Research Square

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

Global Delayed Lung Cancer Burden from Ambient Air Pollution

Bin Wang

binwang@pku.edu.cn

Peking University https://orcid.org/0000-0002-1164-8430

Han Zhang

Peking University

Xiangtao Zhang

University of Electronic Science and Technology of China

Le Zhang

University of Electronic Science and Technology of China

Huizhong Shen

Southern University of Science and Technology https://orcid.org/0000-0003-1335-8477

Bo Fu

Institute of Reproductive and Child Health, School of Public Health, Peking University Health Science Center https://orcid.org/0000-0003-1268-6348

Runmei Ma

National Institutes of Environmental

Qianyi Wang

Peking University

Bo Pan

Department of Environmental Science and Engineering, Kunming University of Science and Technology

Mingliang Fang

Fudan University https://orcid.org/0000-0002-2204-9783

Tiantian Li

National Institute of Environmental Health, Chinese Center for Disease Control and Prevention https://orcid.org/0000-0001-5865-0573

Guofeng Shen

College of Urban and Environmental Sciences, Peking University https://orcid.org/0000-0002-7731-5399

Jingchuan Xue

Guangdong Basic Research Center of Excellence for Ecological Security and Green Development, Key Laboratory for City Cluster Environmental Safety and Green Development of the Ministry of Education,

Zhipeng Bai

Chinese Research Academy of Environmental Sciences

Xianjin Tang

Zhejiang University

Yuanchen Chen

College of Energy and Carbon Neutrality Science and Education Integration, Zhejiang University of Technology

---

Article

Keywords: Lung cancer, Benzo[a]pyrene, Geographical disparity, Intergenerational Inequality, Disease burden

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9647262/v1

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

Additional Declarations: There is **NO** Competing Interest.

Title: Global Delayed Lung Cancer Burden from Ambient Air Pollution

Authors: Han Zhang[1,2,#], Xiangtao Zhang[3,#], Le Zhang[3*], Huizhong Shen[4], Bo Fu[1,2], Runmei Ma[1,2], Qianyi Wang[1,2], Bo Pan[5], Mingliang Fang[6], Tiantian Li[7], Guofeng Shen[8], Jingchuan Xue[9], Zhipeng Bai[10], Xianjin Tang[11], Yuanchen Chen[12*], Bin Wang[1,2,8,13*]

Affiliations:

[1] Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China

[2] Institute of Reproductive and Child Health, School of Public Health Peking University Beijing 100191, China/Key Laboratory of Reproductive Health, National Health and Family Planning Commission of PR China, Beijing 100191, China

[3] School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China

[4] State Key Laboratory of Soil Pollution Control and Safety, Shenzhen Key Laboratory of Precision Measurement and Early Warning Technology for Urban Environmental Health Risks, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China

[5] Yunnan Provincial Key Lab of Soil Carbon Sequestration and Pollution Control, Faculty of Environmental Science and Engineering, Kunming University of Science and Technology, Kunming 650500, China

[6] Department of Environmental Science and Engineering, Fudan University, Shanghai 200433, China

[7] National Key Laboratory of Intelligent Tracking and Forecasting for Infectious Diseases, National Institute of Environmental Health, Chinese Center for Disease Control and Prevention, Beijing 100021, China

[8] College of Urban and Environmental Science, Peking University, Beijing 100871, China

[9] Guangdong Basic Research Center of Excellence for Ecological Security and Green Development, Key Laboratory for City Cluster Environmental Safety and Green Development of the Ministry of Education, School of Ecology, Environment and Ocean, Guangdong University of Technology, Guangzhou, 510006, China

[10] State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences, Beijing 100012, China

[11] Institute of Soil and Water Resources and Environmental Science, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China.

[12] State Key Laboratory of Green Chemical Synthesis and Conversion, Zhejiang Key Laboratory of Clean Energy Conversion and Utilization, Science and Education Integration College of Energy and Carbon Neutralization, Zhejiang University of Technology, Zhejiang, 310014, China

[13] Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing 100191, China

[#] These authors equally contributed to this study.

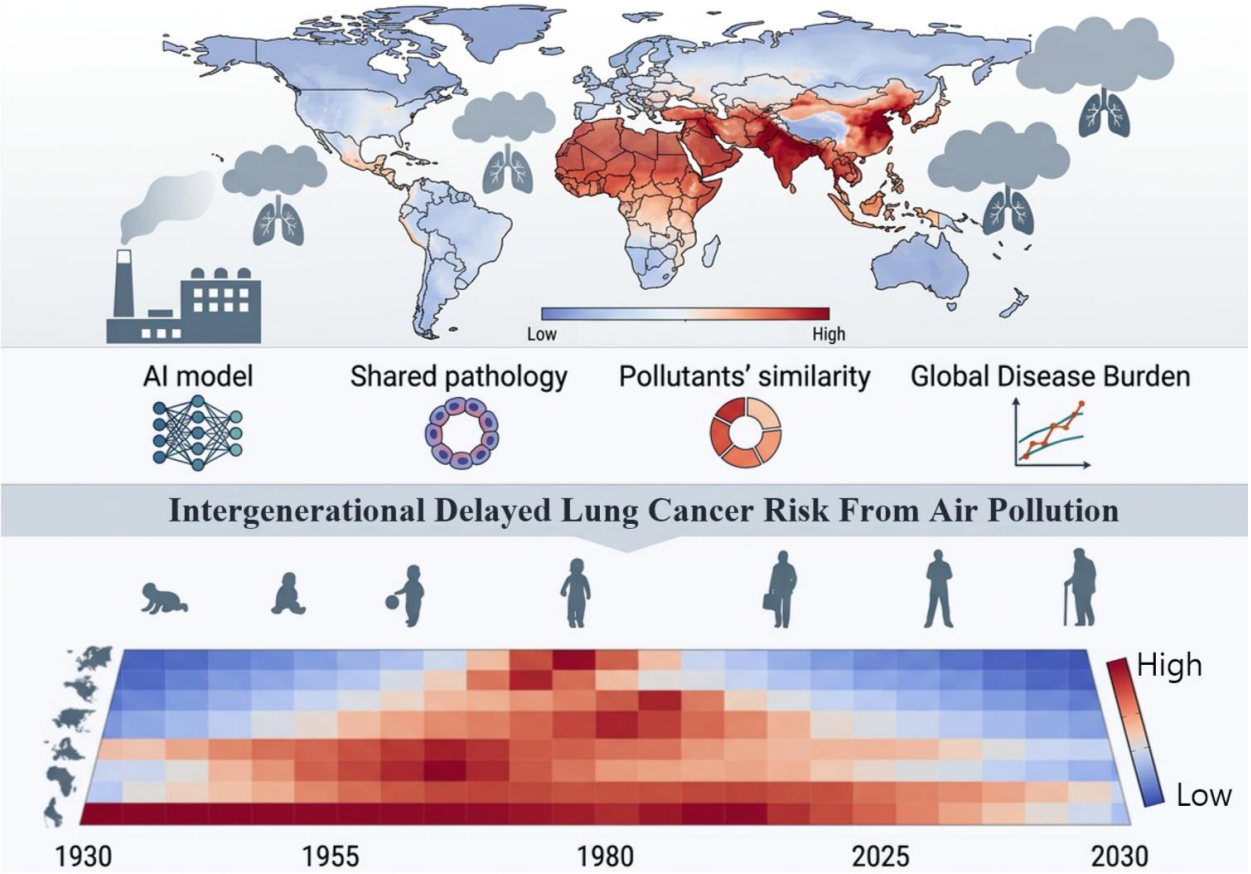
[*]Correspondence and requests for materials should be addressed to Dr. Le Zhang (lezhang@uestc.edu.cn); Dr. Yuanchen Chen (chenyuanchen1988@zjut.edu.cn), Dr. Dr. Bin Wang (binwang@pku.edu.cn)

Abstract:
Ambient air pollution remains a major cause of lung cancer (LC), yet a central paradox persists: why does a substantial LC burden endure even as average pollution levels decline in many parts of the world? Here we address this question by using benzo[a]pyrene (BaP) as a mechanistically anchored marker of carcinogenic air pollution mixtures, supported by shared emission sources, convergent biological pathways, and strong concordance between BaP-based population attributable fractions (PAFs) and air pollution–attributable LC mortality. We reconstructed global ambient BaP concentrations from 1850 to 2100 and found that levels rose approximately fivefold to a peak in 2010–2015, before projected declines of 50–80% by 2050 under future control scenarios. Despite these declines, incorporation of carcinogenic latency and population age structure shifted the peak of attributable LC deaths to 2035–2045, while reducing estimated attributable fractions by 40–70% relative to conventional steady-state assumptions. Birth cohorts exposed during rapid industrial expansion from 1960 to 1990 are projected to experience lifetime attributable risks up to twofold higher than those born after 2000, whereas cohorts born after major regulatory transitions in high-income countries show sustained risk reductions. By integrating historical exposure reconstruction with latency-aware cancer burden modeling, our framework provides the first high-resolution global mapping of air pollution–related LC burden across past, present, and future periods. These findings reveal that the cancer consequences of air pollution are not only delayed, but also structurally unequal across regions and generations, underscoring the need for prevention strategies that account for long-term exposure legacies as well as future emission control.

Keywords: Lung cancer, Benzo[a]pyrene, Geographical disparity, Intergenerational Inequality, Disease burden

One-Sentence Summary: Air pollution leaves an unequal inheritance of lung cancer risk across generations in various regions, even as global pollution levels decline.

Graphical Abstract:



Ambient air pollution is a major driver of LC, but its impacts are shaped by delayed carcinogenic effects and complex pollutant mixtures. Using AI-based exposure reconstruction and benzo[a]pyrene (BaP) as a mechanistic marker, we map the global burden of air pollution–related LC. Incorporating historical exposures, latency, and demographic dynamics reveals spatial hotspots and delayed peaks in risk. Cohorts exposed during rapid industrialization face substantially higher lifetime risks, highlighting persistent intergenerational inequalities.

1. Introduction

Ambient air pollution is among the leading environmental risk factors for lung cancer (LC) worldwide (see Extended Data Fig. 1), contributing substantially to premature mortality and long-term disease burden[1,2]. Air quality has improved in many regions over recent decades, whereas LC attributable to air pollution remains unevenly distributed across populations and places[3]. This persistence raises a first fundamental question: ***How do substantial LC burdens continue to emerge as global average pollution levels decline?*** Birth cohorts may face markedly different lifetime risks depending on when peak exposures coincide with vulnerable life stages. For example, environmental risks are increasingly shaped by birth cohort, with younger generations projected to experience disproportionate lifetime exposure under continued emissions[4]. Distinguishing intergenerational inequalities is therefore critical for understanding the full public health consequences of air pollution[5]. Likewise, global assessments of air pollution-attributable disease burden tend to obscure substantial heterogeneity across regions and populations[6]. Spatial differences in exposure histories, socioeconomic development, and demographic structure can generate pronounced disparities in disease burden, even when global mean exposure declines[7]. Yet such information remains largely unavailable in global cancer assessments below the national level.

The first critical challenge lies in the temporal disconnect between exposure and disease. LC attributable to environmental carcinogens is marked by long latency, often 10 to 30 years between exposure and diagnosis[8,9]. Yet most global assessments implicitly assume that exposure levels in a given year represent lifetime exposure, an assumption that obscures cumulative risk and distorts burden attribution during periods of rapid exposure change[10,11]. This delay obscures the true temporal link between pollution and health outcomes, causing estimation bias of risks. For example, assessments of polycyclic aromatic hydrocarbons (PAHs), a key class of carcinogenic air pollutants, have typically been based on limited regional measurements or static emission inventories, while overlooking the long latency between exposure and disease onset[10,12,13]. This practice can only effectively equate current pollution levels with cancer risk, despite well-established delays in LC development.

Second, attributing LC risk to ambient air pollution remains challenging because air pollution is not a single exposure but a complex mixture of chemically diverse pollutants emitted from multiple sources[1,2,14]. These components differ in toxicity, atmospheric behavior, and dose-response relationships, making it difficult to quantify long-term health impacts in a consistent and mechanistically interpretable way[15,16]. Even fine particulate matter ($PM_{2.5}$), the most widely used regulatory indicator, represents a heterogeneous mixture whose chemical composition varies substantially across sources, regions, and time[14]. This heterogeneity complicates efforts to derive stable and generalizable risk estimates for long-term disease burden assessment. Although epidemiological evidence linking air pollution to LC mortality is robust, mechanistic and dose-response evidence exists for only a limited number of carcinogenic components. This gap highlights the need for a representative marker that captures the carcinogenic dimension of complex air pollution mixtures. Among known pollutants, benzo[a]pyrene (BaP), a prototypical polycyclic aromatic hydrocarbon (PAH) formed during incomplete combustion, has long been used as an indicator of cancer risk from air pollution[10,17]. While previous studies have characterized PAH distributions and associated health risks in specific regions or time periods[10,12,13], translating mechanistic knowledge of individual pollutants into global, long-term estimates of LC burden for complex pollution mixtures remains a major challenge.

Third, reconstructing long-term exposure trajectories is further constrained by data scarcity, especially for the complex constitutes of air pollution[3]. Reliable global measurements of

carcinogenic air pollutants are limited prior to the late twentieth century, and monitoring coverage remains highly uneven across regions. Consequently, many assessments rely on short observational windows or simplified exposure proxies, limiting their ability to capture the legacy of historical emissions. Even for world-widely concerned PAH components, the monitoring data are distributed in urban areas, and these discrete points are difficult to use for spatial-temporal modeling[18,19]. This gap is especially consequential for regions that industrialized rapidly without early pollution controls, where the legacy of past exposures continues to shape present-day health outcomes. Artificial intelligence (AI) or machine learning (ML) provides a transformative advantage by integrating heterogeneous data streams, capturing nonlinear relationship from emission to concentration, and enhancing both spatial and temporal resolution beyond the capacity of traditional inventories[20]. Within this framework, the ML strategies focusing on using the local characteristic to predict the air pollution level are promising to improve the prediction power. As a foundation model using transformer-based prior-data-fitted network for tabular data (TabPFN) leverages pretrained knowledge for small-sample learning, allowing accurate generalization in data-sparse settings[21]. This may critically facilitate reconstructing century-scale BaP concentrations, where observations are limited in both space and time.

Here, we develop an integrated framework to reassess global high-resolution LC burden attributable to ambient air pollution through a latency-aware and inequality-focused lens (Fig. 1). The analysis proceeds in three conceptually linked steps. We first reconstruct global BaP concentrations, a well-established carcinogenic component of air pollution, at high spatiotemporal resolution spanning 1850–2100 using a small-sample foundation AI model that fuses historical measurements, emissions information, and spatial–temporal constraints to generate high-resolution global BaP concentration fields (0.1°×0.1°). We then translate historical exposure trajectories into LC risk by explicitly accounting for disease latency and population age structure, avoiding steady-state lifetime exposure assumptions (Fig. 1A). Second, we extend the BaP-based framework to ambient air pollution more broadly by treating BaP as a mechanistic marker of carcinogenic air pollution mixtures. The validity of this approach is supported by three independent lines of evidence: i.e., (i) strong correlations between BaP and other air pollution components; (ii) shared carcinogenic pathways and biological mechanisms identified through knowledge graph analysis; and (iii) a high degree of concordance between country-level BaP-attributable PAFs and LC mortality attributable to ambient air pollution. In sum, these complementary lines of evidence enable robust projection of high-resolution LC burden attributable to ambient air pollution (Fig. 1B). Finally, leveraging the reconstructed global burden surfaces, we quantify both delayed LC burden and intergenerational inequalities. Regional analyses reveal pronounced geographic disparities, while birth cohort–based assessments uncover persistent and emerging inequalities across time (Fig. 1C).

