



UNIwersytet Jagielloński
w Krakowie

Wydział Fizyki, Astronomii i Informatyki Stosowanej

**Obrazowanie w podczerwieni w klasyfikacji
histopatologicznej i orientacji makromolekuł,
wsparte zoptymalizowanymi metodami wstępnego
przetwarzania danych**

Paulina Koziol

Rozprawa doktorska napisana pod kierunkiem:

dr hab. Tomasza P. Wróbla

prof. dr hab. Pawła Koreckiego

Kraków, 2023

Oświadczenie

Ja, niżej podpisana, Paulina Koziół (nr albumu: 1177147), doktorantka Wydziału Fizyki, Astronomii i Informatyki Stosowanej Uniwersytetu Jagiellońskiego oświadczam, że przedłożona przeze mnie rozprawa doktorska pt. *„Obrazowanie w podczerwieni w klasyfikacji histopatologicznej i orientacji makromolekuł, wsparte zoptymalizowanymi metodami wstępnego przetwarzania danych”* jest oryginalna i przedstawia wyniki badań wykonanych przeze mnie osobiście, pod kierunkiem dr. hab. Tomasza P. Wróbla oraz prof. dr hab. Pawła Koreckiego. Pracę napisałam samodzielnie.

Oświadczam, że moja rozprawa doktorska została opracowana zgodnie z Ustawą o prawie autorskim i prawach pokrewnych z dnia 4 lutego 1994 r. (Dziennik Ustaw 1994 nr 24 poz. 83 wraz z późniejszymi zmianami).

Jestem świadoma, że niezgodność niniejszego oświadczenia z prawdą ujawniona w dowolnym czasie, niezależnie od skutków prawnych wynikających z ww. ustawy, może spowodować unieważnienie stopnia nabytego na podstawie tej rozprawy.

Kraków, ...

(data)

.....

(podpis doktorantki)

Podziękowania

Pragnę złożyć najszczerze podziękowania dla mojego promotora, dr hab. Tomasza P. Wróbla, za czas poświęcony na wprowadzenie mnie do tematyki spektroskopii w podczerwieni oraz analizy danych, pokazanie możliwości tych dziedzin, a finalnie, za wzbudzenie we mnie pasji, która zmotywowała mnie do podjęcia studiów doktoranckich. Dziękuję również za wszystkie cenne dyskusje, motywację oraz wsparcie, zarówno związane z przygotowaniem rozprawy, jak i w trakcie całej kariery naukowej.

Serdecznie dziękuję również promotorowi, prof. dr hab. Pawłowi Koreckiemu, za trzy lata współpracy i wsparcie, jakie okazał w przygotowaniu publikacji, rozprawy doktorskiej oraz będąc godnym przykładem dydaktyka.

Dziękuję również członkom zespołu linii badawczej CIRI, w szczególności mgr Danucie Liberdzie, za wiele wspólnych godzin spędzonych w laboratorium, owocną współpracę i wzajemną motywację.

Dziękuję Fundacji na rzecz Nauki Polskiej (FNP) za wsparcie finansowe w ramach projektu START 2021 oraz Narodowej Agencji Wymiany Akademickiej (NAWA), za umożliwienie odbycia stażu zagranicznego w ramach Programu im. Iwanowskiej.

Praca była wspierana przez Narodowe Centrum Nauki (NCN) w ramach grantu PRELUDIUM-18, nr: UMO-2019/35/N/ST4/02481.

Spis treści

Streszczenie	9
Abstract	12
Wykaz publikacji wchodzących w skład pracy doktorskiej	15
Część teoretyczna	17
1. Spektroskopia w podczerwieni	17
1.1. Oscylacje molekuł	17
1.2. Analiza ilościowa	21
1.3. Transmisyjna spektroskopia FT-IR	22
1.4. Aspekty uzyskiwania obrazu	23
1.5. Spolaryzowana spektroskopia w podczerwieni	25
1.6. Inne metody spektroskopii IR dalekiego pola – O-PTIR i QCL	26
2. Metody wstępnego przetwarzania danych	29
2.1. Metody odsumiania	30
2.2. Metody korekty linii bazowej	34
3. Orientacja molekuł	39
3.1. Orientacja na płaszczyźnie – 2D	40
3.2. Orientacja molekuł – 3D	40
4. Klasyfikacja – Las Losowy	43
Przedruki publikacji	45
Wpływ odsumiania na dane spektroskopowe i skuteczność klasyfikacji	45
Klasyfikacja histopatologiczna tkanek trzustki	83
Wyznaczanie orientacji makromolekuł w oparciu o spolaryzowaną spektroskopię IR	97
Korekta rozpraszania typu Mie dla próbek o cylindrycznym kształcie	115
Referencje	131

Streszczenie

Celem niniejszej rozprawy doktorskiej było opracowanie i optymalizacja metodyki dla obrazowania spektroskopowego w podczerwieni (IR), pod kątem osiągnięcia maksymalnej jakości danych oraz tworzenia modeli klasyfikacji histopatologicznej raka trzustki. Ponadto, podmiotem prowadzonych prac było rozwinięcie technik wyznaczania orientacji makromolekuł, co związane jest z hipotezą mówiącą o dostarczaniu przez nie dodatkowych informacji o chorobie oraz usprawnianiu modeli klasyfikacyjnych. Oba zagadnienia uwzględniają wykorzystanie spektroskopii w podczerwieni, zapewniającej ogrom informacji na temat składu biochemicznego badanej próbki. Co kluczowe, realizacja opisanych powyżej aspektów jest wsparta przez dogłębnie zoptymalizowane procedury wstępnego przetwarzania danych, ze skupieniem na metodach odszumiania oraz korekty linii bazowej.

W związku z nakreślonymi powyżej celami, w pracy pt. *„Obrazowanie w podczerwieni w klasyfikacji histopatologicznej i orientacji makromolekuł, wsparte zoptymalizowanymi metodami wstępnego przetwarzania danych”* można wyodrębnić cztery obszary tematyczne złożone ze zbiorów publikacji. Pierwszy związany jest z szeroko pojętymi optymalizacjami technik odszumiania widm oraz obrazów spektroskopowych, pod kątem podniesienia jakości sygnału oraz poprawy czułości klasyfikacji histopatologicznej. Kolejnym obszarem jest sama klasyfikacja histopatologiczna oraz jej skuteczność w diagnostyce nowotworów trzustki w oparciu o obrazowanie IR. Trzeci obszar badań to wyznaczanie orientacji makromolekuł, w płaszczyźnie (2D) oraz przestrzeni (3D), wykorzystujące spolaryzowaną spektroskopię w podczerwieni, a docelowym zastosowaniem obu metod jest określanie orientacji włókien kolagenowych. Finalnym tematem jest implementacja metody korekty linii bazowej typu EMSC (ang. *Extended Multiplicative Signal Correction*) skierowanej do próbek o cylindrycznym kształcie (np. włókien), wywołujących silne rozpraszanie typu Mie.

W części teoretycznej pracy pokryte zostały podstawy fizyczne zagadnień związanych z badaniami opisanymi w publikacjach wchodzących w skład rozprawy. Rozdział rozpoczęto od opisu spektroskopii IR, skupiając się na podstawach spektroskopii oscylacyjnej oraz kwestiach związanych z analizą ilościową. Scharakteryzowany został układ do transmisyjnej Spektroskopii Fourierowskiej w podczerwieni (FT-IR) z uwzględnieniem układu optycznego i aspektów uzyskiwania obrazu. Przybliżono też dwie dodatkowe metody spektroskopii IR, które były wykorzystane w badaniach – mikroskopie QCL (ang. *Quantum Cascade Laser*) oraz O-PTIR (ang. *Optical-PhotoThermal Infrared*). W dalszej części uwaga skupiona została na konkretnych metodach wstępnego przetwarzania danych. W grupie technik odszumiania spektralnego i przestrzennego znalazły się proste filtry (średniej, medianowy, itp.), częstotliwościowy filtr dolnoprzepustowy, filtr Savitzky'ego-Golay'a, Falki, Analiza Głównych Składowych (PCA) oraz metoda Najmniejszej Frakcji Szumu (MNF). Następnie poruszono aspekty korekty linii bazowej, powstającej w wyniku efektów rozpraszania dominujących w transmisyjnym trybie pomiarowym. Główna uwaga została skupiona na algorytmach typu EMSC, wykorzystujących zarówno pełną teorię rozpraszania typu Mie, jak i uproszczenie van de Hulsta. Szczegółowo opisany został algorytm EMSC służący do korekty rozpraszania przez próbki o cylindrycznym kształcie. Trzeci podrozdział wstępu teoretycznego

opisuje wykorzystanie spolaryzowanej spektroskopii IR do wyznaczania orientacji dipolowych momentów przejścia. Opisano metodę 2D, dającą informację ograniczoną do kierunku rzutu momentu dipolowego na płaszczyznę, oraz metodę 3D, jednocześnie wykorzystującą dwa momenty dipolowe i zapewniającą informację przestrzenną o ich orientacji. Poprzez analizę kierunków oscylacji atomów, orientacja momentów dipolowych jest przenoszona na orientację łańcuchów w makromolekułach. Rozdział sfinalizowano opisem metody klasyfikacji Las Losowy (ang. Random Forest), wykorzystanej w badaniach nad modelem histopatologicznym nowotworów trzustki.

Kluczowym dla niniejszej rozprawy fragmentem jest rozdział przedruków. Grupuje on publikacje, będące podstawą pracy, w cztery wspominane wyżej wątki tematyczne. Każdy z tematów posiada krótki opis kontekstu oraz motywacji do przeprowadzonych badań, jak również kluczowe rezultaty i płynące z nich konsekwencje. Pierwszy i największy zbiór publikacji dotyczył wpływu odszumiania na dane spektroskopowe oraz na skuteczność klasyfikacji histopatologicznej. Dogłębnie zoptymalizowano szereg metod spektralnego i przestrzennego odszumiania w szerokim zakresie ich parametrów. W przypadku zastosowania skierowanego na poprawę jakości sygnału zastosowano symulowane zestawy danych o różnym poziomie szumu oraz rozdzielczości przestrzennej. Umożliwiło to oszacowanie skuteczności metod, zarówno pod kątem poprawy stosunku sygnału do szumu (SNR), jak i zminimalizowania zaburzeń sygnału wprowadzanych przez odszumianie. Uzyskane wyniki wskazały na ogromną skuteczność metod MNF i PCA, uwidaczniając możliwość redukcji czasu pomiarowego do minimum (np. do 4 skanów). Ponadto, skrypty do generacji danych symulowanych, wraz z samymi danymi, zostały udostępnione na zasadach otwartego dostępu. Odszumianie zostało również zoptymalizowane pod kątem poprawienia skuteczności klasyfikacji z wykorzystaniem Lasów Losowych. Pokazano, że po zastosowaniu metody PCA lub MNF do danych FT-IR, czułość klasyfikacji na poziomie piksela wzrasta o nawet 43-44%. W przypadku klasyfikacji opartej o dane pochodzące z mikroskopii QCL, zastosowanie MNF wskazało możliwość redukcji informacji spektralnej do zaledwie 20 zmiennych (pasm). Przytoczone optymalizacje odszumiania miały kluczowy wpływ na opisaną dalej klasyfikację.

W ramach tematyki związanej z klasyfikacją histopatologiczną trzustki, celem badań było stworzenie modeli umożliwiających detekcję nowotworów w oparciu o spektroskopię FT-IR oraz metody uczenia maszynowego (Lasu Losowego). W tym celu zebrano dane dla ponad 600 biopsji trzustki, pochodzących od około 250 pacjentów, wykorzystując zarówno standardowy (15x) jak i wysokorozdzielczy obiektyw (36x), zapewniający bardziej szczegółową informację przestrzenną. Finalnie stworzone zostały modele wyróżniające po 6 klas, co umożliwia detekcję nowotworu lub zapalenia ze skutecznością na poziomie 95%. Dodatkowo, zaprezentowano możliwość detekcji marginesu chirurgicznego uwidaczniającego wyraźną granicę pomiędzy zdrową, a nowotworową tkanką.

Teorie na temat wpływu tkanki włóknistej, będącej mikrootoczeniem nowotworów, na przebieg choroby oraz skuteczności leczenia były motywacją do rozwoju metod wyznaczania orientacji makromolekuł, w oparciu o spolaryzowaną spektroskopię FT-IR. Przedstawiono rezultaty zastosowania metody 2D do tkanek bogatych we włókna kolagenowe, a otrzymane wyniki wyraźnie wskazywały kierunki orientacji molekuł. Co więcej, funkcja Hermansa wykasponowała wysoko- i nisko ustrukturyzowane obszary, odpowiednio dla tkanki włóknistej

oraz komórek nabłonkowych. Było to pierwsze zastosowanie tej metody do tak skomplikowanego układu jak tkanka. W dalszym kroku zaprezentowano pierwszą implementację do danych eksperymentalnych, teoretycznej metody wyznaczania orientacji 3D. Wykorzystano w tym celu dobrze opisany w literaturze układ polimeru polikaprolaktonu w krystalicznej formie sferulitu. Metodę zaaplikowano również do danych O-PTIR, umożliwiających pomiary z submikronową rozdzielczością. Wyznaczona orientacja makromolekuł była zgodna z przewidywaniami teoretycznymi, tworząc współśrodkowe okręgi wokół centrum nukleacji, co przybliżyło potencjał tej metody w nietomograficznym i niedestruktywnym wyznaczaniu orientacji 3D.

Tkanka włóknista analizowana w ramach badań nad orientacją ujawniła wysokie tendencje do rozpraszania światła, w znacznym stopniu zaburzające analizy. Dlatego czwartą i ostatnią poruszoną w rozprawie tematyką jest korekta rozpraszania. Przedstawiono implementację algorytmu EMSC służącego do korekty rozpraszania typu Mie przez próbki o cylindrycznym kształcie. Algorytm rozszerzono również o przypadek liniowej polaryzacji światła, aby jak najlepiej dopasować go do zastosowań związanych z wyznaczaniem orientacji. W celu przyspieszenia obliczeń wykorzystano procesory graficzne, a stworzona w oprogramowaniu MATLAB biblioteka została udostępniona na zasadach otwartego dostępu. Włókno polimeru o silnych właściwościach rozpraszających zostało wykorzystane do walidacji nowego algorytmu, który okazał się skuteczniejszy od metody referencyjnej.

Abstract

This doctoral dissertation was aimed at the development and optimization of a methodology for infrared (IR) spectroscopic imaging, with the goal of achieving maximum data quality and creating models for histopathological classification of pancreatic cancer. Moreover, the research was focused on the development of techniques for the macromolecular orientation determination, which is related to the hypothesis stating, that they can provide additional information about the disease and improve classification models. Both of these issues are based on the use of infrared spectroscopy, which provides a wealth of information about the biochemical composition of the analyzed sample. What is crucial, the implementation of the aforementioned aspects is supported by thoroughly optimized data preprocessing procedures, with a focus on denoising and baseline correction methods.

In relation to the objectives outlined above, the thesis entitled “Infrared Imaging in histopathological classification and macromolecular orientation determination, supported by optimized data preprocessing methods” can be divided into four thematic areas, composed of publication sets. The first area is related to the optimization of spectral and spatial denoising techniques, to improve the quality of the signal and the effectiveness of histopathological classification. The following area focuses on histopathological classification and its effectiveness in diagnosing pancreatic tumors using IR imaging. The third research area is the determination of the in-plane (2D) and in-space (3D) macromolecular orientation using polarized infrared spectroscopy, with the focus pointed at the orientation of collagen fibers. The final topic is the implementation of the EMSC (Extended Multiplicative Signal Correction) baseline correction method for samples with cylindrical domains (e.g. fibers), which exhibit strong Mie-type scattering.

The theoretical part of the thesis covers the physical basis of the aspects related to the research, described in the publications included in the work. The chapter begins with a description of IR spectroscopy, focusing on the fundamentals of vibrational spectroscopy and quantitative analysis. The transmission Fourier Transform Infrared Spectroscopy (FT-IR) setup was characterized, including the optical system and image formation aspects. Two additional far-field IR spectroscopic methods are also introduced, namely Quantum Cascade Laser (QCL) microscopy and Optical-PhotoThermal Infrared (O-PTIR). The following part focuses on specific data preprocessing methods, including spectral and spatial denoising techniques. The group of denoising techniques includes simple filters (mean, median, etc.), low-pass frequency filter, Savitzky-Golay filter, Wavelets, Principal Component Analysis (PCA), and the Minimum Noise Fraction (MNF) method. The aspects of baseline correction, resulting from scattering effects dominant in the transmission measurement mode, are then discussed. The main attention is focused on EMSC-type algorithms, utilizing the full Mie-type scattering theory and the van de Hulst approximation. The EMSC algorithm for scattering correction in samples with cylindrical domains is described in detail. The third subchapter of the theoretical introduction describes the use of polarized IR spectroscopy for the determination of the transition dipole moments orientation. The 2D method provides limited information about the projection of the dipole moment onto the plane, whereas the 3D method utilizes two dipole moments simultaneously and provides spatial information about their orientation. By analyzing

the directions of vibrations, the orientation of transition dipole moments is used to determine the orientation of macromolecular chains. The chapter concludes with a description of the Random Forest classification method, which was used in pancreatic cancer histopathology research.

The key part of this dissertation is the chapter featuring publication reprints. It groups the papers, that form the fundamentals of this thesis, into the four thematic threads mentioned above. Each of the topics has a brief description of the context and motivation for the conducted research, as well as the crucial results and conclusions. The first and largest set of publications is focused on the influence of denoising on spectroscopic data and the histopathological classification results. A range of spectral and spatial denoising methods were thoroughly optimized across a wide range of parameters. In the case of signal quality improvement, simulated data sets with varying levels of noise and spatial resolution were used. This allowed for estimation the effectiveness of the methods, in terms of improving the signal-to-noise ratio (SNR) and minimizing the signal distortion introduced by denoising. The obtained results revealed the power of MNF and PCA methods, highlighting the potential for reduction of measurement time. Furthermore, scripts for generating simulated data, along with the data itself, were published as open access. In addition, denoising was optimized to improve the Random Forests classification accuracy. It was shown that applying PCA or MNF to FT-IR data increased pixel-level classification sensitivity by up to 43-44%. For classification based on data from QCL microscopy, the application of MNF demonstrated the potential for the reduction of spectral information to as few as 20 variables (bands). Such denoising optimizations had a crucial impact on the classification described further.

Pancreatic histopathology was the next research area, with the aim to create models allowing cancer detection based on FT-IR spectroscopy and machine learning methods (Random Forest). To achieve this, data were collected for over 600 pancreatic biopsies, from approximately 250 patients, using both a standard (15x) and a high-resolution objective (36x), with the latter providing more detailed spatial information. Models were trained to distinguish between 6 classes, enabling tumor and inflammation detection with an accuracy of 95%. Additionally, the ability to detect the surgical margin, highlighting a clear boundary between healthy and tumorous tissue, was presented.

The theories regarding the influence of fibrous tissue (forming tumor microenvironment) on disease progression and the effectiveness of treatment, were the motivation for the development of methods for molecular orientation determination, based on polarized FT-IR spectroscopy. Results of 2D method application to collagen-rich tissues were presented, clearly indicating the directions of molecular orientation. Additionally, the Hermans function exposed areas of high and low levels of orientation, respectively for fibrous tissue and epithelial cells. This was the first application of this method to such a complex system as a tissue sample. In the following step, the first implementation of the theoretical 3D orientation determination method to experimental data was presented. For this purpose, a well-described in literature system of polycaprolactone polymer in the form of a spherulite was used as a model sample. Furthermore, the method was applied to O-PTIR, data allowing for sub-micron resolution measurements. The determined orientation of macromolecules was consistent with theoretical predictions, forming concentric circles around the nucleation center, and

highlighting the potential of this method in non-tomographic and non-destructive 3D orientation determination.

Fibrous tissue analyzed in orientation studies revealed a high tendency to scatter light, significantly disrupting the analysis. Therefore, the fourth and final topic addressed in the dissertation is the scattering correction. The implementation of the EMSC algorithm for Mie-type scattering correction in samples with cylindrical domains was presented. Additionally, the algorithm was extended to include linearly polarized light properties, optimizing it for orientation determination applications. To speed up the calculations, graphical processing units (GPUs) were utilized, with prepared MATLAB library later published as an open access. A polymer fiber with strong scattering properties was used for algorithm's validation, and their effectiveness exceeded other methods.

Wykaz publikacji wchodzących w skład pracy doktorskiej

- A) P. Koziol, M.K. Raczowska, J. Skibinska, S. Urbaniak-Wasik, C. Paluszkievicz, W. Kwiatek, T.P. Wrobel, “*Comparison of spectral and spatial denoising techniques in the context of High Definition FT-IR imaging hyperspectral data*”, Scientific Reports, 8 (2018) 1–11
<https://doi.org/10.1038/s41598-018-32713-7>
MNiSW: 140 pkt
IF: 4.011
- B) M.K. Raczowska, P. Koziol, S. Urbaniak-Wasik, C. Paluszkievicz, W. Kwiatek, T.P. Wrobel, “*Influence of denoising on classification results in the context of hyperspectral data: High Definition FT-IR imaging*”, Analytica Chimica Acta, 1085 (2019), 39–47
<https://doi.org/10.1016/j.aca.2019.07.045>
MNiSW: 100 pkt
IF: 5.256
- C) P. Koziol, M.K. Raczowska, J. Skibinska, N.J. McCollum, S. Urbaniak-Wasik, C. Paluszkievicz, W.M. Kwiatek, T.P. Wrobel, “*Denoising influence on discrete frequency classification results for quantum cascade laser based infrared microscopy*”, Analytica Chimica Acta, 1051 (2019) 24–31
<https://doi.org/10.1016/j.aca.2018.11.032>
MNiSW: 100 pkt
IF: 5.256
- D) T.P. Wrobel, P. Koziol, M.K. Raczowska, D. Liberda, C. Paluszkievicz, W.M. Kwiatek, “*Noise-free simulation of an FT-IR imaging hyperspectral dataset of pancreatic biopsy core bound by experiment*”, Scientific Data, 6 (2019) 239
<https://doi.org/10.1038/s41597-019-0260-x>
MNiSW: 140 pkt
IF: 6.785
- E) P. Koziol, D. Liberda, W.M. Kwiatek, T.P. Wrobel, “*Macromolecular Orientation in Biological Tissues Using a Four-Polarization Method in FT-IR Imaging*”, Analytical Chemistry, 92 (2020) 13313–13318
<https://pubs.acs.org/doi/10.1021/acs.analchem.0c02591>
MniSW: 140 pkt
IF: 6.785
- F) P. Koziol, K. Kosowska, D. Liberda, F. Borondics, T.P. Wrobel, “*Super-Resolved 3D Mapping of Molecular Orientation Using Vibrational Techniques*”, Journal of the American Chemical Society, 144 (2022) 14278–14287
<https://doi.org/10.1021/jacs.2c05306>
MNiSW: 200 pkt
IF: 16.383

- G) D. Liberda[#], P. Koziol[#], T. P. Wrobel, “*Comprehensive Histopathology Imaging in Pancreatic Biopsies: High Definition Infrared Imaging with Machine Learning Approach*”, International Journal of Biological Sciences, 2023 (zaakceptowany manuskrypt)
równy wkład
DOI: 10.7150/ijbs.83068
Preprint: <https://doi.org/10.1101/2022.07.01.22277130>
MNiSW: 100 pkt
IF: 10,75
- H) P. Koziol, K. Kosowska, P. Korecki, T. P. Wrobel, “*Scattering correction for samples with cylindrical domains measured with polarized infrared spectroscopy*”,
Preprint: <https://doi.org/10.26434/chemrxiv-2023-cg5dg>
w recenzji w czasopiśmie Analytica Chimica Acta, 2023
MNiSW: 100 pkt
IF: 6,911

Część teoretyczna

1. Spektroskopia w podczerwieni

Podstawą wszystkich rodzajów badań spektroskopowych jest oddziaływanie promieniowania elektromagnetycznego z materią, bazujące na takich efektach jak absorpcja, emisja, czy rozpraszanie fotonów. Jednym z przykładów jest spektroskopia w podczerwieni (IR), której najpowszechniej stosowanym wariantem jest Spektroskopia Fourierowska w Podczerwieni – FT-IR (ang. *Fourier-Transform Infrared Spectroscopy*). W tej metodzie, widmo o charakterystycznym kształcie, zawierające informacje o składzie biochemicznym próbki, powstaje poprzez absorpcje kwantów światła podczerwonego oddziaływujących z próbką. Jakościowe i ilościowe informacje o badanym materiale są na widmach widoczne w postaci pasm o specyficznej intensywności i kształcie, przypadających dla energii dopasowanych do danych poziomów oscylacyjnych. Zwyczajowo stosowaną skalą jest tutaj liczba falowa $\tilde{\nu}$ [cm^{-1}], a najczęściej stosowanym zakresem spektralnym jest tzw. średnia podczerwień MIR (ang. *Mid InfraRed*), od 400 do 4000 cm^{-1} , w skład której wchodzi obszar zwany *fingerprint*, od 400 do 1800 cm^{-1} , umożliwiający identyfikację związków chemicznych poprzez obecność charakterystycznych dla nich pasm [1]. Zakres pomiędzy 50 a 600 cm^{-1} nazywany jest daleką podczerwienią FIR (ang. *Far InfraRed*), natomiast obszar powyżej 4000 cm^{-1} to tzw. bliska podczerwień NIR (ang. *Near InfraRed*) [1].

Intensywny rozwój dziedziny spektroskopii IR na przestrzeni minionych dekad zaowocował powstaniem wielu nowych technik badawczych, między innymi Mikroskopii QCL (ang. *Quantum Cascade Laser*) [2], [3], O-PTIR (ang. *Optical-PhotoThermal Infrared*) [4] oraz szerokiej gamy technik z obszaru nano-skali, np. s-SNOM (ang. *Scattering-type Scanning Near-field Optical Microscopy*) [5]–[7], czy AFM-IR (ang. *Atomic Force Microscope InfraRed*) [8], [9]. Uwaga w tej pracy poświęcona jednak będzie głównie spektroskopii FT-IR.

1.1. Oscylacje molekuł

Informacja biochemiczna zawarta w widmach IR jest zdeterminowana przejściami pomiędzy skwantowanymi energetycznymi stanami oscylacyjnymi molekuł. W przypadku dwuatomowej molekuły, takie oscylacje mogą mieć formę prostych, sprzężonych ruchów każdego z atomów. Sytuacja staje się jednak bardziej skomplikowana dla molekuł wieloatomowych. N – atomowa molekuła posiada $3N$ stopni swobody, z których trzy opisują ruch translacyjny (w osiach X , Y i Z), a trzy kolejne ruchy rotacyjne (wokół osi X , Y i Z). Pozostałe stopnie swobody ($3N-6$) opisują liczbę módów oscylacyjnych, czyli sposobów, w jaki nieliniowa molekuła może wykonywać oscylacje [1]. Wyjątek stanowią molekuły posiadające symetrię liniową, gdzie liczba módów oscylacyjnych sumuje się do $3N-5$ [1]. Podstawą módów oscylacyjnych jest w przybliżeniu harmoniczne wychylenie atomów z ich położenia równowagi, przy czym każdy i -ty mód posiada charakterystyczną częstotliwość drgań ν_i [1]. Finalna wartość energii potencjalnej jest skwantowana i zależy od poziomu energetycznego opisanego oscylacyjną liczbą kwantową ($v = 0, 1, 2, 3 \dots$).

Dwuatomowa molekula posiada pojedynczy mod oscylacyjny. Zakładając, że drga ona ruchem prostym harmonicznym, tzn. spełnia prawo Hooke'a, mówiące, że siła F działająca na ciało jest proporcjonalna do wychylenia x [10]

$$F = -kx, \quad (1)$$

gdzie k to stała siłowa, można wykorzystać to prawo do wyznaczenia operatora energii potencjalnej molekuly [10]

$$V(x) = \frac{1}{2}kx^2. \quad (2)$$

Operator energii potencjalnej jest obecny w niezależnym od czasu równaniu Schrödingera [10]

$$-\frac{\hbar^2}{2\mu} \frac{\partial^2}{\partial x^2} \psi(x) + V(x)\psi(x) = E\psi(x), \quad (3)$$

gdzie $\psi(x)$ to funkcja falowa, E to energia całkowita, μ to masa zredukowana, $\hbar$ to zredukowana stała Planck'a ($\hbar = h/2\pi$) oraz h to stała Planck'a. Podstawienie $V(x)$ i przekształcenie powyższego równania daje równanie różniczkowe drugiego rzędu [10]

$$\frac{\partial^2}{\partial x^2} \psi(x) + \frac{2\mu}{\hbar^2} \left(E - \frac{1}{2}kx^2 \right) \psi(x) = 0, \quad (4)$$

którego rozwiązanie pozwala wyznaczyć wartości energii E .

Skwantowane poziomy energetyczne będące rozwiązaniem równania 4, opisują energię potencjalną modu oscylacyjnego dwuatomowej molekuly [10]

$$E_{iv} = \hbar v_i \left(v + \frac{1}{2} \right), \quad v = 0, 1, 2, 3, \dots \quad (5)$$

gdzie

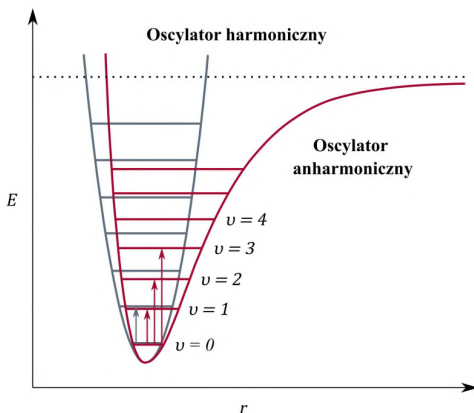
$$v_i = \frac{1}{2\pi} \left(\frac{k}{\mu} \right)^{\frac{1}{2}}, \quad (6)$$

to częstotliwość drgań podstawowych.

Powyższe równania 5-6 są jednak tylko przybliżeniem, zakładającym niewielkie wychylenie atomów z ich położenia równowagi. Energia potencjalna oscylatora może być rozwinięta w szereg Taylora, a założenie niewielkich wychyleń atomów umożliwia ograniczenie formuły na energię jedynie do wyrazów kwadratowych (równanie 2) [10]. W rzeczywistości do opisu energii potencjalnej wykorzystywany jest model oscylatora anharmonicznego, uwzględniający dalsze składowe szeregu (anharmoniczne). Włączenie składowych anharmonicznych do operatora energii potencjalnej w równaniu Schrödingera wymaga podejścia perturbacyjnego [10], w wyniku którego otrzymuje się następujące energie modów:

$$E_{iv}^{anh} = \hbar v_i \left(v + \frac{1}{2} \right) - \hbar v_i x_i \left(v + \frac{1}{2} \right)^2, \quad (7)$$

gdzie x_i jest stałą anharmoniczną zależną od modu oscylacyjnego. Stała ta jest bezwymiarowa i zazwyczaj osiąga wartości z zakresu pomiędzy -0.001 a -0.02 [11]. Wykresy krzywych energetycznych oscylacji dwuatomowej molekuly dla obu typów oscylatorów przedstawione są na rysunku 1. Na pierwszy rzut oka można zaobserwować, że odległości pomiędzy sąsiadującymi stanami energetycznymi nie są stałe, a według modelu anharmonicznego ulegają zmniejszeniu dla wzrastających liczb kwantowych.



Rysunek 1. Przykład skwantowanych poziomów energetycznych dla modu oscylacyjnego dwuatomowej molekuly. Szarą linią wykreślono przybliżenie tej zależności przez model oscylatora harmonicznego, natomiast czerwona linia odpowiada dokładniejszemu modelowi oscylatora anharmonicznego, zakładającego brak spełnienia prawa Hooke’a.

Warunkiem przejścia energii potencjalnej z niższego na wyższy poziom, jest absorpcja kwantu promieniowania o energii dokładnie dopasowanej do różnicy poziomów energetycznych. Dla przykładu, różnica pomiędzy podstawowym ($v = 0$), a pierwszym wzbudzonym ($v = 1$) poziomem energetycznym, dla większości modów oscylacyjnych, jest w zakresie środkowej podczerwieni (od 400 do 4000 cm^{-1}). Warto wspomnieć, że model anharmoniczny (poprzez uwzględnienie odstępów od prawa Hooke’a) rozluźnia reguły wyboru, tj. dopuszcza przejścia dla których liczba kwantowa zmienia się o wartość większą niż 1 ($|\Delta v| > 1$). Dzięki temu, możliwa jest obserwacja nadtonów i pasm kombinacyjnych, zabronionych przez model oscylatora harmonicznego ($|\Delta v| = 1$) [10], [11].

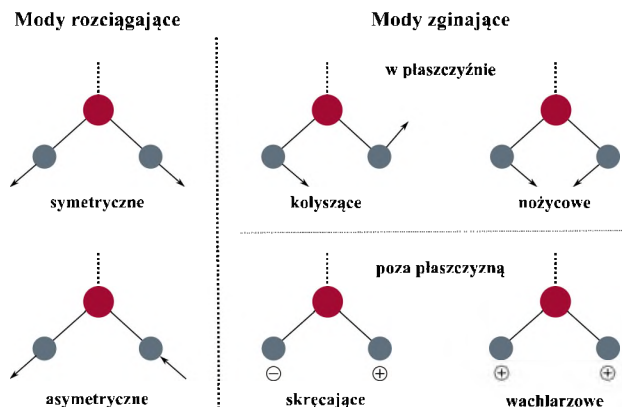
Opisane powyżej reguły, związane ze zmianą liczby kwantowej, nie są jedynymi warunkującymi przejście do stanu wzbudzonego. Molekuła może w wyniku absorpcji przejść na wyższy poziom energetyczny pod warunkiem, że posiada dipol, który w wyniku oddziaływania z polem elektromagnetycznym oscyluje (przynajmniej chwilowo) z tą samą częstotliwością co pole [12]. Oznacza to, że dipolowy moment przejścia μ_{nm} jest niezerowy [1]

$$\mu_{nm} = \langle \psi_n | \hat{\mu} | \psi_m \rangle \neq 0, \quad (8)$$

gdzie ψ_n oraz ψ_m to funkcje falowe stanu początkowego i końcowego, a $\hat{\mu}$ to operator elektrycznego momentu dipolowego. Wartość osiąganą przez dany dipolowy moment przejścia μ_{nm} można utożsamiać z miarą przemieszczania się ładunku, które towarzyszy przejściu [12]. W efekcie, dla molekuł posiadających pewne rodzaje symetrii, niektóre z modów oscylacyjnych mogą być zabronione, a inne mogą być zdegenerowane (więcej niż jeden mod ma tę samą częstotliwość drgań). Przykładem jest liniowa cząsteczka CO_2 , gdzie symetryczne oscylacje atomów tlenu nie niosą za sobą zmiany momentu dipolowego, tym samym pasmo jest nieaktywne w podczerwieni [12]. W związku z tym, obserwowana w IR liczba modów oscylacyjnych jest często inna niż $3N-6$, czego przykładem jest obserwowanie $3N-5$ modów oscylacyjnych dla molekuł z liniową symetrią. Dla takich molekuł, rotacja wokół osi wiązania nie niesie za sobą przesunięcia żadnego z atomów, dlatego jeden z rotacyjnych stopni swobody jest tracony, a liniowa molekula zyskuje dodatkowy mod oscylacyjny.

Podsumowując przedstawione powyżej informacje, zaobserwowanie pasma związanego z danym modem oscylacyjnym na widmie IR jest zdeterminowane trzema regułami wyboru. Po pierwsze, energia absorbowanego promieniowania musi odpowiadać różnicy energii dla poziomów energetycznych molekuly. Ponadto, przejście następuje tylko jeśli liczba kwantowa zmienia się o $\Delta v = +1, +2, +3, \dots$. Finalnie, aktywne w podczerwieni są tylko przejścia przy których zmiana ulega moment dipolowy.

Dla przykładu, na rysunku 2 przedstawiono mody oscylacyjne dla grupy funkcyjnej typu XY_2 , której najczęstszym przykładem jest grupa $-\text{CH}_2$ [11]. Najogólniej, mody oscylacyjne można podzielić na rozciągające, gdzie zmiana ulega długość wiązania chemicznego, oraz zginające, gdzie deformacji ulegają kąty. Dalsze podziały związane są z symetrią (symetryczne i asymetryczne) oraz z płaszczyzną oscylacji (drgania w płaszczyźnie lub poza płaszczyzną grupy funkcyjnej).



Rysunek 2. Mody oscylacyjne grupy funkcyjnej typu XY_2 .

1.2. Analiza ilościowa

Prawo Lamberta-Beera opisuje pochłanianie promieniowania elektromagnetycznego podczas przechodzenia przez dany ośrodek i jest tym samym podstawą wszelkich ilościowych analiz spektroskopowych. Mówi ono, że transmitancja próbki dla danej liczby falowej $\tilde{\nu}$ jest równa stosunkowi mocy promieniowania wychodzącego z próbki $I(\tilde{\nu})$ do mocy promieniowania padającego na próbkę $I_0(\tilde{\nu})$. Inne sformułowanie podaje natomiast, że dla próbki o grubości b , transmitancja dla liczby falowej $\tilde{\nu}$ wynosi [11]

$$T(\tilde{\nu}) = \frac{I(\tilde{\nu})}{I_0(\tilde{\nu})} = \exp[-\alpha(\tilde{\nu}) \cdot b], \quad (9)$$

gdzie $\alpha(\tilde{\nu})$ jest liniowym współczynnikiem absorpcji (cm^{-1}) dla danego $\tilde{\nu}$. Z kolei definicja absorbancji jest następująca [11]

$$A(\tilde{\nu}) = \log_{10} \frac{1}{T(\tilde{\nu})} = \frac{1}{\ln 10} \alpha(\tilde{\nu}) \cdot b = a(\tilde{\nu}) \cdot b. \quad (10)$$

Jeżeli badana próbka jest mieszką i substancji, każda o stężeniu c_i , to absorbancję związaną z każdą z tych substancji indywidualnie można wyrazić jak poniżej [11]

$$A_i(\tilde{\nu}) = a_i(\tilde{\nu}) \cdot b \cdot c_i, \quad (11)$$

co daje następującą absorbancję mieszkanki złożonej z N substancji [11]

$$A(\tilde{\nu}) = \sum_{i=1}^N a_i(\tilde{\nu}) \cdot b \cdot c_i. \quad (12)$$

Analizując powyższe równania można zauważyć, że to absorbancja, a nie transmitancja jest proporcjonalna do stężenia danej substancji w próbce. Jednakże, ani absorbancja ani transmitancja nie są mierzone bezpośrednio. Mierzona jest natomiast intensywność promieniowania docierającego do detektora po przejściu przez próbkę oraz bez jej obecności. Intensywność jest wielkością proporcjonalną do mocy, tak że po wyznaczeniu stosunku tych wielkości (według równania 9) otrzymywana jest transmitancja, a co za tym idzie, wyznaczana jest absorbancja. Warto nadmienić, że w praktyce wykonywane są pomiary ekstynkcji promieniowania, a dopiero po zastosowaniu odpowiednich metod wstępnego przetwarzania danych (opisanych w dalszych rozdziałach) otrzymywana jest absorbancja.

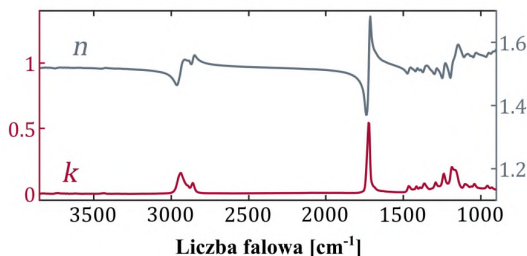
W kontekście analizy ilościowej, poza prawem Lamberta-Beera, fundamentalną rolę w opisie absorpcji promieniowania odgrywają rzeczywisty $n(\tilde{\nu})$ i urojony $k(\tilde{\nu})$ współczynnik załamania światła, będące częściami zespolonego współczynnika załamania $\tilde{n}(\tilde{\nu})$ wyrażonego następująco [11]

$$\tilde{n}(\tilde{\nu}) = n(\tilde{\nu}) + ik(\tilde{\nu}). \quad (13)$$

$n(\tilde{\nu})$ jest często nazywane po prostu współczynnikiem załamania opisywanym przez prawo Snella, z kolei $k(\tilde{\nu})$ opisuje właściwości absorpcyjne próbki, poprzez swój związek z liniowym współczynnikiem absorpcji $\alpha(\tilde{\nu})$ [11]

$$k(\tilde{\nu}) = \frac{\alpha(\tilde{\nu})}{4\pi\tilde{\nu}}. \quad (14)$$

Dla organicznych i nieorganicznych molekuł, posiadających typowe pasma absorpcyjne, współczynnik załamania światła zmienia się w zakresie spektralnych tych pasm (stąd obecność zależności od liczby falowej $\tilde{\nu}$ w równaniach 13 i 14). Stosowny przykład dla polimeru polikaprolaktonu (PCL) jest przedstawiony na rysunku 3, gdzie dla zakresów występowania pasm absorpcyjnych obserwowana jest zdecydowana zmiana współczynnika załamania $n(\tilde{\nu})$. Zjawisko to nazywane jest dyspersją.



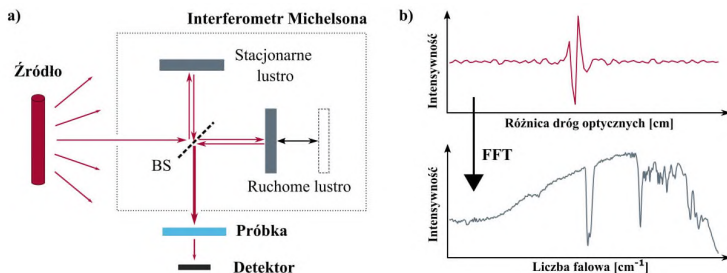
Rysunek 3. Rzeczywisty i urojony współczynnik załamania światła w zakresie średniej podczerwieni dla polimeru PCL.

1.3. Transmisyjna spektroskopia FT-IR

Pośród wielu dostępnych wariantów spektroskopii IR, w tej pracy najwięcej uwagi poświęcone będzie Spektroskopii Fourierowskiej w Podczerwieni – FT-IR. Podstawą tej metody jest operacja matematyczna zwana transformatą Fouriera, umożliwiająca przejście z widma intensywności w dziedzinie czasu (lub położenia) na dziedzinę częstotliwości (liczb falowych). Klasyczny spektrometr składa się ze źródła, interferometru Michelsona, komory próbki i detektora.

Typowym rodzajem źródeł wykorzystywanych w spektrometrach FT-IR są źródła termiczne. Najczęściej stosowane są *globary* (SiC) oraz pręty Nernsta ($ZrO_2+Y_2O_3$), które po podgrzaniu do wysokich temperatur emitują ciągłe widmo, podobne do promieniowania ciała doskonale czarnego [1]. Dla pręta Nernsta jest to zakres od UV do NIR, natomiast emisja *globara* pokrywa obszar od bliskiej, przez średnią, po daleką podczerwień (NIR, MIR oraz FIR). W dalszym etapie światło kierowane jest do interferometru.

W spektroskopii FT-IR intensywność pasm absorpcyjnych nie jest mierzona bezpośrednio w funkcji częstotliwości. Wynika to z faktu, że pierwotnie stosowane siatki dyfrakcyjne, izolujące pojedyncze długości fali, zostały zastąpione przez interferometrię, co znacznie skróciło skalę czasową pomiarów widmowych, zwiększyło czułość badań i poprawiło rozdzielczość spektralną. Budowa większości spektrometrów oparta jest o projekt dwuwijkowego interferometru Michelsona, przedstawionego na rysunku 4a.



Rysunek 4. a) Schemat spektrometru FT-IT, składającego się ze źródła, interferometru Michelsona, komory próbki i detektora. b) Przykładowy interferogram (góra) oraz odpowiadające mu widmo (dół), będące wynikiem szybkiej transformaty Fouriera (FFT).

Szerokopasmowe i w przybliżeniu skolimowane promieniowanie ze źródła termicznego pada na kluczowy element interferometru jakim jest rozdzielacz wiązki (BS, ang. *beamsplitter*). Rozdzielacz jest ustawiony pod kątem 45° do obu luster, dzięki temu, poprzez odbicie światła kieruje połowę wiązki do stacjonarnego lustra, a drugą połowę przepuszcza do lustra ruchomego. Po odbiciu od luster, ramiona podzielonej wiązki wracają z powrotem do rozdzielacza, gdzie ulegają interferencji. Tak powstała wiązka kierowana jest do komory próbki, gdzie zostaje częściowo zaabsorbowana. Najistotniejszym elementem opisanego układu jest ruchome lustro, które przez swój ruch wprowadza różnice dróg optycznych pomiędzy rozdzielonymi wiązkami, zmieniając tym samym warunki zachodzącej interferencji. Intensywność wiązki, odczytywana na detektorze, jest zatem funkcją różnicy dróg optycznych i nazywana jest interferogramem [1], [11]. Aby z otrzymanego interferogramu otrzymać widmo próbki (lub referencyjne) w dziedzinie częstotliwości (liczby falowej), konieczne jest zastosowanie operacji matematycznej zwanej transformatą Fouriera, jak to jest przedstawione na rysunku 4b. Istotną zaletą tej metody jest szybkość zbierania danych, co wynika z faktu, że cały zakres spektralny jest pokrywany jednocześnie. W związku z tym, pomiary interferogramów są wykonywane wielokrotnie, a następnie uśredniane, co skutkuje wysoką jakością sygnału. Szacuje się, że zależność stosunku sygnału do szumu (*SNR*, ang. *signal to noise ratio*) jest następująca [1]

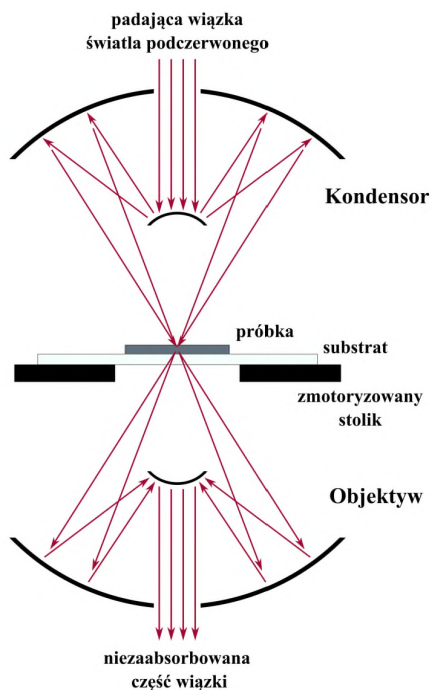
$$SNR \sim \sqrt{N}, \quad (15)$$

gdzie N to liczba uśrednionych interferogramów.

1.4. Aspekty uzyskiwania obrazu

Klasyczne spektrometry FT-IR opisane w poprzednim paragrafie umożliwiają pomiary objętościowe, dając uśredniony sygnał z dużego obszaru próbki (średnica rzędu kilkunastu milimetrów). Dlatego obecnie działają głównie w sprzężeniu z mikroskopami FT-IR, które łączą informacje o składzie chemicznym z informacją przestrzenną oraz zapewniają znacznie korzystniejszą skalę czasową pomiarów, poprzez zastosowanie detektorów macierzowych.

W przypadku takiego układu, wiązka ze spektrometru (interferometru) kierowana jest do mikroskopu FT-IR, którego budowa oparta jest o parę obiektywów Cassegraina, gdzie jeden jest zwyczajowo nazywany kondensorem, a drugi obiektywem [13], [14]. Schemat układu mikroskopowego do pomiarów w trybie transmisji (wykorzystywanym w tej pracy) przedstawiony jest na rysunku 5.



Rysunek 5. Schemat obiektywów Cassegraina do transmisyjnych pomiarów w podczerwieni.

Powiązanie widma z informacją przestrzenną (położeniem X, Y) jest możliwe dzięki zastosowaniu zmotoryzowanego stolika, na którym umieszczony jest substrat z próbką. Po zebraniu widma (lub widm), stolik przesuwany jest do kolejnej pozycji, często tworząc w ten sposób mapę punktów/widm. Liczba widm zbieranych podczas pojedynczego pomiaru uzależniona jest od zastosowanego rodzaju detektora. W przypadku detektorów punktowych zbierane jest pojedyncze widmo, a pokrywanie większego obszaru pomiarowego nazywane jest mapowaniem. Rozmiar pojedynczego piksela jest tutaj uzależniony od niezależnych od obiektywów apertur i jest ustalany przez użytkownika, w zależności od potrzeb pomiarowych. Stosowane są tutaj rozmiary piksela rzędu od $250 \times 250 \mu\text{m}^2$, do nawet $1 \times 1 \mu\text{m}^2$ (kosztem dużego spadku intensywności). Każde z tak zebranych widm reprezentuje uśredniony sygnał z danego

obszaru. Punktowe mapowanie staje się jednak coraz mniej popularne ze względu na dostępność detektorów macierzowych, których zastosowanie znacznie skraca skalę czasową pomiarów i pozwala osiągnąć piksele o rozmiarach rzędu μm , bez strat związanych z jakością sygnału. Najczęściej wykorzystywane są detektory ogniskowej matrycy (FPA, ang. *Focal Plane Array*), posiadające wiele niezależnych elementów detekcyjnych. Klasycznie stosowane rozmiary matrycy to 64×64 lub 128×128 pikseli, co skutkuje jednoczesnym pomiarem 4 096 lub 16 384 widm. Po zebraniu zestawu danych, stolik przemieszcza się do innej lokalizacji. Niezależne zestawy (kafelki) zostają następnie łączone w jeden plik pomiarowy (ang. *stitching*), a opisana procedura pomiarowa jest powszechnie nazywana obrazowaniem. Finalny zestaw danych ma trzy wymiary, gdzie dwa pierwsze (X, Y) dają informację przestrzenną, a trzeci zawiera kontrast chemiczny. Najczęściej stosowanymi typami detektorów, zarówno punktowych jak i macierzowych, są detektory MCT (ang. *Mercury Cadmium Telluride*), wymagające chłodzenia ciekłym azotem.

Rozmiary projekcji pikseli w obrazowaniu FT-IR z wykorzystaniem obiektywów Cassegraina są zdeterminowane przez powiększenie stosowanych obiektywów oraz fizyczny rozmiar każdego piksela detektora (wynoszący $40 \mu\text{m}$). Tym samym powszechnie stosowane obiektywy $15\times$ i $36\times$ dają projekcje pikseli o rozmiarach $2,7 \mu\text{m}$ i $1,1 \mu\text{m}$. Rozmiar piksela nie powinien być jednak mylony z rozdzielczością przestrzenną. Kluczowym parametrem obiektywu jest jego apertura numeryczna NA , która wraz z długością fali padającego promieniowania λ definiuje granicę przestrzennej zdolności rozdzielczej. Ten limit dyfrakcji, wyrażony przez kryterium Rayleigh'a, jest następujący [15]

$$d = \frac{0,61 \cdot \lambda}{NA} \quad (16)$$

i mówi o tym, że obiekty oddalone od siebie o odległość mniejszą niż d są nierozróżnialne. Warto zauważyć, że w związku z powyższym, zdolność rozdzielcza zmienia się w całym zakresie widmowym. Zakładając przy tym, że standardowe wartości NA są w okolicy 0,5, można uznać, że granica dyfrakcji jest porównywalna do długości fali λ . Dla przykładu, w zakresie *fingerprint* będzie to aż 7 do $15 \mu\text{m}$, natomiast w okolicy 3000 cm^{-1} granica dyfrakcji jest na poziomie $3,3 \mu\text{m}$. Stosowane rozmiary pikseli są zdecydowanie mniejsze od limitu zdolności rozdzielczej, co wynika z zastosowania kryterium mówiącego o tym, jaka jest optymalna częstotliwość próbkowania przestrzennego d_c [16]. Dla układu dwóch obiektywów Cassegraina wielkość ta wynosi [16]

$$d_c = \frac{0,25 \cdot \lambda}{NA}, \quad (17)$$

co oznacza, że optymalny rozmiar piksela jest 2.44 razy mniejszy od limitu dyfrakcji. Obiektywy dające częstość próbkowania (rozmiar piksela) spełniającą powyższe kryterium nazywamy obiektywami HD (ang. *High Definition*).

1.5. Spolaryzowana spektroskopia w podczerwieni

Promieniowanie podczerwone padające na próbkę, w przypadku spektroskopii w podczerwieni, może przyjmować różne stany polaryzacji światła. Klasyczne obrazowanie FT-IR wykorzystuje źródło termiczne emitujące niespolaryzowane promieniowanie. Polaryzatory liniowe są jednak standardowym elementem wyposażenia spektrometrów, a ich

zastosowanie umożliwia uzyskanie promieniowania spolaryzowanego liniowo, w dowolnym kierunku prostopadłym do kierunku padania wiązki. [11]

To właśnie składowa elektryczna promieniowania elektromagnetycznego oddziałuje z momentami dipolowymi, a absorbanca osiąga wartość maksymalną w przypadku, gdy kierunek liniowej polaryzacji jest równoległy do wektora dipolowego momentu przejścia danego modu oscylacyjnego [11]. Ta właściwość może być wykorzystana do analizy orientacji molekuł, których ułożenie w próbce przyjmuje pewien preferencyjny kierunek. Takie próbki nazywane są anizotropowymi, a zjawisko zmiennej absorbancji dla prostopadłych kierunków polaryzacji nazywane jest dichroizmem. Szczególnie łatwo można je zaobserwować w przypadku próbek włóknistych lub posiadających długie łańcuchy. Metoda ta była często wykorzystywana do analiz orientacji łańcuchów w rozciąganych foliach polimerowych, poprzez pomiary absorbancji dla polaryzacji równoległej $A_{\parallel}$ i prostopadłej $A_{\perp}$ do kierunku rozciągania oraz wyznaczenie dla nich stosunku dichroicznego $A_{\parallel}/A_{\perp}$ [17], [18]. Izotropowe próbki nie wykazują zależności absorbancji od zmiany kierunku liniowej polaryzacji. [11]

1.6. Inne metody spektroskopii IR dalekiego pola – O-PTIR i QCL

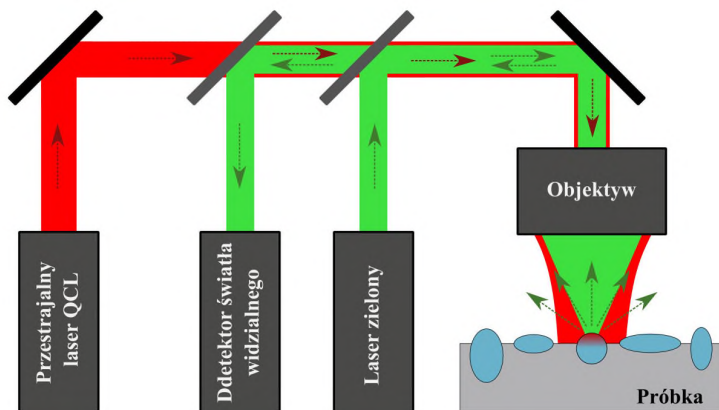
Spektroskopia FT-IR, będąca w głównym stopniu tematem tej rozprawy, pozostaje cały czas najchętniej stosowaną techniką obrazowania IR. Wynika to z faktu dostarczania bogatej informacji o składzie biochemicznym badanej próbki, płynącej z wykorzystania kontrastu chemicznego zapewnionego przez szeroki zakres częstości falowych obecnych w widmach FT-IR. Cecha ta jest następstwem zastosowania szerokopasmowego źródła światła podczerwonego w połączeniu z interferometrycznym układem detekcyjnym, opisanym we wcześniejszych rozdziałach. Jednak jak każda technika pomiarowa posiada swoje ograniczenia, które mogą zostać skompensowane przez inne metody spektroskopowe dające równoważne informacje.

