RESEARCH ARTICLE

Comparison of different machine learning methods and dimensionality reduction for classification astrocytoma and glioblastoma tissues by mass spectra [version 1; peer review: 1 approved with reservations, 1 not approved]

Evgeny S. Zhvansky¹, Anatoly A. Sorokin¹, Denis S. Zavorotnyuk¹, Vsevolod A. Shurkhay^{1,2}, Vasiliy A. Eliferov¹, Denis S. Bormotov¹, Daniil G. Ivanov^{1,3}, Alexander A. Potapov²

¹Moscow Institute of Physics and Technology, Dolgoprudnyi, 141700, Russian Federation

²Federal State Autonomous Institution «N.N. Burdenko National Scientific and Practical Center for Neurosurgery» of the Ministry of Healthcare of the Russian Federation, Moscow, 125047, Russian Federation

³Emanuel Institute for Biochemical Physics of the Russian Academy of Sciences, Moscow, 119334, Russian Federation

V1 First published: 21 Jan 2021, 10:39
<https://doi.org/10.12688/f1000research.28288.1>
 Latest published: 21 Jan 2021, 10:39
<https://doi.org/10.12688/f1000research.28288.1>

Abstract

Background: Recently developed methods of ambient ionization allow rapid obtaining of large mass spectrometric datasets, which have a great application in biological and medical analysis. One of the areas that could employ such analysis is neurosurgery. The fast *in situ* identification of dissected tissues could assist the neurosurgery procedure. The additional information about tumor could help the tumor border monitoring. In this paper, tumor tissues of astrocytoma and glioblastoma are compared, as their identifications during surgery could influence the extent of resection and, hence, the median and overall survival.

Methods: Mass spectrometric profiles of brain tumor tissues contain molecular information, which is rather hard to interpret in terms of identifications of individual molecules. The machine learning algorithms are employed for the fast automated mass spectra classification. Different algorithms of dimensionality reduction are considered to process the mass spectra before the classification task, as the initial dimensionality of mass spectra is too high compared with the number of mass spectra.

Results: Different classifiers are compared for both just preprocessed data and after dimensionality reduction. The Non-Negative Matrix Factorization appears to be the most effective dimensionality reduction algorithm. The random forest algorithm demonstrated the most robust appearance on the tested data. Also, the comparison of the accuracy of the trained classifier on the mass spectra of tissues

Open Peer Review

Approval Status

	1	2
version 1		
21 Jan 2021	view	view

1. **Sergei A. Moshkovskii** , Pirogov Russian National Research Medical University, Moscow, Russian Federation
2. **Konstantin Chingin**, Jiangxi Key Laboratory for Mass Spectrometry and Instrumentation, East China University of Technology, Nanchang, China

Any reports and responses or comments on the article can be found at the end of the article.

measured with different instruments and different resolution is provided in the paper.

Conclusions: Machine learning classifiers overfit the raw mass spectrometric data. The dimensionality reduction allows the classification of both train and test data with 88% accuracy. Positive mode data provides better accuracy. A combination of principal component analysis and AdaBoost algorithms appears to be most robust to changing the instrument and conditions.

Keywords

mass spectra, astrocytoma and glioblastoma tumors, dimensionality reduction, classification, high- and low-resolution

This article is included in the **Oncology** gateway.

This article is included in the **Artificial Intelligence and Machine Learning** gateway.

Corresponding author: Evgeny S. Zhvansky (evgeny.zhvansky@yandex.ru)

Author roles: Zhvansky ES: Conceptualization, Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Sorokin AA: Conceptualization, Methodology, Project Administration, Supervision, Validation, Writing – Review & Editing; Zavorotnyuk DS: Data Curation, Investigation, Validation; Shurkhay VA: Investigation, Resources; Eliferov VA: Investigation; Bormotov DS: Investigation; Ivanov DG: Investigation; Potapov AA: Funding Acquisition, Project Administration, Resources, Supervision

Competing interests: No competing interests were disclosed.

