nature communications

Article

https://doi.org/10.1038/s41467-026-68972-6

Plasma proteome mediates the associations between air pollution exposure and disease risk

Received: 8 May 2025

Accepted: 19 January 2026

Published online: 31 January 2026

Check for updates

Wenran Li[1,5], Kaixuan Li[1,5], Puchen Zhou[1], Yingyu Cheng[1], Yiran Zhao[1,2], Xia Meng[3], Yue Niu[3], Geng Zong[1], Guoqing Zhang[4], Haidong Kan[3] ✉ & Sijia Wang[1] ✉

Air pollution has been increasingly linked to a wide range of diseases, yet the underlying biological mechanisms remain poorly understood. Here, we systematically investigate how circulating proteins mediate the health effects of multiple air pollutants and quantify individual susceptibility. Using large-scale data from the UK Biobank, we identify 30 diseases significantly associated with air pollution exposure. We uncover 1,089 pollutant-associated proteins enriched in immune pathways, among which 296 significantly mediate disease risk. Notably, the mediating pathways of air pollution vary by disease types, involving MAPK signaling in cardiovascular diseases, innate immunity in immune disorders, carbohydrate metabolism in metabolic diseases, and cell proliferation in respiratory conditions. Globally, the proteome mediates 23.69% of the association between air pollution and diseases. Based on these mediators, we develop an Air Pollution Protein Risk Score (APPRS), which shows robust associations with disease risk in the UKBB and two external validation cohorts. APPRS also improves disease prediction when integrated into baseline models. Collectively, our study highlights the central role of circulating proteins in mediating the health impacts of air pollution and introduces APPRS as a tool for personalized risk assessment and precision public health interventions.

Air pollution, including particulate matter ($PM_{2.5}$, PM_{10}) and nitrogen oxides (NO_2, NO_x), is widely recognized as an environmental risk factor for multiple diseases, linked to respiratory, cardiovascular, immune, and metabolic health[1–21]. Epidemiological studies have extensively studied the associations between air pollutants and adverse health effects over the past decades, providing evidence of increased morbidity and mortality through various physiological pathways[22–24]. However, despite significant evidence linking air pollution to numerous health outcomes[25,26], the underlying molecular mechanisms that mediate the effects of air pollution on human health remain largely unexplored[27].

Previous research has identified inflammation, oxidative stress, immune dysregulation, epigenetic alterations, and endocrine disruption as shared pathogenic pathways in pollution-related diseases[28]. To further investigate the molecular mechanisms underlying these processes, recent studies have begun integrating exposure data with

[1]Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China. [2]State Key Laboratory of Genetic Engineering, Human Phenome Institute and Zhangjiang Fudan International Innovation Center, Fudan University, Shanghai, China. [3]School of Public Health, Key Lab of Public Health Safety of the Ministry of Education and NHC Key Lab of Health Technology Assessment, Fudan University, Shanghai, China. [4]Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China. [5]These authors contributed equally: Wenran Li, Kaixuan Li. ✉e-mail: kanh@fudan.edu.cn; wangsijia@sinh.ac.cn

proteomics profiles in narrow contexts[29,30]. For example, controlled exposure experiments have demonstrated dose-dependent changes in cardiovascular-related proteins[29], and cohort studies in young adults have shown that ambient air pollution significantly perturbs the cardiometabolic proteome[30]. However, these studies focus solely on molecular responses to air pollution without linking them to clinical outcomes, leaving the mediating role of proteomic signatures in pollution-related disease development still unclear.

With the release of UKBB proteomic data, plasma proteins have been increasingly linked to various complex diseases[31–33]. Meanwhile, emerging evidence suggests that inflammation-related proteins may mediate the effects of air pollution on health[34–37]. For instance, a recent study has investigated the mediating role of proteins in the association between air pollution and depression[38]. However, the existing studies have been limited to a few specific proteins or a single disease outcome. A systematic investigation is needed to identify the proteomic signatures that underlie the health impacts of air pollution and their potential role in mediating individual disease susceptibility.

The effect of air pollution on populations varies, with significant heterogeneity in individual susceptibility[39]. Previous studies have highlighted these differences, showing that some individuals are more vulnerable to developing diseases than others[40–42]. While one individual has a unique proteomic profile, proteomic mediation analysis offers a promising approach to understanding the individual variability associated with environmental exposures[43,44]. However, despite the potential of leveraging proteomics to better understand the impact of air pollution on diseases, current research has yet to integrate large-scale cohort data and omics approaches to quantify this relationship.

In this work, we address these gaps by assessing the role of proteins in mediating the effects of air pollution on diseases. Leveraging data from UKBB, we systematically investigated the associations between air pollutants, proteomic profiles, and a wide range of diseases. Based on the findings of mediation analysis, we constructed an index named APPRS to quantify individual susceptibility to the health effects of air pollution. Together, our study not only elucidates proteomic pathways that mediate air pollution's impact on human health but also introduces a quantification that may facilitate precision environmental health strategies and personalized disease prevention.

Results

Study design

We conducted our analysis using data from the UK Biobank (UKBB), a large-scale prospective cohort that recruited 506,319 participants between April 4, 2006, and October 1, 2010 (Fig. 1A). For each participant, long-term air pollution exposures were estimated at baseline (2005–2010) based on residential address, and disease outcomes were prospectively ascertained through linkage to health records. Incident cases were defined as diagnoses occurring after 2010, which ensured that exposure assessment preceded disease onset in all analyses. Data on four air pollutants ($PM_{2.5}$, PM_{10}, NO_2, and NO_x) and 36 disease outcomes were collected and processed (Supplementary Data 1, 2). First, a systematic association analysis between air pollutants and 36 diseases was carried out on 502,131 participants who had available data on air pollution and disease but lacked proteomic measurements. Then, in the following analysis, we systematically excluded individuals without proteomic data, resulting in a study population of 49,242 participants with available data on air pollutants, 1,463 circulating proteins, and 36 disease outcomes.

As illustrated in Supplementary Fig. 1, we assessed the paired associations within air pollution ($PM_{2.5}$, PM_{10}, NO_2, NO_x, and an aggregated air pollution index), proteins, and diseases. To gain biological insights, we performed functional enrichment analyses for the identified proteins. Next, to further elucidate the role of proteins in linking air pollution to disease risk, we performed a mediation analysis in 49,242 participants, identifying key proteins that significantly mediated the effects of air pollution on disease development. Besides, to quantify individual susceptibility to air pollution, we developed an Air Pollution Protein Risk Score (APPRS) based on identified mediating proteins and incorporated it into disease prediction models. Finally, to assess the robustness and generalizability of the APPRS, we validated it in two independent external cohorts, demonstrating its predictive performance across diverse populations.

Overview of the impact of air pollutants on diseases

While some of the diseases have been previously studied to be associated with air pollution in UKBB (Supplementary Fig. 2A, B; Supplementary Data 3), the effects of air pollution on 20 of the diseases have not been reported in the UKBB cohort. Therefore, we first systematically assessed the effects of air pollution on various diseases using Cox models. Our findings reveal that mental health disorders (schizophrenia), immune diseases (multiple sclerosis, peripheral artery disease, and systemic lupus erythematosus), cardiovascular diseases (hypertension and anemia), skin diseases, and metabolic disorders (obesity) are affected by different air pollutants (Supplementary Fig. 2C). Besides, we found that distinct diseases were primarily associated with different air pollutants. PM pollutants were predominantly linked to immune diseases, consistent with their ability to induce systemic inflammation and oxidative stress, while NO_2 and NO_x had greater impacts on metabolic diseases and neuron diseases. In contrast, lung diseases and cardiovascular diseases exhibited relatively balanced associations across all pollutants, suggesting potential synergistic interactions across the effects (Supplementary Fig. 3).

Since individuals are often exposed to multiple pollutants simultaneously, we proposed the concept of an air pollution index (AP index) to systematically evaluate the degree of air pollution exposure by integrating multiple pollutants. We observed a significant consistency between our AP index and the Living Environment Score proposed by the UKBB, which reflects the quality of an individual's immediate surroundings both indoors and outdoors ($R_{cor} = 0.524$; Supplementary Fig. 4). Next, we examined the effect of the AP index on diseases and found that 29 diseases were influenced by AP index (Fig. 1B). This suggests that the AP index allows for a more comprehensive understanding of the relationship between air pollution and diseases. Among these associations, we found that air pollution had the strongest impact on pulmonary and immune diseases, while its association with cancer was relatively weaker (Fig. 1C). Besides, we observed discrepancies in statistical significance between individual pollutants and the aggregated AP index across several disease outcomes, potentially due to the integration of heterogeneous exposure signals or differences in statistical power (Supplementary Texts and Supplementary Fig. 5).

Effects of air pollutants on circulating proteins

Of all participants in UKBB, 49,242 individuals (9.8%) had valid proteomic data, which were used to investigate the correlation between air pollutants and the proteome. We performed a proteome-wide association study (PWAS) analysis for each type of air pollutant and identified a total of 852 proteins significantly influenced by air pollution (Fig. 2A and Supplementary Fig. 6; Supplementary Data 4). Additionally, 770 proteins were identified to be associated with the AP index, among which 237 were not captured by the PWAS of individual pollutants (Fig. 2B; Supplementary Data 5). Moreover, we found that while a small subset of proteins was associated with only a single air pollutant—30.66% for $PM_{2.5}$, 36.99% for PM_{10}, 33.05% for NO_2, and 7.05% for NO_x—the majority exhibited with multiple pollutants (Fig. 2C).

The proteins associated with air pollution were primarily enriched in functions related to immune processes and stimuli responses, including cytokine activity, cytokine receptor binding, chemotaxis, and leukocyte migration (Fig. 2D). The pathway enrichment analyses in

