



Article

---

# SEFA: Semantic Embedding-Based Feature Augmentation of Biomedical Language-Model Embeddings Improves Interpretable Metabolomic Prediction of Lung Cancer

---

Jiawen Wu, Jean-François Haince, Rashid A. Bux, Guoyu Huang, Paramjit S. Tappia, Bram Ramjiawan and Maria Vaida

## Special Issue

Bioinformatic Insights into Cancer Biology: Transforming Data into Discovery

Edited by  
Dr. Bofei Wang





## Article

# SEFA: Semantic Embedding-Based Feature Augmentation of Biomedical Language-Model Embeddings Improves Interpretable Metabolomic Prediction of Lung Cancer

Jiawen Wu <sup>1</sup>, Jean-François Haince <sup>2</sup> , Rashid A. Bux <sup>3</sup>, Guoyu Huang <sup>2</sup> , Paramjit S. Tappia <sup>4</sup> ,  
Bram Ramjiawan <sup>4,5</sup> and Maria Vaida <sup>1,\*</sup>

- <sup>1</sup> Department of Data Science, Harrisburg University of Science and Technology, Harrisburg, PA 17101, USA; jwu3@my.harrisburgu.edu
- <sup>2</sup> BioMark Diagnostic Solutions Inc., Quebec, QC G1P 4P5, Canada; jhaince@biomarkdiagnostics.com (J.-F.H.); ghuang@biomarkdiagnostics.com (G.H.)
- <sup>3</sup> BioMark Diagnostics Inc., Richmond, BC V6X 2W2, Canada; rahmed@biomarkdiagnostics.com
- <sup>4</sup> Asper Clinical Research Institute and Albrechtsen Research Centre, St. Boniface Hospital, Winnipeg, MB R2H 2A6, Canada; ptappia@sbr.ca (P.S.T.); bramjiawan@sbr.ca (B.R.)
- <sup>5</sup> Department of Pharmacology & Therapeutics, Max Rady College of Medicine, University of Manitoba, Winnipeg, MB R3E 0T6, Canada
- \* Correspondence: mvaida@harrisburgu.edu

## Abstract

**Background/Objectives:** Feature engineering remains a major challenge in metabolomics-based prediction, particularly when rich biochemical knowledge is available but under-utilized. Conventional metabolomics models rely primarily on measured variables and statistically driven feature selection, overlooking the molecular and pathway context encoded in curated metabolite knowledge bases. We propose SEFA (Semantic Embedding-based Feature Augmentation), a model-agnostic framework that integrates metabolite-level textual knowledge from the Human Metabolome Database (HMDB) into structured metabolomics modeling for lung-cancer prediction. **Methods:** SEFA encodes HMDB metabolite descriptions as 768-dimensional MedBERT vectors and projects measured metabolite concentrations into this semantic space via concentration-weighted aggregation, producing a 928-dimensional candidate feature matrix that concatenates 11 clinical variables, 149 metabolite concentrations, and 768 semantic projection features. Sparse L1-guided feature selection reduced this representation to 24 features (2.6% of candidates) within a leakage-free cross-validation pipeline. Six classifiers were evaluated on a lung-cancer plasma metabolomics cohort of 800 participants (586 cases, 214 controls) with a stratified 80:20 split, and a controlled ablation study compared the augmented representation with a 24-metabolite-only baseline. Pathway-enrichment analysis of the metabolite sets associated with the retained embedding dimensions was performed using MetaboAnalyst 5.0. **Results:** Logistic regression on the 24-feature SEFA representation achieved a test ROC-AUC of 0.969, a precision-recall AUC of 0.987, and an accuracy of 94.4%, competitive with less interpretable approaches. Under the same 24-feature budget, embedding augmentation improved test ROC-AUC by 0.008, precision-recall AUC by 0.004, and accuracy by 3.1 percentage points over the metabolite-only baseline. Five retained embedding dimensions mapped onto coherent metabolic themes—sphingolipids and acylcarnitines, carnitine and glutamine metabolism, one-carbon and nitrogen handling, purine catabolism, and oxidative stress markers—and their associated metabolite sets were enriched for arginine and proline metabolism and glycine, serine, and threonine metabolism, pathways with established roles in lung-cancer biology. **Conclusions:** SEFA demonstrates that semantic embeddings derived from biomedical language models can convert curated metabolite



Academic Editors: Jun Lu and Bofei Wang

Received: 20 June 2026

Revised: 29 July 2026

Accepted: 13 August 2026

Published: 18 August 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

annotations into patient-level features that supply complementary predictive signal while preserving biological interpretability through pathway-level analysis. The present evidence is limited to a single region-specific cohort; external, cross-platform, and cross-disease evaluation is required before broader generalization or clinical application.

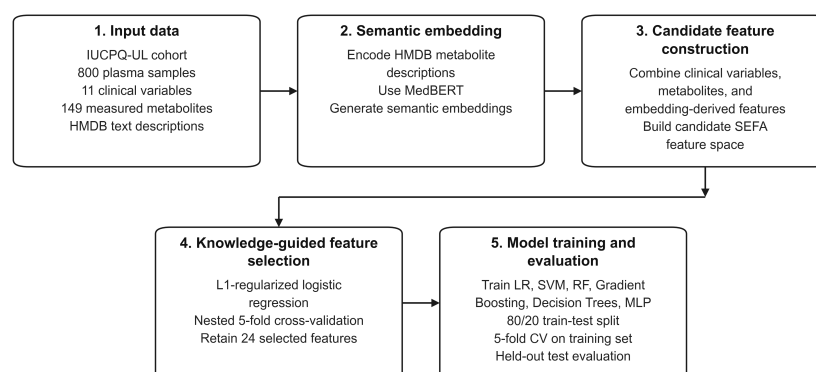
**Keywords:** lung cancer; metabolomics; feature augmentation; biomedical language models; semantic embeddings; pathway enrichment; interpretable machine learning; HMDB

## 1. Introduction

Lung cancer remains the leading cause of cancer mortality worldwide, and outcomes depend strongly on the stage at diagnosis, motivating the search for accurate, minimally invasive biomarkers [1]. Metabolomics, which measures small-molecule end-products of cellular activity, offers a sensitive readout of the metabolic reprogramming that accompanies malignancy and has shown promise for non-invasive lung-cancer detection from blood and other biofluids [2–9].

