SUPPORTING INFORMATION

Prioritizing Chemical Candidates from Non-Targeted Analysis Using Metadata, Spectral Similarity, and Hazard Scoring within INTERPRET NTA

Alex Chao[1‡*], Jeffrey M. Minucci[2‡*], Troy M. Ferland[1,3], E. Tyler Carr[1,4], Greg Janesch[1,4], Safia Rizwan[1,4], Heather D. Whitehead[1], Tommy Cathey[5], Shirley Pu[1,3], Laura D. Brunelle[3,6], Angela L. Batt[6], Jon R. Sobus[1*], and Antony J. Williams[1*]

[1]United States Environmental Protection Agency, Office of Research and Development, Center for Computational Toxicology and Exposure, 109 TW Alexander Dr., Research Triangle Park, NC 27711, United States

[2]United States Environmental Protection Agency, Office of Research and Development, Center for Public Health and Environmental Assessment, 109 TW Alexander Dr., Research Triangle Park, NC 27711, United States

[3]Oak Ridge Institute for Science and Education (ORISE) Participant, Oak Ridge, TN 37831, USA

[4]Oak Ridge Affiliated Universities (ORAU) Student Services Contractor, 109 T.W Alexander Drive, Research Triangle Park, NC 27711, United States

[5]General Dynamics Information Technology, 109 TW Alexander Dr., Research Triangle Park, NC 27711, United States

[6]United States Environmental Protection Agency, Office of Research and Development, Center for Environmental Solutions and Emergency Response, 26 W Martin Luther King Dr., Cincinnati, OH 45268, United States

[‡] Authors contributed equally

[*]Authors to whom correspondence should be addressed:
Alex Chao: chao.alex@epa.gov
Jeffrey Minucci: minucci.jeffrey@epa.gov
Jon Sobus: sobus.jon@epa.gov
Antony Williams: williams.antony@epa.gov

Supporting Information Table of Contents

Section 1.0 – Experimental Details

The following sections provide further experimental details associated with this study.

Section 1.1 – INTERPRET NTA and Associated Database Details

INTERPRET NTA Calculation Engine

INTERPRET NTA (hereafter referred to as I-NTA) is a module of US EPA's Analytical Methods and Open Spectra (AMOS) web application. The backend components of I-NTA (also known as the Calculation Engine) were developed in Python using Django for application programming interfaces (APIs) and the Dask library[1] for distributed and asynchronous processing. When a user submits their data and desired settings to an I-NTA workflow on AMOS, a request is sent to the backend API that triggers job submission to the Dask scheduler. Once the job is distributed to a Dask worker, the worker completes the required calculations and data retrieval, and the output is stored in a central MongoDB database for later retrieval by AMOS to serve to the user (**Figure S1**). The various components of I-NTA's backend are orchestrated and deployed via Kubernetes on an US EPA Enterprise Amazon Web Services (AWS) environment. The source code for I-NTA is publicly available and can be found at https://github.com/quanted/nta_app.

DSSTox MS-Ready Structures

The I-NTA module, and more generally the AMOS application, uses DSSTox as the underlying chemical substance database, facilitating chemical lookup based on measured accurate mass or predicted molecular formula.[2] DSSTox has been accessible to the public for almost a decade via the web-based application titled the "CompTox Chemicals Dashboard" (hereafter, the "Dashboard"), available at https://comptox.epa.gov/dashboard.[3] The Dashboard has been used to great effect for NTA projects, with multiple demonstrations showing benefits of metadata-based candidate prioritization.[4, 5] The majority of users have taken advantage of capabilities for structure-based queries of "MS-Ready" chemical structures via batch interface searches.[6, 7] An MS-Ready structure refers to a chemical structure that has been specifically formatted and processed to accurately represent how a molecule would be observed by mass spectrometry, meaning it is de-salted, de-solvated, and separated into individual components (when in mixtures). The MS-Ready structures associated with the DSSTox structural content are used as input for the generation of *in silico* spectra (see *CFM-ID in silico Spectra* section), which have been shown to aid compound identification in our previous NTA works.[8, 9] In recent years, NTA support provided to the community by EPA's publicly accessible tools has been limited to the Dashboard (via advanced searching capabilities) and open datasets provided as downloads (i.e., *in silico* predicted spectral data, DSSTox MS-Ready structures, etc.). I-NTA development has paralleled the efforts described above to finally integrate all tested and proven approaches for chemical identification and candidate prioritization into an application that also performs necessary NTA data quality review.

AMOS Chemical Metadata

Chemical metadata has been shown to assist the identification of "known unknowns" via ranking of candidate structures.[4, 5, 10] In previous work, we have used the following metadata: 1) number

of associated "data sources"; 2) number of publications; and 3) number of patents. Data sources are equivalent to the "chemical lists" provided in the Dashboard (https://comptox.epa.gov/dashboard/chemical-lists). At the time of writing there are 503 lists, each associated with a specific collection of chemicals registered into the DSSTox database. Lists can represent a specific regulation (e.g., BIOSOLIDS, https://comptox.epa.gov/dashboard/chemical-lists/CWA311HS), a collection of chemicals based on specific formula rules (e.g., PFAS chemicals, https://comptox.epa.gov/dashboard/chemical-lists/PFASSTRUCTV5), or data collections associated with a publication (e.g., suspect screening of chemicals in consumer products, https://comptox.epa.gov/dashboard/chemical-lists/EPACONS).[11-13] Regarding publications and patents, PubChem and PubMed counts have historically been harvested and stored within the Dashboard.[4] However, metadata refresh rates have lagged behind ongoing new chemical registration (in support of NTA projects and other agency activities). Relevant metadata have therefore been added to the AMOS database and are updated on a more frequent basis using the PubChem API. [14] The inclusion of chemical metadata into AMOS, along with analytical method documents, chemical fact sheets (harvested from domestic and international collections), and experimental spectra, has provided expanded capabilities for candidate ranking via I-NTA interactive visualizations (*vide infra*).

CFM-ID in silico Spectra

When available, chemical fragmentation (here MS^2) spectra allow for higher confidence identifications in certain NTA studies.[15] Limitations in the number of chemicals with available experimental MS^2 spectra have led to the generation of *in silico* MS^2 spectra that provide wider chemical coverage in NTA experiments. Our previous work documented the generation of *in silico* MS^2 spectra for the entirety of chemicals within the DSSTox database using the CFM-ID algorithms.[8, 16] It further reported on the performance of this "CFM-ID database" for accurately identifying chemicals in an NTA study.[5, 9, 17] At the time of writing, the CFM-ID database contains predictions generated via CFM-ID for 1.04 million MS-Ready chemical structures; these predictions cover ESI+, ESI-, and EI spectra. While the work reported here demonstrates the utility of the predicted ESI spectra for chemical identification, future work is intended to similarly highlight value of the predicted EI spectra.

CHM Hazard and Exposure Values

In our earliest NTA work, we used *in vitro* bioactivity data and exposure predictions to prioritize chemicals tentatively identified in household dust and drinking water.[18, 19] At that time, our cheminformatics applications and underlying databases were in their infancy. Over the past decade they have advanced significantly to include additional exposure and toxicity information (available databases include Chemical and Products Database [CPDat], Multimedia Monitoring Database [MMDB], Toxicity Forecasting Database [ToxCast], Toxicity Value Database [ToxVal], and Toxicity Reference Database [ToxRef]).[20-24] While the majority of these data are available via the Dashboard in some form, I-NTA primarily utilizes data assembled into the Cheminformatics Hazard Module (CHM), which is part of EPA's publicly available Cheminformatics Modules (https://hcd.rtpnc.epa.gov/#/hazard). CHM data are most desirable for NTA projects, as refreshes are more frequent, with different data streams updated at least annually, allowing more current

chemical information to be utilized for candidate prioritization. CHM is a web-based application (expanding on the work of Vegosen and Martin)[25] that delivers access to human health, ecotoxicity, environmental fate, and exposure information (**Table S4**). It has previously been applied to perform priority-based ranking of chemical candidates identified by NTA. CHM integrates both experimental and predicted (based on molecular structure) toxicity data, along with predicted fate and exposure values, to assess potential risks associated with a given chemical substance. As discussed below, the collective information available from the CHM (hereafter termed "hazard values" for brevity) can be used to prioritize NTA suspects for further investigation.

Section 1.2 – Data Analysis and Results Scoring

NTA Dataset for Proof-of-Concept Demonstration

Study details and initial findings were first described in Brunelle *et al*. (2024).[26] The data used here are based on the subsequent I-NTA QA/QC demonstration work of Sobus *et al.* (2025).[27] Briefly, passive organic chemical integrative samplers (POCIS) were deployed for 27-31 days during one of three sampling campaigns – October 2014, April 2015, or August 2015. POCIS samples were collected at one of four locations, including upstream of a wastewater treatment plant, within the wastewater effluent mixing zone, at the intake tap of a drinking water treatment plant, and at the drinking water tap of the drinking water treatment plant. Samples were extracted using methanol and analyzed using ultra-performance liquid chromatography (UPLC) coupled with HRMS in ESI+ and ESI- modes.

MS^1 data processing

The generation of preprocessed MS^1 feature matrices (one each for ESI+ and ESI- data) using MZmine 3 was previously described.[27] The outputs from MZmine 3 were used as inputs for the I-NTA MS^1 workflow and are provided in Sobus *et al.* (2025).[27] For the current work, MS^1 workflow parameters exactly follow that of Sobus *et al.* (2025), but with additional parameters selected for chemical database searching. Specifically, MS^1 features that met QA/QC criteria were searched by mass (7 ppm tolerance) against the MS-Ready masses of DSSTox structures, retrieving a list of chemical candidates with their associated PubChem (sources, patents, and articles) and PubMed (articles) counts. These candidates were searched by DTXSID against CHM in order to retrieve available hazard values. Full analysis parameters for I-NTA MS^1 data processing are given in **Table S5**.

Additional database searches

Chemical results from the I-NTA MS^1 workflow output were batch searched, via DTXSID, within AMOS to retrieve counts of fact sheets, analytical methods, and experimental spectra associated with each chemical candidate. DTXSIDs were also searched against the Dashboard to identify presence on 50 chemical lists associated with water as a matrix (**Table S6**).

MS^2 data processing

MS^2 spectral data were preprocessed and aligned via Progenesis QI (Nonlinear Dynamics, v2.3) and exported in .msp format.[26] The export files were inputs for the I-NTA MS^2 workflow, which accepts .msp and .mgf files exported from either preprocessed aligned data or individual sample runs (as described in **SI Section 2.0**). Within the MS^2 workflow, input files were parsed to extract a list of individual MS^2 features, where each MS^2 feature is an MS^2 spectrum with an associated precursor mass, retention time, and adduct assignment. When present, adduct assignments were used to calculate a neutral mass from the precursor mass (see **Table S7**); otherwise, neutral masses were calculated with the assumption of $[M+H]^+$ and $[M-H]^-$ ions for positive and negative mode data, respectively. MS^2 features were deduplicated using neutral mass and retention time windows of 10 ppm and 0.2 min, respectively; for any duplicate MS^2 features, the instance with the highest total sum intensity of fragment peaks was retained. Neutral masses of deduplicated MS^2 features were searched against the CFM-ID database using a 10-ppm mass search window. Chemical candidates were retrieved with associated *in silico* spectra at collision energy (CE) levels of 10, 20, and 40 eV. Full analysis parameters for I-NTA MS^2 data processing are given in **Table S5**.

