Respiratory Research

RESEARCH

Open Access

# Multi-omics characterization of early chronic obstructive pulmonary disease

Bolun Li[1†], Jiangfeng Liu[2†], Yinghao Cao[2†], Yiyang Wang[1†], Sinan Wu[6†], Huiyuan Hu[3], Xingqi Xiao[1], Jiantao Hu[4], Qian Wang[4], Junlin Wu[4], Le Luo[5], Yong Liu[5], Qihao Tang[1], Yanjiang Xing[1], Tiantian Zhang[1], Jinyu Zhou[2], Lin Wang[2*], Juntao Yang[2*], Jing Wang[1*] and Chen Wang[1,6]

## Abstract

Chronic obstructive pulmonary disease (COPD) is projected to become the third leading cause of death globally by 2030, accounting for 71.9% of chronic respiratory diseases cases in 2019. Early COPD (ECOPD) diagnosis heavily relies on clinically monitoring of lung functions, with a strong influence from smoking exposures, which may not align well with disease progression. As such, the GOLD 2022–2024 guidelines emphasize the discovery of biological markers over clinical symptoms for early detection. This study explores the biological characteristics of ECOPD in a cohort of 176 adults from China Pulmonary Health Study, consisting 88 healthy controls (HC) and 88 clinically diagnosed ECOPD, matched for age, gender and smoking history. While lung function tests revealed differences between HC and ECOPD, no significant distinctions were observed in routine blood tests. Proteomics analysis identified 377 plasma proteins common to both groups, with low-intensity proteins driving group-specific differences. Univariable logistic regression and gene set enrichment analysis identified 248 proteins associated with ECOPD, particularly those involved in inflammation process. Validation in an independent cohort confirmed the association of 15 proteins with ECOPD. Metabolomics analysis of the plasma identified 1788 metabolites, 137 of which were found linked to ECOPD. Machine learning models indicated that a multi-omics approach provided the best predication of lung function ($R^2 = 0.74$), while proteomics alone effectively diagnosed ECOPD (AUC = 0.949). Similarity network fusion and clustering revealed two ECOPD subgroups: one by markers of inflammatory-immune response, and the other by the presence of those related to hemostasis or the vascular smooth muscle function. These findings underscore the potential of multi-omics integration in distinguishing ECOPD subgroups and predicting disease risk.

†Bolun Li, Jiangfeng Liu, Yinghao Cao, Yiyang Wang, Sinan Wu these authors contribute equally to this work.

*Correspondence:
Lin Wang
linwangZJU@hotmail.com
Juntao Yang
yangjt@pumc.edu.cn
Jing Wang
wangjing@ibms.pumc.edu.cn
[1]State Key Laboratory of Respiratory Health and Multimorbidity, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100005, China
[2]State Key Laboratory of Common Mechanism Research for Major Disease, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China
[3]First Clinical College, Xi'an Jiaotong University, Xi'an 710061, ShanXi, China
[4]Department of Pulmonary and Critical Care Medicine, Qixingguan District People's Hospital, Bijie 551799, Guizhou, China
[5]Department of Pulmonary and Critical Care Medicine, Dafang County People's Hospital, Bijie 551699, Guizhou, China
[6]Institute of Clinical Medical Sciences, Center of Respiratory Medicine, China–Japan Friendship Hospital, Beijing 100029, China

## Introduction

Chronic obstructive pulmonary disease (COPD) is characterized by persistent respiratory symptoms and airflow limitation, caused by abnormalities in the airway or alveolar, typically resulting from prolonged exposure to harmful particles or gases. According to the Global Burden of Disease, in 2019, there were 212 million COPD patients worldwide, contributing to 74.4 million global Disability-Adjusted Life Years (DALYs), which accounted for 71.9% of the burden from chronic respiratory diseases [1]. Although COPD may originate early in life, its clinical manifestations usually take years to emerge, making early identification a challenge [2]. Early COPD (ECOPD) is the initial phase in the pathogenesis of COPD, related to the primary mechanisms that eventually lead to COPD. Clinically, it is defined as individuals under 50 years of age with at least 10-pack-years of smoking exposure and a baseline forced expiatory volume in 1 s (FEV1)/forced vital capacity (FVC) below the lower limit of normal (LLN) [3]. However, only 24% of those diagnosed with ECOPD develop full blown COPD within a decade [4–5]. This indicates that the understanding of ECOPD remains incomplete.

The biological "early" stage of COPD, linked to the starting mechanism of the disease. Therefore, for effective prevention, it's crucial to detect biological early COPD, rather than the clinical early COPD that shows the first-noticed symptoms, functional impairments, or structural abnormalities. Omics approaches, such as transcriptomics, proteomics, and metabolomics, have been applied to uncover molecular mechanisms and define the biological characters of COPD. For example, transcriptomic biomarkers (e.g., ASAH1, CEBPD, FOXP1, TCF7) are linked to lung function [6–7], proteomic markers (e.g., TIMP1, BPIFB1, CNDP1) have been identified [8–10], and metabolomic biomarkers (e.g., sphingolipids) are found associate with exacerbation [6]. Unfortunately, these studies primarily use plasma samples from COPD patients at the late stage, limiting their utility for early screening. There are still no unified and feasible diagnostic biological characters for ECOPD.

In this study, we performed proteomics and metabolomics analysis on plasma samples from the China Pulmonary Health Study (CPHS) cohort [11] and a separate validation cohort to investigate the biological signatures of ECOPD. We identified proteomic and metabolomic signatures associated with clinically defined ECOPD, characterized by lung function and smoking exposures. ECOPD displayed pro-inflammatory proteomic and metabolomic features, including pathways related to leukocyte immunity and aspartate metabolism. Notably, a proteomic least absolute shrinkage and selection operator (LASSO) regression model outperformed other omics-based models in distinguishing healthy controls (HC) from ECOPD. Additionally, 20 individuals in HC group exhibited multi-omics features similar to those of ECOPD, suggesting a higher likelihood of progressing to ECOPD.

## Methods and materials

### Materials and reagents

Acetonitrile (ACN) and water were purchased from Fisher Chemical. 1,4-dithiothreitol (DTT), Uera and iodoacetamide (IAA) were purchased from Sigma-Aldrich. Sequencing Grade Modified Trypsin was purchased from Promega. BCA kit was purchased from Beyotime, and iRT kit was purchased from Biognosys. Protease Inhibitor Cocktail was purchased from Calbiochem. All other chemicals used were of analytical grade or higher.

### Study participants

The participants in discovery cohort were collected from the National Population Health Survey (CPHS) cohort. The CPHS cohort was collected between June 2012 and May 2015 covering participants from both the cities and counties in Guizhou Province. The inclusion and exclusion criteria of healthy controls were confined as the individuals younger than 50 years, baseline FEV1/FVC ratio over the LLN, without comorbidities, including cardiovascular diseases (coronary artery disease, hypertension, heart failure, arrhythmias), cerebrovascular diseases (stroke), and diabetes. The inclusion criteria of ECOPD patients were confined as the individuals younger than 50 years, baseline FEV1/FVC ratio falling below the LLN, with or without smoking exposure. Exclusion criteria of ECOPD encompassed individuals who had undergone thoracic, abdominal, or eye surgery within the previous three months, those admitted to hospital for cardiac conditions in the preceding month, individuals with a heart rate exceeding 120 beats per minute, those receiving antibacterial chemotherapy for tuberculosis, and females who were pregnant or breastfeeding. As a result, the discovery cohort consisting of 88 healthy controls and 88 ECOPD patients was collected, matched by age, gender and smoking exposure. The other validation cohort consisting of 35 healthy control and 35 ECOPD from Guizhou Province was independently collected using the same inclusion and exclusion criteria.

