International Journal of Machine Learning and Computing, Vol. 11, No. 6, November 2021

# A Bi-directional Hierarchical Clustering (BHC) for Peak Matching of Large Mass Spectrometry Data Sets

Nazanin Zounemat Kermani, Xian Yang, Yike Guo, James McKenzie, and Zoltan Takats

***Abstract*—the preprocessing of mass spectrometry (MS) data is a crucial step in every MS study, which not only makes data comparable and manageable but also makes the study more reproducible. However, an essential part of this process, which is often overlooked, is peak matching. Although existing clustering methods have been applied for peak matching, the use of these methods have been limited. For example, the use of hierarchical agglomerative clustering (HAC) for matching of mass/charge signals has been constrained to small-scale MS data sets due to the computational complexity of HAC. In this paper, we reintroduce a bi-directional hierarchical agglomerative clustering (BHC) as a scalable and accurate peak matching technique. As a result, the computational complexity of hierarchical agglomerative clustering for peak matching was optimized by BHC to O(RlogR). BHC was benchmarked against existing peak matching techniques. Finally, we propose a parallelization framework that significantly reduces the peak matching method's computation time.**

***Index Terms*—Mass spectrometry data preprocessing, peak matching, hierarchical agglomerative clustering, parallel computing.**

## I. INTRODUCTION

Mass spectrometry (MS) is a technique that is used for measuring mass-to-charge (m/z) ratios of the compounds when a sample is ionized. It enables high throughput and fast analysis of chemical compositions of biological samples.

Frequently used ionization techniques are Secondary-ion mass spectrometry (SIMS), matrix-assisted laser desorption ionization (MALDI), and desorption electrospray ionization (DESI). The output from a mass spectrometer is a mass spectrum; a mass spectrum consists of a series of paired m/z ratios and intensity values. Mass spectrometry imaging (MSI) considered a significant improvement in MS technology; it allows us to visualize the spatial distribution of molecules [1]. This improvement led to the generation of three dimensional (3D) mass spectrometry imaging data sets. A tissue 3D MSI data sets consist of MSI images of successive tissue sections.

Oetjen *et al*. [2] provided a few benchmark MSI data sets. For example, a 3D MALDI imaging MS dataset of a mouse pancreas consists of 497,225 spectra with 13,312 data points per spectrum. Due to large datasets produced by MS technologies, efficient computational tools are necessary for preprocessing and statistical analysis of such high dimensional datasets [3]. MS research and applications have gained substantial momentum in the recent years [4]-[9], which has made it a priority to develop high-performance computer processing to process such large, multidimensional datasets [10] efficiently.

Preprocessing of MS data consists of multiple steps; typical processing steps involve data transformation, baseline correction, intensity smoothing, spectral alignment, peak picking, normalization, and peak matching. Peak matching is required to account for slight measurement uncertainty as peaks (signals) representing the same compound are not necessarily observed with the same m/z ratio. Which is determined by the instrument mass accuracy that represents the difference between measured and actual mass. It is measured in parts per million (ppm) (e.g., actual mass = 885.554, observed mass = 885.5549, mass error $=\frac{885.5549-885.554}{885.554} \times 10^6 \simeq$ 1.016ppm). Apart from the mass accuracy of the instrument, other reasons such as variability in processing centres and mass shift over a period of time are perceived to cause m/z deviations [11].

Spectral alignment and peak matching techniques sought to eliminate such variations. Techniques such as dynamic time warping (DTW) and derivative dynamic time warping (DDTW) were applied to find optimal alignment between spectra [12], [13]. However, while accurate peak alignment techniques minimize the distance between m/z ratios, even after alignment, there remain variations in m/z ratios due to the differences in peak shapes [14](Fig. 1 b-c).

