SUPRABENCH: A BENCHMARK FOR SUPRAMOLECULAR CHEMISTRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Supramolecular chemistry, which includes the study of non-covalent host–guest assemblies, has shown advances in various applications. However, designing host–guest systems remains a time-consuming process, requiring days of dry-lab verification per candidate pair. Although LLMs have emerged as a fast alternative with strong performance on molecular binding tasks, no benchmark currently systematically evaluates LLMs for host–guest reasoning across fundamental supramolecular chemistry tasks, *e.g.*, binding affinity prediction. To this end, we collaborate with domain experts to release the first **Supra**molecular **Bench**mark, called **SUPRABENCH**, to evaluate LLMs in chemistry reasoning. Specifically, we design four fundamental tasks, *i.e.*, binding affinity prediction, top-binder selection, solvent identification, and host–guest description, plus an auxiliary vision-based task for molecular identification. We also release **SUPRAPMC**, a curated 16M-token corpus of **Supra**molecular chemistry articles distilled from Europe **PMC**, to support the adaptation to the supramolecular domain. We benchmark a broad range of open and proprietary LLMs and find that LLMs leave substantial headroom across all tasks. DAPT over SUPRAPMC transfers cleanly to in-distribution regression but trades off against strict letter-format output. Moreover, the difficulty profile differs sharply across task families, revealing distinct failure modes that indicate specific gaps in current supramolecular chemical reasoning.

1 INTRODUCTION

Supramolecular chemistry includes the study of reversible non-covalent host–guest assemblies and represents a third bond-design paradigm alongside ionic and covalent chemistry, with pioneering work recognized by the 1987 Nobel Prize in Chemistry (Lehn, 1995). Its applications are already widespread, including drug delivery (Loftsson & Brewster, 2010; Webber & Langer, 2017), chemical sensing (You et al., 2015; Kolesnichenko & Anslyn, 2017), and in vivo detoxification of pharmaceutical and anesthetic agents (Brockett et al., 2023; Deng et al., 2020), as illustrated in Figure 1(a). For example, sugammadex (Bom et al., 2002), a γ-cyclodextrin host that selectively encapsulates the muscle relaxant rocuronium, has been approved in 75 countries and has become a standard reversal agent in operating rooms worldwide, reducing postoperative pulmonary complications relative to prior protocols (Olesnicky et al., 2024).

Although these applications reach patients worldwide to reduce morbidity at scale, each success takes decades of expert iteration to deliver, and designing such a host–guest system remains time-consuming. Specifically, when using computational tools to design candidate host scaffolds and guests that match a target binding profile (Thordarson, 2011), each short-listed pair may be verified computationally through classical simulation approaches, *e.g.*, density functional theory (DFT) or molecular dynamics (MD), to estimate the binding free energy before committing to wet-lab synthesis (Yin et al., 2017; Mobley & Gilson, 2017). Each computational verification may require days of computation, even on modern High-Performance Computing clusters (Yin et al., 2017). Over the entire design pipeline, this read-then-simulate loop dominates the calendar long before any molecule reaches the bench.

With recent advances in large language models (LLMs) Ye et al. (2025); Chen et al. (2025a;b), existing studies have attempted to leverage these as a fast alternative to the time-consuming DFT or MD pipelines at the hypothesis-generation stage in molecular binding prediction tasks.

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

Across binding-related tasks, structure–prediction foundation models such as Boltz (Passaro et al., 2025) and follow-ups (Feng et al., 2025) now approach FEP-level affinity, LLM-based methods predict drug-target interactions (Li et al., 2025), and zero-shot kinase-inhibitor binding at accuracies that exceed classical docking (Liu et al., 2024). In particular, Parrilla-Gutiérrez et al. (2024) proposes an electron-density-conditioned GPT that generates candidate host–guest binders for CB[n] and a family of metal–organic cages, and experimentally validates several previously unreported guests. Despite these advances, no benchmark currently evaluates whether modern LLMs can support host–guest reasoning tasks.

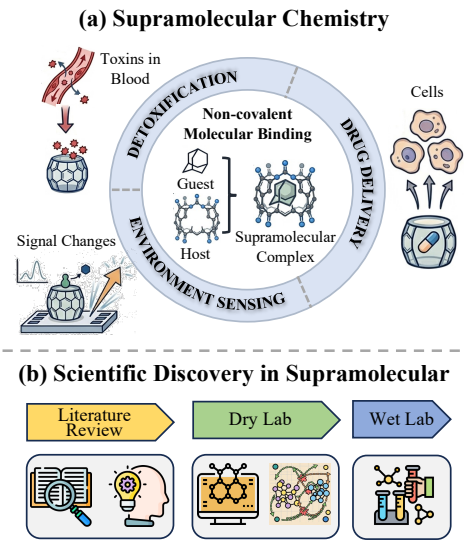

Figure 1: **Overview of supramolecular chemistry.** (a) Studies and applications in supramolecular chemistry. (b) Classic scientific discovery pipeline.

To this end, we introduce the first **Bench**mark for **Supra**molecular chemistry, called **SUPRABENCH**, to evaluate LLMs in host–guest reasoning tasks. Specifically, we collaborate with experts in supramolecular chemistry to carefully define four fundamental tasks, *i.e.*, binding affinity prediction, top-binder selection, solvent identification, and host–guest description, along with one auxiliary vision-based task for molecular identification. To fully isolate the LLMs' ability from the knowledge used to train LLMs, we introduce an expert-annotated set for the binding affinity prediction task that is not included in existing literature. Moreover, we release **SUPRAPMC**, a curated 16M-token corpus of **Supra**molecular chemistry articles obtained from Europe **PMC** (Consortium, 2015), to support future research in domain adaptation. Across the five tasks, we benchmark a broad range of open and proprietary LLMs and find that general-purpose models leave substantial headroom on every task, that domain adaptation closes a measurable but uneven portion of this gap, and that the difficulty profile differs sharply across task families, exposing distinct failure modes. We believe our benchmark SUPRABENCH, along with our released supramolecular chemistry text corpus SUPRAPMC, can contribute to the relevant research communities. Our main contributions are summarized as follows:

- **Supramolecular Benchmark.** We introduce SUPRABENCH, the first supramolecular benchmark comprising four fundamental tasks and an auxiliary vision-based task, to evaluate LLMs under a unified protocol and a single set of metrics.
- **Supramolecular Text Corpus.** We release SUPRAPMC, a 16M-token corpus of supramolecular chemistry articles, that can contribute to the research communities in this field.
- **Systematic Evaluation and Insights.** We evaluate a representative set of open and proprietary-based LLMs, including a domain-adapted variant trained over SUPRAPMC. Among the evaluation results, we reveal actionable insights into the strengths and limitations of existing LLMs for supramolecular chemistry tasks.

2 BACKGROUND

We discuss the key relevant works in this section, and leave detailed related works in Appendix A.

LLMs for Chemistry. LLMs have advanced chemistry along two complementary directions. The first line of work (Zhang et al., 2024; Zhao et al., 2025; Yu et al., 2024; Fang et al., 2024) develops domain-adapted LLMs that trains on chemistry-specific corpora and instruction data. Moreover, several studies introduce chemistry foundation models, such as Galactica (Taylor et al., 2022), MolT5 (Edwards et al., 2022), and nach0 (Livne et al., 2024). The second direction introduces LLM-based agents, equipped with external tools for laboratory automation, *e.g.*, ChemCrow (M. Bran et al., 2024), Coscientist (Boiko et al., 2023), and ChemAgent (Tang et al., 2025). Across both directions, evaluation has primarily focused on small molecule tasks such as property prediction,

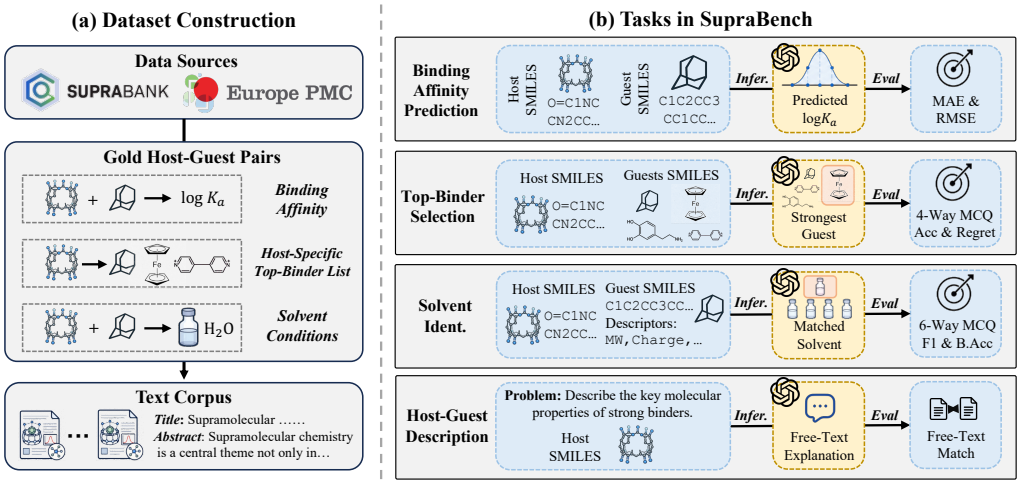

Figure 2: **Overview of SUPRABENCH.** (a) We obtain gold host–guest interaction records from SupraBank, and text corpus of supramolecular chemistry articles from Europe PMC. (b) SUPRABENCH contains four fundamental tasks in supramolecular chemistry reasoning, including binding affinity prediction, top-binder selection, solvent identification, and host–guest description.

retrosynthesis, and reaction yield estimation, leaving the supramolecular host–guest setting that SUPRABENCH targets largely unaddressed.

Chemistry Benchmarks. Chemistry benchmarks have evolved alongside model capabilities. Early suites such as MoleculeNet (Wu et al., 2018) have standardized molecular property prediction, while GuacaMol (Brown et al., 2019) and MOSES (Polykovskiy et al., 2020) provided canonical evaluations for generative chemistry. LLM-specific benchmarks (Guo et al., 2023; Mirza et al., 2025) scale a similar idea to LLM evaluation, and broader scientific benchmarks (Wang et al., 2023; Sun et al., 2024; Laurent et al., 2024) evaluate quantitative and research assistant capabilities. Every existing LLM-targeted benchmark mainly focuses on single-molecule or single-system reasoning. The closest molecular-modeling analog is the SAMPL host–guest blind challenge (Amezcua et al., 2022), which calibrates physics-based free-energy methods on a handful of curated pairs rather than evaluating LLMs. SUPRABENCH fills this gap by evaluating LLMs on supramolecular host–guest reasoning under a unified protocol.

3 SUPRABENCH

3.1 DATASET CONSTRUCTION

Anchor Records. We obtain the anchor host–guest binding records from SupraBank (Biedermann & SupraBank Team, 2020), a public repository that curates the experimentally reported supramolecular interaction records. Specifically, we crawl the host–guest binding measurements associated with the molecular metadata, including host and guest names, identifiers, images, canonical SMILES strings, binding constants, and solvent conditions. Afterward, we obtain 5,362 raw binding records over 2,466 components.

Post-Processing. The collected raw records remain noisy: molecular identifiers are inconsistent across publications, binding constants are reported under heterogeneous experimental conditions (*e.g.*, solvent, temperature, pH), and the same host–guest pair is frequently measured multiple times with occasional outliers. To address these issues, we employ a six-step cleaning pipeline (detailed in Appendix C), *i.e.*, numeric parsing, organic-solvent filtering, default-condition imputation, van't Hoff temperature correction, per-pair averaging, and outlier removal. As a result, we obtain 4,635 high-quality records across 2,008 unique components. Every cleaning step is proposed by domain experts to align with standard practice in supramolecular chemistry, and they further spot-check the cleaned records and review the gold host–guest description answers.

3.2 TASK CONSTRUCTION

SUPRABENCH comprises four fundamental tasks, each targeting a distinct supramolecular reasoning capability, with an auxiliary multimodal task that probes whether the model can identify a molecule from its 2D structural drawing. These tasks are designed by domain experts to reflect the core reasoning capabilities that underpin real-world host–guest discovery workflows, and these tasks are formulated to be answerable from the curated records. The data statistics are provided in Appendix Table 4, and an example record is shown in Appendix Figure 5.

Binding Affinity Prediction (BAP). This task evaluates whether a model can predict the affinity of a host–guest interaction directly from molecular structure, the central design quantity across supramolecular applications, ranging from cyclodextrin-based drug formulations (Loftsson & Brewster, 2010) to cucurbituril-based solubility enhancers (Webber & Langer, 2017). We formulate this task as a regression problem in which the model predicts $\log K_a$ under standard aqueous conditions, given the host and guest SMILES strings.

Top-Binder Selection (TBS). This task challenges a model to reason about relative binding selectivity, *i.e.*, distinguishing the strongest binder among chemically similar candidates, a capability that underpins binder screening and sensor design (You et al., 2015; Kolesnichenko & Anslyn, 2017). We formulate this task as a multiple-choice question, in which the model picks the strongest binder among four candidate guests for a host.

Solvent Identification (SID). This task tests whether a model can infer the experimental measurement context from molecular structure alone, since binding constants are only comparable within a shared solvent regime (Kolesnichenko & Anslyn, 2017). We formulate this task as a six-way classification wherein the model identifies in which of water, DMSO, MeCN, MeOH, $CHCl_3$, or CH_2Cl_2 the reported binding constant is measured.

Host–Guest Description (HGD). This task examines whether a model has physicochemical knowledge of host–guest pairs, a capability that underpins inverse-design workflows such as the macrocyclic sequestrant search that produced sugammadex (Bom et al., 2002) and pillar[6]uril opioid antidotes (Brockett et al., 2023). We formulate this task as an open-ended host–guest QA task with two complementary subtypes: the *reverse* subtype asks the model to describe candidate hosts and their structural features for a given guest, while the *forward* subtype asks the model to describe the property profile and representative examples of guests that bind a given host.

Molecular Identification (MI). This auxiliary task examines whether a multimodal model can recover a molecule's identity from its 2D structural drawing, a vision-grounded capability that complements the four text-based tasks above and exposes precise bond-level reasoning. We formulate it as image-to-SMILES generation in which the model receives a single rendered structure image of a host or guest molecule and emits the corresponding canonical SMILES string. Further task construction details are provided in Appendix D.

3.3 SUPRAPMC CONSTRUCTION

To enable the adaptation for the general-purpose LLMs, we further release **SUPRAPMC**, a 16M-token **Supra**molecular chemistry text corpus obtained from Europe **PMC** (Consortium, 2015), comprising existing articles relevant to the field.

Anchor Corpus. Our text corpus for supramolecular chemistry is obtained from Europe PMC, an open-access repository that indexes biomedical and life sciences citations, via its REST endpoint. We initially retrieve the full abstract index spanning the publication years 1900 through 2026, leading to over forty million articles, of which over eight million contain the full article in XML format.

Supramolecular Text Corpus. With the anchor text corpus, we further filter the supramolecular text corpus by issuing 19 topical queries provided by domain experts that span the supramolecular principle sub-areas, *e.g.*, host–guest chemistry, self-assembly, and molecular recognition, *etc.* As a result, we obtain $420{,}950$ raw filtered articles. Note that these raw, filtered articles still contain residual biomedical contamination, such as papers that mention *host cells* rather than *host molecules*. To this end, we further refine the result with a transparent rule-based filter that combines a positive bank of supramolecular keywords with two reject banks of biomedical contamination. Afterward, we

obtain a high-precision filtered split of 133,867 articles. More details about the filtering logic and keyword lists are discussed in Appendix E.

