metabCombiner Online v1.0 User Manual

Karnovsky Lab
Department of Computational Medicine and Bioinformatics
University of Michigan, Ann Arbor
hhani@umich.edu
https://karnovskylab.dcmb.med.umich.edu/metabCombiner/

August 23, 2023

Contents

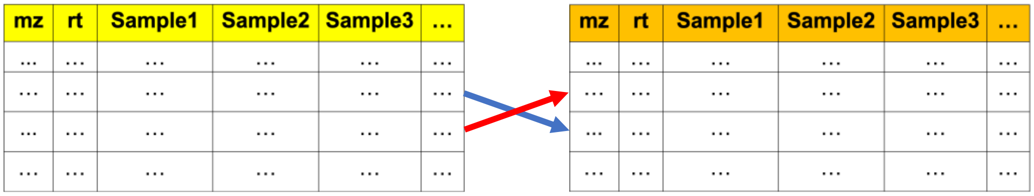
Chapter 1

Introduction

metabCombiner is a computational method and software package for aligning untargeted metabolomics data acquired under non-identical LC-MS conditions, such as by using different mass spectrometers, stationary phase columns, or differences in LC gradients. This method bridges between-laboratory and within-laboratory differences in protocols, as long as the data meets certain assumptions. *metabCombiner Online* is the web-based app version of *metabCombiner* developed using the Shiny R package. This document is a guide to using *metabCombiner Online*.

1.1 Overview

metabCombiner aligns two data sets at a time, one designated the "X" data set and the other the "Y" data set. In each alignment iteration, *metabCombiner* attempts to find the best matches for each feature in the complementary table, based on measured mass-to-charge (m/z), retention time (RT), and per-sample abundances. The respective abundances and other per-feature information are assembled into a combined feature table.



mz	rt	Sample1	Sample2	Sample3	…
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…

mz	rt	Sample1	Sample2	Sample3	…
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…

Figure 1.1: Two pre-processed LC-MS metabolomics tables, an "X" and "Y" data set

The primary inputs to the program are text files (.csv or .txt) representing conventionally preprocessed metabolomics feature tables, as generated by commercial or open-source preprocessing software (e.g. XCMS, MZmine3, MS-DIAL, etc...). The output is a "Combined Table" report listing the ranked matches among all of the paired features (constrained by m/z), as well as a *metabCombiner* object (writable as a .rds file) that may be be used as input for alignment with further metabolomics data sets.

1.2 Method Assumptions

The LC-MS metabolomics data sets to be aligned must share the following characteristics:

1) acquired in the same ionization mode (positive or negative)

2) contain samples representing specimens of a similar metabolic composition

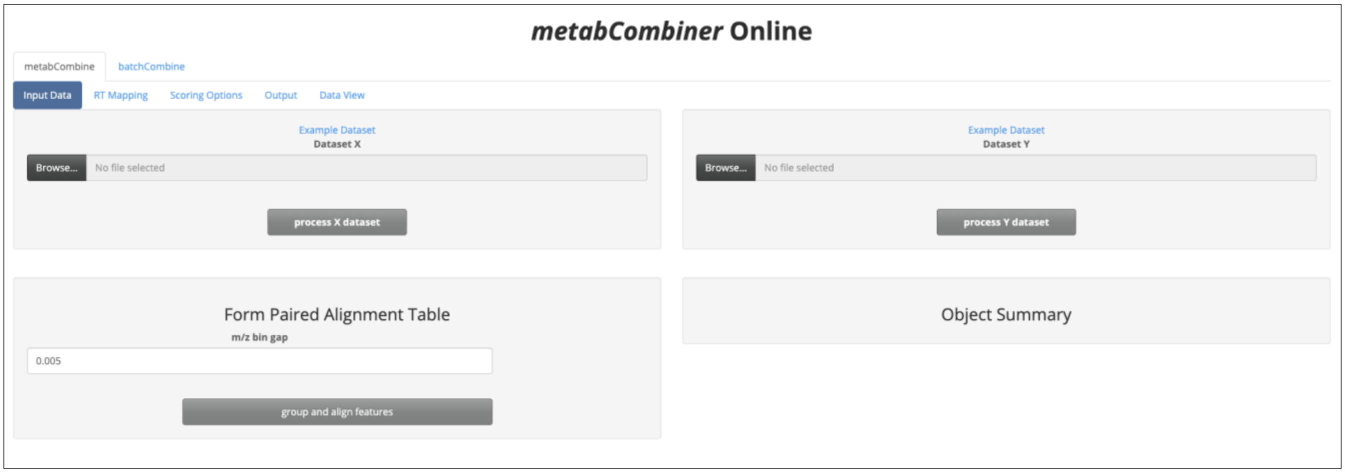
3) generated with similar liquid chromatography methods (elution order similarity)

4) no prior normalization that may distort the ranked order of abundances.

1.3 User Interface

The user interface of *metabCombiner Online* at launch is shown on **Figure 1.2**. The interface consists of two main tabs: *metabCombine* and *batchCombine*. *metabCombine* is designed for pairwise alignment of non-identically acquired LC-MS data. *batchCombine* is an extension of *metabCombiner* for aligning multiple batches of a single untargeted metabolomics experiment, with each batch represented as a separate table. Each tab is further sub-divided into further tabs. For the *metabCombine* method, the tasks must be performed in sequential order to complete the analysis.

In the **Input Data** tab, data sets are loaded into the program for alignment and the initial aligned table is formed. The RT projection model can be customized and viewed in the **RT Mapping** tab. Similarity scoring and feature alignment ranking is determined in the **Scoring Options** Tab. Finally, the **Output** tab contains options for printing both the Combined Table reports and *metabCombiner* objects to file, as well as using the output for further alignment tasks. A limited view of the data sets is displayed in the **Data View** tab.



m/z bin gap: 0.005

Figure 1.2 The User Interface upon launching the metabCombiner Online App

1.4 Formatting Data

There are two example data sets in the Input Data tab. These were constructed from untargeted LC-MS metabolomics of pooled human plasma, using two different LC gradients (30 minute and 20 minute total chromatography). A screenshot of the 30-minute data set is displayed in **Figure 1.3** below. While *metabCombiner* is flexible in how data may be formatted, there are some requirements that each input table must follow to be accepted by the program:

1) As with most pre-processed table inputs, each row must represent an individual feature, with important meta-data and sample abundance values represented by columns.

2) The first row must contain column headers. Avoid having special characters such as asterisks (*), apostrophes ('), or brackets (e.g. (), []. {}) in column names.

3) Each metabolomics data set consists of multiple components, listed below. We recommend placing the first four of these fields (a-d) among the first columns of the table (in any order), whereas the latter (e, f) two can be arranged in any manner throughout the remainder of the table.

a) m/z (mass-to-charge ratio): REQUIRED. Must be a numeric column with no missing or negative values. m/z values should display at least 4 decimal places.

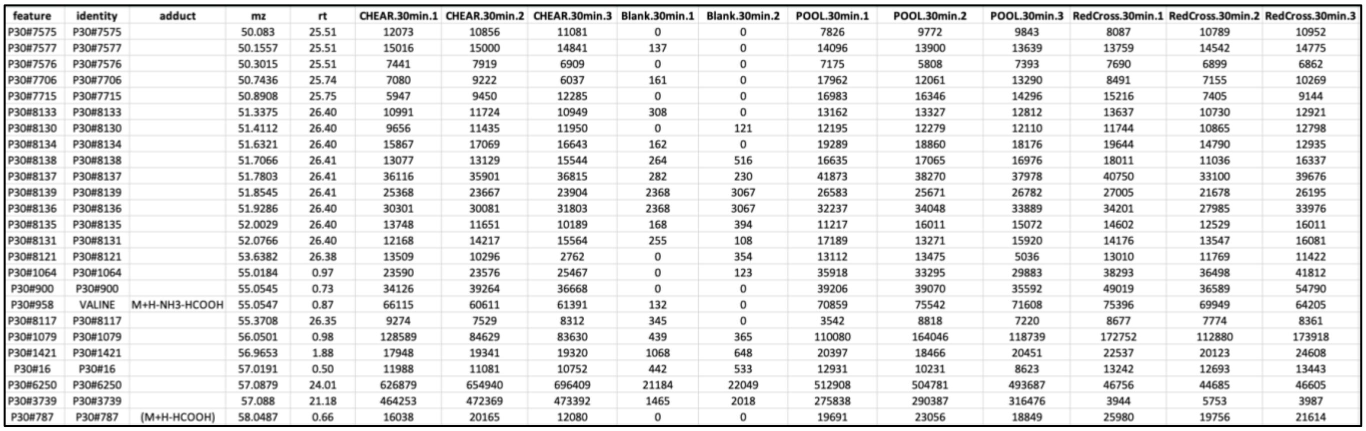
b) rt (retention time): REQUIRED. Must be a numeric column with no missing or negative values, with at least two decimal places. By default, program values assume retention times are in minutes, the recommended units for this application.

c) id (identifier): OPTIONAL. This column may contain character string identifiers for features, whether they be generic labels or known structural identities. Certain program functions can be enhanced with these ids (e.g. if names match between tables). Identifiers enclosed in brackets or parentheses are ignored by the program.

d) adduct: OPTIONAL. This denotes the feature's precursor adduct type (e.g. [M+H]+, [M+Na]+), or alternatively, chemical formulas.

e) samples: REQUIRED. There must be at least one numeric column in the table representing metabolite feature abundances within a sample. Program estimations of percent missingness and relative quantitation (Q) are based on columns in this field. It is best to have unique sample names across all the tables that are to be aligned, but with a common substring. In the example data sets, there are four different sample groups: "CHEAR", "Blank", "POOL", and "RedCross"

f) extra: OPTIONAL. This signifies any additional columns to include in the final combined feature table, but do not affect the alignment process. These may be character string or numeric columns. This may include samples whose columns appear in the final report but are not considered for quantitative comparisons.



feature	identity	adduct	mz	rt	CHEAR.30min.1	CHEAR.30min.2	CHEAR.30min.3	Blank.30min.1	Blank.30min.2	POOL.30min.1	POOL.30min.2	POOL.30min.3	RedCross.30min.1	RedCross.30min.2	RedCross.30min.3
P30#7575	P30#7575		50.083	25.51	12073	10856	11081	0	0	7826	9772	9843	8087	10789	10952
P30#7577	P30#7577		50.1557	25.51	15016	15000	14841	137	0	14096	13900	13639	13759	14542	14775
P30#7576	P30#7576		50.3015	25.51	7441	7919	6909	0	0	7175	5808	7393	7690	6899	6862
P30#7706	P30#7706		50.7436	25.74	7080	9222	6037	161	0	17962	12061	13290	8491	7155	10269
P30#7715	P30#7715		50.8908	25.75	5947	9450	12285	0	0	16983	16346	14296	15216	7405	9144
P30#8133	P30#8133		51.3375	26.40	10991	11724	10949	308	0	13162	13327	12812	13637	10730	12921
P30#8130	P30#8130		51.4112	26.40	9656	11435	11950	0	121	12195	12279	12110	11744	10865	12798
P30#8134	P30#8134		51.6321	26.40	15867	17069	16643	162	0	19289	18860	18176	19644	14790	12935
P30#8138	P30#8138		51.7066	26.41	13077	13129	15544	264	516	16635	17065	16976	18011	11036	16337
P30#8137	P30#8137		51.7803	26.41	36116	35901	36815	282	230	41873	38270	37978	40750	33100	39676
P30#8139	P30#8139		51.8545	26.41	25368	23667	23904	2368	3067	26583	25671	26782	27005	21678	26195
P30#8136	P30#8136		51.9286	26.40	30301	30081	31803	2368	3067	32237	34048	33889	34201	27985	33976
P30#8135	P30#8135		52.0029	26.40	13748	11651	10189	168	394	11217	16011	15072	14602	12529	16011
P30#8131	P30#8131		52.0766	26.40	12168	14217	15564	255	108	17189	13271	15920	14176	13547	16081
P30#8121	P30#8121		53.6382	26.38	13509	10296	2762	0	354	13112	13475	5036	13010	11769	11422
P30#1064	P30#1064		55.0184	0.97	23590	23576	25467	0	123	35918	33295	29883	38293	36498	41812
P30#900	P30#900		55.0545	0.73	34126	39264	36668	0	0	39206	39070	35592	49019	36589	54790
P30#958	VALINE	M+H-NH3-HCOOH	55.0547	0.87	66115	60611	61391	132	0	70859	75542	71608	75396	69949	64205
P30#8117	P30#8117		55.3708	26.35	9274	7529	8312	345	0	3542	8818	7220	8677	7774	8361
P30#1079	P30#1079		56.0501	0.98	128589	84629	83630	439	365	110080	164046	118739	172752	112880	173918
P30#1421	P30#1421		56.9653	1.88	17948	19341	19320	1068	648	20397	18466	20451	22537	20123	24608
P30#16	P30#16		57.0191	0.50	11988	11081	10752	442	533	12931	10231	8623	13242	12693	13443
P30#6250	P30#6250		57.0879	24.01	626879	654940	696409	21184	22049	512908	504781	493687	46756	44685	46605
P30#3739	P30#3739		57.088	21.18	464253	472369	473392	1465	2018	275838	290387	316476	3944	5753	3987
P30#787	P30#787	(M+H-HCOOH)	58.0487	0.66	16038	20165	12080	0	0	19691	23056	18849	25980	19756	21614

Figure 1.3 30-minute untargeted RPLC-MS metabolomics analysis of human plasma

1.5 *metabCombiner* Workflow

Table 1.1 below lists the major steps in the *metabCombiner* workflow. Features from a pair of input tables are first filtered and individually formatted in the ***Input Data*** step. Next, the feature lists are merged and grouped based on m/z similairities in the **Form Aligned Table** step. The pair of steps (**Anchor Selection** and **Model Fitting**) select ordered pairs and construct a warping model between retention times. The **Feature Pair Scoring** step calculates a similarity measure for determining the most likely feature matches. Finally, the **Row Reduction & Annotation** step is an automated determination of the most likely matches based on similarity scores and rankings.

The output of this sequence of operations can then be used as input in the first step. The *batchCombine* workflow performs this sequence of steps multiple times to sequentially align multiple batches, using the results of the preceding step as input for the next alignment between the batches. The remainder of this document provides step-by-step instructions for how to perform and customize parameters for each of these steps.

Step	Description
Data Input	Load input metabolomics data sets and perform feature filters
Form Aligned Table	Construct merged feature consisting of all possible feature pair alignments based on pairwise m/z distances
Anchor Selection	Determine ordered pairs for generating a spline mapping between data set retention times
Model Fitting	Fit a basis spline through selected anchors for retention time mapping
Feature Pair Scoring	Calculate similarity scores based on m/z, RT prediction error, and abundance quantile (Q) differences
Row Annotation & Reduction	Annotate and remove rows representing feature mismatches, retaining the most likely alignments

Table 1.1: List of *metabCombiner* Steps

1.6 Contact

Any questions related to using metabCombiner Online should be directed to hhani@umich.edu. You may also report any issues or errors of this program to the Github page link, https://github.com/hhabra/metabCombiner-Online , under issues.

Chapter 2

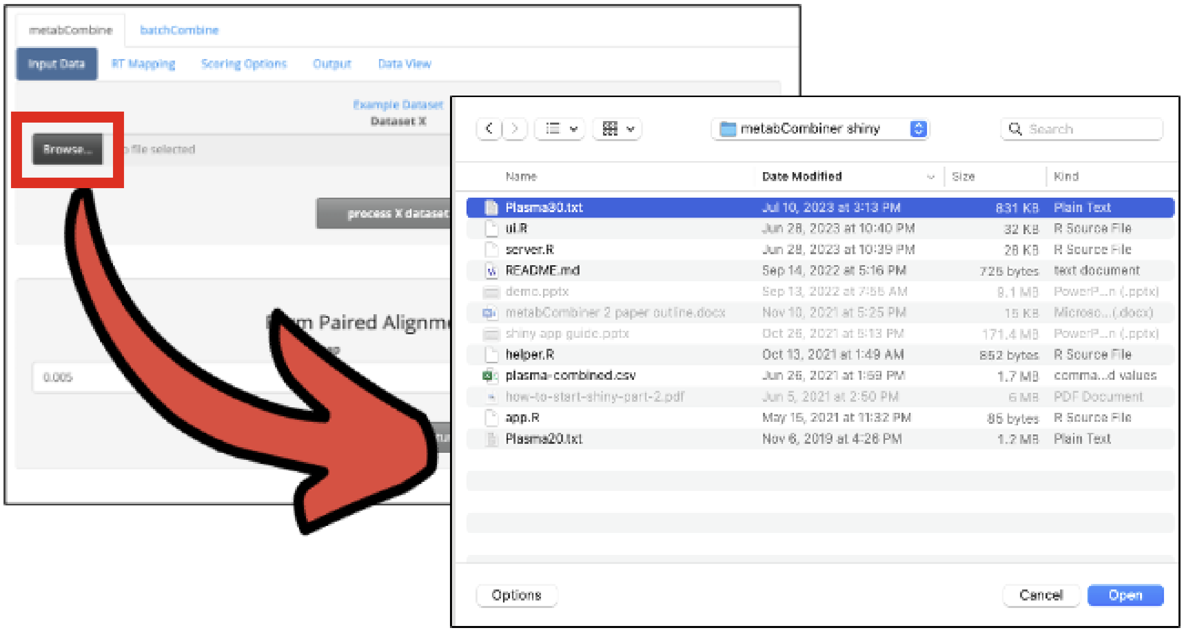
metabCombine

2.1 Input Data Tab

The Input Data tab consists of two data upload panels for the "X" and "Y" Datasets, and one panel for generating the initial combined dataset object consisting of all hypothetical paired alignments. When executed successfully, an object summary will be displayed below or alongside each of the panels.

