International Journal of Software & Hardware Research in Engineering (IJSHRE)
ISSN:2347-4890, ijournals.in/ijshre
Volume 14 Issue 9, September 2026

Plasma Metabolomics Does not Improve Prostate Biopsy Prediction Over the Clinical Variables

Author: Taeyun Ha
Affiliation: Eastlake High School
Email: Hataekorea@gmail.com
https://doi.org/10.26821/ijshre.14.09.2026.140902

ABSTRACT

Prostate cancer detection begins with the PSA blood test, but PSA is prostate specific and not cancer specific, so many men with a raised value are sent to biopsy and turn out benign. This problem is worst in the gray zone, a PSA between 4 and 10 ng/mL, where the test carries little information. Plasma metabolomics has been proposed as a richer signal, and a recent deep learning model on this same cohort reported an AUC of 0.89, which is far above what simple methods reach. In this study I tested whether a Graph Attention Network that connects metabolites through KEGG pathways can predict biopsy outcome better than flat feature models and better than the routine clinical variables that are already available, with a focus on the gray zone where the decision actually matters. I used the Early Detection Research Network cohort of 580 men and 1,169 metabolites. Every model was evaluated with repeated five-fold cross-validation, and all data-dependent steps were fit inside the training folds only. The Graph Attention Network reached an AUROC of 0.559, which was not better than XGBoost at 0.556 and not better than a logistic regression given only age and PSA at 0.563. Removing the graph and keeping only self-loops did not lower performance and slightly improved it. The KEGG graph that the hypothesis proposed was the worst of all the graph variants. The metabolites added nothing over age and PSA. The clinical variables alone reached an AUROC of 0.72 for any cancer, 0.81 for clinically significant cancer, and 0.70 inside the gray zone, while metabolomics stayed at chance in the gray zone. I also showed that the gap with the published 0.89 is largely explained by feature selection performed before cross validation rather than inside it, which inflated the AUROC by 0.14 to 0.20 in my own data. No single metabolite survived multiple testing correction. The conclusion is a clean negative result. Plasma metabolomics does not help predict biopsy outcome in this cohort, and the usable signal is in cheap clinical data.

1. INTRODUCTION

Prostate cancer is one of the most commonly diagnosed cancers in men and the second leading cause of cancer related death among men in the United States, with roughly 333,830 new cases and 36,320 deaths projected for a single year (American Cancer Society, 2026). Because the disease is so common, even a small improvement in how it is detected would affect a large number of patients, and a large amount of research is directed at making early detection both more specific and less invasive. Detection today begins with the Prostate Specific Antigen (PSA) blood test, which measures the concentration of a protein produced by the prostate. An elevated value is treated as a sign of raised risk, and men whose PSA is concerning are referred for a prostate biopsy, where needles are used to remove small cores of prostate tissue for closer examination. Although biopsy is the definitive diagnostic step, it is invasive, uncomfortable, and carries risks of bleeding, infection, and urinary complications, so sending a man to one when he does not have cancer can be harmful.

The weakness of PSA is not accuracy in a general sense but a lack of cancer specificity. PSA is prostate specific rather than cancer specific, so it is released whether or not prostate tissue is malignant, and it rises with benign prostatic hyperplasia, with prostatitis, and simply with age as the gland enlarges. A raised value is therefore a reliable sign that the prostate is producing more antigen and a poor sign of why. This weakness is most consequential in what clinicians call the gray zone, a PSA concentration between 4 and 10 ng/mL (Sun W., 2025).

Within this range a large number of men who go on to biopsy are found to have benign tissue, meaning they underwent the procedure without any benefit, while a large number do have cancer, so the test still cannot be ignored. In the cohort used in this project, 342 of 580 men were within this zone, which shows how much of the clinical decision load sits exactly where PSA lacks specificity. Improving the decision before biopsy inside the gray zone is the main goal of this research.

A second issue is that not all detected cancers are harmful. Prostate cancer tumors are graded using a Gleason score, and tumors graded Gleason 6 are largely indolent and often never progress to symptoms or death, so they require much less attention. Detecting these tumors leads to overtreatment, which exposes men to surgery or radiation and to lasting side effects that were not necessary. Therefore the second question is not only whether a man has cancer but whether he has a cancer that is clinically significant, which I define as Gleason ≥7 versus benign.

Metabolites are the small molecules produced during the metabolism of the body, and because tumors restructure their metabolism as they grow, the concentrations of certain metabolites in blood plasma may carry a signature of malignancy. The best known example for prostate cancer is sarcosine, which was reported to rise during prostate cancer progression and was proposed as a strong candidate marker (Sreekumar et al., 2009). Its value as a marker was contested in later replication work, so I treat it as motivation rather than a proven marker. Following studies have analyzed the difference in metabolites between cancer and benign samples, and also between aggressive and indolent disease, in plasma, urine, and tissue (Yang et al., 2021; Amante et al., 2022; Lima et al., 2021; Huang et al., 2022; Schmidt et al., 2020; Tang et al., 2025). However several papers point to the difficulty of reproducing these signals. Penney and colleagues profiled prostate tumor tissue and serum, and although they recovered individually identified metabolites such as citrate, they were unable to build a reliable metabolite signature for the Gleason score, reaching AUROC values of only about 0.67 in tissue (Penney et al., 2021). This negative prior is important, because it sets up the negative finding of my own study, and I return to it in the Discussion.