Grant information: The research was supported by the Ministry of Science and Higher Education of the Russian Federation (agreement # 075-00337-20-02, project # 0714-2020-0006).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Zhvansky ES *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Zhvansky ES, Sorokin AA, Zavorotnyuk DS *et al.* **Comparison of different machine learning methods and dimensionality reduction for classification astrocytoma and glioblastoma tissues by mass spectra [version 1; peer review: 1 approved with reservations, 1 not approved]** F1000Research 2021, 10:39 <https://doi.org/10.12688/f1000research.28288.1>

First published: 21 Jan 2021, 10:39 <https://doi.org/10.12688/f1000research.28288.1>

Introduction

The extent of tumor resection is important for patients with primary brain tumors in terms of life expectancy since tumor cells can provoke a disease recurrence¹. Recently, we have seen a growing interest in the use of mass spectrometry for the identification of tumor tissues, typing, and detection of tumor boundaries²⁻⁴. Analysis of tumor samples using mass spectrometry is based on the observation that tumor cells differ significantly from normal ones in their metabolic processes and, as a consequence, have a different chemical composition⁵⁻⁸. Identification of the histological type and location of the brain tumor tissue during neurosurgical intervention allows for the correct dissection of the tumor and opens the way to a personalized strategy for further treatment of the patient with chemotherapy, taking into account the molecular features of the tumor. Comparative analysis of tumor types is fundamental, although elucidating tumor boundaries is of the highest priority for neurosurgeons⁹.

Fast mass spectrometric profiling allows rapid clinic and laboratory analyses but faces the problem of classification of high-dimensionality objects.

For the analysis of large mass spectrometry (MS) data, dimension reduction (DR) algorithms are commonly used as a previous step for statistical analysis and visualization. The most widely used DR methods are linear methods such as PLS-DA and PCA¹⁰⁻¹⁶. More advanced nonlinear methods have recently been developed, such as t-SNE and UMAP, to name a few¹⁷. DR methods allow visualization through the major components of the compressed data, e.g. the first three PCA components and three selected ions are the most commonly used characteristics in MS imaging^{18,19}.

Despite the visualization that could produce the results dividing the different types of tissues by their MS, these results are subjective and require further automated classification. Machine learning (ML) methods are commonly used for these purposes²⁰⁻²².

In this paper, we compare the performance of six DR algorithms and ten ML algorithms in their ability to classify the MS profiles of astrocytoma and glioblastoma. Also, we compared the stability of the trained ML models on the data obtained with another instrument and under different conditions. The stability of the trained models to different instruments is very important for the wide spreading of the methods, as the clinical conditions could influence the mass spectra due to different mass analyzers. Different polarities, resolutions, and m/z ranges are also considered.

Methods

Measurements

The two instruments used in our study: for both high and low resolution (HR and LR) under laboratory conditions (Thermo LTQ Orbitrap XL) and only low resolution under clinical conditions (Thermo LTQ XL). Inline cartridge extraction²³ followed by electrospray ionization was used for mass spectrometric profiling of samples. Spectra from Thermo LTQ Orbitrap XL were measured in both positive and negative ion modes with two resolutions: 24,000 at m/z=744 (high resolution using Orbitrap analyzer) and 2000 at m/z=744 (low resolution using LTQ analyzer). All spectra were measured in the m/z 500 — 1000 and m/z 100 — 2000 ranges.

Samples

Tissue samples were provided by the N.N. Burdenko NSPCN and analyzed under a protocol approved by N.N. Burdenko NSPCN Institutional Review Board (order 40 from 12.04.2016, revised with order 131 from 17.07.2018). A signed informed consent form, filled out in accordance with the requirements of the local ethical committee, specifically noting that all removed tissues can be used for further research, was obtained from all patients before surgery. The study was conducted in accordance with the Helsinki Declaration as revised in 2013. All procedures were carried out according to the relevant guidelines and regulations.

Three fragments of tissue taken from a single patient were measured with each instrument to take into account and evaluate inner biological variability. In total, 76 astrocytoma fragments (26 patients) and 89 glioblastomas (31 patients) were measured with Thermo LTQ XL (Ltq) in LR, while 60 astrocytoma fragments (20 patients), 84 glioblastoma fragments (28 patients) were measured with Thermo LTQ Orbitrap XL (Orbitrap) in LR and HR. Samples from 37 patients were measured with both instruments, 20 patients were measured only with Ltq, 11 — only with Orbitrap. The precise schema is given in [Figure 1](#).

