Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Icariin-Informed Nine-Gene Transcriptomic Signature Identifies Mesenchymal–Stromal Recurrence Risk in Resected Colorectal Cancer: External Validation and Single-Cell Localization

Fei Zuo
Changzhou Hospital of Traditional Chinese Medicine, Nanjing University of Chinese Medicine
Jiapan Yan
Changzhou Hospital of Traditional Chinese Medicine, Nanjing University of Chinese Medicine
Mingni Jiang
Changzhou Hospital of Traditional Chinese Medicine, Nanjing University of Chinese Medicine
Tianping Luo
927884643@qq.com
Changzhou Hospital of Traditional Chinese Medicine, Nanjing University of Chinese Medicine

Research Article

Keywords: colorectal cancer, icariin, prognosis, machine learning, consensus molecular subtype, external validation

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11109967/v1

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

Additional Declarations: No competing interests reported.

An Icariin-Informed Nine-Gene Transcriptomic Signature Identifies Mesenchymal–Stromal Recurrence Risk in Resected Colorectal Cancer: External Validation and Single-Cell Localization

Fei Zuo[1], Jiapan Yan[1], Mingni Jiang[2], Tianping Luo[1,*]
[1]Department of General Surgery, Changzhou Hospital of Traditional Chinese Medicine, Nanjing University of Chinese Medicine, Changzhou 213000, Jiangsu, China
[2]Department of Geriatrics, Changzhou Hospital of Traditional Chinese Medicine, Nanjing University of Chinese Medicine, Changzhou 213000, Jiangsu, China
[*]Correspondence: Tianping Luo, Department of General Surgery, Changzhou Hospital of Traditional Chinese Medicine, Nanjing University of Chinese Medicine. E-mail: 927884643@qq.com

Abstract

Background: Recurrence after curative resection of colorectal cancer is stratified largely by stage. Icariin, a prenylated flavonoid glycoside from *Epimedium* spp., defines a target space not yet examined as a source of recurrence-associated transcriptomes.

Methods: Icariin targets intersected with tumour-versus-normal expression (TCGA-COAD, GSE39582) gave an 84-gene candidate pool; a pre-specified 100-combination screen (10 selectors × 10 models) in GSE14333 ($n=226$, 50 events) produced a fixed nine-gene signature. Four conditions fixed before external testing — patient-level independence, a depositor-coded disease-free-survival (DFS) event, the same platform, and the largest event count — selected GSE17536 ($n=145$, 36 events).

Results: The signature (TIMP1/ SPARC/CLDN1 positive; PCNA/ACADM/ DPEP1/MMP3/CASP7/MMP12 negative) stratified DFS in GSE17536 (C-index 0.734; HR 2.30 per SD, $P=2.3\times10^{-8}$) and improved on stage, age and sex (ΔC +0.085, 95% CI 0.023–0.169; likelihood-ratio $P<0.001$), but under-estimated absolute risk (observed-to-expected 1.90). Pooling the eight deduplicated cohorts (181 DFS events) gave a random-effects HR of 1.53 per SD (95% CI 1.13–2.07, $P=0.013$; $I^2=69$%; prediction interval 0.71–3.31). High-risk tumours were mesenchymal/stromal and immune-suppressive; the signal localized mainly to fibroblasts and endothelium through SPARC and TIMP1. A protocol-matched random-pool control placed the signature at the 88th percentile ($P=0.118$), and the stage-dominated TCGA-COAD transfer added no incremental value.

Conclusion: The nine-gene signature is an externally validated relative-risk stratification biomarker derived from an icariin-associated candidate space; all results are in silico and await prospective and experimental validation.

Keywords: colorectal cancer; icariin; prognosis; machine learning; consensus molecular subtype; external validation

1 Introduction

Colorectal cancer (CRC) accounted for approximately 9.6% of newly diagnosed cancers and 9.3% of cancer deaths worldwide in 2022, ranking third in incidence and second in cancer-related mortality [1]. For patients with nonmetastatic colorectal cancer, complete surgical resection remains the cornerstone of curative treatment, and decisions regarding postoperative chemotherapy and surveillance are still based primarily on pathological stage, conventional high-risk features, and mismatch-repair or microsatellite-instability status [2]. However, patients within the same clinicopathological category can experience markedly different outcomes, indicating that anatomical staging incompletely captures the biological determinants of recurrence. Circulating tumour DNA (ctDNA) has emerged as a marker of postoperative molecular residual disease: the updated GALAXY analysis demonstrated a strong association between postoperative ctDNA positivity and recurrence or death in resectable CRC [3], and the five-year results of the randomized DYNAMIC trial supported ctDNA-guided adjuvant management in stage II colon cancer [4]. Tissue-based molecular models nonetheless remain valuable for defining the biological programs associated with recurrence.

Transcriptomic classification has revealed substantial molecular heterogeneity within CRC. The consensus molecular subtypes (CMS) framework separates CRC into four major groups, among which CMS4 is characterized by transforming growth factor-β activation, stromal invasion, angiogenesis, and a mesenchymal transcriptional program [5]; a systematic review and meta-analysis subsequently confirmed that CMS4 is associated with particularly unfavorable survival in localized CRC [6]. Recent single-cell studies have refined this framework by showing that bulk-tumour classifications reflect variable contributions from malignant epithelial cells and the tumour microenvironment. Joanito et al. demonstrated that fibrosis represents a clinically relevant axis partly independent of the intrinsic epithelial state [7], whereas Khaliq et al. and Koncina et al. associated cancer-associated fibroblast and macrophage populations with adverse outcomes, tumour growth and immunotherapy-resistant, immune-excluded microenvironments [8,9]; multicellular immune hubs and extensive malignant–stromal–endothelial–immune communication have since been mapped in CRC [10,11]. A prognostic signature derived from bulk expression data should therefore be interpreted in terms of its cellular origin and microenvironmental context.

Icariin is a prenylated flavonoid glycoside derived from *Epimedium* species, and icaritin is an aglycone and bioactive metabolite of icariin. Preclinical studies suggest that these compounds exert pleiotropic antitumour effects rather than acting through a single molecular target. In colon cancer models, icariin increased sensitivity to tumour necrosis factor-related apoptosis-inducing ligand (TRAIL) through reactive oxygen species, extracellular signal-regulated kinase, CHOP and death-receptor signaling [12], and enhanced the antitumour activity of 5-fluorouracil by inhibiting NF-κB in cells and xenografts [13]. More recently it was reported to promote ferroptosis by activating mitochondrial dysfunction and to synergize with PD-1 inhibitors [14], to re-educate the tumour microenvironment by inhibiting M2 macrophage polarization via PI3K/AKT [15], and to induce mitochondrial-mediated apoptosis through Wnt/β-catenin signaling [16],

while icaritin suppresses invasion and metastasis by modulating epithelial–mesenchymal transition [17] and represses autophagy by disrupting HSP90–TXNDC9 interactions [18]. Icaritin has also entered clinical practice outside CRC: Chinese guidelines report its conditional approval for selected advanced hepatocellular carcinoma patients unsuitable for or declining standard systemic therapy [19,20], and nanomedicine-based icaritin formulations are being explored to boost antitumour immunotherapy [21]. These observations provide a translational precedent for the compound class, but they do not establish clinical efficacy in colorectal cancer, and no study has yet asked whether the target space associated with icariin is enriched for genes that mark recurrence risk or the cellular architecture of human colorectal tumours.

This gap is addressable with transcriptome data, but high-dimensional input, algorithm-dependent feature selection, and thin external validation readily produce unstable or overly optimistic models [22], which is why TRIPOD+AI guidance emphasizes transparent model development, independent performance evaluation, and complete reporting of discrimination and clinical utility [23]. The present study follows a fixed sequence. A multi-source library of putative icariin-associated targets was constrained by differential expression in colon adenocarcinoma (COAD) in two independent cohorts (TCGA-COAD and GSE39582) to define a candidate gene set; 100 combinations of feature-selection and survival-modeling algorithms were then compared under fixed consensus rules to derive a nine-gene model; the model was frozen and validated for disease-free survival (DFS) in an independent external cohort (GSE17536); its direction was estimated across resected cohorts by random-effects meta-analysis; the high-risk phenotype was characterized as a CMS4-like, stromal-rich state; and the cellular source of the signal was localized by single-cell analysis. Icariin is the source of the candidate gene set, and the pharmacological analyses reported here are exploratory. We asked whether this sequence would identify a reproducible adverse-prognosis transcriptomic phenotype and clarify the stromal context of its prognostic signal.

2 Materials and Methods

2.1 Study design and data sources

Thresholds, display rules and model-selection rules were fixed before execution. All datasets are publicly available (TCGA and GEO); no new human samples were collected, and no additional ethics approval was required.

The Cancer Genome Atlas colon adenocarcinoma cohort (TCGA-COAD) provided RNA-seq STAR-Counts (GDC, DR46.0) with GENCODE v36 annotation; after format quality control and de-duplication, 512 samples (471 primary tumours, 41 solid normal) were used for differential expression, and 438 primary-tumour cases with curated clinical endpoints [24] were used for survival analysis. Three Gene Expression Omnibus (GEO) microarray cohorts were used. GSE39582 (GPL570; 566 primary tumours and 19 non-tumoural mucosa) [25] served as the screening/replication layer and was *not* used for survival external validation, to avoid screening–validation circularity. GSE14333 (GPL570; *n=226* patients with 50 disease-free survival [DFS] events) [26] was the signature training cohort. GSE17536 (DFS,

n=145, 36 events) [27] served as the external validation cohort on the DFS endpoint (see *Outcome definitions and harmonization*) and constituted the primary external validation analysis set. It was required to satisfy four design conditions: no sample overlap with the training or screening cohorts under both an expression-fingerprint and a GEO series-relation audit; a depositor-coded DFS endpoint matching the training endpoint; the same platform (GPL570), so that the frozen probe rule and the locked absolute risk threshold transfer without re-standardization; and the largest event count among the cohorts meeting the first three conditions. The further independent cohorts pooled with it (Methods: *Pooled external analysis*) are listed in Table 1. Microarray preprocessing followed cohort-specific pipelines, documented cohort by cohort in Supplementary Table S9: GSE14333 used the depositor's series-matrix values as deposited (simpleaffy MAS5.0 calls followed by quantile normalization) with no local re-normalization; GSE17536 was normalized locally from the deposited CEL files with RMA; GSE39582 used the depositor-provided values (RMA with ComBat correction). No cross-cohort batch correction was applied by us, so the pooled cohorts keep their depositors' differing pipelines (Supplementary Table S9) — one documented source of the between-cohort heterogeneity quantified below. The single-cell reference was the public colorectal-cancer atlas GSE178318 (primary tumour, liver metastasis, and peripheral blood mononuclear cells [PBMC]; 139,543 cells after quality control of ≥200 genes/cell and ≥10 cells/gene). Cohort roles, platforms, sample sizes, endpoints, and risk-score transfer schemes are summarized in Table 1.

Table 1 Cohorts used in this study, their role in the evidence chain, and the fixed-score result in each. Every external cohort is scored with the frozen GSE14333 coefficients and training μ/σ and its own cohort-internal standardisation of the resulting score; the eight cohorts below contribute to the random-effects meta-analysis of Fig. 6e. Cohorts screened and set aside are reported with their reasons above and in Supplementary Tables S6–S8 and are not pooled

Cohort	Role	Platform	Endpoint	*n* (events)	HR per SD (95% CI)
Discovery, screening and primary validation					
TCGA-COAD	transfer / characterization	RNA-seq	OS	438	–
GSE14333	training (signature)	GPL570	DFS	226 (50)	2.43
GSE17536	primary external validation	GPL570	DFS	145 (36)	2.30 (1.72–3.08)
GSE39582	screening only	GPL570	–	566 tumours	–
External cohorts pooled in the random-					

effects meta-analysis (Fig. 6e)					
GSE106584	round 3 (Supplementary Table S7)	GPL17586	DFS	156 (52)	0.99 (0.75–1.32)
GSE17536	primary external validation	GPL570	DFS	145 (36)	2.30 (1.72–3.08)
GSE161158[a]	pre-declared candidate	GPL570	DFS	66 (15)	2.61 (1.54–4.41)
GSE39084	round 3 (Supplementary Table S7)	GPL570	DFS	61 (19)	1.48 (0.95–2.29)
GSE33114[b]	widened (Supplementary Table S6)	GPL570	RFS	89 (18)	1.55 (1.05–2.27)
GSE17537	widened (Supplementary Table S6)	GPL570	DFS	52 (17)	1.09 (0.70–1.69)
GSE143985[c]	widened (Supplementary Table S6)	GPL570	DFS	91 (15)	1.19 (0.78–1.81)
GSE38832	widened (Supplementary Table S6)	GPL570	DFS	92 (9)	2.08 (1.03–4.22)
Pooled (random effects)	*k=8*	–	DFS/RFS	**181**	**1.53 (1.13–2.07)**

OS Kaplan–Meier used *n=419* (94 events) and the multivariable model used *n=411*. GSE39582 is the candidate-screening cohort (source of the third DEG set) and is not counted as a survival external-validation cohort. [a]Duplicate-free subset of GSE161158 (66 of 191 samples; the samples removed in the unified-panel fingerprint audit matched a pooled GSE17536 or GSE17537 sample one-to-one at $\rho \geq 0.90$, see Supplementary Table S6). [b]GSE33114 is the SuperSeries of GSE33113 and contains all of its samples; the cohort is scored once, under the accession whose 89 primary-tumour samples carry the depositor's metastasis-or-recurrence (RFS) event field. [c]GSE143985 is the same cohort as GSE92921 and is counted once. Cohorts excluded by design (biopsy or rectal neoadjuvant radiochemotherapy cohorts), cohorts sharing samples with a cohort already counted, and event-poor cohorts (GSE12945, 5 DFS events) were screened under the same protocol, are reported in Supplementary Tables S6–S8 and are not pooled. The pooled estimate is stable

to their inclusion (Supplementary Table S8), and the leave-one-out re-estimates are shown in Supplementary Fig. S7.

2.2 Outcome definitions and harmonization

All training and external-validation analyses use the disease-free survival (DFS) endpoint exactly as provided by the depositors; events were not re-adjudicated by us. Seven of the eight pooled cohorts deposit a DFS event field and one, GSE33114, a metastasis-or-recurrence field; because the remaining coding is identical we designate these endpoints DFS throughout, write the pooled event count as DFS/RFS, and label GSE33114 RFS where cohorts are listed individually. The term "recurrence" is otherwise used only in its descriptive biological sense. Follow-up times are months from surgery, the depositor-coded binary indicator defining events (1) and censoring at the last documented recurrence-free follow-up (0), and the pre-specified quality-control rule retained patients with complete and positive DFS time and event fields.

Two operational definitions underlie the endpoint. In GSE14333 the series-matrix annotations provide DFS_Time (months) and a binary censoring indicator (DFS_Cens), and the original study analysed recurrence (6-year recurrence in Dukes' stage B/C disease) [26]. For GSE14333, the deposited DFS_Cens variable is a censoring indicator; therefore the analytic event variable was defined as DFS_event = 1 − DFS_Cens. This transformation yielded 50 DFS events among 226 retained patients. In GSE17536 the depositor annotation defines the event explicitly as cancer recurrence and provides death from any cause and cancer-specific death as separate variables. Quality control retained 226/290 (GSE14333) and 145/177 (GSE17536) patients; the patient flow, the stage distribution of the excluded patients and the baseline characteristics of both analysis sets are given in Supplementary Tables S10 and S11. Median recorded observation was 38.5 and 37.3 months (the median of the time-to-event-or-censoring variable, not a median survival time; fewer than half of the retained patients had an event). Because event ascertainment was not centrally reviewed and the handling of second primary tumours, post-recurrence deaths and adjuvant treatment is not uniformly documented, the primary external validation set was confined to a single DFS cohort (GSE17536; *n=145*, 36 events), carrying no cross-series endpoint heterogeneity; further resected cohorts enter only through the pooling protocol described under *Pooled external analysis*, where between-cohort heterogeneity is quantified rather than assumed away.

Four conditions, declared before the search, governed the choice of that cohort, as listed under Study design; the four are not repeated here. The search was exhaustive rather than convenience-based: a survival-term query, a complementary MeSH/recurrence re-query, and parallel ArrayExpress/BioStudies and GPL570 sweeps were screened against these four conditions (Results; Supplementary Data S7). GSE17536 was the only cohort satisfying all four; per-candidate adjudications are given in Supplementary Data S7.

2.3 Icariin compound set, drug-likeness profiling, and target library

Icariin, the major prenylated flavonoid glycoside of *Epimedium* spp., and its aglycone icaritin were profiled with RDKit for physicochemical and drug-likeness descriptors (the

ADMET-relevant subset) from their canonical simplified molecular-input line-entry system (SMILES) strings: molecular weight, Crippen logP, topological polar surface area (TPSA), hydrogen-bond donors/acceptors (HBD/HBA), rotatable bonds, aromatic rings, and fraction sp^3 carbon, together with Lipinski, Veber, Ghose, Egan and Muegge rule evaluations and the Martin bioavailability score. For display, descriptors were normalized to the Lipinski/Veber rule bounds (MW/500, logP/5, TPSA/140, HBD/5, HBA/10, rotatable-bond/10); the rule-limit ring in Fig. 1c marks normalization value 1.0. ADMET values are reproducible descriptor calculations, not external tool scores.

Candidate targets were curated from six independent evidence sources: ChEMBL activity data, PubChem BioAssay, SwissTargetPrediction structural prediction, the Comparative Toxicogenomics Database (CTD), the HERB database, and PharmMapper. Entries were mapped to human gene symbols and consolidated after excluding non-human orthologs into a 551-gene library, with evidence graded A (experimental activity, $n=9$), B (curated or multi-source, $n=211$), or C (computational prediction, $n=331$). The library retains one long non-coding RNA (H19), kept because it is a legitimate curated target rather than an annotation artifact; it was retained by 8 of the 90 non-control screen combinations (frequency 0.089) and does not enter the signature. Multi-source genes are counted once per source in Fig. 1d.

2.4 Differential expression and construction of the candidate pool

TCGA-COAD differential expression used DESeq2 [28] with edgeR [29] as an independent sensitivity analysis; significant genes were defined as Benjamini–Hochberg (BH)-corrected $P<0.05$ and $|\log_2 FC| \geq 1$ (Fig. 2a). Variance-stabilizing transformation (VST)-normalized expression was used for all downstream analyses, and sample quality was summarized by principal-component analysis of the top 8,000 variable genes (Fig. 2d). In the independent screening cohort GSE39582, tumour-versus-normal differential expression was tested with limma under the same thresholds (BH-adjusted $P<0.05$, $|\log_2 FC| \geq 1$; Fig. 2b).

Candidates were defined by the strict three-way intersection of the icariin target library (551 human genes), the TCGA-COAD differentially expressed genes (15,170 with a mapped gene symbol), and the GSE39582 tumour-versus-normal differentially expressed genes (Fig. 2c). The resulting 84-gene candidate pool was used for all subsequent machine-learning and prognostic analyses. Unsupervised clustering of the 84 candidates (VST z-scores; rows clustered by $1-\rho$) was used as a quality check of the tumour/normal separation (Fig. 2e).

