preprint arXiv:2510.10248 (2025).

37. Wang, W. et al. Chem-r: Learning to reason as a chemist. arXiv preprint arXiv:2510.16880 (2025).

38. Narayanan, S. M. et al. Training a scientific reasoning model for chemistry. arXiv preprint arXiv:2506.17238 (2025).

39. Alampara, N. et al. General-purpose models for the chemical sciences: Llms and beyond. arXiv preprint arXiv:2507.07456 (2025).

40. Mswahili, M. E. & Jeong, Y.-S. Transformer-based models for chemical smiles representation: A comprehensive literature review. Heliyon **10** (2024).

41. Yao, S. et al. React: Synergizing reasoning and acting in language models. In The eleventh international conference on learning representations (2022).

42. Liu, N. F. et al. Lost in the middle: How language models use long contexts. Transactions Assoc. for Comput. Linguist. **12**, 157–173, DOI: 10.1162/tacl_a_00638 (2024).

43. Bran, A. M. et al. Chemcrow: Augmenting large-language models with chemistry tools. arXiv preprint arXiv:2304.05376 (2023).

44. M. Bran, A. et al. Augmenting large language models with chemistry tools. Nat. Mach. Intell. **6**, 525–535 (2024).

45. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. Nature **624**, 570–578 (2023).

46. Song, T. et al. A multiagent-driven robotic ai chemist enabling autonomous chemical research on demand. J. Am. Chem. Soc. **147**, 12534–12545 (2025).

47. Feretzakis, G., Papaspyridis, K., Gkoulalas-Divanis, A. & Verykios, V. S. Privacy-preserving techniques in generative ai and large language models: a narrative review. Information **15**, 697 (2024).

48. Guo, H. et al. A survey of large language model for drug research and development. IEEE Access (2025).

49. Muegge, I., Bentzien, J. & Ge, Y. Perspectives on current approaches to virtual screening in drug discovery. Expert. Opin. on Drug Discov. **19**, 1173–1183 (2024).

50. Dorri, A., Kanhere, S. S. & Jurdak, R. Multi-agent systems: A survey. Ieee Access **6**, 28573–28593 (2018).

51. Schick, T. et al. Toolformer: Language models can teach themselves to use tools. Adv. Neural Inf. Process. Syst. **36**, 68539–68551 (2023).

52. Mialon, G. et al. Augmented language models: a survey. arXiv preprint arXiv:2302.07842 (2023).

53. Li, H. et al. Beyond chemical qa: Evaluating llm's chemical reasoning with modular chemical operations (2026). 2505.21318.

54. Xu, Z. et al. Unlocking implicit experience: Synthesizing tool-use trajectories from text. arXiv preprint arXiv:2601.10355 (2026).

55. Wang, H., Wang, J., Leong, C. T. & Li, W. Steca: Step-level trajectory calibration for llm agent learning. In Findings of the Association for Computational Linguistics: ACL 2025, 11597–11614 (2025).

56. Guo, D. et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025).

57. Comanici, G. et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025).

58. Qwen, A. Y. et al. Qwen2. 5 technical report. arXiv preprint (2024).

59. Irwin, R., Dimitriadis, S., He, J. & Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. Mach. Learn. Sci. Technol. **3**, 015022 (2022).

60. Somnath, V. R., Bunne, C., Coley, C., Krause, A. & Barzilay, R. Learning graph models for retrosynthesis prediction. Adv. Neural Inf. Process. Syst. **34**, 9405–9415 (2021).

61. Wang, X. et al. Retroprime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions. Chem. Eng. J. **420**, 129845 (2021).

62. Gramage-Doria, R. & Bruneau, C. Ruthenium-catalyzed c–h bond functionalization in cascade and one-pot transformations. Coord. Chem. Rev. **428**, 213602, DOI: https://doi.org/10.1016/j.ccr.2020.213602 (2021).

63. Knochel, P. & Sämann, C. Amide synthesis via mno2-promoted c–h activation of methylarenes. Synfacts **9**, 1334–1334 (2013).

64. Wei, J., Liu, K.-M. & Duan, X.-F. Cobalt-catalyzed biaryl couplings via c–f bond activation in the absence of phosphine or nhc ligands. The J. Org. Chem. **82**, 1291–1300 (2017).

65. Ding, K. et al. Scitoolagent: a knowledge-graph-driven scientific agent for multitool integration. Nat. Comput. Sci. **5**, 962–972 (2025).

66. Silte, A. A. et al. Design, syntheses, and pharmacological evaluations of core ring expanded fentanyl analogues as potential counteracting agents against fentanyl induced respiratory depression. J. Medicinal Chem. **68**, 22170–22195, DOI: 10.1021/acs.jmedchem.5c00528 (2025). PMID: 41178824.

67. Sun, Y. et al. Discovery of potent, allosteric gsto1 covalent inhibitors with a new binding mode. J. Medicinal Chem. **68**, 24075–24093, DOI: 10.1021/acs.jmedchem.5c01724 (2025). PMID: 41233953.

68. Landrum, G. Rdkit documentation. Release **1**, 4 (2013).

69. Li, H. et al. Freestyleret: Retrieving images from style-diversified queries. In Leonardis, A. et al. (eds.) Computer Vision – ECCV 2024, 258–274 (Springer Nature Switzerland, Cham, 2025).

70. Wu, X. et al. Towards affective evaluation of stem education: Leveraging mllms in project-based learning. IEEE Transactions on Affect. Comput. (2026).

Acknowledgements

We acknowledge the financial support from National Key R&D Program of China (2026ZD0127200). This project is also supported by Shenzhen Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone, under Grant No. HTHZQSWS-KCCYB-2023052. We thank the High-Performance Computing Platform of Peking University for machine learning model training. We also extend our gratitude to all the reviewers for their insightful and constructive feedback.

Author contributions statement

H.L., H.C., Y.L., and L.Y. conceived the basic idea and designed the research study. H.L., H.C. and S.P. generated the dataset for ChemCRAFT. H.L. developed the method. H.C., Z.L. and B.F. further modified the method. Y.W. and Z.Y. conceived the evaluation metric in the experiment. H.L. and S.P. evaluated the performance of the ChemCRAFT model. H.L., H.C., S.P and Z.L. wrote the manuscript. L.Y., Y.T., Y.L., and Y.Z. revised the manuscript. Y.W. and B.F. revised the figures of the manuscript. L.Y. and Y.L. provided the deep learning computing platform.

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ChemCraftsupplementary0605.pdf