From Chemical Language Models to Autonomous Agents: A Survey of Large Language Models in Molecular Science

Anonymous ACL submission

Abstract

The integration of Large Language Models (LLMs) into molecular science marks a paradigm shift from specialized models to generalist, reasoning-capable systems. This survey provides a comprehensive overview of recent advances at the intersection of LLMs and chemistry, systematically tracing the technological evolution from dedicated models to autonomous agents. We propose a structured taxonomy categorizing existing works into four evolutionary stages: (1) Foundations of Chemical Language Models, which treat molecules as sequences for representation learning and generative design; (2) Cross-Modal Alignment and Translation, bridging the semantic gap between natural language descriptions and molecular structures; (3) Adaptation and Reasoning of General LLMs, empowering general-purpose models to tackle complex molecular-science problems via instruction tuning and in-context learning; and (4) Agents and Scientific Autonomy, utilizing LLMs as central controllers for tool orchestration, strategic planning, and automated experimentation. Finally, we discuss critical challenges and future directions for autonomous research in molecular science.

1 Introduction

Molecular science serves as a cornerstone of pharmaceutical discovery and materials innovation (Ansari et al., 2024), yet it faces the challenge of navigating an immense molecular space estimated to contain over 10^{60} drug-like candidates, while contending with the high costs of experimental trial and error. Over the past decade, Artificial Intelligence (AI) has become an accelerator for this exploration. Traditional deep learning approaches, particularly Graph Neural Networks (GNNs) and Convolutional Neural Networks (CNNs), have achieved success in discriminative tasks such as molecular property prediction and virtual screening (Zhang et al., 2025a). However, their supervised nature

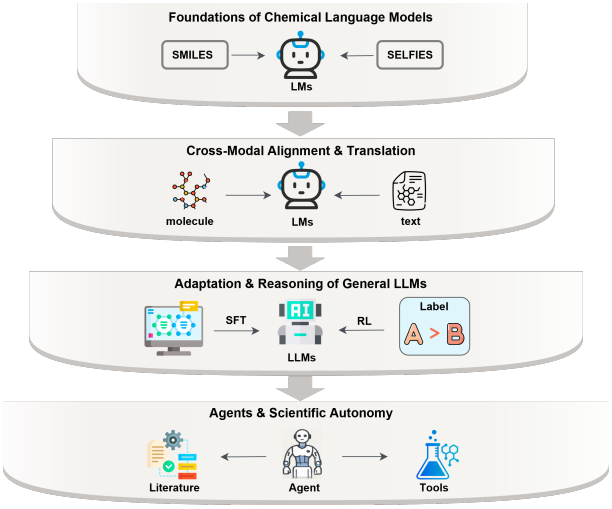

Figure 1: Overview of chemical language models, from molecular representations and cross-modal alignment to LLM adaptation and autonomous agents.

makes them heavily dependent on scarce annotated data. To alleviate this limitation, the field is increasingly turning to self-supervised learning, which leverages large-scale unlabeled data to acquire generalizable molecular knowledge for de novo design and low-data reasoning.

Language models (LMs) offer a natural framework for this shift. By linearizing molecular structures into one-dimensional sequences using notations such as SMILES, researchers can exploit the semantic parallels between molecules and language (Weininger, 1988; Krenn et al., 2020). This analogy frames molecular discovery as a sequence modeling problem, enabling LMs to learn structural patterns and physicochemical constraints from unlabeled corpora. Beyond property prediction, this paradigm also supports the manipulation of molecular representations for structure generation.

The field is now evolving from specialized string-based models toward generalist large language models (LLMs) and autonomous systems. Modern LLMs exhibit reasoning capabilities, enabling

