Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multidimensional proteomics and explainable AI feature selection identify cross-platform lung cancer molecular signature in blood plasma

Peter Jianrui Liu

peter92liu@gmail.com

Oxford Cancer Analytics https://orcid.org/0000-0003-3608-750X

Harriet Ferguson

Oxford Cancer Analytics

Nikola Gushterov

Oxford Cancer Analytics

Benedikt Kessler

University of Oxford https://orcid.org/0000-0002-8160-2446

Geoffrey Liu

University Health Network / Princess Margaret Cancer Centre

Andreas Halner

Oxford Cancer Analytics

Luke Hankey

Oxford Cancer Analytics

Iliyana Kaneva

Oxford Cancer Analytics

Mrunmayee Dupalliwar

Oxford Cancer Analytics

Junetha Syed

Oxford Cancer Analytics

Emma Mi

Oxford Cancer Analytics

Ella Mi

Oxford Cancer Analytics

Daniel Szulc

Oxford Cancer Analytics

Devalben Patel

Princess Margaret Cancer Centre

Luna Zhan

University Health Network / Princess Margaret Cancer Centre / University of Toronto

Roman Fischer

University of Oxford https://orcid.org/0000-0002-9715-5951

Article

Keywords:

Posted Date: October 31st, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7660411/v1

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

Additional Declarations: Yes there is potential Competing Interest. H.R.F, N.G, L.H, I.K, M.D, J.S, Emma.M, Ella.M, D.A.S, A.H, and P.J.L are current/former employees, shareholders, and/or share option holders of Oxford Cancer Analytics Ltd. R.F and B.M.K are share option holders of Oxford Cancer Analytics Ltd. Oxford Cancer Analytics Ltd sponsored this study.

Abstract

Lung cancer is the leading cause of cancer mortality worldwide with 70% diagnosed late stage despite low-dose computed tomography (LDCT) screening availability. We combined data-independent acquisition mass-spectrometry (DIA-MS) and proximity extension assay (PEA) with explainable artificial intelligence (XAI)-led machine learning (ML) for plasma-based biomarker discovery. From a 490 lung cancer and 124 matched-control cohort, ML models were trained to predict lung cancer achieving an AUROC of 0.91 [95% CI: 0.88–0.93] and 0.97 [95% CI: 0.92–0.98] in DIA-MS and PEA, respectively. XAI characterised networks of model-consistent features primarily related to infection and inflammatory responses. We then introduced a DNA-aptamer proteomics method and identified a cross-platform concordance panel, with performances of 0.88 [95% CI: 0.80–0.90] and 0.88 [95% CI: 0.81–0.95] in DIA-MS and PEA, respectively. This study demonstrates that combining multi-dimensional proteomics with XAI-ML can characterise robust biomarker signatures.

Introduction

Lung cancer is the leading cause of cancer mortality worldwide accounting for one fifth of total global cancer mortality and two million deaths per annum globally[1]. Over 70% of lung cancer diagnoses are made at late stages, when regional-to-distant disease have a five-year survivorship of 9–37% for non-small cell lung cancer (NSCLC) and 3–18% for small cell lung cancer (SCLC), in the US, respectively[2]. In contrast, early stage, localised lung cancer five-year survivorship is 65% for NSCLC and 30% for SCLC. Therefore, enabling earlier lung cancer diagnosis has been a priority for improving patient survivorship and outcome.

Low dose computed tomography (LDCT) has been widely studied for lung cancer screening in high-risk populations. A number of studies including the NELSON study in the EU and the National Lung Screening Trial (NLSC) in the United States showed a 20% reduction in mortality in the LDCT screening group[3,4]. However, uptake and adoption have widely been reported to be poor ranging from as low as < 5% in the US to up to ~ 50% in the UK. Limiting factors including radiation exposure, cost, and throughput have been cited to prevent LDCT accessibility and uptake, demonstrating challenges for scaling up[5,6]. Further, as smoking rates fall in Western and Asian countries, a greater fraction of lung cancer are occurring in individuals with a light or never-smoking history[7]; these individuals are ineligible for LDCT screening according to current criteria. Therefore, a minimally invasive and scalable approach for identifying high-risk lung cancer individuals with the potential for population level deployment may help improve current lung cancer screening paradigms.

The blood proteome harbours promising information for the identification of a molecular signature for lung cancer, which can be applied as a liquid biopsy blood test. Prior efforts using a cell-free DNA based approach have shown limited sensitivity and specificity especially in earlier stage and low tumour burden disease, likely due to a low level of circulating tumour DNA[8–10]. In contrast, the blood proteome provides important insight including tumour-specific inflammatory and host response even during early and low-

tumour burden disease, serving as a potent repertoire for novel biomarker discovery[11,12]. Quantification of such circulating proteins may create a molecular signature for lung cancer. To this end, past studies have focused on targeted proteomics approaches such as proximity extension assay (PEA) and parallel/multiple reaction monitoring, liquid-chromatography tandem-mass spectrometry (LC-MS) or untargeted approaches including data-dependent/ data-independent LC-MS. However, different platforms may exhibit biased preferences for different proteins and pathways and show incongruence between quantification platforms, limiting biomarker signatures to a specific platform[13].

To enable the identification of a robust blood biomarker signature, the use of reliable machine learning (ML) models that generalise to unseen data and can model the interactions between the features is critical[14]. In particular it is crucial to aggregate the information and importance of biomarkers across many different models to identify a robust panel that functions regardless of underlying models or model parameters. In addition, the black-box nature of more advanced ML models has been an ongoing concern, especially when clinical guidelines and regulatory approval processes benefit from understanding the rationale behind results.

In this present study, we deployed multidimensional-proteomics combined with an explainable artificial intelligence (XAI)-guided ML approach on a large lung cancer cohort to identify a robust plasma-based molecular signature across diverse platforms. This combined data-independent acquisition mass spectrometry (DIA-MS) as an unbiased proteomics approach with a targeted PEA (Olink) for biomarker discovery from a cohort of 614 individuals (490 lung cancer cases across all stages, histologies, smoking exposures; and 124 age-, sex-matched controls from a research Canadian screening cohort). Using an 80/20 train/holdout strategy with cross-validated Shapley values for interpretability, we identified high-importance model-consistent features and uncovered key feature interactions relevant to lung cancer prediction. A third proteomics approach, targeted DNA-aptamer (SomaScan), was deployed on an expanded subset of patients to characterise a focused cross-platform biomarker signature. This multidimensional proteomics and XAI-driven methodology addresses core challenges in liquid biopsy development, including black-box model transparency, feature generalisability, and platform-specific constraints. We report the first cross-platform plasma-based lung cancer molecular signature identified through integrated multi-platform proteomics and explainable AI.

Results

Synergy between DIA-MS and PEA proteomics approaches enhances proteome coverage

Single-shot DIA-MS and PEA were used for protein quantification of plasma samples from 490 post-diagnosis, treatment-naive lung cancer cases covering all stages and major histologies, and 124 age-/sex-matched cancer-free controls (n = 124) recruited through a lung screening programme, as

described previously[15,16] (Fig. 1a, Extended Data Fig. 1a). Clinical and demographic characteristics are summarised in Table 1.

Table 1
Clinical and demographic characteristics of healthy control vs lung cancer cohort

	Control (n = 124)	**NSCLC (n = 433)**	**SCLC (n = 55)**	**Rare (n = 2)**
Number of females, *n (%)*	63 (50.8)	219 (50.6)	21 (38.2)	1 (50)
Mean age in years, *(sd)*	69 (10)	68 (11)	68 (10)	57 (13)
Age range (10 year intervals), *n (%)*:				
< 50	0 (0)	23 (5.3)	2 (3.6)	1 (50)
50–60	27 (21.8)	76 (17.6)	12 (21.8)	0 (0)
60–70	30 (24.2)	130 (30)	16 (29.1)	1 (50)
70–80	54 (43.5)	148 (34.2)	15 (27.3)	0 (0)
80–90	13 (10.5)	49 (11.3)	10 (18.2)	0 (0)
> 90	0 (0)	7 (1.6)	0 (0)	0 (0)
Ethnicity, *n (%)*:				
Asian/Pacific Islander	5 (4)	66 (15.2)	3 (5.5)	0 (0)
Black/African-Canadian	0 (0)	11 (2.5)	0 (0)	0 (0)
First Nations	0 (0)	2 (0.5)	0 (0)	0 (0)
Latino/Hispanics	0 (0)	2 (0.5)	1 (1.8)	0 (0)
Mixed	0 (0)	4 (0.9)	0 (0)	0 (0)
White/Caucasian	118 (95.2)	268 (61.9)	45 (81.8)	2 (100)
Unknown	1 (0.8)	80 (18.5)	6 (10.9)	0 (0)
Smoking history:				
Mean pack-years *(sd)*	40 (21)	32 (31)	50 (29)	27 (11)
Heavy (20 py)	107 (86.3)	252 (58.2)	49 (89.1)	1 (50)
Non-heavy (< 20 py)	17 (13.7)	175 (40.4)	5 (9.1)	1 (50)
Unknown	0 (0)	6 (1.4)	1 (1.8)	0 (0)
Smoking status at diagnosis, *n (%)*:				
Current smoker	43 (34.7)	128 (29.6)	32 (58.2)	1 (50)
Ex-smoker	81 (65.3)	150 (34.6)	16 (29.1)	1 (50)
Ever smoker, NOS	0 (0)	1 (0.2)	0 (0)	0 (0)

	Control (n = 124)	NSCLC (n = 433)	SCLC (n = 55)	Rare (n = 2)
Light ex-smoker	0 (0)	11 (2.5)	0 (0)	0 (0)
Never smoker	0 (0)	119 (27.5)	2 (3.6)	0 (0)
Stage of cancer, *n (%)*:				
Stage 0	NA	1 (0.2)	0 (0)	0 (0)
Stage 1	NA	139 (32.1)	4 (7.3)	1 (50)
Stage 2	NA	44 (10.2)	2 (3.6)	0 (0)
Stage 3	NA	86 (19.9)	17 (30.9)	1 (50)
Stage 4	NA	161 (37.2)	32 (58.2)	0 (0)
Unknown	NA	2 (0.5)	0 (0)	0 (0)
Clinical morphology, *n (%)*:				
LUAD	NA	251 (58)	NA	2 (100)
LUSC	NA	107 (24.7)	NA	0 (0)
Large cell carcinoma	NA	32 (7.4)	NA	0 (0)
Other	NA	43 (9.9)	55 (100)	0 (0)
Molecular alteration, *n (%)*:				
EGFR-	NA	4 (0.9)	0 (0)	0 (0)
EGFR+	NA	118 (27.3)	0 (0)	0 (0)
EGFR Unknown	NA	3 (0.7)	0 (0)	0 (0)
EGFR NR	NA	308 (71.1)	55 (100)	2 (100)
ALK+	NA	1 (0.2)	0 (0)	0 (0)
ALK-	NA	114 (26.3)	0 (0)	0 (0)
ALK NR	NA	308 (71.1)	55 (100)	2 (100)
ALK Unknown	NA	10 (2.3)	0 (0)	0 (0)
KRAS+	NA	1 (0.2)	0 (0)	0 (0)
TP53+	NA	1 (0.2)	0 (0)	0 (0)

NSCLC, non-small cell lung cancer; SCLC, small cell lung cancer; n, number of; sd, standard deviation; py, pack-years; NOS, not otherwise specified; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma +, positive for mutation; -, negative for mutation; NR, not reported; NA, not applicable.

DIA-MS quantified a wider-range of proteins (n = 3,655) compared with PEA (n = 2,884), with 1,136 quantified by both (Fig. 1b, Supplementary Fig. 1a). Given the high dynamic range of plasma protein concentrations, we assessed available concentrations as reported in the Human Protein Atlas (HPA)[17]. DIA-MS quantified 1,736 mid-low abundance proteins (< 0.1mg/L), while PEA quantified 1,486 (Fig. 1c). Data missingness was higher in DIA-MS (37.07% ± 34.82%) compared with PEA (13.09% ± 24.33%). Both datasets showed a small but statistically significant (FDR adjusted $P < 0.05$) negative Spearman correlation between protein concentration and missingness, suggesting poorer quantification at lower abundance (Fig. 1d). Assessment of secretome location in the HPA showed most proteins were not known to be secreted, with secreted proteins predominantly secreted to the blood across the abundance range (Extended Data Fig. 1b).

Matrix-matched technical replicates were acquired by PEA alongside clinical samples (n = 14), while for DIA-MS a pool of clinical samples was evaluated (n = 42). PEA had lower CVs than DIA-MS with a median CV of 21.4% versus 34.1%, although high CV proteins were not consistent across platforms (Fig. 1e). Analysis of Spearman correlation across protein abundance range indicated higher cross-platform variability for low-abundance proteins (Fig. 1f).