Written informed consent was procured from all participants, and the study was conducted in accordance with the Declaration of Helsinki and approved by the ethics review committee of Beijing Chaoyang Hospital, Capital Medical University (201002008) and Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences (065-2021), along with other collaborating institution, Bijie Qixingguan District People's Hospital (202101).

### Plasma collection and clinical information collection

After the blood routine test, the whole blood in EDTA Vacutainer tubes were immediately placed on ice and centrifuged (3000 r/min, 10 min at 4 °C) within 30 min. The separated plasma was stored at -80 °C until further use. Meanwhile, detailed characteristics (e.g., identification number, age, sex, and clinical indicators) were comprehensively collected.

### Sample Preparation for proteomics and data analysis

#### *Protein digestion and fractionation*

Multiple plasma samples were combined into a pooled mixture for spectral library construction. For each sample, the protein concentration was measured with the BCA protein assay kit (Thermo Fisher Scientific, 23227) according to the instruction provided by the manufacturer. 10 µL plasma was reduced with 5 mM dithiothreitol for 30 min at 56 °C and alkylated with 11 mM iodoacetamide for 15 min at room temperature in darkness. The alkylated samples were transferred to ultrafiltration tubes for FASP digestion. The samples were firstly replaced with 8 M urea for 3 times at 12,000 g at room temperature for 20 min, and then replaced with 100 mM TEAB for 3 times. Trypsin was added at 1:50 trypsin-to-protein mass ratio for digestion at 37 °C overnight. The peptide was recovered by centrifugation at 12,000 g for 10 min at room temperature, and repeated twice. Finally, the combined peptides were desalted by C18 SPE column. Additionally, the sample was then fractionated into fractions by high pH reverse-phase HPLC using Agilent 300 Extend C18 column (5 µm particles, 4.6 mm ID, 250 mm length). As to mixed samples for library construction, peptides were first separated with a gradient of 8–32% acetonitrile in 10 mM ammonium bicarbonate pH 9 over 60 min into 60 fractions. Then, the peptides were combined into 12 fractions and dried by vacuum centrifuging.

### DDA data acquisition and analysis

For data-dependent acquisition (DDA) —LC-MS/MS analysis, the iRT kit was added to all the fractions according to manufacturer's instructions. Next, the tryptic peptides were dissolved in solvent A (0.1% formic acid, 2% acetonitrile), directly loaded onto a home-made reversed-phase analytical column (25-cm length, 100 µm i.d.). Peptides were separated with a gradient from 4 to 24% solvent B (0.1% formic acid in 90% acetonitrile) over 60 min, 24–32% in 29 min and climbing to 80% in 3 min then holding at 80% for the last 3 min, all at a constant flowrate of 500 nL/min on an EASY-nLC 1200 UPLC system (Thermo Fisher Scientific). The separated peptides were analyzed in DDA mode by Q Exploris 480 (Thermo Fisher Scientific) with a nano-electrospray ion source.

For the annotation of DDA data, the resulting DDA data were processed with Spectronaut (v 15.0) to generate the spectral library. Tandem mass spectra were searched against the human SwissProt database (20376 entries, http://www.expasy.ch/sprot), which was concatenated with a reverse decoy database. Trypsin/P was specified as the cleavage enzyme, allowing up to 2 missing cleavages. In the calibration and main searches, the mass tolerance for precursor ions and fragment ions were set "Dynamic" and the correction factor were set 1. Carbamidomethyl on Cys was defined as a fixed modification, while acetylation of the protein N-terminal and oxidation of Met were defined as variable modifications. The false discovery rate (FDR) was adjusted to < 1% for both peptide-spectrum matches (PSMs) and proteins. All other settings were used by default unless otherwise noted.

### DIA data acquisition and analysis

For data-independent acquisition (DIA) —LC-MS/MS analysis, the iRT kit was added to all the samples according to manufacturer's instructions. The LC gradient was kept consistent with those in the spectral library building method. The separated peptides were analyzed in Q Exploris 480 (Thermo Fisher Scientific) with a nano-electrospray ion source. The full MS scan resolution was set to 60,000 for a scan range of 400–1,200 *m/z*. The data acquisition was performed in DIA mode. Each cycle contains one full scan followed by 70 DIA MS/MS scans with predefined precursor *m/z* range. The HCD fragmentation was performed at a normalized collision energy (NCE) of 28%. The fragments were detected in the Orbitrap at a resolution of 15,000. Fixed first mass was set as 20 *m/z*. Automatic gain control (AGC) target was set at 5E5. For the database search of DIA data, all DIA data were analyzed in Spectronaut (v15.0) with the same parameter applied in DDA spectral library construction, imported the established spectral library, and predicted the retention time of peptide segments through nonlinear correction.

### Proteomics data preprocessing and bioinformatic analysis

Proteins with a missing ratio over 50% in both ECOPD and healthy control groups were removed. Protein intensities were quantile normalized, $\log_2$-transformed, and missing values were imputed using the minimal value of entire dataset. In R 4.4.1, the glm function was used to perform logistic regression, adjusting for gender, age, and smoking exposure, to assess the association between normalized protein levels and ECOPD. For differentially protein expression analysis, the limma package in R was employed, using linear models. Adjusted $p$ values were calculated using the Benjamini & Hochberg correction. Significantly altered proteins were filtered based on adjusted $p$ values < 0.05 and absolute $\log_2$ fold

change > $\log_2$(1.3). Gene set enrichment analysis (GSEA) were carried out using the clusterProfiler package (version 4.2.2 in R 4.4.1), with gene ranked by log2 fold change.

### Sample Preparation for metabolomics and data analysis

#### *Metabolites extraction and untargeted metabolomics measurement*

10 µL plasma was extracted with 100 µL Methanol. The solution was kept at -20 ℃ for 10 min and the resulting mixture was transferred into an Eppendorf tube and spun down at 15,000 g for 30 min at 4 °C. The supernatant was taken for LC-MS analysis. LC was performed using a Vanquish UHPLC system (Thermo Fisher) and Xbridge BEH Amide HILIC column (Waters) with 25 min gradient from acetonitrile to pH 9.5 aqueous buffer [12]. LC was coupled by electrospray ionization (± 3.3 kV) to a Orbitrap Exploris 480 mass spectrometer (Thermo Fisher). Injected sample volume was 5 µL.

### Metabolomics data analysis

A reference metabolite spectral library consisting of 800 metabolites was constructed by running the metabolite standards. LC-MS raw data files (.raw) were separately converted to mzXML format using ProteoWizard [13]. To facilitate the process of metabolomic analysis, an automated pipeline with a graphical interface named as MetaPipe (https://github.com/bioinfo-ibms-pumc/MetaPipe) was developed by integrating several popular toolkits such as XCMS, CAMERA, NetID and MetaboAnalystR [14–17]. For this study, the key parameters of the XCMS module for peak picking were set as follows: method = "centWave", ppm = 5; snthr = 2; peakwidth = c [1, 4]. Peaks were then refined based on the raw data by in-house scripts and submitted to the CAMERA module for adducts filtration with default parameters. Last, metabolites were identified by the NetID module taking both the local reference metabolite library with RT information generated by LC-MS and the standard MS2 spectral library of metabolites from HMDB [18]. A manual independent double check was also performed according to the local reference metabolite library using El-Maven with ppm ≤ 5 and an retention time (RT) threshold of 0.5 min. Both peaks from manual selection and MetaPipe were merged to remove redundancy [19].

