

A Bounded Agentic In Silico Framework for Age-Aware Molecular Candidate Prioritization in Breast Cancer

Farah Deeba Memon

`dr.deeba@scnu.edu.cn`

South China Normal University

Ahmad Yar Sukhera

Harbin Institute of Technology

Xingyi Ma

Harbin Institute of Technology

Research Article

Keywords: agentic AI, ageing, breast cancer, molecular prioritization, network biology, transcriptomics

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11201828/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Age-related molecular heterogeneity in breast cancer complicates the prioritization of reproducible research candidates across heterogeneous molecular resources. We developed a bounded agentic in silico framework integrating curated ageing genes, The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA), the independent GSE42568 cohort, STRING protein-protein associations, and Gene Ontology and Kyoto Encyclopedia of Genes and Genomes enrichment. Of 307 curated human ageing-related genes, 292 were expression-testable in 1,079 TCGA-BRCA primary tumors. Continuous-age screening identified 126 genes at a false discovery rate below 0.05, and 80 remained significant after adjustment for estrogen receptor, progesterone receptor, human epidermal growth factor receptor 2, and pathological stage. Among 40 prespecified candidates tested in GSE42568, eight replicated directionally; IL7R, SQSTM1, EGR1, and IGF1 were both cross-cohort replicated and robust across estrogen-receptor strata. Stratified age-label permutation supported the overall discovery signal (empirical $P = 0.001996$). STRING hub rankings were stable across interaction-confidence thresholds (Spearman $\rho = 0.81-0.90$), and 514 significant functional terms were reduced to 30 representative modules. A 760-scenario process-aware benchmark compared static, goal-planning, adaptive, and bounded controllers. The bounded controller achieved 1.000 for policy conformance, feasible-goal completion, action appropriateness, safety conformance, provenance conformance, and escalation conformance, with resource stewardship of 0.988. The framework provides a reproducible computational route for molecular candidate prioritization and hypothesis generation while explicitly excluding autonomous diagnosis, prognosis, treatment recommendation, and clinical biomarker declaration.

1 Introduction

Breast cancer is the most prevalent cancer in women and is molecularly heterogeneous both between patients and within tumors (Bray et al. 2024; Cancer Genome Atlas Network 2012). Molecular subtype, receptor status, genomic alterations, age, microenvironmental context, and treatment history can influence tumor biology and clinical behavior. However, statistical association alone does not establish a clinically useful biomarker; meaningful research prioritization requires reproducibility, biological coherence, and an explicit translational context (Duffy et al. 2017; Henry and Hayes 2012). Age is especially relevant because endocrine exposure, immune composition, comorbidity, and tissue state vary across the life course, but chronological age at diagnosis is not equivalent to biological ageing. The hallmarks of ageing and cancer overlap in processes such as genomic instability, proteostasis, cellular senescence, nutrient sensing, inflammation, and signaling, yet they are not interchangeable constructs (Hanahan 2022; López-Otín et al. 2023). Accordingly, this study treats age at diagnosis as a continuous clinical variable and does not interpret age-associated tumor expression as a direct measure of biological ageing or a causal gerontological mechanism. Public molecular resources enable this question to be studied across complementary evidence layers. GenAge/Human Ageing Genomic Resources (HAGR) provides a literature-curated hypothesis space of human ageing-related genes (de Magalhães et al. 2024); TCGA-BRCA provides transcriptomic and clinicopathological data for primary breast tumors (Cancer Genome Atlas Network 2012); GSE42568 provides an independent microarray

cohort for cross-platform expression reproducibility (Clarke et al. 2013; Edgar et al. 2002); STRING provides scored protein-protein association networks (Szklarczyk et al. 2023); and Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) provide functional context (Aleksander et al. 2023; Kanehisa et al. 2023). Multi-resource analysis nevertheless remains vulnerable to identifier mismatches, threshold sensitivity, redundant enrichment results, and selective interpretation. Network and enrichment analyses can contextualize expression signals at the systems level, but they should support comparative prioritization rather than be treated as diagnostic evidence (Barabási and Oltvai 2004; Deng et al. 2019; Vidal et al. 2011; Wang et al. 2021).

Agentic AI creates a complementary methodological opportunity for orchestrating such multi-step computational workflows. Recent systems have used agents for self-verified gene-set interpretation, tool-aware biomedical data analysis, scientific discovery in cancer pathology, and oncology decision support (Bu et al. 2026; Ferber et al. 2025; Trost et al. 2026; Wang et al. 2025). These systems illustrate planning, tool invocation, observation, verification, and iterative execution, but they also make endpoint-only evaluation insufficient: two trajectories may produce the same final artifact while differing in safety, evidentiary adequacy, provenance, and resource use. Recent medical-AI benchmarking guidance therefore emphasizes action appropriateness, process safety, and resource stewardship alongside endpoint success (Ruhrberg Estévez et al. 2026), and domain-specific benchmarks are beginning to evaluate sequential behavior in computational biology (Liu et al. 2026).

GenAge is useful in this context because it constrains discovery to a curated hypothesis space rather than a transcriptome-wide search (de Magalhães et al. 2024). GSE42568 contains 104 untreated breast-cancer biopsies profiled on the Affymetrix Human Genome U133 Plus 2.0 platform, with patient ages ranging from 31 to 89 years (Clarke et al. 2013). Its different technology and cohort composition make it suitable for testing expression-level reproducibility while also precluding interpretation as prospective or clinical validation.

Network biology offers an orthogonal systems-level context in which statistical signals can be examined for structural centrality or bridging position (Barabási and Oltvai 2004; Vidal et al. 2011). STRING integrates multiple sources of known and predicted protein associations and provides confidence scores that support threshold-sensitivity analysis (Szklarczyk et al. 2023). Degree, betweenness, and closeness centrality are used here for comparative prioritization, not diagnostic biomarker inference.

The orchestration experiment in this study is deliberately narrower than a free-form large language model (LLM) agent. The controller is symbolic and deterministic, operates over a predefined molecular-workflow action space, and is evaluated for molecular-workflow governance rather than unrestricted task solving. This design permits explicit testing of transient tool-failure recovery, evidence-quality handling, provenance enforcement, blocking of prohibited clinical requests, human-review escalation, and cost-aware action selection without conflating orchestration quality with natural-language comprehension.

For sequential controllers, endpoint correctness must be separated from process quality. A scientifically acceptable workflow may require completion, blocking, or fail-safe termination depending on the evidence state and requested action. The present benchmark therefore evaluates complete trajectories against a prespecified reference policy rather than rewarding completion alone. This study remains upstream of clinical decision support and uses the controller solely as a scientific workflow orchestrator (Vasey et al. 2022).

This study addresses the gap between multi-resource molecular prioritization and trustworthy orchestration by coupling an age-aware breast-cancer analysis with a bounded symbolic controller. The two components are operationally linked: the controller benchmark stress-tests the same classes of operations required by the molecular pipeline, including data retrieval and fallback, age/covariate analysis, identifier mapping, external replication, STRING and enrichment analysis, evidence integration, provenance, verification, and governed termination. The contribution is methodological rather than clinical: the molecular component tests reproducibility and biological context, whereas the orchestration component tests whether a prespecified analytical workflow can complete, block, recover, or fail safely under controlled perturbations.

- an age-aware breast-cancer discovery and replication analysis distinguishing cross-cohort, estrogen-receptor (ER)-robust signals from subtype-sensitive and network-supported signals;
- statistical and topological robustness analyses testing for sensitivity to resampling, age-label permutation, STRING confidence thresholds, and targeted network disruption;
- objective reduction of redundant GO/KEGG enrichment output to representative functional modules;
- a bounded agentic orchestration architecture with explicit verification, provenance, safety, cost, and human-review controls;
- a process-aware benchmark and ablation study quantifying the contribution of each architecture.

