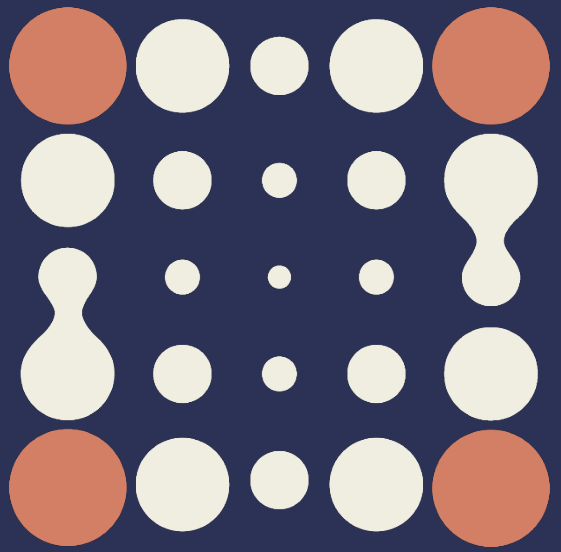
Optimal transport for automatic alignment of untargeted metabolomic data

Marie Breeur[1], George Stepaniants[2], Pekka Keski-Rahkonen[1], Philippe Rigollet[2], Vivian Viallon[1]

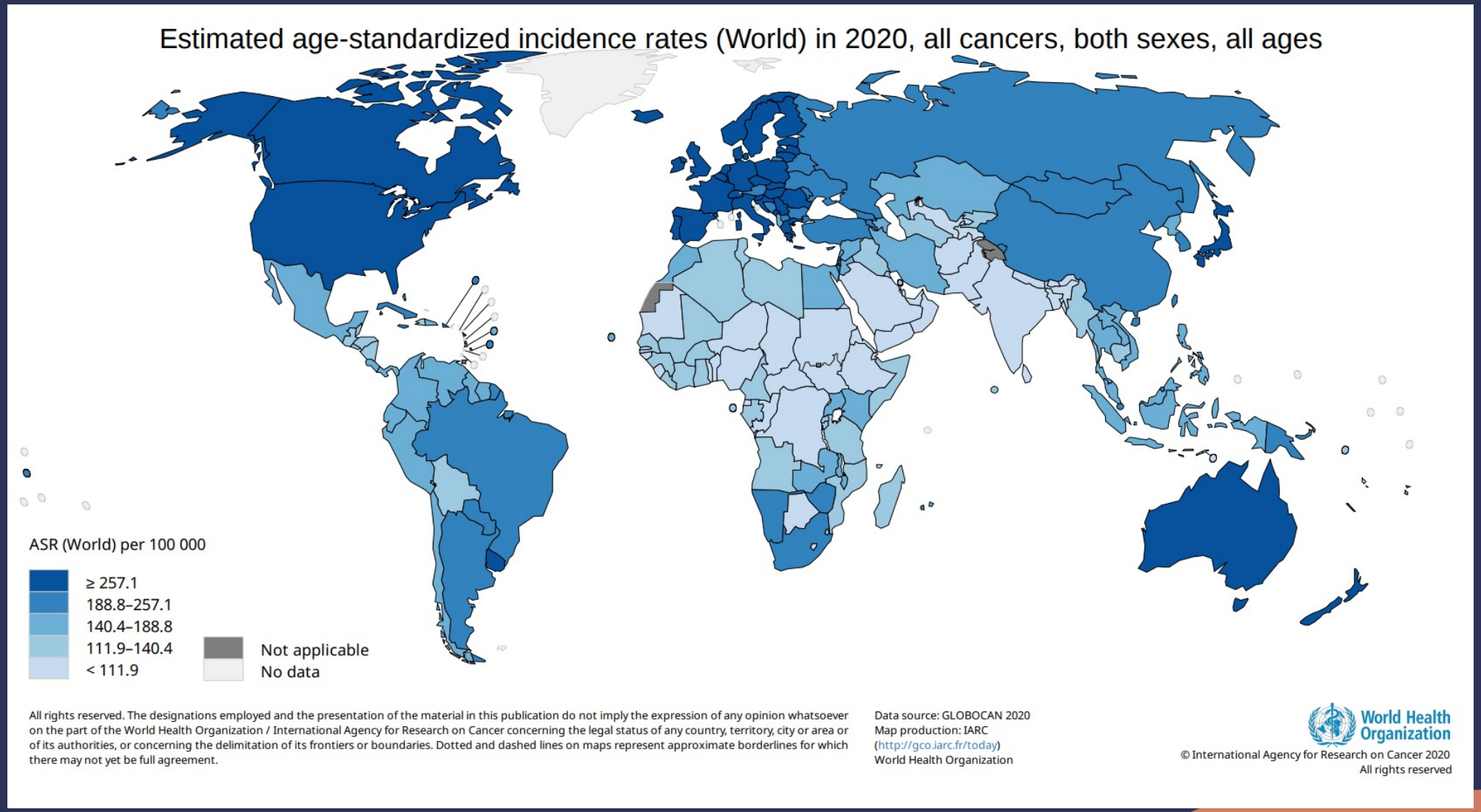
[1] Nutrition and Metabolism (NME) Branch, International Agency for Research on Cancer
[2] Department of Mathematics, Massachusetts Institute of Technology



International Agency for Research on Cancer
World Health Organization

February 23th 2023

Cancer and lifestyle



Cancer and lifestyle

International Agency for Research on Cancer

- Better understand the causes and determinants of cancer, both endogenous and exogenous

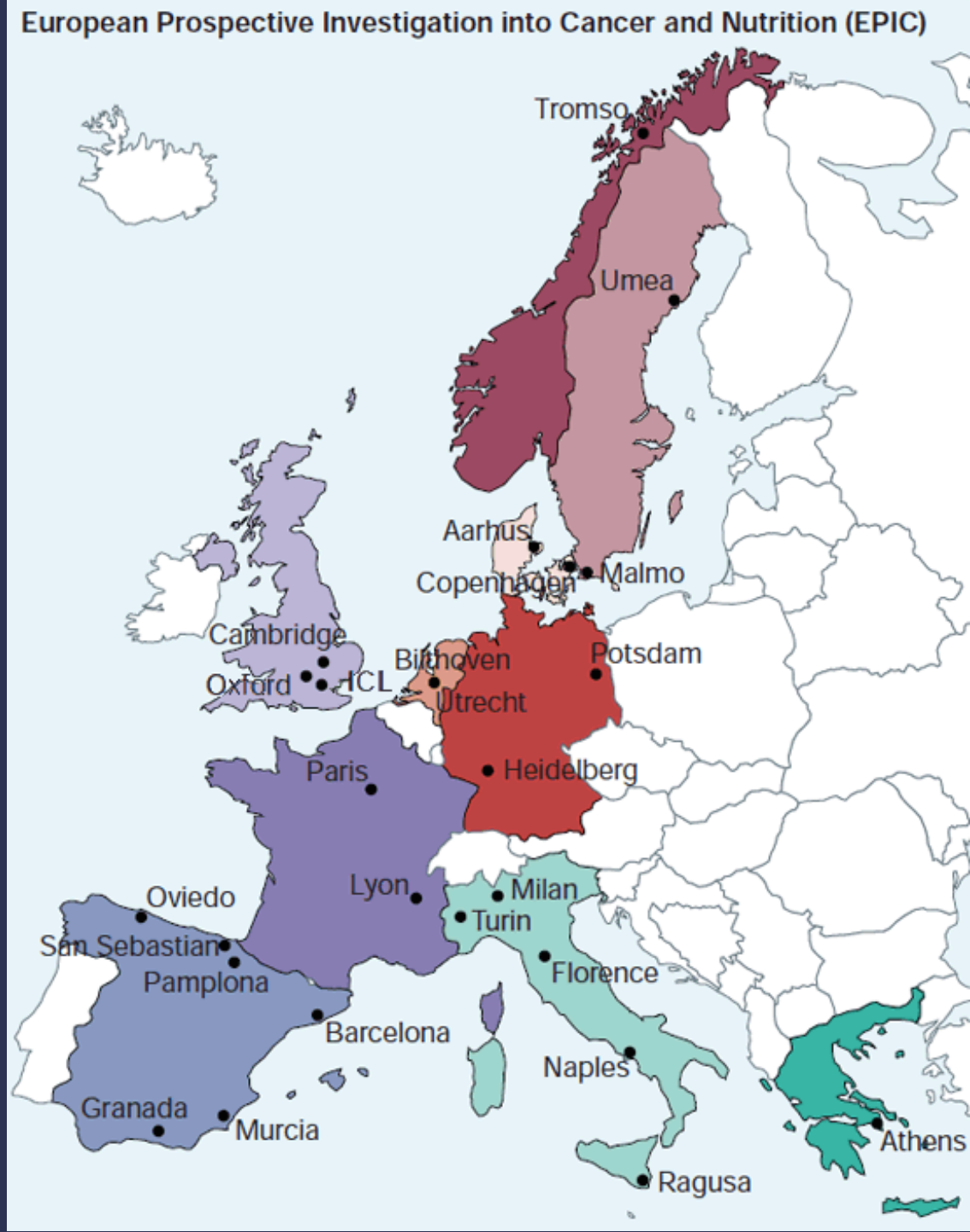
- Nutrition and Metabolism Branch (NME): focuses on lifestyle factors

Cancer and lifestyle

European Prospective Investigation into Cancer and nutrition cohort

- 10 European countries
- ~521K participants recruited around 1990
- Biological samples collected at inclusion

➢ Dietary, lifestyle, metabolomic, genetic data available



Cancer and lifestyle

- Study impact of alcohol on cancer

- Ask about study participants' alcohol intake

Cancer and lifestyle

- Study impact of alcohol on cancer

- Ask about study participants' alcohol intake

Cancer and lifestyle

- Study impact of alcohol on cancer

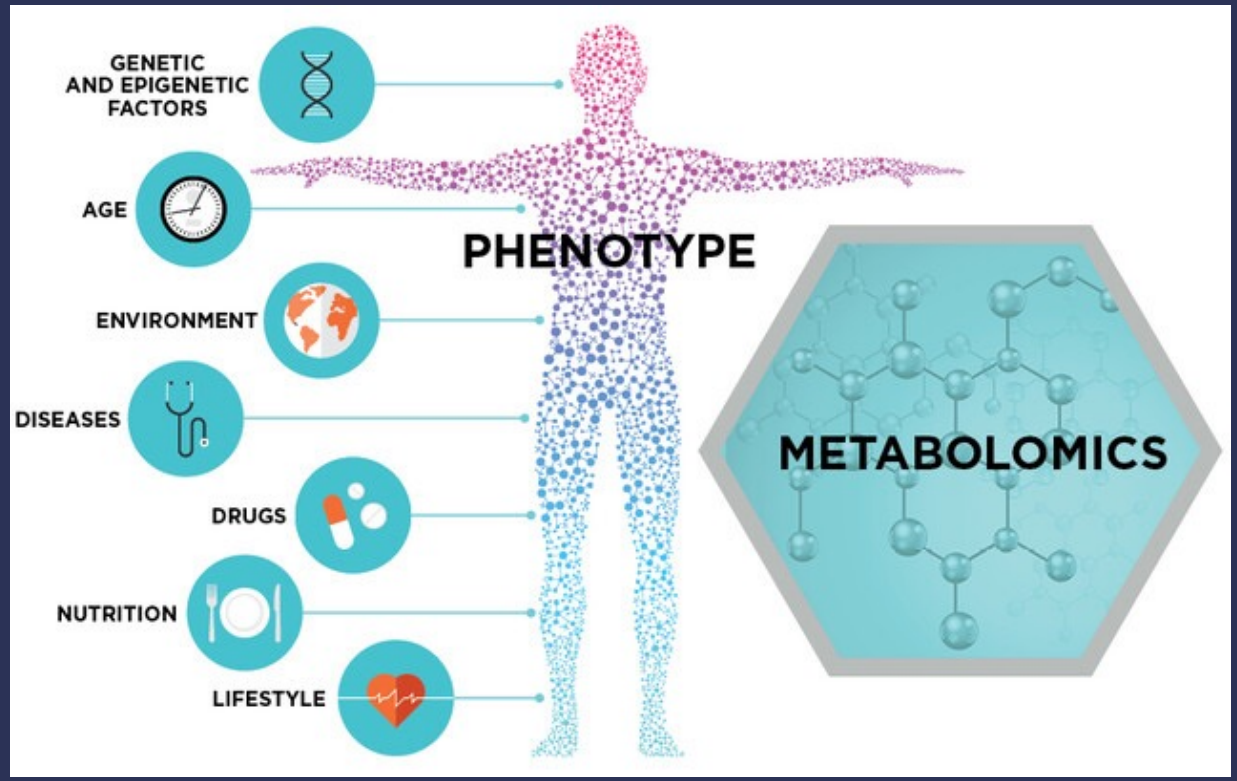
- Ask about study participants' alcohol intake

