



Artificial intelligence-assisted proteomic signatures for discriminating malignant from benign pulmonary nodules

Juan Luo^{a,1}, Qiongye Dong^{a,1}, Jianyi Yang^{b,1}, Hui Shan^a, Qianqian Zhong^c, Yongjian Zhang^c, Guangxi Wang^{d,*}, Jixian Liu^{b,*}, Yuxin Yin^{a,*}

^a Precision Medicine Institute, Peking University Shenzhen Hospital, Shenzhen, Guangdong, 518036, China

^b Thoracic Surgery, Peking University Shenzhen Hospital, Shenzhen, Guangdong, 518036, China

^c Medical Examination Department, Peking University Shenzhen Hospital, Shenzhen, Guangdong, 518036, China

^d Institute of Systems Biomedicine, Department of Pathology, School of Basic Medical Sciences, Peking University Health Science Center, Beijing, 100191, China

ARTICLE INFO

Keywords:

Lung adenocarcinoma (LUAD)
Proteomics
Data-independent acquisition (DIA)
LightGBM
Noninvasive diagnosis

ABSTRACT

Lung cancer remains a major global health burden. Although low-dose CT (LDCT) is effective for early detection, its clinical application is limited by a high false-positive rate and the need for subsequent invasive confirmation, such as biopsy. Meanwhile, the extensive protein-level alterations observed in lung cancer provide a rationale for developing minimally invasive plasma proteomics-based screening a panel of biomarkers. In this study, untar-geted proteomic profiling was performed on a tissue cohort ($n = 61$, paired tumor and normal adjacent tissue samples spanning pre-invasive (AAH/AIS), MIA and IAC), a discovery plasma cohort ($n = 221$), and an independent validation plasma cohort ($n = 144$). Integrative tissue-plasma analysis was used to identify concordant proteins. Classifier training was conducted on the plasma proteomics data through a LightGBM-based pipeline with a two-stage feature selection strategy to derive plasma protein panels and construct binary classifiers that distinguish malignant from benign nodules. The diagnostic performance of the trained models was then evaluated in an independent validation cohort. Furthermore, binary classifiers were constructed to discriminate the invasiveness of lesions among LUADs. A subset of proteins was identified whose expression in both tissue and plasma was significantly associated with tumor invasiveness. The compact pipeline identified reproducible plasma protein panels and trained LightGBM classifiers that robustly discriminated malignant from benign pulmonary nodules. The classifiers maintained strong performance in the independent validation cohort (AUC-ROC: 0.949–0.986). We developed separate classification models to stratify lung adenocarcinoma subtypes by distinct patterns of invasiveness (AUC-ROC: 0.762–0.803). High-depth proteomics combined with LightGBM enables the identification of robust plasma protein panels for discriminating pulmonary nodules and stratifying invasiveness. This generalizable strategy, along with reproducible biomarker panels, supports non-invasive early diagnosis of lung adenocarcinoma.

Introduction

Lung cancer remains a leading cause of cancer-related morbidity and mortality worldwide, as most cases are diagnosed at advanced, less curable stages.^{1,2} Lung adenocarcinoma (LUAD), the predominant histological subtype of lung cancer,³ typically progresses from pre-invasive lesions such as atypical adenomatous hyperplasia (AAH) and adenocarcinoma in situ (AIS), through minimally invasive adenocarcinoma (MIA), to invasive adenocarcinoma (IAC).^{4,5} Early detection is critical in

lung cancer, as the 5-year survival rate exceeds 65% for stage I disease, but falls to 5% for stage IV disease.⁶ Accurate identification and early intervention for indolent pre-invasive and early lesions such as AAH and AIS can therefore substantially improve the long-term survival outcome.⁴ However, these clinical dilemmas are compounded by an incomplete understanding of the protein-level molecular alterations that underlie the sequential progression from AAH and AIS to invasive LUAD.

Low-dose computed tomography (LDCT) has proven effective for lung cancer screening and reducing lung cancer mortality. However, its

* Corresponding authors.

E-mail addresses: guangxiwang@bjmu.edu.cn (G. Wang), 252110465@qq.com (J. Liu), yinyuxin@pku.hsc.edu.cn (Y. Yin).

¹ These authors contributed equally.

clinical utility is limited by difficulties in differentiating malignant from benign pulmonary nodules.^{7,8} In major cohort studies such as the NLST and NELSON trials, the false positive rate of LDCT screening reached approximately 30%. Moreover, more than 95% of screen-detected nodules are ultimately benign, leading to patient anxiety, repeated examinations and unnecessary invasive procedures.^{9,10} In routine clinical practice, management strategies such as serial CT follow-up, PET-CT, minimally invasive biopsy, and multidisciplinary consultation are often required, yet these approaches do not always adequately balance diagnostic accuracy against the risks of overtreatment and delayed diagnosis.

Liquid biopsy modalities offer minimally invasive, repeatable approaches for cancer detection, but each has limitations for population-level early screening.^{11,12} Circulating tumor cells (CTCs) are exceedingly rare in early disease and remain technically difficult to isolate.¹³ Circulating tumor DNA (ctDNA) can capture tumor-specific genomic and epigenomic alterations, but early lesions often release only very limited amounts of ctDNA, resulting in diminished detection sensitivity for early-stage lesions.¹⁴ Circulating metabolite profiles are highly susceptible to dietary intake, medication, comorbidities and pre-analytical confounders.¹⁵ By contrast, circulating proteins, measured through high-depth plasma proteomics or targeted assays, offer a promising complementary strategy for nodule triage and early diagnosis. Plasma proteins may reflect both tumor-derived signals (secreted proteins, EVs, et.al.) and host responses (immune or inflammatory signatures).^{16–18} Indeed, plasma and serum protein biomarkers already have clinical precedent and several have been approved by the FDA for other indications.¹⁹

Recent methodological advances, particularly robust data-independent acquisition (DIA) mass-spectrometry workflows, now enable reproducible, large-scale quantification of thousands of proteins in clinical specimens, supporting the discovery of multi-protein signatures that may outperform single analytes for complex phenotypes.^{20–25} Nevertheless, untargeted proteomics also presents technical challenges, including the wide dynamic range of plasma proteins, batch effects, and missingness, which can complicate biomarker discovery and model development. Generating reproducible and clinically actionable signatures therefore requires careful study design, conservative feature-selection to minimize overfitting and independent validation.

To address these gaps, we performed an integrated, multi-cohort proteomic study. We profiled high-throughput proteomes from three cohorts, a tissue cohort comprising paired tumor and adjacent non-tumor samples, a discovery plasma cohort and an independent validation plasma cohort, spanning the AAH/AIS, MIA, and IAC continuum. A central innovation of our work is a two-stage, repeat-resampled feature selection and modelling pipeline centred on LightGBM.²⁶ LightGBM is a histogram-based, leaf-wise gradient boosting framework that efficiently handles high-dimensional data, captures non-linear interactions among proteins, tolerates missing values without aggressive imputation, and provides regularized training with interpretable feature importance metrics. These properties make LightGBM well suited to proteomic biomarker discovery and clinical classifier development.²⁷

Using this framework, we found that both lung tissue and plasma proteomes undergo substantial remodeling as LUAD progresses from pre-invasive (AAH/AIS) to MIA and IAC, with increasingly extensive alterations. We further established a systematic workflow integrating DIA high-throughput proteomics with the LightGBM artificial intelligence algorithm to identify candidate plasma protein panels for early detection of lung cancer. Potential protein panels were derived in the discovery cohort, using rigorous cross-validation and exhaustive hyperparameter tuning, and the resulting frozen models were evaluated in a blinded independent validation cohort to assess generalizability. We also developed a fully reproducible computational pipeline for robust proteomic biomarker panel discovery. Together, these analyses yielded binary classifiers capable of discriminating malignant lung lesions from benign pulmonary nodules, and stratifying lesions by the degree of

lesion invasiveness. These findings provide a translational foundation for targeted diagnostic assay development and prospective clinical validation.

Results

Study design and clinical characteristics of the collected cohorts

In order to identify plasma protein panels, we implemented a phased discovery-validation proteomics workflow integrating plasma and tissue cohorts (Fig. 1A). Three independent cohorts were collected, including one tissue cohort and two plasma cohorts (a discovery cohort and an independent validation cohort). The tissue cohort comprised paired tumor and adjacent non-tumor tissue samples across major lung cancer subtypes, including pre-invasive (AAH/AIS), MIA, and IAC. The plasma cohorts comprising patients with pathologically confirmed lung cancer, as well as individuals with pulmonary nodules (PNs) detected during routine physical examinations. Patients in the tissue cohort with pre-invasive lesions (AAH/AIS) were younger (mean age 46 years) than those with IAC (mean age 58 years). Similarly, in the plasma cohorts, the benign-nodule and AAH/AIS groups were younger (mean 49 and 47 years, respectively) than the IAC group (mean 56 years). All samples were profiled with an untargeted data-independent acquisition mass-spectrometry (DIA-MS) pipeline to yield proteome-wide measurements.

Using the discovery plasma cohort, we conducted a two-stage feature selection procedure and trained gradient-boosted decision-tree classifiers via LightGBM under rigorous cross-validation to limit overfitting and to derive a parsimonious biomarker panel. The final classifier was frozen upon completion of discovery-phase training and applied to the independent validation cohort without retraining. In the held-out validation cohort set, the model robustly discriminated malignant nodules from benign pulmonary nodules, achieving excellent AUC-ROC performance (Table 1).

Assessment of tissue and plasma proteomic data

Tissue and plasma specimens were analyzed by an untargeted DIA-MS pipeline with comprehensive quality control (QC) throughout acquisition. A commercial A549 digest was repeatedly analyzed as the process QC for the tissue cohort, while a pooled plasma aliquot was run interspersed among samples for the plasma cohort. QC replicates demonstrated high technical reproducibility across the study. The tissue QC runs yielded a mean of 5966 quantified proteins with a technical coefficient of variations (CVs) of 0.72% (Fig. 2a). Across tissue sample groups, we quantified on average 7096 proteins in normal tissue, 7214 in AAH, 7220 in AIS, 7150 in MIA, and 7236 in IAC. Median CVs across all biological replicates were 3–4% for each group, indicating stable performance of the tissue pipeline. Tissue protein intensities spanned more than seven orders of magnitude, and the ten most abundant proteins contributed roughly 24% of the total tissue proteome signal (Fig. 2d).