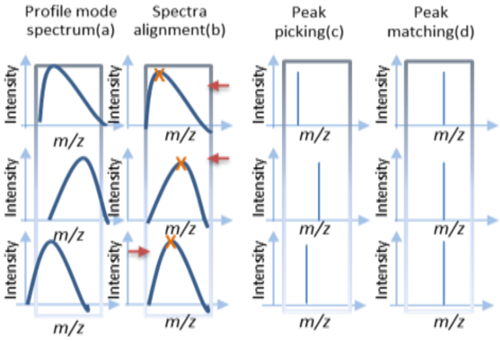

Fig. 1. Spectral alignment, peak picking, and peak matching- (a) show segments from three spectra. Each segment represents an ion measured in profile mode. (b) represents the same segments after spectra alignment. Red arrows show the direction of shift for each segment. (c) shows the same segments after peak picking. (d) Shows $m/z$ ratios after peak matching.

The ideal peak matching technique has to group all the relevant peaks (peaks that represent an identical ion) in one bin. It has to put all the non-related peaks (peaks that represent different structures) in separate bins. In this context, the bin is a term used to refer to a *m/z* range by a number e.g., mass range 200 to 200.001 can be referred to as

---

Manuscript received July 6, 2020; revised March 11, 2021. This work was supported in part by Wellcome Trust under WT ISSF funding.

Nazanin Zounemat Kermani, Xian Yang, Yike Guo are with the Department of Computing, Data Science Institute, Imperial College London.

James McKenzie and Zoltan Takats are with the Faculty of Medicine, Department of Metabolism, Digestion and Reproduction, Imperial College London (e-mail: n.kermani@imperial.ac.uk).

doi: 10.18178/ijmlc.2021.11.6.1064

bin number one. The bin can be specified either as a fixed value or relative to the value of m/z ratios.

Tibshirani *et al*. [15] used hierarchical agglomerative clustering for peak matching of MS proteomics data. However, with the proliferation of MS imaging technologies, the size of MS data sets has been increasing. For example, 3D MALDI imaging MS dataset from a mouse kidney comprised of 1,362,830 spectra, each containing 7,680 pairs of m/z ratios and intensities [2]. Another example is the 3D MALDI imaging MS dataset of a human oral squamous cell Carcinoma consists of 828,558 spectra with 7,680 data points (pairs of *m*/*z* ratios and intensities) per spectrum.

The application of hierarchical agglomerative clustering (HAC) for peak matching of large MSI data sets has been constrained because of the intensive computational and memory requirement of HAC algorithms. This paper aims to introduce BHC, a scalable version of HAC, for peak matching. Besides, we proposed workflow for the parallelization of peak matching methods, which was applied to peak matching function from the MALDIquant R package [16] and peak matching technique by Yasui *et al.* [17]. We showed that the parallelization framework improves the computational time of these algorithms significantly. We benchmarked BHC against these methods. BHC performance on synthetic data is significantly better than the Yasui *et al.* [17] and is on par with MALDIquant R package [16] method.

## II. HIERARCHICAL AGGLOMERATIVE CLUSTERING FOR PEAK MATCHING

Computational and memory/space complexity are indicators of the scalability of a computer algorithm; both are represented by O. Tibshirani *et al.* [15] applied hierarchical agglomerative clustering (HAC) with the complete linkage function for matching low-resolution MS data. In agglomerative (bottom-up) hierarchical clustering, each point assigned to a single cluster, at each iteration two individual closet clusters are merged into one cluster, the distance between individual clusters is defined by a distance metric such as complete linkage function:

$$d(u,v) = Max\{dist(u(i),v(j))\},\ i \in (1,...,n_u),\ j \in (1,...,n_v) \qquad (1)$$

where $n_u$, $n_v$ is the number of data points in cluster $u$ and $v$ respectively. The algorithm stops when all data points are assigned to a single cluster. The computational complexity of hierarchical clustering is affected by calculating pairwise similarity matrix (S) across data points $O(N^2)$ and the time to scan the similarity matrix to find the largest similarity hence $O(N^3)$ ($N$ is the number of data points). Kazmi *et al.* [18] applied HAC with a priority queue algorithm in a sequential manner. This reduces the time complexity of peak matching with HAC to $O(NR^2\log R)$, $R$ is the number of spectra and $R<<N$. While Kazmi *et al.*'s algorithm for peak matching is a considerable improvement, it is still unable to process large MS data sets; it is shown in Table I, the standalone Java program was acquired from Kazmi *et al.*; the program became unresponsive for $N$=10,000 spectra. The purpose of this study is to introduce a bi-directional HAC (BHC) that is a clustering-based technique stemmed from heuristic peak matching by Kazmi *et al.* [18]; it reduces the computational complexity of Kazmi *et al.* method from $O(NR^2\log R)$ to $O(NR\log R)$. This makes the method scalable so that it can be used for very large MS experiments, such as imaging data sets with tens of thousands of spectra.

