# A Survey of Large Language Models for Text-Guided Molecular Discovery: from Molecule Generation to Optimization

Anonymous ACL submission

## Abstract

Large language models (LLMs) are introducing a paradigm shift in molecular discovery by enabling text-guided interaction with chemical spaces through natural language and symbolic notations, with emerging extensions to incorporate multi-modal inputs. To advance this emerging field, this survey provides an up-to-date and forward-looking review of the emerging use of LLMs for two central tasks: molecule generation and molecule optimization. We organize our survey around four fundamental challenges that have emerged as critical evaluation dimensions in recent studies: ensuring validity, enhancing synthesizability, achieving precise property control, and maximizing diversity. Based on this, we systematically analyze how current LLM learning paradigms are applied to tackle each challenge, revealing the distinct capabilities and inherent limitations of each approach. In addition, we include the commonly used datasets and evaluation protocols aligned with these challenges. We conclude by discussing future directions, positioning this survey as a resource for researchers working at the intersection of LLMs and molecular science. A continuously updated reading list is available at https://anonymous.4open.science/r/LLM-Centric-Molecular-Discovery.

## 1 Introduction

Molecular design and optimization are fundamental to multiple scientific disciplines, including drug discovery (Zheng et al., 2024), materials science (Grandi et al., 2025), and synthetic chemistry (Lu et al., 2024; Wang et al., 2025). However, these tasks present significant challenges due to the vast and complex chemical spaces that must be navigated to discover novel compounds with desirable properties while maintaining chemical validity and structural plausibility (Zheng et al., 2024; Yu et al., 2025). Over the years, a range of computational approaches has been developed to achieve these goals, from Variational Autoencoders (Gómez-Bombarelli et al., 2018) and Generative Adversarial Networks (De Cao and Kipf, 2018) to Transformers (Edwards et al., 2022). Despite significant progress, these methods often struggle with generating high-quality, diverse, and synthesizable molecules (Ramos et al., 2025; Sun et al., 2025).

More recently, large language models (LLMs) have emerged as particularly powerful tools for tackling these challenges, drawing increasing research attention (Zheng et al., 2024). These foundation models, characterized by billions of parameters, exhibit emergent capabilities such as advanced reasoning, instruction following, and in-context learning, enabled by extensive pre-training on diverse datasets (Brown et al., 2020; Wei et al., 2022a). Thus, LLMs can leverage their extensive pre-training knowledge to generalize across chemical problems and can be further adapted to specialized tasks through fine-tuning. These unique capabilities have established LLMs as a powerful new paradigm for exploring chemical space and accelerating molecular discovery.

Despite the growing interest in applying LLMs to molecular discovery tasks, existing literature reviews fail to provide a comprehensive analysis of this specific intersection. Most earlier surveys (Cheng et al., 2021; Zeng et al., 2022; Tang et al., 2024; Yang et al., 2024b) focus broadly on general deep generative AI approaches rather than specifically examining LLMs' unique contributions. Other reviews that do mention LLMs (Ramos et al., 2025; Zhang et al., 2025; Guo et al., 2025; AbuNasser, 2024; Janakarajan et al., 2024; Liao et al., 2024) either primarily focus on the general chemical domain or include smaller language models ($<$ 1B parameters) that lack the emergent capabilities of the LLMs central to this survey.

Our survey addresses this critical gap by providing the first overview specifically focused on LLMs in molecular discovery, with particular emphasis on two central tasks: **molecule generation** and **molecule optimization**. We focus on foundation-scale models (>1B parameters) and adopt a multi-dimensional assessment framework based on recent benchmarking studies (Brown et al., 2019; Polykovskiy et al., 2020; Thomas et al., 2024). We organize our survey around four fundamental challenges: **validity** (whether molecules are chemically feasible), **synthesizability** (whether they can be practically synthesized), **property control** (whether they meet desired objectives), and **diversity** (whether they explore chemical space broadly). Unlike prior surveys that categorize studies based on model architectures (AbuNasser, 2024; Janakarajan et al., 2024), we introduce a taxonomy centered on learning paradigms—distinguishing between approaches *without LLM tuning* (Zero-Shot Prompting and In-Context Learning) and those *with LLM tuning* (Supervised Fine-Tuning and Preference Tuning), as illustrated in Fig. 1. To summarize, our main contributions are as follows:

- We introduce a new taxonomy based on learning paradigms, revealing how different approaches address the four fundamental chemical challenges and their respective limitations.
- We provide a systematic summary of commonly used datasets, benchmarks, and evaluation metrics, offering a comprehensive reference for researchers in the field.
- We identify critical challenges and outline promising future research directions to further advance this rapidly evolving domain of LLM-centric molecular discovery.

## 2 Preliminaries

### 2.1 Large Language Models

LLMs distinguish themselves from earlier Pre-trained Language Models (PLMs) like BERT (Devlin et al., 2019) primarily through their massive scale—billions versus millions of parameters—and the resultant emergent capabilities (Zhao et al., 2023; Yang et al., 2023). Pre-trained on vast text corpora using autoregressive objectives, LLMs exhibit capabilities such as in-context learning (Brown et al., 2020), chain-of-thought reasoning (Wei et al., 2022b), and powerful zero-shot generalization that are not consistently observed in smaller models (Wei et al., 2022a). These emergent capabilities make LLMs uniquely suited for complex chemical applications like molecule generation and optimization tasks central to this survey.

### 2.2 Problem Definition and Scope

This survey focuses on LLM-centric approaches to molecular discovery, with two key inclusion criteria: (1) models must have at least **1B parameters** to ensure emergent capabilities, and (2) LLMs must serve as **molecular generators** rather than auxiliary components like feature extraction (Liu et al., 2023) or control (Liu et al., 2024a). Under this scope, we examine two central tasks:

**Problem Definition 1 (LLM-centric Molecule Generation).** *This task leverages LLMs for the de novo design of novel molecular structures based on specified input instructions.*

**Problem Definition 2 (LLM-centric Molecule Optimization).** *This task leverages LLMs to modify or edit a given input molecule, aiming to enhance one or more of its properties while often preserving essential structural characteristics.*

As illustrated in Fig. 2, for both tasks, the input prompt provided to the LLM typically comprises three key components: (1) **Instruction** ($\mathcal{I}$): A textual component that defines the primary guidance and objectives of the task. (2) **Few-Shot Examples** ($E_{fs}$) (Optional): A small set of input-output examples relevant to the task, provided to facilitate in-context learning. (3) **Property Constraints** ($\mathcal{C}_p$) (Optional): Explicit desired values, ranges, or thresholds for specific molecular properties.

### 2.3 Challenges in Molecular Discovery

Based on recent research studies and established evaluation practices (Brown et al., 2019; Polykovskiy et al., 2020; Thomas et al., 2024), we identify four fundamental challenges that comprehensively capture the unique requirements of molecular discovery. These challenges form a multi-dimensional framework for evaluating LLM-based approaches, as they collectively represent the critical aspects that distinguish chemical generation from general text generation:

- **Validity:** Generated molecules must adhere to fundamental chemical rules (e.g., valency) to be structurally meaningful. Unlike grammatically incorrect sentences, an invalid molecule is physically impossible and unusable (Jin et al., 2018).

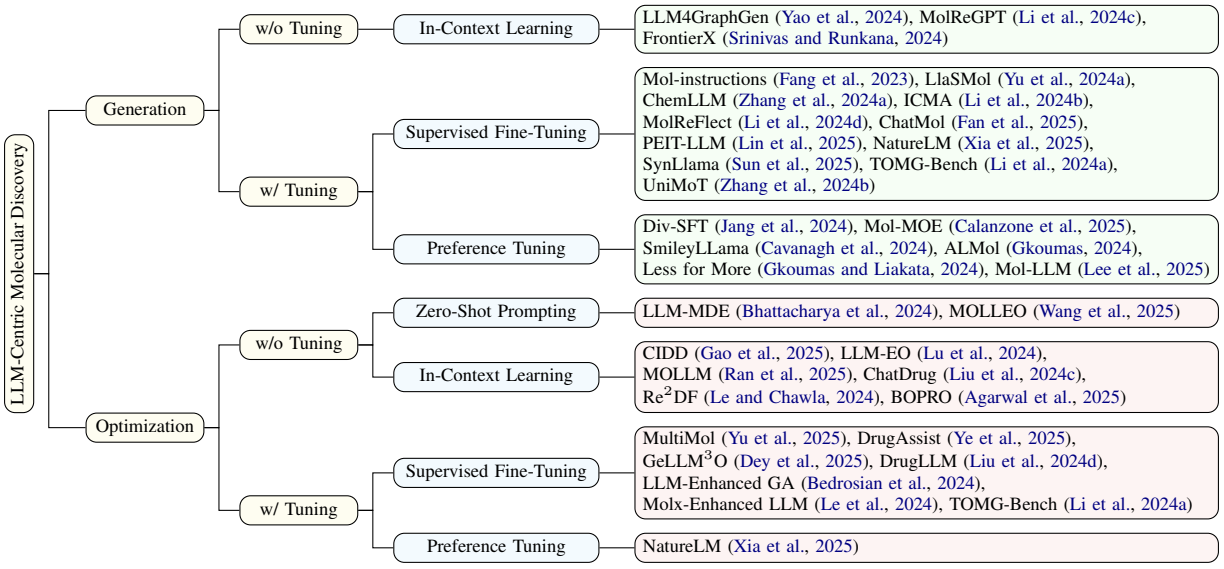


Figure 1: A Taxonomy of LLM-Centric Molecular Discovery.

- **Synthesizability:** A valid molecule must also be practically synthesizable. This requires considering the feasibility and complexity, as a theoretically valid structure may be impossible to create in a lab (Gao and Coley, 2020).
- **Property Control:** The design process must precisely steer molecules toward desired properties, often requiring the simultaneous optimization of multiple, competing objectives (You et al., 2018).
- **Diversity:** To effectively explore the vast chemical space, generated molecules must be structurally diverse, avoiding minor variations of known compounds (Zhavoronkov et al., 2019).

These challenges are interconnected and often conflicting (Gao and Coley, 2020), forming a comprehensive evaluation framework that tests multiple dimensions of LLMs' capabilities in molecular discovery. Throughout this survey, we systematically analyze how different learning paradigms address these competing objectives, revealing their respective strengths and limitations in tackling the full spectrum of molecular design requirements.

### 2.4 Learning Paradigms

The application of LLMs to molecular discovery tasks, as depicted in the taxonomy in Fig. 2, can be broadly categorized based on whether the model's parameters are updated for the specific task. This distinction defines two primary learning paradigms:

**Without LLM Tuning:** These methods utilize pre-trained LLMs directly, guiding their behavior solely through the input prompt $\mathcal{I}$ without modifying the model's weights. This paradigm primarily encompasses strategies like *Zero-Shot Prompting*, where the LLM operates based on instructions alone, and *In-Context Learning (ICL)*, where few-shot examples provided within the prompt guide the model's responses. These approaches avoid computational training but rely heavily on the LLM's inherent capabilities and effective prompt engineering.

**With LLM Tuning:** These methods involve adapting the pre-trained LLM by further training and updating its parameters to specialize it for molecular tasks or align its outputs with desired objectives. This typically includes *Supervised Fine-Tuning (SFT)*, where the model learns from labeled task-specific datasets, and subsequent *Preference Tuning* (or Alignment), where the model is refined based on feedback. While tuning can significantly enhance performance, it requires curated data and computational resources.

## 3 Molecule Generation

Molecule generation, the computational creation of novel molecular structures, is a cornerstone of modern drug discovery and materials science (Elton et al., 2019). This section reviews recent advances in LLM-centric molecule generation, analyzing how different learning paradigms address the four fundamental challenges while creating molecules from scratch.

### 3.1 Molecule Generation without Tuning

#### 3.1.1 In-Context Learning

**Property Control:** Since *Zero-Shot Prompting* is challenging for general-purpose LLMs due to their lack of specialized chemical knowledge, most

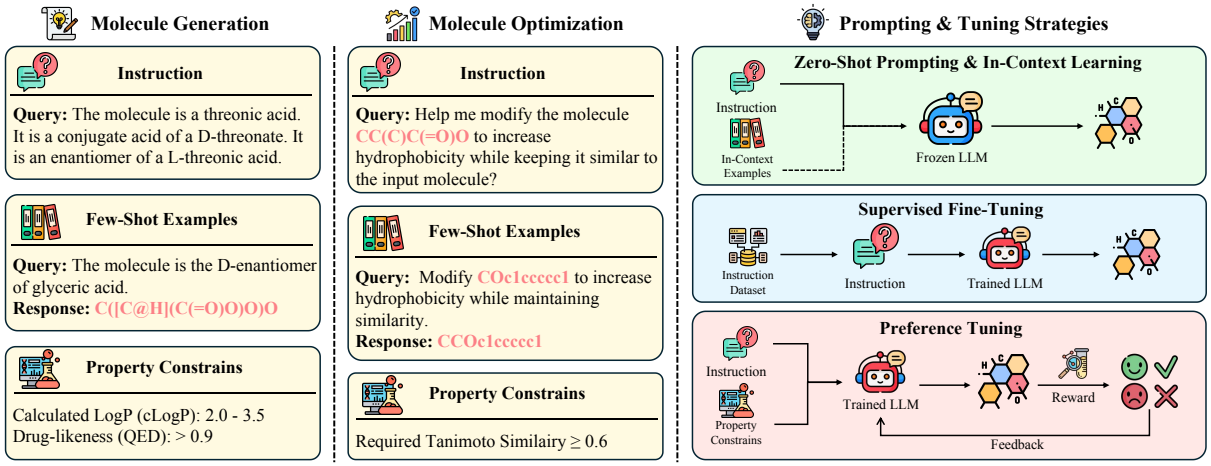


Figure 2: **Overview of LLM-Centric Molecular Discovery.** ***Left:*** Typical input components (Instruction, Few-Shot Examples, Property Constraints) for molecule generation and optimization. ***Right:*** Core learning paradigms for applying LLMs to *Zero-Shot Prompting & In-Context Learning*, *Supervised Fine-Tuning* and *Preference Tuning*.

successful applications in this paradigm rely on In-Context Learning (ICL). This approach primarily addresses the challenge of Property Control by providing high-quality examples to guide generation, as demonstrated in works like *FrontierX* (Srinivas and Runkana, 2024) and *LLM4GraphGen* (Yao et al., 2024). Recognizing that example quality is paramount, a key technical advance is the use of Retrieval-Augmented Generation (RAG). For instance, *MolReGPT* (Li et al., 2024c) incorporates RAG to dynamically retrieve the most relevant molecule-caption pairs, creating a more effective context and significantly boosting performance.