2.1.1 Data Upload

The first step in *metabCombiner* is to upload the data to be aligned (**Figure 2.1**). Click the button captioned "Browse" in the Input Data tab on the left (Dataset X) or right (Dataset Y) side. Use the navigation menu to select one of the data files, which must have a .txt, .csv, or .rds extension. While either of the two data sets to be aligned can be chosen as "Dataset X" or "Dataset Y", the data set with more chromatographic separation/ total time is usually chosen for "X". Once the data file is selected, click "Open" to upload the data set.



Name	Date Modified	Size	Kind
Plasma30.txt	Jul 10, 2023 at 3:13 PM	831 KB	Plain Text
ui.R	Jun 28, 2023 at 10:40 PM	32 KB	R Source File
server.R	Jun 28, 2023 at 10:39 PM	28 KB	R Source File
README.md	Sep 14, 2022 at 5:16 PM	725 bytes	text document
demo.pptx	Sep 13, 2022 at 7:55 AM	9.1 MB	PowerP...n (.pptx)
metabCombiner 2 paper outline.docx	Nov 10, 2021 at 5:25 PM	15 KB	Microso...(.docx)
shiny app guide.pptx	Oct 26, 2021 at 5:13 PM	171.4 MB	PowerP...n (.pptx)
helper.R	Oct 13, 2021 at 1:49 AM	852 bytes	R Source File
plasma-combined.csv	Jun 26, 2021 at 1:59 PM	1.7 MB	comma...d values
how-to-start-shiny-part-2.pdf	Jun 5, 2021 at 2:50 PM	6 MB	PDF Document
app.R	May 15, 2021 at 11:32 PM	85 bytes	R Source File
Plasma20.txt	Nov 6, 2019 at 4:28 PM	1.2 MB	Plain Text

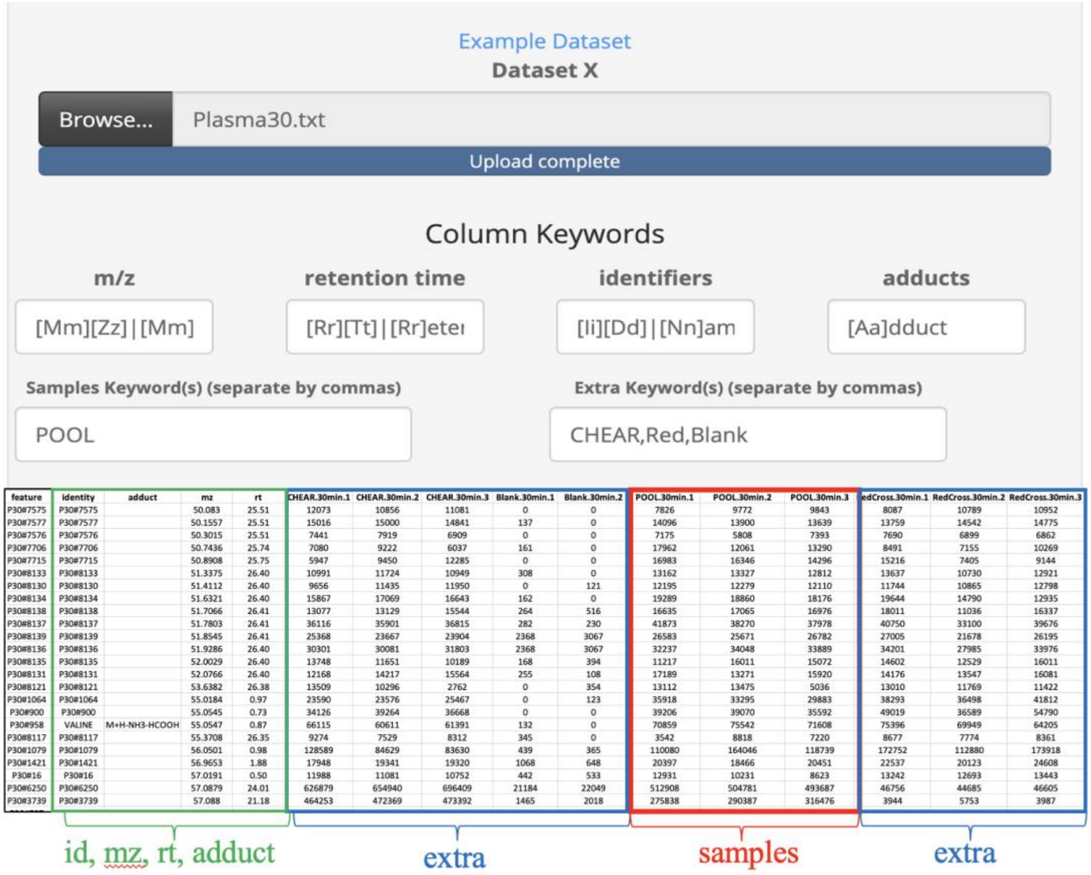
Figure 2.1: File Input Menu in Input Data Tab

Upon upload of a data file, the panel expands with several input options appearing. This panel can be divided into two parts: Column Keyword Mapping and Feature Filters.

2.1.2 Column Keyword Mapping

In the keyword mapping section, the data set components described in Chapter 1.4 are mapped to columns in the uploaded data table. The first four fields (m/z, RT, id, adduct) are assigned to the **first column** whose name contains the indicated keyword(s), whereas the latter two fields (Samples, Extra) are mapped to **all unassigned columns** containing the keyword(s). Keywords can be the full name of the desired column, part of the name, or regular expressions. The default values for m/z, retention time, identifiers, and adducts fields cover a variety of potential column names. For example, the default value for the m/z field "[Mm][Zz]|[Mm]/[Zz]|[Mm].[Zz]|[Mm]ass" matches to the first column whose name contains any of the {"mz", "m/z", "m.z", "mass"}, both in upper and lower-case letters.

The "samples" and "extra" field entries are blank by default and should be customized to accurately be mapped to the desired columns. If the 'samples' keyword field is left blank, the program will set the "samples" component to the longest consecutive stretch of numeric columns. Here, we will use "POOL" as the keyword for samples, which will map to all columns of the input dataset containing the phrase "POOL." The 'extra' keyword will be set to "Red,CHEAR,Blank" (NOTE: no spaces in between!), which map 'extra' to any column containing "Red", "CHEAR", or "Blank" in the header. These columns will be in the final report, but will not affect the alignment results.



Dataset X

Plasma30.txt

Column Keywords

m/z: [Mm][Zz]|[Mm]

retention time: [Rr][Tt]|[Rr]eter

identifiers: [Ii][Dd]|[Nn]am

adducts: [Aa]dduct

Samples Keyword(s) (separate by commas): POOL

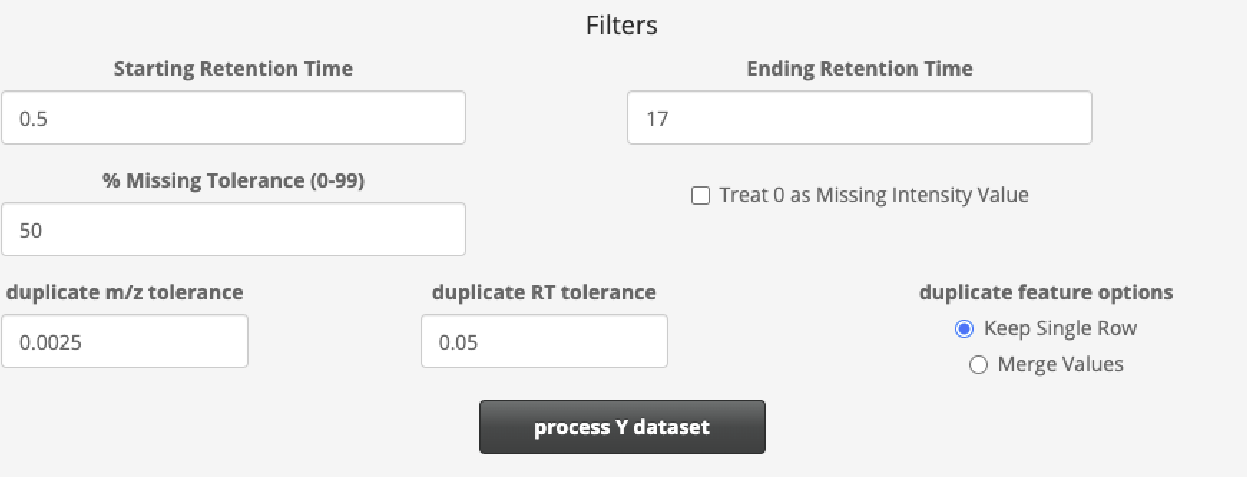
Extra Keyword(s) (separate by commas): CHEAR,Red,Blank

feature	identity	adduct	mz	rt	CHEAR.30min.1	CHEAR.30min.2	CHEAR.30min.3	Blank.30min.1	Blank.30min.2	POOL.30min.1	POOL.30min.2	POOL.30min.3	edCross.30min.1	RedCross.30min.2	RedCross.30min.3
P30#7575	P30#7575		50.083	25.51	12073	10856	11081	0	0	7826	9772	9843	8087	10789	10952
P30#7577	P30#7577		50.1557	25.51	15016	15000	14841	137	0	14096	13900	13639	13759	14542	14775
P30#7576	P30#7576		50.3015	25.51	7441	7919	6909	0	0	7175	5808	7393	7690	6899	6862
P30#7706	P30#7706		50.7436	25.74	7080	9222	6037	161	0	17962	12061	13290	8491	7155	10269
P30#7715	P30#7715		50.8908	25.75	5947	9450	12285	0	0	16983	16346	14296	15216	7405	9144
P30#8133	P30#8133		51.3375	26.40	10991	11724	10949	308	0	13162	13327	12812	13637	10730	12921
P30#8130	P30#8130		51.4112	26.40	9656	11435	11950	0	121	12195	12279	12110	11744	10865	12798
P30#8134	P30#8134		51.6321	26.40	15867	17069	16643	162	0	19289	18860	18176	19644	14790	12935
P30#8138	P30#8138		51.7066	26.41	13077	13129	15544	264	516	16635	17065	16976	18011	11036	16337
P30#8137	P30#8137		51.7803	26.41	36116	35901	36815	282	230	41873	38270	37978	40750	33100	39676
P30#8139	P30#8139		51.8545	26.41	25368	23667	23904	2368	3067	26583	25671	26782	27005	21678	26195
P30#8136	P30#8136		51.9286	26.40	30301	30081	31803	2368	3067	32237	34048	33889	34201	27985	33976
P30#8135	P30#8135		52.0029	26.40	13748	11651	10189	168	394	11217	16011	15072	14602	12529	16011
P30#8131	P30#8131		52.0766	26.40	12168	14217	15564	255	108	17189	13271	15920	14176	13547	16081
P30#8121	P30#8121		53.6382	26.38	13509	10296	2762	0	354	13112	13475	5036	13010	11769	11422
P30#1064	P30#1064		55.0184	0.97	23590	23576	25467	0	123	35918	33295	29883	38293	36498	41812
P30#900	P30#900		55.0545	0.73	34126	39264	36668	0	0	39206	39070	35592	49019	36589	54790
P30#958	VALINE	M+H-NH3-HCOOH	55.0547	0.87	66115	60611	61391	132	0	70859	75542	71608	75396	69949	64205
P30#8117	P30#8117		55.3708	26.35	9274	7529	8312	345	0	3542	8818	7220	8677	7774	8361
P30#1079	P30#1079		56.0501	0.98	128589	84629	83630	439	365	110080	164046	118739	172752	112880	173918
P30#1421	P30#1421		56.9653	1.88	17948	19341	19320	1068	648	20397	18466	20451	22537	20123	24608
P30#16	P30#16		57.0191	0.50	11988	11081	10752	442	533	12931	10231	8623	13242	12693	13443
P30#6250	P30#6250		57.0879	24.01	626879	654940	696409	21184	22049	512908	504781	493687	46756	44685	46605
P30#3739	P30#3739		57.088	21.18	464253	472369	473392	1465	2018	275838	290387	316476	3944	5753	3987

Figure 2.2: keyword input and the associated mapped columns in the example input data set (plasma30.txt)

2.1.3 Feature Filters

After assembling the data set components, the program performs three separate filters on the feature lists. The parameters for each of these filters are customizable in the expandable panel below the Keyword Mapping entries (**Figure 2.3**).



Filters

Starting Retention Time: 0.5

Ending Retention Time: 17

% Missing Tolerance (0-99): 50

□ Treat 0 as Missing Intensity Value

duplicate m/z tolerance: 0.0025

duplicate RT tolerance: 0.05

duplicate feature options

● Keep Single Row

○ Merge Values

Figure 2.3: Feature filter parameters for the example input Y data set (plasma20.txt)

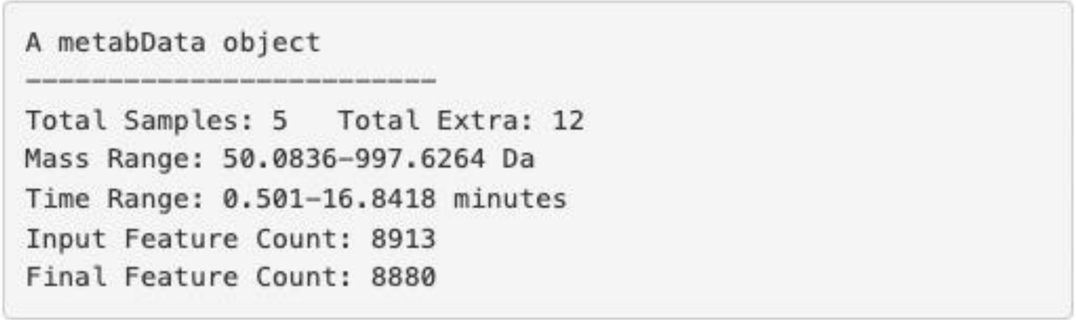
First, the RT range filter constrains the feature list to a minimum and maximum RT. This filter is useful for excising chromatographic regions containing purely non-biological features (e.g. background ions) or significant non-overlaps, thus improving the RT mapping in the later stages. The **Starting Retention Time** should be slightly *before* the first true metabolite's (~0.5 min for both experiments) and the **Ending Retention Time** should be slightly *after* the RT of the last true metabolite or otherwise before the re-calibration to initial LC conditions (30min/ max for plasma30 data set and 17 min for plasma20).

Second, the % missingness filter limits the feature list to compounds that are not present in over the indicated percentage of samples. By default, this parameter is set to 50% and must be from 0 to 100. Optionally, abundance values of zero may be treated as missing for the purposes of calculating missingness and central abundances. Note that columns in the 'extra' field do not factor into calculations of missingness proportions.

Finally, the duplicate feature filter eliminates redundant features with roughly the same m/z and retention time. The **duplicate m/z tolerance** and **duplicate RT tolerance** parameters control how closely features may be in these two dimensions to be considered duplicates. Once two or more features are considered duplicates, there is an option to keep the row with the lowest % missingness or higher median abundance (Keep Single Row) or Merge Values based on highest per-sample abundances.

