HRD-BIOAGENT: EVIDENCE-GROUNDED AGENTIC REASONING FOR MULTIMODAL HOMOLOGOUS RECOMBINATION DEFICIENCY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Homologous recombination deficiency (HRD) is commonly characterized by genomic scars, but its phenotype is also reflected across mutation, transcriptomic, epigenetic, and histomorphologic modalities. Conventional multimodal predictors can estimate HRD status, yet they provide limited support for reasoning about whether biological evidence is consistent with a patient-level prediction, particularly when definitive genomic measurements are unavailable. We introduce **HRD-BioAgent**, an evidence-grounded agentic framework that combines multimodal HRD prediction with adaptive biological evidence acquisition. HRD-BioAgent organizes modality-specific signatures in a provenance-aware **HRD Evidence Map** and uses a **ReACT** workflow to determine when additional genomic evidence is needed. For a patient requiring additional evidence, the agent can invoke a genomic analysis tool to derive SBS3 mutation-signature evidence from raw trinucleotide mutation profiles, while numerical predictions are produced by a calibrated fold-local meta-model rather than an LLM. We separate the value of the evidence that is integrated from the policy that decides when to acquire it. Under an *acquire-all* reference policy, in which every patient receives the complete evidence set, the evidence-integration model reaches AUROC 0.923 on TCGA-BRCA (N=958); on a matched complete-case subset (N=616), adding SBS3 to a non-agent OOF stacking baseline raises AUROC from 0.878 to 0.906 (two-sided paired-bootstrap p=0.0008), and integrating the complete evidence set reaches 0.916, a further numerical increment that is not statistically distinguishable from chance under two-sided testing (ΔAUROC = 0.009, p=0.068)—the paper's contribution therefore rests on SBS3 retrieval, provenance, and auditable concordance rather than on this increment. A *confidence-gated adaptive* policy that requests additional evidence only for patients whose prediction remains uncertain retains AUROC 0.907 against an acquire-all reference of 0.919 on the same cohort (N=898) while scheduling 58% fewer evidence-analysis steps. HRD-BioAgent also provides patient-level reasoning traces, provenance, and prediction–evidence concordance, enabling auditable HRD prediction under incomplete multimodal measurements.

1 INTRODUCTION

Homologous recombination deficiency (HRD) is a biologically heterogeneous state of defective DNA double-strand-break repair (Lord & Ashworth, 2016). A widely used genomic representation summarizes loss of heterozygosity (LOH), large-scale state transitions (LST), and telomeric allelic imbalance (TAI) into an HRD scar score (Marquard et al., 2015; Telli et al., 2016). However, the consequences and correlates of HRD extend beyond the genomic scar. In breast cancer, HRD-positive tumors exhibit mutation, transcriptional, epigenetic, and histologic phenotypes that can be measured on different platforms.

This creates a natural multimodal learning problem. Existing models can learn mappings such as

$$X^{\mathrm{WSI}} \to Y^{\mathrm{HRD}} \quad \text{or} \quad (X^{\mathrm{RNA}}, X^{\mathrm{Mut}}, X^{\mathrm{WSI}}, \ldots) \to Y^{\mathrm{HRD}}, \tag{1}$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

but predictive fusion alone does not answer a second question: *which cross-modal biological observations support, contradict, or fail to inform the HRD prediction for this patient?* This distinction becomes especially important when the defining genomic scar or another molecular assay is unavailable.

We therefore study HRD as both a prediction target and an evidence-reasoning problem. Our central hypothesis is that a model should not simply concatenate all available modalities. Instead, it should identify modality-specific HRD-associated signatures, retain their provenance and uncertainty, and integrate only evidence that is valid for the patient and cohort under consideration. We call this framework **HRD-BioAgent**.

HRD-BioAgent separates three quantities that are often conflated:

$$\boxed{\text{Measured evidence}} \neq \boxed{\text{Model prediction}} \neq \boxed{\text{Secondary biological inference}}. \tag{2}$$

The genomic scar remains the anchor when measured. RNA, mutation, methylation, and WSI signals can support an HRD hypothesis but are not relabeled as genomic scars. The agent organizes these relationships in an HRD Evidence Map and uses a deterministic critic to check fold identifiers, prediction provenance, cancer context, evidence validity, and unverified missing-modality claims before they enter the final output; fold-local training procedures, not the critic itself, are what control data leakage during model development.

Our contributions are: (1) an HRD-centered formulation combining multimodal prediction with explicit biological evidence assessment; (2) a provenance-aware HRD Evidence Map that distinguishes measurement, prediction, and secondary inference; (3) an evidence-acquisition workflow coupled to a fold-local calibrated estimator rather than LLM-generated numerical probabilities; and (4) patient-level traces and evaluations of prediction, evidence ablation, missing modalities, concordance, and deterministic safeguards. Figure 1 provides an overview of the framework.

2 Related Work

HRD prediction. Genomic-scar methods measure LOH/LST/TAI, whereas HRDetect and CHORD draw on whole-genome mutational and structural-variant signatures (Marquard et al., 2015; Davies et al., 2017; Nguyen et al., 2020; Sztupinszki et al., 2018). Histology-based prediction offers an alternative view of molecular phenotypes (Kather et al., 2020; Echle et al., 2021; Coudray et al., 2018), but scalar predictions do not specify whether patient-level molecular evidence supports them.

Multimodal learning and agents. Pathology–omics fusion models combine visual and molecular features (Chen et al., 2022b; 2021; 2022a; Jaume et al., 2024); incomplete assays motivate modality-aware learning (Cheerla & Gevaert, 2019). Pathology agents orchestrate tools and interpret evidence (Weishaupt et al., 2026; Jin et al., 2024; Lu et al., 2024b). HRD-BioAgent instead conditions biological evidence acquisition on an HRD-specific Evidence Map, separates measured evidence from predictions and inference, and reserves numerical probabilities for a calibrated fold-local estimator. Extended comparisons appear in Appendix M.

3 Method

3.1 Problem Formulation

For patient i, let

$$X_i = \{X_i^{\mathrm{WSI}}, X_i^{\mathrm{RNA}}, X_i^{\mathrm{Mut}}, X_i^{\mathrm{Meth}}, X_i^{\mathrm{Gen}}\}, \tag{3}$$

with an availability mask M_i. The primary target $Y_i \in \{0, 1\}$ is genomic HRD status. When the required copy-number components are available, the reference HRD score is

$$\mathrm{HRDsum} = \mathrm{LOH} + \mathrm{LST} + \mathrm{TAI}, \tag{4}$$

with HRD+ defined as $\mathrm{HRDsum} \geq 42$ and HRD− otherwise. The predictive objective is $P(Y_i = 1 \mid X_i, M_i)$, but the agent additionally returns an evidence set E_i, concordance state C_i, uncertainty U_i, and provenance P_i:

$$O_i = (\hat{p}_i^{\mathrm{HRD}}, E_i, C_i, U_i, P_i). \tag{5}$$

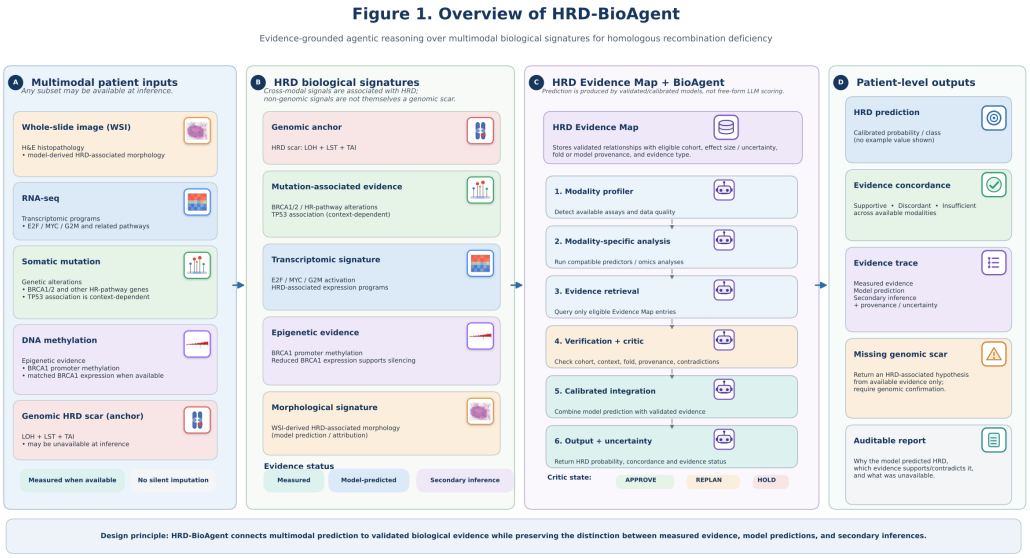

Figure 1: **Overview of HRD-BioAgent.** The framework connects multimodal patient inputs (left) to patient-level outputs (right) through three intermediate stages: (1) modality-specific HRD biological signatures with explicit evidence status (measured, model-predicted, or secondary inference); (2) an HRD Evidence Map that organizes validated relationships and feeds an implementation workflow with eight auditable stages, represented by six higher-level blocks in the schematic—modality profiling, initial prediction, genomic/mutation evidence retrieval, transcriptomic evidence retrieval, epigenetic evidence retrieval, morphology evidence retrieval, evidence integration and concordance analysis, and deterministic critique (APPROVE/REPLAN/HOLD); calibrated integration via meta-model g and uncertainty-aware output are shown as the final blocks; (3) patient-level outputs including HRD prediction, evidence concordance, provenance trace, missing-modality handling, and an auditable report. The design principle is that multimodal predictions are connected to calibrated biological evidence while preserving the distinction between measured evidence, model predictions, and secondary inferences.

