AIXBIO · WORKING PAPER · APRIL 25, 2026

# When Plausibility Isn't Enough - Evaluating AI Systems in Biology Beyond Benchmark Performance

Conceptual governance contribution: why benchmark-plausible outputs fail when applied to biology, and what evaluation discipline is required instead.

**Author** S. Jason Prohaska 0009-0008-8254-8411

SOVEREIGN · SELF-PUBLISHED

Author Manuscript · Print-Ready Draft · Not Peer Reviewed

# When Plausibility Isn't Enough: Evaluating AI Systems in Biology Beyond Benchmark Performance

**Author:** S. Jason Prohaska
**ORCID:** 0009-0008-8254-8411
**Date:** April 25, 2026
**Version:** Author Manuscript / Print-Ready Draft

**Non-Peer-Review Notice.** This is an author manuscript / working paper. It has not been peer reviewed, journal accepted, or institutionally endorsed. It presents a

conceptual governance contribution and does not report new empirical experiments, clinical findings, wet-lab results, or regulatory conclusions.


## Abstract

Artificial intelligence systems are increasingly used to generate biological hypotheses, molecular designs, structural predictions, and drug-discovery leads that appear scientifically credible. Their outputs often resemble expert reasoning: they are structured, fluent, internally coherent, and aligned with recognizable biological concepts. Yet apparent plausibility is not equivalent to empirical validation. In biology, this distinction is consequential because AI-generated outputs may influence experimental design, translational strategy, safety assessment, and downstream research investment.

This paper argues that current evaluation practices for AI systems in biology overemphasize benchmark performance while underemphasizing the gap between plausible outputs and validated claims. Benchmarks remain useful for measuring task-specific capability, but they are insufficient for evaluating performance under distribution shift, uncertainty, incomplete data, user interpretation, and real-world workflow conditions. The central risk is not merely that AI systems may be wrong, but that unsupported outputs may be difficult to distinguish from grounded ones when presented with confidence, fluency, and biological plausibility.

The paper introduces the plausibility-validation gap as a core evaluation problem in AIxBio and proposes a shift toward risk-sensitive evaluation, traceability, provenance preservation, and human-system interaction assessment. It concludes that AI evaluation in biology must move beyond isolated model accuracy toward the broader systems in which biological claims are generated, transformed, trusted, and acted upon.



**Keywords:** AIxBio; artificial intelligence in biology; evaluation; biological AI; benchmark performance; plausibility; validation; provenance; traceability; AI risk; human-AI interaction; drug discovery; protein design


## 1. Introduction

Artificial intelligence systems are becoming increasingly capable of generating biological hypotheses, designs, predictions, and explanations that appear scientifically credible. In structural biology, drug discovery, genomics, synthetic biology, and protein design, AI systems can produce outputs that resemble expert reasoning and can be integrated into research workflows.

These advances are significant. AI tools can accelerate search, compress discovery timelines, and surface candidate hypotheses that would otherwise require extensive manual effort. However, there is a growing gap between plausibility and validation. An AI-generated biological claim may look correct without being empirically grounded. It may be internally consistent, fluently explained, and aligned with known concepts while still resting on weak evidence, hidden assumptions, incomplete data, or distributional artifacts.

In biology, this distinction matters because outputs are often not merely informational. They may guide experiments, prioritize molecules, shape safety decisions, or influence the allocation of research resources. This paper argues that current evaluation practices for AI in biology underemphasize this plausibility-validation gap.

The central claim is simple: AI systems in biology must be evaluated not only by whether they produce accurate answers on known tasks, but by whether their outputs remain traceable, appropriately uncertain, and resistant to misinterpretation when embedded in real scientific workflows.

## 2. Background: AIxBio and the Benchmark Paradigm

AIxBio systems are commonly evaluated through curated datasets, held-out test sets, known structures, annotated biological records, retrospective drug-discovery tasks, or benchmark suites. These evaluation methods are valuable because they allow systems to be compared using shared metrics and reproducible tasks. They help answer whether a model predicts a structure accurately, ranks candidates effectively, classifies a sequence correctly, or recovers a known biological relationship.

The benchmark paradigm has supported major advances. Protein structure prediction benefited from community evaluation settings and curated structural datasets. Yet benchmark evaluation also has limitations. A benchmark is not the world. It is a simplified evaluation environment that constrains the task, input space, metric, and definition of success.

