OXFORD (GIGA)^n SCIENCE

GigaScience, 2022, **11**, 1–9

DOI: 10.1093/gigascience/giac101

TECH NOTE

Alignstein: Optimal transport for improved LC-MS retention time alignment

Grzegorz Skoraczyński *, Anna Gambin and Błażej Miasojedow

Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Stefana Banacha 2, 02-097 Warsaw, Poland
*Correspondence address. Grzegorz Skoraczyński, Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Stefana Banacha 2, 02-097 Warsaw, Poland. E-mail: g.skoraczynski@mimuw.edu.pl

Abstract

Background: Reproducibility of liquid chromatography separation is limited by retention time drift. As a result, measured signals lack correspondence over replicates of the liquid chromatography–mass spectrometry (LC-MS) experiments. Correction of these errors is named retention time alignment and needs to be performed before further quantitative analysis. Despite the availability of numerous alignment algorithms, their accuracy is limited (e.g., for retention time drift that swaps analytes' elution order).

Results: We present the Alignstein, an algorithm for LC-MS retention time alignment. It correctly finds correspondence even for swapped signals. To achieve this, we implemented the generalization of the Wasserstein distance to compare multidimensional features without any reduction of the information or dimension of the analyzed data. Moreover, Alignstein by design requires neither a reference sample nor prior signal identification. We validate the algorithm on publicly available benchmark datasets obtaining competitive results. Finally, we show that it can detect the information contained in the tandem mass spectrum by the spatial properties of chromatograms.

Conclusions: We show that the use of optimal transport effectively overcomes the limitations of existing algorithms for statistical analysis of mass spectrometry datasets. The algorithm's source code is available at https://github.com/grzsko/Alignstein.

Keywords: liquid chromatography–mass spectrometry, retention time alignment, Wasserstein distance, simplex algorithm

Introduction

Advances in liquid chromatography–mass spectrometry (LC-MS) have provided a remarkable insight into the functioning of the organisms, ranging from protein level [1], through tissue [2] to environmental networks [3]. All of these research studies benefit from the possibility to separate complex mixtures in the liquid chromatographic column and then measure the analytes with high-throughput mass spectrometry. Although LC-MS systems provide precise answers to both quantitative and qualitative biological and medical questions, designing algorithms for efficient and precise analysis of LC-MS datasets remains challenging.

One of these challenges is the correction of errors caused by retention time (RT) drift. It limits the reproducibility of LC separation, which is important for experiments usually acquired in many (even hundreds) replicates. RT drift became a significant obstacle with the emergence of high-performance chromatography (HPLC) and ultra-performance chromatography (UPLC) technologies. For example, nanoflow UPLC column separation takes a relatively long time, usually up to several hours. For these experiments, the elution time of peptides may vary up to 5 minutes [4] or even 10 minutes [1].

RT drift can be corrected by the experimental protocol only to a limited extent [5]. It may change the whole gradient or affect only single peaks. These changes may be caused by various reasons such as the unstable mobile phase, the column change or degradation, sample chemical instability, or imprecise experiment setup [6–8].

RT drift requires a correction, usually named the RT alignment. It results in the correspondence of signals across runs [9]. For example, in proteomics, the signal correspondence of the same peptides is needed for further applying label-free quantification (LFQ) for which samples must be measured separately [10, 11]. Moreover, for LFQ techniques, we cannot obtain the correspondence any other way because analytes do not have any additional information, such as metabolic labels, or chemical tags [12, 13].

Here, we present a novel alignment algorithm named Alignstein (cf. Fig. 1). It finds the correspondence of initially detected features (i.e., convex sets of peaks representing the signal of a single analyte). It overcomes the limitations of currently existing algorithms and properly resolves the correspondence of analytes of swapped elution order. To achieve this, we take advantage of the generalization of the Wasserstein distance (GWD) [14] to compare multidimensional features. To obtain the most feasible alignment results, Alignstein has formulated a complex optimization signal-matching problem, for which we use clustering and network flow algorithms to achieve a computationally tractable outcome.

This article is organized as follows. First, we characterize Alignstein and analyze how it deals with the swapped signals. Then, we validate the algorithm on publicly available benchmark datasets. Finally, we show the applicability of our approach to detecting corresponding biomarkers in differing samples.

Received: June 10, 2022. **Revised:** August 24, 2022. **Accepted:** September 30, 2022

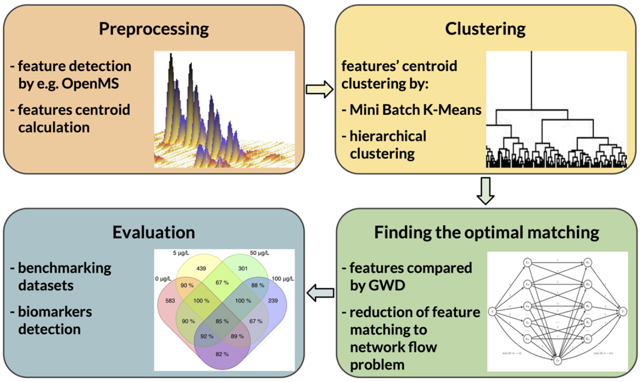
Preprocessing
- feature detection by e.g. OpenMS
- features centroid calculation
Clustering
features' centroid clustering by:
- Mini Batch K-Means
- hierarchical clustering
Evaluation
- benchmarking datasets
- biomarkers detection
Finding the optimal matching
- features compared by GWD
- reduction of feature matching to network flow problem

Figure 1: The outline of the Alignstein algorithm. It starts with feature preprocessing, for which then centroids are computed and clustered. As a next step, the problem of optimal feature matching is solved. The result is obtained with regard to prior clustering and can be further analyzed and verified.

Findings

The problem: resolving swaps

RT drift may swap the order of eluting analytes. In the proteomic experiment (cf. Methods), we analyzed that about 3% of all feature pairs are swapped between two chromatograms. Although many of the available algorithms properly align most signals, still they fail to resolve swaps.

Most approaches to RT alignment are so-called warping algorithms—for example, OpenMS [15], MetAlign [16], MZMine 2 [17], SIMA [18], the solution proposed by Zhang [19], DI-AlignR [20], and the solution proposed by Chiung-Ting Wu et al. [21]. These algorithms consist of applying a warping function that transforms the chromatograms by shifting, stretching, and squeezing. These transformations result in a close distance between corresponding signals. After alignment, however, further feature detection and matching are still required to obtain the signal correspondence. These algorithms' applicability is limited because the warping function is applied under the assumption that ions elute monotonically with RT. Thus, they are not able to deal with elution order swaps.