There tissue samples (total 309) from 68 patients had different brain tumors: 21 anaplastic astrocytomas (WHO Grade III; 9 tumors with IDH-1 R132H mutation), 10 diffuse astrocytomas (WHO Grade II; 7 tumors with IDH-1 R132H mutation), 1 gemistocytic astrocytoma WHO Grade II (IDH-1 with IDH-1 R132H mutation) and 36 glioblastomas (WHO Grade IV; 11 tumors with IDH-1 R132H mutation).

Figure 1. Venn diagrams of astrocytoma and glioblastoma tissues measured with Lttq/Orbitrap (under clinical/laboratory conditions respectively).

Tissue samples were divided into several parts, for annotation and MS analysis. Tissue samples were annotated with routine hematoxylin and eosin staining and further immunohistochemical analysis of its fragment (expression of isocitrate dehydrogenase 1 (IDH-1)). Other fragments of tissue samples were measured with Thermo LTQ XL right after removal or were placed in the normal saline, frozen, and stored at -80°C until measurement with Thermo LTQ Orbitrap XL. Measurement for each sample was carried out continuously with alternating mode, range, and resolution if possible²⁴.

Processing

Spectra were processed with the algorithm similar to described previously^{25,26}. Mass spectra were binned with binning width 0.01 m/z, and then spectra were convoluted with Gaussian (FWHM equals 0.4m/z for high-resolution and 0.2m/z for low-resolution). Spectra of each measurement were filtered by a moving median filter. The width and step of the median filter were chosen to 51. The baseline subtraction was carried out through Kneen and Annegarn's algorithm²⁷. Spectra similarity matrix (SSM) with cosine measure similarity was calculated as described previously²⁶. Both dimensionality reduction and classification were made with Scikit-learn v0.23.1²⁸ machine learning library.

All calculations and visualizations were made using code written by the authors using MATLAB R2019b (GNU Octave 5.2.0 could be used for reproducing results) or Python 3.7. This code is freely available at <http://doi.org/10.5281/zenodo.4307700>⁴³.

Dimensionality reduction

The classification without dimensionality reduction is compared with the method of principal component analysis (PCA), non-negative matrix factorization (NNMF)²⁹, isometric mapping (Isomap)³⁰, Partial least squares discriminant analysis (PLS-DA)³¹, UMAP³², Diffusion map³³. The number of the selected features after dimensionality reduction equals 5.

Classification

Nearest Neighbors³⁴, linear support vector machine SVM³⁵, Radial-basis function (RBF) kernel SVM³⁵, Gaussian Process³⁶, Decision Tree³⁷, Random Forest³⁸, Neural Net (Multi-layer Perceptron classifier with 1 hidden layer with 100 neurons and the log-loss function and "adam" optimizer and L2 regularization), AdaBoost³⁹, Naive Bayes⁴⁰, QDA⁴¹ were used as classifiers. Only one scan from each measurement after median filtration was taken part in the classification task.

Results

SSMs of astrocytoma and glioblastoma in different polarities are presented in Figure 2. There are two clusters of fresh and frozen samples. The first were measured right after surgery, the second after storing samples under -80°C . Figure 2 represents the large-range data (100-2000 m/z) measured in low-resolution.

Frozen glioblastoma MS profiles have a lot of outliers, which is obvious from the SSM Figure 2B. The outliers could be filtered out from the datasets for the classification task, but the classifiers will be created with only fresh samples. Thus, outliers will be tested on the trained classifiers and the results of visual inspection of SSM could be confirmed. The negative mode (Figure 2A) has more variance and spectra of the negative mode have less similarity. At the same time, there are no obvious differences between astrocytoma and glioblastoma. We, therefore, employ the machine learning methods for classifying the MS profiles.

Figure 2. Spectra similarity matrices of fresh and frozen samples. (A) Negative and (B) positive mode.

The accuracy of classifiers for different dimensionality reduction algorithms is presented in [Figure 3](#). The accuracy was calculated as the ratio of correct predictions to the total number of predictions on the given data and labels.

