THE PERSISTENCE TRANSFORMATION; A NEW METHODOLOGY OF TOPOLOGICAL DATA ANALYSIS

Gideon Klaila

A DISSERTATION

in

Graduate College RTG π^3 - Parameter Identification

Presented to the Faculties of the University of Bremen

in

Partial Fulfillment of the Requirements for the

Degree of Doctor rer. nat.

2023

Supervisor of Dissertation

Prof. Dr. Dmitry Feichtner-Kozlov, Chair of Algebra & Geometry, Director of ALTA Institute, Vice Dean of the Faculty of Mathematics and Computer Science

Graduate College Speaker

Prof. Dr. Dr. h.c. Peter Maaß, Speaker of the Graduate College π^3 - Parameter Identification, Leader of the working group Technomathematik

Examination Committee (Colloquium: 29.01.2024)

1.: Prof. Dr. Dmitry Feichtner-Kozlov, Primary Supervisor, First Reviewer
2.: Prof. Dr. Pawel Dlotko, Director of Dioscuri Centre in TDA, Second Reviewer
3.: Prof. Dr. Daniel Schmand, Leiter der AG Diskrete Optimierung
4.: Dr. Tim Haga, Senior Scientist at ALTA
5.: Dr. Pascal Fernsel, Coordinator of the Graduate College RTG π^3
6.: Tjado Edzards, Student of the University of Bremen

THE PERSISTENCE TRANSFORMATION; A NEW METHODOLOGY OF TOPOLOGICAL DATA ANALYSIS

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to several individuals and institutions who have played instrumental roles in the completion of this dissertation.

First and foremost, I extend my heartfelt thanks to my primary supervisor Prof. Dr. Dmitry Feichtner-Kozlov for his guidance, invaluable insights, and dedicated support throughout my research journey. His mentorship has been pivotal in shaping this work.

I am deeply appreciative of the support and resources provided by the Graduate College π^3, which has fostered a conducive research environment, allowing for the exchange of ideas and collaboration with fellow scholars. The financial support received from the Deutsche Forschungsgemeinschaft (DFG) through the Research Training Group (RTG) has been instrumental in the successful execution of this research, and I extend my gratitude to the DFG for this essential funding.

I am grateful to my colleagues, Lena Ranke and Vladimir Vutov, for their exceptional collaborations and contributions to the research. Their dedication and teamwork have been truly remarkable.

I would like to acknowledge Prof. Anastasios Stefanou for his valuable guidance and support in the writing and publication of my papers, which have significantly enhanced the quality and reach of this work.

My heartfelt thanks go to Lukas Mentz for his meticulous revision of my writing, ensuring clarity and precision in the presentation of my research.

Lastly, but most importantly, I extend my profound appreciation to my wife for her unwavering support, understanding, and encouragement throughout this challenging academic journey. Her patience and belief in my work have been my greatest source of strength.

This dissertation would not have been possible without the collective support and guidance of these individuals and institutions. Thank you for being an integral part of this research endeavor.

ABSTRACT

THE PERSISTENCE TRANSFORMATION; A NEW METHODOLOGY OF TDA

The field of Topological Data Analysis (TDA) continues to evolve as a powerful tool for the analysis of complex data. The motivation behind this research lies in the need to extend existing TDA tools to provide more accurate, efficient, and comprehensive analyses of intricate datasets. The primary research problem addressed herein pertains to the limitations of the Persistence Diagram, a fundamental TDA tool that does not inherently incorporate positional information of topological features. The absence of this crucial spatial context can lead to inaccurate results, especially when analyzing low-dimensional data. To tackle this issue, this dissertation introduces the Persistence Transformation, an innovative extension of the Persistence Diagram. It is designed to capture the positional information of topological peaks, enhancing the robustness and depth of TDA analyses. Key findings of this research include a comprehensive analysis of the properties and stability of the Persistence Transformation. Furthermore, a real-world application of this method demonstrates its effectiveness in the classification of MALDI data, highlighting the practical utility of the extension.

The originality and contribution of this work are underscored by the extension of the traditional Persistence Diagram. The introduction of the Persistence Transformation empowers mathematicians and data analysts to tackle a broader spectrum of complex problems, fostering more accurate results across diverse application domains. However, it is important to note that the Persistence Transformation generates results of a higher dimensionality when compared to the Persistence Diagram. While this enables a richer analysis of complex data, it may necessitate additional computational resources. Nevertheless, this research not only advances the domain of TDA but also opens the door to a wider array of analytical possibilities for complex datasets, offering valuable insights across various fields.

Keywords: TDA, Persistence Transformation, Persistence Diagram, Stability, Application

KURZFASSUNG

THE PERSISTENCE TRANSFORMATION; A NEW METHODOLOGY OF TDA

Aus dem Feld der Topologischen Datenanalyse (TDA) entwickeln sich kontinuirlich neue, leistungsstarken Werkzeug für die Analyse komplexer Daten. Die Motivation dieser Forschung liegt in der Notwendigkeit, bestehende TDA-Tools zu erweitern, um genauere, effizientere und umfassendere Analysen komplexer Datensätze zu ermöglichen. Diese Dissertation beschäftigt sich hauptsächlich mit den Limitierungen des Persistence Diagram, eines grundlegenden TDA-Tools, welches die Positionsdaten von topologischen Merkmalen nicht integriert. Das Fehlen dieses wichtigen räumlichen Kontexts kann zu ungenauen Ergebnissen führen, insbesondere bei der Analyse niedrigdimensionaler Daten. Um dieses Problem zu lösen, führt diese Dissertation die Persistence Transformation ein, eine innovative Erweiterung des Persistence Diagram. Diese Methode wurde entwickelt, um die Position von topologischen Feature zu erfassen und die Robustheit und Tiefe der topologischen Datenanalysen zu verbessern. Das wichtigste Ergebniss dieser Forschung umfasst eine umfangreiche Analyse der Eigenschaften und Stabilität der Persistence Transformation. Darüber hinaus demonstriert eine praktische Anwendung dieser Methode ihre Wirksamkeit bei der Klassifikation von MALDI-Daten und hebt die praktische Anwendbarkeit dieser Erweiterung hervor.

Der wissenschaftliche Mehrwert dieser Arbeit liegt in die Erweiterung des traditionellen Persistence Diagram. Die Einführung der Persistence Transformation befähigt Mathematiker und Datenanalysten, ein breiteres Spektrum komplexer Probleme zu bewältigen und genauere Ergebnisse in vielfältigen Anwendungsbereichen zu erzielen. Es ist jedoch wichtig zu beachten, dass die Resultate der Persistence Transformation im Vergleich zum Persistence Diagram eine höhere Dimensionalität haben. Dies ermöglicht eine tiefere Analyse komplexer Daten, kann jedoch zusätzliche Rechenressourcen erfordern. Dennoch fördert diese Forschung nicht nur das Gebiet der TDA, sondern eröffnet auch ein breiteres Spektrum analytischer Möglichkeiten für komplexe Datensätze und bietet in verschiedenen Bereichen wertvolle Erkenntnisse.

Schlüsselwörter: TDA, Persistence Transformation, Persistence Diagram, Stabilität, Anwendung

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

Introduction to Topological Data Analysis

1.1. Preface and Structure

In today's rapidly evolving technological landscape, the sheer expansion of data is both a boon and a challenge. Data is not only growing in volume but is becoming increasingly complex and higher dimensional. Moreover, it often arrives in the form of sparse, incomplete, or noise-corrupted datasets. Harnessing the potential embedded within such impaired data necessitates the deployment of specialized tools that can discern patterns, extract valuable insights, and make sense of the complexity. In response to this growing need for robust data analysis, Topological Data Analysis (TDA) has emerged as a powerful methodology. TDA capitalizes on topological features, unearthing latent structures within intricate and high-dimensional datasets. Within this field, a diverse array of methods offers remarkable resilience to noise, the capacity to visualize both global structures and local features, and the ability to prepare data for subsequent processing, including machine learning, clustering, and classification.

At the core of TDA lies the foundational tool of the Persistence Diagram, a cornerstone for characterizing and comprehending the topological properties of data. However, as the demands of data analysis continue to evolve and diversify, the conventional Persistence Diagram reveals certain limitations, particularly when confronted with intricate datasets like time series data, exemplified by Morse functions. This dissertation embarks on a journey to address these limitations within the TDA framework. The exploration centers on the "Persistence Transformation", a novel tool crafted to extend the capabilities of the Persistence Diagram by harnessing the positional information of topological features. The incorporation of positional data into the Persistence Transformation bridges the gap between topological properties and spatial relationships, revealing important insights, particularly in real-world applications where exact positioning holds a high value. In the following pages, the properties, stability, and practical applications of the Persistence Transformation are examined more closely.

The thesis is structured as follows. The first chapter gives a comprehensive examination of the traditional TDA pipeline, offering insights into the current state-of-the-art tools and methodologies within the field. This foundational understanding sets the stage for recognizing the specific domains to which the Persistence Transformation will make its valuable contribution.

The second chapter embarks on a detailed examination of the Persistence Diagram as an analytical tool for Morse functions, with a particular focus on applications where the absence of positional information could potentially yield inaccurate results. Subsequently, the Persistence Transformation is introduced as a novel approach designed to extend the capabilities of the Persistence Diagram, addressing the identified limitations. The methodology's intricacies, including its properties and its stability, are expounded upon in the subsequent section, which features the submitted paper "Stability of the Persistence Transformation". This section serves as the cornerstone of this dissertation, offering a thorough understanding of the method's reliability and robustness.

The thesis continues in the third chapter with a showcase of specific domains where TDA has been effectively applied, illuminating the versatile nature of this field. This serves as a preamble to a practical application of the introduced Persistence Transformation. The innovation of this tool and its power to extract valuable insights from real-world datasets is highlighted in the published paper titled "Supervised TDA for MALDI Mass Spectrometry Imaging Applications". In this study, MALDI data of cancer cells are prepared with the Persistence Transformation for subsequent machine learning classification.

Finally, the fourth chapter delves into a comprehensive examination of the advantages and disadvantages of the Persistence Transformation. Additionally, potential avenues for further research and development surrounding this methodology are discussed.

The Persistence Transformation represents a methodological advancement towards enhancing the analytical toolkit of TDA. Its ability to transcend the constraints of the traditional Persistence Diagram promises new aspects in data analysis, making it a valuable asset in the evolving landscape of modern data science.

1.2. Contextualizing TDA: A Comprehensive Overview

Topological Data Analysis stands as a revolutionary approach in the realm of data analysis, offering a new perspective on complex datasets by providing underlying structures within the data and relationships and connectivity of data points rather than relying solely on traditional statistic metrics. By delving into the concealed topological intricacies of data, TDA skillfully captures the essence of interconnections, voids, loops, and clusters embedded within. Equipped with this information, the data can be subjected to analysis from an alternative standpoint, thereby unveiling pathways to potential novel results. Real-world datasets often face complexities such as high dimensionality, noise, and incompleteness. However, TDA offers a new way to avoid these problems by considering the intrinsic nature of shapes and relationships that underlie them. By mapping the data to lower-dimensional topological spaces, it is possible to compress the intrinsic information. Moreover, the topological features are assigned a persistence, representing their lifespan. Noise frequently manifests in the data as features with low persistence values. Consequently, TDA presents an opportunity for noise reduction while preserving important features. Lastly, TDA methodologies are frequently accompanied by stability theorems, guaranteeing robustness of the output regarding to minor perturbations in the input. Harnessing these principles, TDA yields lower-dimensional insights that encapsulate the essential information within the datasets.

1.2.1. The TDA Pipeline

Data preparation and analysis

Despite being a relatively new field in data analysis, TDA has already found numerous applications where insights into the underlying structure provide significant value to various subjects. Applications of TDA are generally following a specific pipeline (see 1.1). The first step involves data collection and preparation, where information bearing data is gathered and their suitability for further analysis are checked. Subsequently, Topological Data Analysis tools are applied to the gathered data to extract the inherent structure that might not be apparent through traditional methods. Topological relevant information, such as data point interconnections, loop formations, void regions, or extant clusters within the distribution, are traced throughout the evolution of the

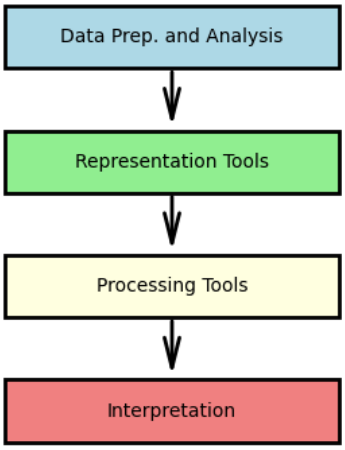


Figure 1.1: The Pipeline of TDA

data. Prominent examples of these tools contain the likes of Morse Theory, Topological Clustering, Persistent Homology and Cohomology and Euler Characteristics. Each of these examples are worth of a closer exploration to comprehend their unique contributions and functionalities.

Morse Theory constitutes a mathematical framework that delves into the critical points of real-valued functions, unraveling pivotal revelations about the underlying topological structure. These critical points assume distinct roles, being associated with distinct topological features such as maxima, minima, or saddle-points, and each offers insights into specific topological characteristics. The persistence of these critical features over varying scales offers insight into the geometric interplay within the data. This versatile tool finds application in diverse areas. In image segmentation, the gray-scale values of images are transformed into the image of a real-valued functions $f : \mathbb{R}^2 \to \mathbb{R}$. Distinct values are assigned to the critical points, subsequently enabling the identification of image segments through connected components within a specified index range. Another application is the shape analysis, where level-sets of the transformed data are analyzed. Finally, Morse Theory can be utilized for a piece-wise function approximation in dependency of the critical points (see among others Kozlov (2020, 2021); Mukherjee (2021); Upadhyay et al. (2022)).

Another tool of Topological Data Analysis is the **Topological Clustering**, which involves constructing structures such as simplicial complexes or Vietoris-Rips complexes based on the data.

These structures encompass topological attributes like loops, voids, and connected components, which exhibit sensitivity to the dataset's shape and interrelationships. The connectivity and interrelation of these features can be assessed through their persistence across various scales. This evaluation aids in discerning authentic structures from noise. Utilizing these insights, the dataset can be divided into clusters, each representing collections of data points that exhibit analogous topological attributes. Unlike traditional clustering methods, which rely solely on geometric distances, Topological Clustering incorporates topological connectivity to define cluster boundaries. Among its various applications, Topological Clustering is prominently employed in tasks such as data grouping and segmentation (see Songdechakraiwut et al. (2021); Perelman and Ostfeld (2011); Majumdar and Laha (2020)).

Persistent Homology emerges as an additional potent tool within the domain of Topological Analysis (see Otter et al. (2017)). Utilizing this tool involves methodically structuring data into filtrations through the gradual modification of a parameter across the dataset, frequently associated with distance or function values. At every stage of the filtration process, the data's homology groups encapsulate topological attributes such as voids, loops, or connected components. These attributes are traced from their initial emergence to their eventual vanishing as the parameter undergoes variation. Features that persist across a range of scales are considered robust and likely to be significant, while those with shorter persistence are more likely to be noise. Persistent Homology frequently finds application in the analysis of images, shapes, and point clouds (see Aktas et al. (2019); Townsend et al. (2020); Chung et al. (2021)).

An equally essential tool in Topological Data Analysis is the **Persistent Cohomology**, complementing Persistent Homology. Persistent Cohomology delves into the interactions between boundaries and cycles within the topological space, offering a dual perspective that illuminates how boundaries encompass cycles within the dataset. This provides a more comprehensive view of the global structure of a space and highlights its global properties. Similarly to Persistent Homology, Persistent Cohomology organizes data into filtrations by incrementally adjusting a parameter across the dataset. Within these filtrations, cohomology groups are computed to capture the connections

between boundaries and cycles. The persistence of these relationships is then monitored as the parameter undergoes variation. This approach is commonly employed in the examination of images, shapes, and networks (examples can be found in Vejdemo-Johansson et al. (2015); Kang et al. (2021); Pokorny and Kragic (2015)).

By integrating Persistent Homology and Persistent Cohomology, a more all-encompassing tool for Topological Data Analysis is formulated, which captures both intrinsic features and their relationships within complex datasets. This combination provides a dual perspective, enhancing the understanding of a dataset's structural nuances by showcasing both local features and global structures. In essence, the synergy of Persistent Homology and Persistent Cohomology enhances the realm of Topological Data Analysis, pushing progress across a spectrum of disciplines ranging from medical imaging to materials science and beyond.

While both Persistent Homology and Persistent Cohomology offer valuable insights, they come with inherent limitations. Their computational demands can be challenging to distribute efficiently, making them less versatile when applied to multifiltrations or large-scale datasets. In contrast, the **Euler Characteristic Curves** for one-parameter filtrations and **Euler Characteristic Profiles** for multi-parameter filtrations, while potentially less robust in one dimension, exhibit notable strengths. They can be computed using efficient distributed algorithms. Additionally, these invariants can be generalized to multifiltrations, extending their practical applicability to big data challenges while maintaining stability. More information about Euler Characteristic Curves and Euler Characteristic Profiles can be found in Dłotko and Gurnari (2022).

Representation Tools

Following data preparation and analysis in the conventional TDA pipeline, the subsequent stage involves transforming features to enhance their visual and structural interpretability by applying representation tools to the analyzed data. The representation of the data offers a way of visualization of the relevant features, aiding in the comprehension of the data's topological structure. This process not only simplifies result summary but also enhances interpretability by offering a concise visual or structural representation, which highlights patterns within the data. Typically representation

tools are Persistence Diagrams, Persistent Barcodes and Persistent Landscapes or the linkage of features with algebraic objects. These particular tools warrant a more in-depth and comprehensive exploration.

Persistent Barcodes leverage the capacity to decompose the topological dataset into a direct sum of feature intervals. Each feature interval represents the lifespan of a topological feature in relation to the filtrations. The intervals are displayed in a way that enables differentiation between feature dimensions and multiplicities (see Xia et al. (2018); Murayama et al. (2023)). Similarly, **Persistence Diagrams** also depend on these intervals. However, instead of representing bars to display the lifespan of a feature, they are depicted as points with coordinates (birth, death), where birth denotes the scale at which the feature emerges, and death signifies the scale at which it vanishes (see Marchese and Maroulas (2018); Chung et al. (2021); Leon et al. (2013)). Expanding upon the Persistence Diagram, the **Persistent Landscape** takes the concept a step further through the application of a transformation that converts the points of the diagram into functions. These functions encapsulate the persistence of features across multiple filtration scales. The resulting landscape is constructed by stacking these functions, creating a visual representation of the persistence values. This approach enhances the ability to display complex details regarding the topological features of the data and their evolution (for more details, see Bubenik and Dłotko (2017); Bubenik (2020)).

Another prominent tool for representing features specifically for Persistent Homology is through the utilization of **Persistent Homology Algebra**. In this approach, the extracted topological information is translated into an algebraic structure known as the Persistent Module. These structure is a collection of vector spaces, with each space corresponding to a distinct interval in the Persistent Homology calculation. These spaces contain the attributes associated with the birth and death of features at specific scales within the dataset. The underlying algebraic structure enables a more systematic and rigorous analysis of the topological features within the data, facilitating the application of algebraic tools for further analysis. Likewise, the features identified through Persistent Cohomology can be depicted using **Persistent Cohomology Algebra** techniques. Through the translation of cohomology features into Persistent Cohomology Modules, it becomes possible to

delve into the data's global structure and investigate the interplay between boundaries and cycles. The integration of algebraic tools amplifies the interpretative potential and analytical capacities of Persistent Cohomology, thus enabling the acquisition of deeper insights into the fundamental characteristics of the underlying data. Enhancing the visual depiction of insights derived from Persistent Homology Algebra and Persistent Cohomology Algebra can be achieved by subsequently employing other representation tools of TDA, such as Persistent Barcodes or Persistence Diagrams (see Bubenik and Milićević (2021)).

Processing Tools

Subsequent to data collection, analysis, and representation, the TDA pipeline proceeds to the application of processing tools. They are developed to take the extracted topological features from the represented data and conduct further analysis, augmentation, or utilization for various objectives. These tools primarily involve processes for transformation, classification, and leveraging the derived topological insights to gain further understanding or predictive capabilities. Examples for further processing tools include Mapper, Interleaving Distances, Kernel Methods, Smoothing and Denoising, and Machine Learning techniques. Each of these examples warrants a more in-depth investigation to understand their individual contributions and functionalities.

The Mapper is specifically engineered for the visualization and analysis of high-dimensional datasets. It operates by segmenting the data into overlapping regions and subsequently constructing a simplified graph-like structure. Within each region, data points form clusters, and the tool calculates topological attributes for each cluster. By summarizing this wealth of topological information of each cluster, the Mapper facilitates a comprehensive grasp of the data's overall structure and underlying features (see Lauric et al. (2022); Thatcher et al. (2021); Feng (2021)).

Another valuable processing tool is the **Interleaving Distance**, tailored to the comparison of algebraic structures that represent the Persistent Homology of a dataset, such as Persistent Modules. This distance metric serves as a guideline for quantifying the similarity between these structures, shedding light on how closely the topological features in one module can align with those in another. Consequently, it enables meaningful comparisons between different data representations.

By examining the correspondence between Persistent Features, the Interleaving Distance unveils the robustness and stability of identified topological structures, even in the presence of slight data perturbations. It finds versatile applications in diverse fields, including shape analysis, clustering validation, and the assessment of the stability of TDA methods (see Lesnick (2015); Gasparovic et al. (2019)).

A notably powerful processing tool for enhancing the performance of subsequent analytical techniques is the utilization of **Kernel Methods**. These methods serve to project data into higher-dimensional spaces, potentially rendering the intrinsic relationships between data points more linear or separable. This transformation can lead to more efficient analyses using methods such as Support Vector Machines (see Ma and Guo (2014); Noble (2004)) or Principal Component Analysis (see Giuliani (2017); Abeywardena (1972)).

To address the challenge posed by data noise, the application of **Smoothing and Denoising** techniques becomes crucial. These processing tools serve to unveil underlying patterns within the data while eliminating spurious topological features that may arise from noise. The elimination of this noise significantly enhances the precision and stability of subsequent analyses (see Ravishanker and Chen (2019); Rosen et al. (2019)). Examples of smoothing and denoising techniques include, among others, Density Estimation (see Phillips et al. (2013)) and Persistent Betti Curve Smoothing (see Ryu (2021)).

Finally, it is worth mentioning that **Machine Learning** techniques can be effectively employed with the topological features. These powerful algorithms have the capacity to discern intricate patterns and relationships within the data, including the topological attributes extracted through TDA. Prominent machine learning algorithms for this purpose encompass Random Forest, Neural Networks, and clustering algorithms such as K-Means and Hierarchical Clustering. These methods enable data-driven predictions and classifications, enhancing the insights drawn from topological data analysis (see Prantzalos et al. (2023); Hensel et al. (2021); Ignacio et al. (2019)).

Interpretation and Visualization

The final step in the TDA pipeline is the interpretation of the results, which can be achieved through various methods. Visualization plays a pivotal role in enhancing the comprehension of these outcomes. Tools like the Persistence Diagram and the Persistent Landscape can once again be leveraged to craft meaningful visual representations. Alternatively, representing the results using graphs and networks not only aids in visualizing the data but also enriches the ability to extract intrinsic information, enabling deeper insights. Moreover, employing machine learning models for interpretation is another valuable approach. After training with the analyzed data, these models can be extended to uncharted data, expanding the potential applications and deepening the understanding of the results. To assess the model's applicability to new, unseen data, it is common practice to set aside a test dataset. This separation allows for the validation of the model's accuracy and provides insights into its effectiveness in handling unfamiliar data.

In conclusion, the Pipeline of TDA stands as a powerful tool in modern data analysis. Its core advantages include the revelation of hidden topological structures in complex datasets, robustness to noise, and the capability to handle high-dimensional data effectively. TDA's versatility spans across various fields, making it indispensable in disciplines such as biology, neuroscience, and image analysis. Stability theorems ensure the pipeline's results can be generalized to new, unseen data.

In practical applications, TDA offers visualization tools like Persistence Diagrams and Landscapes for intuitive data representation. It aids in pattern recognition, data reduction, and data interpretation, making it invaluable for decision-making processes. Furthermore, TDA's role extends to predictive modeling by training machine learning models on its results.

Ultimately, the TDA pipeline serves as a bridge from complex data to actionable insights. Commencing with data representation and advancing to data processing, it concludes with visualization or trained machine learning models. This comprehensive approach equips data analysts with the tools to unveil, process, and interpret intricate data structures, rendering TDA an essential asset in contemporary data analysis.

CHAPTER 2

Methodology and Stability of the Persistence Transformation

2.1. Motivation for Enhancing TDA Methods

Data manifests in diverse forms across real-world applications, and, as demonstrated in the preceding chapter, there are numerous of topological tools to dissect this varied landscape of data. In this chapter, our attention turns to a specific category of data, namely, those that can be represented as Morse functions. These data, encompassing a subset of time series, finds appearance in a multitude of domains. From the realms of signal processing and time series analysis to the frontiers of geospatial, sociological, and biomedical data analysis, Morse functions form the foundational underpinning of a wide range of crucial undertakings. Chapter 3 provides a more in-depth exploration of specific real-world applications.

Among the foremost topological tools employed in the analysis of such data is the Persistence Diagram. This method entails generating filtrations for the dataset's elements, calculating homology classes for each filtration, and subsequently monitoring the persistence of these homology classes. The results are vividly represented within Persistence Diagrams, offering a comprehensive means of conveying invaluable topological insights. Particularly noteworthy among the utilized filtrations are the sub-levelset and upper-levelset filtrations, both of which warrant a more comprehensive examination.

For a Morse function $f : M \to \mathbb{R}$ on a compact set $M \subseteq \overline{\mathbb{R}}$, the sub-levelset filtration and the upper-levelset are defined as follows (cf. Klaila et al. (2023b)):

Definition 2.1.1. For $a \in \mathbb{R}$,

- the **sub-levelsets** are the sets $M_{\leq a} := \{x \in M | f(x) \leq a\}$.
- the **upper-levelsets** are the sets $M_{\geq a} := \{x \in M | f(x) \geq a\}$.

The encapsulation $M_{\leq a} \subseteq M_{\leq a'}$ and $M_{\geq a'} \subseteq M_{\geq a}$ for any $a \leq a'$ defines the corresponding

filtrations.

The topological features tracked in these filtrations pertain to path-connected components. A feature is considered 'born' at time a^* if it first appears in the sets $M_{\leq a^*}$ or $M_{\geq a^*}$. Conversely, it is said to 'die' at time a^+ if it merges with another path-connected component, which exhibits greater persistence, within the sets $M_{\leq a^+}$ or $M_{\geq a^+}$. This merging process aligns with the principles outlined in the Elder Rule (see Edelsbrunner and Harer (2010)).

The persistence of the topological features extracted from these filtrations is visually represented in Persistence Diagrams. In the case of the sub-levelset filtration, these features are denoted as points (a^*, a^+), while in the upper-levelset filtration, they are represented as (a^+, a^*), with multiplicity indicating multiple features sharing the same birth and death values. The distance of these features to the diagonal line $\{(x, f(x) = x) \mid \forall x\}$ corresponds to their persistence. For a more comprehensive understanding of the standard approach to Persistence Diagrams, additional details can be found in Edelsbrunner et al. (2002) and Cohen-Steiner et al. (2007).

An inherent limitation in the analysis of Morse functions, and real-valued functions in general, lies in their one-dimensionality. They are essentially represented as lines devoid of intersections, which translates to a lack of loops, voids, or higher-dimensional topological spaces. Consequently, only zero-dimensional topological features, specifically the path-connected components, can be extracted. This inherent one-dimensionality restricts the applicability of various TDA tools and results in limited diversity in outcomes. An issue frequently encountered when dealing with substantial volumes of data is the presence of numerous features with nearly identical birth and death values, resulting in notably similar Persistence Diagrams for different input datasets. A compelling illustration of this occurs in the case of symmetric functions. Here, the Persistence Diagrams are identical, yet the order of features is exactly reversed, often signifying dissimilar input sets. This complication can yield inaccurate outcomes, particularly in tasks like classification.

Addressing this challenge, the Persistence Transformation emerges as an extension of the traditional Persistence Diagram. Its primary purpose is to not only monitor the persistence of topological fea-

tures but also capture their positional information, enabling the differentiation of extensive datasets represented as Morse functions. By considering the feature positions, even symmetric functions can be readily distinguished from one another. Such information proves invaluable in various applications, particularly in scenarios where distinct peak positions correspond to specific, unique events, setting them apart from other data.

The Persistence Transformation, serving as a representation tool within TDA, can be employed on gathered and prepared data to provide a comprehensive perspective on the distribution and persistence of topological features. Additionally, the outcomes are readily adaptable for subsequent processing stages, including applications in machine learning and clustering. A similar approach to data representation for subsequent analysis is evident in Weis et al. (2020), wherein data transformation into persistence data serves as the preparatory step for applying Support Vector Machines.

In summary, the Persistence Transformation serves as a pivotal tool as extension to the Persistence Diagram, offering valuable insights into the distribution and persistence of topological features. This sets the stage for a closer examination of the method's intricate details in the upcoming section, where we delve into the specifics of its characteristics and stability (cf. Klaila et al. (2023a)).

2.2. Stability of Persistence Transformation

Gideon Klaila, Anastasios Stefanou & Lena Ranke

- **Status:** Submitted on 09.10.2023
- **Journal:** Journal of Applied and Computational Topology
- **Impact Factor:** 2.26
- **Preprint:** Published on arXiv (https://arxiv.org/pdf/2310.05559.pdf, 09.10.2023)
- **Attached:** In Appendix A

The Persistence Transformation is a representation method of Topological Data Analysis designed to expand the capabilities of the Persistence Diagram, a primary tool for analyzing time series data such as Morse functions. The Persistence Diagram of data filtration's, such as the upper-levelset filtration, represents births and deaths of all topological features. The Persistence Transformation further enhances this analysis by incorporating their positional information. The paper "Stability of the Persistence Transformation" offers a comprehensive analysis of the Persistence Transformation, providing a rigorous definition, and establishing its theoretical stability. This stability guarantees consistent and comparable results in real-world applications.

In the context of topological time series analysis, the critical points take center stage, as they encapsulate vital data regarding the birth, death, and positional attributes of topological features. This implies that functions sharing similarities, such as Morse functions related through Morse isotopy, will exhibit identical Persistence Transformations. To extend the applicability of this observation, the paper introduces the methodology of the Persistence Transformation directly within a space of critical points. This innovation broadens the scope of these results, making them applicable to entire isotopy classes of Morse functions.