Most metabolomic classifiers, however, operate purely on the measured concentrations of a metabolite panel and treat each feature as an anonymous number. This discards the substantial prior knowledge that exists for each metabolite in curated resources such as the Human Metabolome Database (HMDB) [10,11], where free-text descriptions summarize biochemical roles, pathway membership, tissue localization, and disease associations. Recent biomedical language models trained on large corpora of clinical and biochemical text encode this knowledge in dense vector representations, and such embeddings have proven transferable across downstream tasks [12–25].

Here we ask whether transferring this textual, knowledge-derived signal into a metabolomic model improves prediction without sacrificing interpretability. We introduce *SEFA*, which concatenates measured metabolite features with semantic projections constructed from HMDB descriptions and then performs sparse selection within the training data. SEFA provides a model-agnostic bridge from metabolite-level text to sample-level numerical features. This differs from studies that use embeddings only for text classification or static similarity analysis. We benchmark SEFA on a lung-cancer metabolomics cohort, quantify the gain attributable to semantic augmentation through ablation, and interpret the selected embedding dimensions through metabolite loadings and pathway-enrichment analysis. An overview of SEFA workflow is shown in Figure 1.



**Figure 1.** Overview of the SEFA pipeline. Clinical variables, measured metabolite concentrations, and patient-level semantic projections derived from HMDB descriptions represent input features. Sparse feature selection and classifier tuning are performed using nested cross-validation, followed by a single evaluation on the held-out test set.

## 2. Materials and Methods

### 2.1. Dataset and Cohort Characteristics

The study used 800 plasma samples from the IUCPQ Biobank–Respiratory Health Research Network, comprising 586 lung-cancer cases and 214 controls. Samples and associated clinical data were obtained under the University of Manitoba Health Research Ethics Board approval (Ethics File H2012:334, approved 12 December 2022), and all participants provided informed consent. The cohort is therefore region-specific and should not be interpreted as representative of all geographic or ancestry groups. A stratified 80:20 split produced 640 development samples and 160 held-out test samples. The test set contained 118 cases and 42 controls.

Each sample included 11 clinical variables (sex, race, smoking status, past smoking, current smoking, age, height, weight, body-mass index, and two pack-year exposure measures) and metabolite measurements. The cohort was approximately balanced by sex (male:  $n = 393$ , 49.1% female:  $n = 407$ , 50.9%), and the distribution by disease status is given in Table 1. Missing values were replaced with the assay limits of detection where available. Columns with more than 30% missingness were excluded. Feature scaling was fitted on the training set only; the held-out test set was transformed using the training-set statistics and contributed to no fitting step. The participant-level dataset cannot be publicly released because of consent and privacy restrictions.

**Table 1.** Demographic and clinical characteristics of the study cohort, stratified by disease status.

| Characteristic                   | Overall<br>( $n = 800$ ) | Control<br>( $n = 214$ ) | Cancer<br>( $n = 586$ ) | $p$    | Q (BH-FDR) | SMD    |
|----------------------------------|--------------------------|--------------------------|-------------------------|--------|------------|--------|
| Sex, $n$ (%)                     | 800                      |                          |                         | 0.097  | 0.117      | 0.139  |
| Female                           | 407 (50.9)               | 98 (45.8)                | 309 (52.7)              |        |            |        |
| Male                             | 393 (49.1)               | 116 (54.2)               | 277 (47.3)              |        |            |        |
| Race/ethnicity, $n$ (%)          | 800                      |                          |                         | <0.001 | <0.001     | 0.455  |
| Other                            | 78 (9.8)                 | 3 (1.4)                  | 75 (12.8)               |        |            |        |
| White                            | 722 (90.2)               | 211 (98.6)               | 511 (87.2)              |        |            |        |
| Smoking status, $n$ (%)          | 800                      |                          |                         | <0.001 | <0.001     | 0.974  |
| Never smoker                     | 140 (17.5)               | 98 (45.8)                | 42 (7.2)                |        |            |        |
| Ever smoker                      | 660 (82.5)               | 116 (54.2)               | 544 (92.8)              |        |            |        |
| Former smoker, $n$ (%)           | 800                      |                          |                         | <0.001 | <0.001     | 0.599  |
| No                               | 304 (38.0)               | 126 (58.9)               | 178 (30.4)              |        |            |        |
| Yes                              | 496 (62.0)               | 88 (41.1)                | 408 (69.6)              |        |            |        |
| Current smoker, $n$ (%)          | 800                      |                          |                         | 0.002  | 0.003      | 0.265  |
| No                               | 636 (79.5)               | 186 (86.9)               | 450 (76.8)              |        |            |        |
| Yes                              | 164 (20.5)               | 28 (13.1)                | 136 (23.2)              |        |            |        |
| Smoking history, $n$ (%)         | 800                      |                          |                         | <0.001 | <0.001     | 0.599  |
| No                               | 304 (38.0)               | 126 (58.9)               | 178 (30.4)              |        |            |        |
| Yes                              | 496 (62.0)               | 88 (41.1)                | 408 (69.6)              |        |            |        |
| Age, years                       | 64.00<br>[58.00–70.00]   | 60.10<br>[53.00–68.75]   | 65.00<br>[60.00–70.00]  | <0.001 | <0.001     | 0.381  |
| Height, m                        | 1.65<br>[1.58–1.72]      | 1.68<br>[1.60–1.73]      | 1.64<br>[1.57–1.70]     | <0.001 | <0.001     | −0.322 |
| Weight, kg                       | 73.00<br>[62.00–84.00]   | 74.00<br>[63.50–86.00]   | 72.00<br>[61.00–84.00]  | 0.198  | 0.198      | −0.085 |
| Body mass index                  | 26.67<br>[23.53–30.09]   | 26.03<br>[23.67–29.17]   | 26.84<br>[23.52–30.37]  | 0.160  | 0.174      | 0.127  |
| Smoking exposure, pack-years     | 30.30<br>[10.00–48.00]   | 3.50<br>[0.00–25.00]     | 40.00<br>[21.38–50.00]  | <0.001 | <0.001     | 1.032  |
| Smoking exposure, cigarettes/day | 20.00<br>[7.00–20.00]    | 0.00<br>[0.00–15.00]     | 20.00<br>[10.00–20.00]  | <0.001 | <0.001     | 0.879  |

Note: Continuous variables are summarized as median [interquartile range] or mean  $\pm$  SD according to the Shapiro–Wilk test for normality; categorical variables as  $n$  (%), with percentages computed within column.  $p$  values are two-sided, from the test named in each row;  $Q$  values are Benjamini–Hochberg false-discovery-rate-adjusted across all variables in the table. SMD, standardized mean difference;  $|SMD| < 0.10$  indicates negligible imbalance.