https://doi.org/10.1038/s41467-026-68972-6

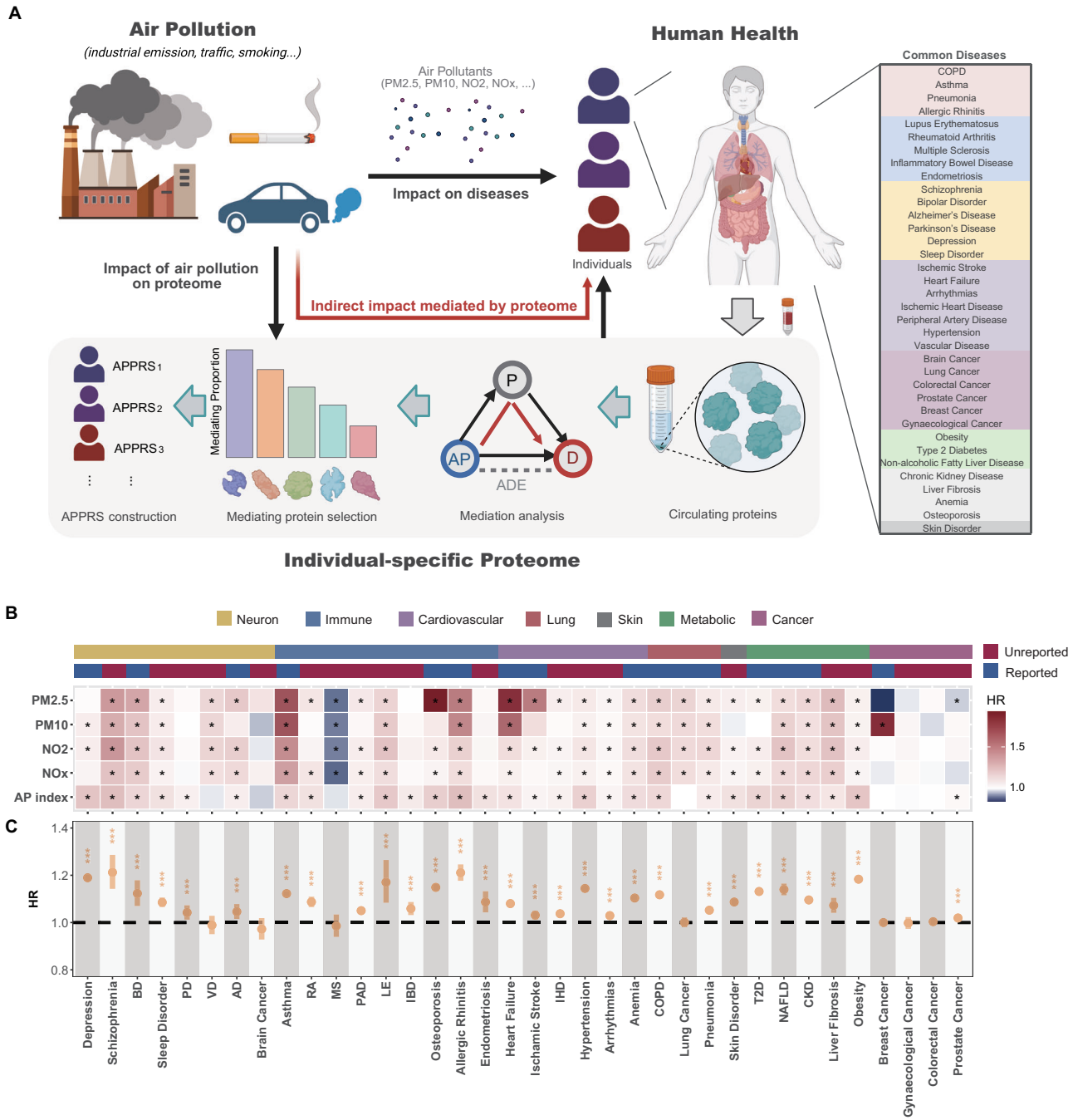


Fig. 1 | Overview of the study design and the associations between air pollutants and diseases. A Overview of the data description and study design, illustrating the schematics of air pollutants, proteomic data, and disease information. Partly created in BioRender. Li, W. (2025) https://BioRender.com/oiqulw5. **B** The associations between various pollutants and multiple diseases. The first colored bar represents the different types of diseases, while the second bar indicates whether an association between the pollutant and the disease has been reported. The color gradient in the heatmap corresponds to the Hazard Ratios (HRs) derived from Cox proportional hazards models. Significant associations ($P < 0.05$) are marked with an asterisk (*). **C** The forest plot illustrates the associations between AP index and various diseases ($n = 502{,}131$ participants). Data are presented as HRs with 95% confidence intervals (CI) derived from Cox proportional hazards models. The AP index is the weighted sum of the air pollutants ($PM_{2.5}$, PM_{10}, NO_2, and NO_x). Significant differences are marked with asterisks (***$P < 0.001$). Source data are provided in the Source Data file.

KEGG[45] and Reactome[46] revealed consistent findings with the functional terms identified in GO (Fig. 2E-F; Supplementary Data 6). Notably, the enrichment observed with the AP index was more significant than that observed for individual pollutants, highlighting the broader impact of cumulative air pollution exposure. Moreover, the specific proteins associated with each individual air pollutant showed distinct enrichment patterns (Supplementary Fig. 7). For example, PM-specific proteins were specifically associated with the MAPK cascade and cell adhesion ($P_{GO:\ 0043410} = 6.5e{-}12$ and $P_{GO:\ 0022407} = 1.9e{-}11$; Supplementary Fig. 7A, B), while NO-specific proteins were primarily related to nervous system functions, such as the ensheathment of neurons and negative regulation of nervous system processes ($P_{GO:\ 0007272} = 5.9e{-}05$ and $P_{GO:\ 0031645} = 5.9e{-}05$; Supplementary Fig. 7C, D). Common proteins shared across the different pollutants were predominantly associated with immune processes and cell chemotaxis (Supplementary Fig. 7E). In general, although the proteins commonly associated

https://doi.org/10.1038/s41467-026-68972-6

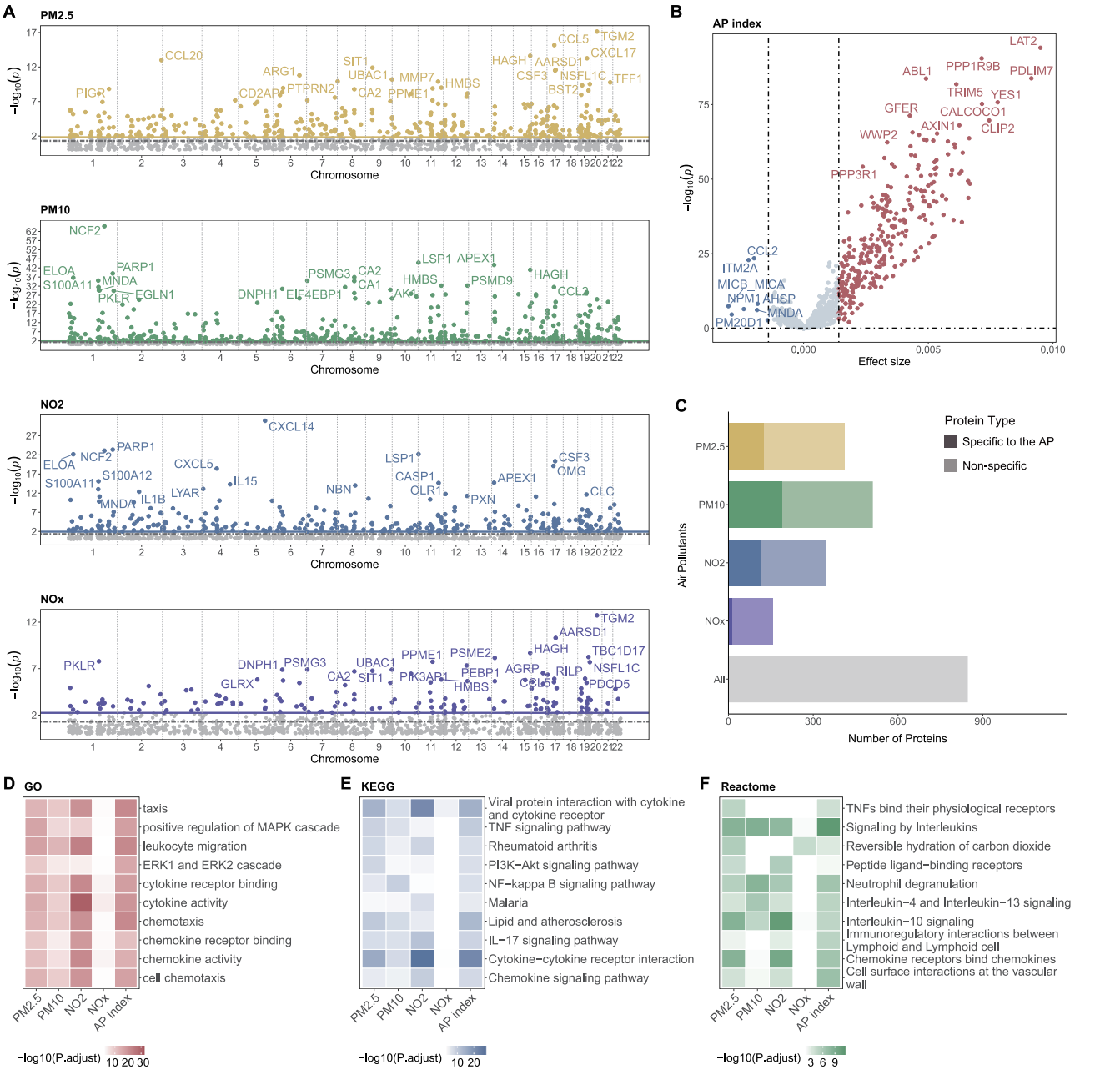


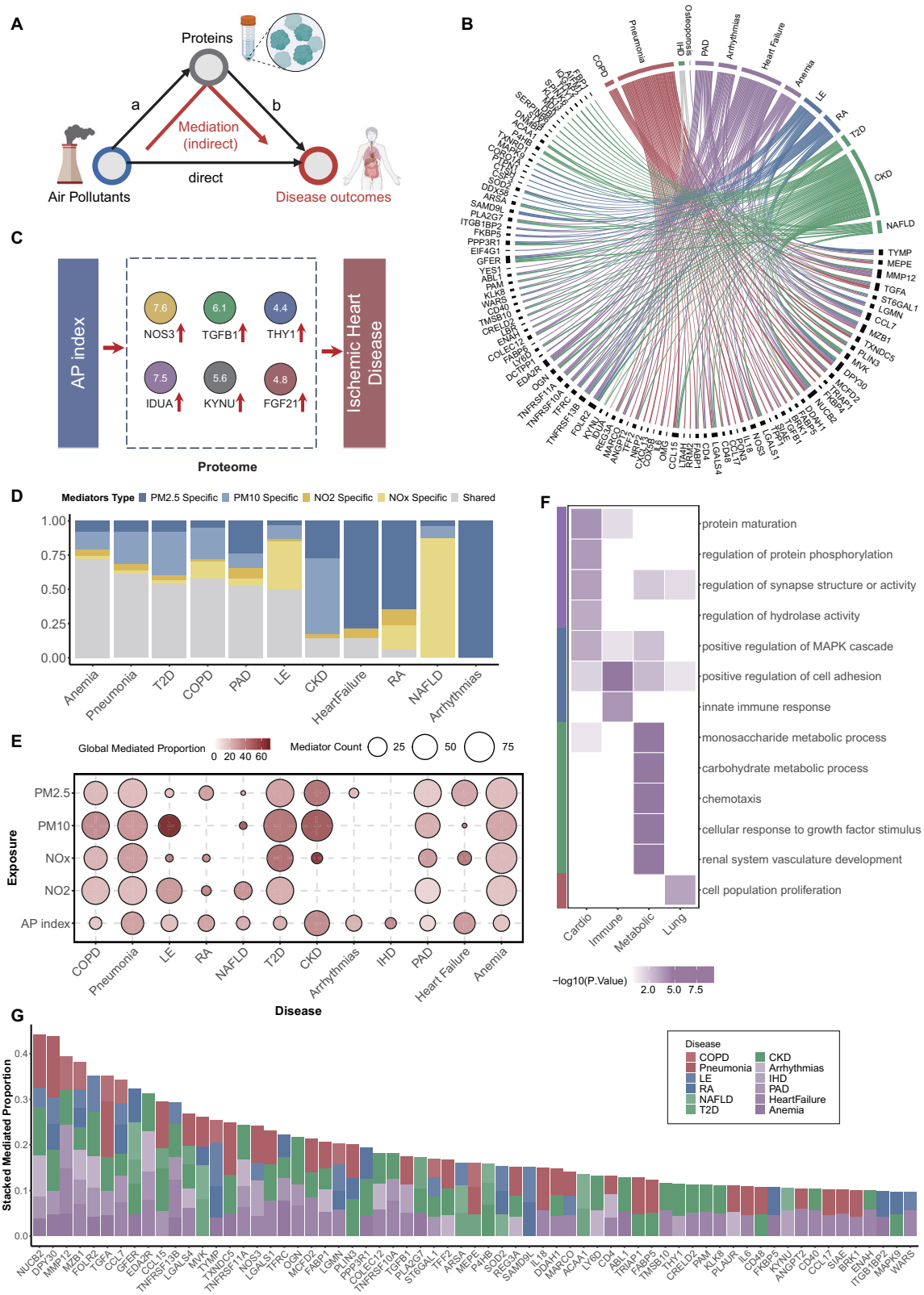
Fig. 2 | Associations between plasma proteins and air pollutants. A Manhattan plot of PWAS results for $PM_{2.5}$, PM_{10}, NO_2, and NO_x ($n = 49{,}242$ participants). Associations were assessed using linear models via the limma package. Colored dots represent proteins significantly associated with the corresponding air pollutant. The $-\log_{10}(p)$ in y-axis represents the statistical significance of the association between proteins and the pollutants. The gray horizontal line represents the significance threshold of $P < 0.05$, while the colored horizontal lines indicate the threshold of FDR < 0.05 (Benjamini-Hochberg correction). **B** Volcano plot of PWAS results for the AP index from linear models. Red dots represent significantly upregulated proteins, and blue dots represent significantly downregulated proteins at the threshold of FDR < 0.05. **C** Number of proteins significantly associated with each air pollutant ($PM_{2.5}$, PM_{10}, NO_2, NO_x). The darker color in each bar represents proteins specific to a single pollutant, while the light color represents proteins associated with multiple pollutants. Functional enrichment analysis of proteins associated with different air pollutants in terms of GO (**D**), KEGG (**E**), and Reactome (**F**) databases. The $-\log_{10}(p.adjust)$ shows the enrichment significance using Fisher's exact tests, adjusted for multiple comparisons using the Benjamini-Hochberg method. Top 10 most significantly and commonly enriched functions and pathways are shown. Source data are provided in the Source Data file.