- **Biomarker (for alcohol)**

Biological molecule found in body that would accurately reflect alcohol intake

Cancer and lifestyle – Metabolomics

- Large-scale study of small molecules (metabolites) in a biological sample
- Reflects the metabolic health of an individual influenced by both genetic and environmental factors
- Untargeted approach: measure as many metabolites as possible in a sample

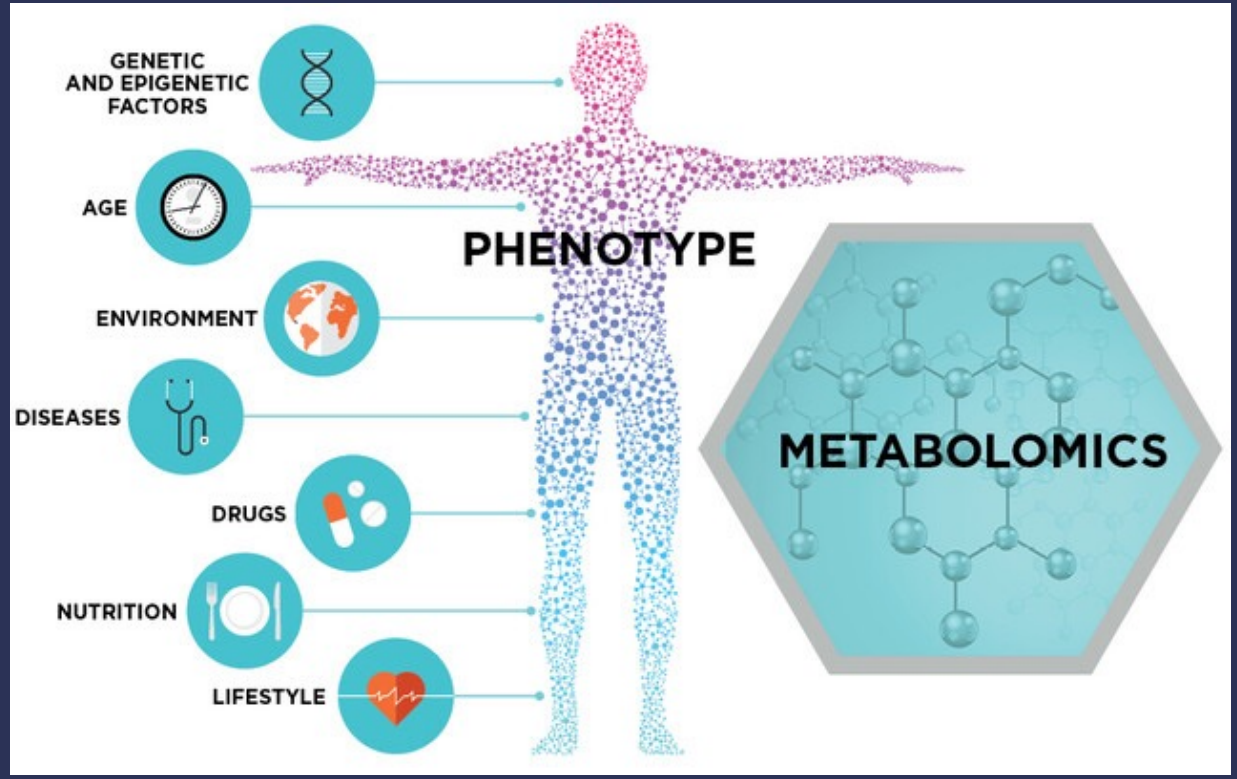


Cancer and lifestyle - Metabolomics

- Large-scale study of small molecules (metabolites) in a biological sample
- Reflects the metabolic health of an individual influenced by both genetic and environmental factors
- Untargeted approach: measure as many metabolites as possible in a sample

✓ **Perfect for biomarker discovery**

✗ **Costly approach, generally low sample size.**



Cancer and lifestyle - Metabolomics

- Large-scale study of small molecules (metabolites) in a biological sample
- Reflects the metabolic health of an individual influenced by both genetic and environmental factors
- Untargeted approach: measure as many metabolites as possible in a sample

✓ **Perfect for biomarker discovery**

✗ **Costly approach, generally low sample size.**

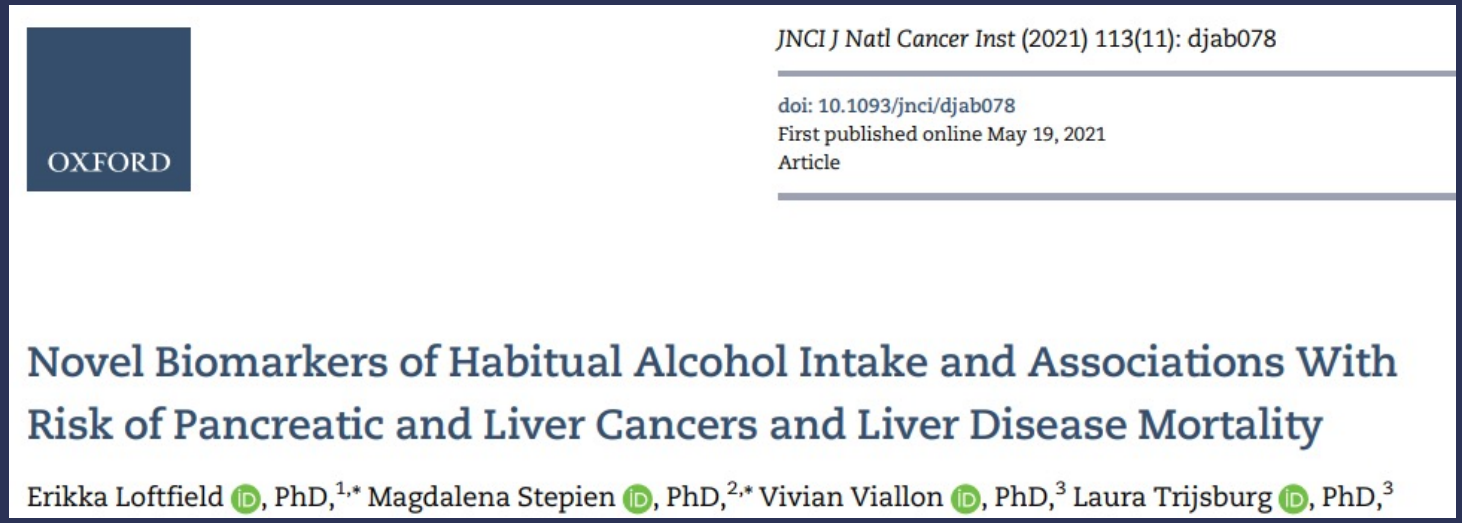
➢ **Pool/meta-analyse data from different sources ?**

Cancer and lifestyle – Metabolomics

Loftfield *et al.* 2021:

- Discover biomarkers associated with alcohol consumption, several features identified

- Untargeted metabolomic data from the EPIC cross-sectional calibration study, EPIC liver study, EPIC pancreas study, and two studies nested in the ATBC cohort



Cancer and lifestyle – Metabolomics

Untargeted metabolomics

- Measure every metabolite in a sample with LC-MS

- Features identified by
  - Mass-to-charge ratio (m/z)
  - Retention time (RT)

Samplename	218.0763@0.5936028	196.0938@0.59344584
Sample003	81951.95	33048.715
Sample004	69366.21	27925.324
Sample005	88970.75	34721.086
Sample006	45261.00	18201.113
Sample007	82271.65	32732.715
Sample008	43436.75	18519.811
Sample009	44902.54	16453.068
Sample010	20530.35	8739.655

Cancer and lifestyle – Metabolomics

Untargeted metabolomics

- Measure every metabolite in a sample with LC-MS

- Features identified by
  - Mass-to-charge ratio (m/z)
  - Retention time (RT)

m/z = 218.0763
RT = 0.5936028

Samplename	218.0763@0.5936028	196.0938@0.59344584
Sample003	81951.95	33048.715
Sample004	69366.21	27925.324
Sample005	88970.75	34721.086
Sample006	45261.00	18201.113
Sample007	82271.65	32732.715
Sample008	43436.75	18519.811
Sample009	44902.54	16453.068
Sample010	20530.35	8739.655

Cancer and lifestyle – Metabolomics

Untargeted metabolomics

- Measure every metabolite in a sample with LC-MS

- Features identified by
  - Mass-to-charge ratio (m/z)
  - Retention time (RT)

➢ Finding and matching features common to several studies is challenging.

m/z = 218.0763
RT = 0.5936028

Can vary across studies

Samplename	218.0763@0.5936028	196.0938@0.59344584
Sample003	81951.95	33048.715
Sample004	69366.21	27925.324
Sample005	88970.75	34721.086
Sample006	45261.00	18201.113
Sample007	82271.65	32732.715
Sample008	43436.75	18519.811
Sample009	44902.54	16453.068
Sample010	20530.35	8739.655

Cancer and lifestyle - Metabolomics

Untargeted metabolomics

- Measure every metabolite in a sample with LC-MS

- Features identified by
  - Mass-to-charge ratio (m/z)
  - Retention time (RT)

➢ Finding and matching features common to several studies is challenging.

- Possible by hand only on a restricted number of features

- Existing methods to align untargeted datasets: metabCombiner, M2S, PAIRUP-MS... Either require prior knowledge, overrely on hyperparameters, or make strict assumptions on the data

➢ **GromovMatcher**

Method overview

Study 1
n_1 samples, p_1 features

	Feat X_1	Feat X_2	Feat X_3	...	Feat X_{p_1}
m/z	743.8	231.1	189.7	...	435.4
RT	0.56	1.58	5.32	...	7.61
Feature intensities	10.6	12.1	8.4	...	9.2
	...	...	...	...	...
	9.5	9.1	13.6	...	10.8

Study 2
n_2 samples p_2 fatures

	Feat Y_1	Feat Y_2	...	Feat Y_{p_2}
m/z	349.0	233.0	...	528.1
RT	0.23	3.47	...	6.82
Feature intensities	12.9	8.9	...	11.2
	13.1	9.9	...	10.3
	...	...	...	...
	13.5	9.1	...	11.4