Jednym z ograniczeń FT-IR jest stosunkowo długi czas trwania pomiarów wynikający z konieczności zbierania danych w pełnym zakresie widmowym. Dlatego w zastosowaniach, gdzie informacja spektralna może być ograniczona do wybranych pasm absorpcyjnych, rozwiązanie stanowią mikroskopy QCL [19]. Podstawy fizyczne działania mikroskopów QCL, tak samo jak FT-IR, są oparte o prawo Lamberta-Beera opisujące absorpcję promieniowania przez próbkę. Kluczowe różnice obserwowane są jednak w budowie obu typów mikroskopów. Mikroskopy QCL jako źródło promieniowania wykorzystują kwantowe lasery kaskadowe (ang. *Quantum Cascade Laser*), charakteryzujące się o rzędu wielkości większą intensywnością emitowanego promieniowania od niekoherentnych źródeł termicznych (np. *globar'ów*) [20]. Ponadto, emitowane jest promieniowanie odpowiadające konkretnej długości fali, z możliwością przestrajania lasera w szerokim zakresie spektralnym, pokrywającym konkretne obszary średniej podczerwieni. Zakres przestrajalności lasera jest uzależniony od dostępnych modów (ang. *chip's*) lasera, a konfiguracja zależy od użytkownika. Emitowana wiązka pada następnie na próbkę, zostaje częściowo zaabsorbowana, a finalnie kierowana jest bezpośrednio do detektora (w przeciwieństwie do FT-IR nie występuje tutaj interferometryczna detekcja). W obrazowaniu w podczerwieni przy użyciu laserów QCL stosowane są detektory punktowe oraz macierzowe (bolometry 480x480 lub FPA 128x128), które dzięki dużej mocy źródła laserowego nie wymagają chłodzenia ciekłym azotem, a czas ekspozycji jest zredukowany do minimum [20], co znacznym stopniu ułatwia to procedury pomiarowe. Probka umieszczona jest na zmotoryzowanym stoliku, dlatego po zebraniu danych dla jednego kafelka pomiarowego, odpowiadającego rozmiarowi matrycy detektora, jest systematycznie przesuwana, umożliwiając tym samym pokrycie dużych obszarów badawczych. Po

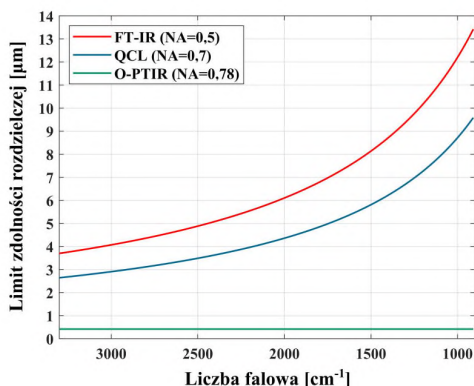
zakończeniu zbierania danych dla danej częstotliwości widmowej, laser jest przestrajany i procedura obrazowania jest powtarzana. Skutkuje to wysokoprzepustowymi pomiarami, skupionymi na konkretnym kontraście chemicznym odpowiadającym wybranym długościom fali (pasmom absorpcyjnym) [20]. Dla przykładu, czas pomiaru 14 częstotliwości dla próbki tkanki o rozmiarach $16 \times 35 \text{ mm}^2$ wynosi około 5 min, przy wykorzystaniu komercyjnego mikroskopu Spero-QT. Jest to szczególnie atrakcyjne w badaniach wykorzystujących tkanki do budowy modeli histopatologicznych, w oparciu o obrazowanie IR i metody uczenia maszynowego [21], [22]. Dodatkową zaletą mikroskopów QCL jest możliwość wykorzystania obiektywów refrakcyjnych, dających lepszą rozdzielczość przestrzenną (NA nawet 0,9) w porównaniu do refleksyjnych obiektywów Cassegraina stosowanych w FT-IR. Różnice pomiędzy mikroskopami QCL i FT-IR dobrze obrazuje fakt mówiący, że przy pomiarach FT-IR mierzone są widma, a obrazy są ich rekonstrukcją, natomiast w przypadku QCL jest odwrotnie – wykonywane jest obrazowanie (pasmo po paśmie), a widma są rekonstruowane. Jednakże, rozwój w dziedzinie technologii laserowej przyczynił się do powstania szybko przestrajalnych laserów, co sprawiło, że obecne procedury pomiarowe umożliwiają również atrakcyjne czasowo pomiary widmowe.

Poza czasem pomiarowym, kolejnym aspektem spektroskopii FT-IR limitującym obszar jej możliwych zastosowań jest rozdzielczość przestrzenna. Granica rozdzielczości przestrzennej opisywana jest przez kryterium Rayleigh'a (równanie 16), co dla środkowej podczerwieni daje możliwość rozróżniania obiektów oddalonych od siebie o dystans rzędu mikrometrów. Poprawa rozdzielczości była zatem motywacją do powstania mikroskopii O-PTIR, wykorzystującej efekt fototermicznej ekspansji próbki oraz unikatowy system detekcyjny oparty o wiązkę światła widzialnego [23], [24]. Mikroskopy O-PTIR jako źródło światła wykorzystują przestrajalne lasery QCL pracujące w trybie pulsacyjnym. Emitowane promieniowanie podczerwone jest skupiane na próbce co skutkuje jego absorpcją. W wyniku absorpcji energia promieniowania jest zamieniona w ciepło, a to wywołuje efekt termicznego rozszerzania się próbki oraz zmianę jej współczynnika załamania światła. Współosiowo do promieniowania podczerwonego biegnie wiązka zielonego lasera (532 nm), również skupiana na powierzchni próbki i wykorzystywana jako sonda detekcyjna, która po odbiciu od powierzchni wraca do detektora. Zmiany w intensywności wiązki światła zielonego (wynikające z ekspansji oraz zmiany współczynnika załamania) są proporcjonalne do absorpcji światła podczerwonego [4]. Skala tych zmian jest jednak bardzo mała, dlatego niezbędne jest użycie wzmacniacza homodynowego (ang. *lock-in amplifier*), który demoduluje sygnał poprzez zastosowanie sygnału referencyjnego dopasowanego do częstotliwości pulsów lasera QCL [4]. Taka demodulacja sygnału, w funkcji długości fali emitowanej przez źródło, umożliwia zebranie widma absorpcyjnego, a zastosowanie zmotoryzowanego stolika pozwala na obrazowanie dowolnych częstotliwości widmowych. Schemat opisanego układu przedstawiony jest na rysunku 6.

Wykorzystanie w metodzie O-PTIR widzialnego lasera jako sondy definiuje limit dyfrakcji na poziomie około $0,45 \text{ }\mu\text{m}$, umożliwiając tym samym obrazowanie z sub-mikronową rozdzielczością przestrzenną. Wartość ta jest stała w całym zakresie widmowym, w przeciwieństwie do mikroskopów QCL i FT-IR. Porównanie granicy rozdzielczości przestrzennej dla wszystkich trzech wspomnianych metod (przy założeniu komercyjnie stosowanych obiektywów) jest przedstawione na rysunku 7. Wyraźnie widoczne są tutaj zalety metody O-PTIR.



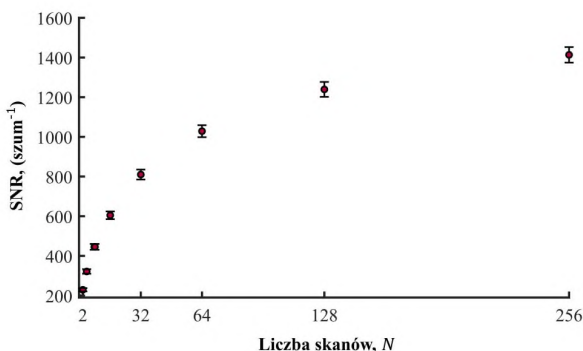
Rysunek 6. Schemat układu O-PTIR, wykorzystującego laser QCL do wywołania fototermicznego rozszerzania próbki, którego detekcja odbywa się przez zastosowanie zielonego lasera jako sondy. Otrzymany sygnał przedstawia kontrast chemiczny równoważny z widmami IR absorpcyj.



Rysunek 7. Porównanie limitu dyfrakcji dla standardowych obiektywów wykorzystywanych w systemach FT-IR, mikroskopach QCL oraz O-PTIR, dla zakresu spektralnego odpowiadającego średniej podczerwieni.

2. Metody wstępnego przetwarzania danych

Widma w podczerwieni będące wynikiem transformaty Fouriera, jak każdy sygnał, nie są idealne - posiadają artefakty wykluczające wykorzystanie ich surowej formy do ilościowych czy jakościowych analiz. Najbardziej podstawowymi i powszechnie stosowanymi krokami wstępnego przetwarzania danych w spektroskopii w podczerwieni jest korekta linii bazowej oraz normalizacja [25], [26]. Korekta linii bazowej ma na celu usunięcie artefaktów związanych z rozpraszaniem. Obecność procesów rozpraszania sprawia, że w eksperymencie dokonuje się pomiaru ekstynkcji, zamiast pożądanego sygnału związanego z absorpcją. Dodatkowo pojawiają się efekty związane z odbiciem czy anomalie sprzętowe, mające wpływ na kształt linii bazowej. Normalizacja polega na podzieleniu widma przez stałą wartość (np. absorpcję jednego z pasm), zapewniając tym samym niezależnienie wyników od grubości mierzonej próbki. Oba opisane kroki nie będą jednak skuteczne jeśli jakość posiadanego sygnału widmowego nie będzie odpowiednia, czego przyczyną jest obecność szumu [27]. Powszechnie stosowanym parametrem oceny jakości sygnału jest SNR . W rozdziale 1.3 wspomniano, że SNR jest proporcjonalny do pierwiastka ilości uśrednionych skanów, dlatego jednym z rozwiązań umożliwiających uzyskanie dobrej jakości widm jest wielokrotne powtarzanie danego pomiaru. Przykład takiej zależności jest przedstawiony na rysunku 8. Jest to jednak bardzo czasochłonna procedura, która może zostać zastąpiona zastosowaniem odpowiednich metod odszumiania.



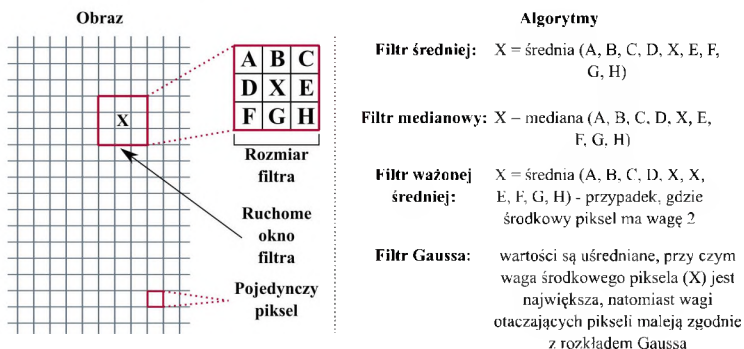
Rysunek 8. Charakterystyka zależności SNR od ilości skanów. Każdy punkt opisuje uśrednioną wartość SNR dla niewielkiego obszaru pomiarowego, a słupki błędów opisują odchylenie standardowe.

Odszumianie powinno być pierwszym krokiem łańcucha wstępnego przetwarzania danych, a jego zastosowanie staje się coraz popularniejsze w środowisku spektroskopowym, ze względu na zyski czasowe oraz pozytywny wpływ na dalsze analizy. Wybór konkretnej metody odszumiania może być uwarunkowany rodzajem szumu obecnego w analizowanych danych.

W przypadku spektroskopii FT-IR obserwowany jest szum multiplikatywny, wynikający z obecności stałego szumu, zarówno w sygnale pochodzącym od próbki jak i referencyjnym, których stosunek daje widmo transmitancji [27]. Dlatego poniższy rozdział przytacza różne rodzaje technik odszumiania spektralnego i przestrzennego, popularne w środowisku przetwarzania sygnałów.

2.1. Metody odszumiania

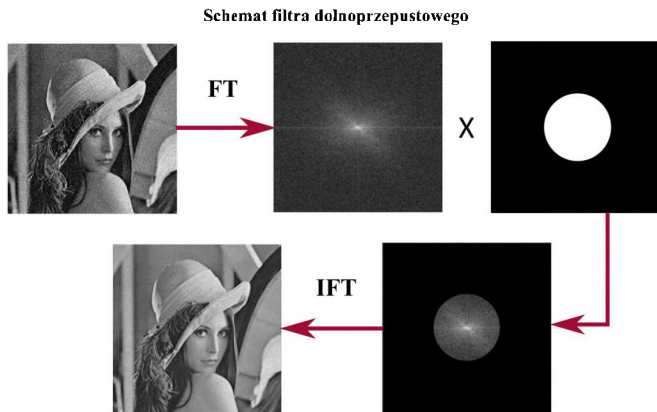
Ze względu na trójwymiarowy charakter danych, inne techniki odszumiania będą stosowane w przypadku widm, a inne w przypadku obrazów, chociaż istnieje szereg technik mających zastosowanie w obu przypadkach. Jednym z takich przykładów są klasyczne filtry, np. średniej, medianowy, Gaussa czy filtr średniej ważonej [28], [29]. Schemat działania tych filtrów jest przedstawiony na rysunku 9, gdzie odzwierciedlone jest działanie algorytmu na obrazach – przypadek 2D. Wygładzanie sygnału za pomocą filtrów jest stosunkowo proste, głównie ze względu na małą ilość parametrów danego filtra – często jest to tylko rozmiar ramki. W przypadku widm zastosowanie jest bardzo podobne, jednak ruchome okno filtra ma kształt wektora danych, a nie macierzy.



Rysunek 9. Mechanizm działania filtrów 2D: średniej, medianowy, średniej ważonej oraz Gaussa.

Kolejnym rodzajem filtrów, które mają zastosowanie w odszumianiu, są klasyczne filtry oparte o transformatę Fouriera (FT) [29]. Transformata Fouriera umożliwia przedstawienie sygnału w dziedzinie częstotliwości, a następnie wzmocnienie lub osłabienie konkretnych składowych, w zależności od zastosowania filtra. W przypadku odszumiania, stosowany jest filtr dolnoprzepustowy, gdzie składowe sygnału o niskich częstotliwościach zostają zachowane, natomiast intensywność składowych wysokoczęstotliwościowych (szumu) zostają stłumione (wycięte). Tak powstałe dane częstotliwościowe zostają poddane odwrotnej transformacji Fouriera (IFT), gdzie zrekonstruowany zostaje sygnał oparty jedynie na niskich częstotliwościach (zostaje pozbawiony szumu). Parametrem wymagającym optymalizacji jest tutaj szerokość filtra, która definiuje zakresy odrzucanych częstotliwości. Taki filtr może być zastosowany zarówno do danych jedno, jak i dwuwymiarowych (widm i obrazów), poprzez

wykorzystanie algorytmu jedno lub dwuwymiarowej transformaty Fouriera. Przykład działania filtru 2D przedstawiony jest na rysunku 10.

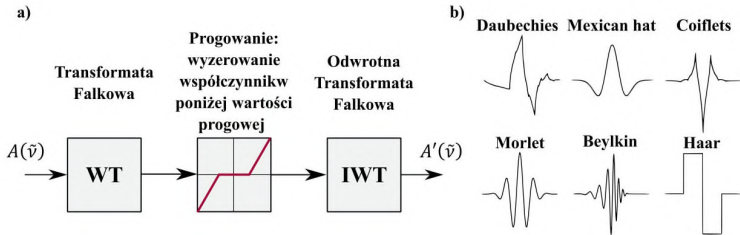


Rysunek 10. Działanie dwuwymiarowego filtra dolnoprzepustowego na przykładzie standardowego obrazu testowego Lenny.

Jedną z prostych i popularnych metod odszumiania w środowisku spektroskopowym, jest zastosowanie filtra Savitzky'ego-Golay'a (SG) [30], [31]. Metoda ta oparta jest o dopasowywanie wielomianu o zadanym stopniu do zadanej ilości punktów pomiarowych (np. fragmentu widma) nazywanych oknem. Okno filtru porusza się wzdłuż sygnału, powtarzając procedurę dopasowania, aż pokryty zostanie cały zakres widmowy. Dopasowanie polega na minimalizacji błędu średniokwadratowego pomiędzy odszumianym sygnałem, a dopasowanym wielomianem. Warto zauważyć, że filtr ten działa na sygnał w sposób lokalny i jest skuteczny w zachowywaniu kształtu sygnału, jednak aby to osiągnąć bardzo ważny jest dobór odpowiedniego rozmiaru okna i stopnia wielomianu. W spektroskopii w podczerwieni jest on najczęściej stosowany jako filtr spektralny, chociaż istnieje możliwość zastosowania go do danych 2D.

Kolejną skuteczną metodą stosowaną dla sygnałów i obrazów jest odszumianie za pomocą falek (ang. *wavelet denoising*) [32], [33]. Jest to technika często wykorzystywana przez środowisko przetwarzania sygnałów, mająca zastosowanie w spektroskopii w podczerwieni. Falka to funkcja matematyczna (rysunek 11b), której celem jest lokalne odzwierciedlenie wzorców o różnej częstotliwości, obecnych w analizowanym sygnale. Procedura odszumiania za pomocą falek składa się z kilku etapów. Po pierwsze, sygnał jest analizowany poprzez wykonanie transformaty falkowej (WT), której rezultatem jest wektor współczynników falkowych odzwierciedlających sygnał wyjściowy. Współczynniki o niskich wartościach są utożsamiane z szumem, dlatego w kolejnym kroku wykonywane jest progowanie mające na celu pozostawienie jedynie współczynników falkowych o wysokich wartościach (utożsamianych z sygnałem). Ostatnim krokiem jest synteza falkowa, która polega na wykonaniu odwrotnej transformaty falkowej (IW^T) wykorzystując współczynniki pozostałe po progowaniu. Kluczowym aspektem, pozwalającym otrzymać skutecznie odszumiony sygnał jest odpowiednie dobranie parametru progowego, odpowiedzialnego za ilość komponentów, na

których oparta jest rekonstrukcja sygnału. Ponadto, dostępne implementacje algorytmu posiadają różne rodziny funkcji falkowych, a wybór konkretnej jest również traktowany jako parametr optymalizacyjny, zależy od analizowanego sygnału. Schemat odszumiania falkowego, wraz z przykładowymi funkcjami falkowymi, jest przedstawiony na rysunku 11. Warto zauważyć, że dużą zaletą tej metody jest dobre zachowanie krawędzi obecnych w sygnałach czy na obrazach.



Rysunek 11. a) Schemat odszumiania za pomocą falek. b) Przykłady powszechnie stosowanych funkcji falkowych.

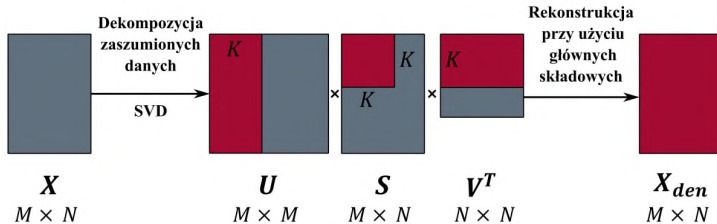
Kolejna z metod, Analiza Głównych Składowych (PCA, ang. *Principal Component Analysis*), jest powszechnie znaną metodą stosowaną w przypadku danych wielowymiarowych do ich eksploracji [34] i redukcji wymiarowości [35], ma też jednak zastosowanie w odszumianiu [36]. Algorytm polega na rozkładzie oryginalnych danych na czynniki zwane głównymi składowymi. Pomocne w opisie algorytmu [37] jest wyobrażenie sobie macierzy początkowych danych składających się z M próbek (obserwacji), każdej opisanej przez N zmiennych (cech), jako chmurę M punktów w przestrzeni N wymiarowej. Celem PCA jest przetransformowanie oryginalnych danych, opisanych przez skorelowane zmienne, w zestaw danych opisanych mniejszą liczbą cech (głównych składowych), które zachowują maksymalną ilość pierwotnej informacji. Szukanie głównych składowych można wyobrazić sobie jako rotację układu współrzędnych tak, aby znaleźć kierunek (oś) dla którego zmaksymalizowana zostaje wariancja danych. Taki kierunek można nazwać pierwszą główną składową (PC1). W dalszym etapie szukana jest kolejna główna składowa (PC2), opisująca maksymalną ilość wariancji nieopisaną przez składową pierwszą. Taka procedura jest kontynuowana maksymalnie N razy. Kluczowym aspektem jest tutaj fakt, że każdy z tak znalezionych kierunków (głównych składowych) jest względem siebie ortogonalny, co zapewnia brak korelacji pomiędzy pierwotnie skorelowanymi zmiennymi. Ponadto, każda z głównych składowych jest liniową kombinacją oryginalnych zmiennych. Algebraicznie, dekompozycja prostokątnej macierzy oryginalnych danych X może być przeprowadzona z wykorzystaniem rozkładu według wartości osłbiwych (SVD, ang. *Singular Value Decomposition*), umożliwiające przedstawienie X jako iloczynu trzech macierzy [38]

$$X = U \cdot S \cdot V^T, \quad (18)$$

gdzie U to macierz lewych wektorów osłbiwych, S to macierz wartości osłbiwych, natomiast V to macierz prawych wektorów osłbiwych. Wartości osłbiwe mówią o tym jaki procent

wariancji jest opisany przez daną główną składową. Macierz będąca iloczynem U oraz S jest w nomenklaturze nazywana macierzą wyników, a jej poszczególne kolumny opisują współrzędne (rzuty) danych próbek na ośiach konkretnych głównych składowych. Dla przykładu, pierwsza kolumna opisuje rzut punktów na oś wyznaczającą kierunek pierwszej głównej składowej. Macierz wyników opisuje więc położenie próbek w układzie współrzędnych głównych składowych, jednak nie daje informacji na temat samego układu. Macierz V opisuje położenie osi głównych składowych względem oryginalnego układu współrzędnych, jest więc niezbędna do rekonstrukcji lub rzutowania zewnętrznych próbek na nową przestrzeń. Kierunek osi każdej z głównych składowych jest wyrażony w V jako cosinus kąta pomiędzy daną osią oraz każdą z osi oryginalnego układu współrzędnych. Wiersze macierzy V^T nazywane są wagami, a każda z wartości opisuje jaki jest wpływ pierwotnej zmiennej na daną główną składową.

Zastosowanie metody PCA do odszumiania związane jest z faktem, że sygnał opisany jest przez zestaw kilku pierwszych głównych składowych, które pokrywają niemal cały zakres wariancji danych, natomiast pozostałe składowe utożsamiane są z szumem. Dlatego w celu odszumienia X wykonywany jest opisujący powyżej rozkład SVD, jednak do rekonstrukcji zestawu danych wykorzystywana jest tylko zadana ilość (K) pierwszych głównych składowych. Ten proces jest schematycznie przedstawiony na rysunku 12. Odszumione dane są tego samego rozmiaru co pierwotne, jednak ze zredukowanym szumem.



Rysunek 12. Schemat odszumiania metodą PCA, wykorzystującą metodę rozkładu według wartości osobliwych.

Metoda Najmniejszej Frakcji Szumu (MNF, ang. *Minimum/Maximum Noise Fraction*) [39], obok PCA, jest kolejną metodą mającą zastosowanie do odszumiania danych wielowymiarowych. Takie też było pierwotne przeznaczenie metody, najczęściej stosowanej do odszumiania danych satelitarnych [39]. Działanie algorytmu jest mocno zbliżone do PCA, ma jednak kilka kluczowych i wyróżniających go kroków. Istnieją różne implementacje algorytmu [40], a wariant opisany tutaj, podobnie jak PCA, bazuje na SVD. Zakładając, że surowe dane zawarte są w macierzy X , pierwszym krokiem algorytmu będzie obliczenie macierzy szumu. Jest kilka sposobów oszacowania macierzy szumu dX , a jeden z nich zakłada, że jest ona wyznaczana przez odjęcie od siebie widm odpowiadających sąsiadującym pikselom. Macierz szumu jest niezbędna do operacji wybielania szumu, poprzez wykorzystanie następującej własności SVD

$$V_1 \cdot S_1^2 \cdot V_1^T = dX^T \cdot dX. \quad (19)$$

Uzyskane poprzez dekompozycję SVD macierze $\mathbf{S}$ oraz $\mathbf{V}$ są następnie wykorzystywane do wybielenia szumu oryginalnych danych w następujący sposób [40]

$$\mathbf{wX} = \mathbf{X} \cdot \mathbf{V}_1 \cdot \mathbf{S}_1^{-1}. \quad (20)$$

Dzięki pomnożeniu oryginalnych danych $\mathbf{X}$ przez $\mathbf{V}_1 \cdot \mathbf{S}_1^{-1}$, szum dla wszystkich zmiennych $\mathbf{wX}$ jest nieskorelowany oraz posiada jednostkowe wartości [40]. Tak otrzymane dane poddawane są docelowej dekompozycji [40]

$$\mathbf{V}_2 \cdot \mathbf{S}_2^2 \cdot \mathbf{V}_2^T = \mathbf{wX}^T \cdot \mathbf{wX}, \quad (21)$$

dzięki której może zostać wyznaczony zestaw wektorów bazowych w przestrzeni MNF [40]

$$\mathbf{Phi} = \mathbf{V}_1 \cdot \mathbf{S}_1^{-1} \cdot \mathbf{V}_2. \quad (22)$$

Aby przetransformować oryginalne dane do tej przestrzeni, wystarczy pomnożyć je przez macierz nowych wektorów bazowych [40]

$$\mathbf{mnfX} = \mathbf{X} \cdot \mathbf{Phi}. \quad (23)$$

Celem powyższych operacji jest odszumienie danych, zatem transformacja odwrotna powinna bazować jedynie na głównych składowych zapewniających jak najwyższy poziom SNR. Jest to zrealizowane za pomocą diagonalnej macierzy maski $\mathbf{R}$, zawierającej jedynki dla składowych wykorzystanych do rekonstrukcji, oraz zera dla składowych opisujących szum [40]

$$\mathbf{denX} = \mathbf{mnfX} \cdot \mathbf{R} \cdot \mathbf{Phi}^{-1}, \quad (24)$$

gdzie $\mathbf{denX}$ to macierz odszumionych danych.

Główną różnicą pomiędzy PCA i MNF jest to, że PCA jest skupione na maksymalizowaniu wariancji danych, natomiast MNF dąży do osiągnięcia jak najwyższych wartości SNR. Warto zwrócić uwagę, że obie metody, MNF i PCA, bazują na całych (wielowymiarowych) zestawach danych otrzymywanych w ramach mapowania lub obrazowania spektroskopowego. Nie nadają się one do odszumiania pojedynczych widm, co odróżnia je od pozostałych opisywanych metod.

2.2. Metody korekty linii bazowej.

Po odszumianiu, kolejnym krokiem wstępnego przetwarzania danych, umożliwiającym wyekstrahowanie z widma jakościowych oraz ilościowych informacji, jest korekta linii bazowej. Dla idealnego widma, w obszarach wolnych od pasm absorpcyjnych, oczekiwane są wartości absorbancji równe zero, ponieważ w teorii żadna część promieniowania podczerwonego (z danego zakresu spektralnego) nie uległa oddziaływaniu z próbką. Tym samym, w teorii całość promieniowania trafia do detektora. Obszar bez pasm absorpcyjnych pozwala określić tak zwaną linię bazową, jednak opisany powyżej przypadek jest wyidealizowany. W praktyce, głównymi artefaktami zaburzającymi widma są efekty interferencji oraz rozpraszania. Interferencja jest efektem przeważającym w transfleksyjnym trybie pomiarowym [41], [42], a w tej pracy stosowany był jedynie tryb transmisyjny, gdzie dominują efekty związane z elastycznym rozpraszaniem polegające na zmianie kierunku padającego promieniowania (tym samym usunięcie go z wiązki padającej na detektor). Korekta

(odjęcie) artefaktów związanych z linią bazową jest zatem kluczowym krokiem wstępnej obróbki danych, a zastosowana metoda zależy od wielu czynników, takich jak kształt linii bazowej, pochodzenie artefaktów, zastosowanie danych, itp.

Powszechnie stosowane są linie bazowe oparte o funkcje liniowe, poczynając od linii nachylonych (ang. *sloping*) opartych o dwa punkty, przez odcinki korygujące fragmenty widma zawarte pomiędzy parami sąsiednich punktów (korekta typu *rubber band*), po lokalne linie dla danych pasm. Zamiast linii prostej stosowana może być też funkcja wielomianowa, przy czym stopień wielomianu ograniczony jest przez ilość punktów wykorzystywanych do dopasowania. Powyższe metody mogą być stosowane zarówno dla artefaktów związanych z rozpraszaniem, jak i interferencją, a ich ogromną zaletą jest prostota obliczeń. To sprawia, że implementacja jest łatwa, a krótki czas obliczeń umożliwia wykorzystanie do korekt dużych zestawów danych. Ich wadą jest jednak ograniczenie możliwości zastosowania do linii bazowych o stosunkowo prostych przebiegach, ponieważ metody te nie są oparte o modele fizyczne.

Różną algorytmów wykorzystujących do korekty linii bazowej informacje na temat parametrów fizycznych próbki są metody oparte o EMSC (ang. *Extended Multiplicative Signal Correction*). Metody te są skierowane do korekty rozpraszania Mie występującego w przypadku, gdy padające promieniowanie oddziałuje ze sferycznym obiektem (np. komórką) o rozmiarze zbliżonym do długości fali. W wyniku tego rozpraszania obserwowana jest silna sinusoidalna linia bazowa, uniemożliwiająca analizę widmową. Wiele wersji algorytmów EMSC zostało zaimplementowanych dla obiektów sferycznych, wykorzystując pełną teorię rozpraszania Mie [43], jak i jej uproszczenia (van de Hulsta) [44]–[46]. Obecnie algorytmy są rozszerzane o implementacje umożliwiające korekty rozpraszania typu Mie dla obiektów o innej geometrii, np. cylindrycznej (rysunek 13b). Ogólny schemat i zasada działania metod typu EMSC zostaną opisane poniżej na podstawie algorytmu RMieS-EMSC (ang. *Resonant Mie Scattering – EMSC*), który jest jednym z powszechnie stosowanych wariantów korekty.

Przewagą algorytmów typu EMSC jest wykorzystanie do korekty rozpraszania fizycznych parametrów badanej próbki, między innymi jej rozmiaru oraz współczynnika załamania światła. Po dekadach rozwoju, popularnym i względnie nowym wariantem metody jest algorytm RMieS-EMSC, będący rezultatem poznania zagadnień związanych z artefaktami wynikającymi z efektów dyspersyjnych [44], [45]. Artefakty te odpowiadają za ostry spadek absorpcji danego pasma po stronie wysokich liczb falowych, co wynika ze zmian wartości części rzeczywistej współczynnika załamania światła, a tym samym wpływa na zmniejszenie wydajności rozpraszania światła [44], [45]. Aby uwzględnić zmiany w rozpraszaniu, wywołane artefaktami dyspersyjnymi, algorytm RMieS-EMSC zawiera transformatę Kramersa-Kroniga (KK). Pozwala ona wyznaczyć rzeczywistą część współczynnika załamania światła na podstawie referencyjnego widma IR badanego materiału (związanego z częścią urojoną współczynnika załamania, jak to opisano w równaniu 14). Kombinacje różnych wartości współczynnika/ów załamania wraz z rozmiarami próbki, są przez algorytm wykorzystywane do wyznaczenia zestawów krzywych opisujących wydajność rozpraszania Q_{scat} . Ogólny model RMieS-EMSC zakłada, że dany sygnał (widmo) $\vec{Z}_{Raw}$ można rozłożyć (skorygować) według następującego równania liniowego [43], [44]

$$\vec{Z}_{Raw} = \vec{c} + m\vec{v} + h\vec{Z}_{Ref} + \sum_{i=1}^{N=7} \vec{p}_i g_i + \vec{E}, \quad (25)$$

gdzie $\vec{c}$ jest stałą a m nachyloną linią bazową, $\vec{v}$ to wektor liczb falowych, h jest czynnikiem skalującym widma referencyjnego $\vec{Z}_{Ref}$, N to liczba głównych składowych (wartość 7 zapewnia satysfakcjonujący poziom opisanej wariancji danych), $\vec{p}_i$ i g_i są wagami oraz ich czynnikami skalującymi opisującymi wpływ wydajności rozpraszania na widmo, natomiast $\vec{E}$ to wektor reszt. Korekta widma wykonywana jest poprzez znalezienie wartości c , m , h oraz g_i , wykorzystując do tego regresję liniową. Jest to jednak iteracyjny proces przedstawiony na rysunku 13d. Pierwsza iteracja, poza korygowanym widmem, wymaga zapewnienia algorytmowi widma referencyjnego (wolnego od artefaktów rozpraszania) dla danej próbki. Kolejne iteracje wykorzystują jako widmo referencyjne widmo będące wynikiem korekty w poprzedniej iteracji. Ten proces jest powtarzany określoną ilość razy (np. 10), lub do momentu ustabilizowania błędu. Kluczowym krokiem każdej iteracji jest znalezienie wartości $\vec{p}_i$, zawierających informacje o rozpraszaniu. W tym celu wyliczany jest zestaw krzywych wydajności rozpraszania Q_{sca} (zazwyczaj 1000 krzywych) dla różnych kombinacji parametrów fizycznych próbki. Taki zestaw jest następnie poddawany dekompozycji PCA, z której otrzymywane są wartości wag $\vec{p}_i$.

W przypadku próbek sferycznych, wydajność rozpraszania może być wyznaczana na dwa sposoby, stosując pełną teorię rozpraszania Mie lub stosując uproszczenie van de Hulsta. W przypadku pełnej teorii, wyznaczenie wydajności rozpraszania jest związane z kosztownymi czasowo obliczeniami sferycznych funkcji Bessela [43]. Ich zastosowanie daje dokładny obraz charakteru rozpraszania, ale równocześnie wymaga wykorzystania procesorów graficznych (GPU) w celu przyspieszenia obliczeń. Dlatego bardzo popularne jest stosowanie teorii van de Hulsta, w przypadku której obliczenie wydajności rozpraszania przez sfery sprowadza się do następującej formy [44]

$$\rho(\vec{v}) = \frac{2\pi d(a + b n_{KK}(\vec{v}) - 1)}{\lambda}, \quad (26)$$

$$Q_{RMieS} = 2 - \left(\frac{4}{\rho(\vec{v})}\right) \sin \rho(\vec{v}) + 4 \left(\frac{4}{\rho(\vec{v})^2}\right) (1 - \cos \rho(\vec{v})), \quad (27)$$

gdzie d to średnica cząstki, a to średni współczynnik załamania światła (rzeczywisty), n_{KK} to rzeczywisty współczynnik załamania światła wyznaczony na podstawie widma referencyjnego poprzez transformatę KK, b to czynnik skalujący dla n_{KK} .

Przypadek rozpraszania przez sfery, na przestrzeni lat, został dobrze poznany i opisany w literaturze. Z tego powodu, uwaga zostaje obecnie skupiona na innych geometriach, które również wywołują silne rozpraszanie typu Mie. Takim przypadkiem jest rozpraszanie przez próbki cylindryczne, których przykładem są np. włókna kolagenowe (rysunek 13a,b). Do wyznaczenia wydajności rozpraszania niezbędne jest tutaj zastosowanie pełnej teorii oraz konieczność uwzględnienia polaryzacji padającego promieniowania. W przypadku próbki umieszczonej w ośrodku o współczynniku załamania n_h , wydajność rozpraszania dla polaryzacji równoległej $Q_{sca,\parallel}$ oraz prostopadłej $Q_{sca,\perp}$ do osi cylindra jest następująca [47]

$$Q_{sca,\parallel} = \frac{2}{x} \left[|b_{0,\parallel}|^2 + 2 \sum_{l=1}^{\infty} (|b_{l,\parallel}|^2 + |a_{l,\parallel}|^2) \right], \quad (28)$$

$$Q_{sca,\perp} = \frac{2}{x} \left[|a_{0,\perp}|^2 + 2 \sum_{l=1}^{\infty} (|a_{l,\perp}|^2 + |b_{l,\perp}|^2) \right], \quad (29)$$

$$x = 2\pi n_h \tilde{v}r, \quad (30)$$

$$a_{l,\parallel} = \frac{C_l V_l - B_l D_l}{W_l V_l + i D_l^2}, \quad (31)$$

$$a_{l,\perp} = -\frac{A_l V_l - i C_l D_l}{W_l V_l + i D_l^2}, \quad (32)$$

$$b_{l,\parallel} = \frac{W_l B_l + i D_l C_l}{W_l V_l + i D_l^2}, \quad (33)$$

$$b_{l,\perp} = -i \frac{C_l W_l + A_l D_l}{W_l V_l + i D_l^2}, \quad (34)$$

$$A_l = i\xi [\xi J_l'(\eta) J_l(\xi) - \eta J_l(\eta) J_l'(\xi)], \quad (35)$$

$$B_l = \xi [\tilde{n}^2 \xi J_l'(\eta) J_l(\xi) - \eta J_l(\eta) J_l'(\xi)], \quad (36)$$

$$C_l = l \cos(\zeta) \eta J_l(\eta) J_l(\xi) \left(\frac{\xi^2}{\eta^2} - 1 \right), \quad (37)$$

$$D_l = l \cos(\zeta) \eta J_l(\eta) H_l^{(1)}(\xi) \left(\frac{\xi^2}{\eta^2} - 1 \right), \quad (38)$$

$$V_l = \xi \left[\tilde{n}^2 \xi J_l'(\eta) H_l^{(1)}(\xi) - \eta J_l(\eta) H_l^{(1)'}(\xi) \right], \quad (39)$$

$$W_l = i\xi \left[\eta J_l(\eta) H_l^{(1)'}(\xi) - \xi J_l'(\eta) H_l^{(1)}(\xi) \right], \quad (40)$$

gdzie r to promień cylindra, J_l i $H_l^{(1)}$ to sferyczne funkcje Besela i Hankela (apostrof oznacza pierwszą pochodną), l to rząd funkcji sferycznej, ζ to kąt pomiędzy kierunkiem padania promieniowania, a osią cylindra, $\xi = \tilde{x} \sin \zeta$, $\eta = \tilde{x} \sqrt{\tilde{n}^2 - (\cos \zeta)^2}$, $\tilde{n} = \frac{n_p}{n_h}$, n_p to współczynnik załamania dla próbki, $\tilde{x} = 2\pi \tilde{v}r$. Ze względu na brak możliwości wykonania

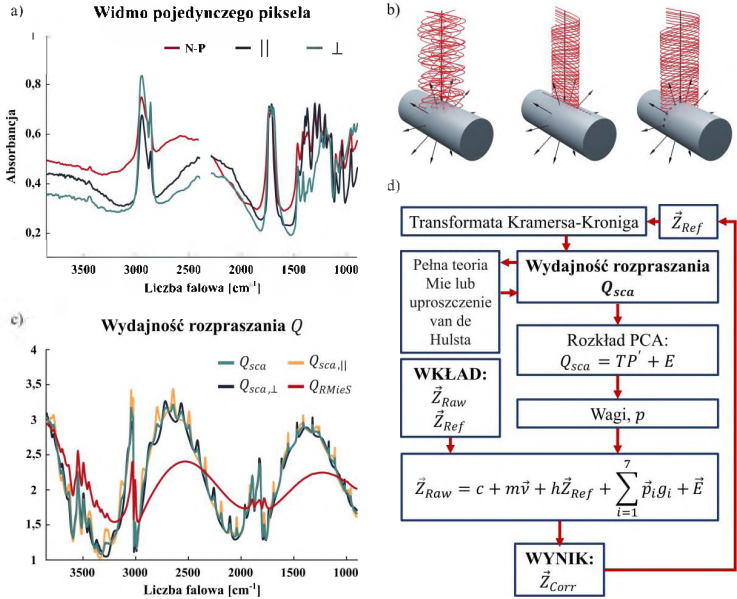
nieskończonego sumowania, przyjmowane jest poniższe kryterium końcowe dla równań numer 28 oraz 29 [47]

$$l_{max} = x + 4.05x^{1/3} + 2. \quad (41)$$

Dla przypadku padającego promieniowania charakteryzującego się brakiem polaryzacji, wydajność rozpraszania jest następująca [47]

$$Q_{sca} = \frac{1}{2} [Q_{sca,\parallel} + Q_{sca,\perp}]. \quad (42)$$

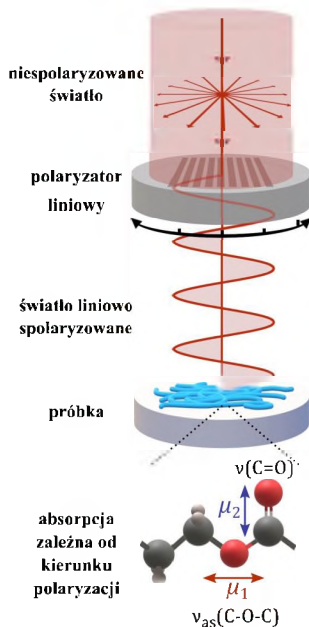
Przykładowe wydajności rozpraszania wyznaczone dla sfery oraz cylindrów (uwzględniając różne przypadki polaryzacji) przedstawione są na rysunku 13c.



Rysunek 13. a) Przykład silnego rozpraszania typu Mie dla tego samego piksela pochodzącego z próbki cylindrycznej oraz różnych wariantów polaryzacji światła (N-P – promieniowanie niespolaryzowane, || – polaryzacja równoległa do osi cylindra, ⊥ – polaryzacja prostopadła do osi cylindra). b) Wizualizacja trzech przypadków polaryzacji, od lewej: N-P, || oraz ⊥. c) Przykłady teoretycznych wydajności rozpraszania dla próbek cylindrycznej i trzech przypadków polaryzacji światła, oraz dla próbki sferycznej (Q_{RMieS}). d) Schemat blokowy algorytmów EMSC opisanych powyżej. [14]

3. Orientacja molekul

Spektroskopia FT-IR oraz jej siostrzane techniki (QCL, O-PTIR, itp.) same w sobie zapewniają bogactwo informacji biochemicznych na temat badanej próbki. Włączenie ponadto kontroli nad polaryzacją padającego światła podczerwonego oraz odpowiedniej teorii pozwala uzyskać nawet dogłębniejsze informacje na temat orientacji makromolekuł, które są szczególnie istotne np. z punktu widzenia biofizyki. Schemat układu FT-IR, umożliwiającego takie badania, przedstawiony jest na rysunku 14. Jak opisano w rozdziale 1.5, absorpcja przyjmuje maksymalną wartość, jeśli kierunek liniowej polaryzacji jest równoległy do kierunku dipolowego momentu przejścia. Dlatego liniowa polaryzacja była od dawna wykorzystywana do określania właściwości dichroicznych materiałów [17], [18]. Stosowana wtedy metoda dwóch polaryzacji (2P) posiada jednak duże ograniczenie, wynikające z konieczności zastosowania polaryzacji równoległej oraz prostopadłej do kierunku łańcucha molekuł. W przeciwnym wypadku otrzymane wartości współczynnika dichroicznego mogą być mylne. Postęp w tym obszarze badań nastąpił po zaproponowaniu metody czterech polaryzacji (4P).



Rysunek 14. Schemat układu do spektroskopii FT-IR z liniową polaryzacją. [F]

3.1. Orientacja na płaszczyźnie – 2D

Zastosowanie metody czterech polaryzacji umożliwia zdefiniowanie orientacji wiązań czy łańcuchów w płaszczyźnie, a co za tym idzie orientacji 2D całych makromolekuł, niezależnie od początkowego wyboru kierunku polaryzacji. Jest ona w głównej mierze oparta o poniższe równanie, opisujące zależność absorbancji $A(\eta)$ od kierunku polaryzacji η [48]–[50]

$$A(\eta) = \frac{A_{max} + A_{min}}{2} + \frac{A_{max} - A_{min}}{2} \cos 2(\eta - \psi), \quad (43)$$

gdzie A_{min} i A_{max} są minimalną i maksymalną absorbancją, ψ jest kierunkiem polaryzacji dla którego $A(\eta) = A_{max}$. Zbieranie czterech, przestrzennie nakładających się zestawów danych, każdy dla innego kierunku liniowej polaryzacji, pozwala znaleźć przebieg funkcji $A(\eta)$ dla każdego piksela. Taka rekonstrukcja wykonywana jest poprzez nieliniowe dopasowanie funkcji cosinus, a następnie znalezienie wartości A_{min} , A_{max} oraz ψ dla danego pasma (dipolowego momentu przejścia). Procedura jest graficznie przedstawiona na rysunku 15a. Rysunek 15a dobrze nakreśla informację na temat orientacji zapewnioną przez tą metodę. Moment dipolowy μ jest zorientowany w przestrzeni, natomiast parametr ψ daje ograniczoną informację na temat kierunku jego rzutu na płaszczyznę, nie odwidkując składowej kierunkowej poza płaszczyznę. Dodatkowa informacja może zostać uzyskana przez wyznaczenie wartości funkcji Hermansa [48]–[50]

$$f = \frac{D - 1}{D + 2} \frac{2}{3 \cos^2 \kappa - 1}, \quad (44)$$

gdzie $D = A_{max}/A_{min}$, natomiast κ to kąt pomiędzy główną osią molekuly, a kierunkiem zmiany momentu dipolowego. κ jest definiowane indywidualnie dla każdego pasma - zakładana jest wartość 0° jeśli kąt jest mniejszy od $\cong 54,73^\circ$ oraz 90° jeśli kąt przekracza wspomnianą wartość [48]–[50]. Funkcja Hermansa przyjmuje wartości od -0,5 do 1, a ze względu na swój związek z ODF (ang. *Orientation Distribution Function*), wartości te mogą informować o poziomie uprządkowania molekuł. Dla przykładu, wartość 1 mówi o tym, że momenty dipolowe są równoległe do osi molekuly, -0,5 oznacza, że są prostopadłe, natomiast wartości w okolicy zera oznaczają, że nie występuje uporządkowanie. Pomimo swoich ograniczeń, metoda 2D zapewnia bogactwo informacji na temat orientacji makromolekuł, co jest potwierdzone analizą próbek polimerowych [51] czy też włókna jedwabiu [52].

3.2. Orientacja molekuł – 3D

Ograniczona informacja o orientacji momentów dipolowych, dostarczona przez metodę 2D, została rozszerzona przez Lee, który sformułował teoretycznie metodę wyznaczania orientacji 3D [53]. Ta technika również bazuje na pomiarach z liniową polaryzacją światła, wykorzystując klasyczną teorię fal elektromagnetycznych oraz informacje o dwóch nierównoległych dipolowych momentach przejścia: głównym μ_1 oraz drugim μ_2 . Równania, o które oparta jest metoda, opisują zależność wartości absorpcyjnej ($\alpha = 1 - T$) od kierunku polaryzacji η dla obu momentów dipolowych [53]

$$\frac{\alpha_1(\eta)}{\alpha_1^0} = \langle P_2 \rangle \sin^2 \theta \cos^2(\eta - \psi) + \frac{1}{3} (1 - \langle P_2 \rangle), \quad (45)$$

$$\frac{\alpha_2(\eta)}{\alpha_2^o} = \langle P_2 \rangle [\cos\theta \cos\phi \cos(\eta - \psi) + \sin\phi \sin(\eta - \psi)]^2 + \frac{1}{3} (1 - \langle P_2 \rangle), \quad (46)$$

gdzie ψ , θ i ϕ to kąty opisujące orientację wektorów μ_1 i μ_2 , $\langle P_2 \rangle$ to parameter uporządkowania (ang. *order parameter*), natomiast α^o zawiera jedynie stałe wartości. Rysunek 15b zawiera wizualizację powyższych parametrów. Podobnie jak w przypadku metody 2D, do punktów eksperymentalnych dopasowywane są nieliniowe funkcje odzwierciedlające równania 45 i 46. Na podstawie otrzymanych funkcji, znajdowane są następujące wielkości: $\alpha_{1,max}$, $\eta_{1,max}$, $\alpha_{1,min}$, $\eta_{1,min}$, $\alpha_{2,max}$, $\eta_{2,max}$ i $\alpha_{2,min}$, $\eta_{2,min}$, według schematu przedstawionego na rysunku 15b. Te wielkości są wykorzystane do wyznaczenia następujących parametrów pośrednich [53]

$$M \equiv \left(\frac{\alpha_{1,max}}{\alpha_{1,min}} - 1 \right) = \frac{3\langle P_2 \rangle}{1 - \langle P_2 \rangle} \sin^2 \theta, \quad (47)$$

$$N \equiv \left(\frac{\alpha_{2,max}}{\alpha_{2,min}} - 1 \right) = \frac{3\langle P_2 \rangle}{1 - \langle P_2 \rangle} (\cos^2 \phi \cos^2 \theta + \sin^2 \phi), \quad (48)$$

$$\Delta\eta = \eta_{2,max} - \eta_{1,max}, \quad (49)$$

$$P = \frac{1}{2} \left[(M + N) + \sqrt{(M - N)^2 + 4MN \cos^2 \theta \Delta\eta} \right]. \quad (50)$$

Stosunki $\frac{\alpha_{1,max}}{\alpha_{1,min}}$ oraz $\frac{\alpha_{2,max}}{\alpha_{2,min}}$ w powyższych równaniach upraszczają wielkości α^o (stałe) obecne w równaniach 45 i 46. Znając parametry pośrednie oraz odpowiednio przekształcając równania, można wyznaczyć orientację wektorów μ_1 i μ_2 oraz parametr uporządkowania [53]

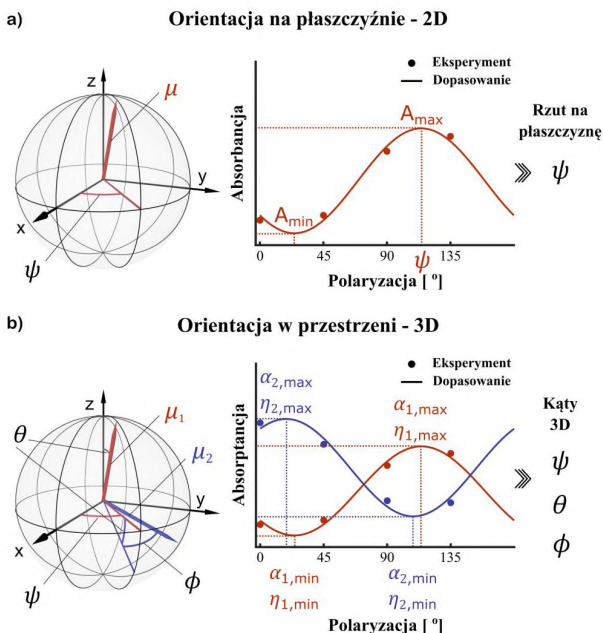
$$\langle P_2 \rangle = \frac{P}{P + 3}, \quad (51)$$

$$\psi = \eta_{1,max}, \quad (52)$$

$$\theta = \sin^{-1} \sqrt{\frac{M}{P}}, \quad (53)$$

$$\phi = \cos^{-1} \left[\pm \sqrt{\frac{(P - N)}{M}} \right]. \quad (54)$$

Znak w równaniu 56 jest zdefiniowany przez znak $\tan \Delta\eta$. Istotne jest, że ze względu na symetrię momentu przejścia względem płaszczyzny polaryzacji, istnieją dwa równoważne rozwiązania powyższych równań. Uwzględniając, że $0 \leq \theta_1 \leq \pi/2$ oraz $0 \leq \phi_1 \leq \pi$, drugie rozwiązanie będzie zdefiniowane następująco: $\theta_2 = \pi - \theta_1$ oraz $\phi_2 = -\phi_1$. Parametr $\langle P_2 \rangle$, podobnie jak funkcja Hermansa, daje informacje na temat ułożenia danego momentu dipolowego względem głównej osi molekuly. Przyjmuje on wartości od $-1/3$ do 1 , gdzie 1 oznacza orientację równoległą, natomiast $-1/3$ prostopadłą.



Rysunek 15. a) Wizualizacja zastosowania metody czterech polaryzacji do wyznaczania orientacji na płaszczyźnie (2D). b) Zastosowanie metody czterech polaryzacji do wyznaczania orientacji makromolekuł w przestrzeni, poprzez jednoczesne wykorzystanie dwóch dipolowych momentów przejścia. [F]

4. Klasyfikacja – Las Losowy

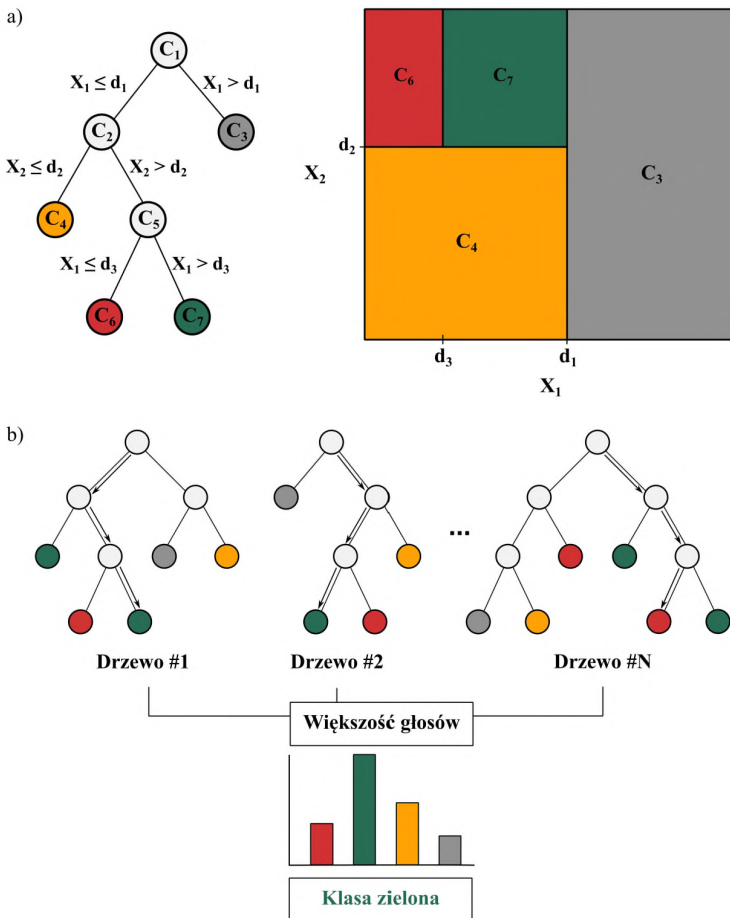
Chemometria oraz uczenie maszynowe zyskały w ostatnich latach ogromną popularność wśród spektroskopistów, ze względu na dużą ilość informacji zawartych w widmach IR, które mogą następnie zostać wykorzystane do trenowania modeli regresji wielowymiarowej lub klasyfikacyjnych. Stosowane metody są uzależnione od celu analizy, ale również od ilości dostępnych danych. Dla przykładu, posiadając kilka do kilkudziesięciu próbek, najstosowniej jest zastosowanie metod regresji wielowymiarowej. Szczególnie popularną jest metoda cząstkowych najmniejszych kwadratów (PLS, ang. *Partial Least Squares*), która poza regresją umożliwia wykorzystanie stworzonego modelu do klasyfikacji. Gdy dostępna liczba próbek przekracza kilkadziesiąt stosowane są często metody klasyfikacji, takie jak sieci neuronowe (NN, ang. *Neural Networks*), Las Losowy (RF, ang. *Random Forest*), czy metoda wektorów wspierających (SVM, ang. *Support Vector Machine*). Sieci neuronowe są potężnym narzędziem pozwalającym trenować skomplikowane modele, jednak aby charakteryzowały się one dobrą mocą predykcyjną, do treningu konieczny jest zbiór wielu (minimum tysiąca) próbek, co w badaniach *life science* jest często nieosiągalne. Dlatego ogromną popularność, szczególnie w klasyfikacji tkanek i tworzeniu modeli histopatologicznych, zyskały Las Losowy.

Las Losowy w swoim działaniu jest oparty o drzewa decyzyjne i polega na wykorzystaniu wielu słabych klasyfikatorów (drzew) w celu stworzenia modelu odpornego na przetrenowanie. Może być wykorzystywany do zagadnień związanych z regresją, ale tutaj uwaga skupia się na klasyfikacji. Do wytrenowania modelu RF wykorzystywany jest zbiór n niezależnych próbek $L = \{ (X_1, Y_1), \dots, (X_n, Y_n) \}$, gdzie każda obserwacja (X_i, Y_i) opisana jest przez wektor b cech (ang. *features*) $X_i = X_{i1}, \dots, X_{ib}$ oraz etykietę Y_i wybraną z zasobu klas. Konieczność zdefiniowania odpowiedzi związana jest z faktem, że RF jest metodą klasyfikacji z nadzorem [54], [55].

