

Optimal transport for automatic alignment of untargeted metabolomic data

Reviewed Preprint

Revised by authors after peer review.

About eLife's process

Reviewed preprint version 2
April 9, 2024 (this version)

Reviewed preprint version 1
December 11, 2023

Posted to preprint server
September 13, 2023

Sent for peer review
September 11, 2023

Marion Beuerle, Stepan Pankovskii, Racheil Rippe, Jilvee Anand, Mallon Nutrition and Metabolism, National Cancer Institute, Cambridge, MA, USA

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

Abstract

Untargeted metabolomic profiling (mass spectrometry (MS) measures a wide range of metabolites within biospecimens, disease diagnosis, and risk prediction. However, challenge for data integration, and experimental cost the merging of multiple datasets. Current data pooling limitations due to batch effects and varying retention times. Here we introduce Gromov-Matcher, a flexible and robust method that combines LC-MS datasets by aligning retention time correlation structures. Gromov-Matcher outperforms existing methods compared to existing approaches. This algorithm simplifies parameter tuning. Applying our method to experimental liver and pancreatic cancer, we discover shared metabolites, demonstrating how Gromov-Matcher facilitates cross-study integration with lifestyle data for cancer risk prediction.

eLife assessment

The authors describe Gromov-Matcher, that can be used for metabolomic data integration. The underlying method and the algorithm is clearly described in the accompanying manuscript.

Introduction

Untargeted metabolomics is a powerful analytical technique used to identify and measure a large number of metabolites in a biological sample without preselecting targets [Patti \(2011\)](#). This approach allows for a comprehensive overview of an individual's metabolic profile, provides insights into the biochemical processes involved in cellular and organismal physiology [Wishart \(2019\)](#); [Pirhaji et al. \(2016\)](#), and allows for the exploration of how environmental factors

impact metabolism [Rappaport et al. \(2014\)](#) [Bedia \(2022\)](#) [Li et al. \(2020\)](#) [Wang et al. \(2011\)](#) [Franzosa et al. \(2019\)](#) [Lofthfield et al. \(2021\)](#) [Li et al. \(2020\)](#) [Lofthfield et al. \(2021\)](#). It creates new opportunities to investigate health-related conditions, including diabetes [Wang et al. \(2011\)](#) [Franzosa et al. \(2019\)](#), and various cancer types [Lofthfield et al. \(2021\)](#) [Li et al. \(2020\)](#) [Lofthfield et al. \(2021\)](#). However, a major challenge in biomarker discovery, metabolic signature identification and other untargeted metabolomic analyses lies in the low throughput of experimental data, necessitating the development of efficient pooling algorithms capable of merging datasets from multiple sources [Lofthfield et al. \(2021\)](#) [Lofthfield et al. \(2021\)](#).

A common experimental technique in untargeted metabolomics is liquid chromatography-mass spectrometry (LC-MS) which assembles a list of thousands of unlabeled metabolic features characterized by their mass-to-charge ratio (m/z), retention time (RT) [Zhou et al. \(2012\)](#) [Zhou et al. \(2012\)](#) [Ivanisevic and Want \(2019\)](#). This problem is further compounded by differing instruments and analytical protocols across laboratories, resulting in seemingly incompatible metabolomic datasets.

Manual matching of metabolic features can be a laborious and error-prone task [Lofthfield et al. \(2021\)](#) [Lofthfield et al. \(2021\)](#). To address this challenge, several automated methods have been developed for metabolic feature alignment. One such method is MetaXCMS, which matches LC-MS features based on user-defined m/z and RT thresholds [Tautenhahn et al. \(2011\)](#) [Tautenhahn et al. \(2011\)](#). More advanced tools use information on feature intensities measured in samples. For instance, PAIRUP-MS uses known shared metabolic features to impute the intensities of all features from one dataset to another [Hsu et al. \(2019\)](#) [Hsu et al. \(2019\)](#). MetabCombiner [Habra et al. \(2021\)](#) [Habra et al. \(2021\)](#) and M2S [Climaco Pinto et al. \(2022\)](#) [Climaco Pinto et al. \(2022\)](#) compare average feature intensities, along with their m/z and RT values, to align datasets without requiring extensive knowledge of shared features. These automated alignment methods have accelerated our ability to pool and annotate datasets as well as extract biologically meaningful biomarkers. However, they demand substantial fine-tuning of user-defined parameters and ignore correlations among metabolic features which provide a wealth of additional information on shared features.

Here we introduce GromovMatcher, a user-friendly flexible algorithm which automates the matching of metabolic features across experiments. The main technical innovation of GromovMatcher lies in its ability to incorporate the correlation information between metabolic feature intensities, building upon the powerful mathematical framework of computational optimal transport (OT) [Peyré et al. \(2019\)](#) [Villani \(2021\)](#) [Villani \(2021\)](#). OT has proven effective in solving various matching problems and has found applications in multiomics analysis [Demetci et al. \(2022\)](#) [Demetci et al. \(2022\)](#), cell development [Schiebinger et al. \(2019\)](#) [Yang et al. \(2020\)](#) [Yang et al. \(2020\)](#), and chromatogram alignment [Skoraczynski et al. \(2022\)](#) [Skoraczynski et al. \(2022\)](#). Here we leverage the Gromov-Wasserstein (GW) method [Mémoli \(2011\)](#) [Solomon et al. \(2016\)](#) [Mémoli \(2011\)](#) [Solomon et al. \(2016\)](#), which matches datasets based on their distance structure and has been seminal applied to spatial reconstruction problems in genomics [Nitzan et al. \(2019\)](#) [Nitzan et al. \(2019\)](#). GromovMatcher builds upon the GW algorithm to automatically uncover the shared correlation structure among metabolic feature intensities while also incorporating m/z and RT information in the final matching process.

To assess the performance of GromovMatcher, we systematically benchmark it on synthetic data with varying levels of noise, feature overlap, and data normalizations, outperforming prior state-of-the-art methods of metabCombiner [Habra et al. \(2021\)](#) [Habra et al. \(2021\)](#) and M2S [Climaco Pinto et al. \(2022\)](#) [Climaco Pinto et al. \(2022\)](#). Next we apply GromovMatcher to align experimental patient studies of liver and pancreatic cancer to a reference dataset and associate the shared metabolic features to each patient's alcohol intake. Through these efforts, we demonstrate how GromovMatcher data pooling improves our ability to discover biomarkers of lifestyle risk factors associated with several types of cancer.

Results

GromovMatcher algorithm

GromovMatcher uses the mathematical framework of OT to find all matching metabolic features between two untargeted metabolomic datasets (Fig. 1a). It accepts two LC-MS datasets with possibly different numbers of metabolic features and samples. Each feature, fx_i , in Dataset 1 and fy_j in Dataset 2, is identified by its m/z , RT, and vector of feature intensities across samples (Fig. 1a). The primary tenet of GromovMatcher is that shared metabolic features have similar correlation patterns in both datasets and can be matched based on the distance/correlations between their feature intensity vectors. Specifically, GromovMatcher computes the pairwise distances between the feature intensity vectors of each metabolic feature in a dataset and saves them into a distance matrix, one per dataset (Fig. 1b). In practice, we use either the Euclidean distance or the cosine distance (negative of correlation) to perform this step (Methods). The resulting distance matrices contain information about the feature intensity similarity within each study. Using optimal transport, we can deduce shared subsets of metabolic features in both datasets which have corresponding feature intensity distance structures.

OT was originally developed to optimize the transportation of soil for the construction of forts Monge (1781) and was later generalized through the language of probability theory and linear programming Kantorovich (2006), leading to efficient numerical algorithms and direct applications to planning problems in economics. The ability of OT to efficiently match source to target locations found applications in data science for the alignment of distributions Courty et al. (2017); Alvarez-Melis et al. (2019) and was generalized by the Gromov-Wasserstein (GW) method Peyré et al. (2016); Alvarez-Melis and Jaakkola (2018) to align datasets with features of differing dimensions.

In practice, a sizeable fraction of the metabolic features measured in one study may not be present in the other. Hence, in most cases only a subset of features in both datasets can be matched. Recent GW formulations for unbalanced matching problems Sejourne et al. (2021) allow for matching only subsets of metabolic features with similar intensity structures (Fig. 1c). To incorporate additional feature information, we modify the optimization objective of unbalanced GW to penalize feature matches whose m/z differences exceed a fixed threshold (Methods, Appendix 1). The optimization of this objective computes a coupling matrix $\hat{\Pi}$ where each entry $\hat{\Pi}_{ij} \geq 0$ indicates the level of confidence in matching metabolic feature fx_i in Dataset 1 to fy_j in Dataset 2.

Differences in experimental conditions can induce variations in RT between datasets that can be nonlinear and large in magnitude Zhou et al. (2012); Climaco Pinto et al. (2022); Habra et al. (2021). In the spirit of previous methods for LC-MS batch or dataset alignment Smith et al. (2006); Brunius et al. (2016); Liu et al. (2020); Vaughan et al. (2012); Habra et al. (2021); Climaco Pinto et al. (2022); Skoraczynski et al. (2022), the learned coupling $\hat{\Pi}$ is used to estimate a nonlinear map (drift function) between RTs of both datasets by weighted spline regression, which allows us to filter unlikely matches from the coupling matrix to obtain a refined coupling matrix $\hat{\Pi}$ (Fig. 1d, Methods). An optional thresholding step removes matches with small weights from the coupling matrix. The final output of GromovMatcher is a binary matching matrix M where M_{ij} is equal to 1 if features fx_i and fy_j are matched and 0 otherwise. Throughout the paper, we refer to the two variants of GromovMatcher, with and without the optional thresholding step as GMT and GM respectively.

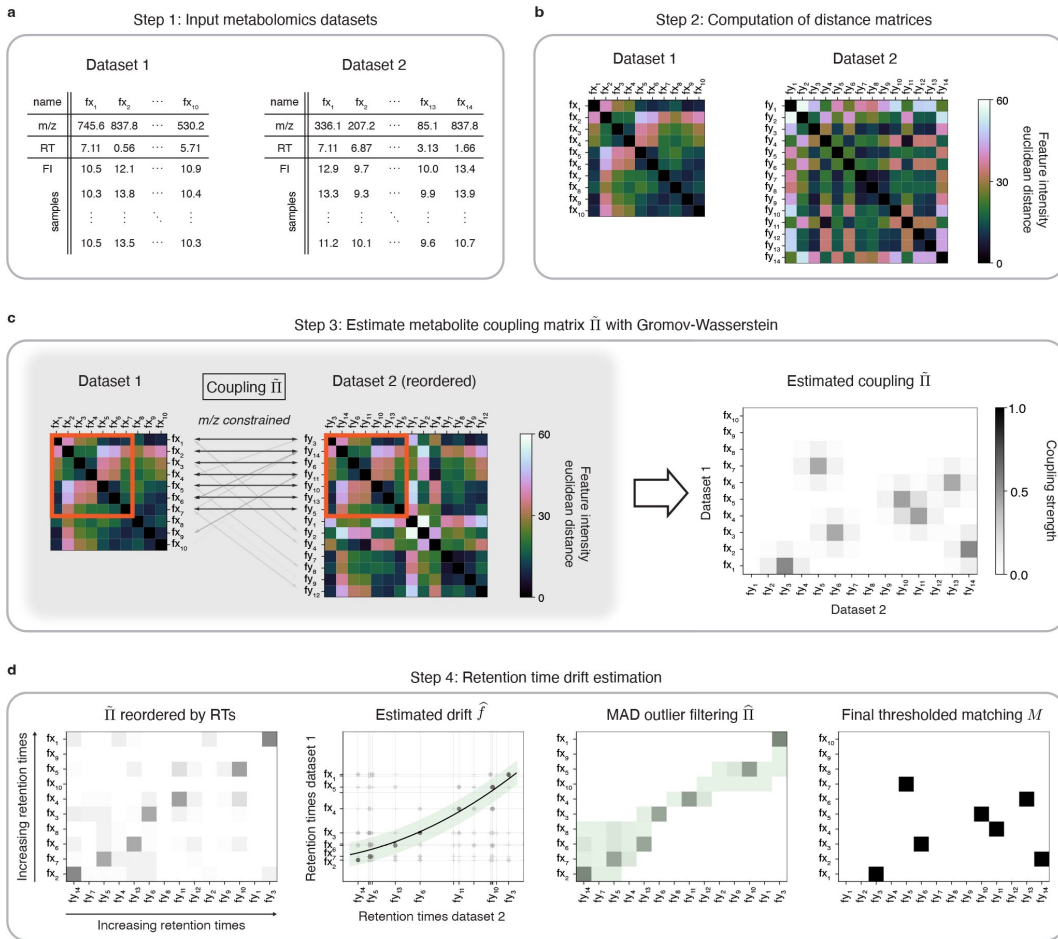


Figure 1.

An optimal pipeline for combining untargeted metabolomic datasets. (a)

(a) Inputs were MS datasets of unlabeled metabolites from two different studies across biospecimen samples. Both studies can have different numbers of metabolites, the sizes of the datasets are not necessarily the same. In this work, we use the Gromov-Wasserstein (GW) algorithm to estimate the coupling matrix $\hat{\Pi}$ between the two datasets. The GW algorithm is a generalization of the optimal transport (OT) algorithm to unbalanced GW algorithm. It finds a coupling matrix $\hat{\Pi}$ that minimizes the cost function $\sum_{i,j} \hat{\Pi}_{ij} d_{ij}$ where d_{ij} is the distance between features i and j in the two datasets. The coupling matrix $\hat{\Pi}$ is then binarized to obtain the final matching matrix M .

Validation on ground truth data

We first evaluate the performance of GromovMatcher using a real-world untargeted metabolomics study of cord blood across 499 newborns containing 4,712 metabolic features characterized by their m/z , RT, and feature intensities [Alfano et al. \(2020\)](#). To generate ground-truth data, we randomly divide the initial dataset into two smaller datasets sharing a subset of features (Fig. 2). We simulate diverse acquisition conditions by adding noise to the m/z and RT of dataset 2, and to the feature intensities in both datasets. Moreover, we introduce an RT drift in dataset 2 to replicate the retention time variations observed in real LC-MS experiments (Methods and Materials). For comparison, we also test M2S [Climaco Pinto et al. \(2022\)](#) and metabCombiner [Habra et al. \(2021\)](#), both of which use m/z , RT, and median or mean feature intensities to match features (Fig. 3). MetabCombiner is supplied with 100 known shared metabolic features to automatically set its hyperparameters, while M2S parameters are manually fine-tuned to optimize the F1-score in each scenario (Appendix 2). We assess the performance of GM, GMT, metabCombiner, and M2S across 20 randomly generated dataset pairs in terms of their precision (fraction of true matches among the detected matches) and recall/sensitivity (fraction of true matches detected) averaged across 20 dataset pairs.

To investigate how the number of shared features affects dataset alignment, we generate pairs of LC-MS datasets with low, medium, and high feature overlap (25%, 50%, and 75%), while maintaining a medium noise level (Methods). Here we find that GM and GMT generally outperform existing alignment methods, with a recall above 0.95 while metabCombiner and M2S tend to be less sensitive (Fig. 3b). All methods drop in precision as the feature overlap is decreased, with GM and GMT still maintaining an average precision above 0.8.

Next we evaluate all four methods at low, moderate, and high noise levels for pairs of datasets with 50% overlap in their features (Methods). Our results show that GMT, GM, and M2S maintain an average recall above 0.89, while metabCombiner's recall drops below 0.6 for high noise. At large noise levels, RT drift estimation becomes more challenging, leading to a higher rate of false matches between metabolites (lower precision) for all four methods (Fig. 3 - figure supplement 1). Nevertheless, GMT obtains a high average precision and recall of 0.86 and 0.92 respectively.

A notable difference between GM, metabCombiner, and M2S lies in their use of feature intensities. MetabCombiner expects that the mean feature intensity rankings are identical across studies, while M2S assumes that shared features have similar median intensities. In contrast, GM uses both the mean feature intensities and their variances and covariances. In practice, differences in experimental assays or study populations can lead to greater variation in feature intensities, making matchings based on these statistics less reliable. Centering and scaling the feature intensities to unit variance avoids potential biases arising from inconsistent feature intensity magnitudes, but preserves correlations that GM leverages.

Exploring this further, we test how sensitive all four methods are to centering and scaling of feature intensities. MetabCombiner and M2S are tuned using the same methodology as for non-centered and non-scaled data. For M2S, we match features solely based on their m/z and RT. In this experiment (Figure 3 - figure supplement 2), the absence of intensity magnitude information significantly affects metabCombiner's performance and, to a lesser extent, M2S. GM and GMT still obtain accurate matchings, due to their use of correlation structures which are preserved under centering and scaling.

Application to EPIC data

Next, we apply GM, metabCombiner and M2S to align datasets from the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort, a prospective study conducted across 23 European centers. EPIC comprises more than 500,000 participants who provided blood samples at

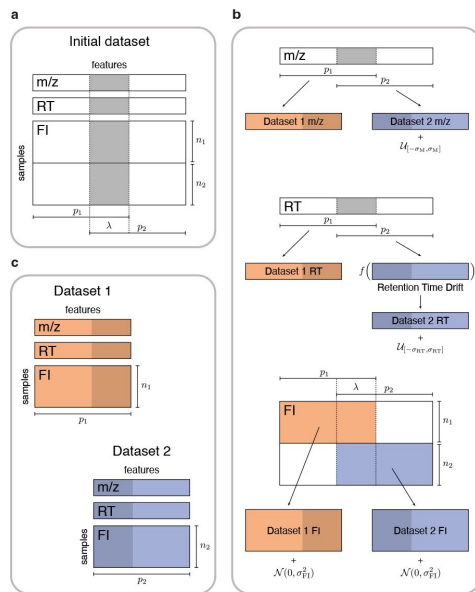


Figure 2.

Simulated data for retention time drift alignment methods.

(a) In the MS/MS dataset for the EXPOSOME COMPACT RT, a dipeptide library of metabolites identified by Li et al. (2019) were used. We randomly split it into two parts $n_1 = 250$ and $n_2 = 250$ respectively and set the retention time of two peaks p_1 and p_2 where $p_2 - p_1 = 0.25, 0.5, 0.75$ (Methods). Retention time drift was applied to the data set 25%, 50%, or 75% of the original data set.

recruitment [Riboli et al. \(2002\)](#). Untargeted metabolomics data were successively acquired in several studies nested within the full cohort.

In the present work, we use LC-MS data from the EPIC cross-sectional (CS) study [Slimani et al. \(2003\)](#) and two matched case-control studies nested within EPIC, on hepatocellular carcinoma (HCC) [Stepien et al. \(2016\)](#), [2021](#) and pancreatic cancer (PC) [Gasull et al. \(2019\)](#). LC-MS untargeted metabolomic data were acquired at the International Agency for Research on Cancer, making use of the same platform and methodology (Methods). The number of samples and features in each study is displayed in Fig. 4a.

[Lofffield et al. \(2021\)](#) previously matched features from the CS, HCC, and PC studies in EPIC for alcohol biomarker discovery. The authors first identified 205 features (163 in positive and 42 in negative mode) associated with alcohol intake in the CS study. These features were then manually matched by an expert to features in both the HCC and PC studies (Methods). In our analysis, we use these features as a validation set and compare each method's matchings to the expert manual matchings on this subset. Due to the imbalance between the number of positive and negative mode features in the validation subset, our main analysis focuses on the alignment results of CS with HCC and CS with PC in positive mode. We delegate the matching results between the negative mode studies to [Appendix 4](#).

In this section, we use the same settings for GM as in our simulation study, and do not apply an additional thresholding step. The parameters of metabCombiner and M2S are calibrated using the validation subset as prior knowledge ([Appendix 2](#)).

Preliminary analysis of the validation subset reveals inconsistencies in the mean feature intensities (Figure 4 - figure supplement 1), but Figure 4b shows that on centered and scaled data, the 90 expert matched features shared between the CS and HCC studies have similar correlation structures. Hence, to avoid potential errors we center and scale the feature intensities which improves the performance of all three methods tested below ([Appendix 4](#), [Appendix 4 - Table 1](#)).

Hepatocellular carcinoma

Here we analyze the quality of the matchings obtained by GM, M2S, and metabCombiner between the CS and HCC datasets in positive mode. Both GM and M2S identify approximately 1000 shared features while metabCombiner finds a smaller number of about 700 shared features. We refer the reader to Figure 4- figure supplement 2a for the precise matched feature sizes and details on the agreement between the feature matchings of all three methods.

We evaluate the performance of metabCombiner, M2S, and GM on the validation subset in positive mode (Fig. 4c), which consist of 90 features from the CS study manually matched to features from the HCC study and 73 features specific to the CS study. MetabCombiner demonstrates precise matching but lacks sensitivity. M2S's precision and recall are comparable with GM, in contrast to its performance on simulated data. This can be attributed to the RT drift shape between the CS and HCC studies ([Appendix 2](#)), which is estimated to be close to linear (Figure 4 - figure supplement 3). Because the parameters of M2S are fine-tuned in the validation subset, it is able to learn this linear drift and apply tight RT thresholds to achieve accurate matchings. In contrast to metabCombiner and M2S, the GM algorithm is not given any prior knowledge of the validation subset, and nevertheless demonstrates the highest precision and recall rates of the three methods (Fig. 4c). Figure 4b shows how GM recovers the majority of the expert matched pairs by leveraging the shared correlations.