Because the biopsy outcome is known for every participant, this cohort is a good candidate for machine learning, and it has already been used this way. Sun and colleagues applied a hybrid transformer and convolutional network named TransConvNet to this data and reported an accuracy of 81.03% and an AUC of 0.89, which they presented as substantially better than conventional machine learning (Sun L. et al., 2024). This gap is the central puzzle of my paper, because my own models on the same data type reach only about 0.55. A model reporting 0.89 while careful baselines sit near chance is a warning sign, and one likely explanation is that features were selected using all of the data before cross validation rather than inside the training folds, which is known to inflate results (Ambroise and McLachlan, 2002). My leakage experiment in Section 2.7 is designed to test exactly this.

Biological function is generally the product of many molecules interacting rather than any single molecule acting alone, and this idea underlies curated pathway resources such as the Kyoto Encyclopedia of Genes and Genomes, which represents metabolism as networks of connected compounds and reactions (Kanehisa and Goto, 2000). Recent work in cancer suggests that encoding pathway structure directly into a model can help. Lee and colleagues modeled each pathway as a graph convolutional network and combined many pathways using attention, which outperformed both conventional classifiers and a model given all 20,531 raw genes (Lee et al., 2020). A related biologically informed network for prostate cancer, P-NET, embedded a pathway hierarchy into a deep model and improved both prediction and interpretability (Elmarakeby et al., 2021). These results were obtained on gene expression, so it is not certain that the same benefit transfers to plasma metabolomics, and my results ultimately question that premise. A Graph Attention Network (GAT) works on graph structured data and uses an attention mechanism to learn how much weight each node should give to each of its neighbors (Veličković et al., 2018). Applied to metabolomics, each metabolite becomes a node, and metabolites that share a KEGG pathway are joined by an edge, so the model can in principle reason over metabolic relationships.

Hypothesis

I hypothesized that a Graph Attention Network encoding metabolite relationships derived from KEGG pathways would classify prostate biopsy outcomes more accurately than existing flat feature models, specifically

logistic regression, random forest, XGBoost, and TransConvNet, in this small sample setting of 580 patients and high dimensional setting of over one thousand metabolites. This is a falsifiable claim, and I note in advance that the setting is a small n and large p regime, meaning few patients but very many features, here 580 patients against more than one thousand metabolites. In this regime an over-parameterized model, one with more free parameters than the data can constrain, is prone to fitting noise, a framing that matters for interpreting the result.

2. MATERIALS AND METHODS

2.1 Dataset

This project uses Metabolomics Workbench Study ST002498, also indexed as project PR001613, which is the Weill Cornell Medicine cohort of the Early Detection Research Network (Benedetti et al., 2023; Sud et al., 2016), accessed in 2025. The study provides two files. The metabolomics file contains the measured metabolite values together with sample and metabolite annotations, and a separate clinical file contains 420 clinical variables per patient. Blood plasma was collected from men at the time they underwent prostate biopsy. Samples were profiled by untargeted liquid chromatography with tandem mass spectrometry across four platform arms covering positive and negative ionization modes. Detected features were matched against a reference library using retention time and accurate mass within about 10 parts per million, and fragmentation patterns were compared against authentic standards to assign identification confidence. The raw profiling reports 1,482 metabolites per patient, of which 1,155 are chemically identified and named while the rest are recorded as unknown features.

After merging the metabolomics and clinical files I had complete metabolomic and clinical data for 580 patients. Of these, 267 were diagnosed with prostate cancer through biopsy and 313 returned benign, which gives a positive rate of 46%. The counts reconcile exactly, since 267 plus 313 equals 580. Among the 267 cancers, 154 were graded Gleason ≥7 and 113 were graded Gleason 6. Looking at the pre biopsy PSA values, 342 of the 580 men were within the 4 to 10 ng/mL gray zone, and of those 156 had cancer and 186 returned benign, which again reconciles since 156 plus 186 equals 342. I keep this 342 figure consistent throughout the paper.

The provided file includes a preprocessed matrix, which I used so that my results are comparable to the earlier work on this cohort. The preprocessing applied by the data providers was probabilistic quotient normalization (Dieterle et al., 2006), then log transformation, then removal of metabolites missing in more than half of samples, then k nearest neighbors imputation of the values that remained, and this filtering reduced the panel from 1,482 to 1,169 metabolites and lowered the dimensionality. I return to a limitation of this file below. The unknown, unidentified features were kept in every flat model, because they carry measured values like any other feature, but they cannot map to a KEGG pathway and therefore carry only self loops in the graph. To check that the provider filtering did not discard the signal, I also reran the metabolomic models on the full raw matrix of 1,482 features, and the result did not change.

The clinical file provides the outcome label and many covariates. The label is a diagnosis variable coded as zero for benign and one for cancer, and I confirmed the coding by checking that every cancer case carries a Gleason score and no benign case does. The pre biopsy variables I used are age, pre biopsy PSA, prostate volume, percent free PSA, digital rectal examination findings, prior biopsy history, and family history. Two of these have notable missingness in the merged cohort, since percent free PSA is missing for about 66% of men and prostate volume for about 5%, and I imputed numeric covariates with the training fold median. A number of columns in the file were recorded after the biopsy, including Gleason scores, tumor stage, and per core pathology, and I excluded these from every predictive model, because they exist only as a consequence of the biopsy result and using them would be a direct leak of the outcome. The Gleason score was used only to define the clinically significant target.