In summary, ICL excels at providing guidance for property control. However, it struggles with validity, synthesizability, and diversity—limitations stemming from its reliance on pattern matching rather than learning chemical principles.

## 3.2 Molecule Generation with Tuning

### 3.2.1 Supervised Fine-Tuning

While non-tuning methods leverage pre-trained knowledge, their capabilities are often limited for specialized generation tasks. SFT addresses this by adapting a pre-trained LLM on labeled datasets, typically pairs of textual instructions and target molecular representations. This approach moves beyond the capabilities of smaller models like MolGPT (Bagal et al., 2021) and MolT5 (Edwards et al., 2022) by harnessing the power of large foundation models.

**Validity:** SFT is the primary paradigm for instilling foundational chemical knowledge into LLMs, making it highly effective for ensuring validity. By fine-tuning on millions of valid molecular structures, the LLM learns the complex "grammar" of chemical representations like SMILES. This foundational training is the focus of several large-scale instruction-tuning efforts, such as *LlaSMol* (Yu et al., 2024a) with its SMolInstruct dataset, *ChemLLM* (Zhang et al., 2024a) with ChemData, Mol-Instructions (Fang et al., 2023), and the OpenMolIns dataset from *TOMG-Bench* (Li et al., 2024a). To further improve structural understanding, multi-modal SFT approaches like *UniMoT* (Zhang et al., 2024b) incorporate 2D graph information directly into the training process by converting molecular graphs into discrete "molecule tokens," enhancing the model's ability to generate valid and complex molecules.

**Property Control:** SFT enables LLMs to learn the intricate mapping between desired properties and molecular structures. This is where instruction tuning truly shines. For instance, *ChatMol* (Fan et al., 2025) directly tackles the need for precise numerical control by using an enhancement technique to improve the model's fidelity to specific quantitative property values. Addressing the need for multi-property optimization, *PEIT-LLM* (Lin et al., 2025) proposes a two-step framework to fine-tune LLMs for multi-constraint generation. To improve the quality of guidance during training, other innovative strategies integrate retrieval directly into the fine-tuning process. *ICMA* (Li et al., 2024b) and *MolReFlect* (Li et al., 2024d), for example, propose In-Context Molecule Tuning (ICMT), which fine-tunes the LLM using relevant retrieved examples to better align outputs with complex instructions.

**Synthesizability:** SFT is beginning to address synthesizability. *SynLlama* (Sun et al., 2025) was

developed to specifically tackle synthetic feasibility by fine-tuning the model to generate not just molecules, but also complete synthetic pathways.

In summary, SFT excels at validity through extensive training on chemical structures, provides strong property control via instruction tuning, and shows emerging capabilities in synthesizability assessment. However, its reliance on training data distributions limits diversity, often causing mode collapse where models generate variations of known scaffolds.

#### 3.2.2 Preference Tuning

Following SFT, which teaches models to mimic static datasets, Preference Tuning techniques offer further refinement by employing feedback-driven learning to shape LLM outputs towards desired characteristics. This is achieved either through RL-based methods (Sutton et al., 1998) that optimize a policy against a reward signal, or offline methods like Direct Preference Optimization (DPO) that learn from "chosen" vs. "rejected" pairs.

**Diversity:** Preference Tuning directly addresses the primary limitation of SFT by excelling at enhancing diversity. By explicitly rewarding novel and varied molecular structures, it encourages exploration of underrepresented chemical spaces. *Div-SFT* (Jang et al., 2024), for example, employs RL with a reward function specifically designed to maximize structural diversity, effectively mitigating SFT's tendency toward mode collapse.

**Property Control:** Preference-based methods also significantly improve multi-property optimization. *SmileyLlama* (Cavanagh et al., 2024) utilizes DPO to improve adherence to property constraints by learning from preferences between correct and incorrect molecules. *Mol-MoE* (Calanzone et al., 2025) uses a preference objective to train a Mixture-of-Experts router, enabling specialization for different property requirements. Contrastive methods like CPO (Xu et al., 2024) also refine molecule quality by learning from comparative data (Gkoumas, 2024; Gkoumas and Liakata, 2024).

**Validity:** Beyond text-based approaches, preference tuning can enhance validity by improving how models utilize structural information. *Mol-LLM* (Lee et al., 2025) addresses the "graph bypass phenomenon" where models ignore 2D structural inputs. Through Molecular Structure Preference Optimization (MolPO), it trains the model to distinguish between correct and perturbed molecular graphs, forcing deeper engagement with structural information and thereby improving the validity.

In summary, Preference Tuning excels at diversity by explicitly rewarding novelty, provides refined multi-property control through comparative learning, and can enhance validity in multi-modal settings. However, it offers no direct improvement to synthesizability and requires substantial effort to obtain high-quality preference data or design appropriate reward functions.

## 4 Molecule Optimization

Molecule optimization is the task of refining molecular structures to improve one or more desired properties, such as solubility, binding affinity, or synthetic accessibility. Unlike molecule generation, optimization starts with an initial molecule and proposes targeted structural modifications to achieve specific goals. This section summarizes LLM-centric molecule optimization methods, analyzing how different learning paradigms address the four fundamental challenges in this more constrained but equally important task.

### 4.1 Molecule Optimization without Tuning

#### 4.1.1 Zero-Shot Prompting

**Property Control:** Zero-Shot Prompting leverages the pre-trained knowledge of LLMs to perform edits based on natural language instructions alone. This paradigm enables flexible property modification through natural language specifications. For example, *LLM-MDE* (Bhattacharya et al., 2024) uses detailed prompts to specify desired property changes and structural constraints, enabling controlled modifications. *MOLLEO* (Wang et al., 2025) integrates LLMs into evolutionary frameworks, using prompt-based sampling to perform mutations and crossovers.

In summary, zero-shot prompting excels at expressing diverse optimization goals flexibly, but its reliance on general pre-trained knowledge results in limited precision for property control and poor performance on validity and synthesizability.

#### 4.1.2 In-Context Learning

**Property Control:** ICL enhances property control by providing examples of successful molecular edits within the prompt. This allows the LLM to learn optimization patterns from context. *CIDD* (Gao et al., 2025) implements a multi-step pipeline of

interaction analysis, design, and reflection, feeding previous designs back into the context. *LLM-EO* (Lu et al., 2024) and *MOLLM* (Ran et al., 2025) integrate LLMs into evolutionary algorithms, where historical data from previous generations serves as in-context examples. *BOPRO* (Agarwal et al., 2025) combines ICL with Bayesian optimization for more sophisticated example selection.

**Validity:** To improve validity, retrieval-augmented methods enhance example quality. *ChatDrug* (Liu et al., 2024c) retrieves structurally similar molecules to inform proposals, while *Re$^2$DF* (Le and Chawla, 2024) incorporates validity feedback from RDKit (Landrum et al., 2013) directly into the prompt to guide the model toward valid outputs.

In summary, ICL offers more guided and iterative control than zero-shot methods through example-based learning, improving both property control and validity. However, its effectiveness depends heavily on example quality, and it still provides limited solutions for ensuring synthesizability or enhancing diversity.

## 4.2 Molecule Optimization with Tuning

### 4.2.1 Supervised Fine-Tuning

SFT adapts pre-trained LLMs for molecule optimization by training them on curated datasets of input molecules paired with their corresponding optimized outputs. This supervision allows the model to learn how to perform controlled structural edits based on specific objectives.

**Property Control:** While smaller Transformer-based chemical language models have shown potential for optimization tasks (Ross et al., 2022, 2024; Wu et al., 2024b; Dai et al., 2025; Liu et al., 2025c), foundation-scale LLMs enable more advanced capabilities through SFT. By training on instruction datasets, models learn precise single- and multi-property optimization. *DrugAssist* (Ye et al., 2025) fine-tunes LLaMA-2-7B-Chat on the MolOpt-Instructions dataset for single/dual-property tasks. *GeLLM$^3$O* (Dey et al., 2025) extends this to multi-property optimization with strong out-of-distribution generalization. *MultiMol* (Yu et al., 2025) employs a collaborative framework where a fine-tuned worker generates candidates and a research agent (GPT-4o) ranks them using literature-derived knowledge. *DrugLLM* (Liu et al., 2024d) introduces group-based molecular representation (GMR) to better align structure and semantics for controlled modifications.

**Diversity:** SFT enables population-based optimization that balances property improvement with diversity. *LLM-Enhanced GA* (Bedrosian et al., 2024) replaces traditional genetic operators with prompt-based sampling from high-performing molecules, incorporating explicit oracle modeling through SFT when performance stagnates to progressively refine understanding of the property landscape.

**Validity:** Multi-modal SFT approaches enhance validity by incorporating richer structural information (Zhang et al., 2024c; Lin et al., 2024; Nakamura et al., 2025). *Molx-Enhanced LLM* (Le et al., 2024) integrates SMILES, 2D graphs, and fingerprints into a unified embedding. Through fine-tuning the multi-modal MolX module, the model captures both global topology and local substructures essential for chemically valid modifications.

In summary, SFT excels at precise property control through explicit instruction-based training and shows promise for diversity in population-based frameworks. Multi-modal SFT further enhances validity by leveraging structural information. However, its effectiveness remains tied to training data quality, with limited inherent capabilities for assessing synthesizability.

### 4.2.2 Preference Tuning

Preference Tuning refines tuned LLMs by aligning them with task-specific goals or preferences (Park et al., 2025; Chen et al., 2025). While RL-based alignment techniques built on smaller Transformer architectures have shown promise (Liu et al., 2025b,d), the application of offline preference methods to large foundation models has enabled more scalable optimization.

**Property Control:** Preference tuning excels at multi-property optimization through comparative learning. *NatureLM* (Xia et al., 2025) exemplifies this approach by augmenting its post-trained 8B model using Direct Preference Optimization (DPO). Instead of training on absolute labels or scalar rewards, the model learns from 179.5k prompt-response pairs, where each instance contains a "preferred" and "rejected" molecular output for the same optimization goal. By learning from these comparative preferences, NatureLM demonstrates improved alignment across nine pharmacologically relevant properties, showcasing DPO's

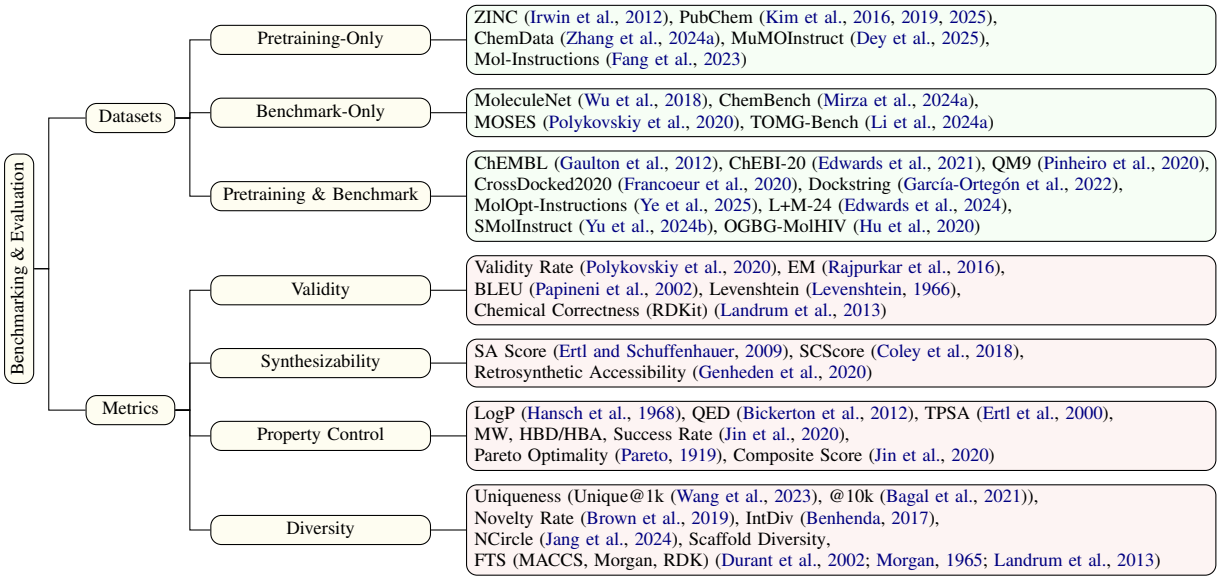


Figure 3: A Taxonomy of Benchmarking & Evaluation in Molecule Discovery.

ability to generalize preference-guided optimization across diverse chemical objectives.