As previously noted, within the practical applications of the Persistence Transformation, method stability assumes a central role. It implies that minor variations in the input will lead to only slight

alterations in the output. In many applications, such as classification, this is beneficial because similar results signify similar inputs. To evaluate the stability of the newly introduced method, it is necessary to define appropriate metrics within the relevant spaces that capture the concept of similarity in both the input and the output, respectively.

The output of the Persistence Transformation results in a discrete set of points within the three-dimensional Euclidean space. Comparing two discrete point sets is a common task, often accomplished using the p-Wasserstein distance. This distance metric calculates the minimal sum of the distances between matched points in the two sets. Points can be matched to corresponding points in the other set or to the diagonal plane, which consists of infinitely many trivial points. For two sets of points $A, B \subseteq \overline{\mathbb{R}}^n$, this metric is formally defined as follows (see Mémoli (2011); Berwald et al. (2018)):

$$d_{W_p}(A,B)^p = \min_{\text{matchings } m:A\times B} \sum_{(a,b)\in m} ||a-b||_\infty^p.$$

The versatility of this distance measure lies in its parameter p, which provides the capability to assign weights to distances, allowing for precise control over the influence of anomalies. For $p \to \infty$, it converges to the bottleneck distance, emphasizing the significance of outliers. The p-Wasserstein distance is vital for proving the stability of the Persistence Diagram, as demonstrated in works like Cohen-Steiner et al. (2005) and Skraba and Turner (2020). For more details on these distance metrics, refer to Cao et al. (2023). This versatile metric effectively captures distances for discrete point sets. Consequently, we also employ it in the formulation of the stability theorem for the Persistence Transformation.

To measure closeness in the input, the Persistence Diagram employs the supremum norm. For two Morse functions f and g mapping from $M \subseteq \overline{\mathbb{R}}$ to $\mathbb{R}$, this norm calculates the distance as $\sup\{|f(x)-g(x)| : x \in M\}$, signifying the maximum difference between the function values across all points. The application of the supremum norm is appropriate as the Persistence Diagram solely encodes topological information related to a single dimension: the height of the features. The Persistence

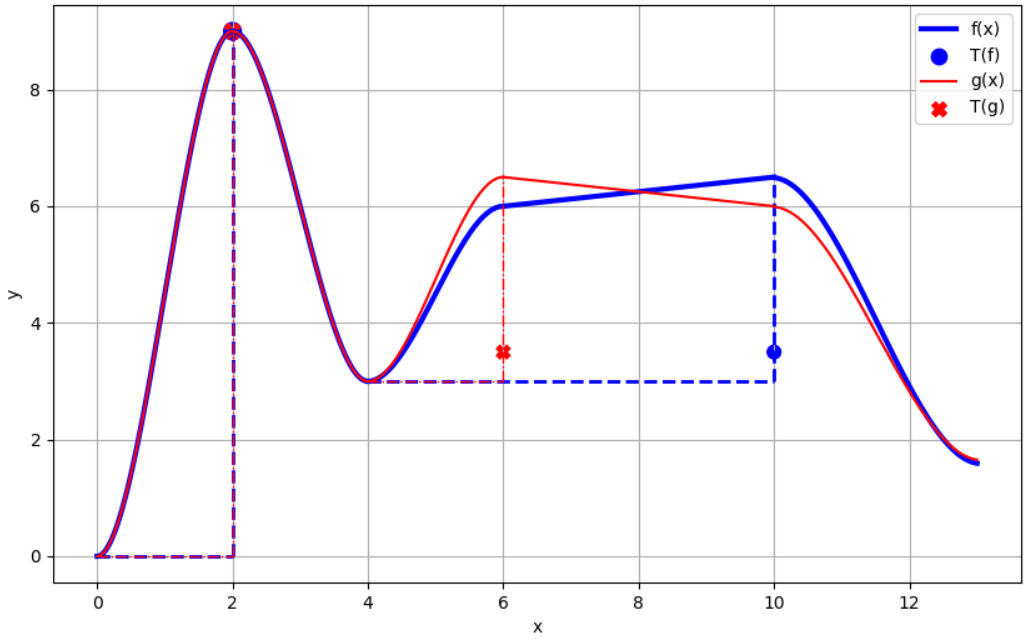


Figure 2.1: Example 1: Influence of an Additional Dimension on the Stability

Transformation, in contrast, seeks to monitor topological changes across two dimensions: both the height and position. Moreover, adhering to the Elder Rule (as outlined in Edelsbrunner and Harer (2010)), any similarity metric used for assessing the stability of the Persistence Transformation must account for a two-dimensional distance between the input values. This requirement becomes clearer when considering some examples.

The first example presented in Figure 2.1 effectively demonstrates the influence of an additional dimension in the topological tracking. It depicts two Morse functions, denoted as f in blue and g in red, which are ε close when considering the supremum norm. However, a significant difference emerges in the topological information of the smaller peak due to an arbitrary shift along the x-axis. This shift results in a noticeable disparity in the positional information captured by the Persistence Transformation features ($T(f)$ and $T(g)$, marked with the corresponding colors). Consequently, this example reveals the instability of the Persistence Transformation when evaluated solely based on the supremum norm. Thus, it underscores the importance of incorporating positional distances

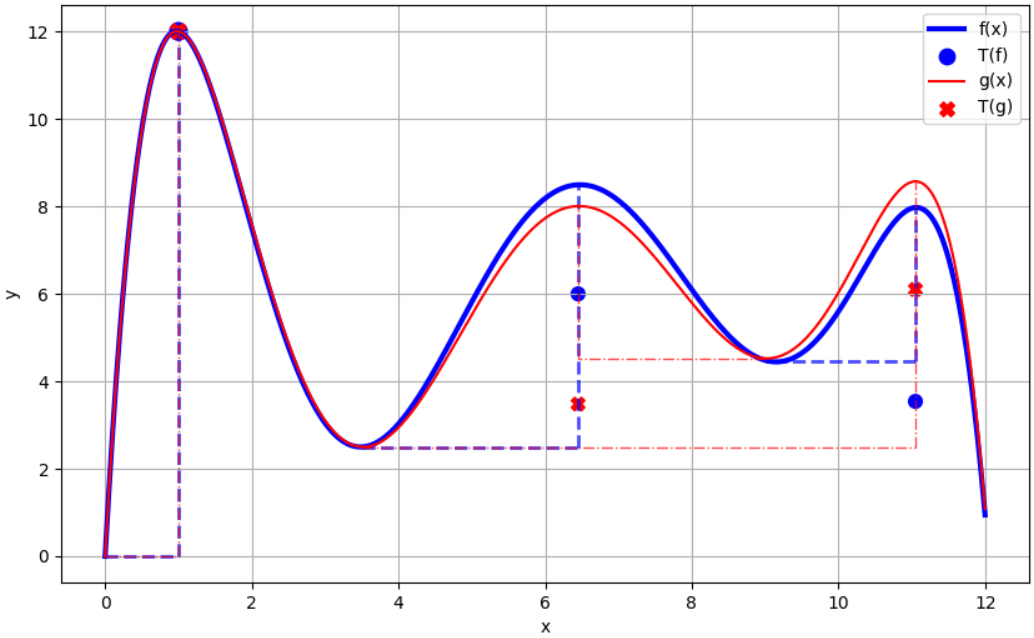


Figure 2.2: Example 2: Influence of the Elder Rule on the Stability

between the peaks into a more suitable metric.

The second example, as depicted in Figure 2.2, emphasizes the pivotal role of the Elder Rule in achieving stability. In this case, two Morse functions (f in blue, g in red) are depicted, which are ε close to each other according to the supremum norm. However, the order of the peak heights in f differs from that in g. As a result, at the merging point of the two topological features, the Elder Rule dictates distinct persisting feature for the two functions. Consequently, the corresponding points in the Persistence Transformation can be arbitrarily far apart, rendering the Persistence Transformation unstable when equipped with the supremum norm. Accordingly, the influence of the Elder Rule should be considered when determining the metric for evaluating the stability of the Persistence Transformation.

These two examples emphasize the necessity of embracing an alternative concept of proximity in relation to the input of the Persistence Transformation. As a solution, we introduce a new ℓ_p metric

for the established space of critical points, which accounts for the additional dimension in the output as well as the influence of the Elder Rule. With this new ℓ_p metric, we establish the stability of the Persistence Transformation as a central element of this paper.

One disadvantage of the Persistence Transformation lies in its additional output dimension. While the Persistence Diagram produces a discrete set of two-dimensional points, the elements of the Persistence Transformation are three-dimensional. As an additional contribution of this paper, we introduce the Reduced Persistence Transformation. This modification aims to reduce the output dimension by focusing on tracking persistence instead of birth and death information for each feature. This reduction in dimensionality comes at the cost of some information loss. However, the paper demonstrates that in certain cases, the omitted details are non-essential, rendering the Reduced Persistence Transformation a suitable choice. We additionally provide a concise comparison between this reduced approach and the traditional Persistence Diagram.

In this paper, we introduce the concept of the Persistence Transformation as suitable extension to the Persistence Diagram of the upper-levelset filtration. By tracking the persistence of peaks, this extension enhances our understanding of data analysis. An intriguing potential modification involves tracking the persistence of valleys instead, aligning the results with the well-established persistence diagram associated with the sub-levelset filtration. The theory pertaining to the properties and stability of the Persistence Transformation, as detailed in this paper, can be seamlessly extended to accommodate this adapted version, providing a comprehensive framework for further analysis and exploration.

In the emergence of the paper, I had a central role in the development and theoretical underpinning of the Persistence Transformation. This encompassed the conception and creation of the Persistence Transformation method, as well as the formulation of the metric applied to the space of critical points. Additionally, I took on the responsibility of establishing the stability theorem and delivering a rigorous proof of its validity.

In conclusion, the paper "Stability of Persistence Transformation" introduces a novel technique in

the field of Topological Data Analysis, extending the conventional Persistence Diagram to capture additional information and enhance existing methodologies. We demonstrate the method's stability and provide an adaptation in the form of a reduced version. Our work highlights the practical significance of these methods over existing approaches and lays the foundation for their real-world applications

CHAPTER 3

Application of the Persistence Transformation

3.1. From Theory to Practice: TDA's Successful Applications

Upon establishing the theoretical stability of the Persistence Transformation, the subsequent chapter serves as a bridge connecting theory to practical application. It commences by introducing various exemplary areas where Topological Data Analysis is commonly applied, ultimately culminating in a real-world application of the Persistence Transformation. This application demonstrates the practical utility of this innovative approach.

The application of TDA methods holds significant relevance in contemporary research and industry, thanks to its remarkable capacity to unveil concealed patterns, extract meaningful features, and provide profound insights into the structures of complex datasets. TDA acts as a vital link between raw data and actionable knowledge, and its successful applications span diverse fields, highlighting its universal utility (see Cheng (2020)).

In the realm of biology and medicine, TDA has enabled breakthroughs in understanding intricate molecular structures. For instance, researchers have utilized TDA methods to analyze protein-protein interaction networks (e.g., Sardiu et al. (2019)), identifying critical nodes and potential drug targets with unprecedented precision. In genetics, TDA aids in deciphering the underlying structures of DNA sequences, giving insights on evolutionary relationships (see Mandal et al. (2020)) and aiding in diagnosing diseases (see Bukkuri et al. (2021)). An example of the application of TDA in the area of medical biology is the prediction of kernel-based microbial phenotype (see Weis et al. (2020)).

In material science, TDA has proven its valuable impact in the analysis of complex materials such as porous structures (see Li et al. (2022); Rudkin et al. (2023)). By representing pore spaces as topological features, researchers have efficiently assessed fluid flow properties in diverse materials, contributing to the optimization of porous materials for various applications, from filtration to

energy storage (e.g., Moon et al. (2017)).

Urban planning and transportation optimization have also benefited from TDA methodologies (e.g., in Carmody and Sowers (2021)). By analyzing urban data, TDA can identify clusters of activity, helping planners in optimizing public transport routes, reducing traffic congestion, and enhancing the overall urban experience (e.g., in Kasatkina et al. (2022)).

Another example is the finance sector, where TDA has been utilized to analyze market trends and portfolio diversification (see Gidea and Katz (2018); Guo et al. (2023)). By representing financial data as topological landscapes, TDA identifies potential market regimes, enabling investors to make informed decisions in a variable environment (see Sokerin et al. (2023)).

Furthermore, TDA plays a pivotal role in neuroscience. Neuroscientists apply TDA to map brain connectivity, revealing functional modules within neural networks and offering insights into disorders such as Alzheimer's and Parkinson's diseases (see Caputi et al. (2021)).

In conclusion, the common thread weaving through these diverse applications is TDA's remarkable ability to capture the underlying geometry and structure concealed within complex datasets. This proficiency transcends the limitations of conventional data analysis methods, enabling us to decipher intricate patterns that might have otherwise remained elusive. By unveiling this geometric perspective, TDA empowers us to navigate the complex tapestry of data, revealing hidden insights and guiding our understanding in unprecedented ways.

As TDA evolves, its scope of potential applications widens, presenting new avenues for comprehending the intricate systems shaping our world. This expansion involves the innovation of analytical techniques and their successful application to real-world datasets. In the forthcoming chapter, we explore this progression by introducing the practical implementation of the Reduced Persistence Transformation, exemplifying its efficacy through an application to MALDI-Images.

3.2. Supervised TDA for MALDI Mass Spectrometry Imaging Applications

Gideon Klaila, Vladimir Vutov & Anastasios Stefanou

- **Status:** Published
- **Journal:** BMC Bioinformatics
- **Impact Factor:** 11.806
- **Date of publication:** 10.07.2023
- **DOI:** https://doi.org/10.1186/s12859-023-05402-0
- **Attached:** In Appendix B

TDA strives to establish a connection among mathematics, data science, and a variety of application domains. Therefore, the relevance of the constructed method depends on the functionality in real world scenarios. An illustrative domain where the benefit of the (Reduced) Persistence Transformation in medical contexts becomes evident is Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry Imaging (MALDI MSI), commonly referred to as MALDI Imaging. In MALDI Imaging, thin biological samples are examined for their molecular composition by collecting mass spectra at multiple discrete positions within the sample. The result can be displayed in an image where each spatial point presents a mass spectrum. The information of the molecular composition of the sample can be utilized for classification of different tumor types.

The paper "Supervised Topological Data Analysis for MALDI Mass Spectrometry Imaging Applications" makes a contribution to this field by preprocessing MALDI MSI Data in a topological way for classification by applying the Reduced Persistence Transformation. The preprocessing is done in almost linear run-time and drastically reduces the amount of data. The significant signal peaks are relieved from noise within the data and are utilized for classification tasks. In this publication, two classifier are applied to deliver competitive results in remarkable short run-times. These classifiers are "random forest" and "logistic regression".

Percentage of Peaks k	Time for Random Forest	Time for Logistic Regression
100%	46 *sec*	28 *sec*
50%	33 *sec*	27 *sec*
40%	30 *sec*	28 *sec*
30%	26 *sec*	28 *sec*
20%	22 *sec*	31 *sec*
10%	18 *sec*	37 *sec*

Table 3.1: Run-time of classification tasks given different percentages of input

A primary advantage of data preprocessed through the Reduced Persistence Transformation is its sparsity, as only significant signal peaks are taken into account. This advantage can be maximized by selecting a classifier proficient in handling sparse data. The data's sparsity can be enhanced by raising the threshold for peaks to become relevant, i.e., decreasing the parameter k, describing the percentage of peaks considered as signals. The Table (3.1) illustrates the run-time of the two classifiers for different percentages k on a standard computer. It can be observed that the "random forest" improves its performance with reduced input, whereas the performance of "logistic regression" remains relatively constant. Further investigation could be dedicated to determining the optimal classifier choice.

Another noteworthy result highlighted in the paper is the capacity to effectively handle noise by carefully choosing the optimal value for the percentage denoising parameter k. Low persistence peaks can be considered as noise and can therefore be removed. The paper demonstrates the impact of this denoising through examples involving synthetic MALDI data, encompassing diverse types and levels of noise.

In this paper, my responsibilities encompassed conducting topological analysis of the data. I conceptualized and developed the Persistence Transformation, creating an algorithm for its efficient implementation. Moreover, I validated the algorithm's complexity and executed its application on the dataset. Additionally, I designed and executed the experiments related to denoising synthetic MALDI data.

In conclusion, this paper stands as a significant stride towards advancing the realm of science

and data analysis. By successfully applying the Persistence Transformation to a real-world data set, we have demonstrated not only its robustness but also its potential for rapid computational analysis. Moreover, the inherent value of this method shines through its applicability in denoising applications, solidifying its relevance and versatility in addressing complex challenges across various domains. Through these achievements, this paper adds to the ever-growing body of knowledge and offers a valuable tool for researchers and practitioners seeking innovative solutions in the intricate landscape of data analysis.

CHAPTER 4

Final Discussion and Conclusion

4.1. Interpretation and Implications

In the preceding chapters, we embarked on a comprehensive exploration of the field of Topological Data Analysis, with a particular focus on a groundbreaking innovation - the Persistence Transformation. This thesis, by establishing the stability of this transformation and showcasing its real-world applications, has firmly positioned this method as a valuable representation tool within the TDA pipeline.

The central goal of this section is to navigate the complexities inherent to the Persistence Transformation. This will be achieved by carefully assessing its merits and limitations, thus shedding light on its valuable contributions and potential constraints. This in-depth discussion will thoroughly explore the contextual advantages that the Persistence Transformation offers to the field of data analysis, alongside the unique challenges it might present. Throughout this exploration, there is a strong awareness of the dynamic nature of modern data analysis, ensuring the analysis is aligned with this ever-evolving landscape. By examining the Persistence Transformation from various angles, this section provides valuable insights tailored to both researchers and practitioners navigating the complex terrain of Topological Data Analysis.

The Persistence Transformation, serving as an extension to the Persistence Diagram, is developed to be applied as a representation tool in the TDA pipeline. It provides efficient means of transforming prepared data into a sparse set of persistence points, which encapsulate essential positional information about the data's topological features. This sparsity is valuable for condensing extensive datasets into a concise representation of relevant information. The resulting points can be visually displayed, offering a comprehensive overview of the data's topology. Importantly, by considering the positional information of features, the Persistence Transformation provides a more profound understanding of both global structures and local characteristics when compared to the Persistence Diagram.

Furthermore, this analysis can be conducted with computational efficiency, operating in nearly linear time relative to the input data or quadratic time concerning the topological features. This level of computational efficiency allows for the analysis of large datasets within short time frames. The sparsity of the output also optimizes subsequent processing tasks, such as machine learning, by utilizing suitable methods. The demonstrated stability of the Persistence Transformation ensures that the analysis remains robust even in the face of minor input perturbations, making it a reliable tool for real-world applications.

Additionally, the analysis can be conducted without the need for extensive parameter tuning, providing a comprehensive analysis. However, there exists the option to introduce a denoising parameter, effectively eliminating minor disturbances in the measured data. This single parameter within the algorithm can be fine-tuned to achieve the most accurate outcomes. Finally, the thesis introduces a semi-algorithmic approach for calculating the persistence of topological features in line with the Elder Rule, and this process can be executed swiftly. Nevertheless, it is worth noting that the direct calculation of the persistence for a single feature is not supported.

While the Persistence Transformation introduces intriguing new features to enrich the domain of TDA, it is essential to acknowledge that the Persistence Transformation does have certain limitations. Its applicability is primarily confined to real-valued functions $f : M \to \overline{\mathbb{R}}$. Although M can represent a higher-dimensional set, this inherent constraint restricts the range of suitable applications. Furthermore, the output of the Persistence Transformation exhibits higher dimensionality compared to the input data. Notably, as the dimension of the input space M increases, the dimension of the resulting output similarly escalates.

One of the most notable limitations of the Persistence Transformation is its sensitivity to the newly introduced ℓ_p metric. As demonstrated earlier, functions that are closely related concerning the supremum norm can yield significantly different results. The requirement for stability solely based on the ℓ_p metric may not be as intuitive as other metrics, which could potentially make it less applicable. In certain applications, this limitation has the potential to introduce inaccuracies into the analysis, thereby presenting challenges in specific contexts.

In spite of these limitations, the Persistence Transformation demonstrates considerable promise as a versatile extension with diverse applications. Particularly, for one-dimensional Morse function-like data, it provides rapid and precise results in relatively low dimensions, making it well-suited for subsequent analyses. An illustrative example of its beneficial application can be found in the case of MALDI Imaging, as evidenced in the paper "Supervised TDA for MALDI Mass Spectrometry Imaging Applications" (3.1). These successes underscore the potential value of the Persistence Transformation in a wide range of applications and suggest that further exploration could yield valuable insights and innovative approaches in the field of data analysis.

4.2. Summary and Future Directions

This dissertation delved into the exploration of a representation tool in the realm of Topological Data Analysis – the Persistence Transformation. It addressed its properties, stability, and practical applications, ultimately positioning it as a valuable extension to the Persistence Diagram within the TDA pipeline.

The journey of this dissertation commenced with a broad introduction to the topic, followed by a comprehensive overview of the state of the art in TDA and the general TDA pipeline. It then focused on the Persistence Diagram, highlighting its potential limitations and the need for a tool like the Persistence Transformation. The stability and other properties of this new tool were rigorously studied and evaluated. Subsequently, the transition from theory to practice was executed by showcasing the broad areas where TDA can be effectively applied. This transition culminated in the demonstration of the Persistence Transformation's remarkable accuracy and computational efficiency in MALDI Imaging. Lastly, the study contemplated the advantages and disadvantages of the Persistence Transformation, providing a balanced perspective.

The research has unveiled that the Persistence Transformation exhibits distinct advantages. It possesses stability properties, adheres to the Elder Rule, and provides richer insights compared to the Persistence Diagram. Moreover, it yields fast computational results, which are ideal for subsequent processing steps, such as machine learning. The practical implications of this method are evident, as exemplified by its successful application in real-world problems like MALDI-Imaging. These findings underscore the potential value and significance of the Persistence Transformation in a variety of applications.

In answer to the central research question, this study affirms that the Persistence Transformation is a valuable and versatile tool. It extends the capabilities of the Persistence Diagram and introduces the TDA pipeline to a method specifically designed to capture positional information. This capability makes it especially valuable in contexts where positional information is of paramount importance, such as the analysis of large Morse functions.

By introducing the Persistence Transformation, this dissertation contributes an additional, purpose-driven component to the TDA pipeline. The method is explicitly designed to capture and represent positional information, making it a valuable addition to the TDA toolkit. Its potential application extends to numerous fields where understanding the spatial relationships within data is vital.

This research goes beyond theory by providing an algorithm for implementing the Persistence Transformation. It outlines a practical path to preparing data for machine learning classification, utilizing the transformative power of this method to enhance the performance of machine learning, specifically via random forest classification.

While this work has illuminated the potential of the Persistence Transformation, several avenues for future research beckon. These include researching the influence of different metrics on the stability assessment, addressing the challenge of high-dimensional output, enabling the direct calculation of feature persistence, fine-tuning denoising parameters for even more precise results, and exploring suitable machine learning algorithms optimized for the sparse output produced by the Persistence Transformation.

In closing, this dissertation not only underscores the potential of the Persistence Transformation in TDA but also sets the stage for further research that promises to enhance its capabilities and widen its array of applications. It embodies the essence of modern data analysis - a dynamic and ever-evolving field with immense potential for future discoveries and innovations.

BIBLIOGRAPHY

Abeywardena, V. (1972). An application of principal component analysis in genetics. *Journal of genetics*, 61:27–51.

Aktas, M. E., Akbas, E., and Fatmaoui, A. E. (2019). Persistence homology of networks: methods and applications. *Applied Network Science*, 4(1):1–28.

Berwald, J. J., Gottlieb, J. M., and Munch, E. (2018). Computing Wasserstein distance for persistence diagrams on a quantum computer. *arXiv preprint arXiv:1809.06433*.

Bubenik, P. (2020). The persistence landscape and some of its properties. In *Topological Data Analysis: The Abel Symposium 2018*, pages 97–117. Springer.

Bubenik, P. and Dłotko, P. (2017). A persistence landscapes toolbox for topological statistics. *Journal of Symbolic Computation*, 78:91–114.

Bubenik, P. and Milićević, N. (2021). Homological algebra for persistence modules. *Foundations of Computational Mathematics*, 21(5):1233–1278.

Bukkuri, A., Andor, N., and Darcy, I. K. (2021). Applications of topological data analysis in oncology. *Frontiers in artificial intelligence*, 4:659037.

Cao, Y., Monod, A., Vlontzos, A., Schmidtke, L., and Kainz, B. (2023). Topological information retrieval with dilation-invariant bottleneck comparative measures. *Information and Inference: A Journal of the IMA*, 12(3):1964–1996.

Caputi, L., Pidnebesna, A., and Hlinka, J. (2021). Promises and pitfalls of topological data analysis for brain connectivity analysis. *NeuroImage*, 238:118245.

Carmody, D. R. and Sowers, R. B. (2021). Topological analysis of traffic pace via persistent homology. *Journal of Physics: Complexity*, 2(2):025007.

Cheng, L. (2020). The application of topological data analysis in practice and its effectiveness. In *E3S Web of Conferences*, volume 214, page 03034. EDP Sciences.

Chung, Y.-M., Hu, C.-S., Lo, Y.-L., and Wu, H.-T. (2021). A persistent homology approach to heart rate variability analysis with an application to sleep-wake classification. *Frontiers in physiology*, 12:637684.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2005). Stability of persistence diagrams. In *Proceedings of the twenty-first annual symposium on Computational geometry*, pages 263–271.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007). Stability of persistence diagrams. *Discret.*

Comput. Geom., 37(1):103–120.

Dłotko, P. and Gurnari, D. (2022). Euler characteristic curves and profiles: a stable shape invariant for big data problems. *arXiv preprint arXiv:2212.01666*.

Edelsbrunner, H. and Harer, J. L. (2010). *Computational topology: an introduction*. American Mathematical Society, Providence, USA.

Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002). Topological Persistence and Simplification. *Discret. Comput. Geom.*, 28(4):511–533.

Feng, J. (2021). Research on tda-effective analytical methods for modern biology. In *2021 3rd International Conference on Intelligent Medicine and Image Processing*, pages 109–115.

Gasparovic, E., Munch, E., Oudot, S., Turner, K., Wang, B., and Wang, Y. (2019). Intrinsic interleaving distance for Merge Trees. *arXiv preprint arXiv:1908.00063*.

Gidea, M. and Katz, Y. (2018). Topological data analysis of financial time series: Landscapes of crashes. *Physica A: Statistical Mechanics and its Applications*, 491:820–834.

Giuliani, A. (2017). The application of principal component analysis to drug discovery and biomedical data. *Drug discovery today*, 22(7):1069–1076.

Guo, H., Ming, Z., and Xing, B. (2023). Topological data analysis of chinese stocks' dynamic correlations under major public events. *Frontiers in Physics*, 11:1253953.

Hensel, F., Moor, M., and Rieck, B. (2021). A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, 4:681108.

Ignacio, P. S., Dunstan, C., Escobar, E., Trujillo, L., and Uminsky, D. (2019). Classification of single-lead electrocardiograms: Tda informed machine learning. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1241–1246. IEEE.

Kang, L., Xu, B., and Morozov, D. (2021). Evaluating state space discovery by persistent cohomology in the spatial representation system. *Frontiers in computational neuroscience*, 15:616748.

Kasatkina, E., Vavilova, D., and Ketova, K. (2022). Optimization of the public transport system using data analysis methods. In *2022 4th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)*, pages 174–177. IEEE.

Klaila, G., Stefanou, A., and Ranke, L. (2023a). Stability of the persistence transformation. *arXiv preprint arXiv:2310.05559*.

Klaila, G., Vutov, V., and Stefanou, A. (2023b). Supervised topological data analysis for maldi mass spectrometry imaging applications. *BMC bioinformatics*, 24(1):279.

Kozlov, D. N. (2020). A combinatorial method to compute explicit homology cycles using discrete Morse theory. *Journal of Applied and Computational Topology*, 4(1):79–100.

Kozlov, D. N. (2021). *Organized collapse: an introduction to discrete Morse theory*, volume 207. American mathematical society.

Lauric, A., Ludwig, C. G., and Malek, A. M. (2022). Topological data analysis and use of mapper for cerebral aneurysm rupture status discrimination based on 3-dimensional shape analysis. *Neurosurgery*, pages 10–1227.

Leon, J. L., Cerri, A., Reyes, E. G., and Diaz, R. G. (2013). Gait-based gender classification using persistent homology. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 18th Iberoamerican Congress, CIARP 2013, Havana, Cuba, November 20-23, 2013, Proceedings, Part II 18*, pages 366–373. Springer.

Lesnick, M. (2015). The theory of the interleaving distance on multidimensional persistence modules. *Foundations of Computational Mathematics*, 15(3):613–650.

Li, A., Bueno-Perez, R., and Fairen-Jimenez, D. (2022). Identifying porous cage subsets in the cambridge structural database using topological data analysis. *Chemical Science*, 13(45):13507–13523.

Ma, Y. and Guo, G. (2014). *Support vector machines applications*, volume 649. Springer.

Majumdar, S. and Laha, A. K. (2020). Clustering and classification of time series using topological data analysis with applications to finance. *Expert Systems with Applications*, 162:113868.

Mandal, S., Guzmán-Sáenz, A., Haiminen, N., Basu, S., and Parida, L. (2020). A topological data analysis approach on predicting phenotypes from gene expression data. In *International Conference on Algorithms for Computational Biology*, pages 178–187. Springer.

Marchese, A. and Maroulas, V. (2018). Signal classification with a point process distance on the space of persistence diagrams. *Advances in Data Analysis and Classification*, 12:657–682.

Mémoli, F. (2011). Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487.

Moon, C., Heath, J. E., and Mitchell, S. A. (2017). Statistical inference for porous materials using persistent homology. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

Mukherjee, S. (2021). Denoising with discrete Morse theory. *The Visual Computer*, 37(9-11):2883–2894.

Murayama, B., Kobayashi, M., Aoki, M., Ishibashi, S., Saito, T., Nakamura, T., Teramoto, H.,

and Taketsugu, T. (2023). Characterizing reaction route map of realistic molecular reactions based on weight rank clique filtration of persistent homology. *Journal of Chemical Theory and Computation*, 19(15):5007–5023.