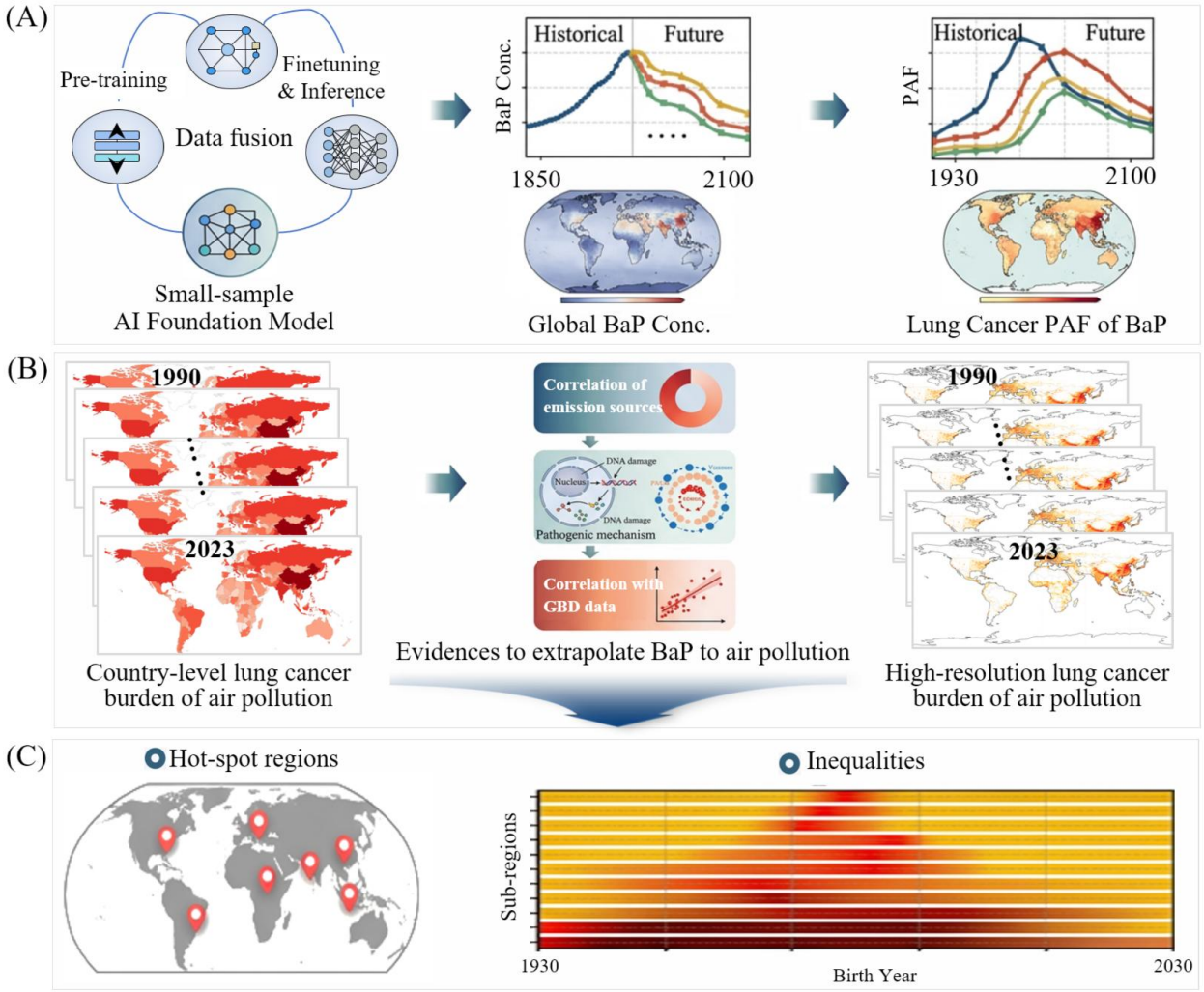


Fig. 1. Study design. (A) Reconstruction of global benzo[a]pyrene (BaP) concentrations (1850–2100) using a small-sample AI foundation model integrating historical emissions and scenario projections. Latency-aware population attributable fractions (PAFs) were applied to translate exposure trajectories into high-resolution global disease burden of lung cancer. (B) Extension from BaP to ambient air pollution–attributable lung cancer burden. BaP is used as a representative carcinogenic marker of ambient air pollution supported by three independent lines of evidence. (C) Quantification of global inequalities across regions and countries.

2. Results

2.1 Century-scale Reconstruction of Global BaP Exposure

Model evaluation during the independent test period demonstrates strong predictive performance of the BaP reconstruction framework (Extended Data Fig. 2). Across the global domain, the TabPFN model consistently achieves higher coefficients of determination (R^2) and lower root mean square errors (RMSE) compared with conventional ensemble models of random Forest (RF) and XGBoost (XGB). Monthly R^2 values generally remain above 0.90 for TabPFN and exhibit substantially reduced seasonal variability relative to RF and XGB, which show more pronounced declines during mid-year periods. Correspondingly, TabPFN maintains RMSE values around ~0.4–0.5 $\log_{10}$ units, whereas RF and XGB typically range between ~0.6–0.8. Performance

patterns are consistent when evaluation is restricted to global land grid cells (Extended Data Fig. 3). Boxplot comparisons indicate that TabPFN also yield the highest R^2 and the lowest RMSE among the tested models. Overall, these results demonstrate that the small-sample foundation model architecture provides improved predictive accuracy and temporal stability for reconstructing global BaP concentrations, particularly under sparse observational constraints.

Fig. 2 summarizes the reconstructed global, high spatiotemporal–resolution concentrations of ambient benzo[a]pyrene (BaP) from 1950 to 2100. Globally, population-weighted median (P50) BaP concentrations increased from approximately 0.05 ng m^{-3} in 1950 to a peak of ~0.26–0.28 ng m^{-3} around 2010–2015, representing a more than fivefold increase over the second half of the 20th century (Fig. 2A), and the detailed data were provided in Table S5 of Supplementary Materials (SM). This rise was accompanied by substantial spatial heterogeneity, with the interquartile range (IQR) expanding from ~0.08 ng m^{-3} in 1950 to >0.45 ng m^{-3} at the global peak, indicating increasingly uneven exposure distributions. Following the peak, projected BaP concentrations diverge markedly across Shared Socioeconomic Pathways (SSP). By 2050, global median concentrations decline to approximately 0.05 ng m^{-3} under SSP-126, ~0.08 ng m^{-3} under SSP-245, ~0.11 ng m^{-3} under SSP-370, and ~0.13 ng m^{-3} under SSP-585, corresponding to reductions of 80%, 70%, 60%, and 50%, respectively, relative to peak levels. However, even under the most stringent scenario, the global IQR remains above 0.15 ng m^{-3}, demonstrating persistent spatial inequality despite declining averages. Spatial reconstructions reveal pronounced regional contrasts. During the historical peak period, BaP concentrations in parts of South and East Asia exceeded 0.8–1.0 ng m^{-3}, more than three times the contemporaneous global median, whereas most high-income regions declined to below 0.1 ng m^{-3} by 2025. Scenario projections shows that spatial distributions of BaP concentrations under SSP-245 for selected years (1950, 2000, 2025, and 2050) illustrate the shifting geography of combustion-related pollution (Fig. 2B-E). In 1950, elevated concentrations (>0.1 ng m^{-3}) were primarily confined to industrial regions in North America and Western Europe, with most of the Global South remaining below 0.03 ng m^{-3}. By 2000, high-exposure zones expanded across East and South Asia, where concentrations exceeded 0.3–0.5 ng m^{-3} in densely populated corridors. In 2025, peak values in parts of northern India and eastern China remained above 0.4 ng m^{-3}, while levels in Europe and North America declined to <0.1 ng m^{-3}. By 2050, projected concentrations fall below 0.1 ng m^{-3} across most regions under SSP-245, though localized hotspots (>0.15 ng m^{-3}) persist in rapidly urbanizing areas, indicating incomplete convergence of global exposure burdens. They reveal the geographic shift of high concentrations from early industrialized regions in North America and Europe toward rapidly developing regions in Asia, with persistent hotspots in densely populated and industrial corridors. Spatial distributions of BaP concentrations under other scenarios (SSP-126, SSP-370, and SSP-585) and selected years were provided in Extended Data Fig. 4. The data for Fig. 2 and Extended Data Fig. 4 were provided in Supplementary Data-1 (SD-1).

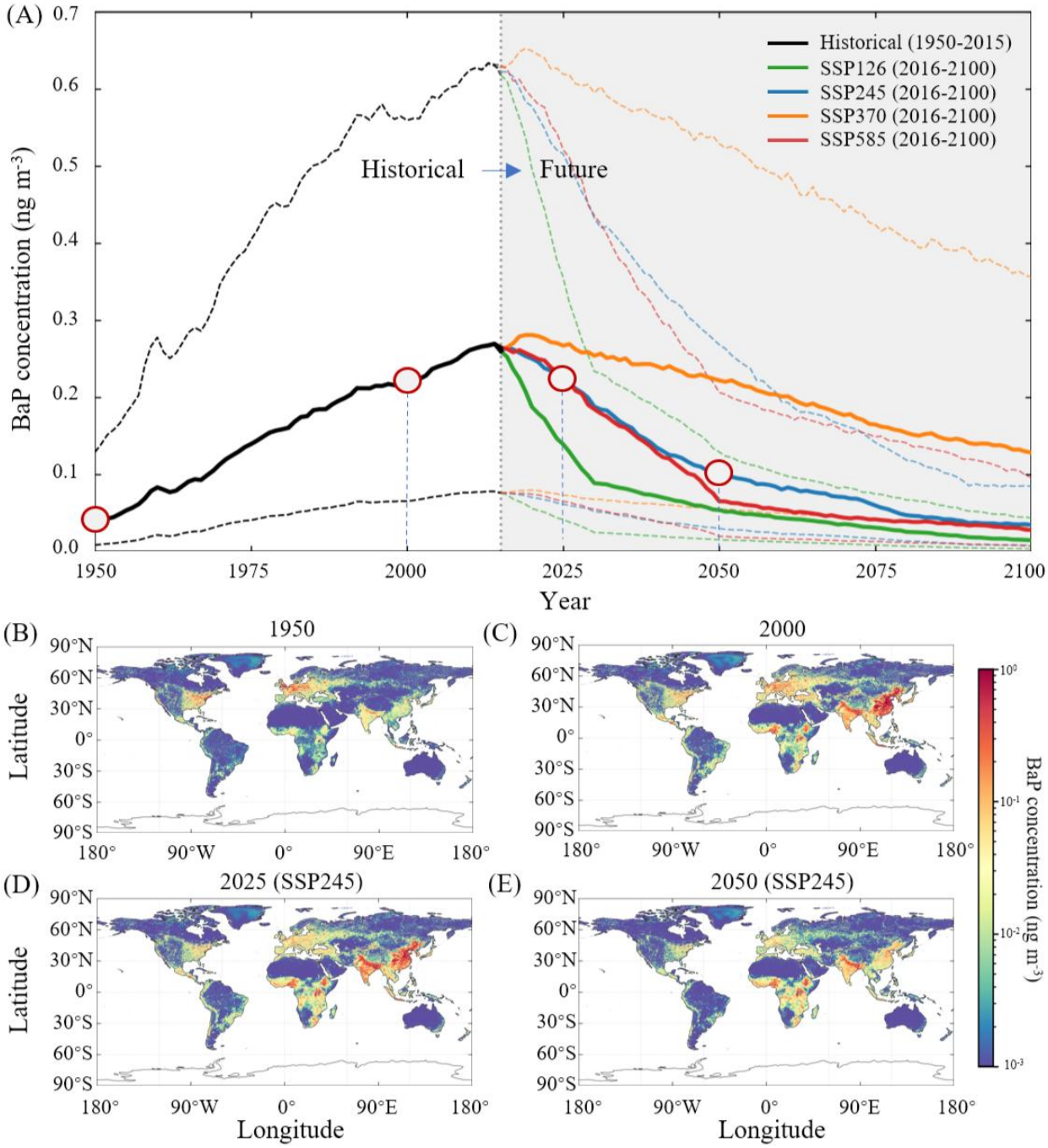


Fig. 2. Global reconstruction and future divergence of ambient benzo[a]pyrene (BaP) concentrations. (A) Global population-weighted median BaP concentrations from 1950 to 2100. The solid line indicates the median concentration, and the dashed lines denote the interquartile range (25th–75th percentile). Future projections are displayed in shade area under four Shared Socioeconomic Pathways (SSP-126, SSP-245, SSP-370, and SSP-585). (B–E) Spatial distributions of BaP concentrations under SSP-245 for selected years (1950, 2000, 2025, and 2050).

2.2 Projecting BaP-attributable LC burden

We selected SSP2-4.5 (SSP-245) as a representative middle-of-the-road scenario reflecting moderate mitigation and socioeconomic development, providing a realistic baseline for evaluating how historical exposure legacies translate into future LC burden. Fig. 3 quantifies the global burden of LC attributable to BaP exposure and demonstrates how accounting for exposure latency fundamentally alters risk estimation under SSP-245. Comparative summary of exposure assessment strategies (S0–S3), which represent progressively refined exposure assessment strategies, were provided in Table S4 (SM). S0 assumes steady-state lifetime exposure, S1 accumulates historical exposures equally, S2 incorporates cohort-specific exposure trajectories, and S3 further integrates latency-weighted accumulation, capturing historical dynamics, demographic structure, and carcinogenic latency in risk estimation. Under SSP-245, incorporating exposure history fundamentally reshapes the temporal profile of LC risk attributable to BaP. Globally (Fig. 3A), static exposure assumptions (S0) suggest an early peak in population attributable fraction (PAF) around 2015–2020, reaching approximately 7×10^{-4}, followed by a rapid decline. In contrast, the latency-aware lagged model (S3) delays the peak by nearly two decades, with maximum global PAF occurring around 2035–2045 at ~2.5×10^{-4}. Models considering only historical exposure intensity (S1) or age structure (S2) produce intermediate trajectories, but consistently misalign the timing of peak burden relative to exposure decline. These differences reflect the prolonged carcinogenic latency of BaP exposure and the demographic momentum embedded in aging populations. Country-level analyses further illustrate the interaction between exposure history and demographic transition (Fig. 3B). In China and India, rapid industrialization during the late twentieth century produces exposure peaks exceeding 1.5×10^{-3} under static assumptions, whereas the lagged model shifts maximal attributable burden into mid-century, despite declining concentrations. In contrast, the United States and United Kingdom exhibit earlier exposure peaks (1960–1980) and a smoother decline in PAF, reflecting earlier regulatory transitions. Brazil and South Africa show mixed trajectories, where demographic growth amplifies delayed burden despite moderate exposure intensity.

Spatially (Fig. 3C–F), the global distribution of BaP-attributable PAF evolves markedly over time. During 1950–2025, elevated PAF ($>5\times10^{-4}$) concentrates in East and South Asia and parts of Eastern Europe. By 2026–2050, risk intensifies in densely populated regions undergoing sustained combustion emissions, while high-income regions show continued decline. After 2050, global convergence emerges under SSP-245, yet residual hotspots persist in rapidly urbanizing corridors. These findings demonstrate that latency-aware modeling reveals a delayed and uneven inheritance of pollution-related cancer risk, in which historical exposure regimes continue to shape future disease burden long after ambient concentrations decline. The similar trends were also identified in the other three scenarios of SSP-126 (Fig. S1, SM), SSP-370 (Fig. S2, SM), and SSP-585 (Fig. S3, SM). The related global PAF of LC due to BaP exposure across four exposure assignment methods and four SSPs were provided in Table S6 (SM). All the data for Fig. 3 and Fig. S1-S3 were provided in SD-2.

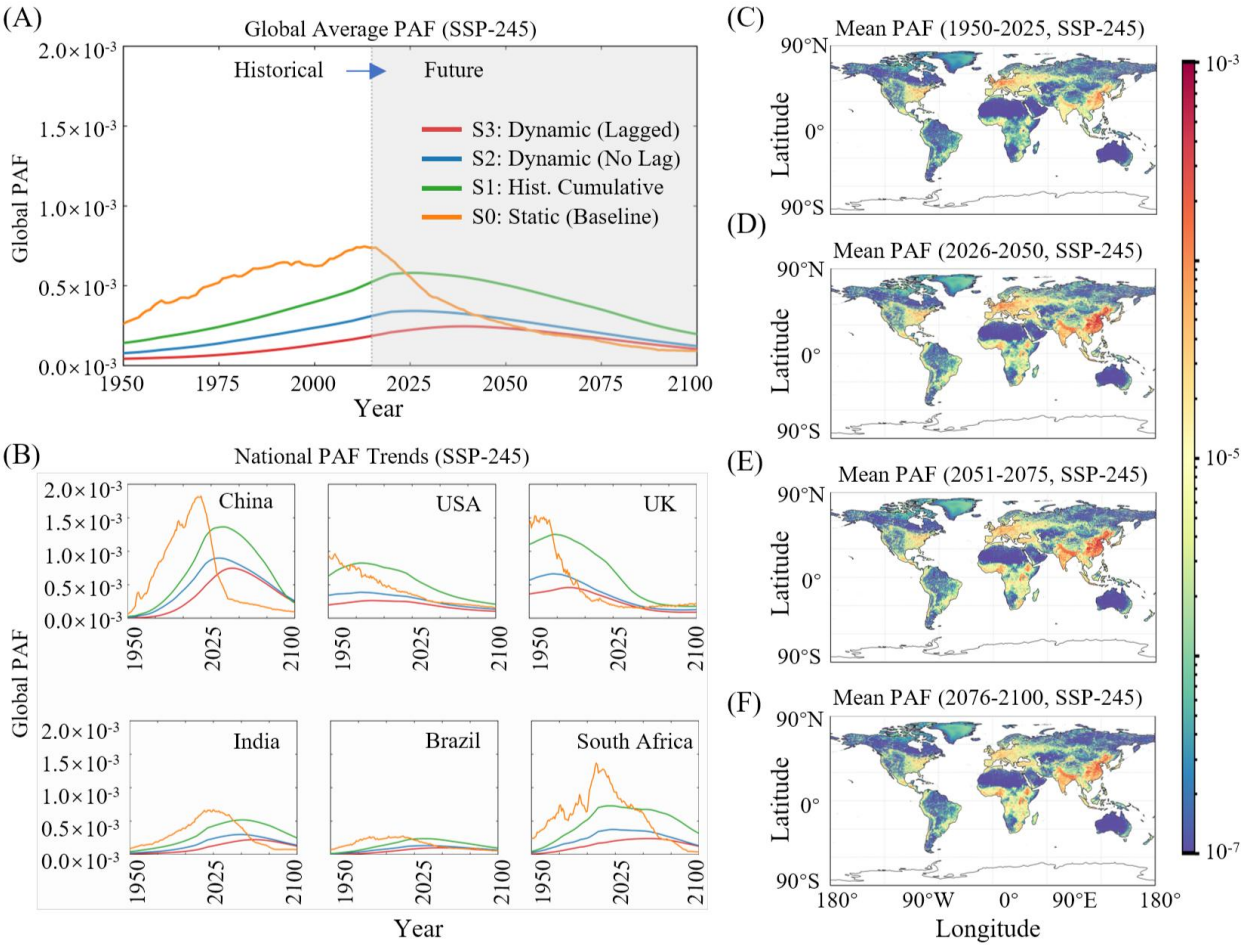


Fig. 3. Latency-aware reconstruction of global lung cancer risk attributable to benzo[a]pyrene (BaP) under SSP-245. (A) Temporal evolution of global BaP-attributable lung cancer PAF under SSP-245, estimated using four alternative exposure accumulation strategies (S0–S3). (B) Country-level comparisons for six representative nations (China, United States, United Kingdom, India, Brazil, and South Africa). Divergence across strategies highlights how historical exposure intensity and demographic transitions interact to shape national risk trajectories. In rapidly industrializing countries, latency-aware modeling reveals a pronounced delay between exposure peaks and maximum attributable burden. (C-F) Spatial distribution of mean BaP-attributable lung cancer PAF under SSP-245 averaged over four time-windows (1950–2025, 2026–2050, 2051–2075, and 2076–2100).