In summary, Preference Tuning, particularly through DPO, provides a powerful and scalable solution for multi-objective property control in optimization tasks. However, it requires significant effort to curate high-quality preference datasets, and its application to other chemical challenges (validity, synthesizability, diversity) remains limited in the optimization domain.

## 5 Benchmarking and Evaluation

Rigorous benchmarking and comprehensive evaluation are crucial for tracking the progress of LLM-centric molecular discovery. This section provides an overview of the evaluation ecosystem, organized around our four fundamental challenges, with comprehensive details available in the appendices.

### 5.1 Datasets

Molecular datasets serve distinct purposes in LLM development, ranging from large-scale pretraining to targeted evaluation. **Pretraining-Only Datasets** like ZINC (Irwin et al., 2012) provide vast chemical structures, while instruction collections like ChemData (Zhang et al., 2024a) offer domain-specific knowledge for teaching chemical reasoning. **Benchmark-Only Datasets** include TOMG-Bench (Li et al., 2024a) for text-guided generation and MOSES (Polykovskiy et al., 2020) for distribution learning. **Dual-Purpose Datasets** such as ChEMBL (Gaulton et al., 2012) support both training and evaluation, enabling consistent benchmarking across different stages of model development. See Appendix B for detailed comparisons

### 5.2 Metrics

Evaluation metrics directly address our four fundamental challenges. **Validity Metrics** include SMILES parsing, uniqueness rates (Unique@1k, Unique@10k), and chemical correctness checks. **Synthesizability Metrics** employ SA Score (Ertl and Schuffenhauer, 2009) and SCScore (Coley et al., 2018) for complexity prediction. **Property Control Metrics** span single-property evaluations (QED (Bickerton et al., 2012), LogP (Hansch et al., 1968), TPSA (Ertl et al., 2000)) and multi-property optimization via success rates and Pareto optimality. **Diversity Metrics** assess chemical space exploration through novelty rate, internal diversity (IntDiv) (Benhenda, 2017), and scaffold analysis. Mathematical definitions and implementation details are provided in Appendix C.

### 5.3 External Tools

Evaluation requires diverse computational tools that bridge chemistry and machine learning. General cheminformatics relies on RDKit (Landrum et al., 2013) for property calculation and validation, OpenBabel (O'Boyle et al., 2011) for format conversion, and CDK (Willighagen et al., 2017) for Java environments. Synthesizability assessment employs AiZynthFinder (Genheden et al., 2020) and ASKCOS (Coley et al., 2019) for retrosynthetic planning. LLM-specific tools like ChemCrow (M. Bran et al., 2024) integrate language models with chemistry tools. Detailed usage guidelines are in Appendix D.

### 5.4 Evaluation Frameworks

Standardized frameworks have evolved from classic to LLM-specific approaches. GuacaMol (Brown et al., 2019) pioneered dual evaluation via distribution learning and goal-directed tasks, while MOSES (Polykovskiy et al., 2020) focused on comprehensive distribution metrics. Recent frameworks address modern needs: MolScore (Thomas et al., 2024) unifies previous benchmarks with modular scoring, TDC (Huang et al., 2021) provides continuously updated leaderboards, and LLM-specific benchmarks like TOMG-Bench (Li et al., 2024a) evaluate instruction-following capabilities. However, all frameworks rely on computational validation without experimental verification—a critical limitation discussed in Appendix E.

## 6 Conclusion and Future Work

This survey presents the first comprehensive review of recent advances in LLM-centric molecular discovery, covering both generation and optimization tasks. We introduced a novel taxonomy that categorizes approaches based on their learning paradigms—distinguishing between methods without LLM tuning (zero-shot prompting and in-context learning) and those with LLM tuning (supervised fine-tuning and preference tuning). Through systematic analysis of how these approaches address four fundamental challenges—validity, synthesizability, property control, and diversity—we uncovered key patterns in the current landscape.

**Key Insights:** Our analysis reveals that no single approach dominates across all challenges, with each exhibiting distinct trade-offs. Zero-Shot prompting offers unmatched flexibility for diverse tasks but struggles with chemical validity and precise property control. ICL improves guidance through carefully selected examples but remains fundamentally limited by example quality and lacks a systematic understanding of chemical principles. SFT excels at ensuring validity through large-scale chemical training and enables precise property control via instruction tuning, yet often suffers from limited diversity due to mode collapse. Preference tuning emerges as the primary solution for diversity through reward-based exploration while maintaining multi-property optimization capabilities. However, across all methods, synthesizability remains the most poorly addressed challenge—current approaches generate molecules that are computationally valid but often practically impossible to synthesize, representing a critical bottleneck for real-world deployment.

Based on these insights and current limitations, we identify three priority areas for advancing the field:

**Prioritizing Synthesizability in Generation:** As illustrated in recent analyses (Walters, 2024), current LLMs frequently produce molecules through string manipulation rather than chemical understanding, resulting in theoretically valid but synthetically inaccessible structures. Future work must move beyond post-hoc SA Score filtering to incorporate synthesizability as a primary constraint during generation. This includes: (i) training on datasets of successfully synthesized molecules; (ii) integrating retrosynthetic planning directly into the generation process; (iii) developing reward functions that explicitly penalize synthetic complexity during preference tuning.

**Multi-Modal Molecular Understanding:** Current LLM approaches predominantly operate on SMILES strings, missing crucial structural information. Future architectures should jointly encode and reason over multiple representations—SMILES strings, 2D molecular graphs, 3D conformations, and quantum chemical properties (Lu et al., 2023; Pirnay et al., 2025). This requires developing unified tokenization schemes that preserve chemical semantics across modalities while enabling efficient transformer processing.

**Unified Benchmarks for LLM-Based Molecular Design:** Current frameworks like MOSES and GuacaMol were designed for traditional generative models and lack standardization for LLM evaluation. We urgently need a unified benchmark with: (i) standardized train/validation/test splits specifically curated for LLMs, preventing data leakage and ensuring fair comparison across models; (ii) comprehensive evaluation metrics that go beyond traditional measures to include LLM-specific capabilities such as instruction-following accuracy, multi-step reasoning ability, and robustness to representation variations (SMILES, IUPAC, natural language); (iii) a continuously updated leaderboard tracking progress in LLM-based molecular design. Such a unified benchmark would provide the community with a clear view of where we stand and where we need to improve in applying LLMs to molecular discovery.

## 7 Limitations

This survey focuses on the use of large language models for two core tasks in text-guided molecular discovery: molecule generation and molecule optimization. These tasks represent the most direct applications of LLMs in molecular design and are the primary scope of current research. We acknowledge that LLMs can also significantly impact other important areas of molecular science, such as reaction prediction, retrosynthesis, protein-ligand modeling, and automated experimentation (Zhang et al., 2024d; Liu et al., 2024b, 2025a). Additionally, while we focus on models with $>$ 1B parameters to ensure emergent capabilities, specialized chemical language models below this threshold remain valuable for specific applications. Given the broad and rapidly evolving landscape, we leave a systematic review of these additional directions to future work. By maintaining this focused scope, we provide a detailed resource for researchers working on LLM-driven molecular generation and optimization, while recognizing that experimental validation of computationally generated molecules remains a critical challenge beyond the scope of computational metrics discussed here.

## References


Raghad AbuNasser. 2024. Large language models in drug discovery: A survey.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Dhruv Agarwal, Manoj Ghuhan Arivazhagan, Rajarshi Das, Sandesh Swamy, Sopan Khosla, and Rashmi Gangadharaiah. 2025. Searching for optimal solutions with llms via bayesian optimization. In *The Thirteenth International Conference on Learning Representations*.

Anthropic. 2024a. Claude-3.

Anthropic. 2024b. Claude-3.5.

Evan R Antoniuk, Shehtab Zaman, Tal Ben-Nun, Peggy Li, James Diffenderfer, Busra Demirci, Obadiah Smolenski, Tim Hsu, Anna M Hiszpanski, Kenneth Chiu, and 1 others. 2025. Boom: Benchmarking out-of-distribution molecular property predictions of machine learning models. *arXiv preprint arXiv:2505.01912*.

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

Menua Bedrosian, Philipp Guevorguian, Tigran Fahradyan, Gayane Chilingaryan, Hrant Khachatrian, and Armen Aghajanyan. 2024. Small molecule optimization with large language models. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*.

Mostapha Benhenda. 2017. Chemgan challenge for drug discovery: can ai reproduce natural chemical diversity? *arXiv preprint arXiv:1708.08227*.

Debjyoti Bhattacharya, Harrison J Cassady, Michael A Hickner, and Wesley F Reinhart. 2024. Large language models as molecular design engines. *Journal of Chemical Information and Modeling*, 64(18):7086–7096.

GR Bickerton, GV Paolini, J Besnard, S Muresan, and AL Hopkins. 2012. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98.

Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. 2019. Guacamol: Benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3):1096–1108.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *NeurIPS*, 33:1877–1901.

Diego Calanzone, Pierluca D'Oro, and Pierre-Luc Bacon. 2025. Mol-moe: Training preference-guided routers for molecule generation. *arXiv preprint arXiv:2502.05633*.

Joseph M Cavanagh, Kunyang Sun, Andrew Gritsevskiy, Dorian Bagni, Thomas D Bannister, and Teresa Head-Gordon. 2024. Smileyllama: Modifying large language models for directed chemical space exploration. *arXiv preprint arXiv:2409.02231*.

Angelica Chen, Samuel D. Stanton, Frances Ding, Robert G. Alberstein, Andrew M. Watkins, Richard Bonneau, Vladimir Gligorijević, Kyunghyun Cho, and Nathan C. Frey. 2025. Generalists vs. specialists: Evaluating llms on highly-constrained biophysical sequence optimization tasks.

Yu Cheng, Yongshun Gong, Yuansheng Liu, Bosheng Song, and Quan Zou. 2021. Molecular design in drug discovery: a comprehensive review of deep generative models. *Briefings in bioinformatics*, 22(6):bbab344.

Connor W Coley, Luke Rogers, William H Green, and Klavs F Jensen. 2018. Scscore: synthetic complexity learned from a reaction corpus. *Journal of chemical information and modeling*, 58(2):252–261.

Connor W Coley, Dale A Thomas III, Justin AM Lummiss, Jonathan N Jaworski, Christopher P Breen, Victor Schultz, Travis Hart, Joshua S Fishman, Luke Rogers, Hanyu Gao, and 1 others. 2019. A robotic platform for flow synthesis of organic compounds informed by ai planning. *Science*, 365(6453):eaax1566.

Zhilian Dai, Jie Zhang, Songyou Zhong, Jiawei Fu, Yangyang Deng, Dan Zhang, Yichao Liu, and Peng Gao.

2025. A zero-shot single-point molecule optimization model: Mimicking medicinal chemists' expertise.

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

Google Deepmind. 2024. Gemini-1.5-pro.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Vishal Dey, Xiao Hu, and Xia Ning. 2025. Gellm$^3$o: Generalizing large language models for multi-property molecule optimization. *arXiv preprint arXiv:2502.13398*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

JL Durant, BA Leland, DR Henry, and JG Nourse. 2002. Reoptimization of mdl keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*.

Carl Edwards, Qingyun Wang, Lawrence Zhao, and Heng Ji. 2024. L+ m-24: Building a dataset for language+ molecules@ acl 2024. *arXiv preprint arXiv:2403.00791*.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2mol: Cross-modal molecule retrieval with natural language queries. In *EMNLP*, pages 595–607.

Daniel C Elton, Zois Boukouvalas, Mark D Fuge, and Peter W Chung. 2019. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering*, 4(4):828–849.

Peter Ertl, Bernhard Rohde, and Paul Selzer. 2000. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry*, 43(20):3714–3717.

Peter Ertl and Ansgar Schuffenhauer. 2009. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1):8.

Chuanliu Fan, Ziqiang Cao, Zicheng Ma, Nan Yu, Yimin Peng, Jun Zhang, Yiqin Gao, and Guohong Fu. 2025. Chatmol: A versatile molecule designer based on the numerically enhanced large language model. *arXiv preprint arXiv:2502.19794*.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*.

Henri A Favre and Warren H Powell. 2014. *Nomenclature of Organic Chemistry: IUPAC Recommendations and Preferred Names 2013*. Royal Society of Chemistry.

Paul G Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B Iovanisci, Ian Snyder, and David R Koes. 2020. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of chemical information and modeling*, 60(9):4200–4215.

