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.

Published gastric cancer extracellular matrix prognostic signatures converge on a shared core-matrisome axis associated with molecular subtype

Viet Anh Nguyen
Vietnam Academy of Science and Technology
Tri Tuyen Luc
tuyenlt.ioit@gmail.com

Vietnam Academy of Science and Technology
Quang Toan Dao
Vietnam Academy of Science and Technology

Research Article

Keywords:

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11531904/v1

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

Additional Declarations: No competing interests reported.

Published gastric cancer extracellular matrix prognostic signatures converge on a shared core-matrisome axis associated with molecular subtype

Viet Anh Nguyen[1], Tri Tuyen Luc[1,*], and Quang Toan Dao[1]

[1]Department of Data Science and Big Data Analytics, Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam

*Correspondence: tuyenlt.ioit@gmail.com

Abstract

Extracellular matrix (ECM) gene-expression signatures are repeatedly reported as prognostic in gastric cancer, yet seldom re-evaluated against a simple matrisome reference. We re-scored all nine signatures meeting prespecified criteria in a structured PubMed search and our evidence-weighted score in The Cancer Genome Atlas stomach adenocarcinoma cohort (TCGA-STAD; 387 patients) and two Korean microarray cohorts, GSE84437 (431 patients) and GSE62254 (300 patients). With frozen weights, the fourteen-module score was associated with survival in both external cohorts (hazard ratio per SD, 1.27 and 1.32) but added no detectable prognostic information beyond an unweighted core-matrisome average. One component explained 70–84% of module-activity variance, with nearly identical loadings across cohorts. Published signatures correlated with the core average at median 0.71, and only one of eleven primary evaluations in cohorts independent of a signature's development or validation added information beyond it. In GSE62254, molecular subtype explained 32% of score variance; the hazard ratio after clinical and subtype adjustment was 1.19 (95% CI 0.98–1.44). The signatures examined largely tracked one shared axis. In the cohort with molecular annotation, this axis was closely associated with mesenchymal subtype; new signatures should demonstrate added value beyond both references.

Keywords: extracellular matrix; matrisome; gastric cancer; prognostic signature; molecular subtype; model evaluation.

1 Introduction

The extracellular matrix is an active component of the tumour microenvironment. Its composition and remodelling shape tissue stiffness, invasion, angiogenesis and the interaction between tumour, stromal and immune cells [1–3]. The matrisome framework organises this biology into core structural components, namely collagens, glycoproteins and proteoglycans, and matrisome-associated regulators, affiliated proteins and secreted factors [4, 5]. In gastric cancer the link between matrix and outcome is clinically tangible: the mesenchymal molecular subtypes described by the Asian Cancer Research Group (ACRG) and by The Cancer Genome Atlas carry a strong stromal programme and the poorest prognosis [6, 7].

This biology has produced a steady stream of ECM-based prognostic scores for gastric cancer. They range from pathway activity of the core matrisome [8] to small fitted panels built from ECM-receptor, basement-membrane or broader ECM gene sets [9–14]. A systematic review collated 39 gastric-cancer prognostic gene signatures, noted that collagen genes recur across them and called

for independent validation, but compared the signatures through the hazard ratios and accuracies reported by each study rather than by applying them to common data [15]. Most are trained and validated in the same few public cohorts, and most are judged by whether their hazard ratio remains significant in another cohort or after adjustment for stage. That criterion is weak. In large cohorts many random gene sets are associated with outcome [16], and transcripts that track stromal content are correlated with one another regardless of how they are selected or weighted. A significant new ECM score may therefore re-express information that a simple matrisome average or an established molecular subtype already carries. Three questions are rarely asked together: whether a carefully constructed ECM score adds anything beyond a simple summary, whether published ECM signatures measure different biology or the same axis, and whether their prognostic value is separable from molecular subtype.

We evaluated these questions at two levels. At the study level, we built a fourteen-module matrisome score in TCGA-STAD whose weights integrate tumour–normal dysregulation, survival association and protein/tissue annotation. We froze it and evaluated it in two independent Korean cohorts measured on different microarray platforms, against flat matrisome averages and the closest published comparators. At the field level, we identified published gastric ECM signatures by a structured search, applied each one in the three cohorts, and asked how strongly they align with a core-matrisome reference and how often they add prognostic information beyond it in cohorts not used to develop or validate them. Both levels were complemented by principal-component analysis of module activity, adjustment for DNA-based tumour purity and methylation-derived leukocyte fraction, adjustment for molecular subtype, and outcome-blind random-gene controls. Decision rules for the external, field-level and subtype analyses were fixed in dated protocols before the corresponding outcomes were evaluated.

The study makes three contributions. It shows that rich evidence weighting yields a transportable survival association but no measurable gain over an unweighted core-matrisome average, and that the fourteen modules collapse onto one dominant activity axis. It provides a cross-signature audit of gastric ECM prognostic scores against a common reference, in which every signature meeting prespecified criteria in a structured search is re-scored in the same cohorts rather than compared through its reported statistics, an analysis that to our knowledge has not been reported; the nine signatures examined largely converge on the same axis and, in the primary analysis, added information beyond it in only one of eleven evaluations in cohorts not used to develop or validate them. And it shows that, in the external cohort with molecular annotation, the association of scores on this axis could not be separated from molecular subtype by the prespecified rule. Together these results motivate a concrete reporting standard: information beyond a core-matrisome average and beyond molecular subtype should be shown, not assumed.

2 Materials and methods

2.1 Study design

The analysis is a retrospective secondary study of public data with four parts (Figure 1). The evidence-weighted prior was constructed in TCGA-STAD, frozen, and transferred to the external cohorts through a gene panel measured on all platforms. Its external association, incremental value and comparison with simpler and published scores were then evaluated. Shared structure among ECM scores was characterised by principal-component and conditional likelihood-ratio analyses, first for the study-derived scores and nearest published comparators and then for all eligible signatures identified by a structured literature search. Finally, the association was examined against tumour composition and molecular subtype. External outcomes were never used to choose genes,

weights, score directions or thresholds. Rules for the external evaluation, the shared-axis and subtype analyses and the structured benchmark were recorded in dated protocols before the corresponding survival results were computed. Analyses chosen after a result was observed are identified as such where they are reported.

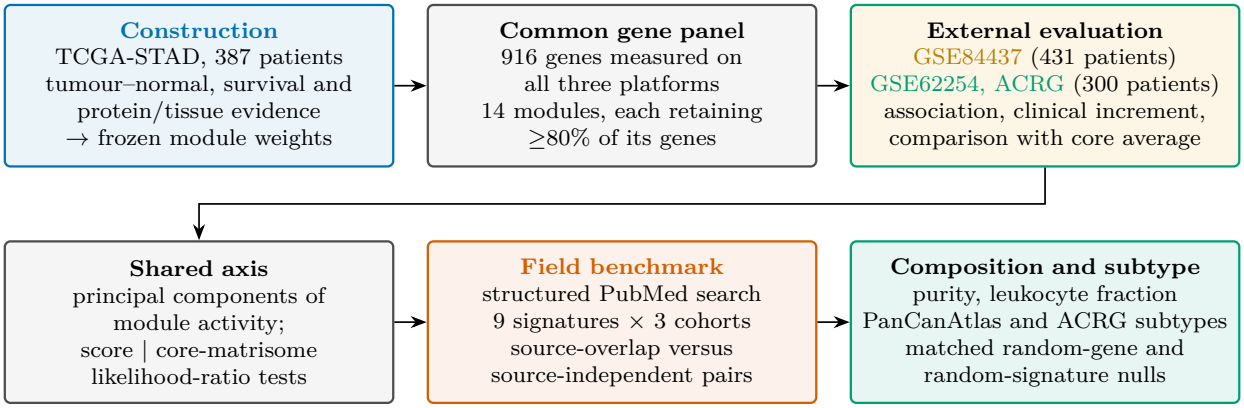


Figure 1: Study design. Colours identify the cohorts (TCGA-STAD blue, GSE84437 orange, GSE62254 green) and the literature benchmark (vermillion) and are used consistently in all figures. The evidence-weighted prior is built only in TCGA-STAD, frozen, and evaluated in two external cohorts on different platforms. The same core-matrisome reference is then used to assess the study-derived scores, published signatures, tumour composition and molecular subtype.