3.4 Evaluation Protocol

Prompt Strategies. To comprehensively evaluate the model performance, we implement three prompting strategies: (i) **Base**, a zero-shot direct prompt that issues the question and answer schema with no exemplars. (ii) **Few-Shot**, which prepends three to five in-domain demonstrations sampled from the pool of the same task. Note that these few-shot examples are not evaluated as a test set. and (iii) **CoT** (Wei et al., 2022), which requires an explicit reasoning before the final answer. Full prompt templates and examples are provided in Appendix H.

Evaluation Metrics. We employ widely used evaluation metrics for tasks in SupraBench. For binding affinity prediction, we report mean absolute error (MAE) and root mean squared error (RMSE) over extracted $\log K_a$ values. For top-binder selection, we report letter accuracy and host–guest regret over the four candidate guests. For solvent identification, we report both class-balanced accuracy and Macro-F_1 to account for class imbalance across solvent labels. For host–guest description, we evaluate open-ended answers with Rouge-1 recall, precision, and F_1. For the molecular identification task, we report SMILES validity, canonical-SMILES exact match, InChIKey first-block match, molecular-formula match, Morgan-fingerprint Tanimoto similarity, and heavy-atom count error (ΔHeavy), so that exact recovery and chemically meaningful near misses are both visible. Full definitions of these evaluation metrics are discussed in Appendix B.2.

4 Experiments

We evaluate a broad range of LLMs, including open-sourced and proprietary model families, including Qwen3.5-{9B, 27B} (Qwen Team, 2026), Llama-3.1-{8B, 70B}-Instruct (Grattafiori et al., 2024), DeepSeek-v4 (DeepSeek-AI, 2026), GPT-5.4-{Mini, Nano} (OpenAI, 2026), and Gemini-3-Flash (Google, 2023). For a rigorous and fair evaluation, we leverage OpenRouter (OpenRouter, 2026) for all model inferences. Detailed discussion about experimental setup is provided in Appendix B.

4.1 Main Results

Table 1 reports the performance on SupraBench over eight LLMs under three prompting strategies. According to the table, we conclude that: (i) Prompting strategy effects are highly task-dependent. Specifically, Few-Shot improves the performance of the host–guest description task for every model, while downgrading on binding affinity prediction, compared with Base. Moreover, CoT consistently underperforms Few-Shot and Base on the host–guest description task. This observation shows that *no single prompting strategy uniformly improves performance across various tasks.* We analyze this phenomenon in Section 4.6. (ii) Gemini-3-Flash achieves the best score on binding affinity prediction, top-binder selection, and solvent identification across every prompting setting, with DeepSeek-v4 consistently in second place and GPT-5.4-Mini close behind. This observation shows that *frontier proprietary LLMs deliver state-of-the-art on supramolecular chemistry tasks.* (iii) We additionally evaluate four small language models, *i.e.*, Qwen3.5-{0.8B, 2B, 4B} and Llama3.2-3B, in Appendix G.2; none comes close to the frontier models, and Few-Shot or CoT prompting does not close the gap, indicating that SupraBench *requires chemistry capabilities that emerge at larger scales.* Appendix G further shows that Gemini-3-Flash is competitive with specialized ML baselines under a host-disjoint split and that a chemistry-specific LLM judge preserves the Rouge-1 ranking on host–guest description.

4.2 Generalization to Novel Host–Guest Pairs

In supramolecular chemistry, the discovery of new guest scaffolds remains a time-consuming and challenging process. Taking the GPT-5.4 model family (with a reported knowledge cutoff of August 2025) as an example, merely four new guest scaffolds have been reported in the literature since, which is too few to serve as a valid temporal-cut test set; scaling this discovery process is precisely the purpose of SupraBench. Since SupraBank and Europe PMC are both public and may overlap with LLM pretraining data, we quantify the resulting memorization risk by asking domain experts

Table 1: Main performance comparison across the four fundamental tasks of SUPRABENCH. For each setting, the best score is shown in **bold**, and the second-best is underlined.

	Model	Binding Affinity		Top-Binder		Solvent		Host–Guest Description		
		MAE ↓	RMSE ↓	ACC ↑	Regret ↓	F1 ↑	B. Acc ↑	Recall ↑	Precision ↑	F1 ↑
Base	Qwen3.5-9B	2.491	3.360	0.379	0.930	0.159	0.166	0.040	0.023	0.043
	Qwen3.5-27B	1.803	2.503	0.404	0.851	0.225	0.364	0.495	0.072	0.122
	Llama3.1-8B	2.699	3.630	0.228	1.281	0.151	0.225	0.266	0.059	0.092
	Llama3.1-70B	1.632	2.149	0.338	1.054	0.118	0.254	0.487	**0.091**	**0.152**
	GPT-5.4-Mini	1.549	2.182	0.428	0.810	0.219	0.274	0.437	0.086	0.137
	GPT-5.4-Nano	1.642	2.169	0.411	0.816	0.182	0.347	0.472	0.062	0.107
	Gemini-3-Flash	**1.248**	**1.679**	**0.498**	**0.647**	**0.350**	**0.470**	**0.506**	0.067	0.118
	DeepSeek-v4	1.433	1.994	0.461	0.730	0.309	0.381	0.500	0.090	0.141
Few-Shot	Qwen3.5-9B	3.650	4.820	0.370	0.951	0.154	0.150	0.513	0.022	0.042
	Qwen3.5-27B	2.258	3.256	0.392	0.889	0.178	0.257	0.636	**0.585**	**0.580**
	Llama3.1-8B	5.504	6.940	0.283	1.227	0.142	0.182	0.655	0.369	0.456
	Llama3.1-70B	1.774	2.359	0.354	1.026	0.144	0.185	0.631	0.474	0.531
	GPT-5.4-Mini	1.958	2.808	0.430	0.824	0.141	0.291	0.542	0.228	0.307
	GPT-5.4-Nano	2.176	2.894	0.419	0.819	0.190	0.270	0.532	0.095	0.152
	Gemini-3-Flash	**1.257**	**1.702**	**0.513**	**0.619**	**0.389**	**0.421**	0.660	0.364	0.448
	DeepSeek-v4	1.618	2.276	0.470	0.713	0.203	0.225	**0.720**	0.303	0.352
CoT	Qwen3.5-9B	3.664	4.885	0.382	0.944	0.167	0.197	0.300	0.039	0.068
	Qwen3.5-27B	2.438	3.468	0.398	0.898	0.254	0.415	**0.526**	0.051	0.092
	Llama3.1-8B	4.911	6.279	0.293	1.220	0.154	0.153	0.380	**0.102**	**0.144**
	Llama3.1-70B	1.833	2.512	0.373	0.985	0.106	0.380	0.421	0.055	0.097
	GPT-5.4-Mini	2.036	2.887	0.429	0.828	0.220	0.282	0.444	0.080	0.129
	GPT-5.4-Nano	2.160	2.881	0.410	0.822	0.174	0.257	0.492	0.056	0.098
	Gemini-3-Flash	**1.261**	**1.723**	**0.510**	**0.609**	**0.331**	**0.432**	0.512	0.062	0.110
	DeepSeek-v4	1.541	2.183	0.445	0.743	0.307	0.414	0.522	0.080	0.134

to hand-curate a set of 63 novel host–guest pairs for binding affinity prediction that have not been reported in the existing literature and to annotate the binding affinity of each pair as the gold-standard value. Table 2 reports the performance on this novel set alongside the original benchmark. According to the table, we find that: (i) Novel-set error stays in the same order of magnitude as the full benchmark, which indicates that *the evaluation reflects model inference capability rather than retrieval of memorized values.* (ii) The strongest models remain the same on the novel set, *i.e.*, Gemini-3-Flash and GPT-5.4-Nano, with Gemini-3-Flash best in 10 of the 12 columns, which demonstrates that *the benchmark measures a transferable capability rather than dataset-specific lookup.* (iii) The original-to-novel gap is smallest at Base and widens under Few-Shot and CoT. Since a memorized value would be retrievable regardless of the prompt, this prompt-dependent widening indicates that the residual gap stems from *numerical miscalibration on harder chemistry problems rather than from a loss of access to memorized answers.*

4.3 MOLECULAR IDENTIFICATION

We additionally evaluate *molecular identification*, and visualize the results in Figure 3. The full numeric values are reported in Appendix Table 5. According to Figure 3, we observe that: (i) Gemini-3-Flash dominates every column under Few-Shot, while GPT-5.4-Nano collapses to ≤ 0.11 Canonical in every prompting setting despite emitting valid SMILES on a majority of prompts. This observation shows that *frontier multimodal training yields a large image-to-SMILES gap that smaller proprietary models fail to close*. (ii) Across all models, the gap between Canonical and InChIKey or Tanimoto remains large, indicating that models recover the molecular scaffold roughly right (high Tanimoto) but miss the exact connectivity (low Canonical). This observation demonstrates that *the visual chemistry knowledge of current LLMs is partial rather than absent.* (iii) CoT consistently degrades identification quality, *e.g.*, Gemini-3-Flash's Canonical drops from 0.593 under Few-Shot to 0.567 under CoT, suggesting that *an explicit reasoning cue destabilizes image-grounded molecular translation rather than improves it.*

Table 2: Binding affinity prediction on the original benchmark versus the 63 expert-curated novel host–guest pairs. The best score within each prompting strategy is shown in bold.

Prompt	Model	Original		Novel	
		MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
Base	GPT-5.4-Nano	1.642	2.169	**1.685**	**2.162**
	Gemini-3-Flash	**1.248**	**1.679**	1.792	2.245
	DeepSeek-v4	1.433	1.994	1.806	2.272
Few-Shot	GPT-5.4-Nano	2.176	2.894	2.562	3.009
	Gemini-3-Flash	**1.257**	**1.702**	**1.813**	**2.181**
	DeepSeek-v4	1.618	2.276	2.537	3.210
CoT	GPT-5.4-Nano	2.160	2.881	2.829	3.215
	Gemini-3-Flash	**1.261**	**1.723**	**1.859**	**2.267**
	DeepSeek-v4	1.541	2.183	2.143	2.802

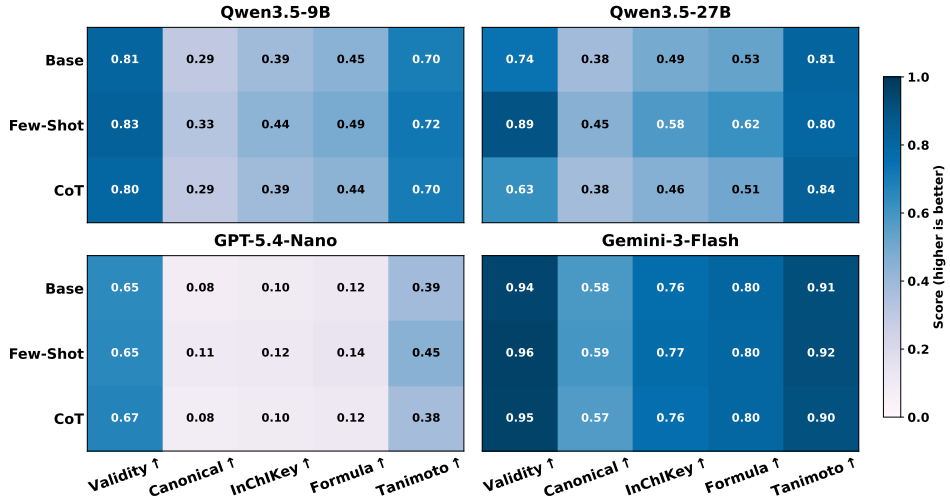

Figure 3: Heatmap for molecular identification task (full numeric values in Table 5).

4.4 Domain Adaptation Analysis

We apply DAPT (Ibrahim et al., 2024; Gururangan et al., 2020) to train three open-weight models, *i.e.*, Qwen3.5-9B, Qwen3.5-27B, and Llama3.1-8B, on our constructed supramolecular chemistry text corpus SupraPMC and evaluate them on all four fundamental tasks under the Base prompting setting. The implementation details are provided in Appendix F, and Table 3 lists the results. We observe that: (i) DAPT substantially improves binding affinity prediction performance for all models, which validates that the SupraPMC *directly transfers to the in-distribution binding affinity regression task.* Moreover, DAPT continues to reduce the binding-affinity error at 27B, mirroring the improvement at 9B, indicating that the gains from domain adaptation *persist as the base scales up.* (ii) DAPT also lifts the model performance in host–guest description tasks, with the smaller Llama3.1-8B showing larger absolute gains than Qwen3.5-9B on most columns. This validates that SupraPMC *exposure transfers beyond regression into open-ended generation*. (iii) The performance on the top-binder selection task degrades for all models, and the Llama variant additionally fails to follow the strict letter format on solvent identification, showing that *DAPT on free-form scientific text trades off against strict letter-format MCQ output*. (iv) *The format-following cost declines with scale.* At 27B the top-binder accuracy is essentially preserved under DAPT, while HGD and SID also improve, indicating that DAPT introduces a net improvement as scale grows. Appendix G.3 ablates the mixing ratio and adapter capacity, showing that this format-following cost is a small-model capacity limitation rather than a recipe artifact.

Table 3: Performance of DAPT on SUPRAPMC for Qwen3.5-9B, Qwen3.5-27B, and Llama3.1-8B under the identical recipe and the Base prompting setting.

Setting	Binding Affinity		Top-Binder		Solvent		Host–Guest Description		
	MAE ↓	RMSE ↓	ACC ↑	Regret ↓	F1 ↑	B. Acc ↑	Recall ↑	Precision ↑	F1 ↑
Qwen3.5-9B	2.491	3.360	**0.379**	0.930	0.159	0.166	0.040	0.023	0.043
+ SUPRAPMC	**2.173**	**2.737**	0.235	**0.873**	**0.161**	**0.167**	**0.053**	**0.048**	**0.050**
Qwen3.5-27B	1.803	2.503	**0.404**	0.851	0.225	0.364	0.495	0.072	0.122
+ SUPRAPMC	**1.757**	**2.376**	0.395	**0.837**	**0.231**	**0.402**	**0.501**	**0.084**	**0.136**
Llama3.1-8B	2.699	3.630	**0.228**	1.281	0.151	0.225	0.266	0.059	0.092
+ SUPRAPMC	**1.636**	**2.204**	0.152	**1.067**	**0.173**	**0.268**	**0.311**	**0.074**	**0.106**

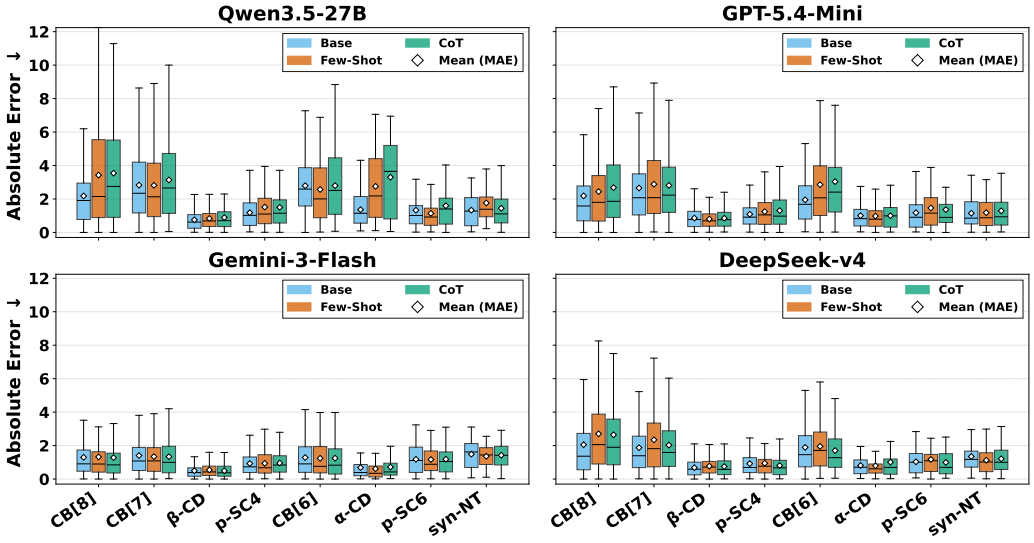

Figure 4: Absolute error distributions for binding affinity prediction grouped by the eight most frequent hosts. Note that CB[n] represents Cucurbit[n]uril, α/β-CD denotes α/β-Cyclodextrin, p-SCn is p-Sulfonatocalix[n]arene, and syn-NT means syn-Amide Naphthotube.

