tidyspec: a gentle framework to handle with spectroscopy data using tidy data philosophy

Marcel Rodrigues Ferreira[1,2] Julia Ferreia Moraes[1]

Luisa Sutter[1] Willian Fernando Zambuzzi[3,*]

[1] Departament of Chemistry and Biochemistry, São Paulo State University (UNESP), Institute of Biosciences, Campus Botucatu
[2] Molecular Genetics and Bioinformatics Laboratory, Experimental Research Unit (Unipex), School of Medicine, São Paulo State University (UNESP)
[3] Departament of Chemistry and Biochemistry, São Paulo State University (UNESP), Institute of Biosciences, Campus Botucatu

[*] Correspondence: Willian Fernando Zambuzzi <w.zambuzzi@unesp.br>

```
<!-- convidar o grupo da USP para trabalhar nisso -->
```

Abstract

Introduction

Spectroscopy

Spectroscopy, the study of interactions between light and matter, has been a cornerstone of scientific exploration for centuries, enabling us to peer into the fundamental properties and intricate structures of materials across various domains. From the early days of visible light observations to the cutting-edge techniques of modern spectroscopy, researchers have continually pushed the boundaries of our understanding, unraveling the mysteries of atoms, molecules, and complex systems. This paper presents a comprehensive exploration of recent advancements in spectroscopic analyses, delving into the multifaceted applications, innovative methodologies, and emergent technologies that have revolutionized our ability to probe, characterize, and manipulate matter at its most fundamental levels. By harnessing the power of spectroscopy, scientists are embarking on a journey to unlock hidden dimensions of matter, uncovering new insights that have far-reaching implications in fields ranging from materials science and

chemistry to astronomy and biology. This paper not only reviews key developments in spectroscopic techniques but also highlights their profound impact on our comprehension of the physical world, underlining the ever-expanding role of spectroscopy as an indispensable tool for scientific discovery and technological innovation.

$$c = \lambda\nu$$

$$E = h\nu$$

$$E = \frac{hc}{\lambda}$$

$$\overline{\nu} = \frac{1}{\lambda}$$

R language

The R programming language has emerged as a versatile and powerful tool in the realm of spectroscopy, offering researchers an extensive suite of specialized packages designed to facilitate data analysis, visualization, and interpretation. Leveraging R's rich statistical capabilities and its seamless integration with a diverse range of data formats, spectroscopists can efficiently process and manipulate complex spectral datasets. Notably, packages such as `{hyperSpec}` [1](Beleites and Sergo) and `{ChemoSpec}` (Hanson 2023) provide dedicated functions for handling hyperspectral data, allowing for preprocessing, spectral alignment, and exploratory data analysis. 'SpecMine' and 'SpecHelpers' further extend R's utility by enabling advanced spectral processing, peak detection, and quantification tasks. Visualization of spectroscopic data is greatly enhanced by packages like `{ggplot2}` (Wickham 2016) and `{plotly}` (Sievert 2020), enabling researchers to create insightful and interactive graphical representations. The extensibility of R also encourages the development of customized algorithms and methods tailored to specific spectroscopic challenges. As the field of spectroscopy continues to expand its horizons, R remains an invaluable asset, empowering scientists to unravel the intricacies of spectral information with precision and depth.

[1]In this paper, the R packages names will be formated as {package}, a format widely used by the R community.

Tidydata and tidyverse

Despite the good functioning of the base version, the creation of `tidyverse` in 2016 revolutionized the R language by offering a new programming ecosystem focused on the entire data analysis cycle (Wickham et al. 2019). Currently, `tidyverse` consists of a core set of 9 packages, `ggplot2` (Wickham 2016), `readr` (Wickham, Hester, and Bryan 2023), `tibble` (Müller and Wickham 2022), `tidyr` (Wickham, Vaughan, and Girlich 2023), `purrr` (Wickham and Henry 2023), `dplyr` (Wickham et al. 2023), `stringr` (Wickham 2022), `forcats` (Wickham 2023) and `lubridate` (Grolemund and Wickham 2011), covering data *import*, *transformation*, *visualization*, *modeling* and *communication*. Tidyverse packages was built under a same philosophy, named tidy data [REF]. A data set must follow three rules to be considered a tidy data: **(1)** Each *column* is a distinguish variable, **(2)** each *row* is a different observation, and **(3)** each *cell* is a single value [REF].