2.2 Cohorts and expression data

Four public cohorts were analysed (Table 1). TCGA-STAD provided RNA-sequencing data, tumour–normal contrasts and survival for construction [7, 17]; the processed matrix is $\log_2(\text{TPM}+1)$ with one primary-tumour profile per patient. GSE84437 is a Korean cohort from Yonsei University profiled on Illumina HumanHT-12 V3 [18]. GSE62254 is the ACRG cohort from Samsung Medical Center profiled on Affymetrix HG-U133 Plus 2.0 [6]; its deposited RMA-normalised values were used without further transformation. GSE26942, also on HumanHT-12 V3 [19], contributed tumour and surrounding normal tissue for an independent tumour–normal comparison only. GEO series were obtained from the GEO repository [20]. Illumina probes were mapped with the platform annotation. Affymetrix probes were mapped with GPL570; probes annotated to several genes were discarded and duplicated genes were represented by the probe with the highest mean expression. In GSE26942, profiles sharing an inferred patient and tissue identifier were averaged before testing.

Overall survival used positive follow-up time and known vital status. In GSE62254, follow-up codes 0 and 1 denote alive and 2 to 4 denote death, according to the source supplement. The clinical models used age, sex and categorical stage in TCGA-STAD and GSE62254, and age, sex, categorical T and categorical N in GSE84437, for which overall stage is not provided. Patients with missing covariates were excluded from adjusted models only, and complete-case numbers are reported with every adjusted estimate. GSE84437 includes 50 profiles deposited separately and without follow-up or stage. They form a distinct expression cluster and never enter survival analyses; because cohort-relative scores use all profiles as the standardisation reference, results were also recomputed with the reference restricted to the 433 profiles with clinical data.

Table 1: Cohorts analysed. Survival patients have positive follow-up and known vital status; follow-up is the reverse Kaplan–Meier median. Age and sex are summarised in the survival sample, and staging counts exclude missing values (stage unavailable for 14 TCGA-STAD patients and ambiguous for two GSE62254 patients). GSE84437 contains 50 additional profiles deposited without follow-up or stage; they are not part of any survival analysis.

	TCGA-STAD	GSE84437	GSE62254 (ACRG)	GSE26942
Source	The Cancer Genome Atlas, multi-institutional [7]	Yonsei University, Seoul, Korea [18]	Asian Cancer Research Group, Samsung Medical Center, Seoul, Korea [6]	Gastric tumour and surrounding normal tissue [19]
Platform	RNA sequencing; $\log_2$(TPM+1)	Illumina HumanHT-12 V3 (GPL6947)	Affymetrix HG-U133 Plus 2.0 (GPL570); RMA	Illumina HumanHT-12 V3 (GPL6947)
Role in this study	construction of weights; internal validation	external evaluation	external evaluation; subtype analysis	tumour–normal comparison only
Expression profiles	387	483	300	217 (205 tumour, 12 normal)
Survival patients / deaths	387 / 160	431 / 207	300 / 152	–
Median follow-up, months	24	116	80	–
Age, median (IQR), years	67 (58–72)	62 (53–68)	64 (55–70)	–
Male, %	65	68	66	–
Clinical staging, n	Stage I/II/III/IV: 52/121/164/36	T1/T2/T3/T4: 11/38/92/290; N0/N1/N2/N3: 80/187/132/32	Stage I/II/III/IV: 30/96/95/77	–
Molecular subtype	PanCanAtlas (CIN, GS, EBV, MSI/POLE)	not available	ACRG (MSS/TP53−, MSS/TP53+, MSI, EMT)	–

2.3 Matrisome modules and gene panels

Fourteen modules were defined. Six correspond to the matrisome categories distributed as MSigDB NABA sets [4, 21]: collagens, ECM glycoproteins, proteoglycans, ECM regulators, ECM-affiliated proteins and secreted factors. The other eight are overlapping process-oriented groupings of matrisome genes: collagen organisation, basement membrane, ECM remodelling enzymes, stromal activation, proteoglycan signalling, cell–matrix adhesion, immune–stroma interaction and invasion-front remodelling. The process modules are heuristic labels rather than validated pathway readouts, and the complete gene lists are released.

Weights were estimated on a construction panel of 995 distinct genes measured without missing values and with non-zero variance in TCGA-STAD and GSE84437 (266 core-matrisome genes; Supplementary Table S2). Evaluation across three platforms requires genes measured in all of them, so each module had to retain at least 80% of its construction genes in all three cohorts. All fourteen modules passed (minimum 86.3%), giving a common panel of 916 genes, 249 of them in the core matrisome, labelled the reduced panel in tables. The common panel with the unchanged construction weights defines the score used in every analysis reported in the main text; construction-panel results are given in the Supplementary Information. Missing expression was never imputed.

2.4 Module activity and the evidence-weighted prior

For a cohort expression matrix X, the activity of module M_k in patient i is the mean of gene-level z-scores,

$$A_{ik} = |M_k|^{-1} \sum_{g \in M_k} (x_{ig} - \bar{x}_g)/s_g,$$

with means and standard deviations taken from the cohort being scored. This statistic has no free parameters and is invariant to positive affine rescaling of each gene, so it can be recomputed identically on any platform. It is cohort-relative: scoring a single new patient would require an external reference, and the study does not address transport of calibrated absolute risk.

Each module received three evidence components in TCGA-STAD. Differential evidence compared 355 tumours from patients without a normal sample with all 36 normal samples by Welch tests on activities computed from a common tumour-plus-normal reference, with Benjamini–Hochberg adjustment across modules [22]. Excluding tumours from patients who also contributed a normal sample avoids partial pairing; the 32 matched pairs were analysed separately. With Δ_k the mean activity difference and q_k its adjusted p-value, $S_k^{\mathrm{diff}} = \mathrm{minmax}_k\{|\Delta_k|[-\log_{10} q_k]\}$. Survival evidence entered each standardised module into an unpenalised Cox model with age, sex and ordinal stage and set $S_k^{\mathrm{surv}} = \mathrm{minmax}_k\{|\hat{\beta}_k|[-\log_{10} p_k]\}$. Protein and tissue support S_k^{prot} averaged four binary Human Protein Atlas annotations over module genes, namely protein-level evidence, tissue annotation, normal-stomach expression and stomach-cancer immunohistochemistry, with relative weights 0.20, 0.10, 0.15 and 0.15 [23, 24]. The raw mixture $0.25S^{\mathrm{diff}} + 0.35S^{\mathrm{surv}} + 0.25S^{\mathrm{prot}}$ was min–max scaled and normalised to sum to one, giving weights w_k; the mixture coefficients were fixed in advance and not optimised. The patient score is $s_i = \sum_k w_k A_{ik}$, with higher values taken to indicate higher risk. Weights were never re-estimated outside TCGA-STAD.

2.5 Survival evaluation

Hazard ratios are reported per standard deviation of the score in the analysis sample, and Harrell's concordance uses the fixed higher-score/higher-risk direction [25, 26]. Association was tested in univariable Cox models, and conditional association by likelihood-ratio tests comparing clinical models with and without the score, fitted without penalty on identical complete cases. Because

clinical coefficients are estimated in each evaluation cohort, the resulting concordance increments are apparent; their uncertainty was summarised by 1,000 patient bootstrap resamples with fixed risk scores. Paired concordance differences between two scores used the same bootstrap. A score was declared superior to a comparator only if the 95% interval of the paired difference lay entirely above zero; failure to meet this criterion is not taken as evidence of equivalence. Proportional hazards were examined with rank-time Schoenfeld residuals. Because submitting site can imprint TCGA data [27], a site-stratified Cox model was fitted in TCGA-STAD. Models were fitted with lifelines [28].

2.6 Comparators and published signatures

The prior was compared with five simpler summaries: the mean z-score of the full matrisome panel and of the core matrisome, equal weights over the fourteen modules, an equal average of the six matrisome categories, and differential-plus-protein weighting without survival evidence. The core-matrisome average serves as the reference summary throughout. The nearest published comparators were fixed before the GSE62254 evaluation: the six-gene Xu signature, the eight-gene Yang signature, the four-gene Deng signature, the nine-gene Yuzhalin panel and a Gaussian-kernel GSVA reconstruction of core-matrisome activity [8–11, 29, 30]. Published coefficients were applied only when every required gene was measured, and z-mean proxies were labelled as such.

To test the field rather than a convenience set, we searched PubMed on 19 October 2026 from inception with a query combining gastric cancer, ECM or matrisome terms, prognostic-signature terms and survival. The full query, all 106 retrieved records and every screening decision are released. Eligible reports were original studies of human gastric cancer that defined a prognostic score from the mRNA expression of at least two protein-coding genes selected from an explicitly ECM, matrisome, collagen, basement-membrane, ECM-remodelling or matrix-stiffness candidate set. Non-coding, single-gene, protein-only and pan-cancer-only signatures, and reports without an accessible formula, were excluded. A core-matrisome activity score known from the literature but not retrieved by the query was also assessed. One investigator screened titles and abstracts, excluding 94 records. Of the 12 retrieved reports and 1 known report assessed in full text, 4 were excluded: for one, a 46-gene model, the full gene list and model formula could not be extracted, two were inaccessible, and one used a candidate set not defined by ECM membership. Nine signatures were retained [8–14, 31, 32].

