Supporting Information

metabCombiner: Paired Untargeted LC-HRMS Metabolomics Feature-Matching and Concatenation of Disparately Acquired Datasets

Hani Habra[1‡], Maureen Kachman[2], Kevin Bullock[3], Clary Clish[3], Charles R. Evans[2*], Alla Karnovsky[1,2*]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI,
[2]Michigan Regional Comprehensive Metabolomics Resource Core, University of Michigan, Ann Arbor MI, [3]Metabolomics Platform, Broad Institute, Cambridge, MA 02142, USA.

Email: hhani@umich.edu

Contents

Table S1: Initial Report Table Size by m/z binGap

S1: Guide to Choosing *A, B, C* Weight Coefficients in calcScores()

S2: Guide to Reducing Feature Pair Alignment Table

Table S2: Conflicting Feature Pair Alignment Examples and Decisions

S3: Human Plasma RPLC-POS Datasets Experimental + Analysis Information

S4: Human Urine HILIC-POS Datasets Experimental + Analysis Information

S5: Rat Muscle RPLC-NEG Datasets Experimental + Analysis Information

S6: Guide to Step-Wise Alignment of Additional Metabolomics Datasets

Figure S1: Plasma Datasets RT Fitting Evaluation Accuracy (Observed RT vs Fit Error)

Figure S2: Urine Datasets RT Fitting Plot Image: Unsupervised vs Semi-supervised

Figure S3: Muscle Datasets RT Fitting Plot Image- Before and After RT Filtering

Table S1.

Total Initial Feature Pair Alignments			
binGap	Plasma Datasets	Urine Datasets	Muscle Datasets
0.001	9055	7416	1929
0.0025	10758	15953	2731
0.005*	12330	31248	3247
0.0075	14024	55224	3603
0.01	16201	95898	4128

Table S1 Initial feature pair alignment (FPA) table sizes for each of the analyses described in this study, as a function of *binGap* parameter in the m/z grouping step. The default *binGap* value of 0.005 is typically sufficient to discover most matching features, but it may miss feature matches with larger m/z distances. Increasing the *binGap* value helps uncover matches with poor m/z agreement, with the disadvantage of increasing initial FPA table sizes.

S1: Guide to Choosing *A, B, C* Weight Coefficients in calcScores()

A, B, C weight arguments penalize differences in m/z, retention time (fitted vs observed), and relative abundance for a proposed pair of complementary dataset features, according to the expression:

$$S(F_x, F_y) = \exp\left(-A|mz_y - mz_x| - B\frac{|rt_y - f(rt_x)|}{range(rt_y)} - C\,|Q_y - Q_x|\right)$$

The choice of weight coefficients A, B, C should be considered carefully based on the underlying data. We divide this guide into two cases: when the datasets contain a sufficiently representative overlap of shared identified compounds (i.e. spanning the chromatogram, high and low relative abundance levels, and approximating the overall m/z deviation patterns), and when they do not.

For the first case, the *evaluateParams* function helps guide selection of coefficients by evaluating an objective function based on known matches. Compounds with matching identity strings (idx = idy, case-insensitive, ignoring empty or bracketed strings) serve as a guide to this optimization. Our goal is to maximize the scores of FPAs of true compound matches, minimize the scores of mismatches, and penalize wherever the score of a mismatch exceeds the true match. We denote *S(X, Y | A, B, C)* as the score between two grouped features X and Y. For shared compounds *i = 1, …, N*, let T_i = S(i, i | A, B, C) be the true FPA alignment score of compound i, and F_i = S(i, j | A, B, C) [i ≠ j] be the score of the highest-scoring mismatch to compound i within the same m/z group (F_i = 0 if there are no misalignments with i). We maximize through a grid search of A, B, and C values an objective function whose expression is as follows:

$$Obj(A,B,C) = \sum_{i}^{N} I(T_i) - I(F_i) - J(F_i > T_i)$$

$$I(x) = \begin{cases} x & if\ x \geq k \\ 0 & otherwise \end{cases} \qquad J(x) = \begin{cases} p & if\ True \\ 0 & if\ False \end{cases}$$

I is analogous to rectified linear unit (RLU) functions, penalizing scores that fall below a fixed threshold *c* (e.g. c = 0.5) and *J* subtracts a constant penalty term whenever a false match (F_i) outscores the true match (T_i). For fixed *c*, increasing the penalty value *p* maximizes the instances of top-scoring true matches among the shared identified compounds. The resulting matrix is sorted in reverse order of total objective score, with the values on top providing the best A, B, C coefficient combinations to use for calcScores. We recommend performing this workflow using different values for *c* and *p* to obtain a better idea of the optimal coefficient region.

In the absence of sufficient prior knowledge of overlapping compounds, here are some qualitative guidelines. As detailed in the text, the most effective range of values used in this study A, B, C are 50 - 120, 5 - 15, & 0 – 0.5, respectively. For the m/z penalty weight *A,* one must consider instrument mass accuracy, instrument similarity, and software m/z calculations used to obtain the datasets. Lower mass accuracy and/or precision should garner smaller *A* values (e.g. 50-70), whereas those obtained with high mass accuracy instruments can afford to be higher (e.g. 100-120). Key indicators of mass accuracy include deviation between observed vs theoretical m/z for known metabolites, deviation between m/z values for shared known metabolites, and deviation of mass differences between adducts from their theoretical values. The rt fit penalty weight *B* can mostly be informed through the plot of the *metabCombiner* fit. For a well-defined mapping, such as that observed in the plasma datasets in this study, we recommend a weight between 13-15 (higher if obtained from close to identical chromatographic conditions). In a model fit for which a high degree of sparsity and fluctuations about the trendline is observed, B should be assigned a lower weight, e.g. 5-7. Those in between should be assigned between 8-12. Smaller values on this continuum should be used for Y datasets with shorter overall chromatography times (< 10 min). Finally, we surveyed values between 0 and 1 for C, largely based on how biologically similar the samples from either dataset are and how similar the acquisition approach is. We mostly used values between 0.2 and 0.5 were used in the three reported analyses.