2.5 Functional enrichment and protein–protein interaction network analysis

Over-representation analysis (ORA) of the 84 candidates was run against Gene Ontology (biological process, cellular component, molecular function), Reactome, and KEGG with clusterProfiler [30] and ReactomePA [31] (BH-adjusted hypergeometric P). The universe was all genes tested in TCGA-COAD differential expression, restricted per system to its annotatable subset; significance was defined as adjusted $P<0.05$. Full term tables are provided in the Supplementary Materials; the main figure shows a representative subset (GO: 6 BP, 4 CC, 4 MF terms; KEGG: 6; Reactome: 8; Fig. 3a, b).

Protein–protein interaction (PPI) networks were built from STRING v12.0 [32] restricted to candidate-only edges (confidence score ≥400, with a sensitivity run at ≥700). Network modules were detected by greedy modularity clustering. Node centrality (degree, betweenness, closeness) was computed with networkx [33], converted to percentile ranks, and averaged into a hub-composite score. Core targets were declared by a pre-specified rule: experimental activity (evidence level A or B) or ≥2 independent source databases, together with a hub-composite ≥0.90 and non-isolated status in the primary network. An integrated evidence matrix (drug-target evidence, differential expression, PPI centrality, significant pathway membership, signature membership) is provided in the Supplementary Materials (Supplementary Data S3).

2.6 Multi-algorithm screen and consensus signature construction

A *10×10* combination screen (10 feature selectors × 10 survival models; Cox-family, regularized Cox, random survival forest, extra trees, gradient boosting, and survival support vector machines) was run on the GSE14333 training cohort. The screen used a single-level five-fold cross-validation whose fold partition was fixed across all 100 combinations (seed 42); within each training fold the feature selector was refitted on the training part only and the model was evaluated on the held-out fold, so no selection or fitting step had access to the evaluation samples. Per-gene scaling statistics (mean/standard deviation) were computed on the full training cohort before fold splitting and transferred to the held-out fold; repeating the entire 100-combination grid with fold-internal scaling changed the median cross-validated C-index by 0.000 (mean -0.001; 12/100 combinations shifted by >0.01, maximum $|\Delta C|=0.026$), with no effect on any conclusion (Supplementary Data S6). All selectors and models ran at fixed a-priori hyperparameters (Supplementary Tables S1 and S2); no hyperparameter search was performed inside the screen. Because the design is a single-level cross-validation with in-fold feature selection, we do not describe it as nested cross-validation; the complete flow, including the inner cross-validation used only for the final consensus model, is shown in Supplementary Fig. S5. The no-selection `all`-feature row served as a control arm and was excluded from consensus counting, because its retained set is the full 84-gene pool by construction; in one of its five folds the no-selection Cox model failed to converge and its score is the mean of the remaining four folds. Genes retained in ≥40% of the 90 non-control combinations were selected (Fig. 4b; per-selector selection frequencies in Supplementary Fig. S6), and a consensus Cox regression with elastic-net penalty (l_1 ratio 0.5; penalty strength selected by an inner five-fold cross-validation over the fixed grid 0.01/0.05/0.1/0.3 on the training cohort, then refitted on the full cohort) was fitted to them. Threshold-selected genes that received a zero coefficient after shrinkage were retained in the figure but carry no weight in the model. Selection stability of the screen and of the final nine-gene model is quantified in Supplementary Fig. S6 and Supplementary Table S3.

2.7 Risk-score transfer and internal characterization in TCGA-COAD

The signature was trained on DFS in GSE14333 and externally validated in GSE17536, so the TCGA-COAD analyses serve here as internal transferability and phenotype-

characterization analyses. For these internal analyses, overall survival (OS) was used as the primary endpoint for power reasons: OS was available for *n=419* patients with 94 events, whereas disease-free interval (DFI) is defined only for patients without metastasis at diagnosis (*n=175*, 22 events) and progression-free interval (PFI) additionally captures new primary tumours; Increments based on OS, PFI, and DFI were all computed and are reported together. OS is affected by competing non-recurrence mortality and adjuvant treatment. For each sample the risk score was $\Sigma_i \beta_i z_i$, where z_i is the per-gene z-score: TCGA-COAD used cohort-internal z-scores on VST-transformed expression, whereas external cohorts used the fixed GSE14333 μ/σ transfer applied without re-estimation. Risk groups were split at the cohort-internal median, and the resulting risk landscape is displayed in Fig. 4c–e. The internal OS analysis used Kaplan–Meier and log-rank tests (*n=419*) and univariable and multivariable Cox models adjusting for composite stage, age, and sex (*n=411*).

Clinical-utility analyses for OS (Fig. 5) used the four-variable Cox model (risk score + composite stage + age + sex). Four summaries were produced: (i) Uno time-dependent receiver-operating-characteristic analysis of the model linear predictor at 1, 3, and 5 years; (ii) calibration of predicted versus observed OS at 1, 3, and 5 years by quartiles of the 3-year predicted OS; (iii) a clinical impact curve and a decision-curve analysis for 3-year mortality (number needed to treat at high risk versus weighted event reduction, inverse probability of censoring weights [IPCW]); and (iv) a nomogram constructed with the `rms` package from the same multivariable model (composite stage, risk score, age) as an illustration of clinical utility in the OS setting, with model coefficients guarded against the fixed Cox table.

2.8 External validation and incremental value

External validation used the GSE17536 DFS cohort described under Study design (Fig. 6). Discrimination used Harrell's C-index with bootstrap 95% confidence intervals (1,000 resamples, seed 42) and Uno IPCW time-dependent AUC and ROC curves at 1/3/5 years; stratification used median-grouped Kaplan–Meier on the transferred score. Incremental value over a clinical baseline of stage, age and sex was quantified as the change in Harrell's C-index (ΔC; 1,000 bootstrap replicates) and, at 36 months, by Uno IPCW-based continuous net reclassification improvement and integrated discrimination improvement. Calibration was assessed by transporting the training cohort's baseline rather than refitting inside the validation cohort: with the score standardized on the training mean and SD, a Cox model fitted in GSE14333 gave a baseline survival $S_0(t)$ that was applied to each GSE17536 patient as $S_0(t)^{\exp(\beta_{\mathrm{tr}} u_i)}$. We report the calibration slope (the validation-cohort coefficient on that scale divided by the training coefficient), observed-to-expected event ratios at 12/36/60 months with 1,000 bootstrap resamples, and the IPCW Brier score at the same times for the clinical baseline (stage, age and sex) versus the same baseline plus the risk score, with the censoring distribution estimated within GSE17536. The external clinical baseline therefore has the same composition as the internal TCGA clinical baseline (stage + age + sex). The equivalent internal increments over the clinical model were computed in TCGA-COAD for OS, PFI and DFI.

Sample independence was audited before validation because partial cohort overlap between GSE14333 and GSE17536 has been alleged in the literature. No GEO accession is shared between the two series, and a full-probe (54,675-probe) expression-fingerprint cross-correlation gave a maximum cross-cohort Pearson ρ of 0.877, below the within-cohort different-sample background (0.952 and 0.988) and clearly separated from duplicate-sample levels (≈0.97–0.99), so no pair reached the pre-declared duplicate thresholds ($\rho \geq 0.90$ or ≥ 0.95).

The audit was then repeated for every cohort entering the pooled analysis under a single preprocessing rule, leaving 17,314 genes scored identically across all ten series. The only sample pairs reaching $\rho \geq 0.90$ lay in four cohort pairs, each a strict one-to-one match: the already-disclosed GSE17536/GSE29621 correspondence, and three pairs joining GSE161158 to GSE17536, GSE17537 and GSE29621 respectively, of which the last two had not been screened before. No other pair of cohorts reached $\rho = 0.90$, the highest remaining value (0.84) lying within the range of the within-cohort background maxima (0.50–0.85); two positive controls run through the same pipeline (known duplicates, and duplicates spiked with Gaussian noise at up to 0.20 of the within-gene SD) were both recovered in full (Supplementary Data S6).

A robustness battery accompanied the external validation and is reported in full in Supplementary Data S6 and S7: duplicate-exclusion and transferred-cut-off sensitivity, stage-restricted and site-stratified analyses, an external clinical-baseline increment, Schoenfeld proportional-hazards tests, treatment adjustment in the training cohort, and the protocol-matched target-library-free null screen.

The third-cohort protocol described next is a fixed-score evaluation. It uses depositor-normalized series-matrix values, the frozen probe rule, the frozen coefficients and training μ/σ (Supplementary Table S5), the locked threshold, Harrell's C-index with bootstrap intervals, locked-threshold and cohort-median Kaplan–Meier comparisons, a ΔC ladder over stage + age, and a fingerprint overlap audit against all three internal cohorts ($\rho \geq 0.90$ flagged as suspected duplicate, with flagged samples removed in sensitivity analyses, for the most affected cohort also at $\rho \geq 0.85$). The signature coefficients and the training μ/σ were never re-fitted on any evaluation cohort, so the reported ΔC compares the transferred fixed score against a cohort-local clinical baseline rather than a re-fitted joint model; Supplementary Table S12 records, for every metric reported in this work, which models are compared and where each of them is estimated. The protocol was applied to three pre-declared candidate cohorts and then to every further public colorectal series that systematic search could identify. Cohorts on non-GPL570 platforms were evaluated under a cross-platform extension. Features were mapped to gene symbols through the platform's official GEO annotation, and per gene the cohort-wide highest-mean feature was retained, an outcome-blind rule that cannot be frozen across platforms; the fingerprint audit is reported descriptively only. Where AJCC stage was not deposited, the ΔC ladder is reported over age alone as an exploratory substitute.

The final, exhaustive round queried the GEO DataSets index through NCBI E-utilities rather than by keyword, retrieved every series annotated as human colorectal cancer with at least 50 samples, and screened each for a time-to-event endpoint; series whose recorded GEO

type was not a genome-wide expression assay were set aside on type alone, before any score was computed. Sample independence was audited structurally through the GEO series-relation metadata field of every series cited here, and three protocol amendments were fixed in advance of that round. The full series-level protocol — the RNA-seq substitution rule (R1), the non-transportability of the frozen threshold and the cohort-standardized sensitivity analysis (R3) — is given in the Supplementary Methods.

2.9 Pooled external analysis

The pooled external estimate summarises the further cohorts carried into it. Each pooled cohort is independent of the training, screening and primary validation cohorts under both audits described above; reports a depositor-coded DFS or RFS endpoint; comprises surgically resected tumours without neoadjuvant radiochemotherapy; and contributes at least six events. Cohorts depositing pre-treatment biopsies, and rectal-cancer cohorts treated with preoperative radiochemotherapy, were excluded by design before scoring, as stated in the footnote to Table 1, and are listed in Supplementary Tables S6 and S7 rather than pooled. Platform was not restricted. Two further cohorts fell outside these conditions — one carrying five events, below the event floor, and one whose metastasis-free endpoint is derived from deposited dates rather than coded by the depositor — and were retained for sensitivity analysis only. The primary analysis set excludes every sample implicated in a duplicate pair, so that each patient contributes once. Each cohort was scored with the frozen coefficients and the training μ/σ and summarised by a Cox proportional-hazards model on the cohort-standardized risk score; the hazard ratio per standard deviation and its standard error (taken from the model's coefficient standard error, rather than back-computed from the *P* value) were then combined in a random-effects model with restricted maximum-likelihood estimation of τ^2 and the Hartung–Knapp adjustment applied to the interval of the pooled effect; DerSimonian–Laird method-of-moments estimation with the conventional normal-theory interval is reported as a comparator. Heterogeneity was summarised by Cochran's Q, τ^2 and I^2, and a 95% prediction interval describes the effect expected in a new cohort. Robustness was assessed by leave-one-out re-estimation of the pooled hazard ratio, by a fixed-effect model, and by adding the two set-aside cohorts back. The eligibility conditions above were pre-specified; the decision to aggregate the qualifying cohorts was taken after the individual cohort results had been read, and is reported as such in the Results.

2.10 Phenotype profiling: consensus molecular subtype, MSI/TMB, and mutation landscape

Consensus molecular subtypes (CMS) were assigned with CMScaller (v0.99.2; nearest-centroid NTP, Entrez-gene input, n_{perm}=*1000*, seed 42) [5,34] in TCGA-COAD (*n=403* of 438 predicted retained at FDR <*0.05*) and independently in GSE39582 (*n=566*, all predictions retained; median-split risk groups, *283*/*283*). Subtype proportions were compared between risk groups with Fisher's exact test per subtype. Microsatellite instability (MSI) status (MANTIS score >*0.4*) and MSI score, and tumour mutation burden (TMB; non-synonymous somatic mutations per Mb, PanCanAtlas annotation), were compared between risk groups (MSI *n=413*; TMB *n=393*). For the mutation landscape, the top 30 recurrently

mutated genes (TCGA-COAD frequency ranking) were drawn as a per-sample oncoprint from the cBioPortal PanCancer Atlas 2018 snapshot (the same TCGA cohort source as the MSI/TMB analyses), distinguishing truncating from missense/other mutations; per-gene association with the risk group was tested with Fisher's exact test and BH correction, and no gene was significant (reported as such). Clinical associations (T, N, M, stage, age, sex) were tested with Mann–Whitney *U* tests (*n=364*). All comparisons are displayed in Fig. 7. Because the score is CMS4-skewed, two further analyses were declared in advance: its discrimination within each CMS class (Harrell C-index, 1,000 within-stratum bootstrap resamples; strata with fewer than 20 events flagged as not evaluable) and its increment over stage, age, sex and CMS (C-index, continuous NRI and 36-month IDI, 1,000 bootstrap replicates, seed 42), computed in TCGA-COAD and in GSE39582.

2.11 Gene-set enrichment analysis and mechanism modules

Pre-ranked gene-set enrichment analysis (GSEA) [35] of Hallmark (MSigDB v2023.2) [36] and KEGG gene sets was run on TCGA-COAD genes ranked by the Spearman correlation between expression and the continuous risk score (fgsea preranked implementation; *n=438* tumours). The display rule selected the top 6 pathways by BH-adjusted *P* within each normalized-enrichment-score (NES) sign (12 rows per panel), sorted by NES; all displayed rows were BH-significant. Positive NES indicates enrichment in the high-risk group (Fig. 8a, b).

Module scores for epithelial–mesenchymal transition (EMT), oxidative phosphorylation (Hallmark and KEGG), and ferroptosis (FerrDb driver/suppressor sets), together with immune and stromal infiltration scores, were computed per TCGA-COAD sample by single-sample GSEA (ssGSEA) and correlated with the risk score (Spearman, BH-corrected, *n=438*). The EMT module shares three genes with the signature (TIMP1, SPARC, MMP3). A de-overlap sensitivity analysis that removes them from the EMT gene set leaves the enrichment unchanged (de-overlap re-run: preranked-GSEA NES 3.110 → 3.152, rank unchanged; ssGSEA Spearman ρ 0.458 → 0.466; the corresponding frozen primary values, reported below, are NES 3.144 and ρ 0.455), so the EMT result does not depend on the signature genes it contains. All 10 modules are reported without filtering. The five pre-specified core targets (TIMP1, PTGS2, COL1A1, HMOX1, MMP9) were tested for Spearman correlation with the risk score with BH correction; all five are reported without outcome-guided selection. Representative module-versus-risk scatter plots (EMT, stromal) show per-sample values, with the Spearman ρ recomputed from the plotted values and asserted equal to the fixed module table (Fig. 8c–f).

2.12 Stromal-content comparator scores and incremental-value analysis

Because MMP3 and MMP12, archetypal pro-tumourigenic matrix metalloproteinases in colorectal cancer, emerged with negative coefficients in the fitted model, we formally tested whether the risk score partly reads out tumour stromal content. Four comparator scores were computed in the TCGA-COAD and GSE14333 expression matrices: the ESTIMATE stromal and immune signatures via single-sample gene-set enrichment analysis (ssGSEA; `gseapy` 1.3.1, rank-based sample normalization) [37], and simplified marker-sum implementations of the MCP-counter fibroblast and endothelial marker panels (the sum of

$\log_2$(expression+*1*) over each panel's marker genes, rather than the original implementation) [38]. In both cohorts, identifiers were collapsed to gene symbols by retaining the probe or Ensembl transcript with the highest mean expression per symbol, as in the main pipeline. Associations with the risk score were quantified by Spearman correlation with Benjamini–Hochberg adjustment over the eight cohort–score tests, and by Mann–Whitney comparison of high- versus low-risk groups. Incremental value was assessed with four nested Cox models (penalizer 0.001, as in the main pipeline): a clinical base model (stage, age, and sex in TCGA-COAD, overall survival, *n=411* with 90 events; Dukes stage, age, and sex in GSE14333, DFS, *n=226* with 50 events), base plus the z-scored ESTIMATE stromal score, base plus the z-scored risk score, and the full model with both. We report the risk-score hazard ratio with and without stromal adjustment and the change in Harrell C-index from adding the risk score over the base model and over the base-plus-stromal model, with bootstrap 95% confidence intervals and two-sided permutation *P* values (1,000 resamples, fixed seed 42). Results are shown in Supplementary Fig. S4A, B (correlations) and S4C, D (incremental value); score tables in Supplementary Data S4.

2.13 Immune deconvolution, checkpoint genes, and TIDE

Immune and stromal populations were quantified by ssGSEA (Bindea signatures; 23 features) and by simplified MCP marker-sum scores (10 features; see stromal-content comparators) and compared between risk groups (Student's *t* test with BH correction across features; all features of both panels reported unfiltered; Fig. 9a, b). Expression of 21 immune-checkpoint genes was compared between risk groups with Welch *t* tests; mean differences (High − Low) are shown with 95% confidence intervals recomputed from per-sample expression, with centres asserted equal to the fixed table, and BH correction across the 21 genes (Fig. 9c, d). Tumour Immune Dysfunction and Exclusion (TIDE, v1.3.9) scores [39] were computed from TCGA-COAD VST expression with `cancer='Other'` (SKCM-rule approximation). Scores were compared between risk groups (Wilcoxon), against the continuous risk score (Spearman), and across the five TIDE model components (T-cell dysfunction, T-cell exclusion, myeloid-derived suppressor cells [MDSC], cancer-associated fibroblasts [CAF], and M2-polarized tumour-associated macrophages [TAM M2]; all reported; stars from the fixed raw Spearman *P*) (Fig. 9e–g).

2.14 Single-cell localization of the signature module

The fixed signature coefficients were applied to the GSE178318 atlas (see Study design). Per-cell module activity was $\Sigma_i \beta_i \cdot \log_{10}(\mathrm{CP10k}+1)$ over the nine signature genes. Cell types were assigned by canonical-marker mean expression (argmax over eight classes). This is a coarse marker-based annotation, not the original study's official labels (the marker list, its lack of overlap with the nine signature genes, and the sensitivity of the stromal contrast to the positivity definition are given in Supplementary Table S13). The atlas is tissue-mixed, so results are interpreted as descriptive cell-type/tissue localization only (Fig. 10). Violin panels display a fixed random sample of 3,000 cells per class (seed 84) for legibility; medians, means and Kruskal–Wallis tests use all cells. Feature plots show, per coefficient sign, the two signature genes with the highest detection rate (pre-declared rule), on a shared linear scale capped at the joint 99th percentile.

Donor-level tissue analysis (the primary single-cell tissue analysis): because the atlas samples six patients across three tissue compartments, cell-level tests overstate evidence (pseudo-replication). For each donor and compartment the detection rate was defined as the fraction of cells with a positive module score; stroma was defined as fibroblasts + endothelial cells. Stromal versus epithelial detection rates were compared within the primary CRC tumours by exact sign test, and primary-versus-metastasis differences by donor-paired exact sign test (two-sided). The cell-level tissue comparison is retained only as a descriptive display (Supplementary Fig. S3E). An additive per-gene decomposition of the stromal contrast and its leave-one-gene-out recomputation are reported in Supplementary Data S7.