2.3 Spatial distribution of LC mortality attributable to ambient air pollution

Fig. 4 illustrates the extrapolation of LC burden attributable to ambient air pollution by leveraging the strong empirical relationship between BaP exposure and air pollution-related LC risk and a dual pattern of spatial disparity and intergenerational inequality in LC mortality attributable to ambient air pollution under SSP-245. Global distribution of LC deaths attributable to ambient air pollution from 1990 to 2023 were depicted in Fig. 4A and the data were provided in Table S7 (SM). The highest attributable mortality is concentrated in East and South Asia, driven by large populations, historically elevated pollution, and delayed effects of past exposures. Moderate burdens occur in parts of Eastern Europe, the Middle East, and Latin America, whereas most high-income regions, including North America and Western Europe, show comparatively lower

mortality. Three evidences were adopted to support extending BaP-based risk estimates to broader air pollution mixtures:

First, statistical coherence analysis demonstrates that BaP is strongly correlated with major combustion-related pollutants, including $PM_{2.5}$, black carbon, organic carbon, and fossil-fuel emission indicators, with absolute correlation coefficients frequently exceeding 0.7–0.8 across spatial and temporal scales (Fig. 4B, Extended Data Fig. 5, and SD-4). Correlation patterns remain stable across industrial combustion, power generation, and transportation sectors, indicating shared emission sources and co-variation in temporal trends. Even in non-combustion sectors, BaP maintains moderate concordance with particulate fractions. This sectoral robustness supports the use of BaP as a mechanistic marker for extrapolating LC risk to broader ambient air pollution mixtures.

Second, mechanistic network analyses using the E-BIO framework (EPPD and EGoD models) reveal dense overlap between BaP-associated proteins, enriched biological pathways, and those linked to broader air pollution components (Fig. 4C, Extended Data Fig. 6-Fig. 7, and Table S8 (SM)). In the EPPD network, BaP shares high-connectivity intermediates, including TP53 and oxidative stress–related proteins, with polycyclic aromatic hydrocarbons, nitrogen oxides, and particulate components. Similarly, the EGoD network demonstrates convergence in gene ontology terms related to DNA damage response, inflammatory signaling, and cell-cycle regulation. The high degree of shared nodes and pathway overlap supports the biological plausibility of extending BaP-based risk estimates to broader ambient air pollution mixtures.

Third, the latency- and demography-aware strategy (S3) integrates historically reconstructed exposures with population age structure and disease latency. The latency weighting function for LC risk following exposure were shown in Extended Data Fig. 8. BaP-attributable PAFs are highly correlated with LC mortality attributable to ambient air pollution (Fig. 4D, Extended Data Fig. 9, and Table S9 (SM)). The near-linear relationship across several orders of magnitude indicates that BaP-based PAF×population estimates capture both the magnitude and ranking of national burdens. Larger-population countries cluster toward the upper-right quadrant, while smaller countries exhibit consistent scaling along the 1:1 reference line. This concordance provides empirical support for extending BaP-anchored risk estimation to broader ambient air pollution mixtures.

Using this relationship, we derive high-resolution global distributions of air pollution–attributable LC mortality for 1990–2023, revealing pronounced spatial heterogeneity (Fig. 4E). At the country level (Fig. 4A), mean annual attributable deaths during 1990–2023 exceed 100,000 in China and 50,000 in India, whereas most high-income countries remain below 10,000 annually, highlighting strong cross-national concentration. However, high-resolution reconstruction demonstrates that within-country heterogeneity is substantial: densely populated industrial corridors in northern India and eastern China exceed 10 deaths per 0.1° grid cell annually, while large rural areas remain below 0.1 deaths per grid (Fig. 4E). Regions in South and East Asia exhibit the highest burdens, whereas most high-income regions show substantially lower attributable mortality. All the data for Fig. 4E were provided in SD-3.

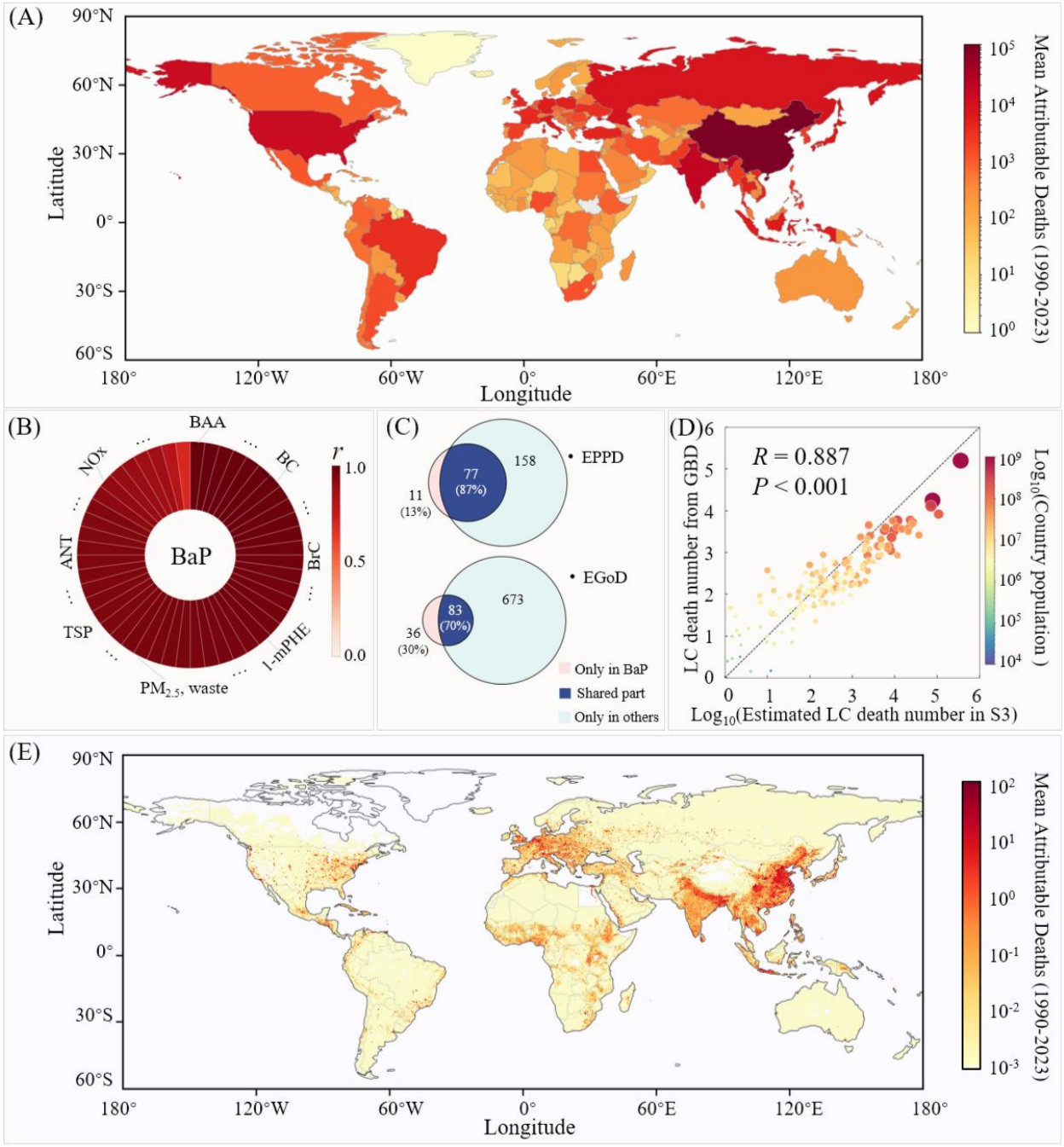


Fig. 4. Spatial reconstruction and mechanistic extrapolation of lung cancer mortality attributable to ambient air pollution under SSP245.

(A) Country-level mean annual lung cancer deaths attributable to ambient air pollution during 1990–2023 based on GBD estimates. (B) Correlation between BaP and major air pollution components. (C) Overlap of mechanistic evidence linking BaP and broader pollutant mixtures using the Exposure–Protein–Protein–Disease (EPPD) and Exposure–Gene Ontology–Disease (EGoD) mode. (D) Country-level agreement between latency-aware BaP-based estimates (S3) and GBD-reported lung cancer deaths. (E) High-resolution global distribution of attributable lung cancer mortality reconstructed using the BaP-anchored framework.

2.4 Intergenerational inequality of air pollution–attributable LC burden

To illustrate spatial heterogeneity across multiple geographic scales, regional analyses were conducted following the hierarchical structure of the GBD framework, including 204 countries and territories nested within 21 global regions. Also, three representative countries of China Mainland, the United States, and India with large population sizes and contrasting industrialization histories were selected to provide informative contrasts in attributable risk. Intergenerational inequality is evident across both global and subnational scales (Fig. 5). Across all countries, lifetime LC PAFs are consistently highest among cohorts born between 1950 and 1990, corresponding to periods of rapid industrialization and rising BaP exposure. In contrast, cohorts born after 2000 show monotonic declines in lifetime attributable risk, particularly in high-income regions (Fig. 5A). Spatial gradients remain pronounced. Eastern Europe and East Asia exhibit persistently elevated cohort risks relative to Oceania and parts of Sub-Saharan Africa (Fig. 5B). Subnational disparities further amplify these patterns: in China mainland, provinces such as Hebei, Henan, and Tianjin show substantially higher cohort risks than western provinces; in the United States, industrial and densely populated states display higher lifetime PAFs than less urbanized regions; and in India, northern and highly populated states bear disproportionately greater cohort burdens (Fig. 5C). These findings demonstrate that declining global averages mask structured intergenerational inheritance of pollution-related LC risk, with both temporal and spatial gradients reinforcing cumulative inequality. The other SSPs were provided in Fig. S4-S6, and all the related data were provided in Table S10.

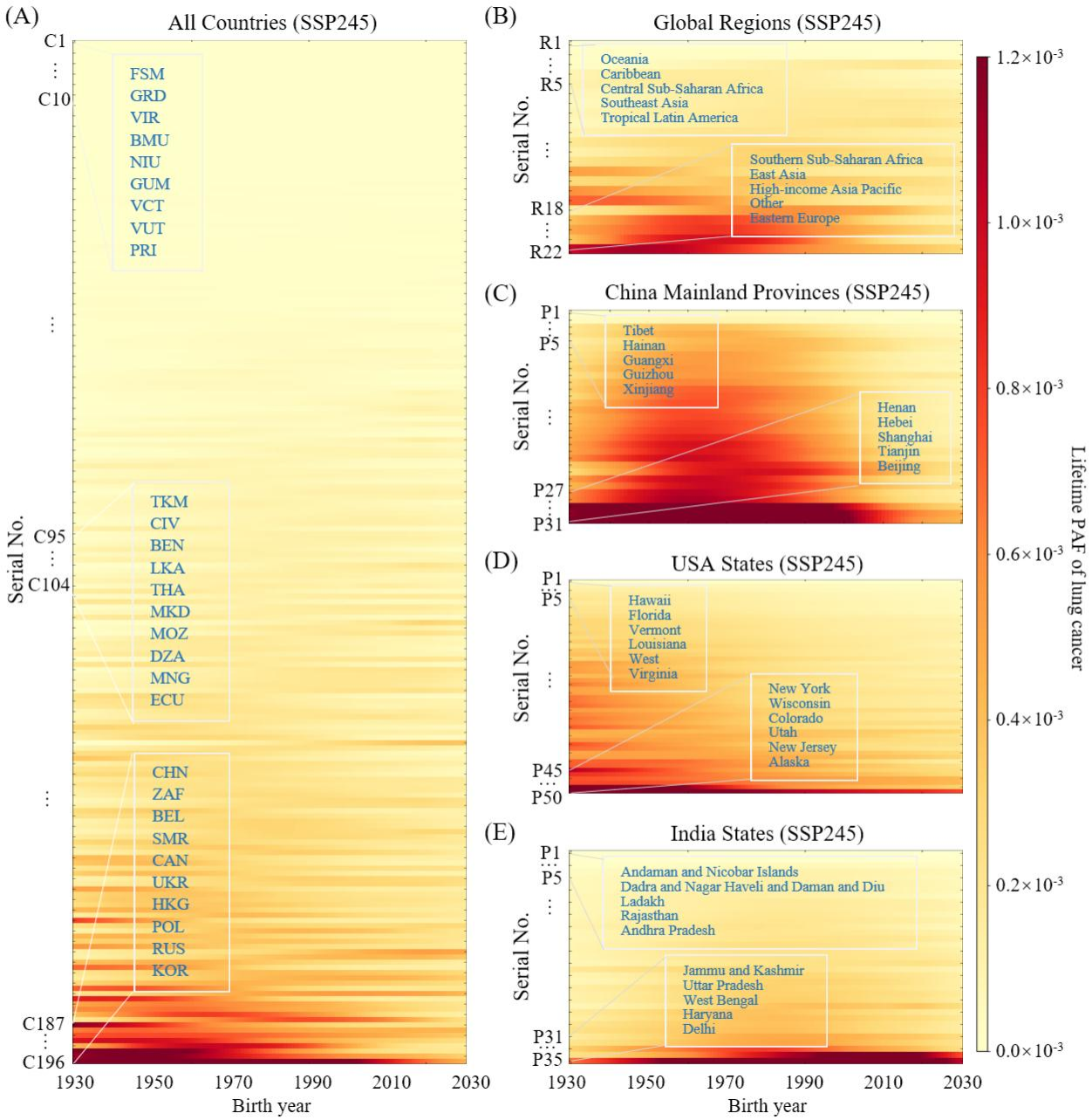


Fig. 5. Intergenerational inequality in lifetime lung cancer risk attributable to ambient air pollution under SSP2–4.5.

(A) Lifetime population attributable fractions (PAFs) of lung cancer across all countries, ordered by cumulative risk. (B–E) Subnational and regional patterns for global regions, China provinces, U.S. states, and India states. Each horizontal band represents a birth cohort (1930–2030), with color intensity indicating lifetime PAF. Cohorts born during periods of rapid industrial expansion exhibit systematically higher lifetime risk, while later cohorts show declining risk following regulatory transitions. Persistent spatial gradients indicate enduring interregional and subnational inequality despite projected declines in average exposure. All the names were provided in Table S7.

3. Discussion

Our study reframes air pollution-attributable LC as a delayed and structurally embedded burden rather than a contemporaneous response to current exposure levels. By integrating century-scale exposure reconstruction, latency-aware risk attribution, and mechanistic anchoring using BaP, we show that the global LC burden attributable to ambient air pollution is shaped by long-term exposure trajectories, demographic transitions, and historically uneven development pathways. This integrated framework reveals why substantial inequalities persist even as average air quality improves: current disease burdens largely reflect the legacy of past emissions and their uneven distribution across birth cohorts. By combining small-sample AI-based exposure reconstruction with explicit quantification of spatial disparities and intergenerational inequalities, our results provide a quantitative explanation for the persistent global imbalance in air pollution-related LC risk.

