

---

# BIORESEARCHER: Scenario-Guided Multi-Agent for Translational Medicine

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 Translational medicine turns underspecified development goals into evidence syn-  
2 thesis that must combine literature, trials, patents, and quantitative multi-omics  
3 analysis while preserving identifiers, uncertainty, and retrievable provenance.  
4 General-purpose foundation models and off-the-shelf tool-augmented or multi-  
5 agent systems are not built for this: they tend to produce single-shot answers or  
6 run open-endedly, and fall short on the auditable, scenario-specific workflows that  
7 heterogeneous biomedical sources demand.

8 This paper introduces BIORESEARCHER, a scenario-guided multi-agent system that  
9 maps queries to versioned research playbooks, delegates to specialized subagents  
10 over 30+ tools and machine-learning endpoints, mixes structured database access  
11 with sandboxed code for genome-scale analyses, and applies claim-level multi-  
12 model reconciliation before editorial assembly.

13 We evaluate BIORESEARCHER at three complementary levels. On 109 single-step  
14 tests covering ontology/entity grounding, qualitative retrieval, and quantitative  
15 computation, the system achieves the best aggregate performance among evaluated  
16 baselines (83.49% pass rate; 0.892 average score; up to 96.58% on quantitative  
17 analysis). On open-ended biomedical analysis benchmarks, it reaches 89.33% on  
18 BixBench-Verified-50, 81.82% on its human-biology subset, and the top mean  
19 score of 0.758 on the 193-question BaisBench Scientific Discovery track. On a  
20 30-query clinical end-to-end benchmark spanning approved therapies, clinical-  
21 stage and preclinical candidates, synthetic-lethality combinations, and prognostic  
22 biomarkers, it reaches the highest positive hit rate ( $74.7\% \pm 3.3\%$ ) and negative  
23 clear rate ( $96.8\% \pm 0.2\%$ ), exceeding OpenAI Deep Research, CellType, and  
24 Medea, while exposing an expected sensitivity-specificity tradeoff for first-pass  
25 hypothesis generation.

26 BIORESEARCHER is not an autonomous clinical decision support, but an auditable  
27 research co-scientist that compresses weeks of qualitative and quantitative evidence  
28 into reviewable dossiers for domain experts, a practical template for expert-facing  
29 biomedical AI.

## 30 1 Introduction

31 Consider the following, exemplar question posed to a translational research team: “Given an antibody-  
32 drug conjugate (ADC) with a topoisomerase-I (TOP1) primary payload, what is the optimal second  
33 payload class for synergistic anti-tumor activity in non-small cell lung cancer (NSCLC)?”

34 This question, representative of real strategic decisions in pharmaceutical R&D, is operationally  
35 hard for several reasons. First, the query is **entity-ambiguous**: “Topo-I,” “topo1,” “topoiso-  
36 merase I,” and “TOP1” all denote the same protein target (ChEMBL ID CHEMBL1781, Ensembl

37 ENSG00000198900), yet different databases index them under different names. At the same time, **on-**  
38 **ontology mismatches** complicate cross-source integration: disease identifiers vary across EFO, DOID,  
39 MONDO, and TCGA conventions; compound identifiers differ between ChEMBL, and PubChem  
40 while accurate drug’s target mapping is noisy and depends on application.

41 Second, relevant evidence is **scattered across heterogeneous sources**: PubMed abstracts, Clini-  
42 calTrials.gov records, Google Patents filings, and DepMap multi-omics datasets, each with distinct  
43 identifiers, access patterns, and biological context. Additional sources such as research documents  
44 and ASCO, AACR, or ESMO abstracts could be integrated where licensing permits.

45 Third, reasoning spans interacting biological scales and high-dimensional measurements, requiring  
46 both qualitative synthesis and quantitative analysis. Fourth, outputs must be auditable dossiers: ranked  
47 hypotheses with mechanistic rationale and provenance. We present BIORESEARCHER, a scenario-  
48 guided multi-agent system that normalizes entities, routes questions to translational playbooks,  
49 delegates retrieval and CodeAct analyses, and reconciles evidence into provenance-preserving reports.

50 Our contributions are:

- 51 (a) **A scenario-guided agent architecture for translational medicine** that frames the problem as  
52 scientific assistance for daily translational workflows: translating broad questions into explicit  
53 research plans, combining qualitative and quantitative evidence, and contextualizing outputs for  
54 clinical development (Sections 2).
- 55 (b) **A layered evidence-synthesis stack with multi-model reconciliation** that integrates ontology-  
56 aware translation, qualitative knowledge synthesis, quantitative dataset analysis, and autonomous  
57 CodeAct-style computation under a single orchestrator, and reconciles outputs via structured  
58 claim-level debate—claim extraction, cross-model grouping, multi-round argumentation, and  
59 quantitative consensus detection—for auditable long-form biomedical report synthesis (Sec-  
60 tion 2).
- 61 (c) **A realistic multi-level evaluation protocol** that separates component verification from hierar-  
62 chical reasoning tests, autonomous custom-analysis tasks, and end-to-end expert assessment  
63 on realistic translational questions, together with ablations against direct LLM generation  
64 (Section 3).

## 65 2 System Architecture

66 BIORESEARCHER is a scenario-guided multi-agent system for expert-facing translational evidence  
67 synthesis. Natural-language questions are typically underspecified and synonym-heavy. Outputs are  
68 auditable dossiers with normalized entities, heterogeneous evidence, ranked hypotheses, mechanistic  
69 links, and provenance sufficient for expert verification (PMIDs, NCT IDs, patent numbers, and related  
70 identifiers).

71 The architecture separates method selection, evidence acquisition, and reconciliation (Figure 1). A  
72 master orchestrator matches a versioned scenario playbook—encoding analyst role, disambiguation  
73 rules, subagent protocols, and mandatory report structure—injects its body without code changes or  
74 fine-tuning, decomposes the query, and delegates to specialized subagents that publish provenanced  
75 artifacts. A reconciliation agent ingests shared artifacts from subagents, plans sections from a  
76 metadata manifest, drafts with models from different provider families, and resolves disagreement  
77 at atomic claim level with verbatim source-support quotes. A final editorial pass harmonizes tone  
78 without altering reconciled claims.

79 Subagents combine schema-stable tool loops, iterative retrieval-heavy synthesis, cross-provider fan-  
80 out for open-web research, and sandboxed code-mediated analysis for genome-scale statistics where  
81 enumerating tools is infeasible.

82 Reconciliation follows the linear procedure in 1: scoped planning and writing mitigate long-context  
83 loss, cross-family drafting exploits decorrelated synthesis biases, semantic claim clustering en-  
84 ables bounded debate with optional frontier reconciliation, and dual stopping on disagreement and  
85 confidence caps cost while preserving auditable transcripts.

86 Entity resolution, retrieval endpoints, sandboxed multi-omics execution, and other text dataset from  
87 RAG are expanded in Appendix B.

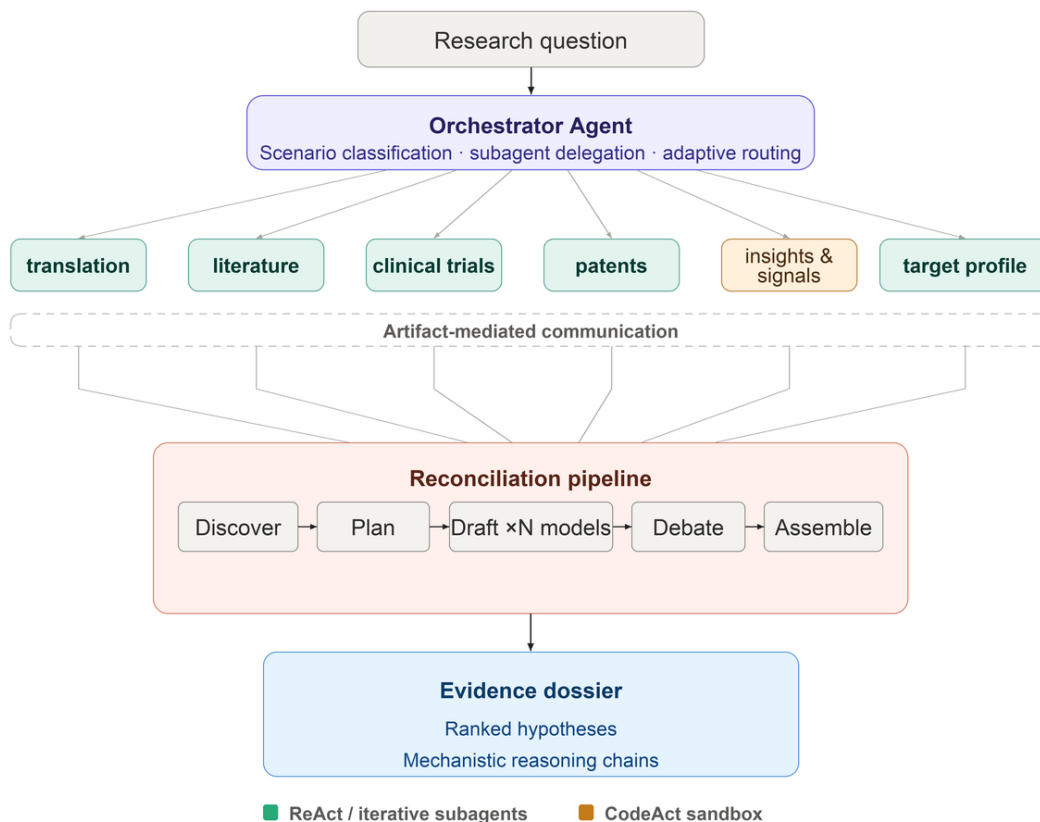


Figure 1: System architecture. The master orchestrator Agent selects a scenario playbook and adaptively delegates self-contained subtasks to specialized retrieval and analysis agents. Agents publish provenanced source artifacts to a shared evidence bus; a reconciliation agent ingests the bus, drafts the report with multiple models, and resolves disagreement at the claim level.

### 88 3 Evaluation

89 To rigorously evaluate capability of or BIORESEARCHER, we designed a benchmark suite at three  
 90 complementary conceptual levels: single-step tests, clinical reasoning & autonomous analysis  
 91 benchmark, and clinical end-to-end benchmark.

#### 92 3.1 Single-Step Tests

93 We assembled a 109-question benchmark that verifies each system layer in isolation: L1 ontology  
 94 and entity resolution, L2 qualitative retrieval, and L3 quantitative computation. That enables to  
 95 verify how the system performs with accurately answering questions about data and facts across drug  
 96 development aspects at every modality scale. The purpose of this benchmark is to see if the agent can  
 97 perform simple factual checks and calculations. One may consider this a sanity check to see if the  
 98 agent is able to perform basic tasks: crucial before evaluating the agent on more complex queries.  
 99 The questions check e.g. the ability to resolve entity IDs, retrieve a PubMed publication, and process  
 100 quantitative data. Each test question asks about a functionality of one sub-agent. Therefore, the tests  
 101 do not only check the ability of the agent to perform single-step reasoning, but also the ability of the  
 102 agent to route a simple question to the correct sub-agent. We use the single-step tests to compare the  
 103 performance of BIORESEARCHER, GPT-5.4-mini, GPT-5.5, and CellType Agent.

##### 104 3.1.1 Benchmark Design

105 The benchmark consists of 109 questions requiring single-step reasoning and is partitioned into three  
 106 subsets (layers), summarized in Table 8. **Ontology & Entity** (L1) challenges entity normalization

---

**Algorithm 1** Claim-level section reconciliation.

---

**Require:** Candidate drafts  $D = \{d_1, \dots, d_M\}$  for section  $s$ , question  $q$ , agreement threshold  $\tau$ , max rounds  $R$ .

**Ensure:** Reconciled section  $y_s$ .

- 1: For each  $d_m$ , extract atomic claims with source-support quote and confidence.
  - 2: Cluster claims across drafts into topic groups labeled {full, partial, conflict}.
  - 3: **for**  $r = 1, \dots, R$  **do**
  - 4:   In parallel, each model issues an updated stance per group with revised claim, explanation, and confidence, conditioned on other agents’ prior positions.
  - 5:   **if** no model disagrees on any group **and** mean confidence  $\geq \tau$  **then**
  - 6:     **break**
  - 7:   **end if**
  - 8: **end for**
  - 9: Synthesize  $y_s$  from drafts, claim groups, and final positions, restricted to the section’s allocated sources.
- 

107 across biomedical ontologies. This layer focuses mostly on the ability of the agent to retrieve  
108 key entities (e.g. genes, diseases) from the free-text input and map them to their canonical IDs.  
109 Another example task is resolving a gene family name to a list of member genes with canonical  
110 IDs. **Qualitative Synthesis** (L2) evaluates retrieval of qualitative knowledge from biomedical  
111 databases. The questions check (among others) for facts about genes, compounds, protein functions  
112 and pathways, clinical trials, patents, and literature. Example tasks are retrieving clinical trial NCT  
113 IDs for a given context or assessing the success of a clinical trial given its NCT number. **Quantitative**  
114 **Analysis** (L3) assesses extraction and computation of key quantitative data points. An example task  
115 is checking the expression level of a given gene in a given cell line.