Alternatively, a rarer implemented approach is feature matching—for example, OpenMS [15] (both warping and matching algorithm), MassUntangler [22], LWBMatch [23], the solution proposed by Wandy et al. [24], MS-Dial [25], and Quandenser [26]. Algorithms by feature matching find the correspondence between initially detected features of 2 or more chromatograms. Corresponding features represent the same analyte and further will be referred to as consensus features. To the best of the authors' knowledge, all matching algorithms reduce multidimensional features to 1-dimensional extracted ion chromatograms or a single point with a monoisotopic peak m/z and average RT value, ignoring the information of isotopic envelope or feature span over the RT dimension. Without feature spatial characteristics and information of coeluting ions, elution order swaps are practically undetectable [8]. The main reason for this simplification lies in the difficulty to find multidimensional feature dissimilarity measures. Typically, Euclidean distance between points or 1-dimensional cosine-like spectra similarity scores is applied [27, 28]. Although the limitations of these scores are known, still there is a shortage of their effective improvements [28, 29].

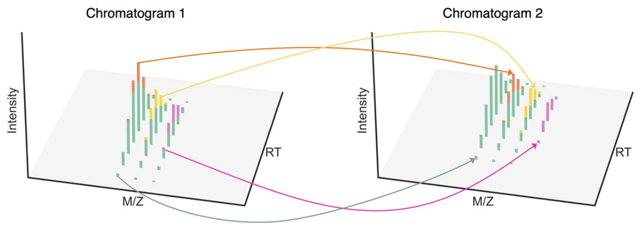

Figure 2: The optimal transport plan between 2 features. The Wasserstein distance captures not only the distance of feature drift along the RT dimension but also spatial differences between features. Here, the left feature consists of 3 ions, and the right feature consists of 4 ions. To properly capture this difference, part of the signal must be transported between different ions (denoted with arrows) and thus the transport cost (the Wasserstein distance) is higher.

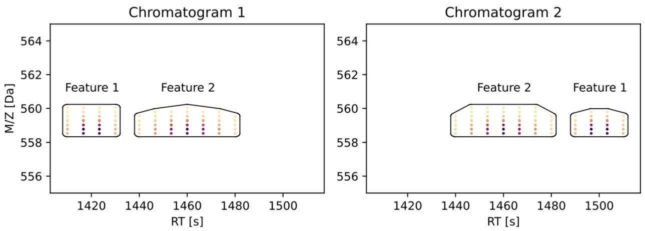

Figure 3: Example of swapped features. They represent 4 times charged peptides: HTALYSSDSVRNVRKKDTTG (Feature 1) and HTAIYSSDSVRNVRKKDTTG (Feature 2). Isotopic envelopes were generated using the IsoSpec tool [30] and smoothed over RT with a Gaussian filter. Retention times were predicted using the Pyteomics package [31]. The Euclidean distance between corresponding shifted features reduced to a point is 0.0 and 80.0 and between noncorresponding features is 40.0 and 40.0, whereas GWD equals 0.3 and 80.3 for corresponding features and 46.3 and 46.3 for noncorresponding features. For such an example, a simple feature-matching algorithm using GWD would match the features correctly, and for the Euclidean distance, this solution would be ambiguous.

The solution: the Alignstein algorithm

Alignstein is the RT alignment algorithm by feature matching that properly deals with features of swapped order. It is possible because the algorithm represents features by all signals contained within their boundaries. To cope with this representation, we use the generalization of the Wasserstein distance as a feature dissimilarity measure. It originates from the optimal transport theory and has been recently attracting growing attention to various problems of mass spectrometry [27, 32–35]. Its design significantly differs from currently existing similarity scores, and thus it overcomes the majority of their limitations. The Wasserstein distance describes the cost of the optimal way how to transform one feature into the other one. The transformations include not only shifting the signal from one feature to another but also splitting or combining the signal between peaks (cf. Fig. 2). The key strength of Wasserstein distance is the ability to compute features' similarity by their spatial shape (cf. Fig. 3). Moreover, it easily scales with dimension. Generalizing the Wasserstein distance allows comparing noisy features by introducing an appropriate penalty. This provides a highly flexible measure for effective computing feature distance and similarity.

Alignstein aligns chromatograms by finding consensus features. It is done in 2 phases (cf. Fig. 1): at first, feature centroids are clustered to find candidates for consensus features, which are then verified by the feature-matching phase. During the

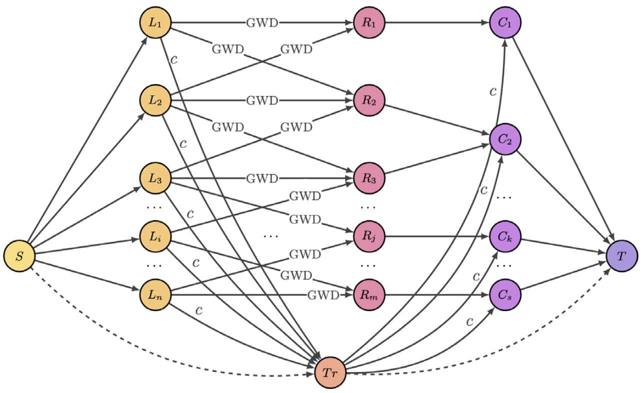

Figure 4: Flow network for finding the optimal feature matching. This matching is computed between selected chromatograms, denoted by n features $L_1, \ldots, L_n$ and m features from the rest of chromatograms, denoted by nodes $R_1, \ldots, R_m$. Clusters are denoted by s nodes $C_1, \ldots, C_s$. Nonzero costs are described by edge labels. The cost between features L_i and features R_j is equal to GWD between them. Additional node Tr ("trash") gives the possibility to not match the feature with cost c. Every edge has capacity equal to 1, except edge between S (source) and Tr and edge between Tr and T (sink) with capacities equal to $\max\{0, s - n\}$ and $\max\{0, n - s\}$, respectively (at most, one of them has nonzero capacity). Edges between $R_1, \ldots, R_m$ and $C_1, \ldots, C_s$ give the restriction that any feature can be matched with at most 1 cluster. As a result, we take all matchings (L_i, C_k). We recognize the consensus feature by its cluster.

latter phase, the algorithm computes the optimal feature matching, which represents the most similar feature pairs throughout all chromatograms (cf. Methods). We solve this problem by reducing it to finding the maximum flow of minimum cost in an appropriate flow network (cf. Fig. 4). Consensus features are then created from optimal feature matching with regard to initial centroid clustering. Such a formulation allows for aligning chromatograms without a requirement for a reference sample or a prior feature identification. It also easily scales with a number of input chromatograms. Finally, this algorithm is not limited to correcting RT perturbations in repeated experimental runs; it also accurately aligns the majority of detected corresponding biomarkers from samples of different experimental treatments.

