UAM

Adam Mickiewicz University in Poznań

**Faculty of Biology**
**Institute of Molecular Biology and Biotechnology**

Poznań, 09.01.2022

**Referee report** concerning the **Doctoral dissertation of M. Eng. Katarzyna Frątczak** (née Bednarczyk) entitled: "***Dimensionality reduction and feature selection methods in the problem of grouping and classification of molecular imaging data***".

### *Overall background and characterization of the doctoral study*

One of the major clinical obstacles in the treatment of various types of cancer is the different behavior of cancer patients undergoing guideline therapies. An important determinant for this occurrence has been coined as inter- and **intertumor heterogeneity**. While intertumor heterogeneity refers to the differences in cancer characteristics between patients, **intratumor heterogeneity** refers to the clonal and non-genetic molecular diversity within a given patient. Mass spectrometry imaging (MSI), including Matrix assisted laser desorption ionization (MALDI) MSI, is a powerful tool for the untargeted but spatially resolved molecular analysis of biological tissues such as solid tumors. As it provides multidimensional datasets by the parallel acquisition of hundreds of mass channels, multivariate data analysis methods can be applied for the automated annotation of tissues. Moreover, it incorporates the histology of the tissue sample, which empowers reviewing the molecular information in a histopathological context (Baluff *et al*., Adv Cancer Res, 2017). Currently, there is no unambiguous software accessible for researchers and clinicians that would enable the fast processing, rapid statistical analysis and machine learning of wide-ranging MSI data. Hence, development of predictive systems for multiple types of cancer may thus enable earlier, faster and more accurate cancer diagnosis.

The doctoral thesis was carried out at Department of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, under the supervision of Prof. dr hab. Joanna Polańska, whose achievements in the field of research on biostatistics, data mining, machine learning, mathematical modelling are widely known and well cited. The joint supervision of the doctoral thesis, was carried out by Associate Professor Monika Pietrowska, at the Cancer Center and Institute of Oncology, Gliwice. Professor Pietrowska is a highly recognized researcher specializing in mass spectrometry-based proteomics of cancer, with MALDI-MSI on tissues being one of her specialties. The research presented in this dissertation was therefore, methodically carefully planned and conducted and well supervised.

The dissertation has a classical layout, although the Materials section followed by combined Results and Discussion sections stems out, although it also an accepted but rather less used way. In an extensive Introduction, the doctoral candidate discusses biological background of the study with chapters on histopathology, clinical proteomics and lipidomics. The other part of Introduction concerns various aspects of mass spectrometry including that on MALDI-MSI. The chapter of Materials deals with subsection on peptide and lipid imaging in **Head and Neck** cancer study (Bednarczyk *et al*., 2019) and **Pan-cancer microarray study** (cases of hormone-dependent cancers including prostate, thyroid, and testicular cancer and environment-dependent cancers encompassing head and neck, colon, and stomach cancers, partly in Kurczyk *et al*. 2020). The objectives (called here Aims) of the work appeared to be well formulated, followed by an extensively descriptive chapters of Results, subdivided into five sections (four to 8) dealing with bioinformatic and statistical analyses intertwined with follow-up discussion. This combination is perhaps a little too confusing judging the complexity of the presented work, although it addressed all major outcomes and considered topics for further research. The Thesis contains overall 118 pages, including **9 Tables** and **31 Figures**, and ends with a relatively short list of approximately 50 literature items, which in Reviewers' opinion, might have well been extended especially concerning the biological descriptions or mass spectrometry technology parts.

***Critical analysis of the obtained results***

The candidate has presented the dissertation related to the development of computational methods dealing with various steps of MALDI mass spectrometry imaging (MALDI-MSI) analysis, including data pre-processing, reduction of dimensionality, extraction of features, and their subsequent selection amenable for selective feature clustering. Such pipelines are extremely useful in the analysis of disease-related phenomena, especially concerning the highly genetically variable and molecularly complex disorders such as cancer. The candidate has focused her research efforts on analyzing the MALDI-MSI peptide and lipid datasets related to various types of cancer, i.e. (head and neck, prostate, testis, thyroid, stomach and colon), to build algorithms and tools amenable of analyzing large-scale data. Importantly, pre-processing efforts presented in this Thesis (i.e., adaptive baseline detection, removal of outliers, normalization and Gaussian mixture model- (GMM)-based peaks extraction, reduced the dimensionality of the data by more than 90%, without major information loss (verifying the **Aim 1** of the dissertation). Furthermore, by implementing the training datasets and scrutinizing the heterogeneity of tissue subtypes largely enhanced the classification quality (at the level of 93% for the weighted accuracy, validating the **Aim 2** of the dissertation). Another valid and important aspect of the dissertation concerned the buildup of a well-fitting feature selection algorithm in combination with adaptive scoring system, in consequence allowing modelling the tumor diversity. The strategy of selecting features followed by multiple random validations in combination with adaptive feature scoring allowed choosing those that were most information-rich within assigned classifier, while largely encompassing the heterogeneity of many neoplastic pathways. The calculated weighted accuracy above 90% proved successful of for the chosen strategy, further pinpointing common processes occurring in each type of cancer, regardless of the tissue origin. Importantly, the strategy proposed in this dissertation verified the presence of common molecular features of cancer, aiding in building of a flexible decision-making strategy tree based on MSI patients' tissue screening. Such classification strategy may support the clinicians in suitable cancer diagnostics strategies.

