MedChemBench: Evaluating complex multimodal reasoning in medicinal chemistry

Jeremy Li[1,†], Andrew Ho[1], Suyash Shringarpure[1,†],
Amelia Liu[1], Paul Humphries[1], Joy Jiao[1]
[1]OpenAI

November 9, 2026

Abstract

Medicinal chemistry involves the design, synthesis, and optimization of chemical compounds to create safe and effective therapeutics, and serves as a cornerstone of modern drug development. It is naturally an interdisciplinary field, involving the synthesis of the biological, chemical, computational, translational, and medical sciences. While artificial intelligence systems have recently begun to accelerate certain components of medicinal chemistry, it remains a primary bottleneck in the modern drug discovery process. Here we introduce MedChemBench (MCB), a retrospective multimodal benchmark comprising 99 medicinal chemistry problems based on cases reported in recent papers and patents. MCB evaluates how models combine molecular and experimental evidence to reason through decisions encountered in medicinal chemistry programs. Generally, each task presents the model with a scientific narrative describing the biological background and context of a drug development task, a description of initial candidate compounds alongside molecular structures, followed by a description of follow-up experiments and their results (*e.g.*, alternative molecular structures, tables of assay results, plots, protein-ligand views), and a task to be completed. MCB tasks evaluate models' abilities to reconstruct and assemble complex molecular structures, reason about stereochemistry and molecular conformation, integrate multimodal evidence, perform mechanistic inference and quantitative prediction, and make decisions under multiple competing constraints.

MCB utilizes a combination of deterministic grading and rubric-based grading to evaluate precise molecular graph reconstruction, numerical prediction, as well as scientific reasoning. The highest mean score was 83.83%, achieved by GPT-6 Astra at `max` effort. At `xhigh` effort, a setting evaluated for all eleven models, scores ranged from 35.02% to 83.33%. MCB therefore measures an emerging and promising capability of frontier models to assist in the drug development process.

[†]Corresponding authors: `jeremiah.li@c-openai.com`; `suyash@openai.com`.

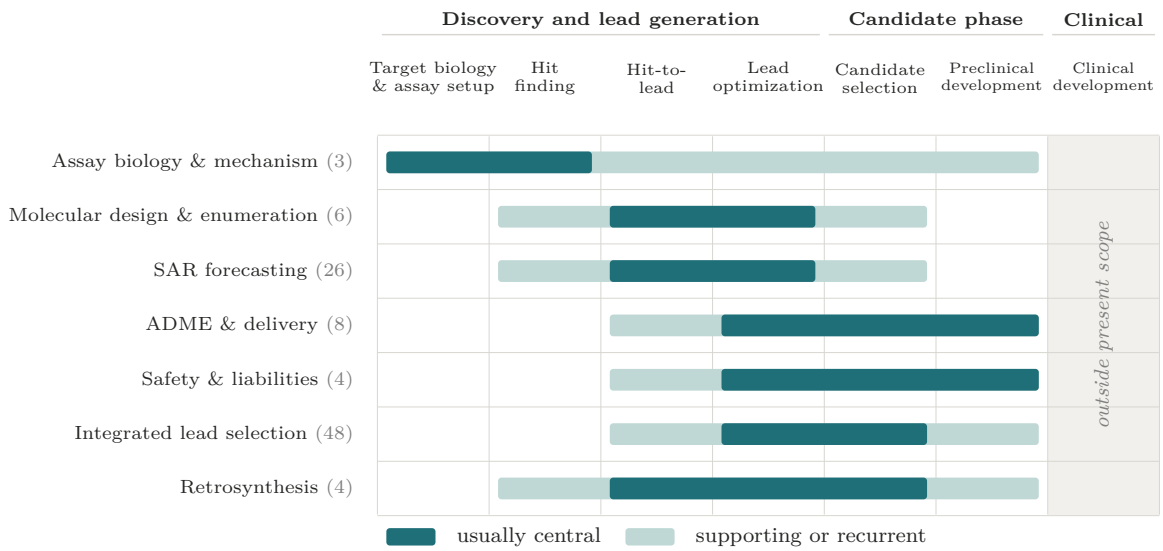


Figure 1: Rough illustration of the scope of the seven MedChemBench task categories in the drug discovery pipeline. Dark bands mark stages where a question is usually central to a program decision supported by medicinal-chemistry evidence; pale extensions mark stages where it commonly recurs or supplies supporting evidence.

Introduction

In the typical drug development process, identifying a promising therapeutic target is only one step toward developing a suitable drug candidate. Human genetics and disease biology can support the choice of target and the desired intervention [1], but do not specify a molecule with the properties needed to achieve that intervention safely and effectively. Drug development often takes more than a decade, and most programs do not lead to an approved drug [2, 3].

Within the broader drug discovery process, medicinal chemistry spans hit identification, hit-to-lead development, and lead optimization (Figure 1) [4, 5]. Initial active compounds, or hits, are identified through screening or other discovery approaches and evaluated to confirm their activity. During hit-to-lead development, medicinal chemists work with biologists and other specialists to explore promising chemical series, establish relationships between molecular structure and biological activity, and identify leads for further optimization [6]. Lead optimization proceeds through repeated cycles of design, synthesis, testing, and analysis. Evidence from biochemical and cellular assays, structural studies, and measurements of drug metabolism and pharmacokinetics guides subsequent molecular changes and decisions about which compounds to advance.

A central challenge in medicinal chemistry is multiparameter optimization: balancing potency, selectivity, exposure, safety, and synthetic feasibility to achieve a compound profile suited to the intended therapeutic application. Potency, which describes how much of a compound is required to produce an effect, can trade off against selectivity, the relative potency of a compound at its intended target compared with relevant off-targets. A useful molecule also often requires cell type-specific activity, appropriate solubility, permeability, and metabolic stability, while remaining practical to synthesize. Because structural changes can affect several properties simultaneously, improving one may compromise another: for example, improving membrane passage can reduce solubility or increase metabolism. Progress therefore involves program-level decisions made based on incomplete evidence and competing goals [7, 8].

Interpreting this evidence requires attention to experimental and biological context. For example, the concentration required to produce 50% inhibition (IC_{50}) depends on assay conditions, and activity against an isolated protein may not translate into activity in relevant cells. Efflux can limit intracellular exposure even when a compound readily crosses cell membranes. Compounds may also interfere with assays through aggregation, chemical reactivity, metal binding, optical effects, or membrane damage [9]. A protein–ligand binding pose can help explain molecular interactions but does not by itself establish potency or exposure. Medicinal chemists must therefore interpret measurements in context and account for uncertainty when making compound-design and selection decisions [10, 11].

These demands also shape what AI systems must be able to do to support medicinal chemistry. Helping researchers select compounds or propose molecular changes requires integrating experimental evidence with biological context, weighing competing objectives, and accounting for uncertainty [12]. Existing benchmarks evaluate several capabilities relevant to this role. MoleculeNet and Therapeutics Data Commons standardize molecular and biomedical prediction tasks [13, 14], while MoleculeACE and Practical Molecular Optimization assess activity cliffs and molecular optimization under fixed scoring functions [15, 16]. Chemistry benchmarks evaluate chemical knowledge, reaction reasoning, and visual interpretation across molecular structures, tables, spectra, and scientific documents [17–21]. More recent drug discovery evaluations address compound prioritization, molecular design, and drug metabolism and pharmacokinetic reasoning [22–27]. LifeSciBench, LABBench2, GeneBench, and GeneBench-Pro extend evaluation toward realistic research tasks and multistage scientific workflows [28–31]. Building on these efforts, we focus on decisions within medicinal chemistry programs that require models to combine molecular and experimental evidence while balancing competing compound properties.

Here we introduce MedChemBench (MCB), a new 99-problem benchmark for AI models with a retrospective setup based on 117 recently published scientific papers and patent publications, spanning 761 distinct graded molecular structures. Each problem presents biological context, molecular structures, and experimental evidence leading up to a selected decision point in a medicinal chemistry program. Models use this evidence to complete a task, such as selecting which compound to advance from a set of candidate leads. Results reported later in the source publication are withheld from the prompt and used to inform reference answers and grading criteria.

MCB covers seven task categories, which Figure 1 maps to stages of the drug discovery process. We begin by discussing the benchmark construction process and scope, the types of tasks it contains and how they are graded, before showing two representative problems in detail. We then discuss results from eleven GPT models across 61 combinations of model and reasoning effort. The highest mean score was 83.83%, achieved by GPT-6 Astra at `max` effort. At the `xhigh` setting evaluated for all eleven models, scores ranged from 35.02% to 83.33%.

Benchmark construction and scope

MedChemBench comprises 99 tasks drawn from 117 recently published papers and patents and spans 761 distinct molecular structures. We selected sources that described a clear medicinal chemistry question and provided enough evidence to reconstruct a compound series or design problem. For each task, we identified a decision point in the source narrative and presented the relevant context and evidence in the prompt. Findings reported later in the source were withheld and used to inform reference answers and grading criteria.

Category	Tasks	Practical question	Evidence used
Integrated lead selection	48	Which molecule or group should advance?	Potency, selectivity, cell assays, drug metabolism, PK, exposure, safety
SAR forecasting	26	What will a local chemical change do?	Highly similar compounds, potency, calculated chemical properties, binding poses
ADME and delivery	8	Will the molecule reach and persist at the site of action?	Solubility, permeability, efflux, metabolic stability or clearance, PK, tissue ratios
Safety and liabilities	4	What could go wrong, and can the risk be reduced?	Unintended target tests, heart ion channels, bile transport, metabolites
Molecular design and enumeration	6	Which valid structures meet the design rules?	Shared chemical cores, variable groups, stereochemistry, property limits
Assay biology and mechanism	3	What does the endpoint measure, and which direction is preferred?	Protein, cell, target-engagement, and mechanism assays
Retrosynthesis	4	How could the target molecule be made using commercially available precursors?	Functional groups, chosen bond breaks, reactions, earlier routes

Table 1: The seven primary task categories.

Each task combines prose with images of molecular structures, assay tables, plots, design diagrams, or protein-ligand complexes. Prompts introduce the biological context and therapeutic goal or target, describe the starting compound or chemical series, and identify a limitation that motivates further work. They then present data on related analogues, including biochemical, cellular, ADME, pharmacokinetic, safety, or structural evidence, before asking the model to make a decision about the next step in the program.

The 99 MCB tasks are grouped into seven primary categories (Table 1): (1) integrated lead selection, (2) SAR forecasting, (3) ADME and delivery, (4) safety and liabilities, (5) molecular design and enumeration, (6) assay biology and mechanism, and (7) retrosynthesis. Individual tasks can contain subtasks that span multiple modalities and scientific domains. Table 2 summarizes scientific-domain coverage. The following sections describe the task categories, answer components, and grading procedures.

Task taxonomy

Each task is assigned a primary category according to its main practical question (Table 1), but can require reasoning across several scientific domains. For example, choosing a lead compound to advance may involve interpreting structure-activity relationships, assay results, and differences in ADME properties. The tasks cover preclinical drug discovery; clinical development is outside the benchmark's scope.

We also annotate each task by its scientific-content domains (Table 2). These domains describe the scientific subject matter and work represented in the prompt, rather than the form of the requested answer. Domain annotations are author-derived and nonexclusive: a task can span potency, ADME, structure-guided reasoning, and other domains. Eighty-five of the 99 tasks span more than one domain.