Noble, W. S. (2004). 3 support vector machine machine applications in computational biology. *Kernel methods in computational biology*, page 71.

Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., and Harrington, H. A. (2017). A roadmap for the computation of persistent homology. *EPJ Data Science*, 6:1–38.

Perelman, L. and Ostfeld, A. (2011). Topological clustering for water distribution systems analysis. *Environmental Modelling & Software*, 26(7):969–972.

Phillips, J. M., Wang, B., and Zheng, Y. (2013). Geometric inference on kernel density estimates. *arXiv preprint arXiv:1307.7760*.

Pokorny, F. T. and Kragic, D. (2015). Data-driven topological motion planning with persistent cohomology. In *Robotics: Science and Systems*.

Prantzalos, K., Upadhyaya, D. P., Shafiabadi, N., Gurski, N., Fernandez-BacaVaca, G., Yoshimoto, K., Sivagnanam, S., Majumdar, A., and Sahoo, S. S. (2023). Matilda: An integrated machine learning and topological data analysis platform for brain network dynamics. *medRxiv*, pages 2023–06.

Ravishanker, N. and Chen, R. (2019). Topological data analysis (tda) for time series. *arXiv preprint arXiv:1909.10604*.

Rosen, P., Suh, A., Salgado, C., and Hajij, M. (2019). Topolines: Topological smoothing for line charts. *arXiv preprint arXiv:1906.09457*.

Rudkin, S., Rudkin, W., and Dłotko, P. (2023). On the topology of cryptocurrency markets. *International Review of Financial Analysis*, 89:102759.

Ryu, H. (2021). *Topological Data Analysis for Studying Brain Functional Connectivity*. PhD thesis, University of Georgia.

Sardiu, M. E., Gilmore, J. M., Groppe, B. D., Dutta, A., Florens, L., and Washburn, M. P. (2019). Topological scoring of protein interaction networks. *Nature communications*, 10(1):1118.

Skraba, P. and Turner, K. (2020). Wasserstein stability for persistence diagrams. *arXiv preprint arXiv:2006.16824*.

Sokerin, P., Kuznetsov, K., Makhneva, E., and Zaytsev, A. (2023). Portfolio selection via topological data analysis. *arXiv preprint arXiv:2308.07944*.

Songdechakraiwut, T., Krause, B. M., Banks, M. I., Nourski, K. V., and Van Veen, B. D. (2021). Fast topological clustering with Wasserstein distance. *arXiv preprint arXiv:2112.00101*.

Thatcher, J., Retchless, D., Thatcher, C., and Jones, K. (2021). Putting mapper on a map: cartographic visualizations of topological data analysis. *Abstracts of the ICA*, 3:1–2.

Townsend, J., Micucci, C. P., Hymel, J. H., Maroulas, V., and Vogiatzis, K. D. (2020). Representation of molecular structures with persistent homology for machine learning applications in chemistry. *Nature communications*, 11(1):3230.

Upadhyay, A., Goldfarb, B., Wang, W., and Ekenna, C. (2022). A new application of discrete Morse theory to optimizing safe motion planning paths. In *International Workshop on the Algorithmic Foundations of Robotics*, pages 18–35. Springer.

Vejdemo-Johansson, M., Pokorny, F. T., Skraba, P., and Kragic, D. (2015). Cohomological learning of periodic motion. *Applicable algebra in engineering, communication and computing*, 26(1):5–26.

Weis, C., Horn, M., Rieck, B., Cuénod, A., Egli, A., and Borgwardt, K. (2020). Topological and kernel-based microbial phenotype prediction from MALDI-TOF mass spectra. *Bioinformatics*, 36(Supplement_1):i30–i38.

Xia, K., Li, Z., and Mu, L. (2018). Multiscale persistent functions for biomolecular structure characterization. *Bulletin of mathematical biology*, 80:1–31.

APPENDIX A

PAPER "STABILITY OF PERSISTENCE TRANSFORMATION"

Gideon Klaila, Anastasios Stefanou & Lena Ranke

- **Status:** Submitted on 09.10.2023
- **Journal:** Journal of Applied and Computational Topology
- **Impact Factor:** 2.26
- **Preprint:** Published on arXiv (https://arxiv.org/pdf/2310.05559.pdf, 09.10.2023)

Stability of the persistence transformation

Gideon Klaila[1*†], Anastasios Stefanou[1†] and Lena Ranke[1†]

[1]Insitute for Algebra, Geometry, Topology and its Applications (ALTA), Department of Mathematics, University of Bremen, Bibliothekstraße 1, 28359, Bremen, Germany.

*Corresponding author(s). E-mail(s): klailag@uni-bremen.de;
Contributing authors: stefanouanastasios@gmail.com;
lranke@uni-bremen.de;
†These authors contributed equally to this work.

Abstract

In this paper, we introduce the persistence transformation, a novel methodology in Topological Data Analysis (TDA) for applications in time series data which can be obtained in various areas such as science, politics, economy, healthcare, engineering, and beyond. This approach captures the enduring presence or 'persistence' of signal peaks in time series data arising from Morse functions while preserving their positional information. Through rigorous analysis, we demonstrate that the proposed persistence transformation exhibits stability and outperforms the persistent diagram of Morse functions (with respect to filtration, e.g., the upper levelset filtration). Moreover, we present a modified version of the persistence transformation, termed the reduced persistence transformation, which retains stability while enjoying dimensionality reduction in the data. Consequently, the reduced persistence transformation yields faster computational results for subsequent tasks, such as classification, albeit at the cost of reduced overall accuracy compared to the persistence transformation. However, the reduced persistence transformation finds relevance in specific domains, e.g., MALDI-Imaging, where positional information is of greater significance than the overall signal height. Finally, we provide a conceptual outline for extending the persistence diagram to accommodate higher-dimensional input while assessing its stability under these modifications.

Keywords: Topological Data Analysis, Time Series, Persistence Homology, Persistence Diagram, Morse functions, Filtrations

1 Introduction

1.1 Motivation and related work

Topological Data Analysis (TDA) stands at the intersection of mathematics, data science, and various application domains, promising a unique perspective on understanding complex datasets (e.g., see Carlsson (2009), Edelsbrunner and Harer (2010)). TDA is a cutting-edge approach that harnesses the power of topology, a branch of mathematics that studies the shape and structure of spaces, to analyze and extract valuable insights from data. In the fast-paced evolution of technical achievements, one of the most valuable resources is data. In scientific, economical and political areas, data are being gathered in a progressively fast pace which amounts to an exponential increase of new information each year. Traditional analytical methods often fall short in capturing the underlying patterns and structures buried within these data. This is where TDA steps in with its innovative approach. By considering data as a collection of points in a high-dimensional space, TDA aims to unveil the intrinsic geometry and relationships present in these datasets.

The significance of TDA spans a wide range of fields, making it a transformative tool across diverse disciplines. In data science, TDA offers an alternative approach to dimensionality reduction, helping to projecting datasets onto representations that retain essential information (e.g., see Nicolau et al (2011), Yu and You (2021)). In biology, TDA aids in understanding complex biological systems, such as protein structures or neural connectivity, by revealing the underlying topological features that govern their behavior (e.g., Koseki et al (2023), Das et al (2023)). Moreover, in image analysis, TDA can decode intricate patterns within images, going beyond pixel-level analysis to uncover hidden structures that might represent critical information (e.g., see Ver Hoef et al (2023)). This is particularly relevant in medical imaging, where TDA's ability to highlight significant features can aid in disease diagnosis and treatment planning (e.g., Singh et al (2023)).

TDA's strength lies in its ability to capture essential characteristics of data that might be overlooked by traditional techniques. By focusing on the shape, connectivity, and arrangement of data points, TDA can identify clusters, holes, voids, and other topological features that convey crucial insights about the data's underlying structure. This capability is especially powerful when dealing with noisy, high-dimensional, or incomplete datasets, where conventional methods often struggle.

The given data types can range from a point cloud to time-series to even visual images and are highly dependent on the subject and method of obtaining said data. In this paper, we focus on spectral data which are typical in fields such as biology and in particular medicine. The most obvious choice for analyzing this type of data utilizing tools of TDA is applying a levelset filtration (e.g., see Edelsbrunner et al (2008)) which extracts important topological information of the underlying space in the form of persistent homology. The results can be represented as barcodes (e.g., see Ghrist (2008)) or persistence diagrams (e.g., see Cohen-Steiner et al (2005)) which can be interpreted individually dependent on the application. However, the levelset filtration has difficulties in differentiating symmetric spectra, sharing the same significant peaks

at different locations. In such scenarios, the generated persistence diagrams might appear identical, disrupting the analysis.

One example for this disturbance can be observed in MALDI imaging, where the importance of positional information in spectral data is linked with the persistence of signal peaks. This insight inspired the closely connected paper "Supervised topological data analysis for MALDI mass spectrometry imaging applications" by G. Klaila, V. Vutov and A. Stefanou (Klaila et al (2023)). In this paper, MALDI imaging was used to classify cancer cells utilizing machine learning and feature extraction with the persistence transformation. This tool helped in pre-processing the signal peaks and their positional information in order to improve the training of the machine learning algorithm. In this example, the position of the peaks could be related to specific molecules which in turn gave information about the type of cancer cell.

Central to our research is the profound significance of stability. While our paper indeed delves into the discrimination of persistence and peaks, a core facet also centers on substantiating the stability of our innovative methodology. In the realm of analytical approaches, stability emerges as a pivotal element, indicating the reliability of outcomes in the face of uncertainties and fluctuations. Stability of the analysis implies robust results related to small perturbations in the input. By demonstrating the stability of our method, we reinforce its credibility and applicability, rendering it a robust tool capable of withstanding the challenges of real-world data.

The first TDA stability theorem by Cohen-Steiner et al. (Cohen-Steiner et al (2005)) proved the stability of persistence diagrams for certain continuous functions. This theorem was expanded to the application of the interleaving distance on persistence modules by Chazal et al. (Chazal et al (2009)), and in succession to other applications. Examples for these are the generalizations of the interleaving distance and thus generalizations of the original stability result (see Bubenik et al (2015); Silva et al (2018)) or the stability of the Euler characteristic curves and its multi-parameter extension, the Euler characteristic profile (see Dłotko and Gurnari (2022)). In this paper, we want to contribute to this already established framework of TDA by stating a stability result for the case of the persistence transformation.

Providing this stability is not just an academic pursuit, but a practical necessity. It enhances the credibility and applicability of our approach, making it a robust tool for real-world data analysis. Our exploration into stability promises to be an essential step forward in the dynamic landscape of modern data analysis, where data complexities are met with steadfast methodologies.

1.2 Overview of our results

During the course of this paper, we explore the utilization of the persistence transformation, a novel method in topological data analysis, and delve into its potential benefits. The paper is organized into several chapters, each contributing to a comprehensive understanding of this methodology.

Section 2 provides the necessary background and lays the foundation for our analysis. We define the space of critical points, called Morse Sets, in 2.2 and establish a ℓ_p-metric to support our subsequent developments in 2.3.

In Section 3, we use this groundwork in order to introduce the persistence transformation in 3.2, showcasing its formulation through the introduction of a matching mechanism on the set of local maxima. Moreover, we place significant emphasis on proving a stability theorem in 3.3, which stands as the pivotal result of our work. To provide context, we conclude this chapter by providing a brief comparison between the persistence transformation and the conventional persistence diagram in 3.4, along with a description of a potential implementation in 3.5.

Our exploration continues in Section 4, where we unveil the reduced persistence transformation - a modified version that offers dimensionality reduction. We carefully describe its structure in 4.1 and demonstrate its stability in 4.2. Additionally, we compare its performance against the traditional persistence diagram in 4.3, shedding light on its strengths and limitations.

Intriguingly, the Section 5 offers a conceptual sketch for extending the persistence transformation to higher-dimensional input. We explore the exciting possibilities and inherent challenges that arise in this endeavor, and we discuss the stability of this extension in light of these new conditions.

As we approach the conclusion, in Section 6 we summarize the key findings presented throughout the paper. We emphasize the significance of the persistence transformation as a powerful and stable tool in topological data analysis. Additionally, we give an outlook by considering potential enhancements and novel applications for this transformative method.

Through these chapters, we aim to establish the persistence transformation as a crucial addition to the arsenal of topological data analysis techniques. We showcase its potential to unlock new insights in various domains and illuminate its stability in diverse scenarios.

2 Morse Sets

2.1 Background

In numerous modern applications, significant data are represented as images of real-valued functions (e.g., science, politics, economy, healthcare, engineering), known as time series. These fields include also domains such as medicine (e.g., Radtke et al (2016)), robotics (e.g., Atkeson and McIntyre (1986)), and climate science (e.g., Ge et al (2010)). While studying these functions, one can conclude that relevant information is often given by the critical values of the functions, i.e., the minima and maxima. For example, in the application of MALDI-Imaging (see Aichler and Walch (2015)), the local maxima of MALDI-Spectra can be associated to specific molecules. This association can help the tumor sub-typing process (see Klaila et al (2023)). On the other hand, values in between these critical points often carry less to none significant information for the analysis of the data. Expanding upon this observation, we establish an abstract space of critical points derived from real-valued functions to facilitate data analysis simplification. In the described context, the abstract space contains similar significant information in a more compact form compared to the original data. Another notable benefit of this abstraction is that the analysis of the data can be performed without any prior knowledge of the original function, thus enabling the

application of these methods to any (data-) set contained within this abstract space. We call these abstract sets *Morse sets*.

2.2 Definition of Morse sets

The term 'Morse set' draws its inspiration from its application in the context of Morse functions. In numerous scenarios, time series data lends itself to interpretation as a real-valued function that satisfies all the prerequisites of a Morse function, as detailed in Milnor (1963). However, it is essential to note that the concept of Morse sets can also extend beyond Morse functions and be defined for a broader class of real-valued functions. In this subsection, we provide a formal definition of the Morse set.

For a compact subset, $M \subseteq \overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$, we define the space $\mathscr{M} := M \times \overline{\mathbb{R}}$ and equip it with the co-lexicographic order $<_{\mathscr{M}}$, which is induced by $<_M$ as follows:

For all $(x, y), (x', y)$ and $(x', y') \in \mathscr{M}$ with $x \neq x'$ and $y \neq y'$ we have:

- $(x, y) <_{\mathscr{M}} (x', y') \Leftrightarrow y <_{\overline{\mathbb{R}}} y'$
- $(x, y) <_{\mathscr{M}} (x', y) \Leftrightarrow x <_M x'$.

Sets K of elements of this space are denoted by as either $k \in K$ or $(x, y) \in K$, where $x \in M$ and $y \in \overline{\mathbb{R}}$. The set $K \subseteq \mathscr{M}$ is referred to as a *Morse set*, if the following conditions are satisfied:

1. Injectivity: For all $x \in M, y, y' \in \overline{\mathbb{R}}$ there is: $(x, y), (x, y') \in K \Rightarrow y = y'$.
2. Disjunction: The set K is a disjoint union of two subsets, i.e., $K = K^+ \sqcup K^-$. To highlight the affiliation to one of these set, we denote the elements with the corresponding symbol, e.g., $k_i^+ \in K^+$ or $(x^-, y^-) \in K^-$.
3. Ordered: For all $i < j$, it holds that $k_i^+ >_{\mathscr{M}} k_j^+$ with $k_i^+, k_j^+ \in K^+$ and $k_i^- <_{\mathscr{M}} k_j^-$ with $k_i^-, k_j^- \in K^-$.
4. Alternation: For all $(x, y), (x', y') \in K^+$ holds: If there is no element $(x^*, y^*) \in K^+$ such that $x^* \in [x, x']$, then there is exactly one element $(\hat{x}, \hat{y}) \in K^-$ such that $(\hat{x}, \hat{y}) < (x, y), (x', y')$ and $\hat{x} \in [x, x']$, and vice versa.

   For all $(x, y), (x', y') \in K^-$ holds: If there is no element $(x^*, y^*) \in K^-$ such that $x^* \in [x, x']$, then there is exactly one element $(\hat{x}, \hat{y}) \in K^+$ such that $(\hat{x}, \hat{y}) > (x, y), (x', y')$ and $\hat{x} \in [x, x']$.
5. Critical Boundary: For all $x \in \partial M$ (i.e., the boundary of $M \subset \mathbb{R}$) there is $k = (x, y) \in K$, with $y \in \overline{\mathbb{R}}$.

Let $\mathscr{K} := \{K \subset \mathscr{M} | K \text{ Morse set}\}$ be the space of all Morse sets. One can think of a Morse set as being the set of non-degenerate critical points for the graph of a Morse function (see Milnor (1963)) $f : M \to \overline{\mathbb{R}}$ together with the boundary points, where the $(x, f(x)) \in K^+$ are the maxima, and the $(x', f(x')) \in K^-$ are the minima.

We denote the cardinality of the subsets $K^+, K^- \subset K$ with $\kappa^+ := |K^+|$ and $\kappa^- := |K^-|$. The condition (4) of 2.2 directly implies the following statement:

$$\begin{aligned} & ||K^+| - |K^-|| \leq 1 \\ \Leftrightarrow \quad & |\kappa^+ - \kappa^-| \leq 1. \end{aligned}$$

A Morse set can be directly obtained by a Morse function f by classifying the non-degenerate critical points in the corresponding subsets K^+ and K^-. We denote the Morse set K_f to indicate the relationship with f. The benefits of employing Morse sets over the original Morse functions become more apparent when considering isotopy, as Morse sets encapsulate information regarding isotopy classes of Morse functions, making them more versatile. Furthermore, when Morse sets are identical, they exhibit shared characteristics related to persistent homology.

According to Kudryavtseva (see Kudryavtseva (2009)), two Morse functions are defined to be isotopic, if

- there are diffeomorphisms h_1 and h_2 such that $f = h_2 \circ g \circ h_1$,
- h_1 preserves the numbering of critical points
- h_1 is homotopic to the identity mapping.

Furthermore, if the critical points are identical, we call the functions *Morse isotopic*. This directly implies the identity of the Morse sets, i.e., if two Morse functions f and g are Morse isotopic, there is $K_f = K_g$. The subsequent theorem demonstrates that working with Morse sets not only encodes more information but also serves as a foundation for deriving concepts such as isotopy and persistent homology for similar Morse functions.

Theorem 1. *If two Morse functions f and g are Morse isotopic with $K_f = K_g$, then the following two statements are satisfied:*

1. *The Morse functions f and g are isotopic.*
2. *The persistent homology classes of the upper levelset filtrations are identical.*

Proof. 1. For K_f, let G_f be the graph obtained by connecting all points $k \in K_f$ with straight lines. The graph G_f is isotopic to the graph G_g^* of f, since the straight lines of G_f can continuous be deformed to the graph G_f^*. Since the critical points of G_f and G_f^* are identical, the deformation will not bring any more critical points. Similarly, a graph G_g for K_g can be constructed, which is isotopic to the graph G_g^* of g. Since $K_f = K_g$, there is equality between G_f and G_g, hence there is an isotopy between G_f^* and G_g^*. Finally, Theorem 2 of Kudryavtseva (see Kudryavtseva (2009)) implies the isotopy of f and g.

2. The persistent homology class of a levelset filtration depends on the path-connected components of the levelsets. For the levelset, a new path-connected component is born, if a critical point is passed. The merging of two path-connected components occurs also during the passing of a critical point. For example. in the upper levelset filtration, a new path-connected component is born in each maxima, and two components are merging in each minima. Since the critical points of f and g are identical, i.e., $K_f = K_g$, the path-connected components of them are also identical in any step. Hence their persistent homology classes are equal, resulting in the same persistent diagram.

□

Utilizing Morse sets allows for the handling of Morse isotopy classes of Morse functions rather than dealing with each Morse function individually. This aids in

categorizing Morse functions within a broader context, ensuring that similar Morse functions exhibit comparable behavior in terms of their persistent homology.

2.3 ℓ_p-Metrics on Morse sets

This work aims at proving stability of the persistence transformation, a method working on the above defined space $\mathscr{K}$. Any stability theorem depends on the choice of metrics on the input and output spaces. One metric often utilized when working with functions is the supremum norm, i.e., the maximal distance between two functions, as can be seen in the proof of the stability theorem of the persistence diagram (see Cohen-Steiner et al (2005)). Translated to two sets $K_f, K_g \in \mathscr{K}$, this would be the maximal distance for the minimal matching between elements $k_f \in K_f$ and $k_g \in K_g$. However, the persistence diagram considers stability only in one dimension, therefore it is enough for the metric to control distance in a single dimension. In contrast, the later defined persistence transformation will consider stability not only in the height, but also in the position of the relevant peaks in accordance to the elder rule (see Edelsbrunner and Harer (2010)). To satisfy these conditions, we define an ℓ_p metric on $\mathscr{K}$ respecting the total order $<_{\mathscr{M}}$ of the elements and controlling the height as well as the position of similar significant peaks in the input. For this metric, these peaks are matched to each other, resulting in a measurable distance for Morse sets.

Define the matching m^* between the Morse sets K and $\hat{K}$ as follows:

- For all $i = 1, \dots, \min\{\kappa^+, \hat{\kappa}^+\}$ there is $(k_i^+, \hat{k}_i^+) \in m^*$.
- For all $j = 1, \dots, \min\{\kappa^-, \hat{\kappa}^-\}$ there is $(k_j^-, \hat{k}_j^-) \in m^*$.
- For all $i = \min\{\kappa^+, \hat{\kappa}^+\} + 1, \dots, \max\{\kappa^+, \hat{\kappa}^+\}$ there is $(0_{\mathscr{M}}, \hat{k}_i^+) \in m^*$ (with $0_{\mathscr{M}} = (0, 0)$).
- For all $j = \min\{\kappa^-, \hat{\kappa}^-\} + 1, \dots, \max\{\kappa^-, \hat{\kappa}^-\}$ there is $(0_{\mathscr{M}}, \hat{k}_j^-) \in m^*$.

With this matching m^*, an ℓ_p metric on $\mathscr{K}$ can be established.

Definition 1. For all $K, \hat{K} \in \mathscr{K}$ with the matching m^*, the distance between K and $\hat{K}$ is defined to be:

$$\begin{aligned} d_{\mathscr{K},p}(K, \hat{K}) :&= \sqrt[p]{\sum_{(k,\hat{k}) \in m^*} ||k - \hat{k}||_\infty^p} \\ &= \sqrt[p]{\sum_{((x,y),(\hat{x},\hat{y})) \in m^*} ||(x, y) - (\hat{x}, \hat{y})||_\infty^p}. \end{aligned}$$

Theorem 2. *The metric $d_{\mathscr{K},p}(K, \hat{K})$ is well-defined.*

Proof. The metric $d_{\mathscr{K},p}(K, \hat{K})$ is well-defined since the following four conditions are satisfied:

- Non-negativity. For all $K, \hat{K} \in \mathscr{K}$, there is:

$$d_{\mathscr{K},p}(K, K') = \sqrt[p]{\sum_{(k,\hat{k}) \in m^*} ||k - \hat{k}||_\infty^p}$$

$$\geq \sqrt[p]{\sum_{(k,\hat{k})\in m^*} 0}$$
$$= \sqrt[p]{0}$$
$$= 0,$$

since $||\cdot,\cdot||_\infty$ is a metric and therefore non-negative.

- Definiteness. For all $K, \hat{K} \in \mathscr{K}$ there is:

$$K = \hat{K} \Rightarrow d_{\mathscr{K},p}(K, \hat{K}) = 0,$$

since m^* is the identity matching on $K = \hat{K}$. On the other hand, assume that

$$d_{\mathscr{K},p}(K, \hat{K}) = 0.$$

Then there is a matching m^*, such that for all $i = 1, \ldots, \min\{\kappa^+, \hat{\kappa}^+\}$ there is $(k_i^+, \hat{k_i}^+) \in m^*$ and for all $j = 1, \ldots, \min\{\kappa^-, \hat{\kappa}^-\}$ there is $(k_j^-, \hat{k_j}^-) \in m^*$. The zero distance between K and $\hat{K}$ implies equality between all these elements, i.e., $k_i^+ = \hat{k_i}^+$ and $k_j^- = \hat{k_j}^-$. Assume now without loss of generality that $\kappa^+ \geq \hat{\kappa}^+$ and $\kappa^- \geq \hat{\kappa}^-$. It holds, that for all $i = \hat{\kappa}^+ + 1, \ldots, \kappa^+$ the elements $k_i^+ \in K^+$ are matched to $0_{\mathscr{M}}$, and similar for all $j = \hat{\kappa}^- + 1, \ldots, \kappa^-$ the elements $k_j^- \in K^-$ are matched to $0_{\mathscr{M}}$. However, since the distance between K and $\hat{K}$ is zero, the elements must be equal to $0_{\mathscr{M}}$. Uniqueness (2.2, (1)) of the sets implies the existence of at most one such element $0_{\mathscr{M}} \in K$. Consequently, there must be equality between all elements $k \in K$ and $\hat{k} \in \hat{K}$, except for at most one element $0_{\mathscr{M}} \in K$. Without loss of generality let $0_{\mathscr{M}} \in K^+$. This element cannot be in the closure of M, because the critical boundary (2.2, (5)) of $\hat{K}$ implies the existence of $\hat{k} \in \hat{K}$ which is in the boundary. Equality of the elements then states the existence of an corresponding element $k \in K$ in the boundary, being different than $0_{\mathscr{M}}$ and hence contradicting the uniqueness (2.2, (1)) of the boundary element. If the element $0_{\mathscr{M}}$ is not in the boundary of M, the alternation of K (2.2, (4)) implies the existence of neighboring elements $k = (x, y), k' = (x', y') \in K^-$ such that $(0_M, 0_{\mathbb{R}}) = 0_{\mathscr{M}} \in K^+$ is a unique element in K^+ with $0_M \in [x, x']$. Per assumption, the elements of K and $\hat{K}$ are the same except for $0_{\mathscr{M}}$, so there are elements $\hat{k}, \hat{k}' \in \hat{K}^-$ such that $k = \hat{k}$ and $k' = \hat{k}'$. The alternation of $\hat{K}$ (2.2, (4)) implies the existence of at least one element $\hat{k}^* = (\hat{x}^*, \hat{y}^*) \in \hat{K}^+$ such that $\hat{x}^* \in [x, x']$. By equality of the elements, there must be an element $k^* = (x^*, y^*) \in K^+$ such that $k^* \neq 0_{\mathscr{M}}$ and $x^* \in [x, x']$, opposing the uniqueness of $0_{\mathscr{M}}$ in this interval. This contradicts the assumption, that there is an element $0_{\mathscr{M}} \in K$ with $0_{\mathscr{M}} \notin \hat{K}$, hence all the elements must be the same, proofing:

$$d_{\mathscr{K},p}(K, \hat{K}) = 0 \Rightarrow K = \hat{K}.$$

- Symmetry. For all $K, \hat{K} \in \mathscr{K}$ there is:

$$\begin{aligned} d_{\mathscr{K},p}(K,\hat{K})^p &= \sum_{(k,\hat{k})\in m^*} ||k-\hat{k}||_\infty^p \\ &= \sum_{(k,\hat{k})\in m^*} ||\hat{k}-k||_\infty^p \\ &= \sum_{(\hat{k},k)\in m^*} ||\hat{k}-k||_\infty^p \\ &= d_{\mathscr{K},p}(\hat{K},K)^p. \end{aligned}$$

- Triangle inequality. For all $K, \hat{K}, \tilde{K} \in \mathscr{K}$ assume without loss of generality $k^+ = \hat{k}^+ = \tilde{k}^+$ and $k^- = \hat{k}^- = \tilde{k}^-$. If not, add elements $0_{\mathscr{M}}$ to the sets with fewer elements. Then:

$$\begin{aligned} & d_{\mathscr{K},p}(K,\tilde{K})^p \\ =& \sum_{(k,\tilde{k})\in m^*_{K,\tilde{K}}} ||k-\tilde{k}||_\infty^p \\ =& \sum_{i=1}^{\kappa^+} ||k_i^+ - \tilde{k}_i^+||_\infty^p + \sum_{j=1}^{\kappa^-} ||k_j^- - \tilde{k}_j^-||_\infty^p \\ =& \sum_{i=1}^{\kappa^+} ||k_i^+ - \hat{k}_i^+ + \hat{k}_i^+ - \tilde{k}_i^+||_\infty^p + \sum_{j=1}^{\kappa^-} ||k_j^- - \hat{k}_i^- + \hat{k}_i^- - \tilde{k}_j^-||_\infty^p \\ \leq& \sum_{i=1}^{\kappa^+} ||k_i^+ - \hat{k}_i^+||_\infty^p + ||\hat{k}_i^+ - \tilde{k}_i^+||_\infty^p + \sum_{j=1}^{\kappa^-} ||k_j^- - \hat{k}_i^-||_\infty^p + ||\hat{k}_i^- - \tilde{k}_j^-||_\infty^p \\ =& \sum_{(k,\hat{k})\in m^*_{K,\hat{K}}} ||k-\hat{k}||_\infty^p + \sum_{(\hat{k},\tilde{k})\in m^*_{\hat{K},\tilde{K}}} ||\hat{k}-\tilde{k}||_\infty^p \\ =& d_{\mathscr{K},p}(K,\hat{K})^p + d_{\mathscr{K},p}(\hat{K},\tilde{K})^p. \end{aligned}$$

□

3 Persistence Transformation

One main objective of topological data analysis is to track the persistence of topological features in the data. Each feature gets assigned a birth value, i.e., the value where the feature arises, and a death value, i.e., the value where the feature merges to other features. In the persistence transformation, the merging process is according to the elder rule (see Edelsbrunner and Harer (2010)). The persistence of a feature is the lifespan of it, in other words, the difference of birth and death.

3.1 Matching

In this paper, we consider features to be the peaks of a Morse function, or in the notation from before, the elements $k^+ = (x^+, y^+) \in K^+$, i.e., the local maxima. The birth of these features is given natural by their height, i.e., by y^+. On the other hand, the merging of two features always happens at a local minimum, or the $(x^-, y^-) = k^- \in K^-$. For each feature, there is a unique $k^- \in K^-$, where it dies, and the death value is given by y^-. To track the death of the feature, we match to each feature $k^+ \in K^+$ the unique element of merging $k^- \in K^-$. The largest feature cannot be merged with another feature according to the elder rule (see Edelsbrunner and Harer (2010)), therefore its death is defined to be $-\infty$. The matching is denoted by $\mu : K^+ \to K^- \cup M \times \{-\infty\}$ and is defined sequential and individual on each path-connected component. The results can later be joined for the persistence transformation.

