Science, Technology and Innovation
ISSN 2544-9125 ▪ 2023; 17 (1-2): 9-29
doi: 10.55225/sti.492

STI • 2023;17 (1-2): 9-29

FTIR fingerprint — testing a new representation of the binary fingerprint based on FTIR spectra in the prediction of physicochemical properties

Kacper Tomaszewski[1] B-F, Rafał Kurczab[2] A-D,G

[1]University of Applied Sciences in Tarnow, Faculty of Mathematics and Natural Sciences, Department of Chemistry, Mickiewicza 8, 33-100 Tarnów, Poland
[2]Maj Institute of Pharmacology, Polish Academy of Sciences, Smętna 12, 31-343 Krakow, Poland

Original article

Abstract

The paper deals with the development of a new method for the generation of binary fingerprints based on the Savitzky-Golay (SG) algorithm and first-order derivatives of FTIR spectra, which are then used to create prediction models for selected the physicochemical properties of chemical compounds. Models based on the FEDS (Functionally-Enhanced Derivative Spectroscopy) transformation and raw spectra were used as a reference to determine whether the use of the SG filter and first-order derivatives was worth to further develop. The FTIR spectra of 103 compounds with theoretically determined values of logP, logD and logS were studied. The Tanimoto coefficient and correlation coefficient were used to compare the fingerprints obtained, while the root mean square error (RMSE) was used to assess the quality of the prediction models. Based on the results, it was found that the use of the SG filter and derivatives had a positive effect on the quality of the prediction models for logP and logS, and a negative effect on the quality of the models for logD, compared to the models based on original spectra and FEDS transformation.

Keywords

- binary fingerprint
- FTIR spectroscopy
- Savitzky-Golay filter
- FEDS
- prediction models
- physicochemical properties

Authors contributions

A – Preparation of the research project
B – Assembly of data for the research undertaken
C – Conducting of statistical analysis
D – Interpretation of results
E – Manuscript preparation
F – Literature review
G – Revising the manuscript

Corresponding author

Rafał Kurczab
e-mail: kurczab@if-pan.krakow.pl
Instytut Farmakologii im. J. Maja PAN
ul. Smętna 12
31-343 Kraków, Poland

Article info

Article history

- Received: 2023-04-23
- Accepted: 2023-05-25
- Published: 2023-09-15

Publisher

University of Applied Sciences in Tarnow
ul. Mickiewicza 8, 33-100 Tarnow, Poland

Financing

This research did not received any grants from public, commercial or non-profit organizations.

Conflict of interest

None declared.

Introduction

In recent years chemistry has evolved from a field based solely on direct work with chemical substances and the use of instrumental methods to science that makes use of the latest advances in mathematics, computer science, and many other branches of science. By using appropriate mathematical calculations and the computational power of modern computers, the process of making new substances in the pharmaceutical, food or chemical industries can be simplified. The combination of well-known instrumental methods with suitable (in silico) mathematical methods offers almost unlimited possibilities [1].

A fingerprint is one of many methods of representing chemical compounds, in this case in the form of a sequence of bits [2,3], where zeros indicate the absence of a given property (i.e. substructure) and ones indicate the presence of a given property. Molecular fingerprints are generally not a unified way of representing a chemical compound (unlike SMILES or SMARTS codes), and up to date many different ways of generating fingerprints have been reported [4,5]. One of the most important application of molecular fingerprints is to use them to create mathematical models to predict physicochemical or biological properties. The development of a relatively reliable model has the potential to simplify the process of developing new therapeutic agents, for example, by predicting some of the properties of a compound without the need for long-lasting, and expensive experimental studies.

One of the biggest obstacles in analyzing a digital signal (e.g. a spectroscopic spectrum or an electrocardiogram) is noise. Noise not only makes analysis difficult, but it also affects the visual aspects of presenting the results. In the case of spectroscopic spectra, they can appear as fluctuations in absorbance (or transmittance). Noise is a type of signal distortion caused by, among other things, errors in the detector itself [6]. For these reasons, it is not possible to eliminate noise from a digital signal by, for example, calibrating the instrument or taking many measurements. One method of eliminating noise is to filter the received signal using an appropriate numerical algorithm. One of the most popular digital filters is the Savitzky-Golay (SG) filter [7]. Due to its simplicity and efficiency, it has found enormous applications in analytical chemistry or chemometrics. It was popularized by Abraham Savitzky and Marcel J.E. Golay in 1964, when they published tables of convolution coefficients for various polynomials and subset sizes [8]. Thanks to the popularity of the Savitzky-Golay filter, the original paper was recognized by the journal Analytical Chemistry as the fifth-best paper published in the journal [9].

The basis of the Savitzky-Golay filter is convolution, which fits a low-degree polynomial to a given point and a predetermined number of neighboring points (window width) using a least-squares method [10]. In practice, this means that the SG filter extracts from a given data set a small subset, centered on the point under study and a predetermined number of neighboring points on either side of said point. A polynomial with a pre-selected degree is then fitted to such a subset (therefore, the width of the window determines the power with which the signal will be smoothed). The SG filter iterates the above steps for each point in the set.

The use of derivatives as a tool for the analysis of spectroscopic spectra dates back to the 1950s. However, it was not until the 1970s that their development and popularity increased [11]. They offer a great deal of convenience in the analysis of spectra, including the ability to easily locate significant parts of the absorption bands. These can include the locations of peaks, inflection points and absorption band edges.

One of the major obstacles in the analysis of infrared spectroscopic spectra is the spectral band overlap (SBO) phenomenon [12,13]. The reason for this phenomenon is the occurrence of absorption bands of different chemical bonds in the same or similar wavenumber intervals. It is particularly noticeable among bonds between the same atoms at different groups of compounds (for example, the N-H bond in amides and amines). There are methods to minimize the effect of SBO, including:

- increasing resolution and minimizing noise;
- modifications to the test sample (e.g. change of solvent);
- mathematical transformations, i.e.: derivatives (the most commonly used are second-order derivatives) [14] and deconvolution [15].

One method of FTIR spectra transformation is Functionally-Enhanced Derivative Spectroscopy (FEDS). The main objective of FEDS is to separate the bands and simplify the spectrum by narrowing the individual bands without significantly changing their position. This is achieved by creating a *P*-function from a series of simple functions [16].

$$P_i = (1 + A_i)\left(\left|\frac{1}{A_i}\right| - \left|\frac{1}{A_{i-1}}\right|\right)^{-\frac{1}{2}} \quad (1)$$

where:
P_i – the *P*-function for the *i*-th point,
A_i – the absorbance for the *i*-th point.

Materials and methods

Data set

The FTIR spectra of 103 compounds (the full list is in Appendix 1) were obtained in our previous work [17,18] and were used in this research. Also, the results for the methods examined in that work were used for comparative purposes.

A Nicolet™ iS™ 5 FTIR spectrophotometer was used to measure the spectra. Each spectrum was obtained by averaging 16 scans taken at a resolution of 2 cm^{-1} over the range of 4000–650 cm^{-1}. The measurements were recorded using OMNIC software, while further processing (smoothing and derivation of the spectrum) was performed using RStudio. Smoothing and derivation were performed with custom scripts using R libraries, i.e. prospectr and rootSolve. Predictive models were developed using the KNIME (KNstanz Information MinEr) environment (Appendix 3). Theoretical and experimental physicochemical properties used for the research were fetched from the chemspider.com website (accessed April 7, 2022).

FTIR fingerprint algorithm

The algorithm of FTIR-based molecular fingerprint generation was proposed in our previous work [17,18]. Herein, we tested the performance of this algorithm by adding a preprocessing step for the original FTIR spectra. Thus, the generation of the binary fingerprints in the present work was performed according to the following protocol:

1. Preprocessing of the FTIR spectra (smoothing and differentiation).
2. Locating the roots of the derivative spectrum (roots refer to the location of the peak in the spectrum).
3. Matching of localized root positions to pre-selected wavenumber intervals corresponding to peak locations for individual absorption bands.
4. Generation of spectral fingerprint (f_s). If the location (i.e. wavenumber) of the root matches any of the wavenumber intervals, the bit corresponding to that wavenumber interval has the value of '1'.
5. Comparison of the spectral fingerprint with a molecular fingerprint (f_m), which is based on the molecular structure of the compound. For example, if the compound is an alcohol, the C–O bond will result in the presence of a bit at positions 80 and 83 in the molecular fingerprint. A table listing our definition of the substructures for each position in the f_m is in Appendix 2.
6. The agreement of a given bit in both fingerprints results in the presence of a bit ('1') in the final fingerprint, while the absence of a bit in one or both fingerprints in a given position results in the absence of a bit ('0') in the final fingerprint.