2.2 Evaluation

All models are assessed using repeated stratified k fold cross validation with five folds and five repetitions, which gives 25 out of fold estimates, where the repetition reduces the influence of any single partition. Every data dependent transformation is fit on the training portion of each fold only and then applied without change to the test portion, which includes imputation, scaling, feature selection, and dimensionality reduction. Performance is

reported as the mean and standard deviation of the AUROC (the area under the receiver operating characteristic curve, where 0.50 means random guessing and 1.0 means perfect separation) across the 25 folds. The tables also report the AUPRC, the area under the precision-recall curve, and the accuracy. I acknowledge one residual leak that I could not fully remove. The provider preprocessed matrix was normalized and imputed across all 580 patients at once, so those specific steps were inherited rather than redone inside folds. I re did scaling, feature selection, and dimensionality reduction inside every fold, but the provider normalization and imputation could not be undone from the released file. This residual leak would, if anything, make the metabolomic estimates slightly optimistic, which strengthens rather than weakens a negative result.

2.3 Graph construction

Three edge definitions were implemented so that the contribution of each could be isolated. The first uses KEGG. Metabolite identifiers were mapped to KEGG pathways using the compound to pathway table from the KEGG REST interface, and two metabolites were joined whenever they shared at least one pathway. Coverage here is low, since 377 of the 1,169 metabolites carry a KEGG identifier and only 279, about 24%, map to at least one pathway, so most nodes end up with only a self loop. I treat this low coverage as a key limitation rather than a detail. The second definition uses the curated sub pathway annotation supplied with the data, which covers a larger fraction of the panel. The third joins metabolites whose absolute correlation across training patients exceeds 0.70. Self loops are always added so that every node keeps its own value, and very large pathway groups are capped at 40 members so that a single dense central metabolism group does not dominate the edge count. The resulting edge counts differ widely across variants and are reported in Table 2, which also shows that sharing any pathway can create dense hubs.

2.4 Graph attention network

The model is a Graph Attention Network implemented directly in PyTorch (Veličković et al., 2018). Each patient is represented as the same fixed graph with node values equal to that patient's metabolite measurements, so the topology is shared across patients while the node features vary. For full reproducibility the architecture is as follows. Each metabolite value is embedded from one dimension into a hidden vector of size 8 by a linear layer. Two graph attention layers follow, each with 4 attention heads and a per head width of 8, with LeakyReLU attention of slope 0.2 and dropout of 0.3 on the attention weights. Node representations are mean pooled into a single patient vector, concatenated with the two clinical values age and PSA, and passed to a small classifier head of one hidden layer of 32 units with ReLU and dropout 0.3, then a single output. Training used the Adam optimizer with learning rate 5e-3 and weight decay 1e-4, a batch size of 32, binary cross entropy loss, and a fixed budget of 30 epochs with no early stopping. The random seed was fixed at 42. I did not run a hyperparameter search, so these settings are fixed choices rather than tuned ones, and I note that with only 580 patients the regularization choices, especially dropout and weight decay, likely drive the near chance behavior.

2.5 Baselines and ablations

The GAT is compared against the flat feature models named in the hypothesis, which are logistic regression, random forest, and XGBoost (Chen and Guestrin, 2016; Pedregosa et al., 2011), all trained on the same folds with the same features. Two ablations isolate the contribution of the graph. The first replaces graph attention with mean aggregation, which is a graph convolutional network, so it removes the attention mechanism while keeping the structure (Kipf and Welling, 2017). The second, and the more important one, removes the graph entirely by keeping only self loops, so the identical network processes each metabolite independently with no message passing. This no graph control is what makes the comparison interpretable, because any difference between it and a graph variant is due to the graph alone.

2.6 Clinical comparators and prediction targets

A metabolomic model is only worthwhile if it improves on information that is already available at no extra cost, so I built clinical comparators. The clinical arm uses variables obtainable before biopsy, which are age, PSA, log PSA, PSA density defined as PSA divided by prostate volume, percent free PSA, prostate volume, digital rectal examination findings, prior biopsy history, and family history. Simple reference models using PSA alone, age alone, and age with PSA establish the floor that any metabolomic model must exceed, and the age plus PSA

model is reported in the main results table so the marginal contribution of the metabolites can be read directly. I use three prediction targets. The primary target is cancer versus benign across all 580 patients. The second is clinically significant cancer, defined as Gleason ≥7 versus benign, which excludes the 113 indolent Gleason 6 cancers. The third restricts the primary target to the 342 men inside the PSA gray zone, which is the population named in the hypothesis. Beyond discrimination, I also report sensitivity, specificity, and the proportion of biopsies avoided across decision thresholds for the clinical gray zone model, because the number of biopsies avoided is the most clinically meaningful output.

2.7 Quantifying the effect of feature selection leakage

Because the published benchmark for this cohort appears to have selected features before rather than inside cross validation, I ran a controlled experiment to quantify what that choice is worth. The same classifier, a logistic regression, and the same folds are used twice. In the first configuration the top k metabolites are selected by a univariate ANOVA F test (the f_classif filter) using all patients, and cross validation is then performed on the selected features. In the second the identical selection is performed inside each training fold only. The difference between the two isolates the inflation attributable to the ordering of these steps, and this is exactly the selection bias described by Ambroise and McLachlan (2002). I report both k equal to 20 and k equal to 50, for the full cohort and for the gray zone.