## 2.2. Semantic Embeddings of Metabolite Descriptions

For each metabolite with an available annotation, the curated HMDB textual description was matched to the corresponding assay variable and encoded as a 768-dimensional MedBERT vector. For each participant, measured concentrations were first transformed as  $\text{sign}(x) \log(1 + |x|)$ . Each participant's transformed profile was shifted to be non-negative by subtracting its row minimum and normalized to sum to one, giving a weight vector over metabolites. The 768-dimensional semantic representation was then formed as  $\mathbf{S} = \mathbf{WE}$ , where  $\mathbf{W}$  is the participant-by-metabolite weight matrix and  $\mathbf{E}$  the metabolite-by-embedding matrix. Because the weights are normalized, the representation is a convex combination of metabolite embeddings and therefore reflects a participant's metabolite composition rather than overall concentration magnitude. These semantic features, together with the clinical and log-transformed metabolite features, were standardized using statistics estimated from the train set. The encoder weights were fixed and were not trained on case/control labels.

## 2.3. Feature Construction and Selection

Clinical, metabolite, and semantic features were concatenated into the candidate feature matrix. Sparse selection was restricted to the test data and was placed inside the cross-validation pipeline so that no held-out fold contributed to selector fitting. The L1 regularization strength and the maximum retained feature count were treated as tuning parameters. Candidate feature budgets from 4 to 40 and regularization values from  $10^{-3}$  to  $10^1$  were considered. The compact model retained 24 of 928 candidate variables (2.6%). This size reflects the cross-validated sparsity/performance trade-off, not a biological claim that exactly 24 variables are universally optimal.

## 2.4. Classification, Tuning, and Evaluation

Six classifiers were evaluated, namely logistic regression, support vector machine, random forest, multilayer perceptron, gradient boosting, and decision tree. All preprocessing, feature selection, and hyperparameter tuning operations that depend on sample statistics were fit within training folds. The final model used an L2 penalty with the liblinear solver, regularization strength  $C = 1.0$ , a maximum of 400 iterations, and random seed 42. Feature selection used an L1-penalized logistic regression (liblinear solver, random seed 0) retaining at most 24 features. Five-fold nested cross-validation optimized ROC-AUC. The held-out test set was evaluated once after model selection. Performance was summarized using ROC-AUC, average precision (PR-AUC), accuracy, and confusion matrix.

The analysis was conducted in Python 3.11 using Pandas, NumPy, SciPy, and Scikit-Learn libraries for data handling and modeling and Matplotlib (v3.10.0) for visualization. The metabolite-text semantic features were generated with the pretrained MedBERT model loaded through the Hugging Face Transformers library. The specific embedding dimensionality is reported in the Materials and Methods. Textual metabolite descriptions are publicly available in HMDB database. Pathway-enrichment significance is reported as  $-\log_{10}(p)$  with false-discovery-rate correction, and an adjusted  $p$ -value  $p < 0.05$  was considered significant.

# 3. Results

## 3.1. Predictive Performance and Model Selection

Across the six classifiers, logistic regression achieved the highest cross-validated ROC-AUC (mean 0.9323,  $\sigma = 0.0213$ ), narrowly ahead of SVM (0.9320), with random forest, MLP and gradient boosting clustered near 0.92 and decision tree lowest at 0.8187

(Table 2). The same ranking held for PR-AUC and accuracy. Per-fold ROC-AUC and the cross-validation-versus-test comparison indicated that the top models generalized well.

The ablation study in Table 3 isolates the effect of adding embeddings under the same 24-feature count. The 24-metabolite-only logistic regression model achieved a cross-validation ROC-AUC of  $0.9219 \pm 0.0367$ , a test ROC-AUC of 0.9609, a test PR-AUC of 0.9832, and a test accuracy of 0.9125. After embedding-derived features were incorporated into the final 24-feature representation, performance gains were observed of 0.0104 in cross-validation ROC-AUC, 0.0082 in test ROC-AUC, 0.0039 in test PR-AUC, and 0.0312 in test accuracy. This gain is statistically and practically smaller than the absolute performance values and should be interpreted as evidence of complementary signal, not as proof of a clinically meaningful advantage. Notably, several selected features were embedding dimensions rather than measured metabolites (e.g., dimensions 368, 495, 514, 609 and 685), indicating that the language-model representation contributed signal that the measured panel did not fully capture. These results show that the embedding features recovered predictive information that would otherwise be lost under aggressive feature reduction, allowing the compact representation to perform modestly better than a metabolite-only model of the same size.

**Table 2.** Classifier comparison on the final 24-feature representation. Bold values indicate the best result in each metric. CV metrics are the mean over 5-fold cross-validation ( $\pm$ standard deviation for ROC-AUC); test metrics are computed once on the held-out set ( $n = 160$ ).

| Model               | CV ROC-AUC                            | CV F1         | Test ROC-AUC  | Test PR-AUC   | Acc.          |
|---------------------|---------------------------------------|---------------|---------------|---------------|---------------|
| Logistic Regression | <b>0.9323 <math>\pm</math> 0.0213</b> | 0.9242        | 0.9691        | <b>0.9871</b> | <b>0.9437</b> |
| SVM                 | 0.9320 $\pm$ 0.0174                   | 0.9202        | 0.9677        | 0.9866        | 0.9187        |
| Random Forest       | 0.9214 $\pm$ 0.0101                   | 0.9201        | 0.9546        | 0.9822        | 0.9000        |
| MLP                 | 0.9213 $\pm$ 0.0192                   | 0.9172        | <b>0.9699</b> | 0.9870        | 0.9375        |
| Gradient Boosting   | 0.9201 $\pm$ 0.0115                   | <b>0.9236</b> | 0.9481        | 0.9798        | 0.9125        |
| Decision Tree       | 0.8187 $\pm$ 0.0392                   | 0.8667        | 0.8676        | 0.9290        | 0.8375        |

**Table 3.** Ablation study for logistic regression under a fixed 24-feature budget. The proposed 24-feature representation with embeddings is compared with a 24-metabolite-only baseline.