Dealing with swapped signal

We assessed that Alignstein properly matches swapped features. For this purpose, we collected over 580 identified features from the chromatograms obtained from Barranger et al.'s work [3] (see Methods). We simulated RT drift by randomly moving features within range (−150 s, 150 s) in the RT dimension and within range (−0.3 Da, 0.3 Da) in the m/z dimension. These 2 sets of features, one with original features and the second with drifted features, represented chromatograms to be aligned. For such a formulation, about 2% (ca. 3,400) of feature pairs were swapped. We aligned these 2 sets and measured a number of properly matched features and a fraction of properly resolved swapped feature pairs. Our tool matched practically all drifted features (96%) and most swapped feature pairs (91%). We compared our results with 2 open-source feature-matching algorithms: OpenMS and LWBMatch. OpenMS had high feature matching precision, and it matched the majority of drifted features (80%). However, its accuracy drastically decreased when analyzing only swapped feature pairs (61%). LWBMatch had a significantly lower matching precision; it matched 24% of drifted features and only 3% of swapped feature pairs.

Algorithm validation on benchmark datasets

We evaluated the accuracy of our method by comparing alignment quality on public benchmark datasets. We reproduced the evaluation protocol from Lange et al. [36] (further referred to as the Critical Assessment of Alignment Procedures [CAAP] study). We analyzed 2 proteomic datasets from CAAP evaluation: P1 and P2, as well as 1 metabolomic: M1. The P1 set contained the analysis of *Escherichia coli* protein extracts and consisted of 6 fractions at different salt bumps, every fraction, in 2 different runs. Analogously, P2 contained the analysis of protein extract from *Mycobacterium smegmatis* in 5 fractions in every 3 replicated runs. M1 contained the analysis of leaf tissue extract from *Arabidopsis thaliana* in 44 repeated runs. To assess the correctness of alignment algorithms, the authors of the CAAP study proposed alignment precision and alignment recall measures (cf. Methods). Moreover, as proposed by the authors of the SIMA algorithm [18], we computed the F-score, which is a harmonic mean of alignment precision and recall.

We analyzed sets P1, P2, and M1 and compared Alignstein with the results of the OpenMS alignment algorithm [15] from the CAAP study. We chose OpenMS because it achieved significantly better results than the other tools and represented a state-of-the-art solution at the time of the original study. Moreover, we included in comparison the available results of algorithms published more recently: MZMine 2 [17], SIMA [18], MassUntagler [22] (only P1 set), and Wandy et al. [24].

Alignstein obtained highly competitive results in the CAAP evaluation. For the P1 dataset, it matched perfectly almost all features; its precision and recall were on average 0.94, similar to MZmine 2 and OpenMS (cf. Table 1, Supplementary Table S1). SIMA obtained slightly worse results, and the rest of the tools obtained lower values than SIMA. Interestingly, all tools achieved average alignment precision and recall no higher than 0.94. It may suggest that improperly matched features are too distant to be matched based on LC-MS information or ground truth is misspecified.

For the P2 set, we achieved the highest average alignment recall (on average 0.82); that is, our approach had a minimal number of unmatched features (cf. Table 1, Supplementary Table S2). It had a lower precision on average equal to 0.73 and was second only to OpenMS. Overall, we obtained the best average *F*-score value, equal to 0.77.

For the M1 dataset, Alignstein achieved competitive results: precision equal to 0.88, recall of 0.91, and *F*-score of 0.89. This confirms that Alignstein scales effectively with the number of input chromatograms. We measured the time of alignment computation; results are presented in Table 2.

Application to the detection of specific biomarkers

Alignstein can detect specific biomarkers in medical applications or biological analysis. To verify this, we analyzed the dataset from Barranger et al. [3]. It contained LC–tandem MS (MS/MS) chromatograms of intestinal protein from marine mussels exposed *in vivo* to various benzo[a]pyrene (BaP) concentrations (0, 5, 50, 100 μg/L).

We checked if Alignstein recognizes MS/MS information by spatial properties of LC-MS features. To assess this, we detected LC-MS features and annotated them with peptide MS/MS identifications. The accuracy of alignment was quantified using proposed identification recall (IR) defined as follows. We chose all repeating identifications that have annotated features and computed a

Table 1: Comparison of alignment precision (P), alignment recall (R), and F-score (F). For P1 and P2 sets, average over fractions is computed. Dash marks result not presented in the original papers.

		Alignstein	OpenMS	MZMine 2	Wandy et al.	SIMA	MassUntangler
P1	P	0.94	0.94	0.94	0.88	0.94	0.87
	R	0.94	0.94	0.94	0.89	0.92	0.79
	F	0.94	0.94	0.94	0.88	0.93	0.83
P2	P	0.74	0.83	0.68	0.72	0.72	—
	R	0.83	0.72	0.75	0.72	0.75	—
	F	0.78	0.77	0.71	0.72	0.74	—
M1	P	0.88	0.69	0.74	—	0.75	—
	R	0.91	0.87	0.91	—	0.92	—
	F	0.89	0.77	0.82	—	0.83	—

Table 2: Alignstein runtimes on benchmark CAAP datasets.

P1	P2	M1
10 s	15 s	15 min 38 s

For the P1 and P2 dataset, wall-time was measured for a single fraction. For the M1 dataset, wall-time was measured for the whole dataset.