2.8 Biomarker discovery

Although the predictive result is negative, I also looked for candidate metabolites, since reporting the individual associations is useful for future work. Within the gray zone I fit a univariate logistic regression of the outcome on each standardized metabolite and report the odds ratio per one standard deviation with a 95% confidence interval, together with a Mann Whitney test and the single metabolite AUROC. All p values are corrected for multiple testing with the Benjamini Hochberg false discovery rate, because more than one thousand metabolites are tested and an uncorrected list would be mostly false positives. I flag whether each top metabolite has been reported in earlier prostate cancer work or is a new candidate. I also fit an XGBoost model and computed SHAP values (Lundberg and Lee, 2017) to rank model based importance. These discovery analyses are exploratory and are reported as hypothesis generating rather than confirmed.

2.9 Statistical analysis and software

Models are compared with the Wilcoxon signed rank test applied to paired fold level AUROC values, together with the count of folds in which one model beats another. Two limitations are acknowledged. Repeated k fold folds reuse the same patients and are therefore not independent, which makes the resulting p values optimistic, so I treat the fold win count as the more robust summary, and where a corrected test is appropriate the Nadeau and Bengio corrected resampled t test (Nadeau and Bengio, 2003) is the suitable choice for this non independence. In addition the signed rank test has a hard lower bound on achievable p values at small fold counts, which is why I use 25 folds rather than 5 for any comparison meant to support a claim. Analyses were carried out in Python using PyTorch, XGBoost, LightGBM (Ke et al., 2017), statsmodels, and SHAP (Pedregosa et al., 2011; Chen and Guestrin, 2016; Lundberg and Lee, 2017).

3. RESULTS

All experiments described in Section 2 were executed. Every figure reported is the mean across held out folds under the protocol in Section 2.2.

3.1 The graph attention network did not beat the flat models

The Graph Attention Network was compared against the flat feature models named in the hypothesis, using identical folds and identical features. The results are shown in Table 1. The GAT ranks highest by a very small amount, but the margin over XGBoost is 0.003 AUROC with a paired p value of 0.93, and every model lies close to 0.55, where 0.50 is chance. An AUROC of 0.55 sits only 0.05 above random guessing and well inside the fold-to-fold spread, so in practice it carries almost no ability to tell cancer from benign. The hypothesis predicted that

the graph model would be substantially more accurate than these baselines, and it is not. A second comparison makes this sharper. The GAT above received age and PSA in addition to the metabolite graph. A logistic regression given nothing but age and PSA reaches 0.563, while the GAT given those same two variables plus 1,169 metabolites arranged into a KEGG graph reaches 0.559. In other words the strongest single statement in this study is that the metabolites add nothing over age and PSA, and a plain logistic regression actually falls from 0.563 to 0.535 when the metabolites are added.

Table 1. Cancer versus benign across all 580 patients. The ± value is the standard deviation of the AUROC across the 25 folds. AUPRC and accuracy are fold means and are shown without SD for readability, since AUROC is the metric used for all claims. The age plus PSA only model is included as the reference floor. Chance AUROC is 0.50 and the majority class accuracy is 0.54

Model	AUROC (mean ± SD)	AUPRC	Accuracy
Age + PSA only (logistic regression)	0.563 ± 0.039	0.551	0.559
GAT (KEGG pathway graph)	0.559 ± 0.053	0.547	0.560
XGBoost	0.556 ± 0.043	0.523	0.544
GCN (attention removed)	0.553 ± 0.044	0.547	0.563
Logistic regression (metabolites + age/PSA)	0.535 ± 0.035	0.509	0.528
Random forest	0.535 ± 0.042	0.507	0.536

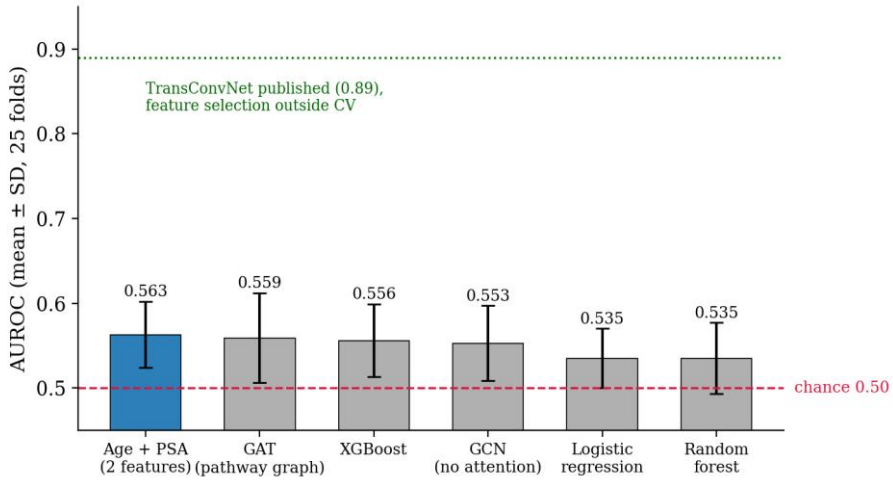

Figure 1. Every model sits just above the 0.50 chance line, and the age plus PSA reference matches the best metabolomic model.