Veronika Ganeeva, Andrey Sakhovskiy, Kuzma Khrabrov, Andrey Savchenko, Artur Kadurin, and Elena Tutubalina. 2024. Lost in translation: Chemical language models and the misunderstanding of molecule structures. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12994–13013.

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. *arXiv preprint arXiv:2503.01376*.

Wenhao Gao and Connor W Coley. 2020. The synthesizability of molecules proposed by generative models. *Journal of chemical information and modeling*, 60(12):5714–5723.

Miguel García-Ortegón, Gregor NC Simm, Austin J Tripp, José Miguel Hernández-Lobato, Andreas Bender, and Sergio Bacallado. 2022. Dockstring: easy molecular docking yields better benchmarks for ligand design. *Journal of chemical information and modeling*, 62(15):3486–3502.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and 1 others. 2012. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107.

Samuel Genheden, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Bjerrum. 2020. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of cheminformatics*, 12(1):70.

Dimitris Gkoumas. 2024. Almol: Aligned language-molecule translation llms through offline preference contrastive optimisation. *arXiv preprint arXiv:2405.08619*.

Dimitris Gkoumas and Maria Liakata. 2024. Less for more: Enhanced feedback-aligned mixed llms for molecule caption generation and fine-grained nli evaluation. *arXiv preprint arXiv:2405.13984*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A

family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

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

Daniele Grandi, Yash Patawari Jain, Allin Groom, Brandon Cramer, and Christopher McComb. 2025. Evaluating large language models for material selection. *Journal of Computing and Information Science in Engineering*, 25(2):021004.

Huijie Guo, Xudong Xing, Yongjie Zhou, Wenjiao Jiang, Xiaoyi Chen, Ting Wang, Zixuan Jiang, Yibing Wang, Junyan Hou, Yukun Jiang, and 1 others. 2025. A survey of large language model for drug research and development. *IEEE Access*.

Corwin Hansch, John E Quinlan, and Gary L Lawrence. 1968. Linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids. *The journal of organic chemistry*, 33(1):347–350.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. 2021. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*.

John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. 2012. Zinc: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, 52(7):1757–1768.

Nikita Janakarajan, Tim Erdmann, Sarath Swaminathan, Teodoro Laino, and Jannis Born. 2024. Language models in molecular discovery. In *Drug Development Supported by Informatics*, pages 121–141. Springer.

Hyosoon Jang, Yunhui Jang, Jaehyung Kim, and Sungsoo Ahn. 2024. Can llms generate diverse molecules? towards alignment with structural diversity. *arXiv preprint arXiv:2410.03138*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2020. Multi-objective molecule generation using interpretable substructures. In *International Conference on Machine Learning*, pages 4849–4859. PMLR.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, and 1 others. 2019. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, and 1 others. 2025. Pubchem 2025 update. *Nucleic Acids Research*, 53(D1):D1516–D1525.

Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, and 1 others. 2016. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213.

Tobias Kreiman and Aditi S Krishnapriyan. 2025. Understanding and mitigating distribution shifts for machine learning force fields. *arXiv preprint arXiv:2503.08674*.

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.

Greg Landrum and 1 others. 2013. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8(31.10):5281.

Khiem Le and Nitesh V Chawla. 2024. Utilizing large language models in an iterative paradigm with domain feedback for molecule optimization. *arXiv preprint arXiv:2410.13147*.

Khiem Le, Zhichun Guo, Kaiwen Dong, Xiaobao Huang, Bozhao Nan, Roshni Iyer, Xiangliang Zhang, Olaf Wiest, Wei Wang, and Nitesh V Chawla. 2024. Molx: Enhancing large language models for molecular learning with a multi-modal extension. *arXiv preprint arXiv:2406.06777*.

Chanhui Lee, Yuheon Song, YongJun Jeong, Hanbum Ko, Rodrigo Hormazabal, Sehui Han, Kyunghoon Bae, Sungbin Lim, and Sungwoong Kim. 2025. Mol-llm: Generalist molecular llm with improved graph utilization. *arXiv preprint arXiv:2502.02810*.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710.

Jiatong Li, Junxian Li, Yunqing Liu, Dongzhan Zhou, and Qing Li. 2024a. Tomg-bench: Evaluating llms on text-based open molecule generation. *arXiv preprint arXiv:2412.14642*.

Jiatong Li, Wei Liu, Zhihao Ding, Wenqi Fan, Yuqiang Li, and Qing Li. 2024b. Large language models

are in-context molecule learners. *arXiv preprint arXiv:2403.04197*.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. 2024c. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *IEEE transactions on knowledge and data engineering*.

Jiatong Li, Yunqing Liu, Wei Liu, Jingdi Le, Di Zhang, Wenqi Fan, Dongzhan Zhou, Yuqiang Li, and Qing Li. 2024d. Molreflect: Towards in-context fine-grained alignments between molecules and texts. *arXiv preprint arXiv:2411.14721*.

Chang Liao, Yemin Yu, Yu Mei, and Ying Wei. 2024. From words to molecules: A survey of large language models in chemistry. *arXiv preprint arXiv:2402.01439*.

Xiaohan Lin, Yijie Xia, Yupeng Huang, Shuo Liu, Jun Zhang, and Yi Qin Gao. 2024. Versatile molecular editing via multimodal and group-optimized generative learning.

Xuan Lin, Long Chen, Yile Wang, Xiangxiang Zeng, and Philip S. Yu. 2025. Property enhanced instruction tuning for multi-task molecule generation with large language models. *arXiv preprint arXiv:2412.18084*.

Gang Liu, Michael Sun, Wojciech Matusik, Meng Jiang, and Jie Chen. 2024a. Multimodal large language models for inverse molecular design with retrosynthetic planning. *arXiv preprint arXiv:2410.04223*.

Pengfei Liu, Jun Tao, and Zhixiang Ren. 2024b. Scientific language modeling: A quantitative review of large language models in molecular science. *arXiv preprint arXiv:2402.04119*, page 3.

Pengfei Liu, Jun Tao, and Zhixiang Ren. 2025a. A quantitative analysis of knowledge-learning preferences in large language models in molecular science. *Nature Machine Intelligence*, pages 1–13.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. 2022. Multi-modal molecule structure-text model for text-based retrieval and editing. *arXiv preprint arXiv:2212.10789*.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. 2023. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457.

Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. 2024c. Conversational drug editing using retrieval and domain feedback. In *The twelfth international conference on learning representations*.

Xianggen Liu, Yan Guo, Haoran Li, Jin Liu, Shudong Huang, Bowen Ke, and Jiancheng Lv. 2024d. Drugllm: Open large language model for few-shot molecule generation. *arXiv preprint arXiv:2405.06690*.

Xuefeng Liu, Songhao Jiang, Siyu Chen, Zhuoran Yang, Yuxin Chen, Ian Foster, and Rick Stevens. 2025b. Drugimprovergpt: A large language model for drug optimization with fine-tuning via structured policy optimization. *arXiv preprint arXiv:2502.07237*.

Xuefeng Liu, Songhao Jiang, Bo Li, and Rick Stevens. 2025c. Controllablegpt: A ground-up designed controllable gpt for molecule optimization. *arXiv preprint arXiv:2502.10631*.

Xuefeng Liu, Songhao Jiang, and Rick Stevens. 2025d. Scaffoldgpt: A scaffold-based large language model for drug improvement. *arXiv preprint arXiv:2502.06891*.

Hao Lu, Zhiqiang Wei, Xuze Wang, Kun Zhang, and Hao Liu. 2023. Graphgpt: A graph enhanced generative pretrained transformer for conditioned molecular generation. *International Journal of Molecular Sciences*, 24(23):16761.

Jieyu Lu, Zhangde Song, Qiyuan Zhao, Yuanqi Du, Yirui Cao, Haojun Jia, and Chenru Duan. 2024. Generative design of functional metal complexes utilizing the internal knowledge of large language models. *arXiv preprint arXiv:2410.18136*.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535.

A Mirza, N Alampara, S Kunchapu, B Emoekabu, A Krishnan, M Wilhelmi, M Okereke, J Eberhardt, AM Elahi, M Greiner, and 1 others. 2024a. Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475*.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, and 1 others. 2024b. Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475*.

HL Morgan. 1965. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113.

Shogo Nakamura, Nobuaki Yasuo, and Masakazu Sekijima. 2025. Molecular optimization using a conditional transformer for reaction-aware compound exploration with reinforcement learning. *Communications Chemistry*, 8(1):40.

Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. 2011. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):33.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Vilfredo Pareto. 1919. *Manuale di economia politica con una introduzione alla scienza sociale*, volume 13. Società editrice libraria.

Jinyeong Park, Jaegyoon Ahn, Jonghwan Choi, and Jibum Kim. 2025. Mol-air: Molecular reinforcement learning with adaptive intrinsic rewards for goal-directed molecular generation. *Journal of Chemical Information and Modeling*, 65(5):2283–2296.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2024. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. *arXiv preprint arXiv:2310.07276*.

Gabriel A Pinheiro, Johnatan Mucelini, Marinalva D Soares, Ronaldo C Prati, Juarez LF Da Silva, and Marcos G Quiles. 2020. Machine learning prediction of nine molecular properties based on the smiles representation of the qm9 quantum-chemistry dataset. *The Journal of Physical Chemistry A*, 124(47):9854–9866.

Jonathan Pirnay, Jan G Rittig, Alexander B Wolf, Martin Grohe, Jakob Burger, Alexander Mitsos, and Dominik G Grimm. 2025. Graphxform: graph transformer for computer-aided molecular design. *Digital Discovery*, 4(4):1052–1065.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, and 1 others. 2020. Molecular sets (moses): a benchmarking platform for molecular generation models. *Frontiers in pharmacology*, 11:565644.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Mayk Caldas Ramos, Christopher J. Collison, and Andrew D. White. 2025. A review of large language models and autonomous agents in chemistry. *Chemical Science*.

Nian Ran, Yue Wang, and Richard Allmendinger. 2025. MOLLM: multi-objective large language model for molecular design - optimizing with experts. *arXiv preprint arXiv:2502.12845*.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. 2022. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264.

Jerret Ross, Samuel Hoffman, Brian Belgodere, Vijil Chenthamarakshan, Youssef Mroueh, and Payel Das. 2024. Learning to optimize molecules with a chemical language model. In *Annual Conference on Neural Information Processing Systems*.

Sakhinana Sagar Srinivas and Venkataramana Runkana. 2024. Crossing new frontiers: Knowledge-augmented large language model prompting for zero-shot text-based de novo molecule design. *arXiv preprint arXiv:2408.11866*.

Kunyang Sun, Dorian Bagni, Joseph M Cavanagh, Yingze Wang, Jacob M Sawyer, Andrew Gritsevskiy, Oufan Zhang, and Teresa Head-Gordon. 2025. Synllama: Generating synthesizable molecules and their analogs with large language models. *arXiv preprint arXiv:2503.12602*.

Richard S Sutton, Andrew G Barto, and 1 others. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Xiangru Tang, Howard Dai, Elizabeth Knight, Fang Wu, Yunyang Li, Tianxiao Li, and Mark Gerstein. 2024. A survey of generative ai for de novo drug design: new frontiers in molecule and protein generation. *Briefings in Bioinformatics*, 25(4):bbae338.

Morgan Thomas, Noel M O'Boyle, Andreas Bender, and Chris De Graaf. 2024. Molscore: a scoring, evaluation and benchmarking framework for generative models in de novo drug design. *Journal of Cheminformatics*, 16(1):64.

Prudencio Tossou, Cas Wognum, Michael Craig, Hadrien Mary, and Emmanuel Noutahi. 2024. Real-world molecular out-of-distribution: Specification and investigation. *Journal of Chemical Information and Modeling*, 64(3):697–711.

Pat Walters. 2024. Silly things large language models do when generating molecules. Practical Cheminformatics Blog. `https://practicalcheminformatics.blogspot.com/2024/10/silly-things-large-language-models-do.html`.

Haorui Wang, Marta Skreta, Cher Tian Ser, Wenhao Gao, Lingkai Kong, Felix Strieth-Kalthoff, Chenru Duan, Yuchen Zhuang, Yue Yu, Yanqiao Zhu, Yuanqi Du, Alan Aspuru-Guzik, Kirill Neklyudov, and Chao Zhang. 2025. Efficient evolutionary search over chemical space with large language models. In *The Thirteenth International Conference on Learning Representations*.

Y. Wang, H. Zhao, S. Sciabola, and W. Wang. 2023. cmolgpt: A conditional generative pre-trained transformer for target-specific de novo molecular generation. *Molecules*, 28(11):4430.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022a. Emergent abilities of large language models. *TMLR*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837.

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.

Egon L Willighagen, John W Mayfield, Jonathan Alvarsson, Arvid Berg, Lars Carlsson, Nina Jeliazkova, Stefan Kuhn, Tomáš Pluskal, Miquel Rojas-Chertó, Ola Spjuth, and 1 others. 2017. The chemistry development kit

(cdk) v2. 0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of cheminformatics*, 9(1):33.

Shirley Wu, Kaidi Cao, Bruno Ribeiro, James Zou, and Jure Leskovec. 2024a. Graphmetro: Mitigating complex graph distribution shifts via mixture of aligned experts. *Advances in Neural Information Processing Systems*, 37:9358–9387.