Implementing dual protein quantification platforms expanded information for biomarker discovery (Fig. 1b). Over-representation analysis (ORA) showed over-representation of proteins associated with cell migration, signalling, haemostasis and immune response in PEA. Also enriched by DIA-MS were cytoskeleton- and chromatin organisation-associated proteins (Fig. 1g). Together, these results show that combining DIA-MS and PEA increases the pool of potential biomarker candidates.

Univariate analysis identified lung cancer-associated plasma proteins and sub-cohort dependent variation

Principal component analysis (PCA) showed a gradual stage-dependent separation of lung cancer cases from cancer-free controls (Fig. 2a,b, Supplementary Fig. 2). In both datasets, early-stage lung cancer cases (Stage I-II), were closer to cancer-free controls compared with late stage (Stage III-IV). PCA loadings showed that separation was distributed across many features (Fig. 2c,d). SERPINA3, PI16 and GSN contributed to separation in both datasets.

Analysis of differentially abundant proteins between cancer-free controls and lung cancer cases showed a larger level of significant downregulation (FDR-adjusted $P < 0.05$, One-way ANOVA, with Log_2 foldchange (Log_2FC) > 1) (Fig. 2e,f). Differential abundance was also evaluated on sub-cohorts including early and late lung cancer, lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), small cell lung cancer (SCLC) and male-/female-only comparisons. Minimal difference in differential abundance was observed across sub-cohorts, with the exception of LUAD/LUSC in both datasets, as well as early lung cancer in PEA (Extended Data Fig. 2a,b).

Gene set enrichment analysis (GSEA) showed the direction of regulation was similar across platforms, although little overlap was observed in significantly regulated pathways (FDR-adjusted $P < 0.05$) (Fig. 2g, Extended Data Fig. 2c), suggesting low concordance/overlap between approaches could generate different interpretations. Regulation of actin cytoskeleton organisation, protein complex assembly, and hydrogen peroxide metabolism associated proteins was observed in DIA-MS, irrespective of sub-cohort. In PEA, pathways associated with coagulation and GTPase/RAS signalling, were significantly downregulated across most sub-cohorts, while RNA metabolism, immune response and chromatin organisation pathways were significantly upregulated in SCLC. Overlapping pathways were related to acute-phase response and actin cytoskeleton organisation, with differences among subtypes largely derived from smaller populations (SCLC and LUSC). Overall, univariate analysis of plasma proteins revealed limited insight into biological processes, particularly in the more heterogeneous sub-populations.

XAI-informed multivariate feature selection to build panels of lung cancer plasma biomarker candidates

We next trained 75 tree-based ML models using combinations of different model architectures, feature de-selection, data augmentation, feature selection, multi-objective optimisation and XAI techniques to identify biomarkers that were consistently predictive of lung cancer across multiple models. All data underwent identical sample stratification, train/holdout splitting with separate pre-processing, minimising data-leakage to unseen holdout data to promote generalisation. All models were evaluated across nine lung cancer sub-cohorts for both datasets, giving rise to 1,350 ML models (Fig. 3a, Extended Data Fig. 3a, Supplementary Table 1).

PEA models had better performance than DIA-MS models; all and late lung cancer, NSCLC, LUSC and LUAD and male-only PEA models achieved better performance than other sub-cohorts, as assessed by area under the receiver operator characteristic curve (AUROC) and Matthews Correlation Coefficient (MCC) (Fig. 3b-e). In DIA-MS models NSCLC performance was higher than SCLC. Male-only models outperformed female-only models in both datasets. Female-only lung cancer cases had a lower incidence of LUSC and SCLC compared with males and higher incidence of LUAD, with comparable age distribution (Supplementary Fig. 3a, b). Interestingly, assessment of an all-lung cancer trained-model predicting sub-cohorts showed similar decreased performance in female-only versus male-only cohort (Extended Data Fig. 3b).

To identify robust feature sets, we next assessed Shapley value feature importance across all lung cancer models[18]. Some high-importance features were consistently selected for a model (WFDC2, CILP, NCF2, MMP7, GLRX) while some high-importance features showed strong model-dependency (SH3BP2, VWA8, KIF20B, CLEC4G and BCAT1) (Fig. 3f). Upon assessment of individual LC models, MB, a feature selected by DIA-MS and PEA, showed an early contribution during recursive feature addition (RFA) for both approaches (Fig. 3g), also showcasing how using Shapley values during RFA promoted incremental

performance increases. Smaller performance increases suggested incorporation of redundancy during feature selection, while minor AUROC changes accompanied by larger changes in MCC highlighted insights from considering two performance metrics.

Univariate analysis is often used to select features for ML[19]. During feature de-selection, features that had low information about the target (mutual information (MI) < 3.5%) were removed. The majority of differentially abundant proteins had MI > 3.5%, however a high number of proteins with high MI were not differentially abundant (Extended Data Fig. 3c,d). Moreover, a large number of high-importance features did not pass the threshold for differential abundance, particularly for DIA-MS models, while some high importance features did not have MI > 3.5%, emphasising the utility in using multiple feature selection-configurations. A direct comparison of univariate with the XAI-informed feature selection approach used here showed that despite containing non-differentially abundant proteins, overall ML-driven feature selection achieved higher performance even with a simple logistic regression model (Extended Data Fig. 3e).

Characteristics of features selected

The Rashomon effect describes the phenomena in ML where similar performances can be achieved with different internal workings of modell[20] We addressed this by comparing feature selection across many different models.

To evaluate feature selection, we considered models that generalised, determined by comparing holdout performance to 95% CI from training data (Extended Data Fig. 3a). Five features were selected in generalisable PEA and DIA-MS models (LRG1, SDC1, MB, LTA4H and FGL1) (Fig. 4a). LTA4H was the only common feature regulated differently across approaches, with minor upregulation (significant in SCLC) in DIA-MS and significant downregulation across PEA sub-cohorts ($P < 0.05$, Log2FC > 1). Other proteins quantified by both methodologies were only selected as features for DIA-MS (n = 25) or PEA (n = 13), albeit mostly regulated similarly across approaches. Like LTA4H, some showed divergent abundance (THBS2, THOP1 and ADAMTSL4) while others showed strong regulation in one dataset with minimal regulation by the alternative method (MUC16, CEACAM5, CILP, HAGH and LXN).

Across generalisable models, high importance features according to Shapley values were similar for all/late lung cancer, NSCLC and LUAD. Variation across sub-cohorts was generally lower for PEA. SCLC and sex-split sub-cohorts diverged most from other models. In female-only DIA-MS models, high importance features for many cohorts (MB, WFDC2, HBA2, TAGLN2 and PPBP) were less informative while other candidates (POSTN, VWA8, HAGH, NAGLU and MUC1) were never selected. Similarly for male-only models in DIA-MS, while the overall order of importance was preserved, many features were never selected. In early lung cancer models, SDC1, POSTN, PPBP and SERPINA3 were less important/absent compared with all/late lung cancer models, suggesting their presence arises from influence of later stage lung cancer. Higher importance features had low missingness in DIA-MS, whereas in PEA two candidates (LTA4H and NEXN) were largely quantified below the LLOD (Extended

Data Fig. 4a-b). High-importance lung cancer features were lower in abundance in PEA models (Extended Data Fig. 4c).

Biological pathways enriched in lung cancer features showed that proteins involved in chemotaxis, cell adhesion, wound healing and immune response (acute-phase, complement activation and humoral immune response) were present at varying levels across sub-cohorts in both datasets (Fig. 4c). This challenged previous GSEA showing immune response pathways enrichment in PEA only (Fig. 2g). Both datasets relied on features associated with cell migration, while in PEA features were enriched for receptor-mediated endocytosis, integrin and cytokine signalling. DIA-MS was also enriched for oxidative stress features (Fig. 4c). In both datasets, if known, secretion to blood was more prevalent (Extended Data 4d).

XAI-driven feature selection captures strong interactions between features.

To further understand how models used features, we investigated how Shapley values individually and through interactions explained differences between cancer-free controls and sub-cohorts.

Feature quantification from a lung cancer model separated cancer-free control samples from early and late lung cancer samples by t-SNE (Extended Data Fig. 4e). Shapley values better separated cancer-free controls from lung cancer samples when compared to protein quantification, with early/late differences being less obvious (Extended Data Fig. 4f). This suggested high-importance lung cancer features could be used regardless of stage. This is further supported by the high performance of lung cancer models predicting early lung cancer (AUROC = 0.89 [95% CI: 0.79–0.98], MCC = 0.65 [95% CI: 0.46–0.82]) (Extended Data Fig. 3b) compared with the performance across early lung cancer models (AUROC = 0.82 [95% CI: 0.77–0.88], MCC = 0.48 [95% CI: 0.34–0.64]) in DIA-MS (Fig. 3b,d).

Analysis of Shapley value model-decision paths showed individual feature contribution towards a prediction (Extended Data Fig. 6a,b), while interactions demonstrated how features are mediated by each other. Shapley value interactions from a lung cancer model identified complex networks of feature interactions (Fig. 4d,e, Extended Data Fig. 6c,d). For DIA-MS, this network contained a higher level of interactions compared with PEA. Shapley value importance scores were comparatively lower and more evenly distributed within the PEA network. For DIA-MS, highly important features were involved in the strongest interactions (MB, WFDC2, HBA2, CILP and SDC1). Increased levels of CILP, a universal feature for all DIA-MS models, interacted with decreased levels of SDC1, SELENBP1 and/or WFDC2 to be predictive of cancer-free controls (Extended Data Fig. 6e). Similarly, decreased levels of SDC1 interacted with increased levels of MB and decreased levels of CEACAM5 and/or MMP7 to be predictive of cancer-free controls in the analysis of top PEA interactions. (Extended Data Fig. 6f). Several interactions (TAGLN2-MUC1, DDTL-ACTB, CASP8-CLIP2 and NCF2-CLIP) identified two groups of controls. These interactions emphasise the importance of using models capable of capturing feature interactions for disease identification.

The identification of a platform-agnostic lung cancer plasma protein signature.

We next introduced a third DNA-aptamer-based proteomic approach (DNA-APT) to further evaluate concordance across platforms. A panel of 100 SomaScan quantified proteins selected from previously identified features (Fig. 5a, Extended Data Fig. 7a,b) was applied to a cohort of 140 patients selected from the previously described cohort (Table 1), 72 cancer-free controls and 68 lung cancer cases.

Variability was low in DNA-APT assay and dimensionality reduction showed separability between cancer-free controls, early and late lung cancer samples (Extended Data Fig. 7c).

The majority of proteins quantified in all three methods (n = 27, 57.1%) had a significant Spearman correlation (FDR-adjusted $P < 0.05$) (Fig. 5b, Extended Data Fig. 7d). Six features showed no concordance with the DNA-APT approach. A group of predominantly PEA features (n = 16) had a strong, significant correlation between PEA/DNA-APT only. Similarly, a group of DIA-MS features (n = 16) had strong correlation between DIA-MS/DNA-APT only. High-importance feature SERPINA3 had opposite correlation in DIA-MS/DNA-APT vs PEA/DNA-APT. Overall, concordance between DNA-APT and PEA was higher than concordance with DIA-MS, with median Spearman *R* of 0.567 versus 0.363 (Fig. 5c).

A panel of proteins with Spearman $R > 0.4$ across all three panels were evaluated (n = 18) and a second panel that contained proteins with Spearman $R > 0.4$ (n = 25) between DIA-MS/DNA-APT or PEA/DNA-APT data were evaluated using DIA-MS and PEA (see methods). Similar performances were seen between DIA-MS and PEA for the three-platform panel, with AUROC of 0.88 [95% CI: 0.80–0.95] and 0.88 [95% CI: 0.81-0. 95], and MCC of 0.49 [95% CI: 0.31–0.67] and 0.60 [95% CI: 0.44–0.77] for DIA-MS and PEA, respectively (Fig. 5d). The PEA/DNA-APT two-platform panel achieved the highest performance, with AUROC of 0.97 [95% CI: 0.94–0.99] and MCC of 0.71 [95% CI: 0.54–0.86]. The DIA-MS/DNA-APT model outperformed the three-platform panel with AUROC of 0.93 [95% CI: 0.86–0.97] and MCC of 0.60 [95% CI: 0.44–0.78].

To evaluate the DNA-APT assay, a partial least squares (PLS) regression was used to show differences between cancer-free controls and lung cancer, which were similarly distinguishable across all platforms (Fig. 5e). Analysis of loadings showed that similar features were responsible for separation on all platforms, separating cancer-free controls (MB, PF4, PPBP, CNTN3 and KIT) from lung cancer (HAGH, ALDH1A1, FGL1, MMP9, C9, CFHR5 and LBP). The same was carried out for the two-platform panels which achieved similar separation by PLS regression (Extended Data Fig. 7e). We show the identification of feature panels compatible across multiple proteomics platforms, as demonstrated by a similar degree and driver(s) of separation between the cross-platform panels.