To maximize plasma proteomic depth, samples were depleted of the 14 most abundant proteins (High Select™ Depletion Spin Columns) prior to MS. Repeated plasma QC runs quantified a mean of 570 proteins (Fig. 2b) and showed excellent agreement (average Pearson correlation coefficient 0.97, range 0.96–0.98) (Fig. 2c) across the entire workflow. In the plasma discovery cohort, average numbers of quantified proteins were 561 (PNs), 569 (AAH), 565 (AIS), 565 (MIA), and 563 (IAC). Median CVs across biological replicates were 2% for QC, and 3–5% across biological groups, supporting high analytical consistency. As expected after depletion, immunoglobulins (IgG) and other removed high-abundance species were not among the top detected proteins. Nevertheless, the ten most abundant plasma proteins accounted for approximately 57.34% (57.45% for the validation cohort) of the total plasma signal (Fig. 2e). The independent validation plasma cohort showed comparable proteomic depth and reproducibility (group means

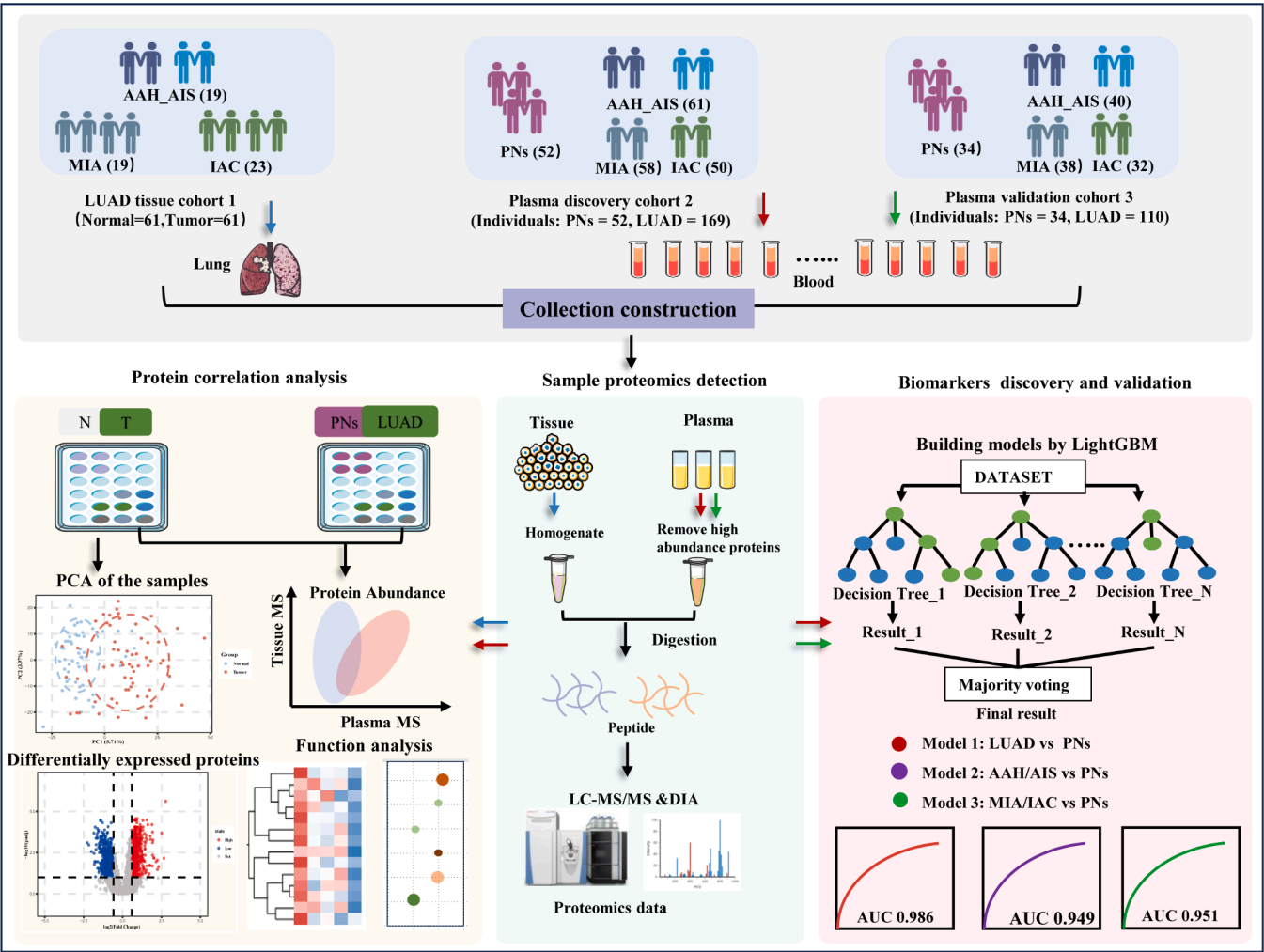


Fig. 1. Study design and overview of proteomics analyses in clinical cohorts. Three cohorts were collected, namely, tissue cohort 1 (N = 61), plasma discovery cohort 2 (N = 221) and plasma validation cohort 3 (N = 144) (top). High-throughput MS-based proteomics were used to profile the collected samples in three cohorts (bottom middle). An integrative analysis of tissue and plasma proteomic data were performed (bottom left). LightGBM was utilized to construct binary predictive models in the plasma discovery cohort to discriminate malignant pulmonary nodule from benign pulmonary nodule samples the models were validated in the independent plasma validation cohort.

Table 1
The classifiers showed modest but measurable discriminative ability in the validation cohort.

Model	Accuracy	Specificity	Sensitivity	Precision	F1_score	AUC
AAH/AIS/MIA/IAC vs PNs	0.931	0.735	0.991	0.924	0.956	0.986
AAH/AIS vs PNs	0.878	0.853	0.900	0.878	0.889	0.949
MIA/IAC vs PNs	0.837	0.676	0.914	0.853	0.883	0.951
AAH/AIS vs MIA	0.718	0.850	0.579	0.786	0.667	0.762
AAH/AIS vs IAC	0.736	0.725	0.750	0.686	0.716	0.799
MIA vs IAC	0.700	0.816	0.563	0.720	0.632	0.803

560–568 proteins; median CVs 2–5%; Fig. S1). Together, these QC metrics demonstrate that our DIA-MS workflow achieves both high proteome coverage and robust technical reproducibility for tissue and plasma profiling, providing a solid foundation for downstream differential analysis and classifier development.

Tissue proteomics analysis reveals progressive remodeling across LUAD subtypes

To characterize proteomic changes across the spectrum of LUAD, we performed large-scale deep untargeted DIA-MS profiling of tumor tissues and matched normal adjacent tissues (NATs) samples from a tissue

cohort (cohort1 in Fig. 1) comprising 61 patients with pre-invasive (AAH/AIS), MIA and IAC subtypes.

Unsupervised principal-component analysis (PCA) separated tumor and NATs samples along PC1 (Fig. 3a), with PC1 explaining approximately 5.9% of the variance. Paired differential abundance analysis (Tumor versus matched NATs) demonstrated extensive proteomic reprogramming in LUAD (Fig. 3b). Pathway enrichment of differentially abundant proteins showed concordant biological themes. Tumor tissue was enriched for translational and metabolic programs, including ribosome, amino acid degradation, and metabolism pathways, consistent with increased biosynthetic and energetic demands. In contrast, pathways related to complement and coagulation cascades, ECM-receptor

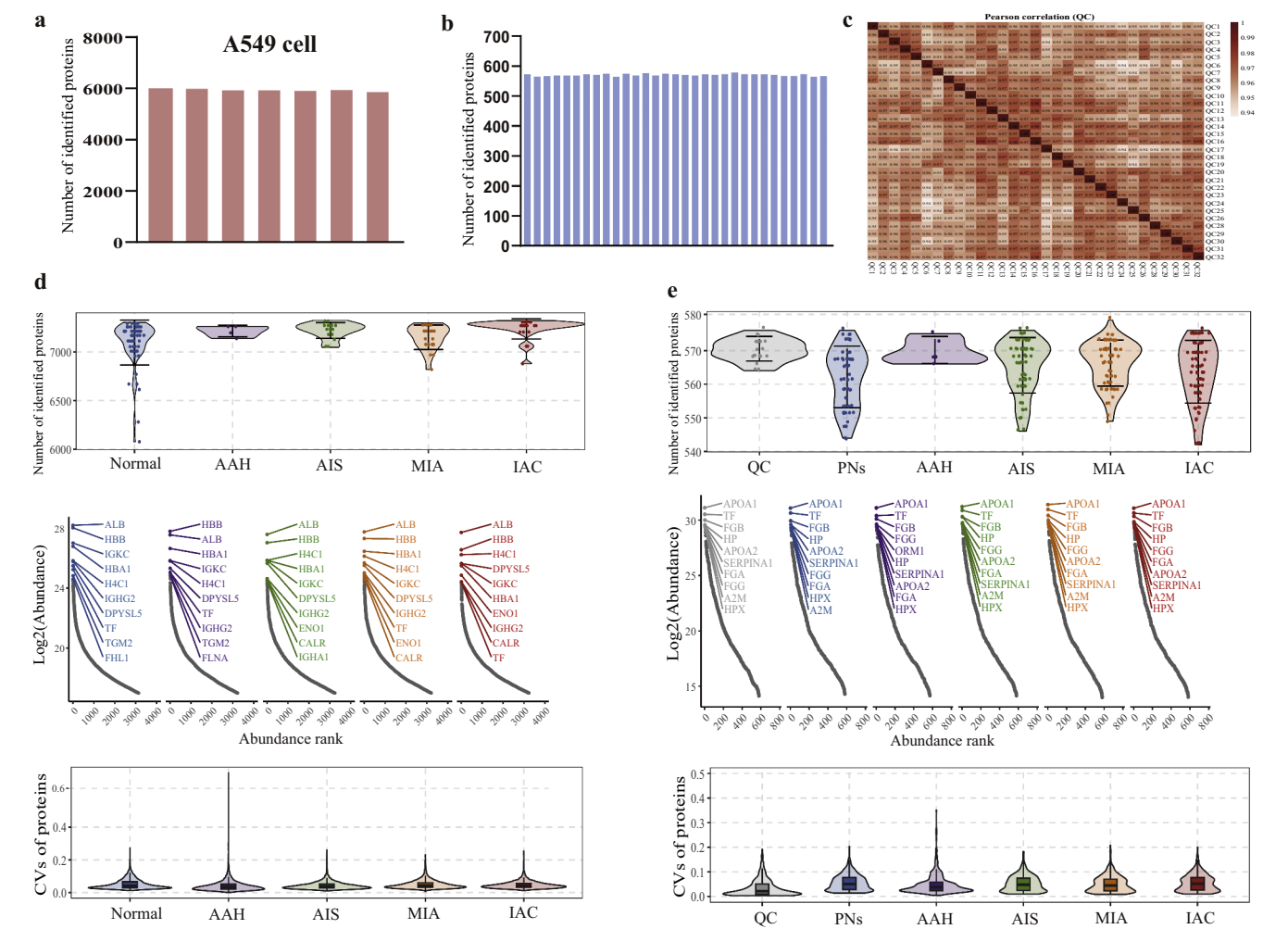


Fig. 2. Quality assessment of MS platform and proteomics data. (a). The proteins number identified in the QC (A549 cell) of tissue cohort. (b). The protein number identified in the QC of plasma cohort. (c). Pearson's correlation coefficients of QC replicates in the plasma proteomics experiments. (d-e). Quality assessment of proteomic data in tissue cohort (d) and plasma discovery cohort (e). Number of proteins discovered in five groups (up), Ranked proteins based on their median intensity within five groups. (middle). Distribution of CVs of biological samples (bottom).

interaction, focal adhesion, and PI3K-Akt signaling pathway were comparatively decreased in tumors, indicating substantial alterations in the remodeling of cell-matrix interactions and signaling, cell proliferation and migration during LUAD procession (Fig. 3c).

