BioSecBench-RespiratoryVirus-Countermeasures

Benchmarking AI agents on disease biology and countermeasures for respiratory viruses

Harmon Bhasin[1], Evan Seeyave[1], Gage Moreno[1], Yuyi Liu[1], Tristan Russell[1], Arya Kaul[1], Ioanna Pagani[1], Lenora Kepler[1], Reema Singh[1], Marcos Parras Molto[1], Tyler K. Chafin[1], Konstantinos Konstantinidis[1], Karl Widney[1], Lucas Yago Melo Ferreira[1], Mikail Kelleci[1], Paul Oluniyi[1], Sara Zufan[1], Cheikh Cambel Dieng[1], Faisal Al-Zaben[1], Naomi Jaramillo[1], Zachary Mudge[1], Dianzhuo Wang[1], Kenny Workman[1]

[1]LatchBio, San Francisco, CA, USA

Correspondence: harmon@latch.bio, kenny@latch.bio

Abstract

Recurrent epidemics and pandemic threats from respiratory viruses demand a deeper understanding of disease mechanisms and rigorous ways to evaluate vaccines and therapeutics. AI agents could support the molecular, cellular, and population analyses underlying this work, but their reliability across these tasks remains uncertain. We introduce BioSecBench-RespiratoryVirus-Countermeasures, a benchmark of 64 evaluations spanning viral structure and function, immunity and antigenic variation, host response and disease biology, antiviral activity and resistance, and clinical and population outcomes. Each evaluation pairs a scientific question and source materials with task-specific grading criteria for numerical, categorical, or structured outputs. Across 22 model–harness configurations spanning 18 models, the highest observed overall pass rate was 30.2% (Gemini 4 Carbon / PI), followed by Claude Sonnet 5.5 / PI (18.8%) and Claude Fable 5.2 / Claude Code (17.7%). Seven models led or tied for the lead in at least one of the 17 displayed evidence-source, decision-type, or failure-class groups, and the highest mean pass rate among the seven failure classes was 10.6%. These results identify limitations in current agents' reliability on the evaluated tasks and provide a baseline for assessing future improvements.

Introduction

Effective vaccines and therapeutics against respiratory viruses remain an urgent need, as these viruses drive recurrent epidemics and can spark pandemics [1, 2]. Developing and evaluating them rests on several kinds of evidence: how viruses cause disease, how immunity provides protection, whether candidate compounds have antiviral activity, and whether interventions improve outcomes. For example, structural biologists use viral protein structures to inform vaccine and drug design [3, 4], while immunologists characterize immune responses and antigenic variation [5, 7]. Pharmacologists evaluate antiviral activity and resistance [4, 8], and clinical researchers and epidemiologists assess whether interventions reduce infection and severe disease [9, 10].

AI agents that combine planning, specialized tools, and code execution could support these analyses [11]. Existing benchmarks evaluate agents on biological data analysis [12] and therapeutic discovery and assessment [13, 14, 15], but their reliability across respiratory-virus disease biology and countermeasure analyses remains uncertain. To assess this, we introduce BioSecBench-RespiratoryVirus-Countermeasures, a benchmark of 64 evaluations spanning viral structure and function, immunity and antigenic variation, host response and disease biology, antiviral activity and resistance, and clinical and population outcomes. Each evaluation asks agents to answer a scientific question and assesses their outputs against task-specific grading criteria. Across 22 model–harness configurations spanning 18 models, the highest observed overall pass rate on these tasks was 30.2%. By exposing failures in how agents interpret and verify scientific evidence, this work can guide improvements toward dependable AI assistance in respiratory virus research and countermeasure development.

Benchmark Construction

The benchmark contains 64 evaluations across five scientific areas: viral structure, function, and engineering (11); immunity and antigenic variation (16); host response and disease biology (13); antiviral activity and resistance (11); and clinical and population outcomes (13). Tasks concern SARS-CoV-2, influenza, respiratory syncytial virus, and other respiratory viruses. They draw on sequencing, structural and functional measurements, transcriptomics, and clinical and population records. Figure 2 summarizes the cohort by scientific area, viral surveillance targets, evidence source, and requested decision.

