Non-Targeted Identification of Reactive Metabolite Protein Adducts

[Supplementary Information]

Michael G. Leeming, [(1)*] William A. Donald, [(2)*] Richard A.J. O'Hair [(1)*]

(1) School of Chemistry and Bio21 Institute of Molecular Science and Biotechnology, The University of Melbourne, Victoria 3010, Australia.

(2) School of Chemistry, University of New South Wales, Sydney, NSW 2052, Australia.

* Corresponding Authors

Mr. Michael G. Leeming
School of Chemistry
Bio21 Research Institute
University of Melbourne
Melbourne, Victoria, 3010
T: +61 3 83442451
E: m.leeming@student.unimelb.edu.au

Dr. William A. Donald
School of Chemistry
University of New South Wales
Dalton Building
Sydney, NSW, 2052
T: +61 2 9385 8827
E: w.donald@unsw.edu.au
F: +61 2 9385-6141

Prof. Richard A. J. O'Hair
School of Chemistry
Bio21 Research Institute
University of Melbourne
Melbourne, Victoria, 3010
T: +61 3 83442452
E: rohair@unimelb.edu.au
F: +61 3 9347-5180

Table of Contents

Additional experimental details

Liquid chromatography-mass spectrometry

Analyte peptides were introduced into the gas phase via nano-ESI using a spray voltage of 1.9 kV, capillary temperature of 275 °C and all MS^1 mass spectra were acquired in the positive ion mode over a range of 400-1500 *m/z* in the Orbitrap analyser operating at a resolution of 120,000. For MS/MS runs, data-dependent MS^2 spectra were collected by subjecting 2+ to 5+ precursor peptide ions with a measured intensity greater than 10^4 to higher-energy collision induced dissociation (HCD) with a normalised collision energy of 35%. Dynamic exclusion was used to prevent the selection of the same precursor ion more than once for 45 s. For each run, 2 μL of sample was injected and samples were held in the autosampler at 5 °C prior to analysis.

Software development

The Xenophile software can be utilised through a graphical user interface developed using PyQt4/Qt4 and contains methods that allow for non-targeted identification of reactive drug metabolites and targeted determination of CRM-modified peptides and proteins. Utilities are also included for interactive review of non-targeted and targeted search results and summary reports of these finding can be generated that contain pertinent data about hits that are accepted by the user. An mzML data file browser is also included that can carry out numerous tasks that are common in this research such as inspecting raw LC-MS/MS data and plotting EICs. These various functions accept LC-MS and LC-MS/MS data in the open-source 'mzML' data file format[1] and the latter datasets are currently designed to be interrogated by the Mascot peptide database search software. Support for additional database searching packages such as Sequest[2, 3] is planned but is not available at this time. The Xenophile source code can be obtained from https://github.com/mgleeming/Xenophile and various screen shots of the user interface are provided in the supplementary information. Descriptions of the algorithms comprising this software can be found in this manuscript, in previous publications[4] or in the documentation accompanying the source code. The software utilises numerous open source libraries.[5-8]

Post-processing of HiTIME data

Raw HiTIME data was subjected to various post-processing steps using the in order to derive a list of possible twin-ion locations. This was achieved using the "Post-processing" tab of the "HiTIME search" tools of the Xenophile software. Specifically, a baseline noise level was estimated by inspecting data heat maps and score distribution histograms then points below this level were removed. Local maxima detection of the remaining data provided sites of probable twin-ions.

It is known that peptide signals in LC-MS data typically display multiple ^{13}C isotope peaks in addition to the monoisotopic peak and, given this, unique HiTIME local maxima should be observed corresponding to each peak in a peptide isotope distribution. This feature can be used to elevate the scores of signals where multiple HiTIME maxima are observed separated by a distance corresponding to the charge state under analysis. For the current data, this feature was enabled by selecting the 'Peptide Isotope Scaling' checkbox in the 'Post-processing' setup tab of the Xenophile software.

The final set of HiTIME local maxima points, with scores that had been scaled according to the presence of ^{13}C isotope peaks, were then manually reviewed using the 'Results Viewer' tab of the Xenophile software. In general, hits that approximately satisfied the condition: $H_S - H_{max}^{raw} > 0$, where H_S is the score of a given hit following peptide isotope scaling and H_{max}^{raw} is the maximum score or the raw (unscaled) HiTIME data were manually reviewed. This ultimately selects data points that also have ^{13}C isotope peaks and are likely to be due to modified peptide signals.

Production of semi-synthetic data sets

Semi-synthetic data sets were produced by computationally creating synthetic data features that mimic the elution and fragmentation of APAP/$^{13}C_6$APAP modified peptides and superimposing these onto experimental data sets produced from analysis of vehicle control samples. There are numerous advantages to this approach that include: 1) the number and location of 'true' twin-ion signals is precisely known, 2) the complexity and confounding features of experimental data are maintained and

accounted for, and 3) numerous properties and parameters of the data analysis describe here can be rapidly assessed.

Semi-synthetic data sets were produced for each of the three vehicle control replicates. For each replicate, two data files were created: 1) an mzML (LCMS) data file containing the synthetic twin-ion features superimposed on experimental MS^1 data, and 2) an MGF file containing all entries from experimental data in addition to those for synthetic peptide signals.

Briefly, the procedure was as follows:

1) Parse the experimental mascot results file and obtain a list of peptides that were confidently assigned as having exactly: (i) no missed enzymatic cleavages, and (ii) exactly one carbamidomethyl-cysteine residue. These requirements were applied simply to constrain the number of peptides.
2) For each peptide, identify the experimentally observed ions in the MS^2 spectrum that were assigned by mascot as corresponding to fragments that contain the cysteine residue.
3) Produce a synthetic MS^2 peak list by incrementing the *m/z* value of these fragments by the desired mass of the reactive metabolite modification. The result of this is an MS^2 spectrum identical to the experimental entry except that mascot-assigned fragments containing cysteine have been offset by the modification mass.
4) MGF entries were created for these synthetic MS^2 spectra which were then combined into the MGF file derived from experimental LC-MS/MS data.
5) For each peptide identified in step 1, a synthetic LC-MS signal was produced using MSSimulator.[7] These signals take into account the predicted isotope distributions from different elemental compositions and intensity values were matched to that of the experimental peptide. Retention times were set at a randomly chosen point between 0.5 and 5 minutes after elution of the experimental peptide.
6) Each synthetic signal was then added into the experimental mzML file. The synthetic data were superimposed on existing experimental data in the target region to replicate aspects of LC-MS experiments such as overlapping peaks.

A new mzML file containing the final semi-synthetic data was written and used in subsequent processing steps.

A Python script utilising the OpenMS library[9, 10] was written to automate the data creation process which can be obtained from the Xenophile project GitHub repository at https://github.com/mgleeming/Xenophile/tree/master/xenophile/extras.

The MGF and mzML files were then processed with Mascot and HiTIME respectively using the parameters described in the main text and the results of these were then subjected to the same post processing routines described in-text.