3.2 BIOLOGICAL EVIDENCE CONSTRUCTION

Biological signature estimation. We use paired molecular data to define candidate HRD-associated signatures. The genomic scar is the anchor rather than one interchangeable modality. For each non-anchor modality m, we estimate a signature $S_{\mathrm{HRD}}^{(m)}$ using training data only and record its cohort, estimator, thresholds, held-out validation, and provenance.

Our existing TCGA-BRCA analysis provides biological priors for this map. HRD-positive versus HRD-negative tumors yielded 2,914 differentially expressed genes, with E2F targets, MYC targets, and the G2M checkpoint among the strongest activated programs. TP53 was strongly enriched overall in HRD-positive breast tumors, although subtype-stratified analyses show that this association is context dependent. An HR-gene-wild-type subgroup with BRCA1 promoter hypermethylation showed lower BRCA1 expression and an HRD-associated transcriptional phenotype. These observations are associations and are not encoded as causal rules.

Provenance-aware Evidence Map. The HRD Evidence Map stores validated relationships as

$$e_j = (S_j \to Y^{\mathrm{HRD}}; \mathcal{C}_j, f_j, \theta_j, V_j, P_j), \tag{6}$$

where $\mathcal{C}_j$ is the eligible population, f_j the estimator or statistical relation, θ_j its prespecified threshold/calibration, V_j held-out validation, and P_j provenance. Evidence is tagged as MEASURED, PREDICTED, or SECONDARY. An edge can be used only when its cancer context, patient eligibility, modality availability, and provenance requirements are satisfied.

3.3 MULTIMODAL HRD PREDICTORS

Each available modality has a predictor

$$\hat{p}_i^{(m)} = f_m(X_i^{(m)}), \tag{7}$$

and a conventional multimodal baseline produces

$$\hat{p}_i^{\text{multi}} = f_{\text{multi}}(X_i^{\text{available}}, M_i). \tag{8}$$

The WSI branch uses CONCH (Lu et al., 2024a) tile embeddings aggregated by ABMIL (Ilse et al., 2018) with patient-level held-out representations and an HRD classifier; RNA, mutation, and methylation branches use fold-local molecular features. All preprocessing, feature selection, calibration, and model selection must occur inside the training fold.

3.4 HRD-BIOAGENT AND REACT

Workflow and critic. The workflow contains eight auditable stages: (1) modality profiling, (2) initial HRD prediction, (3) genomic/mutation evidence retrieval, (4) transcriptomic evidence retrieval, (5) epigenetic evidence retrieval, (6) morphology evidence retrieval, (7) evidence integration and concordance analysis, and (8) deterministic critique. The critic returns APPROVE, REPLAN, or HOLD. HOLD suppresses patient-level interpretation when evidence or provenance requirements fail.

Patient-level evidence acquisition. The baseline fusion models in this study use a predefined feature matrix in which SBS3 is not included. A static model can use SBS3 when precomputed; the agentic question is whether and how the additional analysis is requested for a particular evidence state. We adopt a ReACT (Reasoning + Acting) workflow (Yao et al., 2023) that enables the agent to reason about each patient's biological context and retrieve evidence adaptively:

1. **Think**: assess the patient's initial multi-modal prediction and available modalities.
2. **Act**: retrieve base biological evidence from the Evidence Map (pathway scores, mutation status, methylation, BRCA1 expression).
3. **Observe**: evaluate evidence completeness—is the HRD mechanism explained by known causal variants (BRCA1/2 mutation, BRCA1 methylation)?
4. **Think**: determine whether additional computed evidence is needed. For patients without a known HRD mechanism, the SBS3 mutation signature may reveal HRD through an alternative pathway.
5. **Act**: retrieve SBS3 cosine similarity, computed from the 96-channel trinucleotide mutation profile (MAF files) against the COSMIC SBS3 reference (Alexandrov et al., 2020). This step requires reading raw variant calls and computing a derived biological feature—available as a tool result when requested. For reproducibility, the reported full-evidence result and the deployment-time selective-routing variants are distinguished in Appendix J.
6. **Act**: run the calibrated meta-model g with the assembled evidence to produce the final prediction and reasoning trace.

Each patient receives a structured trace of the evidence considered, tool results, and prediction inputs; its description must not imply that the LLM itself computes the numerical estimate. The SBS3 mutation signature (cosine similarity to COSMIC SBS3; HRD+ mean = 0.465 vs. HRD− mean = 0.358) captures the flat trinucleotide spectrum characteristic of homologous recombination deficiency and has the largest coefficient in the reported evidence model (β=1.39; coefficients are scale-dependent).

Acquisition policy: we distinguish *which* evidence g integrates from *when* it is acquired. Depth $a_i \in \{\text{base}, \text{+mut/meth}, \text{+SBS3}\}$ has cost $c(a_i)$ in scheduled steps (Appendix J); an *acquire-all* reference sets a_i=+SBS3 for all i, while a *confidence-gated adaptive* policy escalates only while $\hat{p}_i$ is uncertain, using the same g and critic throughout. Table 2–3 report the acquire-all reference.

3.5 PREDICTION REFINEMENT AND AUDITABLE OUTPUT

Calibrated evidence integration. To ensure that the agent is more than an explanation layer, we explicitly test whether verified evidence improves prediction. Let $\mathbf{e}_i$ contain fold-local evidence

scores and availability indicators. We define

$$\hat{p}_i^{\text{agent}} = g(\hat{p}_i^{\text{multi}}, \mathbf{e}_i, M_i), \tag{9}$$

where g is a calibrated meta-model (Zadrozny & Elkan, 2002) fit exclusively on training-fold predictions and evidence. The language model is not allowed to assign the numerical HRD probability. Its role is orchestration and evidence synthesis; numerical refinement is performed by the reproducible estimator g. For the ReACT variant, $\mathbf{e}_i$ is extended to include SBS3 cosine similarity and indel fraction (11 features total), enabling the meta-model to integrate mutation signature evidence alongside pathway and epigenetic features. This design allows a direct ablation between conventional fusion and evidence-guided integration.

Prediction–evidence concordance. We independently threshold the model prediction and prespecified evidence summary to assign four analysis groups: concordant positive, prediction-positive/evidence-negative, prediction-negative/evidence-positive, and concordant negative. Discordant groups are not automatically relabeled. Instead, they are analyzed for subtype, mutation, methylation, pathway, and morphology differences to identify model failure modes or biologically unusual subgroups.

Missing-modality states. We evaluate progressive masks over the available modalities. The most important setting withholds the defining genomic-scar measurement while retaining some combination of WSI, RNA, mutation, and methylation. In this setting, the output is explicitly an *HRD-associated hypothesis*; it is never presented as a measured genomic scar. Performance is compared against the hidden reference label only for evaluation.

4 Experiments

4.1 Experimental Setup

The primary development cohort is TCGA-BRCA (Cancer Genome Atlas Network, 2012). Patient-level folds are used throughout to prevent slide-level leakage. Exploratory analyses in OV, UCEC (Cancer Genome Atlas Research Network, 2013), UCS, PAAD, and PRAD are used to study conservation of molecular signatures rather than to claim external validation of the BRCA-trained predictor.

All primary predictive comparisons use patient-level held-out predictions. We distinguish the all-available cohort (N=958) from the matched complete-case cohort (N=616); AUROC differences should only be interpreted on matched patients.

4.2 Biological Grounding

We reproduce the molecular analysis that motivates the Evidence Map: differential expression (Love et al., 2014) and Hallmark enrichment (Liberzon et al., 2015; Korotkevich et al., 2021), mutation enrichment, BRCA1 promoter methylation and expression, and exploratory cross-cancer pathway conservation. Importantly, context-dependent signals such as TP53 are retained with their subgroup restrictions rather than encoded as universal HRD rules.

In TCGA-BRCA, 2,914 genes were differentially expressed between HRD-positive and HRD-negative tumors (Figure 2A). E2F targets (NES = 3.04), MYC targets (2.77), and G2M checkpoint (2.76) were strongly activated, whereas early estrogen response was depleted (NES $= -2.36$). TP53 mutation was enriched overall in HRD-positive tumors (OR $= 8.07$, $p = 1.1 \times 10^{-31}$), but the association collapsed within triple-negative disease, demonstrating why evidence must carry context rather than be treated as a universal rule. The epigenetic analysis identified a small HR-gene-wild-type, BRCA1-promoter-hypermethylated HRD-positive subgroup (11/89 HR-WT HRD+ vs. 1/494 HRD$-$; 450K, $\beta \geq 0.2$). Exploratory extension across six TCGA cohorts (Figure 2B) showed that the E2F/G2M/MYC proliferation core was the most consistently enriched feature, whereas TP53 and estrogen-response associations were tissue dependent. Table 1 summarizes the key molecular findings and Figure 2 shows the differential expression landscape and per-cancer Hallmark enrichment.