3.2 Removing the graph improved performance

Table 2 reports the graph ablation, where the architecture and folds are identical and only the edge set changes. Every graph variant performed worse than the same network with no graph at all, and the KEGG derived graph, which is the one the hypothesis specifically proposed, performed worst of all at 0.545. The no graph control scored 0.557, which is not significantly different from XGBoost at 0.556. The reading is that message passing over these metabolites smooths noisy and weakly informative features into each other and removes what little separation exists, and the low KEGG coverage from Section 2.3 makes this worse, because most nodes have only self loops while the few well connected central metabolism nodes form dense hubs that wash out the attention signal.

Graph variant	Edges	AUROC	vs no graph	p
GAT, no graph (control)	0	0.557 ± 0.051	—	—
GAT, correlation edges	3,442	0.555 ± 0.047	−0.002	0.198
GAT, sub pathway edges	12,094	0.553 ± 0.052	−0.004	0.026
GAT, KEGG + correlation	10,418	0.549 ± 0.053	−0.007	0.121
GAT, KEGG pathway edges	7,302	0.545 ± 0.053	−0.011	0.018

Table 2.

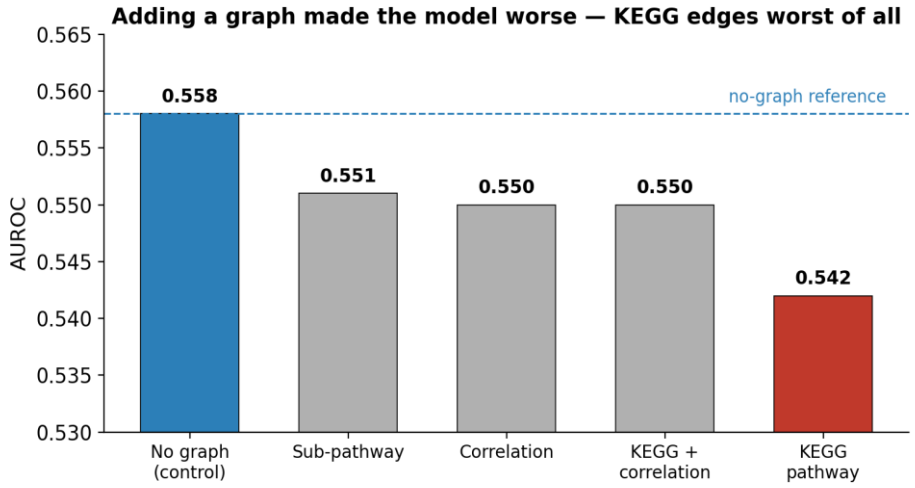

Figure 2. Graph ablation. No graph at all is the best result, and imposing structure only lowers the AUROC, with KEGG pathways the worst.

3.3 The predictive signal is clinical, across all three targets

The clinical comparators tell the opposite story from the metabolites. Using only routine pre biopsy variables, a logistic regression reaches an AUROC of 0.72 for any cancer and 0.81 for clinically significant Gleason ≥7 cancer, both far above the metabolomic ceiling of about 0.55 and consistent with published clinical risk calculators, which suggests the signal is real rather than a fluke of this cohort. Adding the metabolites to the clinical model did not help and slightly hurt it. Table 3 summarizes the three targets. Metabolomics never rises meaningfully above chance on any target, while the clinical model is well above chance on all three.

Prediction target	Clinical AUROC	Clinical accuracy	Best metabolomics AUROC	Majority baseline
Any cancer (n=580)	0.720	0.653	0.565	0.54
Gleason ≥7 vs benign	0.806	0.776	0.539	0.67
PSA gray zone (n=342)	0.702	0.661	0.563	0.54

Table 3. Clinical variables versus metabolites across the three prediction targets. The best metabolomics column is the strongest of logistic regression, XGBoost, and LightGBM on metabolites alone.

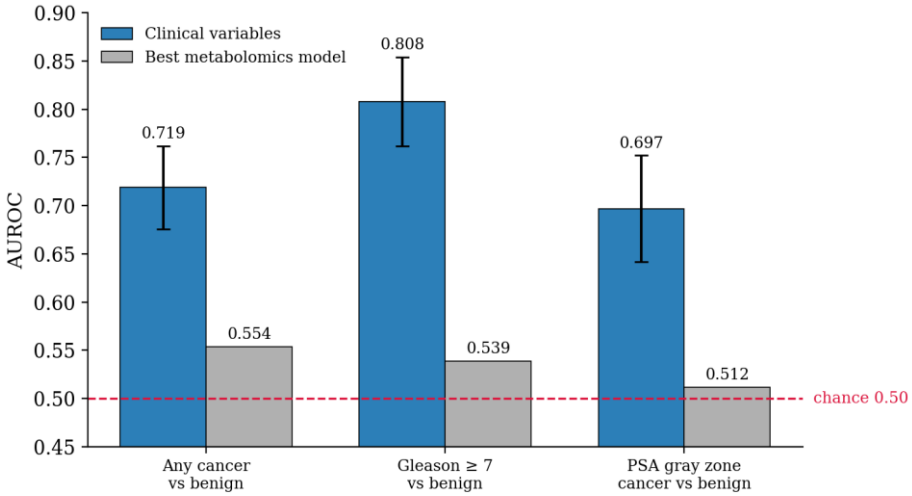

Figure 3. The predictive signal is clinical, not metabolomic, on every target. Metabolomics stays near the chance line while the clinical model is well above it, most clearly for clinically significant cancer.