Discussion

This study deployed unbiased DIA-MS and targeted PEA proteomics as discovery approaches in combination with XAI-guided ML feature selection. On a cohort of 614 patients including 490 lung cancer cases (stages I-IV) covering all major histologies and 124 age-, sex-matched controls we evaluated plasma molecular signatures for lung cancer across 1,350 ML models. Further characterisation using a DNA-aptamer approach identified an eighteen-feature lung cancer plasma

signature with concordance across all platforms, achieving AUROC of 88.2% and 88.3% in DIA-MS and PEA holdout validation, respectively. Signatures with high concordance across two of the three panels achieved even higher AUROC between 92.4% to 97%. PLS regression on the DNA-APT validation dataset clearly distinguished lung cancer from controls.

The proteomics approaches deployed are among the most widely used, DIA-MS offering unbiased discovery, while PEA and DNA-APT are often favoured in clinical research for straightforward data processing[21]. LC-MS based proteomics is increasingly considered for clinical purposes, particularly with regards to cancer detection and disease monitoring due to high-throughput and multiplexing[22]. This study leveraged both the overlap and divergence between platforms to expand coverage and to assess feature robustness. In-line with previous studies, we observed platform-specific discrepancies[21,23,24], potentially arising from either DIA-MS pre-processing or immunoassay epitope-specific biases, reinforcing concerns about the reliability of signatures developed from a single method. Dynamic range is often considered a limitation of DIA-MS. Low abundance proteins showed greater variability and reduced concordance (Extended Data Fig. 4c), however models from all platforms relied on such features to varying degrees and yielded predictive signatures. By focusing on high-importance features with strong cross-platform concordance, we identified a robust plasma signature for lung cancer, supporting the development of transferable, platform-agnostic biomarker panels, facilitating scalable and flexible clinical implementation.

Despite differences across platforms, cytokine/chemokine-mediated signalling were associated with both DIA-MS and PEA biomarker panels, with several proteins previously considered for lung cancer detection, monitoring or prognosis. Many candidates have also been previously associated with cancer pathology but would be considered novel biomarkers of lung cancer prediction (Extended Data Fig. 5a, b). Five proteins were selected across both proteomics platforms (MB, SDC1, LRG1, FGL1 and LTA4H). MB had high importance/concordance across both platforms and had previously been associated with skeletal-muscle wasting in cancer-related cachexia[25]. We observed that low plasma MB coupled with regulation of other candidates (SDC1 and COTL1) were important in lung cancer prediction. SDC1 had high importance across both datasets, with a known role in lung cancer pathology and as a prognostic marker in multiple solid-tumours[26]. We observed significant upregulation of SDC1 in all but early lung cancer sub-cohorts in both datasets, indicating increased circulating SDC1 with stage. To the best of our knowledge, MB and SDC1 have not previously been reported as being predictive of lung cancer. We also characterised high importance lung cancer features associated with immune/inflammatory responses. SERPINA3 and CRP, associated with acute-phase response, have previously been proposed as biomarkers for early NSCLC. We, however, observed SERPINA3 upregulation to be associated with later-stage NSCLC, LUSC/male-only lung cancer. There were however early-stage associated proteins characterised by both approaches. In PEA, the transferrin receptor (TFRC) was important for early lung cancer among other sub-cohorts (Fig. 4b). TFRC upregulation is known in NSCLC and other solid-tumours, also expressed on the surface of activated immune cells[27], potentially with a dual-role in mediating lung cancer progression[28]. Similarly, HAGH showed upregulation in early-stage lung cancer (*P*

< 0.05 in DIA-MS only) (Fig. 4a, Fig. 5e), the only candidate to show an early-dominant response across both platforms, suggesting HAGH upregulation in plasma may occur early and reduce over disease progression. This is in line with HPA data showing HAGH has favourable prognostic value in other cancers including liver, renal and pancreatic cancer[29]. Pathway analysis showed that proteins related to oxidative stress and wound healing were informative, even in early-stage disease. Together, this underscores the potential of protein-based approaches to detect early-stage and low-tumour burden diseases by identifying amplified inflammatory signals, addressing concerns about limited ctDNA signal in such cases[8–10]. Finally, we showed that despite having used cohort-specific models for biomarker selection, it appears that models trained on the complete cohort were similar/superior at understanding the subclassification task, possibly due to additional context from other samples. For example, predictions of early cancer patients in unseen holdout data were improved when late cancer patients were included in training. Future work focused on early detection of cancer may consider the inclusion of late-stage samples to improve separability.

Model performances and feature selection varied between male and female-only models indicative of sex-based variation in plasma molecular signatures (Fig. 3b,c). Male-only candidates had a high association with cell migration and apoptotic pathways in line with other cohorts while some more universal features were not selected in female-only models (LRG1, SELENBP1, KIF20B, SERPINA3, MUC1, SAA1, MUC16, KRT19, CIT and TFRC) (Fig. 4a-c). Moreover, many biological pathways associated with lung cancer were not major candidates in female-only models, particularly immune response-related pathways, aligning with reports at the transcript level[30]. This is consistent with previous studies showing similar differences between sex-separated molecular signatures, proposing the use of different features and models for prediction of sex-based sub-cohorts[30,31]. Although we found lung cancer trained-models predicted female-only sub-cohorts to similar levels as female-only models, further assessment of sex-based differences in plasma molecular signatures and other confounding data in a larger cohort could further characterise the differences. This highlights the potential utility of evaluating sex-based differences in prediction using plasma molecular signatures, particularly in early detection where changes in protein abundance can be small in magnitude.

The strengths in this study lie in the application of a multi-dimensional proteomics approach to a large cohort of lung cancer cases (n = 490) covering all major histologies with a good representation of early-stage cases. Additionally, we applied rigorous XAI and ML methodologies, with separate train/holdout processing where possible, using two complementary metrics, AUROC and MCC and exploring both boosting (XGBoost) and bagging (BRF) algorithms. XGBoost was prioritised to achieve a more concise set of informative features, giving rise to smaller panels of ~ 20 proteins. We also overcame poor model interpretability by using TreeSHAP approximation to efficiently compute cross-validated Shapley values across a large feature set. The resulting stable feature set is expected to maintain strong performance even in simple models. Finally, combining two distinct proteomics approaches for lung cancer circulating biomarker discovery broadened potential candidate coverage and combining three independent protein

detection platforms allowed the assessment of a concordant lung cancer molecular signature across platforms.

A limitation of this study is that, with highly correlated features, Shapley values can distribute across redundant features, diluting the contribution of individual biomarkers. This is a common limitation in multivariate feature selection methods. Similarly, we focused on XGBoost models which can overfit when not carefully designed, and we mitigated these potential limitations by applying unseen-holdout generalisation constraints, regularisation and evaluating multiple models. Another limitation was the over-representation of lung cancer and rarer subtypes. Although this enabled sufficient power to provide reliable insight into subtype specific biomarkers, a more-representative cohort can assess the utility of this ML-approach and feature panels in populations more representative of epidemiological distributions. Similarly, assessment of real-world efficacy and health economic feasibility of plasma molecular signatures were not within the scope of this study. Future work seeks to further evaluate and validate our approach with prospective population level-based detection studies.

In conclusion, we deployed a multidimensional proteomics and XAI-guided ML biomarker discovery approach in a lung cancer cohort composed of early stage, late stage, and all common histologies with age-, sex-matched controls from a screening program cohort. The combined synergy between different proteomics approaches and interpretable ML models identified a cross-platform lung cancer plasma protein signature with explainable feature ranking, contribution, and interaction mechanisms for decision making. We anticipate this signature to be highly versatile, transferable, and scalable across different protein quantification platforms and look forward to future assessment of its cost effectiveness and real-world impact on patient outcomes. Given promising results of sub-cohort analysis demonstrating the ability to detect specific lung cancer histologies and stages, customisable panels may be deployed in diverse clinical use cases including at-risk population risk stratification, screening, and disease classification.

Materials and methods

Cohort selection and collection

Lung cancer cases and healthy controls were recruited from Princess Margaret Cancer Centre's Lung-CALIBRE program (Lung cancer Clinical And Liquid biopsy, Implementation, Breathomics, Radiomics for Early detection). Ethics approval for the CALIBRE program was obtained from The University Health Network Research Ethics Board (REB 06-639). Lung cancer cases were selected using a stage-stratified random process from a biospecimen database of lung cancer patients to sufficiently cover all major histologies and stages. Case participants included in this analysis were recruited sequentially from 2008–2019 which met the conditions of age over 18 years; histological confirmation of lung cancer, adequate plasma specimen (1.8ml vial) taken after initial diagnosis prior to commencing first treatment.

Controls were distribution-matched to age and gender of lung cancer cases from the research lung cancer screening program. This program used a broad eligibility of 10-pack year minimum cigarette smoking history and a minimum age of 45 years old.

Healthy control samples were collected at Princess Margaret Cancer Centre and lung cancer cases collected from Toronto General Hospital and Princess Margaret Hospital. All blood samples used in this study were collected in EDTA tubes and processed by the same operator to plasma using centrifugation within two hours of collection. All processing was carried out at ≤ 4°C and samples were stored at -80°C at Princess Margaret Cancer Centre for further use.

To achieve > 80% power for showing a minimum lung cancer detection sensitivity and specificity of 70% and 90% respectively (compared to a null hypothesis of 50% each, with alpha = 0.05), an overall sample size of 490 cancers and 124 controls, with a 80/20% train/holdout split was selected to satisfy the requirement for a minimum number of holdout sample size of 49 cancers and 12 controls[32].

Sample preparation and processing for DIA-MS

Sample randomisation

The order of the 614 clinical samples to be processed were randomised based on demographic and clinical characteristics. The randomised samples were processed in seven individual 96-well plates. Each plate contained a patient's plasma samples and six “pool plasma” which were later used for quality control. The pool sample was formed by pooling 3 µL of clinical plasma samples of mixed types.

Top 14 most abundant protein depletion

Plasma samples stored at -80°C were thawed at RT. Five µL of neat plasma samples were used for each depletion. Top 14 depletion was performed for each sample using high-selection top 14 depletion resin (A36372, Thermo Scientific) following manufacturer’s instructions. Depleting the plasma high-abundance proteins improves the dynamic range in LC-MS proteomics analysis. Approximately 100 µL of depleted plasma per well were collected for each sample; 25 µL of depleted plasma from each sample were transferred to a fresh 96-well PCR plate (LoBind, Eppendorf) for digestion.

Haemolysis was evaluated as described previously[33] on the unprocessed plasma sample.

SP3 in solution digestion

5 µL of 60 mM TCEP (XH347846, Thermo Scientific), 240 mM CAA (CO267-100G, Merck) mix were added to each sample in the well of a PCR 96-well plate. The plate was sealed with a heat sealer (E0030127838, Eppendorf) on a heat seal (E5391EN100431, Eppendorf). Reduction and alkylation were performed by incubating the samples at 95°C for 15 min on a Thermomixer.

Single pot, solid-phase enhanced sample preparation protocol (SP3) was used to prepare peptide samples. Briefly, Sera-Mag™ Carboxylate-Modified Magnetic Beads (45152105050250 and 65152105050250, Cytiva) were washed twice in HPLC grade water and 2 µL added to each sample along with 80 µL 99.8% EtOH (E/0650/17, Fisher chemical) and incubated at RT for 5 mins. The beads were subsequently washed with 80% EtOH three times. 1 µg MS grade Pierce™ trypsin protease (90058, Thermo Scientific) in 100 mM ammonium bicarbonate (09831-500G, Merck) was added to each sample and incubated overnight at 37°C. Digested samples were transferred to a new plate, lyophilised and stored at -20°C ready for data acquisition.

Sample loading

Dried samples were resuspended in 100 µL 0.1% FA (85170, Thermo Scientific). For analysis on the Evosep One LC system, the sample was loaded onto Evotips (EV2011, Evosep) according to Evotip sample loading protocol. Briefly, the Evotips were rinsed with 20 µL solvent B (85174, Thermo Scientific), conditioned with isopropanol (99.5%, 383920025, Thermo Scientific), then equilibrated with 20 µL of solvent A. 5 µL of digests were loaded onto each Evotip. The Evotips were washed once with 20 µL of solvent A and kept wet with 100 µL of solvent A before LC-MS analysis.

DIA-MS data acquisition

Peptides samples were loaded onto Evosep One (EV-100 S00394, Evosep) coupled to a Orbitrap Exploris 480 mass spectrometry (MA10103C, Thermo Scientific) interfaced with a FAIMS PRO device. A 150um x 15cm PepMap RSLC C18 easy spray column (ES906 Thermo Scientific) was used as the analytical column on Evosep One system.