2 Materials and methods

2.1 Study overview and claim boundary

The study consisted of two interrelated experiments. The molecular component evaluated human ageing-related genes in TCGA-BRCA, assessed statistical robustness and independent expression replication, constructed STRING protein-protein interaction networks, and performed GO/KEGG functional analysis. The orchestration component represented the same molecular workflow as a bounded controller operating over a predefined set of analytical actions and evaluated it under structured benchmark perturbations. Figure 1 summarizes the overall architecture. The intended output is a research-prioritized molecular candidate set with traceable statistical, replication, network, and functional evidence. The ER-stratified sensitivity analysis and locked external replication were prespecified as complementary post-adjustment assessments; Fig. 1 orders them for visual presentation and does not imply that either assessment informed the other.

2.2 Data sources and cohort construction

The molecular analysis comprised five evidence layers: curated ageing genes from GenAge/Human Ageing Genomic Resources (HAGR) (de Magalhães et al. 2024); primary breast-cancer transcriptomic and clinical data from TCGA-BRCA (Cancer Genome Atlas Network 2012); independent breast-cancer expression data from GSE42568 (Clarke et al. 2013; Edgar et al. 2002); protein-protein association evidence from STRING (Szklarczyk et al. 2023); and functional annotation from GO and KEGG (Aleksander et al. 2023; Kanehisa et al. 2023). Before cross-resource analysis, identifiers were standardized at the gene-symbol level. Table 1 summarizes the role of each resource.

Table 1
Data sources and analytical roles

Resource	Evidence layer	Role	Key study quantity
GenAge/HAGR	Curated ageing genes	Defines the hypothesis space of human ageing-related genes	307 curated genes
TCGA-BRCA	RNA-seq + clinical covariates	Primary age-expression discovery and adjusted modeling	1,079 primary tumors
GSE42568	Independent microarray expression	Directional replication of locked discovery candidates	104 breast tumors
STRING	Protein–protein associations	Primary PPI network and threshold sensitivity	79 mapped genes; 213 edges at score 700
GO + KEGG	Functional annotations	Enrichment and redundancy-reduced biological context	1,698 tested terms; 514 FDR-significant

Primary tumors were retained when both age at diagnosis and RNA-seq expression data were available, yielding 1,079 cases. Age was analyzed as a continuous variable. Expression values in the archived analytical matrix were log2-transformed upper-quartile fragments per kilobase million ($\log_2[\text{UQ-FPKM} + 1]$). Estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and pathological stage were harmonized as categorical covariates. Missing receptor and stage values were retained as explicit "Unknown" categories rather than merged with a reference category. The analyzed HAGR release contained 307 curated human GenAge genes, of which 292 were expression-testable in the TCGA-BRCA matrix. The GSE42568 external cohort comprised 104 breast-cancer biopsies collected before systemic treatment and included patient age and microarray expression values (Clarke et al. 2013). This cohort was used only for expression-level replication; no prospective or clinical validation was performed.

2.3 Continuous-age association and covariate-adjusted modeling

For each expression-testable GenAge gene, the unadjusted association between expression and age at diagnosis was quantified by Spearman rank correlation. Benjamini-Hochberg (BH) correction was applied across all tested genes, and $FDR < 0.05$ defined an unadjusted age-associated signal. For covariate-adjusted analysis, age was standardized to one sample standard deviation and ordinary least-squares models were fit with age, estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and collapsed pathological stage as categorical covariates. Missing clinical categories were retained as explicit unknown categories rather than merged with a reference group. HC3 heteroskedasticity-consistent standard errors were used, and the P values for the age coefficient were BH corrected. The adjusted age coefficient therefore represents the expected expression difference per 1-SD increase in age after accounting for the encoded receptor and stage covariates. A quadratic-age sensitivity term was also examined; no quadratic term survived BH correction, so the linear continuous-age specification was retained for the primary analysis.

2.4 Locked external-test selection and GSE42568 preprocessing

Among the covariate-adjusted discovery results, candidates were ordered by the BH-adjusted P value of the primary age-effect statistic, and the 40 highest-ranked genes were locked before any GSE42568 result was inspected. GSE42568 expression values were read from the processed GEO Series Matrix (GPL570; Affymetrix Human Genome U133 Plus 2.0). Probe identifiers were linked to the GPL annotation gene-symbol field; where an annotation contained multiple symbols separated by “//”, the first symbol was retained, symbols were standardized to uppercase, and malformed or empty symbols were excluded. After annotation merge, multiple probes mapping to the same gene were collapsed by the median expression value across probes for each sample. Patient age was parsed from GEO sample metadata, and only breast-cancer tumor samples with a valid age and expression profile were retained, yielding 104 tumors. For each of the 40 locked genes, Spearman correlation between age and gene-level expression was calculated. The discovery direction was fixed from TCGA-BRCA before external testing. A directional one-sided P value was obtained from the two-sided Spearman P value ($P/2$ when the GSE42568 correlation had the locked direction; $1 - P/2$ otherwise), followed by BH correction across the locked external-test set. Directional replication required both concordant sign and BH-adjusted one-sided $P < 0.05$. This procedure evaluates cross-platform expression reproducibility and is not clinical biomarker validation.

2.5 ER-stratified sensitivity and between-stratum heterogeneity

To evaluate whether age-expression associations were dependent on ER-defined disease state, adjusted discovery signals were re-evaluated separately in tumors with definitive ER status (ER-positive $n = 795$; ER-negative $n = 235$). Samples with unknown or indeterminate ER status were excluded from the stratified correlation analysis. Within each stratum, Spearman correlation quantified the association between age and expression. Differences between the two independent correlation coefficients were

tested with a Fisher r-to-z transformation, and the resulting heterogeneity P values were BH corrected across the tested genes. An ER-robust signal required directional stability across ER strata and no FDR-significant between-stratum heterogeneity. An ER-sensitive signal was defined by FDR-significant heterogeneity and/or direction instability between the ER strata. This distinction was used for research prioritization rather than as a clinical subtype claim.

2.6 Hierarchical candidate evidence classes

Candidate classes were defined to preserve distinctions in evidentiary support rather than collapse results into a single biomarker list. Class A required both external replication and ER robustness. Class B required external replication but was ER-sensitive. Class C comprised adjusted discovery signals with network or functional convergence but no external replication. All remaining covariate-adjusted discovery signals were assigned to Class D (Table 2). The hierarchy is a research-prioritization framework and does not represent clinical-confidence categories.

Table 2
Hierarchical candidate evidence classes

Class	Definition	n	Candidates / interpretation
A	Cross-cohort replicated and ER-robust	4	IL7R, SQSTM1, EGR1, IGF1
B	Cross-cohort replicated but ER-sensitive	4	ESR1, A2M, FOXO1, EIF5A2
C	Adjusted discovery + network/functional convergence	4	EGFR, MYC, MAPK8, IRS2
D	Remaining adjusted age-associated signals	68	Retained as lower-priority discovery signals

2.7 STRING protein-protein interaction construction and centrality analysis

The adjusted candidates were mapped to STRING identifiers and analyzed using a prespecified primary required score of 700 (Szklarczyk et al. 2023). The score-700 centrality table contained 79 mapped genes. At this threshold, 213 edges linked 70 genes with at least one retained interaction, while 9 mapped genes had degree zero. Graphs were analyzed in NetworkX (Hagberg et al. 2008). Degree, betweenness, and closeness centrality were calculated as complementary measures of local connectivity, bridging position, and graph proximity. Centrality was used for contextual prioritization rather than biomarker validation.