Each signature was scored in every cohort by its prespecified primary representation: the mean of within-cohort gene z-scores, signed by the published coefficient direction, or unsigned when no coefficients were reported. At least 80% of its genes had to be measured. As a sensitivity analysis, published coefficients were applied when every gene was measured; this shows whether the formula can be computed on the available data, not that the source pipeline is reproduced. A signature–cohort pair was labelled source overlap when the cohort had been used for training or validation in the source report and source independent otherwise.

2.7 Shared-axis analyses

In the survival sample of each cohort, principal-component analysis was applied to the fourteen standardised module activities, with the gene reference restricted to that sample and without outcome input. The first component (PC1) was oriented to correlate positively with the core average. We report its variance share, its loading similarity across cohorts by Tucker congruence, and the effective dimensionality of the module correlation matrix, defined as the participation ratio of its eigenvalues. A dominant shared axis was defined in advance as PC1 explaining at least 50%

of variance in every cohort with pairwise loading congruence of at least 0.85.

Conditional information was assessed by comparing unpenalised Cox models containing the core average with and without a given score, both standardised, by a one-degree-of-freedom likelihood-ratio test; the reverse comparison asks whether the core average adds to the score. Benjamini–Hochberg adjustment was applied across scores and cohorts within each analysis family, and the analyses were repeated with clinical covariates in both models. A score was said to add information beyond the core axis when $q < 0.05$. For the structured benchmark, pair-level passing rates were summarised with Wilson intervals and a Fisher test; both are descriptive because pairs sharing a signature or a cohort are not independent. Two field-level statements were tied to rules fixed in advance: a common axis, if the median Spearman correlation with the core average was at least 0.5, and preferential added information in source-overlap pairs, if at least ten source-independent pairs were available and the Wilson intervals did not overlap or Fisher $p < 0.05$.

For published scores that passed this gate in a source-independent cohort, a matched-signature null asked whether the observed likelihood-ratio statistic exceeded that of comparable random signatures. Each signature gene was replaced, one to one, by one of the 50 genes on the same platform with the closest percentiles of mean and standard deviation and the same matrisome membership, keeping the formula, coefficients or signs fixed. One thousand outcome-blind draws gave an empirical tail probability $(1 + \#\{\mathrm{LRT}_b \geq \mathrm{LRT}_{\mathrm{obs}}\})/1001$. These nulls were specified after the conditional results were observed and are sensitivity analyses.

2.8 Composition and molecular subtype

In TCGA-STAD we used patient-level DNA-based ABSOLUTE tumour purity and methylation-derived leukocyte fraction from the TCGA pan-cancer analyses [33, 34], restricted to primary-tumour samples and aggregated by patient median. These measurements are not derived from the expression of ECM genes and therefore provide a more independent composition adjustment than an expression-based stromal score. Molecular subtypes were the PanCanAtlas gastrointestinal classes for TCGA-STAD [35], obtained through cBioPortal [36, 37], and the ACRG classes for GSE62254 [6], matched by tumour identifier with complete agreement of age and sex. TCGA-STAD classes were chromosomal instability (CIN), genomically stable (GS), Epstein–Barr virus positive (EBV) and a hypermutated group combining microsatellite-instable tumours with seven POLE-mutant tumours; the ACRG classes were MSS/TP53−, MSS/TP53+, MSI and EMT. GSE84437 lacks subtype annotation. We report the variance of each score explained by subtype (η^2) and fitted clinical, clinical-plus-subtype and clinical-plus-subtype-plus-score models on identical complete cases. Sensitivity models added Lauren class, omitted clinical covariates, or excluded the most stromal subtype (EMT in GSE62254 and GS in TCGA-STAD). Because the prior's weights used TCGA-STAD outcomes, subtype results in that cohort are descriptive; the association in GSE62254 was called separable from subtype only if the score's likelihood-ratio p remained below 0.05 after subtype adjustment.

2.9 Internal validation and random-gene controls

Internal validation in TCGA-STAD used three repetitions of five-fold patient splits stratified by event and a five-fold site-held-out scheme. In every training fold the gene reference, differential and survival evidence and module weights were re-estimated, normal samples from test patients were excluded, and test patients were scored with the training reference. Concordance was computed within test folds and averaged without significance testing across repeated folds. Specificity to ECM genes was examined with 1,000 one-to-one replacements of the 916 panel genes by non-

matrisome genes measured in all three cohorts, sampled within bins of TCGA-STAD mean and standard deviation. A single mapping was shared across modules and cohorts, so that module sizes and overlaps were preserved, and the real weights were retained. Leave-one-module-out analysis renormalised the frozen weights after removing each module in turn.

2.10 Software and data availability

Analyses were run in Python with pandas, NumPy, SciPy, scikit-learn and lifelines [28], and the GSVA reconstruction followed Hänzelmann et al. [30]. Every number in the text and tables is generated by scripts from saved outputs. Each public database and record used is listed, with its accession and direct link, in the Data availability statement. The release contains input checksums, analysis protocols, gene lists, construction code, per-patient scores, fold predictions and weights, the PubMed query with all identifiers and screening decisions, signature specifications, all random mappings and draws, bootstrap outputs and fit-failure logs.

2.11 Use of generative artificial intelligence tools

Generative artificial intelligence tools, including OpenAI ChatGPT/Codex and Anthropic Claude, assisted with drafting and debugging analysis code, consistency checks and language editing. The authors made all decisions concerning the study design, analysis, interpretation and conclusions. They reviewed and verified all AI-assisted code and text and take full responsibility for the accuracy and integrity of the work.

3 Results

3.1 Cohorts

The three survival cohorts differ in platform, population and follow-up (Table 1). TCGA-STAD contributed 387 patients with 160 deaths and a median follow-up of 24 months. GSE84437 had 431 patients, 207 deaths and 116 months of follow-up, and GSE62254 had 300 patients, 152 deaths and 80 months. GSE84437 was dominated by T4 tumours, whereas a quarter of GSE62254 patients had stage IV disease. Together with three different expression platforms, these differences make the external cohorts a test of transport rather than a replication within one setting.

3.2 An evidence-weighted ECM prior

Common-reference scoring showed tumour–normal differences in all 14 modules in TCGA-STAD ($q < 0.05$) and in 8 modules in GSE26942 (205 tumour and 11 normal patient-level profiles); the paired analysis is reported in Supplementary Table S1. The three evidence components ranked modules differently (Supplementary Table S2), and their combination concentrated weight on cell–matrix adhesion, ECM regulators, invasion-front remodelling and remodelling enzymes (Supplementary Fig. S1). On the construction panel, module activities in TCGA-STAD had a mean pairwise correlation of 0.82 and an effective dimensionality of 1.41, a first indication that the fourteen modules share most of their variation.

3.3 A transportable association without added discrimination

With frozen weights on the common panel, the prior was associated with overall survival in GSE84437 (HR 1.266 per SD, 95% CI 1.100–1.457; concordance 0.570) and in GSE62254 (HR

1.317, 1.132–1.532; concordance 0.578; Table 2). Clinical adjustment retained the coefficient. In GSE62254, 298 complete cases gave HR 1.268 (1.078–1.491; likelihood-ratio $p = 0.0044$). Apparent concordance, however, moved only from 0.716 to 0.721, with a bootstrap interval for the increment of -0.009 to 0.018, and the corresponding interval in GSE84437 also included zero. The score therefore carries conditional association but not demonstrable added discrimination. The score term showed no departure from proportional hazards in either external cohort. Two nodal categories were flagged in the GSE84437 clinical model, and a model stratified by N gave the same estimate (Supplementary Table S15).

The unweighted core average performed as well as the prior. Its concordance was 0.570 in GSE84437 and 0.593 in GSE62254, and the paired prior-minus-core differences were 0.001 (-0.012–0.013) and -0.015 (-0.033–0.003); neither met the superiority criterion. Restricting the GSE84437 standardisation reference to the 433 profiles with clinical data gave HR 1.215 (1.052–1.403), an adjusted HR of 1.189 (1.026–1.378; $p = 0.020$) and a prior-minus-core difference of -0.001 (-0.012–0.010), so the separately deposited profiles did not drive these conclusions (Supplementary Table S16).