Method overview

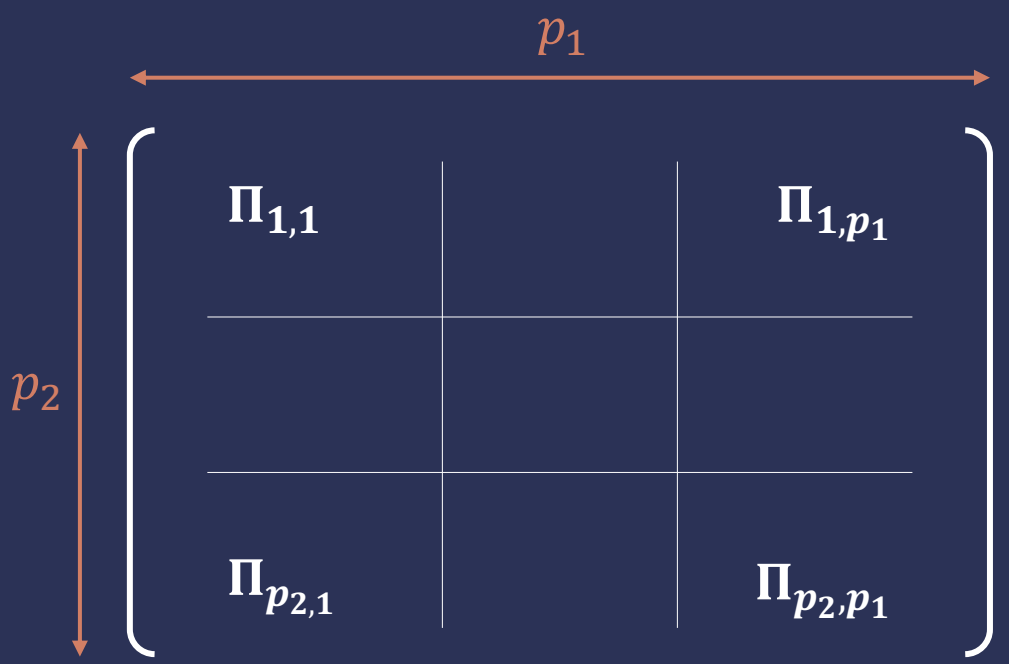
Study 1
$n_1 \times p_1$

	Feat X_1	Feat X_2	Feat X_3	...	Feat X_{p_1}
m/z	743.8	231.1	189.7	...	435.4
RT	0.56	1.58	5.32	...	7.61
Feature intensities	10.6	12.1	8.4	...	9.2
	...	...	...	...	...
	9.5	9.1	13.6	...	10.8

Study 2
$n_2 \times p_2$

	Feat Y_1	Feat Y_2	...	Feat Y_{p_2}
m/z	349.0	233.0	...	528.1
RT	0.23	3.47	...	6.82
Feature intensities	12.9	8.9	...	11.2
	13.1	9.9	...	10.3
	...	...	...	...
	13.5	9.1	...	11.4

- Find *coupling matrix* $\Pi \in [0,1]^{p_1 \times p_2}$ such that $\Pi_{i,j}$ is non-zero iif X_i and Y_j correspond to the same underlying feature, 0 otherwise



Method overview

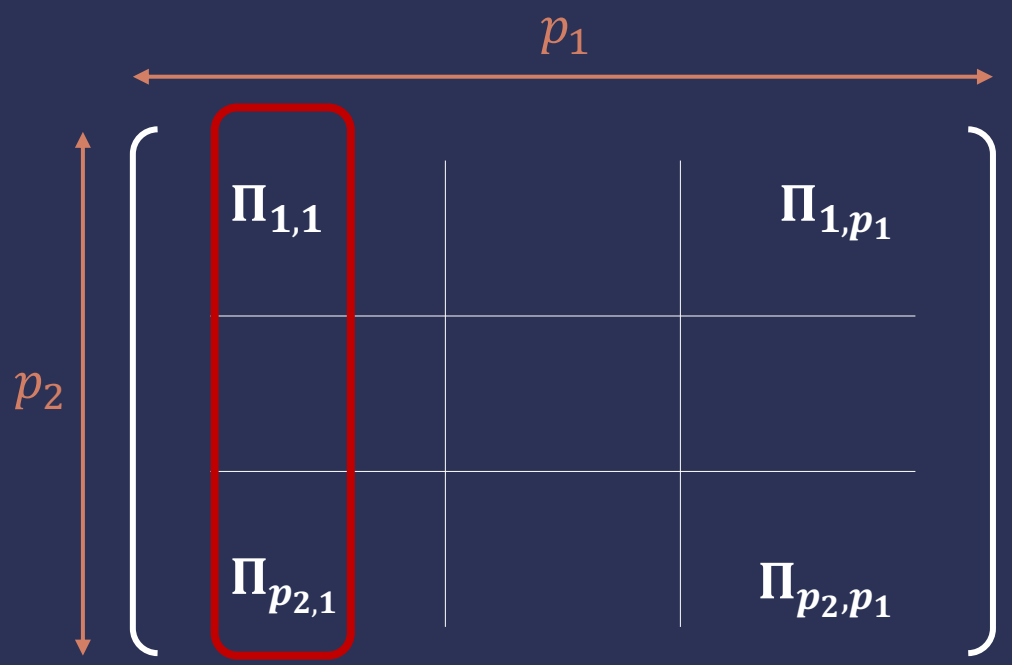
Study 1
$n_1 \times p_1$

	Feat X_1	Feat X_2	Feat X_3	...	Feat X_{p_1}
m/z	743.8	231.1	189.7	...	435.4
RT	0.56	1.58	5.32	...	7.61
Feature intensities	10.6	12.1	8.4	...	9.2
	...	...	...	...	...
	9.5	9.1	13.6	...	10.8

Study 2
$n_2 \times p_2$

	Feat Y_1	Feat Y_2	...	Feat Y_{p_2}
m/z	349.0	233.0	...	528.1
RT	0.23	3.47	...	6.82
Feature intensities	12.9	8.9	...	11.2
	13.1	9.9	...	10.3
	...	...	...	...
	13.5	9.1	...	11.4

- Find *coupling matrix* $\Pi \in [0,1]^{p_1 \times p_2}$ such that $\Pi_{i,j}$ is non-zero iif X_i and Y_j correspond to the same underlying feature, 0 otherwise



Method overview

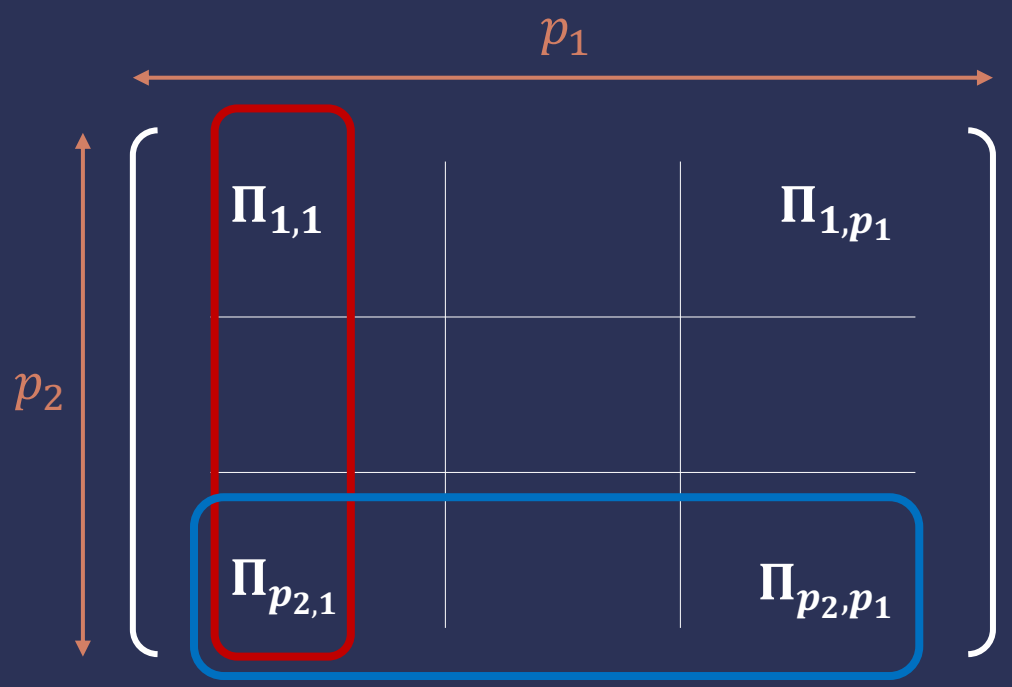
Study 1
$n_1 \times p_1$

	Feat X_1	Feat X_2	Feat X_3	...	Feat X_{p_1}
m/z	743.8	231.1	189.7	...	435.4
RT	0.56	1.58	5.32	...	7.61
Feature intensities	10.6	12.1	8.4	...	9.2
	...	...	...	...	...
	9.5	9.1	13.6	...	10.8

Study 2
$n_2 \times p_2$

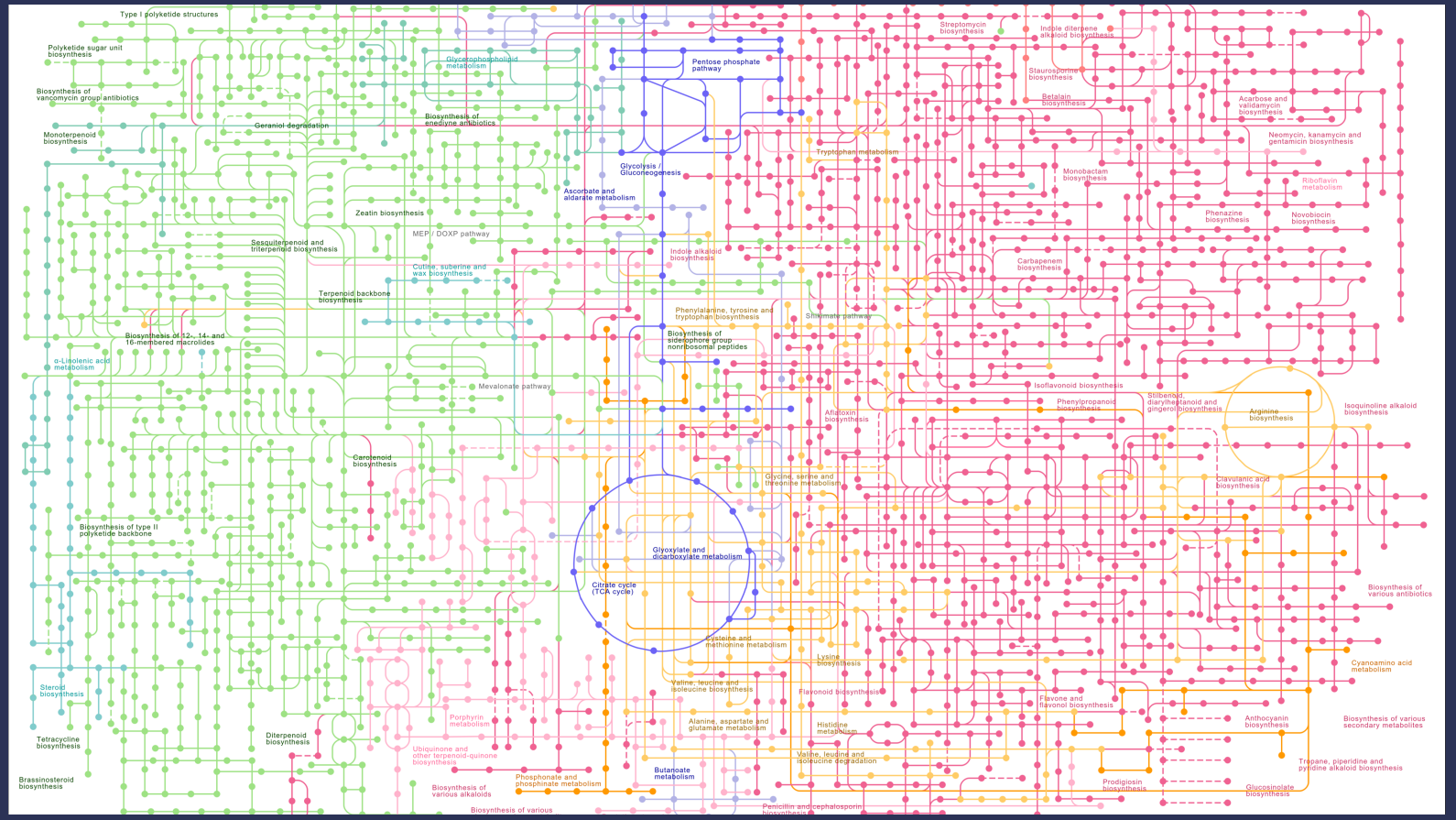
	Feat Y_1	Feat Y_2	...	Feat Y_{p_2}
m/z	349.0	233.0	...	528.1
RT	0.23	3.47	...	6.82
Feature intensities	12.9	8.9	...	11.2
	13.1	9.9	...	10.3
	...	...	...	...
	13.5	9.1	...	11.4

- Find *coupling matrix* $\Pi \in [0,1]^{p_1 \times p_2}$ such that $\Pi_{i,j}$ is non-zero iif X_i and Y_j correspond to the same underlying feature, 0 otherwise



Method overview

m/z and RT are not stable enough, but similar correlation patterns can be expected



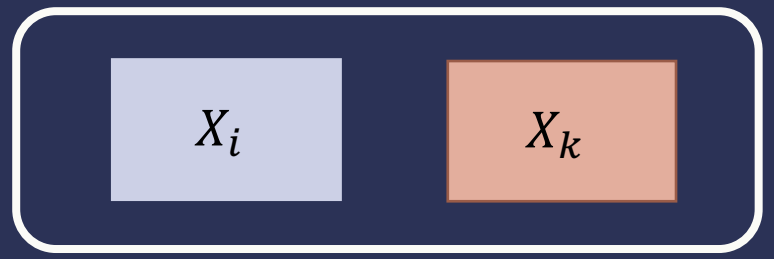
Kanehisa Lab.

