TRIAGE: An AI Scientist for Adversarial Target Falsification

Jiawei Xing [1]

Abstract

Despite massive investment in drug discovery, more than 90% of drug candidates fail in clinical trials, often because the underlying target hypothesis proves ineffective or unsafe in humans. We present TRIAGE (Target Review via Iterative Adversarial Generation and Evaluation), a multi-agent framework that shifts target discovery from hypothesis generation to hypothesis falsification. TRIAGE combines a generative agent for target nomination with two adversarial agents for evidence retrieval and *in silico* perturbation analyses. By systematically challenging proposed targets, TRIAGE aims to filter out false positives before they advance into the costly downstream stages of drug development. We evaluate TRIAGE using benchmarks derived from historical clinical trial outcomes and public biomedical databases.

1. Background

The pharmaceutical industry is among the world's largest and most consequential industries, with an enormous impact on human health. However, drug development remains notoriously difficult, often requiring 10–15 years of work, billions of dollars in investment, and yielding overall success rates of only around 10% (Hay et al., 2014). Much of this inefficiency arises because most programs fail in clinical trials. Drug development programs are typically built on therapeutic hypotheses that perturbing a given target will benefit patients with minimal toxicity. However, more than 90% of such target hypotheses ultimately fail to demonstrate sufficient efficacy or safety in humans, underscoring the central importance of rigorous target validation in drug discovery (Plenge et al., 2013).

Recent advances in agentic AI have enabled increasing automation of the costly and complex drug discovery process (Table 1). Examples include the general-purpose biomedical agent Biomni (Huang et al., 2025), the end-to-end closed-loop agent BioLab (Jin et al., 2025a), the self-evolving agent STELLA (Jin et al., 2025b), and the translational multi-agent system Virtual Biotech (Zhang et al., 2026). However, hallucination and overconfidence in LLMs remain substantial barriers to reliable drug target discovery. To address this challenge, we propose **TRIAGE** (**T**arget **R**eview via **I**terative **A**dversarial **G**eneration and **E**valuation), an agentic target validation system that prioritizes the falsification of weak target hypotheses through historical evidence and *in silico* causal perturbation analyses.

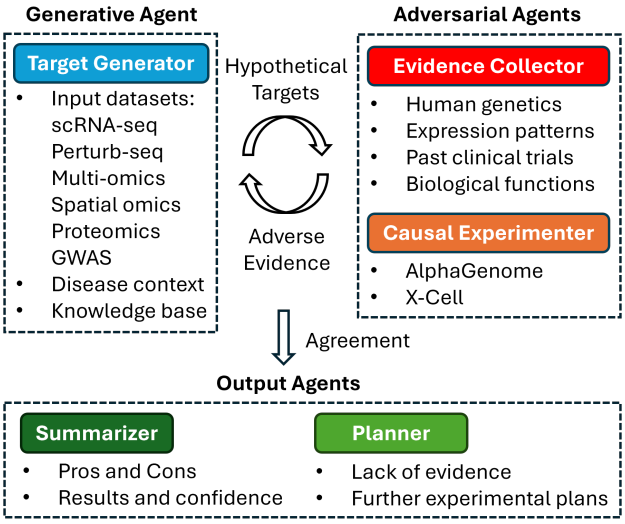

Figure 1. Architecture of TRIAGE.

2. Architecture

TRIAGE comprises three components: (i) a generative agent that proposes candidate targets from multimodal biological inputs, (ii) adversarial agents that iteratively challenge and falsify these hypotheses, and (iii) output agents that summarize the evidence for each target and, when needed, recommend follow-up experiments (Figure 1, Table 2).

2.1. Generation

We use the state-of-the-art multi-task biomedical agent Biomni as the generator to formulate mechanistic biological hypotheses and nominate candidate targets (Huang et al., 2025). **[Datasets]** Powered by hundreds of bioinformatics

[1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11791, USA. Correspondence to: Jiawei Xing <xing@cshl.edu>.

