



Inteligência artificial aplicada ao pré-tratamento e à análise quimiométrica de espectros infravermelho de diferentes famílias de briófitas

Raísa A. L. B. da Silveira¹; Marcelo G. de Barros²; Geovana R. de Moraes¹; Paulo E. A. S. Câmara³; Pedro A. W. R. Brito⁴, Davi A. C. Ferreira¹.

¹*Instituto de Química, Universidade de Brasília*

²*Instituto Nacional de Criminalística, Polícia Federal*

³*Instituto de Biologia, Universidade de Brasília*

⁴*Corpo de Bombeiro Militar do Distrito Federal*

raisalacerda.unb@gmail.com

Palavras-Chave: Assinatura espectral, estatística multivariada, modelagem computacional.

Introdução

Diante da grande diversidade vegetal dos biomas brasileiros e da escassez de bancos de dados espectroscópicos consolidados para identificação e classificação botânica, o desenvolvimento de métodos analíticos mais rápidos e não destrutivos, capazes de subsidiar a análise de acervos já depositados em herbários, possui caráter estratégico. Nesse contexto, a construção de uma biblioteca de caracterização vibracional para famílias de briófitas assume particular importância para a botânica, por representar alternativa não invasiva e complementar às técnicas moleculares, como a análise do Ácido Desoxirribonucleico (ADN) por Reação em Cadeira da Polimerase (PCR) que, embora ofereça maior poder discriminativo em nível varietal, depende de extração e amplificação de material genético (Ferreira, 1998), procedimento nem sempre viável em amostras de herbário antigas.

Briófitas são plantas terrestres avasculares, ou seja, sem tecidos condutores de seiva. Possuem parede celular composta por celulose e não possuem raiz, caule ou folha, apresentando estruturas análogas chamadas rizoide, caulóide e filoide (Groffinet, 2009). São encontradas principalmente em ambientes úmidos e sombreados, onde formam a base de diversos ecossistemas vegetais, sendo encontradas em diferentes tipos de biomas (Câmara, 2005). Dentre os métodos de análise química já consolidados, técnicas de extração por solventes, cromatografia (HPLC, UPLC, GC-MS) e espectroscopia (RMN, FTIR, UV-Vis) já são empregadas para caracterizar metabólitos secundários em plantas de forma geral, incluindo briófitas (Goffinet, 2005, 2009).

A espectroscopia vibracional, especialmente na região do infravermelho (FTIR-IV), consolidou-se como ferramenta poderosa para a caracterização de materiais biológicos, por ser rápida, reprodutível, não destrutiva e capaz de fornecer informações sobre a composição química de tecidos vegetais, como a presença de celulose, lignina, proteínas, lipídios, metabólitos secundários e outros compostos (Türker, 2017). Apesar de suas vantagens, a técnica



não é de detecção molecular única e, portanto, exige padronização, coleta rigorosa de dados e rigor na análise quimiométrica de espectros de IV.

A análise espectral enfrenta desafios ligados a limitações intrínsecas do sinal, como baixo rendimento fotônico, e a perturbações extrínsecas, como oscilações ambientais, que provocam desvios na linha de base, comprometendo a quantificação espectral (Valente *et al.*, 2025). Soma-se a isso a variabilidade entre instrumentos, que amplifica essas distorções e prejudica tanto análises tradicionais quanto modelos estatísticos multivariados usados no pré-processamento dos dados (Workman Jr., 2018)

Dentre os principais desafios para o tratamento e análise de espectros vibracionais, destacam-se os relacionados à flutuações de temperatura e de lâmpada, que provocam desvios no eixo de absorbância ou transmitância (Rinnan, 2009); as diferenças físicas granulométricas das amostras que levam ao espalhamento da luz incidente (Grisanti, 2018); interferência do detector e ruído instrumental, que provocam ruído de fundo (Gautam, 2015); e por fim, a própria dimensionalidade dos dados, ou seja, milhares de variáveis correlacionadas em poucas amostras, desafio este contornado com o uso de ferramentas estatísticas como Análise de Componentes Principais (PCA) (Wold, 1987).

Nesse sentido, métodos de remoção de ruído e correção da linha de base, ambos pelo filtro de Savitzky-Golay, além da correção de espalhamento, obtida pelo método SNV (*Standard Normal Variate*) desempenham papel fundamental para diminuir interferências introduzidas na aquisição de sinais, imagens e demais dados brutos (Matera; Polenta; Napoli, 2024).