Integrated lead selection

Integrated lead selection tasks in MCB require models to select the most promising candidate to advance from a set of compounds, subject to constraints on target potency, selectivity, activity in cells, physicochemical properties, permeability and efflux, metabolism, exposure, and safety. For

Scientific domain	Coverage	What it covers
Potency and selectivity	75/99	Target activity and selectivity over unwanted targets or phenotypes
ADME and exposure	54/99	Solubility, permeability, efflux, metabolism, clearance, pharmacokinetics, and tissue exposure
Structure-guided reasoning	51/99	Protein–ligand poses, binding contacts, conformational effects, and pocket geometry
Cellular and translational pharmacology	39/99	Cell, whole-blood, tissue, or in-vivo evidence connecting biochemical activity to biology
Molecular design	37/99	SAR constraints, design logic, analogue combinations, or proposed chemical changes
Safety and liabilities	11/99	Off-target effects, ion channels, reactive metabolites, and other safety risks
Synthesis	5/99	Reactions, precursors, route logic, and chemical feasibility

Table 2: Scientific-content domains in MedChemBench. Coverage is the number of tasks in each domain divided by the 99 tasks in the benchmark. Domain annotations are author-derived and nonexclusive, so the counts do not sum to 99.

example, one problem presents a series of diacylglycerol kinase (DGK) inhibitors [32] and asks which compounds are predicted to meet four requirements simultaneously: strong DGKα inhibition, strong activity in human whole blood, low human liver microsomal clearance, and adequate permeability across a Caco-2 intestinal cell layer. Models must weigh these properties under uncertainty, track compound identities across experiments and molecular depictions, and support their final choice with a clear scientific rationale.

SAR forecasting

Structure-activity relationship (SAR) forecasting tasks in MCB require models to predict how a local chemical change may alter a measured property, using data from earlier modifications to related compounds. These problems typically describe a promising scaffold that needs further modification to achieve desired properties. Prompts provide biochemical and cellular activity measurements, selectivity data, and images of binding poses for previously studied analogues. For example, one $KRAS^{G12C}$ inhibitor problem links a cyano substituent to changes in the orientation of two linked rings and their contacts with the protein, helping explain observed potency differences [33]. The model then uses this evidence to rank compounds bearing candidate substituents by predicted biochemical potency.

ADME and delivery

The absorption, distribution, metabolism, and excretion (ADME) tasks in MCB require models to predict properties that determine whether a compound can reach its intended site of action and remain there at a useful concentration. For example, one problem describes a CDK4 inhibitor series in which replacing a methyl group with a hydroxymethyl group reduces metabolic clearance but also reduces membrane permeability. The model receives structures and experimental measurements from earlier modifications, then predicts clearance and permeability for additional compounds and identifies those with the best balance of these properties. These tasks require models to weigh coupled properties, since improving one does not necessarily improve the overall delivery profile.

Safety and liability

Safety and liability tasks in MCB ask models to use experimental evidence from related molecules to predict unwanted effects associated with a compound or its metabolites. For example, one problem describes compounds whose metabolism can produce reactive intermediates and provides diagrams of observed and hypothesized metabolic pathways. The model is then shown additional analogues and asked to predict whether they will produce detectable glutathione conjugates in human liver microsomes, an assay readout used to investigate formation of these intermediates. Models must connect structural changes to the mechanism underlying a potential safety concern and explain why compounds may behave differently.

Molecular design and enumeration

The molecular design and enumeration tasks in MCB require models to propose chemical modifications or construct molecular structures under constraints motivated by preceding experimental results. For example, a problem may present assay measurements for compounds modified at two positions on a common scaffold, then ask the model to select favorable substituents at each position and construct a structure for every requested combination. Other problems involve more open-ended design: a model may receive peptide activity data and a crystal structure, then be asked to propose changes that improve stability without disrupting certain interactions important for potency. Models must translate experimental evidence into concrete chemical proposals while preserving the required connectivity, attachment points, and stereochemistry.

Assay biology and mechanism

The assay biology and mechanism tasks in MCB require models to interpret what different experimental measurements reveal about a compound's biological activity before using them to make predictions. For example, one problem presents a set of receptor agonists evaluated through two signaling readouts (cAMP inhibition and β-arrestin recruitment), then asks which additional compounds are likely to favor the desired response. The challenge is to relate experimental results to the underlying biological mechanism while recognizing that the same chemical modification may have different effects across assay readouts.

Retrosynthesis

The retrosynthesis tasks in MCB require models to propose a synthesis for a target compound using only commercially available precursors. Here, the problems provide the model with the target structure and its program context, sometimes together with a synthesis of a related compound, and the models must work backward through simpler intermediates to identify a plausible route. For example, one problem describes the synthesis of an oxadiazole-containing KRAS inhibitor and asks the model to develop a route to a more rigid, spirocyclic analogue. The challenge in these types of problems is to determine which parts of the earlier synthetic strategy can be reused and which require different transformations, while accounting for the target's functional groups, connectivity, and stereochemistry.

Graded subtask	Share of Score	Illustrative task	Rubric criteria
Molecular structure (SMILES)	37.49%	Use the common azaindole scaffold and displayed substituents to enumerate eight MAP4K1 inhibitors, compounds 30 and 32–38, as SMILES.	Reconstruct each molecular graph with the correct attachment points and specified stereochemistry. Each correct structure earns full criterion credit; different SMILES for the same graph are accepted.
Scientific rationale	35.78%	In the same MAP4K1 series, explain how moving the spirocyclopropane from its position in 32 to the oxygen-adjacent position in 33 could change potency and selectivity over ROCK2.	Identify the positional change and connect it to ring conformation, binding contacts, or presentation of the polar headgroup. Give qualified potency and selectivity forecasts based on the supplied evidence, independently of whether the selected compound set is correct.
Ranking and candidate selection	17.51%	Use the supplied early-series observations and eight candidate structures to select MAP4K1 inhibitors predicted to combine potency, ROCK2 selectivity, and oral bioavailability.	Select exactly compound 38 for full criterion credit: MAP4K1 $IC_{50} < 10$ nM, ROCK2/MAP4K1 IC_{50} ratio > 100, and oral bioavailability $F > 30\%$ under the stated study conditions. Partial credit reflects set overlap; missing and extra compounds reduce credit.
Quantity prediction (numerical and categorical)	8.92%	Given the PRMT5 inhibitor structures, earlier activity data, and binding poses, predict cellular potency, permeability, efflux, and microsomal clearance for 16, 23, and 24. In a separate thienopyrimidine task, predict an in-silico Ames mutagenicity label.	*Numerical:* a permeability prediction of 11×10^{-6} cm/s for 23 earns full criterion credit; a twofold error earns half credit, with credit varying continuously with fold error. *Categorical:* the reference positive Ames label earns full criterion credit; a negative label earns zero.

Table 3: Types of graded subtasks, mean shares of the available score, and illustrative tasks and criteria. The molecular-structure, scientific-rationale, and selection examples draw on the representative MAP4K1 task below; numerical prediction uses the representative PRMT5 task, and categorical prediction uses a separate benchmark task. Each task is normalized to a maximum score of 1. Component shares are the mean fractions of that available score across all 99 tasks, including zero where a component is absent. Quantity prediction combines numerical and categorical endpoints. The remaining 0.30% comprises technical response or availability credit, omitted from the rows but included in overall scores.

Grading

Each task is assessed using deterministic checks and model-judged criteria. We group substantive answer requirements into four components: molecular structure, scientific rationale, ranking and candidate selection, and numerical or categorical prediction. These components differ from the scientific content domains in Table 2: for example, potency can be assessed through a numerical prediction, a ranking, or a scientific rationale. Table 3 provides examples of subtasks in each component. A task can require more than one component.

Rubric breakdown

Each task is normalized to a maximum score of 1 and receives equal weight in the benchmark. Averaged across the 99 tasks, molecular structure accounts for 37.49% of the available score, scientific rationale for 35.78%, ranking and candidate selection for 17.51%, and numerical and categorical prediction for 8.92%. Technical checks account for the remaining 0.30%. Table 3 provides example tasks and grading criteria for each component.

The largest average share of the available score comes from molecular-structure criteria, which require models to provide isomeric SMILES strings for the compounds under consideration. These tasks combine visual interpretation, spatial reasoning, and chemical knowledge. Models must read chemical drawings, interpret stereochemical notation at chiral centers, and, where required, assemble scaffolds and substituents into the requested compounds.

Scientific rationale accounts for a similar average share of the available score. These criteria assess whether models draw on the supplied evidence to explain their answers, such as why a particular compound is the most promising lead.

Ranking and candidate selection assess whether models order or choose compounds according to the stated constraints. Quantity prediction assesses numerical values and categorical labels, including experimental measurements and calculated molecular properties reported in the source material.

Aggregation of scores across the benchmark

Each task has its own criteria, weights and points budget. For each model–effort condition, we divide a response's earned points by the task's full budget, average the ten retained responses with completed invalid outputs scored at zero, and then average all 99 tasks equally. Larger rubrics therefore do not give tasks more weight across the benchmark. Table 3 shows each component's average share of the available score.

Grading rules

Molecular structure. We use RDKit to convert submitted SMILES strings to canonical isomeric SMILES, preserving specified stereochemistry and isotopes. We compare each string with the corresponding reference in the same format. An exact match receives full credit; mismatched, invalid, or missing structures receive zero credit.

Ranking and candidate selection. Criteria requiring a complete ordered ranking are evaluated by comparing the predicted order of each compound pair with its reference order. Correctly ordered pairs within the full ordering receive credit of 1, reversed pairs receive 0, and predicted ties receive 0.5. Pairs whose members are interchangeable in the reference ranking are excluded. The ranking score is the mean credit across the remaining pairs; missing, repeated, or unknown compounds receive zero credit. In this approach, a uniformly random ordering has expected credit of 0.5. When a criterion asks for a specific set of candidates, credit is the Jaccard similarity between the predicted set P and reference set R: $|P \cap R|/|P \cup R|$. Selecting exactly the reference set earns full credit; extra or missing candidates reduce credit.

Numerical and categorical prediction. Most numerical predictions receive partial credit based on fold error: an exact prediction earns full credit, while a prediction that is twice or half the reference value earns half credit. This approach makes scores comparable across quantities measured on different scales. For 202 criteria with positive predictions p and reference values t, credit is $\min(p, t)/\max(p, t)$. Four criteria report values as base-10 logarithms; these use the equivalent formula $10^{-|p-t|}$.

Other numerical criteria ask for a specific reported value or require a threshold to be met. These follow the task's stated matching rules, tolerances, or thresholds. When reporting a measurement

given only as a bound, the answer must include both the bound and its direction, such as "less than." Each prediction must identify the requested compound and measurement and include any other required fields. Categorical predictions earn full credit for the reference label and zero otherwise.

Scientific rationale and task-specific constraints. A separate GPT-6 Astra judge at medium reasoning effort scores the scientific rationale criteria. The judge receives the question text and images, the criterion and grading guidance, and the submitted answer. Each criterion earns zero, half, or full credit according to its task-specific scoring anchors. Open-ended design and synthesis items can accept alternative answers that satisfy the task rules, although scoring remains partly tied to a reference design or route.

Representative tasks

We present two representative tasks, one in SAR forecasting and one in integrated lead selection. Both also require molecular structure enumeration. For readability, we show abbreviated JSON response schemas.

Representative problem 1: Prediction of PRMT5 inhibitor potency and ADME properties

This problem, based on [34][1], involves the development of brain-penetrant PRMT5 inhibitors for MTAP-deleted cancers. Given the earlier structure–activity relationships and crystallographic binding poses, the model must reconstruct compounds 16, 23, and 24 and predict cellular potency, permeability, efflux, and human microsomal clearance. The problem tests molecular reconstruction and quantitative forecasting from a shared scaffold with different pyridyl substituents. Numerical predictions are scored separately from molecular structure.

Prompt

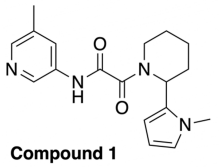
Methylthioadenosine phosphorylase (MTAP) is a key enzyme in the methionine salvage pathway. Loss of MTAP function occurs frequently across multiple cancers, including gliomas, pancreatic cancer, and melanoma.

Protein arginine methyltransferase 5 (PRMT5) is a type II methyltransferase that catalyzes symmetric dimethylation of arginine residues on both histone and non-histone proteins, thereby regulating gene expression, cell-cycle progression, and RNA splicing. PRMT5 is frequently overexpressed in cancer and has emerged as a therapeutic target in oncology.

Genome-wide genetic perturbation studies using shRNA screening have demonstrated that MTAP-deleted (MTAP-null) cancer cell lines exhibit a selective dependency on PRMT5 activity. Loss of MTAP leads to accumulation of methylthioadenosine (MTA), which partially inhibits PRMT5 and sensitizes the cells to additional pharmacological PRMT5 inhibition.