2.15 Molecular docking and interaction analysis

Exploratory drug-sensitivity correlations used the Genomics of Drug Sensitivity in Cancer (GDSC2) release: per drug, the Spearman correlation between the risk score and log IC_{50} across $n=46$ colorectal-cancer cell lines (185 drugs), with Benjamini–Hochberg correction over all drugs. AutoDock Vina 1.2.7 [40] was used to dock icaritin (the aglycone) and its glycoside icariin against the catalytic domains of MMP9 (PDB 1GKC, chain A; co-crystal ligand NFH) and MMP12 (1RMZ, chain A; NGH), plus the extension targets PTGS2 (3LN1, chain A; CEL) and MAPK3 (4QTB, chain A; 38Z). Docking used a 22 Å box centred on the co-crystallized ligand of the relevant chain, exhaustiveness 16, and three seeds (42/7/99) with the co-crystal ligands as same-protocol controls (Fig. 11). Receptors (protein chain plus catalytic metals, retained for MMP9/MMP12) were cleaned of crystallographic water, sulphate, and all non-metal ligands; polar hydrogens were added and Gasteiger charges assigned with Open Babel, and receptors were written as rigid PDBQT. Each protocol was validated by re-docking the co-crystal ligand and comparing it to its crystal pose (heavy-atom RMSD <2.0 Å = valid): MMP9 1.35 Å, MMP12 0.70 Å, PTGS2 0.43 Å, MAPK3 6.71 Å (invalid). MAPK3 additionally scored less favourably than its co-crystal control and was excluded from the docking conclusion; PTGS2, whose protocol validated normally, is reported instead as an explicit negative result (Results). MMP3 (2D1O; FA4) was also docked but failed this validation and is excluded: the original two-chain setup mis-centred the box, and the corrected single-chain re-dock reproduced the binding site but not the crystallographic orientation of the fragment (Supplementary Data S6). The single ligand–target combination not covered by the earlier fixed result sets (MMP9 × icariin) was docked under the identical protocol. The docking conclusion was then re-tested under the identical protocol with the second scoring function Vinardo, with one microstate per free aromatic hydroxyl (the intramolecularly hydrogen-bonded 5-OH retained throughout, the glycosylated 7-OH of icariin unable to ionise), and with three metal treatments (zinc as prepared, deleted, and retyped). Contact geometry was recomputed from the same poses under three distance criteria each for metal coordination, polar contacts, and π-stacking. Scores from different scoring functions are never compared with one another, the within-function ligand-versus-control contrast being recomputed each time; Vina does not read receptor partial charges, so charge-only metal variants are inert, and its scores remain scoring-function estimates rather than affinities. Residue-level contact summaries (hydrogen bonds, metal coordination, π-stacking, hydrophobic contacts) were derived

from the best-scoring poses of the same docking runs; raw scores are in Supplementary Data S2.

2.16 Statistical framework and reproducibility

Thresholds (BH $P<0.05$, $|\log_2 FC| \geq 1$, median grouping, selection frequency ≥0.40) and display rules were recorded in dated analysis protocols. Every figure panel reads a fixed result table, and the rendering scripts hard-assert agreement between the plotted values and the fixed manuscript numbers. Analyses used R 4.3.3 (DESeq2, edgeR, limma, clusterProfiler, fgsea, CMScaller, rms, hgu133plus2.db) and Python 3.12 (scikit-survival, lifelines, statsmodels, RDKit, tidepy, networkx, AutoDock Vina 1.2.7). Negative results are reported alongside positive ones.

3 Results

3.1 Icariin compounds and a six-source target-evidence library

The two study compounds occupy complementary drug-like space. Icariin ($C_{33}H_{40}O_{15}$, MW 676.7) violates the Lipinski bounds on molecular weight, TPSA (238.2 $Å^2$ vs. 140), HBD (8 vs. 5) and HBA (15 vs. 10) with a near-zero logP (0.07), as expected for a diglycoside. Its aglycone icaritin ($C_{21}H_{20}O_6$, MW 368.4) falls within all rule bounds (logP 4.09; TPSA 100.1 $Å^2$; HBD 3; HBA 6) and is evaluated as drug-like by the Ghose, Egan and Muegge filters (Fig. 1a–c). This glycoside-to-aglycone contrast frames icariin as a prodrug-like parent compound whose target library is interpreted at the transcriptomic level (Methods). The curated human target library contained 551 genes from six independent sources (PharmMapper 292, CTD 166, SwissTargetPrediction 99, HERB 79, PubChem BioAssay 28, ChEMBL 9 gene-source pairs; multi-source genes counted once per source), graded A experimental ($n=9$), B curated or multi-source ($n=211$), and C computational prediction ($n=331$) (Fig. 1d).

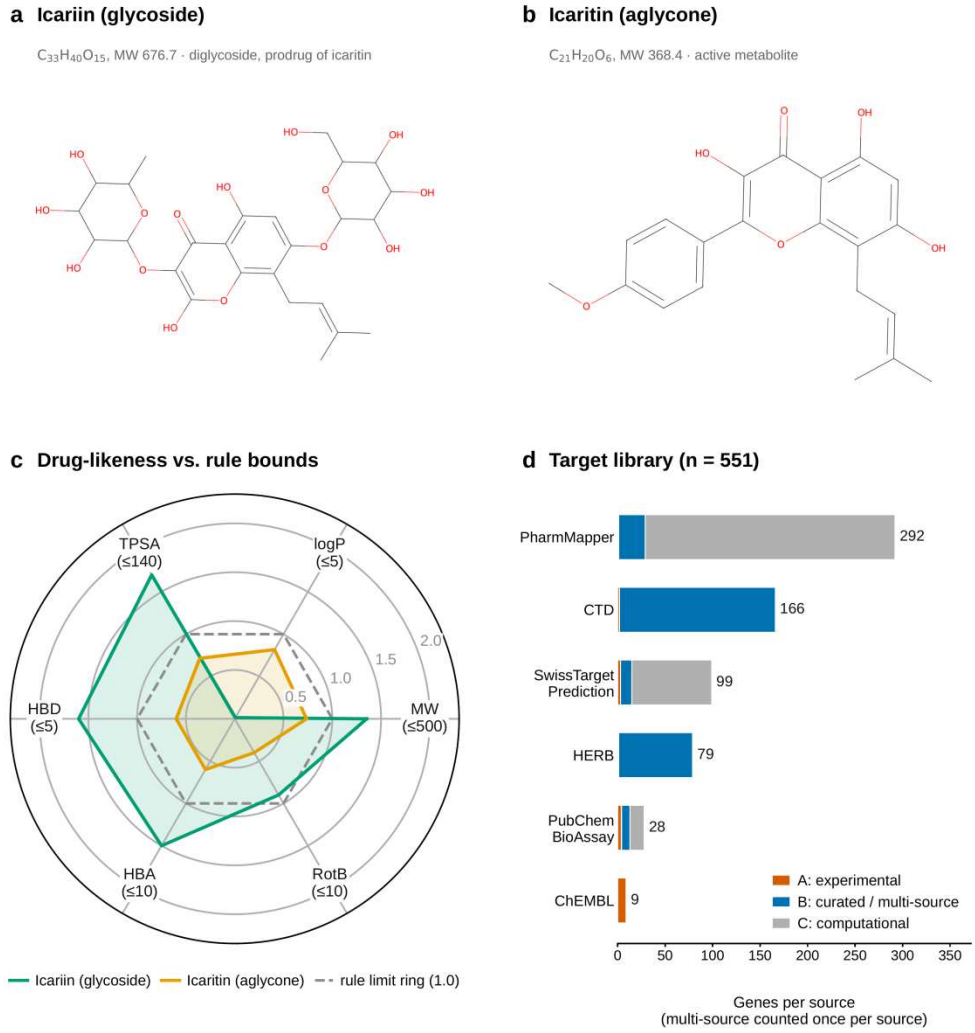


Fig. 1 ***Icariin compounds and target-library composition.*** *(a) Two-dimensional structure of icariin ($C_{33}H_{40}O_{15}$, MW 676.7), the prenylated flavonol diglycoside of Epimedium spp. and prodrug of icaritin. (b) Two-dimensional structure of icaritin ($C_{21}H_{20}O_6$, MW 368.4), the pharmacologically active aglycone. (c) Drug-likeness radar: RDKit-computed physicochemical descriptors normalized to the Lipinski/Veber rule bounds (MW/500, logP/5, TPSA/140, HBD/5, HBA/10, rotatable bonds/10); the grey dashed ring marks normalization value 1.0. Icariin exceeds the bounds on molecular weight, TPSA, HBD and HBA, whereas icaritin lies within all limits. (d) The consolidated 551-gene human icariin target library by evidence source, stacked by evidence grade (A, experimental activity, n=9; B, curated or multi-source, n=211; C, computational prediction, n=331); a gene curated from more than one source is counted once per source*

3.2 An 84-gene candidate pool from the strict three-way intersection

DESeq2 identified 15,305 differentially expressed genes in TCGA-COAD (11,184 up, 4,121 down), 92.5% of which overlapped the edgeR sensitivity analysis (Fig. 2a). In the

independent screening cohort GSE39582, limma identified 1,532 tumour-versus-normal DEGs under the same thresholds (Fig. 2b). The strict three-way intersection of these DEGs with the 551-gene icariin target library defined the 84-gene candidate pool (Fig. 2c). Principal component analysis separated tumour (*n=471*) from normal (*n=41*) samples, and unsupervised clustering of the 84 candidates recapitulated the same dichotomy (a descriptive quality-control observation; PCA is not a formal batch-effect test) (Fig. 2d, e). The pool comprised ECM-remodelling and proteolytic regulators (SPARC, MMP3, MMP9, COL1A1, TIMP1) alongside proliferative markers (MKI67, PLK1, MYC), anticipating the ECM/EMT program detected below.

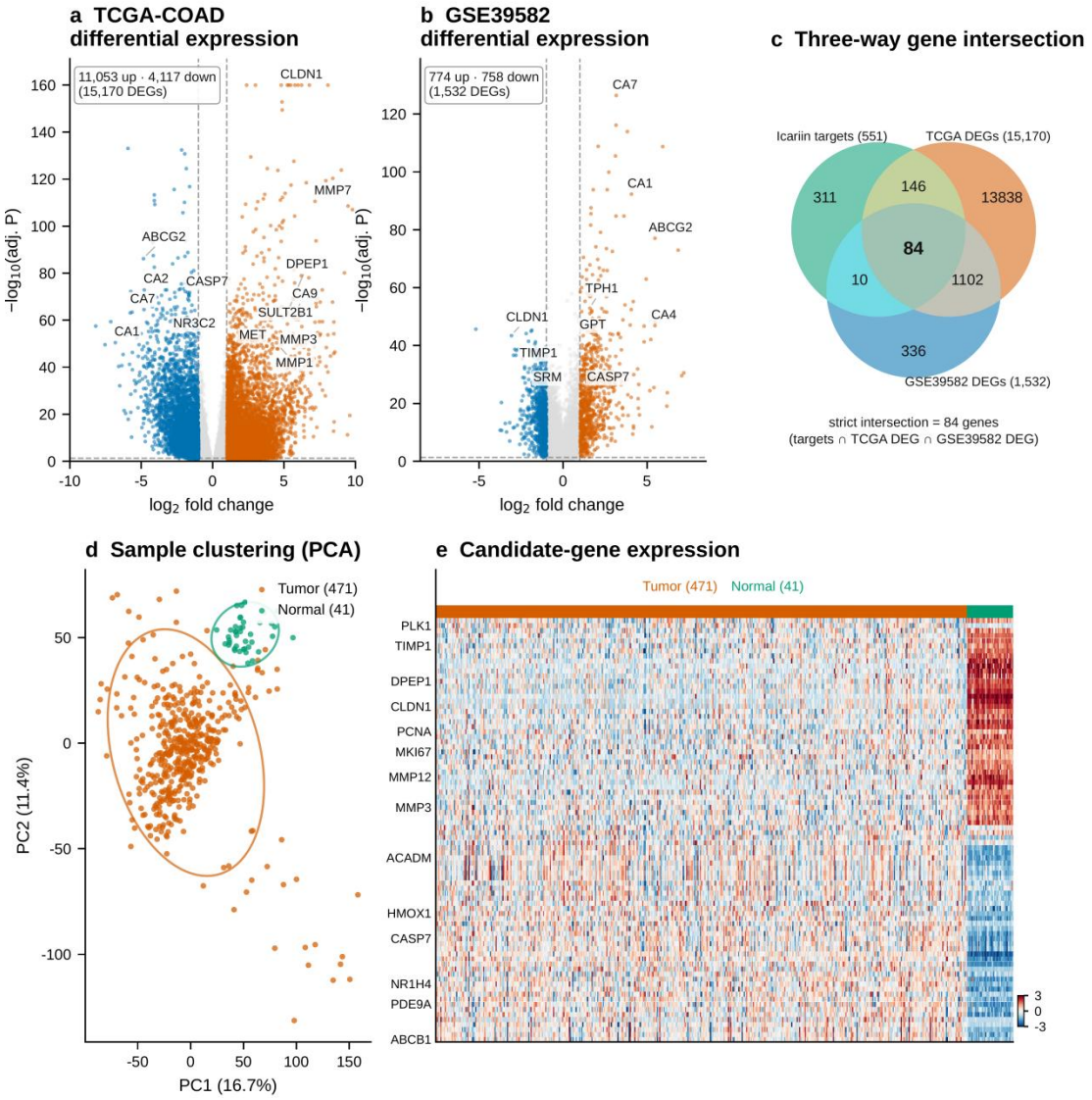


Fig. 2 ***Identification of the 84-gene candidate pool.*** *(a) Volcano plot of TCGA-COAD differential expression (DESeq2; BH-adjusted P<0.05 and |log₂FC|≥1; 15,170 DEGs with mapped gene symbols). (b) Volcano plot of the GSE39582 tumour-versus-normal screening analysis (limma; same thresholds; 1,532 DEGs). (c) Strict three-way intersection of the 551-gene icariin target library with the TCGA-COAD and GSE39582 DEG sets, defining the 84-gene*

candidate pool. (d) Principal-component analysis of TCGA-COAD tumour (n=471) and normal (n=41) samples (top 8,000 variable genes, VST). (e) Unsupervised clustering of the 84 candidates (VST z-scores) recapitulating the tumour/normal separation

3.3 ECM-centred enrichment, PPI architecture, and five core targets

The 84 candidates were dominated by extracellular-matrix remodelling and proteolysis: GO biological-process terms centred on collagen metabolic process and ECM disassembly, molecular-function terms on serine-type endopeptidase activity (362, 30 and 52 significant BP/CC/MF terms overall; representative terms in Fig. 3a), and Reactome/KEGG analysis converged on ECM degradation and matrix-metalloproteinase activation (Fig. 3b). In the STRING v12.0 network (score ≥400), 82 of 84 candidates were mapped and 66 of them carried at least one candidate-only edge (105 edges in total). Greedy modularity resolved these into seven consensus modules (Fig. 3c and its legend): an ECM/stromal module (15 genes, including COL1A1, SPARC, MMP3, MMP7, TIMP1 and CLDN1), a xenobiotic/drug-metabolism module (14, including ABCB1, ABCG2, GSTP1, MYC and MMP12), an inflammation/metastasis module (14, including MMP9, PTGS2, PLAU and MET) and a lipid/redox module (7, including HMOX1 and FABP4), with three smaller modules (energy metabolism, carbonic anhydrases, cell cycle) and a residual five-gene block. Hub ranking by the composite of degree, betweenness, and closeness placed MMP3, TIMP1, PTGS2, COL1A1, MMP7, HMOX1, MMP9 and FABP4 at the top (Fig. 3d). Applying the pre-specified integration rule (experimental or multi-source evidence plus hub-composite ≥0.90 and non-isolated status) declared five core targets: TIMP1, PTGS2, COL1A1, HMOX1, and MMP9. These core targets share exactly one gene with the nine-gene signature (TIMP1), and four of the five correlate significantly with the risk score (TIMP1, COL1A1, and MMP9 positively, PTGS2 inversely; HMOX1 not significant; Fig. 8c); the two gene sets are therefore mechanistically concordant but genomically distinct, and all pharmacological claims below are phrased accordingly (see Discussion).