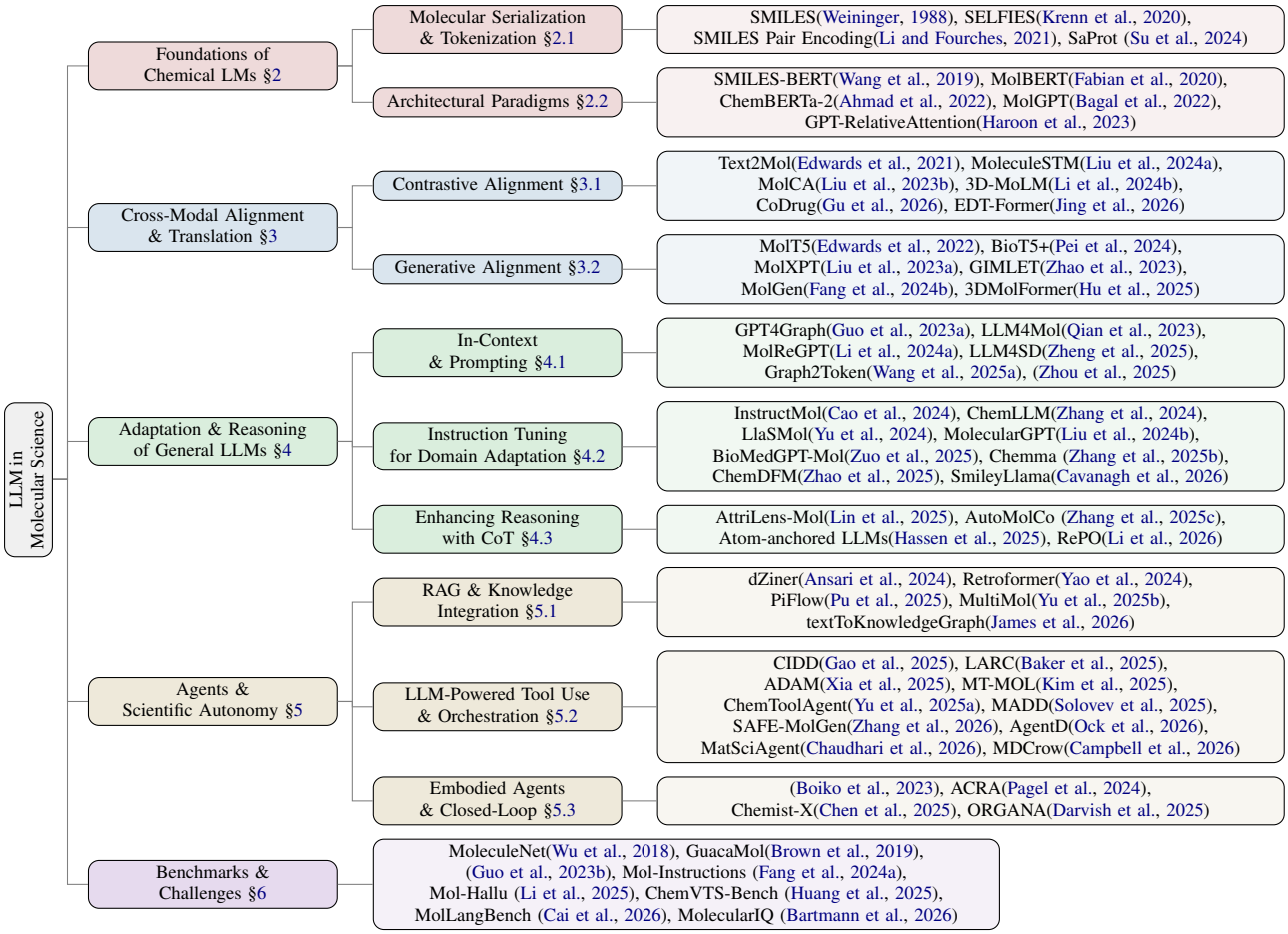

Figure 2: A taxonomy of LLMs in molecular science, spanning foundational pre-training, cross-modal alignment, adaptation, agents, and benchmarks.

high-level molecular-science tasks such as retrosynthetic planning and literature analysis without task-specific training (Yu et al., 2025b). This evolution culminates in LLM-based agents, which shift the focus from standalone prediction to system-level orchestration through tool use and automated experimentation (Darvish et al., 2025). Although prior reviews have covered subdomains such as molecular generation or property prediction (Wang et al., 2025b), a comprehensive survey that maps the evolutionary trajectory from sequence representation to autonomous reasoning remains lacking.

To address this gap, we present an overview of language-model-based approaches in molecular science in Fig. 1 and a detailed taxonomy in Fig. 2, covering foundational pre-training, cross-modal alignment, adaptation strategies, agent systems, and evaluation benchmarks. Following this trajectory, Section 2 introduces molecular serialization and chemical language modeling. Section 3 reviews cross-modal alignment between molecular representations and natural language. Section 4 examines the adaptation of general-purpose LLMs through in-context learning, instruction tuning, and chain-of-thought reasoning. Section 5 surveys autonomous agent systems that integrate retrieval, tools, and embodied components. Finally, Section 6 summarizes datasets and benchmarks, and Section 7 discusses challenges such as hallucination, safety, and future research directions.

2 Foundations of Chemical Language Models

This section reviews the foundations of chemical language models. Since molecules are graph-structured objects governed by strict physicochemical constraints, applying LMs to chemistry requires effective serialization and tokenization strategies. We then summarize representative architectures for molecular modeling, focusing on encoder-only models for representation learning and decoder-only models for generative molecular design.

2.1 Molecular Serialization and Tokenization

The serialization of molecular graphs into linear strings is a prerequisite for applying language models to chemistry, with SMILES serving as the dominant standard. SMILES represents molecules as ASCII strings generated via depth-first graph traversal, efficiently encoding connectivity. However, it suffers from significant syntax fragility in generative tasks; minor character alterations often lead to chemically invalid valences or unclosed rings. To resolve this, SELFIES introduces a 100% robust representation system based on a derivation grammar. Unlike SMILES, SELFIES ensures that every possible sequence of characters corresponds to a valid molecular graph by deriving structure dynamically. This property effectively eliminates the "invalid output" problem common in early generative models and allows the latent space to be fully traversable for molecular optimization.

Beyond the string format, tokenization strategies determine semantic granularity and have recently extended beyond linear sequences to 3D spatial domains. Moving beyond atom-level segmentation, methods such as SMILES Pair Encoding (Li and Fourches, 2021) adapt the Byte Pair Encoding algorithm to merge high-frequency symbols into chemically interpretable substructures, thereby enhancing the model's ability to capture chemical semantics. Furthermore, to enable language models to comprehend the 3D spatial structure of molecules, tokenization strategies have evolved to discretize geometric coordinates or structural conformations. Models such as SaProt (Su et al., 2024) demonstrate that converting structural information (e.g., Foldseek alphabets) into discrete tokens allows standard Transformers to directly learn and generate valid 3D molecular geometries.

2.2 Architectural Paradigms

Encoder-only architectures are characterized by bidirectional attention mechanisms designed to extract dense, context-aware representations from molecular sequences. SMILES-BERT (Wang et al., 2019) pioneers this paradigm by adapting the Transformer encoder for a masked SMILES recovery task, demonstrating that self-supervised learning on large unlabeled corpora can substantially improve downstream property prediction. Building on this foundation, MolBERT (Fabian et al., 2020) introduces additional pre-training objectives, including SMILES equivalence detection and physicochemical property prediction, encouraging the model to capture deeper chemical semantics beyond surface-level structural patterns. Furthermore, ChemBERTa-2 (Ahmad et al., 2022) empirically validates scaling laws in chemical language modeling; by training on 77 million compounds, it highlights the critical role of data scale and model capacity in improving molecular representation quality.

Decoder-only architectures, in contrast, employ unidirectional auto-regressive attention mechanisms and are naturally suited for generative tasks, effectively framing molecular design as a language generation problem. MolGPT (Bagal et al., 2022) exemplifies this paradigm by using causal masked self-attention to predict the next token in a SMILES sequence, while supporting conditional generation to control molecular properties. To address chemical syntax challenges such as long-range dependencies in ring structures, GPT-RelativeAttention (Haroon et al., 2023) replaces absolute positional encodings with relative attention mechanisms, leading to significantly improved structural consistency in generated molecules.

3 Cross-Modal Alignment and Translation

While foundational chemical language models excel at understanding unimodal molecular sequences, real-world scientific discovery requires seamless interaction between molecular structures and natural language instructions. Bridging this gap necessitates cross-modal alignment, which maps the distinct feature spaces of molecules (graphs, 3D geometries) and text into a unified representation. This section reviews the core strategies for achieving this alignment. We categorize existing methodologies into two primary paradigms: contrastive alignment, which learns a shared latent space by minimizing the distance between paired representations, and generative alignment, which unifies modalities by modeling the interaction as a direct sequence-to-sequence generation task.