- For all $i = \kappa^+, \dots, 2$:

$$\begin{aligned}\mu(k_i) := \sup\{(x^-, y^-) = k^- \in K^- | k^- < k_i \\ \wedge \exists k_i \leq k^+ = (x^+, y^+) \in K^+ : x^- \in [\min\{x_i, x^+\}, \max\{x_i, x^+\}] \\ \wedge \forall j = i+1, \dots, \kappa^+ : k^- \neq \mu(k_j)\}.\end{aligned}$$

- For $i = 1$:

$$\mu(k_1) := (x_1, -\infty).$$

This functional determination approach presents an innovative method for calculating persistence's in accordance with the elder rule by utilizing a semi-algorithmic procedure, which departs from the conventional complex algorithms typically employed. An important aspect of this is the simplification of the process of obtaining persistence's, making it more accessible and efficient.

Theorem 3. *The matching μ is injective and is well-defined.*

Proof. Via natural induction over κ^+.

Induction start for $\kappa^+ = 2$, i.e. $K^+ = \{k_1, k_2\}$: The Alternation (2.2, (4)) implies the existence of an element $k^- \in K^-$, such that $k^- < k_2$ and $x^- \in [\min\{x_1, x_2\}, \max\{x_1, x_2\}]$. Hence this element is suitable for $\mu(k_2)$. The element k_1 is matched to $\mu(k_1) = (x_1, -\infty)$, resulting in an injective matching.

For the induction step assume that we can always find an injective matching for $\kappa^+ = n$. For $\kappa^+ = n + 1$ now holds: The Alternation (2.2, (4)) implies with the same reasoning of the induction start the existence of an suitable element k^-, such that $\mu(k_{\kappa^+}) = k^-$. This element is by (2.2, (4)) a unique element of K^- for a suitable $k_i \in K$ with $x^- \in [\min\{x_{\kappa^+}, x_i\}, \max\{x_{\kappa^+}, x_i\}]$, such that there are no other elements $k^+ \in K^+$ in that interval. Therefore, the points k^-, k_{κ^+} can be removed from the set K, resulting in a set $\hat{K}$ with $\hat{\kappa}^+ = n$, which satisfies all conditions of an ordered critical set (2.2). The induction assumption applied to $\hat{K}$ completes the proof. □

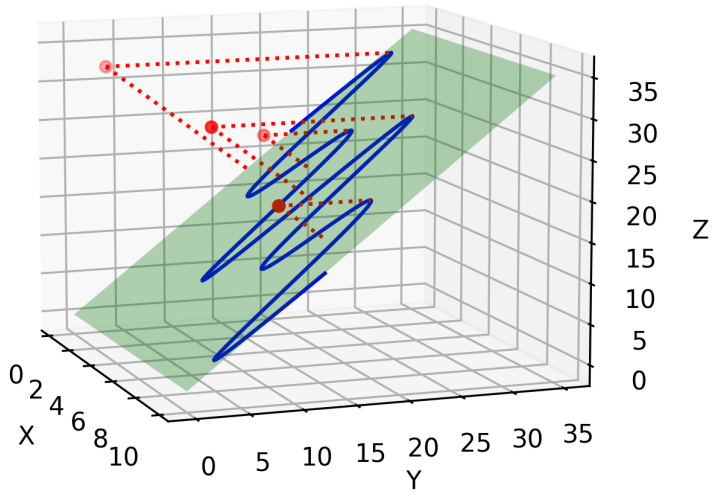


Fig. 1 Example of the persistence transformation. The blue line represents all the trivial features, which are vanishing since they are on the diagonal plane. The red dots represent the relevant features. The distance of the dots to the diagonal plane indicates their persistence.

3.2 Definition of the Persistence Transformation

With the defined matching, the persistence transformation can be defined. It is constructed in such a way, that it tracks the position, the birth and the death of each feature $k \in K^+$.

Definition 2. The **persistence transformation** is a map

$$\begin{aligned} T : \mathscr{K} &\to \mathscr{M} \times \overline{\mathbb{R}} \\ K &\mapsto T_K, \end{aligned}$$

where each element $k = (x, y) \in K^+$ with $\mu(k) = \overline{k} = (\overline{x}, \overline{y})$ is mapped to $t_K := (x, y, \overline{y})$. The elements $k^- = (x^-, y^-) \in K^-$ are mapped to $t_{K^-} = (x^-, y^-, y^-)$ on a diagonal plane.

A graphical representation of the persistence transformation can be seen in Fig. 1. The first coordinate of each element $t_K \in T_K$ corresponds to the position $x \in M$ of a feature. The second coordinate denotes the birth value $y \in \overline{\mathbb{R}}$, while the last coordinate describes the death value, i.e., $\overline{y} \in \overline{\mathbb{R}}$. This results in trivial features for all elements $k^- \in K^-$, since their death value equals their birth value. They are located on the diagonal plane in Fig. 1, which can be compared to the diagonal in the persistence diagram.

3.3 Stability

A necessary characteristic of a useful method in topological data analysis is the stability of the method. The stability is a measure of how changes in the input influence

the output of the analysis and indicates in this sense, how robust the applied method captures the topological information of the data. A stable method ensures that the variation in the output is bound by the variation in the input (see Cohen-Steiner et al (2005)). The subsequent section provides evidence for the stability of the persistence transformation. For this, the p-Wasserstein distance is utilized a the metric on the output of the persistence transformation. This distance is defined as follows (see Mémoli (2011); Berwald et al (2018)):

$$d_{W_p}(A,B)^p = \min_{\text{matchings } m:A\times B} \sum_{(a,b)\in m} ||a-b||_\infty^p.$$

As p approaches infinity, this metric converges to the bottleneck distance, the metric employed in assessing the stability of persistence diagrams (refer to Cohen-Steiner et al (2005)). The bottleneck distance exhibits notable computational efficiency and robustness to outliers when compared to finite p-Wasserstein distances. Furthermore, it offers a more intuitive interpretation by quantifying the maximum separation between elements. Importantly, by establishing the theorem's validity for arbitrary values of p, we ensure that the obtained results are equally applicable to the bottleneck distance.

Theorem 4. *The persistence transformation is stable using the p-Wasserstein distance on* $\mathscr{M} \times \overline{\mathbb{R}}$*:*

$$d_{W_p}(T_K, T_L) \leq d_{\mathscr{K},p}(K,L).$$

Proof. Let $P = T_K \times T_L$ be the set of all matchings for $t_K = (x,y,\hat{y}) \in T_K$ and $t_L = (x',y',\overline{y'}) \in T_L$, given $k = (x,y) \in K$ and $l = (x',y') \in L$. Without loss of generality let $\kappa^+ := \max\{\kappa_K^+, \kappa_L^+\}$ and $\kappa^- = \max\{\kappa_K^-, \kappa_L^-\}$. The stability of the persistence transformation is given by the following inequality:

$$\begin{aligned}
& d_{W_p}(T_K,T_L)^p && (1)\\
=& \min_{m\in P} \sum_{(t_K,t_L)\in m} ||t_K - t_L||_\infty^p && (2)\\
=& \min_{m\in P} \sum_{(t_K,t_L)\in m} ||(x,y,\overline{y}) - (x',y',\overline{y'})||_\infty^p && (3)\\
=& \min_{m\in P} \sum_{(t_K,t_L)\in m} \max\{||(x,y)-(x',y')||_\infty^p, ||\overline{y}-\overline{y'}||_\infty^p\} && (4)\\
\leq& \min_{m\in P} \sum_{(t_K,t_L)\in m} \max\{||(x,y)-(x',y')||_\infty^p, ||(\overline{x},\overline{y})-(\overline{x'},\overline{y'})||_\infty^p\} && (5)\\
\leq& \sum_{(k,l)\in m^*} \max\{||(x,y)-(x',y')||_\infty^p, ||(\overline{x},\overline{y})-(\overline{x'},\overline{y'})||_\infty^p\} && (6)\\
\overset{3}{\leq}& \sum_{i=1}^{\kappa^+} ||k_i^+ - l_i^+||_\infty^p + \sum_{j=1}^{\kappa^-} ||k_j^- - l_j^-||_\infty^p && (7)
\end{aligned}$$

$$= \sum_{(k,l)\in m^*} ||k-l||_\infty^p \tag{8}$$

$$= d_{\mathscr{K},p}(K,L)^p. \tag{9}$$

The inequality of (7) holds due the injectivity (3) of the matching function. This implies, that each element $k^- \in K^-$ and each element $l^- \in L^-$ are matched at most once. □

3.4 Comparison

We conclude this section with a comprehensive comparison between the introduced persistence transformation and the persistence diagram. The persistence diagram captures the information of the persistence in a compact and comprehensive way by displaying the homology classes of specific filtration complexes, e.g., the sub levelset filtration or the upper levelset filtration of a real-valued function. For more information about the standard persistence diagram we refer the reader to Cohen-Steiner et al (2005); Edelsbrunner et al (2002). The method has demonstrated successful applications in various occasions (e.g., Edelsbrunner et al (2002); Otter et al (2017); Nicolau et al (2011)). Nonetheless, in some settings, e.g., MALDI-Images (Klaila et al (2023)), it exhibits a significant drawback: while tracking the persistence of the signal peaks, the persistence diagram forfeits the information of the position of the signal. For instance, the standard persistence diagram of the upper levelset filtration cannot distinguish symmetric inputs, see Figure 2.

In contrast to the persistence diagram of the upper levelset filtration, the persistence transformation is specifically designed to overcome these limitations of the persistence diagram by tracking the position of significant signal peaks as well as their persistence. Even more, the persistence transformation captures all the information the persistence diagram of the upper levelset filtration captures. This implies that it is a strictly stronger method of capturing the necessary information for distinguishing data. However, this increase in performance comes with the cost of higher dimensionality. While the results of the persistence diagram are subsets of $\overline{\mathbb{R}}^2$, i.e., information about the birth and the death, the results of the persistence transformation are subsets of $M \times \overline{\mathbb{R}}^2$. This poses a significant challenge to the desire of improving computational efficiency for analysis. A decision must be made regarding whether to prioritize fast computation or more comprehensive information. In the upcoming section, we introduce a modification to the persistence transformation that decreases its dimensionality, albeit at the expense of some information.

3.5 Implementation

In this section, we present the pseudo-code for a potential implementation of the persistence transformation. Given the critical points K as input, the algorithm efficiently computes the output T_K with quadratic complexity. The resulting T_K can then be further processed in linear complexity to represent the reduced persistence transformation of the next section or the persistence diagram. Additionally, the output is sorted based

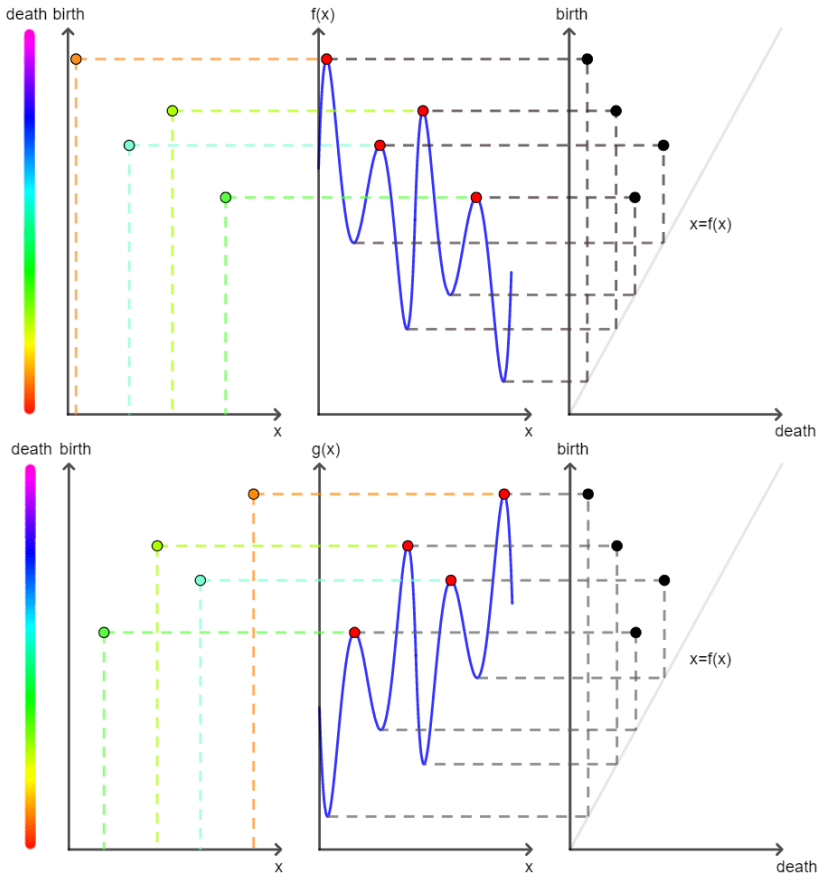


Fig. 2 Problem of the persistence diagram: the middle column displays two symmetric functions $f, g : \mathbb{R} \to \mathbb{R}$. The first column displays the persistence transformation of these functions, with the positional information on the x-axis and the birth values on the y-axis. The death values are encoded in the color scheme. The persistence transformations of the functions can be distinguished. The rightmost column shows the persistence diagram of the upper levelset filtration with the birth value on the y-axis and the death value on the x-axis. The two diagrams are identical. (Graphic: Klaila et al (2023))

on persistence values. By considering low persistence features as noise, it becomes feasible to denoise the output by removing elements below a specific threshold.

For a similar algorithm, its proof and its practical application, refer to the work in (Klaila et al (2023)).

Algorithm 1 Recursion start

Input: $K^+ = \{k_0^+, \dots, k_n^+\}; K^- = \{k_0^-, \dots, k_{n\pm1}^-\}$
Return: $T_K = \{(x_i, y_i, \overline{y_i}) | i = 0, \dots, n\}$
$T_K \leftarrow \varnothing$
$k' = (x', y') \leftarrow K^+.\text{pop}(0)$
$T_K \leftarrow (x', y', -\infty)$
setOne, setTwo $\leftarrow \varnothing$
minimum $\leftarrow \infty$
maximum $\leftarrow -\infty$
for all $k = (x, y) \in K^+$ **do**
  if $x < x'$ **then**
    setOne $\leftarrow k$
    if $x <$ minimum **then**
      minimum $\leftarrow x$
  else
    setTwo $\leftarrow k$
    if $x >$ maximum **then**
      maximum $\leftarrow x$
RecursionStep(minimum, x', setOne, K^-.copy(), T_K)
RecursionStep(maximum, x', setTwo, K^-.copy(), T_K)
return T_K

4 Reduced Persistence Transformation

4.1 Motivation and Definition

As stated in 3.4, the advantages of the persistence transformation are accompanied by an increase in dimensionality. However, in certain applications, dealing with large amounts of data, the subsequent higher dimensionality in the analyzed data results in prolonged computational time, despite not requiring the full extent of the more comprehensive information gathered. Regardless, these application could also benefit from the positional information provided by the persistence transformation. In response to this concern, we develop an adapted version of the persistence transformation known as the reduced persistence transformation. This modified method aims at reducing the dimensionality of the output data while maintaining the positional information of signal peaks. To accomplish the intended adaptation, we opt for storing the persistence of the features, given by the difference of birth and death, instead of storing these values separately.

Definition 3. Building upon the previous definitions of $\mathscr{M}$ (2.2) and the matching μ (3.1), we define the **reduced persistence transformation** as a map

$$\begin{aligned} \tilde{T} : \mathscr{K} &\to \mathscr{M} \\ K^+ &\mapsto \tilde{T}_K \end{aligned}$$

Algorithm 2 Recursion Step

Input: start, end, K^+, K^-, T_K
for all $(x, y) = k \in K^+$ **do**
  if $x \notin [\text{start}, \text{end}]$ **then**
    $K^+ \leftarrow K^+ \backslash k$
if $|K^+| = 0$ **then**
  return
$k' = (x', y') \leftarrow K^+.\text{pop}(0)$
RecursionStep(start, x', K^+.copy(), K^-.copy(), T_K)
for all $(x, y) \in K^-$ **do**
  if $x \notin (x, \text{end})$ **then**
    $K^- \leftarrow K^- \backslash k$
$k^- = (x^-, y^-) \leftarrow K^-.\text{pop}(0)$
$T_K \leftarrow (x', y', y^-)$
RecursionStep(x^-, x', K^+.copy(), K^-.copy(), T_K)
RecursionStep(x^-, end, K^+.copy(), K^-.copy(), T_K)

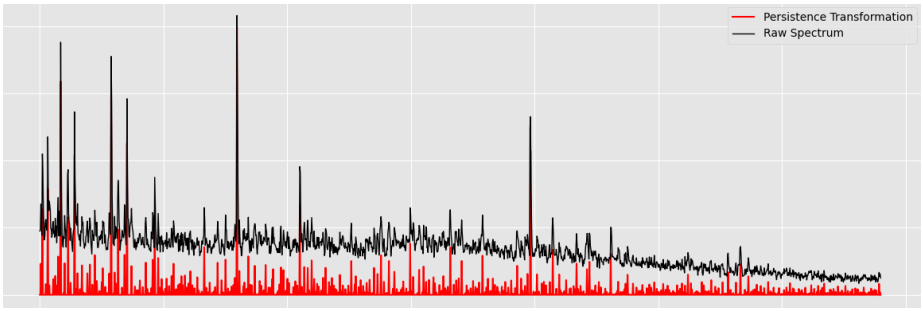


Fig. 3 Example of the reduced persistence transformation. The original spectra is displayed in black, while the values of the reduced persistence transformation are illustrating in red the persistence of each feature.

where each element $k = (x, y) \in K^+$ with $\mu(k) = \overline{k} = (\overline{x}, \overline{y})$ is mapped to $\tilde{t}_K := (x, y - \overline{y})$.

An example of the reduced persistence transformation can be seen in 3

4.2 Stability

Similar to the stability of the persistence transformation in 3.3, the stability theorem can also be applied to the reduced persistence transformation.

Theorem 5. *The reduced persistence transformation satisfies the stability condition using the p-Wasserstein distance in $\mathscr{M}$:*

$$d_{W_p}(\tilde{T}_K, \tilde{T}_L) \leq d_{\mathscr{K},p}(K, L).$$

Proof. Let $P = \tilde{T}_K \times \tilde{T}_L$ be the set of all matchings for $\tilde{t}_K = (x, y - \hat{y}) \in \tilde{T}_K$ and $\tilde{t}_L = (x', y' - \overline{y'}) \in \tilde{T}_L$, given $k = (x, y) \in K$ and $l = (x', y') \in L$. Without loss of generality, let $\kappa^+ := \max\{\kappa_K^+, \kappa_L^+\}$ and $\kappa^- := \max\{\kappa_K^-, \kappa_L^-\}$. The stability of the reduced persistence transformation is given by the following inequality:

$$
\begin{aligned}
& d_{W_p}(\tilde{T}_K, \tilde{T}_L)^p && (10)\\
=& \min_{m \in P} \sum_{(\tilde{t}_K, \tilde{t}_L) \in m} ||\tilde{t}_K - \tilde{t}_L||_\infty^p && (11)\\
=& \min_{m \in P} \sum_{(\tilde{t}_K, \tilde{t}_L) \in m} ||(x, y - \overline{y}) - (x', y' - \overline{y'})||_\infty^p && (12)\\
\leq& \min_{m \in P} \sum_{(\tilde{t}_K, \tilde{t}_L) \in m} ||(x, y) - (x', y')||_\infty^p + ||\overline{y} - \overline{y'}||_\infty^P && (13)\\
\leq& \min_{m \in P} \sum_{(\tilde{t}_K, \tilde{t}_L) \in m} ||(x, y) - (x', y')||_\infty^p + ||(\overline{x}, \overline{y}) - (\overline{x'}, \overline{y'})||_\infty^P && (14)\\
\leq& \sum_{(k,l) \in m^*} ||(x, y) - (x', y')||_\infty^p + ||(\overline{x}, \overline{y}) - (\overline{x'}, \overline{y'})||_\infty^P && (15)\\
\overset{3}{\leq}& \sum_{i=1}^{\kappa^+} ||k_i^+ - l_i^+||_\infty^p + \sum_{j=1}^{\kappa^-} ||k_j^- - l_j^-||_\infty^p && (16)\\
=& \sum_{(k,\hat{k}) \in m^*} ||k - l||_\infty^p && (17)\\
=& d_{\mathscr{K},p}(K, L)^p. && (18)
\end{aligned}
$$

The inequality of line (16) is given by the injectivity (3) of the matching function, which ensures the unique occurrence of each element $k^- \in K^-$ and $l^- \in L-$. □

4.3 Comparison

As mentioned in 3.4, while the persistence transformation provides more comprehensive information than the persistence diagram of the upper levelset, it comes at the cost of increased dimensionality. Conversely, the reduced persistence transformation reduces this dimensionality but loses some information in the process. Consequently, it is crucial to conduct a comparison between the reduced persistence transformation and the persistence diagram. After careful examination, it becomes evident that neither of the mentioned methods outperforms the other, and they share the same dimensionality in the output, i.e., $\mathbb{R}^2$. Giving them a total order is not feasible, as there exist scenarios where one outperforms the other, and vice versa (see Fig. (4). The key distinction of the persistence diagram lies in the total height of birth and death values of a feature, whereas the reduced persistence transformation incorporates the positional information alongside its relative height, i.e., its persistence. This implies that in any application where positional information is crucial, the reduced persistence transformation is a more suitable choice as an analysis method. An instance of such

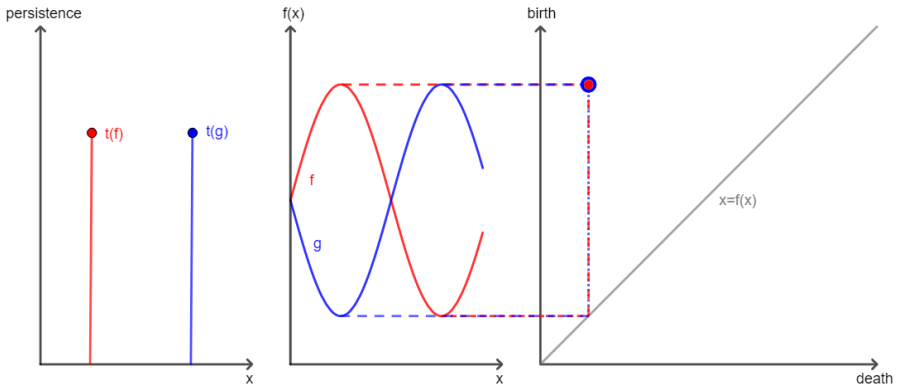


Fig. 4 Example of the reduced persistence transformation: In the middle, two symmetric functions $f, g : \mathbb{R} \to \mathbb{R}$ are displayed. The left-hand side illustrates the reduced persistence transformation of this graphs, with the positional information on the x-axis and the persistence information on the y-axis. The graphs can be distinguished. The right-hand side depicts the persistence diagram of the upper levelset filtration, with the birth value on the y-axis and the death value on the x-axis. The diagrams are identical. (Graphic: Klaila et al (2023))

an application can be observed in Klaila et al (2023), where the reduced persistence transformation is employed to enhance the accuracy of tumor tissue classification. In novel applications, it is imperative to assess the significance of positional relevance to determine the potential benefits of employing the modified methodology in comparison to the original persistence transformation or the persistence diagram as each approach possesses its distinct area of efficiency.

5 Extension to higher dimensions

Until this point, we have made the assumption that $M \subseteq \overline{\mathbb{R}}$; however, it is feasible to relax this condition. In this section, we present a concise overview of the approach to extend the persistent transformation to any arbitrary higher-dimensional compact metric space M, equipped with a total order and a notion of path connectivity. Additionally, M must be a subset of a space including a neutral element, which is essential for the defined metric on $\mathscr{M}$ (1). This requirements can be satisfied by $M \subseteq \overline{\mathbb{R}}^n$, if we equip $\overline{\mathbb{R}}^n$ with a total order, e.g., the lexicographic order.

One crucial adaptations for the increase of dimension is the definition of the ordered critical set. While in $\overline{\mathbb{R}}$ each path $[x, x']$ is unique for all elements $x, x' \in \overline{\mathbb{R}}$, there may exist multiple paths for elements $x, x' \in \overline{\mathbb{R}}^n$. This property breaks the uniqueness of elements $\hat{k} = (\hat{x}, \hat{y}) \in K$ such that $\hat{x} \in [x, x']$, resulting in cases where no minimal element $k \in K^-$ lies on the path.

Another issue that needs to be addressed is the presence of areas with the same height, where all elements within that area are critical compared to any element outside that area. Such situations could lead to an imbalance in the number of elements in the sets K^+ and K^-. To address this, we can utilize the total order of M to select only one element from this area, thus ensuring the satisfaction of the balance equation (1).

Finally, we also need to consider the adaptation of saddle points, which are degenerate critical points in Morse theory (see Milnor (1963)). Although they may not possess a unique gradient direction, they can still act as the highest valley that needs to be traversed to reach the next peak, or in other words, they could be crucial for defining the persistence of peaks. One possibility of handling them is to consider these saddle points as maxima and minima, including them in the sets K^+ and K^-, respectively. In this approach, elements $k \in K$ may have a multiplicity to account for their presence in the analysis.

After implementing all adaptations of the space of critical points, the definitions 1, 3.1, 2 and 3 will hold, resulting in a stable persistence transformation on the higher dimensional set M. The proofs of the theorems are presented in a way that they are independent of the dimensionality of M. However, for the sake of simplicity and clarity, the specific details of the extension to higher dimensions are omitted in this paper.

It is important to note that increasing the dimension of M will also lead to an increase in the dimension of the output space. As previously mentioned, the reduced persistence transformation and the persistence transformation produce an output of dimensions $M \times \overline{\mathbb{R}}$ and $M \times \overline{\mathbb{R}}^2$, respectively. For cases where $M \subseteq \overline{\mathbb{R}}$, the output dimension of the reduced persistence transformation can match the output dimension of the persistence diagram, i.e., $\overline{\mathbb{R}}^2$. However, for any $M \subseteq \overline{\mathbb{R}}^n$, the output dimension of the reduced persistence transformation will have an increased in dimension of $n-1$, One should consider the potential increase in computational time for further processing steps and carefully evaluate the significance of the positional information obtained from the increased output dimensions.

6 Summary and Outlook

In conclusion, this paper has successfully achieved its research objectives by constructing a stable method for analyzing time series data arising from Morse functions in a topological manner while preserving the positional information of signal peaks. We have defined the persistence transformation, established its stability, and demonstrated that it complements nicely the existing machinery of persistent homology on levelset filtrations. Furthermore, we have introduced the reduced persistence transformation as a valuable side result, effectively tracking positional information with reduced dimensionality.

The implications of this research are far-reaching, as the proposed method can be applied across various fields where data is represented as the image of a real-valued function. Its potential to deliver promising results, particularly in domains where positional information is crucial, underscores its significance.

Our hypothesis regarding the successful application of the reduced persistence transformation to real-world problems has been confirmed in a previous study about MALDI-MSI data in (Klaila et al (2023)).

However, we acknowledge certain limitations in the persistence transformation, particularly in terms of its computational complexity. As the dimension of the dataset M increase, the output dimension of the persistence transformation also grows, resulting in longer computational times for subsequent processing tasks.

Looking ahead, the outlook for this research is promising. The complexity of calculating the persistence transformation is quadratic to the number of critical points of the input. Furthermore, by treating low persistence features as noise, we can adapt the algorithm to filter out such peaks with a single hyper-parameter, leading to a denoising functionality of the persistence transformation and a more concise output.

In addition to the outlook presented earlier, exploring the characteristics of critical points offers promising avenues for further improving the algorithms and matchings. Applying Morse theory (Milnor (1963)) to analyze critical points could provide valuable insights and refinements to our approach. Additionally, leveraging Prof. Kozlov's work on optimizing matchings (see Kozlov (2020)) could lead to significant enhancements in the overall concept.

By delving deeper into the nature of critical points and their interplay with the proposed methods, we have the potential to unlock novel techniques and strategies to advance topological data analysis. Such investigations will undoubtedly contribute to the robustness and versatility of our approach and pave the way for even more impactful applications in various domains.

Lastly, the adaptability of this work to specific applications allows for customized implementations that precisely address the unique demands of various domains. As we continue to refine and extend this method, we are confident in its ability to contribute significantly to advancing topological data analysis in diverse real-world scenarios.

Acknowledgements

The authors extend their sincere gratitude to Prof. Dmitry Feichtner-Kozlov at the University of Bremen for supervising GK and LR. Also, we sincerely thank Lukas Mentz for his constructive criticism of the manuscript. The financial support by the German Research Foundation via the RTG 2224, titled "π^3: Parameter Identification - Analysis, Algorithms, Implementations" is gratefully acknowledged.

Authors' information

- Gideon Klaila, klailag@uni-bremen.de, ORCID: 0009-0002-2861-2095
- Anastasios Stefanou, stefanou@uni-bremen.de, ORCID: 0000-0002-5408-9317
- Lena Ranke, lranke@uni-bremen.de, ORCID: 0009-0003-4258-1608

Authors' contributions

This paper is based on joined work of the three main authors as equal contributors. Gideon Klaila developed the theoretical formalism, contributed the main conceptual ideas and performed the analytic calculations. Both Anastasios Stefanou and Lena Ranke verified the results and contributed to the final version of the manuscript while Anastasios Stefanou helped supervise the project.

Competing interests

The authors declare that they have no competing interests.

Declarations

Ethical Approval

Not applicable.

Funding

Deutsche Forschungsgemeinschaft. Grant Number: RTG 2224

Availability of data and materials

Not applicable.