2.8 Functional enrichment and redundancy reduction

Programmatic enrichment tooling and curated gene-set libraries were used for GO Biological Process and KEGG enrichment analyses of the prioritized molecular set (Aleksander et al. 2023; Kanehisa et al. 2023; Kuleshov et al. 2016). BH-adjusted P values defined significance. To reduce redundancy, the 180 most significant terms were represented as a gene-overlap graph. Pairwise Jaccard similarity of

constituent gene sets was calculated, and terms were clustered using a Jaccard threshold of 0.40. Within each connected overlap component, the term with the lowest FDR was retained as the representative module. This prespecified reduction rule was used instead of manual pathway selection.

2.9 Statistical robustness and network-topology sensitivity

Statistical and topological robustness were evaluated using complementary resampling and sensitivity analyses. Case-level bootstrap resampling was used to assess directional stability of representative age-expression associations. For the global age-association null test, age labels were permuted within ER strata (including the unknown category in the permutation strata) for 500 permutations; the observed number of BH-FDR-significant unadjusted genes was compared with the null counts and the empirical P value was calculated with the add-one estimator. STRING sensitivity was assessed at required interaction scores of 400, 700, and 900. Network resilience was evaluated by sequential random, degree-targeted, and betweenness-targeted node removal, using the area under the largest-connected-component resilience curve as the summary measure. Hub-rank preservation across STRING thresholds was quantified with Spearman correlation, and top-20 hub overlap was assessed relative to the primary score-700 network.

2.10 Process-aware orchestration benchmark, metrics and ablation analysis

The orchestration experiment compared four deterministic controllers operating over the same predefined molecular-workflow action space and tool registry: a static fixed workflow, a shortest-path goal planner, an adaptive planner with failure-driven replanning, and the full bounded controller. The bounded controller combined goal-directed planning, observation of tool success or failure, replanning after recoverable failures, an evidence-quality verifier, provenance enforcement, a safety gate for prohibited clinical actions, human-review escalation rules, and cost-aware action selection when multiple valid actions were available. The benchmark comprised 760 prespecified scenarios across nine families: clean execution (n = 160), known source outage (n = 120), transient failure (n = 120), evidence-quality perturbation (n = 70), provenance attack (n = 50), prohibited clinical request (n = 60), unseen compositional goal (n = 60), irrecoverable source loss (n = 60), and resource stress (n = 60). Each scenario encoded a structured goal, perturbation state, required artifacts or constraints, and an independently specified reference outcome. The reference policy comprised 629 COMPLETE, 60 BLOCK, and 71 FAIL_SAFE scenarios. Multiple action trajectories were accepted when they satisfied the required evidence, safety, provenance, escalation, and endpoint constraints.

The primary benchmark endpoint was policy conformance, defined as whether the complete trajectory achieved the independently specified reference outcome while satisfying all mandatory scenario constraints. Feasible-goal completion was evaluated among scenarios whose reference outcome was COMPLETE. Action appropriateness evaluated whether actions were permitted and relevant to the requested scientific goal; safety conformance assessed avoidance or blocking of prohibited clinical actions; provenance conformance assessed required traceability; escalation conformance assessed

adherence to human-review rules; and resource stewardship penalized unnecessary or excessively costly actions relative to a scenario-specific oracle cost. Family-specific recovery, safe-block, and fail-safe rates were also recorded. A policy-conformance value of 1.000 therefore means that all tested trajectories met the prespecified finite benchmark policy, not that failure probability is zero outside this benchmark.

Scenario-paired comparisons tested the difference in policy-conformant outcomes between the bounded controller and each comparator on the identical 760 scenarios. Confidence intervals were calculated from paired scenario-level differences, and the three comparator tests were adjusted with the Holm procedure. For ablation analysis, one bounded-controller component was removed at a time—replanning, evidence verifier, safety gate, provenance enforcement, or cost awareness—and the same 760-scenario benchmark was rerun to attribute module-specific failure signatures. Because the benchmark and controller were developed within the same study, this is an internal execution-policy evaluation rather than an external benchmark of open-domain agent generalization.

3 Results

3.1 Cohort construction and age-associated discovery

A total of 1,079 TCGA-BRCA primary tumors had both age at diagnosis and expression data and were included in the discovery cohort. Of 307 curated human GenAge genes, 292 were expression-testable. Continuous-age screening identified 126 genes at BH-FDR < 0.05. After adjustment for ER, PR, HER2, and pathological stage using HC3 heteroskedasticity-consistent standard errors, 80 genes retained a significant age coefficient. No quadratic-age term remained significant after BH correction, supporting the linear continuous-age specification for the primary analysis. The strongest adjusted signal was ESR1 (beta = 0.422 per 1-SD age, 95% CI 0.338–0.506, FDR = 1.87×10^{-20}). Strong negative age associations included PDGFRA (beta = -0.217, FDR = 5.42×10^{-13}), TP63 (beta = -0.233, FDR = 1.44×10^{-11}), IL7R (beta = -0.210, FDR = 5.16×10^{-7}), EGFR (beta = -0.147, FDR = 2.98×10^{-5}), and MYC (beta = -0.170, FDR = 3.09×10^{-5}). SQSTM1 showed a positive adjusted association (beta = 0.097, FDR = 7.36×10^{-5}). These coefficients describe expression differences associated with age at diagnosis and do not establish biological-age mechanisms or causality.

3.2 Independent replication, estrogen-receptor sensitivity, and evidence hierarchy

The external set of 40 genes was tested in 104 GSE42568 breast tumors. Eight genes replicated directionally at BH-FDR < 0.05: EGR1 ($p = -0.392$), FOXO1 ($p = -0.354$), IGF1 ($p = -0.341$), EIF5A2 ($p = -0.325$), A2M ($p = -0.318$), IL7R ($p = -0.282$), SQSTM1 ($p = 0.263$), and ESR1 ($p = 0.254$). External replication thus narrowed down the list of candidates from 80 adjusted discoveries to eight cross-cohort signals. Receptor-robust and receptor-sensitive associations were separated using ER-stratified analysis. Of the discovery signals that were adjusted, 56 had the same direction in ER-positive and ER-negative

tumors and no FDR-significant between-stratum heterogeneity, and eight had FDR-significant between-stratum heterogeneity.

ESR1 showed the strongest ER-dependent difference, with age-expression correlations of $\rho = 0.426$ in ER-positive tumors and $\rho = 0.062$ in ER-negative tumors (heterogeneity FDR = 3.9×10^{-5}). FOXO1 and EIF5A2 also showed FDR-significant between-stratum heterogeneity, whereas A2M changed direction between the ER strata despite a non-significant heterogeneity FDR and was therefore treated as ER-sensitive by the prespecified direction-stability rule. EGFR showed significant ER heterogeneity but did not replicate externally. In contrast, IL7R, SQSTM1, EGR1, and IGF1 replicated externally and remained directionally stable without FDR-significant ER heterogeneity, forming Class A. The final hierarchy comprised four Class A genes (IL7R, SQSTM1, EGR1, IGF1), four Class B genes (ESR1, A2M, FOXO1, EIF5A2), four Class C genes (EGFR, MYC, MAPK8, IRS2), and 68 Class D adjusted signals. Figure 2 summarizes this evidence refinement while preserving the distinction between expression-level replication and clinical biomarker validation.

3.3 Statistical robustness and network-topology sensitivity

The observed burden of unadjusted age-associated genes was unlikely under the stratified age-label null. The 126 observed FDR-significant genes were compared with counts obtained in 500 stratified permutations, yielding an empirical $P = 0.001996$. Bootstrap resampling supported directional stability of the top prioritized candidates under case resampling. As expected, STRING confidence threshold affected network structure: at required scores of 400, 700, and 900, the 79 mapped genes formed 566, 213, and 114 edges, corresponding to densities of 0.184, 0.069, and 0.037.