Metabolites with over 50% missing ratio in both ECOPD and healthy control groups were removed for the subsequent statistical analyses. The normalization was followed the strategy of MetaboAnalyst [17] using the median normalization and $\log_2$-tranformed as well as imputation of missing value with a tenth of the minimal value. The significantly altered metabolites were filtered using the criteria of adjusted p values less than 0.05 and variable importance in projection (VIP) scores larger than 1, derived from the Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) model. In addition, metabolomic pathway enrichment analysis was also performed employing MetaboAnalyst.

### Multi-omic bioinformatic analysis

The LASSO and elastic net algorithms were conducted using the glmnet package (version 4.1-8 in R 4.4.1), which utilizes a predefined log-scale grid search across a sequence of hyperparameter(s) and combines with cross-validation to identify the optimal regularization parameter. A 10-fold cross-validation was applied to calculate the average mean squared error (MSE). The hyperparameters for LASSO and elastic net analysis was set to the values that minimized the cross-validated MSE. The SNFtool package (version 2.3.1) was used for SNF analysis. Further differential expression (DEP) and gene set enrichment analysis (GSEA) were conducted as previously described.

### Statistical analysis

The quantitative data was analyzed using Graphpad Prism 9.0.0 (GraphPad Software, Inc, https://www.graphpad.com/). For continuous data, The Shapiro–Wilk or Kolmogorov–Smirnov test was used to determine the normally distribution of the data and the Levene test to test the equality of variance. Normally distributed data with equal variances were analyzed by Student t test, while non-normally distributed data or those with unequal variances were analyzed using Mann–Whitney U test. Categorical data were analyzed with the chi-squared ($\chi^2$) test. In addition, multiple hypothesis testing was corrected using the Benjamini–Hochberg (BH) procedure to adjust the $p$ values and control the FDR. Unless otherwise specified, a $p$ value < 0.05 was considered statistically significant.

## Results

### CPHS cohort characteristics

To explore the biological characteristics of ECOPD, a discovery cohort of 176 adults was selected from the China Pulmonary Health Study (CPHS). This cohort consisted of 88 healthy control (HC) and 88 clinically confirmed ECOPD patients, matched by age, gender, and smoking exposure. ECOPD was diagnosed based on a baseline FEV1/FVC ratio below the LLN. The cohort was predominantly male (55.7%), middle-aged (mean ± s.d.: 43.99 ± 2.82) and included a high proportion of smokers (48.9%) (Table 1). Compared to HC, the ECOPD group demonstrated significantly reduced lung function parameters (FEV1, $p < 0.001$; FEV1% predicted, $p < 0.001$; and FEV1/FVC, $p < 0.001$), while no significant differences were observed in other clinical laboratory tests. These results suggest that standard blood tests are insufficient

**Table 1** Demographic table of CPHS population

| | | Healthy control | Early COPD | *p* value |
|---|---|---|---|---|
| | | *N* = 88 | *N* = 88 | |
| Gender (%) | Male | 49 (55.7) | 49 (55.7) | 1 |
| | Female | 39 (44.3) | 39 (44.3) | |
| Age (year, mean (SD)) | | 43.99 (2.82) | 43.99 (2.82) | 1 |
| BMI (kg/m$^2$, mean (SD)) | | 23.96 (3.56) | 23.73 (2.98) | 0.648 |
| Cigarsmoker (%) | Smoker | 43 (48.9) | 43 (48.9) | 1 |
| | Non-smoker | 45 (51.1) | 45 (51.1) | |
| Packyr (year, mean (SD)) | | 8.60 (11.44) | 8.31 (10.59) | 0.86 |
| FEV1 (%, mean (SD)) | | 2.89 (0.52) | 2.49 (0.63) | < 0.001*** |
| FEV1%pred (%, mean (SD)) (SD)) | | 99 (0.12) | 84 (0.15) | < 0.001*** |
| FEV1/FVC (%, mean (SD)) | | 80.96 (4.58) | 66.44 (6.99) | < 0.001*** |
| WBC (10$^9$/L, mean (SD)) | | 6.79 (1.72) | 6.32 (1.65) | 0.071 |
| RBC (10$^9$/L, mean (SD)) | | 4.90 (0.49) | 4.83 (0.55) | 0.339 |
| HGB (g/L, mean (SD)) | | 149.20 (17.71) | 148.18 (19.30) | 0.716 |
| PLT (10$^9$/L, mean (SD)) | | 205.16 (57.74) | 205.65 (69.81) | 0.959 |
| NEU rate (%, mean (SD)) | | 60.77 (10.49) | 58.34 (12.39) | 0.163 |
| EOS rate (%, mean (SD)) | | 2.53 (2.00) | 2.95 (3.19) | 0.29 |
| FBG (mmol/L, mean (SD)) | | 5.32 (1.56) | 5.27 (1.06) | 0.796 |
| TG (mmol/L, mean (SD)) | | 1.78 (1.31) | 1.82 (1.51) | 0.85 |
| TCH (mmol/L, mean (SD)) | | 4.85 (0.87) | 4.75 (0.96) | 0.452 |
| HDL (mmol/L, mean (SD)) | | 1.13 (0.25) | 1.07 (0.22) | 0.085 |
| LDL (mmol/L, mean (SD)) | | 2.55 (0.62) | 2.54 (0.62) | 0.916 |

CPHS, China Pulmonary Health Study; FEV1, Forced expiratory volume in 1 s; FVC, Forced vital capacity; BMI, Body mass index; WBC, White blood cell count; RBC, Red blood cell count; HGB, Hemoglobin; PLT, Platelet; NEU rate, Neutrophil ratio; EOS rate, Eosin ratio; FBG, Fasting blood glucose; TG, Triglyceride; TCH, Total cholesterol; HDL, High-density lipoprotein; LDL, Low-density lipoprotein; SD, Standard Deviation; COPD, Chronic Obstructive Pulmonary Disease