Formulae Determination Metrics

Estimates of the number of candidate formulae assigned to a given CRM mass were produced for a wide variety of molecules. The ZINC15 database[11] of small molecule drugs approved by the United States Food and Drug Administration (containing 1385 chemical entities) was used. Molecules containing fewer than 5 carbon or hydrogen atoms were removed resulting in a list of 1337 compounds ranging in molecular weight from *ca.* 90–1000 Da. For each of these compounds, fragments were generated by disconnection of rotatable bonds as described in the main text. Allowed elemental compositions were computed by taking the lowest and highest number of a given element observed in any fragment generated. These upper limits of these ranges were then extended by addition of $C_8H_{10}O_8$ to allow for the possibility of metabolic reactions that substantially increase the mass of the CRM such as Glucuronidation. The lower limits were decreased by H_2 to allow for oxidations such as formation of NAPQI from APAP. The mass of all combination of stoichiometries within these ranges were then computed to give the set of allowed CRM compositions.

A theoretical CRM mass was then produced by incrementing the mass of the precursor molecule by a value randomly selected from 4 possibilites: 1) -2.01565 (-2H), 2) 13.97926 (+O, -2H), 3) 15.99491 Da (+O), and 4) 176.03209 (+6C 8H 6O). Formulae within 100 ppm of the CRM target mass were then retrieved from the target list and residual mass errors were calculated for each. In analyses of these data that employed RME restrictions, formulae were retained that had RME values of < 20 Da or $160 < \text{RME} < 180$ Da. These allow for common modifications that include various combinations of +O, -2H and glucuronic acid conjugation.

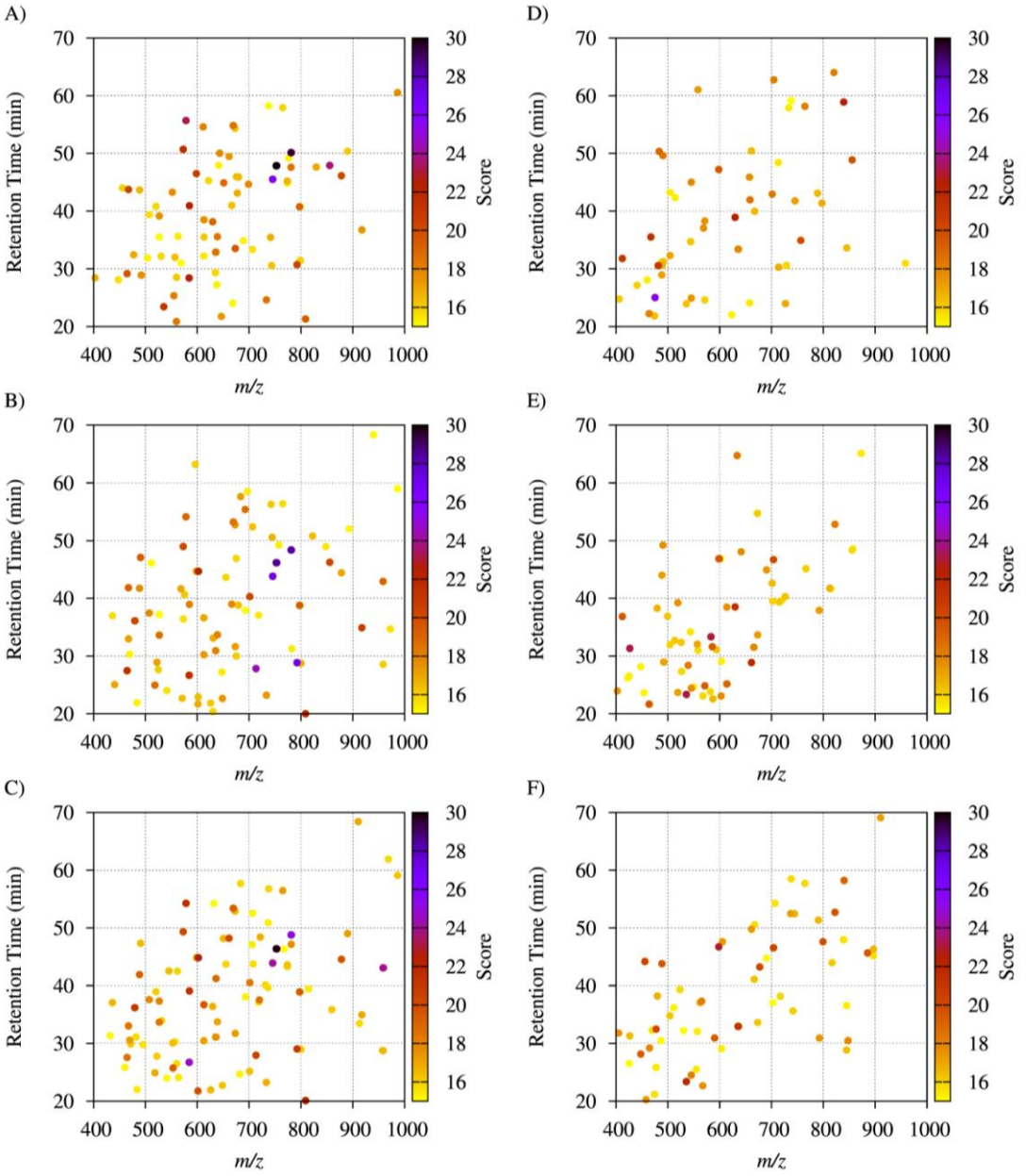

Figure S1. Heat maps produced by HiTIME scoring of microsomal protein digest LC-MS data sets with a doublet spacing of 3.01005 aimed at mining doubly charged peptides modified by an APAP metabolite. A-C) APAP treatment, D-F) control. Data points with a weighted score less than 15 have been omitted for clarity. For convenience, the data shown here in panels A and D are reproduced from Figure 2A and 2B respectively in the main text.