A central methodological advance of this work is the reconstruction of global BaP concentrations at high spatiotemporal resolution over an unprecedented temporal span (1850–2100). Previous global assessments of air pollution-related cancer risk have largely relied on short observational windows or steady-state assumptions, implicitly treating exposure as temporally static or spatially homogeneous[10,13,18,20,22]. Such approaches are ill-suited to capture the delayed effects of historical emissions, particularly in regions that experienced rapid industrialization followed by delayed or uneven implementation of pollution controls. By leveraging a small-sample foundation AI modeling strategy[23], our framework enables consistent global inference under severe data scarcity, revealing coherent long-term exposure patterns that are difficult to resolve using conventional approaches, like the RF and XGB models.

Incorporating disease latency fundamentally alters how air pollution-related LC risk is attributed. LC is characterized by long induction and latency periods, yet many existing population-level risk assessments implicitly assume that exposure levels in a given year are representative of lifetime exposure. This steady-state assumption can substantially distort risk estimates, particularly in periods of rapid exposure change[10-13]. By explicitly modeling latency and integrating historical exposure trajectories with population age structure, our approach links experienced exposure histories to disease outcomes without imposing unrealistic assumptions about lifetime exposure constancy[8]. The resulting estimates are more conservative and epidemiologically plausible, avoiding the implausibly high PAFs that can arise under steady-state formulations. More importantly, this latency-aware perspective highlights a key conceptual insight that LC attributable to air pollution is a delayed manifestation of past exposure regimes interacting with demographic transitions. Recent evidence indicates that improvements in air quality do not automatically translate into proportional declines in pollution-attributable mortality because demographic shifts, particularly rapid population aging, can offset exposure reductions[24]. Our study can precisely explain the temporo-spatial trend of LC disease burdern of most countries/region in the word attributable to air pollution[6], which can't be accomplished by the previous studies due to lacking of long-term, high-resolution exposure data to air pollutants.

Building on this latency-aware framework, we extend BaP-based risk attribution to ambient air pollution mixtures by treating BaP as a mechanistic marker of carcinogenicity. There is clear and substantial evidence of a link between outdoor ambient air pollution[16], and particularly $PM_{2.5}$[19,25]. However, air pollution is a complex mixture of gases and particles. Across cohorts, dramatic heterogeneity of the effect value (e.g., relative risk, hazard ratio) was identified and its relationship with the $PM_{2.5}$ concentration is difficult to build [19]. In addition, current GBD estimates quantify air pollution-attributable LC primarily at national resolution[6], limiting insight into

subnational heterogeneity. Such coarse aggregation constrains precision governance and obscures localized hotspots. Relatively, Armstrong et al. established a log-linear exposure-response relationship between cumulative BaP exposure and LC, providing a more explicitly quantified risk function than typically available for $PM_{2.5}$[17]. Our results demonstrate that BaP-based PAFs are strongly aligned with air pollution-attributable LC mortality across countries, reflecting shared exposure patterns, overlapping carcinogenic pathways, and consistent epidemiological relationships. This convergence across concentration data, biological mechanisms, and population-level outcomes provides a robust basis for extrapolating LC risk from BaP to broader air pollution mixtures. Our high-resolution reconstruction addresses this gap, enabling spatially explicit risk assessment to inform targeted intervention and equitable resource allocation.

The global decline in average air pollution masks a more complex moral landscape. The high-resolution global burden of LC revealed that LC attributable to ambient air pollution is structured along inequality: the unequal inheritance of risk across generations. These patterns are not random fluctuations of exposure but reflect the cumulative imprint of industrialization, energy transitions, and delayed regulatory action[4,14,26,27]. Intergenerational inequality introduces a temporal dimension of justice. Exposure histories become biologically embedded, transforming past emissions into future disease. In this sense, air pollution produced delayed health effects, in which historical benefits of industrial growth are decoupled from the long-term health costs borne by subsequent generations. Recent evidence shows that air quality policies that reduce average exposure may fail to eliminate, and even exacerbate, relative disparities across socioeconomic groups[5]. Our findings clearly demonstrated that without explicitly quantifying spatial and intergenerational inequality, mitigation efforts risk perpetuating structurally uneven LC burdens. Scenario-based retrospection across countries (e.g., China, USA, and India) highlights the long-term health dividends of early emission control. In countries where peak combustion emissions occurred earlier and regulatory interventions were implemented promptly, LC burden declines sooner and more steadily. In contrast, delayed mitigation shifts peak attributable mortality decades into the future because of disease latency. These findings underscore a critical policy insight: ***Today's emission reductions translate into tomorrow's health gains, but with substantial temporal lag*****.** For developing countries where emissions have not yet peaked, early intervention offers an important opportunity to avoid locking future generations into preventable cancer risk. It has been proved that national climate action can reshape international air pollution inequalities, yet historical emissions and demographic momentum sustain uneven cancer burdens[28]. As for China, it was proposed that even substantial reductions in air pollution do not guarantee immediate declines in attributable mortality, as demographic aging and disease latency can offset exposure gains[29]. Our BaP-based framework shows that LC reflects structurally embedded, transboundary, and intergenerational legacies, requiring coordinated mitigation that addresses both historical responsibility and future vulnerability.

The implications of these findings extend beyond methodological refinement to questions of environmental governance and justice. Policies aimed solely at reducing current pollution levels may substantially underestimate the ongoing and future health burdens rooted in historical exposures[4]. Equity-oriented interventions must therefore account for the delayed consequences of past pollution, particularly in regions and populations that experienced rapid industrialization without commensurate health protections. From a global perspective, the persistence of inequalities underscores the need for international cooperation and responsibility-sharing, as the health consequences of historical emissions are unevenly distributed across regions and generations. More broadly, this study illustrates how integrating foundation modeling, causal time

structures, and inequality metrics can transform environmental health risk assessment. Rather than treating exposure, risk, and burden as static quantities, our framework emphasizes their dynamic and path-dependent nature. This approach is readily extendable to other pollution-related diseases and environmental stressors characterized by long latency and uneven exposure histories. As global environmental change accelerates, such dynamic and equity-focused frameworks will be increasingly necessary to inform policy decisions that are both scientifically robust and socially just.

Several limitations warrant consideration. Treating BaP as a representative marker of air pollution-related carcinogenicity inevitably simplifies the complexity of pollutant mixtures[16]. Future work incorporating additional components and mixture-specific toxicological data could further refine risk estimates. In addition, while our framework captures long-term exposure trajectories and demographic dynamics, uncertainties remain in historical exposure reconstruction, particularly in regions with sparse monitoring data, as well as those greatly affected by the other emissions, e.g., wild fire, industrial activities [27]. Nevertheless, the consistency of patterns across multiple lines of evidence suggests that the central conclusions regarding disparity and inequality are robust. By linking long-term exposure reconstruction with population dynamics and structural inequalities, it demonstrates that air pollution-related LC is not merely a declining environmental risk, but a persistent and uneven legacy of historical environmental change. Recognizing and quantifying this legacy is essential for developing policies that address not only pollution control, but also the enduring health inequalities it has produced. Beyond environmental governance, these results carry important implications for public health planning and risk management systems. In particular, the delayed and cohort-specific nature of pollution-related cancer risk suggests that future healthcare demand and insurance liabilities may be strongly influenced by historical exposure patterns rather than current pollution levels alone. Therefore, integrating exposure history into health risk assessment could improve long-term forecasting of disease burden, guide targeted prevention strategies, and inform more equitable allocation of healthcare and insurance resources in aging populations.

4. Materials and Method

4.1 Reconstruction of long-term global BaP concentrations (1850–2100)

4.1.1 Model Construction

The data sources to build the BaP prediction model were provided in the section "1. Data sources" of Supplementary Sections (SM), including BaP concentration, emission densities of various air pollutants, meteorological drivers, and the selection of Shared Socioeconomic Pathways (SSPs). To evaluate and improve global benzo[a]pyrene (BaP) concentration prediction, we developed a hierarchical modeling framework and conducted systematic benchmarking across multiple algorithms. All models were trained and evaluated using an identical temporal split to ensure comparability and prevent information leakage. Specifically, monthly BaP concentration fields (1°×1° resolution) from 1998–2005 (first 96 months) were used exclusively for model fitting and cross-validation, while data from 2006–2007 (final 24 months) were reserved as an independent test set for one-time prediction. Three modeling approaches were implemented: two conventional ensemble machine-learning models and one transformer-based foundation model. Among them, Random Forest and Extreme Gradient Boosting (XGBoost) were selected as strong nonlinear baselines due to their capacity to capture complex feature interactions. There parameters were set as below:

(i) XGBoost. We implemented the XGBRegressor with the following hyperparameters:

n_estimators = 300, learning_rate = 0.05, max_depth = 8, subsample = 0.8, colsample_bytree = 0.8, tree_method = "hist", random_state = 42. Five-fold time-series cross-validation (TimeSeriesSplit, n_splits = 5) was applied to the 96-month training period. Predictions from each fold were averaged to construct the final model, which was then evaluated on the 24-month independent test set.

(ii) Random Forest. The RandomForestRegressor was implemented with hyperparameters: n_estimators = 1000, max_features = 0.8, max_samples = 0.8, min_samples_split = 5, min_samples_leaf = 2, random_state = 42, n_jobs = −1. The same five-fold time-series cross-validation strategy was applied. Fold-wise predictions were averaged to form the final estimator before evaluation on the independent test set.

(iii) TabPFN Foundation Model. TabPFN (Tabular Prior-Fitted Network) is a transformer-based foundation model designed for small-sample supervised learning [23]. By leveraging pre-trained priors learned from synthetic tasks, the model enables efficient transfer learning without conventional gradient-based retraining. Pre-trained weights ("*tabpfn-v2-regressor.ckpt*") were obtained from the official repository (https://github.com/automl/TabPFN). We implemented the TabPFNRegressor using the pretrained checkpoint without additional hyperparameter tuning. To accommodate global-scale gridded data, we adopted a spatial partitioning strategy. The global 180×360 grid was divided into 900 subgrids (30 latitudinal×30 longitudinal partitions), each containing 6×12 = 72 grid cells. For each subgrid, the first 96 months of data (72 cells×96 months = 6,912 samples) were directly provided to the TabPFN model via a fit operation, enabling adaptation to local feature distributions without traditional parameter optimization or cross-validation. The trained model for each subgrid then generated predictions for the subsequent 24 months (72 cells×24 months = 1,728 samples). Predicted subgrid outputs were spatially reassembled to reconstruct the global BaP field, thereby enabling external temporal validation.

4.1.2 Performance Comparison

Model evaluation during the independent test period (2006–2007) shows strong predictive performance of the BaP reconstruction framework across both global and land domains (Extended Data Fig. 2 and Extended Data Fig. 3). The TabPFN model consistently achieves higher coefficients of determination (R^2) and lower root mean square errors (RMSE) than conventional ensemble models. Across months, TabPFN maintains R^2 values generally above 0.9 with reduced seasonal variability, while RF and XGBoost exhibit larger mid-year declines. Similar patterns are observed over global land grid cells, where TabPFN shows the highest median R^2 and lowest RMSE. Density scatterplots further indicate tighter clustering of predicted versus observed $\log_{10}$(BaP) concentrations along the 1:1 line. These results demonstrate that the small-sample foundation model improves predictive accuracy and temporal stability for global BaP reconstruction.

4.1.3 Model Projection and High-Resolution Reconstruction

Based on benchmarking results, TabPFN demonstrated the highest predictive accuracy and temporal stability during the independent test period (2006–2007) and was therefore selected as the primary reconstruction model for long-term BaP estimation. The model was applied to generate monthly global BaP concentrations for 1850–2100. Predictor variables included sector-resolved emission fluxes (8-sector BC, 8-sector OC, and 8-sector NOx; 24 emission channels in total) and five meteorological variables (tas, huss, psl, sfcWind, and bldep). All inputs were harmonized to a 1°×1° spatial resolution. The global 180×360 grid was divided into 900 independent subgrids (30 latitudinal×30 longitudinal partitions), and each subgrid generated its own prediction sequence. Subgrid outputs were subsequently mosaicked to reconstruct complete global fields.

For the historical period (1850–2014), reconstructions were driven by CEDS historical emissions and CMIP6 historical meteorology (Table S3). For the future period (2015–2100), projections were based on ScenarioMIP emission extensions under the SSP pathways combined with CMIP6 multi-model ensemble meteorology. To restore fine-scale spatial heterogeneity, we applied the spatial weight mask constructed during preprocessing. This mask, derived from the 1998–2007 time-averaged 0.1° concentration field, represents the relative contribution of each 0.1° pixel to its corresponding 1° grid-cell mean. Predicted 1°×1° concentration fields were disaggregated to 0.1°×0.1° resolution through pixel-level weighting, ensuring mass conservation while preserving empirically constrained subgrid spatial patterns.

The final product consists of a continuous global annual mean BaP concentration dataset at 0.1°×0.1° resolution for 1850–2100, providing a high-resolution exposure foundation for subsequent population-level assessment and health risk quantification.

4.2. Population Data and Age Structure

4.2.1 Population Attributable Fractions

Gridded global population data were obtained from a publicly available Zenodo dataset (https://zenodo.org/records/10968817) providing historical reconstructions (1870–2010) and future projections (2020–2100) under various SSPs. Age-structured population data were obtained from the United Nations World Population Prospects (UN WPP) standard projections (https://population.un.org/wpp/). The detailed process methods were provided in the section "2. Gridded Global Population" (SM). We quantified the lung cancer burden attributable to long-term BaP exposure using the population attributable fraction (PAF) framework. For each 0.1°×0.1° grid cell, relative risk (*RR*) was estimated using a power-law exposure–response function based on cumulative exposure:

$$RR = URR^{\left(\frac{CE}{100}\right)}$$

$$PAF = 1 - \frac{1}{RR} = 1 - URR^{-\left(\frac{CE}{100}\right)}$$

where *URR* denotes the unit relative risk. Based on epidemiological cohort evidence from Xuanwei, China, we set URR to 4.49 for a lifetime (70-year) cumulative exposure of 100 μg m^{-3}·year, representing the multiplicative increase in lung cancer risk relative to a clean-air baseline [10]; The constant 100 (μg m^{-3}·year) serves as the reference exposure level used to normalize cumulative exposure; *CE* denotes cumulative exposure (μg m^{-3}·year), defined as the total inhaled exposure accrued over a specified time window; *PAF* ranges in 0–1, quantifying the proportion of lung cancer cases within each grid cell attributable to BaP inhalation. Accurate estimation of CE is central to robust PAF quantification. To address exposure timing and latency effects, we implemented four progressively refined exposure assessment strategies (S0–S3).

4.2.3 Evaluation of PAFs Under Alternative Exposure Strategies

(1) S0: Steady-State Lifetime Exposure

The S0 strategy represents the conventional steady-state assumption frequently used in burden-of-disease assessments, which can be considered as the "conventional baseline". Under this approach, the exposure level observed in a given year is assumed to remain constant throughout an individual's entire lifetime (typically 70 years). Lung cancer risk is therefore calculated by treating the current annual BaP concentration as if it had applied uniformly across all prior decades. This steady-state framework ignores temporal variability in historical exposure trajectories and does not explicitly account for latency between exposure and disease onset. Conceptually, it equates

contemporary pollution levels with lifetime exposure history. Because of its simplicity and minimal data requirements, S0 has been widely adopted in large-scale assessments. However, during periods of declining pollution, this assumption may substantially overestimate the PAF, as it implicitly attributes past higher exposures to the present population regardless of actual exposure timing. The complete computational formulation is provided below.

$$CE_{S0} = C_y \times 70$$

$$PAF_{S0} = 1 - URR^{-\left(\frac{CE_{S0}}{100}\right)}$$

where CE_{S0} denotes the lifetime cumulative exposure assigned under the steady-state assumption; C_y represents the annual mean BaP concentration (μg m^{-3}) in the target evaluation year *y* (i.e., the current year) for the given grid cell. The constant 70 corresponds to the conventional epidemiological assumption of a standard lifetime duration (years); PAF_{S0} denotes the grid-level PAF estimated under the S0 steady-state exposure strategy.

(2) S1: Simple Accumulation Over the Previous 70 Years

The S1 strategy estimates lung cancer risk using cumulative exposure accrued over the 70-year period preceding the evaluation year, thereby incorporating temporal variability in historical BaP concentrations. In contrast to the steady-state assumption (S0), S1 explicitly sums year-to-year exposure changes across the window from birth to the evaluation year. However, S1 assigns equal weight to all exposures within the 70-year window and does not account for carcinogenic latency or population age structure. Implicitly, it treats the population as if all individuals experience the same lifetime exposure history equivalent to that of a 70-year-old cohort, and assumes that early-life and recent exposures contribute identically to cancer risk. This simplification neglects demographic heterogeneity and compresses complex latency-dependent carcinogenic processes into an unweighted cumulative metric. The complete computational formulation is provided below.

$$CE_{S1} = \sum_{i=0}^{69} C_{y-i}$$

$$PAF_{S1} = 1 - URR^{-\left(\frac{CE_{S1}}{100}\right)}$$

where CE_{S1} denotes the cumulative historical exposure assigned under the S1 strategy for the population within a given grid cell; C_y represents the annual mean BaP concentration in the grid cell *i* years prior to the target evaluation year y; PAF_{S1} denotes the grid-level PAF estimated under the S1 cumulative exposure framework.

