Machine-Learning Classification of Colorectal Cancer from Targeted Serum Metabolomics: Discriminative Metabolites and Concordance with Known Disease Alterations

Abstract

Background. Colorectal cancer (CRC) is the second leading cause of cancer death worldwide, yet participation in existing screening programmes remains limited, motivating the search for minimally invasive, blood-based detection strategies. Serum metabolomics offers a functional readout of the metabolic reprogramming that accompanies tumourigenesis and is therefore an attractive substrate for biomarker discovery. **Methods.** We re-analysed a publicly available targeted serum metabolic-profiling dataset (Metabolomics Workbench study ST000284; 224 participants; 113 named metabolites quantified by LC–MS/MS) and built classifiers for the cancer-versus-healthy-control contrast. The 76 colorectal-polyp samples were deliberately held aside, leaving 64 CRC and 84 healthy sera (cancer prevalence 0.43). The abundance matrix was processed with half-minimum imputation, median (sample-wise) normalization, $\log_2$ transformation and per-metabolite auto-scaling. An L_1-regularized logistic-regression model (Lasso) and a random forest were evaluated with stratified 5-fold cross-validation using out-of-fold predictions, and discriminative features were ranked by mean model weight and by univariate statistics computed on the pre-scaling $\log_2$ data. **Results.** Both models achieved comparable, moderate discrimination on out-of-fold data (Lasso: AUROC 0.807, AUPRC 0.803; random forest: AUROC 0.813, AUPRC 0.806; both AUPRC values far exceeding the prevalence baseline of 0.43). At a threshold calibrated to 90% specificity, the Lasso detected 64% of cancers. Sixteen of 113 metabolites were significant at a 5% false-discovery rate under both parametric and rank-based tests. The most discriminative features included depleted histidine ($\downarrow$) and elevated glutamic acid, proline, hydroxyproline, kynurenate, ornithine, conjugated bile acids and hippurate ($\uparrow$)—a pattern that recapitulates independently reported CRC serum-metabolite alterations. **Conclusions.** A compact, interpretable serum-metabolite signature separates CRC patients from healthy controls with performance consistent with a single-analyte-class assay, and its top-ranked features are biologically coherent with tumour amino-acid reprogramming, tryptophan–kynurenine activation and bile-acid dysregulation. The analysis is fully reproducible; per-sample predicted probabilities and code are provided.

Keywords: colorectal cancer; serum metabolomics; biomarkers; machine learning; Lasso; random forest; liquid biopsy; LC–MS/MS

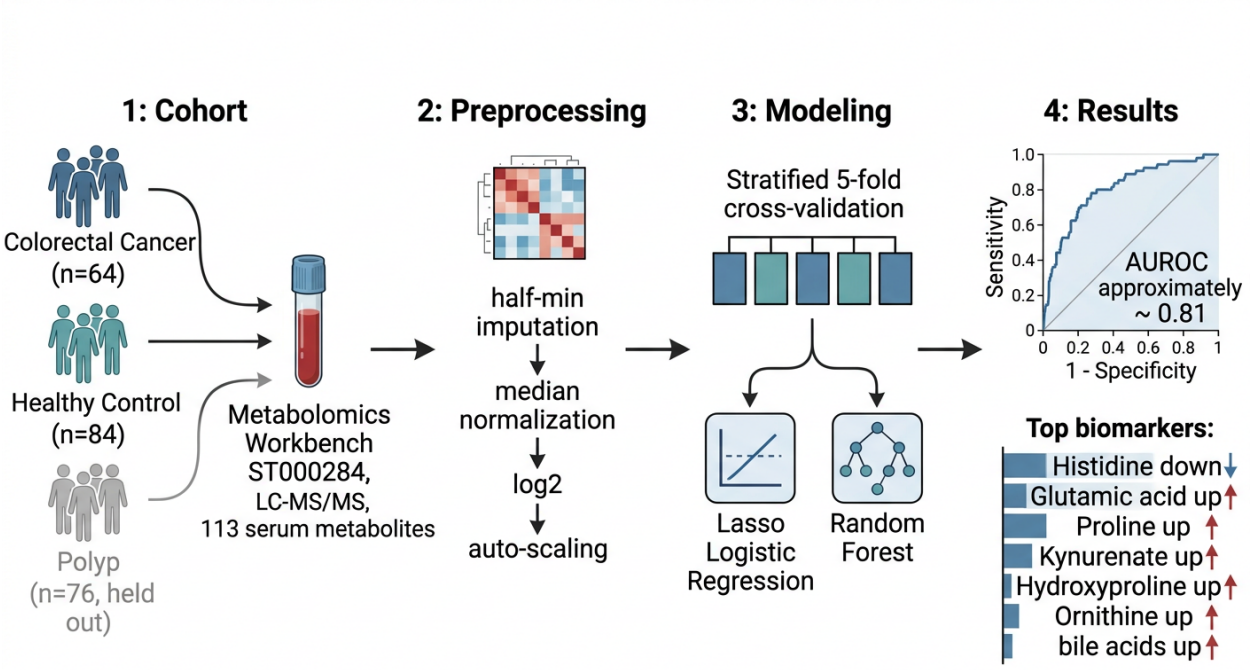


Figure 1: Graphical abstract. Study workflow from cohort assembly through preprocessing, cross-validated modelling and biomarker ranking. Targeted LC–MS/MS serum profiles of 64 colorectal-cancer and 84 healthy participants (the 76 polyp samples were held aside) were normalized and modelled with Lasso logistic regression and a random forest under stratified 5-fold cross-validation, yielding an out-of-fold AUROC of ≈ 0.81 and a compact, biologically coherent panel of discriminative metabolites.

1 Introduction

Colorectal cancer (CRC) is among the most common malignancies and a leading cause of cancer-related mortality. Global estimates place the annual burden at roughly 1.9 million new cases and close to one million deaths, with incidence rising in many regions as populations age and lifestyle exposures shift [1, 2]. Analyses of long-term trends within the Global Burden of Disease framework similarly document a more than two-fold increase in incident CRC over recent decades, underscoring the scale of the public-health challenge [3, 4]. Because CRC develops slowly through a well-characterized adenoma–carcinoma sequence, it is, in principle, highly amenable to early detection and prevention [5].

The clinical difficulty is not the absence of effective screening but its uneven uptake. Colonoscopy is sensitive and permits removal of premalignant lesions, yet it is invasive, resource-intensive and subject to substantial non-adherence; stool-based tests such as the faecal immunochemical test improve participation but trade away sensitivity for advanced neoplasia [6]. These limitations have spurred intense interest in blood-based alternatives that might be integrated into routine care. Circulating tumour DNA has demonstrated compelling performance for detecting minimal residual disease after surgery [7], and multi-cancer early-detection assays based on cell-free DNA methylation have shown that a single blood draw can flag many tumour types, albeit with stage-dependent and modest sensitivity for early gastrointestinal disease [8]. Nucleic-acid liquid biopsies, however, report primarily on tumour genotype and shedding; they capture little of the systemic physiological state of the host. A complementary strategy is to interrogate the small-molecule end-products of

metabolism, which integrate genetic, transcriptional, enzymatic and microbial influences into a functional phenotype [9–11].