3.1 Contrastive Alignment

Contrastive alignment maps molecular and textual modalities into a shared semantic space, usually through modality-specific encoders and contrastive objectives. Early methods focus on molecule–text matching: Text2Mol (Edwards et al., 2021) aligns molecular graph embeddings with textual descriptions for natural language–based re-

trieval, while MoleculeSTM (Liu et al., 2024a) scales this paradigm to large molecule–text corpora and enables zero-shot retrieval and instruction-guided molecule editing. Recent work expands this setting to richer molecular-science contexts. CoDrug (Gu et al., 2026), for example, aligns natural language, molecular SMILES, and protein sequences with modality-specific encoders and InfoNCE objectives, supporting zero-shot text-driven virtual screening and multi-property molecular optimization without requiring high-resolution 3D complex structures. This progression shows that contrastive alignment is moving beyond simple molecule–description matching toward grounding molecular semantics in biological targets, functional contexts, and optimization goals.

A second line of work focuses on LLM-oriented modality bridging, where molecular representations are projected into the embedding space of general-purpose LLMs. MolCA (Liu et al., 2023b) maps 2D molecular graph features into Galactica through a query-based projector, while 3D-MoLM (Li et al., 2024b) extends this strategy to 3D molecular geometry by injecting learned spatial representations into Llama-2 (Touvron et al., 2023). EDT-Former (Jing et al., 2026) further improves this connector-based paradigm by replacing fixed-length query tokens with entropy-guided dynamic tokens extracted from informative molecular patches. These methods suggest that graph–LLM alignment depends not only on cross-modal similarity but also on how molecular structures are selected, compressed, and exposed to LLMs. Since molecular semantics are unevenly distributed across graphs, faithful and compact molecular tokenization becomes crucial for downstream reasoning and generation.

3.2 Generative Alignment

While contrastive methods align modalities in a shared latent space, generative alignment achieves cross-modal interaction by modeling it as a sequence generation task. This paradigm trains models to generate the target modality sequence from the source, effectively unifying the representation through the generation process. One approach within this paradigm is the encoder-decoder framework, which treats scientific tasks as "text-to-text" translation problems. MolT5 (Edwards et al., 2022) pioneers this by pre-training on massive datasets to enable bidirectional translation between chemical SMILES and natural language descriptions. Building on this, BioT5+ (Pei et al., 2024) incorporates IUPAC nomenclature and biological context into the instruction tuning process, enhancing generalization across diverse biochemical tasks. Similarly, MolGen (Fang et al., 2024b) leverages the BART (Lewis et al., 2020) architecture combined with SELFIES serialization and reinforcement learning, specifically optimizing the validity and controllability of generated molecules to improve transfer from synthetic molecules to natural products.

Alternatively, the unified-decoder paradigm integrates both molecular data and natural language into a single autoregressive stream, employing GPT-style models for comprehensive generation. MolXPT (Liu et al., 2023a) exemplifies this by employing a "text-wrapped molecule" strategy, where molecular SMILES are seamlessly contextualized within scientific texts to unify tasks such as property prediction and generation. Moving beyond 1D sequences, GIMLET (Zhao et al., 2023) proposes a unified graph-text Transformer. By utilizing generalized position encoding, it models molecular graphs and instructions directly without a separate graph encoder, achieving impressive zero-shot performance. Furthermore, 3DMolFormer (Hu et al., 2025) extends this unification to 3D geometry using a dual-channel Transformer. By representing 3D protein pockets and ligands as parallel sequences, it effectively unifies the dual tasks of docking and pocket-aware drug design.

4 Adaptation and Reasoning of General LLMs

Building on the specialized architectures and cross-modal alignments discussed above, we now examine how general-purpose LLMs are adapted to molecular-science tasks. Although general LLMs offer broad knowledge, few-shot generalization, and emergent reasoning, they often lack chemical expertise and struggle to satisfy strict molecular syntax constraints, as required for valid SMILES generation. As shown in Fig. 3, existing methods mainly follow three paradigms: (a) in-context learning (ICL), which elicits task behavior from frozen LLMs through prompts and few-shot examples; (b) supervised fine-tuning (SFT) and reinforcement learning, which inject molecular knowledge and task preferences into model parameters; and (c) Chain-of-Thought (CoT) reasoning, often coupled with external verification or feedback.

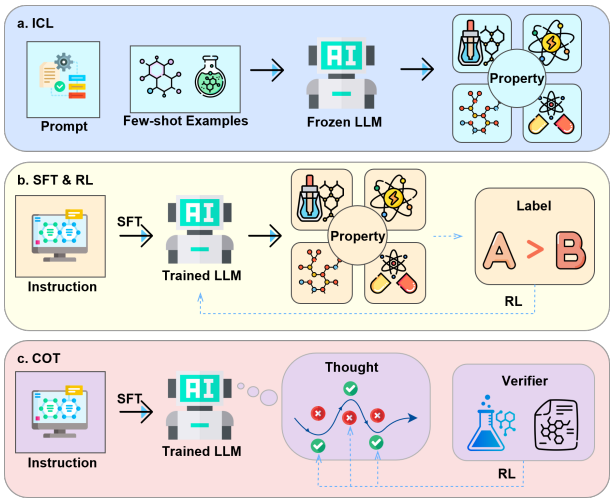

Figure 3: LLM adaptation strategies for molecular science: in-context learning, supervised fine-tuning/RL, and verified chain-of-thought reasoning.

4.1 In-Context Learning and Prompt Engineering

As illustrated in Fig. 3(a), ICL represents the least intrusive strategy for adapting general-purpose LLMs to molecular science. By keeping model parameters frozen, ICL elicits task-specific behavior through task descriptions, prompt design, and few-shot demonstrations. This setting tests whether general LLMs can interpret molecular representations, follow scientific instructions, and generate chemically meaningful outputs without explicit domain training. Its effectiveness, however, is highly sensitive to prompt format, demonstration selection, and the textualization of molecular structures.