(3) S2: Age-Structured Accumulation Along Realized Exposure Histories

The S2 strategy incorporates population age structure by accumulating exposure along the realized historical trajectory experienced by each age group, from birth to the evaluation year. This approach corrects the key limitation of S1, which implicitly assumes that all individuals have fully experienced the same 70-year exposure window. By weighting exposure histories according to the observed age distribution, S2 produces a demographic-consistent estimate of cumulative exposure at the grid level. However, S2 still applies an equal-weight accumulation across the exposure window and does not explicitly represent latency-induced delays between exposure and lung cancer occurrence. As a result, while S2 improves demographic realism, it remains limited in capturing the temporal alignment between exposure timing and cancer risk manifestation. The complete computational formulation is provided below.

$$CE_{S2,a} = \sum_{t=y-a+1}^{y} C_t$$

$$PAF_{S2} = \sum_{a=1}^{100} \left[\left(1 - URR^{-\left(\frac{CE_{S2,a}}{100}\right)} \right) \times p_a \right]$$

where *a* denotes the age-group index, ranging from 1 to 100 years; "y-a+1" represents the birth year corresponding to individuals aged *a* in evaluation year y. For example, individuals aged 30 in 2025 are assumed to have begun exposure accumulation in 1996; *t* denotes the absolute calendar year of exposure, spanning from the birth year to the evaluation year y; C_t represents the annual mean BaP concentration in the grid cell during year *t*; $CE_{S2,a}$ denotes the lifetime cumulative exposure experienced by individuals of age a based on their realized historical exposure trajectory; p_a denotes the population weight (proportion) of age group *a* in the target year *y* within the corresponding country, derived from the standard projections by United Nations World Population Prospects 2024 (UN-WPP) (https://population.un.org/wpp/). By definition, the sum of p_a across all age groups equals 1; PAF_{S2} denotes the grid-level *PAF* under the S2 framework after age-structure standardization.

(4) S3: Age-Structured Exposure with Latency-Weighted Risk

The S3 strategy constitutes the "Core Strategy" of this study. It integrates age-structured exposure histories with latency-weighted risk assignment using a Weibull distribution to approximate the temporal dynamics of lung cancer development. Under this approach, exposure contributions are weighted according to the time elapsed since exposure occurrence, allowing risk to vary as a function of biologically plausible latency. Exposures occurring within etiologically relevant windows (e.g., approximately 10-20 years prior to diagnosis) receive greater weight than very recent or remote exposures. By combining reconstructed historical BaP trajectories with observed cohort-specific population distributions, S3 avoids the steady-state assumption and equal-weight accumulation inherent in S0–S2. This framework provides a demographically and biologically grounded estimate of attributable lung cancer risk, aligning exposure timing with the known latency structure of carcinogenesis.

(i) Latency Probability Distribution and Derivation of the Weighting Scheme

To quantify how exposures occurring at different historical time points contribute differentially to current disease risk, we first defined a probability density function (PDF) for the lung cancer latency duration in the population:

$$f(lag; k, \lambda) = \frac{k}{\lambda}\left(\frac{lag}{\lambda}\right)^{k-1} \exp\left[-\left(\frac{lag}{\lambda}\right)^{k}\right]$$

where *f*(lag) denotes the latency probability density, the probability density that the lung cancer latency duration in the population equals lag years At the population level, whether exposure occurring in year *t* contributes to disease incidence in evaluation year *y* depends on whether the biological latency duration is less than or equal to the elapsed physical time lag (lag = *y*-*t*). Accordingly, we adopt the integral form of the PDF, i.e., the cumulative distribution function (CDF), as the effectiveness weight assigned to exposure occurring in year *t*. The annual exposure weight w_{y-t} is therefore defined as the cumulative probability that the latency duration does not exceed the elapsed time since exposure:

$$w_{y-t} = \int_0^{y-t} f(t')dt' = 1 - \exp\left[-\left(\frac{y-t}{\lambda}\right)^{k}\right]$$

where w_{y-t} represents the cumulative probability that the latency duration has been completed after $y-t$ years since exposure, which can also be interpreted as the proportion of exposure assumed to be associated to the carcinogenic effect by evaluation year y. For recent exposures (small *y*-*t*, e.g., within 5 years), most exposed individuals remain within the biological latency phase, and clinical

manifestation has not yet occurred; accordingly, w_{y-t} approaches 0, indicating that only a very small proportion of exposure risk was caused by BaP exposure. For more remote exposures (large *y*-*t*, e.g., ≥ 20 years), the majority of exposed individuals have completed the latency process, and the carcinogenic potential has been largely realized; in this case, w_{y-t} approaches 1.

(ii) Parameter Selection and Standardization

Parameterization of the latency distribution followed internationally recognized standards for air pollution health impact assessment, including the European Environment Agency report (ETC/ATNI 04/2020) [30] and the previous study [9]. The shape parameter (k) of the Weibull distribution was specified as 2.0. The scale parameter (λ) was derived from the expectation formula:

$$MLD = \lambda\gamma(1+1/k)$$

such that the distribution yields a mean latency duration (MLD) of 13.6 years, resulting in $\lambda \approx 15.35$ (Extended Data Fig. 8). This parameterization produces an inverted U-shaped PDF, in which risk density initially increases with time since exposure, reaches a peak (approximately at lag ≈ 10 years, with a maximum density near 0.06), and subsequently declines. The corresponding probability density function and cumulative distribution function used to characterize lung cancer latency are illustrated below.

(iii) S3 Strategy: Full Computation

Building on the latency-weighting framework described above, we first computed the latency-adjusted "effective cumulative exposure" for each age group. These age-specific estimates were then aggregated using the population age-structure proportions to derive the grid-level PAF under the S3 strategy.

$$CE_{S3,a} = \sum_{t=y-a+1}^{y} (C_t \times w_{y-t})$$

$$PAF_{S3} = \sum_{a=1}^{100} \left[\left(1 - URR^{-\left(\frac{CE_{S3,a}}{100}\right)} \right) \times p_a \right]$$

where w_{y-t} denotes the time-lag weight, representing the carcinogenic effectiveness of exposure occurring in historical year *t* as manifested in evaluation year *y*; Its value ranges from 0 to 1; *y*-*t* denotes the exposure lag duration, defined as the elapsed time between the historical exposure year *t* and the current evaluation year *y*; CE_{S3} denotes the latency-adjusted effective cumulative exposure for age group *a*, defined as the total exposure dose accumulated by individuals aged a that contributes meaningfully to clinical carcinogenic risk in the evaluation year; p_a denotes the population age proportion, representing the share of individuals aged a within the total population of the corresponding country for the given grid cell in the evaluation year. This term enables the transformation from cohort-level risk to population-level attributable fraction. PAF_{S3} denotes the final Population Attributable Fraction estimated under the S3 latency-weighted framework.

(5) Methodological Summary of S0–S3

Table S4 summarizes the conceptual progression from S0 to S3, illustrating increasing epidemiological realism and methodological complexity. The steady-state assumption (S0) treats current exposure as constant over a lifetime, while S1 incorporates historical variability but assumes equal weighting across the exposure window. S2 further improves demographic realism by aligning cumulative exposure with age-specific birth cohorts, yet still applies uniform weighting across time. In contrast, S3 integrates both age structure and biologically informed latency weighting, allowing exposure contributions to evolve dynamically according to a Weibull distribution. Although computationally more demanding and dependent on accurate historical data,

S3 provides the most mechanistically grounded and epidemiologically consistent framework for estimating population attributable fractions.

(6) Spatial Aggregation of PAF at the National and Regional Levels

To upscale risk estimates from the grid level to the national or regional level, simple arithmetic averaging of grid-level PAF values is inappropriate. Instead, spatial aggregation was performed using lung cancer baseline case counts as weights to ensure epidemiological consistency. For a given country or region R, the overall Population Attributable Fraction (PAF_R) was calculated as a weighted aggregation of grid-level PAF values, where weights correspond to the baseline lung cancer burden within each grid cell.

$$PAF_R = \frac{\sum_{i \in R}(PAF_i \times I_R \times Pop_i)}{\sum_{i \in R}(I_R \times Pop_i)} = \frac{\sum_{i \in R}(PAF_i \times Pop_i)}{\sum_{i \in R} P\,op_i}$$

where PAF_i denotes the grid-level attributable fraction calculated under a given strategy (S0–S3) for grid cell i within region R. Pop_i represents the total population of grid cell i in the evaluation year; I_R denotes the baseline lung cancer incidence rate for country or region, assumed to be spatially homogeneous within the same national boundary. The numerator of the aggregation formula represents the total number of BaP-attributable lung cancer cases across all grid cells within region, while the denominator represents the total lung cancer incidence in that region. Because I_R is treated as constant within a given country or region, it cancels out algebraically, and the national-level PAF_R reduces to a population-weighted average of grid-level PAF values reflecting the spatial distribution of population exposure.

4.3. Extrapolating Lung Cancer Burden from BaP to Ambient Air Pollution

4.3.1 Adopting BaP as A Mechanistic Anchor

To evaluate the suitability of benzo[a]pyrene (BaP) as a mechanistic anchor for extrapolating air-pollution–related lung cancer risk, we integrated emission correlation analysis, knowledge-graph–based biological pathway concordance, and country-level epidemiological validation within a unified analytical framework. The detailed description was provided in the section "3. Rationale for BaP as A Mechanistic Anchor" (SM). In briefly, we firstly quantified global co-emission structures between BaP and major atmospheric pollutants using Pearson correlation analysis. Emission inventories from 1990–2021 were log-transformed to stabilize variance and address right-skewed distributions. Correlation coefficients were calculated across countries and sectors to identify pollutants with consistent emission coupling to BaP, and temporal synchrony of emission trajectories was further examined in six representative economies to assess whether BaP and co-emitted pollutants share stable long-term emission dynamics. Second, mechanistic concordance between pollutants and lung cancer pathways was evaluated using Exposure–Biology–Disease knowledge graphs constructed in the EBIO module of ExposomeX platform (https://www.exposomex.cn/). Two complementary network modes were implemented: Exposure–GO–Disease (EGoD) and Exposure–Protein–Protein–Disease (EPPD). For each exposure–disease pair, all mechanistic paths were enumerated and summarized using pathway diversity scores based on the degree centrality of mediating GO terms or proteins, capturing pathway redundancy and hub-mediated biological connectivity. Third, we conducted country-level validation by comparing BaP-derived model-estimated burden indices with Global Burden of Disease (GBD) observed lung-cancer deaths attributable to air pollution. Pearson correlation analysis in log space was used to evaluate spatial and temporal agreement, assessing whether BaP-based attribution reproduces real-world mortality patterns.

4.3.2 Spatial Downscaling of National Air Pollution–Attributable Lung Cancer Deaths

Ambient air pollution is a complex mixture of multiple carcinogenic components. Among them, BaP is a well-established marker of combustion-related particulate pollution and has been widely used as a proxy for carcinogenic polycyclic aromatic hydrocarbons (PAHs). At the national scale, we observed a strong correspondence between BaP-based attributable fractions (PAF_{BaP}) and total air pollution–attributable fractions for lung cancer, indicating that spatial variation in BaP exposure captures a substantial share of the heterogeneity in air pollution-related lung cancer risk. Leveraging this correspondence, we used PAF_{BaP} as a spatially resolved risk surface to downscale national air pollution–attributable lung cancer deaths to finer spatial units. National lung cancer deaths attributable to ambient air pollution (D_c^{ap}) were estimated as described in the main analysis and treated as fixed national constraints for spatial allocation. The spatial downscaling was conducted on a global grid at a resolution of 0.1×0.1. This resolution balances the need to capture meaningful spatial heterogeneity in air pollution exposure while avoiding artificial precision beyond the resolution of available exposure and population datasets. Each grid cell (g) within country (c) is characterized by a population size $POP_{\mathrm{c,g}}$ and a BaP-based attributable fraction $PAF_{c,g}^{\mathrm{bap}}$. The same allocation framework applied at the provincial scale demonstrated improved agreement with external mortality benchmarks when population weighting was incorporated, particularly under lagged exposure assumptions consistent with lung cancer latency. These results provide empirical support for extending the same allocation logic to finer spatial units. Here, the spatial downscaling approach relies on the following four assumptions:

(i) Spatial heterogeneity in air pollution–attributable lung cancer burden within countries is primarily driven by variation in BaP-related attributable fractions and population density;

(ii) BaP-based attributable fractions adequately capture the spatial pattern of carcinogenic air pollution relevant to lung cancer risk;

(iii) Residual spatial heterogeneity in baseline lung cancer mortality unrelated to air pollution (e.g., smoking, age structure, healthcare access) does not systematically dominate the fine-scale spatial patterns implied by $PAF_{c,g}^{\mathrm{bap}} \times POP_{c,g}$, which were supported by Extended Data Fig. 12;

National air pollution-attributable lung cancer deaths provide a reliable constraint for spatial allocation. For each grid cell, we defined a spatial weight as:

$$w_{c,g} = PAF_{c,g}^{\mathrm{bap}} \times POP_{c,g}$$

This formulation follows directly from the definition of attributable burden, in which attributable deaths in a subpopulation are proportional to both the attributable fraction and the size of the population at risk. National air pollution–attributable lung cancer deaths were downscaled to grid cells by proportionally redistributing the national total according to the grid-level weights:

$$D_{c,g}^{\mathrm{ap}} = D_c^{\mathrm{ap}} \times \frac{w_{c,g}}{\sum_{g \in c} w_{c,g}}$$

By construction, this procedure ensures exact mass conservation at the national level:

$$\sum_{g \in c} D_{c,g}^{\mathrm{ap}} = D_c^{\mathrm{ap}}$$

The spatial allocation does not explicitly account for spatial variation in smoking prevalence, age structure, or healthcare access, which may modify absolute lung cancer mortality risk, but rather describe the spatial pattern of attributable burden consistent with national totals. Consequently, grid-level estimates should be interpreted as indicative of relative spatial patterns of air pollution–attributable burden rather than precise local mortality counts.

4.4. Quantifying spatial disparities and intergenerational inequalities

In this study, "*Intergenerational inequality*", captures systematic differences in cumulative lifetime risk between birth cohorts. It reflects how long-term air pollution exposure trajectories translate into systematically unequal lifetime lung cancer burdens across birth cohorts. Unlike spatial disparities, intergenerational inequalities explicitly incorporate a temporal and ethical dimension, addressing who bears the long-term consequences of past and present air pollution. We quantified spatial disparities across multiple nested geographic scales to capture both global and within-country heterogeneity in air pollution–attributable lung cancer mortality. At the global scale, analyses were conducted for 21 Global Burden of Disease (GBD) subregions and for all individual countries worldwide, enabling systematic comparison across major world regions and national contexts. At the national scale, we further examined country–province (or equivalent subnational) units for selected exemplar countries with diverse emission histories and regulatory trajectories. This multiscale framework allows assessment of whether apparent national-level improvements obscure persistent or widening disparities in lung cancer mortality burden at finer spatial resolutions.

To quantitatively evaluate intergenerational differences in environmental exposure, we developed a birth-cohort lifetime risk assessment module. This framework estimates the cohort-specific lifetime PAF (PAF_c), capturing the cumulative lung cancer risk attributable to BaP exposure over a standardized 70-year lifespan for individuals born in a given year. Unlike period-based burden estimation, this cohort-based approach does not incorporate the contemporaneous population age structure during risk calculation; demographic weights are applied only during spatial aggregation. At the 0.1°×0.1° grid scale, cohort-specific cumulative exposure (CE_c) and corresponding PAF_c are computed directly from the reconstructed lifetime concentration trajectory. For a cohort born in year t, the cumulative effective exposure over its 70-year lifespan, $CE_c(t)$, is determined by convolving the annual BaP concentration trajectory experienced from birth to age 69 with a biologically informed latency weighting function, thereby aligning exposure timing with carcinogenic risk development:

$$CE_c(t) = \sum_{a=0}^{69} C\,(t+a) \cdot W(a)$$

where $C(t+a)$ denotes the environmental BaP concentration experienced by the cohort at age a, where concentration fields include both historical reconstructions and future projections under the SSP scenarios. $W(a)$ represents the age-specific latency weight applied to exposure at age a, reflecting its relative contribution to carcinogenic risk. The same weighting function used in the cross-sectional S3 strategy was adopted to ensure consistency between period-based and cohort-based estimations. Based on the unit relative risk (*URR*) framework, the lifetime attributable risk for the cohort born in year t, denoted as $PAF_c(t)$, is calculated accordingly.

$$PAF_c(t) = 1 - URR^{-\left(\frac{CE_c(t)}{100}\right)}$$

Using gridded population data corresponding to each birth year as weights, cohort-specific grid-level $PAF_c(t)$ values were aggregated to the national scale to represent the true societal exposure burden experienced by each generation as:

$$PAF_{c,National}(t) = \frac{\sum(PAF_{c,grid}(t) \cdot Population_{grid}(t))}{\sum Population_{grid}(t)}$$

Acknowledgments: We would like to express our gratitude to the working group of environmental exposure and human health of the China Cohort Consortium (http://chinacohort.bjmu.edu.cn/).