2.1.4 Data Processing

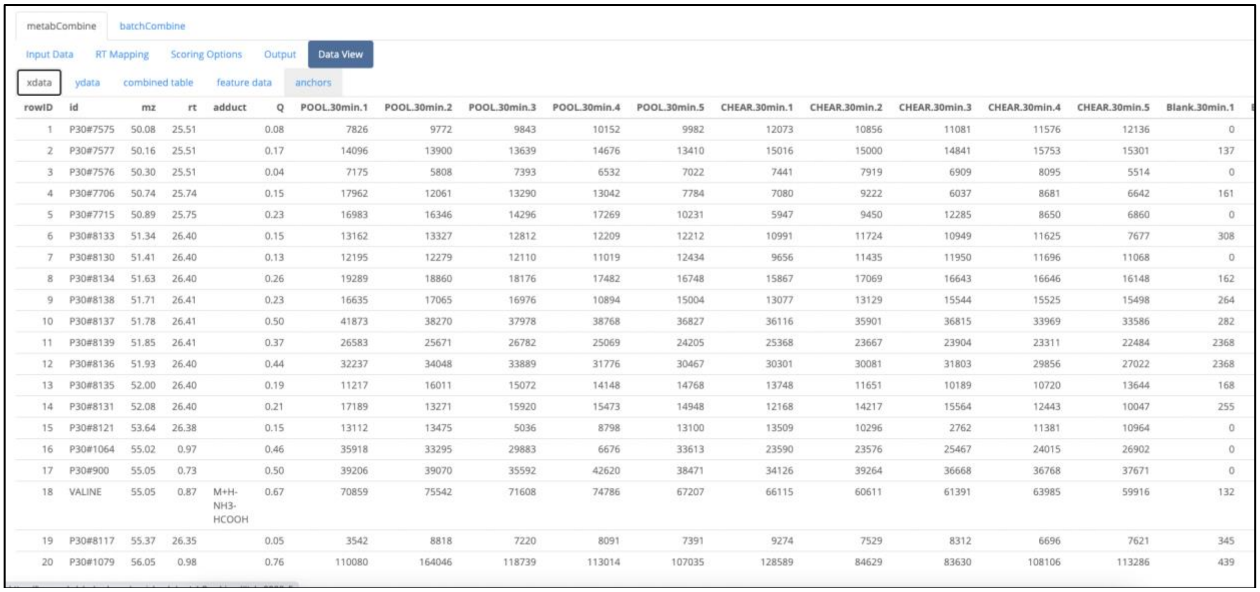
After customizing the necessary parameters in each Input Data panel, click on the corresponding **process X dataset** or **process Y dataset** button. If all was done correctly, a text box will appear providing a summary of the pre-processed data set, exemplified by **Figure 2.4**. listing the number of samples and extra columns, the m/z and retention time ranges, and the initial and final feature counts.



```
A metabData object
------------------------
Total Samples: 5   Total Extra: 12
Mass Range: 50.0836-997.6264 Da
Time Range: 0.501-16.8418 minutes
Input Feature Count: 8913
Final Feature Count: 8880
```

Figure 2.4: Single data set summaries for example input table (plasma20.txt)

The Data View tab is updated to display a limited view of the uploaded and processed data sets, as illustrated in **Figure 2.5**. In addition to the columns derived from the input data set, there is a new column **(Q)**, a relative quantitation measure calculated from ranking the median abundances from 0 (least abundant) to 1 (most abundant).



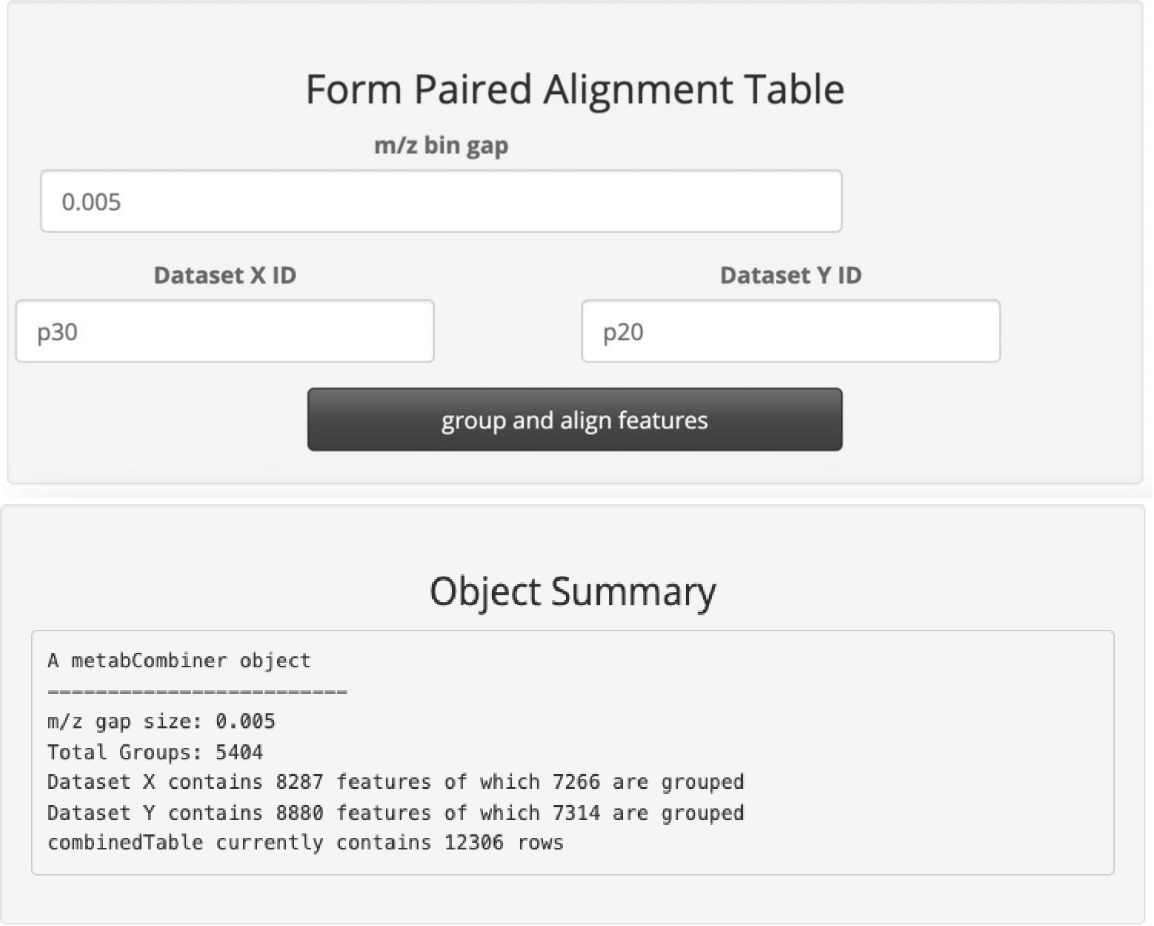
rowID	id	mz	rt	adduct	Q	POOL.30min.1	POOL.30min.2	POOL.30min.3	POOL.30min.4	POOL.30min.5	CHEAR.30min.1	CHEAR.30min.2	CHEAR.30min.3	CHEAR.30min.4	CHEAR.30min.5	Blank.30min.1
1	P30#7575	50.08	25.51		0.08	7826	9772	9843	10152	9982	12073	10856	11081	11576	12136	0
2	P30#7577	50.16	25.51		0.17	14096	13900	13639	14676	13410	15016	15000	14841	15753	15301	137
3	P30#7576	50.30	25.51		0.04	7175	5808	7393	6532	7022	7441	7919	6909	8095	5514	0
4	P30#7706	50.74	25.74		0.15	17962	12061	13290	13042	7784	7080	9222	6037	8681	6642	161
5	P30#7715	50.89	25.75		0.23	16983	16346	14296	17269	10231	5947	9450	12285	8650	6860	0
6	P30#8133	51.34	26.40		0.15	13162	13327	12812	12209	12212	10991	11724	10949	11625	7677	308
7	P30#8130	51.41	26.40		0.13	12195	12279	12110	11019	12434	9656	11435	11950	11696	11068	0
8	P30#8134	51.63	26.40		0.26	19289	18860	18176	17482	16748	15867	17069	16643	16646	16148	162
9	P30#8138	51.71	26.41		0.23	16635	17065	16976	10894	15004	13077	13129	15544	15525	15498	264
10	P30#8137	51.78	26.41		0.50	41873	38270	37978	38768	36827	36116	35901	36815	33969	33586	282
11	P30#8139	51.85	26.41		0.37	26583	25671	26782	25069	24205	25368	23667	23904	23311	22484	2368
12	P30#8136	51.93	26.40		0.44	32237	34048	33889	31776	30467	30301	30081	31803	29856	27022	2368
13	P30#8135	52.00	26.40		0.19	11217	16011	15072	14148	14768	13748	11651	10189	10720	13644	168
14	P30#8131	52.08	26.40		0.21	17189	13271	15920	15473	14948	12168	14217	15564	12443	10047	255
15	P30#8121	53.64	26.38		0.15	13112	13475	5036	8798	13100	13509	10296	2762	11381	10964	0
16	P30#1064	55.02	0.97		0.46	35918	33295	29883	6676	33613	23590	23576	25467	24015	26902	0
17	P30#900	55.05	0.73		0.50	39206	39070	35592	42620	38471	34126	39264	36668	36768	37671	0
18	VALINE	55.05	0.87	M+H-NH3-HCOOH	0.67	70859	75542	71608	74786	67207	66115	60611	61391	63985	59916	132
19	P30#8117	55.37	26.35		0.05	3542	8818	7220	8091	7391	9274	7529	8312	6696	7621	345
20	P30#1079	56.05	0.98		0.76	110080	164046	118739	113014	107035	128589	84629	83630	108106	113286	439

Figure 2.5: processed X Dataset (xdata) in Data View tab

2.1.5 Form Aligned Table

After uploading and processing the X and Y data sets, it is time to form the initial aligned feature table. The "Form Paired Alignment Table" panel shown in **Figure 2.6** has three parameter inputs. The first, **m/z bin gap**, determines the distance in m/z space for a complementary pair of features to be considered a possible matching metabolite ion species. Features are binned into separate groups based on this parameter value, with the best values between 0.005 and 0.01 Da. The latter entries, **Dataset X ID** and **Dataset Y ID**, are customizable identifiers for new single data sets. Data set identifiers must be unique and cannot be "x" or "y". Dataset IDs are more important for aligning more than two data sets (e.g. using a combined table as input; see Section 3.1).

Here, we leave the bin gap at the default value (0.005) and assign the data sets as "p30" and "p20." Clicking on the **"group and align features"** button generates the combined feature table as well as a text box summarizing the *metabCombiner* object combining the data sets.



Form Paired Alignment Table

m/z bin gap: 0.005

Dataset X ID: p30

Dataset Y ID: p20

Object Summary

```
A metabCombiner object
-------------------------
m/z gap size: 0.005
Total Groups: 5404
Dataset X contains 8287 features of which 7266 are grouped
Dataset Y contains 8880 features of which 7314 are grouped
combinedTable currently contains 12306 rows
```

Figure 2.6: Panels for Combined Data class construction and initial data size

2.2 RT Mapping Tab

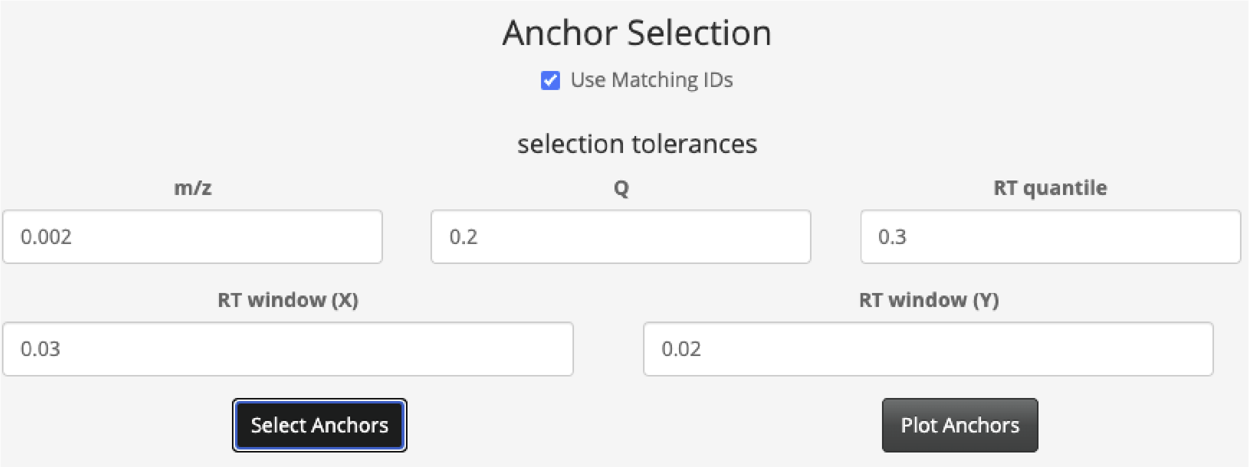
After completing all tasks in the Input Data tab, switch to the RT Mapping tab. Here, RTs will be mapped from data set X to data set Y for subsequent pairwise comparisons. There are two parts to the RT Mapping process: Anchor Selection and Model-Fitting. Both steps generate and update a plot of the mapping, with a separate sub-panel and buttons for customizing the plot.

2.2.1 Anchor Selection

Anchor selection is the process by which the program chooses ordered pair features for mapping between retention times. Since most (if not all) features are unidentified in LC-MS metabolomics data sets, this process instead relies on selecting highly abundant compounds with small inter-dataset differences in m/z, Q, and RT quantile (the latter defined as the RT as a linear proportion of total chromatography time).

Threshold differences are determined through the **selection tolerances** parameters. Alongside these parameters are **RT windows** around which all points are excluded for a given anchor point. Generally, smaller RT windows leads to more selected anchors. It may be useful to experiment with different parameter values and observe how it affects the plotted points after clicking the **Select Anchors** button. Recommended values follow the 1-5 rule (i.e., 0.001 - 0.005 Da for m/z tolerance, 0.1 - 0.5 for Q tolerance, 0.1 - 0.5 for RT quantile tolerance, 0.01 - 0.05 for RT windows).

In addition, the **Use Matching IDs** option enables selection of feature pairs with string-matched identifiers. Identity-matched pairs may exceed any of the tolerances and window parameters listed in the panel and provides a slight performance advantage, provided the features are correctly annotated in both data sets.



Anchor Selection

☑ Use Matching IDs

selection tolerances

m/z: 0.002

Q: 0.2

RT quantile: 0.3

RT window (X): 0.03

RT window (Y): 0.02

Figure 2.7: Anchor Selection panel

Clicking on the **Select Anchors** button in the generates a plot of ordered pair retention times, such as the one illustrated in **Figure 2.8** below. Ideally these point should provide a path for a curve fit, to be generated in the model fitting step. In addition, the anchors generated in this step are viewable in the Data View panel, as depicted in **Figure 2.9**.