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. *Chemical science*, 9(2):513–530.

Zhenxing Wu, Odin Zhang, Xiaorui Wang, Li Fu, Huifeng Zhao, Jike Wang, Hongyan Du, Dejun Jiang, Yafeng Deng, Dongsheng Cao, and 1 others. 2024b. Leveraging language model for advanced multiproperty molecular optimization via prompt engineering. *Nature Machine Intelligence*, pages 1–11.

Yingce Xia, Peiran Jin, Shufang Xie, Liang He, Chuan Cao, Renqian Luo, Guoqing Liu, Yue Wang, Zequn Liu, Yuan-Jyue Chen, and 1 others. 2025. Naturelm: Deciphering the language of nature for scientific discovery. *arXiv preprint arXiv:2502.07527*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Nianzu Yang, Huaijin Wu, Kaipeng Zeng, Yang Li, Siyuan Bao, and Junchi Yan. 2024b. Molecule generation for drug design: a graph learning perspective. *Fundamental Research*.

Nianzu Yang, Kaipeng Zeng, Qitian Wu, Xiaosong Jia, and Junchi Yan. 2022. Learning substructure invariance for out-of-distribution molecular representations. *Advances in Neural Information Processing Systems*, 35:12964–12978.

Yang Yao, Xin Wang, Zeyang Zhang, Yijian Qin, Ziwei Zhang, Xu Chu, Yuekui Yang, Wenwu Zhu, and Hong Mei. 2024. Exploring the potential of large language models in graph generation. *arXiv preprint arXiv:2403.14358*.

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

Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. 2018. Graph convolutional policy network for goal-directed molecular graph generation. *Advances in neural information processing systems*, 31.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, and 1 others. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024a. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*.

Botao Yu, Frazier N. Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024b. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*.

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

Xiangxiang Zeng, Fei Wang, Yuan Luo, Seung-gu Kang, Jian Tang, Felice C Lightstone, Evandro F Fang, Wendy Cornell, Ruth Nussinov, and Feixiong Cheng. 2022. Deep generative molecular design reshapes drug discovery. *Cell Reports Medicine*, 3(12).

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, and 1 others. 2024a. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*.

Juzheng Zhang, Yatao Bian, Yongqiang Chen, and Quanming Yao. 2024b. Unimot: Unified molecule-text language model with discrete token representation. *arXiv preprint arXiv:2408.00863*.

Odin Zhang, Haitao Lin, Hui Zhang, Huifeng Zhao, Yufei Huang, Chang-Yu Hsieh, Peichen Pan, and Tingjun Hou. 2024c. Deep lead optimization: Leveraging generative ai for structural modification. *Journal of the American Chemical Society*, 146(46):31357–31370.

Qiang Zhang, Keyan Ding, Tianwen Lv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, and 1 others. 2025. Scientific large language models: A survey on biological & chemical domains. *ACM Computing Surveys*, 57(6):1–38.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024d. A comprehensive survey of scientific large language models and their applications in scientific discovery. *arXiv preprint arXiv:2406.10833*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Alex Zhavoronkov, Yan A Ivanenkov, Alex Aliper, Mark S Veselov, Vladimir A Aladinskiy, Anastasiya V Aladinskaya, Victor A Terentiev, Daniil A Polykovskiy, Maksim D Kuznetsov, Arip Asadulaev, and 1 others. 2019. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature biotechnology*, 37(9):1038–1040.

Yizhen Zheng, Huan Yee Koh, Maddie Yang, Li Li, Lauren T May, Geoffrey I Webb, Shirui Pan, and George Church. 2024. Large language models in drug discovery and development: From disease mechanisms to clinical trials. *arXiv preprint arXiv:2409.04481*.

## A Data Modalities for Molecular LLMs

LLMs used for molecular generation and optimization interface with structured molecular data in various modalities. Each modality offers distinct structural or physicochemical information, with different implications for model performance and capabilities. As shown in Fig. 4, commonly used molecular representations can be categorized into the following three formats:

- **1D Sequence Representations** ($S$): These are linear string encodings of molecular structures. Common formats include:
  - *SMILES* (Simplified Molecular Input Line Entry System) (Weininger, 1988): Most widely used due to direct compatibility with LLM tokenizers, but sensitive to representation choices (canonical vs. randomized)
  - *SELFIES* (Self-Referencing Embedded Strings) (Krenn et al., 2020): Guarantees validity through constrained grammar but at the cost of longer sequences
  - *IUPAC nomenclature* (Favre and Powell, 2014): Systematic chemical names used as auxiliary representations

  **Advantages**: Direct LLM compatibility, compact representation, human-readable
  **Limitations**: Loss of spatial information, multiple valid representations for same molecule, difficulty capturing stereochemistry

- **2D Graph Representations** ($G$): A molecule is represented as a graph $G = (V, E)$, where nodes $v \in V$ correspond to atoms and edges $e \in E$ correspond to chemical bonds. Node and edge features encode atom types, bond orders, aromaticity, and other topological attributes.
  - Integration approaches include: hybrid LLM-GNN architectures (e.g., UniMoT (Zhang et al., 2024b)), graph serialization methods, and cross-attention mechanisms (e.g., MvMRL)
  - Recent work shows significant improvement in molecular discovery when combining graphs with SMILES (Zhang et al., 2024b)

  **Advantages**: Captures topological connectivity, invariant to atom ordering, explicit bond information
  **Limitations**: Requires specialized architectures, computational overhead, potential "graph bypass phenomenon" where LLMs ignore structural information (Lee et al., 2025)

- **3D Geometric Representations** ($X$): These representations capture atomic coordinates in three-dimensional space. Formally, $X = \{(a_i, \vec{r}_i)\}_{i=1}^{N}$, where $a_i$ denotes the atomic species and $\vec{r}_i \in \mathbb{R}^3$ specifies the Cartesian coordinates of atom $i$.
  - Critical for: stereochemistry determination, conformational analysis, binding affinity prediction
  - Integration methods: learned 3D embeddings, auxiliary conformer generation models (e.g., RDKit), geometric deep learning approaches

  **Advantages**: Captures spatial relationships, essential for stereochemistry, enables interaction modeling
  **Limitations**: High computational cost, multiple conformers per molecule, challenging to tokenize for LLMs

## B Datasets

Datasets are crucial resources for advancing LLM-centric molecule design, serving extensively in both the training and evaluation phases of model development. Table 1 provides a comprehensive summary of commonly utilized molecule datasets, detailing their key features. For each dataset listed, the table specifies its **Last Update** year, approximate **Scale** (number of entries), whether it includes natural language **Instruction** components, and its suitability for **Pretraining** LLMs or as a **Benchmark** for evaluation. Furthermore, the table indicates the types of **Molecule Representations** available within each dataset, such as SMILES, IUPAC names, ready-to-dock formats (**Dock**), graph structures (**Graph**), 3D coordinates (**3D**), or formal chemical ontologies (**Ontology**). Finally, it highlights whether a dataset supports **Generation** or **Optimization** tasks, lists **Other Tasks** it is commonly used for (e.g., property prediction, translation), and provides a **Link** to access the resource.

The subsequent subsections categorize these datasets based on their primary application focus, aligning with the classification used in Section 5 of the main text.

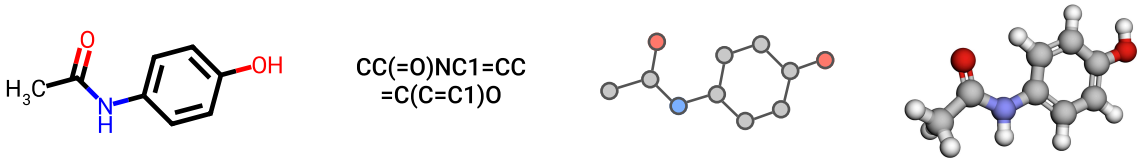


Figure 4: **Illustration of an example molecule and its representation in different data modalities.** From left to right following the 2D chemical structure diagram: its 1D SMILES string representation, a simplified 2D graph view, and its 3D ball-and-stick model.

### B.1 Pretraining-Only Datasets

Pretraining-only datasets typically contain diverse molecular structures and associated property information, designed to support broad generalization capabilities when pretraining LLMs for downstream tasks. These datasets generally do not include explicit natural language instructions or task-specific labels for direct supervised learning of specific generation or optimization objectives.

- **ZINC:** ZINC (Irwin et al., 2012) is a public and comprehensive database containing over 20 million commercially available molecules presented in biologically relevant representations. These molecules can be downloaded in popular ready-to-dock formats and various subsets, making ZINC widely used for distribution learning-based and goal-oriented molecule generation tasks.
- **PubChem:** PubChem (Kim et al., 2016, 2019, 2025) serves as a vast public chemical information repository, holding over 750 million records. It covers a wide array of data, including chemical structures, identifiers, bioactivity outcomes, genes, proteins, and patents, and is organized into three interlinked databases: Substance (contributed chemical information), Compound (standardized unique structures), and BioAssay (biological experiment details).
- **ChemData:** ChemData (Zhang et al., 2024a) is a large-scale dataset specifically curated for fine-tuning chemical LLMs, containing 7 million instruction query-response pairs. Derived from various online structural datasets like PubChem and ChEMBL, it encompasses a broad range of chemical domain knowledge and is frequently used for tasks in molecule understanding, chemical process reasoning, and other domain-specific applications.
- **Mol-Instructions:** Mol-Instructions (Fang et al., 2023) is a large-scale, diverse, and high-quality dataset designed for the biomolecular domain, featuring over 2 million carefully curated biomolecular instructions. It is structured around three core components: molecule-oriented instructions (148.4K across six tasks focusing on properties, reactions, and design), protein-oriented instructions (505K samples across five task categories related to protein structure, function, and design), and biomolecular text instructions (53K for bioinformatics and chemoinformatics NLP tasks like information extraction and question answering).
- **MuMOInstruct:** MuMOInstruct (Dey et al., 2025) is presented as the first high-quality instruction-tuning dataset focused on complex, multi-property molecular optimization tasks. Unlike datasets such as MolOpt-Instruction (Ye et al., 2025) that primarily target single- or dual-property tasks, MuMOInstruct emphasizes tasks involving at least three properties, facilitating the evaluation of LLMs in both in-domain and out-of-domain settings.

### B.2 Benchmark-Only Datasets

Benchmark-only datasets are specifically curated for the evaluation of models, particularly in generative molecular tasks. These datasets often feature structured input-output pairs, such as instruction-molecule pairings, and are typically smaller in scale, manually verified, and tailored to specific evaluative purposes.

- **MoleculeNet:** A large-scale benchmark compendium, MoleculeNet (Wu et al., 2018) is derived from multiple public databases. It comprises 17 curated datasets with over 700,000 compounds, represented textually (e.g., SMILES) and in 3D formats. Covering a wide array of properties categorized into quantum mechanics, physical chemistry, biophysics, and physiology, it serves as a standard for evaluating molecular property prediction models.
- **ChemBench:** ChemBench (Mirza et al., 2024a) offers a comprehensive framework for benchmarking the chemical knowledge and reasoning

Table 1: Summary of commonly used molecule datasets and their features. **Dock** denotes the "ready-to-dock" format; **Ontology** denotes the structured representation of the molecule; **Captioning** denotes molecule captioning task; **Docking** denotes molecule docking (a way to find correct molecule binds for proteins); **Translation** denotes the translation from textual knowledge to molecular features; **Conversion** denotes the translation between different representations of a molecule's identity; **Prediction** denotes property prediction, forward reaction prediction and retrosynthesis tasks; **QM** denotes hybrid quantum mechanics.

| Datasets | Last Update | Scale | Instruc-tion | Pretrain-ing | Bench-mark | Molecule Representations | | | | | | Genera-tion | Optimi-zation | Other Tasks | Link |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | SMILES | IUPAC | Dock | Graph | 3D | Ontology | | | | |
| PubChem (Kim et al., 2016, 2019, 2025) | 2025 | 119M | ✗ | ✓ | ✗ | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | Property Prediction & Biology Domain | Link |
| ChEMBL (Gaulton et al., 2012) | 2024 | >20M | ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✓ | ✓ | Prediction & ML Benchmark | Link |
| CrossDocked2020 (Francoeur et al., 2020) | 2024 | 22.5M | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ | ✓ | Docking Datasets | Link |
| ZINC (Irwin et al., 2012) | 2023 | >980M | ✗ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | Ligand Discovery | Link |
| Dockstring (García-Ortegón et al., 2022) | 2022 | >260k | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | Virtual Screening | Link |
| ChEBI-20 (Edwards et al., 2021) | 2021 | 33k | ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ | ✗ | Translation & Classification & Captioning | Link |
| OGBG-MolHIV (Hu et al., 2020) | 2020 | ∼41k | ✗ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✓ | ✗ | Graph Property Prediction | Link |
| MOSES (Polykovskiy et al., 2020) | 2020 | ∼1.9M | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | De novo Design | Link |
| MoleculeNet (Wu et al., 2018) | 2019 | 700k | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✗ | ✓ | ✓ | ML Benchmark | Link |
| QM9 (Pinheiro et al., 2020) | 2014 | 134k | ✗ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✗ | ✓ | ✓ | Hybrid QM/ML Modeling | Link |
| TOMG-Bench (Li et al., 2024a) | 2025 | 5k | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | Molecule Editing | Link |
| MuMOInstruct (Dey et al., 2025) | 2025 | 873k | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | — | Link |
| ChemData (Zhang et al., 2024a) | 2024 | 7M | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | ✓ | Conversion & Prediction & Reaction | Link |
| ChemBench (Mirza et al., 2024a) | 2024 | 4k | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | Reaction Benchmark & Virtual Screening | Link |
| Mol-Instructions (Fang et al., 2023) | 2024 | 2M | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | Translation, Retrosynthesis | Link |
| MolOpt-Instructions (Ye et al., 2025) | 2024 | 1M | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | — | Link |
| L+M-24 (Edwards et al., 2024) | 2024 | 148k | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✓ | ✗ | Captioning | Link |
| SMolInstruct (Yu et al., 2024b) | 2024 | 3.3M | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | Captioning & Prediction | Link |

abilities of LLMs. It consists of thousands of manually curated question-answer pairs from diverse sources, focusing on three core aspects: Calculation, Reasoning, and Knowledge.

