IDSL.UFA assigns high confidence molecular formula annotations for untargeted LC/HRMS datasets in metabolomics and exposomics

Sadjad Fakouri Baygi[1], Sanjay K Banerjee[2], Praloy Chakraborty[2], Yashwant Kumar[2], Dinesh Kumar Barupal[1]*

[1] Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA

[2] Non-communicable Diseases Division, Translational Health Science and Technology Institute, Faridabad, Haryana, 121001, India

* Corresponding author: Address: CAM Building, 3rd floor, 17 E 102nd St, New York, NY 10029 Email: dinesh.kumar@mssm.edu, phone: +1-530-979-4354

Abstract

Untargeted LC/HRMS assays in metabolomics and exposomics aims to characterize the small molecule chemical space in a biospecimen. To gain maximum biological insights from these datasets, LC/HRMS peaks should be annotated with chemical and functional information including molecular formula, structure, sub-structure, chemical class and metabolic pathways. Among these, a molecular formula can be assigned to a majority of LC/HRMS peaks using the theoretical and observed isotopic profiles (MS1) of the underlying ionized compound. For this, we have developed the Integrated Data Science Laboratory for Metabolomics and Exposomics – Ultimate Formula Annotation (IDSL.UFA) R package. In the untargeted metabolomics validation tests, IDSL.UFA assigned 54.31%-85.51% molecular formula for true positive annotations as the top hit, and 90.58%-100% within the top five hits. Molecular formula annotations were further validated by MS/MS data. We have implemented novel strategies to 1) generate formula sources and their theoretical isotopic profiles 2) optimize the formula hits ranking on the individual and the aligned peak lists and 3) scale IDSL.UFA based workflows for studies with larger sample sizes. Annotating the raw data for a publicly available pregnancy metabolome study using IDSL.UFA highlighted hundreds of new pregnancy related compounds, and also suggested presence of chlorinated perfluorotriether alcohols (Cl-PFTrEAs) in human specimens. IDSL.UFA is useful for human metabolomics and exposomics studies where we need to minimize the loss of biological insights in untargeted LC/HRMS datasets. The IDSL.UFA package is available in the R CRAN repository https://cran.r-project.org/package=IDSL.UFA. Detailed documentation and tutorials are also provided at www.ufa.idsl.me.

Introduction

Untargeted LC/HRMS analyses of human specimens enable studying the metabolome and exposome in an unbiased manner[1,2].They have delivered many novel biomarkers and mechanisms for diseases and have improved our understating of basic metabolic pathways[3-5] . These assays are unique in nature since they record all the mass to charge (m/z) ratio signals and their intensities for ionized compounds in a sample[6]. This makes the collected data a rich source of information with great opportunities to generate novel hypotheses about metabolome and exposome. It is critical for the promises that untargeted assay offer, that the data are utilized in an inclusive way to not miss any discovery opportunities.

A key post-data acquisition step in the untargeted LC/HRMS assays is to annotate the detected peaks with a range of structural and functional information which can enable a biological interpretation[3, 7, 8]. This information includes a chemical structure, sub-structure, molecular formula, chemical class and metabolic pathway[1, 9, 10]. These annotations may help in understanding the nature, origin and function of the chemical structure underlying a peak. Among these information, molecular formula can be assigned to a LC/HRMS peak using the observed and theoretical isotopic profiles for a chemical compound.[11] Isotopic profiles are distinguishable mass spectral signature that represent atomic masses and their natural abundances in the molecular formulas of a compound.[12] Given recent advancements in mass spectrometry instruments, experimental isotope profiles can be almost identical to their theoretical counterpart[13] and thus can be used to annotate LC/HRMS peaks with molecular formula[14]. Peak annotation by isotopic profile matching should be performed using efficient computational strategies to accounts for instrumental errors, multi-sample studies, biological plausibility and chemical diversity.[15]

There has been a great deal of efforts to develop computational tools in academic and commercial software to assign molecular formulas to the detected peaks in a LC/HRMS dataset. SIRIUS[16] and ZODIAC[17], NetID[18] assigns molecular formula to MS/MS fragmentation data. However, implementing it for larger studies where only MS1 data are available for every sample remains to be challenging due to the ranking of formula hits on individual and aligned peak tables, scalable computation and various sources for formulas. Many other successful tools have been developed to detect Br/Cl containing compounds which had a characteristics isotopic profile in the MS1 data.[11-15, 19] MFAssignR[20] can assign molecular formula using MS1 data using Kendrick mass defect combined with isotopic signatures of selected elements. But these tools were not designed and tested for large-scale human biospecimens studies which is routinely seen in the exposome research. Moreover, there is a paucity of formula assignment solutions for developing LC/HRMS data processing workflows in the R computing environment[21].

Here, we have developed a scalable, user-friendly, thoroughly tested R package, the IDSL.UFA to assign molecular formula with high confidence to LC/HRMS peaks in untargeted LC/HRMS datasets from large-scale studies. IDSL.UFA covers major possible situations in which a molecular formula can be assigned to LC/HRMS peaks. We propose that processing LC/HRMS data with IDSL.UFA can find new opportunities for hypotheses and biomarker discoveries for studying the role of metabolism and exposome in human diseases.

Methods

IDSL.UFA R package has been submitted to the R-CRAN repository at https://cran.r-project.org/package=IDSL.UFA. Tutorials with detailed parameter spreadsheets are provided at (https://ufa.idsl.me). Figure 1 (simplified) and Figure S.1 (detailed) provide overview of steps to assign molecular formulas using the IDSL.UFA software.

Publicly available LC/HRMS datasets: The raw LC/HRMS data for human and mouse biospecimen studies (MTBLS1684[22], MTBLS2542[23], ST001683[24], ST001430[25], ST001154[26], ST002044 and reference authentic standards (MSV000088661) were downloaded from the National Institute of Health (NIH) Metabolomics WorkBench (https://www.metabolomicsworkbench.org/), Massive UCSD (https://www.massive.ucsd.edu), and European Bioinformatics Institute MetaboLights repositories (https://www.ebi.ac.uk/metabolights) using the study accession identifier (Table S.1). Sample preparation and data collection procedures are available at entry pages for these studies in the repositories. Raw data files were converted to the mzML format using the ProteoWizard msconvert utility[27].

IDSL.IPA data processing: IDSL.IPA R package (https://cran.r-project.org/package=IDSL.IPA) R package was utilized to generate peak intensity matrices. IDSL.IPA computed chromatographic characteristics for the masses of interests for each LC/HRMS file. It generated individual peaks lists and an aligned peak table for each study. Data processing parameter files and IDSL.IPA results for each study are provided in the Zenodo repository (Table S.1).

Generating the isotopic profile database (IPDB). To annotate IDSL.IPA peak lists with molecular formula, IDSL.UFA required theoretical isotopic profiles of candidate molecular formulas. The reference stable isotope masses and abundances for elements in the periodic table were extracted from the PubChem entries[28] which had IUPAC source[29]. User-provided ionization pathways (e.g. $[M-H]^-$, $[M+H]^+$) were required to generate ionized molecular formula.

