AI News Digest

11/09/2026 - 11/15/2026

WEEKLY DIGEST: BIOLOGICAL & BIOMEDICAL SCIENCES

AI Agent Benchmarks Redefine Biomedical Data Evaluation

A wave of new benchmarks is measuring how well AI agents can reason over real biomedical data, from pathogen structures to single-cell genomics, tissue microscopy, viral countermeasures, and neuroscience recordings. BioSecBench-Structure and the Respiratory Virus Countermeasures suite test structural and virological reasoning, while AlphaGenome 2 leverages over 100 million single-cell profiles to predict variant effects across dozens of tissues. Early results reveal both impressive strides-such as multi-cell-type expression prediction-and stark gaps, with only a small fraction of runs achieving high accuracy on complex imaging and neuroscience tasks. These efforts aim to create shared standards that will accelerate the development of truly research-ready AI agents.

KEY FINDINGS

✓ BioSecBench-Structure introduces 58 pathogen-related structural biology tasks, evaluated across 20 models and 35 configurations.

✓ AlphaGenome 2 is trained on more than 100 million single-cell profiles from 68 human tissues, enabling variant-impact predictions in hundreds of cell types.

✓ Tissue-MicroscopyBench comprises 49 visual-recognition and spatial-reasoning tasks; among 4,151 graded runs, only 13.2% achieve top-tier performance.

IMPLICATIONS

→ Standardized, multimodal benchmarks will push AI developers toward models that can integrate experimental data, not just recall facts.

→ Accurate, cell-type-specific variant effect prediction could speed discovery of disease mechanisms and personalized therapeutic targets.

→ Low success rates on imaging and neuroscience tasks highlight the need for stronger multimodal reasoning and better data-fusion architectures before AI agents can be trusted in critical biomedical workflows.

> Key Metrics: 58 structural biology tasks (BioSecBench-Structure); 20 AI models and 35 configurations evaluated; >100 million single-cell profiles across 68 tissues (AlphaGenome 2); 49 tissue-microscopy tasks, 30 model-harness configurations; 4,151 graded runs; 13.2% high-performance rate; 64 respiratory-virus countermeasure questions; 91 evaluations from 43 neuroscience studies across 34 configurations

1. BioSecBench-Structure: Can AI Agents Analyze Molecular Structures?

Category: Research
Departments: Computer Science, Biological Sciences
Topics: Science & Research, Molecular Structure Reasoning, Density Map Interpretation, Reference Structure Selection, AI Refusal Mitigation

Impact Score: 28.0
Rubric Scores

Research Relevance: 5	Educational Value: 4	Innovation / Novelty: 4
Practical Impact: 5	Interdisciplinary Potential: 5	Ethical / Policy Implications: 5

The paper introduces BioSecBench-Structure, a benchmark of 58 pathogen-related structural biology tasks designed to evaluate AI reasoning on experimental data such as crystallography, cryo-EM, predicted structures, and HDX-MS. Across 20 models and 35 configurations, the highest overall pass rate was 43.7% achieved by the Fable 5.2 / Claude Code configuration, with other top scores ranging from 31.0% to 41.4% and confidence intervals overlapping substantially. Disabling model refusals increased countermeasure task performance by 14-15 percentage points for Fable 5.2 and Opus 5.5, while pathogen-characterization scores changed little, highlighting persistent errors in density interpretation and reference-structure selection. The benchmark also reports resource usage, showing wide cost and time variations among models.

https://blog.latch.bio/p/biosecbench-structure-can-ai-agents

2. AlphaGenome 2: Deciphering how cells read DNA to better understand disease

Category: Research
Departments: Biological Sciences, Computer Science
Topics: Science & Research, Variant Effect Prediction, Single-cell Genomics Modeling, Genomic Language Model

Impact Score: 27.0
Rubric Scores

Research Relevance: 5	Educational Value: 4	Innovation / Novelty: 5
Practical Impact: 5	Interdisciplinary Potential: 5	Ethical / Policy Implications: 3

AlphaGenome 2 is a next-generation DNA sequence model that predicts the impact of genetic variants on gene expression in hundreds of human cell types and states. It is trained on more than 100 million single-cell profiles from 68 tissues, enabling cell-type-specific variant effect predictions unavailable to prior broad-tissue models. Compared with the original AlphaGenome, the new model identifies roughly twice as many expression-changing variants and shows a step-change improvement in predictive performance. AlphaGenome 2 is released for non-commercial research via an API and in private preview for commercial use on Google Cloud.