3.4 Metabolites do not help inside the gray zone

The gray zone is the population named in the hypothesis, so it deserves a direct test. Among the 342 gray zone men, 156 had cancer and 186 were benign, which gives a majority baseline of 54%. Metabolomic models trained and tested within the gray zone stayed at chance, with the best of logistic regression, XGBoost, and LightGBM reaching only 0.563 AUROC, while the clinical model reached 0.702, and combining the two was worse than the clinical model alone. This mirrors the whole cohort result and shows that restricting to the harder gray zone does not reveal a hidden metabolomic signal. The graph ablation was also repeated inside the gray zone, and the outcome is reported in Table 4.

Table 4 reports the same graph ablation restricted to the gray zone, where each model was trained on the full cohort and evaluated on the gray zone cases of every test fold. The pattern is identical to the whole cohort. Every model sits at chance, the no graph control at 0.549 is not beaten by any graph variant, and the KEGG graph is again the worst at 0.529. XGBoost on metabolites is the highest at 0.558 but that is still chance against the 54% majority. So even in the exact population the hypothesis targeted, the graph does not help and the metabolites do not separate cancer from benign.

Model (gray zone, n = 342)	AUROC (mean ± SD)	AUPRC	Accuracy
XGBoost (metabolites + age/PSA)	0.558 ± 0.062	0.535	0.549
GAT, no graph (control)	0.549 ± 0.081	0.528	0.549
GAT, correlation edges	0.543 ± 0.086	0.521	0.549
GAT, sub pathway edges	0.540 ± 0.086	0.521	0.556
GAT, KEGG + correlation	0.530 ± 0.080	0.513	0.550
GAT, KEGG pathway edges	0.529 ± 0.085	0.514	0.552
Logistic regression (metabolites)	0.528 ± 0.051	0.501	0.526

Table 4. Graph ablation inside the PSA gray zone (156 cancer, 186 benign, majority baseline 0.544), 15 folds. Models were trained on the full cohort and scored on the gray zone cases of each held out fold. Every metabolomic model is at chance, the no graph control is not beaten by any graph, and the KEGG graph is the worst, which mirrors the whole cohort result in Table 2.

3.5 Feature selection leakage explains the gap with the published benchmark

The single most important experiment for reconciling my results with the published 0.89 is the leakage test, and its result is clear. When I select metabolites using all patients before cross validation, the AUROC jumps well above chance even though no real signal was added, and when I select inside each training fold the same pipeline returns to chance. Table 5 shows the effect. Selecting features outside the cross validation loop inflated the AUROC by 0.14 to 0.20 in my own data, and in the gray zone with k equal to 50 it reached 0.707 from what is otherwise pure noise. This does not fully reach 0.89, so I do not claim that leakage is the only factor in the published number, but it is a large and sufficient mechanism, and it shows how a pipeline that selects features before validation can manufacture strong looking results from data that carries no signal.

Cohort	Feature selection	AUROC, k = 20	AUROC, k = 50
All cancer (n=580)	Outside CV (leaky)	0.655	0.646
All cancer (n=580)	Inside CV (honest)	0.512	0.499
PSA gray zone (n=342)	Outside CV (leaky)	0.674	0.707
PSA gray zone (n=342)	Inside CV (honest)	0.499	0.504

Table 5. Feature selection leakage experiment. The same logistic regression and the same folds are used throughout, and only the position of the univariate ANOVA F selection changes. Selecting outside the cross validation loop inflates the AUROC by 0.14 to 0.20 with no real signal added, which is the selection bias of Ambroise and McLachlan (2002).

3.6 Candidate metabolites and metabolite structure

Even though the predictive result is negative, I report the individual metabolite associations inside the gray zone as leads for future work, and I am careful to frame them as exploratory. No metabolite survived the Benjamini Hochberg correction at a false discovery rate of 0.10, so none of these can be called a confirmed marker, and this is itself an honest finding rather than a failed search. The strongest single metabolites by univariate AUROC included 2-hydroxysebacate, 5-hydroxymethyl-2-furoic acid, and N6,N6-dimethyllysine, which are not on the usual prostate cancer marker lists, together with sphingosine 1-phosphate and sphinganine 1-phosphate, which have appeared in earlier work. The XGBoost SHAP ranking highlighted a piperine glucuronide, mannose, a lyso phospholipid, and retinol. Figure 4 shows the strongest candidates by single metabolite AUROC. The overall message is that the associations are weak, unconfirmed, and not usable as biomarkers in this cohort, which is consistent with the negative predictive result.