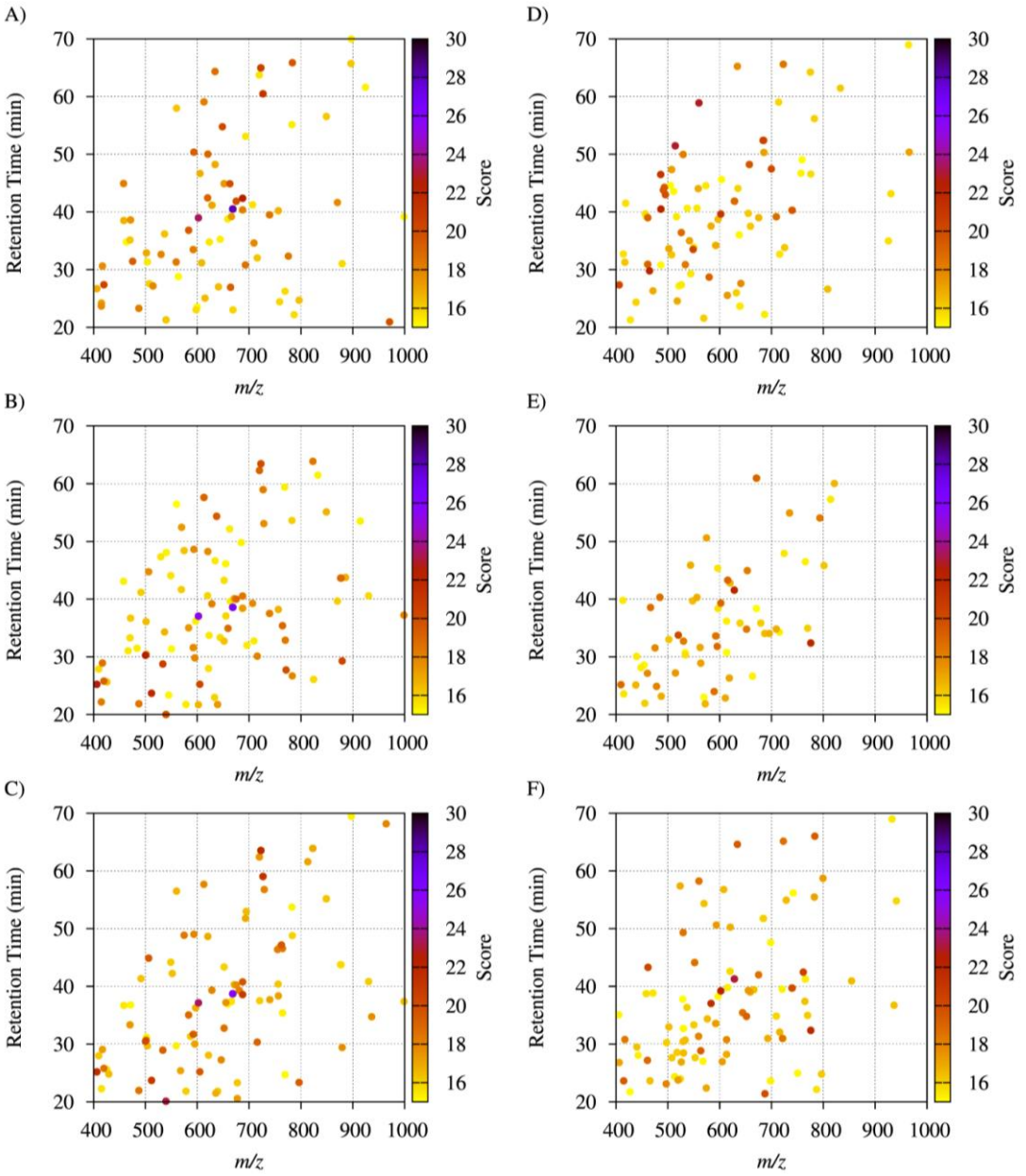

Figure S2. Heat maps produced by HiTIME scoring of microsomal protein digest LC-MS data sets with a doublet spacing of 2.0067 aimed at mining triply charged peptides modified by an APAP metabolite. A-C) APAP treatment, D-F) control. Data points with a weighted score less than 15 have been omitted for clarity.

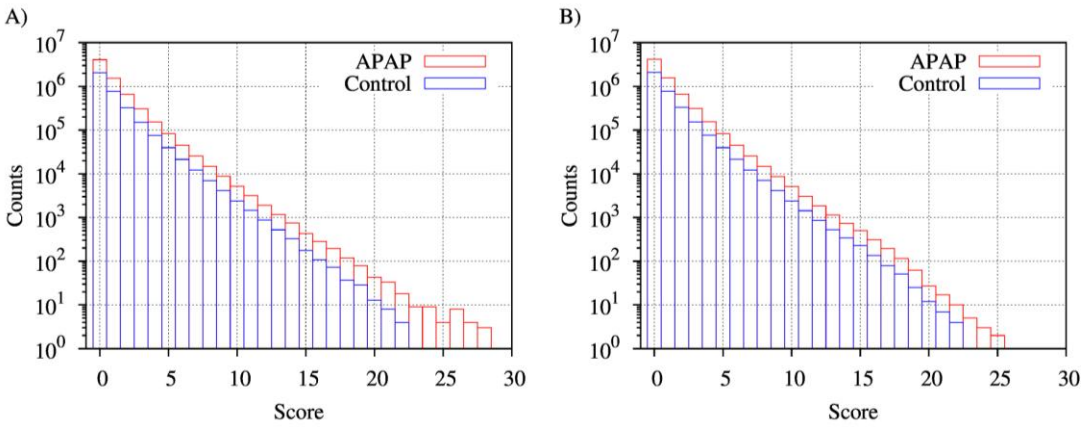

Figure S3. Histogram of the distribution of scores from HiTIME analysis of APAP treatment and control microsomal protein digest LC-MS data using twin-ion spacing settings of A) 3.01005, and B) 2.0067.

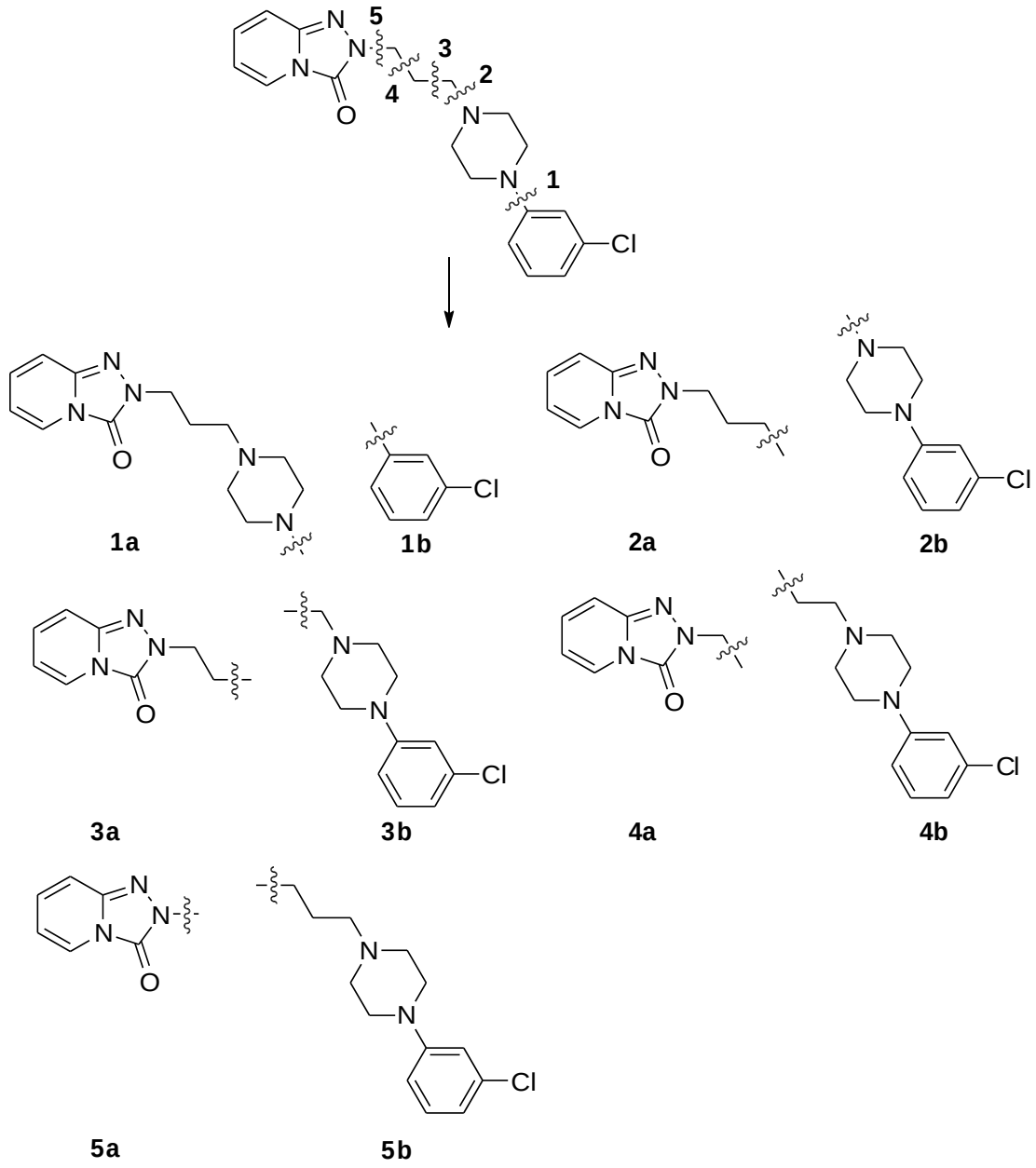

Figure S4. The fragments generated by disconnection of rotatable bonds in the antidepressant compound trazodone.