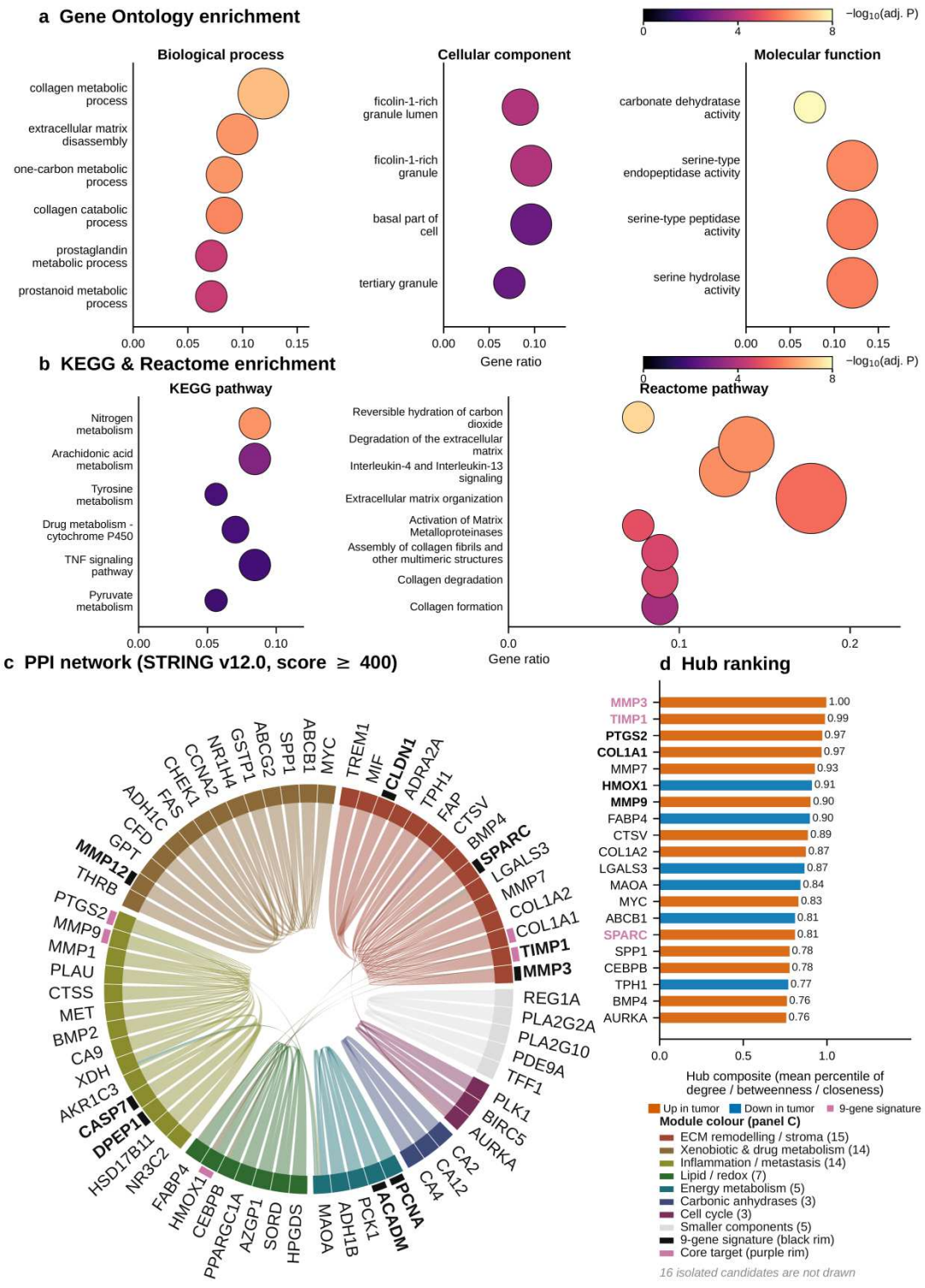


Fig. 3 ***Functional characterization of the candidate pool.*** *(a) Representative Gene Ontology terms (biological process, cellular component, molecular function; BH-adjusted hypergeometric P<0.05; 362/30/52 significant BP/CC/MF terms overall). (b) Representative KEGG and Reactome pathways converging on ECM degradation and matrix-*

metalloproteinase activation. (c) The STRING v12.0 candidate-only interaction network (confidence score ≥400) drawn as a chord diagram: the ring holds the 66 candidates that carry at least one such edge (105 edges in total), grouped into the seven consensus modules plus a residual block of five genes (block colours and gene counts in the legend); ribbon width scales with the STRING combined score, normalised within each gene's sector. Black outer rim, the nine signature genes; purple outer rim, the five core targets. The 16 candidates with no edge at this threshold are not drawn (AHCY, BCHE, CA1, CA7, CDA, EPHX2, IMPA1, LGALS2, MKI67, OTC, QPCT, RNASEL, SIGMAR1, SLCO1B3, SRM, SULT2B1). (d) Hub ranking by the composite of degree, betweenness and closeness centrality; the integration rule defined five core targets (TIMP1, PTGS2, COL1A1, HMOX1, MMP9)

3.4 A 100-combination screen converges on a nine-gene signature and a reproducible risk landscape

Across the *10×10* algorithm-combination grid on the GSE14333 training cohort, the median cross-validated C-index of the 90 non-control combinations was 0.613 (IQR 0.596–0.631); the best non-control combination (univariable filter × Cox-Ridge) reached 0.656, comparable to the no-selection CoxPH control (0.654; mean of the four convergent folds), while the single highest score in the grid was achieved by an `all`-feature model (extra-trees survival forests on the full 84-gene pool; 0.668) (Fig. 4a). This top-scoring entry was excluded from signature construction by the pre-specified consensus rule, which restricts consensus counting to the 90 selection-based combinations: an all-feature model cannot yield a parsimonious signature, and adopting the single best of 100 correlated cross-validation point estimates would import the corresponding selection optimism into the final model. That optimism is quantified by nesting the whole pipeline: re-running selection, consensus and the final elastic-net fit inside 15 held-out folds (outer stratified five-fold, three repeats) gave a nested C-index of 0.60 on average (median 0.62, range 0.48–0.70), against an apparent 0.75 for the same pipeline scored on the cohort that built it (Supplementary Data S5). The exclusion is inconsequential for the external-validation conclusions: across three consensus thresholds and seven final-model variants, re-evaluated under the current external-validation framework, the external C-index in GSE17536 ranged 0.727–0.745 for all Cox-family variants (0.704–0.749 including the random-forest and linear survival-SVM variants) with all Kaplan–Meier separations significant (Supplementary Table S4). Ten genes met the 40% retention threshold; one of them (MKI67) received a zero coefficient after elastic-net shrinkage and is annotated as such, yielding the final nine-gene model: three genes with positive coefficients (TIMP1, SPARC, CLDN1) and six with negative coefficients (PCNA, ACADM, DPEP1, MMP3, CASP7, MMP12), dominated by ECM/stromal components (Fig. 4b). Splitting the 438 TCGA-COAD samples at the median risk score (219 per group; Fig. 4c), deaths accumulated preferentially in the high-risk half across the follow-up window rather than in a short-term burst (Fig. 4d). Over the same sample ordering the three positive-coefficient genes rose and the six negative-coefficient genes fell reciprocally (rows ordered by Spearman correlation with the risk score; Fig. 4e).

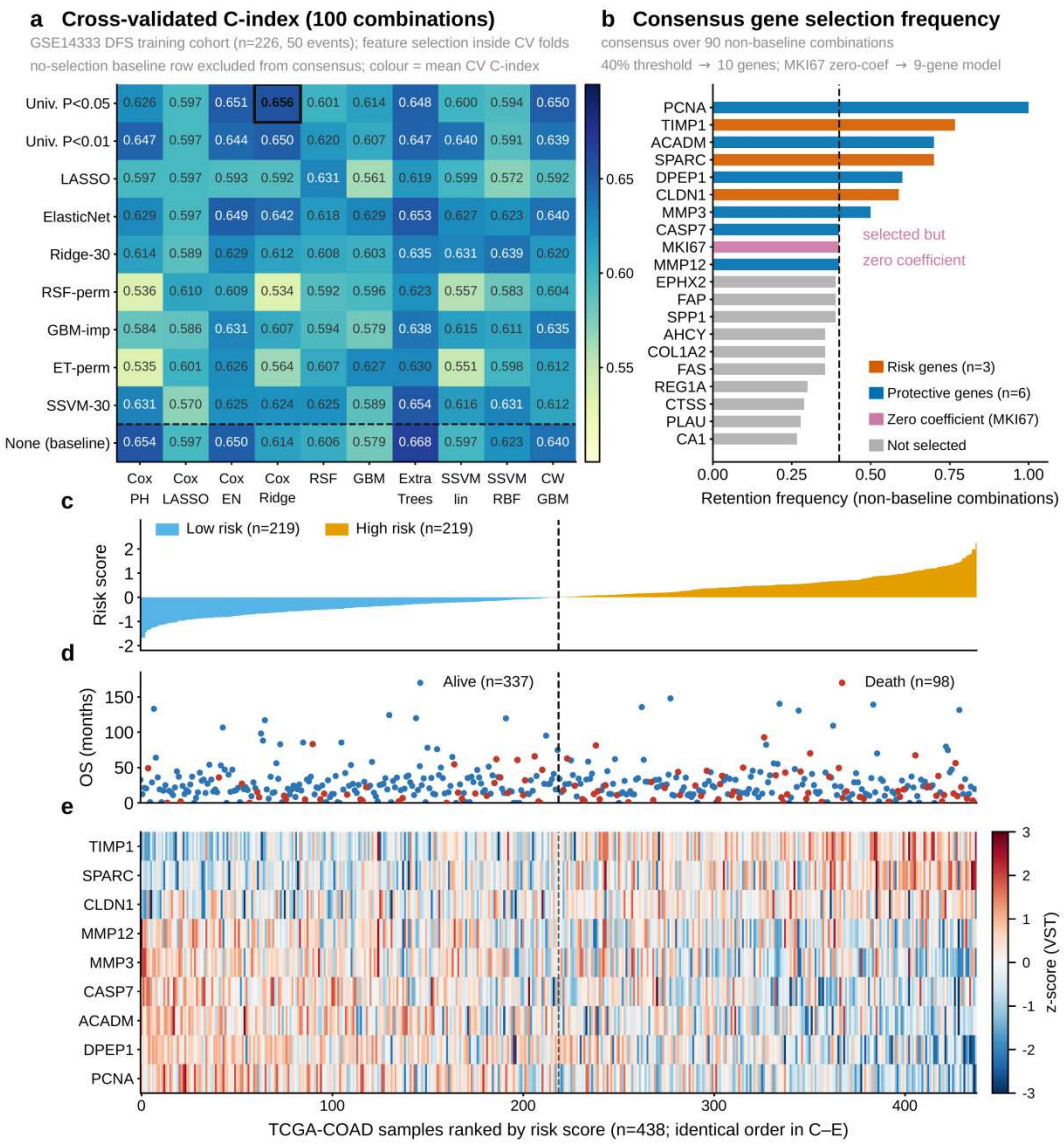


Fig. 4 ***A 100-combination screen converges on the nine-gene consensus signature.*** *(a) Cross-validated C-index across the 10×10 algorithm-combination grid on the GSE14333 training cohort, from a single-level five-fold cross-validation with feature selection inside each training fold (rows: feature selectors; columns: survival models); the no-selection CoxPH control and the excluded top-scoring* `all`*-feature extra-trees model are annotated. (b) Gene selection frequency across the 90 non-baseline combinations (top 20 shown); the dashed line marks the 40% retention threshold, and MKI67 (retained by frequency but shrunk to a zero elastic-net coefficient) is annotated. (c) Risk-score distribution of the 438 TCGA-COAD samples in ascending order, median-split into high- and low-risk groups (219 each). (d) Survival status (OS) over the same ordering, showing the 435 of the 438 samples that carry an OS follow-up time (98 deaths, 337 alive at last follow-up; three samples without a follow-up time are omitted), whereas the OS Kaplan–Meier analysis below uses the 419 samples with complete*

follow-up and the multivariable Cox models use the 411 with complete covariate data. (e) VST z-score heatmap of the nine signature genes over the same ordering, rows ordered by Spearman correlation with the risk score

3.5 Internal transfer and clinical characterization in TCGA-COAD

The TCGA-COAD analysis is an internal transferability check; the primary prognostic evidence is the external DFS validation below. The median-grouped signature stratified overall survival in TCGA-COAD (OS $n=419$; 94 events; log-rank $P=0.0017$; Fig. 5a). In the multivariable Cox model adjusting for composite stage, age, and sex ($n=411$), the risk score remained significant (HR 1.27 per SD, 95% CI 1.03–1.57, $P=0.024$), with composite stage the dominant covariate (HR 2.46 per level, 95% CI 1.90–3.18, $P<0.001$) (Fig. 5b). The OS effect was weaker than the external DFS effect and the Kaplan–Meier separation was driven by an absolute excess of early events in the high-risk group; whether these early deaths are recurrence-related is speculative because cause-of-death information was not available. The four-variable model — a joint model of the risk score with composite stage, age, and sex — reached time-dependent AUCs of 0.782/0.801/0.741 at 1/3/5 years (Fig. 5d); the score-alone analysis is provided in Supplementary Fig. S1 (AUCs 0.610/0.568/0.615). Calibration was adequate at 1 and 3 years; at 5 years predicted survival overestimated observed survival in the mid-to-upper risk quartiles, consistent with sparse late follow-up (Fig. 5e). The 3-year clinical impact curve (Fig. 5f) reports how many patients the four-variable model classifies as high-risk at each threshold and how many observed deaths fall within that group. In decision-curve analysis with the three models refitted in one common cohort and net benefit weighted for censoring (Supplementary Fig. S2; Supplementary Table S12), the clinical model of stage, age and sex kept a positive net benefit relative to "treat none" across the whole 0.01–0.50 range. Adding the score changed net benefit by at most $+0.014$ (at a threshold of 0.17) and lowered it beyond about 0.25. The decision curve therefore shows the same modest, stage-dominated signal. The nomogram confirmed stage as the dominant contributor (0–100 points) ahead of the risk score (0–58 points per SD) and age (0–78 points over the observed age range) (Fig. 5c). Internally, adding the signature to the clinical model did not significantly improve discrimination for OS (Harrell $\Delta C=+0.016$, $P=0.050$ — two-sided, 25 of 1,000 bootstrap resamples with $\Delta C \leq 0$, at the pre-specified threshold and treated as not significant; 36-month IDI -0.004, $P=0.738$; 36-month NRI $+0.054$, $P=0.590$), for PFI ($\Delta C=+0.006$, $P=0.200$), or for DFI ($\Delta C=+0.024$, $P=0.464$); the internal baseline additionally includes age and sex.

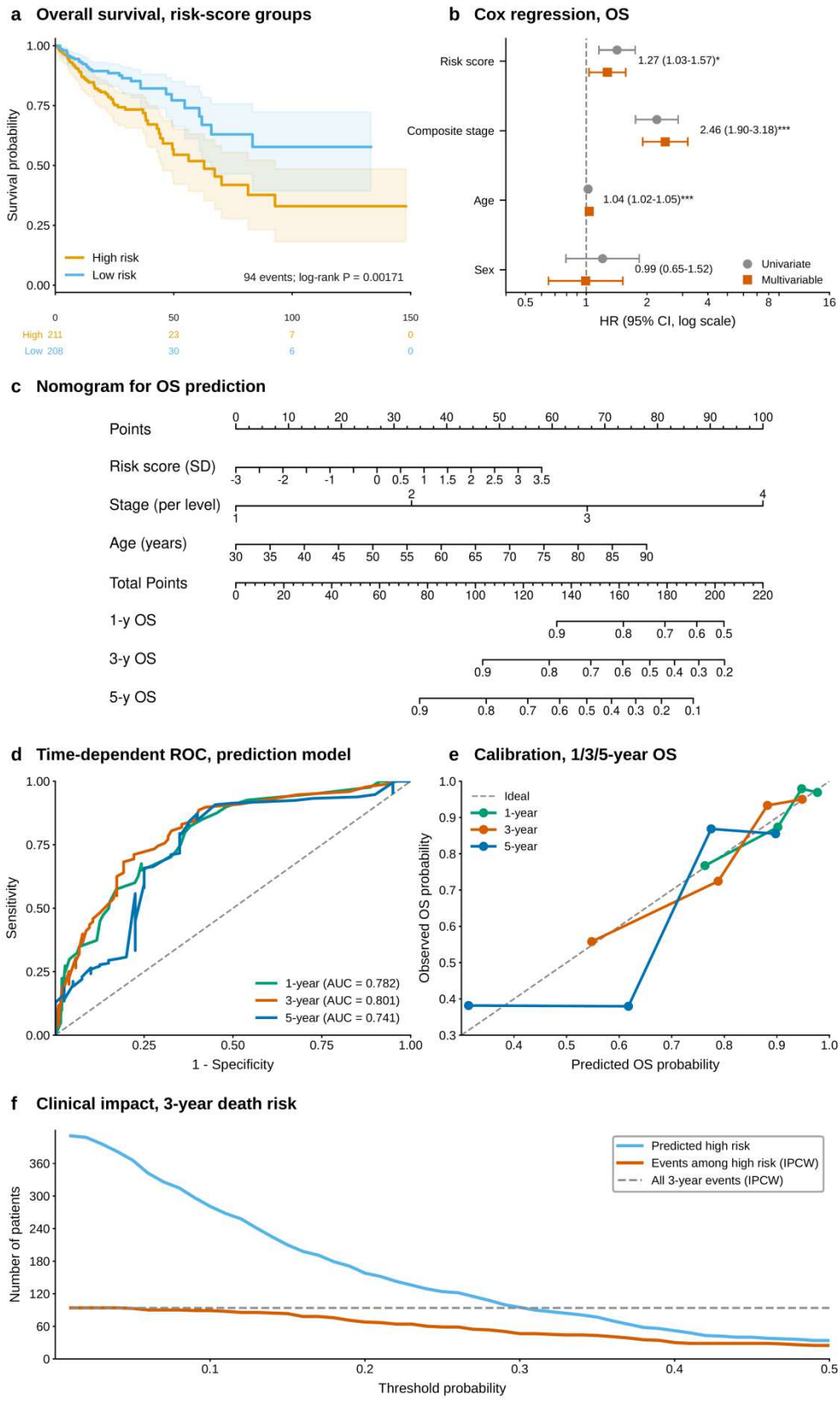

a Overall survival, risk-score groups
Survival probability
1.00
0.75
0.50
0.25
0.00
High risk
Low risk
94 events; log-rank P = 0.00171
0
50
100
150
High 211
23
7
0
Low 208
30
6
0
b Cox regression, OS
Risk score
1.27 (1.03-1.57)*
Composite stage
2.46 (1.90-3.18)***
Age
1.04 (1.02-1.05)***
Sex
0.99 (0.65-1.52)
Univariate
Multivariable
0.5
1
2
4
8
16
HR (95% CI, log scale)
c Nomogram for OS prediction
Points
0 10 20 30 40 50 60 70 80 90 100
Risk score (SD)
-3 -2 -1 0 0.5 1 1.5 2 2.5 3 3.5
Stage (per level)
1 2 3 4
Age (years)
30 35 40 45 50 55 60 65 70 75 80 85 90
Total Points
0 20 40 60 80 100 120 140 160 180 200 220
1-y OS
0.9 0.8 0.7 0.6 0.5
3-y OS
0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2
5-y OS
0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1
d Time-dependent ROC, prediction model
Sensitivity
1.00
0.75
0.50
0.25
0.00
1-year (AUC = 0.782)
3-year (AUC = 0.801)
5-year (AUC = 0.741)
0.25
0.50
0.75
1.00
1 - Specificity
e Calibration, 1/3/5-year OS
Observed OS probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Ideal
1-year
3-year
5-year
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Predicted OS probability
f Clinical impact, 3-year death risk
Number of patients
360
300
240
180
120
60
0
Predicted high risk
Events among high risk (IPCW)
All 3-year events (IPCW)
0.1
0.2
0.3
0.4
0.5
Threshold probability

Fig. 5 ***Internal transfer and clinical utility in TCGA-COAD (overall survival).*** *(a) Kaplan–Meier overall survival of median-split risk groups (n=419; 94 events; log-rank P=0.0017) with numbers at risk. (b) Univariable and multivariable Cox regression for the risk score, composite stage, age, and sex (n=411). (c) Nomogram of the multivariable OS model (composite stage, risk score, age; sex, retained but negligible — HR 0.99, P=0.98 — is omitted), constructed with the R* `rms` *package; predicted 1-, 3-, and 5-year OS are read from the total-points axis. (d) Uno time-dependent ROC of the clinical prediction model (linear predictor) at 1, 3, and 5 years (AUC 0.782/0.801/0.741). (e) Calibration of predicted versus observed OS at 1, 3, and 5 years by quartiles of the 3-year predicted OS; the 60-month point is affected by the ≈2-year median follow-up and is shown as reference only. (f) Clinical impact curve for 3-year death risk: at each threshold, the number classified high-risk by the four-variable model, the weighted deaths among them, and total deaths (n=411)*

3.6 External DFS validation: primary cohort, pooled multi-cohort evidence, and incremental value over stage, age and sex

In the external validation cohort GSE17536 (*n=145*, 36 events), the transferred signature stratified disease-free survival (log-rank $P=2.0\times10^{-4}$; HR 2.30 per SD, $P=2.3\times10^{-8}$) with a Harrell C-index of 0.734 (95% CI 0.657–0.810) and 1/3/5-year time-dependent AUCs of 0.751/0.755/0.743 (Fig. 6a–c); the training cohort GSE14333 attained an apparent C-index of 0.750. Adding the signature to a clinical baseline of stage, age and sex improved discrimination in GSE17536: the full-follow-up Harrell *ΔC=+0.085* (95% CI 0.023–0.169, *P=0.004*), the likelihood-ratio test for adding the score to that baseline $\chi^2=25.6$ (*P<0.001*), and the 36-month IPCW Brier score 0.147 versus 0.173 for the baseline alone (Fig. 6d; Tables 2 and 3). The 36-month Uno IPCW IDI and continuous NRI are reported as secondary measures in Table 3. Against a baseline refitted inside this cohort, calibration was acceptable in three of four predicted-probability quartiles (the lowest quartile somewhat optimistic, predicted 0.504 versus observed 0.427; calibration slope 1.00, 36-month observed-to-expected 0.98). The Brier-score improvement held at 60 months as well (0.155 versus 0.183). Both come from that locally refitted model, not from the transport analysis below.