Metabolic reprogramming is now recognized as a core hallmark of cancer [12, 13]. Proliferating tumour cells rewire central carbon metabolism—the aerobic glycolysis captured by the Warburg effect [14, 15]—and place heavy demands on amino-acid supply for biosynthesis, redox balance and signalling [16]. Such intracellular changes propagate into the circulation, where altered fluxes of amino acids, organic acids, bile acids and nucleotide derivatives can be detected in serum or plasma. This premise has been repeatedly validated in CRC. Targeted and untargeted blood-metabolomics studies have identified reproducible signatures separating cases from controls, frequently implicating amino-acid metabolism: histidine is consistently depleted, whereas glutamate, proline and tryptophan-pathway metabolites are commonly elevated [17–20]. Discovery–replication designs in large prospective cohorts have reinforced a subset of these associations and highlighted bile acids—including glycochenodeoxycholate—as circulating correlates of CRC risk [21, 22]. Collectively, these observations support the biological plausibility of a serum-metabolite classifier and provide an external yardstick against which any newly derived signature can be judged.

Translating metabolomic signals into a classifier nonetheless demands methodological care. Metabolomics matrices are high-dimensional relative to sample size, exhibit heavy-tailed and heteroscedastic intensity distributions, and are vulnerable to sample-to-sample dilution effects; each of these must be addressed before modelling [23–25]. Equally, the modest sample sizes typical of single-cohort studies make performance estimates fragile: feature selection performed on the full dataset, or hyperparameters tuned on the same folds used for evaluation, produce optimistically biased accuracy [26, 27]. Rigorous, leakage-free cross-validation and honest reporting of class-imbalance-aware metrics are therefore essential.

In this work we revisit a well-curated, publicly deposited targeted serum-metabolomics dataset of CRC detection and construct classifiers for the cancer-versus-healthy contrast under a transparent, reproducible pipeline. We pursue three objectives. First, we quantify how well a compact serum-metabolite signature discriminates CRC from health, reporting the area under the receiver-operating-characteristic curve (AUROC), the area under the precision–recall curve (AUPRC) relative to prevalence, sensitivity at a clinically motivated 90% specificity, and the confusion matrix. Second, we identify the metabolites the models rank most discriminative and specify their direction of change in cancer. Third, we ask whether these data-driven features overlap the serum-metabolite alterations previously reported for CRC, thereby testing the internal signal against the accumulated external literature. Throughout, we favour interpretable models and provide per-sample predicted probabilities and code to support scrutiny and reuse.

2 Materials and Methods

2.1 Data source and cohort definition

Data were obtained from the NIH Common Fund Metabolomics Workbench [28], study `ST000284` ("Colorectal Cancer Detection Using Targeted Serum Metabolic Profiling"; University of Washington; *Homo sapiens*; LC–MS; released under a CC BY 4.0 licence). The study corresponds to a targeted serum metabolic-profiling effort for CRC detection [29]. The processed named-metabolite abundance table was retrieved programmatically through the public REST API (`/summary`, `/factors` and `/data` endpoints; access date 12 July 2026) and aligned to the participant factor metadata by local sample identifier. The retrieved matrix comprised 224 serum samples × 113 named metabolites,

with abundances expressed as raw peak intensities (counts per second). Group labels resolved to 64 colorectal-cancer (CRC), 76 colorectal-polyp and 84 healthy-control participants, matching the expected distribution; no missing or zero values were present in the processed table.

Because the analytical target of this study is the diagnostic separation of manifest cancer from health, the 76 polyp samples were *held aside* rather than merged into either class. Polyps are a heterogeneous, intermediate biological state along the adenoma–carcinoma continuum; assigning them to "healthy" would contaminate the negative class with premalignant physiology, whereas assigning them to "cancer" would inflate the positive class with non-malignant lesions. Holding them aside yields a clean, well-posed binary contrast and keeps the polyp group available for future adenoma-focused analyses. The modelling cohort therefore consisted of 148 participants (64 CRC, 84 healthy), giving a cancer prevalence of 0.432.

2.2 Preprocessing and normalization

The abundance matrix was processed with a standard metabolomics pipeline applied strictly in the following order; the rationale for each step follows established practice for LC–MS data [23–25].

1. **Half-minimum imputation (safeguard).** Any zero or missing value was replaced by half of the smallest non-zero value observed for that metabolite. As expected from the curated table, no such values were present; the step is retained as a defensive guard for reproducibility.
2. **Median normalization.** Each sample was divided by its own median abundance and rescaled by the grand mean of all sample medians. This corrects for systematic per-sample differences in injected amount and dilution. Sample medians ranged from 2.31×10^5 to 5.98×10^5 (grand mean 3.19×10^5).
3. $\log_2$ **transformation.** Intensities were $\log_2$-transformed to stabilize variance, compress the dynamic range and render the heavy-tailed distributions approximately symmetric.
4. **Auto-scaling (unit variance).** Each metabolite was mean-centred and divided by its standard deviation, so that all metabolites contribute on a comparable scale and the analysis is not dominated by the most abundant species.

To avoid inflating fold-change or effect-size estimates, all univariate biological statistics (below) were computed on the median-normalized, $\log_2$-transformed matrix *before* auto-scaling, preserving the natural units of biological variation; a reconstruction check confirmed that this pre-scaling matrix reproduced the modelling input exactly (maximum absolute difference 8.9×10^{-16}). Principal-component analysis was used only for exploratory visualization of sample structure before and after preprocessing.

2.3 Classifiers and cross-validation

Two complementary classifiers were trained on the 148×113 auto-scaled matrix. The first was an L_1-regularized (Lasso) logistic-regression model, which performs embedded feature selection by shrinking uninformative coefficients to zero [30]; the inverse regularization strength C was tuned by nested (inner) stratified 5-fold cross-validation over a logarithmic grid of 20 values spanning 10^{-3} to 10^2, optimizing inner-fold AUROC (`liblinear` solver). The second was a random forest of 100 trees (maximum depth 5, minimum three samples per leaf, $\sqrt{p}$ features considered per split), an

ensemble able to capture non-linear structure and interactions [31]. All models were implemented in `scikit-learn` [32].

Performance was estimated by *outer* stratified 5-fold cross-validation (shuffled, fixed random seed 42). For every sample, the predicted probability used for evaluation was the one produced when that sample lay in the held-out test fold, so that no sample contributed to the model that scored it. Crucially, Lasso hyperparameter tuning occurred within each training fold only, eliminating the selection-bias that arises when tuning or feature selection is performed on the full dataset [26, 27]. Pooling these out-of-fold probabilities yields an unbiased, leakage-free estimate of generalization for the full cohort of 148 participants.