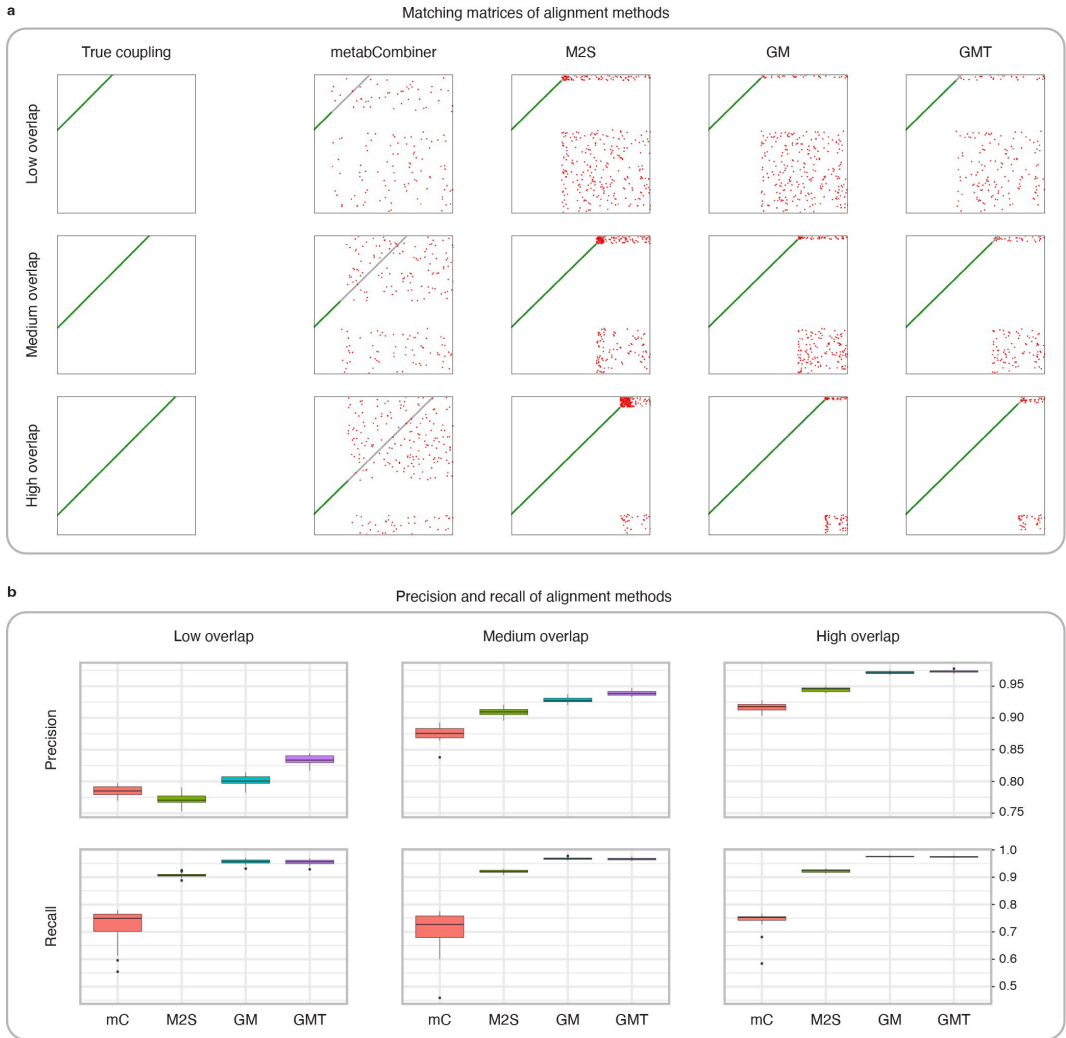


Figure 3.

Comparison of MetabCombiner, M2S, and simulated data.

(a) Ground-truth matchings, and simulated data for M2S, GM, and GMT. Pairs of foreground and background (medium and high), with a medium noise level. (b) Precision and recall of alignment methods (false positive rate and false negative rate) for each method. The y-axis represents precision and recall. The x-axis represents the overlap level (low, medium, and high) with a medium noise level.

P a n c r e a t i c c a n c e r

Matching features between the CS and PC studies in positive mode, GM and M2S identify approximately 1000 common features, while metabCombiner detects approximately 600 matches (Figure 4- figure supplement 2b [↗](#)). We examine the performance of all three methods on the validation subset consisting of 66 manually matched features between CS and PC along with 97 features specific to the CS study. As before, GM and M2S have high recall while the recall of metabCombiner is less than 0.5.

A decrease in precision is observed for both GM and M2S compared to the previous CS-HCC matchings. We therefore manually inspect the false positive matches; the set of CS features matched by the method to the PC study but explicitly examined and left unmatched in the expert manual matching. Assessing the GM results, we identify 7 false positive feature matches. Upon secondary inspection, 3 pairs are revealed as correct matches that were not initially identified in the expert matching. M2S finds 11 false positive matches which include the 7 false positives recovered by GM. Manual examination of the 4 remaining pairs reveals 2 clear mismatches. These results highlight the advantage of using automated methods for data alignment, as both GM and M2S detect correct matches that were not identified by experts, with GM being more precise than M2S.

I l l u s t r a t i o n f o r a l c o h o l b i o m a r k e r

Lofffield et al. [Lofffield et al. \(2021\) \[↗\]\(#\)](#) identified biomarkers of habitual alcohol intake by first performing a discovery step, where they examined the relationship between alcohol intake and metabolic features in the CS study. They then manually matched the significant features in CS to features from the HCC and PC studies, and repeated the analysis with samples from the HCC and PC studies to determine whether the association with alcohol intake persisted. This led to the identification of 10 features possibly associated with alcohol intake (Fig. 5a [↗](#)).

To extend this analysis and illustrate the benefit of GM automatic matching for biomarker discovery, we use GM to pool features from the CS, HCC, and PC studies, and examine the relationship between metabolic features and alcohol intake in the pooled study (Methods and Fig. 5b [↗](#)).

Applying an FDR correction on the pooled study, we identify 243 features associated with alcohol intake, including 185 features consistent with the discovery step of Lofffield et al. [Lofffield et al. \(2021\) \[↗\]\(#\)](#), and 55 newly discovered features (Fig. 5c [↗](#)). Using the more stringent Bonferroni correction on the pooled data, we identify 36 features shared by all three studies that are significantly associated with alcohol intake. These features include all 10 features identified in Lofffield et al. (Fig. 5c [↗](#)). These findings highlight the potential benefits of using GM automatic matching for biomarker discovery in untargeted metabolomics data. Additional information regarding the methodology and findings of our GM and Lofffield et al. analyses can be found in Methods and Appendix 4 [↗](#).

D i s c u s s i o n

LC-MS metabolomics has emerged as an increasingly powerful tool for biological and biomedical research, offering promising opportunities for epidemiological and clinical investigations. However, integrating data from different sources remains challenging. To address this issue, we introduce GromovMatcher, a method based on optimal transport that automatically aligns LC-MS

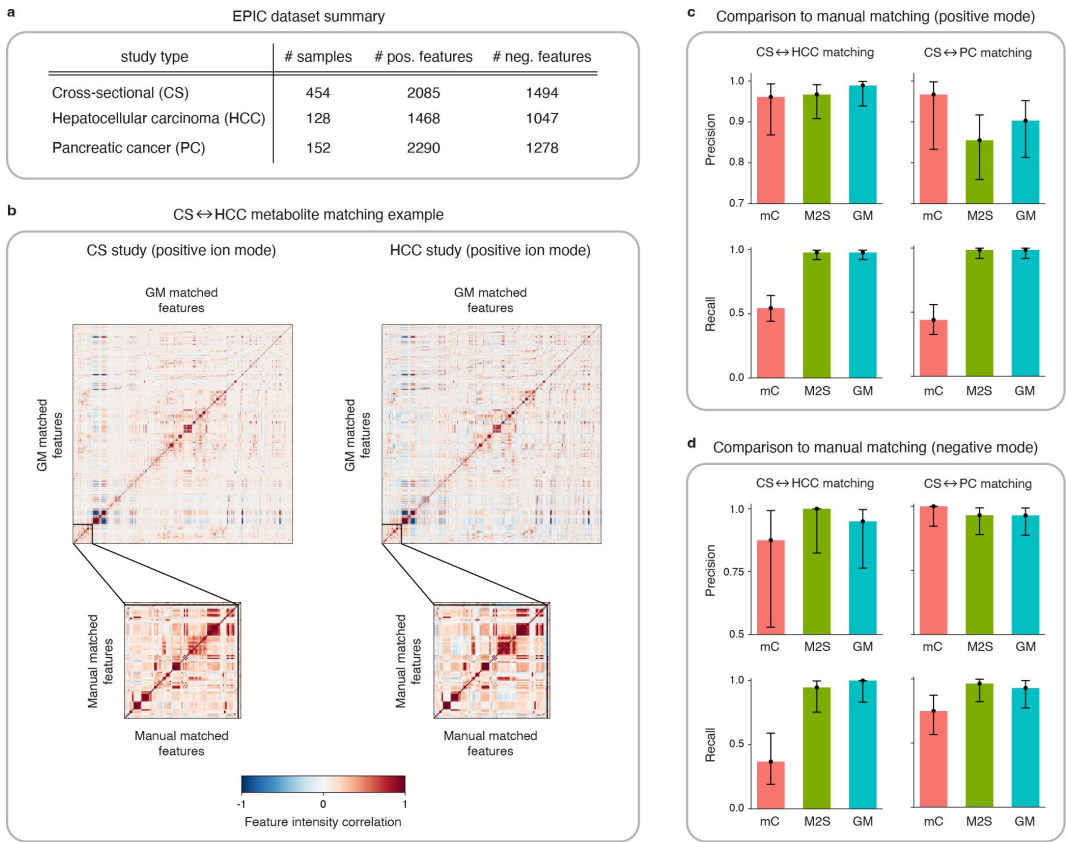


Figure 4.

Application of MetComber and comparison methods on EPIC dataset.

(a) Dimensions of the studies used. For each study, the scaling was performed with the hepatic ionization mode (HCC) and the pancreatic ionization mode (PC). The manual matching was performed using the manual matching method (GM) matching between the two datasets. Experts manually matched features in the field and the two datasets. The two datasets have similar features (bottom two matrices). The GM method between the two datasets showed similar correlation (top two matrices). What is the reason for this? The number of features in the sets of features. Performance of metabolite combiner (mC), M2S and GM in positive and negative mode on a validation subset of the dataset. The 95% confidence intervals were computed using model fitting. The performance of mC, M2S and GM in the negative mode was measured on a validation subset of the dataset. The 95% confidence intervals were modified. We list the results in Table 3 for comparison, and confidence intervals for positive and negative mode.

| Study | Manual matches found in positive mode | Manual matches found in negative mode |
|--------------------------------|---------------------------------------|---------------------------------------|
| Hepatocellular carcinoma (HCC) | 90 | 19 |
| Pancreatic cancer (PC) | 66 | 28 |

Table 1

Results from the manual matching conducted by the field et al.

Features from the CS study (6 positive features, 1 negative feature) were investigated for matches in the HCC studies.

| Method | CS ↔ HCC | | CS ↔ PC | |
|----------------------|----------------------|----------------------|----------------------|----------------------|
| | Precision | Recall | Precision | Recall |
| GromovMatcher | 0.989 (0.939, 0.999) | 0.978 (0.923, 0.996) | 0.903 (0.813, 0.952) | 0.985 (0.919, 0.999) |
| M2S | 0.967 (0.908, 0.991) | 0.978 (0.923, 0.996) | 0.855 (0.759, 0.917) | 0.985 (0.919, 0.999) |
| metabCombiner | 0.961 (0.868, 0.993) | 0.544 (0.442, 0.643) | 0.967 (0.833, 0.998) | 0.439 (0.326, 0.559) |

Table 2

Precision-recall on the EPIC validation subset in positive mode
 95% confidence intervals were computed using modified Wilson's score test (Agresti & Coull (1998))

| Method | CS ↔ HCC | | CS ↔ PC | |
|----------------------|----------------------|----------------------|----------------------|----------------------|
| | Precision | Recall | Precision | Recall |
| GromovMatcher | 0.950 (0.764, 0.997) | 1.000 (0.832, 1.000) | 0.929 (0.774, 0.987) | 0.929 (0.774, 0.987) |
| M2S | 1.000 (0.824, 1.000) | 0.947 (0.754, 0.997) | 0.931 (0.780, 0.988) | 0.964 (0.823, 0.998) |
| metabCombiner | 0.875 (0.529, 0.993) | 0.368 (0.191, 0.590) | 1.000 (0.845, 1.000) | 0.750 (0.566, 0.873) |

Table 3

Precision-recall on the EPIC validation subset in positive mode
 95% confidence intervals were computed using modified Wilson's score test (Agresti & Coull (1998))

data from pairs of studies. Our method exhibits superior performance on both simulated and real data when compared to existing approaches. Additionally, it presents a user-friendly interface with few hyperparameters.

While GromovMatcher is robust to noise and variations in data, it may face limitations when aligning LC-MS studies from populations with different characteristics, where the correlation structures between features may be inconsistent across studies. In this case, the base assumption of GromovMatcher can be relaxed by focusing on subsamples with similar characteristics, as exemplified in a recent study [Gomari et al. \(2022\)](#).

A current limitation is that GromovMatcher does not account for more than two datasets simultaneously, although this can be overcome by aligning multiple studies to a chosen reference dataset, as demonstrated in our biomarker experiments. The extension of Gromov-Wasserstein to multiple distributions [Beier et al. \(2022\)](#) is another promising approach for generalizing GromovMatcher to multiple dataset alignment. Further improvements can be made by incorporating existing knowledge about the studies being matched, such as known shared features, samples in common, or MS/MS data.

The results obtained from GromovMatcher are highly promising, opening the door for various analyses of metabolomic datasets acquired in different experimental laboratories. Here we demonstrated the potential of GromovMatcher in expediting the combination and meta-analysis of data for biomarker and metabolic signature discovery. The matchings learned by GromovMatcher also allow for comparison between experimental protocols by assessing the drift in m/z , RT, and feature intensities across studies. Finally, inter-institutional annotation efforts can directly benefit from incorporating this method to transfer annotations between aligned datasets. Bridging the gap between otherwise incompatible LC-MS data, GromovMatcher enables seamless comparison of untargeted metabolomics experiments.

Methods and Materials

GromovMatcher methodology

GromovMatcher accepts as input two feature tables from separate LC-MS untargeted metabolomics studies. Each feature table for dataset 1 and dataset 2 consists of n_1 , n_2 biospecimen samples respectively and p_1 , p_2 metabolic features respectively detected in the study. Features in dataset 1 are given the label fx_i for $i = 1, \dots, p_1$. Every feature is characterized by a mass-to-charge ratio (m/z) denoted by m_i^x , a retention time (RT) denoted by RT_i^x , and a vector of intensities across all samples written as $X_i \in \mathbb{R}^{n_1}$. Similarly, features in dataset 2 are labeled as fy_j for $j = 1, \dots, p_2$ and are characterized by their m/z m_j^y , retention time RT_j^y , and a vector of intensities across all samples $Y_j \in \mathbb{R}^{n_2}$.

Our goal is to identify pairs of indexes (i, j) with $i \in \{1, \dots, p_1\}$ and $j \in \{1, \dots, p_2\}$, such that fx_i and fy_j correspond to the same metabolic feature. More formally, we aim to identify a matching matrix $M \in \{0, 1\}^{p_1 \times p_2}$ such that $M_{ij} = 1$ if fx_i and fy_j correspond to the same feature, hereafter referred to as matched features. Otherwise, we set $M_{ij} = 0$.

Because the m/z and RT values of metabolomic features are often noisy and subject to experimental bias, our matching algorithm leverages metabolite feature intensities X_i, Y_j to produce accurate dataset alignments. The GromovMatcher method is based on the idea that signal

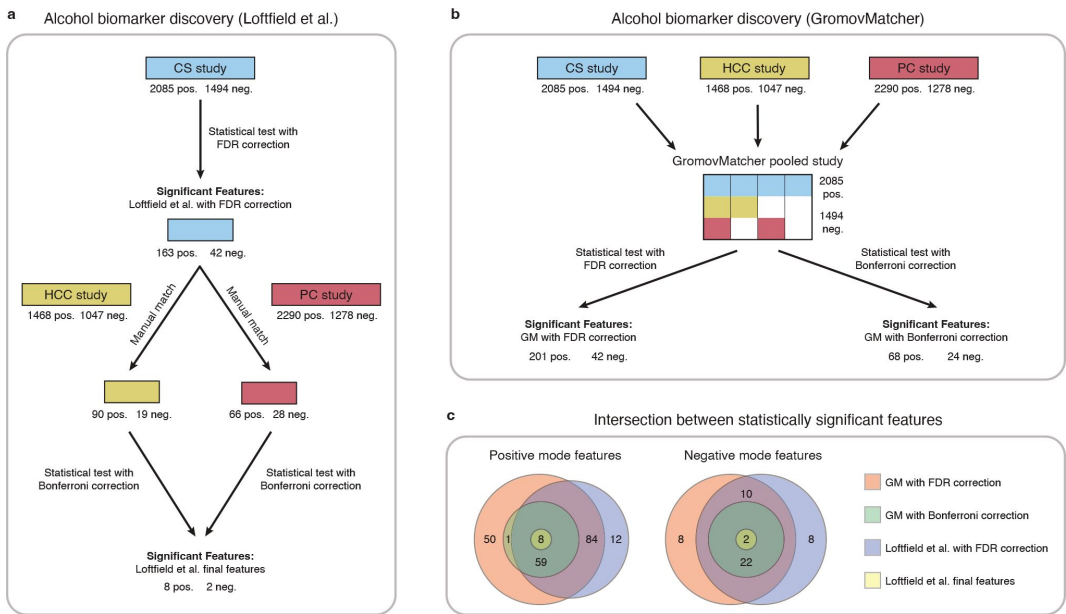


Figure 5.

Comparison of GromovMatcher and Lofffield et al. (2021) analysis for alcohol biomarker discovery.

(a) Lofffield study workflow. In this comparison between Lofffield et al. and GromovMatcher, the analysis was repeated using GromovMatcher on the HCC and PC studies. After a statistical test, we identified 163 positive and 42 negative features in the CS study. These significant features were then compared with the HCC and PC studies. After a statistical test, we identified 8 positive and 2 negative features in the CS study. (b) GromovMatcher analysis of the HCC and PC studies. After a statistical test, we identified 201 positive and 42 negative features in the pooled study. After a statistical test, we identified 68 positive and 24 negative features in the pooled study. (c) Comparison of the features found to be associated with alcohol consumption (positive and negative mode) found to be associated with alcohol consumption.

intensities of the same metabolites measured in two different studies should exhibit similar correlation structures, in addition to having compatible m/z and RT values. Here we define the Pearson correlation for vectors $u, v \in \mathbb{R}^n$ as

$$\text{corr}(u, v) = \frac{\langle u - \bar{u}, v - \bar{v} \rangle}{\|u - \bar{u}\| \|v - \bar{v}\|} \quad (1)$$

where we define

$$\bar{u} = \frac{1}{n} \sum_{i=1}^n u_i, \quad \|u\| = \sqrt{\sum_{i=1}^n u_i^2}, \quad \langle u, v \rangle = \sum_{i=1}^n u_i v_i \quad (2)$$

as the mean value, Euclidean norm and inner product respectively. If measurements X_i, Y_j correspond to the same underlying feature, and similarly, measurements X_k, Y_l share the same an underlying feature, we expect that

$$\text{corr}(X_i, X_k) \approx \text{corr}(Y_j, Y_l). \quad (3)$$

This idea that the feature intensities of shared metabolites have the same correlation structure in both datasets also holds more generally for distances, under a suitable choice of distance. For example, the correlation coefficient $\text{corr}(u, v)$ can be turned into a dissimilarity metric by defining

$$d^{\text{cos}}(u, v) = \sqrt{1 - \text{corr}(u, v)} \quad (4)$$

commonly referred to as the cosine distance. Preservation of feature intensity correlations then trivially amounts to the preservation of cosine distances.

Another classical notion of distance between vectors $u, v \in \mathbb{R}^n$ is the normalized Euclidean distance

$$d^{\text{euc}}(u, v) = \frac{1}{\sqrt{n}} \|u - v\| = \sqrt{\frac{1}{n} \sum_{i=1}^n (u_i - v_i)^2} \quad (5)$$

which is equal to the cosine distance (up to constants) when the vectors u, v are centered and scaled to have zero mean and a standard deviation of one. The Euclidean distance depends on the magnitude or mean intensity of metabolic features, and hence is a useful metric for matching metabolites as long as these mean feature intensities are reliably collected.

To summarize, the main tenant of GromovMatcher is that if measurements X_i, Y_j correspond to the same feature and X_k, Y_l correspond to the same feature, then for suitably chosen distances $d_x : \mathbb{R}^{n_1} \times \mathbb{R}^{n_1} \rightarrow \mathbb{R}$ and $d_y : \mathbb{R}^{n_2} \times \mathbb{R}^{n_2} \rightarrow \mathbb{R}$, these distances are preserved

$$d_x(X_i, X_k) \approx d_y(Y_j, Y_l) \quad (6)$$

across both datasets. In this paper, the distances d_x, d_y are taken to be the normalized Euclidean distances in (5). We take care to specify those experiments where the metabolic features X and Y are centered and scaled. In these cases, implicitly the Euclidean distance between normalized feature vectors becomes the cosine distance (4) between the original (unnormalized) feature vectors.

Unbalanced Gromov-Wasserstein

The goal of GromovMatcher is to learn a matching matrix $M \in \{0, 1\}^{p_1 \times p_2}$ that gives an alignment between a subset of metabolites in both datasets. However, searching over the combinatorially large set of binary matrices would be an inefficient approach for dataset alignment. The mathematical framework of optimal transport [Peyré et al. \(2019\)](#) instead enlarges this space of binary matrices to the set of coupling matrices with real nonnegative entries $\Pi \in \mathbb{R}_+^{p_1 \times p_2}$. The entries Π_{ij} with large weights indicate that feature fx_i in dataset 1 and feature fy_j in dataset 2 are a likely match. Taking inspiration from (6), we minimize the following objective function

$$\mathcal{E}(\Pi) = \sum_{i,k=1}^{p_1} \sum_{j,l=1}^{p_2} \Pi_{ij} \Pi_{kl} |d_x(X_i, X_k) - d_y(Y_j, Y_l)| \quad (7)$$

to estimate the coupling matrix Π .

A standard approach is to optimize this objective over all coupling matrices Π under exact marginal constraints $\Pi \mathbf{1}_{p_2} = \frac{1}{p_1} \mathbf{1}_{p_1}$, $\Pi^T \mathbf{1}_{p_1} = \frac{1}{p_2} \mathbf{1}_{p_2}$. Here we define $\mathbf{1}_n$ is the ones vector of length n , and $\Pi_{\cdot j} = \Pi \mathbf{1}_{p_2}$, $\Pi_{i \cdot} = \Pi^T \mathbf{1}_{p_1}$ denote the column and row sums of the coupling matrix. Objective (7) under these exact marginal constraints defines a distance between the two sets of metabolic feature vectors $\{X_i\}_{i=1}^{p_1}$, $\{Y_j\}_{j=1}^{p_2}$ known as the Gromov-Wasserstein distance [Mémoli \(2011\)](#), a generalization of optimal transport to metric spaces. Note that for pairs X_i, Y_j and X_k, Y_l , for which $d_x(X_i, X_k) \approx d_y(Y_j, Y_l)$, the entries Π_{ij}, Π_{kl} are penalized less and hence matches between features fx_i, fy_j and features fx_k, fy_l are more favored. In our optimization, we avoid enforcing exact marginal constraints on the marginal distributions $\Pi_{\cdot j}$ and $\Pi_{i \cdot}$ of our coupling matrix as this would enforce that all metabolites in both datasets are matched ([Appendix 1](#)). However, without any marginal constraints on the coupling Π , the objective function (7) is trivially minimized by $\Pi = 0$, leaving all metabolites in both datasets unmatched.