Each evaluation asks agents to answer a scientific question using a published dataset available online. We tested 22 model–harness configurations spanning 18 models, with three offline runs per evaluation using PI, Claude Code, and Codex. We graded the requested outputs using task-specific numerical tolerances, categorical matching, and structured-output checks. We report pass rates overall and by scientific category, evidence source, decision type, and observed failure class.

Results

Leaderboard and efficiency

The highest observed mean pass rate was 30.2% (Gemini 4 Carbon / PI), and all other model–harness configurations scored below 20% (Figure 1A). Claude Sonnet 5.5 / PI followed at 18.8%, then Claude Fable 5.2 / Claude Code at 17.7%. These three models accounted for the five highest-scoring configurations, with Sonnet and Fable each represented under both PI and Claude Code.

Among the four models evaluated under two harnesses, the higher-scoring harness differed by model. Claude Code exceeded PI by 0.8 percentage points for Claude Fable 5.2 and 1.0 point for Claude Opus 5.5. Conversely, PI exceeded Claude Code by 2.1 points for Claude Sonnet 5.5 and Codex by 3.6 points for GPT-6 Astra.

The highest-scoring configuration used fewer output tokens and tool calls than the runner-up, but cost more and took longer (Figure 1B–E). Among completed, non-refused runs, Gemini 4 Carbon / PI averaged 84.8 thousand output tokens and 25.9 tool calls, compared with 142.3 thousand tokens and 46.7 calls for Claude Sonnet 5.5 / PI. Carbon's mean reported cost and runtime were $4.96 and 72.5 minutes per run, compared with $2.97 and 41.6 minutes for Sonnet.