### *Pros of a Doctoral dissertation*

The reviewed dissertation presents a modern, comprehensive panel of computational approaches in the search for best suited, bioinformatic and statistical approaches amenable for large-scale scrutiny of MSI data. The doctoral student showed her skills in the computational analysis of multidimensional data and the use of this knowledge to analyze the molecular heterogeneity of cancer, both on the global scale as well topically. The work shows well the doctoral student's skills in the field of design and gathering the large-scale data of multiple MSI experiments (related to peptide and lipid imaging on tissues), buildup of suitable computational pipelines and finally choosing the right strategies for follow-up computational MS and functional validation analyses. The obtained results may contribute to the development of more effective methods aiding the cancer clinicians. The dissertation presents the good theoretical knowledge of the doctoral student as well as numerous skills in biostatistical analysis of multiple data.

***Critical comments and questions***
*Notes on planning and conducting experiments and interpretation of results*

*Overall, the manuscript is relatively well designed and written, but the misalignment in the styles of paragraphs related to the biological description and mass spectrometry methodologies are marked. The computational analysis of the data is clearly a major focus of the Doctoral candidate, and as such more precisely formulated.*

*1)* **The accompanying CV sheet** encompasses five publications of the doctoral candidate (4 peer-reviewed and one from proceedings), with the medium impact factor. The first important publication constituting the part of the doctoral study (Bednarczyk *et al.*) is wrongly cited. It shall be listed as *Journal of Molecular Histology*! Importantly, the publications used in the Thesis shall be listed separately as part of the Thesis and the contribution of the author explained in more detail. Furthermore, the name of the Doctoral candidate shall be marked and the change of name (née Bednarczyk*)* explained for clarity.

2) ***Introduction*, part Motivation**. The candidate introduces several important ideas and research questions, which are not supported by any reference. For example, "*More homogeneous tumours, which are easy to interpret histologically, show high spectral similarity and compactness*" is correct, but not explicitly clarified. Similarly, "*Studies have shown that regardless of the cancer heterogeneity, it is possible to distinguish between healthy and cancerous tissue*

*with high accuracy".* One would expect to strengthen such statements with the literature use. Furthermore, the link to *"The Cancer Genome Atlas"* shall be given.

3) **Biological background, chapter 2**. Importantly, on page 9 "Pan-cancer proteomes" shall be clarified. The link to WHO page with the accession date shall be provided (page 9). While introducing the MALDI-MSI technique one would be encouraged to provide an aiding scheme (pages 9-10).

4) **Histopathology, chapter 2.1**. The brief introduction of the pros and cons of fresh frozen (FF) vs formalin fixed paraffin embedded (FFPE) specimens would have been an advantage, especially concerning the recovery of signals in MSI technique. The term FFPE is not introduced, while cross-linking only briefly mentioned. The sentence "*Additionally, frozen tissues are also less known to pathologists, who are generally more convenient to diagnose FFPE upon microscopic analysis of the tissue"* is not truly accurate.

5) **Clinical proteomics, chapter 2.2**. Page 13, "*Due to numerous problems and limitations in microarray technology, sequencing for transcriptome analysis has increased in recent years.*", reference is missing. Page 14, "*The 35,000 genes in the human genome can encode at least ten times as many proteins; in extreme cases, a single gene can encode more than 1000.*" This statement provides an inaccurate information. End-to-end sequencing of the human genome (*Science*, 31.03.2022), includes 6 articles that lists **19.969 protein coding genes**. One shall rather talk about various proteoforms which are derived from the single gene due to posttranscriptional events and various posttranslational modifications, PTMs.

