









## Nontarget mass spectrometry and in silico molecular characterization of air pollution from the Indian subcontinent

Stefano Papazian <sup>1</sup>, Lisa A. D'Agostino<sup>1</sup>, Ioannis Sadiktsis <sup>1,5</sup>, Jean Froment <sup>1</sup>, Bénilde Bonnefille <sup>1</sup>, Kalliroi Sdougkou <sup>1</sup>, Hongyu Xie<sup>1</sup>, Ioannis Athanassiadis <sup>1</sup>, Krishnakant Budhavant<sup>2,3</sup>, Sanjeev Dasari<sup>4,6</sup>, August Andersson <sup>4</sup>, Örjan Gustafsson<sup>4</sup> & Jonathan W. Martin <sup>1</sup>✉

Fine particulate-matter is an important component of air pollution that impacts health and climate, and which delivers anthropogenic contaminants to remote global regions. The complex composition of organic molecules in atmospheric particulates is poorly constrained, but has important implications for understanding pollutant sources, climate-aerosol interactions, and health risks of air pollution exposure. Here, comprehensive nontarget high-resolution mass spectrometry was combined with in silico structural prediction to achieve greater molecular-level insight for fine particulate samples ( $n = 40$ ) collected at a remote receptor site in the Maldives during January to April 2018. Spectral database matching identified 0.5% of 60,030 molecular features observed, while a conservative computational workflow enabled structural annotation of 17% of organic structures among the remaining molecular dark matter. Compared to clean air from the southern Indian Ocean, molecular structures from highly-polluted regions were dominated by organic nitrogen compounds, many with computed physicochemical properties of high toxicological and climate relevance. We conclude that combining nontarget analysis with computational mass spectrometry can advance molecular-level understanding of the sources and impacts of polluted air.

<sup>1</sup> Department of Environmental Science (ACES, Exposure & Effects), Science for Life Laboratory, Stockholm University, Stockholm 106 91, Sweden. <sup>2</sup> Maldives Climate Observatory at Hanimaadhoo (MCOH), Hanimaadhoo 02020, Maldives. <sup>3</sup> Divecha Centre for Climate Change, Indian Institute of Science (IISc), Bangalore 560012, India. <sup>4</sup> Department of Environmental Science (ACES, Biogeochemistry), Bolin Centre for Climate Research, Stockholm University, Stockholm 106 91, Sweden. <sup>5</sup> Present address: Department of Materials and Environmental Chemistry, Stockholm University, 106 91 Stockholm, Sweden. <sup>6</sup> Present address: Institute of Environmental Geosciences, University Grenoble Alpes, CNRS, IRD, Grenoble INP, 38000 Grenoble, France. ✉email: [jon.martin@aces.su.se](mailto:jon.martin@aces.su.se)

Particulate matter (PM) is a major component of air pollution that impacts health and global climate. The fine fraction ( $<2.5\ \mu\text{m}$ ,  $\text{PM}_{2.5}$ ) is responsible for millions of premature deaths annually<sup>1,2</sup> and is a risk factor for chronic illness and cancer<sup>3,4</sup>. Ambient levels in low- and middle-income countries of Asia have been declared public health emergencies<sup>1,5,6</sup>.  $\text{PM}_{2.5}$  has intercontinental spatial impacts and transports persistent organic contaminants from populated regions to remote global regions<sup>7–9</sup>. PM furthermore affects aerosol-sunlight and aerosol-cloud interactions<sup>10</sup> and is generally believed to contribute to climate cooling through light-scattering and cloud condensation, although black carbon and brown carbon (BrC; light-absorbing organic matter) components may also lead to warming through absorption of solar radiation<sup>11,12</sup>. These opposing factors remain poorly constrained in climate models<sup>11–13</sup>. Organic molecules can be a major mass fraction of total PM<sup>14</sup>, thus a comprehensive molecular characterization of  $\text{PM}_{2.5}$  could contribute to improved understanding of global air pollution sources, climate impacts, and health effects.

High-resolution mass spectrometry (HRMS) is an established instrumental technique that can reveal the molecular complexity of  $\text{PM}_{2.5}$  organic compounds, however, most substances remain uncharacterized beyond assignment of molecular formula or the presence of certain functional groups<sup>15–22</sup>. Characterization of  $\text{PM}_{2.5}$  samples by HRMS also creates high demands on data processing which has limited previous detailed studies to only a few atmospheric samples. In metabolomics and proteomics, high-throughput workflows for batch-processing of full-scan HRMS chromatographic data (i.e.  $\text{MS}^1$ ) and the associated fragmentation spectra ( $\text{MS}^2$ ) are now applied to explore the chemical structures of ‘molecular dark matter’ in biological systems<sup>23–26</sup>. Such methods have yet to be applied in atmospheric research, but could open new molecular-level windows for studies of air pollution. We hypothesized that new insights into the molecular composition and effects of  $\text{PM}_{2.5}$  organic compounds could be achieved by combining comprehensive nontarget HRMS analysis with computational workflows<sup>23–26</sup>.

Here,  $\text{PM}_{2.5}$  was continuously collected throughout January–April by high-volume sampling ( $n = 40$  samples) at the Maldives Climate Observatory at Hanimaadhoo (MCOH) as part of the South Asian Pollution Experiment, 2018 (SAPOEX-18)<sup>11</sup>. During these months, the MCOH enables sampling of highly polluted plumes originating from the Indian subcontinent<sup>4,11</sup>, and occasional pristine air from the Southern Indian Ocean<sup>27,28</sup>. Polluted air in this geographical hotspot leads to millions of premature deaths<sup>1,2,4</sup> and the associated ‘Atmospheric Brown Cloud’ extends south of the equator, influencing atmospheric energy balances over a vast region<sup>29</sup>. Previous studies of PM in this region have highlighted the major fraction of sunlight-absorbing BrC<sup>30</sup>.

To achieve a broad molecular characterization of organic compounds, each  $\text{PM}_{2.5}$  sample was extracted with a range of solvents and analyzed by gas-chromatography (GC)-HRMS electron ionization (EI) and negative chemical ionization (NCI), or by high-performance liquid chromatography (LC)-HRMS electrospray ionization (ESI) in positive and negative mode. This approach resulted in six unique molecular profiles per sample (Fig. 1a), and revealed high molecular complexity (Fig. 1b–d). Known anthropogenic contaminants were confirmed (Level 1), including legacy persistent organic pollutants, polycyclic aromatic hydrocarbons (PAHs), plasticizers, pesticides, and associated transformation intermediates. However, these identifications represented only a minor portion of all detected molecules. Hence, by integration of open-source cheminformatics and computational workflows, including molecular networking<sup>25</sup>,  $\text{MS}^2$ -guided *in silico* structural predictions<sup>31</sup>, and physicochemical property

estimation of optical and toxicological relevance<sup>32–34</sup>, we proceeded to characterize thousands of structures among the remaining unknown molecules. The molecular properties of these structures were evaluated with consideration of potential health and climate impacts.

## Results and discussion

**Comprehensive nontarget analysis of  $\text{PM}_{2.5}$ .** After quality control and field-blank correction, the combined analyses of 40  $\text{PM}_{2.5}$  samples (120 extracts) revealed 60,030 molecular features (Supplementary Data 2—Dataset). Each feature is defined by a retention time (Rt) in the chromatographic dimension for GC and LC, and for GC analyses a mass spectrum dimension corresponding to mass-to-charge ratio ( $m/z$ ) with a base-peak ion and deconvoluted  $\text{MS}^1$  spectrum (EI and NCI), and for LC analyses by a precursor  $\text{MS}^1$  (full-scan) and corresponding deconvoluted data-independent (DIA)  $\text{MS}^2$  spectrum.

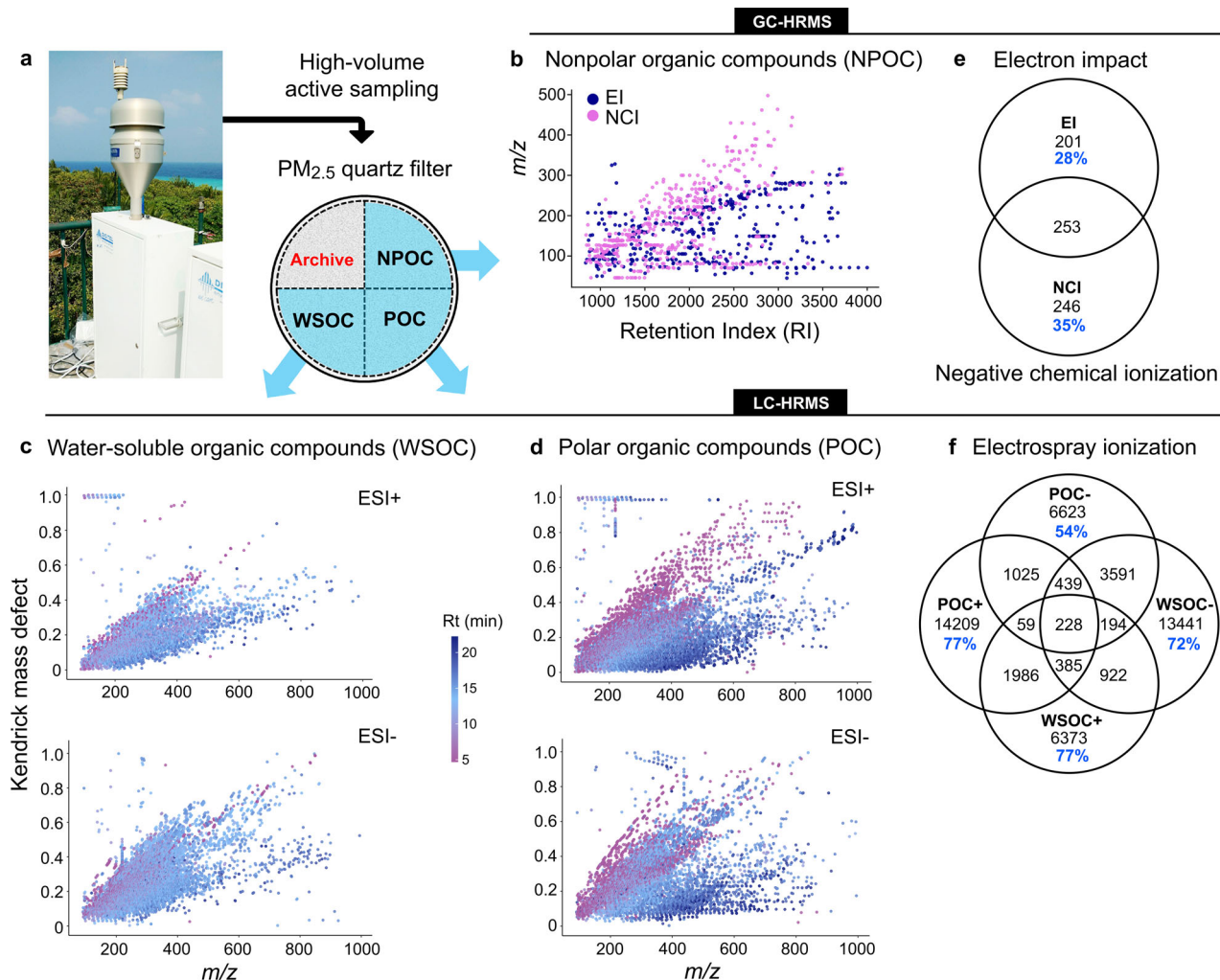
The number of features detected in water-soluble, polar, and nonpolar organic compound extracts (WSOC, POC, and NPOC, respectively) across the four modes of instrumental analysis (Fig. 1b, c, d GC-EI, GC-NCI, LC-ESI+, LC-ESI-) indicated that a great proportion of molecules were unique to each dataset (Fig. 1e, f), and thus that a battery of approaches was important for achieving comprehensive analysis of  $\text{PM}_{2.5}$  (see criteria in Supplementary methods—Estimation of unique features).

The greatest molecular complexity was found in the WSOC and POC extracts (98% of all features). This was not unexpected because hydrophilic compounds represent the largest fraction (up to 50–100%) of aerosol organic compounds<sup>11</sup>. Moreover, the remote MCOH sampling location allows substantial atmospheric oxidation of transported pollution to occur prior to collection<sup>11</sup>. Altogether, the large number of samples, and the battery of extracts and analytical modes employed here resulted in a greater view of molecular complexity than reported in previous analyses of atmospheric PM<sup>18,22,35</sup>.