Method overview

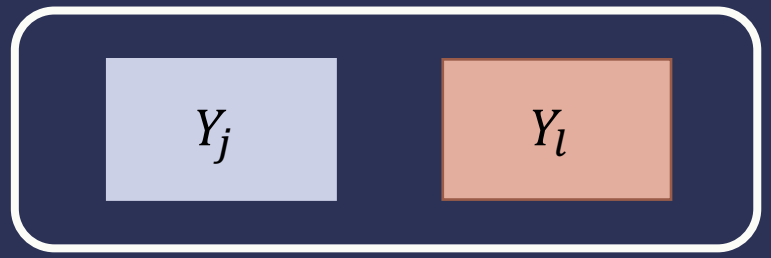
Metabolite A

Metabolite B

Study 1: n_1 samples, p_1 features



Study 2: n_2 samples, p_2 features



$$Corr(X_i, X_k) \approx Corr\ (Y_j, Y_l)$$

Method overview

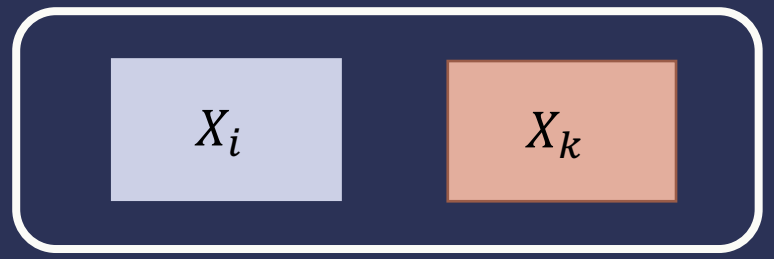
Metabolite A

Metabolite B

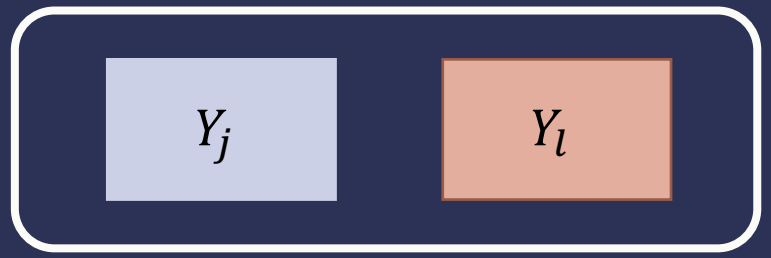
- For $d_i(x, x') = \frac{1}{\sqrt{n_i}} \left|\left|x - x'\right|\right|_2$,

$$d_1(X_i, X_k) \approx d_2(Y_j, Y_l)$$

Study 1: n_1 samples, p_1 features



Study 2: n_2 samples, p_2 features

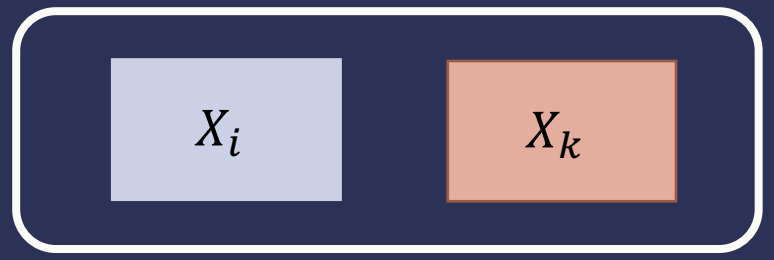


Method overview

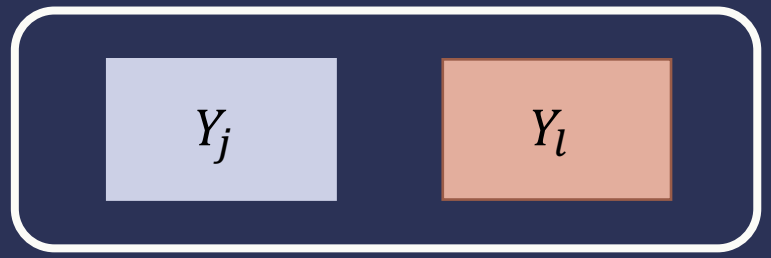
Metabolite A

Metabolite B

Study 1: n_1 samples, p_1 features



Study 2: n_2 samples, p_2 features



- For $d_i(x, x') = \frac{1}{\sqrt{n_i}} \left|\left|x - x'\right|\right|_2$,

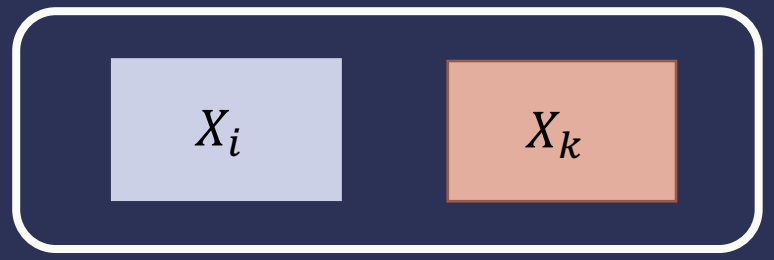
$$\left|d_1(X_i, X_k) - d_2(Y_j, Y_l)\right| \approx 0$$

Method overview

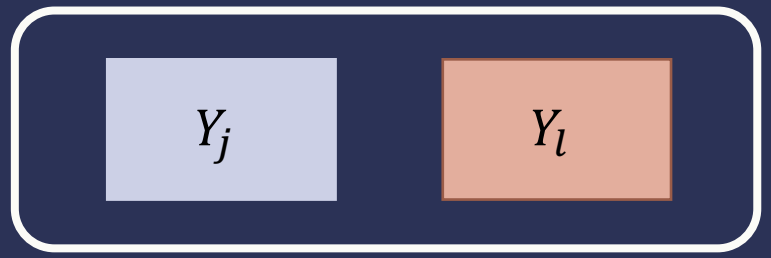
Metabolite A

Metabolite B

Study 1: n_1 samples, p_1 features



Study 2: n_2 samples, p_2 features



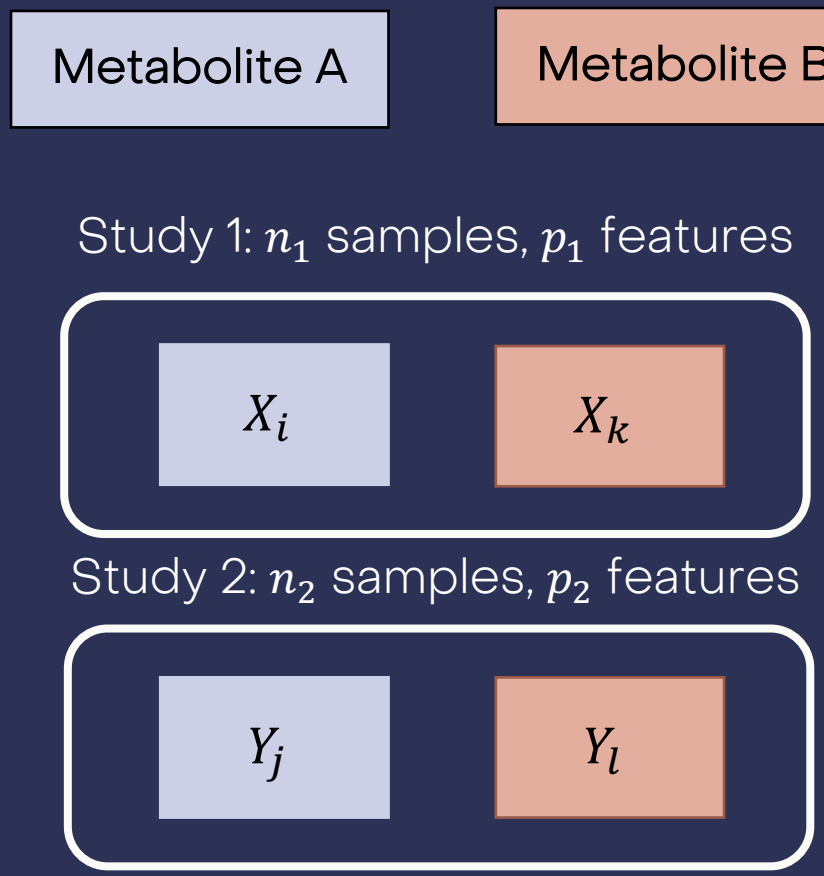
- For $d_i(x, x') = \frac{1}{\sqrt{n_i}} \left|\left|x - x'\right|\right|_2$,

$$\left|d_1(X_i, X_k) - d_2(Y_j, Y_l)\right| \approx 0$$

- **Gromov-Wasserstein [Memoli, 2011]:**

$$\hat{\Pi} = \underset{\boldsymbol{\Pi} \in \mathbb{U}}{\operatorname{argmin}} \sum_{i,j,k,l} \Pi_{i,j} \Pi_{k,l} \left|d_1(\boldsymbol{X_i}, \boldsymbol{X_k}) - d_2(\boldsymbol{Y_j}, \boldsymbol{Y_l})\right|^2$$