Table 2: Reduced-panel association of the frozen prior and clinical adjustment. Hazard ratios are per score SD. Concordance increments are apparent and their intervals use 1,000 fixed-risk patient bootstraps.

Cohort	Score	n/events	HR (95% CI)	C	adjusted HR (95% CI)	LRT p	C_{clin}	C_{full}	ΔC (95% CI)
STAD	Evidence prior	387/160	1.294 (1.105–1.515)	0.560	1.307 (1.106–1.545)	0.0017	0.637	0.649	0.012 (-0.013–0.039)
STAD	Core average	387/160	1.304 (1.119–1.520)	0.567	1.338 (1.134–1.579)	6.1e-04	0.637	0.655	0.018 (-0.009–0.047)
GSE84437	Evidence prior	431/207	1.266 (1.100–1.457)	0.570	1.230 (1.063–1.423)	0.0052	0.674	0.682	0.008 (-0.004–0.021)
GSE84437	Core average	431/207	1.265 (1.104–1.450)	0.570	1.239 (1.075–1.428)	0.0031	0.674	0.683	0.009 (-0.005–0.022)
GSE62254	Evidence prior	300/152	1.317 (1.132–1.532)	0.578	1.268 (1.078–1.491)	0.0044	0.716	0.721	0.005 (-0.009–0.018)
GSE62254	Core average	300/152	1.392 (1.198–1.618)	0.593	1.340 (1.134–1.583)	6.4e-04	0.716	0.721	0.005 (-0.011–0.020)

3.4 Simpler and published summaries perform alike

The prior did not consistently exceed any comparator (Supplementary Table S3). In GSE84437 it exceeded the full-matrisome average and the direct Xu coefficient score by the paired criterion. It did not exceed the core average, equal module weights, the Xu z-mean, the Deng scores, the Yuzhalin proxy or the core-matrisome GSVA reconstruction, and in GSE62254 no comparison met the criterion. The highest concordance in GSE62254 belonged to the Deng score, but GSE62254 was its development cohort. The Yang formula could not be computed in GSE84437 or GSE62254 because SV2B and VTN, respectively, were not measured; missing genes were not dropped.

3.5 ECM scores converge on one module-activity axis

The similarity of these concordances is mirrored in the structure of module activity. In every cohort, PC1 explained 0.70–0.84 of the variance of the fourteen module activities, and its loadings were nearly identical across cohorts (Tucker congruence at least 0.997; Figure 2a,b; Supplementary Table S4). Effective dimensionality on the common panel was 1.39–1.96 of a possible fourteen, and both prespecified criteria for a dominant shared axis were met. The core average correlated with PC1 at Spearman $\rho = 0.91$–0.96, the prior at 0.956–0.989, and the four study-derived summaries together at 0.936–0.998 (Figure 2c). In TCGA-STAD, PC1 correlated with DNA-based purity ($\rho = -0.39$) and leukocyte fraction ($\rho = 0.35$), and in GSE62254 ACRG subtype explained 0.26 of its variance. The axis is therefore related to tumour composition and subtype, although neither identifies a single cell type.

Once the core average was in the model, none of the study-derived summaries added prognostic information in any cohort (all $q \geq 0.05$; Supplementary Table S5); their correlations with the core average were 0.87–0.96. The reverse did not hold: in GSE62254, the core average added information to the evidence prior and the other module-weighted scores in unadjusted models. Among the nearest published comparators, the scores that added information beyond the core mostly did so in cohorts used in their own development. Two exceptions occurred in source-independent cohorts. The Deng coefficient score added information in GSE84437 (conditional HR 1.241, 95% CI 1.070–1.439; $q = 0.022$; with clinical covariates HR 1.265, $q = 0.025$), and the Xu z-mean did so in GSE62254 only after clinical adjustment (HR 1.455; $q = 0.040$; unadjusted $q = 0.16$). Against matched random signatures (Figure 4a,b), the Deng result reached empirical $p = 0.058$ with the core alone and 0.015 with clinical covariates, and the Xu result 0.182 and 0.049 (Supplementary Table S7). The two Deng representations were among the scores least aligned with the core axis ($\rho = 0.33$–0.46 in GSE84437).

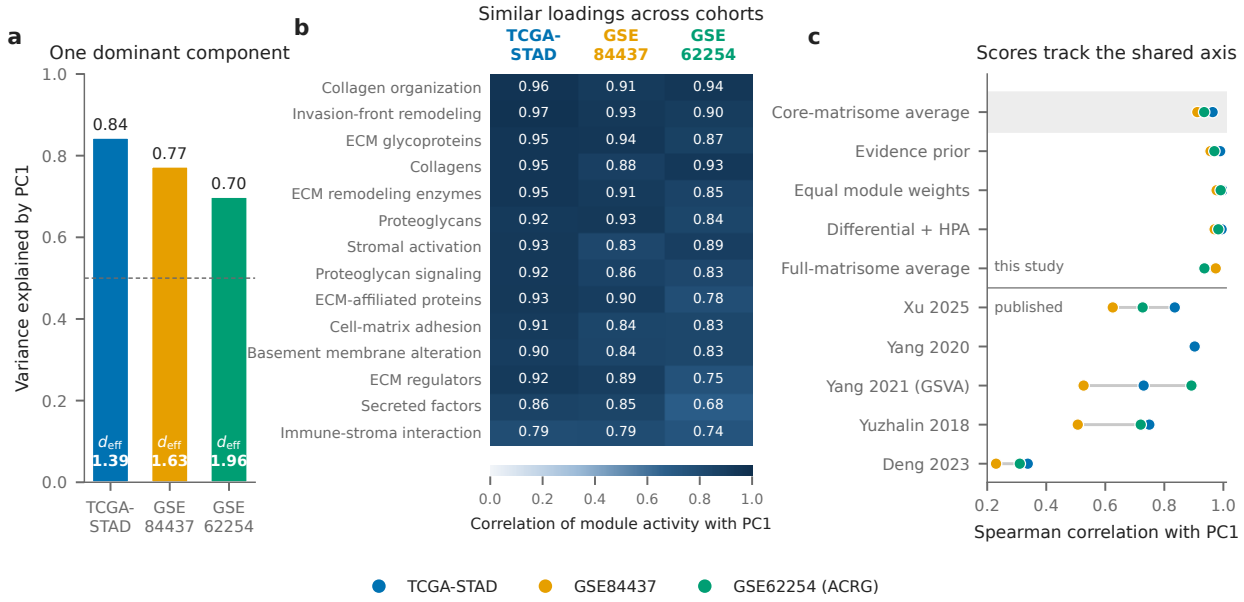


Figure 2: A shared module-activity axis. (a) Proportion of variance of the fourteen standardised module activities explained by the first principal component (PC1) in the survival sample of each cohort; the dashed line marks the prespecified threshold of 0.5, and d_{eff} is the effective dimensionality (participation ratio of the eigenvalues, maximum fourteen). (b) Correlation of each module activity with PC1, estimated separately in each cohort and oriented to the core-matrisome average. (c) Spearman correlation with PC1 of the scores built in this study and of the nearest published scores; each point is one cohort, and the line spans the three cohorts.

3.6 Published gastric ECM signatures share the axis

The structured search retained 9 signatures (Table 3): three built from basement-membrane genes, one from ECM-receptor interaction genes, one core-matrisome activity score and four from broader ECM-related candidate sets. All 27 signature–cohort pairs met the coverage rule. The published-coefficient formula could be computed for 18 pairs; one of these relies on reading a typographical symbol in the source formula as SPARC, as indicated by the source text. Sixteen pairs were source

overlap, because most signatures had been developed or validated in TCGA-STAD or the ACRG cohort, and only eleven were source independent.

The median Spearman correlation of the signatures with the core average was 0.71, meeting the common-axis rule (Figure 3a). The core-matrisome activity score was almost identical to the reference ($\rho = 0.999$), as expected from its definition; without it, the median over the remaining 24 pairs was 0.66. Only 1/11 source-independent pairs added information beyond the core (Figure 3; Supplementary Table S6): the six-gene basement-membrane signature of Wang et al. in GSE62254. This signature also exceeded its matched-signature null, both with the core alone (empirical $p = 0.019$) and with clinical covariates (0.021; Figure 4c; Supplementary Table S7). Source-overlap pairs passed more often (6/16, Wilson 95% interval 0.18–0.61) than source-independent pairs (1/11, 0.02–0.38; Fisher $p = 0.18$; Figure 3b), but the intervals overlapped and the prespecified rule for a preferential effect in source cohorts was not met. With clinical covariates the counts were 5/16 and 2/11 (Fisher $p = 0.66$), and with published coefficients, available for fewer pairs, 7/11 and 2/7 (Fisher $p = 0.33$). Across source-independent pairs, stronger alignment with the core tended to accompany less conditional information ($\rho = -0.46$, $p = 0.15$).