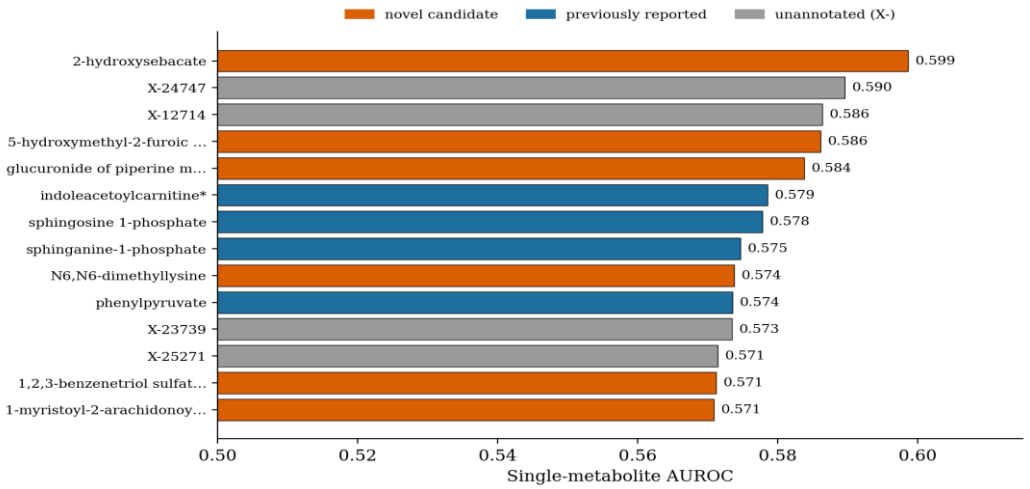

Figure 4. Exploratory candidate metabolites inside the gray zone, ranked by single metabolite AUROC. Orange marks candidates not on prior prostate cancer marker lists, blue marks previously reported markers, and grey marks unannotated (X-) metabolites. None survive false discovery rate correction, so all are hypothesis generating only.

3.7 Clinical operating points and biopsies avoided

Because the number of biopsies avoided is what matters clinically, I report operating points for the clinical model inside the gray zone. At a threshold that keeps sensitivity at 0.83, the model would avoid about 32% of biopsies while missing about 17% of cancers, and at a more cautious threshold with sensitivity 0.87 it would avoid about 23% of biopsies while missing about 13% of cancers. These numbers are modest and would need external validation before any clinical use, but they show that even a simple clinical model carries usable information in exactly the population where PSA alone does not, which is more than the metabolites achieved.

4. DISCUSSION

This study set out to test a specific and hopeful hypothesis, that a graph attention network encoding KEGG pathway relationships among plasma metabolites would predict prostate biopsy outcome better than flat models and better than routine clinical data, especially in the PSA gray zone. The hypothesis was not supported. The graph model sat at about 0.55 AUROC, it did not beat XGBoost, it did not beat a logistic regression given only age and PSA, and removing the graph made it slightly better rather than worse. The metabolites added nothing over age and PSA, and inside the gray zone they stayed at chance while the clinical model reached 0.70.

The most important thing to reconcile is the gap with Sun and colleagues, who reported an AUC of 0.89 on this data type (Sun L. et al., 2024). My leakage experiment offers a concrete and sufficient explanation. Selecting features across all patients before cross validation inflated my own AUROC by 0.14 to 0.20 with no real signal added, and reached 0.707 in the gray zone from pure noise. This is the classic selection bias documented by Ambroise and McLachlan (2002). I cannot audit the published pipeline directly, so I do not claim leakage is the whole story, but it is a large mechanism that turns chance level data into strong looking numbers, and it is the most likely reason careful baselines and a published deep model can differ by so much on the same data. My negative finding also sits comfortably next to Penney and colleagues, who could not build a reliable metabolite signature for the Gleason score in tissue and serum (Penney et al., 2021), and next to the contested history of sarcosine. The pattern across the field is that individual metabolite associations are real but weak, and that turning them into a reliable classifier is much harder than early reports suggested.

Why did the graph hurt rather than help? I think there are three reasons that reinforce each other. First, the setting is a small n and large p regime, with 580 patients and more than one thousand features, where an over parameterized message passing model has every opportunity to fit noise. Second, the KEGG coverage is low, since only about 279 of 1,169 metabolites map to any pathway, so most nodes carry only a self loop while a few central metabolism nodes become dense hubs, and this uneven structure gives the attention little useful to attend to. Third, message passing averages a node with its neighbors, and when the neighbors are noisy and weakly informative this smoothing removes the small separation that exists rather than sharpening it. The pathway prior that helped on gene expression in Lee and colleagues and in P-NET (Lee et al., 2020; Elmarakeby et al., 2021) does not transfer to this plasma metabolomics setting, and my results specifically question that transfer.

This work has clear limitations. The provider preprocessed file was normalized and imputed across all patients at once, so a residual leak remains that I could not remove from the released data, although it would make the metabolomic estimates optimistic and therefore does not rescue them. The KEGG coverage is low, which weakens the central hypothesis test even on its own terms. The sample size is small for a deep model, and I did not run a hyperparameter search, so I cannot rule out that some untried configuration performs better, although the convergence of five different model families on the same ceiling argues against it. Percent free PSA was missing for about two thirds of the men, which limits the clinical arm. Finally this is a single cohort from one center, so the clinical numbers need external validation before any use.

For future work, the productive direction is not a bigger metabolomic model but better information. Imaging based variables such as multiparametric MRI and PI-RADS are the largest lever in modern prostate risk and are absent from this dataset, and adding them to the cheap clinical model is the natural next step. The unidentified metabolites could be chemically annotated so that any signal they carry becomes interpretable, although my test showed they carry little on their own. Most importantly, the leakage experiment argues that the field needs to

report feature selection inside cross validation as a default, because otherwise the literature will keep producing numbers that do not hold up.