Table 2 External validation of the nine-gene signature in the GSE17536 DFS cohort

Cohort (endpoint)	*n* (events)	C-index (95% CI)	*P* (log-rank)	HR/SD (95% CI)	*P* (HR)	AUC*
GSE14333 (training; DFS)	226 (50)	0.750	*<0.001*	2.43	*<0.001*	0.796
GSE17536 (validation; DFS)	**145 (36)**	**0.734 (0.657–0.810)**	2.0×10^{-4}	**2.30 (1.72–3.08)**	2.3×10^{-8}	**0.755**

*Time-dependent AUC at 36 months (Uno IPCW). GSE17536 is the primary external validation cohort (DFS); GSE14333 is the training cohort shown as an internal reference, not an external validation. The further independent cohorts pooled with GSE17536 are listed in Table 1.

Table 3 Incremental predictive value of the nine-gene signature (1,000 bootstrap replicates). ΔC is the full-follow-up Harrell concordance increment; IDI and NRI are 36-month Uno inverse-probability-of-censoring-weighted measures. For each measure, the value, its 95% bootstrap CI (parentheses), and the bootstrap P are shown on successive lines; the bootstrap P is two-sided, computed as twice the smaller tail proportion of the resampled statistic relative to zero

Cohort (endpoint)	Baseline	*ΔC*	IDI	NRI
External DFS (GSE17536; *n=145*, 36 events)	stage/age/sex	+*0.085* (0.023–0.169) *P=0.004*	+*0.141* (0.048–0.269) *P<0.001*	+*0.653* (0.250–1.135) *P<0.001*
TCGA-COAD OS (*n=411*)	stage/age/sex	+*0.016* (*0.000–0.044*) *P=0.050*	−*0.004* (*−0.013–0.021*) *P=0.738*	+*0.054* (*−0.245–0.426*) *P=0.590*
TCGA-COAD PFI (*n=410*)	stage/age/sex	+*0.006* (*−0.003–0.026*) *P=0.200*	+*0.001* (*−0.005–0.019*) *P=0.768*	+*0.068* (*−0.212–0.351*) *P=0.600*
TCGA-COAD DFI (*n=175*)	stage/age/sex	+*0.024* (*−0.016–0.095*) *P=0.464*	+*0.003* (*−0.005–0.045*) *P=0.488*	+*0.211* (*−0.411–0.775*) *P=0.484*

All increments use a baseline of stage with age and sex; the external row also reports the likelihood-ratio χ^2 for adding the score to that baseline (25.6, *P<0.001*) and the IPCW Brier score at 36 months (0.147 with the score versus 0.173 for the baseline alone). DFI is defined only for patients without metastasis at diagnosis, so the analysis set covers stages I–III (*n=175*, 22 events).

A robustness battery supported these results (Supplementary Data S6, S7). An expression-fingerprint audit found no sample overlap with GSE14333, and excluding the 14 GSE17536 samples with the highest cross-cohort correlation left performance unchanged (C-index 0.751, HR 2.51 per SD). Scoring the external cohort with the training-cohort median threshold transferred as a fixed cut-off (no re-medianing) gave 41 high-risk / 104 low-risk patients (log-rank $P=1.6\times10^{-6}$). The signature remained significant restricted to stage I–III (*n=135*, 32 events; C-index 0.718), with stage entered as a categorical covariate (risk HR 2.21 vs. 2.19), and with the same increment over stage, age and sex as reported above. Proportional hazards held for the risk score (Schoenfeld *P=0.095*) and stage (*P=0.355*), and the 36-month observed-to-expected ratio against a within-cohort fitted clinical baseline was 1.073 (0.785–1.360). Transporting the training baseline instead of refitting within the validation cohort (Supplementary Fig. S8) gave a calibration slope of 1.03 (95% CI 0.75–1.40), so the transferred relative risk shows no evidence of over-fitting, whereas absolute risk was under-predicted: observed-to-expected ratios were 1.39 (0.53–2.41), 1.90 (1.34–2.45) and 1.79 (1.32–2.24) at 12, 36 and 60 months. Adding adjuvant-chemotherapy exposure to the training base model changed neither the risk-score HR (2.16 → 2.16) nor the increment (*ΔC=−0.0005*).

A final pre-specified check asked whether this external performance could be reproduced from an arbitrary 84-gene pool. One hundred and one unrelated pools and the real pool were run through the same *full* 100-combination protocol. The null pools gave external C-index median 0.689 (IQR 0.658–0.715, 95th percentile 0.749); the real pool scored 0.735, the 88th percentile (empirical $P=0.118$). Two readings obtained under earlier protocol variants are withdrawn, with their provenance recorded in the Supplementary Methods (protocol revisions). A permuted-outcome control returned external C-index 0.50, giving no indication of outcome leakage; a negative control of this kind complements, but does not replace, a full audit of the processing pipeline. Transcriptome-wide prognostic signal in colon cancer is therefore dense enough that an arbitrary 84-gene panel propagates to external DFS discrimination of the same order, so this screen does not establish target-library specificity; the signature-specific evidence rests on its increment over stage and on the single-cell and pharmacological localisation (Supplementary Data S7).

Three further screening rounds searched for additional independent cohorts under the protocol described in the Methods; every candidate is reported regardless of outcome (Supplementary Data S7; Supplementary Tables S6, S7). The fingerprint audit proved decisive. GSE161158, the strongest performer on naive metrics in that round ($n=191$ primary tumours, 46 events; C-index 0.721), shared 66 of its 250 deposited samples with GSE17536 itself, so it cannot serve as independent validation. It stayed positive in every de-duplicated subset: $n=66$, 15 events under the unified-panel audit (C-index 0.716; HR 2.61 per SD, $P=3.5\times10^{-4}$), and $n=143$ and $n=99$ under the earlier pass at $\rho\geq0.90$ and $\rho\geq0.85$ (HR 2.40 and 2.53). GSE37892 ($n=130$, 37 events) was direction-consistent but weaker (C-index 0.589; HR 1.38, 95% CI 1.01–1.86, $P=0.040$), and GSE71222 deposited no time-to-event field and was not evaluable. Widening the screen to every further public colorectal series identified six more with an evaluable endpoint, of which one independent cohort reproduced the result at the declared significance level: GSE38832 ($n=92$, 9 events; C-index 0.706, bootstrap 95% CI 0.519–0.875; HR 2.08, $P=0.042$; locked-threshold KM $P=0.012$). GSE29621 produced the largest naive effect (C-index 0.812) but re-analyses the Vanderbilt GSE17536 series: 63 of its 65 samples are the same patients, matched one-to-one at $\rho\geq0.90$ under the unified panel, leaving a duplicate-free remainder of two patients, so it cannot stand as a ninth cohort. The remaining resected cohorts were direction-consistent without reaching significance (nine cohorts, C-index 0.525–0.648; Supplementary Tables S7). The AMC stage II cohort, deposited as GSE33114 whose primary-tumour subset is the same patients as GSE33113 and is therefore scored once ($n=89$, 18 events), reached a significant Cox effect (C-index 0.648; HR 1.55 per SD, $P=0.027$) but a locked-threshold KM $P=0.079$, which the pre-declared rule scores as direction-consistent. An exhaustive third round read 32 further series, ten of which proved evaluable; **no new cohort met the pre-declared criterion**, its strongest result (GSE12945, C-index 0.716, five events) being too event-poor to count by the six-event floor applied throughout.

Outside resected disease the signature does not transfer. Across the two parts of a first-line FOLFOX/FOLFIRI trial (GSE72968, GSE72969) the fixed score fell at or below chance in all four runs (PFS C-index 0.453 and 0.450; OS 0.495 and 0.489). In GSE72968 PFS the locked-threshold split was nominally significant in the *reverse* direction ($P=0.019$, high-risk group faring better), a failure of the score in a below-0.5 cohort, recorded as one. In a liver-

metastasis resection cohort (GSE159216, 77 events) the point estimate was 0.555 with a confidence interval spanning 0.5, and the neoadjuvant setting yields two weakly positive and one null result (GSE87211 0.577; GSE40492 0.575; GSE56710 0.464). Fingerprint audits were clean for every other cohort evaluated (largest pairwise ρ 0.60–0.76). The signature therefore carries prognostic information in resected stage I–III disease, consistent with its stromal/CMS4 phenotype, but shows no transferability to metastatic or perioperative-systemic-therapy populations.

Pooled across the independent cohorts. Individually these cohorts are underpowered — six of the eight carry fewer than 30 events — so we summarised across them rather than selecting among them. A cohort enters the pooling only if it satisfies the pre-specified design criteria (Methods); the decision to aggregate was taken after the individual cohorts had been read and is reported as such. The unified-panel audit, turned on the pooled set itself, found 77 samples of the already deduplicated GSE161158 subset still matching a pooled GSE17536 or GSE17537 sample one-to-one at $\rho \geq 0.90$. The primary analysis therefore excludes every sample involved in such a pair, leaving eight cohorts and 181 DFS/RFS events with each patient counted once (Table 1). A random-effects model using restricted maximum likelihood for τ^2 and a Hartung–Knapp interval for the pooled effect gave an HR of 1.53 per SD (95% CI 1.13–2.07, $P=0.013$; Fig. 6e, Table 1), with seven of the eight cohorts pointing in the same direction; the one that did not is GSE106584 (52 events, HR 0.99). Heterogeneity is substantial ($I^2=68.6\%$, $\tau^2=0.090$) and the 95% prediction interval for a new cohort includes 1 (0.71–3.31). Leave-one-out re-estimates span HR 1.40–1.65 and all eight remain significant ($P \leq 0.038$; Supplementary Fig. S7). Estimator and set-size sensitivity analyses move none of this: the DerSimonian–Laird estimator with a normal-theory interval, reported as a comparator, gives 1.53 (1.17–2.01; $I^2=71.7\%$); a fixed-effect model gives 1.49 (1.30–1.71); removing both members of every duplicate pair gives 1.55 (1.17–2.03; 166 events); and adding back the two cohorts that fell outside the design criteria gives 1.51 (1.20–1.92; $k=10$, 223 events; Supplementary Table S8). The pooled effect is smaller than the training-cohort (2.43 per SD) and primary external (2.30) estimates, the shrinkage expected on transfer across platforms and endpoints; we report it as the more realistic summary.

A fourth and final round closed the search surface rather than widening it once more: the GEO DataSets index was re-queried with complementary MeSH-based and recurrence/follow-up term sets, and the 368 native ArrayExpress/BioStudies studies, all 64 human GPL570 colorectal series with ≥50 samples and the 15 GPL570 datasets of E-MTAB-6698 were swept in parallel (Supplementary Data S7). **None of these yielded a new evaluable cohort.** Together the four rounds close the search surface for public cohorts meeting the study's four design conditions.

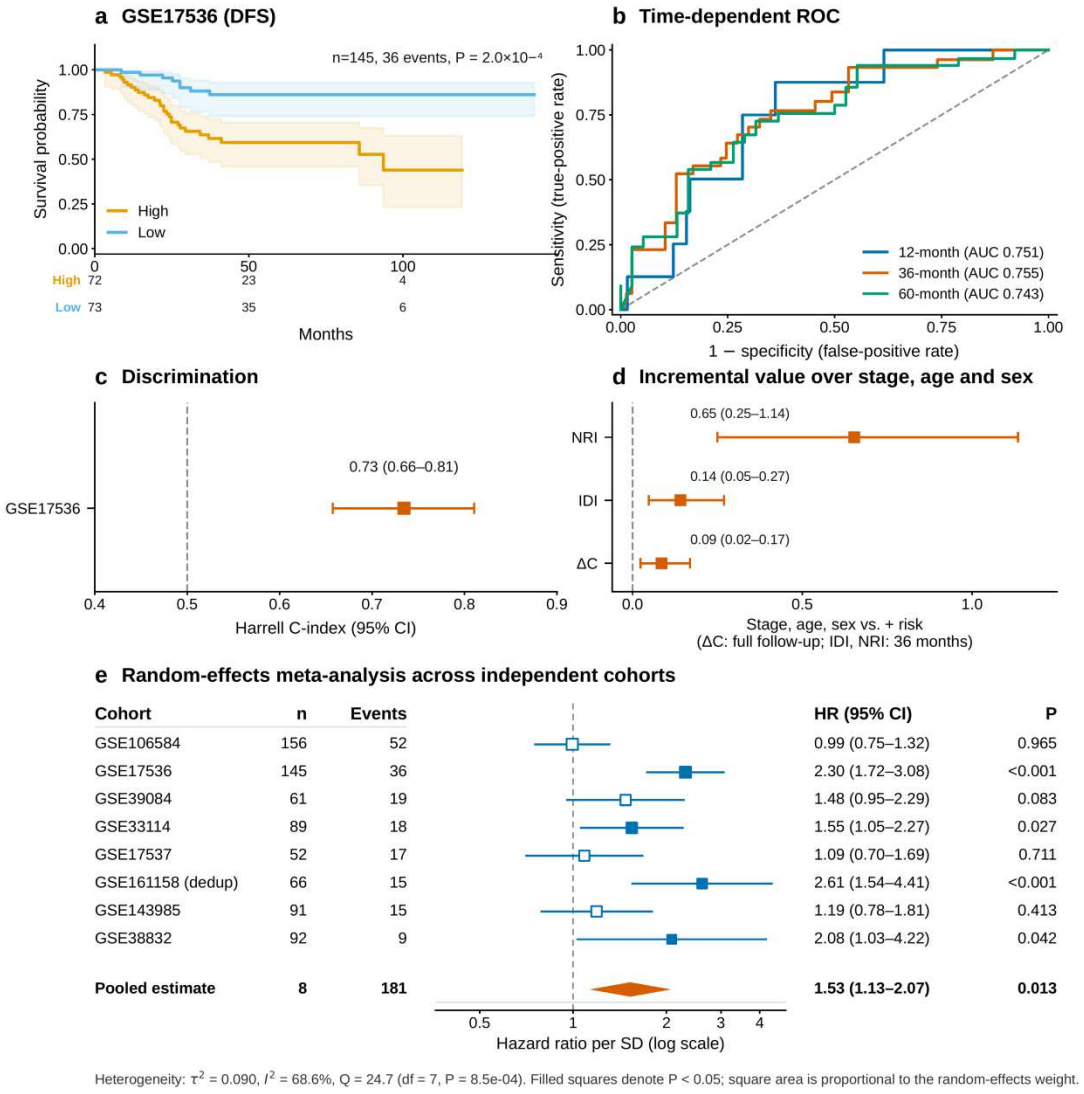


Fig. 6 ***External validation of the nine-gene signature.*** *Panels a–d use the primary external-validation cohort, GSE17536 DFS (n=145, 36 events), with the fixed GSE14333 μ/σ transfer. (a) Disease-free survival by median split (numbers at risk below the axis). (b) Time-dependent ROC at 12, 36, and 60 months (Uno IPCW; AUCs annotated). (c) Harrell C-index with bootstrap 95% CI (1,000 resamples). (d) Incremental value over stage, age and sex: ΔC is the full-follow-up Harrell increment; the 36-month Uno IPCW IDI and continuous NRI are shown as secondary measures (1,000 bootstrap replicates). (e) Random-effects meta-analysis of the eight cohorts carried into the pooled analysis after removal of every sample implicated in a duplicate pair, each satisfying the four design conditions and scored with the frozen coefficients and training μ/σ (181 DFS/RFS events in total; Table 1 gives cohort identities, Supplementary Table S8 the sensitivity analyses). Squares are cohort hazard ratios per SD with 95% CIs, sized by random-effects weight and filled when P<0.05; the diamond is the pooled estimate (restricted maximum likelihood, Hartung–Knapp interval; 1.53, 95% CI 1.13–2.07). Leave-one-out re-estimates: Supplementary Fig. S7*

3.7 The high-risk phenotype is mesenchymal (CMS4), replicated externally, and not associated with mutation burden

High-risk tumours were strongly enriched for the CMS4/mesenchymal subtype (50.2% vs. 14.0% in high- vs. low-risk; $P=2.6\times10^{-15}$, Fisher's exact test; $n=403$), with reciprocal enrichment of CMS2 (21.2% vs. 44.5%, $P=8.7\times10^{-7}$) and CMS3 (11.3% vs. 22.0%, $P=4.8\times10^{-3}$) in the low-risk group and no difference in CMS1 ($P=0.61$) (Fig. 7a). The CMS4 skew replicated independently in GSE39582 ($n=566$, median-split 283/283): CMS4 50.5% vs. 6.0% ($P=1.4\times10^{-34}$), CMS3 6.0% vs. 27.6% ($P=3.3\times10^{-12}$), CMS2 27.9% vs. 44.9% ($P=3.8\times10^{-5}$), and CMS1 not significant ($P=0.083$) (Fig. 7b). In contrast, no significant association with genome instability was observed: MSI-H proportion (17.1% vs. 22.8%, $P=0.175$), TMB (median 3.82 vs. 3.97 mut/Mb; Mann–Whitney $P=0.914$, $n=393$), and the MANTIS MSI score ($P=0.50$) did not differ between risk groups (Fig. 7c–e). In the mutation landscape of the top 30 recurrently mutated genes (cBioPortal PanCancer Atlas cohort; truncating vs. missense/other; Fig. 7f), no gene reached significance after BH correction (per-gene Fisher tests), i.e. the risk stratification captures expression programs rather than a specific mutational driver. Clinically, high-risk tumours were enriched in higher T, stage, and N categories (T $P=2.3\times10^{-4}$; stage $P=0.019$; N $P=0.036$; $n=364$), with no association for M, age, or sex. Because the score is so strongly CMS4-skewed, we asked next whether it is merely a CMS4 read-out. Within CMS4 it still ranked survival, reaching nominal significance where that stratum is adequately powered (in GSE39582, a screening cohort used here merely as a powered sample — RFS, $n=143$, 55 events: $C-index$ 0.593, HR 1.47 per SD, $P=0.0034$; OS, $n=158$, 65 events: 0.614, HR 1.45, $P=0.0031$), and the direction held without significance in the event-poorer TCGA-COAD CMS4 stratum ($n=120$, 31 events: 0.605, HR 1.35, $P=0.13$). Within-CMS4 progression-free interval, the event-poorest endpoint, was not discriminated ($C-index$ 0.543, HR 1.10, $P=0.57$). Its increment over a model already containing stage, age, sex and the CMS class was nevertheless marginal: on TCGA-COAD OS, $\Delta C=+0.014$ ($P=0.084$), continuous NRI 0.049 ($P=0.77$), IDI 0.001 ($P=0.85$), with only the likelihood-ratio test and the per-SD hazard ratio nominally positive (likelihood-ratio $P=0.024$; HR 1.34 per SD, $P=0.023$). The score thus carries prognostic information that is not reducible to subtype, but little of it is independent of standard clinical variables once CMS is known (Supplementary Data S6).