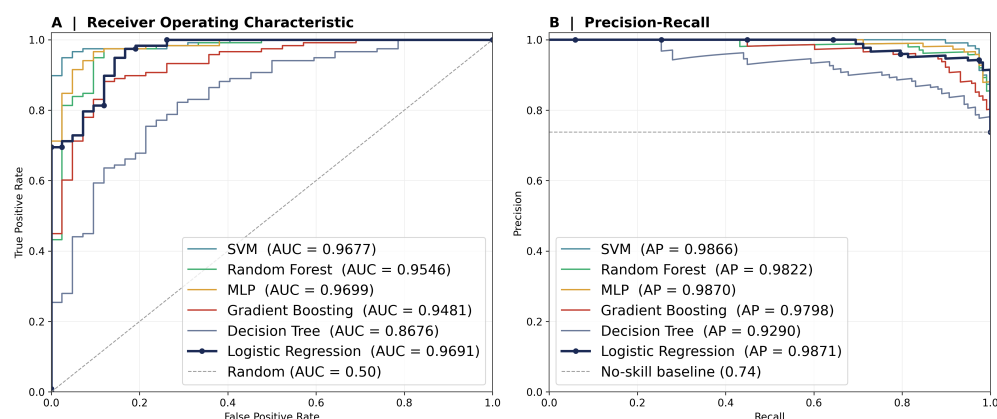
| Feature Set                      | CV ROC-AUC          | Test ROC-AUC | Test PR-AUC | Acc.    |
|----------------------------------|---------------------|--------------|-------------|---------|
| 24 metabolites only              | 0.9219 $\pm$ 0.0367 | 0.9609       | 0.9832      | 0.9125  |
| 24 features + embeddings         | 0.9323 $\pm$ 0.0213 | 0.9691       | 0.9871      | 0.9437  |
| $\Delta$ (with emb. – met. only) | +0.0104             | +0.0082      | +0.0039     | +0.0312 |

The regularized logistic-regression model selected 24 features and achieved a held-out test ROC-AUC of 0.9691 and a PR-AUC of 0.9871 (Figure 2). The confusion matrix contained 6 false-positive controls and 3 false-negative cancer cases and had strong precision, recall and F1 score on the positive class (Figure 3A,B).

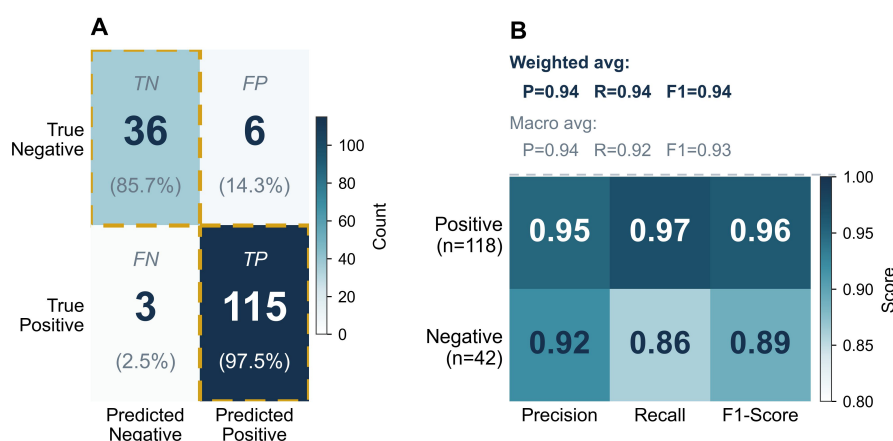
### 3.2. Interpretation of Selected Embedding Dimensions

To understand what the retained embedding dimensions encode, we ranked the metabolites loading most strongly on each (Supplementary Figure S1) and summarized their description vocabulary as word clouds (Supplementary Figure S2). These loading and word-cloud displays are exploratory attribution aids, without claiming any mechanistic or causal results. The dimensions captured lung cancer-associated metabolites. Dimension 368 was dominated by sphingolipids and acylcarnitines [26,27], dimension

495 by carnitine/glutamine-related terms [28], dimension 514 by GABA, sarcosine and creatine-related species [29], dimension 609 by trimethylamine and hypoxanthine [30], and dimension 685 by urate, spermine and methionine-related chemistry [31–33].



**Figure 2.** Test-set performance curves for six classifiers trained on the final 24-feature SEFA representation. **(A)** Receiver Operating Characteristic (ROC) curves. Logistic Regression (AUC = 0.9691) and MLP (AUC = 0.9699) achieve the highest area under the curve. **(B)** Precision-Recall (PR) curves. Logistic Regression attains the highest average precision (AP = 0.9871).

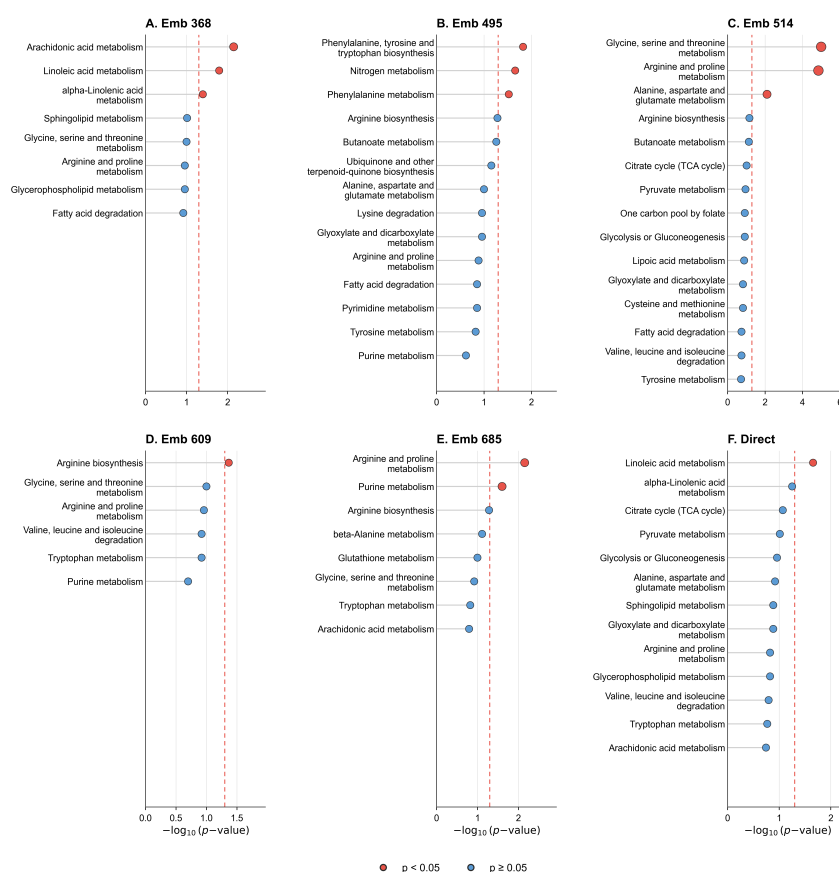


**Figure 3.** Evaluation of the final logistic regression model on the 24-feature representation. Panel **(A)** shows the confusion matrix on the held-out test set. Panel **(B)** shows the class-specific precision, recall, and F1-score together with macro and weighted averages.