Scoring INTERPRET NTA Outputs

Chemical metadata scoring

Metadata categories used for candidate scoring include (1) PubChem source counts (i.e., the number of depositors in PubChem associated with a chemical)[28] , (2) PubChem patent counts, (3) PubChem article counts, (4) PubMed article counts, (5) AMOS experimental spectra counts, (6) AMOS analytical method counts, (7) AMOS fact sheet counts, and (8) Dashboard water list counts (see **Table S6** for considered Dashboard lists). A total metadata score was calculated and assigned to each MS-Ready candidate returned from the I-NTA mass query. Scores were calculated via a three-step process (see **SI section 3.0** for an illustrative example). First, for each MS-Ready candidate with multiple associated substances, a sum value across substances was separately calculated for each metadata category. Second, each metadata category score was normalized to a value between 0 and 1 at the feature level (by dividing each candidate score by the max score for that feature). Third, normalized metadata values for each candidate assigned to a given feature were summed into a total metadata score, with a potential range of 0 to 8 (a value of 8 would indicate a feature candidate had the highest counts in all eight metadata categories).

MS2 in silico spectra scoring

Scoring between MS^2 features and corresponding candidates considered all three CE levels of predicted *in silico* spectra, following the method of Chao *et al.* (2020).[9] Specifically, predicted spectra were scored against experimental spectra using a cosine dot-product algorithm[29], with fragment peaks matched on a 0.02 Da window, and the three similarity scores for each candidate (from 10, 20, and 40 eV predictions) summed into a total score. For each MS^2 feature, total scores for all candidates were normalized into quotient scores by dividing each candidate score by the highest candidate score for that feature.

Hazard value scoring

Hazard value scoring was based on CHM hazard levels (i.e., very high, high, medium, low, inconclusive, no data) and authority source designations (i.e., authoritative, screening, QSAR model) across 20 endpoints (**Table S4**). Following the procedures of Newmeyer *et al.* (2024), hazard levels and authority source designations were converted to numeric values and used to calculate a quality-adjusted hazard (QAH) score for each candidate substance.[30] An accompanying completeness score was also calculated at the substance level. Briefly, average hazard and quality scores were calculated for each candidate substance, which were then multiplied to create the QAH score. Completeness scores were calculated by dividing the number of endpoints with data for each candidate substance by the total number of endpoints (n=20). Scores at the chemical substance level were collapsed to the MS-Ready chemical structure level by taking only the highest QAH score, and associated completeness score, amongst all associated chemical substances.

Ratio calculations

Using data for all *known chemical features* (i.e., those associated with the 77 *known chemicals*), distributions of ratios were calculated using metadata field values, the overall metadata scores, the MS^2 scores, and the QAH scores. For each *known chemical feature*, exactly one correct candidate exists, and a series of incorrect candidates exists. Thus, for each pairing of correct candidate and incorrect candidate, ratio values were calculated as the correct candidate score divided by the incorrect candidate score. Zero values for any score category did not allow for meaningful ratio calculations. Thus, imputations were performed in instances where scores were zero. Imputed values were set as the lowest nonzero value for a given score category divided by the square root of two. No imputation was performed if the correct candidate for a *known chemical feature* was the only candidate with a nonzero score.

Section 2.0 – Input Requirements for MS^2 Workflow

The I-NTA MS^2 workflow currently accepts .mgf and .msp formats as input files. Shown below are examples of file formats for both, as well as a description of how they are parsed by I-NTA when read in.

Example MGF format file

```
BEGIN IONS
PEPMASS=105.0423685
CHARGE=1+
TITLE=EntactEnv_Pos_MS1_Dust1IDA_01.d, MS/MS of 105.0423685  1+ at 0.04745  mins
RTINSECONDS=2.847
86.5515288395944        10.08654
86.5781883146149        11.06618
87.4602539495444        12.05921
103.348974721758        10.08654
103.996592740677        13.04167
110.133284589395        13.11111
126.802093185387        12.20833
136.071037424052        11.08654
144.893286462799        13.125
154.376888744377        10.075
157.374623516955        11.16667
158.92422332612         10
185.559080046904        14.1
194.428548644229        38.80357
214.063863547247        10.47115
216.511031353531        11.125
218.149310353634        12.06618
221.991928276189        11.00961
263.123839285519        11.68056
318.116692565632        11.03571
332.775479092216        13.00735
337.009514276059        16.00833
347.614828778035        23.07143
386.061720707835        11.00833
466.52480913612         13.11111
474.667515677988        13
END IONS
```

- "BEGIN IONS" – Indicates the beginning of a new MS^2 spectrum. This will trigger I-NTA to create a new MS^2 feature for which to populate with the relevant spectrum/feature information.

- "PEPMASS" – The numeric value in this line will be saved as the precursor mass for the MS^2 feature.
    - Example value: 105.0423685

- "RTINSECONDS" – The numeric value in this line will be saved as the retention time (in seconds).
    - Example value: 2.847

- Any numeric value at the beginning of the line – Indicates a line that has a fragment m/z, fragment intensity pair. The numeric values in the line are split, where the first is added to a list of fragment m/z values and the second is added to a list of fragment intensity values.
    - Example value: [86.5515288395944], [10.08654]

- "END IONS" – This indicates the end of an MS^2 spectrum in the input file. The precursor mass, retention time, and list of fragment m/z and intensity pairs are saved as an MS^2 feature in a list.

Example MSP format file

```
Name: Unknown (0.72_101.9637n)
Precursor_type: [M+H-H2O]+
PrecursorMZ: 84.9604
Comment: 0.72_101.9637n
Num Peaks: 25
53.0199 80.6951
61.0080 2561.4187
61.5094 102.5312
62.0240 190.7465
62.5263 46.6838
63.0048 47.1720
66.0197 485.2085
68.0167 46.0241
68.9532 51.1917
68.9831 824.8460
70.0131 2110.7041
70.5147 92.6136
70.9805 170.6099
71.0303 130.9467
72.0108 49.6613
73.5328 1239.6155
74.0334 178.1040
77.0206 159.1474
79.0187 499.7516
81.5214 2441.2715
82.0148 2040.3814
82.5377 1487.0464
83.0386 287.4457
83.5182 47.6695
84.9604 16908.8770
```

- "Name" – Indicates the beginning of a new MS^2 spectrum. This will trigger I-NTA to create a new MS^2 feature for which to populate with the relevant spectrum/feature information.

- "Precursor_type" – Where present, the text value in this line will be saved as the adduct type of the MS^2 feature. If this line is not present, the adduct type is assumed to be $[M+H]^+$ or $[M-H]^-$ for positive or negative mode, respectively.
    - Example value: "[M+H-H2O]+"

- "PrecursorMZ:" – The numeric value in this line will be saved as the precursor mass for the MS^2 feature. If there is an adduct type present for the MS2 feature, the precursor mass is adjusted (See **Table S7** for specific mass adjustments performed by I-NTA)
    - Example value: 84.9604 + 18.010565 = 102.970965

- "Comment:" – The first numeric value in this line prior to the underscore will be saved as the retention time (in minutes).
    - Example value: 0.72

- "Num Peaks:" – The numeric value in this line represents if there are fragment peaks associated with the spectrum. A "has_fragments" Boolean is created checking if this number is greater than zero.

- Any numeric value at the beginning of the line – Indicates a line that has a fragment m/z, fragment intensity pair. The numeric values in the line are split, where the first is added to a list of fragment m/z values and the second is added to a list of fragment intensity values.
    - Example value: [53.0199], [80.6951]

- Blank space line: This indicates the end of an MS^2 spectrum in the input file. If the MS^2 spectrum has any fragments (has_fragments == True), then the precursor mass, retention time, and list of fragment m/z and intensity pairs are saved as an MS^2 feature in a list.

Section 3.0 – Chemical metadata scoring process

The following text describes how overall metadata scores were calculated for chemical structures from their associated substance-related metadata retrieved from DSSTox.

Step 1: Collapse chemical substances (identified by a DTXSID) into chemical structures (identified by a DTXCID) by summing metadata fields for each substance.

All Chemical Substances Mapped from Chemical Structure DTXCID0027983

Chemical Substance Name	DTXSID	PubChem Patents	PubChem Articles	PubMed Articles	PubChem Sources	AMOS spectra	AMOS methods	AMOS fact sheets	Dash-board Water lists
Acesulfame potassium	DTXSID1030606	15842	813	286	21	12	15	1	1
Acesulfame	DTXSID0048006	27072	752	346	31	75	8	0	3
Aspartame acesulfame	DTXSID30909991	111	0	0	10	0	0	0	0
Calcium bis(6-methyl-2,2-dioxo-2H-1,2lambda~6~,3-oxathiazin-4-olate)	DTXSID50968888	25	283	283	5	0	0	0	0

Sum within each metadata field (+ + + + + + + +)

Chemical Structure (DTXCID)	PubChem Patents	PubChem Articles	PubMed Articles	PubChem Sources	AMOS spectra	AMOS methods	AMOS fact sheets	Dash-board Water lists
DTXCID0027983	43050	1848	915	67	87	23	1	4

Step 2: Within each feature, normalize all the metadata field counts for all associated chemical structures by dividing by the maximum metadata field count.

DTXCID	PubChem Patents	PubChem Articles	PubMed Articles	PubChem Sources	AMOS spectra	AMOS methods	AMOS fact sheets	Dash-board Water lists
DTXCID0027983	43050	1848	915	67	87	23	1	4
DTXCID10618574	11	0	0	2	0	0	0	0
DTXCID20585208	14	0	0	2	0	0	0	0
DTXCID60300210	1	8	0	6	3	0	0	0
DTXCID70618575	0	0	0	2	0	0	0	0
DTXCID80401596	20	0	0	3	0	0	0	0

Divide values by maximum value within each metadata field

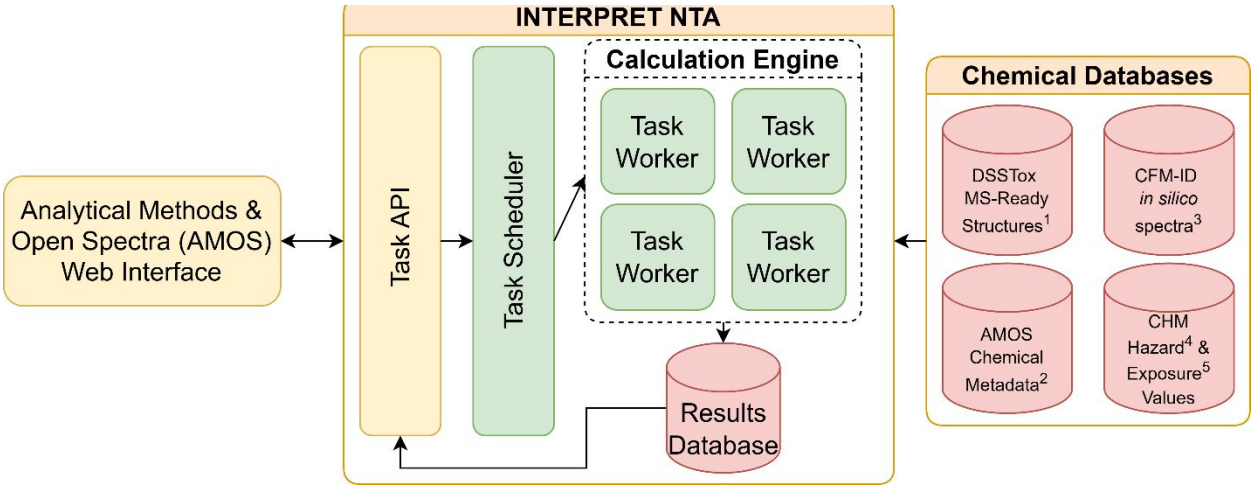
÷ 43050 ÷ 1848 ÷ 915 ÷ 67 ÷ 87 ÷ 23 ÷ 1 ÷ 4

DTXCID	PubChem Patents	PubChem Articles	PubMed Articles	PubChem Sources	AMOS spectra	AMOS methods	AMOS fact sheets	Dash-board Water lists
DTXCID0027983	1	1	1	1	1	1	1	1
DTXCID10618574	0.00026	0	0	0.030	0	0	0	0
DTXCID20585208	0.00033	0	0	0.030	0	0	0	0
DTXCID60300210	0.00002	0.0043	0	0.090	0.034	0	0	0
DTXCID70618575	0.00000	0	0	0.030	0	0	0	0
DTXCID80401596	0.00046	0	0	0.045	0	0	0	0

Step 3: Sum across the eight metadata field normalized values for each chemical structure (DTXCID) to create an overall metadata score for each chemical structure.