When benchmark data are close to training distributions, contaminated by leakage, or derived from similar biological domains, performance estimates may overstate generalization. In drug discovery and molecular design, the gap between computational performance and prospective biological validation is especially important.

The core limitation is not that benchmarks are useless. They are necessary. The problem is that they are often treated as more complete than they are. Benchmark success can show that a system performs well under specified conditions. It does not show that a claim is biologically true, experimentally actionable, safe to operationalize, or robust under novel conditions.

## 3. The Plausibility-Validation Gap

The plausibility-validation gap is the distance between an output that appears scientifically reasonable and a claim that has been adequately supported by evidence.

This gap is especially important in AIxBio because biological plausibility is often easy to imitate but difficult to verify. Biological systems are complex, context-dependent, and noisy. A generated explanation can cite recognizable mechanisms, pathways, binding intuitions, sequence motifs, or structural relationships while still being weakly grounded.

The risk is not simply hallucination in the ordinary sense. The more subtle risk is credible insufficiency: an output may be plausible enough to influence action while lacking the evidentiary support required for that action.

A simple workflow illustrates the problem:

1. A model proposes a drug-target interaction.

2. The output is summarized in a report.
3. The summary is used to guide experimental planning.
4. The experimental plan is presented as if the original claim were more grounded than it was.

At each step, uncertainty can be compressed. Assumptions can disappear. Confidence language can harden. Caveats can be omitted. By the time the claim reaches decision-making, it may no longer be clear whether it originated from validated evidence, model inference, speculative generation, weak analogy, benchmark-correlated pattern recognition, or a mixture of these.

This is not only a model-performance issue. It is a claim-lifecycle issue. The risk emerges as an AI-generated output moves through documents, summaries, presentations, human interpretation, and organizational workflows.

## 4. Limits of Current Evaluation Practices

Most evaluation frameworks focus on accuracy within predefined tasks. This is appropriate for measuring narrow performance, but insufficient for assessing real-world biological risk.

### 4.1 Distribution Shift

A model may perform well when test inputs resemble training data but degrade when inputs differ. In biology, distribution shift may arise from novel protein families, rare variants, underrepresented organisms, unfamiliar assay conditions, unusual molecular scaffolds, sparse disease contexts, or experimental settings not reflected in curated datasets.

Benchmark performance under distributional similarity should not be mistaken for robustness under novelty. Many valuable use cases involve the unknown, rare, or underexplored cases where benchmark similarity is weakest.

### 4.2 Noisy, Incomplete, and Ambiguous Data

Biological data are often incomplete, uncertain, contradictory, or context-dependent. Experimental conditions vary. Datasets contain bias. Labels may be uncertain.

Negative results may be underreported. Assays may not generalize.

A model that performs well on cleaned benchmark data may behave differently when exposed to real laboratory uncertainty. Evaluation should therefore include sensitivity to missing data, noisy labels, conflicting evidence, and ambiguous biological context.

### 4.3 Unsupported but Plausible Claims

Current metrics often do not ask whether a model is generating claims that are convincing but insufficiently supported. Unsupported claims may be more dangerous when they are fluent and plausible than when they are obviously wrong.

A visibly absurd output is easy to reject. A plausible but weakly grounded output is harder to detect.

### 4.4 Human Interpretation

Benchmarks typically evaluate outputs against labels or metrics, not against human interpretation. Humans do not respond only to correctness. They respond to structure, confidence, fluency, explanation style, and alignment with prior beliefs.

In AIxBio, the presentation of a claim can become part of the risk profile. A biologically plausible statement written with confident fluency may be treated as stronger evidence than it actually is.

## 5. Human-System Interaction as a Biological Risk Factor

Evaluation is not only about model performance. It is also about how outputs are used.

Researchers, analysts, and decision-makers may give more weight to outputs that are clearly structured, confidently expressed, visually polished, or consistent with prior expectations. In high-complexity domains such as biology, fluency can be mistaken for expertise and coherence can be mistaken for evidence.

This creates a feedback loop: the model generates a plausible output, the human finds it understandable, the output is summarized or operationalized, uncertainty is

reduced or removed, and the claim gains authority through repetition and formatting.