Pierwszym etapem trenowania Lasu Losowego jest agregacja *bootstrap'a*. Polega ona na losowym wyborze (ze zwracaniem) próbek z zestawu L tak, aby stworzyć podzestaw próbek służących do treningu jednego drzewa. Dla lasu losowego złożonego z N drzew, proces ten jest powtarzany N razy. Ponadto, do treningu każdego z drzew wykorzystywany jest losowy podzestaw cech, a ich ilość wynosi zazwyczaj $\sqrt{b}$. Losowy wybór próbek i cech tworzy indywidualny zestaw danych treningowych dla każdego z drzew. Sam proces treningu każdego drzewa rozpoczyna się od tak zwanego węzła korzenia (ang. *root node*), w którym zawarte są wszystkie próbki. Węzeł następnie ulega podziałowi bazując na wartościach danej cechy. Procedura ta jest powtarzana aż osiągnięte zostaną liście drzewa [54], [55]. Algorytm wyboru cech i ich wartości progowych, w oparciu o które dokonywany jest podział, bazują na współczynniku Gini lub entropii i są szeroko opisane w literaturze [56]. Na rysunku 16a przedstawiona jest budowa prostego drzewa decyzyjnego opartego o dwie cechy i klasyfikującego próbki do czterech klas (szara, żółta, czerwona i zielona). Węzeł C_1 to korzeń, gdzie zawarte są wszystkie próbki, natomiast węzły C_3 , C_4 , C_6 oraz C_7 to liście odpowiadające finalnemu przypisaniu do klasy przez dane drzewo. Tak wytrenowany zestaw N drzew tworzy finalny model Lasu Losowego oraz może zostać wykorzystany do klasyfikacji próbki o nieznannej etykietce. Odbywa się to przez zebranie głosów od wszystkich drzew z lasu (rysunek 16b) i przyjęcie, że prawdziwą jest klasa, która zebrała ich największą ilość.

Wykorzystanie przez RF zespołu indywidualnie wytrenowanych drzew dobrze przeciwdziała przetrenowaniu wywołanemu skupianiem się modelu na szczegółowych

cechach, jak to często bywa w przypadku prostych drzew decyzyjnych. Wynika to w znacznym stopniu z zastosowania algorytmu *bootstrap'a* i prowadzi do powstania silnych i dokładnych modeli.



Rysunek 16. a) Przykładowe drzewo decyzyjne wytrenowane w oparciu o dwie cechy. b) Schemat wyboru klasy dla próbki nieznaney w oparciu o model RF zbudowany z N drzew.

Przedruki publikacji

Wpływ odszumiania na dane spektroskopowe i skuteczność klasyfikacji

Pierwsze badania naukowe realizowane w ramach niniejszej pracy doktorskiej związane były z udziałem w grancie Homing pt. „Pełna histopatologia raka trzustki przy zastosowaniu obrazowania chemicznego”, kierowanym przez dr hab. Tomasza P. Wróbla. W ramach projektu metody uczenia maszynowego, takie jak Lasy Losowy, wykorzystywane były do stworzenia modelu histopatologicznego dla raka trzustki, w oparciu o obrazowanie w podczerwieni. Do realizacji powyższych zagadnień niezbędne było zebranie odpowiedniej ilości danych FT-IR, umożliwiających uzyskanie dużej statystyki pacjentów, na którą składało się kilkaset biopsji igłowych w formie mikromacierzy tkanek (TMAs, ang. *Tissue MicroArrays*) oraz dodatkowe materiały resekcyjne. Pierwotnie zastosowana została procedura pomiarowa zakładająca wielokrotne skanowanie w celu osiągnięcia satysfakcjonującego poziomu SNR. Pomiar jednego kafelka, uwzględniający 32 skany, trwał około 52 s, co przy dodatkowych ograniczeniach związanych z pracą detektora (około 3,5 h bez uzupełniania ciekłego azotu) skutkowało czasem trwania pomiarów jednej TMA rzędu 5,6 tygodnia dla obiektywu 15x, oraz 8,5 miesiąca dla 36x. W ramach projektu planowane były pomiary sześciu TMA, dlatego pierwotne oszacowania znacznie wykraczały poza ramy czasowe projektu. Bezpośrednią konsekwencją tej obserwacji było gruntowne zoptymalizowanie praktyk związanych ze zbieraniem danych, zaczynając od zastosowania metod odszumiania w celu poprawy jakości sygnału, a tym samym skrócenia czasu skanowania. Po wspomnianej redukcji czasu skanowania dominującym blokiem stała się Transformata Fouriera, która z tego powodu została przeniesiona na zewnętrzny komputer. Zastosowanie dodatkowo nowych skryptów pomiarowych (pokonujących ograniczenia pamięciowe oprogramowania) oraz regularna konserwacja detektora (odpompowanie dewara), umożliwiły pomiary 24 godziny na dobę z prędkością około 12,5 s na kafelek. Skróciło to czas pomiaru jednej TMA do 2 dni oraz 2 tygodni, odpowiednio dla obiektywów 15x i 36x, umożliwiając tym samym uzyskanie planowanej statystyki pacjentów w projekcie Homing. Jednorazowe pomiary dużych obszarów próbki (o wymiarach rzędu cm) wiązały się z zauważalną względną zmianą pozycji ogniska wiązki. Wynikało to z niedoskonałości uchwytów i skutkowało przechyleniem próbki, mającym negatywny wpływ na jakość otrzymywanych danych. W celu korekty, zaimplementowany został algorytm korygujący pozycję ogniska na bieżąco w trakcie trwania pomiarów.

Wspomniane powyżej optymalizacje metod odszumiania, kluczowe dla redukcji skali czasowej pomiarów, rozpoczęły serię tematycznych publikacji skupionych na poprawie jakości danych oraz skuteczności klasyfikacji. Charakterystyka wielowymiarowych danych, pochodzących z obrazowania FT-IR, umożliwia zastosowanie zarówno metod spektralnego jak i przestrzennego odszumiania. Aby liczbowo oszacować wpływ każdej z metod na jakość sygnału lub obrazu, wykorzystane zostały symulowane dane w postaci rdzenia (biopsji igłowej) tkanki trzustki [A]. Dane o różnej rozdzielczości przestrzennej zostały zasymulowane

specjalnie do tego zastosowania, uwzględniając surowy sygnał (niezaburzony szumem) oraz zestawy danych o różnym poziomie szumu, odpowiadające konkretnym czasom skanowania. Takie podejście umożliwiło nie tylko oszacowanie poprawy jakości sygnału (SNR), ale również oszacowanie zniekształcenia sygnału związanego z odszumianiem, poprzez zastosowanie stworzonego w tym celu algorytmu [A]. W publikacji [A] zawarte są wyniki skuteczności wszystkich wykorzystanych metod, z uwzględnieniem optymalizacji parametrów pod kątem poziomu początkowego szumu, rozdzielczości przestrzennej oraz skali odszumianych danych (liczb widm). Otrzymane wyniki ujawniły korzyści płynące z zastosowania odszumiania, zwłaszcza w przypadku takich metod jak MNF oraz PCA. Ich główne zalety to możliwość zmniejszenia czasu skanowania do kilku skanów przy zachowaniu poziomu SNR przekraczającego odpowiednik kilkuset skanów [A]. Pozytywna reakcja środowiska na powyższą publikację oraz chęć wsparcia podobnych badań poskutkowały opublikowaniem na zasadach otwartego dostępu danych symulowanych, wraz ze skryptami do ich generacji, w czasopiśmie Scientific Data [D].

Korzyści płynące z odszumiania zauważalne są nie tylko przez analizę parametrów takich jak SNR, ale również w wysokiej jakości wynikach analiz ilościowych i jakościowych. Dlatego, poza optymalizacją metod odszumiania i ich parametrów wyłącznie pod kątem osiągnięcia jak najlepszej jakości sygnału, przeprowadzono optymalizacje umożliwiające osiągnięcie wysokiej skuteczności klasyfikacji danych spektroskopowych za pomocą Lasów Losowych. Pierwotnie ta analiza ta skupiona była na modelach opartych o dane FT-IR [B], co miało szczególny wydźwięk w kontekście trwającego projektu Homing. Metody MNF oraz PCA poprawiły skuteczność klasyfikacji na poziomie piksela nawet o 43-44%.

W późniejszym czasie do projektu Homing włączone zostały dane pochodzące z mikroskopu QCL, zapewniające korzystniejszą skalę czasową pomiarów kosztem ograniczonej informacji spektralnej. Zaowocowało to rozszerzeniem prac związanych z optymalizacją odszumiania również na tą metodę obrazowania w podczerwieni [C]. Rezultaty przedstawione w tej publikacji wskazały metody oraz ich parametry, najskuteczniejsze w przypadku klasyfikacji opartej o kontrast chemiczny pochodzący od zredukowanej liczby pasm (obrazów) [C]. W tym przypadku jedynie MNF zapewniał znaczącą poprawę klasyfikacji, przy założeniu zastosowania przynajmniej 20 zmiennych spektralnych.

Mój wkład w publikacje:

- [A] związany był z eksperymentalną częścią zbierania danych FT-IR, symulacjami wolnego od szumu zestawu danych, udziałem w implementacji metod odszumiania, pisanem skryptów do analizy wpływu odszumiania na jakość danych oraz pisanem części manuskryptu;
- [B] związany był z eksperymentalną częścią zbierania danych FT-IR, udziałem w implementacji skryptów do odszumiania i analizy danych oraz pisanem części manuskryptu;
- [C] związany był z implementacją metod odszumiania, trenowaniem modeli klasyfikacyjnych, pisanem skryptów do analizy danych związanych z badaniem wpływu odszumiania na klasyfikację oraz pisanem części manuskryptu;
- [D] związany był z eksperymentalną częścią zbierania danych FT-IR, implementacją skryptów do symulacji danych, symulacją danych, przygotowaniem danych oraz powstałej biblioteki do publikacji oraz pisanem części manuskryptu.

SCIENTIFIC REPORTS

Corrected: Author Correction

OPEN

Comparison of spectral and spatial denoising techniques in the context of High Definition FT-IR imaging hyperspectral data

Paulina Koziół¹, Magda K. Raczowska^{1,2}, Justyna Skibinska^{1,3}, Sławka Urbaniak-Wasik⁴, Czesława Paluszkievicz¹, Wojciech Kwiatek¹ & Tomasz P. Wrobel¹

The recent emergence of High Definition (HD) FT-IR and Quantum Cascade Laser (QCL) Microscopes elevated the IR imaging field very close to clinical timescales. However, the speed of acquisition and data quality are still the critical factors in reaching the clinic. Denoising offers aid in both aspects if performed properly. However, there is a lack of a direct comparison of the efficiency of denoising techniques in IR imaging in general. To achieve such comparison within a rigorous framework and obtaining the critical information about signal loss, a simulated dataset strongly bound by experimental parameters was created. Using experimental structural and spectral information and experimental noise levels data as an input for the simulation, a direct comparison of spatial (Fourier transform, Mean Filter, Weighted Mean Filter, Gauss Filter, Median Filter, spatial Wavelets and Deep Neural Networks) and spectral (Savitzky-Golay, Fourier transform, Principal Component Analysis, Minimum Noise Fraction and spectral Wavelets) denoising schemes was enabled. All of these techniques were compared on the simulated dataset, taking into account SNR gain, signal distortion and sensitivity to tuning parameters as comparison metrics. Later, the best techniques were applied to experimental data for validation. The results presented here clearly show the benefit of using hyperspectral denoising schemes such as PCA and MNF which outperform other methods.

Signal processing is a very rich field with new denoising techniques created every year. These are becoming more and more precise and adaptive in nature, however, most of them are focused on photography, medical imaging and satellite imagery¹. Spectroscopic hyperspectral imaging in this context is relatively new and denoising techniques have not yet been fully optimized for the specific structure and characteristic of IR data^{2–6}.

IR imaging provides information about the total chemical composition of the sample along with specific information about the conformation of certain molecules, e.g. proteins, DNA. This chemical contrast can be used to enable more precise models of pathology classification for many diseases, e.g. cancers, atherosclerosis or diabetes^{7–12}. In recent years there has been a tremendous progress in instrumentation and image formation theory, leading to new modalities among which High Definition (HD) pushed the projected pixel size to diffraction limit sampling¹³. Obtaining the full spatial information requires using a 1.1 μm projected pixel size, which is smaller than more commonly utilized 5.5 μm . This leads to a 25-times smaller sampling volume and light throughput, resulting in a reduction of signal to noise ratio (SNR). This significantly increases the acquisition times in order to obtain reasonable SNR in HD FT-IR and calls for denoising techniques¹⁴.

The current literature is lacking in a direct comparison of the efficiency of denoising techniques in IR imaging in general – this is especially difficult since every approach risks distorting the original signal when removing noise. Since it is not possible experimentally to obtain a noise-free dataset it is impossible to measure actual signal distortion. To circumvent this problem and allow a rigorous framework for comparing denoising efficiency, including the critical information about signal loss, we created a simulated dataset, strongly bound by

¹Institute of Nuclear Physics Polish Academy of Sciences, PL-31342, Krakow, Poland. ²Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, Mickiewicza 30, Krakow, Poland. ³Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering, AGH University of Science and Technology, Mickiewicza 30, Krakow, Poland. ⁴NZO Pathology Department, Jagiellońska 70, Kielce, Poland. Correspondence and requests for materials should be addressed to T.P.W. (email: tomasz.wrobel@ifj.edu.pl)

experimental parameters. Using actual structural and spectral information and experimental noise levels data as an input for the simulation, we enabled a direct comparison of spatial and spectral denoising schemes.

As IR imaging grew from the spectroscopic community, the natural tendency was to use standard spectral denoising techniques such as Savitzky-Golay¹⁵ or Fourier-filtering¹⁶. More advanced approaches are based on Principal Component Analysis (PCA)¹⁷ and may take a form of Minimum Noise Fraction (MNF)^{18,19}. Up to date, spatial denoising in IR imaging has been rarely used with a few exceptions of Fourier filtering¹⁶. This stems from the fact that IR imaging is a relatively high SNR technique and unless used in HD-IR mode with a thermal source it provides data of reasonable quality. The emergence of HD-IR has opened the need for efficient denoising algorithms^{20–22}. The developments in Quantum Cascade Laser Microscopes²³ are enabling Discrete Frequency imaging and that puts even more interest in denoising techniques based on a limited number of frequencies, i.e. spatial denoising⁷. Median, Mean and Gauss filters⁷ are commonly used in data processing of imaging data, while wavelets have been suggested as very promising techniques for IR as well²⁴. Finally, Artificial Neural Networks can be used for denoising²⁵ and recently the first Deep Neural Network (DNN) denoising approach was enabled in Matlab packages and is applied here to IR data for the first time.

All of these techniques were compared on the simulated dataset taking into account SNR gain, signal distortion, sensitivity to tuning parameters and influence of specific FT-IR data structure with different pixel sizes of 1.1 and 5.5 μm representing HD and standard definition imaging options. The results presented here create a benchmark comparison and facilitate the choice of denoising methods for any FT-IR dataset, which reduces the optimization required to be done by the end user.

Materials and Methods

Experimental. A paraffin embedded serial pancreatic tissue microarray sections were obtained from US Biomax, Inc. 5 μm thick section of the TMA was placed on a BaF₂ salt plate for transmission IR imaging. The section was deparaffinized using a 24h hexane bath. FT-IR measurements were performed using a Bruker Vertex70v Spectrometer in transmission mode coupled to Hyperion 3000 microscope equipped with an MCT FPA 64 $\times$ 64 detector and 15x and 36x objectives, giving pixel projections of 2.7 microns and 1.1 microns, respectively. 5.5 microns image was obtained by 5x binning of the corresponding 36x image and served as an example of typical IR imaging projected pixel size. The images were acquired in the range from 3900 cm^{-1} to 900 cm^{-1} with the resolution of 4 cm^{-1} . A subset of the core, i.e. a single 64 $\times$ 64 pixels tile, was scanned with both 15x and 36x objectives with 2, 4, 8, 16, 32, 64, 128 and 256 scans. This data was used for the experimental noise level (8 levels) estimation to be used in the simulations corresponding to the experimental number of scans.

Simulations. A single pancreatic cancer tissue core served as an input for simulations. Fuzzy C-Means clustering was performed on the 1.1 μm image (4 scans, empty pixels were discarded) with 3 chosen classes. The spatial distribution of these classes were used for initial concentration profiles of three spectral profiles (resembling proteins, lipids and DNA/RNA profiles), each containing 16 bands (1030, 1080, 1172, 1242, 1280, 1342, 1400, 1462, 1550, 1650, 1754, 2352, 2850, 2920, 2962, 3300 cm^{-1}). Each band was simulated as a 50:50% mixture of Gaussian and Lorentzian peak shapes. The heights, band positions and widths were all randomized separately for each of the pixels using a normal distribution with 5%, 0, 2%–1, 8% and 1% variability, respectively, simulating experimental variability of components and concentrations. Afterward, a Gaussian filter was used for each of the variables in the spectral range, leading to simulating the dependence of IR resolution from the wavelength of light. The result of simulation was over 700 000 spectra arranged spatially in a tissue core structure, with an additional 600 000 empty space spectra. Raw signal was corrupted by randomized baseline and various levels of multiplicative noise (up to 40% for the most intense bands as reported⁴), to model different SNR levels of experimental data.

Performance of each denoising method in the group of spectral and spatial denoising techniques was evaluated based on two metrics representing data quality and distortion. For spectral denoising, data improvement was represented by Signal to Noise Ratio (SNR) with a signal value of unity and a standard deviation (after slope correction) of 2150–2080 cm^{-1} region (free of absorption bands) as a noise measure. Each SNR was divided by the SNR of a noisy signal to yield SNR Gain. Second parameter – Signal Distortion (SD), was defined as a sum of differences between denoised and clean (simulated) spectrum, but only for variables where the difference is higher than noise added to a clean signal and is presented as the total signal %. Examination of spatial denoising was carried out based on commonly used in image processing parameters, such as Peak Signal to Noise Ratio (pSNR) and Structural Similarity Index (SSIM). Similarly to SNR, pSNR takes the ratio of squared maximum value and Mean Square Error (MSE) between input and reference image. Since pSNR is usually defined on a decibel scale it was transformed into a linear scale and then divided by the corresponding transformed pSNR of the noisy signal to yield pSNR Gain. SSIM index is a multiplicative combination of the structure, contrast and luminance term²⁶, giving the level of similarity to the reference image.

Data processing and classification were done in OPUS and Matlab software using built-in and home written procedures.

Results and Discussion

Figure 1 presents the workflow leading to a much more precise comparison of a range of denoising techniques. Experimental data served as an input for defining the structure of the image as well as relative concentrations of major biochemical classes (proteins, lipids, carbohydrates/DNA/RNA) and their spectral profiles. This allowed simulating a noise-free dataset with variability levels comparable or higher than that of a typical tissue sample. A noise-free signal was then corrupted by the addition of spectral multiplicative noise at levels corresponding to experimental values of 2 to 256 scans (8 different levels). This dataset was then subjected to spatial and spectral denoising algorithms, each optimized separately for best SNR Gain vs SD (Supplementary Materials S3–S4) and

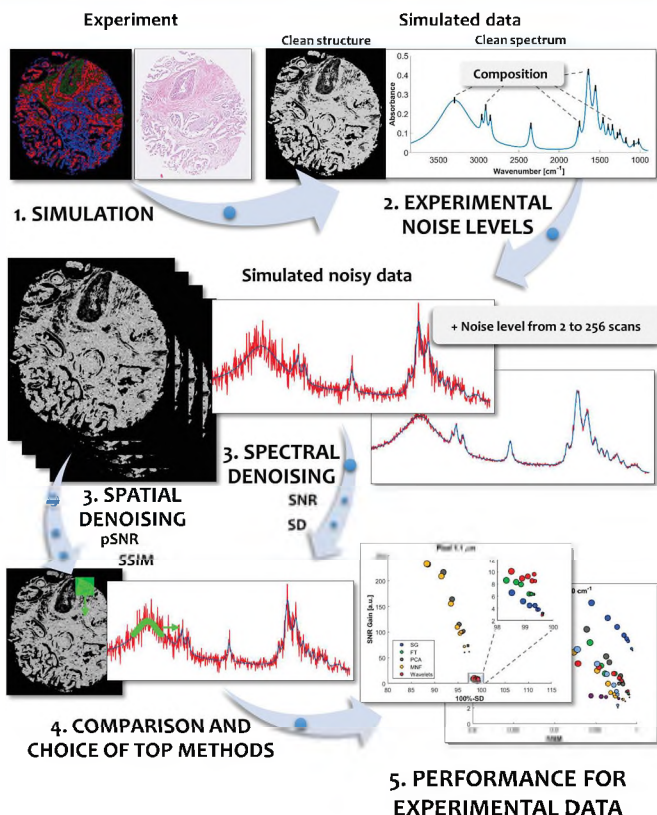


Figure 1. A schematic representation of the workflow designed to compare denoising techniques for IR imaging. An experimental IR image is used to define the structure and composition profiles of the dataset. Afterward, multiplicative noise is introduced to the system with levels corresponding to experimental 2 to 256 scans (8 levels). In the next step, the noisy data is subjected to different denoising approaches using spectral and spatial methods which are optimized for different noise levels. Finally, optimized methods are compared using Signal to Noise Ratio (SNR) and Signal Distortion (SD) metrics for spectral data and corresponding peak Signal to Noise Ratio (pSNR) and Structural Similarity Index (SSIM). The best methods are then applied again to experimental data to show their actual capability.

pSNR Gain vs SSIM (Supplementary Materials S5–S8). Finally, the best methods were again applied to experimental data to validate their effectiveness.

Spectral denoising. The results of SNR gain and SD for all of the spectral methods tested are shown in Fig. 2 and single pixel spectra are shown in Fig. 3. The best SNR Gain is usually obtained for the noisiest data for a given technique. The most commonly used in spectroscopy denoising algorithm is the Savitzky–Golay smoothing filter (blue) is giving the smallest SD values in the range of 0.5–1.5% at a reasonable increase of SNR of up to 6 times in the HD resolution (with optimized parameters of polynomial degree equal to 3 and the smoothing points equal to 15 and 17). The higher the initial noise level the higher the gain, but more signal was lost. The increase

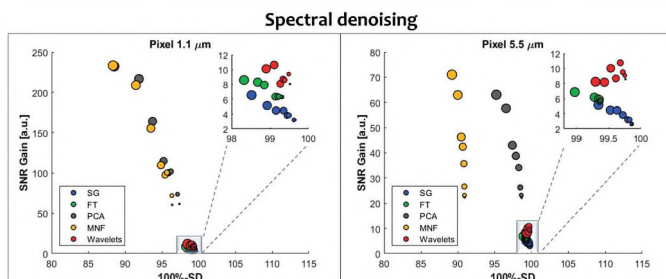


Figure 2. A comparison of spectral denoising techniques as measured by SNR gain (normalized to SNR of the input noisy data) and SD (presented as 100%-SD) for two different pixel sizes representing HD and standard definition imaging magnifications. Five methods were tested: Savitzky-Golay (SG), Fourier transform (FT), Principal Component Analysis (PCA), Minimum Noise Fraction (MNF) and spectral wavelets (Wavelets). The size of the dots corresponds to the initial noise level, starting with the highest noise for the largest dot (2 scans) and going down to the smallest (256 scans). The optimal parameters for each of the methods and noise levels are given in Supplementary Materials.

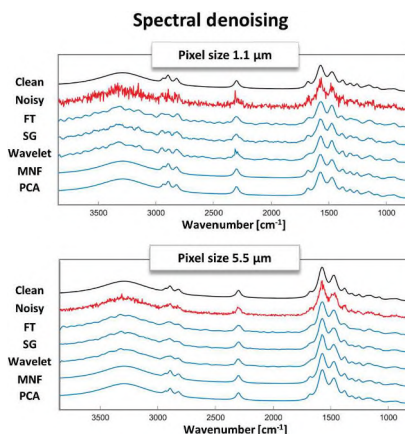


Figure 3. Comparison of the spectral result of denoising for a single pixel at noise levels of 2 scans for HD and standard definition modes – noisy spectrum in red and denoised spectrum in blue.

could be higher if the user would be satisfied with higher SD which are related in a nice, almost linear fashion (Supplementary Materials S3–S4). The next method, the Fourier transform, is showing an increased SNR gain behavior at a slightly higher cost of SD. On the downside, it is more sensitive to optimization parameters having a well-defined maximum for a given noise level (Supplementary Materials S3–S4). Wavelet denoising is also based on Fourier filter and different shapes (families) can be used with different levels of hardness, leading to 450 combinations that were tested (Supplementary Materials Tables 1 and 2). Since the best wavelets significantly flattened the spectra, Pearson correlation coefficient was used to support the choice of the best set. The latter performed better than Savitzky-Golay and Fourier transform in terms of SNR gain having values of 8–10 with SD at a comparable level, however, significant optimization was required and the choice of the best ones is somewhat arbitrary. The methods discussed up to now work on a single spectrum level, while the next two methods take advantage of the multivariate nature of the data using the whole data-cube with thousands of spectra at once. PCA denoising performs a matrix decomposition of the whole dataset at once and MNF has an additional step of noise whitening

Adjacent pixels analysis - 1650 cm^{-1}

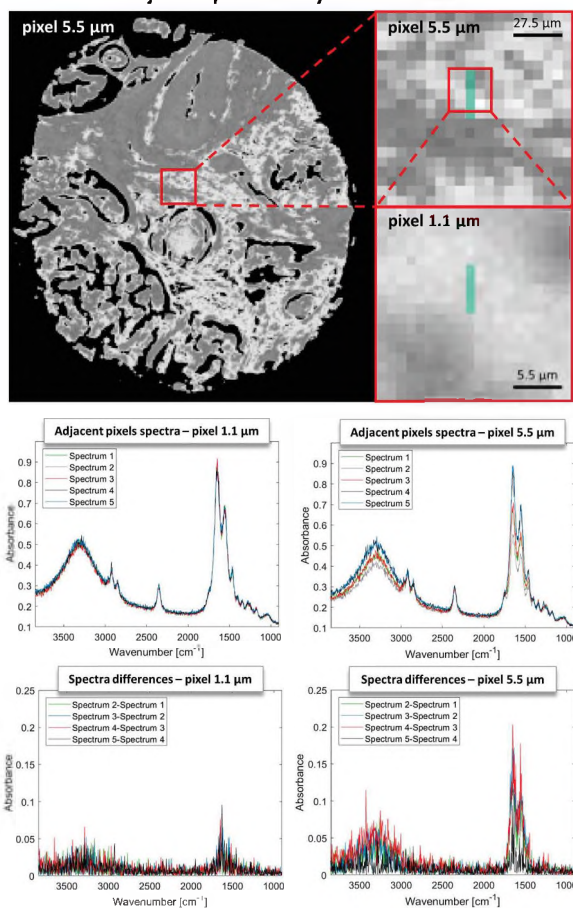


Figure 4. Example of calculation of the shift difference stat in MNE. The spectra of 5 adjacent pixels and their differences in 1.1 μm and 5.5 μm image of the same area are shown.

before the same matrix decomposition as PCA. This additional step provides an estimate of the noise structure and helps to order the new directions in space in decreasing SNR. For HD data both PCA and MNF outclass every other method with the SNR gain in the range of 60 to 240 times, however, this comes at a cost of larger SD values in the 3–12% range. Moreover, the choice of optimal parameters is not obvious, since the more PCA or MNF bands are used for reconstruction the less signal is lost but the SNR Gain drops (Supplementary Materials S3–S4).

The situation for standard definition imaging at 5.5 μm is quite similar, however, the starting noise level is already smaller than in HD and the SNR gains are also smaller. Savitzky-Golay, Fourier transform and Wavelets have slightly smaller SNR gains than in HD, but compensating in smaller SD values. PCA performs in a similar

Spatial denoising

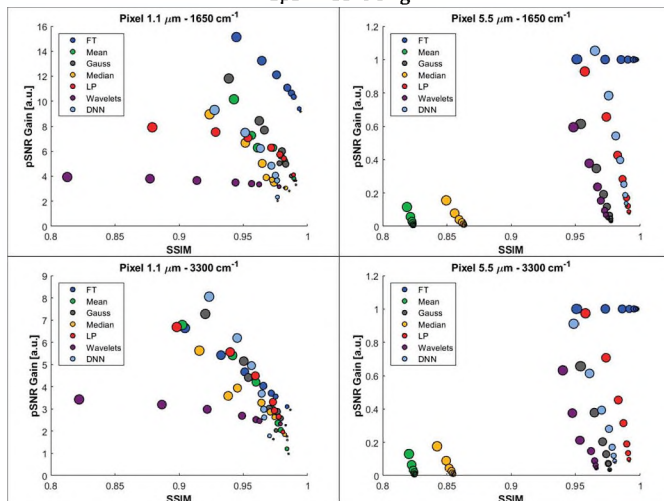


Figure 5. A comparison of spatial denoising techniques as measured by pSNR Gain (normalized to pSNR of the input data) and SSIM for two different pixel sizes representing HD and standard definition imaging magnifications and two spectral bands of 3300 cm^{-1} and 1650 cm^{-1} , corresponding to 3 and $6\text{ }\mu\text{m}$ wavelengths. Seven methods were tested: Fourier transform (FT), Mean Filter (Mean), Gauss Filter (Gauss), Median Filter (Median), Weighted Mean (W. Mean), spatial Wavelets (Wavelets) and Deep Neural Networks (DNN). The size of the dots corresponds to the initial noise level, starting with the highest noise for the largest dot (2 scans) and going down to the smallest (256 scans). The optimal parameters for each of the methods and noise levels are given in Supplementary Materials.

fashion as in HD, however, MNF has a significantly higher SD, which is almost constant at around 10%, regardless of the noise level. This is a bit surprising but can be understood by analyzing the way the noise estimate is obtained. A shift difference stat is most commonly used, which calculates the difference between adjacent pixels, assuming that the only difference is noise. Results of such calculation and this effect are shown in Fig. 4.

The differences between adjacent pixels show a presence of the signal in both projected pixel size images. However, the assumption about only noise being present is met in HD data to a greater extent than in standard definition. The origin of this is the different spatial sampling of the actual structure and the fact, that HD is point-spread function limited and standard definition is not. This phenomenon is moreover wavelength dependent since a wavelength of $6\text{ }\mu\text{m}$ will be better sampled with $5.5\text{ }\mu\text{m}$ pixels than a wavelength of $3\text{ }\mu\text{m}$. This leads to higher signal losses at the high-wavenumber end of the spectra than in the fingerprint and can be seen in Fig. 3, where exemplary single pixel spectra are shown before and after denoising. Interestingly, this is not lowered with the inclusion of more MNF bands as happens in PCA.

Spatial denoising. Spatial denoising was performed with 7 different approaches and the results are summarized in Fig. 5. In general, the wavelength dependence of the image structure and contrast coming from spatial resolution required to focus on two examples – 3300 cm^{-1} and 1650 cm^{-1} bands, corresponding to 3 and $6\text{ }\mu\text{m}$ and having a different contrast. For HD imaging case the lowest pSNR gains were obtained for wavelets and surprisingly also the highest signal loss (low SSIM values) for both bands. This means that wavelets are not properly estimating the structure of human tissue sections as in photography this approach with modifications is very efficient. For the 1650 cm^{-1} band Fourier transform outperformed all other methods. This comes from the fact that image at this wavelength is highly oversampled and the corresponding structure frequency is significantly different from the noise frequency. The pSNR gains range from 9 to 16 with a very small signal loss as measured by SSIM. The 3300 cm^{-1} band results show that most methods perform similarly, however, for the highest noise levels Deep Neural Networks obtained the highest pSNR gain. The standard definition imaging counterpart shows interesting results, i.e. that most of the methods decrease pSNR values – the exception being Fourier transform. Deeper investigation of this surprising result shows that again the structure sampled as $5.5\text{ }\mu\text{m}$ vary highly

Spatial denoising

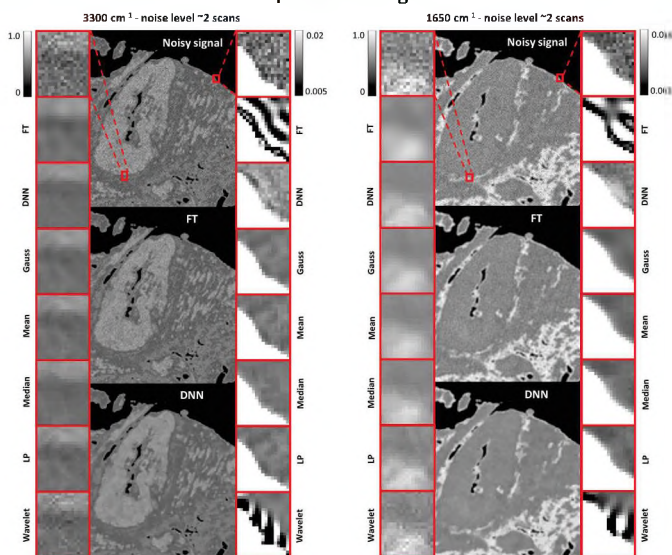


Figure 6. Details of spatial denoising results for HD data with Fourier transform (FT) and Deep Neural Networks (DNN) shown in details as they performed relatively the best (all images are given in Supplementary Materials S10). Zoom-ins on tissue structure and edges are shown to better highlight noise rejection and potential artifacts introduced by a given method.

from pixel to pixel causing mixing of the structure signal with the noise component. This leads to signal removal and is not recommended for standard definition data. Profiles of the structure before and after denoising showing this effect are presented in Supplementary Materials (Figure S11).

pSNR and SSIM are very popular metrics for estimating the efficiency of denoising methods in the signal processing community, they are not, however, without drawbacks. The HD images before and after denoising for all of the methods are shown in Fig. 6. It can be seen that even though Fourier transform had the best values for pSNR and SSIM it leads to artifact creation at the edges – an expected feature of this denoising method. Wavelets as methods based on Fourier transform share a similar artifact, however, to a smaller degree. Human eye suggests that the Deep Neural Network actually performs better than Fourier transform and Mean, Median, W. Mean and Gauss filters are quite comparable.

Interestingly, when pSNR and SSIM were calculated for the same variables after spectral PCA and MNF denoising, the results are in the order of 33 and 113 of pSNR gain and 0.95–0.99 for SSIM for PCA and 37 to 30 pSNR gain and 0.96–0.97 for SSIM for MNF. This indicates that these spectral methods perform better spatially than any other spatial filter. To be exact, MNF does use spatial information about the noise structure, therefore it is not strictly spectral.

Experimental data denoising. Every simulation is burdened by an assumption which may or may not be met in real life and therefore, it is always important to verify theoretical findings with experimental data. A healthy pancreatic tissue core was used for this purpose and the results of the best denoising methods are shown in Figs 7 and 8. It can be seen that a popular Median Filter is quite good at reducing noise, but significantly distorts the structure especially at the edges. FT Spatial Filter on the hand preserves more structural detail but leads to blurring and an impression of defocusing. Deep Neural Network provides a milder noise removal at the gain of preservation of the original structure. The balance between noise removal and structure preservation is apparent here. The multivariate methods of PCA and MNF do seem to perform the best, with PCA actually highlighting some edges more than can be seen in the original image. This comes from structure propagation from other spectral bands that influence the result as well. MNF also performs very well, however, edges do get smeared out due to the way the noise estimate is calculated as was explained in the simulation section.

Experimental data - 3300 cm^{-1}

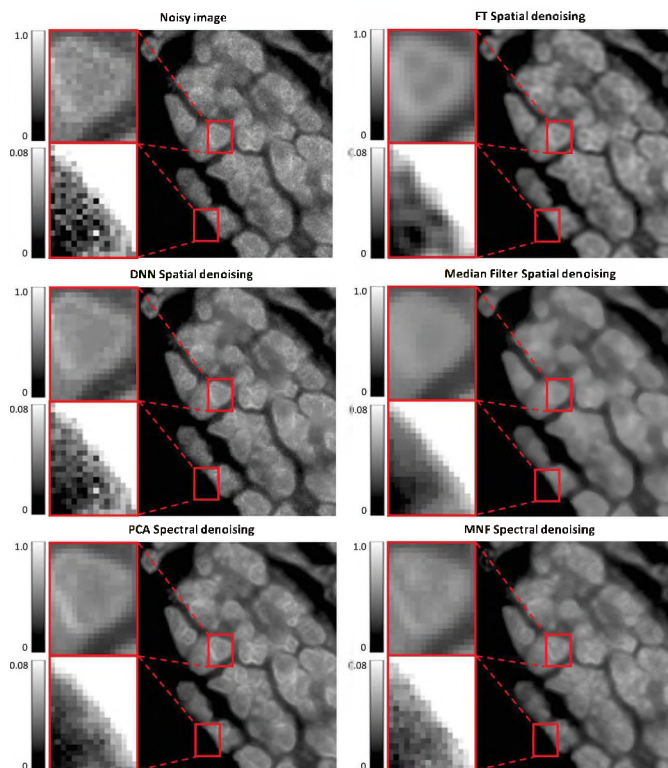


Figure 7. Results of spatial and spectral denoising on the image quality of a healthy pancreatic tissue sample. Median, FT, DNN, PCA and MNF denoising methods were applied to an experimental image acquired with 4 scans.

Image quality is important, however, the most useful information coming from IR imaging is the spectral information about the chemical content. It is this information that is used for understanding changes occurring in the sample and is the basis for almost all disease classification approaches based on IR imaging. Both PCA and MNF outperformed every other technique in both spectral and spatial denoising efficiency and a closer comparison on experimental data was done. The same pancreatic tissue core, which fragment is shown in Fig. 6 was reconstructed using a different number of bands (MNF bands and PCA loadings) and the resulting SNR Gain was calculated and is presented in Fig. 8. The overall SNR Gain is in the 30–60 range, unlike the simulation, where it reached over 200. This can be explained by a non-perfect simulation model and confirms the need to test experimental data. Still, the SNR gain is very high and drops with the incorporation of more bands, which is expected since consecutive bands will have an increasing noise component. It can be seen that MNF is more efficient at ordering the bands according to SNR, while PCA uses variance as a criterion for ordering. Up to band 10, MNF reconstructions have a better SNR gain and outperform PCA. Interestingly, the SNR Gain drops much faster if too many bands are taken for reconstruction, while PCA remains relatively constant. Therefore, if properly optimized MNF will yield a better SNR but for an unknown signal PCA is safer. Overall, both methods are yielding very good results as can be seen by the single spectra. The differences between PCA and MNF reconstructed spectra

PCA and MNF comparison

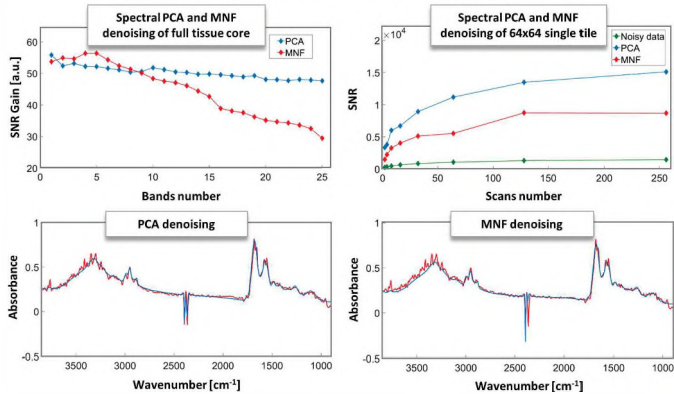


Figure 8. (Upper left) Comparison of SNR Gains for PCA and MNF denoising with a different number of bands taken for reconstructions for an experimental whole tissue core shown in Fig. 7. (Bottom row) Single spectra taken from the same experimental pixel from Fig. 7. (Upper right) Actual SNR values of an experimental 2 to 256 scans of a single 64 × 64 tile of a tissue compared with PCA and MNF denoised values of the same image, showing the clear gain from denoising regardless of the actual number of scans.

Spectral denoising technique	PROS	CONS
Savitzky-Golay	Very low signal distortion, short computational time	SNR gain lower than one order of magnitude, two parameters optimization
Fourier-Transform	Very low signal distortion, easy to implement, short computational time, easy to optimize	SNR gain lower than one order of magnitude
PCA	Significant SNR gain and reasonable signal distortion	Medium difficulty algorithm, time and memory consuming computations
MNF	Significant SNR gain and reasonable signal distortion	Difficult algorithm, hard to implement, time and memory consuming computations
Wavelets	Very low signal distortion	SNR gain around one order of magnitude, time consuming calculation and optimization
Spatial denoising technique	PROS	CONS
Fourier-Transform	Good pSNR gain, high SSIM, easy to implement, reasonable computation time, easy to optimize	Image artifacts
Mean Filter	Good SSIM, easy to implement, low computational time	Mild pSNR gain
Median Filter	Good SSIM, easy to implement, low computational time	Mild pSNR gain
Gauss Filter	Good SSIM, easy to implement, low computational time	Mild pSNR gain
Weighted Mean Filter	Good SSIM, easy to implement, low computational time	Mild pSNR gain
Wavelets	Noticeable pSNR gain	Low pSNR gain and SSIM, time consuming calculation and optimization, image artifacts
Deep Neural Networks	Reasonable pSNR and SSIM	Difficult algorithm to train

Table 1. A summary of the tested methods with a description of their performance.

can be seen in the high wavenumber region, where MNF has an offset with regards to the experimental signal. Also, the CO₂ bands at 2300–2400 cm⁻¹ are reconstructed in a slightly different way, showing that some of the signal loss is irrelevant since these bands are not of use in interpretation. Finally, when looking at SNR plot for the different number of scans it is clear that no matter how many scans are taken experimentally, denoising will improve it beyond the experimentally reasonable limits. This measurement was done on a smaller area, therefore, the SNR Gain is smaller than for a full tissue core, where much more pixels are available for noise estimation and results in lower MNF performance than that of PCA. A summary of the pros and cons of the different techniques is given in Table 1.

Conclusions

We investigated the denoising efficiency and signal distortion properties of a series of spectral and spatial noise removal techniques, using HD FT-IR data as an input. The results clearly show that multivariate based techniques of PCA and MNP outperform any other spatial or spectral method, taking full advantage of the broadband nature of the data. For single pixel level denoising, Wavelet denoising performed the best, while for a single frequency spatial denoising the answer was less clear, but Deep Neural Networks performed very well across different inputs and suggest that this method has good potential if trained properly. Overall, we show that any kind of denoising done on even a small number of scans for imaging will give a better outcome than extended acquisition times.

References

1. Motwani, M. C., Gadiya, M. C. & Motwani, R. C. Survey of image denoising techniques. *Proc. Glob. Signal Process. Expo Conf.* (2004).
2. Wrobel, T. P. & Bhargava, R. Infrared Spectroscopic Imaging Advances as an Analytical Technology for Biomedical Sciences. *Anal. Chem.* **90**, 1444–1463 (2018).
3. Wrobel, T. P., Kole, M. R. & Bhargava, R. Emerging trends and opportunities in discrete frequency infrared and Raman spectroscopic imaging. *Spectroscopy* **31**, 6, 28–45 (2016).
4. Mark, H. L. & Griffiths, P. R. Analysis of Noise in Fourier Transform Infrared Spectra. *Appl. Spectrosc.* **56**, 633–639 (2002).
5. Kwak, J. T., Reddy, R., Sinha, S. & Bhargava, R. Analysis of variance in spectroscopic imaging data from human tissues. *Anal. Chem.* **84**, 1063–9 (2012).
6. Piqueras, S. et al. Handling Different Spatial Resolutions in Image Fusion by Multivariate Curve Resolution-Alternating Least Squares for Incomplete Image Multisets. *Anal. Chem.* <https://doi.org/10.1021/acs.analchem.8b00630> (2018).
7. Wrobel, T. P. et al. In *Optical Spectroscopy and Computational Methods in Biology and Medicine*, M. Baranska (Ed.) (ed. Baranska, M. 1505–535. https://doi.org/10.1007/978-94-007-7832-0_17 (Springer Netherlands, 2014).
8. Wrobel, T. P., Majzner, K. & Baranska, M. Protein profile in vascular wall of atherosclerotic mice analyzed *ex vivo* using FT-IR spectroscopy. *Spectrochim. Acta Part A: Mol. Biomol. Spectrosc.* **96**, 940–945 (2012).
9. Wrobel, T. P. et al. Effects of a Low Carbohydrate High Protein (LCHP) diet on atherosclerotic plaque phenotype in ApoE^{-/-}/LDLR^{-/-} mice: FT-IR and Raman imaging. *Sci. Rep.* **1–9**, <https://doi.org/10.1038/srep14002> (2015).
10. Leslie, L. S. et al. High Definition Infrared Spectroscopic Imaging for Lymph Node Histopathology. *PLoS One* **10**, e0127238 (2015).
11. Majzner, K., Wrobel, T. P., Fedorowicz, A., Chlopicki, S. & Baranska, M. Secondary structure of proteins analyzed *ex vivo* in vascular wall in diabetic animals using FT-IR spectroscopy. *Analyst* **138**, 7400–7410 (2013).
12. Perez-Guaita, D. et al. High Resolution FTIR Imaging Provides Automated Discrimination and Detection of Single Malaria Parasite Infected Erythrocytes on Glass. *Faraday Discuss.* 341–352, <https://doi.org/10.1039/C5FD00181A> (2015).
13. Reddy, R. K., Walsh, M. J., Schumacher, M. V., Carney, P. S. & Bhargava, R. High-definition infrared spectroscopic imaging. *Appl. Spectrosc.* **67**, 93–105 (2013).
14. Reddy, R. K. & Bhargava, R. Accurate histopathology from low signal-to-noise ratio spectroscopic imaging data. *Analyst* **135**, 2818–25 (2010).
15. Zimmermann, B. & Kohler, A. Optimizing Savitzky-Golay parameters for improving spectral resolution and quantification in infrared spectroscopy. *Appl. Spectrosc.* **67**, 892–902 (2013).
16. Lash, P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemom. Intell. Lab. Syst.* **117**, 100–114 (2012).
17. Chen, G. & Qian, S. E. Denoising and dimensionality reduction of hyperspectral imagery using wavelet packets, neighbour shrinking and principal component analysis. *Int. J. Remote Sens.* **30**, 4889–4895 (2009).
18. Green, A., Berman, M., Switzer, P. & Craig, M. D. A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Trans. Geosci. Remote Sens.* **26**, 65–74 (1988).
19. Berisha, S. et al. SIProc: an open-source biomedical data processing platform for large hyperspectral images. *Analyst* **142**, 1350–1357 (2017).
20. Davis, B. J., Carney, P. S. & Bhargava, R. Theory of Midinfrared Absorption Microspectroscopy: I. Homogeneous Samples. *Anal. Chem.* **82**, 3474–3486 (2010).
21. Davis, B. J., Carney, P. S. & Bhargava, R. Theory of Mid-infrared Absorption Microspectroscopy: II. Heterogeneous Samples. *Anal. Chem.* **82**, 3487–3499 (2010).
22. Wiercigroch, E. et al. FT-IR Spectroscopic Imaging of Endothelial Cells Response to Tumor Necrosis Factor- α : To Follow Markers of Inflammation Using Standard and High-Magnification Resolution. *Anal. Chem.* **90**, 3727–3736 (2018).
23. Wrobel, T. P., Mukherjee, P. & Bhargava, R. Rapid visualization of macromolecular orientation by discrete frequency mid-infrared spectroscopic imaging. *Analyst* **142**, 75–79 (2017).
24. Alsborg, B. K., Woodward, A. M., Winsom, M. K., Rowland, J. & Kell, D. B. Wavelet denoising of infrared spectra. *Analyst* **122**, 645–652 (1997).
25. Fichou, D. & Morlock, G. E. Powerful Artificial Neural Network for Planar Chromatographic Image Evaluation, Shown for Denoising and Feature Extraction. *Anal. Chem.* <https://doi.org/10.1021/acs.analchem.8b01296> (2018).
26. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).

Acknowledgements

This research was supported from the “Pancreatic cancer comprehensive histopathology based on IR chemical imaging” project, which is carried out within the Homing programme (grant no. Homing/2016-2/20) of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund. This research was performed using equipment purchased in the frame of the project co-funded by the Malopolska Regional Operational Program Measure 5.1 Krakow Metropolitan Area as an important hub of the European Research Area for 2007–2013, project no. MRPO.05.01.00-12-013/15.

Author Contributions

Study design: T.P.W., P.K., W.M.K., C.P., S.U.W. Development of methodology: T.P.W., P.K., M.K.R., J.S. Acquisition of spectroscopic data: T.P.W., P.K., M.K.R., J.S. Analysis and interpretation of data: T.P.W., P.K., W.M.K., C.P., M.K.R., J.S. Writing the manuscript: P.K., T.P.W.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-018-32713-7>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018

SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Noise-free simulation of an FT-IR imaging hyperspectral dataset of pancreatic biopsy core bound by experiment

Tomasz P. Wrobel^{1*}, Paulina Koziol¹, Magda K. Raczowska^{1,2}, Danuta Liberda¹, Czesława Paluszkievicz² & Wojciech M. Kwiątek¹

A noise-free hyperspectral FT-IR imaging dataset of a pancreatic tissue core was simulated based on experimental data that allows to test the performance of various data analysis and processing algorithms. A set of experimental noise levels was also added and used for denoising approaches comparison, which due to the noise-free reference signal enables to truly observe signal distortion caused by different approaches.

Background & Summary

IR Imaging and spectroscopy is a fast developing field with new applications and branches of research emerging every year^{1–3}. Some of the applications are well within the current capabilities of the technique, while others are still pushing the limit. High speed, High Definition imaging would be desirable for a clinical translation^{4,5}, however, obtaining useful Signal to Noise Ratio (SNR) usually requires averaging multiple acquisitions. Pre-processing^{6,7} and denoising in particular plays a major role here. However, few approaches give full insight into their influence on signal, since any noise rejection may result in distorted data. Moreover, every experimental signal will have a noise component even after long acquisitions. In order to have a true insight into signal distortion we decided to create simulated datasets with raw and noisy data^{8,9}.

To ensure the highest reliability of the simulations, experimental data was taken to provide spectral and spatial information used as the initial starting point. A tissue biopsy core was measured and cluster analysis was done to determine three main classes distribution. Afterwards, these spatial concentration profiles were assigned spectral data profiles of proteins, nucleic acids and lipids, which were based on spectroscopic knowledge and heavily randomized to ensure high variability as expected in a biological material. Still, only three profiles were used and a larger number of classes may be used to capture more of real biological material heterogeneity. Finally, the summed spectral profiles of varying spatial concentrations formed an initial dataset, which was later corrupted with randomized baseline. However, this dataset did not take into account the dependence of spatial resolution from the wavelength of light (the spectral frequency), since the concentration profiles obtained from cluster analysis removed it during clustering. In the next step such dependence was introduced and a full noise-free dataset was obtained. Afterwards, varying levels of noise, derived from experimental measurements, corresponding to increasing number of acquisitions were introduced to form different SNR datasets for denoising.

The motivation behind this research was to investigate available denoising approaches, not only by estimating improvement done by noise reduction but also by having insight into the degree of distortion done to a signal. This kind of analysis was never done before, especially on that scale. We recently published results, where among others we presented optimization process of five spectral denoising techniques⁵. By making this data available to other researchers we aim to encourage them to perform similar optimizations, which is possible not only for denoising but might be applied to diversity of preprocessing techniques. These datasets are a perfect platform to test and objectively compare state-of-the-art and new algorithms of any kind of data treatment with denoising being the intended use for this simulations.

¹Institute of Nuclear Physics Polish Academy of Sciences, PL-31342, Krakow, Poland. ²Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, Mickiewicza 30, Krakow, Poland. *email: tomasz.wrobel@ifj.edu.pl

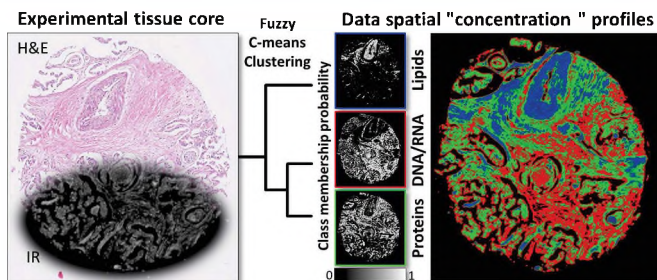


Fig. 1 Graphical illustration of the process leading to data spatial structure characterization. FT-IR data (left) served as a base for Fuzzy C-means Clustering, which led to obtaining three spatial profiles (right) including proteins, nucleic acids and lipids. Black and white images with color frames present the class membership probability of each profile separately, and RGB image shows a combination of all.

Methods

A typical FT-IR imaging dataset consists of a data cube, which spectral dimension is usually approximately 800 or 1600 variables depending on the spectral range and resolution, 8 or 4 cm^{-1} , respectively. The spatial x and y dimension vary a lot – depending on the samples and magnification optics, dictating the projected pixel size – still a range from hundreds to dozens of thousands reported in the literature.

In our datasets, the complete process of obtaining a simulated FT-IR dataset consists of two main parts: experiment and simulation. The purpose of the experiment was to obtain an actual estimation of the data structure, which will form the basis of the simulation. The briefly outlined here process lead to creation of simulated FT-IR datasets directly based on experimental parameters.

1. Pancreatic biopsy tissue core was measured to create a full image, and thus obtain the spatial structure of the dataset.
2. Small subset (a 64×64 pixel area) was measured with varying number of acquisitions giving information about spectral noise levels.
3. Initial spectra were simulated with Voigt profiles and randomized.
4. Baseline was added to achieve more realistic (disturbed by artifacts) data.
5. Wavelength dependence of the spatial resolution was introduced using Gaussian filter imitating an Airy's disc.
6. Various levels of noise were added creating range of simulated data with different SNR.

More detailed description is presented below.

Experimental. The experimental part was based on a Tissue Micro Array (TMA) with pancreatic cancer tissue cores, purchased from US Biomax Inc., which also holds informed consents from human subjects for research. A paraffin embedded, 5 μm thick section was mounted on BaF₂ salt window for transmission measurements mode. 24h long hexane bath was performed for sample deparaffinization. The equipment used for FT-IR imaging was Bruker Vertex70v Spectrometer coupled to Hyperion 3000 microscope with Focal Plane Array (FPA) detector 64×64 pixels and 36x objective (projected pixel size of 1.1 μm).

Data spatial structure. A single pancreatic tissue core with a diameter of approximately 1 mm was measured in transmission mode, giving around 700 000 sample pixels and 600 000 background pixels. Data was acquired in the range from 3850 cm^{-1} to 900 cm^{-1} with spectral resolution of 8 cm^{-1} and 4 scans averaging. The normally used in the Fourier transform zero filling factor of 1 results in interpolated data and the final spectral spacing is equal to 4 cm^{-1} . Data were written in a binary file with single precision giving a file size of 3.8 GB. Unsupervised Machine Learning in the form of simple Cluster Analysis is known to show good results in FT-IR Imaging^{10,11}. Therefore, in order to prepare data for further analysis, Principal Component Analysis (PCA) denoising with 50 components was applied. After empty pixels exclusion, Fuzzy C-means Clustering (FCM) with D-Values as a distance measure (modified Pearson correlation coefficient)¹⁰ was performed using algorithm implemented in CytoSpec software. The number of classes was arbitrarily chosen as three, since the goal was to obtain spatial profiles of three main biochemical classes. The general scheme and outcome of this process are shown in Fig. 1. Two more spatially similar classes were assigned to protein and DNA/RNA classes, respectively, while the more distinct one to lipids, as is often the case in biological tissues. It later served as a base for spatial arrangement (concentration scaling factors) during simulation process. Additionally, the image for Amide I band (1650 cm^{-1}) of experimental data was used for tissue mask creation. The applied threshold was equal to 0.15. Pixels with absorbance above this limit were marked with 0 (tissue region) and 1 if opposite (background region).

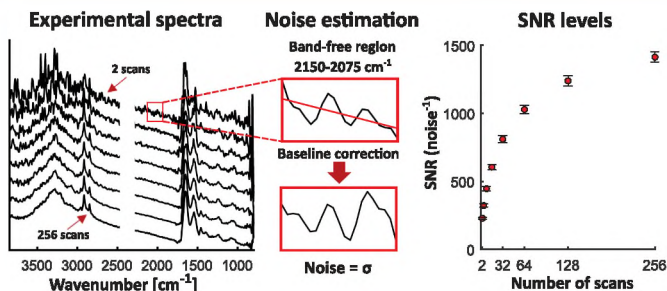


Fig. 2 Procedure for experimental noise estimation. Examples of experimental spectra with different noise level corresponding to 2–256 scans averaging (left). (middle) Presentation of band free region, with a fitted straight line, used for noise estimation. After baseline correction, standard deviation gives numerical noise value (middle). SNR presented as an average of results for sample region pixels (spectra), for datasets corresponding to the number of scans presented on X axis (right).