## III. DATASETS

### A. Sample Preparation and Data Acquisition

The collection of the breast cancer data set (137 samples), as well as a colorectal cancer data set (28 samples), was approved by the institutional review board at Imperial College Healthcare National Health Service Trust (research ethics committee reference no REC 11/LO/0686 and 07/H0712/112 respectively). All tissue samples were stored at 80°C after collection. Each sample was cryosectioned to a thickness of 10 μm, mounted onto a glass slide, and stored at 80°C before DESI-MSI analysis. This analysis was carried out using a home-built DESI ion source coupled to an Exactive Fourier transform Orbitrap mass spectrometer (Thermo Scientific, Bremen, Germany). The mass analysis was carried out in negative ion mode over the mass range m/z 200–1000 at a nominal mass resolution of 100,000 and a mass accuracy of ± 2 ppm. The DESI−MSI was operated at a spatial resolution of 100 μm (injection time of 1000 ms) with a nitrogen pressure of 7.0 bar, sprayer voltage of 4.5 kV, capillary voltage 50 V, the capillary temperature of 250°C, tube lens voltage of 150 V, and skimmer voltage of 40 V and a solvent mixture of methanol and water in a signal 95:5 with a flow rate of 1.5 *μL/min*. Following DESI-MSI acquisition, tissue sections were stained with hematoxylin and eosin (HE). They underwent histological examination by a pathologist, where the different tissue types present in each sample was determined. Raw mass spectrometric images were converted to imzML format using imzML Converter (version 1.0.5) and imported into MATLAB (R2014a) for pre-processing and analysis.

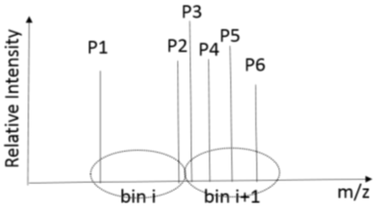

Fig. 2. Outliers in complete linkage HAC. The six peaks are shown $P_1...P_6$. HAC with complete linkage function creates the two clusters ($bin_i$,$bin_{i+1}$) shown as ellipses. The most intuitive two-cluster clustering is $\{\{P_1\},\{P_2,P_3,P_4,P_5,P_6\}\}$, but in complete-linkage HAC, the outlier $\{P_1\}$ splits $\{P_2,P_3,P_4,P_5,P_6\}\}$ into $\{P_1,P_2\},\{P_3,P_4,P_5,P_6\}\}$ [19].

### B. Synthetic Data

To generate m/z signals that resemble technical replicates from the ground truth spectrum, peaks on the spectrum were randomly shifted bidirectionally according to a variable ($S_{ppm}$) sampled from a truncated normal distribution ($\mu$ = 0, $\sigma^2$ = 2.5).

An MSI sample was chosen randomly from the breast cancer data. It consists of 4884 spectra. After peak matching using BHC ($E_{ppm}$ = 2.5), a mean spectrum was generated. This spectrum consists of 89,694 m/z signals. To study the practical scalability and performance of the methods, random subsets of this spectrum were generated and used as the ground truth for evaluation. [100, 1000, 5000, 7500, 10,000, 12500, 15,000, 20,000, 25,000, 50,000] number of *m/z* were randomly selected (reflects the number of m/z signals). For each of these spectra [50, 100, 200, 250, 500, 1000, 2000, 5000] (reflects the number of spectra) synthetic variations were generated.