6) **Clinical proteomics, chapter 2.2**. Page 16, while talking about protein modifications the term PTM should have been introduced,

7) **Lipidomics, chapter 2.3,** the scheme concerning MSI of various biologicals, including lipids is missing (could have been combined with one at page 9/10).

8) While writing about **mass-to-charge ratios** (***m/z***) it shall be written in italics throughout the whole text, page 21.

9) While describing various types of ionization techniques, formed ions or MS analyzers, an informative scheme portraying various ion analysis options might have been more appropriate instead of just listing the various types (pages 22-25),

*10)* Page 26, Mass spectrometry method is not quantitative...This sentence is imprecise as quantitative methods are being used (i.e. based on labels). Furthermore, single/multiple reaction monitoring (SRM/MRM) methods are widely used for quantitative assessment. This could have been mentioned here.

*11)* Page 27. Proteome analysis of blood/plasma or other biological fluids is usually performed by immunological methods, i.e. ELISA, allowing for quantitative targeted assessment of biologicals (proteins, peptides),

12) Page 27, analysis of blood serum. *"The basic detected proteins are albumin…".* This statement is entirely imprecise. Albumins, immunoglobulins, apolipoproteins, etc. constitute over 90% of the proteome. The so called, depletion kits, are widely used as the first step of proteomic analysis of serum (i.e. removing top 14 or 20 contaminating proteins).

13) Page 28, ***chapter 3.1.*** The more appropriate nomenclature of compounds used as MALDI matrices shall be taken into account. For example, alpha-cyano-4-hydroxycinnamic acid (α-CN) is later in the text (on page 39) called as $\alpha$-CHCA ($\alpha$-CCA),

14) Page 29, we shall rather talk about good *sample solubility* not *homogenization*, for MALDI MS preparation.

15) Page 32, **Protein identification strategies in MALDI-MS**. The name of this chapter is rather inaccurate and should have been revised since the candidate talk about various measurement strategies. The digested peptides do not have to be eluted from gels, but the digestion process is usually performed in solution. The term shot-gun proteomics might have been introduced here as well as SRM/MRM targeted strategies. Page 33 on the bottom, what does the statement "…*different ions give the highest bending in the MS spectrum*…" relates to?

16) Page 32, "*Mass spectrometry has been used to analyze the blood serum proteome in diagnosing cancers of the head and neck region, breast, ovary, uterus, prostate, lung, colon, pancreas, thyroid, kidney, bladder, and liver.*" is not supported by any literature reference.

17) Page 33, MALDI enables measuring the distribution of a large number of analytes at one time without destroying the sample“, term “partially destroying” is more accurate here,

18) Page 33, “*MALDI-MSI is characterized by a relatively high sensitivity of ionic mass measurements…*”. More appropriate would be to say about *m/z* ratios. “The matrix must absorb *the light* at the laser wavelength and *desorb* and ionize the analyte” These two terms are missing,

19) Page 34, the part relating to MALDI-MSI and MRI technique requires a citation. Similarly, the sentence “*Similar results were obtained in the MALDI-TOF-MSI analysis of the lipid profile of prostate cancer, where prostate cancer was associated with fatty acid synthesis and lipid oxidation*” requires an adequate citation,

20) **Chapter 4**, **Methods** shall be renamed to **Material and Methods**. Importantly more information about the experimental setup should have been introduced. The types of experiments should be have been listed as labelled with appropriate literature positions of the candidate,

21) **Chapter 4.1**. The fragment starting with the sentence until “*Worldwide, head and neck cancer account for approximately 900,000 cases and over 400,000 deaths annually*, until …*laryngeal cancer is 20 times higher*.” is missing several references!

22) ENT examination should have been spelled out as “ears, nose, throat and neck examination”. This term is missing in the list of abbreviations,

23) Were the **images reproduced with the permission of a publisher**? **Figure 4 and onwards**? Such information must be added into the Figure legends.

24) The name of the matrix, DHB has been introduced already earlier.

25) The information about the tissue trypsin digestion step in the procedure is missing here (introduced in Bednarczyk *et al*., 2019), why was omitted here?

26) Page 37, instead of saying “removal of template” one shall refer to “matrix removal”. Please note, Bruker FlexAnalysis 1.4 software, is this version correct?

27) Page 39, TMA- tissue microarray, Heat induced antigen retrieval (HIAR), acetonitrile (ACN). The names should have been properly spelled out and listed in the Abbreviations part. Same page: Bruker format should be spelled as Bruker-specific data format,

28) Page 40, ROI- region of interest shall be specified,

29) Page 40, what does the term "nucleus" refer to in terms of spectra?