References

Aichler M, Walch A (2015) Maldi imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. Laboratory investigation 95(4):422–431

Atkeson C, McIntyre J (1986) Robot trajectory learning through practice. In: Proceedings. 1986 IEEE International Conference on Robotics and Automation, IEEE, pp 1737–1742

Berwald JJ, Gottlieb JM, Munch E (2018) Computing wasserstein distance for persistence diagrams on a quantum computer. arXiv preprint arXiv:180906433

Bubenik P, De Silva V, Scott J (2015) Metrics for generalized persistence modules. Foundations of Computational Mathematics 15:1501–1531

Carlsson G (2009) Topology and data. Bulletin of the American Mathematical Society 46(2):255–308

Chazal F, Cohen-Steiner D, Glisse M, et al (2009) Proximity of persistence modules and their diagrams. In: Hershberger J, Fogel E (eds) Proceedings of the 25th ACM Symposium on Computational Geometry, Aarhus, Denmark, June 8-10, 2009. ACM, pp 237–246, https://doi.org/10.1145/1542362.1542407, URL https://doi.org/10.1145/1542362.1542407

Cohen-Steiner D, Edelsbrunner H, Harer J (2005) Stability of persistence diagrams. In: Proceedings of the twenty-first annual symposium on Computational geometry, pp 263–271

Das S, Anand DV, Chung MK (2023) Topological data analysis of human brain networks through order statistics. Plos one 18(3):e0276419

Dłotko P, Gurnari D (2022) Euler characteristic curves and profiles: a stable shape invariant for big data problems. arXiv preprint arXiv:221201666

Edelsbrunner, Letscher, Zomorodian (2002) Topological persistence and simplification. Discrete & Computational Geometry 28:511–533

Edelsbrunner H, Harer JL (2010) Computational topology: an introduction. American Mathematical Society, Providence, USA

Edelsbrunner H, Harer J, et al (2008) Persistent homology-a survey. Contemporary mathematics 453(26):257–282

Ge QS, Zheng JY, Hao ZX, et al (2010) Temperature variation through 2000 years in china: An uncertainty analysis of reconstruction and regional difference. Geophysical Research Letters 37(3)

Ghrist R (2008) Barcodes: the persistent topology of data. Bulletin of the American Mathematical Society 45(1):61–75

Klaila G, Vutov V, Stefanou A (2023) Supervised topological data analysis for maldi mass spectrometry imaging applications. BMC bioinformatics 24(1):279

Koseki J, Hayashi S, Kojima Y, et al (2023) Topological data analysis of protein structure and inter/intra-molecular interaction changes attributable to amino acid mutations. Computational and Structural Biotechnology Journal 21:2950–2959

Kozlov DN (2020) A combinatorial method to compute explicit homology cycles using discrete morse theory. Journal of Applied and Computational Topology 4(1):79–100

Kudryavtseva EA (2009) Uniform morse lemma and isotopy criterion for morse functions on surfaces. Moscow University Mathematics Bulletin 64(4):150–158

Mémoli F (2011) Gromov–wasserstein distances and the metric approach to object matching. Foundations of computational mathematics 11:417–487

Milnor JW (1963) Morse theory. 51, Princeton university press

Nicolau M, Levine AJ, Carlsson G (2011) Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proceedings of the National Academy of Sciences 108(17):7265–7270

Otter N, Porter MA, Tillmann U, et al (2017) A roadmap for the computation of persistent homology. EPJ Data Science 6:1–38

Radtke JP, Schwab C, Wolf MB, et al (2016) Multiparametric magnetic resonance imaging (mri) and mri–transrectal ultrasound fusion biopsy for index tumor detection: correlation with radical prostatectomy specimen. European urology 70(5):846–853

Silva V, Munch E, Stefanou A (2018) Theory of interleavings on categories with a flow. Theory and Applications of Categories 33:583–607

Singh Y, Farrelly CM, Hathaway QA, et al (2023) Topological data analysis in medical imaging: current state of the art. Insights into Imaging 14(1):1–10

Ver Hoef L, Adams H, King EJ, et al (2023) A primer on topological data analysis to support image analysis tasks in environmental science. Artificial Intelligence for the Earth Systems 2(1):e220039

Yu B, You K (2021) Shape-preserving dimensionality reduction: An algorithm and measures of topological equivalence. arXiv preprint arXiv:210602096

APPENDIX B

PAPER "SUPERVISED TOPOLOGICAL DATA ANALYSIS FOR MALDI MASS SPECTROMETRY IMAGING APPLICATIONS"

Gideon Klaila, Vladimir Vutov & Anastasios Stefanou

- **Status:** Published
- **Journal:** BMC Bioinformatics
- **Impact Factor:** 11.806
- **Date of publication:** 10.07.2023
- **DOI:** https://doi.org/10.1186/s12859-023-05402-0

BMC Bioinformatics

RESEARCH

Open Access

Supervised topological data analysis for MALDI mass spectrometry imaging applications

Gideon Klaila[1*†], Vladimir Vutov[2†] and Anastasios Stefanou[1†]

†Gideon Klaila, Vladimir Vutov and Anastasios Stefanou have contributed equally to this work

*Correspondence: klailag@uni-bremen.de

[1] Institute for Algebra, Geometry, Topology and their Applications (ALTA), University of Bremen, 28359 Bremen, Germany
[2] Institute for Statistics, University of Bremen, 28359 Bremen, Germany

Abstract

Background: Matrix-assisted laser desorption/ionization mass spectrometry imaging (MALDI MSI) displays significant potential for applications in cancer research, especially in tumor typing and subtyping. Lung cancer is the primary cause of tumor-related deaths, where the most lethal entities are adenocarcinoma (ADC) and squamous cell carcinoma (SqCC). Distinguishing between these two common subtypes is crucial for therapy decisions and successful patient management.

Results: We propose a new algebraic topological framework, which obtains intrinsic information from MALDI data and transforms it to reflect topological persistence. Our framework offers two main advantages. Firstly, topological persistence aids in distinguishing the signal from noise. Secondly, it compresses the MALDI data, saving storage space and optimizes computational time for subsequent classification tasks. We present an algorithm that efficiently implements our topological framework, relying on a single tuning parameter. Afterwards, logistic regression and random forest classifiers are employed on the extracted persistence features, thereby accomplishing an automated tumor (sub-)typing process. To demonstrate the competitiveness of our proposed framework, we conduct experiments on a real-world MALDI dataset using cross-validation. Furthermore, we showcase the effectiveness of the single denoising parameter by evaluating its performance on synthetic MALDI images with varying levels of noise.

Conclusion: Our empirical experiments demonstrate that the proposed algebraic topological framework successfully captures and leverages the intrinsic spectral information from MALDI data, leading to competitive results in classifying lung cancer subtypes. Moreover, the framework's ability to be fine-tuned for denoising highlights its versatility and potential for enhancing data analysis in MALDI applications.

Keywords: Topological persistence, Persistence transformation, Peaks detection, Data denoising, Data compression, Logistic regression, Random forest

Background

Matrix-assisted laser desorption/ionization mass spectrometry imaging (MALDI MSI), also known as MALDI Imaging, is a label-free tool for spatially furnishing molecular weight information of compounds like proteins, peptides, and many others (see, e.g., [1, 2]). Provided a thin biological sample (usually a tissue section [3, 4]), MALDI collects mass spectra at multiple discrete positions within the biological sample. As a result, an image is obtained where each spatial spot presents a mass spectrum [5]. The latter depicts the relative abundances of ionizable molecules with a significant number of mass-to-charge ratio (m/z) values, ranging from a couple of hundreds to a few tens of thousands of m/z values (see [3, 5, 6]). MALDI has been demonstrated to be a valuable instrument for many pathological applications (for more details, see [7, 8]), which is possible because of its feasibility in examining formalin-fixed paraffin-embedded (FFPE) tissue samples. In other words, the MALDI tool allows analyses of multiple tumor cores across many patients by aggregating them in a single tissue microarray (TMA) (cf. [2]). As discussed in [9], "the pathological diagnosis of a tumor found in a tissue specimen, including the determination of the tumor origin and genetic subtype" [9] is essential for adequate treatment of patients. This study reports our findings on a MALDI dataset based on two lung cancer (LC) subtypes.

As pointed out by several studies (e.g., in [4, 6, 10]), a plain approach to discovering meaningful m/z values relies upon the idea of identifying significant signal peaks, also known in the literature as peak detection. Focusing on the relevant peaks, one can neglect those highly associated with noise, as acknowledged in [5]. Significant peaks are assumed to provide information for discriminating mass spectra from different cancer subtypes (see, among others, in [11, 12]). Different peak-detection algorithms have been compared in [13]. Furthermore, in [14], a more recent and novel approach proposes to incorporate an isotope pattern [15] around the chosen peaks, which can boost the peak detection methodology.

Other methods aim to extract "characteristic spectral patterns (CSP)" from MALDI-MSI data (cf. [4]). Namely, these methods combine spectral information from different correlated spectral features into a lower dimensional subspace of the data. Afterward, classification models are performed on the extracted feature vectors to classify data units into class labels, i.e., the tumor types or subtypes. Some other frameworks tend to perform feature selection first. Then, based on the selected features, such frameworks execute supervised classification methods to classify observational units into response labels (e.g., [16]). In the context of variable selection, in [10, 17], the authors have proposed approaches by means of large-scale simultaneous testing so as to identify the most associative m/z values with the (cancerous) outcome variables.

The usual challenge modeling methods face is a significant amount of spectral data. As more and more data are being gathered, analyzing the data efficiently with short computational time becomes increasingly more challenging. Topological data analysis (TDA) is a contemporary scientific area that arose from diverse works in applied topology and computational geometry (see [18, 19]). TDA offers an algebraic way of reducing the dimensionality of datasets and extracting essential features in short computational times.

TDA is a relatively novel field of data analysis, and more theoretical work needs to be done to improve the methods (cf. [20]). Even so, TDA has been successfully employed in various fields of science, e.g., in physics, chemistry, and bio-medicine [21], as well as in oncology (see [22, 23]). Motivated by these approaches, this study's objective is to propose a novel framework based on the algebraic topology of MALDI imaging to get improved classification results in a shorter computational time.

In the context of MALDI, one can take advantage of TDA by filtering out the most relevant part of the data, namely the peak-related information. We hypothesize that the importance of a peak increases with its relative height, also known as topological "persistence". Accordingly, low persistent peaks are more likely to be noise. To this end, one can benefit from utilizing our topological framework due to its superior characteristics of denoising and compression.

A general approach to determining a peak's persistence is the upper-level set filtration, where each peak corresponds to a topological feature. These topological features are tracked from their appearance until they merge with a larger feature. This way of analyzing the data is fast in its computational time but has a significant drawback. While tracking the persistence of each peak, one loses the information about their positions, which is paramount to carrying out data analysis applications. For example, in the context of MALDI-related studies, the locations of the biomarkers (cf. [4, 10]) are relevant information for the analysis. To circumvent this limitation, we introduce a different analysis method: the persistence transformation (cf. [24]). The proposed approach keeps track of each peak's position while determining its persistence. This enables the application of this methodology to spectral data.

Topological data analysis

MALDI data structure

The first step of our approach is to transform the input MALDI data in order to reflect the topological persistence. MALDI-MSI datasets are commonly stored in an $n \times q$ matrix, denoted by X, and X takes its values in $\mathbb{R}_{\geq 0}^{n \times q}$, where each data record corresponds to an intensity value. Usually, mass spectra are stored as rows. While every data column corresponds to an intensity plot for a certain m/z value (see in [9, 25]), n corresponds to the number of mass spectra, and q is the number of m/z values within each mass spectrum. Furthermore, the data records are non-negative since MALDI data presents information on molecular masses of ionizable molecules (cf. [5]). For the convenience of notation, we assume that each mass spectrum is represented as a set of tuples $M := \{(x_1, s_1), (x_2, s_2), \cdots (x_q, s_q)\}$, where x_j is the j-th m/z value and s_j is the j-th intensity value for $1 \leq j \leq q$. Note that this set M is equipped with a real-valued function f corresponding to the projection to the second coordinate. That means that the intensity values induce a map $f : M \to \mathbb{R}$ with $f(x_j) := s_j$ for $1 \leq j \leq q$ within each mass spectrum.

To sum up, our approach processes each mass spectrum (defined by the set M) individually in order to extract the topological properties of the data, i.e., the topological persistence.

Topological persistence

Let $f : M \to \mathbb{R}$ be a real-valued function on a compact set, then for $a \in \mathbb{R}$ the upper-level set can be defined as $M_a := \{x \in M | f(x) \geq a\}$. Note that $M_a \subseteq M_{a'}$ for any $a > a'$. This yields the "upper-level set filtration" $M_{a_1} \subseteq \cdots \subseteq M_{a_i}$ for $a_1 > \cdots > a_i \in \mathbb{R}$ (for more details; see Chapter 18 in [20]). In this study, we aim at utilizing the upper-level set filtration instead of the common alternative, i.e., the sublevel set filtration (defined as $M_{<a} := \{x \in M | f(x) \leq a\}$). Since the latter, filtration detects local minima and tracks them until they merge with other minima. Conversely, the maxima are highly important in MALDI applications because they correspond to the underlying spectral peaks. As mentioned, peaks provide the necessary information to distinguish mass spectra from different cancerous subtypes [5, 10]. To this end, the upper-level set filtration is of interest in this study since it tracks the local maxima.

Let $x, x' \in M$ with a path-connection, i.e. there exists a continuous function $\rho : [0, 1] \to M$ such that $\rho(0) = x$ and $\rho(1) = x'$. We denote the image of the map ρ by $[x, x']_\rho$. Then we say that x and x' are path-connected in M_a, and we write $x \sim_a x'$, if there exists a path connection ρ of x, x', such that $\forall \hat{x} \in [x, x']_\rho : \hat{x} \in M_a$, i.e. $f(\hat{x}) \geq a$, for all $\hat{x} \in [x, x']_\rho$. To track the homology in the upper-level set filtration, we now identify all path-connected points to each other. The degree of the 0-the homology group h_0 is given by $b_0(M_a) = |M_a / \sim_a|$, i.e. the number of different path-connected components in M_a. These components are called "topological features". Henceforth, we refer to a (topological) feature in this pure topological sense, not a feature in a statistical sense, like a covariate or an explanatory variable. Remark that there are no features of dimension one or higher since each data unit (mass spectra; see Fig. 5) is represented as a curve.

In the upper-level set filtration, a topological feature is detected for x at a^* if and only if $x \in M_{a^*}$ and $\forall x' \in M_{a^*} : x \not\sim_{a^*} x'$, i.e., x is not path-connected to any other element in M_{a^*}. The feature in x merges with another feature in M_{a^+}, if $a^+ = \max\{a | x \in M_a \wedge \exists x' \in M_a : x \sim_a x' \wedge f(x') > f(x)\}$, i.e., the largest upper-level set in which the peak gets path-connected to a larger peak. We call a^* the *birth* of the feature, a^+ the *death* of the feature, and $p := a^* - a^+$ the *persistence* of the feature. Since the largest peak (the global maximum) does not merge with any other feature, its death is defined as the global minimum. The induced merging order is according to the "elder rule" (see [26]).

To encode all information about features and their persistence in a meaningful and comparable way, the persistence diagram is utilized. Figure 1 displays an example of the encoding process for the upper-level set filtration. Here, the birth and death axis are swapped since the birth values are always greater or equal to the death values. In this way, all the feature points appear above the diagonal line $x = f(x)$ (cf. Section 4 in [27]). In the persistence diagram, each feature is represented by a point (a^*, a^+) with a multiplicity for similar features. The closer a point is to the diagonal of the diagram $\{(x, f(x) = x) \mid \forall x\}$, the lesser its persistence is and vice versa. For more details of the standard persistence diagram approach, we refer to [28, 29].

Given two real-valued functions on a compact set $f, g : M \to \mathbb{R}$, the resulting persistence diagrams $\mathbf{dgm}(f)$ and $\mathbf{dgm}(g)$ can be compared with a suitable metric on the diagrams, e.g., the bottleneck distance (cf. [29, 30]). This metric defines the closeness of persistence diagrams and can also indicate the closeness of the corresponding functions

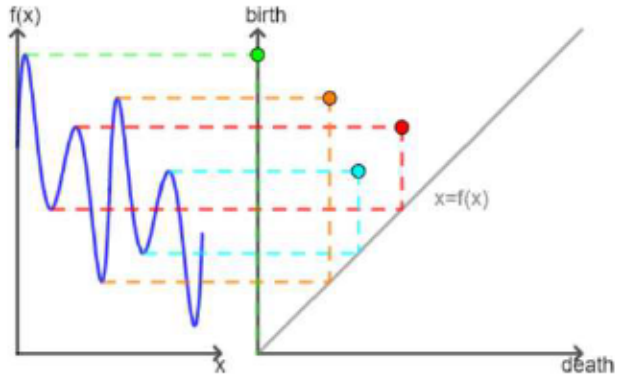


Fig. 1 The persistence diagram. **Left hand**: A real-valued function $f : M \to \mathbb{R}$ is plotted with $x \in M$ on the x-axis and $f(x)$ on the y-axis. **Right hand**: The corresponding persistence diagram is illustrated. Here, the y-axis represents the birth values, while the x-axis represents the death values. Each point stands for a topological feature for the function f. The first feature, corresponding to the global maxima, never dies. To indicate this, we mark its death value with 0

in the sense that if the persistence diagrams are different, the functions are different. In the context of MALDI-MSI data, the persistence diagram can be used as a proxy for distinguishing cancer (sub-)types. The opposite case that similar persistence diagrams imply similar functions does not hold, as illustrated in Fig. 4.

Persistence transformation

Using persistence homology to track the features and applying the bottleneck distance on the resulting persistence diagrams is a viable way to distinguish functions. However, there is a disadvantage to this approach. While tracking the persistence of each feature in a comprehensible way, its position is not being tracked. Nevertheless, as mentioned above, in data-driven applications, including MALDI Imaging, the position of variables is a required property.

To process the information regarding the positions along with the persistence of peaks, we approach the topological manipulation of the spectral data differently. Instead of creating a point $(a^*, a^+) \in \mathbb{R}^2$ for each feature and displaying it in the persistence diagram, we introduce a new dimension to track the position. For each feature, this approach gives a point $(x, a^*, a^+) \in M \times \mathbb{R}^2$, where $x \in M$ is the position of each peak. We define a pairing function $\mu : M \to \mathbb{R}$ with $\mu(x) = a$. The value a is defined to be the highest value smaller or equal to $f(x)$, which upper-level set M_a contains a point x' having a greater function value (similarly as in [24]):

$$\mu(x) := \sup\{a \leq f(x) | \exists x' \in M_a : f(x') > f(x) \wedge x \sim_a x'\}. \tag{1}$$

For the global maximum $\hat{x}$ the pairing value $\mu(\hat{x})$ is not defined, so we define $\mu(\hat{x}) := \min\{a \in \mathbb{R} | \exists x : f(x) = a\}$ instead. Then the birth of a topological feature is given by $a^* = f(x)$, while the death is calculated by $a^+ = \mu(x)$. The persistence p of a point x can now be defined to be

$$p(x) := f(x) - \mu(x) = a^* - a^+.$$

For any point x not being a local maximum there is a local maximum $x' \in M_{f(x)}$ with $f(x') > f(x)$ and $x \sim_{f(x)} x'$. Then the pairing of x is trivial, i.e. $\mu(x) = f(x)$ with the resulting persistence of $p(x) = f(x) - \mu(x) = 0$. Alternatively, for a point x being a local maximum, the pairing is non-trivial, i.e., there is $\mu(x) < f(x)$, which results in a non-zero persistence $p(x) = f(x) - \mu(x) > 0$. This pairing value always corresponds to $f(\tilde{x})$ for a unique local minimum $\tilde{x}$, i.e.

$$\mu(x) = f(\tilde{x}). \tag{2}$$

The persistence transformation $t : M \to M \times \mathbb{R}^2$ can then be defined for each $x \in M$ to be $t(x) = (x, f(x), \mu(x)) = (x, a^*, a^+) \in M \times \mathbb{R}^2$. For each $x \in M$, the feature triple $t(x) = (x, a^*, a^+)$ consists of the position, the birth value, and the death value. The storage can be reduced to $3 \cdot m$, with m being the number of peaks, by neglecting the trivial tuples, resulting in the "persistence transformation vector".

Similar to the persistence diagram of the upper-level set filtration, the merging of features in the persistence transformation occurs according to the elder rule (see [26]), i.e., the topological feature with the higher birth value persists when merged to another feature. The process can be illustrated in the corresponding merge tree (e.g., in Fig. 2).

Application

The notion of the persistence transformation is, in theory, a great way to store more topological information about a graph in general. But many applications do not need the whole information provided by the persistence feature. In these cases, the greater dimensionality of the persistence transformation is rather disadvantageous in calculations. This could be evaded by introducing a "reduced persistence transformation", where instead of $t(x) = (x, a^*, a^+)$ only the position and the persistence are stored:

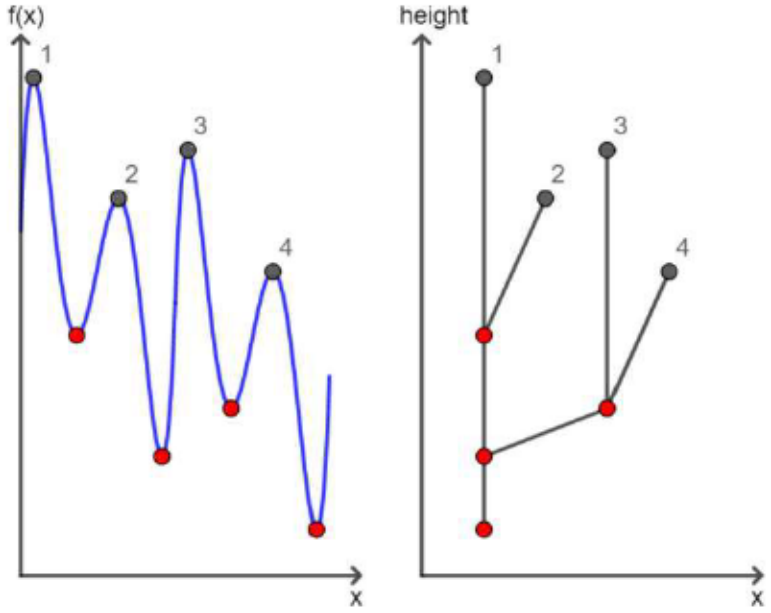


Fig. 2 The merge tree. A merge tree of a real-valued function $f : M \to \mathbb{R}$ is depicted. On the x-axis are the $x \in M$ values. On the y-axis is $f(x) \in \mathbb{R}$ (left-hand side) and the height (right-hand side)

$\tilde{t}(x) = (x, a^* - a^+) \in M \times \mathbb{R}$ for $x \in M$. This computational reduction has two benefits. First, it compresses the persistence vector. Second, it can track the persistence of each topological feature and its position.

Another computational improvement for applications can be made by associating low persistent features with noise. By omitting the possible noise by only considering the k% most persistent features, the accuracy of the analysis can be improved. In addition, reducing the number of stored features might also improve the run-time of subsequent approaches.

Finally, in many applications, there exists a total order on the set M, e.g., if $M \subseteq \mathbb{R}$. This order can be passed to the feature space, such that the elder rule for features with equal birth value can be applied deterministically by using the induced order.

Comparison

The persistence transformation is strictly a better invariant in distinguishing two functions $f, g : M \to \mathbb{R}$ than the persistence diagram of their 0-dimensional upper-level set filtration. By taking the projection, $p_i((x, a^*, a^+)) = (a^*, a^+)$ of the persistence transformation, the persistence diagram of the upper-level set filtration can be obtained. Hence, all functions that can be distinguished by the persistence diagram can also be distinguished by the persistence transformation. Furthermore, there are cases where the persistence transformation differentiates two functions successfully while the zero-dimensional persistence diagram of the upper-level set filtration fails to do so (see Fig. 3).

The reduced persistence transformation, on the other hand, is not strictly better than the persistence diagram of the upper-level set filtration. However, there are cases where the reduced persistence transformation outperforms the persistence diagram (see Fig. 4). For MALDI-related applications, one seeks to utilize exactly the advantages that the (reduced) transformation offers. For example, the total height of a peak is less interesting in these applications than the relative height, and the position of the peak can be used to backtrack molecules. To this end, in the MALDI-MSI applications, the reduced persistence transformation performs better in analyzing the spectral data than the persistence diagram of the upper-level set filtration.

The benefit of the persistence transformation can be utilized in different scientific areas where the position of the peaks is of interest (e.g., [31]). In the context of TDA, data analyses generally only work reliably if the applied method is stable, meaning that a slight change in the data (given a suitable metric) only leads to small changes in the numerical results. The persistence diagram has been proven stable (see [29]). Furthermore, there are stability theorems for other topological methods (see [32, 33]), but to the best of our knowledge, there has not yet been a proven stability theorem for the persistence transformation. This may be done in further work.

Implementation and analysis of the Algorithm

To analyze the MALDI dataset, we implemented a custom-made computer algorithm for the reduced persistence transformation. The recursive algorithm is based on pairing peaks with their unique local minimum (cf. Eq. (2)) to determine their persistence. The pseudo code with a detailed analysis by run-time and storage usage is given in the "Additional file 1", proving the following theorem:

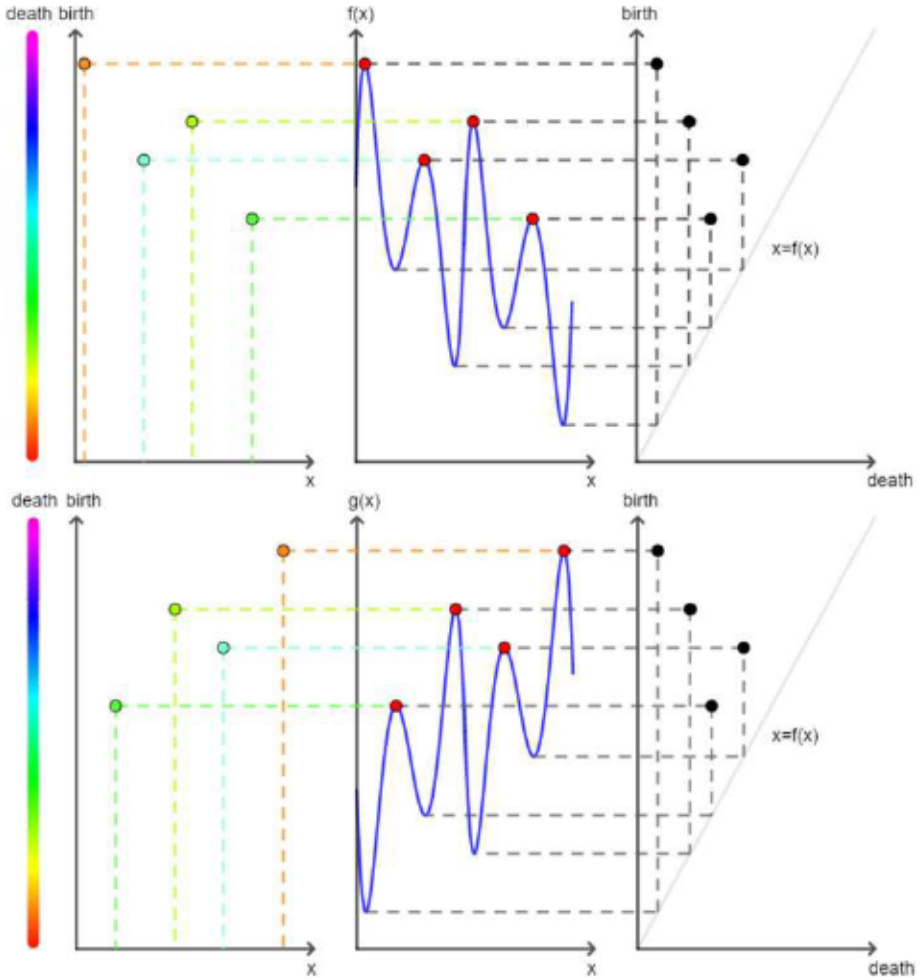


Fig. 3 The persistence transformation in comparison to the persistence diagram. **Mid**: A real-valued function $f : M \to \mathbb{R}$ (top row) and its mirror equivalent $g : M \to \mathbb{R}$ (bottom row) are illustrated, with $x \in M$ on the x-axis and $y \in \mathbb{R}$ on the y-axis. **Left hand:** The persistence transformation of the functions f and g are shown with x values on the x-axis and birth values on the y-axis. The color coding of the points corresponds to the third dimension, i.e., the magnitude of death values. The features (i.e., the points) of the functions f and g can be distinguished clearly. **Right hand**: The persistence diagram of the upper-level set filtration of the functions f and g is shown with the birth values on the y-axis and the death values on the x-axis. In contrast to the persistence transformation, the persistence diagrams are identical

Theorem 1 *The algorithm always terminates and has a complexity of* $\sigma(q) + \sigma(m^2)$*, where* m *is the number of peaks for each spectrum and returns all features with their persistence.*

Note that the implementation of the algorithm stores for each feature the tuple $(x, p(x))$, where $p(x) = a^* - a^+$ is the persistence of the feature. Without the further cost of calculation, the algorithm could calculate the persistence transformation instead of the reduced persistence transformation by storing the triple (x, a^*, a^+). Furthermore, the algorithm could be adjusted to determine the persistence diagram of the upper-level set filtration instead by storing the tuple (a^*, a^+).

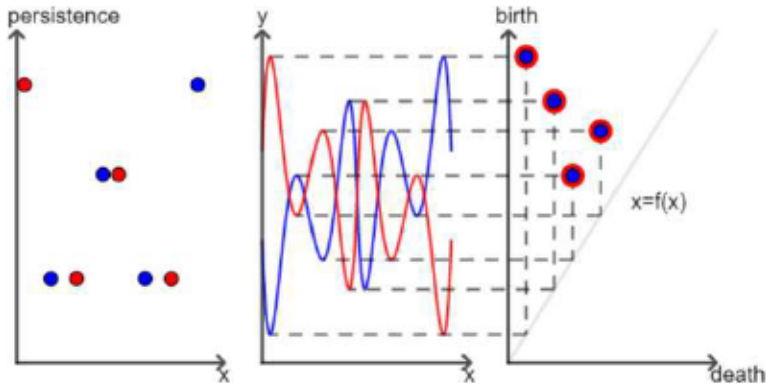


Fig. 4 The reduced persistence transformation in comparison to the persistence diagram. **Mid**: There are two real-valued functions $f : M \to \mathbb{R}$ (blue) and $g : M \to \mathbb{R}$ (red) plotted with $x \in M$ on the x-axis and $y \in \mathbb{R}$ on the y-axis. **Left hand**: The persistence transformation of the functions f (blue) and g (red) are displayed with the x values on the x-axis and the persistence values on the y-axis. The features of g are distinct from the features of f (blue). **Right hand**: The persistence diagram of the functions f (blue) and g (red) are shown. The y-axis indicates the birth values, while the x-axis shows the death values. The features of g (red) cannot be distinguished from the features of f (blue)

By considering only the peak information while tracking the position of the peaks, the storage space can be compressed to $2 \cdot m$, i.e. the x value and the p value.