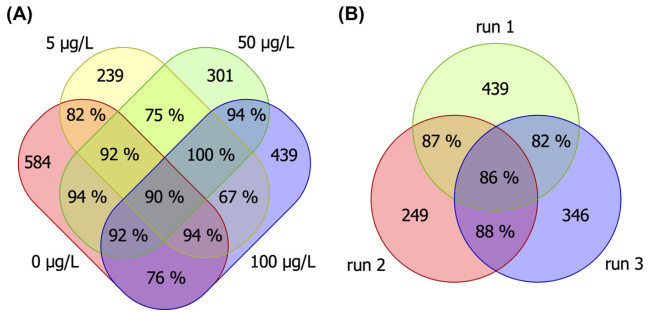

Figure 5: Identification recall calculated separately for every chromatogram subset. (A) For aligned chromatograms over all BaP concentrations. (B) For replicates (1, 2, 3) of the sample of 100 μg/L BaP concentration. Sets represent chromatograms, intersections contain identification recall, and the nonoverlapping part of the sets contains the number of feature-annotated identifications.

fraction of them that were properly aligned (cf. Methods). For every BaP concentration, we computed IR for all aligned technical replicates of the sample. We achieved satisfactory results IR equal to 81%, 78%, 85%, and 86%, respectively, for BaP concentrations of 0, 5, 50, and 100 μg/L. As a baseline, we repeated this analysis for the OpenMS algorithm, which achieved similar results with IR equal to 81%, 76%, 85%, and 83%. Moreover, we calculated the IR separately for every subset of all aligned chromatograms (see Methods). This demonstrated that our approach uniformly treats all chromatograms (cf. Fig. 5A and Supplementary Fig. S1).

Moreover, we checked if Alignstein can detect corresponding biomarkers for LC-MS measurements of samples under different experimental conditions. For this purpose, we repeated the analysis above by aligning chromatograms across all BaP concentrations. The overall IR was equal to 85%. Contrary to the previous experiment, IR for OpenMS had fallen to 0.75%. Analogously as earlier, Alignstein's results were uniform over all chromatogram subsets (cf. Fig. 5B) with IR values not lower than 67%, reaching even 100% for some subsets of repeated identifications. This proves that, despite the varying experimental conditions, our solution is able to correctly align most corresponding features without accuracy loss. Finally, this experiment shows that it may be applied as a tool for biomarker screening in LC-MS analysis.

Discussion

Alignstein is a novel, original algorithm for LC-MS alignment based on the GWD feature dissimilarity measure. This allows for incorporating not only distances between features but also their spatial differences and thus more accurate feature alignment. The GWD emerges to be a key solution for correctly aligning signals with a swapped elution order, as demonstrated in the previous sections.

In addition to correctly resolving feature swaps, Alignstein has more advantages over the majority of alignment algorithms. It requires no prior feature identification, so LC-MS data without additional tandem mass spectra suffice as input to the algorithm. Moreover, our approach makes no assumptions about the characteristics of the analyzed chromatograms, so it is not limited to one type of data (e.g., proteomic or metabolomic). Still, specific properties of the analyzed data (e.g., maximum drift size) can be passed as algorithm parameters. Finally, it treats uniformly all analyzed chromatograms, and thus it does not require a reference sample.

Alignstein requires only the prior feature detection as a data preprocessing step. Although approaches with this requirement are criticized [8, 20], we argue that the analysis with detected features is more accurate than the analysis of raw chromatograms. Properly executed feature detection effectively discriminates regions of high signal-to-noise ratio from chromatograms [37]. Moreover, multidimensional feature detection is crucial for collecting information about coeluting ions (e.g., isotopic envelopes of compounds). Without this, any alignment algorithm might yield inaccurate results by aligning signals across isotopic envelopes.

Besides advantages, Alignstein has also limitations. It correctly matches most features, but it happens to fail to match distant features. This mismatch can be explained by interpreting GWD as a sum of 2 costs: the cost of transporting the feature along the RT (to eliminate drift) and the cost of transformation (to incorporate feature–feature spatial differences). For a pair of distant, corresponding features, the cost of transport along the RT far exceeds the cost of transformation. For this reason, even highly dissimilar but much closer features may camouflage the correct feature correspondence. This can be particularly troublesome for complex datasets having a significant number of features, which are densely packed within chromatograms. This limitation can be only partially corrected by adjusting GWD parameters because most corresponding feature pairs have RT differences of less than 10 seconds (cf. Supplementary Fig. S2), and thus the GWD parameters must be optimized for small feature distances. One of the possible solutions is to incorporate additional information for alignment, for example, MS/MS data. Thus, we plan to extend our algorithm to deal with LC-MS/MS datasets in a data-independent acquisition mode.

In conclusion, Alignstein correctly aligns chromatograms, as we have shown in the biomarker detection experiment, by reproducing the CAAP evaluation study, as well as in swaps resolving computational comparison. Its highly competitive matching accuracy is the result of applying the GWD as a feature dissimilarity measure, which allows matching features without reducing feature spatial information or the dimension of data. Thus, Alignstein is capable of detecting nonobvious signal patterns and finding optimal alignment. Our solution provides a solid basis for further applications of optimal transport theory to the multidimensional problems of automated analysis in mass spectrometry. We hope that the optimal transport-based distances will become a new paradigm as a measure of spectra dissimilarity and will allow the construction of highly effective, robust, and accurate algorithms for mass spectrometry analysis.

Methods

Feature dissimilarity measure

The most common approach to comparing mass spectra is a cosine-like similarity score [28, 38]. Despite its popularity, this class of scores is not applicable to feature alignment, because they are not scalable with dimension and cannot effectively compare spectra of significantly different molecules [29]. To address these limitations, we propose the Wasserstein distance [39] with additional generalizations [14, 40] as a feature dissimilarity measure.