Early studies primarily examine whether LLMs can directly handle molecular or graph-structured tasks from textual prompts. GPT4Graph (Guo et al., 2023a) evaluates LLMs on graph-related problems, showing that they can perform simple graph reasoning but still lag behind specialized GNNs. To improve task grounding, MolReGPT (Li et al., 2024a) introduces retrieval-based few-shot prompting, where molecular fingerprints and text retrieval are used to select similar examples for molecule–text translation. Graph2Token (Wang et al., 2025a) further strengthens the graph–LLM interface by encoding molecular graphs into vocabulary-level "Graph Tokens" and combining them with IUPAC-enhanced prompts. These studies suggest that ICL performance depends not only on the reasoning ability of LLMs but also on whether molecular structures can be converted into prompt-compatible semantic units.

Beyond direct prediction, ICL also enables LLMs to act as knowledge or feature generators for specialized molecular models. LLM4Mol (Qian et al., 2023) explores zero-shot prompting and a "Caption as New Representation" strategy, where LLM-generated explanations are used as molecular features. LLM4SD (Zheng et al., 2025) prompts LLMs to synthesize literature-derived and data-driven rules as interpretable vectors for property prediction. Similarly, LLM-Enhanced MPP (Zhou et al., 2025) uses LLMs to generate domain-specific code and feature rules, which are then fused with GNN-based structural representations. Together, these methods show that prompt-based LLMs can serve as knowledge translators, converting scientific priors into explicit features, rules, or explanations. Nevertheless, their outputs remain sensitive to prompt design and often require retrieval modules or structural models to maintain chemical validity and robustness. Future ICL evaluation should also account for potential data leakage, since general LLMs may have encountered benchmark molecules, task templates, or textual descriptions during pre-training, leading to inflated zero-shot or few-shot performance.

4.2 Instruction Tuning for Domain Adaptation

As illustrated in Fig. 3(b), instruction tuning is a parameter-based adaptation strategy that explicitly trains general-purpose LLMs to acquire chemical knowledge and align their behavior with domain-specific objectives. Unlike ICL, which keeps model parameters frozen, instruction tuning updates LLMs with curated instruction–response data, enabling them to internalize chemical concepts, constraints, and reasoning patterns.

Instruction tuning injects domain knowledge mainly through SFT and, in some cases, reinforcement learning. Its effectiveness largely depends on the quality and coverage of training data. LlaSMol (Yu et al., 2024) constructs SMolInstruct, a large-scale dataset containing over 3 million samples across 14 chemical tasks, enabling open-source models to achieve expert-level performance. Similarly, ChemDFM (Zhao et al., 2025) and ChemLLM (Zhang et al., 2024) adopt a two-stage strategy that combines continued pre-training on chemical corpora with instruction tuning on structured chemical tasks converted into dialogue formats, improving performance from molecular recognition to open-ended chemical conversation.

Beyond text-only supervision, recent studies extend instruction tuning to multimodal understanding, controllable generation, and planning. InstructMol (Cao et al., 2024) aligns molecular structural encoders with Llama-2 to support molecular understanding and question answering. SmileyLlama (Cavanagh et al., 2026) adapts Llama for directed chemical space exploration, using property-prompted SFT and DPO to generate drug-like molecules under multi-property constraints, and integrates with iMiner to optimize 3D target binding affinity. BioMedGPT-Mol (Zuo et al., 2025) combines SFT with RL for end-to-end retrosynthetic planning, while Chemma (Zhang et al., 2025b) embeds a fine-tuned LLaMA-2 into an active learning loop for reaction-space exploration. MolecularGPT (Liu et al., 2024b) further improves few-shot generalization through structure-aware retrieval. Overall, these works suggest that chemical instruction tuning is moving beyond static domain adaptation toward controllable generation, planning, and feedback-driven scientific optimization.

4.3 Enhancing Reasoning with Chain-of-Thought

As illustrated in Fig. 3(c), CoT reasoning represents a reasoning-oriented adaptation strategy that improves LLM inference by decomposing complex chemical problems into intermediate steps. Rather than focusing only on knowledge injection, CoT-based methods expose intermediate decision processes, thereby improving reliability, interpretability, and controllability in tasks involving molecular structures, reactions, and constraints.

Many chemical tasks require multi-step reasoning grounded in molecular structures rather than factual recall. Atom-anchored LLMs (Hassen et al., 2025) introduce a training-free framework that anchors CoT reasoning to atom-mapping indices in SMILES strings, decomposing retrosynthesis into position prediction and transformation generation to improve chemical validity. AutoMolCo (Zhang et al., 2025c) adopts concept-based reasoning by prompting LLMs to generate and label molecular concepts via code or tools, thereby forming an interpretable concept bottleneck for property prediction. These studies show that explicit reasoning traces can make LLM predictions transparent but their reliability still depends on chemically grounded representations and verification signals.

Recent work further optimizes reasoning trajectories through reinforcement learning. RePO (Li et al., 2026) proposes reference-guided policy optimization for molecular instruction optimization, combining RLVR/GRPO-style exploration over reasoning trajectories with supervised reference guidance to mitigate SFT-induced reasoning collapse and reward sparsity under similarity constraints, without requiring human-written step-by-step trajectories. AttriLens-Mol (Lin et al., 2025) employs attribute-guided RL rewards to encourage reasoning paths that discuss relevant molecular attributes, enabling the model to discover effective prediction strategies without handcrafted prompts. Overall, CoT-based adaptation is shifting from eliciting explicit explanations to optimizing verifiable reasoning processes. This trend suggests that future chemical reasoning models should evaluate not only final molecular validity or prediction accuracy, but also the faithfulness, chemical grounding, and optimizability of intermediate reasoning traces.

5 Agents and Scientific Autonomy

The shift from static language models to LLM-based agents marks a move from passive molecular prediction to active scientific workflows. Standard LLMs encode broad knowledge but cannot, by themselves, access up-to-date evidence, execute precise calculations, or interact with laboratory environments. Agents address these limits by using LLMs as planners that coordinate retrieval, external tools, computational models, and experimental interfaces. This paradigm enables systems to synthesize literature, design molecules, run simulations, and support closed-loop experimentation. We categorize these agentic systems into three functional levels: retrieval-augmented generation (RAG) for knowledge grounding, tool use and orchestration for computational action, and embodied agents for physical experimentation.