Method overview



- For $d_i(x, x') = \frac{1}{\sqrt{n_i}} \left|\left|x - x'\right|\right|_2$,

$$\left|d_1(X_i, X_k) - d_2(Y_j, Y_l)\right| \approx 0$$

- **Gromov-Wasserstein [Memoli, 2011]:**

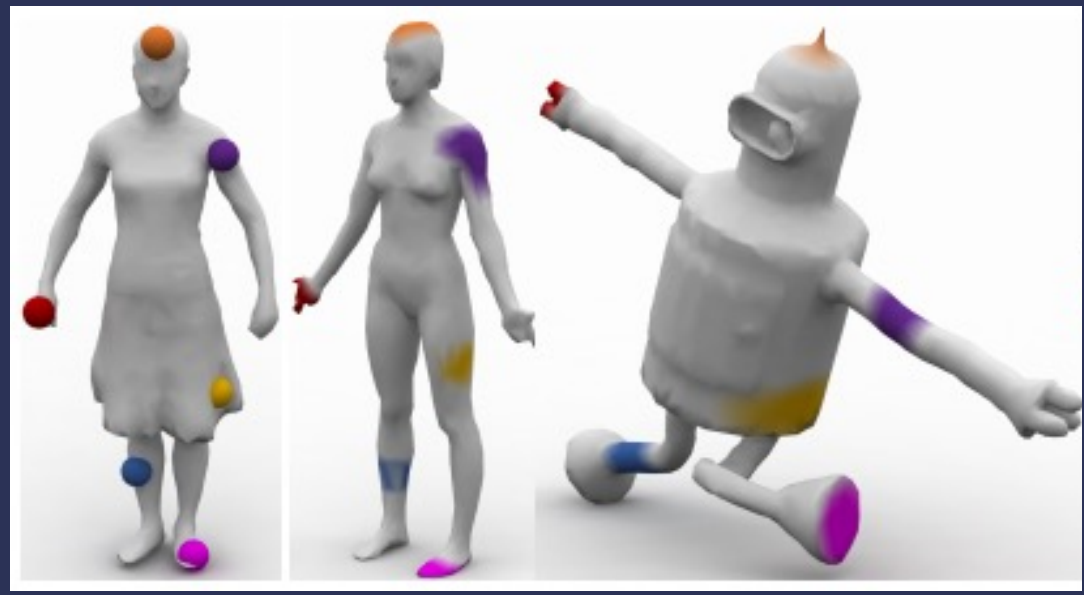
$$\hat{\Pi} = \underset{\boldsymbol{\Pi} \in \mathbb{U}}{\operatorname{argmin}} \sum_{i,j,k,l} \Pi_{i,j} \Pi_{k,l} \left|d_1(\boldsymbol{X_i}, \boldsymbol{X_k}) - d_2(\boldsymbol{Y_j}, \boldsymbol{Y_l})\right|^2$$

with $\mathbb{U} = \left\{\Pi \in \mathbb{R}_+^{p_1 \times p_2} : \Pi \mathbb{1}_{p_2} = \frac{1}{p_1} \mathbb{1}_{p_1} \text{ and } \Pi^T \mathbb{1}_{p_1} = \frac{1}{p_2} \mathbb{1}_{p_2}\right\}$

Method overview - Gromov-Wasserstein

- Expands optimal transport framework to sets living in different spaces: shape-wise matching

- ✓ Use distance profile to characterize the 'shape' of the sets
  - Versatile, adapts to every setting where a distance can be set between the points to match.



Solomon et al. 2016

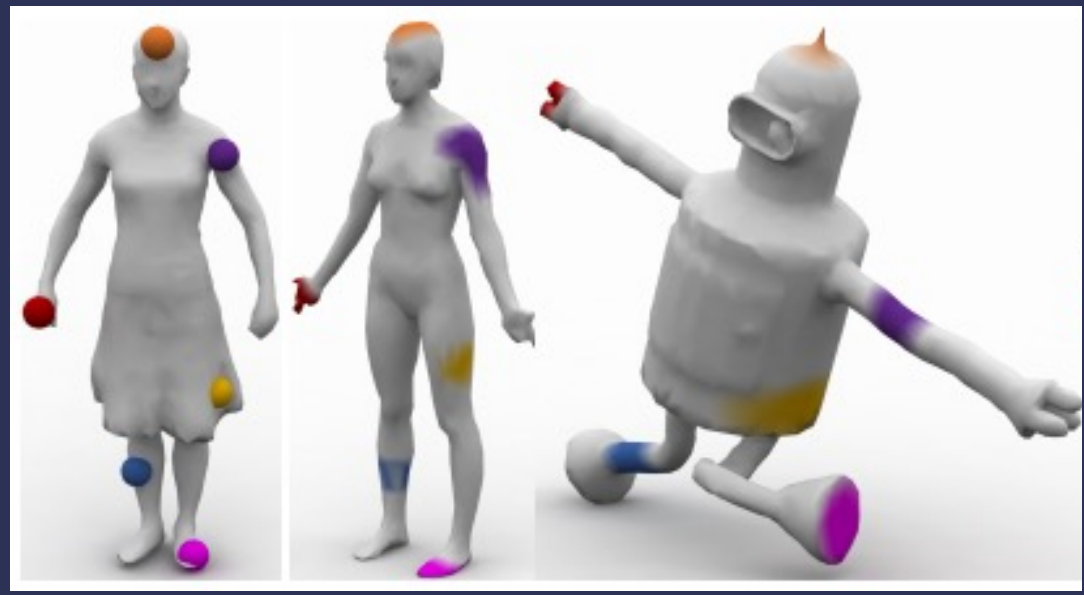
Method overview - Gromov-Wasserstein

- Expands optimal transport framework to sets living in different spaces: shape-wise matching

✓ Use distance profile to characterize the 'shape' of the sets
  - Versatile, adapts to every setting where a distance can be set between the points to match.

✗ Has a hard constraint ($\Pi \in \mathbb{U}$)
  - Will match every point in both sets

✗ Does not take into account additional knowledge on the points it's looking at
  - m/z and RT are not accounted for at all



Solomon et al. 2016

Method overview - Constraints

Use the information contained in the feature tags:

- m/z are relatively stable.

➢ Only couple features if their m/z difference is less than a user-specified threshold

Method overview - Constraints

Use the information contained in the feature tags:

- m/z are relatively stable.

➢ Only couple features if their m/z difference is less than a user-specified threshold

$$\hat{\Pi} = \underset{\Pi \in \mathcal{S}}{\operatorname{argmin}} \sum_{i,j,k,l} \Pi_{i,j}\Pi_{k,l}\left|d_1(X_i, X_k) - d_2(Y_j, Y_l)\right|^2$$

with $\mathcal{S} = \{\Pi \in \mathbb{U}: \Pi_{i,j} = 0 \text{ if } \left|m_i^1 - m_j^2\right| > M_{gap}\}$

Method overview - Constraints

Use the information contained in the feature tags:

- m/z are relatively stable.

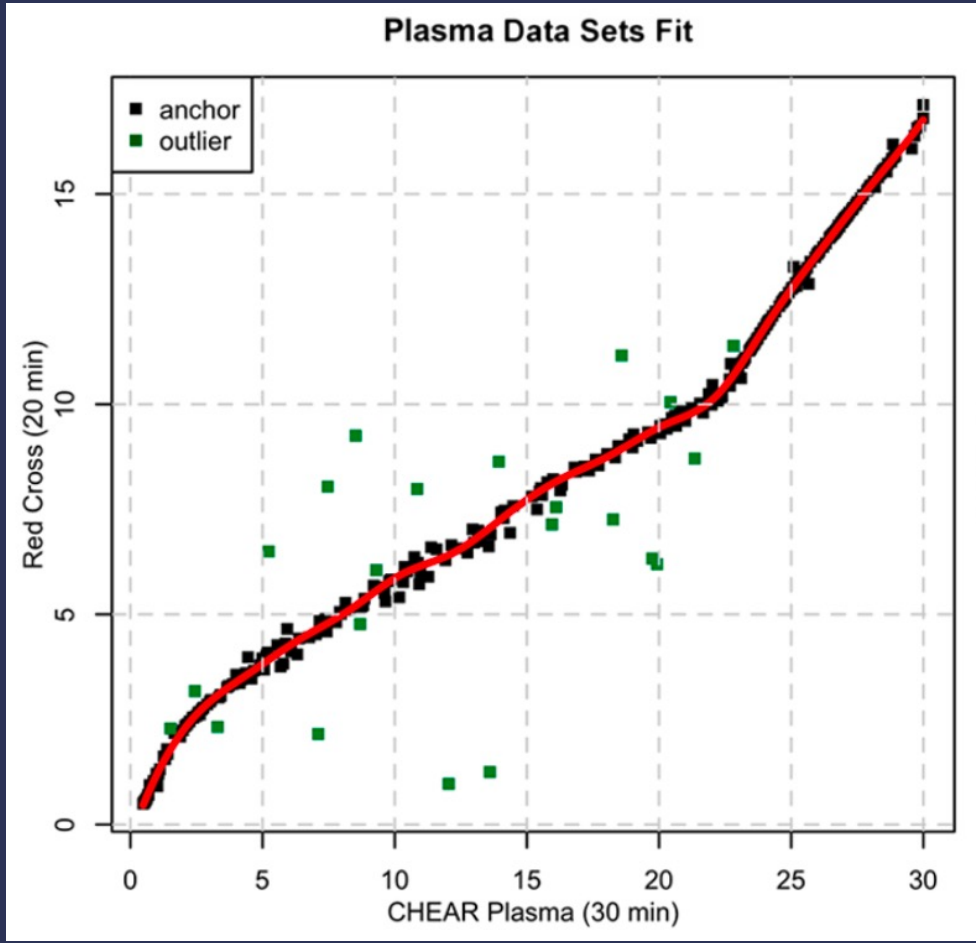
➢ Only couple features if their m/z difference is less than a user-specified threshold

$$\hat{\Pi} = \underset{\Pi \in \mathcal{S}}{\operatorname{argmin}} \sum_{i,j,k,l} \Pi_{i,j}\Pi_{k,l}\left|d_1(X_i, X_k) - d_2(Y_j, Y_l)\right|^2$$

with $\mathcal{S} = \{\Pi \in \mathbb{U}: \Pi_{i,j} = 0 \text{ if } \left|m_i^1 - m_j^2\right| > M_{gap}\}$