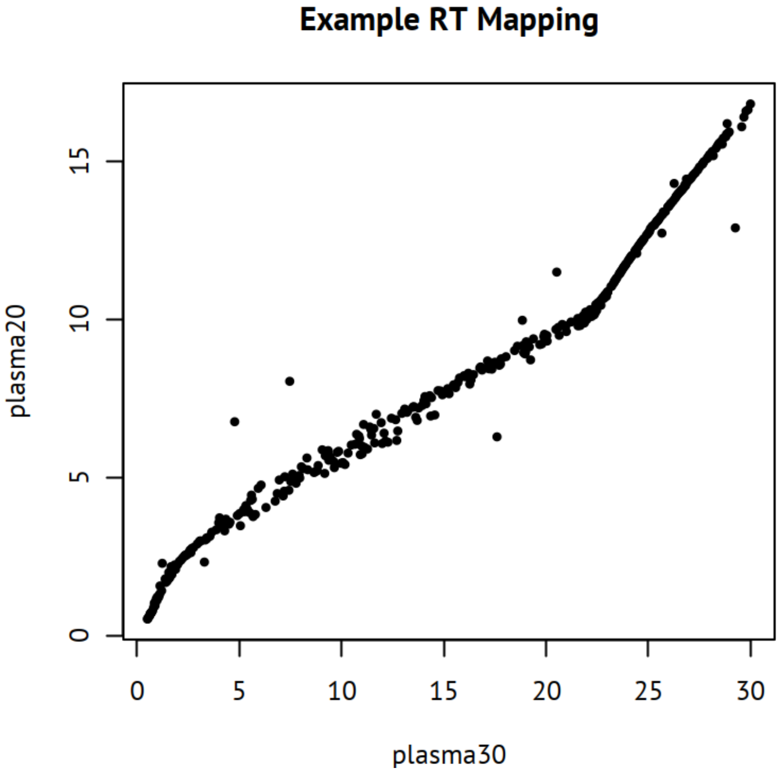
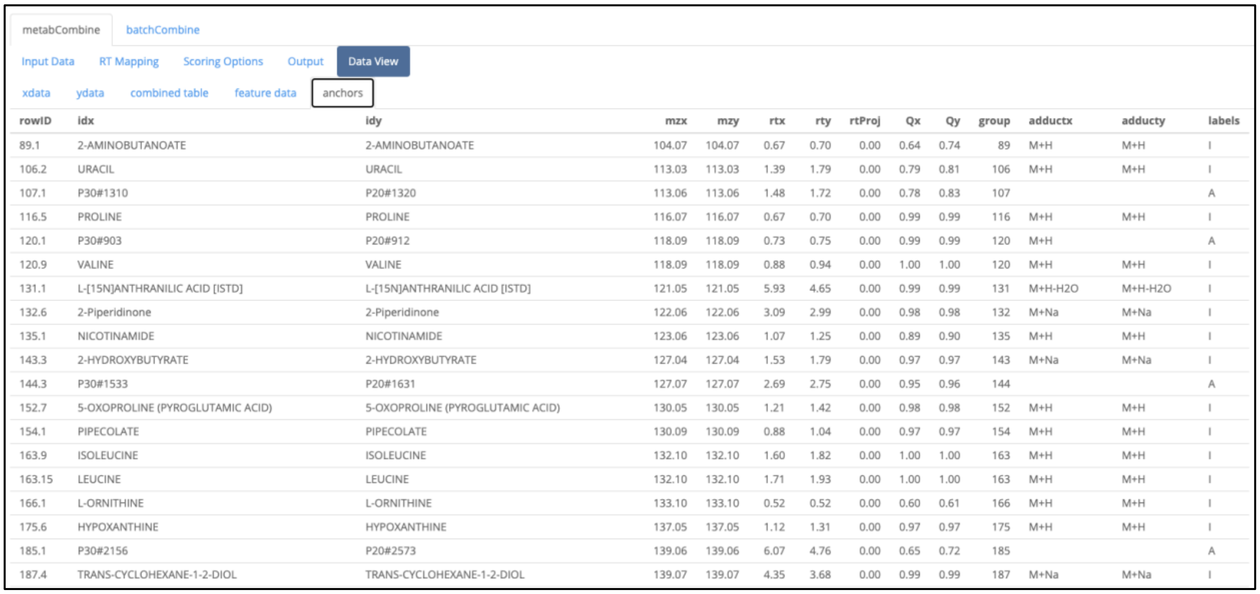


Figure 2.8: Plotted RT Times (in min) of Selected Anchors



rowID	idx	idy	mzx	mzy	rtx	rty	rtProj	Qx	Qy	group	adductx	adducty	labels
89.1	2-AMINOBUTANOATE	2-AMINOBUTANOATE	104.07	104.07	0.67	0.70	0.00	0.64	0.74	89	M+H	M+H	I
106.2	URACIL	URACIL	113.03	113.03	1.39	1.79	0.00	0.79	0.81	106	M+H	M+H	I
107.1	P30#1310	P20#1320	113.06	113.06	1.48	1.72	0.00	0.78	0.83	107			A
116.5	PROLINE	PROLINE	116.07	116.07	0.67	0.70	0.00	0.99	0.99	116	M+H	M+H	I
120.1	P30#903	P20#912	118.09	118.09	0.73	0.75	0.00	0.99	0.99	120	M+H		A
120.9	VALINE	VALINE	118.09	118.09	0.88	0.94	0.00	1.00	1.00	120	M+H	M+H	I
131.1	L-[15N]ANTHRANILIC ACID [ISTD]	L-[15N]ANTHRANILIC ACID [ISTD]	121.05	121.05	5.93	4.65	0.00	0.99	0.99	131	M+H-H2O	M+H-H2O	I
132.6	2-Piperidinone	2-Piperidinone	122.06	122.06	3.09	2.99	0.00	0.98	0.98	132	M+Na	M+Na	I
135.1	NICOTINAMIDE	NICOTINAMIDE	123.06	123.06	1.07	1.25	0.00	0.89	0.90	135	M+H	M+H	I
143.3	2-HYDROXYBUTYRATE	2-HYDROXYBUTYRATE	127.04	127.04	1.53	1.79	0.00	0.97	0.97	143	M+Na	M+Na	I
144.3	P30#1533	P20#1631	127.07	127.07	2.69	2.75	0.00	0.95	0.96	144			A
152.7	5-OXOPROLINE (PYROGLUTAMIC ACID)	5-OXOPROLINE (PYROGLUTAMIC ACID)	130.05	130.05	1.21	1.42	0.00	0.98	0.98	152	M+H	M+H	I
154.1	PIPECOLATE	PIPECOLATE	130.09	130.09	0.88	1.04	0.00	0.97	0.97	154	M+H	M+H	I
163.9	ISOLEUCINE	ISOLEUCINE	132.10	132.10	1.60	1.82	0.00	1.00	1.00	163	M+H	M+H	I
163.15	LEUCINE	LEUCINE	132.10	132.10	1.71	1.93	0.00	1.00	1.00	163	M+H	M+H	I
166.1	L-ORNITHINE	L-ORNITHINE	133.10	133.10	0.52	0.52	0.00	0.60	0.61	166	M+H	M+H	I
175.6	HYPOXANTHINE	HYPOXANTHINE	137.05	137.05	1.12	1.31	0.00	0.97	0.97	175	M+H	M+H	I
185.1	P30#2156	P20#2573	139.06	139.06	6.07	4.76	0.00	0.65	0.72	185			A
187.4	TRANS-CYCLOHEXANE-1-2-DIOL	TRANS-CYCLOHEXANE-1-2-DIOL	139.07	139.07	4.35	3.68	0.00	0.99	0.99	187	M+Na	M+Na	I

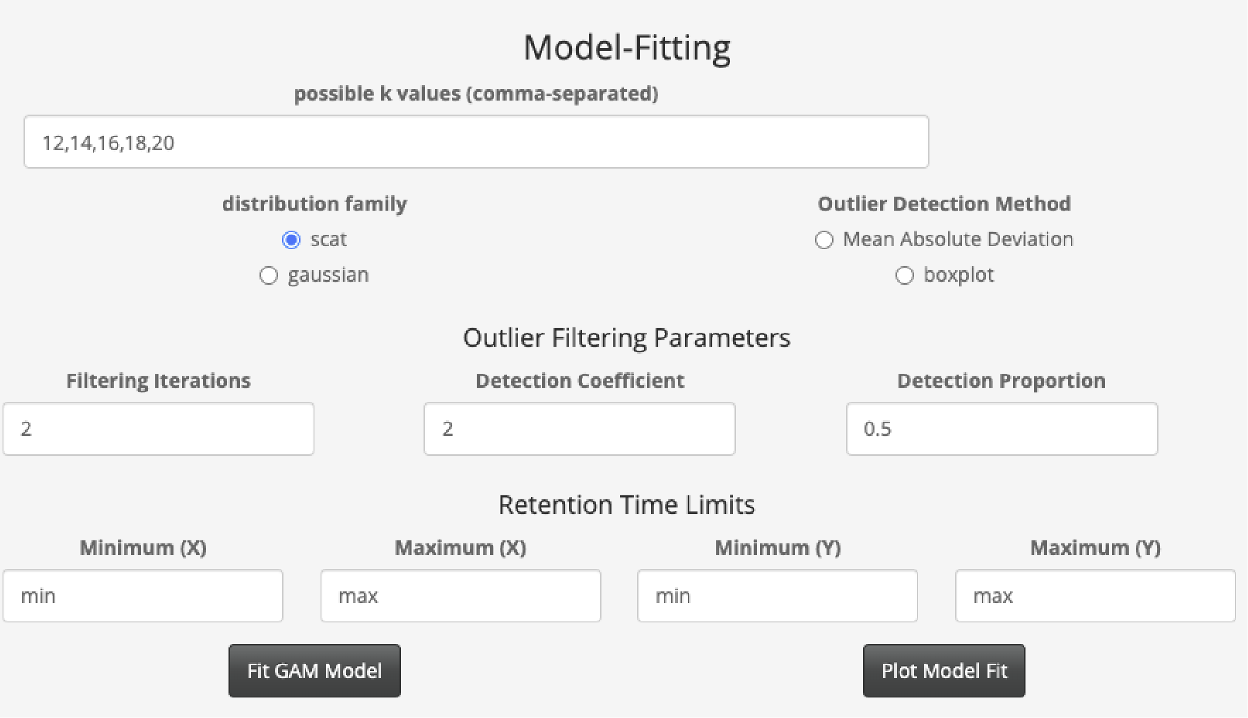
Figure 2.9: Selected Anchors in Data View tab

2.2.2 Model Fitting

After determining the ordered pair RTs in the anchor selection step, the model fitting step generates a mapping between retention times, using a generalized additive model (GAM). The complexity of this GAM is determined by a hyperparameter, k. Users may supply one or multiple **possible k values**, with the most tested values falling between 10 and 20. A single value is chosen by 10 fold cross validation, unless only a single value is provided.

As illustrated in **Figure 2.8** above, there are inevitably numerous ordered pairs that do not fall along the intended path of the RT mapping. Multiple parameters control the process of filtering these outlying points. There are two choices for **distribution family**: Gaussian and SCAT (scaled t). Compared to Gaussian, SCAT is more robust to outliers, though computationally more intensive. There are two **Outlier Detection Method**s: mean absolute outliers are determined based on a) **detection coefficient** times *mean absolute deviation* or b) *boxplot*-based, with the **detection coefficient** determining the IQR multiplier. Outliers are detected and eliminated from the model in multiple rounds, based on the **Filtering iterations** parameter. An ordered pair must be flagged in a proportion of fits greater than the **detection proportion** to be eliminated as an outlier.

In certain cases, the retention time range of the model should be restricted if the boundaries are poorly mapped. For this, the **Retention Time Limits** parameters allows for setting minimum and maximum retention times for the X and Y data sets. Clicking the **Fit GAM Model** button generates the spline fit.



Model-Fitting

possible k values (comma-separated): 12,14,16,18,20

distribution family: ● scat ○ gaussian

Outlier Detection Method: ○ Mean Absolute Deviation ○ boxplot

Outlier Filtering Parameters

Filtering Iterations: 2

Detection Coefficient: 2

Detection Proportion: 0.5

Retention Time Limits

Minimum (X): min

Maximum (X): max

Minimum (Y): min

Maximum (Y): max

Figure 2.10: Model-Fitting panel under *metabCombine*

A text will appear below the plot to indicate the progress of the model fit. Once completed, the plot will update to show the model fit, as shown in **Figure 2.11**. Note that the **Plot Anchors** button in the *Select Anchors* panel restores the previous plot, whereas the **Plot Model Fit** re-generates the plotted model fit without the need to re-compute the model.

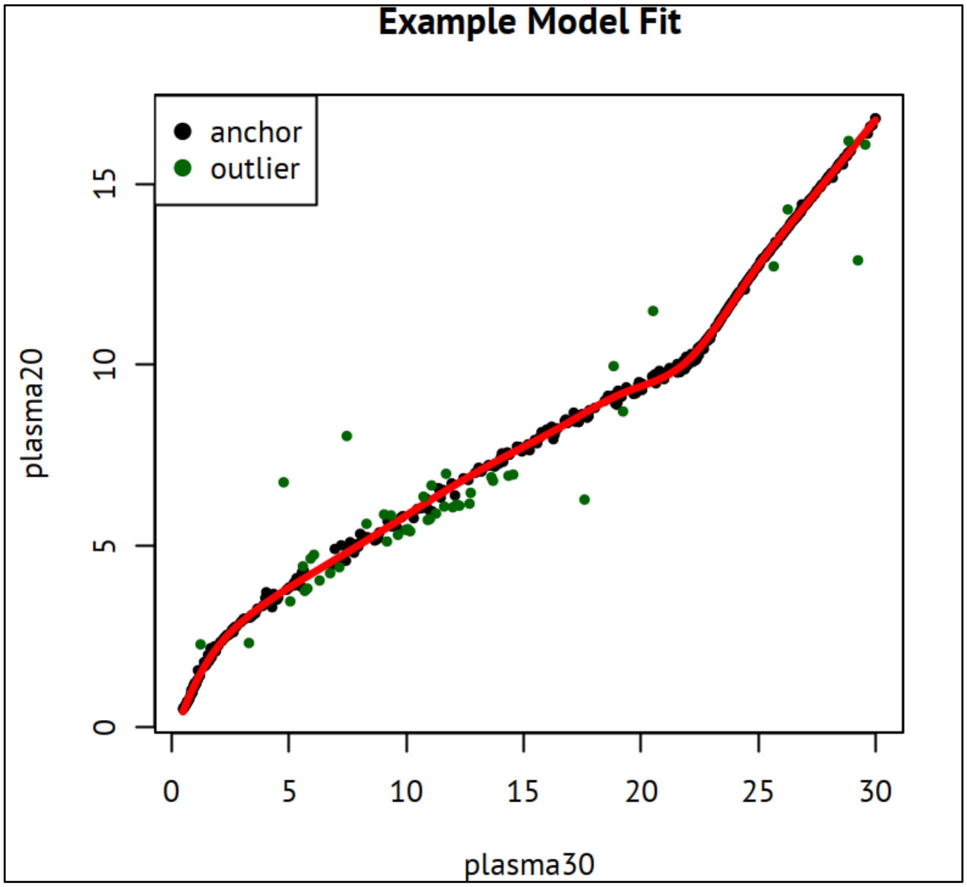
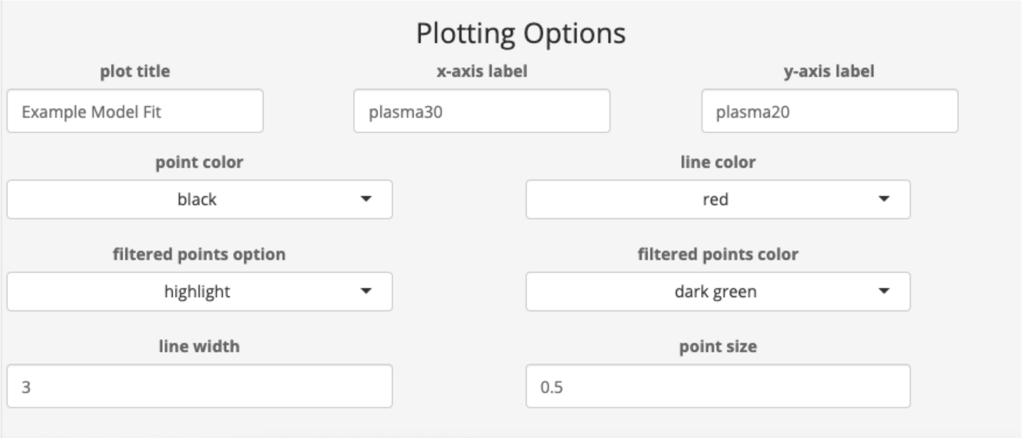


Figure 2.11: Example RT mapping model fit between plasma data sets

2.2.3 Model Fitting

Lower down on the panel of the RT Mapping tab is a section for customizing the plot, as depicted in **Figure 2.12** below. The various plot characteristics that can be modified are the plot titles, axis labels, colors, line width, size of the points, and how outlier points are handled within the plot (highlighting, showing, or removing). The highlighting option triggers a legend in the upper left corner. Changes to any of these parameters leads to instant changes to the plot.



Plotting Options

plot title: Example Model Fit

x-axis label: plasma30

y-axis label: plasma20

point color: black

line color: red

filtered points option: highlight

filtered points color: dark green

line width: 3

point size: 0.5

Figure 2.12: Plotting Options panel

2.3 Scoring Options

The next stage of metabCombine is to computationally determine the most likely feature pair alignments, using pairwise alignment scores. The *Scoring Options* tab has two parts: Feature Pair Scoring and Row Annotation & Reduction.