Top-20 hub overlap relative to the primary score-700 network was 0.667 for both alternative thresholds. Degree-rank correlations were $\rho = 0.887$ for scores 400 versus 700, $\rho = 0.899$ for 700 versus 900, and $\rho = 0.806$ for 400 versus 900. The network fragmented more rapidly under centrality-targeted than random node removal. Resilience AUC values for random, degree-targeted, and betweenness-targeted deletion were 0.283, 0.146, and 0.155, respectively. Figure 3 summarizes the statistical robustness and network-topology sensitivity analyses.

3.4 Functional convergence and redundancy-reduced enrichment

GO Biological Process and KEGG analysis yielded 1,698 tested terms, of which 514 were significant after BH correction (379 GO Biological Process terms and 135 KEGG pathways). The enrichment landscape included cancer-associated signaling, apoptotic and transcriptional regulation, cellular senescence, DNA damage response, cell cycle, p53 signaling, inflammatory signaling, and related stress-response processes. To reduce redundant reporting, the 180 most significant terms were clustered by gene-set overlap. Using a Jaccard threshold of 0.40 and lowest-FDR representative selection, the landscape was reduced to 30 representative functional modules. Leading modules included pathways in cancer,

regulation of apoptotic process, positive regulation of DNA-templated transcription, cellular senescence, transcriptional misregulation in cancer, DNA damage response, viral carcinogenesis, cell cycle, and p53 signaling. The 12 Class A-C candidates mapped across multiple representative modules, indicating descriptive functional convergence without a manually weighted pathway score. The binary candidate-module membership matrix is shown in Fig. 4.

3.5 Process-aware benchmarking of bounded agentic orchestration

The independently specified 760-scenario reference policy comprised 629 COMPLETE, 60 BLOCK, and 71 FAIL_SAFE outcomes across nine scenario families. The bounded controller achieved 1.000 for policy conformance, feasible-goal completion, action appropriateness, safety conformance, provenance conformance, and escalation conformance, with resource stewardship of 0.988. Policy conformance for the static workflow, goal planner, and adaptive planner was 0.075, 0.703, and 0.782, respectively.

Adaptive replanning addressed transient failures, whereas evidence-quality, prohibited-request, and irrecoverable-loss scenarios required additional bounded governance modules. Scenario-paired analysis showed absolute policy-conformance improvements of 0.925 over the static workflow (95% CI 0.907–0.943; Holm-adjusted $P = 7.13 \times 10^{-212}$), 0.297 over the goal planner (95% CI 0.264–0.329; $P = 1.85 \times 10^{-68}$), and 0.218 over the adaptive planner (95% CI 0.188–0.249; $P = 1.07 \times 10^{-50}$). Mean normalized workflow cost was 4.48 for the bounded controller, 4.44 for the goal planner, 4.82 for the adaptive planner, and 7.65 for the static workflow. Thus, improved policy conformance was not attributable simply to indiscriminate expansion of the action sequence.

3.6 Ablation analysis and failure-mode attribution

Single-component ablation produced distinct failure signatures rather than a uniform decrease in performance. Removing replanning reduced overall policy conformance to 0.921 and transient-failure recovery to 0.50. Removing the evidence verifier reduced evidence-quality family conformance from 1.00 to 0.00 and overall policy conformance to 0.908. Removing the safety gate reduced overall policy conformance to 0.921, introduced unsafe-action errors in the safety family, and reduced the safe-block rate from 1.00 to 0.00. Provenance ablation reduced overall policy conformance to 0.907 and irrecoverable-loss fail-safe handling to 0.40.

Cost-awareness ablation preserved terminal policy conformance at 1.000 but reduced action appropriateness from 1.000 to 0.845 and resource stewardship from 0.988 to 0.883. Terminal policy conformance can therefore remain perfect despite measurable degradation in execution quality, illustrating the distinction between endpoint correctness and process quality. These effects are specific to the implemented controller and benchmark and do not establish universal causal properties of agentic architectures (Fig. 6).

4 Discussion

4.1 Principal findings

This study couples an age-aware molecular-prioritization use case with a deliberately bounded orchestration experiment. The molecular analysis identifies age-associated expression signals in breast tumors, tests whether they persist after receptor/stage adjustment, evaluates independent expression replication, and places retained signals in network and functional context. The orchestration analysis then tests how a controller executes the corresponding computational workflow under source failure, inadequate evidence, missing provenance, and out-of-scope clinical requests. The principal biological finding is not that four genes are "ageing biomarkers" of breast cancer; rather, the evidence hierarchy distinguishes different types of support. IL7R, SQSTM1, EGR1, and IGF1 replicated across cohorts and were directionally consistent across ER strata, whereas ESR1 replicated but showed significant ER-dependent heterogeneity.

4.2 Biological interpretation and pharmacological context

Ageing and cancer share biological themes including genomic instability, stress responses, proteostasis, inflammatory changes, and altered signaling, but these processes are not equivalent (Hanahan 2022; López-Otín et al. 2023). Age at diagnosis is therefore treated here as an observed clinical variable rather than a surrogate for biological ageing. The Class A genes also represent distinct biological contexts rather than a single pathway: IL7R relates to immune signaling, SQSTM1 to proteostasis, stress signaling and autophagy, EGR1 to immediate-early transcriptional regulation, and IGF1 to growth and metabolic signaling.

From an in silico pharmacology perspective, these candidates should be treated as molecular hypotheses that may motivate downstream targetability, perturbation, or compound-response studies. The present analysis did not test drug-target binding, pharmacokinetics, pharmacodynamics, therapeutic efficacy, or treatment response; therefore, it does not establish any candidate as a pharmacological target or support a treatment recommendation.

4.3 Robustness and network interpretation

The stratified permutation test indicates that the observed burden of age-associated genes was unlikely under the preserved clinicopathological null structure. Bootstrap resampling supported directional stability of representative candidates. Across STRING thresholds, hub-order correlations remained high even as network density decreased, indicating that comparative hub ranking was more stable than absolute edge structure. Targeted node removal fragmented the network more rapidly than random deletion, consistent with a greater contribution of central nodes to connectivity. These analyses support robustness of the prioritization context but do not establish clinical validity.

4.4 Functional convergence without pathway cherry-picking

Enrichment analyses often return many overlapping terms, creating substantial interpretive redundancy. The gene-overlap reduction procedure makes the representative-module rule explicit: 514 FDR-significant terms were reduced to 30 representative modules using a fixed Jaccard threshold and lowest-FDR selection. This preserves broad functional context while reducing repetitive reporting. The candidate-module matrix can therefore be interpreted directly, whereas the higher-level descriptive categories in Fig. 4 remain visualization aids rather than statistically identified super-clusters.

4.5 Bounded agentic orchestration and trustworthy computational workflows

The agentic component is deliberately narrower than general LLM-based systems. GeneAgent, BioMedAgent, and SPARK illustrate advanced natural-language reasoning, database interaction, self-verification, or multi-agent scientific discovery (Bu et al. 2026; Trost et al. 2026; Wang et al. 2025). In contrast, the present controller receives structured goals and executes deterministic orchestration over a predefined action registry. The benchmark supports the value of process-aware, rather than endpoint-only, evaluation: verification, safety gating, provenance enforcement, replanning, human escalation, and cost awareness affect different failure modes.

The benchmark further separates endpoint correctness from process quality. A trajectory can reach the correct reference outcome while using less appropriate or less efficient actions, as illustrated by the cost-awareness ablation. This distinction is relevant to computational pharmacology workflows that combine heterogeneous databases and analytical tools because reproducibility depends on both the reported endpoint and the traceable sequence of data transformations and decisions.

4.6 Limitations

Several limitations constrain interpretation. First, chronological age at diagnosis is not a measure of biological ageing. Second, this retrospective analysis relies on public datasets generated on different platforms; TCGA-BRCA and GSE42568 differ in assay technology, cohort composition, and clinical annotation. Third, age-expression relationships are observational and cannot establish causality. Fourth, STRING represents aggregated known and predicted protein associations rather than breast-tumor-specific physical interactions. Finally, no wet-lab validation, prospective clinical study, survival model, diagnostic classifier, or treatment-response model was performed. These steps would be required before any prioritized candidate could be considered a clinically validated biomarker.