S2. Guide to Reducing Feature Pair Alignment Table

Reducing the table of FPAs by eliminating misaligned feature pairs is the final step in the *metabCombiner* analysis pipeline. A simplistic approach would be to retain only the top-ranked FPAs above some threshold score; however, sometimes lower-ranked matches have to be taken into account. A fully automated approach to eliminating misaligned pairs is difficult to implement, given the complex nature of untargeted LC-MS metabolomics data; therefore, we provide some guidelines here on how to reduce the table of FPAs as efficiently and accurately as possible.

The *labelRows* package function provides an automated row annotation method, assigning one of four values to each FPA: a) *"IDENTITY"* whenever identity strings match; b) *"REMOVE"* for FPAs that meet at least one of the removal criteria; c) *"CONFLICT"*, a flag for conflicting FPAs that may need further examination; d) "" (empty) whenever a FPA neither meets removal criteria, nor do they conflict with other FPAs. Feature pairs with matching identity strings are detected and labeled first; these shall not be regarded as removable, regardless of whether or not they meet exclusion criteria. Each m/z group is processed separately, using alignment score and pairwise rank (rankX & rankY) to determine row labels. First, the program determines whether a group is "balanced", defined as a group containing an equal number of features from both datasets with each feature assigned its own individual top-ranked FPA (i.e. rankX = 1 & rankY = 1). An example of this is shown in *Figure 3* in the main text, where three features from dataset X are aligned to three from dataset Y. In groups meeting this definition, all FPAs apart from top-ranked FPAs are labeled "REMOVE". Following this step, the program applies user-defined thresholds for alignment score and pairwise rank. If possible, thresholds should be guided by knowledge of shared known compounds; in our study, we have mostly used thresholds of 0.5 for score and 3 for ranks.

Subsequently, the remaining FPAs are assigned to subgroups if top-ranked matches and at least one alternative FPA (rankX > 1 or rankY > 1) meet certain conditions. There are two methods of subgroup assignment: 1) score-based 2) m/z & RT-based. In the first, a subgroup is assigned whenever the alignment score difference between a top-ranked and alternative FPAs sharing a feature is within some small threshold (e.g. Δscore = 0.2); in the second, candidate features with small m/z and RT distances between their unshared feature is within a set threshold. Examples of this process are illustrated in Table 2A & 2B. In Table 2A, two features from the X dataset are previewed as alignments to a single Y dataset feature labeled THAM (Tris(hydroxymethyl)aminomethane). One feature <122.081, 7.375> is a better candidate match to THAM, but users may consider a second feature <122.0806, 7.136> a potential match to this compound as well. The difference in scores between these candidate FPAs is small (0.05) and the m/z and RT distances of the two X dataset features are also very small (0.0004Da & 0.24 min). Therefore, both rows are flagged as "CONFLICT". Users may elect to discard the lower-ranked FPA, keep both rows, merge the X feature measurements (when applicable) or validate the correct match if necessary. In Table 2B, we have the similar situation that a single feature (labeled Phenylacetylglutamine) in datatset Y is compared to two possible candidate X dataset features; this time however, the difference in candidate FPA scores is much greater and their m/z & RT differences are further apart. Here the second-ranked FPA is annotated as "REMOVE."

While this process removes the majority of the misalignments, some additional criteria can help achieve further table reduction. Some of these may need to be customized according to the specific analyses.

1. Multiple Conflicts: FPA rows assigned to multiple conflict subgroups are nearly always fit for removal. A second column (alt) is provided for FPAs falling in multiple subgroups.

2. Sum of Ranks: While thresholds are applied in the above process for rankX or rankY individually, no rank sum criteria is implemented. A rank sum threshold of 4 would retain rank ordered pairs (1,1), (1,2), (2,1), (2,2), (1,3) and (3,1), efficiently removing all other lower-ranked feature pairs.

3. RT error tolerances: RT fitting errors can be used to either remove or flag certain candidate FPAs. The maximum allowable error may be guided by shared known compounds, if there are any. These may also be chromatographic region-specific; for example, if the early retention times are well-fit, then no fitting errors should exceed a specific number.

4. High m/z / RT / Q deviations: FPAs whose scores are on the border between inclusion or exclusion may require further scrutiny. Moderate to high errors in two or all three of these categories may indicate poor evidence of a matching compound and the FPA can be removed accordingly.

5. Adducts/ fragment information: the presence of multiple adducts can provide additional confidence of a proposed alignment or may indicate a potential erroneous match. In Table 2C below, Azelaic Acid [M+2Na-H]+ is aligned to four possible candidate features in the Y dataset; the presence of the Azelaic Acid [M+H]+ to a feature with the same retention time as the top-scoring candidate provides additional confidence to the top-scoring candidate FPA. The remaining FPAs may be discarded.

6. Retention Order: In certain cases, RT order may be used to correct and resolve conflict subgroups. An example is shown in Table 2D below for the two compounds, CAR 5:0 (valerylcarnitine) and CAR 4:0(3Me) (Isovalerylcarnitine), whose scored alignment order is erroneously rearranged. The retention order of the two compounds can be used to choose the two correct FPAs in the middle.

A number of FPAs may remain ambiguous without further laborious experimental validation or spectral visualization but taken together these steps are the most efficient guidelines for reducing the table to the most confident and accurate FPAs between matching metabolomics features.

Tables S2

A.

idx	idy	mzx	mzy	rtx	rty	rtProj	Qx	Qy	Score	rankX	rankY	label
	THAM	122.081	122.0811	7.375	8.052	8.115	0.905	0.543	0.844	1	1	CONFLICT
	THAM	122.0806	122.0811	7.136	8.052	7.93	0.897	0.543	0.784	1	2	CONFLICT

B.

idx	idy	mzx	mzy	rtx	rty	rtProj	Qx	Qy	Score	rankX	rankY	label
	Phenylacetylglutamine	265.1184	265.1182	4.781	5.979	5.9395	0.998	0.999	0.961	1	1	
	Phenylacetylglutamine	265.1161	265.1182	4.195	5.979	5.618	0.978	0.999	0.678	1	2	REMOVE

C.