Isotope profile for each ionized formula was generated by combining naturally occurring isotope abundances for each element. Theoretical profile intensity cutoffs may vary for different molecular formulas; therefore IDSL.UFA can handle conditional cutoffs for different molecular formula configurations using generic R commands. For example, R commands (ex: cutoff = 'if (s>0) {min(c(c, 10, s*4))} else {min(c(c, 10))}') may be used to preserve [M+2] isotopologues for sulfonated compounds (where c and s represent number of carbon and sulfur atoms in the molecular formula, respectively). A peak spacing parameter is also used to merge neighboring theoretical isotopologues depending on the HRMS instrument mass accuracy.[11] IPDBs were generated using various sources of molecular formula databases. Once, an IPDB generated, it may be reused in similar studies. IDSL.UFA can generate an IPDB of up to 10^8 candidate molecular formulas due to limits of current computational resources.

Molecular formula source – 1 (databases): Chemical compounds lists for four key databases in metabolomics and exposomics databases included the blood exposome (chemicals expected in a mammalian blood specimen)[30], RefMetDB (measured and expected small molecules in biological organisms)[31], Lipid Maps (known lipid molecules)[32] and the US-FDA substance registry[33]. These databases were combined into a single compound list referenced as IDSL.ExposomeDB in this manuscript and also provided at the Zenodo repository (https://zenodo.org/record/5823455). Charged compounds, isotope-labeled compounds, multi-components were excluded while compiling the IDSL.ExposomeDB. Unique molecular formulas from this consolidated database were used for IPDB generation. IPDBs for these four databases and CompTox Chemicals Dashboard[34] are available at Zenodo repository (https://zenodo.org/record/5823455).

Molecular formula source – 2 (enumerated chemical space with constraints): Molecular formulas were enumerated using a set of combinatorial and filtering rules using C, H, As, B, Br, Cl, F, I, K, N, Na, O, P, S, Se, and Si elements. These 16 elements were able to cover 93.76% of carbon-containing compounds (50 ≤ mass ≤ 2000) in the IDSL.ExposomeDB combined with EPA chemistry Dashboard[34]. Enumerated chemical spaces (ECSs) can be represented using equation (1).

$$C_c H_h As_{as} B_b Br_{br} Cl_{cl} F_f I_i K_k N_n Na_{na} O_o P_p S_s Se_{se} Si_{si} \quad (1)$$

where the subscripts of elements represent the number of atoms. ECSs should be constrained according to HRMS instruments and metabolomics and exposomics assays. The ECS boundaries and rules in the parameter files are provided in the Zenodo repository (https://zenodo.org/record/5838603). For screening of halogenated and phthalate like compounds different IPDBs were generated using different parameters files.

A fully combinatorial chemical space from above-mentioned 16 elements can have ~ 1.6×10^{27} formulas, which is impractical to be managed by current computational resources. Therefore, we derived and coded in R a set of four rules to constrain the ECSs which were inspired from the seven golden rules approach[35]. These rules were tested using the unique carbon containing molecular formulas (N = 244,990) from IDSL.ExposomeDB and the CompTox Chemicals Dashboard[34]. A combination of these four rules can cover 95.46% of molecular formulas.

Rule 1) C/N chemical space ((c/2-n-1) ≤ (h+cl+br+f+i) ≤ (2c+3n+6)) may be used to set elemental boundaries for the organic compounds to ensure entire moieties are bond to carbon and nitrogen atoms. There are 99.43% carbon-containing compounds in the blood exposome

database that follow this rule. For example, the top range of hydrogens may set as h ≤ (2c+3n+6-(br+cl)) to focus on realistic molecular formulas.

Rule 2) Extended SENIOR may be used to ensure that the molecular formulas completely filled s- and p- valence electron shells.[35] This rule was able to preserve 95.92% carbon-containing compounds. A random subset of 20 molecular formulas that did not observe this rule indicated 95.92% of them were in fact radical ions.

Rule 3) Maximum number of halogens may be used to focus on halogenated compounds to generate the molecular formulas library. The first halogen rule is the maximum number of (br+cl) ≤ 8 for 99.85% carbon-containing compounds (mass ≤ 2,000). The second halogen rule is the maximum number of ((br+cl+f+i) ≤ 31) for 99.9% carbon-containing compounds (mass ≤ 2,000).

Rule 4) Maximum number of elements may be used to skip unrealistically complex molecular formulas generated through molecular formula enumeration. There 98.13% carbon-containing molecular formulas have ≤ 6 elements in their molecular formula configurations. For example, the maximum number of elements for glucose ($C_6H_{12}O_6$) is three (C, H, and O).

Molecular formula source – 3 (exhaustive chemical enumeration for a single m/z value):

Use of larger ECSs (ex: $<10^8$ molecular formulas) for formula assignment tasks is computationally challenging. Thus, we developed an exhaustive chemical enumeration (ECE) approach to generate specific IPDBs for each single m/z using upper boundaries covering 99% of the chemical space. ECE utilized a subset sum solution provided by the RccpAlgos R package (https://CRAN.R-project.org/package=RcppAlgos) for each m/z and the four molecular formula enumeration rules. We recommend to only annotate a selected list of ions in a peak list using ECE. The ECE approach was implemented in a separate R package, IDSL.UFAx (https://cran.r-project.org/package=IDSL.UFAx).

Isotopic profile matching on individual IDSL.IPA peak lists: To annotate molecular formulas, the IDSL.UFA software requires IDSL.IPA peak lists to access properties of chromatographic peaks including peak boundaries, ^{12}C m/z, and ratio of cumulated intensity of ^{12}C to ^{13}C ($R^{13}C$). IDSL.UFA screens individual mass spectra across chromatographic peaks for the whole isotopic profiles using a mass accuracy. The IDSL.UFA software calculates cumulated intensities and average masses for each isotopologue using equations (2) and (3) across the chromatographic peaks.

$$\overline{Int} = \sum_{t=t_0}^{t=t_{end}} Int_t$$	(2)
$$\overline{m/z} = \frac{\sum_{t=t_0}^{t=t_{end}} m/z_t * Int_t}{\overline{Int}}$$	(3)

where m/z_t and Int_t represent mass and intensity of the matched isotopologue in individual scans across the chromatographic peak from t_0 to t_{end}.

Profile cosine similarity ($\overline{PCS}$) is determined to quantify profile similarity between experimental and theoretical isotopic profiles using equation (4). Normalized Euclidean mass error ($\overline{NEME}$) is calculated to measure the instrument accuracy for all isotopologues using equation (5).[11]

$$\overline{PCS} = \sum_{i=1}^{S} \frac{I_i^{theor} I_i^{exptl}}{\sqrt{\sum_{i=1}^{S}\left(I_i^{theor}\right)^2}\sqrt{\sum_{i=1}^{S}\left(I_i^{exptl}\right)^2}} \quad (4)$$

$$\overline{NEME} = \sqrt{\frac{\sum_{i=1}^{S}\left(M_i^{theor} - M_i^{exptl}\right)^2}{S}} \quad (5)$$

where I_i, M_i, and S represent the intensity of the isotopologue, mass of the isotopologues, and number of isotopologues in the isotopic profile, respectively. Superscripts of *theor* and *exptl* also represent theoretical and experimental isotopic profiles, respectively. $\overline{PCS}$ and $\overline{NEME}$ parameters enable the IDSL.UFA software to compare several candidate molecular formulas match for a peak regardless of the number of isotopologues in isotopic profile models. Rule-based criteria may also be used to define thresholds for *maxNEME* and *minPCS* to mitigate number of candidate molecular formulas.