- RT vary way more, in a non-linear fashion

Method overview – RT drift

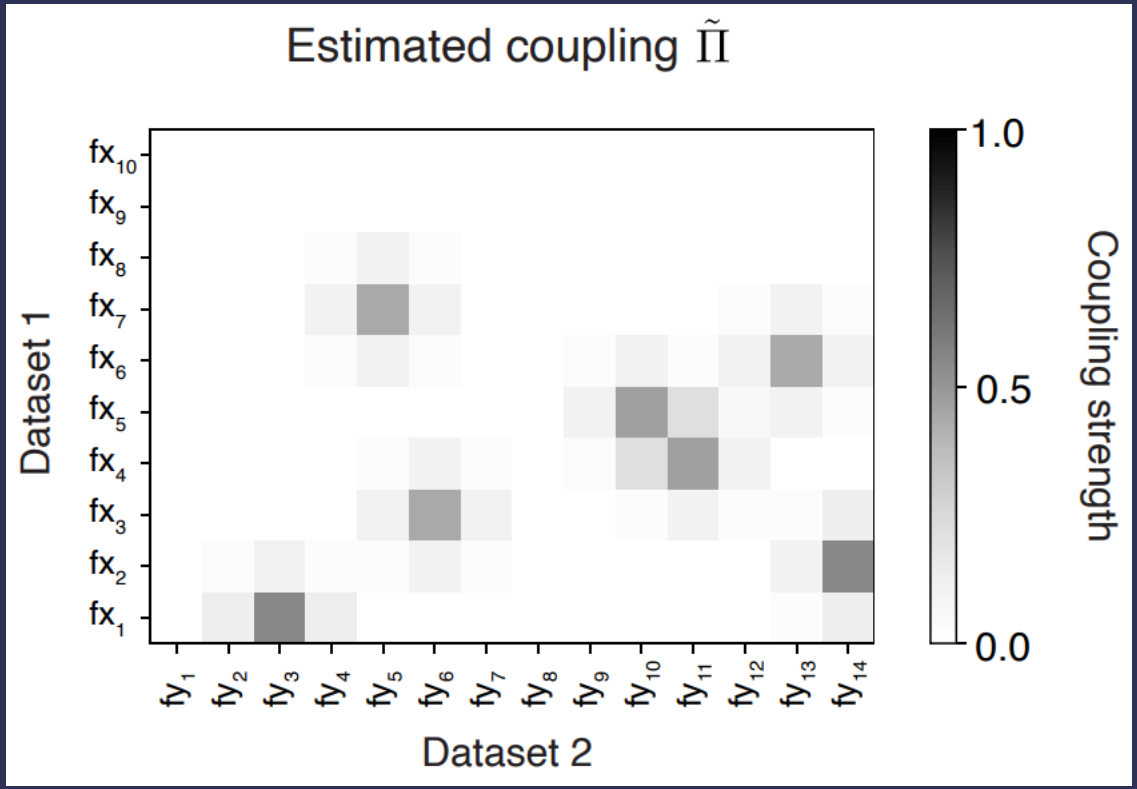


Habra et al. 2021

RT drift can be non-linear and of high amplitude

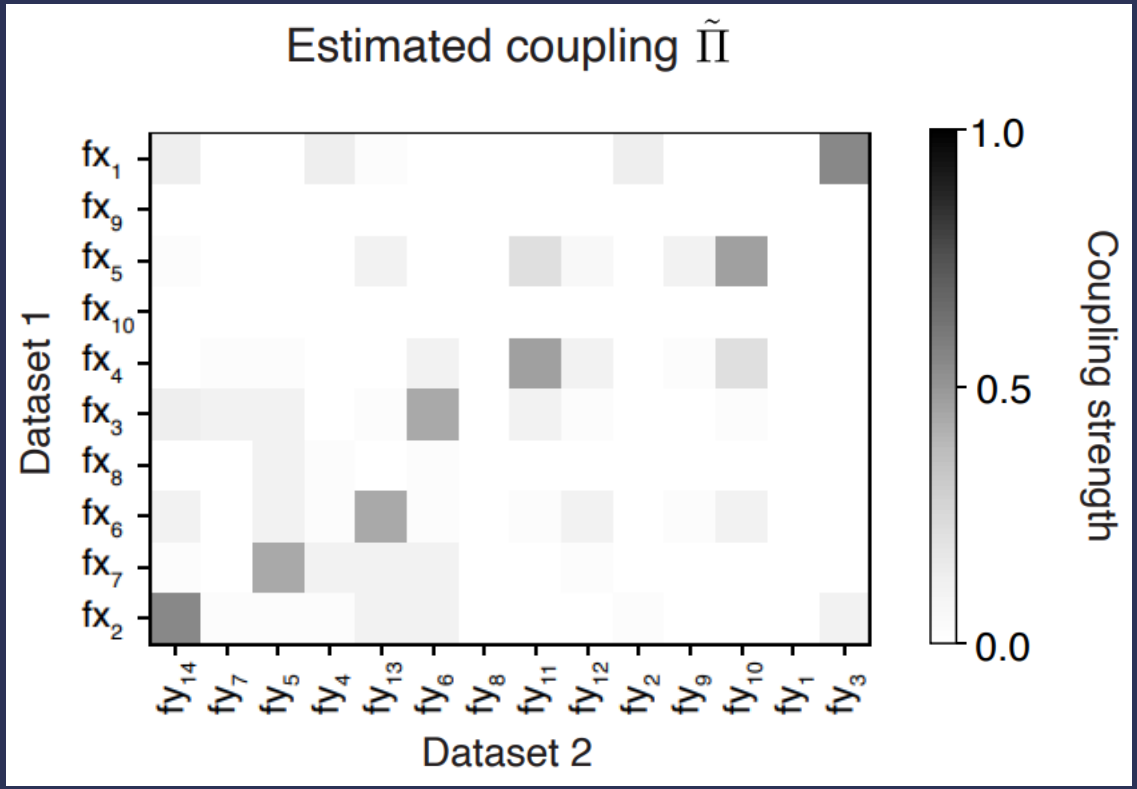
- Difficult to account for with an a priori constraint like the m/z
- Estimate the drift a posteriori and discard matched pairs that have incompatible RTs

Method overview – RT drift



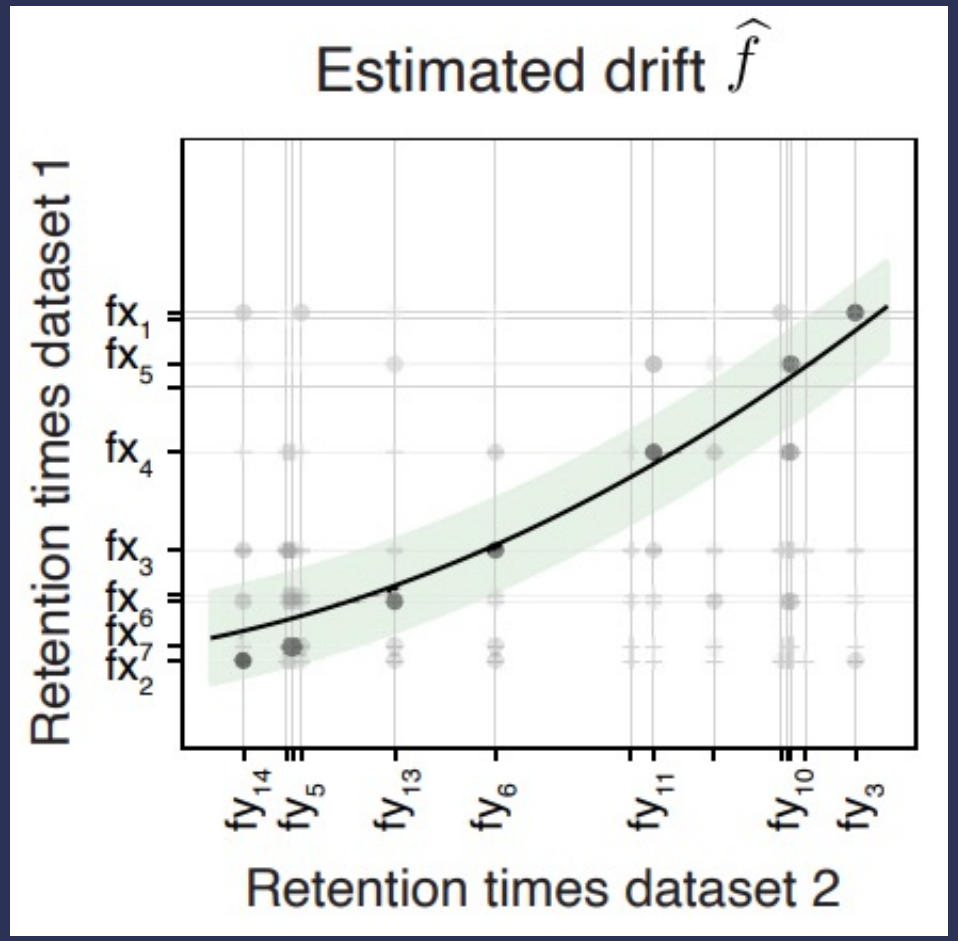
1. Solve m/z constrained GW problem

Method overview – RT drift



1. Solve m/z constrained GW problem

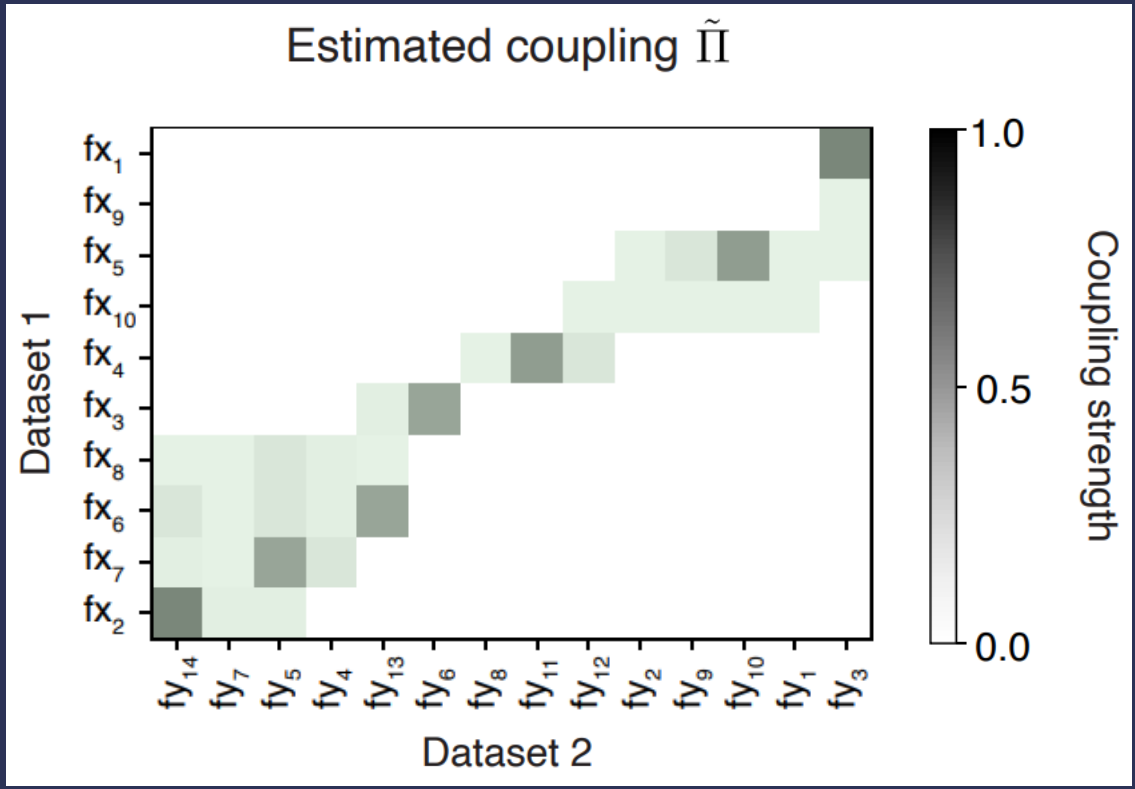
Method overview – RT drift



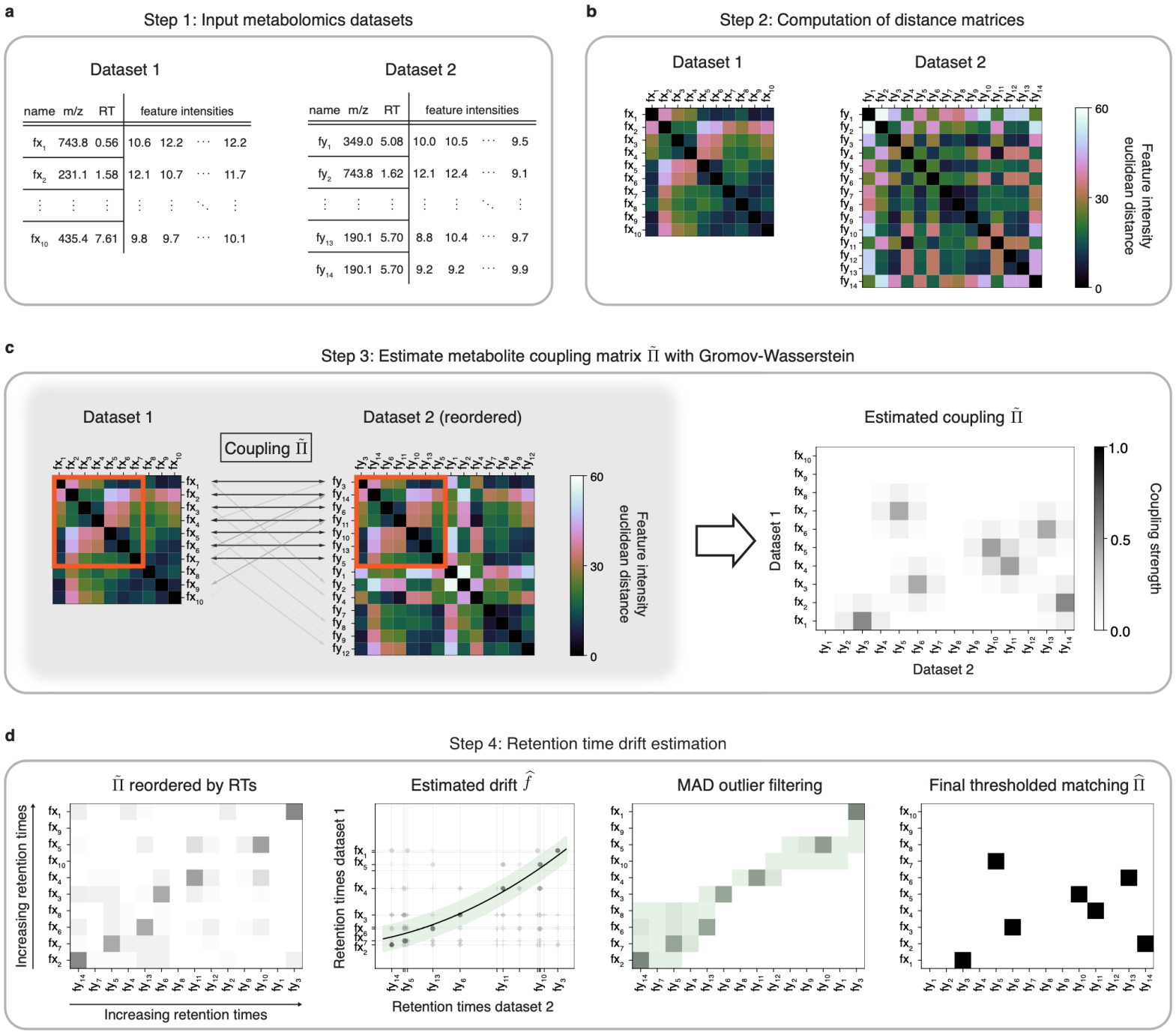
1. Solve m/z constrained GW problem

2. Estimate RT drift f such that $rt^Y = f(rt^X)$
   - Weighted cubic B-spline with k knots, k selected by CV

Method overview – RT drift



1. Solve m/z constrained GW problem
2. Estimate RT drift f such that $rt^Y = f(rt^X)$
   - Weighted cubic B-spline with k knots, k selected by CV
3. Discard the outlying pairs
   - Discard pairs whose residual is higher than the MAD