To account for this, we follow the ideas of unbalanced Gromov-Wasserstein (UGW) [Sejourne et al. \(2021\)](#) and add three regularization terms to our objective

$$\begin{aligned} \mathcal{L}_{\rho, \varepsilon}(\Pi) = \mathcal{E}(\Pi) + \rho D_{\text{KL}}(\Pi_{\cdot j} \otimes \Pi_{i \cdot}, a \otimes a) \\ + \rho D_{\text{KL}}(\Pi_{\cdot j} \otimes \Pi_{i \cdot}, b \otimes b) \\ + \varepsilon D_{\text{KL}}(\Pi \otimes \Pi, (a \otimes b)^{\otimes 2}) \end{aligned} \quad (8)$$

where $\rho, \varepsilon > 0$ and we define $a = \mathbf{1}_{p_1}$, $b = \mathbf{1}_{p_2}$. Here \otimes denotes the Kronecker product. We define D_{KL} as the Kullback-Leibler (KL) divergence between two discrete distributions $\mu, \nu \in \mathbb{R}_+^p$ by

$$D_{\text{KL}}(\mu, \nu) = \sum_{i=1}^p \mu_i \ln \left(\frac{\mu_i}{\nu_i} \right) - \sum_{i=1}^p \mu_i + \sum_{i=1}^p \nu_i \quad (9)$$

which measures the closeness of probability distributions.

The first two regularization terms in (8) enforce that the row sums and column sums of the coupling matrix Π do not deviate too much from a uniform distribution, leading our optimization to match as many metabolic features as possible. The magnitude of the regularizer ρ roughly enforces the fraction of metabolites in both datasets that are matched where large ρ implies most metabolites are matched across datasets. The final regularization term ε in (8) controls the smoothness (entropy) of the coupling matrix Π where larger values of ε encourage to put uniform weights on many of its entries, leading to less precision in the metabolite matches. However, increasing ε also leads to better numerical stability and a significant speedup of the

alternating minimization algorithm used to optimize the objective function (Appendix 1). In our implementation, we set ϵ and δ to the lowest possible values under which our optimization converges, with $\epsilon = 0.05$ and $\delta = 0.005$.

Our full optimization problem can now be written as

$$\text{UGW}_{\rho, \epsilon} = \min_{\Pi \in \mathbb{R}_+^{p_1 \times p_2}} \mathcal{L}_{\rho, \epsilon}(\Pi). \quad (10)$$

The UGW objective function is optimized through alternating minimization based on the code of Sejourne et al. (2021) using the unbalanced Sinkhorn algorithm Séjourné et al. (2019) from optimal transport (Appendix 1).

Constraints

Matched metabolic features must have compatible m/z so we enforce that $|m_i^x - m_j^y| \leq m_{\text{gap}}$ where m_{gap} is a user-specified threshold. Based on prior literature Lofftfield et al. (2021); Hsu et al. (2019); Climaco Pinto et al. (2022); Habra et al. (2021); Chen et al. (2021) we set $m_{\text{gap}} = 0.01$ ppm. Note that m_{gap} is not explicitly used in (10) but is rather enforced in each iteration of our alternating minimization algorithm for the UGW objective (Appendix 1).

Unlike the m/z ratios discussed above, RTs often exhibit a non-linear deviation (drift) between studies so we cannot enforce compatibility of RTs directly in our optimization. Instead, in the following step of our pipeline we ensure matched metabolite pairs have compatible RTs by estimating the drift function and subsequently using it to filter out metabolite matches whose RT values are inconsistent with the estimated drift.

Estimation of drift and filtering

Estimating the drift between RTs of two studies is a crucial step in assessing the validity of metabolite matches and discarding those pairs which are incompatible with the estimated drift.

Let $\hat{\Pi} \in \mathbb{R}_+^{p_1 \times p_2}$ be the minimizer of (10) obtained after optimization. We seek to estimate the RT drift function $f: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ which relates the retention times of matched features between the two studies. Namely, if feature fx_i and feature fy_j correspond to the same metabolic feature, then we must have that $RT_j^y \approx f(RT_i^x)$.

We propose to learn the drift f through the weighted spline regression

$$\min_{f \in \mathcal{B}_{n,k}} \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \hat{\Pi}_{ij} |f(RT_i^x) - RT_j^y| \quad (11)$$

where $\mathcal{B}_{n,k}$ is the set of n -order B-splines with k knots. All pairs (RT_i^x, RT_j^y) in objective (11) are weighted by the coefficients of $\hat{\Pi}$ so that larger weights are given to pairs identified with high confidence in the first step of our procedure. The order of the B-splines was set to $n = 3$ by default, while the number of knots k was selected by 10-fold cross-validation.

Pairs identified as incompatible with the estimated RT drift are then discarded from the coupling matrix. To do this, we first take the estimated RT drift \hat{f} , and the set of pairs $S = \{(i, j) : \hat{\Pi}_{ij} \neq 0\}$ recovered in $\hat{\Pi}$. We then define the residual associated with $(i, j) \in S$ as

$$r_f(i, j) = |\hat{f}(RT_i^x) - RT_j^y|. \quad (12)$$

The 95% prediction interval and the median absolute deviation (MAD) of these residuals are given by

$$\begin{aligned} \text{PI} &= 1.96 \times \text{std}(\{r_f(i, j), (i, j) \in S\}) \\ \text{MAD} &= \text{median}(\{|r_f(i, j) - \mu_r|, (i, j) \in S\}) \\ \mu_r &= \text{median}(\{|r_f(i, j)|, (i, j) \in S\}) \end{aligned} \quad (13)$$

where $|S|$ is the size of S and the functions std and median denote the standard deviation and median respectively. Similar to the approach in [Climaco Pinto et al. \(2022\)](#), we create a new filtered coupling matrix $\hat{\Pi} \in \mathbb{R}_+^{p_1 \times p_1}$ given by

$$\hat{\Pi}_{ij} = \begin{cases} \tilde{\Pi}_{ij} & \text{if } r_f(i, j) < \mu_r + r_{\text{thresh}} \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

where r_{thresh} is a given filtering threshold. Following [Habra et al. \(2021\)](#), the estimation and outlier detection step can be repeated for multiple iterations, to remove pairs that deviate significantly from the estimated drift and improve the robustness of the drift estimation. In our main algorithm, we use two preliminary iterations where estimate the RT drift and discard outliers outside of the 95% prediction interval by setting $r_{\text{thresh}} = \text{PI}$. We then re-estimate the drift and perform a final filtering step with the more stringent MAD by setting $r_{\text{thresh}} = 2 \times \text{MAD}$.

At this stage, it is possible for $\hat{\Pi}$ to still contain coefficients of very small magnitude. As an optional postprocessing step, we discard these coefficients by setting all entries smaller than $\epsilon_{\max}(\hat{\Pi})$ to zero, for some user-defined $\epsilon \in [0, 1]$. Lastly, a feature from either study could have multiple possible matches, since $\hat{\Pi}$ can have more than one non-zero coefficient per row or column. Although reporting multiple matches can be helpful in an exploratory context, for the sake of simplicity in our analysis, the final output of GromovMatcher returns a one-to-one matching, as we only keep those metabolite pairs (i, j) where the entry $\hat{\Pi}_{ij}$ is largest in its corresponding row and column. All nonzero entries of $\hat{\Pi}$ which do not satisfy this criterion are set to zero. Finally, we convert $\hat{\Pi}$ into a binary matching matrix $M \in \{0, 1\}^{p_1 \times p_2}$ with ones in place of its nonzero entries and this final output is returned to the user.

As a naming convention, we use the abbreviation GM for our GromovMatcher method, and use the abbreviation GMT when running GromovMatcher with the optional ϵ -thresholding step with $\epsilon = 0.3$.

Metrics for dataset alignment

Every alignment method studied in this paper returns a binary partial matching matrix $M \in \{0, 1\}^{p_1 \times p_2}$ which has at most one nonzero entry in each row and column. Specifically, $M_{ij} = 1$ if metabolic features i and j in both datasets correspond to each other and $M_{ij} = 0$ otherwise. In our simulated experiments, we compare the partial matching M to a known ground-truth partial matching matrix $M^* \in \{0, 1\}^{p_1 \times p_2}$.

To do this, we first compute the number of true positives, false positives, true negatives, and false negatives as

$$\begin{aligned}
 \text{TP} &= \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \mathbf{1}_{M_{ij}=1} \mathbf{1}_{M_{ij}^*=1} \\
 \text{FP} &= \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \mathbf{1}_{M_{ij}=1} \mathbf{1}_{M_{ij}^*=0} \\
 \text{TN} &= \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \mathbf{1}_{M_{ij}=0} \mathbf{1}_{M_{ij}^*=0} \\
 \text{FN} &= \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \mathbf{1}_{M_{ij}=0} \mathbf{1}_{M_{ij}^*=1}
 \end{aligned} \tag{15}$$

where 1 denotes the indicator function. Then we use these values to compute the precision and recall as

$$\begin{aligned}
 \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\
 \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}.
 \end{aligned} \tag{16}$$

Precision measures the fraction of correctly found matches out of all discovered metabolite matches, while recall, also known as sensitivity, measures the fraction of correctly matched pairs out of all truly matched pairs. These two statistics can be summarized into one metric called the F1-score by taking their harmonic mean

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

These three metrics, precision, recall, and the F1-score, are used throughout the paper to assess the performance of dataset alignment methods, both on simulated data where the ground-truth matching is known, and on the validation subset in EPIC, using results from the manual examination as the ground-truth benchmark.

Validation on simulated data

To assess the performance of GromovMatcher and compare it to existing dataset alignment methods, we simulate realistic pairs of untargeted metabolomics feature with known ground-truth matchings. This allows us to analyze the dependence of alignment methods on the number of shared metabolites, dataset noise level, and feature intensity centering and scaling.

Dataset generator

Our pairs of synthetic feature tables are generated from one real untargeted metabolomics study of 500 newborns within the EXPOSOMICS project, which uses reversed phase liquid chromatography-quadrupole time-of-flight mass spectrometry (UHPLC-QTOF-MS) system in positive ion mode [Alfano et al. \(2020\)](#). The original dataset is first preprocessed following the procedure detailed in [Alfano et al. \(2020\)](#), resulting in $p = 4,712$ features measured in $n = 499$ samples available for subsequent analysis. Features and samples from the original study are then divided into two feature tables of respective size (n_1, p_1) and (n_2, p_2) , with $n_1 + n_2 = n$ and $p_1, p_2 \leq p$. In order to do this, $n_1 = \lfloor n/2 \rfloor$ randomly chosen samples from the original study are

placed into dataset 1 and the remaining $n_2 = \lceil n/2 \rceil$ samples from the original study are placed into dataset 2. Here $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ denote integer floor and ceiling functions. The features of the original study are randomly assigned to dataset 1, dataset 2, or both, allowing the resulting studies to have both common and study-specific features (Fig. 2). Specifically, for a fixed overlap parameter $\rho \in [0,1]$, we assign a random subset of ρp features into both dataset 1 and dataset 2 while the remaining $(1 - \rho)p$ features are divided equally between the two studies such that $p_1 = p_2$. We choose $\rho \in \{0.25, 0.5, 0.75\}$ corresponding to low, medium and high overlap. For more detailed information on how the dataset split is performed and for additional validation experiments with unbalanced dataset splits (e.g. $n_1 \neq n_2, p_1 \neq p_2$) we refer the reader to Appendix 3.

After generating a pair of studies, random noise is added to the m/z, RT and intensity levels of features in dataset 2 to mimic variations in data acquisition across two different experiments. The noise added to each m/z value in study 2 is sampled from a uniform distribution on the interval $[-M, M]$ with $M = 0.01$ (Climaco Pinto et al. (2022)). The RTs of dataset 2 are first deviated by the function $f(x) = 1.1x + 1.3 \sin(1.2\sqrt{x})$, corresponding to a systematic inter-dataset drift (Habra et al. (2021); Climaco Pinto et al. (2022); Brunius et al. (2016)). A uniformly distributed noise on the interval $[-RT, RT]$ is added to the deviated RTs of dataset 2, with $RT \in \{0.2, 0.5, 1\}$ (in minutes) corresponding to low, moderate and high variations (Climaco Pinto et al. (2022); Habra et al. (2021); Vaughan et al. (2012)). Finally, we add a Gaussian noise $\mathcal{N}(0, \sigma_F^2)$ to the feature intensities of both studies where σ_F is the scalar variance of the noise. This noise perturbs the correlation matrices of dataset 1 and dataset 2, making matching based on feature intensity correlations more challenging. We vary σ_F over the set of values $\{0.1, 0.5, 1\}$.

Given this data generation process, we test the performance of the four alignment methods (M2S, metabCombiner, GM, and GMT) under the parameter settings described below.

Dependence on overlap

We first assess how the performance of the four methods is affected by the number of metabolic features shared in both datasets. For each value of $\rho \in \{0.25, 0.5, 0.75\}$ (low, medium and high overlap), we randomly generate 20 pairs of datasets with noise on the m/z, RT and feature intensities set to $M = 0.01$, $RT = 0.5$, $\sigma_F = 0.5$. The precision and recall of each method at low, medium, and high overlap is recorded for each of the repetitions.

Noise robustness

Next we test the robustness to noise of each method by fixing the metabolite overlap fraction at $\rho = 0.5$ and generating 20 random pairs of datasets at low ($RT = 0.2$, $\sigma_F = 0.1$), medium ($RT = 0.5$, $\sigma_F = 0.5$), and high ($RT = 1$, $\sigma_F = 1$) noise levels. Similarly, the precision and recall of each method is saved for each noise level across the 20 repetitions.

Feature intensity centering and scaling

In order to test how all four methods are affected when the mean feature intensities and variance are not comparable across studies, we assess their performance when the feature intensities in both studies are mean centered and standardized to have unit standard deviation across all samples. We again generate 20 random pairs of datasets with medium overlap and medium noise, normalize the feature intensities in each pair of datasets, and compute the precision and recall of each method across the 20 repetitions.

EPIC data

We also evaluate our method on data collected within the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort, an ongoing multicentric prospective study with over 500,000 participants recruited between 1992 and 2000 from 23 centers in 10 European countries, and who

provided blood samples at the inclusion in the study [Riboli et al. \(2002\)](#). In EPIC, untargeted metabolomics data were successively acquired in several studies nested within the full cohort.

In the present work, we use untargeted metabolomics data acquired in three studies nested in EPIC, namely the EPIC cross-sectional (CS) study [Slimani et al. \(2003\)](#) and two matched case-control studies nested within EPIC, on hepatocellular carcinoma (HCC) [Stepien et al. \(2016, 2021\)](#) and pancreatic cancer (PC) [Gasull et al. \(2019\)](#), respectively. All data were acquired at the International Agency for Research on Cancer, making use of the same platform and methodology: UHPLC-QTOF-MS (1290 Binary Liquid chromatography system, 6550 quadrupole time-of-flight mass spectrometer, Agilent Technologies, Santa Clara, CA) using reversed phase chromatography and electrospray ionization in both positive and negative ionization mode.

In a previous analysis aiming at identifying biomarkers of habitual alcohol intake in EPIC, the 205 features associated with alcohol intake in the CS study were manually matched to features in both the HCC and PC studies [Lofffield et al. \(2021\)](#). The results from this manual matching are presented in [Table 1](#). This matching process was based on the proximity of m/z and RT, using a matching tolerance of ± 15 ppm and ± 0.2 min, and on the comparison of the chromatograms of features in a quality control samples from both studies.

Pre processing

In the HCC and PC studies, samples corresponding to participants selected as cases in either study (i.e., participants selected in the study because of a diagnosis of incident HCC or PC) are excluded. Indeed, the metabolic profiles of participants selected as controls are expected to be more comparable across studies than those of cases, especially if certain features are associated with the risk of HCC or PC. Apart from this additional exclusion criterion, the untargeted metabolomics data of each study is pre-processed following the steps described in [Lofffield et al. \(2021\)](#), to eliminate unreliable features and samples, impute missing values and minimize technical variations in the feature intensity levels.

Alcohol biomarker discovery

[Lofffield et al. \(2021\)](#) used the untargeted metabolomics data of the CS, HCC and PC studies in their alcohol biomarker discovery study in EPIC, without being able to automatically match their common features and pool the 3 datasets. Instead, the authors first implemented a discovery step, examining the relationship between alcohol intake and metabolic features measured in the CS study and accounting for multiple testing using a false discovery rate (FDR) correction. This led to the identification of 205 features significantly associated with alcohol intake in the CS study. In order to gauge the robustness of these associations, the authors of [Lofffield et al. \(2021\)](#) then implemented a validation step using data from two independent test sets. The first test set was composed of data from the EPIC HCC and PC studies, while the second was derived from the Finnish Alpha-Tocopherol, Beta-Carotene Cancer Prevention (ATBC) study. The 205 features identified in the discovery step were manually investigated for matches in the EPIC test set, and 67 features were effectively matched to features in the HCC or PC study, or both. The authors then evaluated the association between alcohol intake and those 67 features, applying a more conservative Bonferroni correction to determine whether the association with alcohol intake persisted. This step led to the identification of 10 features associated with alcohol intake ([Extended Data Fig. 5a](#)). The second test set was then used to determine whether those 10 features were also significant in the ATBC population, which was indeed the case.

To conduct a more in-depth investigation of the matchings produced by the GromovMatcher algorithm, we build upon the analysis previously conducted by [Lofffield et al. \(2021\)](#) by exploring potential alcohol biomarkers using a pooled dataset created from the CS, HCC, and PC studies. Our goal is to assess whether pooling the data leads to increased statistical power and allows for the detection of more features associated with alcohol intake. Namely, we

generate the pooled dataset by aligning a chosen reference dataset (CS study) with the HCC and PC studies successively using the GM matchings computed in both positive and negative mode (Methods and Extended Data Fig. 5b [↗](#)). Features that are not detected in either the HCC or PC studies are designated as 'missing' in the final pooled dataset for samples belonging to the respective studies where the feature is not found.

To evaluate the potential relationship between alcohol consumption and pooled metabolic features, we use a methodology akin to that of Loftfield et al. [Loftfield et al. \(2021\) ↗](#). The self-reported alcohol intake data is adjusted for various demographic and lifestyle factors (age, sex, country, body-mass-index, smoking status and intensity, coffee consumption, and study) via the residual method in linear regression models. Feature intensities are also adjusted for technical variables (plate number and position within the plate) via linear mixed effect models. The significance of the association is assessed using correlation coefficients computed from the residuals for both self-reported alcohol intake and feature intensities. P-values are corrected using either false discovery rate (FDR) or Bonferroni correction to account for multiple testing. Corrected p-values less than 5% are considered significant.

A c k n o w l e d g e m e n t s

We thank Jörn Dunkel for helpful advice on our manuscript. We acknowledge the MIT Super-Cloud and Lincoln Laboratory Supercomputing Center [Reuther et al. \(2018\) ↗](#) for providing HPC resources that have contributed to the research results reported within this paper. G.S. acknowledges support through a National Science Foundation Graduate Research Fellowship under Grant No. 1745302. P.R. is supported by NSF grants IIS-1838071, DMS-2022448, and CCF-2106377. We are grateful to the Principal Investigators of each of the EPIC centers for sharing the data for our experimental application.

D a t a a v a i l a b i l i t y

The LC-MS data used to generate our simulated validation experiments is located at the bottom of the "Files" section in <https://www.ebi.ac.uk/metabolights/MTBLS1684/files> [↗](#) under filename 'metabolomics_normalized_data.xlsx'. The EPIC data is not publicly available, but access requests can be submitted to the Steering Committee <https://epic.iarc.fr/access/index.php> [↗](#).

All code for the data preprocessing, figure generation, as well as the GromovMatcher algorithm and its comparison to other methods are available at: <https://github.com/sgstepaniants/GromovMatcher> [↗](#). Instructions and examples for how to run the GromovMatcher method are provided in the Github repository. The metabCombiner implementation written by the original authors was taken from their Github codebase: <https://github.com/hhabra/metabCombiner> [↗](#). The M2S implementation of the original authors was taken from their Github codebase: <https://github.com/rjdossan/M2S> [↗](#).

A u t h o r c o n t r i b u t i o n s

M.B. and G.S. contributed equally and are joint first authors. P.R. and V.V. conceived the project. M.B. and G.S. developed the algorithms as well as performed the comparison to other alignment methods. P.K.R. analyzed and postprocessed the EPIC data and the analysis of our method on this data was performed by M.B. The manuscript was prepared by M.B., G.S., P.K.R., P.R., and V.V.

Materials and Methods

All correspondence and material requests should be addressed to V.V.

IARC disclaimer

Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer/World Health Organization.

Appendix 1

In this paper, we study how to match metabolic features across two datasets where Dataset 1 has p_1 metabolic features measured across n_1 patients and Dataset 2 has p_2 metabolic features measured across n_2 patients. Our goal is to identify pairs of indexes (i, j) with $i \in \{1, \dots, p_1\}$ and $j \in \{1, \dots, p_2\}$, such that feature i in Dataset 1 and feature j in Dataset 2 correspond to the same metabolic feature. More formally, we aim to identify a matching matrix $M^* \in \{0, 1\}^{p_1 \times p_2}$ such that $M_{ij}^* = 1$ if features i in Dataset 1 and feature j in Dataset 2 correspond to the same feature, hereafter referred to as matched features. Otherwise we set $M_{ij}^* = 0$ otherwise. We emphasize that a matching matrix M^* can have at most one nonzero entry in each row and column.

Both of the datasets we aim to match are obtained from liquid chromatography-mass spectrometry (LC-MS) experiments. Hence, for Dataset 1 each metabolite $i \in [p_1]$ is labeled with a mass-to-charge (m/z) ratio m_i^1 , as well as a retention time (RT) given by RT_i^1 . Additionally, each metabolite has a vector of intensities across patients denoted by $X_i \in \mathbb{R}^{n_1}$. Similarly, each metabolite $j \in [p_2]$ in Dataset 2 is labeled by its m/z ratio m_j^2 , its retention time RT_j^2 and its vector of intensities across samples $Y_j \in \mathbb{R}^{n_2}$.