across pollutants are primarily linked to immune functions, different air pollutants can specifically impact distinct biological pathways (Supplementary Fig. 7F).

Proteome mediates the association between air pollution and diseases

To investigate the role of the proteome in mediating the effects of air pollutants on disease risk, we calculated the associations between proteins and diseases using Cox models, selecting proteins that are significantly causal for each disease (Bonferroni-corrected $P < 0.05$; Supplementary Texts, Supplementary Figs. 8–10, and Supplementary Data 7). By intersecting these proteins with those associated with air pollutants identified in our study, we obtained 479 proteins that were simultaneously associated with both air pollution exposure and corresponding diseases. Then, we performed mediation analysis to assess whether the proteins serve as mediators between air pollutants and

https://doi.org/10.1038/s41467-026-68972-6



diseases (Fig. 3A; Supplementary Data 8). Among the 36 diseases examined, we found 23 diseases were significantly influenced by air pollution through proteomic mediation (Supplementary Fig. 11), of which 12 were influenced by AP index (Fig. 3B). For example, six proteins, including NOS3, TGFB1, IDUA, KYNU, THY1, and FGF21, mediated the association between AP index and ischaemic heart disease (IHD). 5 of the 6 proteins (NOS3, TGFB1, THY1, FGF21, and KYNU) were involved in the cellular catabolic process and regulation of transmembrane transport, which are the dysregulated processes resulting from oxidative stress triggered by air pollution (Fig. 3C).

Besides, a large overlap was observed among the mediating proteins across different air pollutants (Fig. 3D). The proportion of mediation varied on the specific air pollutants, and different diseases exhibited distinct mediation patterns influenced by air pollutants

Fig. 3 | Mediation analysis of air pollution, the proteome, and disease outcomes. A A schematic diagram illustrating the mediation analysis framework between air pollutants, the proteome, and disease outcomes. Created in BioRender. Li, W. (2025) https://BioRender.com/18w5kd1. **B** Mediation results of the AP index on various diseases through different proteins. Diseases included pulmonary diseases (red), cardiovascular diseases (purple), immune diseases (blue), and metabolic diseases (green). **C** An example illustration of the mediation process, illustrating how proteins link air pollution to disease via specific biological processes. **D** Classification of mediator proteins across different air pollutants and diseases. Y-axis represents the proportion of mediating proteins for each disease. Blue-shaded proteins are primarily mediators for PM exposures ($PM_{2.5}$ and PM_{10}), yellow-shaded proteins are mediators for nitrogen oxides (NO_2 and NO_x), and unshaded proteins are shared mediators. **E** The number of mediators for different diseases (circle size) and the overall mediation proportion of the proteome across different air pollutants (circle color). **F** Functional enrichment of disease-specific mediators. Significance was determined using Fisher's exact tests, and P values were adjusted for multiple comparisons using the Benjamini-Hochberg method. Pathways are grouped by disease category and color-coded for clarity (cardiovascular, immune, metabolic, lung). Partial overlap of pathways between diseases reflects shared molecular components in enrichment analyses. **G** Stacked mediation proportions of mediator proteins across diseases. Proteins are ranked based on their cumulative mediation proportions between the AP index and various diseases. Source data are provided in the Source Data file.

(Supplementary Data 9). While many diseases were influenced by common mediating proteins, some diseases were predominantly mediated by PM-related proteins, such as chronic kidney disease (CKD), heart failure, and arrhythmia. In contrast, other diseases, such as non-alcoholic fatty liver disease (NAFLD), were primarily mediated by NO-associated proteins (Fig. 3D).

The number of proteins acting as mediators varied across diseases, with an average of 29 proteins significantly mediating the associations between air pollutants and diseases (Fig. 3E; Supplementary Fig. 12). To assess the global mediation effect of the proteome, we computed the top 10 principal components of protein abundances and used the weighted sum of mediation proportions as a global mediation measure, following the approach described in refs. 47,48. On average, the proteome mediated 23.69% of the total effects of air pollution on diseases, with the highest mediation proportions observed in pneumonia (27.55%), IHD (36.73%), and CKD (35.33%; Fig. 3E; Supplementary Data 10).

To further investigate the specific pathways mediating the associations between air pollution and various diseases, we performed functional enrichment analyses on disease-specific mediator proteins and those shared across multiple diseases (Fig. 3F; Supplementary Fig. 13A). Our results revealed that mediators shared by two or more diseases were enriched in processes related to protein phosphorylation and immune responses (Supplementary Fig. 13B). Notably, stratified analyses revealed distinct enrichment patterns across disease categories. Cardiovascular disease-specific mediators were particularly associated with regulation of the MAPK cascade and protein phosphorylation pathways, while those unique to immune diseases were enriched in innate immune response and positive regulation of cell adhesion (Supplementary Fig. 13C, D). Metabolic disease mediators demonstrated strong involvement in carbohydrate and monosaccharide metabolic processes, and lung disease mediators were predominantly linked to cell population proliferation (Supplementary Fig. 13E, F).

Next, to identify key mediating proteins, we calculated the stacked mediation proportion across diseases for each mediator and ranked the mediators based on their contribution to the mediation proportion. The top ranked proteins exhibited the greatest mediating effects of AP index on disease outcomes (Fig. 3G). Among them, NUCB2, DPY30, and MMP12 were the most prominent mediators, with stacked proportions of 44.19%, 43.73%, and 39.33%, respectively. A comparison between healthy and unhealthy individuals revealed significant differences in the expression of these key mediating proteins at baseline, with the diseased population showing distinct expression patterns (Supplementary Fig. 14). Functionally, NUCB2 is the precursor of nesfatin-1, a peptide involved in cardiovascular regulation and stress response; DPY30 is a core component of an epigenetic modification, histone methyltransferase complex, which is easily influenced by environmental factors; MMP12 is an enzyme belonging to the matrix metalloproteinase family, which plays a key role in inflammation and immune response. Besides, by querying the DrugBank database[49], we found that 23 out of the 65 identified proteins have been previously documented as drug targets associated with different diseases (Supplementary Fig. 15; Fisher's exact test $P = 2.73e{-}82$).

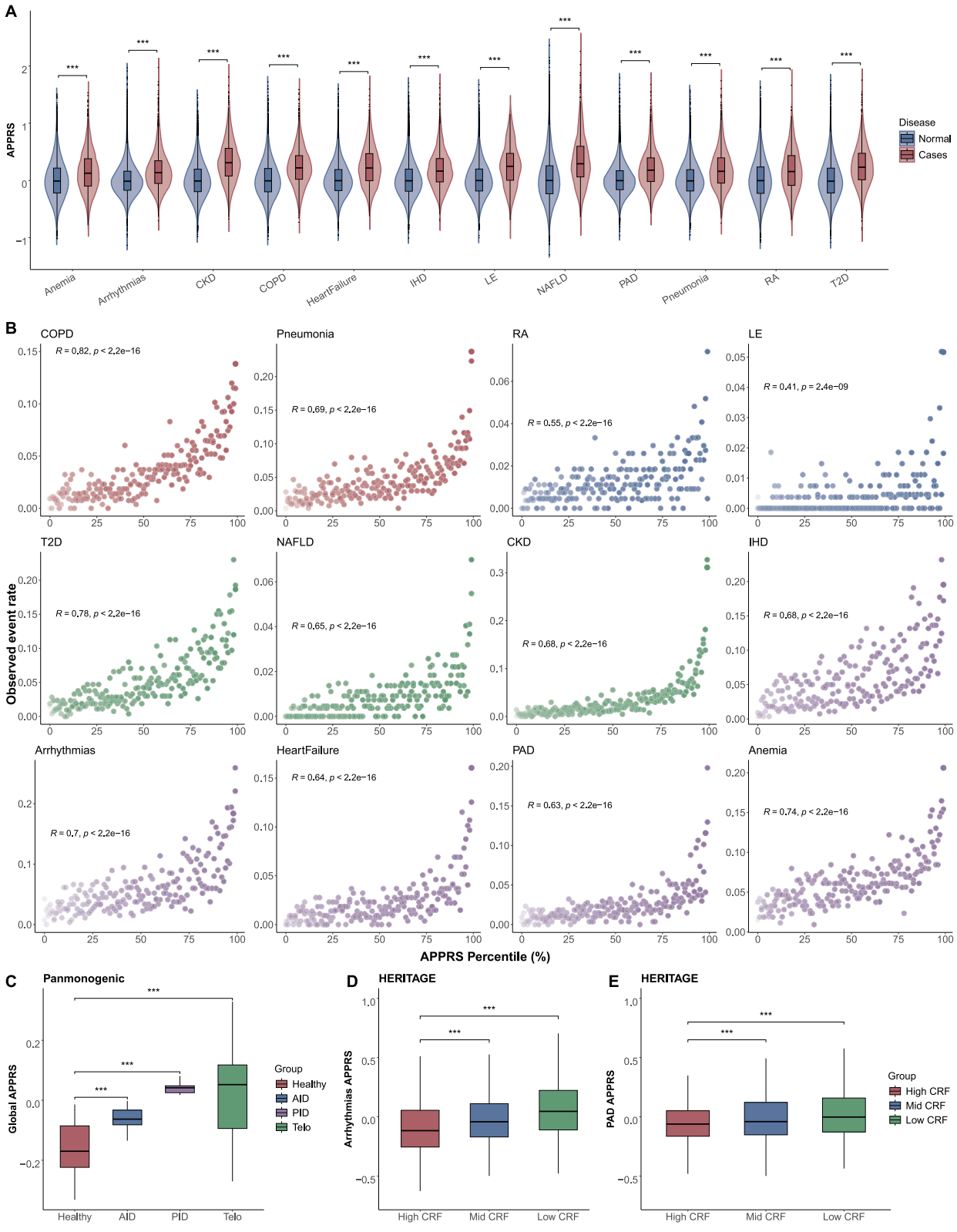
APPRS represents individual sensitivity to the health effects of air pollution

The design of our mediation analysis quantified the role of individual proteomic profiles in mediating the effects of air pollution on diseases. The mediating proteins identified in our study provided insights into how environmental exposure influences disease through proteomic alterations. Thus, to quantify individual susceptibility to the effects of air pollution, we developed an air pollution-related protein risk score (APPRS), integrating proteins identified as mediators of the effects of pollutants on diseases using a designed statistical model (detailed in Methods). A total of 65 proteins, identified as mediators in at least two diseases, were selected to construct the APPRS (Fig. 3G; Supplementary Data 11). The coefficients used for APPRS construction were derived from the mediation effects of each protein on the association between air pollution and a specific disease, thereby making the APPRS disease-specific (Supplementary Fig. 16).