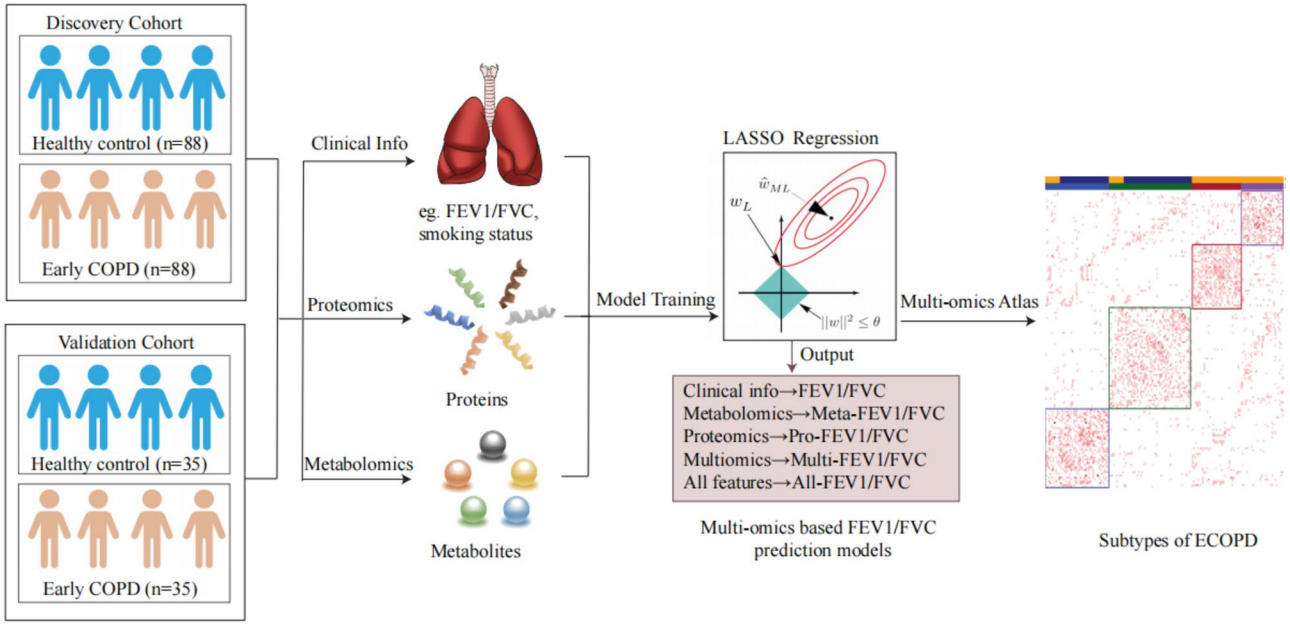


**Fig. 1** Schematic diagram of current study and analysis workflow

for differentiating ECOPD from HC. Furthermore, no strong correlation was found between lung function and smoking exposure, indicating that increased smoking exposure in individuals under 50 does not necessarily result in lung function impairment. Subsequently, we recruited another 35 HC and 35 ECOPD patients from Guizhou Province as a separate validation cohort. Proteomics and metabolomics analysis were performed in both cohorts to find the biological molecular characteristics of ECOPD (Fig. 1).

### Proteomic signatures of ECOPD

Mass spectrometry (MS)-based data independent acquisition (DIA) was employed to perform plasma proteomics analysis on the CPHS cohort. Of the 902 proteins measured in total, 377 were detected in both HC and ECOPD group. High-abundance proteins shared between both groups included ALB, IGLC2, IGKV3D-11, IGKC and APOA1. In contrast, low-abundance proteins varied: ANPEP, CST3, LPA, VWF and ISLR are low in HC, while MSN, SOD3, CHL1, LRP1 and ISLR are low in ECOPD, highlighting differences between the groups (Fig. 2A, Table S1). To analyze protein signatures, logistic regression, adjusted for gender, age, and smoking exposure, was used to assess the association between normalized protein levels and ECOPD. This identified 248 proteins significantly linked to ECOPD (Table S2).

A comparison between smokers and non-smokers in the CPHS cohort revealed 17 differentially expressed proteins (DEPs). Of these, 13 were associated with ECOPD, with KRT13 and DCD consistently identified in both discovery and validation cohorts, suggesting enhanced airway epithelial differentiation and antimicrobial activity in smokers (Fig. 2B and Figure S1). Gene-set enrichment analysis (GSEA) of the ECOPD linked proteins revealed that proteins involved in "leukocyte-mediated immunity", such as CLU, AZGP1, and F2, were upregulated in ECOPD, indicating increased inflammation (Fig. 2C and Figure S2). To validate this finding, we also performed plasma proteomics analysis in the validation cohort using the same MS-based DIA method. Notebably, 15 proteins remained significantly associated with ECOPD after adjusting for baseline factors in both discovery

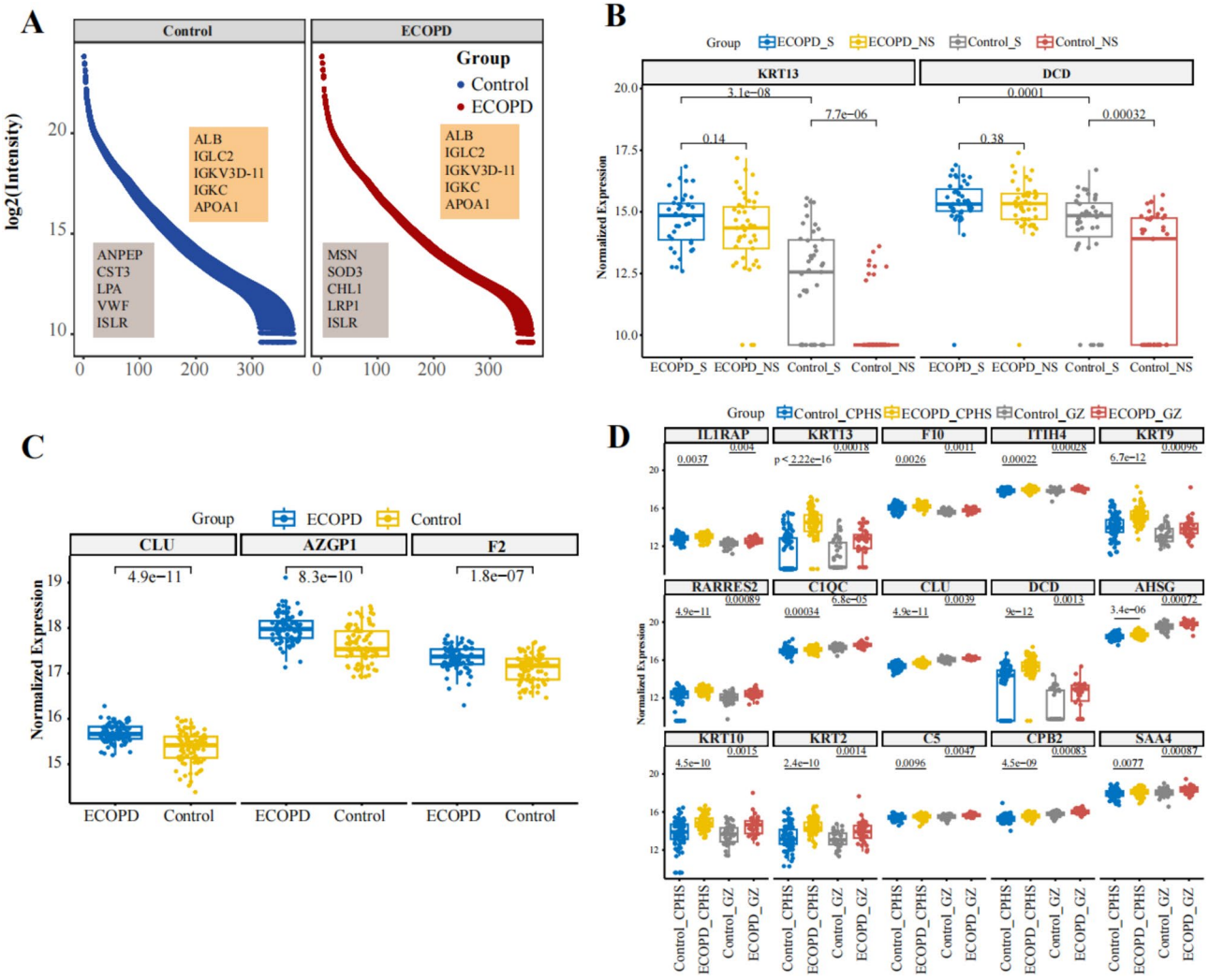


**Fig. 2** Proteomic signatures of biological ECOPD. (**A**) Overview of the proteomic profiles of CPHS populations. Protein abundances from plasma of ECOPD (red) and HC (blue) are shown, with quantification using $\log_2$ transformed quantile normalized intensity. The highest- and lowest-abundance proteins are highlighted in the boxed region. (**B**) Expression levels of KRT13 and DCD in both the CPHS cohort and validation cohort. (**C**) Expression changes of CLU, AZGP1 and F2 in HC and ECOPD in the CPHS cohort. (**D**) Validation of observed protein expression changes in the independent validation cohort

and validation cohort. These included keratin proteins (KRT2, KET9, KRT10, KRT13), complement proteins (C1QC, C5), and pro-inflammatory factors (IL1RAP, ITIH4, RARRES2, SAA4) (Fig. 2D).