First 70% of astrocytomas and 70% of glioblastomas were taken as the training data, the rest as the test data. Neither of the mass spectra of the samples corresponding to the single patient was in training and test data simultaneously to prevent the overfitting. Thus 46% of training data and 46% of test data were astrocytoma class, whereas 54% of the training and test data were glioblastoma class. Many of the combinations of classifiers and dimensionality reduction algorithms score about 46% or 54% of accuracy.

The PLS-DA is overfitted in all cases except probably negative mode and m/z 500-1000 ([Figure 3E and F](#)). The highest values of accuracy are achieved with NNMf for positive mode MS profiles. NNMf produced better results for a wide m/z range (100-2000) since classifiers produce similar accuracy in this case. The best result of NNMf and random forest combination for the positive mode, m/z 500-1000 is achieved for 5 components of NNMf. The most stable results were produced for the 100-2000 m/z range in positive mode. These conditions are considered below

The robustness of each combination of DR and ML algorithms is tested on the data obtained with another instrument (Orbitrap) both in low and high resolution ([Figure 4](#)). In this setup, the classifiers were taken from the previous step, so they were fitted with Ltq data. High-resolution data was roughed through Gaussian convolution to decrease resolving power to low-resolution data.

[Figure 4](#) demonstrates the lack of classifier robustness to the different instruments despite the rather high similarity of MS profiles from [Figure 2](#). The combinations of the algorithms that seem to be overfitted from [Figure 3C and D](#) (Diffusion map, PLS-DA) leads to almost random classification for both low-resolution and high-resolution converted to low-resolution data.

Considering data without outliers provides better results ([Figure 5](#)) for overfitted models, whereas it has an almost negligible influence on the other models' accuracy. The exclusion of outliers with the SSM by indicating low-similarity blue strips mostly in glioblastoma samples reduces the number of samples from 144 to 116 for low-resolution data and from 144 to 113 for high-resolution. It increased the accuracy, for example, of NNMf+Linear SVM by five percent.

The combination of the PCA and AdaBoost approach demonstrates here the most robustness to the instrument.

Discussion

The SSMs in [Figure 2](#) demonstrate different variance of MS profiles in negative and positive mode. The negative mode has less similarity between MS profiles especially in the m/z range 100-2000. The positive mode has similar MS profiles excepting the outliers that are not contained in the negative mode^{24,42}. Thus, the positive mode could be used for looking

Figure 3. Accuracy scores matrices. (A-D) 100-2000 m/z; (E-H) 500-1000 m/z.

Figure 4. Accuracy by validation data. (A) Low-resolution; (B) high-resolution converted to low-resolution.

Figure 5. Accuracy by filtered validation data. (A) Low resolution; (B) high resolution.

for general distinctions between classes, whereas the negative mode could contain fine distinctions in the class as well. The classifiers seem to have difficulties in application to such high variation objects as MS profiles have in negative mode.

Preliminary feature selection (dimensionality reduction) is necessary for classifiers to not be overfitted, as the dimensionality of the object is equal to 7600 features in the given preprocessing (binning for m/z 100-2000 range) and the number of samples is about 150 (about 50 patients) in the dataset from the single instrument. The dimensionality reduction is expected to work rather effectively in the first several components only in case the variation is caused by interclass variability with classes characterized by a single Gaussian distribution. In the presented case, each class has a set of multi-dimensional Gaussian distributions. Also, astrocytoma tissues can have smooth changes of grades up to glioblastoma due to its origin. The whole variability of the data could be explained by six factors: tumor type, tumor localization, mutation status, interpatient variability, intratumor variability, and batch effect. Usually, interpatient variability is comparable with interclass variability. So, the presented data is unlikely to have most of its variability in the first components due to tumor type. Thus, both for the data with low (positive mode, Figure 2), and with high variance data (negative mode Figure 2) the first components of DR could be useless in terms of classification if the intraclass variability is higher than interclass.

PCA (Figure 3) provides good results for positive mode: most classifiers achieve similar accuracy of about 80%. PCA is ineffective for negative mode, as the accuracy falls to ~46%/54%, which corresponds to the ratio of the classes in the dataset. At the same time accuracy in the training data is much higher, so PCA provides specific only for the training data, which leads to overfitting. PCA and further classifiers predict mostly one class for any test data in negative mode.