Given the high prevalence of MTAP deletion in patients with Glioblastoma (GBM), as well as in cancers that frequently metastasize to the brain, our objective was to develop brain-penetrant, MTA-cooperative PRMT5 inhibitors.

A high-throughput screening campaign led to the identification of hit compound **1**, which demonstrated a K_i of 600 nM in an MTA-dependent peptide displacement assay with fivefold selectivity, along with a K_d of 400 nM in an SPR assay conducted in the presence of MTA.

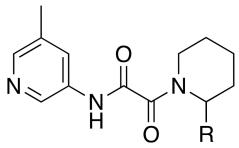


Compound 1

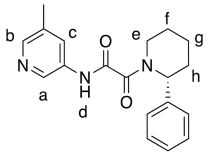
	+MTA (Me_0)	+SAM (Me_2)
IC_{50},	4 μM	4 μM
K_i	0.6 μM	3 μM
Selectivity (K_i, (+MTA)/K_i, (+SAM)) = 5		

Racemic compound **1** was separated by supercritical fluid chromatography (SFC) to afford the enantiomerically pure isomers **1R** and **1S**. Evaluation in the MTA-dependent peptide displacement biochemical assay revealed a clear stereochemical preference, with the **1R** isomer exhibiting substantially greater potency than the **1S** isomer (IC_{50} values of 2 μM and 33 μM, respectively). These results indicated that the R configuration at the 2-position of the piperidine ring is strongly preferred for activity.

[1]Source images: Cottrell et al. (2024), © 2024 The Authors, published by the American Chemical Society, licensed under CC BY 4.0. Figures and tables reproduce the benchmark prompt images and are resized for display; the binding-pose panels are from source Figure 5.



A broad exploration of substituents at the 2-position of the piperidine ring was undertaken to identify replacements for the pyrrole group. Compound **8** exhibited potency comparable to that of compound **1** while also showing lower cLogP and improved selectivity. Compound **10** demonstrated both greater potency and enhanced selectivity relative to compound **1**, despite lacking an aromatic substituent. Smaller substituents such as the methyl group in compound **9** resulted in a substantial loss of activity. We selected the phenyl-substituted piperidine motif, exemplified by compound **2R**, as the preferred scaffold.

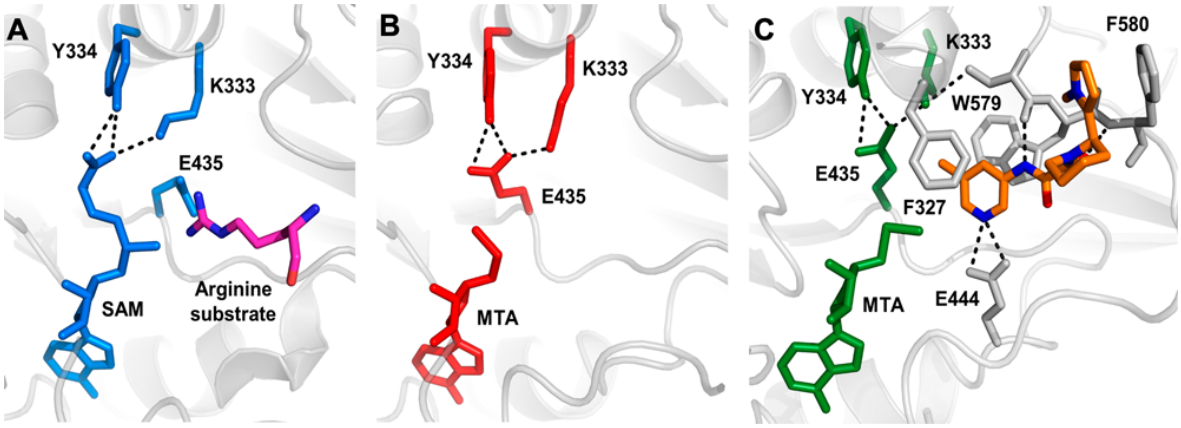


We conducted a systematic methyl scan across the scaffold. Most methyl substitutions either reduced activity or produced little effect, as observed for compounds **11**, **14**, **17**, **18**, **19**, **20**, and **21**. Several methyl substitutions resulted in substantial potency gains. Introduction of a cis-methyl group at the 6-position of the piperidine ring, as in compound **15**, improved biochemical potency approximately threefold, although this enhancement was insufficient to generate measurable cellular activity. Incorporation of a methyl group at the 5-position of the piperidine ring trans to the 2-phenyl substituent, as in compound **16**, produced an approximately twentyfold increase in biochemical potency in the presence of MTA while maintaining selectivity. Compound **16** provided the first evidence of selective cellular activity within the series, demonstrating MTAP-null SDMA IC_{50} values of 3.5 μM while remaining inactive in MTAP wild-type cells ($IC_{50} > 10$ μM). Combining the methyl substitutions from compounds **15** and **16** yielded the dimethyl analogue **22**, which was essentially equipotent to the corresponding monomethyl derivatives. Introduction of a methyl substituent at the 6-position of the pyridine ring, as in compound **12**, improved potency sixteenfold and produced selective cellular activity (MTAP-null SDMA $IC_{50} = 3.5$ μM; MTAP WT SDMA $IC_{50} > 10$ μM).

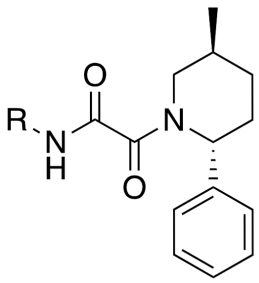
Compound	R	PRMT5 biochemical IC_{50} (μM)[a]	
		+ MTA (Me_0 peptide)[b]	+ SAM (Me_2 peptide)[c]
1R		2	2
1S		33	56
2R		1	2
2S		36	49
3[d]		7	10
4[d]		3	5
5[d]		6	25
6[d]		5	13
7[d]		7	15
8[d]		4	13
9[d]	CH3	38	38
10[d]		1	6

		biochemical IC_{50} (μM) a		SDMA IC_{50} HAP1 (μM) d	
compound	**methyl position**	**MTA (Me_0) b**	**SAM (Me_2) c**	**MTAP-null**	**MTAP WT**
11	a	>125	>125		
12	b	0.06	0.20	3.5	>10
13	c	0.9	11		
14	d	>125	>125		
15	e, cis	0.3	0.5		
16	f, trans	0.05	0.08	3.5	>10
17	f, cis	2	3		
18	g, trans	0.5	6		
19	g, cis	0.5	3		
20	h trans	3	5		
21	h cis	4	40		
22	e, f (2*R*,3*S*,6*R*)	0.2	0.6	4.5	>10

We obtained a co-crystal structure of **1R** bound to the PRMT5–MTA complex. The pyridine ring formed van der Waals contacts with MTA and participated in π-stacking interactions with the aromatic side chains of Phe327 and Trp579, while the pyridine nitrogen formed a hydrogen bond with the carboxylate of Glu444. The 5-methyl substituent on the pyridine projected toward Lys333 and Glu435 and stabilized Glu435 in the preferred rotameric conformation observed in the MTA-bound state. This sterically prevented Glu435 from adopting the alternative rotamer required for SAM binding and likely contributed to the observed selectivity profile. The structure revealed a small pocket adjacent to the 6-position of the pyridine ring, providing a structural rationale for the improved potency observed with the 6-methylpyridine analogue **12**. The NH group of the oxamide formed a hydrogen bond with the carbonyl of Ser578, while the two oxamide carbonyl groups engaged in additional hydrogen-bonding interactions: one directly with the backbone NH of Phe580 and the other indirectly through a water-mediated interaction with Leu312. The piperidine ring and its pyrrole substituent occupied an adjacent hydrophobic pocket, with the pyrrole ring engaging in π-stacking interactions with Phe580.



Given the above info, enumerate (provide SMILES) the below small molecules. Predict (with reasoning) the SDMA (MTAP-null) IC_{50}, permeability, efflux ratio and human microsomal intrinsic clearance (CLint) for each cpd.



Compound	R
16	
23	
24	

Using only the information in the question, return a JSON object with a single top-level key `answer`. Under `answer`, provide an `overall_reasoning` string plus three objects named `cpd16`, `cpd23`, and `cpd24`. For each compound object, include: `smiles` (string), `predicted_sdma_mtap_null_ic50_uM` (number_or_null), `predicted_permeability_1e-6_cm_per_s` (number_or_null), `predicted_efflux_ratio` (number_or_null), `predicted_human_microsomal_clint_uL_min_mg` (number_or_null) for `cpd16` and `cpd23`. For `cpd24`, instead use `predicted_human_microsomal_clint_upper_bound_uL_min_mg` (number_or_null) to report the numeric limit of the censored `<` result. For every compound, include `reasoning` with short strings for `enumeration`, `sdma_ic50`, `permeability`, `efflux_ratio`, and `microsomal_clint`. Put any medicinal-chemistry reasoning inside these JSON fields only; do not include text outside the JSON.

Output schema (abridged for display). The compound fields repeat for 16, 23, and 24. Compound 24 reports the numeric upper limit of its censored clearance result using the distinct upper-bound field.

```
{"answer": {
  "cpd16": {
    "smiles": "string",
    "predicted_sdma_mtap_null_ic50_uM": 0.0,
    "predicted_permeability_1e-6_cm_per_s": 0.0,
    "predicted_efflux_ratio": 0.0,
    "predicted_human_microsomal_clint_uL_min_mg": 0.0,
    "reasoning": {
      "enumeration": "string", "sdma_ic50": "string",
      "permeability": "string", "efflux_ratio": "string",
      "microsomal_clint": "string"}},
  "cpd23": "same fields as cpd16",
  "cpd24": {
    "predicted_human_microsomal_clint_upper_bound_uL_min_mg": 0.0,
    "...": "remaining fields as cpd16"},
  "overall_reasoning": "string"
}}
```

Do not include markdown fences or any text outside the JSON.

Reference answer: The source literature reported substantially greater cellular potency and lower human microsomal clearance for compounds 23 and 24 than for compound 16. Permeability remained in a broadly similar range across the three compounds, while efflux ratios increased from 16 to 23 to 24. The required SMILES preserve the depicted (2R,5S) piperidine scaffold, with its ring nitrogen linked to the oxamide and the indicated pyridyl substituent.

Models are graded on the three molecular structures (20% of the available credit) and on their predictions of potency (20%), permeability (20%), efflux (20%), and microsomal clearance (20%).

Numerical forecasts receive partial credit based on fold error relative to the reported measurements. The potency value already supplied for compound 16 and the reported clearance bound for compound 24 must be reproduced faithfully. Explanations carry no separate credit in this problem.

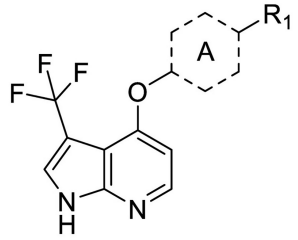
Representative problem 2: Selection of MAP4K1 inhibitors across potency, selectivity, and exposure

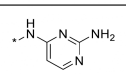
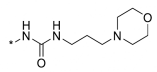
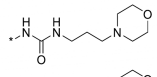
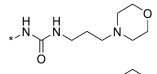
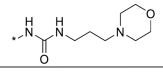
This problem, based on [35][2], involves the optimization of MAP4K1 inhibitors for cancer immunotherapy. Given the early optimization history, assay results, and structures of eight candidates, the model must predict which compounds meet thresholds for MAP4K1 potency, selectivity over ROCK2, and oral bioavailability, and which are limited specifically by bioavailability. The graded outputs are the eight molecular structures, two candidate sets, and compound-specific reasoning. The prompt asks for structural comparisons and for kinase binding to be considered separately from oral exposure, including the limits of attributing selectivity to size or rigidity.

Prompt

MAP4K1 (mitogen-activated protein kinase kinase kinase kinase 1), also known as hematopoietic progenitor kinase 1 (HPK1), is a serine/threonine kinase that serves as a key negative regulator of T-cell receptor (TCR) signaling. MAP4K1 suppresses both T-cell activation and dendritic cell function, making it an attractive therapeutic target for cancer immunotherapy.