2.3.1 Feature Pair Scoring

metabCombiner uses three measures to score feature pairs: absolute m/z differences, mapped RTs *f(rt*$_x$*)* vs observed RTs *rt*$_y$, and relative abundance (Q) comparisons. This is calculated according to the following equation:

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

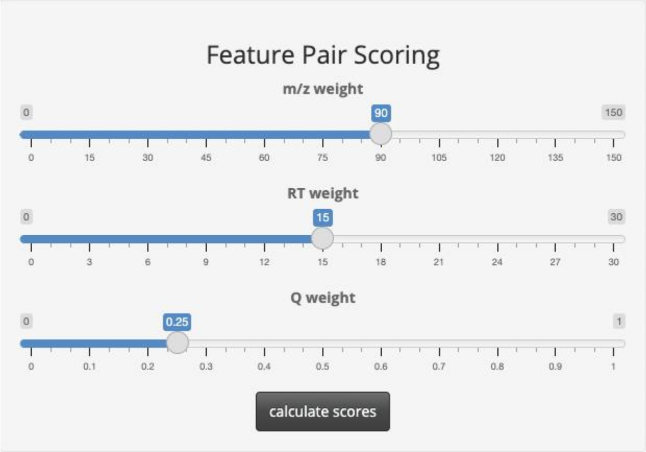
where mz_x, rt_x, & Q_x are the respective m/z, RT, and Q values of feature F_x; mz_y, rt_y, & Q_y are the respective m/z, RT, and Q values of feature F_y ; *f* denotes the computed RT mapping function, with prediction errors normalized by the range of Y dataset RT values. *A, B, C* are positive weight parameters penalizing differences in feature m/z, rt, and Q, respectively. Scores range from 0 (feature mismatch) to 1 (likely feature match). The values of the weight parameters are determined in the first panel (**Figure 2.13**). Here are a few brief guidelines for selecting the weight parameters:

- **m/z weight** (0 to 150, default: 75). The default is generally a useful value for most analyses. For high mass accuracy instrumentation, higher values (e.g. 90-120) are useful; lower values (50-70) may be necessary if high mass errors are observed

- **RT weight** (0 to 30, default: 10). In most cases, the closest feature in mapped RT space will be the highest scoring match; this parameter determines how highly the RT prediction factors into choosing the best match. Values should depend on how closely RT modeling curve maps the anchors (see plot in RT Mapping). Only use values greater than 20 if the LC methods are roughly identical. If the anchors are closely mapped, such as in **Figure 2.11**, the RT weight should be higher than 10 (e.g. 13-17). For spline curves with sparse or poorly mapped chromatographic regions, use lower values (e.g. 5-8).

- **Q weight** (0 to 1, default: 0.25). This is generally the least influential factor, but it may prove decisive for comparisons between high and low abundance species. The more similar the specimens, the higher this value should be (typically up to 0.5-0.7), though ionization factors should also be considered. Use lower values (e.g. 0.1) for data representing dissimilar or multiple specimens.

Figure 2.14 shows an example of a scored feature group with calculated ranks (from the original Habra et al. (2021) paper).



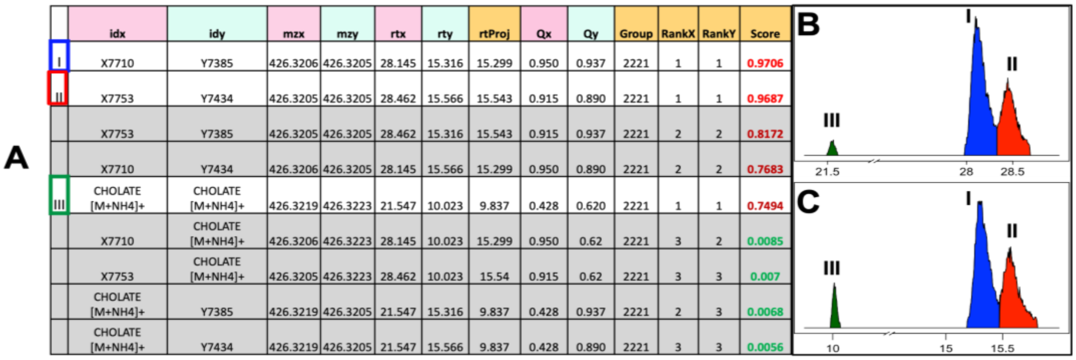
Feature Pair Scoring

m/z weight: 90 (0 to 150)

RT weight: 15 (0 to 30)

Q weight: 0.25 (0 to 1)

Figure 2.13: Feature Pair Scoring weight panel



	idx	idy	mzx	mzy	rtx	rty	rtProj	Qx	Qy	Group	RankX	RankY	Score
I	X7710	Y7385	426.3206	426.3205	28.145	15.316	15.299	0.950	0.937	2221	1	1	0.9706
II	X7753	Y7434	426.3205	426.3205	28.462	15.566	15.543	0.915	0.890	2221	1	1	0.9687
	X7753	Y7385	426.3205	426.3205	28.462	15.316	15.543	0.915	0.937	2221	2	2	0.8172
	X7710	Y7434	426.3206	426.3205	28.145	15.566	15.299	0.950	0.890	2221	2	2	0.7683
III	CHOLATE [M+NH4]+	CHOLATE [M+NH4]+	426.3219	426.3223	21.547	10.023	9.837	0.428	0.620	2221	1	1	0.7494
	X7710	CHOLATE [M+NH4]+	426.3206	426.3223	28.145	10.023	15.299	0.950	0.62	2221	3	2	0.0085
	X7753	CHOLATE [M+NH4]+	426.3205	426.3223	28.462	10.023	15.54	0.915	0.62	2221	3	3	0.007
	CHOLATE [M+NH4]+	Y7385	426.3219	426.3205	21.547	15.316	9.837	0.428	0.937	2221	2	3	0.0068
	CHOLATE [M+NH4]+	Y7434	426.3219	426.3205	21.547	15.566	9.837	0.428	0.890	2221	3	3	0.0056

Figure 2.14 Example scored feature group output with matching peaks. (A) Columns with pink and turquoise headings are feature metadata provided by the input datasets. Orange headings- rtProj (model-predicted RT), Group number, Score and Feature Ranks with respect to alternative matches- are generated by the program. Rows in grey contain mismatched features and should be discarded. Rows 1 & 2 are unidentified isomers; row 5 was previously identified in both datasets as Cholate [M+NH4]+. (B) and (C) are the plotted chromatographic peaks corresponding to these features in datasets X & Y, respectively.

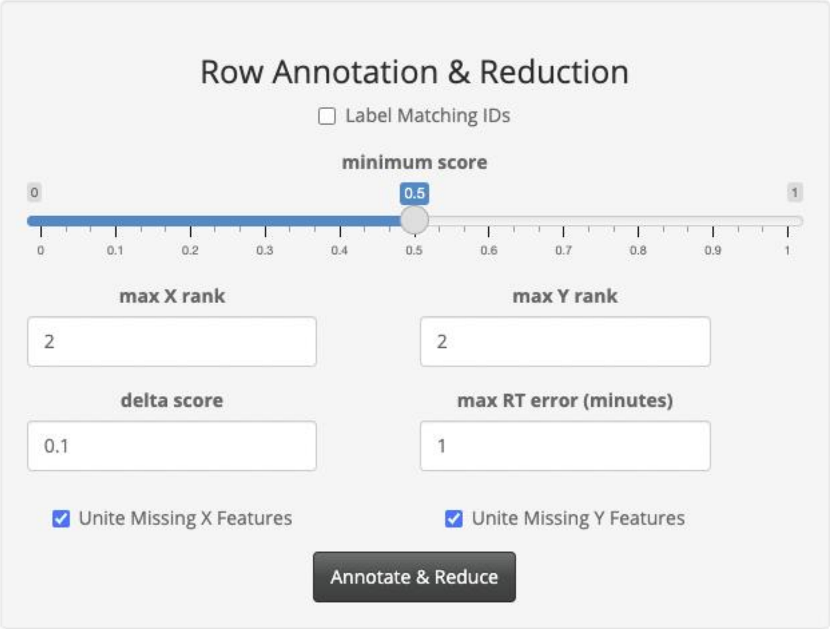
2.3.2 Row Annotation & Reduction

The final step of the analysis is to determine which feature pairs in the combined table report correspond to true feature alignments between identical analytes and which are less plausible. In the background, *metabCombiner* generates two separate reports: 1) a full table with column listing labels of removable ("REMOVE") and conflicting ("CONFLICT") feature alignment rows; 2) a final table with all removable rows are eliminated, and all features reduced to at most one complementary match.

The Row Annotation & Reduction panel is illustrated in **Figure 2.15**. Feature pairs with similarity scores below the **minimum score** or higher pairwise ranks above the indicated thresholds (**max X rank** and **max Y rank**) are treated as removable, along with any RT prediction errors above the indicated **max RT error**. **Delta score** is the difference between the top-ranking hypothetical alignment (rankX = 1, rankY = 1) and the next highest pair for a shared compound; score differences smaller than the indicated parameter results in both alignments flagged as "CONFLICT" in the full table, and "REMOVE" otherwise. For the final table, the best combination of rows is labeled as "RESOLVED."

The **Unite Missing X Features** and **Unite Missing Y Features** options creates new rows in the final table for all non-intersected X and Y data set features. This is recommended if planning to align the results to further data sets. The **Label Matching IDs** option annotates string-matched ID rows as "IDENTITY."

Click the **"Annotate & Reduce"** button to run this module.



Row Annotation & Reduction

☐ Label Matching IDs

minimum score: 0.5

max X rank: 2

max Y rank: 2

delta score: 0.1

max RT error (minutes): 1

☑ Unite Missing X Features

☑ Unite Missing Y Features

Figure 2.15: Row Annotation & Reduction panel

Below are examples of this row annotation and reduction process in **Tables 2.1**. In (A), one Y feature (m/z = 265.1182, RT = 5.979) is matched to two possible X features (m/z = 265.1184, RT = 4.781) and (m/z = 265.1161, RT = 4.195). The delta score is very high (0.283), leading to this feature being labeled "REMOVE". In **(B)** and **(C)** show possible matches between two isomers from the two input tables. Due to the closeness in scores, all four are designated as a common subgroup of conflicting alignments **(B)**; subsequently, they are resolved (and labeled as such) by choosing the pair of rows with 1-1 alignments that respect elution order as well as attained the highest sum of scores (resolveScore) **(C)**.

A.

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

B.

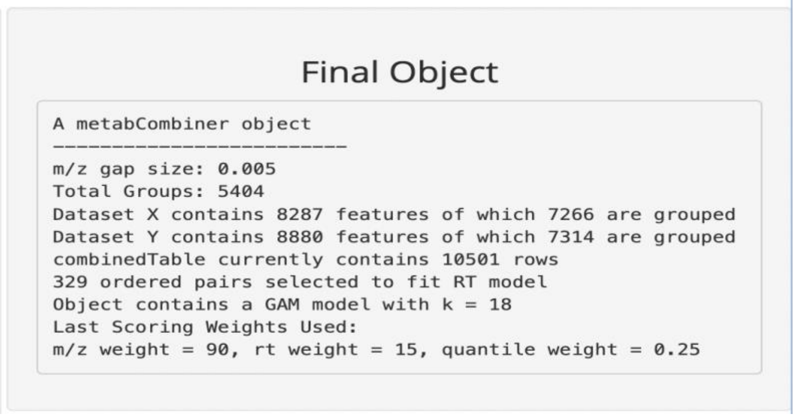
idx	idy	mzx	mzy	rtx	rty	rtProj	Qx	Qy	score	rankX	rankY	label
X	YY	246.1706	246.1707	5.263	3.985	4.014	0.826	0.926	0.936	1	1	CONFLICT
XX	YY	246.1707	246.1707	5.395	3.985	4.089	0.898	0.926	0.903	1	2	CONFLICT
X	Y	246.1706	246.1706	5.263	3.911	4.014	0.826	0.884	0.895	2	1	CONFLICT
XX	Y	246.1707	246.1706	5.395	3.911	4.089	0.898	0.884	0.84	2	2	CONFLICT

C.

idx	idy	mzx	mzy	rtx	rty	rtProj	Qx	Qy	score	rankX	rankY	label	resolveScore
X	YY	246.1706	246.1707	5.263	3.985	4.014	0.826	0.926	0.936	1	1	REMOVE	0.936
XX	YY	246.1707	246.1707	5.395	3.985	4.089	0.898	0.926	0.903	1	2	RESOLVED	1.798
X	Y	246.1706	246.1706	5.263	3.911	4.014	0.826	0.884	0.895	2	1	RESOLVED	1.798
XX	Y	246.1707	246.1706	5.395	3.911	4.089	0.898	0.884	0.84	2	2	REMOVE	0.84

Tables 2.1: Annotated feature pair alignment rows. Rows annotated as "REMOVE" are automatically eliminated in the "final" results table, leaving unannotated rows and rows with "RESOLVED" labels.

Once this process is completed, there should be a panel to the right of this tab, similar to **Figure 2.16** shown below. This is the same summary as previously generated in the Input Data panel (see: **Figure 2.6**), but with new information about the RT prediction model, the last used score weight parameters, and the updated feature row counts. This completes the *metabCombiner* alignment analysis between a pair of metabolomics data sets.



Final Object

```
A metabCombiner object
-------------------------
m/z gap size: 0.005
Total Groups: 5404
Dataset X contains 8287 features of which 7266 are grouped
Dataset Y contains 8880 features of which 7314 are grouped
combinedTable currently contains 10501 rows
329 ordered pairs selected to fit RT model
Object contains a GAM model with k = 18
Last Scoring Weights Used:
m/z weight = 90, rt weight = 15, quantile weight = 0.25
```

Figure 2.16: Final Object summary panel in Scoring Options tab

2.4 Output Tab

The Output tab of *metabCombiner Online* provides multiple ways of reviewing the alignment results and using the combined data set for further analyses. **Figure 2.17** below is a screenshot of the various panels within this tab. This section summarizes the various outputs that can be obtained from the app.

As previously discussed in section 2.3, *metabCombiner Online* generates two reports for each alignment analysis:

- **Final Results**: The finalized table of feature pair alignments in which each individual data set feature is matched to at most one feature from the other data set

- **Full Results**: The full annotated list of m/z - grouped feature pairs, useful for manual user inspection of alignment results.

For both generated reports, there are four outputs. Depending on the analysis requirements, only a subset of these needs to be downloaded. The outputs are as follows:

- **Combined Table** (.csv): Main report containing the pairwise alignment information for the X and Y data sets as well as the samples and 'extra' columns. A subset of rows can be viewed in *Data View* under *combined table*

- **Combined Table** (special format): This is a specially generated format for the main report that separates rows by feature groups, facilitating user inspection

- **Feature Data** (.csv): This lists aligned feature meta-data under the respective data set identifiers. A preview can be viewed in *Data View* under *combined table*.

- **Download Object** (.rds): This is the *metabCombiner* object file. It stores the results of the analysis and can be used as input for further alignment tasks.

Both the *Combined Table* and *Feature Data* files can be linked together through the *rowID* column that both reports share. Saving objects is a recommended step for retaining progress for longer analyses involving multiple data seta (see **Section 2.5**).

In addition to the combined data set, the formatted X and Y tables used for input can also be saved to files. Data tables (.csv) can be obtained by clicking the **xdata Table (.csv)** or **ydata Table (.csv)** buttons. Associated objects (.rds) are accessible with the **xdata Object (.rds)** and **ydata Object (.rds)** buttons. Either format for these data sets can be used as input in the **Input Data** tab.

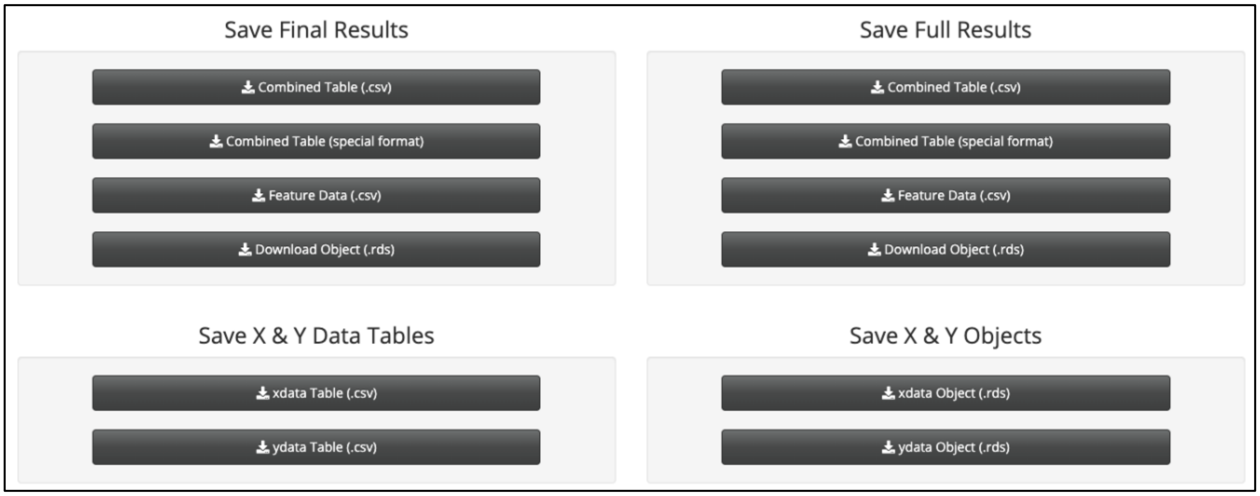


Figure 2.17 Output Data panel

2.5 Alignment of Multiple Data Sets

While *metabCombiner* strictly aligns pairs of feature tables, there is functionality for aligning additional data sets beyond two. This is achieved by aligning "combined" feature table objects (the results of a *metabCombiner* analysis) with new metabolomics data sets in stepwise merging operations. This process makes use of the organization of constituent datasets with identifiers specified by the user. Due to significant inter-experiment differences in measured retention times, *metabCombiner* currently supports an "all-to-one" alignment, i.e. aligning all data sets to a common "master" list with one set of retention times, but not an "all to all" alignment. Consequently, features missing from the chosen "master" data set but present in other tables are reported but not aligned in the current workflow. This section covers how to perform multi-stage alignment.