Representation can matter for an individual signature. The Deng signature passed in GSE84437 in its published-coefficient form ($q = 0.011$ within the coefficient family) but not as a signed z-mean ($q = 0.17$). The two forms share genes and signs and differ only in the published weights and the unstandardised source scale. Such divergence is one reason why a signature should be reported with its exact formula and expression scale.

Table 3: Gastric ECM prognostic signatures retained by the structured PubMed search. The primary cross-platform representation was prespecified; "coefficients" counts cohorts in which all genes needed for a published-coefficient calculation were measured. Source overlap means that the cohort was used for training or validation in the source publication.

Signature	Candidate basis	Genes	Primary representation	Coefficients	Source-overlap cohorts
Xu et al. 2025 [11]	ECM/WGCNA	6	signed z-mean	3/3	STAD
Yang et al. 2020 [9]	ECM–receptor	8	signed z-mean	1/3	STAD, GSE62254
Deng et al. 2023 [10]	ECM/PDTX	4	signed z-mean	3/3	GSE62254, STAD
Yang et al. 2021 [8]	core matrisome	275	unsigned z-mean	0/3	STAD, GSE84437, GSE62254
Tang et al. 2024 [12]	basement membrane	6	signed z-mean	3/3	STAD, GSE84437, GSE62254
Liu and Xin 2023 [13]	basement membrane	4	signed z-mean	3/3	STAD, GSE62254
Wang et al. 2023 [14]	basement membrane	6	signed z-mean	2/3	STAD
Wang et al. 2025 [31]	ECM-related	2	signed z-mean	3/3*	STAD
Liu et al. 2020 [32]	ECM-related	7	unsigned z-mean	0/3	STAD

*The source formula prints "S" where the surrounding text identifies SPARC; this coefficient calculation is an explicitly flagged interpretation, not a literal reproduction. Published-coefficient calculations also retain cohort-specific source scales and therefore do not reproduce complete original pipelines.

3.7 Composition and molecular subtype

In TCGA-STAD the prior correlated inversely with DNA-based purity ($\rho = -0.374$) and positively with leukocyte fraction ($\rho = 0.333$). Adjusting for both, together with clinical covariates, attenuated but did not remove the association (356 patients; HR 1.264, 95% CI 1.052–1.519; $p = 0.012$; Supplementary Table S8), and the apparent concordance increment was 0.004 (-0.017 to 0.025).

Molecular subtype accounted for more of the score in the external cohort than in TCGA-STAD. In GSE62254, ACRG subtype explained 0.32 of the variance of the prior and 0.44 of that of the core average, with the highest values in EMT tumours (Figure 4d); in TCGA-STAD the PanCanAtlas subtype explained 0.05 and 0.13, with GS tumours highest (Figure 4e). Subtype added to the clinical model in GSE62254 (298 patients, 152 deaths; $p = 0.0063$). With subtype

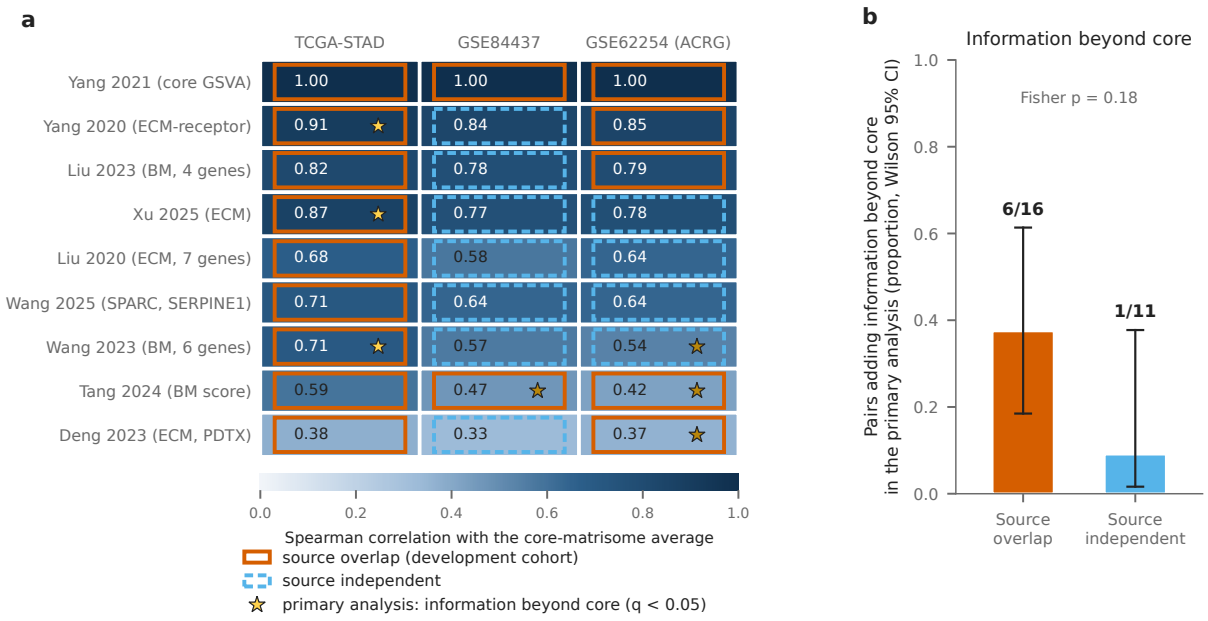


Figure 3: Field-level audit of published gastric ECM signatures. (a) Each cell is one signature–cohort pair; its shade and number give the Spearman correlation of the signature with the core-matrisome average. A solid vermillion frame marks a cohort used in the signature's development or validation (source overlap) and a dashed blue frame a source-independent cohort. A star marks a pair meeting the primary unadjusted information-beyond-core gate (likelihood-ratio test, Benjamini–Hochberg $q < 0.05$, hazard ratio above one). Signatures are ordered by median correlation. BM, basement membrane; PDTX, patient-derived tumour xenograft. (b) Proportion of pairs meeting the primary unadjusted gate, with Wilson 95% intervals.

in the model, the prior's hazard ratio fell to 1.186 (0.977–1.439; $p = 0.083$) and its concordance increment to 0.000 (-0.007 to 0.008); the core average behaved similarly (HR 1.235, 0.997–1.531; $p = 0.052$). By the prespecified rule, the association in GSE62254 is not separable from molecular subtype (Figure 4f; Supplementary Table S9). The conclusion held when Lauren class was added (HR 1.198; $p = 0.076$) and when EMT tumours were excluded (HR 1.155, 0.963–1.384; $p = 0.12$; $n = 252$). Without clinical covariates, the score remained associated after subtype adjustment (HR 1.211, 1.008–1.456; $p = 0.042$). In TCGA-STAD, where the result is descriptive, the association persisted after clinical and subtype adjustment (343 patients; HR 1.281, 1.075–1.526; $p = 0.0057$) and more weakly after excluding GS tumours (HR 1.207; $p = 0.046$).