Table 1: Single-step test subsets, layer coverage, and a representative query per subset. Each subset targets a specific system layer and is scored independently with its own rubric.

Subset	Layer Tested	N	Representative Query
Ontology & Entity	Entity resolution / grounding	33	“What is the Ensembl gene ID and ChEMBL target ID for CD340?”
Qualitative Synthesis	Literature / trials / patents retrieval	37	“Provide the title of the publication with PMID 25439351.”
Quantitative Analysis	DepMap / omics computation	39	“What is the CRISPR gene-effect score for PARP1 in MCF7?”
<b>Total</b>		<b>109</b>	

116 **Test-case format.** Each test case is a pair of a question (a natural-language query for the agent)  
117 and an expected output (example natural-language response that the agent is expected to return). For  
118 example, the expected output for the question “What is the Ensembl gene ID and ChEMBL target ID  
119 for CD340?” is as follows.

120 Ensembl ID: ENSG00000141736  
121 ChEMBL target ID: ChEMBL1824

122 Each expected output was manually verified by an expert to be correct.

### 123 3.1.2 Evaluation Protocol

124 **Execution.** Every question is submitted end-to-end to the full master graph: the top-level orchestra-  
125 tor dispatches it to the appropriate sub-agent, which invokes tools as needed. No sub-agent is tested  
126 in isolation, so the single-step tests also probe the orchestrator’s routing decisions. The agent’s *actual*  
127 *output* for scoring is the natural-language content of the final message emitted by the master graph.

128 **Judge.** We use DeepEval [1], an LLM-as-a-judge framework, to score open-ended natural-language  
129 outputs against the expected outputs. For each test case, the judge is shown the question, the actual

130 output, and the expected output, and assigns a score between 0 and 1, together with a free-text  
131 justification. Each test case is paired with a custom DeepEval metric that instructs the judge how  
132 to score the output. A metric specification consists of (i) natural-language evaluation criteria, (ii) a  
133 rubric, i.e. a list of tier descriptions mapped to disjoint score bands on  $[0, 1]$ , and (iii) a threshold  
134 score (the minimum score required for the test to pass). An example criteria block is:

135 Evaluation dimensions:

- 136 1. Canonical entity: When the expected output labels a canonical name (e.g. drug name,  
137 disease name), is it present and unambiguously identified in the actual output?
- 138 2. Identifier match: Do all identifiers labelled in the expected output appear  
139 in the actual output with matching values?
- 140 3. No fabrication: The actual output does not introduce alternative identifiers  
141 that contradict the expected ones, and does not confuse the entity with a same-string  
142 alias of a different gene/disease/drug.

143 A three-tier rubric yields the bands  $[0.0, 0.2]$ ,  $[0.4, 0.6]$ ,  $[0.8, 1.0]$  (detailed in Table 2); the gaps make  
144 tier boundaries unambiguous and the judge commits to one tier per test case. In each custom metric,  
145 the threshold score is set so that a test passes if the judge assigns a score in the highest band (e.g. 0.8  
146 for three-tier rubrics). An example three-tier rubric for entity-resolution questions (like the example  
147 for layer *Ontology & Entity* in Table 8) is shown in Table 2. We used DeepEval’s default LLM model,  
148 which is GPT 4.1, for all metrics.

Table 2: Example rubric: the judge assigns one tier, yielding a score in the corresponding band.

Score band	Expected outcome specification
$[0.0, 0.2]$	Identifiers absent or contradicting the expected ones, or canonical entity (when labelled in the expected output) wrong or missing.
$[0.4, 0.6]$	Most identifiers correct but some missing or partially mismatched, or canonical entity (when labelled) partially off. Additional information contradicts the expected output.
$[0.8, 1.0]$	All expected identifiers present and matching, canonical entity (when labelled) correct, no fabricated alternatives. If additional information is present, it doesn’t contradict the expected output.

149 **Metric.** The primary per-subset metric is the *pass rate*: the fraction of test cases in that subset  
150 whose judge-assigned score satisfies  $\text{score} \geq \text{threshold}$  (i.e. such that were classified into the top tier  
151 of the rubric). Another metric is the average metric score: the average score of the test cases.

### 152 3.2 Quantitative Reasoning & Autonomous Analysis Benchmark

153 We evaluate BIORESEARCHER’s ability to analyze custom datasets and propose dedicated au-  
154 tonomous analyses when the required computation is not fully specified in advance, using two  
155 complementary benchmarks: *BixBench* [2] and *BaisBench* (Scientific Discovery).

156 *BixBench.* We use a subset of BixBench restricted to *Homo sapiens* data, which probes open-ended  
157 biomedical data analysis under partial specification. We report BixBench-style open-answer accuracy  
158 on *phylobio/BixBench-Verified-50* [2] (Table 4) for the full 50-question set and a 22-question  
159 *human* subset of questions that concern human biology. For fairness, package-augmented Claude  
160 Code variants were provided the same data-analysis package list available to BIORESEARCHER,  
161 ensuring comparable minimal execution context.

162 *BaisBench.* We use the Scientific Discovery (BAIS-SD) track of BaisBench [3]: 193 data-driven  
163 questions on real single-cell RNA-seq datasets, grounded in biological conclusions from 41 published  
164 single-cell studies, where agents analyze the provided data to select answers in single- and multi-  
165 answer formats as in the benchmark. Scoring is trichotomous (0, 0.5, or 1) per the SD task: single-  
166 answer items receive 0 or 1; multi-answer items receive 1 if all correct options are selected and no  
167 incorrect one, 0.5 if at least one correct option is selected and no incorrect one (but not the full set),  
168 and 0 otherwise. We report the mean  $S_{SD}$  in Table 4.

169 **3.3 Clinical End-to-End Benchmark.**

170 We constructed a realistic benchmark centered on the translational medicine scenario family, where  
 171 each query is paired with expert-authored ground truth. The benchmark tests system’s end-2-end  
 172 performance, i.e. whether the system can reason, retrieve evidence, and generalize across a domain  
 173 hierarchy: approved therapies, therapies in clinical trials, preclinical therapies, synthetic-lethality  
 174 combinations, and prognostic biomarkers for specific indications. This tiered design evaluates not  
 175 only recall of established knowledge but also generalization to less mature therapeutic settings.

176 **4 Results**

177 **4.1 Single-Step Tests**

178 Evaluation results for BIORESEARCHER, GPT-5.4-mini, GPT-5.5, and CellType Agent are shown  
 179 in Table 3. As expected, GPT-5.4-mini has the lowest pass rates and average metric score across all  
 180 layers and subsets. GPT-5.5 has significantly higher performance than GPT-5.4-mini, in all metrics.  
 181 CellType Agent has the highest pass rate in layer L2 (Qualitative Synthesis). BIORESEARCHER with  
 182 GPT-5.4 as core model achieved the highest average metric score in layers L1 (Ontology & Entity)  
 183 and L3 (Quantitative Analysis), and the highest total pass rate and average metric score overall.

Table 3: Single-step test results, averaged over 3 runs. The first four columns report pass rates, defined as the percentage of cases in the top rubric band (mean  $\pm$  std over 3 runs) per layer and overall. The last two columns report the judge’s average metric score and the average token usage on the entire evaluation set (mean  $\pm$  std over 3 runs). Model suffixes indicate reasoning effort, and BIORESEARCHER variants indicate the core model used. BIORESEARCHER (mini) stands for BIORESEARCHER with core model GPT-5.4-mini, BIORESEARCHER (low) stands for BIORESEARCHER with core model GPT-5.4 with reasoning effort set to “low”.

System	L1 (%)	L2 (%)	L3 (%)	Total (%)	Avg score	Token usage
GPT-5.4-mini	26.26 $\pm$ 6.31	17.12 $\pm$ 4.13	11.11 $\pm$ 3.92	17.74 $\pm$ 2.80	0.316 $\pm$ 0.022	19167 $\pm$ 3948
GPT-5.4 (low)	40.85 $\pm$ 3.94	26.13 $\pm$ 3.12	20.51 $\pm$ 0.00	28.53 $\pm$ 1.84	0.448 $\pm$ 0.008	61824 $\pm$ 2341
GPT-5.5 (medium)	65.66 $\pm$ 1.75	37.84 $\pm$ 4.68	28.21 $\pm$ 5.13	42.81 $\pm$ 2.12	0.585 $\pm$ 0.009	221829 $\pm$ 7968
GPT-5.5 (high)	67.68 $\pm$ 1.75	40.04 $\pm$ 4.67	30.77 $\pm$ 2.56	45.09 $\pm$ 0.24	0.591 $\pm$ 0.005	447181 $\pm$ 20637
Gemini 3.1 Pro	47.38 $\pm$ 3.24	27.17 $\pm$ 1.21	12.82 $\pm$ 0.00	27.94 $\pm$ 1.10	0.481 $\pm$ 0.009	176974 $\pm$ 7500
CellType	81.82 $\pm$ 3.03	<b>77.48 <math>\pm</math> 1.56</b>	46.60 $\pm$ 7.32	67.80 $\pm$ 1.70	0.731 $\pm$ 0.003	51953004 $\pm$ 931350
<b>BIORESEARCHER (mini)</b>	77.78 $\pm$ 1.75	74.77 $\pm$ 3.12	91.45 $\pm$ 1.48	81.65 $\pm$ 0.92	0.889 $\pm$ 0.015	2975729 $\pm$ 8001
<b>BIORESEARCHER (low)</b>	<b>84.85 <math>\pm</math> 3.03</b>	68.47 $\pm$ 3.12	<b>96.58 <math>\pm</math> 1.48</b>	<b>83.49 <math>\pm</math> 1.59</b>	<b>0.892 <math>\pm</math> 0.004</b>	2646126 $\pm$ 29309

184 **4.2 Quantitative Reasoning & Autonomous Analysis Benchmark**

185 **BixBench and BaisBench.** We report BixBench-style open-answer accuracy on  
 186 phylobio/BixBench-Verified-50 [2] and mean  $S_{SD}$  on the Scientific Discovery (BAIS-  
 187 SD) track of BaisBench [3] in Table 4. The BixBench columns cover the full 50-question  
 188 set and a 22-question *human* subset of questions that concern human biology. For fairness,  
 189 package-augmented Claude Code variants were provided the same data-analysis package list  
 190 available to BIORESEARCHER, ensuring minimal execution context. Task setup and scoring are  
 191 specified in subsection 3.2.

192 **4.3 Clinical End-to-End Benchmark**

193 BIORESEARCHER vs. baselines on all evaluation criteria. Statistical significance via paired bootstrap  
 194 or permutation test. Table 5 reports end-to-end benchmark results using positive hit rate and negative  
 195 clear rate.

Table 4: Quantitative reasoning and autonomous analysis benchmark results (mean  $\pm$  std over 3 runs). BixBench-Verified-50 reports open-answer accuracy (%) on the full set and the human-biology subset; BaisBench Scientific Discovery (BAIS-SD) reports mean  $S_{SD}$  on the 193-question SD track. BIORESEARCHER uses core model GPT-5.4 with reasoning effort set to “medium”; No results is bolded since all results are statistically indistinguishable from each other. CellType and Claude Code are powered by Claude Opus 4.7.

System	BixBench		BaisBench
	Full (%)	Human (%)	$S_{SD}$
BIORESEARCHER (GPT-5.4 medium)	89.33 $\pm$ 1.15	81.82 $\pm$ 0.00	0.758 $\pm$ 0.005
CellType (Claude Opus 4.7)	89.33 $\pm$ 2.31	83.33 $\pm$ 6.94	0.641 $\pm$ 0.082
Claude Code (Claude Opus 4.7)	90.00 $\pm$ 2.00	83.33 $\pm$ 2.62	0.759 $\pm$ 0.012

Table 5: End-to-end benchmark results (30 queries): positive hit rate and negative clear rate (%). BIORESEARCHER, CellType, Medea, and OpenAI Deep Research: mean  $\pm$  std over 3 runs. Medea is powered by GPT-5.4 with reasoning effort set to “low”.

System	Positive rate	Negative rate
<b>BIORESEARCHER</b>	<b>74.7 <math>\pm</math> 3.3</b>	<b>96.8 <math>\pm</math> 0.2</b>
CellType	61.7 $\pm$ 2.3	83.3 $\pm$ 0.0
Medea	33.7 $\pm$ 5.3	93.3 $\pm$ 3.3
OpenAI Deep Research	68.9 $\pm$ 7.7	81.1 $\pm$ 1.9

## 196 5 Related Work

### 197 5.1 Agentic Systems in Biomedical Research

198 LLM-based agent systems in the life sciences span autonomous experimental design, clinical de-  
 199 cision support, and broader biological investigation. In experimental design, Coscientist [4] and  
 200 ChemCrow [5] demonstrated LLM-driven robotic synthesis and tool-grounded chemical reasoning,  
 201 with later work extending to retrosynthesis [6], in-silico lead generation [7], protein binder design [8],  
 202 multi-objective protein engineering [9, 10, 11, 12], and early clinical validation of an AI-designed  
 203 kinase inhibitor [13]. These systems typically address well-scoped molecular or synthetic design  
 204 tasks rather than multi-source evidence integration.