Multiple discovery programs have been undertaken to identify potent and selective MAP4K1 inhibitors. Our efforts began with a high-throughput screening campaign that yielded several promising hits, including the azaindole-based lead compound **1** (BAY-755). Although compound **1** demonstrated meaningful MAP4K1 inhibitory activity (IC_{50} = 82 nM), it exhibited poor kinase selectivity and suboptimal DMPK properties.

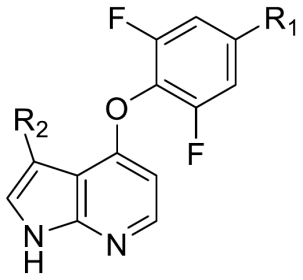


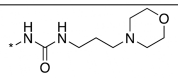
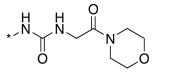
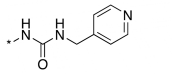
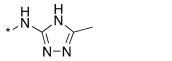
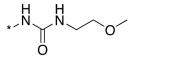
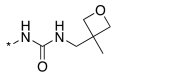
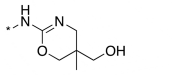
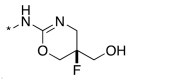
Cpd	A	R1	MAP4K1 IC_{50} (nM)[a]	ratio ROCK2[b]	ratio IRAK4[b]	ratio MST1[b]
1 **BAY-755**			75	0.03	184	3
2			7	n.d.	24	6
3			15	0.3	13	3
4			29	0.4	8	3
5			29	0.5	15	3

Compound **1** belongs to a class of molecules known to inhibit members of the Rho-associated kinase family, including ROCK1 and ROCK2. In biochemical assays, compound **1** exhibited single-digit nanomolar potency against ROCK2 (IC_{50} = 2 nM), representing approximately 40-fold greater potency than that observed for MAP4K1. Our initial efforts to improve selectivity over ROCK2 focused on modifications to the pyrimidine region and the difluorophenyl A-ring. Replacing the pyrimidine with urea-based motifs, as exemplified by compound **2**, resulted in an approximately tenfold increase in MAP4K1 potency, yielding single-digit nanomolar inhibitors (IC_{50} = 7 nM). X-ray crystallographic studies confirmed that the urea moiety formed a bidentate hydrogen-bonding interaction with the Asp101 side chain in MAP4K1 (corresponding to Asp112 in the surrogate kinase MST1). Despite this substantial gain in potency, improving selectivity over ROCK2 remained challenging. To further investigate the contribution of the A-ring, analogues **3**–**5** were synthesized. Removal of a single fluorine atom, as in compound **3**, produced an approximately twofold reduction in potency, while removal of the second fluorine atom in compound **4** resulted in an additional twofold loss. The relative positioning of the fluorine substituents also proved important, as the 2,5-difluoro isomer **5** showed a fourfold decrease in potency compared with compound **2**. These modifications to the A-ring also led to diminished metabolic stability, suggesting that the 2,6-difluoro substitution pattern plays an important role in both target potency and pharmacokinetic performance. Despite these structural changes, selectivity over ROCK2 remained poor, while selectivity against other kinases tested, including IRAK4 and MST1, either worsened or remained unchanged.

[2]Source images: Mowat et al. (2024), © 2024 The Authors, published by the American Chemical Society, licensed under CC BY 4.0. Figures and tables reproduce the benchmark prompt images, including its masked development-code label, and are resized for display.

Rat pharmacokinetic studies showed that compound **2** exhibited moderate in vivo clearance together with poor oral bioavailability (F = 8%). A series of modifications to the urea side chain was explored. Reducing the basicity of the morpholine ring by converting it to the corresponding amide, as in compound **6**, maintained MAP4K1 potency comparable to that of compound **2**. Compound **6** demonstrated more than a fourfold reduction in in vivo clearance together with improved oral bioavailability (F = 28%).



Cpd	R_2	R_1	MAP4K1 IC_{50} (nM)[a]	ratio ROCK2[b]	Rat *in vivo* CL_b (L/h/kg)[c]	F[d] (%)
2	CF_3		7	n.d.	1.6	8
6	CF_3		12	0.4	0.35	28
7	CF_3		9	0.2	0.88	2
8	Cl		161	0.02	0.18	29
9	CF_3		11	0.25	0.38	25
10	CF_3		14	0.23	-	-
11	CF_3		3	1.5	3.7	26
12 BAY-712	CF3		6	1.5	1.3	69

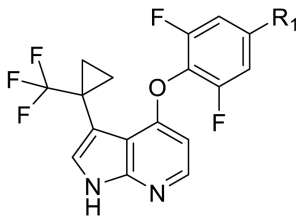
Side chains containing aromatic substituents, such as compound **7**, were also well tolerated and retained strong potency (IC_{50} = 9 nM). Compound **7** exhibited nearly threefold higher in vivo clearance than compound **6** and showed markedly reduced bioavailability (F = 2%). Evaluation of five-membered heterocycles designed to preserve the dual hydrogen-bond donor motif resulted in encouraging in vivo pharmacokinetic properties, although potency was less favorable. Compound **8** displayed an IC_{50} of 161 nM against MAP4K1. Ether-containing side chains, as exemplified by compound **9**, produced a more balanced overall profile like that observed for compound **6**. The racemic oxazine analogue **11** demonstrated excellent potency (IC_{50} = 3 nM), although it also showed high in vivo clearance (CL_b = 3.7 L/h/kg) and only moderate bioavailability (F = 26%). Compound **11** was the first analogue in the series to display greater activity against MAP4K1 than against ROCK2. The oxetane-containing urea analogue **10** exhibited a less favorable overall profile. Fluorine substitution on the oxazine scaffold to generate compound **12** produced a comparable potency and selectivity profile while significantly reducing in vivo clearance (CL_b = 1.3 L/h/kg) and improving oral bioavailability (F = 69%).

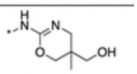
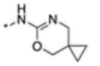
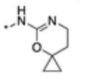
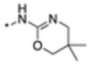
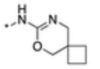
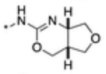
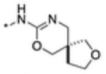
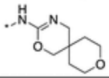
Given the above info, enumerate (provide SMILES) the below molecules. Predict (with reasoning) which cpds exhibit MAP4K1 IC_{50} < 10 nM, ROCK2 ratio > 100 AND F > 30%.

Return a JSON object matching the provided schema. Use the medicinal chemistry information in the prompt to: (1) enumerate the SMILES for compounds 30, 32, 33, 34, 35, 36, 37, and 38; (2) identify which compounds you predict satisfy all three criteria simultaneously: MAP4K1 IC50 < 10 nM, ROCK2 ratio > 100, and oral bioavailability F > 30%; (3) identify any compounds that nearly satisfy the profile but fail specifically because F is below 30%; and (4) provide a concise rationale with SAR/PK observations supporting your predictions. Use compound IDs as strings (for example, "30").

For the oral-bioavailability forecasts, use the compound-specific studies in male Wistar rats: a single oral dose of 0.6 mg/kg for 34, 35 and 37, and 1.0 mg/kg for 38. The oral formulation specified for 38 is water/Solutol/ethanol (50/40/10); formulations for the other compounds are not established by the information supplied here. For clearance comparisons, the corresponding single intravenous doses are 0.3 mg/kg for 34, 35 and 37 and 1.0 mg/kg for 38. These are study-condition-specific forecasts, not a common-dose or dose-independent comparison. Do not assume the unspecified formulations or missing study conditions are equivalent.

In `rationale.summary` and `rationale.sar_observations`, compare the azaindole C3 substituent and oxazine variants shown for 30 and 32-38 with the supplied early series. Cover substitution position (32 versus 33), the 32/34/35 substitutions relative to 30, fused versus spiro rings (36 versus 37), 37 versus 34 for binding and exposure, and spiro ring size (37 versus 38). For 34 versus 35, go beyond a size/shape forecast: explain why the available evidence cannot isolate size from rigidity or conformation, and propose a matched analogue comparison that could separate those effects on selectivity. Paired MAP4K1/ROCK2 assays can distinguish on-target from off-target contributions to the ratio, but cannot alone attribute them to size or rigidity. Separate target potency, ROCK2 selectivity and oral bioavailability. For proposed F-limited near misses, explain the evidence and uncertainty behind that call. Distinguish in vitro metabolic stability, in vivo clearance and oral F; none is interchangeable with another. Do not present unshown assay values as measurements.



Cpd	R1
30	
32	
33	
34	
35	
36	
37	
38	

Output schema (abridged for display). The `compounds` object contains one `smiles` field for each of 30, 32, 33, 34, 35, 36, 37, and 38.

```
{"answer": {
  "compounds": {
    "30": {"smiles": "string"},
    "...": "same fields for 32,33,34,35,36,37,38"},
  "qualifying_compounds": ["<compound_id>"],
  "near_miss_f_below_30": ["<compound_id>"],
  "rationale": {
    "summary": "string",
    "sar_observations": ["string"]}
}}
```

Do not include markdown fences or any text outside the JSON.

Reference answer: Compound 38 is the benchmark reference candidate satisfying all three thresholds for potency, ROCK2 selectivity, and oral bioavailability. Compounds 32, 34, and 35 are treated as bioavailability-limited near misses in the benchmark reference. The eight required SMILES are formed by joining the common scaffold to the displayed substituents while preserving their attachment points and specified stereochemistry.

Models are also graded on their reasoning about structure–activity relationships and ADME tradeoffs. This includes relating changes in substitution and ring geometry to separate forecasts of potency, selectivity, and oral exposure, and recognizing the limits of attributing selectivity to size or rigidity alone. Forecasts should draw on the supplied early-series observations and study conditions, with predictions distinguished from measurements. Candidate selections receive partial credit for overlap with the reference sets, while sound but incomplete explanations receive partial credit independently of compound selection.

MAP4K1 rubric	Full-credit requirement	Maximum credit
Molecular structures	Eight graph-equivalent SMILES for 30 and 32–38, with correct attachment points and specified stereochemistry; scored separately.	8/25
Qualifying compound	Select 38 as satisfying all three conditions.	1/25
Bioavailability near misses	Match {32, 34, 35}	3/25
C3 substituent	Distinguish cyclopropyl–CF_3 in 30 from directly attached CF_3; explain chemical implications and limits of forecasting selectivity.	1/25
Oxazine substitutions	Compare all of 32/34/35 with 30 using early oxazine evidence, connecting structural/conformational changes to separate potency and selectivity forecasts.	3/25
Substitution position	Compare 33 with 32; link position to contacts or conformation, with qualified potency and selectivity predictions.	1/25
Exposure in 34 and 35	Use an early clearance/bioavailability comparison to assess both compounds; explain why stability or lipophilicity alone does not establish oral bioavailability.	2/25
Size versus rigidity	Explain coupled changes in gem-dimethyl 34 versus spirocyclobutyl 35. Propose a matched-analogue comparison and contrasting outcomes isolating size from rigidity/conformation; paired kinase assays alone cannot do so.	2/25
Fused-ring geometry	Compare constraints in fused 36 with 30 or spiro 37, linking geometry to a binding hypothesis and conditional selectivity forecast.	1/25
Binding versus exposure	Compare spiro-THF 37 with 34: weigh conformation and polarity to support separate binding/selectivity and exposure forecasts.	1/25
Spiro-ring size	Compare the five- and six-membered ethers in 37/38: polar-group presentation, competing fit/entropy effects, and a grounded forecast.	1/25
Exposure in 38	Link its structure and an early clearance/bioavailability example to metabolism and absorption/solubility. Distinguish clearance from condition-specific oral bioavailability; avoid unshown measurements.	1/25
Total		**1**

Results

We evaluated eleven GPT models at all reasoning effort settings available at the time of evaluation: GPT-5.2, GPT-5.4, GPT-5.5, GPT-5.6 Luna, GPT-5.6 Terra, GPT-5.6 Sol, GPT-6 Luna, GPT-6 Sol, GPT-6 Astra, GPT-6.1 Sol, and GPT-6.1 Terra. Each of the 61 model–effort conditions contains ten retained responses for each of 99 tasks, giving 60,390 responses. Scores are the average proportion of available rubric points earned per task replicate, averaged over all replicates with completed invalid responses scored at zero. The intervals reported below are 95% bootstrap intervals; see Methods for details on the statistical analysis.