idx	idy	mzx	mzy	rtx	Rty	rtProj	Qx	Qy	score	rankX	rankY	label
Azelaic Acid [M+H]+		187.097	187.097	10.821	6.02	6.006	0.973	0.960	0.981	1	1	
Azelaic Acid [M+2Na-H]+		209.0785	209.079	10.821	6.02	6.006	0.741	0.749	0.942	1	1	CONFLICT
Azelaic Acid [M+2Na-H]+		209.0785	209.079	10.821	5.76	6.006	0.741	0.5	0.806	2	1	CONFLICT
Azelaic Acid [M+2Na-H]+		209.0785	209.079	10.821	5.68	6.006	0.741	0.5335	0.792	3	1	CONFLICT
Azelaic Acid [M+2Na-H]+		209.0785	209.079	10.821	6.16	6.006	0.741	0.1381	0.788	4	1	CONFLICT

D.

idx	idy	mzx	mzy	rtx	rty	rtProj	Qx	Qy	score	rankX	rankY	label
CAR 4:0(3Me)	CAR 5:0	246.1706	246.1707	5.263	3.985	4.014	0.826	0.926	0.936	1	1	CONFLICT
CAR 5:0	CAR 5:0	246.1707	246.1707	5.395	3.985	4.089	0.898	0.926	0.903	1	2	CONFLICT
CAR 4:0(3Me)	CAR 4:0(3Me)	246.1706	246.1706	5.263	3.911	4.014	0.826	0.884	0.895	2	1	CONFLICT
CAR 5:0	CAR 4:0(3Me)	246.1707	246.1706	5.395	3.911	4.089	0.898	0.884	0.84	2	2	CONFLICT

Table S2 (A-D) Examples of conflicting FPA candidates. Rows highlighted in grey imply that these FPA rows should be eliminated. In (A) two candidate FPAs to the THAM (Tris(hydroxymethyl)aminomethane) compound are grouped, without a definitive one-to-one match assigned; in (B) the lower-scoring candidate is eliminated based on the existence of the higher-scoring FPA; in (C), the presence of a high-scoring [M+H]+ adduct match for Azelaic Acid enables a more definitive one-to-one assignment to the feature Azelaic Acid [M+2Na-H]+; in (D), the conflict subgroup containing CAR 5:0 & CAR 4:0(3Me) can be resolved by considering retention order.

S3 Plasma Datasets Information

Experimental Details

5 CHEAR & 5 Red Cross human plasma samples were thawed on ice prior to processing. For deproteinization in preparation for LC-MS analysis, 100 µL of plasma was combined with 400 µL 1:1:1 methanol:acetone:water containing the internal standards L-[15N] Anthranilic acid(5 uM), L-Epibrassinolide (20 uM). The sample was vortexed, then centrifuged (10 min at 15,000 x g). For reversed phase (RPLC)-MS analysis, the supernatant was transferred to a clean vial and dried under a stream of nitrogen gas. The dried sample was reconstituted at 50 µL MeOH: Water (50:50) containing Zeatin (1 uM). Samples were analyzed on an Agilent Infinity Lab II LC / 6545 qTOF MS system (Agilent Technologies, Inc., Santa Clara, CA USA) using a Waters Acquity HSS T3 1.8 µ, 100 mm column (Waters Corporation, Milford, MA). Mobile phase A is 100% water with 0.1% formic acid and mobile phase B was 100% methanol with 0.1% formic acid. The gradient for the 30-minute method is 0-2 min 2% B, 2-20 minutes 2-75% B (linear), 20-22 min 75-98% B (linear), followed by a 7 minute re-equilibration at starting conditions. The gradient for the 20-minute method is: 1-16 minutes 0-99% B (linear), 16-20 min 99% B (hold); 20 min return to 1% B, followed by a 4 minute re-equilibration at starting conditions. The flow rate for both methods is 4.50 mL/min and the column temperature was 40°C. Mass spectrometry was performed by electrospray ionization with an Agilent Jetstream ion source, with full-scan mass spectra acquired over the m/z range 50-1000 Da. Source parameters were: drying gas temperature 350°C, drying gas flow rate 10 L/min, nebulizer pressure 30 psi, sheath gas temp 350°C and flow 11 L/min, and capillary voltage 3500V, with internal reference mass correction. Five pooled plasma aliquots and two negative control blanks were analyzed alongside the experimental samples.

Data Processing

All Agilent .d files were converted to .mzML format using the MSConvert tool in Proteowizard [1] and processed with XCMS v. 3.6.1 [2]. Peak picking is performed with the Centwave algorithm [3], with peak width = 3-30s, noise = 1000, ppm = 20, prefilter = c(1,3000), snthresh = 10, integrate = 1, fitgauss = TRUE. RT correction follows with the obiwarp [4] method, with profstep = 0.5, and then peak grouping using the default density method, bw = 2, mzwid = 0.03, and minfrac = 0.5. Gap-filling is then applied using fillPeaks(). We apply negative control blank sample filtering, removing all features whose median pooled intensity to median blank intensity ratio is less than 2.5. We then search for and remove C^{13} isotopologues, defined by m/z differences of 1.0033 / z for charges (z = +1, +2, +3, +4, +5), log-scaled intensity value correlation > 0.5, RT tolerance of 0.03 min (1.8s), and m/z tolerance of 0.003 Da; median isotoplogue intensity values must also meet similar theoretical thresholds as those defined in the CAMERA R package[5]. Additional isotopologues, including Cl^{37} and S^{34}, are later annotated and removed using *Binner* v. 1.0.0 [6]. The feature counts generated are 8910 and 8286 for the 20- and 30-minute datasets, respectively.