GromovMatcher

Unbalanced Gromov-Wasserstein distance with entropic regularization [Séjourné et al. 2020]

- Allows for features to be dropped during the matching
- Computationally faster

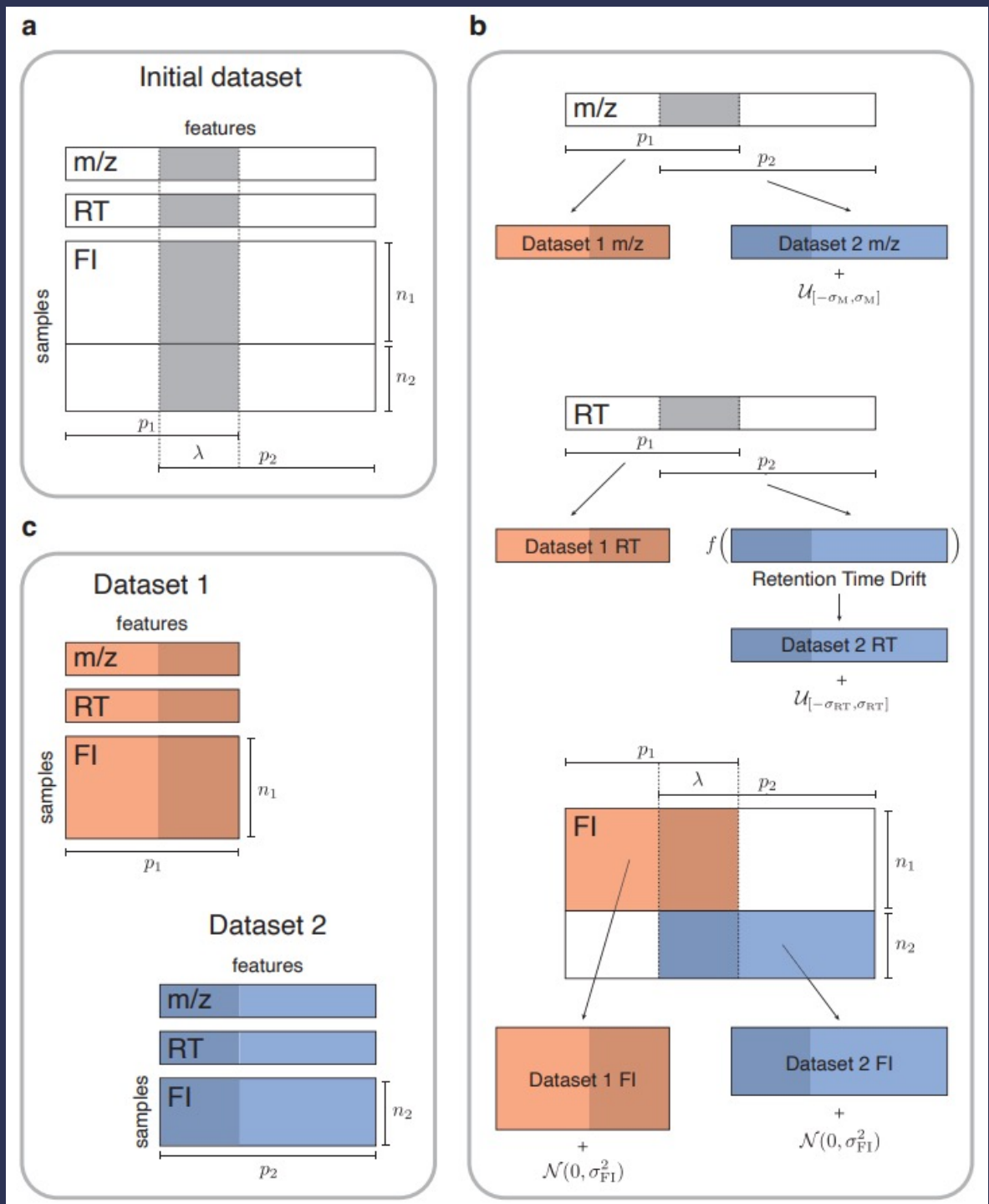
Implemented in Python

Runtime depends on the number of features. Typically less than 10 minutes for ~5000 features

Simulated data

Replicate a situation with 2 studies sharing a known set of features using an existing dataset of untargeted metabolomics on newborns

- Various setting investigated
- Compared with metabCombiner [Habra et al. 2021] and M2S [Pinto et al. 2022]

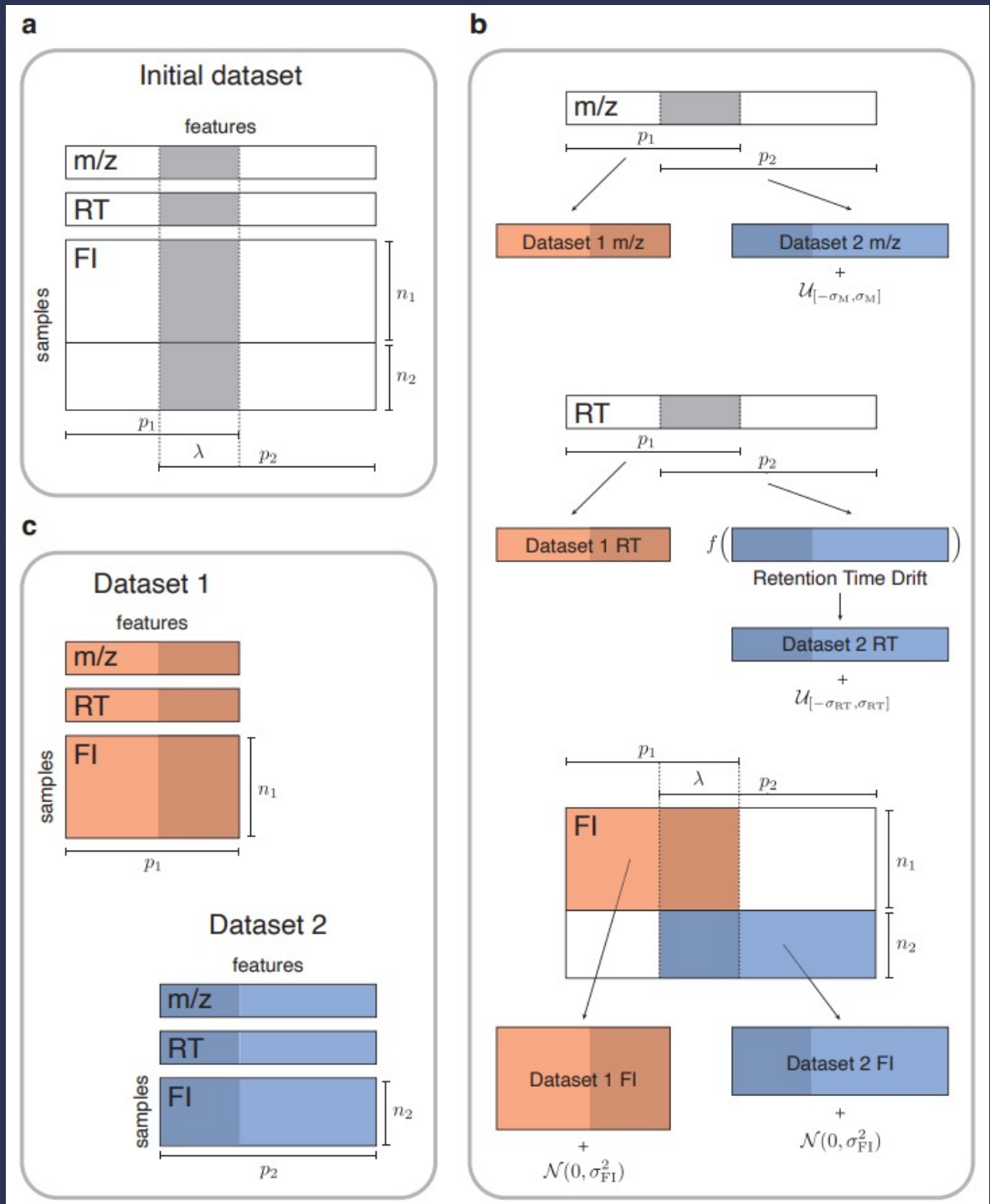


Simulated data

Replicate a situation with 2 studies sharing a known set of features using an existing dataset of untargeted metabolomics on newborns

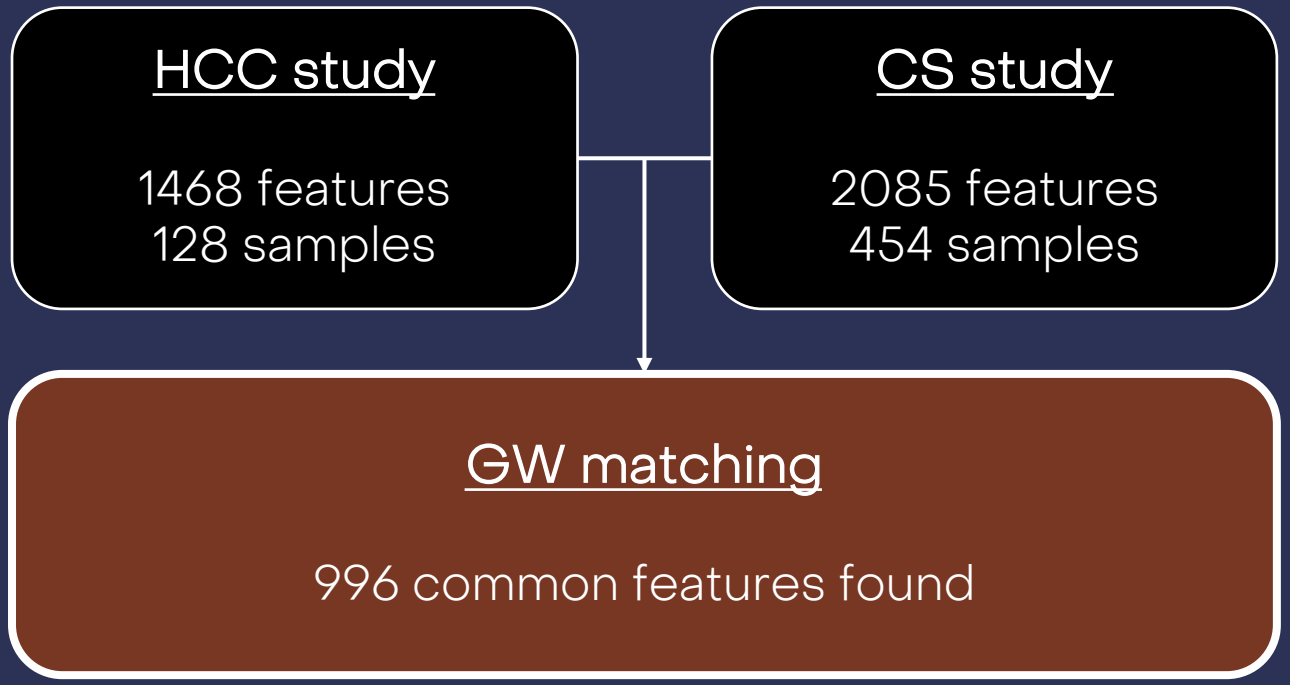
- Various setting investigated
- Compared with metabCombiner [Habra et al. 2021] and M2S [Pinto et al. 2022]