The distribution of APPRS follows a normal distribution and shows a positive correlation with the AP index (Supplementary Figs. 17–19). To check heterogeneity in individual susceptibility, we conducted subgroup analyses of APPRS across key demographic strata. Significant inter-group differences were observed (Supplementary Texts and Supplementary Figs. 20–24). Then, we compared APPRS distributions between healthy individuals and those with diseases. Consistently, APPRS was significantly elevated in individuals with diseases compared to those in a healthy state (Fig. 4A). This finding supports the notion that APPRS represents individual susceptibility to the effects of air pollution, with substantial inter-individual variability; higher APPRS values were associated with poorer health conditions. In addition, we investigated the associations between APPRS and disease incidence rates. Across the 12 diseases, individuals with higher APPRS percentiles at baseline exhibited elevated observed event rates (Fig. 4B). Notably, the strongest associations between APPRS and disease risk were observed in lung diseases, such as chronic obstructive pulmonary disease (COPD) and pneumonia ($P_{COPD} = 3.93e{-}49$; $P_{pneumonia} = 3.54e{-}30$), while the associations were slightly weaker in metabolic diseases such as T2D and NAFLD ($P_{T2D} = 8.91e{-}43$; $P_{NAFLD} = 3.97e{-}25$).

To further assess the generalizability of APPRS, we validated its practicality in two independent external cohorts: the Panmonogenic cohort and the HERITAGE cohort. In the Panmonogenic cohort, we found that individuals with immune diseases exhibited significantly higher global APPRS compared to healthy individuals (Fig. 4C), reinforcing the link between air pollution susceptibility and immune dysfunction. In the HERITAGE cohort, participants were stratified into three groups based on their cardiorespiratory fitness (CRF) levels. Since CRF was closely associated with cardiovascular efficiency and vascular function[50,51], both arrhythmias APPRS and peripheral artery disease (PAD) APPRS were calculated and compared across these groups (Fig. 4D-E). Notably, individuals with higher CRF levels

https://doi.org/10.1038/s41467-026-68972-6



exhibited lower arrhythmias APPRS values, consistent with previous findings that higher CRF is associated with improved cardiopulmonary function and more efficient oxygen utilization[52]. Likewise, individuals with higher CRF levels also aligned with lower PAD APPRS values, likely due to enhanced vascular health and endothelial function associated with better CRF[53]. We also found that APPRS did not show significant differences in phenotypes unrelated to air pollution, supporting its specificity in capturing pollution-associated disease risk (Supplementary Texts; Supplementary Fig. 25).

Next, to investigate whether air pollution and APPRS interact in influencing disease incidence, we analyzed the association between APPRS and disease risk under varying levels of air pollution exposure. Participants were stratified into three groups based on their air pollution exposure levels, and within each group, we compared disease

Fig. 4 | Validation of APPRS in relation to disease risk and external cohorts. **A** Comparison of APPRS between healthy individuals and those with diseases ($n = 49,242$). Statistical significance was determined using two-sided Wilcoxon rank-sum tests. Significant differences are marked with asterisks (*** $P < 0.001$). **B** Associations between APPRS and disease incidence rates ($n = 49,242$). The x-axis represents the percentiles of APPRSs in the population, while the y-axis shows the observed event rates of different diseases. The colors in the dot plot correspond to the different types of diseases: pulmonary diseases (red), cardiovascular diseases (purple), immune diseases (blue), and metabolic diseases (green). External validation of APPRS in the Panmonogenic and HERITAGE cohorts. (**C**) Comparison of APPRSs between healthy individuals and those with diseases in the Panmonogenic cohort ($n = 228$ participants across 22 immune-related diseases). Significance was assessed using two-sided Wilcoxon rank-sum tests. In the HERITAGE cohort, Arrhythmias APPRS (**D**) and PAD APPRS (**E**) were compared among individuals with varying levels of CRFs using two-sided Wilcoxon rank-sum tests. For all box plots, data are presented as the median (centre), 25th and 75th percentiles (bounds of box), and minima/maxima (whiskers). Source data are provided in the Source Data file.

incidence between individuals with high versus low APPRS. The results showed that, within the same exposure category, individuals with higher APPRS consistently exhibited a greater disease incidence. Moreover, increased air pollution exposure was independently associated with a higher disease risk (Supplementary Fig. 26). The most significant patterns were observed in lung diseases (COPD and pneumonia) and immune diseases (RA and LE), whereas the trends were less notable for CKD, NAFLD, and cardiovascular diseases (IHD and arrhythmia). Moreover, similar trends were observed across different pollutants (Supplementary Figs. 27–30), further supporting the general interaction between APPRS and air pollution.

APPRS improves prediction performance for the early onset of diseases

Building on the understanding that APPRS is positively associated with disease risk, we developed predictive models that integrate APPRS with basic features (age, sex, smoking status, alcohol consumption, race, and BMI) and clinical variables (ApoB/ApoA, albumin, creatinine, HbA1c, and glucose) to assess the early onset of diseases related to air pollution (Fig. 5A). Models using only basic or clinical information were established as baseline models. We compared the performance of models incorporating APPRS with that of baseline models. The results showed that incorporating APPRS as an additional feature consistently improved model performance, leading to average AUROC increases of 3.64% and 6.67% compared to basic and clinical models, respectively (Fig. 5B, C; Supplementary Data 12). The final model, which integrated APPRS with both basic and clinical features, achieved the highest predictive performance, with an average AUROC of 77.25% (Fig. 5D; Supplementary Fig. 31).

To assess the sensitivity of APPRS in the prediction model, we designed three comparative models. The first comparative model constructed a random protein score by randomly selecting the same number of proteins and derived this score by applying the same statistical models as APPRS. In the second comparative model, proteins were still randomly selected but restricted to those significantly associated with the corresponding diseases.

We then compared the performance of prediction models incorporating APPRS with those incorporating either type of random protein score in the baseline model. The results showed that across all diseases, models with APPRS significantly outperformed those with random protein scores, illustrating the predictive value of APPRS in disease prediction (Supplementary Fig. 32A, B). Next, we compared the predictive performance of models incorporating AP index versus APPRS as features in the baseline model. The results demonstrated that APPRS consistently improved prediction performance and outperformed the AP index across all diseases (Supplementary Fig. 32C). These comparative models highlight the superior ability of APPRS to quantify individual susceptibility to improve disease prediction.

To evaluate the contribution of APPRS to the prediction, we examined the model coefficients of all predictive features. Among all predictors, we found APPRS had the most significant impact, accounting for 23.34% of the variation in disease outcomes (Fig. 5E). Additionally, we validated the predictive utility of APPRS in an external cohort, the HERITAGE cohort. While the Panmonogenic cohort only provided average protein expression data across disease subtypes, the HERITAGE cohort offered individual-level proteomic data and CRF measurements, making this cohort suitable for validation. In the HERITAGE cohort, we compared the performance of the baseline model, which was built using only age, sex, race, and BMI, with the model that incorporated APPRS in addition to the baseline features (Fig. 5F). Results showed that adding APPRS improved 2.77% of AUROC in model performance, with APPRS contributing 75.41% to the prediction (Fig. 5G).

Discussion

In this study, we explored the potential mediating role of proteins in the associations between air pollution and diseases, utilizing data from 49,242 participants with available proteomic data in the UKBB. We identified significant associations between air pollution and 30 diseases, including 14 new associations not previously reported in UKBB. Our findings demonstrate that the proteome mediates the effects of air pollution on disease susceptibility, with a global mediation proportion of 23.69%. Additionally, we developed an APPRS score, which successfully quantified individual susceptibility to air pollution exposure and predicted various disease risks. Moreover, APPRS was validated in two independent cohorts and showed strong associations with specific health outcomes, providing a valuable tool for predicting the onset of diseases related to air pollution.

In recent decades, UKBB has been extensively used to investigate associations between air pollution and various health outcomes[1–21]. However, most of these studies have only focused on the associations with a particular disease outcome, without exploring the differential impact of specific air pollutants on different disease categories. In this study, we systematically examined the associations of multiple air pollutants with 36 diseases and identified the dominant pollutants for each disease. Our results revealed that PM pollutants were predominantly associated with immune-related diseases, consistent with their known role in inducing systemic inflammation and oxidative stress[54]. Nitrogen oxides were more strongly linked to metabolic and neurological disorders, aligning with previous findings suggesting their involvement in metabolic dysregulation[55] and cognitive decline[56]. More interestingly, lung and cardiovascular diseases exhibited relatively balanced and significant associations across all pollutant types, which may suggest their broad susceptibility to air pollution.