Metabolite Identification

Metabolites were identified by matching the retention time (+/- 0.1 min), mass (+/- 10 ppm) and isotope profile (peak height and spacing) to authentic standards. Adduct and fragments of these known compounds were annotated using *Binner* and a custom R script searching for known based on mass-based rules, with a m/z tolerance of up to 0.01 Da and RT tolerance from the main peak of 0.03 min. Annotations were manually inspected for validity, inserting and correcting annotations to maximize the overlapping features for subsequent evaluation.

metabCombiner Analysis

One set of samples (CHEAR or Red Cross) of the 30-minute dataset is aligned to the complementary subset in the 20-minute dataset. Two analyses (unsupervised & semi-supervised) were performed in each case, with four analyses in total. No features were filtered from the 30-minute dataset; in the 20-minute dataset, the max retention time is set to 17.25, removing 3 features in the tail region. The m/z grouping *binGap* value is set to 0.0075, generating 14024 possible FPAs. Compounds are split into 50% training and 50% test sets, with test set compound identity strings enclosed in brackets (e.g. {Leucine}), rendering them invisible to *metabCombiner* string-matching operations. In selectAnchors, *useID* is set to TRUE (semi-supervised analysis using all training set compounds) or FALSE (no prior knowledge used), depending on the analysis; other parameters are windx = 0.03, windy = 0.03, tolmz = 0.003, and tolQ = 0.3. In the GAM-fitting step, iterFilter (# outlier filtering iterations) is set to 2, with {12,14,16,18,20} as possible values for k. In calcScores, score coefficients were set to A = 100, B = 15, and C= 0.3. We evaluated RT fitting and feature matching performance on test set compounds as described in the main text and proceeded with table reduction for the best-performing alignment (semi-supervised CHEAR 30-minute to Red Cross 20-minute samples). *labelRows* parameters maxRankX = 3, maxRankY = 2, minScore = 0.5, method = “score”, and delta = 0.2. Applying additional rules described in S2, such as removal of FPAs assigned to multiple conflict subgroups, rank sums greater than 4, harboring RT errors in excess of the highest error (0.35) or higher than 0.2 in well-predicted chromatographic regions, as well as accounting for retention order, resulted in a reduction to 6862 rows.

S4 Urine Metabolomics Datasets Information

IC43 Dataset

Experimental Details
Experimental details have been published previously in Blaženović et al. (2019) [7].

Data Processing
Sample files were downloaded from Metabolomics Workbench [8] accession ST0001122 and processed using MZMine2 [9] version 2.42 with the following steps and parameters: *Mass Detection*: detector mode set to "centroid", noise level =10000; *ADAP Chromatogram Builder*: min Group in # Scans = 5, Group Intensity Threshold = 20000, Min Highest Intensity = 50000, m/z Tolerance = 0.01; *Chromatogram Deconvolution*: Algorithm = "Wavelets (ADAP)", S/N Threshold = 100, Coefficient/Area Threshold = 120, Peak Duration Range = 0.05-0.75, RT Wavelet Range = 0.03-0.5 [10]; *Isotope Peak Grouper*: m/z tolerance = 0.01, Retention Time tolerance = 0.05, Representative Isotope set to "minimum m/z"; *RANSAC Peak Aligner*: m/z tolerance = 0.01, RT Tolerance (Before Correction) = 0.15, RT Tolerance (After Correction) = 0.1, Threshold Value = 0.1; Same RT and m/z Range Gap Filler: m/z tolerance = 0.01. *Duplicate Peak Filter*: Mode set to "SINGLE", m/z tolerance = 0.01, RT tolerance = 0.05. This generated an initial feature count of 22313.

Metabolite Identification
Metabolites were identified at MSI levels 1 & 2 as previously described [7]. The m/z and RT values of these metabolites were searched in the processed dataset using a custom R script. Compound identities were drawn from Table S2 of the published manuscript, labeled as either mzrt matches or MS2 matches; an in-house spectral library (*uclib*) provided by Dr. Blaženović was also used in this search. We require the presence of a common adduct (e.g. [M+H]+, [M+Na]+, [M+H-H2O]+, …) for inclusion of a named compound into this analysis. m/z and RT tolerance parameters for the search were set to 0.005Da and 0.2 min. The automated feature matches were carefully examined for duplicate and conflicting identity assignments, weighing m/z and RT errors as well as library type. "mzrt" named matches were prioritized, followed by MS2, and lastly uclib library matches. The names found in each category were 101 mzrt, 199 MS2, and 144 uclib matches.

B3N3 Dataset

Experimental Details
Human urine samples were analyzed using a method similar to that described by Blaženović et al. (2019), with slight but nevertheless significant alterations. Three replicates of pooled human urine samples were used from two separate sources for a total of six samples. The first source is National Institute of Standards and Testing (NIST) Standard Reference Material SRM3673. The second source is a pooled healthy human urine sample obtained from BioIVT (Westbury, NY). Both samples were thawed and extracted using a biphasic aqueous / methanol tert-butyl ether solvent system exactly as previously described; only the aqueous layer was used for subsequent HILIC analysis. Samples were analyzed using an Agilent 6545 LC-qTOF mass spectrometer (as opposed to a Thermo Q-Exactive LC-MS as in the above reference). The chromatographic column was a Waters Xbridge Amide 1.7 um, 2.1 mm ID, which was 100 mm in length as opposed to the prior reference's 150 mm. Chromatographic approach, including gradient length and mobile phase composition, were replicated as previously described. For MS1 sample analysis, MS source conditions were as follows: Agilent Dual Jetstream ESI, positive ion mode, source gas temp 275 C, drying gas 12 L/min, nebulizer 45psi, sheath gas temp 325 C, sheath gas flow 12 L/min, capillary voltage 4000, MS scan range 50-1200 Da, 2 spectra/sec, reference mass correction enabled. For MS2 analysis, all parameters were the same except the MS2 scan range was 25-1200 Da with a rate of 2 spectra/sec, isolation width was narrow, collision energy was 20, 3 precursor ions were allowed per cycle with active exclusion enabled after 2 spectra for 0.5 minutes, with abundance dependent accumulation off. Four runs of iterative MS/MS (rolling-precursor ion exclusion between replicate LC injections of a sample) were used with a mass error tolerance of 20ppm and RT tolerance of +/- 0.5 min.