Accepted by ICML 2026 AI for Science Workshop, Seoul, South Korea.

tools, the generator can analyze diverse input modalities, including bulk and single-cell RNA-seq, Perturb-seq, multi-omics, spatial transcriptomics, proteomics, and genetic variant data. **[Knowledge]** In addition, the generator has access to comprehensive literature resources spanning diverse diseases and genes, as well as structured biological knowledge such as gene sets, regulatory networks, and protein-protein interaction networks. The generator produces candidate target hypotheses for a disease of interest in the form: "Target X in Context Y drives Phenotype Z through Pathway A".

2.2. Adversary

The core function of TRIAGE is to falsify candidate target hypotheses through two adversarial agents. The first agent gathers counterevidence using specialized sub-agents. **[Genetics]** Human genetic support from genome-wide association studies (GWAS) is associated with a higher success rate for drug targets (Plenge et al., 2013). **[Safety]** Gene expression patterns from Genotype-Tissue Expression (GTEx) and Cancer Dependency Map (DepMap) provide evidence on toxicity risk, as safer targets tend to exhibit greater specificity in expressions (GTEx Consortium, 2020; Arafeh et al., 2025). **[Clinical trials]** Historical clinical trial outcomes analyzed by Virtual Biotech provide direct evidence from past target failures (Zhang et al., 2026). **[Functions]** Finally, gene sets and biological annotations help identify likely bystanders involved in stress responses, generalized inflammation, and housekeeping programs.

Notably, nearly half of the targets fail because of insufficient efficacy, often due to reactive associations between the target and disease phenotype rather than true causal relationships (Plenge et al., 2013). **[Causality]** Therefore, the second adversarial agent focuses on falsifying the underlying causal hypotheses using two state-of-the-art predictive models. AlphaGenome is used to predict changes in target gene expression *in-silico* under mutagenesis of fine-mapped GWAS variants (Avsec et al., 2026). In addition, X-Cell is used to perform *in silico* perturbation tests, where knockout of a candidate gene target is expected to shift cell states toward healthy phenotypes (Wang et al., 2026).

2.3. Outcome

After sufficient rounds of generation and adversarial evaluation, the adversarial agents are expected to converge with the generative agent on the most credible target hypotheses. The summary agent then records the outcome of each round of falsification in long-term memory and generates a target card for each candidate (Table 3). Each card summarizes the supporting and opposing evidence across the preceding analyses and provides a recommendation of "pass", "fail", or "need data" along with a confidence level. In addition, when the available evidence is insufficient for a definitive decision, the planning agent proposes follow-up experiments needed to further evaluate the target.

3. Evaluation

TRIAGE is evaluated in three complementary settings: recovery of known targets, elimination of false targets, and architectural ablations, using various metrics (Table 4).

3.1. Target recovery

To evaluate TRIAGE, we construct benchmark datasets from target–disease associations in the Open Targets Platform (Buniello et al., 2025). Highly validated associations serve as proxy ground-truth positives, and performance is measured using Top-k recall. Low-evidence associations serve as decoys for evaluating elimination performance, including elimination rate and the adversarial rationale generated across iterations. Human experts will review both the validation process and the resulting target hypotheses.

3.2. False-target elimination

To evaluate TRIAGE's ability to reject invalid targets, we curate a false-target set enriched for candidates with limited genetic support, non-specific expression patterns, and prior failures in clinical trials. We measure whether these targets are eliminated during validation, when they are eliminated, and the adversarial rationale for their elimination.

3.3. Ablation

To assess the value of the dual-critic architecture, we compare TRIAGE against generator-only and single-adversary variants, as well as versions with one perturbation module removed. To measure the contribution of each evidence source, we perform knowledge ablations by withholding access to key databases, including GWAS, GTEx, DepMap, and clinical trial records. To evaluate the role of iterative falsification, we analyze intermediate outputs from each validation round and compare performance across iterations.

4. Governance

TRIAGE is designed to reduce false outputs during target evaluation. All evidence and analyses produced by agents are grounded in established literature and databases, with explicit references and citations. To mitigate risks from LLM-generated code, all executions are sandboxed with read-only data access, immutable logging, and no credential exposure. Output agents are restricted to summarizing target evaluations and proposing follow-up experiments under human expert oversight, without access to real-world laboratory devices or biological samples. For private commercial or patient data, TRIAGE supports local deployment with

confidential memory and secure data storage.