Supervised methods

The second step of the proposed methodology is to carry out a classification method on the resulting persistences to classify observational units into class labels, i.e., LC subtypes. To do so, we consider two classifiers, logistic regression (LR) and random forest (RF). Note that our goal is to investigate the performance of the proposed topological framework in the context of MALDI modeling, not a benchmark study that compares RF versus LR. For benchmark studies, see, e.g., in [34, 35].

Logistic regression

Throughout the remainder, we denote the topologically transformed matrix by $Z = (z_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le q}}$ where the entry z_{ij} corresponds to a persistence value for the i-th mass spectrum at the j-th m/z value. Let Y be a (random) binary outcome variable, meaning that it takes its values in $\{0, 1\}$, in our application, describing two LC subtypes. Further, we denote by Z_j the j-th persistence vector corresponding to its j-th m/z value alternative. The aim of the logistic regression is to model and estimate the effects of the available covariates on the conditional probability, $\pi_i = P(Y_i = 1|Z_{i1}, \ldots, Z_{iq})$ for the outcome variable $(Y_i)_{1 \le i \le n}$ and the numerical realizations of the covariates $(Z_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le q}}$. In this setting, the observations $(Y_i, Z_{ij})_{1 \le j \le q}$ are assumed to be independent and identically distributed for all $i \in \{1, \ldots, n\}$. LR models combine the probability π_i with the linear predictor η_i via a "structural" (functional) component given in the linear form $\pi_i = h(\eta_i) = h(\beta_0 + \beta_1 Z_{i1} + \cdots + \beta_q Z_{iq})$ (for more details, see Section 2 in [36]).

In this study, we consider the logit (canonical) link function. Then, the logistic response function is given by

$$\pi_i = h(\eta_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}.$$

Correspondingly, the logit link function can be expressed as

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 Z_{i1} + \cdots + \beta_q Z_{iq} = Z_i^T \beta. \tag{3}$$

Here, $\beta := (\beta_0, \beta_1, \ldots, \beta_q)^T$ and $Z_i := (1, Z_{i1}, \ldots, Z_{iq})^T$, where the first coordinate corresponds to an intercept term and for all $i \in \{1, \ldots, n\}$. The unknown parameters are (usually) estimated by the maximum (log-) likelihood principle. The log-likelihood function related to model (3) is expressed by

$$l(\beta) = \sum_{i=1}^{n} \{Y_i[\log(\pi_i) - \log(1-\pi_i)] + \log(1-\pi_i)\}. \tag{4}$$

For the aforementioned (logit) model, by plugging

$$\pi_i = \frac{\exp(Z_i^T\beta)}{1+\exp(Z_i^T\beta)} \text{ along with } 1-\pi_i = \frac{1}{1+\exp(Z_i^T\beta)}$$

in (4), it yields that

$$l(\beta) = \sum_{i=1}^{n} \{Y_i\left(Z_i^T\beta\right) - \log\left(1+\exp(Z_i^T\beta)\right)\}. \tag{5}$$

Finally, the probability that an unobserved (new) observation is assigned to class 1 is estimated by substituting $(\beta_0, \ldots, \beta_q)$ by their fitted counterparts and Z's by their numerical realizations for the considered new observation in the conditional $P(Y = 1|Z_0, \ldots, Z_q) = \frac{\exp(Z^T\beta)}{1+\exp(Z^T\beta)}$, where we have $q+1$ covariates since the first coordinate corresponds to the intercept term. Respectively, the new observation is assigned to class $Y = 1$ if the conditional probability, $P(Y = 1) > c$, is greater than a pre-specified threshold c, and oppositely to class $Y = 0$. In this study, we set $c = 0.5$ – a commonly used threshold (cf. [34]). To obtain the numerical results, we adopted the "LogisticRegression()" function in *scikit-learn, v. 1.2.1* [37] (with no penalty) to obtain our numerical results.

Random forest

The RF algorithm has become an established non-parametric procedure for regression and classification tasks. It has been broadly used in various scientific disciplines [38–40], including subtyping of lung cancer [41]. RF was originally introduced by Leo Breiman in [42] and it presents an "ensemble learning" approach constituting the aggregation of a collection of a great number of decision trees [43]. RF takes advantage of numerous decision trees, which leads to a reduction of empirical variance in comparison to a single (decision) tree and significant enhancements in its prediction accuracy [35]. RF utilizes decision trees in order to calculate the majority votes in the leaf nodes when deciding a class label for each observational unit [35, 42]. In essence, RF consists of two steps. The

first step is to build an RF tree. The following step is to classify the data on the basis of an RF tree that has been generated in the first step. For more details, e.g., see in [39, 44].

In this study, we employ the original variant of RF (see [42]), where each tree of the RF algorithm is constructed on a bootstrap sample drawn arbitrarily from the data by employing the classification and regression trees method and minimizes the Gini impurity (GI) regarding the splitting criterion. When constructing each tree (for each split), solely a pre-specified number of randomly selected (data) covariates are deemed as candidates for splitting.

An important step when using RF is selecting hyperparameters, also called tuning parameters. Their values have to be optimized attentively since the optimal quantities depend on the data at hand. An essential concept regarding tuning optimization is "overfitting". In other words, tuning parameters related to complex rules be inclined to "overfit" the training data. As a result, they produce prediction rules overly specific to the training data, performing well for that (training) data but potentially underperforming when applied to independent data. As discussed in [45], the choice of less-than-optimal parameter quantities can be (at least) partially prevented by utilizing a test set or cross-validation (CV) procedures for tuning. However, it is out of the scope of this study to identify the (most) optimal tuning parameters in the context of MALDI modeling. Instead, we are predominantly interested in evaluating the performance of the proposed TDA approach. To this end, we select the "typical default values" for the RF algorithm, as listed in Table 1 in [45]. Specifically, we set the tuning parameters in our numerical experiments as follows: the number of trees equals 1000. The number of drawn candidate variables per split is equal to $\sqrt{q}$ (often referred to as "mtry", *max_features* in *scikit-learn*). The splitting criterion in the nodes is the Gini impurity. The minimum number of samples in a terminal node is equal to one (*min_samples_leaf* in *scikit-learn*). Regarding the sampling scheme, the number of observational units that are (randomly) drawn for training each tree is determined by the sample size parameter. The default value corresponds to n (i.e., to the overall number of data samples), respectively, observational units are drawn with replacement when generating each tree. The seed for all experiments was set to 1234.

We carried out the RF approach on the basis of the resulting algebraic vectors $Z := (Z_{i1}, \ldots, Z_{iq})^T$ and the binary outcome Y_i for all $i \in \{1, \ldots, n\}$. Notice that a unit vector has not been considered, i.e., without an intercept term, as in the case of LR. We adopted the function '*RandomForestClassifier()*' in [37] (*version 1.2.1*) to yield the numerical results.

Real data analysis

Description of the MALDI-MSI data

Here, we present the empirical results obtained by applying multiple classification schemes to MALDI-MSI data based on different levels of persistence extraction. Several studies have previously analyzed this dataset (cf. [4, 6, 7, 9, 10]). We refer to [4, 7, 9] for an in-depth description of the aforementioned dataset concerning its acquisition protocols, tissue sections, tissue blocks, etc. Here, we provide only a brief outline of the dataset.

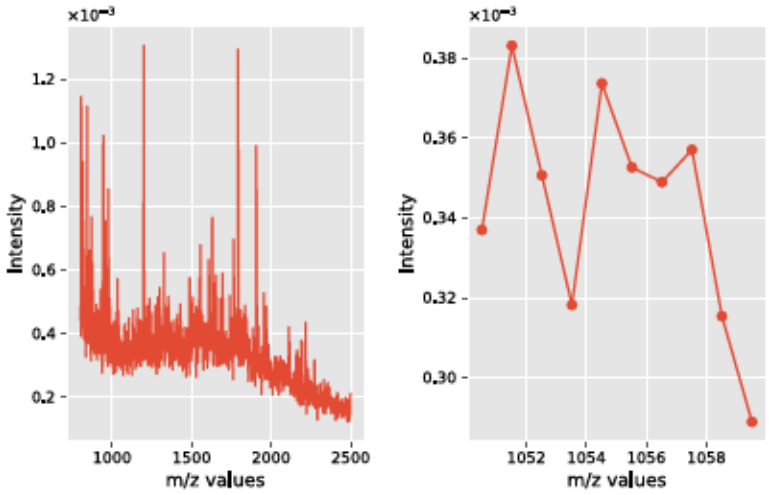


Fig. 5 An example of a mass spectrum. **Left hand:** An example of a mass spectrum from a single cancerous spot within a patient tissue. **Right hand:** A closer look at spectral data following spectral filtering

Table 1 Descriptive statistics for each TMA

TMA	Number of spectra	Ratio ADC: SqCC (%)
TMA_1	680	59.3
TMA_2	437	60.4
TMA_3	563	41.2
TMA_4	601	49.3
TMA_5	512	63.5
TMA_6	650	76.8
TMA_7	536	50.4
TMA_8	690	54.3

Cylindrical tissue cores (CTCs) of non-small cell lung cancer were taken from 304 patients. Specifically, 168 patients were associated with primary lung adenocarcinoma (ADC), while 136 patients were associated with primary squamous cell carcinoma (SqCC). CTCs of all patients were gathered into eight tissue microarray (TMA) blocks (for descriptive statistics, see Table 1). As discussed in [9], the tumor status and subtyping for all CTCs were affirmed by standard "histopathological examination". Furthermore, this dataset has been generated only on annotated subregions called regions-of-interest (cf. [6]), i.e., subregions comprising only tumor cells.

For illustrative purposes, Fig. 5 depicts an example of an output from a MALDI experiment taken from a single spatial location in the provided tissue. In the left panel of Fig. 5, the m/z values are illustrated on the x-axis, while the intensity values of "ionizable molecules" are charted on the vertical axis. This spatial information can be used in two directions, namely, for the determination of the subtyping (i.e., the cancer subtypes) or the identification of the source of the tumor in tissue. The right panel of Fig. 5 illustrates the data granularity following one of the data-processing steps, i.e., spectral filtering. Namely, the latter means that m/z values were centered

around their expected peptide masses (for more details, see [9] and the references therein). Other data-processing steps applied to this dataset were baseline correction and total ion count (TIC) normalization.

As pointed out in different studies (e.g., [46, 47]), LC is the primary cause of cancer-related fatalities globally; for example, there were 1.59 million reported deaths in 2012 (see [47]). Two major LC categories are recognized, i.e., small cell lung cancer (SCLC) and non-small cell lung cancer (NSLC). The latter constitutes approximately 85% of all LC cases as reported. The two prevailing NSLC entities are ADC and SqCC, compromising approx. 50% and 40% of all lung-related cancers, respectively [46, 47]. As discussed in [9], the distinction between these two common subtypes is of great importance for the therapy choice of patients.

The used dataset can be found on Gitlab, as provided in [9]. Note we did not apply any other data-processing steps on this dataset. Specifically, this dataset contains $n = 4669$ (observational units, the number of mass spectra), and the number of m/z values is $q = 1699$.

Classification evaluation

To evaluate the performance of the classification schemes, we mimic the realistic scenario proposed in [9]. Namely, the data is split into training and test sets. The upcoming results were derived by performing k-fold cross-validation (CV) on a TMA level. We followed both scenarios as proposed in [9], specifically 8-fold CV and 2-fold CV. Regarding the 8-fold CV, eight distinct test subsets were created based on each TMA from the overall set of eight TMA blocks. Then, in each of the eight CV folds, each classification scheme was applied to seven TMAs and predicted on the remaining test set—not considered in the training process. Likewise, we carried out 2-fold CV, creating two subsets $A := \{TMA_1, \ldots, TMA_4\}$ and $B := \{TMA_5, \ldots, TMA_8\}$. We reported the obtained results on the basis of all test sets for the 2-fold and 8-fold CVs.

The classification accuracy was assessed by computing the balanced accuracy. The latter metric is computed as the average proportions of correctly classified spectra for each class separately. As a result, this metric is independent with respect to imbalanced binary categories, i.e., when one of the target classes appears far more often than the other in the test set. To illustrate the numerical performances appertaining to the persistence transformation, we set a tuning parameter k based on different percentages of peaks extraction.

Data-analysis results

Figure 6 depicts an example of the proposed topological feature extraction. Namely, the top row illustrates an example of a raw spectrum (cf. Figure 5), whereas the next subplots depict extracted persistence values. Apart from the first row, the m/z values are plotted on the horizontal axes, while the vertical axes show the derived persistence values, not the intensity values as in the first row of the figure.

Tables 2, 3 summarize the classification results obtained by applying LR and RF classifiers based on several levels of the extracted persistence vectors. The percentage of extracted persistence employed for the evaluation, denoted by k, ranges in a grid of prespecified percentages—$k \in \{10\%, 20\%, 25\%, 30\%, 40\%, 50\%\}$. By compressing the raw

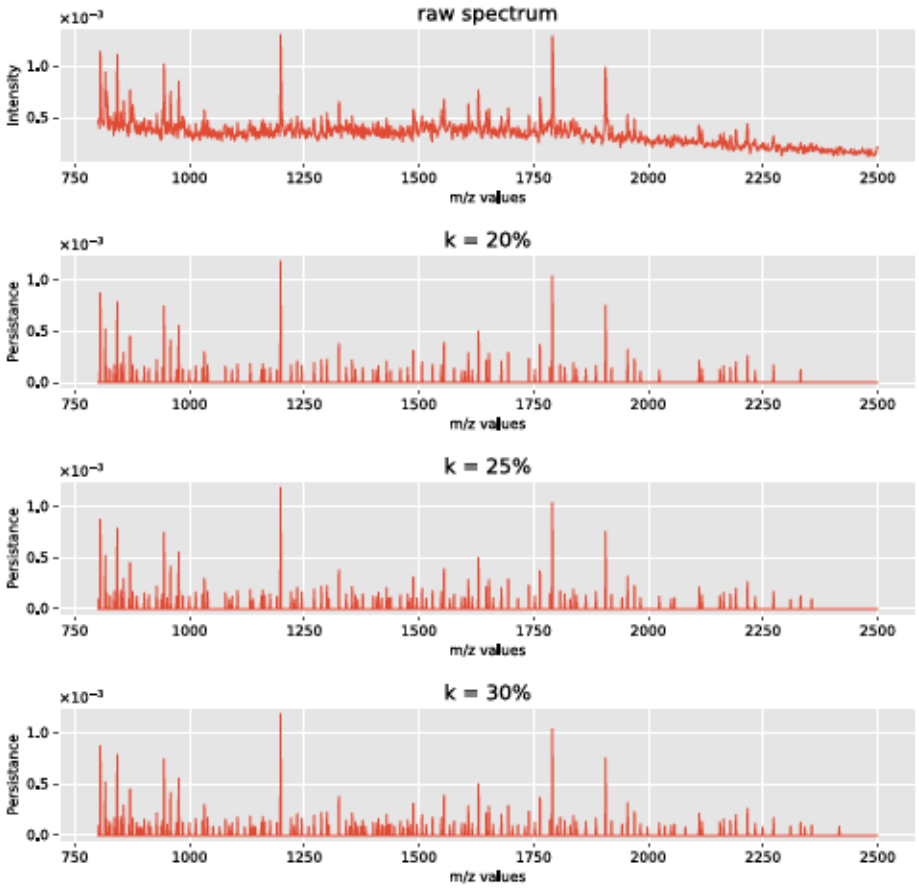


Fig. 6 Application of the reduced persistence transformation. A visualizing inspection of the proposed topological framework. The top row subplot depicts the raw spectrum, while the following subplots illustrate simplified representations of this raw spectrum given different percentages of peak extraction (abbreviated with *k*%). Namely, the raw spectrum is transformed to sparse tuples based on a pre-specified k level of peaks extraction

Table 2 Comparison table of 8-fold cross-validation. The table tabulates the obtained balanced accuracy results given different percentages of peaks extraction and non-extracted data based on the LR classifier

Statistic	k = 10%	k = 20%	k = 25%	k = 30%	k = 40%	k = 50%
mean	0.826	0.845	0.859	0.866	0.859	0.859
min	0.722	0.716	0.724	0.743	0.737	0.723
max	0.917	0.937	0.931	0.950	0.935	0.935
median	0.827	0.869	0.886	0.888	0.880	0.878
std	0.074	0.079	0.077	0.076	0.076	0.078

data with respect to different k, we observed a significant gain in the computational time for executing RF. For example, to execute the 8-fold CV task on 7 CPUs on a standard machine, it takes 35 seconds when using $k = 5\%$ and 70 seconds when using $k = 50\%$, in contrary to the 215 seconds it takes using the raw data, i.e., the original variant where each data entry corresponds to intensity value.

Table 3 Comparison table of 8-fold cross-validation. The table tabulates the obtained balanced accuracy results given different percentages of peaks extraction and non-extracted data based on the RF classifier

Statistic	k = 10%	k = 20%	k = 25%	k = 30%	k = 40%	k = 50%
mean	0.831	0.863	0.871	0.878	0.877	0.868
min	0.702	0.734	0.760	0.774	0.776	0.746
max	0.936	0.945	0.940	0.959	0.930	0.949
median	0.849	0.873	0.880	0.892	0.900	0.890
std	0.083	0.076	0.066	0.067	0.061	0.079

To put our results in the context of other competitors for this dataset, we performed a comparison with a popular method for retrieving informative parts of the spectral data and executing automated cancer (sub-)typing. Briefly put, in [9], the authors proposed novel supervised non-negative matrix factorization (NMF) methods: the classification tasks are executed in parallel to feature extractions (in the context of NMF extraction), which differs from the more classical NMF-related scenario (cf. [4]). The authors introduced 13 distinct classification schemes. From these, we selected the top 2 competitors; for the remaining ones, we refer interested readers to Figure 3 and Figure 4 in [9]. These top 2 competitors from [9] are: *Flog_int Flog_log*, where the number of NMF "features" is 60, as suggested in the provided code. Our competitors are the persistence transformation where k is either 30% or 40% and the FR classifier. These schemes are abbreviated to *PT_ RF_*30% and *PT_ RF_*40%, respectively.

Table 4 illustrates that *Flog_ int* and *Flog_ log* perform slightly better than our best performers for this dataset for the 8-fold CV task and 2-fold Train B. However, the topological-based competitors outperformed the other 11 NMF-based schemes for this dataset, even in some scenarios *PT_ RF_*40% produces numerically similar results vis-à-vis all NMF-based competitors, cf. Bal. Acc. (2-fold) Train A. The authors [9] concluded that apart from the top three classification schemes, most of the other methods achieved, on average, balanced accuracy values below 80%. This table stands for the proof of concept that our topological framework accompanying the RF classifier can produce competitive results. Moreover, the RF (non-linear) algorithm can operate in pure high-dimensional scenarios, i.e., when covariates exceed the observational units $n \ll q$. Therefore, the proposed classification scheme $PT + RF$ can also be applied in different data regimes.

Image denoising with persistence transformation

Simulation setup

A (big) challenge in examining real-world applications is the presence of noise that can corrupt the data and lead to incorrect data-analysis results (see [14, 48]). To this end, researchers have to be careful when analyzing datasets with a possibility of noise and address it appropriately to improve the accuracy of the results (see [10, 17, 49]).

To assess the effectiveness of the proposed topological framework given the presence of noise, we proceeded as follows. First, we simulated multiple synthetic mass spectrometry (MS) images, where each pixel of these images corresponds to a unique mass spectrum. Specifically, each pixel presents an average value for its respective mass spectrum.

Table 4 Performance of the proposed classification algorithms vis-à-vis the best classification competitors from [9]. *PT_ RF_k%* stands for Persistence Transformation, using the Random Forest classifierer with *k* as hyperparameter. According to [9], Flog stands for the Frobenius norm utilizing the logistic regression classifier. The suffix '*_int*' denotes the Integrated approach, while '*_log*' stands for the Optimized approach

Method	Avg. Bal. Acc. (8-Fold)	Bal. Acc. (2-Fold) Train A (%)	Bal. Acc. (2-Fold) Train B (%)
PT_ RF_30%	87.8% ± 6.70	90.75	84.40
PT_ RF_40%	87.7% ± 6.01	91.15	83.53
Flog_int	89.8% ± 4.35	90.8	89.1
Flog_log	88.8% ± 6.36	91.1	87.1

Second, we artificially contaminated the spectral data that generated the MS images by adding different types and levels of noise. Finally, for each figure, we plotted the ground truth, the noise image, and two variants of denoised MS images in a row. As a result, one can pictorially identify the ability of our TDA approach to differentiate signal from noise in the images. Two of these results are displayed in Fig. 7 and in Fig. 9.

We utilized the *"Cardinal"* package ([50] *v. 3.0.1*) in *R* ([51] to simulate noiseless (ground truth) MS images. Accordingly, we employed the "SimulateImage()" function with the following parameters: the preset image is two (i.e., there are two figures, a circle in the top-left corner and a square in the bottom-right corner), the m/z range lies in $500 - 2000$ (resulting in 3466 m/z values), the number of peaks $k^* = 50$, and a noiseless image (i.e., "sdnoise" equals to zero). We aim to demonstrate the efficacy of our methodology with varying baseline levels so as to cover baseline value ranges $\in \{0, 5, 15\}$. Based on these parameters, we simulated multiple MS images with sizes $\{30 \times 30, 42 \times 42, 60 \times 60\}$. Following the simulation of the ground truth images, we contaminated the spectral data by adding either Gaussian or Poisson noise. We chose increasing values for the standard deviation for the Gaussian noise and λ for the Poisson noise, i.e., artificially contaminated synthetic data more and more. These distribution parameters are given in the caption of each subplot and can serve as a proxy for different signal-to-noise ratio variants. Due to its relevance to our real-world data application, we proceeded by picking a percentage of the most significant peaks – signals outside these fractions were considered noise.

Adding artificial noise to the MALDI spectra has two significant effects. First, the height of the existing signal peaks could be altered. Second, new noise peaks can be established. For low levels of noise, the added noise peaks are less persistent than the signal peaks. By optimizing the tuning parameter k, our algorithm can differentiate the bulk of the signal peaks by neglecting the noisy ones in spectral data. A good example of low-level noise is the Gaussian noise, which is displayed in Figs. 7 and 8. As can be seen in Fig. 8, most of the signal peaks can be distinguished from the noise peaks by their height. Hence, the original shapes from the ground truth images can be reconstructed in the denoised images in Fig. 7 when the tuning parameter k is chosen accordingly.

Higher levels of noise, on the other hand, can be challenging for the persistence transformation. By adding noise peaks larger than the signal peaks, the ground truth can be compromised so that the persistence transformation can not reconstruct the original

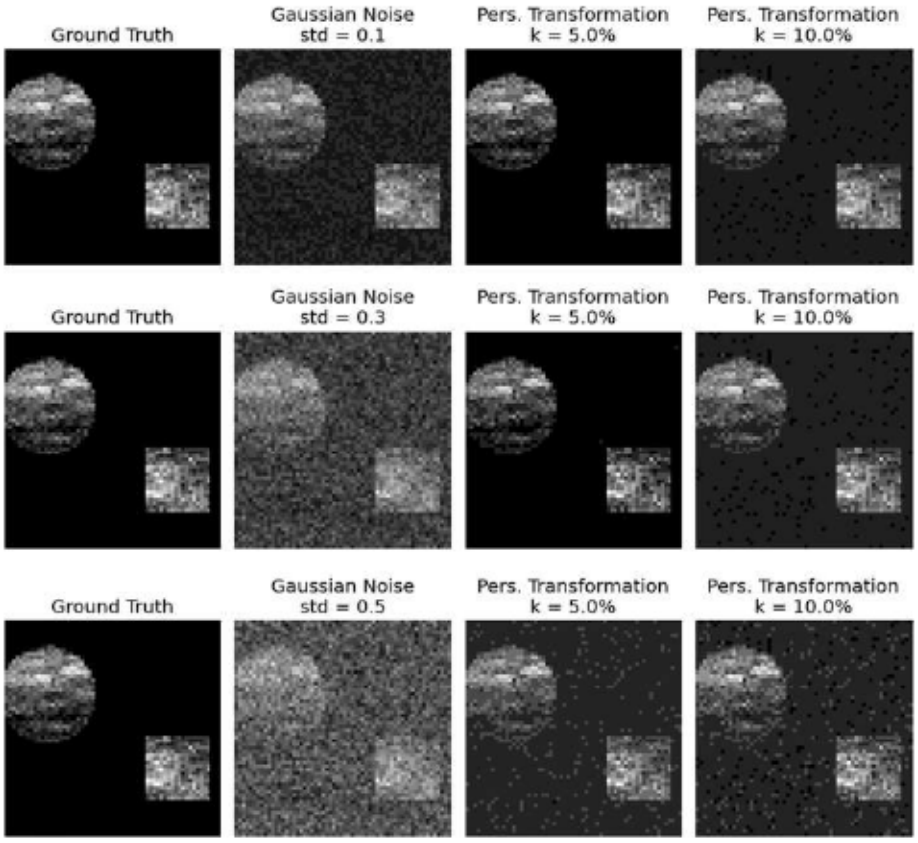


Fig. 7 Denoising with the persistence transformation. On the leftmost column, the ground truths of synthetic MALDI-Images are displayed. In the second column, distinct levels of Gaussian noise are added to the spectral data. The processed images based on two choices of k are displayed in the third and fourth columns

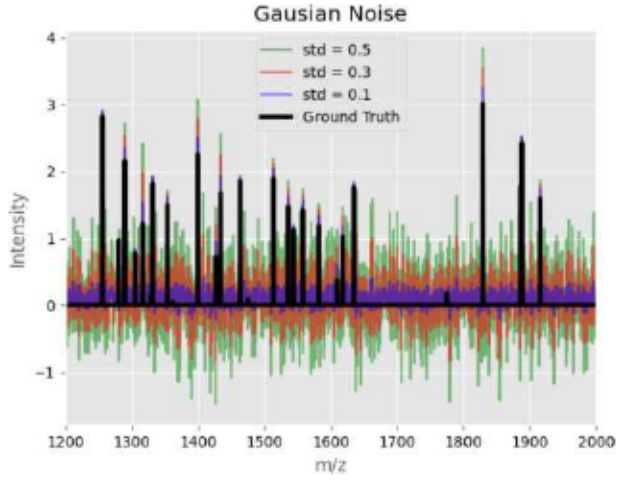


Fig. 8 An example of a synthetic spectrum of the images in Fig. 7. The ground truth spectrum is displayed in black, and its different noisy counterparts are displayed in color

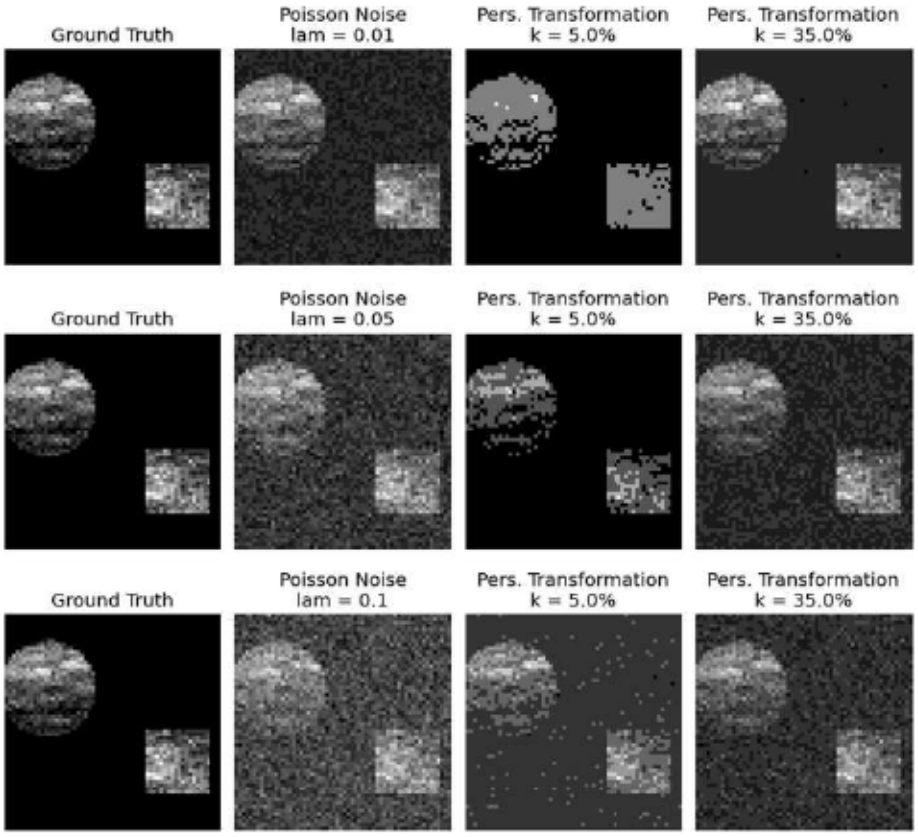


Fig. 9 Denoising with the persistence transformation. On the leftmost column, the ground truths of synthetic MALDI-Images are displayed. In the second column, distinct levels of Poisson noise are added to the spectral data. The processed images based on two choices of k are displayed in the third and fourth columns

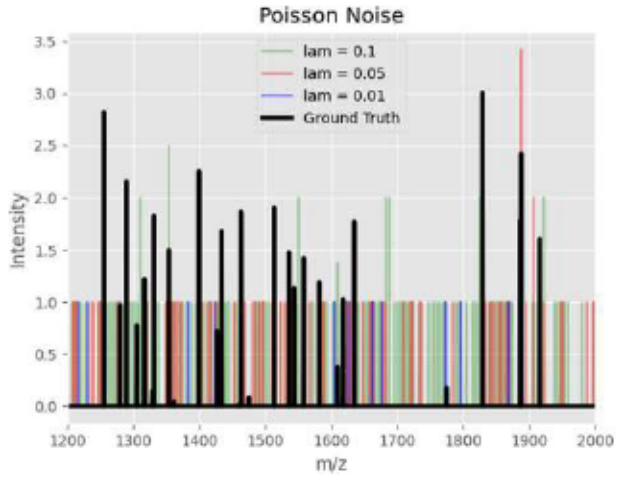


Fig. 10 An example of a synthetic spectrum of the images in Fig. 9. The ground truth spectrum is displayed in black, and its different noisy counterparts are displayed in color

shapes. An example of such kind of noise is the Poisson noise, and it is displayed in Figs. 9 and 10. In Fig. 10, it is shown that most of the noise spectra are exceeding the signal peaks. Even so, with an appropriate k value, the shapes of the ground truth can be reconstructed to some degree in Fig. 9.