Data Processing
The six Agilent .d files were converted into .mzML format using the MSConvert tool and processed using MZMine2 v. 2.42, like with *IC43*. Parameter settings used to process datasets are listed as follows: Mass Detection: mode set to "centroid" with noise level = 500; ADAP Chromatogram Builder: min Group in # Scans = 3, Group Intensity Threshold =

2000, Min Highest Intensity = 3000, m/z Tolerance = 0.01; Chromatogram Deconvolution: Algorithm used is "Wavelets(ADAP)", S/N Threshold = 20, Coefficient/Area Threshold = 100, Peak Duration Range = 0.03-0.6, RT Wavelet Range = 0.02-0.25; Isotope Peak Grouper: m/z tolerance = 0.01, Retention Time tolerance = 0.05, Representative Isotope set to "minimum m/z"; RANSAC Peak Aligner: m/z tolerance = 0.01, RT Tolerance (Before Correction) = 0.1, RT Tolerance (After Correction) = 0.05, Threshold Value = 0.1; Gap Filling Peak Finder: m/z tolerance = 0.01, RT tolerance = 0.1. Duplicate Peak Filter: Mode set to "SINGLE", m/z tolerance = 0.01, RT tolerance = 0.1. This generated an initial set of 10624 features.

Metabolite Identification

For compound annotation, all MS2 data were loaded into Masshunter Qualitative Workflows and features were detected using the Find by Auto MS/MS tool. The resulting MS/MS spectra were then exported in MGF format. These data were then simultaneously searched against two MS/MS libraries using the NIST MSPepSearch [11] software tool (02/22/2019 version): the NIST 2017 tandem MS library and the MoNA LC-MS/MS positive mode library (http://massbank.us, downloaded 12/2019). The resulting MS/MS hits were considered "identified" (MSI level 2) if the NIST score was >650, the dot product score was >750, and visual review confirmed that the spectrum was a good match with multiple well-aligned fragment ions. We then used a custom R script to search the processed dataset for the metabolites named in the MS/MS search list. Like previously, library hits derived from lower-order fragments were not included in the search. The associated m/z and RT values were searched with m/z and RT tolerances of 0.007 Da and 0.25 min. In addition to the main adduct form annotated in the MS/MS search, other adduct and fragment variants of these metabolites were searched using their associated mass-based rules. These include [M]+, [M+H]+, [M+Na]+, [M+2Na-H]+, [M+NH4]+, [M+K]+, [2M+H]+, [2M+Na]+, [M+H-H2O]+, [M+H-NH3]+ and [M+H-HCOOH]+. Compound and adduct annotations were carefully examined for consistency and validity (e.g. the retention time & m/z deviated from theoretical m/z and rt, retention time deviation between variants). Annotations for which there was some uncertainty were eliminated or bracketed using {} braces.

***metabCombiner* Analysis**

We select *B3N3* as the "X" dataset and *IC43* as the "Y" dataset. In initial dataset filtering steps, RT ranges were set to 0.5 to 10.5 min (*B3N3*) and 0.5 to 11 min (*IC43*), excising sparse head and tail regions and reducing by 213 and 268 features, respectively. The default missingness threshold 50% reduced *B3N3* by an additional 88 features and by over 11000 features in IC43, which contains features lacking detected peaks across the majority of the samples. The final feature count stands at 10320 and 11014 for *B3N3* and *IC43*. This pair of datasets was analyzed in two phases: first, without using identity information (unsupervised) and second, using features with identity agreement for anchor selection and RT mapping (semi-supervised). In both phases, the m/z grouping *binGap* value is set to 0.01, generating an initial set of 95898 possible FPAs. In the first round, *selectAnchors* argument values are windx = 0.02, windy = 0.03, tolmz = 0.003, tolQ = 0.3, useID = FALSE, selecting 66 anchors; GAM fitting proceeded with with iterFilter = 1, {12,14, 16, 18, 20} as possible k values, with k = 16 chosen through cross validation. *calcScores* arguments in the initial analysis were set to A = 60, B = 8, C = 0.3. The table of FPAs was narrowed to alignments scoring above 0.4 for inspection of assigned identities. Feature pairs with identity agreement were used to enhance analysis in the second round. *selectAnchors* argument values are the same in the second round, but windx is changed to 0.03 and useID = TRUE, selecting 98 anchors; GAM fitting proceeded with iterFilter increased to 2 and the same possible *k* values used, with k = 18 determined through cross validation; *calcScores* arguments were the same as before, but B is changed to 7; *labelRows* arguments are method = "score", minScore = 0.5, delta = 0.2, maxRankX = 3 and maxRankY = 3. This reduces the table to around 4100 – 4200 FPAs. Another set of FPAs were removed if they belonged to more than one conflict subgroups, if the RT error exceeded 0.6-0.7 min, if both m/z and Q deviation were deemed high, if the sum of ranks exceeded 4, or if the B3N3 feature RT is greater than the IC43 feature RT (since the column length in the former is shorter, smaller RTs are expected). This results in a further reduction of about 800-900 FPAs, with a final count of 3265 FPAs.

S5 Test Case 3 Datasets Information

MiSE10 Dataset

Experimental Details

Frozen tissue samples were weighed into chilled, pre-tared Eppendorf tubes. An ice-cold mixture of of 8:1:1 methanol:water:chloroform containing a mix of stable isotope labeled internal standards was added, one sample at a time, at a ratio of 1 mL solvent per 50 mg frozen tissue. Immediately after solvent addition samples were homogenized and extracted by sonication with a Branson 450 probe sonicator (power level 4, 40% duty cycle) for 30 seconds. Extracted samples were allowed to rest on ice for 10 minutes, after which vials were centrifuged at 15,000x for 5 minutes. 200 uL of supernatant were transferred to glass vials with flat-bottom inserts and were dried under a stream of nitrogen gas. Samples were reconstituted in 50 uL of 8:2 water:methanol and were analyzed by LC-MS using an Agilent 1290 LC / 6530 qTOF MS system (Agilent Technologies, Inc., Santa Clara, CA USA). Apart from the mass spectrometer used, chromatography and mass spectrometry instrumentation and parameters are exactly as described for reversed-phase analysis of plasma samples (30-minute method) in S4.

Data Processing

Agilent .d files were converted to .mzML format using the MSConvert tool and processed with XCMS v. 3.6.1. Peak picking is performed with the Centwave algorithm, with peak width = 2-25s, noise = 250, ppm = 30, prefilter = c(1,500), snthresh = 5, integrate = 1, fitgauss = TRUE, mzCenterFun = "wmean." RT alignment follows with the obiwarp method, with profstep = 0.5. Peak grouping follows using the default density method, with bw = 3, mzwid = 0.025, and minfrac = 0.4. Gap-filling is then applied with the fillPeaks() method. All other XCMS arguments are set to their default values. We then search for and remove C^{13} isotopologues, defined by m/z differences of 1.0033 / z for charges (z = -1, -2, -3), log-scaled intensity value correlation > 0.5, RT tolerance of 0.03 min (1.8s), and m/z tolerance of 0.003 Da; isotoplogues must also meet similar theoretical intensity thresholds as those defined in the CAMERA R package. This generated a total set of 5335 features for this dataset.