Based on the obtained fingerprints and the known values of logP, logS, and logD, the predictive models were built using linear regression and regression tree algorithm. A genetic feature (variable) preselection strategy with iterations of 1000 and a population size of 50 was used for the calculations for both the regression tree and the linear regression algorithm. In addition, the data for linear regression was pre-processed using PCA (Principal Component Analysis) with a target dimension of 50. The quality of a given prediction model was assessed by comparing the true value of a given physicochemical parameter with its predicted value. This was obtained by the use of the root mean square error (RMSE) parameter.

Results and discussion

Basic methods of FTIR spectra preprocessing

Firstly, the FTIR fingerprints were investigated based on pure experimental spectra modified by a first-order derivative without any pre-processing. However, this type of approach renders the use of spectroscopic spectra completely meaningless, as all the noise and interference present in the spectrum are interpreted as absorption bands. The spectral fingerprint obtained in this way was characterized by a high and unjustified number of bits on ('1'), since the number of roots in the derivative spectrum is also high (as shown in Figure 2, the differentiation of the pure spectrum resulted in 293 roots).

Based on these results, it can be concluded that the use of a suitable filter is necessary to eliminate as much noise as possible. In the present study, it was decided to use an SG filter because of the nature of the noise present in the experimental spectra. If the spectrum is measured correctly, the noises that can be observed are local fluctuations in absorbance that tend to accumulate at the end of the spectrum and in areas where there are no absorption bands. During the research, it was concluded that direct application of the SG filter across the spectrum is a significant complication due to the different densities of absorption bands in different areas of the spectrum.

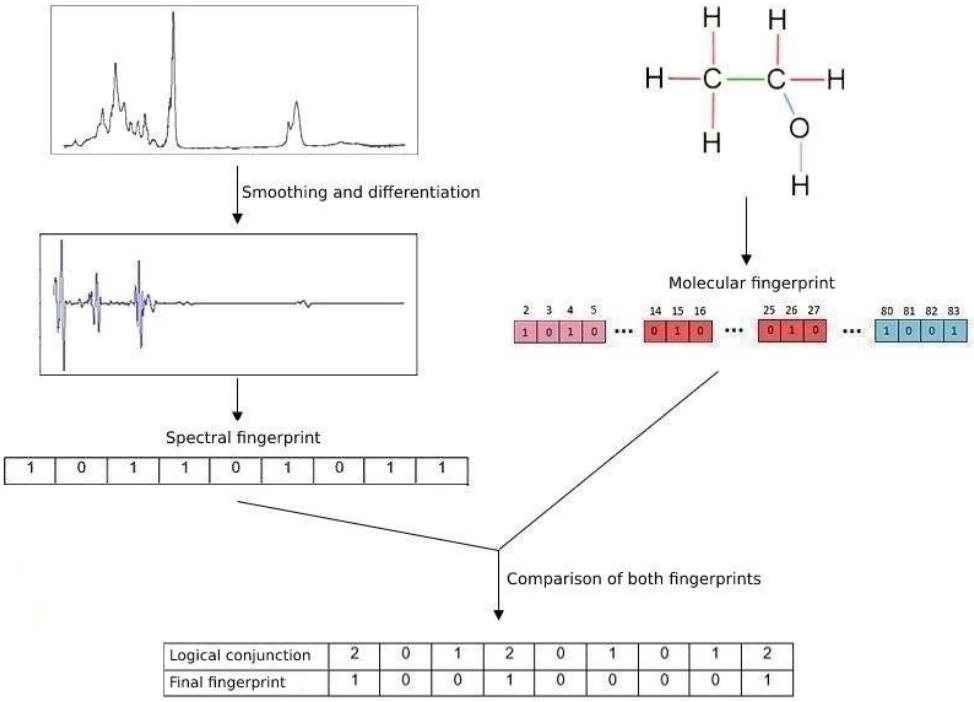


Figure 1. FTIR fingerprint generation scheme

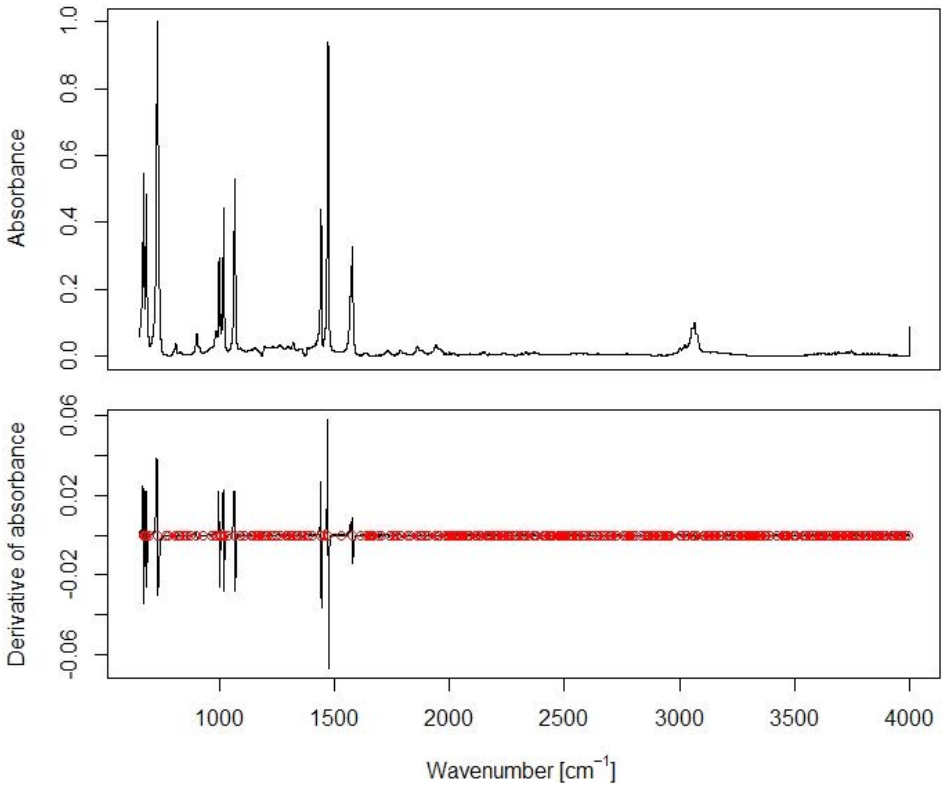


Figure 2. FTIR spectrum and its 1st derivative of bromobenzene. Red dots indicate roots

In general, FTIR spectra can be divided into two regions: the fingerprint region, which has no specific conventional limit but is considered in this article to be between 650 and 1500 cm^{-1}, and the functional group region, which is between 1500 and 4000 cm^{-1} [19]. The fingerprint region is characterized by a dense packing of absorption bands [20], so the application of a high-power filter can cause the merging of the separate absorption bands. However, for the functional group region, which is characterized by a lower packing of absorption bands and a higher presence of noise, a relatively higher power filter would be preferable. To avoid this phenomenon when processing the spectra, the two areas were processed separately with different window width.

After a thorough visual analysis of the spectra, as well as an examination of the influence of the window width on the number of detected roots in the fingerprint and the function group area, it was concluded that the best smoothing effect with the SG filter is achieved for the window width of 200 assigned to the fingerprint area, and 400 to the function group area.