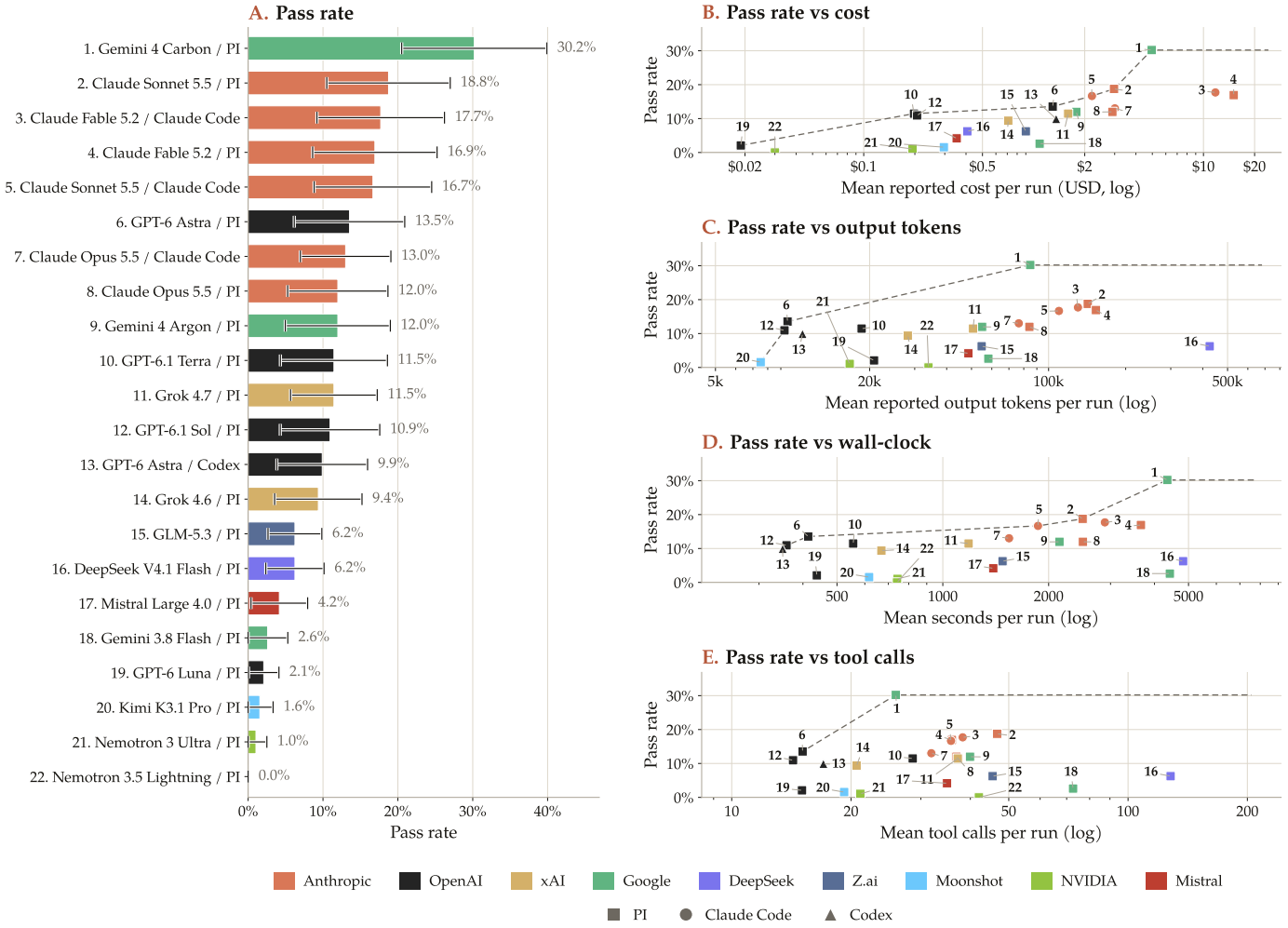


Figure 1: Pass rate by configuration. (A) Mean pass rate over the 64 evaluations, with 95% Student's t intervals calculated across per-evaluation pass rates; color identifies the model's lab. Refusals and errors count as failures. (B–E) Pass rate against mean reported cost, output tokens, wall-clock time and tool calls per completed, non-refused run, on log scales. The dashed line marks the efficiency frontier.