- **TOMG-Bench:** As the first benchmark dedicated to the open-domain molecule generation capabilities of LLMs, TOMG-Bench (Text-based Open Molecule Generation Benchmark) (Li et al., 2024a) contains 45,000 samples. It is structured around three primary tasks: molecule editing (MolEdit), molecule optimization (MolOpt), and customized molecule generation (MolCustom).

- **MOSES:** MOSES (Molecular Sets) (Polykovskiy et al., 2020) is a task-specific resource designed for both training and benchmarking molecule generation models in drug discovery. Containing approximately 1.9 million molecules in SMILES format derived from the ZINC Clean Leads dataset, it also furnishes training, testing, and scaffold-split subsets, along with built-in evaluation metrics.

### B.3 Datasets for Pretraining & Benchmark Applications

A distinct category of datasets offers the flexibility to be used for both pretraining LLMs and for subsequent benchmarking. These resources often combine substantial scale with features amenable to diverse evaluation scenarios.

- **ChEMBL:** ChEMBL (Gaulton et al., 2012) is a manually curated, open-access database focusing on drug-like bioactive molecules. It houses 5.4 million bioactivity measurements for over 1 million compounds and 5,200 protein targets, effectively integrating chemical, bioactivity, and genomic data to support drug discovery and the translation of genomic insights into therapeutics.

- **ChEBI-20:** ChEBI-20 (Edwards et al., 2021), derived from the ChEBI database, is a freely available, manually curated dictionary of molecular entities concentrated on small chemical compounds. It includes over 20,000 molecules represented by SMILES strings, natural language descriptions, and ontology terms, widely em-

ployed in molecule generation and instruction-based tasks requiring chemical understanding.

- **CrossDocked2020:** CrossDocked2020 (Francoeur et al., 2020) is a large-scale dataset specifically geared towards structure-based drug design (SBDD). It features over 22 million 3D docked poses of protein-ligand pairs, making it a valuable resource for tasks like pocket-conditioned 3D molecule generation.
- **Dockstring:** Dockstring (García-Ortegón et al., 2022) provides a large-scale, well-curated dataset for molecular docking. It encompasses an extensive collection of docking scores and poses for more than 260,000 ligands against 58 medically relevant targets, and includes pharmaceutically relevant benchmark tasks such as virtual screening and the *de novo* design of selective kinase inhibitors.
- **QM9:** QM9(The Quantum Mechanics 9) dataset (Pinheiro et al., 2020) is a public quantum chemistry resource containing approximately 134,000 small organic molecules (composed of H, C, N, O, F; up to nine non-hydrogen atoms). It provides SMILES representations, 3D geometries, and quantum chemical properties, widely utilized for training and evaluating molecular property prediction models.
- **SMolInstruct:** SMolInstruct (Yu et al., 2024b) is a large-scale, comprehensive, and high-quality dataset for instruction tuning LLMs in chemistry. It consists of 3.3 million language-molecule pairs and 1.6 million distinct molecules, covering four types of molecular representations and 14 different tasks, with molecules represented in SMILES or SELFIES format.
- **OGBG-MolHIV:** OGBG-MolHIV (Hu et al., 2020), part of the Open Graph Benchmark, is an open-access, task-specific dataset for binary molecular property prediction, specifically for classifying HIV inhibition. It contains 41,127 unique molecules in graph format, where nodes (atoms) have 9 numerical features and edges (bonds) have 3-dimensional features (type, stereochemistry, conjugation). It is derived from MoleculeNet and preprocessed using RDKit.
- **MolOpt-Instructions:** MolOpt-Instructions (Ye et al., 2025) is an instruction-based dataset tailored for molecule optimization, containing over 1 million molecule-molecule pairs. It was constructed by selecting molecules from ZINC and using MMPDB to generate and filter for highly similar pairs, covering six molecular properties including solubility, BBBP, and hERG inhibition.

**Molecule Discovery**

**Molecule Generation**

**Query:** Build a molecule that meets the requirement: The molecule is a phenolate anion obtained by deprotonation of the 7-hydroxy group of noreugenin. It is the major microspecies at pH 7.3 (according to Marvin v 6.2.0.). It has a role as a plant metabolite. It is a conjugate base of a noreugenin.
**Response:** Here is a potential molecule:
CC1=CC(=O)C2=C([O-])C=C(O)C=C2O

**Molecule Optimization**

**Query:** Help me increase the water solubility value of the molecule Cc1cc(C(COCc2ccccc2)sc(C)(C)C)sn1
**Response:** Here is a potential molecule:
Cc1cc(C(N)COCc2ccccc2)sn1

Figure 5: Visualization of the Instruction dataset of molecule generation and optimization task.

- **L+M-24:** L+M-24 (Language + Molecules 24 Tasks) (Edwards et al., 2024) is a large-scale, multi-task instruction dataset designed to leverage the benefits of natural language (compositionality, functionality, abstraction) in molecule design. Derived from PubChem and other sources, it contains over 148,000 language-molecule pairs spanning 24 distinct molecule design tasks across various application domains.

## C Evaluation Metrics

Evaluation metrics for LLM-centric molecular discovery are organized around the four fundamental challenges identified in our survey. Each category of metrics addresses specific aspects of molecular generation and optimization quality, reflecting the unique requirements of chemical tasks compared to general text generation.

### C.1 Validity Metrics

Validity metrics assess whether generated molecules adhere to fundamental chemical rules and structural constraints. Unlike grammatically incorrect text, invalid molecules are physically impossible and unusable.

- **Validity Rate** (Polykovskiy et al., 2020): Fraction of generated molecules that are chemically valid (parsable by RDKit). High validity rates (>90%) indicate successful learning of chemical grammar.

- **Exact Match (EM)** (Rajpurkar et al., 2016): Measures perfect sequence matching between generated and target molecules. Critical for tasks requiring precise molecular replication.

- **BLEU Score** (Papineni et al., 2002): Adapted from NLP, measures n-gram overlap between generated and reference SMILES. Higher scores indicate better sequence-level fidelity.

- **Levenshtein Distance** (Levenshtein, 1966): Minimum edit distance between molecular strings. Lower values indicate closer structural similarity.

- **Chemical Correctness** (Landrum et al., 2013): RDKit-based validation checking valency rules, ring systems, and aromatic systems. Essential for filtering chemically impossible structures.

**Performance Benchmarks:** State-of-the-art LLMs achieve 85-95% validity on standard benchmarks, with multi-modal approaches reaching 95-99%. However, validity alone is insufficient—many valid molecules are practically useless.

### C.2 Synthesizability Metrics

Synthesizability metrics evaluate whether valid molecules can be practically synthesized in a laboratory setting. This addresses the critical gap between theoretical validity and practical utility.

- **SA Score** (Ertl and Schuffenhauer, 2009): Synthetic Accessibility score (1-10 scale, lower is better) based on molecular complexity and fragment contributions. Molecules with SA > 6 are typically considered difficult to synthesize.

- **SCScore** (Coley et al., 2018): Synthetic Complexity score learned from reaction databases. More accurate than SA Score but computationally intensive.

- **Retrosynthetic Accessibility** (Genheden et al., 2020): Evaluates synthesizability through automated retrosynthetic planning. Molecules with viable synthetic routes are considered accessible.

**Current Limitations:** Most LLM-generated molecules have poor synthesizability (average SA Score > 4.5), highlighting a major gap in current approaches. Only specialized models like SynLlama directly address this challenge.

### C.3 Property Control Metrics

Property control metrics assess the model's ability to generate molecules with desired physicochemical or biological properties, often requiring multi-objective optimization.

#### C.3.1 Single-Property Metrics

- **LogP** (Hansch et al., 1968): Octanol-water partition coefficient, indicating hydrophobicity. Target ranges vary by application (e.g., -0.4 to 5.6 for oral drugs).

- **QED** (Bickerton et al., 2012): Quantitative Estimate of Drug-likeness (0-1 scale). Combines multiple properties; scores > 0.67 indicate drug-like molecules.

- **TPSA** (Ertl et al., 2000): Topological Polar Surface Area. Values < 140 Ų correlate with oral bioavailability.

- **Molecular Weight (MW)**, **HBD/HBA**: Basic descriptors for drug-likeness (Lipinski's Rule of Five).

#### C.3.2 Multi-Property Metrics

- **Success Rate** (Jin et al., 2020): Fraction of molecules meeting all specified property constraints. Typical success rates: 60-80% for single properties, 20-40% for multiple properties.

- **Pareto Optimality** (Pareto, 1919): Identifies solutions optimal across multiple objectives. Essential for understanding trade-offs between competing properties.

- **Composite Score** (Jin et al., 2020): Weighted combination of multiple properties. Allows single-objective optimization of multi-property goals.

**Benchmarking Insights:** Instruction-tuned models show 15-25% improvement in property control over base models. Multi-property optimization remains challenging, with success rates dropping exponentially with constraint count.

### C.4 Diversity Metrics

Diversity metrics evaluate the breadth of chemical space explored, preventing mode collapse and encouraging novel discoveries.

- **Uniqueness** (Wang et al., 2023; Bagal et al., 2021): Fraction of non-duplicate valid molecules.

Measured at different scales:

  - Unique@1k: Short-term diversity (typical: 95-99%)
  - Unique@10k: Long-term diversity (typical: 85-95%)

- **Novelty Rate** (Brown et al., 2019): Fraction of generated molecules not in training set. Low novelty (<50%) indicates overfitting.
- **Internal Diversity (IntDiv)** (Benhenda, 2017): Average pairwise dissimilarity within generated set:

$$\text{IntDiv}_p(S) = 1 - \left( \frac{1}{|S|^2} \sum_{s_i, s_j \in S} T(s_i, s_j)^p \right)^{\frac{1}{p}}$$

- **NCircle** (Jang et al., 2024): Largest subset with pairwise Tanimoto similarity below threshold. Higher values indicate better structural diversity.
- **Scaffold Diversity**: Number of unique Bemis-Murcko scaffolds. Critical for avoiding "decoration" of known structures.
- **Fingerprint Tanimoto Similarity (FTS)**: Structural similarity using various fingerprints:
  - MACCS keys (Durant et al., 2002): 166-bit structural keys
  - Morgan fingerprints (Morgan, 1965): Circular fingerprints
  - RDKit fingerprints (Landrum et al., 2013): Topological fingerprints

**Key Findings:** Supervised fine-tuning often reduces diversity (IntDiv drops 20-30%). Preference tuning methods like Div-SFT successfully restore diversity while maintaining other properties.

### C.5 Integrated Evaluation Framework

No single metric captures all aspects of molecular quality. We recommend:

1. **Minimum requirements**: Validity > 90%, Uniqueness@1k > 95%
2. **Task-specific priorities**: Weight metrics based on application (e.g., prioritize synthesizability for lead optimization)
3. **Multi-metric reporting**: Always report all four categories to reveal trade-offs
4. **Baseline comparisons**: Compare against both random generation and domain-specific baselines

## D External Tools

The evaluation of molecular generation and optimization models relies on a comprehensive ecosystem of computational tools that bridge chemistry, machine learning, and specialized assessment frameworks. These tools can be categorized into three main groups based on their primary functions.

### D.1 General Cheminformatics Libraries

**RDKit** (Landrum et al., 2013) has become the de facto standard in the field, providing extensive functionality for molecular representation, property calculation, and structure validation. It handles SMILES parsing, canonicalization, and validation; calculates physicochemical properties including logP, molecular weight, TPSA, and hydrogen bond donors/acceptors; generates various molecular fingerprints (Morgan/ECFP, MACCS, RDK topological); performs substructure searching and Bemis-Murcko scaffold extraction; and validates chemical structures including aromatic system detection. Nearly all major benchmarks including MOSES and GuacaMol rely heavily on RDKit for their metric calculations.

**OpenBabel** (O'Boyle et al., 2011) serves as the "universal translator" of chemical file formats, supporting over 110 formats and providing critical interoperability between different computational chemistry software. While it also offers descriptor calculation and structure manipulation, its primary strength lies in format conversion, accessible through the PyBel Python interface. This capability is essential when integrating diverse chemical data sources or connecting different software tools in evaluation pipelines.