Noise levels estimation. Six areas of 64×64 pixels (corresponding to a single tile), partly covering pancreatic tissue and partly background area were measured in transmission mode. In this case, the spectral range was from 3850 cm^{-1} to 800 cm^{-1} , with 8 cm^{-1} resolution. These measurements were repeated 8 times, for 2, 4, 8, 16, 32, 64, 128 and 256 scans averaging (and 64 for background measurement) and giving experimental spectra with different signal quality as presented in the left part of Fig. 2.

In order to numerically estimate noise levels corresponding to each dataset, a noise describing metric was implemented. Like it is shown in the middle part of Fig. 2, noise estimation consists of two steps. First, a straight line is fitted to the band free region of the spectrum ($2150\text{--}2075 \text{ cm}^{-1}$). Afterwards, a line is subtracted and the absolute value is taken:

$$x_i = |y_i - b_i|$$

where: y_i is a raw data point and b_i is a corresponding data point of the fitted line. Secondly, the results from the previous step served for the standard deviation calculations giving noise values for a given spectrum.

$$\text{noise} = \sqrt{\frac{\sum_{i=1}^K (x_i - \bar{x})^2}{N - 1}}$$

where: $\bar{x}$ is a mean value of x_i , i and K are the indexes of $2150\text{--}2075 \text{ cm}^{-1}$ data range and N is a number of points. A MATLAB implementation of the presented above algorithm is provided with the code set. The obtained results of Signal to Noise Ratio (SNR) with a signal of unity (signal value taken as 1) are presented in Fig. 2. For better statistics and results stability the SNRs for pixels in the sample region were averaged for each dataset. We observed here a good SNR shape, which was expected to be proportional to the square root of number of scans. This result was later used for the estimation of the amount of noise added to simulated spectra.

Simulations. The core part of this data descriptor is the process of simulation of a reliable dataset being a good imitation of FT-IR imaging data. The whole process consists of four main steps described below with a general workflow presented in Fig. 3. This part of analysis was entirely done in MATLAB Software.

Spectral profiles generation. In this part of the simulation, each spectrum of the hyperspectral dataset was simulated individually and later arranged in a spatial manner via spatial “concentration” profiles, which are linear combination coefficients of the three profiles. An important remark to make is the definition of single band shape. In general, IR bands are with a good approximation similar to Gaussian or Lorentzian-like curves¹². However, more specific sources claim that closer estimation is given by a Voigt profile¹³, which is defined as a convolution of Gauss function and Lorentz function^{14,15}. Here the Voigt profile implementation made by Sanjar M. Abrarov and Brendan M. Quine from York University in Canada was used. Individually simulated profiles contain 16 bands with defined position and width of Voigt functions (50% Gaussian and 50% Lorentz functions). These profiles are later multiplied by corresponding scaling factors obtained from cluster analysis of experimental data and summed giving the shape of final spectrum. In other words, it is a sum of three components corresponding to data profiles determined in the experimental section – proteins, nucleic acid and lipids. Spectral characteristics of the base profiles used in simulation are presented in Table 1. They were defined by an experienced spectroscopist to give typical shapes and ratios of bands of the main profiles – typical profiles are shown in Fig. 4. During the simulation process, all parameters were randomized with various levels of standard deviation (Table 1). Spectra were simulated in the range from around 3850 cm^{-1} to 900 cm^{-1} with 4 cm^{-1} resolution, although in the provided

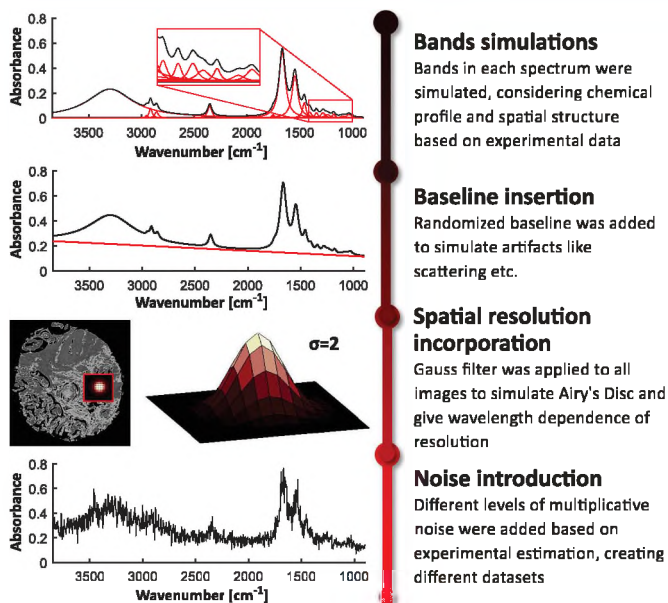


Fig. 3 Scheme describing the workflow of the simulation process. First part included creation of spectra with determined chemical profile using Voigt profiles as bands approximation. Next step was aimed to insert randomized baseline in order to introduce artifacts to the system. Afterwards, spatial resolution dependence was introduced by using a Gauss filter. Finally, different levels of multiplicative noise were added to clean data.

Band position [cm ⁻¹]	Randomization of position - in term of SDT	Band width [cm ⁻¹]	Randomization of width - in term of SDT	Voigt profile area - DNA/RNA	Voigt profile area - proteins	Voigt profile area - lipids
3300	1.8%	308	1%	4,000	3,000	1,800
2962	0.7%	15	1%	0,010	0,010	0,030
2920	0.6%	19	1%	0,030	0,060	0,080
2850	0.6%	23	1%	0,030	0,040	0,070
2352	0.4%	27	—	0,100	0,100	0,100
1754	0.3%	19	1%	0,020	0,010	0,070
1650	0.5%	42	1%	0,400	1,000	0,300
1550	0.4%	39	1%	0,300	0,600	0,200
1462	0.3%	23	1%	0,040	0,120	0,080
1400	0.2%	23	1%	0,100	0,037	0,060
1342	0.2%	19	1%	0,045	0,032	0,050
1280	0.2%	27	1%	0,060	0,050	0,010
1242	0.2%	39	1%	0,100	0,040	0,020
1172	0.2%	19	1%	0,012	0,020	0,040
1080	0.2%	39	1%	0,100	0,020	0,015
1030	0.2%	35	1%	0,120	0,035	0,020
Randomization with SDT of 5%						

Table 1. Information/data used in spectral profiles simulations.

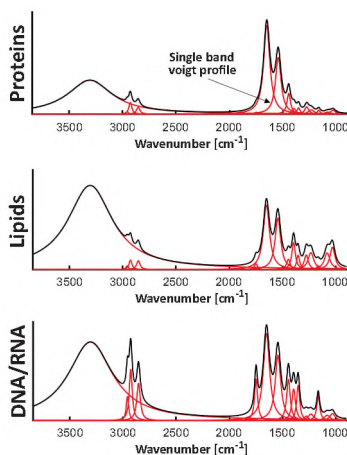


Fig. 4 Simulated spectral profiles of the three main components: proteins, lipids and DNA/RNA.

scripts bands positions and widths were converted from wavenumbers into points. After generation and randomization of the three profiles for each of the pixels, the three profiles were summed according to spatial concentration profiles. An exception was made for CO_2 band (2352 cm^{-1}), since it is a random component, not related in concentration to any spatial structure. Therefore, its concentration profile was completely random spatially, while not randomized in terms of width. In order to reach more realistic tissue section absorbance values whole the dataset was multiplied by the factor of 10. Some additional minor steps were also applied, but it can be concluded from provided scripts.

Baseline insertion. Experimental data are almost never artifact free and this fact must be included in the simulation. A group of artifacts distorting spectra with an appearance of a varying baseline is related to scattering effects and strongly depend on sample geometry and morphology. Scattering is approximately proportional to the wavenumber¹⁶ (light frequency) apart from situation where measured object is spherical with diameter close to the wavelength (for example cells)^{17–19}. This is not the case for tissue measurements, therefore, a baseline in the form of a straight line was introduced here. Both line coefficients were randomized, and once again based on the spectroscopic knowledge, slope range was set from 0.15 to 0.25 and y-intercept was in the range from -0.00026 to 0. Baseline added to background spectra was divided by the factor of 15. This comes from the fact that the lack of sample greatly reduces scattering effects. An exemplary baseline is presented in the proper section of Fig. 3.

Spatial resolution dependence on wavelength. Since the simulated hyperspectral data should be representative of IR imaging technique, it was essential to take into account image formation physics related to the dependence of spatial resolution from the wavelength of light at the diffraction limit. Light spot coming from circular aperture is described by Airy's disc⁶, with radius limited by the first minimum and described by:

$$r = 1.22 \cdot \frac{\lambda \cdot R}{a}$$

where λ is wavelength, a is aperture diameter and $R = \frac{a}{2 \cdot \text{NA}}$ is a distance of focus from aperture described by numerical aperture NA. This leads to a more commonly known expression of.

$$r = \frac{\lambda}{2 \cdot \text{NA}}$$

With a typical NA of FT-IR imaging in the 0.5 range this leads to almost direct proportionality of the wavelength of light and the radius of an Airy's disk. A reasonable approximation of the disk can be done with a Normal Distribution, therefore a Gauss spatial filter was applied to all images with individually computed filter parameter:

$$\sigma = 34.1\% \cdot r.$$

Estimation of σ was done, by assuming that r is covered by three standard deviations of Gaussian distribution (3σ). This means that for a wavelength of $6\mu\text{m}$ (amide I band at 1650 cm^{-1}) the total distribution should cover $12\mu\text{m}$, which means it is covering almost 11 pixels in x and 11 in y (projected pixel size of $1.1\mu\text{m}$ was used to obtain the spatial structure). In this step the data gained another important source of variability coming from the dependence of images (structures) on wavelength. Gauss filtering was the final step in the creation of raw FT-IR dataset simulating clean (noise free) signal.

Noise insertion. The main purpose of this simulation was to obtain datasets with different levels of noise and strongly bound by experimental data. Based on knowledge about the multiplicative character of noise in FT-IR spectroscopy²⁰, noise level for maximal absorbance in the dataset was set to be 40% higher than for the minimal absorbance value. To execute this condition a Gaussian noise was simulated at first. Noise values were later linearly scaled with coefficients found from a system of two linear equations obtained for maximal and minimal value of absorbance:

$$a = \text{noise}_G \cdot \frac{\text{MIN} - 1.4 \cdot \text{MAX}}{\text{MIN} - \text{MAX}},$$

$$b = \text{noise}_G \cdot a \cdot \text{MAX},$$

where a is slope for noise scaling, b is y -intercept, noise_G is Gaussian noise and MIN and MAX are minimal and maximal values in whole raw dataset. Final noise added to the system was:

$$\text{noise}_{\text{fit}} = a \cdot A_{ijk} + b,$$

where A_{ijk} is absorbance for a given data point of hyperspectral data. The parameter that allowed to manipulate the final level of noise was the standard deviation of initially generated Gaussian noise. To reach the desired noise level (corresponding to experimental dataset) an iterative process was conducted. The standard deviation of Gaussian distribution used in noise generation was set to the value of noise for the experimental dataset. After introduction of generated noise into raw data, noise estimation was performed with described earlier SNR metric. If the obtained value differed from the experimental one, process was repeated with proper correction in standard deviation of Gauss noise. Due to the size of raw dataset (around 4 GB) this process might be time consuming and it is recommended to conduct it on a computer with sufficient RAM memory. The approach described here allows simulation of any noise level. 8 datasets with signal quality corresponding to 2, 4, 8, 16, 32, 64, 128 and 256 scans averaging were created.

Data Records

Simulated datasets and above-mentioned scripts and algorithms are available at the Open Science Framework⁹.

Main data including raw and noisy datasets were stored in Matlab.mat format. Each file is around 3.2 GB and after loading into workspace has three dimensions. First and second corresponds to a spatial position (pixels), third dimension refers to spectral variables. For example, extracted image of any spectral variable gives matrix with size of 1180×1100 , but extracted single spectrum gives vector of size $1 \times 1 \times 765$ and needs reshape (or squeeze) function to be plotted.

Besides final datasets we also provided results of experiment, which are directly needed for the simulations. Results of cluster analysis giving scaling factors for each data profile are available in .mat format. This file contains four variables, three variables with size of 1180×1100 corresponding nucleic acids (firs class), proteins (second class) and lipids (third class). Last variable is a combination of all three with size of $1180 \times 1100 \times 3$ and the same order (might be used for RGB image). Results of noise estimating experiment are provided in excel file with proper description inside.

Data mask determining sample and background regions is in the form of .mat file with logical data type where 0 corresponds to sample pixel and 1 to background pixel.

Wavelength vector (.mat file) with dimensions of 765×1 is provided for the potential user in case if determining spectral regions is needed.

Scripts and functions have Matlab .m format, but might be open with any text editor.

Technical Validation

Principal component analysis. The main potential limitation of trying to simulate spectra of biological origin is to have too little variability, as there are millions of different molecules contributing to the acquired signal. PCA is a convenient way of comparing variance sources^{21,22} in two datasets and was therefore performed on the original experimental dataset and a simulated one with noise levels of the same order (the same number of scans). The scatter plot of PC1 vs. PC2 is shown in Fig. 5.

It can be seen that there is a large overlap between the two datasets, especially on PC1, which accounts for 80% of variance. There is some shift in PC2 as was expected, since simulation will not capture all of the variability. The values of the simulated dataset were scaled by 1.6 to have a total max absorption of 1.6 at amide I position as was found in the experimental dataset.

Usage Notes

A huge advantage of the provided here data is a presence of clean (noise free) dataset. Since preprocessing applied to more realistic noisy data might affect them in an unexpected way, therefore it is useful to compare introduced changes with the initial signal. This was exactly the case in the study investigating the influence of different denoising techniques on FT-IR data⁸. To numerically estimate level of deformation done by denoising, a Signal

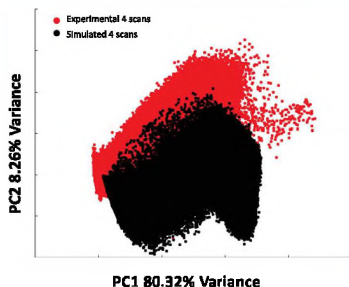


Fig. 5 PC1 (80.32% variance) vs. PC2 (8.26% variance) scatter plot of the simulated dataset with 4 scans and of the experimental dataset with 4 scans.

Distortion (SD) metrics was defined and it consists of the following steps. Initial step was to find an absolute value of noise introduced to the system, as:

$$noise_{ijk, abs} = |A_{ijk, noisy} - A_{ijk, clean}|$$

where $A_{ijk, noisy}$ is absorbance for data point of noisy dataset and $A_{ijk, clean}$ is a corresponding data point from raw dataset. Next, an absolute value of noise left after denoising was estimated:

$$noise_{ijk, abs, red} = |A_{ijk, denoised} - A_{ijk, noisy}|$$

where $A_{ijk, denoised}$ is absorbance for data point of denoised dataset. Final part was to sum values of:

$$SD_{ijk} = |noise_{ijk, abs, red} - noise_{ijk, abs}|$$

but only for ijk data points where $noise_{ijk, abs, red} > noise_{ijk, abs}$. This additional condition was introduced to avoid situation where the distortions caused by denoisings were smaller than the original noise. This custom made metrics was created for the purpose of denoising but it might be easily adapted for other applications and we recommend to use whenever reference signal is needed. MATLAB script for SD computation is also provided.

Code availability

Simulated datasets and mentioned above scripts and algorithms are available at the Open Science Framework⁹.

MATLAB Software version R2017b was used for the creation of scripts with additional toolboxes: Statistics and Machine Learning and Image Processing.

CystoSpect Software version 2.00.01 was used for PCA denoising and FCM clustering.

Voigt profile computation algorithm implemented by Sanjar M. Abrarov and Brendan M. Quine is available at <https://www.mathworks.cn/matlabcentral/fileexchange/47801-the-voigt-complex-error-function-second-version>.

Received: 8 July 2019; Accepted: 27 September 2019;

Published online: 29 October 2019

References

1. Wróbel, T. P. & Bhargava, R. Infrared Spectroscopic Imaging Advances as an Analytical Technology for Biomedical Sciences. *Anal. Chem.* **90**, 1444–1463 (2018).
2. Hermes, M. *et al.* Mid-IR hyperspectral imaging for label-free histopathology and cytology. *J. Opt. (United Kingdom)* **20**, 23002 (2018).
3. Wróbel, T. P., Kole, M. R. & Bhargava, R. Emerging trends and opportunities in discrete frequency infrared and Raman spectroscopic imaging. *Spectroscopy* **31**, 28–44 (2016).
4. Wróbel, T. P., Kwak, J. T., Kadajsky-Balla, A. & Bhargava, R. High-definition Fourier transform infrared spectroscopic imaging of prostate tissue. In *Progress in Biomedical Optics and Imaging - Proceedings of SPIE* **9791**, 97911D (2016).
5. Leslie, L. S. *et al.* High Definition Infrared Spectroscopic Imaging for Lymph Node Histopathology. *PLoS One* **10**, e0127238 (2015).
6. Lasch, P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemom. Intell. Lab. Syst.* **117**, 100–114 (2012).
7. Berisha, S. *et al.* SIProc: an open-source biomedical data processing platform for large hyperspectral images. *Analyst* **142**, 1350–1357 (2017).
8. Kozioł, P. *et al.* Comparison of spectral and spatial denoising techniques in the context of High Definition FT-IR imaging hyperspectral data. *Sd. Rep.* **8**, 1–11 (2018).
9. Wróbel, T. P. *et al.* IR Denoising Data Descriptor. *Open Science Framework*. <https://doi.org/10.17605/OSF.IO/WVF7P> (2019).

10. Wrobel, T. P., Mateuszuk, L., Kostogrys, R. B., Chlopicki, S. & Baranska, M. Quantification of plaque area and characterization of plaque biochemical composition with atherosclerosis progression in ApoE/LDLR(−/−) mice by FT-IR imaging. *Analyst* **138**, 6645–52 (2013).
11. Nallala, J., Lloyd, G. R., Shepherd, N. & Stone, N. High-resolution FTIR imaging of colon tissues for elucidation of individual cellular and histopathological features. *Analyst* **141**, 630–639 (2016).
12. Stuart, B. H. *Infrared Spectroscopy: Fundamentals and Applications*, <https://doi.org/10.1002/0470011149> (John Wiley & Sons, Ltd, 2004).
13. Stancik, A. L. & Brauns, E. B. A simple asymmetric lineshape for fitting infrared absorption spectra. *Vib. Spectrosc.* **47**, 66–69 (2008).
14. Abrarov, S. M. & Quine, B. M. Efficient algorithmic implementation of the Voigt/complex error function based on exponential series approximation. *Appl. Math. Comput.* **218**, 1894–1902 (2011).
15. Abrarov, S. M. & Quine, B. M. On the Fourier expansion method for highly accurate computation of the Voigt/complex error function in a rapid algorithm. *arXiv* **1205**, 1768 (2012).
16. Hecht, E. *Optics (4th Edition)*. (Addison-Wesley, 2012).
17. Bassan, P. et al. Resonant Mie scattering in infrared spectroscopy of biological materials – understanding the ‘dispersion artefact’. *Analyst* **134**, 1386–1393 (2009).
18. Bassan, P. et al. Resonant Mie scattering (RMieS) correction of infrared spectra from highly scattering biological samples. *Analyst* **135**, 268–77 (2010).
19. Bassan, P. et al. Reflection contributions to the dispersion artefact in FTIR spectra of single biological cells. *Analyst* **134**, 1171–5 (2009).
20. Mark, H. L. & Griffiths, P. R. Analysis of Noise in Fourier Transform Infrared Spectra. *Appl. Spectrosc.* **56**, 633–639 (2002).
21. Rygula, A., Jekiel, K., Szostak-Kot, J., Wrobel, T. P. & Baranska, M. Application of FT-Raman spectroscopy for *in situ* detection of microorganisms on the surface of textiles. *J. Environ. Monit.* **13**, 2983–2987 (2011).
22. Bonnier, F. & Byrne, H. J. Understanding the molecular information contained in principal component analysis of vibrational spectra of biological systems. *Analyst* **137**, 322–32 (2012).

Acknowledgements

This research was supported from the “Pancreatic cancer comprehensive histopathology based on IR chemical imaging” project, which is carried out within the Homing programme (grant no. Homing/2016-2/20) of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro P6000 GPU used for this research. This research was performed using equipment purchased in the frame of the project co-funded by the Malopolska Regional Operational Program Measure 5.1 Krakow Metropolitan Area as an important hub of the European Research Area for 2007–2013, project no. MRPO.05.01.00-12-013/15.

Author contributions

Study design: T.P.W., P.K., W.M.K. C.P., M.K.R., D.L. Development of methodology: T.P.W., P.K., M.K.R., D.L. Acquisition of spectroscopic data: T.P.W., P.K., M.K.R. D.L. Analysis and interpretation of data: T.P.W., P.K., W.M.K. C.P., M.K.R., D.L. Writing the manuscript: T.P.W., P.K.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to T.P.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2019



Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca



Influence of denoising on classification results in the context of hyperspectral data: High Definition FT-IR imaging



Magda K. Raczowska^{a, b}, Paulina Koziol^a, Slawka Urbaniak-Wasik^c,
Czesława Paluszkiwicz^a, Wojciech M. Kwiatek^a, Tomasz P. Wrobel^{a, *}

^a Institute of Nuclear Physics Polish Academy of Sciences, PL-31342, Krakow, Poland

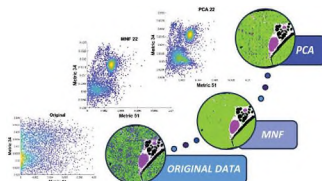
^b Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, Mickiewicza 30, Krakow, Poland

^c NZOZ Pathology Department, Jagiellonska 70, Kielce, Poland

HIGHLIGHTS

- Denoising process is demonstrated to significantly impact classification performance of FT-IR imaging data.
- Increase in pixel-level classification accuracy reach up to 45%.
- PCA and MNF have significantly improved classification results and outperform other methods.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:

Received 16 April 2019

Received in revised form

16 July 2019

Accepted 22 July 2019

Available online 25 July 2019

Keywords:

FT-IR imaging

PCA

MNF

Deep neural network

Wavelets

Random forest

ABSTRACT

Owing to the high information content about the biochemical composition of the sample, the implementation of Fourier-Transform Infrared Spectroscopy (FT-IR) in the clinic is currently under investigation by many researchers. Cancer biology with the use of histopathological models is one of the most explored application areas. Most of the publications show sensitivity of the method to be above 90%, however, it is still often not enough for clinical standards. Robust denoising techniques with an optimized classification model allow to shorten the experimental acquisition times which still are a bottleneck for FT-IR translation into the clinic. The main premise of this work is to evaluate denoising impact on classification results using spectral techniques: Savitzky Golay (SG), Wavelets (WV), Principal Component Analysis (PCA) and Minimum Noise Fraction (MNF); and spatial techniques: Deep Neural Network (DNN), Median Filter. Using denoising methods, especially MNF and PCA, gave significant improvement of the classification and prediction results. Moreover, the increase in pixel level accuracy for High Definition data (1.1 μm projected pixel size) was found to be dependent on the complexity of the histopathological class and reached even 43–44% level, while core level increase reached around 28%. Moreover, we investigated the impact of denoising methods on the spectral input to better understand the mechanism of such large improvement. The results presented here highlight the benefits and the importance of proper denoising for classification purposes of FT-IR imaging data.

© 2019 Elsevier B.V. All rights reserved.

* Corresponding author.

E-mail address: tomasz.wrobel@fj.edu.pl (T.P. Wrobel).

1. Introduction

Fourier transform infrared (FT-IR) imaging is rapidly developing as a field both in instrumentation [1–3] and in applications [4–6]. High information content about biochemical composition makes FT-IR imaging perfectly suitable for biological applications [7]. Part of the applications is concerned with exploring new systems with chemical imaging [8,9] while biomedical applications often use *a priori* knowledge about the system resulting in classification approaches [10–15]. The multivariate characteristic of IR spectroscopic data is fundamental for the use of machine learning. Moreover, the progress in instrumentation and resulting speed enhancement enables to use millions of objects for training and even furthermore justifies the use of machine learning, which requires a large model data set. The range of classification approaches is large, e.g. PLS-DA, KNN, SVM, Neural Networks, with Random Forest recently gaining attention due to its simple optimization, high accuracy, robustness and variable importance measure. The variety of available tools caused that the idea of translation FT-IR into the clinic, especially for the tissue diagnostic purposes, is gaining serious attention through the last years [16–19]. One of the most explored fields is cancer biology, with series of histopathological models created for, e.g. lung [20], prostate [21], breast [22] or lymph node [10] cancers. Most of the publications show sensitivity of the method above 90%, although it is still too slow to reach clinical timescales. Sufficient data throughput might be achieved by introducing QCL microscopes, allowing discrete frequencies imaging, in particular after widefield imaging systems became available [23,24]. First reports of QCL based histopathology are very promising [12,25]. Still, FT-IR systems offer the highest quality data and optimization of classifier performance is therefore crucial for the future adaptation of FT-IR based models for QCLs ones.

Several factors influence the classification performance of a given model, *i.e.* model training parameters, number of samples and variables, complexity of the approach (linear or non-linear) and data transformation techniques. The latter is usually termed pre-processing and includes all sorts of corrections taking into account the experimental effects and artifacts. In IR imaging the corrections are typically baseline corrections, normalization, metric calculation or denoising. Recently, we have done a thorough comparison of spatial and spectral denoising techniques in the context of FT-IR imaging [26] as well as in Quantum Cascade Laser IR imaging [12] data, looking into data-specific structures that influence the results. By looking at noise reduction from just signal processing angle (Signal to Noise Ratio Gain and Signal Distortion), PCA and MNF clearly outperformed all other approaches. In QCL IR imaging data such measures of noise removal were not available due to Discrete Frequency [27] nature of the data and a classification model was used to determine which methods perform the best. It was clear that only MNF was having any positive effect on the classification accuracy.

Here we extended our approach to FT-IR data, which has a different data-structure than QCL IR imaging and is expected to behave differently. Since the vast majority of research is still done with FT-IR instruments it is crucial to fully take advantage of the obtained signal. Therefore, we have chosen specifically a histopathological model of low-accuracy to observe which of the top denoising schemes will have the highest effect on pixel-level classification and on the prediction of new biological specimen. Moreover, an insight into the actual effect of denoising methods on the classification input space was done.

2. Materials and methods

2.1. Sample preparation and measurement

FT-IR measurements were performed on a paraffin-embedded serial pancreatic tissue microarray (TMA) purchased from Biomax. The TMA was sectioned into 5 μm slices and placed on BaF₂ salt plate for transmission IR imaging, deparaffinized with 24 h hexane bath. Bruker Vertex70v Spectrometer in transmission mode combined with Hyperion 3000 microscope equipped with an MCT FPA 64 $\times$ 64 detector and 36 $\times$ objective, with a pixel projection of 1.1 μm was used for data acquisition in the range from 3850 cm^{-1} to 900 cm^{-1} with the resolution of 8 cm^{-1} . 4 and 64 scans were averaged for sample and background measurements, respectively. Based on the High Definition (HD) data a more common IR imaging pixel size (5.5 μm) was obtained by binning the HD dataset.

2.2. Classification and denoising techniques

In the first step of analysis raw spectra (further referred to as original) and images were denoised with the following spectral techniques: Savitzky Golay (SG), Wavelets (WV), Principal Component Analysis (PCA) and Minimum Noise Fraction (MNF); and spatial techniques: Deep Neural Network (DNN), Median Filter. Those methods were applied with different parameters, but only results for parameters with the best performance are presented. Further classification was performed on spectral metrics, defined as bands integration, maximum value and centroid (m1), normalized to amide I band integration, maximum value and centroid, respectively.

The SG method is based on least-squares polynomial approximation, which is fitted to a moving window of an odd number of points, with the middle point replaced by the polynomial approximated one. In this work polynomial degree 3 and frame length of 11 was used [28]. Fourier transform (FT) in the spectral dimension is based on fitting a number of sine and cosine functions with varying frequency and offsets to the whole spectrum. After the fit, the coefficients of highest frequency functions are removed and rest is used for reconstruction of the signal [29]. The Wavelet transformation (WV) is similar to the FT method however, the main difference lies in the simultaneous localization of time and frequency domains of wavelet basis functions. WV has the local operation nature, where the wavelet function moves along the tested spectra, which is different from the global sine and cosine functions which are used in the Fourier transform. The WV can be interpreted as a combination of bandpass filters because allows the filtering of both low and high frequencies. The wavelet decomposition consists of distributing the signal into a series of functions that each of the decomposition levels contains high-frequency detail and approximation resulting from the subtraction of high-frequency details (low frequency part) [30]. Bior wavelet family with 4 decomposition levels was applied in this work. PCA and MNF methods are commonly used for hyperspectral denoising. PCA consists of the principal components constructed iteratively to maximize the description of data variance. Each subsequent component describes the variance unexplained by the previous one, which means that the first component has the largest possible variance [26]. Typically noise is expected not to occupy first PCs, therefore, only the first ones are used for signal reconstruction. Whereas MNF is basically PCA transform performed on noise-whitened data [31]. The median filter technique works based on the median from neighboring pixels

calculation. A window width determines the number of pixels. In this work noise removal results using filter width (kernel) of 3 and 5 were presented [32]. The second method of spatial denoising was possible thanks to the deep neural network implemented in Matlab software.

A subset of the TMA containing 12 cores was adapted, since classification on HD images is a computationally costly process. The set consists of 6 cores representing normal pancreatic acini and 6 inflammatory tissue cores. Histopathological regions (marked by an experienced histopathologist) of interest were marked in the form of a corresponding four classes tissue masks using ImageJ software. The selected classes were: normal pancreatic acini, inflammation tissue, fiber tissue and a group of other structures, e.g. nerves, pancreatic islet, vessels, secretory duct. Data processing, as well as classification, were done in MATLAB software using built-in and home written routines. The Random Forest (RF) classifier has been chosen due to high accuracy and robustness in biomedical imaging. A classifier was built with 50 trees. Two stages can be distinguished in the classification process: training and dataset validation. A pixel level validation was performed by a random 50:50 split of pixels from each of the classes from all cores. A core-level validation was done by Leave one out cross-validation, therefore, dataset was divided for training set of 11 cores and one core for validation. In a classification process, a model based on 11 cores was created and applied to the validation core. This process was repeated 12 times and obtained validation results were averaged. Cross-validation validates the stability of machine learning as well as gives information about the predictive capability of the model. From the 60 metrics of original data, the two major were selected to create figures showing the pixel distribution of classes within each of the denoising methods.

3. Results and discussion

3.1. Pixel level optimization

The first step in testing the denoising impact on FT-IR based classification was to check if there are any differences in the classification results after applying various noise removal methods. Comparison of classification accuracies on a pixel level for the selected approaches is shown in Fig. 1.

The first things to notice are the relatively low pixel-level classification accuracies for original data (44%–73%), which were expected, as a relatively difficult histopathological model has been attempted. Such a choice was required to allow the noise removal effect observation, since many of IR based models are in the above 95% accuracy range and improvements are marginal. Analyzing Fig. 1 it can be noticed that histological class strongly influences the results, not only initial classification accuracy level but also which techniques perform the best. The highest sensitivity was obtained for fiber tissue whereas the smallest for others group class, which is expected since fiber tissue has a characteristic spectral profile, while others class contain multiple cell types. Denoising methods implementation gave better classification outcomes than the original data in almost all cases, only the SG method is an exception in inflammation tissue and other group class. The results for DNN and Median Filter with kernel size 3 performed in a similar way, giving a positive contribution to true positive ratio improvement but the effect is minimal. Surprisingly, Median Filter with kernel size 5 gave high classification results of 92.5% for normal pancreatic acini and 93.9% for fiber tissue. The best increase in true positive ratio was obtained for PCA and MNF techniques, which parameters were selected based on the optimization curves, showing the dependence of the classification outcome on the number of com-

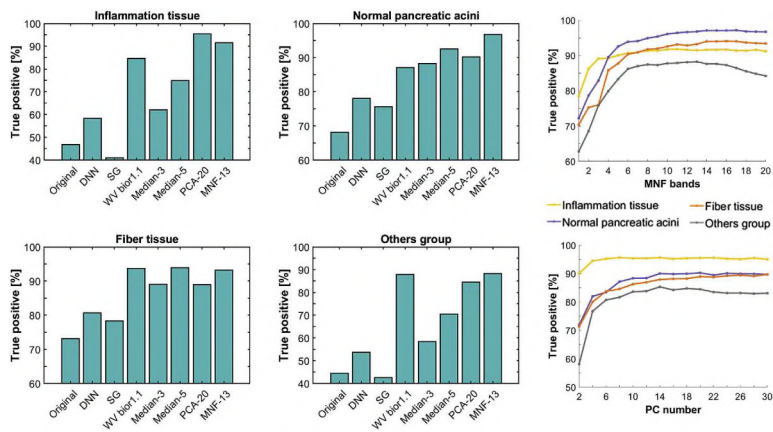


Fig. 1. A comparison of pixel-level classification results of spectral and spatial denoising based on the projected pixel size of 1.1 μm . Six methods: DNN, SG, Wavelets, Median Filter, PCA and MNF using different parameters, are set together with the original data. Figures show classification results for four histological classes of pancreatic tissue. Charts placed on the right show denoising parameters optimization for MNF and PCA method depend on classification results.

ponents (loadings in PCA and bands in MNF) used for reconstruction. The curves for both methods have the same character, for a small number of components classification results increase until it reaches a plateau area where only small fluctuations in values occur. The greatest results of 91.5%, 96.8%, 93.2% and 88.3% were obtained for MNF with 13 bands for each of the class, respectively. A clear rise of classifier performance was observed in the range of 2–20 MNF components. PCA characteristic was tested in the range from 1 to 100 (only part from 2 to 30 was presented due to decreasing curve character above 30 PC). For further calculation first 20 loadings were chosen, which gave results comparable with MNF; however, for inflammation tissue the true positive ratio is even higher. Accuracy for other classes was around 90.3%, 88.95 and 84.48% for PCA for normal pancreatic acini, fiber tissue and others group, respectively. In almost all classes a gain for Wavelets is observed, however, the method's characteristic is not consistent. For the fiber tissue class the obtained values are comparable with two the best methods, i.e. PCA and MNF, while for normal pancreatic acini small improvements comparable with Median Filter with kernel size 3 are observed. Results for Wavelets are strongly dependent on the input data. To summarize, the highest increase for each of the histopathological classes was as follows: inflammation tissue improvement by 44.7% done with PCA, normal pancreatic acini 28.1% with MNF, fiber tissue 20% with MNF, others group 43.7% with MNE.

Analogous analysis (as shown above in Fig. 1) was carried out for data with projected pixel size of 5.5 μm . The results are shown in Fig. 2.

Classification results are significantly different from those for 1.1 μm projected pixel size. The original data starting level of accuracy is much higher, which gives fewer opportunities for improvements in observed true positive ratio among considered histological classes and denoising methods. The variation in the

obtained results for classes is smaller than for data of projected pixel size of 1.1 μm but with similar importance for particular noise removal approaches. Thus the best denoising method achieved an improvement of only about 14.33% for inflammation tissue, 6.91% for normal pancreatic acini, 4.82% for fiber tissue and 13.92% for others group. The range of 2–30 MNF components was explored and in this region, a clear rise and a plateau of classifier performance were observed. PCA characteristic was tested in the range from 1 to 100 (only part from 2 to 30 was presented due to the decreasing curve character above 30 PC). For the final calculation 22 components were taken for both MNF and PCA. MNF method gave the best result of 97.5% for inflammatory tissue, 94.5% for normal pancreatic acini, 96.7% for fiber tissue and 92.91% for others group. PCA as well achieved high values, but comparable to Median Filter with kernel size 5. DNN, SG, Wavelets and Median Filter with kernel size 3 performed similar, without a clear increase in true positive ratio for each of the classes. Overall, on a pixel level MNF and PCA had the best performance, while Median with kernel size of 5 and Wavelets obtained very good results in some cases. This is a different situation than the one we observed for Quantum Cascade Laser (QCL) data [12]. There, only MNF was able to actually increase the classification accuracy of raw data, while here more approaches do work.

3.2. Core level optimization

Figs. 1 and 2 present the general classification for different methods based on the randomly split pixels, however, this may lead to overly optimistic performance measure as was recently shown by Perez-Guaita et al. [33] If pixels from the same cell are split to training and validation sets the result will be biased. This is especially relevant for spatial denoising methods and for HD imaging. Therefore, in order to get a better estimate of the performance for

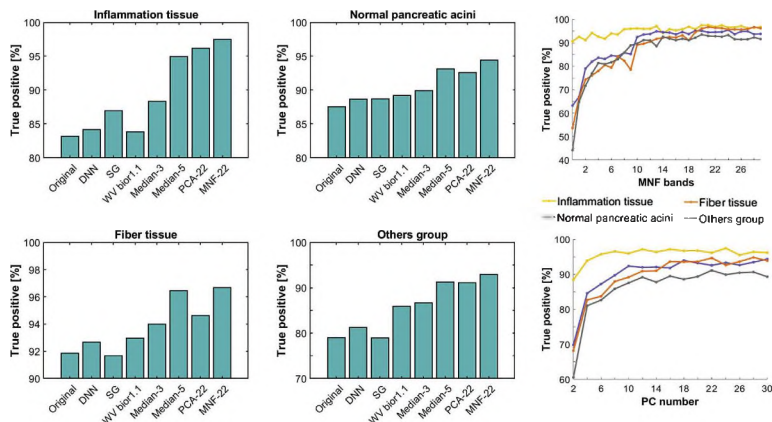


Fig. 2. A comparison of pixel-level classification results of spectral and spatial denoising based on the projected pixel size of 5.5 μm . Six methods: DNN, SG, Wavelets, Median Filter, PCA and MNF using different parameters, are set together with the original data. Figures show classification results for four histological classes of pancreatic tissue. Charts placed on the right show denoising parameters optimization for MNF and PCA method depend on classification results.

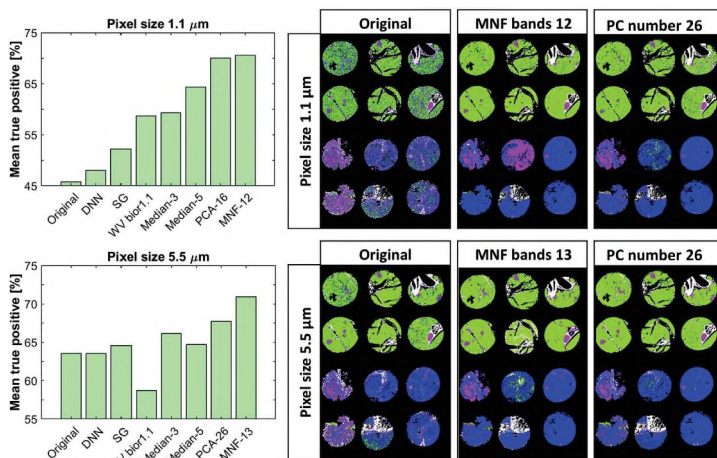


Fig. 3. (Left) Bar plots present comparison of mean cross-validation classification results for all of the histological classes for spectral and spatial denoising, based on the projected pixel size of (top) 1.1 μm and (bottom) 5.5 μm . On the x-axis, all used denoising methods are listed. (Left) Prediction results for original, MNF and PCA denoised data of each class for two pixel size (1.1 μm and 5.5 μm).

new biological samples (and not only regions of the samples as is happening in pixel level validation), a core level cross-validation was performed. Mean cross-validation classification results for all of the histological classes for DNN, SG, Wavelets, Median Filters with kernel size 3 and 5, PCA and MNF are presented in Fig. 3.

As before, a projected pixel size of 1.1 μm has a smaller original input data accuracy (45.8%) than the pixel size of 5.5 μm (63.6%). The number of components taken for MNF and PCA was chosen experimentally to maximize accuracy. PCA-16 and MNF-12 for a projected pixel size of 1.1 μm and PCA-26, MNF-13 for 5.5 μm data were chosen. For HD data the best predictive value, with the result of 70.6% for MNF and 70.1% for PCA, was obtained. Median Filter 5, which previously gave high classification results also performed well at 64.4% level. A surprisingly good result was obtained for Wavelets, comparable with Median 3 giving an improvement of 13.5%. Prediction values for DNN and SG are the smallest but still improve mean true positive ratio of original data by 2.25% (for DNN) and 5.4% (for SG). Using a well-optimized denoising method is able to greatly improve positive predictive value, even about 25% (value for MNF with 12 bands used for reconstruction).

For projected pixel size of 5.5 μm , DNN method gave no improvement for prediction reaching result equal with the original data of 63.5%, as well as SG and Median 5 which increase the accuracy by a small percent of 1.2%. Results for Median Filter 3 and 5 were surprising considering previous values presented in Fig. 2, where Median Filter 3 always had lower influence than Median Filter 5. The best effect was observed for PCA - 67.73% and MNF - 70.94%. These methods gave small but still leading among other methods mean true positive ratio improvement of 7.39% for MNF and 4.18% for PCA. The only method which did not contribute to

increasing the value and even lowered it was the Wavelets approach, reaching 58.72% of mean true positive value.

In order to visualize the cross validation results not only on selected regions, the full cores prediction is presented. Data are presented for both spatial resolutions. Differences can be noticed between the projected pixel sizes of the data as well as between the original and the denoised data. In the case of original data for a pixel size of 1.1 μm , significant errors in pixel classification are visible, both normal pancreatic acini and inflammatory tissue are wrongly classified in several regions. The use of MNF and PCA methods allowed eliminating inflammation tissue classification presence on healthy tissue, as well as better detection of histological structure like pancreatic islet from the others group class. A similar trend of improvement is noticeable for a pixel size of 5.5 μm data. Depending on the pixel size results, the improvement of MNF and PCA methods is better for classification of the inflammatory tissue class for 5.5 μm data. These differences are mainly coming from fiber tissue and others group as a consequence of highly histopathological complexity of inflammation tissue.

3.3. Influence of denoising on the spectral input

An increase in classification accuracy in the order of 30–50% may seem unrealistic and it is important to validate the method's influence on the data itself. We have previously shown [12] the effect of MNF on average spectra of QCL imaging data and it is hard visually to see spectral differences. Therefore, we decided to look more into the actual input structure (metrics) rather than spectra. Random Forest offers the ability to estimate the importance of input variables and this information enables to observe the

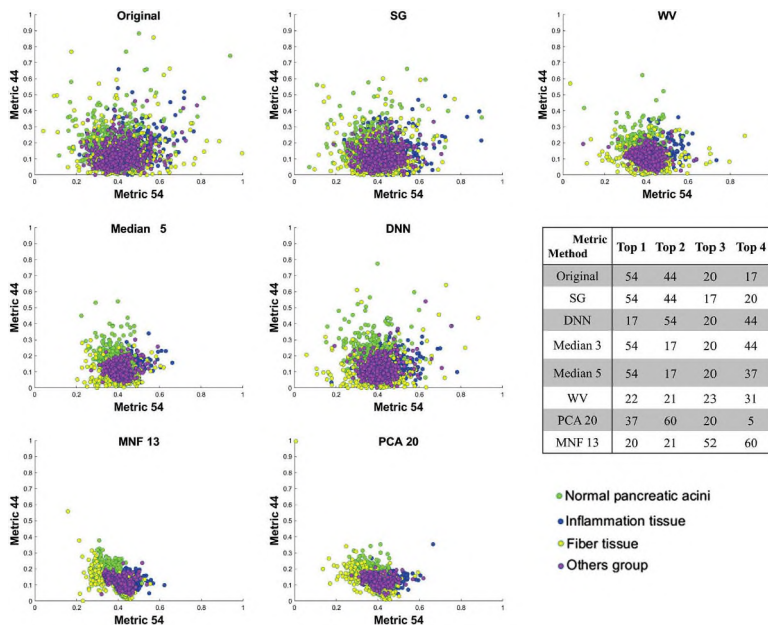


Fig. 4. Scatter plots show the pixels distribution for each histological class based on most important original data metrics. On the axes, the first two top metrics for the original data were presented -vertically metrics 44, horizontally 54. The results were presented for all denoising methods (DNN, SG, Wavelets, Median Filter, PCA and MNF) for a projected pixel size of 1.1 μm . The table contains four top metrics for each of the methods.

influence of denoising on the top variables (metrics) selected by the classification algorithm. Thus, two of the highest important metrics numbers of original data were selected to present pixels distribution for each the histological classes of the marked regions (used for pixel level classification). Combinations of these metrics for each of the considered techniques are presented in scatter plots in Fig. 4 and Fig. 5 for a projected pixel size of 1.1 μm and 5.5 μm , respectively.

Tables in the right bottom corners present the list of the four most important metrics selected by the various denoising methods. Pixels belonging to each of a histopathological class are marked according to the colors described in the attached below plots legend.

The presented in Fig. 4 differences for data of projected pixel size of 1.1 μm between scatter plots are visible at a first glance. Original pixels distributions show a strong overlap between the classes and also a vast pixel spread around classes. Class distinction is hard to achieve. A similar effect was obtained for the SG method. Comparing the original data and after SG denoising metrics importance, it could be noticed that the first top metrics are the same, what explains similarity of pixels distribution. Another group with comparable results is taken for DNN, Median Filters with a

kernel 3 (plot not included) and 5. In this case histopathological classes are more apparent, however pixels are still strongly co-localized with others. Visually, the distribution of Wavelets is similar to this group, however, the metrics choice is quite different from the original data. The largest influence in the ability of class distinction was obtained for MNF and PCA methods, pixels were compacted within assigned histopathological classes the most. Moreover, such good results for MNF and PCA were obtained for completely different metrics than their most optimal, for which the discrimination would be even better.

Distributions presented in Fig. 5 are not as clear as in the previous case. Visually all methods are close to each other and a distinction of histopathological classes is possible. However, the worst results were achieved for Wavelets, what has been already observed during classification (Figs. 2 and 3). Both original data and after denoising by SG, DNN and Median Filters with kernel 3 and 5 have almost the same metrics chosen as the best as is presented for data of pixel size of 1.1 μm . However, again metrics selected by the PCA and MNF are different from original data. What is more, metrics for PCA and MNF are not identical for data of projected pixel size of 1.1 μm and 5.5 μm , which suggests that these methods are depended on the exact input.

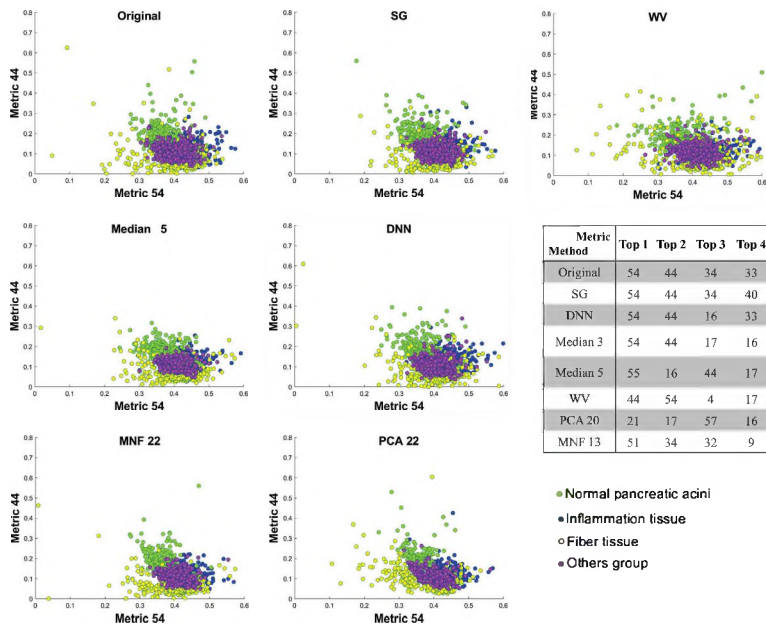


Fig. 5. Scatter plots show the pixels distribution for each histological class based on the most important original data metrics. On the axes, the first two top metrics for the original data were presented –vertically metrics 44, horizontally 54. The results were presented for all denoising methods (DNN, SG, Wavelets, Median Filter, PCA and MNF) for a projected pixel size of 5.5 μm . The table contains four top metrics for each method.

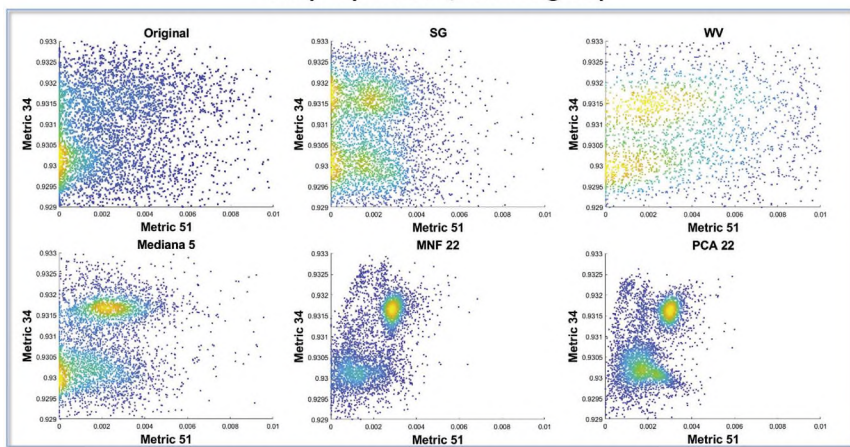
To better highlight the influence of MNF denoising on the 5.5 μm data, scatter plots for two classes were prepared (remaining two are shown in Fig. S2 in Supplementary Materials) using the best metrics for this method and are shown in Fig. 6.

Images show density distributions, which means that the colors correspond to numbers of pixels at a given point. The DNN method was not included because of a very similar character to SG. Density distributions for original data, SG and WV in the others group show that a large portion of the pixels is dispersed. The best results were obtained for MNF and PCA methods, which distributions for others group class are condensed and clearly show pixels agglomeration. Similar results are observed for the second type class – normal pancreatic acini. For original data, SG, WV and Median Filters the place with the highest density is extended as well as the pixels are highly dispersed. PCA and MNF methods again cause the best pixel concentration. It can be seen that MNF and PCA have the highest influence on the classification space working primarily by condensing the distributions of a given class. Moreover, since regardless of projected pixel size the optimal metrics chosen are different for PCA and MNF than for other methods, it is clear that crucial information for classification as hidden in the noise level and now can be fully leveraged.

4. Conclusions

Overall, MNF and PCA (both based on similar mathematical principles) are outperforming other denoising methods in their capability to increase classification accuracy of FT-IR based histopathological models using Random Forest. The increase in pixel level accuracy for High Definition data (1.1 μm projected pixel size) was found to be dependent on the complexity of the histopathological class and reached even 43–44% level, while core level increase reached around 28%. The results for 5.5 μm projected pixel size were smaller in magnitude, but still very significant, with the same methods outperforming others. Median filter and Wavelets in some cases gave a high increase, but it was dependent on the class and projected pixel size. Moreover, for prediction of new cores, these methods did not perform in a stable way. The effect of MNF and PCA on the classification input has been shown to be primarily in narrowing the inter-class spread while also leveraging more noisy spectral features to be used in classification. Overall, the results strongly favor the use of multivariate denoising techniques of PCA and MNF over any other approach for classification purposes of FT-IR imaging data. This is different from Quantum Cascade Laser IR imaging data, where only MNF was found previously to increase the

5.5 μm pixel size, others group



5.5 μm pixel size, normal pancreatic acini

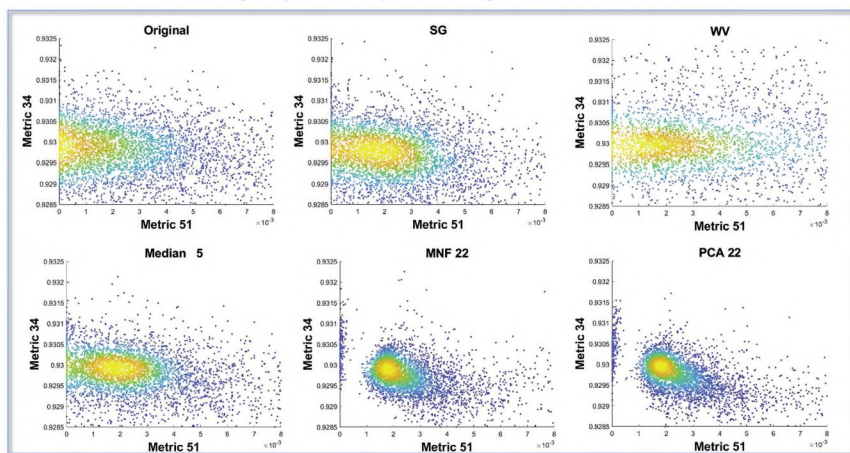


Fig. 6. Scatter plots show the pixels density distribution for two histological classes, others group (top) and normal pancreatic acini (bottom), based on major MNF data metrics for a projected pixel size of 5.5 μm presented for following denoising methods: SG, Wavelets, Median Filter, PCA and MNF and original data.

classification potential.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Conflicts of interest

The authors declare no conflicts of interest.

Acknowledgements

This research was supported from the „Pancreatic cancer comprehensive histopathology based on IR chemical imaging“ project, which is carried out within the Homing programme (grant no. Homing/2016-2/20) of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund. This research was performed using equipment purchased in the frame of the project co-funded by the Malopolska Regional Operational Program Measure 5.1 Krakow Metropolitan Area as an important hub of the European Research Area for 2007–2013, project no. MRPO.05.01.00-12-013/15.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.aca.2019.07.045>.