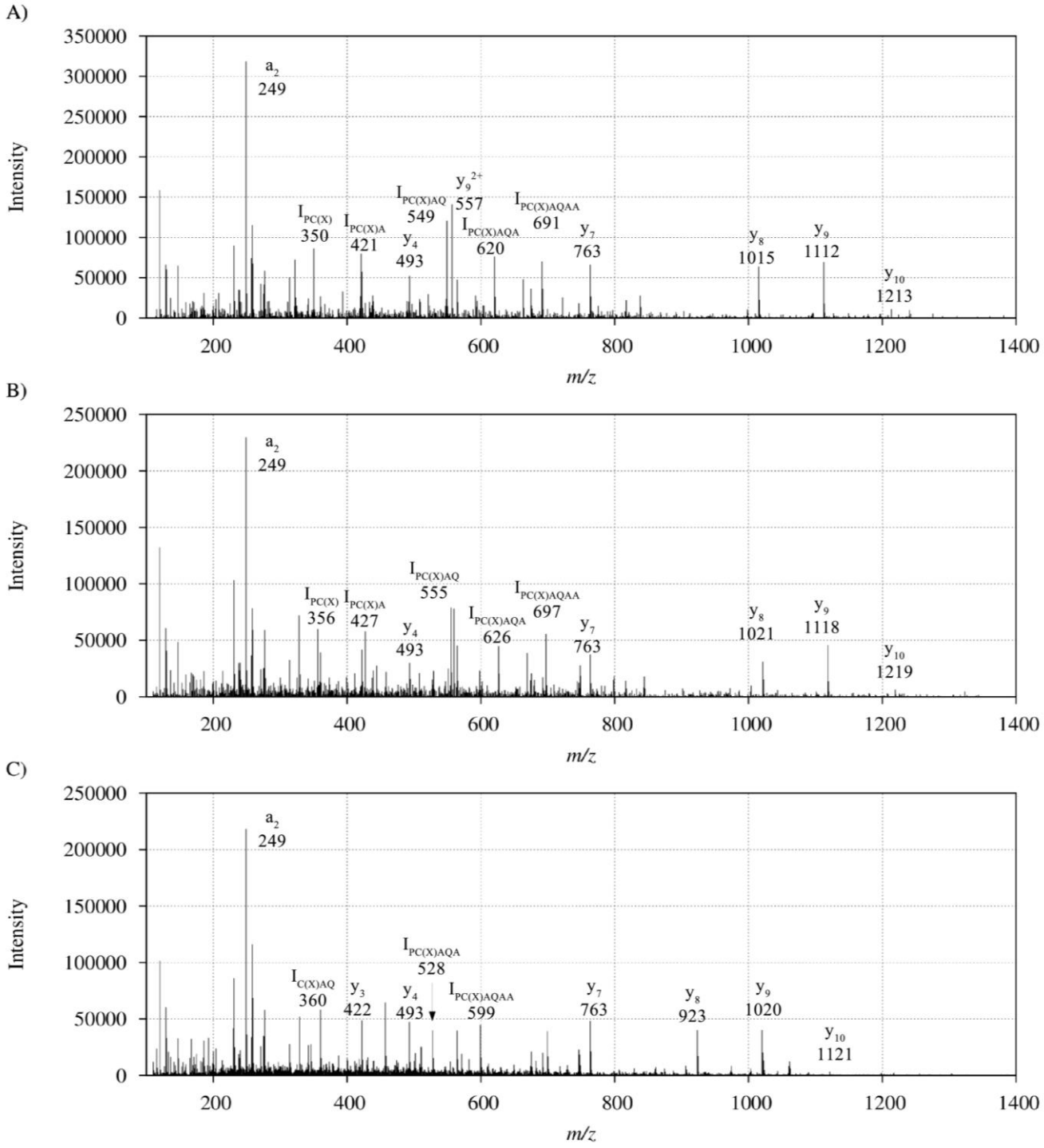

Figure S5. HCD MS^2 spectra of peptides assigned by targeted correlation of HiTIME and Mascot peptide assignments. EFTPC(**X**)AQAAFQK^{2+} where A) X = APAP, B) X = $^{13}C_6$ APAP and C) X = carbamidomethyl (C_2H_3NO, from iodoacetamide treatment of reduced proteins).

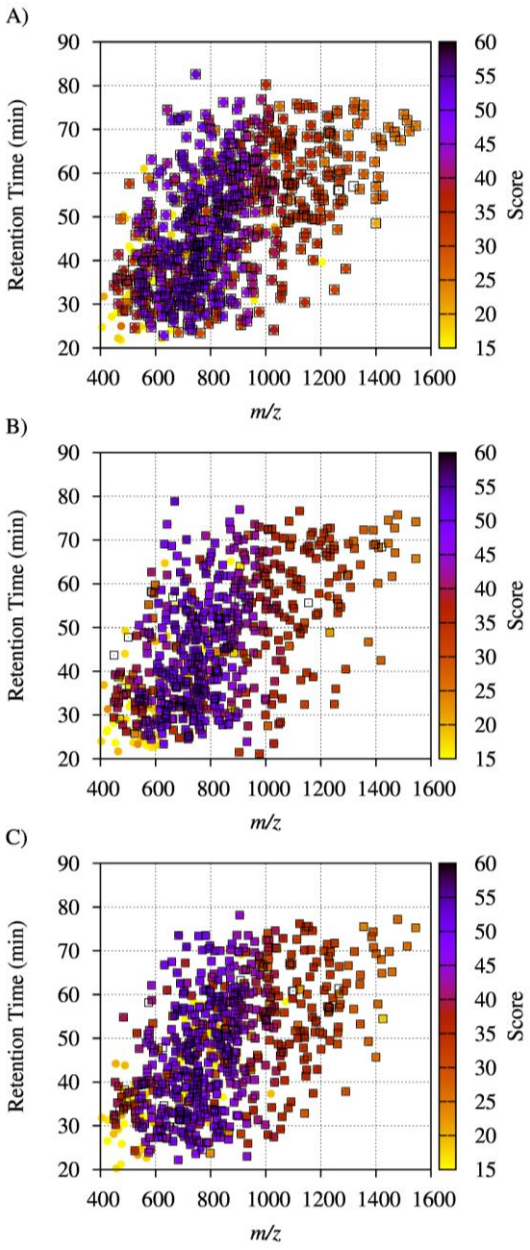

Figure S6. Heat maps produced by HiTIME scoring of the replicates of semi-synthetic LC-MS data sets (A-C) with a doublet spacing of 3.01005 aimed at mining doubly charged twin-ion peptides. Black boxes indicate locations of artificially introduced twin-ion signals.