Considering subtype-specific patterns, a total of 1,148 proteins were significantly upregulated in tumors relative to NATs in at least one subtype (including 617 in AAH/AIS, 366 in MIA, and 728 in IAC), representing approximately 15.6% of the quantified tissue proteome. Conversely, 904 proteins were significantly downregulated (including 583 in AAH/AIS, 293 in MIA and 679 in IAC) in at least one subtype (representing 12.3% of the proteome) (Fig. 3d). Across AAH/AIS, MIA and IAC, 157 proteins were consistently upregulated and 195 consistently downregulated relative to NATs (Table S2). These shared alterations highlight common biological programs associated with the progression from pre-invasive to invasive disease.

Group-level heatmaps of mean expression revealed a stage-wise gradient of proteomic change from NATs and pre-invasive lesions (AAH/AIS) through MIA to IAC (Figs. 3e, S2d). Many proteins upregulated in tumors showed stepwise increases with advancing invasiveness, whereas downregulated proteins showed progressive loss in IAC, supporting a model of stepwise proteomic remodeling during LUAD progression. Notably, a subset of 23 co-dysregulated proteins (Fig. 3f) exhibited large (>2-fold) and statistically robust (FDR-adjusted p-values

< 0.01) changes in tumors versus NATs across three comparisons. Several of these are predicted secreted or plasma-accessible proteins, encoded by cancer-associated genes, or potential therapeutic targets.

Integrated analysis of plasma and tissue proteomics in LUAD

To identify proteins whose abundance changes are concordant between tumor tissue and the circulation, we performed an integrated analysis combining tissue proteomes (paired tumor and normal adjacent tissues across LUAD subtypes) with plasma proteomes from patients with pulmonary nodules and lung cancer (cohort 2 in Fig. 1). A total of 455 proteins were confidently quantified in both tissue and plasma datasets (Fig. 4a, Table S3). We then compared mean tumor tissue abundance to mean plasma levels for each protein and visualized the relationship in a scatter plot (Fig. 4b). Two principal clusters emerged, a diagonal cluster of proteins showing strong tissue-plasma concordance (blue box), and a vertical cluster of proteins that were abundant in tissue but present at low levels in plasma (red box). This partitioning is consistent with patterns previously reported in an integrated analysis of tissue-plasma in the context of liver diseases.²⁸

To determine whether proteomic alterations tracked with histopathological invasiveness, we assigned each sample an ordinal invasiveness score (NATs in the tissue cohort and PNs in the plasma cohort,

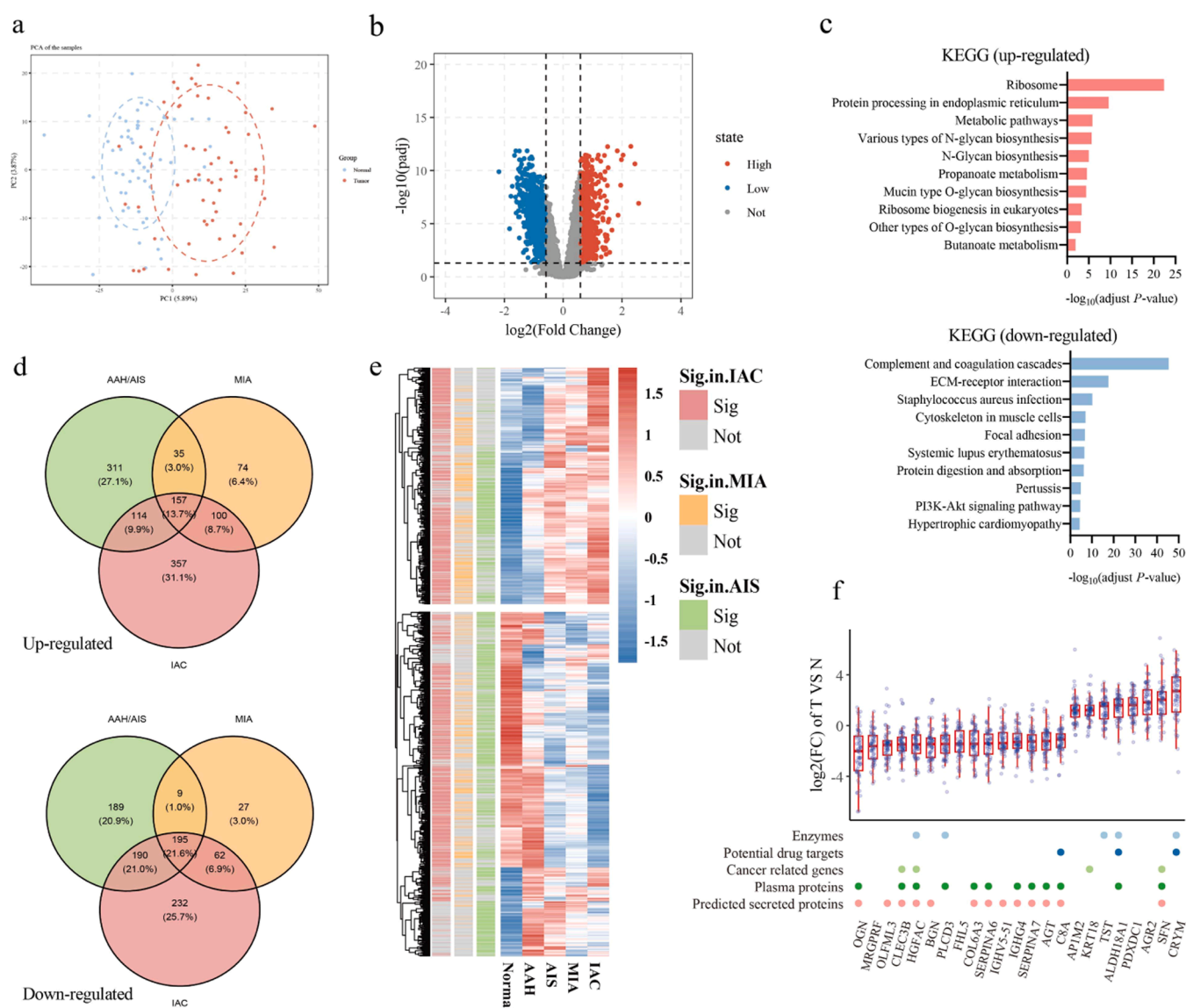


Fig. 3. Tissue proteomic analysis reveals global alterations across lung adenocarcinoma subtypes. (a). Principal-component analysis (PCA) of tissue proteomes. (b). Differential protein abundance between tumor and matched NATs. (c). Pathway enrichment of differentially expressed proteins. (d). Subtype-specific counts of differentially expressed proteins. (e). Heatmap of union of differentially expressed genes in three subtypes shows stage-wise alterations. (f). Boxplot showing $\log_2$ fold change between tumors and paired NATs for LUAD-associated proteins annotated with potential clinical utility by the Human Protein Atlas.

followed by AAH/AIS, MIA, and IAC) and calculated Spearman's correlations between invasiveness and protein abundance separately in tissue and plasma. We then intersected proteins that showed significant correlation with invasiveness in both of plasma and tissue datasets. This analysis identified 11 proteins with consistent positive correlations and 16 proteins with consistent negative correlations across tissue and plasma (Figs. 4c, S3a, b). The concordant directions and effect sizes supported a biological link between tumor processes and their circulating proteomic readouts.

Boxplot of these concordant proteins in different groups in the tissue and plasma cohorts, together with protein annotations (Figs. 4d, S3c), demonstrate that many tumor-associated perturbations are reflected systemically and detectable in plasma. Most of the 27 correlated proteins were annotated as plasma proteins or predicted secreted proteins. Among the 11 positive regulated proteins, most have been reported in the literature to be upregulated in lung cancer or associated with invasive phenotypes. Mechanistically, they converge on extracellular matrix remodeling (TIMP1, CTSD, FAP),^{29–31} metabolic and redox adaptation (GAPDH, LDHB, PRDX6),^{32–34} altered glycosylation and receptor

signaling (GALNT2, MOGS),^{35,36} and cell cycle progression and tumor formation (S100A6),³⁷ all processes that can promote tumor migration and invasion. Particularly, we observed that tissue LDHB levels increased from AAH/AIS to MIA before decreasing in IAC. In contrast, plasma LDHB levels increased continuously throughout the progression from AAH/AIS to IAC, suggesting that plasma LDHB likely reflects tissue leakage that escalates with the degree of infiltration. This phenomenon may be associated with enhanced exosome release from tumor cells — highly expressed LDHB is secreted into the circulatory system.³⁸ Overall, our integrated tissue–plasma analysis identified a subset of proteins whose expression in both tissue and plasma is significantly associated with tumor invasiveness. These concordant changes reflect underlying tumor biology and support the potential of plasma proteomics as a promoting non-invasive source of candidate biomarkers for further identification.