2.4 Evaluation metrics

Discrimination was summarized by the AUROC and, because the classes are imbalanced, by the AUPRC, which we interpret against the no-skill baseline equal to the cancer prevalence (0.432) [33]. To reflect a screening-oriented operating point that limits false positives, we identified the probability threshold delivering at least 90% specificity on the out-of-fold predictions and report the corresponding sensitivity. Confusion matrices are reported at the default 0.5 threshold and at the 90%-specificity threshold. Accuracy, sensitivity and specificity accompany the default-threshold matrices.

2.5 Biomarker ranking and literature integration

Discriminative importance was quantified two ways. For the Lasso, coefficients were extracted from each of the five outer training folds and averaged; the sign of the mean coefficient encodes the direction of association with cancer (positive = higher odds of cancer), and the fraction of folds in which a coefficient was non-zero (selection frequency) indexes stability. For the random forest, mean impurity-decrease feature importances were averaged across folds. In parallel, for every metabolite we computed the $\log_2$ fold change between cancer and healthy groups on the pre-scaling data, Welch's t-test and the Mann–Whitney U test, Cohen's d, and Benjamini–Hochberg false-discovery-rate (FDR) adjusted p-values. Finally, each metabolite was cross-referenced against a curated set of previously reported CRC serum/plasma metabolite alterations drawn from the primary literature [17–19, 21, 22, 29], and its data-driven direction of change was annotated as consistent or inconsistent with the reported direction.

2.6 Reproducibility and availability

The complete analysis is deterministic under the fixed seed. Per-sample out-of-fold predicted probabilities for both classifiers, the full ranked-biomarker table (all 113 metabolites with model weights, fold changes, effect sizes and FDR values), and the analysis scripts accompany this manuscript. The underlying data are openly available from the Metabolomics Workbench under study `ST000284`.

3 Results

3.1 Cohort, data quality and preprocessing

The curated named-metabolite table contained no missing or zero entries across the 224×113 matrix, confirming the maturity of the deposited processing. After holding aside the 76 polyp samples, the binary modelling cohort comprised 64 CRC and 84 healthy sera. Median normalization removed the substantial per-sample intensity offsets (sample medians spanning a 2.6-fold range), and the subsequent $\log_2$ transformation and auto-scaling produced the symmetric, unit-variance feature distributions required for regularized and tree-based modelling. Exploratory principal-component analysis showed that raw intensities were dominated by a few high-variance components reflecting technical scale (the first five components explained 88% of variance), whereas after preprocessing the variance was distributed across many components (first five components 49%), consistent with the removal of dominant technical structure and the exposure of finer biological variation.

3.2 Classification performance

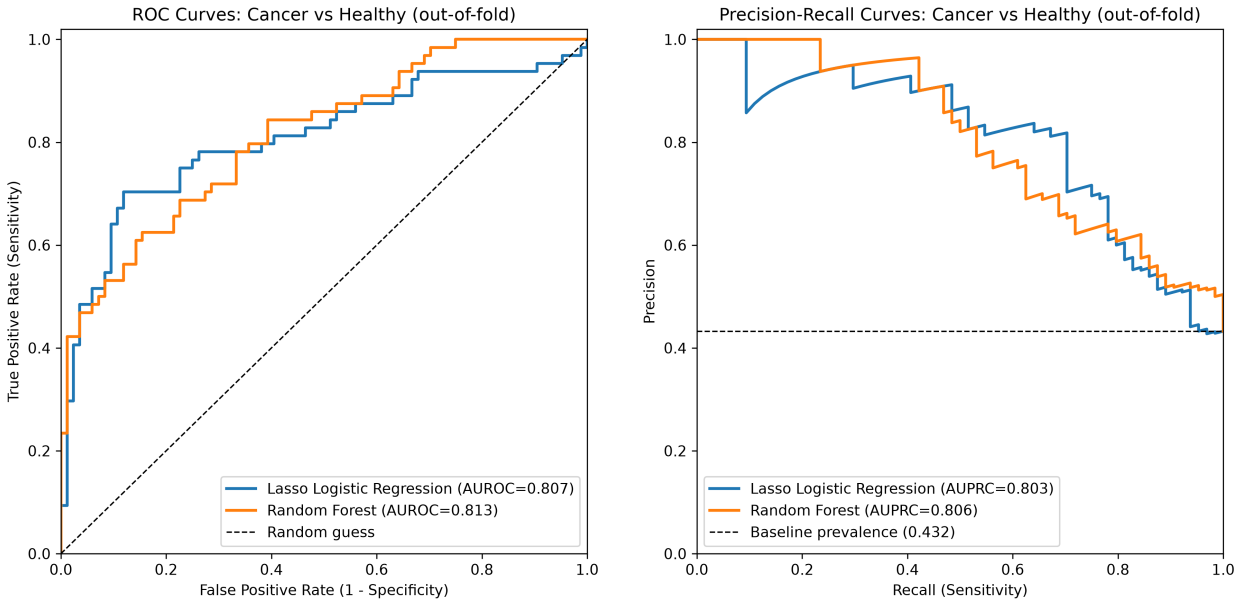
Both classifiers achieved moderate and closely matched discrimination on out-of-fold predictions (Table 1; Figs. 2a–3). The random forest attained an AUROC of 0.813 and the Lasso 0.807; their AUPRC values (0.806 and 0.803, respectively) sit far above the prevalence baseline of 0.432, indicating that positive predictions are substantially enriched for true cancers relative to chance. The near-superimposable ROC curves (Fig. 2a) and precision–recall curves (Fig. 2b) indicate that a sparse linear model and a non-linear ensemble extract essentially the same discriminative signal, which argues that the separability of CRC and healthy sera in this dataset is largely additive rather than driven by higher-order interactions.

Table 1: Out-of-fold classification performance for the cancer-versus-healthy contrast ($n = 148$; 64 cancer, 84 healthy; prevalence = 0.432). Metrics are pooled across the five stratified test folds.

Metric	**Lasso logistic regression**	**Random forest**
AUROC	0.807	0.813
AUPRC	0.803	0.806
AUPRC baseline (prevalence)	0.432	0.432
Sensitivity at 90% specificity	0.641	0.531
At default threshold = 0.5:		
Accuracy	0.757	0.730
Sensitivity	0.734	0.563
Specificity	0.774	0.857

The two models differ mainly in calibration and in how they trade sensitivity against specificity. At a threshold tuned to 90% specificity—an operating point appropriate for a screening adjunct where false positives carry cost—the Lasso recovered 64% of cancers, outperforming the random forest (53%) at the same specificity target. At the default 0.5 threshold the random forest was more conservative (specificity 0.857, sensitivity 0.563) while the Lasso was better balanced (specificity 0.774, sensitivity 0.734, accuracy 0.757). The corresponding confusion matrices (Fig. 3) make the error structure explicit: at threshold 0.5 the Lasso misclassified 19 of 84 healthy participants and 17 of 64 cancers, whereas the random forest reduced false positives to 12 at the cost of 28 missed cancers. For a detection setting the better-calibrated, more sensitive Lasso operating point is preferable, and

its interpretable coefficients further recommend it as the primary model.