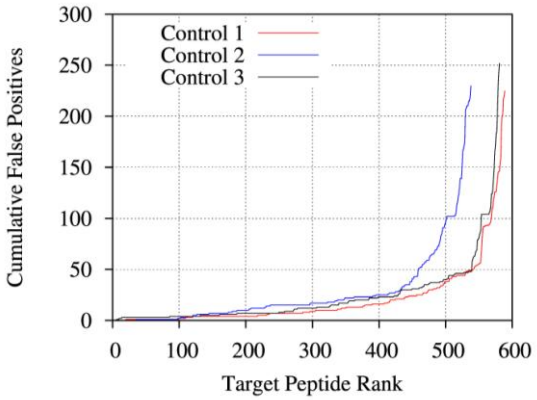

Figure S7. Cumulative number of false positive HiTIME hits as a function of target rank for synthetic twin-in peptides incorporated into experimental microsomal protein digests LC-MS data from control samples.

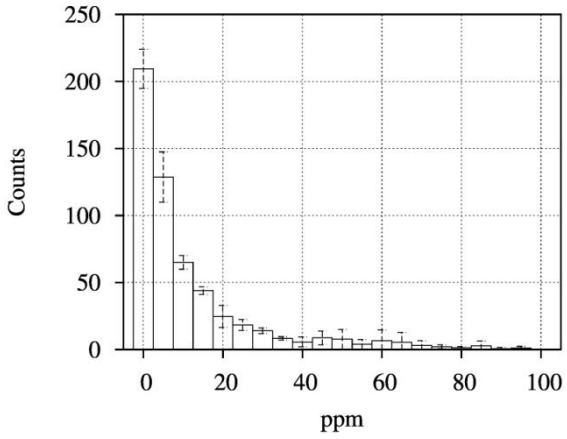

Figure S8. Histogram of CRMs mass ppm errors assigned by non-targeted reactive metabolite searching from the true value of 149.04713 Da. Counts show average values from across the three replicates of semi-synthetic twin-ion data and errors are quoted as +/- one standard deviation.

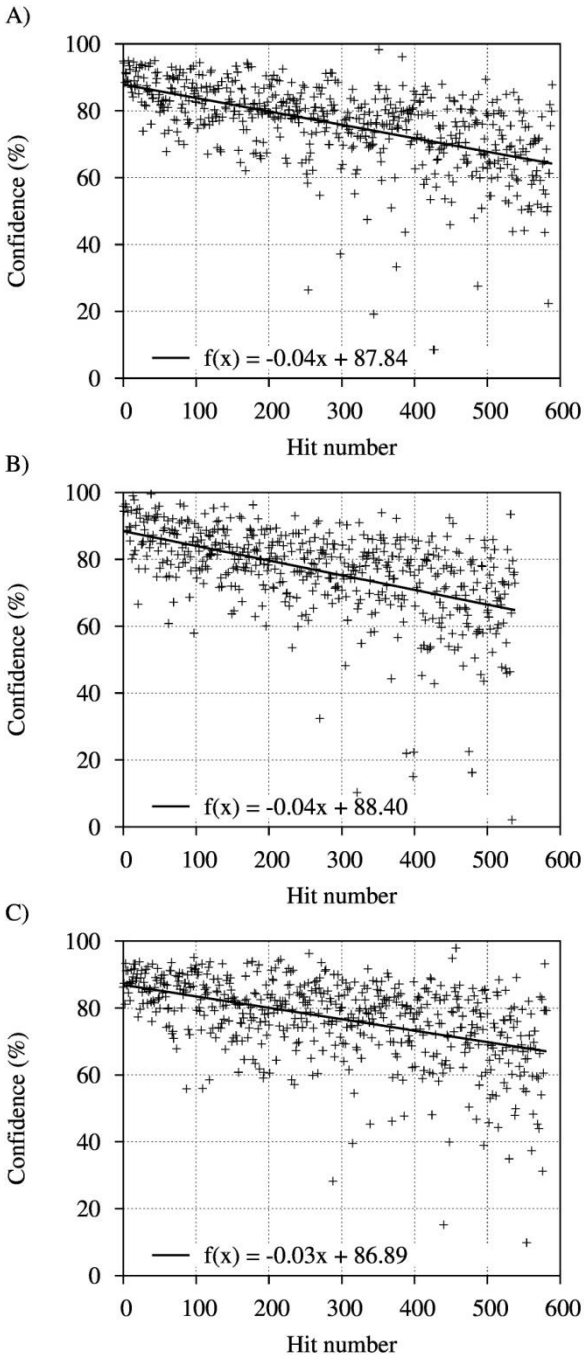

Figure S9. Confidence in the CRM masses assigned by the non-targeted reactive metabolite detection algorithm for three replicates (A-C) of semi-synthetic twin-ion data. Confidence values are calculated from the normalized peptide correlation scores according to $\frac{C1-C2}{C1} * 100$, where C1 and C2 are the correlation scores for the highest and second highest ranked hits respectively.

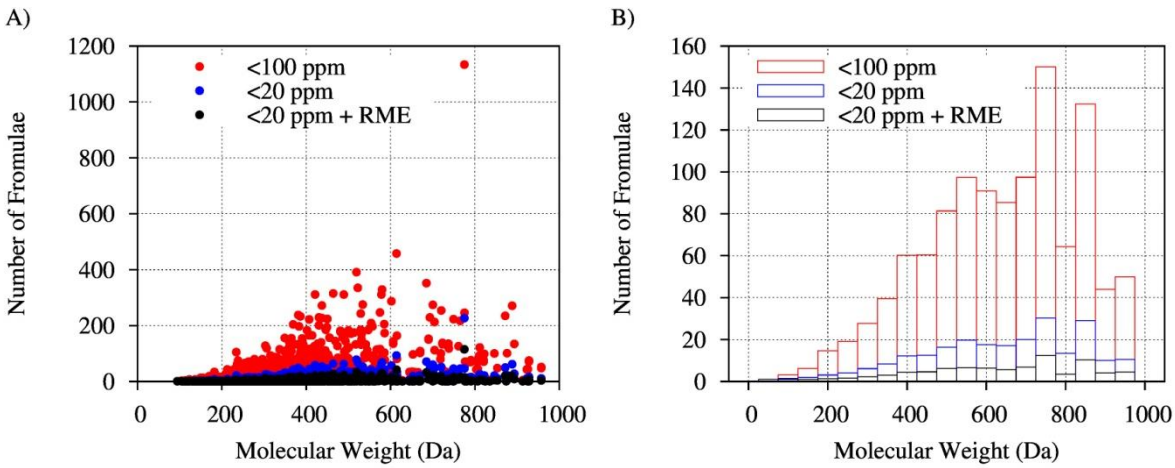

Figure S10. A) Number of formulae for theoretical reactive metabolites generated from 1337 molecules in the ZINC15 FDA approved drugs list. B) Histogram summarising the average number of possible formulae for molecules in in 50 Da bins.

Table S2. Parameters used for postprocessing of HiTIME scoring results

Parameter	Value	Unit
Minimum HiTIME Score	15	Dimensionless
m/z Width	0.15	*m/z*
RT Width	1	Minutes
RT Exclusion	0	Minutes
mzDelta	3.01005	Daltons
EIC Width	0.03	*m/z*
Peptide Isotope Scaling	True	N/A

Table S3. Theoretical '*b*' and '*y*' series ions generated via the rolling modification function for the exemplary peptide GLCGLR with a reactive metabolite of mass 100.0 Da. The notation '–X(M)-' is used to indicate the hypothetical modification of residue 'X' by reactive metabolite 'M'.