Metabolite Identification

For metabolite identification, features were annotated based on accurate mass and retention time data derived from analysis of an in-house library of authentic reference standards using this chromatographic method. Once the base peak of each metabolite has been identified, adducts and fragments were annotated using a similar custom R script to that used in other datasets, with a RT tolerance of 0.05 and m/z tolerance of 0.005. The adducts and fragments searched include [M-H]-, [M+Cl]-, [M+COOH]-, [M-H-H2O]-, [M+Na-2H]-, [M+K-2H]-, [M-H+NaCOOH]-, [2M-H]-, [2M+Na-2H]-, and [2M+K-2H]-, with the appropriate mass relationship rules to the neutral mass and base peak retention time of the known metabolite. All annotations were manually checked for validity.

BrSE10 Dataset

Experimental Details

This is the "C18-neg: Reversed-phase C18 chromatography/negative ion mode MS detection to measure free fatty acids, bile acids, and metabolites of intermediate polarity" method as applied by the Broad Institute**.** Analyses of free fatty acids and bile acids were conducted using an LC-MS system comprised of a Shimadzu Nexera X2 U-HPLC (Shimadzu Corp.) coupled to a Q Exactive hybrid quadrupole orbitrap mass spectrometer (Thermo Fisher Scientific). Muscle samples were first homogenized in water using a TissueLyserII bead mill (4 uL water per mg of tissue), then extracted using 90 uL of methanol containing 15R-15-methyl-PGA_2, 15R-15-methyl-PGF_{2alpha}, 15S-15-methyl-PGD_2, 15S-15-methyl-PGE_1, and 15S-15-methyl-PGE_2 (Cayman Chemical Co.) internal standards and centrifuged (10 min, 9,000 x g, 4°C). The samples were injected onto a 150 x 2 mm ACQUITY BEH C18 column (Waters). The column was eluted isocratically at a flow rate of 400 µL/min with 60% mobile phase A (0.1% formic acid in water) for 4 minutes followed by a linear gradient to 100% mobile phase B (acetonitrile with 0.1% acetic acid) over 8 minutes. MS analyses were carried out in the negative ion mode using electrospray ionization, full scan MS acquisition over 70-850 m/z, and a resolution setting of 70,000. Other MS settings were: sheath gas 45, sweep gas 5, spray voltage -3.5 kV, capillary temperature 320°C, S-lens RF 60, heater temperature 300°C, microscans 1, automatic gain control target 1e6, and maximum ion time 250 ms.

Data Processing & Identification

Raw data were processed using TraceFinder software (Thermo Fisher Scientific) for targeted peak integration and manual review of a subset of identified metabolites, using Progenesis QI (Nonlinear Dynamics) for peak detection and integration of both metabolites of known identity and unknowns. Metabolite identities were confirmed using authentic reference standards. Adducts and fragments of detected metabolites were searched using a custom script similar to *MiSE10*, with a RT tolerance of 0.1 min and m/z tolerance = 0.005; annotations were manually inspected afterwards for validity.

***metabCombiner* Analysis**

Data columns corresponding to the exercised rat muscle samples (10 each) were selected in both datasets. Analysis is limited to RTs between 0.5 and 24 min for MiSE10 and between 0.5 and 17 min for BrSE10, eliminating 1177 and 1262 features, respectively; 19 and 866 features are further reduced from BrSE10 due to missingness and duplicate feature filters. The final feature counts are 4158 and 6426 for MiSE10 and BrSE10. m/z grouping binGap value is kept at the default 0.005Da, generating a list of 3247 possible FPAs. We performed a grid search of RT fitting parameters, optimizing for the mean prediction error of fourteen shared compound identities with additional guidance from the RT mapping plot; this generated a best parameter set of windx = 0.04, windy = 0.01, tolmz = 0.001, tolQ = 0.2 in *selectAnchors* & bs = "ps", iterFilter = 4, k = (10,12,14,16,18), family = "gaussian", and method = "GCV.Cp" (all other parameters kept to their default values). A total of 80 ordered pairs were selected for RT modeling and an optimal k value of 10 chosen through cross-validation. *calcScores* coefficients are set to A = 100, B = 7, C = 0.2 and *labelRows* arguments set to minScore = 0.35, maxRankX = 5, and delta = 0.25. Additional row inspection criteria, including multiple conflicts, row sums greater than 7, retention time fitting accuracy, and so forth reduced the set of FPAs from 3247 to 984 FPAs.

S6: Guide to Step-Wise Alignment of Additional Metabolomics Datasets

metabCombiner is designed for the pairwise alignment of two untargeted LC-MS metabolomics datasets. Additional functionality for the implementation of multi-dataset alignment is planned for metabCombiner, however the evaluation of these approaches is beyond the scope of the current study.

For the current package implementation, it is possible to perform stepwise alignment of additional datasets using the following procedure: for a given *metabCombiner* object containing the completed alignment analysis of two datasets, X and Y, extract the combined data table with the *combinedTable*() method, e.g.

```
combined.x.y <- combinedTable(object)
```

where "object" is the metabCombiner object containing the results of the alignment of datasets X & Y. Next, call *metabData()* with a combined data table as input to create a new *metabData* (single dataset object), setting the "mz", "rt", "id", "adduct", "Q", "samples" arguments to the X metadata or the Y metadata, and use the "extra" argument to bring forth any additional dataset columns that must be brought forth into the final resulting table, e.g.,

```
combined.x <- metabData(combined.x.y, mz = "mzx", rt = "rtx", id = "idx", adduct = "adductx", Q = "Qx",
samples = getSamples(object, "x"), extra = getSamples(object, "y"), (additional arguments))
```

Then call metabCombiner() with this metabData and another dataset "z" as inputs, e.g.

```
combined.x.y.z <- metabCombiner(xdata = combined.x, ydata = new.dataset.z, binGap = 0.005)
```

Next, follow the workflow steps outlined in the Fig 1 and the manuscript text. In the above example, this aligns intersected XY features with Z features, using X feature meta-data and Z feature meta-data, with Z becoming the new "Y"; the original Y feature meta-data is no longer used, but the Y dataset samples are included in the final table. One trade-off to this approach is that if a combinedTable is used as input into the *metabData* function, only the top-scoring alignment of X or Y will be used by default, removing information about alternative, lower-scoring matches. Moreover, information about features not present in at least one of the datasets may be lost as only the intersection of datasets is reported.