30) Page 43, were the images of MS spectra taken from Bednarczyk *et al.*, 2019 publication, it shall be specified,

31) Page 46, "threshold proposed by Chavenet", literature reference?

32) Page 51, the name Gaussian mixture model, GMM shall be spelled out,

33) Page 51, **Figure 8**. Mean spectrum (black lines)? Is this correct?

34) Page 51, protein quantity estimation should be protein abundance estimation, based on peptide abundances? **Figure 9**, peptides abundance for masses? This is rather confusing and should have been harmonized throughout text and the Figure legend,

35) Page 55, the scheme would be a clear asset in this part. For example, Histogram of inverse amplitude → decomposition to GMM component → selection of the optimal number of BIC criteria → intersection 3/4th component of the model,

36) Page 56, **Figure 13**, the legend is incomplete and not informative, **Figure 10**, peptides abundance for masses? Similarly as concerning the Figure 9, it is rather confusing and should have been harmonized throughout text and the Figure legend,

37) Page 59, another scheme should have been in place here. For example, GMM model → Adaptive feature filtering technique → components modelling → data reduction by 99%

38) Page 61, **Figure 17**. The *y*-axis should have been labelled as *Features,*

39) Page 69, **table 3**. Why weighted accuracy (wACC) is much lower for Thyroid cancer patients, this issue has not been addressed in the text,

40) Page 72, LOO CV abbreviation should have been introduced and spelled out as Leave One Out cross validation,

41) Page 83, internal MRV shall be specified as multiple random validations,

42) Same page, *curse of dimensionality shall* be called as *bias,*

43) Page 90, **table 6**. NPV (negative predictive value) and PPV (positive predictive value shall be specified in the legend,

44) Page 90, "The data we collect is often severely shaken" is rather misfortunate term. We talk about perturbation of data,

45) Page 91, "In this study, for each type of tissue, we observed unbalanced samples resulting from the detachment of the tissue cores from the plaque at the stage of biological preparation." This is unclear, what does the detachment of tissue cores relates to? I assume that there was severe tissue loss when embedding them?

46) Page 92, ROC curve= Receiver operating characteristic curve, the name shall be specified,

47) Page 93, although commonly used, the abbreviations in the equations 20-22 should have been spelled out and added. TP- true positives, TN- True negatives, FP- false positives, FN-False negatives, Se- sensitivity, Sp-specificity, CZ(c)= – concordance probability method.

48) Page 93, There is a description of a **Figure 24** as a plot of NPV vs. 1-PPV while in the Figure legend *y*-axis represents PPV and x-axis 1-NPV?

49) Page 99, when referring to the results of Pan-cancer analysis (first chapter) the publication by Chen *et al.*, *Nature Commun*, 2019 shall be listed,

50) Page 104, **Figure 29**. The information provided in this Figure is not optimal. Ideally the color images should have been provided, and the Figure merged with the **Figure 4**. The legend of the figure should have been revised accordingly.

51) Same page. Formula 19 is placed in the chapter 6.6, and not in the chapter 6.5, this shall be corrected,

***Editorial concerns***

1) *Page 8, shall be "All were assessed according…."*
2) *Page 12, phrasing "Currently, histological techniques are used not only to diagnose patients but also to form the basis of all development research." is not precise enough,*

3) *Page 12, format of citation by KW & MA, 2002 is incorrect,*
4) *Page 15 shall be "…metabolomics is presented in Figure 1.",*
5) *Page 16, the sentence shall be rephrased "After properly processing, various samples such as urine, cerebrospinal fluid and tissues provide a complementary, system-wide approach to explaining the interaction between environment and health…",*
6) *Page 17, phrase "The three main types of lipids that serve different roles in the human body:" is incomplete, same page "around your body's cells.." shall be "around one's body cells,*
7) *Page 18, in situ shall be in italics. "in high sensitivity equipment,…" shall be using high sensitivity…, "metabolic molecules" are simply metabolites,*
8) *Page 19, the sentence is rigged "However, a major challenge in interpreting the lipid distribution detected by MALD….". "I-MSI", what does that mean?,*
9) *Page 21, "the ionization of the molecules…" shall be "ionization of molecules",*
10) *Page 23, "Secondary Ion Mass Spectrometry, SIMS) - use " shall be "uses",*
11) *Page 24, phrase "This beam is directed to the load mass analyzer." shall be revised,*
12) *Page 26, phrase "Analyzing the mass spectra of the protein profiles of sick and healthy people makes it possible to study their differences." is imprecise and should be revised,*
13) *Page 32, phrase "Research on identifying proteins with use has an extensive application in diagnosing various diseases." is incorrectly built,*
14) *Page 34, "…bioinformatics methods that would allow us to deal with the non-ideality of the method."should be rephrased,*
15) *Page 35, "multiple two-dimensional sets." Term sets was missing.*
16) *Page 36, "the slides of tissues were scanned…", space is missing,*
17) *Page 37, shall be "in the Table 1" and (25.009 spectra),*
18) *Page 39, shall be "…steps in processing of biological material…",*
19) *Page 40, shall be pre-processing,*
20) *Page 40, shall be "The first step after data acquisition is data cleaning,…",*
21) *Page 42, shall be Savitzky-Golay filter not Savicki Goley,*
22) *Page 42, shall be "Jenkins, 1999). Baseline correction…",space is missing,*