NNMF components correspond to the highest accuracy values in most conditions. It seems to be not overfitted in positive mode. NNMF produces the best results for the m/z range 500-1000 in combination with random forest or decision tree models. NNMF and PCA repeat the accuracy of the test data with the KNN classifier for positive mode.

UMAP seems to be perspective, whereas Isomap, Diffusion map, PLS-DA algorithms demonstrate worse results compared with others. Classifiers are the most overfitted if no DR algorithm is applied, as expected. Although the AdaBoost model has about 70% accuracy on the test data for each experimental condition for non-preprocessed data (Figure 3).

PLS-DA is used for dimensionality reduction here, but it is common practice to use it as a classifier. PLS-DA appears to be overfitted for all cases except the 500-1000 m/z range in negative mode, where all other combinations are overfitted. Random forest demonstrates the best result for this case but still seems to be slightly overfitted, as the accuracy on the test data decreases by 20% compared with accuracy on the train data.

Neural net and AdaBoost models demonstrated the best results for uncompressed data in negative mode.

Algorithms were applied to a similar dataset of frozen samples, measured with another instrument (Orbitrap). As the measurements were carried out in two resolutions, the accuracy matrices are calculated for both resolutions. Figure 4 shows a lower level of accuracy compared to Figure 3, e.g. for PCA+AdaBoost. This is explained by the outliers, which could be seen in Figure 2B for frozen glioblastomas as the blue crossing stripes correspond to MS profiles that don't have similar MS profiles in the dataset. Excluding the outliers from the consideration reveals the accuracy improvements, but the accuracy is still not so high. The combination PCA+AdaBoost demonstrated the most robust result of classification over all the data in positive mode (Figure 3-5). Thus, the most independent combination of dimensionality reduction and classification algorithm to the instrument, resolution, and freeze-thaw process is supposed to be PCA+AdaBoost. This combination accuracy is not changing for positive mode from train to test data on any instrument and equals to about 70-80%.

Conclusions

1. Positive mode mass spectra provide better accuracy for astrocytoma and glioblastoma classification by mass spectrometric profiles of samples without sample preparation.
2. Astrocytoma and glioblastoma could be classified with 88% accuracy in low-resolution (NNMF+random forest) by positive-mode mass spectrometric profiles.
3. The PCA and AdaBoost combination appeared to be most stable in positive mode while transferring the classifier from the Lsq's to Orbitrap's data. The accuracy of classification is about 65-70% for validation data.
4. The dimensionality reduction algorithms combined with the classification models can process the outliers from the SSM as normal data.

Data availability

Zenodo: Data and code for comparison of different machine learning methods and dimensionality reduction for classification astrocytoma and glioblastoma tissues by mass spectra, <https://doi.org/10.5281/zenodo.4307700>⁴³.

This project contains the following underlying data:

Datasets of mass spectrometric profiles for different instruments, ranges, polarities, and resolutions.

Software files for figure replication.

Data are available under the terms of the [Creative Commons Attribution 4.0 International license](#) (CC-BY 4.0).

Acknowledgments

The research used the equipment of Shared Research Facilities of N.N. Semenov Federal Research Center for Chemical Physics of the Russian Academy of Sciences.

References

1. Ermolaev AY, Kravets LY, Smetanina SV, *et al.*: **Cytologic control of the resection margins of hemispheric gliomas and metastases.** *Zh. Vopr. Neirokhir. Im. N N Burdenko* 2020; **84**: 33–42.
[PubMed Abstract](#) | [Publisher Full Text](#)
2. Agar NYR, Golby AJ, Ligon KL, *et al.*: **Development of stereotactic mass spectrometry for brain tumor surgery.** *Neurosurgery* 2011; **68**: 280–89; discussion 290.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Clark AR, Calligaris D, Regan MS, *et al.*: **Rapid discrimination of pediatric brain tumors by mass spectrometry imaging.** *J Neurooncol* 2018; **140**: 269–279.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Eberlin LS, Norton I, Orringer D, *et al.*: **Ambient mass spectrometry for the intraoperative molecular diagnosis of human brain tumors.** *Proc Natl Acad Sci USA* 2013; **110**: 1611–6.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