205 Clinical agents have been applied to therapeutic question answering over curated knowledge  
 206 bases [14], adaptive multi-agent triage [15], multimodal precision oncology [16], patient-to-trial  
 207 matching [17], and histopathology pipeline construction [18]. They operate within structured work-  
 208 flows such as diagnosis, treatment selection, and eligibility screening, but are not primarily designed  
 209 for open-ended translational evidence synthesis across heterogeneous sources.

210 Broader biological agent platforms include Virtual Lab [19], Biomni [20], BioLab [21], OriGene [22],  
 211 and Medea [23], alongside systems for bioinformatics workflow automation [24, 25], self-verified  
 212 gene-set analysis [26], and knowledge-graph construction [27]. Our system falls within this category  
 213 but focuses on translational drug-discovery use cases, integrating evidence from literature, clinical  
 214 trials, patents, and multi-omics datasets into structured research dossiers for preclinical and clinical  
 215 development decisions.

## 216 6 Analysis and Discussion

### 217 6.1 Interpreting the Results

218 BIORESEARCHER obtains the best overall single-step pass rate (83.49%) and average judge score  
 219 (0.892) on the 109-question suite (Table 3), surpassing the specialized CellType agent (67.80%,  
 220 0.731) and the strongest frontier baseline GPT-5.5 (high) (45.09%, 0.591). The largest margin  
 221 is at L3 (Quantitative Analysis), where BIORESEARCHER reaches 91.45–96.58% against 46.60%  
 222 for CellType and 28.21–30.77% for GPT-5.5, consistent with the architectural choice to expose  
 223 genome-wide DepMap-style analyses through a CodeAct sandbox rather than a fixed set of pre-

224 enumerated tool functions. At the same time, BIORESEARCHER is not uniformly dominant at every  
225 layer: CellType achieves a higher L2 pass rate (77.48% vs. 68.47–74.77%) on qualitative-synthesis  
226 questions, and frontier LLMs plateau near a 67% L1 ceiling regardless of reasoning effort. For narrow  
227 fact retrieval, specialized systems can therefore match or exceed a broader orchestrator; the advantage  
228 of BIORESEARCHER appears when retrieval steps must be chained with entity normalization, multi-  
229 source provenance, quantitative cohort analysis, and structured synthesis.

230 The open-ended analysis benchmarks reinforce this interpretation. On BixBench, BIORESEARCHER  
231 reaches 89.33% accuracy on the full set and 81.82% on the human-biology subset, competitive with  
232 package-augmented coding agents (Table 4). On BaisBench Scientific Discovery, BIORESEARCHER  
233 achieves a mean score of  $0.758 \pm 0.005$ , exceeding CellType ( $0.641 \pm 0.082$ ) and matching Claude  
234 Code ( $0.759 \pm 0.012$ ; Table 4). These results do not imply that orchestration alone solves autonomous  
235 biological discovery. Rather, they show that a translational agent can retain competitive data-analysis  
236 ability while adding the surrounding machinery that scientific users need: entity grounding, scenario-  
237 specific methodology, source separation, and final dossier construction.

238 The end-to-end benchmark exposes the central tradeoff of the present system. BIORESEARCHER  
239 achieves the highest positive hit rate ( $74.7\% \pm 3.3\%$ ; Table 5) and the highest negative clear rate  
240 ( $96.8\% \pm 0.2\%$ ), exceeding CellType, Medea, and OpenAI Deep Research on both axes. Higher  
241 sensitivity is useful for hypothesis generation, where missing a plausible therapeutic or biomarker  
242 signal is costly, but lower negative clearance is a risk for exclusionary decisions: the system may  
243 preserve weak or exploratory hypotheses that a more conservative pipeline would reject. This argues  
244 for reporting the system as a first-pass evidence-dossier generator, not as an autonomous decision-  
245 maker, and for treating negative claims and absence-of-evidence conclusions as requiring explicit  
246 expert review.

## 247 6.2 Where the System Helps Most

248 The system provides the greatest value on queries with four properties. First, it helps when the  
249 input is **synonym-heavy or ontology-heavy**, such that entity resolution is a first-class part of the  
250 problem rather than preprocessing. Second, it helps when the **answer depends on fusing at least**  
251 **three evidence types**, for example literature, clinical trials, patents, and multi-omics data. Third,  
252 it helps when the **expected output is a long-form research artifact with provenance, ranked**  
253 **hypotheses, and uncertainty** rather than a short answer. Fourth, it helps when the **research plan**  
254 **includes computations that cannot be specified fully in advance**, such as cohort-specific survival  
255 analysis, mutation-stratified expression testing, or genome-wide dependency exploration.

256 The ATR biomarker case study illustrates this regime described in Appendix D. The system translated  
257 a broad synthetic-lethality question into a multi-stage analysis over TCGA and METAPRISM  
258 cohorts, separated pan-cancer from subtype-specific signals, and distinguished mechanistic pathway  
259 activation from candidate therapeutic sensitivity. Its final ranking did not simply maximize statistical  
260 signal: TP53 loss was treated as a robust marker of ATR pathway activation, ATM loss as the most  
261 mechanistically and clinically supported synthetic-lethal biomarker, APC loss as an exploratory  
262 colorectal-cancer-specific candidate, and ARID1A loss as biologically plausible but weaker in the  
263 available patient data. This behavior is aligned with the intended use case: producing a reviewable  
264 translational dossier that exposes evidence, caveats, and prioritization logic.

## 265 6.3 Why Evidence-Centric Artifacts Matter

266 Intermediate artifacts are not merely implementation details; they define the trust boundary of the  
267 system. Each subagent writes a source artifact with its query scope, retrieved identifiers, source  
268 metadata, and local conclusions. The reconciliation agent then operates over these artifacts rather  
269 than over an undifferentiated conversation transcript. This design has four consequences. First,  
270 it **improves auditability**: claims in the final report can be traced back to evidence streams with  
271 PMIDs, NCT IDs, patent numbers, cohort names, or analysis outputs. Second, it **supports modular**  
272 **debugging**: an error in patent retrieval need not contaminate the clinical-trial or quantitative-analysis  
273 artifacts. Third, it **enables partial reproducibility**: downstream drafting and reconciliation can be  
274 rerun over a fixed evidence bus without re-executing every expensive or rate-limited search. Fourth,  
275 it **gives domain experts something more useful than a final answer to review**: they can inspect the  
276 intermediate artifacts that determined the recommendation.

277 This artifact boundary also clarifies the role of claim-level debate. The debate mechanism should  
278 be understood as a reconciliation layer over imperfect evidence, not as a guarantee of truth. It can  
279 reduce single-model drafting artifacts by forcing competing drafts to expose atomic claims and source  
280 support, but it cannot recover evidence that no subagent retrieved and it can still preserve an upstream  
281 hallucination if the cited support is itself wrong or misinterpreted. For this reason, debate is best  
282 viewed as a precision-improving mechanism inside a broader evidence workflow, not as a substitute  
283 for retrieval quality, entity grounding, or expert review.

## 284 **6.4 Human-AI Collaboration**

285 We position BIORESEARCHER as a research copilot for translational scientists: a system that produces  
286 first-draft evidence dossiers, not final scientific or clinical decisions. This framing matters because  
287 translational drug discovery routinely involves uncertain evidence, conflicting incentives, and high  
288 downstream cost. The system is useful when it compresses weeks of preliminary evidence gathering  
289 into a structured dossier that experts can challenge, edit, and extend. It is unsafe if the same dossier is  
290 treated as a self-validating recommendation.

291 The desired workflow is therefore artifact-mediated human review. Experts should inspect the selected  
292 scenario, canonical entities, evidence coverage, quantitative assumptions, and ranked conclusions.  
293 The system should make uncertainty operational by distinguishing established evidence, conflicting  
294 evidence, negative evidence, and exploratory hypotheses. The case study demonstrates this pattern:  
295 the system surfaces TP53, ATM, APC, and ARID1A as different types of signals rather than collapsing  
296 them into a single unqualified biomarker answer.

## 297 **6.5 Limitations and Risks**

298 Several limitations bound the present results. First, the evaluation remains finite and partly single-run:  
299 109 single-step questions, BixBench-Verified-50, the 193-question BaisBench Scientific Discovery  
300 track, and a 30-query end-to-end benchmark. These benchmarks cover important capabilities, but they  
301 do not yet provide confidence intervals over repeated agent runs, systematic per-failure frequencies, or  
302 full ablations of scenario routing, entity grounding, CodeAct analysis, and claim-level debate. Second,  
303 the end-to-end metrics are coarse. Positive hit rate and negative clear rate are useful signals, but they  
304 do not fully measure citation faithfulness, mechanistic plausibility, novelty, actionability, or the cost  
305 of false positives in real portfolio decisions. Third, the single-step benchmark uses LLM-as-judge  
306 scoring; although the expected outputs and rubrics are manually specified, judge errors and rubric  
307 incompleteness remain possible.

308 BIORESEARCHER is bounded by its sources, tools, and intended use. Coverage depends on external  
309 APIs and local snapshots; patents and trial registries overrepresent disclosed, indexable evidence and  
310 often miss proprietary or negative results. The current stack is oncology-centered, so new domains  
311 require new playbooks, ontologies, datasets, and validation tasks. CodeAct analyses are powerful but  
312 sensitive to cohort definition, missingness, multiple testing, and model-chosen analysis paths.

313 The main risks are overconfident synthesis, false completeness, and misuse. Structured dossiers can  
314 look complete even when searches fail or evidence is absent, and weak hypothesis-generation signals  
315 must not drive exclusionary or clinical decisions without expert adjudication. The same workflow  
316 also enables competitive intelligence over public trials, patents, and disclosures; BIORESEARCHER is  
317 not a clinical decision-support system and must not be used to select treatment for individual patients.

## 318 **6.6 Synthesis**

319 The broader lesson is that translational-medicine assistance is not well captured by single-turn  
320 biomedical question answering. Useful systems must translate underspecified questions into method-  
321 ology, normalize entities across biomedical ontologies, gather heterogeneous evidence, run custom  
322 analyses, preserve provenance, and reconcile disagreement under uncertainty. BIORESEARCHER  
323 addresses this regime through a two-tier orchestrator, scenario methodology injection, specialized  
324 retrieval and CodeAct subagents, an evidence bus, and claim-level reconciliation. The results support  
325 the promise of this design most strongly for evidence-rich, analysis-heavy hypothesis generation.  
326 They also identify the next technical bar: stronger negative-evidence handling, systematic ablations,  
327 repeated-run robustness, calibrated uncertainty, and larger expert-audited end-to-end evaluation.

## References

- 328
- 329 [1] Confident AI. DeepEval: The open-source LLM evaluation framework. <https://github.com/confident-ai/deepeval>, 2024. Accessed 2026.
- 330
- 331 [2] Ludovico Mitchener, Jon M. Laurent, Alex Andonian, Benjamin Tenmann, Siddharth Narayanan,  
332 Geemi P. Wellawatte, Andrew White, Lorenzo Sani, and Samuel G. Rodriques. BixBench: a  
333 Comprehensive Benchmark for LLM-based Agents in Computational Biology, October 2025.  
334 arXiv:2503.00096 [q-bio].
- 335 [3] Erpai Luo, Jinneng Jia, Yifan Xiong, Xiangyu Li, Xiaobo Guo, Baoqi Yu, Minsheng Hao,  
336 Lei Wei, and Xuegong Zhang. Benchmarking AI scientists for omics data driven biological  
337 discovery, January 2026. arXiv:2505.08341 [cs].
- 338 [4] Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical  
339 research with large language models. *Nature*, 624(7992):570–578, December 2023.
- 340 [5] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe  
341 Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelli-*  
342 *gence*, 6(5):525–535, May 2024.
- 343 [6] Chenglong Kang, Xiaoyi Liu, and Fei Guo. RetroInText: A Multimodal Large Language Model  
344 Enhanced Framework for Retrosynthetic Planning via In-Context Representation Learning.  
345 October 2024.
- 346 [7] Reza Averly, Frazier N. Baker, Ian A Watson, and Xia Ning. LIDDIA: Language-based  
347 Intelligent Drug Discovery Agent. In Christos Christodoulopoulos, Tanmoy Chakraborty,  
348 Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical*  
349 *Methods in Natural Language Processing*, pages 12004–12028, Suzhou, China, November  
350 2025. Association for Computational Linguistics.
- 351 [8] Martin Pacesa, Lennart Nickel, Christian Schellhaas, Joseph Schmidt, Ekaterina Pyatova, Lucas  
352 Kissling, Patrick Barendse, Jagrity Choudhury, Srajan Kapoor, Ana Alcaraz-Serna, Yehlin Cho,  
353 Kourosh H. Ghamary, Laura Vinué, Brahm J. Yachnin, Andrew M. Wollacott, Stephen Buckley,  
354 Adrie H. Westphal, Simon Lindhoud, Sandrine Georgeon, Casper A. Goverde, Georgios N.  
355 Hatzopoulos, Pierre Gönczy, Yannick D. Muller, Gerald Schwank, Daan C. Swarts, Alex J.  
356 Vecchio, Bernard L. Schneider, Sergey Ovchinnikov, and Bruno E. Correia. One-shot design of  
357 functional protein binders with BindCraft. *Nature*, 646(8084):483–492, October 2025.
- 358 [9] Mingming Zhu, Jiahua Rao, Xiaoyu Chen, Qianmu Yuan, and Yuedong Yang. Advancing  
359 Protein Design via Multi-Agent Reinforcement Learning with Pareto-Based Collaborative  
360 Optimization, January 2026. ISSN: 2692-8205 Pages: 2026.01.13.699365 Section: New  
361 Results.
- 362 [10] Manvitha Ponnampati, Sam Cox, Cade W. Gordon, Michael J. Hammerling, Siddharth Narayanan,  
363 Jon M. Laurent, James D. Braza, Michaela M. Hinks, Michael D. Skarlinski, Samuel G.  
364 Rodriques, and Andrew White. ProteinCrow: A Language Model Agent That Can Design  
365 Proteins. July 2025.
- 366 [11] Yusuf H. Roohani, Andrew H. Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang,  
367 Alexander Marson, Percy Liang, and Jure Leskovec. BioDiscoveryAgent: An AI Agent for  
368 Designing Genetic Perturbation Experiments. October 2024.
- 369 [12] Alireza Ghafarollahi and Markus Buehler. ProtAgents: Protein discovery via large language  
370 model multi-agent collaborations combining physics and machine learning. March 2024.
- 371 [13] Zuojun Xu, Feng Ren, Ping Wang, Jie Cao, Chunting Tan, Dedong Ma, Li Zhao, Jinghong  
372 Dai, Yipeng Ding, Haohui Fang, Huiping Li, Hong Liu, Fengming Luo, Ying Meng, Pinhua  
373 Pan, Pingchao Xiang, Zuke Xiao, Sujata Rao, Carol Satler, Sang Liu, Yuan Lv, Heng Zhao,  
374 Shan Chen, Hui Cui, Mikhail Korzinkin, David Gennert, and Alex Zhavoronkov. A generative  
375 AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial.  
376 *Nature Medicine*, 31(8):2602–2610, August 2025.