DTXCID	PubChem Patents	PubChem Articles	PubMed Articles	PubChem Sources	AMOS spectra	AMOS methods	AMOS fact sheets	Dash-board Water lists	*Sum values for all fields*	Overall Metadata Score
DTXCID0027983	1	1	1	1	1	1	1	1	→	8
DTXCID10618574	0.00026	0	0	0.030	0	0	0	0	→	0.030
DTXCID20585208	0.00033	0	0	0.030	0	0	0	0	→	0.030
DTXCID60300210	0.00002	0.0043	0	0.090	0.034	0	0	0	→	0.128
DTXCID70618575	0.00000	0	0	0.030	0	0	0	0	→	0.030
DTXCID80401596	0.00046	0	0	0.045	0	0	0	0	→	0.045

Section 4.0 – Supplemental Figures



[1] Includes MS-Ready chemical structures linked to parent substances and chemical identifiers[3, 7]
[2] Includes fact sheets, methods, patent counts, publication counts, and open spectra[31]
[3] Includes predicted electron ionization (EI) and electrospray ionization (ESI) spectra based on the CFM-ID algorithm[8, 9, 17]
[4] Includes hazard (measured and predicted) scores and authority levels (Table S4)[20-25, 30, 32-34]
[5] Includes exposure predictions based on the systematic empirical evaluation of models (SEEM3) framework[35]

Figure S1: Application diagram illustrating the various components of the INTERPRET NTA application and its database connections.

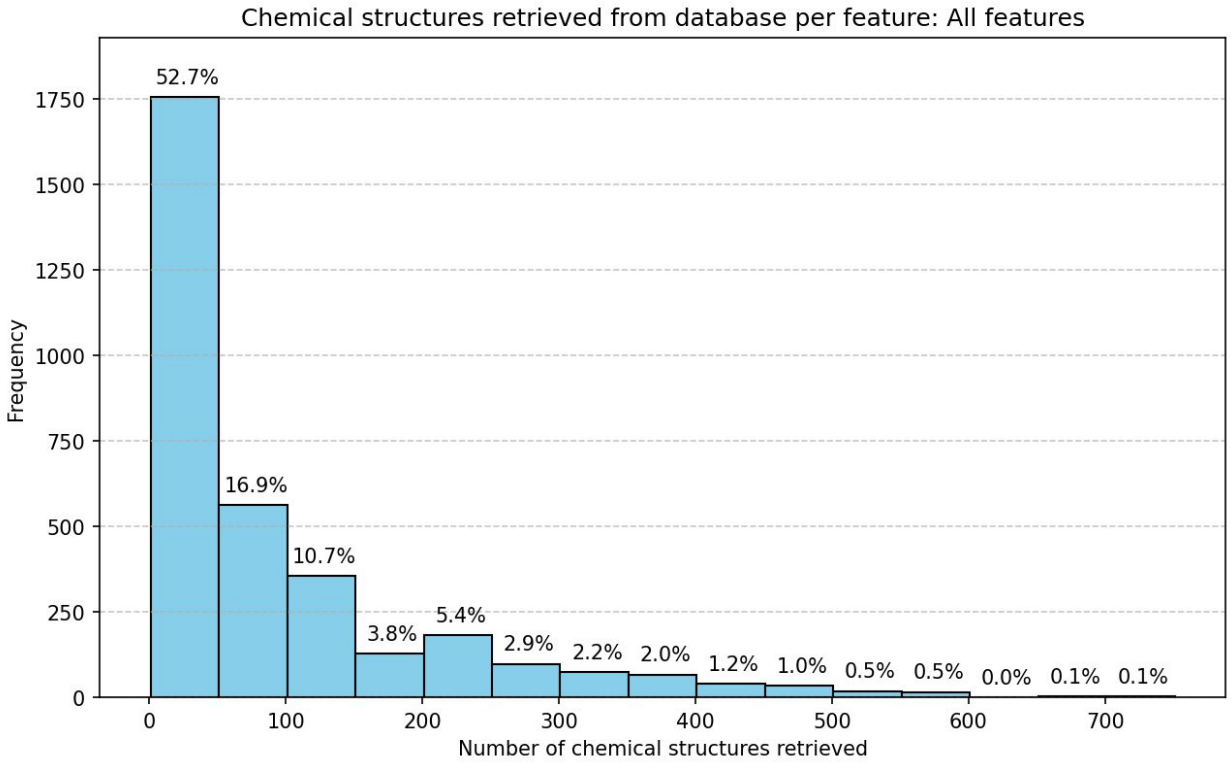


Figure S2a. The number of chemical structures retrieved for all features based on mass searches against the DSSTox chemical database.

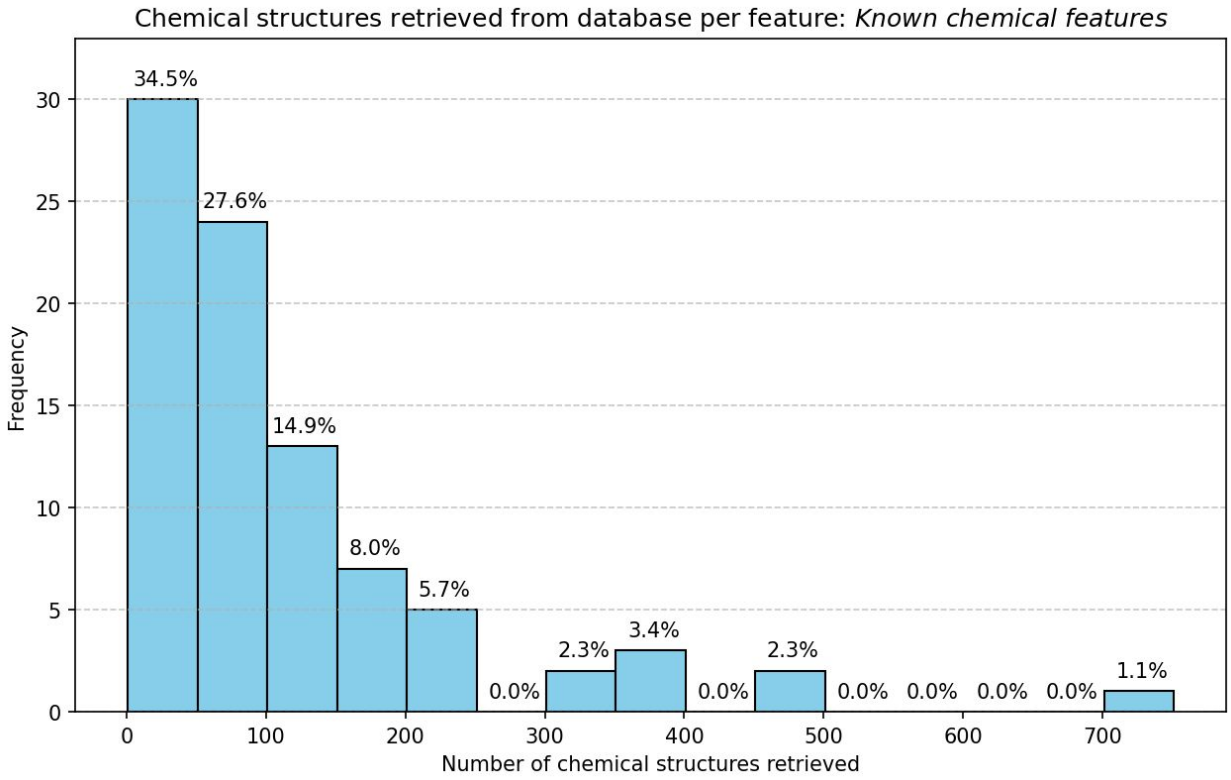


Figure S2b. The number of chemical structures retrieved for *known chemical features* based on mass searches against the DSSTox chemical database.

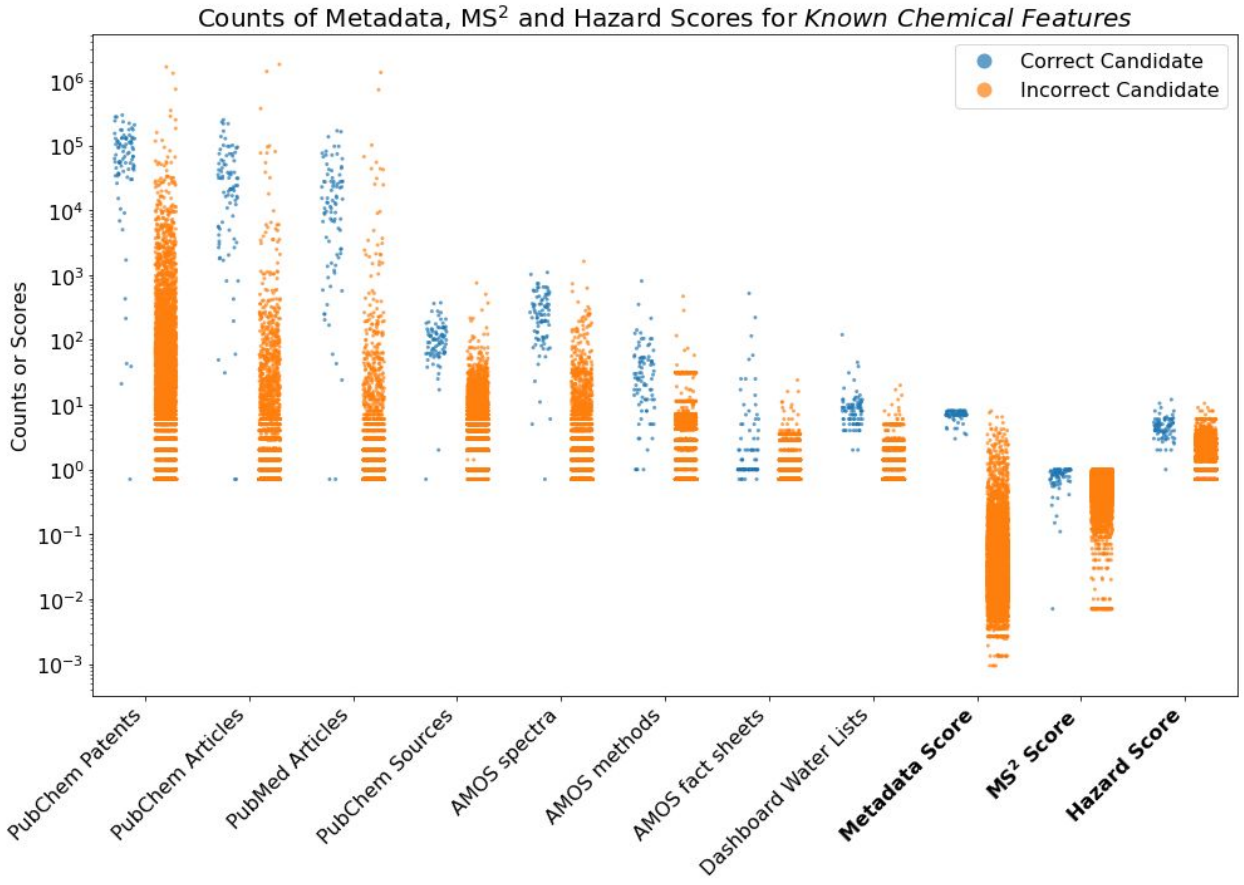


Figure S3. Absolute counts for chemical metadata fields, total metadata score, MS^2 score, and hazard score. Distributions are separated between correct and incorrect candidates for *known chemical features*.