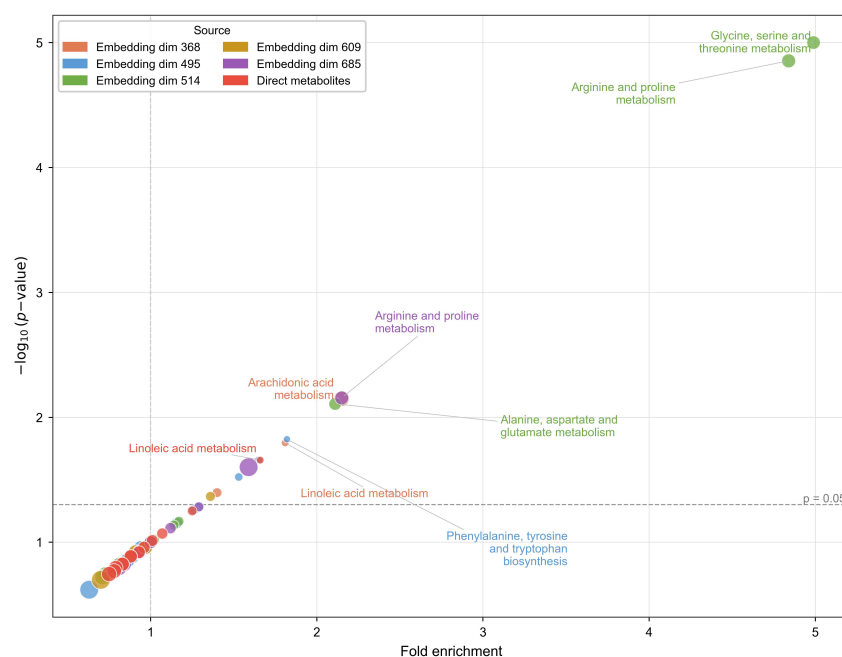
### 3.3. Pathway Enrichment of Embedding-Derived Metabolite Sets

Pathway-enrichment analysis of the metabolite sets associated with each embedding dimension, compared against direct metabolite selection, recovered lung-cancer-relevant biology (Figures 4 and 5). Arginine and proline metabolism was the most broadly shared pathway, appearing across all six sources, followed by arginine biosynthesis and glycine, serine and threonine metabolism (Figure 6B). Embedding dimension 514 produced the strongest enrichment signals—glycine, serine and threonine metabolism (~5-fold) and arginine and proline metabolism (~4.8-fold), exceeding those obtained from direct metabolites.

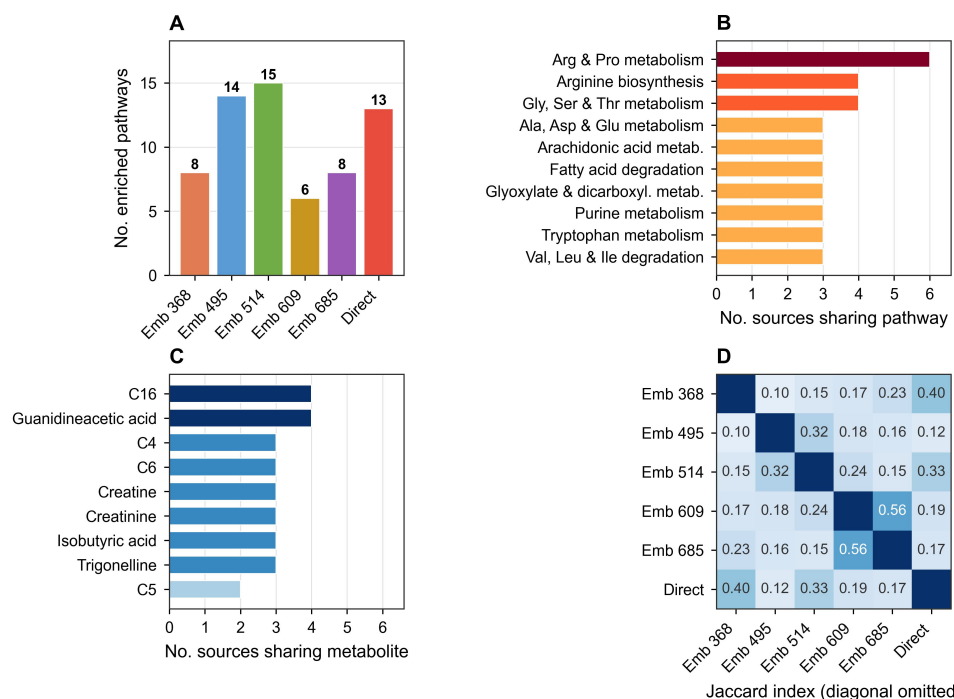
The number of significantly enriched pathways varied by source (15 for embedding 514, 14 for 495, 13 for direct metabolites, and 6–8 for the remaining dimensions. Figure 6A), and pairwise Jaccard similarity showed that different embedding dimensions captured partly non-overlapping pathway sets (Figure 6D), consistent with the embeddings supplying complementary information.



**Figure 4.** Metabolic pathway enrichment for each feature source. Panels (A–F) correspond to the five selected MedBERT embedding dimensions and the directly selected metabolites. Each row is an enriched KEGG pathway, positioned by  $-\log_{10}(p)$  from Fisher’s exact test against the 149-metabolite background. The red dashed line marks  $p = 0.05$  ( $-\log_{10} p = 1.30$ ). Points to its right are enriched at nominal significance.



**Figure 5.** Combined significance-versus-enrichment scatter across all sources. Each point is a pathway colored by its source. The dashed line marks  $p = 0.05$ . Glycine, serine and threonine metabolism and arginine and proline metabolism (embedding 514) emerge as the most significant, high-fold-enrichment pathways.