**CDK (Chemistry Development Kit)** (Willighagen et al., 2017) provides a comprehensive Java-based cheminformatics library with mature graph algorithms for structural analysis and 3D molecular modeling. Its Java foundation makes it particularly suitable for integration into enterprise-level applications, offering robust APIs for custom chemical informatics solutions.

### D.2 Synthesizability Assessment Tools

Given that computational validity does not guarantee practical synthesizability, specialized tools

have emerged to bridge this critical gap.

**AiZynthFinder** (Genheden et al., 2020) employs neural network-guided Monte Carlo tree search for retrosynthetic planning. It evaluates synthesizability by attempting to find viable synthetic routes from commercially available starting materials, providing both binary feasibility assessments and synthetic accessibility scores. The tool has become increasingly important as the field recognizes that many computationally valid molecules remain synthetically inaccessible.

**ASKCOS** (Coley et al., 2019) (Automated System for Knowledge-based Continuous Organic Synthesis) offers a comprehensive platform that integrates multiple machine learning models for forward reaction prediction, retrosynthetic route planning, condition recommendation, and synthetic complexity evaluation. This unified approach provides more reliable synthesizability assessments by considering multiple aspects of the synthetic process simultaneously.

### D.3 LLM-Specific Integration Tools

The emergence of LLMs has necessitated new tools that bridge natural language processing with chemical computation.

**ChemCrow** (M. Bran et al., 2024) represents a paradigm shift by augmenting LLMs with 17 expert-designed chemistry tools. It enables LLMs to execute chemical calculations they cannot perform natively, access real-time chemical databases, perform safety checks on generated molecules, and plan and evaluate synthetic routes. This tool-augmented approach addresses the fundamental limitation that LLMs, while excellent at pattern recognition, lack the ability to perform precise chemical calculations or access up-to-date chemical information.

**ChemBench Package** (Mirza et al., 2024b) provides a modular, extensible framework specifically designed for benchmarking LLM performance on chemical tasks. It offers standardized evaluation pipelines through automated model querying, answer parsing, and report generation, significantly simplifying the process of evaluating LLMs on chemical reasoning and generation tasks.

## E Evaluation Frameworks

The evolution of evaluation frameworks in molecular generation reflects the field's progression from statistical distribution matching to instruction-following and multi-objective optimization. Each framework addresses specific limitations of its predecessors while introducing new evaluation paradigms.

### E.1 Classical Generation Frameworks

**MOSES (Molecular Sets)** (Polykovskiy et al., 2020) established the foundation for standardized evaluation by providing a carefully filtered dataset of 1.9M drug-like molecules from ZINC, a comprehensive metric suite including validity, uniqueness, novelty, FCD, and fragment/scaffold similarity, baseline implementations of multiple architectures (CharRNN, VAE, AAE, ORGAN, JT-VAE), and standardized train/test splits to ensure fair comparison. MOSES primarily focuses on distribution learning—the ability of models to replicate the statistical properties of the training set. Its key contribution was creating a unified, reproducible testing ground for comparing different generative architectures.

**GuacaMol** (Brown et al., 2019) significantly expanded the evaluation scope by introducing both distribution learning tasks using KL divergence and Fréchet ChemNet Distance, and goal-directed benchmarks comprising 20 tasks ranging from simple property maximization to complex multi-parameter optimization (MPO). These tasks were specifically designed to mirror real drug discovery scenarios, such as generating molecules similar to celecoxib but with improved properties. This dual approach better reflects the practical needs of molecular design, where both exploration (distribution learning) and exploitation (goal-directed optimization) are crucial.

### E.2 Modern Unified Frameworks

**MolScore** (Thomas et al., 2024) addresses the fragmentation issue in molecular optimization evaluation through its modular architecture supporting over 40 scoring functions, unified interface for diverse molecular optimization algorithms, flexible aggregation methods for multi-objective optimization, and extensive configuration options via JSON/YAML. Its key innovation lies in decoupling scoring from optimization, allowing researchers to mix and match components freely while maintain-

ing consistent evaluation protocols.

**TDC (Therapeutics Data Commons)** (Huang et al., 2021) takes a community-driven approach by providing 66+ datasets across 22+ therapeutic tasks, continuously updated leaderboards with standardized evaluation protocols, and realistic data splits (scaffold-based, temporal, and combination splits) that better reflect real-world deployment scenarios. The framework's APIs enable easy integration and benchmarking, making it particularly valuable for researchers seeking to evaluate their methods against established baselines on therapeutically relevant tasks.

### E.3 LLM-Specific Evaluation Frameworks

The emergence of LLMs necessitated entirely new evaluation paradigms that assess instruction-following and reasoning capabilities rather than just statistical properties.

**TOMG-Bench** (Li et al., 2024a) pioneered open-domain molecule generation evaluation with three task categories: MolEdit for component manipulation (adding, removing, or replacing functional groups), MolOpt for property optimization (LogP, QED, molecular refractivity), and MolCustom for constrained generation based on specific requirements. The framework provides 45,000 test samples with diverse instructions and employs weighted accuracy metrics that combine task success with chemical similarity or novelty scores. Its automated evaluation system directly assesses whether generated molecules adhere to the given instructions while maintaining chemical validity.

**ChemBench** (Mirza et al., 2024b) focuses on evaluating chemical reasoning capabilities through a question-answering format covering calculation tasks, chemical reasoning, and factual knowledge. The framework enables direct comparison with human expert performance, includes safety evaluation components to assess potentially harmful outputs, and supports multi-modal queries involving both text and molecular structures. This comprehensive approach reveals that while LLMs can match or exceed human experts on certain knowledge tasks, they still struggle with deep chemical reasoning requiring multi-step inference.

**AMORE (Augmented Molecular Retrieval)** (Ganeeva et al., 2024) further probes the robustness of chemical language models by assessing if they truly understand the underlying molecular structure rather than memorizing textual patterns. This zero-shot framework evaluates a model's chemical awareness through a retrieval task based on molecular augmentations that preserve chemical identity, such as canonicalization, explicit hydrogen addition, kekulization, and cycle renumbering. The model is tasked with matching the embedding of an original SMILES string to the embedding of its chemically equivalent but textually different augmentation. Key findings reveal that many LLMs are not robust to these variations, showing significant performance degradation on both the retrieval task and downstream property prediction tasks when presented with augmented inputs. This indicates that models often overfit to specific string representations, highlighting a critical gap in their chemical understanding.

### E.4 Recommendations for Framework Selection

For researchers navigating this landscape, framework selection should align with specific evaluation needs. MOSES provides the most standardized comparison for distribution learning tasks. GuacaMol or MolScore offer comprehensive evaluation for goal-directed optimization, with MolScore providing greater flexibility for custom objectives. TDC excels when therapeutic relevance is paramount, offering realistic data splits that better predict real-world performance. For LLM evaluation, TOMG-Bench effectively assesses generation capabilities while ChemBench evaluates reasoning and knowledge. Comprehensive evaluation often requires combining multiple frameworks to capture different aspects of model performance.

## F Qualitative and Quantitative Analysis

To synthesize the discussions from previous sections, we present a qualitative and quantitative analysis of the different learning paradigms. We first offer a qualitative summary of how each paradigm addresses the core challenges, visualized in a comparative radar chart. Subsequently, to ground these observations in empirical data, we leverage a recent benchmark (Li et al., 2024a) as a quantitative case study, focusing on its insights into property control and model fine-tuning.

### F.1 Qualitative Comparison of Learning Paradigms

Based on our survey of existing literature, the strengths and weaknesses of the primary learn-

Table 2: Performance of various LLMs on the Molecule Optimization (MolOpt) task.

| Models | LogP | | | MR | | | QED | | |
|---|---|---|---|---|---|---|---|---|---|
| | SR | Similarity | Validity | SR | Similarity | Validity | SR | Similarity | Validity |
| GPT-4o (Achiam et al., 2023) | 0.7190 | 0.6586 | 0.8796 | 0.6864 | 0.6420 | 0.8352 | 0.3952 | 0.6180 | 0.8570 |
| GPT-4-turbo (Achiam et al., 2023) | 0.7662 | 0.6984 | 0.9048 | 0.7388 | 0.6821 | 0.8848 | 0.3946 | 0.6587 | 0.9050 |
| GPT-3.5-turbo (Achiam et al., 2023) | 0.4048 | 0.6327 | 0.8540 | 0.4120 | 0.6263 | 0.8486 | 0.3316 | 0.5635 | 0.8354 |
| Claude-3.5 (Anthropic, 2024b) | 0.7970 | 0.7124 | 0.9422 | 0.6962 | 0.7112 | 0.9110 | 0.5361 | 0.7042 | 0.8604 |
| Claude-3 (Anthropic, 2024a) | 0.7984 | 0.6067 | 0.9096 | 0.6094 | 0.6398 | 0.9062 | 0.4678 | 0.5855 | 0.9044 |
| Gemini-1.5-pro (Deepmind, 2024) | 0.7712 | 0.7022 | 0.9274 | **0.7876** | 0.6744 | 0.8926 | 0.4704 | 0.6077 | 0.9484 |
| Llama3-70B-Instruct (Int4) (Dubey et al., 2024) | 0.5984 | 0.6028 | 0.6482 | 0.5684 | 0.6032 | 0.6272 | 0.2774 | 0.4828 | 0.6340 |
| Llama3-8B-Instruct (Dubey et al., 2024) | 0.4642 | 0.3658 | 0.6086 | 0.4332 | 0.4793 | 0.5704 | 0.2568 | 0.4547 | 0.6112 |
| Llama3.1-8B-Instruct (Dubey et al., 2024) | 0.3990 | 0.4235 | 0.5122 | 0.4336 | 0.5257 | 0.5910 | 0.2655 | 0.4499 | 0.6158 |
| Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) | 0.2220 | 0.4501 | 0.2802 | 0.1908 | 0.2578 | 0.3795 | 0.1210 | 0.3244 | 0.2532 |
| Qwen2-7B-Instruct (Yang et al., 2024a) | 0.0000 | 0.2923 | 0.0004 | 0.0002 | 0.4123 | 0.0004 | 0.0000 | 0.0000 | 0.0000 |
| Yi-1.5-9B (Young et al., 2024) | 0.2884 | 0.5461 | 0.4927 | 0.2050 | 0.3724 | 0.4126 | 0.1064 | 0.6596 | 0.4526 |
| Chatglm-9B (GLM et al., 2024) | 0.3666 | 0.6902 | 0.4736 | 0.3514 | 0.6820 | 0.5000 | 0.1832 | 0.6506 | 0.4342 |
| Llama-3.2-1B-Instruct (Dubey et al., 2024) | 0.0644 | 0.5055 | 0.1664 | 0.0822 | 0.4410 | 0.1604 | 0.0714 | 0.4757 | 0.1796 |
| MolT5-small (Edwards et al., 2022) | 0.2158 | 0.1052 | 0.4302 | 0.2316 | 0.1011 | 0.4420 | 0.2214 | 0.1031 | 0.4326 |
| MolT5-base (Edwards et al., 2022) | 0.2074 | 0.1051 | 0.4168 | 0.1856 | 0.1073 | 0.3796 | 0.2358 | 0.1054 | 0.4536 |
| MolT5-large (Edwards et al., 2022) | 0.4244 | 0.1015 | 0.8156 | 0.4496 | 0.1072 | 0.8678 | 0.4654 | 0.1190 | 0.9214 |
| BioT5-base (Pei et al., 2024) | 0.5158 | 0.1526 | 1.0000 | 0.5060 | 0.1597 | 1.0000 | 0.5068 | 0.1580 | 1.0000 |
| Llama-3.2-1B (OpenMolIns-large) | 0.2898 | 0.5951 | 0.3850 | 0.2644 | 0.5956 | 0.3678 | 0.1996 | 0.5849 | 0.3490 |
| Llama-3.1-8B (OpenMolIns-large) | **0.8054** | 0.6678 | 0.8720 | 0.7122 | 0.6548 | 0.8514 | 0.5224 | 0.6398 | 0.8802 |
| Galactica-125M (OpenMolIns-light) | 0.3202 | 0.6547 | 0.6416 | 0.3508 | 0.6435 | 0.6358 | 0.2690 | 0.6521 | 0.6380 |
| Galactica-125M (OpenMolIns-small) | 0.4172 | 0.6420 | 0.5568 | 0.3958 | 0.6452 | 0.5338 | 0.2956 | 0.6385 | 0.5376 |
| Galactica-125M (OpenMolIns-medium) | 0.5904 | 0.5812 | 0.7890 | 0.5874 | 0.5873 | 0.7384 | 0.4608 | 0.5859 | 0.7768 |
| Galactica-125M (OpenMolIns-large) | 0.6454 | 0.5927 | 0.8198 | 0.6388 | 0.5973 | 0.8028 | 0.4950 | 0.5962 | 0.8100 |
| Galactica-125M (OpenMolIns-xlarge) | 0.7362 | 0.5744 | 0.8902 | 0.7124 | 0.5697 | 0.8612 | **0.5786** | 0.5677 | 0.8626 |

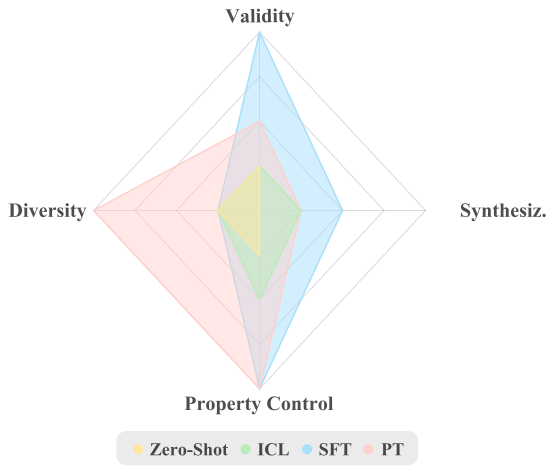


Figure 6: A qualitative comparison of learning paradigms. Abbreviations: Supervised Fine-Tuning (SFT), Preference Tuning (PT), and In-Context Learning (ICL). Each paradigm demonstrates distinct trade-offs in addressing the challenges of molecular discovery.

ing paradigms can be qualitatively summarized as shown in Figure 6.