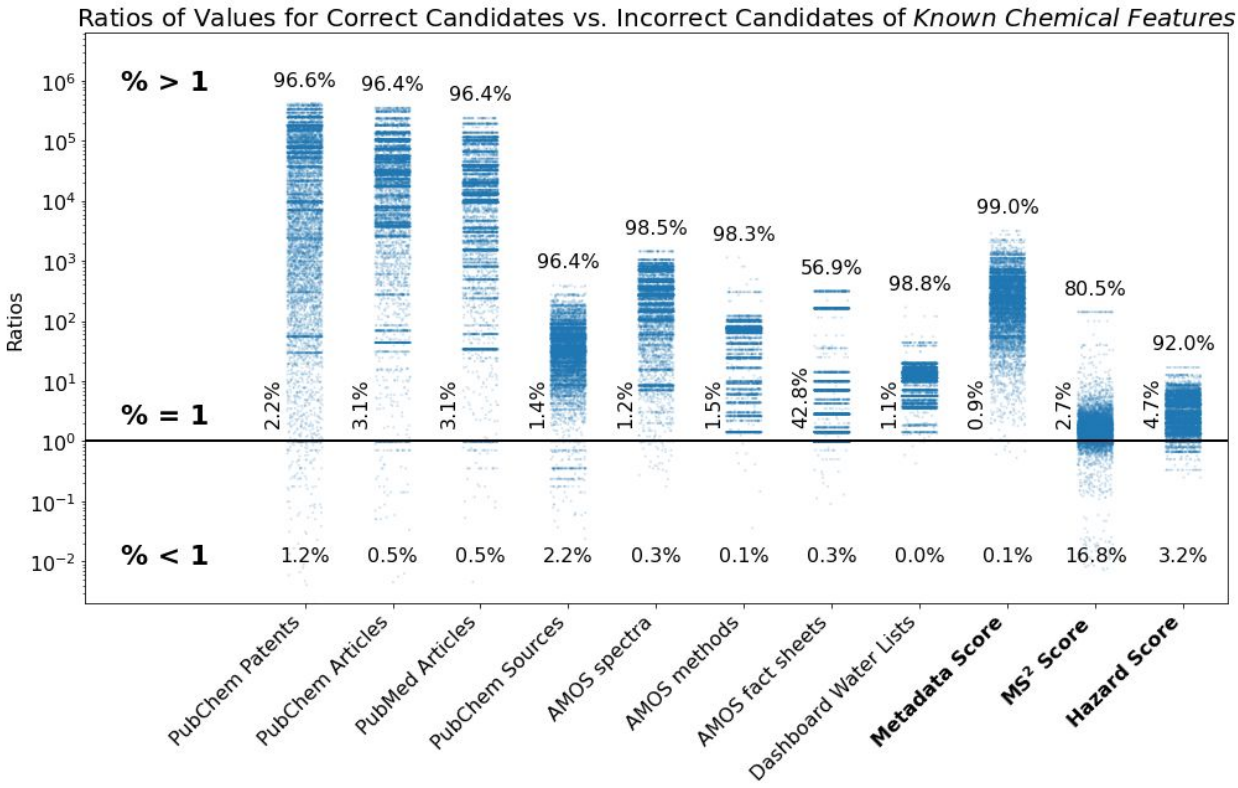


Figure S4. Score ratios between the correct candidate and all associated incorrect candidates for each *known chemical feature* (n=87). Ratios greater than one indicate higher scores for correct candidates. Percentage values indicate the proportion of ratios greater than, equal to, and less than one.

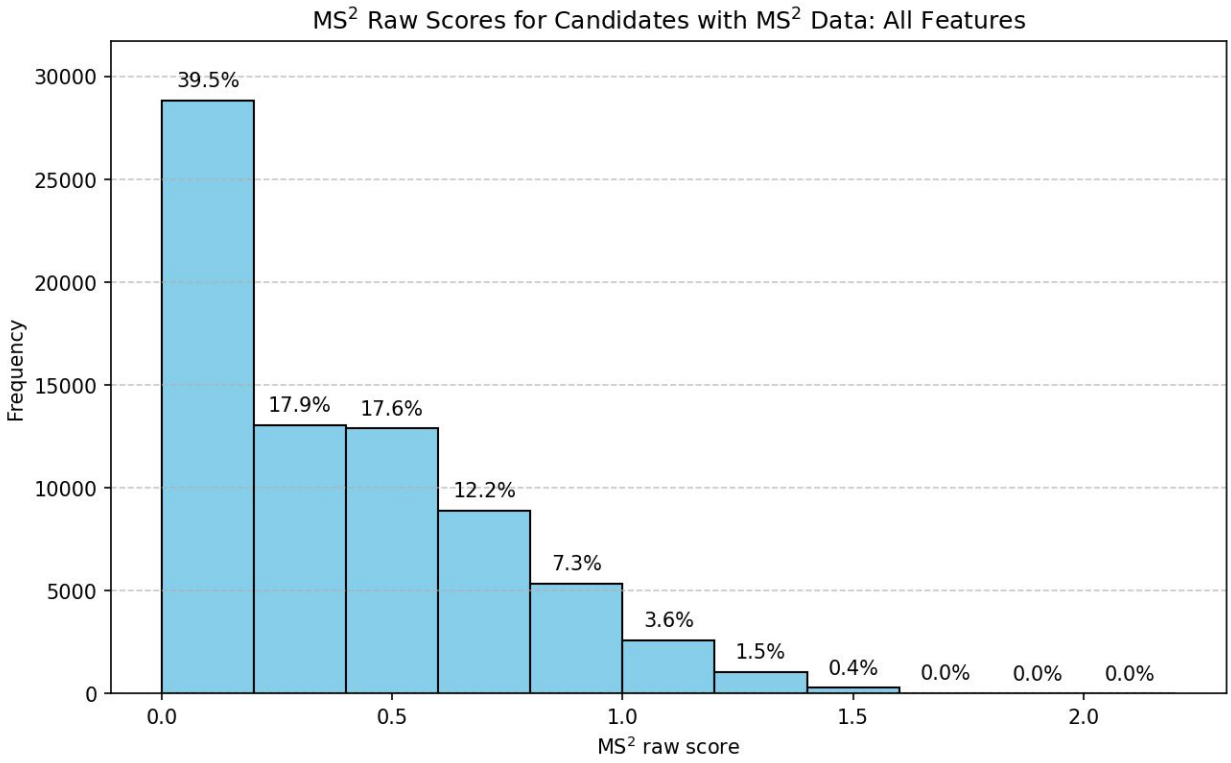


Figure S5a. Distribution of MS^2 raw scores returned for all candidates of all features with acquired MS^2 data.

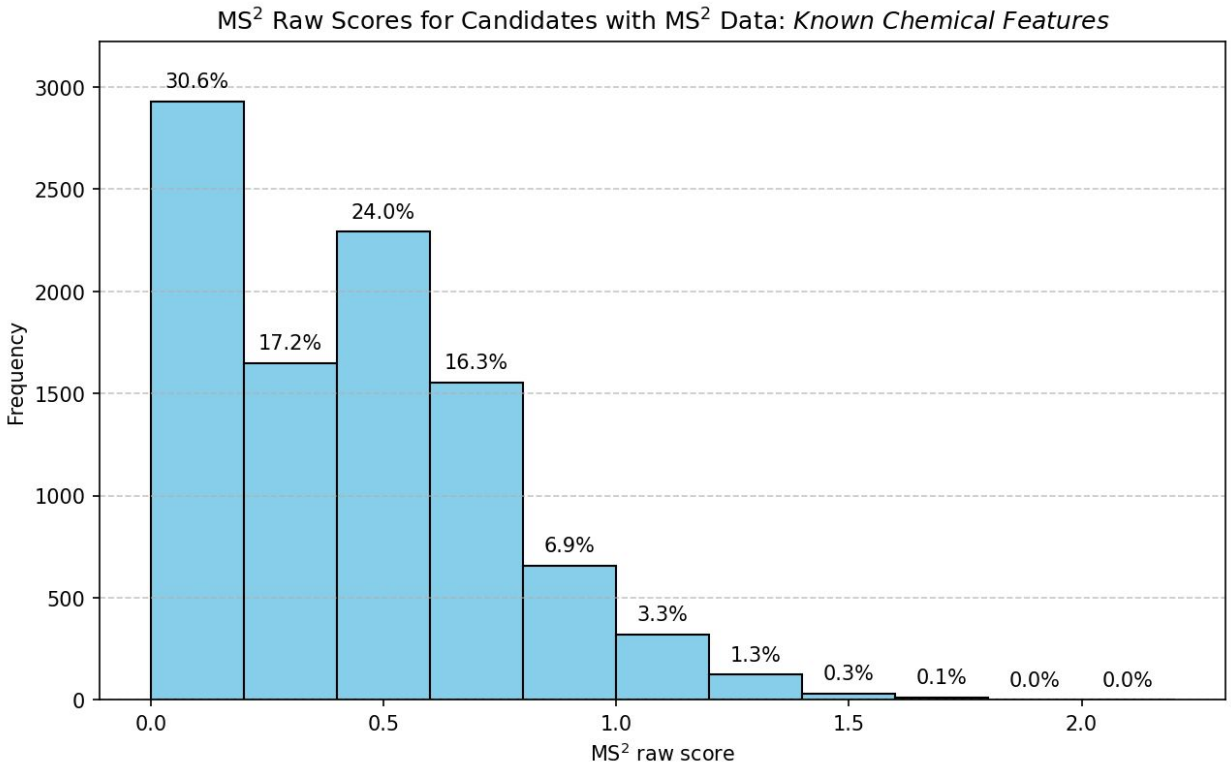


Figure S5b. Distribution of MS^2 raw scores returned for all candidates of *known chemical features* with acquired MS^2 data.

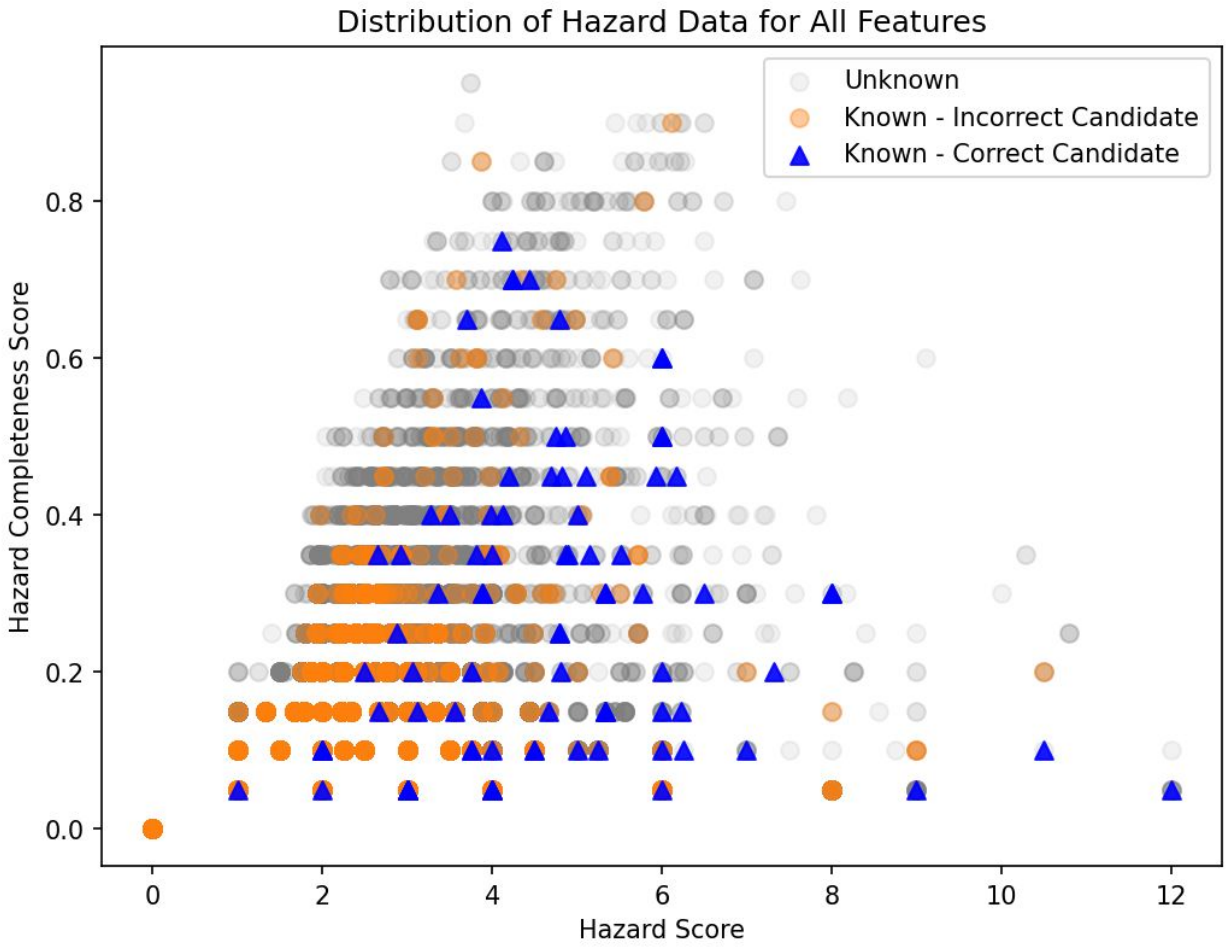


Figure S6. Distribution of calculated hazard scores for candidates of unknown chemical features (grey), and correct (blue) and incorrect (orange) candidates of *known chemical features*.