## IV. METHODS

### A. Parallelization Framework

To speed up the peak matching methods, a parallelization framework was implemented. Intervals are generated based on the rationale that if there is a gap between *m/z* signals bigger than the instrument accuracy, this gap divides the super spectrum into two *m/z* intervals that have no overlapping regions before and after peak matching. The underlying idea of this framework is that the super spectrum can be partitioned into smaller intervals (i.e., *m/z* ranges), and each interval can be peak matched independently of others (Fig. 5). The peak matching is finished when all the peaks from all intervals were exhausted. The set of intervals ($S = \{s_1, s_2, ,...\}$) is a partition set of the super spectrum ($S$) such that:

1) *Intervals* are pairwise disjoint: Ê$S1, S2 \in S$: $S1 \cap S2 = I$ when $S1 \neq S2$
2) The union of S forms the whole set $S$: $\cup S = S$

None of the elements of $S$ is empty: Ê$T \in S$: $T \neq f$.

### B. A Bi-directional Hierarchical Agglomerate Clustering (BHC)

For the peak matching of mass spectrometry data, clustering has to be performed on one-dimensional data (only on *m/z* signals). Therefore computational complexity of HAC can be reduced from $O(NR^2\log R)$ to $O(NRlogR)$. BHC is described in Algorithm 1. Here, the super-spectrum undergoes a bi-directional HAC (BHC). One priority queue and a single passage processing of this priority queue are enough for clustering. At each iteration of the algorithm, a *m/z* signal (or cluster of *m/z* signals) can be merged with either of its immediate adjacent signals (one bigger or smaller hence bi-directional), such that only distances between adjacent *m/z* signals are calculated (hence only one priority queue instead on $N$ priority queues for HAC). This reduces the number of distances to be calculated from $R(R-1)/2$ to $R-1$.

In the pseudo-code for Algorithm 1 'min index' returns the index of the smallest element. 'concatenate' returns a concatenation of all its arguments. 'find(A,x)' returns the indices $i$ for which $A[i] == x$, 'replace(A,x,v)' replace all values in A equal to $x$ with $v$. 'sum(A)' sums over all elements in A. 'length(A)' returns the number of elements in A.

### C. Split Correction

HAC with complete linkage function is sensitive to outliers. Outlier data points are points that do not fit well into the global structure of the cluster. HAC, with complete linkage function, does not find the most intuitive cluster structure in these cases (Fig. 2).

We observed that in such cases, the centroids of two consecutive bins are very close, and the distribution of m/z ratios in the bins is skewed. Such splitting can be corrected by merging bins whose centroids are closer than the $E_{ppm}$. Equation 2 shows the formula to calculate the centroid of $bin_i$. where $x = \{m/z | m/z$ ▯ $bin_i\}$ and $f(x) = intensity(x)$.

$$centriod(bin_i) = \frac{\sum x f(x)}{\sum f(x)} \tag{2}$$

**A** DESI Mass Spectra data set **B** Super-Spectrum **C** Intervals

Fig. 3. From Mass spectra data set to the intervals. A shows three mass spectra. B illustrates the collated spectrum (super-spectrum). C shows the intervals for the super-spectrum.

### D. Validation of Scalability and Performance

To compare the scalability Kazmi *et al*. method with BHC, an MSI sample was chosen randomly from MSI data sets. This MSI sample consists of 4884 spectra; a mass spectrum was randomly selected from this MSI sample. From this mass spectrum, 100, 500, 1000, and 10,000 spectra were generated using the method explained in the synthetic data section. Then the peak matching was performed. To compare execution time with Kazmi *et al*., the output of 'cputime' function Matlab (R2015a) were averaged over multiple runs (5 runs).

The performance of methods was measured by precision, recall, and F1 score. For each bin, the precision and recall were calculated based on the following formulas.

$$Precision(bin_i|method_j) = \frac{|retrievedpeaks_{bin_i}^{method_j} \bigcap relevantpeaks_{bin_i}|}{|retrievedpeaks_{bin_i}^{method_j}|}$$

$$Recall(bin_i|method_j) = \frac{|retrievedpeaks_{bin_i}^{method_j} \bigcap relevantpeaks_{bin_i}|}{|relevantpeaks_{bin_i}|}$$

$$Precision(method_j) = \frac{1}{M_j} \Sigma_{i=1}^{i=M_j} Precision(bin_i|method_j)$$

$$Recall(method_j) = \frac{1}{M_j} \Sigma_{i=1}^{i=M_j} Recall(bin_i|method_j)$$

$$F1(method_j) = 2 * \frac{Precision(method_j) * Recall(method_j)}{Precision(method_j) + Recall(method_j)}$$

cancer datasets were calculated. Each MS imaging sample is a collection of many spectra (> 1000) acquired from the predefined areas of a tissue section.