Figure S1

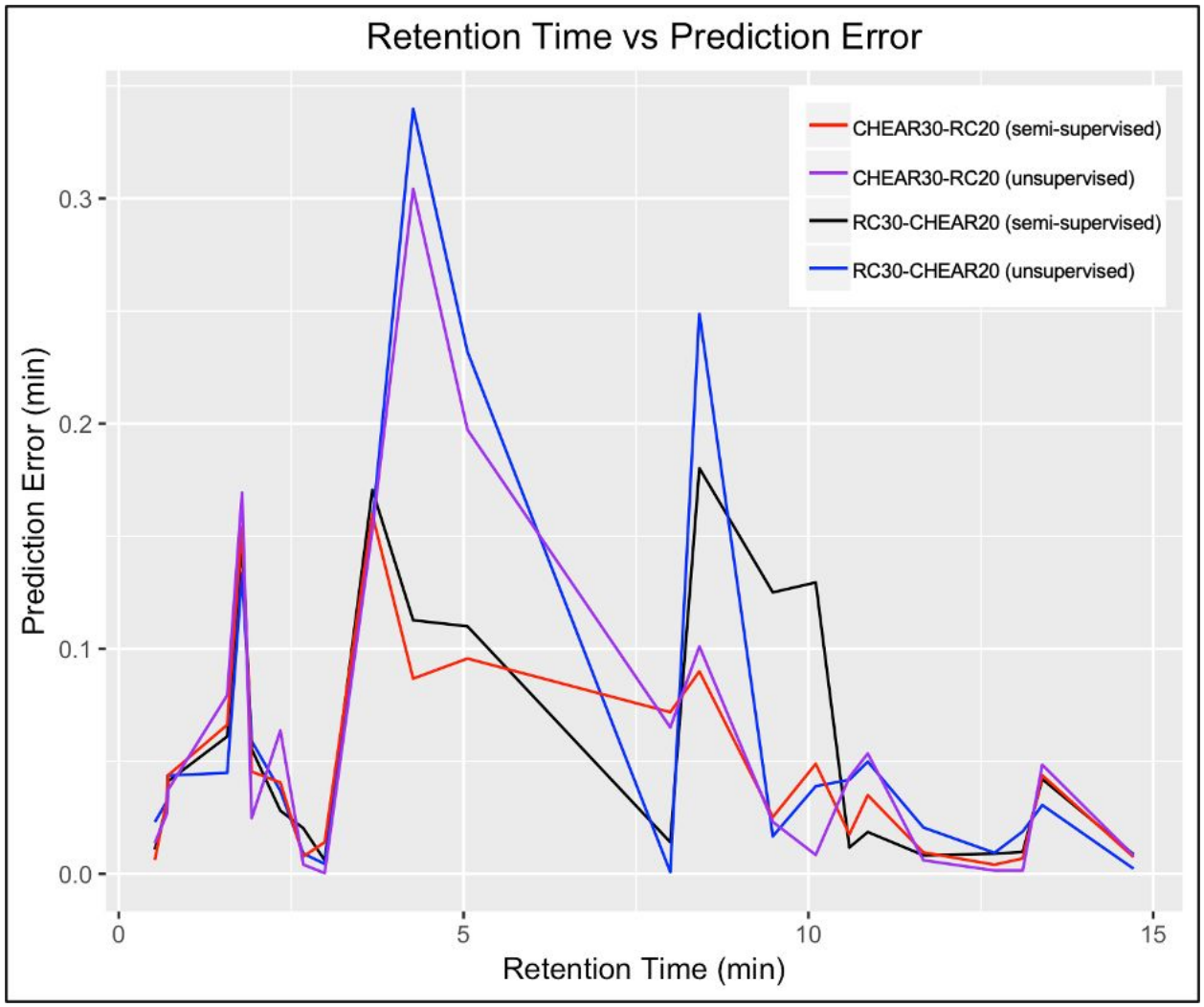

Figure S1 Comparison of RT fitting errors for a selected subset of known compounds in the plasma datasets. Different sample subsets (CHEAR or Red Cross) of the 30-minute dataset are aligned to the other sample subset in the 20-minute dataset. Fitted vs observed 20-minute dataset RTs are used for evaluation. RT errors vary by chromatographic region, with the highest observed errors observed along the gradient region between 4-10min. While fit errors vary by selected sample subsets, semi-supervised models utilizing all prior known compounds to select anchors improved model fits over unsupervised fits.