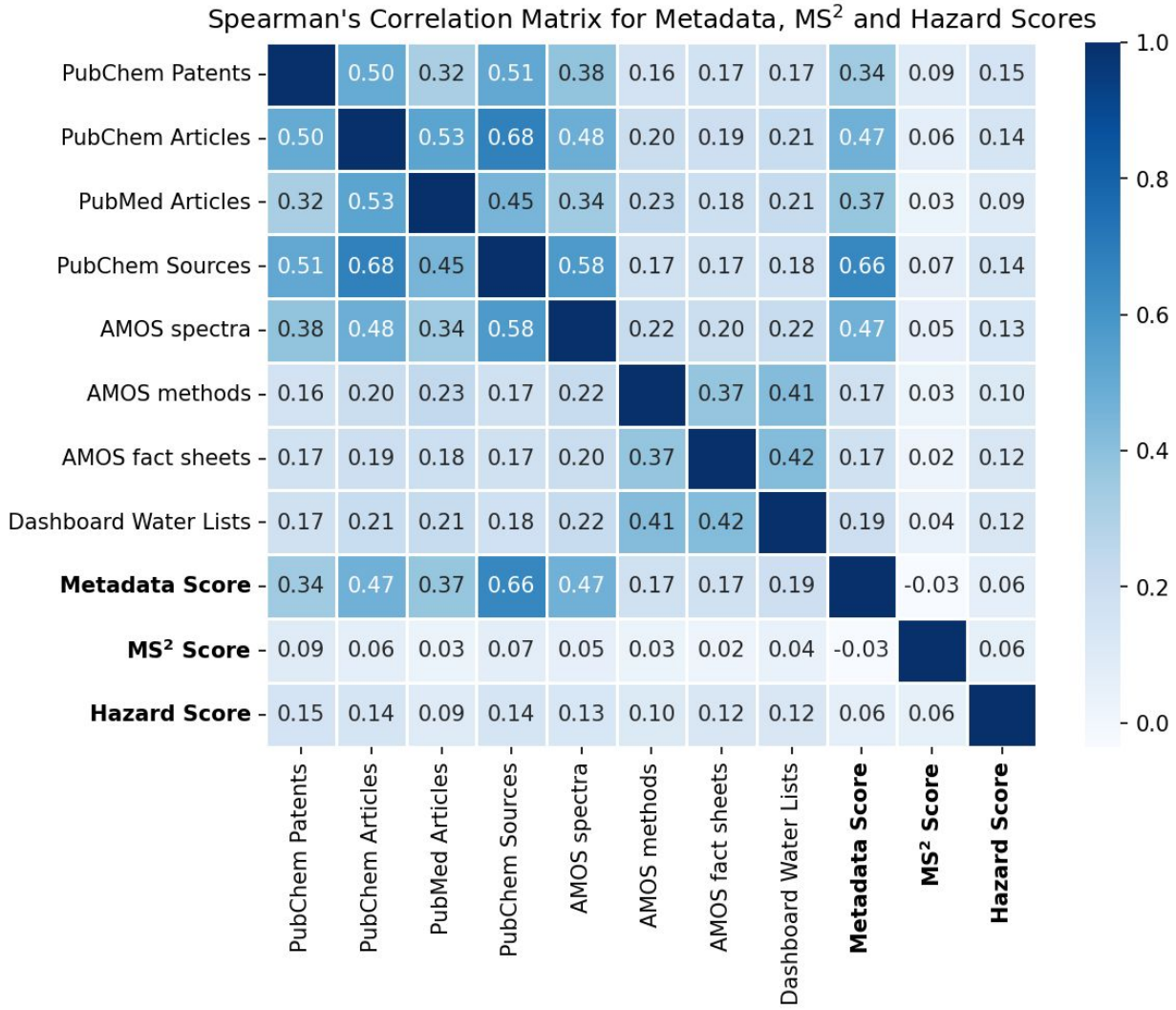
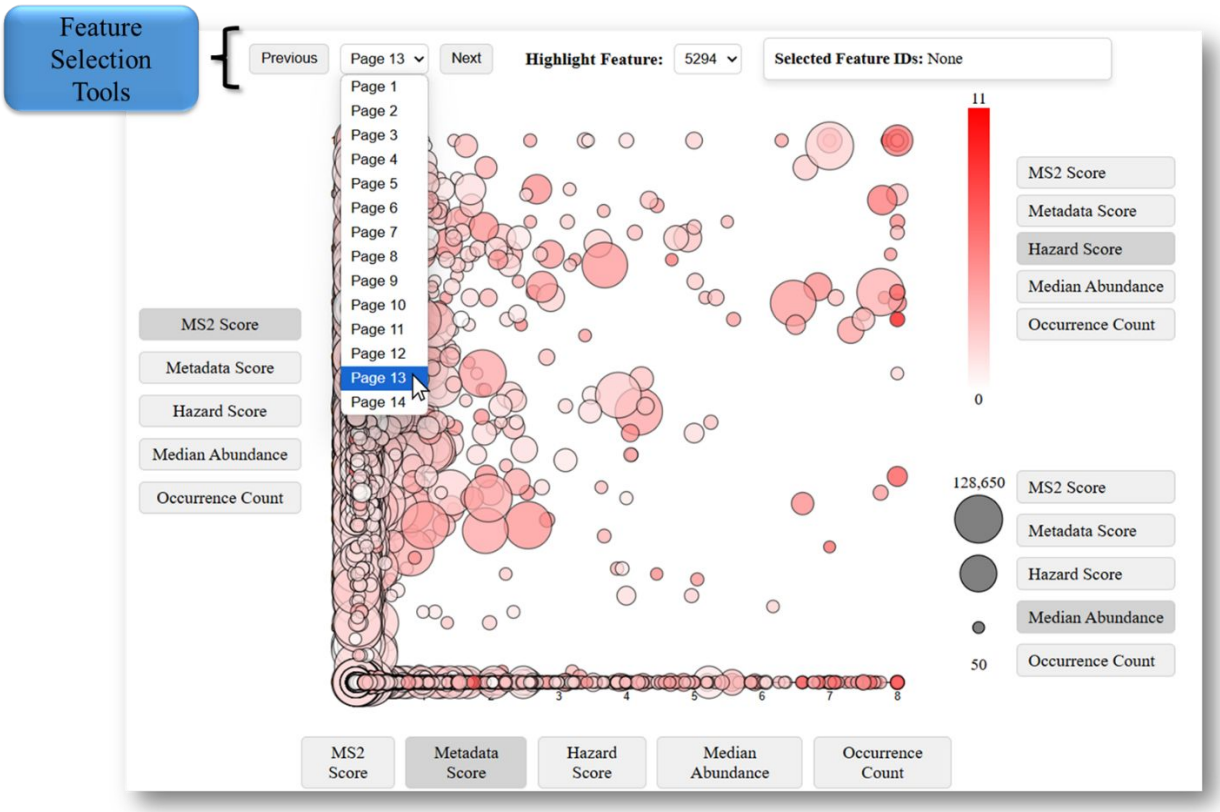


Figure S7. Spearman correlation matrix results for absolute counts of metadata fields, and the total metadata, MS^2, and hazard scores, with Spearman correlation values for each field/score pair.

Section 5.0 – Interactive Chemical Results Visualizations

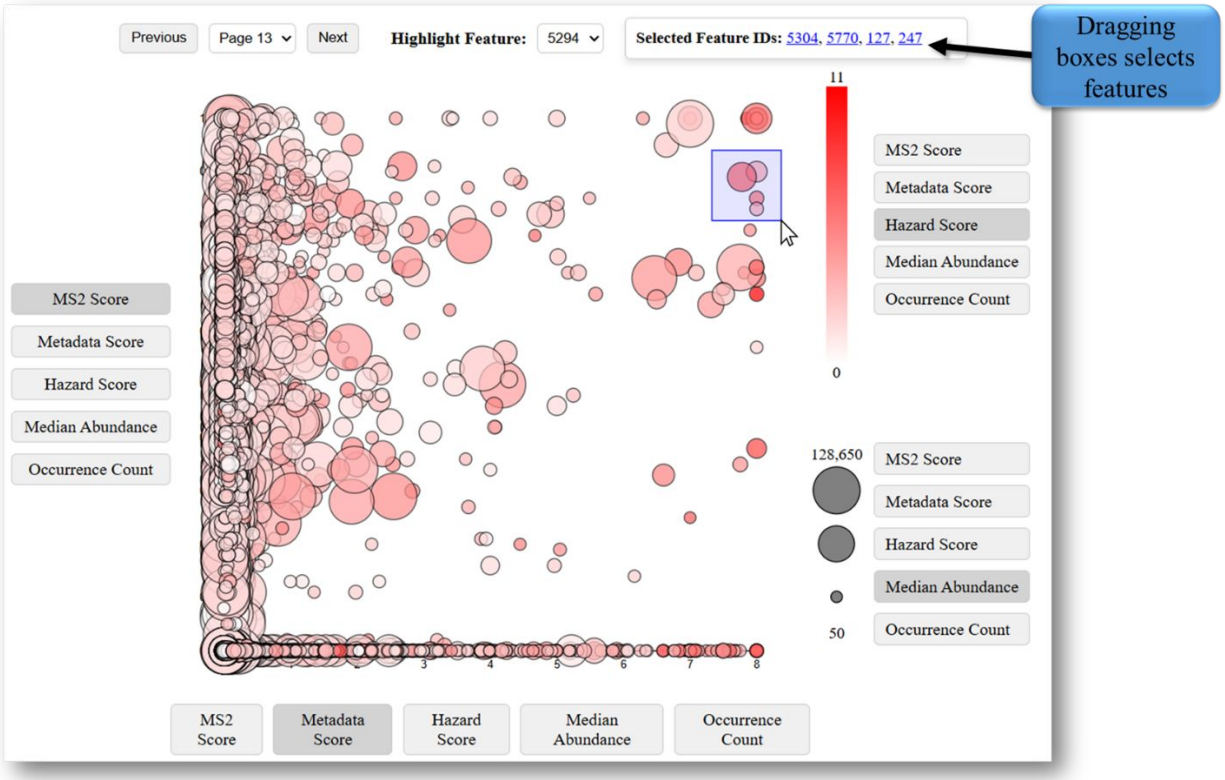
The following sections describe the usage and functionality of interactive visualizations developed within I-NTA to display, control, and interrogate chemical results generated by I-NTA. These visualizations are generated using data from the chemical results sheet output of the I-NTA's MS[1] workflow. Visualizations were developed in JavaScript and utilized the D3.js and AG Grid JavaScript libraries.