The radar chart illustrates the distinct trade-offs:

- **Supervised Fine-Tuning (SFT)** is highly effective for instilling foundational chemical knowledge, leading to high **Validity** and precise **Property Control**, but often at the cost of reduced **Diversity**.
- **Preference Tuning (PT)** directly addresses the limitations of SFT by rewarding novelty, making it the strongest paradigm for enhancing **Diversity**. It also maintains excellent **Property Control** through feedback-driven learning.
- Both **In-Context Learning (ICL)** and **Zero-Shot** prompting offer tuning-free application but provide limited guarantees. ICL's performance is highly dependent on example quality, while Zero-Shot methods struggle with chemical nuances.
- Critically, **Synthesizability** remains the most significant unresolved challenge across all paradigms, indicating a crucial area for future research.

### F.2 Quantitative Insights

To validate these qualitative observations with concrete data, we analyze results from the Molecule Optimization (MolOpt) task of a recent benchmark TOMG-Bench (Li et al., 2024a). This task evaluates an LLM's ability to modify molecules according to textual instructions, using metrics like Success Rate (SR) for property control and Validity for chemical correctness. The benchmark also provides a domain-specific instruction dataset, OpenMolIns, enabling a direct comparison between general-purpose and specialized models.

### F.2.1 Key Findings

The detailed results for the MolOpt task, presented in Table 3, reveal several critical findings that quantitatively support our qualitative analysis.

1. **Baseline Performance of General-Purpose LLMs:** Without task-specific fine-tuning, leading proprietary models like GPT-4o and Claude-3.5 already demonstrate strong baseline capabilities. For instance, Claude-3.5 achieves a high Success Rate (SR) of 0.7970 on LogP optimization, showcasing the powerful out-of-the-box reasoning and instruction-following abilities of state-of-the-art LLMs for property control. However, open-source generalist models tend to lag behind, indicating that while powerful, zero-shot performance is not guaranteed.

2. **Significant Performance Gains from SFT:** The data provides striking evidence for the effectiveness of Supervised Fine-Tuning (SFT) with domain-specific data. The general-purpose `Llama-3.1-8B-Instruct` model achieves a modest SR of 0.3990 on LogP optimization. However, after being fine-tuned on the domain-specific `OpenMolIns` dataset, the same model's SR more than doubles to **0.8054**, outperforming many larger, proprietary models. This quantitatively demonstrates that SFT is a crucial step for elevating a generalist model to an expert-level performer, enabling high-fidelity property control.

In summary, the quantitative results strongly corroborate our qualitative assessment. They confirm that while powerful generalist models provide a strong baseline via zero-shot or few-shot prompting, achieving state-of-the-art performance in molecular tasks requires dedicated, domain-specific tuning (SFT).

## G Distribution Shift and Out-of-Distribution Generalization

A critical challenge in molecular discovery is **distribution shift**, where models trained on known molecules fail to generalize to novel, out-of-distribution (OOD) compounds necessary for true innovation. Recent benchmarks reveal that molecular ML models exhibit OOD errors **3× larger** than in-distribution performance (Antoniuk et al., 2025), with performance degradations of 20-60% in real-world scenarios (Tossou et al., 2024). This problem is a primary cause of "mode collapse" and directly limits the **Diversity** discussed in the main text (Tossou et al., 2024). Effectively navigating this shift is essential for moving beyond rediscovery to genuine design.

To address this, machine learning models have developed distinct strategies. Traditional methods like GNNs and VAEs often focus on learning invariant representations, for instance, through Mixture-of-Experts (MoE) architectures that handle specific data domains (Wu et al., 2024a) or by disentangling molecules into "causal" and "spurious" substructures to improve robustness (Yang et al., 2022). However, these approaches often require substantial data to avoid spurious correlations.

LLMs leverage different learning paradigms with unique advantages. While standard Supervised Fine-Tuning (SFT) can overfit to the training distribution, **Preference Tuning (PT)** directly encourages OOD exploration by explicitly rewarding novelty and diversity, as exemplified by models like `Div-SFT` (Jang et al., 2024). Furthermore, advanced **Instruction Tuning** on complex, multi-property tasks (using datasets like `MuMOInstruct`) enables the model to learn more generalizable chemical reasoning for unseen tasks.

**Test-time adaptation** represents a particularly promising direction, with methods like TAIP achieving 30% error reduction through self-supervised learning during inference (Kreiman and Krishnapriyan, 2025). Finally, **Agentic frameworks** like `MultiMol` (Liu et al., 2022) contribute by incorporating external, out-of-distribution knowledge from scientific literature to guide the generation process. Together, these LLM-centric techniques represent a key frontier in developing models that can truly innovate, though significant challenges remain in ensuring synthesizability and practical utility of OOD-generated molecules.

## H Method Summary

This section provides a consolidated overview of representative LLM-based methods for molecular discovery, as detailed in Table 3. The table organizes these approaches primarily by the two core task categories central to this survey: molecule generation and molecule optimization. Within each task, methods are further sub-categorized by their primary learning Strategy (referred to as "Category" and "Technique" in the table), encompassing

approaches without LLM tuning (such as zero-shot prompting and in-context learning) and those with LLM tuning (supervised fine-tuning and preference tuning).

Table 3 details several key aspects for each listed **Method**:

- **Venue**: The publication venue or preprint archive where the method was reported.
- **Input Type**: Specifies the primary format of molecular data and instructions provided to the LLM (e.g., SMILES strings, textual instructions, few-shot examples, or multi-modal inputs like graphs).
- **Base Model**: Indicates the foundational LLM architecture (e.g., GPT-4, LLaMA variants, Mistral) upon which the method is built or applied.
- **Dataset**: Lists the key molecular corpora or benchmarks used for training the model (if applicable) or for its evaluation in the context of the reported work.
- **Repository**: Provides a link to the public code or resource repository, if available.

This structured presentation aims to offer a clear comparative landscape of the current methodologies in the field.

Table 3: Summary of LLM-based methods for molecule generation and optimization. Each row corresponds to a method, organized by **Task** (generation or optimization), and **Technique**. **Input Type** denotes the molecular data format provided to the model. **Base Model** denotes the large language model architecture used. **Dataset** denotes the molecular corpus or benchmark used for training or evaluation.

| Task | Category | Technique | Method | Venue | Input Type | Base Model | Dataset | Repository |
|---|---|---|---|---|---|---|---|---|
| Molecule Generation | w/o Tuning | ICL | LLM4GraphGen (Yao et al., 2024) | Arxiv | Instruction + Few shot | GPT-4 | OGBG-MolHIV | Link |
| | | | MolReGPT (Li et al., 2024c) | TKDE | Instruction + Few shot | GPT-3.5-turbo/ GPT-4 | ChEBI-20 | Link |
| | | | FrontierX (Srinivas and Runkana, 2024) | Arxiv | Instruction | GPT-3.5 | ChEBI-20 | N/A |
| | w/ Tuning | SFT | Mol-instructions (Fang et al., 2023) | ICLR | Instruction | LLaMA-7B | Mol-Instructions | Link |
| | | | LlaSMol (Yu et al., 2024a) | COLM | Instruction | Galactica 6.7B/ LLaMA-2-7B/ Mistral-7B | SMolInstruct | Link |
| | | | ChemLLM (Zhang et al., 2024a) | Arxiv | Instruction | InternLM2-7B-Base | ChemData | N/A |
| | | | ICMA (Li et al., 2024b) | TKDE | Instruction + Few shot | Mistral-7B | PubChem & ChEBI-20 | N/A |
| | | | MolReFlect (Li et al., 2024d) | Arxiv | Instruction + Few shot | Mistral-7B | ChEBI-20 | Link |
| | | | ChatMol (Fan et al., 2025) | Arxiv | Instruction | LLaMA-3-8B | ZINC | Link |
| | | | PEIT-LLM (Lin et al., 2025) | Arxiv | Instruction | LLaMA-3.1-8B/ Qwen2.5-7B | ChEBI-20 | Link |
| | | | NatureLM (Xia et al., 2025) | Arxiv | SMILES + Instruction | NatureLM-8B | ChEMBL & MoleculeNet | Link |
| | | | SynLlama (Sun et al., 2025) | Arxiv | Instruction | LLaMA-3.1-8B / LLaMA-3.2-1B | ChEMBL | Link |
| | | | TOMG-Bench (Li et al., 2024a) | Arxiv | Instruction | LLaMa-3.1-8B | TOMG-Bench | N/A |
| | | | UniMoT (Zhang et al., 2024b) | Arxiv | Instruction | LLaMA-2-7B | Mol-Instructions | Link |
| | | Preference Tuning | Div-SFT (Jang et al., 2024) | Arxiv | Instruction | LLaMA-7B | ChEBI-20 | N/A |
| | | | Mol-MOE (Calanzone et al., 2025) | Arxiv | Instruction | LLaMA-3.2-1B | ChEMBL & ZINC & MOSES | Link |
| | | | SmileyLLama (Cavanagh et al., 2024) | NeurIPS Workshop | Instruction | LLaMA-3.1-8B | ChEMBL | N/A |
| | | | ALMol (Gkoumas, 2024) | ACL Workshop | Instruction | Meditron-7B | L+M-24 | N/A |
| | | | Less for More (Gkoumas and Liakata, 2024) | Arxiv | Instruction | Meditron-7B | L+M-24 | N/A |
| | | | Mol-LLM (Lee et al., 2025) | Arxiv | Instruction | Mistral-7B | ChEBI-20 | N/A |
| Molecule Optimization | w/o Tuning | Zero-Shot Prompting | LLM-MDE (Bhattacharya et al., 2024) | JCIM | SMILES + Instruction | Claude 3 Opus | ZINC | N/A |
| | | | MOLLEO (Wang et al., 2025) | ICLR | SMILES + Instruction | GPT-4 | ZINC | Link |
| | | ICL | CIDD (Gao et al., 2025) | Arxiv | SMILES + Interaction report | GPT-4o | CrossDocked2020 | N/A |
| | | | LLM-EO (Lu et al., 2024) | Arxiv | SMILES + Ligands Pool | Claude 3.5 Sonnet / OpenAI o1-preview | TMC dataset | Link |
| | | | MOLLM (Ran et al., 2025) | Arxiv | SMILES + Instruction | GPT-4o | ZINC | N/A |
| | | | ChatDrug (Liu et al., 2024c) | ICLR | SMILES + Instruction | Galactica / LLaMA-2 / ChatGPT | ZINC | Link |
| | | | $Re^2DF$ (Le and Chawla, 2024) | Arxiv | SMILES + Instruction | LLaMA-3.1-8B/ LLaMA-3.1-70B | ZINC | Link |
| | | | BOPRO (Agarwal et al., 2025) | ICLR | SMILES + Instruction | Mistral-Large-Instruct-2407 | Dockstring | Link |
| | w/ Tuning | SFT | MultiMol (Yu et al., 2025) | Arxiv | SMILES + Instruction | Qwen2.5-7B / LLaMA-3.1-8B / Galactica 6.7B | PubChem | Link |
| | | | DrugAssist (Ye et al., 2025) | Brief Bioinform | SMILES + Instruction | LLaMA-2-7B-Chat | MolOpt-Instructions | Link |
| | | | $GeLLM^3O$ (Dey et al., 2025) | Arxiv | SMILES + Instruction | Mistral-7B-Instruct / LLaMA-3.1-8B-Instruct | MuMOInstruct | Link |
| | | | DrugLLM (Liu et al., 2024d) | Arxiv | Group-based Molecular Representation | LLaMA-2-7B | ZINC & ChEMBL | N/A |
| | | | TOMG-Bench (Li et al., 2024a) | Arxiv | Instruction | LLaMa-3.1-8B | TOMG-Bench | N/A |
| | | | LLM-Enhanced GA (Bedrosian et al., 2024) | NeurIPS Workshop | JSON Objects | Chemma / Chemlactica | PubChem | Link |
| | | | Molx-Enhanced LLM (Le et al., 2024) | Arxiv | SMILES + Graph + Instruction | LLaMA-2-7B | PubChem | N/A |
| | | Preference Tuning | NatureLM (Xia et al., 2025) | Arxiv | SMILES + Instruction | NatureLM-8B | ChEMBL & MoleculeNet | N/A |