Construction of classification models from untargeted plasma proteomics

To derive non-invasive plasma protein panels that discriminate lung

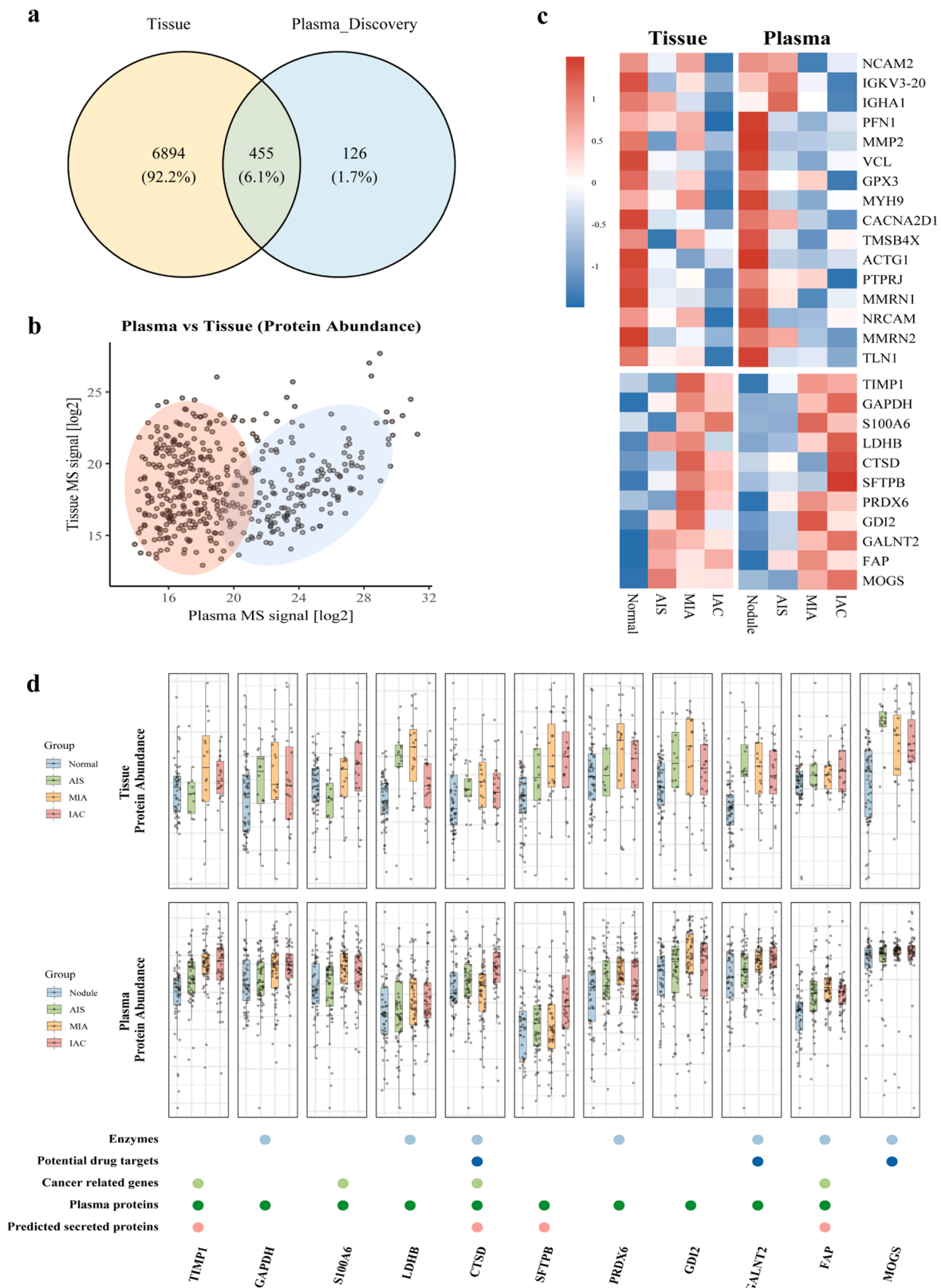


Fig. 4. Integrated tissue–plasma proteomic analysis identifies tissue–plasma concordant proteins associated with LUAD invasiveness. (a). Overlap of proteins confidently quantified in tissue and plasma datasets. A total of 455 proteins were detected in both compartments. (b). Scatter plot comparing mean tumor tissue abundance (x-axis) with mean plasma abundance (y-axis) for proteins quantified in both datasets. (c). Heatmap of mean protein abundance in different groups of the proteins with consistent positive correlations in both tissue and plasma ($n = 11$) and consistent negative correlations ($n = 16$). (d). Boxplots of selected concordant positive proteins across histological groups in tissue (top) and plasma (bottom), annotated by the Human Protein Atlas.

cancer from indeterminate pulmonary nodules, we trained gradient-boosted tree classifiers (LightGBM) on the discovery plasma proteome (cohort 2 in Fig. 1) using a conservative, cross-validated feature selection pipeline and rigorous internal validation. Raw untargeted protein quantifications were first filtered to retain proteins detected in >80% of samples, then log-transformed and normalized (see Methods). From the

processed discovery data, we trained three LightGBM classifiers and applied them (without further retraining) to an independent validation cohort. Fig. 5a provides a schematic overview of the overall machine learning workflow. The three binary tasks were: (1) lung adenocarcinoma versus pulmonary nodules, (2) non-invasive lesions (AAH/AIS) versus pulmonary nodules, and (3) invasive lesions (MIA/IAC) versus

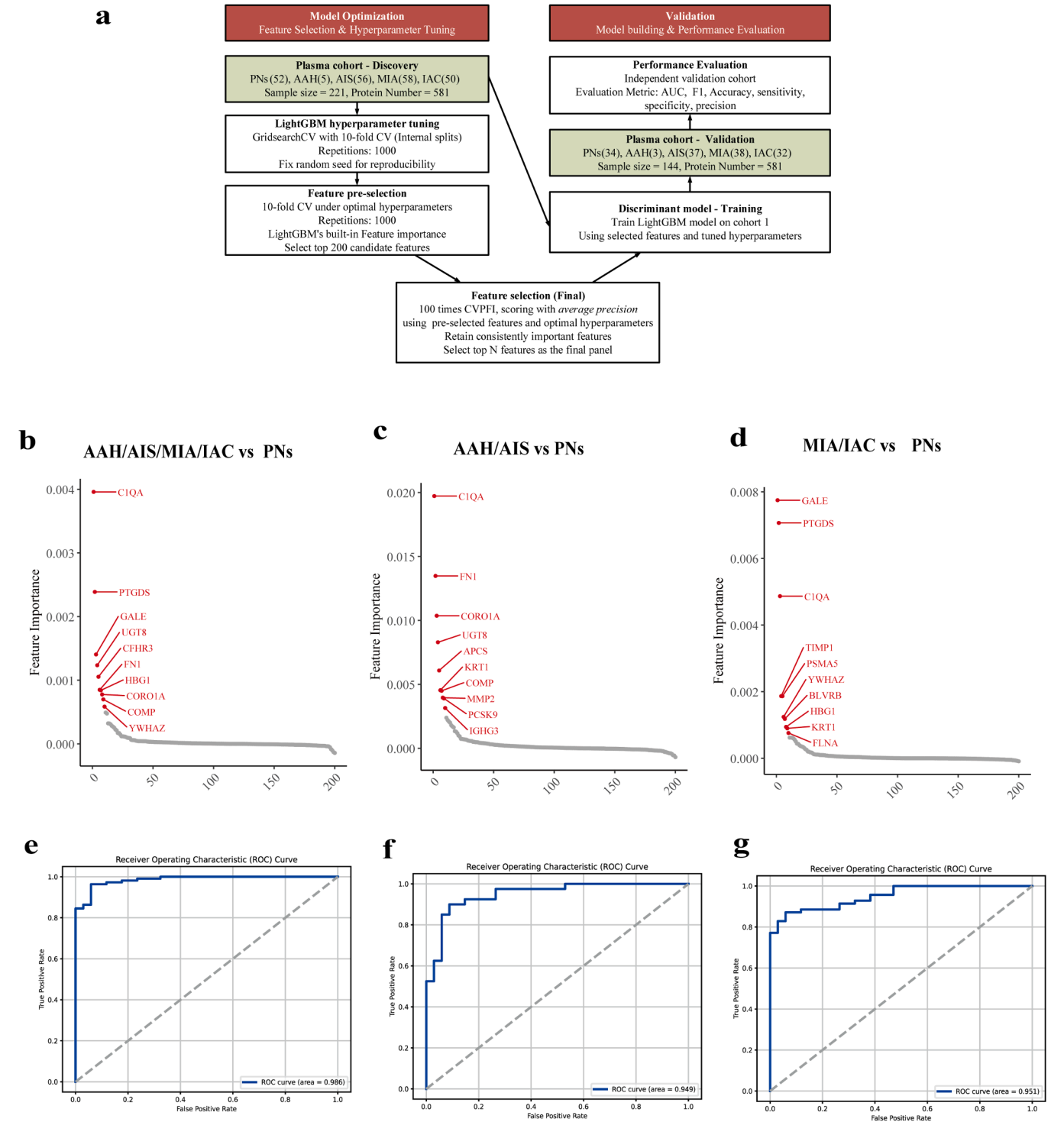


Fig. 5. Feature ranking and validation performance of LightGBM classifiers built from untargeted plasma proteomics. (a). Flowchart of model optimization (hyperparameter tuning and feature selection) and validation (performance evaluation). (b–d) Mean CVPFI rankings of mean permutation importance across CVPFI repeats for feature selection for each binary task. (e–g) Receiver-operating-characteristic (ROC) curves for the three frozen classifiers when applied without retraining to the independent validation cohort. b, e. Adenocarcinoma (all stages) vs pulmonary nodules. c, f. Non-invasive lesions (AAH/AIS) vs pulmonary nodules. d, g. Invasive lesions (MIA/IAC) vs pulmonary nodules.

pulmonary nodules. Feature ranking was performed using mean cross-validated permutation feature importance (CVPFI). The top-ranked proteins are shown in Fig. 5b–d and the top ten proteins were selected to build the classifiers. These proteins not only changed (increased or decreased) with disease progression but also exhibited significant variation across different histological subtypes (Fig. S4), as exemplified by C1QA, APCS, IGHG3, and FN1. The corresponding ROC curves on the validation cohort are shown in Fig. 5e–g. On the independent validation set, the models achieved AUCs of 0.986, 0.946, and 0.951, with mean F1 scores of 0.956, 0.889, and 0.883, respectively. In summary, a compact ten-protein panel derived from untargeted DIA-MS and selected via a conservative, cross-validated pipeline yields highly discriminative classifiers that generalize to an independent cohort, supporting the promise of plasma proteomics for non-invasive lung cancer discrimination from benign pulmonary nodules.