Several studies have investigated the relationship between air pollution and proteomic alterations. However, these studies typically focus on a limited number of proteins, such as C-reactive protein (CRP) and other inflammation-related biomarkers[57]. For instance, a study from Stockholm County found that ambient air pollution exposure is associated with changes in inflammation-related protein levels in young children[58]. While informative, these studies were mostly limited to specific functional proteins and were conducted in relatively small cohorts, lacking a comprehensive exploration of the circulating proteome. Our study systematically assessed the effects of various air pollutants on a wide spectrum of circulating proteins, revealing potential biological mechanisms underlying pollutant-specific impacts. PM-specific proteins were observed to be related to MAPK cascades and cell adhesion, consistent with previous reports that PM exposure induces reactive oxygen species (ROS) production and enhances adhesion molecule expression[59,60]. Proteins uniquely

https://doi.org/10.1038/s41467-026-68972-6

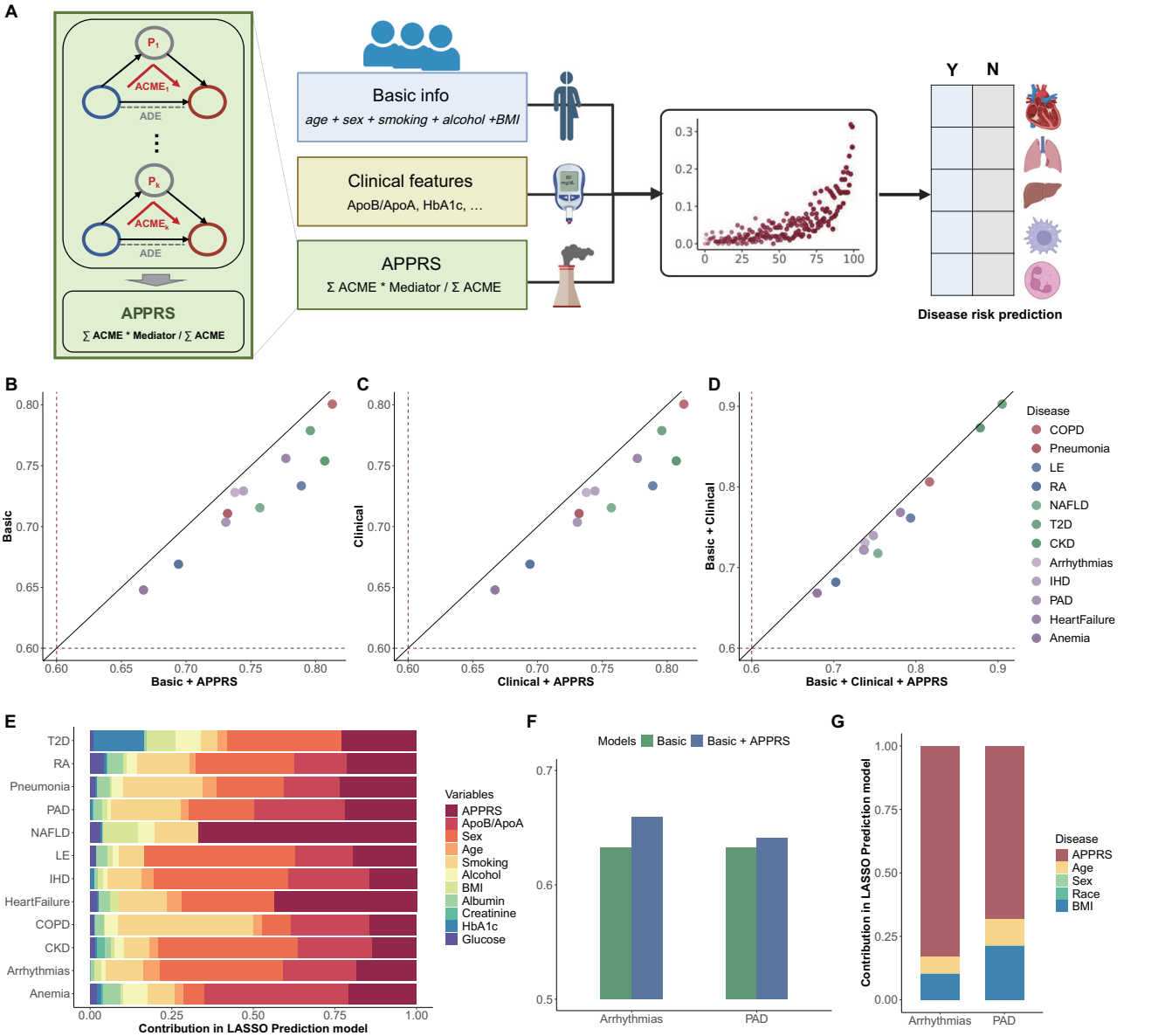


Fig. 5 | Integration of APPRS into disease prediction models. A Schematic framework of the disease prediction model incorporating APPRS. In addition to basic demographic (age, sex, smoking status, alcohol consumption, race, and BMI) and clinical information (ApoB/ApoA, albumin, creatinine, HbA1c, and glucose), APPRS is included as a feature to enhance disease risk prediction associated with air pollution exposure. Partly Created in BioRender. Li, W. (2025) https://BioRender.com/i8vhsv3. AUROC performance comparison of models with and without APPRS as an input feature: Basic model *vs.* Basic + APPRS model (**B**), Clinical model *vs.* Clinical + APPRS model (**C**), and Basic + Clinical model *vs.* Basic + Clinical + APPRS model (**D**). **E** Contribution of different features in disease prediction models across various diseases. **F** Validation of APPRS contribution to disease prediction in an external cohort. Model performance was compared between the baseline model and models incorporating Arrhythmias APPRS or PAD APPRS. Due to data availability constraints, the baseline model was built only using age, sex, race, and BMI. **G** Evaluation of APPRS contributions in the external validation cohort. Source data are provided in the Source Data file.

associated with nitrogen oxides were mainly involved in neurological functions. This finding aligns with previous studies highlighting the role of nitric oxide in modulating neural networks[61]. Besides, consistent with previous findings[62,63], the proteins commonly affected by different air pollutants were predominantly related to immune responses and chemotaxis, suggesting shared pathways of systemic immune activation in response to air pollution. Notably, while most proteins showed consistent association directions between individual pollutants and the AP index, a subset exhibited opposite patterns (Supplementary Fig. 33). This likely reflects that the AP index, constructed as a weighted composite of multiple correlated pollutants, may display reversed directionality, as its weights were based on self-rated overall health rather than disease-specific associations. While single-pollutant analyses reveal pollutant-specific effects, the AP index reflects the overall pollution burden, providing a complementary system-level perspective on air pollution–disease associations.

With the release of UKBB proteomics data in 2023, the associations between proteins and diseases have gained increasing attention. Recent UKBB studies have demonstrated that plasma proteomic profiles predict disease risk beyond traditional factors[31,32,64], highlighting the critical role of proteins in disease pathogenesis. Building on this, our study integrates the impact of air pollution on the proteome with the proteome-disease associations to systematically investigate the mediating role of proteins in linking air pollution to multiple disease outcomes. Our analysis identified distinct biological pathways across AP-affected diseases—such as MAPK signaling in cardiovascular diseases, innate immune responses in immune disorders, monosaccharide and carbohydrate metabolic processes in metabolic

diseases, and cell proliferation in respiratory diseases—suggesting both the shared and unique mechanisms through which pollutants are associated with different organ systems. To strengthen the biological support for these findings, we reviewed the literature on the 65 key mediating proteins used in constructing APPRS and found evidence for 19 of them linking air pollutant exposure to relevant pathways and disease outcomes (Supplementary Data 13). Collectively, our work goes beyond existing studies by providing a comprehensive and disease-specific functional map of protein mediators, enabling a more mechanistic understanding of how air pollution contributes to diverse health outcomes.

The APPRS developed in this study represents a novel statistical index for quantifying individual susceptibility to air pollution-induced diseases based on circulating protein mediators. When integrated into baseline disease prediction models, APPRS consistently improved predictive performance across 12 outcomes. Notably, the magnitude of improvement varied, with some diseases showing significant gains while others exhibited more limited enhancements. This variability may be due to the strong predictive power of baseline clinical models, especially for chronic diseases, as well as differences in disease sensitivity to air pollution. These results suggest that APPRS may be particularly informative for conditions where air pollution plays a major etiological role. Unlike static polygenic risk scores (PRS) derived solely from genetic variation, APPRS captures dynamic, environmentally responsive molecular changes, reflecting both baseline genetic risk and real-time physiological responses to environmental exposures. Compared to other emerging responsive scores like the gut microbiome health index (GMHI)[65] and Immune Health Metric (IHM)[66], APPRS reflects the current biological state shaped by both environmental exposures and internal regulatory mechanisms. The dynamic nature of APPRS highlights its potential as a biomarker for real-time health monitoring, early intervention, and precision prevention.

Our study presents several key advances over previous research. First, to our knowledge, this is the first study to utilize a large-scale cohort to explore the relationship among the air pollution exposure, proteome, and a broad range of disease outcomes. We presented a systematic proteomic analysis of the mediation pathways linking air pollution to a broad range of diseases, providing insights into underlying mechanisms. Second, we investigated the pollutant-specific effects on diseases, identified proteins uniquely associated with different air pollutants, and explored disease-specific mediation pathways linking air pollution and different types of diseases. This offers a deeper understanding of the molecular mechanisms by which distinct pollutants contribute to various disease types. Third, the development of APPRS represents an advancement in understanding and quantifying individual susceptibility to air pollution. By integrating the expression levels of proteins that mediate air pollution's impact on diseases, APPRS offers a personalized risk assessment tool that captures inter-individual variability in pollution sensitivity. Notable associations were observed between APPRS and several diseases, with especially strong links observed for respiratory conditions such as COPD and pneumonia, underscoring the respiratory system's particular vulnerability to air pollution. Furthermore, the application of APPRS in disease prediction models and the validation in external cohorts support its broader applicability and implies its potential as a biomarker for personalized risk stratification and targeted prevention in population health.

Several limitations should be acknowledged in interpreting our findings. First, this study relies on single-time point measurements of proteomic data, which may not capture long-term responses in protein levels. Future studies incorporating longitudinal exposure data with finer temporal granularity, early biomarkers of disease onset, and causal modeling approaches will be essential to identify disease-specific exposure windows. Longitudinal proteomic data and disease-specific sensitive windows would enable a more detailed examination of temporal changes and adaptive responses in protein levels, providing insight into both acute and chronic impacts of air pollution. Second, exposure estimates were based on residential addresses and may not fully reflect individual-level variations (e.g., indoor exposure or commuting), potentially leading to misclassification. Personal exposure monitoring could help refine these assessments. Third, although we adjusted for common confounders, residual confounding from factors such as lifestyle or co-exposures cannot be excluded. In addition, although we verified the robustness of results using AP indexes derived from objective health indicators, potential bias from self-rated health cannot be entirely ruled out. Fourth, we used a subset of coding proteins available from UKBB, which may not fully reflect the entire range of proteins relevant to air pollution response. Additionally, the use of a single cohort for developing the APPRS may introduce biases related to the specific population sampled. Although we validated the APPRS in two external cohorts, further validation in diverse populations with different environmental exposures would enhance its generalizability. Finally, our study focused on the mediating role of the proteome, whereas emerging novel frameworks that integrate environmental mixtures with biological networks[67,68] suggest promising directions for developing biologically informed APPRS models in future research.

Other omics layers (e.g., transcriptome, metabolome) may also play a role in the response to air pollution, and should be integrated in future research.

In conclusion, our study provides insights into the pathways by which air pollution exposure contributes to disease risk, highlighting the mediating role of the proteome. Our work supports the hypothesis that the proteome may serve as a bridge between environmental exposure and disease, revealing potential biomarkers for assessing disease risk linked to air pollution. We provided a personalized score for assessing susceptibility to air pollution-related health risks and demonstrated its value in disease prediction. By validating the APPRS in external cohorts, we underscore its potential for application in diverse populations and settings. Our findings not only advance our understanding of the molecular mechanisms underlying air pollution's health impacts but also highlight proteomic biomarkers that may serve as targets for early intervention and risk stratification.