**Figure 6.** Cross-source comparison. (A) Number of enriched pathways per source. (B) Pathways ranked by the number of sources sharing them. (C) Metabolites ranked by sharing across sources. (D) Pairwise Jaccard similarity between sources, showing that embedding dimensions capture partly distinct pathway sets.

#### 4. Discussion

SEFA shows that semantic embeddings of biomedical text can be folded into a metabolomic classifier as additional features and that they yield an improvement in lung-cancer prediction (+0.008 ROC-AUC, +0.031 accuracy) while keeping the final model compact and interpretable. This gain should be interpreted as evidence of a complementary signal, not as proof of a clinically meaningful advantage. The best model—an L2-regularized logistic regression using only 24 of 928 candidate features—reached a test ROC-AUC of 0.969 and 94.4% accuracy, performance competitive with far less transparent approaches.

The selected latent dimensions mapped onto coherent metabolite groups (sphingolipids, acylcarnitines, carnitine/glutamine chemistry, GABA/creatine species), and their metabolite sets were enriched for pathways with established links to malignant metabolic reprogramming, most consistently arginine and proline metabolism and glycine, serine, and threonine metabolism. The partly non-overlapping pathway coverage across dimensions (low pairwise Jaccard similarity) supports the view that the embeddings contribute complementary information to the measured panel, which is consistent with the ablation result.

The pathways most consistently recovered have established links to lung-cancer biology. Arginine and proline metabolism is implicated in lung adenocarcinoma prognosis and immune escape, and proline catabolism has been shown to drive progression in non-small-cell lung cancer, consistent with the reduced arginine availability reported in cancer more broadly [34–36]. Glycine, serine and threonine metabolism feed one-carbon metabolism, a recognized hallmark of proliferating tumors [37], while purine metabolism has been linked to macrophage-mediated immunosuppression in non-small-cell lung cancer [38]. These convergent signals, together with the broader lung-cancer metabolomic disturbances reported previously [7,8,39,40], support the biological plausibility of the embedding-derived features.

This study is subject to several limitations. All participants were drawn from a single biobank network, and the cohort was imbalanced toward cancer cases. Differences in geography, ancestry, sex, clinical stage, and acquisition protocol may therefore constrain the transportability of the model. The held-out partition provides an estimate of internal generalization only and cannot substitute for temporal, cross-site, or prospective validation. The semantic features are contingent on the completeness, phrasing, and update cycle of the HMDB annotations, as well as on the choice of language model and the metabolite-name mapping procedure. In addition, both the selected feature-set size and the specific variables retained may vary across resampling iterations, indicating a degree of selection instability. The pathway-enrichment results and embedding loadings are associative and do not establish causal mechanisms. Finally, the observed augmentation gain is modest and may not warrant the additional implementation complexity in all settings.

Accordingly, SEFA should presently be viewed as a research framework rather than a deployable diagnostic system. Future work will evaluate it on independent cohorts, other cancer and non-cancer metabolomics datasets, different assay platforms, and prospectively collected samples. Such studies should include subgroup performance, calibration, decision-curve analysis, and prespecified clinical thresholds before any claim of clinical utility.

## 5. Conclusions

SEFA demonstrates that biomedical language-model embeddings of metabolite descriptions can be converted into patient-level features that supply biologically coherent and statistically complementary information to measured metabolite concentrations and clinical variables. Within this single-cohort study, the augmented representation improved lung-cancer prediction modestly but consistently, and the retained embedding dimensions captured known hallmarks of malignant metabolic reprogramming through pathway-enrichment analysis. SEFA should be regarded as a research framework whose clinical utility, transportability, and scalability require independent evaluation. The results nonetheless position semantic knowledge augmentation as a principled and reproducible strategy for enriching metabolomics-based predictive models with biochemical domain knowledge.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/biomedicines14081856/s1>, Figure S1: Top 10 metabolites per selected embedding dimension, ranked by cosine-similarity loading. Each panel lists the metabolites whose HMDB-description embeddings align most strongly with embedding dimensions 368, 495, 514, 609 and 685. These loadings are exploratory attribution aids that suggest the chemical themes associated with each latent dimension. Figure S2: Word clouds of HMDB description vocabulary associated with each selected embedding dimension (368, 495, 514, 609, 685). Term prominence reflects frequency within the descriptions of the top-loading metabolites.

**Author Contributions:** Conceptualization, J.W., R.A.B., J.-F.H. and M.V.; data curation, J.-F.H.; formal analysis and interpretation, J.W. and M.V.; funding acquisition, R.A.B., J.-F.H. and P.S.T.; methodology, J.W., J.-F.H., R.A.B. and P.S.T.; project administration, R.A.B.; project manager, G.H.; resources, R.A.B., G.H., J.-F.H. and P.S.T.; writing—original draft, J.W., M.V., R.A.B. and J.-F.H.; writing—review and editing, R.A.B., J.-F.H., P.S.T., B.R. and M.V. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** This study was conducted in accordance with the Declaration of Helsinki and approved by the University of Manitoba Health Research Ethics Board (Ethics File #: H2012:334; approval date 12 December 2022) prior to study implementation.

**Informed Consent Statement:** Informed consent was obtained from all subjects prior to sample donation to the IUCPQ Biobank–Respiratory Health Research Network, Canada.

**Data Availability Statement:** Participant-level data are not publicly available because of privacy, consent, and institutional restrictions. A de-identified data dictionary and access requests may be directed to the corresponding author and data custodian, subject to ethics and institutional approval. All code implementing the SEFA method is publicly available at <https://github.com/miliana/SEFA>, accessed on 28 July 2026. The repository provides a complete, self-contained reference implementation of semantic embedding feature augmentation, including construction of the abundance-weighted embedding, the L1-penalized feature selection, and the cross-validation and held-out evaluation procedure. It runs on synthetic data generated by the accompanying script, so the pipeline can be executed and inspected without access to participant-level data.

**Acknowledgments:** Infrastructure support was provided by the St. Boniface Hospital Foundation, the University of Manitoba, the Institut Universitaire de Cardiologie et de Pneumologie de Québec—Université Laval (IUCPQ), and the Cooperative Health Tissue Network (USA), which provided the plasma samples and patient data.