- 377 [14] Shanghua Gao, Richard Zhu, Zhenglun Kong, Ayush Noori, Xiaorui Su, Curtis Ginder,  
378 Theodoros Tsiligkaridis, and Marinka Zitnik. TxAgent: An AI Agent for Therapeutic Reasoning  
379 Across a Universe of Tools, March 2025.
- 380 [15] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeon-  
381 hoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. MDAgents: An Adaptive  
382 Collaboration of LLMs for Medical Decision-Making. November 2024.
- 383 [16] Dyke Ferber, Omar S. M. El Nahhas, Georg Wölflein, Isabella C. Wiest, Jan Clusmann,  
384 Marie-Elisabeth Leßmann, Sebastian Foersch, Jacqueline Lammert, Maximilian Tschochohei,  
385 Dirk Jäger, Manuel Salto-Tellez, Nikolaus Schultz, Daniel Truhn, and Jakob Nikolas Kather.  
386 Development and validation of an autonomous artificial intelligence agent for clinical decision-  
387 making in oncology. *Nature Cancer*, 6(8):1337–1349, August 2025.
- 388 [17] Qiao Jin, Zifeng Wang, Charalampos S. Floudas, Fangyuan Chen, Changlin Gong, Dara  
389 Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. Matching patients  
390 to clinical trials with large language models. *Nature Communications*, 15(1):9074, November  
391 2024.
- 392 [18] Anurag Jayant Vaidya, Felix Meissen, Daniel C. Castro, Shruthi Bannur, Tristan Lazard,  
393 Drew FK Williamson, Faisal Mahmood, Javier Alvarez-Valle, Stephanie Hyland, and Kenza  
394 Bouzid. NOVA: An Agentic Framework for Automated Histopathology Analysis and Discovery.  
395 November 2025.
- 396 [19] Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The Virtual Lab of  
397 AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646(8085):716–723, October 2025.
- 398 [20] Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan  
399 Li, Lin Qiu, Gavin Li, Junze Zhang, Di Yin, Shruti Marwaha, Jennefer N. Carter, Xin Zhou,  
400 Matthew Wheeler, Jonathan A. Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael  
401 Snyder, Le Cong, Aviv Regev, and Jure Leskovec. Biomni: A General-Purpose Biomedical AI  
402 Agent. *bioRxiv: The Preprint Server for Biology*, page 2025.05.30.656746, June 2025.
- 403 [21] Ruofan Jin, Yucheng Guo, Yuanhao Qu, Ming Yang, Chun Shang, Qirong Yang, Linlin Chao,  
404 Yi Zhou, Ruilai Xu, Ziyao Xu, Ruhong Zhou, Zaixi Zhang, Mengdi Wang, Xiaoming Zhang,  
405 and Le Cong. BioLab: End-to-End Autonomous Life Sciences Research with Multi-Agents  
406 System Integrating Biological Foundation Models, September 2025. ISSN: 2692-8205 Pages:  
407 2025.09.03.674085 Section: New Results.
- 408 [22] Zhongyue Zhang, Zijie Qiu, Yingcheng Wu, Shuya Li, Dingyan Wang, Zhuomin Zhou,  
409 Duo An, Yuhan Chen, Yu Li, Yongbo Wang, Chubin Ou, Zichen Wang, Jack Xiaoyu Chen,  
410 Bo Zhang, Yusong Hu, Wenxin Zhang, Zhijian Wei, Runze Ma, Qingwu Liu, Bo Dong, Yuexi  
411 He, Qiantai Feng, Lei Bai, Qiang Gao, Siqi Sun, and Shuangjia Zheng. OriGene: A Self-  
412 Evolving Virtual Disease Biologist Automating Therapeutic Target Discovery, June 2025.  
413 Pages: 2025.06.03.657658 Section: New Results.
- 414 [23] Pengwei Sui, Michelle M. Li, Shanghua Gao, Wanxiang Shen, Valentina Giunchiglia, Andrew  
415 Shen, Yepeng Huang, Zhenglun Kong, and Marinka Zitnik. Medea: An omics AI agent for  
416 therapeutic discovery. *bioRxiv: The Preprint Server for Biology*, page 2026.01.16.696667,  
417 January 2026.
- 418 [24] Houcheng Su, Weicai Long, and Yanlin Zhang. BioMaster: Multi-agent System for Automated  
419 Bioinformatics Analysis Workflow, January 2025. Pages: 2025.01.23.634608 Section: New  
420 Results.
- 421 [25] Joshua Pickard, Ram Prakash, Marc Andrew Choi, Natalie Oliven, Cooper Stansbury, Jillian  
422 Cwycyshyn, Alex Gorodetsky, Alvaro Velasquez, and Indika Rajapakse. Language Model  
423 Powered Digital Biology with BRAD, December 2024. arXiv:2409.02864 [cs].
- 424 [26] Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping  
425 Day, Christina Ross, Robert Leaman, and Zhiyong Lu. GeneAgent: self-verification language  
426 agent for gene-set analysis using domain databases. *Nature Methods*, 22(8):1677–1685, August  
427 2025.

- 428 [27] Yuxing Lu, Wei Wu, Xukai Zhao, Rui Peng, and Jinzhuo Wang. KARMA: Leveraging Multi-  
429 Agent LLMs for Automated Knowledge Graph Enrichment. October 2025.
- 430 [28] Timothy A. Yap, Elisa Fontana, Elizabeth K. Lee, David R. Spiegel, Martin Højgaard, Stephanie  
431 Lheureux, Niharika B. Mettu, Benedito A. Carneiro, Louise Carter, Ruth Plummer, Gregory M.  
432 Cote, Funda Meric-Bernstam, Joseph O'Connell, Joseph D. Schonhoft, Marisa Wainszelbaum,  
433 Adrian J. Fretland, Peter Manley, Yi Xu, Danielle Ulanet, Victoria Rimkunas, Mike Zinda,  
434 Maria Koehler, Ian M. Silverman, Jorge S. Reis-Filho, and Ezra Rosen. Camonsertib in DNA  
435 damage response-deficient advanced solid tumors: phase 1 trial results. *Nature Medicine*,  
436 29(6):1400–1411, June 2023.
- 437 [29] Emily Clough, Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F  
438 Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman,  
439 et al. NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update.  
440 *Nucleic Acids Research*, 52(D1):D138–D144, 2024.

## 441 **A Case Study: Generation of Novel Biomarker Hypotheses using Synthetic** 442 **Lethality Concept and Patient Data**

443 We evaluated the TRESR trial [28] (NCT04497116), which is based on a synthetic-lethality (SL)  
444 hypothesis linking ATR inhibition to selected loss-of-function (LoF) alterations. While the study  
445 refers to “ATR inhibitor–sensitizing mutations,” these are not explicitly defined. Additional research  
446 identified a report suggesting a focus on DNA damage response genes identified via chemogenomic  
447 CRISPR screens [PMID: 37277454].

448 Thus, we tasked BIORESEARCHER with generating SL biomarker hypotheses for ATR inhibition  
449 based on next-generation sequencing and clinical data (input specification in Figure 2). The agent  
450 autonomously selected TCGA and METAPRISM cohorts, identified LoF events associated with  
451 ATR expression, and evaluated candidates through survival analyses. Specifically, the system  
452 executed an end-to-end translational pipeline within a LangGraph-based deep agent orchestration  
453 framework (total runtime  $\approx 1352$  s; see Figure 7 for the full pipeline view). Scenario classification  
454 routed the query to the `combination_discovery` workflow, after which the `translation` subagent  
455 canonicalized ATR and the relevant clinical indications. Subsequently, the `insights_and_signals`  
456 (CodeAct) subagent conducted downstream multi-omic and survival analyses across multiple cohorts  
457 and selected patients with advanced solid tumors, leveraging iterative model invocations and tool-  
458 mediated operations to complete the analysis under the system-generated methodology summarized  
459 in Figure 3. This very closely resembled the clinical trial design.

460 The BIORESEARCHER output combined patient-level data with literature and clinical context, evalu-  
461 ating two signal types: (i) LoF-associated ATR upregulation and (ii) survival patterns in low-ATR  
462 tumors as a proxy for drug-induced ATR inhibition. A condensed reasoning trace integrating these  
463 signals into a candidate ranking is shown in Figure 4.

464 Across both cohorts, increased ATR expression was most robustly associated with TP53 loss, suggest-  
465 ing a pan-cancer marker of ATR pathway activation, though likely reflecting general replication stress  
466 rather than specific therapeutic sensitivity. In contrast, the strongest evidence for a synthetic-lethal  
467 interaction emerged from ATM loss, which - despite lacking compensatory ATR upregulation -  
468 showed the clearest survival signal in low-ATR tumors, consistent with prior preclinical and clinical  
469 findings.

470 Subtype-specific signals further highlighted APC loss in colorectal cancer as the most prominent data-  
471 driven discovery, supported by both elevated ATR expression and favorable low-ATR survival patterns,  
472 although it lacks prior clinical validation. Finally, ARID1A loss demonstrated only modest support in  
473 patient datasets but remains biologically plausible given consistent evidence from external studies,  
474 warranting consideration as a secondary translational candidate. The agent-generated biological  
475 interpretation across these candidates is summarized in Figure 5.

476 Overall, the integrated evidence supports the following ranked hypotheses: TP53 LoF as the strongest  
477 pan-cancer marker of ATR pathway activation; ATM LoF as the most clinically and mechanistically  
478 supported synthetic-lethal biomarker; APC LoF in colorectal cancer as an unexpected data-driven  
479 candidate; and ARID1A LoF as a literature-supported but modestly recovered clinical candidate  
480 (clinically actionable prioritization in Figure 6). The framework’s key strength lies in combining  
481 autonomous multi-cohort analysis with calibrated restraint: where survival models were sparse or  
482 unstable, the agent avoided over-interpretation, preserving the reliability of the resulting translational  
483 dossier.

## 484 **B Domain-Specific Components and Tool Specifications**

485 This appendix expands the domain-specific component stack summarized in Section 2 and provides  
486 parameter schemas for representative tools from each major category. All tools are implemented  
487 as Python functions with type annotations; a `@tool` decorator exposes them to agents as structured  
488 function-calling tools with auto-generated JSON schemas.

## 489 B.1 Entity Resolution and Biomedical Knowledge Grounding

490 The **translation** subagent is the foundation for all downstream research. It resolves entity ambiguity  
491 by mapping natural-language mentions to canonical identifiers using a ReAct agent with eight tools  
492 backed by local SQLite databases (ChEMBL 36, Open Targets) with API fallback:

- 493 • `target_name_to_chembl_id`: gene/protein name → ChEMBL target ID
- 494 • `molecule_name_to_chembl_id`: drug name → ChEMBL molecule ID
- 495 • `target_to_drug_chembl_ids`: target → associated drugs (Phase  $\geq 2$ )
- 496 • `compound_smiles_to_name_and_chembl_id`: SMILES → compound identity
- 497 • `get_drug_indications`: drug → approved/investigated indications
- 498 • `disease_name_to_efo_id`: disease name → EFO/MONDO ontology ID
- 499 • `molecule_smiles_to_molecule_synonyms`: SMILES → all known synonyms (Pub-  
500 Chem)
- 501 • `get_gene_id_to_ensembl_id_mapping`: gene symbol → Ensembl ID

502 Local SQLite is preferred over API access for latency and reproducibility; a swappable backend  
503 via `translation_factory.py` supports both modes. The output is a structured entity table with  
504 canonical IDs consumed by all downstream subagents, ensuring consistent entity resolution across all  
505 evidence streams.