4.5 PER-HOST ANALYSIS

To understand the performance breakdown in the binding affinity prediction task, we conduct an experiment to evaluate the results grouped by the eight most frequent hosts, and report the results in Figure 4. According to Figure 4, we observe that: (i) β-CD is the easiest host for every setting, while 18 of the 20 worst cells fall on CB[n], showing that *per-host difficulty is highly heterogeneous and CB[n] hosts drive the headline error.* (ii) Switching from Base to Few-Shot or CoT leaves the performance of Gemini-3-Flash's error remain unchanged, and degrades the other three models, *i.e.*, Qwen3.5-27B, GPT-5.4-Mini, and DeepSeek-v4. This observation shows that the prompting strategy interacts strongly with base-model capacity rather than uniformly improving it.

4.6 WHY PROMPT STRATEGIES HURT?

To diagnose the CoT regressions reported in Table 1, we walk through a representative failure of DeepSeek-v4 on binding affinity prediction, with the full response shown in Figure 11. For the host–guest pair (*4,7,13,18-Tetraoxa-1,10-diazabicyclo[8.5.5]icosane*, Ba^{2+}) with reference $\log K_a = 2.00$, the Base prompt yields a near-perfect prediction of 2.10, while CoT yields 11.00, off by nine orders of magnitude in the association constant. Inspecting the CoT trace reveals the failure mode: the model confidently asserts that "the binding constant for Ba^{2+} with such bicyclic diaza-crown ethers typically falls in the range of $\log K_a \approx 10$–12" and lands on 11.0 as a "reasonable estimate" from "typical literature values for similar cryptands", while the actual literature value sits near 2.0.

This failure mode generalizes the prompt-strategy regression observed across binding affinity prediction and molecular identification. CoT prompts the model to articulate a reasoning chain, but articulation is not knowledge, *i.e.*, when invited to cite "typical literature values", the model fabricates a confident range that does not exist, and the qualitative chemistry it does recover (preorganized cavity, favorable ion–dipole interactions) has no calibrated mapping to $\log K_a$ magnitudes. Appendix G.5 confirms that this miscalibration holds across every host: models rank hosts correctly (Spearman ≥ 0.75) yet compress the $\log K_a$ scale by roughly a factor of two. In short, *CoT only helps when the model can actually reason about the underlying chemistry, and supramolecular reasoning at scientist-level rigor remains beyond the reach of current LLMs.*

5 INSIGHTS

We close by distilling the key insights from SUPRABENCH throughout the paper.

① **Frontier scale dominates, but every task leaves headroom.** Gemini-3-Flash leads binding affinity prediction, top-binder selection, and solvent identification under every prompting strategy (Table 1), confirming that frontier proprietary multimodal training transfers to supramolecular reasoning. Even so, the best $\log K_a$ MAE remains 1.25, the best top-binder accuracy plateaus at 0.513, and host–guest description Rouge-1 F_1 stays below 0.6 for every model, leaving substantial room for improvement on every task.

② **No prompting strategy is universally helpful.** Few-Shot improves host–guest description for nearly every model but degrades binding affinity prediction, *e.g.*, Llama3.1-8B inflates from MAE 2.699 to 5.504, and CoT consistently hurts binding affinity prediction and molecular identification. The effect is mediated by base-model capacity, *i.e.*, Gemini-3-Flash is largely robust across prompting strategies, while smaller models pay a noticeable penalty. Practitioners should therefore choose prompting strategies per task and per base model, not as a one-size-fits-all default.

③ **CoT amplifies, rather than fixes, the underlying reasoning gap.** The case study in Section 4.6 shows that when a model lacks the supramolecular knowledge to ground its qualitative chemistry in $\log K_a$ magnitudes, asking it to articulate a reasoning chain yields fluent but uncalibrated text and chemically nonsensical predictions. CoT is therefore not a substitute for domain knowledge in supramolecular chemistry, and targeted domain adaptation or external chemistry tools are likely more productive directions for closing the gap.

④ **Domain adaptation has uneven transfer.** A single DAPT recipe on the SUPRAPMC corpus improves binding affinity prediction MAE for both adapted models, *i.e.*, Qwen3.5-9B and Llama3.1-8B (Table 3), but degrades top-binder selection accuracy for both, and the Llama variant collapses on solvent identification due to format-following failures. Free-form scientific text adaptation transfers cleanly to regression but not to strict letter-format MCQ output, suggesting that domain-adaptive pretraining and instruction-format preservation need to be optimized jointly.

⑤ **Molecular identification recovers scaffolds but not exact connectivity.** On the molecular identification task (Figure 3), every model shows a large gap between Canonical SMILES match and the looser InChIKey or Tanimoto similarity metrics, indicating that models recover the molecular scaffold roughly right but miss the exact connectivity. This split signals that the bottleneck is not abstract visual chemistry comprehension but precise bond-level reasoning, an axis on which CoT appears to degrade performance rather than improve it.

6 CONCLUSION

We introduce the first host–guest supramolecular chemistry benchmark for LLMs, called SUPRABENCH, that contains four fundamental tasks, *i.e.*, binding affinity prediction, top-binder selection, solvent identification, and host–guest description, along with an auxiliary vision-based task for molecular identification. Moreover, we release a 16M-token text corpus, SUPRAPMC, comprising Europe PMC supramolecular chemistry articles that can contribute future studies in this domain. We evaluate eight LLMs and reveal several key findings. These results highlight that supramolecular chemistry remains a genuine bottleneck for current LLMs, and we hope SUPRABENCH catalyzes future research on chemistry-grounded LLMs.

ETHICS STATEMENT

Data and licensing. The host–guest binding records in SUPRABENCH are derived from SupraBank (Biedermann & SupraBank Team, 2020), a public repository of experimentally reported supramolecular interactions, and are redistributed under its public-distribution terms. The SUPRAPMC text corpus is built only from open-access Europe PMC articles, and each article remains subject to its individual license; we release document identifiers and the filtering pipeline so that the corpus can be rebuilt from the source. The 63 novel host–guest pairs are annotated by domain experts who collaborate on this work. The benchmark contains no human-subjects data and no personally identifiable information, and no IRB approval was required.

Intended use and potential misuse. SUPRABENCH is a research benchmark for evaluating LLM reasoning on host–guest supramolecular chemistry. It is not intended to drive clinical, regulatory, or production drug-design decisions without expert oversight and experimental validation. LLMs can produce confident, fluent, but chemically incorrect predictions, especially under CoT prompting; downstream users should treat model outputs as hypotheses to be verified against the literature and established computational or experimental methods. We do not foresee dual-use risks beyond those already present in the public literature from which the benchmark is derived.

REPRODUCIBILITY STATEMENT

We release our datasets and source code, including the evaluation code, all prompt templates, the SUPRAPMC corpus construction pipeline, and the DAPT recipe. Appendix C details the six-step cleaning pipeline, Appendix E the corpus collection and filtering, Appendix B the evaluation metrics and compute environment, and Appendix F the DAPT hyperparameters. Random number generators are seeded throughout task construction and evaluation. All base models, open-weight and proprietary, are queried through OpenRouter; the DAPT variants are evaluated locally from our released LoRA adapters. OpenRouter does not pin proprietary model versions across requests, so exact reproduction of those numbers depends on provider-side version stability beyond our control; we record the model identifier and request date with each inference run.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementation. LLM-generated code was verified and tested for correctness. LLMs are also part of our evaluation methodology, as reported in the paper, i.e., an LLM judge validates the topical precision of SUPRAPMC (Appendix E.2) and a chemistry-specific LLM judge scores host–guest descriptions (Appendix G.4). Both are evaluation instruments whose outputs we report as results, not as substitutes for our own analysis. We also used generative AI tools to polish the paper, and we have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Martin Amezcua, Jeffry Setiadi, Yunhui Ge, and David L Mobley. An overview of the sampl8 host–guest binding challenge. *Journal of computer-aided molecular design*, 36(10):707–734, 2022.

Katsuhiko Ariga, Hiroshi Ito, Jonathan P Hill, and Hiroshi Tsukube. Molecular recognition: from solution science to nano/materials technology. *Chemical Society Reviews*, 2012.

Peter William Atkins, Julio De Paula, and James Keeler. *Atkins' physical chemistry*. Oxford university press, 2023.

Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):20, 2015.

Frank Biedermann and SupraBank Team. SupraBank: An open resource for intermolecular interactions. `https://suprabank.org`, 2020.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 2023.

Anton Bom, Mark Bradley, Ken Cameron, John K Clark, Jan Van Egmond, Helen Feilden, Elizabeth J MacLean, Alan W Muir, Ronald Palin, David C Rees, et al. A novel concept of reversing neuromuscular block: chemical encapsulation of rocuronium bromide by a cyclodextrin-based synthetic host. *Angewandte Chemie International Edition*, 2002.

Adam T Brockett, Weijian Xue, David King, Chun-Lin Deng, Canjia Zhai, Michael Shuster, Shivangi Rastogi, Volker Briken, Matthew R Roesch, and Lyle Isaacs. Pillar [6] maxq: A potent supramolecular host for in vivo sequestration of methamphetamine and fentanyl. *Chem*, 2023.

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

Chaoran Chen, Zhiping Zhang, Bingcan Guo, Shang Ma, Ibrahim Khalilov, Simret Gebreegziabher, Yanfang Ye, Ziang Xiao, Yaxing Yao, Tianshi Li, et al. The obvious invisible threat: Llm-powered gui agents' vulnerability to fine-print injections. In *Soups*. The Twenty-First Symposium on Usable Privacy and Security (SOUPS), 2025a.

Chaoran Chen, Daodao Zhou, Yanfang Ye, Toby Jia-jun Li, and Yaxing Yao. Clear: Towards contextual llm-empowered privacy policy analysis and risk generation for large language model applications. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 277–297, 2025b.

Matheus C Colaço, Vinícius A Glitz, Amanda K Jacobs, Vinícius C Port, and Giovanni F Caramori. Supramolecular chemistry: Exploring the use of electronic structure, molecular dynamics, and machine learning approaches. *European Journal of Organic Chemistry*, 2024.

Europe PMC Consortium. Europe pmc: a full-text literature database for the life sciences and platform for innovation. *Nucleic acids research*, 43(D1):D1042–D1048, 2015.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL `https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro/blob/main/DeepSeek_V4.pdf`.

Chun-Lin Deng, Steven L Murkli, and Lyle D Isaacs. Supramolecular hosts as in vivo sequestration agents for pharmaceuticals and toxins. *Chemical Society Reviews*, 49(21):7516–7532, 2020.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In *EMNLP*, 2022.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *ICLR*, 2024.

Bin Feng, Zijing Liu, Mingjun Yang, Junjie Zou, He Cao, Yu Li, Lei Zhang, and Sheng Wang. A foundation model for protein-ligand affinity prediction through jointly optimizing virtual screening and hit-to-lead optimization. *bioRxiv*, pp. 2025–02, 2025.

Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3 (4):128–135, 1999.

Google. Gemini 3: A new era of intelligence, 2023. URL https://blog.google/products/gemini/gemini-3/.

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

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *NeurIPS*, 2023.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8342–8360, 2020.

Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, and Igor Pletnev. Inchi-the worldwide chemical structure identifier standard. *Journal of cheminformatics*, 5(1):7, 2013.

Stephen R Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. Inchi, the iupac international chemical identifier. *Journal of cheminformatics*, 7(1):23, 2015.

Edwin A Hill. On a system of indexing chemical literature; adopted by the classification division of the us patent office. *Journal of the American Chemical Society*, 22(8):478–494, 1900.

David C Hoaglin and Boris Iglewicz. Fine-tuning some resistant rules for outlier labeling. *Journal of the American statistical Association*, 82(400):1147–1149, 1987.

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

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. Simple and scalable strategies to continually pre-train large language models. *arXiv preprint arXiv:2403.08763*, 2024.

Mingxuan Ju, Wenhao Yu, Tong Zhao, Chuxu Zhang, and Yanfang Ye. Grape: Knowledge graph enhanced passage reader for open-domain question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 169–181, 2022.

Mingxuan Ju, Tong Zhao, Wenhao Yu, Neil Shah, and Yanfang Ye. Graphpatcher: mitigating degree bias for graph neural networks via test-time augmentation. *Advances in Neural Information Processing Systems*, 36:55785–55801, 2023.

Igor V Kolesnichenko and Eric V Anslyn. Practical applications of supramolecular chemistry. *Chemical Society Reviews*, 46(9):2385–2390, 2017.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. 2024.

Greg Landrum. Rdkit: Open-source cheminformatics software. 2016. URL https://github.com/rdkit/rdkit/releases/tag/Release_2016_09_4.

Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

Namkyeong Lee, Dongmin Hyun, Gyoung S. Na, Sungwon Kim, Junseok Lee, and Chanyoung Park. Conditional graph information bottleneck for molecular relational learning. In *International Conference on Machine Learning*, pp. 18852–18871, 2023.

Jean-Marie Lehn. *Supramolecular chemistry: concepts and perspectives*. John Wiley & Sons, 1995.

Chen-An Li and Hung-Yi Lee. Examining forgetting in continual pre-training of aligned large language models. *arXiv preprint arXiv:2401.03129*, 2024.

Tianyi Li, Zhengyu Fang, Xiaoge Zhang, Kaiyu Tang, Huiyuan Chen, Zhimeng Jiang, Tianxiang Zhao, Rong Xu, Feixiong Cheng, Xiao Li, et al. Druglm: a unified framework to enhance drug-target interaction predictions by incorporating textual embeddings via language models. *bioRxiv*, pp. 2025–07, 2025.

Kaifeng Liu, Xiangyu Yu, Huizi Cui, Wannan Li, and Weiwei Han. Gpt4kinase: high-accuracy prediction of inhibitor-kinase binding affinity utilizing large language model. *International Journal of Biological Macromolecules*, 282:137069, 2024.

Micha Livne, Zulfat Miftahutdinov, Elena Tutubalina, Maksim Kuznetsov, Daniil Polykovskiy, Annika Brundyn, Aastha Jhunjhunwala, Anthony Costa, Alex Aliper, Alán Aspuru-Guzik, et al. nach0: multimodal natural and chemical languages foundation model. *Chemical Science*, 2024.

Thorsteinn Loftsson and Marcus E Brewster. Pharmaceutical applications of cyclodextrins: basic science and product development. *Journal of pharmacy and pharmacology*, 2010.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. Fineweb-edu: the finest collection of educational content, 2024. URL `https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu`.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nat. Mach. Intell*, 2024.

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

David L Mobley and Michael K Gilson. Predicting binding free energies: frontiers and benchmarks. *Annual review of biophysics*, 46:531–558, 2017.

Harry L Morgan. 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, 1965.

Benjamin L Olesnicky, Clare Farrell, Phoebe Clare, Shelly Wen, Kate Leslie, and Anthony Delaney. The effect of sugammadex on patient morbidity and quality of recovery after general anaesthesia: a systematic review and meta-analysis. *British Journal of Anaesthesia*, 132(1):107–115, 2024.