GPT-6 Astra at `max` achieved the highest observed mean score, 83.83% (interval, 81.27–86.20%). At the same effort setting, GPT-6.1 Sol and GPT-6.1 Terra earned 82.89% (80.52–85.23%) and 77.56% (74.98–79.95%), respectively. GPT-6 Sol and GPT-6 Luna at `max` earned 69.07% (65.52–72.38%) and 40.10% (36.35–43.88%), respectively. The best performing GPT-5-family model was GPT-5.6 Sol at `max` reasoning effort, with 59.77% (56.09–63.52%; Figure 2). At `xhigh`, the highest evaluated reasoning effort shared by all eleven models, scores ranged from 35.02% for GPT-6 Luna to 83.33% for GPT-6 Astra; GPT-5.6 Sol earned 58.23%, exceeding GPT-5.5 by 8.97 percentage points.

Models differed most in molecular structure performance

At `xhigh`, molecular structure scores showed the widest range across models: GPT-6 Astra earned 95.37% of available structure credit, GPT-6.1 Sol and Terra earned 93.70% and 92.18%, respectively, GPT-6 Sol earned 65.85%, GPT-6 Luna earned 12.42%, GPT-5.6 Sol earned 47.31%, and GPT-5.5 earned 28.70%, compared with 8.74–17.00% for the other four models (Figure 2C–F). The ranges across all eleven models were much narrower for scientific rationale (50.94–91.05%), ranking and candidate selection (43.58–52.78%), and numerical prediction (37.69–47.02%). GPT-6 Astra's corresponding values were 91.05%, 50.42%, and 47.02%, respectively.

GPT-5.6 Luna and GPT-5.6 Terra performed poorly on molecular-structure tasks relative to GPT-5.5, GPT-5.6 Sol, and Astra, consistent with weaker spatial reasoning in reconstructing and assembling chemical structures.

Figure 3 illustrates a typical structural reconstruction problem from MCB with large differences in model performance. The task is drawn from the program that led to daraxonrasib [36] and requires attaching substituents from two sets to a macrocyclic RAS(ON) inhibitor scaffold. The example shown, compound 17, combines this scaffold with an N-methylpiperazine group and a stereodefined fused bicyclic carboxamide. The correct structure requires preserving the macrocyclic ester and amide connections, the piperazic N–N bond, heteroaromatic attachment positions, and specified stereochemistry while incorporating both substituents at the specified locations. At `xhigh`, GPT-6 Astra reproduced the correct molecular graph and specified stereochemistry in all 10 responses, whereas GPT-5.6 Sol and GPT-5.6 Luna did not manage to reconstruct the correct graph in any of their attempts.

Depending on the model, the errors could often be egregious, as shown in the figure. For example, the illustrated GPT-5.6 Sol response adds a fused cyclopropane to the piperazine ring and replaces N-methyl with N-isopropyl. The resulting molecule has 67 heavy atoms rather than 60, and its macrocycle contains 15 ring atoms rather than 18. The selected GPT-5.6 Luna response displayed even worse results by returning two entirely disconnected fragments and completely omitting the required assembled macrocycle. Its piperazine fragment also has an N-isopropyl group

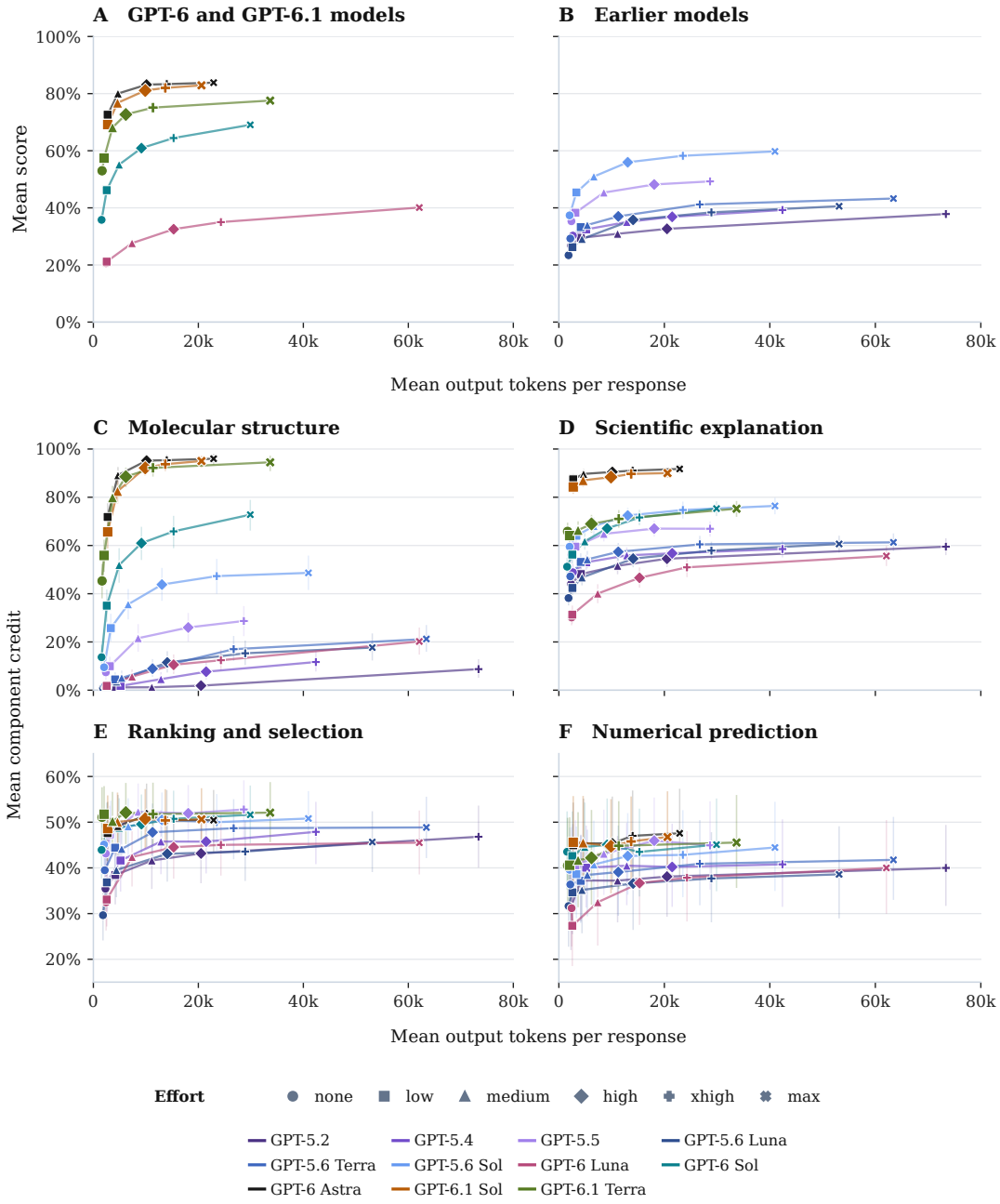


Figure 2: Overall and component-level MedChemBench performance. **A–B,** Mean score versus total output-token use across eleven models and 61 model–effort conditions, with GPT-6 and GPT-6.1 models on the left and earlier models on the right. Both panes use the same axes. **C–F,** Component performance versus mean output-token use, with pointwise 95% whole-task bootstrap intervals. Token use in every panel is averaged over all 99 tasks. Component points are normalized within each applicable task, averaged across its responses, and then averaged equally across applicable tasks (92, 91, 86, and 21 tasks in C–F, respectively). Panels E–F use a 15–65% vertical scale; panels A–D use 0–100%. Colors identify models; symbols identify effort in every panel. Lines connect the tested effort settings for each model. Numerical values are reported in Supplementary Tables S1 and S2.

and an additional ring methyl, while the bicyclic ether has different connectivity and no specified stereochemistry.

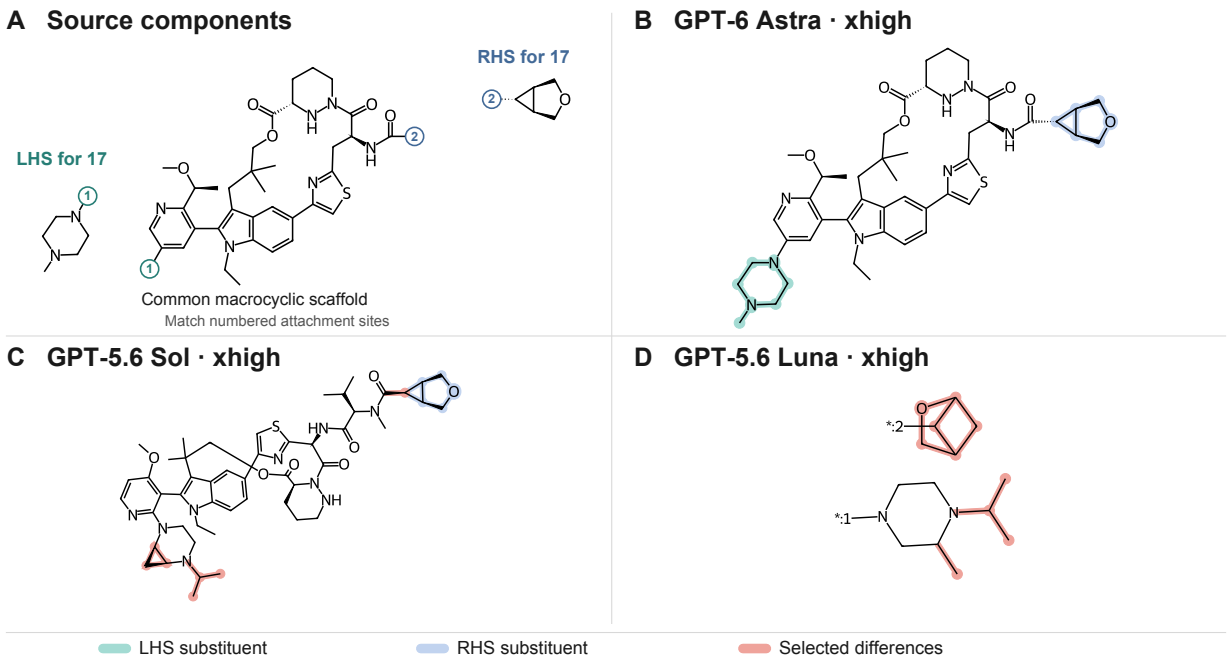


Figure 3: Assembly of macrocyclic RAS(ON) inhibitor 17 from source evidence and saved `xhigh` responses. **A,** Common scaffold and the two substituents specified for compound 17, with matching numbered markers identifying the attachment sites to be joined. **B,** GPT-6 Astra reproduces the reference structure, including its specified stereochemistry. **C,** GPT-5.6 Sol introduces a cyclopropane fused to the piperazine ring and an N-isopropyl group in place of N-methyl, among other changes to the molecular graph. **D,** GPT-5.6 Luna returns two disconnected, altered fragments instead of the assembled macrocycle. In A, matching teal 1 and blue 2 markers identify the left-hand-side (LHS) and right-hand-side (RHS) attachment sites. In B, teal and blue highlights identify the LHS and RHS substituents. In C, blue highlights show the preserved RHS ring connectivity; red marks the inverted stereochemistry at its attachment carbon and selected LHS differences. Red highlights in D mark selected connectivity differences. Red highlights mark selected errors; unhighlighted regions are not necessarily correct.

This example shows how molecular reconstruction depends on maintaining local substituent identity and stereochemistry together with connectivity across a large ring system. It also illustrates differences between models in interpreting multimodal evidence, spatial reasoning, and chemical reasoning.

Differences in fidelity of scientific reasoning

Scientific rationale, which accounts for the second largest average share of the available score, also showed substantial variation across models. At the shared `xhigh` setting, GPT-6 Astra earned 91.05% of the available score (95% interval, 89.02–92.96%), compared with 89.68% for GPT-6.1 Sol (87.50–91.73%), 74.72% for GPT-5.6 Sol (71.14–78.20%), 71.61% for GPT-6 Sol, 71.02% for GPT-6.1 Terra (67.37–74.67%), and 66.94% for GPT-5.5 (63.67–70.02%). GPT-6 Luna earned 50.94%. GPT-5.2 and GPT-5.6 Luna earned 59.51% and 57.88%, respectively; GPT-5.4 and GPT-5.6 Terra earned 58.52% and 60.46%.