Table 1: **HRD biological signature summary** (TCGA-BRCA, HRD+ vs. HRD−). NES values are BRCA-specific; pan-cancer tissue-invariant mean NES across 6 cancers is 2.10/1.85/2.04 for E2F/MYC/G2M respectively.

Analysis	Main finding	Effect size	Significance	Scope
DESeq2 DEG	2,914 ($\lvert \log_2 \text{FC} \rvert \geq 1.5$)	1,370 up / 1,544 down	FDR < 0.05	958 patients
Hallmark GSEA — E2F	Enriched in HRD+	NES = 3.04	FDR $< 10^{-45}$	50 pathways
Hallmark GSEA — MYC_V1	Enriched in HRD+	NES = 2.77	FDR $< 10^{-30}$	50 pathways
Hallmark GSEA — G2M	Enriched in HRD+	NES = 2.76	FDR $< 10^{-30}$	50 pathways
Mutation co-occurrence	TP53 enriched in HRD+	OR = 8.07	$p = 1.1 \times 10^{-31}$	Fisher's exact
BRCA1 promoter meth.	11/89 HR-WT HRD+ meth.†	12.4%	450K, $\beta \geq 0.2$	cg04658354

†Complete-case methylation-analysis cohort. The 10/102 figure reported in Appendix F is from the matched cross-modal subset with stricter eligibility.

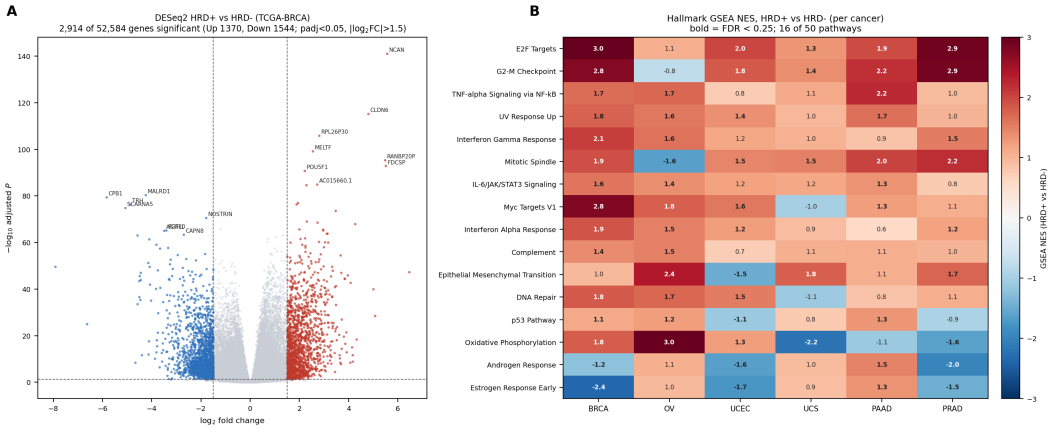

Figure 2: **HRD biological landscape** (cohort-level analysis). **A.** DESeq2 volcano for HRD+ vs. HRD− (TCGA-BRCA): 2,914 of 52,584 genes significant at adjusted $P < 0.05$ and $\lvert \log_2 \text{FC} \rvert \geq 1.5$ (Up 1,370; Down 1,544). **B.** Hallmark GSEA NES for the same contrast across six TCGA cohorts (16 of 50 pathways, bold: FDR < 0.25). The proliferation core (E2F, G2M, MYC) is enriched across all tissues.

4.3 Predictive Performance and Evidence Integration

Prediction baselines. We compare WSI-only, RNA-only, mutation-only, methylation-aware, pairwise multimodal, and all-available multimodal predictors. Metrics are AUROC, AUPRC, weighted F1, Brier score, and expected calibration error. Where class prevalence differs, AUPRC and calibration are interpreted alongside AUROC.

Evidence-guided integration. Because the baselines exclude HRD-causal mutations (BRCA1/2, PALB2, RAD51C/D) from prediction features, we evaluate three agent variants to separate the contributions of agentic organization, privileged feature access, and evidence reasoning:

$$\begin{aligned} \text{BioAgent}_{\text{fair}} &: g(\hat{p}^{\text{multi}}, \mathbf{e}_{\text{fair}}, M), \\ \text{BioAgent}_{\text{biology}} &: g(\hat{p}^{\text{multi}}, \mathbf{e}_{\text{all}}, M), \\ \text{BioAgent}_{\text{AcqAll}} &: g(\hat{p}^{\text{multi}}, \mathbf{e}_{\text{all+SBS3}}, M), \end{aligned} \tag{10}$$

where $\mathbf{e}_{\text{fair}}$ contains only features also accessible to baselines (4 pathway Z-scores, BRCA1 expression, TP53 mutation; 6 features), $\mathbf{e}_{\text{all}}$ additionally includes BRCA1/2 mutation status and BRCA1 promoter methylation (9 features), and $\mathbf{e}_{\text{all+SBS3}}$ further adds SBS3 cosine similarity and indel fraction retrieved through the ReACT workflow (11 features). This progression tests changes in the evidence set and its integration, not a pure agent-reasoning effect. The static SBS3 control is informative but does not contain the exact same eleven evidence features as the full system. To further disentangle the SBS3 feature contribution from the agent's evidence integration, we compare against a non-agent OOF stacking baseline augmented with the same static SBS3 feature (Table 3). Paired bootstrap confidence intervals and paired tests are computed on identical held-out patients.

Table 2: **HRD prediction comparison** (TCGA-BRCA, 5-fold patient-level CV; $\text{BioAgent}_{\text{fair/biology/AcqAll}}$ as defined above). HRD-causal mutations (BRCA1/2, PALB2, RAD51C/D) and TMB excluded from baseline mutation features. Panel A: all available patients per modality. Panel B: complete-case subset (N=616) with bootstrap 95% CI. $\text{BioAgent}_{\text{AcqAll}}$ is the acquire-all reference (Section 3.4); the confidence-gated adaptive policy is reported in Appendix J and Table 3.

Method	Modalities	N	AUROC	AUPRC	Brier↓	ECE↓
A. All available patients						
WSI-only (ABMIL (Ilse et al., 2018))	WSI	958	.833	.518	.169	.212
Mutation-only (MLP)	Mut	898	.816	.487	.122	.033
Methylation-only (MLP)	Meth	616	.809	.556	.112	.053
RNA-only (MLP)	RNA	958	.892	.654	.097	.030
Early fusion (concat)	RNA+Mut+Meth	616	.865	.591	.101	.059
Late fusion (stacking)	RNA+Mut+Meth	616	.884	.635	.095	.034
RNA+Mut (residual)	RNA+Mut	958	.885	.609	.130	.139
$\text{BioAgent}_{\text{fair}}$	Pred.+evidence†	958	.899	.707	.092	.034
$\text{BioAgent}_{\text{biology}}$	Pred.+evidence	958	.900	.696	.090	.033
$\text{BioAgent}_{\text{AcqAll}}$	**Pred.+evidence‡**	**958**	**.923**	**.749**	**.082**	**.023**
B. Complete-case (N=616, fair comparison)						
RNA-only (MLP)	RNA	616	.876 (.839–.910)	.582	.097	.032
Late fusion (stacking)	RNA+Mut+Meth	616	.884 (.849–.915)	.635	.095	.034
OOF stacking	All OOF preds	616	.878 (.840–.910)	.623	.096	.036
OOF stacking + SBS3	All OOF preds + SBS3	616	.906 (.876–.932)	.678	.087	.038
$\text{BioAgent}_{\text{fair}}$	Pred.+evidence†	616	.881	.628	.095	.045
$\text{BioAgent}_{\text{biology}}$	Pred.+evidence	616	.884	.654	.090	.025
$\text{BioAgent}_{\text{AcqAll}}$	**Pred.+evidence‡**	**616**	**.916** (.887–.941)	**.703**	**.083**	.039

†Evidence features restricted to match baseline feature availability: 4 pathway Z-scores + BRCA1 expression + TP53 mutation (no BRCA1/2 mutation, no BRCA1 methylation).

‡Acquire-all reference policy (Section 3.4): SBS3 and indel features are retrieved for every patient. The confidence-gated adaptive policy retains AUROC 0.907 against an acquire-all reference of 0.919 (N=898) while scheduling 58% fewer evidence-analysis steps (Appendix J).

Table 3: **Evidence decomposition ablation** (complete-case N=616, 5-fold patient-level CV). The progression quantifies SBS3 augmentation and the additional full-pipeline increment, but does not independently identify a pure agent effect; p-values are two-sided.

Configuration	AUROC (95% CI)	AUPRC	Brier↓	ΔAUC	p
OOF stacking (no SBS3)	.878 (.840–.910)	.623	.096	—	—
OOF stacking + SBS3 (static)	**.906** (.876–.932)	**.678**	**.087**	**+.029**	**.0008**
$\text{BioAgent}_{\text{AcqAll}}$ (N=616)	.916 (.887–.941)	.703	.083	+.009	.068 (n.s.)

ΔAUC and p-values are sequential: stacking+SBS3 vs. stacking, and AcqAll vs. stacking+SBS3. Paired bootstrap n=2000, two-sided. The all-available N=958 estimate is reported separately in Table 2.