OpenAI. Introducing gpt-5.4. 2026. URL `https://openai.com/index/introducing-gpt-5-4/`.

OpenRouter. OpenRouter: A Unified Interface for LLMs. `https://openrouter.ai`, 2026.

Juan M Parrilla-Gutiérrez, Jarosław M Granda, Jean-François Ayme, Michał D Bajczyk, Liam Wilbraham, and Leroy Cronin. Electron density-based gpt for optimization and suggestion of host–guest binders. *Nature computational science*, 4(3):200–209, 2024.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

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

Zhenting Qi, Fan Nie, Alexandre Alahi, James Zou, Himabindu Lakkaraju, Yilun Du, Eric Xing, Sham Kakade, and Hanlin Zhang. Evolm: In search of lost language model training dynamics. In *NeurIPS*, 2025.

Yiyue Qian, Chunhui Zhang, Yiming Zhang, Qianlong Wen, Yanfang Ye, and Chuxu Zhang. Co-modality graph contrastive learning for imbalanced node classification. *Advances in Neural Information Processing Systems*, 35:15862–15874, 2022.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Nadine Schneider, Roger A Sayle, and Gregory A Landrum. Get your atoms in order: An open-source implementation of a novel and robust molecular canonicalization algorithm. *Journal of chemical information and modeling*, 55(10):2111–2120, 2015.

Jonathan W Steed and Jerry L Atwood. *Supramolecular chemistry*. John Wiley & Sons, 2022.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *AAAI*, 2024.

Xiangru Tang, Tianyu Hu, Muyang Ye, Yanjun Shao, Xunjian Yin, Siru Ouyang, Wangchunshu Zhou, Pan Lu, Zhuosheng Zhang, Yilun Zhao, et al. Chemagent: Self-updating memories in large language models improves chemical reasoning. In *ICLR*, 2025.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

Pall Thordarson. Determining association constants from titration experiments in supramolecular chemistry. *Chemical Society Reviews*, 40(3):1305–1323, 2011.

John Wilder Tukey et al. *Exploratory data analysis*, volume 2. Springer, 1977.

Jacobus Henricus Van't Hoff. *Etudes de dynamique chimique*, volume 1. Muller, 1884.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.

Matthew J Webber and Robert Langer. Drug delivery by supramolecular design. *Chemical Society Reviews*, 46(21):6600–6620, 2017.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

David Weininger. 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, 1988.

David Weininger, Arthur Weininger, and Joseph L Weininger. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of chemical information and computer sciences*, 29(2):97–101, 1989.

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

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019.

Yanfang Ye, Zheyuan Zhang, Tianyi Ma, Zehong Wang, Yiyang Li, Shifu Hou, Weixiang Sun, Kaiwen Shi, Yijun Ma, Wei Song, et al. Llms4all: A review of large language models across academic disciplines. *arXiv preprint arXiv:2509.19580*, 2025.

Jian Yin, Niel M Henriksen, David R Slochower, Michael R Shirts, Michael W Chiu, David L Mobley, and Michael K Gilson. Overview of the sampl5 host–guest challenge: Are we doing better? *Journal of computer-aided molecular design*, 31(1):1–19, 2017.

Lei You, Daijun Zha, and Eric V Anslyn. Recent advances in supramolecular analytical chemistry using optical sensing. *Chemical reviews*, 2015.

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

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

Jianan Zhao, Qianlong Wen, Shiyu Sun, Yanfang Ye, and Chuxu Zhang. Multi-view self-supervised heterogeneous graph embedding. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 319–334. Springer, 2021.

Jianan Zhao, Qianlong Wen, Mingxuan Ju, Chuxu Zhang, and Yanfang Ye. Self-supervised graph structure refinement for graph neural networks. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, pp. 159–167, 2023.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu, Su Zhu, et al. Developing chemdfm as a large language foundation model for chemistry. *Cell Rep. Phys. Sci.*, 2025.

Appendix Contents

A Detailed Related Works **17**

A.1 LLMs for Chemistry 17

A.2 Chemistry Benchmarks 17

A.3 Supramolecular Chemistry 18

B Implementation Details. **18**

B.1 Environment 18

B.2 Evaluation Metrics. 18

C Dataset Construction **20**

D Task Construction **21**

D.1 Binding Affinity Prediction 21

D.2 Top-Binder Selection. 22

D.3 Solvent Identification 22

D.4 Host-Guest Description 22

D.5 Molecular Identification 22

E Text Corpus Construction Details **23**

E.1 Text Collection. 23

E.2 Corpus Quality Validation 23

F Implementation Details for Domain Adaptation **24**

G Additional Experiments **24**

G.1 Comparison with Specialized ML Baselines 24

G.2 Performance of Small Language Models 25

G.3 DAPT Recipe Ablation 25

G.4 Chemistry-Aware Evaluation for Host–Guest Description 26

G.5 Systematic Failure Analysis 26

H SUPRABENCH Examples **27**

A Detailed Related Works

A.1 LLMs for Chemistry

Although AI-driven methods Zhao et al. (2023); Ju et al. (2022); Qian et al. (2022); Zhao et al. (2021); Ju et al. (2023) have long been applied to a wide range of chemistry problems, recent advances in LLMs have significantly expanded the scope of what AI systems can achieve in chemical understanding and reasoning. LLMs can now read, write, and reason about chemistry with growing fluency. Early efforts treated chemistry as one slice of a broader scientific corpus: Galactica pretrained a single model on millions of papers, textbooks, and reference works, and showed that a generalist LM can recall chemical facts and manipulate SMILES strings with surprising fidelity (Taylor et al., 2022). A parallel line of work has framed chemistry as translation between natural language and molecular notation: MolT5 (Edwards et al., 2022) jointly pre-trains a sequence-to-sequence model on text and SMILES, and nach0 (Livne et al., 2024) extends this idea into a unified multitask foundation model that handles named entity recognition, property prediction, and forward and retro reaction prediction within a single decoder. Building on these foundations, subsequent work has pushed in two complementary directions. The first is domain-adapted LLMs that continue training on chemistry-specific corpora and instruction data; representative examples include ChemLLM, ChemDFM, LlaSMol, and the Mol-Instructions resource, which together show that targeted adaptation of mid-scale open models (7–13B) can match or exceed much larger general-purpose models on molecular description, property prediction, and reaction tasks (Zhang et al., 2024; Zhao et al., 2025; Yu et al., 2024; Fang et al., 2024). The second is LLM-driven chemistry agents that couple an LLM controller to external tools and laboratory automation. ChemCrow (M. Bran et al., 2024) interfaces a planner with retrosynthesis, property-prediction, and web-search modules to execute multi-step chemical workflows; Coscientist (Boiko et al., 2023) drives a cloud lab end-to-end, autonomously planning and running palladium-catalysed cross-coupling optimisations; and ChemAgent (Tang et al., 2025) maintains a self-updating set of planning, execution, and knowledge memories that improves chemical reasoning as the agent accumulates experience. Despite this progress, existing systems are almost exclusively evaluated on small-molecule tasks such as single-molecule property prediction, retrosynthesis, and reaction-yield estimation. None target supramolecular reasoning, which involves a pair (or set) of molecules interacting through non-covalent forces and therefore stresses different capabilities, including host–guest geometry, binding thermodynamics, and application-aware design. SupraBench addresses this gap with tasks that span affinity prediction, host–guest description, application-level inference, and property-conditioned generation.

A.2 Chemistry Benchmarks

Benchmarks have been a central driver of progress in machine-learning chemistry. For graph- and fingerprint-based deep models, MoleculeNet (Wu et al., 2018) unified property-prediction tasks across quantum mechanics, physiology, and toxicity, Therapeutics Data Commons (Huang et al., 2021) extended the same idea to drug-discovery pipelines, and GuacaMol (Brown et al., 2019) and MOSES (Polykovskiy et al., 2020) standardized the evaluation of generative models for de novo molecule design. For LLMs specifically, ChemLLMBench (Guo et al., 2023) provided an early eight-task suite spanning name prediction, property classification, and reaction prediction, and ChemBench (Mirza et al., 2025) subsequently scaled the idea to over 2,700 expert-annotated questions and showed that frontier proprietary models can rival or exceed expert chemists on textbook-style queries. Broader scientific suites complement these chemistry-specific resources: SciBench and SciEval (Wang et al., 2023; Sun et al., 2024) probe quantitative reasoning across chemistry, physics, and biology, while LAB-Bench (Laurent et al., 2024) evaluates research-assistant capabilities such as literature search, protocol planning, and figure interpretation in the life sciences. Across this landscape, however, every benchmark targeted at LLMs focuses on single-molecule or single-system reasoning: property prediction for one molecule, generation of one molecule, or knowledge questions about a single concept or protocol. The closest analog on the molecular-modeling side is the SAMPL host–guest blind challenge (Amezcua et al., 2022), but it is designed to calibrate physics-based free-energy methods on a handful of curated pairs rather than to evaluate language models, and it does not cover application-level outcomes such as drug-delivery vehicle selection, sensor design, or toxin sequestration. SupraBench fills this gap by evaluating LLMs on supramolecular host–guest reasoning under a unified protocol, with task families that span both foundational competencies and application-level inference.

A.3 SUPRAMOLECULAR CHEMISTRY

Supramolecular chemistry studies molecular assemblies held together not by covalent bonds but by reversible non-covalent interactions, including hydrogen bonding, hydrophobic effects, π-stacking, electrostatic forces, and metal coordination (Lehn, 1995; Steed & Atwood, 2022). A typical system pairs a host, often a macrocycle with a well-defined cavity, with a guest bound through molecular recognition (Ariga et al., 2012). The dominant host families in current practice include cyclodextrins, cucurbiturils, calixarenes, pillararenes, and crown ethers, as well as extended porous architectures such as metal-organic and covalent-organic frameworks (MOFs and COFs). These platforms underpin three major clinical applications of supramolecular chemistry. In drug delivery, cyclodextrin-based formulations improve the solubility and bioavailability of hydrophobic active ingredients (Loftsson & Brewster, 2010). In sensing, host–guest recognition coupled with optical or electrochemical reporters enables detection of analytes ranging from pollutants to disease biomarkers (You et al., 2015). In toxin sequestration, macrocyclic receptors are being developed as in-vivo antidotes for drugs of abuse and other toxins, building on the sugammadex precedent (Deng et al., 2020); an acyclic pillararene host, for example, sequesters both methamphetamine and fentanyl in vivo and reverses their pharmacological effects (Brockett et al., 2023).

The same properties that make supramolecular systems powerful also make them difficult to model. Binding affinities are highly sensitive to solvent, pH, ionic strength, and counter-ion identity, so values reported by different laboratories are often not directly comparable. Both hosts and guests can adopt multiple conformations, and a single pair may bind in more than one geometry (*e.g.*, inclusion vs. external complexation), which complicates experimental fitting and computational simulation alike. High-fidelity tools such as DFT and molecular dynamics can in principle resolve these ambiguities, but they are slow, expensive, and require expert setup; even in the carefully curated SAMPL host–guest blind challenges, leading free-energy methods still incur RMSE on the order of 1–2 kcal/mol on small benchmark sets (Amezcua et al., 2022). As a consequence, curated binding-affinity datasets remain small and fragmented relative to the small-molecule property-prediction datasets that drive mainstream ML chemistry, and recent reviews have identified data scarcity and heterogeneity as the central bottleneck for machine learning on supramolecular systems (Colaço et al., 2024). SUPRABENCH is designed to expose and quantify these difficulties for modern LLMs.

B IMPLEMENTATION DETAILS

B.1 ENVIRONMENT

All experiments are conducted on a Linux OS equipped with four Nvidia A100 GPUs. The models are implemented using PyTorch 2.4.0 with CUDA 12.1 and Python 3.11.5.

B.2 EVALUATION METRICS

In this section, we discuss the employed evaluation metrics in SUPRABENCH. For a test sample i, $\hat{y}_i$ and y_i denote the extracted model prediction and the gold reference, respectively.

MAE and RMSE. For scalar predictions $\hat{y}_i$ against scalar gold references y_i,

$$\text{MAE} \;=\; \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|, \tag{1}$$

$$\text{RMSE} \;=\; \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}. \tag{2}$$

Letter Accuracy. For categorical predictions $\hat{\ell}_i$ drawn from a discrete option set $\{A, B, C, D, \ldots\}$ against gold labels ℓ_i,

$$\text{ACC} \;=\; \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\left[\hat{\ell}_i = \ell_i\right]. \tag{3}$$

Binding-Affinity Regret. For the top-binder selection task, accuracy treats every wrong option equally even when some wrong choices are close to the strongest binder. Let $\mathcal{O}_i$ denote the candidate option set for item i, let $k_i^\star = \arg\max_{k \in \mathcal{O}_i} \log K_{a,i}^{(k)}$ be the strongest candidate, and let $\hat{k}_i$ be the model-selected candidate. We define binding-affinity regret as the mean loss in $\log K_a$ caused by selecting $\hat{k}_i$ instead of $k_i^\star$,

$$\text{Regret} \;=\; \frac{1}{N} \sum_{i=1}^{N} \left(\log K_{a,i}^{(k_i^\star)} - \log K_{a,i}^{(\hat{k}_i)} \right). \tag{4}$$

This metric distinguishes chemically near-optimal mistakes from selections that miss the strongest binder by several orders of magnitude.

Macro-F_1. For multi-class classification over a class set $\mathcal{C}$, Macro-F_1 is the unweighted mean of per-class F_1,

$$\text{Macro-}F_1 \;=\; \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{2\, P_c\, R_c}{P_c + R_c}, \tag{5}$$

where P_c and R_c are the per-class precision and recall.

Balanced Accuracy. For the solvent identification task, the label distribution is imbalanced because aqueous measurements dominate the dataset. We therefore also report balanced accuracy, defined as the unweighted mean of per-class recall over the solvent class set $\mathcal{C}$,

$$\text{Balanced-ACC} \;=\; \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} R_c. \tag{6}$$

Balanced accuracy penalizes majority-class collapse and directly measures whether a model recovers minority solvent regimes, rather than only exploiting the dominant water prior.

Rouge-1 Recall, Precision, and F_1. Let $T(\cdot)$ denote the unigram token bag of a text after Porter stemming and case folding. For a prediction $\hat{a}_i$ and gold answer a_i, the per-row Rouge-1 quantities are

$$R_i = \frac{|T(\hat{a}_i) \cap T(a_i)|}{|T(a_i)|}, \tag{7}$$

$$P_i = \frac{|T(\hat{a}_i) \cap T(a_i)|}{|T(\hat{a}_i)|}, \tag{8}$$

$$F_{1,i} = \frac{2\, P_i\, R_i}{P_i + R_i}. \tag{9}$$

We report the unweighted means $\overline{R}$, $\overline{P}$, $\overline{F_1}$ across the N examples, and empty predictions contribute zero to all three.

Validity. This metric computes the fraction of predicted SMILES strings $\hat{s}_i$ that parse successfully (Polykovskiy et al., 2020; Brown et al., 2019; Weininger, 1988). Let $\mathrm{Mol}(\cdot)$ denote RDKit's SMILES parser, where $\mathrm{Mol}(s) = \perp$ indicates a parse failure, validity is computed as:

$$\text{Validity} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\mathrm{Mol}(\hat{s}_i) \neq \perp]. \tag{10}$$

We additionally define $\mathcal{V} = \{i : \mathrm{Mol}(\hat{s}_i) \neq \perp \wedge \mathrm{Mol}(s_i) \neq \perp\}$ as the set of examples on which both the prediction and the reference parse, used by Tanimoto and ΔHeavy below.