506 **Concrete Example.** For the query “*Topo-I inhibitor combinations in NSCLC*,” the translation  
507 subagent resolves: (i) “Topo-I” → gene symbol TOP1 → ChEMBL target CHEMBL1781 → Ensembl  
508 ENSG00000198900; (ii) “NSCLC” → EFO\_0003060 (cross-mapped to TCGA “LUAD”/“LUSC”  
509 and DepMap disease contexts); (iii) associated compounds: Irinotecan (CHEMBL481), Topotecan  
510 (CHEMBL84), Trastuzumab Deruxtecan (CHEMBL4297564), among 24 Phase  $\geq 2$  drugs retrieved  
511 via `target_to_drug_chembl_ids`. This single grounding step prevents downstream subagents  
512 from searching for “topoisomerase I” in one database and “TOP1” in another without recognizing  
513 them as equivalent.

## 514 B.2 Tool Ecosystem

515 The system integrates 30+ tools spanning six ML endpoints and multiple external data sources  
516 (Table 6). Table 6 provides a categorized overview.

Table 6: Tool ecosystem organized by category, data source, and interaction pattern. The 30+ tools span structured function-calling (Pattern 1–3 subagents) and CodeAct sandbox functions (Pattern 4). ML model endpoints provide predictions for drug synergy.

Category	Tools	Data Source	Pattern
Entity Resolution	8 translation tools	ChEMBL 36, Open Targets, PubChem	Local SQLite / REST
Literature	<code>search_pubmed</code> , <code>fetch_abstracts</code> , etc. (4)	PubMed / NCBI E-utilities	REST API
Literature	<code>get_target_disease_evidence</code> , etc. (4)	Open Targets	GraphQL API
Text data RAG	<code>search_conference_abstracts</code> (2)	Licensed text data sources	ChromaDB
Clinical Trials	<code>search_clinical_trials</code> (2)	ClinicalTrials.gov v2 / AACT	REST / SQLite
Patents	<code>search_patents</code> (2)	Google Patents	SerpAPI
Target Profile	<code>get_basal_protein_info</code> , etc. (4)	DepMap, HPA	Parquet + API
Genome-Wide	Dependency, expression, mutation tools (11)	DepMap 24Q4 multi-omics	CodeAct sandbox
PPI/Pathways	<code>get_gene_pair_annotations</code> , etc. (3)	STRING/Reactome/GO	CodeAct sandbox
Drug Synergy	<code>predict_drug_combination</code> (1)	DrugComb v2	HF Gradio API

## 517 B.3 Conference Abstract and other licensed data sources - RAG

518 Conference abstracts and other licensed data sources contain the earliest reports of clinical results,  
519 often 6–12 months before peer-reviewed publication. This latency gap is critical for translational  
520 research.

521 The indexing pipeline processes conference PDFs via PyMuPDF with conference-specific chunking  
522 strategies. Metadata fields are extracted per chunk: conference name, year, page, abstract ID, section  
523 header (e.g., “Breast Cancer: Metastatic”), inferred disease type (mapped to 12 canonical categories  
524 including `nsc1c`, `colorectal_cancer`, `melanoma`), gene symbols (regex-matched against 33 key  
525 oncology genes including KRAS, EGFR, BRAF, TOP1, BRCA1/2), drug mentions, trial phase

526 (regex-extracted), NCT IDs (pattern `NCT\d{8}`), and abstract type (late-breaking, oral, poster).  
527 Embeddings are computed via SentenceTransformer (`all-MiniLM-L6-v2`) and stored in ChromaDB  
528 with metadata filtering, enabling queries such as “*find late-breaking abstracts mentioning TOP1 in*  
529 *NSCLC.*”

#### 530 B.4 CodeAct Sandbox for Genome-Wide Analysis

531 The `insights_and_signals` subagent operates a sandboxed Python execution environment pre-loaded  
532 with nine DepMap 24Q4 Parquet datasets spanning  $\sim 19\text{K}$  genes  $\times$   $\sim 1,600$  cell lines:

- 533 • **Dependency screens:** CRISPR gene effect (Chronos), CRISPR dependency probability,  
534 shRNA gene effect (DEMETER2), shRNA dependency probability
- 535 • **Omics:** Gene expression ( $\log_2(\text{TPM}+1)$ ), copy number ( $\log_2(\text{CN ratio}+1)$ ), proteomics  
536 (normalized abundance), somatic mutations (binary), loss-of-function mutations (binary)

537 The namespace includes pre-loaded `pandas`, `numpy`, and `scipy.stats`, plus a `THRESHOLDS` dic-  
538 tionary encoding expert-curated biological cutoffs: CRISPR likely dependent ( $< -0.5$ ), strongly  
539 dependent ( $< -1.0$ ), dependency probability thresholds (dependent  $> 0.6$ , resistant  $< 0.4$ ), copy-  
540 number gain ( $> 1.5$ ), amplification ( $> 1.9$ ), loss ( $< 0.6$ ), FDR significance ( $< 0.1$ ), and minimum  
541 sample size ( $\geq 3$ ). These thresholds ground the LLM’s statistical analyses in accepted DepMap  
542 conventions, preventing arbitrary cutoff selection.

543 **Security Model.** Table 7 summarizes the sandbox security constraints.

Table 7: CodeAct sandbox security model.

Constraint	Details
Blocked builtins	<code>exec</code> , <code>eval</code> , <code>compile</code> , <code>__import__</code> , <code>open</code> , <code>input</code> , <code>breakpoint</code>
Execution timeout	600 s per code block
Round cap	Maximum 12 code blocks per session
Output truncation	8,000 chars (stdout), 4,000 chars (stderr)
Variable persistence	Only <code>msgpack</code> -serializable primitives ( <code>str</code> , <code>int</code> , <code>float</code> , <code>bool</code> , <code>list</code> , <code>dict</code> ); DataFrames and numpy arrays ex- cluded
Namespace isolation	Filtered <code>__builtins__</code> copy; no filesystem or network access

544 **Edge-Case Handling.** The sandbox addresses three failure modes: (1) *resource exhaustion*: code  
545 exceeding the 600 s timeout or 8,000-char output limit is killed mid-execution, and the LLM receives  
546 truncated output with a diagnostic message; (2) *blocked operations*: attempts to call restricted builtins  
547 raise `NameError`, which the LLM can observe and self-correct; (3) *serialization failures*: variables  
548 that cannot be `msgpack`-serialized (e.g., `matplotlib` figures) are silently dropped between code blocks,  
549 preventing state pollution.

550 **Supported Analyses.** 18 callable tool functions span genome-wide dependency correlations,  
551 expression–dependency associations, mutation stratification (Mann–Whitney  $U$ ), synthetic lethality  
552 analysis (absolute thresholds + percentile ranking), PPI network analysis, GEO [29] expression  
553 outlier detection, and drug combination synergy prediction.

#### 554 B.5 Parameter Schemas for Representative Tools

555 **Patent Search** (`search_patents`). The most complex tool schema, supporting both textual and  
556 chemical structure queries:

557 SMILES inputs are canonicalized via `RDKit StandardizeSmiles`; InChIKeys are validated by  
558 regex. Backend: `SerpAPI google_patents` engine with field-code query construction (e.g., `TI=` for  
559 title, `SSS=(struct)` for substructure).

Parameter	Type	Description
text_terms	List[TextTerm]?	Structured query terms. Each TextTerm specifies value (phrase), where $\in$ {TITLE, ABSTRACT, CLAIM, FULL_DOCUMENT}, match $\in$ {PARTIAL, EXACT}
chemical_terms	List[ChemicalTerm]?	Chemical structure queries. Each ChemicalTerm specifies value (SMILES or InChIKey), molecule_type $\in$ {SMILES, INCHI_KEY}, chemical_search_type $\in$ {EXACT, SIMILAR, SUBSTRUCTURE}
text_connector	QueryLogic	AND/OR for combining text terms
assignees	List[str]?	Patent assignee (company) names
status	StatusTypes	GRANT / APPLICATION / ALL
start_date, end_date	str?	Date range (YYYY/MM/DD format)
limit	int	Results cap (10–100, default 10)

Parameter	Type	Description
condition	str?	Disease/condition; supports AND/OR syntax, e.g., ((NSCLC) OR (lung cancer)) AND (EGFR)
intervention	str?	Drug or treatment name; supports AND/OR
term	str?	Free-text keyword search across all study fields
status	List[str]?	Phase filter: RECRUITING, COMPLETED, ACTIVE_NOT_RECRUITING, etc. (14 valid values)
date_range	str?	Primary completion date range, e.g., 2020-01-01..2023-12-31
limit	int	Max results (default 50, max 1000)

560 **Clinical Trial Search** (`search_clinical_trials`). Default returned fields: NCTId, BriefTitle, OverallStatus, BriefSummary, Condition, Phase, InterventionName. Backend: ClinicalTrials.gov REST API v2. A local AACT SQLite variant with identical signature is available for  
561  
562 offline/reproducible evaluation.  
563

Parameter	Type	Description
query	str	PubMed query using native syntax: field tags ([Title/Abstract], [MeSH Terms], [Gene]), Boolean operators, quoted phrases
max_results	int	Result limit (default 80)
sort	relevance   date	Sort order
min_date, max_date	str?	Publication date range (YYYY/MM/DD)

564 **PubMed Search** (`search_pubmed`). Returns PMID, title, journal, and year per result. On zero  
565 results, the tool returns PubMed’s `querytranslation` and lists unrecognized terms as diagnostic  
566 feedback, enabling the agent to reformulate queries. Backend: NCBI E-utilities (`esearch`  $\rightarrow$   
567 `esummary`) with exponential backoff on rate limiting.

## 568 C Benchmarks Construction

### 569 C.1 Single-Step Tests: Subsets

Table 8: Single-step test subsets, layer coverage, and a representative query per subset. Each subset targets a specific system layer and is scored independently with its own rubric.

Subset	Layer Tested	N	Representative Query
Ontology & Entity	Entity resolution / grounding	33	“What is the Ensembl gene ID and ChEMBL target ID for CD340?”
Qualitative Synthesis	Literature / trials / patents retrieval	37	“Provide the title of the publication with PMID 25439351.”
Quantitative Analysis	DepMap / omics computation	39	“What is the CRISPR gene-effect score for PARP1 in MCF7?”
<b>Total</b>		<b>109</b>	

## 570 C.2 Single-Step Tests: Benchmark Design

571 The benchmark consists of 109 questions requiring single-step reasoning and is partitioned into three  
572 subsets (layers), summarized in Table 8. **Ontology & Entity** (L1) challenges entity normalization  
573 across biomedical ontologies. This layer focuses mostly on the ability of the agent to retrieve  
574 key entities (e.g. genes, diseases) from the free-text input and map them to their canonical IDs.  
575 Another example task is resolving a gene family name to a list of member genes with canonical  
576 IDs. **Qualitative Synthesis** (L2) evaluates retrieval of qualitative knowledge from biomedical  
577 databases. The questions check (among others) for facts about genes, compounds, protein functions  
578 and pathways, clinical trials, patents, and literature. Example tasks are retrieving clinical trial NCT  
579 IDs for a given context or assessing the success of a clinical trial given its NCT number. **Quantitative**  
580 **Analysis** (L3) assesses extraction and computation of key quantitative data points. An example task  
581 is checking the expression level of a given gene in a given cell line.

582 **Test-case format.** Each test case is a pair of a question (a natural-language query for the agent)  
583 and an expected output (example natural-language response that the agent is expected to return). For  
584 example, the expected output for the question “*What is the Ensembl gene ID and ChEMBL target ID*  
585 *for CD340?*” is as follows.

586 Ensembl ID: ENSG00000141736  
587 ChEMBL target ID: ChEMBL1824

588 Each expected output was manually verified by an expert to be correct.

## 589 C.3 Single-Step Tests: Evaluation Protocol

590 **Execution.** Every question is submitted end-to-end to the full master graph: the top-level orchestra-  
591 tor dispatches it to the appropriate sub-agent, which invokes tools as needed. No sub-agent is tested  
592 in isolation, so the single-step tests also probe the orchestrator’s routing decisions. The agent’s *actual*  
593 *output* for scoring is the natural-language content of the final message emitted by the master graph.