Funding: This research was supported by the National Key Research & Development Program of Ministry of Science and Technology of China (Grant No. 2023YFC3708305; 2024YFE0100700), National Natural Science Foundation of China (Grant No. 52388101; 42477455; 42521004; 42177390; 42077390), Beijing Municipal Ecology and Environment Bureau (No. BJST20250209), the Strategy Priority Research Program (Category B) of the Chinese Academy of Sciences (XDB0750300), National Natural Science Foundation of China Key International (Regional) Joint Research Program (W2511067), Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM909).

Author contributions: B.W., L. Z. Y. C. conceived the study. H. Z., Y. C., X. Z., B. W., and B. F. curated data and performed the analysis. H. Z., Y. C., X. Z., and B.W. wrote the first draft of the manuscript. H. Z., Y. C., X. Z., B. W., and B. F., H. S., R. M., Q. W., B. P., M. F., T. L., G. S., J. X., Z. B., and Y. C. reviewed and edited the manuscript. B.W., L. Z., and Y. C. managed the program.

Competing interests: Authors declare no competing interests.

Data and Code materials availability: The demo code used in this study can be accessible on GitHub (https://github.com/ExposomeX/Data-and-Code-for-BAP-Lung-Cancer-Manuscript.git), or can publicly available via cloud storage at:
https://pan.baidu.com/s/17F4P1Kxf5QTxGPhqyX_Llg?pwd=bbw3 (Access code: bbw3).
The data can publicly available via cloud storage at:
https://pan.baidu.com/s/1RIC1JSblLHcxejqPxUtWHQ?pwd=4npn (Access code: 4npn).
Readers can contact Dr. Bin Wang for more detailed information, Email: binwang@pku.edu.cn

List of Supplementary Materials:

The following information was provided in the file of Supplementary Materials, i.e., Supplementary Sections, Supplementary Data, Supplementary Code, Figs. S1 to S6, and Tables S1 to S10. Their detailed titles were listed as below:

- Supplementary Sections
- Supplementary Data
- Supplementary Code
- Table S1. Selected Climate Models Included in the CMIP6 Multi-Model Ensemble Mean
- Table S2. Classification and Definitions of Emission Sectors
- Table S3. Detailed Description of Predictor Variables
- Table S4. Comparative Summary of Exposure Assessment Strategies (S0–S3)
- Table S5. Global Population-Weighted Benzo[a]pyrene (BaP) Concentrations from 1950 to 2100 Under Different Emission Scenarios
- Table S6. Global Population Attributable Fractions of Lung Cancer Induced by Benzo[a]pyrene (BaP)
- Table S7. Global Distribution of Lung Cancer Deaths Attributable to Ambient Air Pollution During 1990–2023
- Table S8. Exposure–Biological Mechanism–Disease Network Built By E-BIO Module
- Table S9. Country-Level Lung Cancer Deaths From the GBD Database and Latency-Aware Benzo[a]pyrene-Based Estimates Under the S3 Scenario
- Table S10. Global Lifetime Population Attributable Fractions of Lung Cancer Induced by Benzo[a]pyrene
- Fig. S1. Latency-Aware Reconstruction of Global Lung Cancer Risk Attributable to Benzo[a]pyrene (BaP) Under the SSP1–2.6 Scenario

- Fig. S2. Latency-Aware Reconstruction of Global Lung Cancer Risk Attributable to Benzo[a]pyrene (BaP) Under the SSP3–7.0 Scenario
- Fig. S3. Latency-Aware Reconstruction of Global Lung Cancer Risk Attributable to Benzo[a]pyrene (BaP) Under the SSP5–8.5 Scenario
- Fig. S4. Intergenerational Inequality in Lifetime Lung Cancer Risk Attributable to Ambient Air Pollution Under the SSP1–2.6 Scenario.
- Fig. S5. Intergenerational Inequality in Lifetime Lung Cancer Risk Attributable to Ambient Air Pollution Under the SSP3–7.0 Scenario
- Fig. S6. Intergenerational Inequality in Lifetime Lung Cancer Risk Attributable to Ambient Air Pollution Under the SSP5–8.5 Scenario

Reference:

1 Fajersztajn, L., Veras, M., Barrozo, L. V. & Saldiva, P. Air pollution: a potentially modifiable risk factor for lung cancer. *Nat Rev Cancer* **13**, 674-678 (2013). https://doi.org/10.1038/nrc3572

2 Cohen, A. J. & Pope, C. I. I. I. Lung cancer and air pollution. *Environ Health Persp* **103**, 219-224 (1995). https://doi.org/10.2307/3432314

3 Leiter, A., Veluswamy, R. R. & Wisnivesky, J. P. The global burden of lung cancer: current status and future trends. *Nat Rev Clin Oncol* **20**, 624-639 (2023). https://doi.org/10.1038/s41571-023-00798-3

4 Thiery, W. *et al.* Intergenerational inequities in exposure to climate extremes Young generations are severely threatened by climate change. *Science* **374**, 158-160 (2021). https://doi.org/10.1126/science.abi7339

5 Wang, Y. Z. *et al.* Air quality policy should quantify effects on disparities. *Science* **381**, 272-274 (2023). https://doi.org/10.1126/science.adg9931

6 Vollset, S. E. *et al.* Burden of disease scenarios for 204 countries and territories, 2022-2050: a forecasting analysis for the Global Burden of Disease Study 2021. *Lancet* **403**, 2204-2256 (2024). https://doi.org/10.1016/S0140-6736(24)00685-8

7 Yang, H., Huang, X. Y., Westervelt, D. M., Horowitz, L. & Peng, W. Socio-demographic factors shaping the future global health burden from air pollution. *Nat Sustain* **6**, 58-68 (2023). https://doi.org/10.1038/s41893-022-00976-8

8 Archer, V. E., Coons, T., Saccomanno, G. & Hong, D. Y. Latency and the lung cancer epidemic among United States uranium miners. *Health Phys* **87**, 480-489 (2004). https://doi.org/10.1097/01.hp.0000133216.72557.ab

9 Nadler, D. & Zurbenk, I. Estimating Cancer Latency Times Using a Weibull Model. *Adv Epidemiol* **2014**, 746769 (2014). https://doi.org/doi.org/10.1155/2014/746769

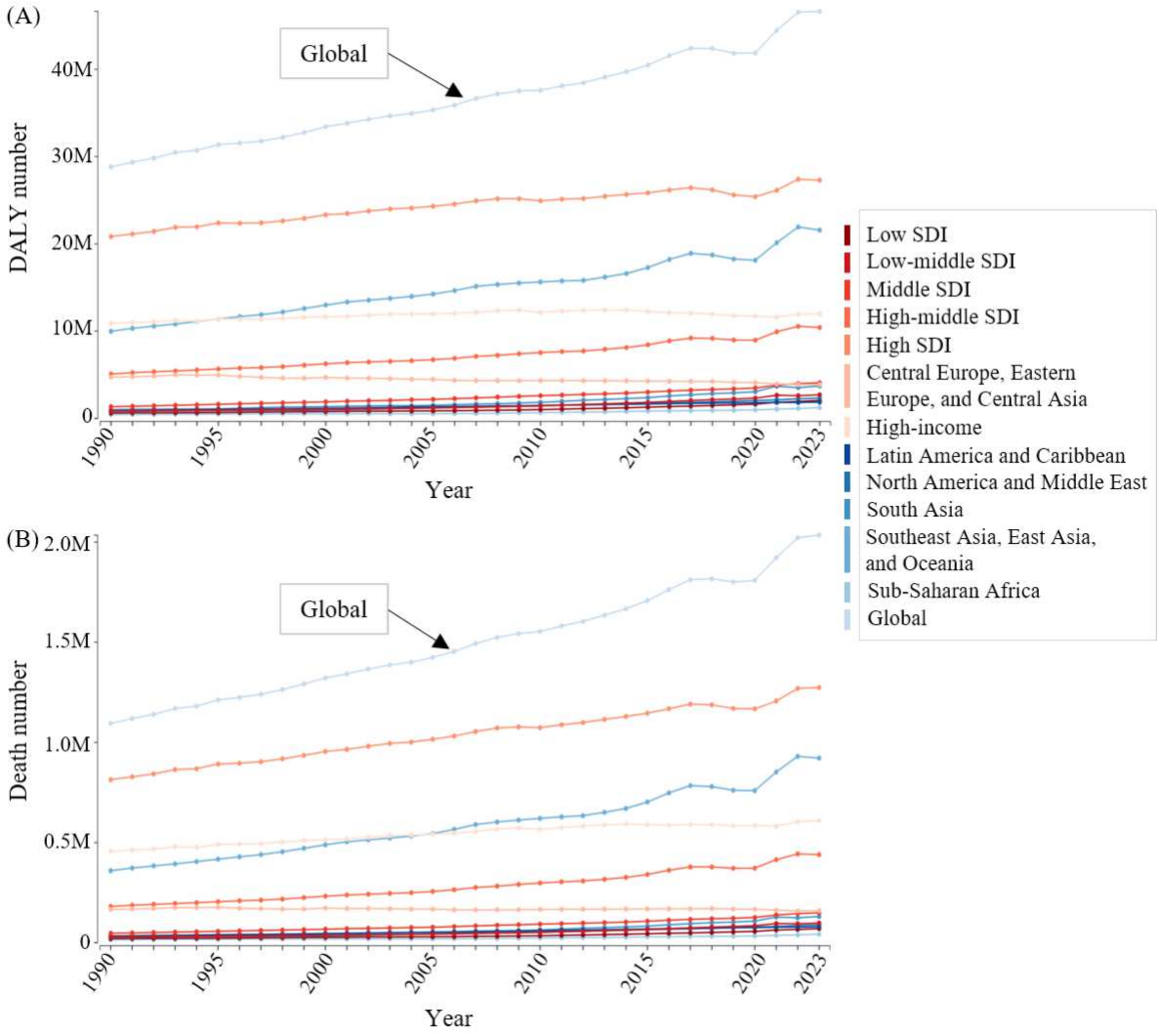
10 Zhang, Y. X., Tao, S., Shen, H. Z. & Ma, J. M. Inhalation exposure to ambient polycyclic aromatic hydrocarbons and lung cancer risk of Chinese population. *P Natl Acad Sci USA* **106**, 21063-21067 (2009). https://doi.org/10.1073/pnas.0905756106

11 Wu, Z. C. *et al.* Modeling of polycyclic aromatic hydrocarbons (PAHs) from global to regional scales: model development (IAP-AACM_PAH v1.0) and investigation of health risks in 2013 and 2018 in China. *Geosci Model Dev* **17**, 8885-8907 (2024). https://doi.org/10.5194/gmd-17-8885-2024

12 Shen, H. Z. *et al.* Global lung cancer risk from PAH exposure highly depends on emission sources and individual susceptibility. *Sci Rep-Uk* **4**, 6561 (2014). https://doi.org/10.1038/srep06561

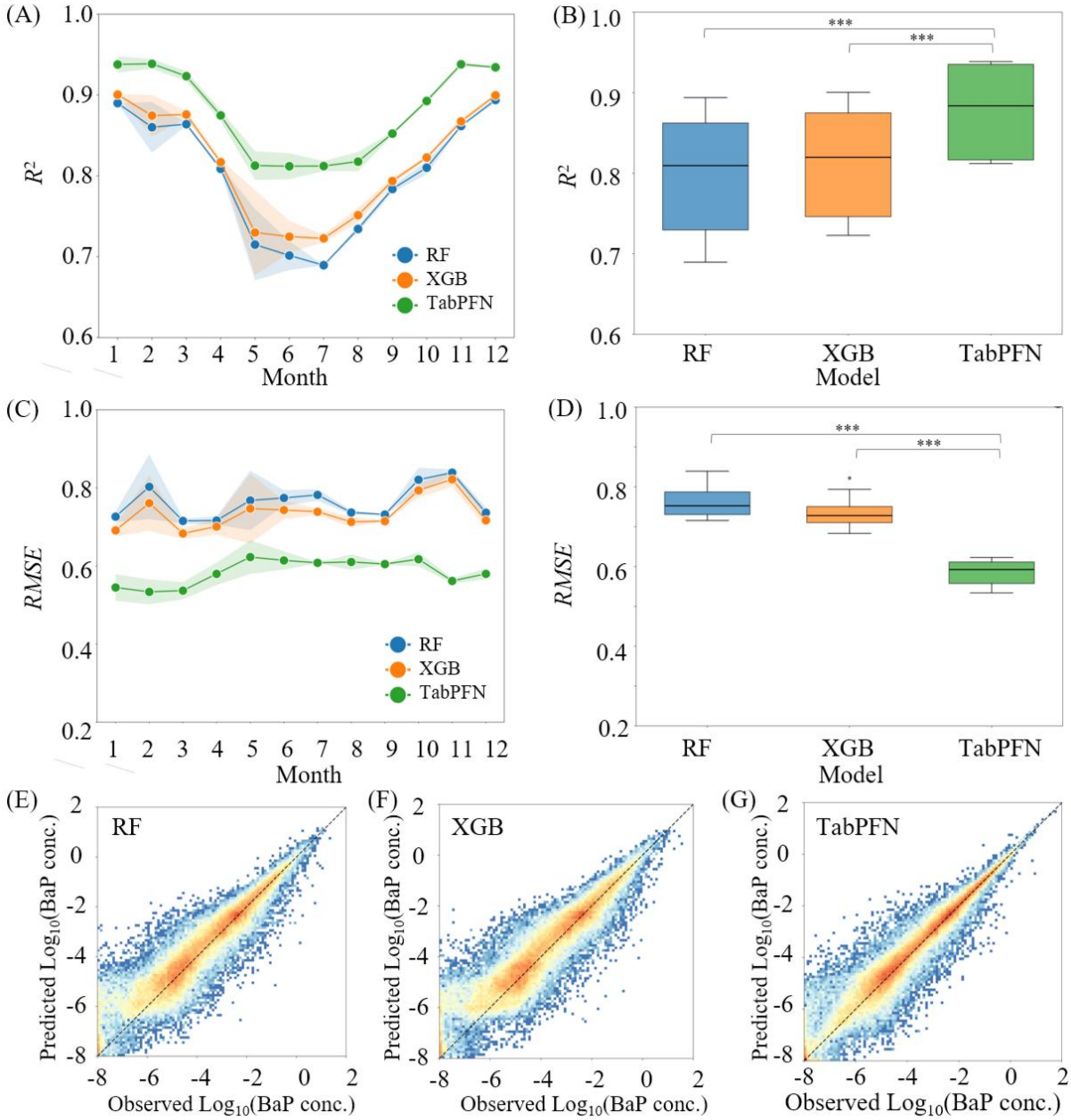
13 Yu, Y. X. *et al.* Risk of human exposure to polycyclic aromatic hydrocarbons: A case study in Beijing, China. *Environ Pollut* **205**, 70-77 (2015). https://doi.org/10.1016/j.envpol.2015.05.022
14 Zheng, H. T. *et al.* Control of toxicity of fine particulate matter emissions in China. *Nature* **643**, 404-411 (2025). https://doi.org/10.1038/s41586-025-09158-w
15 Zhang, Q. *et al.* Integrated Evidence for Lysosomal Dysfunction-Mediated Iron Dysregulation induced by PM2.5 Exposure. *Environ Sci Technol* **60**, 5324-5336 (2026). https://doi.org/PMID 21315510.1021/acs.est.5c14185
16 Hill, W. *et al.* Lung adenocarcinoma promotion by air pollutants. *Nature* **616**, 159–167 (2023). https://doi.org/10.1038/s41586-023-05874-3
17 Armstrong, B., Hutchinson, E., Unwin, J. & Fletcher, T. Lung cancer risk after exposure to polycyclic aromatic hydrocarbons: A review and meta-analysis. *Environ Health Persp* **112**, 970-978 (2004). https://doi.org/10.1289/ehp.6895
18 Zhu, Y. *et al.* Environmental Distributions of Benzo[]pyrene in China: Current and Future Emission Reduction Scenarios Explored Using a Spatially Explicit Multimedia Fate Model. *Environ Sci Technol* **49**, 13868-13877 (2015). https://doi.org/10.1021/acs.est.5b00474
19 Turner, M. C. *et al.* Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations. *Ca-Cancer J Clin* **70**, 460-479 (2020). https://doi.org/10.3322/caac.21632
20 Zhao, Y. *et al.* Deep Learning Prediction of Polycyclic Aromatic Hydrocarbons in the High Arctic. *Environ Sci Technol* **53**, 13238-13245 (2019). https://doi.org/10.1021/acs.est.9b05000
21 Bodnar, C. *et al.* A foundation model for the Earth system. *Nature* **641**, 1180-1187 (2025). https://doi.org/10.1038/s41586-025-09005-y
22 Shrivastava, M. *et al.* Global long-range transport and lung cancer risk from polycyclic aromatic hydrocarbons shielded by coatings of organic aerosol (vol 114, pg 1246, 2017). *P Natl Acad Sci USA* **114**, E2263-E2263 (2017). https://doi.org/10.1073/pnas.1702221114
23 Hollmann, N. *et al.* Accurate predictions on small data with a tabular foundation model. *Nature* **637**, 319-326 (2025). https://doi.org/10.1038/s41586-024-08328-6
24 Xu, F. J. *et al.* The challenge of population aging for mitigating deaths from PM air pollution in China. *Nat Commun* **14**, 5222 (2023). https://doi.org/10.1038/s41467-023-40908-4.
25 Chang, C. Y. *et al.* Chronic exposure to carbon black ultrafine particles reprograms macrophage metabolism and accelerates lung cancer. *Sci Adv* **8**, eabq0615 (2022). https://doi.org/10.1126/sciadv.abq0615
26 Tao, C. L., Zhang, Y. Q., Shindell, D. & Zhang, H. L. Inequality in air pollution-attributable mortality by income level between and within countries. *P Natl Acad Sci USA* **122**, e2511394122 (2025). https://doi.org/10.1073/pnas.2511394122
27 Connolly, R. *et al.* Mortality attributable to PM from wildland fires in California from 2008 to 2018. *Sci Adv* **10**, eadl1252 (2024). https://doi.org/doi: 10.1126/sciadv.adl1252
28 Nawaz, O. & Henze, D. National climate action can ameliorate, perpetuate, or exacerbate international air pollution inequalities. *Nature Communications* **17**, 1649 (2026). https://doi.org/10.1038/s41467-026-68827-0
29 Yue, H. B., He, C. Y., Huang, Q. X., Yin, D. & Bryan, B. A. Stronger policy required to substantially reduce deaths from PM pollution in China. *Nat Commun* **11**, 1462 (2020). https://doi.org/10.1038/s41467-020-15319-4