Over time, this can lead to provenance loss. The origin, assumptions, uncertainty, and transformation history of a claim are no longer visible. Once provenance is lost, downstream users may not know whether a claim came from experimental evidence, literature synthesis, model prediction, speculative generation, or human interpretation.

Thus, human-system interaction should be treated as a core evaluation dimension. Evaluation should ask not only whether the model was correct, but whether users understood uncertainty, distinguished prediction from validation, preserved assumptions through summarization, and could trace the claim back to its source.

## 6. Risk-Sensitive Evaluation

Addressing the plausibility-validation gap does not require perfect models. It requires better structure around how AI-generated biological outputs are evaluated and used.

Not all outputs require the same evidentiary threshold. A low-stakes brainstorming suggestion does not require the same validation as a claim used to select experimental protocols, nominate therapeutic targets, assess pathogenicity, or support safety-relevant biological design.

A risk-sensitive evaluation framework should consider downstream impact, reversibility, novelty, and dual-use or misuse potential. Claims that influence experiments, safety decisions, clinical hypotheses, or resource allocation should require stronger validation than claims used for exploratory ideation.

Novel contexts require stronger caution. When inputs are far from known distributions, systems should provide more uncertainty, not more confidence.

## 7. Traceability and Transparency

AI-generated outputs should retain information about their evidentiary basis and transformation history.

At minimum, AIxBio workflows should preserve source inputs, model or system version, relevant datasets or references, assumptions, uncertainty limitations, transformation steps, human edits, downstream use context, and validation status.

A traceable claim should make clear whether it is experimentally validated, literature-supported, computationally predicted, analogically inferred, generated as a hypothesis, speculative, or unknown in evidentiary status.

This classification prevents a speculative model output from gradually becoming treated as validated knowledge.

## 8. Toward a Plausibility-Validation Evaluation Framework

A practical evaluation framework for AIxBio should combine benchmark performance with workflow-aware safeguards.

### 8.1 Benchmark Performance

Models should still be evaluated on task-specific benchmarks. Accuracy, precision, recall, ranking performance, structural similarity, calibration, and other domain metrics remain valuable. Benchmark performance should be treated as one layer of evidence, not the whole case.

### 8.2 Novelty and Distribution Testing

Evaluation should include stress tests for out-of-distribution inputs, rare cases, underrepresented domains, and difficult boundary conditions. Where novelty is high, outputs should be labeled as less certain unless independently validated.

### 8.3 Claim Support Assessment

For each actionable output, the system should identify what supports the claim. A claim without traceable support should be treated as hypothesis, not evidence.

### 8.4 Uncertainty Retention

Uncertainty should be preserved across summaries and downstream documents. If a model output begins as speculative, later reports should not convert it into a firm

statement without validation.

### 8.5 Human Reliance Testing

Evaluation should measure how users interpret outputs. Controlled studies can assess outputs with varying confidence language, explanation styles, and provenance labels. The goal is to identify whether presentation causes overreliance.

### 8.6 Workflow Auditability

AIxBio systems should support audit trails. A downstream reviewer should be able to reconstruct how a claim was generated, modified, approved, and used.

### 8.7 Validation Thresholds by Use Case

Organizations should define explicit thresholds for moving from AI-generated hypothesis to experimental action. These thresholds should vary by risk class.

## 9. Practical Governance Implications

### 9.1 For AI Developers

Developers should not present biological outputs as if benchmark performance alone establishes real-world reliability. Interfaces should distinguish predictions from validated findings and preserve uncertainty in downstream exports.

### 9.2 For Laboratories

Laboratories should treat AI outputs as claims with provenance. Experimental plans should record whether an input came from validated literature, computational inference, AI generation, or human synthesis.

### 9.3 For Publishers

Journals and repositories should encourage disclosure of AI-generated biological claims, including model identity, version, input context where appropriate, validation status, and human review process.

### 9.4 For Funders

Funders should require risk-sensitive evaluation plans in AIxBio proposals, especially when projects involve translational biology, synthetic biology, biosecurity-relevant tools, or clinical-adjacent claims.

### 9.5 For Regulators and Standards Bodies

Standards bodies should develop guidance for claim traceability, validation thresholds, and AI-generated evidence classification in biological research workflows. This paper offers conceptual governance guidance, not legal or regulatory advice.