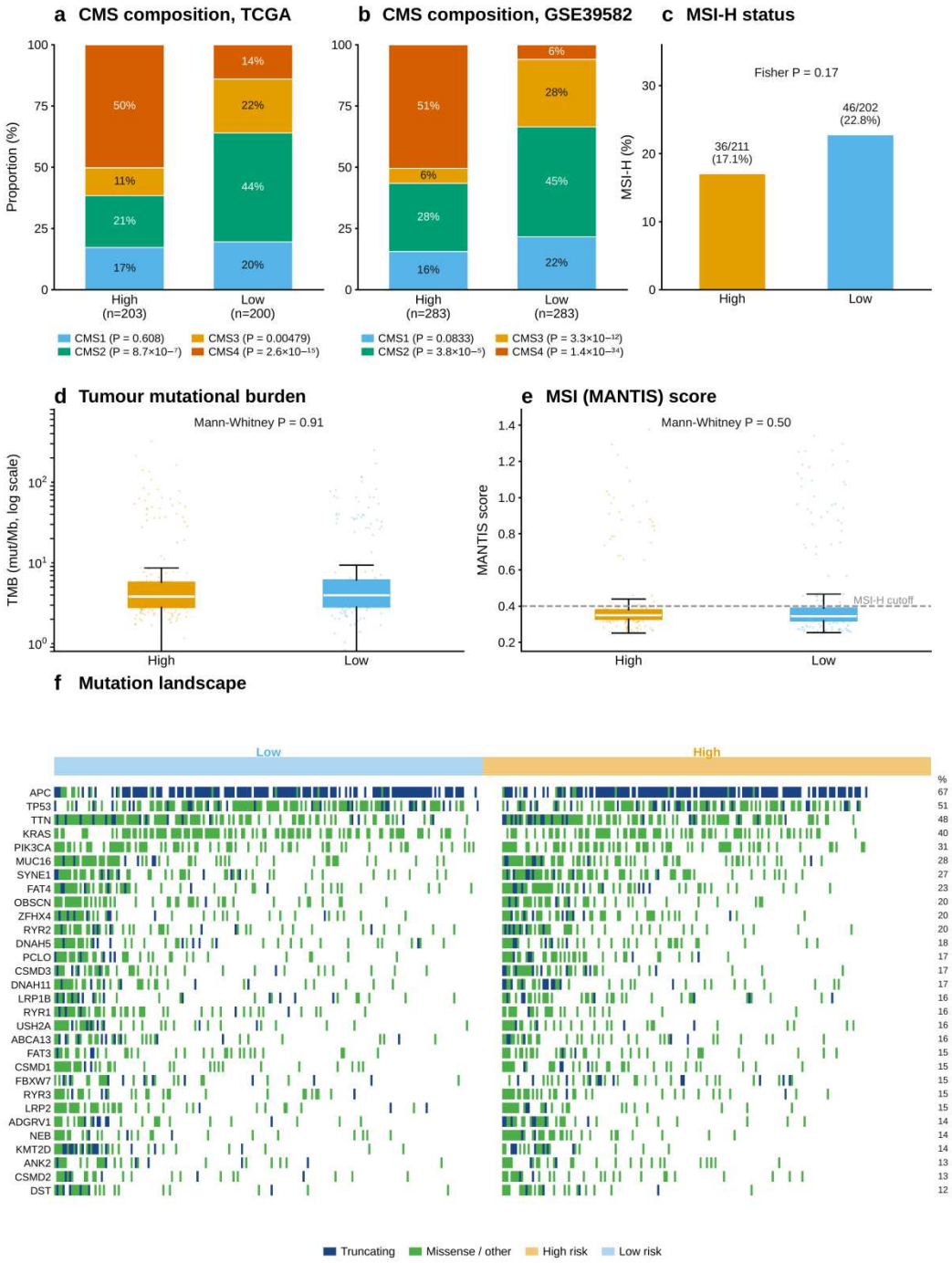


Fig. 7 ***High-risk phenotype: consensus molecular subtype, MSI/TMB, and mutation landscape.*** *(a) CMS composition of high- versus low-risk TCGA-COAD tumours (CMScaller, n=403; Fisher's exact test per subtype). (b) CMS composition in the independent GSE39582 replication cohort (n=566, median-split 283/283). (c) MSI-H fraction (MANTIS score >0.4, n=413). (d) Tumour mutation burden (median 3.82 vs. 3.97 mut/Mb; Mann–Whitney P=0.914; n=393). (e) MANTIS MSI score (n=413). (f) Mutation landscape of the top 30 recurrently mutated genes (cBioPortal PanCancer Atlas 2018 snapshot; truncating versus*

missense/other); no gene differed between risk groups after BH correction (per-gene Fisher tests)

3.8 Pathway-level mechanism: EMT/stromal activation with oxidative phosphorylation suppression

Pre-ranked GSEA on genes ranked by their Spearman correlation with the continuous risk score identified EMT as the pathway most strongly enriched in high-risk tumours (NES +*3.14*; a de-overlap re-analysis excluding the three signature genes contained in the EMT set gave NES +*3.15*, rank unchanged), together with myogenesis, apical-junction, and angiogenesis programs, whereas E2F targets, MYC targets, and oxidative phosphorylation were depleted (NES −*3.95*, −*3.89*, and −*3.84*), consistent with a non-proliferative, mesenchyme-driven tumour (Fig. 8a; KEGG representative rows in Fig. 8b). Among the five core targets, COL1A1 and TIMP1 correlated most strongly with the risk score (ρ=+*0.484* each; BH $P=1.1\times10^{-26}$) and MMP9 more weakly (+*0.156*; BH $P=1.3\times10^{-3}$). PTGS2 showed the opposite direction (−*0.156*; BH $P=1.3\times10^{-3}$), and HMOX1 was not associated (+*0.06*, not significant). All five are reported without selection (Fig. 8c). Only TIMP1 is shared between the two sets, a deliberate separation of the prediction tool from the mechanistic hypothesis. Across the 10 mechanism modules (all reported unfiltered), EMT was the strongest correlate (ρ=+*0.455*) and the stromal score followed (+*0.331*). Oxidative phosphorylation was negatively correlated (Hallmark −*0.366*; KEGG −*0.254*), the ferroptosis driver and suppressor arms rose modestly (+*0.218* and +*0.194*; reported as overall pathway activation rather than a pro-ferroptotic shift), and the T-cell score did not correlate (−*0.038*, not significant) (Fig. 8d–f).

The negative coefficients of MMP3 and MMP12 prompted a formal test of whether the risk score partly reads out tumour stromal content rather than a purely malignant-epithelial program (Supplementary Fig. S4). The score correlated positively with the ESTIMATE stromal score in both cohorts (Spearman ρ=+*0.33* and +*0.49*; Benjamini–Hochberg $P<0.001$), with fibroblast MCP marker-sum scores (ρ=+*0.43* and +*0.46*) and endothelial cells (ρ=+*0.30* and +*0.30*; both $P<0.001$), and not with the ESTIMATE immune score (ρ=+*0.06*, $P=0.18$ and ρ=+*0.13*, $P=0.052$). In multivariable Cox models the score retained an adjusted hazard ratio of 1.27 per SD in TCGA-COAD after stromal adjustment ($P=0.029$), but its increment over the clinical base model plus the stromal score was small and borderline (ΔC=+*0.013*, bootstrap 95% CI −*0.000–0.041*, $P=0.052$; over the base model alone, +*0.016*, 95% CI *0.000–0.044*, $P=0.050$). In GSE14333, the apparent (in-sample) increment over base plus stromal remained significant (ΔC=+*0.100*, 95% CI *0.040–0.162*, $P<0.001$; hazard ratio 2.16 → 2.44 per SD). The same comparison in the external DFS cohort, where ESTIMATE and MCP marker-sum scores are computable on the official full-probe matrix, gave a risk-score–stromal Spearman ρ=+*0.493* (MCP fibroblasts +*0.449*, endothelial cells +*0.405*) and retained an increment over stage plus the stromal score ($\Delta C'$=+*0.072*, bootstrap 95% CI 0.010–0.145, $P=0.016$; in the joint model the risk score HR =*2.53*, $P=2.3\times10^{-6}$, while the stromal score was not significant, HR *0.81*, $P=0.26$). Stromal content is therefore a major carrier of the risk signal, but not its sole carrier.

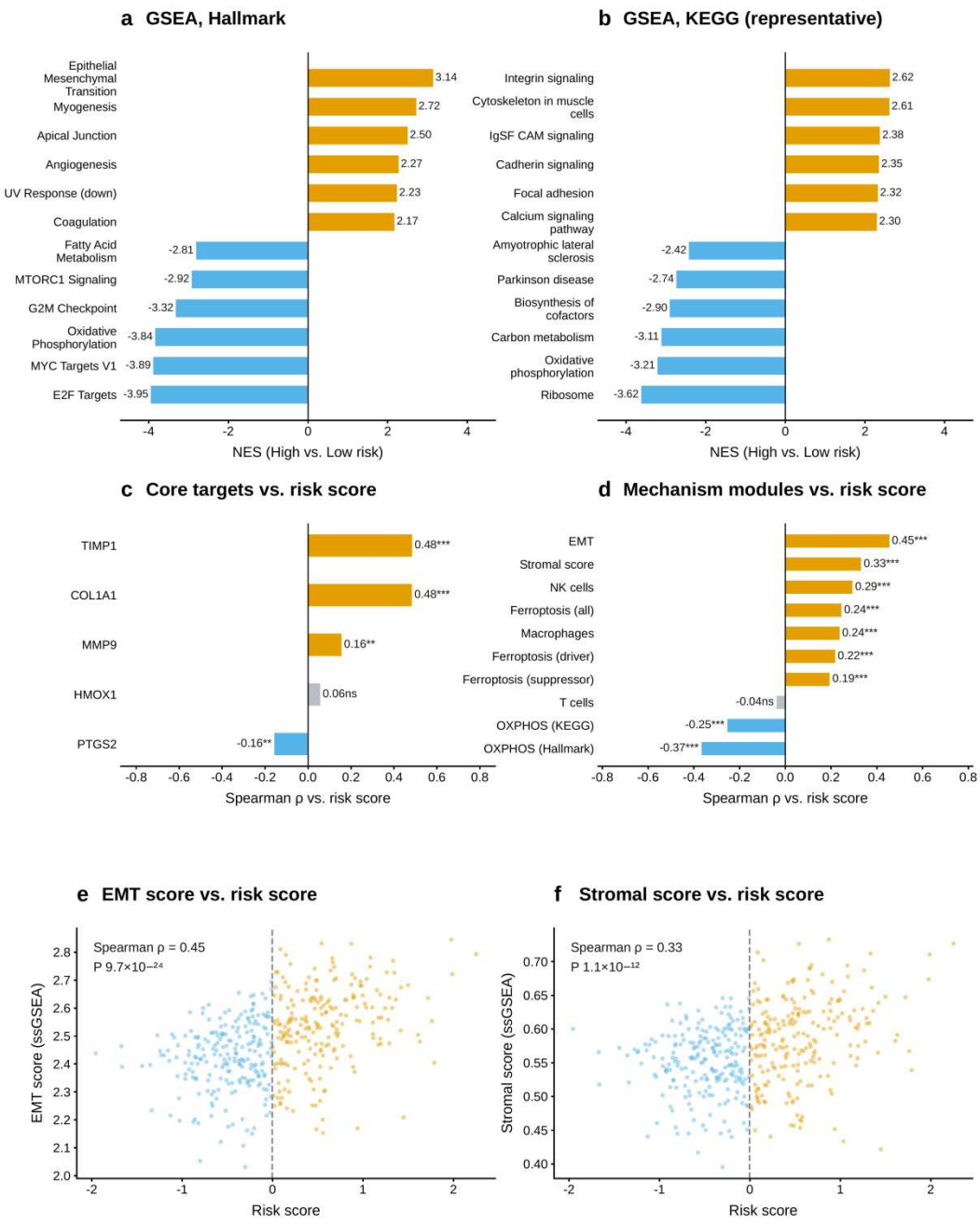


Fig. 8 ***Pathway-level mechanism: EMT/stromal activation with oxidative-phosphorylation suppression.*** *(a) Pre-ranked GSEA of Hallmark gene sets (fgsea; genes ranked by Spearman correlation with the risk score, n=438 tumours; display rule: top 6 pathways per NES sign by BH-adjusted P; all displayed rows BH-significant; positive NES = enriched in the high-risk group). (b) Representative KEGG GSEA results (same rule). (c) Spearman correlation of the five core targets with the risk score (all five reported without selection; BH correction). (d) Spearman correlation of all 10 mechanism modules with the risk*

score (unfiltered). (e, f) Per-sample EMT (Hallmark ssGSEA) and stromal-score versus risk-score scatter plots

3.9 An immune-suppressive, stroma-rich phenotype with TIDE-scored immune exclusion and co-elevated B7-H3, CD39 and CD73

Immune deconvolution showed higher stromal, fibroblast, endothelial, macrophage, and monocytic-lineage scores in the high-risk group (BH $P=3.4\times10^{-8}$ to 5.0×10^{-14}), whereas CD8 T cells and Th1 cells did not differ and Th2 and T-helper cells were lower (BH $P=2.8\times10^{-6}$ and 2.0×10^{-3}); the same pattern was recovered by the MCP marker-based simplified scores (Methods), with endothelial (+*1.63*) and fibroblast (+*0.73*) populations significantly higher in the high-risk group (Fig. 9a, b; all 33 features reported unfiltered). Among the 21 immune-checkpoint genes, six were significantly elevated in the high-risk group — CD276 (B7-H3), ENTPD1 (CD39), NT5E (CD73), CD70, TNFRSF4, and ICOSLG (BH $P=2.8\times10^{-6}$ to 4.6×10^{-2}) — while the remaining 15, including PDCD1, CD274 (PD-L1), CTLA4, and LAG3, did not differ. The high-risk phenotype is thus marked by an abundant fibroblast- and endothelial-rich stroma, by immune exclusion as scored by TIDE, and by co-elevated B7-H3, CD39 and CD73 rather than by classical checkpoint up-regulation, and these are expression-level associations only, not functional pathway activity (Fig. 9c, d). TIDE scores were markedly higher in the high-risk group (Wilcoxon $P=3.0\times10^{-13}$; Spearman $\rho=+0.398$ against the continuous risk score). The difference was driven by the cancer-associated fibroblast ($\rho=+0.52$) and exclusion (+*0.39*) components and by dysfunction (+*0.27*), rather than by TAM M2 (−*0.10*, raw $P=0.044$ in the correlation but not significant between risk groups) or MDSC (−*0.06*, not significant) (Fig. 9e–g).

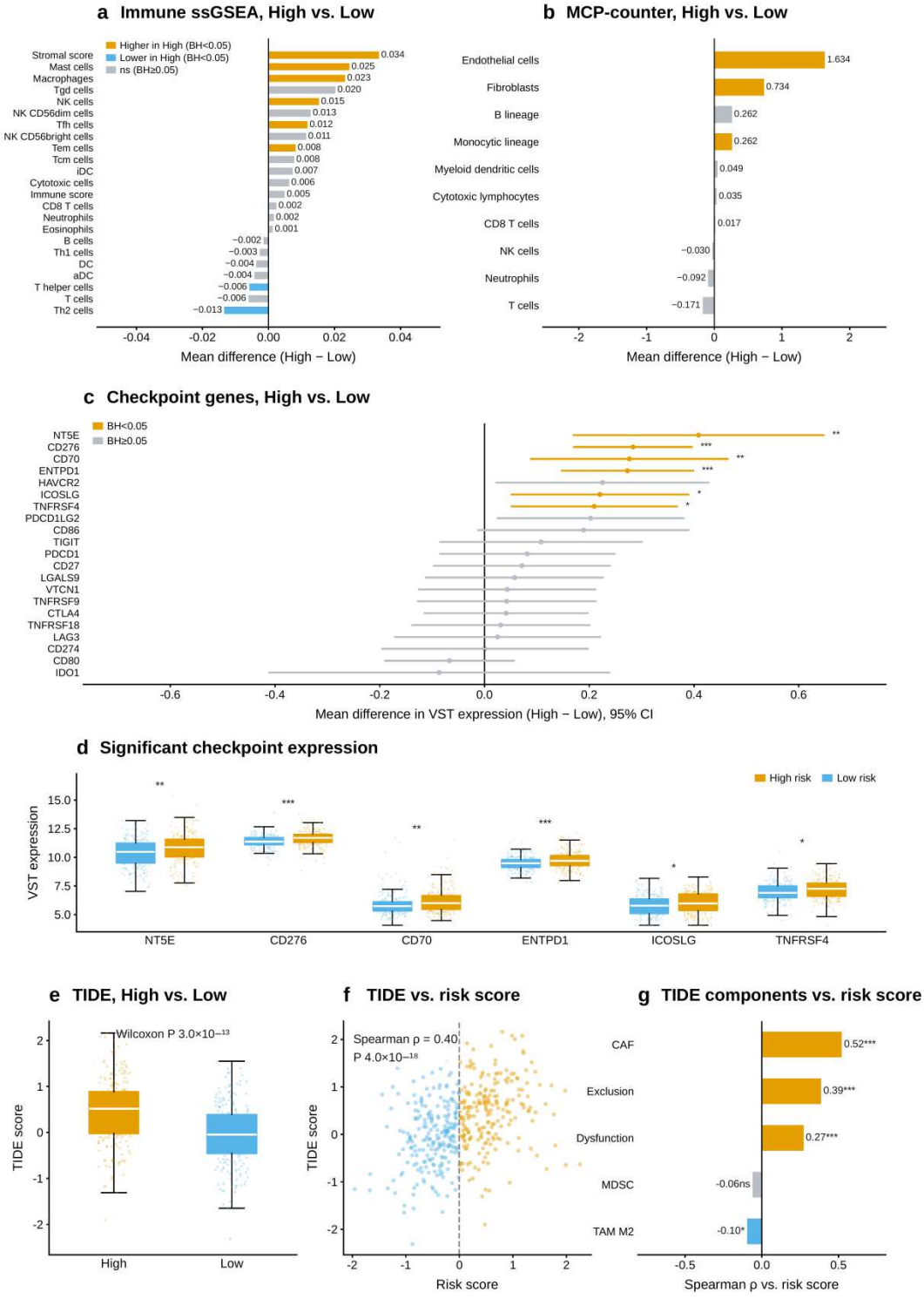


Fig. 9 ***An immune-suppressive, stroma-rich phenotype with TIDE-scored immune exclusion and co-elevated B7-H3, CD39 and CD73.*** *(a) Immune ssGSEA (Bindea signatures; all 23 features reported) between risk groups. (b) MCP marker-based simplified scores (all 10 features). (c) Forest of the mean VST-expression difference (high − low) for 21 immune-*

checkpoint genes with 95% CIs (Welch t tests; BH correction). (d) VST expression of the six BH-significant checkpoint genes (CD276/B7-H3, ENTPD1/CD39, NT5E/CD73, CD70, TNFRSF4, ICOSLG), all up-regulated in the high-risk group. (e) TIDE score between risk groups (n=438; Wilcoxon $P=3.0\times10^{-13}$). (f) TIDE score versus the continuous risk score (Spearman $\rho=+0.398$). (g) Spearman correlation of the five TIDE model components (dysfunction, exclusion, MDSC, CAF, TAM M2; fixed raw-P annotation)