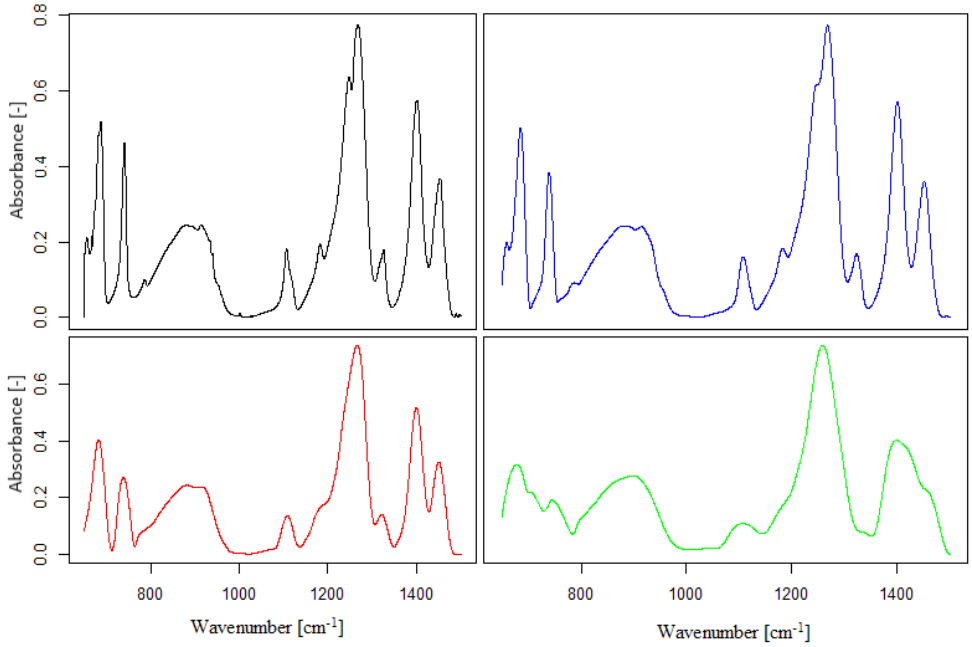


Figure 3. Comparison of the influence of the window width used in the SG filter on the shape of the FTIR fingerprint region. Black indicates a slice of the original spectrum, blue with a window width of 100, red with a window width of 200, and green with a window width of 400

Advanced methods of FTIR spectra preprocessing

Secondly, four methods of processing the FTIR spectra were developed and for each method, two cut-off levels were used to generate spectral fingerprints. The selection criterion was to make the cut-offs as representative as possible, e.g. if both the higher and lower cut-offs did not significantly affect the number of bits in the final fingerprint or were similar in terms of the Tanimoto coefficient, both were omitted:

a) Baseline (BS) – all absorbance values below a given cut-off are removed (Figure 4). Used cut-offs are 0.15 and 0.10;
b) Pre-Smoothing Fragmentation (PrSF) – the spectrum is split into fragments which are either smoothed using the regular SG filter (window width = 400) or with a far stronger SG filter (window width = 1500), if the amplitude of the values in a given fragment is less than the cut-off (Figure 5). Used cut-offs are 0.15 and 0.25;
c) Post-Smoothing Fragmentation (PoSF) – the smoothed spectrum is split into fragments and the given fragment is either left unchanged or all absorbance values in the fragment are set to 0 if the amplitude of the values in a given fragment is less than the cut-off (Figure 6). Used cut-offs are 0.15 and 0.25;

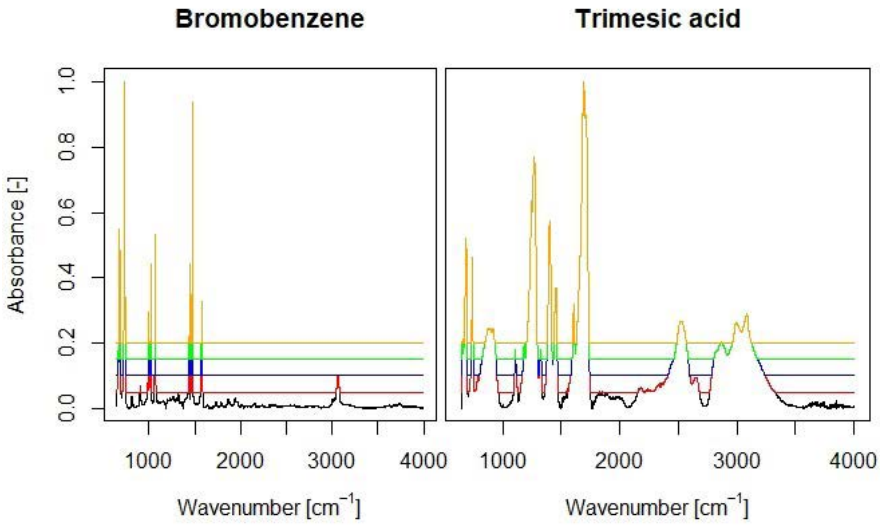


Figure 4. Comparison of cut-offs in the baseline (BS) method (colors of line marks the cut-off applied: black – original spectrum; red – 0.05; blue – 0.10; green – 0.15; yellow – 0.20).

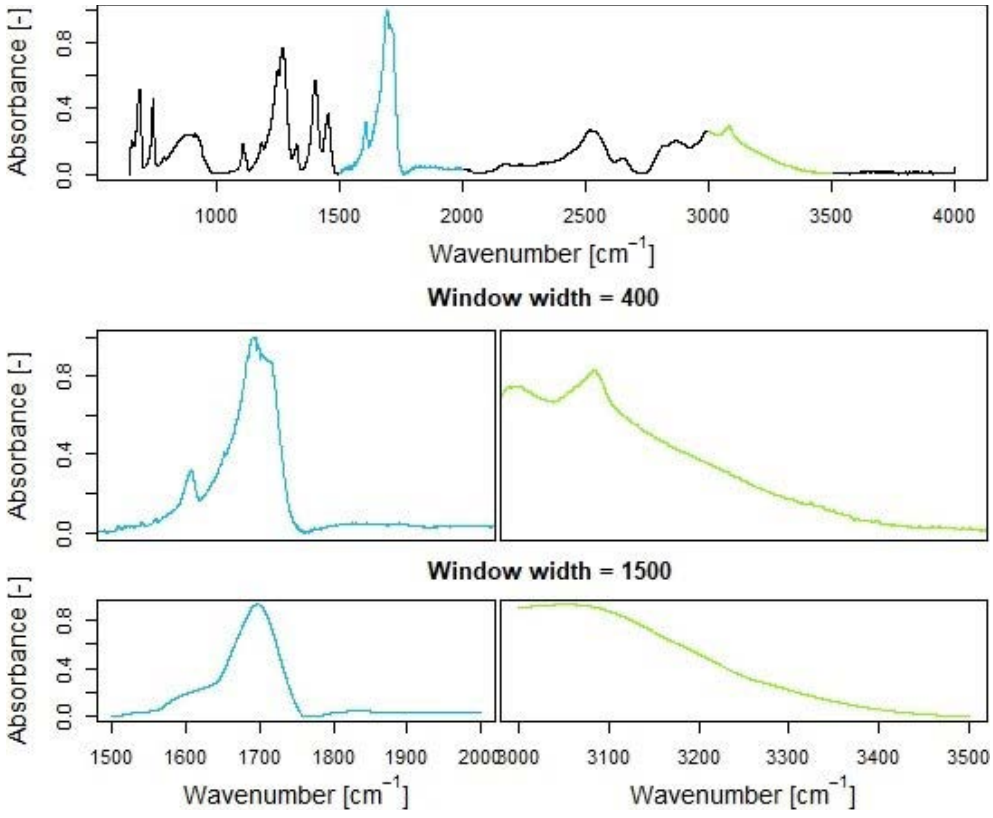


Figure 5. The effect of window widths of 500 and 1500 on the smoothing power of selected parts of the trimesic acid spectrum