Methods

Study cohort and participants

We retrieved data from the UK Biobank under Application Number 77803. The UK Biobank is a large prospective cohort project (https://www.ukbiobank.ac.uk/)[69]. This project complies with all relevant ethical regulations. All participants provided informed consent. UKBB participants included both males and females. Sex was based on self-report and a genetic sex check by UKBB. The analytical sample consisted of 506,040 participants recruited between April 4, 2006, and October 1, 2010, and 502,131 of them had data on air pollutants ($PM_{2.5}$, PM_{10}, NO_2, NO_x; Supplementary Data 1) and health outcomes. Air pollution exposure estimates were calculated using Land Use Regression (LUR) models developed by the European Study of Cohorts for Air Pollution Effects (ESCAPE), which integrated traffic-related pollution variables and provided high spatial resolution (100 × 100 m, i.e., 0.01 km^2 grid cells). Annual air pollution estimates for each year were calculated for all addresses, measured in micrograms per cubic meter ($\mu g/m^3$). The levels of NO_x and $PM_{2.5}$ were available only for the year 2010, we used the data for 2010 to calculate baseline exposure. NO_2 data were available for 2005-2007 and 2010, and PM_{10} data were available for 2007 and 2010. Given stable spatial patterns in UKBB air pollution[17,70] and validation against UK-AIR[8], we calculated the average values across the available years to represent the long-term exposure for each participant at baseline.

Disease outcomes encompass 36 diseases across various categories reported by refs. 31,64, including 8 neurological diseases, 9

immune diseases, 6 cardiovascular diseases, 3 lung diseases, 1 skin disease, 5 metabolic diseases, and 4 cancers. Detailed information on 36 diseases, including their prevalence rates and ICD-10 codes were shown in Supplementary Data 2. Disease outcomes were defined as time to first diagnosis after January 1, 2010, and recurrent events were not considered. Individuals with complete data on air pollutants and health outcomes were used for air pollution-disease association analysis.

For the investigation of the effects of air pollutants on diseases through proteome, we utilized data from 49,242 UKBB participants with available proteome profiles. This proteomics cohort was shown to closely mirror the full UKBB in key demographic, behavioral, and biochemical characteristics (Supplementary Data 14), indicating its broad representation of the wider UKBB population[71]. Proteomic data, generated from April 2021 to February 2022, were obtained using the Olink Explore platform, covering 1463 unique coding proteins. Data were provided as $\log_2$-scaled Normalized Protein eXpression (NPX) values following Olink's standard preprocessing and quality control pipeline. For analysis, NPX values were further standardized using a z-score transformation to ensure a mean of 0 and a standard deviation of 1 across participants. We followed the standard QC pipeline, including data filtering, normalization, and missing value imputation[71,72]. Ethical approval for the UKBB was obtained from the National Health Service National Research Ethics Service, and all participants provided electronic informed consent.

Construction of AP index

To derive a joint evaluation for air pollution quality, our study constructed an air pollution (AP) index by integrating the effects of several air pollutants (NO_2, NO_x, $PM_{2.5}$, and PM_{10}). The self-reported overall health ratings (Field ID: 2178) used to construct the AP index were collected at baseline assessment during participant recruitment (2006–2010). We quantified the relationship between each air pollutant (independent variable) and self-reported overall health ratings (dependent variable) by a linear regression model, adjusting for baseline age, sex, body-mass index (BMI), smoking status, alcohol assumption and race. Then, the AP index was calculated by summing the product of each pollutant's regression coefficient (β) and its corresponding exposure level, normalized by the sum of the β coefficients. The formula for the AP index is as follows:

$$AP\ index = \frac{\sum_{i=1}^{K} \beta_i \times AP_i}{\sum_{i=1}^{K} \beta_i} \times K \quad (1)$$

where K is the number of pollutants. The resulting AP index represents a joint exposure to multiple air pollutants, incorporating the effects of NO_2, NO_x, $PM_{2.5}$, and PM_{10}, and is intended to capture the combined influence of these pollutants on overall health ratings.

The Living Environment Scores were collected from UKBB (Filed ID: 26417), which were composed of four indicators: social and private housing in poor condition, presence of central heating in houses, air quality, and road traffic accidents. Pearson's correlation test was performed to assess the consistency between the AP index and the Living Environment Scores. Sensitivity analysis was also performed by comparing the AP index constructed from overall health rating with those derived using handgrip strength and walking pace following the same method (Supplementary Texts and Supplementary Fig. 34).

PWAS analysis

A proteome-wide association study (PWAS) was conducted to identify proteins associated with air pollution exposure. We applied general linear regression models using the "limma" R package to assess the relationship between the expression levels of each protein and air pollutants ($PM_{2.5}$, PM_{10}, NO_2, NO_x, and AP index). The limma package employs an empirical Bayes approach to borrow strength across proteins when estimating residual variances, thereby reducing uncertainty in variance estimates and yielding greater statistical power than conventional single-protein t tests[73]. To account for potential confounding, the regression models were adjusted for baseline age, sex, BMI, smoking status, alcohol assumption, and race. Additionally, for each pollutant's PWAS analysis, we included the other pollutants as covariates to adjust for their effects. We used false discovery rate (FDR) correction, the Benjamini-Hochberg (BH) method, to control multiple testing across proteins for each pollutant[74].

Functional enrichment analysis

Enrichment analysis and functional annotation of air pollutants-associated proteins were performed using the "clusterProfiler" R package[75]. Biological processes, metabolic pathways, and biochemical reactions associated with these proteins were characterized by searching against the Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome databases using Metascape[76]. Statistical significance of enriched terms was determined using a false discovery rate (FDR) correction for multiple testing. Terms with FDR < 0.05 and containing at least 10 associated proteins were considered statistically significant. Identical analytical workflows were applied to proteins linked to each pollutant, and the top 10 commonly enriched pathways across all pollutants were prioritized for visualization and downstream interpretation.

To capture pollutant-specific biological effects, proteins uniquely associated with individual pollutants were subjected to Gene Ontology (GO) enrichment analysis. Pathways with FDR < 0.01 were defined as significantly enriched, with the top 10 pathways selected for visualization. To further characterize shared mechanisms among pollutants, proteins co-occurring across pollutants were independently analyzed through identical enrichment workflows.

Cox proportional hazards regression analysis

Cox proportional hazards regression analysis was performed to analyze the associations between air pollution and disease outcomes, adjusting for potential confounders such as baseline age, sex, BMI, smoking status, alcohol assumption and race. Since the air pollution data provided by UKBB was collected in or before 2010, the time to disease onset (duration) was measured from the baseline date (January 1, 2010) to the date of confirmed diagnosis. Follow-up was censored on January 1, 2020, due to the COVID-19 pandemic. Our Cox models considered only first events, in line with standard UKBB practice[31,64]. Cox models were developed for every single disease, and Hazard ratios (HRs) and 95% confidence intervals (CIs) were estimated for each pollutant. Analyses were conducted using the "survival" package in R.

Mediation analysis

Mediation analysis was conducted to investigate the role of proteins in mediating the relationship between air pollution and disease outcomes. Proteins identified as being associated with both air pollution exposure and disease risk were tested for their mediating effect using the "*mediation*" package in R. Proteins involved in the significant indirect effects of air pollution on disease risk were considered as mediators. The calculation process can be formulated as:

$$\begin{cases} M = a \times AP + d \times X + \epsilon_1 \\ D = c' \times AP + b \times M + e \times X + \epsilon_2 \\ D = c \times AP + f \times X + \epsilon_3 \end{cases} \quad (2)$$

where a is the regression coefficient for air pollution in predicting protein levels, b is the effect of the mediator on disease risk, c is the total effect of air pollution on disease risk and d, e, f are the effects of covariates X which includes baseline age, sex, BMI, smoking status,

alcohol assumption and race. The indirect effect $c - c' = a \times b$, which tests the effect of air pollution on disease risk mediated by the mediator. The mediation proportion is calculated as $\frac{c'}{c}$, which represents the ratio of the mediation effect to the total effect. Although the function *mediate()* in the "*mediation*" package does not natively support Cox models, the Poisson-with-offset specification is equivalent to a Cox proportional hazards model for time-to-event data, allowing valid survival-time adjustment and effect decomposition[77]. The indirect effect reported by *mediate()* is the average causal mediation effect (ACME)[78]. We performed false discovery rate (FDR) multiple testing correction for the mediation analysis results of all pollutant-protein-disease pairs. A protein was considered a mediator if its ACME FDR < 0.05, total effect FDR < 0.05, and the mediation proportion was greater than 4% (median value).

To compute the global mediation proportion across all proteins, we performed dimensionality reduction on the proteomic data using principal component analysis (PCA) and derived uncorrelated components that best capture the variability in the proteome data. The advantage of this approach is that it transforms correlated protein variables into low-dimensional orthogonal components by performing spectral decomposition of the covariance matrix of 1,463 proteins, allowing the stepwise estimation of effects through a low-dimensional model. We selected the top 10 principal components (PCs) as mediators for global mediation, which explained 44.80% of the variance in the proteome. A PC was defined as having a significant mediation effect if the p-value for the indirect effect was <0.05 and the direction of the indirect effect was consistent with the effect of air pollution on the disease. The global total indirect effect was calculated as the weighted sum of the mediation effects of the significant PCs, reflecting the contributions of each principal component to the overall mediation pathway between air pollution exposure and disease outcomes. The overall mediation proportion of the proteome was calculated as the ratio of the global total indirect effect to the total effect of air pollution on the disease, which represents how much of the total effect of air pollution on disease risk is mediated by the proteome.

Enrichment analysis of tissue-specific mediators

Mediators that exclusively linked AP index to a single disease type were classified as disease-specific mediators, while those mediating associations across disease types were designated as common mediators. Subsequent GO enrichment analyses were performed separately on each mediator subset using a cutoff of $P < 0.01$ to delineate disease-specific biological pathways.

Development of APPRS

The air pollution protein risk score (APPRS) was developed by integrating proteins identified as mediators in the mediation analysis between AP index and disease outcomes. We included 65 proteins to construct the APPRS based on their mediation performance across multiple diseases. The inclusion criteria were: (a) FDR < 0.05 for both the ACME and total effect in at least two or more AP index-disease mediation models; (b) Mediation proportion > 4%. These thresholds were designed to ensure both the statistical significance and to reflect proteins with broader mediation roles across diseases. To address potential arbitrariness in APPRS construction, we conducted sensitivity analyses using alternative construction criteria, details were provided in Supplementary Texts and Supplementary Fig. 35. Then, APPRS was calculated as a weighted sum of the protein expressions associated with air pollution, with the weights derived from their average causal mediation effect (ACME), which can be denoted as:

$$\mathrm{APPRS} = \frac{\sum_i^M (ACME_i \times Protein_i)}{\sum_i^M ACME_i} \quad (3)$$

where $ACME_i$ is the coefficient of indirect effects for ith protein, $Protein_i$ the expression level of ith protein, M the total number of selected proteins. Almost all ACME values used to construct the APPRS were positive, whereas protein expression values included both positive and negative numbers due to normalization. For each disease, we derived a disease-specific APPRS by combining protein levels using the corresponding disease-specific ACME coefficients from the mediation models as weights.

Association between APPRS and observed disease rates

To examine the relationship between APPRS and disease incidence, we divided the study population into 100 groups based on APPRS quantiles, calculated the disease incidence for each group, and generated a figure depicting the results. The Pearson correlation coefficient was used to assess the relationship between APPRS percentiles and disease incidence rates.