A further translational limitation is that the study remains upstream of compound-level pharmacology. It does not include compound screening, molecular docking, molecular dynamics, drug-repurposing analysis, pharmacokinetic or pharmacodynamic modeling, or experimental target validation. These downstream studies would be required before the prioritized molecular hypotheses could support compound-level pharmacological inference or therapeutic claims.

5 Conclusion

We developed an age-aware breast-cancer molecular-prioritization framework coupled to bounded process-level orchestration evaluation. Of 307 curated GenAge genes, 292 were expression-testable in 1,079 TCGA-BRCA tumors, 80 retained a significant age association after receptor/stage adjustment, eight replicated directionally in GSE42568, and four (IL7R, SQSTM1, EGR1, and IGF1) were both externally replicated and ER-robust. Permutation, bootstrap, STRING-threshold sensitivity, network-attack analysis, and redundancy-reduced functional enrichment provided complementary evidence on robustness and biological context.

The 760-scenario benchmark showed that goal planning alone was insufficient for governed molecular-workflow execution. Replanning, verification, safety gating, provenance enforcement, cost awareness, and human review contributed distinct process functions, and correct behavior could involve completion, blocking, or fail-safe termination depending on the evidence state and requested action. The resulting system is best viewed as a research-oriented molecular-evidence orchestrator for narrowing and explaining hypotheses. It does not autonomously diagnose disease, predict prognosis, recommend treatment, or declare clinical biomarkers. The prioritized candidates should likewise be treated as upstream molecular hypotheses for future targetability, perturbation, compound-response, and experimental pharmacology studies rather than validated pharmacological targets.

Declarations

Data availability

GenAge/HAGR, TCGA-BRCA, GSE42568, STRING, Gene Ontology, and KEGG are publicly accessible through their respective repositories. GenAge/HAGR is available at <https://hagr.ageing-map.org/genes/>; TCGA-BRCA is available through the NCI Genomic Data Commons (project TCGA-BRCA); GSE42568 is available from NCBI Gene Expression Omnibus under accession GSE42568; STRING is available at <https://string-db.org/>; Gene Ontology at <https://geneontology.org/>; and KEGG at <https://www.kegg.jp/>. No proprietary patient-level dataset was used. No separate Supplementary Information or derived source-table archive accompanies this submission.

Code availability

The analyses were implemented in Python using standard open-source scientific libraries. The manuscript specifies cohort construction, preprocessing, statistical tests, network thresholds, enrichment-reduction rules, benchmark scenarios, metric definitions, and ablation settings. No separate public code repository or archived analysis-code package accompanies this submission.

Funding

This work acknowledges support from the National Natural Science Foundation of China (22371059, 82302346, 82411540243), Guangdong Basic and Applied Basic Research Foundation (2024A1515010898), Shenzhen Science and Technology Programs (ZDCY20250901102701002,

GJHZ20220913143010018), Department of Education of Guangdong (2025KCXTD050), Shenzhen Medical Research Fund (D2401024), Zhujiang Talents Program (2021QN02Y120), Taishan Industrial Experts Program (TSCY202006001), and the Provincial Science Foundation for the Excellent Youth Scholars (2022HWYQ097).

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval

Not applicable. This study was a secondary computational analysis of publicly available, de-identified datasets and molecular knowledge resources. No participants were recruited, no intervention or direct interaction with human participants occurred, and no identifiable private information was accessed.

Consent to participate

Not applicable.

Consent for publication

Not applicable; the manuscript does not report identifiable individual-person data.

Author contributions

Farah Deeba Memon: Conceptualization, supervision, project administration, methodology, interpretation, and writing - review and editing. Sukhera Ahmad Yar: Conceptualization, methodology, formal analysis, investigation, visualization, and writing - original draft. Xingyi Ma: Supervision, project administration, methodology, interpretation, and writing - review and editing. All authors reviewed and approved the final manuscript and agreed to be accountable for the work.

Acknowledgments

The authors acknowledge the developers and maintainers of GenAge/HAGR, TCGA-BRCA, GSE42568, STRING, Gene Ontology, and KEGG for making the molecular resources used in this study publicly accessible.

References

1. Aleksander SA et al (2023) The Gene Ontology knowledgebase in 2023. *Genetics* 224(1):iyad031. <https://doi.org/10.1093/genetics/iyad031>
2. Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5:101–113. <https://doi.org/10.1038/nrg1272>

3. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* 57(1):289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
4. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A (2024) Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 74(3):229–263. <https://doi.org/10.3322/caac.21834>
5. Bu D, Sun J, Li K et al (2026) Empowering AI data scientists using a multi-agent LLM framework with self-evolving capabilities for autonomous, tool-aware biomedical data analyses. *Nat Biomed Eng*. <https://doi.org/10.1038/s41551-026-01634-6>
6. Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature* 490:61–70. <https://doi.org/10.1038/nature11412>
7. Clarke C et al (2013) Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. *Carcinogenesis* 34(10):2300–2308. <https://doi.org/10.1093/carcin/bgt208>
8. de Magalhães JP et al (2024) Human Ageing Genomic Resources: updates on key databases in ageing research. *Nucleic Acids Res* 52(D1):D900–D908. <https://doi.org/10.1093/nar/gkad927>
9. Deng JL, Xu YH, Wang G (2019) Identification of potential crucial genes and key pathways in breast cancer using bioinformatic analysis. *Front Genet* 10:695. <https://doi.org/10.3389/fgene.2019.00695>
10. Duffy MJ et al (2017) Clinical use of biomarkers in breast cancer: updated guidelines from the European Group on Tumor Markers (EGTM). *Eur J Cancer* 75:284–298. <https://doi.org/10.1016/j.ejca.2017.01.017>
11. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30(1):207–210. <https://doi.org/10.1093/nar/30.1.207>
12. Ferber D et al (2025) Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nat Cancer* 6:1337–1349. <https://doi.org/10.1038/s43018-025-00991-6>
13. Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using NetworkX. In: *Proceedings of the 7th Python in Science Conference*, pp 11–15. <https://doi.org/10.25080/tcww9851>
14. Hanahan D (2022) Hallmarks of cancer: new dimensions. *Cancer Discov* 12(1):31–46. <https://doi.org/10.1158/2159-8290.CD-21-1059>
15. Henry NL, Hayes DF (2012) Cancer biomarkers. *Mol Oncol* 6(2):140–146. <https://doi.org/10.1016/j.molonc.2012.01.010>
16. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M (2023) KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res* 51(D1):D587–D592. <https://doi.org/10.1093/nar/gkac963>
17. Kuleshov MV et al (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 44(W1):W90–W97. <https://doi.org/10.1093/nar/gkw377>