Section 5.1 – Interactive Scatterplot of Chemical Results

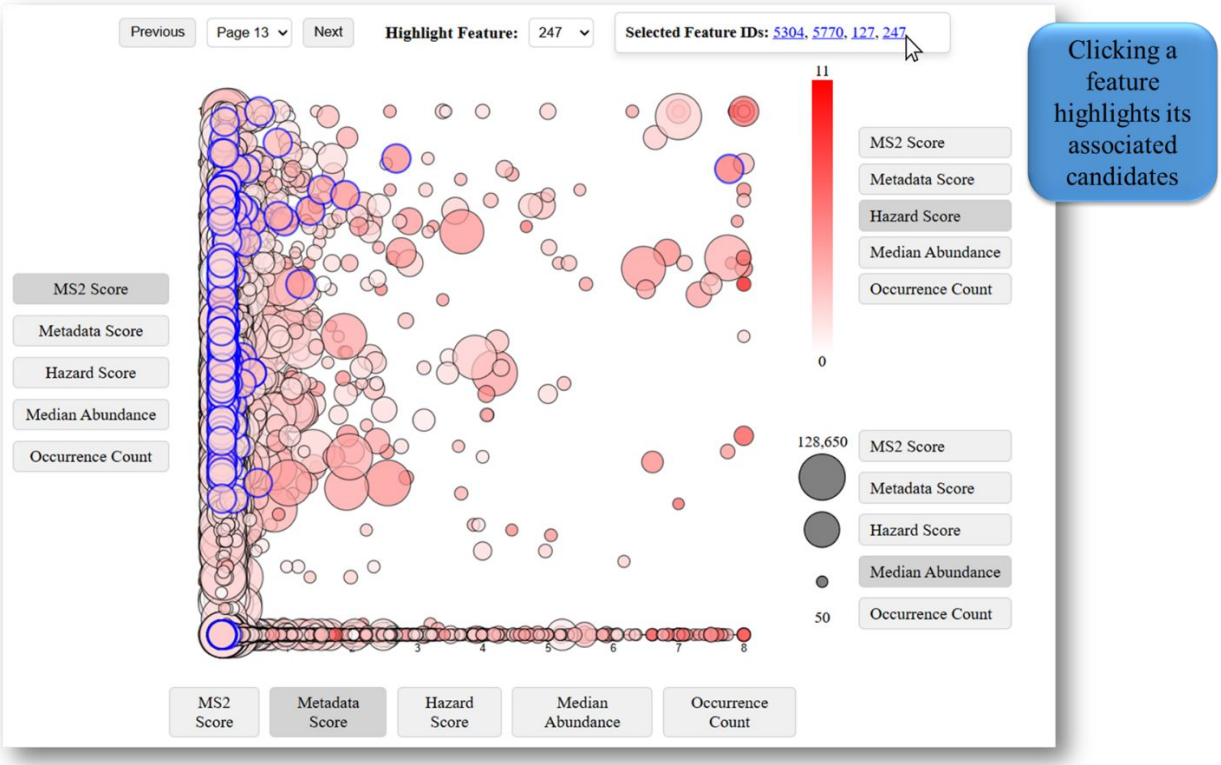


An interactive scatterplot displays the chemical results for chemical structure candidates retrieved for NTA study features, where each circle marker on the plot represents a single chemical structure candidate. A subset of all study features is displayed on a given page for purposes of clarity and visualization speed. Users may toggle through pages of features and their respective candidates using either "Previous" and "Next" buttons, or with a dropdown menu.

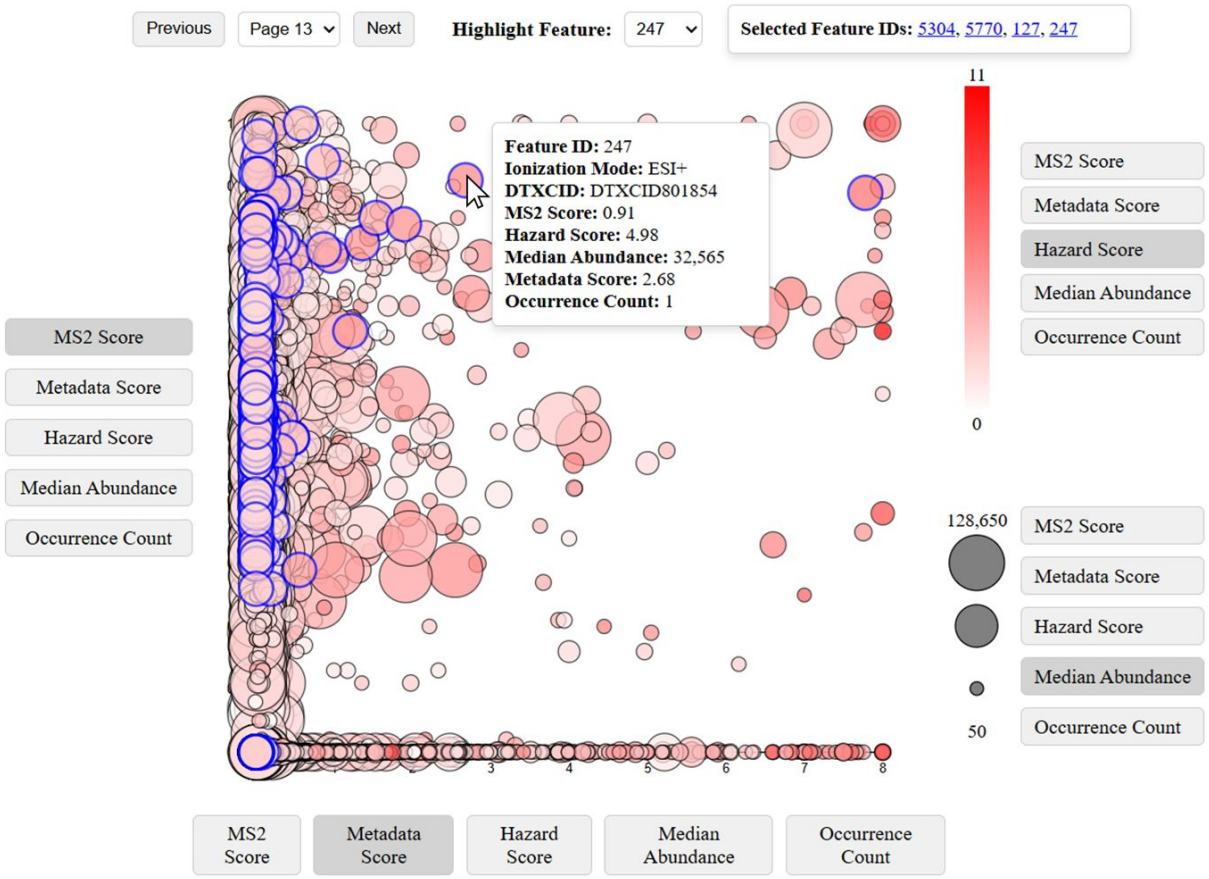
Value selectors for the x-axis and y-axis allow users to select what chemical or feature data to plot on either plot axis, ultimately controlling how candidates are sorted and grouped in space. Value selectors for the marker color and marker size allow users to select what chemical or feature data is represented by the color and size of the marker, respectively, ultimately controlling what markers stand out visually wherever they are located.



After configuring parameters of the scatterplot, users may select candidates/markers and their respective features by dragging a box around data markers. This reveals features associated with the selected markers in the “Selected Feature IDs” box.

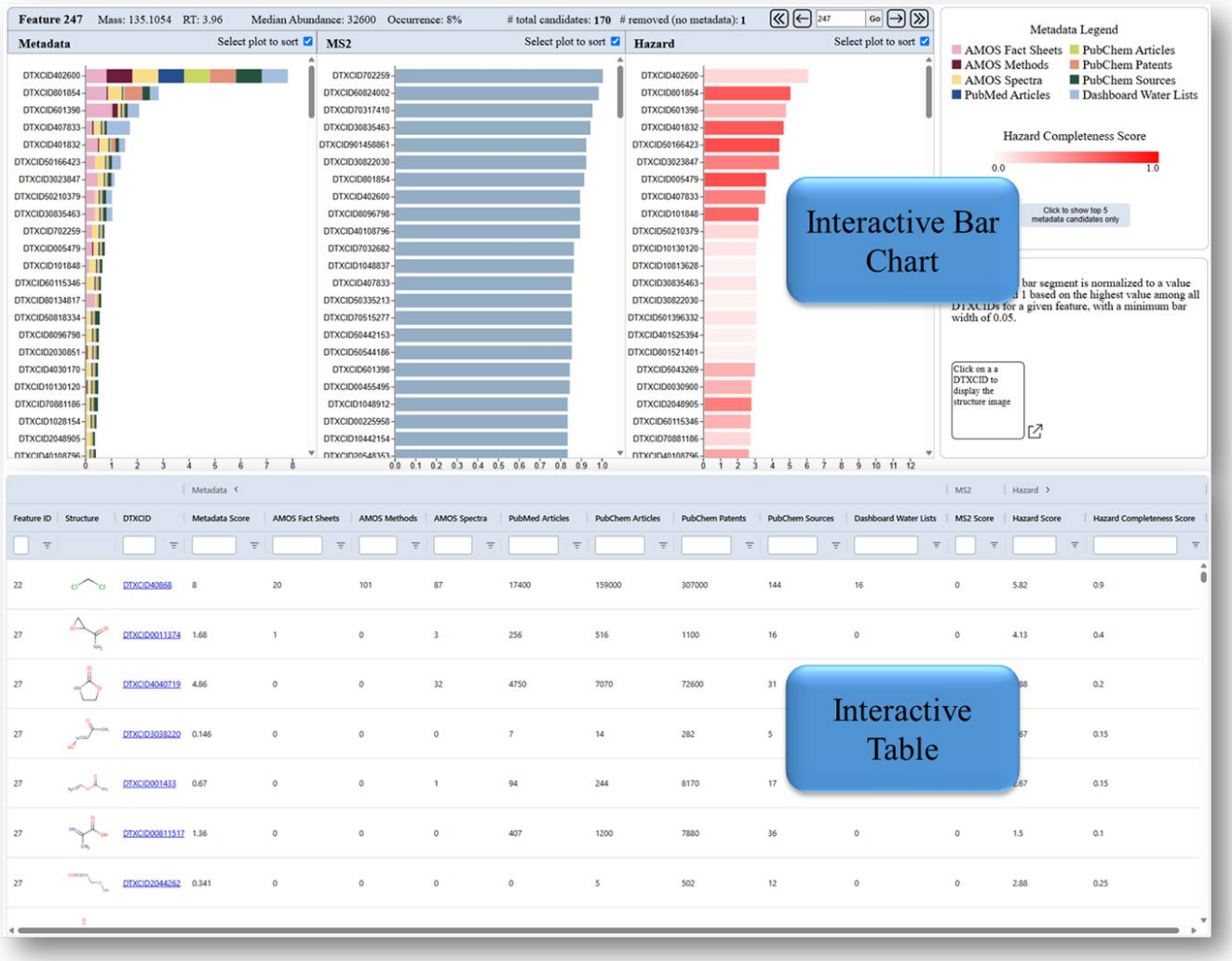


A feature may be selected by either clicking on its ID in the "Selected Feature IDs" box or by selecting it within the "Highlight Features" dropdown menu. Selecting a feature highlights all the candidates associated with the feature.