2.5.1 Combined Data Inputs

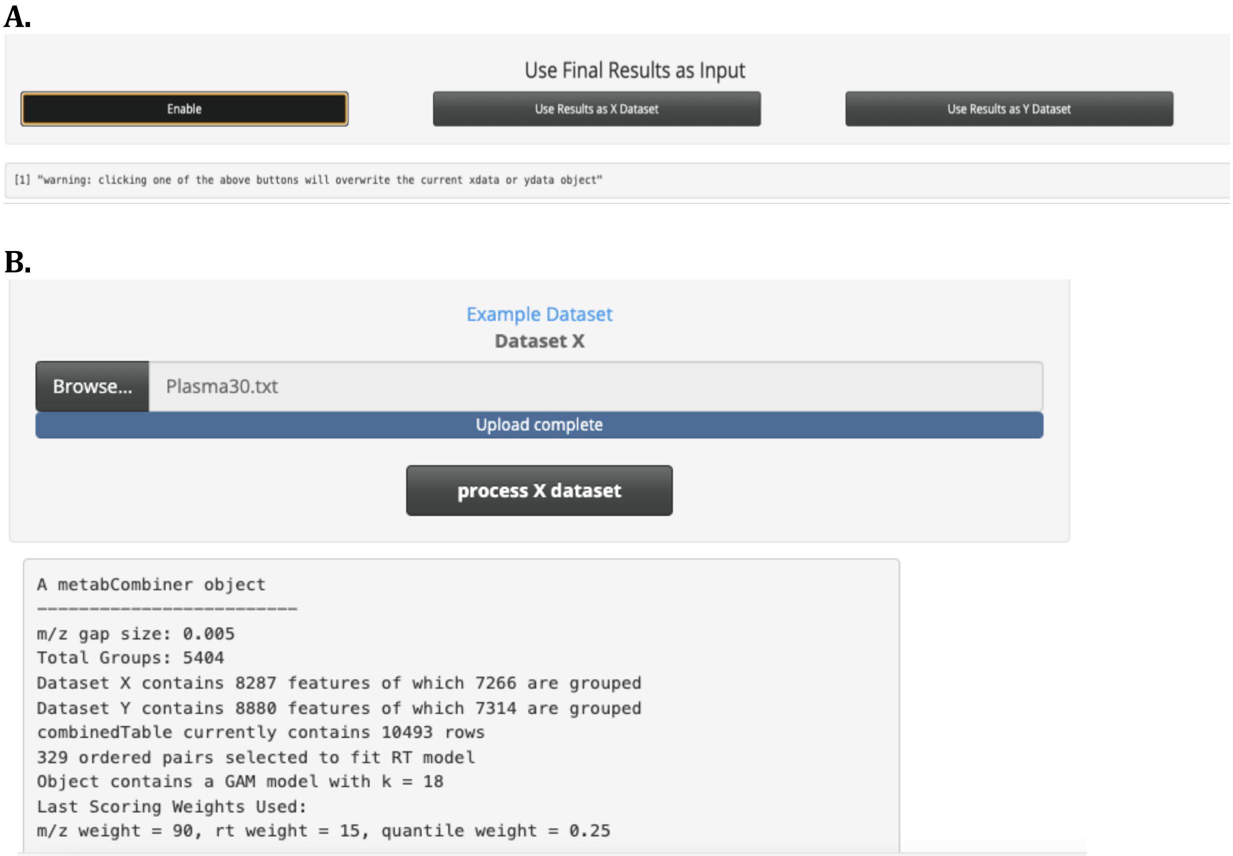
There are three methods for to using combined data as inputs for multi-stage alignment workflows: 1) using the Output tab panel (easiest method); 2) uploading an object .rds file; 3) uploading the table as a single data set (.csv) file.

Method #1: Output Tab Panel

In the Output tab, there is a panel near the bottom of the page (depicted in **Figure 2.18 A**) with buttons for **Using Results as X Dataset** and **Using Results as Y Dataset**. By default, these buttons are disabled. There is a button alongside titled "**Enable**". Clicking on it will remove the lock on both buttons, along with printing a warning that the present action will overwrite the current X dataset or Y dataset. Consider saving all necessary objects to retain analysis progress (see **Section 2.4**).

Clicking on either button below is the simplest method for performing multi-stage analysis.

For this example, clicking on **Use Results as X Dataset** leads to two major changes in the *Input Data* tab under Dataset X: 1) the panel (with parameters for Column Keyword input and Filters) has collapsed as these are no longer relevant for this object; 2) the data summary has been altered to represent a combined table, i.e. the results of the most recent analysis in the same user session.



```
[1] "warning: clicking one of the above buttons will overwrite the current xdata or ydata object"
```

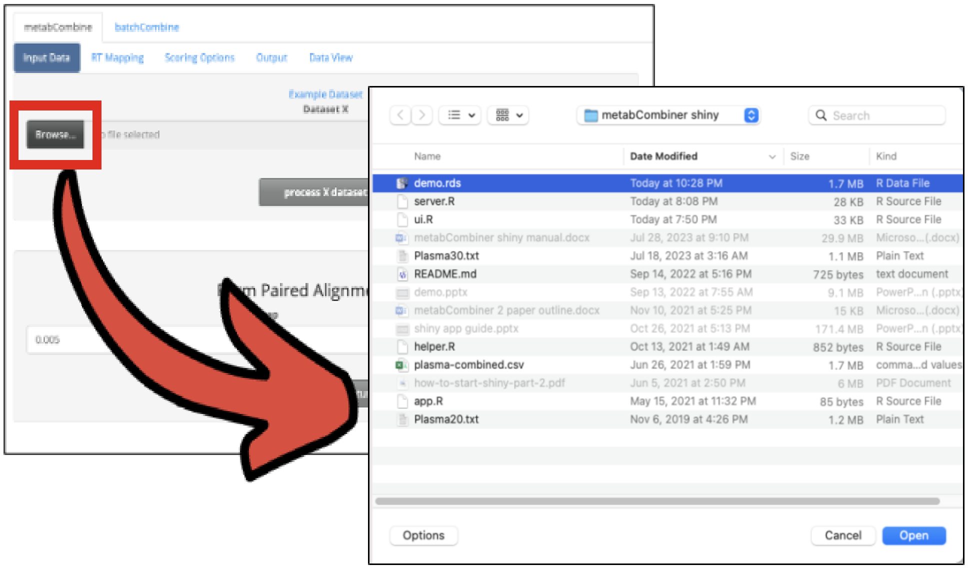
```
A metabCombiner object
-------------------------
m/z gap size: 0.005
Total Groups: 5404
Dataset X contains 8287 features of which 7266 are grouped
Dataset Y contains 8880 features of which 7314 are grouped
combinedTable currently contains 10493 rows
329 ordered pairs selected to fit RT model
Object contains a GAM model with k = 18
Last Scoring Weights Used:
m/z weight = 90, rt weight = 15, quantile weight = 0.25
```

Figure 2.18 A) Using *Final Results as Input* panel in Output tab under *metabCombine*; **B)** After clicking *"Use Results as X Dataset"*, the summary in the Input Data tab is changed to a *metabCombiner* object, replacing the previous xdata object

Method #2: Uploading Object Files (.rds)

To perform this method, save results objects in the *Output* tab to a .rds file. Here, the results of the analysis were saved to demo.rds. Then head back to the *Input Data* tab and upload the file using the Browse button under either X Dataset or Y Dataset, as illustrated in **Figure 2.19**. Once the file is uploaded, click the **Process X Dataset** dataset.

Like method #1, the panel that appears for single data set inputs remains collapsed. This approach requires slightly more inputs than the easiest method, but has the advantage of using saved results without repeating previous analyses, e.g. in new sessions.



Name	Date Modified	Size	Kind
demo.rds	Today at 10:28 PM	1.7 MB	R Data File
server.R	Today at 8:08 PM	28 KB	R Source File
ui.R	Today at 7:50 PM	33 KB	R Source File
metabCombiner shiny manual.docx	Jul 28, 2023 at 9:10 PM	29.9 MB	Microso...(.docx)
Plasma30.txt	Jul 18, 2023 at 3:16 AM	1.1 MB	Plain Text
README.md	Sep 14, 2022 at 5:16 PM	725 bytes	text document
demo.pptx	Sep 13, 2022 at 7:55 AM	9.1 MB	PowerP...n (.pptx
metabCombiner 2 paper outline.docx	Nov 10, 2021 at 5:25 PM	15 KB	Microso...(.docx)
shiny app guide.pptx	Oct 26, 2021 at 5:13 PM	171.4 MB	PowerP...n (.pptx
helper.R	Oct 13, 2021 at 1:49 AM	852 bytes	R Source File
plasma-combined.csv	Jun 26, 2021 at 1:59 PM	1.7 MB	comma...d values
how-to-start-shiny-part-2.pdf	Jun 5, 2021 at 2:50 PM	6 MB	PDF Document
app.R	May 15, 2021 at 11:32 PM	85 bytes	R Source File
Plasma20.txt	Nov 6, 2019 at 4:26 PM	1.2 MB	Plain Text

Figure 2.19 Loading .rds Files as metabCombiner data inputs

Method #3: Uploading Combined Table Files (.csv)

The combined table report (.csv) can be uploaded instead of the object, with the same workflow as previously described in **Section 2.1**. Column keywords will have to be re-mapped to the correct columns, which should map feature descriptors (m/z, RT, id, adduct) to one set of columns, samples to one set of data columns, and extra set to include keywords from all other constituent data sets.

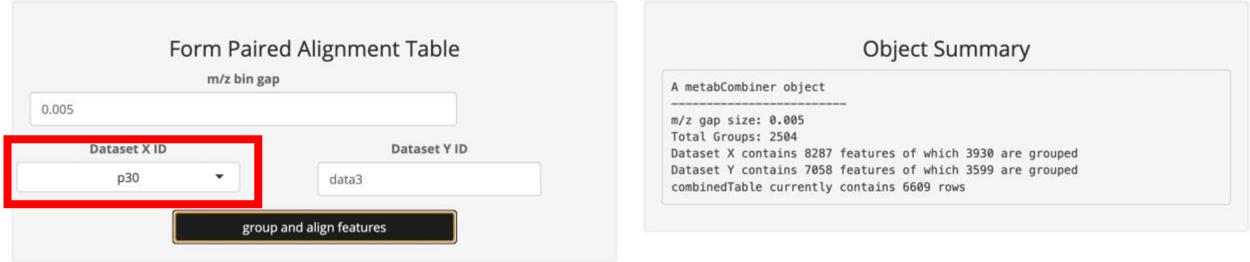
Of the three listed methods, this is the most tedious to perform for multi-dataset alignment workflows. It also would not allow for all non-intersected rows to be reported due to the requirement of non-missing m/z and RT values. This method is mainly useful for allowing manual corrections to *metabCombiner* results, such as choosing alternative alignment hypotheses in "full results" report to those assigned in the "final results" report (i.e. by eliminating the latter). Note that as duplicate features are eliminated by default, only one row (at most) will be accepted for the chosen data set.

2.5.2 Combined Data Merging

Once the combined table object is loaded (using methods #1 or #2 described in Section 2.5.1), upload a new complementary LC-MS metabolomics feature table to be aligned to this table. The new table must adhere to the assumptions of *metabCombiner* (as listed in **Section 1.2**). For this example, another plasma metabolomics feature table was uploaded as the Y Dataset using the same process as outlined in **Section 2.1**.

When arriving at the bottom of the Input Data tab, there is one notable difference to the previous analysis, as illustrated in **Figure 2.20** below. Under **Dataset X ID** for the combined data set input, there is a dropdown menu containing the previously specified data identifiers (p30 and p20 in this example). This menu allows users to specify the "master" or "central" data set that will be used to represent the combined entity. In the current implementation, we recommend using the same "master" data set for all stepwise alignments to generate an "all to one" alignment. **Dataset Y ID** is unchanged from the previous workflow, except that the default identifier value is incremented according to the number of data sets in the combined object.

Click on the **group and align features** button to generate the initial aligned table. A new object summary will appear to the right of the panel.



m/z bin gap: 0.005

Dataset X ID: p30

Dataset Y ID: data3

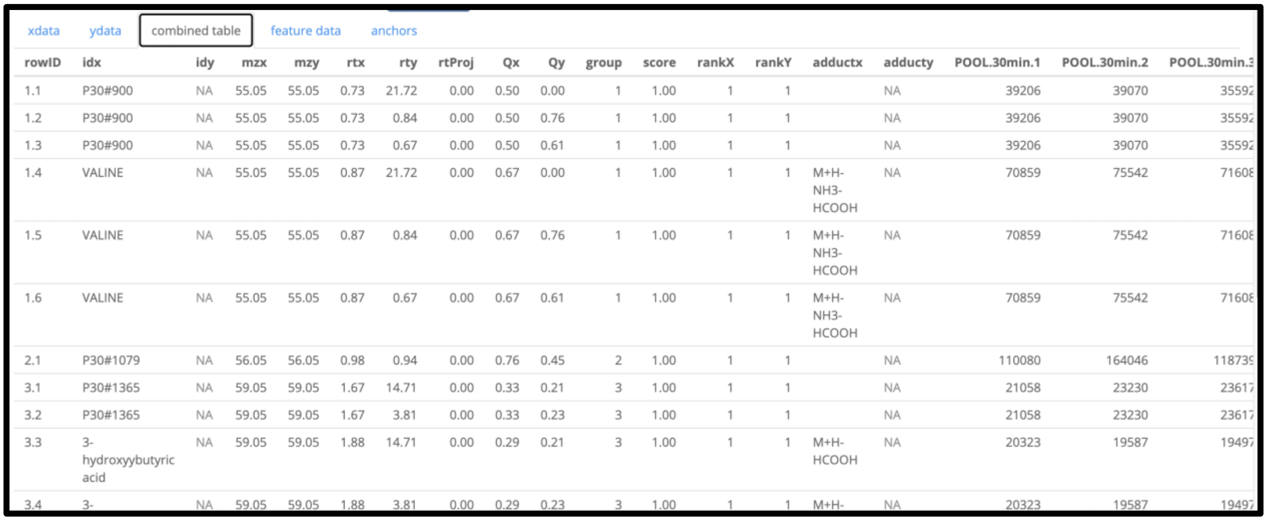
```
A metabCombiner object
-------------------------
m/z gap size: 0.005
Total Groups: 2504
Dataset X contains 8287 features of which 3930 are grouped
Dataset Y contains 7058 features of which 3599 are grouped
combinedTable currently contains 6609 rows
```

Figure 2.20 Form Paired Alignment Table panel with Combined Data inputs

The rest of the alignment steps proceeds exactly according to the steps laid out from **section 2.2** through **section 2.4**. The results of this process can then be incrementally aligned with further data sets in the same manner.