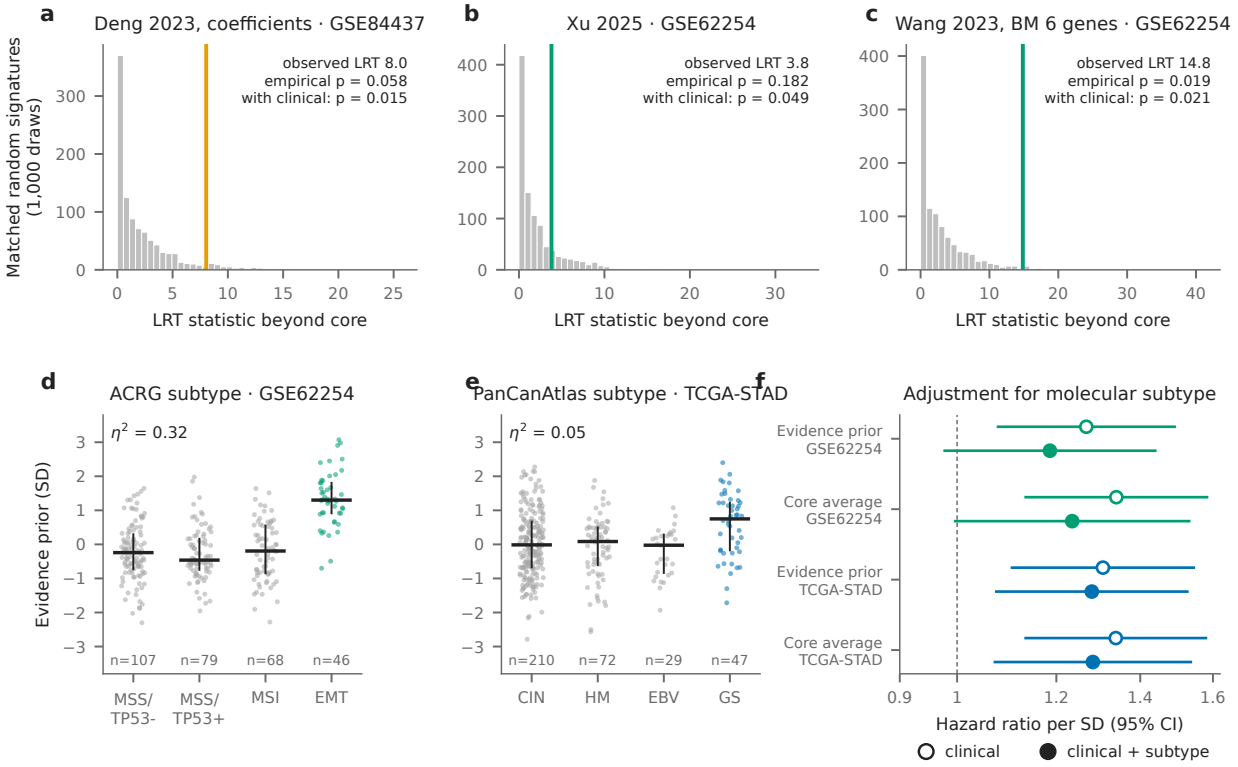


Figure 4: Matched random-signature controls and molecular subtype. (a–c) Likelihood-ratio statistic for information beyond the core average of 1,000 random signatures matched gene by gene on expression mean, variability and matrisome class (grey), against the observed signature (vertical line), for the three source-independent pairs that added information under any specification; empirical p values are given without and with clinical covariates. All three null analyses were specified after the conditional results were observed and are sensitivity analyses. (d, e) Evidence prior by molecular subtype in GSE62254 (ACRG subtypes) and TCGA-STAD (PanCanAtlas subtypes); bars show the median and interquartile range, the stromal subtype is coloured, and η^2 is the proportion of score variance explained by subtype. (f) Hazard ratio per standard deviation for the evidence prior and the core average with clinical covariates (open) and with clinical covariates and subtype (filled). TCGA-STAD estimates are descriptive because the prior weights used its outcomes.

3.8 Internal validation and random-gene controls

When the whole construction was repeated inside training folds, mean test-fold concordance was 0.562 for patient splits and 0.572 for site-held-out splits, and the mean increments over the clinical model were 0.0057 and 0.004 (Supplementary Table S10), compared with an apparent concordance of 0.560 for the final prior in TCGA-STAD. The site-stratified model gave HR 1.213 (1.000–1.472; $p = 0.0494$). Against 1,000 expression-matched non-matrisome gene sets with identical module structure, the observed concordance had empirical tail probabilities of 0.0010 in TCGA-STAD, 0.0010 in GSE84437 and 0.0030 in GSE62254 (Figure 5); the TCGA-STAD comparison is descriptive because the weights used its outcomes. Removing any single module changed concordance little (Supplementary Table S17), consistent with the dominant shared axis.

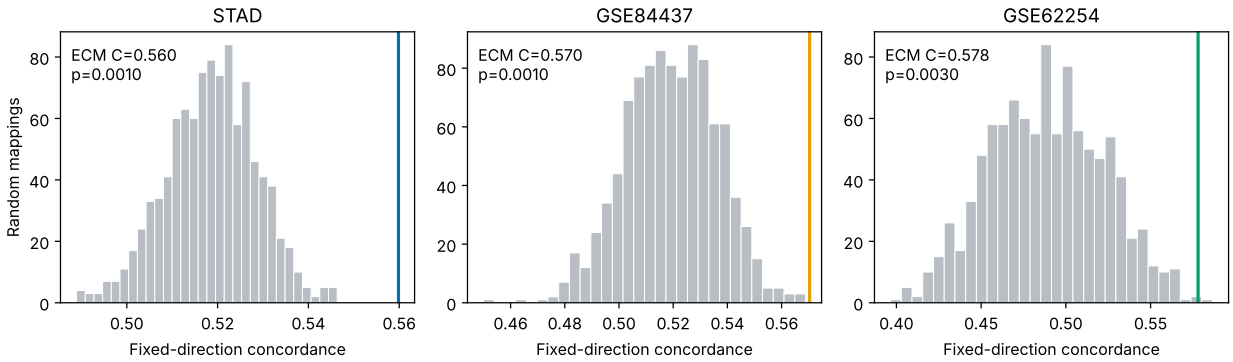


Figure 5: Observed concordance of the prior (vertical line) against 1,000 expression-matched random gene mappings that preserve module sizes and overlaps. One outcome-blind mapping is shared across modules and cohorts; the TCGA-STAD panel is descriptive because the weights used its outcomes.

4 Discussion

This study asked what ECM expression scores add to gastric-cancer prognosis, rather than whether another such score can be made significant. A score enriched with tumour–normal, survival and tissue evidence transported across sequencing and two microarray platforms and remained associated with survival after clinical adjustment. It nevertheless did no better than an unweighted average of core-matrisome genes. This is consistent with the organisation of gastric ECM expression along one dominant axis. Fourteen biologically labelled modules had an effective dimensionality below two, their first component had nearly identical composition in three cohorts, and none of the study-derived summaries added detectable information beyond the core average. In these cohorts, evidence weighting did not sharpen a signal that was effectively one-dimensional, and the stability of the score under reweighting or module removal is better read as a consequence of this redundancy than as evidence of robustness.

The field-level benchmark extends the conclusion beyond our own construction. Published gastric ECM signatures, built from basement-membrane, ECM-receptor or broader ECM gene sets with different statistical procedures, had a median correlation of 0.71 with the core average, and in the primary analysis only one of eleven evaluations in cohorts not used to develop or validate a signature showed prognostic information beyond it. Signatures passed more often in cohorts that overlapped their development data (6/16 against 1/11; Fisher $p = 0.18$), a pattern that would

be consistent with optimistic selection, but with eleven source-independent pairs the difference remains a trend rather than an established property of the field. The exceptions are informative. The basement-membrane signature of Wang et al. added information in the ACRG cohort and exceeded comparable random signatures, and the Deng signature behaved similarly in GSE84437 in its coefficient form. Both were among the signatures least aligned with the core axis, and each has only one core-matrisome gene. These two exceptions, examined after the primary results were known, suggest without establishing that prognostic information beyond the shared axis may reside in genes outside the dominant matrisome programme rather than in finer weighting of matrisome genes.

Matrix-linked genes also appear in gastric-cancer signatures framed around other biological themes. A recent macrophage-associated model selected GPX3, SERPINE1 and SPARC, whereas an unrestricted systems-biology workflow produced an FN1–SRSF7 score [38, 39]. These models were outside the prespecified benchmark because ECM membership did not define their candidate sets, but they illustrate how stromal or matrix-associated signal can enter prognostic models under different labels.

The subtype analysis indicates what the shared axis represents. Its correlation with purity and leukocyte fraction and its concentration in EMT and GS tumours point to stromal abundance and to the mesenchymal phenotype described by the ACRG and others [6, 19]. In the ACRG cohort the prior's association no longer met the prespecified threshold after subtype adjustment. This needs careful reading. The EMT subtype was itself defined from mesenchymal gene expression, so adjusting for it partly adjusts for what the score measures, and the result shows overlapping information rather than a proxy of a separate entity. In TCGA-STAD, where subtypes rest on genomic features, the association persisted, but that cohort supplied the weights. DNA-based composition measures attenuated but did not abolish the association. Bulk expression cannot determine whether the remainder reflects matrix remodelling, fibroblast states or their spatial organisation.

These findings suggest what a new gastric ECM signature should demonstrate. Significance in another cohort and after stage adjustment can be achieved by scores that track stromal content, and is therefore not sufficient evidence of novelty. A new signature should report its correlation with, and conditional information beyond, a core-matrisome average; its relation to molecular subtype in a cohort where subtype is known; and its performance against matched random signatures. It should also state which public cohorts were used in its development, because most available validation cohorts are shared across studies, and it should give a computable formula with the expression scale on which its coefficients apply. In our benchmark, published coefficients could be applied in only 18 of 27 signature–cohort pairs, two signatures reported no coefficients, one formula contained a typographical symbol, and one signature changed its conclusion with representation. These checks are inexpensive and could change the interpretation of several published scores.