5.1 Retrieval-Augmented Generation and Knowledge Integration

While LLMs possess broad internal knowledge, they often suffer from hallucinations and outdated information. RAG addresses this limitation by grounding model outputs in external, verifiable knowledge sources. dZiner (Ansari et al., 2024) exemplifies this direction by using GPT-4o to retrieve explicit design guidelines from scientific literature and combining them with numerical surrogate models for rational inverse design of surfactants and MOFs. MultiMol (Yu et al., 2025b)

further introduces a dual-agent framework, where a Data-Driven Worker generates molecules and a Literature-Guided Researcher retrieves supporting evidence, enabling knowledge-grounded filtering for multi-objective molecular optimization.

Beyond direct textual retrieval, recent systems increasingly structure scientific evidence into reusable knowledge representations. textToKnowledgeGraph (James et al., 2026) uses LLMs such as GPT-4o-mini to extract molecular interactions from unstructured scientific literature into Biological Expression Language (BEL) knowledge graphs, integrates with Cytoscape for network exploration, and supports GraphRAG to improve the traceability and reliability of LLM question answering. PiFlow (Pu et al., 2025) introduces a principle-aware multi-agent framework that dynamically balances the exploitation of known scientific principles with the exploration of novel spaces, improving discovery efficiency. Retroformer (Yao et al., 2024) further enables agents to learn from retrieval interactions through retrospective reinforcement learning, where a lightweight retrospective model generates reflective prompts from environmental feedback. Overall, RAG-based chemical agents are evolving from passage-level retrieval toward structured, interactive knowledge integration. This shift highlights the importance of not only retrieving relevant evidence but also organizing it into verifiable representations that can support reasoning, feedback, and long-horizon scientific planning.

5.2 LLM-Powered Tool Use and Orchestration

While RAG connects agents to external knowledge, tool use enables them to act by invoking computational APIs, predictive models, simulation engines, and safety checkers. One group of systems targets end-to-end drug discovery workflows. AgentD (Ock et al., 2026) and MADD (Solovev et al., 2025) organize LLM agents around early discovery pipelines: AgentD modularizes data retrieval, literature-grounded QA, molecule generation, property prediction, property-based optimization, and 3D protein–ligand modeling, while MADD combines decomposer, orchestrator, and summarizer agents with generative and AutoML models for automated hit generation. Other agents specialize in scientific computation and simulation. ADAM (Xia et al., 2025), MDCrow (Campbell et al., 2026), and MatSciAgent (Chaudhari et al., 2026) translate natural language goals into executable workflows across computational biophysics and materials science, including antibody design, molecular dynamics simulation, materials retrieval, continuum simulation, and crystal structure generation.

At the molecular design level, tool orchestration links LLM reasoning with domain-specific evaluation. CIDD (Gao et al., 2025) couples LLMs with structure-based drug design models to iteratively balance binding affinity and drug-likeness, while SAFE-MolGen (Zhang et al., 2026) integrates LLM-based generation with cheminformatics tools, ML predictors, and QM calculations for designing selective f-element extractant ligands. Scaling this idea to broader molecular optimization, MT-MOL (Kim et al., 2025) assigns role-based agents, such as Analyst, Scientist, and Reviewer, and equips them with 154 RDKit tools for structured reasoning. However, tool access also introduces orchestration and safety challenges: ChemToolAgent (Yu et al., 2025a) shows that tools can improve specialized chemistry tasks but may increase cognitive load and weaken general reasoning, whereas LARC (Baker et al., 2025) uses LLM-based supervision to enforce safety constraints in retrosynthesis planning. Overall, these systems suggest that the key challenge is to shift from merely providing tools to deciding when, why, and how agents should invoke, verify, and safely compose them in long-horizon scientific workflows.

5.3 Embodied Agents and Closed-Loop Discovery

The frontier of chemical agents lies in embodied systems that connect digital reasoning with physical experimentation. Early demonstrations by Boiko et al. (Boiko et al., 2023) show that GPT-4-based agents, equipped with web search, Python execution, and robotic interfaces, can autonomously plan and execute laboratory syntheses such as ibuprofen preparation and Suzuki reactions. ACRA (Pagel et al., 2024) standardizes this transition by extracting synthesis protocols from literature via RAG, translating them into machine-executable chemical description language, and executing them on the Chemputer platform. More recent systems incorporate richer perception and feedback: ORGANA (Darvish et al., 2025) combines LLM reasoning, visual perception, and task-and-motion planning to automate operations such as solubility screening, recrystallization, and pH measurement, while Chemist-X (Chen et al., 2025) closes the reaction optimization loop through

a retrieve–recommend–verify pipeline that links RAG, condition recommendation, ML/CAD models, and wet-lab execution. Despite these advances, embodied chemical agents remain far from being general-purpose autonomous chemists. Their success often depends on constrained laboratory setups, well-specified protocols, reliable hardware abstractions, and carefully engineered safety boundaries. Moreover, errors can accumulate across perception, planning, execution, and analysis, making reproducibility and accountability difficult. Future progress requires not only more capable LLM planners but also standardized machine-readable protocols, robust uncertainty estimation, real-time failure recovery, and human-auditable safety mechanisms for closed-loop experimentation.

6 Datasets and Benchmarks

As LLMs in molecular science evolve from property predictors to multimodal assistants and autonomous agents, evaluation must move beyond task accuracy toward testing whether models understand molecular structures, instructions, and constraints. Earlier benchmarks provide useful but partial coverage: MoleculeNet (Wu et al., 2018) and GuacaMol (Brown et al., 2019) evaluate property prediction and molecular generation, while ChemLLMBench (Guo et al., 2023b), Mol-Instructions (Fang et al., 2024a), ChemVTS-Bench (Huang et al., 2025), and Mol-Hallu (Li et al., 2025) extend evaluation to instruction following, multimodal reasoning, and hallucination detection.