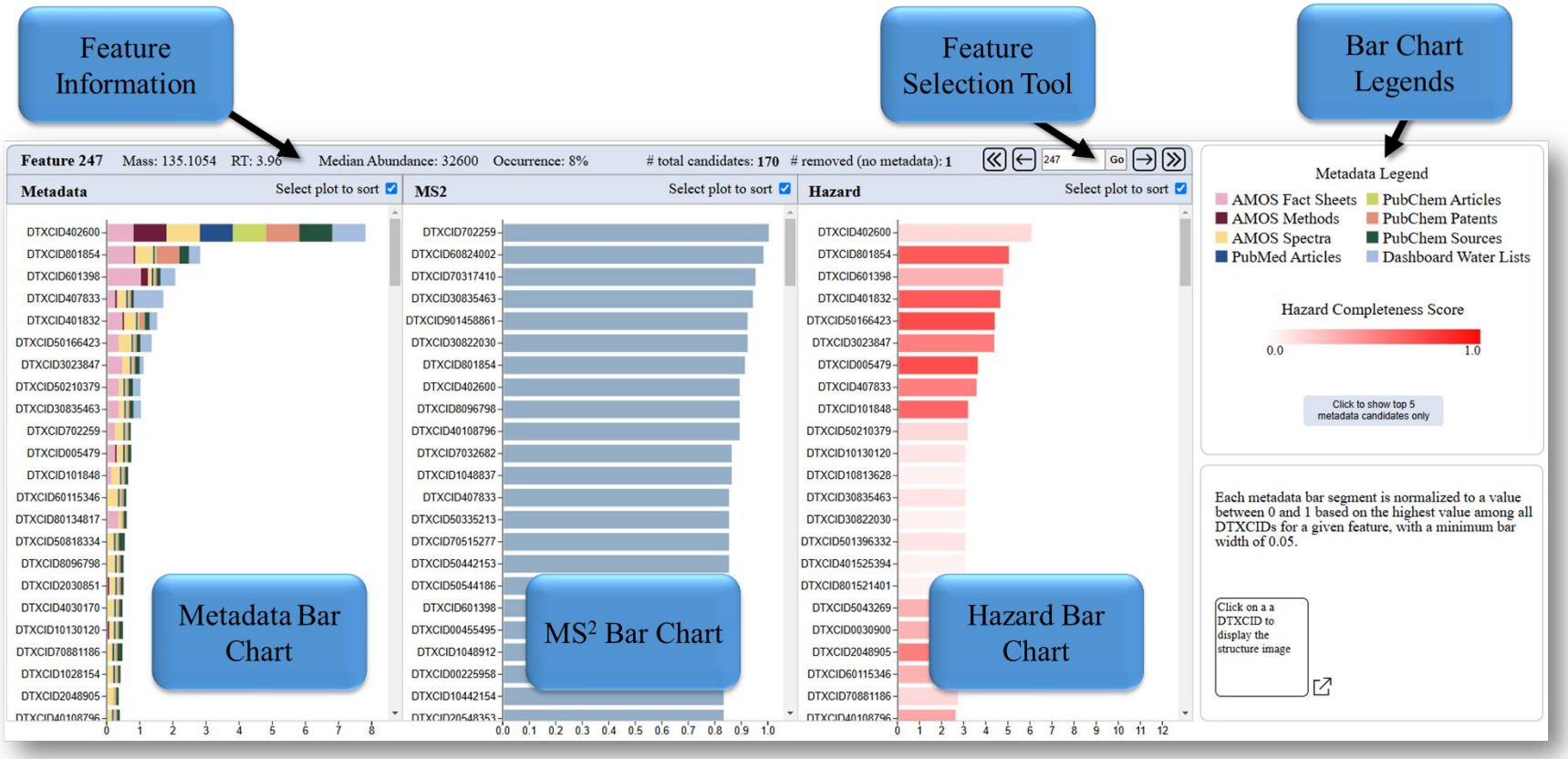


Finally, mousing over any marker in the plot brings up a textbox containing the relevant feature and candidate information for that specific candidate. Altogether, the user can drive how the data are sorted spatially and highlighted via color and size to visually determine the specific data points that are of interest. To explore further, the user can mouseover markers to see underlying feature and candidate information.

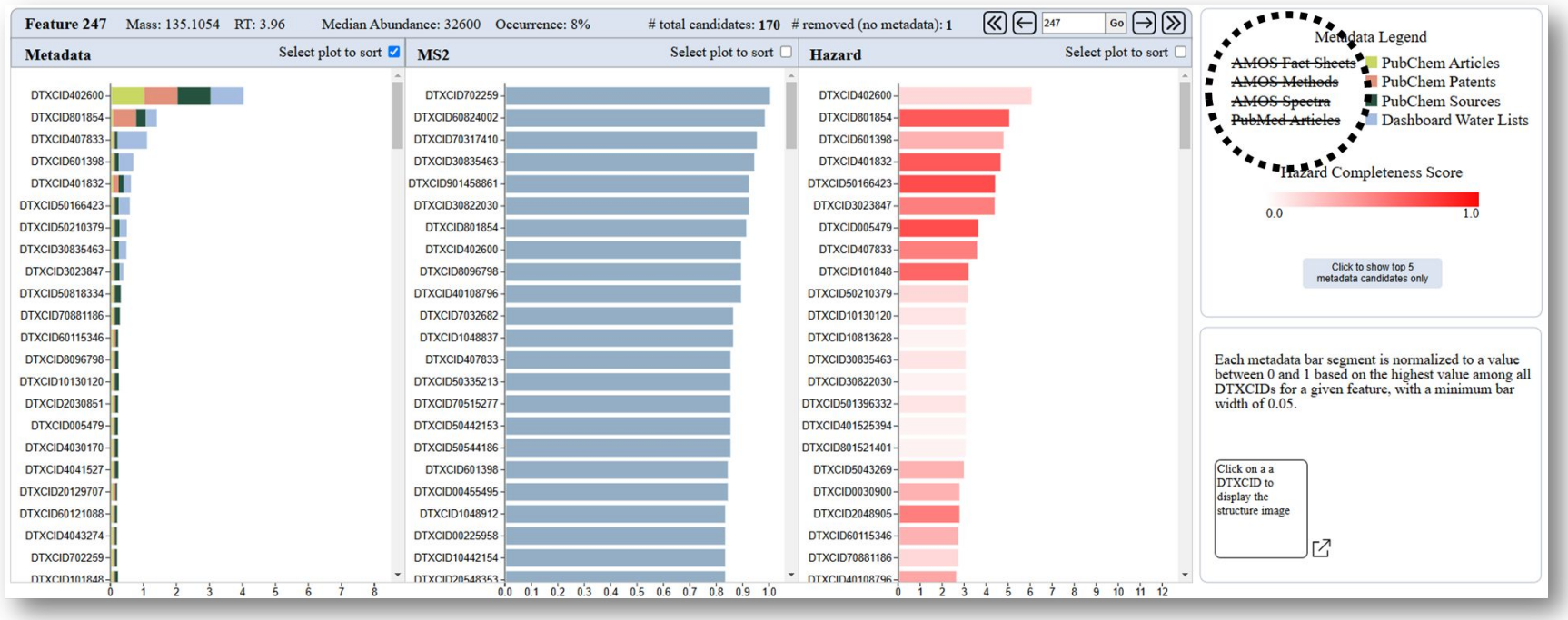
Section 5.2 – Interactive Bar Chart and Grid Table of Chemical Results



A combined interactive bar chart and grid table visualization allows for further investigation of the underlying data once features and candidates of interest have been identified from the previous interactive scatterplot. The bar chart section displays chemical information graphically, whereas the grid table contains the data underlying the visuals.



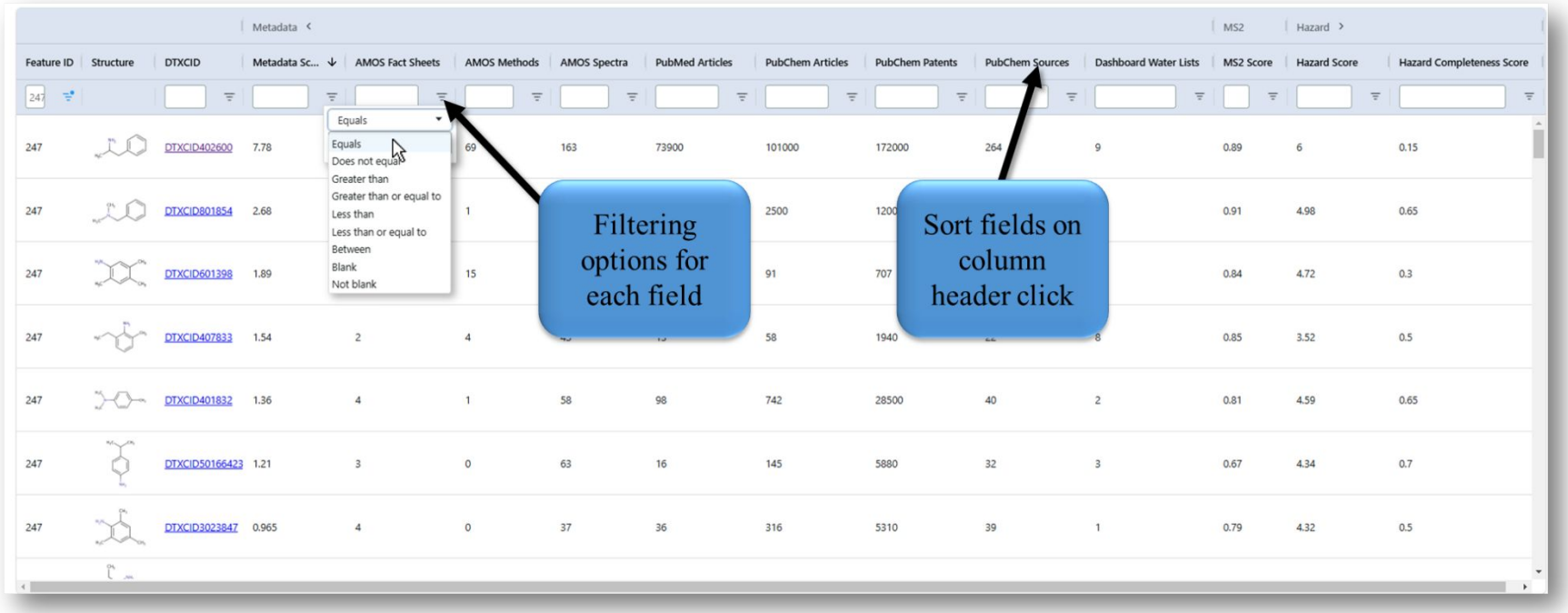
In the title block of the bar chart plot, following the feature ID, mass, retention time, median abundance and occurrence percentage, the total number of candidates for that feature and the number of those candidates with no associated metadata are displayed (NOTE: candidates with no metadata do not appear in the bar plot). There is also a feature selection tool to scroll through the features, as well as to select a specific feature ID. When a feature is selected, three bar charts are populated for the associated candidate structures of that feature; a metadata, MS^2, and hazard bar chart. Each bar chart's bar represents that respective category's score value. The metadata bar chart displays the normalized contributions from each metadata field broken out by color, and the hazard bar chart displays the completeness score of the hazard score value based on coloration as well. To ensure that metadata bar segments are wide enough to mouse-over, segments with a normalized value of less than 0.05 are adjusted to have a width of 0.05. A bar chart legend on the right explains the colors for these two bar charts.



The user can select the metadata fields to display in the metadata bar chart by interacting with the metadata legend. As a metadata field is selected or de-selected, the bars will re-sort in descending order of the sum of the selected metadata scores only.

Using the “Click to show top 5 metadata candidates” button beneath the legends, the user can choose to toggle between displaying all candidates for a feature or displaying only the five candidates with the highest total metadata scores. Clicking the button to show only the top 5 candidates will cause the grid to populate with only the top 5 candidates for each feature.