Patient-level ReACT execution. Figure 3 illustrates the actual patient TCGA-A8-A07L with held-out HRDsum 71. The initial late-fusion probability was 0.358, within the confidence-gated policy's escalation band ($0.15 \leq \hat{p} \leq 0.85$; Section 3.4); RNA-associated pathway observations and the available variant calls motivated an SBS3 analysis (cosine 0.569), after which the calibrated evidence model returned 0.838. The case illustrates a traceable computation consistent with the adaptive policy's own trigger condition, not a population-level agent effect. Methylation was unavailable and is not interpreted as a measured negative result. Additional data and the cohort-level provenance are provided in Appendix F.

4.4 Agent-Specific Evaluation

Component ablation. We ablate components of the evidence-guided meta-model g to quantify each contribution (Table 4).

The curated evidence-selection control, including the competitive L1 baseline, is retained in Appendix D; it should not be interpreted as proof of superiority over feature selection.

Deterministic critic reliability. We inject 85 synthetic failures across 10 prespecified categories into valid pipeline states and measure whether the deterministic critic correctly issues HOLD (Appendix E, Table 8). To validate specificity, 54 diverse clean states (varying patient counts, fold

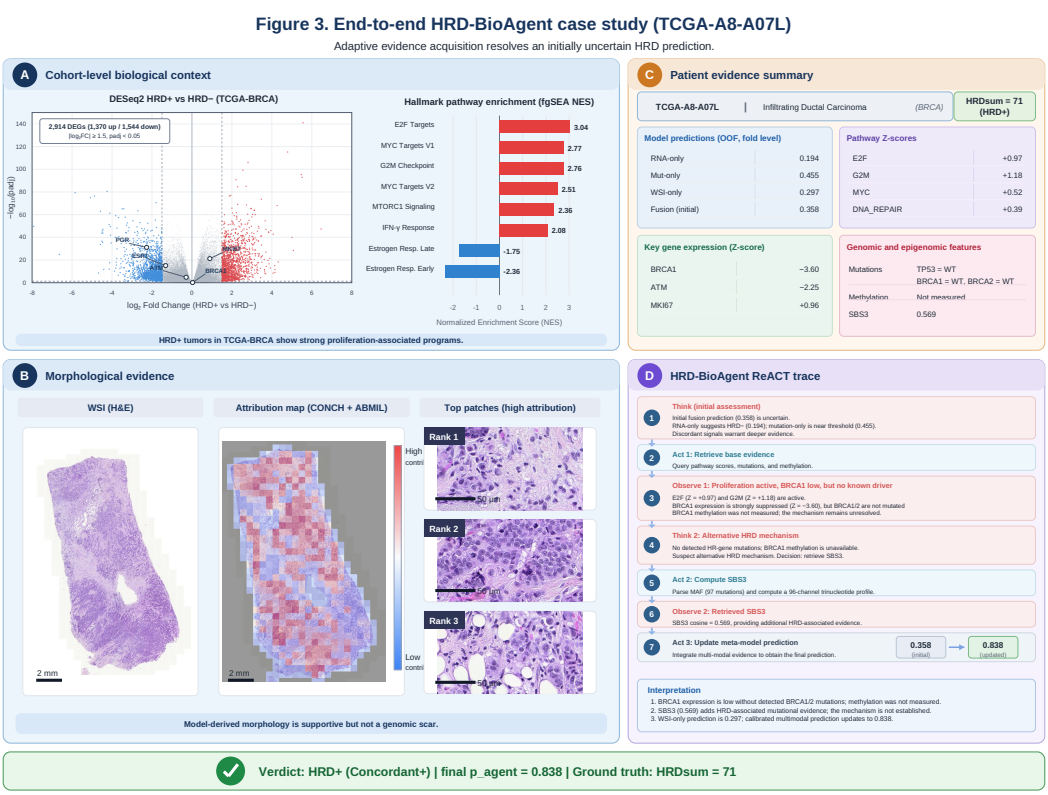

Figure 3: **End-to-end HRD-BioAgent case study for TCGA-A8-A07L.** Panels A–D show cohort-level biological context, model-derived WSI attribution, patient-level observations, and the ReACT evidence trace. The agent retrieves SBS3 (cosine 0.569) and updates the calibrated prediction from 0.358 to 0.838; HRDsum = 71 is the held-out reference (HRD+ denotes $\mathrm{HRDsum} \geq 42$). Methylation is *not measured* for this patient, and morphology attribution is not spatially resolved genomic truth. The example illustrates an execution trace, not a population-level effect size.

Table 4: **Agent ablation.** Full agent (ReACT) = OOF predictions + evidence + SBS3 + modality masks. RNA-only (no meta) is the best single-modality predictor for reference.

Configuration	N	AUROC	AUPRC	Brier↓	ECE↓	Δ
Full agent (acquire-all)	958	**.923**	**.749**	**.082**	**.023**	—
Agent (biology, no SBS3)	958	.900	.696	.090	.033	−.023
OOF stacking only	958	.899	.680	.094	.033	−.024
OOF + evidence (no SBS3)	958	.899	.695	.091	.021	−.024
Evidence only	958	.853	.610	.107	.013	−.070
RNA-only (no meta)	958	.892	.654	.097	.030	−.031

assignments, probability values, model types, evidence subsets, and modality combinations) are tested for false HOLD. Reliability claims are limited to the tested failure classes.

The critic held 85/85 injected failures across the prespecified ten categories and falsely held 0/54 clean states; these results quantify only the tested cases.

4.5 Robustness and Failure Analysis

Progressive modality masking. We evaluate full available data, removal of each modality, and reduced combinations (Appendix Table 13): *retrained* fits a separate fold-local meta-model per masking condition, while *mask-aware* trains one model with random modality dropout and tests it under every mask without retraining. The underlying biology-model results are in Appendix L; this does not establish full-pipeline robustness.

Evidence-discordance analysis. For each held-out patient we classify by prediction ($p \geq 0.5$) and evidence support (≥ 2 of: TP53 mutated, E2F $Z \geq 0.3$, BRCA1 methylated, BRCA1/2 mutated). Table 5 compares molecular characteristics across the four groups.

Table 5: **Prediction–evidence concordance.** Concordant+: high prediction + strong evidence. Discordant−: low prediction but evidence supports HRD, flagged for review.

Group	N (%)	HRD+/HRD−	$\overline{\text{HRDsum}}$	TP53	E2F $\bar{Z}$	BRCA1 meth	Acc.
Concordant+	105 (11)	78/27	55.8	95%	0.77	11%	.743
Discordant+	36 (4)	29/7	52.0	33%	0.33	3%	.806
Discordant−	103 (11)	26/77	34.5	97%	0.60	0%	.748
Concordant−	714 (75)	48/666	18.4	15%	−0.22	0%	.933

The Discordant− group (103 patients, 11%) was classified as prediction–evidence discordant under the prespecified rule, with 97% TP53 mutated and E2F Z=0.60. Among them, 26/103 (25%) are truly HRD-positive. The Discordant+ group (36 patients, 4%) predicts HRD-positive without strong evidence support, yet 29/36 (81%) are truly HRD-positive, suggesting predictive features beyond the curated evidence set. TP53 dominates the current concordance rule (subgroup analysis in Appendix A).

5 Discussion

HRD-BioAgent connects multimodal prediction to an explicit record of biological observations, evidence provenance, and possible additional analyses. The main matched-cohort sequence ($0.878 \rightarrow 0.906 \rightarrow 0.916$) shows SBS3 is a robust, significant predictor (two-sided p=0.0008), but the further increment from complete evidence integration is not distinguishable from chance under two-sided testing (p=0.068). We do not read this as evidence integration having no value—the estimate stays positive and the comparator is not feature-matched—but the paper's contribution rests on SBS3 retrieval, provenance, and auditable concordance rather than this increment; these acquire-all numbers upper-bound the evidence set's value, not a deployed policy's performance (Section 3.4; Appendix J).

The biomedical value of the framework is therefore not restricted to discrimination. A patient-level report distinguishes measured assays from model-predicted and secondary evidence; the critic checks provenance and workflow errors; and concordance groups expose disagreements requiring review rather than automatic relabeling. RNA scores and the RNA predictor are not statistically independent (same assay), and the WSI attribution map identifies model-associated regions, not spatially measured genomic HRD; the matched and all-available cohorts must remain separate in quantitative interpretations.

Limitations. All cohorts come from TCGA; independent paired-assay and treatment-response validation is lacking. BRCA evidence priors were not fully selected within outer training folds. Concordance depends strongly on TP53 and is not a clinical rule. The acquire-all reference upper-bounds discrimination; the adaptive policy's cost–AUROC trade-off (Appendix J) has no measured runtime or financial savings. Pan-cancer findings are heterogeneous (Appendix H). A feature-matched static control is still needed to isolate a reasoning effect within the acquire-all comparison.

6 Conclusion

HRD-BioAgent links calibrated HRD prediction to provenance-aware, patient-level evidence retrieval and auditing. Under an acquire-all reference policy, AUROC is 0.923 (N=958); in matched cases, $0.878 \rightarrow 0.906 \rightarrow 0.916$ shows SBS3 is a significant contributor (two-sided p=0.0008), while the further evidence-integration increment is not (two-sided p=0.068). A confidence-gated adaptive policy recovers most of this value at substantially lower cost (Appendix J).

REFERENCES

Ludmil B. Alexandrov, Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, et al. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, 2020.