With the growing expansion of the `tidyverse` and the interest in machine learning tools, in 2019 a set of packages inspired by the tidy data philosophy called `tidymodels` was released (Giorgi, Ceraolo, and Mercatelli 2022)[REF].

Data sets

UV:

FTIR:

Raman: `{hyperSpec}` package `chodro` data is a special S4 R object named ***`hyperSpec`*** class which contains 850 Raman spectra with x (`chondro$x`) and y (`chondro$y`) spatial coordinates (from a grid conformation of 25 x 35), in addition the `chondor$spc` has 300 data points measured from 600 cm^{-1} to 1800 cm^{-1} for each spectra.

Overview of `tidyspec` package

The `tidyspec` package was design to enable the data analysis of spectroscopy data (as IR, Raman, NMR) with the tidydata format. Similar to packages like `{stringr}` (Wickham 2022) and `{forcats}` (Wickham 2023), the tidyspec package contains a naming pattern at the beginning of its function name, in this case `spec_`. Since spectra are values in function of *wavelength*, *wavenumber*, *frequency* or *energy* values, a `wn_col` argument has been defined in order to map these values to the input data. ESCREVER SOBRE OS PACOTES QUE COMPOE AS FUNÇÕES

The functions in the `tidyspec` package were created by wrapping functions of `tidyverse` packages in order to simplify the code syntax. Package performance is guaranteed by using the `recipes` package (Kuhn and Wickham 2022), which provides structure for data processing.

The functions were designed to solve 6 different problems in spectroscopy:

- **Transformation**: Convert data from absorbance to transmittance (`spec_abs2trans`) & from transmittance to absorbance (`spec_trans2abs`). The conversion is based on the following equation, where A means absorbance and T (values from 0 to 1) or T(%) (values from 0 to 100) for transmittance.

$$A = 2 - \log_{10} T(\%)$$

or

$$A = -log_{10} T$$

- **Normalization**: Normalize the data to range 0-1 (`spec_norm_01`), normalize between a custom range (`spec_norm_minmax`), or normalize to have a standard deviation of one (`spec_norm_var`).
- **Baseline correction**: Correct the baseline using the the *rolling ball* (Kneen and Annegarn 1996) algorithm (`spec_blc_rollingBall`). The function `spec_bl` return the baseline vectors (`spec_bl_rollingBall`).
- **Smooth correction**: Smooth the data using the average window (`spec_smooth_avg`) or using the *Savitzky-Golay* algorithm (Savitzky and Golay 1964) (`spec_smooth_sga`).
- **Derivative**: Create differential data from the spectra (`spec_diff`).
- **Preview**: Preview your data while applying changes statically (`spec_smartplot`) or interactively (`spec_smartplotly`).

Case studies

UV-VIS data

```
library(tidyspec)
library(tidyverse)
library(modeldatatoo)
library(viridis)
library(baseline)

#UV <- data_chimiometrie_2019()
```

FTIR

Raman

For the purpose of this work the `chondor$spc` data was converted into a tidy data format with the first column the wavenumbers and the following columns each spectra. This data was called `Raman_tidy`.