(a) Receiver-operating-characteristic curves.

(b) Precision–recall curves.

Figure 2: Out-of-fold discrimination. (a) ROC curves for the Lasso (AUROC 0.807) and random forest (AUROC 0.813); the diagonal denotes chance. (b) Precision–recall curves; both classifiers lie well above the prevalence-defined no-skill level (0.432), the appropriate baseline for AUPRC on imbalanced data.

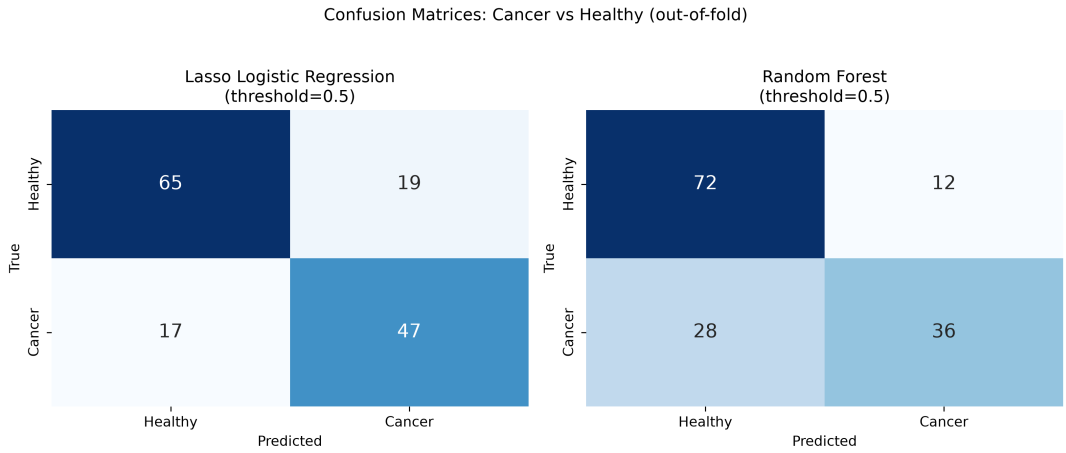


Figure 3: Confusion matrices at the default 0.5 **threshold** (out-of-fold). The Lasso (left) achieves a balanced error profile (65 true-negative, 19 false-positive, 17 false-negative, 47 true-positive), whereas the random forest (right) is more specific but misses more cancers (72, 12, 28, 36).

3.3 Most discriminative metabolites

Feature selection was highly consistent across folds: the Lasso retained a compact subset (between 27 and 56 of 113 metabolites per fold; median 40), confirming that a minority of metabolites carries the discriminative signal. Ranking metabolites by mean Lasso coefficient (Fig. 4) placed **histidine**

as by far the strongest single feature and the dominant marker of the healthy class (mean coefficient -2.27, selected in 100% of folds), i.e. histidine is *depleted* in cancer. The features most strongly associated with cancer (positive coefficients) included **epinephrine**, **maleic acid**, **ornithine**, **glycochenodeoxycholate**, **hippuric acid**, **4-hydroxyphenylpyruvate**, **hydroxyproline** and **kynurenate**; additional negatively weighted (healthy-associated) features included **glycerate**, **methionine** and **malonic acid/3-hydroxybutyrate**.

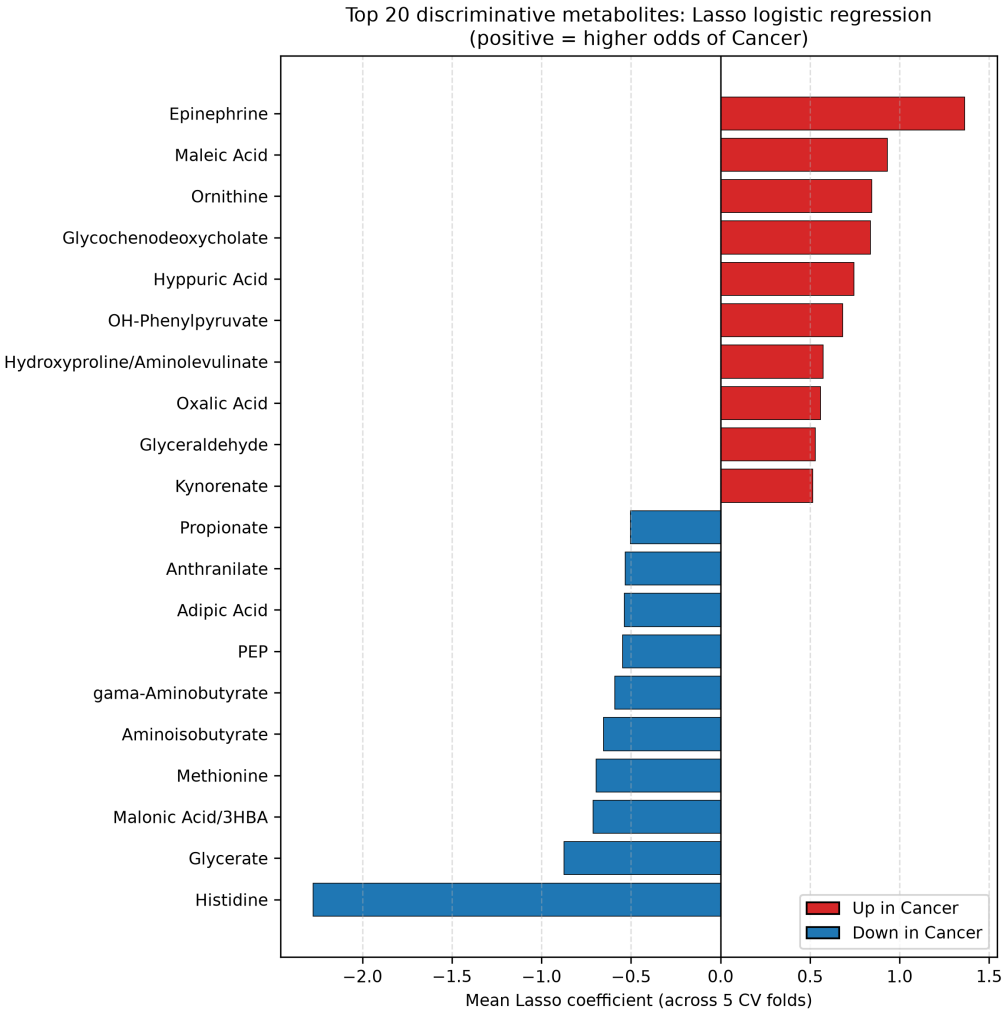


Figure 4: Top-20 metabolites by mean Lasso coefficient (averaged across the five outer training folds). Positive coefficients (red) indicate higher odds of cancer; negative coefficients (blue) indicate association with the healthy class. Histidine is the single strongest discriminator and is depleted in cancer.