An optimised data independent acquisition (DIA) method was used for MS data acquisition to maximise the depth of plasma proteome. The MS1 scan instrument setting: orbitrap resolution 120000; scan range (m/z): 345–1055; FAIMS CV (V): three sequential CVs – 40, -55, -70; Normalised AGC target (%): 300; Maximum injection time mode: Auto. The DIA MS2 scan the instrument setting: Orbitrap resolution: 30000; DIA window and m/z range: 38 variable DIA windows covering from 350 to 1050; FAIMS CV (V): three sequential CVs – 40, -55, -70 the same as MS1 FAIMS setting; RF length (%): 40; Normalised AGC target (%): 1000, Customed Maximum injection time (ms): 54; Data type: Profile; Loop control N = 13.

Samples were run using Evosep One 15 sample per day method (gradient of 88min). Master pool samples were loaded as a QC after each plate row of samples.

Data pre-processing

In total, 714 DIA MS files were acquired including 614 lung cancer samples, 42 pooled quality control plasma samples and 58 master pool digests. Data was first converted to .htrms format using HTRMS converter software available with Spectronaut V18.5 (Biognosys). All 714 HTRMS files were loaded into Spectronaut V18.5 (Biognosys) for library-free directDIA analysis. Unless specified, default settings parameters were used. Default BSG factory settings were used for peptide/protein identification and quantification. Qvalue cutoff was 0.01 for both peptides and protein identification. LFQ method for

protein quantitation at MS1 level and cross run global normalisation were enabled. Pulsar search workflow: directDIA+(Deep) was used with peptide fixed modification cysteine carbamidomethylation, and variable modification: acetyl (protein N-term) and oxidation (M). Two miss cleavages were allowed for trypsin cleavage. Human uniport FASTA file (UP000005640_9606_Hsapiens, Feb 2023) was used for *in silico* spectra library generation and protein inference.

Proximity extension assay (PEA) data acquisition and processing

Sample selection

For PEA analysis, 600 samples were selected prioritising samples with low haemolysis score. The 600 selected samples were processed in two stages; an initial study of 88 samples (one plate), PEA 01, consisting of 29 healthy controls and 59 lung cancer cases, was followed by a larger cohort of 528 samples across six plates. PEA 02 consisted of 100 healthy controls and 428 lung cancer, with 16 bridging samples (6 healthy control, 10 LC) used across the two studies.

Olink Explore 3072 Panel

The Olink® Explore 3072/384 assay (referred to as PEA) consists of eight panels: Cardiometabolic, cardiometabolic II, inflammation, inflammation II, neurology, neurology II, oncology and oncology II. Each panel targets a maximum of 384 proteins, which quantify protein abundance using pairs of antibodies with complimentary oligonucleotide tags targeting the same protein. When both antibodies are bound to the protein of interest (i.e. in close proximity) the oligonucleotide tags hybridise, allowing DNA polymerase-dependent extension. The extended dsDNA is then tagged with barcodes and quantified using Next Generation Sequencing. Samples were randomised across the seven plates. Each assay is spiked with internal controls that act as incubation controls, extension controls and detection controls. Alongside internal controls, each plate contains sample controls, negative samples and inter-plate controls, used for plate normalisation, limit of detection (LOD) estimation and variation assessment respectively. The Olink® Explore 3072/384 assay was carried out by Randox Laboratories Ltd. Six µL of plasma from each sample was mixed with the antibodies from the eight panels and the resulting dsDNA specific for target protein extended, labelled with barcoded sequences and sequenced using NovaSeq 6000 Sequencing System (Illumina) for relative quantification.

Data pre-processing

Raw data was processed using NGS2COUNTS and imported into Olink® NPX software. Raw NGS reads were converted into normalised protein expression (NPX) values through a series of normalisation steps. Firstly, samples are normalised to extension controls and converted to Log_2 scale. Secondly, plate-control normalisation was performed by normalising the median of each sample to the median of plate controls. Intensity normalisation was subsequently performed, centering the median of each sample to the median of all samples (excluding controls) of the plate. These NPX values were used for all downstream analysis.

DNA-aptamer data acquisition and pre-processing

Sample selection

A sub-cohort of the Princess Margaret patient cohort was selected for protein quantification on a third platform. This cohort consisted of 72 healthy control patients that were matched to 32 early lung cancer and 32 late lung cancer according to age, sex and smoking status (Extended Data Fig. 7a).

DNA-aptamer quantification of feature subset

A DNA-aptamer approach, SomaScan 11K V5.0 (SomaLogic), was used to assess feature concordance on a third proteomics technology. From the 11K SomaScan panel, 100 proteins from PEA and DIA-MS feature selection were chosen (Extended Data Fig. 7b). Samples were randomised across three plates according to cancer status (cancer-free, early lung cancer and late lung cancer). The SomaScan assays consist of SOMAmer probes composed of fluorophores, biotin and a photocleavable linker, which bind specific proteins. Once bound, biotinylated protein complexes are captured on Streptavidin beads and washed to remove unbound/weakly bound proteins. Photocleavage dissociates SOMAmers from protein complexes, non-specific complexes are dissociated and prevented from reforming using polyanionic competitors. Remaining SOMAmer-protein complexes are recaptured on streptavidin, eluted and binding of complementary DNA sequences to a microarray allows measurement of fluorescence.

Data pre-processing

Data processing of the SomaScan 11K V5.0 assay was performed using SomaScan in-house software. Results are reported in relative fluorescence units (RFU) and processed as follows: Firstly, using the 12 control SOMAmers spiked into each sample prior to microarray hybridisation normalisation was carried out. This was followed by intraplate median signal normalisation using plasma calibrator controls run alongside patient samples to account for dilution differences. Plate scaling was also carried out using calibrator samples, normalising the medians of calibrator reference ratios. Calibration was performed to normalise calibrator controls to reference values for each SOMAmer. Finally, Adaptive Normalisation by Maximum Likelihood (ANML) was used to normalise signal across samples. Briefly, SOMAmer reagents within a sample that are within two population standard deviations from reference are used to generate scaling factors; this was iterated over until convergence. The 11K panel was used for pre-processing while the data for the 100 chosen SOMAmers plus the 12 control SOMAmers were reported.

Bioinformatic analysis

Sample and protein filtering

DIA-MS and PEA sample filtering are described in Extended Data Fig. 1a. Data filtering and quality assessment were performed in the R environment (version 4.2.2).

For DIA-MS, any raw PG.Quantity values < 100 were replaced with NA. Samples with high haemolysis (≥ 2) were removed, along with samples not quantified by PEA. One clinical sample (cancer) that was later identified to be post-treatment was also removed. A minimum of 70% coverage across all remaining clinical samples was applied.

PEA data from both studies (PEA 01 and PEA 02) were loaded and bridge-normalisation performed using the OlinkAnalyse package (version 3.8.2) available in R. Assays with QC warning were removed for all patients, as well as assays with incomplete data across the two studies. For bridging samples, data for PEA 02 was discarded and for assays with repeat measurements across panels, measurements with a warning were removed and the mean assay calculated across the panels. Patients with high haemolysis were removed to give the same population for comparison with DIA-MS.

DNA-APT ADAT file was loaded using the SomaDataIO package (version 6.1.0) in R; all 140 samples passed Somalogics quality assessment and 20% dilution values were taken forward for further analysis.

Data imputation and normalisation

For non-ML related analysis, DIA-MS data for the 563 clinical samples were imputed using the MissForest algorithm with default parameters from the missingpy package (version 0.2.0, https://github.com/epsilon-machine/missingpy) available in the Python environment and adapted to be compatible with the ML workflow. MissForest is a python implementation of the missForest package in R[34] that imputes data using random forests typically used for imputation of data that is considered missing at random (MAR)[34]. Assessment of missingness often relies on grouping of samples according to clinical characteristics. In this case where introduction of bias from imputation should be avoided and not be dependent on ML target, a target-independent model that can be used to impute unseen (unlabeled) data was needed. Given missForest has previously been shown to be a high performing imputation algorithm for missing data imputation in LC-MS based proteomics[35] and does not depend on knowing clinical characteristics of each instance, this method of imputation was selected for DIA-MS imputation. Following imputation, data was Log_2 transformed and normalised using median centering to account for differences in sample loading.

Imputation was not required for PEA or DNA-APT data. The NPX values for PEA are already on the Log_2 scale, while DNA-APT data was Log_2 transformed.

Post-analysis quality, missingness and dynamic range assessment

For DIA-MS, CVs were calculated using pooled plasma samples run across seven plates without normalisation/Log2 transformation using standard CV formula (standard deviation/mean*100). For PEA, NPX data are both normalised and on Log_2 scale making the previous CV formula unsuitable[36]. For PEA and DNA-APT Log_2 transformed ANML values, the following formula was used:

$$CV_i = 100 \bullet \sqrt{e^{Sln_i{}^2} - 1}\,, \quad where \;\; Sln_i = ln2 \bullet \; SD_{isk}\,.$$

Missingness in DIA-MS data was assessed using all pre-treatment clinical samples with an NA for PG.Quantity. For PEA, missing data is not reported. Instead, values below the lower limit of detection (LLOD) were replaced with NA and classed as missing. Average missingness for each protein was reported as mean ± standard deviation.

To assess dynamic range, known concentrations of proteins according to the Human Protein Atlas (HPA) (v24.0)[17] were mapped to UniProt IDs in PEA and DIA-MS data. For ProteinGroups in DIA-MS data the first protein was taken. Where blood concentration from mass spectrometry was not available, immunoassay known concentrations were used. “Secretome location” was used to identify secretion location of proteins predicted to be secreted. Prediction of secreted protein in HPA was performed using amino acid sequence and location derived from literature/available data[17].

Univariate and exploratory analysis

Principal component analysis (PCA) was performed using prcomp() available in the stats package available in R (version 4.4.1). Default parameters were used except for scale = TRUE to equalise variance contribution of features and prevent bias towards high-variance features.

Partial least squares regression (PLS) was used to distinguish lung cancer cases from healthy controls using the plsr() function from the pls package in R, with cross-validation (validation = “CV”) and scale = TRUE.

T-distributed stochastic neighbourhood embedding (t-SNE) was performed using Rtsne from Rtsne (version 0.17) package in R. Uniform Manifold Approximation and Projection (UMAP) was performed using umap() from umap package (version 0.2.10.0) in R.

For univariate analysis a linear model was fit to each protein using lmFit() from limma, followed by empirical Bayes adjustment using eBayes() from limma to improve estimation of variance[37]. Differentially expressed proteins were extracted using decideTests() with p.value = 0.05, lfc = 1 and adjustment.method = “fdr”. Summary statistics were extracted using topTable(). Proteins were considered differentially expressed if Log2 fold change > 1 and FDR-adjusted p-value < 0.05. Results are visualised as volcano plots (Fig. 2e,f and Extended Data Fig. 3c,d) and heatmaps (Supp Fig. 2B, 2C and Fig. 6A).

Over-representation analysis (ORA) and gene set enrichment analysis (GSEA)[38] of Gene Ontology (GO) biological processes[39] were carried out using enirchGO() and gseGO() from clusterProfiler package (version 4.12.2) available in R[40]. The analysis was performed across sub-cohorts: For protein groups the first ID was used, pvalueCutoff = 0.05 and pAdjustMethod = “fdr”, human database was used (org.Hs.eg.db) and minGSsize = 5. Semantic similarity analysis was used to remove redundant terms,

applied using the mgoSim() and clusterSim() functions available from clusterProfiler. Data was visualised as heatmaps.

Spearman correlation across platforms was chosen due to it not being dependent on scale and calculated using cor() function in R for the 152 patient cohort quantified using all three methods. For assessing significance of correlation *p*-values were FDR-adjusted. For DNA-APT data a second method of normalisation was used (median centering) as an alternative to ANML. Spearman correlation was calculated on both median-centered and ANML normalised datasets, features with Spearman correlation $R < 0.4$ in the appropriate comparison were removed and only the intersection between the two normalisation methods were considered. This would account for bias potentially introduced following ANML normalisation[41].

Unsupervised, hierarchical clustering of differentially abundant proteins, ORA/GSEA results, Shapley value interactions and Spearman correlation co-efficient were performed using dist() and hclust() available in stats package in R and clusters generated using cutree() function in R.

Data visualisation

All data visualisations were generated in the R environment with the exception of networks. For heatmaps the heatmap.2() function from the gplots package was used with the exception of Fig. 4C visualised using ggplot2. Venn diagrams were generated using ggvenn. All other plots were generated using ggplot2.

ML analysis for feature selection

All ML data generation was performed using Python (version 3.10). All models undergo the same sample stratification, train/holdout split as well as normalisation and imputation. Following this, different combinations of model selection, feature de-selection, augmentation, feature selection and model optimisation are trained on training data only to give rise to the 75 different models produced for each sub-cohort of each dataset, as described in Extended Data Fig. 3a. Each step in the ML process is detailed below.