## Back-trajectories and geographical sources of air pollution.

Throughout the campaign, back-trajectories showed that sampled air originated from four geographical regions (Fig. 2a, b, Fig. S8 and Supplementary Data 3—Back-trajectories), including three regions to the north that cover reaches of the Indian subcontinent (i.e. Arabian Sea, Indo-Gangetic Plain, and Peninsular India), and a fourth region originating in the Southern Indian Ocean (Fig. 2a, b). The frequency contributions of air from these four back-trajectories (Fig. 2a) were used to model the chemical variation observed in each 48 hr sample; considering all combined features from all fractions (NPOC, POC, WSOC). The resulting multivariate model (Fig. 2c) explained variation among molecular profiles by geographical source, in particular, the first latent variable significantly separated the Southern Indian Ocean cluster from the three subcontinental clusters. Based on satellite data, the back-trajectories of air coming from the three subcontinental regions coincided with much higher tropospheric nitrogen dioxide ( $\text{NO}_2$ ) concentrations (Fig. 2d), a generic indicator of air pollution<sup>1,36</sup>. Consistent with this, samples dominated by air from any of the three subcontinental regions had significantly higher levels of combustion-derived polycyclic aromatic compounds (PACs) (Fig. 2e and Supplementary Data 4—Identifications GC–NPOC), in particular for air originating in the Indo-Gangetic Plain, a global hotspot for air pollution during the dry winter monsoon<sup>4,13</sup>.

**Molecular annotation and identification.** Based on spectral library searches, a total of 318 features (across both GC- and LC-HRMS analyses) were highlighted as putative anthropogenic or



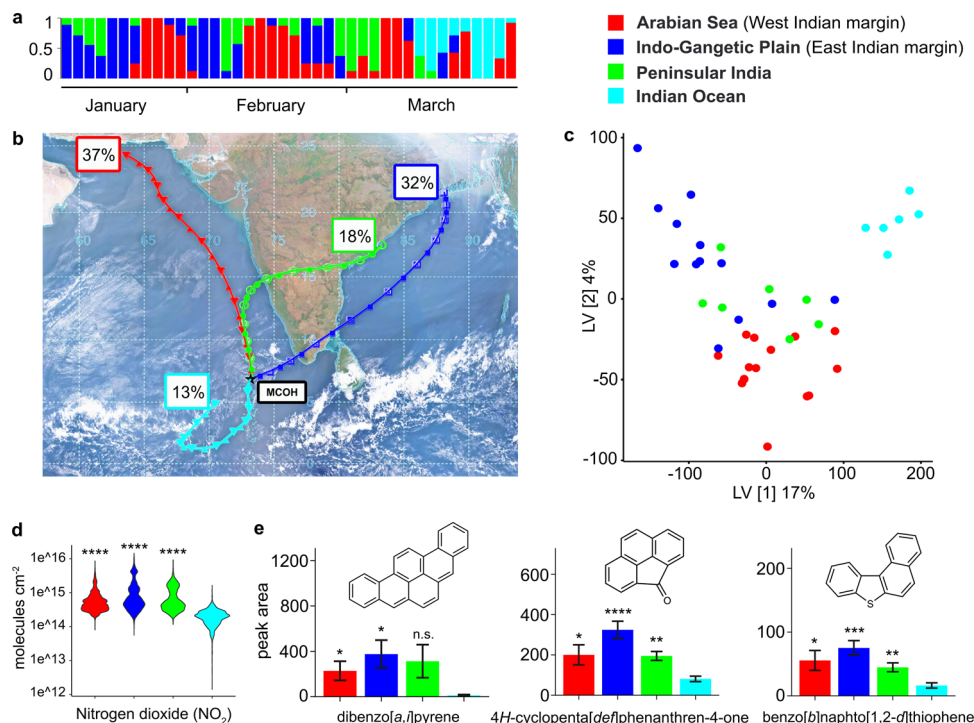
**Fig. 1** Sampling and comprehensive nontarget HRMS analysis of  $PM_{2.5}$ . **a** High-volume sampling ( $500\text{ L min}^{-1}$ ) of  $PM_{2.5}$  at MCOH onto quartz-fiber filters that were sectioned and extracted by three distinct protocols. **b** The nonpolar organic compound (NPOC) extract was analyzed with GC-HRMS by electron ionization (EI) and negative chemical ionization (NCI), shown here plotted by base peak  $m/z$  and Kovats retention index (RI). **c** Water-soluble organic compound (WSOC) extract and **d** polar organic compound (POC) extract were analyzed with LC-HRMS in data-independent  $MS^2$  acquisition (DIA) with electrospray positive (ESI+) and negative (ESI-) mode, shown here as Kendrick plots with color shading by retention time. **e**, **f** Venn diagrams showing the percent of molecular features that were unique to each extract and mode of analysis, demonstrating that a battery of approaches was important to achieve comprehensive molecular analysis of  $PM_{2.5}$ .

biogenic compounds, up to Level 2a identification<sup>37</sup> (see criteria in Methods). Annotations of putative anthropogenic substances were selected for confirmation by comparing orthogonal evidence under identical analytical conditions (Rt,  $MS^1$  and  $MS^2$ ) to reference standards (Supplementary methods). Across GC and LC datasets, 89 compounds were ultimately confirmed with highest confidence (Level 1; 53 compounds with Rt shift < 0.2 min or RI < 30) or as closely related isomers (36 compounds with Rt shift < 0.4 min or RI < 250) (Supplementary data 4—‘Identifications’, and Supplementary data 6—‘Spectral matches’). For the NPOC extracts, these consisted mostly of n-alkanes and PACs detected by GC-EI-HRMS (Fig. 3), including four oxy-PACs (e.g. 4*H*-cyclopenta[*def*]phenanthren-4-one ( $C_{15}H_8O$ ), 9-anthracenecarboxaldehyde ( $C_{15}H_{10}O$ ), and 7*H*-benz[*de*]anthracene-7-one ( $C_{17}H_{10}O$ )), six sulfur-containing PACs (i.e.  $C_{16}H_{10}S$ ,  $C_{18}H_{12}S$ , and  $C_{20}H_{12}S$  isomers), and a benzocarbazole isomer ( $C_{16}H_{11}N$ ). Several persistent organic pollutants were detected in the same extracts using GC-NCI-HRMS, including 11 chlorinated compounds e.g., polychlorinated dioxins, and polychlorinated biphenyls (PCBs), and seven brominated

flame retardants (BDEs) (Fig. 3). Compounds confirmed by LC-HRMS encompassed a wider variety of chemical classes, including parent commercial substances such as tris-2-butoxyethyl-phosphate ( $C_{18}H_{39}O_7P$ ) (Fig. 3), degradation products of commercial substances, or products of combustion and/or atmospheric oxidation, such as monoethyl phthalate ( $C_{10}H_{10}O_4$ ), phthalic acid ( $C_8H_6O_4$ ), 4-nitrophenol ( $C_6H_5NO_3$ ), 2,4-dinitrophenol ( $C_6H_4N_2O_5$ ), benzimidazole ( $C_7H_6N_2$ ), and 2-hydroxybenzimidazole ( $C_7H_6N_2O$ ) (Fig. 3). Various herbicides and insecticides were also confirmed (Fig. 3), e.g., DEET ( $C_{12}H_{17}NO$ ), prometon ( $C_{10}H_{19}N_5O$ ), malaoxon ( $C_{10}H_{19}O_7PS$ ), methamidophos ( $C_2H_8NO_2PS$ ), as well as their environmental transformation products, e.g. simazine-2-hydroxy ( $C_7H_{13}N_5O$ ) and atrazine-2-hydroxy ( $C_8H_{15}N_5O$ )<sup>37</sup>.

Many of the legacy persistent organic pollutants are semi-volatile and partition to the gas-phase, particularly at high ambient temperatures as recorded in the current campaign (mean =  $32.9^\circ\text{C}$ ), but several PCBs and BDEs were nevertheless detected sporadically from polluted back-trajectories, suggesting continued emissions in South Asia, with highest



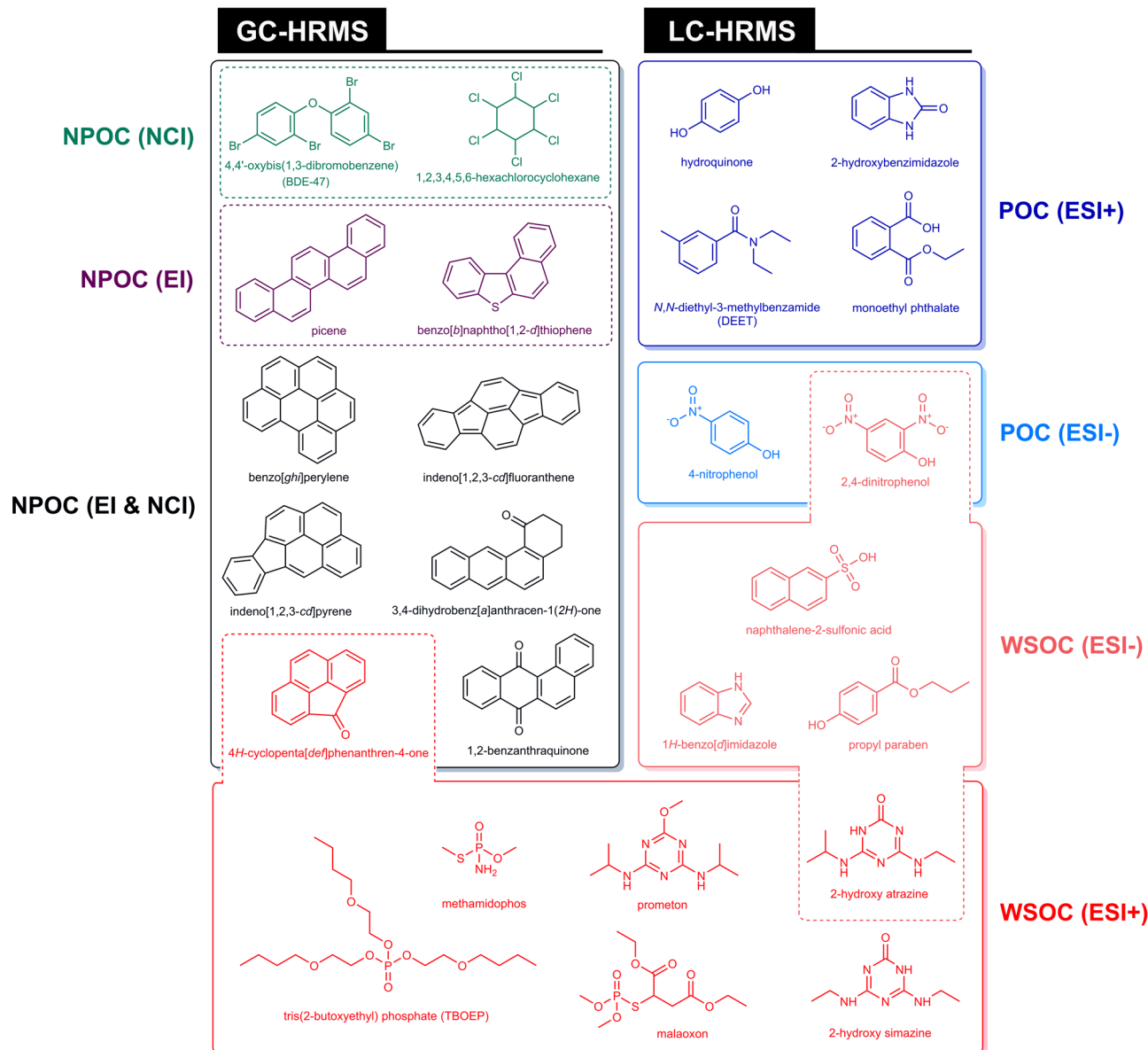


**Fig. 2 Geographical sources of PM<sub>2.5</sub> and modelling of molecular profiles.** **a** Relative contribution of air masses for each 48 hr sample ( $n = 40$  time-points) originating from the Arabian Sea, Peninsular India, Indo-Gangetic Plain or Indian Ocean, and **b** the mean path of these four 10-day back-trajectory clusters around the receptor site MCOH (6.80°N, 73.20°E). The satellite image corresponds to the monitoring campaign day 2nd of February, 2018 (NASA Worldview; <https://worldview.earthdata.nasa.gov/>). **c** Scores of the orthogonal partial least square (OPLS; 2 + 1 + 0) multivariate model correlating the back-trajectories to the nontarget (LC+GC) HRMS chemical profiles of PM<sub>2.5</sub> samples, with Indian Ocean back-trajectory cluster cross-validated [CV]-ANOVA  $p$  value = 0.0013. See also principal component analysis (PCA) in Fig. S10, and details in Supplementary data 3—Models and statistics OPLS 2 + 1). Each sample is colored according to the most dominant back-trajectory within the 48 h window. **d** Tropospheric nitrogen dioxide (NO<sub>2</sub>) levels measured along the back-trajectories by satellite remote sensing (OMI/Aura; NASA). **e** Polycyclic aromatic compound abundance ( $\pm$ SE) in GC-HRMS for analytes identified at Level 1 (dibenzo[*a,i*]pyrene, 4*H*-cyclopenta[*def*]phenanthren-4-one, benzo[*b*]naphto[1,2-*d*]thiophene). Error bars indicate S.E. Significant differences relative of levels in Arabian Sea ( $n = 14$ ), Peninsular India ( $n = 13$ ), and Indo-Gangetic Plain ( $n = 7$ ) to the Indian Ocean ( $n = 6$ ) are shown by the Welch  $t$ -test (two-sided): \* $p \leq 0.05$ ; \*\* $p \leq 0.01$ ; \*\*\* $p \leq 0.001$ ; and \*\*\*\* $p \leq 0.0001$  (d.f. = 18, 17, 11).