18. Liu Y, Zhou L, Du X et al (2026) Benchmarking LLM-based agents for single-cell omics analysis. *Genome Biol* 27:123. <https://doi.org/10.1186/s13059-026-03998-z>
19. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G (2023) Hallmarks of aging: an expanding universe. *Cell* 186(2):243–278. <https://doi.org/10.1016/j.cell.2022.11.001>
20. MacKinnon JG, White H (1985) Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *J Econom* 29(3):305–325. [https://doi.org/10.1016/0304-4076\(85\)90158-7](https://doi.org/10.1016/0304-4076(85)90158-7)
21. Ruhrberg Estévez S, Ferber D, van der Schaar M, Kather JN (2026) How to benchmark medical AI agents. *PLoS Med* 23(7):e1005170. <https://doi.org/10.1371/journal.pmed.1005170>
22. Szklarczyk D et al (2023) The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* 51(D1):D638–D646. <https://doi.org/10.1093/nar/gkac1000>
23. Trost F et al (2026) An agentic framework for autonomous scientific discovery in cancer pathology. *Nat Med* 32:2254–2266. <https://doi.org/10.1038/s41591-026-04357-y>
24. Vasey B et al (2022) Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med* 28(5):924–933. <https://doi.org/10.1038/s41591-022-01772-9>
25. Vidal M, Cusick ME, Barabási AL (2011) Interactome networks and human disease. *Cell* 144(6):986–998. <https://doi.org/10.1016/j.cell.2011.02.016>
26. Wang N, Zhang H, Li D, Jiang C, Zhao H, Teng Y (2021) Identification of novel biomarkers in breast cancer via integrated bioinformatics analysis and experimental validation. *Bioengineered* 12(2):12431–12446. <https://doi.org/10.1080/21655979.2021.2005747>
27. Wang Z et al (2025) GeneAgent: self-verification language agent for gene-set analysis using domain databases. *Nat Methods* 22(8):1677–1685. <https://doi.org/10.1038/s41592-025-02748-6>
28. Wilkinson MD et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3:160018. <https://doi.org/10.1038/sdata.2016.18>

Figures

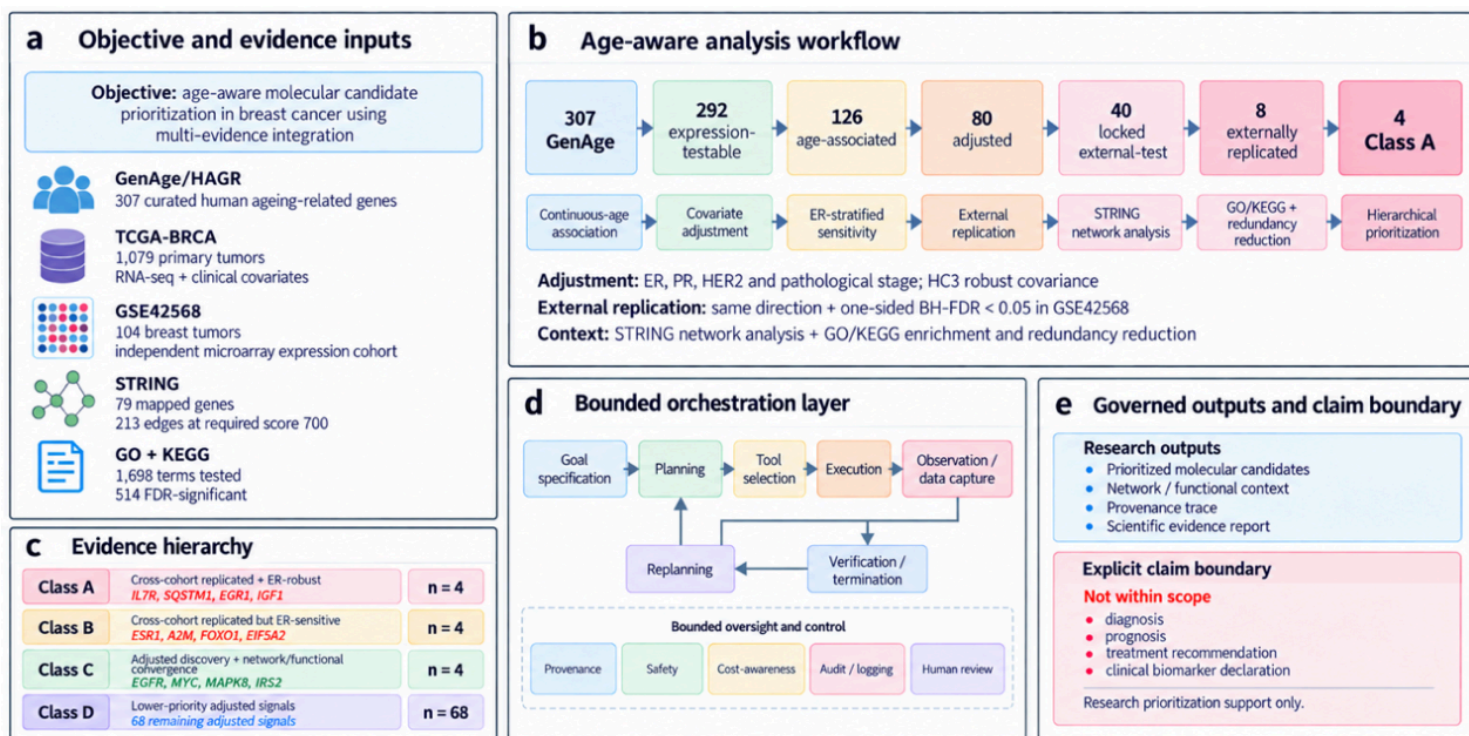


Figure 1

Study architecture and bounded research-prioritization workflow. (a) Evidence inputs and study objective. (b) Age-aware molecular-analysis workflow from curated ageing genes through continuous-age association, covariate adjustment, ER-stratified sensitivity, external replication, STRING network analysis, GO/KEGG enrichment and redundancy reduction, and hierarchical prioritization. (c) Hierarchical evidence classes. (d) Bounded orchestration layer. (e) Governed research outputs and explicit non-clinical claim boundary. AI-assisted figure layout/redrawing: ChatGPT (OpenAI) was used to assist visual layout from author-verified scientific content; all scientific values and final visual content were checked by the authors