External validation of APPRS

To assess the generalizability of the APPRS and its association with health outcomes, we validated it within two independent cohorts: the Panmonogenic cohort and the HERITAGE Family Study cohort. Both cohorts provided proteomic profiles and at least one relevant health phenotype. The validation analyses using the HERITAGE and Panmonogenic cohorts relied on publicly available data. The Panmonogenic cohort provides proteomics data of individuals with 22 monogenic immune-mediated diseases[66]. The HERITAGE Family Study cohort reports the cardiorespiratory fitness (CRF) levels of 481 sedentary adults from 99 families[79]. For both cohorts, 12 disease-specific APPRSs were constructed for each cohort using the same set of mediating proteins and weights as in the UKBB cohort. Missing values among the 65 mediating proteins were imputed as zero, and APPRSs were computed following the same methodological framework applied in UKBB.

In the Panmonogenic cohort, a global APPRS was calculated by averaging the 12 disease-specific scores to quantify overall susceptibility to air pollution-related protein-mediated disease risk. This global APPRS was then used to compare risk levels between healthy individuals and those with immune-related diseases (Autoimmune Diseases, AID; Primary Immunodeficiency, PID; Telomere Diseases, Telo). In the HERITAGE cohort, individuals' maximal oxygen uptake (VO_2max)—a direct measure of cardiorespiratory fitness (CRF)—was recorded. Participants were stratified into tertiles based on VO_2max levels, representing low, intermediate, and high CRF groups. We subsequently assessed the association between APPRSs and these CRF categories.

Interaction between APPRS and air pollution on disease risk

To evaluate the joint interaction effects of APPRS with air pollution exposure on disease risk, we performed a stratified Cox proportional hazards regression analysis. Participants were categorized into tertiles based on their air pollution exposure levels (low, intermediate, high), and the APPRS was dichotomized into low and high groups using the median as the cutoff. To avoid potential collinearity, we examined the correlation between APPRS and air pollution exposures and found that their associations were too weak to cause multicollinearity (Supplementary Fig. 18). Then, a combined stratification variable was generated by jointly stratifying participants based on their surrounding exposure and APPRS levels (3 exposure levels × 2 APPRS levels). Cox proportional hazards regression analysis was performed to examine the incidence of 12 diseases, while the lowest-risk stratum (APPRS-low and exposure-low) serves as the reference. All models were adjusted for relevant covariates, including age, sex, BMI, smoking status, and alcohol consumption.

Construction of disease prediction models

LASSO prediction regression models were developed to assess the early onset of diseases associated with air pollution exposure, using the "glmnet" R package, predictors incorporated APPRS, basic features (age, sex, smoking, alcohol consumption, race, and BMI), and clinical features (ApoB/ApoA, albumin, creatinine, HbA1c, and glucose). To assess the predictive power of the Air Pollution Protein Risk Score (APPRS) for disease risk, we established three baseline models: (1) basic demographic features alone, (2) clinical features alone, and (3) a combination of basic and clinical features. Each of these baseline models was then augmented with the APPRS to evaluate its added value in predicting 12 distinct disease outcomes. The analysis was conducted using baseline data from both healthy individuals and patients. Model training and evaluation were performed using 10-fold cross-validation to ensure robustness and generalizability. Predictive performance was measured by the area under the receiver operating characteristic curve (AUROC), and improvements in classification accuracy were assessed by comparing the baseline models with their corresponding APPRS-enhanced counterparts.

Evaluation of APPRS contribution in the prediction model

To quantify the contribution of predictors, particularly the APPRS, to disease prediction, we extracted the absolute values of the final feature coefficients from the combined model incorporating basic features, clinical features, and APPRS. Before fitting the model, all features were standardized, ensuring that the coefficient magnitudes are directly comparable across variables. These coefficients, derived from the LASSO regression, reflect the relative importance of each predictor in the model, with larger absolute values indicating stronger associations with disease risk. Then we calculated the proportion of each feature's absolute coefficient relative to the total sum of all absolute coefficients in the combined model (basic + clinical + APPRS). These proportions were then visualized using bar plots. This approach quantifies the weight of each variable in the model's decision-making process, with higher proportions indicating stronger contributions to disease risk stratification.

Sensitivity analysis of APPRS to the prediction models

To confirm that the predictive contribution of the APPRS was neither stochastic nor confounded by the AP index, we conducted sensitivity analyses by constructing three alternative LASSO models. Firstly, we randomly selected 65 proteins from the 1463 proteins and calculated the APPRS using their corresponding ACME coefficients. Secondly, we restricted our selection to those proteins showing significant associations with the corresponding disease ($P < 0.05$), again selecting 65 proteins for APPRS calculation. 1000 iterations of random selection were performed, and the final value was derived as the mean value across all iterations. Thirdly, the APPRS was substituted with the AP index alone to isolate its specific contribution from APPRS. These values were trained with basic and clinical information, and models were evaluated under identical cross-validation and frameworks as the primary analysis. By comparing the performance (AUROC) in these alternative models against the APPRS-enhanced model, we rigorously assessed whether the predictive power attributed to the APPRS arose from systematic biological signals rather than random chance or redundancy with the AP index.

External validation of the APPRS contribution

To validate the generalizability of the APPRS in predicting health outcomes, we performed external validation in the HERITAGE cohort. The Panmonogenic cohort was excluded from external validation due to the unavailability of publicly accessible individual-level protein expression data. We focused on cardiorespiratory fitness (CRF) as the health outcome. Basic models were constructed by including age, sex, race, and BMI as predictors. Arrhythmias and peripheral artery disease APPRS were further added to the basic models for comparison. Model performance was evaluated and compared using the area under the receiver operating characteristic curve (AUROC). This approach quantified the incremental predictive power conferred by the APPRS. The method for quantifying and visualizing the contribution of predictors was the same as that of the UKBB study.

Statistical analysis

All statistical analyses were conducted using R (version 4.1.3) and relevant R packages ("survival" [v3.6-4] for Cox proportional hazards regression, "limma" [v3.50.3] for PWAS analysis, "mediation" [v4.5.0] for mediation analysis, "glmnet" [v4.1-8] and "caret" [v6.0-94] for LASSO regression models, "stats" for basic statistical analyses, "prcomp" [v4.4.1] for PCA analysis, and "pROC" [v1.18.5] for AUROC calculation).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Datasets generated in this study are made available in the Supplementary Data files. Source data are provided with this paper. This research has been conducted using the UK Biobank Resource under Application Number 77803. Proteomics data, air pollution data, and disease outcomes are available as part of the UK Biobank. The data can be accessed through the UK Biobank Research Analysis Portal (https://www.ukbiobank.ac.uk/enable-your-research). The HERITAGE cohort proteomics dataset is publicly available via the MoTrPAC Data Hub (https://motrpac-data.org/related-studies/heritage-proteomics). Publicly accessible proteomics summary tables from the Pan-monogenic cohort are available through the Pan-Monogenic dashboard (https://panmonogenic.yale.edu/). Source data are provided with this paper.

Code availability

The custom code used for data processing and statistical analysis in this study has been deposited on GitHub at https://github.com/kaixuanli99/AirPollution-Proteome-Diseases-Mediation, and stored at https://doi.org/10.5281/zenodo.17774180[80].

References

1. Gao, X., Jiang, M., Huang, N., Guo, X. & Huang, T. Long-term air pollution, genetic susceptibility, and the risk of depression and anxiety: a prospective study in the UK biobank cohort. *Environ. Health Perspect.* **131**, 017002 (2023).
2. Tian, F. et al. Air pollution associated with incident stroke, post-stroke cardiovascular events, and death. *Neurology* **99**, e2474–e2484 (2022).
3. Ma, Y. et al. Genetic susceptibility modifies relationships between air pollutants and stroke risk: a large cohort study. *Stroke* **55**, 113–121 (2024).
4. Zhang, J. et al. Association of combined exposure to ambient air pollutants, genetic risk, and incident rheumatoid arthritis: a prospective cohort study in the UK Biobank. *Environ. Health Perspect.* **131**, 037008 (2023).
5. Wang, X. et al. Ambient air pollution associated with incident asthma, subsequent cardiovascular disease and death: A trajectory analysis of a national cohort. *J. Hazard Mater.* **460**, 132372 (2023).
6. Liu, W. et al. Exposure to ambient air pollutants during circadian syndrome and subsequent cardiovascular disease and its subtypes and death: A trajectory analysis. *Sci. Total Environ.* **944**, 173777 (2024).
7. Wu, G. et al. Ambient air pollution and incidence, progression to multimorbidity and death of hypertension, diabetes, and chronic