The Wasserstein distance is a metric based on optimal transport theory. It describes how to optimally transform one feature into the other one. These transformations may include shifting the signal as well as splitting or combining the signal between peaks (cf. Fig. 2). Formally, suppose that we have 2 discrete features, μ and ν, so that $\mu(x)$ is the intensity of μ at m/z value x. Then we define the transport plan T so that $T(x, y)$ corresponds to the amount of signal that is transported from a peak x of feature μ to peak y of feature ν. The transport cost is the sum of amounts of transport between all pairs of peaks multiplied by the distance between peaks:

$$\sum_{x,y} T(x, y) \cdot d(x, y), \qquad (1)$$

where $d(x, y)$ is a distance between peaks x and y. For this setup, we have chosen $d(x, y)$ to be ℓ_1 distance (a Manhattan distance). The Wasserstein distance W is the minimal transport cost of all possible transport plans T:

$$W(\mu, \nu) = \min_{T} \sum_{x,y} T(x, y) \cdot d(x, y). \qquad (2)$$

Besides effectiveness, we observed that Wasserstein distance unsatisfactorily deals with noisy features. To overcome this limitation, we use a GWD as proposed by Chizat et al. [40]. GWD differs mainly from Wasserstein distance by the possibility of omitting the transporting part of the signal with a constant penalty. More specifically, GWD allows omitting the transport of signal on a distance larger than the user-defined λ parameter with a constant penalty proportional to λ and the amount of not transported signal:

$$W(\mu, \nu) = \min_{T} \sum_{x,y} \Big(T(x, y) \cdot d(x, y) + \lambda \cdot F(T_\mu, \mu) + \lambda \cdot F(T_\nu, \nu) \Big), \qquad (3)$$

where T_μ and T_ν are the marginals of the transport plan. F is a divergence chosen so that the approximation of the transport plan T to features μ and ν is possible. To compute GWD, we regularize it with the entropic term, which allows for fast and numerically stable computation, using a scaling Sinkhorn–Knopp approximation algorithm [41]. Fully formal distance derivation is available in Supplementary Material sections 1 and 2.

Alignstein algorithm scheme

Alignstein is an algorithm for LC-MS alignment. Here, the alignment is formulated as finding the correspondence of detected features, which represent the same chemical entities (e.g., ions, compounds). Specifically, the algorithm takes chromatograms with detected features as an input, and the outcome of the algorithm is a list of consensus features. Consensus features are sets of corresponding features from distinct chromatograms. The algorithm outline is depicted in Fig. 1 and pseudocode is available in Algorithm 1.

Algorithm 1: Alignstein algorithm

```
Input: chromatograms $ch_1, \ldots, ch_n$,
Result: consensus features $c_1, \ldots, c_s$
// Preprocessing phase
$features_1, \ldots, features_n \leftarrow$
  preprocessFeatures($ch_1, \ldots, ch_n$)
// Centroid clustering phase
$centroids \leftarrow \emptyset$
forall $features_i$ do
    $centroids \leftarrow$
      $centroids \cup$ {centroids of all features from $features_i$}
$clusters \leftarrow$ cluster $centroids$
// Matching phase
forall $features_i$ do
    $matching_i \leftarrow$ match features from $features_i$ to
      $\bigcup_{j \in \{1,\ldots,n\} \setminus i} features_j$
$c_1 \ldots, c_m \leftarrow$
  createConsensusFromMatchings($matching_1, \ldots, matching_n$)
```

Alignstein starts with feature preprocessing. If the features are not provided by the user, it detects them using the Feature Finder algorithm from the OpenMS package. Features are represented as a set of all signal peaks contained within the boundaries of detected features. For further processing, Alignstein normalizes features and scales the RT so that the RT dimension variation becomes a similar order of magnitude as the m/z dimension variation. Scaling is done by dividing the RT by a factor proportional to the ratio of the average feature length (along the RT axis) and the average feature width (along the m/z axis).

After preprocessing, alignment consists of 2 phases: the centroid clustering phase and then the feature-matching phase. During the former one, centroids of features from all chromatograms are clustered using Mini-Batch K-Means [42] and hierarchical clustering algorithms. Clustering is computed to create candidates for consensus features, which are then verified by the feature-matching phase. During this phase, the algorithm searches for pairs of the most similar features across all chromatograms. It is done by finding the feature matching of minimal cost, where the cost is equal to the sum of GWDs between the matched features. We interpret this problem as finding the maximal flow of minimum cost in an appropriately designed flow network, in which we compare iteratively every chromatogram with the rest of the chromatograms. In Fig. 4, we show a flow network for a single chromatogram. The flow of minimum cost is obtained using the network simplex [43] algorithm. This minimization problem has formulated restrictions, such that features may be not matched with a constant penalty, or at most, 1 feature may be matched to features within 1 cluster. These restrictions ensure that features are matched appropriately as required—for example, the algorithm treats uniformly all input chromatograms, at most one feature from every chromatogram would be chosen to consensus features, and so on. Finally, consensus features are created via clusters that have been matched as most similar in optimal matching. A more detailed algorithm description is available in Supplementary Material section 3.

In the special case, when only 2 chromatograms are aligned, the clustering phase is omitted and consensus features are created by finding the optimal matching between 2 feature sets (cf. Supplementary Fig. S3).

Implementation details

Alignstein is implemented as a Python 3 package and available at [44]. It uses C++ implementation of GWD in the MassSinkhornmetry package available at [45]. For centroid clustering, we used clustering algorithms implemented in the scikit-learn package [46, 47]. For solving the minimum cost flow problem, we used the data structures and algorithms implemented in the NetworkX [48, 49] package.

Alignstein benchmarking details

We validated the Alignstein algorithm by reproducing the evaluation protocol from the CAAP study [36]. It was the analysis and comparison of 6 alignment algorithms: OpenMS [15], msInspect [50], MZmine 1 [51], SpecArray [52], XAlign [53], and XCMS [54].

We analyzed 2 proteomic datasets (P1 and P2) and 1 metabolomic dataset (M1) from the CAAP study. For all sample sets, preparation and analysis protocols are described in the original study. For the metabolomic set as well as for every fraction at different salt bumps (0, 20, 40, 60, 80, and 100 mM ammonium chloride) of both proteomic sets, the authors prepared a set of ground-truth consensus features, which represent feature correspondence over chromatograms of significantly high confidence.

To assess the accuracy of alignment, the authors of the CAAP study proposed the generalization of precision and recall as alignment precision and alignment recall. Alignment precision measures how the given ground-truth consensus feature was split over tool consensus features (i.e., it reflects the number of false positives). Alignment recall measures how many features of a given ground-truth consensus feature are found by the algorithm (i.e., it reflects the number of false negatives). Both alignment precision and recall are calculated as an arithmetic mean over all ground-truth consensus features. Furthermore, the authors of SIMA [18] and Wandy et al. [24] proposed the *F*-score, which is the harmonic mean of alignment precision and alignment recall ($\frac{2 \cdot P \cdot R}{P+R}$, where *P* is alignment precision and *R* is alignment recall) to express the balance of alignment precision and alignment recall.