References

- [1] T.P. Wrobel, R. Bhargava, Infrared spectroscopic imaging advances as an analytical technology for biomedical sciences, *Anal. Chem.* 90 (2018) 1444–1463, <https://doi.org/10.1021/acs.analchem.7b05330>.
- [2] T.P. Wrobel, A. Vichi, M. Baranska, S.G. Kazarian, Mid-attenuated total reflection fourier transform infrared (FT-IR) spectroscopic imaging with variable angles of incidence, *Appl. Spectrosc.* 69 (2015) 8–13, <https://doi.org/10.1366/15-07963>.
- [3] K. Yeh, D. Lee, R. Bhargava, Multicolor discrete frequency infrared spectroscopic imaging, *Anal. Chem.* 90 (2019) 2177–2185, <https://doi.org/10.1021/acs.analchem.8b04749>.
- [4] M. Hermes, R.B. Morrish, L. Huo, L. Meng, S. Junaedi, J. Tomko, G.R. Lloyd, W.T. Masselink, P. Tidemand-Lichtenberg, C. Pedersen, F. Palombo, N. Stone, Mid-IR hyperspectral imaging for label-free histopathology and cytology, *J. Opt.* 20 (2018) 10.1088/2040-8986/aa36b6.
- [5] T.P. Wrobel, K.M. Marzec, S. Chlopicki, E. Maslak, A. Jasztal, M. Franczyk-Zarow, I. Czyzyska-Gichon, T. Moszkowski, R.B. Kostogryns, M. Baranska, Effects of Low Carbohydrate High Protein (LCHP) diet on atherosclerotic plaque phenotype in ApoE/LDLR^{-/-} mice: FT-IR and Raman imaging, *Sci. Rep.* 5 (2015) doi:10.1038/srep14002.
- [6] K. Isensee, N. Kröger-Lui, W. Petrich, Biomedical applications of mid-infrared quantum cascade lasers – a review, *Analyst* 143 (2018) 5888–5911, <https://doi.org/10.1039/C8AN01306C>.
- [7] J.A. Kimber, S.G. Kazarian, Spectroscopic imaging of biomaterials and biological systems with FTIR microscopy or with quantum cascade lasers, *Anal. Bioanal. Chem.* 409 (2017) 1–8, <https://doi.org/10.1007/s00216-017-0574-5>.
- [8] K. Augustyniak, K. Chrabaszcz, A. Jasztal, M. Smeda, G. Quintas, J. Kuligowski, K.M. Marzec, K. Malek, High- and Ultra-High definition of IR spectral histopathology gives an insight into chemical environment of lung metastases in breast cancer, *J. Biophot.* (2018) e201800345, <https://doi.org/10.1002/jbio.201800345>.
- [9] H.H. Chen, T.T. Lee, A. Chen, Y. Hwu, C. Petibois, 3D digital pathology for a chemical-functional analysis of glomeruli in health and pathology, *Anal. Chem.* 90 (2018) 3811–3818, <https://doi.org/10.1021/acs.analchem.7b04265>.
- [10] L.S. Leslie, T.P. Wrobel, D. Mayerich, S. Bindra, R. Emmadi, R. Bhargava, High definition infrared spectroscopic imaging for lymph node histopathology, *PLoS One* 10 (2015) e0127238, <https://doi.org/10.1371/journal.pone.0127238>.
- [11] B. Bird, J. Rowlette, A protocol for rapid, label-free histochemical imaging of fibrotic liver, *Analyst* 1208 (2017), <https://doi.org/10.1039/C6AN02080A>.
- [12] P. Kozioł, M.K. Raczkowska, J. Skibińska, N.J. McCollum, S. Urbanik-Wasik, C. Paluszkievicz, W.M. Kwiatek, T.P. Wrobel, Denoising influence on discrete frequency classification results for quantum cascade laser based infrared microscopy, *Anal. Chim. Acta* 1051 (2019) 24–31, <https://doi.org/10.1016/j.aca.2018.11.032>.
- [13] S. Berisha, M. Lotfollahi, J. Jahanpour, I. Gurcan, M.J. Walsh, R. Bhargava, H. Nguyen, D. Mayerich, Deep learning for FTIR histology: leveraging spatial and spectral features with convolutional neural networks, *Analyst* (2018) 10–13, <https://doi.org/10.1039/C8AN01495G>.
- [14] P. Lasch, M. Stämmler, M. Zhang, M. Baranska, A. Boschi, K. Majzner, FT-IR hyperspectral imaging and artificial neural network analysis for identification of pathogenic bacteria, *Anal. Chem.* 90 (2018) 8896–8904, <https://doi.org/10.1021/acs.analchem.8b01024>.
- [15] U. Sharaha, E. Rodriguez-Diaz, O. Sagi, K. Riesenber, I. Lapidot, Y. Segal, I.J. Bigio, M. Huelshel, A. Salman, Detection of extended-spectrum β -lactamase-producing *Escherichia coli* using infrared microscopy and machine-learning algorithms, *Anal. Chem.* (2019), <https://doi.org/10.1021/acs.analchem.8b05497>.
- [16] D.C. Fernandez, R. Bhargava, S.M. Hewitt, I.W. Levin, Infrared spectroscopic imaging for histopathologic recognition, *Nat. Biotechnol.* 23 (2005) 469–474, <https://doi.org/10.1038/nbt1080>.
- [17] M. Diem, A. Mazur, K. Lénau, J. Schubert, B. Bird, C. Krafft, Molecular pathology via IR and Raman spectral imaging, *J. Biophot.* 886 (2013) 855–886, <https://doi.org/10.1002/jbio.201300131>.
- [18] M.J. Pilling, A. Henderson, J.H. Shanks, M.D. Brown, N.W. Clarke, P. Gardner, Infrared spectral histopathology using haematoxylin and eosin (H&E) stained glass slides: a major step forward towards clinical translation, *Analyst* 142 (2017) 1258–1266, <https://doi.org/10.1039/C6AN02244c>.
- [19] M.J. Baker, H.J. Byrne, J. Chalmers, P. Gardner, R. Goodacre, A. Henderson, S.G. Kazarian, F.L. Martin, J. Moger, N. Stone, J. Sulé-Suso, Clinical applications of infrared and Raman spectroscopy: state of play and future challenges, *Analyst* 143 (2018) 1735–1757, <https://doi.org/10.1039/C7AN01871a>.
- [20] S.H. Remiszewski, C.M. Thompson, D.J. Raz, A. Akalin, X. Mu, M.A. Kon, M. Diem, Classification of malignant and benign tumors of the lung by infrared spectral histopathology (SHP), *Lab. Invest.* 95 (2015) 406–421, <https://doi.org/10.1038/abinvest.2015.1>.
- [21] J.T. Kwak, S.M. Hewitt, S. Sinha, R. Bhargava, Multimodal microscopy for automated histologic analysis of prostate cancer, *BMC Canc.* 11 (2011) 62, <https://doi.org/10.1186/1471-2407-11-62>.
- [22] M.J. Walsh, S.E. Holton, A. Kajdacsy-Balla, R. Bhargava, Vibrational Spectroscopy Attenuated total reflectance Fourier-transform infrared spectroscopic imaging for breast histopathology, *Vib. Spectrosc.* 60 (2012) 23–28, <https://doi.org/10.1016/j.vibspec.2012.01.010>.
- [23] B. Bird, J. Rowlette, High definition infrared chemical imaging of colorectal tissue using a Spero QCL microscopy, *Analyst* 142 (2017) 1381–1386, <https://doi.org/10.1039/C6AN01916a>.
- [24] K. Yeh, S. Kenkel, J. Liu, R. Bhargava, Fast infrared chemical imaging with a quantum cascade laser, *Anal. Chem.* 87 (2015) 485–493, <https://doi.org/10.1021/acs.analchem.5b02751>.
- [25] C. Kaeppner, A. Kalenb, H. Jaette, A. Tannapfel, F. Großeschenckamp, K. Gerwert, Quantum cascade laser-based infrared microscopy for label-free and automated cancer classification in tissue sections, *Sci. Rep.* (2018) 1–10, <https://doi.org/10.1038/s41598-018-26098-w>.
- [26] P. Kozioł, M.K. Raczkowska, J. Skibińska, S. Urbanik, C. Paluszkievicz, W.M. Kwiatek, T.P. Wrobel, Comparison of spectral and spatial denoising techniques in the context of High Definition FT-IR imaging hyperspectral data, *Sci. Rep.* 8 (2018) 1–11, <https://doi.org/10.1038/s41598-018-32713-7>.
- [27] T.P. Wrobel, M.R. Kole, R. Bhargava, Emerging Trends and Opportunities in Discrete Frequency Infrared and Raman Spectroscopic Imaging, *Spectroscopy* 31 (2016) 28–45.
- [28] B. Zimmermann, A. Kohler, Optimizing Savitzky-Golay parameters for improving spectral resolution and quantification in infrared spectroscopy, *Appl. Spectrosc.* 67 (2013) 892–902, <https://doi.org/10.1366/12-06723>.
- [29] P. Lasch, Spectral pre-processing for biomedical vibrational spectroscopic imaging, *Chemometr. Intell. Lab. Syst. J.* 117 (2012) 100–114, <https://doi.org/10.1016/j.chemolab.2012.03.011>.
- [30] B.K. Alsberg, A.M. Woodward, M.K. Winslow, J. Rowland, D.B. Kel, Wavelet denoising of infrared spectra, *Analyst* 122 (1997) 645–652, <https://doi.org/10.1039/A608255F>.
- [31] A.A. Green, M. Berman, P. Switzer, M.D. Craig, A transformation for ordering multispectral data in terms of image quality with implications for noise removal, *IEEE Trans. Geosci. Remote Sens.* 26 (1988) 65–74, <https://doi.org/10.1109/36.3001>.
- [32] J.C. Church, Y. Chen, S. V. Rice, A Spatial Median Filter for noise removal in digital images, in: *IEEE Southeastcon Vol* 2008, 2008, pp. 618–623, <https://doi.org/10.1109/SECON.2008.4494449>.
- [33] D. Pérez-Guaita, J. Kuligowski, B. Lendl, B.R. Wood, G. Quintas, Assessment of discriminant models in infrared imaging using constrained repeated random sampling – cross validation, *Anal. Chim. Acta* 1033 (2018) 156–164, <https://doi.org/10.1016/j.aca.2018.05.019>.



Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Denoising influence on discrete frequency classification results for quantum cascade laser based infrared microscopy



Paulina Koziol^a, Magda K. Raczowska^{a, b}, Justyna Skibinska^{a, c}, Nicholas J. McCollum^d,
Sławka Urbaniak-Wasik^e, Czesława Paluszkiwicz^a, Wojciech M. Kwiatek^a,
Tomasz P. Wrobel^{a, *}

^a Institute of Nuclear Physics Polish Academy of Sciences, PL-31342 Krakow, Poland

^b Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, Mickiewicza 30, Krakow, Poland

^c Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering, AGH University of Science and Technology, Mickiewicza 30, Krakow, Poland

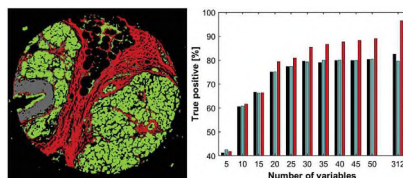
^d DRS Daylight Solutions, 15378 Avenue of Science, Suite 200, San Diego, USA

^e NZOZ Pathology Department, Jagiellońska 70, Kielce, Poland

HIGHLIGHTS

- QCL approach enables a truly Discrete Frequency approach speeding up IR imaging by a factor of 100–1000.
- Effects of spatial and spectral denoising on classification accuracy of QCL data were studied.
- Most important factor in classification accuracy is the number of experimental spectral variables.
- Only MNF has significantly improved classification results, but only when more than 20 variables were denoised.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:

Received 14 August 2018

Received in revised form

14 November 2018

Accepted 15 November 2018

Available online 21 November 2018

Keywords:

Quantum cascade laser

PCA

MNF

Deep neural network

Wavelets

Random forest

ABSTRACT

Currently, there is great interest in bringing the application of IR spectroscopy into the clinic. This however will require a significant reduction in measurement time as Fourier Transform Infrared (FT-IR) imaging takes hours to days to scan a clinically relevant specimen. A potential remedy for this issue is the use of Quantum Cascade Laser Infrared (QCL IR) microscopy performed in Discrete Frequency (DF) mode for maximum speed gain. This gain could be furthermore improved by applying a proper denoising algorithm that takes into account the specific data structure. We have recently compared spectral and spatial denoising techniques in the context of Fourier Transform IR (FT-IR) imaging and showed that the optimal methods depend heavily on the exact data structure. In general multivariate denoising methods such as Principal Component Analysis (PCA) and Minimum Noise Fraction (MNF) are the most effective for a dataset containing multiple bands. Histologic classification of QCL IR images of pancreatic tissue using Random Forest was therefore performed to investigate which denoising schemes are the most optimal for such experimental data structure. This work is the first to show the effects of denoising on

* Corresponding author.

E-mail address: tomasz.wrobel@ifj.edu.pl (T.P. Wrobel).

classification accuracy of QCL data and is likely to be transferable to other QCL microscopes and other modalities using DF imaging, e.g. AFM-IR or CARS/SRS imaging.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Infrared (IR) imaging is a technique providing information about the chemical composition of the sample in a spatially-resolved manner and is gaining attention especially in the biomedical field [1–4]. The field of IR imaging is rapidly evolving with new technical advances opening avenues previously unattainable [5–7]. One good example is the emergence of Quantum Cascade Lasers (QCL) [8] as a source of IR radiation, which not only provides higher energy flux but also permits truly Discrete Frequency (DF) imaging [6,9–11] enabling high-speed imaging of samples [12–17]. The acquisition speed enhancement factor can vary from ~100 to ~1000 depending on the QCL microscope design [9,10,17].

Noise removal algorithms in IR imaging have mostly originated in traditional IR spectroscopy [18] and only a limited number of spatial denoising techniques were tested on IR imaging data. Therefore, we have recently compared spectral and spatial denoising techniques in the context of Fourier Transform Infrared (FT-IR) imaging [19] and showed that the optimal methods depend heavily on the exact data structure. In general multivariate denoising methods such as Principal Component Analysis (PCA) and Minimum Noise Fraction (MNF) [20–23] were the most effective for a dataset containing approximately 800 bands (a typical FT-IR imaging dataset). Including more pixels results in a higher denoising efficiency. This poses an interesting question for Discrete Frequency imaging as a similar behavior is expected with the number of spectral variables. Few variables should be less efficiently denoised than e.g. 200. It may be the case that up to a few variables denoising in the spatial domain (inherently based on a single spectral variable/frequency) will outperform PCA/MNF approaches. Since the principle of DF imaging is to acquire as many frequencies as possible for a given system and problem, it is important to determine at what number of frequencies the multivariate benefit of PCA/MNF will outperform spatial denoising.

However, there is no good metric of spectral Signal to Noise Ratio (SNR) if only a few spectral variables are present while peak Signal to Noise Ratio (pSNR) does not provide as much insight about useful information content. To overcome this we decided to focus on the histologic classification of QCL IR images of pancreatic tissue using Random Forest, which offers the ability to use only the most important variables/frequencies. Using classification accuracy enables us to investigate which denoising schemes are the most optimal for such experimental data structure.

2. Materials and methods

2.1. Sample preparation and measurement

A paraffin embedded five micron-thick Tissue Micro Array section (purchased from US Biomax, Inc) served as a source of single pancreatic tissue core of 1.5 mm in diameter. The sample was mounted on BaF₂ transmission window and deparaffinized with 24 h hexane bath. Measurements were performed by DRS Daylight Solutions, CA, USA using Spero QT – a QCL based microscope with 0.7 NA objective and Microbolometer Focal Plane Array (FPA) 480 × 480 detector, which gives a projected pixel size of 1.3 μm and a 650 μm field of view. Therefore, each analyzed core consists of 9

image tiles. Images were acquired in transmission mode with the spectral range from 1176 cm⁻¹ to 1800 cm⁻¹ and with 2 cm⁻¹ spectral resolution, giving 312 variables/frequencies. The total acquisition time for a single core with a High Resolution objective was approx. 5 min – a corresponding High Definition FT-IR image would take approx. 1.5–2 h with a minimal number of scans.

2.2. Signal processing and classification

The high accuracy in biomedical imaging, robustness and the ability to estimate each variable importance in classification, makes Random Forest (RF) [23–27] most appropriate in DF imaging spectroscopy. ImageJ software was used to generate tissue masks corresponding to analyzed classes of normal pancreatic acini, secretory duct wall and stroma, which further in the process allowed creating spectral profiles of each class. The class masks were marked under supervision of a trained histopathologist. Data processing, as well as classification, were done in MATLAB software using built-in and home written routines. During pre-processing, all spectra were baseline corrected (linear baseline between first and last variables) and normalized to amide I band. In the first part of the project, classification was performed on full preprocessed (further referred as original) spectra as well as spectra denoised with different techniques and parameters: Deep Neural Network (DNN), Wavelets (WV), Median Filter, PCA and MNF. For the spatial techniques (DNN, WV, Median Filter), denoising was applied to all images like it was used for all spectra in the case of PCA and MNF. Different techniques with different parameters were tested, and the choice was made based on results reported in Ref. [19].

To provide precise information about Gini variable importance, a single classification with 1000 trees parameter was performed. Based on that, the dependence of classification results on the number of most important variables taken into classification was investigated, both for pre-processed and denoised data (denoising was applied after variables number reduction). In all cases, input spectra were randomly split for training and validation datasets, and all RFs classifiers, apart from the previously mentioned one, were built with 50 trees. Such split of the pixels in the case of projected pixel size (1.3 μm) lower than the spatial resolution may lead to pixels from the same cell to be used for training and validation leading to over optimistic validation results as was recently noticed by Perez-Guaita et al. [28]. However, the aim of this study was to look at denoising effect on accuracy and not to create the most reliable and accurate models, therefore the absolute values of performance should be taken with caution.

3. Results and discussion

The first step in investigating the effect of denoising on QCL data was to check whether the specific data and noise structure alters the way different approaches remove noise. In order to achieve that, a classifier was trained using all of the 312 variables using both the original data and after applying all of the denoising algorithms. Results of classification accuracies with different pretreatments are shown in Fig. 1.

For Wavelets, different families and levels were tested, based on the best performances for FT-IR reported in Ref. [19]. Presented

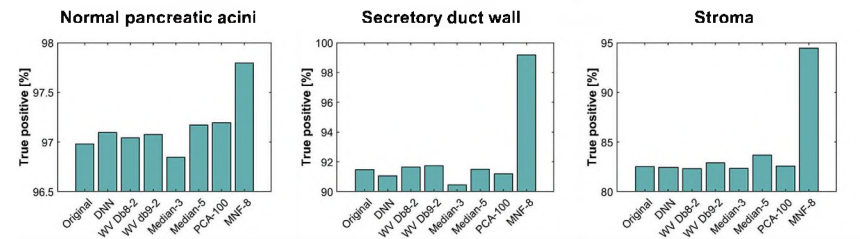


Fig. 1. Comparison of the full spectrum (312 variables) based classification results (using the correct classification, i.e. the true positive, as a measure) for original and denoised datasets for three histologic classes of pancreatic tissue. Five denoising methods were explored: Deep Neural Networks, Wavelets, Median Filter, PCA and MNF, with Wavelets and Median using different sets of parameters.

here are Daubechies wavelets of order 8 level 2 and order 9 level 2, which gave the best, but still small improvements in the tested group. Other spatial techniques like DNN and Median Filter with kernel size 3 and 5 performed similarly, without a clear increase in true positive rate for either of the classes. Almost no improvement was observed in true positive rate for normal pancreatic acini class, which is not surprising because there was not much room for improvement when results exceed 95%. Even though for the two other classes the results based on original data were much lower, a clear improvement from spatial denoising is not observed. It may be a result of a relatively good single frequency image quality (Fig. 2). PCA and MNF methods have principally only one parameter to optimize, i.e. the number of components in PCA and bands in MNF that are used for reconstruction. PCA denoising with a low number of bands distorted classification, therefore, only results with 100 bands are shown here. No rise in classification accuracy was observed for PCA, even though it gives good noise reduction for FT-IR data as measured by SNR gain. This may be explained by the fact, that noise is not homogenous spectrally and PCA is not taking that into account. This may lead to noise of high variance to be retained in reconstruction. The best results of 97.8%, 99.2% and 94.5% were obtained for MNF with 8 bands. The range of 2–25 MNF components was explored and in this region, a clear rise and fall of classifier performance were observed (results not shown). This significant increase of 8% and 12% for secretory duct wall and stroma classes respectively for MNF with a low number of components taken into reconstruction suggests, that this technique gives the biggest potential for classification improvement after variables number reduction.

Fig. 2 presents the tissue core which served as an input data along with classification results using MNF. H&E stain part shows various tissue regions which are easily identifiable on the QCL Microscope image. Since results presented in the previous figure clearly show that the biggest gain in true positive rate was reached for MNF, a comparison of prediction results for original and MNF denoised data is presented here. Original data based prediction allows definite distinction of all three classes, with only random pixels mixed in between. Denoising resulted in an obvious increase in numerical values, but prediction improvement was not as clear. Only after a closer examination, pixels mixed in secretory duct wall and stroma classes were smoothed, furthermore, structures gained more sharpness on the edges.

Mean spectra of original and MNF denoised data revealed differences in nearly the whole spectral range. Denoising, however, caused secretory duct wall and stroma spectra to almost fully overlap with some separation in DNR/RNA and protein regions. This

could have been expected since both classes are types of fiber tissue. Pancreatic acini class is more differentiated from the others making it easier to classify and the results reflect this. Overall, the spectral differences between classes seem smaller after MNF denoising; however, the original profiles mostly differ in unwanted spectral distortions, such as baseline effects and scattering. Denoising leads to more homogenous groups which can be more easily separated in a systematic way.

MNF and PCA techniques are conceptually very similar with an additional step of noise whitening for MNF. Their performance for FT-IR data was found to be comparable as shown in Ref. [19] in terms of SNR. Based on that, it was expected that these two methods will both give a good increase in classification results. However, only MNF presented positive outcome and PCA was only able to reach initial levels of true positive rates. This might be caused by additional noise calculation step in MNF, which allows a good estimation of data-specific noise structure. Despite that, it seemed reasonable to further investigate both denoising techniques and their influence on classification and prediction with a reduced number of variables since PCA may reach the classification power of the original data with fewer variables and thus it could still reduce acquisition times.

The analysis was carried out for PCA and MNF with the number of reconstruction bands in the range from 2 to 20 and with a number of spectral variables ranging from 5 to 50. As it was mentioned in the materials and methods section, variables were chosen based on importance for 1000 trees classification. The obtained results are presented in Fig. 3 and also include results with 312 variables (full spectrum).

For both methods with the increasing components number, a rapid rise in classification results for all data series was observed. The peak value for MNF occurred around 8–10 bands, while PCA reached a plateau a little later – 10–14 bands. Prompt improvement is also observed for increasing variables taken into classification. The difference between 5 and 10 variables is significant and what is more, 20–30 variables enabled reliable classification results. It was shown earlier that MNF gives much better results in the whole spectrum case, it is also depicted in Fig. 3, but what is interesting, is that these MNF advantages are observed only for 20–30 variables and more. Below that both techniques acted similarly. It is also interesting to note, that each class has a specific data structure that seems to be more difficult for PCA to capture, e.g. stroma for the full spectrum (sea green color) up to 5 reconstruction bands has no improvement in classification accuracy, while MNF with just 1 band improved the accuracy. Overall, MNF results were much more robust to data structure. Interestingly,

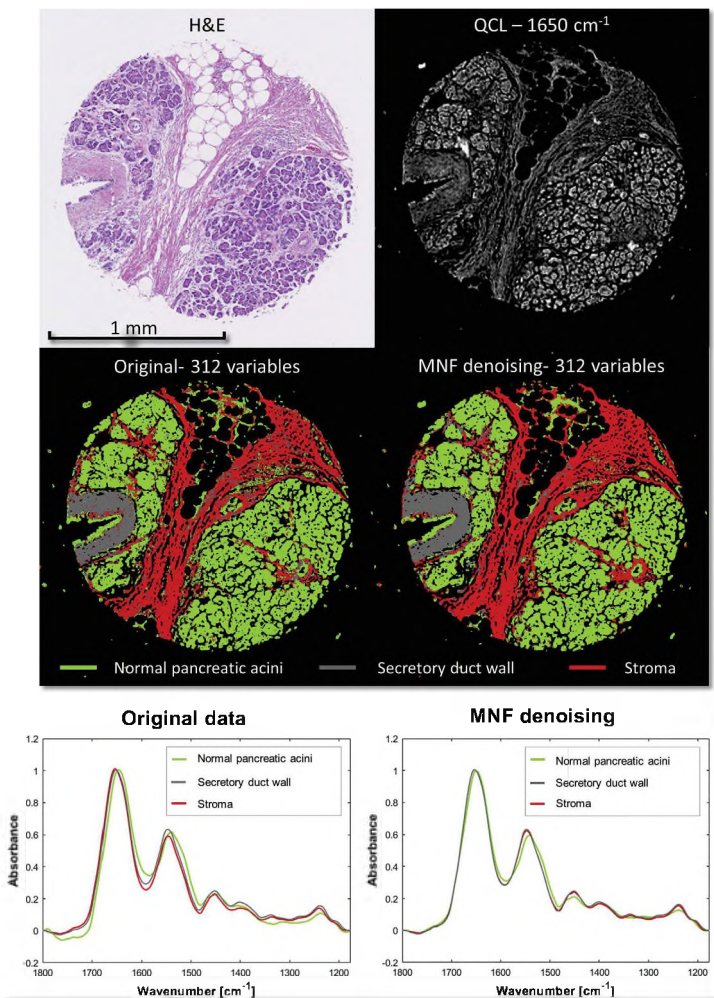


Fig. 2. (Top) H&E and QCL microscope (1650 cm^{-1}) based image of examined tissue core. (middle) Prediction results for original and MNF denoised data for full spectrum classification. (bottom) Mean spectra of each class for original and MNF denoised data. For data presented above, reconstruction in MNF denoising was done with 8 components.

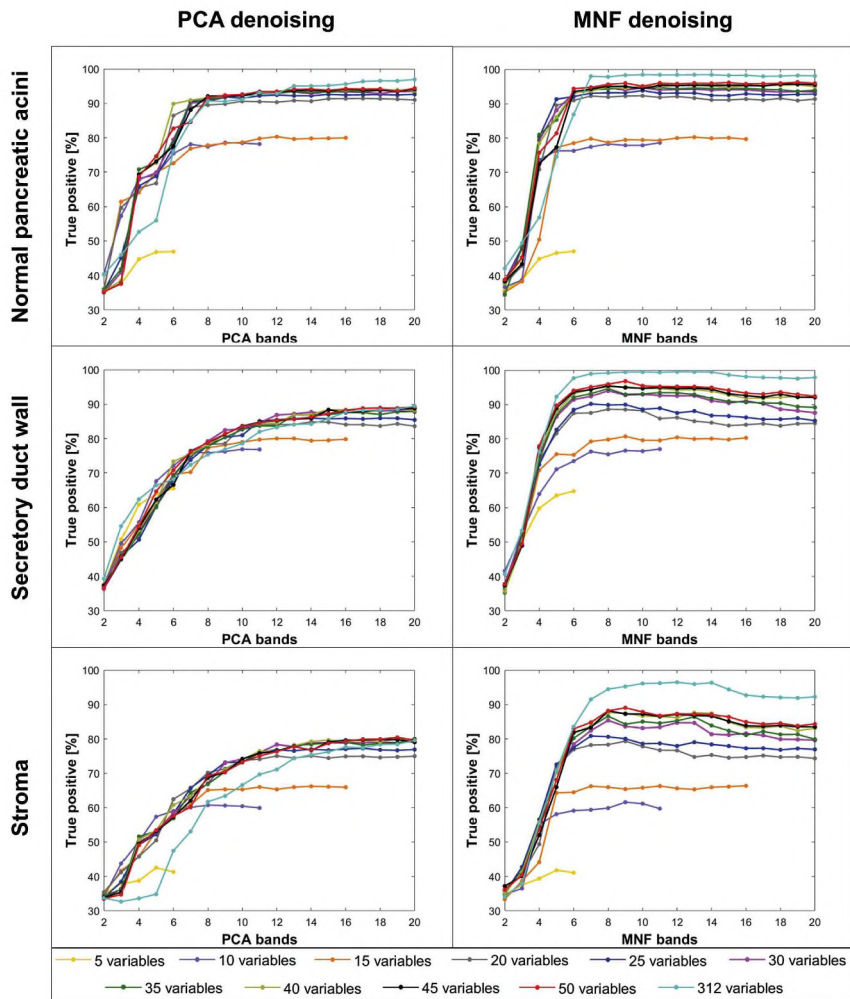


Fig. 3. PCA (left column) and MNF (right column) denoising influence on classification results for different numbers of reconstruction bands (x-axis) and input variables (different series) for each class with a comparison to full spectrum denoising.

using too many MNF bands for reconstruction lowers the classification accuracy in a more significant way than it occurs for PCA. We have observed very similar behavior of these techniques in terms of SNR gain [19] and we conclude that the decreasing SNR gain in MNF

explains the lowering classification accuracy.

To underline results improvement delivered by denoising, the maxima reached for each data series in Fig. 3 for both techniques were compared with the original data with the reduced number of input variables – results are shown in Fig. 4. Starting from 20 variables, there is a clear improvement in the results caused by MNF denoising. It seems that the data reaches a plateau around 30 variables and further increasing the number of variables creates only slight changes. Surprisingly, for 5 variables PCA denoising gives a small increase of results (class dependent), so it could be suspected that for a very limited variables number this method might give some advantages but this assumption needs further investigation. After all, this data confirms results presented in Fig. 1 – PCA denoising was not beneficial for classification based on QCL microscopic data, but MNF gave a significant rise in true positive rates.

To confirm the prediction power of classification based on the reduced number of variables, prediction images for 5, 10 and 30 variables are presented in Fig. 5. It is clear that for top pictures (5 variables) prediction results were poor, especially for normal pancreatic acini and secretory duct wall classes where it is not possible to distinguish between structures. Additionally, it is known that one of the major QCL based microscopes problems are coherence effects. One of the effects is caused by the sample itself and the other by laser instabilities, which leads to differences between background and sample measurements and therefore illumination based shapes (concentric rings) are appearing in the measurement area. This effect is clearly seen in the top part of Fig. 5. For 10 variables this effect was much weaker and almost absent for 30 variables, hence it can be concluded that more variables are able to counter this effect. Even though new generation of QCL microscopes have minimized this effect [17], it still can be crucial for classification based on a small number of variables and should be a focus of future technical improvements.

True positives rate for each class (Fig. 4) shows that there is a big jump in quality of results between 5 and 10 variables and it is also visible on the prediction images (middle part of Fig. 5). It is definitely possible to identify all three classes, although a lot of pixels are intermixed. 30 variables based prediction results are also presented (bottom part) and it confirmed earlier results that 30 is a reasonable number of variables to generate a reliable classification. It should be noticed that unlike in the two upper cases, here the MNF denoising produced prediction improvement which can be observed in the zoomed areas. Mixed pixels were strongly reduced (especially in the secretory duct wall class – gray) and some structures were more clearly exposed (stromal features – red). To better visualize the changes in pixel assignment, a class change was calculated for regions corresponding to ground truth (Fig. 5 right panel) between original and MNF denoised data. The corresponding pixel percentage changes were summarized in Table 1. It can be seen that only for 30 variables there is an improvement in prediction.

Overall, the number of variables used had a larger influence on the classification results than on the denoising efficiency of MNF. Above 20 input variables, the classification benefits from MNF, but up to 30 variables the classification quality is more based on the number of experimental variables. Above 30 variables classification is mostly improved by the multivariate nature of data and MNF.

4. Conclusions

Currently, there is great interest in bringing the application of IR spectroscopy into the clinic. This however will require a significant reduction in measurement time as Fourier Transform Infrared (FT-IR) imaging takes hours to days to scan a clinically relevant

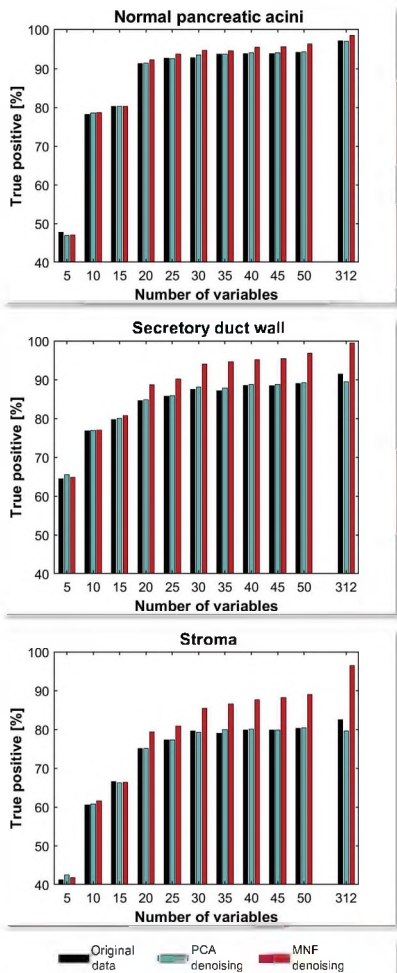


Fig. 4. Comparison of classification results dependence on spectral variables number for original dataset, PCA and MNF. Results shown for denoising methods are maximum values obtained for each series presented in Fig. 3.

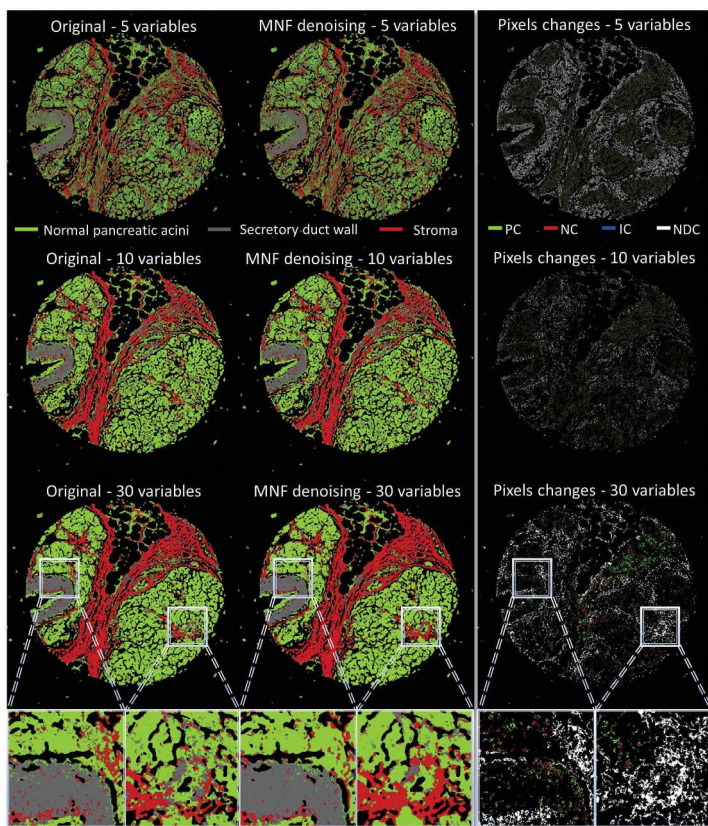


Fig. 5. Prediction comparison for a different number of variables taken into classification, for original and MNF denoised data (6, 8 and 8 bands used for reconstruction for 5, 10 and 30 variables, respectively). In the third column pixel class assignment changes between original and MNF denoised images were calculated by comparison with ground truth used to create the model. Positive Changes (PC – green), from any incorrect class to a correct one. Negative Changes (NC – red), from correct class to any incorrect. Irrelevant Changes (IC – blue), from any incorrect class to other incorrect class and Non Determined Changes (NDC – white), changes in regions without ground truth. To better underline advantages coming from denoising, zoomed areas are shown for 30 variables case. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

specimen. A potential remedy for this issue is the use of QCL IR microscopy performed in Discrete Frequency mode for maximum speed gain. This gain could be furthermore improved by applying a proper denoising algorithm that takes into account the specific data structure. The results suggest that the most important factor in obtaining a reasonable classification accuracy (which was reached at 30 frequencies for pancreas histology using a Spero QT QCL microscope) is the number of experimental spectral variables. Furthermore, out of all tested spatial and spectral denoising algorithms only MNF has significantly improved classification results, but only when more than 20 variables were denoised and used for

Table 1

Percentages of pixels that changed after MNF denoising corresponding to Fig. 5 right panel. Total change indicates pixels that changed compared to all pixels in the tissue.

Variables	5	10	30
Positive Change [%]	44.0	45.9	60.2
Negative Change [%]	43.9	46.7	36.3
Irrelevant Change [%]	12.1	7.4	3.5
Total Change [%]	37.8	16.9	16.7

classification. This work is the first to show the effects of denoising

on classification accuracy of QCL data which are likely to be transferable to other QCL microscopes and other modalities using discrete frequency imaging, e.g. AFM-IR or CARS/SRS imaging.

Acknowledgements

This research was supported from the “Pancreatic cancer comprehensive histopathology based on IR chemical imaging” project, which is carried out within the Homing programme (grant no. Homing/2016-2/20) of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund. The authors declare no conflicts of interest.

References

- [1] M.J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H.J. Butler, K.M. Dorling, et al., Using Fourier transform IR spectroscopy to analyze biological materials, *Nat. Protoc.* 9 (2014) 1771–1791, <https://doi.org/10.1038/nprot.2014.110>.
- [2] K.M. Marzec, T.P. Wrobel, A. Rygiel, E. Maslak, A. Jaszał, A. Fedorowicz, et al., Visualization of the biochemical markers of atherosclerotic plaque with the use of Raman, IR and AFM, *J. Biophot.* 7 (2014), <https://doi.org/10.1002/jbio.201400014>.
- [3] E. Wiercigroch, E. Staniszevska-Slezak, K. Szkaradek, T. Wojcik, Y. Ozaki, M. Baranska, et al., FT-IR spectroscopic imaging of endothelial cells response to tumor necrosis factor- α to follow markers of inflammation using standard and high-magnification resolution, *Anal. Chem.* 90 (2018) 3727–3736, <https://doi.org/10.1021/acs.analchem.7b03089>.
- [4] T.P. Wrobel, K.M. Marzec, S. Chlopicki, E. Maslak, A. Jaszał, Effects of Low Carbohydrate High Protein (LCHP) diet on atherosclerotic plaque phenotype in ApoE/LDLR $-/-$ mice: FT-IR and Raman imaging, *Sci. Rep.* (2015) 1–9, <https://doi.org/10.1038/srep14002>.
- [5] T.P. Wrobel, R. Bhargava, Infrared spectroscopic imaging advances as an analytical technology for biomedical sciences, *Anal. Chem.* 90 (2018) 1444–1463, <https://doi.org/10.1021/acs.analchem.7b05330>.
- [6] T.P. Wrobel, M.R. Kole, R. Bhargava, Emerging Trends and Opportunities in Discrete-Frequency Infrared and Raman Spectroscopic Imaging, *Spectrosc. Santa Monica*, 2016, p. 31.
- [7] M. Hermes, R.B. Morrish, L. Huot, L. Meng, S. Junaid, J. Tomko, et al., Mid-IR hyperspectral imaging for label-free histopathology and cytology, *J. Optic.* 20 (2018), <https://doi.org/10.1088/2049-8986/aa2365>.
- [8] Y. Yao, A.J. Hoffman, C. Gnani, Mid-infrared quantum cascade lasers, *Nat. Photon.* 6 (2012) 432–439, <https://doi.org/10.1038/nphoton.2012.143>.
- [9] T.P. Wrobel, P. Mukherjee, R. Bhargava, Rapid visualization of macromolecular orientation by discrete frequency mid-infrared spectroscopic imaging, *Analyst* 142 (2017), <https://doi.org/10.1039/c6an01086e>.
- [10] K. Yeh, S. Kenkel, J.-N. Liu, R. Bhargava, Fast infrared chemical imaging with a quantum cascade laser, *Anal. Chem.* 87 (2015) 485–493, <https://doi.org/10.1021/acs502753>.
- [11] J.A. Kimber, S.G. Kazarian, Spectroscopic imaging of biomaterials and biological systems with FTIR microscopy or with quantum cascade lasers, *Anal. Bioanal. Chem.* 409 (2017) 1–8, <https://doi.org/10.1007/s00216-017-0574-5>.
- [12] N. Kröger-Lui, N. Gretz, K. Haase, B. Kränzlin, S. Neudecker, A. Pucci, et al., Rapid identification of goblet cells in unstained colon thin sections by means of quantum cascade laser-based infrared microspectroscopy, *Analyst* 140 (2015) 2086–2092, <https://doi.org/10.1039/c4an02001d>.
- [13] K. Haase, N. Kröger-Lui, A. Pucci, A. Schönnhals, W. Petrich, Real-time mid-infrared imaging of living microorganisms, *J. Biophot.* 9 (2016) 61–66, <https://doi.org/10.1002/jbio.201500264>.
- [14] N. Kröger, A. Egl, M. Engel, N. Gretz, K. Haase, I. Herpich, et al., Quantum cascade laser-based hyperspectral imaging of biological tissue, *J. Biomed. Optic.* 19 (2014) 111607, <https://doi.org/10.1117/1.JBO.19.11.111607>.
- [15] A.V. Ewing, S.G. Kazarian, Infrared spectroscopy and spectroscopic imaging in forensic science, *Analyst* 142 (2017) 257–272, <https://doi.org/10.1039/c6an02244h>.
- [16] R.M. DeVetter, S. Kenkel, S. Mittal, R. Bhargava, T.P. Wrobel, Characterization of the structure of low-e substrates and consequences for IR transfection measurements, *Vib. Spectrosc.* 91 (2017) 119–127, <https://doi.org/10.1016/j.vibspec.2016.09.001>.
- [17] C. Kuepper, A. Kallenbach-Thielges, H. Juetter, A. Tannapfel, F. Großerueschkamp, K. Gerwert, Quantum cascade laser-based infrared microscopy for label-free and automated cancer classification in tissue sections, *Sci. Rep.* 8 (2018) 1–10, <https://doi.org/10.1038/s41598-018-26098-w>.
- [18] P. Lasch, Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging, *Chemosens. Intell. Lab. Syst.* 117 (2012) 100–114, <https://doi.org/10.1016/j.chemolab.2012.03.011>.
- [19] P. Kozioł, M.K. Raczowska, J. Skibinska, S. Urbaniak, C. Paluszkievicz, W.M. Kwiatek, et al., Comparison of spectral and spatial denoising techniques in the context of High Definition FT-IR imaging hyperspectral data, *Sci. Rep.* 8 (2018) 1, <https://doi.org/10.1038/s41598-018-32713-7>.
- [20] S. Berisha, S. Chang, S. Sak, D. Daeinejad, Z. He, R. Mankar, et al., Sproc: an open-source biomedical data processing platform for large hyperspectral images, *Analyst* 142 (2017) 1350–1357, <https://doi.org/10.1039/c6an02082h>.
- [21] R.K. Reddy, R. Bhargava, Accurate histopathology from low signal-to-noise ratio spectroscopic imaging data, *Analyst* 135 (2010) 2818–2825, <https://doi.org/10.1039/c0an00350f>.
- [22] A.A. Green, M. Berman, P. Switzer, M.D. Craig, A transformation for ordering multispectral data in terms of image quality with implications for noise removal, *IEEE Trans. Geosci. Rem. Sens.* 26 (1988) 65–74, <https://doi.org/10.1109/36.3001>.
- [23] L.S. Leslie, T.P. Wrobel, D. Mayerich, S. Bindra, R. Emmadi, R. Bhargava, High definition infrared spectroscopic imaging for lymph node histopathology, *PLoS One* 10 (2015), e0127238, <https://doi.org/10.1371/journal.pone.0127238>.
- [24] S. Mittal, T.P. Wrobel, L.S. Leslie, A. Kadajacy-Balla, R. Bhargava, A four class model for digital breast histopathology using high-definition Fourier transform infrared (FT-IR) spectroscopic imaging, *Prog. Biomed. Opt. Imaging - Proc. SPIE* 9791 (2016), <https://doi.org/10.1117/12.2217358>.
- [25] T.P. Wrobel, J.T. Kwak, A. Kadajacy-Balla, R. Bhargava, High-definition Fourier transform infrared spectroscopic imaging of prostate tissue, *Prog. Biomed. Opt. Imaging - Proc. SPIE* 9791 (2016), <https://doi.org/10.1117/12.2217341>.
- [26] I.M. Scott, W. Lin, M. Liakata, J.E. Wood, C.P. Vermeer, D. Allaway, et al., Merits of random forests engine in evaluation of chemometric classifiers by external validation, *Anal. Chim. Acta* 801 (2013) 22–33, <https://doi.org/10.1016/j.jaca.2013.09.027>.
- [27] J.H. Huang, H.L. Xie, J. Yan, H.M. Lu, Q.S. Xu, Y.Z. Liang, Using random forest to classify T-cell epitopes based on amino acid properties and molecular features, *Anal. Chim. Acta* 804 (2013) 70–75, <https://doi.org/10.1016/j.jaca.2013.10.003>.
- [28] D. Pérez-Gualta, J. Kuligowski, B. Lendi, B.R. Wood, G. Quintás, Assessment of discriminant models in infrared imaging using constrained repeated random sampling cross validation, *Anal. Chim. Acta* (2018) 1–9, <https://doi.org/10.1016/j.jaca.2018.05.019>.

Klasyfikacja histopatologiczna tkanek trzustki

Rak trzustki jest jedną z najczęstszych przyczyn śmierci związanych z nowotworami, zarówno wśród kobiet jak i mężczyzn. Wynika to często z bardzo późnej diagnozy, którą dodatkowo utrudnia skomplikowany w analizie obraz histopatologiczny. Dodatkowo, środowisko histopatologów zmagą się z problemem właściwego odróżnienia wczesnego etapu nowotworu od zapalenia, co skutkuje błędnymi diagnozami i nietrafionymi procedurami leczenia. Wynikiem tego jest śmiertelność na poziomie 80% w ciągu pierwszego roku po diagnozie. Tak fatalne statystyki są ogromną motywacją do znalezienia skutecznej metody diagnostyki wspierającej histopatologów.

Wspominany poprzednio projekt Homing miał na celu stworzenie modelu histopatologicznego raka trzustki w oparciu o obrazowanie IR. W początkowej fazie projektu skupiono się na zgromadzeniu odpowiedniej statystyki pacjentów z wykorzystaniem metody FT-IR. Zebrano tym samym dane dla ponad 600 biopsji igłowych, na które składało się ponad 250 pacjentów. Pomiary przeprowadzono zarówno z wykorzystaniem obiektywu 15x, jak również wysokorozdzielczego (HD) obiektywu 36x, zapewniającego maksymalną ilość informacji przestrzennej. Zmierzone ponad 672 000 000 widm FT-IR zawierających kontrast chemiczny pochodzący od różnego typu tkanek trzustki. Po odpowiednim wstępnym przetwarzaniu danych, uwzględniającym zastosowanie zoptymalizowanych pod tym kątem metod odszumiania, przygotowane dane były wykorzystane do stworzenia modelu Lasu Losowego.

Finalne modele oparte o dane 15x i 36x wyróżniały po 6 klas - kontrola, nowotwór, zapalenie, martwica, tkanka włóknista oraz krew [G]. Otrzymane wyniki umożliwiły detekcję nowotworów oraz zapalenia ze skutecznością na poziomie $AUC=95\%$ [G], co odpowiada poziomowi czułości oraz specyficzności wymaganemu w zastosowaniach klinicznych. Ponadto, ukazano możliwość detekcji śródoperacyjnego marginesu chirurgicznego dla tkanek resekcyjnych [G]. Dla tych tkanek zarysowana została wyraźna granica przechodzenia zdrowej tkanki w nowotworową, z widocznymi naciekami zapalnymi pomiędzy nimi [G].

Powyższe badania są pierwszym krokiem ku wprowadzeniu tej metody do praktyki klinicznej. Ich ogromną zaletą jest wykorzystanie informacji o składzie biochemicznym próbki, zawartych w widmach IR, w przeciwieństwie do innych metod bazujących jedynie na obrazach pochodzących z mikroskopu widzialnego. Ponadto, kontrast chemiczny umożliwia postawienie diagnozy bez wykonywania wielu żmudnych i długotrwałych barwień histopatologicznych.

Badania te wchodzą w skład serii publikacji prezentujących możliwości zastosowania obrazowania IR w praktyce klinicznej. Korzystając z kompromisu (zapewnionego przez mikroskopy QCL) pomiędzy szybkimi pomiarami dużych obszarów tkanek, a ograniczoną informacją spektralną, stworzony został również model do szybkiej detekcji marginesu chirurgicznego w oparciu o 14 częstości spektralnych. Ponadto, wytrenowano model stopniujący tkankę trzustki w skali od zdrowej, przez neoplazję (PanIN), do nowotworowej. Oba te zagadnienia są obecnie przygotowywane do publikacji.

Mój wkład w publikację [G] związany był z eksperymentalną częścią zbierania danych FT-IR, zarówno z wykorzystaniem standardowego (15x) jak i wysokorozdzielczego obiektywu (36x), przeprowadzeniem wstępnego przetwarzania danych i obliczeniem metryk będących

baza dla modeli klasyfikacyjnych, implementacją algorytmów do obliczeń AUC oraz pisaniem części manuskryptu.

Comprehensive Histopathology Imaging in Pancreatic Biopsies: High Definition Infrared Imaging with Machine Learning Approach

Danuta Liberda^{1,2#}, Paulina Koziol^{2,3#}, Tomasz P. Wrobel^{2*}

¹ Jagiellonian University, Doctoral School of Exact and Natural Sciences, Prof. St. Łojasiewicza 11, PL30348, Cracow, Poland

² Solaris National Synchrotron Radiation Centre, Jagiellonian University, Czerwone Maki 98, 30-392 Krakow, Poland.

³ Institute of Physics, Jagiellonian University, Łojasiewicza 11, 30-348 Krakow, Poland

*email: tomek.wrobel@uj.edu.pl

equal contribution

Preprint available at: <https://www.medrxiv.org/content/10.1101/2022.07.01.22277130v2>

Abstract

Infrared (IR) based histopathology offers a new paradigm in looking at tissues and can provide a complimentary information source for more classical histopathology, which makes it a noteworthy tool given possible clinical application. This study aims to build a robust, pixel level machine learning model for pancreatic cancer detection using IR imaging. In this article, we report a pancreatic cancer classification model based on data from over 600 biopsies (coming from 250 patients) imaged with IR diffraction-limited spatial resolution. To fully research model's classification ability, we measured tissues using two optical setups, resulting in Standard and High Definitions data. This forms one of the largest IR datasets analyzed up to now, with almost 700 million spectra of different tissue types. The first six-class model created for comprehensive histopathology achieved pixel (tissue) level AUC values above 0.95, giving a successful technique for digital staining with biochemical information extracted from IR spectra.

Keywords

Pancreatic cancer, infrared imaging, machine learning, high definition

Introduction

Pancreatic ductal adenocarcinoma (PDAC), the most common pancreatic cancer type, accounts for less than 3% of all cancers, but it remains the third leading cause of cancer-related deaths, both for men and women in the United States with 53,000 estimated new cases and an 80% death rate in one year [1]. Furthermore, PDAC is projected to become the second cause of cancer-related death by 2030 [2]. The dreadful prognosis of patients with this disease, including a less than 8% 5-year survival after diagnosis, is due to the lack of early symptoms and/or specific biomarkers for early diagnosis and the paucity of available chemotherapy. In PDAC, three types of biomarkers are desirable: those that help in the detection of the disease onset (diagnosis biomarkers); those that predict responses to treatments (predictive biomarkers) and those that forecast the likely course of the disease, including survival and recurrence patterns (prognosis biomarkers).

Infrared (IR) based histopathology offers a new paradigm in looking at tissues and can provide a complementary information source for more classical histopathology as being label-free and non-destructive (1–4). Especially, correct understanding of the involved inflammation, fibrosis and neoplasia is required for a method to be successful. IR histopathology is being developed for an increasing number of cancer types, e.g. breast (5), colon (6), lung (7), brain (8) prostate (9), esophagus (10) - highlighting the increasing potential of the method. The technological developments, especially on the speed and spatial resolution front, make this approach feasible in the clinic (11). The key to the optimization of IR measurements and reducing analysis time is the spectral information reduction: whether it is selection of specific spectroscopic bands (12) or significant reduction of spectral resolution (13). Both conditions can be met using Quantum Cascade Lasers (QCL) based microscopes. It has been reported that a QCL data collection needed for classification of a centimeter-scale size biopsies could be done in the order of minutes (13,14) with a commercially available Spero-QT (Daylight Solutions, CA, USA) microscope. However, to optimize a method for a fast screening or imaging modality, a proper foundation needs to be laid beforehand even after initial proof of concept studies such as this one (15,16).

Due to the Rayleigh criterion, spatial resolution in IR imaging depends on the wavelength and objective numerical aperture (NA). In this study we applied two optical setups:

- 15x objective with 0.4 NA, giving a projected pixel size of 2.7 μm ,
- 36x objective with 0.5 NA, giving a projected pixel size of 1.1 μm .

Considering under-sampling in high wavenumber region (above 2314 cm^{-1}) for applied 15x objective, data measured with this optical setup are termed Standard Definition (SD). In the case of the 36x objective, spatial sampling criteria are met and measurements are Rayleigh-limited in the whole measured spectral range (3850-900 cm^{-1}), therefore, data has High Definition (HD) label. In our previous research (17) we showed that HD data reveal small tissue components like thin fibers or blood cells squeezed in the vessels. However, SD gives more homogenous predictions within tissue classes while also being faster than HD measurements.

In this study, we applied both mentioned definitions to check their influence on classification on the pixel (tissue) and patient level. It is especially interesting in the case of inflammations that are composed of lymphocytes (size of 7-8 μm). Reported results for pancreatic cancer histopathology were based on FT-IR imaging and machine learning, providing a classification model created with data from over 600 biopsies (coming from 250 patients). Data collection was done with IR diffraction-limited spatial resolution, forming one of the largest IR datasets analyzed up to now, with almost 700 million spectra of different tissue types.

Results and discussion

Patient pathology recognition methodology

This research was performed using Tissue Micro Arrays (TMAs), which are typically applied in studies on tissue composition and differentiation. This is caused by the high variability of TMAs due to many biopsies that are taken from patients differentiated by age, sex, pathology diagnosis, grade, etc.. After the typical procedure of needle biopsies collection, to prevent their damage they are embedded with paraffin and cut into thin 5 μm slices (Figure 1 A). In this

research, we used two neighboring slices. Pathology recognition needs to be done based on histopathological annotations performed by an experienced histopathologist (Figure 1B). Therefore, the first TMA slice is H&E (Hematoxylin and Eosin) stained and imaged with a visible light microscope. The second slice is imaged with an IR spectrometer. The result of this measurement is data in the form of a cube containing IR images in the X and Y plane, and spectra in the Z direction (Figure 1C). Based on histopathological annotations mentioned before, corresponding regions are marked on the IR image – spectra (after preprocessing) that come from this area are used for final histopathological model creation. This model is later applied to digitally describe histopathological regions in IR image and such process is called prediction. Patient pathology is finally recognized using Receiver Operating Characteristic (ROC) curves, based on the number of pixels (from a chosen class/classes) present in predicted images (Cancer class in the Patient pathology classification panel in Figure 1D). Model prediction power is evaluated using Area Under the Curve (AUC).

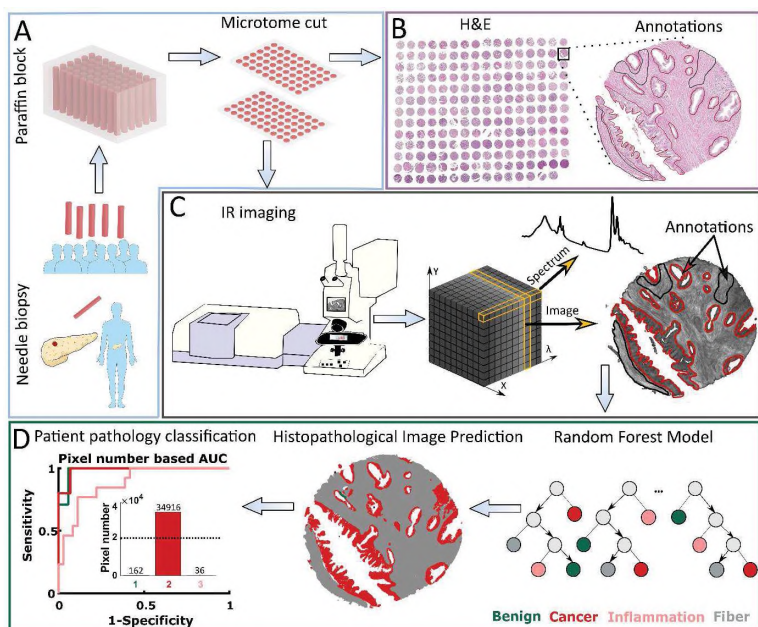


Figure 1. Patient pathology recognition pipeline: A) Tissue Micro Array preparation, B) Histopathological annotation on H&E stained biopsies images, C) Biopsies' Infrared Imaging and obtained data structure, D) Model creation and patient pathology recognition.

Comprehensive histopathological model

The FT-IR histopathological Random Forest model was based on 6 TMAs (more information in Supplementary Materials Table S1). Our Standard Definition based model can differentiate six classes of pancreatic tissue: Benign, Cancer, Necrosis, Inflammation, Fiber, and Blood.

Differences between analyzed classes are to some extent reflected in their mean spectra (Figure 2B). At the first glance, there are visible differences in shape and relative intensity of Amide A and B ($3550 - 3000 \text{ cm}^{-1}$) protein bands and DNA/RNA bands ($1300 - 1000 \text{ cm}^{-1}$). Patients in TMAs were divided into model, validation, and test sets (Supplementary Materials Table 3) for model optimization and prediction power evaluation. In Figure 2A predicted TMAs (for Standard Definition) images with zooms on single cores and regions of interest are presented. H&E stained images with zooms are presented in Supplementary Materials (S2-S3). Benign biopsies with green color can be easily differentiated from remaining cases in all TMAs. Class benign can be described as a combination of lobules and islands annotations, coming from normal pancreas tissue (for example single core from PA1002a TMA presented in Figure 2A), and normal adjacent to cancer tissue. The microenvironment of cancer is stroma-rich, which is visible among others in the single core from PA2072a TMA presented in Figure 2A. In both presented cancerous cores in the upper part of Figure 2A there can be also found necrotic regions associated with cancer changed biopsy tissue. Frequently, we observed lymphocyte clusters like in the single biopsy from PA961e TMA presented in Figure 2A, infiltrating into the cancerous tissue surrounded by stroma. Most chronic pancreatitis cases are visible in BBS14011 TMA. In the single biopsy predicted image, one may observe both: lymphocyte infiltration into the tissue and inflamed changed lobules. In the single core from BIC14011a TMA in Figure 2A, there is visible very well classified aggregation of erythrocytes in the affected stroma. The classification results on the pixel level (tissue level) were quantified using AUC metric in Figure 2C. All classes reached AUC values higher than 0.95, which is a very good result, especially taking into account that AUC value equal to one corresponds to the situation where each pixel was correctly classified.