Kevin M. Boehm, Pegah Khosravi, Rami Vanguri, Jianjiong Gao, and Sohrab P. Shah. Harnessing multimodal data integration to advance precision oncology. *Nature Reviews Clinical Oncology*, 19(7):448–464, 2022.

Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73, 2013.

Anika Cheerla and Olivier Gevaert. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics*, 35(14):i446–i454, 2019.

Richard J. Chen, Ming Y. Lu, Wei-Hung Weng, Tiffany Y. Chen, Drew F. K. Williamson, et al. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4025, 2021.

Richard J. Chen, Ming Y. Lu, Jingwen Wang, Drew F. K. Williamson, Scott J. Rodig, et al. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4):757–770, 2022a.

Richard J. Chen, Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Jana Lipkova, et al. Pancancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8): 865–878, 2022b.

Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, 2018.

Helen Davies, Dominik Glodzik, Sandro Morganella, Lucy R. Yates, Johan Staaf, et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nature Medicine*, 23(4):517–525, 2017.

Amelie Echle, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander T. Pearson, and Jakob Nikolas Kather. Deep learning in cancer pathology: a new generation of clinical biomarkers. *British Journal of Cancer*, 124(4):686–696, 2021.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the International Conference on Machine Learning*, pp. 2127–2136, 2018.

Guillaume Jaume, Anurag Vaidya, Richard J. Chen, Drew F. K. Williamson, Paul Pu Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11579–11590, 2024.

Jiuhai Jin, Hao Zhu, Jiaqi Chen, et al. AgentMD: Empowering language agents for clinical medicine. arXiv preprint arXiv:2410.13553, 2024.

Jakob Nikolas Kather, Lara R. Heij, Heike I. Grabsch, Chiara Loeffler, Amelie Echle, et al. Pancancer image-based detection of clinically actionable genetic alterations. *Nature Cancer*, 1(8): 789–799, 2020.

Gennady Korotkevich, Vladimir Sukhov, Nikolay Budin, Boris Shpak, Maxim N. Artyomov, and Alexey Sergushichev. Fast gene set enrichment analysis. bioRxiv, 060012, 2021.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdottir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6): 417–425, 2015.

Jana Lipkova, Richard J. Chen, Bowen Chen, Ming Y. Lu, Matteo Barbieri, et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*, 40(10):1095–1110, 2022.

Christopher J. Lord and Alan Ashworth. BRCAness revisited. *Nature Reviews Cancer*, 16(2):110–120, 2016.

Christopher J. Lord and Alan Ashworth. PARP inhibitors: Synthetic lethality in the clinic. *Science*, 355(6330):1152–1158, 2017.

Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024a.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Melissa Zhao, et al. A multimodal generative AI copilot for human pathology. *Nature*, 634(8033):466–473, 2024b.

Andrea Marion Marquard, Aron Charles Eklund, Tejal Joshi, Marcin Krzystanek, Francesco Favero, et al. Pan-cancer analysis of genomic scar signatures associated with homologous recombination deficiency suggests novel indications for existing cancer drugs. *Biomarker Research*, 3:9, 2015.

Luan Nguyen, John W. M. Martens, Arne Van Hoeck, and Edwin Van Cuppen. Pan-cancer landscape of homologous recombination deficiency. *Nature Communications*, 11(1):5584, 2020.

Zsofia Sztupinszki, Miklos Diossy, Marcin Krzystanek, Lilla Reiniger, Istvan Csabai, et al. Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. *npj Breast Cancer*, 4:16, 2018.

Melinda L. Telli, Kirsten M. Timms, Julia Reid, Bryan Hennessy, Gordon B. Mills, et al. Homologous recombination deficiency (HRD) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast cancer. *Clinical Cancer Research*, 22 (15):3764–3773, 2016.

Lukas L. Weishaupt, Chengkuan Chen, Drew F. K. Williamson, Richard J. Chen, Guillaume Jaume, et al. Evidence-based diagnostic reasoning with multi-agent copilot for human pathology. arXiv preprint arXiv:2506.20964v2, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002.

A TP53 Concordance Sensitivity

TP53 sensitivity and weighted evidence. Since TP53 mutation is context-dependent (Section 4.2), we repeat concordance analysis under two alternative evidence rules: (1) TP53 weight halved (w_{TP53}=0.5, requiring TP53 plus at least one other signal), and (2) TP53 excluded entirely, requiring $\geq$1 non-TP53 signal. Under rule (2), Discordant$-$ increases from 103 to 190 patients (many current Concordant+ patients lose evidence support), confirming that TP53 is the dominant concordance signal. Under the original rule, Concordant+ contains 105 patients; without TP53, only 25 retain evidence support from pathway or BRCA1 features alone. This does not invalidate the analysis—TP53 is genuinely enriched in HRD+ tumors (OR=8.07)—but underscores that the current evidence set is heavily weighted toward mutation-driven biology. A weighted evidence score $E_i = \sum_j w_j e_j$ with subtype-aware TP53 weighting (e.g., $w_{\text{TP53}}\approx 0$ in TNBC) would align better with the Evidence Map concept and is a direction for future work.

B Implementation and Agent Specification

B.1 Agent Configuration

The HRD-BioAgent planner receives the following task specification:

```
{
  "cancer_type": "BRCA",
  "analysis_mode": "cohort",
  "modality_priority": ["RNA", "Mutation", "Methylation", "WSI"],
  "target": "HRD",
  "hrd_cutoff": 42,
  "hrd_label": "HRDsum >= 42",
  "excluded_from_prediction": ["TMB","BRCA1","BRCA2",
                               "PALB2","RAD51C","RAD51D"],
  "evidence_features": ["pw_E2F_TARGETS","pw_G2M_CHECKPOINT",
      "pw_MYC_TARGETS_V1","pw_DNA_REPAIR","BRCA1_expr",
      "mut_TP53","mut_BRCA1","mut_BRCA2","BRCA1_meth",
      "SBS3_cosine","indel_fraction"],
  "react_workflow": "Think-Act-Observe loop per patient",
  "meta_model": "CalibratedClassifierCV(LogisticRegression,
                 isotonic, fold-local)",
  "analysis_steps": ["modality_profiling","hrd_prediction",
      "evidence_retrieval","concordance","critique"]
}
```

B.2 Deterministic Critic Rules

The critic validates each pipeline output state and returns Approve, Replan, or Hold. The following checks are applied:

- **Cancer type**: must be non-empty and equal to `BRCA`.
- **Probability**: must be a numeric value in $[0, 1]$; rejects `None`, booleans, `NaN`, $\pm\infty$, and out-of-range values.
- **Fold**: must be an integer in $\{0, 1, 2, 3, 4\}$.
- **Model**: must be in {`rna_only`, `mut_only`, `meth_only`, `wsi_only`, `rna_plus_mut`, `early_fusion`, `late_fusion`, `hrd_bioagent`}.
- **OOF provenance**: all non-WSI predictions must have provenance `retrospective_out_of_fold`.
- **Evidence features**: must be from the 11-feature allowed set (9 base + SBS3 cosine + indel fraction); mutation and methylation features must be binary.
- **Evidence provenance**: evidence sources must match the declared cancer type and cohort (`TCGA-BRCA`).
- **Treatment**: any treatment recommendation triggers Hold (research predictions cannot authorize treatment).

- **Patient–task match**: predicted patient set must equal the requested patient set.
- **Meta-model**: type `llm_freeform` is rejected; the meta-model g must be a calibrated estimator.

Any violation triggers HOLD with a list of issues; recoverable violations (e.g., incomplete evidence retrieval) trigger REPLAN to re-enter the ReACT loop; only states passing all checks receive APPROVE. Table 8 reports detection of 85 synthetic failures from ten prespecified categories; this does not validate unseen error classes.

B.3 HRD EVIDENCE ENTRY SCHEMA

```
{
  "feature": "pw_E2F_TARGETS",
  "value": 0.77,
  "type": "MEASURED",
  "source": "Hallmark GSEA (TCGA-BRCA RNA-seq VST)",
  "cohort": "TCGA-BRCA",
  "threshold": "Z >= 0.3 -> HRD-associated",
  "provenance": "fold-local z-score"
}
```

Each evidence entry records the feature name, patient-level value, measurement type (MEASURED, PREDICTED, or SECONDARY), data source, cohort restriction, prespecified threshold, and provenance. The critic rejects entries from mismatched cohorts or cancer types.

C PREDICTOR PORTFOLIO

Table 6: HRD prediction model portfolio with required modalities and cohort sizes.

Model	**Required modalities**	N	**HRD AUC**
RNA-only (MLP)	RNA	958	.892
RNA+Mut (residual)	RNA, Mut	958	.885
Late fusion (stacking)	RNA, Mut, Meth	616	.884
Early fusion (concat)	RNA, Mut, Meth	616	.865
WSI-only (ABMIL)	WSI	958	.833
Mutation-only (MLP)	Mut	898	.816
Methylation-only (MLP)	Meth	616	.809
HRD-BioAgent (biology)	All OOF + evidence	958	.900
HRD-BioAgent (acquire-all)	All OOF + evidence + SBS3	958	**.923**

D EVIDENCE-SELECTION CONTROLS

Tool accessibility ablation. To test whether the agent's performance comes from *which* evidence it selects (biological reasoning) or merely from having access to evidence features (tool accessibility), we compare the agent's 11 biologically curated features against alternative feature selection strategies given a 32-candidate evidence pool (Table 7). All conditions share the same base inputs (five OOF predictions and modality masks); these base inputs are outside the 32 candidate evidence features.