5. Sorokin A, Shurkhay V, Pekov S, *et al.*: **Untangling the metabolic reprogramming in brain cancer: discovering key molecular players using mass spectrometry.** *Curr. Top. Med. Chem.* 2019; **19**: 1521–1534.
[PubMed Abstract](#) | [Publisher Full Text](#)
6. Carpinteiro A, Dumitru C, Schenck M, *et al.*: **Ceramide-induced cell death in malignant cells.** *Cancer Lett.* 2008; **264**: 1–10.
[PubMed Abstract](#) | [Publisher Full Text](#)
7. Wymann MP, Schneider R: **Lipid signalling in disease.** *Nat. Rev. Mol. Cell Biol.* 2008; **9**: 162–76.
[PubMed Abstract](#) | [Publisher Full Text](#)
8. Hannun YA, Obeid LM: **Principles of bioactive lipid signalling: lessons from sphingolipids.** *Nat. Rev. Mol. Cell Biol.* 2008; **9**: 139–150.
[PubMed Abstract](#) | [Publisher Full Text](#)
9. Lau D, Hervey-Jumper SL, Han SJ, *et al.*: **Intraoperative perception and estimates on extent of resection during awake glioma surgery: overcoming the learning curve.** *J. Neurosurg.* 2018; **128**: 1410–1418.
[PubMed Abstract](#) | [Publisher Full Text](#)
10. Povey JF, O'Malley CJ, Root T, *et al.*: **Rapid high-throughput characterisation, classification and selection of recombinant mammalian cell line phenotypes using intact cell MALDI-ToF mass spectrometry fingerprinting and PLS-DA modelling.** *J. Biotechnol.* 2014; **184**: 84–93.
[PubMed Abstract](#) | [Publisher Full Text](#)
11. Pereira HV, Amador VS, Sena MM, *et al.*: **Paper spray mass spectrometry and PLS-DA improved by variable selection for the forensic discrimination of beers.** *Anal. Chim. Acta* 2016; **940**: 104–12.
[PubMed Abstract](#) | [Publisher Full Text](#)
12. Cajka T, Smilowitz JT, Fiehn O: **Validating Quantitative Untargeted Lipidomics Across Nine Liquid Chromatography-High-Resolution Mass Spectrometry Platforms.** *Anal. Chem.* 2017; **89**: 12360–12368.
[PubMed Abstract](#) | [Publisher Full Text](#)
13. Anderson TJ, Jones RW, Ai Y, *et al.*: **High-resolution time-of-flight mass spectrometry fingerprinting of metabolites from cecum and distal colon contents of rats fed resistant starch.** *Anal. Bioanal. Chem.* 2014; **406**: 745–756.
[PubMed Abstract](#) | [Publisher Full Text](#)
14. Zhou W, Xia L, Huang C, *et al.*: **Rapid analysis and identification of meat species by laser-ablation electrospray mass spectrometry (LAESI-MS).** *Rapid Commun. Mass Spectrom.* 2016; **30**(Suppl 1): 116–121.
[PubMed Abstract](#) | [Publisher Full Text](#)
15. Cortés M, Pareja E, Castell JV, *et al.*: **Exploring mass spectrometry suitability to examine human liver graft metabonomic profiles.** *Transplant. Proc.* 2010; **42**: 2953–2958.
[PubMed Abstract](#) | [Publisher Full Text](#)
16. Hänel L, Kwiatkowski M, Heikaus L, *et al.*: **Mass spectrometry-based intraoperative tumor diagnostics.** *Future Science OA* 2019; **5**: FSO373.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. Moon KR, van Dijk D, Wang Z, *et al.*: **Visualizing structure and transitions in high-dimensional biological data.** *Nat. Biotechnol.* 2019; **37**: 1482–1492.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
18. Race AM, Bunch J: **Optimisation of colour schemes to accurately display mass spectrometry imaging data based on human colour perception.** *Anal. Bioanal. Chem.* 2015; **407**: 2047–2054.
[PubMed Abstract](#) | [Publisher Full Text](#)
19. Abramowski P, Kraus O, Rohn S, *et al.*: **Combined application of RGB marking and mass spectrometric imaging facilitates detection of tumor heterogeneity.** *Cancer Genomics Proteomics* 2015; **12**: 179–187.
[PubMed Abstract](#)
20. Mascini NE, Teunissen J, Noorlag R, *et al.*: **Tumor classification with MALDI-MSI data of tissue microarrays: A case study.** *Methods* 2018; **151**: 21–27.
[PubMed Abstract](#) | [Publisher Full Text](#)
21. Chagovets VV, Starodubtseva NL, Tokareva AO, *et al.*: **Validation of breast cancer margins by tissue spray mass spectrometry.** *Int. J. Mol. Sci.* 2020; **21**.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