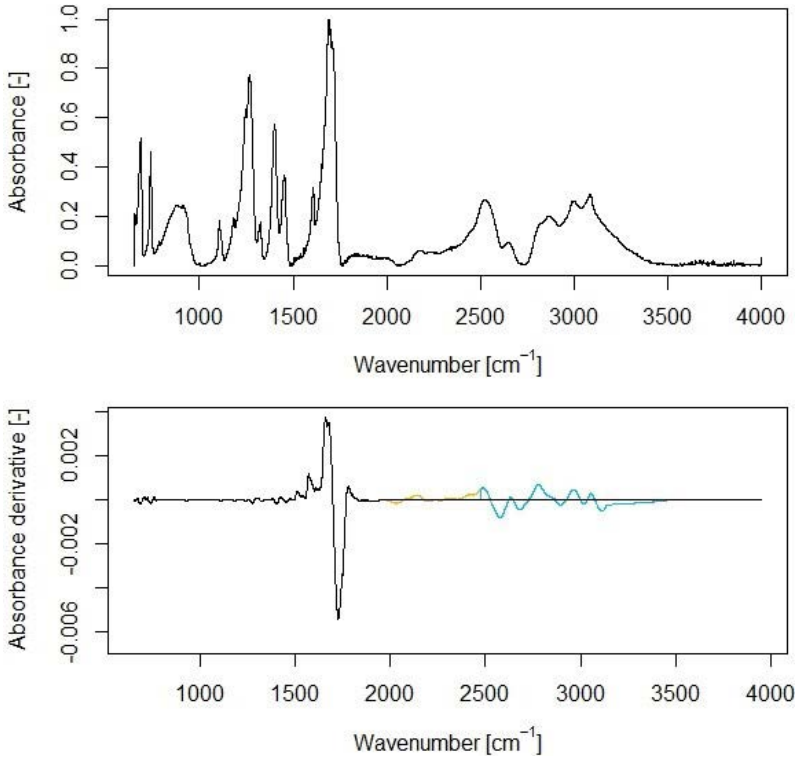


Figure 6. Effect of cut-offs on the spectrum of trimesic acid derivatives in the PoSF method with plots of the original spectra as reference. Cut-offs: 0.15 – yellow; 0.25 – blue; 0.35 – black

d) Derivative Fragmentation (DF) – the derived spectrum is fragmented and the fragment is either left unchanged or all absorbance values in the fragment are set to 0, if the amplitude of the values in a given fragment is less than the cut-off (Figure 7). Used cut-offs are 0.0050 and 0.0005.

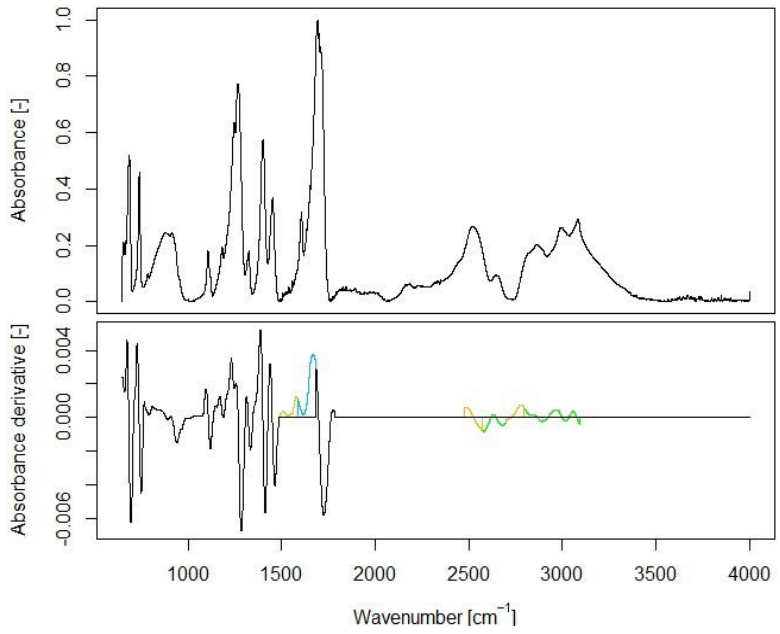


Figure 7. Effect of cut-offs on the spectrum of trimesic acid derivatives in the DF method with graphs of the original spectra as reference. Cut-offs: 0.005 – black; 0.002 – blue; 0.001 – yellow; 0.0005 – green

Prediction models evaluation

Based on the above-described methods, final fingerprints were generated according to the algorithm presented in the previous section and then used to create models to predict physicochemical properties. Predictive models were based on linear regression and a regression tree algorithm.

During the study, it was observed that in almost all cases the models based on the regression tree were more accurate in predicting physicochemical properties. As can be seen in Figure 8, the prediction models for logP and logS obtained by the regression tree method show better quality than those obtained by linear regression (except for PoSF with a cut-off 0.15, which showed a 2.45% worse result). The prediction models for logD showed a significantly higher degree of similarity between the results of the two methods, but the regression tree still proved to be the better method (by 6.11% on average). Therefore, the results obtained using the regression tree method will be used to compare the methods themselves.

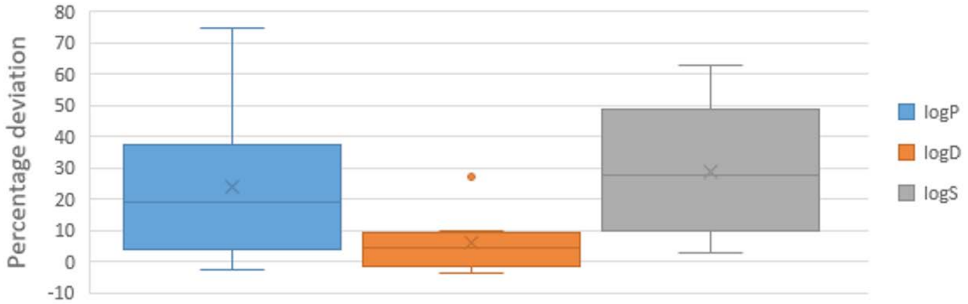


Figure 8. The percentage deviation between the RMSE values for the models obtained using a regression tree and linear regression

As can be seen in Table 1, it is not possible to determine the best method for processing spectroscopic spectra to generate fingerprints in the context of creating predictive models. To compare the methods, the deviation from the lowest error in the group was determined for each RMSE value obtained. The mean of the results for each method and the cut-off was then calculated. It can be seen that fingerprints obtained by the baseline (BS) and derivative fragmentation (DF) methods showed relatively the highest performance. In the case of fingerprints obtained using the baseline method (BS), the RMSE values showed an average deviation from the best result of 23.53% (with a deviation for a cut-off of 0.15 of only 14.41%). For the derived fragmentation method, the same parameter was 29.69%. In contrast, the PrSF method consistently showed some of the worst results among all the physicochemical properties tested, with an average deviation of 54.90%. The results for the PoSF method proved highly variable.

Table 1. Summary of minimum RMSE error values for regression tree models predicting logP, logD and logS values. The lowest RMSE error values for the prediction models based on the FEDS method and the original spectra for each physicochemical property are shown for a reference [17,18]

logD			logP			logS		
Method	**Cut-off**	**RMSE**	**Method**	**Cut-off**	**RMSE**	**Method**	**Cut-off**	**RMSE**
PoSF	0.15	1.1980	BS	0.15	0.6215	BS	0.10	0.6023
DF	0.0050	1.2610	DF	0.0050	0.8419	PoSF	0.15	0.6864
BS	0.15	1.4212	PoSF	0.25	0.9978	DF	0.0005	0.6998
DF	0.0005	1.4652	BS	0.10	1.0394	PrSF	0.25	0.7390

logD			logP			logS		
Method	**Cut-off**	**RMSE**	**Method**	**Cut-off**	**RMSE**	**Method**	**Cut-off**	**RMSE**
PrSF	0.15	1.4682	DF	0.0005	1.0691	BS	0.15	0.7506
PrSF	0.25	1.5645	PoSF	0.15	1.1124	DF	0.0050	0.7645
BS	0.10	1.5658	PrSF	0.25	1.2185	PrSF	0.15	0.8716
PoSF	0.25	1.6310	PrSF	0.15	1.3222	PoSF	0.25	0.9127
FEDS	0.05	1.2200	FEDS	0.05	0.9260	FEDS	0.15	0.8370
Original spectra	0.20	1.1020	Original spectra	0.15	0.6480	Original spectra	0.35	0.8330

Compared to the methods developed in our previous paper, the methods discussed here showed lower RMSE error values for the prediction models of logP and logS. This was particularly the case for logS, where all methods and cut-offs tested, except the PRSF method, showed lower RMSE error values. For logD, the methods discussed in this paper showed slightly higher RMSE error values.