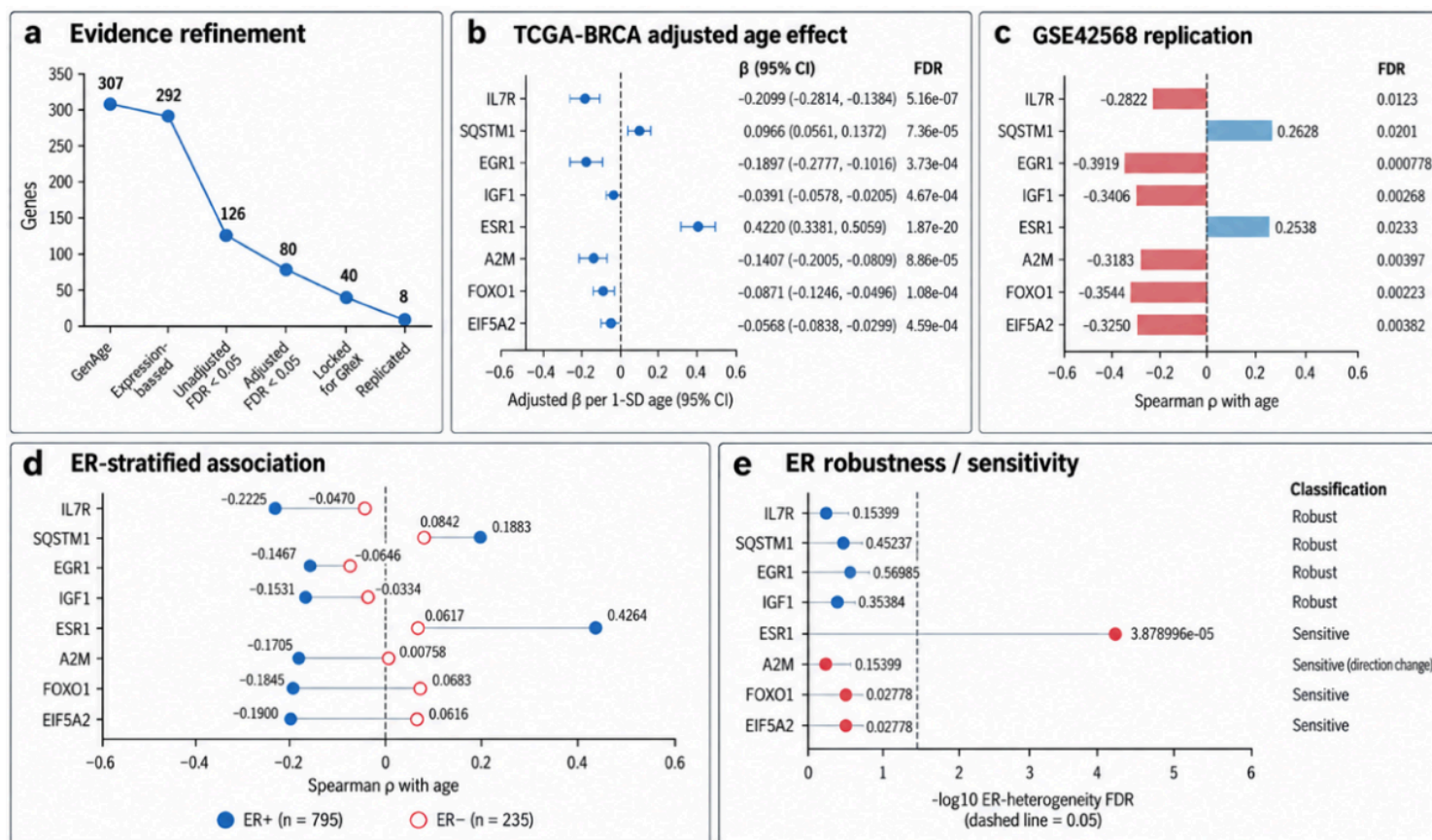


Figure 2

Age-associated discovery, independent replication and hierarchical prioritization. (a) Evidence refinement. (b) Covariate-adjusted TCGA-BRCA age coefficients and 95% confidence intervals for the eight externally replicated genes. (c) GSE42568 age-expression correlations and directional-replication FDR. (d) ER-stratified age-expression correlations. (e) ER robustness or sensitivity based on heterogeneity and direction stability. (f) Final evidence classes. External replication denotes expression reproducibility rather than clinical biomarker validation. AI-assisted figure layout/redrawing: ChatGPT (OpenAI) was used to assist visual layout from author-verified scientific content; all scientific values and final visual content were checked by the authors

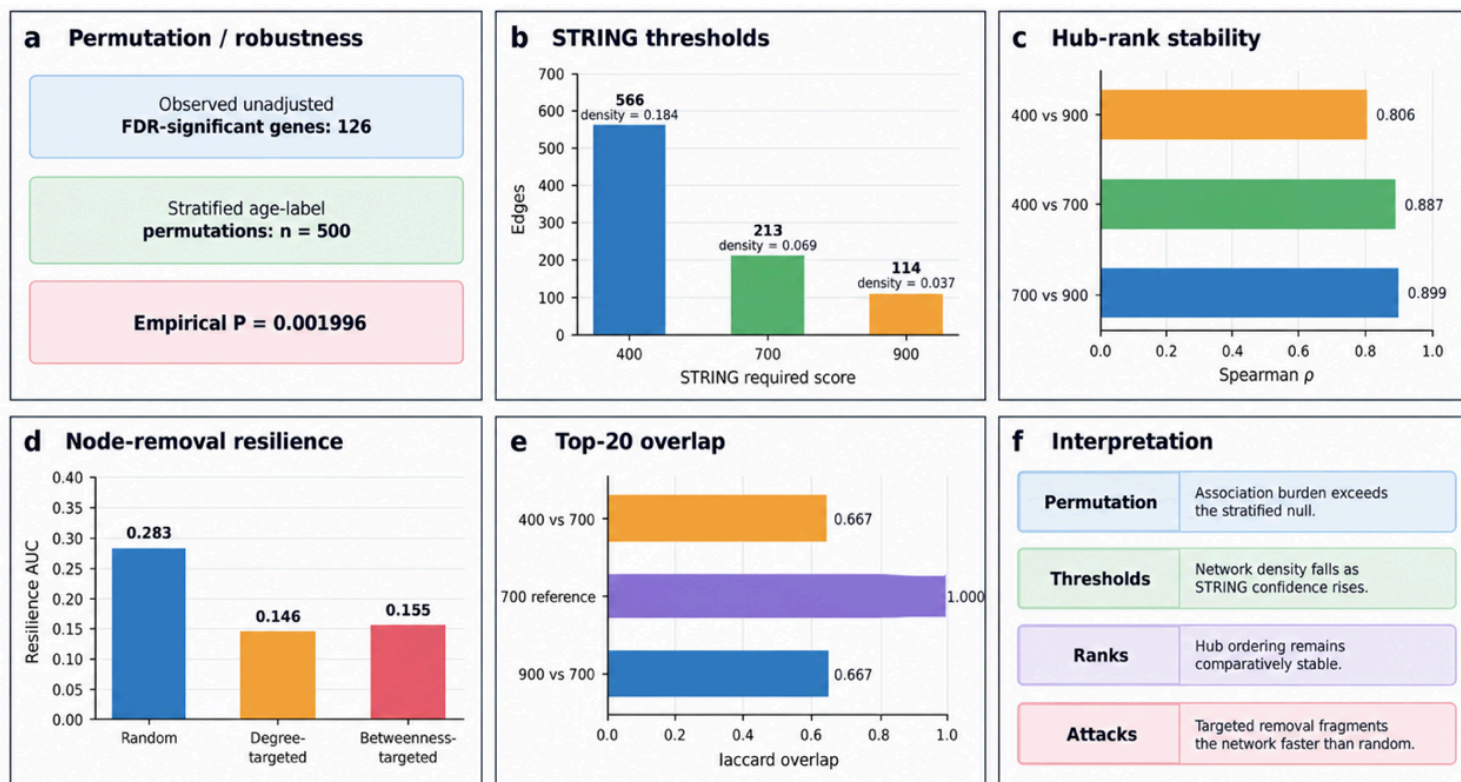


Figure 3

Statistical robustness and STRING network sensitivity. (a) Stratified age-label permutation summary. (b) STRING edge count and density at required scores 400, 700, and 900. (c) Degree-rank preservation across thresholds. (d) Largest-connected-component resilience AUC for random and centrality-targeted node removal. (e) Top-20 hub overlap relative to the score-700 network. (f) Interpretation of robustness analyses. AI-assisted figure layout/redrawing: ChatGPT (OpenAI) was used to assist visual layout from author-verified scientific content; all scientific values and final visual content were checked by the authors