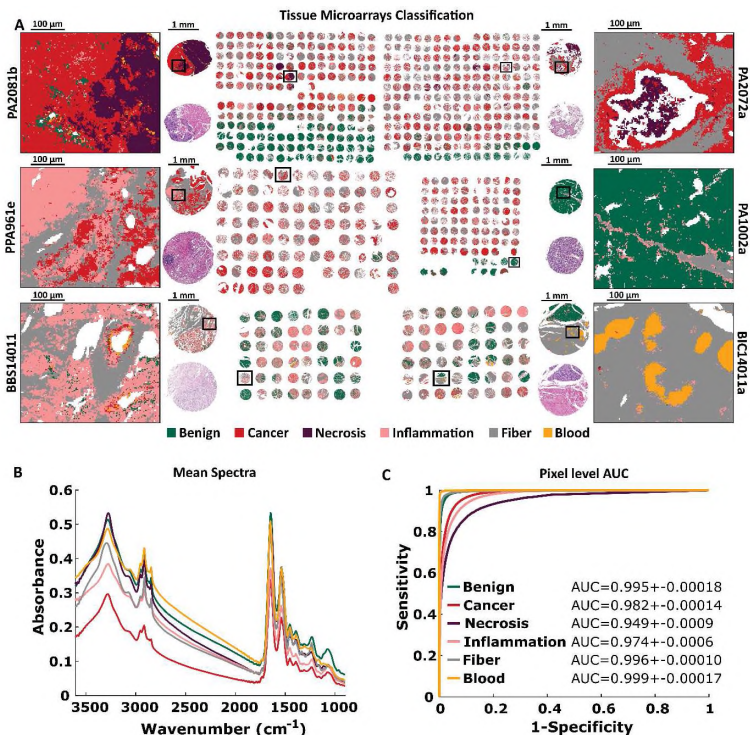


Figure 2. Standard Definition based histopathological recognition of six pancreatic tissue classes with Random Forest classification for: A) Tissue Micro Arrays with zooms on single cores with chosen regions. B) Tissue Micro Arrays raw mean spectra, and C) Pixel level AUC values of recognized tissue classes.

Finally, we used intraoperative biopsy (taken in clinical conditions) as the external sample for Standard Definition model assessment (Figure 3A). The most valuable result achieved for this sample is a very well classified transition from benign to cancerous state with inflammation infiltration. Zooms on interesting intraoperative biopsy regions are also presented (Figure 3B-D). In the predicted image in Figure 3 intraoperative margin can be easily marked.

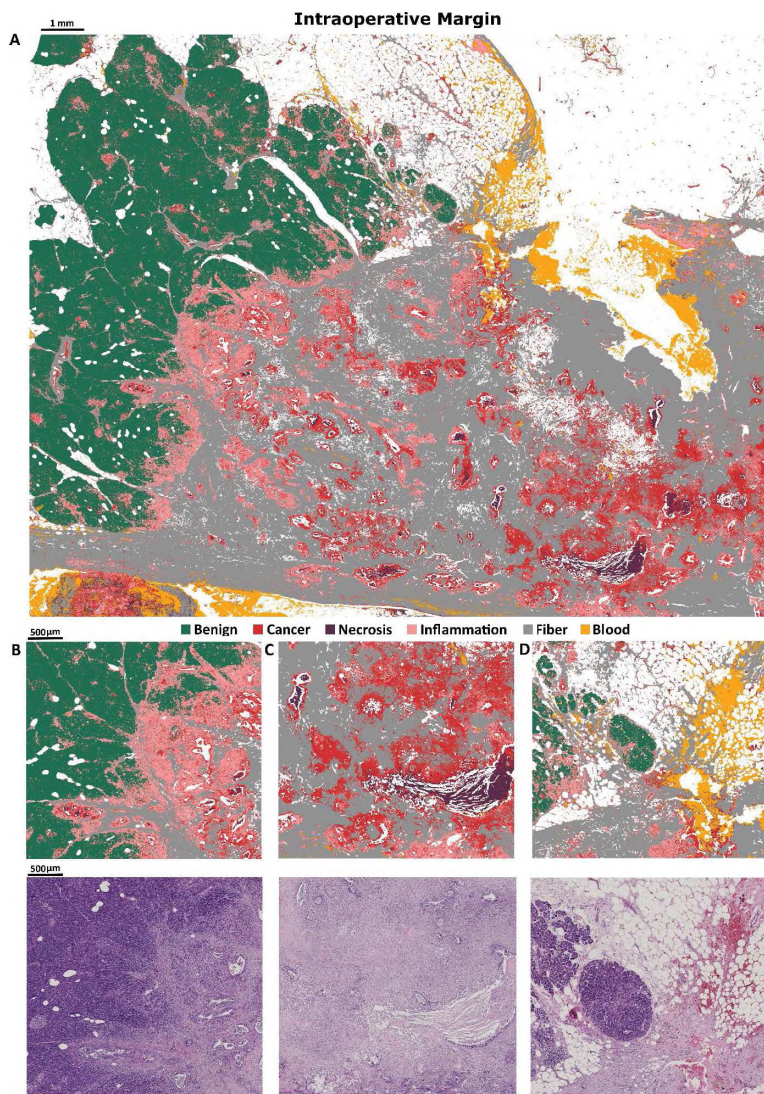


Figure 3. Predicted image of intraoperative biopsy A), with zooms on: B) margin between benign and cancerous tissue, with clearly visible inflammation between, C) big necrosis region

surrounded by cancerous changed tissue, D) benign tissue with small inflammation and blood infiltration.

High definition histopathological model

A similar model based on the same histopathological annotations and TMAs was built for High Definition data. The projected pixel size was $1.1 \mu\text{m}$ ensuring sufficient spatial sampling (for 0.5 NA objective) to retrieve the maximum of spatial information. In comparison to Standard Definition, this model unraveled small cells (i.e. lymphocytes) located between tissue. In general, predicted classes are not as consistent as in Standard Definition, there are cases with pixels that are highly mixed (Figure 4A). This is reflected in AUC values – there are three classes with results below 0.95 (Figure 4C). The mean spectra (and differences between them) for the HD data set (Figure 3B) are similar to those calculated for the SD data set in Figure 2B.

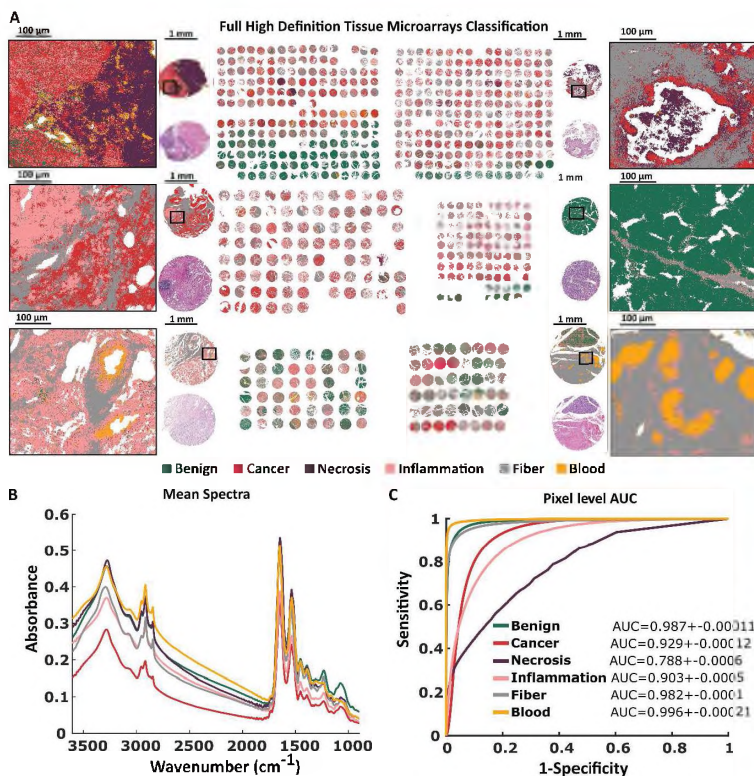


Figure 4. High Definition based histopathological recognition of 6 pancreatic tissue classes with Random Forest classification for A) Tissue Micro Arrays with zooms on single cores with

chosen regions. B) Tissue Micro Arrays raw mean spectra, and C) Pixel level AUC values of recognized tissue classes.

We also calculated AUC metrics on the core and patient level for test set (49 patients) giving high AUC results in the range 0.99 (Figure 5) regardless of the definition.

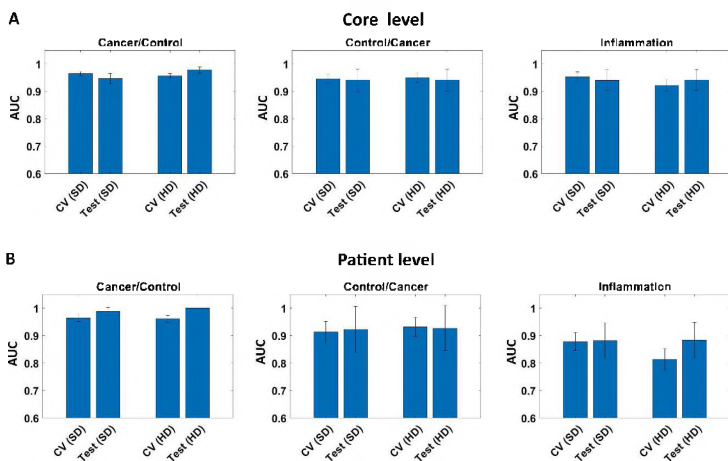


Figure 5. AUC for model (cross validation set - CV) and test set on: A) Core level, and B) Patient level for Standard (SD), and High Definitions (HD).

The Standard and High Definition images seem to provide a very similar contrast for this 6 class models. Similar results have been found for other pathologies (18) which means that for most cases High Definition is not required for a robust model. The increased noise level actually makes IID models slightly less stable, while still having very high performance.

Conclusions

The infrared imaging combined with the machine learning algorithm was applied to develop a comprehensive model predicting pancreatic histopathology. In the first step FT-IR Random Forest model was built based on a broad and reach in biochemical information spectra. This model has enabled recognition of six tissue classes: Benign, Cancer, Necrosis, Inflammation, Fiber, and Blood for both Standard and High Definition. In predictions based on High Definition data it was possible to differentiate inflammatory cells in more detail, however, predicted images were of better quality for Standard Definition. Pathology recognition that can be performed based on this model allows differentiation of cancerous, benign, and chronic pancreatitis cases with sensitivity higher than 95%. Moreover, in the event of a nonobvious case, an image of a predicted biopsy can be analyzed further to diagnose a patient finding specific regions of affected tissue, i.e. immune cells and erythrocyte infiltration, fibrosis, changes in a

number of characteristic tissue structures. Model for Standard Definition achieved similar AUC values for cancerous patient detection to High Definition – that is positive in the context of research continuation using fast QCL system (that is equipped with objective giving projected pixel size down to 1.3 μm). This study opens the way for more subtle pathology recognition in pancreatic tissues.

Methods

All calculations described in this study were done in MATLAB software.

Biological material

Human pancreatic tissue was used in this study, with sample set consisting of six TMAs: PA2081b, PA2072a, PA1002a, PA961e, BBS14011, BIC14011a purchased from Biomax Inc. and two surgical resections. Since biopsy collection and personal data anonymization was done by Biomax Inc. no written consent from patients was requested. TMAs are assembled from tissue cores coming from needle biopsies with diameters between 1-1.5 mm (depending on the TMA). Resection material was specifically selected to present border between cancer and control tissue (surgical margin), therefore sample sizes are roughly 16x35 mm and 16x13 mm. Samples were placed in a paraffin block and cut using microtome for 5 μm thick slices, with one slice placed on IR transparent BaF₂ salt plate for transmission measurement, and consecutive slice mounted on glass for Hematoxylin and Eosin (H&E) staining allowing histopathological annotations. Initially paraffin covered IR samples were deparaffinized with 24 hours hexan bath to avoid paraffin absorption around 1462 cm^{-1} .

High patient statistic was achieved by the use of TMAs with 663 tissue cores in total. However, in most cases, more than one tissue core came from a single patient. Additionally, overlapping cases existed between some TMAs, thus, final patient cohort comprised of 250 cases (with two additional patients represented by surgical resection). Ethical approval was granted by Ethics committee at Jagiellonian University in Krakow (no. 1072.6120.304.2020).

Fourier Transform Infrared Imaging

Bruker Vertex 70v spectrometer coupled with Hyperion 3000 microscope and 64x64 FPA MCT detector was used for FT-IR measurements. Spectra were measured within 3850-900 cm^{-1} range, with 8 cm^{-1} spectral resolution and zero filling factor of 2, giving 765 spectral points. Signal was co-averaged 4 times for a sample region and 64 times for background collection. Two types of objectives were used during measurements – 15x (SD) and 36x (HD), giving projected pixel sizes of 2.7 μm and 1.1 μm , respectively. All TMAs were measured with both resolutions, but surgical resections samples were only measured using 15x objective. Total number of sample spectra collected with FT-IR imaging in this study reached 672 407 632 (including both resolutions). Additional information on experimental time optimization is provided in Supplementary Materials.

The first step of preprocessing was data denoising using Minimum Noise Fraction (MNF) method with 20 bands used for reconstruction (19,20). Due to partially spatial character of MNF's noise estimation and RAM memory constraints, TMAs processing was done in a core-by-core manner. In case of resection material, denoising was done for region of 2x2 measurement tiles. Such approach allows to prevent information leakage between patients, which could cause over optimistic results in further classification. For the classification

purpose, FT-IR spectra were transferred into so-called metrics. Metrics provide spectral information extracted from selected IR bands defined by experienced spectroscopist (list of spectral regions presented in Table S2 in Supplementary Materials). Metrics calculation for each spectral region includes local linear baseline removal followed by calculation of band integration, maximum value and center of gravity. Finally, those three characteristics are normalized to the corresponding value of Amide I. Total number of metrics extracted from single spectrum is 123.

Random Forest Classification

Classification model creation requires multiple steps, starting from tissue types annotation and Region of Interests (ROIs) definition for chosen classes, which are required to extract data later used to feed the classifier. Annotations of Infrared images for all TMAs were done utilizing ground truth annotations provided by an experienced Histopathologist, based on visual inspection of H&E images. Such annotations were later transformed into ROI masks, allowing correct pixels (metrics/spectra) selection. Furthermore, to assure classification using only tissue originating pixels, a tissue mask was created after thresholding absorbance for 1650 cm^{-1} (most intensive band). Such a mask was used to remove background spectra from ROI masks and to select samples for final prediction images.

To prevent from Random Forest classifier being biased, all classes had the same number of samples (pixels) during a particular model training process. This number was determined by the number of samples available for the smallest class. Moreover, to ensure training data variability, equal number of samples from each TMA was taken, unless the number of available samples was insufficient. In such case, lacking samples were evenly topped up from other TMAs. In all cases, samples were always randomly selected.

FT-IR based six-class models (SD and HD) were built with 50 Trees, which was sufficient to prevent overtraining, but at the same time did not unnecessarily complicate the model. A crucial factor for stable model creation is its proper validation, assuring its stability and lack of overtraining. For FT-IR based models, patients were divided according to their annotations (benign, inflammation, cancer) in a random manner into model, validation and test sets. Subset of patients from each TMA was assigned to a test set, keeping in mind overlapping cases between TMAs. Number of patients in validation and test sets are presented in Supplementary Materials Table S3. A four-fold cross validation (using model data) was used during the model optimization, which is a long and iterative process. To assess models' performance during the cross validation, confusion matrix along with Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) were calculated. Finally, when created model reached satisfactory parameters, a test set was used for its final evaluation with ROC and AUC defining final diagnostic ability. Described validation approach was applied to both, SD and HD datasets. Nonetheless, adopted validation method is strongly dependent from available patient cohort. Additional information on models validation is provided in Supplementary Materials.

Author Contributions

Study design: T.P.W., P.K., D.L.. Development of methodology: T.P.W., P.K., D.L.. Acquisition of spectroscopic data: T.P.W., P.K., D.L.. Data processing and analysis: P.K., D.L.. Writing the manuscript D.L., P.K., T.P.W..

Acknowledgements

Authors wanted to acknowledge Slawka Urbaniak for histopathological annotations. This research was supported by „Pancreatic cancer comprehensive histopathology based on IR chemical imaging” project, which was carried out within the Homing programme (grant no. Homing/2016-2/20) of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund. This research was performed using equipment purchased in the frame of the project co-funded by the Malopolska Regional Operational Program Measure 5.1 Krakow Metropolitan Area as an important hub of the European Research Area for 2007-2013, project no. MRPO.05.01.00-12-013/15. The authors acknowledge the support from National Synchrotron Radiation Centre SOLARIS, financed by the project of Polish Ministry of Education and Science under contract no. 1/SOL/2021/2. The Figure 1 and Graphical Abstract were partly generated using pancreas (modified with dot) and blue human (with added pancreas, and to generate crowd) images - using Servier Medical Art, provided by Servier, licensed under a Creative Commons Attribution 3.0 unported license (21).

Competing Interests

The Authors declare no Competing Financial or Non-Financial Interests.

Data Availability

Data available on a reasonable request from the authors.

References

1. Wrobel TP, Bhargava R. Infrared Spectroscopic Imaging Advances as an Analytical Technology for Biomedical Sciences. *Anal Chem.* 2018;90(3):1444–63.
2. Baker MJ, Byrne HJ, Chalmers J, Gardner P, Goodacre R, Henderson A, et al. Clinical applications of infrared and Raman spectroscopy: State of play and future challenges. *Analyst.* 2018;143(8):1735–57.
3. Fernandez DC, Bhargava R, Hewitt SM, Levin IW. Infrared spectroscopic imaging for histopathologic recognition. *Nat Biotechnol.* 2005 Apr;23(4):469–74.
4. Levin IW, Bhargava R. Fourier Transform Infrared Vibrational Spectroscopic Imaging: Integrating Microscopy and Molecular Recognition*. *Annu Rev Phys Chem.* 2004;56(1):429–74.
5. Mittal S, Wrobel TP, Walsh M, Kajdacsy-Balla A, Bhargava R. Breast cancer histopathology using infrared spectroscopic imaging: The impact of instrumental configurations. *Clin Spectrosc.* 2021;3(February):100006.
6. Tiwari S, Kajdacsy-Balla A, Whiteley J, Cheng G, Hewitt SM, Bhargava R. INFORM: INFRared-based ORGANIZational Measurements of tumor and its microenvironment to predict patient survival. *Sci Adv.* 2021;7(6).
7. Großerueschkamp F, Kallenbach-Thieltges A, Behrens T, Brüning T, Altmayer M, Stamatis G, et al. Marker-free automated histopathological annotation of lung tumour subtypes by FTIR imaging. *Analyst.* 2015;140(7):2114–20.
8. Hands JR, Abel P, Ashton K, Dawson T, Davis C, Lea RW, et al. Investigating the rapid diagnosis of gliomas from serum samples using infrared spectroscopy and cytokine and angiogenesis factors. *Anal Bioanal Chem.* 2013;405(23):7347–55.

9. Kwak JT, Kajdacsy-Balla A, Macias V, Walsh M, Sinha S, Bhargava R. Improving prediction of prostate cancer recurrence using chemical imaging. *Sci Rep.* 2015;5:1–10.
10. Liberda D, Hermes M, Koziol P, Stone N, Wrobel TP. Translation of an esophagus histopathological FT-IR imaging model to a fast quantum cascade laser modality. *J Biophotonics.* 2020;13(8):1–9.
11. Schnell M, Mittal S, Falahkheirkhah K, Mittal A, Yeh K, Kenkel S, et al. All-digital histopathology by infrared-optical hybrid microscopy. *Proc Natl Acad Sci U S A.* 2020;117(7):3388–96.
12. Mittal S, Yeh K, Suzanne Leslie L, Kenkel S, Kajdacsy-Balla A, Bhargava R. Simultaneous cancer and tumor microenvironment subtyping using confocal infrared microscopy for all-digital molecular histopathology. *Proc Natl Acad Sci U S A.* 2018;115(25):E5651–60.
13. Kuepper C, Kallenbach-Thieltges A, Juette H, Tannapfel A, Großesuechskamp F, Gerwert K. Quantum Cascade Laser-Based Infrared Microscopy for Label-Free and Automated Cancer Classification in Tissue Sections. *Sci Rep.* 2018;8(1):1–10.
14. Goertzen N, Pappesch R, Fassunke J, Brüning T, Ko YD, Schmidt J, et al. Quantum Cascade Laser-Based Infrared Imaging as a Label-Free and Automated Approach to Determine Mutations in Lung Adenocarcinoma. *Am J Pathol.* 2021;191(7):1269–80.
15. Szymoński K, Lipiec E, Sofińska K, Skirlińska-Nosek K, Milian-Ciesielska K, Szpor J, et al. Spectroscopic screening of pancreatic cancer. *Clin Spectrosc.* 2021;3:100016.
16. Notarstefano V, Sabbatini S, Conti C, Pisani M, Astolfi P, Pro C, et al. Investigation of human pancreatic cancer tissues by Fourier Transform Infrared Hyperspectral Imaging. *J Biophotonics.* 2020;13(4):1–10.
17. Liberda D, Kosowska K, Koziol P, Wrobel TP. Spatial sampling effect on data structure and Random Forest classification of tissue types in High Definition and Standard Definition FT-IR imaging. *Chemom Intell Lab Syst.* 2021;217(April):104407.
18. Mittal S, Wrobel TP, Walsh M, Kajdacsy-Balla A, Bhargava R. Breast cancer histopathology using infrared spectroscopic imaging: The impact of instrumental configurations. *Clin Spectrosc.* 2021;3(July 2020):100006. Available from: <https://doi.org/10.1016/j.clispe.2021.100006>
19. Raczowska MK, Koziol P, Urbaniak S, Paluszkievicz C, Kwiatek WM, Wrobel TP. Influence of denoising on classification results in the context of hyperspectral data : High Definition FT-IR imaging. *Anal Chim Acta.* 2019;1085:39–47.
20. Koziol P, Raczowska MK, Skibinska J, Urbaniak-Wasik S, Paluszkievicz C, Kwiatek W, et al. Comparison of spectral and spatial denoising techniques in the context of High Definition FT-IR imaging hyperspectral data. *Sci Rep.* 2018;8(1):1–11.
21. SMART - Servier Medical ART, URL <https://smart.servier.com/> (accessed 20.02.2022).

Wyznaczanie orientacji makromolekuł w oparciu o spolaryzowaną spektroskopię IR

Praca na modelami histopatologicznymi umożliwiającymi detekcję nowotworów trzustki ujawniła pewną cechę tkanki nowotworowej – zaobserwowano, że mikro-środowisko raka trzustki jest bogate w otaczającą go tkankę włóknistą zbudowaną z kolagenu. Podobna sytuacja występuje również w przypadku innych nowotworów, np. piersi. W związku z tym wysnutych zostało szereg teorii związanych z charakterystyką tkanki włóknistej. Podejrzewa się, że może ona determinować wiele aspektów związanych z chorobą, między innymi może zostać wykorzystana jako czynnik prognostyczny, może determinować skuteczność penetracji leków (chemioterapii) czy może być wykorzystana jako czynnik pozwalający określić zaawansowanie choroby. Szczególny wpływ na te elementy może mieć orientacja włókien oraz ich poziom uporządkowania. Z tego powodu dalsze prace skupione były głównie na badaniach związanych z orientacją makromolekuł, z wykorzystaniem spolaryzowanego promieniowania podczerwonego, ze szczególną uwagą skierowaną na tkankę włóknistą bogatą w kolagen. Była to część grantu Sonata pt. „Trójwymiarowa orientacja makromolekuł przy użyciu obrazowania w podczerwieni (IR) oraz jej znaczenie w mikrośrodowisku raka” dr hab. Tomasza P. Wróbla. Wykorzystanie polaryzacji przełamuje tutaj pewną barierę postawioną przez brak informacji o organizacji próbek.

Prace wykorzystujące cztery lub więcej polaryzacji w celu wyznaczenia orientacji molekuł w płaszczyźnie (2D) były do tej pory skupione głównie na próbkach polimerów posiadających dobrze opisaną w literaturze orientację. Nie były jednak badane tą metodą próbki bardziej skomplikowanych systemów, jakimi dla przykładu są tkanki. Praca [E] była pierwszą wykorzystującą spektroskopię FT-IR, w połączeniu z metodą czterech polaryzacji, do wyznaczenia orientacji 2D molekuł dla bogatych w kolagen tkanek trzustki. Ponadto, wykorzystując funkcję Hermansa wyznaczony został poziom uporządkowania włókien [E]. Otrzymane wyniki były zgodne z przewidywaniami teoretycznymi opartymi na budowie molekuł kolagenu (bogatych w wiązania Amidowe). Zaobserwowana orientacja molekuł, zdeterminowana kierunkiem dipolowego momentu przejścia dla Amidu II oraz Amidu III, była równoległa do kierunku włókien widocznych na obrazach histopatologicznych [E]. Odnotowano też charakterystyczne, w przybliżeniu prostopadłe, ułożenie Amidu I w stosunku do głównej osi włókien [E]. Również funkcja Hermansa wskazywała na wysokie uporządkowanie w rejonie tkanki włóknistej, oraz przypadkową orientację wiązań w obszarach komórek nabłonkowych.

Tak obiecujące wyniki orientacji 2D stały się jeszcze większą motywacją do rozszerzenia tej metody o możliwość wyznaczania orientacji 3D. Metoda wykorzystująca FT-IR do wyznaczania orientacji molekuł w przestrzeni była jednak do tej pory zdefiniowana jedynie teoretycznie, bez udokumentowanego zastosowania do danych eksperymentalnych. W związku z tym, pierwsza aplikacja tej metody wymagała wykorzystania dobrze poznanego układu. Sferulity PCL, które w związku z długimi łańcuchami i ich dobrze scharakteryzowanym w literaturze ułożeniem, okazały się wdzięcznym modelem do badań nad orientacją. Uzyskane wyniki orientacji w przestrzeni były w pełni zgodne z przewidywanym uporządkowaniem polimeru. Zaobserwowano prostopadłe ułożenie łańcuchów do kierunku fibryli, co na obrazach orientacji utworzyło formę współśrodkowych okręgów wokół centrum nukleacji sferulitu [F]. Ponadto, wiązania C=O były ułożone prostopadłe do płaszczyzny wzrostu sferulitu, co było

niemożliwe do zaobserwowania w przypadku metody 2D [F]. Sukcesem prowadzonych badań było również rozszerzenie metody na dane O-PTIR, poprzez odpowiednie zmodyfikowanie algorytmów wykorzystywanych do obliczeń. Dane O-PTIR pokrywają się z wynikami FT-IR, a ze względu na swoją sub-mikronową rozdzielczość przestrzenną umożliwiły obserwację dodatkowych struktur, szczególnie tych związanych z kształtem fibryli [F].

Ogromną zaletą powyższych metod, a szczególnie 3D, jest ich niedestrukcyjny oraz nietomograficzny charakter. Nie wymagają one manipulacji próbką ani padającą iluminacją. Jedynym zmiennym elementem jest kierunek padającej polaryzacji, który może być kontrolowany przez prosty element optyczny (polaryzator czy półfalówkę) wsunięty w wiązkę. Kolejne badania będą obejmowały zastosowanie metody 3D do próbek tkanek oraz wykorzystanie informacji o orientacji i poziomie uporządkowania w dalszych pracach, np. w celu poprawy skuteczności klasyfikacji histopatologicznej.

Mój wkład w publikacje:

- [E] związany był z eksperymentalną częścią zbierania danych FT-IR (z uwzględnieniem czterech przypadków polaryzacji), aplikacją algorytmu wyznaczania orientacji w płaszczyźnie (2D), pisanem skryptów do analizy danych oraz pisanem części manuskryptu;

- [F] związany był z eksperymentalną częścią zbierania danych FT-IR (z uwzględnieniem czterech przypadków polaryzacji), udziałem w zbieraniu danych O-PTIR, aplikacją algorytmu wyznaczania orientacji w przestrzeni (3D), pisanem skryptów do analizy danych oraz pisanem części manuskryptu.

Macromolecular Orientation in Biological Tissues Using a Four-Polarization Method in FT-IR Imaging

Paulina Koziol, Danuta Liberda, Wojciech M. Kwiatak, and Tomasz P. Wrobel*



Cite This: *Anal. Chem.* 2020, 92, 13313–13318



Read Online

ACCESS |

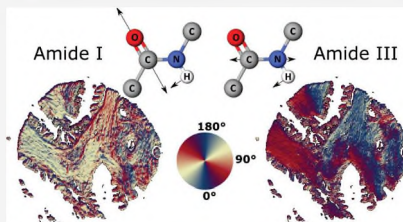


Metrics & More



Article Recommendations

ABSTRACT: Fourier transform infrared spectroscopy has emerged as a powerful tool for tissue specimen investigation. Its nondestructive and label-free character enables direct determination of biochemical composition of samples. Furthermore, the introduction of polarization enriches this technique by the possibility of molecular orientation study apart from purely quantitative analysis. Most of the molecular orientation studies focused on polymer samples with a well-defined molecular axis. Here, a four-polarization approach for Herman's in-plane orientation function and azimuthal angle determination was applied to a human tissue sample investigation for the first time. Attention was focused on fibrous tissues rich in collagen because of their cylindrical shape and established amide bond vibrations. Despite the fact that the tissue specimen contains a variety of molecules, the presented results of molecular ordering and orientation agree with the theoretical prediction based on sample composition and vibration directions.



INTRODUCTION

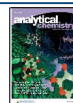
Infrared (IR) spectroscopic imaging offers a wealth of information about the chemical composition of the sample in a spatially resolved manner, which is particularly attractive in biomedical applications.^{1,2} However, structural aspects are still waiting to be incorporated into this modality.^{3,4} Taking advantage of the anisotropic orientation of molecules in solid samples, it is possible to obtain different material responses when changing linear polarization of incoming IR light. Using this fact, the determination of molecular orientation with Fourier transform IR (FT-IR) was extensively explored, but mostly for polymer samples.^{5,6} For example, polarized IR spectroscopy turned out to be an effective tool to examine the orientation and relaxation of drawn poly(ethylene terephthalate).^{6–8} A more quantitative approach in the form of Herman's function with a two-polarization approach was used to determine the level of molecular ordering.^{9,6,9,10} Later, a four-polarization method was proposed,^{11–13} which gave a great tool for further polymer analysis,^{14,15} especially when combined with imaging modalities.^{12,13,16} A variety of other techniques may be used to determine molecular orientation: nuclear magnetic resonance, X-ray diffraction, polarized visible microscopy, fluorescence microscopy, second-harmonic generation (SHG), or Raman spectroscopy (RS).¹⁷ However, each has specific requirements or drawbacks, such as a specific sample form (liquid or crystallized), lack of biochemical information, a complex instrumentation setup, or the need for labeling. Most of these can be met or outrun with FT-IR

imaging. A specimen of biological origin adds a level of difficulty to this approach, and among all FT-IR studies determining molecular orientation, only a few were focused on biological samples, especially tissues.^{16–21} Although the most recent study used Stokes vector and a quantum cascade laser microscope to investigate fiber orientation in breast tissue surgical section,²² that approach does not offer insights into vibrational mode orientation. This is especially challenging because a biological specimen contains a multitude of different molecules and deciphering contributions from each is a daunting task. However, the possibility of obtaining a vibrational mode-specific orientation map is very tempting and will open new avenues for research. So far, RS techniques are closest to reaching this goal, with a wealth of research done for tissue samples using coherent anti-Stokes Raman scattering (CARS) and stimulated Raman scattering (SRS) microscopy.^{23,24} The differences in selection rules make FT-IR more sensitive to changes in the main molecular chain while RS to amino acid side chains. Furthermore, RS-based techniques are more sensitive to lipids while FT-IR to proteins. Another

Received: June 17, 2020

Accepted: August 28, 2020

Published: August 28, 2020



example of mapping method sensitive to vibrational modes is phase-sensitive vibrationally resonant sum-frequency generation (PSVR-SFG) microscopy derived from SHG—also used for tissue samples.²⁵ PSVR-SFG, CARS and SRS use relatively high power lasers to achieve specific vibrational transitions and are thus called discrete imaging techniques, while FT-IR measures the full spectrum and thus provides wide spectral information. Also, the use of lasers increases the risk of sample burn. On the other hand, the strength of CARS and SRS is the possibility to measure the second- and fourth-order parameter of the orientational distribution function, when FT-IR and PSVR-SFG can measure only the second-order parameter. Finally, FT-IR has an order of magnitude lower spatial resolution (in the far field) than the techniques based on visible light, which does not allow obtaining a high spatial detail but offers a large-scale sample coverage. Summing up, all techniques have their pros and cons and selection of a measurement method should be done including the sample type and targeted chemical information.

In this work, for the first time, we applied the four-polarization method to a human tissue (mainly fibrous and collagen-rich). The four-polarization approach allows obtaining unique information about the system—an in-plane orientation of dipole moment transitions (Herman's orientation function) along with the azimuthal angle.^{11,12,16} Combining this computational method with FT-IR allows characterization of different vibrational modes.

MATERIALS AND METHODS

Sample Preparation and Measurements. A paraffin-embedded pancreatic tissue microarray (TMA) PA2072a was purchased from Biomax Inc. The TMA was sliced into 5 μm thick samples, one was placed on a regular glass and stained with hematoxylin and eosin (H&E) for histopathological inspection and the other was mounted on a BaF₂ salt plate for IR transmission measurements. The second sample was deparaffinized with 24 h hexane bath. Three tissue cores, later referred to as C1, D14, and J14 (Biomax naming convention), were selected for this research. All cores came from different patients with malignant ductal adenocarcinoma and were chosen because of high fibrous tissue content—adequate for molecular orientation study.

FT-IR measurements were performed in transmission mode with a Bruker Vertex 70v spectrometer coupled with a Hyperion 3000 microscope with 15 $\times$ objective (2.7 μm projected pixel size, NA = 0.4) and a 64 $\times$ 64 FPA MCT detector. Two linear polarizers were inserted in default slots (first above the top objective and second under the bottom one) and were simultaneously rotated to remain in parallel to each other and obtain four target polarizations: 0, 45, 90, and 135°. Spectra were acquired from 3850 to 900 cm^{-1} , with an 8 cm^{-1} spectral resolution and a zero filling factor of 1. Sample and background signals were covered 4 and 64 times, respectively. Furthermore, the stage was programmed to repeat the same positions for each polarization measurement, which led to four pixel-to-pixel spatially aligned data sets.

Data Preprocessing. To increase the spectral and spatial quality, data were denoised with the minimum noise fraction method, with 15 bands used for reconstructions.^{26,27} Rubber-band baseline correction was applied, where each spectrum was divided into three segments: 3845–2750, 2746–1774, and 1770–900 cm^{-1} , determining ends for linear baseline subtraction. Specific band absorption was characterized by

band integration rather than intensities for a single wavelength, which in our opinion provided a more stable outcome (results not shown).

Herman's Orientation Function and Azimuth Angle Theory. Anisotropic samples, such as polymers or fibers, absorb linearly polarized light to a different extent—depending on the polarization direction. Such an effect is called linear dichroism and is being often applied in the studies of molecular chain orientation. During IR absorption measurements with polarized light, if a specific band is to be observed, its dipole moment derivative must exhibit a nonzero component in the direction of light polarization. This is a condition for a band to be present in the spectrum. As a result, bands with dipole moments changing parallel to the polarization will appear as most intense, with the opposite situation for perpendicular bands.²⁸ A dichroic ratio was introduced to quantify this effect. It requires measurements with two orthogonal polarizations and is defined as^{11,28}

$$D = \frac{A_{0^\circ}}{A_{90^\circ}} \quad (1)$$

where A_{0° and A_{90° are the absorbance for linear polarizations of 0 and 90°, respectively. On its own, this parameter provides information about the dipole moment orientation distribution with respect to the orientation axis but might also be employed in other parameter calculations.

A mathematical description of molecular orientation can be done utilizing orientation distribution function, which in the case of molecules with cylindrical symmetry is being approximated by a series of Legendre polynomials (with respect to the symmetry axis).²⁹ Depending on the mechanism of optical interaction, different measurement techniques can determine a different number of Lagrange polynomial coefficients.²⁹ FT-IR can provide a second-order coefficient,²⁹ which is also called Herman's orientation function or in-plane orientation function, calculated as^{11,15}

$$f = \frac{D - 1}{D + 2} \frac{2}{3 \cos^2 \alpha - 1} \quad (2)$$

where D is the dichroic ratio and α is the angle between the transition dipole moment (of a specific absorption band) and the main molecular chain axis. The α angle needs to be defined individually for each absorption band; nonetheless, an approximation can be applied. If α is smaller than $\approx 54.73^\circ$, the band can be interpreted as parallel, whereas in the other case, the band can be interpreted as perpendicular. Herman's function f exhibits results in the range from -0.5 to 1 , where -0.5 and 1 correspond to perfectly perpendicular and parallel bands, respectively. Reaching 0 indicates the isotropic character of a sample. However, 0 might also be obtained when polarization 0° forms a 45° angle with the main molecular axis ($D = 1$)—when two probing points are symmetrical on the sides of the cosine function. Such a result implies random orientation of a sample which, in fact, might be highly oriented. This matter was nicely discussed in the literature,¹¹ where an influence of 0 and 90° polarization direction on the in-plane orientation function was shown.

To aid this issue, a four-polarization method was developed, where a degree of molecular orientation can be determined independently from the choice of the direction of incident radiation polarization.^{11,15} A simple formula to describe the absorbance A_λ under linearly polarized light was derived¹¹

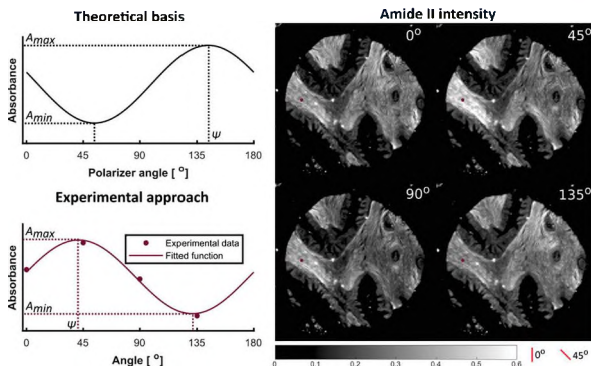


Figure 1. (Top left) Theoretical presentation of absorbance dependence from polarization. $A_{\max}$ and $A_{\min}$ correspond to the maximum and minimum absorption, respectively, and ψ represents the polarization when the absorbance reaches $A_{\max}$. (Right) Intensity images of amide II absorbance under different polarizations. Purple dots indicate the pixel used for experimental approach demonstration. (Left bottom) Presentation of the experimental approach, where a cosine function indicated by equation 3 was fitted to absorbance for a single-pixel absorbance of amide II under four polarizations: 0, 45, 90, and 135°. Dashed lines are indicators of parameters of interest.

$$A_{\gamma} = \frac{A_{\max} + A_{\min}}{2} + \frac{A_{\max} - A_{\min}}{2} \cos 2(\gamma - \psi) \quad (3)$$

where $A_{\max}$ and $A_{\min}$ are the maximum and minimum absorbance, γ is a polarizer angle, and ψ is a polarizer angle when $A_{\gamma} = A_{\max}$. This formula is graphically shown in the left top part of Figure 1, whereas visible differences in absorbance for four polarizations are presented on the right. This clearly shows how different regions of a sample are being highlighted when polarization matches the direction of a transition dipole moment. Different approaches of utilizing formula 3 were proposed,^{11–13} with a nonlinear fitting method being preferred for this study. An exemplary calculation is shown in the left bottom part of Figure 1. Four experimental points describing absorbance under different polarizations (for a single pixel) are used as experimental results. The cosine function is later fitted allowing to find $A_{\max}$, $A_{\min}$, and ψ . This way, a new dichroic ratio might be determined in the form of

$$D_{\max} = \frac{A_{\max}}{A_{\min}} \quad (4)$$

which, used for Herman's function calculation, provides results independent of the initial choice of orthogonal polarizer angles.

Another major source of information is delivered by the azimuthal angle ψ . This parameter points out polarization which, for a specific band, results in maximum absorption. In other words, polarization ψ is parallel to a transition dipole moment; therefore, it might be considered as a chemical bond direction or in-plane orientation indicator. An example is demonstrated in Figure 1, where purple dots in the right panel indicate pixel with absorbance plotted on the left side. The fitted curve shows ψ close to 45°, which agrees with the fact that this pixel comes from a region where the absorbance value was the highest for 45° polarization.

RESULTS AND DISCUSSION

As mentioned in the **Materials and Methods** section, samples with high fibrous tissue content were selected for this study to ease the interpretation. Histological images (H&E staining) of all three tissue sections are shown in the top part of Figure 2. Fibrous tissue with its cylindrical shape suits perfectly for molecular orientation studies. Moreover, high collagen content allows to determine the spectral regions and bands useful for this study. A single collagen molecule is built from three polypeptide chains, arranged into a helical conformation.^{15,30} Therefore, three most important regions for FT-IR spectroscopic analysis are^{15,18,30} amide I band (C=O stretching vibrations in the 1700–1600 cm^{-1} range), amide II band (C–N stretching and N–H bending vibrations in the 1590–1490 cm^{-1} range), and amide III band (N–H bending, C–N stretching, and C–H stretching vibrations in the 1290–1210 cm^{-1} range).

Herman's Function. Using eq 2 and amide III band integration, Herman's function was calculated for all pixels of the three tissue sections and results are presented in the middle part of Figure 2. The brightest regions indicate the areas with the highest molecular ordering. Comparing the images of in-plane orientation function with those of H&E shows that highly oriented regions originate mainly from fibrous tissues. Therefore, a random orientation is observed for epithelial cells and parts where fibers and cells are mixed together. Spectra from pixels with high and low (random) orientation were further investigated and are presented in the bottom part of Figure 2. The analysis of Herman's orientation function was carried out for amide III band (assuming $\alpha = 0^\circ$) and spectral differences come to an agreement with what is being displayed above. Highly oriented pixels show significant differences between polarizations, while spectra from pixels with random orientation are almost overlapping in the amide III region. However, an interesting observation is that the same pixels that exhibit low orientation for the amide III band might show substantial differences for the amide I and II bands, as is shown

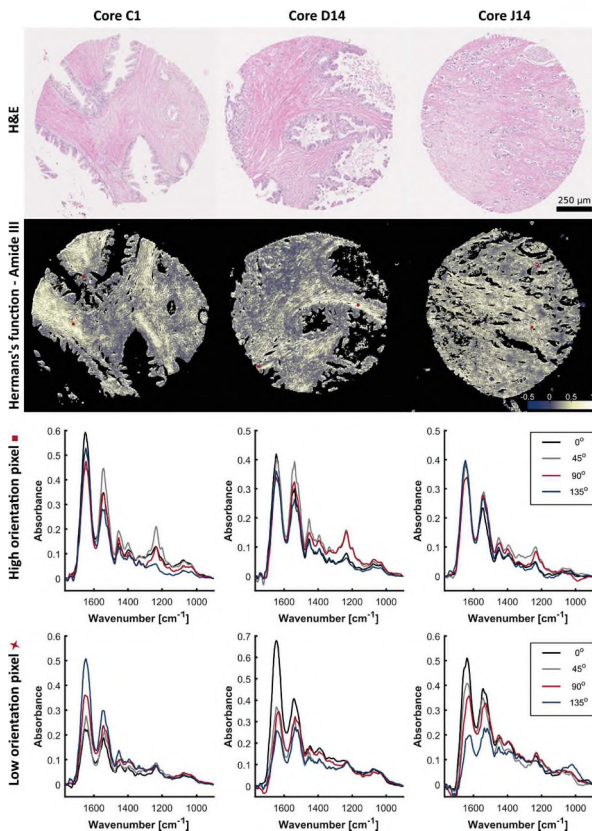


Figure 2. (Top row) Histological H&E stained image of samples. Pink regions are associated with fibrous tissues while purple with nuclei (cancerous cells). (Middle row) Herman's orientation function results calculated for amide III band integration. Red squares and crossed point pixels with high and low orientation, respectively. (Bottom) Spectra corresponding to pixels with a high and low orientation of amide III band. The pixels with low amide III band orientation were chosen to have at the same time high orientation of amide I band.

in the bottom spectra of Figure 2. The different behavior of spectral regions originating from the same protein contrast is very interesting. This implies different biomolecules contributing to the spectral signal in a distinct manner. Moreover, individual conformations of the same molecule may have unique ordering characteristics. To understand this spectral behavior, model systems of known composition will need to be investigated; however, it is very promising that such large differences can be observed in a biological, highly heterogeneous matrix.

Azimuthal Angle. An azimuthal angle ψ describes the in-plane orientation of specific molecular bonds, or being more specific, the direction of transition dipole moment. The results

for amide I, amide II, and amide III of all three samples are presented in Figure 3. The results span from 0 to 180° because of the fact that phase shifting by π gives the same results. A circular chart representing color-angle correspondence is shown in Figure 3. The most interesting feature to be inspected is the main molecular chain axis (fiber direction), which was observed in histological images (Figure 2). Amide II and amide III results in Figure 3 show a very similar outcome for all three samples. Observed angles come to a complete agreement with fiber structure in H&E images. The explanation of this comes from analyzing the type of vibrations involved in amide II and amide III spectral regions. Both include C–N stretching and N–H bending vibrations, which

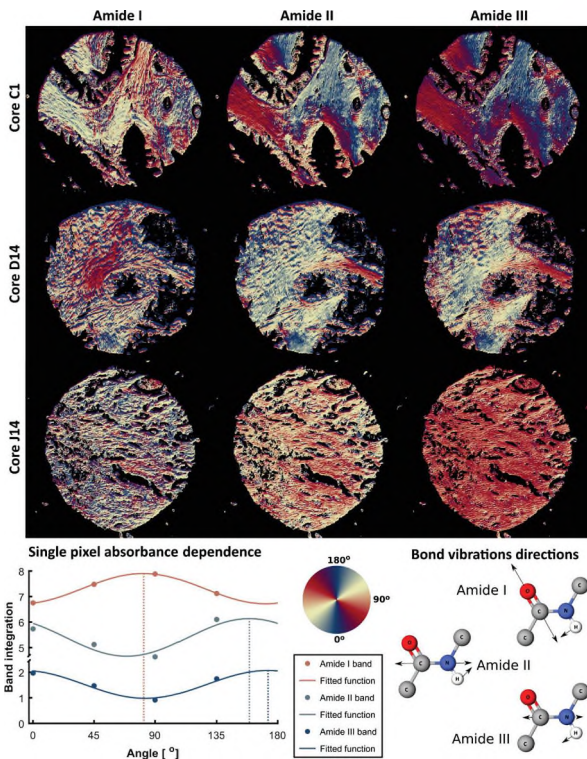


Figure 3. (Top) Orientation of amide I, amide II, and amide III bands represented by the azimuthal angle ψ . (Bottom left) Absorbance dependence from polarization for three analyzed spectral regions, for a single pixel from core C1. (Bottom right) Theoretical representation of amide bond vibration directions.

are present in collagen polypeptide chains. Especially, stretching of the C–N bond might be a great indicator of the main molecular chain orientation of collagen. The spatial sensitivity of this method is particularly visualized by C1 core results, with fibers changing directions throughout the whole sample. Furthermore, one of the most interesting observations is that the spatial orientation coming from amide I band is almost perpendicular to that of amide II and amide III bands, as predicted by the theoretical models of bond vibration directions. Absorbance dependence from the polarization of all the three spectral regions was plotted for a single pixel from core C1. The results are shown at the bottom of Figure 3. This plot is color-coded, so each series color corresponds to a pixel color in the top image. It is clear that the maximum absorbance (indicated by the dashed line) is shifted by $\sim 90^\circ$ for amide II and amide III, with respect to amide I. Another strong feature of this method is pointed out by the results of orientation for epithelial regions—which were found to be mostly random.

This is caused by the lack of collagen or other molecules with the oriented amide III band. However, random orientation might also be achieved by collagen in two specific cases: first, when fiber direction is parallel to the optical axis—perpendicular to the focal plane and second, when bundles of fibers exhibiting different values of the azimuthal angle meet at one point and are below the spatial resolution capabilities of the system.

These examples show that even in a complex biological matrix containing multiple components, it is possible to observe macromolecular ordering. Moreover, the four-polarization method enables vibration-specific observation of spatially resolved orientation of dipole moment transitions. This opens up a new area of research, not only focused on polymers but also on naturally occurring macromolecules.

■ CONCLUSIONS

The field of spatially resolved IR polarimetry is rapidly expanding, and more advanced approaches allow the identification of new structural parameters in addition to just chemical composition. The first application of the four-polarization method to human tissue showed that calculating Herman's orientation function and obtaining the azimuthal angle of selected vibrational modes is feasible and opens up new avenues in studying biological materials.

■ AUTHOR INFORMATION

Corresponding Author

Tomasz P. Wrobel – *Solaris National Synchrotron Radiation Centre, Jagiellonian University, Krakow 30-392, Poland; Institute of Nuclear Physics, Polish Academy of Sciences, Krakow PL-31342, Poland; orcid.org/0000-0003-2053-0570; Email: tomek.wrobel@uj.edu.pl*

Authors

Paulina Koziol – *Solaris National Synchrotron Radiation Centre, Jagiellonian University, Krakow 30-392, Poland; Institute of Nuclear Physics, Polish Academy of Sciences, Krakow PL-31342, Poland*

Danuta Liberda – *Solaris National Synchrotron Radiation Centre, Jagiellonian University, Krakow 30-392, Poland; Institute of Nuclear Physics, Polish Academy of Sciences, Krakow PL-31342, Poland*

Wojciech M. Kwiatek – *Institute of Nuclear Physics, Polish Academy of Sciences, Krakow PL-31342, Poland; orcid.org/0000-0002-2197-8572*

Complete contact information available at:
<https://pubs.acs.org/10.1021/acs.analchem.0c02591>

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was supported by the National Science Centre, Poland ("three-dimensional macromolecule orientation by means of IR imaging and its significance in cancer micro-environment", grant no. 2018/31/D/ST4/018). This research was performed using the equipment purchased in the frame of the project cofunded by the Malopolska Regional Operational Program Measure 5.1 Krakow Metropolitan Area as an important hub of the European Research Area for 2007–2013, project no. MRPO.05.01.00-12-013/15. The authors declare no conflicts of interest.

■ REFERENCES

- Wrobel, T. P.; Mateuszuk, L.; Kostogrys, R. B.; Chlopicki, S.; Baranska, M. *Analyst* **2013**, *138*, 6645–6652.
- Walsh, M. J.; Reddy, R. K.; Bhargava, R. *IEEE J. Sel. Top. Quantum Electron.* **2012**, *18*, 1502–1513.
- Wrobel, T. P.; Bhargava, R. *Anal. Chem.* **2018**, *90*, 1444–1463.
- de los Santos Pereira, A.; Cernescu, A.; Svoboda, J.; Sivkova, R.; Romanenko, I.; Bashta, B.; Keilmann, F.; Pop-Georgievski, O. *Anal. Chem.* **2020**, *92*, 4716–4720.
- Everall, N. J.; Chalmers, J. M.; Local, A.; Allen, S. *Vib. Spectrosc.* **1996**, *10*, 253–259.
- Lu, X.; Hay, J. N. *Polymer* **2001**, *42*, 8055–8067.
- Branciforti, M. C.; Silva, L. B.; Bretas, R. E. S. *J. Appl. Polym. Sci.* **2006**, *102*, 2241–2248.
- Duchesne, C.; Kong, X.; Brisson, J.; Pérolet, M.; Prud'homme, R. E. *Macromolecules* **2002**, *35*, 8768–8773.

- Albunia, A. R.; Di Masi, S.; Rizzo, P.; Milano, G.; Musto, P.; Guerra, G. *Macromolecules* **2003**, *36*, 8695–8703.
- Albunia, A. R.; Musto, P.; Guerra, G. *Polymer* **2006**, *47*, 234–242.
- Hikima, Y.; Morikawa, J.; Hashimoto, T. *Macromolecules* **2011**, *44*, 3950–3957.
- Hikima, Y.; Morikawa, J.; Hashimoto, T. *Macromolecules* **2012**, *45*, 8356–8362.
- Hikima, Y.; Morikawa, J.; Hashimoto, T. *Macromolecules* **2013**, *46*, 1582–1590.
- Ryu, M.; Balçiyti, A.; Wang, X.; Vongevirut, J.; Hikima, Y.; Li, J.; Tobin, M. J.; Juodkazis, S.; Morikawa, J. *Sci. Rep.* **2017**, *7*, 7419.
- Riaz, T.; Zeeshan, R.; Zarif, F.; Ilyas, K.; Muhammad, N.; Sali, S. Z.; Rahim, A.; Rizvi, S. A. A.; Rehman, I. U. *Appl. Spectrosc. Rev.* **2018**, *53*, 703–746.
- Wrobel, T. P.; Mukherjee, P.; Bhargava, R. *Analyst* **2017**, *142*, 75–79.
- Cheng, J.-X.; Xie, X. S. *Science* **2015**, *350*, aaa8870.
- Camacho, N. P.; West, P.; Torzilli, P. A.; Mendelsohn, R. *Biopolymers* **2001**, *62*, 1–8.
- Hsu, W.-L.; Davis, J.; Balakrishnan, K.; Ibn-Elhaj, M.; Kroto, S.; Brock, N.; Pau, S. *Opt. Express* **2015**, *23*, 4357.
- Choi, J.-H.; Cho, M. J. *Phys. Chem. B* **2014**, *118*, 12837–12843.
- Miller, L. M.; Dumas, P. *Biochim. Biophys. Acta* **2006**, *1758*, 846–857.
- Phal, Y.; Yeh, K. L.; Bhargava, R. *Polarimetric Infrared Spectroscopic Imaging Using Quantum Cascade Lasers. Advanced Chemical Microscopy for Life Science and Translational Medicine*; SPIE, 2020; p 1125210.
- Dubois, J.; Berto, P.; Gasecka, P.; Bioud, F.-Z.; Ferrand, P.; Rigneault, H.; Brasselet, S. *J. Phys. Chem. B* **2015**, *119*, 3242–3249.
- de Vito, G.; Bilone, A.; Piazza, V. *Opt. Express* **2012**, *20*, 29369.
- Han, Y.; Raghunathan, V.; Feng, R.-r.; Maekawa, H.; Chung, C.-Y.; Feng, Y.; Potma, E. O.; Ge, N.-H. *J. Phys. Chem. B* **2013**, *117*, 6149–6156.
- Koziol, P.; Raczowska, M. K.; Skibinska, J.; Urbaniak-Wasik, S.; Paluszkievicz, C.; Kwiatek, W.; Wrobel, T. P. *Sci. Rep.* **2018**, *8*, 14351.
- Raczowska, M. K.; Koziol, P.; Urbaniak-Wasik, S.; Paluszkievicz, C.; Kwiatek, W. M.; Wrobel, T. P. *Anal. Chim. Acta* **2019**, *1085*, 39–47.
- Griffiths, P. R.; De Haseth, J. A. *Fourier Transform Infrared Spectrometry*; Wiley-Interscience, 2007.
- Lee, Y. J. *Opt. Express* **2018**, *26*, 24577.
- Belbachir, K.; Noreen, R.; Gouspillou, G.; Petitbois, C. *Anal. Bioanal. Chem.* **2009**, *395*, 829–837.