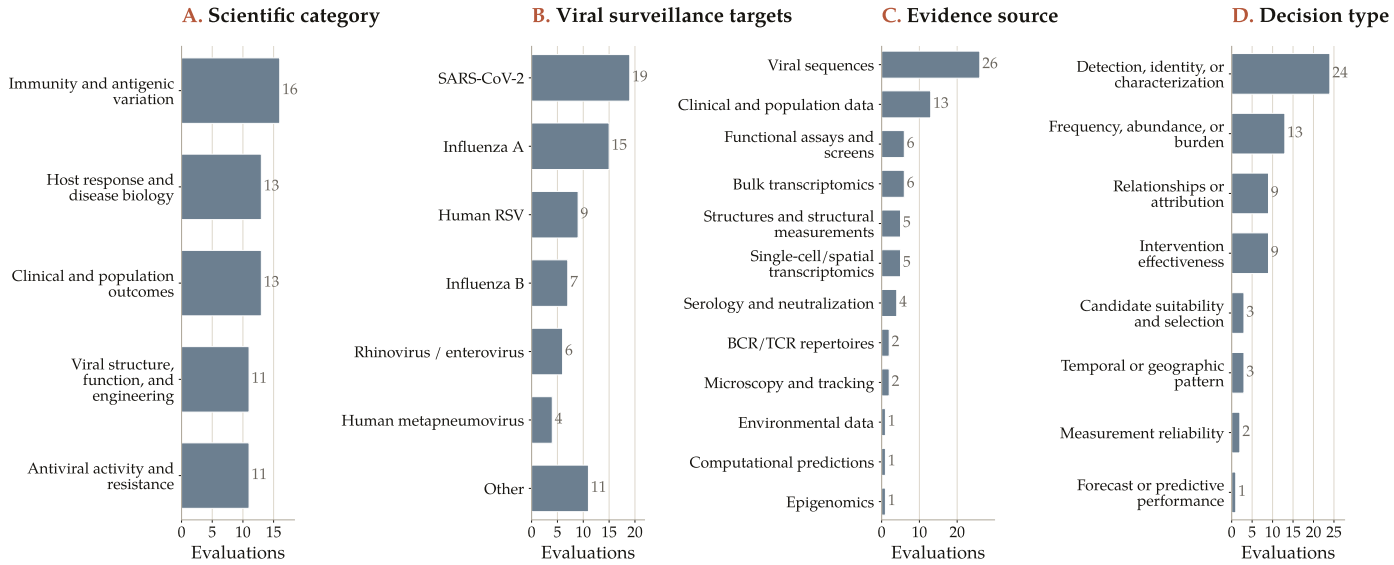


Figure 2: Benchmark anatomy. All 64 selected evaluations by (A) scientific category, the area of disease biology or countermeasure analysis; (B) viral surveillance targets, named viruses or defined surveillance groups; (C) evidence source, the types of observations or measurements analyzed; and (D) decision type, the requested analytical endpoint. Categories and decision types have one label per evaluation. Targets and evidence sources permit multiple labels, with each evaluation counted once per applicable label. Other includes targets represented in only one evaluation, unresolved viral targets, and evaluations without a virus-specific endpoint.

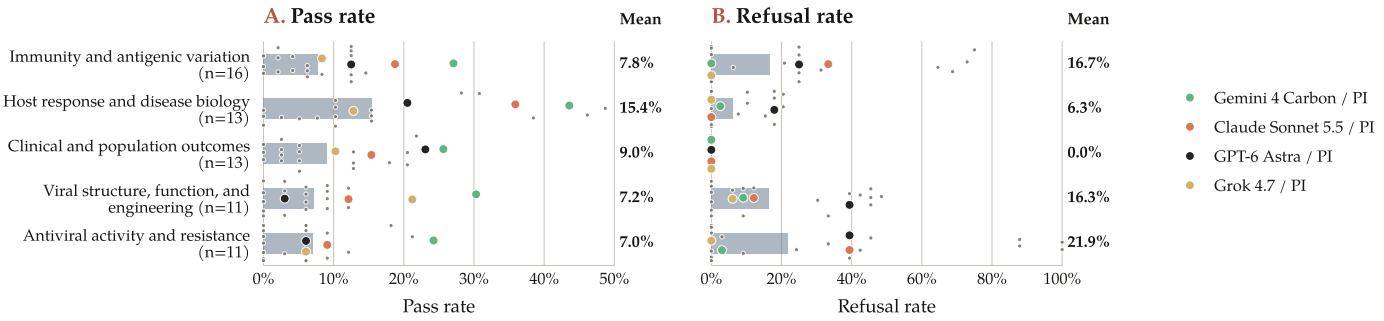


Figure 3: Pass and refusal rates by scientific category. (A) Pass rate. (B) Refusal rate. Bars show the mean across 18 models, pooling available harness counts within each model and evaluation before averaging evaluation rates. Dots show individual model–harness configurations. In both panels, color highlights the best configuration from each of the top four labs, ranked by overall pass rate; other configurations are gray. Category labels give evaluation counts. Refusals and errors count as failures in A.

Performance by category

Mean pass rates were low across all five scientific categories (7.0–15.4%), with the highest mean in host response and disease biology (Figure 3A). The remaining four categories had mean pass rates of 7.0–9.0%. Even within host response and disease biology, the highest configuration-level pass rate was 48.7%, achieved by Claude Fable 5.2 / Claude Code.