Correlations and distances between metabolites

Features cannot be aligned based on their m/z and RT alone as they are often too inconsistent across studies. Our method is based on the idea that, in addition to their m/z and RT being compatible, the signal intensities of metabolites measured in two different studies should exhibit similar correlation structures, or more generally exhibit similar distances between their intensity vectors. In other words, if feature intensity vectors $X_i \in \mathbb{R}^{n_1}, Y_j \in \mathbb{R}^{n_2}$ correspond to the same underlying feature ($M_{ij}^* = 1$) and similarly if $X_k \in \mathbb{R}^{n_1}, Y_l \in \mathbb{R}^{n_2}$ correspond to the same feature ($M_{kl}^* = 1$), then we expect that

$$\text{corr}(X_i, X_k) \approx \text{corr}(Y_j, Y_l) \quad \text{if} \quad \Pi_{ij}^* = \Pi_{kl}^* = 1. \quad (18)$$

Here we define $\text{corr}(u, v)$ to be the Pearson correlation coefficient between two feature intensity vectors $u, v \in \mathbb{R}^n$ by

$$\text{corr}(u, v) = \frac{\langle u - \bar{u}, v - \bar{v} \rangle}{\|u - \bar{u}\| \|v - \bar{v}\|} \quad (19)$$

where we define

$$\bar{u} = \frac{1}{n} \sum_{i=1}^n u_i, \quad \|u\| = \sqrt{\sum_{i=1}^n u_i^2}, \quad \langle u, v \rangle = \sum_{i=1}^n u_i v_i \quad (20)$$

as the mean value, Euclidean norm and inner product respectively. More generally, with d_x and d_y denoting two given distances on \mathbb{R}^{n_1} and \mathbb{R}^{n_2} respectively, we expect that

$$d_x(X_i, X_k) \approx d_y(Y_j, Y_l) \quad \text{if} \quad \Pi_{ij}^* = \Pi_{kl}^* = 1. \quad (21)$$

Throughout this paper, we use the normalized Euclidean distance defined for any $u, v \in \mathbb{R}^n$ as

$$d^{\text{euc}}(u, v) = \frac{1}{\sqrt{n}} \|u - v\| \quad (22)$$


where for d_x and d_y we take $n = n_1, n_2$ respectively. If the signal intensity vectors u, v are mean centered and normalized by their standard deviation as

$$u \mapsto \sqrt{n} \cdot \frac{u - \frac{1}{n} \sum_{i=1}^n u_i}{\left\| u - \frac{1}{n} \sum_{i=1}^n u_i \right\|} \quad (23)$$

and likewise for v , then it follows that

$$d^{\text{euc}}(u, v) = \sqrt{2(1 - \text{corr}(u, v))} = \sqrt{2} d^{\text{cos}}(u, v) \quad (24)$$

where we denote $d^{\text{cos}}(u, v) = \sqrt{1 - \text{corr}(u, v)}$ as the cosine distance. For the purposes of this paper, we will always assume that d_x and d_y denote the normalized Euclidean distance from (22). As shown above, this will be implicitly equal to the cosine distance from (24) on centered and scaled data.

The goal of metabolomic feature matching is to learn the binary matching matrix M^* that aligns the distances between pairs of features in the most consistent way possible as shown in (21). To formalize this notion into a practical algorithm, we use the mathematical theory of optimal transport [Peyré et al. \(2019\)](#)  which we discuss next.

Optimal transport

Optimal transport (OT) applies in the setting when the points $\{X_i\}_{i=1}^{p_1}$ and $\{Y_j\}_{j=1}^{p_2}$ and $\{Y_j\}_{j=1}^{p_2}$ being matched live in the same dimensional space $n_1 = n_2 = n$. It aims to find a matching between each point X_i and its corresponding point Y_j such that the sum of distances between matches is minimized. Matches between each pair of points can be stored in a matching matrix $M \in \{0, 1\}^{p_1 \times p_2}$ such that $M_{ij} = 1$ if X_i and Y_j are matched, and $M_{ij} = 0$ otherwise. Again we note that M must have at most one nonzero entry in each row and column to be a valid matching matrix.

Instead of searching over this space of binary matching matrices, optimal transport places masses $a_i \geq 0$ at all points X_i for $i = 1, \dots, p_1$ and masses $b_j \geq 0$ at all points Y_j for $j = 1, \dots, p_2$ and optimizes over the space of probabilistic couplings $\Pi \in \mathbb{R}_+^{p_1 \times p_2}$ which move a Π_{ij} amount of mass from X_i to Y_j . We assume here for simplicity that the sum of masses in both datasets are equal to one $\sum_{i=1}^{p_1} a_i = \sum_{j=1}^{p_2} b_j = 1$ and that the coupling Π transports all mass from a into b . More formally, optimal transport optimizes over the constrained set of couplings

$$\mathcal{U}(a, b) = \{ \Pi \in \mathbb{R}_+^{p_1 \times p_2} : \Pi \mathbf{1}_{p_2} = a \text{ and } \Pi^T \mathbf{1}_{p_1} = b \} \quad (25)$$

where $\mathbf{1}_p$ denotes the all ones vector of length p . In practice, the points X_i and Y_j in each dataset are all treated the same and the masses placed on the data are chosen to be uniform $a = \frac{1}{p_1} \mathbf{1}_{p_1}$ and $b = \frac{1}{p_2} \mathbf{1}_{p_2}$.

The cost function which optimal transport minimizes is the sum of squared distances of its transported mass

$$\mathcal{E}(\Pi) = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} d(X_i, Y_j) \quad (26)$$

where $d(u, v) = d^{\text{euc}}(u, v) = \frac{1}{\sqrt{n}} \|u - v\|$ is the Euclidean distance. The distance matrix $d(X_i, Y_j)$ in the OT objective can be replaced more generally with a cost matrix $C \in \mathbb{R}^{p_1 \times p_2}$ that is not necessarily a distance matrix. In this case the cost function becomes

$$\mathcal{E}(\Pi) = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} C_{ij} \quad (27)$$

When the transport cost C_{ij} is a distance, the OT optimization defines a valid distance metric known as the optimal transport distance between discrete distributions $\{(a_i, X_i)\}_{i=1}^{p_1}$ and $\{(b_j, Y_j)\}_{j=1}^{p_2}$ in \mathbb{R}^n given by

$$\text{OT}(a, b) = \min_{\Pi \in \mathcal{U}(a, b)} \mathcal{E}(\Pi). \quad (28)$$

When $d(u, v)$ is Euclidean, this OT distance is also referred to as the L^1 optimal transport distance, the Wasserstein 1-distance, or the Earth mover's distance. As formulated, the computation of the optimal transport objective involves an optimization over coupling matrices which can be solved by linear programming [Peyré et al. \(2019\)](#). The OT optimization problem becomes time consuming for problems with many points $p_1, p_2 \gg 1$. We show in the next section how augmenting this distance with a regularization term leads to a more efficient algorithm for learning the optimal coupling.

Entropy regularization

Define the Kullback-Leibler (KL) divergence between two positive vectors $\mu, \nu \in \mathbb{R}_+^p$ as

$$D_{\text{KL}}(\mu, \nu) = \sum_{i=1}^p \mu_i \ln \left(\frac{\mu_i}{\nu_i} \right) - \sum_{i=1}^p \mu_i + \sum_{i=1}^p \nu_i \quad (29)$$

Given fixed marginals $a \in \mathbb{R}^{p_1}$ and $b \in \mathbb{R}^{p_2}$ from the previous section, we can define the entropy of a coupling matrix $\Pi \in \mathbb{R}_+^{p_1 \times p_2}$ with respect to these fixed marginals as

$$D_{\text{KL}}(\Pi, a \otimes b) = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} \ln \left(\frac{\Pi_{ij}}{a_i b_j} \right) - \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} + \sum_{i=1}^{p_1} a_i + \sum_{j=1}^{p_2} b_j. \quad (30)$$

where $(a \otimes b)_i = a_i b_j$ denotes the outer product. This can be further simplified as

$$\begin{aligned} D_{\text{KL}}(\Pi, a \otimes b) &= \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} \ln \left(\frac{\Pi_{ij}}{a_i b_j} \right) - \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} + \left(\sum_{i=1}^{p_1} a_i \right) \left(\sum_{j=1}^{p_2} b_j \right) \\ &= \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} \ln \left(\frac{\Pi_{ij}}{a_i b_j} \right) \\ &= H(\Pi) + \ln(p_1) + \ln(p_2) \end{aligned} \quad (31)$$

where we define $H(\Pi)$ by

$$H(\Pi) = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} \ln(\Pi_{ij}). \quad (32)$$

In the second line of the derivation above, we used the fact that the entries of a , b , and Π summed to one, and in the third line we used the fact that the marginals a and b were uniform. Under these assumptions, we see that the KL divergence $D_{\text{KL}}(\Pi, a \otimes b)$ is independent of the values of the marginals a , b and is equal to $H(\Pi)$ up to constants.

Although here the general definition of entropy through the KL divergence reduces to the simpler formula of $H(\Pi)$, in the following sections we will need to extend our analysis to cases when a , b , and Π have positive values that do not sum to one (i.e. not distributions). In this context, we will no longer have that $D_{\text{KL}}(\Pi, a \otimes b) = H(\Pi) + \text{const}$ but we will still be able to use $D_{\text{KL}}(\Pi, a \otimes b)$ as a general notion of entropy for Π .

The entropy of a coupling $D_{\text{KL}}(\Pi, a \otimes b)$ is an important notion because it quantifies how uniform or smooth Π is with respect to the product distribution $a \otimes b$. In particular, if a and b are set to uniform distributions as commonly done in practice, then $D_{\text{KL}}(\Pi, a \otimes b)$ is small when Π has close to uniform entries and is large otherwise. This notion of smoothness allows us to use $D_{\text{KL}}(\Pi, a \otimes b)$ as a regularizer in our optimal transport distance as

$$\mathcal{L}_\epsilon(\Pi) = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} C_{ij} + \epsilon D_{\text{KL}}(\Pi, a \otimes b) \quad (33)$$

where ϵ is a small regularization parameter. Note that here we have denoted the transport cost matrix by $C \in \mathbb{R}^{p_1 \times p_2}$ which is not necessarily a distance matrix. The introduction of the regularizer $D_{\text{KL}}(\Pi, a \otimes b)$ gives us an efficient iterative algorithm known as the Sinkhorn algorithm for optimizing \mathcal{L}_ϵ which we describe in the following sections.

Unbalanced optimal transport

Before we introduce the Sinkhorn algorithm, we introduce a final modification to our optimal transport distance that allows us to learn couplings between distributions $a, b \in \mathbb{R}_+^p$ that do not preserve mass. In other words, the coupling Π is not required to perfectly satisfy the marginal constraints $\Pi \mathbf{1}_{p_2} = a$ and $\Pi^T \mathbf{1}_{p_1} = b$. In our metabolite matching problem, this is particularly useful as not all metabolites in one dataset necessarily appear in the other dataset and hence should be left unmatched. This modification of optimal transport, known as unbalanced optimal transport (UOT) Chizat et al. (2018) [optimizes the following cost function](#)

$$\mathcal{L}_{\rho, \epsilon}(\Pi) = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} C_{ij} + \rho D_{\text{KL}}(\Pi \mathbf{1}_{p_2}, a) + \rho D_{\text{KL}}(\Pi^T \mathbf{1}_{p_1}, b) + \epsilon D_{\text{KL}}(\Pi, a \otimes b) \quad (34)$$

where we have added two KL terms with regularization parameter ρ to enforce that the marginals of the coupling $\Pi \mathbf{1}_{p_2}, \Pi^T \mathbf{1}_{p_1}$ are approximately close to the prescribed marginals a, b respectively. We have also kept the smoothness/entropy regularizer $D_{\text{KL}}(\Pi, a \otimes b)$ from the previous section.

Unbalanced Sinkhorn algorithm

Now we are ready to present the unbalanced Sinkhorn algorithm [Peyré et al. \(2019\)](#) for optimizing the unbalanced optimal transport cost defined above. First we rewrite our optimization as

$$\begin{aligned} \min_{\Pi \in \mathbb{R}_+^{p_1 \times p_2}} \mathcal{L}_{\rho, \varepsilon}(\Pi) &= \min_{\Pi \in \mathbb{R}_+^{p_1 \times p_2}} \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} C_{ij} + \rho D_{\text{KL}}(\Pi \mathbf{1}_{p_2}, a) + \rho D_{\text{KL}}(\Pi^T \mathbf{1}_{p_1}, b) + \varepsilon D_{\text{KL}}(\Pi, a \otimes b) \\ &= \min_{u \in \mathbb{R}_+^{p_1}, v \in \mathbb{R}_+^{p_2}} \min_{\substack{\Pi \in \mathbb{R}_+^{p_1 \times p_2} \\ \Pi \mathbf{1}_{p_2} = u, \Pi^T \mathbf{1}_{p_1} = v}} \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} C_{ij} + \rho D_{\text{KL}}(u, a) + \rho D_{\text{KL}}(v, b) + \varepsilon D_{\text{KL}}(\Pi, a \otimes b). \end{aligned}$$

The inner minimization can be solved exactly by introducing dual variables $f \in \mathbb{R}^{p_1}, g \in \mathbb{R}^{p_2}$ and writing out the Lagrange dual problem

$$\begin{aligned} \min_{\substack{\Pi \in \mathbb{R}_+^{p_1 \times p_2} \\ \Pi \mathbf{1}_{p_2} = u, \Pi^T \mathbf{1}_{p_1} = v}} \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} C_{ij} + \varepsilon D_{\text{KL}}(\Pi, a \otimes b) - f^T(\Pi \mathbf{1}_{p_2} - u) - g^T(\Pi^T \mathbf{1}_{p_1} - v) &= \\ \max_{f \in \mathbb{R}^{p_1}, g \in \mathbb{R}^{p_2}} \min_{\Pi \in \mathbb{R}_+^{p_1 \times p_2}} \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} C_{ij} + \varepsilon D_{\text{KL}}(\Pi, a \otimes b) - f^T(\Pi \mathbf{1}_{p_2} - u) - g^T(\Pi^T \mathbf{1}_{p_1} - v). \end{aligned}$$

where we have removed the terms $D_{\text{KL}}(u, a)$ and $D_{\text{KL}}(v, b)$ since they do not depend on Π . Taking the gradient in Π in the inner minimization and setting it to zero we get

$$C_{ij} + \varepsilon \log\left(\frac{\Pi_{ij}}{a_i b_j}\right) - f_i - g_j = 0$$

which implies that

$$\Pi_{ij} = a_i b_j \exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right)$$

Now we can substitute this expression for Π back into our Lagrange dual problem. First we compute

$$\varepsilon D_{\text{KL}}(\Pi, a \otimes b) = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} (f_i + g_j - C_{ij}) - \varepsilon \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} + \varepsilon \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} a_i b_j$$

which implies that

$$\begin{aligned} \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} C_{ij} + \varepsilon D_{\text{KL}}(\Pi, a \otimes b) - f^T(\Pi \mathbf{1}_{p_2} - u) - g^T(\Pi^T \mathbf{1}_{p_1} - v) &= \\ = u^T f + v^T g - \varepsilon \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} a_i b_j \exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right) + \varepsilon \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} a_i b_j. \end{aligned}$$

Hence, the outer maximization in our Lagrange dual problem for f and g can now be written as

$$\max_{f \in \mathbb{R}^{p_1}, g \in \mathbb{R}^{p_2}} u^T f + v^T g - \varepsilon \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} a_i b_j \exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right)$$

where we have removed the last constant sum in $a_i b_j$. Finally we can rewrite our entire minimization from the start of this section as

$$\begin{aligned} \min_{\Pi \in \mathbb{R}_+^{p_1 \times p_2}} \mathcal{L}_{\rho, \varepsilon}(\Pi) &= \min_{u \in \mathbb{R}_+^{p_1}, v \in \mathbb{R}_+^{p_2}} \min_{\substack{\Pi \in \mathbb{R}_+^{p_1 \times p_2} \\ \Pi \mathbf{1}_{p_2} = u, \Pi^T \mathbf{1}_{p_1} = v}} \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} C_{ij} + \rho D_{\text{KL}}(u, a) + \rho D_{\text{KL}}(v, b) + \varepsilon D_{\text{KL}}(\Pi, a \otimes b) \\ &= \min_{u \in \mathbb{R}_+^{p_1}, v \in \mathbb{R}_+^{p_2}} \max_{f \in \mathbb{R}^{p_1}, g \in \mathbb{R}^{p_2}} u^T f + v^T g - \varepsilon \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} a_i b_j \exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right) + \rho D_{\text{KL}}(u, a) + \rho D_{\text{KL}}(v, b). \end{aligned}$$

By strong duality, we can interchange the minimum and maximum above to write

$$\begin{aligned} \min_{\Pi \in \mathbb{R}_+^{p_1 \times p_2}} \mathcal{L}_{\rho, \varepsilon}(\Pi) &= \max_{f \in \mathbb{R}^{p_1}, g \in \mathbb{R}^{p_2}} \min_{u \in \mathbb{R}_+^{p_1}, v \in \mathbb{R}_+^{p_2}} u^T f + v^T g - \varepsilon \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} a_i b_j \exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right) + \rho D_{\text{KL}}(u, a) + \rho D_{\text{KL}}(v, b) \\ &= U^*(f) + V^*(g) - \varepsilon \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} a_i b_j \exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right) \end{aligned}$$

where we define the functions

$$\begin{aligned} U^*(f) &= \min_{u \in \mathbb{R}_+^{p_1}} u^T f + \rho D_{\text{KL}}(u, a) \\ V^*(g) &= \min_{v \in \mathbb{R}_+^{p_2}} v^T g + \rho D_{\text{KL}}(v, b). \end{aligned} \tag{35}$$

In fact, we can solve the minimizations in U^* and V^* in closed form to get the minimizers $u^* = a \exp(-f / \rho)$ and $v^* = b \exp(-g / \rho)$ which we can substitute back in to get

$$\begin{aligned} U^*(f) &= \sum_{i=1}^{p_1} u_i^* f_i + \rho \sum_{i=1}^{p_1} u_i^* \ln\left(\frac{u_i^*}{a_i}\right) - \rho \sum_{i=1}^{p_1} u_i^* + \rho \sum_{i=1}^{p_1} a_i \\ &= \sum_{i=1}^{p_1} u_i^* f_i - \sum_{i=1}^{p_1} u_i^* f_i - \rho \sum_{i=1}^{p_1} a_i \exp(-f_i / \rho) + \rho \sum_{i=1}^{p_1} a_i \\ &= -\rho \sum_{i=1}^{p_1} a_i \exp(-f_i / \rho) + \rho \sum_{i=1}^{p_1} a_i. \end{aligned}$$

Likewise we can see that

$$V^*(g) = -\rho \sum_{i=1}^{p_2} b_i \exp(-g_i / \rho) + \rho \sum_{i=1}^{p_2} b_i.$$

Thus, we can rewrite our full optimization as

$$\begin{aligned} \min_{\Pi \in \mathbb{R}_+^{p_1 \times p_2}} \mathcal{L}_{\rho, \varepsilon}(\Pi) &= \max_{f \in \mathbb{R}^{p_1}, g \in \mathbb{R}^{p_2}} -\rho \sum_{i=1}^{p_1} a_i \exp(-f_i / \rho) - \rho \sum_{i=1}^{p_2} b_i \exp(-g_i / \rho) - \varepsilon \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} a_i b_j \exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right) \\ &= \min_{f \in \mathbb{R}^{p_1}, g \in \mathbb{R}^{p_2}} \rho \sum_{i=1}^{p_1} a_i \exp(-f_i / \rho) + \rho \sum_{i=1}^{p_2} b_i \exp(-g_i / \rho) + \varepsilon \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} a_i b_j \exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right) \end{aligned}$$

where we have removed the terms independent of f and g .

Note that now we can optimize the cost function above by performing an alternating minimization on the dual variables f and g . Taking the gradient in f and setting it to zero we see that

$$-a_i \exp(-f_i / \rho) + a_i \sum_{j=1}^{p_2} b_j \exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right) = 0$$

which implies that

$$f_i = -\frac{\varepsilon \rho}{\varepsilon + \rho} \ln \left(\sum_{j=1}^{p_2} b_j \exp \left(\frac{g_j - C_{ij}}{\varepsilon} \right) \right).$$

Similarly, we can write out

$$g_j = -\frac{\varepsilon \rho}{\varepsilon + \rho} \ln \left(\sum_{i=1}^{p_1} a_i \exp \left(\frac{f_i - C_{ij}}{\varepsilon} \right) \right).$$

We are now ready to write out the full unbalanced Sinkhorn algorithm which performs an alternating minimization on the dual potentials f, g as outlined above. We remind the reader that the coupling matrix can be recovered from the dual potentials by the formula

$$\Pi_{ij} = a_i b_j \exp \left(\frac{f_i + g_j - C_{ij}}{\varepsilon} \right).$$

The unbalanced Sinkhorn algorithm proceeds as follows.

The final output of the Sinkhorn algorithm optimization is a real-valued coupling matrix $\Pi \in \mathbb{R}_+^{p_1 \times p_2}$. In some cases, it is desirable to transform the coupling matrix into a binary-valued matching matrix $M \in \{0, 1\}^{p_1 \times p_2}$ with possibly an added restriction that there is at most one nonzero element in each row and column (to obtain a valid partial matching). This can be done by either thresholding the real matrix or by assigning all maximal entries in each row (or column) to one and setting the remaining entries to zero. For our metabolomics matching problem, we describe our procedure for transforming our real-valued coupling into a binary matching matrix in the section on the GromovMatcher algorithm below.

Gromov - Wasserstein

Now that we have introduced the general formulation of unbalanced optimal transport and its corresponding Sinkhorn algorithm, we can extend this formulation to matching problems between distributions of points that live in different dimensional spaces. In our metabolomics setting, we aim to match two datasets of p_1 and p_2 metabolic features respectively where each feature in a dataset is associated with a feature intensity vector $\{X_i\}_{i=1}^{p_1} \subset \mathbb{R}^{n_1}$ and $\{Y_j\}_{j=1}^{p_2} \subset \mathbb{R}^{n_2}$ respectively across samples. We assume that there exists a true matching matrix $M^* \in \{0, 1\}^{p_1 \times p_2}$ with at most one nonzero entry in each row and column such that two metabolites (i, j) are matched if $M_{ij}^* = 1$.