detection frequency for 2,2',4,4'-tetrabromodiphenyl ether (BDE-47; C<sub>12</sub>H<sub>6</sub>Br<sub>4</sub>O) (Fig. 3 and Fig. S11). Samples with subcontinental back-trajectories, particularly those associated with the Arabian Sea and Indo-Gangetic Plain, consistently had higher levels of PACs, plasticizers, biocides, and herbicides; at least 2- to 10-fold higher than in air masses from the Southern Indian Ocean (Fig. 2e, and Fig. S11). Simazine-2-hydroxy (C<sub>7</sub>H<sub>13</sub>N<sub>5</sub>O) (Fig. 3) was detected at highest levels in two samples from the Indian Ocean, suggesting local use of simazine in the Maldives. Conversely, atrazine-2-hydroxy (C<sub>8</sub>H<sub>15</sub>N<sub>5</sub>O) was detected at higher levels (up to 10-fold) in samples from subcontinental regions (Fig. 3, Fig. S11 and S16–S17), and we are not aware of any previous reports of this substance in ambient air.

**Characterization of PM<sub>2.5</sub> molecular dark matter.** Only a minor proportion (0.5%) of all molecular features in polluted PM<sub>2.5</sub> were identified or putatively annotated by MS<sup>2</sup> spectral database matching (see criteria in Methods). Higher but still low annotation rates (up to 1–2%) have been reported for environmental water analysis<sup>38,39</sup> and metabolomics<sup>24</sup> despite larger and specialized databases. The vast majority of molecules in polluted air could not be matched to known compounds, not only because spectral databases mostly cover biogenic compounds (e.g., anthropogenic substances account for approximately 15% of records in both NIST20 and MassBankEU), but also because the

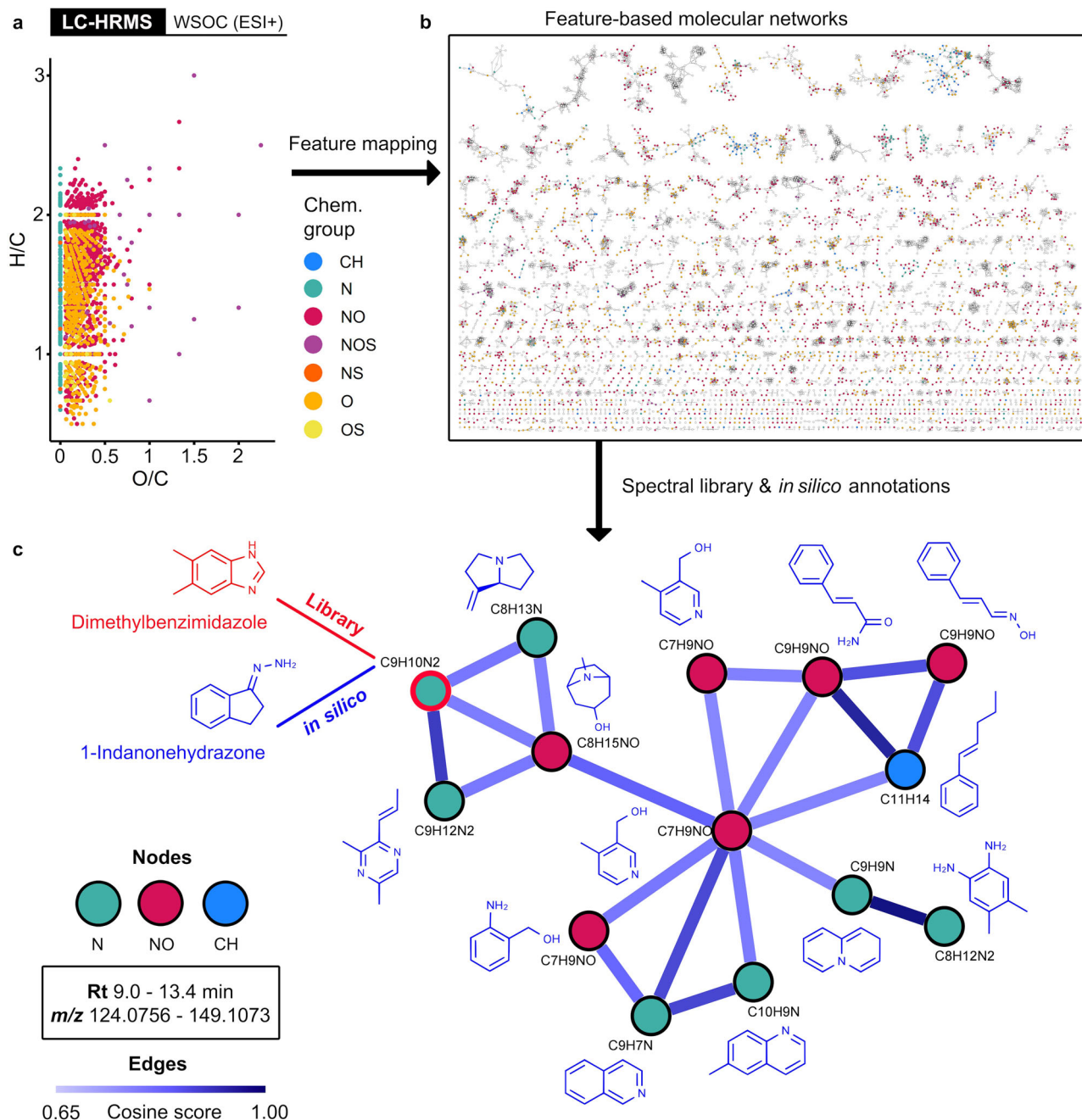
major sources of air pollution include the incomplete combustion of complex fuels and heterogeneous biomass that are prone to rapid transformations in the atmosphere (oxidation, photolysis, hydrolysis) to yield molecular byproducts and secondary organic aerosols<sup>40</sup>. Hence, we combined cheminformatics and computational strategies that leverage data-rich MS<sup>1</sup> and MS<sup>2</sup> information of unknown features, and structural information of annotated molecules, to characterize the remaining unknown features in PM<sub>2.5</sub> through in silico predictions of structure and physicochemical properties<sup>38,39</sup>. For all LC-HRMS features, we first calculated molecular formulae by combining two independent approaches, MFAssignR<sup>41</sup> and SIRIUS<sup>42</sup>. Consensus between the two methods resulted in >19,000 features being assigned a molecular formula (i.e., 33% of the LC-HRMS datasets, Level 4 identification)<sup>37</sup> (Fig. 4a). Next, features were clustered by molecular networks constructed in GNPS (Global Natural Products Social Molecular Networking)<sup>25,26</sup>, whereby neighboring molecular features (nodes) are linked by pairwise MS<sup>2</sup> spectral similarity (edges) representing an inferred structural analogy (Fig. 4b, c). We then performed a high-throughput structural elucidation of all molecules in the networks with the MS<sup>2</sup>-guided in silico Network Annotation Propagation (NAP) GNPS workflow<sup>31</sup>. This workflow leverages the molecular network topology to re-rank the in silico predicted candidates based on joint similarity within a molecular family cluster (e.g., ten first-neighbors), and by combining in silico predictions via the core algorithm MetFrag<sup>43</sup> to structural information from MS<sup>2</sup>



**Fig. 3 Anthropogenic substances confirmed in PM<sub>2.5</sub>.** Examples of environmental contaminants detected across different extracts and modes of analyses, all of which were confirmed by comparison with authentic standards (i.e., identification Level-1). Dashed-line intersections indicate example analytes detected in multiple extracts or modes of analysis. Nonpolar hydrocarbons, including PAHs and oxy-PAHs, were detected in the NPOC extracts by GC-HRMS with electron ionization (EI) and/or by negative chemical ionization (NCI). Sulfur-containing PACs were detected and confirmed by GC-EI, while other persistent pollutants (e.g., PCBs, BDEs) by GC-NCI. The WSOC and POC fractions were analyzed by LC-HRMS with each extract injected twice for separate acquisitions in electrospray positive (ESI+) and negative (ESI-) modes. ESI- is optimal for weak organic acids, while ESI+ reveals nitrogenous bases and a range of polar neutral molecules such as alcohols, aldehydes, and ketones. These analyses combined allowed the detection of anthropogenic pollutants representing a wide variety of chemical classes, some potentially originating from atmospheric oxidation of incomplete combustion sources, others synthesized for commercial or industrial use, e.g., plasticizers, generic biocides, insecticides, herbicides, and their transformation products.

spectral library matches. The structures of 30,389 molecules were predicted in this way to achieve Level 3 identification<sup>37</sup> (Supplementary data 4—Identifications NAP). While the NAP workflow already increases the reliability of the *in silico* first-candidates<sup>31</sup>, we further proceeded to consider only those predicted structures for which a matching molecular formula had been consistently assigned by all three computational steps (i.e., MFAssignR, SIRIUS, and NAP). By this conservative approach, 10,256 structures (out of 30,389 initial predictions; 34%) were carried forward as a relatively reliable *in silico* portfolio of small molecules in PM<sub>2.5</sub> (Fig. 4b, c and Supplementary data 4—NAP + formula\_consensus).

**Molecular hallmarks of polluted and clean air.** After formula assignment and thousands of structural predictions, the contrasting molecular profiles across back-trajectory regions (Fig. 2) presented an opportunity to investigate what types of organic molecules are most characteristic of clean and polluted air. Thus, in a second supervised multivariate model, we collapsed the four back-trajectory matrices into one vector expressing each sample's polluted fraction (i.e., air originating from any of the three polluted subcontinental trajectories) versus clean air of the Southern Indian Ocean (Fig. 5a). The fraction of molecules that most significantly correlated with polluted or clean air back-trajectories were selected for further investigation (Fig. 5b). These top-VIP