Diante do exposto, o presente estudo se insere em um cenário mais amplo da espectroscopia vibracional: o uso da inteligência artificial (IA) para o pré-tratamento e análise estatística dos espectros. Até recentemente, o tratamento e interpretação de espectros exigia conhecimento especializado e pressupostos sobre a estrutura da amostra; hoje, modelos de IA já processam diretamente dados hiperespectrais brutos e retornam resultados calibrados em milissegundos (Workman, 2025). Segundo o autor, a IA vem se consolidando como fator determinante para a área, representando ponto de inflexão na análise espectroscópica, na calibração de modelos e na interpretação de dados obtidos por diferentes plataformas, ajudando a superar gargalos tradicionais no processamento de dados e viabiliza a análise de espectros complexos.

Portanto, este trabalho tem como objetivo central investigar o potencial de desenvolvimento de um banco de dados de assinaturas espectrais vibracionais de famílias de briófitas, utilizando FTIR-IV associada a criação de uma ferramenta de IA para pré-tratamento e análise estatística multivariada (PCA e HCA).



Material e Métodos

Amostras das famílias de briófitas secas foram cedidas pelo herbário do Departamento de Botânica da Universidade de Brasília. Os espectros de infravermelho médio (4000 a 400 cm^{-1}) foram obtidos no espectrômetro Bruker Alpha II, equipado com interferômetro RockSolid™ e detector DTGS. Os espectros IV obtidos foram pré-processados e analisados estatisticamente utilizando ferramenta de IA desenvolvida pelos autores no Gemini Canva, intitulada “*Análise Espectral Avançada 1.20 - Raman/NIRS*”. A aplicação opera de forma autônoma no lado do cliente (client-side), com motores de álgebra linear e modelos estatísticos implementados nativamente em JavaScript.

A ferramenta desenvolvida de é estruturada em três módulos funcionais (separadores), os quais:

(i) Pré-processamento e visualização: Destina-se ao tratamento inicial dos espectros brutos via (a) imputação de dados ausentes, por meio de interpolação dos pontos faltantes no conjunto espectral; (b) aplicação de limiar de ruído (noise threshold), para supressão de sinais de baixa intensidade associados a ruído de fundo; (c) suavização e derivação pelo algoritmo de Savitzky-Golay (Savitzky; Golay, 1964), empregada para atenuar desvios e deslocamentos da linha de base; e (d) normalização dos dados por Variação Normal Padrão (Standard Normal Variate, SNV) e por centragem na média, procedimentos que visam assegurar a comparabilidade entre amostras ao minimizar efeitos de espalhamento de luz e variações multiplicativas de intensidade (Yan, 2025). A visualização dos espectros é realizada por meio de gráficos bidimensionais interativos, com possibilidade de exportação das representações gráficas em formato de imagem (PNG).

(ii) Análise Hierárquica de Agrupamentos (HCA): Investiga padrões de similaridade calculando a matriz de correlação de Pearson (Brereton, 2005) entre os espectros. Os resultados fundamentam a construção de um dendrograma e de um mapa de calor (heatmap), cuja intensidade cromática expressa o grau de correlação entre os pares de amostras.

(iii) Análise de Componentes Principais (PCA) e Detecção de Outliers: O terceiro módulo compreende ferramentas estatísticas multivariadas voltadas à redução de dimensionalidade e à identificação de amostras anômalas. A Análise de Componentes Principais é realizada por meio do algoritmo NIPALS (Nonlinear Iterative Partial Least Squares) (Wold, 1987), implementado internamente na própria ferramenta, o qual permite reduzir a elevada dimensionalidade dos dados espectrais às três primeiras componentes principais (PC1, PC2 e PC3), representadas em gráfico tridimensional interativo. A detecção de valores anômalos é efetuada com base no cálculo da estatística T^2 de Hotelling e dos resíduos Q (SPE), conforme metodologia de Jackson e Mudholkar (1979), associados aos respectivos limites de confiança de 95% e 99%, permitindo identificar amostras com comportamento espectral discrepante.

Resultados e Discussão

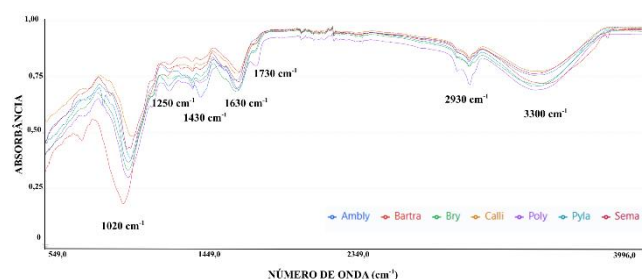
As sete famílias de briófitas e suas respectivas espécies estão apresentadas na Tabela 1. Os espectros no infravermelho médio, apresentados na Figura 1, confirmam que o perfil geral das amostras é carboidrato-dominante, com contribuição de proteína e traço de lipídio, padrão característico de tecido de briófitas.