E CRITIC STRESS TEST DETAILS

F PATIENT-LEVEL SUPPORTING DETAIL

Figure 3 presents the patient-level evidence trace for TCGA-A8-A07L, including SBS3 retrieval and calibrated integration. This case reports a known HRDsum for evaluation only; it is not supplied as a predictor input. The methylation modality is unavailable for the depicted patient and cannot be interpreted as a negative assay.

Table 7: **Tool accessibility ablation** (N=898, 5-fold patient-level CV). All conditions use the same 5 OOF predictions and modality masks as base; only the evidence features differ. The 32-candidate pool includes 4 pathway Z-scores, BRCA1 expression, 24 mutation features, BRCA1 methylation, SBS3 cosine, and indel fraction. The 32-feature evidence pool excludes OOF predictions and masks; random-11 is evaluated in 100 draws.

Feature selection	K	AUROC	AUPRC	Brier↓	ECE↓	Note
No evidence	0	.890	.670	.097	.032	Baseline
Kitchen sink (all)	32	.913	.715	.088	.023	All features
Random-11 (mean±sd)	11	.898±.014	—	—	—	100 trials
Random-11 (best of 100)	11	.921	.741	.083	.032	Selected extreme
L1 automated	29	.920	.731	.086	.024	C=0.5
Curated biological set	11	.919	.737	.084	.019	Curated

In this exploratory experiment the curated set is near the upper tail of 100 random draws. Its differences from the all-feature set (ΔAUC$= +.006$, $p = .135$) and the 29-feature L1 control (ΔAUC$= -.001$, $p = .58$) are not statistically significant. The result supports compact, interpretable evidence selection, not superiority or a separately isolated agent-reasoning effect.

Table 8: **Critic stress test.** 85 synthetic failures across 10 prespecified categories injected into valid pipeline states; the measured detection rate does not establish performance on unseen errors. Detection = fraction correctly held. False HOLD measured on 54 diverse clean states.

Failure category	Cases	Detection
Missing cancer type	3	100%
Wrong cancer type	10	100%
Invalid probability	10	100%
Wrong fold ID	10	100%
Unsupported model	10	100%
Missing OOF provenance	10	100%
Cross-cancer cached evidence	10	100%
Treatment authorization	10	100%
Patient–task mismatch	10	100%
LLM as meta-model	2	100%
Total corrupted	**85**	**100%**

False HOLD rate: 0/54 (0%). 54 diverse clean states correctly approved.

Step 1: Cohort-level biological signature discovery (DESeq2 (Love et al., 2014)/fGSEA (Korotkevich et al., 2021)). Cohort-level HRD vs. HRD− differential expression analysis supplies a source of candidate Evidence Map entries (Figure 2); Figure 3 focuses on the patient-level execution. DESeq2 identified 2,914 DEGs (1,370 up / 1,544 down; $|\log_2 \mathrm{FC}| \geq 1.5$, $p_{\mathrm{adj}} < 0.05$) between HRD+ and HRD− groups. fGSEA on ranked DEGs revealed 36/50 Hallmark pathways significantly enriched ($p_{\mathrm{adj}} < 0.05$): E2F Targets (NES=3.04), MYC Targets V1 (NES=2.77), G2M Checkpoint (NES=2.76), DNA Repair (NES=1.84) were the top positively enriched pathways, while Estrogen Response Early (NES$= -2.36$) was the most strongly depleted. Fisher's exact test identified TP53 co-mutation (OR=8.07, p=1.1×10^{-31}) as the dominant somatic event. BRCA1 promoter methylation prevalence was 10/102 (9.8%) in the matched cross-modal HRD+ subset and 11/89 (12.4%) in the more narrowly eligible HR-WT HRD+ methylation-analysis cohort (Table 1); the denominators differ. These cohort-level signatures provide candidate evidence types and associations; patient-level assay availability and provenance determine the admissible retrieved observations.

G Pan-Cancer Pathway Conservation

The E2F/G2M/MYC proliferation core is consistently enriched in HRD-positive tumors across all six cohorts, supporting its use as a tissue-invariant evidence feature. TP53 and estrogen-response associations are tissue-dependent and therefore carry cancer-type restrictions in the Evidence Map.

H Exploratory Pan-Cancer HRD Prediction

To evaluate whether the framework generalizes beyond breast cancer, we replicate the evidence decomposition (baseline → +SBS3 → ReACT) on three additional TCGA cohorts with sufficient HRD-positive cases: OV, UCEC, and UCS (Table 10). In the reported fold-nested design, each can-

Table 9: Tissue-invariant Hallmark pathway enrichment in HRD+ vs HRD− across 6 TCGA cancer types (BRCA, OV, UCEC, UCS, PAAD, PRAD).

Pathway	Direction	Mean NES	Cancers enriched
E2F_TARGETS	UP	2.29	6/6
G2M_CHECKPOINT	UP	2.18	6/6
MYC_TARGETS_V1	UP	1.94	6/6
DNA_REPAIR	UP	1.76	5/6
MITOTIC_SPINDLE	UP	1.53	4/6
ESTROGEN_RESPONSE_EARLY	DOWN	−1.95	2/6

cer's evidence map is intended to be discovered from outer-training partitions, and we additionally compare against direct BRCA map transfer to isolate the value of cancer-specific adaptation.

Fold-nested cancer-specific evidence maps. A key design principle of HRD-BioAgent is that the agent selects evidence features based on the biological context rather than applying a fixed feature set across cancer types. The stated procedure derives the evidence map from each outer training split: Wilcoxon rank-sum tests select differentially enriched Hallmark pathways (Liberzon et al., 2015) (FDR < 0.05) and Fisher's exact test selects HRD-associated mutations (FDR < 0.05). The following cross-fold selection counts are descriptive, with consensus defined as selection in at least 50% of folds. Test-fold inference should use only the map constructed from the corresponding training partition, not a map aggregated across test folds:

- **BRCA** (N=616, 17% HRD+): E2F, G2M, MYC_V1, DNA_REPAIR (all 5/5 folds) + TP53, BRCA1, BRCA2 mutations (5/5). Highly stable across folds.
- **OV** (N=166, 55% HRD+): Angiogenesis (3/5 folds), PIK3CA mutation (3/5), and recurrent but non-consensus OxPhos (2/5)—different from BRCA's E2F/G2M core. TP53 is mutated in 95% of OV patients and does not discriminate HRD status.
- **UCEC** (N=482, 10% HRD+): E2F (5/5), MYC_V2 (5/5), G2M (4/5) + TP53, PTEN, ARID1A mutations (all 5/5). PTEN and ARID1A are newly discovered cancer-specific HRD-associated mutations not present in the BRCA evidence map. Additionally, POLE and mismatch-repair (MMR) gene mutations (MSH2, MSH6, MLH1, PMS2), extracted from MAF files, serve as multi-step reasoning targets (see below).
- **UCS** (N=56, 34% HRD+): EMT (2/3 folds) is the only consensus pathway; no mutations reach significance. The evidence map is unstable due to small sample size (N=56, 3-fold CV).

Key observations. **OV**: SBS3 provides a large improvement over stacking (ΔAUC = +.132, p=.001), consistent with the known role of HR deficiency in high-grade serous ovarian cancer. The fold-nested cancer-specific map (Angiogenesis + PIK3CA; OxPhos recurrent in 2/5) achieves .803, comparable to the BRCA map transfer (.805) and all-pathway agent (.802)—suggesting that in OV the primary signal comes from SBS3 rather than pathway features. **UCEC**: Cancer-specific ReACT (.891) significantly outperforms both stacking (ΔAUC = +.058, p=.004) and stacking+SBS3 (ΔAUC = +.033, p=.030), and is numerically higher than BRCA-map transfer (.874; ΔAUC = +.017, p=.085, not significant). PTEN and ARID1A were selected as UCEC-associated candidates; the present ablations do not isolate their individual causal contribution. **UCS** (N=56): BRCA map transfer (.669) outperforms cancer-specific (.615), likely because the small sample yields an unstable fold-nested evidence map (EMT only); the richer BRCA map compensates despite tissue mismatch. All UCS comparisons are non-significant due to wide CIs.

UCEC multi-step reasoning: POLE/MSI screening. UCEC comprises four molecular subtypes with distinct genomic instability mechanisms: POLE ultra-mutated, MSI-high (MMR deficient), copy-number low (endometrioid), and copy-number high (serous-like). The serous-like subtype is more frequently associated with the HRD-scar phenotype in this analysis; subtype annotations and mutational flags are not interchangeable. Extracting POLE and MMR gene (MSH2, MSH6, MLH1, PMS2) mutations from MAF files reveals that 119/482 (24.7%) of UCEC patients carry at least one such mutation flag (not a direct MSI assay), with an HRD-positive rate of only 1.7% compared to 12.4% in non-POLE/MSI patients (Fisher's exact OR = 0.121, p=2.5×10^{-4}).

Table 10: **Pan-cancer HRD prediction with fold-nested evidence maps** (within-cancer stratified CV). The described protocol derives maps from each outer training partition using Wilcoxon/Fisher tests (FDR < 0.05). The across-fold consensus frequencies are descriptive, not an input to outer-test predictions. "BRCA map" transfers the BRCA evidence map without cancer-specific adaptation; "Agent (all pw)" uses all 11 pathways; "Agent ReACT" denotes the reported cancer-specific configuration. Implementation-level verification of map construction and test-fold isolation is required before interpreting these estimates as leakage-free.