Sample stratification and data split

The 563 clinical samples that passed pre-processing were stratified using sex > age-range (10 year intervals) > stage (control vs early vs late LC) > histology (NSCLC vs SCLC) > smoking history (non-heavy vs heavy-smoker). An 80/20% split was chosen for generating train/holdout data; Training data was used for training models and analysis, while the 20% holdout dataset was used to assess model generalisability and evaluate feature importances.

This single train/holdout split for healthy control vs lung cancer samples was used to generate sub-cohorts for all controls vs early only lung cancer patients (Stage 0–2), late lung cancer patients (Stage

3–4), NSCLC only, SCLC only, LUAD only and LUSC only. For male only and female only models, only male and female healthy controls and lung cancer patients were included, respectively. Using the same split for different cohorts allowed comparison of sub-cohort analyses as well as maintaining the unseen nature of the holdout dataset.

Normalisation, imputation and scaling for ML analysis

Data was median centered and Log2 transformed. missForest() from missingpy was used to impute training data as described previously. This model was fit on the training data only, and the subsequent model used to impute the training data and unseen holdout data. Fitting the imputation model on the training data only and applying to holdout avoids data-leakage between the two datasets and subsequent over-inflated ML performances.

Similarly, scaling using StandardScaler() from scikit-learn (version 1.3.0, used for all scikit-learn functions) whereby the fit was applied to training only and used to transform train and holdout.

Model selection

Tree-based models are suitable for handling high-dimension datasets, maintaining feature interactions and capturing non-linear relationships, unlike linear models often used to build ML models on omics datasets. Extreme gradient boosting (XGBoost), a tree-based model, was chosen to classify healthy controls vs lung cancer (and sub-cohorts). XGBClassifier() function from the XGBoost Python model (version 1.7.6) was used.

Balanced random forest (BRF) were chosen as an alternative tree-based model to XGBoost. BRF handles class imbalances by randomly under-sampling the majority class and maintaining balanced class representation during training, unlike random forests (RF) which favour the majority class. The use of bootstrapping in RF/BRF also prevents overfitting. BalancedRandomForestClassifier() from the Python package imblearn (version 0.12.2, https://github.com/scikit-learn-contrib/imbalanced-learn) was used to build BRF models.

Hyperparameters were determined during model-optimisation.

Feature de-selection

Feature de-selection is the process by which features with low information content about the target (control vs lung cancer) are removed, with the aim of improving ML performance and reducing overfitting. Two methods of feature de-selection were used along with no feature de-selection.

The first method uses mutual information (MI) that assesses the dependency between two variables (cancer status and protein abundance) that is capable of capturing non-linear relationships. The mutual_info_classif() was used to compute MI and was cross-validated using RepeatedStratifiedKFold(), stratified according to target (healthy control vs LC) with n_splits = 2 and n_repeats = 25. Any feature with

cross-validated MI less than 3.5% was considered to not contain information related to classification, in this case lung cancer case, and was subsequently removed from the feature panel.

Alternatively, BorutaShap a two-step algorithm that combines the Boruta Algorithm[42] with Shapley values[43] was used to perform feature de-selection. Boruta works by creating shadow features by random permutation using actual features. By comparing feature importances of shadow and actual features, features where shadow features out-perform can be removed. In the case of BorutaShap, feature importances are calculated using Shapley values for each sample rather than model importances[18]. Shapley values were derived from cooperative game theory and applied to ML models to assess feature importance. Shapley values fairly attribute a "reward" among features based on their individual contributions to the prediction. This approach was implemented using the BorutaShap function from the BorutaShap package (version 1.0.17, (https://github.com/Ekeany/Boruta-Shap), using an XGBoost model with importance_measure = “shap”.

Augmentation

In some models augmentation was applied to balance classes between healthy controls and lung cancer cases by up-sampling the minority class. A two-step process was implemented to augment the minority class. Firstly, synthetic minority over-sampling technique (SMOTE) was used to generate synthetic samples from the minority class. SMOTE selects a random point from the minority class and interpolates a new datapoint between it and a neighbour actual sample, maintaining correlations/relationships between features in the synthetic data[44]. Tomek links are then used to remove a portion of the majority class with overlapping decision boundaries with the augmented minority class. SMOTE-Tomek augmentation was applied to training data following feature de-selection using the SMOTETomek() function available in imbalanced-learn (version 0.12.2).

Recursive feature addition (RFA) with Shapley values

Proteomics models that require large feature sets have limited clinical utility if needing to be translated onto alternative platforms. For example, immunoassay approaches require small panels for single-plex assays (such as ELISA) up to tens of analytes for multiplexed assays. Alternatively targeted mass-spectrometry methods such as single-reaction, multiple-reaction and parallel reaction monitoring (SRM, MRM and PRM, respectively) can be used, however such methods depending on instrumentation can also be limited to as little as 15 peptides[45].

Recursive feature addition (RFA) was used to build models that rely on a small subset (< 30) of proteins. RFA first selected the highest importance feature according to cross validated Shapley values. Using mean absolute Shapley value, high importance features were sequentially added and model performance evaluated to select the smallest subset of features that achieves the highest performance.

RFA was coupled with repeated stratified cross validation (RSCV), as described for BorutaShap, to give more robust performance estimates across different feature sets.

Traditionally, model feature importances (such as those that come from XGBoost or BRF) are used to rank features and determine order of feature addition in RFA. The use of Shapley values instead of model feature importances captures both global (overall) and local (sample-specific) attributions, as opposed to only global. This ensures a fair and consistent distribution of feature importance and fairly distributes importance among correlated features. Another key advantage is their ability to account for feature interactions, revealing how features work together to give a prediction rather than evaluating them in isolation. Shapley values were calculated using the TreeExplainer() function from the Shap Python package, a fast implementation of tree-based Shapley value (TreeSHAP) generation using interventional perturbations to maintain dependencies between features. As an alternative Shapley value method, the package EjectSHAP was also used[46]. EjectSHAP works similarly to TreeSHAP with the exception of how it attributes importances to features that never contributed to a prediction; if a branch or single node in a tree is not used to make a prediction, the importance of features within that tree are zero (i.e. “ejected”).

Two metrics are used to evaluate model performance, area under the receiver operator characteristic curve (AUROC) and Matthews correlation coefficient (MCC). These metrics were evaluated on the top 30 features in XGBoost models and top 50 features in BRF, which often required higher numbers of features to converge during RFA. ROC curves plot the true positive rate (TPR/sensitivity) vs false positive rate (FPR) at different thresholds. AUC measures the area under the ROC curve; higher AUROC indicates higher model performance. MCC on the other hand utilises true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), giving a more meaningful performance in cases where data is imbalanced. MCC can be interpreted as the correlation between the true labels and the predicted labels for binary classification problems[47,48]. MCC was calculated using the matthews_corrcoef() function in scikit-learn, using the following formula:

$$MCC = \frac{TP \bullet TN - FP \bullet FN}{\sqrt{(TP+FP) \bullet (TP+FN) \bullet (TN+FP) \bullet (TN+FN)}} .$$

MCC and AUROC performance gives rise to five different models. The first maximising mean AUROC across folds in RFA, the second maximising mean MCC across folds, the third approach maximises the stability of MCC across folds. MCC stability was calculated by dividing the mean of MCC across folds by the standard deviation across folds. This will give a subset of features for which MCC were minimally impacted by the fold used in RSCV. Finally, a ttest MCC and ttest AUROC feature set are generated to give a smaller subset of features, using ttest_ind_from_stats() available in scipy (version 1.11.1).

In some cases, mean AUROC, MCC or MCC stability, or ttest MCC and AUROC might converge on the same subset of features and therefore produce less than five models, while in some cases all five RFA performance metrics give rise to models with different feature sets.

Model optimisation

XGBoost models require fine-tuning of model hyperparameters to avoid overfitting/underfitting on training data. For this, NSGAIISampler() from Optuna (version 3.6.1) Python module was used for multi-objective optimisation of model hyperparameters.

RFA using baseline XGBoost or BRF hyperparameters were used to select top 30 features. Baseline hyperparameters were chosen to mitigate the overfitting potential of the models. For the XGBoost models we limited the max_depth to 2, the n_estimators to 200, the learning rate to 0.05, lambda to 0, and setting the scale_pos_weight as the square root of the ratio of number of controls to cancer cases. For the BRF models we limited the max_depth to 10, the n_estimators to 500 and used sampling without replacement.

A model using these features was then optimised in Optuna to maximise three combinations of model performance metrics: MCC only, MCC + sensitivity at 99% specificity (Sens@90%Spec) and MCC + AUROC. If followed by optimisation, RFA was repeated using the optimised model hyperparameters to give a selection of features as described previously. Optuna was selected over other optimisation tools (such as GridSearchCV from Scikit-learn) as a result of its efficiency in searching higher-dimensional parameters spaces due to its Bayesian optimisation underpinning. Optimised model-hyperparameters were chosen from a local neighbourhood of similarly performing models, improving the stability of model sensitivity to minor hyperparameter perturbations.

Model evaluation

Following training of the models, each model was evaluated using the unseen holdout test set. Assessment of model generalisability, rather than performance, promoted the refinement of features based on similar performance across train/holdout data with the aim of improving generalisability in future unseen data. To do this in a high throughput manner (1,350 models to evaluate), 95% confidence intervals (CI) were generated for the training data during model training. These 95% CI evaluated model performance during training using RSCV with n_splits = 4 (equivalent in size to holdout data) and n_repeats = 25. 95% CI were calculated for the following performance metrics: MCC, AUROC, Sens@90%Spec, Sens@95%Spec and Sens@99%Spec. These specificities were chosen to ensure reliable prediction in a low-prevalence setting, where high specificity reduces FPRs while maintaining high TPR. If four out of five of the performance metrics in holdout were within the 95% CI of training data a model was considered to have generalised, and was taken forward for further interpretation.

Model interpretability (XAI)

Shapley values for each model were calculated as described during RFA. By calculating the mean of absolute Shapley values the contribution of each feature to all predictions can be interpreted. Within each generalisable model, features were ranked from largest mean absolute Shapley value to lowest. The average rank of a feature across generalisable models was used to understand the most important features for each sub-cohort. All data used in plotting Shapley value-related data comes from the SHAP module in Python.

To assess the interactions between features, rather than the importances of a single feature, Shapley value interactions were used. Shapley value interactions capture pairwise interactions by assessing how the presence of one feature influences the contribution of a different feature to model prediction. Shapley value interactions aim to improve model-interpretability while identifying synergies and redundancies between features[39]. For a single representative healthy control vs lung cancer model, Shapley value interactions were calculated using the shap_interaction_values method from the TreeExplainer class from the SHAP module in Python. High values for Shapley value interaction equate to a higher proportion of Shapley value explained by the interaction between the two features. Shapley value interactions were normalised to the sum of all Shapley value interactions for both DIA-MS and PEA data to account for variation in scale of Shapley values between the two datasets. Normalised Shapley value interactions are visualised using a heatmap (Extended Data Fig. 6c,d), while a network of interactions between features accounting for more than 5% of actual Shapley value was visualised using Cytoscape (version 3.10.3)[49] (Fig. 4d,e).

ML analysis for cross-platform concordance

ML models were used to assess the ability of features with concordance across the three platforms to predict lung cancer using DIA-MS and PEA data. Two panels were fit to both datasets; the first panel contained 18 features that were concordant (definition of concordance used: Spearman $R^2 > 0.4$) across DIA-MS, PEA and DNA-APT. The second panel for DIA-MS were features concordant between DIA-MS and DNA-APT that if present were also concordant in PEA, and for PEA those that were concordant between PEA and DNA-APT that if present were also concordant in DIA-MS. This accounted for non-overlapping features between DIA-MS and PEA.

A logistic regression was chosen as the model for assessing concordance due to its interpretability and low number of parameters to fit, allowing easier comparison of performances across datasets and/or features. The LogisticRegressionCV class from Scikit-learn was used to fit a balanced logistic regression on training data from PEA/DIA-MS data using relevant features. Performances were evaluated on holdout data not used in training.

Comparing univariate feature selection with ML-based feature selection

To compare ML-dependent feature selection with a univariate approach, as has been used in other studies[19,50], an analysis of variance (ANOVA) was used to select features had the smallest FDR-adjusted p-values when comparing lung cancer cases with healthy controls in all data. Twenty features were selected for each cohort that had the smallest adjusted p-values according to an f_test from Scikit-learn and compared with the twenty features that had the highest Shapley value rank in that cohort. A balanced logistic regression was used to assess the ability of these features to predict lung cancer as previously described. MCC and AUROC from the two logistic regression models were compared with their respective metric from an XGBoost model (Extended Data Fig. 3e).

Declarations

Data and code availability

All data analysed and code used in this study is available upon request.