The IDSL.UFA software also has a minimum threshold for the top 80% of number of scans with the confirmed whole isotopic profile (NDCS) within the chromatography peak. A minimum percentage of NDCS within a chromatography peak (RCS (%)) is also defined to scale NDCS values.

When all candidate molecular formulas have been screened, matching scores are calculated for all the candidate molecular formulas using equation (6) to rank candidate molecular formulas. This score function is a multiplicative score function where the product of the “greater is desirable” parameters are divided by the product of the “lower is desirable” parameters to boost score values for the true candidate matches.[11] The score values should be used only to rank molecular formulas related to a peak and cannot be used to evaluate distinct peaks.

$$Score = \left(\frac{S^{coeff[1]} * \left(\frac{\overline{PCS}}{1000}\right)^{coeff[2]} * \left(\frac{RCS}{100}\right)^{coeff[3]}}{\left(\frac{\overline{NEME}}{maxNEME}\right)^{coeff[4]} * \left(\exp\left(\left|\ln\left(\frac{\overline{R^{13}C_{PL}}}{R^{13}C_{IP}}\right)\right|\right)\right)^{coeff[5]}} \right) \quad (6)$$

where $\overline{R^{13}C_{PL}}$ and $R^{13}C_{IP}$ indicate experimental and theoretical $R^{13}C$ values, respectively. $R^{13}C$ values represent ratio of the general ^{13}C isotopologue [M+1] relative to ^{12}C isotopologue [M] on the most abundant mass. *coeff[1-5]* are powers of the parameters to apply different magnitude of each variable in different studies.

Score coefficients optimization: The IDSL.UFA utilizes coefficients of 1 to rank candidate molecular formulas by default. However, to improve the ranking of candidate molecular formulas, the IDSL.UFA package may compute optimized coefficients according to the instrument performance especially when many molecular formulas were included in an IPDB. When needed, *coeff[1-5]* were optimized using objective functions described in the section S.1 and the genetic algorithm R package[36].

Molecular formula scoring at the aligned table: This unique feature facilitated prediction of molecular formula of unknown masses using statistical methods especially in population-size studies. This module collects top rank molecular formulas from individual annotated IDSL.IPA peak lists and filters top hits with high detection frequency and ranks from the peak alignment table of the IDSL.IPA pipeline. The order of candidate molecular formula hits in the aligned table may be modified using the following formula: $\sqrt{frequency}/rank$.

Molecular formula class detection: Many compounds in the mass spectrometry analyses appeared in a class such as polychlorinated biphenyl (PCBs), polybrominated diphenyl ethers (PBDEs), polycyclic aromatic hydrocarbons (PAHs), perfluoroalkyl substances (PFAS), lipids, phthalates, etc; therefore, an R function '*detect_formula_sets'* in the IDSL.UFA R package was developed to group molecular formulas having repeated unit patterns which may represent a class. Tables S.2 and S.3 demonstrated constant ΔH/ΔC ratios for polymeric (ΔH/ΔC = 2) and cyclic (ΔH/ΔC = 1/2) chain progressions, respectively in a class of compounds. Then, Table S.4 indicated these ratios may be utilized to detect polymeric and cyclic classes according to the repeated units regardless of the fundamental organic group of the class. Additionally, Table S.5 also showed a constant number of carbons and fixed summation of hydrogens and halogens (Σ(H+Br+Cl+F+I)) may be used to detect classes with fixed structures similar to PCBs, PBDEs.

Correlation analysis for gestational age: The ST001430 study[25] includes weekly blood samples of 30 pregnancies acquired in 781 samples each processed in positive and negative modes to predict gestational age. To reduce batch effects, the peak heights were adjusted by raw total ion chromatograms (TICs) in each sample, and then, the positive and negative aligned peak height tables were stacked to generate a comprehensive list of peaks. We computed a Pearson correlation coefficient between gestational age and peak height data for each pregnancy. For the peaks that had a Pearson correlation coefficient of > 0.65, a hierarchal clustering dendrogram was generated using an average linkage in the hclust function in R. Formula annotations from the aligned table were used a label in the dendrogram.

Unsupervised clustering of LC/HRMS peaks: The combined data matrix for ESI (+) and ESI (-) modes were used an input for an unsupervised clustering analysis using the umap[37] and HDBSCAN[38] which relies on both linear and non-linear patterns to find clusters in the data. UMAP parameters were min_dist ≤ 0.11, n_neighbors ≤ 15 and matrix = Pearson. The HDBSCAN parameter minPts was 15. Annotated formula table was used to provide labels on each point in the UMAP-HDBSCAN plot. We visualized the LC/HRMS peaks that were potentially halogenated by using a mass defect of <0.80. Figure S.2 provides a flowchart of the analysis for pregnancy study (ST001430)[25].

Results and discussion

We have engineered a new software, IDSL.UFA, to annotate LC/HRMS peaks in an untargeted metabolomics or exposomics study with molecular formulas. In this approach, IDSL.UFA computes theoretical isotopic profiles for molecular formulas, matches theoretical isotopic profiles on experimental LC/HRMS data using a set of criteria and then ranks the candidate molecular formulas by frequency and order in multiple samples (aligned annotated peak table) in a study. The IDSL.UFA software has been implemented as an R package and made publicly available via the R-CRAN repository and www.ufa.idsl.me site.

IDSL.UFA natively uses IUPAC isotope table data[28] to calculate theoretical isotopic profiles. We demonstrated IDSL.UFA calculated almost identical isotopic profiles to that obtained from the *envi*Pat package[39] (Table S.6). Negligible mass and profile errors (NEME ≤ 0.69 mDa and PCS ≥ 999.99‰) were observed for formula $[C_8F_{17}O_3S]^-$ between IDSL.UFA and *envi*Pat[39]. However, by directly utilizing IUPAC tables in conjunction with few constraints for an enumerated chemical space (ECS), we could scale the computation of isotopic profiles for thousands of molecular formulas for large-scale untargeted LC/HRMS studies. Isotopic Profile Databases (IPDBs) for various molecular formula databases and scenarios were provided at the Zenodo repository and were utilized in the IDSL.UFA software.

Section 1) Development and validation of IDSL.UFA results: To demonstrate the validity of our new approach to assign molecular formulas, we have utilized datasets with true positive annotations and show their ranks in the IDSL.UFA result matrices.

Analysis of authentic reference standards: First, we evaluated performance of the IDSL.UFA software to detect molecular formulas in LC/HRMS data for authentic reference standards. We found that the average $\overline{NEME}$ (indicator of mass difference) was 0.70 mDa and $\overline{PCS}$ (indicator of isotope profile similarity) were 999.68‰ for 367 authentic standard compounds of common metabolites. This indicated that the observed isotopic profiles were perfectly matched to the theoretically counterparts for these reference standards and suggested that molecular formula can be reliably assigned to untargeted data generated by the commonly used LC/HRMS instruments. The theoretical and experimental integrated isotopic profile spectra across chromatography for these standards are provided at Zenodo repository accession (https://zenodo.org/record/5803968) and an example compound (Selenomethionine ion $[C_5H_{12}NO_2Se]^+$) is shown in Figure 2 and Figure S.3.