Limitations

This is an Analysis of existing public data, and its conclusions are bounded by the size and nature of those data. The evidence-weighted prior was constructed in a single cohort, TCGA-STAD (387 patients with survival data, 160 deaths), and was evaluated externally in two cohorts, GSE84437 (431 patients, 207 deaths) and GSE62254 (300 patients, 152 deaths); the tumour–normal comparison in GSE26942 rested on 205 tumour and only 12 normal profiles. These cohorts are moderate in size, so modest differences in discrimination between scores could have gone undetected, and the absence of a significant improvement over the core-matrisome average should not be read as proof of equivalence. Both external cohorts are from Korean centres, and molecular subtype annotation

was available for only one of them, so the subtype result rests on a single external cohort and its generalisability to other populations is untested.

No new validating study was performed. The work is retrospective, uses no prospectively collected samples, and includes no laboratory, histological or experimental validation of the module activities, the shared axis or its interpretation as stromal abundance; the composition analysis relies on published DNA-based estimates for TCGA-STAD only. Although decision rules were fixed before the relevant outcomes were evaluated, all cohorts and the related literature were public. Scores are cohort-relative and do not establish single-patient use or calibrated risk, and concordance does not address calibration or clinical net benefit. The literature benchmark is small: the search used one database and one screener and yielded nine eligible signatures and eleven source-independent evaluations, pair-level statistics are descriptive because pairs share signatures and cohorts, published coefficients were applied on scales that differ from the source studies, and the matched-signature nulls were specified after the conditional results were seen. Confirmation in larger, independent and ethnically diverse cohorts with molecular annotation, and direct measurement of matrix architecture, for example by spatial transcriptomics or quantitative histology of collagen organisation, will be needed to determine whether ECM biology carries prognostic information beyond stromal abundance.

5 Conclusion

A biologically weighted ECM expression score for gastric cancer transported across platforms and remained prognostic. It nevertheless tracked the same dominant module-activity axis as an unweighted core-matrisome average and did not add detectable information beyond it. In the external cohort with molecular annotation, its association could not be separated from molecular subtype by the prespecified rule. The nine published gastric ECM signatures examined largely shared this axis and, in the primary analysis, added information beyond it in only one of eleven evaluations in cohorts not used to develop or validate them. The findings support reporting information beyond a core-matrisome reference and beyond molecular subtype, together with matched random-signature controls and fully computable formulas. Although the shared-axis finding is specific to the gastric-cancer cohorts examined, the benchmarking framework could serve as a testable template for assessing pathway-based prognostic signatures in other cancers and complex diseases.

Declarations

Ethical Approval and Accordance

Not applicable. This study did not involve the recruitment of human participants, the collection of human tissue or the use of animals. It is a secondary analysis of publicly available, de-identified datasets, listed under Data availability, whose original collection was approved by the ethics committees of the source studies; formal ethics committee approval was therefore not required for the present work.

Consent to participate

Not applicable.

Consent to publish

Not applicable.

Clinical trial number

Not applicable.

Competing interests

The authors declare no competing interests.

Funding

This work forms part of the Vietnam Academy of Science and Technology science and technology task ĐHCNCL.02/26-29, funded through its Science and Technology Development Fund.

Author contributions

N.V.A. supervised the study and reviewed its design, analyses and interpretation. L.T.T. developed the methodological framework, designed the study and analyses, and led the interpretation of the results and preparation of the manuscript. D.Q.T. executed the computational workflows and verified the analysis code. All authors jointly determined the analytical steps, reviewed and confirmed the reported results, revised and approved the submitted manuscript, and accept responsibility for the work.

Data availability

All data analysed in this study are publicly available; no new data were generated. The Cancer Genome Atlas stomach adenocarcinoma cohort (project TCGA-STAD) was obtained from the NCI Genomic Data Commons (https://portal.gdc.cancer.gov/projects/TCGA-STAD): open-access RNA-sequencing gene expression quantification files (Transcriptome Profiling, STAR - Counts workflow) and the corresponding clinical data were retrieved through the GDC application programming interface with TCGAbiolinks on 7 July 2026. Three gene-expression series were obtained from the NCBI Gene Expression Omnibus: GSE84437 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84437), GSE62254 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62254) and GSE26942 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26942). Probes were annotated with the GEO platform records GPL6947 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL6947) and GPL570 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL570).

Molecular subtypes for TCGA-STAD were taken from the cBioPortal study `stad_tcga_pan_can_atlas_2018` (https://www.cbioportal.org/study/summary?id=stad_tcga_pan_can_atlas_2018; file `data_clinical_patient.txt` from https://github.com/cBioPortal/datahub/tree/master/public/stad_tcga_pan_can_atlas_2018, downloaded 19 October 2026), and ACRG subtypes and follow-up codes for GSE62254 from Supplementary Table 1 of Cristescu et al. [6] (https://doi.org/10.1038/nm.3850). Tumour purity and leukocyte fraction for TCGA-STAD were taken from the TCGA PanCanAtlas files `TCGA_mastercalls.abs_tables_JSedit.fixed.txt` (https://api.gdc.cancer.gov/data/4f277128-f793-4354-a13d-30cc7fe9f6b5) and `TCGA_all_leuk_estimate.masked.20170107.tsv` (https://api.gdc.cancer.gov/data/6f75c9d7-5134-4ed1-b8f3-72856c98a4e8), listed at https://gdc.cancer.gov/about-data/publ

ications/pancanatlas. Matrisome categories were the Matrisome Project gene sets distributed in the Molecular Signatures Database and retrieved with the msigdbr package: NABA_COLLAGENS (https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/NABA_COLLAGENS.html), NABA_ECM_GLYCOPROTEINS (https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/NABA_ECM_GLYCOPROTEINS.html), NABA_PROTEOGLYCANS (https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/NABA_PROTEOGLYCANS.html), NABA_ECM_REGULATORS (https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/NABA_ECM_REGULATORS.html), NABA_ECM_AFFILIATED (https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/NABA_ECM_AFFILIATED.html) and NABA_SECRETED_FACTORS (https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/NABA_SECRETED_FACTORS.html). Protein and tissue annotations were taken from the Human Protein Atlas files proteinatlas.tsv (https://www.proteinatlas.org/download/proteinatlas.tsv.zip, accessed 7 July 2026), and normal_tissue.tsv and pathology.tsv of release 22 (https://v22.proteinatlas.org/download/normal_tissue.tsv.zip and https://v22.proteinatlas.org/download/pathology.tsv.zip). The literature benchmark used PubMed, searched on 19 October 2026; the exact query and the PubMed identifiers of all 106 retrieved records are available in the Zenodo archive (https://doi.org/10.5281/zenodo.23529709, file results/paper1_signatures_D/pubmed_search_D.json), together with every screening decision (file results/paper1_signatures_D/screening_log_D.csv), so that the search can be rerun and checked.

Code, analysis protocols, gene lists, derived per-patient scores, signature specifications, checksums of every downloaded input, and all random mappings and bootstrap outputs are available at Zenodo (https://doi.org/10.5281/zenodo.23529709). Redistribution of the source data follows the terms of the original providers. Additional results are provided in the Supplementary Information.

References

[1] Richard O Hynes. The extracellular matrix: not just pretty fibrils. *Science*, 326(5957):1216–1219, 2009. doi: 10.1126/science.1176009.

[2] Pengfei Lu, Valerie M Weaver, and Zena Werb. The extracellular matrix: a dynamic niche in cancer progression. *Journal of Cell Biology*, 196(4):395–406, 2012. doi: 10.1083/jcb.201102147.

[3] Michael W Pickup, Janna K Mouw, and Valerie M Weaver. The extracellular matrix modulates the hallmarks of cancer. *EMBO Reports*, 15(12):1243–1253, 2014. doi: 10.15252/embr.201439246.

[4] Alexandra Naba, Karl R Clauser, Sebastian Hoersch, Hui Liu, Steven A Carr, and Richard O Hynes. The matrisome: in silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices. *Molecular & Cellular Proteomics*, 11(4):M111.014647, 2012. doi: 10.1074/mcp.m111.014647.

[5] Xinhao Shao, Clarissa D Gomez, Nandini Kapoor, et al. MatrisomeDB 2.0: 2023 updates to the ECM-protein knowledge database. *Nucleic Acids Research*, 51(D1):D1519–D1530, 2023. doi: 10.1093/nar/gkac1009.