22. Eberlin LS, Norton I, Dill AL, *et al.*: **Classifying human brain tumors by lipid imaging with mass spectrometry.** *Cancer Res.* 2012; **72**: 645–654.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
23. Pekov SI, Eliferov VA, Sorokin AA, *et al.*: **Inline cartridge extraction for rapid brain tumor tissue identification by molecular profiling.** *Sci. Rep.* 2019; **9**: 18960.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
24. Zhvansky ES, Eliferov VA, Sorokin AA, *et al.*: **Assessment of variation of inline cartridge extraction mass spectra.** *J. Mass Spectrom.* 2020; e4640.
[PubMed Abstract](#) | [Publisher Full Text](#)
25. Zhvansky ES, Sorokin AA, Pekov SI, *et al.*: **Unified representation of high- and low-resolution spectra to facilitate application of mass spectrometric techniques in clinical practice.** *Clin Mass Spectrom* 2019; **12**: 37–46.
[Publisher Full Text](#)
26. Zhvansky ES, Pekov SI, Sorokin AA, *et al.*: **Metrics for evaluating the stability and reproducibility of mass spectra.** *Sci. Rep.* 2019; **9**: 914.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
27. Kneen MA, Annegarn HJ: **Algorithm for fitting XRF, SEM and PIXE X-ray spectra backgrounds.** *Nucl Instrum Methods Phys Res B* 1996; **109–110**: 209–213.
[Publisher Full Text](#)
28. Pedregosa F, Varoquaux G, Gramfort A: **Scikit-learn: Machine learning in Python.** In: *J Mach Learn Res* 2011.
29. Cichocki A, Phan A-H: **Fast local algorithms for large scale nonnegative matrix and tensor factorizations.** *IEICE T Fund Electr* 2009; **E92-A**: 708–721.
[Publisher Full Text](#)
30. Tenenbaum JB, de Silva V, Langford JC: **A global geometric framework for nonlinear dimensionality reduction.** *Science* 2000; **290**: 2319–2323.
[PubMed Abstract](#) | [Publisher Full Text](#)
31. Barker M, Rayens W: **Partial least squares for discrimination.** *J. Chemom.* 2003; **17**: 166–173.
[Publisher Full Text](#)
32. Becht E, McInnes L, Healy J, *et al.*: **Dimensionality reduction for visualizing single-cell data using UMAP.** *Nat. Biotechnol.* 2018; **37**: 38–44.
[PubMed Abstract](#) | [Publisher Full Text](#)
33. Berry T, Harlim J: **Variable bandwidth diffusion kernels.** *Appl. Comput. Harmon. Anal.* 2016; **40**: 68–96.
[Publisher Full Text](#)
34. Altman NS: **An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression.** *Am. Stat.* 1992; **46**: 175–185.
[Publisher Full Text](#)
35. Chang CC, Lin CJ: **LIBSVM: A library for support vector machines.** *ACM Trans. Intell. Syst. Technol.* 2011; **2**: 1–27.
[Publisher Full Text](#)
36. Rasmussen CE, Williams CKI: **Gaussian Processes for Machine Learning.** Cambridge: MIT Press; 2005; Mass p. 266; ISBN 0-262-18253-X.
37. Breiman L, Friedman JH, Olshen RA, *et al.*: **Classification and regression trees.** *Biometrics* 1984; **40**: 874.
[Publisher Full Text](#)
38. Breiman L: **Random Forests.** Springer Science and Business Media LLC 2001.
[Publisher Full Text](#)
39. Hastie T, Rosset S, Zhu J, *et al.*: **Multi-class AdaBoost.** *Stat. Interface* 2009; **2**: 349–360.
[Publisher Full Text](#)
40. Rennie JD, Shih L, Teevan J, *et al.*: **Tackling the poor assumptions of naive bayes text classifiers.** In: *Proceedings of the 20th ...* 2003.
41. Hastie T, Tibshirani R, Friedman J: **The Elements of Statistical Learning.** In: *Springer Series in Statistics* 2nd ed.; New York, NY: Springer New York; 2009; pp. 106–119; ISBN 978-0-387-84857-0.
42. Eliferov VA, Zhvansky ES, Sorokin AA, *et al.*: **The role of lipids in the classification of astrocytoma and glioblastoma using MS tumor profiling.** *Biomed. Khim.* 2020; **66**: 317–325.
[PubMed Abstract](#) | [Publisher Full Text](#)
43. Zhvansky E, Sorokin A, Shurkhay V, *et al.*: **Data and code for comparison of different machine learning methods and dimensionality reduction for classification astrocytoma and glioblastoma tissues by mass spectra [Data set].** Zenodo 2020.
[Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 30 November 2021