A mouseover of any candidate DTXCID on the y-axis will display a tooltip containing the candidate structure, and also highlight the candidate in all three bar charts. Clicking on a candidate DTXCID will bring up the structure of the candidate in a hyperlinked diagram on the lower right that will load the CompTox Chemicals Dashboard page for the specified candidate. Additionally, a mouseover of any bar segment in the plot displays the actual value for that metadata/MS^2/hazard field.



The grid table portion of this visualization contains all the underlying data used to generate the bar charts. This includes all the counts for each metadata field, the MS^2 quotient scores, and each individual hazard endpoint's hazard and quality score. Numerical filters can be applied to any metadata column. The grid can be sorted in ascending or descending order of any of the data field values by clicking on the column headers. Sort and filter changes applied to the grid table will be reflected in the bar plot.

REFERENCES

(1) *Dask: Library for dynamic task scheduling*; Dask Development Team, 2016. http://dask.pydata.org.

(2) Grulke, C. M.; Williams, A. J.; Thillanadarajah, I.; Richard, A. M. EPA's DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research. *Computational Toxicology* **2019**, *12*, 100096. DOI: https://doi.org/10.1016/j.comtox.2019.100096.

(3) 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.; et al. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *Journal of Cheminformatics* **2017**, *9* (1), 61. DOI: 10.1186/s13321-017-0247-6.

(4) McEachran, A. D.; Sobus, J. R.; Williams, A. J. Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. *Analytical and Bioanalytical Chemistry* **2017**, *409* (7), 1729-1735. DOI: 10.1007/s00216-016-0139-z.

(5) McEachran, A. D.; Chao, A.; Al-Ghoul, H.; Lowe, C.; Grulke, C.; Sobus, J. R.; Williams, A. J. Revisiting Five Years of CASMI Contests with EPA Identification Tools. In *Metabolites*, 2020; Vol. 10.

(6) Lowe, C. N.; Williams, A. J. Enabling High-Throughput Searches for Multiple Chemical Data Using the U.S.-EPA CompTox Chemicals Dashboard. *Journal of Chemical Information and Modeling* **2021**, *61* (2), 565-570. DOI: 10.1021/acs.jcim.0c01273.

(7) McEachran, A. D.; Mansouri, K.; Grulke, C.; Schymanski, E. L.; Ruttkies, C.; Williams, A. J. "MS-Ready" structures for non-targeted high-resolution mass spectrometry screening studies. *Journal of Cheminformatics* **2018**, *10* (1), 45. DOI: 10.1186/s13321-018-0299-2.

(8) McEachran, A. D.; Balabin, I.; Cathey, T.; Transue, T. R.; Al-Ghoul, H.; Grulke, C.; Sobus, J. R.; Williams, A. J. Linking *in silico* MS/MS spectra with chemistry data to improve identification of unknowns. *Scientific Data* **2019**, *6* (1), 141. DOI: 10.1038/s41597-019-0145-z.

(9) Chao, A.; Al-Ghoul, H.; McEachran, A. D.; Balabin, I.; Transue, T.; Cathey, T.; Grossman, J. N.; Singh, R. R.; Ulrich, E. M.; Williams, A. J.; et al. *In silico* MS/MS spectra for identifying unknowns: a critical examination using CFM-ID algorithms and ENTACT mixture samples. *Analytical and Bioanalytical Chemistry* **2020**, *412* (6), 1303-1315. DOI: 10.1007/s00216-019-02351-7.

(10) Little, J. L.; Williams, A. J.; Pshenichnov, A.; Tkachenko, V. Identification of "Known Unknowns" Utilizing Accurate Mass Data and ChemSpider. *Journal of the American Society for Mass Spectrometry* **2012**, *23* (1), 179-185. DOI: 10.1007/s13361-011-0265-y.

(11) Richman, T.; Arnold, E.; Williams, A. J. Curation of a list of chemicals in biosolids from EPA National Sewage Sludge Surveys & Biennial Review Reports. *Scientific Data* **2022**, *9* (1), 180. DOI: 10.1038/s41597-022-01267-9.

(12) Gaines, L. G. T.; Sinclair, G.; Williams, A. J. A proposed approach to defining per- and polyfluoroalkyl substances (PFAS) based on molecular structure and formula. *Integrated Environmental Assessment and Management* **2023**, *19* (5), 1333-1347. DOI: 10.1002/ieam.4735.

(13) Phillips, K. A.; Yau, A.; Favela, K. A.; Isaacs, K. K.; McEachran, A.; Grulke, C.; Richard, A. M.; Williams, A. J.; Sobus, J. R.; Thomas, R. S.; et al. Suspect Screening Analysis of Chemicals in Consumer Products. *Environmental Science & Technology* **2018**, *52* (5), 3125-3135. DOI: 10.1021/acs.est.7b04781.

(14) *Programmatic Access*. PubChem, 2025. https://pubchem.ncbi.nlm.nih.gov/docs/programmatic-access (accessed 2025 April 10th).

(15) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environmental Science & Technology* **2014**, *48* (4), 2097-2098. DOI: 10.1021/es5002105.
(16) Allen, F.; Greiner, R.; Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **2015**, *11* (1), 98-110. DOI: 10.1007/s11306-014-0676-4.
(17) Wang, F.; Liigand, J.; Tian, S.; Arndt, D.; Greiner, R.; Wishart, D. S. CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. *Analytical Chemistry* **2021**, *93* (34), 11692-11700. DOI: 10.1021/acs.analchem.1c01465.
(18) Rager, J. E.; Strynar, M. J.; Liang, S.; McMahen, R. L.; Richard, A. M.; Grulke, C. M.; Wambaugh, J. F.; Isaacs, K. K.; Judson, R.; Williams, A. J.; et al. Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring. *Environment International* **2016**, *88*, 269-280. DOI: 10.1016/j.envint.2015.12.008.
(19) Newton, S. R.; McMahen, R. L.; Sobus, J. R.; Mansouri, K.; Williams, A. J.; McEachran, A. D.; Strynar, M. J. Suspect screening and non-targeted analysis of drinking water using point-of-use filters. *Environmental Pollution* **2018**, *234*, 297-306. DOI: 10.1016/j.envpol.2017.11.033.
(20) Dionisio, K. L.; Phillips, K.; Price, P. S.; Grulke, C. M.; Williams, A.J.; Biryol, D.; Hong, T.; Isaacs, K. K. The Chemical and Products Database, a resource for exposure-relevant data on chemicals in consumer products. *Scientific Data* **2018**, *5* (1), 180125. DOI: 10.1038/sdata.2018.125.
(21) Isaacs, K. K.; Wall, J. T.; Williams, A. R.; Hobbie, K. A.; Sobus, J. R.; Ulrich, E.; Lyons, D.; Dionisio, K. L.; Williams, A. J.; Grulke, C.; et al. A harmonized chemical monitoring database for support of exposure assessments. *Scientific Data* **2022**, *9* (1), 314. DOI: 10.1038/s41597-022-01365-8.
(22) Richard, A. M.; Judson, R. S.; Houck, K. A.; Grulke, C. M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.; Martin, M. T.; Wambaugh, J. F.; et al. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chemical Research in Toxicology* **2016**, *29* (8), 1225-1251. DOI: 10.1021/acs.chemrestox.6b00135.
(23) Feshuk, M.; Kolaczkowski, L.; Watford, S.; Paul Friedman, K. ToxRefDB v2.1: update to curated in vivo study data in the Toxicity Reference Database. *Frontiers in Toxicology* **2023**, *5*, Mini Review.
(24) Wall, T.; Sayre, R.; Smith, D.; Winter, S.; Groover, M.; Hope, J.; Webb, A.; Friedman, K.; Feshuk, M.; Williams, A.; et al. Development of the Toxicity Values Database, ToxValDB: A Curated Resource for Experimental and Derived Human Health-Relevant Toxicity Data. *Submitted* **2025**.
(25) Vegosen, L.; Martin, T. M. An automated framework for compiling and integrating chemical hazard data. *Clean Technologies and Environmental Policy* **2020**, *22* (2), 441-458. DOI: 10.1007/s10098-019-01795-w.
(26) Brunelle, L. D.; Batt, A. L.; Chao, A.; Glassmeyer, S. T.; Quinete, N.; Alvarez, D. A.; Kolpin, D. W.; Furlong, E. T.; Mills, M. A.; Aga, D. S. De facto Water Reuse: Investigating the Fate and Transport of Chemicals of Emerging Concern from Wastewater Discharge through Drinking Water Treatment Using Non-targeted Analysis and Suspect Screening. *Environmental Science & Technology* **2024**, *58* (5), 2468-2478. DOI: 10.1021/acs.est.3c07514.
(27) Sobus, J. R.; Sayre-Smith, N. A.; Chao, A.; Ferland, T. M.; Minucci, J. M.; Carr, E. T.; Brunelle, L. D.; Batt, A. L.; Whitehead, H. D.; Cathey, T.; et al. Automated QA/QC reporting for

non-targeted analysis: a demonstration of “INTERPRET NTA” with de facto water reuse data. *Analytical and Bioanalytical Chemistry* **2025**. DOI: 10.1007/s00216-025-05771-w.
(28) *Data Sources*. PubChem, https://pubchem.ncbi.nlm.nih.gov/source (accessed 2025 April 10th).
(29) Stein, S. E.; Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry* **1994**, *5* (9), 859-866. DOI: 10.1016/1044-0305(94)87009-8.
(30) Newmeyer, M. N.; Lyu, Q.; Sobus, J. R.; Williams, A. J.; Nachman, K. E.; Prasse, C. Combining Nontargeted Analysis with Computer-Based Hazard Comparison Approaches to Support Prioritization of Unregulated Organic Contaminants in Biosolids. *Environmental Science & Technology* **2024**, *58* (27), 12135-12146. DOI: 10.1021/acs.est.4c02934.
(31) Janesch, G.; Carr, E. T.; Sivasupramaniam, S.; Charest, N.; Williams, A. J. Applying Cheminformatics to Develop a Structure Searchable Database of Analytical Methods. *Submitted.* **2025**.
(32) Sloop, J. T.; Chao, A.; Gundersen, J.; Phillips, A. L.; Sobus, J. R.; Ulrich, E. M.; Williams, A. J.; Newton, S. R. Demonstrating the Use of Non-targeted Analysis for Identification of Unknown Chemicals in Rapid Response Scenarios. *Environ Sci Technol* **2023**, *57* (8), 3075-3084. DOI: 10.1021/acs.est.2c06804 From NLM.
(33) Brueck, C. L.; Xin, X.; Lupolt, S. N.; Kim, B. F.; Santo, R. E.; Lyu, Q.; Williams, A. J.; Nachman, K. E.; Prasse, C. (Non)targeted Chemical Analysis and Risk Assessment of Organic Contaminants in Darkibor Kale Grown at Rural and Urban Farms. *Environmental Science & Technology* **2024**, *58* (8), 3690-3701. DOI: 10.1021/acs.est.3c09106.
(34) Ring, C. L.; Arnot, J. A.; Bennett, D. H.; Egeghy, P. P.; Fantke, P.; Huang, L.; Isaacs, K. K.; Jolliet, O.; Phillips, K. A.; Price, P. S.; et al. Consensus Modeling of Median Chemical Intake for the U.S. Population Based on Predictions of Exposure Pathways. *Environmental Science & Technology* **2019**, *53* (2), 719-732. DOI: 10.1021/acs.est.8b04056.
(35) Wambaugh, J. F.; Bare, J. C.; Carignan, C. C.; Dionisio, K. L.; Dodson, R. E.; Jolliet, O.; Liu, X.; Meyer, D. E.; Newton, S. R.; Phillips, K. A.; et al. New approach methodologies for exposure science. *Current Opinion in Toxicology* **2019**, *15*, 76-92. DOI: https://doi.org/10.1016/j.cotox.2019.07.001.