Analysis of untargeted LC/HRMS data with structurally annotated peaks: IDSL.UFA correctly assigned a majority of molecular formulas for true positives annotations in untargeted LC/HRMS data using a rank strategy at the aligned peak table level. For four publicly available studies (ST001154[26], ST001683[24], MTBLS1684[22], and MTBLS2542[23]). The IDSL.UFA software assigned 61.85%, 54.31%, 70.58% and 85.51% molecular formula as the top hit, and 96.90%, 90.58%, 100% and 99.29% molecular formulas in the top 5 hits in the aligned table. These results were generated using the IPDBs of the IDSL.ExposomeDB with 209,592 and 129,122 ion formulas in positive and negative modes from multiple ionization pathways, respectively representing 83,951 unique intact molecular formulas (http://zenodo.org/deposit/5838709). When IDSL.UFA software was executed using an ECS that covers the true positives annotations, the assignment rate were – 52.74%, 53.36%, 79.41% and 51.08% molecular formula as the top hit, and 95.60%, 84.45%, 100% and 91.66% molecular formula in the top 5 hits in the aligned table (Figure S.4). Generally, the IDSL.UFA software annotated 924 (90.14%) and 877 (85.56%) molecular formulas across all four studies using IDSL.ExposomeDB and ECS IPDBs, respectively. There are 603 (65.18%) and 470 (53.59%) first order hits on the aligned molecular formula tables using IDSL.ExposomeDB and study specific ECS methods, respectively shown in Figure S.4 for each study. These findings demonstrate that IDSL.UFA is a sensitive approach to cover all formulas for chemicals detectable in a biospecimen.

Ranking in the aligned peak table for the formula assignment. We propose that the ranking of annotated formula in the aligned peak table instead of individual peak list should be used for the formula assignment purpose. For example, a lipid peak (*m/z* = 766.5815, RT = 8.859 min) was putatively annotated as $[C_{44}H_{80}NO_7P+H]^+$ in the MTBLS1684 study. However, IDSL.UFA

suggested alternative molecular formula ion ($[C_{42}H_{82}NO_7P+Na]^+$) for this peak using the aligned annotated peak table strategy. The $[C_{42}H_{82}NO_7P+Na]^+$ ion was detected 492 times (98.59%) with median rank 1, while the $[C_{44}H_{80}NO_7P+H]^+$ ion was detected 37 times (7.41%) with median rank 2 using the IDSL.ExposomeDB IPDB. Generally, $[C_{42}H_{82}NO_7P+Na]^+$ has better mean $\overline{PCS}$ (991.47‰ > 895.07‰) and $\overline{NEME}$ (1.72mDa < 3.47mDa) values relative to $[C_{44}H_{80}NO_7P+H]^+$ shown in Figure S.5. When top 20 formula hits in the aligned table for three distinct sample types (ST002044, ST001683, and ST001430) studies were investigated for overlaps, only ~1/3 of them were found to be common among them, indicating the IDSL.UFA results at the aligned table were not random and reflected the sample specificity. Figures S.6 (a) and (b) show there are 37.55% and 34.27% commonalities among these cohorts in positive and negative modes, respectively. To our knowledge IDSL.UFA is the only tool that assigns molecular formulas based upon the aligned annotated molecular formula table on population-size studies.

Additional validation of molecular formula assignment by MS/MS: To further ensure that IDSL.UFA can assign high confidence molecular formula for untargeted LC/HRMS data, we utilized data from ST002044 study which has high quality MS/MS data collected in the data dependent mode. Of the 73 total hits that could be confirmed by matching their spectra to the NIST 2020 MS/MS library (https://chemdata.nist.gov/), IDSL.UFA assigned 78.75% of them within a median rank of ≤ 2 in the aligned peak table generated using the IDSL.ExposomeDB formula IPDB (Table S.7 and Figure S.7) even though a set of relaxed criteria (generally, mass accuracy ≤ 0.0075 Da, $\overline{NEME}$ ≤ 7.5 mDa, and $\overline{PCS}$ ≥ 950‰) was used for IDSL.UFA. Table S.8 shows the MS/MS library spectra matching results and the molecular formula that were ranked top by IDSL.UFA for 13 selected candidates (out of 14). These results provided additional supports to confidence in the molecular formula assignment by the IDSL.UFA software using the IDLS.ExposomDB IPDB.

Rank score optimization: IDSL.UFA utilized a number of chromatographic-mass spectrometry parameters to compute the rank of a molecular formula for a peak in the individual peak list. By default, a score coefficient of 1 is used which worked sufficiently in most situations. However, the rank can be further improved by an optimization strategy that utilize the true positive, curated and high-quality structure annotations for each data file as input. This can be achieved by running a mixture of reference standards using the same analytical method or by annotating peaks using MS/MS, RT and isotopic profile matching using stringent criteria. For THSTI IROA standards (MSV000088661) and THSTI-CVD(ST002044) studies, we have observed a significant improvement in the ranking of molecular formulas when optimized score coefficients were utilized in the IDSL.UFA software (Table S.9).

Covering a larger chemical space: One of the central aims of an untargeted LC/HRMS analysis is to identify compounds that are first time reported for a biospecimen. Many peaks in LC/HRMS datasets remained to be unannotated with any formulas from the IDSL.ExposomeDB. To cover larger chemical spaces beyond the IDSL.ExposomeDB, we can use broader ranges of element counts for creating ECS. However, this may generate an IPDB of many gigabytes in size ($<10^8$ molecular formulas) which is impractical for our aligned formula table strategy due to computing constraints. So, we tried a generative chemistry type approach using a sub-set sum problem to enumerate formula for each m/z value in the aligned peak-table, it took several days to annotate a single LC/HRMS data file even with parallel computing with 20 threads. However, it did provide the most comprehensive molecular formula annotation for a single sample from studies ST001430 (https://zenodo.org/record/5804527) and ST002044

(https://zenodo.org/record/5803970) and resulted with 940,041 annotated molecular formulas for 4,824 peaks. The Venn diagram (Figure S.8) indicated there are 889,130 (94.58%) unique molecular formulas only annotated by the generative chemistry approach that the ECS and IDSL.ExposomeDB were not able to cover on a test file (1.NEG.mzXML from ST001430). This subset sum-based approach requires a very high computing resource for annotating all the peak in a single sample, therefore it is more suitable for selected high priority peaks that have been identified using a statistical or bioinformatics analyses.

Targeting molecular formula pattern for a chemical class: As an alternative to the computationally expensive approach, we propose a middle strategy that utilize a sub-structure pattern search in conjunction with formula filtering rules, in which ECS is generated using a formula pattern such as alkanes or phthalates. This allows us to specifically look for certain chemical classes having formula patterns as shown in the figure S3-6. Using this approach, we have assigned 510 classes groups of related molecular formula including phthalates for the ST002044 study. However, these assignments were not validated yet and will require further MS/MS experiments. Nonetheless, IDSL.UFA provided a high-confidence list of formula for a class that can be helpful in interpreting the results for many LC/HRMS peaks which were otherwise left under-utilized due the lack of any annotation.