23) *Page 50, shall be "Gaussian mixture of model distributions", same page "each distribution used for data represents one submolecule of the modeled phenomenon",*

24) *Page 51, shall be "peptide vs. lipids dataset",*

25) *Page 55, phrase "Even traditional peak identification methods then no longer find any finds." should be revised,*

26) *Page 62, should be "49% of peptides.", and "only ten features with a large effect..",*

27) *Page 64, Figure 18 shall be Orthogonal Projection,*

28) *Page 65, shall be "..in recent years  (van der Maaten..", space was missing,*

29) *Page 67, shall be ".., this is a very often an undesirable effect…",*

30) *Page 68, shall be "a histogram was created for amplitude and variance/", and "Then they were decomposed…",*

31) *Page 68,* ***Figure 20*** *legend shall be corrected to "Histogram decomposition of estimated peptide abundance and variance with a Gaussian mixture model."*

32) *Same page, sentence above shall be Gaussian mixture model,*

33) *Page 75, sentence shall be as "When choosing a method of classification…",*

34) *Page 76, the phrase "When developing an ML model using training data, evaluate the model's performance." shall be revised,*

35) *Page 78, the phrase shall read "..or evolve, and are called intra-tumour heterogeneity (Bedard et al., 2013).*

36) *Page 85, Equation 17, shall be as "data given",*

37) *Page 88, shall be plural for ""the number of analyzed features",*

38) Page 92, (Youden, 1950), please check the reference style,

39) *Page 94 "is" redundant in the sentence "…the created model is crucial before implementation.",*

40) *Page 97, shall be "70% training / 30% testing",*

41) *Page 99, shall be "basis of carcinogenesis",*

42) *Page 105, Figure 30 legend shall be "….results in the validation showing…*

### *Conclusions*

Practical comments and questions as well as indicated technical reservations do not, of course, affect my overall positive assessment of the dissertation. A quality of computational work presented by the Doctoral candidate, the multiplicity of techniques and statistical analyzes used arouse recognition and indicate on one hand the significance of the analytical work done by the PhD student, and on the other hand, very good computational skills. The method of preparation of the dissertation also proves ability to critically analyze results, both published and own. The dissertation presented for evaluation brings an important information to the knowledge about the mechanisms of cancer and points to converging pathways that may be involved taken into account when improving the methods of cancer diagnosis.

Based on the conducted formal analysis, I conclude that the reviewed doctoral dissertation of M. Eng. **Katarzyna Frątczak's** meets the requirements set out in the Act of March 14, 2003 on academic degrees and titles and degrees and titles in the field of art (Journal of Laws No. 65, item 595, as amended) (in accordance with Article 175.1 of the Act of July 3 2018, provisions introducing the Act - Law on Higher Education and Science (Journal L. 2018, item 1669). On this basis, I submit to the Scientific Council of the Discipline of Biomedical Engineering of the Silesian University of Technology in Zabrze for the admission of Ms Katarzyna Frątczak for further stages of the doctoral thesis.

Sincerely yours,
*Maciej M. Łałowski*

Maciej Lalowski, Ph.D., D.Sc., visiting professor
Department of Gene Expression
Institute of Molecular Biology and Biotechnology
Faculty of Biology,
Adam Mickiewicz University
Uniwersytetu Poznanskiego 6 St.
61-614 Poznań, Poland
**E-mail:** maciej.lalowski@amu.edu.pl

Associate professor
Finnish Proteomics Society, President
Principal investigator, Medicum
Biochemistry/Developmental Biology
Meilahti Clinical Proteomics Core Facility
PO Box 63 (Haartmaninkatu 8), Room C214a,
FI-00014 Uni. of Helsinki, Finland
Tel. (office) +358-294125203
Tel. (mobile) +358-407790950
**E-mail:** maciej.lalowski@helsinki.fi