The plot is generated by adding one MSI sample at a time and calculating the number of intervals cumulatively. This plot shows that by increasing the number of spectra to 250,000, the number of intervals increases and stabilizes at around 150,000 intervals with approximately 100,000 spectra. The average size of intervals is 0.5-1 percent of the size of the super spectrum. Therefore the parallelization framework introduced here would significantly speed up the peak matching methods for MS data. We observed that for the peak matching methods, the additional time for making the intervals and merging the intervals is negligible.

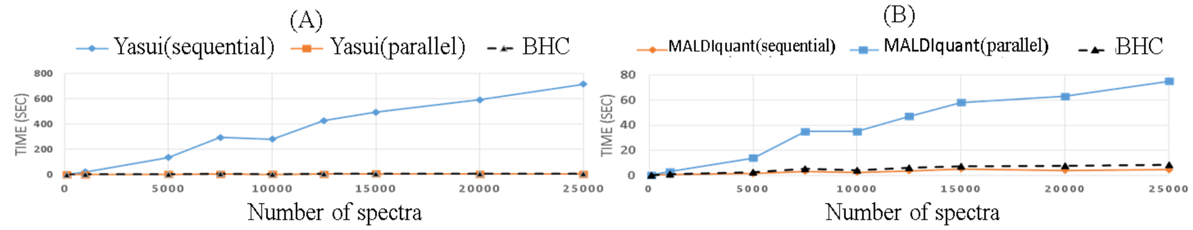

Fig. 6. Execution time as a function of a number of spectra. Y-axis represents the run-time in seconds and X-axis shows the number of spectral. The solid blue line represent the execution times without using the parallel framework, i.e., sequential. The orange line shows the run-time after utilizing the parallel framework. The dashed black line shows the BHC's execution time.

### D. *Parallelization Framework*

Fig. 6 (orange lines) shows that the peak matching routines can be executed substantially faster in the parallel framework (blue lines versus orange lines). Run-time of the BHC (black dashed lines) is on par with the other methods in parallel mode.

### E. *Split Correction*

MS image visualizes the spatial intensity distribution of a particular m/z signal. Phosphatidylinositol is an abundant lipid in tissues that could be captured by DESI-MSI. Here a case study was described that shows how a group of *m/z* signals that correspond to the same compound split into 2 clusters (bins). Fig. 7 left panel shows an MS image for *m/z* = 885.55 [Phosphatidylinositol-H]. The middle and the Right panels show the MS images (first row) and distribution of *m/z* signals (second row). The left panel displays the ion image (first row) and the distribution of *m/z* signals (second row) after split correction. From the complementary MS images, it could be observed that peaks separated into two bins represents one compound (left panel). Fig. 7 illustrates that in such cases, the centroids of two split bins are very close, and the *m/z* distributions in each bin are highly skewed.

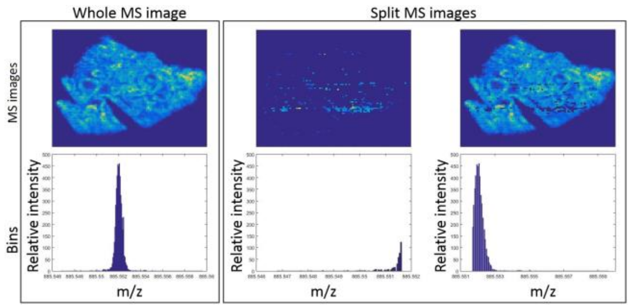

Fig. 7. Split correction- The plots on the top row illustrate MS images; the first plot shows a complete MS image for m/z = 885.55, and the second and third plots show the split MS images for the same signal. The bottom row represents the m/z distributions of signals corresponding to the MS images. If the split correction option was enabled, then it was performed for each interval after peak matching.