Super-Resolved 3D Mapping of Molecular Orientation Using Vibrational Techniques

Paulina Koziol, Karolina Kosowska, Danuta Liberda, Ferenc Borondics, and Tomasz P. Wrobel*

Cite This: *J. Am. Chem. Soc.* 2022, 144, 14278–14287

Read Online

ACCESS |



Metrics & More

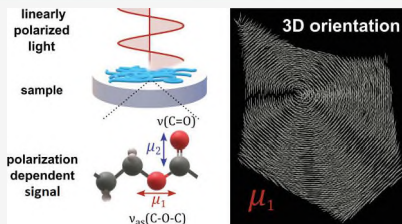


Article Recommendations



Supporting Information

ABSTRACT: When a sample has an anisotropic structure, it is possible to obtain additional information controlling the polarization of incident light. With their straightforward instrumentation approaches, infrared (IR) and Raman spectroscopies are widely popular in this area. Single-band-based determination of molecular in-plane orientation, typically used in materials science, is here extended by the concurrent use of two vibration bands, revealing the orientational ordering in three dimension. The concurrent analysis was applied to IR spectromicroscopic data to obtain orientation angles of a model polycaprolactone spherulite sample. The applicability of this method spans from high-resolution, diffraction-limited Fourier transform infrared (FT-IR) and Raman imaging to super-resolved optical photothermal infrared (O-PTIR) imaging. Due to the nontomographic experimental approach, no image distortion is visible and nanometer scale orientation domains can be observed. Three-dimensional (3D) bond orientation maps enable in-depth characterization and consequently precise control of the sample's physicochemical properties and functions.



■ INTRODUCTION

Chemical characterization and visualization of nanoscale structures and interphases in materials like semicrystalline polymers (spherulites, nanofibers), nanocomposites, and biological systems (proteins, peptides) are critical for a deeper understanding of the correlation between the “superstructure” of materials and its physicochemical and biological properties. Reorganization at the nano-structural level even makes it possible to obtain metal-like thermal conductors from polymers, commonly known as insulators. As an example, Hu et al. presented a polyethylene film with parallel aligned nanofibers with thermal conductivity two orders of magnitude greater than the normal polymer.¹ Research is also ongoing into flexible energy storage based on polymer nanohydrogels,² e-skin for medicine,³ and intelligent (biomimetic) scaffolds to induce a cellular response, e.g., stem cell differentiation,⁴ all of which originate from the molecular organization at the nano- and meso-scale. Several methods can be used to evaluate the structural and compositional properties of nanomaterials. Techniques such as X-ray diffraction probe only crystalline phases, while polarization modulation-infrared reflection-absorption spectroscopy (PM-IRRAS), calorimetry, or nuclear magnetic resonance provide very limited spatial resolution. Electron microscopy (EM) and atomic force microscopy (AFM) offer high spatial resolution at the cost of specificity. Optical spectroscopic techniques (fluorescence, Raman, infrared, or X-ray) seem to be the middle ground, providing a wealth of information while being spatially diffraction-limited,⁵

however, not offering fully three-dimensional (3D) resolved information apart from single analytes.

Fourier transform infrared spectroscopy (FT-IR) in combination with chemometrics is an intensively developed, powerful tool for the investigation of a material's chemical structure with simultaneous microscopic visualization.⁶ One of the shortcomings of the FT-IR imaging method is the inherent low spatial resolution ($\sim 2\text{--}10\ \mu\text{m}$), restricted by the diffraction limit close to the wavelength (λ) of light. An emerging technique that overcomes this limitation by detecting photothermal expansion caused by IR absorption using an optical probe (a visible laser) has recently been commercialized under the name of optical photothermal infrared spectroscopy (O-PTIR).⁷ This far-field microscopic technique overcomes several limitations of traditional IR microspectroscopy: sample thickness/reflective properties, optical properties of the sample substrate, or the presence of water while offering super-resolved imaging with an effective 400 nm spatial resolution.⁸

Received: May 18, 2022

Published: July 26, 2022



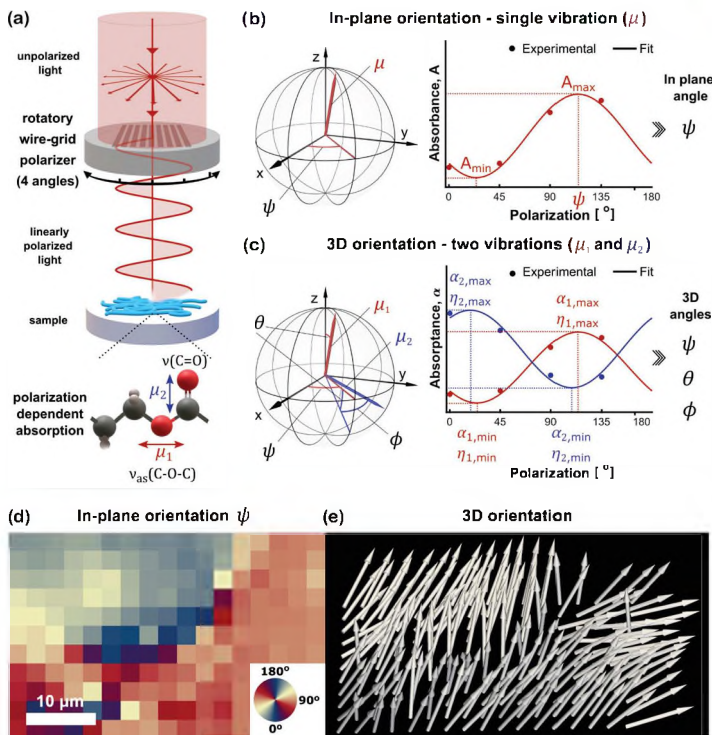


Figure 1. Orientation determination approach (a) Schematic drawing of polarized FT-IR experimental setup including absorbance dependence of a molecule from the orientation of incident polarization. (b) Visualization of the 4P method used to determine in-plane orientation (2D) of transition dipole moment μ . (c) Visualization of the 4P method used to determine 3D orientations of transition dipole moments μ_1 and μ_2 . (d, e) Examples of visualization results for 2D and 3D methods, for the same sample region, respectively.

IR spectroscopy with linear polarization can provide new and useful information about the orientation of molecules.⁹ When a sample has an anisotropic structure, changing the polarization of incident light in a microscopic investigation can reveal additional information. The in-plane direction of transition dipole moments of vibrational bands can be described by the azimuthal angle (ψ), when a linear polarization-based method is applied. The IR⁹ and Raman¹⁰ two-polarization (2P) methods were used to examine the orientation in polymeric samples for decades, however, with significant limitations. Next, the four-polarization method (4P) has been proposed, which enabled the study of the in-plane organization of molecules in complex systems. The first applications of IR imaging with 4P on polymeric systems^{11–14} were followed by active compound visualization¹⁵ and recently by human tissue microarray analysis by Koziol et al. in 2020.^{16,17}

Until now, experiments focused on using a single vibrational band to retrieve the in-plane orientation and internal ordering of samples. 4P is experimentally a simple (no sample tilting) framework that offers much more with proper mathematical treatment. A concurrent analysis of two vibrational bands of roughly perpendicular transition moment orientations was suggested and results in full 3D angles (in-plane and out-of-plane) retrieval.¹⁸ We show here, to the best of our knowledge, the first application of concurrent analysis (4P-3D) to IR imaging data and obtain orientation angles of a model spherulite polycaprolactone sample. Independently, Xu et al. followed with two examples of concurrent 3D orientation application to IR imaging.^{19,20} Moreover, we show that this method can be applied to diffraction-limited FT-IR and Raman imaging and even to super-resolved O-PTIR imaging. Due to the stationary experimental approach, no image distortion is visible and nanometer scale domains can be observed.

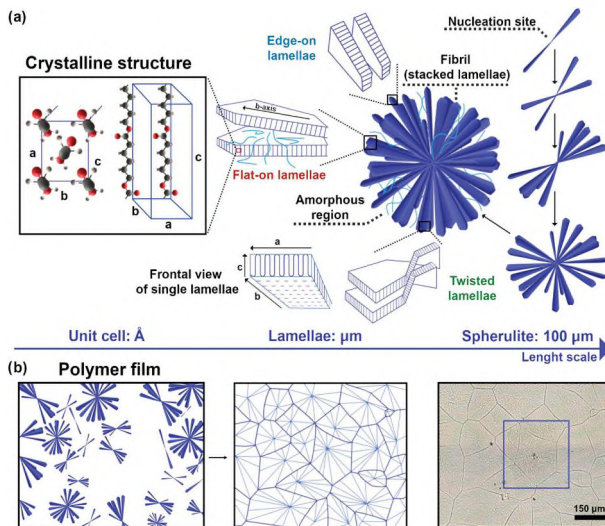


Figure 2. Schematic representation of the hierarchical structure of the model PCL polymer spherulite. In panel (a), the orientation of polymer chains and crystalline unit cell of PCL is indicated together with the crystal axes of the lamella made of tightly folded chains. A spherulite is composed of highly ordered lamellae, connected by amorphous regions, in different orientations: the “edge-on” orientation, “flat-on” orientation, and twisted form. Spherulite growth during polymer crystallization starts in the nucleation site. The size is controlled by the nucleation of spherulites and is in the tens of μm range with an approx. $10\ \mu\text{m}$ thickness. Panel (b) shows the schematic illustration of polymer film crystallization and optical microscopic image of a PCL film with a marked spherulite selected for detailed study. With a large number of nucleation sites, spherulites interact with each other upon growth.

RESULTS AND DISCUSSION

Figure 1 presents the concept of the 4P-3D analysis (briefly described in the Materials and Methods section) leading to the retrieval of 3D orientation. Four different linear polarization angles allow probing a sinusoidal function that arises from orientation-dependent absorbance. The analysis of a single vibration leads to obtaining a 2D in-plane angle of orientation, while simultaneous analysis of two near-orthogonal vibrations retrieves 3D angles.

PCL Spherulite Model Sample. To present the capabilities of this approach a model polymer was selected and measured using FT-IR, O-PTIR, and Raman imaging. Poly- ϵ -caprolactone (PCL) is often used as a model system for polymer crystallization studies due to its well-known chemical and crystal structure. PCL has an orthorhombic crystal structure with a planar, zig-zag chain conformation.²¹ In the left panel of Figure 2a the unit cell of PCL is presented. Undercooling a solution or melt of semi-crystalline polymers often produces spherulites, each with a nucleation site at its center. However, in most polymer films, spherulites appear in the flat form with radial symmetry and grow until an impact stops the process (Figure 2b). Long fibrils which are clearly visible under an optical microscope, are made of lamellae oriented in parallel and separated by amorphous regions. The free space between fibrils can be filled with the amorphous phase or remain empty. To uniformly fill the space, new fibrils

must be generated during growth. Processes like heterogeneous nucleation on existing fibrils or branching generate new, neighboring fibrils (right panel of Figure 2a). Several AFM studies showed that in the early stages of growth, spherulites appear in the form of sheaves of lamellae.^{23–24} During growth crystals splay apart and fill the space.²⁵ A single lamella is made of highly oriented chains folded several times, so a single chain passes through the lamellar structure more than once. This is presented in the frontal view of a single lamella section in Figure 2a. The amorphous regions between lamellae are created of folds, chain ends, and unordered structures. Chains connecting neighboring lamellae are described as “tie chains”. Lamellae grow along crystallographic axis *b* and they can exhibit different types of orientations on the substrate, which can change from “flat-on” to “edge-on” as shown in the middle part of Figure 2a.

PCL Spherulite Imaging. Data for the polycaprolactone spherulite sample described above and characterized in Figure 2 were collected using three spectroscopic techniques: FT-IR, O-PTIR, and Raman spectroscopy. Each technique possesses individual features highly influencing the measurement time, e.g., detector size or measurement step, integration time, and objective magnification. Furthermore, the measurement time is strongly increased (in comparison with experiments done with unpolarized light) by the requirement of four-polarization dataset collection. Hence, an FT-IR spectrometer with a matrix

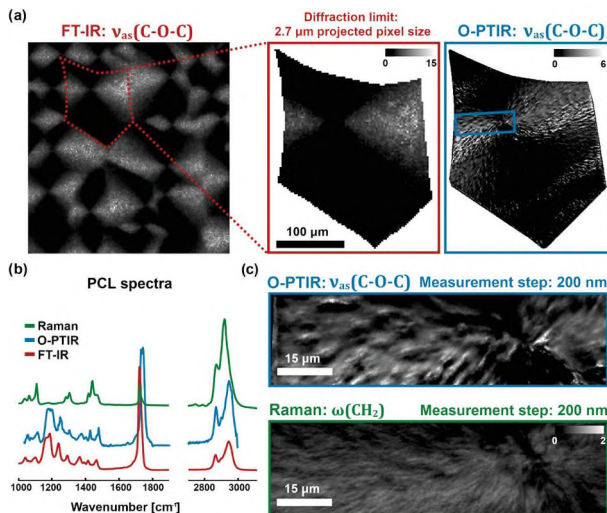


Figure 3. PCL spherulite imaging results. (a) Ratio of intensities corresponding to $\nu_{as}(\text{C-O-C})$ vibrations of the PCL sample, collected with two perpendicular polarizations of the incident light, using FT-IR and O-PTIR microscopes. Larger polymer film area (covering many spherulites) was acquired for FT-IR, with further attention focused on a single spherulite structure. (b) Examples of PCL spectra for FT-IR, O-PTIR, and Raman techniques (unpolarized), presenting correspondence between FT-IR and O-PTIR and the lack of some bands in the Raman spectrum. (c) Zoom of results shown above (a) for O-PTIR (blue rectangle) with corresponding results (ratio of intensities for two perpendicular polarizations) of the same sample region collected with Raman for the $\omega(\text{CH}_2)$ band.

detector and lowest spatial resolution allowed the measurements of a larger area of the polymer film than other modalities, offering at the same time full spectral coverage. O-PTIR on the other hand provides significantly higher spatial detail, but to cover a region of a full spherulite, only 5 discrete frequencies were chosen for measurements. Therefore, from the whole PCL film, a single spherulite (optical image in Figure 2b) was selected for detailed analysis. Raman measurements were narrowed to a smaller spherulite region, but still covered the nucleation site and the spherulite edge. The measurement steps were kept at 200 nm in both O-PTIR and Raman and allowed us to directly compare the two datasets as shown in Figure 3c. To visualize results, including sample heterogeneity, images of intensity ratios for two perpendicular polarizations are presented in Figure 3a,c as would typically be done in the 2P approach. Specific bands were selected for each method, $\nu_{as}(\text{C-O-C})$ for FT-IR and O-PTIR, and $\omega(\text{CH}_2)$ for Raman—both with transition moment changes approximately parallel to the main molecular axis. It is clearly visible that there is strong absorption dependence from the incident polarization direction—some regions are highlighted. Different spherulites might be recognized based on the FT-IR intensity image of the PCL film. Moreover, the nucleation site is clearly visible for all techniques. Exemplary spectra (unpolarized) from the PCL spherulite sample are presented in Figure 3b, with a good correspondence between FT-IR and O-PTIR techniques. Additionally, polarized FT-IR spectra from a single pixel are available in the supplementary data (Figure S20).

3D Orientation. Collected datasets were used to determine the orientation of the transition dipole moment and polarizability tensor (which is simplified to a vectorial value, due to the assumptions of the theoretical derivation),²⁶ which can be used to conclude on chemical bonds and macromolecular orientations. Even though it is not the focus of this article, the results of in-plane orientation along with Hermans in-plane orientation function are presented in the supplementary materials for all three spectroscopic techniques (Figures S1–S3).

As described in the theoretical section, 3D orientation calculations were done based on pairs of band intensities, corresponding to transition moments having approximately perpendicular orientations (with respect to each other). Exemplary results are presented in Figure 4 for $\nu_{as}(\text{C-O-C})$ used as a primary vector ($\nu(\text{C-C})$ in the case of Raman) and $\nu(\text{C=O})$ as the secondary vector. The primary vector was always chosen to be approximately parallel to the polymer chain. Unfortunately, the size of the large 3D datasets makes them hard to visualize, therefore, for clarity, in some cases, only every second point is presented. Moreover, the large pixel density along with acquired data size for O-PTIR (1500×1500 pixels) made it impossible to present the whole spherulite at once, therefore, the comparison to Raman data is the focus in Figure 4. Results for other pairs of spectroscopic bands and larger sample regions in O-PTIR are visualized in Figures S4–S19 in the Supporting Material.

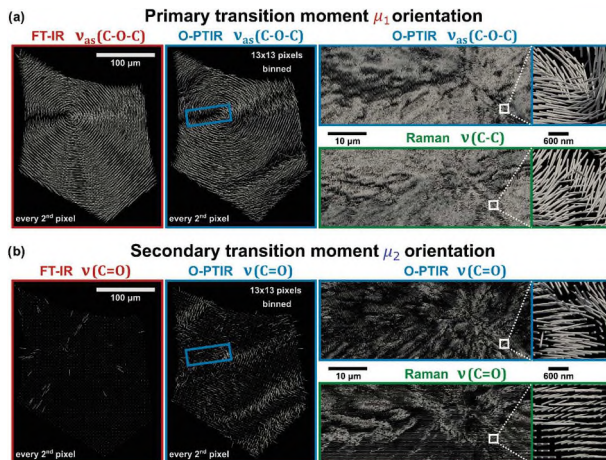


Figure 4. 4P-3D orientation results. Visualization of the primary (a) and secondary (b) transition moment orientation calculated for FT-IR, O-PTIR, and Raman. To compare FT-IR and O-PTIR results, O-PTIR data were binned (13×13 pixels) to yield a similar pixel size as FT-IR. Orientation results calculated using binned intensities are presented in the middle column. To preserve visualization clarity, results for every second pixel are presented in FT-IR and O-PTIR comparison sections. Orientations for a smaller area of spherulite (blue frame covering the nucleation center) are presented for O-PTIR (without binning) and Raman in the right column, using their high detail content coming from a diffraction-limited image probed with a 532 nm laser.

Considering the crystalline structure of lamellae in the PCL spherulite presented in Figure 2, conclusions could be done regarding theoretical transition moment orientations. With fibrils being organized radially from the nucleation center, the direction of the main molecular axis (polymer chain) is expected to be perpendicular to the fibril axis. This prediction is confirmed by 3D orientation results for $\nu_{as}(\text{C-O-C})$ and the best visualization is provided by FT-IR and binned O-PTIR shown in the left part of Figure 4a. The purpose of O-PTIR data binning yielding a similar spatial detail (pixel size) as FT-IR was to validate and compare results provided by the two techniques. Numerical evaluation is provided in Figure S8 in the form of Euclidean distance. Orientation vectors used for data presentation form concentric circles with the center located exactly within the nucleation site area, which confirms that the polymer chain is perpendicular with respect to the fibril's axis. The two techniques significantly differ in the optical setup and artifacts coming from scattering and varying sample thickness are also not the same. Still, the obtained results are consistent with each other, which is confirmed by the Euclidean distance results. However, the exact orientation of primary transition moment depends on lamellae alignment, which can vary from flat-on to twisted to edge-on (Figure 2a). Based on FT-IR and binned O-PTIR orientation of $\nu_{as}(\text{C-O-C})$ one may conclude that the whole spherulite consists of fibrils formed with edge-on lamellae (flat-on lamellae should exhibit an orientation perpendicular to the figure plane). Nonetheless, in this case, the pixel size ($\sim 2.7 \mu\text{m}$) is close to the size of lamellae, therefore, the averaging of signal coming from this region causes a significant loss of details observed. O-PTIR and Raman with much higher spatial resolution come to

aid, with results for a region of the spherulite including the nucleation site (blue rectangle) presented in the right part of Figure 4a. As expected, the comparison of $\nu_{as}(\text{C-O-C})$ for O-PTIR and $\nu(\text{C=O})$ for Raman shows great similarities. The general structure observed in FT-IR is also preserved in O-PTIR, but there is a striking increase in detail. Heterogeneity of the sample is clearly visible with wavy structures probably corresponding to fibril shapes. At such magnification, it cannot be stated if the spherulite is built exclusively from edge-on lamellae, because both parallel and out of the figure plane vectors are observed in different sample regions.

Interesting results are also observed for the secondary transition moment (Figure 4b). The $\nu(\text{C=O})$ vibration should be approximately perpendicular to the molecular chain, but it is hard to predict their exact theoretical orientation even if lamellae alignment (edge-on or flat-on) would be determined. Therefore, it is interesting that for FT-IR and binned O-PTIR (left sector of Figure 4b) vectors are consistently oriented perpendicularly to the figure plane. There is a slight tilt in O-PTIR results in comparison with FT-IR, but the vectors' general trend is the same, and differences could be again caused by the dissimilarities in optical setups. Raman and O-PTIR once more show analogous results, preserving a perpendicular trend with respect to the primary vector. Although, vectors are no longer consistently perpendicular to the figure plane and heterogeneous wavy shapes can be again observed. A detailed relation between primary and secondary vectors might be observed in zoomed-in areas for O-PTIR and Raman.

Combining orientation information of primary and secondary vectors, one may conclude that the investigated spherulite

is primarily built out of edge-on lamellae. First, for all observed techniques, primary vectors are mostly aligned or slightly out of the figure plane, which theoretically should be expected for edge-on units. This hypothesis is supported by the perpendicular character (to the figure plane) of secondary vectors. Obviously, there are regions exhibiting those properties to a different extent, but this is expected since the crystallization process is very subtle, and its conditions have a major influence on lamellae alignment.

Valid information about cylindrical symmetry and molecular ordering is provided by the $\langle P_2 \rangle$ parameter, with results shown in supplementary materials (Figures S4–S7 and S9–S19). As mentioned in the 3D part of the Theoretical Approach section (Materials and Methods), $\langle P_2 \rangle$ is used as a metric describing the breadth of orientation distribution function (ODF), e.g., $\langle P_2 \rangle = 1$ for infinitely narrow distribution parallel to the symmetry axis, whereas $\langle P_2 \rangle = 0$ in the case of an isotropic distribution. Negative values reaching $\langle P_2 \rangle = -1/3$ characterize distributions perpendicular with respect to the symmetry axis. Considering that primary transition moments used in this study were always chosen to be approximately parallel to the main molecular axis, the analysis of $\langle P_2 \rangle$ results may lead to conclusions in which bands are the best indicator of the main molecular axis. Based on FT-IR and O-PTIR results (Figures S4–S7 and S9–S14), the highest $\langle P_2 \rangle$ values are reached by $\nu(\text{C}=\text{O})$ vibrations. Additionally, close to zero results ($\langle P_2 \rangle = 0$) are observed (FT-IR data) around the nucleation center, which confirms its amorphous character. The smallest breadth of ODF in the case of Raman is observed for the $\nu(\text{C}=\text{C})$ vibration associated with 1037 cm^{-1} .

Molecular orientation (main polymer axis) and lamellae alignment based on O-PTIR were so far only discussed on a limited region of the sample corresponding to the Raman measurement area to allow for a direct comparison. A larger area presenting the orientation of $\nu_{\text{as}}(\text{C}=\text{O}-\text{C})$ is shown in Figure 5a with more structural insight. Once more, the presented region includes the nucleation site marked with a white circle, as it is the most characteristic spot in spherulites. Having a look at spherulite's larger area one can notice that there are two significantly different domains with zoom-ins marked with a white box. Interestingly, after overlaying the orientation image presented here (right zoom) with the optical microscopy image, fibrils are overlapping with structures formed by vectors (results not shown). However, such structures are not observed in the left zoom where vectors are more parallel to each other. This is regardless of the fact that fibrils observed in the optical image have the same structure, with no visible differences. This may lead to a conclusion that there are significant differences within fibril structures on the molecular level, probably associated with growth phases. As explained earlier (Figure 2a), at the early stages of spherulite formation there is a single fiber, which later becomes surrounded by other radially oriented fibers and branches until full spherulite is formed. For a polymer film with a thickness larger than 500 nm, the edge-on orientation of lamellae is expected. Single lamella is built of tightly packed and folded polymer chains. In the commonly accepted model, the chains are aligned normal to the plane of the lamella, so the orientation of chains determines the orientation of the lamella. In the spherulites, lamellae are stacked together, creating thicker structures, the fibrils (Figure 2). The unambiguous determination of lamellae orientation requires further development of the 4P-3D analysis method, however, obtained results

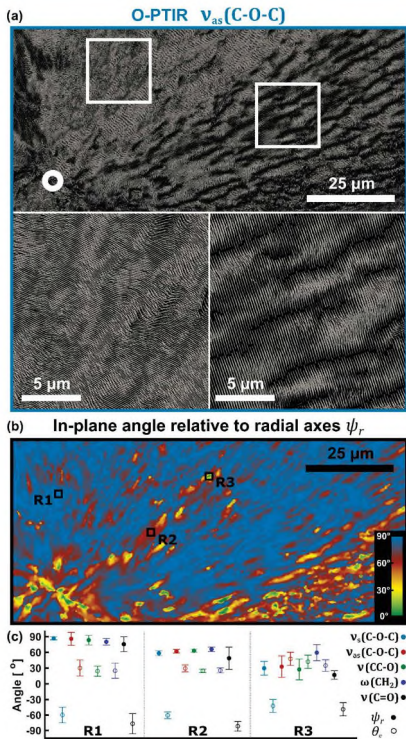


Figure 5. 4P-3D orientation analysis based on O-PTIR. (a) Orientation of the $\nu_{\text{as}}(\text{C}=\text{O}-\text{C})$ dipole moment based on the O-PTIR dataset. Zoomed areas are pointing to two clearly distinguishable sample regions, which may be related to areas of different spherulite growth phases. Nucleation site is marked with a white circle. (b) In-plane orientation relative to the radial axes ψ_r for $\nu_{\text{as}}(\text{C}=\text{O}-\text{C})$. (c) Mean values of ψ_r and θ_e for three regions (R1, R2, and R3) for different dipole moments along with standard deviations within that region.

give us partial information. Normal alignment of the secondary vectors to the image plane indicates more edge-on than flat-on orientation of lamellae. However, the tilt angle of lamellae requires additional modeling methods.²⁷ Yet, the observed dissimilarities can arise directly from growth stages.

To have a numerical quantity allowing to visualize and compare some regions of spherulite, a relative in-plane orientation ψ_r was defined as the difference between the ψ angle and the direction of fibers. ψ_r describes an in-plane angle relative to the radial axes formed by spherulite's fibrils. Exemplary results for the $\nu_{\text{as}}(\text{C}=\text{O}-\text{C})$ vibration (O-PTIR) are shown in Figure 5b. The presented region directly corresponds to the orientation in Figure 5a. As expected, the

majority of results exhibit values of around 90° (R1 region), apart from the nucleation center, which is amorphous. Nevertheless, there are some regions with different orientations—R2 and R3. Mean values of ψ , for all three regions are presented in Figure 5c, including the orientation of other vibrations. Additionally, the values for the $\theta = \frac{\pi}{2} - \theta$ angle (elevation) are also presented. A combination of these results allows us to analyze relative orientations of specific transition dipole moments. Orientation of $\nu_{\text{as}}(\text{C}=\text{O}-\text{C})$ is parallel to the PCL polymer main chain axis, whereas $\nu_{\text{s}}(\text{C}=\text{O}-\text{C})$ should be perpendicular.²⁸ Here, both vectors have similar values of ψ , but θ_{c} differs by around 90° for all three regions, which confirms the perpendicular character of those two. Very similar results are observed for the $\nu_{\text{as}}(\text{C}-\text{O}-\text{C})$ and $\nu(\text{C}=\text{O})$ pair. The $\nu(\text{CC}-\text{O})$ vibration is slightly more complex, coming from C–O and C–C stretching in the crystalline phase,²⁸ but the obtained results show that the transition dipole moment directions are parallel to the main chain axis, similar to $\nu_{\text{as}}(\text{C}-\text{O}-\text{C})$. The same orientation is also observed for $\omega(\text{CH}_2)$. Thus, those three may be considered as vibrations describing the main molecular axis orientation.

■ CONCLUSIONS

Here, we successfully demonstrated for the first time the four-polarization methods coupled to concurrent vibration analysis (4P-3D) on a spherulite polymer sample. We showed that this framework gives consistent results across the spectroscopic modalities, allowing diffraction-limited and even super-resolved imaging. Our research opens the way to applications for spatially resolved orientation studies of a wide range of thin film samples, in a nondestructive, label-free manner. We expect the 4P-3D method to impact both material and biological sciences as previously unattainable information can be extracted without a complicated experimental setup.

■ MATERIALS AND METHODS

Experimental Approach. Spherulite Sample Preparation. Polycaprolactone (PCL) of $M_n = 80\,000$ g/mol was purchased from Sigma Aldrich Sp. z o.o. (Poznan, Poland). Chloroform for analysis (99.8%), amylene stabilized was obtained from Idalia SpJ (Radom, Poland). PCL has a melting point (T_m) of 60°C and glass transition temperature (T_g) of around -60°C . The PCL film was prepared from solution (12 g of granular polymer in 80 mL of chloroform). The system was left on a magnetic stirrer for 24 h at room temperature to allow complete dissolution. The PCL film with a thickness of around $10\ \mu\text{m}$ was prepared by casting solution onto a calcium fluoride (CaF_2) slide. It was placed in a fume hood to evaporate the solvent overnight. PCL was melted at 75°C to erase the structure on a hot stage and crystallized under isothermal conditions at 35°C . After 5 h of spherulite growth, the sample was rapidly cooled down to room temperature.

FT-IR Imaging. FT-IR measurements were performed using a Bruker Vertex70v spectrometer coupled with a Hyperion 3000 microscope equipped with a 64×64 element FPA MCT array detector. A 15 $\times$ Cassegrain objective ($\text{NA} = 0.4$) was used giving a $2.7\ \mu\text{m} \times 2.7\ \mu\text{m}$ projected pixel size. To perform measurements with linearly polarized light, two wire grid polarizers were inserted before the objective and at the bottom of the condenser. Simultaneous rotation of both polarizers aligned in parallel allowed achieving desired polarizations—four datasets were collected, with relative polarizations corresponding to 0, 45, 90, and 135° . To provide pixel-to-pixel spatial alignment of all datasets, a motorized sample stage was programmed to repeat the same positions (for each polarization). Data acquisition was done in the $3850\text{--}900\ \text{cm}^{-1}$ spectral range, with $8\ \text{cm}^{-1}$ spectral resolution and zero filling factor of 1. Sample and

background spectra were co-averaged 4 and 64 times, respectively. Background datasets were collected individually for each polarization, in a blank region of the CaF_2 slide.

O-PTIR Imaging. O-PTIR imaging was done using a mIRage O-PTIR microscope. A quantum cascade laser (QCL) was used as a pump beam with a pulse rate of 100 kHz in a 2.4% duty cycle. Images were collected at five frequencies: 1736, 1368, 1296, 1246, and $1170\ \text{cm}^{-1}$ with corresponding QCL laser powers: 4, 24, 24, 18, and 18%, respectively. A 532 nm laser with 0.5% power (about 7.5 mW power on the sample) was used as a probe beam. Measurements were performed in the reflection mode using an avalanche photodiode detector (APD) and 200 nm steps. The background was measured once, on a low-e slide (Kevley Technologies). Since QCL laser is intrinsically linearly polarized, to obtain different relative polarizations the sample was rotated roughly by 45° increments to obtain a four-polarization dataset at relative 0, 45, 90, and 135° angles. The measurement region was adjusted each time the sample was rotated, to cover the whole spherulite of interest.

Raman Imaging. Raman measurements were performed using a WITec confocal CRM α 300 Raman microscope with a 50 $\times$ objective. The spectrometer was equipped with an air-cooled solid-state laser operating at 532 nm (with linear polarization) and a charge-coupled device (CCD) detector. The laser power measured before the objective was approximately 50 mW. The spectra were recorded with $4\ \text{cm}^{-1}$ spectral resolution and 0.1 s integration time. An achromatic half-wave plate was placed above the objective to control the natural laser polarization direction. Four polarizations of interest (-30 , 0, 60, and 120°) were achieved by the rotation of the wave plate. A rectangular fragment of the spherulite from the border of the neighboring spherulite to the nucleation site was measured with 200 nm steps (size: 150×450 steps).

Preprocessing. FT-IR Data. Spectral and spatial data quality was increased using minimum noise fraction (MNF) denoising, with 15 bands used for reconstructions.^{29,50} Furthermore, a local linear baseline was applied to the spectral ranges corresponding to vibrations of crucial chemical bonds, which allowed the mitigation of scattering artifacts. Single intensities (within corrected ranges) were used for transition dipole moment orientation calculations, with the attention focused on 1720, 1365, 1292, 1238, and $1165\ \text{cm}^{-1}$, coinciding with the O-PTIR data. Mentioned band identification for both IR methods is presented in Tables S1 and S2.

O-PTIR Data. Since for O-PTIR imaging different linear polarization angle data were collected by sample rotation (not by polarization control), an additional preprocessing step was required to spatially align four datasets. The final sample orientation was defined by the 90° polarization dataset (closest to FT-IR results orientation), whereas 0, 45, and 135° datasets were warped by global rotation and translation of each image. Calculation of rotation angles (85.7 , 136.9 , 42.5°) used for image transformation was done using the line marked on the sample before measurements. Following, images were translated referring to characteristic sample structures—spherulite edges. During data manipulation, we noticed that intensities for the last polarization dataset (135°) were significantly higher and therefore not comparable to other polarizations. To overcome this artifact, all images collected with the O-PTIR technique were divided by their median value. Finally, aiming to prepare data for further absorbance calculations in the 4P-3D orientation method, intensities for each spectral variable (cumulative data for all four polarizations) were rescaled to span within the intensity range corresponding to absorption values exhibited by FT-IR data. Furthermore, additional datasets were generated by binning 13×13 pixels (giving $2.6\ \mu\text{m}$ pixel), with the aim to later compare O-PTIR and FT-IR data ($2.7\ \mu\text{m}$ pixel).

Raman Data. The data treatment started with cosmic spike removal and Raman shift calibration to Si bands. This was followed by denoising by principal component analysis (PCA), with 15 principal components used for reconstructions, which significantly increased spectral and spatial data quality. Global baseline correction was done using the rubberband method with 13 spectral ranges used for linear corrections. One of the polarization datasets exhibited slightly higher

global intensities (distorting orientation calculations), therefore, each spectrum was normalized to its sum. Similar to O-PTIR, Raman spectra also required intensity rescaling for further calculations. However, due to the absence of some bands present in the FT-IR spectra, data were scaled so that intensities for the most intensive band in Raman spectra (2917 cm^{-1}) are similar to the absorbance of the most intensive bands in FT-IR data (1720 cm^{-1}). Identification of Raman bands used in further analysis is available in Table S3.

Theoretical Approach. Anisotropic materials, such as fibers or polymers with long molecular chains exhibit high absorption dependence on the polarization direction of the incoming light, both in IR spectroscopy and other spectroscopic techniques. This effect, called linear dichroism, is extensively applied in molecular orientation studies. In polarized infrared spectroscopy (FT-IR and O-PTIR), absorption by a specific molecular vibration is the highest when the transition dipole moment $\frac{\partial \mu}{\partial Q}$ vector is aligned with the polarization of light.^{3,1} A schematic example of an FT-IR instrumentation setup is shown in Figure 1a. Similarly, in polarized Raman spectroscopy scattering reaches a maximum for the best alignment of polarization with the highest polarizability tensor change $\frac{\partial \alpha}{\partial Q}$, resulting in maximum band intensity observed in the Raman spectrum.³⁷ Hence, the analysis of spectra collected for different incident light polarizations can provide information about the chemical bond and molecular orientation, with the dichroic ratio (D) being one of the most widely used parameters. D is defined as the ratio of absorbance (or intensity) for parallel and perpendicularly polarized light. However, in the case of Raman, we define D as a ratio of scattering intensities for perpendicular polarizations of the incident green laser used for Raman excitation, as opposed to the depolarization ratio, which compares the scattered photons. Thus, using only two polarizations introduces a specific measurement condition—parallel and perpendicular orientation of transition moment with respect to polarization. In other cases, D would exhibit misleading values, like in a situation with 45° angle, where both polarizations would be affecting the sample, in the same way, giving $D = 1$ and incorrectly suggesting isotropic properties. Such measurement geometry makes this method appropriate only for systems with molecules aligned alongside each other and well-known chemical bond orientation, excluding systems with molecules changing directions throughout the sample.

Four-Polarization Method for 2D Orientation. The disadvantages of the two-polarization method have been aided by the development of the four-polarization approach, which allows molecular bond orientation determination, independently from the choice of incident light polarization. Absorbance dependence from linear polarization may be described by a simple formula, as follows:^{12,13,35}

$$A(\eta) = \frac{A_{\min} + A_{\max}}{2} + \frac{A_{\max} - A_{\min}}{2} \cos 2(\eta - \psi)$$

where $A_{\min}$ and $A_{\max}$ are minimum and maximum absorbance, η is the incident polarization angle and ψ is the polarization angle when $A(\eta) = A_{\max}$. Collecting data with four different linear polarizations (for each pixel) gives the possibility to experimentally find the shape of $A(\eta)$, as shown in Figure 1b. The mathematical approach applied here is based on a nonlinear fit of a cosine function to extract $A_{\min}$, $A_{\max}$, and ψ . Considering dipole moment μ being oriented in 3D space, ψ gives limited information about the in-plane orientation of this vector. Exemplary result visualization (Figure 1d) clearly shows the disadvantage of the 2D method, where extracted information fully omits the out-of-plane component of the orientation. Based on $A_{\min}$ and $A_{\max}$, one can also compute Hermans (in-plane) orientation function defined as:^{12,13,33}

$$f = \frac{D - 1}{D + 2} \frac{2}{3 \cos^2 \kappa - 1}$$

where $D = A_{\max}/A_{\min}$ and κ describe the angle between the main molecular axis and transition moment (used values are available in Tables S1–S3). Hermans function is equal to the second-order

coefficient of a series of Legendre polynomials approximating the orientation distribution function (ODF) of a molecule with cylindrical symmetry.^{12,18} Therefore, it can be used to describe the level of molecular orientation of the sample. This approach is equally suitable for the analysis of FT-IR and Raman measurements, with absorbance and scattering intensity being the corresponding results.

Four-Polarization in 3D Orientation. Out-of-plane orientation issues, mentioned in the above section, were recently approached by Lee, who proposed a mathematical method allowing determination of 3D angles and order parameter for molecular orientation imaging with FT-IR.¹⁸ Derivation is established on the grounds of classical electromagnetic wave theories and concurrently uses two nonparallel transition dipole moments: the primary (μ_1) associated with the main molecular axis and a secondary (μ_2) representing nonparallel vibration. Based on this, the absorbance ($\alpha = 1 - T$) dependence from the polarization direction η for vectors μ_1 and μ_2 is described as:¹⁸

$$\begin{aligned} \frac{\alpha_1(\eta)}{\alpha_2} &= \langle P_2 \rangle \sin^2 \theta \cos^2(\eta - \psi) + \frac{1}{3}(1 - \langle P_2 \rangle) \\ \frac{\alpha_2(\eta)}{\alpha_2} &= \langle P_2 \rangle [\cos \theta \cos \phi \cos(\eta - \psi) + \sin \phi \sin(\eta - \psi)]^2 \\ &\quad + \frac{1}{3}(1 - \langle P_2 \rangle) \end{aligned}$$

where ψ , θ , and ϕ are angles describing the 3D orientation of μ_1 and μ_2 vectors (shown in Figure 1c), $\langle P_2 \rangle$ is the order parameter, whereas α^2 contains only nonorientational quantities. Similar to the 2D approach, nonlinear functions were fitted to four-polarization datasets. This allowed the reconstruction of the experimental shape of absorbance dependence on incident polarization. As presented in Figure 1c, the following parameters pairs were extracted from the obtained experimental functions: $\alpha_{1,\max}$, $\eta_{1,\max}$, $\alpha_{1,\min}$, $\eta_{1,\min}$, $\alpha_{2,\max}$, $\eta_{2,\max}$, and $\alpha_{2,\min}$, $\eta_{2,\min}$. Formulas needed for ψ , θ , and ϕ calculations are presented in supplementary materials. However, one needs to be aware of the fact, that this approach gives two undistinguishable orientation results, mirror symmetrical with respect to the xy plane as described in supplementary materials. The order parameter ($\langle P_2 \rangle$), which is equal to the second-order coefficient of an ODF of a molecule with cylindrical symmetry, is used as a quantity describing the breadth of ODF.¹⁸ An example of 3D orientation results is shown in Figure 1e. It should be noted that all vectors have the same length and only vector directions are of significance. Since sections d and e of Figure 1 show results of corresponding sample regions, it is clearly visible how important advantage one gains using the method for 3D orientation determination. Even though this method was developed for FT-IR, the adaptation for O-PTIR and Raman spectra was done by transferring intensities to reach values corresponding to FT-IR absorbance, as it was briefly described in the preceeding section.

■ ASSOCIATED CONTENT

● Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/jacs.2c05306>.

3D orientation calculation approach; spectroscopic band identification; in-plane orientation (2D) results; 3D orientation results; exemplary polarized FT-IR spectra (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Tomasz P. Wrobel – Solaris National Synchrotron Radiation Centre, Jagiellonian University, 30-392 Krakow, Poland;
 orcid.org/0000-0003-2053-0570;
 Email: tomek.wrobel@uj.edu.pl

Authors

Paulina Koziol – Solaris National Synchrotron Radiation Centre, Jagiellonian University, 30-392 Krakow, Poland; Institute of Physics, Jagiellonian University, 30-348 Krakow, Poland

Karolina Kosowska – Solaris National Synchrotron Radiation Centre, Jagiellonian University, 30-392 Krakow, Poland

Danuta Liberda – Solaris National Synchrotron Radiation Centre, Jagiellonian University, 30-392 Krakow, Poland

Ferenc Borondics – Synchrotron SOLEIL, L'Orme Des Merisiers, 91192 Gif-Sur-Yvette, France; orcid.org/0000-0001-9975-4301

Complete contact information is available at:
<https://pubs.acs.org/10.1021/jacs.2c05306>

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was supported by the National Science Centre, Poland (Grant No. 2018/31/D/ST4/01833). Part of this study was performed using equipment purchased in the frame of the project co-funded by the Malopolska Regional Operational Program Measure 5.1 Krakow Metropolitan Area as an important hub of the European Research Area for 2007–2013, project no. MRPO.05.01.00-12-013/15. O-PTIR measurements were done at the SMIS beamline of the SOLEIL synchrotron facility. FT-IR measurements were in part done at the CIRI (SOLAIR) beamline of the SOLARIS synchrotron facility. The authors declare no conflicts of interest.

■ REFERENCES

- (1) Xu, Y.; Kraemer, D.; Song, B.; Jiang, Z.; Zhou, J.; Loomis, J.; Wang, J.; Li, M.; Ghasemi, H.; Huang, X.; Li, X.; Chen, G. Nanostructured Polymer Films with Metal-like Thermal Conductivity. *Nat. Commun.* **2019**, *10*, No. 1771.
- (2) Zhang, W.; Feng, P.; Chen, J.; Sun, Z.; Zhao, B. Electrically Conductive Hydrogels for Flexible Energy Storage Systems. *Prog. Polym. Sci.* **2019**, *88*, 220–240.
- (3) Kang, J.; Son, D.; Wang, G.; Niu, L.; Lopez, J.; Kim, Y.; Oh, J.; Y. Katsumata, T.; Mun, J.; Lee, Y.; Jin, L.; Tok, J. B.-H.; Bao, Z. Tough and Water-Insensitive Self-Healing Elastomer for Robust Electronic Skin. *Adv. Mater.* **2018**, *30*, No. 1706846.
- (4) Zhang, K.; Wang, S.; Zhou, C.; Cheng, L.; Gao, X.; Xie, X.; Sun, J.; Wang, H.; Wei, M.; D.; Reynolds, M.; A.; Zhang, N.; Bai, Y.; Xu, H. K. Advanced Smart Biomaterials and Constructs for Hard Tissue Engineering and Regeneration. *Bone Res.* **2018**, *6*, No. 31.
- (5) Hecht, E. *Optics*, 4th ed.; Addison-Wesley, 2012.
- (6) Wrobel, T.; P. Bhargava, R. Infrared Spectroscopic Imaging Advances as an Analytical Technology for Biomedical Sciences. *Anal. Chem.* **2018**, *90*, 1444–1463.
- (7) Zhang, D.; Li, C.; Zhang, C.; Slipchenko, M.; N.; Ekins, G.; Cheng, J.-X. Depth-Resolved Mid-Infrared Photothermal Imaging of Living Cells and Organisms with Submicrometer Spatial Resolution. *Sci. Adv.* **2016**, *2*, No. e1600521.
- (8) Su, Y.; Hu, X.; Tang, H.; Lu, K.; Li, H.; Lin, S.; Xing, B.; Ji, R. Steam Disinfection Releases Micro(Nano)Plastics from Silicone-Rubber Baby Teats as Examined by Optical Photothermal Infrared Microspectroscopy. *Nat. Nanotechnol.* **2022**, *17*, 76–85.
- (9) Looijmans, S. F. S.; P.; Carmeli, E.; Puskas, L.; Ellis, G.; Cavallo, D.; Anderson, P.; D.; Breemen, L. C. A Polarization Modulated Infrared Spectroscopy, A Pragmatic Tool for Polymer Science and Engineering. *Polym. Cryst.* **2020**, *3*, No. e10138.
- (10) Bocz, K.; Decsov, K.; E.; Farkas, A.; Vadas, D.; Bárány, T.; Wacha, A.; Bóta, A.; Marosi, G. Non-Destructive Characterisation of

All-Polypropylene Composites Using Small Angle X-Ray Scattering and Polarized Raman Spectroscopy. *Composites, Part A* **2018**, *114*, 250–257.

(11) Ryu, M.; Balçitis, A.; Wang, X.; Vongsavut, J.; Hikima, Y.; Li, J.; Tobin, M.; J.; Juodkazi, S.; Morikawa, J. Orientational Mapping Augmented Sub-Wavelength Hyper-Spectral Imaging of Silk. *Sci. Rep.* **2017**, *7*, No. 7419.

(12) Hikima, Y.; Morikawa, J.; Hashimoto, T. Wavenumber Dependence of FT-IR Image of Molecular Orientation in Banded Spherulites of Poly(3-Hydroxybutyrate) and Poly(L-Lactic Acid). *Macromolecules* **2013**, *46*, 1582–1590.

(13) Hikima, Y.; Morikawa, J.; Hashimoto, T. Imaging of Two-Dimensional Distribution of Molecular Orientation in Poly(Ethylene Oxide) Spherulite Using IR Spectrum and Birefringence. *Macromolecules* **2012**, *45*, 8356–8362.

(14) Wrobel, T.; P.; Mukherjee, P.; Bhargava, R. Rapid Visualization of Macromolecular Orientation by Discrete Frequency Mid-Infrared Spectroscopic Imaging. *Analyst* **2017**, *142*, 75–79.

(15) Honda, R.; Ryu, M.; Balçitis, A.; Vongsavut, J.; Tobin, M.; J.; Juodkazi, S.; Morikawa, J. Paracetamol Micro-Structure Analysis by Optical Mapping. *Appl. Surf. Sci.* **2019**, *473*, 127–132.

(16) Koziol, P.; Liberda, D.; Kwiatek, W.; M.; Wrobel, T. P. Macromolecular Orientation in Biological Tissues Using a Four-Polarization Method in FT-IR Imaging. *Anal. Chem.* **2020**, *92*, 13313–13318.

(17) Kosowska, K.; Koziol, P.; Liberda, D.; Wrobel, T. P. Spatially Resolved Macromolecular Orientation in Biological Tissues Using FT-IR Imaging. *Clin. Spectrosc.* **2021**, *3*, No. 100013.

(18) Lee, Y. J. Concurrent Polarization IR Analysis to Determine the 3D Angles and the Order Parameter for Molecular Orientation Imaging. *Opt. Express* **2018**, *26*, 24577–24590.

(19) Xu, S.; Rowlette, J.; Lee, Y. J. Imaging 3D Molecular Orientation by Orthogonal-Pair Polarization IR Microscopy. *Opt. Express* **2022**, *30*, 8436–8447.

(20) Xu, S.; Snyder, C. R.; Rowlette, J.; Lee, Y. J. Three-Dimensional Molecular Orientation Imaging of a Semicrystalline Polymer Film under Shear Deformation. *Macromolecules* **2022**, *55*, 2627–2635.

(21) Androsch, R.; Di Lorenzo, M. L.; Schick, C.; Wunderlich, B. Mesophases in Polyethylene, Polypropylene, and Poly(1-Butene). *Polymer* **2010**, *51*, 4639–4662.

(22) Buti, D.; Rosi, F.; Brunetti, B. G.; Milani, C. In-Situ Identification of Copper-Based Green Pigments on Paintings and Manuscripts by Reflection FTIR. *Anal. Bioanal. Chem.* **2013**, *405*, 2699–2711.

(23) Li, L.; Chan, C.-M.; Yeung, K. L.; Li, J.-X.; Ng, K.-M.; Lei, Y. Direct Observation of Growth of Lamellae and Spherulites of a Semicrystalline Polymer by AFM. *Macromolecules* **2001**, *34*, 316–325.

(24) Xu, J.; Guo, B.-H.; Zhang, Z.-M.; Zhou, J.-J.; Jiang, Y.; Yan, S.; Li, L.; Wu, Q.; Chen, G.-Q.; Schultz, J. M. Direct AFM Observation of Crystal Twisting and Organization in Banded Spherulites of Chiral Poly(3-Hydroxybutyrate- Co -3-Hydroxyhexanoate). *Macromolecules* **2004**, *37*, 4118–4123.

(25) Nguyen-Tr, P.; Ghassemi, P.; Carriere, P.; Nanda, S.; Assadi, A. A.; Nguyen, D. D. Recent Applications of Advanced Atomic Force Microscopy in Polymer Science: A Review. *Polymers* **2020**, *12*, No. 1142.

(26) Lee, Y. J. Determination of 3D Molecular Orientation by Concurrent Polarization Analysis of Multiple Raman Modes in Broadband CARS Spectroscopy. *Opt. Express* **2015**, *23*, 29279–29295.

(27) Crist, B.; Schultz, J. M. Polymer Spherulites: A Critical Review. *Prog. Polym. Sci.* **2016**, *56*, 1–63.

(28) Elzein, T.; Nasser-Eddine, M.; Delaite, C.; Bistac, S.; Dumas, P. FTIR Study of Polycaprolactone Chain Organization at Interfaces. *J. Colloid Interface Sci.* **2004**, *273*, 381–387.

(29) Raczkowska, M.; K.; Koziol, P.; Urbaniak-Wasik, S.; Paluszkievicz, C.; Kwiatek, W.; M.; Wrobel, T. P. Influence of Denoising on Classification Results in the Context of Hyperspectral

Data: High Definition FT-IR Imaging. *Anal. Chim. Acta* **2019**, *1085*, 39–47.

(30) Koziol, P.; Raczkowska, M. K.; Skibinska, J.; Urbaniak-Wasik, S.; Paluszkievicz, C.; Kwiatek, W.; Wrobel, T. P. Comparison of Spectral and Spatial Denoising Techniques in the Context of High Definition FT-IR Imaging Hyperspectral Data. *Sci. Rep.* **2018**, *8*, No. 14351.

(31) Griffiths, P. R.; de Haseth, J. A. *Fourier Transform Infrared Spectrometry*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2007.

(32) Atkins, P. de Paula Molecular Spectroscopy 1: Rotational and Vibrational Spectra. *Atkins' Physical Chemistry*, 8th ed.; W. H. Freeman and Company, 2006; pp 430–480.

(33) Hikima, Y.; Morikawa, J.; Hashimoto, T. FT-IR Image Processing Algorithms for In-Plane Orientation Function and Azimuth Angle of Uniaxially Drawn Polyethylene Composite Film. *Macromolecules* **2011**, *44*, 3950–3957.

Recommended by ACS

3D Orientation Imaging of Polymer Chains with Polarization-Controlled Coherent Raman Microscopy

Shuyu Xu, Young Jong Lee, *et al.*

DECEMBER 07, 2022
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

READ 

Imaging Orientation of a Single Molecular Hierarchical Self-Assembled Sheet: The Combined Power of a Vibrational Sum Frequency Generation Microscopy and Neural Network

Jackson C. Wagner, Wei Xiong, *et al.*

SEPTEMBER 13, 2022
THE JOURNAL OF PHYSICAL CHEMISTRY B

READ 

Cellulose Templating for π -Conjugated Polymer Orientation: An In Situ Time-Resolved Spectroscopy Exploration

Toshiki Sakata, Ken-ichi Saitow, *et al.*

SEPTEMBER 15, 2022
ACS APPLIED POLYMER MATERIALS

READ 

Three-Dimensional Molecular Orientation Imaging of a Semicrystalline Polymer Film under Shear Deformation

Shuyu Xu, Young Jong Lee, *et al.*

MARCH 31, 2022
MACROMOLECULES

READ 

Get More Suggestions >

Korekta rozpraszania typu Mie dla próbek o cylindrycznym kształcie

Badania nad orientacją tkanki włóknistej ujawniły jej wysoką skłonność do rozpraszania promieniowania podczerwonego, co bez zastosowania odpowiedniej korekty zaburza wyniki badań nad orientacją makromolekuł. Analizując rozmiar włókien kolagenu, obserwowany promień jest zbliżony do długości fali padającego promieniowania. Takie sprzyjające warunki na zajęcie rozpraszania typu Mie sugerują, że jest to źródło zaistniałych artefaktów. Korekta rozpraszania Mie przez obiekty sferyczne jest szeroko przestudiowana i posiada w literaturze wiele wariantów algorytmów umożliwiających korekty dla takich próbek jak np. komórki. Poza teorią nie było jednak dostępnej implementacji korekty rozpraszania dla próbek o cylindrycznym kształcie. Stało się to motywacją do zaaplikowania, a następnie zrealizowania grantu Preludium-18 pt. „Ulepszanie wyznaczania orientacji 3D makromolekuł na podstawie spolaryzowanego obrazowania chemicznego IR poprzez optymalizację algorytmów usuwania rozpraszania”.

W ramach badań, prowadzonych w trakcie trwania grantu, wykorzystano teoretyczny model opisujący wydajność rozpraszania dla obiektów cylindrycznych i zaimplementowano w oprogramowaniu MATLAB korektę rozpraszania bazującą na metodzie RMieS-EMSC [H]. Dostępny model rozpraszania rozszerzono o przypadek liniowej polaryzacji padającego promieniowania, co miało na celu jak najlepsze dostosowanie metody do przyszłych obliczeń związanych z wyznaczaniem orientacji molekuł [H]. Ze względu na intensywność obliczeń, powyższe implementacje wymagały zastosowania procesorów graficznych (GPU), które znacznie skróciły skalę czasową wykonywania korekty [H]. Ponadto, w celu wsparcia środowiska spektroskopowego, stworzona biblioteka jest udostępniona w otwartym repozytorium. Skuteczność metody została zweryfikowana z wykorzystaniem specjalnie przygotowanej próbki włókna polimeru o rozmiarach indukujących silne rozpraszanie typu Mie, a uzyskane rezultaty przewyższały inne zastosowane korekty [H].

Mój wkład w publikację [H] związany był ze sformułowaniem celu badawczego, zaplanowaniem eksperymentów, przeprowadzeniem pomiarów FT-IR uwzględniających polaryzację liniową oraz jej brak, aplikacją algorytmu EMSC do korekty próbek cylindrycznych, dostosowaniem algorytmu do zastosowań związanych liniową polaryzacją, adaptacją algorytmów do obliczeń z wykorzystaniem procesorów graficznych, wykonanie korekt dla danych eksperymentalnych, analizą danych i porównaniem do referencyjnej metody, pisanie części manuskryptu oraz przygotowaniem i udostępnieniem biblioteki funkcji na zasadach otwartego dostępu.

Scattering correction for samples with cylindrical domains measured with polarized infrared spectroscopy

Paulina Koziol^{1,2}, Karolina Kosowska¹, Pawel Korecki^{1,2}, Tomasz P. Wrobel^{1*}

¹ SOLARIS National Synchrotron Radiation Centre, Jagiellonian University, Czerwone Maki 98, 30-392 Krakow, Poland.

² Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, Lojasiewicza 11, 30-348 Krakow, Poland

*email: tomek.wrobel@uj.edu.pl

Abstract

Scattering artifacts are one of the most common effects distorting transmission spectra in Fourier-Transform Infrared spectroscopy. Their increased impact, strongly diminishing the quantitative and qualitative power of IR spectroscopy, is especially observed for structures with a size comparable to the radiation wavelength. To tackle this problem, a wide range of preprocessing techniques based on the Extended Multiplicative Scattering Correction method was developed, using physical properties to remove scattering presence in the spectra. However, until recently those algorithms were mostly focused on spherically shaped samples, for example, cells. Here, an algorithm for samples with cylindrical domains is described, with additional implementation of a linearly polarized light case, which is crucial for the growing field of polarized IR imaging and spectroscopy. An open-source code with GPU based implementation is provided, with a calculation time of several seconds per spectrum. Optimizations done to improve the throughput of this algorithm allow the application of this method into the standard preprocessing pipeline of small datasets.