Figure S2

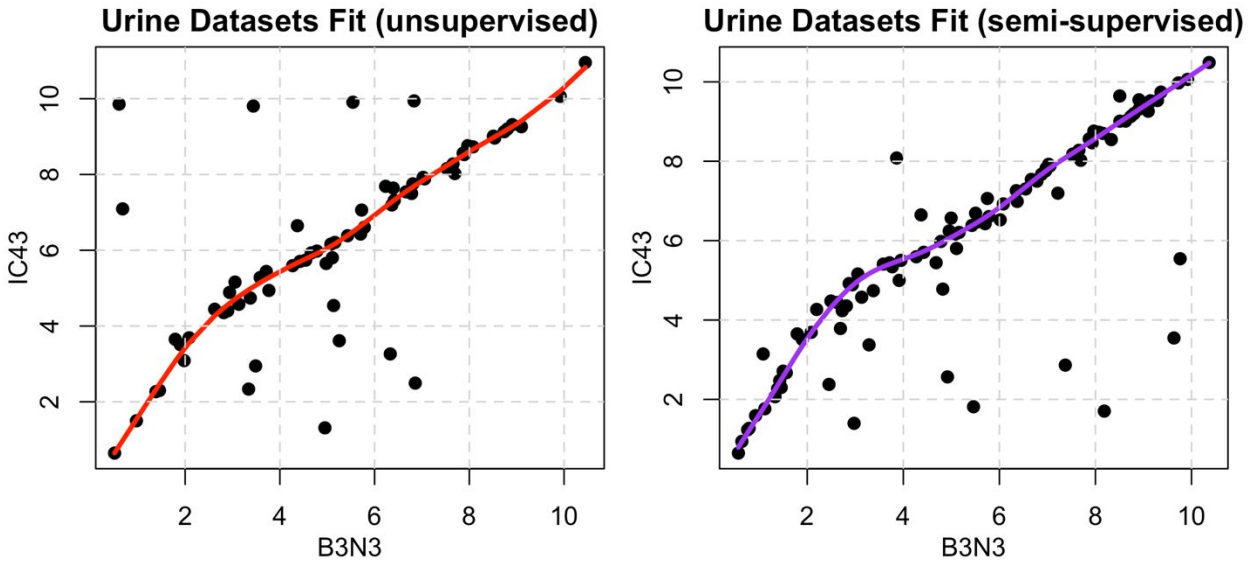

Figure S2 A side by side comparison of RT fits in the urine datasets analysis. In the unsupervised analysis (left image), no compound identity information was used; in the semi-supervised analysis (right image), features with identity agreement were incorporated as anchors. Using shared identities refines the RT mapping, especially in the sparsely-anchored early chromatographic regions.

Figure S3

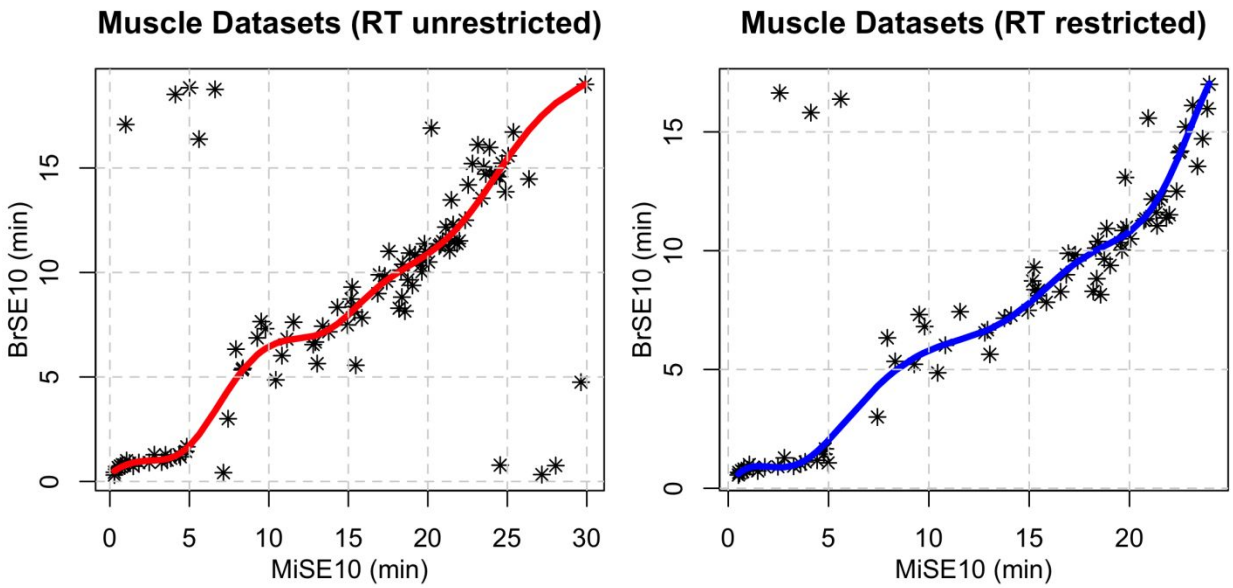

Figure S3 RT mapping image with and without RT range restrictions. Due to the lack of shared nonpolar metabolite coverage, the later RT regions (which contain most of the shared identified compounds) are poorly fit, regardless of anchor selection or GAM-fitting parameters. One effective solution is to excise the later chromatographic regions by setting a maximum RT in the early filtering steps. Setting a max RT at 24 min gives a more accurate mapping, illustrated in the image on the right.

References

(1) Kessner D; Chambers M, Burke R, Agus D, Mallick P. *Bioinformatics* **2008**, 24(21), 2534-2536.

(2) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Analytical Chemistry* 2006, 78, 779-787.

(3) Tautenhahn, R.; Bottcher, C.; Neumann, S. *BMC Bioinformatics* 2008, 9, 504.

(4) Prince, J. T.; Marcotte, E. M. *Analytical Chemistry* **2006**, 78, 6140-6152.

(5) Kuhl C.; Tautenhahn R.; Böttcher C.; Larson T.R.; Neumann S. *Analytical Chemistry* **2012**, 84(1), 283-289.

(6) Kachman, M.; Habra, H.; Duren, W.; Wigginton, J.; Sajjakulnukit, P.; Michailidis, G.; Burant, C.; Karnovsky, A. *Bioinformatics* **2020**, *36*, 1801-1806.

(7) Blaženović, I.; Kind, T.; Sa, M. R.; Ji, J.; Vaniya, A.; Wancewicz, B.; Roberts, B. S.; Torbašinović, H.; Lee, T.; Mehta, S. S.; Showalter, M. R.; Song, H.; Kwok, J.; Jahn, D.; Kim, J.; Fiehn, O. *Analytical Chemistry* **2019**, *91*, 2155-2162.

(8) Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K. S.; Sumner, S.; Subramaniam, S. *Nucleic Acids Res* **2016**, *44*, D463-470.

(9) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinformatics* **2010**, *11*, 395.

(10) Myers, O. D.; Sumner, S. J.; Li, S.; Barnes, S.; Du, X. *Analytical chemistry* **2017**, *89*, 8696-8703.

(11) Stein, S. E.; Scott, D. R. *Journal of the American Society for Mass Spectrometry* **1994**, *5*, 859-866.