Conclusions

This article discusses the use of the Savitzky-Golay filter and derivatives in the creation of a new type of molecular representation based on molecular structure and FTIR spectra. This is a particularly important topic due to the growing interest of *in silico* methods in the pharmaceutical, cosmetic, and agrochemical industries, among others. The use of *in silico* methods can significantly reduce the time and cost of developing new compounds with desired bioactive properties.

In summary, the use of the Savitzky-Golay filter and derivatives had a positive effect on the quality of the resulting prediction models compared to the FEDS transformation for all of the physicochemical parameters compared. Compared to models based on original spectra, the positive effect of the SG filter and derivatives was evident only for predictive models for logS and to a lesser extent for logP. All these observations show that the application of the SG filter and derivatives for new fingerprint representation has great potential and is a good direction for the development of this particular *in silico* method.

Acknowledgments

This study was supported by the University of Applied Sciences in Tarnów, Poland. I would also like to thank Dominika Golonka for providing her master's thesis along with all data and developed R and KNIME scripts.

References

[1] Terstappen GC, Reggiani A. In silico research in drug discovery. Trends in Pharmacological Sciences. 2001;22(1):23–26. HTTPS://DOI.ORG/10.1016/S0165-6147(00)01584-4.

[2] Willett P, Barnard JM, Downs GM. Chemical similarity searching. Journal of Chemical Information and Computer Sciences. 1998;38(6):983–996, 1998, https://doi.org/10.1021/ci9800211.

[3] Bajorath J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. Journal of Chemical Information and Computer Sciences. 2001;41(2):233–245. https://doi.org/10.1021/ci0001482.

[4] Wigh DS, Goodman JM, Lapkin AA. A review of molecular representation in the age of machine learning. WIREs: Computional Molecular Science. 2022;12(5):1–19. https://doi.org/10.1002/wcms.1603.

[5] Zagidullin B, Wang Z, Guan Y, Pitkänen E, Tang J. Comparative analysis of molecular fingerprints in prediction of drug combination effects. Briefings in Bioinformatics. 2021;22(6):1–15. https://doi.org/10.1093/bib/bbab291.

[6] Ball DW. Field Guide to Spectroscopy [Internet]. Bellingham: SPIE Press; 2006. [cited 2022 June 21]. Available form: https://spie.org/Publications/Book/682726.

[7] Luo J, Ying K, Bai J. Savitzky-Golay smoothing and differentiation filter for even number data. Signal Processing. 2005;85(7):1429–1434. https://doi.org/10.1016/j.sigpro.2005.02.002.

[8] Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry. 1964;36(8):1639–1643. https://doi.org/10.1021/ac60214a048.

[9] Riordon ANJ, Zubritsky E. Top 10 articles. Analytical Chemistry. [Internet] 2000 [cited 2022 June 21]. Available form: http://pubs.acs.org/ac.

[10] Schafer RW. What is a Savitzky-Golay filter? [Lecture notes]. IEEE Signal Processing Magazine. 2011;28(4):111–117. https://doi.org/10.1109/MSP.2011.941097.

[11] Cygański A. Metody spektroskopowe w chemii analitycznej. Warszawa: Wydawnictwo WNT; 2017.

[12] de Aragão BJG, Messaddeq Y. Peak separation by derivative spectroscopy applied to FTIR analysis of hydrolized silica. Journal of the Brazilian Chemical Society. 2008;19(8):1582–1594. https://doi.org/10.1590/S0103-50532008000800019.

[13] Palencia M. Functional transformation of Fourier-transform mid-infrared spectrum for improving spectral specificity by simple algorithm based on wavelet-like functions. Journal of Advanced Research. 2018;14:53–62. Htttps://doi.org/10.1016/J.JARE.2018.05.009.

[14] Rieppo L, Saarakkala S, Närhi T, Helminen HJ, Jurvelin JS, Rieppo J. Application of second derivative spectroscopy for increasing molecular specificity of fourier transform infrared spectroscopic imaging of articular cartilage. Osteoarthritis and Cartilage. 2012;20(5):451–459. https://doi.org/10.1016/J.JOCA.2012.01.010.

[15] Yukihiro O, Slobodan Š, Jiang JH. How can we unravel complicated near infrared spectra? Recent progress in spectral analysis methods for resolution enhancement and band assignments in the near infrared region. Journal of Near Infrared Spectroscopy. 2001;9(2). https://doi.org/10.1255/jnirs.2.

[16] Otálora A, Palencia M. Application of functionally-enhanced derivative spectroscopy (FEDS) to the problem of the overlap of spectral signals in binary mixtures: Triethylamine-acetone. Journal of Science with Technological Applications. 2019;6:96–107. https://doi.org/10.34294/J.JSTA.19.6.44.

[17] Golonka D. Development of a new chemical compound representation based on FTIR spectrum for prediction of physicochemical properties of potential therapeutic substances [master thesis]. Kraków: Jagiellonian University; 2021.

[18] Kurczab R, Golonka D. A new approach to encoding the chemical structure based on the FTIR spectra of compound. In: Xth Conversatory on Medicinal Chemistry in Lublin; 2021. https://doi.org/10.13140/RG.2.2.18264.01284.

[19] Gorzynski Smith J. Chapter 13: Mass spectrometry and infrared spectroscopy. In: Organic Chemistry. 3rd ed. New York: McGraw-Hill; 2011. p. 463–488.

[20] Kennepohl D, Farmer S, Reusch W. 11.5: Infrared Spectra of Some Common Functional Groups. In: LibreTexts: Chemistry [Internet]. [cited 2023, March 11]. Available from: https://chem.libretexts.org/Bookshelves/Organic_Chemistry/Map%3A_Organic_Chemistry_(Wade)_Complete_and_Semesters_I_and_II/Map%3A_Organic_Chemistry_(Wade)/11%3A_Infrared_Spectroscopy_and_Mass_Spectrometry/11.05%3A_Infrared_Spectra_of_Some_Common_Functional_Groups.