Section 2) Application of IDSL.UFA for a pregnancy study

To demonstrate an application of IDSL.UFA software to characterize the metabolome and exposome for blood specimens, we have re-processed a publicly available study ST001430[25] (n=781) which has weekly blood samples analyzed for 30 pregnancies to accurately predict gestational age (GA in weeks). Raw data were processed using the IDSL.IPA software to generate the individual peak lists and the aligned peak table (https://zenodo.org/record/5804527). On average, (3,416 ESI^- and 6,978 ESI^+) peaks were detected across individual peak lists for this study and a total of (89,174 ESI^- and 143,712 ESI^+) peaks were reported in the aligned peak table by the IDSL.IPA software. The IDSL.UFA software using the IDSL.ExposomeDB IPDB annotated (80,957 ESI^- and 124,647 ESI^+) peaks in the aligned peak table with at least one molecular formula having a median rank of ≤ 5.

First, we identify the peaks that were associated with GA by computing a spearman correlation coefficient between normalized peak-height for each peak and GA. On a spearman cutoff of (p-value ≤ 0.05, $|\rho| \geq 0.65$, “two.sided” alternative), 274 peaks with a detection frequency of ≥ 5 within each subject were found to be significantly associated with GA (only ≤ 36 weeks). We observed 242 (red) and 32 (blue) ascending and descending correlations patterns with GA, which were consistent with the patterns reported in the original paper[25] and corresponded to chemicals related to steroid hormone biosynthesis and long-chain fatty acids. However, we found many peaks with molecular formula annotations related to GA in comparison the original paper, highlighted that IDSL.UFA comprehensive characterize the pregnancy related metabolic changes (Figure 4).

We explored an unsupervised clustering of the entire dataset using umap[37] and hdbscan[38] methods to detect peak clusters based on linear and non-linear inter-chemical correlation patterns. A total of 340 clusters were detected by this approach, provided a view of the biochemical landscape that underlies pregnancy. On average 58 peaks were found across these clusters with a minimum size of 15 and maximum size of 737. Peaks that were associated with GA were clustered into cluster #105 (descending) and cluster #126 (ascending) shown in

Figure S.9. Formula assignments by IDSL.ExposomeDB IPDB suggested that these clusters have steroid hormone related compounds. Other clusters corresponded to relevant patterns such as compounds that were detected only in the second trimester period or compounds having a wave pattern across the pregnancy. These clusters provide a repertoire of hypotheses to further expand the knowledge about biochemical changes during the pregnancy. The user-interactive clustered umap figure annotated by the IDSL.ExposomeDB and ECS IPDBs are available at (https://ufa.idsl.me/st001430). Next, we have utilized a CF_2 mass defect technique (CF_2 mass defect = 50/49.99681*m/z - floor (m/z)) that can suggest, on a CF_2 mass defect values ≥ 0.80 if a peak was potentially heavily halogenated. It suggested that 32.64% and 52.35% of clusters can have ≥ 1 heavily halogenated peaks using the IDSL.ExposomeDB and ECS IPDBs, respectively (Figures S.10(a-b)). Further validation of the identity of these putative halogenated compounds is required considering they were detected in pregnant women's blood specimens. These findings also show that IDSL.UFA software can detect chemicals of public health concerns in a human biospecimen and can expand the existing database of exposome chemicals.[30]

To flag the potential peaks related to chemical exposures, we first assigned molecular formula using an ECS that may cover diverse halogenated compounds that were not found in the IDSL.ExposomDB formula list. IDSL.UFA resulted with 199,837 unique molecular formulas on the aligned table (top rank ≤ 30 and number of hits ≤ 30) in the ST001430 study. Grouping these formulas by class detection rules highlighted that 7,615, 18,452, and 32,107 distinct formula classes can be associated with polymeric chain increment ($\Delta H/\Delta C = 2$), cyclic chain increment ($\Delta H/\Delta C = 1/2$) and constant numbers of heteroatoms ($\Sigma(H+Br+Cl+F+I)$), respectively. For instance, a class of heavily halogenated compounds, $C_nHClF_{2n}O_4$ (n=10-12), known as chlorinated perfluorotriether alcohols (Cl-PFTrEAs) was detected for human specimens in this study. Cl-PFTrEAs was previously only reported in air samples from eastern China.[40] These compounds may represent a new ubiquitous global contaminant class. IDSL.UFA can only confirm isotopic profiles match (Figure S.11); however, a confirmatory fragment ($[M-C_3F_6O]^-$) was observed for the chemistry of C_{12} in MS1 which was consistent with the published MS/MS fragmentation (Figure S.12).[40] Levels of Cl-PFTrEAs were similar to the commonly known legacy halogenated compounds[14] for human serum samples (Figure 6).

Conclusion

IDSL.UFA enabled a comprehensive characterization of the chemical space that was detected by an untargeted LC/HRMS assay to study the metabolome and exposome and its role in human health. The unique feature of the IDSL.UFA software is to utilize statistics of rank and frequently of detected molecular formulas on the aligned annotated molecular formula table to confidently assign molecular formulas. It can complement the other peak annotation efforts that use mainly MS/MS data to annotate peak, and thus lower the number of false negative reporting of peaks and minimize the under-utilization of the untargeted LC/HRMS datasets. We provided various scenarios to obtain molecular formula from a known database and enumeration strategies to assign a formula to almost every peak in a LC/HRMS dataset. These new computational strategies for molecular formula assignment can greatly expand the quality of untargeted LC/HRMS data matrices and their analyses especially when MS/MS data are not available.