30 Schucht, S. *et al.* Costs of air pollution from European industrial facilities 2008–2017. Report No. ETC/ATNI 2020/4, (European Environment Agency; European Topic Centre on Air pollution,transportnoise and industrial pollution, 2021).
31 Lin, N. *et al.* Age disparities in PFAS-associated influenza-like illness among adult females at low exposure levels. *J Hazard Mater* **504**, 141225 (2026). https://doi.org/10.1016/j.jhazmat.2026.141225
32 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). https://doi.org/10.1021/acs.est.5c05956



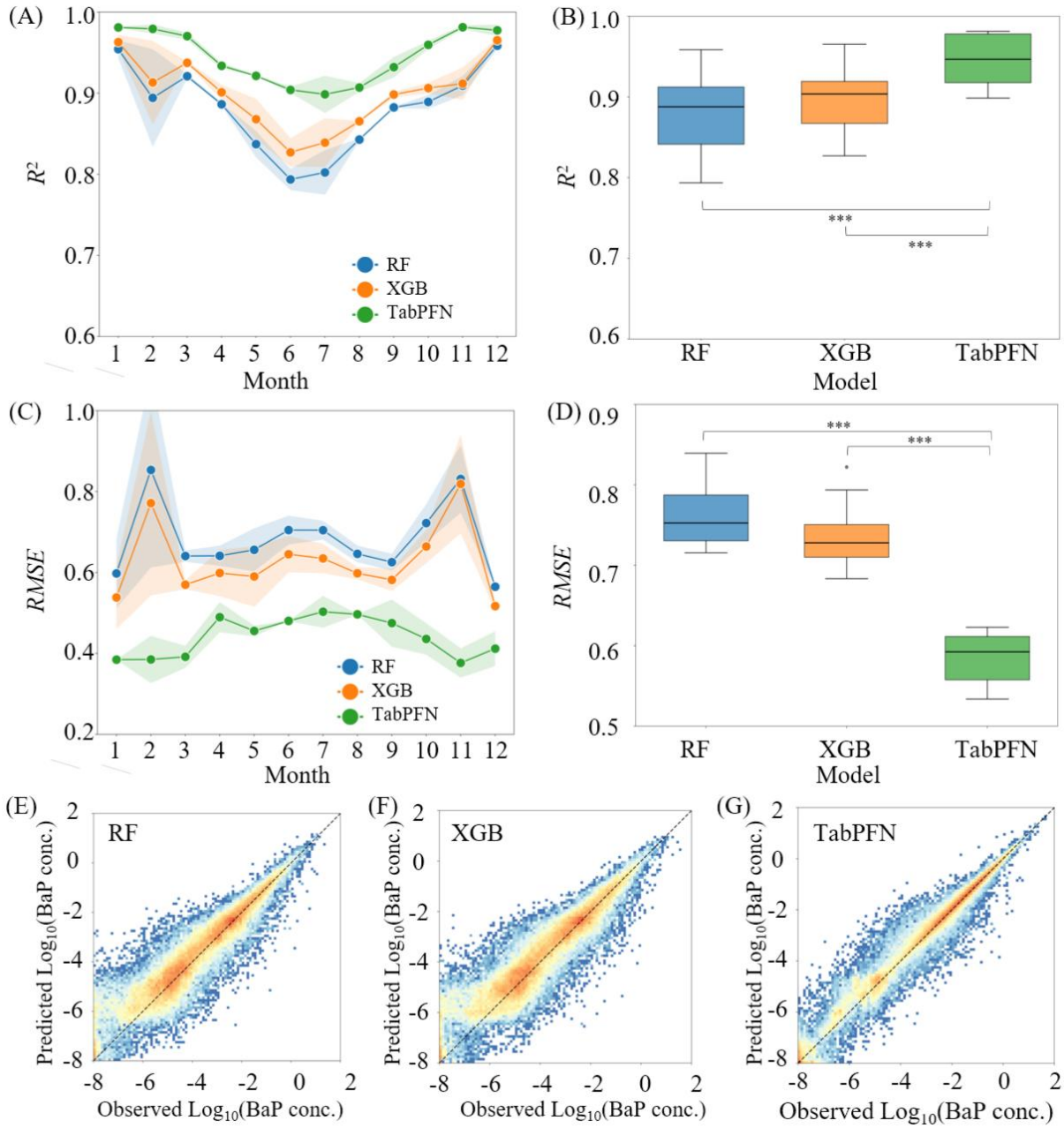
Extended Data Fig. 1. Temporal Trends in Global Lung Cancer Burden Across the World and Major Regions (1990–2023)

(A) Disability-adjusted life years (DALYs) attributable to lung cancer. (B) Corresponding number of deaths attributable to lung cancer. Lines represent estimates from the Global Burden of Disease (GBD) study stratified by Socio-demographic Index (SDI) categories and major geographic regions. Global totals are shown for reference. Data were obtained from the Global Health Data Exchange visualization tool: https://vizhub.healthdata.org/gbd-compare/.

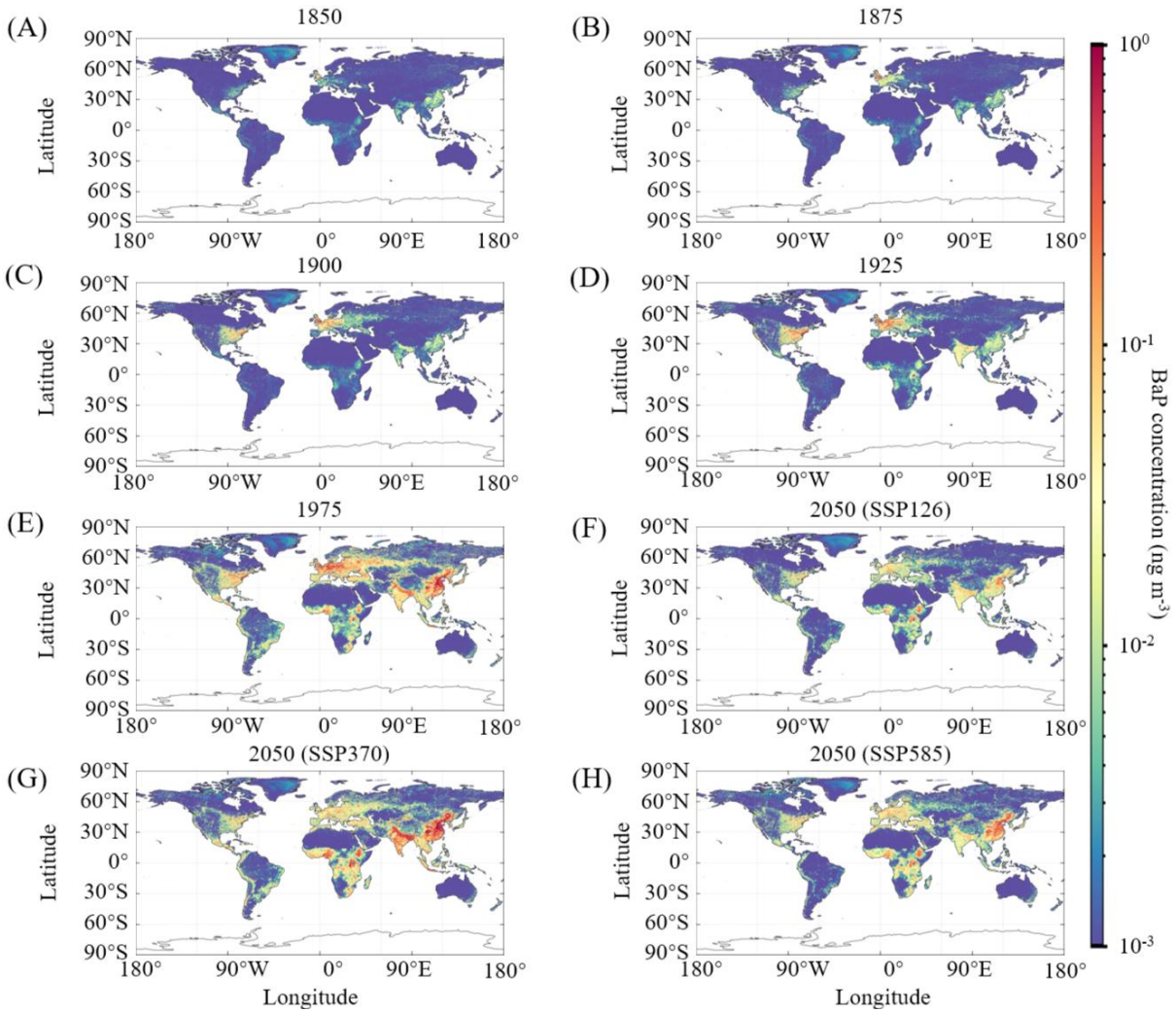


Extended Data Fig. 2. Global Predictive Performance of Benzo[a]pyrene (BaP) Reconstruction Models During the Independent Test Period (2006–2007).

(A, C) Monthly coefficient of determination (R^2) and root mean square error (RMSE) between predicted and reference BaP concentrations across the global domain. Concentration unit: ng m^{-3}. (B, D) Summary statistics of model performance for land grid cells. (E–G) Density scatterplots comparing predicted and observed log_{10}-transformed BaP concentrations for Random Forest (RF), Extreme Gradient Boosting (XGB), and Tabular Prior-data Fitted Network (TabPFN) models. *** $P < 0.001$, Student's T-test.

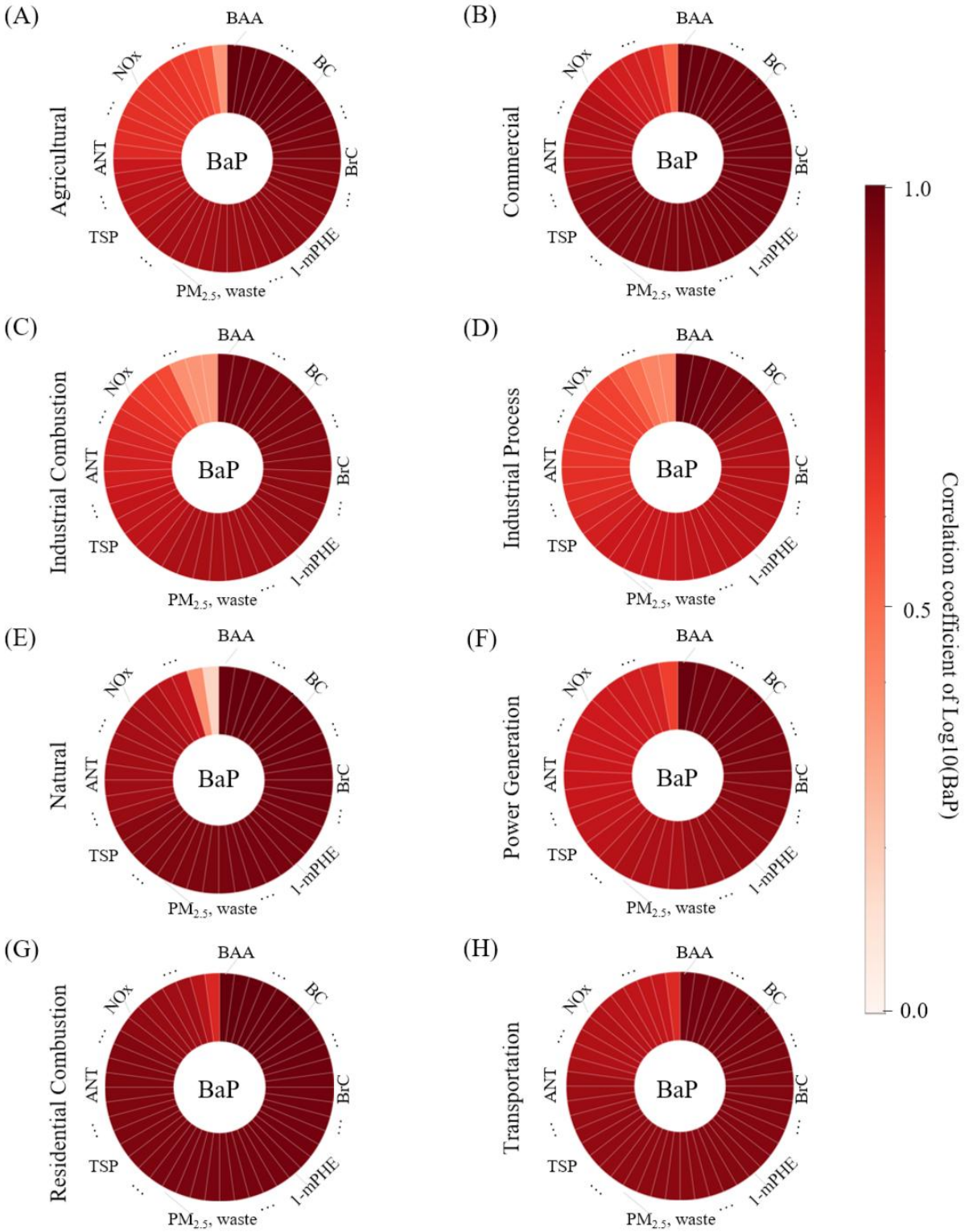


Extended Data Fig. 3. Predictive Performance of Global Benzo[a]pyrene (BaP) Reconstruction Models Over Land During the Independent Test Period (2006–2007). (A, C) Monthly coefficient of determination (R^2) and root mean square error (RMSE) between predicted and reference BaP concentrations across global land grid cells. Concentration unit: ng m^{-3}. (B, D) Distribution of model performance across land grid cells. (E–G) Density scatterplots comparing predicted and observed log_{10}-transformed BaP concentrations for Random Forest (RF), Extreme Gradient Boosting (XGB), and Tabular Prior-data Fitted Network (TabPFN) models. *** $P < 0.001$, Student's *T*-test.