https://deepmind.google/blog/alphagenome-2-deciphering-how-cells-read-dna-to-better-understand-disease/

3. Tissue-MicroscopyBench: understanding how well AI agents work with real tissue images

Category: Research
Departments: Computer Science, Biological Sciences
Topics: Computer Vision, Spatial Reasoning, Rare Feature Detection, Agent Benchmarking, Cost-Performance Correlation

Impact Score: 24.0
Rubric Scores

Research Relevance: 5	Educational Value: 4	Innovation / Novelty: 4
Practical Impact: 4	Interdisciplinary Potential: 5	Ethical / Policy Implications: 2

TissueMicroscopyBench is a new benchmark comprising 49 visual recognition and spatial-reasoning tasks drawn from four public tissue-microscopy datasets, each evaluated in triplicate across 30 model-harness configurations. Across 4,151 graded runs, only 13.2 % of attempts passed, with the highest pass rate of 34.0 % achieved by Fable 5.2 on the Pi harness; Claude-based models were the next best performers. Agents performed better on large, visually distinct structures and struggled with small, rare features and multi-step spatial reasoning, showing high variability-Fable 5.2 (Pi) solved 43 % of tasks at least once but only 22 % on all three runs. Pass rates increased moderately with cost (Spearman ρ = 0.61), while average partial scores for frontier LLMs ranged between 52 % and 64 %.

https://blog.latch.bio/p/tissue-microscopybench-understanding

4. Can AI Agents Evaluate Vaccines and Treatments for Respiratory Viruses?

Category: Research
Departments: Computer Science, Biological Sciences, Pharmaceutical Sciences, Public Health Sciences
Topics: Healthcare AI, AI Agent Benchmarking, Biomedical Question Answering, Antiviral Activity Prediction, Resistance Mechanism Assessment

Impact Score: 23.0
Rubric Scores

Research Relevance: 5	Educational Value: 4	Innovation / Novelty: 4
Practical Impact: 3	Interdisciplinary Potential: 4	Ethical / Policy Implications: 3

The authors introduced BioSecBench-RespiratoryVirus-Countermeasures, a benchmark of 64 questions covering viral structure, immunity, host response, antiviral activity, resistance, and clinical outcomes, each requiring AI agents to analyze source material and meet task-specific grading criteria. They evaluated 18 models across 22 model-harness configurations with three attempts per question (4,224 trials), finding a maximum overall pass rate of 30.2 % for Gemini 4 Carbon on the Pi harness, while all other setups scored below 20 %. Performance was poorest on antiviral activity and resistance tasks (mean pass rate 7.0 %), and cost per run did not consistently predict accuracy. No model achieved high success across all scientific categories.

https://blog.latch.bio/p/can-ai-agents-evaluate-vaccines-and

5. NeuroscienceBench: understanding AI agents' ability to make the right decisions from neuroscience data

Category: Research
Departments: Computer Science, Biological Sciences
Topics: Science & Research, Neuroscience Benchmark, Multimodal Data Interpretation, Causal Reasoning in AI, Omics Data Analysis

Impact Score: 22.0
Rubric Scores

Research Relevance: 4	Educational Value: 3	Innovation / Novelty: 4
Practical Impact: 3	Interdisciplinary Potential: 5	Ethical / Policy Implications: 3

The article introduces NeuroscienceBench, a benchmark comprising 91 evaluations from 43 peer-reviewed neuroscience studies designed to test AI agents' ability to analyze and interpret multimodal real-world data rather than merely recall information. Across 34 model configurations, no agent achieved more than a 33% pass rate, with Claude Opus 5.5 on the Pi harness attaining the highest score of 32.2% and the overall mean pass rate at 21.0%. Performance varied markedly by data type, ranging from 14% on rare omics to roughly 25% on behavior, EEG, and sleep scoring, and errors were predominantly conceptual-such as misreading correlations as causation, overlooking confounders, and falling for quality-control traps. The benchmark highlights the current limitations of frontier AI models in reliably interpreting complex neuroscience datasets.

https://blog.latch.bio/p/neurosciencebench-understanding-ai