In conclusion, this is a rigorous negative result with a diagnosed cause. Plasma metabolomics, whether arranged into a KEGG graph, passed through attention, or handed to strong flat models, does not predict prostate biopsy outcome better than age and PSA in this cohort, inside or outside the gray zone. The famous 0.89 is best explained by feature selection leakage. The usable signal for this decision is in routine clinical data that is already collected, and that is where a practical model should be built.

5. REFERENCES

[1] Amante, E., Cerrato, A., Alladio, E., Capriotti, A. L., Cavaliere, C., Marini, F., et al. (2022). Comprehensive biomarker profiles and chemometric filtering of urinary metabolomics for effective discrimination of prostate carcinoma from benign hyperplasia. Scientific Reports, 12(1), 4361. https://doi.org/10.1038/s41598-022-08435

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

[3] American Cancer Society. (2026). Cancer Facts and Figures 2026. Atlanta: American Cancer Society.

[4] Benedetti, E., Chetnik, K., Flynn, T., Barbieri, C. E., Scherr, D. S., Loda, M., and Krumsiek, J. (2023). Plasma metabolomics profiling of 580 patients from an Early Detection Research Network prostate cancer cohort. Scientific Data, 10, 830. https://doi.org/10.1038/s41597-023-02750-7

[5] Chen, T., and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

[6] Dieterle, F., Ross, A., Schlotterbeck, G., and Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Analytical Chemistry, 78(13), 4281–4290. https://doi.org/10.1021/ac051632c

[7] Elmarakeby, H. A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., et al. (2021). Biologically informed deep neural network for prostate cancer discovery. Nature, 598, 348–352. https://doi.org/10.1038/s41586-021-03922-4

[8] Huang, J., Mondul, A. M., Weinstein, S. J., Karoly, E. D., Sampson, J. N., and Albanes, D. (2022). Metabolomic profile of prostate cancer-specific survival among 1812 Finnish men. BMC Medicine, 20, 362. https://doi.org/10.1186/s12916-022-02561-4

[9] Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, 28(1), 27–30. https://doi.org/10.1093/nar/28.1.27

[10] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. Y. (2017). LightGBM: a highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

[11] Kipf, T. N., and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations. https://arxiv.org/abs/1609.02907

[12] Lee, S., Lim, S., Lee, T., Sung, I., and Kim, S. (2020). Cancer subtype classification and modeling by pathway attention and propagation. Bioinformatics, 36(12), 3818–3824. https://doi.org/10.1093/bioinformatics/btaa203

[13] Lima, A. R., Pinto, J., Amaro, F., Bastos, M. L., Carvalho, M., and Guedes de Pinho, P. (2021). Advances and perspectives in prostate cancer biomarker discovery in the last 5 years through tissue and urine metabolomics. Metabolites, 11(3), 181. https://doi.org/10.3390/metabo11030181

[14] Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1705.07874

[15] Nadeau, C., and Bengio, Y. (2003). Inference for the generalization error. Machine Learning, 52(3), 239–281. https://doi.org/10.1023/A:1024068626366

[16] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[17] Penney, K. L., Tyekucheva, S., Rosenthal, J., El Fandy, H., Carelli, R., Borgstein, S., et al. (2021). Metabolomics of prostate cancer Gleason score in tumor tissue and serum. Molecular Cancer Research, 19(3), 475–484. https://doi.org/10.1158/1541-7786.MCR-20-0548

[18] Schmidt, J. A., Fensom, G. K., Rinaldi, S., Scalbert, A., Appleby, P. N., Achaintre, D., et al. (2020). Patterns in metabolite profile are associated with risk of more aggressive prostate cancer. International Journal of Cancer, 146(3), 720–730. https://doi.org/10.1002/ijc.32314

[19] Sreekumar, A., Poisson, L. M., Rajendiran, T. M., Khan, A. P., Cao, Q., Yu, J., et al. (2009). Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature, 457(7231), 910–914. https://doi.org/10.1038/nature07762

[20] Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., et al. (2016). Metabolomics Workbench: an international repository for metabolomics data and metadata. Nucleic Acids Research, 44(D1), D463–D470. https://doi.org/10.1093/nar/gkv1042

[21] Sun, L., Fan, X., Zhao, Y., Zhang, Q., and Jiang, M. (2024). Deep learning based metabolomics data study of prostate cancer. BMC Bioinformatics, 25, 391. https://doi.org/10.1186/s12859-024-06016-w

[22] Sun, W. (2025). The research progress on diagnostic indicators related to prostate specific antigen gray zone prostate cancer. BMC Cancer, 25, 1264. https://doi.org/10.1186/s12885-025-14505-1

[23] Tang, C., Guo, X., Zou, Q., Wang, J., Tang, Y., Tang, H., et al. (2025). Untargeted metabolomics revealed urinary metabolic pattern for discriminating prostate cancer from benign prostatic hyperplasia in Chinese participants. Frontiers in Oncology, 15, 1604169. https://doi.org/10.3389/fonc.2025.1604169

[24] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks. International Conference on Learning Representations. https://arxiv.org/abs/1710.10903

[25] Yang, B., Zhang, C., Cheng, S., Li, G., Griebel, J., and Neuhaus, J. (2021). Novel metabolic signatures of prostate cancer revealed by 1H-NMR metabolomics of urine. Diagnostics, 11(2), 149. https://doi.org/10.3390/diagnostics11020149