We used input chromatograms as mzML and mzXML files and features as featureXML files provided by authors of the CAAP study. We measured alignment precision and recall using an evaluation script written in R programming language by the authors of this study. Computation was done on a computer with a Linux operating system and 24 Intel Xeon E5-2620 2.10 GHz processors + 62 GB RAM. We measured wall time using the Linux built-in `time` command. More details on CAAP benchmarking are provided in the Supplementary Material section 4.

Mussels toxicological response experiment summary

For assessment of Alignstein's ability to detect specific biomarkers, we analyzed chromatograms originally created in Barranger et al. [3]. The original study aimed to measure the effects of polluting the environment of marine mussels (*Mytilus galloprovincialis*) with fullerene (C60) and BaP. For this purpose, the authors performed a proteomic analysis.

Mussels were collected in Trebarwith Strand, Cornwall, UK, and were exposed *in vivo* to C60 and BaP at concentrations 0, 5, 50, and 100 μg/L as described in the original study. For proteomic analysis, mussel intestinal proteins were collected. After digestion and purification, the peptides were analyzed by the LC-MS/MS system with the data-dependent acquisition (DDA) mode as described in Sequiera et al. [55]. In summary, peptides were separated on a Dionex, Camberly, UK Ultimate 3000 RSLC nanoflow system: Acclaim PepMap C18 nano column (75 μm × 25 cm, 3 μm, 100 Å), plus bypass, including a linear gradient of 96% buffer A (0.5% acetic acid) and 4% buffer B (80% acetonitrile in 0.5% acetic acid) to 60% buffer A and 40% buffer B, with a flow rate of 300 mL/min for 120 minutes. Separated analytes were analyzed in an Orbitrap Velos Pro FTMS (Thermo Finnigan, Bremen, Germany) with positive ion mode ionization with a Proxeon, Thermo Fisher Hemel, UK nanospray ESI source. In each run, the 10 most abundant ions were further analyzed with additional collision-induced dissociation (CID) fragmentation (30% collision energy) in a linear ion trap spectrometer. For every BaP concentration from 0, 5, 50, to 100 μg/L, 3 replicates were obtained. Collected chromatograms for all BaP exposure levels were deposited in the ProteomeXchange Consortium PRIDE repository (PXD013805) [56, 57].

Data analysis for detection of repeating biomarkers

In downloaded chromatograms, we identified peptides using Comet [58, 59]. We obtained the database for peptide identification from the original work (taxa Mollusca, subcategory Bivalvia from Uniprot KnowledgeBase, and contaminants from the Global Proteome Machine [60]). The most important Comet search parameters were peptide mass tolerance of 10 ppm, trypsin as search enzyme, concatenated decoy search, and allowed missed enzyme cleavages no higher than 2.

We detected features in chromatograms using the OpenMS algorithm Feature Finder in the Centroided version. We annotated the detected LC-MS features with MS/MS Comet identifications. Peptide MS/MS identifications were represented in LC-MS by retention time in seconds and the ratio of the precursor neutral mass to the assumed charge. The feature was annotated with identification when LC-MS representation of identification was enclosed within feature boundaries. For further analysis, we considered annotated features.

For calculating IR, we computed the number of repeating identifications over chromatograms. For every repeating identification, we checked if annotated features were properly matched by Alignstein. IR was calculated as a ratio of the number of correctly aligned annotated repeating identifications and the total number of annotated repeating identifications.

Number of swaps estimation

We analyzed 2 replicates of 0 μg/L BaP concentration in the dataset described in the previous section. Computation was done for all pairs of annotated features with repeating identification in both chromatograms. We computed the fraction of these pairs that were swapped (i.e., a feature pair was considered a swap when the computed feature RT means of the same identifications in 2 replicates were in a different order).

Availability of Source Code and Requirements

- Project name: Alignstein
- Project homepage: [44]
- Operating systems: Linux, macOS
- Programming language: Python 3
- Other requirements: Python 3.6 or higher; dependency packages: MassSinkhornmetry, pyOpenMS, NumPy, SciPy, NetworkX, scikit-learn
- License: MIT
- Any restrictions to use by nonacademics: none
- RRID: SCR_022483
- bio.tools ID: alignstein

Data Availability

The Marine Mussels dataset was obtained from ProteomeXchange Consortium PRIDE repository under accession no. PXD013805. Benchmark datasets (P1, P2, M1), as well as evaluation script, were obtained from the CAAP webpage at [61]. Datasets P1 and P2 are originally available in Open Proteomic Database [62]. Snapshots of Alignstein source code and other data further supporting this work are openly available in the *GigaScience* repository, GigaDB [63].

Additional Files

Supplementary Fig. S1. Identification recall calculated separately for identifications repeating in every chromatogram subsets. (A) For replicates of the sample with 0 μg/L BaP. (B) For replicates of the sample with 5 μg/L BaP. (C) For replicates of the sample with 50 μg/L BaP. Sets represent replicates (chromatograms) of the same experiments, the inconjunct part of the set contains the number of feature-annotated identifications, and conjunctions contain identification recall.

Supplementary Fig. S2. Histogram of RT centroid differences between feature pairs annotated with the same identification. The histogram is computed for chromatograms from Barranger et al. [3], including replicates of a sample with 0 μg/L BaP. For better readability, outliers over 200 seconds are omitted. Most RT differences are not greater than 10 seconds.

Supplementary Fig. S3. Flow network for finding the optimal feature matching between n features of 1 chromatogram denoted by nodes $L_1, \dots, L_n$ and m features from the other chromatogram, denoted by nodes $R_1, \dots, R_m$. Nonzero costs are described by edge labels. The cost between features L_i and features R_j is equal to the GWD between them. Additional node T_r ("trash") gives the possibility to not match the feature with cost c. Every edge has capacity equal to 1, except edge between S (source) and T_r and edge between T_r and T (sink) with capacities equal to $\max\{0, s-n\}$ and $\max\{0, n-s\}$, respectively (at most, one of them has nonzero capacity). As a result, we take all matchings (L_i, R_j).

Supplementary Table S1. Detailed results for P1 set in CAAP comparison. P stands for alignment precision, R stands for alignment recall, and F stands for *F*-score.