Cancer	Method	N	HRD+ (%)	AUROC	AUPRC	Brier↓	ECE↓
OV	OOF stacking	166	55	.655	.709	.230	.085
	OOF stacking + SBS3	166	55	.787	.774	.185	.067
	Agent (BRCA map)	166	55	.805	.835	.185	.105
	Agent (all pw)	166	55	.802	.814	.188	.107
	Agent ReACT	**166**	**55**	**.803**	**.822**	**.185**	**.085**
UCEC	OOF stacking	482	10	.834	.283	.086	.047
	OOF stacking + SBS3	482	10	.858	.305	.081	.042
	Agent (BRCA map)	482	10	.874	.321	.080	.044
	Agent (all pw)	482	10	.894	.378	.070	.026
	Agent ReACT	482	10	.891	.360	.076	.040
	+ POLE/MSI reasoning	**482**	**10**	**.892**	**.362**	**.074**	**.032**
UCS	OOF stacking	56	34	.607	.414	.222	.128
	OOF stacking + SBS3	56	34	.573	.509	.230	.136
	Agent (BRCA map)	56	34	**.669**	**.492**	**.216**	.143
	Agent (all pw)	56	34	.594	.381	.235	.137
	Agent ReACT	56	34	.615	.483	.226	.129
BRCA	OOF stacking	616	17	.878	.623	.096	.036
	OOF stacking + SBS3	616	17	.906	.678	.087	.038
	Agent (all pw)	616	17	.897	.658	.088	.043
	Agent ReACT	**616**	**17**	**.916**	**.703**	**.083**	.039

Bootstrap (n=2000): OV ReACT vs. stacking ΔAUC = +.148, p=.001; UCEC ReACT vs. stacking ΔAUC = +.058, p=.004; vs. stacking+SBS3 ΔAUC = +.033, p=.030; vs. BRCA map ΔAUC = +.017, p=.085; UCEC POLE/MSI reasoning vs. stacking ΔAUC = +.059, p=.011; vs. ReACT ΔAUC = +.001, p=.453 (n.s.). UCS BRCA map (.669) outperforms cancer-specific (.615) but CIs overlap (p=.726) due to small N.

Adding POLE/MSI features statically to the evidence map does not improve AUROC (.891 → .891), indicating no observed additional discrimination from unconditional augmentation in this comparison. However, multi-step reasoning—where Stage 2 re-evaluates uncertain patients ($0.15 \leq \hat{p} \leq 0.65$; n=110) with POLE/MSI evidence—achieves the best calibration (ECE .032 vs. .040) and reduces false positives from 21 to 15 (-29%). The AUROC improvement over the evidence map is marginal (ΔAUC = +.001, p=.45) because the expression-based evidence model already captures most POLE/MSI discrimination through pathway correlates: only 1 of 119 POLE/MSI patients receives $\hat{p} > 0.5$. The available subtype-correlated evidence already provides substantial discrimination; the multi-step variant shows an exploratory calibration change for borderline predictions.

The SBS3 mutation signature (COSMIC SBS3 cosine) shows cancer-dependent discrimination: HRD+ mean = 0.512 vs. HRD− = 0.438 in OV; 0.431 vs. 0.396 in UCEC; 0.447 vs. 0.407 in UCS (vs. 0.465 vs. 0.358 in BRCA). SBS3 is most discriminative in BRCA and OV, consistent with the known prevalence of HR-deficient mutational processes in these cancers.

The observations, which are exploratory within TCGA rather than external validation, are: (1) the evidence decomposition pattern (baseline → +SBS3 → +cancer-specific evidence) generalizes across cancer types with varying effect sizes; (2) the described fold-local map discovery produces different tissue-specific candidate profiles (OV: Angiogenesis/PIK3CA; UCEC: PTEN/ARID1A); (3) UCEC shows a numerical, non-significant advantage for locally discovered evidence over BRCA map transfer, whereas BRCA-map transfer performs numerically better than the local map in small UCS; no automated fallback was implemented; and (4) multi-step reasoning with cancer-specific screening targets (UCEC POLE/MSI) improves calibration and specificity even when the primary evidence map already captures most subtype discrimination, showing that selective retrieval can alter calibration for UCEC even where AUROC changes little.

Together with the tool accessibility ablation (Table 7), these pan-cancer results illustrate cancer-dependent utility of candidate evidence sources; they do not by themselves isolate the effect of language-model reasoning.

I Adaptive Cross-Modal Evidence Discovery

To evaluate whether the agent's Evidence Map can be discovered from data rather than manually curated, we implement fold-local cross-modal evidence discovery using statistical tests with FDR correction.

Method. Within each outer CV fold (k=1, . . . , 5), we test 25 candidate features (20 mutation indicators from frequently mutated genes + BRCA1 methylation + 4 pathway Z-scores) for association with HRD status. Binary features are tested with Fisher's exact test; continuous features with Wilcoxon rank-sum test. All p-values are FDR-corrected (Benjamini–Hochberg) within each fold. Features with FDR < 0.05 are selected. SBS3 cosine similarity is always retrieved via the ReACT workflow.

Feature stability. Six features are discovered in all 5/5 folds: TP53 (OR = 7.7–8.4), PIK3CA (OR = 0.25–0.28), CDH1 (OR = 0.08–0.17), GATA3 (OR = 0.22–0.38), MAP3K1 (OR = 0.06–0.20), and BRCA1 methylation (OR = 41–46). BRCA1 and BRCA2 mutations are discovered in 3/5 folds each, reflecting their lower prevalence and borderline significance in some folds. ARID1A is discovered in 1/5 folds only.

Biological interpretation. The adaptively discovered features partition into two groups: (1) *HRD-associated* (TP53, BRCA1 methylation) with OR $\gg 1$, reflecting known HR deficiency mechanisms; and (2) *inversely HRD-associated* (PIK3CA, CDH1, GATA3, MAP3K1) with OR $\ll 1$, reflecting luminal-enriched genes whose mutations are inversely associated with genomic instability. These inverse associations may partly reflect breast cancer subtype composition and are not causal protective effects. Two stable signals (TP53 and BRCA1 methylation) belong to the prespecified 11-feature set. PIK3CA, CDH1, GATA3, and MAP3K1 are additional candidates discovered in this exploratory screen, not components of the evaluated 11-feature model. These results motivate possible map extension; they do not establish independent biological validation.

J Cost-Efficiency Analysis

This appendix instantiates the two acquisition policies formalized in Section 3.4. The *acquire-all* reference (a_i = +SBS3 for all i) supplies all 11 evidence features when available (six defined analysis steps: four pathway scores, mutation/methylation lookup, and SBS3 computation); it upper-bounds the value of the evidence set and is not a measurement of selective retrieval at deployment. Two *confidence-gated adaptive* variants are compared below; their step counts proxy analysis effort rather than measured wall-clock time or financial cost. Both reduce computation for patients whose prediction is already confident, i.e., for whom escalating a_i is unlikely to change the outcome.

Tier-based routing. Patients are stratified into confidence tiers based on the RNA OOF prediction: *easy-negative* ($\hat{p} < 0.10$; N=618, 68.8%), *easy-positive* ($\hat{p} > 0.80$; N=53, 5.9%), *moderate-positive* ($0.50 \leq \hat{p} \leq 0.80$; N=15, 1.7%), and *hard* (remaining; N=212, 23.6%). Easy patients use only OOF predictions (1 step); moderate patients add mutation/methylation (3 steps); hard patients receive the full pipeline (6 steps). Average steps: 2.2 (63.1% savings). AUROC: .902 ($\Delta = -.017$ vs. full).

Confidence-based escalation. All patients start with OOF predictions only. If confidence is below threshold ($0.15 \leq \hat{p} \leq 0.85$), mutation/methylation evidence is retrieved. If still uncertain ($0.25 \leq \hat{p} \leq 0.75$), SBS3 is computed. Average steps: 2.5 (58.3% savings). AUROC: .907 ($\Delta = -.012$ vs. full). This strategy outperforms tier routing by .005 AUC with 5% more computation, suggesting that the adaptive escalation order (OOF $\rightarrow$ mutations $\rightarrow$ SBS3) better preserves performance on borderline cases.

Both strategies reduce the number of scheduled evidence steps at a measurable AUROC cost. Real tool latency and prospective operating cost were not measured.

Table 11: **Cost-efficiency analysis of the two acquisition policies from Section 3.4** (N=898). Acquire-all retrieves every evidence feature for every patient; the adaptive variants escalate acquisition depth only for patients whose prediction remains uncertain.

Strategy	AUROC	Avg. steps	Savings	ΔAUC vs. full
HRD-BioAgent (acquire-all)	.919	6.0	0%	—
HRD-BioAgent (adaptive, tier-based)	.902	2.2	63.1%	−.017
HRD-BioAgent (adaptive, confidence-gated)	.907	2.5	58.3%	−.012

Table 12: **HRDsum cutoff sensitivity analysis. A.** Prediction accuracy by HRDsum range. Errors concentrate at the cutoff boundary (42–49: 43% accuracy) while extreme ranges approach perfect classification. **B.** AUROC after excluding symmetric bands around the HRDsum = 42 cutoff. Removing label-ambiguous patients raises AUROC from .923 to .974 (±10) or .994 (±20), consistent with better discrimination in more extreme HRDsum groups; exclusion of borderline cases also makes the classification problem easier and does not establish incorrect labels.