## 10. Discussion

AI systems in biology are powerful not only because they generate correct outputs, but because they generate convincing ones. This creates a distinctive evaluation challenge. The appearance of scientific reasoning can travel faster than validation.

The problem is not solved by rejecting AI tools. Nor is it solved by relying only on larger benchmarks. The appropriate response is a more mature evaluation culture that distinguishes capability from evidence, plausibility from validation, and prediction from decision support.

The plausibility-validation gap is likely to grow as models become more fluent, multimodal, agentic, and integrated into laboratory workflows. As systems become better at producing structured biological explanations, the need for traceability becomes stronger, not weaker.

## 11. Conclusion

AI systems in biology are increasingly capable of producing outputs that look scientifically credible. This creates opportunity, but also risk.

Current evaluation practices capture performance within known settings, but they do not fully address how systems behave in uncertain environments or how their outputs are interpreted by humans. The central challenge is not only improving model accuracy. It is ensuring that plausible outputs are not mistaken for validated ones.

Biology requires empirical grounding. AI-generated outputs should therefore be treated as part of a claim lifecycle: generated, interpreted, transformed, trusted, acted upon, and eventually validated or rejected. Evaluation must follow that lifecycle.

The future of AIxBio evaluation should move beyond benchmark performance alone. It should include risk-sensitive standards, traceability, provenance preservation, uncertainty retention, human reliance testing, and explicit validation thresholds.

Plausibility is useful. It is not enough.

## Author Statement

This manuscript was prepared by S. Jason Prohaska on April 25, 2026 as an author manuscript for scholarly dissemination and ORCID record inclusion. The manuscript presents a conceptual and governance-oriented argument concerning AI evaluation in biology. It does not report new empirical experiments, human-subjects research, clinical findings, or laboratory results.

## Ethics and Integrity Statement

This paper does not involve human-subjects data, animal research, clinical intervention, or wet-lab experimentation. Any future empirical extension involving user studies, laboratory workflows, dual-use biological systems, or institutional data should undergo appropriate ethical, legal, biosafety, and institutional review before execution.

## Conflict of Interest Statement

The author may develop related academic, governance, consulting, educational, or commercial frameworks concerning AI evaluation, constitutional runtime and evidence control layer architectures, research traceability, and human-sovereign AI systems. Readers should interpret the paper as an author manuscript and conceptual contribution, not as a peer-reviewed empirical study.

## Suggested Citation

Prohaska, S. J. (2026). *When plausibility isn't enough: Evaluating AI systems in biology beyond benchmark performance*. Author manuscript. April 25, 2026. ORCID: 0009-0008-8254-8411.

## References


Abramson, J., Adler, J., Dunger, J., and colleagues. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630, 493-500.

Jumper, J., Evans, R., Pritzel, A., and colleagues. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596, 583-589.

AlphaFold Protein Structure Database. (2026). Open access protein structure predictions. Reference metadata requires final validation before external citation.

BETA / Benchmarking Evaluation Test for AlphaFold. (2025). Regularly updated benchmark sets for statistically correct evaluations of AlphaFold-based analyses. Reference metadata requires final validation before external citation.

Ferreira, F. J. N., & Carneiro, A. S. (2025). AI-driven drug discovery: A comprehensive review. *ACS Omega*. Reference metadata requires final validation before external citation.

RAND Europe. (2025). Global Risk Index for AI-enabled Biological Tools: Summary Assessment & Methods Report. Reference metadata requires final validation before external citation.

PLOS Computational Biology. (2025). Dual-use capabilities of concern of biological AI models. Reference metadata requires final validation before external citation.

Romeo, G., & Conti, D. (2025/2026). Exploring automation bias in human-AI collaboration: A review and implications for explainable AI. *AI & Society*. Reference metadata requires final validation before external citation.

Scientific Reports. (2026). Examining human reliance on artificial intelligence in decision-making contexts. Reference metadata requires final validation before

external citation.





## CITATION

Prohaska, S. J. (2026). When Plausibility Isn't Enough - Evaluating AI Systems in Biology Beyond Benchmark Performance. ETHRAEON Papers. https://papers.ethraeon.ai/aixbio-plausibility/



## PROVENANCE

ORCID 0009-0008-8254-8411

Prior CodePen RTP: n/a

`EUDS v4.1`

`Slug aixbio-plausibility`