Such splitting can be removed by merging bins that their centroids are closer than the window of potential shift. This is optional, and the user can choose whether to perform split correction or not.

### F. *Performance Evaluation*

Table II shows the precision, recall, and F1 percentages for all techniques. Yasui *et al.* technique have a lower precision compared to other techniques, while BHC and MALDIquant have close to 100 percent precision. This means that bins generated by BHC and MALDIquant are collections of relevant peaks. BHC consistently outperforms other methods' recall and F1 score (P value<0.001). The second and third columns of Table II showed a slight improvement of the recall and F1 score when a split correction was performed on top of BHC; however, the improvement is not statistically significant (Table II column 6). Despite this improvement, more precise split correction methods than the one that is explored here can be devised by defining the bin boundaries more precisely.

TABLE II: PERFORMANCE COMPARISON OF PEAK MATCHING TECHNIQUES - PEAK MATCHING TECHNIQUES AND THEIR PAIRWISE COMPARISON ARE SHOWN. PRECISION, RECALL, AND F1 SCORE WAS MEASURED OVER 10 SPECTRA WITH 5000 VARIATIONS FOR EACH. MEAN (STANDARD DEVIATIONS) ARE SHOWN FOR EACH METHOD; ZERO SIGNIFIES A VALUE SMALLER THAN 0.00001. THE P-VALUES GENERATED BY THE KRUSKAL-WALIS RANK-SUM TEST

| ***Methods*** | ***Precision*** | ***Recall*** | ***F1 score*** |
|---|---|---|---|
| **BHC** | *99.99(±0)* | *99.76(±0)* | *99.87(±0)* |
| **BHC NC** | 99.99(±0) | 99.72(±0) | 99.86(±0.05) |
| **MALDIquant** | 99.99(±0) | 96.59(±0.001) | 98.26(±0.1) |
| **Yasui** | 77.94(±0.005) | 77.85(±0.005) | 77.89(±3.5) |
| **Pairwise comparison** | **Group Comparison(p-value)** | | |
| **BHC vs. BHC NC** | 0.2 | 0.1 | 0.1 |
| **BHC vs. MALDIquant** | 0.3 | <.001 | <.001 |
| **BHC vs. Yasui** | <.001 | <.001 | <.001 |
| **BHC NC vs. MALDIquant** | 0.9 | <.001 | <.001 |
| **BHC NC vs. Yasui** | <.001 | <.001 | <.001 |
| **MALDIquant vs. Yasui** | <.001 | <.001 | <.001 |

## VI. DISCUSSIONS

We have discussed an agglomerative hierarchical clustering for peak matching of large MS datasets. BHC is an agglomerative hierarchical clustering method that works for one-dimensional data. In addition to BHC, a framework is introduced that can be used for the parallelization of peak matching methods. However, the improvement in the speed and scalability is straightforward to validate, but the performance of the methods is harder to compare. F1 score, precision, and recall were used on the synthetic data to compare the methods' performance. None of the methods were perfect. We observed that the MALDIquant method tends to split bins, and the Yasui method tends to merge bins. BHC is sensitive to noise; however, a split correction method is proposed and tested, but more concise methods would be beneficial. There have been a plethora of methods designed for spectra alignment of liquid chromatography-mass spectrometry (LC-MS) [20], [21] and nuclear magnetic resonance (NMR) [22]; some these methods can be used for MS peak matching. However, most of these methods rely upon a reference spectrum to align the spectra.

In summary, our approach provides an efficient framework on which peak matching with hierarchical agglomerative clustering can be scaled to 100,000 of mass spectra with millions of *m*/*z* signals. Future work is required to validate the accuracy of the method on larger mass ranges and other MS technologies; we demonstrated the performance of BHC for DESI-MS on mass range <1000.

The parallelization framework showed a high potential for speeding up the peak matching techniques. BHC has made the peak matching with HAC scalable to large MS data sets. Correction of split bins due to the sensitivity if the HAC to outliers improves the performance of HAC for peak matching. BHC performance on the synthetic data is on par or better than the peak matching methods devised by Yasui *et al.* [11] and Gibb *et al.* [17]. However, finding the best peak matching technique remains a challenge as each method has its idiosyncrasies.