3.10 Single-cell localization of the signature module in fibroblasts and endothelial cells

The atlas contains six patients (six primary tumours, six liver metastases from the same patients, and three PBMC samples). Single-cell module scores apply the fixed signature coefficients to $\log_{10}$(counts per 10,000 + 1) without per-gene cohort standardization, since no cohort-matched per-gene σ exists in scRNA-seq; the scores are comparable across cells within the atlas but are not on the same scale as the bulk risk score. Applying the fixed coefficients to that atlas, the signature module was concentrated in fibroblasts (median 1.224) and endothelial cells (0.525), with intermediate activity in myeloid cells (0.321) and near-zero activity in lymphocytes (Kruskal–Wallis *P* over cell types shown in the panel; Fig. 10a–c). The cell-level tissue comparison is demoted to a descriptive display (Supplementary Fig. S3E) because it ignores within-patient correlation and is subject to pseudo-replication: it was dominated by the upper tail of the distribution (99th percentile: primary tumour 1.303 > liver metastasis 0.961 > PBMC 0.527) with zero medians in all tissues, indicating a higher *proportion* of high-module cells in primary tumours rather than a uniform elevation. The primary analysis is donor-level (Methods): in the six primary CRC tumours, the stromal detection rate exceeded the epithelial detection rate in 6/6 donors (exact sign test *P=0.031*; median difference +*0.58*), driven by fibroblasts (6/6, *P=0.031*) rather than endothelial cells (5/6, *P=0.219*); the direction was unchanged under every alternative positivity definition tested (Supplementary Table S13). An independent donor-level pseudobulk analysis (mean expression per donor and compartment, rather than a per-cell detection rate) agrees and sharpens this: in the six primary tumours the module exceeded the same donor's epithelium in fibroblasts (6/6, *P=0.031*; median difference +*1.06*) and in endothelial cells (6/6, *P=0.031*; +*0.54*), with TIMP1 (+*2.07* and +*0.74*) and SPARC (+*2.31* and +*1.07*) each individually significant in both compartments, while T and B cells showed no difference (3+/3− each, *P=1.0*). Under this analysis the endothelial contrast is as consistent across donors as the fibroblast one, and myeloid cells were also higher than epithelium (6/6, +*0.30*; TIMP1 +*1.14*), so the donor-level signal is stromal-predominant rather than stromal-exclusive (per-donor values in Supplementary Data S7). In the paired primary-versus-liver-metastasis comparison no difference reached significance (fibroblasts 2+/4−, *P=0.688*; endothelial 4+/1−, *P=0.375*; epithelium 3+/3−, *P=1.0*). A per-gene decomposition of the stromal contrast (Supplementary Data S7) shows that the nine weighted contributions to that difference sum exactly to the module difference, but that it is carried by two genes: SPARC (+*0.61*) and TIMP1 (+*0.37*), with $|\Delta|\leq0.08$ for each of the other seven, four of which contribute negatively. SPARC alone separated the compartments at least as strongly as the composite (detection 0.78 vs 0.04, 6/6, *P=0.031*), the contrast survived removal of any single gene (all nine leave-one-out modules 6/6, *P=0.031*; $\rho\geq0.66$ with the full module), and MMP3 and MMP12 sit at the detection floor (<*2.5*%), so their

near-zero contributions cannot be separated from dropout. The localisation therefore rests on SPARC and TIMP1; the remaining seven genes contribute little to it. Feature plots of the highest-detected signature genes per coefficient sign (TIMP1, SPARC for $\beta>0$; PCNA, ACADM for $\beta<0$) localized to the same fibroblast/endothelial regions (Fig. 10e). A single-cell overview of the cohort is provided in Supplementary Fig. S3.

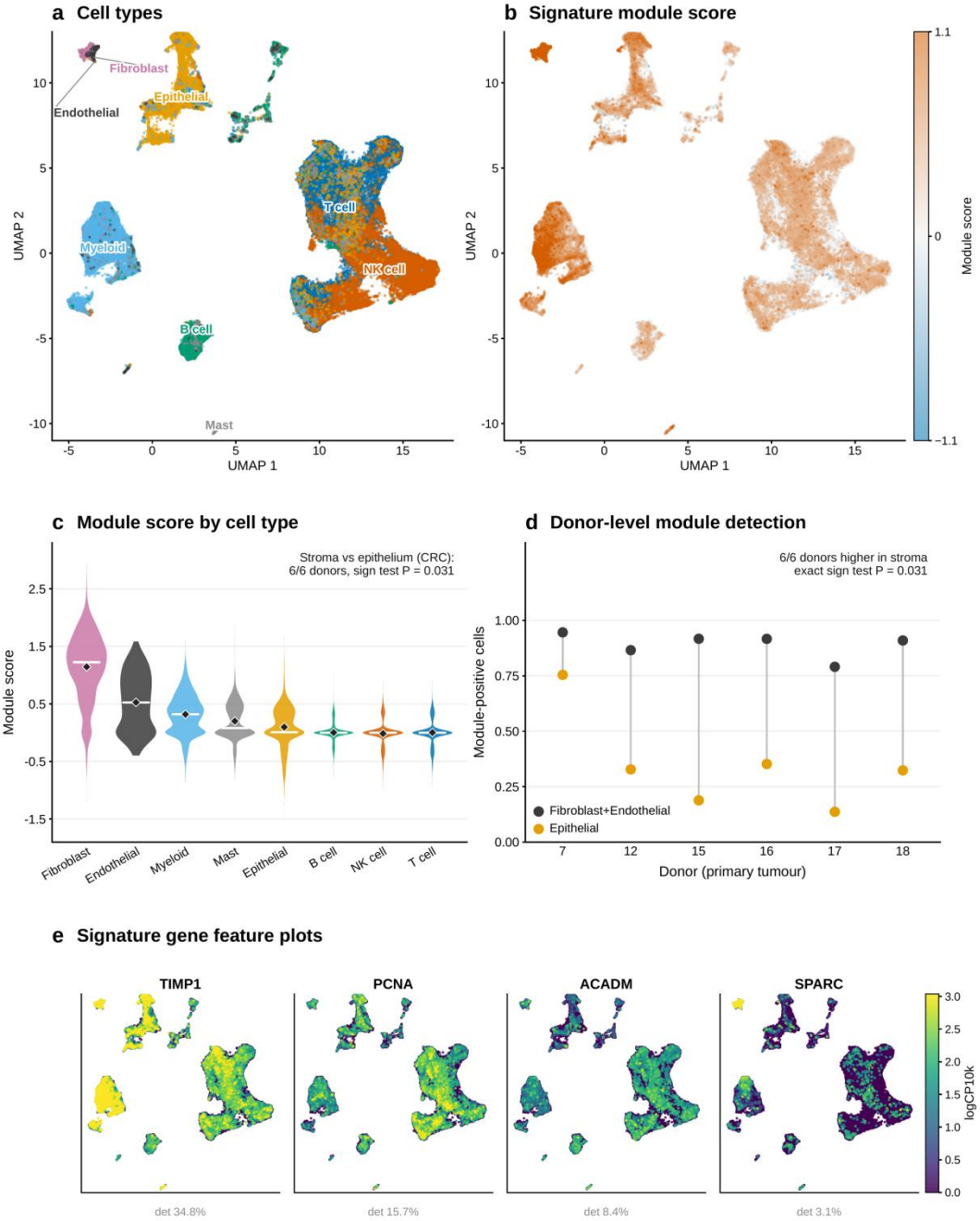


Fig. 10 ***Single-cell localization of the signature module (GSE178318; 139,543 cells after QC).*** *(a) UMAP coloured by coarse canonical-marker cell type (display annotation only). (b)*

UMAP coloured by the nine-gene module score. (c) Module score by cell type (3,000-cell random display sample per class, seed 84; medians, means and Kruskal–Wallis tests use all cells). (d) Donor-level paired comparison of the module-positive detection rate (stroma = fibroblasts + endothelial vs. epithelium) in the six primary tumours; grey lines pair donors; 6/6 higher in stroma (exact sign test P=0.031). (e) Feature plots of the highest-detected signature genes per coefficient sign (TIMP1, SPARC for β>0; PCNA, ACADM for β<0), shared linear scale capped at the joint 99th percentile

3.11 Exploratory pharmacology

As an exploratory, hypothesis-generating pharmacological readout, the risk score was correlated with drug sensitivity (log IC_{50}) across 185 compounds profiled in 46 colorectal-cancer cell lines (GDSC2): no drug passed Benjamini–Hochberg *P<0.05*, including the standard-of-care agents 5-fluorouracil, oxaliplatin, and irinotecan (Fig. 11a) — an honest null that precludes any claim of score-guided drug selection. Molecular docking (AutoDock Vina; three seeds, same-protocol co-crystal controls) gave the strongest interaction for icaritin with MMP9 (*−10.41* vs. *−6.86* kcal/mol for the co-crystal control; re-docking RMSD 1.35 Å), followed by MMP12 (*−8.70* vs. *−6.90* kcal/mol; RMSD 0.70 Å); the contact geometry of both poses, including a predicted catalytic-zinc coordination in MMP9 that appears only at the most permissive distance criterion, is given in Fig. 11c–e. A sensitivity battery (Supplementary Data S7) re-tested these contrasts: icaritin–MMP9 and icaritin–MMP12 held under the second scoring function Vinardo (margins *−2.61* and *−0.60* kcal/mol), whereas the icariin contrasts reversed there, so the weaker-glycoside reading is scoring-function dependent, and deprotonation, zinc deletion and zinc retyping moved no verdict. Two further targets were excluded for the protocol reasons detailed in Methods. PTGS2, the remaining extension target, is reported here as an explicit negative result: its protocol re-docked cleanly (RMSD 0.43 Å) and yet icaritin scored 4.9–5.0 kcal/mol *worse* than the co-crystal control under both scoring functions. Docking scores are scoring-function estimates requiring experimental confirmation.

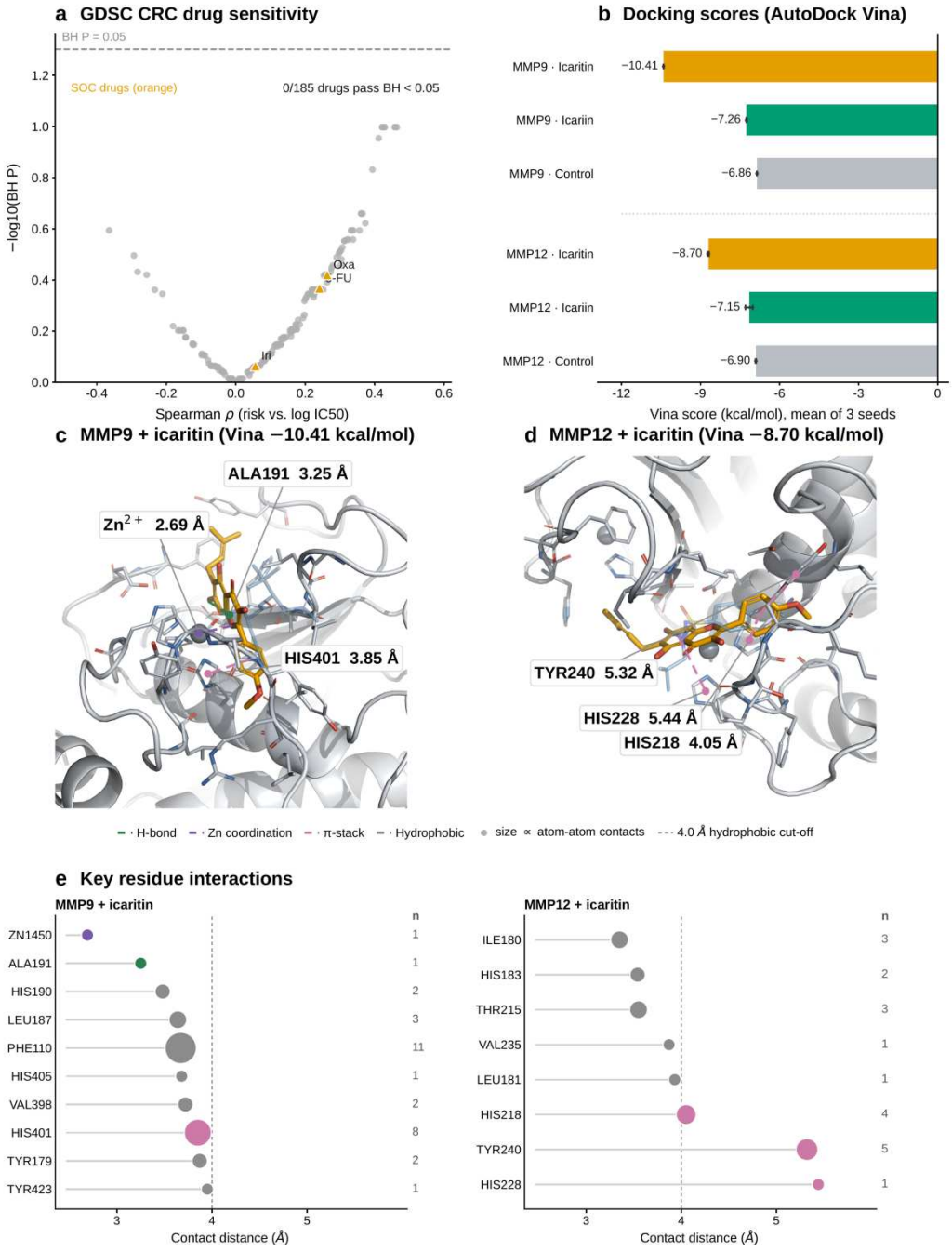


Fig. 11 ***Exploratory pharmacology.*** *(a) GDSC2 drug-sensitivity correlations: Spearman ρ (risk score vs. log IC_{50}; n=46 CRC cell lines, 185 drugs) against $-\log_{10}$(BH P); no drug passes BH P<0.05 (dashed line), including the standard-of-care drugs (triangles). (b) AutoDock Vina scores (three seeds; mean ± SD) for icaritin, icariin, and the co-crystal controls against MMP9 and MMP12; the MMP9 × icariin combination was docked under the identical pre-declared protocol. (c, d) Ray-traced docking poses of icaritin in the MMP9 (c; hydrogen bond to ALA191, a predicted catalytic-zinc coordination that appears only at the most permissive distance criterion, π-stack to HIS401) and MMP12 (d; π-stacks to HIS218, HIS228, and*

TYR240; no hydrogen bond or zinc coordination) catalytic pockets; semi-transparent sticks show co-crystallized reference ligands. (e) Residue-level contact summary of the same poses: one row per contacted residue (MMP9, ten; MMP12, eight) at the distance of its highest-priority interaction, coloured by type and numbered by its atom–atom contact count (n); the dashed line is the 4.0 Å hydrophobic cut-off (π-stack distances follow their own centroid-and-angle criteria)

4 Discussion

This study linked a multi-source library of putative icariin targets to cross-cohort differential expression, defined an 84-gene candidate pool, and ran 100 feature-selection and survival-modeling combinations under fixed consensus rules to derive a nine-gene risk signature. In the independent external validation cohort GSE17536, the transferred signature stratified disease-free survival (Harrell C-index 0.734; HR 2.30 per SD) and added incremental value over stage, age and sex (ΔC=+*0.085*, 95% CI 0.023–0.169; likelihood-ratio *P<0.001*; 36-month Brier score 0.147 versus 0.173). It also marked out a coherent high-risk phenotype: strong enrichment of the CMS4/mesenchymal subtype, EMT/stromal pathway activation with oxidative-phosphorylation suppression, an immune-suppressive microenvironment with an abundant stroma, immune exclusion as scored by TIDE, and co-elevated B7-H3, CD39 and CD73, and single-cell localization of the signature module in fibroblasts and endothelial cells. Exploratory molecular docking further connected the ECM-centred target space to icariin pharmacology.

The high-risk state we identified matches the CMS4/mesenchymal subtype, which is defined by transforming growth factor-β activation, stromal invasion, and angiogenesis [5] and carries the least favourable survival in localized colorectal cancer in meta-analysis [6]. Single-cell and spatial studies indicate that such bulk-level classes partly reflect the composition of the tumour microenvironment rather than the malignant epithelial state alone: fibrosis constitutes a clinically relevant axis that is partly independent of epithelial state [7], and cancer-associated fibroblast and macrophage populations are associated with adverse outcomes and therapy-resistant, immune-excluded microenvironments [8,9,10,11]. Within the score, COL1A1 and TIMP1 correlated most strongly with risk (ρ=+*0.484* each), EMT was the strongest of the ten mechanism modules (ρ=+*0.455*), oxidative phosphorylation was the strongest negative correlate (ρ=−*0.366*), and the three risk genes (TIMP1, SPARC, CLDN1) are canonical stromal and ECM-remodelling markers. In a multivariable model that also carries the positive-weighted stromal genes, the negative coefficients of MMP3 and MMP12 are conditional associations within this score. A simpler compositional explanation is that matrix metalloproteinases reported as pro-tumourigenic in most colorectal-cancer studies are largely stromal-cell-derived, so in a stroma-weighted score low expression of them marks tumour-cell-rich, stroma-poor samples. The score correlates with established stromal, fibroblast, and endothelial scores. Adjusting for the ESTIMATE stromal score left the risk-score effect essentially unchanged in TCGA OS and strengthened it in the training cohort, which indicates that the score carries information beyond stromal content. The same holds in the external cohort, where the increment over stage plus the stromal score remains significant ($\Delta C'$=+*0.072*, *P=0.016*; Supplementary

Data S6). The high-risk immune phenotype is suppressive through immune exclusion rather than classical checkpoint up-regulation: six checkpoint genes were elevated (B7-H3, CD39, CD73, CD70, TNFRSF4, ICOSLG) whereas PDCD1, CD274, CTLA4, and LAG3 did not differ, and TIDE scores were markedly higher ($P=3.0\times10^{-13}$), driven by the CAF and exclusion components. This pattern matches the CAF-driven immune exclusion described in mesenchymal colorectal cancer [8,9,41]. TIDE is expression-derived under a non-COAD rule. If this phenotype is causal, high-risk tumours are the candidates for stroma- and myeloid-directed combinations; whether the icariin compound class modulates this phenotype is untested in signature-defined disease, although reported activities bear on it, including re-education of the microenvironment through inhibition of M2 macrophage polarization [15], ferroptosis promotion with synergistic efficacy of PD-1 inhibitors [14], MDSC-mediated immunosuppression as an emerging therapeutic axis [20], and the immunodynamic survival signal of icaritin in advanced hepatocellular carcinoma [19].

The icariin relationship is the candidate origin of the gene set. The signature is a prediction model built from that target space, whereas the five core targets (TIMP1, PTGS2, COL1A1, HMOX1, MMP9) form a separate pharmacological hypothesis; the two axes are directionally concordant. The signature is dominated by ECM-remodelling genes, and exploratory docking placed icaritin–MMP9 first among the combinations tested (*−10.41* vs. *−6.86* kcal/mol for the co-crystal control), although that ranking was not stable across scoring functions and PTGS2 re-docked as an explicit negative. Reported icariin and icaritin activities bear on this phenotype from a distance, from EMT and Wnt/β-catenin regulation [17,16] and TRAIL sensitization [12] to enhanced 5-fluorouracil activity via NF-κB inhibition [13] and autophagy repression through HSP90–TXNDC9 disruption [18]; they illustrate the candidate space from which the signature was built.