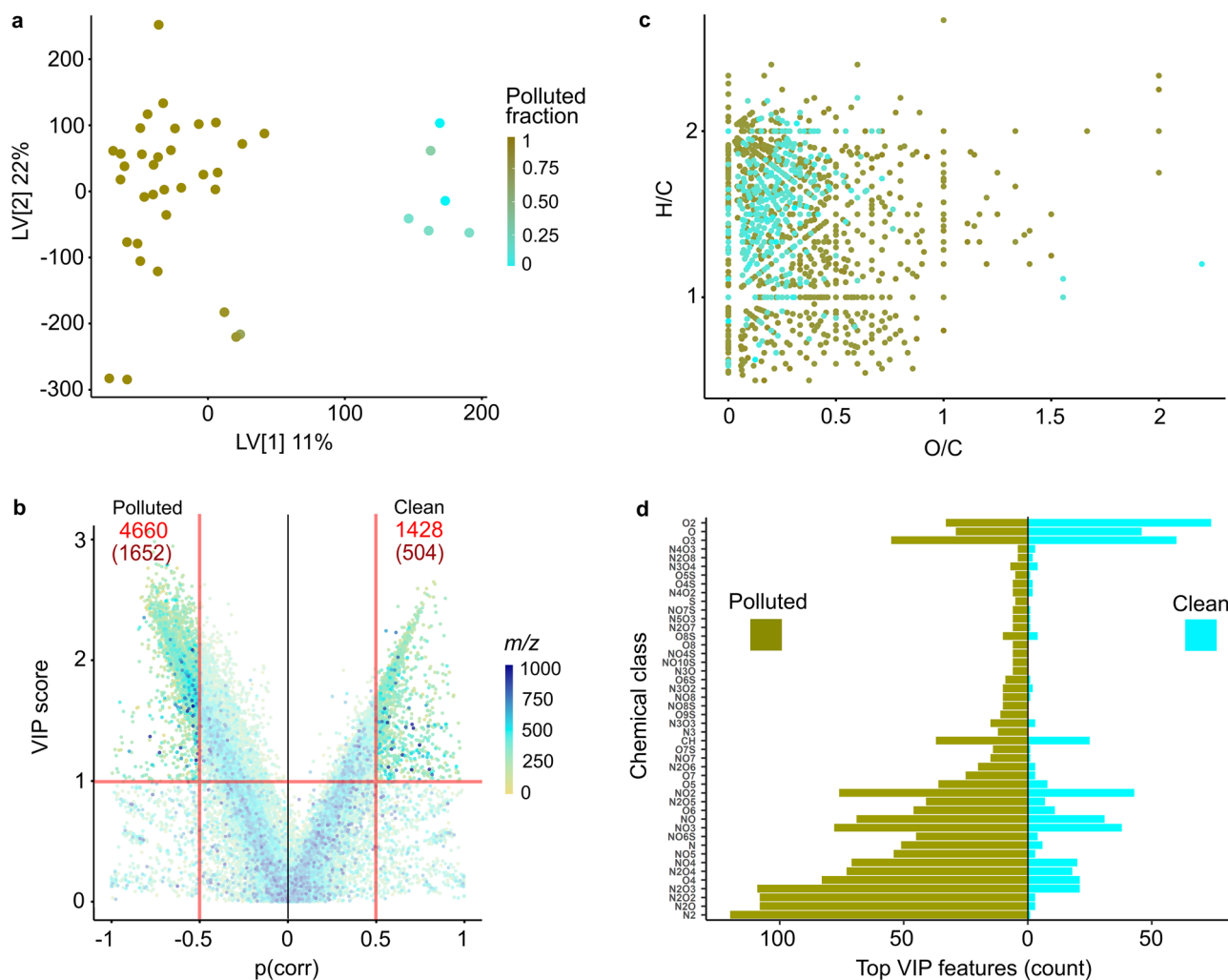


**Fig. 4 Molecular-level characterization of complex PM<sub>2.5</sub> extracts.** High-throughput characterization of PM<sub>2.5</sub> molecules was performed by combining information from MS<sup>1</sup> and MS<sup>2</sup> HRMS data. The integration of multiple cheminformatics approaches is illustrated here for molecular features detected in the WSOC extracts by LC-HRMS with ESI+ (see also Fig. S6). **a** Molecular formulae were assigned to a proportion of features (45% in this example, see Fig. S5) allowing color-coded visualization in van Krevelen diagrams, and in **b** molecular networks (GNPS). In the network, each feature is shown as a node linked to other nodes by edges indicating the degree of similarity among deconvoluted MS<sup>2</sup> spectra (minimum cosine score = 0.65). In this example, 10,051 molecular features are clustered into 1064 molecular families having inferred structural analogy. **c** Zoom-in on a molecular family cluster of 15 nitrogen-containing benzenoids. The first-candidate structures from re-ranked *in silico* predictions (NAP/MetFrag) are shown in blue. The highlighted node (red outline) shows two putative annotations for the same molecular formula (C<sub>9</sub>H<sub>10</sub>N<sub>2</sub>), i.e., 5,6-dimethylbenzimidazole (red structure) from the GNPS library match (Fig. S18), and 1-indanonehydrazone (blue structure) as the top *in silico* first-candidate ranked by the network consensus (Fig. S20). The structure of 5,6-dimethylbenzimidazole was also predicted *in silico*, but ranked as the sixth candidate.

(variable importance for the projection) features accounted for 10% of the dataset (6088), 35% (2156) of which had been successfully assigned a molecular formula, and 17% (1049) of which had been assigned a structure consistent with the formula (Supplementary data 4—Identifications).

Overall, molecules in PM<sub>2.5</sub> that correlated with polluted back-trajectories (4660 features; 1652 formulae; 775 structures) were

three times more numerous than those correlating with clean air (1428 features; 504 formulae; 274 structures) (Fig. 5b), and occupied a broader and more oxidized chemical space (median; polluted = O/C 0.29 ± 0.28 SD, H/C 1.43 ± 0.41 S.D.; clean = O/C 0.20 ± 0.18 S.D., H/C 1.55 ± 0.30 S.D.) (Fig. 5c). The most numerous heteroatomic formula classes in polluted air corresponded to molecules containing one or two nitrogen atoms (N<sub>2</sub>,



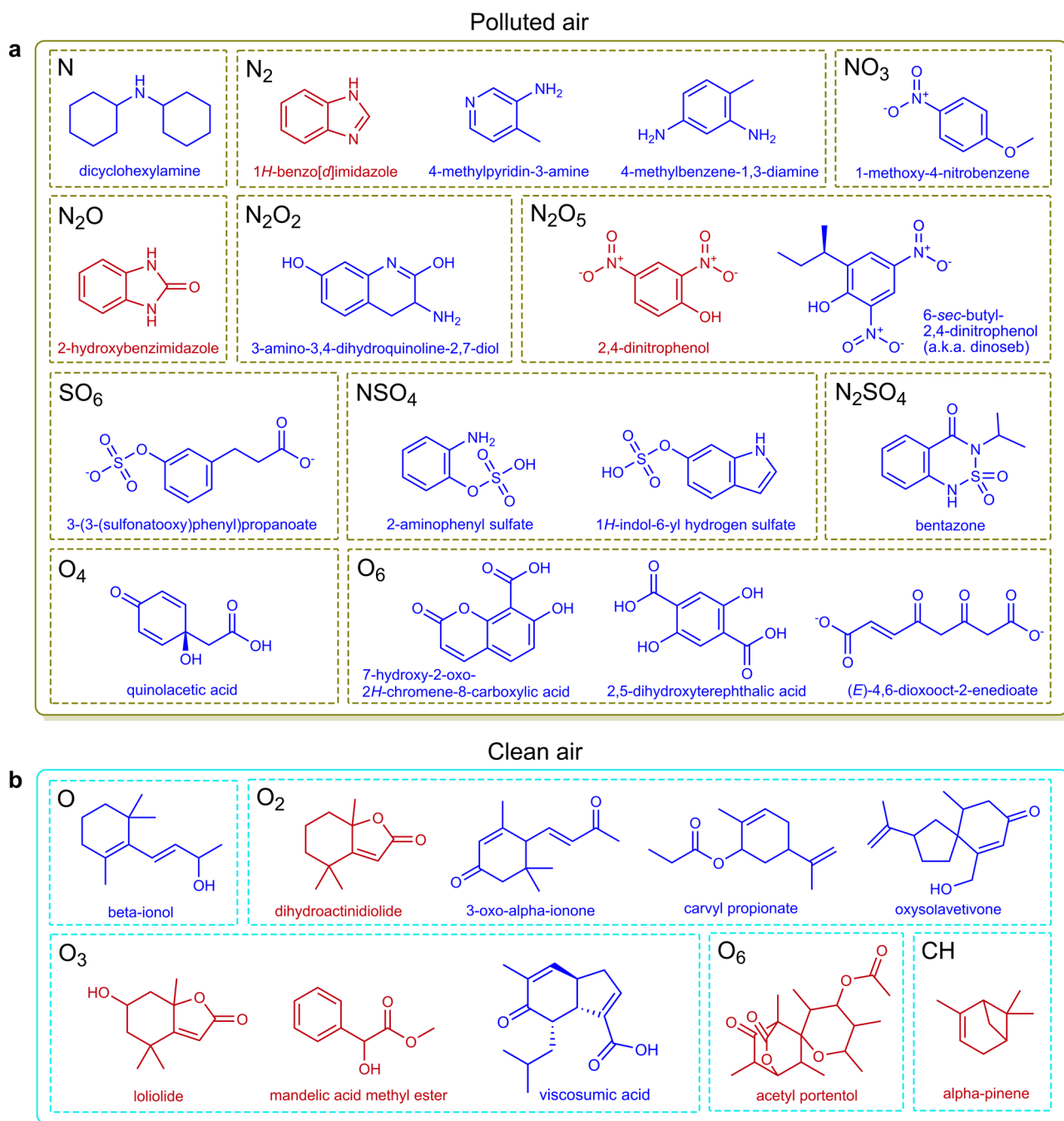
**Fig. 5 Molecular markers of polluted and clean PM<sub>2.5</sub>.** **a** Multivariate OPLS (1 + 1 + 0) model explaining chemical profiles of polluted (subcontinental) and clean air (Indian Ocean). Samples are clustered based on similarity of molecular feature profiles, and colored according to the relative contribution of polluted (brown) or clean (cyan) air back-trajectory frequencies in each sample (CV-ANOVA  $p$  value =  $9e-06$ ; see details in Supplementary data 3 - Models and statistics OPLS 1 + 1). **b** From the OPLS model, the contribution of each molecular feature ( $n = 60,030$ ) to either polluted or clean air profiles is highlighted in the volcano plot, showing molecules with the highest model-correlation coefficients ( $|p[\text{corr}]| > 0.50$ ) and variable importance for the projection scores ( $\text{VIP} > 1.0$ ). Features are colored by their  $m/z$  values. The total number of top-VIP features (i.e.,  $\text{VIPs} > 1.00$ ,  $|p[\text{corr}]| > 0.50$ ) as significant markers of polluted and clean air are shown (red font) with the proportion of unambiguous molecular formulae assigned (in brackets). **c** Distribution of these top -VIP features in van Krevelen space, colored by their model-correlation with either polluted (brown) or clean (cyan) air. **d** Total counts of top-VIP features by heteroatomic formula classes (minimum of five features per class) ordered by difference between polluted and clean air. See also Fig. S12 for the relative contribution of each extract (NPOC, POC, WSOC) and data sources in Supplementary data 3 - Models and statistics, Top VIP  $x$  corr).

$\text{N}_2\text{O}_x$  classes, Fig. 5d), and were represented by in silico predicted structures of e.g., nitrophenols,  $N$ -heterocycles, imidazoles, quinazolines, and diazine derivatives (Fig. 6a). These are compound classes previously highlighted as contributors to the light-absorption of BrC in atmospheric aerosols<sup>15</sup>. Polluted air also included relatively more molecules containing sulfur (S and  $\text{O}_x\text{S}$  classes) or mixed sulfur and nitrogen ( $\text{NO}_x\text{S}$ ), e.g., the predicted structure of the herbicide bentazone ( $\text{C}_{10}\text{H}_{12}\text{N}_2\text{O}_3\text{S}$ ) (Fig. 6a), and highly oxygenated compounds ( $\text{O}_4\text{O}_8$  classes) (Fig. 5d) such as quinolacetic acid ( $\text{C}_8\text{H}_8\text{O}_4$ ) and dihydroxyterephthalic acid ( $\text{C}_8\text{H}_6\text{O}_6$ ) (Fig. 6). The  $\text{O}_x\text{S}$  class in polluted samples also included organosulfates (Fig. 6a) which are implicated in cloud condensation processes<sup>44</sup>.

In contrast, clean air PM<sub>2.5</sub> was distinguished by mono-, di-, and tri-oxygenated molecules ( $\text{O}-\text{O}_3$  classes), of which many were

annotated through the GNPS spectral library (Level 2a) or predicted in silico (Level 3) as derivatives of alpha-pinene ( $\text{C}_{10}\text{H}_{16}$ ), such as dihydroactinidiolide ( $\text{C}_{11}\text{H}_{16}\text{O}_2$ ) and loliolide ( $\text{C}_{11}\text{H}_{16}\text{O}_3$ ), or other biogenic volatiles, e.g., mandelic acid-methyl ester ( $\text{C}_9\text{H}_{10}\text{O}_3$ ), and viscosomic acid ( $\text{C}_{15}\text{H}_{20}\text{O}_3$ ), a sesquiterpene produced by *Polygonum sp.*<sup>45</sup>, native to South East Asia (Fig. 6b and Fig. S11). An interesting natural product with higher oxygen content was identified (Level 1) as acetyl portentol ( $\text{C}_{19}\text{H}_{28}\text{O}_6$ ; Fig. 6b and Fig. S21), a polyketide produced by marine lichens of *Rocella sp.* native to Indian coastal habitats<sup>46,47</sup>. Biogenic volatiles, such as (mono)terpenes, that are photochemically oxidized in the atmosphere (including to  $\text{O}-\text{O}_3$  class substances) can contribute to particle nucleation in the absence of pollution<sup>48</sup>, and are chromophoric components of secondary organic aerosols, for example, oxidized indole derivatives<sup>49</sup>.



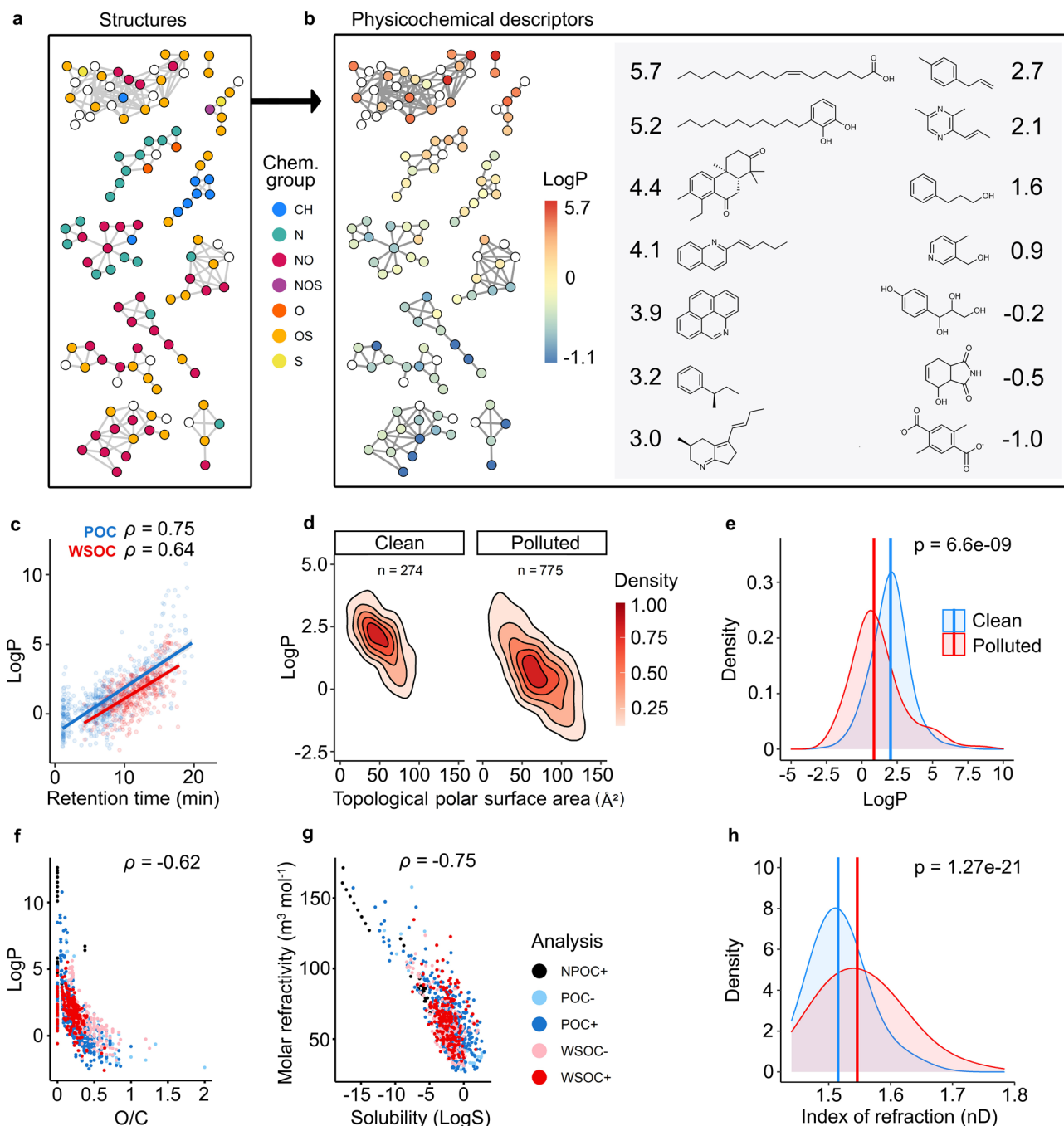


**Fig. 6 Representative molecular structures among the major heteroatomic formula classes in polluted and clean air.** Molecular markers in polluted and clean air (OPLS; Fig. 5) were structurally characterized by MS<sup>2</sup> spectral library matching (red structures; Level 2) or by MS<sup>2</sup> in silico prediction (blue structures; Level 3). Representative example structures are shown to illustrate the results (see also Supplementary data 4 - Identifications). **a** Polluted air was distinguished by a mixture of nitrogen- and sulfur-containing compounds, and highly oxygenated organic molecules. Several of these structures corresponded to aromatic *N*-heterocycles (e.g. nitrophenols, imidazoles, and diazine derivatives), which are structures that have been previously linked to light-absorbing BrC. The three structures in red from polluted air (N<sub>2</sub>, N<sub>2</sub>O, N<sub>2</sub>O<sub>5</sub>) were confirmed Level 1 by authentic reference standards. **b** Molecular hallmarks of clean air were mono-, di-, and tri-oxygenated organic molecules (O-O<sub>3</sub> class), many of which were predicted in silico (Level 3) or annotated on the GNPS spectral library (Level 2) as biogenic compounds. These included terpenes, such as derivatives of alpha-pinene (CH) and other plant volatiles. An interesting exception of a compound in clean air with higher oxygen content (O<sub>6</sub>) was confirmed as acetyl portentol (Level 1; Fig. S21), a polyketide produced by the marine lichens of *Roccella* sp. native to Indian coastal habitats. Another biogenic compound was predicted in silico (Level 3) as viscosomic acid, a sesquiterpene produced by the annual herbs of *Polygonum* sp. native to Nepal, Bangladesh, and north-eastern India.

**Physicochemical properties of molecules in clean and polluted air.** To gain molecular insights to the impacts of PM<sub>2.5</sub> on human health and climate, we next employed the structures of top-VIP molecular markers to estimate physicochemical descriptors of toxicological and environmental relevance, e.g., lipophilicity

(logP), topological polar surface area (TPSA), water solubility (logS), and molar refractivity (MR). A concordance between physicochemical properties and clusters of molecular families in the molecular networks was observed, whereby closely related structures showed similar values of e.g., logP (Fig. 7a, b).





**Fig. 7 Analysis of physicochemical properties of clean and polluted air molecules.** Structures of molecules in clean and polluted air (i.e. VIPs > 1.00,  $p[\text{corr}] > \pm 0.50$ ; Fig. 5) were used to compute physicochemical descriptors of toxicological and environmental relevance. **a, b** A subset of the WSOC+ molecular network (see Fig. 4) is shown to highlight the tight relationship between structural analogy and physicochemical properties, with features color-coded according to **(a)** chemical group, and **(b)** predicted lipophilicity (logP). In **b**, example structures from the same network are reported with corresponding logP values. **c** Positive Pearson's correlation between the predicted logP and the liquid chromatography retention time (Rt) on a C18 stationary phase for POC and WSOC extracts (ESI+ and ESI- data combined). **d** Two-dimensional density plots visualizing the distribution of logP and topological polar surface area (TPSA, Å<sup>2</sup>) in clean and polluted air. **e** Density distribution and median values of logP in clean and polluted air molecules. **f, g** Scatterplots with molecular features color-coded according to detection in the analyzed extracts, showing a negative Pearson's correlation between the **(f)** LogP and O/C, and **(g)** between MR and solubility (logS). **h** Density distribution and median values of the index of refraction (nD) in clean and polluted air molecules. All Pearson's correlation coefficients ( $\rho$ ) are shown with  $p$  values < 2.2e-16. Significant differences between clean and polluted air profile distributions are reported for unpaired Student's  $t$ -test  $p$ -values (two-sided).

Moreover, for features detected by LC-HRMS (WSOC, POC), the computed logP of predicted molecular structures were strongly correlated with empirical measurements of their hydrophobicity: the reversed-phase HPLC retention times in our analyses

(Pearson's correlation,  $\rho = 0.64$ – $0.75$ , Fig. 7c). These important results demonstrate the reliability of the molecular networking approach and of the *in silico* predicted structures (Supplementary data 4 – Identifications, Descriptors).

Compared to clean air, molecules in polluted air occupied a broader physicochemical space, including for logP and TPSA (Fig. 7d) which influence bioavailability and tendency to cross biological membranes. Molecules in polluted air had lower median logP (polluted = 0.85 logP; clean = 2.02 logP;  $p$  value =  $6.6e-09$ ) (Fig. 7d), but the distribution was bimodal and polluted air also had a higher frequency of structures with extreme lipophilic values ( $> 5$  logP, polluted 10.4%; clean = 5.8%) (Fig. 7e). The most highly lipophilic substances included  $n$ -alkanes and PACs detected by GC-HRMS in NPOC extracts (confirmed, Level 1), and related alcohols, aldehydes, fatty acids, and amides predicted in silico (Level 3) by LC-HRMS. The latter substances were often detected by LC-ESI+ in POC extracts, e.g. 8-dotriacontenoic acid ( $C_{32}H_{62}O_2$ ), 22-oxononacosanoic acid ( $C_{29}H_{56}O_3$ ), and docosanamide ( $C_{22}H_{45}NO$ ), but also in other fractions and ionization modes, e.g. tricosanoylglycine ( $C_{25}H_{49}NO_3$ ) and dimethyl octadecanedioate ( $C_{20}H_{38}O_4$ ) detected by LC-ESI- in POC and WSOC, respectively.

The overall trend of lower logP (i.e. increased polarity) among molecules from polluted regions of the Indian subcontinent (Fig. 7d, e) is consistent with photochemical oxidation of water-soluble BrC during transport from the Indo-Gangetic Plain<sup>11</sup>. We reported a higher O/C among molecules in polluted air (Fig. 5c, d), and here we further observed a strong inverse correlation between logP and molecular content of oxygen ( $\rho = -0.62$ ; Fig. 7f). As an illustrative example of oxidation, fluoranthene ( $C_{16}H_{10}$ ; logP = 4.53) was among the PAHs detected by GC-EI-HRMS (Level 1), while by LC-HRMS (ESI-, WSOC) our in silico workflow predicted the hydroxy-PAH 9H-fluoren-9-ol ( $C_{13}H_{10}O$ ; logP = 2.52, Level 3). The bulk of relatively polar substances revealed by LC-HRMS at this receptor site may constitute secondary organic aerosols derived from atmospheric processing and photooxidation of anthropogenic and biogenic precursors (Fig. 6). With the current approach, these distributions could be examined along spatial transects from source to receptor regions in future.

**Significance to human health and global climate.** In silico workflows may be useful for future research into the health impacts of PM<sub>2.5</sub> exposure. While PAHs are lipophilic carcinogenic molecules<sup>50</sup>, semi-polar byproducts of their atmospheric processing span a wider range of physicochemical properties and can be more acutely toxic. Here, a quantitative structure-activity relationship analysis based on descriptors used to predict human absorption, distribution, metabolism, and excretion (e.g. MW, logP, logS, TPSA; Fig. 7c, d)<sup>33</sup> revealed that more than a third of the top-VIP molecular markers in polluted air (35%, 272/774 structures) had high predisposition for gastro-intestinal absorption and permeation through the human blood-brain barrier, thus representing potential gut inflammatory<sup>51</sup> and neurotoxic<sup>52</sup> components of PM<sub>2.5</sub>. Of these, the majority (147 molecules) were also predicted to be inhibitors of cytochrome P450 enzymes, with representative structures highlighted in Fig. 8. Several of these were confirmed (Level 1) as sulfur-containing PACs (e.g. benzo[*b*]naphtho[1,2-*d*]thiophene), oxy-PAHs (e.g. benzanthrone and 4H-cyclopenta[*def*]phenanthren-4-one) and imidazoles (e.g. benzimidazole and 2-hydroxybenzimidazole), while other in silico predicted *N*-heterocycles included azoles and azaarenes, such as 11H-Indeno[1,2-*b*]quinoline ( $C_{16}H_{11}N$ ), non-ylpyrazole ( $C_{12}H_{22}N_2$ ), and 4-azapyrene ( $C_{15}H_9N$ ) (Fig. 8) (Supplementary data 4 – Identifications, Descriptors).

From a climate perspective, many top-VIP molecular markers of polluted air confirmed here (Level 1), such as PACs (detected by GC-HRMS), and imidazoles and nitrophenols (detected by LC-HRMS) are known light-absorbing chromophores in BrC

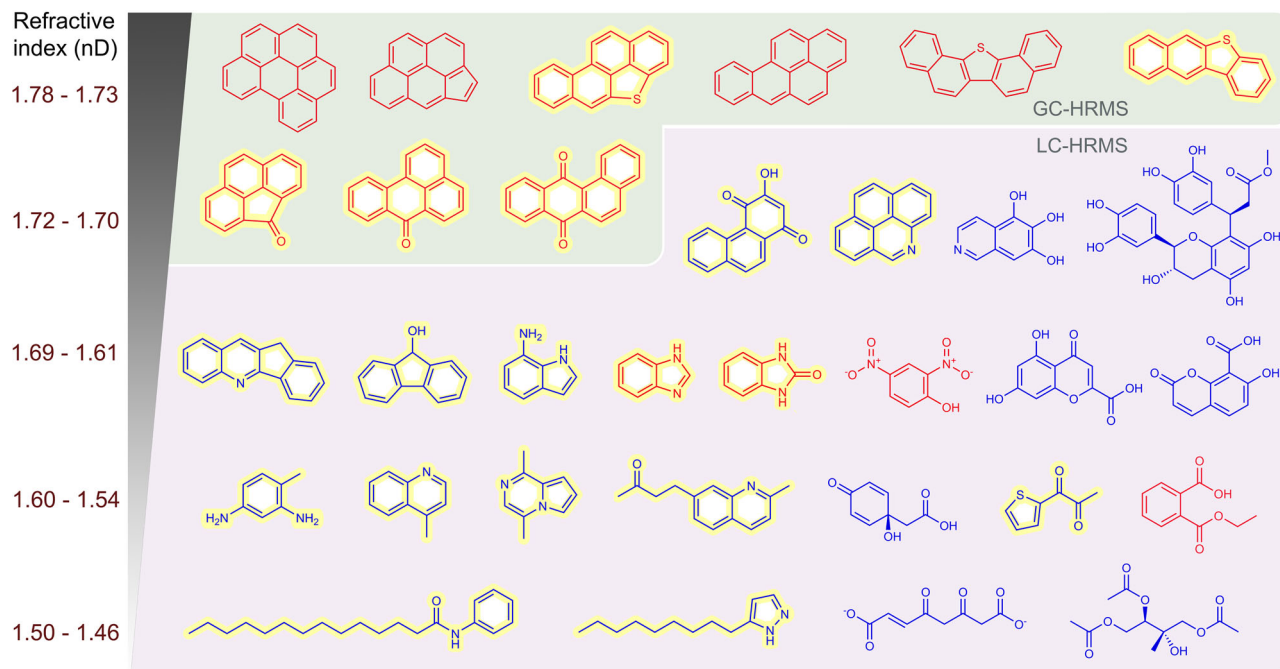
aerosols<sup>53</sup> (Fig. 8). Given that atmospheric aging rapidly alters the molecular components of such organic aerosols, leading to high uncertainty of their optical properties<sup>54</sup>, high-throughput workflows and in silico prediction of molecular structures could be exploited in future studies to gain further insight to the climate impact for complex mixtures of organic substances in PM<sub>2.5</sub>. For instance, molecules in polar and water-soluble (WSOC, POC) extracts of PM<sub>2.5</sub> at this receptor site had predicted water solubilities (logS) that negatively correlated with molar refractivity (MR,  $\rho = -0.75$ ; Fig. 7g); a measure of polarizability and tendency for molecules to interact with light (e.g. driving Rayleigh-scattering)<sup>55</sup>. A significant difference was evident in the optical properties of molecular hallmarks in clean and polluted air ( $p$  value =  $1.3e-21$ ) (Fig. 7h), highlighted by the predicted light-scattering capacity (i.e. expressed by the real part of the complex refractive index (nD))<sup>34,54</sup>. Molecules with high-scattering capacity ( $>1.55$  nD) were nearly three times more abundant in polluted air (polluted = 48.1%; clean = 16.9%). Lower-scattering molecules (1.44–1.55 nD) were abundant in clean air profiles, but some were associated with polluted air (e.g. oxy-PAHs, indoles, and derivatives of nitrobenzene, benzimidazole, and quinoline), whereby the lowest nD values were for small-molecules and byproducts with linear aliphatic structures (Fig. 8). This latter trend may indicate atmospheric processing and photooxidation in secondary organic aerosols<sup>49</sup>, as suggested by previous laboratory experiments<sup>54</sup>.

## Conclusions

Through a battery of comprehensive extractions and complementary LC- and GC- nontarget analyses of PM<sub>2.5</sub> from South Asia (720 analyses of 120 extracts from 40 air samples), we resolved and characterized greater molecular complexity of atmospheric aerosols than previously reported. In the environmental context, the intensive analytical workflow facilitated molecular discoveries and high-throughput characterization of thousands of unidentifiable substances across wide spatial scales of air mass origins throughout the continuous 3-month SAPOEX-18 campaign. The molecular complexity and relative profile of 60,030 molecular features varied by source region and pollution levels, and chemical class hallmarks of polluted and clean air were revealed at this receptor site in the Indian Ocean. The dominant nitrogenous organic molecules in polluted air are likely of relevance to health and climate but must be confirmed through further studies using approaches such as those described here.

As anticipated at this remote receptor site, only a small fraction of molecular structures could be confidently annotated by matching to spectral databases (Level 2) or identified with authentic standards (Level 1). We demonstrated that in silico predictions based on the underlying information-rich MS<sup>2</sup> spectra can be exploited for the high-throughput structural characterization of thousands of substances in atmospheric samples. High-throughput molecular structure prediction remains an imperfect tool, however, predictions herein were aided by the molecular network topology and validated by strong and statistically significant correlations between the structures' predicted physicochemical properties (i.e. logP) and our LC retention times.

Overall, the ranges of molecular formula classes, predicted structures, and properties (physicochemical and toxicological) were wider in samples from polluted regions than in pristine air from the Indian Ocean. The molecular profile of organic chemicals in PM<sub>2.5</sub> originating from polluted regions were more oxidized (higher O/C and lower H/C), reflecting atmospheric processing and secondary organic aerosol formation<sup>56,57</sup>.



**Fig. 8 Molecular hallmarks of polluted air with climate and human health relevance.** Example of structures and estimated toxicological and optical properties of organic compounds detected by GC- and LC-HRMS in  $PM_{2.5}$ , as molecular markers of polluted air. Structures were confirmed with authentic standards (in red; Level 1 or Level 2 as close isomers) or predicted *in silico* (in blue; Level 3). Molecules are ordered in rows by decreasing light-scattering capacity expressed as the real component of the complex refractive index (nD). Some of the resulting molecular byproducts include sulfur-, oxygen- and nitrogen-containing molecules (highlighted in yellow) with predicted toxicity to multiple endpoints, i.e. simultaneously CYP450 inhibitors and with high predisposition to human gastro-intestinal absorption and blood-brain barrier permeability.

Persistent anthropogenic compounds were also confirmed, and their ubiquitous observation at a relatively remote site 2500 km from the outflow of the Indo-Gangetic Plain is evidence of their long-range atmospheric transport potential<sup>58</sup>. Higher levels of many compounds, including various PACs (oxy-, nitro-, and sulfur-containing) in polluted air from the Indian subcontinent likely come from massive emissions of incomplete combustion processes, characteristic of the South Asian Atmospheric Brown Cloud, such as from biomass burning (household biofuel and burning of agricultural crop residue) and small-scale fossil fuel combustion (e.g., traffic, kerosene lamps, and diesel generators)<sup>59–61</sup>. Sources of other anthropogenic contaminants may include fugitive releases from industry or urban areas, such as plasticizers, resins, textile dyes, and flame retardants. Biocides, herbicides, and their metabolites (e.g. atrazine-2-hydroxy) may derive directly from agricultural spraying and soil erosion, or burning of post-harvest biomass residues in agricultural regions across the Indian subcontinent<sup>62</sup>. Atrazine-2-hydroxy is a major metabolite of atrazine in soils<sup>63</sup>, but has to our knowledge never been reported in air. Previous work at MCOH and in South Asia demonstrate that marine sources of organic carbon are minor contributors to PM and its water-soluble organics<sup>11,64</sup>. It is plausible that marine sources could be of greater relative importance for molecular signals detected here in  $PM_{2.5}$  from air masses coming from southern Indian Ocean. The sources and global relevance of contaminants emitted from these high-emission regions of South Asia deserves much more attention.

In addition to sources, the present approach opens up multiple avenues for deeper understanding of the atmospheric chemistry of aerosols at the molecular level, of central relevance to health and climate. Simultaneous identification or characterization of known and unknown emissions and transformation products could allow the following of coupled reaction pathways, which may also be linked to mesoscale chemical information, e.g. from

aerosol mass spectrometry<sup>14</sup>. PM contributes to some of the largest uncertainties in our current understanding of the climate, and thus molecular markers of photochemical aging or secondary formation (e.g., carboxylic acids or dicarbonyls) may be comprehensively tracked to better resolve complex photochemistry which is shown to attenuate light-absorption of climate warming BrC in the South Asian outflow<sup>11</sup>. Here, we observed several hallmark chromophores (e.g., nitro-phenols and PAHs), opening up possibilities for the broad-scale understanding of the molecular origins of light absorption beyond targeted analysis. Furthermore, the present work suggests links between the molecular composition and the real refractive index, a key to computing the scattering properties and overall climate cooling effects of organic aerosols. We also observed and semiquantified molecules with known strong impacts on cloud condensation, e.g. organosulfates, thus there is great potential to explore molecular-level connections to climate with present approaches.

From a health perspective, the wide variety of molecules discovered or described here for polluted South Asian air may contribute to mortality and physiological stress and disease<sup>65,66</sup>, including adverse birth outcomes<sup>67</sup>, asthma<sup>68</sup>, or even increased susceptibility to respiratory infections such as COVID-19<sup>69</sup>. The most toxic and bioavailable substances in polluted air have yet to be identified in toxicology or health studies, but may be confounded by toxicological interactions between primary emissions (e.g. lipophilic biocides and drug-like molecules) and secondary organics with diverse potentials to activate adverse outcome pathways. Comprehensive, detailed and high-throughput molecular analyses will be necessary to uncover these relationships. Altogether, these results highlight how nontarget analyses and *in silico* structure predictions can be implemented as advanced tools to explore deeper molecular-level insights and hypotheses on the health and climate impacts of complex organic compound mixtures in airborne PM.



## Methods

**High-volume PM<sub>2.5</sub> sampling.** PM<sub>2.5</sub> was sampled continuously in 48 h intervals between January 11th and April 4th, 2018, at MCOH (Hanimaadho, Haa Dhalu atoll, Maldives, 6.77 °N, 73.18 °E) onto pre-cleaned quartz fiber filters (150 mm Ø) using a high-volume sampler equipped with a PM<sub>2.5</sub> selective inlet (DH-77, Digital Elektronik AG, Volketswil, Switzerland) operating at 500 L min<sup>-1</sup>. To minimize sampling of local air, a wind-censored system interrupted the sampling when the wind was below 1.2 m s<sup>-1</sup> or coming from the southwest (180–270°)<sup>41</sup>. Field-blanks (*n* = 4) consisted of PM<sub>2.5</sub> filters placed in the air samplers with the pump turned off. Samples and field-blanks were stored frozen in pre-cleaned aluminum envelopes inside sealed bags, and shipped to Stockholm University for analysis.

**Sample preparation.** PM<sub>2.5</sub> samples were cut and extracted by three different protocols (Fig. 1a and Supplementary methods). An accelerated solvent extraction (ASE-350, Thermo Scientific Dionex ASE) was used with hexanes and toluene for nonpolar organic compounds (NPOCs), and with methanol and toluene for polar organic compounds (POCs), and extracts concentrated under nitrogen gas. Water-soluble organic compounds (WSOCs) were extracted by sonication in 40 mL HPLC grade water, followed by centrifugation (see Supplementary methods)<sup>41</sup>. Multiple isotope-labeled internal standards (Supplementary data 1 – Standards) were spiked to all PM<sub>2.5</sub> samples, field blanks, and urban dust reference samples (NIST SRM 1649b) prior to extraction. Sample preparation was performed in a positive pressure clean laboratory.

**GC- and LC-HRMS.** After silica cleanup, the NPOC extracts (2 µL injection) were analyzed with gas-chromatography (DB5 column) and HRMS (Q Exactive GC Orbitrap, Thermo Scientific) using electron ionization (EI) or negative chemical ionization (NCI) with full scan (44–700 *m/z*) and 60,000 resolution full-width half-maxima (FWHM) at 200 *m/z*. For POC and WSOC, extracts were filtered (0.45 and 0.2 µm, respectively) and analyzed with ultra-high-pressure liquid chromatography (UHPLC, Ultimate 3000) and HRMS (Q Exactive Orbitrap HF-X, Thermo Fisher Scientific) using electrospray ionization (ESI) in positive and negative mode. POC extracts (10 µL) were injected directly to the column (Waters Acquity UPLC BEH C18), while WSOC extracts (1000 µL) were injected to online solid-phase extraction prior to analytical separation. The mobile phases were 10 mM ammonium acetate in water (A) and methanol (B) and flow rate 0.4 mL/min (Supplementary methods). LC-HRMS was operated with alternating full scan (90–1000 *m/z*, 120,000 resolution FWHM at 200 *m/z*) and four MS<sup>2</sup> data-independent analysis (DIA) scans (30,000 FWHM) with variable *m/z* precursor windows.

**Data pre-processing.** GC- and LC-HRMS raw data were pre-processed using MS-DIAL (v4.24)<sup>70</sup>, allowing chromatographic alignment across all samples, basic data reduction (e.g. grouping of C<sub>13</sub> isotopes), spectral deconvolution, peak integration, and field-blank filtering (Supplementary data 1 – MS-DIAL parameters). All features were blank filtered in MS-DIAL based on a fivefold difference between sample maximum and the average in field blanks (*n* = 4). For semi-quantitative analysis, integrated peak areas from MS-DIAL were normalized using the areas of different isotope-labeled internal standards (Supplementary data 1 – Standards). All normalized feature areas were blank-subtracted by the average area of the corresponding feature detected in the field blanks (negative values were set to zero). Finally, the feature areas were normalized to the air volume accounted by the portion of PM<sub>2.5</sub>-filter extracted each sample.

**Molecular formula assignments.** The R Package ‘MFAssignR’<sup>41</sup> and the software SIRIUS (v4.5)<sup>42</sup> were used for molecular formula assignment (mass accuracy < 5 ppm). MFAssignR applies element heuristics on the MS<sup>1</sup>-level, then subtracts non-oxygen heteroatoms to solve for low-mass moieties (CHO), and finally assigns formula extensions via nested loops of homologous series<sup>41</sup>. SIRIUS similarly generates molecular formulae for the MS<sup>1</sup> and then leverages MS<sup>2</sup> fragmentation decision trees (i.e. shared neutral losses) to rank the candidates<sup>42</sup>. Consensus results were retained, corresponding to every unambiguous formula assigned by MFAssignR (Fig. S5) that matched the first-candidate assigned by SIRIUS, and later in the workflow by NAP (Supplementary data 4 – Identifications, NAP + formula\_consensus).

**Spectral library annotations.** For LC-HRMS (WSOC/POC; ESI + /ESI – modes), spectral library search was performed on the open-access platform GNPS (<http://gnps.ucsd.edu>) and third-party libraries (including MoNA, <https://mona.fiehnlab.ucdavis.edu/>; and MassBankEU; <https://massbank.eu/>) using a minimum of two shared MS<sup>2</sup> fragments (cosine ≥ 0.60) and later filtered for an MS<sup>1</sup> threshold of 5 ppm (See Fig. S15). For GC-HRMS (NPOC; EI/NCI modes), annotations were performed using a combination of high-throughput spectral library search (MSPepSearch; <https://chemdata.nist.gov/>) on the NIST20 and our in-house Orbitrap-HRMS library of environmental contaminants, and candidates were considered only for spectral match factors ≥ 700.

**Molecular networks and in silico structural elucidation.** Feature-based molecular networks<sup>25,26</sup> were built in GNPS (ver. 28.2) and visualized using Cytoscape

v.3.8.2. For LC-HRMS (WSOC and POC) datasets, parameters were: MS<sup>1</sup> and MS<sup>2</sup> tolerances of 0.02 Da, minimum spectra similarity cosines of ≥ 0.65, and a minimum of four shared spectral peaks. For in silico structural prediction with the GNPS/NAP workflow<sup>31</sup>, the following parameters were used: 10 first-neighbors, 5 ppm accuracy, cosine score ≥ 0.65, 10 maximum candidates from structural databases (GNPS, HMDB, SUPNAT, CHEBI), and Consensus + Fusion ranking algorithm. The above workflow was only partly applicable to GC-HRMS data. GC-EI molecular networks were built in GNPS (ver. 30)<sup>71</sup> using an ion tolerance of 0.4 Da, spectra similarity cosines ≥ 0.50, and a minimum of five shared spectral peaks, and were used to assist identifications (Fig. S7), together with formula assignments (MFAssignR), Kovats RI, Lee index, and GC-NCI data (Supplementary data 4 – Identifications ‘GC-NPOC’).

**Physicochemical properties.** Molecular formulae from the in silico predicted structures were translated using the open-source cheminformatics API OpenBabel<sup>32</sup> (<http://openbabel.org>). OpenBabel was also used to compute physicochemical descriptors and derive toxicological endpoints within the pharmacokinetics platform SwissADME<sup>33</sup> (<http://www.swissadme.ch/>). The real (i.e. light-scattering) component of the complex refractive index was computed in Python using the model developed by Bouteloup & Mathieu<sup>34</sup>.

**Back-trajectories and satellite measurements.** Ten-day back-trajectories were calculated every six hr using the HYSPLIT model (version 4) of the National Oceanic and Atmospheric Administration (NOAA), at 0:00, 06:00, 12:00, and 18:00 h GMT for 10 d into the past and 100 m height at MCOH (6.80°N, 73.20°E) (Fig. S8 and Supplementary data 3 – Back-trajectories). A model was selected with four mean back-trajectories using the clustering algorithm in HYSPLIT. Tropospheric NO<sub>2</sub> concentrations were averaged over the period of the campaign and prior 10 days of the first back-trajectory in 0.25° resolution using the Giovanni web application (<https://giovanni.gsfc.nasa.gov/giovanni/>) to access the National Aeronautics and Space Administration (NASA) OMI/Aura NO<sub>2</sub> Cloud-Screened Total and Tropospheric Column dataset<sup>72</sup>, for a region including all back-trajectories (40°E to 108°E, 40°N to 10°S). Aerosol optical density satellite measurements for the same period are reported in Fig. S9.

**Statistics.** In total, 41 PM<sub>2.5</sub> samples were initially collected during the campaign, but one filter was excluded due to technical problems with the pump at the time of collection. For the 40 samples included in the analysis, quality assessment of sample variation by PCA showed no outliers (Fig. S10). Multivariate analyses (i.e. PCA and OPLS models) were performed in SIMCA v.16 (Umetrics/Satorius); See Supplementary Information, Chemometrics, and Supplementary data 3 – Models and statistics. The R Packages ‘ggpubr’ and ‘ggplot2’ were used for other statistics and data visualization.

## Data availability

All supporting data are available in the Supplementary and on the Figshare repository under the DOI identifier: <https://doi.org/10.6084/m9.figshare.18517874>. Mass spectrometry (MS<sup>1</sup> and MS<sup>2</sup>) datasets have been deposited at the GNPS / Mass Spectrometry Interactive User Environment (MassIVE) database and made public under the access numbers: LC-HRMS WSOC ESI(+) [MSV000087675](https://massive.ucsd.edu/MSV000087675) and ESI(-) [MSV000087679](https://massive.ucsd.edu/MSV000087679); LC-HRMS POC ESI(+) [MSV000087681](https://massive.ucsd.edu/MSV000087681) and ESI(-) [MSV000087682](https://massive.ucsd.edu/MSV000087682); GC-HRMS NPOC EI(+) [MSV000087683](https://massive.ucsd.edu/MSV000087683) and NCI(-) [MSV000087684](https://massive.ucsd.edu/MSV000087684). See details in Supplementary Information for links to molecular networking and data visualization in the GNPS Dashboard<sup>73</sup>. As an illustrative example, the peak of atrazine-2-hydroxy (*m/z* 198.1352 [M + H]<sup>+</sup> at Rt 11.8 min) – an herbicide metabolite that to our knowledge has never been reported in air – is shown for a polluted air sample (WSOC LC-ESI+) associated with the Indo-Gangetic Plain back-trajectory (<https://bit.ly/3Lbteb0>; XIC Tolerance 0.005 Da). See also spectral library hit with the GNPS/MassBank record (<https://bit.ly/3spg4ys>; see View Mirror Match)

## Code availability

Code (R and Python) used in this study for the calculation of Kendrick’s mass defects (KMD), estimation of feature overlap (GC and LC), and prediction of refractive indexes from molecular structures, can be found on the Figshare repository under the DOI identifier: <https://doi.org/10.6084/m9.figshare.18517874>.

Received: 30 August 2021; Accepted: 27 January 2022;

Published online: 18 February 2022

## References

1. Landrigan, P. J. et al. The Lancet Commission on pollution and health. *Lancet*. **391**, 462–512 (2018).
2. World Health Organization (WHO), 7 million deaths linked to air pollution annually NIEHS: new WHO collaborating centre for environmental health