### Metabolomic signatures of ECOPD

Next, untargeted metabolomics analysis was performed to assess the metabolic changes in the plasma of HC and ECOPD patients. To facilitate high-throughput analysis, an integrated pepeline called MetaPipe was developed ( https://github.com/bioinfo-ibms-pumc/MetaPipe). After manually validating the peaks generated by MetaPipe against a local reference metabolite library, 1,788 metabolites were identified in both HC and ECOPD groups within the discovery cohort (Table S3). To further explore metabolomic alterations, variable importance in projection (VIP) scores, derived from the OPLS-DA model, were combined with a t-test to identify 137 metabolites significantly associated with ECOPD (adjusted $p < 0.05$, VIP score $\geq 1$, Table S3). Pathway enrichment analysis revealed an upregulation in amino acid metabolism such as alanine and asparagine (Fig. 3A and Figure S3). In the validation cohort, four metabolites: tryptophan, adrenic acid, 2-hydroxyethanesulfonate, and lysine exhibited a similar decreasing trend in ECOPD compared to HC in both the discovery and validation cohorts (Fig. 3B).

### FEV1/FVC prediction models

To investigate the relationship between lung function and biological features, machine learning models were developed to predict baseline FEV1/FVC using proteomic dataset, metabolic dataset, clinical features or their combinations. These models included: Meta-FEV1/FVC based on 137 associated metabolites, Pro-FEV1/FVC based on 248 associated proteins, Multi-FEV1/FVC using both metabolites and proteins, All-FEV1/FVC incorporating metabolites, proteins and 12 clinical features as a comprehensive model. To identify key features and reduce multicollinearity among the features, the LASSO and elastic net (ENET) algorithms with 10-fold cross-validation were applied, generating sparse models for each category. In metabolomics analysis, LASSO and ENET performed similarly. However, in proteomics and multi-omics analysis, LASSO outperformed ENET (Fig. 4A). Therefore, LASSO models were used for subsequent analysis.

Single-omic models (metabolomics or proteomics) provided good FEV1/FVC prediction, with $R^2$values of 0.62 for the metabolites-based model (Meta-FEV1/FVC) and 0.69 for the proteins-based model (Pro-FEV1/FVC). Combing metabolomics and proteomics improved prediction accuracy ($R^2$ = 0.72, Multi- FEV1/FVC), and integrating clinical data achieved the best performance ($R^2$ = 0.74, All- FEV1/FVC). This indicates a strong association between multi-omics data and lung function (Fig. 4B). Regarding the area under curve (AUC) of receiver operating characteristic (ROC) curve among all the FEV1/FVC models, the one using only proteomics data performed best in identifying ECOPD (Figure S4), highlighting the robustness of proteomics for ECOPD diagnosis. This might be because metabolite levels fluctuate rapidly,

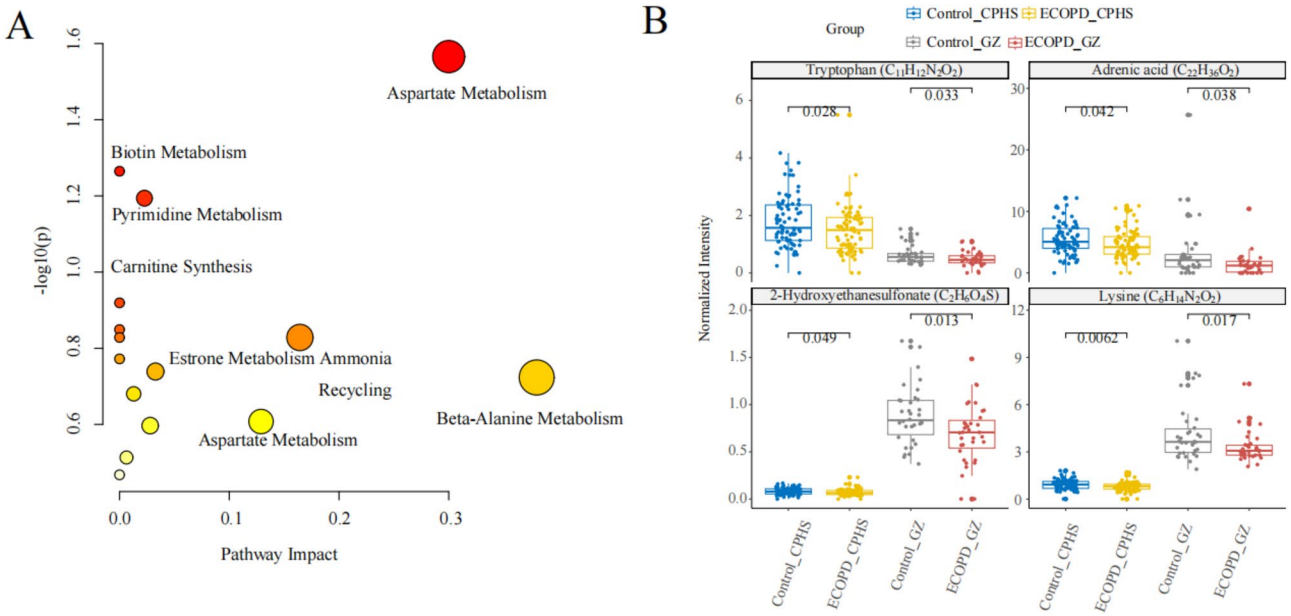


**Fig. 3** Metabolic characterizations of biological ECOPD. (**A**) Pathway enrichment analysis of metabolites related to ECOPD. (**B**) Changes in tryptophan, adrenic acid, 2-hydroxyethanefulfonate and lysine in the validation cohort