MS imaging technologies advanced to generate three-dimensional views of tumors and tissues by stacking MS images of successive tissue sections, hence larger MSI data sets. These data sets and large cohort studies can benefit from BHC. BHC peak matching method is tested for small MS range (<1000 Da), but the application of the method to the larger mass range is considered as future research. As the result of the parallelization framework, the time required for peak matching techniques is significantly reduced, this allows not only to run multiple peak matching instances simultaneously but also allows us to run multiple techniques quickly. An ensemble method that aggregates the results of various peak matching techniques could be considered as future research.

The instrument mass accuracy is reflected in the data. However, the exact precision of the recorded data drifts over time (declines), a further study should include an automated technique to adjust for this. Another potential that can be utilized for MS imaging datasets is to incorporate spatial information for improved peak matching. Finally, a speed-up may come from the implementation of the methods on graphical processing units.

## VII. Conclusions

BHC peak matching algorithm allows the application of peak matching by agglomerate hierarchical clustering on large mass spectrometry datasets. The benchmarking against two other peak matching methods proved that the performance of the BHC is either on par with other methods or is better. The parallel framework introduced here significantly improves the current peak-matching methods runtime. These improvements potentially contribute positively to a more accurate and efficient preprocessing pipeline for MSI data.

## Conflict of Interest

The authors declare no conflict of interest.

## Author Contributions

NZK, conducted the research. JMK and XY helped with the writing the paper. All authors had approved the final version.

## Acknowledgment

Authors wish to acknowledge Anna K Mroz and Nicole Strittmatter for colorectal MSI, and Sabine Guenther for breast MSI data acquisition and Luisa Doria for providing information about the data acquisition platform. The authors would like to thank Adrian Mallinue and Timothy Ebbels for constructive comments on the pseudo-code and algorithm.

## References

[1] J. D. Watrous, T. Alexandrov, and P. C. Dorrestein, "The evolving field of imaging mass spectrometry and its impact on future biological research," *Journal of Mass Spectrometry*, vol. 46, no. 2, pp. 209-222, Feb. 2011.

[2] J. Oetjen, K. Veselkov, J. Watrous, J. S. McKenzie *et al.*, "Benchmark datasets for 3D MALDI-and DESI-imaging mass spectrometry," *GigaScience*, vol. 4, no. 1, p. 20, Dec. 2015.

[3] D. Trede, J. H. Kobarg, J. Oetjen, H. Thiele, P. Maass, and T. Alexandrov, "On the importance of mathematical methods for the analysis of MALDI-imaging mass spectrometry data," *Journal of Integrative Bioinformatics*, vol. 9, no. 1, p. 1, Mar. 1, 2012.

[4] W. M. Abdelmoula, B. Balluff, S. Englert *et al.*, "Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data," *Proceedings of the National Academy of Sciences*, vol. 113, no. 43, pp. 12244-12249, Oct. 25, 2016.

[5] T. P. Siegel, G. Hamm, J. Bunch, J. Cappell, J. S. Fletcher, and K. Schwamborn, "Mass spectrometry imaging and integration with other imaging modalities for greater molecular understanding of biological tissues," *Molecular Imaging and Biology*, vol. 20, no. 6, pp. 888-901, 2018.

[6] L. S. Eberlin, I. Norton, A. L. Dill, A. J. Golby, K. L. Ligon, S. Santagata, R. G. Cooks, and N. Y. Agar, "Classifying human brain tumors by lipid imaging with mass spectrometry," *Cancer Research*, vol. 72, no. 3, pp. 645-654, 2012.

[7] S. Rauser, C. Marquardt, B. Balluff *et al*., "Classification of HER2 receptor status in breast cancer tissues by MALDI imaging mass spectrometry," *Journal of Proteome Research*, vol. 9, no. 4, pp. 1854-1863, 2010.

[8] A. R. Clark, D. Calligaris, M. S. Regan *et al*., "Rapid discrimination of pediatric brain tumors by mass spectrometry imaging," *Journal of Neuro-oncology*, vol. 140, no. 2, pp. 269-279, 2018.