Recent benchmarks therefore focus on molecular-language interaction and verifiable structural reasoning. MolLangBench (Cai et al., 2026) evaluates LLMs and VLMs on three language-prompted tasks: molecular structure recognition, editing, and generation, revealing limitations in precise molecular understanding. MolecularIQ (Bartmann et al., 2026) complements this by using symbolically verifiable molecular-graph tasks, such as feature counting, indexing, and constrained generation, to diagnose whether models parse molecular structures under varying complexity. Together, these benchmarks show that fluent chemical language does not necessarily imply faithful molecular reasoning. Future evaluation should therefore emphasize leakage-resistant splits, deterministic structural verification, calibrated uncertainty, and process-level diagnostics for reasoning and agentic workflows.

7 Conclusion and Future Directions

This survey traces the evolution of LLMs in molecular science from chemical language models and cross-modal alignment to the adaptation of general-purpose LLMs and autonomous agents. The field is moving from specialized predictors toward generalist systems that interpret molecular representations, follow scientific instructions, reason over constraints, and coordinate tools or experiments. Yet this transition exposes a central tension: fluent chemical language does not necessarily imply faithful molecular understanding, reliable reasoning, or experimentally executable synthesis plans.

A first priority is trustworthiness. Chemical LLMs remain vulnerable to hallucination, overconfident reasoning, and weak molecular grounding, consistent with analyses of hallucination in language models (Kalai et al., 2025). These issues are amplified by limited data provenance and benchmark leakage: many general LLMs are trained on opaque web-scale corpora and may have encountered benchmark molecules, reaction templates, or textual task descriptions before evaluation. As a result, zero-shot and few-shot performance can be difficult to interpret. Future benchmarks should therefore incorporate provenance-aware curation, deduplication, temporal or scaffold-based splits, uncertainty calibration, and process-level verification rather than rely only on final-answer accuracy.

A second priority is real-world executability and safety. Autonomous chemical agents should not be evaluated only by whether they produce plausible digital plans, because actual synthesis may fail due to reagent availability, unstable intermediates, side reactions, low yields, purification barriers, scale-up constraints, missing experimental metadata, or hardware errors. These practical bottlenecks make closed-loop chemistry fundamentally harder than text-based planning. At the same time, safety risks become more serious as agents gain access to synthesis planners, tools, and robotic platforms. Benchmarks such as SciSafeEval (Huang et al., 2024) and attacks such as SMILES-Prompting (Wong et al., 2024) show that safety mechanisms must move beyond simple refusal policies toward domain-aware risk assessment, route-level monitoring, and human-auditable oversight. The future of LLMs in molecular science therefore depends not only on larger models but also on systems that can reason, verify, abstain, and act safely within real chemical workflows.

Limitations

This survey provides a structured overview of large language models in molecular science, but it has several limitations. Because the field is evolving rapidly, especially in areas such as LLM-based agents, molecular reasoning benchmarks, and autonomous laboratory systems, very recent or concurrently developed methods may not be fully reflected. The proposed taxonomy organizes heterogeneous work around the progression from molecular sequence modeling to autonomous agents, but it may underrepresent task-oriented perspectives, and some methods span multiple categories. We also synthesize published papers and preprints rather than conducting a unified empirical comparison, so reported results are difficult to compare due to differences in datasets, protocols, model scales, prompting strategies, validity checks, data leakage, and reproducibility. Finally, current autonomous chemical agents are often evaluated in constrained settings, and real-world deployment still depends on reagent availability, hardware reliability, safety constraints, and human oversight. Thus, this survey should be viewed as a snapshot of an emerging research trajectory rather than a definitive account of mature autonomous molecular science.

References

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2022. Chemberta-2: Towards chemical foundation models. *Preprint*, arXiv:2209.01712.

Mehrad Ansari, Jeffrey Watchorn, Carla E. Brown, and Joseph S. Brown. 2024. dziner: Rational inverse design of materials with ai agents. *Preprint*, arXiv:2410.03963.

Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. 2022. MolGPT: Molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076.

Frazier N. Baker, Daniel Adu-Ampratwum, Reza Averly, Botao Yu, Huan Sun, and Xia Ning. 2025. Larc: Towards human-level constrained retrosynthesis planning through an agentic framework. *Preprint*, arXiv:2508.11860.

Christoph Bartmann, Johannes Schimunek, Mykyta Ielanskyi, Philipp Seidl, Günter Klambauer, and Sohvi Luukkonen. 2026. MolecularIQ: Characterizing chemical reasoning capabilities through symbolic verification on molecular graphs. In *The Fourteenth International Conference on Learning Representations*.

Daniil A. Boiko, Robert MacKnight, and Gabe Gomes. 2023. Emergent autonomous scientific research capabilities of large language models. *Preprint*, arXiv:2304.05332.

Nathan Brown, Marco Fiscato, Marwin H. S. Segler, and Alain C. Vaucher. 2019. GuacaMol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3):1096–1108.

Feiyang Cai, Jiahui Bai, Tao Tang, Guijuan He, Joshua Luo, Tianyu Zhu, Srikanth Pilla, Gang Li, Ling Liu, and Feng Luo. 2026. MolLangBench: A comprehensive benchmark for language-prompted molecular structure recognition, editing, and generation. In *The Fourteenth International Conference on Learning Representations*.

Quintina Campbell, Sam Cox, Jorge Medina, Brittany Watterson, and Andrew D. White. 2026. MDCrow: Automating molecular dynamics workflows with large language models. *Machine Learning: Science and Technology*, 7(2):025037.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2024. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *Preprint*, arXiv:2311.16208.

Joseph M. Cavanagh, Kunyang Sun, Andrew Gritsevskiy, Dorian Bagni, Yingze Wang, Thomas D. Bannister, and Teresa Head-Gordon. 2026. SmileyLlama: Modifying large language models for directed chemical space exploration. *Nature Computational Science*.

Akshat Chaudhari, Janghoon Ock, and Amir Barati Farimani. 2026. Modular large language model agents for multi-task computational materials science. *Communications Materials*.

Kexin Chen, Jiamin Lu, Junyou Li, Xiaoran Yang, Yuyang Du, Kunyi Wang, Qiannuan Shi, Jiahui Yu, Lanqing Li, Jiezhong Qiu, Jianzhang Pan, Yi Huang, Qun Fang, Pheng Ann Heng, and Guangyong Chen. 2025. Chemist-x: Large language model-empowered agent for reaction condition recommendation in chemical synthesis. *Preprint*, arXiv:2311.10776.