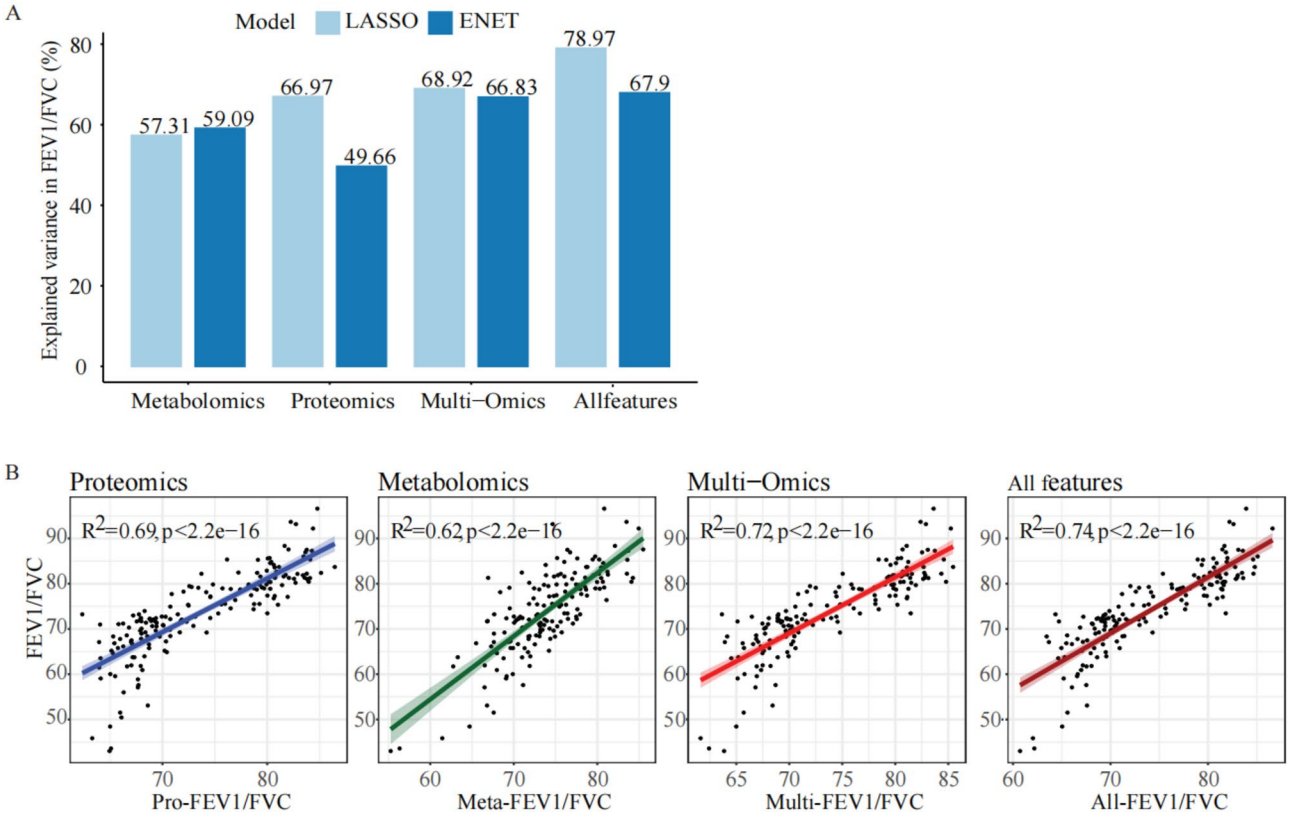


**Fig. 4** Multi-omics analysis for lung function prediction and ECOPD diagnosis. (**A**) Performance comparison of LASSO and ENET algorithms in different FEV1/FVC prediction models. (**B**) Correlation between predicted FEV1/FVC in different models and measured FEV1/FVC

as they are influenced by both disease state and environmental or genetic factors.

### Multi-omics atlas revealed the biological subtypes of ECOPD

To comprehensively characterize the biological signatures of ECOPD, proteomics and metabolomics data from the China Pulmonary Health Study (CPHS) cohort were integrated using similarity network fusion (SNF). Sample-by-sample similarity networks for each omics platform were calculated and iteratively updated to create a fused similarity network reflecting multi-omics information (Fig. 5A). This process identified four distinct subgroups (Fig. 5B). ECOPD patients clustered into Groups 1 and 2, alongside 20 healthy controls, demonstrating significant similarities in biological characteristics with ECOPD (Fig. 5B).

Further differential expression (DEP) and gene set enrichment analysis (GSEA) indicated that SNF-cluster 1 was characterized by a high expression of acute inflammation-related proteins (e.g., OSMR, SAA1, VCAM1, CD163, ICAM1) and proteins involved in the aminoglycan metabolic process (e.g., NAGLU, CTBS, ITIH3) (Fig. 5C, Table S4), suggesting an infection and inflammation profile. In contrast, SNF-cluster 2 was enriched in proteins associated with platelet activation and blood coagulation (e.g., FERMT3, TLN1), indicating a coagulation-active ECOPD subgroup.

Differential metabolite analysis revealed that most metabolites were overexpressed in SNF-cluster 2, particularly those linked to vascular smooth muscle contraction and galactose metabolism (e.g., 14,15-EET and stachyose) (Fig. 5D, Table S5). Only a few metabolites, such as $C_{19}H_{41}O_6P$ and $C_{37}H_{72}O_5$, were highly expressed in SNF-cluster 1. The 20 healthy controls in SNF-clusters 1 and 2, with metabolomic and proteomic profiles resembling those of ECOPD, were considered at risk of developing the disease.

Comparisons of FEV1/FVC, Omics-based FEV1/FVC, Multi-FEV1/FVC, and All-FEV1/FVC among at-risk populations and HC in in SNF-clusters 3 and 4 (Fig. 5E) revealed that the FEV1/FVC values for those at risk fell between those of healthy controls and ECOPD patients. This indicates that these individuals share similar biological characteristics with ECOPD, suggesting a heightened likelihood of developing the condition.

In summary, multi-omics integration can stratify ECOPD into two groups: one characterized by infection and inflammation features, and the other by coagulation and vascular smooth muscle contraction features.