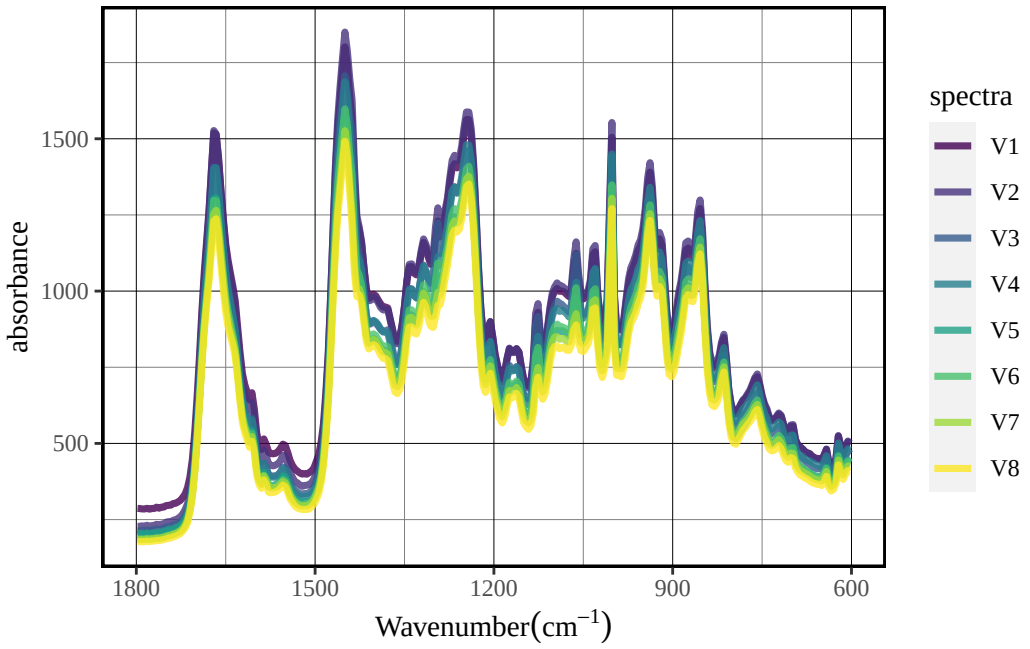
```
library(hyperSpec)
Raman <- chondro

dim(Raman$spc)
```

```
[1] 875 300
```

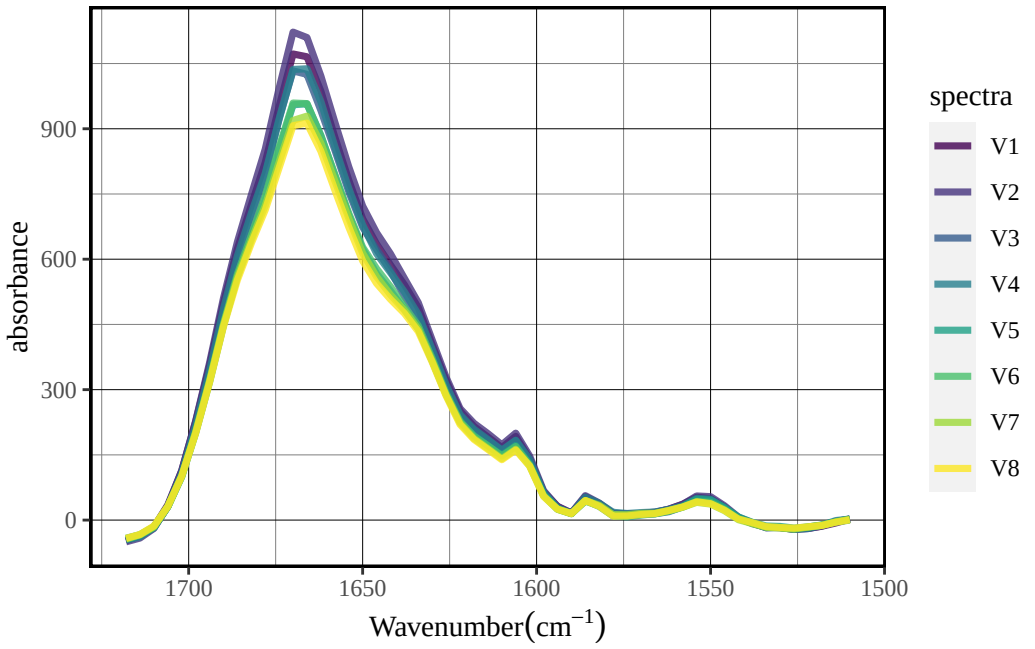
```
Raman_tidy <- tibble(Wn = as.numeric(colnames(Raman$spc))) %>%
  bind_cols(as_tibble(t(Raman$spc),.names_repair = "minimal"))

print(Raman_tidy[1:10,1:8])
```

```
# A tibble: 10 x 8
      Wn    V1    V2    V3    V4    V5    V6    V7
   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
 1   602  502.  500.  466.  477.  439.  436.  425.
 2   606  505.  508.  475.  482.  445.  442.  426.
 3   610  489.  489.  456.  465.  429.  426.  412.
 4   614  466.  465.  437.  445.  411.  408.  397.
 5   618  492.  491.  458.  470.  433.  431.  421.
 6   622  524.  526.  490.  501.  461.  458.  446.
 7   626  452.  451.  423.  431.  397.  394.  383.
 8   630  428.  424.  395.  405.  373.  371.  363.
 9   634  425.  419.  391.  401.  368.  366.  357.
10   638  438.  435.  406.  416.  382.  378.  369.
```

```
Raman_tidy %>%
  select(1:9) %>%
  spec_smartplot()
```




```
Raman_tidy %>%
  select(1:9) %>%
  spec_blc_rollingBall(Wn_min = 1508,
                       Wn_max = 1718,
                       wm = 15,
                       ws = 15) %>%
  spec_smartplot(geom = "point")
```