References

1. Needham, B. D.; Adame, M. D.; Serena, G.; Rose, D. R.; Preston, G. M.; Conrad, M. C.; Campbell, A. S.; Donabedian, D. H.; Fasano, A.; Ashwood, P.; Mazmanian, S. K., *Biol Psychiatry* **2021,** *89* (5), 451-462.
2. Gonzalez-Dominguez, R.; Jauregui, O.; Queipo-Ortuno, M. I.; Andres-Lacueva, C., *Anal Chem* **2020,** *92* (20), 13767-13775.
3. Shen, B.; Yi, X.; Sun, Y.; Bi, X.; Du, J.; Zhang, C.; Quan, S.; Zhang, F.; Sun, R.; Qian, L.; Ge, W.; Liu, W.; Liang, S.; Chen, H.; Zhang, Y.; Li, J.; Xu, J.; He, Z.; Chen, B.; Wang, J.; Yan, H.; Zheng, Y.; Wang, D.; Zhu, J.; Kong, Z.; Kang, Z.; Liang, X.; Ding, X.; Ruan, G.; Xiang, N.; Cai, X.; Gao, H.; Li, L.; Li, S.; Xiao, Q.; Lu, T.; Zhu, Y.; Liu, H.; Chen, H.; Guo, T., *Cell* **2020,** *182* (1), 59-72 e15.
4. Wozniak, J. M.; Mills, R. H.; Olson, J.; Caldera, J. R.; Sepich-Poore, G. D.; Carrillo-Terrazas, M.; Tsai, C. M.; Vargas, F.; Knight, R.; Dorrestein, P. C.; Liu, G. Y.; Nizet, V.; Sakoulas, G.; Rose, W.; Gonzalez, D. J., *Cell* **2020,** *182* (5), 1311-1327 e14.
5. Wang, L. B.; Karpova, A.; Gritsenko, M. A.; Kyle, J. E.; Cao, S.; Li, Y.; Rykunov, D.; Colaprico, A.; Rothstein, J. H.; Hong, R.; Stathias, V.; Cornwell, M.; Petralia, F.; Wu, Y.; Reva, B.; Krug, K.; Pugliese, P.; Kawaler, E.; Olsen, L. K.; Liang, W. W.; Song, X.; Dou, Y.; Wendl, M. C.; Caravan, W.; Liu, W.; Cui Zhou, D.; Ji, J.; Tsai, C. F.; Petyuk, V. A.; Moon, J.; Ma, W.; Chu, R. K.; Weitz, K. K.; Moore, R. J.; Monroe, M. E.; Zhao, R.; Yang, X.; Yoo, S.; Krek, A.; Demopoulos, A.; Zhu, H.; Wyczalkowski, M. A.; McMichael, J. F.; Henderson, B. L.; Lindgren, C. M.; Boekweg, H.; Lu, S.; Baral, J.; Yao, L.; Stratton, K. G.; Bramer, L. M.; Zink, E.; Couvillion, S. P.; Bloodsworth, K. J.; Satpathy, S.; Sieh, W.; Boca, S. M.; Schurer, S.; Chen, F.; Wiznerowicz, M.; Ketchum, K. A.; Boja, E. S.; Kinsinger, C. R.; Robles, A. I.; Hiltke, T.; Thiagarajan, M.; Nesvizhskii, A. I.; Zhang, B.; Mani, D. R.; Ceccarelli, M.; Chen, X. S.; Cottingham, S. L.; Li, Q. K.; Kim, A. H.; Fenyo, D.; Ruggles, K. V.; Rodriguez, H.; Mesri, M.; Payne, S. H.; Resnick, A. C.; Wang, P.; Smith, R. D.; Iavarone, A.; Chheda, M. G.; Barnholtz-Sloan, J. S.; Rodland, K. D.; Liu, T.; Ding, L.; Clinical Proteomic Tumor Analysis, C., *Cancer Cell* **2021,** *39* (4), 509-528 e20.
6. Franzosa, E. A.; Sirota-Madi, A.; Avila-Pacheco, J.; Fornelos, N.; Haiser, H. J.; Reinker, S.; Vatanen, T.; Hall, A. B.; Mallick, H.; McIver, L. J.; Sauk, J. S.; Wilson, R. G.; Stevens, B. W.; Scott, J. M.; Pierce, K.; Deik, A. A.; Bullock, K.; Imhann, F.; Porter, J. A.; Zhernakova, A.; Fu, J.; Weersma, R. K.; Wijmenga, C.; Clish, C. B.; Vlamakis, H.; Huttenhower, C.; Xavier, R. J., *Nat Microbiol* **2019,** *4* (2), 293-305.
7. Shen, X.; Wang, R.; Xiong, X.; Yin, Y.; Cai, Y.; Ma, Z.; Liu, N.; Zhu, Z. J., *Nat Commun* **2019,** *10* (1), 1516.
8. Wang, L.; Xing, X.; Chen, L.; Yang, L.; Su, X.; Rabitz, H.; Lu, W.; Rabinowitz, J. D., *Anal Chem* **2019,** *91* (3), 1838-1846.
9. Senan, O.; Aguilar-Mogas, A.; Navarro, M.; Capellades, J.; Noon, L.; Burks, D.; Yanes, O.; Guimera, R.; Sales-Pardo, M., *Bioinformatics* **2019,** *35* (20), 4089-4097.
10. Uppal, K.; Walker, D. I.; Jones, D. P., *Anal Chem* **2017,** *89* (2), 1063-1067.
11. Fakouri Baygi, S.; Fernando, S.; Hopke, P. K.; Holsen, T. M.; Crimmins, B. S., *Anal Chem* **2019,** *91* (24), 15509-15517.
12. Fakouri Baygi, S.; Crimmins, B. S.; Hopke, P. K.; Holsen, T. M., *Environ Sci Technol* **2016,** *50* (17), 9460-8.
13. Fakouri Baygi, S.; Fernando, S.; Hopke, P. K.; Holsen, T. M.; Crimmins, B. S., *Environ Sci Technol* **2020,** *54* (22), 14352-14360.
14. Fakouri Baygi, S.; Fernando, S.; Hopke, P. K.; Holsen, T. M.; Crimmins, B. S., *Environ Sci Technol* **2021,** *55* (6), 3765-3774.
15. Fakouri Baygi, S.; Hutinet, S.; Cariou, R.; Fernando, S.; Hopke, P. K.; Holsen, T. M.; Crimmins, B. S., *Int J Environ Sci Technol* **2022,** *27*.

16. Duhrkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Bocker, S., *Nat Methods* **2019,** *16* (4), 299-302.
17. Ludwig, M.; Nothias, L. F.; Duhrkop, K.; Koester, I.; Fleischauer, M.; Hoffmann, M. A.; Petras, D.; Vargas, F.; Morsy, M.; Aluwihare, L.; Dorrestein, P. C.; Bocker, S., *Nat Mach Intell* **2020,** *2* (10), 629-+.
18. Chen, L.; Lu, W.; Wang, L.; Xing, X.; Chen, Z.; Teng, X.; Zeng, X.; Muscarella, A. D.; Shen, Y.; Cowan, A.; McReynolds, M. R.; Kennedy, B. J.; Lato, A. M.; Campagna, S. R.; Singh, M.; Rabinowitz, J. D., *Nat Methods* **2021,** *18* (11), 1377-1385.
19. Leon, A.; Cariou, R.; Hutinet, S.; Hurel, J.; Guitton, Y.; Tixier, C.; Munschy, C.; Antignac, J. P.; Dervilly-Pinel, G.; Le Bizec, B., *Anal Chem* **2019,** *91* (5), 3500-3507.
20. Schum, S. K.; Brown, L. E.; Mazzoleni, L. R., *Environ Res* **2020,** *191*, 110114.
21. Stanstrup, J.; Broeckling, C. D.; Helmus, R.; Hoffmann, N.; Mathe, E.; Naake, T.; Nicolotti, L.; Peters, K.; Rainer, J.; Salek, R. M.; Schulze, T.; Schymanski, E. L.; Stravs, M. A.; Thevenot, E. A.; Treutler, H.; Weber, R. J. M.; Willighagen, E.; Witting, M.; Neumann, S., *Metabolites* **2019,** *9* (10).
22. Alfano, R.; Chadeau-Hyam, M.; Ghantous, A.; Keski-Rahkonen, P.; Chatzi, L.; Perez, A. E.; Herceg, Z.; Kogevinas, M.; de Kok, T. M.; Nawrot, T. S.; Novoloaca, A.; Patel, C. J.; Pizzi, C.; Robinot, N.; Rusconi, F.; Scalbert, A.; Sunyer, J.; Vermeulen, R.; Vrijheid, M.; Vineis, P.; Robinson, O.; Plusquin, M., *Metabolism* **2020,** *110*, 154292.
23. Wu, P.; Chen, D.; Ding, W.; Wu, P.; Hou, H.; Bai, Y.; Zhou, Y.; Li, K.; Xiang, S.; Liu, P.; Ju, J.; Guo, E.; Liu, J.; Yang, B.; Fan, J.; He, L.; Sun, Z.; Feng, L.; Wang, J.; Wu, T.; Wang, H.; Cheng, J.; Xing, H.; Meng, Y.; Li, Y.; Zhang, Y.; Luo, H.; Xie, G.; Lan, X.; Tao, Y.; Li, J.; Yuan, H.; Huang, K.; Sun, W.; Qian, X.; Li, Z.; Huang, M.; Ding, P.; Wang, H.; Qiu, J.; Wang, F.; Wang, S.; Zhu, J.; Ding, X.; Chai, C.; Liang, L.; Wang, X.; Luo, L.; Sun, Y.; Yang, Y.; Zhuang, Z.; Li, T.; Tian, L.; Zhang, S.; Zhu, L.; Chang, A.; Chen, L.; Wu, Y.; Ma, X.; Chen, F.; Ren, Y.; Xu, X.; Liu, S.; Wang, J.; Yang, H.; Wang, L.; Sun, C.; Ma, D.; Jin, X.; Chen, G., *Nat Commun* **2021,** *12* (1), 4543.
24. Han, S.; Van Treuren, W.; Fischer, C. R.; Merrill, B. D.; DeFelice, B. C.; Sanchez, J. M.; Higginbottom, S. K.; Guthrie, L.; Fall, L. A.; Dodd, D.; Fischbach, M. A.; Sonnenburg, J. L., *Nature* **2021,** *595* (7867), 415-420.
25. Liang, L.; Rasmussen, M. H.; Piening, B.; Shen, X.; Chen, S.; Rost, H.; Snyder, J. K.; Tibshirani, R.; Skotte, L.; Lee, N. C.; Contrepois, K.; Feenstra, B.; Zackriah, H.; Snyder, M.; Melbye, M., *Cell* **2020,** *181* (7), 1680-1692 e15.
26. Barupal, D. K.; Zhang, Y.; Shen, T.; Fan, S.; Roberts, B. S.; Fitzgerald, P.; Wancewicz, B.; Valdiviez, L.; Wohlgemuth, G.; Byram, G.; Choy, Y. Y.; Haffner, B.; Showalter, M. R.; Vaniya, A.; Bloszies, C. S.; Folz, J. S.; Kind, T.; Flenniken, A. M.; McKerlie, C.; Nutter, L. M. J.; Lloyd, K. C.; Fiehn, O., *Metabolites* **2019,** *9* (5).
27. Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P., *Nat Biotechnol* **2012,** *30* (10), 918-20.
28. Kim, S.; Gindulyte, A.; Zhang, J.; Thiessen, P. A.; Bolton, E. E., *Chem Teach Int* **2021,** *3* (1), 57-65.
29. Currie, L. A., *Pure Appl Chem* **1995,** *67* (10), 1699-1723.
30. Barupal, D. K.; Fiehn, O., *Environ Health Perspect* **2019,** *127* (9), 97008.
31. Fahy, E.; Subramaniam, S., *Nat Methods* **2020,** *17* (12), 1173-1174.
32. Fahy, E.; Sud, M.; Cotter, D.; Subramaniam, S., *Nucleic Acids Res* **2007,** *35* (Web Server issue), W606-12.

33. FDA structured product labelling.
34. Williams, A. J.; Grulke, C. M.; Edwards, J.; McEachran, A. D.; Mansouri, K.; Baker, N. C.; Patlewicz, G.; Shah, I.; Wambaugh, J. F.; Judson, R. S.; Richard, A. M., *J Cheminform* **2017,** *9* (1), 61.
35. Kind, T.; Fiehn, O., *BMC Bioinformatics* **2007,** *8*, 105.
36. Scrucca, L., *J Stat Softw* **2013,** *53* (4), 1-37.
37. Bengio, Y.; Delalleau, O.; Le Roux, N.; Paiement, J. F.; Vincent, P.; Ouimet, M., *Neural Comput* **2004,** *16* (10), 2197-219.
38. Hahsler, M.; Piekenbrock, M.; Doran, D., *J Stat Softw* **2019,** *91* (1), 1-30.
39. Loos, M.; Gerber, C.; Corona, F.; Hollender, J.; Singer, H., *Anal Chem* **2015,** *87* (11), 5738-44.
40. Yu, N.; Wen, H.; Wang, X.; Yamazaki, E.; Taniyasu, S.; Yamashita, N.; Yu, H.; Wei, S., *Environ Sci Technol* **2020,** *54* (6), 3103-3113.

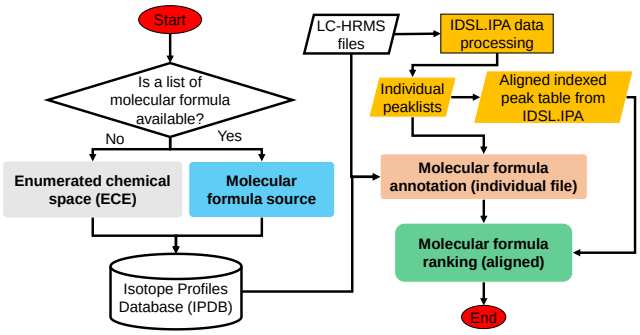

Figure 1. A simplified flowchart of the IDSL.UFA software.

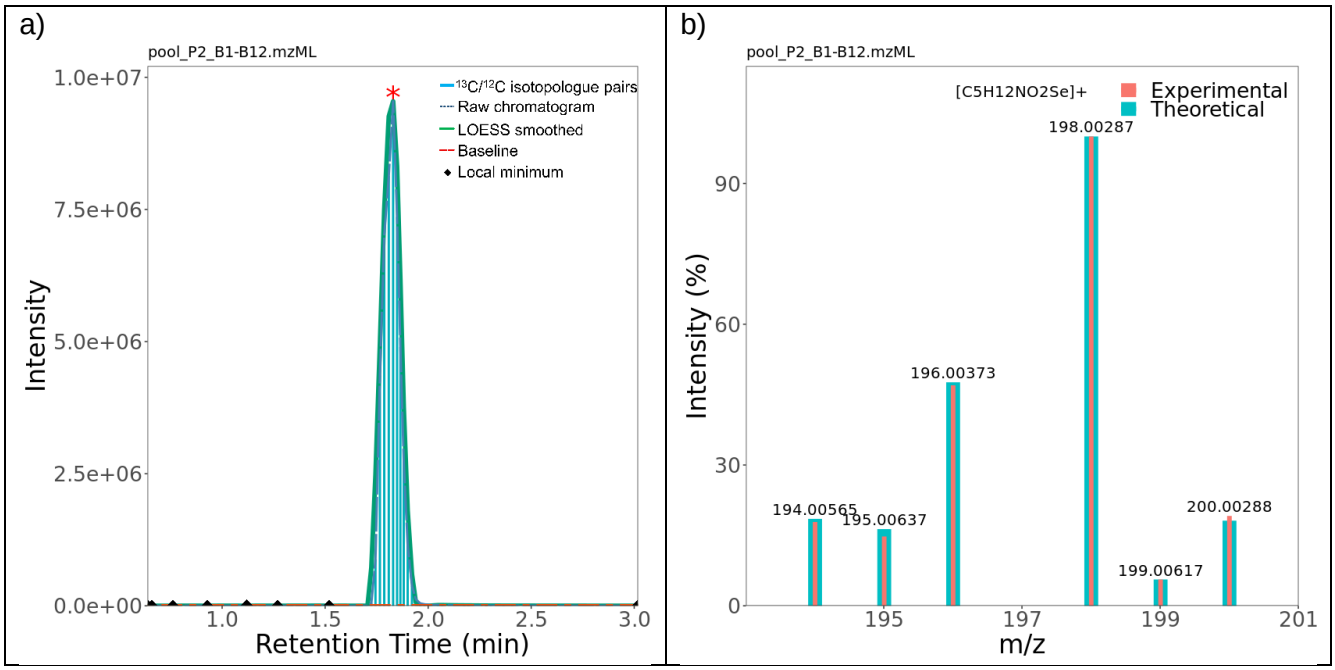

Figure 2. a) A chromatographic peak generated by the IDSL.IPA pipeline for Selenomethionine ion ($[C_5H_{12}NO_2Se]^+$) to detect peak boundaries. b) Comparison between the theoretical isotopic profile and integrated spectra across the chromatographic peak after molecular formula annotation using IDSL.UFA.