594 **Judge.** We use DeepEval [1], an LLM-as-a-judge framework, to score open-ended natural-language  
595 outputs against the expected outputs. For each test case, the judge is shown the question, the actual  
596 output, and the expected output, and assigns a score between 0 and 1, together with a free-text  
597 justification. Each test case is paired with a custom DeepEval metric that instructs the judge how  
598 to score the output. A metric specification consists of (i) natural-language evaluation criteria, (ii) a  
599 rubric, i.e. a list of tier descriptions mapped to disjoint score bands on [0, 1], and (iii) a threshold  
600 score (the minimum score required for the test to pass). An example criteria block is:

601 Evaluation dimensions:

- 602 1. Canonical entity: When the expected output labels a canonical name (e.g. drug name,  
603 disease name), is it present and unambiguously identified in the actual output?
- 604 2. Identifier match: Do all identifiers labelled in the expected output appear  
605 in the actual output with matching values?
- 606 3. No fabrication: The actual output does not introduce alternative identifiers  
607 that contradict the expected ones, and does not confuse the entity with a same-string  
608 alias of a different gene/disease/drug.

609 A three-tier rubric yields the bands [0.0, 0.2], [0.4, 0.6], [0.8, 1.0]; the gaps make tier boundaries  
610 unambiguous and the judge commits to one tier per test case. In each custom metric, the threshold  
611 score is set so that a test passes if the judge assigns a score in the highest band (e.g. 0.8 for three-tier  
612 rubrics). An example three-tier rubric for entity-resolution questions (like the example for layer  
613 *Ontology & Entity* in Table 8) is shown in Table 9. We used DeepEval’s default LLM model, which  
614 is GPT 4.1, for all metrics.

615 **Metric.** The primary per-subset metric is the *pass rate*: the fraction of test cases in that subset  
616 whose judge-assigned score satisfies  $\text{score} \geq \text{threshold}$  (i.e. such that were classified into the top tier  
617 of the rubric). Another metric is the average metric score: the average score of the test cases.

Table 9: Example rubric: the judge assigns one tier, yielding a score in the corresponding band.

Score band	Expected outcome specification
[0.0, 0.2]	Identifiers absent or contradicting the expected ones, or canonical entity (when labelled in the expected output) wrong or missing.
[0.4, 0.6]	Most identifiers correct but some missing or partially mismatched, or canonical entity (when labelled) partially off. Additional information contradicts the expected output.
[0.8, 1.0]	All expected identifiers present and matching, canonical entity (when labelled) correct, no fabricated alternatives. If additional information is present, it doesn't contradict the expected output.

#### 618 C.4 BixBench

619 We use a subset of BixBench restricted to *Homo sapiens* data, which probes open-ended biomedical  
 620 data analysis under partial specification. We report BixBench-style open-answer accuracy on  
 621 phylobio/BixBench-Verified-50 [2] (Table 4) for the full 50-question set and a 22-question  
 622 *human* subset of questions that concern human biology. For fairness, package-augmented Claude  
 623 Code variants were provided the same data-analysis package list available to BIORESEARCHER,  
 624 ensuring comparable minimal execution context.

#### 625 C.5 BaisBench

626 We use the Scientific Discovery (BAIS-SD) track of BaisBench [3]: 193 data-driven questions on  
 627 real single-cell RNA-seq datasets, grounded in biological conclusions from 41 published single-cell  
 628 studies, where agents analyze the provided data to select answers in single- and multi-answer formats  
 629 as in the benchmark. Scoring is trichotomous (0, 0.5, or 1) per the SD task: single-answer items  
 630 receive 0 or 1; multi-answer items receive 1 if all correct options are selected and no incorrect one,  
 631 0.5 if at least one correct option is selected and no incorrect one (but not the full set), and 0 otherwise.  
 632 We report the mean  $S_{SD}$  in Table 4.

633 **Item construction.** Candidate items were drafted by domain experts and cross-checked against  
 634 at least one primary source per item (regulatory document, registered trial with reported biomarker  
 635 analysis, or peer-reviewed publication). Items were retained only if (i) the positive answer was  
 636 unambiguously supported by the primary source and (ii) the pairing was not trivially derivable from  
 637 the query string itself (e.g., we excluded items where the biomarker name appeared verbatim in  
 638 the drug's INN or trial title). Tiers 2 and 3 were deliberately included to stress-test failure modes  
 639 observed in pilot evaluations: target/biomarker conflation in Tier 2, and the gap between mechanistic  
 640 plausibility and clinical translation in Tier 3.

#### 641 C.6 End-2-End Clinical Biomarkers Benchmark Construction

642 **Scope and design rationale.** The benchmark targets a single, recurring translational scenario:  
 643 given a therapy (approved or investigational), an MOA, or an indication, identify a biomarker with  
 644 documented clinical or preclinical relevance. Items were stratified by evidential maturity so that ag-  
 645 gregate scores can be decomposed into (a) recall of canonical, label-level facts and (b) generalization  
 646 to settings where signal is distributed across heterogeneous sources (trial registries, mechanistic liter-  
 647 ature, functional-genomics screens, observational prognostic studies). Tier-level expected accuracy  
 648 was set *a priori* based on the breadth and consistency of available primary evidence.

649 **Category specification.** The 30 items are distributed across seven tiers as summarized in Table 10.

650 **Negative controls.** For each item, an expert-authored negative biomarker was selected from adjacent  
 651 biology—same pathway, same indication class, or same therapeutic modality—such that surface-level  
 652 retrieval (keyword overlap, co-occurrence in literature) would plausibly surface it. This design  
 653 penalizes systems that retrieve by topical proximity rather than evidence-grounded reasoning.

Table 10: Tier structure of the translational-medicine biomarker benchmark. Tier-level expected recall reflects *a priori* difficulty estimates based on evidential maturity and source breadth.

Tier	Construct tested	Source archetype	n
1	Clinically-approved biomarker for an approved therapy (canonical pairing)	FDA/EMA label, NCCN guideline	4
2	Clinically-approved biomarker – testing detailed description	Label + pivotal trial	4
3	Synthetic-lethality biomarker with clinical evidence, no full approval	Phase I/II trials, mechanistic papers	4
4	Biomarker reported in trials of non-approved therapies	ClinicalTrials.gov, trial publications	4
5	Biomarker with extensive preclinical validation (non-approved therapies)	Peer-reviewed preclinical studies	5
6	Biomarker from synthetic-lethality discovery studies	CRISPR/RNAi screens, DepMap-class evidence	5
7	Prognostic biomarker for a specific indication	Observational and meta-analytic literature	4

654 **Query template.** Each benchmark item is submitted to the system using the fixed prompt below,  
 655 with the placeholders [Target (MoA)] and [Indication] replaced by the corresponding fields  
 656 from the item record:

657 *Your task is:*

658 Given a drug with its known molecular target(s): [Target (MoA)], and an indication  
 659 [Indication], return a prioritized list of candidate molecular biomarkers that may predict  
 660 treatment response, resistance, sensitivity, or patient enrichment.

661 *Definitions:*

- 662 • *Target* — the molecular entity or pathway directly modulated by the drug.
- 663 • *Biomarker* — a molecular feature distinct from the direct drug target, such as mutation,  
 664 amplification, deletion, fusion, expression change, pathway activation state, homologous  
 665 recombination deficiency marker, co-mutation pattern, or other molecular alteration  
 666 associated with response or resistance.
- 667 • *Indication* — disease / tumor type / clinical context.

668 **Scoring.** For each item  $i$  with positive answer  $p_i$ , negative control  $n_i$ , and a system returning a  
 669 Top-10 list  $L_i$ , we report two metrics computed separately and then averaged across all  $N$  benchmark  
 670 items: the *positive sample rate*

$$\text{PSR} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[p_i \in L_i], \quad (1)$$

671 i.e. the fraction of items whose positive answer is recovered, and the *negative sample rate*

$$\text{NSR} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[n_i \notin L_i], \quad (2)$$

672 i.e. the fraction of items whose negative control is correctly excluded. The whole benchmark is run  
 673 three times per system, and the reported PSR and NSR are averaged across the three runs. Tier-level  
 674 subscores are reported alongside the aggregate to expose where systems trade off canonical recall  
 675 against generalization to less mature evidence.

Table 11: Full specification of the 30-item translational-medicine biomarker benchmark (Tiers 1–4). Each item pairs a therapy/MOA/indication with an expert-validated positive biomarker and an expert-authored negative control drawn from adjacent biology.

#	Tier	Drug (modality)	Target (MOA)	Indication	Biomarker (positive)	(posi- Negative	Reference
1	T1	Trametinib	MEK (MAP2K1, MAP2K2)	Melanoma	BRAF	TP53	NCT01245062
2	T1	Trastuzumab deruxtecan (T-DXd)	ERBB2, TOP1	Breast, NSCLC, gastric	HER2	TACSTD2	NCT03734029; NCT01275677
3	T1	Vismodegib	SMO	Basal cell carcinoma; medulloblastoma	PTCH1	CTNNB1	NCT00833417
4	T1	Palbociclib	CDK4/CDK6	Advanced breast cancer	ER, HER2	CCND1	NCT00721461; NCT01740427
5	T2	Osimertinib	EGFR	Metastatic & adjuvant NSCLC	EGFR activating mutations (ex19del, L858R; T790M for 2L)	BRCA1	NCT02296125
6	T2	Erdafitinib	FGFR1/2/3/4	Metastatic urothelial carcinoma	FGFR3 mutations (R248C, G370C, S249C, Y373C); FGFR3–TACC3 fusion	CD274	NCT03390504
7	T2	Mirvetuximab soravtansine	FOLR1, TUBB	Ovarian cancer	FR $\alpha$ IHC $\geq$ 75% cells at $\geq$ 2+/3+	BRCA1/2	NCT02631876; NCT04296890
8	T2	Elacestrant	ESR1	Advanced breast cancer	ESR1 LBD missense mutations (ctDNA)	ER	NCT03778931
9	T3	Entrectinib	ROS1	Lobular breast, gastric, TNBC	CDH1	RHOA	NCT04551495; NCT03620643
10	T3	CPI-0209	EZH2	Advanced solid tumours & lymphomas	ARID1A	KMT2D	NCT04104776
11	T3	RP-6306	PKMYT1	Advanced solid tumours	FBXW7	TP53	NCT04855656
12	T3	Onvansertib	PLK1	Metastatic CRC	KRAS	APC	NCT03829410
13	T4	Everolimus	mTOR	Thyroid cancer	TSC1	ESR1	PMID:25295501
14	T4	Olaparib	PARP1, PARP2	Prostate cancer	BRIP1	PARG	NCT02987543
15	T4	Vemurafenib	BRAF	Pancreatic cancer	KRAS	PTEN	NCT01524978
16	T4	Palbociclib	CDK4/6	Myeloma	CDKN2A	MDM2	NCT00555906

Table 12: Full specification of the 30-item translational-medicine biomarker benchmark (Tiers 5–7), continued from Table 11.

#	Tier	Drug (modality)	Target (MOA)	Indication	Biomarker (positive)	(posi- Negative	Reference
17	T5	—	BRAF	Melanoma	MAP2K1	BRCA1/2	NCT01006980
18	T5	—	HSP90	Acute lymphoblastic leukemia	JAK1	CDKN2A	PMID:22271575
19	T5	—	AURKA	NSCLC	SMARCA4	TP53	PMID:28102363
20	T5	—	CDK1/2/5/9	Osteosarcoma	CCNE1	VEGFA	PMID:30266815
21	T5	—	FANCM	Pan-cancer	SMARCAL1	POLE2	PMID:39510066
22	T6	—	PRKDC	Pan-cancer	ATM	POLR2G	PMID:23761041
23	T6	—	PIK3CB	Pan-cancer	PTEN	ARID1A	PMID:18594509
24	T6	—	ROCK2	Pan-cancer	BRCA2	IDH1	PMID:37073955
25	T6	—	LIG1	Pan-cancer	BRCA2	POLD1	PMID:41070620
26	T6	—	FEN1	Pan-cancer	WDR48	RUVBL2	PMID:40205037
27	T7	—	—	Myelodysplastic syndromes	NRAS	PTEN	PMID:38319256
28	T7	—	—	Colorectal cancer	POLD1	ESR1	PMID:38777726
29	T7	—	—	Invasive breast carcinoma	CCND1	IDH1	PMID:18823530
30	T7	—	—	Medulloblastoma	CTNNB1	ERBB2	NCT01878617

## 676 **D Case Study: Generation of Novel Biomarker Hypotheses using Synthetic** 677 **Lethality Concept and Patient Data**

678 We evaluated the TRESR trial [28] (NCT04497116), which is based on a synthetic-lethality (SL)  
679 hypothesis linking ATR inhibition to selected loss-of-function (LoF) alterations. While the study  
680 refers to “ATR inhibitor–sensitizing mutations,” these are not explicitly defined, though subsequent  
681 reports suggest a focus on DNA damage response genes identified via chemogenomic CRISPR  
682 screens [PMID: 37277454].

683 Thus, we tasked BIORESEARCHER with generating SL biomarker hypotheses for ATR inhibition  
684 based on clinical data (input specification in Figure 2). The agent autonomously selected TCGA  
685 and METAPRISM cohorts, identified LoF events associated with ATR expression, and evaluated  
686 candidates through survival analyses. Specifically, the system executed an end-to-end translational  
687 pipeline within the multi-agent orchestration framework (total runtime  $\approx 1352$  s; see Figure 7 for  
688 the full pipeline view). Scenario classification routed the query to the `combination_discovery`  
689 workflow, after which the `translation` subagent canonicalized ATR and the relevant clinical  
690 indications. Subsequently, the `insights_and_signals` (CodeAct) subagent conducted downstream  
691 multi-omic and survival analyses across multiple cohorts, leveraging iterative model invocations  
692 and tool-mediated operations to complete the analysis under the system-generated methodology  
693 summarized in Figure 3.