Residue	N-Terminal Fragment	N-Terminal Fragment Mass	C-terminal fragment	C-terminal Fragment Mass
Unmodified 'native' peptide GLCGLR				
1	G	58.03	LCGLR	561.32
2	GL	171.11	CGLR	448.23
3	GLC	274.12	GLR	345.22
4	GLCG	331.14	LR	288.2
5	GLCGL	444.23	R	175.12
Modification at site 1 G(M)-LCGLR				
1	G(M)	158.03	LCGLR	561.32
2	G(M)-L	271.11	CGLR	448.23
3	G(M)-LC	374.12	GLR	345.22
4	G(M)-LCG	431.14	LR	288.2
5	G(M)-LCGL	544.23	R	175.12
Modification at site 2 G-L(M)-CGLR				
1	G	58.03	L(M)-CGLR	661.32
2	G-L(M)	274.11	CGLR	448.23
3	G-L(M)-C	374.12	GLR	345.22
4	G-L(M)-CG	431.14	LR	288.2
5	G-L(M)-CGL	544.23	R	175.12
Modification at site 3 GL-C(M)-GLR				
1	G	58.03	L-C(M)-GLR	661.32
2	GL	171.11	C(M)-GLR	548.23
3	GL-C(M)	374.12	GLR	345.22
4	GL-C(M)-G	431.14	LR	288.2
5	GL-C(M)-GL	544.23	R	175.12
Modification at site 4 GLC-G(M)-LR				
1	G	58.03	LC-G(M)-LR	661.32
2	GL	171.11	C-G(M)-LR	548.23
3	GLC	274.12	G(M)-LR	445.22
4	GLC-G(M)	431.14	LR	288.2
5	GLC-G(M)-L	544.23	R	175.12

Table S4. Parameters used for non-targeted reactive metabolite detection

Parameter	Value	Units
Peptide Selection		
Threshold	Identity	
Charge state range	2-4	
HiTIME Selection		
MS^2 *m/z* offset	0.5	*m/z*
MS^2 RT offset	2	minutes
MS^2 correlation parameters		
Match *m/z* Tolerance	0.5	*m/z*
Match Ion Types	b,y	N/A
Reactive Residues	C, W, Y, M, K	N/A
CRM parameters		
m/z band	76-360	Da
Max. RME	100	Da
Ppm tolerance	20	Ppm

Table S5. Atom ranges used in non-targeted CRM detection

Element	Minimum	Maximum
C	6	15
N	0	4
O	0	4
H	3	30

Table S6. Three highest ranked reactive metabolite assignments produced for one of the APAP replicates using the non-targeted reactive metabolite identification algorithm. (APAP replicate 2 of 3)

Hit	CRM mass (Da)	Sequence	Modification Site	Formula	Ppm	Residual mass
1	149.0471	VFANPEDCAGFGK	C (8)	$C_8 H_7 N O_2$	-4	2
				$C_6 H_5 N_4 O$	5	86
2	149.0475	EFTPCAQAAFQK	C (5)	$C_8 H_7 N O_2$	-2	2
				$C_6 H_5 N_4 O$	7	86
3	149.0468	TIQLNVCNSEEVEK	C (7)	$C_6 H_5 N_4 O$	3	86
				$C_8 H_7 N O_2$	-6	2

Table S7. Three highest ranked reactive metabolite assignments produced for one of the APAP replicates using the non-targeted reactive metabolite identification algorithm. (APAP replicate 3 of 3)

Hit	CRM mass (Da)	Sequence	Modification Site	Formula	Ppm	Residual mass
1	149.0506	VFANPEDCAGFGK	C (8)	C_8 H_7 N O_2	20	2
2	149.0493	TIQLNVCNSEEVEK	C (7)	C_8 H_7 N O_2	11	2
3	151.0261	KDAQTLYDAGEK	G(10)	C_7 H_5 N O_3	-6	32

Table S8. Parameters for targeted detection of NAPQI-modified peptides

Parameter	Value	Units
Peptide Selection		
Threshold	Identity	
Charge state range	2-4	
HiTIME Selection		
MS^2 *m/z* offset	0.5	*m/z*
MS^2 RT offset	2	minutes
EIC toleranc	0.03	*m/z*
Neutral isotope mass difference	6.0201	Da

Table S9. Summary of the peptide signals created for each data set and the discovery rate of these peptides for HiTIME and non-targeted searching (NTS) analysis.

Experiment	Simulated Peptides	Post HT-Processing	NTS < 20 ppm	NTS < 100 ppm
Control 1	620	590 (95 %)	526 (89 %)	579 (98 %)
Control 2	560	539 (96 %)	467 (87 %)	530 (98 %)
Control 3	610	582 (95 %)	523 (90 %)	574 (99 %)

$$\text{Score} = \text{n} \sum \text{i}_\text{m}$$

Equation S1. Equation used to score the fit between the experimental MS^2 spectra of HiTIME hit peptides and theoretical ion sets generated from Mascot assignments where n is the number of fragment ions matched for a given pair of spectra and $\sum \text{i}_\text{m}$ is the total intensity of matched ions.

[!$([NH]!@C(=O))&!D1&!$(*#*)]-&!@[!$([NH]!@C(=O))&!D1&!$(*#*)]

Equation S2. SMARTS string used to identify rotatable bonds in the structure of administered drug substrates.

$$RME = \sum \left| n_f^i - n_d^i \right| \cdot m^i$$

Equation S3. Calculation of residual mass error for rank ordering of candidate molecular formulae where $\left| n_f^i - n_d^i \right|$ is the absolute difference in stoichiometry of element *i* between the candidate CRM formula *f* and the input drug fragment *d* and m^i is the molar mass of element *i*.

Image 1. Screen capture of the Postprocessing setup dialog from the Xenophile software.