Acknowledgements

We thank all participants in the CALIBRE study. We also thank members of the Princess Margaret Cancer Centre, Toronto General Hospital and Princess Margaret Hospital for their contribution to the study. We would like to thank members of the Target Discovery Institute, in particular Dr Iolanda Vendrell for their technical advice in this study. We thank team members at Randox for their support collecting and processing the PEA (Olink) data. We would like to thank all former and current employees of Oxford Cancer Analytics Ltd for their guidance in the preparation of this manuscript, especially Dr Honglei Huang and Dr Heinrich Roder.

B.M.K. and R.F. were supported by the Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Science (CIFMS), China (grant number: 2024-I2M-2-001-1).

Author contributions

H.R.F and N.G performed data analysis and developed the ML workflow, produced all figures and tables and wrote the text of the manuscript. L.H developed the ML workflow and contributed to interpretation of the data and data analysis. I.K developed and optimised the DIA-MS workflow and experimental design, carried out sample preparation, data acquisition and pre-processing of DIA-MS dataset. Ella.M and Emma.M contributed to data analysis, interpretation including cohort selection and stratification. D.A.S and J.S contributed to conceptualisation of the study, cohort selection and experimental design. G.L, L.J.Z and D.P provided insights into clinical cohorts, sourced the clinical samples and performed clinical data abstraction in this study. M.D contributed to proteomic sample selection and preparation. B.K and R.F provided insights into proteomic methodologies and analysis. A.H contributed to conceptualisation and methodology of the study. P.J.L conceptualised, supervised and guided the design, analysis, interpretation of the study, and wrote the text of the manuscript. All authors contributed to the drafting of the manuscript.

Competing interests

H.R.F, N.G, L.H, I.K, M.D, J.S, Emma.M, Ella.M, D.A.S, A.H, and P.J.L are current/former employees, shareholders, and/or share option holders of Oxford Cancer Analytics Ltd.

R.F and B.M.K are share option holders of Oxford Cancer Analytics Ltd.

Oxford Cancer Analytics Ltd sponsored this study.

References

1. World Health Organisation. Lung cancer. *World Health Organisation* https://www.who.int/news-room/fact-sheets/detail/lung-cancer (2023).
2. American Cancer Society. Lung Cancer Survival Rates. https://www.cancer.org/cancer/types/lung-cancer/detection-diagnosis-staging/survival-rates.html https://www.cancer.org/cancer/types/lung-cancer/detection-diagnosis-staging/survival-rates.html (2024).
3. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *New England Journal of Medicine* 365, (2011).
4. de Koning, H. J. *et al.* Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *New England Journal of Medicine* 382, (2020).
5. Jemal, A. & Fedewa, S. A. Lung cancer screening with low-dose computed tomography in the United States – 2010 to 2015. *JAMA Oncol* 3, (2017).
6. Dickson, J. L. *et al.* Uptake of invitations to a lung health check offering low-dose CT lung cancer screening among an ethnically and socioeconomically diverse population at risk of lung cancer in the UK (SUMMIT): a prospective, longitudinal cohort study. *Lancet Public Health* 8, (2023).
7. LoPiccolo, J., Gusev, A., Christiani, D. C. & Jänne, P. A. Lung cancer in patients who have never smoked — an emerging disease. *Nature Reviews Clinical Oncology* vol. 21 Preprint at https://doi.org/10.1038/s41571-023-00844-0 (2024).
8. Chin, R. I. *et al.* Detection of Solid Tumor Molecular Residual Disease (MRD) Using Circulating Tumor DNA (ctDNA). *Molecular Diagnosis and Therapy* vol. 23 Preprint at https://doi.org/10.1007/s40291-019-00390-5 (2019).
9. Chaudhuri, A. A. *et al.* Early detection of molecular residual disease in localized lung cancer by circulating tumor DNA profiling. *Cancer Discov* 7, (2017).
10. Chabon, J. J. *et al.* Integrating genomic features for non-invasive early lung cancer detection. *Nature* 580, (2020).
11. Kelly-Spratt, K. S. *et al.* Plasma proteome profiles associated with inflammation, angiogenesis, and cancer. *PLoS One* 6, (2011).
12. Pitteri, S. J. *et al.* Tumor microenvironment-derived proteins dominate the plasma proteome response during breast cancer induction and progression. *Cancer Res* 71, (2011).
13. Bhardwaj, M., Terzer, T., Schrotz-King, P. & Brenner, H. Comparison of proteomic technologies for blood-based detection of colorectal cancer. *Int J Mol Sci* 22, (2021).
14. Ng, S., Masarone, S., Watson, D. & Barnes, M. R. The benefits and pitfalls of machine learning for biomarker discovery. *Cell and Tissue Research* vol. 394 Preprint at https://doi.org/10.1007/s00441-023-03816-z (2023).
15. Shen, S. Y. *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* 563, (2018).

16. Khodayari Moez, E. *et al.* Circulating proteome for pulmonary nodule malignancy. *JNCI: Journal of the National Cancer Institute* 115, (2023).
17. Uhlén, M. *et al.* The human secretome. *Sci Signal* 12, (2019).
18. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2, (2020).
19. Liang, H. *et al.* LcProt: Proteomics-based identification of plasma biomarkers for lung cancer multievent, a multicentre study. *Clin Transl Med* 15, e70160 (2025).
20. Müller, S. *et al.* An Empirical Evaluation of the Rashomon Effect in Explainable Machine Learning. *ArXiv* (2023) doi:https://doi.org/10.48550/arXiv.2306.15786.
21. Katz, D. H. *et al.* Proteomic profiling platforms head to head: Leveraging genetics and clinical traits to compare aptamer- And antibody-based methods. *Sci Adv* 8, (2022).
22. Birhanu, A. G. Mass spectrometry-based proteomics as an emerging tool in clinical laboratories. *Clinical Proteomics* vol. 20 Preprint at https://doi.org/10.1186/s12014-023-09424-x (2023).
23. Eldjarn, G. H. *et al.* Large-scale plasma proteomics comparisons through genetics and disease associations. *Nature* 622, (2023).
24. Petrera, A. *et al.* Multiplatform Approach for Plasma Proteomics: Complementarity of Olink Proximity Extension Assay Technology to Mass Spectrometry-Based Protein Profiling. *J Proteome Res* 20, (2021).
25. Weber, M. A. *et al.* Myoglobin plasma level related to muscle mass and fiber composition - A clinical marker of muscle wasting? *J Mol Med* 85, (2007).
26. Czarnowski, D. Syndecans in cancer: A review of function, expression, prognostic value, and therapeutic significance. *Cancer Treatment and Research Communications* vol. 27 Preprint at https://doi.org/10.1016/j.ctarc.2021.100312 (2021).
27. Dinh, H. Q. *et al.* Coexpression of CD71 and CD117 Identifies an Early Unipotent Neutrophil Progenitor Population in Human Bone Marrow. *Immunity* 53, (2020).
28. Daniels, T. R., Delgado, T., Rodriguez, J. A., Helguera, G. & Penichet, M. L. The transferrin receptor part I: Biology and targeting with cytotoxic antibodies for the treatment of cancer. *Clinical Immunology* vol. 121 Preprint at https://doi.org/10.1016/j.clim.2006.06.010 (2006).
29. Human Protein Atlas. HAGH: Cancer. *HAGH* https://www.proteinatlas.org/ENSG00000063854-HAGH/cancer (2024).
30. Yang, W. & Rubin, J. B. Treating sex and gender differences as a continuous variable can improve precision cancer treatments. *Biol Sex Differ* 15, 35 (2024).
31. Budnik, B., Amirkhani, H., Forouzanfar, M. H. & Afshin, A. Novel proteomics-based plasma test for early detection of multiple cancers in the general population. *BMJ Oncology* 3, (2024).
32. Bujang, M. A. & Adnan, T. H. Requirements for Minimum Sample Size for Sensitivity and Specificity Analysis. *JOURNAL OF CLINICAL AND DIAGNOSTIC RESEARCH* 10, YE01–YE06 (2016).

33. Searfoss, R. *et al.* Impact of hemolysis on multi-OMIC pancreatic biomarker discovery to derisk biomarker development in precision medicine studies. *Sci Rep* 12, (2022).
34. Stekhoven, D. J. & Bühlmann, P. Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, (2012).
35. Jin, L. *et al.* A comparative study of evaluating missing value imputation methods in label-free proteomics. *Sci Rep* 11, (2021).
36. Canchola, J. A. Correct Use of Percent Coefficient of Variation (%CV) Formula for Log-Transformed Data. *MOJ Proteom Bioinform* 6, (2017).
37. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43, (2015).
38. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102, (2005).
39. Aleksander, S. A. *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* 224, (2023).
40. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* 2, (2021).
41. Pietzner, M. *et al.* Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. *Nat Commun* 12, (2021).
42. Kursa, M. B. & Rudnicki, W. R. Feature selection with the boruta package. *J Stat Softw* 36, (2010).
43. Lundberg, S. M. & Lee, S. I. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. pp. 4765–4774 (2017). *NIPS-2017 Advances in Neural Information Processing Systems* 32, (2017).
44. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16, (2002).
45. Van Bentum, M. & Selbach, M. An introduction to advanced targeted acquisition methods. *Molecular and Cellular Proteomics* vol. 20 Preprint at https://doi.org/10.1016/J.MCPRO.2021.100167 (2021).
46. Campbell, T. W., Roder, H., Georgantas III, R. W. & Roder, J. Exact Shapley values for local and model-true explanations of decision tree ensembles. *Machine Learning with Applications* 9, (2022).
47. Chicco, D. & Jurman, G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Min* 16, (2023).
48. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21, (2020).
49. Shannon, P. *et al.* Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res* 13, (2003).
50. Davies, M. P. A. *et al.* Plasma protein biomarkers for early prediction of lung cancer. *EBioMedicine* 93, (2023).

Figures

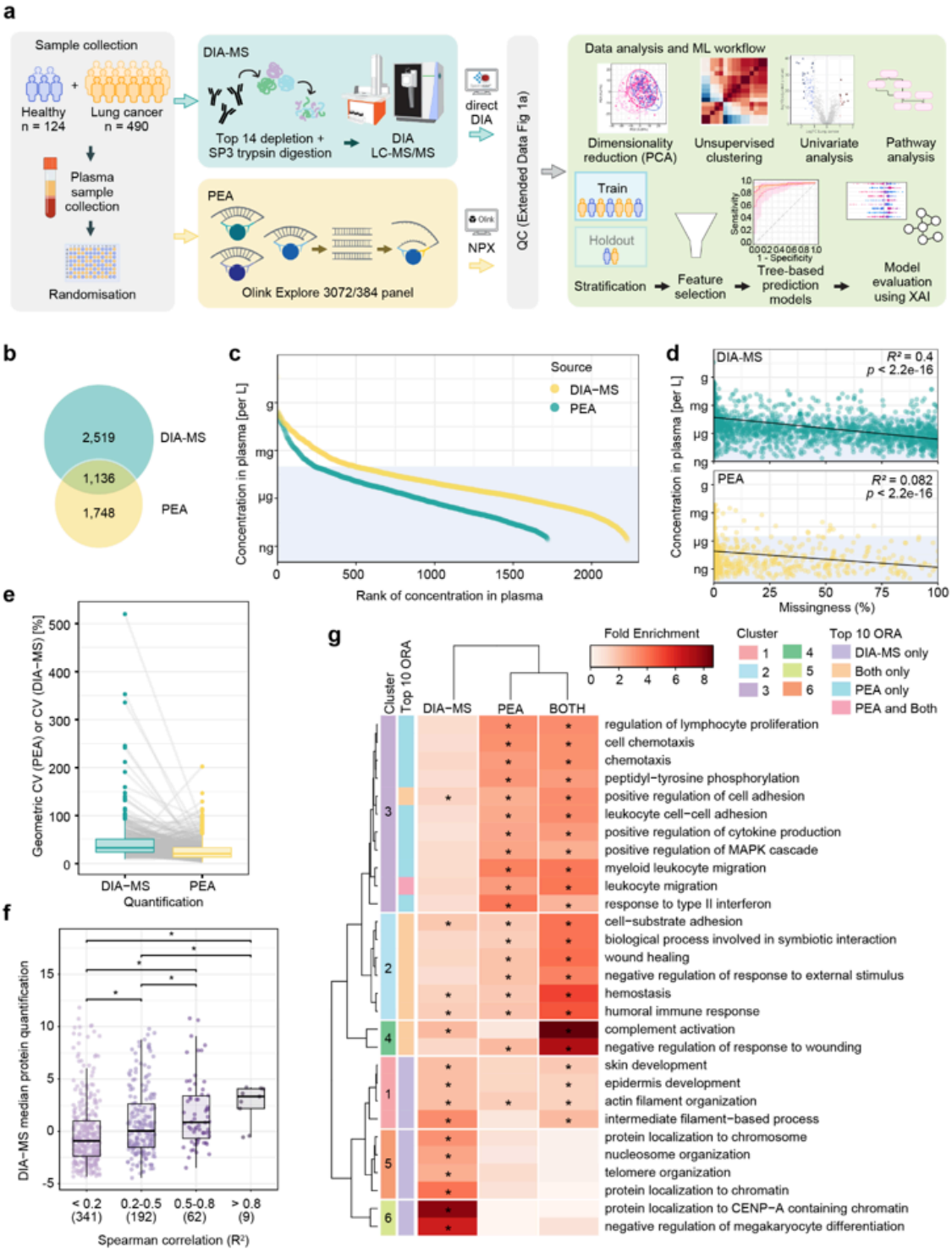


Figure 1

Characterisation of the blood plasma proteome using DIA-MS and PEA. a, Proteomics and clinical workflow. **b**, Overlap in proteins identified by DIA-MS (blue) and PEA (yellow). **c**, Distribution of proteins with known concentrations according to Human Protein Atlas (HPA), proteins with concentration below 0.1 mg/L (shaded blue) are considered in the mid-low abundance range. **d**, Concentration vs missingness (%) of proteins with known concentration in HPA showing Spearman correlation with FDR-