**Conflicts of Interest:** R.A.B. is the President and CEO of BioMark Diagnostics Inc. and is a shareholder. G.H. is the President of BioMark Diagnostic Solutions Inc. J.-F.H. is the Executive Director of BioMark Diagnostic Solutions Inc. P.S.T. is a minor shareholder of BioMark Diagnostics Inc. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

## Abbreviations

The following abbreviations are used in this manuscript:

|             |                                                           |
|-------------|-----------------------------------------------------------|
| SEFA        | Semantic Embedding-Based Feature Augmentation             |
| HMDB        | Human Metabolome Database                                 |
| ROC-AUC     | Area Under the Receiver Operating Characteristic Curve    |
| PR-AUC      | Area Under the Precision-Recall Curve (Average Precision) |
| CV          | Cross-Validation                                          |
| FDR         | False Discovery Rate                                      |
| MLP         | Multilayer Perceptron                                     |
| SVM         | Support Vector Machine                                    |
| TP/TN/FP/FN | True/False Positives and Negatives                        |

## References

1. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2018. *CA Cancer J. Clin.* **2018**, *68*, 7–30. [[CrossRef](#)] [[PubMed](#)]
2. Yang, R.; Tsigelny, I.F.; Kesari, S.; Kouznetsova, V.L. Colorectal cancer detection via metabolites and machine learning. *Curr. Issues Mol. Biol.* **2024**, *46*, 4133–4146. [[CrossRef](#)] [[PubMed](#)]
3. Vaida, M.; Arumalla, K.K.; Tatikonda, P.K.; Popuri, B.; Bux, R.A.; Tappia, P.S.; Huang, G.; Haince, J.-F.; Ford, W.R. Identification of a novel biomarker panel for breast cancer screening. *Int. J. Mol. Sci.* **2024**, *25*, 11835. [[CrossRef](#)] [[PubMed](#)]
4. Sun, L.; Fan, X.; Zhao, Y.; Zhang, Q.; Jiang, M. Deep learning-based metabolomics data study of prostate cancer. *BMC Bioinform.* **2024**, *25*, 391. [[CrossRef](#)] [[PubMed](#)]
5. Ogunjobi, T.T.; Ohaeri, P.N.; Akintola, O.T.; Atanda, D.O.; Orji, F.P.; Adebayo, J.O.; Abdul, S.O.; Eji, C.A.; Asebebe, A.B.; Shodipe, O.O.; et al. Bioinformatics applications in chronic diseases: A comprehensive review of genomic, transcriptomics, proteomic, metabolomics, and machine learning approaches. *Medinformatics* **2024**, *3*, 110–127. Available online: <https://ojs.bonviewpress.com/index.php/MEDIN/article/view/2335> (accessed on 1 April 2026). [[CrossRef](#)]
6. Zhao, D.; Grist, J.T.; Rose, H.E.L.; Davies, N.P.; Wilson, M.; MacPherson, L.; Abernethy, L.J.; Avula, S.; Pizer, B.; Gutierrez, D.R.; et al. Metabolite selection for machine learning in childhood brain tumour classification. *NMR Biomed.* **2022**, *35*, e4673. [[CrossRef](#)] [[PubMed](#)]
7. Tang, Z.; Liang, D.; Deubler, E.L.; Sarnat, J.A.; Chow, S.S.; Diver, W.R.; Wang, Y. Lung cancer metabolomics: A pooled analysis in the Cancer Prevention Studies. *BMC Med.* **2024**, *22*, 262. [[CrossRef](#)] [[PubMed](#)]

8. Qi, S.A.; Wu, Q.; Chen, Z.; Zhang, W.; Zhou, Y.; Mao, K.; Li, J.; Li, Y.; Chen, J.; Huang, Y.; et al. High-resolution metabolomic biomarkers for lung cancer diagnosis and prognosis. *Sci. Rep.* **2021**, *11*, 11805. [[CrossRef](#)] [[PubMed](#)]
9. Galal, A.; Talal, M.; Moustafa, A. Applications of machine learning in metabolomics: Disease modeling and classification. *Front. Genet.* **2022**, *13*, 1017340. [[CrossRef](#)] [[PubMed](#)]
10. Wishart, D.S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A.C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; et al. HMDB: The Human Metabolome Database. *Nucleic Acids Res.* **2007**, *35*, D521–D526. [[CrossRef](#)] [[PubMed](#)]
11. Wishart, D.S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B.L.; et al. HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids Res.* **2022**, *50*, D622–D631. [[CrossRef](#)] [[PubMed](#)]
12. Yuan, Z.; Liu, Y.; Tan, C.; Huang, S.; Huang, F. Improving biomedical pretrained language models with knowledge. In *Proceedings of the 20th Workshop on Biomedical Language Processing, Online, 11 June 2021*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 180–190.
13. Tan, J.; Hu, J.; Dong, S. Incorporating entity-level knowledge in pretrained language model for biomedical dense retrieval. *Comput. Biol. Med.* **2023**, *166*, 107535. [[CrossRef](#)] [[PubMed](#)]
14. Biryukov, M.; Groues, V.; Satagopam, V.; Schneider, R. BioKB—Text mining and semantic technologies for biomedical content discovery. In *Proceedings of the 10th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, Rome, Italy, 4–7 December 2017*. [[CrossRef](#)]
15. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [[CrossRef](#)] [[PubMed](#)]
16. Chang, D.; Balažević, I.; Allen, C.; Chawla, D.; Brandt, C.; Taylor, A. Benchmark and best practices for biomedical knowledge graph embeddings. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, Online, 9 July 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 167–176.
17. Jha, K. Knowledge-base enriched word embeddings for biomedical domain. *arXiv* **2021**, arXiv:2103.00479.
18. Alachram, H.; Chereda, H.; Beißbarth, T.; Wingender, E.; Stegmaier, P. Text mining-based word representations for biomedical data analysis and protein–protein interaction networks in machine learning tasks. *PLoS ONE* **2021**, *16*, e0258623. [[CrossRef](#)] [[PubMed](#)]
19. Vithanage, D.; Yu, P.; Wang, L.; Deng, C. Contextual word embedding for biomedical knowledge extraction: A rapid review and case study. *J. Healthc. Inform. Res.* **2024**, *8*, 158–179. [[CrossRef](#)] [[PubMed](#)]
20. Vasantharajan, C.; Tun, K.Z.; Thi-Nga, H.; Jain, S.; Rong, T.; Siong, C.E. MedBERT: A pre-trained language model for biomedical named entity recognition. In *Proceedings of the 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, 7–10 November 2022*; pp. 1482–1488. [[CrossRef](#)]
21. Michalopoulos, G.; Wang, Y.; Kaka, H.; Chen, H.; Wong, A. UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 1744–1753.
22. Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; Liu, T.-Y. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.* **2022**, *23*, bbac409. [[CrossRef](#)] [[PubMed](#)]
23. Taylor, R.; Kardas, M.; Cucurull, G.; Scialom, T.; Hartshorn, A.; Saravia, E.; Poulton, A.; Kerkez, V.; Stojnic, R. Galactica: A large language model for science. *arXiv* **2022**, arXiv:2211.09085.
24. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S.R.; Cole-Lewis, H.; et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **2025**, *31*, 943–950. [[CrossRef](#)] [[PubMed](#)]
25. Jung, S.-J.; Kim, H.; Jang, K.-S. LLM based biological named entity recognition from scientific literature. In *Proceedings of the 2024 IEEE International Conference on Big Data and Smart Computing (BigComp), Bangkok, Thailand, 18–21 February 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 433–435.
26. Ni, J.; Xu, L.; Li, W.; Zheng, C.; Wu, L. Targeted metabolomics for serum amino acids and acylcarnitines in patients with lung cancer. *Exp. Ther. Med.* **2019**, *18*, 188–198. [[CrossRef](#)] [[PubMed](#)]
27. Hagihara, M.; Kato, H.; Yamashita, M.; Shibata, Y.; Umemura, T.; Mori, T.; Hirai, J.; Asai, N.; Mori, N.; Mikamo, H. Lung cancer progression alters lung and gut microbiomes and lipid metabolism. *Heliyon* **2024**, *10*, e23509. [[CrossRef](#)] [[PubMed](#)]
28. Vanhove, K.; Derveaux, E.; Graulus, G.-J.; Mesotten, L.; Thomeer, M.; Noben, J.-P.; Guedens, W.; Adriaenssens, P. Glutamine addiction and therapeutic strategies in lung cancer. *Int. J. Mol. Sci.* **2019**, *20*, 252. [[CrossRef](#)] [[PubMed](#)]
29. Dong, Y.; Wang, G.; Nie, D.; Xu, Y.; Bai, X.; Lu, C.; Jian, F.; Wang, H.; Zheng, X. Tumor-derived GABA promotes lung cancer progression by influencing TAMs polarization and neovascularization. *Int. Immunopharmacol.* **2024**, *126*, 111217. [[CrossRef](#)] [[PubMed](#)]
30. Kim, A.W.; Batus, M.; Myint, R.; Fidler, M.J.; Basu, S.; Bonomi, P.; Faber, L.P.; Wightman, S.C.; Warren, W.H.; McIntire, M.; et al. Prognostic value of xanthine oxidoreductase expression in patients with non-small cell lung cancer. *Lung Cancer* **2011**, *71*, 186–190. [[CrossRef](#)] [[PubMed](#)]