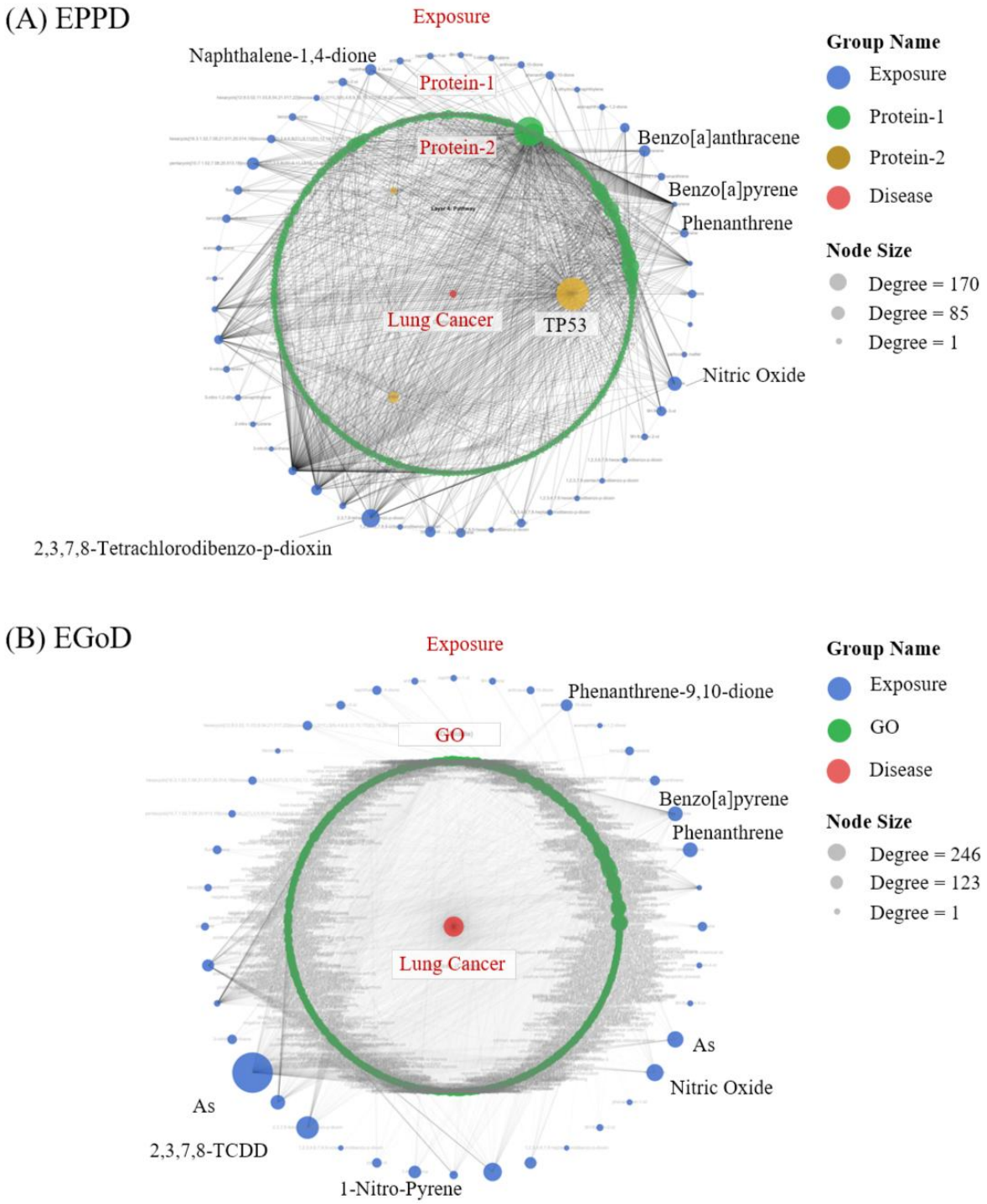
Extended Data Fig. 4. Global Distribution of Benzo[a]pyrene (BaP) Concentrations at 0.1° × 0.1° Resolution for Historical Reconstruction and Future Scenarios.

Global BaP concentrations are shown for selected years: (A) 1850, (B) 1875, (C) 1900, (D) 1925, and (E) 1975 based on historical reconstruction. Future projections for 2050 are presented under the Shared Socioeconomic Pathway scenarios (F) SSP-126, (G) SSP-370, and (H) SSP-585. Colors denote BaP concentrations (ng m^{-3}, logarithmic scale). Abbreviations: Shared Socioeconomic Pathway 1-2.6 (SSP-126); Shared Socioeconomic Pathway 2-4.5 (SSP-245); Shared Socioeconomic Pathway 3-7.0 (SSP-370); Shared Socioeconomic Pathway 5-8.5 (SSP-585)



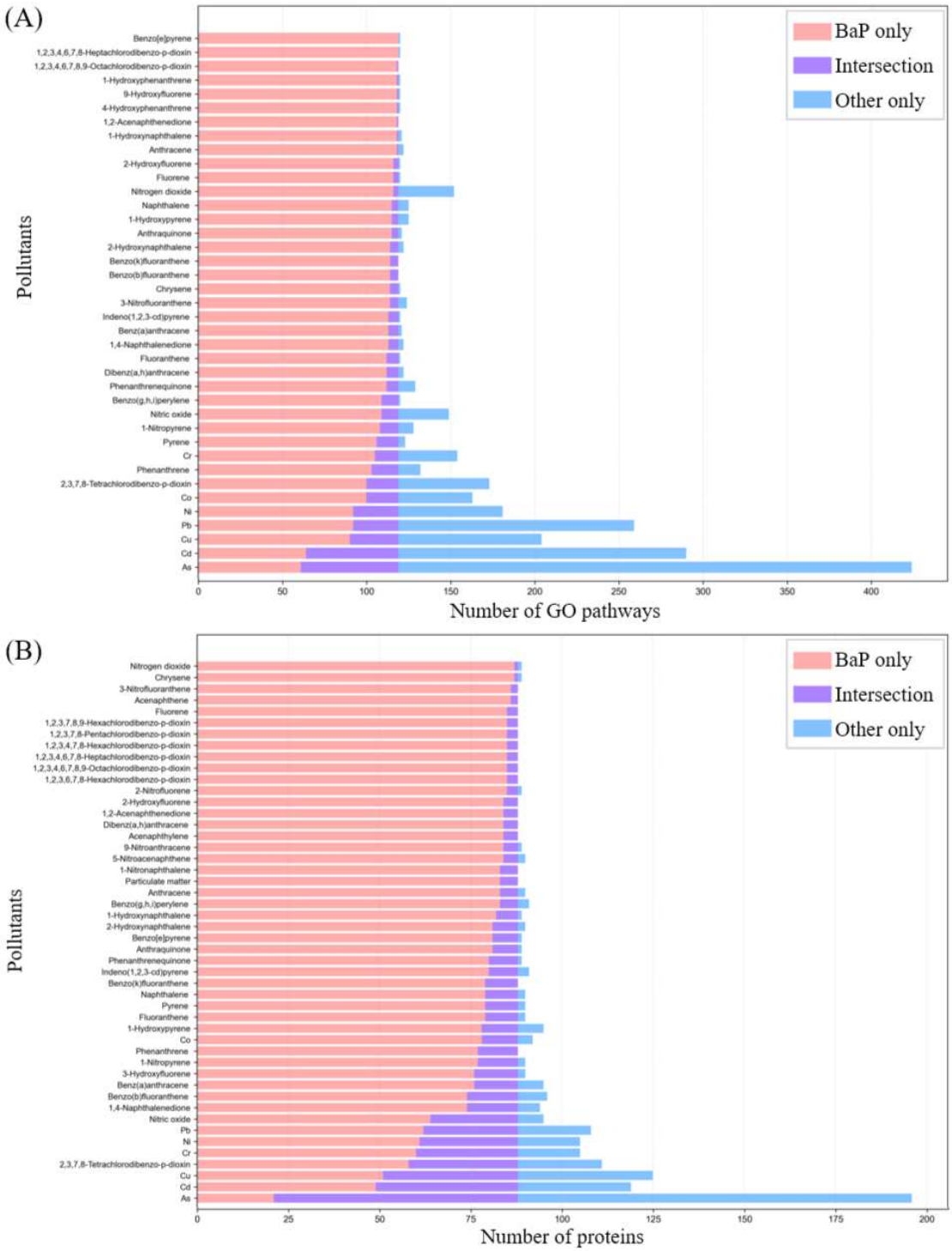
Extended Data Fig. 5. Sector-Specific Concordance Between Benzo[a]pyrene (BaP) and Major Particulate Matter (PM) Components.

(A) agriculture, (B) commercial activities, (C) global aggregate emissions, (D) industrial combustion, (E) industrial processes, (F) natural sources, (G) power generation, and (H) transportation sectors.



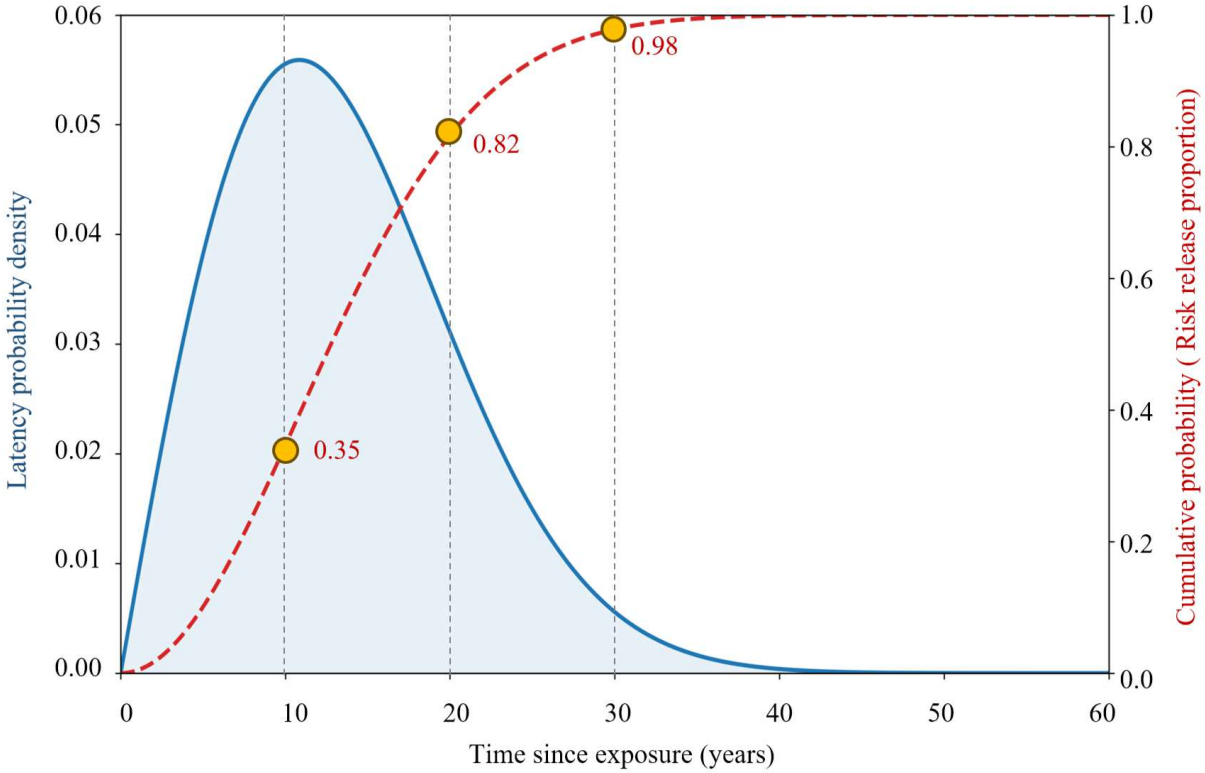
Extended Data Fig. 6. Shared Molecular Pathways Linking Benzo[a]pyrene (BaP) and Broader Air Pollution Components to Lung Cancer

(A) Exposure–Protein–Protein–Disease (EPPD) network illustrating interactions among combustion-related pollutants, intermediate proteins (e.g., TP53), and lung cancer. (B) Exposure–Gene Ontology–Disease (EGoD) network highlighting overlapping biological processes associated with BaP and other air pollution components. Node size represents connectivity (degree), and edge density reflects shared mechanistic associations.



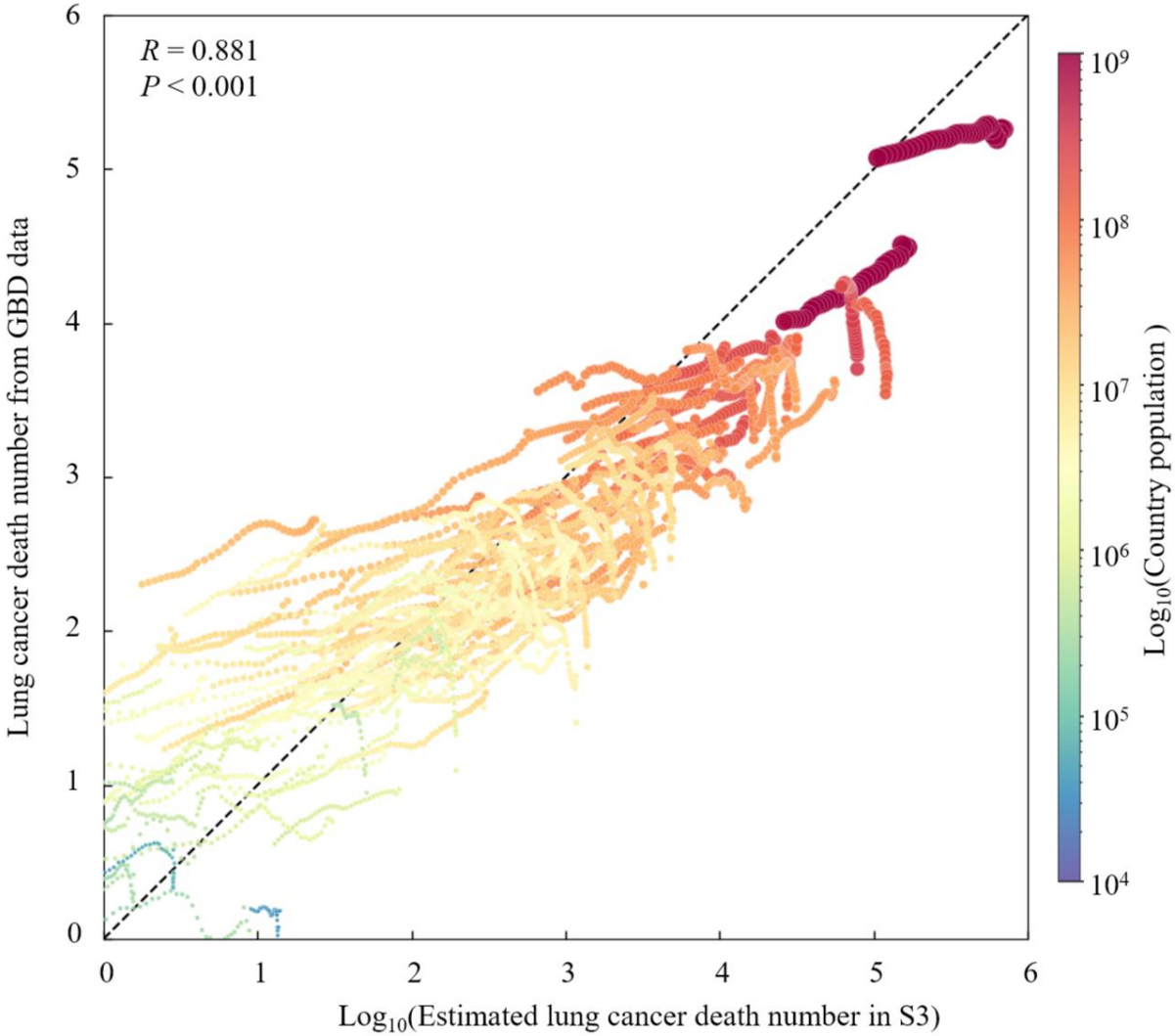
Extended Data Fig. 7. Shared Biological Mechanisms Linking Benzo[a]pyrene (BaP) and Other Air Pollutants to Lung Cancer.

(A) Number of Gene Ontology (GO) pathways associated with each pollutant-lung cancer relationship within the EBIO framework. (B) Number of lung cancer-related proteins linked to each pollutant. Bars represent BaP-specific associations (BaP only), shared mechanisms with BaP (intersection), and pollutant-specific associations independent of BaP (other only).



Extended Data Fig. 8. Latency Weighting Function for Lung Cancer Risk Following Exposure

The latency structure linking historical exposure to lung cancer risk was modeled using a Weibull distribution. The blue curve represents the probability density function (PDF), indicating the relative contribution of exposures occurring at different lag times since exposure. The red dashed curve represents the cumulative distribution function (CDF), showing the cumulative probability that the latency period has been completed. Vertical reference lines mark key latency milestones: approximately 35% of cumulative risk occurs within 10 years after exposure, 82% within 20 years, and 98% within 30 years.



Extended Data Fig. 9. Concordance Between Latency-Aware Benzo[a]pyrene (BaP)-Based Estimates and Global Burden of Disease (GBD) Lung Cancer Deaths

Country-level log₁₀-transformed lung cancer deaths estimated using the latency-aware BaP-based framework (S3) are compared with corresponding estimates from the Global Burden of Disease (GBD) study for 1990–2023. The dashed line indicates the 1:1 reference. Strong agreement is observed across countries (Pearson correlation coefficient $R = 0.881$, $P < 0.001$).

Supplementary Files

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

- SupplementaryMaterials.docx