➢ Precision/recall were better in a majority of settings



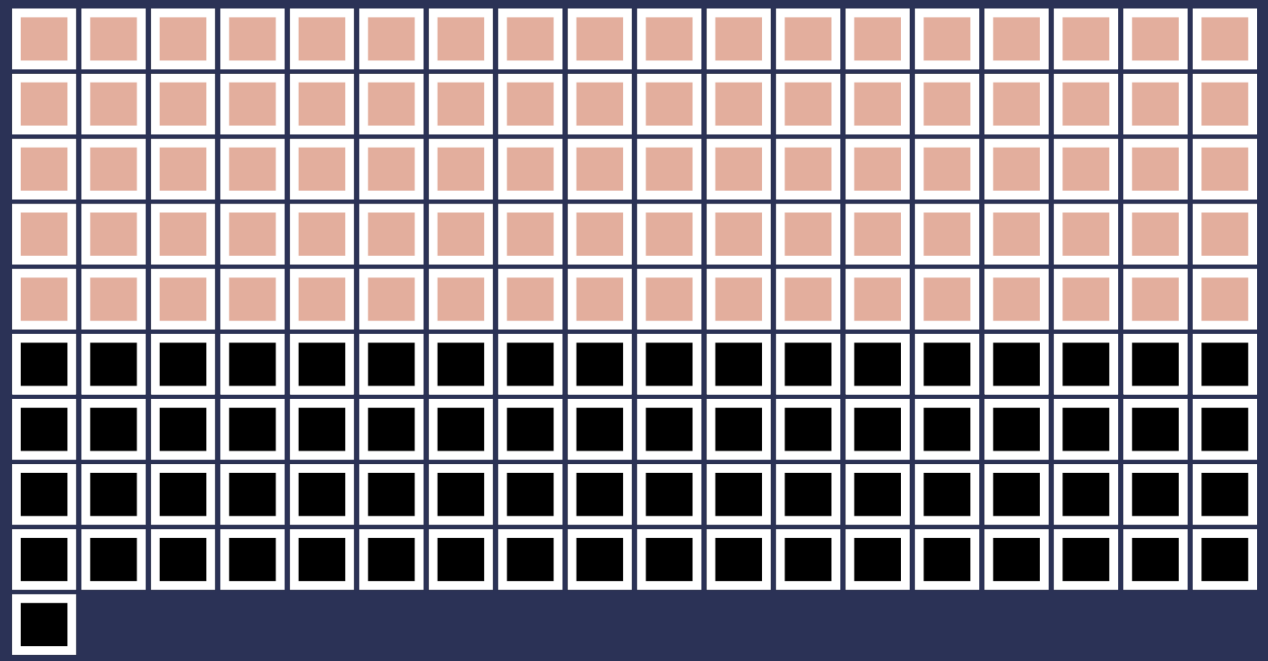
Application to EPIC data

Data from two EPIC studies used for alcohol biomarker discovery:
Cross-sectional (CS) study and hepatocellular carcinoma (HCC) study.



Application to EPIC data and validation

Data from two EPIC studies: Cross-sectional study and liver cancer study.

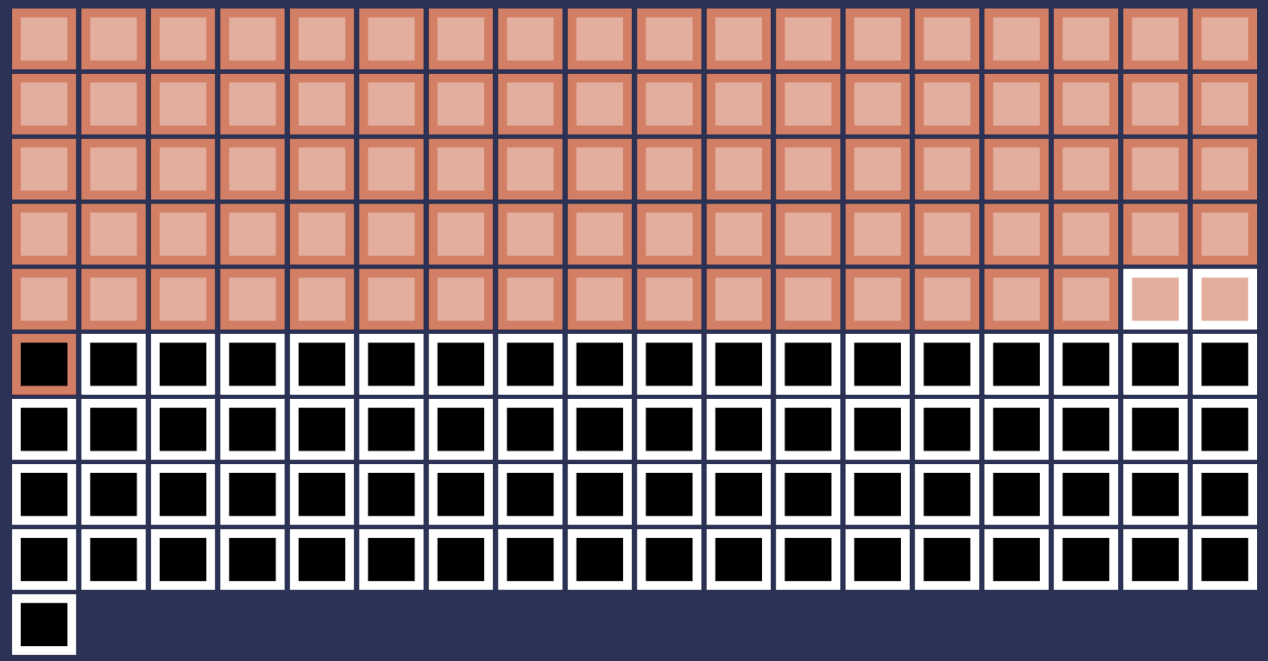


Manual examination (Loftfield *et al.*):
163 features from CS examined:

- 90 features also found in Liver
- 73 features unique to the CS study

Application to EPIC data and validation

Data from two EPIC studies: Cross-sectional study and liver cancer study.



Manual examination (Loftfield *et al.*):
163 features from CS examined:

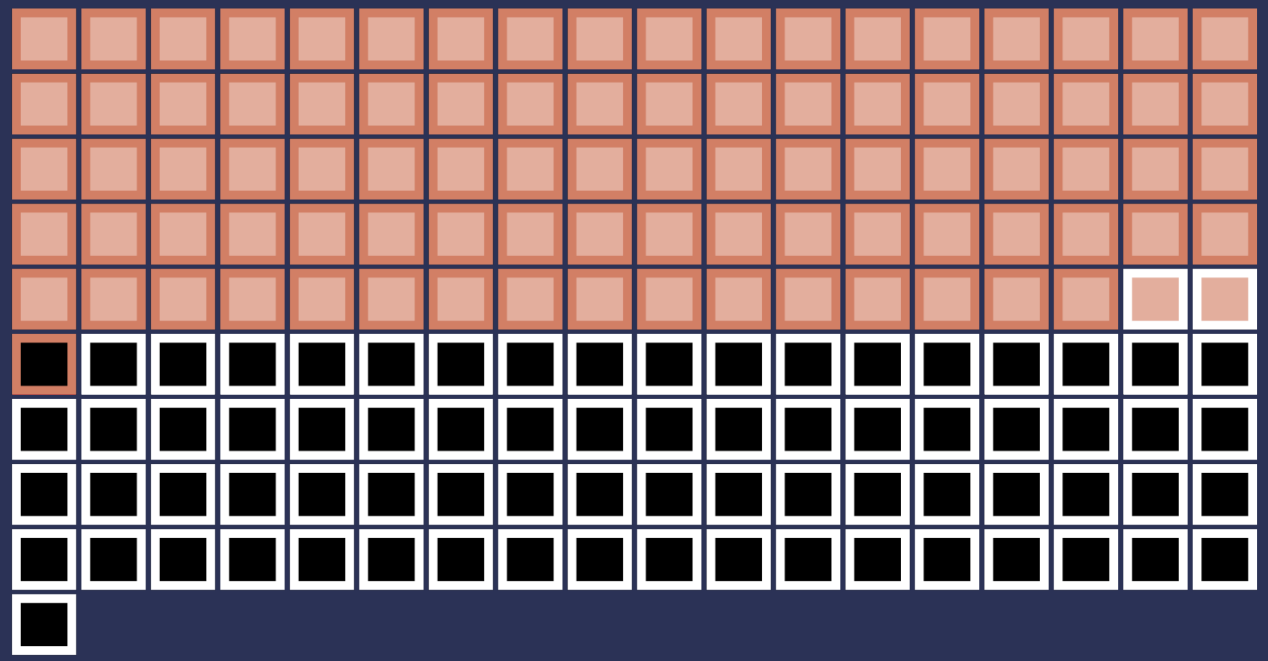
- 90 features also found in Liver
- 73 features unique to the CS study

- **89 common features found by GM**

(Recall: 0.98, precision: 0.99)

Application to EPIC data and validation

Data from two EPIC studies: Cross-sectional study and liver cancer study.



Manual examination (Loftfield *et al.*):
163 features from CS examined:

- 90 features also found in Liver
- 73 features unique to the CS study

- **89 common features found by GM**
- metabCombiner performed poorly (~20 matches recovered), M2S's optimal parameter combination was on par with GM

Application to EPIC data and validation

Data from two EPIC studies: Cross-sectional study and pancreatic cancer study.

- 987 common features found

- 65 out of 66 pairs recovered (same as M2S for optimal parameter tuning)
  - Recall: 0.98
- 7 additional pairs (11 for M2S)
  - Precision: 0.89

Application to EPIC data and validation

Data from two EPIC studies: Cross-sectional study and pancreatic cancer study.

- 987 common features found

- 65 out of 66 pairs recovered (same as M2S for optimal parameter tuning)
  - Recall: 0.98
- 7 additional pairs (11 for M2S)
  - Precision: 0.89

- Manual assessment found 2 good matches amongst the 7, the others were uncertain

Discussion

- Better performance than existing approaches
  - Compared with metabCombiner (Habra et al. 2021) and M2S (Pinto et al. 2022)
  - Better performance on simulated data and on EPIC data

- Perspectives
  - Extension to data where isotopic peaks/prior knowledge are available
  - Application to annotated data for EPIC Norfolk samples
  - Assess performance when data come from different studies, using different platforms

Acknowledgements

- George Stepaniants, Philippe Rigollet, Vivian Viallon
- Collaborators from IARC: Pekka Keski-Rahkonen, Mazda Jenab, Augustin Scalbert

References

- Mémoli, F. (2011) Gromov–Wasserstein Distances and the Metric Approach to Object Matching. *Found Comput Math* 11,17–487

- Habra H. *et al.* (2021) metabCombiner: Paired Untargeted LC-HRMS Metabolomics Feature Matching and Concatenation of Disparately Acquired Data Sets *Analytical Chemistry* 93(12), pp.5028-5036
- Pinto RC. *et al.* (2022) Finding Correspondence between Metabolomic Features in Untargeted Liquid Chromatography–Mass Spectrometry Metabolomics Datasets *Analytical Chemistry* 94(14), pp.5493-5503

International Agency for Research on Cancer

World Health Organization