We make the further assumption that if feature vectors X_i, Y_j are matched and feature vectors X_k, Y_l are matched under M^* , then we expect that

$$d_x(X_i, X_k) \approx d_y(Y_j, Y_l) \quad \text{if} \quad M_{ij}^* = M_{kl}^* = 1. \quad (36)$$

where d_x is a distance metric on \mathbb{R}^{n_1} and d_y is a distance metric \mathbb{R}^{n_2} . In practice, we always choose these distance metrics to be the normalized Euclidean distance defined for any $u, v \in \mathbb{R}^n$ as

$$d^{\text{euc}}(u, v) = \frac{1}{\sqrt{n}} \|u - v\| \quad (37)$$

which is equal to the cosine distance d^{cos} (i.e. one minus the correlation) for centered and scaled data. Given these two distance matrices $D^x = [d_x(X_i, X_k)]_{i,k=1}^{p_1} \in \mathbb{R}^{p_1 \times p_1}$ and $D^y = [d_y(Y_j, Y_l)]_{j,l=1}^{p_2} \in \mathbb{R}^{p_2 \times p_2}$ we would like to infer the true matching matrix M^* by solving an optimization problem.

Algorithm 1: UnbalancedSinkhorn

input : Transport cost C , marginals a, b , marginal relaxation ρ , entropic regularization ϵ

output: Return the coupling matrix Π

Initialize $g = 0$

while (f, g) has not converged **do**

 Set $f_i \leftarrow -\frac{\epsilon\rho}{\epsilon+\rho} \ln \left(\sum_{j=1}^{p_2} b_j \exp \left(\frac{g_j - C_{ij}}{\epsilon} \right) \right)$ for $i \in [p_1]$

 Set $g_j \leftarrow -\frac{\epsilon\rho}{\epsilon+\rho} \ln \left(\sum_{i=1}^{p_1} a_i \exp \left(\frac{f_i - C_{ij}}{\epsilon} \right) \right)$ for $j \in [p_2]$

Return the coupling matrix $\Pi_{ij} = a_i b_j \exp \left(\frac{f_i + g_j - C_{ij}}{\epsilon} \right)$ for $i \in [p_1]$ and $j \in [p_2]$

Algorithm 1

UnbalancedSinkhorn

Consider the following objective function

$$\mathcal{E}(M) = \sum_{i,k=1}^{p_1} \sum_{j,l=1}^{p_2} M_{ij} M_{kl} |D_{ik}^x - D_{jl}^y|, \quad (38)$$

where the matching matrices $M \in \{0, 1\}^{p_1 \times p_2}$ we optimize over are constrained to satisfy marginal constraints $\mathbf{1}_{p_2}^T M > 0$ and $M^T \mathbf{1}_{p_1} > 0$. These marginal constraints simply impose that there is at least one nonzero entry in each row and column (i.e. each metabolite in both datasets has at least one corresponding match). Searching for the minimizing χ , ψ consists of putting the nonzero entries in M such that the distance profiles of the matched features are similar, so that the minimizer of this criterion provides a good candidate estimate of χ^* . This is closely related to the Gromov-Hausdorff distance [Gromov \(2001\)](#), an extension of optimal transport to the case where the sets to be coupled do not lie in the same metric space.

In practice, it is often desirable to optimize over a different set of matrices in order to make the optimization problem more tractable. Here we take intuition from optimal transport, and search over the set of coupling matrices with marginal constraints

$$\mathbf{U}(a, b) = \{\Pi \in \mathbb{R}_+^{p_1 \times p_2} : \Pi \mathbf{1}_{p_2} = a \text{ and } \Pi^T \mathbf{1}_{p_1} = b\}. \quad (39)$$

where as before, $a \in \mathbb{R}_+^{p_1}$ and $b \in \mathbb{R}_+^{p_2}$ are desired marginals which are typically set to be uniform distributions $a = \frac{1}{p_1} \mathbf{1}_{p_1}$ and $b = \frac{1}{p_2} \mathbf{1}_{p_2}$. These marginal vectors can be interpreted as distributions of masses a_i and b_j on the feature vectors X_i and Y_j respectively for $i \in [p_1], j \in [p_2]$.

Coupling matrices in $\mathbf{U}(a, b)$ transport the distribution of masses a in the first dataset to the distribution of masses b in the second dataset. Now we can formulate the Gromov-Wasserstein (GW) distance, introduced by [Memoli \(2011\)](#), as

$$\text{GW}(a, b) = \min_{\Pi \in \mathbf{U}(a, b)} \mathcal{E}(\Pi) \quad (40)$$

By optimizing this objective, each entry Π_{ij} now reflects the strength of the matched pair (X_i, Y_j) . Optimizing $\text{GW}(a, b)$ then amounts to placing larger entries in Π whose paired features have similar distance profiles. Before we develop an algorithm to optimize this objective, we first modify it to allow for unbalanced matchings where marginal constraints are not enforced exactly (e.g. features in both datasets can remain unmatched).

Unbalanced Gromov-Wasserstein

In an untargeted context, all features measured in one study are not necessarily observed in another, either because these features are truly not shared or because of measurement error. However, the constraint $\Pi \in \mathbf{U}(a, b)$ in the original GW optimization criterion (40) ensures that all the mass is transported from one set to another, resulting in all features being matched across studies. In order to discard study-specific features during the GW computation, we use the unbalanced Gromov-Wasserstein (UGW) distance with an additional entropic regularization for computational purposes, described in [Séjourné et al. \(2021\)](#). The optimization problem therefore reads

$$\begin{aligned} \mathcal{L}_{\rho, \varepsilon}(\Pi) = & \mathcal{E}(\Pi) + \rho D_{\text{KL}}(\Pi \mathbf{1}_{p_2} \otimes \Pi \mathbf{1}_{p_2}, a \otimes a) \\ & + \rho D_{\text{KL}}(\Pi^T \mathbf{1}_{p_1} \otimes \Pi^T \mathbf{1}_{p_1}, b \otimes b) \\ & + \varepsilon D_{\text{KL}}(\Pi \otimes \Pi, (a \otimes b)^{\otimes 2}) \end{aligned} \quad (41)$$

$$\text{UGW}_{\rho, \varepsilon} = \min_{\Pi \in \mathbb{R}_+^{p_1 \times p_2}} \mathcal{L}_{\rho, \varepsilon}(\Pi) \quad (42)$$

with $\epsilon > 0$. Here D_{KL} is the Kullback–Leibler divergence defined in the previous sections and we define the tensor product $(P \otimes Q)_{i_1, \dots, i_p, j_1, \dots, j_q} = P_{i_1, \dots, i_p} Q_{j_1, \dots, j_q}$. Here we set the desired marginal constraints to $a = \frac{1}{p_1} \mathbf{1}_{p_1}$ and $b = \frac{1}{p_2} \mathbf{1}_{p_2}$ as before.

As in the case of unbalanced optimal transport Chizat et al. (2018), the regularization ϵ times the Kullback–Leibler divergences allows for the relaxation of the marginal constraints $\Pi \mathbf{1}_{p_2} = a$ and $\Pi^T \mathbf{1}_{p_1} = b$. The value of $\epsilon > 0$ controls the extent to which we allow for mass destruction. Smaller values of ϵ tend to lessen the constraint on the marginals of Π , while balanced GW is recovered when $\epsilon \rightarrow 0$. As proposed in the original paper Séjourné et al. (2021), our UGW cost modifies the UOT formulation by using the quadratic Kullback–Leibler divergence in $\Pi \mathbf{1}_{p_2} \otimes \Pi \mathbf{1}_{p_2}$ and $\Pi^T \mathbf{1}_{p_1} \otimes \Pi^T \mathbf{1}_{p_1}$ instead, hence preserving the quadratic form of the GW cost function (42).

The term $\epsilon D_{KL}(\Pi \otimes \Pi, (a \otimes b)^{\otimes 2})$ serves as an entropic regularization, inspired again by optimal transport. Adding such a penalty is a standard way to compute an approximate solution to the optimal transport problem using the Sinkhorn algorithm as we shall show in the following section. Here again, we modify the entropic penalty in UGW to have a quadratic form in Π to agree with the quadratic form of the GW cost (42). The parameter ϵ controls the smoothness (entropy) of the coupling matrix Π where larger values of ϵ encourage Π to put uniform weights on many of its entries, leading to less precision in the feature matches. However, increasing ϵ also leads to better numerical stability and a significant speedup of the alternating Sinkhorn algorithm used to optimize the objective function described below.

UGW optimization algorithm

Now we are ready to write out an algorithm to optimize the UGW objective in (42). First write our objective as

$$\begin{aligned} \mathcal{L}_{\rho, \epsilon}(\Pi) = & \sum_{i,k=1}^{p_1} \sum_{j,l=1}^{p_2} \Pi_{ij} \Pi_{kl} |D_{ik}^x - D_{jl}^y| + \rho D_{KL}(\Pi \mathbf{1}_{p_2} \otimes \Pi \mathbf{1}_{p_2}, a \otimes a) \\ & + \rho D_{KL}(\Pi^T \mathbf{1}_{p_1} \otimes \Pi^T \mathbf{1}_{p_1}, b \otimes b) \\ & + \epsilon D_{KL}(\Pi \otimes \Pi, (a \otimes b)^{\otimes 2}). \end{aligned} \quad (43)$$

Using the quadratic nature of our cost function, we aim to perform an alternating minimization in the two copies of Π . For the moment, let's differentiate these two copies by Γ and Π and write the new cost

$$\begin{aligned} \mathcal{F}_{\rho, \epsilon}(\Pi, \Gamma) = & \sum_{i,k=1}^{p_1} \sum_{j,l=1}^{p_2} \Pi_{ij} \Gamma_{kl} |D_{ik}^x - D_{jl}^y| + \rho D_{KL}(\Pi \mathbf{1}_{p_2} \otimes \Gamma \mathbf{1}_{p_2}, a \otimes a) \\ & + \rho D_{KL}(\Pi^T \mathbf{1}_{p_1} \otimes \Gamma^T \mathbf{1}_{p_1}, b \otimes b) \\ & + \epsilon D_{KL}(\Pi \otimes \Gamma, (a \otimes b)^{\otimes 2}). \end{aligned} \quad (44)$$

Before we expand this cost, we introduce the notation $m(\cdot)$ to denote the sum of the elements of which can be a vector, matrix or tensor. In general, for four positive distributions $\pi, a \in \mathbb{R}_+^p$ and $\gamma, b \in \mathbb{R}_+^q$ we have that the KL satisfies the tensorization property

$$\begin{aligned} D_{KL}(\pi \otimes \gamma, a \otimes b) &= \sum_{i=1}^p \sum_{j=1}^q \pi_i \gamma_j \ln \left(\frac{\pi_i \gamma_j}{a_i b_j} \right) - \sum_{i=1}^p \sum_{j=1}^q \pi_i \gamma_j + \sum_{i=1}^p \sum_{j=1}^q a_i b_j \\ &= m(\gamma) \sum_{i=1}^p \pi_i \ln \left(\frac{\pi_i}{a_i} \right) + m(\pi) \sum_{j=1}^q \gamma_j \ln \left(\frac{\gamma_j}{b_j} \right) - m(\pi) m(\gamma) + m(a) m(b) \\ &= m(\gamma) D_{KL}(\pi, a) + m(\pi) D_{KL}(\gamma, b) + (m(\pi) - m(a))(m(\gamma) - m(b)). \end{aligned} \quad (45)$$

Specifically, if we remove those terms that do not depend on γ we are left with

$$D_{\text{KL}}(\pi \otimes \gamma, a \otimes b) = m(\pi)D_{\text{KL}}(\gamma, b) + m(\gamma) \sum_{i=1}^p \pi_i \ln \left(\frac{\pi_i}{a_i} \right) + \text{const.} \quad (46)$$

This allows us to write for the marginal constraints $a \in \mathbb{R}_+^{p_1}$, $b \in \mathbb{R}_+^{p_2}$ and couplings $\Pi, \Gamma \in \mathbb{R}_+^{p_1 \times p_2}$ that

$$\begin{aligned} D_{\text{KL}}(\Pi \mathbf{1}_{p_2} \otimes \Gamma \mathbf{1}_{p_2}, a \otimes a) &= m(\Pi)D_{\text{KL}}(\Gamma \mathbf{1}_{p_2}, a) + m(\Gamma) \sum_{i=1}^{p_1} (\Pi \mathbf{1}_{p_2})_i \ln \left(\frac{(\Pi \mathbf{1}_{p_2})_i}{a_i} \right) + \text{const.} \\ D_{\text{KL}}(\Pi^T \mathbf{1}_{p_1} \otimes \Gamma^T \mathbf{1}_{p_1}, b \otimes b) &= m(\Pi)D_{\text{KL}}(\Gamma^T \mathbf{1}_{p_1}, b) + m(\Gamma) \sum_{j=1}^{p_2} (\Pi^T \mathbf{1}_{p_1})_j \ln \left(\frac{(\Pi^T \mathbf{1}_{p_1})_j}{b_j} \right) + \text{const.} \\ D_{\text{KL}}(\Pi \otimes \Gamma, (a \otimes b)^{\otimes 2}) &= m(\Pi)D_{\text{KL}}(\Gamma, a \otimes b) + m(\Gamma) \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} \ln \left(\frac{\Pi_{ij}}{a_i b_j} \right) + \text{const.} \end{aligned}$$

where in the expansions above we have removed all terms that are independent of γ . Finally, expanding out γ , (γ, γ) and keeping only those terms that depend on γ we get

$$\mathcal{F}_{\rho, \epsilon}(\Pi, \Gamma) = \sum_{k=1}^{p_1} \sum_{l=1}^{p_2} \Gamma_{kl} C_{kl}^{\Pi} + \rho m(\Pi)D_{\text{KL}}(\Gamma \mathbf{1}_{p_2}, a) + \rho m(\Pi)D_{\text{KL}}(\Gamma^T \mathbf{1}_{p_1}, b) + \epsilon m(\Pi)D_{\text{KL}}(\Gamma, a \otimes b) \quad (47)$$

where the cost matrix $C^{\Pi} \in \mathbb{R}^{p_1 \times p_2}$ is defined as

$$C_{kl}^{\Pi} = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} |D_{ik}^x - D_{jl}^y| + \rho \sum_{i=1}^{p_1} (\Pi \mathbf{1}_{p_2})_i \ln \left(\frac{(\Pi \mathbf{1}_{p_2})_i}{a_i} \right) + \rho \sum_{j=1}^{p_2} (\Pi^T \mathbf{1}_{p_1})_j \ln \left(\frac{(\Pi^T \mathbf{1}_{p_1})_j}{b_j} \right) + \epsilon \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \Pi_{ij} \ln \left(\frac{\Pi_{ij}}{a_i b_j} \right). \quad (48)$$

where we have hidden the dependence of C on the distance matrices D^x, D^y , the marginals a, b , and the regularization parameters ρ, ϵ for ease of notation.

Remarkably, the cost above in (48) for fixed Π is in the form of an unbalanced optimal transport problem which can be solved through unbalanced Sinkhorn iterations (Algorithm 1). Note that in our derivation above, it did not matter whether we optimized γ with Π fixed or vice versa because the cost $C(\gamma, \gamma)$ is symmetric in both of its arguments.

Our iterative algorithm for solving the unbalanced GW problem will proceed at each iteration by optimizing γ to minimize the cost above using the unbalanced Sinkhorn method, setting Π equal to γ and repeating. With each iteration, we expect this iterative procedure to make smaller and smaller updates to γ until convergence. By definition, at the end of each iteration we assign $\Pi = \gamma$ so the minimizer of $C(\gamma, \gamma)$ we converge to should also be a minimizer of the original UGW cost (48) in the sense that the relaxation of $C(\gamma, \gamma)$ to $C(\Pi, \Pi)$ is tight. This is proven rigorously under strict mathematical assumptions in [Sejourne et al. \(2021\)](#). We state the full UGW optimization algorithm below.

Following the implementation of the UGW algorithm in [Sejourne et al. \(2021\)](#), we initialize both Π and Γ to be the product distribution of the marginals $a \otimes b / \sqrt{m(a)m(b)}$ before we begin the optimization. Also, we note that if (Π, Γ) is a minimizer of our UGW objective (48), then so is $(\frac{1}{s}\Pi, s\Gamma)$ for any scale factor $s > 0$. Hence, we can set $m(\frac{1}{s}\Pi) = m(s\Gamma)$ by choosing $s = \sqrt{m(\Pi)/m(\Gamma)}$. This motivates the final step in the while loop of the UGW algorithm where the rescaling of Π by the factor $\sqrt{m(\Pi)/m(\Gamma)}$ leads to mass equality $m(\Pi) = m(\Gamma)$ and also stabilizes the convergence of the algorithm.

Algorithm 2: UnbalancedGromovWasserstein

input : Distance matrices D^x, D^y , marginals a, b , marginal relaxation ρ , entropic regularization ϵ

output: Return the coupling matrix Π

Initialize $\Pi = \Gamma = a \otimes b / \sqrt{m(a)m(b)}$

while (Π, Γ) has not converged **do**

 Update $\Pi \leftarrow \Gamma$

 Update $\Gamma = \text{UnbalancedSinkhorn}(C^\Pi, a, b, \rho m(\Pi), \epsilon m(\Pi))$

 Rescale $\Gamma \leftarrow \sqrt{m(\Pi)/m(\Gamma)}\Gamma$

Update $\Pi \leftarrow \Gamma$

Return Π .

Algorithm 2

Unbalanced Gromov Wasserstein

Returning to our metabolomics matching problem, we further guide our UGW optimization procedure by discouraging it from matching metabolic feature pairs whose mass-to-charge ratios are incompatible. Namely, we choose a value m_{gap} such that for all pairs (i, j) with $i \in [p_1], j \in [p_2]$ and mass-to-charge ratios m_i^x, m_j^y we enforce that

$$|m_i^x - m_j^y| > m_{\text{gap}} \implies \Pi_{ij} = 0. \quad (49)$$

In practice, this is done by taking the optimal transport cost C in every iteration of the UGW algorithm and premultiplying it elementwise by a factor $W \in \mathbb{R}_+^{p_1 \times p_2}$ given by

$$C^\Pi \rightarrow W \odot C^\Pi, \quad W_{ij} = 99 \cdot \mathbf{1}_{\{|m_i^x - m_j^y| > m_{\text{gap}}\}} + 1 \quad (50)$$

where $\mathbf{1}_x$ denotes the indicator function that is one when the condition x is satisfied and zero otherwise. Such a prefactor changes the transport cost to be very large for feature matches with incompatible mass-to-charge ratios, and hence, the entries of C set small weights at these entries. Our weighted UGW algorithm is rewritten below.

As mentioned before, the coupling matrix returned by our weighted UGW algorithm is a realvalued matrix rather than a binary matching matrix. In the next section, we describe how we incorporate metabolite retention time information to filter out unlikely pairs in our coupling matrix and transform it into a valid one-to-one matching of features across two datasets.

Retention time drift estimation and filtering

To filter out unlikely matches from the coupling matrix returned by Algorithm 3 above, we use the retention times (RTs) of the metabolites in both datasets. We remind the reader that RTs were not incorporated into the weighted UGW algorithm since they often exhibit a non-linear deviation between datasets, and hence are not directly comparable. However, using the metabolite coupling $\tilde{\Pi} \in \mathbb{R}_+^{p_1 \times p_2}$ obtained from Algorithm 3, it is possible to estimate this RT drift. The estimated RT drift $\hat{f} : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ allows us to assess the plausibility of the pairs recovered by the restricted UGW coupling $\tilde{\Pi}$, and discard pairs incompatible with the estimated drift.

We propose to learn the drift f through the weighted spline regression

$$\min_{f \in \mathcal{B}_{n,k}} \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \tilde{\Pi}_{ij} |f(RT_i^x) - RT_j^y| \quad (51)$$

where $\mathcal{B}_{n,k}$ is the set of n -order B-splines with k knots. All pairs (RT_i^x, RT_j^y) in objective (51) are weighted by the coefficients of $\tilde{\Pi}$ so that larger weights are given to pairs identified with high confidence in the first step of our procedure.

Pairs identified as incompatible with the estimated RT drift are then discarded from the coupling matrix. To do this, we first take the estimated RT drift \hat{f} , and the set of pairs $S = \{(i, j) : \tilde{\Pi}_{ij} \neq 0\}$ recovered in $\tilde{\Pi}$ with nonzero entries. We then define the residual associated with $(i, j) \in S$ as

$$r_f(i, j) = |\hat{f}(RT_i^x) - RT_j^y|. \quad (52)$$

The 95% prediction interval and the median absolute deviation (MAD) of these residuals are given by

$$\begin{aligned} \text{PI} &= 1.96 \times \text{std}(\{r_f(i, j), (i, j) \in S\}) \\ \text{MAD} &= \text{median}(\{|r_f(i, j) - \mu_r|, (i, j) \in S\}) \\ \mu_r &= \text{median}(\{|r_f(i, j)|, (i, j) \in S\}) \end{aligned} \quad (53)$$

Algorithm 3: WeightedUnbalancedGromovWasserstein

input : Distance matrices D^x, D^y , marginals a, b , marginal relaxation ρ , entropic regularization ϵ , mass-to-charge ratios m^x, m^y , mass-to-charge ratio gap m_{gap}

output: Return the coupling matrix Π

Initialize $\Pi = \Gamma = a \otimes b / \sqrt{m(a)m(b)}$

Set $W_{ij} = 99 \cdot \mathbf{1}_{\{|m_i^x - m_j^y| > m_{\text{gap}}\}} + 1$ for $i \in [p_1]$ and $j \in [p_2]$

while (Π, Γ) has not converged **do**

 Update $\Pi \leftarrow \Gamma$

 Update $\Gamma = \text{UnbalancedSinkhorn}(W \odot C^\Pi, a, b, \rho m(\Pi), \epsilon m(\Pi))$

 Rescale $\Gamma \leftarrow \sqrt{m(\Pi)/m(\Gamma)}\Gamma$

Update $\Pi \leftarrow \Gamma$

Return Π .

Algorithm 3

WeightedUnbalancedGromovWasserstein

where $|S_r|$ is the size of S_r and the functions std , median denote the standard deviation and median respectively. Following Habra et al. (2021), we then create a new filtered coupling matrix $\hat{\Pi} \in \mathbb{R}_+^{p_1 \times p_2}$ given by

$$\hat{\Pi}_{ij} = \begin{cases} \tilde{\Pi}_{ij} & \text{if } r_f(i, j) < \mu_r + r_{\text{thresh}} \\ 0 & \text{otherwise} \end{cases} \quad (54)$$

where $r_{\text{thresh}} \in \mathbb{R}_+$ given filtering threshold. The procedure of estimating the drift function f in (51) and filtering the coupling can be repeated for multiple iterations, to improve the drift and coupling estimation. In our main algorithm, we use two preliminary iterations where we estimate the RT drift and discard outliers with $r_{\text{thresh}} = 5 \times \text{PI}$, defined as points falling outside of the 95% prediction interval. We then re-estimate the drift and perform a final filtering step with the more stringent MAD by setting $r_{\text{thresh}} = 2 \times \text{MAD}$.