adjusted p-value. **e**, Inter-plate CV (%) of proteins quantified by both DIA-MS and PEA across pooled plasma samples. Lines between boxplots link matching protein IDs. **f**, Boxplot showing relationship between median protein quantification by DIA-MS and strength of Spearman correlation of proteins quantified by DIA-MS and PEA. Spearman correlation is shown as R^2, and divided into four groups. * FDR-adjusted *p*-value < 0.05. **g**, Heatmap of hierarchical clustering (Manhattan average) of the top 10 over-represented gene ontology biological pathways in proteins quantified by DIA-MS only (DIA-MS), PEA only (PEA) or quantified by both DIA-MS and PEA (BOTH). Top 10 ORA colour key indicates which analysis identified pathway as in top 10 overrepresented pathways. Purple = top 10 ORA pathway in DIA-MS only proteins, orange Top 10 pathways only in proteins quantified by both DIA-Ms and PEA, blue = top 10 pathway in proteins quantified by PEA only; pink = top 10 pathway in proteins quantified by PEA only and quantified by both PEA and DIA-MS. * indicates significant over-representation (FDR adjusted p-value < 0.05).

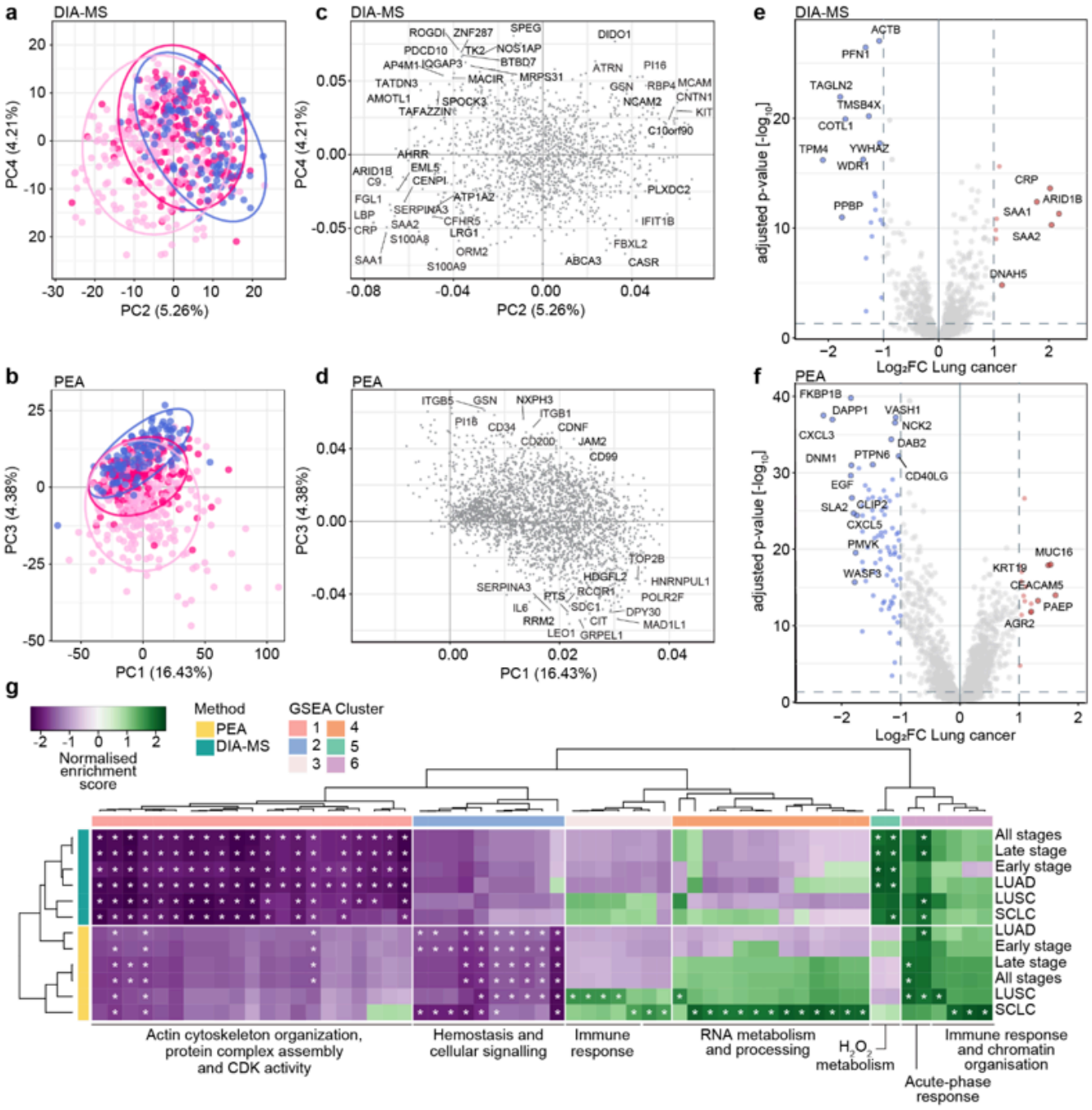


Figure 2

Characterisation of blood plasma proteins associated with lung cancer by stage and common histological subtypes. PCA of patient samples using QC qualified DIA-MS quantification (**a**) or PEA quantification (**b**) of proteins across patients. Blue = control, dark pink = early lung cancer (LC) and light pink = late LC. Ellipsis drawn assuming multivariate t-distribution, confidence level = 0.95. Loadings for PCA of DIA-MS (**c**) and PEA (**d**); highly separated genes are labelled. Volcano plot of DIA-MS (**e**) and PEA (**f**) protein differential expression in lung cancer patients compared with control patients. Blue = decreased expression in lung cancer patients compared with controls, red = increased expression in lung

cancer patients compared with controls, light grey = not differentially expressed in lung cancer patients compared with controls. Criteria for differential expression marked by dashed line; proteins with FDR-adjusted p-value < 0.05 and Log2 fold change (LogFC) > 1 are considered differentially expressed. **e**, Heatmap of biological processes differentially regulated according to Gene set enrichment analysis (GSEA) of proteins within different patient sub-cohorts. Rows and columns ordered according to hierarchical clustering (Manhattan average) of Normalised enrichment score in sub-cohort compared with controls, identifying six clusters of pathways. Pathways GSEA clustering. SCLC = Small cell lung cancer, LUSC = Lung Squamous Cell Carcinoma, LC = All lung cancer, ELC = Early (Stage I-II) lung cancer patients only, LLC = Late (stage III-IV) lung cancer patients only, LUAD = Lung Adenocarcinoma patients only.

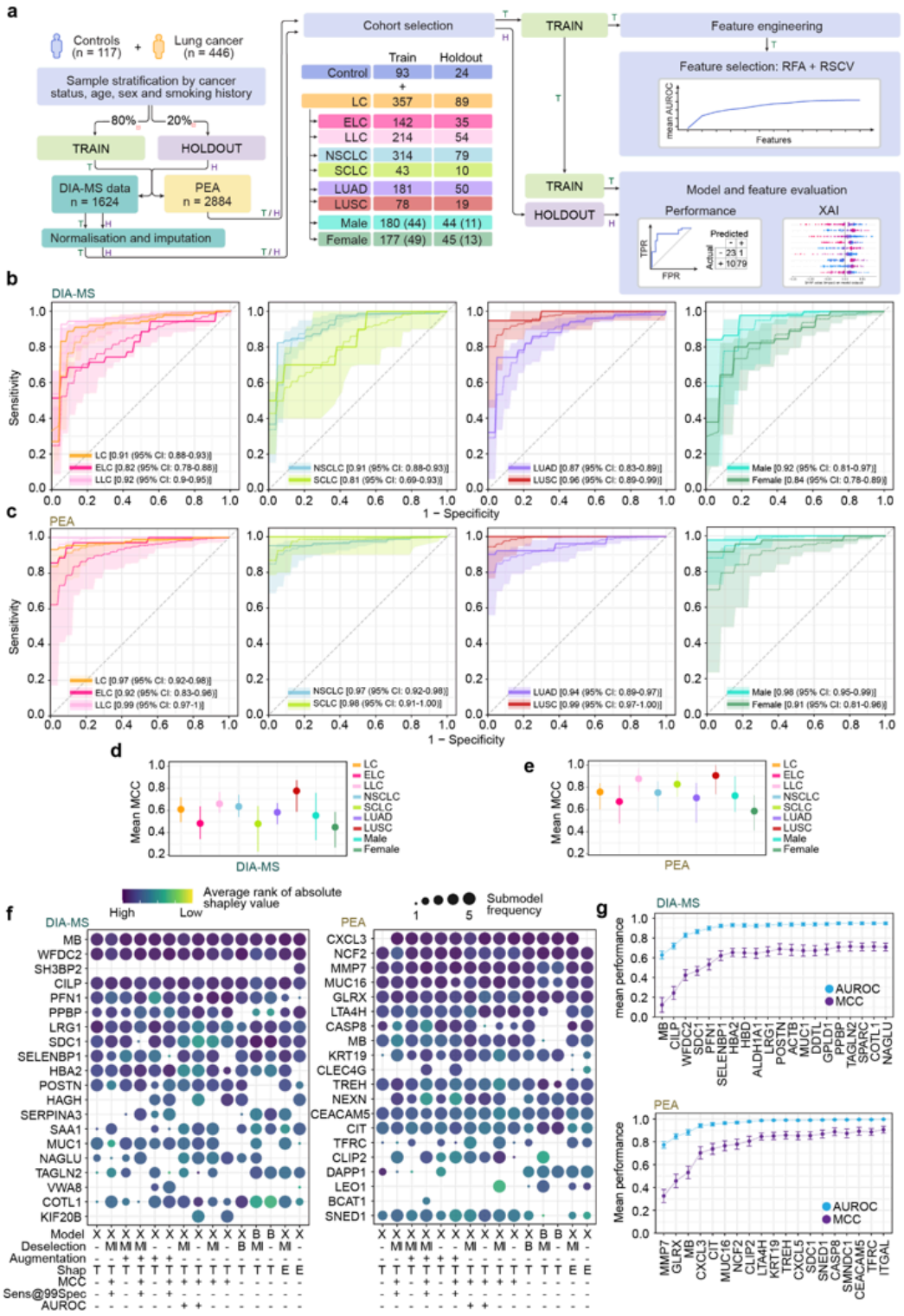


Figure 3

Explainable machine learning identifies focused plasma molecular signature for lung cancer. a, ML workflow used to classify lung cancer and subtypes using proteomic plasma signatures. Models were trained on 80% of stratified samples with a 20% unseen holdout test set. Sub-cohorts with the same sample stratification and train/holdout split applied to both DIA-MS and PEA datasets. Feature engineering was used to reduce feature lists, followed by feature selection using recursive feature

addition (RFA) with repeated stratified cross validation (RSCV) to produce a refined set of features. The model (a reduced set of features with model hyperparameters) was trained on the training data and applied to holdout data. Model performance consistency between training data and holdout data was used to assess generalisability of models, using explainable AI (XAI), artifacts and performances. T = Training data only, H = Holdout data only. LC = all lung cancer, NSCLC = non-small cell lung cancer, SCLC = small cell lung cancer, LUAD = lung adenocarcinoma, LUSC = lung squamous cell carcinoma, ELC = early lung cancer, LLC = late lung cancer. Area under the receiver operator characteristic curve (AUROC) of all DIA-MS models (**b**) and PEA models (**c**), with 95% confidence intervals (CI) across all models (shaded area) and mean performance (solid line) for all sub-cohort comparisons as indicated by colour key. Mean MCC with 95% CI for all sub-cohort models for DIA-MS (**d**) and PEA (**e**). **f**, Dot plot showing frequency of top 20 most consistent features in the 75 control vs lung cancer models evaluated for DIA-MS (left) and PEA (right) datasets. Genes are ordered from high to low (top to bottom) global mean of rank based on MAD (mean absolute deviation) of Shapley value. Nodes are coloured according to mean rank of MAD of Shapley value within five sub-models (mean MCC, mean AUROC, stability of MCC, ttest AUROC and ttest MCC) assessed with each feature engineering, selection and model combination (grid below plot). + indicates use in model training; - indicates absence in model training; X: XGBoost; B: Balanced random forest; MI: Mutual information; B: BorutaSHAP; T: TreeSHAP; E: EjectSHAP; Sens@99: Sensitivity at 99% specificity; AUROC, Sens@99 and MCC refer to multi-objective optimisation. **g**, Plot showing change in the representative DIA-MS (left) and PEA (right) model mean ROC AUC (blue) and mean MCC (purple) during RFA+RSCV as additional features are considered in the model.