The MAP4K1 lead-selection task in Representative problem 2 illustrates two differences in fidelity to the supplied evidence (Table 4). The task asked models to predict which inhibitors would meet the specified requirements for MAP4K1 potency, selectivity over ROCK2, and oral bioavailability, and which would fall short specifically because of low bioavailability. Models also had to explain their predictions using the supplied structures and early-series evidence.

Criterion and context	Comparator response	GPT-6 Astra response
Oxazine substitution in 32/33. Both compounds have a spirocyclopropane, but 33 places it next to the ring oxygen. Explain how this positional change could affect conformation and kinase recognition.	**GPT-5.2: zero credit.** Bases its comparison on a carbamoylated ring nitrogen that is not present. Excerpt for 32/33: "moving the cyclopropyl from the carbon α to the carbamoylated N (32) to the distal carbon near the ring oxygen (33) changes how bulk is presented toward the MAP4K1 pocket"	**Full credit.** Identifies the actual positional change and its conformational consequences. Excerpt for 32/33: "32 has its spirocyclopropane at oxazine C5, whereas 33 places it at C6, adjacent to oxygen. This changes local steric demand and ring puckering without changing gross composition."
Spiro ether in 37 versus 34. Compare replacement of gem-dimethyl substitution with a spiro tetrahydrofuran, assessing kinase binding and oral exposure separately.	**GPT-5.6 Luna: half credit.** Recognizes added polarity but leaves the separate kinase-binding predictions underdeveloped. Excerpt for 37: "The extra oxygen and altered three-dimensional shape are expected to improve polarity and potentially exposure while retaining the hinge-binding and bidentate interaction topology."	**Full credit.** Separates a conditional binding hypothesis from competing exposure effects. Excerpt for 37: "For exposure, additional polarity could improve dissolution yet reduce permeability or introduce metabolic liability. My preference for 37 as an F $>$ 30% candidate assumes a favorable balance rather than equating an ether oxygen with better PK."

Table 4: Scientific-rationale excerpts for the MAP4K1 lead-selection task in Representative problem 2, using the first saved response at `xhigh` for each model. F denotes oral bioavailability and PK pharmacokinetics. Excerpts are verbatim apart from typesetting.

GPT-5.2 based its comparison of compounds 32 and 33 on a carbamoylated ring nitrogen that is absent from the supplied structures. Both compounds instead contain a spirocyclopropane, but at different positions on the oxazine ring. GPT-6 Astra, on the other hand, identified the movement of the spiro center to the carbon next to oxygen and explained how this could change ring conformation and presentation of the polar headgroup. As a result, GPT-5.2 received zero credit for the corresponding positional-comparison reasoning criterion, while GPT-6 Astra received full credit.

Additionally, the responses differed in how they applied general property trends to individual compounds. For the replacement of gem-dimethyl substitution in compound 34 with a spiro tetrahydrofuran in 37, GPT-5.6 Luna recognized added polarity and potentially improved exposure but left the separate arguments for MAP4K1 potency and ROCK2 selectivity underdeveloped. GPT-6 Astra proposed a conditional difference in kinase recognition and explained how greater polarity could improve dissolution while reducing permeability or introducing metabolic liability. These differences concern the selected reasoning criteria, not the correctness of the final candidate sets; Astra also selected an incorrect qualifying set in this response.

Sensitivity to source identification

Although the sources for the 99 problems are relatively recent (median publication year, 2025), we were concerned that memorization could contribute to the stronger performance of more recent models. As a heuristic check, we reran all 99 problems with the source DOI or patent identifier added to each prompt. We analyzed ten responses per problem and prompt condition for each of nine models at `xhigh` effort, pairing scores by problem. For GPT-6 Sol and Luna, we reused their canonical unmodified responses as historical controls; for the other seven models, both prompt conditions were generated anew. Completed invalid responses remained eligible, so selection depended on API

completion rather than answer quality. The original problem content and grading rules remained unchanged in this experiment, and no additional source text or retrieval tools were supplied.

Mean overall score changes after adding the source DOI or patent identifier ranged from −0.23 to +0.80 percentage points. All model-specific 95% confidence intervals included zero (10,000 paired bootstrap resamples of source groups), and model rankings remained unchanged. Thus, explicitly identifying the source produced little average performance benefit, but this test cannot establish whether source material was present in training data.

Discussion

We present MedChemBench, a retrospective benchmark of evidence integration and decision making in medicinal chemistry. Its 99 tasks span seven categories of work across hit finding, lead optimization, candidate selection, and supporting preclinical studies. Each task is based on a recently published paper or patent describing a drug program. It presents a scientific narrative interleaved with selected experimental evidence, including molecular structures, assay results, plots, and binding views, and asks the model to make a program decision. This case-based assessment approach complements broader evaluations of chemical knowledge and multimodal scientific reasoning by evaluating these capabilities within a shared medicinal chemistry context based on real historical work [18, 19].

Across eleven models in the GPT family, with effort sweeps for all models, the highest mean score was 83.83%, achieved by GPT-6 Astra at `max` effort. At the highest reasoning effort shared across all models, `xhigh`, model scores ranged from 35.02% to 83.33%. Across all eleven models, increasing test-time compute was associated with mean score gains of 8.97–33.26 percentage points, while substantial differences between models remained. Among the grading components, molecular structure showed the widest score range: GPT-6 Astra earned 95.37% of available structure credit at `xhigh`, compared with 47.31% for GPT-5.6 Sol. Scientific rationale also showed substantial differences across models, while score ranges were narrower for ranking and candidate selection and for numerical prediction.

MCB's retrospective design reduces concerns about task realism compared with a synthetic setup. However, many tasks, especially those involving numerical prediction, require models to predict held-out experimental results. Performance therefore primarily measures agreement with a single historical case under the stated conditions, which may not be fully representative. Contamination or memorization also remains a potential concern: the small average benefit from explicit source identifiers does not rule out training-data exposure. The benchmark is also close to saturation in its current form: GPT-6 Astra achieved 83.83% at `max` reasoning effort. Future versions will need to be more challenging to continue distinguishing model performance.

Additionally, open-ended molecular design and synthesis problems remain partly tied to reference answers, which may penalize alternative but equally plausible designs. Further validation should assess whether benchmark grading agrees with independent medicinal chemists' evaluations of alternative approaches and distinguish incomplete or incorrect answers from workable alternatives not described in the source material. Finally, the single-turn protocol measures performance with a fixed evidence package and does not assess recovery through computation, additional evidence, or iterative revision.

Future work should address these limitations through broader task and domain coverage, explicit treatment of reference assay uncertainty (as most reference answers are given as point estimates), expert adjudication of alternative answers, and validation on newly collected or prospective cases. Richer multimodal inputs could include raw machine-readable assay data and spectra or interactive

three-dimensional molecular and protein structures. A complementary agent-based track should allow models to retrieve information, calculate molecular properties, verify structures, and revise proposals over multiple steps, building on emerging evaluations of coding agents in drug discovery [27]. Such extensions would test whether improvements in molecular representation support more reliable medicinal-chemistry decisions when models can inspect evidence, compute, and iterate.

Methods

Model evaluation

Each model response was generated in a single turn. Models received text and images in the specified order and were instructed to return strict JSON with one top-level answer object. The scoring rules are described in Grading. API-error and incomplete runs were recovered using unchanged requests. For GPT-6 Luna at `none` effort, one completed replacement response was selected uniformly at random before grading when multiple replacements were available. Completed invalid responses and refusals were retained at zero credit; missing responses or grades were not scored as zero. Recovery did not screen responses for answer quality, and these procedures condition the results on captured completion. Token-use summaries describe retained responses, count reasoning tokens once within total output tokens, and exclude input tokens, unsuccessful or unselected recovery attempts, and judge calls.

GPT-6 Sol and Luna were evaluated before the provider's September 2026 image-encoding fix; its effect on these results was not measured (OpenAI API changelog).

Statistical analysis

For component analyses, we use six principal components: molecular structure, scientific rationale, ranking and candidate selection, numerical prediction, categorical prediction, and technical checks. Together with two small model-judged selection/assignment components, these form an eight-component partition. All analyses include ten retained responses per task and model–effort condition, including invalid responses at zero credit. Response scores are first averaged within each task.

Overall scores normalize earned points by each task's full rubric budget and then average equally across all 99 tasks. Composition is the mean over all 99 tasks of each component's available points divided by the full task budget. Subtask attainment instead normalizes earned points by the component's available points within each applicable task, then averages equally across applicable tasks. A component's additive contribution to the overall score is the mean over all 99 tasks of its earned points divided by the full task budget; it is not generally the product of attainment and composition.

Pointwise 95% percentile intervals use 5,000 nonparametric task-bootstrap samples, with independent resamples for each model–effort condition and metric. Overall-score and additive-contribution intervals resample all 99 tasks with replacement; subtask-attainment intervals resample directly from the applicable tasks. The per-task normalized scores and replicate means are held fixed. These are intervals for individual scores, not paired model contrasts. Response replicates are not separately resampled, and tasks rather than source publications are the resampling unit.

References

[1] Minikel EV, Painter JL, Dong CC, Nelson MR. Refining the impact of genetic evidence on clinical success. *Nature*. 2024;629:624–629. doi:10.1038/s41586-024-07316-0.

[2] Congressional Budget Office. Research and Development in the Pharmaceutical Industry. *Congressional Budget Office Report*. 2021. official report.

[3] Zhou Y, Zhang Y, Xu H, et al. Dynamic clinical trial success rates for drugs in the 21st century. *Nature Communications*. 2025;16:9537. doi:10.1038/s41467-025-64552-2.

[4] Wermuth CG, Ganellin CR, Lindberg P, Mitscher LA. Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). *Pure and Applied Chemistry*. 1998;70:1129–1143. doi:10.1351/pac199870051129.

[5] Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. *British Journal of Pharmacology*. 2011;162:1239–1249. doi:10.1111/j.1476-5381.2010.01127.x.

[6] Quancard J, Bach A, Borsari C, et al. The European Federation for Medicinal Chemistry and Chemical Biology (EFMC) Best Practice Initiative: Hit to Lead. *ChemMedChem*. 2025;20:e202400931. doi:10.1002/cmdc.202400931.

[7] Plowright AT, Johnstone C, Kihlberg J, et al. Hypothesis driven drug design: improving quality and effectiveness of the design–make–test–analyse cycle. *Drug Discovery Today*. 2012;17:56–62. doi:10.1016/j.drudis.2011.09.012.

[8] Rácz A, Mihalovits LM, Beckers M, et al. The changing landscape of medicinal chemistry optimization. *Nature Reviews Drug Discovery*. 2025;24:870–887. doi:10.1038/s41573-025-01225-1.

[9] Tan L, Hirte S, Palmacci V, et al. Tackling assay interference associated with small molecules. *Nature Reviews Chemistry*. 2024;8:319–339. doi:10.1038/s41570-024-00593-3.

[10] Kenakin T. Know your molecule: pharmacological characterization of drug candidates to enhance efficacy and reduce late-stage attrition. *Nature Reviews Drug Discovery*. 2024;23:626–644. doi:10.1038/s41573-024-00958-9.

[11] Brennan RJ, Jenkinson S, Brown A, et al. The state of the art in secondary pharmacology and its impact on the safety of new medicines. *Nature Reviews Drug Discovery*. 2024;23:525–545. doi:10.1038/s41573-024-00942-3.

[12] Bender A, Thomas MC, Scannell JW, et al. Artificial intelligence in drug discovery—what it is, where we stand and the path forward. *Nature Reviews Drug Discovery*. 2026. doi:10.1038/s41573-026-01496-2.

[13] Wu Z, Ramsundar B, Feinberg EN, et al. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science*. 2018;9:513–530. doi:10.1039/C7SC02664A.

[14] Huang K, Fu T, Gao W, et al. Artificial intelligence foundation for therapeutic science. *Nature Chemical Biology*. 2022;18:1033–1036. doi:10.1038/s41589-022-01131-2.