Tabela 1 - Famílias e espécies de briófitas selecionadas para análise espectroscópica no infravermelho.

Família	Espécie	Local de coleta	Código
<i>Polytrichaceae</i>	<i>Notoligotrichum trichodon</i>	Antártida	Poly
<i>Pylaisiadelphaceae</i>	<i>Microcalpe subsimplex</i>	Brasil	Pyla
<i>Sematophyllaceae</i>	<i>Hydropogon fontinaloides</i>	Brasil	Sema
<i>Bryaceae</i>	<i>Bryum pseudotriquetrum</i>	Antártida	Bry
<i>Calliergonaceae</i>	<i>Sarmentypnum exannulatum</i>	Suécia	Calli
<i>Amblystegiaceae</i>	<i>Hygroamblystegium tenax</i>	Estados Unidos	Ambly
<i>Bartramiaceae</i>	<i>Bartramia patens</i>	Antártida	Bartra

Fonte: os autores.

Figura 1. Espectros de absorção no infravermelho brutos das amostras pertencentes às sete famílias de briófitas analisadas.



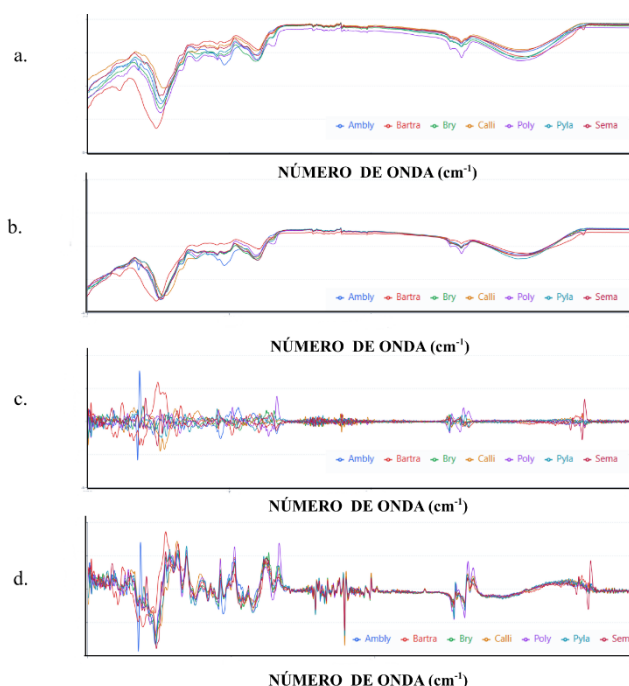
Fonte: os autores.

A banda larga centrada em aproximadamente 3300 cm^{-1} é atribuída ao estiramento sobreposto de grupos O–H e N–H, evidenciando a presença conjunta de carboidratos estruturais e compostos proteicos (Cao, 2023). Em 2930 cm^{-1} , observa-se banda de intensidade média a fraca referente ao estiramento assimétrico de C–H em grupos CH_2/CH_3 , associada à fração lipídica e cerosa da cutícula. Na região de 1630 cm^{-1} , de intensidade média, verifica-se sobreposição entre a banda de Amida I (C=O de proteínas) e vibrações do esqueleto aromático associadas a compostos fenólicos; nessa mesma faixa espectral, é possível que ocorra a banda de Amida II, em torno de 1550 cm^{-1} , referente à deformação N–H (Kļaviņa, 2012). A banda em 1430 cm^{-1} , também de intensidade média, é atribuída à deformação angular de O–H e CH_2 de polissacarídeos da parede celular. Além disso, duas faixas espectrais podem estar presentes nas amostras e servir como potenciais marcadores para diferenciar famílias de briófitas: a região entre 1710 e 1735 cm^{-1} , associada ao estiramento C=O de carbonilas em ésteres saturados (como lipídios de membrana e pectina esterificada), e a região entre 1505 e 1515 cm^{-1} ,

associada a vibrações do esqueleto aromático de compostos fenólicos (Hu, 2011) (Maksimova, 2013). Por fim, o pico mais fino e de forte intensidade em 1020 cm^{-1} é atribuído ao estiramento C–O de polissacarídeos estruturais (celulose, hemicelulose e oligossacarídeos), reforçando o caráter carboidrato-dominante da matriz espectral, com possível contribuição adicional das regiões de $1260\text{--}1270\text{ cm}^{-1}$ e 1160 cm^{-1} , associadas a vibrações C–C, C–O e C–N de polissacarídeos e proteínas (Cao, 2023).