Separability of lung adenocarcinoma invasive subtypes by plasma proteomics

To determine whether plasma proteomes can resolve tumor invasiveness within lung adenocarcinoma, we applied the same LightGBM-based, cross-validated feature selection pipeline used above to pairwise subtype comparisons, using a slightly less parsimonious feature set to capture subtler differences between subtypes. After preprocessing and hyperparameter tuning (Methods), we retained the top 35 features (by CVPFI ranking) to train binary classifiers for the following pairwise tasks: AAH/AIS vs MIA, AAH/AIS vs IAC, and MIA vs IAC (Tables S4–S6). Models were trained in the discovery cohort and evaluated as frozen predictors on the independent validation cohort. Ranked proteins and the AUC plot were presented (AUCs of 0.762, 0.799, and 0.803), and the top 10 of these proteins are displayed (Fig. 6). In the

independent validation cohort, the classifiers demonstrated modest but measurable discriminative performance (Table 1). Analysis of the 35-protein panels in each model revealed significant variation in the expression of some proteins across histological subtypes (Tables S7–S9). These results indicate that plasma proteomes carry information related to tumor invasiveness at the population level, yet pairwise discrimination among adjacent stages is substantially less accurate than distinguishing cancer versus benign nodules. Likely contributors to this reduced accuracy include the biological continuity across the adenocarcinoma progression spectrum, dilution of tumor-derived signals in plasma, inter-patient heterogeneity, and limited sample sizes for individual subtypes. In summary, plasma proteomics shows promise for indicating tumor invasiveness across cohorts, but reliable discrimination of adjacent invasive subtypes for individual patients will require larger, better-powered datasets and further methodological refinement.

Discussion

Current early lung cancer screening strategies have important limitations in terms of accuracy, accessibility and the potential for over-diagnosis.^{39,40} Although LDCT enables sensitive detection of pulmonary nodules, its clinical utility is limited by suboptimal specificity. Many benign lesions detected by LDCT require additional imaging evaluations, prolonged surveillance, or even invasive diagnostic procedures, which may increase patient anxiety and impose an incremental healthcare burden.⁴¹ In addition, LDCT has limited ability to discriminate tumor invasiveness or molecular subtypes, potentially delaying appropriate treatment decisions. Moreover, cost and resource constraints further limit access to population screening in some regions, compounding these challenges.⁴²

Given the complexity of lung cancer, plasma proteomics offers a

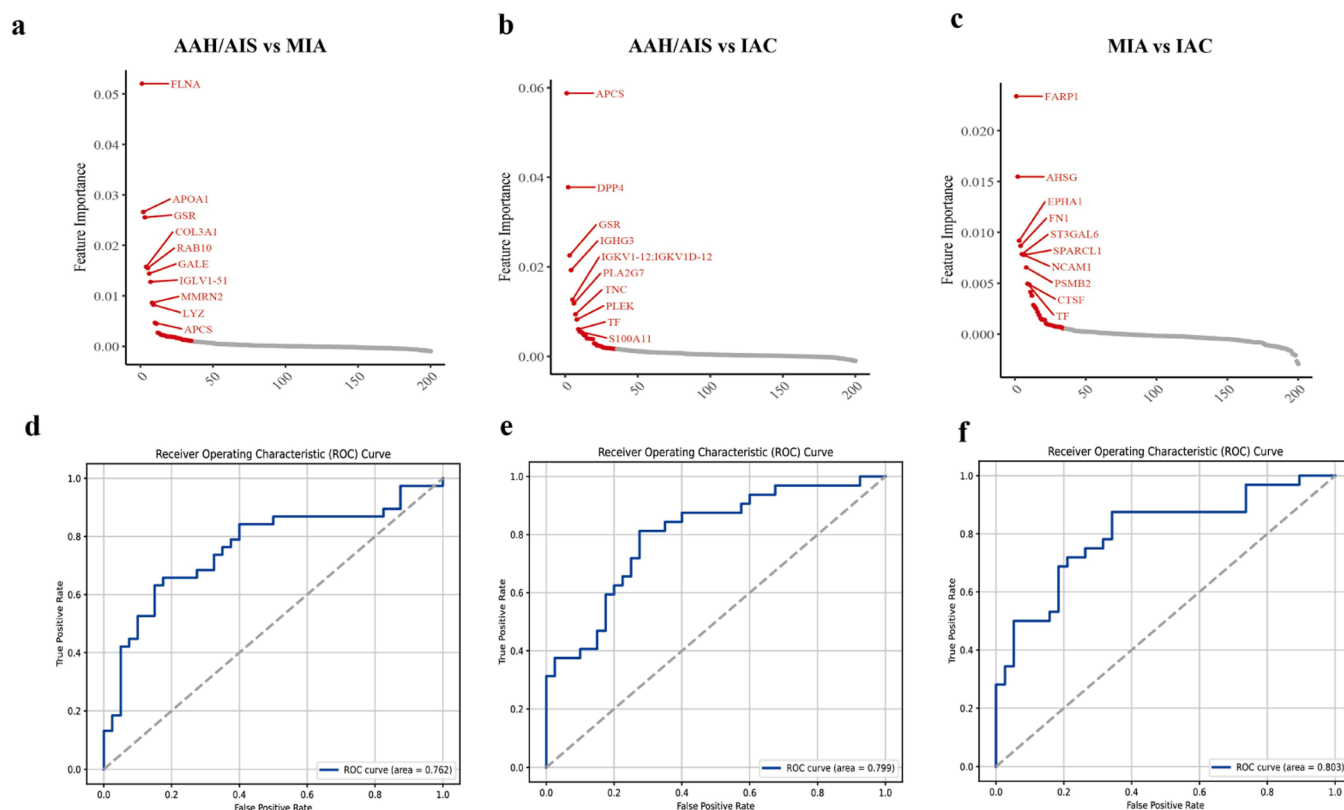


Fig. 6. Feature ranking and validation performance of LightGBM classifiers built from untargeted plasma proteomics. (a–c) Mean CVPFI rankings of mean permutation importance across CVPFI repeats for feature selection for each binary task. (d–f) Receiver-operating-characteristic (ROC) curves for the three frozen classifiers when applied without retraining to the independent validation cohort. a, d. Non-invasive lesions (AAH/AIS) vs MIA; b, e. Non-invasive lesions (AAH/AIS) vs IAC; c, f. MIA vs IAC.

complementary, minimally invasive approach that may help address several of these gaps.⁴³ Plasma is readily obtainable, allows repeat sampling for longitudinal monitoring, and can reveal circulating proteins reflecting the underlying biology of diseases.^{44,45} A validated plasma proteomic classifier enables accurate discrimination between malignant and benign pulmonary nodules, which reduces unnecessary biopsies, follow-up surveillance and downstream diagnostic testing for benign lesions, relieves patient anxiety, improves early detection of malignant lesions,^{43,46} and supports subtype-specific management and referrals for targeted therapies.⁴⁷ Plasma proteomics—by providing minimally invasive, repeatable measurements of circulating proteins—offers a promising route to improve nodule classification. Moreover, discovery of subtype-specific protein signatures for lung adenocarcinoma could inform precision treatment: different histologic or molecular subtypes may exhibit differential sensitivity to targeted therapies or immunotherapy, and proteomic biomarkers could help stratify patients for the most effective interventions. Taken together, plasma proteomic profiling has the potential to enhance diagnostic accuracy, enable earlier detection, and guide personalized management strategies for patients.⁴⁶

Here, we conducted MS-based proteomic profiling of plasma and integrated these data with a paired tissue proteome (tumor and adjacent non-tumor tissue) spanning multiple lung cancer subtypes. The sample sizes of the three cohorts were determined according to their respective study objectives and sample availability. Because tissue specimens are difficult to obtain due to limited clinical resources and stringent inclusion criteria, the tissue cohort was relatively small. By contrast, the plasma discovery cohort was larger to reduce the influence of technical and biological variability and improve the robustness of biomarker screening. The validation cohort was used solely to test predefined biomarkers, thereby substantially reducing the burden of multiple comparisons and allowing a small sample size to be sufficient. Importantly, this cohort imbalance did not compromise the subsequent proteomic analysis or model development, consistent with previous studies.^{24,25,28} The result reveals a stage-wise gradient of proteomic changes during LUAD progression, characterized by the progressive dysregulation of proteins with increasing invasiveness. In addition, our integrated tissue–plasma analysis identified a subset of proteins whose expression in both tissue and plasma is significantly associated with tumor invasiveness. Specifically, most of these 27 correlated proteins were annotated as plasma proteins or predicted secreted proteins. It has been reported that genomics and transcriptomics have established the molecular basis of LUAD by identifying driver mutations and transcriptional networks underlying subtype heterogeneity.⁴⁸ Moreover, metabolomic analysis of resected lesions and plasma from invasive LUAD patients has delineated key metabolic reprogramming events.⁴⁹ However, proteomic studies that integrate tissue and plasma profiles to characterize protein remodeling remain underdeveloped. Here, we bridge this gap through an integrated research model, which provides direct evidence for a stepwise proteomic remodeling process from pre-invasive lesions to invasive adenocarcinoma.

At the same time, we investigated the feasibility of detecting pulmonary nodules using a state-of-the-art, MS-based proteomics workflow, and trained LightGBM classifiers in a rigorously cross-validated discovery set, and then evaluated the final locked model, with all pre-processing procedures and model parameters fixed, in an independent validation cohort. The resulting AI-assisted proteomic signatures reproducibly discriminated malignant from benign pulmonary nodules, and supported subtype-level classification. As the plasma proteome spans a very wide dynamic range, high-abundance proteins (for example immunoglobulins) can mask lower-abundance proteins that may carry clinical signal.⁵⁰ To mitigate this, we applied high-abundance protein depletion together with state-of-the-art mass spectrometry methods (DIA) to increase depth and reproducibility of detection.^{51,52} To evaluate and mitigate potential batch effects, we pooled the discovery and validation cohorts for systematic assessment and incorporated regularly

injected pooled QC samples during acquisition to monitor instrument performance and support inter-batch correction for the untargeted DIA proteomics. The number of proteins identified across all QC samples were comparable, with median CVs below 5%. Furthermore, we applied a two-stage feature-selection pipeline. We first performed a rapid pre-selection using LightGBM's built-in feature-importance (gain) metric,²⁶ followed by a model-agnostic, CVPFI ranking to obtain stable importance estimates. The reduced feature set was used to train a LightGBM classifier with repeated cross-validated hyperparameter tuning and resampling to stabilize performance estimates. The final locked model was then evaluated on the independent validation cohort to quantify generalizability. This QC-guided, two-step selection and independent-validation workflow was designed to minimize batch-driven artifacts and overfitting, stabilize feature ranking, and provide conservative, reproducible estimates of classifier performance, thereby improving confidence in downstream clinical translation.