A. Accuracy by HRDsum range

HRDsum	N	FN	FP	Acc.
$[0, 20)$	460	–	2	99.6%
$[20, 30)$	172	–	13	92.4%
$[30, 35)$	69	–	11	84.1%
$[35, 42)$	76	–	17	77.6%
$[42, 49)$	44	25	–	43.2%
$[49, 60)$	56	31	–	44.6%
$[60, 80)$	72	16	–	77.8%
$[80, \infty)$	9	0	–	100%

B. AUC by exclusion band

Exclusion	N	HRD+	AUC	Δ
None (all)	958	181	.923	—
±3	894	151	.948	+.026
±5	869	145	.954	+.031
±7	838	137	.959	+.036
±10	776	117	.974	+.051
±15	684	94	.985	+.063
±20	577	77	.994	+.072

Spearman ρ=0.72 ($p<10^{-155}$) between HRD-BioAgent continuous predictions and HRDsum, indicating strong monotonic tracking of genomic instability across the full range.

K Threshold-Sensitivity Analysis (Exploratory)

Prediction errors cluster near the binary HRDsum $\geq$ 42 cutoff (Table 12, Figure 4). Of 72 false negatives, 25 (34.7%) fall in the borderline HRDsum 42–49 range, and 46 (63.9%) fall in the 42–55 range. Of 43 false positives, 17 (39.5%) have HRDsum 35–42, just below the cutoff. By contrast, the model achieves 99.6% accuracy for HRDsum $<$ 20 and 100% for HRDsum $\geq$ 80.

As an exploratory sensitivity analysis, excluding patients near the cutoff increases measured discrimination: removing the ±10 band (HRDsum 32–52) raises AUROC from .923 to .974, and removing ±20 yields .994 (Table 12). The retained cohort is easier to discriminate; this result is not evidence by itself that the excluded labels are erroneous. The continuous correlation between HRD-BioAgent predictions and HRDsum is strong (Spearman ρ=0.72, $p<10^{-155}$), indicating that the model tracks genomic instability well across the full range. Molecularly, borderline false negatives (HRDsum 42–49) show weak HRD phenotypes: TP53 mutation 52% (vs. 83% in true positives), E2F Z=0.16 (vs. 0.65), BRCA1/2 mutation 0% (vs. 17%), and mean SBS3 cosine 0.417 (vs. 0.467). These patients accumulate modest genomic scars without the molecular hallmarks that expression-based models rely on.

L Additional Missing-Modality Results

M Extended Related Work

HRD prediction from histology and molecular data. Homologous recombination deficiency determines eligibility for PARP inhibitor therapy (Lord & Ashworth, 2017), making accurate HRD prediction clinically actionable. Prior work has shown that HRD-related molecular phenotypes can be predicted from H&E whole-slide images (Kather et al., 2020; Echle et al., 2021; Coudray et al., 2018) and from molecular measurements. HRDetect (Davies et al., 2017) uses six whole-genome sequencing features—including mutational signatures SBS3/SBS8 and structural variant patterns—to classify BRCA1/2-deficient tumors with AUC $>$ 0.98, but requires whole-genome sequencing unavailable in most clinical settings. CHORD (Nguyen et al., 2020) extends this to a pan-cancer

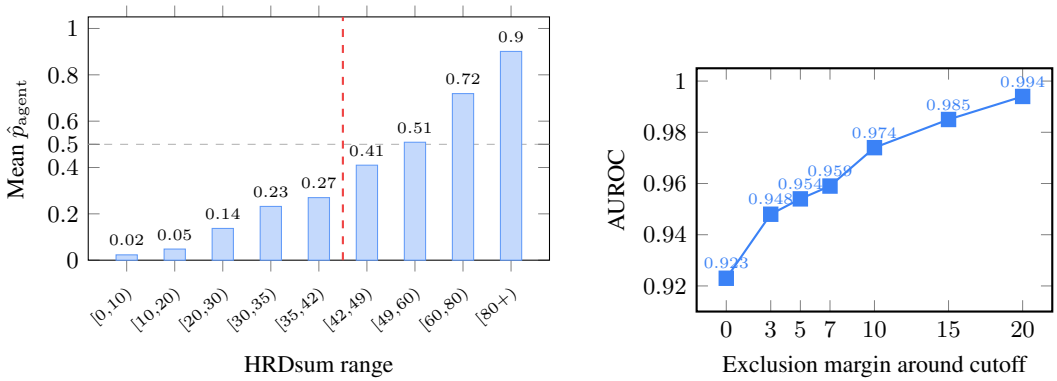

Figure 4: **HRDsum cutoff sensitivity. Left:** Mean agent prediction by HRDsum range. Predictions increase monotonically with HRDsum, but patients near the cutoff ([42,49): mean $\hat{p}$=0.41) fall in the uncertain zone where the binary label is most ambiguous. **Right:** AUROC after excluding symmetric bands around HRDsum = 42. Performance rises steeply as label-ambiguous patients are removed, reaching .994 at ±20 margin, indicating sensitivity of discrimination to the exclusion of borderline HRDsum patients, without attributing the entire difference to label noise.

Table 13: **Missing-modality stress test.** Retrained: a separate meta-model is trained per condition. Mask-aware: a single model trained with modality dropout, tested under each mask (no retraining at test time).

	Retrained per condition				Mask-aware (single model)			
Condition	AUC	AUPRC	Brier	ECE	AUC	AUPRC	Brier	ECE
Full	.900	.696	.090	.033	.898	.686	.092	.033
−WSI	.899	.693	.091	.026	.892	.670	.093	.025
−Meth	.900	.713	.090	.022	.901	.700	.089	.031
−Mut	.875	.664	.097	.030	.874	.648	.098	.027
−RNA	.872	.638	.102	.028	.858	.589	.108	.016
RNA+Mut	.900	.711	.090	.019	.893	.680	.091	.021
RNA only	.878	.676	.095	.022	.865	.635	.101	.042
WSI only	.827	.499	.115	.015	.810	.461	.121	.091

classifier using structural variant context, again requiring WGS. scarHRD (Sztupinszki et al., 2018) computes LOH/LST/TAI scar scores from SNP arrays or targeted panels, providing the widely used HRDsum that we adopt as our reference label.

These approaches establish HRD as a learnable phenotype, but each operates within a single modality—WGS features, copy-number scars, or histology alone—and prediction probability does not encode whether orthogonal molecular evidence supports the same patient-level conclusion. HRD-BioAgent differs in two respects: (1) it integrates across four modalities (RNA, mutation, methylation, WSI) without requiring WGS, and (2) it explicitly reasons about whether biological evidence corroborates or contradicts the prediction, rather than returning a scalar probability.

Multimodal and missing-modality learning. Multimodal oncology systems have progressed from feature concatenation to attention-based architectures. PORPOISE (Chen et al., 2022b) integrates histology and genomic features via a multimodal co-attention framework for pan-cancer survival prediction and achieves superior prognostication compared to unimodal baselines. MCAT (Chen et al., 2021) uses cross-attention between WSI regions and genomic pathways, enabling interpretable survival prediction from gigapixel images. Pathomic Fusion (Chen et al., 2022a) employs gated attention and tensor fusion for diagnosis and prognosis. SurvPath (Jaume et al., 2024) models dense interactions between biological pathways and histological features.

These systems optimize a unified objective (typically survival) across modalities via learned fusion. Missing modalities are handled through modality dropout (Cheerla & Gevaert, 2019) or imputation (Boehm et al., 2022; Lipkova et al., 2022). Our approach differs structurally: rather than jointly optimizing a fused representation, HRD-BioAgent uses modality-specific predictors as inputs to a

meta-model that integrates them alongside curated biological evidence. Unavailable assays are not synthesized or imputed; instead, the system reasons over validated signatures that remain observable and labels any conclusion about an unmeasured assay as secondary inference.

Biomedical agents and the deterministic agent distinction. Recent pathology agents have demonstrated tool orchestration, WSI navigation, verification, and biological discovery (Weishaupt et al., 2026; Jin et al., 2024). PathChat (Lu et al., 2024b) uses vision–language models for interactive pathology interpretation. These systems leverage LLM capabilities for flexible reasoning and natural-language interaction.

HRD-BioAgent adopts the agent abstraction—tool orchestration, conditional evidence retrieval, structured reasoning traces—but enforces a critical constraint: **all numerical predictions are produced by calibrated, deterministic estimators, not by an LLM**. The meta-model g (Eq. 9) is a logistic regression with isotonic calibration, fit fold-locally with fixed random seeds; re-running the pipeline on identical inputs produces identical outputs. The LLM's role is restricted to orchestration (deciding *which* evidence to retrieve) and synthesis (generating reasoning traces); it never assigns HRD probabilities. This design enables full reproducibility—a requirement for clinical biomarker validation—while retaining the agent's adaptive evidence retrieval capability. Deterministic critics (Appendix B.2) further enforce provenance, fold isolation, and cancer-context constraints, providing auditable guardrails that complement rather than replace the statistical estimator.