More results can be found in the "Additional file 2". We want to highlight the time used for the (denoising) analysis. It can be seen that doubling the number of pixels results in doubled time for the algorithm, e.g., for Gaussian noise with a standard deviation of 0.1, the analysis of a 30×30 image takes approx. 29 seconds, for a 42×42 image it takes approx. 61 seconds, and for a 60×60 image it takes approx. 113 seconds on a standard computer. This illustrates the almost linearity of the implementation.

Discussion

Summary

Motivated by the MALDI classification studies, the objective of this study has been to propose a novel custom-made approach for modeling MALDI-MSI data. In general, the study's methodology consists of two steps. First, we carry out the introduced topological framework to obtain the intrinsic information from each mass spectrum, given thousands of m/z values. Generally speaking, this step can be considered as a data-compression method for MALDI-MSI data. Second, we execute two supervised classification methods based on the resulting persistence vectors so as to classify the observational units into lung cancer subtypes.

The usefulness of the proposed topological framework consists of three perspectives. First, our numerical classification results illustrate that the topological framework extracts the necessary information, which can be used for further classification tasks. The obtained results are competitive with other data-analysis methods for this dataset (cf. [9]). Second, the proposed framework compresses MALDI-MSI data, resulting in a significant computational gain for the RF classifier. Third, we have demonstrated its effectiveness in retrieving the informative parts of spectral signals under different noisy scenarios. The proposed topological framework can be adopted in a computationally efficient algorithm depending on a single tuning parameter, i.e., the fraction of used peaks.

Outlook

The persistence transformation is a novel tool for topological data analysis, which can be employed in different real-world applications. Future work to extend the introduced framework might include the application of the heat equation to reduce the impact of noisy peaks (see [24]), which might increase the classification performance. Alternatively, one can use the persistence pairs as input points for a second persistence homology analysis. However, any computational improvement would lead to the extension of the computational time. Furthermore, the information on the position of the peaks might be used for backtracking to identify (biologically) relevant molecules given as peaks. Finally, the algorithm used in this paper shows similarity to Morse Theory [52] and especially to the matching theorems in [20]. This presents an interesting topic for follow-up research.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05402-0.

Additional file 1. The file contains the pseudo-code of the introduced algorithm and the proof of its complexity and correctness.

Additional file 2. The file contains synthetic MALDI-images. Distinct types and levels of noise are added to the ground truth and displayed. Finally, the results of the denoising with the persistence transformation is depicted.

Acknowledgements
We thank the editor-in-chief, the associate editor, and two anonymous reviewers for their reading of the paper and for their constructive suggestions, which have improved the presentation. The authors extend their sincere gratitude to Prof. Dmitry Feichtner-Kozlov and Prof. Thorsten Dickhaus at the University of Bremen for supervising GK and VV. Also, we sincerely thank Lena Ranke, Friederike Preusse, and Lukas Mentz for their constructive criticism of the manuscript. The financial support by the German Research Foundation via the RTG 2224, titled "π^3: Parameter Identification—Analysis, Algorithms, Implementations" is gratefully acknowledged.

Author contributions
This paper is based on joined work of the three main authors as equal contributors. GK implemented the algorithm and proved the complexity. VV took care of the statistical treatment. AS helped clarify the notion of persistent transformation and its connection to the persistence diagram via the elder rule and proposed examples illustrating this connection. All authors have written the manuscript and approved the final version.

Funding
Open Access funding enabled and organized by Projekt DEAL. Deutsche Forschungsgemeinschaft. Grant Number: RTG 2224

Availability of data and materials
Python code with the numerical results presented in the paper, including the MALDI data, is available at: https://github.com/klailag/SupervisedTDAMethodForMALDI

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 24 February 2023 Accepted: 26 June 2023
Published online: 10 July 2023

References

1. Mortier T, Wieme AD, Vandamme P, Waegeman W. Bacterial species identification using MALDI-TOF mass spectrometry and machine learning techniques: a large-scale benchmarking study. Comput Struct Biotechnol J. 2021;19:6157–68. https://doi.org/10.1016/j.csbj.2021.11.004.
2. Caprioli RM, Farmer TB, Gile J. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. Anal Chem. 1997;69(23):4751–60.
3. Alexandrov T, Bartels A. Testing for presence of known and unknown molecules in imaging mass spectrometry. Bioinformatics (Oxford, England). 2013;29(18):2335–42. https://doi.org/10.1093/bioinformatics/btt388.
4. Boskamp T, Lachmund D, Oetjen J, Cordero Hernandez Y, Trede D, Maass P, Casadonte R, Kriegsmann J, Warth A, Dienemann H, Weichert W, Kriegsmann M. A new classification method for MALDI imaging mass spectrometry data acquired on formalin-fixed paraffin-embedded tissue samples. Biochim Biophys Acta Proteins Proteom. 2017;1865(7):916–26. https://doi.org/10.1016/j.bbapap.2016.11.003.
5. Alexandrov T. MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. BMC Bioinform. 2012;13(Suppl 16):11. https://doi.org/10.1186/1471-2105-13-S16-S11.
6. Behrmann J, Etmann C, Boskamp T, Casadonte R, Kriegsmann J, Maaß P. Deep learning for tumor classification in imaging mass spectrometry. Bioinformatics. 2018;34(7):1215–23. https://doi.org/10.1093/bioinformatics/btx724.
7. Kriegsmann J, Kriegsmann M, Casadonte R. MALDI TOF imaging mass spectrometry in clinical pathology: a valuable tool for cancer diagnostics (review). Int J Oncol. 2015;46(3):893–906. https://doi.org/10.3892/ijo.2014.2788.
8. Veselkov KA, Mirnezami R, Strittmatter N, Goldin RD, Kinross J, Speller AVM, Abramov T, Jones EA, Darzi A, Holmes E, Nicholson JK, Takats Z. Chemo-informatic strategy for imaging mass spectrometry-based hyperspectral profiling of lipid signatures in colorectal cancer. Proc Natl Acad Sci. 2014;111(3):1216–21. https://doi.org/10.1073/pnas.1310524111.

9. Leuschner J, Schmidt M, Fernsel P, Lachmund D, Boskamp T, Maass P. Supervised non-negative matrix factorization methods for MALDI imaging applications. Bioinformatics. 2019;35(11):1940–7. https://doi.org/10.1093/bioinformatics/bty909.
10. Vutov V, Dickhaus T. Multiple two-sample testing under arbitrary covariance dependency with an application in imaging mass spectrometry. Biom J. 2023;65(2):2100328.
11. Wijetunge CD, Saeed I, Boughton BA, Roessner U, Halgamuge SK. A new peak detection algorithm for MALDI mass spectrometry data based on a modified Asymmetric Pseudo-Voigt model. BMC Genomics. 2015;16(Suppl 12):12. https://doi.org/10.1186/1471-2164-16-S12-S12.
12. Timm W, Scherbart A, Böcker S, Kohlbacher O, Nattkemper TW. Peak intensity prediction in MALDI-TOF mass spectrometry: a machine learning study to support quantitative proteomics. BMC Bioinform. 2008. https://doi.org/10.1186/1471-2105-9-443.
13. Yang C, He Z, Yu W. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. BMC Bioinform. 2009. https://doi.org/10.1186/1471-2105-10-4.
14. Lieb F, Boskamp T, Stark HG. Peak detection for MALDI mass spectrometry imaging data using sparse frame multipliers. J Proteomics. 2020;225: 103852. https://doi.org/10.1016/j.jprot.2020.103852.
15. Slawski M, Hussong R, Tholey A, Jakoby T, Gregorius B, Hildebrandt A, Hein M. Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching. BMC Bioinform. 2012;13:291. https://doi.org/10.1186/1471-2105-13-291.
16. von Schroeder J. Stable Feature Selection with Applications to MALDI Imaging Mass Spectrometry Data. Preprint; 2020, available via arXiv:2006.15077.
17. Vutov V, Dickhaus T. Multiple multi-sample testing under arbitrary covariance dependency. Stat Med. 2023. https://doi.org/10.1002/sim.9761.
18. Otter N, Porter MA, Tillmann U, Grindrod P, Harrington HA. A roadmap for the computation of persistent homology. EPJ Data Sci. 2017;6(1):17. https://doi.org/10.1140/epjds/s13688-017-0109-5.
19. Chazal F, Michel B. An introduction to topological data analysis: fundamental and practical aspects for data scientists. Front Artif Intell. 2021;4: 667963. https://doi.org/10.3389/frai.2021.667963.
20. Kozlov DN. A combinatorial method to compute explicit homology cycles using Discrete Morse Theory. J Appl Comput Topol. 2020;4(1):79–100. https://doi.org/10.1007/s41468-019-00042-x.
21. Skaf Y, Laubenbacher RC. Topological data analysis in biomedicine: a review. J Biomed Inform. 2022;130:104082. https://doi.org/10.1016/j.jbi.2022.104082.
22. Bukkuri A, Andor N, Darcy IK. Applications of topological data analysis in oncology. Front Artif Intell. 2021;4: 659037. https://doi.org/10.3389/frai.2021.659037.
23. Loughrey CF, Fitzpatrick P, Orr N, Jurek-Loughrey A. The topology of data: opportunities for cancer research. Bioinformatics. 2021;37(19):3091–8. https://doi.org/10.1093/bioinformatics/btab553.
24. Weis C, Horn M, Rieck B, Cuénod A, Egli A, Borgwardt KM. Topological and kernel-based microbial phenotype prediction from MALDI-TOF mass spectra. Bioinformatics. 2020;36(Supplement–1):30–8. https://doi.org/10.1093/bioinformatics/btaa429.
25. Fernsel P. Spatially coherent clustering based on orthogonal nonnegative matrix factorization. J Imaging. 2021;7(10):194.
26. Edelsbrunner H, Harer JL. Computational topology: an introduction. Providence: American Mathematical Society; 2010.
27. Fasy B, Lecci F, Rinaldo A, Wasserman L, Balakrishnan S, Singh A. Statistical inference for persistent homology: confidence sets for persistence diagrams. 2013. https://doi.org/10.1214/14-AOS1252.
28. Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. Discret Comput Geom. 2002;28(4):511–33. https://doi.org/10.1007/s00454-002-2885-2.
29. Cohen-Steiner D, Edelsbrunner H, Harer J. Stability of persistence diagrams. Discret Comput Geom. 2007;37(1):103–20. https://doi.org/10.1007/s00454-006-1276-5.
30. Vlontzos A, Cao Y, Schmidtke L, Kainz B, Monod A. Topological data analysis of database representations for information retrieval. CoRR; 2021. arXiv:2104.01672.
31. Grélard F, Legland D, Fanuel M, Arnaud B, Foucat L, Rogniaux H. Esmraldi: efficient methods for the fusion of mass spectrometry and magnetic resonance images. BMC Bioinform. 2021;22(1):56. https://doi.org/10.1186/s12859-020-03954-z.
32. Contessoto M, Mémoli F, Stefanou A, Zhou L. Persistent cup-length; 2021. arXiv preprint arXiv:2107.01553.
33. Mémoli F, Stefanou A, Zhou L. Persistent cup product structures and related invariants; 2022. arXiv preprint arXiv:2211.16642
34. Couronné R, Probst P, Boulesteix A. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinform. 2018;19(1):270–127014. https://doi.org/10.1186/s12859-018-2264-5.
35. Kirasich K, Smith T, Sadler B. Random forest vs logistic regression: binary classification for heterogeneous datasets. SMU Data Sci Rev. 2018;1(3):9.
36. Fahrmeir L, Tutz G, Hennevogl W, Salem E. Multivariate statistical modelling based on generalized linear models, vol. 425. Berlin: Springer; 1994.
37. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
38. Pal M. Random forest classifier for remote sensing classification. Int J Remote Sens. 2005;26(1):217–22.
39. Noshad Z, Javaid N, Saba T, Wadud Z, Saleem MQ, Alzahrani ME, Sheta OE. Fault detection in wireless sensor networks through the random forest classifier. Sensors. 2019;19(7):1568. https://doi.org/10.3390/s19071568.
40. Shrestha B, Stephen H, Ahmad S. Impervious Surfaces Mapping at City Scale by Fusion of Radar and Optical data through a random forest classifier. Remote Sens. 2021;13(15):3040. https://doi.org/10.3390/rs13153040.
41. Neumann JM, Freitag H, Hartmann JS, Niehaus K, Galanis M, Griesshammer M, Kellner U, Bednarz H. Subtyping non-small cell lung cancer by histology-guided spatial metabolomics. J Cancer Res Clin Oncol. 2022;148(2):351–60.

42. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
43. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer Series in Statistics. Springer, New York; 2009. https://doi.org/10.1007/978-0-387-84858-7.
44. Liaw A, Wiener M, et al. Classification and regression by randomforest. R News. 2002;2(3):18–22.
45. Probst P, Wright MN, Boulesteix A. Hyperparameters and tuning strategies for random forest. WIREs Data Min Knowl Discov. 2019. https://doi.org/10.1002/widm.1301.
46. Sugár S, Bugyi F, Tóth G, Pápay J, Kovalszky I, Tornóczky T, Drahos L, Turiák L. Proteomic analysis of lung cancer types–a pilot study. Cancers. 2022. https://doi.org/10.3390/cancers14112629.
47. Kriegsmann M, Casadonte R, Kriegsmann J, Dienemann H, Schirmacher P, Hendrik Kobarg J, Schwamborn K, Stenzinger A, Warth A, Weichert W. Reliable entity subtyping in non-small cell lung cancer by matrix-assisted laser desorption/ionization imaging mass spectrometry on formalin-fixed paraffin-embedded tissue specimens. Mol Cell Proteomics. 2016;15(10):3081–9. https://doi.org/10.1074/mcp.m115.057513.
48. Krutchinsky AN, Chait BT. On the nature of the chemical noise in MALDI mass spectra. J Am Soc Mass Spectrom. 2002;13(2):129–34.
49. Trede D, Kobarg JH, Oetjen J, Thiele H, Maass P, Alexandrov T. On the importance of mathematical methods for analysis of MALDI-imaging mass spectrometry data. J Integr Bioinform (JIB). 2012;9(1):1–11.
50. Bemis KD, Harry A, Eberlin LS, Ferreira C, van de Ven SM, Mallick P, Stolowitz M, Vitek O. Cardinal: an R package for statistical analysis of mass spectrometry-based imaging experiments. Bioinformatics. 2015. https://doi.org/10.1093/bioinformatics/btv146.
51. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria; 2021. R Foundation for Statistical Computing. https://www.R-project.org/.
52. Milnor J. Morse Theory. (AM-51), Volume 51. Princeton University Press, Princeton; 1963. https://doi.org/10.1515/9781400881802.

Publisher's Note

Additional file 1 — The introduced algorithm

This additional file provides the pseudo-code of the algorithm introduced in the paper *Supervised topological data analysis for MALDI mass spectrometry imaging applications*. Furthermore, it contains the proof of Theorem 1 of the paper.

Algorithm 1 Recursion start

Input: $[f(x_0), \dots, f(x_{q-1})]$
Return: $[(\hat{x}, p(\hat{x})), \dots]$
maxima, minima, featurePairs $\leftarrow \varnothing$
for all $x_j \in [x_0, x_{q-1}]$ **do**
  if x_j is maximum **then**
    maxima $\leftarrow (x_j, f(x_j))$
  else if x_j is minimum **then**
    minima $\leftarrow (x_j, f(x_j))$
SORT(maxima, f(x), $>$)
SORT(minima, f(x), $<$)
$(\hat{x}, f(\hat{x})) \leftarrow$ maxima.pop(0)
featurePairs $\leftarrow (\hat{x}, f(\hat{x}) -$ minima[0][1])
RecursionStep($x_0, \hat{x}$, maxima.copy(), minima.copy(), featurePairs)
RecursionStep($x_n, \hat{x}$, maxima.copy(), minima.copy(), featurePairs)
return featurePairs

Algorithm 2 Recursion Step

Input: start, end, maxima, minima, featurePairs
for all $(x_j, f(x_j)) \in$ maxima **do**
  if $x_j \notin [\text{start}, \text{end})$ **then**
    maxima $\leftarrow$ maxima$\backslash(x_j, f(x_j))$
if $|\text{maxima}| = 0$ **then**
  return
$(\hat{x}, f(\hat{x})) \leftarrow$ maxima.pop(0)
RecursionStep(start, $\hat{x}$, maxima.copy(), minima.copy(), featurePairs)
for all $(x_j, f(x_j)) \in$ minima **do**
  if $x_j \notin (\hat{x}, \text{end})$ **then**
    minima $\leftarrow$ minima$\backslash(x_j, f(x_j))$
$(x', f(x')) \leftarrow$ minima.pop(0)
featurePairs $\leftarrow (\hat{x}, f(\hat{x}) - f(x'))$
RecursionStep(x', $\hat{x}$, maxima.copy(), minima.copy(), featurePairs)
RecursionStep(x', end, maxima.copy(), minima.copy(), featurePairs)

Proof of Theorem 1: The algorithm to calculate the reduced persistence transformation is divided into two parts: the recursion start (Algorithm 1) and the recursion step (Algorithm 2).

The *recursion start* (Algorithm 1) gets as input the list of all the intensity values for each m/z value, marked as $f(x_j)$. Let m be the number of maxima in this mass spectrum. In the beginning, empty lists are created for the maxima, the minima, and the results (called "featurePairs"). The latter list stores the x value, i.e., the position, as well as the persistence of each peak. Notice that each recursion step updates the list instead of returning results.

In the next step, all the minima and the maxima are stored in the corresponding lists in tuples of the form $(x, f(x))$. For this, the algorithm iterates through the list of the $f(x_j)$. If the value is larger than its neighbors, it is marked as maximum and stored in the corresponding list. Correspondingly, values that are smaller than their neighbors are marked as a minimum. This identification of extremal points can be made in linear run-time since the list is traversed just once, resulting in a complexity of $\sigma(q)$. The two lists *maxima* and *minima* are then sorted by their corresponding value $f(x_j)$ (the *minima* list inverted) with a complexity of $\sigma(m \cdot \log m)$.

For the largest feature, the global maximum, the persistence is defined to be the difference to the global minimum (see Equation (1) of the paper *Supervised topological data analysis for MALDI mass spectrometry imaging applications*). These values are the first elements of their corresponding lists. After calculating the persistence, the maximum is removed from the list, and the found feature $(x, p(x))$ is stored in the list *persistencePairs*. The *recursion step* is called afterwards with the intervals $[x_0, \hat{x})$ and $[x_{q-1}, \hat{x})$. Notice that the second interval is reversed. As input, the *recursion step* gets a copy of the two lists minima and maxima as well as the original list *featurePairs*. All these computations can be done in constant time, i.e., $\sigma(1)$. After the last *recursion step*, all the features are detected and stored in the *featurePairs* list and can be returned.

The input for the *recursion step* (Algorithm 2) consists of two indices, namely *start* and *end* (indicating the part of the data which is processed in the current recursion step, i.e., the positions of m/z values), a list of *maxima* and a list of *minima*, and the shared list of *featurePairs*. In the first step, the routine removes all the maxima not in the currently processed part of the data. There are at most m elements in the *maximum* list, so the complexity of this task is $\sigma(m)$. If the list is empty after the removing step, the *recursion step* reaches the end and can return. If not, it removes the first element $\hat{x}$ from the list. This is the most persistent feature (in terms of topology) in the processed part of the data. The elder rule (cf. [1]) states that the feature can only merge with a feature with a larger persistence, which is per construction at the index *end*. The corresponding minimum (cf. Equation (2) of the paper *Supervised topological data analysis for MALDI mass spectrometry imaging applications*) to $\hat{x}$ can only be in the interval $(\hat{x}, \mathit{end})$ so the *minima* list can be filtered in a similar fashion to the *maxima* list with the same complexity. Since there are more possible features in the interval $(\mathit{start}, \hat{x})$, the *recursion step* is called once more for this interval.

The values $\hat{x}$ and the smallest minimum x' from the *minima* list generate a topological feature. This feature is updated in the original *featurePairs* list, and the recursion step can be repeated with the two intervals $(x', \hat{x})$ and (x', end).

The recursion step is processed at least once for each maximum with three extra calls after no more maxima are left, i.e., it runs at most $4 \times m$ times. Given the complexity of each step of $\sigma(m)$, the complexity of all the recursion steps together is $\sigma(m^2)$. This gives an overall complexity of the algorithm of

$$\sigma(q) + \sigma(m^2) + \sigma(m \cdot \log m) = \sigma(q) + \sigma(m^2).$$

For each maximum, there is a tuple stored which contains the information of the position and the persistence, resulting in an overall storage use of $2m$ elements.

The algorithm always terminates since, at each recursion step, one maximum is removed from the list of maxima —if it is not already empty. Likewise, each part of the input list is being processed. Since there is only a finite number of elements in the *maxima* list (i.e., m), the algorithm terminates after all are processed. Even more, the algorithm returns all the features with their persistence. Each maximum creates a feature, and all maxima are processed. They are paired with the correct minimum between themselves and a feature with a higher persistence according to the elder rule (see [1]). Hence, the algorithm always terminates and returns the correct solution in $\sigma(q)+\sigma(m^2)$ run-time. □

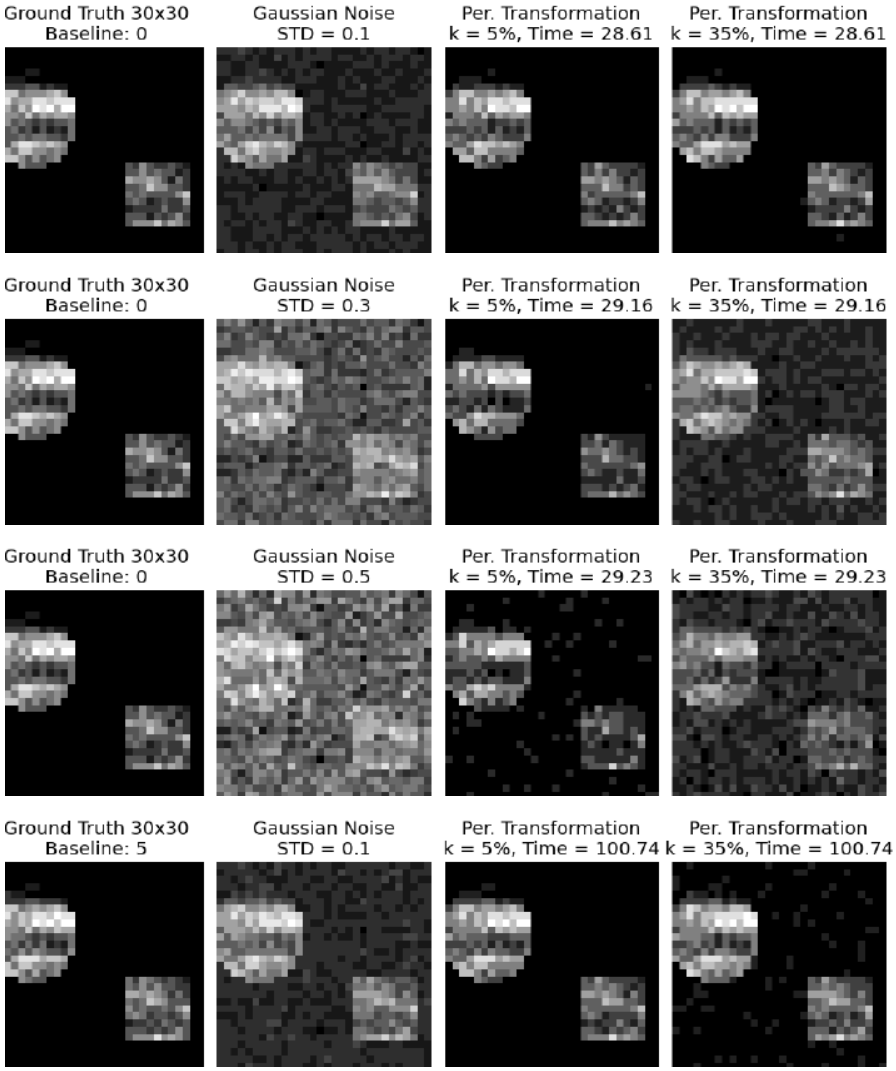
References

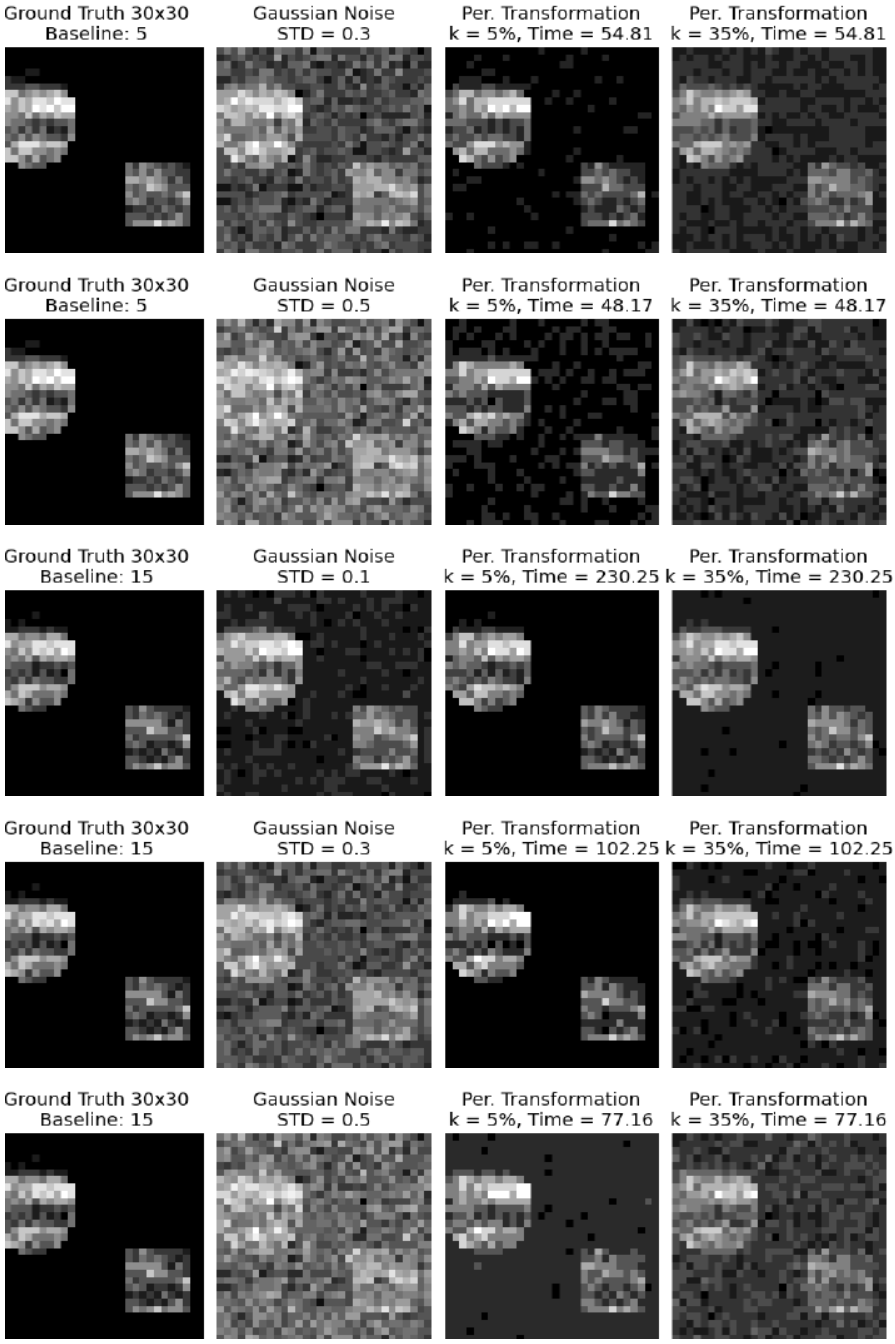
[1] Edelsbrunner, H., Harer, J.L.: Computational Topology: an Introduction. American Mathematical Society, Providence, USA (2010)

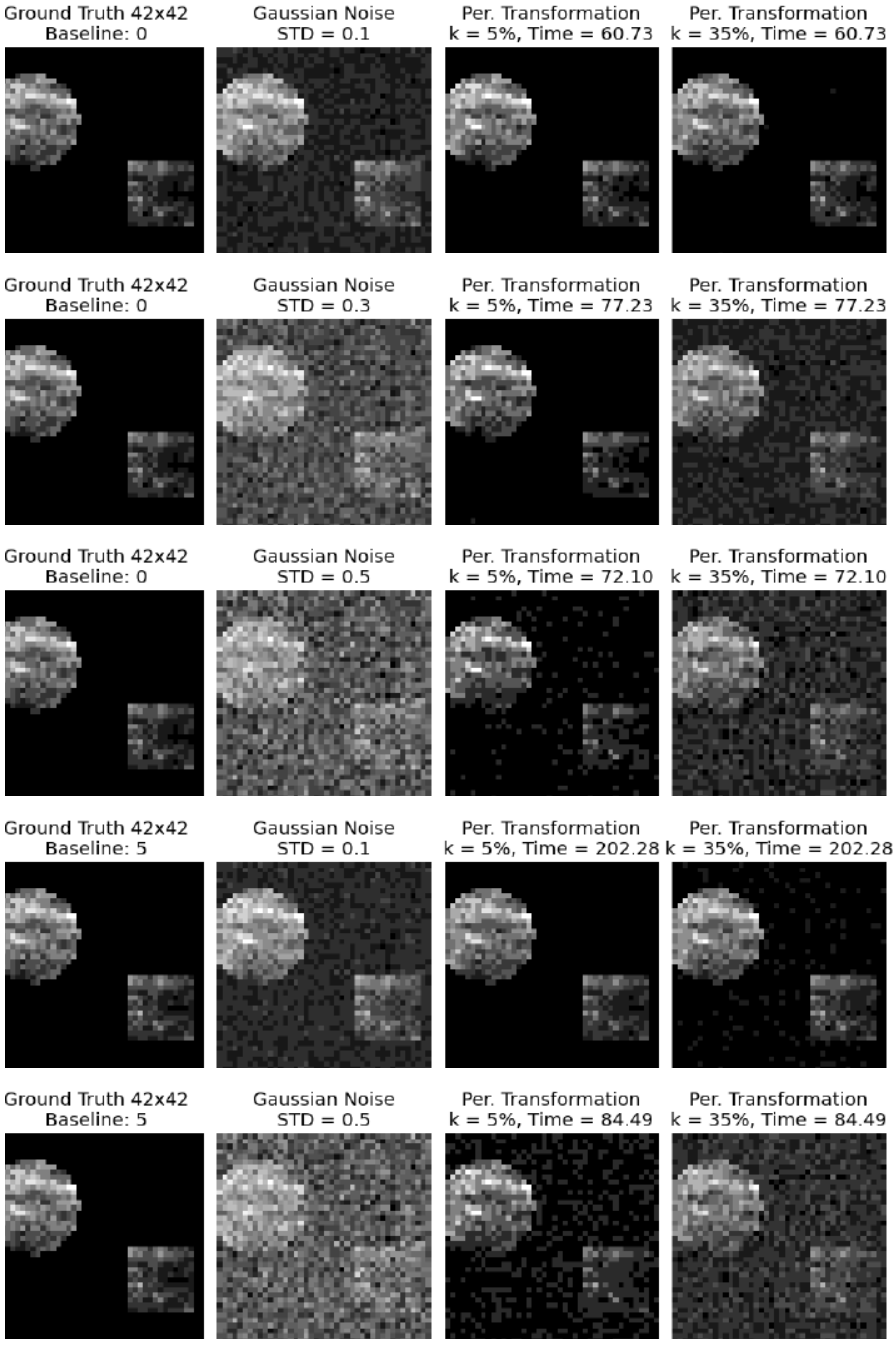
Additional file 2 — Additional simulation results