Kourosh Darvish, Marta Skreta, Yuchi Zhao, Naruki Yoshikawa, Sagnik Som, Miroslav Bogdanovic, Yang Cao, Han Hao, Haoping Xu, Alán Aspuru-Guzik, Animesh Garg, and Florian Shkurti. 2025. Organa: A robotic assistant for automated chemistry experimentation and characterization. *Preprint*, arXiv:2401.06949.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2Mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin H. S. Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. 2020. Molecular representation learning with language models and domain-relevant auxiliary tasks. *Preprint*, arXiv:2011.13230.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2024a. Mol-Instructions: A large-scale biomolecular instruction dataset for large language models. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Yin Fang, Ningyu Zhang, Zhuo Chen, Lingbing Guo, Xiaohui Fan, and Huajun Chen. 2024b. Domain-agnostic molecular generation with chemical feedback. In *International Conference on Learning Representations*, pages 53993–54018.

Bowen Gao, Yanwen Huang, Yiqiao Liu, Wenxuan Xie, Wei-Ying Ma, Ya-Qin Zhang, and Yanyan Lan. 2025. Pushing the boundaries of structure-based drug design through collaboration with large language models. *Preprint*, arXiv:2503.01376.

Rui Gu, Yingxu Liu, Bingxing Zhu, Li Liang, Haichun Liu, Yanmin Zhang, and Yadong Chen. 2026. CoDrug: A text-driven molecular virtual screening and multiproperty optimization framework via multimodal language model. *Journal of Chemical Information and Modeling*, 66(3):1582–1593.

Jiayan Guo, Lun Du, Hengyu Liu, Mengyu Zhou, Xinyi He, and Shi Han. 2023a. Gpt4graph: Can large language models understand graph structured data ? an empirical evaluation and benchmarking. *Preprint*, arXiv:2305.15066.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, and Xiangliang Zhang. 2023b. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In *Advances in Neural Information Processing Systems*, volume 36, pages 59662–59688. Curran Associates, Inc.

Suhail Haroon, Hafsath C.A., and Jereesh A.S. 2023. Generative pre-trained transformer (GPT) based model with relative attention for de novo drug design. *Computational Biology and Chemistry*, 106:107911.

Alan Kai Hassen, Andrius Bernatavicius, Antonius P. A. Janssen, Mike Preuss, Gerard J. P. van Westen, and Djork-Arné Clevert. 2025. Atom-anchored llms speak chemistry: A retrosynthesis demonstration. *Preprint*, arXiv:2510.16590.

Xiuyuan Hu, Guoqing Liu, Can Chen, Yang Zhao, Hao Zhang, and Xue Liu. 2025. 3DMolFormer: A dual-channel framework for structure-based drug discovery. In *International Conference on Learning Representations*, pages 27562–27583.

Zhiyuan Huang, Baichuan Yang, Zikun He, Yanhong Wu, Fang Hongyu, Zhenhe Liu, Lin Dongsheng, and Bing Su. 2024. Scisafeeval: A comprehensive benchmark for safety alignment of large language models in scientific tasks. *Preprint*, arXiv:2410.03769.

Zhiyuan Huang, Baichuan Yang, and 1 others. 2025. Chemvts-bench: Evaluating visual-textual-symbolic reasoning of multimodal large language models in chemistry. *Preprint*, arXiv:2511.17909.

Favour James, Dexter Pratt, Christopher Churas, and Augustin Luna. 2026. textToKnowledgeGraph: Generation of molecular interaction knowledge graphs using large language models for exploration in Cytoscape. *Bioinformatics*, 42(2):btag031.

Zihao Jing, Qiuhao Zeng, Ruiyi Fang, Yan Sun, Boyu Wang, and Pingzhao Hu. 2026. Entropy-guided dynamic tokens for graph-LLM alignment in molecular understanding. In *Proceedings of the Fourteenth International Conference on Learning Representations*.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *Preprint*, arXiv:2509.04664.

Hyomin Kim, Yunhui Jang, and Sungsoo Ahn. 2025. Mt-mol:multi agent system with tool-based reasoning for molecular optimization. *Preprint*, arXiv:2505.20820.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. 2020. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Hao Li, Liuzhenghao Lv, He Cao, Zijing Liu, Zhiyuan Yan, Yu Wang, Yonghong Tian, Yu Li, and Li Yuan. 2025. How to detect and defeat molecular mirage: A metric-driven benchmark for hallucination in llm-based molecular comprehension. *Preprint*, arXiv:2504.12314.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. 2024a. Empowering molecule discovery for molecule-caption translation with large language models: A ChatGPT perspective. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6071–6083.

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. 2024b. Towards 3d molecule-text interpretation in language models. *Preprint*, arXiv:2401.13923.

Xinhao Li and Denis Fourches. 2021. Smiles pair encoding: A data-driven substructure tokenization algorithm for deep learning. *Journal of Chemical Information and Modeling*, 61(4):1560–1569.

Xuan Li, Zhanke Zhou, Zongze Li, Jiangchao Yao, Yu Rong, Lu Zhang, and Bo Han. 2026. Reference-guided policy optimization for molecular optimization via LLM reasoning. In *Proceedings of the Fourteenth International Conference on Learning Representations*.

Xuan Lin, Long Chen, and Yile Wang. 2025. Attrilensmol: Attribute guided reinforcement learning for molecular property prediction with large language models. *Preprint*, arXiv:2508.04748.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. 2024a. Multi-modal molecule structure-text model for text-based retrieval and editing. *Preprint*, arXiv:2212.10789.

Yuyan Liu, Sirui Ding, Sheng Zhou, Wenqi Fan, and Qiaoyu Tan. 2024b. Moleculargpt: Open large language model (llm) for few-shot molecular property prediction. *Preprint*, arXiv:2406.12950.

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023a. MolXPT: Wrapping molecules with text for generative pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1606–1616, Toronto, Canada. Association for Computational Linguistics.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023b. MolCA: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15623–15638, Singapore. Association for Computational Linguistics.

Janghoon Ock, Radheesh Sharma Meda, Srivathsan Badrinarayanan, Neha S. Aluru, Achuth Chandrasekhar, and Amir Barati Farimani. 2026. Large language model agent for modular task execution in drug discovery. *Journal of Chemical Information and Modeling*, 66(4):2055–2068.