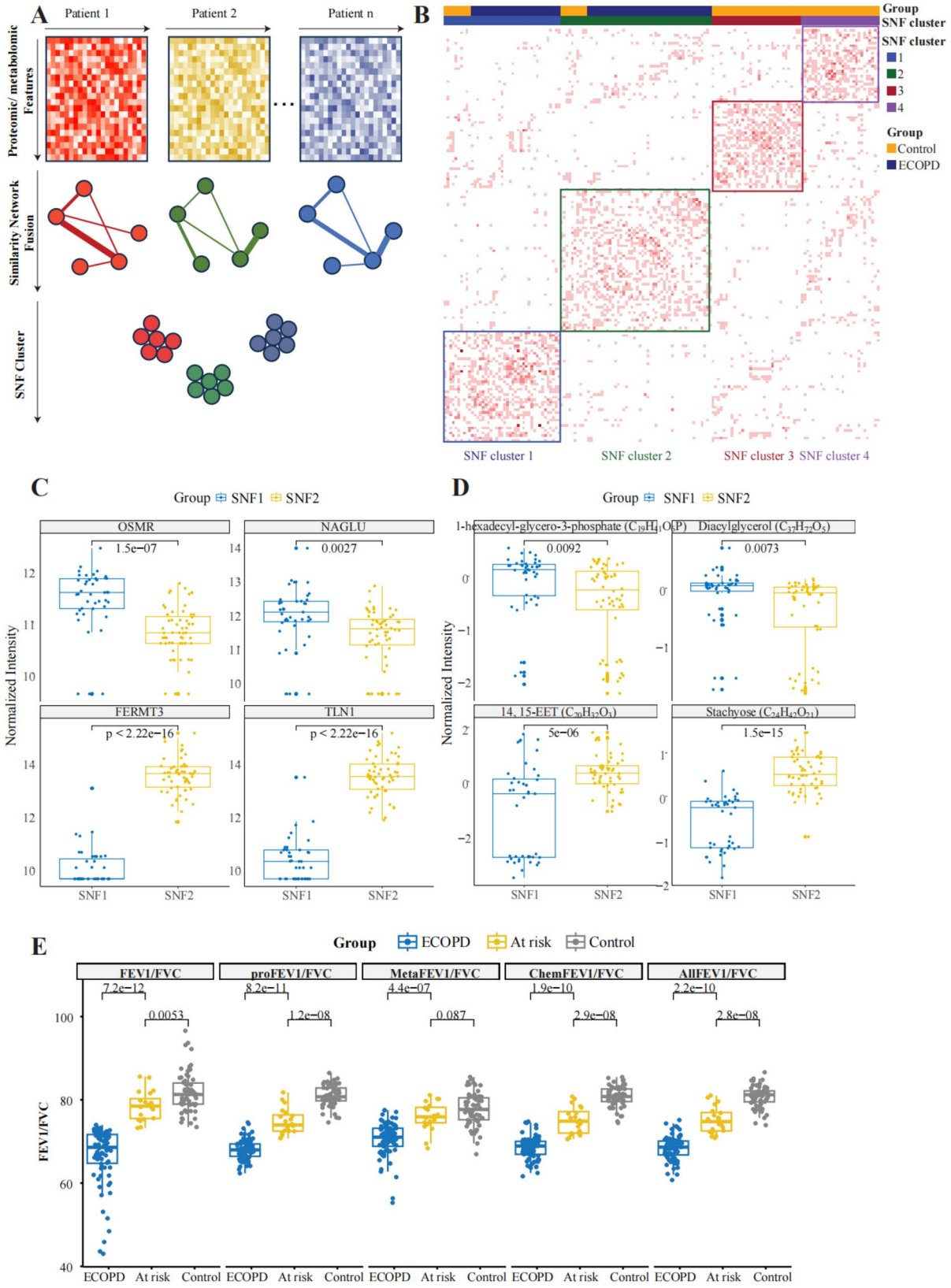


**Fig. 5** Multi-omics analysis reveals ECOPD subtypes. (**A**) SNF analysis of metabolomics and proteomics data from CPHS. (**B**) Cluster analysis reveals four subgroups within the CPHS cohort. (**C**) Proteins highly expressed in SNF-cluster1 and SNF-cluster2. (**D**) Elevated metabolites in SNF-cluster1 and SNF-cluster2. (**E**) FEV1/FVC ratios for HC, ECOPD and at-risk individuals

Additionally, multi-omics-based FEV1/FVC metrics can further indicate the potential risk for ECOPD. The identified biomarkers and molecular pathways offer promising targets for early diagnosis and personalized treatment strategies, ultimately improving patient outcomes. Our findings could aid in screening individuals with ECOPD for early intervention.

## Discussion

In recent years, early COPD (ECOPD) has been defined as occurring in individuals under 50 years old with a smoking history of ≥ 10 pack-years and a baseline FEV1/FVC ratio below the lower limit of normal (LLN). The 2023 Global Initiative for COPD [20] report emphasized that ECOPD primarily refers to the biological early stages of the disease. However, current diagnostic criteria focus predominantly on lung function and exposure, overlooking important biological characteristics. This study offers critical insights into the biological features of ECOPD through multi-omics integration, including proteomics and metabolomics.

Most ECOPD patients showed a significant decline in lung function parameters (FEV1, FEV1% predicted, and FEV1/FVC) compared to healthy controls. However, disease progression is also influenced by factors such as smoking status, physiological traits, and clinical performance. To explore these variations, our findings highlight distinct molecular signatures between ECOPD patients and healthy controls, demonstrating the value of multi-omics data in improving the understanding and diagnosis of ECOPD. We identified 248 proteins associated with ECOPD, many of which are linked to inflammation-related pathways. The validation of specific proteins, such as keratins, complement proteins, and pro-inflammatory factors, in an independent cohort supports their potential as biomarkers for ECOPD detection.

Metabolomic profiling revealed changes in metabolites related to aspartate metabolism, indicating that metabolic dysregulation is a hallmark of ECOPD and offering new therapeutic possibilities. Multi-omics models significantly outperformed single-omics models in predicting lung function ($R^2 = 0.74$), highlighting the added value of integrating diverse biological data. Key proteins and metabolites associated with FEV1/FVC variability show strong diagnostic potential. Through Similarity Network Fusion (SNF) and clustering, we identified two distinct ECOPD subgroups: one characterized by infection and inflammation, and the other by coagulation and vascular smooth muscle contraction, suggesting potential for tailored treatment approaches.

This study identifies multi-omics signatures of ECOPD, develop a proteomics-based model for FEV1/FVC prediction, and perform multi-omics SNF-cluster analyses. It provides a comprehensive understanding of ECOPD and proposes a novel definition for individuals at risk based on multi-FEV1/FVC models. Our data show that plasma-based multi-omics studies mirror lung tissue pathogenesis, with proteomics revealing airway inflammation markers such as leukocyte-mediated immunity, complement proteins, pro-inflammatory factors, and keratins. These findings align with previous studies describing COPD as characterized by airway inflammation, small airway remodeling, and emphysema [21–22]. The roles of inflammation and epithelial development in airway remodeling are consistent with our findings of enriched inflammasome and epithelial cell differentiation pathways.

Untargeted metabolomics offers unbiased insights into metabolic changes related to physiological and pathological states. However, challenges remain, such as the resolution of mass spectrometry and metabolite complexity. The MetaPipe pipeline, which integrates multiple metabolite analysis tools, streamlines data processing and putative metabolite identification. Aspartate metabolism was particularly highlighted in ECOPD, suggesting its involvement in systemic inflammation and impaired pulmonary function.

Despite the strength of our findings, limitations include the relatively small cohort size, which may affect generalizability, and the cross-sectional design, which precludes causal inferences. Larger, multi-center longitudinal studies are necessary to validate these findings and establish temporal relationships between molecular changes and disease progression. In addition, our current findings were limited to Chinese population, which need further validations on cohorts from multiple ethnic populations with different gene background. Future research should focus on functionally validating the identified biomarkers and incorporating additional omics layers for a more comprehensive understanding of ECOPD pathogenesis.

In conclusion, this study demonstrates the power of multi-omics integration in elucidating the complex biology of ECOPD. The identified biomarkers and molecular pathways offer promising targets for early diagnosis and personalized treatment strategies, ultimately improving patient outcomes. Our findings could facilitate the early screening and intervention for individuals at risk of ECOPD.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12931-025-03250-5.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

**Acknowledgements**
We acknowledge the use of mass spectrometry at the Core Facility of Instrument, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, School of Basic Medicine Peking Union Medical College. We also thanks to all members of the Wang laboratories for the discussion.

**Author contributions**
B.L., S.W., L.W., J.Y, J.W. and C.W. designed the study and interpreted the data. B.L., J.L., Y.C., and J. Z. analyzed the data and wrote the first draft. Y.W. reanalyzed the data for revision. Y. W., S. W., T. Z. and L. W revised the manuscript. B.L., H.H., X.X. and Q.T. collected the samples and baseline data of validation cohort. S. W., J.H., Q.W., J.W., L.L. and Y.L. recruited the volunteers in local hospitals. S.W. kindly provided the samples and baseline data of CPHS cohort. # These authors contribute equally to this work.

**Funding**
Supported by Major Program of the National Natural Science Foundation of China, No. 82090010, 82090011 (to C.W.), the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences, No. 2021-RC350-008 (to L.W.), Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences, No. 2023-I2M-2-001 (to C.W.) and No. 2021-I2M-1-049 (to J.W.), Beijing Natural Science Foundation, NO. Z240014 (to L. W.).

**Data availability**
No datasets were generated or analysed during the current study.

**Code availability**
The code of this article can be shared upon reasonable requests to Dr. Jing Wang (wangjing@ibms.pumc.edu.cn).