At this stage, it is possible for $\hat{\Pi}$ to still contain coefficients of very small magnitude. As an optional postprocessing step, we discard these coefficients by setting all entries smaller than $\epsilon \times \max(\hat{\Pi})$ to zero for some scaling constant $\epsilon \in [0, 1]$. Lastly, a feature from either study could have multiple possible matches, since $\hat{\Pi}$ can have more than one non-zero coefficient per row or column. Although reporting multiple matches can be helpful in an exploratory context, for the sake of simplicity in our analysis, the final output of GromovMatcher returns a one-to-one matching. Consequently, we only keep those metabolite pairs (i, j) where the entry $\hat{\Pi}_{ij}$ is largest in its corresponding row and column. All nonzero entries of $\hat{\Pi}$ which do not satisfy this criterion are set to zero. Finally, we convert $\hat{\Pi}$ into a binary matching matrix $M \in \{0, 1\}^{p_1 \times p_2}$ with ones in place of its nonzero entries and this final output is returned to the user.

As a naming convention, we use the abbreviation GM for our GromovMatcher method, and use the abbreviation GMT when running GromovMatcher with the optional ϵ -thresholding step.

GromovMatcher algorithm summary

In summary, our full GromovMatcher algorithm consists of (1) UGW optimization followed by (2) retention time drift estimation and filtering.

The tuning of ϵ and PI was computationally driven and the two parameters were set as low as possible, with $\epsilon = 0.05$ and $\text{PI} = 0.005$. Based on literature [Lofffield et al. \(2021\)](#); [Hsu et al. \(2019\)](#); [Climaco Pinto et al. \(2022\)](#); [Habra et al. \(2021\)](#); [Chen et al. \(2021\)](#) and what is considered to be a plausible variation of a feature's m/z , we set $m_{\text{gap}} = 0.01 \text{ ppm}$. For RT drift estimation, the order of the B-splines was set to $n = 3$ by default, while the number of knots k was selected by 10-fold cross-validation. If the optional thresholding step was applied in GMT, we set $\epsilon = 0.3$. Otherwise, we let $\epsilon = 0$ which gives the unthresholded GM algorithm.

Appendix 2

Here we discuss existing metabolomic alignments methods and the hyperparameter experiments we perform on these methods. We consider two existing alignment methods for comparison, [metabCombiner Habra et al. \(2021\)](#) and [M2S Climaco Pinto et al. \(2022\)](#). Both of them take the same kind of input as GromovMatcher, i.e. feature tables with features identified with their m/z , RT, and intensities across samples.

Algorithm 4: GromovMatcher

input : Distance matrices D^x, D^y , marginals a, b , marginal relaxation ρ , entropic regularization ϵ , mass-to-charge ratios m^x, m^y , mass-to-charge ratio gap m_{gap} , retention times RT^x, RT^y , B-spline order n , filtering threshold τ

output: Return the matching matrix M and the retention time drift \hat{f}

Step 1: Weighted UGW optimization

Compute $\tilde{\Pi} = \text{WeightedUnbalancedGromovWasserstein}(D^x, D^y, a, b, \rho, \epsilon, m^x, m^y)$

Step 2: Retention time drift estimation and filtering

for $i = 1 : 3$ **do**

 Perform weighted spline regression (51) for RT drift $\hat{f} \in \mathcal{B}_{n,k}$ where k is chosen by 10-fold cross validation

 Initialize $r_{\text{thresh}} = 0$

if $i < 3$ **then**

 └ Set $r_{\text{thresh}} = \text{PI}$ from (53)

else

 └ Set $r_{\text{thresh}} = 2 \times \text{MAD}$ from (53)

 └ Set $\tilde{\Pi} = \hat{\Pi}$

Compute $\mathcal{U} = \max(\tilde{\Pi})$

Set $\hat{\Pi}_{ij} = 0$ if $\tilde{\Pi}_{ij} < \tau \mathcal{U}$ for $i \in [p_1]$ and $j \in [p_2]$

Set $\hat{\Pi}_{ij} = 0$ if $i \neq \text{argmax}_k \hat{\Pi}_{k,j}$ or $j \neq \text{argmax}_k \hat{\Pi}_{i,k}$ for $i \in [p_1]$ and $j \in [p_2]$

Define the binarized matching $M_{ij} = \mathbf{1}_{\{\hat{\Pi}_{ij} > 0\}}$

Return M and \hat{f} .

Algorithm 4

Gromov Matcher

MetabCombiner a novel appearance

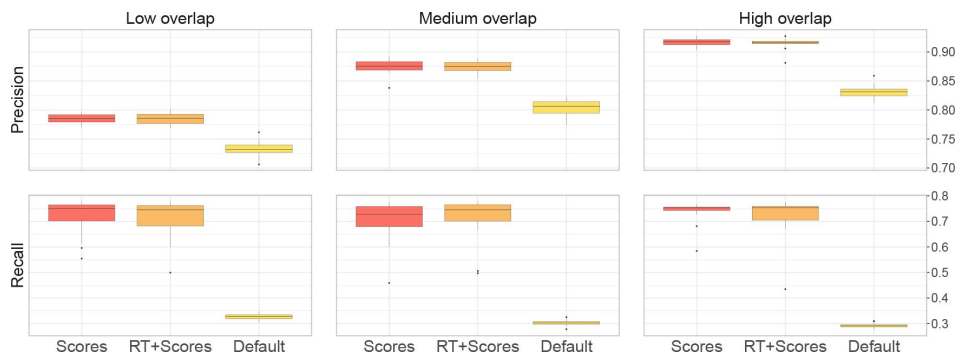
MetabCombiner [Habra et al. \(2021\)](#) is a three-step process that begins by grouping features based on their m/z within user-specified bins. This creates a search space for potential feature pairs. In the second step, MetabCombiner estimates the RT drift using the potential feature pairs identified in the first step, and eliminates outlying pairs over several iterations. This step can incorporate prior knowledge by identifying shared features and marking them as anchors, which are not discarded. In the final step, MetabCombiner scores the remaining feature pairs based on their m/z , RT, and relative intensity compatibility to discriminate between multiple matches for one feature. The scoring system relies on weights assigned to m/z , RT, and feature intensities, with the magnitude of those weights reflecting the reliability of the corresponding measurements across studies.

MetabCombiner [Habra et al. \(2021\)](#) includes adjustable parameters throughout the pipeline. We set most of them to default values unless otherwise stated. MetabCombiner first establishes candidate pairs by binning features in the m/z dimension with a width of `binGap`, and pairing the features sorted by relative intensities. The `binGap` parameter sets the m/z tolerance of metabCombiner, similar to `mgap` in GromovMatcher. We used the same value of 0.01 as in GromovMatcher.

MetabCombiner then estimates the RT drift using basis splines, and removes pairs associated with a high residual (twice the mean model error) from the candidate set.

In our main experiment, the RT drift is estimated exclusively using candidate pairs selected by the pipeline. However, it is also possible to include known ground truth pairs as 'anchors' to estimate the RT drift. We choose not to rely on prior knowledge for drift estimation as [Habra et al. \(2021\)](#) show their drift estimation to be efficient and robust, even without prior knowledge. To confirm this claim, we conduct a sensitivity analysis comparing the results obtained in our main experiment with those obtained when supplying metabCombiner with known shared metabolites to anchor the RT drift estimation. We randomly select 100 anchors from the ground truth matching and compute the metabCombiner matchings with otherwise identical settings as in our main experiment. The results from this analysis (reported in Appendix 2 - Figure 1) show that the unsupervised RT drift estimation (using anchors selected by the pipeline only) performs as well as the supervised RT drift estimation, showing the drift estimation to be very consistent, with or without shared entities.

After establishing candidate pairs and filtering out those that contradict the estimated RT drift, metabCombiner discriminates between multiple matches using a scoring system that considers m/z , RT, and rankings of the median feature intensities. Each dimension has a specific weight that can be left at default, manually adjusted, or automatically tuned using known matched pairs. [Habra et al. \(2021\)](#) provide qualitative guidelines for tuning the weights manually, mainly based on the experimental conditions and visual inspection of the RT drift plot. Since this approach is difficult to implement in the various settings we consider for our simulation study, we rely on the quantitative tuning function included in the metabCombiner pipeline. This function takes into account known shared features and tunes the weights to optimize the scores of those known matches. We randomly select 100 known true matches to define the objective function metabCombiner maximizes. We search over the recommended range of values, with the m/z weight $A \in [50, 150]$, the RT weight $B \in [5, 20]$ and the feature intensities weight $C \in [0, 1]$. Appendix 2 - Figure 1 presents the results obtained with the weights set at default values ($A = 100$, $B = 15$, $C = 0.5$), as a sensitivity analysis.



Appendix 2—figure

Performance of metabCombiner compared to the selected integrator

The first settings explored led to the design of 1000 randomly assembled pairwise pairs supplied to metabCombiner to set the scoring weights. In addition, the integrator was labelled '+Scores', metabCombiner set to 100 true pairs not only to set the estimate the RT drift. Default settings will be of course and RT drift estimate keep the scoring weights' default values.

M2S hyperparameter experiments

Pinto et al. [Climaco Pinto et al. \(2022\)](#) introduce M2S as a more versatile alternative to metabCombiner, while still adhering to most of its core principles. Like metabCombiner, M2S follows a three-step process. First it searches for matches within user-defined thresholds for m/z, retention time, and mean feature intensity. Next, M2S estimates m/z, RT and feature intensity drifts between datasets and removes any outlier pairs. Finally, M2S selects the best match using a scoring system that weighs each measurement, similar to metabCombiner. M2S notably stands out by providing greater flexibility in the methods and measurements used at each step of the procedure, resulting however in a larger number of parameters that require manual fine-tuning. To address this, we adopt two different approaches for the simulation study and the EPIC study alignment. In the simulation study, we set the initial thresholds to oracle values and investigate technical parameters. For the EPIC study alignment, we use the combination of technical parameters with the best average F1-score in the simulation study and select the best threshold values based on the performance on the validation subset.

More precisely, M2S first matches all pairs of metabolic features whose absolute difference in m/z, RT, and median of $\log_{10}FI$ are within the user-defined thresholds 'MZ_intercept', 'RT_intercept' and 'log10FI_intercept'. On simulated data experiments, we set these thresholds to MZ_intercept = 0.01, RT_intercept = 3.5 and log10FI_intercept = 0.2 which are large enough to not exclude any true feature matches in any of the scenarios for our simulated data under low, medium, and high overlap/noise (see Methods). M2S also offers more detailed options to match features whose absolute difference stays within two lower and upper bound lines with a given slope where the intercepts of these lines are defined using the values above. In our analysis, we set the slopes of these linear boundaries to zero so as to not remove any true matches. Because the reference and target studies we are matching in the simulated analysis are on the same scale, we set the FI adjust method to 'none'.

The second step of M2S involves calculating penalization scores for every pair of matches which are used to determine the best set of matches between metabolic features of both datasets. This step depends on a set of hyperparameters which we perform a grid search over to optimize the performance of M2S. For estimating the m/z, RT, and FI drift, the hyperparameters are the percentage of neighbors 'nrNeighbors', the neighborhood shape 'neighMethod', and the LOESS span percentage 'pctPointsLoess' used to smooth the estimated drift functions. After the drifts are estimated, they are normalized using a method specified by 'residPercentile' that puts the m/z, RT, and FI residuals on the same scale. We always fix residPercentile = NaN which defaults to the standard $2 \times MAD$ normalization. Next, for every remaining metabolic feature match, the residuals/drifts of the m/z, RT, and FI are added together by taking the weighted square root sum of squares. For unnormalized data where feature intensity magnitudes are important, we weight all three drifts equally using $W = (1, 1, 1)$ and for data with normalized feature intensities we set the FI drift weight to zero such that $W = (1, 1, 0)$. Finally, using these weighted penalization scores, M2S selects the best matched pair within a multiple match cluster to obtain a one-to-one matching between datasets.

The third and final step of M2S involves removing those remaining matches which have large differences in m/z, RT, or FI. This can be performed using several methods indicated by the hyperparameter 'methodType'. Each method excludes those matched pairs whose differences in m/z, RT, or FI exceed a certain number of median absolute deviations indicated by the parameter 'nr-MAD'. The remaining one-to-one metabolic feature matches are returned as the final result of the M2S algorithm.

| Metric | Low overlap | Medium overlap | High overlap |
|-----------|-------------|----------------|--------------|
| Precision | 0.831 | 0.917 | 0.947 |
| Recall | 0.934 | 0.933 | 0.939 |

Appendix 2—table 1.

Performance of M2S in the setting where between studies is linear

To optimally tune M2S on our simulated experiments, we determine the optimal M2S parameter combination for each individual simulation setting (low, medium, high overlap and noise) by performing a grid search over the product of parameter lists

- nrNeighbors = [0.01, 0.05, 0.1, 0.5, 1]
- neighMethod = ['cross', 'circle']
- pctPointsLoess = [0, 0.1, 0.5]
- methodType = ['none', 'scores', 'byBins', 'trend_mad', 'residuals_mad']
- nrMAD = [1, 3, 5]

Each parameter combination for M2S is tested across 20 randomly generated datasets at the same overlap and noise settings. For each setting, the combination of parameters above with the best average F1-score across these 20 trials is used as the optimal parameter choice.

M2S applies initial RT thresholds to search for candidate pairs, which may favor settings where the RT drift follows a linear trend. Therefore, as a sensitivity analysis, we apply M2S to simulated data with a linear drift. The simulation process is identical to that of our main simulation study, except for the deviation of the RT in dataset 2. Specifically, for a given overlap value, we divide the original real-world dataset into two smaller datasets and introduce random noise to the m/z, RT and intensities of the features, without introducing a systematic deviation to the RT in dataset 2. M2S parameters are kept identical to the ones used in our main analysis in comparable settings. The results obtained by M2S on three pairs of datasets generated for three overlap values (0.25, 0.5 and 0.75) and a medium noise level are reported in [Appendix 2 - Table 1](#). While the results obtained in a high overlap setting are close to those obtained in our main analysis M2S demonstrates better performance in a low overlap setting when the RT drift is linear than in our main analysis. This observation is consistent with the results obtained by M2S on EPIC data, considering the relatively low estimated overlap between the aligned EPIC studies in our main analysis.

For the EPIC data, we select the parameter combination that yields the highest F1-score across all simulated settings. However, due to the unavailability of oracle values for setting initial thresholds, we perform a search over several MZ intercept values (0.01, 0.05, and 0.1), RT intercept values (0.1, 0.5, 1, and 5), and logFI intercept values (1, 10, and 100).

Appendix 3

In this section, we study the sensitivity of all three alignment methods GMT, M2S, and mC to the validation dataset split when creating two validation studies for matching. As described in the section "Validation on ground-truth data" and depicted in [Fig. 2](#) of the main text, we generate two datasets to be matched by splitting an initial LC-MS dataset with p features and n samples into two smaller overlapping datasets. The first dataset has p_1 features and n_1 samples while the second dataset has p_2 features and n_2 samples. The sets of samples in both datasets are disjoint such that $n_1 + n_2 = n$. However, the dataset split is constructed such that both datasets share λp of their features where $\lambda \in [0, 1]$ is an overlap fraction. Namely, this is done by defining the dataset feature sizes as

$$p_1 = \left\lfloor \left(\lambda + \lambda_f(1 - \lambda) \right) p \right\rfloor, \quad p_2 = \left\lfloor \left(\lambda + (1 - \lambda_f)(1 - \lambda) \right) p \right\rfloor \quad (55)$$

and the dataset sample sizes as

$$n_1 = \lfloor \lambda_s n \rfloor, \quad n_2 = n - \lfloor \lambda_s n \rfloor. \quad (56)$$

As before, $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ denote integer floor and ceiling functions. Then taking the original LC-MS dataset and randomly permuting its samples and features, the first p_1 features and first n_1 samples are placed into dataset 1 while the last p_2 features and last n_2 samples are placed into dataset 2. It is indeed easy to check here that with such a splitting procedure, the feature overlap between both datasets is $p_1 + p_2 - p$.

Here $f \in [0.5, 1]$ controls the fraction of features in dataset 1 that is not shared with dataset 2 and $s \in (0, 1)$ controls the fraction of samples in dataset 1 vs. dataset 2. In particular, if $f = 1$ then the features in dataset 2 are entirely a subset of those in dataset 1. In the experiments described in the main text, we always set $f = s = 0.5$ as to balance the number of features and samples in both resulting datasets.

Now we study how the performance of all three alignment methods changes when f , s and the feature overlap are varied. Here we vary the feature overlap $\in \{0.25, 0.5, 0.75\}$, the feature fraction $s \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, and the sample fraction $s \in \{0.1, 0.2, \dots, 0.9\}$. In Figs. 1, 2, 3 we show how the precision and recall of GMT, M2S, and mC depend on these parameters. Here we use the same unnormalized validation data and experimental setup as described in the main text section "Validation on ground-truth data" and in the Methods and Materials section "Validation on simulated data". For each triple (\cdot, f, s) we randomly generate 20 dataset splits with these parameters and show the average precision and recall for each method over these trials. Our method GMT (thresholded GromovMatcher) is applied out-of-box with the default hyperparameter settings. The algorithm hyperparameters for mC and M2S are chosen optimally for each individual triple of dataset parameters (\cdot, f, s) to maximize the average F1 score in each setting. The hyperparameters searched over when optimizing mC and M2S are described in detail in Appendix 2.

Consistent with prior validation experiments, we find that GromovMatcher outperforms both mC and M2S in all dataset regimes, for low overlap and high overlap as well as for varying balances of features f and samples s . Remarkably, all three methods exhibit the same sensitivity to variations of (\cdot, f, s) . All methods exhibit a monotonic decrease in their precision as f drops from 0.9 to 0.5. In other words, the most challenging setting for matching both datasets is when dataset 1 and dataset 2 both have an equal number of unshared features (e.g. $f = 0.5$). Likewise, the simplest setting for matching is when the features in dataset 2 are exactly a subset of the features in dataset 1 (e.g. $f = 1$). Sensitivity to this parameter f is most noticeable at low feature overlap $= 0.25$.

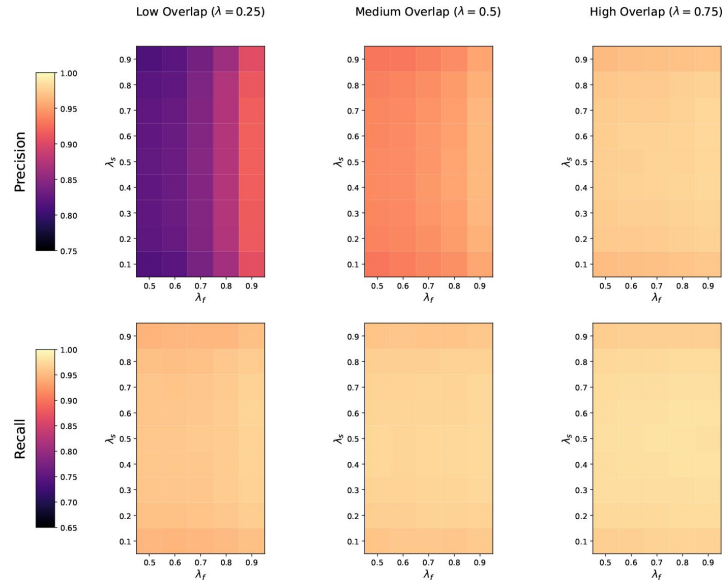
Appendix 4

Here we describe additional preprocessing details and analyses of the EPIC data.

Centered and scaled data - Negative mode

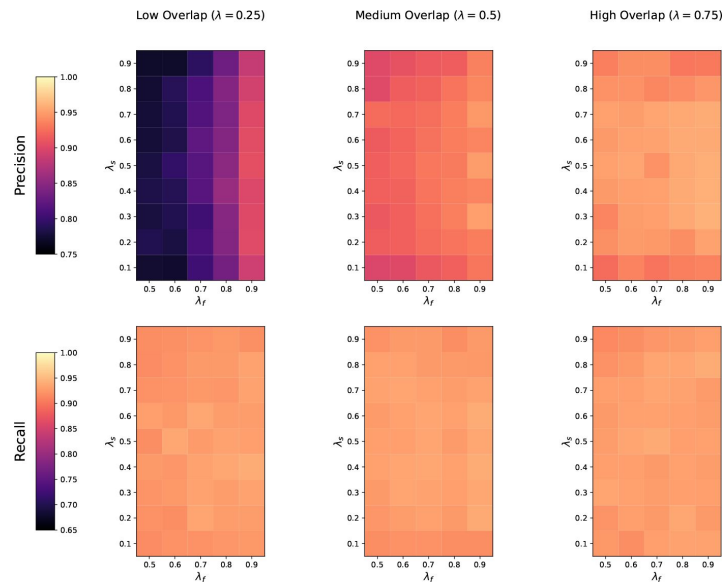
In this section, we present the results obtained on centered and scaled EPIC data in negative mode, shown in Figure 4 of our main paper. However, due to the smaller size of the validation subset (42 features examined in negative mode compared to 163 in positive mode), the evaluation of the performance of the three methods may be less reliable than in positive mode.

First, we align the CS and HCC studies in negative mode and detect a total of 449, 492, and 180 matches with GM, M2S, and metabCombiner, respectively. Similar to the positive mode analysis, we evaluate the precision and recall of the three methods on the 42 feature validation subset, of which 19 were manually matched. GM and M2S demonstrate identical F1-scores of 0.98, while metabCombiner performs poorly in comparison. GM is able to recover all 19 true matches and identified only 1 false positive, while M2S recovers no false positives but missed 1 true positive.