Sebastian Pagel, Michael Jirasek, and Leroy Cronin. 2024. Validation of the scientific literature via chemputation augmented by large language models. *Preprint*, arXiv:2410.06384.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. 2024. BioT5+: Towards generalized biological understanding with IUPAC integration and multi-task tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1216–1240, Bangkok, Thailand. Association for Computational Linguistics.

Yingming Pu, Tao Lin, and Hongyu Chen. 2025. Piflow: Principle-aware scientific discovery with multi-agent collaboration. *Preprint*, arXiv:2505.15047.

Chen Qian, Huayi Tang, Zhirui Yang, Hong Liang, and Yong Liu. 2023. Can large language models empower molecular property prediction? *Preprint*, arXiv:2307.07443.

Gleb Vitalevich Solovev, Alina Borisovna Zhidkovskaya, Anastasia Orlova, Nina Gubina, Anastasia Vepreva, Rodion Golovinskii, Ilya Tonkii, Ivan Dubrovsky, Ivan Gurev, Dmitry Gilemkhanov, Denis Chistiakov, Timur A. Aliev, Ivan Poddiakov, Galina Zubkova, Ekaterina V. Skorb, Vladimir Vinogradov, Alexander Boukhanovsky, Nikolay Nikitin, Andrei Dmitrenko, and 2 others. 2025. MADD: Multi-agent drug discovery orchestra. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6956–6998, Suzhou, China. Association for Computational Linguistics.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. 2024. Saprot: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations*, volume 2024, pages 6987–7009.

Hugo Touvron, Louis Martin, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Runze Wang, Mingqi Yang, and Yanming Shen. 2025a. Bridging molecular graphs and large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. SMILES-BERT: Large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 429–436, New York, NY, USA. Association for Computing Machinery.

Ziqing Wang, Kexin Zhang, Zihan Zhao, Yibo Wen, Abhishek Pandey, Han Liu, and Kaize Ding. 2025b. A survey of large language models for text-guided molecular discovery: from molecule generation to optimization. *Preprint*, arXiv:2505.16094.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36.

Aidan Wong, He Cao, Zijing Liu, and Yu Li. 2024. Smiles-prompting: A novel approach to llm jailbreak attacks in chemical synthesis. *Preprint*, arXiv:2410.15641.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: A benchmark for molecular machine learning. *Preprint*, arXiv:1703.00564.

Yijie Xia, Xiaohan Lin, Zicheng Ma, Jinyuan Hu, Yanheng Li, Zhaoxin Xie, Hao Li, Li Yang, Zhiqiang Zhao, Lijiang Yang, Zhenyu Chen, and Yi Qin Gao. 2025. Large language models as ai agents for digital atoms and molecules: Catalyzing a new era in computational biophysics. *APL Computational Physics*, 1(1):010902.

Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh Murthy, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, Ran Xu, Phil Mui, Huan Wang, Caiming Xiong, and Silvio Savarese. 2024. Retroformer: Retrospective large language agents with policy gradient optimization. *Preprint*, arXiv:2308.02151.

Botao Yu, Frazier N. Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024. LlaSMol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. In *Proceedings of the First Conference on Language Modeling*.

Botao Yu, Frazier N. Baker, Ziru Chen, Garrett Herb, Boyu Gou, Daniel Adu-Ampratwum, Xia Ning, and Huan Sun. 2025a. Chemtoolagent: The impact of tools on language agents for chemistry problem solving. *Preprint*, arXiv:2411.07228.

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

Baosen Zhang, Thomas J. Summers, Logan J. Augustine, Michael G. Taylor, Andreas Geist, Rebecca Li, Enrique R. Batista, Danny Perez, Ping Yang, and Joshua Schrier. 2026. Augmenting large language models for automated discovery of F-element extractants. *Journal of the American Chemical Society*, 148(5):5520–5532.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, Dongzhan Zhou, Shufei Zhang, Mao Su, Han-Sen Zhong, and Yuqiang Li. 2024. Chemllm: A chemical large language model. *Preprint*, arXiv:2402.06852.

Odin Zhang, Haitao Lin, Xujun Zhang, Xiaorui Wang, Zhenxing Wu, Qing Ye, Weibo Zhao, Jike Wang, Kejun Ying, Yu Kang, Chang-Yu Hsieh, and Tingjun Hou. 2025a. Graph neural networks in modern AI-aided drug discovery. *Chemical Reviews*, 125(20):10001–10103.

Yu Zhang, Yang Han, Shuai Chen, Ruijie Yu, Xin Zhao, Xianbin Liu, Kaipeng Zeng, Mengdi Yu, Jidong Tian, Feng Zhu, Xiaokang Yang, Yaohui Jin, and Yanyan Xu. 2025b. Large language models to accelerate organic chemistry synthesis. *Preprint*, arXiv:2504.18340.

Zimin Zhang, Qianli Wu, Botao Xia, Fang Sun, Ziniu Hu, Yizhou Sun, and Shichang Zhang. 2025c. Automated molecular concept generation and labeling with large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6918–6936, Abu Dhabi, UAE. Association for Computational Linguistics.

Haiteng Zhao, Shengchao Liu, Chang Ma, Hannan Xu, Jie Fu, Zhi-Hong Deng, Lingpeng Kong, and Qi Liu. 2023. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. *Preprint*, arXiv:2306.13089.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, Kai Yu, and Xin Chen. 2025. Developing ChemDFM as a large language foundation model for chemistry. *Cell Reports Physical Science*, 6(4).

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh T. N. Nguyen, Lauren T. May, Geoffrey I. Webb, and Shirui Pan. 2025. Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence*, 7(3):437–447.

Peng Zhou, Lai Hou Tim, Zhixiang Cheng, Kun Xie, Chaoyi Li, Wei Liu, and Xiangxiang Zeng. 2025. Enhancing molecular property prediction with knowledge from large language models. *Preprint*, arXiv:2509.20664.

Chenyang Zuo, Siqi Fan, and Zaiqing Nie. 2025. Biomedgpt-mol: Multi-task learning for molecular understanding and generation. *Preprint*, arXiv:2512.04629.