Ensemble tree-based learners are well suited to proteomic classifiers because they capture non-linear relationships and interactions among features and are relatively robust to noisy predictors. We chose LightGBM²⁶ for classifier construction: a histogram-based gradient boosting framework that grows tree leaf-wise and natively accommodates missing values by optimally assigning missing-value directions during split evaluation. These properties reduce the need for aggressive imputation and increase computational efficiency for high-resolution proteomics data. To further control overfitting and stabilize performance, we carried out extensive repeated cross-validated hyperparameter tuning and cross-validated permutation feature importance, selecting a concise panel of proteins that showed consistent predictive value. In the study, we developed and validated plasma protein panels and trained LightGBM classifiers that robustly discriminated malignant from benign pulmonary nodules. Final classifiers maintained strong performance in the independent validation cohort (AUC-ROC: 0.949–0.986), and each diagnostic panel comprised ten plasma proteins. Our study also established separate classification models for stratifying LUAD invasiveness subtypes, with robust performance (AUC-ROC: 0.762–0.803) while maintaining a consistent panel size of 35 plasma proteins per model. This approach highlights the potential of plasma protein biomarkers for distinguishing malignant from benign pulmonary lesions and supports their application as a minimally invasive adjunct for early lung cancer detection and pulmonary nodule assessment. Notably, our aim is not to replace LDCT-based screening, but rather to complement existing imaging approaches by providing additional molecular information that may improve risk stratification and diagnostic decision-making for LDCT-detected pulmonary nodules.

Several inherent limitations associated with this work ought to be acknowledged. Although the discovery and validation cohorts were independently collected, both were of modest size and derived from a single center. Consequently, the generalizability of our findings requires further evaluation in larger, prospective, blinded, and multicenter studies. Future studies are also needed to assess classifier performance across diverse populations and clinical settings before routine clinical implementation. Moreover, cases featuring multiple lesions that include benign, pre-invasive, and invasive subtypes are uncommon. It is also challenging to track a single patient from the initial detection of a pulmonary nodule to the eventual development of lung cancer while obtaining serial samples. Consequently, larger cohorts with long-term follow-up are essential to validate our findings. In addition, biological confounders that we reduced by design may still influence protein abundances, and technical variability between platforms or laboratories could affect classifier performance. Lastly, while untargeted discovery identifies candidate biomarkers, translation into a routine clinical test will likely require development of a targeted, analytically validated assay (for example, a multiplexed SRM/PRM or immunoassay) and prospective clinical evaluation.^{53,54} With further validation and transition to targeted assays, plasma proteomic classifiers may ultimately improve the risk stratification of pulmonary nodules, reduce

unnecessary interventions and inform personalized management strategies for patients with pulmonary nodules.

Conclusions

In conclusion, we applied an enhanced DIA-MS workflow to three independent cohorts. A subset of proteins was identified whose expression in both tissue and plasma was significantly associated with tumor invasiveness. Furthermore, we developed a two-stage, LightGBM-centered feature selection pipeline. In the discovery cohort, we then trained classifiers through stringent repeated cross-validation and extensive hyperparameter tuning. This approach identified compact plasma protein panels capable of robustly discriminating malignant from benign pulmonary nodules, as well as models for invasiveness stratification. The overall strategy, combining a reproducible DIA-MS platform with a generalizable LightGBM pipeline, effectively supports the non-invasive early diagnosis of lung adenocarcinoma.

Methods

Ethics

The LUAD samples for this study were collected from Peking University Shenzhen Hospital in China between January 2024 and February 2025, and patients were randomly selected at their first visit and had not received any anti-cancer treatments prior to surgery. The control samples were collected from donors with pulmonary nodules that were identified during routine physical examinations via CT or LDCT scans. The protocol was approved by institutional review boards and the Ethics Committee of Peking University Shenzhen Hospital (approval number: 2023 – 178). The study was conducted according to the principles of the Declaration of Helsinki, and oral and written informed consent was obtained from all participants. We ensured that the data users or all persons involved in this study agreed to protect the confidentiality of the participants when handling data containing potentially identifiable information about all patients.

Samples collection and characteristics

All samples in this study were divided into three cohorts: tissue

cohort 1 (N = 61), plasma discovery cohort 2 (N = 221) and plasma validation cohort 3 (N = 144). In the discovery and internal validation cohorts, plasma of 86 cases of pulmonary nodules was obtained from individuals undergoing annual physical examination at physical examination center, and cases of patients with lung adenocarcinoma were collected from the thoracic surgery departments. Subjects of pulmonary nodules were those with stable evaluations by follow-up CT scans over 1-3 years at the time of analysis, and lung adenocarcinoma cases were obtained before lung resection surgery and confirmed by pathological diagnosis, which contained 86 patients with AAH and AIS, 96 patients with MIA, and 82 patients with IAC. Their associated clinical data (including gender, age, tumor size, histological subtype, differentiation degree, TNM stage, tumor purity, smoking status) were collected (Table 2). The inclusion criteria were an age range of 24-75 years and no history of any other pulmonary diseases. Exclusion criteria included pregnancy, history of malignancy, concurrent pneumonia or other pulmonary diseases, and any prior radiotherapy or chemotherapy. All tissues were surgically resected, and the corresponding normal adjacent tissues (more than 3 cm apart from tumor edge) were collected and immediately placed into sterile cryovials. The vials were then immersed in liquid nitrogen and stored at –80°C until subsequent analysis. In addition, plasma sample was collected using Vacutainer EDTA tubes and centrifuged at 3500 rpm for 10 min, then stored at –80°C for analysis.

Lung tissue proteome sample preparation

The collected tissue samples were thoroughly washed two times with phosphate-buffered saline (PBS) to remove residual blood and cellular debris. Subsequently, these samples were lysed in ice-cold lysis buffer (4%SDS, 100 mM Tris-HCl, pH 8.0) containing protease inhibitors for 10 minutes. Samples were homogenized on ice (total 2 min, in 60 s bursts) and centrifuged at 16,000 × g for 15 min at 4°C to pellet debris. Supernatants were collected and protein concentrations measured by BCA assay. Extracted proteins were reduced with 10 M dithiothreitol (DTT) for 30 min and alkylated with 200 mM iodoacetamide (IAA) for 30 min at room temperature in the dark. Proteins were precipitated with 10% (w/v) trichloroacetic acid (TCA) at 4°C for 30 min and pelleted by centrifugation at 16,000 × g for 15 min. Pellets were washed twice with cold acetone to remove residual TCA, air-dried briefly, and resuspended for enzymatic digestion. Proteins were digested with trypsin at an

Table 2
Detailed characteristics of all patients.

Characteristic	Tissue cohort			Plasma cohort			
	AHH/AIS(n = 19)	MIA(n = 19)	IAC(n = 23)	PNs (n = 86)	AHH/AIS(n = 86)	MIA(n = 96)	IAC(n = 82)
Age(mean±SD)	46.00±12.74	53.00±7.62	58.87±11.40	49.06±10.66	47.57±11.09	47.48±11.20	56.63±11.49
Gender							
Male	10(52.63%)	6(31.58%)	12(52.17%)	51(59.30%)	25(29.07%)	24(25.00%)	36(43.90%)
Female	9(47.37%)	13(68.42%)	11(47.83%)	35(40.7%)	61(70.93%)	72(75.00%)	46(56.10%)
Smoking history							
Smoking	1	3(15.79%)	5(21.74%)	15(17.44%)	10(11.63%)	11(11.46%)	22(26.83%)
Nosmoking	18	16(84.21%)	18(78.26%)	71(82.56%)	76(88.37%)	85(88.54%)	60(73.17%)
Nodule type							
GGO	10(52.63%)	10(52.63%)	/		77(89.53%)	76(79.17%)	6(7.32%)
PSN<50%	8(42.11%)	4(21.05%)	6(26.09%)		9(10.47%)	16(16.67%)	27(32.93%)
PSN>50%	/	4(21.05%)	9(39.13%)		/	3(3.13%)	20(24.39%)
SN	1(5.26%)	1(5.26%)	8(34.78%)		/	1(1.04%)	29(35.37%)
Nodule size							
<10mm	13(68.42%)	5(26.32%)	/		64(74.42%)	69(71.88%)	5(6.10%)
10–20 mm	6(31.58%)	13(68.42%)	10(43.48%)		22(25.58%)	27(28.13%)	52(63.41%)
20–30mm	/	1(5.26%)	9(39.13%)				13(15.85%)
>30 mm	/	/	4(17.39%)				12(14.63%)
Stage							
0	19(100%)				82(95.35%)	1(1.04%)	
I		19(100%)	18(78.26%)		4(4.65%)	95(98.96%)	70(85.37%)
II			3(13.04%)				5(6.10%)
III			2(8.70%)				6(7.32%)
IV							1(1.22%)

enzyme-to-substrate ratio of 1:50 at 37°C for 16 h. Resulting peptides were desalted using SepPak C18 cartridges (Waters, MA), concentrated by vacuum centrifugation (SpeedVac), and subjected to LC–MS/MS analysis.

Plasma proteome sample preparation

Collected plasma samples were first depleted of high-abundance proteins using High Select™ Depletion Spin Columns (Thermo Scientific High Select HAS, A36366) according to the manufacturer's instructions. Depleted plasma was processed using a modified Filter-Aided Sample Preparation (FASP) protocol. Briefly, the depleted plasma was transferred to a 10 kDa cutoff Microcon filter unit (Millipore) and centrifuged at $14,000 \times g$ at 20°C for 15 min. Subsequently, 100 µL lysis buffer (8 M urea in 100 mM Tris, pH 8.5) was added to the filter to wash the plasma twice by centrifugation at $14,000 \times g$ at 20°C for 15 min per wash. Proteins retained on the filter were reduced with 100 µL 100 mM dithiothreitol (DTT) at 37°C for 60 min and alkylated with 100 µL 200 mM iodoacetamide (IAA) at 37°C for 60 min in the dark. The precipitate in the filter was washed twice with 200 µL 100 mM triethylammonium bicarbonate (TEAB) and centrifuged at $14,000 \times g$ at 20°C for 15 min. Proteins were digested with trypsin at an enzyme-to-substrate ratio of 1:50 at 37°C for 16 hours. After digestion, peptides were recovered, desalted using SepPak C18 cartridges (Waters, MA), and vacuum-dried by Speed Vac centrifuge at 35°C prior to analysis.