Supplementary Table S2. Detailed results for P2 set in CAAP comparison. P stands for alignment precision, R stands for alignment recall, and F stands for *F*-score.

Abbreviations

BaP: benzo[a]pyrene; CAAP: Critical Assessment of Alignment Procedures; C60: fullerene; CID: collision-induced dissociation; DDA: data-dependent acquisition; GWD: generalized Wasserstein distance; HPLC: high-performance liquid chromatography; IR: identification recall; LC-MS: liquid chromatography–mass spectrometry; MS/MS: tandem mass spectrometry; *m/z*: mass-to-charge ratio; RT: retention time; UPLC: ultra-performance liquid chromatography.

Competing Interests

The authors declare that they have no competing interests.

Funding

G.S. was supported by Polish National Science Center grant number 2019/33/N/ST6/02949. A.G. and B.M. were supported by Polish National Science Center grant number 2018/29/B/ST6/00681.

Authors' Contributions

G.S. implemented and verified the algorithm. A.G. conceived the idea of the project and discussed the results. B.M. designed the algorithm and supervised the work. G.S., A.G., and B.M. cowrote the manuscript.

Acknowledgments

The authors thank Marta Nowak for her comments on the manuscript.

References

1. Runkle, KB, Kharbanda, A, Stypulkowski, E, *et al.* Inhibition of DHHC20-mediated EGFR palmitoylation vreates a dependence on EGFR signaling. *Mol Cell* 2016;**62**(3):385–96.
2. Sethi, MK, Thaysen-Andersen, M, Kim, H, *et al.* Quantitative proteomic analysis of paired colorectal cancer and non-tumorigenic tissues reveals signature proteins and perturbed pathways involved in CRC progression and metastasis. *J Proteomics* 2015;**126**:54–67.
3. Barranger, A, Langan, LM, Sharma, V, *et al.* Antagonistic interactions between benzo[a]pyrene and fullerene (C60) in toxicological response of marine mussels. *Nanomaterials* 2019;**9**(7):987.
4. Tomechko, SE, Liu, G, Tao, M, *et al.* Tissue specific dysregulated protein subnetworks in type 2 diabetic bladder urothelium and detrusor muscle. *Mol Cell Proteomics* 2015;**14**(3):635–45.
5. Zhou, B, Xiao, JF, Tuli, L, *et al.* LC-MS-based metabolomics. *Mol BioSyst* 2012;**8**(2):470–81.
6. Snyder, LR, Kirkland, JJ, Dolan, JW. *Introduction to Modern Liquid Chromatography*. New York: John Wiley, 2009.
7. Magnus Åberg, K, Alm, E, Torgrip, RJO. The correspondence problem for metabonomics datasets. *Anal Bioanal Chem* 2009;**394**(1):151–62.