Keywords

Scattering correction, Linear Polarization, FT-IR Imaging, Cylindrical domains, EMSC

Introduction

Fourier-Transform Infrared Spectroscopy (FT-IR) and Imaging are rapidly developing and highly utilized methods, mostly due to the wealth of biochemical information encoded in infrared (IR) spectra. Reasonable speed of measurements, high Signal-to-Noise Ratio (SNR) and good sensitivity cause IR based techniques to be widely used in such fields as material science [1], [2], biology and medicine [3]–[5], archeology [6], [7], environmental science [8], forensic [9]–[11] and others. Apart from qualitative analysis targeted at compound identification, there is also a strong quantitative aspect of IR spectroscopy. Both areas of applications demand spectra to contain only absorption related information. Unfortunately, this requirement is very often not fulfilled because of the scattering artifacts. Such artifacts are especially pronounced for the spherical and cylindrical samples with a size comparable to the wavelength which is often the case for cells, tissue fibers and others. Scattering theory for spherical samples and small particles [12], [13] has been developed decades ago and explains the basics of the artifacts observed in IR spectra, allowing to distinguish between absorptive and scattering origin of absorbance variations. Many implementations have been developed with e.g. Extended Multiplicative Signal Correction (EMSC) algorithms based on full Mie theory or van de Hulst approximation targeted at scattering effects removal [14]–[17]. In the past, those algorithms mostly focused on cells, which due to their morphology, can create perfect conditions for Mie scattering. Yet, over time, they were also extended to aid scattering by tissue samples [18]–[20]. By nature, EMSC based algorithms are relatively computationally expensive, which despite the hardware development, makes their application on a large scale still problematic. Thus new advances using Neural Networks and augmentation approaches are also explored [21], [22].

A second important geometry for analysis, that is gaining attention, are samples with cylindrical domains [23]–[25]. From collagen fibers in tissues, hair to microfibers, the cylindrical shape will distort the analytical IR signal. The theoretical approach for cylinder's scattering is well described [26], [27]. However, the application of full Mie type scattering theory for infinite cylinders in EMSC algorithm was done only lately by Rasskazov et al. [28], considering unpolarized light and including a new workflow to optimize computation time. Yet, the polarized light case was still not addressed, which, considering cylinder's reduced symmetry, cannot be omitted as it is in the case of spherically symmetrical samples. Figure 1 presents how predominant are scattering artifacts for cylindrical samples considering unpolarized (a), parallelly (b), and perpendicularly (c) polarized light. This is an example of a Polycaprolactone (PCL) fiber model sample, specifically made for this study, with its dimensions strongly inducing scattering artifacts.

IR studies with linear polarization control are currently of interest [29], as many researchers apply it in molecular orientation studies [30]–[33], including studies focused on cylindrically shaped samples [34]. Not only spectra themselves are relevant, but more advanced analysis of the integral intensity of vibrational bands allow calculation of in-plane orientation functions and more recently, retrieval of full 3D angles of a transition moment orientation [35]–[37]. Considering that, there is a growing need for proper correction of polarized IR light scattering from cylindrical domains, in this research, an open-source, EMSC based algorithm for scattering artifacts correction in cylindrical samples has been developed. The algorithm considers different states of linear polarizations as an extension of the unpolarized light case reported by Rasskazov et al. [28]. Optimization of the potentially extended calculation time was done via GPU based calculations allowing to introduce this method into the preprocessing pipeline of relatively small datasets. The developed algorithm was also tested with an isolated collagen fiber sample. An open-source, GPU based, version of MATLAB implementation is available, and referenced in the Code availability section.

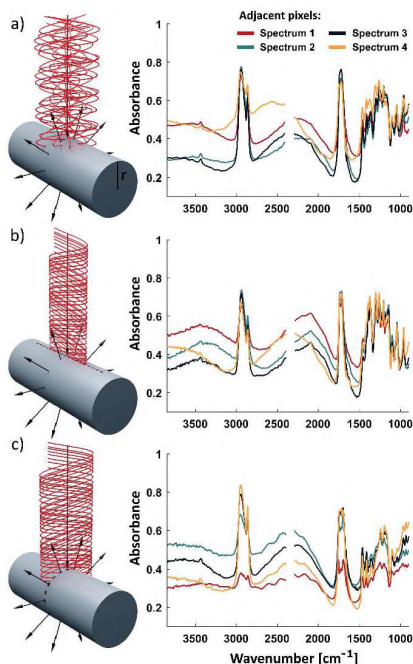


Figure 1. Spectra from a PCL fiber sample highlighting scattering artifacts, for unpolarized (a), parallel (b), and perpendicular (c) polarization of incident IR light with respect to the cylinder's axis. Spectra of each color come from the same location (the same chemical composition) as all datasets are spatially aligned.

Materials and Methods

PCL fiber sample preparation

Polycaprolactone (PCL) of $M_n = 80\,000$ g/mol was purchased from Sigma Aldrich Sp. z o.o. (Poznan, Poland). Chloroform for analysis (99.8%), amylene stabilized was obtained from Idalia Sp. j. (Radom, Poland). In the first step, a polymer solution was obtained by dissolving 0.2 g of PCL in 40 mL of chloroform. The system was left on a magnetic stirrer for 24 h at room temperature to allow complete dissolution. The solution was cast onto a glass laboratory dish and placed in a fume hood for evaporation for the next 24 h. The obtained thin film was melted at 80°C as PCL has a melting point of about 60°C . At the same time, a calcium fluoride (CaF_2) slide was heated up to 60°C . The thin PCL fiber was manually pulled out using a steel needle from the melt and placed on the CaF_2 slide. Next, the sample was rapidly cooled down to room temperature so the fiber would not melt.

Tissue fiber sample

Breast tissue microarray (TMA) was purchased from Biomax Inc. One slice was placed on a BaF₂ plate for transmission measurements, whereas the adjacent slice was mounted on a regular glass and H&E stained for visual inspection and structures identification. An isolated fiber from one of the tissue cores was selected for further FT-IR analysis.

IR measurements and data preprocessing

Samples were measured in transmission mode with Bruker Vertex80v spectrometer coupled to a Hyperion3000 Microscope with 64x64 FPA MCT detector and 15x objective (NA = 0.4) giving 2.7 μ m projected pixel size. Spectral range cover region from 3850 cm⁻¹ to 900 cm⁻¹, with 8 cm⁻¹ resolution and zero filling factor of 2, giving a spectrum with 765 points. Data were collected for unpolarized light and with four linear polarizations: 0°, 45°, 90°, and 135° (corresponding to the fiber's orientation). The stage was specifically programmed to repeat the same positions, thus all five datasets (for each sample) are spatially aligned (to the accuracy of the stage movement, i.e. below 1 μ m) and spectral information might be compared between measurements. During PCL fiber measurements sample and background signals were co-averaged 32 and 64 times, respectively, whereas tissue sample was scanned 4 times for the signal and 64 times for the background. Background measurements were done individually for each polarization. Datasets were individually denoised with Principal Component Analysis with 15 components used for signal reconstruction [38].

Resonant Mie scattering EMSC

The current status of EMSC based algorithms for Mie scattering have decades of development behind them, starting from simple cases to later on extending the original algorithm to be more accurate and tackle more realistic cases in terms of physics. In that context, Resonant Mie Scattering (RMieS) EMSC algorithm has been relatively recently developed, as a result of the progress in the comprehension of dispersion artifact [15], [39], responsible for the sharp decrease in absorbance band on the high wavenumber side. As described by Bassan *et al.*, a change in real refractive index near the absorption band causes a reduction of resonant Mie Scattering efficiency [15], [39]. Thus, the non-resonant Mie scattering algorithm, previously proposed by Kohler *et al.* [16], was extended by the implementation of the Kramers-Kronig (KK) transform for real refractive index calculations. An example of the KK transform for the determination of the real refractive index n based on IR spectrum for PCL is presented in Figure 2b. A simplification has been done by denoting IR spectrum as k , due to its relationship with the imaginary part (absorptive) of the complex refractive index. This, along with ranges of possible values for average refractive index and particle radius are later used to calculate a set of scattering efficiency Q_{sca} curves – one curve for each specific combination of mentioned parameters.

General model for RMieS-EMSC allows to decompose (correct) recorded signal $\vec{Z}_{Raw}$ as follows [14], [15]:

$$\vec{Z}_{Raw} = \vec{c} + m\vec{v} + h\vec{Z}_{Ref} + \sum_{i=1}^{N=7} g_i \vec{p}_i + \vec{E}, \quad (1)$$

where $\vec{c}$ is the constant baseline, m is the gradient of the sloping baseline, $\vec{v}$ is the wavenumber vector, h is the magnifying factor of the reference spectrum $\vec{Z}_{Ref}$, N is the number of principal components (7 gives satisfactory level of explained variance), $\vec{p}_i$ and $\vec{g}_i$ are loading and its weights (describing scattering efficiency impact), respectively, and $\vec{E}$ are residuals. In general, correction of the recorded spectrum is done by finding values of c , m , h and g_i using linear regression. However, this is an iterative process, graphically presented in Figure 2a. The first iteration requires a reference spectrum for the refractive index calculations, whereas, further iterations use information from the previous step. Crucial step in each iteration is finding $\vec{p}_i$ (loadings) which contain the scattering information. This is done by

PCA decomposition of the set of scattering curves Q_{sca} , usually 1000 curves, calculated for permutations of physical parameters such as particle radius and refractive index. This RMieS-EMSC algorithm has been thoroughly described in previous publications, thus more detailed explanation can be found in [14], [15].

Scattering efficiency Q_{sca} for cylindrical samples

Scattering efficiency calculations are the crucial part of described RMieS-EMSC method, providing physical background for the correction algorithm. Up to recently, most papers focused on Mie scattering correction for spherical samples, as it has a wide range of applications, for example during cells measurements. In this case, two possible options for Mie scattering efficiency calculations are available. First of all, using full Mie theory [12], which is physically more accurate, but at the same time very computationally expensive due to the need for Bessel's spherical functions calculations [14]. On the other hand, there is a van der Hulst approximation [40] giving major a time advantage, yet, being less accurate [15], [16].

Currently, a focus has been pointed on the extension of the algorithm to an infinite cylinders case, which could be a good approximation of fibrous samples. Recently, Rasskazov et al. published an EMSC based approach where the theory for an infinite cylinders scattering is applied [28]. One of the differences from spherical samples is that cylinders have rotational symmetry, thus, light polarization can no longer be omitted in this case. Even though neither discussed nor applied, theoretical approach for cylinders' scattering of linearly polarized light was provided in [28] and it is as follows:

$$Q_{sca,\parallel} = \frac{2}{x} \left[|b_{0,\parallel}|^2 + 2 \sum_{l=1}^{\infty} (|b_{l,\parallel}|^2 + |a_{l,\parallel}|^2) \right], \quad (2)$$

$$Q_{sca,\perp} = \frac{2}{x} \left[|a_{0,\perp}|^2 + 2 \sum_{l=1}^{\infty} (|a_{l,\perp}|^2 + |b_{l,\perp}|^2) \right], \quad (3)$$

where $Q_{sca,\parallel}$ and $Q_{sca,\perp}$ are scattering efficiencies per infinite cylinder unit length for the parallel and transversal illumination, respectively, x is a size parameter and l is a spherical function order. Other parameters - a_l and b_l along with x are described in the supplementary data. Presented formulas are also used to calculate the scattering efficiency for unpolarized light [28]:

$$Q_{sca} = \frac{1}{2} [Q_{sca,\parallel} + Q_{sca,\perp}]. \quad (4)$$

An extension for linear polarizations other than parallel and transversal is also provided in Supplementary Materials. An example of scattering efficiencies (Q_{sca} , $Q_{sca,\perp}$ and $Q_{sca,\parallel}$) for a theoretical PCL cylinder with 7.5 μm radius is shown in Figure 2c. It is also compared to a scattering of a spherical sample of the same radius, calculated through the van de Hulst approximation used in RMieS-EMSC [15]. It is clearly visible, that there is a major difference between sphere's and cylinder's scattering and how different polarization cases influence scattering. Note that for a cylinder, a high frequency fine structure resulting for scattering is superimposed on the spectrum. Such a high frequency signal may overlap with absorption bands and hinder quantitative spectral analysis. More examples of quantitative and qualitative differences in spheres' and cylinders' scattering character, including the impact of the varying angle of incidence ζ , are provided in [28].

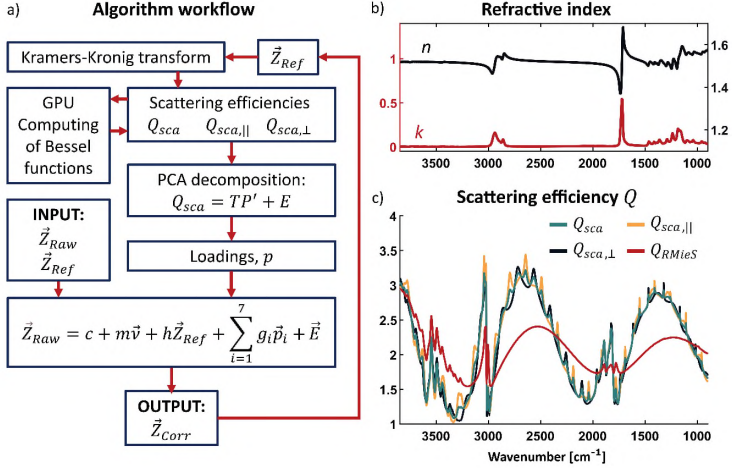


Figure 2. a) Workflow of the Cylinders EMSC algorithm. b) Examples of the real refractive index calculated from IR spectrum of PCL (k) using KK transform. c) Theoretical scattering efficiency for cases of cylindrical sample (unpolarized Q_{sca} and with parallel $Q_{sca,||}$ and perpendicular $Q_{sca,\perp}$ polarizations) and spherical sample Q_{RMieS} of the same radius.

Algorithm development and GPU acceleration

The development of a scattering correction algorithm for cylindrical samples was based on an open source code provided by Bassan et al. for RMieS-EMSC. The key part was to replace van de Hulst approximation of Q for spheres, with the scattering formulas for cylinders described in the previous paragraphs. Also, a necessary parameter ζ , describing the angle of incidence was added and the normal incidence was assumed for all further calculations ($\zeta = 90^\circ$). A general workflow of the new algorithm is presented in Figure 2a. Apart from data to be corrected, its input requires a reference spectrum, with PCL film spectrum (almost free from scattering) used in this case.

Implementation of the full theory for scattering efficiency (Mie type) calls for advanced programming solutions due to the time and computing power needed for Bessel's spherical functions calculations. Rasskazov et al. [28] suggested to use a new way based on interpolation of Q_{sca} from data coarsely simulated for a wide range of x and mean refractive index. Here, a classical approach with GPU calculations was applied, considering major progress in graphic card development. It is worth mentioning, that not only Bessel and Hankel's functions were transferred to GPU, but also other parts of code required for Q_{sca} calculations (differentiation and coefficients calculations). This required further optimizations to apply code vectorization, but more information is provided and discussed in the Supplementary Material.

Calculations and time estimation were done using two machines:

- 1) Desktop working station Dell Precision 5820 with Intel Core i9-10900X processor, 256 GB RAM memory and NVIDIA Quadro P6000 24 GB GPU
- 2) Server Dell PowerEdge R740 with 2x Intel Xeon Gold 6254, 12x 64GB RDIMM RAM memory and single NVIDIA Tesla v100 32GB GPU.

As the Quadro P6000 GPU is not an optimal card for double precision scientific calculations, the algorithms were developed using the first (desktop) computing station, but later their final performance was tested with a Tesla v100 card, dedicated to clusters in supercomputing centers. Assuming a spectrum of 765 variables, the calculation time per single spectrum with a Tesla card was 3.98 s, which gives 0.398 s per iteration (10 iterations), and 9.62 s for Quadro (0.96 s per iteration). This time scale is already optimized, but it will give even faster results for samples with smaller radii (the size parameter defines the order of Bessel's functions as described in Supplementary Materials). Here, the calculations were done assuming the cylinder's radius varied within the 5 to 15 μm range. Whereas, for smaller samples, for example from 2 to 8 μm range, the calculation time was 2.6 s per spectrum for Tesla.

Results and discussion

As was mentioned in the Materials and Method section, a model PCL fiber sample with a radius varying from 7.5 μm to 12 μm , was crafted specifically to exhibit strong scattering properties. An optical microscope image of the sample is presented in Figure 3a. The goal here is to test the developed algorithms for unpolarized and linearly polarized IR radiation using a highly scattering model sample, observing at the same time if the scattering pattern is similar to scattering efficiency calculated theoretically. Additionally, results are compared to the RMieS-EMSC method for spherical samples, which also has an open source code available.

Unpolarized IR radiation

Figure 3 presents the correction done for the unpolarized light case. As the cylindrical sample has a varying size (radius varying from 7.5 to 12 μm), it makes perfect conditions to analyze scattering with different physical conditions (different radii). Here, a case of 7.5 μm and 11 μm radius was analyzed and is shown in Figure 3 a-d and Figure 3 e-h, respectively. Based on theoretically calculated scattering efficiency for a 7.5 μm radius of the cylindrical sample shown in Figure 3a, major scattering artifacts were expected to appear in the experimental spectra. This is exactly what is observed for non-corrected spectra coming from the cylinder's cross-section and shown in Figure 3b (legend for spectra is provided in Figure 3a, along with spatial location marked on IR image). As expected, strong scattering is observed for those spectra, having the specific shape predicted by Q curves. Artifacts are most prominent in the 3000-1700 cm^{-1} region, with a small number of absorption bands. However, the fingerprint region is also strongly distorted, although, the distortion is covered by absorption bands. After visual inspection of correction results, it might be concluded that the algorithm eliminated most of the scattering component (Figure 3c), which is indicated by the absence of characteristic shape and strong intensity reduction observed around 1000 cm^{-1} . Moreover, Cylinders EMSC results look very similar to the RMieS-EMSC algorithm results (Figure 3d) with subtle qualitative and qualitative differences in absorption bands' intensities. Both algorithms had a problem with spectrum number 5 (grey), as the strong scattering region falls around CH_2 absorption bands. RMieS-EMSC seems to be doing slightly better in this case. On the other hand, both algorithms seem to struggle in the 3500-3000 cm^{-1} region, with Cylinders EMSC doing a little better. Very similar tendencies are observed for a larger radius case. Theoretical scattering efficiency curves changed their sine-shaped characteristics for higher frequency (Figure 3e), which is also clearly observed in the experimental spectra (Figure 3f). Smaller wavy distortions appearing on top of a main characteristic, became sharper and narrower. Thus, less likely to be observed in the spectra. As more spectra have a scattering peak around the CH_2 band, both correction algorithms are not ideal in this case, with Cylinders EMSC doing definitely better in the 3500-2500 cm^{-1} region. Overall, considering extremely high scattering artifacts (enhanced by almost perfect physical conditions for their occurrence) present in experimental spectra, the developed algorithm for cylindrical samples is highly effective. Its introduction into the preprocessing pipeline allows performing quantitative analysis using corrected spectra – which was not possible before.

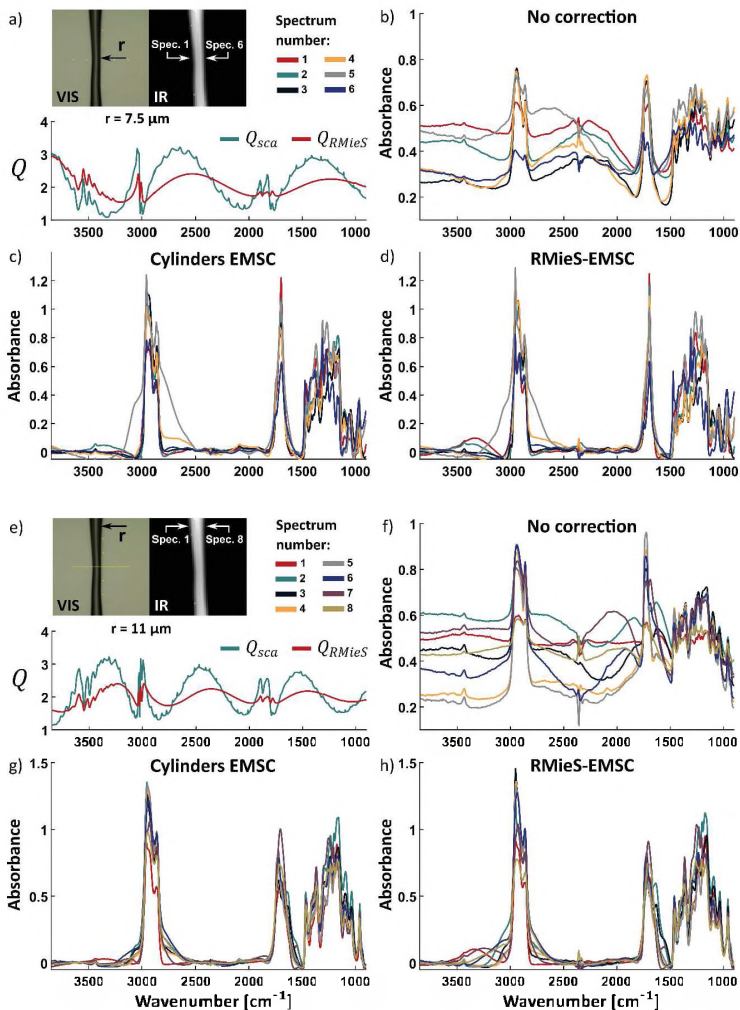


Figure 3. Scattering correction results for model PCL cylinder sample, for two cases of sample's radius. Spectra from cylinder's cross-sections are analyzed from left to right, with 2.7 μm step (pixel size). a)-d) Results for 7.5 μm radius region of the sample. e)-h) Results for 11 μm radius region of the sample. Legends for spectra in plots b)-d) and f)-h) are provided in sections a) and e) respectively. Scales on visible and IR images are the same and the dimensions might be referenced using the fiber's radius.

Linear polarization of IR radiation

Analysis similar to the one presented for unpolarized light was conducted for linear polarization of incident IR radiation case. Again, scattering from the cylindrical sample with radii equal to 7.5 μm and 11 μm is analyzed. However, instead of the cylinders cross section, spectra from the same pixel coming from different polarization datasets are presented in Figure 4. This requires for cylinder's orientation to be well defined to determine the polarization orientation. Interesting differences in scattering efficiency are observed for different polarization cases (Figure 4a and e). Those differences might be crucial for the spectroscopic bands' appearance in experimental absorption spectra. Observed structures also show how distinct is RMieS-EMSC scattering model for spheres and expose the need for cylindrical model applications. The characteristic structure (wave) most observed for perpendicular polarization can be observed in Figure 1c. In both cases, raw spectra show extremely high scattering artifacts with the sinusoidal pattern well predicted by the theory. The second case of 11 μm radius exhibits a higher frequency of scattering pattern changes (Figure 4f) than is observed in Figure 4b. Also, as the spectra come from the same spatial location, the phase of the sinusoidal shape is very similar for all polarization datasets. Analyzing the effectiveness of the algorithms, PCylinders EMSC seems to be doing definitely better than RMieS-EMSC, especially in the 3500-3000 cm^{-1} region. Furthermore, for the 11 μm radius case, the carbonyl stretching region around 1720 cm^{-1} seems to be problematic for both algorithms (Figure 4 g-h), but with PCylinders EMSC doing better.

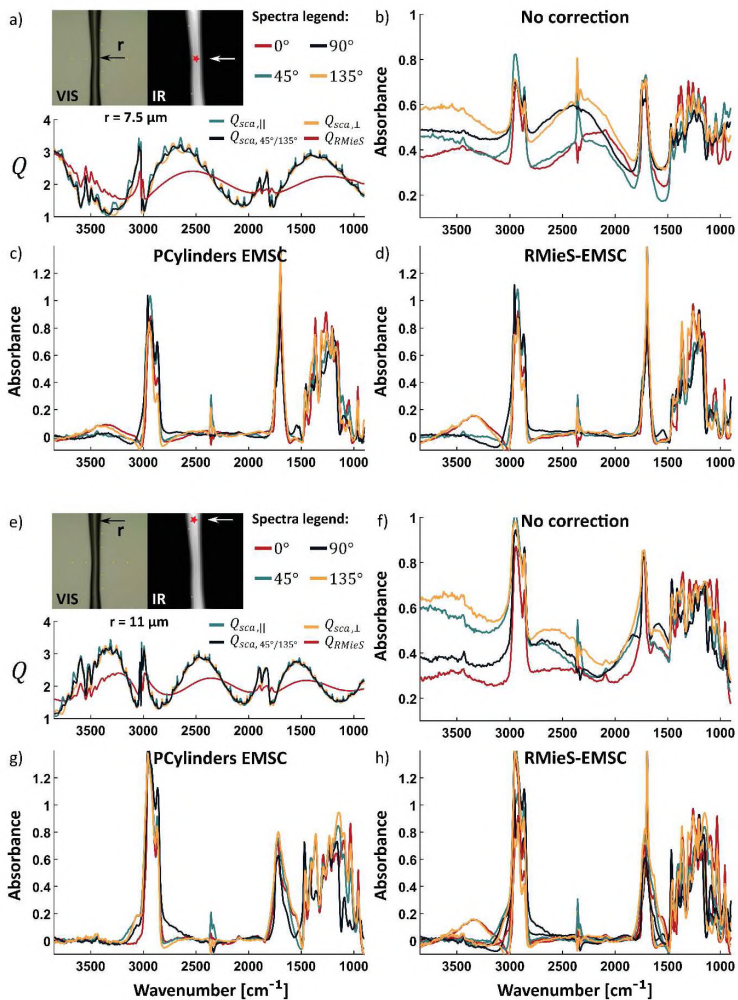


Figure 4. Scattering correction results for linearly polarized light. A model PCL cylinder sample with varying radii was used as a reference and a single pixel (marked with star) with different polarizations is shown. a)-d) Results for 7.5 μm radius region of the sample. e)-h) Results for 11 μm radius region of sample. Legends for spectra in plots b)-d) and f)-h) are provided in sections a) and e) respectively. Scales on visible and IR images are the same and the dimensions might be referenced using the fiber's radius.

PCylinders EMSC algorithm implementation with linear polarization of incident light was tested using a tissue sample. The region of isolated collagen fiber, presented in Figure 5a, was used for evaluation due to its high scattering properties. Spectra from a single pixel measured with different linear polarization are presented in Figure 5b. Despite the fact that strong scattering is observed in this case, there is no specific sinusoidal shape in the baseline. However, spectra perpendicular (90°) to the fiber's axis show the biggest scattering level, whereas parallel one (0°) is the smallest. 45° and 135° polarization have very similar scattering characteristics – as expected since the physical conditions are the same. The PCylinders EMSC algorithm manages to correct those spectra perfectly, which is mostly visible in the $2700\text{--}1800\text{ cm}^{-1}$ region (Figure 5c), and provides slightly better correction than RMieS-EMSC (Figure 5d). Furthermore, as polarized light absorption depends on its alignment with the dipole moment, all spectra should show variability in the absorbing parts, especially in fingerprint regions. In general, PCylinders EMSC seems to be doing very well in the case of tissue fiber, preserving absorptive differences in spectra coming from polarization orientation with respect to dipole moments orientation. This highlights the potential of this method in molecular orientation studies, based on polarized FT-IR.

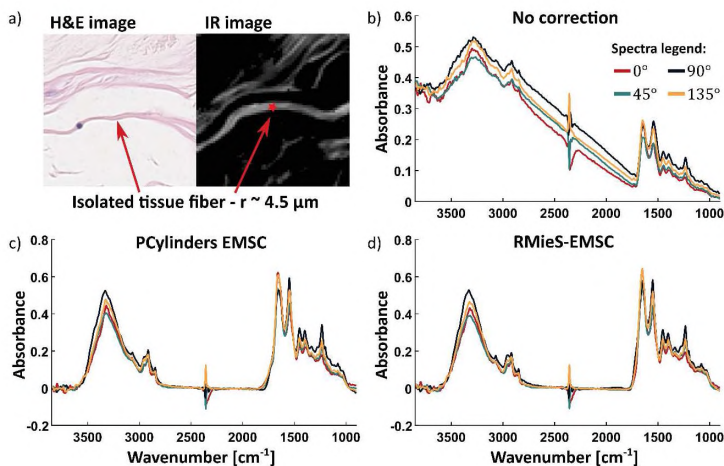


Figure 5. Scattering correction results for tissue fiber sample (single pixel marked with star) and linearly polarized light. a) Tissue sample used to validate the developed algorithm. b) Raw spectra coming from the single pixel (marked with a red star) and different polarization sets. c) Correction results for PCylinders EMSC algorithm. d) Correction results for RMieS-EMSC algorithm.

Conclusions

As the polarized FT-IR field is rapidly gaining new applications, especially in molecular orientation studies, targeted preprocessing methods considering physical properties and parameters need to be developed. Therefore, an EMSC based scattering correction algorithm using full Mie-type scattering theory for samples with cylindrical domains (Cylinders EMSC) and considering the state of light linear

polarization (PCylinders EMSC) was developed, with MATLAB open source code available. Due to computational requirements, GPU was used for the code acceleration resulting in a calculation time of several seconds per spectrum. This allows the introduction of this algorithm into a standard preprocessing pipeline of small FT-IR datasets. Algorithm performance was tested with a specifically prepared PCL model sample, exhibiting strong scattering properties, along with isolated tissue fiber. Comparison with the widely used RMieS-EMSC algorithm shows an advantage of the new algorithm.

Code availability

The developed code along with FT-IR data for model PCL fiber sample is available at https://github.com/paulinakozio/Cylinders_EMSC, under open source license (GNU Lesser General Public License v3.0).

Acknowledgements

This research was financed by the National Science Centre, Poland (“Improving 3D macromolecule orientation determination based on polarized IR chemical imaging by optimization of scattering removal algorithms”, Grant No. 2019/35/N/ST4/02481).

FT-IR measurements were performed at CIRI beamline under the provision of the Polish Ministry and Higher Education project Support for research and development with the use of research infra-structure of the National Synchrotron Radiation Centre SOLARIS” under contract nr 1/SOL/2021/2. This research was also supported by the NVIDIA Academic Hardware Grant Program providing a Quadro P6000 GPU unit. NSRC SOLARIS infrastructure was used for the Tesla v100 GPU based computations.

References

- [1] Y. Zhang *et al.*, “Revealing electrolyte oxidation: Via carbonate dehydrogenation on Ni-based oxides in Li-ion batteries by in situ Fourier transform infrared spectroscopy,” *Energy Environ. Sci.*, vol. 13, no. 1, pp. 183–199, 2020, doi: 10.1039/c9ee02543j.
- [2] X. Li *et al.*, “Waterproof-breathable PTFE nano- and Microfiber Membrane as High Efficiency PM2.5 Filter,” *Polymers (Basel)*, vol. 11, no. 4, 2019, doi: 10.3390/polym11040590.
- [3] M. J. Baker *et al.*, “Using Fourier transform IR spectroscopy to analyze biological materials,” *Nat. Protoc.*, vol. 9, no. 8, pp. 1771–91, Aug. 2014, doi: 10.1038/nprot.2014.110.
- [4] T. P. Wrobel and R. Bhargava, “Infrared Spectroscopic Imaging Advances as an Analytical Technology for Biomedical Sciences,” *Anal. Chem.*, vol. 90, no. 3, pp. 1444–1463, 2018, doi: 10.1021/acs.analchem.7b05330.
- [5] H. Tieman, B. Byrne, and S. G. Kazarian, “ATR-FTIR spectroscopy and spectroscopic imaging for the analysis of biopharmaceuticals,” *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.*, vol. 241, no. 118636, 2020, doi: 10.1016/j.saa.2020.118636.
- [6] S. H. Bayari, K. Özdemir, E. H. Sen, C. Araujo-Andrade, and Y. S. Erdal, “Application of ATR-FTIR spectroscopy and chemometrics for the discrimination of human bone remains from different archaeological sites in Turkey,” *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.*, vol. 237, 2020, doi: 10.1016/j.saa.2020.118311.
- [7] E. Smidt *et al.*, “Infrared spectroscopy refines chronological assessment, depositional environment and pyrolysis conditions of archeological charcoals,” *Sci. Rep.*, vol. 10, no. 12427, 2020, doi: 10.1038/s41598-020-69445-6.
- [8] S. Veerasingam *et al.*, “Contributions of Fourier transform infrared spectroscopy in microplastic pollution research: A review,” *Crit. Rev. Environ. Sci. Technol.*, vol. 51, no. 22, pp. 2681–2743, 2021, doi: 10.1080/10643389.2020.1807450.

- [9] S. Giuliano, E. Mistek-Morabito, and I. K. Lednev, "Forensic phenotype profiling based on the attenuated total reflection Fourier transform-infrared spectroscopy of blood: Chronological age of the donor," *ACS Omega*, vol. 5, no. 42, pp. 27026–27031, 2020, doi: 10.1021/acsomega.0c01914.
- [10] E. Mistek-Morabito and I. K. Lednev, "Discrimination between human and animal blood by attenuated total reflection Fourier transform-infrared spectroscopy," *Commun. Chem.*, vol. 3, no. 1, pp. 1–6, 2020, doi: 10.1038/s42004-020-00424-8.
- [11] T. Iwai *et al.*, "Forensic Discrimination of Drug Powder Based on Drug Mixing Condition Determined Using Micro Fourier Transform Infrared Spectroscopy," *ACS Omega*, 2022, doi: 10.1021/acsomega.2c07573.
- [12] G. Mie, *Contributions to the optics of turbid media, particularly of colloidal metal solutions*. 1908.
- [13] "Light scattering by small particles. By H. C. van de Hulst. New York (John Wiley and Sons). London (Chapman and Hall), 1957. Pp. xiii, 470; 103 Figs.; 46 Tables. 96s." *Q. J. R. Meteorol. Soc.*, 1958, doi: 10.1002/qj.49708436025.
- [14] P. Bassan *et al.*, "RMieS-EMSC correction for infrared spectra of biological cells: extension using full Mie theory and GPU computing," *J. Biophotonics*, vol. 3, no. 8–9, pp. 609–20, Aug. 2010, doi: 10.1002/jbio.201000036.
- [15] P. Bassan *et al.*, "Resonant Mie scattering (RMieS) correction of infrared spectra from highly scattering biological samples," *Analyst*, vol. 135, no. 2, pp. 268–77, Feb. 2010, doi: 10.1039/b921056c.
- [16] A. Kohler *et al.*, "Estimating and correcting Mie scattering in synchrotron-based microscopic fourier transform infrared spectra by extended multiplicative signal correction," *Appl. Spectrosc.*, vol. 62, no. 3, pp. 259–266, 2008, doi: 10.1366/000370208783759669.
- [17] T. Konevskikh, R. Lukacs, R. Blumel, A. Ponomosov, and A. Kohler, "Mie scatter corrections in single cell infrared microscopy spectroscopy," *Faraday Discuss.*, vol. 187, pp. 235–257, Apr. 2016, doi: 10.1039/c5fd00171d.
- [18] J. H. Solheim, E. Gunko, D. Petersen, F. Großeruschkamp, K. Gerwert, and A. Kohler, "An open-source code for Mie extinction extended multiplicative signal correction for infrared microscopy spectra of cells and tissues," *J. Biophotonics*, vol. 12, no. 8, pp. 1–14, 2019, doi: 10.1002/jbio.201800415.
- [19] T. Van Dijk, D. Mayerich, P. S. Carney, and R. Bhargava, "Recovery of absorption spectra from fourier transform infrared (FT-IR) microspectroscopic measurements of intact spheres," *Appl. Spectrosc.*, vol. 67, no. 5, pp. 546–552, 2013, doi: 10.1366/12-06847.
- [20] T. Konevskikh, R. Lukacs, and A. Kohler, "An improved algorithm for fast resonant Mie scatter correction of infrared spectra of cells and tissues," *J. Biophotonics*, vol. 11, no. 1, 2018, doi: 10.1002/jbio.201600307.
- [21] U. Blazhko, V. Shapaval, V. Kovalev, and A. Kohler, "Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra," *Chemom. Intell. Lab. Syst.*, vol. 215, no. 104367, 2021, doi: 10.1016/j.chemolab.2021.104367.
- [22] E. A. Magnussen *et al.*, "Deep convolutional neural network recovers pure absorbance spectra from highly scatter-distorted spectra of cells," *J. Biophotonics*, vol. 13, no. 12, 2020, doi: 10.1002/jbio.202000204.
- [23] Y. Huo *et al.*, "In Vivo Stable Allogenic Cartilage Regeneration in A Goat Model Based on Immunoisolation Strategy Using Electrospun Semipermeable Membranes," *Adv. Healthc. Mater.*, vol. n/a, no. n/a, p. 2203084, doi: <https://doi.org/10.1002/adhm.202203084>.

- [24] C. Jia *et al.*, “Highly compressible and anisotropic lamellar ceramic sponges with superior thermal insulation and acoustic absorption performances,” *Nat. Commun.*, vol. 11, no. 1, pp. 1–13, 2020, doi: 10.1038/s41467-020-17533-6.
- [25] S. Woo, H. R. Park, J. Park, J. Yi, and W. Hwang, “Robust and continuous oil/water separation with superhydrophobic glass microfiber membrane by vertical polymerization under harsh conditions,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, 2020, doi: 10.1038/s41598-020-78271-9.
- [26] S. Zhang, W. Zhang, J. Dong, and L. Liu, “Optical theorem of an infinite circular cylinder in weakly absorbing media,” *Phys. Rev. A*, vol. 105, no. 2, pp. 1–12, 2022, doi: 10.1103/PhysRevA.105.023516.
- [27] B. J. Davis, P. Scott Carney, and R. Bhargava, “Theory of infrared microspectroscopy for intact fibers,” *Anal. Chem.*, vol. 83, no. 2, pp. 525–532, 2011, doi: 10.1021/ac102239b.
- [28] I. L. Rasskazov, R. Singh, P. S. Carney, and R. Bhargava, “Extended Multiplicative Signal Correction for Infrared Microspectroscopy of Heterogeneous Samples with Cylindrical Domains,” *Appl. Spectrosc.*, vol. 73, no. 8, pp. 859–869, 2019, doi: 10.1177/0003702819844528.
- [29] G. Bakir *et al.*, “Orientation Matters: Polarization dependent IR spectroscopy of collagen from intact tendon down to the single fibril level,” *Molecules*, vol. 25, no. 18, 2020, doi: 10.3390/molecules25184295.
- [30] Y. Hikima, J. Morikawa, and T. Hashimoto, “FT-IR Image Processing Algorithms for In-Plane Orientation Function and Azimuth Angle of Uniaxially Drawn Polyethylene Composite Film,” *Macromolecules*, vol. 44, no. 10, pp. 3950–3957, May 2011, doi: 10.1021/ma2003129.
- [31] Y. Hikima, J. Morikawa, and T. Hashimoto, “Imaging of Two-Dimensional Distribution of Molecular Orientation in Poly(ethylene oxide) Spherulite Using IR Spectrum and Birefringence,” *Macromolecules*, vol. 45, no. 20, pp. 8356–8362, Oct. 2012, doi: 10.1021/ma3010372.
- [32] P. Koziol, D. Liberda, W. M. Kwiatek, and T. P. Wrobel, “Macromolecular Orientation in Biological Tissues Using a Four-Polarization Method in FT-IR Imaging,” *Anal. Chem.*, vol. 92, no. 19, pp. 13313–13318, Oct. 2020, doi: 10.1021/acs.analchem.0c02591.
- [33] T. P. Wrobel, P. Mukherjee, and R. Bhargava, “Rapid visualization of macromolecular orientation by discrete frequency mid-infrared spectroscopic imaging,” *Analyst*, vol. 142, no. 1, pp. 75–79, 2017, doi: 10.1039/c6an01086e.
- [34] M. Ryu *et al.*, “Orientational Mapping Augmented Sub-Wavelength Hyper-Spectral Imaging of Silk,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, 2017, doi: 10.1038/s41598-017-07502-3.
- [35] P. Koziol, K. Kosowska, D. Liberda, F. Borondics, and T. P. Wrobel, “Super-Resolved 3D Mapping of Molecular Orientation Using Vibrational Techniques,” *J. Am. Chem. Soc.*, vol. 144, no. 31, pp. 14278–14287, 2022, doi: 10.1021/jacs.2c05306.
- [36] S. Xu, J. Rowlette, and Y. J. Lee, “Imaging 3D molecular orientation by orthogonal-pair polarization IR microscopy,” *Opt. Express*, vol. 30, no. 6, pp. 8436–8447, Mar. 2022, doi: 10.1364/OE.449667.
- [37] Y. J. Lee, “Concurrent polarization IR analysis to determine the 3D angles and the order parameter for molecular orientation imaging,” *Opt. Express*, vol. 26, no. 19, p. 24577, 2018, doi: 10.1364/OE.26.024577.
- [38] P. Koziol *et al.*, “Comparison of spectral and spatial denoising techniques in the context of High Definition FT-IR imaging hyperspectral data,” *Sci. Rep.*, vol. 8, no. 1, pp. 1–11, 2018, doi: 10.1038/s41598-018-32713-7.

- [39] P. Bassan, H. J. Byrne, F. Bonnier, J. Lee, P. Dumas, and P. Gardner, “Resonant Mie scattering in infrared spectroscopy of biological materials--understanding the ‘dispersion artefact’ .,” *Analyst*, vol. 134, no. 8, pp. 1586–93, Aug. 2009, doi: 10.1039/b904808a.
- [40] H. C. van de Hulst, “Light scattering by small particles.,” *New York (John Wiley Sons)*, 1957, doi: 10.1002/qj.49708436025.

Referencje

- [1] K. Malek, Ed., *Vibrational spectroscopy from theory to practice*. Warszawa: Wydawnictwo Naukowe PWN, 2016.
- [2] K. Yeh, M. Schulmerich, and R. Bhargava, "Mid-infrared microspectroscopic imaging with a quantum cascade laser," *Proc. SPIE*, vol. 8726, no. 217, p. 87260E, May 2013, doi: 10.1117/12.2015984.
- [3] P. Bassan, M. J. Weida, J. Rowlette, and P. Gardner, "Large scale infrared imaging of tissue micro arrays (TMAs) using a tunable Quantum Cascade Laser (QCL) based microscope.," *Analyst*, vol. 139, pp. 3856–3859, 2014, doi: 10.1039/c4an00638k.
- [4] D. Zhang, C. Li, C. Zhang, M. N. Slipchenko, G. Eakins, and J.-X. Cheng, "Depth-resolved mid-infrared photothermal imaging of living cells and organisms with submicrometer spatial resolution," *Sci. Adv.*, vol. 2, no. 9, pp. 1–8, 2016.
- [5] L. Wang and X. G. Xu, "Scattering-type scanning near-field optical microscopy with reconstruction of vertical interaction," *Nat. Commun.*, vol. 6, 2015, doi: 10.1038/ncomms9973.
- [6] S. G. Stanciu *et al.*, "Scattering-type Scanning Near-Field Optical Microscopy of Polymer-Coated Gold Nanoparticles," *ACS Omega*, vol. 7, no. 13, pp. 11353–11362, 2022, doi: 10.1021/acsomega.2c00410.
- [7] K. G. Wirth *et al.*, "Tunable s-SNOM for Nanoscale Infrared Optical Measurement of Electronic Properties of Bilayer Graphene," *ACS Photonics*, vol. 8, no. 2, pp. 418–423, 2021, doi: 10.1021/acsp Photonics.0c01442.
- [8] A. Dazzi, R. Prazeres, F. Glotin, and J. M. Ortega, "Local infrared microspectroscopy with subwavelength spatial resolution with an atomic force microscope tip used as a photothermal sensor," *Opt. Lett.*, vol. 30, no. 18, pp. 2388–2390, 2005, doi: 10.1364/OL.30.002388.
- [9] A. Dazzi and C. B. Prater, "AFM-IR: Technology and applications in nanoscale infrared spectroscopy and chemical imaging," *Chem. Rev.*, vol. 117, no. 7, pp. 5146–5173, 2017, doi: 10.1021/acs.chemrev.6b00448.
- [10] D. A. McQuarrie and J. D. Simon, *Physical Chemistry: A Molecular Approach*. University Science Books, 1997.
- [11] P. R. Griffiths and J. A. De Haseth, *Fourier Transform Infrared Spectrometry, Second Edition*. John Wiley & Sons, Inc., 2007.
- [12] P. W. Atkins, *Chemia fizyczna*. Wydawnictwo Naukowe PWN, 2012.
- [13] B. J. Davis, P. S. Carney, and R. Bhargava, "Theory of Mid-infrared Absorption Microspectroscopy : II . Heterogeneous Samples," *Anal. Chem.*, vol. 82, no. 9, pp. 3487–3499, 2010, doi: 10.1021/ac902067p.10.1021/ac902068e.
- [14] B. J. Davis, P. S. Carney, and R. Bhargava, "Theory of Midinfrared Absorption Microspectroscopy : I. Homogeneous Samples," *Anal. Chem.*, vol. 82, no. 9, pp. 3474–3486, 2010, doi: 10.1021/ac902068e.
- [15] E. Hecht, *Optics (4th Edition)*. Addison-Wesley, 2012.
- [16] R. K. Reddy, M. J. Walsh, M. V. Schulmerich, P. S. Carney, and R. Bhargava, "High-definition infrared spectroscopic imaging.," *Appl. Spectrosc.*, vol. 67, no. 1, pp. 93–105, Jan. 2013, doi: 10.1366/11-06568.

- [17] M. C. Branciforti, L. B. Silva, and R. E. S. Bretas, "Molecular orientation of extruded PET/LCP blend films. Part I: Polarized infrared spectroscopy," *J. Appl. Polym. Sci.*, vol. 102, no. 3, pp. 2241–2248, 2006, doi: 10.1002/app.24395.
- [18] C. Duchesne, X. Kong, J. Brisson, M. Pérolet, and R. E. Prud'homme, "Molecular orientation and relaxation of poly(ethylene terephthalate) by polarization Modulation infrared spectroscopy," *Macromolecules*, vol. 35, no. 23, pp. 8768–8773, 2002, doi: 10.1021/ma020469+.
- [19] K. Yeh, S. Kenkel, J. Liu, and R. Bhargava, "Fast Infrared Chemical Imaging with a Quantum Cascade Laser," *Anal. Chem.*, vol. 87, no. 1, pp. 485–493, 2015, doi: 10.1021/ac5027513.
- [20] M. R. Kole, R. K. Reddy, M. V. Schulmerich, M. K. Gelber, and R. Bhargava, "Discrete frequency infrared microspectroscopy and imaging with a tunable quantum cascade laser.," *Anal. Chem.*, vol. 84, no. 23, pp. 10366–72, Dec. 2012, doi: 10.1021/ac302513f.
- [21] M. J. Baker *et al.*, "Using Fourier transform IR spectroscopy to analyze biological materials.," *Nat. Protoc.*, vol. 9, no. 8, pp. 1771–91, Aug. 2014, doi: 10.1038/nprot.2014.110.
- [22] C. Kuepper, A. Kallenbach-Thieltges, H. Juetz, A. Tannapfel, F. Großeschenkamp, and K. Gerwert, "Quantum Cascade Laser-Based Infrared Microscopy for Label-Free and Automated Cancer Classification in Tissue Sections," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, 2018, doi: 10.1038/s41598-018-26098-w.
- [23] J. A. Reffner, "Advances in Infrared Microspectroscopy and Mapping Molecular Chemical Composition at Submicrometer Spatial Resolution," *Spectroscopy*, vol. 33, no. 9, pp. 12–17, 2018.
- [24] M. Kansiz, L. M. Dowling, I. Yousef, O. Guaitella, F. Borondics, and J. Sulé-Suso, "Optical Photothermal Infrared Microspectroscopy Discriminates for the First Time Different Types of Lung Cells on Histopathology Glass Slides," *Anal. Chem.*, vol. 93, no. 32, pp. 11081–11088, 2021, doi: 10.1021/acs.analchem.1c00309.
- [25] R. Gautam, S. Vanga, F. Ariese, and S. Umapathy, "Review of multidimensional data processing approaches for Raman and infrared spectroscopy," *EPJ Tech. Instrum.*, vol. 2, no. 8, 2015, doi: 10.1140/epjti/s40485-015-0018-6.
- [26] P. Lasch, "Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging," *Chemom. Intell. Lab. Syst.*, vol. 117, pp. 100–114, 2012, doi: 10.1016/j.chemolab.2012.03.011.
- [27] H. L. Mark and P. R. Griffiths, "Analysis of Noise in Fourier Transform Infrared Spectra," *Appl. Spectrosc.*, vol. 56, no. 5, pp. 633–639, 2002.
- [28] P. Deepa and M. Suganthi, "Performance Evaluation of Various Denoising Filters for Medical Image," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 4205–4209, 2014, [Online]. Available: www.ijcsit.com.
- [29] A. K. Jain, "Fundamentals of digital image processing. Anil K. Jain.," *Vision*, no. November. Prentice Hall, p. 569, 1989.
- [30] W. H. Press and S. A. Teukolsky, "Savitzky-Golay Smoothing Filters," *Comput. Phys.*, vol. 4, no. 6, pp. 669–672, 1990, doi: 10.1063/1.4822961.
- [31] B. Zimmermann and A. Kohler, "Optimizing savitzky-golay parameters for improving spectral resolution and quantification in infrared spectroscopy," *Appl. Spectrosc.*, vol. 67, no. 8, pp. 892–902, 2013, doi: 10.1366/12-06723.
- [32] B. K. Alsberg, A. M. Woodward, M. K. Winson, J. Rowland, and D. B. Kell, "Wavelet denoising of infrared spectra," *Analyst*, vol. 122, no. 7, pp. 645–652, 1997, doi:

- [33] S. D. Ruikar and D. D. Doye, "Wavelet Based Image Denoising Technique," *Int. J. Adv. Comput. Sci. Appl.*, vol. 3, no. 2, 2011, doi: 10.14569/IJACSA.2011.020309.
- [34] F. L. Gewers *et al.*, "Principal component analysis: A natural approach to data exploration," *ACM Comput. Surv.*, vol. 54, no. 4, 2021, doi: 10.1145/3447755.
- [35] B. M. Salih Hasan and A. M. Abdulazez, "A Review of Principal Component Analysis Algorithm for Dimensionality Reduction," *J. Soft Comput. Data Min.*, vol. 02, no. 01, pp. 20–30, 2021, doi: 10.30880/jscdm.2021.02.01.003.
- [36] L. Zhang, W. Dong, D. Zhang, and G. Shi, "Two-stage image denoising by principal component analysis with local pixel grouping," *Pattern Recognit.*, vol. 43, no. 4, pp. 1531–1549, 2010, doi: 10.1016/j.patcog.2009.09.023.
- [37] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1–3, pp. 37–52, 1987, doi: 10.1016/0169-7439(87)80084-9.
- [38] D. P. Berrar, W. Dubitzky, and M. Granzow, Eds., *A Practical Approach to Microarray Data Analysis*. KLUWER ACADEMIC PUBLISHERS, 2003.
- [39] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "A Transformation for Ordering Multispectral Data in Terms of Image Quality with Implications for Noise Removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 1, pp. 65–74, 1988, doi: 10.1109/36.3001.
- [40] S. Gupta, S. Mittal, A. Kajdacsy-Balla, R. Bhargava, and C. Bajaj, "A fully automated, faster noise rejection approach to increasing the analytical capability of chemical imaging for digital histopathology," *PLoS One*, vol. 14, no. 4, pp. 1–15, 2019, doi: 10.1371/journal.pone.0205219.
- [41] D. Liberda, P. Koziol, M. K. Raczkowska, W. M. Kwiatek, and T. P. Wrobel, "Influence of interference effects on the spectral quality and histological classification by FT-IR imaging in transfection geometry," *Analyst*, vol. 146, no. 2, pp. 646–654, 2021, doi: 10.1039/d0an01565b.
- [42] J. Filik, M. D. Frogley, J. K. Pijanka, K. Wehbe, and G. Cinque, "Electric field standing wave artefacts in FTIR micro-spectroscopy of biological materials," *Analyst*, vol. 137, no. 4, pp. 853–861, 2012, doi: 10.1039/c2an15995c.
- [43] P. Bassan *et al.*, "RMieS-EMSC correction for infrared spectra of biological cells: extension using full Mie theory and GPU computing," *J. Biophotonics*, vol. 3, no. 8–9, pp. 609–20, Aug. 2010, doi: 10.1002/jbio.201000036.
- [44] P. Bassan *et al.*, "Resonant Mie scattering (RMieS) correction of infrared spectra from highly scattering biological samples," *Analyst*, vol. 135, no. 2, pp. 268–77, Feb. 2010, doi: 10.1039/b921056c.
- [45] P. Bassan, H. J. Byrne, F. Bonnier, J. Lee, P. Dumas, and P. Gardner, "Resonant Mie scattering in infrared spectroscopy of biological materials—understanding the 'dispersion artefact'," *Analyst*, vol. 134, no. 8, pp. 1586–93, Aug. 2009, doi: 10.1039/b904808a.
- [46] J. H. Solheim, E. Gunko, D. Petersen, F. Großeruschkamp, K. Gerwert, and A. Kohler, "An open-source code for Mie extinction extended multiplicative signal correction for infrared microscopy spectra of cells and tissues," *J. Biophotonics*, vol. 12, no. 8, pp. 1–14, 2019, doi: 10.1002/jbio.201800415.
- [47] I. L. Rasskazov, R. Singh, P. S. Carney, and R. Bhargava, "Extended Multiplicative Signal Correction for Infrared Microspectroscopy of Heterogeneous Samples with Cylindrical Domains," *Appl. Spectrosc.*, vol. 73, no. 8, pp. 859–869, 2019, doi: 10.1177/0003702819844528.

- [48] Y. Hikima, J. Morikawa, and T. Hashimoto, "FT-IR Image Processing Algorithms for In-Plane Orientation Function and Azimuth Angle of Uniaxially Drawn Polyethylene Composite Film," *Macromolecules*, vol. 44, no. 10, pp. 3950–3957, May 2011, doi: 10.1021/ma2003129.
- [49] Y. Hikima, J. Morikawa, and T. Hashimoto, "Imaging of Two-Dimensional Distribution of Molecular Orientation in Poly(ethylene oxide) Spherulite Using IR Spectrum and Birefringence," *Macromolecules*, vol. 45, no. 20, pp. 8356–8362, Oct. 2012, doi: 10.1021/ma3010372.
- [50] Y. Hikima, J. Morikawa, and T. Hashimoto, "Wavenumber dependence of FT-IR image of molecular orientation in banded spherulites of poly(3-hydroxybutyrate) and poly(l-lactic acid)," *Macromolecules*, vol. 46, no. 4, pp. 1582–1590, 2013, doi: 10.1021/ma302560q.
- [51] T. P. Wrobel, P. Mukherjee, and R. Bhargava, "Rapid visualization of macromolecular orientation by discrete frequency mid-infrared spectroscopic imaging," *Analyst*, vol. 142, no. 1, 2017, doi: 10.1039/c6an01086e.
- [52] M. Ryu *et al.*, "Orientalional Mapping Augmented Sub-Wavelength Hyper-Spectral Imaging of Silk," *Sci. Rep.*, vol. 7, no. 1, Dec. 2017, doi: 10.1038/s41598-017-07502-3.
- [53] Y. J. Lee, "Concurrent polarization IR analysis to determine the 3D angles and the order parameter for molecular orientation imaging," *Opt. Express*, vol. 26, no. 19, p. 24577, 2018, doi: 10.1364/oe.26.024577.
- [54] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
- [55] R. Genuer, J. M. Poggi, C. Tuleau-Malot, and N. Villa-Vialaneix, "Random Forests for Big Data," *Big Data Res.*, vol. 9, pp. 28–46, 2017, doi: 10.1016/j.bdr.2017.07.003.
- [56] R. Genuer, J. M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognit. Lett.*, vol. 31, no. 14, pp. 2225–2236, 2010, doi: 10.1016/j.patrec.2010.03.014.