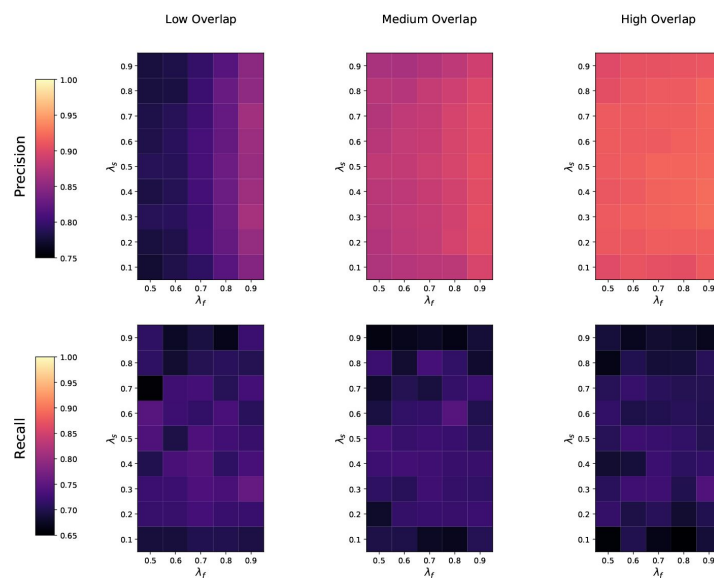
Appendix 3—Figure 1

Sensitivity of the Matcher (GMT) to different feature imbalance ratios and sample imbalances between two datasets being matched.



Appendix 3—Figure 2

Sensitivity of the Matcher to different feature imbalance ratios and sample imbalances between two datasets being matched.



Appendix 8—Bigur

Sensitivity of meta-bCombiner (a function of λ_T and λ_B) to the combination of λ_T and λ_B for different combinations of λ_T and λ_B and sample size n were determined by matching two datasets being matched.

Next, we align the CS and PC studies in negative mode and detect a total of 485, 569, and 314 matches with GM, M2S, and metabCombiner, respectively. Again, we evaluate the precision and recall of the three methods on the 42 feature validation subset, of which 26 were manually matched. MetabCombiner performs better than in the other EPIC pairings with an F1-score of 0.857, but is still outperformed by the other two methods. GM is slightly outperformed by M2S in this setting, with an almost identical precision of 0.93, but a slightly higher recall for M2S due to detecting 1 additional true positive. However, this remains a good performance for GM since M2S was optimally tuned using the validation subset itself.

Non-centered and non-scaled data

As a sensitivity analysis, we apply the three methods to EPIC data that has not been centered or scaled. The detailed results can be found in Appendix 4 - Table 1 [↗](#).

M2S was tuned manually on the validation subset to ignore feature intensities in both cases. As a result, it maintains its performance compared to our main experiment. On the other hand, the performance of GM and metabCombiner is affected by the lack of consistency in feature intensities. MetabCombiner's recall drops slightly but its precision remains comparable to that of our main experiment, with the method clearly favoring the latter. Although GM's recall decreases slightly in positive mode, it remains more precise than the optimally tuned M2S, and it balances precision and recall better than metabCombiner. Interestingly, GM's results in negative mode are improved compared to our main experiment, and it outperforms both mC and M2S. However, since the validation subset in negative mode is relatively small, these differences may not be significant. Nonetheless, GM maintains a good performance, similar to that of the optimally tuned M2S.

Similar to the analysis we conducted on centered and scaled data, we find a high number of false positives when aligning the CS study and the PC study in positive mode. Therefore, we manually examine the matches recovered by GM. Our examination reveals 2 false positives, 4 unclear matches, and 3 additional good matches that GM also identifies in our main analysis. This demonstrates that the lack of centering and scaling results in two additional false positives for GM that are not present in our main results.

Illustration for alcohol discovery mark

Lofffield et al. [Lofffield et al. \(2021\) ↗](#) identified 205 features associated with alcohol intake in the CS study, using a false discovery rate (FDR) correction to account for multiple testing. By applying an FDR correction in our pooled analysis, we identify 243 features associated with alcohol intake. Out of those 243 features, 185 are consistent with the features identified in the discovery step of Lofffield et al. [Lofffield et al. \(2021\) ↗](#), while 55 features are newly discovered (Fig. 5c [↗](#)). We examine the 20 features identified as significant in Lofffield et al.'s discovery analysis but that are not significant in our pooled analysis. Both manual and GM matching yield identical results for these features, indicating that the loss of significance is not due to incorrect matching. Upon further investigation, we find that these features do not demonstrate a meaningful association with alcohol intake in the HCC and PC studies. This observation is reinforced by the fact that none of these features are among the 10 features that persisted after the validation step in Lofffield et al.

Out of the 205 features initially discovered in Lofffield et al. [Lofffield et al. \(2021\) ↗](#), 10 are replicated in the EPIC HCC and PC studies using the more stringent Bonferroni correction. When using a Bonferroni correction in our pooled analysis, we find significant association between alcohol intake and 92 features, 36 of which are effectively shared by the three studies. Notably, these features include all 10 features that were retained in Lofffield et al. (Fig. 5c [↗](#)).

| Method | CS ↔ HCC | | CS ↔ PC | |
|----------------------|----------------------|----------------------|----------------------|----------------------|
| | Precision | Recall | Precision | Recall |
| GromovMatcher | 0.988 (0.937, 0.999) | 0.944 (0.876, 0.997) | 0.873 (0.776, 0.932) | 0.939 (0.854, 0.976) |
| M2S | 0.967 (0.908, 0.991) | 0.978 (0.923, 0.996) | 0.855 (0.759, 0.917) | 0.985 (0.919, 0.999) |
| metabCombiner | 0.979 (0.889, 0.999) | 0.511 (0.410, 0.612) | 0.926 (0.766, 0.987) | 0.379 (0.271, 0.499) |

(a) Positive mode

| Method | CS ↔ HCC | | CS ↔ PC | |
|----------------------|----------------------|----------------------|----------------------|----------------------|
| | Precision | Recall | Precision | Recall |
| GromovMatcher | 0.950 (0.764, 0.997) | 1.000 (0.832, 1.000) | 0.964 (0.823, 0.998) | 0.964 (0.823, 0.998) |
| M2S | 1.000 (0.824, 1.000) | 0.947 (0.754, 0.997) | 0.931 (0.780, 0.988) | 0.964 (0.823, 0.998) |
| metabCombiner | 1.000 (0.566, 1.000) | 0.263 (0.118, 0.488) | 1.000 (0.785, 1.000) | 0.500 (0.326, 0.674) |

Appendix 4—table 1.

Precision and recall on the EPIC validation subset for unnormalized data in (a) positive mode, and (b) negative mode. 95% confidence intervals are reported using modified Wilson (1998) and

This analysis illustrates how GromovMatcher can be used in the context of biomarker discovery, and its potential to allow for increased statistical power.

Appendix 5

Here we investigate how the choice of the reference dataset influences the discovery of metabolites shared across the CS, HCC and PC EPIC studies by GromovMatcher. All three methods considered in this paper, GromovMatcher, M2S, and metabCombiner, are limited to the comparison of two datasets. However, they can still be used to compare and pool multiple datasets using a multi-step procedure. Namely, this can be done by designating a 'reference' dataset and aligning all studies to it one by one. We take this exact approach in our analysis when aligning the CS, HCC, and PC studies of the EPIC data in positive mode. Namely, the HCC and PC studies are both aligned to the CS study (see main text Fig. 5b [↗](#)). However, this method raises two critical questions: (i) how does the use of a reference dataset affect matching results, and (ii) how is the matching affected by the choice of reference dataset.

To address these questions, we compare the features identified as common to the three studies using two different studies as references: the CS study used as reference in the main analysis, and the HCC study. For simplicity, let's denote M_{s_1, s_2} the matching matrix obtained when aligning study 1 and study 2.

Changes in results when the reference dataset is used

Concerning question (i), we compare two matchings: HCC to CS to PC (the matrix product $M_{CS, HCC}^T M_{CS, PC}$) which we will refer to as the reference matching, and the direct matching of PC to HCC ($M_{HCC, PC}$). Note that these matchings are not fully comparable as the former considers only features found in CS, potentially missing unique HCC and PC matches. We can however compare the two matchings on the subset of 706 features common to all three studies, as determined by the reference matching. We find that the direct matching supports 683 out of them, indicating that the matching via a reference still yields good results compared to the direct matching (see Appendix 5 [↗](#) Fig. 1 [↗](#)).

Effect of reference dataset choice on matching

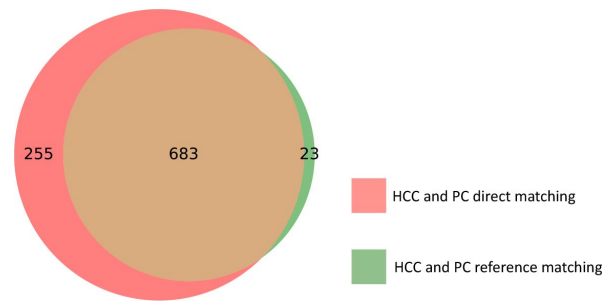
Concerning question (ii), we compare the features identified as common to the three studies using two different studies as references: the CS study used in the paper, and the HCC study. We find that they identify 706 and 708 common features respectively, with an overlap of 640 features (see Appendix 5 [↗](#) Fig. 2 [↗](#)). This highlights that the choice of reference does matter to some extent. In the paper, choosing CS as a reference was informed by CS's sample size, and study population.

References

Agresti, C. A. (1993). Approximate is better than interval estimation for binomial proportions. *Statistical Science*, 18(2), 119–126. [DOI: 10.1214/ss.1993.18.2.119](#)

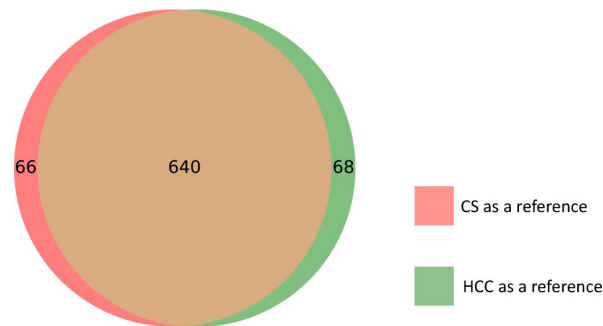
et al. (2020). Multi-omic analysis of blood metabolites in underlying metabolic disease. *Nature*, 584(7821), 1–6. [DOI: 10.1038/s41586-020-2192-2](#)

Alvarez Mel, J. & Koka (2021). Gromov-Wasserstein Alignment in Defining Space. *EMNLP*: 1881–1890. [DOI: 10.18653/v1/2021.emnlp-main.148](#)



Appendix 5—figure

Overlap between the 700 of the HCC studies identified via direct matching, and the 938 of the HCC studies identified via reference matching.



Appendix 5—figure

Overlap between the 100 identified as common to the three studies using either Chestnut or the HCC.

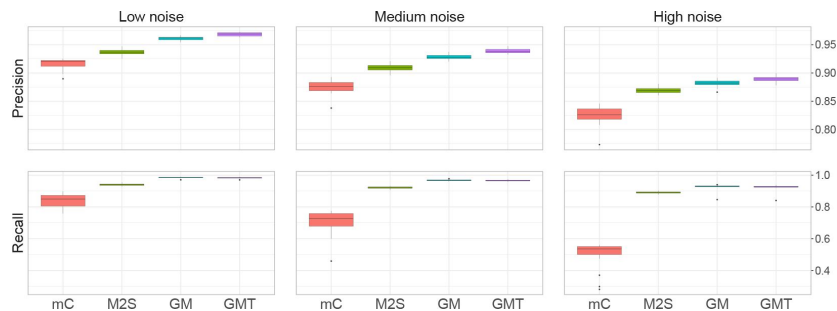
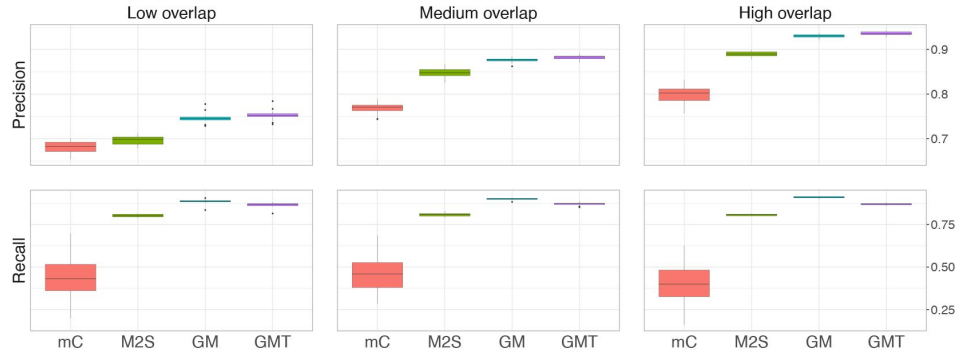


Figure 3—figure supplement 1.

Average precision can be obtained on simulated data with fixed noise levels of Low, medium and high, and on the 100 (0.2, 0.05, 0.15), respectively, 200 simulations for each setting.

a. Overlap influence



b. Noise influence

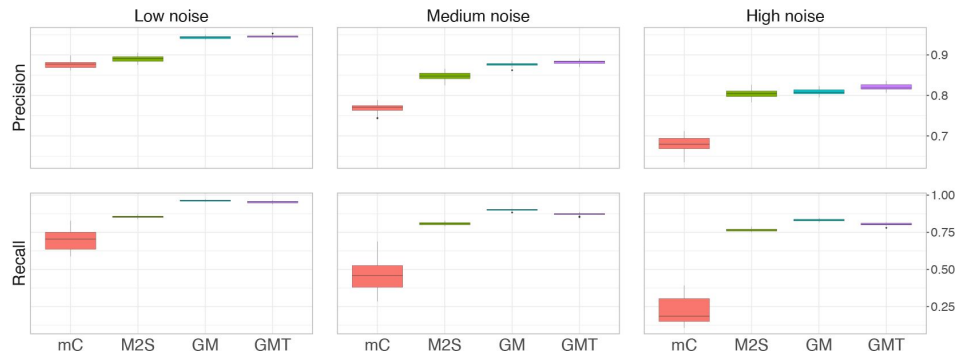


Figure 3—feature supplement 2.

Performance on scaled data.

The feature intensities of each dataset are standardized to have mean 0 and standard deviation 1. The average precision and recall of the methods are computed on 2000 random pairs of datasets, for (a) levels of overlap (medium and strong) of 0.5 and 0.75, respectively, with a precision (medium noise) $(F_1) = (0.5, 0.8)$ and a recall (low) of 0.5, respectively, and (b) levels of overlap (medium and strong) of 0.5 and 0.75, respectively, with a precision (medium noise) $(F_1) = (0.2, 0.1)$, respectively, and (1, 1), respectively.

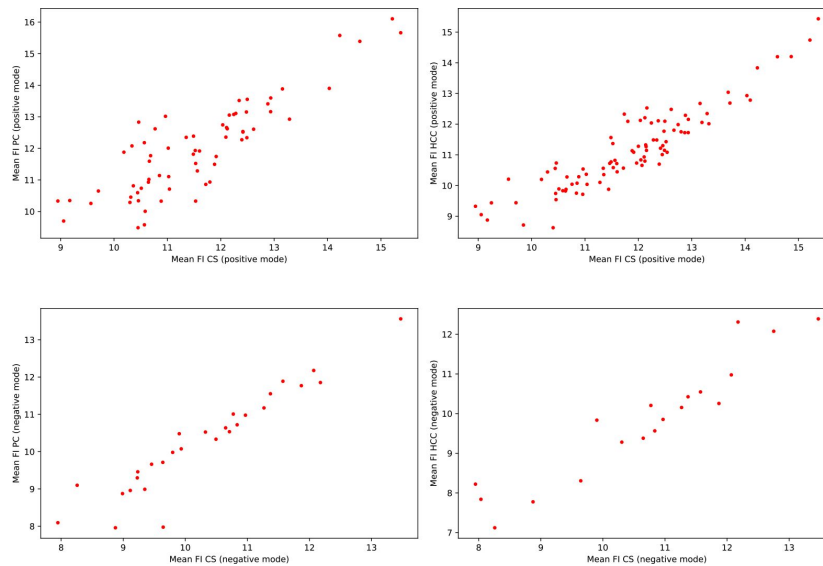
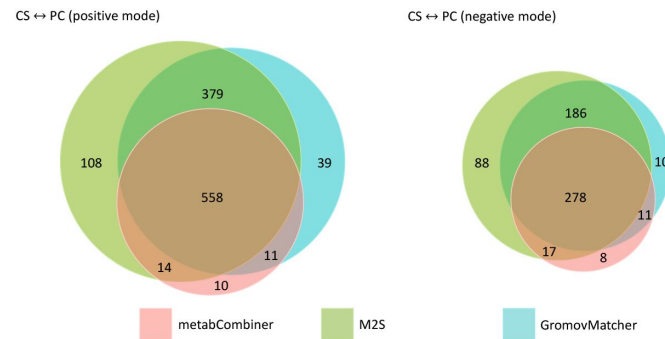


Figure 4—fig supplement 1.

Consistency of the mean FI scores in EPI C.

Each scatter plot represents the mean FI scores of a pair of modalities (PC or HCC) in a subset of the data. Each point represents a pair of modalities (PC or HCC) in a subset of the data. The x-axis represents the mean FI score in the first study, and the y-axis represents the mean FI score in the second study.

a. Matching results obtained for the cross-sectional (CS) and hepatocellular carcinoma (HCC) studies, in positive and negative ionization mode.



b. Matching results obtained for the cross-sectional (CS) and pancreatic cancer (PC) studies, in positive and negative ionization mode.

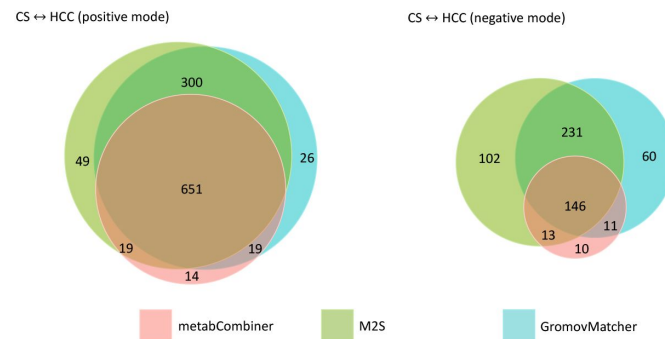


Figure 4—fig supplement 2.

Overlap between results obtained by metabCombiner, M2S and GromovMatcher in RPLASG is not up to scale.

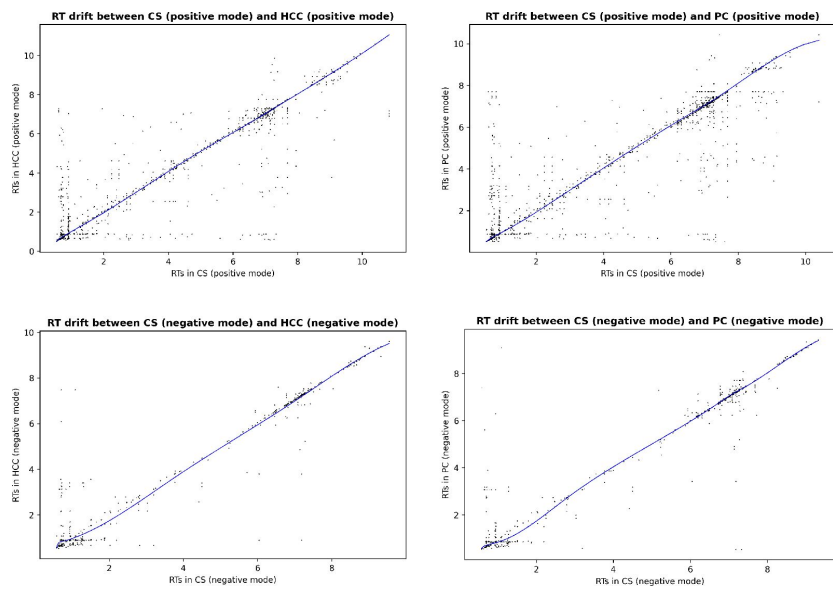


Figure 4—figure supplement 3.

Estimated RT drift between the EPXPC sites aligned in the
 Each dot corresponds to a candidate matched pair from the GWAS in the
 drift estimation and RT based filtering.