Clinically, the signature is a relative-risk stratifier. It could complement anatomical stage and emerging ctDNA-based residual-disease assessment [3,4] by separating a CMS4/mesenchymal group at higher relative risk, and it may be most informative where current tools are least conclusive, for example in stage II disease where biomarker-guided post-operative management remains an unmet need [42]. The random-effects synthesis of the eight cohorts carried into the pooled analysis is significant ($P=0.013$) and stable to leave-one-out, and we report it as a directional confirmation at attenuated magnitude; its quantitative caveats are gathered under Limitations. The strongest initial candidate proved partially duplicated with GSE17536 (a unified-panel re-audit removed 77 further duplicates from the pooled set without moving the estimate), and the further cohorts disagree in magnitude and, in one case, in direction, so the wider screen adds supporting evidence.

Three design features support these conclusions. Model selection was fixed in advance: the median cross-validated C-index across the 90 non-control combinations was 0.613 and fell to 0.60 when the whole pipeline was re-run inside an outer resampling, and the external conclusions do not depend on the consensus rule's exclusion of the higher-scoring `all`-feature model (0.668; Supplementary Table S4). The signature was transferred with frozen coefficients, validated for disease-free survival in an independent cohort with a formal incremental-value analysis, and its high-risk phenotype is corroborated across bulk subtyping, immune deconvolution, TIDE and single-cell localization, with negative and

boundary results reported alongside positive ones [23]. The conclusions were also stable across anatomical site (right colon 101, left colon 93, rectum 30, unannotated 2): the score performed in the right- and left-sided subgroups (C-index 0.731, HR 2.10, *P=0.016* and C-index 0.776, HR 3.08), and refitting the whole pipeline on the 196 colon-only samples changed no coefficient sign (9/9), reproduced the frozen coefficients (Pearson *r=0.985*), and moved the training and external C-index from 0.750 and 0.734 to 0.755 and 0.739; only the rectal subgroup (*n=30*, 7 events) was too small to test separately (HR 1.82, log-rank *P=0.114*).

4.1 Limitations

Primary external validation rests on a single cohort (GSE17536; C-index 0.734, *n=145* with 36 DFS events), so that estimate is imprecise (C-index 95% CI 0.657–0.810; HR 95% CI 1.72–3.08), and of the further screening rounds only GSE38832 met the pre-declared significance level, four of five runs in metastatic and perioperative-systemic-therapy populations falling at or below chance. The pooled synthesis is significant and stable to leave-one-out, but its point estimate (1.53 per SD) is materially below the training (2.43) and primary external (2.30) values, heterogeneity is substantial (I^2*=69*%), the 95% prediction interval for a new cohort includes 1 (0.71–3.31), and absolute risk transported from the training baseline is under-estimated roughly twofold, so the score requires recalibration before any absolute risk is quoted. Within TCGA-COAD the nomogram remained stage-dominated (multivariable HR 1.27 per SD), adding the signature to stage, age, sex and CMS did not significantly improve internal discrimination for OS, PFI or DFI (*ΔC=+0.014*, *P=0.084*), and the internal endpoint (OS) differs from the DFS endpoints used for training and external validation, being influenced by competing non-recurrence mortality and adjuvant treatment; the excess of early events in the high-risk group is consistent with a recurrence-derived signal but cannot be verified without cause-of-death data, so the incremental-value claim rests on the external DFS validation. The primary external cohort further includes 10 of 145 patients (6.9%) with stage IV disease, wider than the stage II–III training cohort, and its endpoint is a depositor-provided DFS variable (see *Outcome definitions and harmonization*) that could not be re-adjudicated centrally, so residual definitional heterogeneity cannot be excluded; the increment over the clinical base model plus the ESTIMATE stromal score was borderline in TCGA-COAD (*P=0.052*), so the score is a mesenchymal-risk program that partly measures, and partly adds to, tumour stromal content. The study is retrospective and entirely computational, no new experimental data were generated, no formal PROBAST+AI risk-of-bias assessment is presented [22], and the signature is derived from bulk transcriptomes and remains a cohort-level relative-risk marker pending prospective validation. MSI status and TMB did not differ significantly between risk groups, and TIDE scores rely on a non-melanoma (SKCM-rule) approximation without immune-checkpoint-outcome validation. The single-cell annotation is coarse and tissue-mixed, and its stromal contrast is carried mainly by SPARC and TIMP1 rather than by all nine genes. The signature is also not target-library specific: under the identical 100-combination protocol the real 84-gene pool scored at the 88th percentile of 101 arbitrary pools of the same size (external C-index 0.735 versus a null median of 0.689; empirical *P=0.118*), so it stands as a reproducible prognostic model built from, but not specific to, the icariin target library. Finally, the target library contains a large

computational-prediction component (evidence grade C, *n=331*), and no icariin-based intervention has been tested against the signature-defined phenotype.

5 Conclusions

An icariin-informed nine-gene signature built from a multi-source target library is externally validated for disease-free survival in resected colorectal cancer, where it adds to stage, age and sex, and it marks a mesenchymal, immune-suppressive phenotype with an abundant stromal compartment whose signal localizes to fibroblasts and endothelium. Random-effects synthesis across the deduplicated cohorts supports the same direction at attenuated magnitude, and the protocol-matched random-pool control shows that this prognostic performance is not specific to the icariin target space, which is best described as the candidate origin of the gene set. Translation will require independent experimental work — viability and apoptosis assays, target-axis qPCR in signature-high versus signature-low models, and patient-derived systems — together with prospective evaluation.

Declarations

Ethics approval and consent to participate

All datasets analysed here are publicly available, de-identified secondary data (TCGA-COAD and the GEO series listed under *Study design and data sources*); no new human participants, human material, or animals were involved. Ethical approval and informed consent were therefore not required for this work and were obtained by the investigators of each source study under their own institutional review boards.

Consent for publication

No identifiable individual participant information is presented; consent for publication is therefore not applicable.

Availability of data and materials

This study reanalysed publicly available data; no new experimental data were generated. All primary data sources are openly accessible.

Tumour expression and clinical data. TCGA-COAD (RNA-seq STAR-Counts, GDC data release DR46.0) from the Genomic Data Commons, https://portal.gdc.cancer.gov/. The colorectal GEO series used here are listed with their accession numbers, roles and endpoints in Table 1 and Supplementary Tables S6–S7 and are freely available from the Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/. The single-cell atlas is GSE178318, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178318. The one screened series held outside GEO, E-MTAB-6698, is available from ArrayExpress, https://www.ebi.ac.uk/biostudies/arrayexpress.

Icariin target library. ChEMBL https://www.ebi.ac.uk/chembl/; PubChem BioAssay https://pubchem.ncbi.nlm.nih.gov/; SwissTargetPrediction http://www.swisstargetprediction.ch/; the Comparative Toxicogenomics Database https://ctdbase.org/; HERB http://herb.ac.cn/; and PharmMapper https://lilab-ecust.cn/pharmmapper/.

Interaction and pharmacology resources. STRING v12.0 for the protein–protein interaction network, https://string-db.org/; the GDSC2 release for drug-sensitivity correlations, https://www.cancerrxgene.org/.

Protein structures. RCSB PDB entries 1GKC, 1RMZ, 3LN1, 4QTB and 2D1O, https://www.rcsb.org/.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the Natural Science Foundation of Nanjing University of Chinese Medicine (grant no. XZR2024157).

Authors' contributions

Conceptualization, Fei Zuo and Tianping Luo; methodology, Fei Zuo and Tianping Luo; formal analysis and software, Fei Zuo; validation, Fei Zuo and Mingni Jiang; investigation, Fei Zuo, Jiapan Yan, Mingni Jiang, and Tianping Luo; data curation, Fei Zuo, Jiapan Yan, and Mingni Jiang; writing – original draft preparation, Fei Zuo; writing – review and editing, Mingni Jiang and Tianping Luo; visualization, Fei Zuo; supervision, project administration, and funding acquisition, Tianping Luo; resources, Tianping Luo. All authors have read and approved the final version of the manuscript.

The bioinformatics analyses and the initial manuscript draft were prepared with the assistance of an AI writing/analysis agent (ClawsGO Science) under the supervision of the corresponding author.

Acknowledgements

Not applicable.

References

[1] Freddie Bray, Mathilde Laversanne, Hyuna Sung, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74 (3): 229–263, 2024. https://doi.org/10.3322/caac.21834.

[2] C. Pericay, C. Montagut, J. J. Reina, et al. SEOM-GEMCAD-TTD clinical guidelines for the adjuvant treatment of colon cancer (2023). *Clinical & Translational Oncology*, 26 (11): 2812–2825, 2024. https://doi.org/10.1007/s12094-024-03559-5.

[3] Y. Nakamura, J. Watanabe, N. Akazawa, et al. ctDNA-based molecular residual disease and survival in resectable colorectal cancer. *Nature Medicine*, 30 (11): 3272–3283, 2024. https://doi.org/10.1038/s41591-024-03254-6.

[4] J. Tie, Y. Wang, S. N. Lo, et al. Circulating tumor DNA analysis guiding adjuvant therapy in stage II colon cancer: 5-year outcomes of the randomized DYNAMIC trial. *Nature Medicine*, 31 (5): 1509–1518, 2025. https://doi.org/10.1038/s41591-025-03579-w.

[5] Justin Guinney, Rodrigo Dienstmann, Xin Wang, Aurélien de Reyniès, Andreas Schlicker, Charlotte Soneson, Laetitia Marisa, Paul Roepman, Gift Nyamundanda, Paolo Angelino, et al. The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, 21 (11): 1350–1356, 2015. https://doi.org/10.1038/nm.3967.

[6] S. Ten Hoorn, T. R. de Back, D. W. Sommeijer, and L. Vermeulen. Clinical value of consensus molecular subtypes in colorectal cancer: a systematic review and meta-analysis. *Journal of the National Cancer Institute*, 114 (4): 503–516, 2022. https://doi.org/10.1093/jnci/djab106.

[7] I. Joanito, P. Wirapati, N. Zhao, et al. Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. *Nature Genetics*, 54 (7): 963–975, 2022. https://doi.org/10.1038/s41588-022-01100-4.

[8] A. M. Khaliq, C. Erdogan, Z. Kurt, et al. Refining colorectal cancer classification and clinical stratification through a single-cell atlas. *Genome Biology*, 23 (1): 113, 2022. https://doi.org/10.1186/s13059-022-02677-z.

[9] E. Koncina, M. Nurmik, V. I. Pozdeev, et al. IL1R1^{+} cancer-associated fibroblasts drive tumor development and immunosuppression in colorectal cancer. *Nature Communications*, 14 (1): 4251, 2023. https://doi.org/10.1038/s41467-023-39953-w.

[10] K. Pelka, M. Hofree, J. H. Chen, et al. Spatially organized multicellular immune hubs in human colorectal cancer. *Cell*, 184 (18): 4734–4752.e20, 2021. https://doi.org/10.1016/j.cell.2021.08.003.

[11] J. Xiao, X. Yu, F. Meng, et al. Integrating spatial and single-cell transcriptomics reveals tumor heterogeneity and intercellular networks in colorectal cancer. *Cell Death & Disease*, 15 (5): 326, 2024. https://doi.org/10.1038/s41419-024-06598-6.

[12] B. Kim, J. H. Seo, K. Y. Lee, and B. Park. Icariin sensitizes human colon cancer cells to TRAIL-induced apoptosis via ERK-mediated upregulation of death receptors. *International Journal of Oncology*, 56 (3): 821–834, 2020. https://doi.org/10.3892/ijo.2020.4970.

[13] D. B. Shi, X. X. Li, H. T. Zheng, et al. Icariin-mediated inhibition of NF-κb activity enhances the in vitro and in vivo antitumour effect of 5-fluorouracil in colorectal cancer.

Cell Biochemistry and Biophysics, 69 (3): 523–530, 2014. https://doi.org/10.1007/s12013-014-9827-5.

[14] Haoyue Wang, Kexiang Sun, Tan-Wei Shan, Jiamin Gao, Luyun Yuan, Junkai Wen, and Wanli Deng. Icariin promoted ferroptosis by activating mitochondrial dysfunction to inhibit colorectal cancer and synergistically enhanced the efficacy of PD-1 inhibitors. *Phytomedicine*, 136: 156224, 2025. https://doi.org/10.1016/j.phymed.2024.156224.

[15] Y. Chen, Y. Qi, Y. Jiang, Y. Li, S. Yang, L. Wang, M. Li, K. Chai, and Y. Wang. Icariin modulates the tumor microenvironment in colorectal cancer by targeting M2 macrophage polarization via PI3K/AKT pathway. *Bioorganic & Medicinal Chemistry*, 129: 118317, 2025. https://doi.org/10.1016/j.bmc.2025.118317.

[16] X Wang, R-Y Lu, H-B Wang, Y-J Li, L-F Guo, T-H Lan, C Pang, C-H Wang, J Zhang, F-J Shang, L-M Meng, and Z-R Zhao. Mechanism of icariin induces mitochondrial-mediated apoptosis by Wnt/β-catenin signaling pathway in colorectal cancer. *Frontiers in Pharmacology*, 17: 1857797, 2026. https://doi.org/10.3389/fphar.2026.1857797.

[17] X Li, Y-J Li, D-D Ren, M Yuan, Y-R Qi, H-W Li, Y Zhang, and R-Y Wang. Icaritin suppresses colorectal cancer invasion and metastasis through regulation of epithelial-mesenchymal transition and the Wnt/β-catenin signaling pathway. *World Journal of Gastrointestinal Oncology*, 18 (5): 115697, 2026. https://doi.org/10.4251/wjgo.v18.i5.115697.

[18] D. He, S. Chen, X. Wang, X. Wen, C. Gong, L. Liu, and G. He. Icaritin represses autophagy to promote colorectal cancer cell apoptosis and sensitized low-temperature photothermal therapy via targeting HSP90-TXNDC9 interactions. *Advanced Science*, 12 (20): e2412953, 2025. https://doi.org/10.1002/advs.202412953.

[19] Shu-Kui Qin, Qing Li, Jian-Ming Xu, Jun Liang, Ying Cheng, Ying Fan, Jun Jiang, Hao Ye, Huimin Tao, Lian Li, Limin Zheng, Zhaohui Wei, Shu Li, Kun Meng, Bin Ye, and Yan Sun. Icaritin-induced immunomodulatory efficacy in advanced hepatitis B virus-related hepatocellular carcinoma: immunodynamic biomarkers and overall survival. *Cancer Science*, 111 (11): 4218–4231, 2020. https://doi.org/10.1111/cas.14641.

[20] Y Ou, H Wei, C Peng, J Li, K Wei, C Zhan, and Z Zhang. Overcoming MDSC-mediated immunosuppression in hepatocellular carcinoma: from mechanisms to novel immunotherapeutic approaches. *Cancers*, 18 (6): 980, 2026. https://doi.org/10.3390/cancers18060980.

[21] Y. Lu, Y. Gao, H. Yang, Y. Hu, and X. Li. Nanomedicine-boosting icaritin-based immunotherapy of advanced hepatocellular carcinoma. *Military Medical Research*, 9 (1): 69, 2022. https://doi.org/10.1186/s40779-022-00433-9.

[22] P. Dhiman, J. Ma, C. L. Andaur Navarro, et al. Overinterpretation of findings in machine learning prediction model studies in oncology: a systematic review. *Journal of Clinical Epidemiology*, 157: 120–133, 2023. https://doi.org/10.1016/j.jclinepi.2023.03.012.

[23] G. S. Collins, K. G. M. Moons, P. Dhiman, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385: e078378, 2024. https://doi.org/10.1136/bmj-2023-078378.

[24] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173 (2): 400–416.e11, 2018. https://doi.org/10.1016/j.cell.2018.02.052.

[25] Laetitia Marisa, Aurélien de Reyniès, Alex Duval, Janick Selves, Marie Pierre Gaub, Laure Vescovo, Marie-Christine Etienne-Grimaldi, Renaud de Schiappa, Dominique Guenot, Mira Ayadi, et al. Gene expression classification of colon cancer into molecular subtypes: Characterization, validation, and prognostic value. *PLoS Medicine*, 10 (5): e1001453, 2013. https://doi.org/10.1371/journal.pmed.1001453.

[26] Robert N Jorissen, Peter Gibbs, Michael Christie, Sivakumar Prakash, Lara Lipton, Jayesh Desai, David Kerr, Lauri A Aaltonen, Diego Arango, Mogens Kruhøffer, et al. Metastasis-associated gene expression changes predict poor outcomes in patients with Dukes stage B and C colorectal cancer. *Clinical Cancer Research*, 15 (24): 7642–7651, 2009. https://doi.org/10.1158/1078-0432.CCR-09-1431.

[27] Jennifer J Smith, Natasha G Deane, Feiyan Wu, Nipun B Merchant, Bing Zhang, Aixiang Jiang, Pengcheng Lu, Jeffrey C Johnson, Carl Schmidt, Christian E Bailey, et al. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology*, 138 (3): 958–968, 2010. https://doi.org/10.1053/j.gastro.2009.11.005.

[28] 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. https://doi.org/10.1186/s13059-014-0550-8.

[29] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26 (1): 139–140, 2010. https://doi.org/10.1093/bioinformatics/btp616.

[30] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofiler: an R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, 16 (5): 284–287, 2012. https://doi.org/10.1089/omi.2011.0118.

[31] Guangchuang Yu and Qing-Yu He. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*, 12 (2): 477–479, 2016. https://doi.org/10.1039/C5MB00663E.

[32] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51 (D1): D638–D646, 2023. https://doi.org/10.1093/nar/gkac1000.

[33] Aric A Hagberg, Daniel A Schult, and Pieter J Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, pages 11–15, 2008.

[34] Peter W. Eide, Jonas Bruun, Ragnhild A. Lothe, and Anita Sveen. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Scientific Reports*, 7: 16618, 2017. https://doi.org/10.1038/s41598-017-16747-x.

[35] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102 (43): 15545–15550, 2005. https://doi.org/10.1073/pnas.0506580102.

[36] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Systems*, 1 (6): 417–425, 2015. https://doi.org/10.1016/j.cels.2015.12.004.

[37] K. Yoshihara, M. Shahmoradgoli, E. Martínez, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature Communications*, 4: 2612, 2013. https://doi.org/10.1038/ncomms3612.

[38] E. Becht, N. A. Giraldo, L. Lacroix, et al. Estimating the abundance of cell populations in tumour tissue from bulk expression data with MCP-counter. *Genome Biology*, 17: 218, 2016. https://doi.org/10.1186/s13059-016-1070-5.

[39] Peng Jiang, Shixiang Gu, Deng Pan, Jingxin Fu, Avinash Sahu, Xihao Hu, Ziyi Li, Nicole Traugh, Xiaojing Bu, Bo Li, et al. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nature Medicine*, 24 (10): 1550–1558, 2018. https://doi.org/10.1038/s41591-018-0136-1.

[40] Oleg Trott and Arthur J Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31 (2): 455–461, 2010. https://doi.org/10.1002/jcc.21334.

[41] A. Hijazi and J. Galon. Principles of risk assessment in colon cancer: immunity is key. *Oncoimmunology*, 13 (1): 2347441, 2024. https://doi.org/10.1080/2162402X.2024.2347441.

[42] A. A. A. Alnakli, A. Mohamedali, B. Heng, C. Chan, J. S. Shin, M. Solomon, P. Chapuis, G. J. Guillemin, M. S. Baker, and S. B. Ahn. Protein prognostic biomarkers in stage II colorectal cancer: implications for post-operative management. *BJC Reports*, 2 (1): 13, 2024. https://doi.org/10.1038/s44276-024-00043-z.