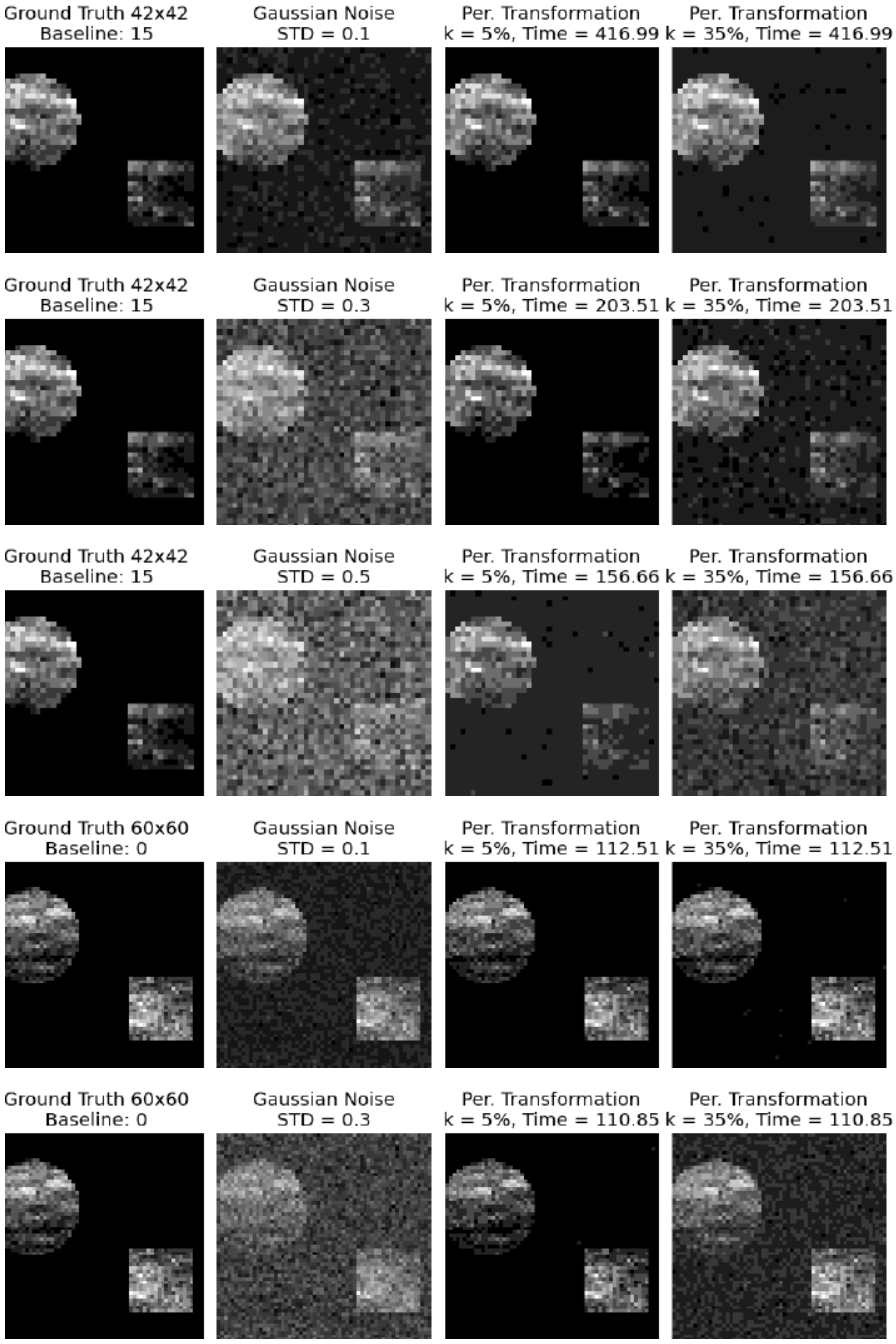
The additional file contains synthetic MALDI-images. Distinct types and levels of noise are added to the ground truth and displayed. Finally, the results of the denoising with the persistence transformation is depicted.

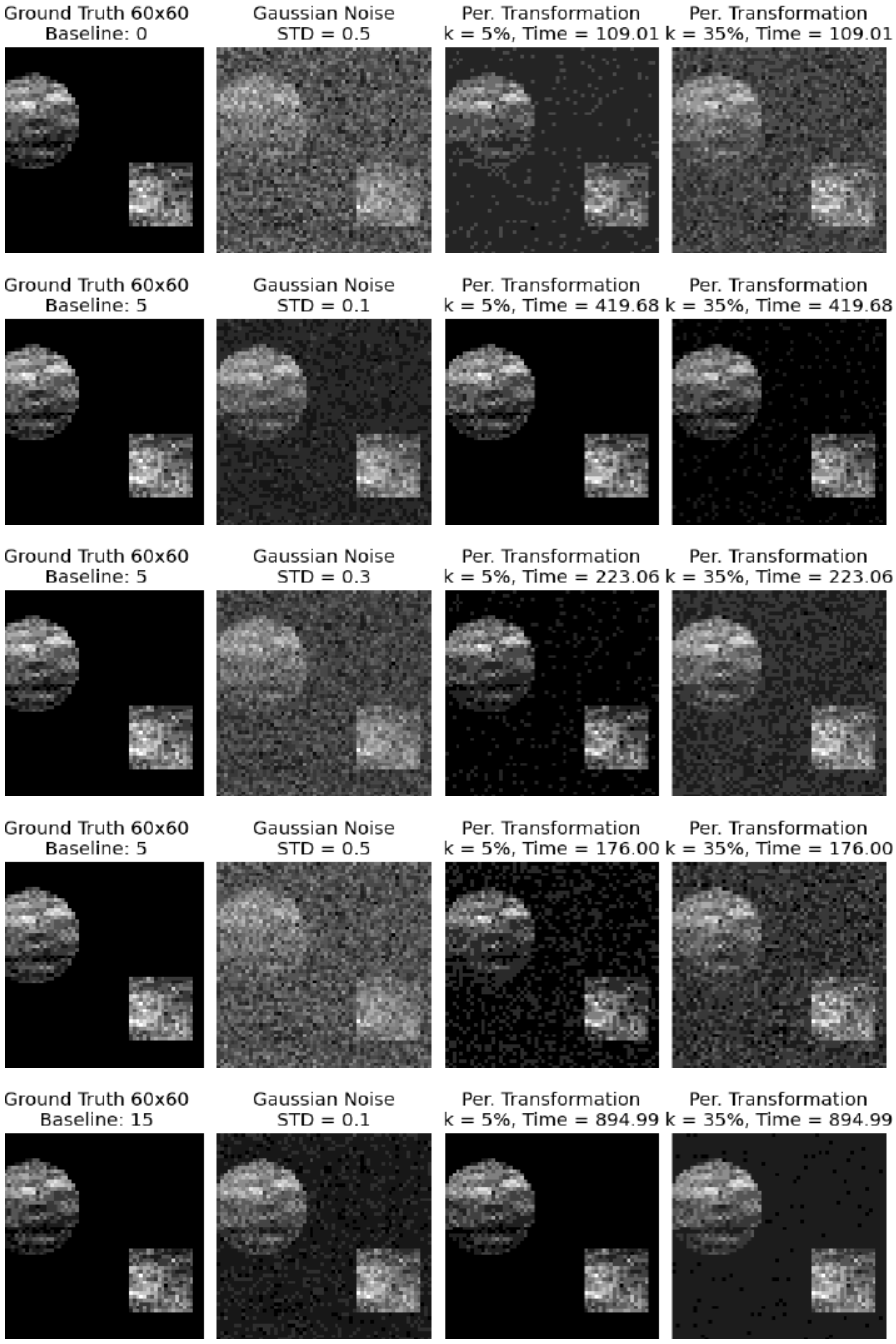
Gaussian Noise

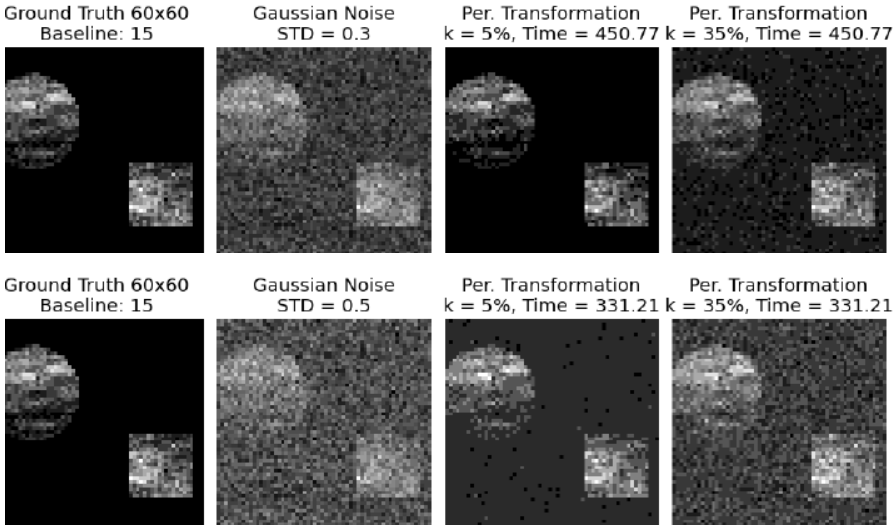



Ground Truth 30x30
Baseline: 5
Gaussian Noise
STD = 0.3
Per. Transformation
k = 5%, Time = 54.81
Per. Transformation
k = 35%, Time = 54.81
Ground Truth 30x30
Baseline: 5
Gaussian Noise
STD = 0.5
Per. Transformation
k = 5%, Time = 48.17
Per. Transformation
k = 35%, Time = 48.17
Ground Truth 30x30
Baseline: 15
Gaussian Noise
STD = 0.1
Per. Transformation
k = 5%, Time = 230.25
Per. Transformation
k = 35%, Time = 230.25
Ground Truth 30x30
Baseline: 15
Gaussian Noise
STD = 0.3
Per. Transformation
k = 5%, Time = 102.25
Per. Transformation
k = 35%, Time = 102.25
Ground Truth 30x30
Baseline: 15
Gaussian Noise
STD = 0.5
Per. Transformation
k = 5%, Time = 77.16
Per. Transformation
k = 35%, Time = 77.16

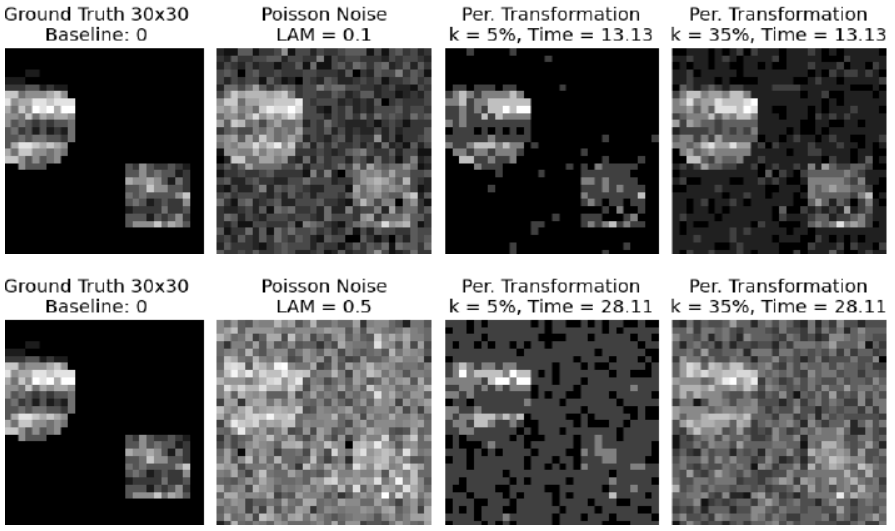

Ground Truth 42x42
Baseline: 0
Gaussian Noise
STD = 0.1
Per. Transformation
k = 5%, Time = 60.73
Per. Transformation
k = 35%, Time = 60.73
Ground Truth 42x42
Baseline: 0
Gaussian Noise
STD = 0.3
Per. Transformation
k = 5%, Time = 77.23
Per. Transformation
k = 35%, Time = 77.23
Ground Truth 42x42
Baseline: 0
Gaussian Noise
STD = 0.5
Per. Transformation
k = 5%, Time = 72.10
Per. Transformation
k = 35%, Time = 72.10
Ground Truth 42x42
Baseline: 5
Gaussian Noise
STD = 0.1
Per. Transformation
k = 5%, Time = 202.28
Per. Transformation
k = 35%, Time = 202.28
Ground Truth 42x42
Baseline: 5
Gaussian Noise
STD = 0.5
Per. Transformation
k = 5%, Time = 84.49
Per. Transformation
k = 35%, Time = 84.49

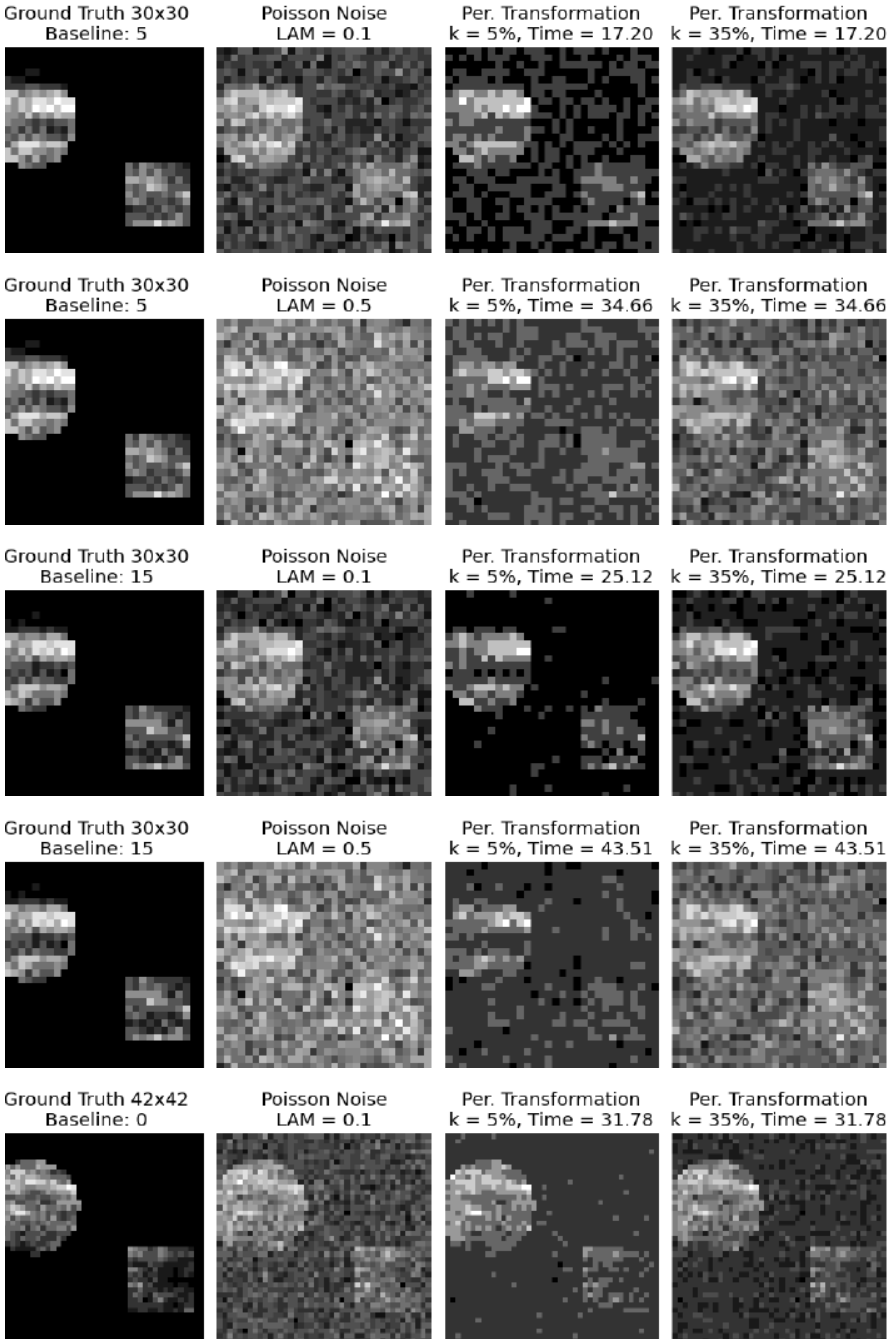

Ground Truth 42x42
Baseline: 15
Gaussian Noise
STD = 0.1
Per. Transformation
k = 5%, Time = 416.99
Per. Transformation
k = 35%, Time = 416.99
Ground Truth 42x42
Baseline: 15
Gaussian Noise
STD = 0.3
Per. Transformation
k = 5%, Time = 203.51
Per. Transformation
k = 35%, Time = 203.51
Ground Truth 42x42
Baseline: 15
Gaussian Noise
STD = 0.5
Per. Transformation
k = 5%, Time = 156.66
Per. Transformation
k = 35%, Time = 156.66
Ground Truth 60x60
Baseline: 0
Gaussian Noise
STD = 0.1
Per. Transformation
k = 5%, Time = 112.51
Per. Transformation
k = 35%, Time = 112.51
Ground Truth 60x60
Baseline: 0
Gaussian Noise
STD = 0.3
Per. Transformation
k = 5%, Time = 110.85
Per. Transformation
k = 35%, Time = 110.85

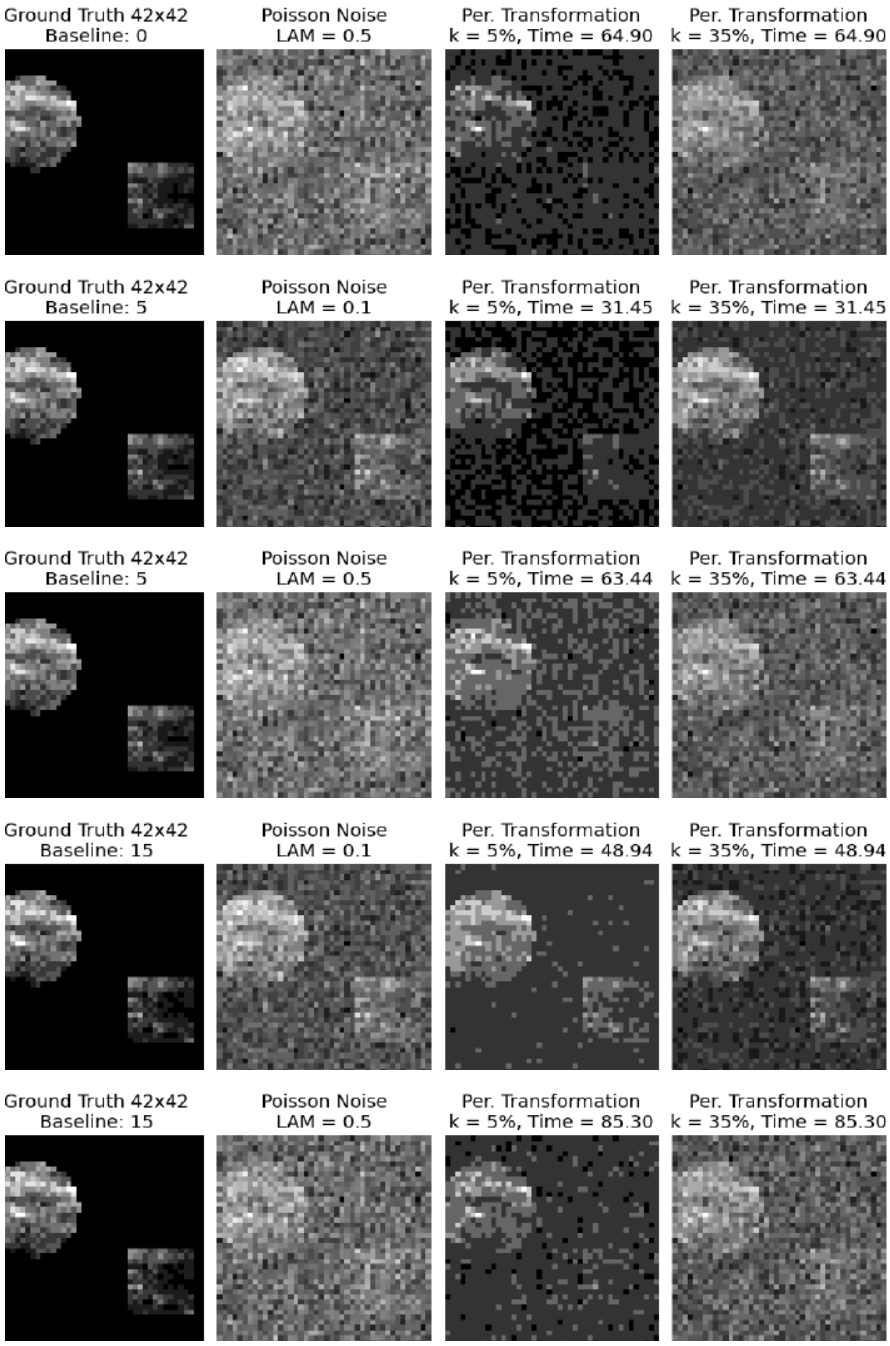

Ground Truth 60x60
Baseline: 0
Gaussian Noise
STD = 0.5
Per. Transformation
k = 5%, Time = 109.01
Per. Transformation
k = 35%, Time = 109.01
Ground Truth 60x60
Baseline: 5
Gaussian Noise
STD = 0.1
Per. Transformation
k = 5%, Time = 419.68
Per. Transformation
k = 35%, Time = 419.68
Ground Truth 60x60
Baseline: 5
Gaussian Noise
STD = 0.3
Per. Transformation
k = 5%, Time = 223.06
Per. Transformation
k = 35%, Time = 223.06
Ground Truth 60x60
Baseline: 5
Gaussian Noise
STD = 0.5
Per. Transformation
k = 5%, Time = 176.00
Per. Transformation
k = 35%, Time = 176.00
Ground Truth 60x60
Baseline: 15
Gaussian Noise
STD = 0.1
Per. Transformation
k = 5%, Time = 894.99
Per. Transformation
k = 35%, Time = 894.99

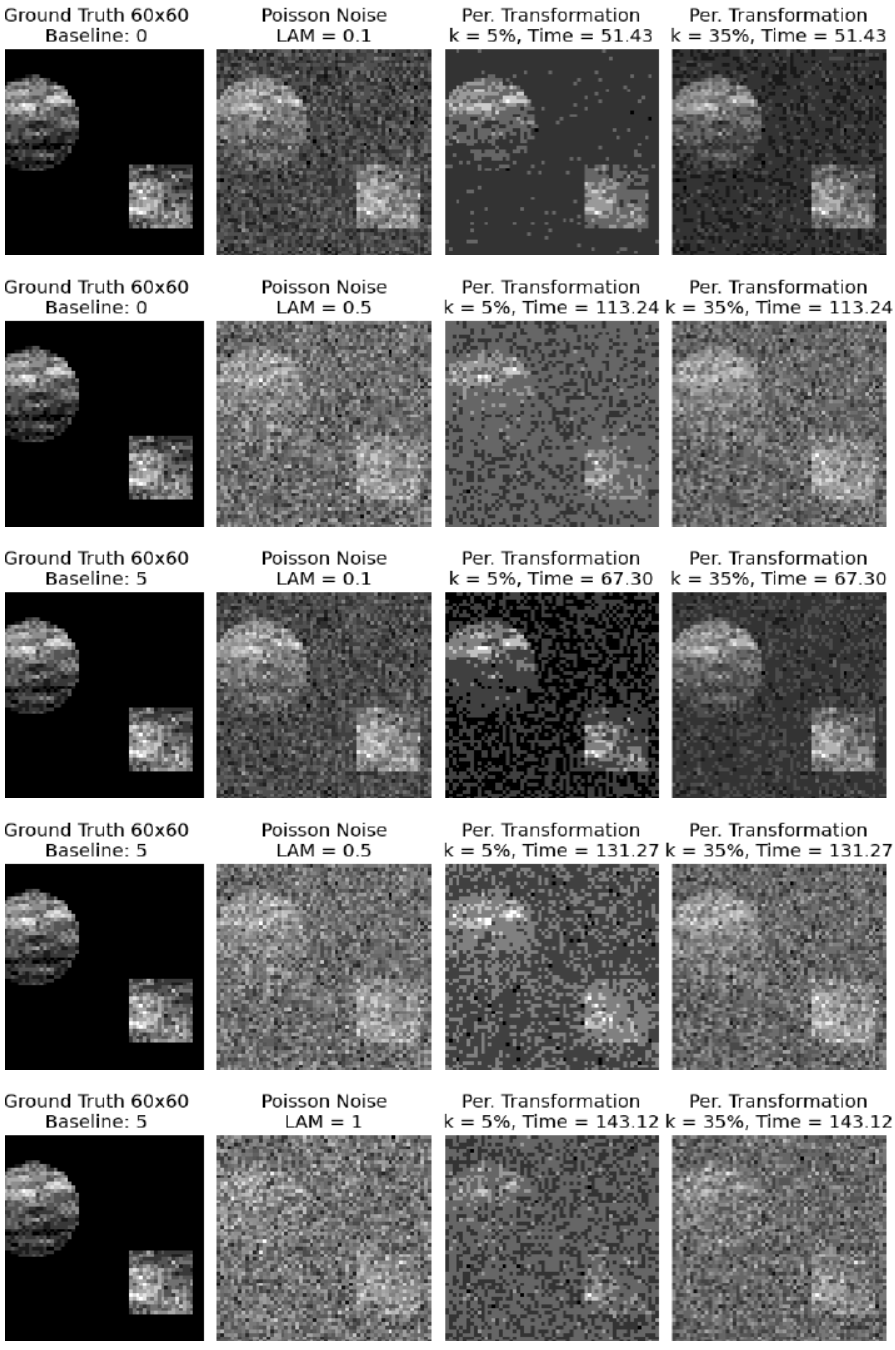


Poisson Noise




Ground Truth 30x30
Baseline: 5
Poisson Noise
LAM = 0.1
Per. Transformation
k = 5%, Time = 17.20
Per. Transformation
k = 35%, Time = 17.20
Ground Truth 30x30
Baseline: 5
Poisson Noise
LAM = 0.5
Per. Transformation
k = 5%, Time = 34.66
Per. Transformation
k = 35%, Time = 34.66
Ground Truth 30x30
Baseline: 15
Poisson Noise
LAM = 0.1
Per. Transformation
k = 5%, Time = 25.12
Per. Transformation
k = 35%, Time = 25.12
Ground Truth 30x30
Baseline: 15
Poisson Noise
LAM = 0.5
Per. Transformation
k = 5%, Time = 43.51
Per. Transformation
k = 35%, Time = 43.51
Ground Truth 42x42
Baseline: 0
Poisson Noise
LAM = 0.1
Per. Transformation
k = 5%, Time = 31.78
Per. Transformation
k = 35%, Time = 31.78


Ground Truth 42x42
Baseline: 0
Poisson Noise
LAM = 0.5
Per. Transformation
k = 5%, Time = 64.90
Per. Transformation
k = 35%, Time = 64.90
Ground Truth 42x42
Baseline: 5
Poisson Noise
LAM = 0.1
Per. Transformation
k = 5%, Time = 31.45
Per. Transformation
k = 35%, Time = 31.45
Ground Truth 42x42
Baseline: 5
Poisson Noise
LAM = 0.5
Per. Transformation
k = 5%, Time = 63.44
Per. Transformation
k = 35%, Time = 63.44
Ground Truth 42x42
Baseline: 15
Poisson Noise
LAM = 0.1
Per. Transformation
k = 5%, Time = 48.94
Per. Transformation
k = 35%, Time = 48.94
Ground Truth 42x42
Baseline: 15
Poisson Noise
LAM = 0.5
Per. Transformation
k = 5%, Time = 85.30
Per. Transformation
k = 35%, Time = 85.30


Ground Truth 60x60
Baseline: 0
Poisson Noise
LAM = 0.1
Per. Transformation
k = 5%, Time = 51.43
Per. Transformation
k = 35%, Time = 51.43
Ground Truth 60x60
Baseline: 0
Poisson Noise
LAM = 0.5
Per. Transformation
k = 5%, Time = 113.24
Per. Transformation
k = 35%, Time = 113.24
Ground Truth 60x60
Baseline: 5
Poisson Noise
LAM = 0.1
Per. Transformation
k = 5%, Time = 67.30
Per. Transformation
k = 35%, Time = 67.30
Ground Truth 60x60
Baseline: 5
Poisson Noise
LAM = 0.5
Per. Transformation
k = 5%, Time = 131.27
Per. Transformation
k = 35%, Time = 131.27
Ground Truth 60x60
Baseline: 5
Poisson Noise
LAM = 1
Per. Transformation
k = 5%, Time = 143.12
Per. Transformation
k = 35%, Time = 143.12