[15] van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of Chemical Information and Modeling*. 2022;62:5938–5951. doi:10.1021/acs.jcim.2c01073.

[16] Gao W, Fu T, Sun J, Coley CW. Sample efficiency matters: a benchmark for practical molecular optimization. *Advances in Neural Information Processing Systems*. 2022;35:21342–21357. doi:10.52202/068431-1551.

[17] Guo T, Guo K, Nan B, et al. What can large language models do in chemistry? A comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*. 2023;36:59662–

59688. official proceedings.

[18] Mirza A, Alampara N, Kunchapu S, et al. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nature Chemistry*. 2025;17:1027–1034. doi:10.1038/s41557-025-01815-x.

[19] Alampara N, Schilling-Wilhelmi M, Ríos-García M, et al. Probing the limitations of multimodal language models for chemistry and materials research. *Nature Computational Science*. 2025;5:952–961. doi:10.1038/s43588-025-00836-3.

[20] Huang Y, Zhang R, He X, et al. ChemEval: A Multi-level and Fine-grained Chemical Capability Evaluation for Large Language Models. *The Fourteenth International Conference on Learning Representations (ICLR)*. 2026. OpenReview:JrqjSkEPrX; earlier preprint: doi:10.48550/arXiv.2409.13989.

[21] Li H, Fang X, Li Y, et al. RxnBench: a multimodal benchmark for evaluating large language models on chemical reaction understanding from the scientific literature. *Journal of Chemical Information and Modeling*. 2026;66:5747–5756. doi:10.1021/acs.jcim.6c00286.

[22] Mas P, Filoche-Rommé B, Bianciotto M, Vuilleumier R. Back to the future of lead optimization: benchmarking compound prioritization strategies. *Digital Discovery*. 2026;5:2983–2997. doi:10.1039/D5DD00387C.

[23] Li J, Chen B, Zou Z, et al. DMPKBench: A Multi-Modal Benchmark for Evaluating LLMs and Agents in Drug Discovery DMPK Tasks. *NeurIPS 2025 Workshop: AI for Science*. 2025. OpenReview:1NSnXVTxNR.

[24] Han K, Zhang R, Wei K, et al. SMDD-Bench: Can LLMs Solve Real-World Small Molecule Drug Design Tasks? *arXiv*. 2026;arXiv:2605.21740. doi:10.48550/arXiv.2605.21740.

[25] Li C, Tan H, Zhu W, et al. MolOpt-Eval: Can Frontier LLMs Perform Structure-Based Hit-to-Lead Optimization? *ICML 2026 Workshop on Generative and Agentic AI for Biology*. 2026. official OpenReview paper.

[26] Le H, Ramasamy R, Urrutia A, et al. TxBench-PP: Analyzing AI Agent Performance on Small-Molecule Preclinical Pharmacology. *arXiv*. 2026;arXiv:2606.19245. doi:10.48550/arXiv.2606.19245.

[27] Akyürek AF, Tu X, Gutmanstein A, et al. DrugDiscoveryBench: Can Coding Agents Assist Early-Stage Drug Discovery? *Scale Labs and Phylo Technical Report*. 2026. official report.

[28] Liu A, Ho A, Droste AM, et al. LifeSciBench: Evaluating Language Models on Realistic, Expert-Level Tasks in the Life Sciences. *bioRxiv*. 2026;2026.08.13.744657. doi:10.64898/2026.08.13.744657.

[29] Laurent JM, Bou A, Pieler M, et al. LABBench2: An Improved Benchmark for AI Systems Performing Biology Research. *arXiv*. 2026;arXiv:2604.09554. doi:10.48550/arXiv.2604.09554.

[30] Li J, Ho A. GeneBench: Assessing AI Agents for Multi-Stage Inference Problems in Genomics and Quantitative Biology. *bioRxiv*. 2026;2026.04.22.720113. doi:10.64898/2026.04.22.720113.

[31] Li JH, Shringarpure S, Wong E, Ho AJ. GeneBench-Pro: Evaluating Multistage Statistical Reasoning in Genomics, Quantitative Biology, and Translational Biomedicine. *bioRxiv*. 2026;2026.06.29.735386. doi:10.64898/2026.06.29.735386.

[32] Wei B, Li X, Lacharity JJ, et al. Discovery of INCB191358: A Potent and Selective DGKα/ζ Dual Inhibitor. *Journal of Medicinal Chemistry*. 2026;69:10545–10578. doi:10.1021/acs.jmedchem.5c03768.

[33] Huang D, Manoni F, Sun Z, et al. Identification of Structurally Novel KRAS^{G12C} Inhibitors through Covalent DNA-Encoded Library Screening. *Journal of Medicinal Chemistry*. 2025;68:4801–4817. doi:10.1021/acs.jmedchem.4c03071.

[34] Cottrell KM, Briggs KJ, Whittington DA, et al. Discovery of TNG908: A Selective, Brain Penetrant, MTA-Cooperative PRMT5 Inhibitor That Is Synthetically Lethal with MTAP-Deleted Cancers. *Journal of Medicinal Chemistry*. 2024;67:6064–6080. doi:10.1021/acs.jmedchem.4c00133.

[35] Mowat J, Carretero R, Leder G, et al. Discovery of BAY-405: An Azaindole-Based MAP4K1 Inhibitor for the Enhancement of T-Cell Immunity against Cancer. *Journal of Medicinal Chemistry*. 2024;67:17429–17453. doi:10.1021/acs.jmedchem.4c01325.

[36] Cregg J, Edwards AV, Chang S, et al. Discovery of Daraxonrasib (RMC-6236), a Potent and Orally Bioavailable RAS(ON) Multi-selective, Noncovalent Tri-complex Inhibitor for the Treatment of Patients with Multiple RAS-Addicted Cancers. *Journal of Medicinal Chemistry*. 2025;68:6064–6083. doi:10.1021/acs.jmedchem.4c02314.

Author contributions

Paul Humphries designed the first version of the benchmark problems. Andrew Ho and Jeremy Li refined, edited, and improved them.

Supplementary Information

Model	Effort	Mean output tokens per response	Mean score (95% interval)
GPT-5.2	none	2.3k	26.8% (24.36%–29.28%)
GPT-5.2	low	4.17k	29.51% (27.1%–32.01%)
GPT-5.2	medium	11.1k	30.92% (28.32%–33.58%)
GPT-5.2	high	20.5k	32.65% (30.24%–35.19%)
GPT-5.2	xhigh	73.4k	37.83% (35.03%–40.72%)
GPT-5.4	none	2.66k	30.27% (27.64%–32.93%)
GPT-5.4	low	5.18k	32.45% (29.93%–35.04%)
GPT-5.4	medium	12.9k	35.04% (32.36%–37.86%)
GPT-5.4	high	21.5k	36.86% (34%–39.74%)
GPT-5.4	xhigh	42.4k	39.24% (36.18%–42.52%)
GPT-5.5	none	2.38k	35.31% (32.66%–38.14%)
GPT-5.5	low	3.14k	38.29% (35.51%–41.22%)
GPT-5.5	medium	8.53k	45.35% (42.2%–48.65%)
GPT-5.5	high	18.1k	48.19% (44.95%–51.62%)
GPT-5.5	xhigh	28.7k	49.26% (45.92%–52.77%)
GPT-5.6 Luna	none	1.84k	23.42% (21%–25.93%)
GPT-5.6 Luna	low	2.57k	26.25% (23.91%–28.63%)
GPT-5.6 Luna	medium	4.39k	29.1% (26.66%–31.72%)
GPT-5.6 Luna	high	14.1k	35.8% (32.83%–39.04%)
GPT-5.6 Luna	xhigh	28.9k	38.42% (35.07%–41.84%)
GPT-5.6 Luna	max	53.1k	40.57% (37.12%–44.11%)
GPT-5.6 Terra	none	2.17k	29.25% (26.85%–31.74%)
GPT-5.6 Terra	low	4.15k	33.25% (30.46%–36.22%)
GPT-5.6 Terra	medium	5.4k	33.99% (31.08%–36.89%)
GPT-5.6 Terra	high	11.3k	37.01% (34%–40.1%)
GPT-5.6 Terra	xhigh	26.7k	41.19% (37.84%–44.51%)
GPT-5.6 Terra	max	63.4k	43.26% (39.89%–46.69%)
GPT-5.6 Sol	none	2.04k	37.37% (34.53%–40.31%)
GPT-5.6 Sol	low	3.32k	45.4% (41.81%–49.07%)
GPT-5.6 Sol	medium	6.64k	50.96% (47.24%–54.66%)
GPT-5.6 Sol	high	13.1k	55.96% (52.27%–59.57%)
GPT-5.6 Sol	xhigh	23.5k	58.23% (54.4%–61.93%)
GPT-5.6 Sol	max	41k	59.77% (56.09%–63.52%)
GPT-6 Astra	low	2.72k	72.61% (69.28%–75.91%)
GPT-6 Astra	medium	4.68k	80.02% (77.29%–82.69%)
GPT-6 Astra	high	10.1k	83.14% (80.65%–85.49%)
GPT-6 Astra	xhigh	14k	83.33% (80.71%–85.71%)
GPT-6 Astra	max	22.9k	83.83% (81.27%–86.2%)
GPT-6 Sol	none	1.57k	35.81% (32.47%–39.16%)
GPT-6 Sol	low	2.55k	46.19% (42.41%–49.93%)
GPT-6 Sol	medium	4.91k	55.13% (51.44%–58.79%)
GPT-6 Sol	high	9.16k	60.92% (57.29%–64.51%)
GPT-6 Sol	xhigh	15.3k	64.44% (60.72%–67.94%)
GPT-6 Sol	max	29.9k	69.07% (65.52%–72.38%)

Continued on the next page.

Table S1 continued from the previous page.

Model	Effort	Mean output tokens per response	Mean score (95% interval)
GPT-6 Luna	`none`	2.42k	20.64% (18.31%–23.11%)
GPT-6 Luna	`low`	2.55k	21.16% (18.57%–23.86%)
GPT-6 Luna	`medium`	7.37k	27.64% (24.74%–30.54%)
GPT-6 Luna	`high`	15.3k	32.53% (29.41%–35.86%)
GPT-6 Luna	`xhigh`	24.3k	35.02% (31.64%–38.53%)
GPT-6 Luna	`max`	62.1k	40.1% (36.35%–43.88%)
GPT-6.1 Sol	`low`	2.73k	69.16% (65.62%–72.64%)
GPT-6.1 Sol	`medium`	4.6k	76.74% (73.69%–79.75%)
GPT-6.1 Sol	`high`	9.89k	81.09% (78.41%–83.58%)
GPT-6.1 Sol	`xhigh`	13.7k	81.99% (79.47%–84.37%)
GPT-6.1 Sol	`max`	20.6k	82.89% (80.52%–85.23%)
GPT-6.1 Terra	`none`	1.65k	52.98% (49.3%–56.75%)
GPT-6.1 Terra	`low`	2.04k	57.4% (53.85%–61.03%)
GPT-6.1 Terra	`medium`	3.65k	68.11% (65.1%–71.02%)
GPT-6.1 Terra	`high`	6.19k	72.7% (69.84%–75.41%)
GPT-6.1 Terra	`xhigh`	11.4k	75.1% (72.39%–77.74%)
GPT-6.1 Terra	`max`	33.7k	77.56% (74.98%–79.95%)

Table S1: Numerical values corresponding to Figure 2A–B. Each row gives mean total output-token use (reasoning and final answer) per response and mean score with its pointwise 95% task-bootstrap interval. The analysis includes 60,390 graded responses across 61 model–effort conditions and all 99 tasks. Every condition contains ten retained responses per task. Scores are normalized by each task's rubric budget, averaged within each task, and then averaged equally across tasks. Token means use the same retained responses; unsuccessful or unselected recovery attempts and judge calls are excluded. Intervals use 5,000 whole-task bootstrap resamples. Token counts are shown in thousands and rounded to three significant figures; scores and interval bounds are rounded to four significant figures.