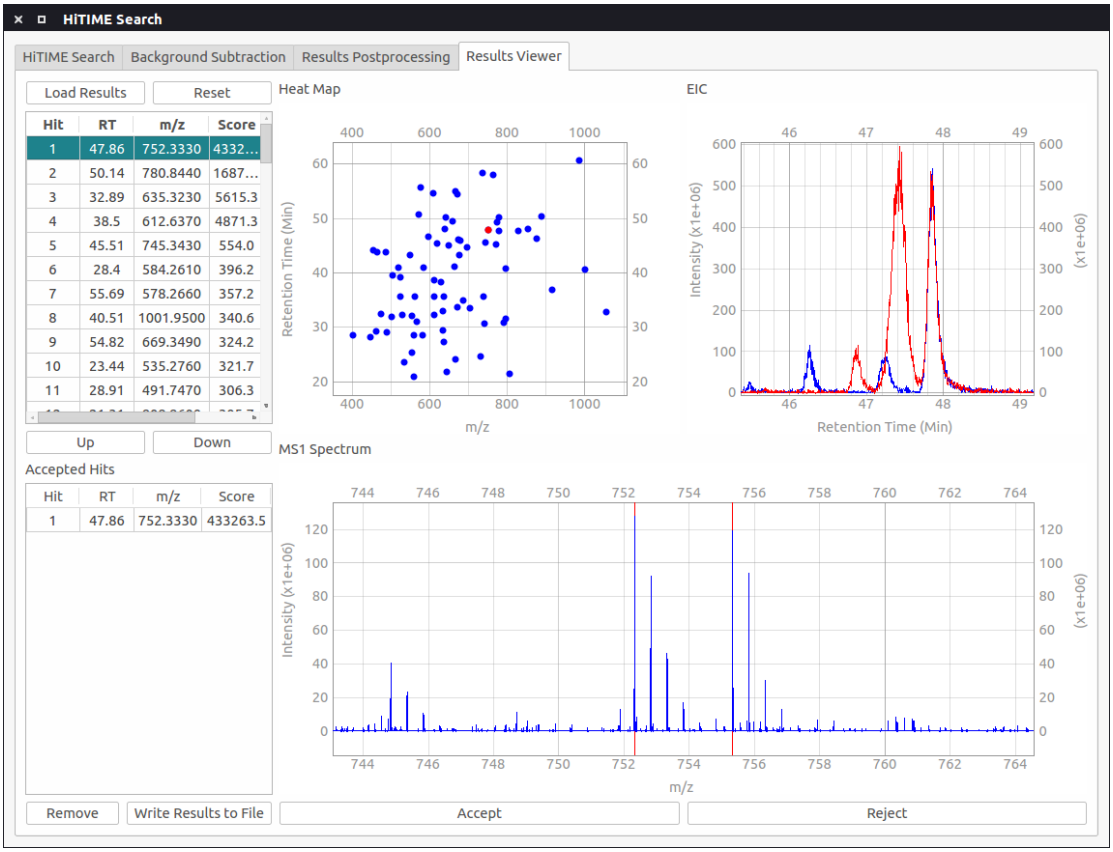

Image 2. Screen capture of non-targeted CRM search setup dialog from the Xenophile software.

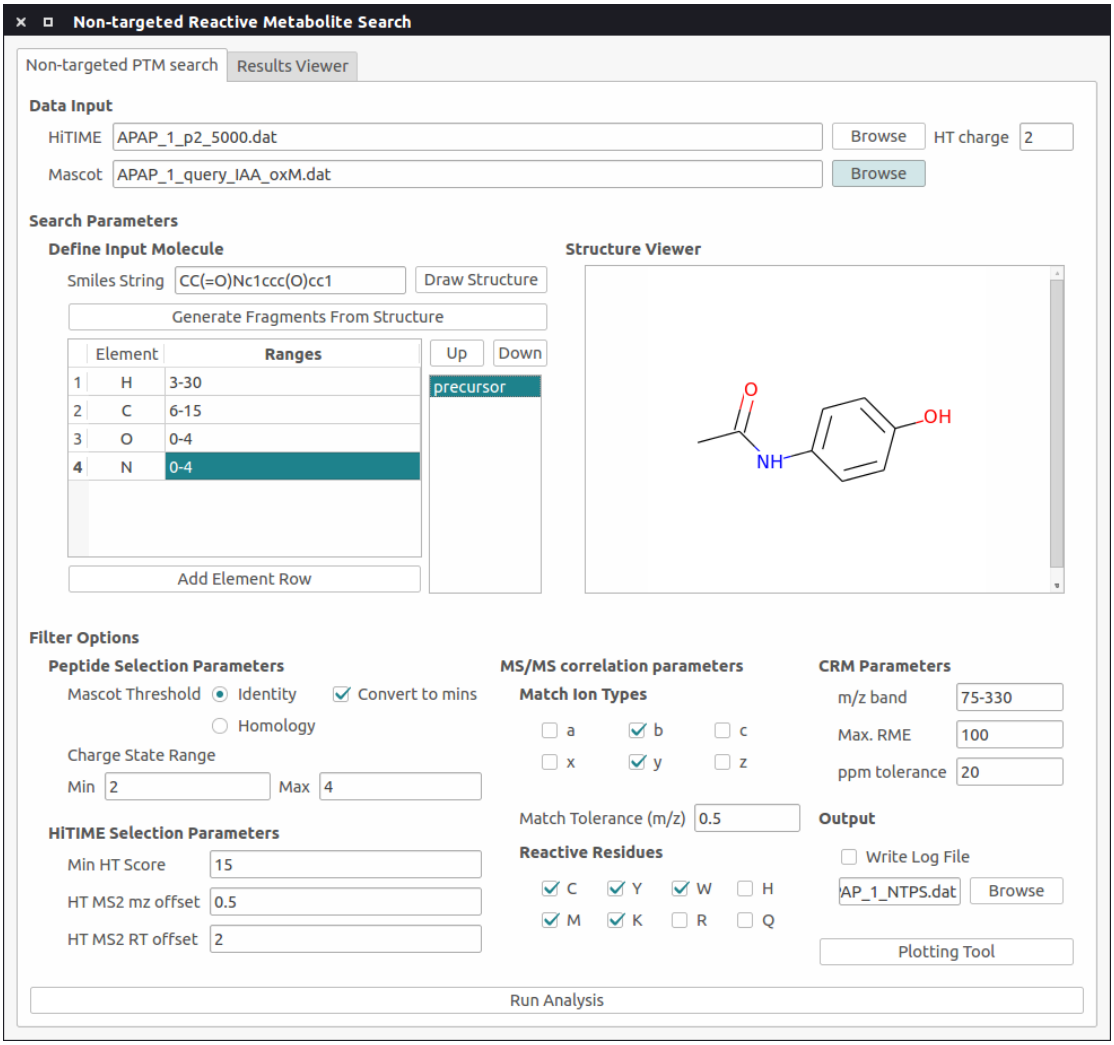

Image 3. Screen capture of non-targeted CRM search setup dialog from the Xenophile software.