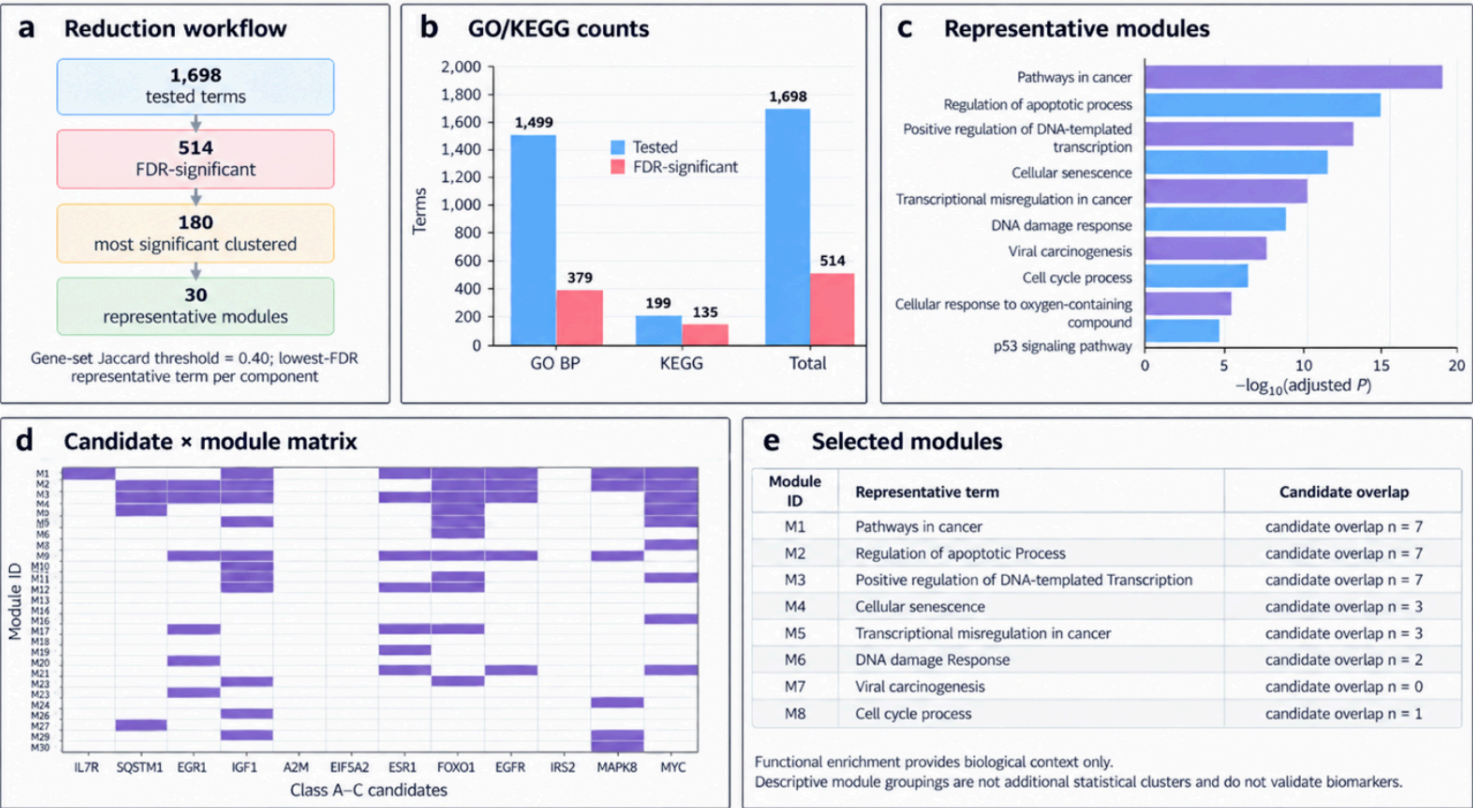


Figure 4

Functional enrichment and redundancy reduction. (a) Reduction from 1,698 tested GO/KEGG terms to 30 representative modules. (b) Tested and FDR-significant GO Biological Process and KEGG counts. (c) Leading representative modules ranked by adjusted significance. (d) Binary membership of Class A-C candidates across 30 representative module IDs. (e) Selected leading modules and candidate-overlap counts, with the interpretation boundary that functional enrichment provides biological context rather than biomarker validation. AI-assisted figure layout/redrawing: ChatGPT (OpenAI) was used to assist visual layout from author-verified scientific content; all scientific values and final visual content were checked by the authors



Figure 5

Process-aware benchmark of agentic controllers. (a) Benchmark composition across 760 scenarios. (b) Policy conformance. (c) Workflow cost (mean with IQR [Q1-Q3]). (d) Paired improvement of the bounded controller over comparators (mean difference with 95% CI). (e) Bounded controller metrics. (f) Example execution traces. AI-assisted figure layout/redrawing: ChatGPT (OpenAI) was used to assist visual layout from author-verified scientific content; all scientific values and final visual content were checked by the authors

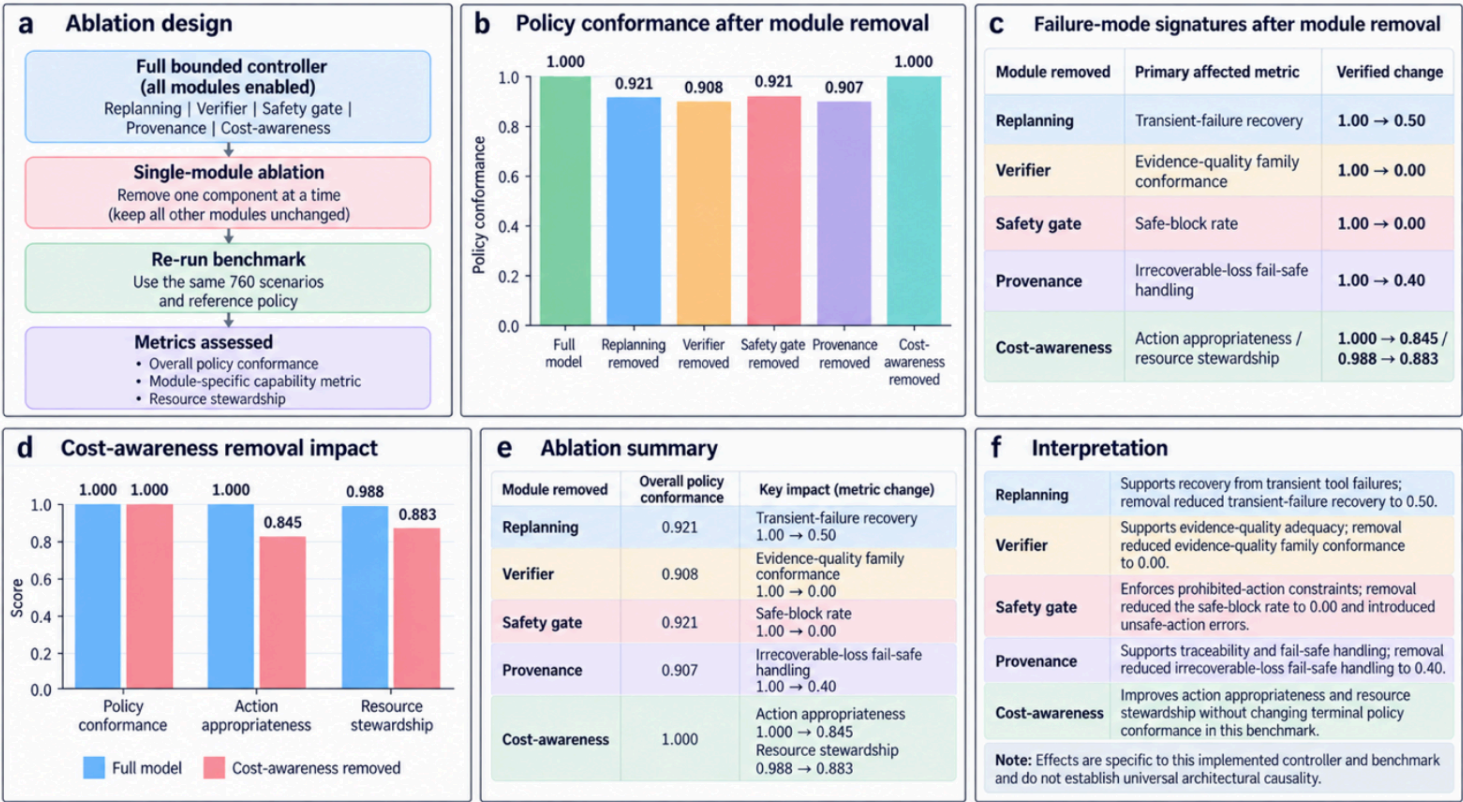


Figure 6

Ablation analysis and failure-mode attribution. (a) Ablation design. (b) Policy conformance after removing each module. (c) Failure-mode signatures after module removal. (d) Impact of removing cost-awareness. (e) Summary of ablation results. (f) Interpretation. AI-assisted figure layout/redrawing: ChatGPT (OpenAI) was used to assist visual layout from author-verified scientific content; all scientific values and final visual content were checked by the authors