References

Arafeh, R., Shibue, T., Dempster, J. M., Hahn, W. C., and Vazquez, F. The present and future of the cancer dependency map. *Nature Reviews Cancer*, 25(1):59–73, 2025.

Avsec, Ž., Latysheva, N., Cheng, J., Novati, G., Taylor, K. R., Ward, T., Bycroft, C., Nicolaisen, L., Arvaniti, E., Pan, J., et al. Advancing regulatory variant effect prediction with alphagenome. *Nature*, 649(8099):1206–1218, 2026.

Buniello, A., Suveges, D., Cruz-Castillo, C., Llinares, M. B., Cornu, H., Lopez, I., Tsukanov, K., Roldán-Romero, J. M., Mehta, C., Fumis, L., et al. Open targets platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic acids research*, 53(D1):D1467–D1475, 2025.

GTEx Consortium. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369 (6509):1318–1330, 2020.

Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., and Rosenthal, J. Clinical development success rates for investigational drugs. *Nature biotechnology*, 32(1):40–51, 2014.

Huang, K., Zhang, S., Wang, H., Qu, Y., Lu, Y., Roohani, Y., Li, R., Qiu, L., Li, G., Zhang, J., et al. Biomni: A general-purpose biomedical ai agent. *bioRxiv*, 2025. doi: 10.1101/2025.05.30.656746.

Jin, R., Guo, Y., Qu, Y., Yang, M., Shang, C., Yang, Q., Chao, L., Zhou, Y., Xu, R., Xu, Z., et al. Biolab: End-to-end autonomous life sciences research with multi-agents system integrating biological foundation models. *bioRxiv*, 2025a. doi: 10.1101/2025.09.03.674085.

Jin, R., Zhang, Z., Wang, M., and Cong, L. Stella: Self-evolving llm agent for biomedical research. *arXiv preprint arXiv:2507.02004*, 2025b.

Mehandru, N., Hall, A. K., Melnichenko, O., Dubinina, Y., Tsirulnikov, D., Bamman, D., Alaa, A., Saponas, S., and Malladi, V. S. Bioagents: Bridging the gap in bioinformatics analysis with multi-agent systems. *Scientific Reports*, 15(1):39036, 2025.

Plenge, R. M., Scolnick, E. M., and Altshuler, D. Validating therapeutic targets through human genetics. *Nature reviews Drug discovery*, 12(8):581–594, 2013.

Qu, Y., Huang, K., Yin, M., Zhan, K., Liu, D., Yin, D., Cousins, H. C., Johnson, W. A., Wang, X., Shah, M., et al. Crispr-gpt for agentic automation of gene-editing experiments. *Nature Biomedical Engineering*, 10(2):245–258, 2026.

Wang, C., Karimzadeh, M., Ravindra, N. G., Bounds, L. R., Alerasool, N., Huang, A. C., Ma, S., Gulbranson, D. R., Cui, H., Lee, Y., et al. X-cell: Scaling causal perturbation prediction across diverse cellular contexts via diffusion language models. *bioRxiv*, 2026. doi: 10.64898/2026.03.18.712807.

Wang, Z., Jin, Q., Wei, C.-H., Tian, S., Lai, P.-T., Zhu, Q., Day, C.-P., Ross, C., Leaman, R., and Lu, Z. Geneagent: self-verification language agent for gene-set analysis using domain databases. *Nature Methods*, 22(8):1677–1685, 2025.

Zhang, H. G., Eckmann, P., Miao, J., Mahon, A. B., and Zou, J. The virtual biotech: A multi-agent ai framework for therapeutic discovery and development. *bioRxiv*, 2026. doi: 10.64898/2026.02.23.707551.

Table 1. Comparison of TRIAGE and other biomedical agents.