The random-forest importance ranking (Fig. 5) agreed closely at the top, foregrounding metabolites with the largest effect sizes: **glycochenodeoxycholate** ($\log_2$FC $+1.26$), **hippuric acid** ($+1.07$), **glyceraldehyde** ($+0.38$), **linolenic acid** (-0.45) and **malonic acid/3-hydroxybutyrate** (-0.87), followed by **glycocholate**, **proline**, **glutamic acid**, **histidine** and **kynurenate**. The concordance between an L_1-linear model and an impurity-based ensemble on the identity and direction of the leading features is reassuring, because the two methods weight features by different criteria (regression log-odds versus impurity reduction) and are susceptible to different artefacts.

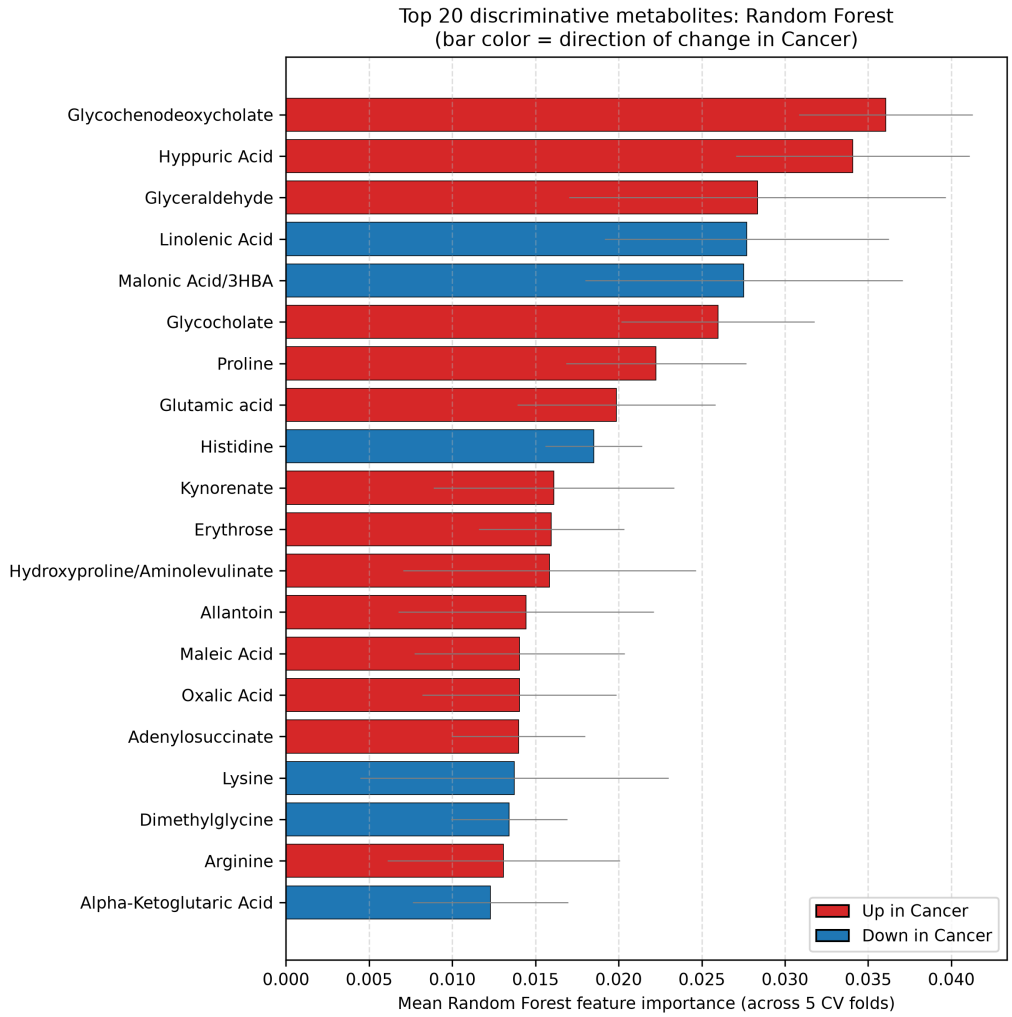


Figure 5: Top-20 metabolites by mean random-forest importance (averaged across folds). Bar colour encodes the direction of change in cancer (red, elevated; blue, reduced). Conjugated bile acids and hippurate show the largest importances, followed by amino acids including proline, glutamate and histidine.

3.4 Univariate significance and direction of change

Sixteen of the 113 metabolites reached significance at a 5% FDR, and—reassuringly—the same 16 were significant under both the parametric Welch test and the non-parametric Mann–Whitney test, indicating that the signals are not artefacts of distributional assumptions. The volcano plot (Fig. 6) integrates effect size and significance and shows that the largest and most significant changes are dominated by increases in cancer, notably the conjugated bile acids glycochenodeoxycholate and glycocholate, hippurate, glyceraldehyde, hydroxyproline, maleate, glutamate, proline and kynurenate, alongside decreases in histidine, malonate/3-hydroxybutyrate and linolenate. Table 2 consolidates the leading discriminative metabolites with their direction, effect size, significance and, where applicable, their status as previously reported CRC serum markers.