## Declarations

**Ethics declarations**
The study was conducted in accordance with the Declaration of Helsinki and approved by the ethics review committee of Beijing Chaoyang Hospital, Capital Medical University (201002008) and Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences (065-2021), along with other collaborating institution, Bijie Qixingguan District People's Hospital (202101).

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

**Clinical trial number**
Not applicable.

Received: 2 November 2024 / Accepted: 21 April 2025
Published online: 28 April 2025

**References**

1. Adeloye D, Song P, Zhu Y, Campbell H, Sheikh A, Rudan I, Unit NRGRH. Global, regional, and National prevalence of, and risk factors for, chronic obstructive pulmonary disease (COPD) in 2019: a systematic review and modelling analysis. Lancet Respir Med. 2022;10:447–58.
2. Soriano JB, Polverino F, Cosio BG. What is early COPD and why is it important. Eur Respir J. 2018;52(6):1–11.
3. Martinez FJ, Han MK, Allinson JP, Barr RG, Boucher RC, Calverley PMA, Celli BR, Christenson SA, Crystal RG, Fagerås M, Freeman CM, Groenke L, Hoffman EA, Kesimer M, Kostikas K, Paine R 3rd, Rafii S, Rennard SI, Segal LN, Shaykhiev R, Stevenson C, Tal-Singer R, Vestbo J, Woodruff PG, Curtis JL, Wedzicha JA. At the root: defining and halting progression of early chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2018;197(12):1540–51.
4. Çolak Y, Afzal S, Nordestgaard BG, Lange P, Vestbo J. Importance of early COPD in young adults for development of clinical COPD: findings from the Copenhagen general population study. Am J Respir Crit Care Med. 2021;203(10):1245–56.
5. Yang W, Li F, Li C, Meng J, Yang Y. Focus on early COPD: definition and and early lung development. Int J Chon Obstruct Pulmon Dis. 2021;16:3217–28.
6. Cruickshank-Quinn CI, Jacobson S, Hughes G, Powell RL, Petrache I, Kechris K, Bowler R, Reisdorph N. Metabolomics and transcriptomics pathway approach reveals outcome-specific perturbations in COPD. Sci Rep. 2018;8:17132.
7. Titz B, Sewer A, Schneider T, Elamin A, Martin F, Dijon S, Luettich K, Guedj E, Vuillaume G, Ivanov NV, Peck MJ, Chaudhary NI, Hoeng J, Peitsch MC. Alterations in the sputum proteome and transcriptome in smokers and early-stage COPD subjects. J Proteom. 2015;128:306–20.
8. Nicholas BL, Skipp P, Barton S, Singh D, Bagmane D, Mould R, Angco G, Ward J, Guha-Niyogi B, Wilson S, Howarth P, Davies DE, Rennard S, O'Connor CD, Djukanovic R. Identification of Lipocalin and Apolipoprotein A1 as biomarkers of chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2010;181:1049–60.
9. Zarei S, Mirtar A, Morrow JD, Castaldi PJ, Belloni P, Hersh CP. Subtyping chronic obstructive pulmonary disease using peripheral blood proteomics. Chronic Obstr Pulm Dis. 2017;4:97–108.
10. Serban KA, Pratte KA, Bowler RP. Protein biomarkers for COPD outcomes. Chest. 2021;159(6):2244–53.
11. Wang C, Xu J, Yang L, Xu Y, Zhang X, Bai C, Kang J, Ran P, Shen H, Wen F, Huang K, Yao W, Sun T, Shan G, Yang T, Lin Y, Wu S, Zhu J, Wang R, Shi Z, Zhao J, Ye X, Song Y, Wang Q, Zhou Y, Ding L, Yang T, Chen Y, Guo Y, Xiao F, Lu Y, Peng X, Zhang B, Xiao D, Chen CS, Wang Z, Zhang H, Bu X, Zhang X, An L, Zhang S, Cao Z, Zhan Q, Yang Y, Cao B, Dai H, Liang L, He J. China pulmonary health study G. Prevalence and risk factors of chronic obstructive pulmonary disease in China (the China pulmonary health [CPH] study): a National cross-sectional study. Lancet. 2018;391:1706–17.
12. Wang L, Xing X, Chen L, Yang L, Su X, Rabitz H, Lu W, Rabinowitz JD. Peak annotation and verification engine for untargeted LC-MS metabolomics. Anal Chem. 2018;91(3):1838–46.
13. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak MY, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol. 2012;30:918–20.
14. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem. 2006;78:779–87.
15. Kuhl C, Tautenhahn R, Böttcher C, Larson TR, Neumann S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal Chem. 2012;84:283–9.
16. Chen L, Lu W, Wang L, Xing X, Chen Z, Teng X, Zeng X, Muscarella AD, Shen Y, Cowan A, McReynolds MR, Kennedy BJ, Lato AM, Campagna SR, Singh M, Rabinowitz JD. Metabolite discovery through global annotation of untargeted metabolomics data. Nat Methods. 2021;18:1377–85.
17. Pang Z, Chong J, Zhou G, de Lima Morais DA, Chang L, Barrette M, Gauthier C, Jacques P, Li S, Xia J. MetaboAnalyst 5.0: narrowing the gap between Raw spectra and functional insights. Nucleic Acids Res. 2021;49:W388–96.
18. Wishart DS, Guo A, Oler E, Wang F, Anjum A, Peters H, Dizon R, Sayeeda Z, Tian S, Lee BL, Berjanskii M, Mah R, Yamamoto M, Jovel J, Torres-Calzada C, Hiebert-Giesbrecht M, Lui VW, Varshavi D, Varshavi D, Allen D, Arndt D, Khetarpal N, Sivakumaran A, Harford K, Sanford S, Yee K, Cao X, Budinski Z, Liigand J, Zhang L, Zheng J, Mandal R, Karu N, Dambrova M, Schiöth HB, Greiner R, Gautam V. HMDB 5.0: the human metabolome database for 2022. Nucleic Acids Res. 2022;50:D622–31.
19. Agrawal S, Kumar S, Sehgal R, George S, Gupta R, Poddar S, Jha A, Pathak S. El-MAVEN: A Fast, Robust, and User-Friendly Mass Spectrometry Data Processing Engine for Metabolomics. *Methods Mol Biol* 2019; 1978: 301–321.
20. Agusti A, Celli BR, Criner GJ, Halpin D, Anzueto A, Barnes P, Bourbeau J, Han MK, Martinez FJ, de Oca MM, Mortimer K, Papi A, Pavord I, Roche N, Salvi S, Sin DD, Singh D, Stockley R, Lopez Varela MV, Wedzicha JA, Vogelmeier CF. Global initiative for chronic obstructive lung disease 2023 report: GOLD executive summary. Eur Respir J 2023.
21. Belgacemi R, Luczka E, Ancel J, Diabasana Z, Perotin JM, Germain A, Lalun N, Birembaut P, Dubernard X, Merol JC, Delepine G, Polette M, Deslee G, Dormoy

V. Airway epithelial cell differentiation relies on deficient Hedgehog signalling in COPD. EBioMedicine. 2020;51:102572.
22. Bodas M, Moore AR, Subramaniyan B, Georgescu C, Wren JD, Freeman WM, Brown BR, Metcalf JP, Walters MS. Cigarette smoke activates NOTCH3 to promote goblet cell differentiation in human airway epithelial cells. Am J Respir Cell Mol Biol. 2021;64:426–40.

## Publisher's note