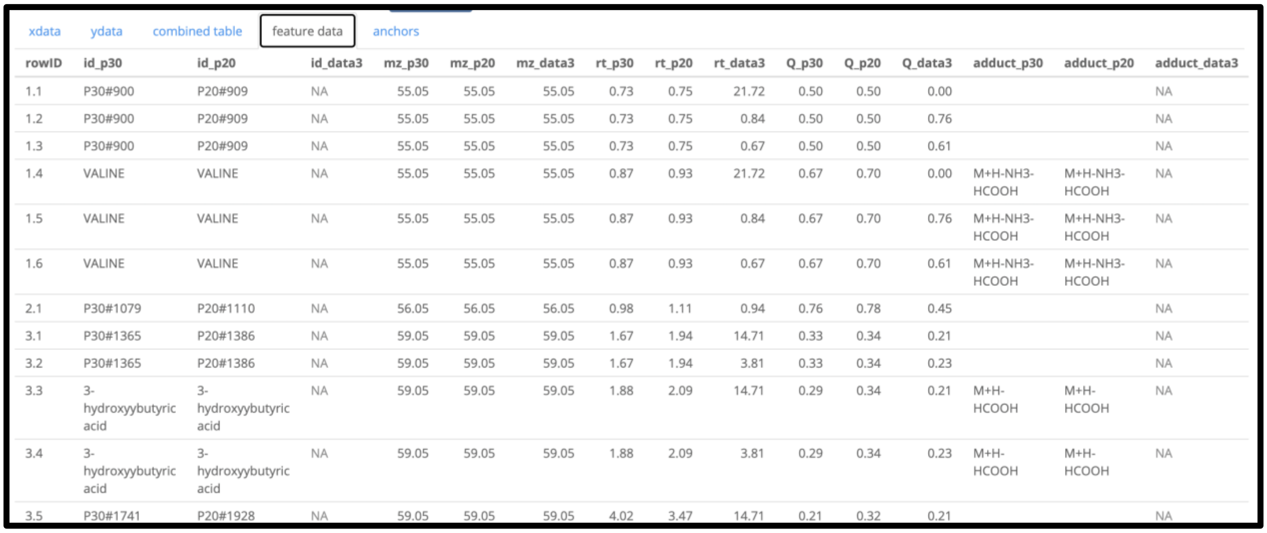
2.5.3 Outputs

While the procedure is largely the same for single data and combined data inputs, the alignment of three or more data sets changes how the output files are interpreted. For the combined table exemplified by **Figure 2.21**, the combined dataset assembled from feature tables "p30" and "p20" is aligned to a new single feature table, "data3". The feature descriptor columns {idx, mzx, rtx, Qx, adductx} are extracted from the "master" feature list (p30), whereas feature descriptors from other constituent dataset(s), e.g. p20, are not displayed in the combined table. Instead, they are provided in the feature data table (**Figure 2.22**), which is organized by descriptor type and data set identifier (provided by the user). The 'samples' and 'extra' columns of all data sets are found in the combined table, to the right of all meta-data columns. As before, the program also provides a results object file that can be used for further alignments, as described in this section.



rowID	idx	idy	mzx	mzy	rtx	rty	rtProj	Qx	Qy	group	score	rankX	rankY	adductx	adducty	POOL.30min.1	POOL.30min.2	POOL.30min.3
1.1	P30#900	NA	55.05	55.05	0.73	21.72	0.00	0.50	0.00	1	1.00	1	1		NA	39206	39070	35592
1.2	P30#900	NA	55.05	55.05	0.73	0.84	0.00	0.50	0.76	1	1.00	1	1		NA	39206	39070	35592
1.3	P30#900	NA	55.05	55.05	0.73	0.67	0.00	0.50	0.61	1	1.00	1	1		NA	39206	39070	35592
1.4	VALINE	NA	55.05	55.05	0.87	21.72	0.00	0.67	0.00	1	1.00	1	1	M+H-NH3-HCOOH	NA	70859	75542	71608
1.5	VALINE	NA	55.05	55.05	0.87	0.84	0.00	0.67	0.76	1	1.00	1	1	M+H-NH3-HCOOH	NA	70859	75542	71608
1.6	VALINE	NA	55.05	55.05	0.87	0.67	0.00	0.67	0.61	1	1.00	1	1	M+H-NH3-HCOOH	NA	70859	75542	71608
2.1	P30#1079	NA	56.05	56.05	0.98	0.94	0.00	0.76	0.45	2	1.00	1	1		NA	110080	164046	118739
3.1	P30#1365	NA	59.05	59.05	1.67	14.71	0.00	0.33	0.21	3	1.00	1	1		NA	21058	23230	23617
3.2	P30#1365	NA	59.05	59.05	1.67	3.81	0.00	0.33	0.23	3	1.00	1	1		NA	21058	23230	23617
3.3	3-hydroxyybutyric acid	NA	59.05	59.05	1.88	14.71	0.00	0.29	0.21	3	1.00	1	1	M+H-HCOOH	NA	20323	19587	19497
3.4	3-	NA	59.05	59.05	1.88	3.81	0.00	0.29	0.23	3	1.00	1	1	M+H-	NA	20323	19587	19497

Figure 2.21 Combined Table in Data View tab. Feature values are drawn from the two data sets being compared.



rowID	id_p30	id_p20	id_data3	mz_p30	mz_p20	mz_data3	rt_p30	rt_p20	rt_data3	Q_p30	Q_p20	Q_data3	adduct_p30	adduct_p20	adduct_data3
1.1	P30#900	P20#909	NA	55.05	55.05	55.05	0.73	0.75	21.72	0.50	0.50	0.00			NA
1.2	P30#900	P20#909	NA	55.05	55.05	55.05	0.73	0.75	0.84	0.50	0.50	0.76			NA
1.3	P30#900	P20#909	NA	55.05	55.05	55.05	0.73	0.75	0.67	0.50	0.50	0.61			NA
1.4	VALINE	VALINE	NA	55.05	55.05	55.05	0.87	0.93	21.72	0.67	0.70	0.00	M+H-NH3-HCOOH	M+H-NH3-HCOOH	NA
1.5	VALINE	VALINE	NA	55.05	55.05	55.05	0.87	0.93	0.84	0.67	0.70	0.76	M+H-NH3-HCOOH	M+H-NH3-HCOOH	NA
1.6	VALINE	VALINE	NA	55.05	55.05	55.05	0.87	0.93	0.67	0.67	0.70	0.61	M+H-NH3-HCOOH	M+H-NH3-HCOOH	NA
2.1	P30#1079	P20#1110	NA	56.05	56.05	56.05	0.98	1.11	0.94	0.76	0.78	0.45			NA
3.1	P30#1365	P20#1386	NA	59.05	59.05	59.05	1.67	1.94	14.71	0.33	0.34	0.21			NA
3.2	P30#1365	P20#1386	NA	59.05	59.05	59.05	1.67	1.94	3.81	0.33	0.34	0.23			NA
3.3	3-hydroxyybutyric acid	3-hydroxyybutyric acid	NA	59.05	59.05	59.05	1.88	2.09	14.71	0.29	0.34	0.21	M+H-HCOOH	M+H-HCOOH	NA
3.4	3-hydroxyybutyric acid	3-hydroxyybutyric acid	NA	59.05	59.05	59.05	1.88	2.09	3.81	0.29	0.34	0.23	M+H-HCOOH	M+H-HCOOH	NA
3.5	P30#1741	P20#1928	NA	59.05	59.05	59.05	4.02	3.47	14.71	0.21	0.32	0.21			NA

Figure 2.22 Feature Data in Data View tab. Feature values from all data sets {p30, p20, data3} are displayed.

Chapter 3

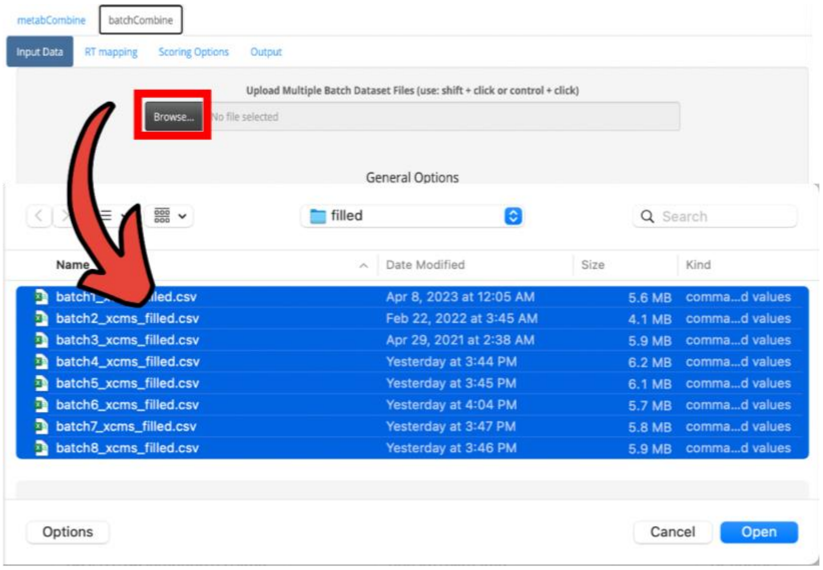
batchCombine

3.1 Input Data Tab

This section describes the initial steps to performing large-scale metabolomics data alignment, where each LC-MS batch is represented as a separate feature table. The Input Data tab contains one panel with modules for uploading and formatting feature tables. With a few notable differences, this tab under *batchCombine* operates very similarly to the Input Data tab of *metabCombine*. The panel is divided into five sections: Data Upload, General Options, Keyword Mapping, Feature Filters, and Stepwise Averaging.

3.1.1 Data Upload

After pre-processing the raw data of each batch separately, print each to a .csv or .txt file, placing all batch files within one folder. Order all the batches such that the first batch appears first and the final batch appears last. Navigate to batchCombine and click the Browse button in the Input Data tab. This will open a navigation menu, such as depicted in **Figure 3.1**. Using "shift + click" or "control + click", select all batch data files to upload them into the program for alignment, then click "Open". Once successful, a blue bar will appear that says, "Upload complete". **Note**: multiple files must be selected for this process to proceed. Choosing one file will lead to a silent error and no file will be uploaded.



Upload Multiple Batch Dataset Files (use: shift + click or control + click)

No file selected

Name	Date Modified	Size	Kind
batch1_[illegible]led.csv	Apr 8, 2023 at 12:05 AM	5.6 MB	comma...d values
batch2_xcms_filled.csv	Feb 22, 2022 at 3:45 AM	4.1 MB	comma...d values
batch3_xcms_filled.csv	Apr 29, 2021 at 2:38 AM	5.9 MB	comma...d values
batch4_xcms_filled.csv	Yesterday at 3:44 PM	6.2 MB	comma...d values
batch5_xcms_filled.csv	Yesterday at 3:45 PM	6.1 MB	comma...d values
batch6_xcms_filled.csv	Yesterday at 4:04 PM	5.7 MB	comma...d values
batch7_xcms_filled.csv	Yesterday at 3:47 PM	5.8 MB	comma...d values
batch8_xcms_filled.csv	Yesterday at 3:46 PM	5.9 MB	comma...d values

Figure 3.1 Multi-batch file upload in *batchCombine*

3.1.2 General Options

There are two parameters to customize in the General Options section: the **batch ID prefix** and **operation**. The batch ID prefix is a common character string for naming the batch data sets in the output files. For example, if the prefix is set to "example", the batches will be named "example1", "example2", "example3", etc... The operation parameter determines whether to take the union or intersection of all features. The *intersection* option removes any features absent in one or more batches, whereas the *union* aligns all features detected in at least one batch.

3.1.3 Column Keyword Mapping

Column keyword mapping in *batchCombine* works similarly to *metabCombine*, as described in **Section 2.1.2**. One notable difference is that all batch feature tables provided as input must be similarly formatted, i.e. columns are named similarly so that keywords are mapped to the correct columns in all batch tables. The example data sets in *metabCombine* provide a template for naming the meta-data (id, mz, rt, adduct) columns. If using pooled samples for relative quantitation ("samples"), include a similar term (e.g. "pool") in all names of columns representing pooled samples in the batch feature tables, then provide one or multiple keywords to match all other columns in the "extra" field.

3.1.4 Feature Filters

As with *metabCombine*, there are three feature filters performed on input tables:

1) retention time range filters set the maximum and minimum feature RTs, excluding all features outside this range

2) missingness filter eliminates any feature missing in over the specified percentage of samples, as defined in the keyword mapping of 'samples'. Optionally, zero values may be treated as missing by the program.

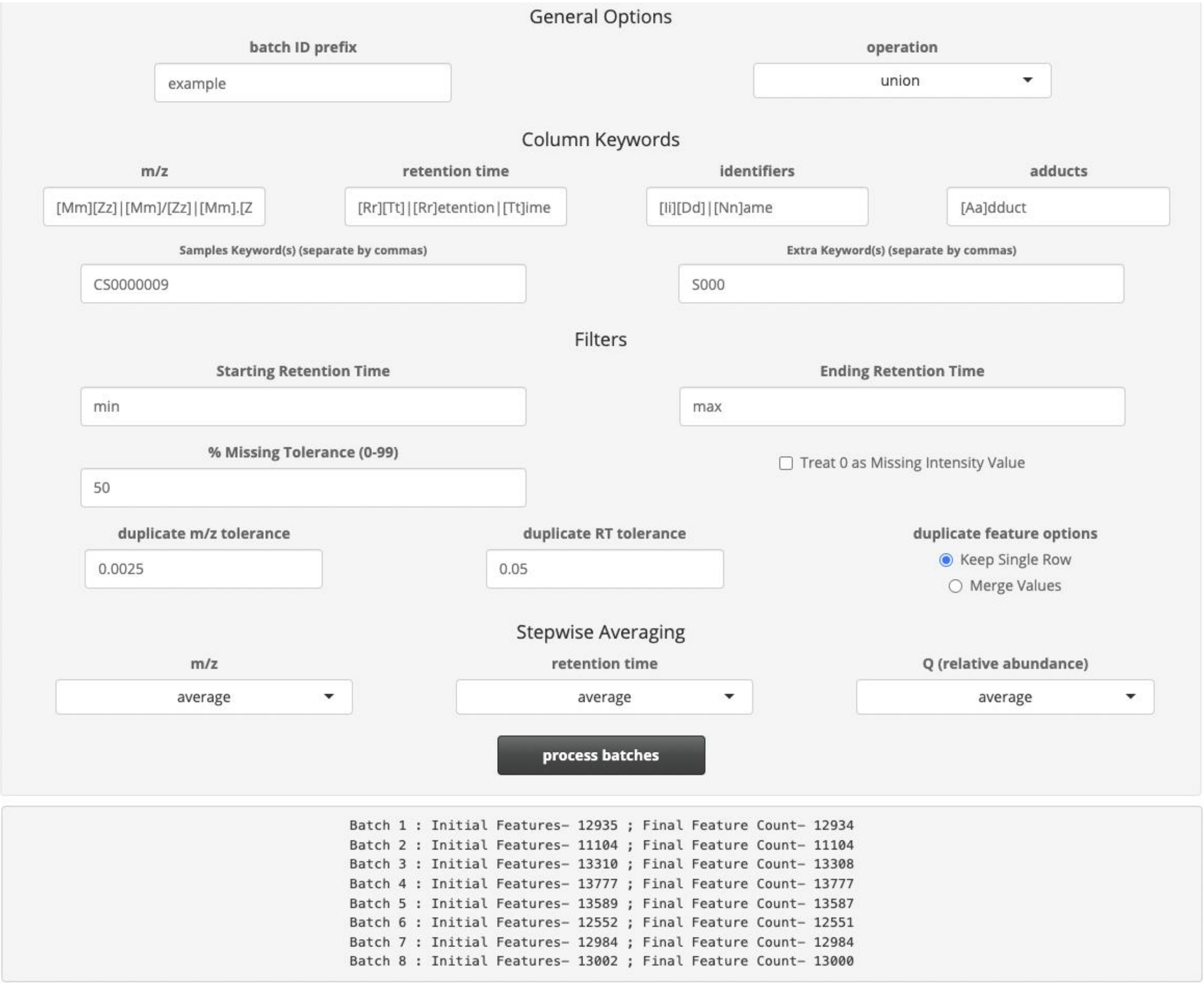
3) duplicate feature filter merges or eliminates features with the same m/z and RT values (within specified tolerances)

The same filter parameters are applied to all batch feature tables prior to alignment.

3.1.5 Stepwise Averaging

Near the bottom of the Input Data panel is a trio of options for stepwise averaging. When the batches are sequentially aligned (i.e. batch 1 with batch 2, followed by batches 1 - 2 with 3, batches 1-2-3 with 4, etc...), the features of the batch-merged table are compared to the next batch using the "observed" m/z, RT, Q values of the latter aligned batch or with the averaged value(s) of previously merged batches. Stepwise averaging is often useful for data with random RT and m/z shifts observed for one or multiple batches.