[9] K. Margulis, A. S. Chiou, S. Z. Aasi, R. J. Tibshirani, J. Y. Tang, and R. N. Zare, "Distinguishing malignant from benign microscopic skin lesions using desorption electrospray ionization mass spectrometry imaging," *Proceedings of the National Academy of Sciences*, vol. 115, no. 25, pp. 6347-6352, 2018.

[10] E. A. Jones, R. J. van Zeijl, P. E. Andrén, A. M. Deelder, L. Wolters, and L. A. McDonnell, "High speed data processing for imaging MS-based molecular histology using graphical processing units," *Journal of the American Society for Mass Spectrometry*, vol. 23, no. 4, pp. 745-752, 2012.

[11] K. R. Coombes, J. S. Morris, J. Hu, S. R. Edmonson, and K. A. Baggerly, "Serum proteomics profiling — a young technology begins to mature," *Nature Biotechnology*, vol. 23, no. 3, p. 291, 2005.

[12] Q. P. He, J. Wang, J. A. Mobley, J. Richman, and W. E. Grizzle, "Self-calibrated warping for mass spectra alignment," *Cancer Informatics*, vol. 10, CIN-S6358, 2011.
[13] G. Tomasi, F. V. D. Berg, and C. Andersson, "Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data," *Journal of Chemometrics*, vol. 18, no. 5, pp. 231-241, 2004.
[14] J. W. Wong, C. Durante, and H. M. Cartwright, "Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets," *Analytical Chemistry*, vol. 77, no. 17, pp. 5655-5661, 2005.
[15] R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, and Q. T. Le, "Sample classification from protein mass spectrometry, by 'peak probability contrasts," *Bioinformatics*, vol. 20, no. 17, pp. 3034-3044, 2004.
[16] S. Gibb and K. Strimmer, "MALDIquant: A versatile R package for the analysis of mass spectrometry data," *Bioinformatics*, vol. 28, no. 17, pp. 2270-2271, 2012.
[17] Y. Yasui, D. McLerran, B. L. Adam, M. Winget, M. Thornquist, and Z. Feng, "An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers," *BioMed Research International*, no. 4, pp. 242-248, 2003.
[18] S. A. Kazmi, S. Ghosh, D. G. Shin, D. W. Hill, and D. F. Grant, "Alignment of high resolution mass spectra: development of a heuristic approach for metabolomics," *Metabolomics*, vol. 2, no. 2, pp. 75-83, 2006.
[19] D. Kornack and P. Rakic, "Cell proliferation without neurogenesis in adult primate neocortex," *Science*, vol. 294, pp. 2127-2130, 2001.
[20] Y. Xian, "Analysing datafied life," PhD thesis, Imperial College London, 2015.
[21] E. Lange, R. Tautenhahn, S. Neumann, and C. Gröpl, "Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements," *BMC Bioinformatics*, vol. 9, no. 1, p. 375, 2008.
[22] T. N. Vu and K. Laukens, "Getting your peaks in line: A review of alignment methods for NMR spectral data," *Metabolites*, vol. 3, no. 2, pp. 259-276, 2013.

**Nazanin Zounemat Kermani** is a research associate. She is now working at the Data Science Institute, Imperial College London. Her research interests are machine learning in healthcare and genomics.

**Xian Yang** is an assistant professor at the Department of Computer Science, Hong Kong Baptist University. She worked as a research associate in the Data Science Institute at Imperial College London from 2012 to 2018. Her research interests are artificial intelligence in healthcare, modern medicine and cloud computing.

**Yike Guo** is a professor of computing science in the Department of Computing at Imperial College London. He is the founding director of the Data Science Institute at Imperial College. Professor Guo has published over 250 articles, papers and reports.

**James McKenzie** is a research associate at the Faculty of Medicine, Department of Metabolism, Digestion and Reproduction, Imperial College London. His research interest is the development of analytical and machine learning solution for mass spectrometry data.

**Zoltan Takats** is a professor of Analytical Chemistry at the Faculty of Medicine, Department of Metabolism, Digestion and Reproduction, Imperial College London. He is one of the founders of the field of 'Ambient Mass Spectrometry'. He is the author of more than 78 publications.