Appendix 1

Table A1. 103 compounds used in this paper as a dataset

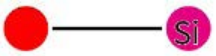
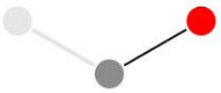
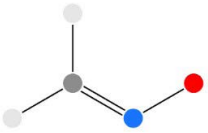
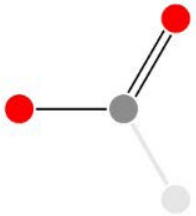
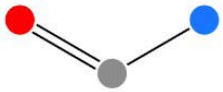
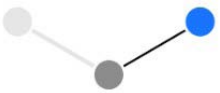
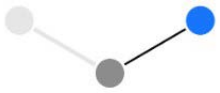
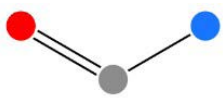
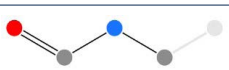
No.	SMILES code	logP	logD	logS
1	CC(C)(\N=N\C(C)(C)C#N)C#N	1.27	1.27	−1.42
2	NC1=C(O)C=C(C2=C1C=CC=C2)S(O)(=O)=O	0.79	−1.37	0.01
3	NC1=CC(O)=C(C=C1)C(O)=O	0.83	−2.23	1.08
4	CCOC(=O)C(NC(C)=O)C(=O)OCC	−0.33	−0.33	−1.29
5	NC1=CC2=C(C=CC=C2)C(=C1)S(O)(=O)=O	1.09	−1.06	0.01
6	COC1=C(N)C=CC(=C1)N(=O)=O	0.93	0.93	−1.68
7	CCOC(CCCN)OCC	0.67	−1.93	2.05
8	OC1(OC2C(=O)NC(=O)NC2=O)C(=O)NC(=O)NC1=O	−2.85	−5.35	1.16
9	CC(=O)NC1=CC=C(N)C=C1	0.38	0.38	−1.36
10	[H]C(=O)C1=CC=C(NC(C)=O)C=C1	0.92	0.92	−1.49
11	CC1=C(N)C=CC(=C1)N(=O)=O	1.6	1.6	−1.93
12	CC(=O)NC1=CC=CC=C1	1.21	1.21	−1.52
13	CC(=O)OC1=CNC2=CC=C(Br)C=C12	2.45	2.45	−3.69
14	[H]C(=O)C1=CC=C(Br)C=C1	2.45	2.45	−2.47
15	CC1=CC(Br)=CC=C1	3.26	3.26	−3.04
16	[O−]C1=CNC2=CC=C(Br)C(Cl)=C12	3.14	3.12	−3.7
17	CCOC(=O)CCC(=O)OCC	0.61	0.61	−0.38
18	[Na+],[O−]S(=O)(=O)C1=CC=CC=C1	1.15	−1.22	−1.8
19	BrC1=CC2=C(NC=C2)C=C1	2.84	2.84	−3.45
20	O=C(OOC(=O)C1=CC=CC=C1)C1=CC=CC=C1	3.95	3.95	−4.29
21	O=C(C1=CC=CC=C1)C1=CC=CC=C1	3.43	3.43	−3.73
22	ClS(=O)(=O)C1=CC=CC=C1	1.92	1.92	−2.17
23	BrC1=CC=CC=C1	2.74	2.74	−2.51
24	C(NC1=CC=CC=C1)C1=CC=CC=C1	3.17	3.17	−3.04
25	C(N1CCNCC1)C1=CC=CC=C1	1.38	−0.47	1.08
26	CCOC(=O)C1=CC=CC=C1	2.33	2.33	−2.03
27	[H]C(=O)C1=NC2=C(C=CC=C2)C=C1	2.51	2.52	−2.16
28	CC1=C(C=CC=C1Cl)N(=O)=O	3.03	3.03	−2.88
29	OC(=O)C1=CC2=C(C=CC=C2)N=C1C(O)=O	0.37	−4.01	0.01
30	BrC1CCCCC1	2.82	2.82	−2.38
31	[H]C(=O)C1=CC=C(Cl)C=C1	2.29	2.29	−2.07
32	ClC(=O)C1=CC=C(Cl)C=C1	2.77	2.77	−3.2
33	OC(=O)C1=CC=CN=C1Cl	1.24	−2.1	0.21
34	[H]C(=O)C1=CC=NC2=C1C=CC=C2	1.84	1.84	−1.93
35	ClC1=C(Cl)C(=O)C(Cl)=C(Cl)C1=O	2.58	2.58	−3.94
36	ClC1=NC=CC=N1	0.96	0.96	−1.65
37	OC(=O)C1=C(Cl)C=CC=C1	2.23	−1.23	0.01

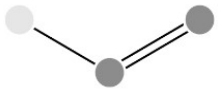
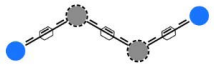
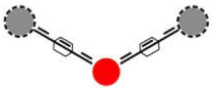
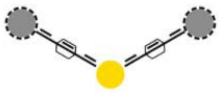
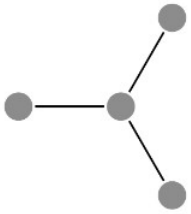
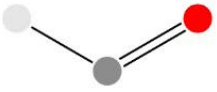
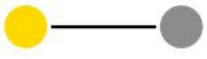
No.	SMILES code	logP	logD	logS
38	O=C(N1C=CN=C1)N1C=CN=C1	−0.95	−0.95	−0.59
39	CN1C(=O)CC(=O)N(C)C1=O	−0.82	−3.14	1.38
40	C(N1CCNCC1)C1=CC=CC=C1	1.41	1.2	−0.64
41	C(C1=CC=CC=C1)C1=CC=CC=C1	4.07	4.07	−3.35
42	NC1=CC2=C(C=CC=C2)C=C1N	1.3	1.3	−2.93
43	ClC1=CC=CC(=C1Cl)N(=O)=O	3.12	3.12	−3.36
44	CC1=CC(=C(O)C(=C1)C(C)(C)C)C(C)(C)C	5.27	5.27	−4.38
45	NC1=C(Cl)C=CC=C1Cl	2.35	2.35	−2.75
46	[H]N(C1CCCCC1)C1CCCCC1	3.41	0.29	0.01
47	ClC1=C(Cl)C(=O)C(C#N)=C(C#N)C1=O	1.43	1.43	−3.03
48	CCCCCCCCCCCCCCCCCCCC(O)=O	8.03	5.62	−6.22
49	COC1=C(O)C=CC(CC=C)=C1	2.61	2.61	−2.26
50	NC(CC1=CC=CC=C1)C(O)=O	−1.18	−1.19	−1.37
51	OC(=O)C1=C(C=CC=C1)C(O)=O	1.29	−4.1	0.39
52	NC(CCC(=O)NC(CS)C(=O)NCC(O)=O)C(O)=O	−4.88	−8.08	0.95
53	COC1=CC=CC(C=O)=C1O	1.87	1.86	−0.94
54	NC(CC1=CNC2=CC=C(O)C=C12)C(O)=O	−1.39	−1.4	−1.81
55	OCCN1CCOCC1	−0.72	−0.73	1.05
56	OCCN(CCO)CC(O)=O	−4.44	−4.45	0.98
57	OC(=O)CC1=CNC2=C1C=CC=C2	1.71	−0.96	0.1
58	OC(=O)CCCC1=CNC2=C1C=CC=C2	2.6	0.1	0.01
59	N1C=CN=C1	−0.15	−0.24	0.04
60	CC(C)C1=NC=CN1	1.22	1.01	−0.42
61	NC1=CC=C(I)C=C1	2.07	2.07	−2.32
62	O,[H][C@@]12CC[C@](CS(O)(=O)=O)(C(=O)C1)C2(C)C	0.98	−1.39	0.82
63	[H][C@]12CC[C@](C)(C(=O)C1)C2(C)C	2.55	2.55	−1.95
64	CN1C=NC2=C1C(=O)N(C)C(=O)N2C	−0.55	−0.55	−0.44
65	OC1C2=CC=CC=C2OC2=C1C=CC=C2	2.52	2.52	−3.24
66	CNCC(O)C(O)C(O)C(O)CO	−3.4	−5.11	2.94
67	COC1=C(N)C=CC(=C1)N(=O)=O	0.93	0.93	−1.68
68	CCCCCCCCCCCCCCCCCCCC(=O)OCC(O)CO	6.86	6.86	−8.51
69	COC1=CC=C(CO)C=C1	1.05	1.05	−0.93
70	OCC1=CC=C(C=C1)N(=O)=O	1.15	1.15	−1.34
71	ClC(=O)C1=CC=C(C=C1)N(=O)=O	2.1	2.1	−2.87
72	OC(=O)C1=C(C=CC=C1)N(=O)=O	1.57	−1.96	0.07
73	ClCC1=C(C=CC=C1)N(=O)=O	2.5	2.5	−2.63
74	O=C1OC(=O)C2=C1C=CC=C2N(=O)=O	1.36	1.36	−2.85
75	OC(=O)C1=CC=CN=C1	−0.17	−2.88	1.72
76	OC1(O)C(=O)C2=CC=CC=C2C1=O	0.45	0.4	−2.18
77	CCC1=C(C=CC=C1)N(=O)=O	2.87	2.87	−2.61