System	Task	Approach	Reference
TRIAGE	Hypothesis generation and **target validation**.	Iterative targets generation and falsification by multiple agents, including information retrieving and causal analyses.	This study.
Biomni	**General-purpose** tool execution and biomedical workflow automation.	Retrieving tools, writing codes, and generating answers.	(Huang et al., 2025)
BioLab	End-to-end **closed-loop** biomedical research automation.	Integration of dry lab and wet lab.	(Jin et al., 2025a)
STELLA	**Self-evolving** problem solving and lab control.	Multi-agent for tool creation.	(Jin et al., 2025b)
CRISPR-GPT	**Gene-editing** experiment design and analysis.	Domain-specific planning and execution.	(Qu et al., 2026)
GeneAgent	**Gene-set** analyses with better accuracy.	Critic agent retrieving information from databases for self-verification.	(Wang et al., 2025)
Virtual Biotech	Organization-scale **translational** reasoning.	Multiple specialist agents for clinical data analyses.	(Zhang et al., 2026)
BioAgents	**Bioinformatics** workflows.	Multi-agent system with retrieval augmented generation (RAG).	(Mehandru et al., 2025)

Table 2. Agentic design of TRIAGE framework.

Component	Task	Approach
Generator	Generating hypothetical mechanisms and candidate targets for the disease of interest.	1. Use Biomni with multimodal bioinformatics tools for input data analyses. 2. Retrieve biomedical literature, pathways, and interaction networks for the disease of interest. 3. Output mechanistic hypotheses in target-context-phenotype format.
Adversary 1	Retrieving counterevidence to falsify candidate targets from the generator.	1. Retrieve fine-mapped data for the disease of interest from GWAS Catalog. Check if variants affect candidate targets. 2. Retrieve gene expression patterns from GTEx and DepMap. Check if candidate targets have tissue-specific expressions. 3. Search historical failed clinical trials and conflicting literature using Virtual Biotech. 4. Remove non-causal bystanders such as stress-response genes based on gene annotations at NCBI database.
Adversary 2	Performing *in silico* perturbation analyses to verify the causal relationships.	1. AlphaGenome predicts effects of regulatory variants on target genes, especially those intronic and intergenic. 2. X-Cell simulates gene perturbation and cell-state transitions. Check whether target knockout shifts cells toward healthy cell states.
Outcome	Output target cards for each gene and plan follow-up experiments.	1. Summarize analyses for each gene into target cards. Assign decisions for each gene based on analyses. 2. For candidate targets with limited evidence, make plans for follow-up experiments.

Table 3. Example target card as TRIAGE output.

Category	Example Output
Disease	Idiopathic pulmonary fibrosis
Candidate target	TGFB1
Hypothesis	TGFB1 drives fibroblast activation and extracellular matrix deposition.
Genetic evidence	Strong support from GWAS loci near TGF-β signaling genes.
Safety evidence	Moderate risk due to broad tissue expression; monitor cardiac toxicity.
Clinical evidence	Prior pathway inhibitors showed partial efficacy with tolerability concerns.
Functional evidence	Cytokine regulating cell growth and differentiation; likely not a bystander.
In silico perturbation	Knockdown predicted reduction of profibrotic fibroblast state.
Overall verdict	**Pass (confidence: Medium)**
Recommended experiments	CRISPR Perturb-seq in patient fibroblasts; selective delivery strategy.

Table 4. Evaluation metrics for TRIAGE.

Metric	Description
Top-*k* Recall	Fraction of known validated targets ranked within top-k candidates.
Precision of Pass Calls	Fraction of approved targets among all pass recommendations.
False Target Elimination Rate	Percentage of weak or historically failed targets rejected.
Early Kill Iteration	Average iteration step when false targets are removed.
Failure Attribution	Primary evidence source leading to target rejection.
Expert Agreement	Consistency between TRIAGE verdicts and human expert review.
Calibration	Alignment between confidence levels and true success rates.
Ablation Sensitivity	Performance drop after removing an adversary or knowledge source: • Generator • Generator + Adversary 1 • Generator + Adversary 2 • Generator + Adversary 1 + AlphaGenome • Generator + Adversary 1 + X-cell • Generator + GWAS/GTEx/DepMap/Clinical Trials + Adversary 2 • Generator + Adversaries within 1, 5, and 10 iterations