Low pass rates persisted in scientific categories with few or no refusals (Figure 3A,B). Clinical and population outcomes had a mean pass rate of 9.0% despite having no refusals. Host response and disease biology had a mean pass rate of 15.4% and a mean refusal rate of 6.3%. Refusal rates were higher in the remaining three categories: 21.9% for antiviral activity and resistance, 16.7% for immunity and antigenic variation, and 16.3% for viral structure, function, and engineering.

Across evidence sources and decision types, Gemini 4 Carbon led or tied for the highest pass rate in four of six evidence-source groups and three of four decision-type groups (Figure 4). Claude Fable 5.2 led on bulk transcriptomics and tasks assessing relationships or attribution. On functional assays and screens, however, Mistral Large 4.0 scored 22.2%, compared with 5.6% for Gemini 4 Carbon.

Performance was highest on transcriptomic evaluations among the displayed evidence-source groups (Figure 4). Mean pass rates were 16.5% for bulk transcriptomics and 13.5% for single-cell/spatial transcriptomics, compared with 6.0–9.0% across the other four groups. The highest model scores in the two transcriptomic groups were 55.6% and 53.3%, respectively; no model exceeded 26.7% in any other evidence-source group.

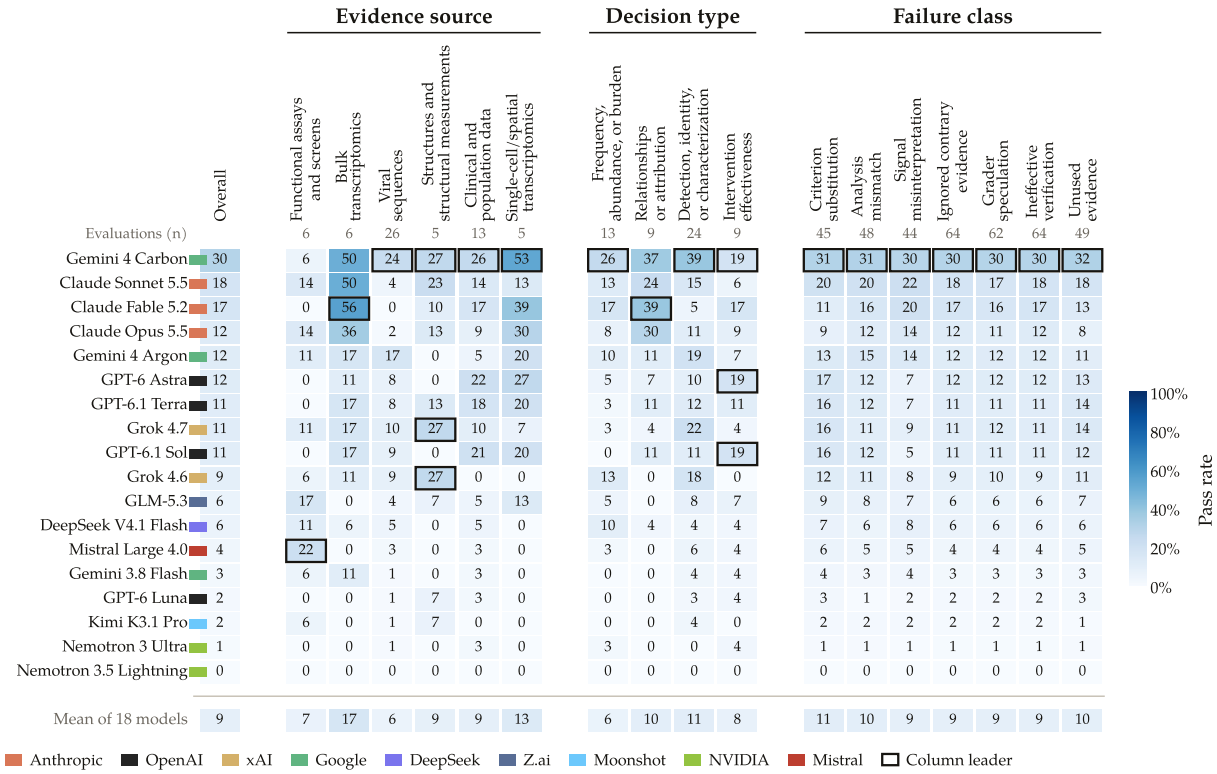


Figure 4: Pass rate by evidence source, decision type and failure class. All 18 models, with available harnesses pooled within each evaluation, ordered by overall pass rate. Evidence sources and failure classes can overlap; decision type assigns one label per evaluation. Each evaluation belongs to every failure class observed in its reviewed failed runs. Outlined cells mark column leaders, including ties, and the bottom row is the mean over models. Columns with fewer than five selected evaluations are omitted and cells with fewer than five observed evaluations are blank. Refusals and errors count as failures.

Failure decomposition

Mean pass rates were below 11% across all seven overlapping failure classes (Figure 5). Group means ranged from 9.1% for signal misinterpretation to 10.6% for criterion substitution. In every group, at least 15 of the 18 models had pass rates below 20% (Figure 4).