Liquid chromatography-mass spectrometry acquisition

Plasma sample acquisition order was randomized to minimize batch effects. Dried peptides were reconstituted in 0.1% formic acid and loaded onto a trap column (Acclaim™ PepMap™ 100 75 µm × 2 cm column, nanoViper 2kv C18, 3 µm, 100A, Thermo Fisher Scientific), and separated on an analytical column (Acclaim™ PepMap™ RSLC C18 75 µm × 25 cm column, Thermo Fisher Scientific). Tissue samples were analyzed on a single-column setup with an Aurora Ultimate analytical column (25 cm × 75 µm, C18 1.7 µm) maintained at 50°C. Solvent A was 0.1% formic acid in water and solvent B was 80% acetonitrile with 0.1% formic acid. The flow rate was 300 nL/min and the gradient was: 0 min, 3%B; 0–3 min, 3–8%B; 3–86 min, 8–35%B; 86–96 min, 35–40%B; 96–100 min, 90%B; 100–120 min, 90%B.

MS data were acquired on an Orbitrap Exploris 480 (ThermoFisher Scientific, United States) operated in data-independent acquisition (DIA) mode. Instrument parameters were: RF lens frequency 40%; pos spray voltage: 25kV; ion transfer tube temperature: 320°C. Full MS scans were acquired from 350 to 1200 m/z in the Orbitrap mass analyzer at a high resolution of 120,000 with an auto matic gain control (AGC) target of $3E+06$ and a maximum ion injection time of 50 ms. MS2 scans were acquired at a 30,000 resolution with AGC value of $2E+06$ and a maximal ion injection time of 50 ms. HCD fragmentation was used with stepped normalized collision energy of 25%, 30%, and 35%. Data were recorded in profile mode. The DIA isolation scheme (centered isolation window) was: 361.5–517.0 m/z, 524.0–657.0 m/z, 664–797 m/z, 807.5–1181.5 m/z (wide window). Raw data were collected with Xcalibur v4.4.16.14 (Thermo Fisher Scientific).

Data preprocessing and protein quantification for tissue and plasma untargeted proteomics

Raw DIA data acquired from tissue and plasma samples were processed using Spectronaut2 (v17.0) following the vendor-recommended workflow.⁵⁵ DirectDIA search parameters were: enzyme specificity trypsin, a maximum of two missed cleavage, fixed modification with carbamidomethylation on C, variable modifications with oxidation on M and N-terminus protein acetylation. Proteotypic peptides only were retained and false discovery rates (FDR) were controlled at 1% for peptide-spectrum matches (PSMs), peptides and proteins. The search

database was the reviewed human UniProt FASTA (Homo sapiens, n = 20,433 reviewed entries; 2024 release). Label-free quantification was enabled with a minimum ratio count of 2, and intensity-based absolute quantification (iBAQ) values were calculated for protein-level quantification. Proteins with missing values in >20% of plasma samples or >25% of tissue samples were excluded from downstream analyses. After filtering, 581 protein groups remained in the plasma cohort and 7,349 protein groups remained in the tissue cohort. Samples with protein-group counts outside the whiskers of the boxplot (statistical outliers) were excluded from downstream analysis. Protein abundance was log2-transformed, and median normalization (normalizeMedianAbsValues from limma package) was applied to generate the final normalized protein matrix for subsequent analysis. Data were processed using R (v4.5.1).

Mass-spectrometry quality control and assessment

Instrument and workflow performance were monitored throughout data acquisition. For tissue-proteome QC, an A549 cell protein digest was repeatedly measured across the run to assess the reproducibility. For plasma-proteome QC, a pooled quality control (QC) sample, generated by combining aliquots of tryptic-digested plasma peptides from multiple donors, was injected regularly among analytical samples to monitor instrument stability. Pearson correlation coefficients between QC runs were calculated in R (v4.5.1), and the resulting correlation matrix was visualized as a heatmap using the pheatmap R package. All other statistical plots were generated using GraphPad Prism v8.0.

Grouping paired differential-protein and functional enrichment analysis in tissue cohort

Because of the limited sample size of the AAH group in the tissue cohort, AAH and AIS samples were combined into a single non-invasive group in the analysis. Tumor and adjacent normal (paired) samples were analyzed as matched pairs. For each of the three groups (AAH/AIS [non-invasive], MIA, and IAC), differential abundance was assessed using two-sided paired t-tests. Raw P-values were adjusted for multiple testing by the Benjamini–Hochberg procedure to control the false discovery rate (FDR). Differentially expressed proteins were defined as $|\log_2 \text{fold change}| > \log_2(1.5)$ and $\text{FDR} < 0.05$. Proteins that were consistently up-regulated (or consistently down-regulated) across all three group comparisons were taken forward for downstream functional enrichment analysis. Enrichment analyses were performed in R using the clusterProfiler package with appropriate human annotation (e.g., org.Hs.eg.db), and significance was assessed on FDR-adjusted P-values ($q < 0.05$). Visualizations—including volcano plots and heatmaps—were generated with ggplot2 and pheatmap. All statistical analyses were conducted in R (v4.5.1).

Integrated analysis of plasma and tissue proteomes

A total of 455 proteins were detected in both the plasma and tissue proteomic datasets and carried forward for integrative analysis. To assess relationships between protein abundance and disease malignancy and invasiveness, samples in each cohort were assigned an ordinal score with four levels. For the plasma cohort, samples were scored as: 0 = benign pulmonary nodule, 1 = non-invasive (AAH/AIS), 2 = MIA, 3 = IAC, while for the tissue cohort samples were scored as: 0 = distal adjacent normal, 1 = non-invasive (AAH/AIS), 2 = MIA, 3 = IAC. For each of the 455 shared proteins, we computed the association between abundance and the assigned ordinal score separately in the plasma and tissue cohorts using Spearman's rank correlation (two-sided). Spearman's rho (ρ) was estimated and interpreted as the strength of monotonic association. Proteins with an absolute correlation coefficient $|\rho| > 0.15$ were considered correlated with malignancy/invasiveness. Correlation results were summarized as correlation matrices and visualized

with heatmaps (pheatmap), and per-group distributions were plotted with boxplots generated by ggplot2. All analyses were performed in R (v4.5.1).

Classifier construction using with LightGBM and performance evaluation for LUAD early diagnosis biomarkers

In this study, we developed a binary classifier based on the LightGBM gradient boosting algorithm with decision tree base learners to identify robust predictive biomarkers from plasma proteomic profiles. Model training was performed using normalized plasma protein abundances as input features, with sample phenotypes assigned as classification labels. To evaluate the generalizability of the model and minimize overfitting, the modeling cohort was randomly split into a discovery set and an independent validation set at a 6:4 ratio. Specifically, among the 365 samples in the modeling cohort (86 PNs, 101 AAH/AIS, 96 MIA, and 82 IAC), 221 samples were allocated to the discovery set (52 PNs, 61 AAH/AIS, 58 MIA, and 50 IAC) and 144 samples to the internal validation set (34 PNs, 40 AAH/AIS, 38 MIA, and 32 IAC). After model training and internal cross-validation, feature importance scores were generated for individual plasma proteins and used to prioritize candidate biomarkers with the strongest discriminatory power. The finalized model was then evaluated on the blinded internal validation set without any data leakage or post hoc parameter adjustment, ensuring an unbiased assessment of performance and stability. The model objective was set to binary classification, and both L1 (alpha) and L2 (lambda) regularization were enabled to mitigate overfitting. Hyperparameter optimization was performed via grid search nested within a ten-fold cross-validation (CV) loop. To stabilize performance estimates, the entire ten-fold CV procedure was repeated 1,000 times with random fold assignments, and the hyperparameter set that maximized the mean AUC on held-out folds was selected. Random seeds were fixed where applicable to ensure reproducibility. The final tuned hyperparameters are listed in table S1.

With hyperparameters fixed, we applied a two-step feature selection strategy to derive a small, robust biomarker panel. In step one, we performed a rapid pre-selection using LightGBM's built-in feature-importance metric (gain). Importance scores were recorded across 1,000 repeated CV runs and averaged. The top 200 features by mean importance were retained for the next step. In step two, we applied cross-validated permutation feature importance (CVPFI), a model-agnostic permutation-based approach that produces stable importance estimates by averaging fold-specific permutation importances and that can account for feature correlations by grouping and permuting highly correlated features together. CVPFI was computed with 10-fold CV and repeated 100 times to stabilize estimates. Mean CVPFI score across repeats was used to rank features. The ten proteins with the most consistently high CVPFI scores were selected as the final panel.

All feature-selection steps and model evaluations used area under the receiver operating characteristic curve (AUC–ROC) as the performance metric, while CVPFI used average precision as the scoring metric. Hyperparameter tuning was implemented with *GridSearchCV* (*scikit-learn*), the classifier was implemented with the package *LightGBM* v3.2.1 and the CVPFI was implemented with *permutation importance* (*scikit-learn*). All implementations relied on *python* v3.8.0. Code is available upon request.

Model performance was quantified using AUC–ROC throughout hyperparameter tuning, feature selection, and final model evaluation. In addition, CVPFI employed average precision (AP), the area under the precision–recall curve, as its scoring metric to better capture performance on imbalanced classes. The ROC curve plots sensitivity (true positive rate) against 1-specificity (false positive rate) across decision thresholds. AUC–ROC summarizes a model's discriminative ability independent of class prevalence, while AP emphasizes precision at different recall levels and is therefore complementary for imbalanced data.

CRedit authorship contribution statement

Juan Luo: Writing – original draft, Project administration, Methodology, Conceptualization. **Qiongye Dong:** Writing – original draft, Software, Project administration, Funding acquisition. **Jianyi Yang:** Writing – original draft, Investigation, Data curation. **Hui Shan:** Resources, Methodology, Formal analysis. **Qianqian Zhong:** Validation, Methodology, Investigation. **Yongjian Zhang:** Validation, Resources, Investigation. **Guangxi Wang:** Writing – review & editing, Supervision, Project administration. **Jixian Liu:** Writing – review & editing, Supervision, Project administration. **Yuxin Yin:** Writing – review & editing, Supervision, Methodology, Funding acquisition.

Declaration of competing interest

All authors have read the journal's policy on disclosure of potential conflicts of interest. The authors declare no competing conflicts of interest.

Acknowledgements

We extend our gratitude to the patients in the study and their family members. We acknowledge the staff of all departments for their assistance in this study. This work was supported by the Basic Research of Shenzhen Municipal Science and Technology Innovation Commission. The grant numbers are JCY20220818102811024, JCYJ20230807095109019 and JCYJ20250604183730039. All authors have read the journal's authorship agreement, and the manuscript has been reviewed and approved by all named authors.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.trsl.2026.07.013.