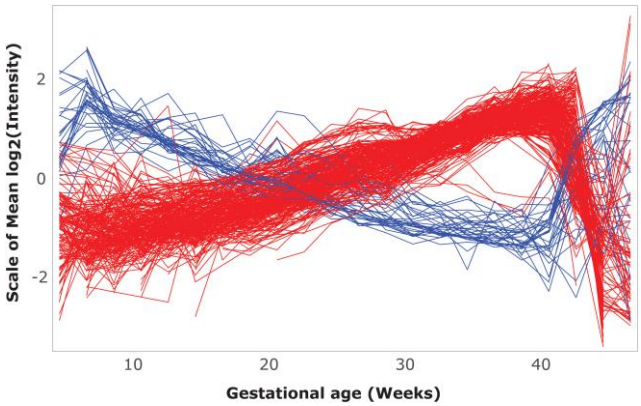

Figure 3. Trends of 274 peaks associated with pregnancy dynamics. (A user interactive plot is available at https://idsl.me/trends.)

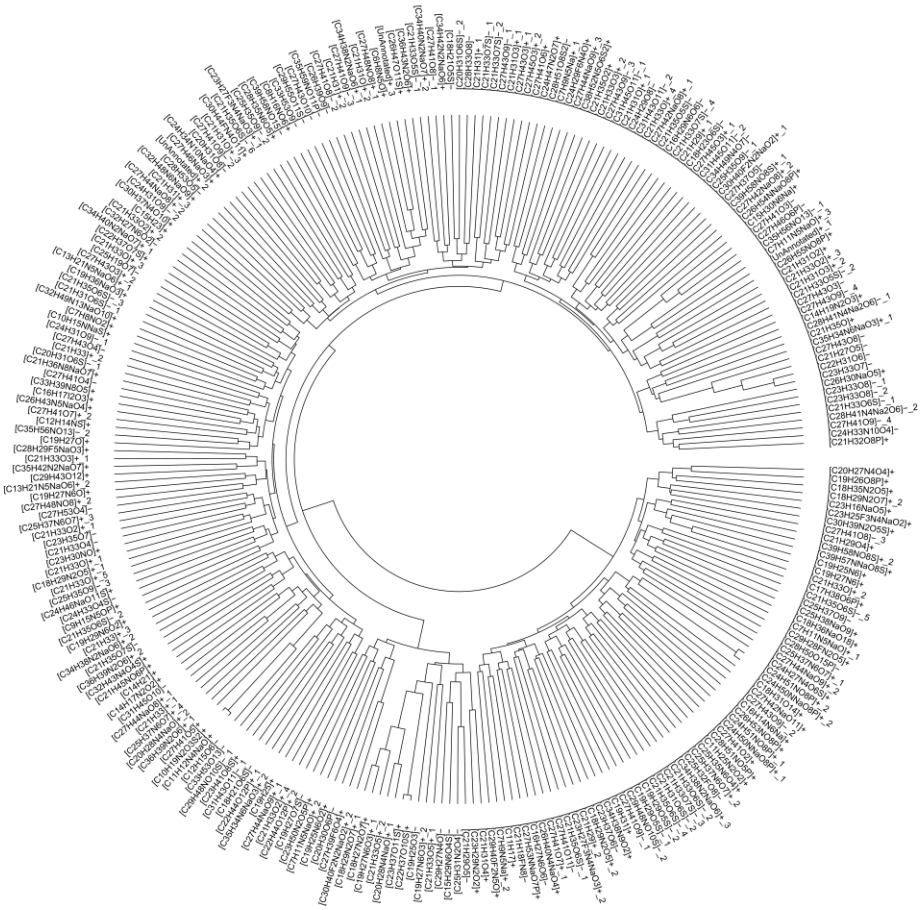

Figure 4. Unsupervised clustering of the related peaks represented by the most frequent candidate molecular formulas detected by IDSL.UFA from IDSL.ExposomeDB. These peaks have shown correlation with gestational age in at least 10 subjects (n = 30).

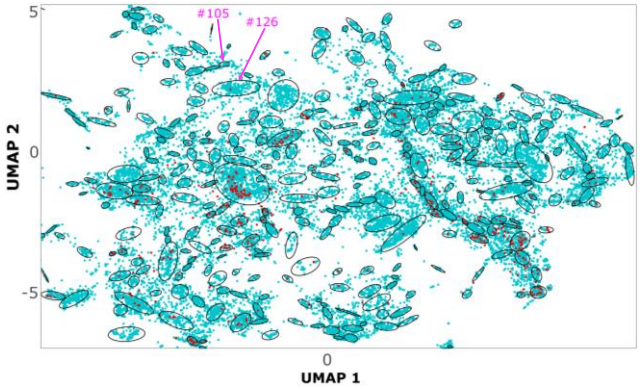

Figure 5. Clustered UMAP metabolome network of pregnancy study consisting of 340 clusters. Heavily halogenated compounds (CF_2 mass defect ≥ 0.80) represented by red points. (Two user interactive figures are available at https://idsl.me/umap with annotated peaks from the IDSL.ExposomeDB and ECS.)

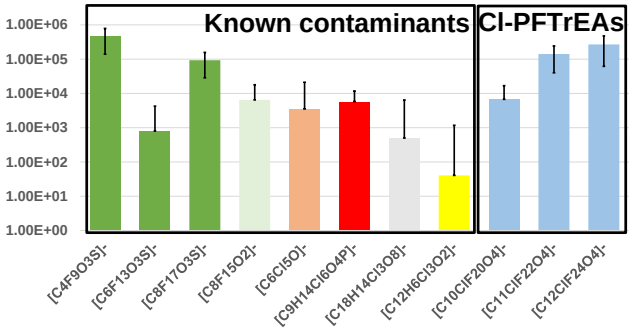

Figure 6. Peak area of halogenated contaminants in human blood ($[C_nF_{2n+1}O_3S]^-$ (n = 4, 6, 8), $[C_8F_{15}O_2]^-$, $[C_6Cl_5O]^-$, $[C_9H_{14}Cl_6O_4P]^-$, $[C_{18}H_{14}Cl_3O_8]^-$, $[C_{12}H_6Cl_3O_2]^-$) and Cl-PFTrEAs$[C_nClF_{2n}O_4]^-$ (n = 10-12) across 781 negative samples in the ST001430 study.