Conclusion and perspectives

Acknowledgments

The authors thanks FAPESP (n 2018/05731-7, 2021/14271-2) for the funding and to MSc. Viviane Oliveira for the valuable suggestions.

References

Beleites, Claudia, and Valter Sergo. "hyperSpec: A Package to Handle Hyperspectral Data Sets in r." https://github.com/r-hyperspec/hyperSpec.

Giorgi, Federico M., Carmine Ceraolo, and Daniele Mercatelli. 2022. "The R Language: An Engine for Bioinformatics and Data Science." *Life* 12 (5): 648. https://doi.org/10.3390/life12050648.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with {Lubridate}" 40. https://www.jstatsoft.org/v40/i03/.

Hanson, Bryan A. 2023. "ChemoSpec: Exploratory Chemometrics for Spectroscopy." https://CRAN.R-project.org/package=ChemoSpec.

Kneen, M. A., and H. J. Annegarn. 1996. "Algorithm for Fitting XRF, SEM and PIXE X-Ray Spectra Backgrounds." *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* 109-110 (April): 209–13. https://doi.org/10.1016/0168-583x(95)00908-6.

Kuhn, Max, and Hadley Wickham. 2022. "Recipes: Preprocessing and Feature Engineering Steps for Modeling." https://CRAN.R-project.org/package=recipes.

Müller, Kirill, and Hadley Wickham. 2022. "Tibble: Simple Data Frames." https://CRAN.R-project.org/package=tibble.

Savitzky, Abraham., and M. J. E. Golay. 1964. "Smoothing and Differentiation of Data by Simplified Least Squares Procedures." *Analytical Chemistry* 36 (8): 1627–39. https://doi.org/10.1021/ac60214a047.

Sievert, Carson. 2020. "Interactive Web-Based Data Visualization with r, Plotly, and Shiny." https://plotly-r.com.

Wickham, Hadley. 2016. "Ggplot2: Elegant Graphics for Data Analysis." https://ggplot2.tidyverse.org.

———. 2022. "Stringr: Simple, Consistent Wrappers for Common String Operations." https://CRAN.R-project.org/package=stringr.

———. 2023. "Forcats: Tools for Working with Categorical Variables (Factors)." https://CRAN.R-project.org/package=forcats.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. "Dplyr: A Grammar of Data Manipulation." https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Lionel Henry. 2023. "Purrr: Functional Programming Tools." https://CRAN.R-project.org/package=purrr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. "Readr: Read Rectangular Text Data." https://CRAN.R-project.org/package=readr.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2023. "Tidyr: Tidy Messy Data." https://CRAN.R-project.org/package=tidyr.