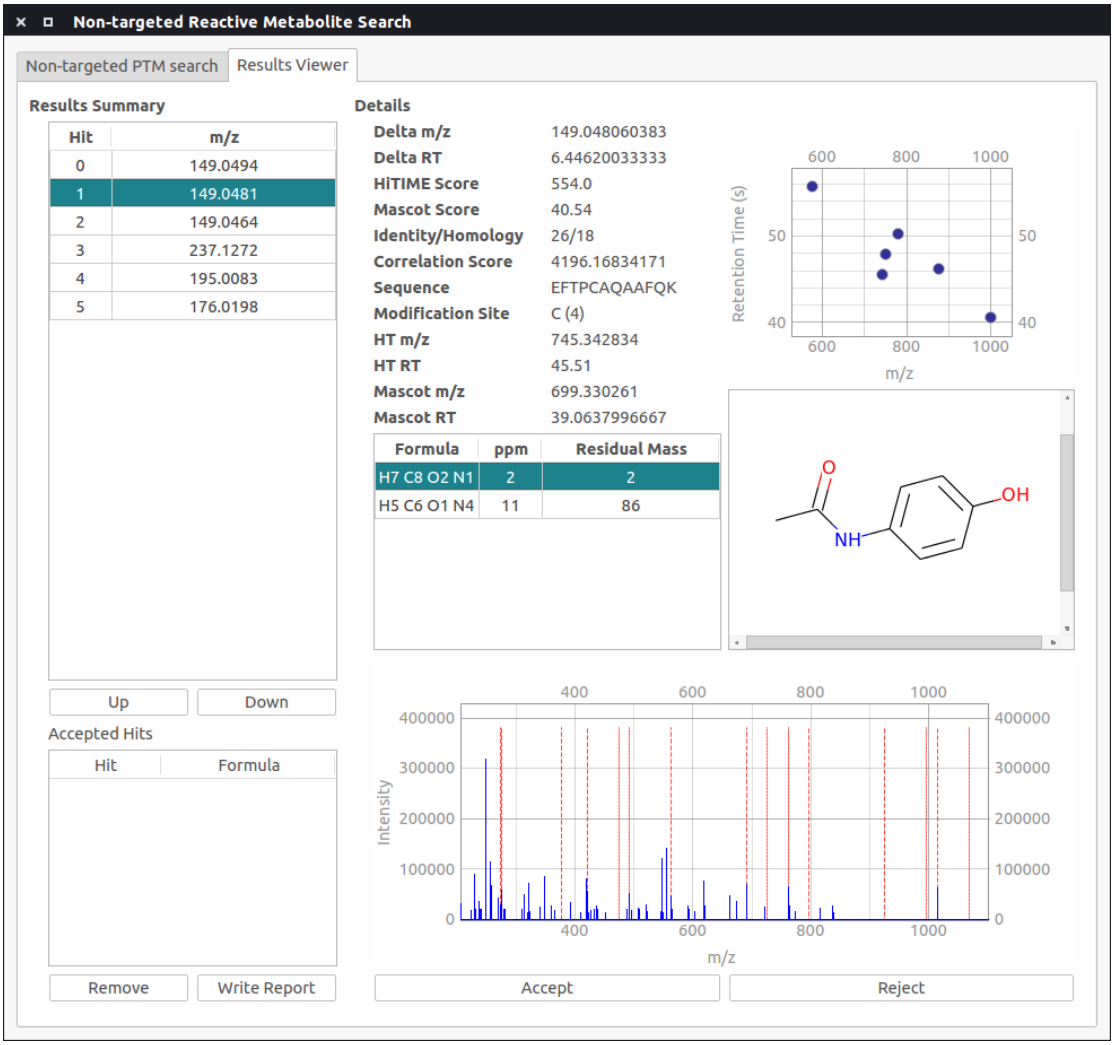

Image 4. Screen capture of non-targeted CRM search results analysis dialog from the Xenophile software.

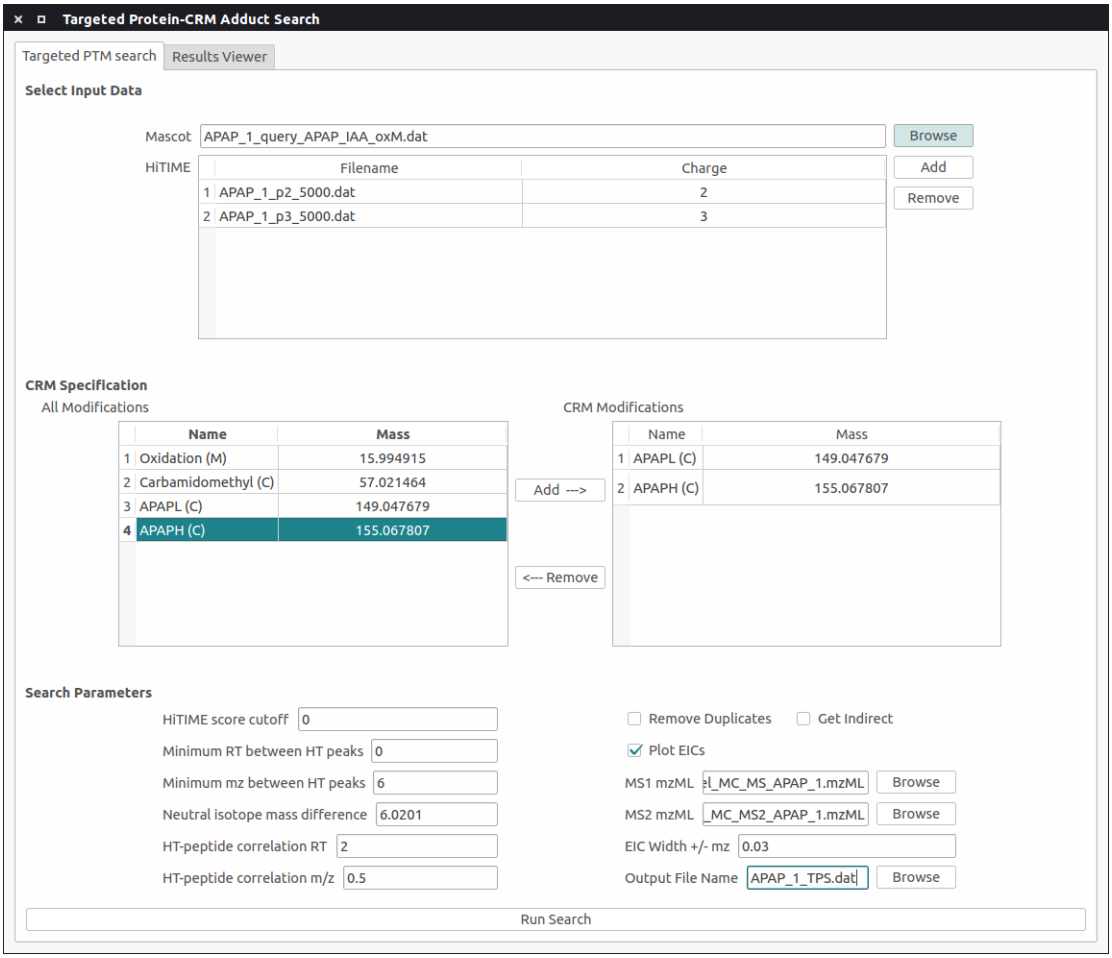

Image 5. Screen capture of targeted correlation search setup dialog from the Xenophile software.

Image 6. Screen capture of targeted correlation search results analysis dialog from the Xenophile software.

References to the supplementary information

(1) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Römpp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P.-A.; Deutsch, E. W., *Mol. Cell. Proteomics* **2011,** *10*, R110-000133.
(2) Eng, J. K.; McCormack, A. L.; Yates, J. R., *J. Am. Soc. Mass. Spectrom.* **1994,** *5*, 976-989.
(3) Yates, J. R.; Eng, J. K.; McCormack, A. L.; Schieltz, D., *Anal. Chem.* **1995,** *67*, 1426-1436.
(4) Leeming, M. G.; Isaac, A. P.; Pope, B. J.; Cranswick, N.; Wright, C. E.; Ziogas, J.; O'Hair, R. A. J.; Donald, W. A., *Anal. Chem.* **2015,** *87*, 4104-4109.
(5) Bald, T.; Barth, J.; Niehues, A.; Specht, M.; Hippler, M.; Fufezan, C., *Bioinformatics* **2012,** *28*, 1052-1053.
(6) Goloborodko, A. A.; Levitsky, L. I.; Ivanov, M. V.; Gorshkov, M. V., *J. Am. Soc. Mass. Spectrom.* **2013,** *24*, 301-304.
(7) Bielow, C.; Aiche, S.; Andreotti, S.; Reinert, K., *J. Proteome Res.* **2011,** *10*, 2922-2929.
(8) Bienfait, B.; Ertl, P., *J. Cheminform* **2013,** *5*, 1-6.
(9) Sturm, M.; Bertsch, A.; Gröpl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O., *BMC Bioinformatics* **2008,** *9*, 1-11.
(10) Röst, H. L.; Schmitt, U.; Aebersold, R.; Malmström, L., *Proteomics* **2014,** *14*, 74-77.
(11) Sterling, T.; Irwin, J. J., *Journal of Chemical Information and Modeling* **2015,** *55*, 2324-2337.