<https://doi.org/10.5256/f1000research.31286.r99306>

© 2021 Chingin K. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Konstantin Chingin

Jiangxi Key Laboratory for Mass Spectrometry and Instrumentation, East China University of Technology, Nanchang, China

I fully agree with the first reviewer's comments.

On top of that, I suggest that the Introduction should be written in more detail. Also, the authors should provide more background on the determination of tumour boundaries by MS for other tissues as reported in earlier studies. Are there specific difficulties related to brain as compared to other types of tissues?

Language should be improved with the help of a native English speaker.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: mass spectrometry

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 15 March 2021

<https://doi.org/10.5256/f1000research.31286.r80908>

© 2021 Moshkovskii S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sergei A. Moshkovskii

Pirogov Russian National Research Medical University, Moscow, Russian Federation

A paper by Zhvansky et al aims to distinguish between glioblastoma and astrocytoma tumor tissues by analysis of mass-spectrometry tissue profiling. Mass-spectra recorded from tissues intra- and extra-surgery, using low and high resolution mass spectrometry. Authors tried to select an optimal mode for spectra recording and processing, to use, in perspective, a mass spectrometry during surgery. Similar efforts are recognized in the field. However, the paper may add to the prior art some new ways of data processing. In its present form, it is still difficult for perception by the readership and has to be crucially improved.

1. Abstract. Please specify what certain mass spectrometry methods were used. Actually, they were two, a linear ion trap and Orbitrap, not just 'different', as it was stated. The abstract must better illustrate a specific content of the work. Please check the language, a 'train' in abstract has nothing to do with a railway.
2. Introduction. A text neglects needs of readers. Many abbreviations are given without deciphering. Check all these PLS-D, tSNE, UMAP et al, these are not established terms in the field of biomedicine. In the end, better define the novelty of your approach. Also specify the aim more clear. Is it to classify between glioblastoma and astrocytoma? Why, again, do we need this during surgery?
3. A pipeline (clinics, analytics, data processing) is not quite clear. We have here two tumors with subclasses inside them, frozen vs. fresh tissues, two resolutions by two MS machines, and two modes of spectra recording. Please provide a figure (maybe instead of the Venn diagrams in Fig.1 or in addition) strictly designating what exactly was done. From the conclusion, it is only partly clear what type of pipeline gave the best result.
4. What was done with the spectra produced from the same tumor? If they are used by classifiers independently, that is a mistake. Then, the model mixed intertumoral and intratumoral variability, which is not possible. Further, what was finally done with the outliers? An explanation in page 4, the last paragraph, is difficult to understand. Please

clarify.

5. The idea of transition of the results between LTQ and Orbitrap seems to be not feasible. Even if ionization is the same, these ion traps act differently and may record different ions with different intensity, with inevitable loss in model performance. Please discuss with examples of similar use of classification models in mass spectrometry.
6. Discussion. Page 7, paragraph 2. A big piece of text with several unobvious statements was written without references, please provide them.
7. Figures. All legends are short and unclear for the reader who looks through the paper, sometimes with abbreviations not disclosed. Please make the legends self-explanatory.
8. Language. Many places are unclear due to uncertain English. Please ask the confident English speaker and writer to go through the text.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: mass spectrometry, proteomics, proteogenomics, molecular medicine

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research