References

- Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024;74(3):229–263. May–Jun.
- Siegel RL, Miller KD, Wagle NS, et al. Cancer statistics, 2023. *CA Cancer J Clin.* 2023; 73(1):17–48. Jan.
- Leiter A, Veluswamy RR, Wisnivesky JP. The global burden of lung cancer: current status and future trends. *Nat Rev Clin Oncol.* 2023;20(9):624–639. Sep.
- Inamura K. Clinicopathological characteristics and mutations driving development of early lung adenocarcinoma: tumor initiation and progression. *Int J Mol Sci.* 2018; 19(4). Apr 23.
- Travis WD, Brambilla E, Nicholson AG, et al. The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. *J Thorac Oncol.* 2015;10(9):1243–1260. Sep.
- Miller KD, Nogueira L, Devasia T, et al. Cancer treatment and survivorship statistics, 2022. *CA Cancer J Clin.* 2022;72(5):409–436. Sep.
- Oudkerk M, Liu S, Heuvelmans MA, et al. Lung cancer LDCT screening and mortality reduction - evidence, pitfalls and future perspectives. *Nat Rev Clin Oncol.* 2021;18 (3):135–151. Mar.
- Mazzone PJ, Silvestri GA, Souter LH, et al. Screening for lung cancer: CHEST guideline and expert panel report. *Chest.* 2021;160(5):e427–e494. Nov.
- Aberle DR, Adams AM, Berg CD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med.* 2011;365(5):395–409. Aug 4.
- Horeweg N, van der Aalst CM, Vliegenthart R, et al. Volumetric computed tomography screening for lung cancer: three rounds of the NELSON trial. *Eur Respir J.* 2013;42(6):1659–1667. Dec.
- Li W, Liu J-B, Hou L-K, et al. Liquid biopsy in lung cancer: significance in diagnostics, prediction, and treatment monitoring. *Mol Cancer.* 2022;21(1):25, 2022.
- Ren F, Fei Q, Qiu K, et al. Liquid biopsy techniques and lung cancer: diagnosis, monitoring and evaluation. *J Exp Clin Cancer Res.* 2024;43(1):96. Apr 1.
- Deng Z, Wu S, Wang Y, et al. Circulating tumor cell isolation for cancer diagnosis and prognosis. *EBioMedicine.* 2022;83, 104237. Sep.
- Reina C, Sabanovic B, Lazzari C, et al. Unlocking the future of cancer diagnosis - promises and challenges of ctDNA-based liquid biopsies in non-small cell lung cancer. *Transl Res.* 2024;272:41–53. Oct.
- Wang G, Qiu M, Xing X, et al. Lung cancer scRNA-seq and lipidomics reveal aberrant lipid metabolism for early-stage diagnosis. *Sci Transl Med.* 2022;14(630), eabk2756. Feb 2.

16. Cohen JD, Li L, Wang Y, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*. 2018;359(6378):926–930. Feb 23.
17. Liu C, Xiang X, Han S, et al. Blood-based liquid biopsy: insights into early detection and clinical management of lung cancer. *Cancer Lett*. 2022;524:91–102. Jan 1.
18. Li S, Xin K, Pan S, et al. Blood-based liquid biopsy: insights into early detection, prediction, and treatment monitoring of bladder cancer. *Cell Mol Biol Lett*. 2023;28(1):28. Apr 4.
19. Hanash SM, Pitteri SJ, Faca VM. Mining the plasma proteome for cancer biomarkers. *Nature*. 2008;452(7187):571–579. Apr 3.
20. Karayel O, Virreira Winter S, Padmanabhan S, et al. Proteome profiling of cerebrospinal fluid reveals biomarker candidates for Parkinson's disease. *Cell Rep Med*. 2022;3(6), 100661. Jun 21.
21. Li Y, Wang B, Ma F, et al. Proteomic characterization of the colorectal cancer response to chemoradiation and targeted therapies reveals potential therapeutic strategies. *Cell Rep Med*. 2023;4(12), 101311. Dec 19.
22. Niu L, Geyer PE, Wewer Albrechtsen NJ, et al. Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease. *Mol Syst Biol*. 2019;15(3), e8793. Mar 1.
23. Goncalves E, Poulos RC, Cai Z, et al. Pan-cancer proteomic map of 949 human cell lines. *Cancer Cell*. 2022;40(8). Aug 8835–849 e8.
24. Xing X, Cai L, Ouyang J, et al. Proteomics-driven noninvasive screening of circulating serum protein panels for the early diagnosis of hepatocellular carcinoma. *Nat Commun*. 2023;14(1):8392. Dec 18.
25. Li Y, Wang B, Yang W, et al. Longitudinal plasma proteome profiling reveals the diversity of biomarkers for diagnosis and cetuximab therapy response of colorectal cancer. *Nat Commun*. 2024;15(1):980. Feb 1.
26. Ke G, Meng Q, Finley T, et al. Lightgbm: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;2017:30.
27. Climente-Gonzalez H, Oh M, Chajewska U, et al. Interpretable machine learning leverages proteomics to improve cardiovascular disease risk prediction and biomarker identification. *Commun Med (Lond)*. 2025;5(1):170. May 19.
28. Niu L, Thiele M, Geyer PE, et al. Noninvasive proteomic biomarkers for alcohol-related liver disease. *Nat Med*. 2022;28(6):1277–1287. Jun.
29. Dantas E, Murthy A, Ahmed T, et al. TIMP1 is an early biomarker for detection and prognosis of lung cancer. *Clin Transl Med*. 2023;13(10), e1391. Oct.
30. Gemoll T, Epping F, Heinrich L, et al. Increased cathepsin D protein expression is a biomarker for osteosarcomas, pulmonary metastases and other bone malignancies. *Oncotarget*. 2015;6(18):16517–16526. Jun 30.
31. Yanagawa N, Sugai M, Shikanai S, et al. High expression of fibroblast-activating protein is a prognostic marker in non-small cell lung carcinoma. *Thorac Cancer*. 2022;13(16):2377–2384. Aug.
32. Deng H, Gao Y, Trappetti V, et al. Targeting lactate dehydrogenase B-dependent mitochondrial metabolism affects tumor initiating cells and inhibits tumorigenesis of non-small cell lung cancer by inducing mtDNA damage. *Cell Mol Life Sci*. 2022;79(8):445. Jul 25.
33. Ge H, Malsiu F, Gao Y, et al. Inhibition of LDHB suppresses the metastatic potential of lung cancer by reducing mitochondrial GSH catabolism. *Cancer Lett*. 2024;611, 217353. Nov 28.
34. Chen Z, Lin J, Wang H, et al. Expression and clinical role of PRDX6 in lung adenocarcinoma. *J Int Med Res*. 2024;52(3). Mar3000605241236276.
35. Hu Q, Tian T, Leng Y, et al. The O-glycosylating enzyme GALNT2 acts as an oncogenic driver in non-small cell lung cancer. *Cell Mol Biol Lett*. 2022;27(1):71. Sep 4.
36. Zhou S, Huang H, Zheng Z, et al. MOGS promotes stemness acquisition and invasion via enhancing NOTCH1-glycosylation dependent NOTCH pathway in colorectal cancer. *Am J Cancer Res*. 2023;13(12):5996–6010.
37. Wang T. S100A6 promotes carcinogenesis of lung adenocarcinoma cells. *Gene*. 2025;965, 149652. Sep 15.
38. Koh YW, Lee SJ, Park SY. Prognostic significance of lactate dehydrogenase B according to histologic type of non-small-cell lung cancer and its association with serum lactate dehydrogenase. *Pathol Res Pract*. 2017;213(9):1134–1138. Sep.
39. Pasello G, Scattolin D, Bonanno L, et al. Secondary prevention and treatment innovation of early stage non-small cell lung cancer: impact on diagnostic-therapeutic pathway from a multidisciplinary perspective. *Cancer Treat Rev*. 2023;116, 102544. May.
40. Choi HK, Mazzone PJ. Lung cancer screening. *Med Clin North Am*. 2022;106(6):1041–1053. Nov.
41. de Koning HJ, van der Aalst CM, de Jong PA, et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. *N Engl J Med*. 2020;382(6):503–513. Feb 6.
42. Jonas DE, Reuland DS, Reddy SM, et al. Screening for lung cancer with low-dose computed tomography: updated evidence report and systematic review for the US Preventive Services Task Force. *JAMA*. 2021;325(10):971–987. Mar 9.
43. Gupte TP, Azizi Z, Kho PF, et al. Plasma proteomic signatures for type 2 diabetes mellitus and related traits in the UK Biobank cohort. *medRxiv*. 2024. Sep 15.
44. You J, Guo Y, Zhang Y, et al. Plasma proteomic profiles predict individual future health risk. *Nat Commun*. 2023;14(1):7817. Nov 28.
45. Eldjarn GH, Ferkingstad E, Lund SH, et al. Large-scale plasma proteomics comparisons through genetics and disease associations. *Nature*. 2023;622(7982):348–358. Oct.
46. Oh HS, Rutledge J, Nachun D, et al. Organ aging signatures in the plasma proteome track health and disease. *Nature*. 2023;624(7990):164–172. Dec.
47. Zhao Y, Xue Q, Wang M, et al. Evolution of mass spectrometry instruments and techniques for blood proteomics. *J Proteome Res*. 2023;22(4):1009–1023. Apr 7.
48. Liang P, Peng M, Tao J, et al. Development of a genome atlas for discriminating benign, preinvasive, and invasive lung nodules. *MedComm (2020)*. 2024;5(8), e644. Aug.
49. Nie M, Yao K, Zhu X, et al. Evolutionary metabolic landscape from preneoplasia to invasive lung adenocarcinoma. *Nat Commun*. 2021;12(1):6479. Nov 10.
50. Jacobs JM, Adkins JN, Qian WJ, et al. Utilizing human blood plasma for proteomic biomarker discovery. *J Proteome Res*. 2005;4(4):1073–1085. Jul-Aug.
51. Geyer PE, Kulak NA, Pichler G, et al. Plasma proteome profiling to assess Human health and disease. *Cell Syst*. 2016;2(3):185–195. Mar 23.
52. Zheng R, Matzinger M, Mayer RL, et al. A high-sensitivity low-nanoflow LC-MS configuration for high-throughput sample-limited proteomics. *Anal Chem*. 2023;95(51):18673–18678. Dec 26.
53. Song E, Gao Y, Wu C, et al. Targeted proteomic assays for quantitation of proteins identified by proteogenomic analysis of ovarian cancer. *Sci Data*. 2017;4, 170091. Jul 19.
54. Liu Q, Ye L, Li SG, et al. A simplified direct on-chip forward or reverse immunoassay for evaluating protein-protein interactions in the serum. *Biotechnol J*. 2023;18(3), e2200427. Mar.
55. Lou R, Cao Y, Li S, et al. Benchmarking commonly used software suites and analysis workflows for DIA proteomics and phosphoproteomics. *Nat Commun*. 2023;14(1):94. Jan 6.