A Figura 2 apresenta, além dos espectros brutos das amostras (Figura 2a), os espectros de infravermelho obtidos após a aplicação sequencial dos pré-tratamentos espectrais empregados: correção de linha de base por Variação Normal Padrão (SNV) (Figura 2b), primeira derivada de Savitzky-Golay (Figura 2c) e, por fim, centralização na média (mean-centering) (Figura 2d).

Figura 2. Espectros de infravermelho das amostras: (a) brutos; (b) após correção de linha de base por SNV; (c) após 1ª derivada de Savitzky-Golay; (d) após centralização na média.



Fonte: os autores.

A ferramenta desenvolvida aplica sequencialmente as etapas de pré-tratamento espectral: inicialmente, o SNV (Figura 2b) normaliza cada espectro individualmente, removendo efeitos multiplicativos de espalhamento de luz decorrentes de diferenças físicas entre amostras (tamanho de partícula, textura) e corrigindo deslocamentos de linha de base, resultando em espectros com média zero e variância unitária, comparáveis entre si. Em seguida, a ferramenta aplica a primeira derivada de Savitzky-Golay (1964) (Figura 2c), que suprime deslocamentos residuais de linha de base e realça bandas sobrepostas, aumentando a resolução



aparente; essa etapa é executada após o SNV para operar sobre um sinal já livre de efeitos multiplicativos grosseiros. Por fim, o centrar na média (Figura 2d) é aplicado imediatamente antes da PCA e da HCA, centralizando cada variável (número de onda) em torno de sua média no conjunto amostral — operação complementar ao SNV, pois este atua ao longo do eixo espectral de cada amostra, enquanto a centragem atua ao longo do eixo das amostras para cada variável. Assim, a ferramenta assegura que a análise quimiométrica capture a variância relevante entre amostras, e não a magnitude absoluta do sinal.

A matriz de correlação de Pearson, Figura 3, revelou alta similaridade (coloração verde) entre as famílias *Calliergonaceae* e *Pylaisiadelphaceae*, bem como entre *Sematophyllaceae* e *Pylaisiadelphaceae*. As famílias *Calliergonaceae* e *Bartramiaceae*, assim como *Bartramiaceae* e *Pylaisiadelphaceae*, apresentaram similaridade inversa (coloração vermelha).

Figura 3. Matriz de correlação de Pearson representada por mapa de calor (*heatmap*)



Fonte: os autores.

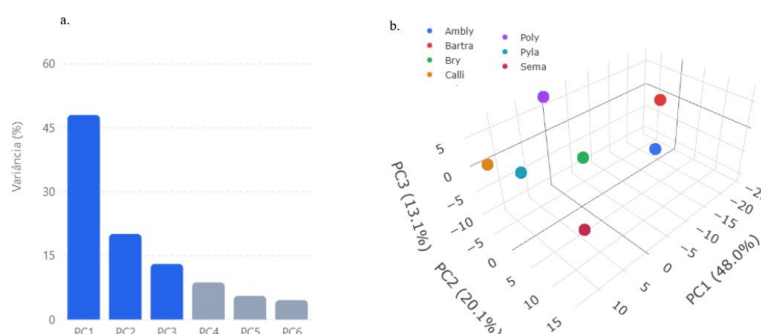
Esse padrão pode ser interpretado devido ao perfil ecológico tradicionalmente associado a cada família: *Sematophyllaceae* e *Pylaisiadelphaceae* são famílias pleurocárpicas de distribuição predominantemente tropical com estreita parentela filogenética frequentemente associadas a crescimento sobre troncos, cascas de árvores e madeira em decomposição em florestas úmidas, o que pode justificar a maior similaridade espectral entre ambas (Carvalho-Silva, 2017). Por outro lado, *Calliergonaceae* reúne espécies tipicamente higrófilas, associadas a ambientes brejosos, turfeiras e corpos d'água (frequentemente submersas ou flutuantes), enquanto *Bartramiaceae* agrupa espécies predominantemente rupícolas e terrícolas, associadas a afloramentos rochosos, fendas de rocha e solos, em condições de umidade mais variável e irregular (Oliván, 2005) (Castro, 2002). Essa divergência de hábito ecológico poderia explicar a baixa similaridade espectral observada entre essas duas famílias, uma vez que diferentes regimes hídricos e substratos tendem a favorecer composições químicas distintas na parede celular e em compostos de proteção.