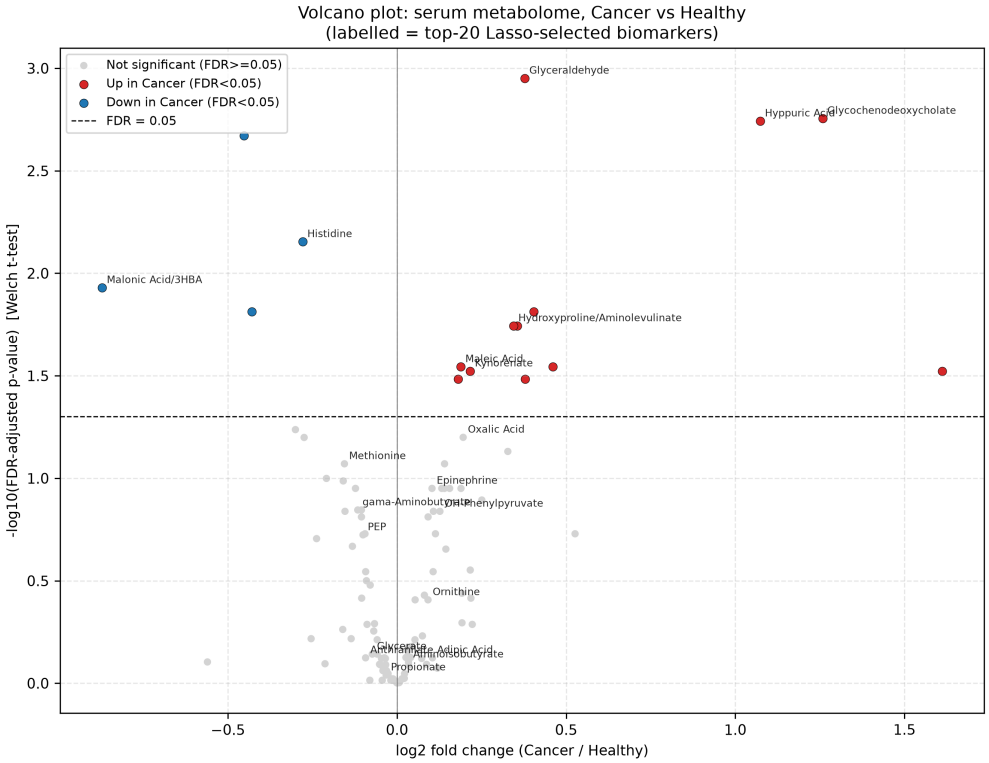


Figure 6: Volcano plot of the serum metabolome (cancer versus healthy). The horizontal axis is the $\log_2$ fold change (cancer/healthy) computed on median-normalized, $\log_2$ pre-scaling data; the vertical axis is $-\log_{10}$ of the FDR-adjusted Welch p-value. Red and blue points are significant increases and decreases in cancer (FDR < 0.05); labelled points are the top-20 Lasso-selected features. The dashed line marks FDR $= 0.05$.

Table 2: Leading discriminative serum metabolites, direction of change in cancer, univariate statistics (computed on pre-scaling $\log_2$ data), and concordance with previously reported CRC serum/plasma alterations. Arrows: ↑ elevated, ↓ depleted in cancer. FDR is the Benjamini–Hochberg-adjusted Welch p-value. "Known CRC marker" flags metabolites with a directional report in the cited literature.

Metabolite	**Dir.**	$\log_2$**FC**	**Cohen's** d	**FDR**	**Lasso coef.**	**Known CRC**
Histidine	↓	−0.28	−0.61	0.007	−2.27	Yes (consistent)
Glycochenodeoxycholate	↑	+1.26	+0.70	0.002	+0.84	—
Hippuric acid	↑	+1.07	+0.73	0.002	+0.74	—
Glyceraldehyde	↑	+0.38	+0.79	0.001	+0.53	—
Malonic acid / 3-HBA	↓	−0.87	−0.61	0.012	−0.71	—
Glycocholate	↑	+1.61	+0.47	0.030	—	—
Hydroxyproline	↑	+0.34	+0.55	0.018	+0.57	Yes (consistent)
Glutamic acid	↑	+0.36	+0.52	0.018	+0.33	Yes (consistent)
Proline	↑	+0.18	+0.49	0.033	—	Yes (consistent)
Kynurenate	↑	+0.22	+0.51	0.030	+0.51	Yes (consistent)
Maleic acid	↑	+0.19	+0.51	0.029	+0.93	—
Allantoin	↑	+0.41	+0.55	0.015	—	—
Ornithine	↑	+0.09	+0.22	0.392	+0.84	Yes (consistent)

3.5 Overlap with known CRC serum-metabolite alterations

Of the 113 profiled metabolites, 30 carried a directional annotation in the curated CRC literature. Twenty of these (67%) changed in the direction reported previously; the ten discordant cases were, without exception, statistically non-significant with near-zero fold changes, so their nominal direction reflects noise rather than genuine disagreement. Critically, the concordance was strongest precisely among the features the models ranked highest. Histidine—the top Lasso feature—was significantly depleted, matching the well-replicated histidine decrease in CRC [17, 29]. Elevations in glutamic acid and proline reproduce reported increases in these amino acids [17, 19], hydroxyproline is consistent with heightened collagen turnover in the tumour stroma, kynurenate elevation is consistent with tryptophan–kynurenine pathway activation, and ornithine elevation aligns with urea-cycle remodelling. Thus the internal, data-driven signal and the accumulated external evidence converge on the same biological picture (see Discussion), which materially strengthens confidence that the ≈ 0.81 AUROC reflects true metabolic reprogramming rather than dataset-specific artefact.

4 Discussion

4.1 Principal findings

Working from an openly deposited targeted serum-metabolomics dataset, we built leakage-free classifiers that separate colorectal-cancer patients from healthy controls with an out-of-fold AUROC of approximately 0.81 and an AUPRC nearly double the prevalence baseline. Two very different learning algorithms—a sparse logistic regression and a random forest—delivered near-identical discrimination and, more importantly, converged on the same compact set of discriminative metabolites. The signature is biologically coherent: histidine is depleted, while glutamate, proline, hydroxyproline, kynurenate, ornithine, conjugated bile acids and hippurate are elevated in cancer. Because these directions match independently reported CRC serum alterations, the result is best read not as a stand-alone diagnostic claim but as a reproducible, mechanistically interpretable demonstration that circulating metabolites encode a detectable CRC phenotype.

4.2 Biological interpretation of the top biomarkers

The leading features map onto several well-established axes of tumour metabolism. **Histidine depletion** is one of the most consistently reported serum changes in CRC and featured here as the single strongest discriminator. Beyond serving as a proteinogenic substrate, histidine catabolism modulates one-carbon/tetrahydrofolate pools; enhanced flux through this pathway in tumours draws down circulating histidine and has direct therapeutic resonance, since histidine catabolism is a determinant of methotrexate sensitivity [34]. The coordinated **elevation of glutamate and proline** reflects the amino-acid demands of proliferating cells: glutamine/glutamate anaplerosis feeds the tricarboxylic-acid cycle and nitrogen metabolism, while proline biosynthesis and the proline cycle support redox balance and collagen production [14, 16]. The rise in **hydroxyproline**, a near-unique constituent of collagen, is a plausible circulating footprint of the accelerated extracellular-matrix remodelling and collagen turnover that characterize the desmoplastic CRC stroma.

The **kynurenate** increase points to activation of the tryptophan–kynurenine pathway, a central node linking tumour metabolism to immune evasion. Up-regulation of indoleamine-2,3-dioxygenase-1 in colorectal tumours diverts tryptophan toward kynurenine and its derivatives; higher tumoural expression of this enzyme is associated with more aggressive disease and worse outcome, and the

resulting metabolites reshape the immune microenvironment [35]. A serum kynurenate elevation is therefore consistent with, and a candidate peripheral readout of, this immunometabolic reprogramming. The **ornithine** increase is compatible with urea-cycle and polyamine-pathway remodelling that supports rapid proliferation.

Among the metabolites with the largest effect sizes were the **conjugated bile acids** glycochenodeoxycholate and glycocholate. Circulating bile acids have repeatedly been linked to colorectal carcinogenesis; a large prospective serum-metabolomics analysis identified glycochenodeoxycholate as positively associated with CRC risk [22], and bile-acid dysregulation reflects the interlinked host–microbiome axis implicated in colonic tumour promotion. The concomitant elevation of **hippurate**—a host–microbial co-metabolite of dietary aromatic compounds—similarly points to a shifted gut-microbial milieu. That these microbially influenced metabolites ranked among the strongest features underscores the value of serum metabolomics as an integrated readout of tumour, host and microbiome, information not accessible to tumour-DNA-based liquid biopsies [7, 8].