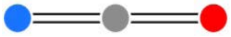
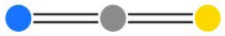
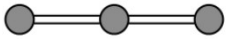
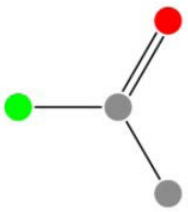
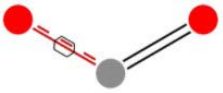
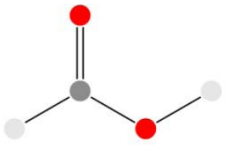
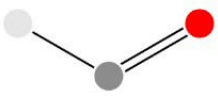
No.	SMILES code	logP	logD	logS
78	OCC1CCCCN1	0.03	−2.35	2.25
79	OC1=C(C=C(C=C1N(=O)=O)N(=O)=O)N(=O)=O	1.49	−0.29	0.01
80	O=CC1=CC=C2OCOC2=C1	1.31	1.31	−1.8
81	O1C(=CN=C1C1=CC=C(C=C1)C1=NC=C(O1)C1=CC=C-C=C1)C1=CC=CC=C1	5.04	5.04	−7.24
82	[Na]OS(=O)(=O)CCN1CCN(CCS(=O)(=O)O[Na])CC1	−0.55	−0.59	0.29
83	CC(C)(C)C(O)=O	1.58	−0.86	1.5
84	COC(=O)C1=CC=CC=C1O	2.32	2.32	−1.29
85	NC(=O)C1=C(O)C=CC=C1	1.17	1.11	−1.1
86	NC1=CC=C(C=C1)S(N)(=O)=O	−0.25	−0.25	−1.17
87	OC(=O)C1=CC(=CC=C1O)S(O)(=O)=O	1.16	−4.72	1.17
88	OC(=O)C(Cl)(Cl)Cl	1.53	−2	0.07
89	N[C@@H](CC1=CNC2=C1C=CC=C2)C(O)=O	−1.09	−1.09	0.6
90	O=C1CCC2=CC=CC=C2C1	2.25	2.25	−1.49
91	CC1=C(C=CC=C1)S(O)(=O)=O	1.67	−0.71	0.6
92	CC1=CC=C(C=C1)S(Cl)(=O)=O	2.43	2.43	−2.7
93	OC(=O)C1=CC(=CC(=C1)C(O)=O)C(O)=O	0.95	0.03	0.26
94	OC(=O)C1=CC=C(C=C1)C(O)=O	1.29	−4.91	0.39
95	CN1C2=C(NC=N2)C(=O)N(C)C1=O	−0.77	−0.89	−0.82
96	OC(C1=CC=CC=C1)(C1=CC=CC=C1)C1=CC=CC=C1	4.64	4.64	−4.46
97	CCCCCCCCN(CCCCCCCC)CCCCCCCC	9.5	6.34	−5.7
98	[H]C(=O)C1=CC(OC)=C(OC)C=C1	1.37	1.37	−1.39
99	[H]C(=O)C1=CC(OC)=C(O)C=C1	1.22	1.08	−0.81
100	BrCCN1C(=O)C2=CC=CC=C2C1=O	1.77	1.77	−3.46
101	COC1=CC=C(CCN)C=C1OC	1.07	−1.24	0.86
102	CN1N(C(=O)C=C1C)C1=CC=CC=C1	1.22	1.22	−1.58
103	C1CN2CCN1CC2	−0.13	−2.47	2.82

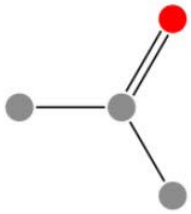
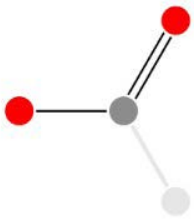
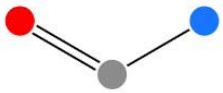
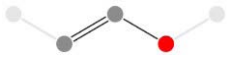
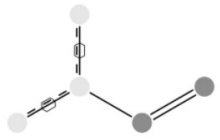
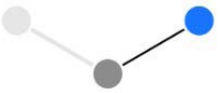
Appendix 2

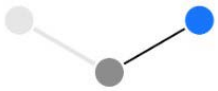
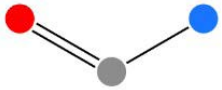
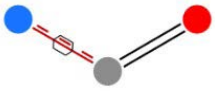
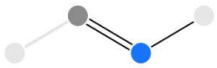
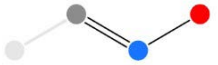
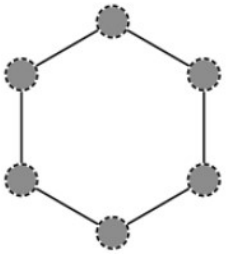
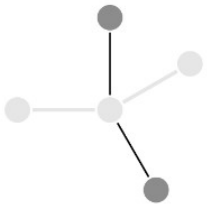
Table A2. The dictionary with the definition of each bit position in the proposed FTIR fingerprint and the corresponding vibration ranges of the functional groups (ν – stretching vibration; δ – deformation vibration)

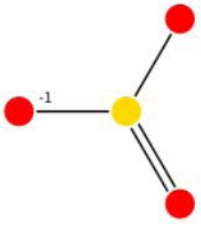
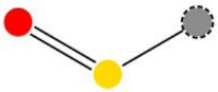
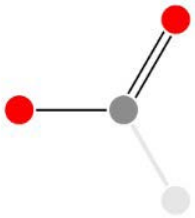
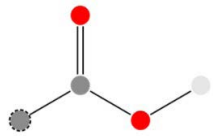
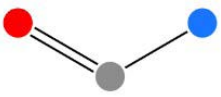
Position number	Chemical bond	Absorption band interval [cm^{-1}]	SMARTS code	SMARTS code visualization
1	Any band	4000–3700		
2	**νO–H**	3700–2500	*~[#8]	
3	νO–H in the SiOH group	3700–3200	[#14]–[#8]	
4	νO–H diluted alcohols and phenols (without HB)	3700–3580	*~[#6]–[#8]	
5	νO–H diluted alcohols and phenols (HB)	3550–3200	*~[#6]–[#8]	
6	νO–H diluted oximes	3650–3590	*–[#6](–[*])=[#7][#8]	
7	νO–H diluted carboxylic acid dimers	3300–2500	*~[#6](=[#8])[#8]	
8	**ν N–H**	3500–2800	*~[#7]	
9	νN–H primary amides	3450–3200	[#8]=[#6]–[#7]	
10	νN–H secondary amides	3500–3400	[#8]=[#6]–[#7][#6]~*	
11	νN–H primary amines	3550–3300	*~[#6]–[#7]	
12	νN–H secondary amines	3350–3300	*~[#6]–[#7][#6]~*	
13	νN–H solid primary amines	3400–3100	[#8]=[#6]–[#7]	
14	νN–H solid secondary amines	3330–3060	[#8]=[#6]–[#7][#6]~*	

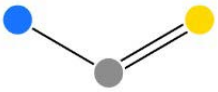
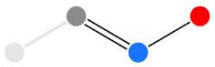
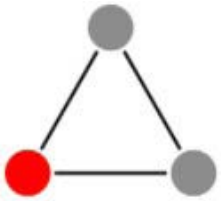
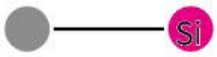
Position number	Chemical bond		Absorption band interval [cm^{-1}]	SMARTS code	SMARTS code visualization
15	**νC–H**		3100–2700	*~[#6]	
16	νC–H alkynes		3300–3250	*–[#6]#[#6]	
17	νC–H alkenes		3100–2950	*–[#6]=[#6]	
18	νC–H	cycloalkane	3100–2990	[#6;R]	
19		romatic com-pound		[c;R]	
20	νC–H heteroaromatic compounds	pyridines	3080–3000	[c]@;:[#7]	
21		pyrazine		[#7]@;:[c]@;:[c] @;:[#7]	
22		pyrroles		[c]@;:[#7]@;:[c]	
23		furans		[c]@;:[#8]@;:[c]	
24		thiophene		[c]@;:[#16]@;:[c]	
25	νC–H tertiary groups		2990–2880	[#6](–[#6])(–[#6])(–[#6])	
26	νC–H alkanes		2970–2840	*~[#6]	
27	νC–H in aldehyde groups		2830–2695	*–[#6]=[#8]	
28	νS–H thiols		2830–2700	[#6]–[#16]	
29	νC≡N nitriles		2300–2200	[#6]#[#7]	

Position number	Chemical bond	Absorption band interval [cm^{-1}]	SMARTS code	SMARTS code visualization
30	νX≡Y X=Y=Z (X,Y,Z=N,C,O,S)	2280–2000	[#6]#[#7]	
31			[#7]#[#16]	
32			[#6]#[#16]	
33			[#6]=[#16]=[#8]	
34			[#7]=[#6]=[#8]	
35			[#6]=[#16]=[#7]	
36			[#7]=[#6]=[#16]	
37			[#6]=[#7]=[#8]	
38			[#7]=[#16]=[#8]	
39	νC≡C alkynes	2270–2100	[C]#[C]	
40	νC=C=C cumulative alkenes	2000–1900	[C]=[C]=[C]	
41	νC=O	1870–1540	[#6]=[#8]	
42	νC=O acyl chlorides	1815–1750	[#6]–[#6](=[#8])[#17]	
43	νC=O lactones	1800–1730	[#8]@;!:[#6]=[#8]	
44	νC=O acid esters	1740–1720	[#6](=[#8])–[#8]–*	
45	νC=O aldehydes	1740–1720	*–[#6]=[#8]	