[6] Razvan Cristescu, Jeeyun Lee, Michael Nebozhyn, Kyoung-Mee Kim, Jason C. Ting, Swee Seong Wong, et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. *Nature Medicine*, 21(5):449–456, 2015. doi: 10.1038/nm.3850.

[7] The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, 513(7517):202–209, 2014. doi: 10.1038/nature13480.

[8] Zhi Yang, Feifei Xue, Minhuan Li, Xingya Zhu, Xiaofeng Lu, Chao Wang, En Xu, Xingzhou Wang, Liang Zhang, Heng Yu, Chuanfu Ren, Hao Wang, Yizhou Wang, Jie Chen, Wenxian Guan, and Xuefeng Xia. Extracellular matrix characterization in gastric cancer helps to predict prognosis and chemotherapy response. *Frontiers in Oncology*, 11:753330, 2021. doi: 10.3389/fonc.2021.753330.

[9] Xiangchou Yang, Liping Chen, Yuting Mao, Zijing Hu, and Muqing He. Progressive and prognostic performance of an extracellular matrix-receptor interaction signature in gastric cancer. *Disease Markers*, 2020:8816070, 2020. doi: 10.1155/2020/8816070.

[10] Ziqian Deng, Ting Guo, Jiwang Bi, Gangjian Wang, Ying Hu, Hong Du, Yuan Zhou, Shuqin Jia, Xiaofang Xing, and Jiafu Ji. Transcriptome profiling of patient-derived tumor xenografts suggests novel extracellular matrix-related signatures for gastric cancer prognosis prediction. *Journal of Translational Medicine*, 21:638, 2023. doi: 10.1186/s12967-023-04473-0.

[11] Nan Xu, Taojing Zhang, Weiwei Sun, Chenxiao Ye, and Huamiao Zhou. Identification of an extracellular matrix signature for predicting prognosis and sensitivity to therapy of patients with gastric cancer. *Scientific Reports*, 15:7464, 2025. doi: 10.1038/s41598-025-88376-8.

[12] Xin Tang, Yu Liu, Jiarong Zhao, Changfang Fu, and Wulin Yang. Subtyping of gastric cancer based on basement membrane genes that stratifies the prognosis, immune infiltration and therapeutic response. *Discover Oncology*, 15(1):362, 2024. doi: 10.1007/s12672-024-01238-z.

[13] Zhi-Yang Liu and Lin Xin. Identification of a basement membrane-related genes signature to predict prognosis, immune landscape and guide therapy in gastric cancer. *Medicine*, 102(39): e35027, 2023. doi: 10.1097/MD.0000000000035027.

[14] Jinyun Wang, Dingwei Liu, Qixuan Wang, and Yong Xie. Identification of basement membrane-related signatures in gastric cancer. *Diagnostics*, 13(11):1844, 2023. doi: 10.3390/diagnostics13111844.

[15] Longxiang Xie, Linghao Cai, Fei Wang, Lu Zhang, Qiang Wang, and Xiangqian Guo. Systematic review of prognostic gene signature in gastric cancer patients. *Frontiers in Bioengineering and Biotechnology*, 8:805, 2020. doi: 10.3389/fbioe.2020.00805.

[16] David Venet, Jacques E. Dumont, and Vincent Detours. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Computational Biology*, 7 (10):e1002240, 2011. doi: 10.1371/journal.pcbi.1002240.

[17] Antonio Colaprico, Tiago C Silva, Catharina Olsen, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, 44(8):e71, 2016. doi: 10.1093/nar/gkv1507.

[18] Jae-Ho Cheong, Han-Kwang Yang, Hyunki Kim, Woo Ho Kim, Young-Woo Kim, Myeong-Cherl Kook, et al. Predictive test for chemotherapy response in resectable gastric cancer: a multi-cohort, retrospective analysis. *The Lancet Oncology*, 19(5):629–638, 2018. doi: 10.1016/S1470-2045(18)30108-6.

[19] Sang Cheul Oh, Bo Hwa Sohn, Jae-Ho Cheong, Sang-Bae Kim, Jae Eun Lee, Ki Cheong Park, et al. Clinical and genomic landscape of gastric cancer with a mesenchymal phenotype. *Nature Communications*, 9(1):1777, 2018. doi: 10.1038/s41467-018-04179-8.

[20] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, 2013. doi: 10.1093/nar/gks1193.

[21] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, et al. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6):417–425, 2015. doi: 10.1016/j.cels.2015.12.004.

[22] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57 (1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

[23] Mathias Uhlén, Linn Fagerberg, Björn M Hallström, et al. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015. doi: 10.1126/science.1260419.

[24] Mathias Uhlén, Cheng Zhang, Sunjae Lee, et al. A pathology atlas of the human cancer transcriptome. *Science*, 357(6352):eaan2507, 2017. doi: 10.1126/science.aan2507.

[25] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–202, 1972. doi: 10.1111/j.2517-6161.1972.tb00899.x.

[26] Frank E Harrell, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, 1996. doi: 10.1002/(sici)1097-0258(19960229)15:4<361::aid-sim168>3.0.co;2-4.

[27] Frederick M. Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I. Olopade, Jakob N. Kather, Nicole Cipriani, Robert L. Grossman, and Alexander T. Pearson. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications*, 12(1):4423, 2021. doi: 10.1038/s41467-021-24698-1.

[28] Cameron Davidson-Pilon. lifelines: survival analysis in Python. *Journal of Open Source Software*, 4(40):1317, 2019. doi: 10.21105/joss.01317.

[29] Arseniy E. Yuzhalin, Tomas Urbonas, Michael A. Silva, Ruth J. Muschel, and Alex N. Gordon-Weeks. A core matrisome gene signature predicts cancer outcome. *British Journal of Cancer*, 118(3):435–440, 2018. doi: 10.1038/bjc.2017.458.

[30] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, 14:7, 2013. doi: 10.1186/1471-2105-14-7.

[31] Lu Wang, Meng-Han Wang, Yao-Hong Yuan, Rui-Ze Xu, Lu Bai, and Mi-Zhu Wang. Identification and validation of extracellular matrix-related genes in the progression of gastric cancer with intestinal metaplasia. *World Journal of Gastrointestinal Oncology*, 17(6):105160, 2025. doi: 10.4251/wjgo.v17.i6.105160.

[32] Chenbin Liu, Linyi Deng, Jinrong Lin, Jianjun Zhang, Shu Huang, Jinglin Zhao, Peipei Jin, Peiqing Xu, Peihua Ni, Dakang Xu, Le Ying, and Yiqun Hu. Zinc finger protein CTCF

regulates extracellular matrix (ECM)-related gene expression associated with the Wnt signaling pathway in gastric cancer. *Frontiers in Oncology*, 10:625633, 2020. doi: 10.3389/fonc.2020.625633.

[33] Katherine A. Hoadley, Christina Yau, Toshinori Hinoue, Denise M. Wolf, Alexander J. Lazar, Esther Drill, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304.e6, 2018. doi: 10.1016/j.cell.2018.03.022.

[34] Vesteinn Thorsson, David L. Gibbs, Scott D. Brown, Denise Wolf, Dante S. Bortone, Tai-Hsien Ou Yang, et al. The immune landscape of cancer. *Immunity*, 48(4):812–830.e14, 2018. doi: 10.1016/j.immuni.2018.03.023.

[35] Yang Liu, Nilay S. Sethi, Toshinori Hinoue, Barbara G. Schneider, Andrew D. Cherniack, Francisco Sanchez-Vega, et al. Comparative molecular analysis of gastrointestinal adenocarcinomas. *Cancer Cell*, 33(4):721–735.e8, 2018. doi: 10.1016/j.ccell.2018.03.010.

[36] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404, 2012. doi: 10.1158/2159-8290.CD-12-0095.

[37] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S. Onur Sumer, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6(269):pl1, 2013. doi: 10.1126/scisignal.2004088.

[38] Rongbo Han, Fei Wang, Xiujuan Wang, and Benxin Zhao. Bioinformatics analysis of macrophage-associated genes reveals prognostic signatures and immune landscape in gastric cancer. *Discover Oncology*, 16:2254, 2025. doi: 10.1007/s12672-025-04019-4.

[39] Reza Akhavan, Mahdi Foroughian, and Forouzan Amerizadeh. Integrative systems biology and drug repurposing reveal key regulatory hubs and a prognostic signature in gastric cancer. *Discover Oncology*, 17:907, 2026. doi: 10.1007/s12672-026-04989-z.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf