

**SUPPLEMENTARY MATERIAL FOR:**

**ENVIRONMENTAL CHEMICAL BURDEN IN DIFFERENTIATED  
THYROID CANCER**

4 Joshua D. Preston<sup>1,2,3</sup>, Yongliang Liang<sup>3</sup>, Thomas Szabo Yamashita<sup>4</sup>, Sami N. Teeny<sup>3</sup>, Jaclyn Weinberg<sup>3,5</sup>, William J.  
5 Crandall<sup>3,5</sup>, Zachery R. Jarrell<sup>3</sup>, Xin Hu<sup>3,6</sup>, Susan A. Sailey<sup>4</sup>, Jennifer M. Robertson<sup>7</sup>, ViLinh Tran<sup>3</sup>, Anee S.  
6 Jackson<sup>4</sup>, Snehal G. Patel<sup>4</sup>, Collin J. Weber<sup>4</sup>, Jyotirmay Sharma<sup>4</sup>, Neil D. Saunders<sup>4</sup>, Young-Mi Go<sup>3</sup>, Dean P. Jones<sup>3</sup>,  
7 and M. Ryan Smith<sup>3,8</sup>

8 1- Medical Scientist Training Program, Emory University School of Medicine, Atlanta, GA 30322

9 2- Nutrition and Health Sciences, Laney Graduate School, Emory University, Atlanta, GA 30322

10 3- Division of Pulmonary, Allergy, Critical Care and Sleep Medicine, Department of Medicine, Emory University School of Medicine, Atlanta,  
11 GA, USA 30322

12 4- Division of General and GI Surgery, Department of Surgery, Emory University School of Medicine, Atlanta, GA 30322

13 5- Molecular and Systems Pharmacology, Laney Graduate School, Emory University, Atlanta, GA 30322

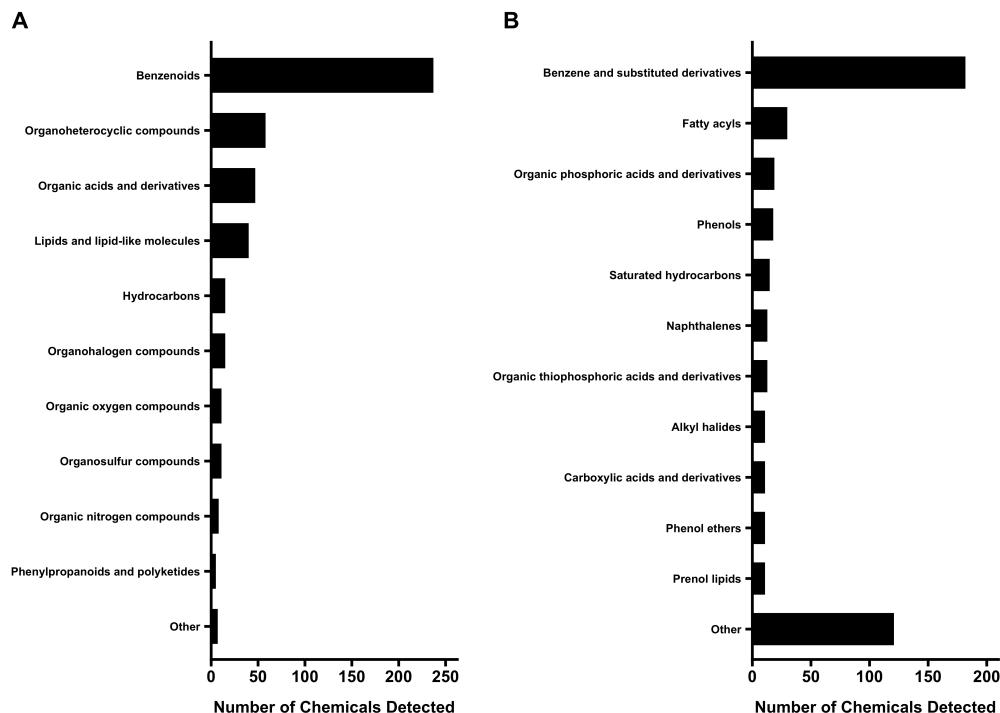
14 6- Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA 30322

15 7- Winship Cancer Institute, Emory University, Atlanta, GA 30322

16 8- VA Healthcare System of Atlanta, Decatur, GA 30033

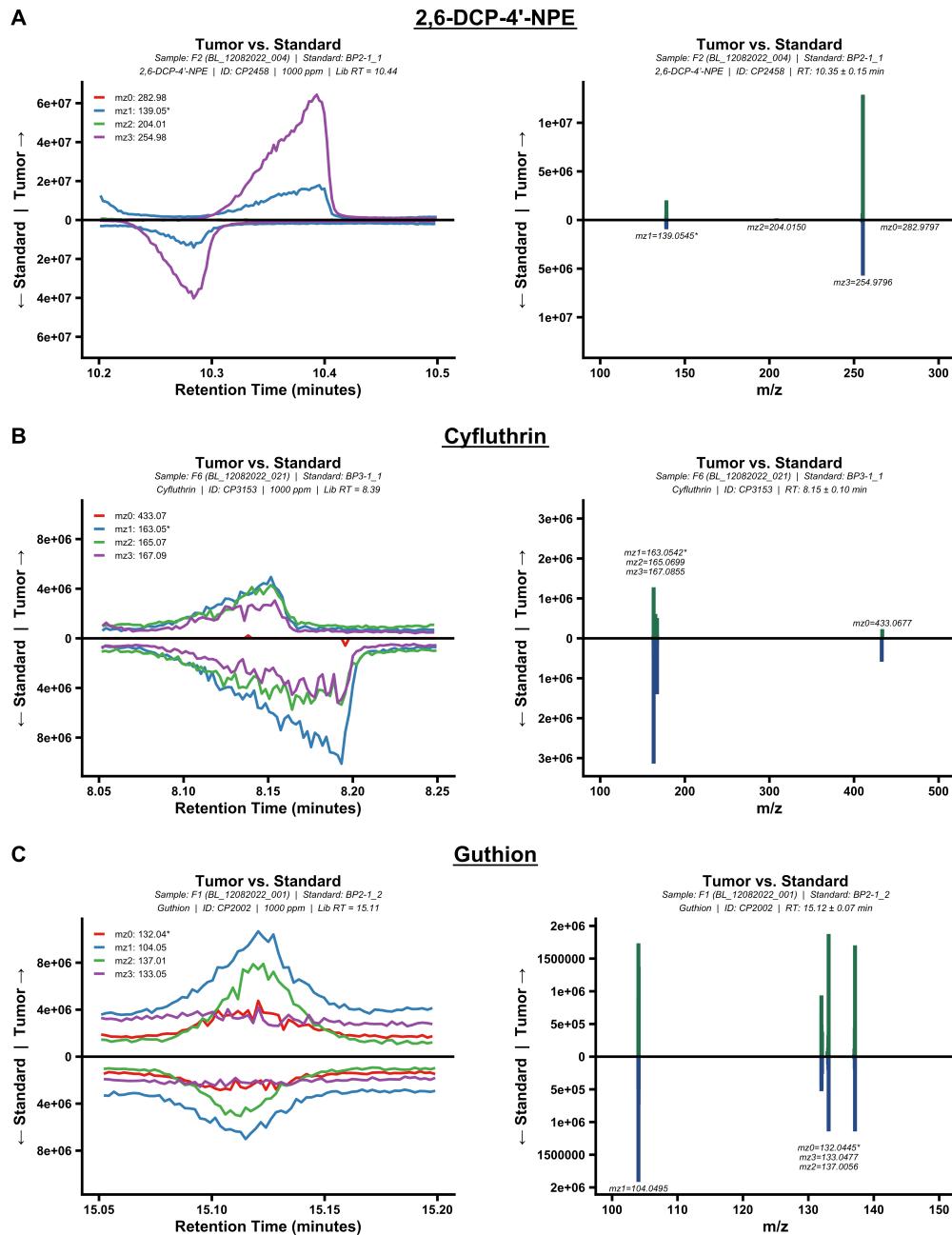
**TABLE OF CONTENTS**

18 <b>SUPPLEMENTARY FIGURE 1</b> .....	2
19 <b>SUPPLEMENTARY FIGURE 2</b> .....	3-4
20 <b>SUPPLEMENTARY FIGURE 3</b> .....	5
21 <b>EXPANDED METHODS</b> .....	6-11
22 <i>High-Resolution Exposomics</i> .....	6
23 <i>Data Extraction, Feature Annotation, and Signal Identification</i> .....	6
24 <i>Preprocessing of Exposomics Data</i> .....	7
25 <i>Metadata Annotation of Environmental Chemicals</i> .....	7
26 <i>Core Statistical Analysis and Data Visualization</i> .....	7
27 <i>Online Code and Data Repository</i> .....	8
28 <i>Exposome-Wide Association Study</i> .....	8
29 <i>Advanced Carcinogenicity Classification of Select Chemicals</i> .....	9
30 <i>Quantitative Estimates of Chemical Concentrations in Tissues</i> .....	9
31 <i>Quantitative Comparisons of Tumors and Non-Cancer Cadaver Thyroids</i> .....	11
32 <i>Comparing Observed Concentrations to Literature Values for Select Chemicals</i> .....	11
33 <b>SUPPLEMENTARY TABLE 1</b> .....	12-31
34 <b>SUPPLEMENTARY TABLE 2</b> .....	32-45
35 <b>SUPPLEMENTARY TABLE 3</b> .....	46
36 <b>SUPPLEMENTARY TABLE 4</b> .....	47
37 <b>TABLE ABBREVIATION DICTIONARY</b> .....	48-51
38 <b>REFERENCES CITED IN SUPPLEMENTARY MATERIAL</b> .....	52

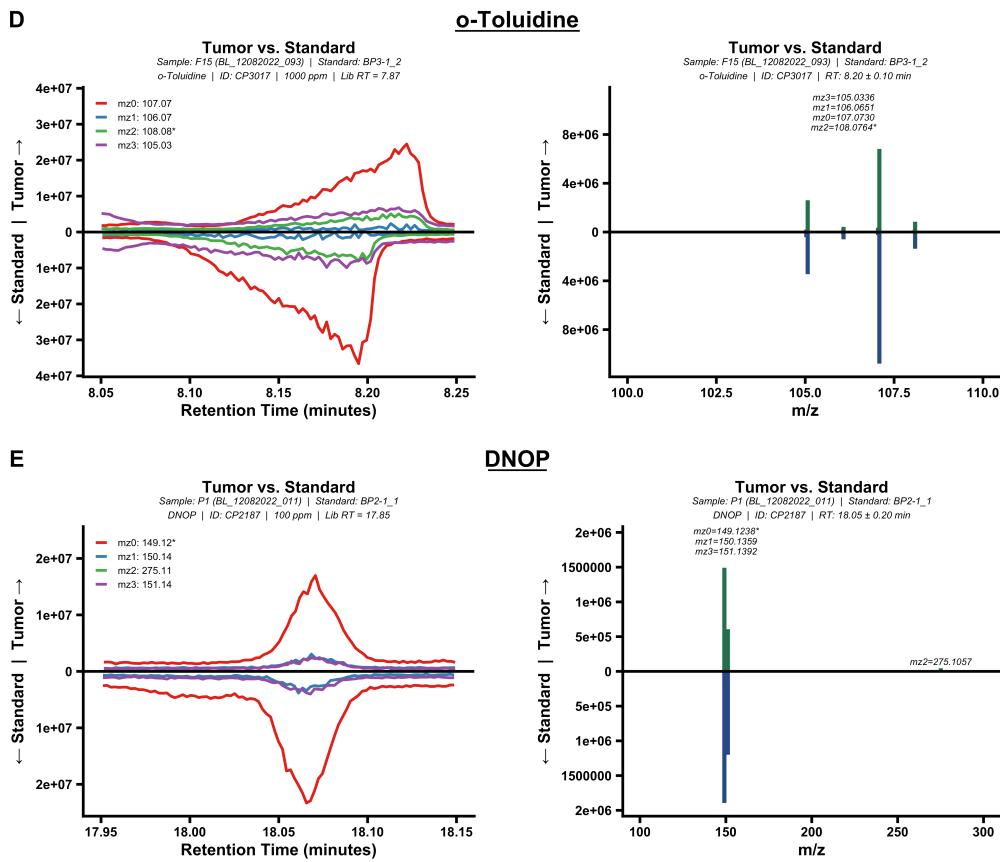
39 **SUPPLEMENTARY FIGURE 1**40 **Classification of detected chemicals.** Chemical superclasses (A) and subclasses (B) of all detected chemicals.

42 **SUPPLEMENTARY FIGURE 2**

43 **Spectral validation of select chemical identifications in thyroid tumor samples.** Mirrored plots comparing sample  
 44 spectra (top, positive y-axis) to reference standard spectra (bottom, negative y-axis, flipped) for the top five  
 45 quantitative mode chemicals from figure 3A: (A) 2,6-DCP-4'-NPE, (B) Cyfluthrin, (C) Guthion, (D) o-Toluidine,  
 46 and (E) DNOP. For each chemical, the left panel shows retention time (RT) chromatograms with extracted ion  
 47 chromatograms (EIC) for library m/z fragments, while the right panel shows mass spectra extracted within a RT  
 48 window around the target peak. The Y axis on all plots represents intensity. Asterisks (\*) indicate the fragment used  
 49 for statistical comparison and displayed in Figure 3A. Figures A-C are displayed below and D-E are displayed on the  
 50 following page.

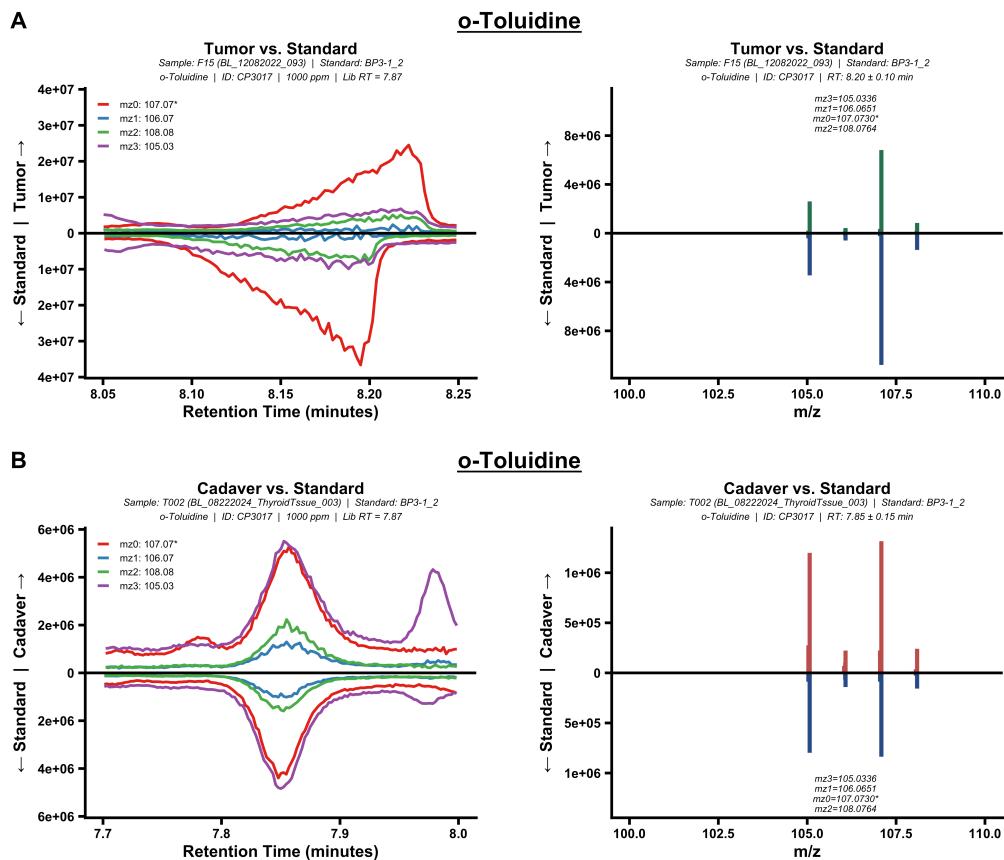


## 52 SUPPLEMENTARY FIGURE 2 (CONTINUED)



54 **SUPPLEMENTARY FIGURE 3**

55 **Spectral validation of o-Toluidine in tumor and non-cancer cadaver thyroid tissues.** Mirrored plots comparing  
 56 sample spectra (top, positive y-axis) to reference standard spectra (bottom, negative y-axis, flipped) for o-Toluidine  
 57 detected in (A) thyroid tumor tissue and (B) non-cancer cadaver thyroid tissue. Note that tumor and cadaver samples  
 58 were analyzed in separate batches. For each tissue type, the left panel shows retention time (RT) chromatograms with  
 59 extracted ion chromatograms (EIC) for library m/z fragments, while the right panel shows mass spectra extracted  
 60 within a RT window around the target peak. The Y axis on all plots represents intensity. Asterisks (\*) indicate the  
 61 fragment used for statistical comparison and displayed in Figure 3D.



## 63 EXPANDED METHODS

### 64 High-Resolution Exposomics

65 Prior to processing samples, several individual reference standards were prepared by spiking pure chemical standards  
66 into pooled reference plasma to achieve a common concentration of 0·47 ng/mL. A majority of all chemical standards  
67 were purchased from AccuStandard (New Haven, CT, USA), Restek (Bellefonte, PA, USA), MilliporeSigma  
68 (Burlington, MA, USA), Santa Cruz Biotechnology (Dallas, TX, USA), Wellington Laboratories (Guelph, ON,  
69 Canada), Cambridge Isotope Laboratories (Tewksbury, MA, USA). Plasma reference standards and tissue samples  
70 were processed using a modified version of the express liquid extraction (XLE) method.<sup>1</sup> The XLE is optimized for  
71 biofluids; however, it was necessary to modify it to ensure analytical consistency across plasma standards and tumor  
72 tissue samples, as detailed below.

73 First, ~50 mg of tumor or non-tumor cadaver thyroid tissue was cut and placed into a glass vial (13\$×\$100 mm),  
74 and exact tissue weights were recorded (mean ± SD = 52·7 ± 11·3 mg for tumor tissue, 43·1 ± 4·5 mg for  
75 non-cancer cadaver thyroid tissue). Next, 200 µL of each pooled reference plasma standard (an in-house standard  
76 referred to as QStd; see Go et al.<sup>2</sup> for a full description of a previous iteration of this standard) was transferred to  
77 glass vials (13\$×\$100 mm). For tissues, 500 µL of an extraction buffer comprised of acetone and petroleum ether  
78 (1:1 v/v) (MilliporeSigma, Catalog #184519) with ~2% internal standard mix, was added to the glass vials  
79 containing tissues. For plasma, 200 µL of extraction buffer containing 2% internal standard mix was added to the  
80 glass vials containing the pooled reference plasma. 50 µL of formic acid (Emprove® Essential DAC, 98-100% pure,  
81 MilliporeSigma, Catalog #1002631000) was then added to tissue and plasma standard vials. To the tissue vials, 20  
82 mg of NaCl was added, and 45 ± 5 mg of MgSO<sub>4</sub> (\$≥\$99·99% trace metals basis, Sigma-Aldrich, Catalog #203726)  
83 was added to tissue and plasma standard vials.

84 Tissue samples were then homogenized in this solution using a tissue-tearor, followed by water bath sonication for 30  
85 minutes at room temperature. Plasma samples were shaken vigorously on ice using a multitube vortexer (VWR  
86 VX-2500) for 60 min. Following this, both tissue and plasma homogenates were centrifuged at 3000 rpm × 4 °C ×  
87 10 min. Supernatants were then transferred to new glass vials (13 × 100 mm) and dried using a vacuum centrifuge  
88 for 60 min at 35 °C. Finally, dried samples were reconstituted in 50 µL of isoctane. Reconstituted samples were  
89 transferred to autosampler vials with 150 µL vial inserts for analysis. Samples were prepared in batches containing  
90 20 samples along with standard reference material (SRM) samples (NIST 1957, 1958, and several in-house SRMs  
91 discussed above). In addition, pooled reference plasma samples, a retention-time batch comparison standard, a  
92 method solvent blank that underwent solvent extraction, and a solvent blank were prepared in tandem as part of  
93 quality control measures. It should be noted that tumor tissues and non-cancer cadaver thyroid tissues were analyzed  
94 at different timepoints and in separate batches, although the same method was used for each tissue. As a result, direct  
95 comparison of raw spectral intensities between non-cancer thyroids and tumors was limited; however, quantitative  
96 estimates of chemical concentrations in tissues (see the section ‘Quantitative Estimates of Chemical Concentrations  
97 in Tissues’ below) did enable some degree of direct comparison, though these comparisons should be interpreted  
98 conservatively and cautiously given the high potential for batch effects.

99 Following preparation, samples were analyzed in duplicate with a Q Exactive GC hybrid quadrupole Orbitrap mass  
100 spectrometer (ThermoScientific), allowing for 25 minutes of data collection on retention times and spectral intensities  
101 from m/z 85-850 collected at 60k resolution. Samples were injected at 2 µL and subjected to a capillary DB-5MS  
102 column (15 m × 0·25 mm × 0·25 µm film thickness) with a gradient as follows: 75 °C for 1 min then 25 °C/min to  
103 180 °C, 6 °C/min to 250 °C, 20 °C/min to 300 °C, with a final 5 min hold. Helium gas flow rate was 1 mL/min.  
104 Positive electron ionization was set at 70 eV, the ion source was set at 250 °C, and the transfer line was set at 280 °C.

### 105 Data Extraction, Feature Annotation, and Signal Identification

106 Raw data were extracted on a sample-by-sample basis using MZmine2.<sup>3</sup> Data were interpreted and combined using  
107 an in-house algorithm for feature annotation and the identification and quantification of chemicals in GC-MS  
108 exposomics data. Briefly, the algorithm utilizes a large library of environmental chemicals (EC) with known  
109 concentrations in standards, each of which has been manually quantified to ascertain retention times within the  
110 context of the laboratory procedure. In total, 738 unique standards are in the library; however, 28 of these represent  
111 chemicals that can be considered largely endogenous and not ECs. The remaining standards can be considered  
112 entirely exogenous (700) or both endogenous and exogenous (10) chemicals, resulting in a total of 710 ECs in the  
113 library, which can be screened by parallel analysis of standards with samples. This library is compared against sample  
114 data to find the best matching signal for ECs if any appropriate candidates exist. Shifts in elution time due to column  
115 degradation and the presence of co-occurring isomers pose significant challenges for standard annotation procedures;  
116 however, the algorithm addresses this issue by employing time-warping techniques, along with corrective algorithms

that mimic manual chemical identification procedures, to overcome these challenges. Chemical identification and accuracy are achieved by leveraging the co-elution and correlation of chemical fragments generated from the same chemical. Standards are run in tandem with samples, which allows not only for the identification and relative quantification of all chemicals in the library of standards but also for the absolute quantification of identified metabolites using the comparative intensity of peak areas. The algorithm specifically assigns “quality fractions” based on the simultaneous detection of multiple fragments of the same chemical. Following the application of the algorithm for annotation, the quality fraction was used to label features as “annotations” (quality fraction of 0) while the remaining signals were labeled “identifications.” It is important to note that identification versus annotation status was determined solely in reference to tumor samples and not non-cancer cadaver thyroid samples. Thus, when annotations or identifications are denoted in any figures or tables, these refer to the determination made in the data from thyroid tumors rather than non-cancer cadaver samples. Information on annotation versus identification status in non-cancer cadaver samples can be found in the “primary\_data.xlsx” spreadsheet located in the GitHub repository (see below for further details). The algorithm will be published as a Python or R package in the coming 1-2 years. For further details on the algorithm, readers are welcome to contact the corresponding author.

### 131 Preprocessing of Exposomics Data

A targeted exposomics feature table (see the associated GitHub repository) was derived from the feature annotation and signal identification process described above. The proportion of missing values (PMV) was then calculated for each feature (total number of missing values/60 total samples). Next, features with PMV > 30% missing values were separated for analysis in “qualitative” mode. The individual spectral intensity values for qualitative features were converted to “1”, representing detection, whereas missing values were converted to “0”, representing non-detection. Any qualitative features with 100% “0” values (indicating detection in reference standards but complete non-detection in samples) were eliminated and not considered for further analysis. Alternatively, the remaining features with PMV ≤ 30% were separated for analysis in “quantitative” mode. The missing values for quantitative features were imputed using the half minimum method (study-wide).<sup>4</sup> Following imputation, all spectral intensity values for quantitative features were log<sub>2</sub>-transformed before further analysis. Finally, to determine the total number of unique chemicals detected per sample, features were consolidated and grouped by their respective Chemical Abstracts Service registry numbers (CAS number), regardless of analytical mode (qualitative versus quantitative), such that the detection of at least one fragment for a given CAS number was considered as detection of that unique chemical. This list of chemicals was then compiled, and chemicals that were mainly endogenous (i.e., carnitines, cholesterol, arachidonic acid metabolites, bile acids, hormones, and hormone precursors) were removed and not considered for exposomics analysis (23 total). Chemicals that had both exogenous and endogenous dispositions (7 total) were retained for exposomics analysis. The number of unique chemicals detected was then tallied for each sample. The median number of chemicals detected per variant (prior to imputation) was compared using the Kruskal-Wallis test.

### 150 Metadata Annotation of Environmental Chemicals

The name, IUPAC name, CAS number, PubChem Compound Identification (CID), SMILES, InChIKey, InChI, Toxin and Toxin-Target Database<sup>5</sup> ID (when applicable), monoisotopic mass, and formula were obtained for each chemical detected in at least one sample. For the purposes of data visualization and simplified naming in the manuscript text, a short name or abbreviation was assigned to a large portion of chemicals. The Toxin-Toxin-Targeted Database<sup>5</sup> (T3DB) ID and the International Agency for Research on Cancer (IARC) carcinogen group<sup>6</sup> were assigned to chemicals when applicable and possible (when this information could not be ascertained, the T3DB ID was listed as “NA” and the IARC Group was listed as “Not Classified”). Classification as either a potential endocrine-disrupting chemical (EDC) or non-EDC was assigned to each chemical, per the PARCEDC list.<sup>7</sup> Inclusion on the list resulted in chemicals being marked as potential EDCs; non-inclusion resulted in classification as a non-EDC. The single most common use or best-fit chemical class was identified for each chemical. To facilitate data visualization, broader categories were also assigned to each chemical. For example, if a chemical’s common-use class was “Insecticide/Pesticide (Pyrethroid)”, this was simplified to “Insecticide/Pesticide” for data display. In the broader categories, metabolites or degradation products of certain chemicals were simply considered as members of the class to which their precursor belonged. For example, aldicarb sulfone, a breakdown product of the carbamate pesticide aldicarb, was classified as an insecticide/pesticide. However, the inclusion of breakdown products is indicated in figures and tables where applicable. ClassyFire<sup>8</sup> was used to designate superclasses, classes, subclasses, direct parents, and molecular frameworks. All chemical metadata are available below in Supplementary Table 1.

### 168 Core Statistical Analysis and Data Visualization

All data were compiled and structured using Microsoft Excel (Mac v16.96, Microsoft Corporation, 2025). All code

170 was written and executed using R (v4·3·1)<sup>9</sup> in Visual Studio Code (v1·82·2, Microsoft Corporation, 2025). Source  
171 code can be found on the GitHub repository (`source_code.R`; see below for details). Data visualization was  
172 performed using GraphPad Prism (Mac v10·4·0, GraphPad Software, 2024) and the R package, `ggplot2`.<sup>10</sup> All figures  
173 were compiled and edited using BioRender.com (see ‘BioRender Publication Licenses’ below).

#### 174 **Online Code and Data Repository**

175 All source code is available in the accompanying GitHub repository, accessible at the following URL:  
176 <https://github.com/jdpreston30/thyroid-exposomics-2025>.

#### 177 **Exposome-Wide Association Study**

178 Electron ionization generates multiple fragments for each chemical prior to detection; thus, in many cases, multiple  
179 fragments were annotated for the same chemical. In keeping with our exploratory and descriptive approach and to  
180 maximize coverage, we treated each fragment as an individual observation during statistical analysis. For qualitative  
181 features, Fisher’s exact test for count data was employed on the binary exposomics data to determine if any chemical  
182 was overrepresented in its detection within any specific variant. Corresponding detection fractions were calculated  
183 for each variant (sum of samples with detection of chemical/20 samples per variant). For quantitative features,  
184 one-way ANOVA was used to assess differences in mean spectral intensities across variants, with corresponding  
185 p-values calculated for each chemical. Post-hoc testing was performed using Tukey’s HSD test, which included  
186 correction for multiple comparisons. Given the pilot design and exploratory nature of this study, no false discovery  
187 rate corrections were applied beyond those used in post-hoc testing. Mean spectral intensities were scaled by  
188 converting them to z-scores within each chemical. Chemicals with a  $p < 0\cdot05$  from either test were considered  
189 significant and were compiled, tabulated, and further classified as described below. Four chemicals had multiple  
190 fragments showing significant differences between variants. In these cases, the quantitative fragment was selected if  
191 available; otherwise, the fragment with the highest frequency of detection across all samples, irrespective of variant,  
192 was selected. The top 5 quantitative features were visualized via violin plots (of z-scored spectral intensities), and the  
193 top 10 qualitative features via a heatmap (of detection fraction/percent). A balloon plot was constructed based on  
194 these results, wherein significant ECs were grouped by usage classes and variants, being counted ‘once each  
195 time it was highest,’ whether by qualitative or quantitative measures in a variant. In cases where the detection  
196 percentages were equally high in two variants, a count was assigned to each variant.

#### 197 **Advanced Carcinogenicity Classification of Select Chemicals**

198 To identify any potential carcinogens that have not been evaluated by the IARC, additional research was performed  
199 on the chemicals that showed significant differences in presence or concentration between variants (63 in total). To  
200 accomplish this, the Global Harmonized System of Classification and Labelling of Chemicals (GHS) statements were  
201 screened for any listed H350 (“may cause cancer”), H350i (“may cause cancer by inhalation”), or H351 (“suspected  
202 of causing cancer”) status (and the corresponding consensus/confidence percentages of data sources as listed on  
203 PubChem in section 12·1·1, “GHS Classification”). Systematic logic that considered both the IARC grouping and the  
204 GHS carcinogen statement was then applied to categorize chemicals into the groups “Known Carcinogen”, “Likely  
205 Carcinogen”, “Possible Carcinogen”, or “Uncertain Risk” (for the function accomplishing this, see the blocks of code  
206 under “#+ Carcinogen classification based on GHS and IARC (3E)” in the `source_code.R` file available in the GitHub  
207 repository). Chemicals with no IARC group and GHS statement were not classified. Briefly, the following rules were  
208 applied for classification:

- 209 • IARC group 1 chemicals were automatically assigned “Known Carcinogen”, and group 2A “Likely  
210 Carcinogen”. These conditions overrode any other considerations.
- 211 • If IARC grouping was 2B, 3, or not evaluated, but H350 or H350i 50%, then the chemical was assigned  
212 “Likely Carcinogen”.
- 213 • If IARC grouping was 2B, 3, or not evaluated, but  $H351 > 0\%$  and/or  $0\% < H350/H350i < 50\%$ , then the  
214 chemical was assigned “Possible Carcinogen”.
- 215 • IARC group 3 chemicals with no supporting GHS H350, H350i, or H351 statement listed were assigned  
216 “Uncertain Risk”.
- 217 • Chemicals with no IARC evaluation and no supporting GHS H350, H350i, or H351 statement listed were  
218 considered “Unclassified” as to their carcinogenicity.

<sup>219</sup> These classifications, along with other relevant chemical metadata, can be found on the GitHub repository  
<sup>220</sup> ([chemical\\_metadata.xlsx](#)).

## <sup>221</sup> Quantitative Estimates of Chemical Concentrations in Tissues

<sup>222</sup> The SRM used for annotation and identification had a concentration of 0·47 ng/mL for each individual EC. Thus, we  
<sup>223</sup> used the spectral intensities associated with these known concentrations to establish estimates of chemical  
<sup>224</sup> concentrations within tissues in parts per million (PPM) and parts per billion (PPB). It should be noted that the  
<sup>225</sup> original data processing pipeline used a nominal concentration of 0·5 ng/mL for calibration calculations; therefore, a  
<sup>226</sup> correction factor of 0·94 (0·47/0·5) was applied to all concentration estimates prior to PPM and PPB calculations to  
<sup>227</sup> account for the true SRM concentration. Importantly, we relied on the following experimental features and  
<sup>228</sup> assumptions:

- <sup>229</sup> • Tissues are processed via solvent extraction and subsequent solvent evaporation, followed by reconstitution in  
<sup>230</sup> 50  $\mu$ L of isoctane. For the purposes of calculating concentrations in the original samples, we assume 100%  
<sup>231</sup> extraction efficiency from tissues and plasma, such that the entire amount of the EC present in the sample is  
<sup>232</sup> recovered in the isoctane.
- <sup>233</sup> • A chemical concentration of 0 ng/mL results in an intensity of 0 (i.e., the origin of (0,0)). This is used in  
<sup>234</sup> combination with the intensity from the known concentration in the SRM (0·47 ng/mL) to establish a two-point  
<sup>235</sup> standard curve, which assumes linearity.
- <sup>236</sup> • While each chemical has a concentration in the SRM of 0·47 ng/mL, the concentration in 50  $\mu$ L of isoctane is  
<sup>237</sup> 1·88 ng/mL. This is because 200  $\mu$ L of SRM plasma is used, and therefore, 0·094 ng of the total EC of interest  
<sup>238</sup> is extracted. This is dried after extraction and reconstituted in 50  $\mu$ L of isoctane, resulting in a concentration  
<sup>239</sup> of 1·88 ng/mL in the injected solution.

<sup>240</sup> Our calculations to estimate tissue concentrations were performed as follows for each individual chemical:

1. A linear model is fit based on the assumptions discussed above, such that:

$$y = \beta x$$

<sup>241</sup> where  $y$  is the known concentration of the EC in the SRM (0·47 ng/mL for all ECs analyzed),  $x$  is the observed  
<sup>242</sup> spectral intensity for the given EC, and  $\beta$  is the slope of the calibration curve, defined by the line through the  
<sup>243</sup> origin (0,0) and the point  $(x, y)$ .

2. Once the slope ( $\beta$ ) for the specific chemical is established, it is then used to estimate concentration (ng/mL) in the sample ( $C_e$ ) using the following equation:

$$C_e = \beta I$$

<sup>244</sup> where  $I$  is the observed spectral intensity for the given EC in the sample. However, to determine the true  
<sup>245</sup> concentration in the isoctane solvent ( $C_{es}$ ), we must scale  $C_e$  by a factor of 4 to account for the concentration  
<sup>246</sup> that occurs during extraction. Specifically, the 200  $\mu$ L of plasma standard extracted contains 0·094 ng total of  
<sup>247</sup> the EC (0·47 ng/mL). Assuming 100% extraction, this 0·094 ng was then resuspended in 50  $\mu$ L isoctane  
<sup>248</sup> (1·88 ng/mL). Therefore, we use a 4 $\times$  scaling factor:

$$C_{es} = 4C_e$$

3. Next, the  $C_{es}$  is converted from ng/mL to mg/L:

$$C_{es} (\text{mg/L}) = C_{es} (\text{ng/mL}) \times \frac{10^3 \text{ mL}}{\text{L}} \times \frac{10^{-6} \text{ mg}}{\text{ng}}$$

$$C_{es} (\text{mg/L}) = C_{es} (\text{ng/mL}) \times 10^{-3}$$

4. Next, the total mass of the chemical of interest present in 50  $\mu\text{L}$  of isoctane ( $M_{EC}$ ) is determined:

$$M_{EC} \text{ (mg)} = C_{es} \text{ (mg/L)} \times 50 \text{ } \mu\text{L} \text{ solvent} \times \frac{10^{-6} \text{ L}}{\mu\text{L}}$$

$$M_{EC} \text{ (mg)} = 5 \times 10^{-5} \text{ (L)} \times C_{es} \text{ (mg/L)}$$

5. To derive the mass fraction of the EC in tissue, we divide  $M_{EC}$  (mg) by  $M_T$  (mg) and scale to PPM or PPB:

$$\text{PPM} = \frac{M_{EC}}{M_T} \times 10^6$$

$$\text{PPB} = \frac{M_{EC}}{M_T} \times 10^9$$

Alternatively, when algorithm outputs are used directly without unit conversions, the following formula can be used, provided  $C_e$  is in ng/mL and  $M_T$  is in mg:

$$\text{PPM} = \frac{C_e \times 10^2}{M_T}$$

$$\text{PPB} = \frac{C_e \times 10^5}{M_T}$$

#### **244 Quantitative Comparisons of Tumors and Non-Cancer Cadaver Thyroids**

245 Targeted quantitative feature tables, which were additionally filtered for detection in relevant standards, were  
 246 generated for tumors and non-cancer cadaver thyroids, and conversion to PPM was performed as described above.  
 247 Once spectral intensities were converted to PPM, direct comparisons of chemical concentrations were made only for  
 248 fragments that were detected in both the tumor and non-cancer thyroid tissues. In cases where multiple fragments  
 249 were annotated for the same chemical in both tumors and non-cancer thyroids, an ideal fragment was first selected  
 250 based on the highest percentage of detection in tumors, then the highest percentage of detection in non-cancer  
 251 thyroids, and finally the greatest mean intensity in tumors. For all matching fragments, the mean, maximum, and  
 252 theoretical minimum (i.e., half the minimum detectable concentration in either the tumors or non-cancer thyroids)  
 253 were calculated for tumors (all combined) and non-cancer thyroids. These data, along with percentage detection in  
 254 tumors and non-cancer thyroids, are reported as PPB below in Supplementary Table 3. It is important to note that  
 255 means were calculated using half-minimum imputed data, even if features were detected at ‘qualitative’ analysis  
 256 thresholds. Finally, these values were converted to PPM for a direct comparison of tumors to non-cancer thyroids of  
 257 all detected IARC Group 1 carcinogens (Figure 3D); however, this comparison was only reported for features that  
 258 met quantitative criteria in both the tumor and non-cancer thyroid datasets. T-tests were run on log<sub>2</sub>-transformed data  
 259 to compute p-values for comparing the means between tumors and non-cancer thyroids for these chemicals.

#### **260 Comparing Observed Concentrations to Literature Values for Select Chemicals**

261 To compare the quantification estimates generated as described above with estimated values in other tissues and  
 262 matrices, we identified values for select chemicals in the published literature and tabulated our data alongside these  
 263 reported concentrations (Supplementary Table 2). We prioritized chemicals that were IARC Group 1 carcinogens and  
 264 polycyclic aromatic hydrocarbon combustion byproducts. For quantitative estimates of 245 individual chemicals, see  
 265 Supplementary Table 3.

266 **SUPPLEMENTARY TABLE 1**

267 **The full library of xenobiotic chemicals employed for chemical identification.** The library of 710 confirmed xenobiotic chemicals employed for chemical  
268 identification. All chemicals were present in pooled reference plasma at a concentration of 0·47 ng/mL. There are 710 total unique chemicals (i.e., unique CAS  
269 numbers), but for some chemicals, there are multiple fragments from different standards used for identification, thus resulting in 892 total rows in the table. The  
270 individual mz columns indicate typical fragments observed for the given chemical.

271 [INSERT ST1 HERE - TO BE GENERATED PROGRAMMATICALLY]

**272 SUPPLEMENTARY TABLE 2**

**273 Metadata for all chemicals detected in samples.** Both the long-form chemical name and alias or abbreviation are listed if there is sufficient space. However, for  
274 chemical names that are too long or redundant, only the alias or abbreviation has been listed. The column ‘variant diff.’ specifies if the chemical had differential  
275 abundance or detection between the three variants.

**276 [INSERT ST2 HERE - TO BE GENERATED PROGRAMMATICALLY]**

**277 SUPPLEMENTARY TABLE 3**

**278 Observed concentrations versus reported literature values.** All values originally published as ng/g,  $\mu$ g/L, ng/mL  
279 are listed as PPB. Values originally published as pg/mL were converted to ng/mL and then listed as PPB. All values  
280 are rounded to the nearest integer or are listed as < 1 PPB when applicable. The corresponding reference from which  
281 the comparison value is derived is cited next to the listed concentration.

**282 [INSERT ST3 HERE - TO BE GENERATED PROGRAMMATICALLY]**

**283 SUPPLEMENTARY TABLE 4**

**284 Quantitative estimates of chemicals in non-cancer thyroids and tumors.** Data table containing quantitative  
285 estimates of chemical concentrations in thyroid tissues.

**286 [INSERT ST4 HERE - TO BE GENERATED PROGRAMMATICALLY]**

**287 TABLE ABBREVIATION DICTIONARY**

288 The supplementary tables have a substantial number of abbreviations, largely for chemical names. A full dictionary  
289 of abbreviations relevant to all supplementary tables can be found below:

- 290 • 9Cl-PF3ONS = 9-Chlorohexadecafluoro-3-oxanone-1-sulfonic acid
- 291 • BDCPP = bis(1,3-Dichloro-2-propyl) phosphate
- 292 • BDE = brominated diphenyl ether
- 293 • BDPP = Bis(2,3-dibromopropyl) hydrogen phosphate
- 294 • Bromo-TMP-Phenol = 2-Bromo-4-(2,4,4-trimethylpentan-2-yl)phenol
- 295 • CAS = Chemical Abstracts Service (Number)
- 296 • CDC = Centers for Disease Control and Prevention
- 297 • CID = Compound ID (PubChem)
- 298 • Compds. = compounds
- 299 • DBaA = Dibenz(a,h)anthracene
- 300 • DCP = Dichlorophenyl
- 301 • DCPMN = 4-(2,4-dichlorophenoxy)-2-methyl-1-nitrobenzene
- 302 • DDD = Dichlorodiphenyldichloroethane
- 303 • DDE = Dichlorodiphenyldichloroethylene
- 304 • DDT = Dichlorodiphenyltrichloroethane
- 305 • DFTPP = Decafluorotriphenylphosphine
- 306 • DTPAs = Dithiophosphoric Acids
- 307 • EPN = Ethyl p-nitrophenyl phenylphosphorothioate
- 308 • EtFOSAA = N-Ethylperfluoro-1-octanesulfonamidoacetic acid (linear)
- 309 • Furaneol = 4-Hydroxy-2,5-dimethyl-3(2H)-furanone
- 310 • HpCDD = Heptachlorodibenzo-p-dioxin
- 311 • HpCDF = Heptachlorodibenzofuran
- 312 • HxCDD = Hexachlorodibenzo-P-dioxin
- 313 • HxCDF = Hexachlorodibenzofuran
- 314 • IARC = International Agency for Research on Cancer
- 315 • IMHP = 2-Isopropyl-6-methyl-4-pyrimidinol
- 316 • Lin. = linear
- 317 • LLMs = lipid-like molecules
- 318 • LOD = limit of detection
- 319 • MBOT = 4,4'-Methylenebis(o-toluidine)
- 320 • MCPA = 2-Methyl-4-chlorophenoxyacetic acid
- 321 • MEcPP = Mono(5-carboxy-2-ethylpentyl) phthalate

- 322 • MEHHP = Mono(2-ethyl-5-hydroxyhexyl) phthalate
- 323 • MEOHP = Mono(2-ethyl-5-oxohexyl) phthalate
- 324 • MGK-264 = McLaughlin Gormley King-264 (also known as N-2-Ethylhexylbicycloheptenedicarboximide)
- 325 • min = minutes
- 326 • MOCA = 4,4'-Methylenebis(2-chloroaniline)
- 327 • mz = mass-to-charge ratio
- 328 • N-MeFOSAA = N-Methylperfluoro-1-octanesulfonamidoacetic acid (linear)
- 329 • NHANES = National Health and Nutrition Examination Survey
- 330 • NPE = nitrophenyl ether
- 331 • o-Dianisidine = 3,3'-Dimethoxybenzidine
- 332 • OD-PABA = Octyl-dimethyl-p-aminobenzoic acid
- 333 • Org = organic
- 334 • Org. Heterocycl. = organoheterocyclic
- 335 • p-Chlorocresol = 4-Chloro-3-methylphenol
- 336 • PAH = polycyclic aromatic hydrocarbon
- 337 • PBB = polybrominated biphenyl
- 338 • PCB = polychlorinated biphenyl
- 339 • PCDF = Pentachlorodibenzofuran
- 340 • PeCDD = pentachlorodibenzo-p-dioxin
- 341 • PKs = polyketides
- 342 • PPB = parts per billion
- 343 • RT = retention time
- 344 • SDs = Substituted Derivatives
- 345 • TBBPA-BAE = Tetrabromobisphenol A bis(allyl ether)
- 346 • TCDD = tetrachlorodibenzo-p-dioxin
- 347 • TCDF = tetrachlorodibenzofuran
- 348 • TCP = Trichlorophenyl
- 349 • TCP-4'-NPE = TCP-4'-NPE
- 350 • TCPP = Tris(1-chloro-2-propyl) phosphate
- 351 • TDCPP = Tris(1,3-dichloro-2-propyl)phosphate
- 352 • TEEP = Tetraethyl ethylenediphosphonate
- 353 • TPAs = Thiophosphoric acids
- 354 • TTBNPP = Tris(tribromoneopentyl) phosphate

**355 REFERENCES CITED IN SUPPLEMENTARY MATERIAL**

- 356 1 Hu X, Walker DI, Liang Y, *et al.* A scalable workflow to characterize the human exposome. *Nature Communications* 2021; **12**: 5575.
- 357 2 Go Y-M, Walker DI, Liang Y, *et al.* Reference Standardization for Mass Spectrometry and High-resolution Metabolomics Applications to Exposome Research. *Toxicological Sciences* 2015; **148**: 531–43.
- 358 3 Pluskal T, Castillo S, Villar-Briones A, Orešić M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 2010; **11**: 395.
- 359 4 Wei R, Wang J, Su M, *et al.* Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Scientific Reports* 2018; **8**: 663.
- 360 5 Wishart D, Arndt D, Pon A, *et al.* T3DB: The toxic exposome database. *Nucleic Acids Research* 2015; **43**: D928–34.
- 361 6 Agents Classified by the IARC Monographs, Volumes 1–136.
- 362 7 Andres S, Dulio V. S109 | PARCEDC | List of 7074 potential endocrine disrupting compounds (EDCs) by PARC T4.2. 2024; published online April. DOI:10.5281/ZENODO.10944198.
- 363 8 Djoumbou Feunang Y, Eisner R, Knox C, *et al.* ClassyFire: Automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics* 2016; **8**: 61.
- 364 9 (2023) RCT. R: A Language and Environment for Statistical Computing.
- 365 10 Wickham H. Ggplot2: Elegant graphics for data analysis, Second edition. Switzerland: Springer, 2016.