After customizing all necessary parameters, click the **process batches** button. A small panel will appear briefly summarizing the feature counts of each batch table.



batch ID prefix
example

operation
union

m/z
[Mm][Zz]|[Mm]/[Zz]|[Mm].[Z

retention time
[Rr][Tt]|[Rr]etention|[Tt]ime

identifiers
[Ii][Dd]|[Nn]ame

adducts
[Aa]dduct

Samples Keyword(s) (separate by commas)
CS0000009

Extra Keyword(s) (separate by commas)
S000

Starting Retention Time
min

Ending Retention Time
max

% Missing Tolerance (0-99)
50

☐ Treat 0 as Missing Intensity Value

duplicate m/z tolerance
0.0025

duplicate RT tolerance
0.05

duplicate feature options
● Keep Single Row
○ Merge Values

Stepwise Averaging

m/z
average

retention time
average

Q (relative abundance)
average

```
Batch 1 : Initial Features- 12935 ; Final Feature Count- 12934
Batch 2 : Initial Features- 11104 ; Final Feature Count- 11104
Batch 3 : Initial Features- 13310 ; Final Feature Count- 13308
Batch 4 : Initial Features- 13777 ; Final Feature Count- 13777
Batch 5 : Initial Features- 13589 ; Final Feature Count- 13587
Batch 6 : Initial Features- 12552 ; Final Feature Count- 12551
Batch 7 : Initial Features- 12984 ; Final Feature Count- 12984
Batch 8 : Initial Features- 13002 ; Final Feature Count- 13000
```

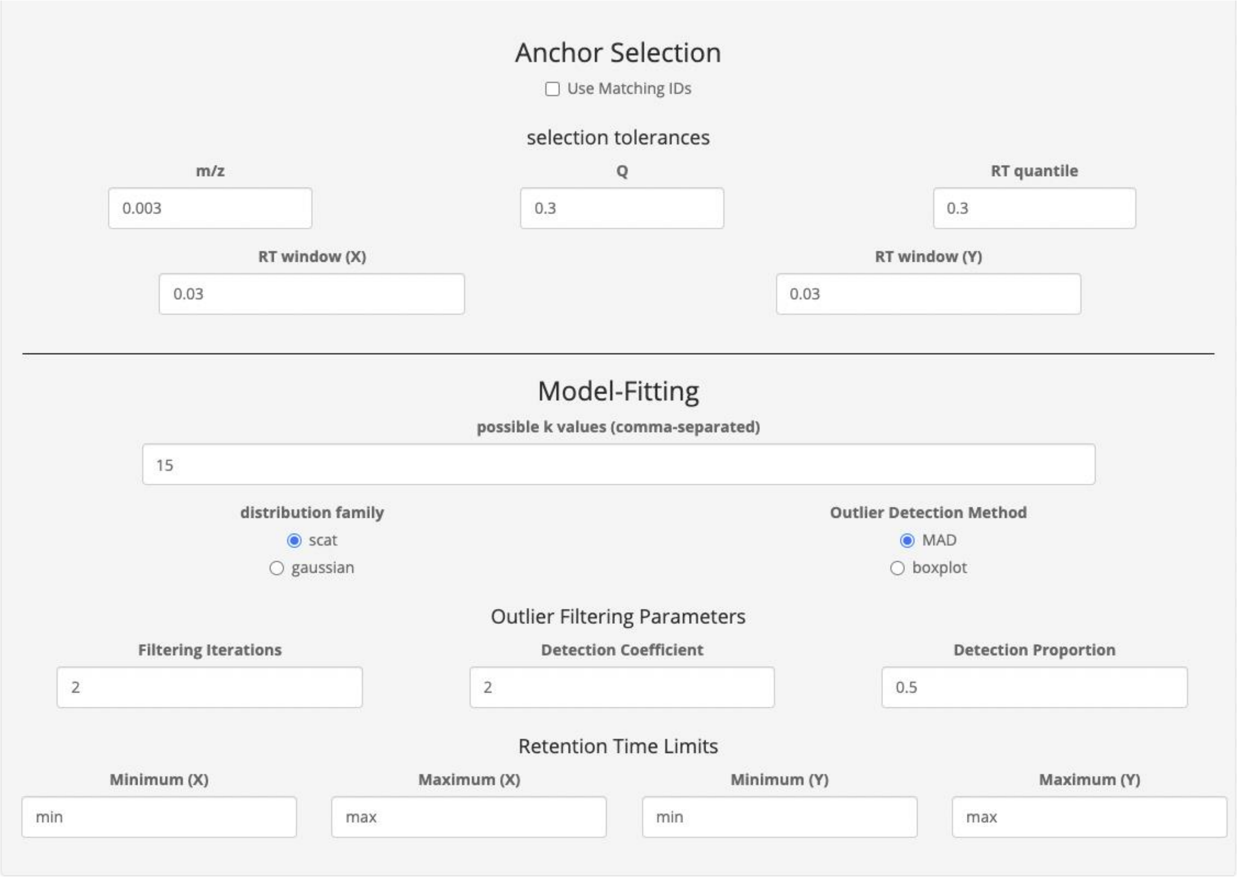
Figure 3.2 *batchCombine* Input Data panel. Lower panel summarizes batches after clicking "process batches" button. Note: 'CS0000009' is a common keyword for pooled samples and every sample name contains 'S000' in this example.

3.2 RT Mapping Tab

This tab contains all parameters for the mapping between retention times of the batches. All parameters were previously described in **Section 2.2**, with no differences except that the RT shift is expected to be much smaller between metabolomics batches as compared to non-identically acquired LC-MS methods. The default parameters for anchor selection and will be suitable in nearly all *batchCombine* applications

RT mapping is typically the bottleneck in these multi-batch alignments. The analysis can be sped up by providing one value for **possible k values** (between 10 and 20 is recommended), skipping cross-validation.

There are no buttons in this tab as the RT mapping will be performed sequentially with other downstream steps.



Anchor Selection

□ Use Matching IDs

selection tolerances

m/z: 0.003

Q: 0.3

RT quantile: 0.3

RT window (X): 0.03

RT window (Y): 0.03

Model-Fitting

possible k values (comma-separated): 15

distribution family: ● scat ○ gaussian

Outlier Detection Method: ● MAD ○ boxplot

Outlier Filtering Parameters

Filtering Iterations: 2

Detection Coefficient: 2

Detection Proportion: 0.5

Retention Time Limits

Minimum (X): min

Maximum (X): max

Minimum (Y): min

Maximum (Y): max

Figure 3.3 RT Mapping panel under *batchCombine*.

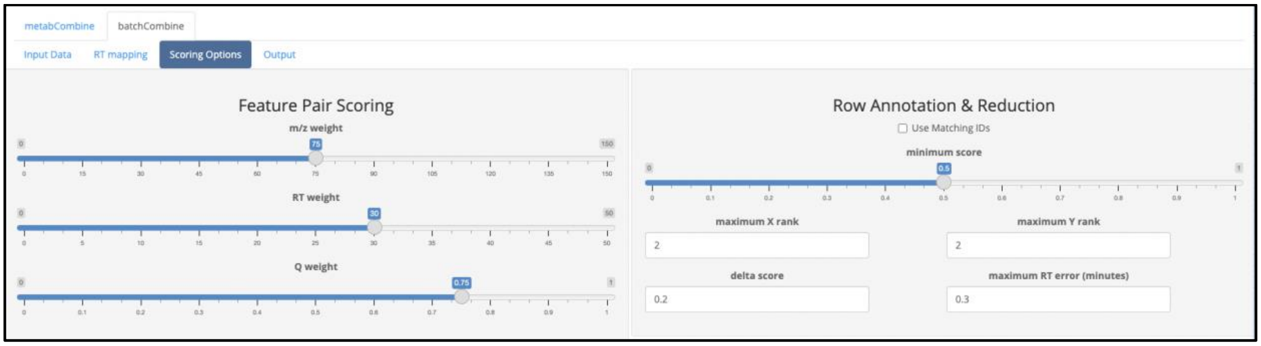
3.3 Scoring Options

See **Section 2.3** for a more in-depth explanation of how feature pair scoring is performed in *metabCombiner*. The main differences in scoring options for *batchCombine* compared to *metabCombine* are:

1. The higher default **RT weight** value of 30, along with a higher maximum value (50) compared to *metabCombine* applications

2. A higher default Q weight of 0.5
3. A default maximum RT error value of 0.5 (under Row Annotation & Reduction).

As with the RT mapping tab, there are no buttons to click here as all scoring steps are performed sequentially once the program is run.



Feature Pair Scoring

m/z weight: 75 (0–150)

RT weight: 30 (0–50)

Q weight: 0.75 (0–1)

Row Annotation & Reduction

□ Use Matching IDs

minimum score: 0.5 (0–1)

maximum X rank: 2

maximum Y rank: 2

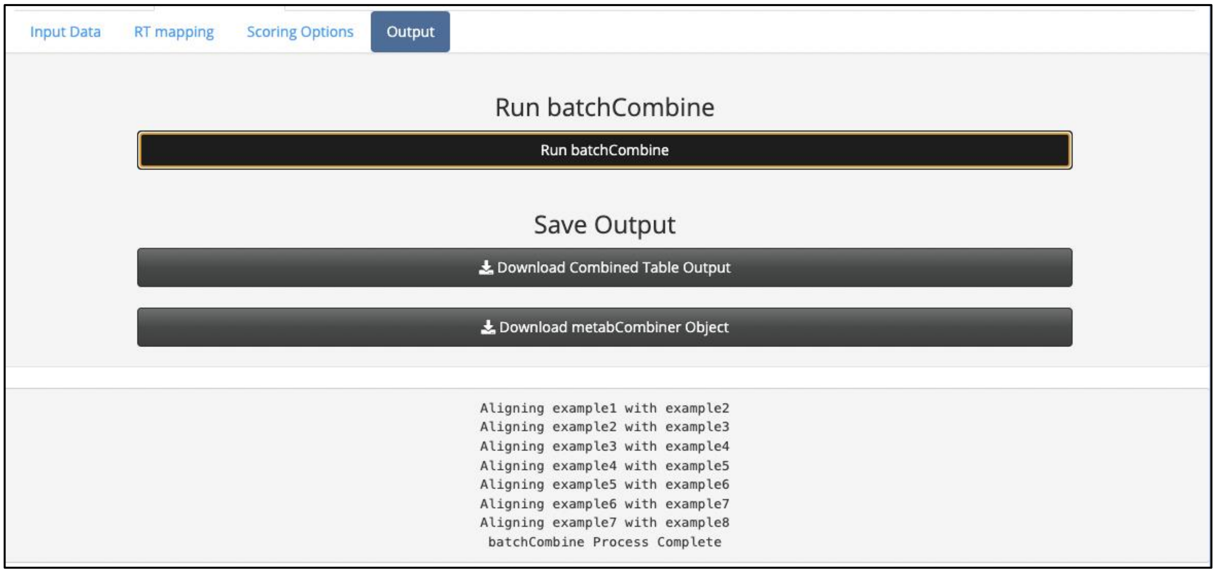
delta score: 0.2

maximum RT error (minutes): 0.3

Figure 3.4 Scoring Options panel under *batchCombine*

3.4 Output

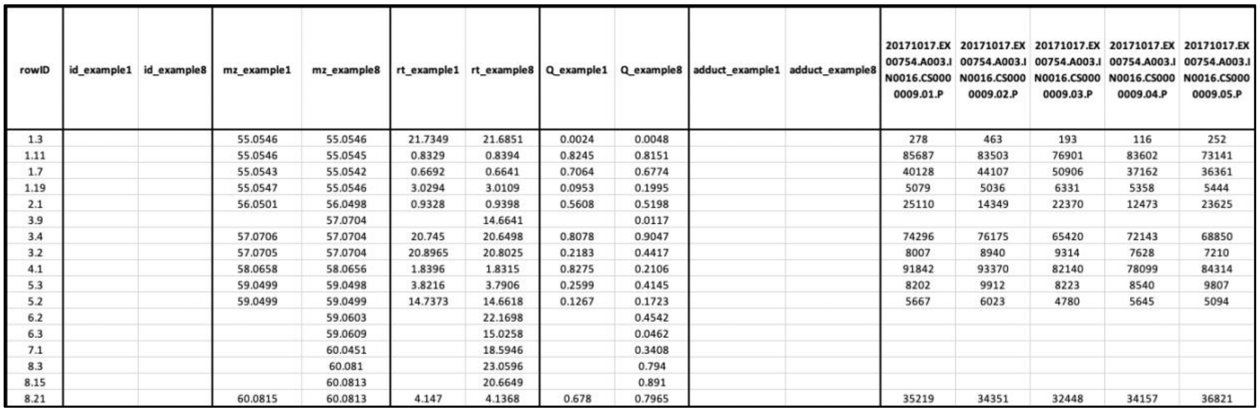
The Output tab under *batchCombine* has a simple layout consisting of three buttons. When first launching the program, all three buttons are de-activated. Upon loading the feature tables in the Input Data tab, the **Run batchCombine** button is enabled. Clicking on this button will start the *batchCombine* process, guided by parameters in the three other tabs. A panel will appear below, providing the current progress of the alignment. Each step of the multi-batch alignment will use the latest aligned batch's feature descriptors or the averaged values of previously merged batches to compare with the next, until the last of the stepwise alignments is complete. A "*batchCombine* Process Complete" message will appear at the end of the progress panel.



```
Aligning example1 with example2
Aligning example2 with example3
Aligning example3 with example4
Aligning example4 with example5
Aligning example5 with example6
Aligning example6 with example7
Aligning example7 with example8
 batchCombine Process Complete
```

Figure 3.4 Output panel under *batchCombine*

Completion of *batchCombine* will enable the two buttons for downloading the output file objects. The **Download Combined Table Output** button can be used to download the batch-merged feature table, which includes the meta-data of all aligned features from each batch followed by the samples and 'extra' columns. **Figure 3.6** shows an example output from the example in this demonstration. The **Download *metabCombiner* Object** button generates a metabCombiner object that can be used in *metabCombine*, with each of the individual batches treated as separate data sets.



rowID	id_example1	id_example8	mz_example1	mz_example8	rt_example1	rt_example8	Q_example1	Q_example8	adduct_example1	adduct_example8	20171017.EX00754.A003.IN0016.CS0000009.01.P	20171017.EX00754.A003.IN0016.CS0000009.02.P	20171017.EX00754.A003.IN0016.CS0000009.03.P	20171017.EX00754.A003.IN0016.CS0000009.04.P	20171017.EX00754.A003.IN0016.CS0000009.05.P
1.3			55.0546	55.0546	21.7349	21.6851	0.0024	0.0048			278	463	193	116	252
1.11			55.0546	55.0545	0.8329	0.8394	0.8245	0.8151			85687	83503	76901	83602	73141
1.7			55.0543	55.0542	0.6692	0.6641	0.7064	0.6774			40128	44107	50906	37162	36361
1.19			55.0547	55.0546	3.0294	3.0109	0.0953	0.1995			5079	5036	6331	5358	5444
2.1			56.0501	56.0498	0.9328	0.9398	0.5608	0.5198			25110	14349	22370	12473	23625
3.9				57.0704		14.6641		0.0117							
3.4			57.0706	57.0704	20.745	20.6498	0.8078	0.9047			74296	76175	65420	72143	68850
3.2			57.0705	57.0704	20.8965	20.8025	0.2183	0.4417			8007	8940	9314	7628	7210
4.1			58.0658	58.0656	1.8396	1.8315	0.8275	0.2106			91842	93370	82140	78099	84314
5.3			59.0499	59.0498	3.8216	3.7906	0.2599	0.4145			8202	9912	8223	8540	9807
5.2			59.0499	59.0499	14.7373	14.6618	0.1267	0.1723			5667	6023	4780	5645	5094
6.2				59.0603		22.1698		0.4542							
6.3				59.0609		15.0258		0.0462							
7.1				60.0451		18.5946		0.3408							
8.3				60.081		23.0596		0.794							
8.15				60.0813		20.6649		0.891							
8.21			60.0815	60.0813	4.147	4.1368	0.678	0.7965			35219	34351	32448	34157	36821

Figure 3.6 Example *batchCombine* output for 8 aligned batches (columns for example2 through example7 hidden in image, but present in results)

Bibliography

[1] Habra, H.; Kachman, M.; Bullock, K.; Clish, C.; Evans, C. R.; Karnovsky, A. *metabCombiner* : Paired Untargeted LC-HRMS Metabolomics Feature Matching and Concatenation of Disparately Acquired Data Sets. *Anal. Chem.* **2021**, *93* (12), 5028–5036. https://doi.org/10.1021/acs.analchem.0c03693.