4.3 Concordance with the literature and its significance

The overlap analysis provides an important sanity check. Two-thirds of previously annotated CRC markers moved in the expected direction, and every apparent discordance was a non-significant, near-null change—so the classifier is not trading on idiosyncratic features but on the same biology reported by independent cohorts using different platforms and populations [17–19, 21]. This convergence matters because single-cohort metabolomics models are notoriously vulnerable to overfitting and batch effects [26, 27]; agreement with external evidence is among the most persuasive forms of validation available in the absence of an independent replication set.

4.4 Comparison with prior performance and the role of the operating point

The ≈ 0.81 AUROC obtained here is more modest than the 0.9-plus figures reported by some CRC metabolomics studies [17, 18, 20]. Several factors reconcile this. First, our estimates derive strictly from out-of-fold predictions with hyperparameter tuning confined to training folds; performance measured this way is systematically lower than, but more honest than, estimates from models tuned or selected on the evaluation data. Second, some high headline figures come from panels optimized and reported on the same samples, or from cohorts with different case mix and stage distribution. Third, our deliberate exclusion of polyps sharpens the biological contrast but removes the "easy" separability sometimes gained by contrasting cancer against a maximally healthy reference. We regard the value reported here as a realistic single-cohort estimate. The choice of operating point is equally consequential: at 90% specificity the Lasso detected 64% of cancers, a profile suited to a triage or adjunct role that could enrich referral to colonoscopy rather than replace it.

4.5 Strengths and limitations

The chief strengths of this study are methodological transparency and reproducibility: a public dataset, an explicit and conventional preprocessing pipeline, leakage-free nested cross-validation, class-imbalance-aware metrics interpreted against prevalence, and the release of per-sample predictions and code. The convergence of two dissimilar models and the concordance with external literature further buttress the findings. Several limitations temper interpretation. The cohort is single-centre and modest in size ($n = 148$), so the estimates carry appreciable sampling variance and require external, ideally prospective, validation before any clinical inference. The targeted panel

of 113 metabolites, while curated, is narrower than contemporary untargeted platforms and may miss informative species. A handful of features are reported as isobaric or co-eluting bundles (for example, hydroxyproline with aminolevulinate, or leucine with isoleucine), so annotations reference the dominant compound and warrant confirmatory identification. Clinical covariates such as stage, tumour location, age, sex, body-mass index, diet and medication were not modelled here and could confound or, if incorporated, sharpen the signal. Finally, cross-sectional case–control data cannot establish whether these metabolic changes precede diagnosis and are therefore useful for early detection, or arise as a consequence of established disease; only prospective pre-diagnostic sampling can resolve that question [21, 22].

4.6 Future directions

Natural extensions include external validation in independent cohorts and harmonized multi-cohort training to test transportability across platforms and populations; incorporation of the held-out polyp group to model the full adenoma–carcinoma progression and evaluate detection of premalignant disease; integration of metabolomics with orthogonal liquid-biopsy modalities such as cell-free-DNA methylation, which may be complementary rather than redundant [8]; and adjustment for clinical covariates together with formal assessment of clinical utility through decision-curve analysis. Confirmatory identification and absolute quantification of the leading features—particularly the bile acids, kynurenate and hydroxyproline—would move promising markers toward assay development.

4.7 Conclusion

A compact, interpretable serum-metabolite signature distinguishes colorectal-cancer patients from healthy controls with performance consistent with a single-analyte-class blood assay, and its most discriminative features—histidine depletion alongside elevations in glutamate, proline, hydroxyproline, kynurenate, ornithine, conjugated bile acids and hippurate—recapitulate the metabolic reprogramming, immunometabolic activation and microbiome dysregulation established for colorectal cancer. Delivered with full code and per-sample predictions, the analysis offers a transparent, reproducible reference point for serum-metabolomics-based CRC detection and a biologically grounded shortlist of candidate biomarkers for prospective evaluation.

Data and Code Availability

The primary data are openly available from the NIH Metabolomics Workbench (`https://www.metabolomicsworkbench.org`) under study `ST000284` (CC BY 4.0). Per-sample out-of-fold predicted probabilities (`per_sample_predictions.csv`), the full ranked-biomarker table for all 113 metabolites (`ranked_biomarkers.csv`), and the analysis scripts accompany this manuscript.

Conflicts of Interest

The authors declare no competing interests.

References

[1] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021. doi: 10.3322/caac.21660.

[2] Eileen Morgan, Melina Arnold, Andrea Gini, Valentina Lorenzoni, Citadel J. Cabasag, Mathieu Laversanne, Jerome Vignat, Jacques Ferlay, Neil Murphy, and Freddie Bray. Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN. *Gut*, 72(2):338–344, 2023. doi: 10.1136/gutjnl-2022-327736.

[3] GBD 2019 Colorectal Cancer Collaborators. Global, regional, and national burden of colorectal cancer and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Gastroenterology & Hepatology*, 7(7):627–647, 2022. doi: 10.1016/S2468-1253(22)00044-9.

[4] Rebecca L. Siegel, Angela N. Giaquinto, and Ahmedin Jemal. Cancer statistics, 2024. *CA: A Cancer Journal for Clinicians*, 74(1):12–49, 2024. doi: 10.3322/caac.21820.

[5] Evelien Dekker, Pieter J. Tanis, Jasper L. A. Vleugels, Pashtoon M. Kasi, and Michael B. Wallace. Colorectal cancer. *The Lancet*, 394(10207):1467–1480, 2019. doi: 10.1016/S0140-6736(19)32319-0.

[6] Jennifer S. Lin, Leslie A. Perdue, Nora B. Henrikson, Sarah I. Bean, and Paula R. Blasi. Screening for colorectal cancer: Updated evidence report and systematic review for the US preventive services task force. *JAMA*, 325(19):1978–1998, 2021. doi: 10.1001/jama.2021.4417.

[7] Jeanne Tie, Yuxuan Wang, Cristian Tomasetti, Lu Li, Simeon Springer, Isaac Kinde, Natalie Silliman, Mark Tacey, Hui-Li Wong, Michael Christie, et al. Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. *Science Translational Medicine*, 8(346):346ra92, 2016. doi: 10.1126/scitranslmed.aaf6219.

[8] Eric A. Klein, David Richards, Allen Cohn, Mohan Tummala, Roger Lapham, Douglas Cosgrove, Gina Chung, Jessica Clement, Jane Gao, Nathan Hunkapiller, et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Annals of Oncology*, 32(9):1167–1177, 2021. doi: 10.1016/j.annonc.2021.05.806.

[9] Jeremy K. Nicholson and John C. Lindon. Systems biology: Metabonomics. *Nature*, 455(7216):1054–1056, 2008. doi: 10.1038/4551054a.