Canonical-SMILES Exact Match. Canonical (Schneider et al., 2015; Weininger et al., 1989) is the strictest correctness criterion, *i.e.*, a prediction counts only when its RDKit canonical SMILES coincides with the gold,

$$\text{Canonical} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\mathrm{Canon}(\hat{s}_i) = \mathrm{Canon}(s_i)]\,. \tag{11}$$

This metric rewards recovery of the exact constitution, bond orders, and (where written) stereochemistry of the reference; predictions that differ only in non-semantic SMILES variation (atom ordering, branch direction, aromaticity notation) canonicalize to the same form and are scored correctly.

InChIKey First-Block Match. InChIKey relaxes Canonical along the stereo and isotope dimensions (Heller et al., 2015; 2013). Let $\mathrm{IK}(s)$ denote the first 14-character block of the InChIKey of s, which is a hash over the molecule's main InChI layers (formula, connectivity, and fixed-hydrogen), so molecules that differ only in stereodescriptors or isotopic labeling collapse to the same prefix,

$$\mathrm{InChIKey} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\mathrm{IK}(\hat{s}_i) = \mathrm{IK}(s_i)]. \tag{12}$$

The gap between Canonical and InChIKey quantifies how often the model recovers connectivity but misses stereo.

Molecular-Formula Match. Formula is the most lenient identity criterion, *i.e.*, a prediction counts when its Hill-system molecular formula $F(\cdot)$ equals the gold. We follow Hill (1900) to implement this metric, which is computed as:

$$\mathrm{Formula} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[F(\hat{s}_i) = F(s_i)]. \tag{13}$$

A high Formula with low InChIKey indicates that the model reads atom counts off the image but cannot recover bonding.

Tanimoto Similarity. Tanimoto (Morgan, 1965; Bajusz et al., 2015; Rogers & Hahn, 2010) is a soft structural similarity that smooths the binary identity metrics into a continuous score. For each $i \in \mathcal{V}$, let $f_i = \mathrm{FP}(\hat{s}_i)$ and $g_i = \mathrm{FP}(s_i)$ denote Morgan circular fingerprints with radius r=2 and 2048-bit folding for the prediction and reference. Pairs where either fingerprint is undefined ($i \notin \mathcal{V}$) contribute 0, so the metric is normalized over the full split,

$$\mathrm{Tanimoto} = \frac{1}{N}\sum_{i\in\mathcal{V}} \frac{|f_i \cap g_i|}{|f_i \cup g_i|}. \tag{14}$$

This graded similarity lets a near-miss prediction (*e.g.*, correct scaffold with one substituent wrong) earn partial credit it would not receive from exact-match metrics.

Heavy-Atom Count Difference (ΔHeavy). ΔHeavy (Landrum, 2016) measures how far the prediction deviates from the reference in molecular size, irrespective of connectivity. Let $H(\cdot)$ denote the number of non-hydrogen atoms. We report the mean absolute difference over parseable pairs,

$$\Delta\mathrm{Heavy} \;=\; \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} |H(\hat{s}_i) - H(s_i)|, \tag{15}$$

in units of heavy atoms, with lower values being better.

C Dataset Construction

We construct SupraBench from experimentally reported host–guest binding records collected from SupraBank (Biedermann & SupraBank Team, 2020), a public repository[1] of peer-reviewed interactions in supramolecular chemistry. We crawl each interaction together with its associated molecular metadata, *i.e.*, host and guest names, identifiers, two-dimensional structure images, canonical SMILES strings, binding constants, solvent conditions, temperatures, pH, and the original literature citation, and de-duplicate by the SupraBank interaction ID. The raw data contains 5,362 measurements spanning 2,466 unique molecular components from diverse supramolecular host families, including cucurbiturils, cyclodextrins, calixarenes, pillararenes, cavitands, and naphthotubes. To obtain a clean, maintainable, and reproducible benchmark, we apply a six-step cleaning pipeline, and the final cleaned data contains 4,635 records over 2,008 components. Every cleaning step is proposed by domain experts to align with standard practice in supramolecular chemistry.

Step 1: Numeric Parsing. SupraBank stores binding constants, temperature, and pH as free-form strings with heterogeneous notation, *e.g.*, `"7.76·10^4"`, `"1.12·10^7 M^-1"`, `"25.0°C"`. We implement a regex-based parser that normalizes these into numeric fields, *i.e.*, K_a, $\log K_a$, and pH, handling scientific notation, units, and unicode multiplication marks.

[1] https://suprabank.org/

(a) # Sample per task

Task	# Samples
BAP	2,609
TBS	2,264
SID	2,172
HGD	135

(b) # Sample in top-4 hosts

Host	BAP	TBS	SID
CB[8]	261	200	571
CB[7]	217	200	217
β-CD	201	200	264
p-SC4	144	144	225

Table 4: **Dataset statistics for SUPRABENCH.** Here, BAP, TBS, SID, and HGD denote binding affinity prediction, top-binder selection, solvent identification, and host–guest description tasks, respectively.

Step 2: Organic-Solvent Filtering. Binding constants measured in organic solvents are thermodynamically incomparable to those measured in aqueous media. We discard records in which the field contains any of the common organic-solvent tokens, *e.g.*, {methanol, acetonitrile, DMSO, chloroform, dichloromethane, toluene, acetone, . . .}, and retain water-based systems, *e.g.*, water, buffer, and D_2O.

Step 3: Default-Condition Imputation. Many literature reports omit temperature and pH when measurements are taken under standard conditions. We impute missing values with the most-common literature defaults: $T = 25\,^{\circ}\mathrm{C}$ for 2,238 records missing temperature, and $\mathrm{pH} = 7.0$ for 3,358 records missing pH.

Step 4: van't Hoff Temperature Correction. To make binding constants comparable across studies, we correct all K_a values to a single reference temperature, *i.e.*, $T_{\mathrm{ref}} = 298.15$ K, through the van't Hoff relation (Atkins et al., 2023; Van't Hoff, 1884),

$$\ln \frac{K_a(T_{\mathrm{ref}})}{K_a(T)} \;=\; -\frac{\Delta H^{\circ}}{R}\left(\frac{1}{T_{\mathrm{ref}}} - \frac{1}{T}\right). \tag{16}$$

Since most of the records are missing the ΔH° field, we employ literature-averaged ΔH° values for the most common hosts, *e.g.*, CB[7] as -40 kJ/mol, CB[8] as -35 kJ/mol, β-CD as -20 kJ/mol, *etc.*

Step 5: Per-Pair Averaging. For a measurement tuple, *i.e.*, (`host`, `guest`, `pH`, `solvent`), we compute the geometric mean of K_a, which is equivalent to the arithmetic mean of $\log K_a$. Note that we bin pH to 0.5 units, *e.g.*, pH 6.8 and pH 7.1 collapse to `pH bin`$= 7.0$, while pH 6.0 and pH 7.0 remain distinct.

Step 6: Outlier Removal. Within each host–guest pair, we further employ Tukey's $1.5\,\mathrm{IQR}$ rule (Tukey et al., 1977; Hoaglin & Iglewicz, 1987) on $\log K_a$ to flag outliers. Specifically, a record is removed if it falls outside $[Q_1 - 1.5\,\mathrm{IQR},\ Q_3 + 1.5\,\mathrm{IQR}]$.

Final Output. The cleaning pipeline above yields $4{,}635$ records over $2{,}008$ components Each record carries K_a, $\log K_a$, and ΔG values at $25\,^{\circ}\mathrm{C}$, full provenance fields (host/guest names, PubChem CIDs, SMILES, original literature citation), and a complete pipeline trail for reproducibility. An example record is shown in Figure 5 in Appendix H.

D TASK CONSTRUCTION

D.1 BINDING AFFINITY PREDICTION

Each row in the cleaned binding data, *i.e.*, a host–guest pair with a numeric $\log K_a$ value and both host and guest SMILES available, becomes one regression question. The prompt provides the host and guest names together with their canonical SMILES strings and asks the model to return a single $\log K_a$ value under standard aqueous conditions. We hold out three exemplars drawn at the lowest, the median, and the highest $\log K_a$ in the pool as shared few-shot demonstrations and exclude them from the evaluation set. The gold answer is preserved at four decimals, while few-shot demonstrations are rounded to one decimal so that the model is not biased toward a specific precision. Random number generators are seeded for reproducibility.

D.2 TOP-BINDER SELECTION

We first drop hosts that admit fewer than four distinct guests. Each question is then constructed by sampling four distinct guests uniformly at random from the remaining pool. The sample is rejected if the spread between the highest and lowest $\log K_a$ falls below 0.5, so that the gold answer leads the runner-up by a non-trivial margin, and rejected if the same four-guest set has already been sampled for this host. The four candidates are randomly permuted across the answer letters A through D, and the correct letter is taken as the position of the maximum-$\log K_a$ guest among the displayed options. We cap the number of questions per host at 200 to prevent over-represented hosts from dominating the evaluation. Three exemplars produced by the same procedure on the first three eligible hosts in alphabetical order are held out as shared few-shot demonstrations and excluded from the evaluation set.

D.3 SOLVENT IDENTIFICATION

We construct this task from the same cleaned binding records as above, but use the experimental solvent annotation rather than the binding affinity itself. Each raw solvent string is mapped to one of six canonical classes (water, DMSO, MeCN, MeOH, $CHCl_3$, CH_2Cl_2) via a curated synonym map that absorbs deuterated and salt-buffered variants such as D_2O, buffer, DMSO-d_6, and CD_2Cl_2, and rows whose primary solvent reads "complex" are rescued through the secondary listing where possible; rows that match no canonical class are dropped. Each question is a six-way multiple-choice question with the candidate solvent classes in fixed letter order A through F, and the prompt embeds a short domain guidance text relating qualitative host features, *i.e.*, cavity size, functional groups, charge state, hydrophobicity, to the operating solvent regime, so the model is asked to reason from molecular structure rather than retrieve memorized study metadata. The few-shot pool contains one example per solvent class, preferring rows where the guest also has a SMILES string and the host has not yet appeared in the pool; these examples are excluded from the evaluation set.

D.4 HOST-GUEST DESCRIPTION

The task has two complementary subtypes that share the same source. For the *forward* subtype, we group records by host, deduplicate guests by name, and keep only hosts that admit at least ten distinct guests with a known SMILES. For each remaining host, we compute molecular-weight, hydrogen-bond-donor, hydrogen-bond-acceptor, ring-count, and formal-charge descriptors of every guest via RDKit, then take the top 30% by $\log K_a$ as the high-affinity set. The reference answer aggregates this set into a single paragraph that reports the mean molecular weight, the dominant formal charge, and the average counts of ring, H-bond donor, and H-bond acceptor sites, followed by a list of five representative guest names. For the *reverse* subtype, we group records by guest, drop guests with fewer than five distinct hosts, and again take the top 30% by $\log K_a$. The reference answer lists the top five hosts with their measured $\log K_a$ values, reports the maximum observed $\log K_a$, and summarizes the structural families of those hosts, *e.g.*, cucurbituril, cyclodextrin, and calixarene, *etc.*

D.5 MOLECULAR IDENTIFICATION

Based on the cleaned interaction data, we additionally construct a multimodal molecular identification benchmark using crawled two-dimensional molecular structure images from SupraBank. In this task, the model predicts the molecular name or canonical SMILES from a single molecular structure image. To improve evaluation robustness, we aggregate aliases from multiple molecular metadata sources, including common names, abbreviations, and IUPAC names. Moreover, to ensure reliable alignment between molecular images and molecular annotations, all image records are linked via interaction-level identifiers rather than string-based name matching. The final multimodal benchmark contains 1,773 unique molecular images and corresponding gold SMILES.

	Model	Molecular Identification					
		Validity ↑	Canonical ↑	InChIKey ↑	Formula ↑	Tanimoto ↑	ΔHeavy ↓
Base	Qwen3.5-9B	0.805	0.295	0.390	0.445	0.696	1.549
	Qwen3.5-27B	0.738	0.384	0.489	0.529	0.814	**0.531**
	GPT-5.4-Nano	0.648	0.081	0.099	0.118	0.385	6.014
	Gemini-3-Flash	**0.945**	**0.584**	**0.764**	**0.801**	**0.914**	0.839
Few-Shot	Qwen3.5-9B	0.832	0.329	0.436	0.488	0.722	1.426
	Qwen3.5-27B	0.894	0.450	0.576	0.620	0.798	0.685
	GPT-5.4-Nano	0.653	0.106	0.116	0.140	0.449	4.076
	Gemini-3-Flash	**0.958**	**0.593**	**0.770**	**0.800**	**0.915**	**0.369**
CoT	Qwen3.5-9B	0.799	0.294	0.388	0.438	0.700	4.271
	Qwen3.5-27B	0.626	0.380	0.455	0.507	0.838	**0.378**
	GPT-5.4-Nano	0.669	0.084	0.098	0.115	0.376	6.325
	Gemini-3-Flash	**0.955**	**0.567**	**0.755**	**0.800**	**0.904**	0.526

Table 5: Performance comparison for molecular identification task. The best score is shown in **bold**, and the second-best is underlined.

E Text Corpus Construction Details

E.1 Text Collection

Anchor Corpus. We pull the Europe PMC abstract index through the public REST search endpoint, spanning publication years 1900 through 2026 and yielding approximately 40 million abstract records. In parallel, we mirror the bulk open-access full-text XML corpus, which covers approximately 8 million articles, together with the auxiliary PMID, PMCID, and DOI identifier-mapping tables that link abstract metadata to full-text bodies.

Supramolecular Topical Queries. To obtain the domain-specific text corpus, we leverage 19 topical Europe PMC search queries shared by domain experts that span the field's principal sub-areas, listed in Table 6. We deduplicate the union of query results by PMID, yielding a raw split of 420,950 unique articles. Since Europe PMC matches against MeSH headings and indexed keyword lists in addition to titles and abstracts, the raw split retains residual biomedical contamination, such as articles that mention *host cells* in immunology rather than *host molecules* in supramolecular chemistry, which motivates us for the following two-stage rule-based filtering.

Two-Stage Rule-Based Filter. We refine the raw split with a transparent filter built from three keyword banks listed in Table 6. The *positive bank* of approximately 440 supramolecular-specific keywords is organized into five tiers spanning core terms, named entities, and qualified self-assembly phrases. The *hard-reject bank* of approximately 140 biomedical keywords in the title conclusively indicates off-topic content regardless of context. The *conditional-reject bank* of approximately 50 keywords fires only when no supramolecular anchor is present in the title. Given these banks, each article is deterministically assigned to one of five clusters based on where positive or negative signals appear in its title and abstract: (A) title-positive, kept; (B) abstract-positive with title-neutral, kept; (C) hard-reject, dropped; (D) conditional-reject, dropped; (E) no signal in either field, excluded. The two kept clusters together yield a filtered split of 133,867 high-precision supramolecular articles.

E.2 Corpus Quality Validation

To quantify how much the two-stage filter actually concentrates supramolecular content, we employ an LLM-as-judge validation on randomly drawn samples from both splits. Specifically, we randomly draw 5,000 articles from the raw and filtered splits, present the title and abstract of each article to `Claude-Haiku-4.5`, and ask the model to classify the paper as relevant to the supramolecular chemistry domain on a strict centrality criterion: The paper is centrally about supramolecular chemistry only if its primary subject is non-covalent host–guest association, molecular recognition by macrocyclic hosts, non-covalent inclusion complexes, or supramolecular self-assembly. Table 7 reports the result.

Bank	Representative keywords
Topical queries (19)	supramolecular, host–guest chemistry, self-assembly, molecular recognition, cyclodextrin, crown ether, calixarene, cucurbituril, pillararene, rotaxane/catenane, dendrimer, metal-organic framework, non-covalent interaction, molecular cage, inclusion complex, supramolecular polymer, metallosupramolecular, macrocyclic chemistry, binding affinity.
Positive bank ($\sim$440)	supramolecular, host–guest, cucurbit[n]uril, calix[n]arene, pillar[n]arene, cryptand, cavitand, foldamer, mechanically interlocked, MOF, COF, β-CD, CB[7], MIL-53, MOF-5, HKUST-1, UiO-66, polyoxometalates, fullerenes, ionophores, . . .
Hard-reject bank ($\sim$140)	SARS-CoV-2, T-cell receptor, bacterial infection, RNA splicing, viral capsid, immune response, . . .
Conditional-reject bank ($\sim$50)	host cell, drug delivery, nanoparticle, liposome, protein-protein interaction, DNA binding, . . .

Table 6: Keyword banks used for the Europe PMC supramolecular corpus.

Split	YES ↑	BORDERLINE	NO ↓	Positive ↑
Raw	15.2%	3.1%	81.7%	18.3%
Filtered	62.9%	10.9%	26.2%	73.8%

Table 7: Corpus quality validation over 5,000 randomly sampled articles via `Claude-Haiku-4.5`.

F Implementation Details for Domain Adaptation

Training Corpus. Existing studies (French, 1999; Li & Lee, 2024) have shown that DAPT on a narrow domain in isolation tends to degrade general-purpose competence. To this end, we follow EvoLM (Qi et al., 2025) that construct an $80/15/5$ token-fraction mix, where 80% of the tokens come from the filtered supramolecular split we discussed in Section E.1, 15% from FineWeb-Edu (Lozhkov et al., 2024) for general-domain replay, and 5% from Tulu-3 SFT mixture (Lambert et al., 2024) flattened into question-answer pairs.

Training Recipe. We employ LoRA for DAPT, with parameters rank 32, alpha 64 and dropout rate 0.05 to every transformer block. Training runs for one epoch at a peak learning rate of 1×10^{-5} with a cosine schedule and 5% warmup in bf16, sharded with FSDP across four A100 GPUs. The per-device batch size is 1 with gradient accumulation 4, giving an effective batch size of 16 at a sequence length of 4096. We apply the identical recipe to both Qwen3.5-9B and Llama3.1-8B, holding the optimization budget fixed across bases so that downstream score differences in Table 3 can be attributed to the base model rather than to the adaptation procedure.

G Additional Experiments

G.1 Comparison with Specialized ML Baselines

To examine how LLMs stack up against existing methods, we evaluate two specialized molecular models, *i.e.*, ChemProp (Yang et al., 2019) and CGIB (Lee et al., 2023), on binding affinity prediction, top-binder selection, and solvent identification. As these are drug–target models, we adapt their input to the host–guest setting by replacing the protein encoder with a second molecule encoder, so each model reads host and guest as two molecules; top-binder selection is derived from the trained affinity predictor. Since SupraBench does not contain a dedicated training set for ML methods, we split each task into five folds via a host-disjoint split, *i.e.*, each test fold contains only hosts absent from the training folds, so a model is always evaluated on unseen hosts. Each LLM is scored on the same folds by reusing its per-example predictions to enable a fair comparison. Table 8 reports the results. The strongest LLM, *i.e.*, Gemini-3-Flash, is competitive with the specialized ML baselines: it trails only marginally on BAP RMSE and is on par or slightly ahead on TBS and SID, indicating that LLMs are

Method	BAP RMSE ↓	TBS ACC ↑	SID B. Acc ↑
Specialized ML baselines			
ChemProp (Yang et al., 2019)	$\mathbf{1.56 \pm 0.20}$	0.47 ± 0.04	0.69 ± 0.26
CGIB (Lee et al., 2023)	1.61 ± 0.23	0.45 ± 0.04	0.70 ± 0.26
LLMs (Base)			
Gemini-3-Flash	1.62 ± 0.14	$\mathbf{0.50 \pm 0.04}$	$\mathbf{0.74 \pm 0.20}$
DeepSeek-v4	2.04 ± 0.43	0.45 ± 0.02	0.56 ± 0.17
GPT-5.4-Mini	1.93 ± 0.44	0.42 ± 0.02	0.43 ± 0.16
GPT-5.4-Nano	2.09 ± 0.34	0.41 ± 0.03	0.45 ± 0.13
Qwen3.5-27B	2.47 ± 0.29	0.40 ± 0.03	0.50 ± 0.12

Table 8: Comparison with specialized ML baselines under five-fold host-disjoint cross-validation (mean ± std over folds). LLMs are scored on the same folds by reusing their per-example predictions.

Setting	Model	BAP MAE ↓	TBS Regret ↓	SID B. Acc ↑	HGD Recall ↑
Base	Qwen3.5-0.8B	3.301	1.529	0.167	0.369
	Qwen3.5-2B	7.503	1.476	0.166	0.408
	Qwen3.5-4B	4.232	1.474	**0.178**	**0.409**
	Llama3.2-3B	**2.647**	**1.454**	0.129	0.389
Few-Shot	Qwen3.5-0.8B	8.725	**1.428**	0.167	0.620
	Qwen3.5-2B	6.973	1.507	**0.173**	0.620
	Qwen3.5-4B	**2.276**	1.466	0.102	**0.659**
	Llama3.2-3B	7.023	1.482	0.167	0.651
CoT	Qwen3.5-0.8B	8.622	1.502	**0.167**	0.371
	Qwen3.5-2B	6.319	**1.481**	0.146	0.388
	Qwen3.5-4B	**2.225**	1.501	0.162	0.382
	Llama3.2-3B	4.075	1.488	0.149	**0.400**

Table 9: Performance of small language models (0.8B–4B) on SUPRABENCH. The best score within each prompting strategy is shown in bold.

genuinely useful here rather than lagging far behind specialized methods. Note that these absolute scores differ from Table 1 because the evaluation is restricted to the shared host-disjoint folds rather than the full benchmark.

G.2 PERFORMANCE OF SMALL LANGUAGE MODELS

To study the trade-off between model size and reasoning capability, we additionally evaluate four small language models, *i.e.*, Qwen3.5-{0.8B, 2B, 4B} and Llama3.2-3B, and report the results in Table 9. According to the table, we conclude that: (i) No small LLM comes close to the frontier models in Table 1. On binding affinity prediction, top-binder selection, and solvent identification, every sub-4B model is far worse than Gemini-3-Flash and DeepSeek-v4 (*e.g.*, BAP MAE ≥ 2.2 vs. 1.25, TBS Regret ≥ 1.43 vs. 0.65, and SID B. Acc ≤ 0.18 vs. 0.47), indicating that SUPRABENCH *requires chemistry capabilities that emerge at larger scales.* (ii) Few-Shot and CoT prompting do not recover small-model performance and often degrade it, validating that *the gap is a fundamental capacity limitation rather than a prompting or format artifact*, so in-context strategies cannot substitute for scale on supramolecular reasoning.

G.3 DAPT RECIPE ABLATION

Recipe ablation. To test whether the trade-off between scientific-text adaptation and strict letter-format output is an artifact of the specific recipe, we ablate the DAPT recipe on Qwen3.5-9B, independently varying the corpus mixing ratio and the adapter capacity. Specifically, we (a) double the instruction/format-anchor stream ($5\% \rightarrow 10\%$, +anchor), (b) increase the general-domain replay stream ($15\% \rightarrow 25\%$, +replay), and (c) double the LoRA rank ($32 \rightarrow 64$, +capacity), evaluating each variant under the identical protocol. Table 10 reports the results. We observe that: (i) At 9B,

Recipe	Mix (domain/replay/anchor)	LoRA rank	BAP MAE ↓	TBS ACC ↑	HGD F1 ↑
Standard	80/15/5	32	2.173	**0.235**	0.050
+anchor	80/10/10	32	2.198	**0.235**	0.030
+replay	70/25/5	32	**2.168**	0.234	**0.055**
+capacity	80/15/5	64	2.172	**0.235**	0.046

Table 10: DAPT recipe ablation on Qwen3.5-9B, varying the corpus mixing ratio and the adapter capacity.

Model	Rouge-1 R ↑	Rouge-1 P ↑	Rouge-1 F1 ↑	Judge ↑
Llama3.1-8B	0.266	0.059	0.092	2.5
GPT-5.4-Mini	0.437	0.086	0.137	**3.9**
Gemini-3-Flash	**0.506**	0.067	0.118	3.7
DeepSeek-v4	0.500	**0.090**	**0.141**	**3.9**

Table 11: Host–guest description under the Base prompt: Rouge-1 versus a chemistry-specific LLM judge (GPT-5.5, 1–5 scale).

TBS accuracy stays similarly degraded across every recipe, indicating that *the format-following drop is not an artifact of the specific* $80/15/5$ *mix or rank-32 adapter*, since no variant recovers it. (ii) Neither doubling the format-anchor stream nor increasing general-domain replay recovers TBS, and the mixing ratio instead only trades off open-ended generation quality (HGD), indicating that *the trade-off is orthogonal to the mix composition.* (iii) Doubling the adapter capacity leaves all tasks essentially unchanged, indicating that *the trade-off is not a low-capacity artifact.* In conclusion, the recipe ablation shows that the 9B format-following drop is not a tuning artifact, while the Qwen3.5-27B results in Section 4.4 show that it is not intrinsic to DAPT either: the cost shrinks sharply with scale, with TBS essentially preserved at 27B while BAP, HGD, and SID all improve. Together, this indicates that *the format-following cost is a small-model capacity limitation* that larger bases alleviate, rather than a general or recipe-specific property of DAPT.

G.4 Chemistry-Aware Evaluation for Host–Guest Description

Rouge-1 measures lexical overlap rather than chemical correctness, so a model could generate fluent but chemically incorrect descriptions and still receive a reasonable score. To directly assess chemical correctness, we conduct an additional chemistry-specific LLM-as-judge evaluation over the host–guest description task with an independent chemistry judge (GPT-5.5). For each (method, model, sample) triple, the judge only sees the question, the gold reference answer, the gold structured chemical facts, and the model answer, and rates the result from 1 (wrong) to 5 (fully correct). Table 11 reports the results. The judge preserves the Rouge-1 ranking, *i.e.*, both order the models as DeepSeek-v4 $\approx$ GPT-5.4-Mini $>$ Gemini-3-Flash $>$ Llama3.1-8B. This indicates that *Rouge-1 is a valid evaluation metric for the host–guest description task*, given its lower cost compared to LLM-as-judge.

G.5 Systematic Failure Analysis

To understand whether errors trace back to output parsing, missing supramolecular knowledge, or numerical estimation more broadly, we systematically categorize every failure mode on binding affinity prediction and find that:

(1) Parsing is not the bottleneck for frontier models; the cause of failure is abstention rather than malformed syntax. On the output side, parsing fails on only 0.0 to 0.2% of frontier-model completions (Gemini-3-Flash 0.0%, GPT-5.4-Mini 0.0%, DeepSeek-v4 0.2%), so neither is a meaningful error source at this scale. Where output parsing does fail, the dominant mechanism is explicit abstention, not format error. Specifically, for Llama3.1-8B (34.1% failures on BAP), 97% are outright refusals, such as "I can't answer that. I'm not capable of predicting binding affinities between specific host and guest molecules."

(2) The dominant failure is not missing chemistry but a compressed affinity scale. We aggregate each model's predictions within each host and compare them to that host's true mean value. Models

Model	Spearman ↑	Regression slope
Gemini-3-Flash	0.75	0.48
DeepSeek-v4	0.81	0.40
GPT-5.4-Mini	0.81	0.50
GPT-5.4-Nano	0.81	0.57
Llama3.1-70B	0.80	0.25
Ground truth	1.00	1.00

Table 12: Per-host calibration of binding affinity prediction (Base prompt). Spearman measures the rank correlation between predicted and true per-host mean $\log K_a$; the regression slope of predicted against true values quantifies compression of the affinity scale.

rank hosts correctly, *i.e.*, they recognize that Cucurbit[8]uril binds more strongly than β-cyclodextrin, while they compress the affinity scale by roughly a factor of two, under-predicting strong hosts and over-predicting weak ones. Table 12 quantifies this effect with the per-host Spearman rank correlation and the regression slope of predicted against true values; on the 87 samples with $\log K_a \geq 8$, mean predictions fall in the 5.5 to 6.6 range. The numerical estimation failure is therefore not arithmetic but a collapsed dynamic range, which validates, across every host in the benchmark, the miscalibration illustrated in Section 4.6. We conclude that the failures on SUPRABENCH are not superficial parsing artifacts but a genuine chemical limitation, *i.e.*, models recover the correct qualitative ordering yet lack a calibrated mapping to absolute $\log K_a$ magnitudes, which is precisely the capability our benchmark is designed to measure.

H SUPRABENCH EXAMPLES

This appendix presents one worked example per task, including the full prompt sent to the model and the model's response. We use Gemini-3-Flash under the Base prompting strategy as the running model. For CoT, we keep the question unchanged and add "Let's think step by step" to the system prompt. For Few-Shot, the demonstrations are fixed per task and excluded from the test set; Figures 12–16 show one Few-Shot example per task with the demonstrations embedded in the prompt. All prompt templates are also provided in our released source code. Figure 5 additionally shows one cleaned binding record from SUPRABENCH.

We further include in Figure 11 the full Base and Chain-of-Thought traces behind the failure case analyzed in Section 4.6, where Chain-of-Thought reasoning destabilizes an otherwise near-perfect binding affinity prediction.

SUPRABENCH Record-444913-637563

Host Identifiers	
Name	α-cyclodextrin
PubChem CID	444913
SMILES	`OC[C@H]1O[C@@H]2O...`
Guest Identifiers	
Name	*trans*-anethole
PubChem CID	637563
SMILES	`C/C=C/C1=CC=C(OC)C=C1`
Measurement	
Binding constant K_a	$710\ \mathrm{M}^{-1}$
$\log K_a$	2.85
Binding free energy ΔG	$-16.27\ \mathrm{kJ/mol}$
pH	7.0
Temperature	298 K

Images



Host: α-cyclodextrin **Guest:** *trans*-anethole

Figure 5: **A binding record example.**

Binding Affinity Prediction Example

Prompt: Predict the binding affinity ($\log K_a$) between the host molecule p-Sulfonatocalix[4]arene and the guest molecule Gadolinium(III) chloride, SMILES: Cl[Gd](Cl)Cl in aqueous solution at standard conditions. Provide a single numerical value. Put your final answer between `<answer></answer>`.
Reference $\log K_a$**:** 3.94

Gemini-3-Flash response: To predict the binding affinity ($\log K_a$) between p-sulfonatocalix[4]arene ($sC4$) and gadolinium(III) chloride ($GdCl_3$), we consider the supramolecular interactions in aqueous environment. p-Sulfonatocalix[4]arene is a macrocyclic host with a hydrophobic cavity capped with four sulfonic acid groups at the upper rim and four phenolic hydroxyl groups at the lower rim; in aqueous solution near neutral pH, the sulfonic acid groups are deprotonated and give the host a charge of -4. Gadolinium(III) is a trivalent lanthanide cation, and binding between $sC4$ and lanthanide ions is primarily driven by strong electrostatic interactions between the negatively charged sulfonated rim and the highly charged Gd^{3+} cation, supplemented by outer-sphere coordination to the sulfonic and phenolic oxygens. Experimental data for the binding of lanthanide(III) ions with $sC4$ typically show $\log K_a$ in the range of 3.5 to 5.0 depending on buffer and ionic strength.
`<answer>4.15</answer>`

Figure 6: Example for binding affinity prediction.

Top-Binder Selection Example

Prompt: Which of the following guest molecules has the strongest binding affinity (highest $\log K_a$) with the host molecule (1,1',4,4')-Tetraoxy-propylsulphite-bisnaftlen-methylen-bisglycouril Sodium in aqueous solution at standard conditions?

A. Azure A
B. N-Methylquinolinium iodide
C. Methylene Blue
D. sodium 3,3'-(1,3,6,8-tetraoxo-...-2,7-diyl)dipropionate

Put your final answer between `<answer></answer>`.
Reference: C (Methylene Blue, $\log K_a = 7.60$). Distractor affinities: A 6.77, B 3.90, D 3.79.

Gemini-3-Flash response: The host molecule, an acyclic naphthalene-walled cucurbituril analogue functionalized with sulfonate groups, possesses a hydrophobic cavity expanded by naphthalene walls and a strongly anionic rim that provides electrostatic complementarity to cationic guests. Azure A (A) and Methylene Blue (C) are cationic phenothiazine dyes that bind well to such hosts via combined hydrophobic and electrostatic interactions; N-methylquinolinium iodide (B) is a smaller cationic guest whose smaller surface area limits hydrophobic contact, and the sodium dipropionate-functionalized naphthalene diimide in (D) is an anion that experiences electrostatic repulsion. Methylene Blue in particular forms a rigid planar tricyclic system that fits perfectly between the naphthalene walls, yielding an exceptionally high $\log K_a$ exceeding that of the other three candidates.
`<answer>C</answer>`

Figure 7: Example for top-binder selection.

Host-Guest Description Forward Examp

Prompt: For the host molecule Cucurbit[8]uril, describe the key molecular properties of guest molecules that exhibit strong binding affinity (high $\log K_a$) in aqueous solution at standard conditions. Include: approximate molecular weight range, typical formal charge, number of ring systems, and any characteristic functional groups. Put your final answer between `<answer></answer>`.
Reference: High-affinity guests of Cucurbit[8]uril ($\log K_a \geq 6.2$) typically have a molecular weight of 358 ± 292 g/mol, are neutral (formal charge $+0$), and contain on average 3.2 ring systems, 2.3 H-bond donors, and 3.2 H-bond acceptors. Representative guests: L-phenylalanyl-L-valine, pentacyclo[7.3.1.1^{4,12}.0^{2,7}.0^{6,11}]tetradecane-4,9-diol, L-Tyrosyl-L-Phenylalanyl-L-alanine, 3,6-Diaminoacridin-10-ium, Tri-Trp Gly$_4$ target.

Gemini-3-Flash response: Cucurbit[8]uril (CB[8]) is a macrocyclic host composed of eight glycoluril units with a large hydrophobic cavity ($\sim$479 Å^3) and two carbonyl-fringed portals. Strong binders fall into two regimes: monomeric guests with molecular weights between 200 and 500 Da that are large enough to maximize van der Waals contact with the cavity, and ternary complexes in which two smaller guests (100–250 Da each) jointly fill the cavity. The carbonyl portals provide strong ion-dipole stabilization, so high-affinity guests are almost always cationic (formal charge $+1$ to $+2$, occasionally higher for polyamines). Guests typically contain two to four fused aromatic rings, with viologen and methylviologen scaffolds, naphthalenediimide, and dicationic aromatic dyes (*e.g.*, methylene blue, acridine) as archetypes. Characteristic functional groups include quaternary ammonium centers, aromatic π-rich cores, and amide carbonyls capable of hydrogen bonding to the portals.

Figure 8: Example for the forward subtype of host–guest description.

Host-Guest Description Reverse Example

Prompt: For the guest molecule 2-phenylethylamine hydrochloride (SMILES: NCCc1ccccc1.[H]Cl), describe the key molecular properties of host molecules that exhibit strong binding affinity (high $\log K_a$) in aqueous solution at standard conditions. Include: host structural type, approximate cavity size or geometry, typical charge, and any characteristic binding interactions. Put your final answer between <answer></answer>.
Reference: High-affinity hosts for 2-phenylethylamine hydrochloride ($\log K_a \geq 1.9$) include 5,11,17,23-tetrakis(dihydroxyphosphoryl)-25,26,27,28-tetrapropoxycalix[4]arene ($\log K_a = 2.2$) and a sulfonatobutoxy-substituted tris-oxazoline triethylbenzene host ($\log K_a = 1.9$). The highest recorded $\log K_a$ is 2.2. Favorable hosts are predominantly calixarene macrocycles (cone-shaped, phenol-derived).

Gemini-3-Flash response: 2-Phenylethylamine hydrochloride is amphiphilic, with a hydrophobic phenyl ring and a hydrophilic, positively charged $-NH_3^+$ group. Effective hosts must accommodate both moieties. The preeminent candidate is Cucurbit[7]uril (CB[7]), a macrocycle of seven glycoluril units forming a barrel-shaped cavity ($\sim$279 Å^3, $\sim$5–7 Ådiameter, $\sim$9.1 Åheight); the hydrophobic cavity encapsulates the phenyl ring with high packing density, the carbonyl-lined portals stabilize the ammonium by ion-dipole interactions, and CB[7]-arylalkylammonium complexes typically reach $\log K_a$ between 5 and 7. Cyclodextrins (especially β-CD) and sulfonated calixarenes are weaker but viable alternatives: β-CD provides a $\sim$6.0 Åcavity for hydrophobic inclusion of the phenyl ring, and p-sulfonatocalix[4]arene supplements the hydrophobic effect with cation-π interactions and electrostatic stabilization of the ammonium.

Figure 9: Example for the reverse subtype of host–guest description.

Solvent Identification Example

Prompt: Given a host–guest complex, predict which solvent environment is most appropriate for measuring its binding constant. The host molecule is provided as a SMILES string; analyze its structure (cavity size, functional groups, charge, hydrophobicity) to determine the solvent class. The candidate solvents are

A. Water
B. DMSO
C. MeCN
D. MeOH
E. $CHCl_3$
F. CH_2Cl_2

Host SMILES: *[skipped for simplicity]*. Put your final answer between <answer></answer>.
Reference: A (water).

Gemini-3-Flash response: The provided SMILES represents p-sulfonatocalix[4]arene. Its structure is a calix[4]arene scaffold of four phenolic rings linked by methylene bridges that defines a hydrophobic cavity; the upper rim is substituted with four sulfonic acid groups ($-SO_3H$) that are typically deprotonated to sulfonates ($-SO_3^-$) in solution. The four sulfonate groups make the molecule highly anionic and extremely soluble in polar, aqueous environments. p-Sulfonatocalix[4]arenes are archetypal water-soluble hosts: they drive guest inclusion through the hydrophobic effect into the cavity, while the sulfonate rim provides water solubility and electrostatic stabilization of cationic guests. By the operating-medium rule for sulfonated macrocycles with polar portals, the appropriate solvent is water.
<answer>A. water</answer>

Figure 10: Example for solvent identification.

Chain-of-Thought Failure on Binding Affinity Prediction (DeepSeek-v4)

Prompt: Predict the binding affinity ($\log K_a$) between the host molecule 4,7,13,18-Tetraoxa-1,10-diazabicyclo[8.5.5]icosane (SMILES: `C1COCCN2CCOCCN1CCOCCOCC2`) and the guest molecule Barium cation (SMILES: `[Ba+2]`) in aqueous solution at standard conditions. Provide a single numerical value. Put your final answer between `<answer></answer>`.
Reference $\log K_a$**:** 2.00

Base response: `<answer>2.1</answer>` **(correct)**

CoT response: The host molecule is a diaza-crown ether derivative, specifically a bicyclic structure with two nitrogen atoms and four ether oxygen atoms. This type of ligand is known to bind strongly to alkali and alkaline earth metal cations, particularly Ba^{2+}, due to its preorganized cavity and favorable ion–dipole interactions. In aqueous solution, the binding constant for Ba^{2+} with such bicyclic diaza-crown ethers typically falls in the range of $\log K_a \approx 10$–12, depending on ring size and solvation effects. Given the structure (a [8.5.5] system), it is a relatively good fit for Ba^{2+}. Considering standard aqueous conditions and typical literature values for similar cryptands, a reasonable estimate is:
`<answer>11.0</answer>` **(off by nine orders of magnitude in K_a)**

Figure 11: Case study of a Chain-of-Thought regression. On the same host–guest pair, the Base prompt returns a near-perfect $\log K_a$, while CoT articulates fluent but uncalibrated chemistry and fabricates a "typical literature" range that does not exist, landing nine orders of magnitude away in the association constant. See Section 4.6.

Binding Affinity Prediction (Few-Shot) Example

Prompt: You are an expert in supramolecular chemistry. Your task is to answer the following question. Question: Predict the binding affinity (logKa) between the host molecule p-Sulfonatocalix[4]arene (SMILES: *[skipped for simplicity]*) and the guest molecule Gadolinium(III) chloride (SMILES: *[skipped for simplicity]*) in aqueous solution at standard conditions. Provide a single numerical value. Below are some examples that you should follow. Question: Predict the binding affinity (logKa) between the host molecule α-Cyclodextrin (SMILES: *[skipped for simplicity]*) and the guest molecule 1-adamantane carboxylate (SMILES: *[skipped for simplicity]*) in aqueous solution at standard conditions. Provide a single numerical value. `<answer>-3.5</answer>` Question: Predict the binding affinity (logKa) between the host molecule β-Cyclodextrin (SMILES: *[skipped for simplicity]*) and the guest molecule (1S,2D,5S)-(-)-2-Hydroxy-3-pinanone (SMILES: *[skipped for simplicity]*) in aqueous solution at standard conditions. Provide a single numerical value. `<answer>3.4</answer>` Question: Predict the binding affinity (logKa) between the host molecule Cucurbit[8]uril (SMILES: *[skipped for simplicity]*) and the guest molecule 3,6-Diaminoacridin-10-ium (SMILES: *[skipped for simplicity]*) in aqueous solution at standard conditions. Provide a single numerical value. `<answer>15.3</answer>` Put your final answer between `<answer></answer>`
Reference $\log K_a$**:** 3.94

Gemini-3-Flash response: The interaction between p-sulfonatocalix[4]arene and lanthanide ions like Gadolinium (Gd^{3+}) is driven by the strong electrostatic attraction between the negatively charged sulfonate groups at the upper rim (and the phenolic oxygens at the lower rim) and the highly positive trivalent cation. p-Sulfonatocalix[4]arene provides a well-defined cavity and multiple anionic sites that facilitate the formation of stable complexes with lanthanide aqua-ions in aqueous solution. Literature values for the stability constants ($\log K_a$ or $\log \beta_1$) of p-sulfonatocalix[4]arene with trivalent lanthanide ions (including Gd^{3+}) typically fall in the range of 3.0 to 4.5, depending on the specific pH and ionic strength. For Gd^{3+} specifically, the coordination to the sulfonate rim or the inclusion-like behavior leading to $1:1$ stoichiometry results in a binding affinity of approximately 3.4. `<answer>3.4</answer>`

Figure 12: Few-shot example for binding affinity prediction.

Top-Binder Selection (Few-Shot) Example

Prompt: You are an expert in supramolecular chemistry. Your task is to answer the following question. Question: Which of the following guest molecules has the strongest binding affinity (highest logKa) with the host molecule (1,1',4,4')-Tetraoxy-propylsulphite-bisnaftlen-methylen-bisglycouril Sodium in aqueous solution at standard conditions? A. Spermine (fully protonated form) (SMILES: *[skipped for simplicity]*). B. Azure A (SMILES: *[skipped for simplicity]*). C. N-Methylacridinium iodide (SMILES: *[skipped for simplicity]*). D. 1,4-Benzodiazine. Below are some examples that you should follow. Question: Which of the following guest molecules has the strongest binding affinity (highest logKa) with the host molecule (1,1',4,4')-Tetraoxy-propylsulphite-bisnaftlen-methylen-bisglycouril Sodium in aqueous solution at standard conditions? A. Thioflavin T (SMILES: *[skipped for simplicity]*). B. hexane-1,6-diaminium chloride (SMILES: *[skipped for simplicity]*). C. 1,3-phenylenedimethanaminium chloride (SMILES: *[skipped for simplicity]*). D. Methylene Blue (SMILES: *[skipped for simplicity]*). `<answer>D</answer>` Question: Which of the following guest molecules has the strongest binding affinity (highest logKa) with the host molecule 1,3,5-tris[[(4s)-4-[3-[4-(sodiosulfo)butoxy]phenyl]-2-oxazoline-2-yl]methyl]-2,4,6-triethylbenzene in aqueous solution at standard conditions? A. 4-aminobutyric acid (SMILES: *[skipped for simplicity]*). B. 2-aminoethanol hydrochloride (SMILES: *[skipped for simplicity]*). C. Epinephrine hydrochloride (SMILES: *[skipped for simplicity]*). D. Dopamine hydrochloride (SMILES: *[skipped for simplicity]*). `<answer>D</answer>` Ques-

tion: Which of the following guest molecules has the strongest binding affinity (highest logKa) with the host molecule 1,3,5-tris[[(4s)-4-[3-[4-(sodiosulfo)butoxy]phenyl]-2-oxazoline-2-yl]methyl]-2,4,6-trimethylbenzene in aqueous solution at standard conditions? A. Tyramine hydrochloride (SMILES: *[skipped for simplicity]*). B. Epinephrine hydrochloride (SMILES: *[skipped for simplicity]*). C. Noradrenaline hydrochloride (SMILES: *[skipped for simplicity]*). D. 4-aminobutyric acid (SMILES: *[skipped for simplicity]*). `<answer>A</answer>` Put your final answer between `<answer></answer>`
Reference: B (Azure A).

Gemini-3-Flash response: The host molecule described, (1,1',4,4')-Tetraoxy-propylsulphite-bisnaftlen-methylen-bisglycouril Sodium, is a specific member of the **Cucurbit[n]uril (CB[n])** family analogues, specifically a **naphthalene-walled acyclic cucurbituril** (often referred to as a "molecular clip" or "molecular container" like the ones developed by the Isaacs group). These hosts are characterized by: 1. A hydrophobic cavity defined by the aromatic (naphthalene) sidewalls. 2. Negative charge provided by the sulfonate (propylsulphite) groups, which aids in water solubility and provides strong electrostatic attraction for cationic guests. Evaluating the guests: * **A. Spermine (fully protonated):** A tetra-cationic aliphatic amine. While it has strong electrostatic interactions, it lacks the aromatic surface area to maximize π-stacking and hydrophobic interactions with the naphthalene walls compared to large dyes. * **B. Azure A:** A tricyclic aromatic cationic dye. It has a high surface area for π-stacking and a positive charge. * **C. N-Methylacridinium iodide:** A planar, tricyclic cationic aromatic molecule. * **D. 1,4-Benzodiazine (Quinoxaline):** Neutral and relatively small; it bind much more weakly than cationic guests in these negatively charged hosts. Historically and experimentally, for this specific class of acyclic naphthalene-walled glycouril hosts (like the "Calix[4]arene-like" acyclic cucurbiturils), large tricyclic cationic dyes like **Azure A** and **Methylene Blue** exhibit exceptionally high binding constants (often $\log K_a > 7 - 8$) due to the high complementarity between the naphthalene sidewalls and the extended π-system of the dyes, combined with strong ion-ion interactions. Between B and C, Azure A presents more opportunities for hydrogen bonding and has a more extended conjugation suitable for these specific molecular "clips." `<answer>B</answer>`
`<answer>B</answer>`

Figure 13: Few-shot example for top-binder selection.