Gemini 4 Carbon led every failure class, but its pass rates remained below 32% (Figure 4). Claude Sonnet 5.5 ranked second in every group, with pass rates of 16.9–21.6%. Carbon's lead over Sonnet ranged from 8.0 percentage points for signal misinterpretation to 14.3 points for unused evidence.

Failures to act on contrary evidence and verify conclusions occurred throughout the benchmark. Ignored contrary evidence and ineffective verification each appeared in failed runs from all 64 evaluations, whereas the other five classes appeared in 44–62 evaluations (Figure 5). Agents sometimes submitted an answer despite having identified a better approach or a reason to doubt their result. In other cases, agents tried to verify their answers by repeating calculations with the same faulty assumptions, or omitted contradictory results from the final report.

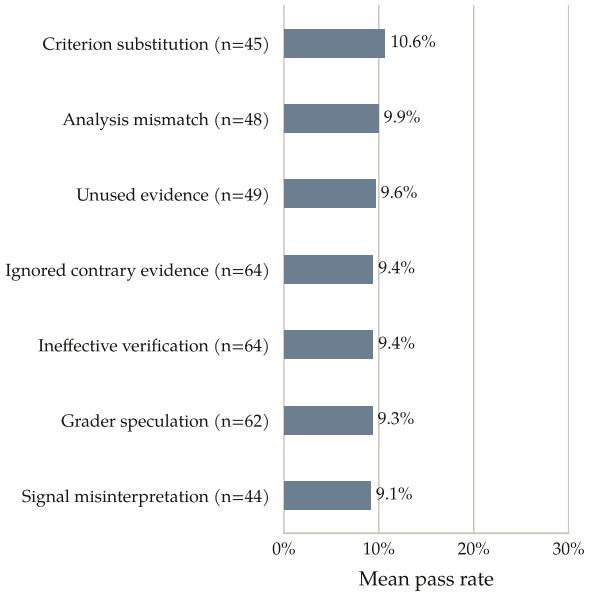


Figure 5: Mean pass rate by failure class. Mean over the 18 models of each model's pass rate on the evaluations in the class, with available harnesses pooled within each evaluation. Evaluations can belong to multiple classes; n is the number of evaluations in each class. Refusals and errors count as failures.

Discussion

BioSecBench-RespiratoryVirus-Countermeasures measures whether AI agents can make the analytical decisions required to study diseases caused by respiratory viruses and evaluate countermeasures against them. Across 64 evaluations, the highest observed overall pass rate was 30.2%, indicating that reliable completion remains difficult on these tasks. Refusals contributed to low scores but did not fully explain them: clinical and population outcomes had a mean pass rate of 9.0% despite having no recorded refusals. Review of failed executions also showed that some agents submitted answers despite recognizing problems with their analyses or computing better-supported alternatives. In other cases, agents tried to verify their answers by repeating calculations that relied on the same faulty assumptions. These observations motivate evaluating whether agents can use contradictory evidence to revise their conclusions and perform checks capable of detecting errors in their analyses.