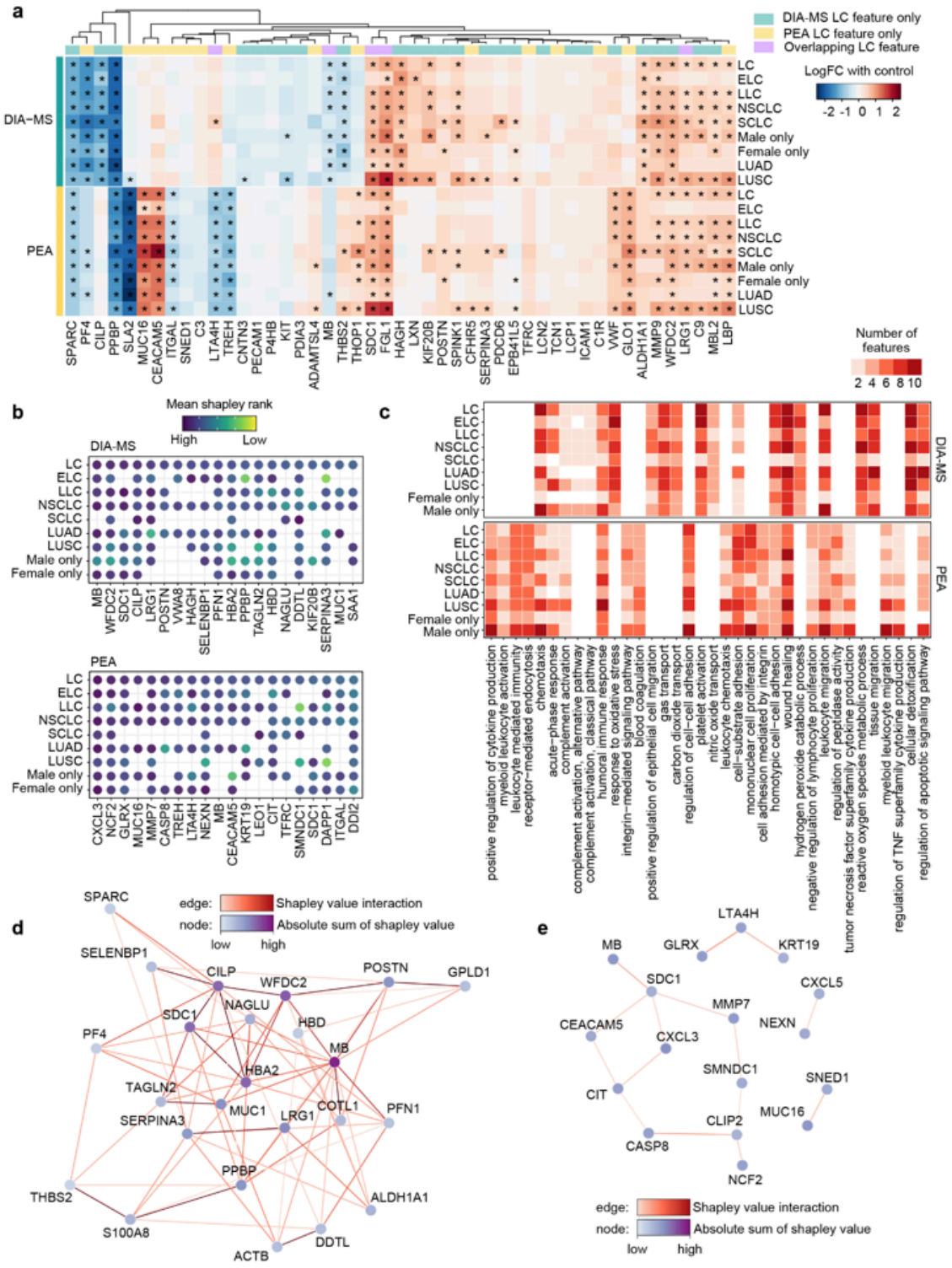


Figure 4

Explainable AI integrating Shapley values identifies key features and pathways associated with lung cancer detection using plasma proteins. a, Heatmap of differentially expressed (*; LogFC > 0.5 and FDR adjusted p-value < 0.05) features quantified by both DIA-MS (upper panel) and PEA (lower panel). Each row of heatmap is a different sub-cohort comparison with controls. Colour indicates Log2 fold change (LogFC) of sub-cohort compared with appropriate controls. Column colour shows if a lung cancer

feature was used by DIA-MS lung cancer models only (pale turquoise), PEA lung cancer models only (pale yellow) or used by both datasets (pale purple) for lung cancer models. **b**, Average rank of absolute Shapley value within individual models for DIA-MS (upper) and PEA (lower) for each sub-cohort comparison. High average rank = on average higher Shapley value within models. Missing points indicate that a feature was never selected in generalisable models. Features are ordered on x-axis, from left to right, according to average rank in all lung cancer models, followed by sub-cohorts. Only the top 20 features according to rank of absolute Shapley value from each sub-cohort are shown. **c**, Heatmap of biological pathways over-represented in all DIA-MS and PEA selected features. Top 20 features were taken from all models; top 20 pathways are shown. ORA was performed on all features. Each cell represents the number of features assigned to each pathway. Order of pathways is based on hierarchical clustering of number of features within each pathway to group similar pathways/patterns together. Shapley value interaction network built using Shapley values from a DIA-MS (**d**) and PEA (**e**) control vs lung cancer model. Shapley value interaction networks were built using only interactions between two proteins that explain 5% of the average absolute Shapley value within a model. Dark red edge = high relative Shapley interaction, pale pink = low relative Shapley value interaction. Nodes are coloured according to each protein's normalised absolute Shapley value. Dark purple node = high absolute Shapley value, light blue node = low absolute Shapley value.

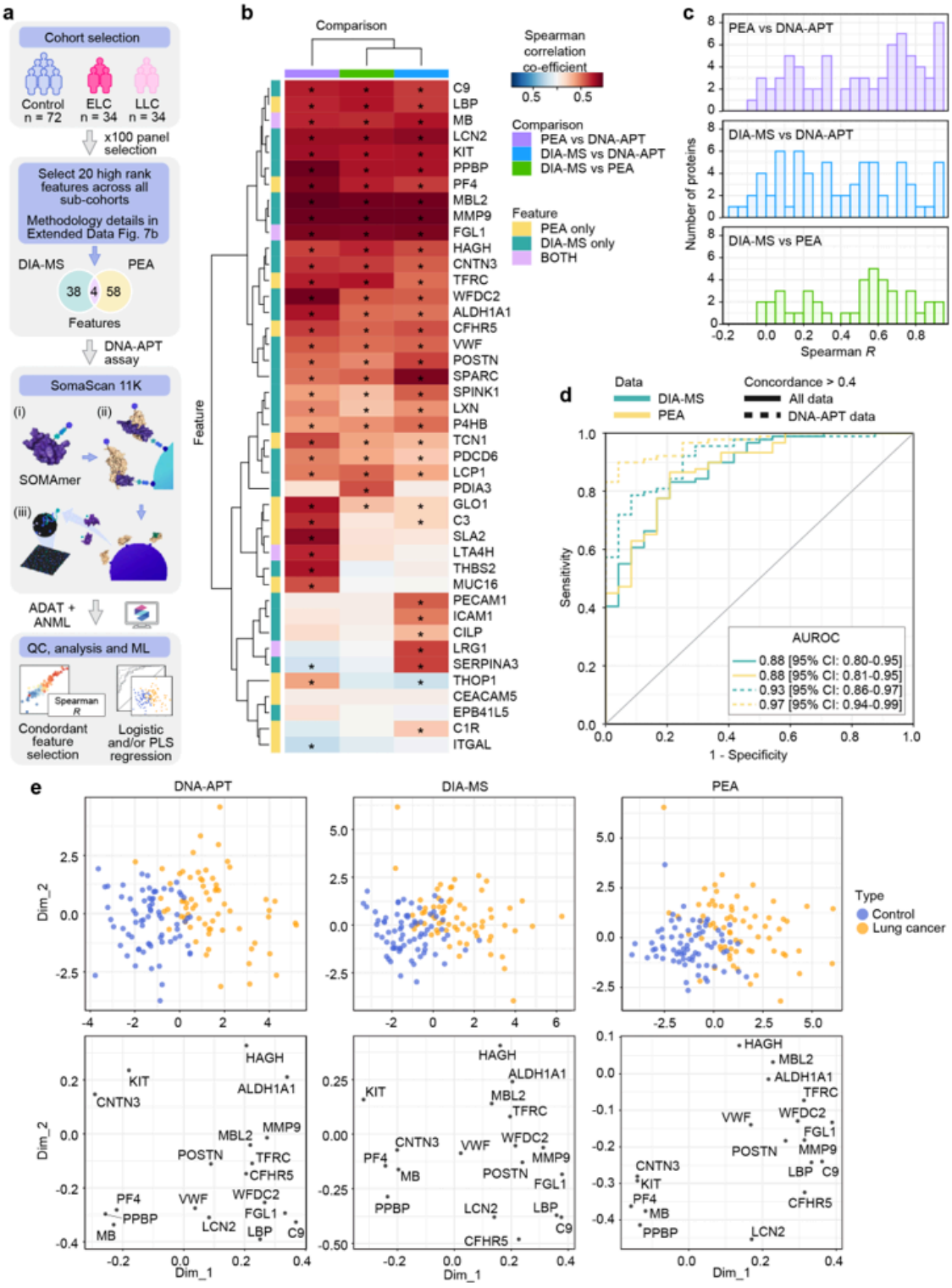


Figure 5

Characterisation of a platform agnostic lung cancer molecular signature from blood plasma across DIA-MS, PEA, and DNA-aptamer platforms. a, Workflow for DNA-APT characterisation of features selected by DIA-MS and PEA. Briefly, a cohort from the 563 individuals was selected to assess platform concordance. Twenty highly ranked features were then selected according to Shapley values across generalisable models were selected for each sub-cohort/platform. Additional explanation of DNA-

aptamer candidate selection available in Extended Data Fig. 7b. SomaScan 11K panel was used to quantify these 100 proteins and processed in line with SomaLogic recommendations. Following QC, concordance and ML performance was evaluated on complete DIA-MS and PEA cohorts using a logistic regression, and using partial least squares (PLS) regression across all three platforms. **b**, Hierarchical clustering (Manhattan average) of Spearman correlation co-efficient of protein quantification across the three platforms (DIA-MS, PEA and DNA-APT) for 42 proteins quantified by all three methods. Columns correspond to comparison (purple = PEA vs DNA-APT, green = DIA-MS vs PEA and blue = DIA-MS vs DNA-APT); rows correspond to ML features (lilac = feature in PEA and DIA-MS models, turquoise = feature in DIA-MS models only, yellow = feature in PEA models only); * = FDR adjusted p-value < 0.05 for significance of Spearman correlation. **c**, Histogram of Spearman correlation coefficients across the three comparisons. **d**, ROC curves of holdout performance for logistic regression models trained on DIA-MS (turquoise) or PEA (yellow) training data using either features with > 0.4 spearman correlation coefficient of protein quantification between DIA-MS and DNA-APT or PEA and DNA-APT for DIA-MS (n = 25) and PEA (n = 25) models respectively (dashed line), or > 0.4 spearman correlation coefficient across all three platforms (n = 18). AUROC for each model is shown in plot. **e**, PLS plots (upper) and corresponding PLS-loadings (lower) trained with features with concordance > 0.4 spearman correlation coefficient across all three platforms, using data from DNA-APT (left), DIA-MS (middle) or PEA > 0.4 (right). PLS plots show separability between control samples (blue) (n = 70) and lung cancer samples (orange) (n = 69). Loadings are labelled with gene names.

Supplementary Files

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

- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx
- SupplementaryTable3.xlsx
- SupplementaryTable4.xlsx
- ExtendedData.docx