[10] Caroline H. Johnson, Julijana Ivanisevic, and Gary Siuzdak. Metabolomics: beyond biomarkers and towards mechanisms. *Nature Reviews Molecular Cell Biology*, 17(7):451–459, 2016. doi: 10.1038/nrm.2016.25.

[11] David S. Wishart. Emerging applications of metabolomics in drug discovery and precision medicine. *Nature Reviews Drug Discovery*, 15(7):473–484, 2016. doi: 10.1038/nrd.2016.32.

[12] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, 2011. doi: 10.1016/j.cell.2011.02.013.

[13] Douglas Hanahan. Hallmarks of cancer: New dimensions. *Cancer Discovery*, 12(1):31–46, 2022. doi: 10.1158/2159-8290.CD-21-1059.

[14] Matthew G. Vander Heiden, Lewis C. Cantley, and Craig B. Thompson. Understanding the Warburg effect: The metabolic requirements of cell proliferation. *Science*, 324(5930):1029–1033, 2009. doi: 10.1126/science.1160809.

[15] Maria V. Liberti and Jason W. Locasale. The Warburg effect: How does it benefit cancer cells? *Trends in Biochemical Sciences*, 41(3):211–218, 2016. doi: 10.1016/j.tibs.2016.01.004.

[16] Natalya N. Pavlova and Craig B. Thompson. The emerging hallmarks of cancer metabolism. *Cell Metabolism*, 23(1):27–47, 2016. doi: 10.1016/j.cmet.2015.12.006.

[17] Yang Yang, Zhen Wang, Xiaotong Li, Jia Lv, Rui Zhong, Shanshan Gao, Fan Zhang, and Wei Chen. Profiling the metabolic disorder and detection of colorectal cancer based on targeted amino acids metabolomics. *Journal of Translational Medicine*, 21(1):824, 2023. doi: 10.1186/s12967-023-04604-7.

[18] Farshad Farshidfar, Aalim M. Weljie, Karen A. Kopciuk, Robert Hilsden, S. Elizabeth McGregor, W. Donald Buie, Anthony MacLean, Hans J. Vogel, and Oliver F. Bathe. A validated metabolomic signature for colorectal cancer: exploration of the clinical value of metabolomics. *British Journal of Cancer*, 115(7):848–857, 2016. doi: 10.1038/bjc.2016.243.

[19] Chuandong Zhang, Shengke Zhou, Hong Chang, Fengfeng Zhuang, Yanhua Shi, Lei Chang, Wenpeng Ai, Juan Du, Wenqi Liu, Hongyuan Liu, Xin Zhou, Zhiwei Wang, and Tao Hong. Metabolomic profiling identified serum metabolite biomarkers and related metabolic pathways of colorectal cancer. *Disease Markers*, 2021:6858809, 2021. doi: 10.1155/2021/6858809.

[20] Yao Yi, Jing Wang, Cheng Liang, Chunyan Ren, Xin Lian, Chengjian Han, and Wei Sun. LC-MS-based serum metabolomics analysis for the screening and monitoring of colorectal cancer. *Frontiers in Oncology*, 13:1173424, 2023. doi: 10.3389/fonc.2023.1173424.

[21] Anne J. M. R. Geijsen, Stefanie Brezina, Pekka Keski-Rahkonen, Andreas Baierl, Thomas Bachleitner-Hofmann, Michael M. Bergmann, Juergen Boehm, Hermann Brenner, Jenny Chang-Claude, Franzel J. B. van Duijnhoven, et al. Plasma metabolites associated with colorectal cancer: A discovery-replication strategy. *International Journal of Cancer*, 145(5):1221–1231, 2019. doi: 10.1002/ijc.32146.

[22] Amanda J. Cross, Steven C. Moore, Simina Boca, Wen-Yi Huang, Xiaoqin Xiong, Rachael Stolzenberg-Solomon, Rashmi Sinha, and Joshua N. Sampson. A prospective study of serum metabolites and colorectal cancer risk. *Cancer*, 120(19):3049–3057, 2014. doi: 10.1002/cncr.28799.

[23] Robert A. van den Berg, Huub C. J. Hoefsloot, Johan A. Westerhuis, Age K. Smilde, and Mariët J. van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7:142, 2006. doi: 10.1186/1471-2164-7-142.

[24] Jianguo Xia, Nick Psychogios, Nelson Young, and David S. Wishart. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Research*, 37(suppl_2):W652–W660, 2009. doi: 10.1093/nar/gkp356.

[25] Zhiqiang Pang, Yao Lu, Guangyan Zhou, Fiona Hui, Lei Xu, Charles Viau, Aline F. Spigelman, Patrick E. MacDonald, David S. Wishart, Shuzhao Li, and Jianguo Xia. MetaboAnalyst 6.0: towards a unified platform for metabolomics data analysis. *Nucleic Acids Research*, 52(W1):W398–W406, 2024. doi: 10.1093/nar/gkae246.

[26] Christophe Ambroise and Geoffrey J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, 99(10): 6562–6566, 2002. doi: 10.1073/pnas.102102699.

[27] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006. doi: 10.1186/1471-2105-7-91.

[28] Manish Sud, Eoin Fahy, Dawn Cotter, Kenan Azam, Ilango Vadivelu, Charles Burant, Arthur Edison, Oliver Fiehn, Richard Higashi, K. Sreekumaran Nair, Susan Sumner, and Shankar Subramaniam. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Research*, 44(D1):D463–D470, 2016. doi: 10.1093/nar/gkv1045.

[29] Jiangjiang Zhu, Danijel Djukovic, Lingli Deng, Haiwei Gu, Farhan Himmati, E. Gabriela Chiorean, and Daniel Raftery. Colorectal cancer detection using targeted serum metabolic profiling. *Journal of Proteome Research*, 13(9):4120–4130, 2014. doi: 10.1021/pr500494u.

[30] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

[31] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[32] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. doi: 10.5555/1953048.2078195.

[33] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

[34] Naama Kanarek, Heather R. Keys, Jason R. Cantor, Caroline A. Lewis, Sze Ham Chan, Tenzin Kunchok, Monther Abu-Remaileh, Elizaveta Freinkman, Lawrence D. Schweitzer, and David M. Sabatini. Histidine catabolism is a major determinant of methotrexate sensitivity. *Nature*, 559 (7715):632–636, 2018. doi: 10.1038/s41586-018-0316-7.

[35] Liesbeth Ferdinande, Christine Decaestecker, Laurine Verset, Anne Mathieu, Xavier Moles Lopez, Ana-Maria Negulescu, Tom Van Maerken, Isabelle Salmon, Claude A. Cuvelier, and Pieter Demetter. Clinicopathological significance of indoleamine 2,3-dioxygenase 1 expression in colorectal cancer. *British Journal of Cancer*, 106(1):141–147, 2012. doi: 10.1038/bjc.2011.513.