Alvarez-Mel, J, Segal, J, Kulkarni (2017) Towards optimal growth with global invariance. *PLoS Comput Biol* 13(10): e1005470. doi:10.1371/journal.pcbi.1005470

Bedi, A, Metzger, M, Kishimoto, A, et al. (2016) Metabolomic network biology: Applications and Trends. *Emerging Infectious Diseases* 22(10): 1611-1618. doi:10.1186/s12874-016-0261-1

Beier, B, Frenkel, R, et al. (2017) Multi-omic analysis of user taste and baroque nature. *PLoS One* 12(10): e0182056. doi:10.1371/journal.pone.0182056

Brown, L, Datta, G, et al. (2017) Interval Estimation for Sparse Data. *Biometrics* 73(1): 101-113. doi:10.1111/biomet.12886

Bruni, S, Shi, C, and others (2016) Large-scale metabolomics data collection using between-label network and cluster-based intensity correction. *Metabolites* 6(1): 1-14. doi:10.3390/met6010001

Chen, L (2017) Metabolic diversity in global annotation of untargeted metabolomics data. *Metabolites* 7(1): 1-13. doi:10.3390/met7010001

Chizari, P, Schmitt, W, and others (2017) Unbalanced dynamic and kinetic formulation. *PLoS One* 12(10): e0182056. doi:10.1371/journal.pone.0182056

Climaco, P, et al. (2017) Finding correspondence between metabolomic Untargeted Lipidomics and Mass Spectrometry Metabolomic Data. *ChemRxiv* 1510.03411v1. doi:10.26434/chemrxiv-2017-15100

Court, F, Lima, H, and others (2017) Joint distribution of transport for domain adaptation. *PLoS One* 12(10): e0182056. doi:10.1371/journal.pone.0182056

Demetriou, P, et al. (2017) Single-Cell Multi-Omic Alignment via Shared Information. *PLoS One* 12(10): e0182056. doi:10.1371/journal.pone.0182056

et al. (2017) Gut microbiome serum and metabolic activity in disease. *Metabolites* 7(1): 1-13. doi:10.3390/met7010001

et al. (2017) Methodological issues in the analysis of persister cell populations and their risk within the environment. *PLoS One* 12(10): e0182056. doi:10.1371/journal.pone.0182056

Gomari, S, Dweic, K, et al. (2017) A multi-omic approach to the study of variation in auto-antibodies in patients with rheumatoid arthritis. *PLoS One* 12(10): e0182056. doi:10.1371/journal.pone.0182056

Gromov, V (2007) Metric Structures for Riemannian and Non-Riemannian Geometry. *Journal of Mathematical Sciences* 142(6): 803-833. doi:10.1007/s12220-007-9111-1

Harbar, M, Acharya, B, Miki, S, et al. (2017) MetabCombiner: Paired Untargeted-Targeted Metabolomic Data Integration and Concatenation. *PLoS One* 12(10): e0182056. doi:10.1371/journal.pone.0182056

Hsu, Yeh H (2019) PAIRUP-MS: Pathway analysis and winnowing of protein abundance profiles using mass spectrometry-based metabolite abundance data. *BMC Bioinformatics* 20:1-12. doi:10.1186/s12859-019-1673-4

Ivanis, Vahid (2019) From Samples to Insights in Metabolomics: A Review of Biological and Clinical Applications. *Metabolites* 9:1-12. doi:10.3390/met9120308

Kantowich, David (2019) The Evolution of Metabolomics: From Clinical Chemistry to Systems Biology. *Metabolites* 9:1-12. doi:10.3390/met9120308

Li, Zhen, Xu, Yiqi, Li, Yan, Wu, Hui, Xian (2019) Metabolomics-Based Discovery of Molecular Targets in a Novel Natural Product in Asian Populations. *Frontiers in Molecular Biosciences* 6:1-12. doi:10.3389/fmolb.2019.00058

Liu, Walek, Datta, Iliana, Tran, Livi, Jones, Yu D (2019) Addressing the Bias in Metabolomics Data: A Review of Statistical Methods. *Frontiers in Molecular Biosciences* 6:1-12. doi:10.3389/fmolb.2019.00050

et al. (2019) Novel biomarkers for hepatocellular carcinoma and pancreatic and liver cancers and their clinical significance. *Journal of Cellular Biochemistry* 155:1-12. doi:10.1002/jcb.23678

Mérol, Grom - Wasserstein Distances and the Optimal Transport Problem. *Mathematics* 7:1-12. doi:10.3390/math7010093

Monge (1781) Mémoire sur la théorie des courbes à double courbure. *Annales de Chimie et de Physique* 17:1-12. doi:10.1016/j.cpl.1781.01.004

Nitz, Kamil, Friedman, Rajiv (2019) Genexpression analysis using deep learning. *Journal of Cellular Biochemistry* 155:1-12. doi:10.1002/jcb.23673

Patti, Sepati, et al. (2019) The effect of the environment on the development of the brain. *Journal of Cellular Biochemistry* 155:1-12. doi:10.1002/jcb.23660

Peyé, Gutur, Solim (2019) Gromov-Wasserstein distance and its application in matrix completion. *Journal of Cellular Biochemistry* 155:1-12. doi:10.1002/jcb.23671

et al., Peyé, Gutur (2019) Computational approaches for the analysis of high-dimensional data. *Journal of Cellular Biochemistry* 155:1-12. doi:10.1002/jcb.23673

Pirhalla, Liei, Cucun, Atvilia - Pachiesch, Wh, E, S, H, A, E (2019) Revealing disease-associated pathways in metabolomics. *Journal of Cellular Biochemistry* 155:1-12. doi:10.1002/jcb.23640

Rappaport, Bar, Sump, Wi, Dika, V, N, B, S, S, A, P, B (2019) The Blood Exposure Levels of Various Chemicals in the Human Body. *Journal of Cellular Biochemistry* 155:1-12. doi:10.1002/jcb.23615

Reuthert (2019) Inactive components in the 1000 machine learning and data analysis. *Journal of Cellular Biochemistry* 155:1-12. doi:10.1002/jcb.23629

et al. (2009) European Prospective Investigation into Cancer study populations and data. *Health and Nutrition Research* 10:1079-1084. <https://doi.org/10.1002/1465-1959.nh2002394>

et al. (2019) Optimal support analysis of sparse high-dimensional data. *Journal of the Royal Society Series B* 17:6928-6943. <https://doi.org/10.1098/rspb.2019.0106>

Séjour, F, & Vignat, A. (2019) Sinkhorn distances for unbalanced optimal transport. *arXiv preprint arXiv:1905.05488*. <https://arxiv.org/abs/1905.05488>

Sejour, M, & Vignat, A. (2021) The Unbalanced Wasserstein Distance Formulation and Relaxation. *SIAM Journal on Matrix Analysis and Applications* 62:717-739. <https://doi.org/10.1137/20M1102958>

Skoaczynski, G, & Miś, A. (2021) Alignment: a novel approach to LC-MS retention time prediction. *Journal of Cheminformatics* 13:1-10. <https://doi.org/10.1186/s13065-021-00101-0>

Slama, J, & G. (2003) Groupwise validation of peak identification recall and dietary intake assessment in a prospective investigation into cardiovascular mortality. *Epidemiology and Infection* 131:784-795. <https://doi.org/10.1017/S0950268803001010>

Smith, W, & T. (2006) MS: Processing Mass Spectrometry Data for Metabolite Identification and Quantification. *Journal of Proteomics* 9:1051-1064. <https://doi.org/10.1016/j.jprot.2006.05.003>

Solomon, G, & M. (2019) Bipartite Metric Alignment for Coarse Grained Protein-Protein Interactions. *Journal of Machine Learning Research* 20:5903-5923. <https://doi.org/10.48550/jmlr.2019.20.5903>

et al. (2016) Alteration of amino acid and biogenic amine hepatobiliary cancer risk. *Journal of Hepatology* 64:360-368. <https://doi.org/10.1016/j.jhep.2015.11.018>

et al. (2021) Metabolic perturbations in primary cholangiocarcinoma. *Journal of Hepatology* 74:1048-1058. <https://doi.org/10.1016/j.jhep.2021.05.036>

Tautenhahn, R, & G. (2019) Metabolite identification in XCMS: a metabolite identification pipeline for metabolomics. *Journal of Proteomics* 20:836-846. <https://doi.org/10.1016/j.jprot.2019.02.009>

Vaughan, D, & A. (2019) Metabolite identification in metabolomics. *Journal of Proteomics* 20:836-846. <https://doi.org/10.1016/j.jprot.2019.02.009>

Villa, A. (2021) Topics in applied mathematics. *Mathematical Society of America*. <https://doi.org/10.1090/S0025-5718-2021-0002618>

Wang, T. (2019) Metabolite identification and metabolite identification. *Journal of Proteomics* 20:836-846. <https://doi.org/10.1016/j.jprot.2019.02.009>

Wishart, D. (2019) Metabolomics for Investigating Physiology. *Journal of Proteomics* 20:836-846. <https://doi.org/10.1016/j.jprot.2019.02.009>

Yang, D, Dmo daanr, Vē n k a t a c h a l, S a g l a t z n o g I S u s h a i v a s h a t h k l a e r G C .
(2024) Predicting cell lineages using a deep generative model. *eLife* 13:e100782. doi:10.7554/eLife.100782

Zhou, X, B a o T u J F R e l s s o m (2024) LC - MS - based method for identifying and quantifying metabolites. *eLife* 13:e100782. doi:10.7554/eLife.100782

Article and author information

Marie E Burr

Nutrition and Metabolism, Boston Children's Hospital, Boston, Massachusetts, United States

Georgy Stepaniants

Massachusetts Institute of Technology Department of Mathematics, Cambridge, Massachusetts, United States

Pekka Keskinen-Rahk

Nutrition and Metabolism, Boston Children's Hospital, Boston, Massachusetts, United States

Philippe Rigollet

Massachusetts Institute of Technology Department of Mathematics, Cambridge, Massachusetts, United States

Vivian Viallon

Nutrition and Metabolism, Boston Children's Hospital, Boston, Massachusetts, United States

For correspondence: viallon@bwh.harvard.edu

Copyright

© 2024 Burr et al.

This article is distributed under a Creative Commons Attribution License, which permits unrestricted use and redistribution in any medium, provided the original author and source are credited.

Editors

Reviewing Editor

Paula Fernandez

INAT, Buenos Aires, Argentina

Senior Editor

Murim Choi

Seoul National University, Seoul, Republic of Korea

Reviewer #1 (Public Review):

Summary:

The authors have implemented a novel method for comparing LC/MS features from different datasets.

proteomics field for performing meta-analysis of

Strengths:

The main strength is that GromovMatcher achieves s compared to other existing methods. The authors h that GromovMatcher performs well.

Weaknesses:

The authors might need to add the limitation of da tool using simulatædt dæst awênl the abstr

<http://doi.org/10.7554/eLife.91597.a1>

Reviewer #2 (Public Review):

Summary

The goal of untargeted metabolomics is to identify different biological samples.

Untargeted metabolomics identifies fætaitur(ers/ zw) i thn d retention)ti Met (R) ing those to specific metabolite from databases is layso rpi osussi balned, y nveit d h d ias f wlr comp samples on the level of unmatched features are cr The main purpose of the GromovMatcher method pres untargeted metabolomes from different experiments used to advance biological analyses, for example, markrs.

The main problem that complicates merging d i f f e r e slightly for the same feature (metabolite).

The main idea behind the GromovMatcher is built o match between two datasets (that feature i from d and feature k from dataset 1 matches feature l fr distances between the two features within each of similar. The authors b s e r s e i t h e n e G h o m o v o W f i n d t h matrix from these data.

The variation in m/z between the same features in value and it is initially set to T O d e o v l i a p t i m o n s t , h e s r o e t estimates a non-linear Tr b d e e t v w e a e t n i o t n w o (d s r t i u f d t i) e s o f G R o m o v the drift between two studies, and then discards deviate significantly from the estimate. It learn

The authors validate the performance of their Gro cord blood. They use 20 different splits and comp GMT i a t t e r o n s , y w G M T e b u s e s t h e d e v i d a t i f d n t f o r d m l matching matrix) with two other matching methods:

The second validation was done using a (scaled an cancer datasets from the EPIC co h y o r a t n t e h x a p t e r v t e r e T h m i a s was also used to show that using automated method associated with a particular group m a n s a l m p m a s c h i n a Specific, i d h e l y a u t h o r s i d e n t i f y a d d i t i o n a l f e a t u r e s c

Strengths:

I see the main strength of this work in its combination of first- and higher-order information on correlations between features. The use of information that is appropriate for the measure, using the Gromov-Matcher method to match the features

The authors of the paper identify two main shortcomings of methods that attempt to match features from different datasets: they require fine-tuning a model for each dataset, and they do not take into account correlations between features. The main strength of the paper is the information on distances between the features.

Weaknesses:

The main weakness is that there is no data that can be used for validation. It will, therefore, be important to generate data for validating and improving their methods in the future.

The second weakness is that the experimental setup is not clear. The authors set up now can be directly used only to compare their method with other methods. The authors will successfully implement novel algorithms in the future.

<https://doi.org/10.7554/eLife.91597.2>

Author Response

The following is the authors' response to the original

eLife assessment

This paper presents important findings when identifying the difference between the two datasets. The authors' assumption of featuring and matching complex data reflects a solid methodology for validation with datasets.

We are grateful to the eLife editor for taking the time

We are however unsure of why the results are not as good as they could be. The datasets used for validation are scarce. Most public datasets do not have sufficient information to establish a ground truth or a reference. Combiner can be tested. Even in the case where a ground truth can be established, it is extremely time-consuming and requires very specific expertise. This highlights the need for automatic alignment methods like Gromov-Matcher.

We do agree that the performance of Gromov-Matcher needs to be validated further, and we plan to continue validation as soon as the method becomes available in EPIC and other cohorts. With the validation data we have, we can see which an external validation

more comprehensive than previous validations, experts believe reflect real-life data (arXiv:1803.09833). We would like to stress that this allows us to highlight previously published methods, MetabCombiner and M2S.

We wish to thank the editor and reviewers for their comments on the manuscript which led to the following additions to our paper:

- Performed an additional sensitivity analysis (A) the number of features or samples between two studies affects the quality of the match found by MetabCombiner.
- Investigated how changing or removing the reference study (metabolic data to EPIC data), affects the results.
- Improved alignment matrix visualizations in Figure 3 validation data, to highlight more clearly which metabolites were missed.

The revised paper is uploaded as a pre-proof. Changes are highlighted in blue. New additions are not highlighted.

Summary:

The authors have implemented a new method for comparing metabolomic datasets. This paper gains the metabolomics field for performing meta-analysis of metabolomic data.

Strengths:

The new method, Match, achieves significant performance compared to other existing methods. The authors have shown that Match performs well.

Weaknesses:

There are two weaknesses.

- (1) When the number of metabolites is small ($n < 10$), the precision drops to ~ 0.8 .

We would like to clarify that this drop in precision occurs only for a small proportion of metabolites. The proportion of shared metabolites between datasets is 25%. In our metabolomic datasets, only 25% of the features are shared. The automated matching method used in the validation data, as the vast majority (75%) of metabolites remain unmatched.

In such settings, the reviewer correctly observes that the performance of the algorithms (GM and Match) drops (Figure 3). Such a precision of 0.8 or larger is still competitive with MetabCombiner (mC) and M2S whose precisions drop in the validation panel).

Summary:

The authors have implemented a new tool for comparing metabolite datasets. This paper gains the metabolomics field for performing meta-analysis of

Strengths:

The new tool, **Match**, achieves significant performance compared to other existing methods. The authors have claimed that **Match** performs well.

Weaknesses:

There are two weaknesses.

- (1) When the number of shared features drops to ~ 0.8 .

We would like to clarify that this drop in precision occurs only for a small proportion of metabolites. The proportion of shared features - was 25% in our metabolite datasets share only 25% of their features with the automated matching method as the vast majority (75%) remain unmatched.

In such settings, the reviewer correctly observes that the algorithms (GM and GMT) are outperformed by **Match** (Figure 1). Such a precision of 0.8 or larger is still competitive with **MetabCombiner** (mC) and **M2S** whose precisions drop (see panel 1).

Precision is measured as the number of metabolite matches identified divided by the total number of matches. In other words, even in the case of a 25% overlap, upwards of 80% of the true matches are correctly identified. This is a very good result.

- (2) How applicable is the method for other non-human species?

We thank the reviewer for this question. The crux of the issue is the application of the method to animal data. The metabolites in two different studies are compared under similar principles and the same set of reactions. Consistent with human population studies, the same principles hold to some extent.

However, for application to animal data, we have smaller sample sizes and often stem from individuals subjected to specific diets or treatments. In addition, the structures of metabolites can vary significantly between species. In response to the reviewer's question, we have searched through the two repositories (**MetaboLights** and **Metabolic Pathways**)

Summary:

The authors have implemented a tool for comparing LC-MS/MS datasets. This paper gains the proteomics field for performing meta-analysis of

Strengths:

The main strength of the MATCHer achieves significant performance compared to other existing methods. The authors have claimed that MATCHer performs well.

Weaknesses:

There are two weaknesses.

- (1) When the new sample is added, the precision drops to ~0.8.

We would like to clarify that this drop in precision occurs only for a small proportion of metabolites. The proportion of shared features - was 25% in our metabolic datasets share only 25% of their features. Our automated matching method as the vast majority 75% remain unmatched.

In such settings, the reviewer correctly observes that the algorithm (GM and GM+) drops to ~0.8 (Figure 1). Such a precision of 0.8 or larger is still competitive with MetabCombiner (mC) and M2S whose precisions drop in panel 1).

Precision is measured as the number of metabolite matches identified. In other words, even in the number of shared features (true matches) between two samples, upwards of 80% of the matches are correct. This is a very good result.

- (2) How applicable is the method for other non-human

We thank the reviewer for this question. The crux of the application to animal data is that the vast majority of metabolites in two different studies are shared. The metabolites under similar principles are the same in both of reactions. Core metabolites in human populations should hold to some extent.

However, application to animal data is more difficult. We have smaller sample sizes and often stem from individuals subjected to specific diets or treatments. The metabolite structures are different. In response to the reviewer's question, we have searched through the two repositories (MetaboLight and Metabolomics Workbench) for comparable animal studies due to the reasons mentioned. To navigate this issue could entail regressing the metabolite

notably differ between the two animal populations intensities. This would be an interesting direct validation would be needed to test the robustness

Reviewer #2 (Review): i c Re

Summary:

The goal of untargeted metabolomics is to identify biological samples. Untargeted metabolomics is a mass-to-charge ratio (m/z) and retention time (RT) based approach. Matching those metabolites based on their molecular weights and retention times is why methods for comparing samples are so crucial.

The main purpose of the paper is to merge and compare untargeted metabolomics experiments. These large datasets could then be used to advance biological research. A major challenge in metabolomics is that comparing metabolites from different experiments is difficult because of differences in m/z and RT values.

The main idea of the paper is to build a method that can match between two datasets based on their m/z and RT values. The authors propose a method that uses a combination of m/z and RT values to match metabolites between two datasets. This method is based on the assumption that the m/z and RT values of a metabolite are unique and can be used to identify it.

The variation in m/z and RT values is a common problem in metabolomics. The authors propose a method that can correct for this variation. The method uses a combination of m/z and RT values to match metabolites between two datasets. This method is based on the assumption that the m/z and RT values of a metabolite are unique and can be used to identify it.

The authors validate their method using a combination of experimental and computational data. The authors compare their method to other methods and show that their method is more accurate and robust. The authors also show that their method can be used to identify metabolites in a large dataset.

The second validation was done using a dataset of metabolites from a cancer cohort. The authors show that their method can be used to identify metabolites in a large dataset. The authors also show that their method can be used to identify metabolites in a large dataset.

Strengths:

I see the strength of this work in its combination of m/z and RT values to match metabolites between two datasets. The authors propose a method that uses a combination of m/z and RT values to match metabolites between two datasets. This method is based on the assumption that the m/z and RT values of a metabolite are unique and can be used to identify it.

We thank the reviewer for acknowledging this strength of the method.

The authors of the paper identify two main strengths of the methods that attempt to improve the performance: a) all of the methods require fine-tuning of user-defined parameters, and do not consider combinations between the features. However, the GromovMatcher incorporates the information (on distances) by using m/z and RT).

Weaknesses:

The first, minor, weakness is the lack of publicly available curated datasets that could be used for validation.

We thank the reviewer for the issue concerning the availability of

Manual layout datasets available for validation purposes from the laborious nature of matching features across datasets. Automatic matching methods like GromovMatcher approach as more data becomes accessible.

The scarcity of real-life publicly available datasets is the reason we conducted an extensive search for "ground-truth data" and Appendix 3). It is notably more than previous validations, utilizes real-life untargeted metabolomics data originates from distinct untargeted metabolomics parameters encoding feature intensities. Our study comprehensively explores the performance of GromovMatcher including in challenging realistic settings where varying levels of feature overlaps between studies as well as imbalances in the number of features and matched.

The second is also emphasized by the authors in the current setup is not representative of actual datasets.

This is indeed a limitation that is common to all methods. However, all these methods, GromovMatcher, M2S, and compare and pool multiple datasets using a single reference by designating a 'reference' dataset. We highlight this exact approach in our paper when aligning the CS, positive mode (pH 7.0) to the EPI C (pH 7.0) and the PE C (pH 7.0) aligned to the CS is not used by GromovMatcher twice, and after matchings, our analysis is restricted to those features in the study.

After the reviewer's comment, we have added an additional section 5, to compare the results of GromovMatcher depending on the reference dataset by setting the reference study to, either the GromovMatcher identified 706 and 708 common features. This highlights that the choice of reference dataset in the original analysis of the EPIC data, which is the largest original analysis of the EPIC data, chosen for the CS had the largest sample size (compared to HCC and

PC were already expected to the CS study which we c
Loftfield et al. (2021). J Natl Cancer Inst.).

As mentioned in the discussion section of our man
multimarginal as Gromov-Wasser algorithm (Beier, F., Bein
Information and Inference) could potentially allo
matched using one optimization routine (e.g. with
for matching). We explored this possibility in d
for multi-marginal in Athesei, r siuncfha nncyl multi-marginal
computational coupling or matching matrices th
number of dimensions is equal to the number of da
multimarginal methods have large memory costs, wh
application for the matching of multiple metabolo

Reviewer #2 (Recommendations For The Authors):

(1) I was struck by the points in the Figure points
over the points on a distribution. It is visually quant
eyes mainly focused on the pattern of the r

Figure 3a has been modified. We improve the conditions
reordered the rows and columns of the coupling ma
(green points) are the part of the negative match
fraction of true positive and false negative match
eye in Figure 3a.

(2) would add the caveat that I cannot judge
other two methods with the M2S and
metabCombiner) optimally. But I also do not see
Hopefully one of these is the

We appreciate the reviewer for highlighting the c
to the other existing methods M2S and MetabCombin
tensy parameters each with a discrete or continuou
properly optimized. The coupling matrices between d
Appendix 2 show parameters of the M2S and mC methods a
achieve the best possible performance on, the h
the simulation study and on EPIC data. We report
the M2S and mC methods are the best in the simulation
F1 score and we found that the M2S method is the best
optimization was performed for our Gromov-Matcher
the main text, we find that Gromov-Matcher outper
when parameters of M2S and mC are tuned to predi

Given the large computational cost, we believe
thoroughly tested the combinations that user
would like to explore in their own research.

(3) Validation

(3a) The first validation dataset used to see if the
the paper however is its use of the dataset split.

We are grateful for the reviewer and have included new
3 which show how the results of Gromov-Matcher, M2

the dataset split. In our original manuscript, our with an untargeted metabolomic dataset consisting metabolic features which is split equally into two samples $n_1 = n_2$ and an equal number of metabolic equal-sized datasets $w_{p_1} = w_{p_2}$. matched b

Now in Appendix 3 (Figs. 1-3) we show the sensitivity (GromovMatcher, M2S, and MetabCombiner) of w_{p_1} over dataset 1 over dataset 2. In the overlap in shared features the fraction of metabolic features in dataset 1 that the feature sizes of w_{p_1} in dataset 1. w_{p_1} maintain a consistent precision and recall. w_{p_1} are varied. GromovMatcher achieves a higher precision MetabCombiner for all choices of dataset split, a results from the main text (see main text Fig. 3) precision (without dropping in recall) when dataset of unshared features (e.g. when $p_1 = p_2$). Therefore Appendix 3 show that our results in the main text setting for the dataset split.

(B) The second validation was done on a dataset of (metabolic) from cancer dataset EPIIC cohort that was manually. Here the authors observe that the drift in RT between the two experiments. M2S has a very similar performance to GromovMatcher. The fact that the drift in RT between the two experiments affect the M2S performance in its validation dataset the drift in RT is not significant.

We thank the reviewer for their question. At the time of writing this manuscript we did not have a validation study with a nonlinear retention time shift.

Nevertheless, we performed an additional analysis - "M2S performance experiments" in Appendix 2. We show a decrease in M2S performance when the simulated drift in retention time is large. In Appendix 2-1 in a low overlap setting with a linear drift in retention time, precision and recall, were set to 0.3 and 0.7 respectively in the main analysis where the drift was nonlinear.