Model	Effort	Mean score (95% interval) C. Molecular structure (SMILES) (n=92)	D. Scientific rationale (n=91)	E. Ranking and candidate selection (n=86)	F. Numerical prediction (n=21)
GPT-5.2	none	1.051% (0.08452%–2.777%)	44.05% (40.65%–47.48%)	35.44% (28.84%–42.33%)	31.42% (22.02%–42.04%)
GPT-5.2	low	1.163% (0.188%–2.515%)	48.14% (44.75%–51.51%)	38.44% (32.01%–45.21%)	37.25% (28.31%–47.26%)
GPT-5.2	medium	1.206% (0.4727%–2.242%)	51.57% (48.06%–55.02%)	41.62% (35.38%–47.7%)	37.21% (28.07%–46.98%)
GPT-5.2	high	1.869% (0.7133%–3.519%)	54.52% (51.12%–57.89%)	43.16% (36.64%–49.52%)	38.09% (29.3%–47.22%)
GPT-5.2	xhigh	8.735% (5.142%–12.85%)	59.51% (56.01%–63%)	46.81% (40.06%–53.64%)	39.99% (31.68%–49.4%)
GPT-5.4	none	1.117% (0.2312%–2.55%)	48.89% (45.42%–52.29%)	37.07% (30.45%–44.14%)	40.75% (32.27%–49.61%)
GPT-5.4	low	1.77% (0.5555%–3.36%)	53.03% (49.81%–56.12%)	41.61% (34.81%–48.34%)	40.04% (31.33%–49.27%)
GPT-5.4	medium	4.579% (2.179%–7.498%)	55.89% (52.42%–59.2%)	45.76% (38.62%–52.63%)	40.48% (31.85%–49.85%)
GPT-5.4	high	7.659% (4.451%–11.49%)	56.75% (53.45%–60.1%)	45.75% (38.78%–52.47%)	40.2% (31.47%–49.68%)
GPT-5.4	xhigh	11.67% (7.746%–16.14%)	58.52% (55.06%–61.94%)	47.88% (40.82%–54.45%)	40.77% (31.47%–50.62%)
GPT-5.5	none	7.405% (4.308%–10.92%)	56.13% (52.71%–59.51%)	43.14% (36.84%–49.46%)	40.33% (32.4%–48.9%)
GPT-5.5	low	9.941% (6.425%–13.86%)	59.52% (56.43%–62.62%)	47.32% (41.36%–53.6%)	40.56% (31.57%–50.43%)
GPT-5.5	medium	21.56% (15.96%–27.4%)	64.82% (61.68%–67.93%)	52.27% (45.92%–58.47%)	43.09% (34.09%–52.74%)
GPT-5.5	high	26% (20.19%–32.03%)	66.99% (63.73%–70.28%)	51.95% (45.6%–58.12%)	45.96% (37.74%–55.41%)
GPT-5.5	xhigh	28.7% (22.42%–34.87%)	66.94% (63.67%–70.02%)	52.78% (46.42%–59.15%)	44.93% (35.25%–54.58%)
GPT-5.6 Luna	none	0.8389% (0.04831%–1.905%)	38.2% (35.16%–41.31%)	29.65% (24.12%–35.34%)	31.64% (22.75%–41.77%)
GPT-5.6 Luna	low	1.417% (0.4374%–2.696%)	42.39% (39.42%–45.47%)	36.83% (31.24%–42.89%)	34.64% (25.51%–44.36%)
GPT-5.6 Luna	medium	3.567% (1.633%–5.867%)	46.73% (43.45%–49.79%)	39.53% (33.58%–45.46%)	35.17% (25.69%–45.36%)
GPT-5.6 Luna	high	11.51% (7.324%–16.16%)	54.54% (50.97%–58.07%)	43.09% (36.98%–49.56%)	36.59% (26.43%–47.37%)
GPT-5.6 Luna	xhigh	15.31% (10.42%–20.56%)	57.88% (54.12%–61.57%)	43.58% (37.14%–50.25%)	37.69% (28.1%–47.81%)
GPT-5.6 Luna	max	17.67% (12.33%–23.54%)	60.63% (57.01%–64.05%)	45.66% (39.08%–52.39%)	38.59% (28.94%–49.11%)
GPT-5.6 Terra	none	1.642% (0.4831%–3.089%)	47.19% (44.1%–50.26%)	39.49% (33.18%–45.65%)	36.36% (27.55%–46.05%)
GPT-5.6 Terra	low	4.457% (2.031%–7.399%)	53.17% (49.59%–56.77%)	44.42% (38.53%–50.42%)	37.17% (27.97%–47.08%)
GPT-5.6 Terra	medium	5.112% (2.547%–8.131%)	54.25% (50.57%–57.72%)	44.14% (37.73%–50.57%)	38.46% (29.42%–48.35%)
GPT-5.6 Terra	high	8.924% (5.55%–12.88%)	57.37% (53.74%–61.05%)	47.79% (41.37%–54.07%)	39.09% (29.51%–49.27%)
GPT-5.6 Terra	xhigh	17% (12.14%–22.5%)	60.46% (56.74%–64.01%)	48.69% (41.88%–55.02%)	40.9% (31.82%–50.46%)
GPT-5.6 Terra	max	21.23% (15.92%–27.02%)	61.3% (57.7%–65.01%)	48.85% (42.14%–55.57%)	41.75% (32.99%–51.16%)
GPT-5.6 Sol	none	9.501% (5.704%–13.84%)	59.49% (55.73%–63.15%)	45.11% (39.25%–51.05%)	39.58% (30.8%–48.82%)
GPT-5.6 Sol	low	25.74% (20.2%–31.69%)	64.04% (60.51%–67.73%)	48.78% (42.72%–54.71%)	38.63% (29%–49.06%)
GPT-5.6 Sol	medium	35.6% (29.35%–41.95%)	68.04% (64.35%–71.64%)	49.08% (42.84%–55.24%)	40.73% (31.43%–50.35%)
GPT-5.6 Sol	high	43.79% (37.05%–50.71%)	72.42% (69.05%–75.65%)	50.5% (44.46%–56.5%)	42.6% (34.14%–51.98%)
GPT-5.6 Sol	xhigh	47.31% (39.98%–54.43%)	74.72% (71.14%–78.2%)	50.01% (43.83%–56.16%)	42.85% (33.6%–52.64%)
GPT-5.6 Sol	max	48.64% (41.56%–55.81%)	76.41% (72.95%–79.8%)	50.8% (44.69%–56.81%)	44.46% (34.94%–54.47%)
GPT-6 Astra	low	71.83% (65.27%–78.08%)	87.43% (85.07%–89.68%)	47.61% (40.64%–54.44%)	44.95% (35.21%–54.29%)
GPT-6 Astra	medium	89.13% (85.24%–92.53%)	89.67% (87.53%–91.79%)	49.16% (42.41%–55.98%)	45.37% (35.58%–55.14%)
GPT-6 Astra	high	95.19% (92.5%–97.63%)	90.49% (88.46%–92.46%)	51.52% (44.57%–58.46%)	45.36% (35.26%–55.61%)
GPT-6 Astra	xhigh	95.37% (92.58%–97.63%)	91.05% (89.02%–92.96%)	50.42% (43.76%–57.03%)	47.02% (37.12%–56.95%)
GPT-6 Astra	max	95.97% (93.5%–97.99%)	91.74% (89.88%–93.54%)	50.44% (43.44%–57.06%)	47.57% (38.08%–57.32%)
GPT-6 Sol	none	13.68% (9.012%–18.84%)	51.21% (47.61%–54.91%)	43.93% (37.63%–50.21%)	43.53% (35.27%–52.37%)
GPT-6 Sol	low	35.05% (28.39%–41.71%)	56.22% (52.45%–60.07%)	47.69% (41.54%–54.27%)	42.59% (33.51%–52.36%)
GPT-6 Sol	medium	51.87% (44.57%–58.9%)	61.69% (58.25%–65.18%)	48.78% (42.16%–55.24%)	44.57% (35.14%–54.31%)
GPT-6 Sol	high	60.97% (53.99%–67.75%)	67.06% (63.7%–70.38%)	49.57% (42.84%–56.07%)	45.05% (35.82%–55.19%)
GPT-6 Sol	xhigh	65.85% (58.88%–72.3%)	71.61% (68.36%–74.82%)	50.74% (44.39%–56.9%)	43.5% (33.74%–52.97%)
GPT-6 Sol	max	72.78% (66.17%–78.91%)	75.3% (72.2%–78.31%)	51.63% (45.35%–58.03%)	45.08% (34.89%–55.19%)
GPT-6 Luna	none	1.03% (0.1594%–2.304%)	30.15% (26.92%–33.34%)	32.47% (26.29%–38.97%)	31.18% (22.79%–40.92%)

Continued on the next page.

Table S2 continued from the previous page.

Model	Effort	C. Structure	D. Rationale	E. Ranking	F. Numerical
GPT-6 Luna	`low`	1.787% (0.3257%–3.641%)	31.33% (27.72%–35.09%)	33.08% (27.18%–39.41%)	27.33% (18.52%–37.96%)
GPT-6 Luna	`medium`	5.629% (3.169%–8.618%)	40.05% (36.07%–43.94%)	42.38% (35.93%–48.79%)	32.48% (23%–43.4%)
GPT-6 Luna	`high`	10.54% (6.73%–14.83%)	46.6% (42.52%–50.89%)	44.54% (37.66%–51.44%)	36.67% (27.53%–46.56%)
GPT-6 Luna	`xhigh`	12.42% (8.267%–16.79%)	50.94% (46.86%–55.09%)	45.03% (38.32%–51.71%)	37.85% (28.27%–48.05%)
GPT-6 Luna	`max`	20.2% (14.76%–25.99%)	55.64% (51.49%–59.79%)	45.51% (38.57%–52.56%)	39.99% (29.96%–50.47%)
GPT-6.1 Sol	`low`	65.63% (58.75%–72.33%)	84.28% (81.4%–86.98%)	48.61% (41.52%–55.87%)	45.6% (35.14%–55.72%)
GPT-6.1 Sol	`medium`	82.51% (77.39%–87.28%)	87% (84.38%–89.57%)	50.04% (43%–56.49%)	45.5% (35.22%–55.75%)
GPT-6.1 Sol	`high`	92.18% (88.97%–95.01%)	88.36% (86.15%–90.42%)	50.76% (44.27%–57.25%)	44.73% (34.43%–55.05%)
GPT-6.1 Sol	`xhigh`	93.7% (90.88%–96.08%)	89.68% (87.5%–91.73%)	50.39% (43.53%–57.22%)	45.79% (35.84%–55.63%)
GPT-6.1 Sol	`max`	95.02% (92.39%–97.15%)	90.02% (88.01%–91.99%)	50.64% (44.17%–57.44%)	46.76% (37.38%–56.76%)
GPT-6.1 Terra	`none`	45.31% (38.07%–52.41%)	65.84% (62.21%–69.47%)	51.07% (44.23%–57.65%)	40.48% (30.48%–50.96%)
GPT-6.1 Terra	`low`	55.86% (49.06%–62.42%)	64.11% (60.43%–67.72%)	51.74% (45.34%–57.9%)	40.58% (30.29%–51.06%)
GPT-6.1 Terra	`medium`	79.86% (74.44%–84.7%)	66.35% (62.6%–70.03%)	50.22% (43.83%–56.63%)	41.69% (31.38%–52.19%)
GPT-6.1 Terra	`high`	88.52% (84.17%–92.33%)	68.94% (65.35%–72.69%)	52.11% (45.22%–58.59%)	42.2% (31.93%–52.68%)
GPT-6.1 Terra	`xhigh`	92.18% (88.56%–95.23%)	71.02% (67.37%–74.67%)	51.78% (45.03%–58.65%)	44.83% (35.37%–54.6%)
GPT-6.1 Terra	`max`	94.49% (90.83%–97.26%)	75.25% (71.88%–78.45%)	52.09% (45.64%–58.78%)	45.55% (35.61%–55.95%)

Table S2: Numerical values for the component performance points in Figure 2C–F. Component points are normalized by the available points for that component within each task, averaged across its responses, and then averaged equally across applicable tasks. All conditions contain every applicable task; n gives the number of such tasks. Every condition contains ten retained responses per task. Bounds are pointwise 95% percentile intervals from 5,000 whole-task bootstrap resamples. Values are rounded to four significant figures.