694 The BIORESEARCHER output combined patient-level data with literature and clinical context, evalu-  
695 ating two signal types: (i) LoF-associated ATR upregulation and (ii) survival patterns in low-ATR  
696 tumors as a proxy for drug-induced ATR inhibition. A condensed reasoning trace integrating these  
697 signals into a candidate ranking is shown in Figure 4.

698 Across both cohorts, increased ATR expression was most robustly associated with TP53 loss, suggest-  
699 ing a pan-cancer marker of ATR pathway activation, though likely reflecting general replication stress  
700 rather than specific therapeutic sensitivity. In contrast, the strongest evidence for a synthetic-lethal  
701 interaction emerged from ATM loss, which - despite lacking compensatory ATR upregulation -  
702 showed the clearest survival signal in low-ATR tumors, consistent with prior preclinical and clinical  
703 findings.

704 Subtype-specific signals further highlighted APC loss in colorectal cancer as the most prominent data-  
705 driven discovery, supported by both elevated ATR expression and favorable low-ATR survival patterns,  
706 although it lacks prior clinical validation. Finally, ARID1A loss demonstrated only modest support in  
707 patient datasets but remains biologically plausible given consistent evidence from external studies,  
708 warranting consideration as a secondary translational candidate. The agent-generated biological  
709 interpretation across these candidates is summarized in Figure 5.

710 Overall, the integrated evidence supports the following ranked hypotheses: TP53 LoF as the strongest  
711 pan-cancer marker of ATR pathway activation; ATM LoF as the most clinically and mechanistically  
712 supported synthetic-lethal biomarker; APC LoF in colorectal cancer as an unexpected data-driven  
713 candidate; and ARID1A LoF as a literature-supported but modestly recovered clinical candidate  
714 (clinically actionable prioritization in Figure 6). The framework’s key strength lies in combining  
715 autonomous multi-cohort analysis with calibrated restraint: where survival models were sparse or  
716 unstable, the agent avoided over-interpretation, preserving the reliability of the resulting translational  
717 dossier.

718 The remainder of this appendix shows the artifacts produced at each stage of the pipeline.

```

# Identification of Biomarkers for ATR Inhibition
#----- BIORESEARCHER - master -----
## Background
I want to explore which genetic changes in cancer might make tumors sensitive to ATR inhibitors, lethality:
* ATR helps cancer cells survive DNA damage
* Loss of certain genes may make tumors dependent on ATR
* Tumors may avoid having both:
  + a gene loss
  + low ATR activity (proxy for drug effect)
Possible signals in data:
* Gene loss - higher ATR expression (compensation)
* Avoidance of gene loss + low ATR
* Survival differences depending on gene loss and ATR levels
## Goal
Using available genomic and clinical data from advanced solid tumors, identify genes that:
* May form synthetic lethal relationships with ATR
* Could act as biomarkers for ATR inhibitor response
You are free to choose datasets, methods, and approaches. Please justify your choices and describe methodology.
## Task 1: Find Candidate Biomarkers
* Identify genes where loss-of-function is linked to higher ATR expression in patients.
* Look for avoidance of gene loss with low ATR expression
* Perform:
  + Pan-cancer analysis
  + Cancer-type specific analyses
  + Rank genes and propose testable hypotheses (e.g., gene X loss increases ATR expression)
## Task 2: Check Overall Survival Relevance
* Test whether candidate gene loss is linked to patient survival, especially depending on ATR expression levels
* Look for patterns consistent with ATR dependency or synthetic lethality
* Perform both:
  + Pan-cancer
  + Cancer-type specific analyses
## Task 3: Check Survival Relevance at Fixed Timepoint
* In patients with low ATR expression, check whether gene loss affects survival at 12, 24 and 36 months
* Compare across definitions of "low ATR"
* Perform both:
  + Pan-cancer
  + Cancer-type specific analyses
## Validation
* Cross-check key findings with:
  + Scientific literature
  + Public databases
  + Clinical trials
* Highlight biolical plausibility and clinical relevance
## Aim
Develop clear, testable hypotheses about:
* ATR dependency
* Synthetic lethal partners
* Biomarkers for ATR inhibitor response
Focus on interpretable and biologically meaningful insights.

```

719

Figure 2: **Research goals for automated ATR inhibitor biomarker discovery.** Input specification to the BIORESEARCHER system defining the biological context (ATR synthetic lethality), expected signals (e.g., gene loss linked to high ATR expression, depletion in low-ATR tumors, survival patterns), and analysis workflow. The pipeline comprises three tasks: (1) candidate biomarker identification via pan-cancer and subtype-specific loss-of-function and ATR expression analyses; (2) survival association conditioned on ATR levels; and (3) fixed-timepoint survival analysis in low-ATR subgroups. A final validation step cross-references results with literature, databases, and clinical trials to ensure biological and clinical relevance. The specification acts as a programmatic interface linking user intent to agent execution, enabling reproducible and interpretable biomarker discovery while constraining the search space.

<p><b>Methodology</b></p> <p><b>Cohorts</b></p> <ul style="list-style-type: none"> <li>- Advanced TCGA: 2,534 advanced-cancer patients; 2,486 with expression; 2,094 with LoF matrix</li> <li>- METAPRISM: 1,831 patients; 932 with expression; 321 with LoF matrix</li> </ul> <p><b>Discovery tests</b></p> <ul style="list-style-type: none"> <li>- Pan-cancer and cancer-type-specific tests of gene LoF vs ATR expression</li> <li>- Tool: run_target_expression_vs_genomewide_lof_tool</li> <li>- Test: Mann-Whitney U</li> <li>- Alternative: greater, meaning LoF group has higher ATR expression</li> <li>- Minimum LoF carriers: <ul style="list-style-type: none"> <li>- 20 for pan-cancer advanced TCGA</li> <li>- 15 for pan-cancer METAPRISM</li> <li>- 5 for subtype analyses</li> </ul> </li> </ul> <p><b>Significance rule</b></p> <ul style="list-style-type: none"> <li>- Toolkit significance uses raw <math>p &lt; 0.01</math> or FDR-adjusted <math>p &lt; \text{THRESHOLD}["\text{fdr\_significance}"]</math></li> <li>- In practice, interpret FDR as the stronger evidence standard</li> </ul>	<p><b>Low-ATR definitions</b></p> <ul style="list-style-type: none"> <li>- Bottom 10%, 25%, and 33% ATR expression</li> <li>- Biologic meaning: <ul style="list-style-type: none"> <li>- q10 = stringent low ATR</li> <li>- q25 = moderate low ATR</li> <li>- q33 = broader low ATR</li> </ul> </li> </ul> <p><b>Survival analyses</b></p> <ul style="list-style-type: none"> <li>- Overall survival interaction modeling with Cox PH</li> <li>- Fixed-timepoint analyses at 12, 24, 36 months in low-ATR subsets</li> <li>- For fixed-timepoint analyses, Fisher exact tests compared event odds in LoF vs WT among low-ATR tumors</li> </ul> <p><b>Canonical ATR / replication-stress genes used for focused interpretation</b></p> <ul style="list-style-type: none"> <li>- ATR pathway: ATR, ATRIP, TOPBP1, ETAA1, RAD17, HUS1, CLSPN, CHEK1</li> <li>- DDR / HR / FA genes assessed as focused candidates: ATR, ATRIP, ATRIP1, ATRIP2, ATRIP3, ATRIP4, ATRIP5, ATRIP6, ATRIP7, ATRIP8, ATRIP9, ATRIP10, ATRIP11, ATRIP12, ATRIP13, ATRIP14, ATRIP15, ATRIP16, ATRIP17, ATRIP18, ATRIP19, ATRIP20, ATRIP21, ATRIP22, ATRIP23, ATRIP24, ATRIP25, ATRIP26, ATRIP27, ATRIP28, ATRIP29, ATRIP30, ATRIP31, ATRIP32, ATRIP33, ATRIP34, ATRIP35, ATRIP36, ATRIP37, ATRIP38, ATRIP39, ATRIP40, ATRIP41, ATRIP42, ATRIP43, ATRIP44, ATRIP45, ATRIP46, ATRIP47, ATRIP48, ATRIP49, ATRIP50, ATRIP51, ATRIP52, ATRIP53, ATRIP54, ATRIP55, ATRIP56, ATRIP57, ATRIP58, ATRIP59, ATRIP60, ATRIP61, ATRIP62, ATRIP63, ATRIP64, ATRIP65, ATRIP66, ATRIP67, ATRIP68, ATRIP69, ATRIP70, ATRIP71, ATRIP72, ATRIP73, ATRIP74, ATRIP75, ATRIP76, ATRIP77, ATRIP78, ATRIP79, ATRIP80, ATRIP81, ATRIP82, ATRIP83, ATRIP84, ATRIP85, ATRIP86, ATRIP87, ATRIP88, ATRIP89, ATRIP90, ATRIP91, ATRIP92, ATRIP93, ATRIP94, ATRIP95, ATRIP96, ATRIP97, ATRIP98, ATRIP99, ATRIP100</li> <li>- Additional recurrent clinical hits: TP53, APC, CDKN2A</li> </ul>
--	--

720

**Figure 3: System-generated methodology for automated ATR inhibitor biomarker discovery.** Methodological specification produced by BIORESEARCHER. Cohort composition (TCGA, METAPRISM) and sample sizes for expression and LoF data is shown. Discovery is defined as genome-wide testing of LoF vs. ATR expression using directional Mann–Whitney  $U$  tests, with cohort-specific carrier thresholds and significance assessed via  $p$ -values and FDR (primary criterion). Low-ATR subgroups are defined by expression quantiles (bottom 10%, 25%, 33%) for downstream analyses. Survival analysis combines Cox interaction models with fixed-timepoint analyses (12, 24, 36 months), supplemented by Fisher’s exact tests in sparse settings. Results are contextualized using canonical ATR and DNA damage response genes. The framework exposes all analytical choices, ensuring reproducibility and interpretability.

Task 1 – Candidate biomarkers

A. Pan-cancer genes where LoF is linked to higher ATR expression

Advanced TCGA pan-cancer

- 2,050 evaluable patients
- 281 genes tested
- Only 2 genes met the toolkit's significance rule, but only 1 was FDR-significant

Top results:

- TP53:  $p = 9.6e-5$ , FDR = 0.0269, effect = 0.00798
- APC:  $p = 0.0065$ , FDR = 0.914, effect = 0.00218

Interpretation:

- TP53 is the only robust pan-cancer hit
- APC is only raw-p significant, not FDR-supported pan-cancer

METAPRISM pan-cancer

- 263 evaluable patients
- only 11 genes testable because of limited LoF-gene coverage
- TP53 again significant:
  - $p = 0.00319$
  - FDR = 0.0351
  - effect = 0.0468

721

Figure 4: **Multi-signal biomarker prioritization from agent reasoning trace.** Condensed execution trace highlighting how BIORESEARCHER integrates heterogeneous evidence to prioritize ATR biomarker candidates. TP53 loss emerges as the only reproducible pan-cancer signal associated with increased ATR expression across cohorts. Subtype-specific analysis identifies APC loss as a strong colorectal cancer signal. In contrast, survival-based analyses reveal ATM loss as the most synthetic-lethality-like candidate, showing improved outcomes in low-ATR tumors despite lacking expression association. The system explicitly downweights unstable Cox interaction models and deprioritizes inconsistent signals (e.g., CDKN2A), illustrating calibrated uncertainty handling. This figure demonstrates cross-task synthesis, where expression, stratified survival, and cohort-specific analyses jointly inform hypothesis ranking.

## Biological interpretation

Most likely interpretations by gene:

### TP53

- Best supported by patient data for ATR compensation
- Likely reflects replication stress / checkpoint dependence
- Good biomarker of ATR biology, but not necessarily the most specific predictive marker for ATR inhibitors

### ATM

- Best supported for a low-ATR survival interaction
- Did not show higher ATR expression
- Could still be a **true** ATRi biomarker because dependency need not be transcriptionally compensated

### APC

- Strongest subtype hit in COAD and favorable low-ATR fixed-timepoint signal
- Data-driven but not yet canonical
- Worth testing, especially in colorectal cancer

### ARID1A

- Modest support in metastatic disease and bladder cancer
- Strong external biologic plausibility, but this dataset did not robustly recover it pan-cancer

722

Figure 5: **Agent-generated biological interpretation and hypothesis prioritization.** Post-analysis interpretation of candidate ATR biomarkers, summarizing gene-specific evidence across expression and survival signals. The system distinguishes between robust, data-supported signals (e.g. TP53, ATM) and exploratory or context-specific candidates (e.g. APC, ARID1A), producing biologically grounded, testable hypotheses.