A main limitation in this benchmark is the number and breadth of tasks within each scientific category. Each category contains 11–16 evaluations, and coverage was constrained by the availability of suitable published datasets. Some respiratory viruses, evidence sources, and analytical decisions are therefore less represented than others. More tasks drawn from different datasets would help establish whether observed weaknesses recur across analyses or reflect difficulties specific to individual evaluations. Broader viral coverage would also allow us to examine whether agents can apply similar analytical methods across biological systems with different properties.

This benchmark provides a tool for researchers assessing AI agents for respiratory virus disease biology and countermeasure evaluation. Task-level results show how agents perform on specific scientific analyses, while failure analysis characterizes the errors observed in unsuccessful runs. As agents improve, the benchmark can help researchers assess whether those advances translate into more reliable analyses of disease biology, immunity, vaccines, and therapeutics in respiratory virus research.

References

1. Morens DM, Taubenberger JK, Fauci AS. Rethinking next-generation vaccines for coronaviruses, influenzaviruses, and other respiratory viruses. *Cell Host & Microbe*. 2023;31:146–157.
2. Cassetti MC, Pierson TC, Patterson LJ, et al. Prototype pathogen approach for vaccine and monoclonal antibody development: a critical component of the NIAID plan for pandemic preparedness. *Journal of Infectious Diseases*. 2023;227:1433–1441.
3. McLellan JS, Chen M, Joyce MG, et al. Structure-based design of a fusion glycoprotein vaccine for respiratory syncytial virus. *Science*. 2013;342:592–598.
4. Owen DR, Allerton CMN, Anderson AS, et al. An oral SARS-CoV-2 M^{pro} inhibitor clinical candidate for the treatment of COVID-19. *Science*. 2021;374:1586–1593.
5. Stephenson E, Reynolds G, Botting RA, et al. Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine*. 2021;27:904–916.
6. Planas D, Saunders N, Maes P, et al. Considerable escape of SARS-CoV-2 Omicron to antibody neutralization. *Nature*. 2022;602:671–675.
7. Smith DJ, Lapedes AS, de Jong JC, et al. Mapping the antigenic and genetic evolution of influenza virus. *Science*. 2004;305:371–376.
8. Iketani S, Mohri H, Culbertson B, et al. Multiple pathways for SARS-CoV-2 resistance to nirmatrelvir. *Nature*. 2023;613:558–564.
9. Hammond J, Leister-Tebbe H, Gardner A, et al. Oral Nirmatrelvir for High-Risk, Nonhospitalized Adults with Covid-19. *New England Journal of Medicine*. 2022;386:1397–1408.
10. Dagan N, Barda N, Kepten E, et al. BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting. *New England Journal of Medicine*. 2021;384:1412–1423.
11. Huang K, Zhang S, Wang H, et al. Autonomous biomedical research with an artificial intelligence agent. *Science*. 2026;393:eadz4351.
12. Koch Z, Wassie AT, Valdes-Aleman J, et al. BixBench3: Benchmarking AI agents on research-study-scale computational biology tasks. arXiv:2608.25286, 2026. Preprint.
13. Le H, Ramasamy R, Urrutia A, et al. TxBench-PP: Analyzing AI Agent Performance on Small-Molecule Preclinical Pharmacology. arXiv:2606.19245, 2026. Preprint.
14. Ramasamy R, Le H, Brougher J, et al. TxBench–Antibody Discovery: Benchmarking AI Agents on Experimentally Grounded Decisions in Therapeutic Antibody Discovery. LatchBio, 2026. Research manuscript.
15. Akyürek AF, Tu X, Gutmanstein A, et al. DrugDiscoveryBench: Can Coding Agents Assist Early-Stage Drug Discovery? Scale AI and Phylo, 2026. Technical report.