8. Smith, R, Ventura, D, Prince, JT. LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Briefings Bioinformatics* 2013;**16**(1):104–17.
9. Lai, X, Wang, L, Witzmann, FA. Issues and applications in label-free quantitative mass spectrometry. *Int J Proteomics* 2013;**2013**:1–13.
10. Claudia, L, Nikolas, T, Franziska, H, *et al.* Strategies in relative and absolute quantitative mass spectrometry based proteomics. *Biol Chem* 2017;**398**(5–6):687–99.
11. Cox, J, Hein, MY, Luber, CA, *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* 2014;**13**(9):2513–26.
12. Dowell, JA, Wright, LJ, Armstrong, EA, *et al.* Benchmarking quantitative performance in label-free proteomics. *ACS Omega* 2021;**6**(4):2494–504.
13. Yunong, L, Liang, L. Retention time shift analysis and correction in chemical isotope labeling liquid chromatography/mass spectrometry for metabolome analysis. *Rapid Commun Mass Spectrometry* 2020;**34**(Suppl 1):e8643.
14. Peyré, G, Cuturi, M. Computational optimal transport: with applications to data science. *Foundations Trends Machine Learn* 2019;**11**(5–6):355–607.
15. Lange, E, Gröpl, C, Schulz-Trieglaff, O, *et al.* A geometric approach for the alignment of liquid chromatography—mass spectrometry data. *Bioinformatics* 2007;**23**(13):i273–81.
16. Lommen, A. MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal Chem* 2009;**81**(8):3079–86.
17. Pluskal, T, Castillo, S, Villar-Briones, A, *et al.* MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 2010;**11**:395.
18. Voss, B, Hanselmann, M, Renard, BY, *et al.* SIMA: simultaneous multiple alignment of LC/MS peak lists. *Bioinformatics* 2011;**27**(7):987–93.
19. Zhang, Z. Retention time alignment of LC/MS data by a divide-and-conquer algorithm. *J Am Soc Mass Spectrometry* 2012;**23**(4):764–72.
20. Gupta, S, Ahadi, S, Zhou, W, *et al.* DIAlignR provides precise retention time alignment across distant runs in DIA and targeted proteomics. *Mol Cell Proteomics* 2019;**18**(4):806–17.
21. Wu, CT, Wang, Y, Wang, Y, *et al.* Targeted realignment of LC-MS profiles by neighbor-wise compound-specific graphical time warping with misalignment detection. *Bioinformatics* 2020;**36**(9):2862–71.
22. Ballardini, R, Benevento, M, Arrigoni, G, *et al.* MassUntangler: a novel alignment tool for label-free liquid chromatography–mass spectrometry proteomic data. *J Chromatogr A* 2011;**1218**(49):8859–68.
23. Wang, J, Lam, H. Graph-based peak alignment algorithms for multiple liquid chromatography-mass spectrometry datasets. *Bioinformatics* 2013;**29**(19):2469–76.
24. Wandy, J, Daly, R, Breitling, R, *et al.* Incorporating peak grouping information for alignment of multiple liquid chromatography-mass spectrometry datasets. *Bioinformatics* 2015;**31**(12):1999–2006.
25. Tsugawa, H, Cajka, T, Kind, T, *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods* 2015;**12**(6): 523–6.
26. Matthew, T, Lukas, K. Focus on the spectra that matter by clustering of quantification data in shotgun proteomics. *Nat Commun* 2020;**11**:3234.
27. Moorthy, AS, Kearsley, AJ. *Pattern Similarity Measures Applied to Mass Spectra*. Cham, Switzerland: Springer International Publishing, 2020.
28. Kim, S, Zhang, X. Comparative analysis of mass spectral similarity measures on peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry. *Comp Math Methods Med* 2013;**2013**:1–12.
29. Huber, F, Ridder, L, Verhoeven, S, *et al.* Spec2Vec: improved mass spectral similarity scoring through learning of structural relationships. *PLoS Comp Biol* 2021;**17**(2):e1008724.
30. Lacki, MK, Valkenborg, D, Startek, MP. IsoSpec2: ultrafast fine structure calculator. *Anal Chem* 2020;**92**(14):9472–5.
31. Levitsky, LI, Klein, JA, Ivanov, MV, *et al.* Pyteomics 4.0: five years of development of a Python proteomics framework. *J Proteome Res* 2018;**18**(2):709–14.
32. Ciach, MA, Miasojedow, B, Skoraczyński, G, *et al.* Masserstein: linear regression of mass spectra by optimal transport. *Rapid Commun Mass Spectrometry* 2020;e8956.
33. Permiakova, O, Guibert, R, Kraut, A, *et al.* CHICKN: extraction of peptide chromatographic elution profiles from large scale mass spectrometry data by means of Wasserstein compressive hierarchical cluster analysis. *BMC Bioinformatics* 2021;**22**:68.
34. Seifert, NA, Prozument, K, Davis, MJ. Computational optimal transport for molecular spectra: the fully discrete case. *J Chem Phys* 2021;**155**(18):184101.
35. Seifert, NA, Prozument, K, Davis, MJ. Computational optimal transport for molecular spectra: the semi-discrete case. *J Chem Phys* 2022;**156**(13):134117.
36. Lange, E, Tautenhahn, R, Neumann, S, *et al.* Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics* 2008;**9**:375.
37. Zohora, FT, Rahman, MZ, Tran, NH, *et al.* DeepIso: a deep learning model for peptide feature detection from LC-MS map. *Sci Rep* 2019;**9**:17168.
38. Frank, AM, Monroe, ME, Shah, AR, *et al.* Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nat Methods* 2011;**8**(7):587–91.
39. Kantorovich, LV. Mathematical methods of organizing and planning production. *Management Sci* 1960;**6**(4):366–422.
40. Chizat, L, Peyré, G, Schmitzer, B, *et al.* Scaling algorithms for unbalanced optimal transport problems. *Math Comput* 2018;**87**(314):2563–609.
41. Knopp, P, Sinkhorn, R. Concerning nonnegative matrices and doubly stochastic matrices. *Pac J Math* 1967;**21**(2):343–8.
42. Sculley, D. Web-scale k-means clustering. In: M Rappa, P Jones, J Freire, S Chakrabarti, editors. *Proceedings of the 19th international conference on World Wide Web—WWW '10 New York*. New York: ACM Press; 2010. p. 1177–8.
43. Király, Z, Kovács, P. Efficient implementations of minimum-cost flow algorithms. *Acta Univ Sapientiae Inform* 2012;**4**(1):67–118.
44. Alignstein. Version: 1.0. https://github.com/grzsko/Alignstein. Accessed 8 June 2022.
45. MassSinkhornmetry. https://github.com/grzsko/MassSinkhornmetry. Accessed 8 June 2022.
46. Pedregosa, F, Varoquaux, G, Gramfort, A, *et al.* Scikit-learn: machine learning in Python. *J Machine Learn Res* 2011;**12**:2825–30.
47. scikit-learn, Version: 1.1.1. https://scikit-learn.org/stable/index.html. Accessed 8 June 2022.
48. Hagberg, AA, Schult, DA, Swart, PJ. Exploring network structure, dynamics, and function using NetworkX. In: G Varoquaux, T Vaught, J Millman, editors. *Proceedings of the 7th Python in Science Conference*. Pasadena, CA: 2008. p. 11–5. https://conference.scipy.org/proceedings/SciPy2008/SciPy2008_proceedings.pdf.

49. NetworkX, Version: 2.8.3. https://networkx.org/. Accessed 8 June 2022.
50. Bellew, M, Coram, M, Fitzgibbon, M, *et al.* A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics* 2006;**22**(15):1902–9.
51. Katajamaa, M, Orešič, M. Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics* 2005;**6**:179.
52. Li, Xj, Eugene, CY, Kemp, CJ, *et al.* A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry∗ S. *Mol Cell Proteomics* 2005;**4**(9):1328–40.
53. Zhang, X, Asara, JM, Adamec, J, *et al.* Data pre-processing in liquid chromatography–mass spectrometry-based proteomics. *Bioinformatics* 2005;**21**(21):4054–9.
54. Smith, CA, Want, EJ, O'Maille, G, *et al.* XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem* 2006;**78**(3):779–87.
55. Sequiera, GL, Sareen, N, Sharma, V, *et al.* High throughput screening reveals no significant changes in protein synthesis, processing, and degradation machinery during passaging of mesenchymal stem cells. *Can J Physiol Pharmacol* 2019;**97**(6): 536–43.
56. Vizcaíno, JA, Deutsch, EW, Wang, R, *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* 2014;**32**(3):223–6.
57. Perez-Riverol, Y, Csordas, A, Bai, J, *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* 2018;**47**(D1):D442–50.
58. Eng, JK, Jahan, TA, Hoopmann, MR. Comet: an open-source MS/MS sequence database search tool. *Proteomics* 2012;**13**(1): 22–4.
59. Eng, JK, Hoopmann, MR, Jahan, TA, *et al.* A deeper look into comet—implementation and features. *J Am Soc Mass Spectrometry* 2015;**26**(11):1865–74.
60. The Global Proteome Machine. https://www.thegpm.org/. Accessed 28 September 2022.
61. IPB. Critical assessment of alignment procedures data and evaluation scripts. https://msbi.ipb-halle.de/msbi/caap. Accessed 28 September 2022.
62. Open Proteomics Database. http://data.marcottelab.org/MSdata/OPD/. Accessed 28 September 2022.
63. Skoraczyński, G, Gambin, A, Miasojedow, B. Supporting data for "Alignstein: optimal transport for improved LC-MS retention time alignment." *GigaScience Database*. 2022. http://dx.doi.org/10.5524/102267.