31. Horsfall, L.J.; Hall, I.P.; Nazareth, I. Serum urate and lung cancer: A cohort study and Mendelian randomization using UK Biobank. *Respir. Res.* **2021**, *22*, 179. [[CrossRef](#)] [[PubMed](#)]
32. Fahrman, J.F.; Grapov, D.D.; Wanichthanarak, K.; DeFelice, B.C.; Salemi, M.R.; Rom, W.N.; Gandara, D.R.; Phinney, B.S.; Fiehn, O.; Pass, H.; et al. Integrated metabolomics and proteomics highlight altered nicotinamide- and polyamine pathways in lung adenocarcinoma. *Carcinogenesis* **2017**, *38*, 271–280. [[CrossRef](#)] [[PubMed](#)]
33. Naughton, K.J.; Song, X.; Childress, A.R.; Skaggs, E.M.; Byrd, A.L.; Gosser, C.M.; Esoe, D.P.; DuCote, T.J.; Plaughner, D.R.; Lukyanchuk, A.; et al. Methionine restriction reduces lung cancer progression and increases chemotherapy response. *bioRxiv* **2024**. [[CrossRef](#)] [[PubMed](#)]
34. Wang, Z.; Zhang, J.; Shi, S.; Ma, H.; Wang, D.; Zuo, C.; Zhang, Q.; Lian, C. Predicting lung adenocarcinoma prognosis, immune escape, and pharmacomic profile from arginine and proline-related genes. *Sci. Rep.* **2023**, *13*, 15198. [[CrossRef](#)] [[PubMed](#)]
35. Liu, Y.; Mao, C.; Wang, M.; Liu, N.; Ouyang, L.; Liu, S.; Tang, H.; Cao, Y.; Liu, S.; Wang, X.; et al. Cancer progression is mediated by proline catabolism in non-small cell lung cancer. *Oncogene* **2020**, *39*, 2358–2376. [[CrossRef](#)] [[PubMed](#)]
36. Engelen, M.P.; Safar, A.M.; Bartter, T.; Koeman, F.; Deutz, N.E. Reduced arginine availability and nitric oxide synthesis in cancer is related to impaired endogenous arginine synthesis. *Clin. Sci.* **2016**, *130*, 1185–1195. [[CrossRef](#)] [[PubMed](#)]
37. Amelio, I.; Cutruzzolà, F.; Antonov, A.; Agostini, M.; Melino, G. Serine and glycine metabolism in cancer. *Trends Biochem. Sci.* **2014**, *39*, 191–198. [[CrossRef](#)] [[PubMed](#)]
38. Yang, L.; Li, A.; Yu, W.; Wang, H.; Zhang, L.; Wang, D.; Wang, Y.; Zhang, R.; Lei, Q.; Liu, Z.; et al. Blockade of purine metabolism reverses macrophage immunosuppression and enhances anti-tumor immunity in non-small cell lung cancer. *Drug Resist. Updat.* **2025**, *78*, 101175. [[CrossRef](#)] [[PubMed](#)]
39. Singh, A.; Prakash, V.; Gupta, N.; Kumar, A.; Kant, R.; Kumar, D. Serum metabolic disturbances in lung cancer investigated through an elaborate NMR-based serum metabolomics approach. *ACS Omega* **2022**, *7*, 5510–5520. [[CrossRef](#)] [[PubMed](#)]
40. Gui, Q.; Li, H.; Wang, A.; Zhao, X.; Tan, Z.; Chen, L.; Xu, K.; Xiao, C. The association between gut butyrate-producing bacteria and non-small-cell lung cancer. *J. Clin. Lab. Anal.* **2020**, *34*, e23318. [[CrossRef](#)] [[PubMed](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.