### Clinical relevance

If you want clinically interpretable, testable hypotheses from these data, I would prioritize:

#### Highest-priority hypotheses

1. TP53 LoF marks tumors with compensatory ATR upregulation
  - best for biology / stratification
  - weaker as a direct predictive marker from survival alone
2. ATM LoF may define a subgroup that does poorly with intact ATR but better when ATR is low
  - strongest low-ATR survival interaction candidate
  - high external plausibility
3. APC LoF in colorectal cancer may mark an ATR-relevant subgroup
  - strongest subtype-specific discovery
  - should be validated in DepMap and treatment-response datasets before clinical use
4. ARID1A LoF in bladder / metastatic settings remains a biologically credible secondary candidate
  - but current patient evidence is modest

723

Figure 6: **Translation of hypotheses into clinically actionable prioritization.** Agent-generated ranking of ATR biomarker hypotheses based on clinical interpretability, integrating expression, survival, and external plausibility signals. The system distinguishes between stratification markers (e.g. TP53), synthetic-lethal candidates (e.g. ATM), and exploratory, subtype-specific hypotheses (e.g. APC, ARID1A).

724

```
[DEBUG] [RECONCILE] Discovered 37 source files
[DEBUG] [RECONCILE] Total length of the descriptions: 47107
[DEBUG] [RECONCILE] Total length of the manifest: 163840
[NODE] Node: plan_report
[DEBUG] [RECONCILE] Report plan: 10 sections, covering 37 unique files
[DEBUG] [RECONCILE] Assembled reconciled sections into frankenstein draft.
[DEBUG] [RECONCILE] Final review complete.

----- SUBAGENT OUTPUT: reconcile -----
Final reconciled report has been generated and saved. Report length: 57058 characters.

----- TODO LIST -----
• [completed] Classify request and retrieve scenario guidance
• [completed] Run biomarker/survival analyses via specialized subagents
• [completed] Validate top findings with literature and clinical trials
• [completed] Synthesize and present ranked hypotheses and methodology

----- FINAL OUTPUT -----
The comprehensive report has been generated.

Top ranked ATR biomarker hypotheses from the analysis:

1 TP53 loss → higher ATR expression pan-cancer
2 ATM loss + low ATR → strongest low-ATR survival interaction
3 APC loss in colorectal cancer → subtype-specific ATR-relevant signal
4 ARID1A loss → plausible but weaker support
5 CDKN2A loss → ATR biology signal, but not a clean sensitivity marker

If you want, I can now give you:
• a 1-page executive summary,
• the ranked biomarker table,
• or a methods-focused version with exact analysis choices and caveats.

Results saved to: results/20260427_231244_b8ad7688
Thread ID: 20260427_231244_b8ad7688 (reuse with --thread-id to ask follow-up questions)
```

725

**Figure 7: End-to-end execution and hypothesis synthesis in BIORESEARCHER.** Multi-agent pipeline for ATR biomarker discovery comprising four stages: (1) task decomposition, (2) specialized biomarker and survival analyses, (3) validation with literature and clinical trials, and (4) reconciliation and synthesis. The system integrates 37 intermediate artifacts into a single report (~57k characters), completing all steps autonomously. Final outputs are ranked, testable hypotheses, including TP53 loss (pan-cancer ATR activation), ATM loss (top synthetic-lethal candidate), APC loss (colorectal-specific), and ARID1A loss (weaker signal), while inconsistent findings (e.g., CDKN2A) are deprioritized. The framework demonstrates end-to-end automation, large-scale result synthesis, and calibrated uncertainty handling.

## 726 E Asset Provenance, Licenses, and Terms of Use

727 This appendix consolidates the third-party assets used by BIORESEARCHER and in our evaluation.  
728 For each asset we report the version (where applicable), the original creator or maintainer, a citation,  
729 a URL, the license under which the asset is made available, and a brief note on how we use it. Asset  
730 usage was verified against each license at the time of submission.

731 **DepMap compliance note.** The Broad Institute DepMap Terms of Use permit AI/ML model use  
732 with or on the data *for internal research use, including non-profit sharing of methodologies*, but  
733 restrict commercial use and explicitly restrict use of the data to “train, develop, or enhance machine  
734 learning or AI models other than for internal research use.” BIORESEARCHER’s sandbox loads  
735 DepMap 24Q4 Parquet files for on-the-fly statistical analysis (correlations, Mann–Whitney tests,  
736 percentile ranking, survival analysis). DepMap data are *not* used to fine-tune any model in this work,  
737 are not redistributed in our released artifacts, and are re-fetched by the user from the official DepMap  
738 portal when reproducing our results.

Asset	Version	Owner / Maintainer	Citation & URL	License / Terms	Use in this work
<i>(a) Biomedical databases and ontologies</i>					
ChEMBL	36	EMBL-EBI	<a href="https://www.ebi.ac.uk/chembl/">https://www.ebi.ac.uk/chembl/</a>	CC-BY; EBI Terms of Use	Local SQLite; entity resolution and target/drug lookup.
Open Targets Platform	26.03	Open Targets / EMBL-EBI	<a href="https://platform.opentargets.org/">https://platform.opentargets.org/</a>	Terms of use for the Open Targets Platform - updated June 2021	Local SQLite + GraphQL API; target-disease evidence.
PubChem	2026-04-29	NCBI / NLM	<a href="https://pubchem.ncbi.nlm.nih.gov/">https://pubchem.ncbi.nlm.nih.gov/</a>	NCBI/NLM public domain; NCBI Website and Data Usage Policies	Compound synonym resolution from SMILES.
Ensembl	Release 115	EMBL-EBI	<a href="https://www.ensembl.org/">https://www.ensembl.org/</a>	Ensembl Privacy Notice v 3.0.0	Gene-symbol to Ensembl-ID mapping.
EFO		EMBL-EBI	<a href="https://www.ebi.ac.uk/efo/">https://www.ebi.ac.uk/efo/</a>	Licensing of EMBL-EBI data resources	Disease-name to ontology-ID resolution.
MONDO	v2026-04-07	Monarch Initiative	<a href="https://mondo.monarchinitiative.org/">https://mondo.monarchinitiative.org/</a>	CC BY 4.0	Cross-mapped disease IDs.
DOID		Disease Ontology consortium	<a href="https://disease-ontology.org/">https://disease-ontology.org/</a>		Cross-mapped disease IDs.
Human Protein Atlas	25.0	SciLifeLab / KTH	<a href="https://www.proteinatlas.org/">https://www.proteinatlas.org/</a>	CC BY-SA 4.0	Basal protein characteristics and for Uniprot ID mapping.
STRING	12.0	SIB / EMBL	<a href="https://string-db.org/">https://string-db.org/</a>	CC BY 4.0	Protein-protein interaction graph.
Reactome	v96	Reactome consortium	<a href="https://reactome.org/">https://reactome.org/</a>	CC BY 4.0	Pathway annotations.
GEO (via ARCHS4 pipeline)	ARCHS4 pipeline Apache 2.0	NCBI / NLM	<a href="https://www.ncbi.nlm.nih.gov/geo/">https://www.ncbi.nlm.nih.gov/geo/</a>	GEO data: public domain; ARCHS4 pipeline: Apache 2.0	GEO expression outlier detection
Gene Ontology	2026-03-25	GO Consortium	<a href="http://geneontology.org/">http://geneontology.org/</a>	CC BY 4.0	Functional annotations.
DrugComb	2	Changzhou University and CSU	<a href="http://drugcombdb.denglab.org/">http://drugcombdb.denglab.org/</a>		Drug-pair synergy prediction endpoint.
ClinicalTrials.gov / AACT		NLM / CTTI	<a href="https://aact.ctti-clinicaltrials.org/">https://aact.ctti-clinicaltrials.org/</a>	AACT user agreement	Trial registration data.
PubMed	live API	NCBI / NLM	<a href="https://www.ncbi.nlm.nih.gov/">https://www.ncbi.nlm.nih.gov/</a>	NCBI/NLM Disclaimer	Literature retrieval.
Google Patents (SerpAPI)	live API	SerpAPI (third-party)	<a href="https://serpapi.com/google-patents-api">https://serpapi.com/google-patents-api</a>	SerpAPI Terms of Service (commercial API)	Patent search.

Table 13: Third-party assets used in BIORESEARCHER and in evaluation.

Table 13 (continued).

Asset	Version	Owner / Main-tainer	Citation & URL	License / Terms	Use in this work
<i>(b) Functional-genomics and patient cohorts</i>					
DepMap (CRISPR, shRNA, omics, proteomics, mutations)	25Q3	Broad Institute	<a href="https://depmap.org/portal/">https://depmap.org/portal/</a>	DepMap Terms of Use	Sandboxed Python analyses (dependency, expression, mutation stratification).
TCGA	45.0	NCI / NHGRI	<a href="https://portal.gdc.cancer.gov/">https://portal.gdc.cancer.gov/</a>	NIH Genomic Data Sharing Policy	Survival analyses; ATR case study. We do not redistribute controlled-access data.
MSK-CHORD	2024	MSK Cancer Data Science Initiative Group	<a href="https://www.cbioportal.org/study/summary?id=msk_chord_2024">https://www.cbioportal.org/study/summary?id=msk_chord_2024</a>	BY-NC-ND 4.0	Survival analyses; ATR case study.
METAPRISM	30/10/2022	European Bioinformatics Institute (EMBL-EBI)	<a href="https://www.ega-archive.org/">https://www.ega-archive.org/</a>		Survival analyses; ATR case study.

Table 13 (continued).

Asset	Version	Owner / Main-tainer	Citation & URL	License / Terms	Use in this work
<i>(d) Foundation models, agents, and APIs</i>					
GPT-5.4-mini, GPT-5.4, GPT-5.5	GPT-API: 2.14.0	OpenAI	<a href="https://platform.openai.com/">https://platform.openai.com/</a>	OpenAI Business Terms; usage policies	Core models for BIORESEARCHER and baselines.
GPT-4.1	GPT-API: 2.14.0	OpenAI	<a href="https://platform.openai.com/">https://platform.openai.com/</a>	OpenAI Business Terms	LLM-as-a-judge for single-step tests.
OpenAI Codex (CLI)	0.128.0	OpenAI	<a href="https://openai.com/">https://openai.com/</a>	OpenAI Business Terms	Baseline.
OpenAI Deep Research	API: 2.14.0	OpenAI	<a href="https://openai.com/">https://openai.com/</a>	OpenAI Business Terms	Baseline.
Gemini 3.1 Pro	API: 1.73.0	Google	<a href="https://ai.google.dev/">https://ai.google.dev/</a>	Google APIs Terms of Service	Baseline.
Claude Code	2.1.128	Anthropic	<a href="https://www.anthropic.com/">https://www.anthropic.com/</a>	Anthropic Commercial Terms of Service	Baseline.
CellType Agent	28-03-2026	Jiawen Chen, Jianghao Zhang, Huaxiu Yao, Yun Li	<a href="https://github.com/jianghao-zhang/CellTypeAgent/">https://github.com/jianghao-zhang/CellTypeAgent/</a>		Baseline.
Medea	26-03-2026	[23]	<a href="https://github.com/mims-harvard/MEDEA">https://github.com/mims-harvard/MEDEA</a>	Apache-2.0	Baseline.
all-MiniLM-L6-v2	2.0	SentenceTransformers	<a href="https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2">https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2</a>	Apache 2.0	Embedding model for the conference RAG index.

Table 13 (continued).

Asset	Version	Owner / Main-tainer	Citation & URL	License / Terms	Use in this work
<i>(e) Software libraries and frameworks</i>					
DeepEval	3.9.7	Confident AI	[1]; <a href="https://github.com/confident-ai/deepeval">https://github.com/confident-ai/deepeval</a>	Apache 2.0	LLM-as-a-judge framework.
ChromaDB	1.5.7	Chroma	<a href="https://github.com/chroma-core/chroma">https://github.com/chroma-core/chroma</a>	Apache 2.0	Vector store for the conference RAG index.
RDKit	2026.3.1	RDKit consortium	<a href="https://www.rdkit.org/">https://www.rdkit.org/</a>	BSD-3-Clause	SMILES standardisation.
pandas, numpy, scipy	2.3.3 / 2.1.3 / 1.17.1	/ NumFOCUS /	<a href="https://numpy.org/">https://numpy.org/</a> , etc.	BSD-3-Clause	Sandbox numerics.
LangGraph LangChain	/ 1.1.6 / 1.2.15	/ LangChain Inc.	<a href="https://github.com/langchain-ai/langgraph">https://github.com/langchain-ai/langgraph</a>	MIT	Orchestration framework for the agent graphs.

Table 13 (continued).

Asset	Version	Owner / Main-tainer	Citation & URL	License / Terms	Use in this work
<i>(f) Benchmarks</i>					
BixBench-Verified-50	50-question subset	FutureHouse / Phylobio	[2]; <a href="https://huggingface.co/datasets/phylobio/BixBench-Verified-50">https://huggingface.co/datasets/phylobio/BixBench-Verified-50</a>	Apache 2.0	Quantitative-reasoning evaluation.
BaisBench (Scientific Discovery, BAIS-SD)	193-question track	Luo et al.	[3]; <a href="https://github.com/EperLuo/BaisBench">https://github.com/EperLuo/BaisBench</a>	MIT	Open-ended biological-discovery evaluation.