Host-Guest Description Forward (Few-Shot) Example

Prompt: You are an expert in supramolecular chemistry. Your task is to answer the following question. Question: For the host molecule Cucurbit[8]uril (SMILES: *[skipped for simplicity]*), describe the key molecular properties of guest molecules that exhibit strong binding affinity (high logKa) in aqueous solution at standard conditions. Include: approximate molecular weight range, typical formal charge, number of ring systems, and any characteristic functional groups. Below are some examples that you should follow. Question: For the host molecule (1,1',4,4')-Tetraoxy-propylsulphite-bisnaftlen-methylen-bisglycouril Sodium, describe the key molecular properties of guest molecules that exhibit strong binding affinity (high logKa) in aqueous solution at standard conditions. Include: approximate molecular weight range, typical formal charge, number of ring systems, and any characteristic functional groups. `<answer>High-affinity guests of (1,1',4,4')-Tetraoxy-propylsulphite-bisnaftlen-methylen-bisglycouril Sodium (logKa` $\geq$ `5.3) typically have a molecular weight of 318±99 g/mol, are neutral (formal charge +0), contain on average 3.4 ring system(s), 1.2 H-bond donor(s), and 2.0 H-bond acceptor(s). Representative guests: N,N'-bis(2-dimethylammonioethyl)-Naphthalene diimide, Methyl viologen, N-Methylacridinium iodide, Methylene Blue, Neutral Red protonated.</answer>`
Question: For the host molecule 4-sulfocalix[4]arene sodium salt (SMILES: *[skipped for simplicity]*), describe the key molecular properties of guest molecules that exhibit strong binding affinity (high logKa) in aqueous solution at standard conditions. Include: approximate

molecular weight range, typical formal charge, number of ring systems, and any characteristic functional groups. <answer>High-affinity guests of 4-sulfocalix[4]arene sodium salt (logKa $\geq$ 4.4) typically have a molecular weight of 400±364 g/mol, are neutral (formal charge +0), contain on average 2.4 ring system(s), 3.0 H-bond donor(s), and 3.1 H-bond acceptor(s). Representative guests: N,N'-dimethyl-1,4-diazabicyclo[2.2.2]octane, 1,3-Dimethylimidazolium, N-methyl pyridinium, Lucigenin, methyl viologen diiodide.</answer> Question: For the host molecule 5,11,17,23,29-pentaguanidinium-31,32,33,34,35-penta (4-methylpentloxy) calix[5]arene, describe the key molecular properties of guest molecules that exhibit strong binding affinity (high logKa) in aqueous solution at standard conditions. Include: approximate molecular weight range, typical formal charge, number of ring systems, and any characteristic functional groups. <answer>High-affinity guests of 5,11,17,23,29-pentaguanidinium-31,32,33,34,35-penta (4-methylpentloxy) calix[5]arene (logKa $\geq$ 7.1) typically have a molecular weight of 486±376 g/mol, are negatively charged (formal charge -4), contain on average 1.0 ring system(s), 0.5 H-bond donor(s), and 6.2 H-bond acceptor(s). Representative guests: clodronate, etidronate, lysophosphatidic acid, rose bengal.</answer> Question: For the host molecule Calix[4]arene (SMILES: *[skipped for simplicity]*), describe the key molecular properties of guest molecules that exhibit strong binding affinity (high logKa) in aqueous solution at standard conditions. Include: approximate molecular weight range, typical formal charge, number of ring systems, and any characteristic functional groups. <answer>High-affinity guests of Calix[4]arene (logKa $\geq$ 4.5) typically have a molecular weight of 149±49 g/mol, are neutral (formal charge +0), contain on average 0.1 ring system(s), 1.4 H-bond donor(s), and 1.5 H-bond acceptor(s). Representative guests: 1,4-Butanediammonium, N6-Dimethyllysine, Pentane-1,5-diamine, Lanthanum cation, Tetraethylammonium iodide.</answer> Question: For the host molecule Cucurbit[5]uril (SMILES: *[skipped for simplicity]*), describe the key molecular properties of guest molecules that exhibit strong binding affinity (high logKa) in aqueous solution at standard conditions. Include: approximate molecular weight range, typical formal charge, number of ring systems, and any characteristic functional groups. <answer>High-affinity guests of Cucurbit[5]uril (logKa $\geq$ 4.0) typically have a molecular weight of 312±83 g/mol, are neutral (formal charge +0), contain on average 2.4 ring system(s), 0.0 H-bond donor(s), and 5.8 H-bond acceptor(s). Representative guests: Sanguinarine, Lanthanum(III) nitrate hexahydrate, Coptisine, Barium nitrate, Strontium nitrate.</answer> Question: For the host molecule Cucurbit[6]uril (SMILES: *[skipped for simplicity]*), describe the key molecular properties of guest molecules that exhibit strong binding affinity (high logKa) in aqueous solution at standard conditions. Include: approximate molecular weight range, typical formal charge, number of ring systems, and any characteristic functional groups. <answer>High-affinity guests of Cucurbit[6]uril (logKa $\geq$ 4.0) typically have a molecular weight of 168±91 g/mol, are neutral (formal charge +0), contain on average 0.9 ring system(s), 0.6 H-bond donor(s), and 0.6 H-bond acceptor(s). Representative guests: 1-dodecyl-3-methylimidazol-3-ium bromide, cyclopropylmethanaminium, 4-(2-pyridyl)aniline, Iron(III) chloride, 1-heptyl-3-methylimidazol-3-ium bromide.</answer> Put your final answer between <answer></answer>

Reference: High-affinity guests of Cucurbit[8]uril (logKa $\geq$ 6.2) typically have a molecular weight of 358±292 g/mol, are neutral (formal charge +0), contain on average 3.2 ring system(s), 2.3 H-bond donor(s), and 3.2 H-bond acceptor(s). Representative guests: L-phenylalanyl-L-valine, pentacyclo[7.3.1.14,12.02,7.06,11]tetradecane-4,9-diol, L-Tyrosyl-L-Phenylalanyl-L-alanine, 3,6-Diaminoacridin-10-ium, Tri-Trp Gly4 target.

Gemini-3-Flash response: <answer>High-affinity guests of Cucurbit[8]uril (logKa $\geq$ 5.0) typically have a molecular weight of 285±145 g/mol, are positively charged (formal charge +1 or +2), contain on average 2.5 ring system(s), 0.8 H-bond donor(s), and 1.5 H-bond acceptor(s). Cucurbit[8]uril

is unique for its large cavity, often accommodating two aromatic guests (ternary complexes) or large dicationic species. Representative guests: Methyl viologen, Naphthalene derivatives, Tryptophan, Acridine Orange, and various Ru(II) or Pt(II) polypyridyl complexes.</answer>

Figure 14: Few-shot example for the forward subtype of host–guest description.

Host-Guest Description Reverse (Few-Shot) Example

Prompt: You are an expert in supramolecular chemistry. Your task is to answer the following question. Question: For the guest molecule 2-phenylethylamine hydrochloride (SMILES: *[skipped for simplicity]*), describe the key molecular properties of host molecules that exhibit strong binding affinity (high logKa) in aqueous solution at standard conditions. Include: host structural type, approximate cavity size or geometry, typical charge, and any characteristic binding interactions. Below are some examples that you should follow. Question: For the guest molecule 1,2-Dihydroxybenzene (SMILES: *[skipped for simplicity]*), describe the key molecular properties of host molecules that exhibit strong binding affinity (high logKa) in aqueous solution at standard conditions. Include: host structural type, approximate cavity size or geometry, typical charge, and any characteristic binding interactions. <answer>High-affinity hosts for 1,2-Dihydroxybenzene (logKa $\geq$ 3.3) include: Tetrahydropyridazinophthalazine-tetraone functionalized Cucurbit[6]uril (logKa=5.5), Cucurbit[8]uril (logKa=4.2), Cyclobis(paraquat-1,4-phenylene) tetrakis(hexafluorophosphate) (logKa=3.6), (2-(((anthracen-9-ylmethyl)(methyl)amino)methyl)phenyl)boronic acid (logKa=3.3). The highest recorded logKa is 5.5. Favorable hosts are predominantly cucurbituril macrocycle (barrel-shaped, hydrophobic cavity, carbonyl-lined portals); macrocyclic host molecule.</answer> Question: For the guest molecule 1,4-phenylenedimethanaminium chloride (SMILES: *[skipped for simplicity]*), describe the key molecular properties of host molecules that exhibit strong binding affinity (high logKa) in aqueous solution at standard conditions. Include: host structural type, approximate cavity size or geometry, typical charge, and any characteristic binding interactions. <answer>High-affinity hosts for 1,4-phenylenedimethanaminium chloride (logKa $\geq$ 7.3) include: sodium sulfonated pillar[6]arene (logKa=8.3), carboxylate pillar[6]arene (logKa=7.8). The highest recorded logKa is 8.3. Favorable hosts are predominantly pillar[n]arene macrocycle (cylindrical, electron-rich cavity).</answer> Question: For the guest molecule 1-Adamantanamine (SMILES: *[skipped for simplicity]*), describe the key molecular properties of host molecules that exhibit strong binding affinity (high logKa) in aqueous solution at standard conditions. Include: host structural type, approximate cavity size or geometry, typical charge, and any characteristic binding interactions. <answer>High-affinity hosts for 1-Adamantanamine (logKa $\geq$ 7.9) include: Cucurbit[8]uril (logKa=8.9), cucurbit[7]uril-tetraethylene glycol-berberine conjugate (logKa=8.0). The highest recorded logKa is 8.9. Favorable hosts are predominantly cucurbituril macrocycle (barrel-shaped, hydrophobic cavity, carbonyl-lined portals).</answer> Question: For the guest molecule 1-Adamantylammonium chloride (SMILES: *[skipped for simplicity]*), describe the key molecular properties of host molecules that exhibit strong binding affinity (high logKa) in aqueous solution at standard conditions. Include: host structural type, approximate cavity size or geometry, typical charge, and any characteristic binding interactions. <answer>High-affinity hosts for 1-Adamantylammonium chloride (logKa $\geq$ 6.7) include: Monohydroxycucurbit[7]uril (logKa=11.0), carboxylate pillar[6]arene (logKa=7.3). The highest recorded logKa is 11.0. Favorable hosts are predominantly cucurbituril macrocycle (barrel-shaped, hydrophobic cavity, carbonyl-lined portals); pillar[n]arene macrocycle (cylindrical, electron-rich cavity).</answer> Question: For the guest molecule 2,3-Diazabicyclo(2.2.2)oct-2-ene (SMILES: *[skipped*

for simplicity]), describe the key molecular properties of host molecules that exhibit strong binding affinity (high logKa) in aqueous solution at standard conditions. Include: host structural type, approximate cavity size or geometry, typical charge, and any characteristic binding interactions. <answer>High-affinity hosts for 2,3-Diazabicyclo(2.2.2)oct-2-ene (logKa $\geq$ 3.0) include: Cucurbit[8]uril (logKa=4.2), tetrasulfonato-tetrakis(n-butyl)-calix[4]arene (logKa=3.0), β-Cyclodextrin (logKa=3.0). The highest recorded logKa is 4.2. Favorable hosts are predominantly cucurbituril macrocycle (barrel-shaped, hydrophobic cavity, carbonyl-lined portals); calixarene macrocycle (cone-shaped, phenol-derived); cyclodextrin oligosaccharide (truncated-cone shape, hydrophobic interior).</answer> Question: For the guest molecule 2,3-Diazabicyclo[2.2.1]-hept-2-ene (SMILES: *[skipped for simplicity]*), describe the key molecular properties of host molecules that exhibit strong binding affinity (high logKa) in aqueous solution at standard conditions. Include: host structural type, approximate cavity size or geometry, typical charge, and any characteristic binding interactions. <answer>High-affinity hosts for 2,3-Diazabicyclo[2.2.1]-hept-2-ene (logKa $\geq$ 2.9) include: Cucurbit[8]uril (logKa=3.4), Cucurbit[6]uril (logKa=3.1), tetrasulfonato-tetrakis(n-butyl)-calix[4]arene (logKa=2.9). The highest recorded logKa is 3.4. Favorable hosts are predominantly cucurbituril macrocycle (barrel-shaped, hydrophobic cavity, carbonyl-lined portals); calixarene macrocycle (cone-shaped, phenol-derived).</answer> Put your final answer between <answer></answer>

Reference: High-affinity hosts for 2-phenylethylamine hydrochloride (logKa $\geq$ 1.9) include: 5,11,17,23-Tetrakis(dihydroxyphosphoryl)-25,26, 27,28-tetrapropoxycalix[4]arene (logKa=2.2), 1,3,5-tris[[(4s)-4-[3-[4-(sodiosulfo)butoxy]phenyl]-2-oxazoline-2-yl]methyl]-2,4,6-triethylbenzene (logKa=1.9). The highest recorded logKa is 2.2. Favorable hosts are predominantly calixarene macrocycle (cone-shaped, phenol-derived); macrocyclic host molecule.

Gemini-3-Flash response: <answer>High-affinity hosts for 2-phenylethylamine hydrochloride (logKa $\geq$ 4.0) include: Cucurbit[7]uril (logKa=6.1), Cucurbit[8]uril (logKa=5.1), and various sulfonated calixarenes such as p-sulfonatocalix[4]arene (logKa=4.1). The highest recorded logKa is 6.1. Favorable hosts are predominantly cucurbituril macrocycles (barrel-shaped, hydrophobic cavity, carbonyl-lined portals) and calixarene macrocycles (cone-shaped, often negatively charged with sulfonate groups). Binding is driven by the hydrophobic effect as the phenethyl group resides in the cavity, and strong cation-dipole or electrostatic interactions between the guest's ammonium group and the host's portals (carbonyl oxygens in cucurbiturils or sulfonate groups in calixarenes).</answer>

Figure 15: Few-shot example for the reverse subtype of host–guest description.

Solvent Identification (Few-Shot) Example

Prompt: You are an expert in supramolecular chemistry. Your task is to answer the following question. Question: Given a host–guest complex, predict which solvent environment is most appropriate for measuring its binding constant. You are given the host molecule as a SMILES string. Analyze its structure (cavity size, functional groups, charge, hydrophobicity) to determine the solvent class. General principles: - Large hydrophobic cavities with polar portals (e.g. glycoluril-based, sulfonated) $\rightarrow$ water - Neutral macrocycles with aromatic walls, no charged groups $\rightarrow$ CHCl3 or CH2Cl2 - Polyether / aza-crown scaffolds $\rightarrow$ MeOH or MeCN - Hydrogen-bond donors/acceptors without water solubility $\rightarrow$ DMSO Host SMILES: *[skipped for simplicity]*)

Reference: A (water).

Gemini-3-Flash response: The host molecule provided is a sulfonated calix[4]arene (specifically, p-sulfonatocalix[4]arene). Analyzing its structure: 1. **Macrocycle**: It is a calixarene, which forms a bucket-shaped cavity. 2. **Functional Groups**: It contains four phenolic hydroxyl groups (-OH) at the "lower rim" and four sulfonic acid groups (-S(=O)(=O)O) at the "upper rim." 3. **Charge/Polarity**: The sulfonic acid groups are highly polar and typically deprotonated in solution, making the molecule polyanionic and highly water-soluble. 4. **Cavity**: It has a hydrophobic aromatic cavity capable of binding guests, but the presence of the sulfonated groups (polar portals) strongly indicates its use in aqueous environments. 5. **Guest**: The guest is Gadolinium(III) chloride (GdCl3), a charged inorganic salt. According to the general principles: - Large hydrophobic cavities with polar portals (e.g., glycoluril-based, sulfonated) $\rightarrow$ water. - Sulfonated calixarenes are classic water-soluble hosts used to bind various guests (including metal ions like Gd^{3+}) in aqueous solutions. Therefore, the most appropriate solvent for measuring the binding constant of this host–guest complex is water. <answer>A</answer>
<answer>A. water</answer>

Figure 16: Few-shot example for solvent identification.