Position number	Chemical bond		Absorption band interval [cm^{-1}]	SMARTS code	SMARTS code visualization
46	νC=O ketones		1730–1700	[#6]-[#6](=[#8])[#6]	
47	νC=O carbonyl acids (dimers)		1720–1700	*~[#6](=[#8])[#8]	
48	νC=O secondary amides		1710–1660	[#8]=[#6]-[#7][#6]~*	
49	νC=O primary amides		1690–1620	[#8]=[#6]-[#7]	
50	νCH–CF_2 and CF=CF_2		1790–1750	[#6]~[#6]~[#9]	
51	νC=C		1680–1550	[#6]=[#6]	
52	νC=C alkenes		1680–1630	[C]=[C]	
53	νC=C	cycloalkenes	1660–1550	[#6]=;@;!:[#6]	
54		vinyl ethers		*~[#6]=[#6][#8]~*	
55	νAr–C=C		1650–1600	* ~@:~*(@:*)[#6]=[#6]	
56	**νNO2**		1661–1260	[#8]~[#7]~[#8]	
57	νNO2 asymmetric		1661–1500	[#8]~[#7]~[#8]	
58	νNO2 symmetric		1390–1260	[#8]~[#7]~[#8]	
59	**δN–H**		1650–1500	*~[#7]	
60	δN–H primary amines		1650–1580	*~[#6]-[#7]	

Position number	Chemical bond		Absorption band interval [cm^{-1}]	SMARTS code	SMARTS code visualization
61	δN–H secondary amines		1650–1550	*~[#6]–[#7][#6]~*	
62	δN–H primary amides		1620–1590	[#8]=[#6]–[#7]	
63	δN–H secondary amides		1570–1510	[#8]=[#6]–[#7][#6]~*	
64	δN–H lactams		1600–1500	[#7]@;!:[#6]=[#8]	
65	νC=N	imines	1650–1500	*~[#6]=[#7]-*	
66		oximes		*~[#6]=[#7]-[#8]	
67	ν Ar		1620–1560	c1ccccc1	
68	**δC–H**		1470–1350	*~[#6]	
69	δC–H cycloalkanes		1470–1440	[#6;R]	
70	δC–H geminal dimethyl groups		1400–1350	[#6]–*(–[#6])(~*)(~*)	
71	δC–H	alkanes	1470–1350	*~[#6]	
72		alkynes		*-[#6]#[#6]	
73	δO–H		1430–1330	*~[#8]	
74	**νS=O**		1350–1030	[#16]=[#8]	
75	νS=O asymmetric sulfones		1350–1300	[#8]=[#16]=[#8]	

Position number	Chemical bond		Absorption band interval [cm^{-1}]	SMARTS code	SMARTS code visualization
76	νS=O symmetric sulfones		1160–1120	[#8]=[#16]=[#8]	
77	νS=O sulfonates		1195–1150	[#8]=[#16](-[#8])[#8-]	
78	νS=O stretching	sulfoxides alkyl	1070–1030	[#8]=[#16]-[C]	
79		aryl		[#8]=[#16]-[c]	
80	νC–O		1320–1000	[#6]–[#8]	
81	νC–O stretching	carbonyl acids	1320–1210	*~[#6](=[#8])[#8]	
82		aromatic acid esters		[c]-[#6](=[#8])[#8]-*	
83	νC–O alcohols		1260–1000	[#6]–[#8]	
84	νC–O aliphatic esters		1150–1000	[#6]–[#8]–[#6]	
85	**νC–N stretching**		1420–1020	[#6]–[#7]	
86	νC–N primary amides		1420–1400	[#8]=[#6]–[#7]	
87	νC–N secondary amides		1300–1200	[#8]=[#6]–[#7][#6]~*	
88	νC–N diluted aromatic amines		1340–1260	[c]–[#7]	

Position number	Chemical bond	Absorption band interval [cm^{-1}]	SMARTS code	SMARTS code visualization
89	νC–N diluted aliphatic amines	1250–1020	[C]–[#7]	
90	νC=S thioamides	1100–1050	[#7]–[#6]=[#16]	
91	νC–F monofluorinated compounds	1100–1000	[#6]–[#9]	
92	νSi–O–Si siloxanes	1100–1000	[#14]–[#8]–[#14]	
93	νP–O–C aliphatic phosphates	1090–1000	[#15]–[#8]–[C]	
94	νN–O oximes	960–930	*~[#6]=[#7]–[#8]	
95	νC–O–C epoxides	950–810	[#6]1–[#8]–[#6]1	
96	νSi–CH_3	860–750	[#14]–[#6]	
97	νC–X (X=Cl, I, S, Br)	850–460	[#6]–[#17,#53,#16,#35]	
98	νC–Cl	850–550	[#6]–[#17]	
99	νC–S	700–570	[#6]–[#16]	
100	νC–Br	700–500	[#6]–[#35]	
101	νC–I	600–460	[#6]–[#53]	

Appendix 3

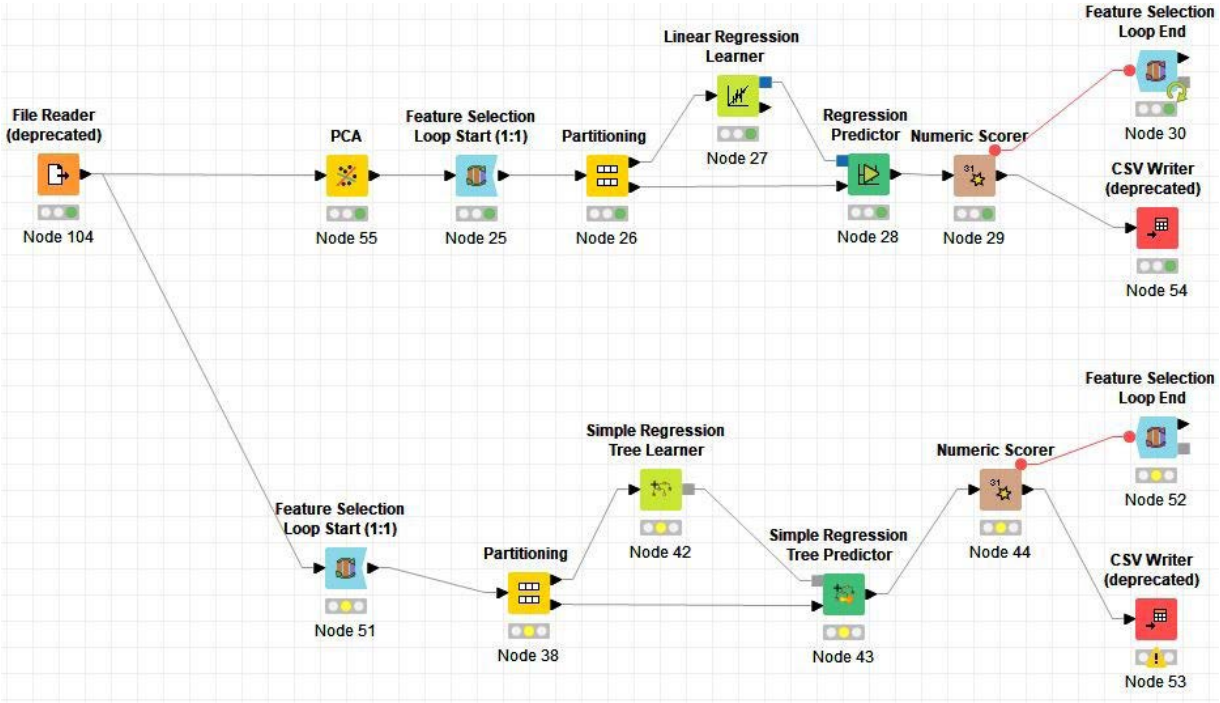


Figure A3. The KNIME workflow used in the study