- Network to advance progress in children's environmental health Launch of WHO International Scheme to Evaluate Household Water Treatment Technology (2014).
- Apte, J. S., Brauer, M., Cohen, A. J., Ezzati, M. & Pope, C. A. Ambient PM<sub>2.5</sub> Reduces Global and Regional Life Expectancy. *Environ. Sci. Tech. Let.* **5**, 546–551 (2018).
  - Lelieveld, J. et al. Loss of life expectancy from air pollution compared to other risk factors: a worldwide perspective. *Cardiovasc. Res.* **116**, 1910–1917 (2020).
  - Sharma, D. C. No clear way ahead: smog in northern India. *Lancet.* **394**, 1891–1892 (2019).
  - Balakrishnan, K. et al. The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the Global Burden of Disease Study 2017. *Lancet. Planet. Heal.* **3**, e26–e39 (2019).
  - Chin, M., Diehl, T., Ginoux, P. & Malm, W. Intercontinental transport of pollution and dust aerosols: Implications for regional air quality. *Atmos. Chem. Phys.* **7**, 5501–5517 (2007).
  - Süßing, R. et al. Organophosphate esters in Canadian Arctic air: Occurrence, levels and trends. *Environ. Sci. Technol.* **50**, 7409–7415 (2016).
  - Liu, Y. et al. Heterogeneous OH initiated oxidation: A possible explanation for the persistence of organophosphate flame retardants in air. *Environ. Sci. Technol.* **48**, 1041–1048 (2014).
  - Boucher, O. D. et al. Clouds and aerosols. in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 571–657 <https://doi.org/10.1017/cbo9781107415324.016> (2013).
  - Dasari, S. et al. Photochemical degradation affects the light absorption of water-soluble brown carbon in the South Asian outflow. *Sci. Adv.* **5**, 1–11 (2019).
  - Seinfeld, J. H. et al. Improving our fundamental understanding of the role of aerosol-cloud interactions in the climate system. *Proc. Natl. Acad. Sci.* **113**, 5781–5790 (2016).
  - Shamjad, P. M. et al. Contribution of Brown Carbon to Direct Radiative Forcing over the Indo-Gangetic Plain. *Environ. Sci. Technol.* **49**, 10474–10481 (2015).
  - Jimenez, J. L. et al. Evolution of Organic Aerosols in the Atmosphere. *Science.* **326**, 1525–1529 (2009).
  - Laskin, A., Laskin, J. & Nizkorodov, S. A. Chemistry of Atmospheric Brown Carbon. *Chem. Rev.* **115**, 4335–4382 (2015).
  - Ditto, J. C. et al. An omnipresent diversity and variability in the chemical composition of atmospheric functionalized organic aerosol. *Commun. Chem.* **1**, 75 (2018).
  - Ditto, J. C. et al. Nontargeted Tandem Mass Spectrometry Analysis Reveals Diversity and Variability in Aerosol Functional Groups across Multiple Sites, Seasons, and Times of Day. *Environ. Sci. Tech. Let.* **7**, 60–69 (2020).
  - Johnston, M. V. & Kerecman, D. E. Molecular Characterization of Atmospheric Organic Aerosol by Mass Spectrometry. *Annu. Rev. Anal. Chem.* **12**, 247–274 (2019).
  - Lin, P., Fleming, L. T., Nizkorodov, S. A., Laskin, J. & Laskin, A. Comprehensive Molecular Characterization of Atmospheric Brown Carbon by High Resolution Mass Spectrometry with Electrospray and Atmospheric Pressure Photoionization. *Anal. Chem.* **90**, 12493–12502 (2018).
  - An, Y. et al. Molecular characterization of organic aerosol in the Himalayas: Insight from ultra-high-resolution mass spectrometry. *Atmos Chem Phys* **19**, 1115–1128 (2019).
  - Laskin, J. et al. High-resolution desorption electrospray ionization mass spectrometry for chemical characterization of organic aerosols. *Anal Chem* **90**, 12493–12502 (2010).
  - Wang, X. et al. Chemical Characteristics and Brown Carbon Chromophores of Atmospheric Organic Aerosols Over the Yangtze River Channel: A Cruise Campaign. *J. Geophys. Res. Atmos.* **125**, 32497 (2020).
  - Skinner, O. S. & Kelleher, N. L. Illuminating the dark matter of shotgun proteomics. *Nat. Biotechnol.* **33**, 717–718 (2015).
  - Silva, R. R. D., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proc. Natl. Acad. Sci.* **112**, 12549–12550 (2015).
  - Aron, A. T. et al. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat. Protoc.* **15**, 1954–1991 (2020).
  - Nothias, L. F. et al. Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods.* **17**, 905–908 (2020).
  - Hamilton, D. S. et al. Occurrence of pristine aerosol environments on a polluted planet. *Proc. Natl. Acad. Sci.* **111**, 18466–18471 (2014).
  - Uetake, J. et al. Airborne bacteria confirm the pristine nature of the Southern Ocean boundary layer. *Proc. Natl. Acad. Sci.* **117**, 13275–13282 (2020).
  - Cressey, D. Brown clouds boost global warming. *Nature.* **448**, 575–578 (2007).
  - Lawrence, M. G. & Lelieveld, J. Atmospheric pollutant outflow from southern Asia: a review. *Atmos Chem Phys* **10**, 11017–11096 (2010).
  - Silva, R. R. da et al. Propagating annotations of molecular networks using in silico fragmentation. *PLoS Comput. Biol.* **14**, 1006089 (2018).
  - O'Boyle, N. M. et al. Open Babel: An open chemical toolbox. *J. Cheminformatics.* **3**, 33 (2011).
  - Daina, A., Michielin, O. & Zoete, V. SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep* **7**, 42717 (2017).
  - Bouteloup, R. & Mathieu, D. Improved model for the refractive index: Application to potential components of ambient aerosol. *Phys. Chem. Chem. Phys.* **34**, 22017–22026 (2018).
  - Tang, J. et al. Molecular compositions and optical properties of dissolved brown carbon in biomass burning, coal combustion, and vehicle emission aerosols illuminated by excitation-emission matrix spectroscopy and Fourier transform ion cyclotron resonance mass spectrometry analysis. *Atmos. Chem. Phys.* **20**, 2513–2532 (2020).
  - Lelieveld, J. et al. Cardiovascular disease burden from ambient air pollution in Europe reassessed using novel hazard ratio functions. *Eur. Heart J.* **40**, 1590–1596 (2019).
  - Schymanski, E. L. et al. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environ. Sci. Technol.* **48**, 2097–2098 (2014).
  - Petrás, D. et al. Non-targeted tandem mass spectrometry enables the visualization of organic matter chemotype shifts in coastal seawater. *Chemosphere.* **271**, 129450 (2021).
  - Peisl, B. Y. L., Schymanski, E. L. & Wilmes, P. Dark matter in host-microbiome metabolomics: Tackling the unknowns—A review. *Anal. Chim. Acta.* **1037**, 13–27 (2018).
  - Pospisilova, V. et al. On the fate of oxygenated organic molecules in atmospheric aerosol particles. *Sci. Adv.* **6**, aax8922 (2020).
  - Schum, S. K., Brown, L. E. & Mazzoleni, L. R. MFAssignR: Molecular formula assignment software for ultrahigh resolution mass spectrometry analysis of environmental complex mixtures. *Environ. Res.* **191**, 110114 (2020).
  - Dührkop, K. et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
  - Ruttikies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. *J. Cheminformatics.* **8**, 3 (2016).
  - Vogel, A. L. et al. Aerosol Chemistry Resolved by Mass Spectrometry: Linking Field Measurements of Cloud Condensation Nuclei Activity to Organic Aerosol Composition. *Environ. Sci. Technol.* **50**, 10823–10832 (2016).
  - Datta, B. K., Datta, S. K., Rashid, M. A., Nash, R. J. & Sarker, S. D. A sesquiterpene acid and flavonoids from *Polygonum viscosum*. *Phytochemistry* **54**, 201–205 (2000).
  - Shukla, V. et al. Lichen Diversity in Different Lichenogeographical Regions of India. In *Lichens to Monitor the Environment*. Publisher: Springer India, [https://doi.org/10.1007/978-81-322-1503-5\\_4](https://doi.org/10.1007/978-81-322-1503-5_4) (2014).
  - Parrot, D. et al. Qualitative and spatial metabolite profiling of Lichens by a LC-MS approach combined with optimised extraction. *Phytochem. Analysis.* **26**, 23–33 (2015).
  - Kirkby, J. et al. Ion-induced nucleation of pure biogenic particles. *Nature.* **533**, 521–526 (2016).
  - Montoya-Aguilera, J. et al. Secondary organic aerosol from atmospheric photooxidation of indole. *Atmos. Chem. Phys.* **17**, 11605–11621 (2017).
  - Melendez-Colon, V. J., Luch, A., Seidel, A. & Baird, W. M. Cancer initiation by polycyclic aromatic hydrocarbons results from formation of stable DNA adducts rather than apurinic sites. *Carcinogenesis.* **20**, 1885–1891 (1999).
  - Salim, S. Y., Kaplan, G. G. & Madsen, K. L. Air pollution effects on the gut microbiota. *Gut. Microbes.* **5**, 215–219 (2013).
  - MohanKumar, S. M. J., Campbell, A., Block, M. & Veronesi, B. Particulate matter, oxidative stress and neurotoxicity. *Neurotoxicology.* **29**, 479–488 (2008).
  - Hems, R. F., Schnitzler, E. G., Liu-Kang, C., Cappa, C. D. & Abbatt, J. P. D. Aging of Atmospheric Brown Carbon Aerosol. *Accs. Earth Space Chem.* **5**, 722–748 (2021).
  - He, Q. et al. Evolution of the Complex Refractive Index of Secondary Organic Aerosols during Atmospheric Aging. *Environ. Sci. Technol.* **52**, 3456–3465 (2018).
  - Tomasi, C., Vitale, V., Petkov, B., Lupi, A. & Cacciari, A. Improved algorithm for calculations of Rayleigh-scattering optical depth in standard atmospheres. *Appl. Optics.* **44**, 3320–3341 (2005).
  - Chen, Q. et al. Elemental composition of organic aerosol: The gap between ambient and laboratory measurements. *Geophys. Res. Lett.* **42**, 4182–4189 (2015).
  - Tu, P., Hall, W. A. & Johnston, M. V. Characterization of Highly Oxidized Molecules in Fresh and Aged Biogenic Secondary Organic Aerosol. *Anal. Chem.* **88**, 4495–4501 (2016).
  - Scheringer, M. Long-range transport of organic chemicals in the environment. *Environ. Toxicol. Chem.* **28**, 677–690 (2009).

59. Andersson, J. T., Hegazi, A. H. & Roberz, B. Polycyclic aromatic sulfur heterocycles as information carriers in environmental studies. *Anal. Bioanal. Chem.* **386**, 891–905 (2006).
60. Tomaz, S. et al. Sources and atmospheric chemistry of oxy- and nitro-PAHs in the ambient air of Grenoble (France). *Atmos. Environ.* **161**, 144–154 (2017).
61. Daellenbach, K. R. et al. Sources of particulate-matter air pollution and its oxidatise potential in Europe. *Nature.* **587**, 414–419 (2020).
62. Gustafsson, Ö. et al. Brown Clouds over South Asia: Biomass or Fossil Fuel Combustion? *Science.* **323**, 495–498 (2009).
63. Mandelbaum, R. T., Wackett, L. P. & Allan, D. L. Rapid Hydrolysis of Atrazine to Hydroxyatrazine by Soil Bacteria. *Environ. Sci. Technol.* **27**, 1943–1946 (1993).
64. Kirillova, E. N. et al. 13C- and 14C-based study of sources and atmospheric processing of water-soluble organic carbon (WSOC) in South Asian aerosols. *J. Geophys. Res. Atmos.* **118**, 614–626 (2013).
65. Liang, D. et al. Use of high-resolution metabolomics for the identification of metabolic signals associated with traffic-related air pollution. *Environ. Int.* **120**, 145–154 (2018).
66. Vermeulen, R., Schymanski, E. L., Barabási, A. L. & Miller, G. W. The exposome and health: Where chemistry meets biology. *Science.* **367**, 392–396 (2020).
67. Lamichhane, D. K., Leem, J.-H., Lee, J.-Y. & Kim, H.-C. A meta-analysis of exposure to particulate matter and adverse birth outcomes. *Environ. Heal. Toxicol.* **30**, 11 (2015).
68. Guarnieri, M. & Balmes, J. R. Outdoor air pollution and asthma. *Lancet.* **383**, 1581–1592 (2014).
69. Martelletti, L. & Martelletti, P. Air Pollution and the Novel Covid-19 Disease: a Putative Disease Risk Factor. *Sn. Compr. Clin. Medicine* **2**, 383–387 (2020).
70. Tsugawa, H. et al. MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **12**, 523–526 (2015).
71. Aksenov, A. A. et al. Auto-deconvolution and molecular networking of gas chromatography–mass spectrometry data. *Nat. Biotechnol.* **39**, 169–173 (2021).
72. Krotkov, N. A. et al. and the OMI core team. Cloud-Screened Total and Tropospheric Column L3 Global Gridded 0.25 degree × 0.25 degree V3, *Goddard Space Flight Center, Goddard Earth Sciences Data and Information Services Center (GES DISC)*, <https://doi.org/10.5067/aura/omi/data3007> (2019).
73. Petras, D. et al. GNPS Dashboard: collaborative exploration of mass spectrometry data in the web browser. *Nat. Methods* **1–3**, 88 (2021).

## Acknowledgements

This research was supported by grants from the Swedish Research Council for Sustainable Development, Formas (Grants 2017–00567 and 2020-01917) and the Swedish Research Council (Grants 2018-03409, 2017 – 01601). MCOH is operated by the Maldives Meteorological Service (MMS) and funding for the operation of the site comes from Formas (Grant 942-2015-1061) and the Swedish Research Council (Grants 2015-03279 and 2017-01601). We thank the technical staff at MCOH for collecting and shipping air samples and quality controls. We thank Jan T. Andersson (University of Münster, Germany) for donation of sulfur-containing PAH standards, and Joël Boustie (Rennes Institute of Chemical Sciences, France) for donation of acetyl portentol standard. We thank Hiroshi Tsugawa (Tokyo University of Agriculture and Technology, Japan) for support with MS-DIAL to facilitate the analysis of environmental contaminants.

## Author contributions

S.P. performed raw data pre-processing, chemical confirmations, data analyses and statistics, evaluated and interpreted the results, created the figures, and drafted the main paper. L.A.D. and I.S. performed sample extractions and HRMS analyses, raw data pre-processing, data analyses and evaluated and interpreted the results. J.F. performed raw data pre-processing and chemical confirmations. B.B., K.S., and H.X. performed HRMS analyses and chemical confirmations. I.A. performed sample extractions and HRMS analyses. K.B. and S.D. calculated back-trajectories. K.B., S.D., A.A., Ö.G., and J.W.M. established field sampling. J.W.M. and Ö.G. conceived the project. J.W.M. coordinated the research and contributed to data interpretation and writing. All authors commented or edited in the final version of the paper.

## Funding

Open access funding provided by Stockholm University.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s43247-022-00365-1>.

**Correspondence** and requests for materials should be addressed to Jonathan W. Martin.

**Peer review information** *Communications Earth & Environment* thanks Daniel Petras and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Yinon Rudich and Clare Davis. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022