kidney disease: A national prospective cohort. *Sci. Total Environ.* **881**, 163406 (2023).
8. Wang, M. et al. Joint exposure to various ambient air pollutants and incident heart failure: a prospective analysis in UK Biobank. *Eur. Heart J.* **42**, 1582–1591 (2021).
9. Parra, K. L., Alexander, G. E., Raichlen, D. A., Klimentidis, Y. C. & Furlong, M. A. Exposure to air pollution and risk of incident dementia in the UK Biobank. *Environ. Res.* **209**, 112895 (2022).
10. Li, Z. H. et al. Long-term air pollution exposure, habitual physical activity, and incident chronic kidney disease. *Ecotoxicol. Environ. Saf.* **265**, 115492 (2023).
11. Wang, J. et al. Long-term Exposure to Ambient Air Pollutants and Increased Risk of Pneumonia in the UK Biobank. *CHEST* **164**, 39–52 (2023).
12. Li, D. et al. Long-term exposure to ambient air pollution, genetic susceptibility, and the incidence of bipolar disorder: A prospective cohort study. *Psychiatry Res.* **327**, 115396 (2023).
13. Chen, T. et al. Dietary polyunsaturated fatty acids intake, air pollution, and the risk of lung cancer: A prospective study in UK Biobank. *Sci. Total Environ.* **882**, 163552 (2023).
14. Smotherman, C. et al. Association of air pollution with postmenopausal breast cancer risk in UK Biobank. *Breast Cancer Res* **25**, 83 (2023).
15. Li, F. R. et al. Long-term exposure to air pollution and incident nonalcoholic fatty liver disease and cirrhosis: A cohort study. *Liver Int.* **43**, 299–307 (2023).
16. Liu, X. et al. Associations of reproductive risk score and joint exposure to ambient air pollutants with chronic obstructive pulmonary disease: a cohort study in UK Biobank. *Environ. Health Prev. Med.* **28**, 76 (2023).
17. Huang, Y. et al. Air Pollution, Genetic Factors, and the Risk of Lung Cancer: A Prospective Study in the UK Biobank. *Am. J. Respir. Crit. Care Med.* **204**, 817–825 (2021).
18. Xu, C. et al. Association of air pollutants and osteoporosis risk: The modifying effect of genetic predisposition. *Environ. Int* **170**, 107562 (2022).
19. Li, F. R. et al. Long-term exposure to air pollution and risk of incident inflammatory bowel disease among middle and old aged adults. *Ecotoxicol. Environ. Saf.* **242**, 113835 (2022).
20. Li, X. et al. Obesity and the relation between joint exposure to ambient air pollutants and incident type 2 diabetes: A cohort study in UK Biobank. *PLoS Med* **18**, e1003767 (2021).
21. Luo, P. et al. Air Pollution and Allergic Rhinitis: Findings from a Prospective Cohort Study. *Environ. Sci. Technol.* **57**, 15835–15845 (2023).
22. Zhu, Y. et al. Ambient air pollution, lifestyle, and genetic predisposition on all-cause and cause-specific mortality: A prospective cohort study. *Sci. Total Environ.* **933**, 173120 (2024).
23. Ji, J. S., Zhu, A., Lv, Y. & Shi, X. Interaction between residential greenness and air pollution mortality: analysis of the Chinese Longitudinal Healthy Longevity Survey. *Lancet Planet Health* **4**, e107–e115 (2020).
24. Fuller, R. et al. Pollution and health: a progress update. *Lancet Planet Health* **6**, e535–e547 (2022).
25. Tainio, M. et al. Air pollution, physical activity and health: A mapping review of the evidence. *Environ. Int.* **147**, 105954 (2021).
26. Dominski, F. H. et al. Effects of air pollution on health: A mapping review of systematic reviews and meta-analyses. *Environ. Res.* **201**, 111487 (2021).
27. Wu, H., Eckhardt, C. M. & Baccarelli, A. A. Molecular mechanisms of environmental exposures and human disease. *Nat. Rev. Genet.* **24**, 332–344 (2023).
28. Peters, A., Nawrot, T. S. & Baccarelli, A. A. Hallmarks of environmental insults. *Cell* **184**, 1455–1468 (2021).
29. Mookherjee, N. et al. Defining the effects of traffic-related air pollution on the human plasma proteome using an aptamer proteomic array: A dose-dependent increase in atherosclerosis-related proteins. *Environ. Res.* **209**, 112803 (2022).
30. Liao, J. et al. Ambient air pollution and serum cardiometabolic proteome in young adults with obesity history. *ISEE Conference Abstracts* **2023**, https://doi.org/10.1289/isee.2023.PK-038.
31. You, J. et al. Plasma proteomic profiles predict individual future health risk. *Nat. Commun.* **14**, 7817 (2023).
32. Carrasco-Zanini, J. et al. Proteomic signatures improve risk prediction for common and rare diseases. *Nat. Med.* **30**, 2489–2498 (2024).
33. Deng, Y. T. et al. Atlas of the plasma proteome in health and disease in 53,026 adults. *Cell* **188**, 253–271.e257 (2025).
34. Pilz, V. et al. C-reactive protein (CRP) and long-term air pollution with a focus on ultrafine particles. *Int. J. Hyg. Environ. Health* **221**, 510–518 (2018).
35. Tian, F. et al. Air pollution, APOE genotype and risk of dementia among individuals with cardiovascular diseases: A population-based longitudinal study. *Environ. Pollut.* **347**, 123758 (2024).
36. Cheng, H. et al. Ambient air pollutants and traffic factors were associated with blood and urine biomarkers and asthma risk. *Environ. Sci. Technol.* **56**, 7298–7307 (2022).
37. de Prado-Bert, P. et al. Short- and medium-term air pollution exposure, plasmatic protein levels and blood pressure in children. *Environ. Res* **211**, 113109 (2022).
38. Pan, C. et al. Large-scale plasma proteomics uncovers novel targets linking ambient air pollution and depression. *Mol. Psychiatry*, https://doi.org/10.1038/s41380-025-02953-x (2025).
39. Ma, Y. et al. Air pollution, genetic susceptibility, and the risk of atrial fibrillation: A large prospective cohort study. *Proc. Natl. Acad. Sci. USA* **120**, e2302708120 (2023).
40. Hooper, L. G. & Kaufman, J. D. Ambient Air Pollution and Clinical Implications for Susceptible Populations. *Ann. Am. Thorac. Soc.* **15**, S64–s68 (2018).
41. Zhong, P. et al. Individual-level modifiers of the acute effects of air pollution on mortality in Wuhan, China. *Glob. Health Res. Policy* **3**, 27 (2018).
42. Kingma, B., Sullivan-Kwantes, W., Castellani, J., Friedl, K. & Haman, F. We are all exposed, but some are more exposed than others. *Int. J. Circumpolar Health* **82**, 2199492 (2023).
43. Cristobal, A. et al. Personalized proteome profiles of healthy and tumor human colon organoids reveal both individual diversity and basic features of colorectal cancer. *Cell Rep.* **18**, 263–274 (2017).
44. Kosnik, M. B., Enroth, S. & Karlsson, O. Distinct genetic regions are associated with differential population susceptibility to chemical exposures. *Environ. Int.* **152**, 106488 (2021).
45. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–d592 (2023).
46. Gillespie, M. et al. The Reactome Pathway Knowledgebase 2022. *Nucleic Acids Res.* **50**, D687–d692 (2022).
47. Clark-Boucher, D. et al. Methods for mediation analysis with high-dimensional DNA methylation data: Possible choices and comparisons. *PLoS Genet.* **19**, e1011022 (2023).
48. Clark-Boucher, D. et al. Methods for Mediation Analysis with High-Dimensional DNA Methylation Data: Possible Choices and Comparison. *medRxiv* https://doi.org/10.1101/2023.02.10.23285764 (2023).
49. Knox, C. et al. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Res.* **52**, D1265–d1275 (2024).
50. Haidar, A. & Horwich, T. Obesity, Cardiorespiratory Fitness, and Cardiovascular Disease. *Curr. Cardiol. Rep.* **25**, 1565–1571 (2023).

51. Größer, V. et al. Association of cardiorespiratory fitness level with vascular function and subclinical atherosclerosis in the elderly. *Eur. J. Appl Physiol.* **124**, 1487–1497 (2024).
52. Laukkanen, J. A., Lavie, C. J., Khan, H., Kurl, S. & Kunutsor, S. K. Cardiorespiratory Fitness and the Risk of Serious Ventricular Arrhythmias: A Prospective Cohort Study. *Mayo Clin. Proc.* **94**, 833–841 (2019).
53. Keshvani, N. et al. Midlife Cardiorespiratory Fitness and the Development of Peripheral Artery Disease in Later Life. *J. Am. Heart Assoc.* **10**, e020841 (2021).
54. Barbier, E. et al. Oxidative stress and inflammation induced by air pollution-derived PM(2.5) persist in the lungs of mice after cessation of their sub-chronic exposure. *Environ. Int.* **181**, 108248 (2023).
55. López-Sánchez, L. M., Aranda, E. & Rodríguez-Ariza, A. Nitric oxide and tumor metabolic reprogramming. *Biochem. Pharm.* **176**, 113769 (2020).
56. Ma, Y. H. et al. Association of Long-term Exposure to Ambient Air Pollution With Cognitive Decline and Alzheimer's Disease-Related Amyloidosis. *Biol. Psychiatry* **93**, 780–789 (2023).
57. Liu, Q. et al. Ambient particulate air pollution and circulating C-reactive protein level: A systematic review and meta-analysis. *Int. J. Hyg. Environ. Health* **222**, 756–764 (2019).
58. He, S. et al. Ambient air pollution and inflammation-related proteins during early childhood. *Environ. Res* **215**, 114364 (2022).
59. Gao, Y. et al. Extracellular vesicles derived from PM2.5-exposed alveolar epithelial cells mediate endothelial adhesion and atherosclerosis in ApoE(-/-) mice. *FASEB J.* **36**, e22161 (2022).
60. Rui, W., Guan, L., Zhang, F., Zhang, W. & Ding, W. PM2.5-induced oxidative stress increases adhesion molecules expression in human endothelial cells through the ERK/AKT/NF-κB-dependent pathway. *J. Appl. Toxicol.* **36**, 48–59 (2016).
61. Chachlaki, K. & Prevot, V. Nitric oxide signalling in the brain and its control of bodily functions. *Br. J. Pharm.* **177**, 5437–5458 (2020).
62. Carberry, C. K. & Rager, J. E. The impact of environmental contaminants on extracellular vesicles and their key molecular regulators: A literature and database-driven review. *Environ. Mol. Mutagen* **64**, 50–66 (2023).
63. Niu, Y. et al. Ozone exposure and prothrombosis: Mechanistic insights from a randomized controlled exposure trial. *J. Hazard Mater.* **429**, 128322 (2022).
64. Gadd, D. A. et al. Blood protein assessment of leading incident diseases and mortality in the UK Biobank. *Nat. Aging* **4**, 939–948 (2024).
65. Lin, L. et al. The airway microbiome mediates the interaction between environmental exposure and respiratory health in humans. *Nat. Med.* **29**, 1750–1759 (2023).
66. Sparks, R. et al. A unified metric of human immune health. *Nat. Med.* **30**, 2461–2472 (2024).
67. Ren, M. et al. Using environmental mixture exposure-triggered biological knowledge-driven machine learning to predict early pregnancy loss. *Environ. Sci. Technol.* **59**, 19691–19704 (2025).
68. Wang, B. et al. ExposomeX: Development of an Integrative Exposomic Platform to Expedite Discovery of the "Exposure–Biology –Disease" Nexus. *Environ. Sci. Technol.* **59**, 13251–13263 (2025).
69. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
70. Chen, L. et al. Risk/benefit tradeoff of habitual physical activity and air pollution on chronic pulmonary obstructive disease: findings from a large prospective cohort study. *BMC Med.* **20**, 70 (2022).
71. Sun, B. B. et al. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
72. Eldjarn, G. H. et al. Large-scale plasma proteomics comparisons through genetics and disease associations. *Nature* **622**, 348–358 (2023).
73. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
74. Wei, Y. et al. Short term exposure to fine particulate matter and hospital admission risks and costs in the Medicare population: time stratified, case crossover study. *BMJ* **367**, l6258 (2019).
75. Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovations* **2**, 100141 (2021).
76. Zhou, Y. et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).
77. Austin, P. C. A Tutorial on Multilevel Survival Analysis: Methods, Models and Applications. *Int. Stat. Rev.* **85**, 185–203 (2017).
78. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. Mediation: R Package for Causal Mediation Analysis. *J. Stat. Softw.* **59**, 1–38 (2014).
79. Robbins, J. M. et al. Human plasma proteomic profiles indicative of cardiorespiratory fitness. *Nat. Metab.* **3**, 786–797 (2021).
80. Li, W. et al. Plasma proteome mediates the associations between air pollution exposure and disease risk. *Zenodo*, https://doi.org/10.5281/zenodo.17774180 (2025).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC; 32325013 to SW, 32200472 to WL, 92043301 to HK, and 32530022 to SW), Chinese Academy of Sciences Young Team Program for Stable Support of Basic Research (YSBR-077 to SW), Shanghai Science Technology and Innovation Action Plan (24JS2810300 to SW), the National Key Research and Development Project (2024YFC3405802 to SW), Shanghai Science and Technology Commission Excellent Academic Leaders Program (22XD1424700 to SW), and the Strategic Priority Research Program of Chinese Academy of Sciences (XDB38020400 to SW).

Author contributions

S.W., H.K., and W.L. conceived and designed the study. W.L. and K.L. did statistical analyses and wrote the manuscript. G.Zh., P.Z., Y.C., Y.Z., X.M., and Y.N. contributed to data collection and the overall structure of the study. S.W., H.K., and G.Zo. reviewed and edited the manuscript. S.W. and H.K. supervised the study. All authors read and approved the final manuscript and had final responsibility for the decision to submit for publication.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-68972-6.

Correspondence and requests for materials should be addressed to Haidong Kan or Sijia Wang.

Peer review information *Nature Communications* thanks Xiao Wu, Bin Wang, and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.