Por fim, a análise por PCA permitiu reduzir a dimensionalidade dos dados espectrais a três componentes principais, que juntas explicaram aproximadamente 78% da variância total do conjunto de dados (Figura 4a), indicando que a maior parte da variabilidade espectral entre



as amostras pôde ser capturada em um espaço reduzido de apenas três dimensões. No gráfico de scores, amostras posicionadas mais próximas entre si indicam maior semelhança em relação ao perfil espectral no infravermelho, enquanto amostras mais distantes entre si refletem maior divergência.

Figura 4. Análise PCA dos perfis espectrais. (a) Percentual de variância explicada pelas componentes principais (PCs); (b) *Scoreplot* 3D evidenciando a separação espacial das amostras no sistema de eixos PC1 × PC2 × PC3.



Fonte: os autores.

Embora a ferramenta utilizada não disponibilize gráficos de *loadings*, que impossibilita a identificação direta e quantitativa das variáveis que mais contribuem para cada componente, a análise das regiões onde os sinais oscilam acima e abaixo do ponto zero do eixo espectral (Figura 2c) permite inferir, de forma qualitativa, que a região compreendida entre 1000 e 1700 cm^{-1} é a principal responsável pela variância observada entre as amostras. Essa constatação é coerente com o perfil químico discutido anteriormente, uma vez que essa faixa espectral concentra justamente as bandas mais informativas quanto à composição da parede celular e de metabólitos associados: as vibrações de deformação e estiramento de polissacarídeos, as bandas de carbonila de ésteres e carboxilatos, as bandas de Amida I e II de proteínas e a região de vibração do esqueleto aromático de compostos fenólicos. Ou seja, por reunir simultaneamente marcadores de carboidratos estruturais, proteínas, lipídios esterificados e compostos fenólicos, essa faixa espectral corresponde à chamada região de impressão digital (*fingerprint region*).

Conclusões

A ferramenta de IA utilizada permitiu uma análise consideravelmente mais rápida dos dados espectrais em comparação com o fluxo de trabalho tradicional em *toolboxes* quimiométricas consolidadas. Essa agilidade decorreu principalmente da possibilidade de integrar, em um único ambiente, as etapas de pré-tratamento espectral, análise multivariada e interpretação dos resultados, reduzindo a necessidade de intervenção manual e de transição entre diferentes softwares. Dessa forma, ferramentas como a utilizada neste estudo podem representar uma alternativa promissora, sobretudo



para triagens exploratórias rápidas, ainda que não substituam necessariamente a robustez e a validação consolidadas de *toolboxes* tradicionais.

Este trabalho teve como objetivo investigar essa ferramenta perante a possibilidade de análise estatística de amostras espectrais, e, apesar de ter sido possível identificar regiões espectrais características de cada família e observar, por meio da análise por PCA, uma separação consistente entre os grupos em determinadas faixas do espectro IV, faz-se necessário análises contendo um número maior de espectros por amostra. Ainda, para que os resultados alcancem maior robustez estatística, é necessário realizar testes cegos que permitam avaliar a capacidade preditiva e a confiabilidade da ferramenta de forma mais rigorosa e isenta de vies.

Adicionalmente, identifica-se como limitação a ausência, na versão atual da ferramenta, do módulo de *loadings* da PCA, o que restringiu a interpretação mais aprofundada da origem da separação espectral observada entre as famílias de briófitas. Assim, ficam como oportunidades para trabalhos futuros a ampliação do banco de espectros por amostra e por família, a condução de testes cegos para validação da ferramenta, e a implementação do cálculo e da visualização dos *loadings* de PCA, possibilitando associar diretamente as regiões espectrais discriminantes às bandas de absorção.

Agradecimentos

Agradecimento ao Laboratório de Análise Instrumental (LAI) do Instituto de Química da Universidade de Brasília pelo uso dos equipamentos espectroscópicos; à colaboração intelectual com o corpo técnico do Instituto de Criminalística da Polícia Federal; ao acesso ao Herbário do Departamento de Botânica do Instituto de Biologia da Universidade de Brasília.

Referências

- BRERETON, R. G. Chemometrics: data analysis for the laboratory and chemical plant. **Journal of Analytical Chemistry**, v. 60, n. 10, p. 994-996, oct. 2005.
- CÂMARA, P. E. A. S.; OLIVEIRA, J. R. M. P.; SANTIAGO, M. M. M. A checklist of the bryophytes of Distrito Federal (Brasília, Brazil). **Tropical Bryology**, v. 26, p. 133-140, 2005.
- CARVALHO-SILVA, M.; STECH, M.; SOARES-SILVA, L. H.; BUCK, W. R.; WICKETT, N. J.; LIU, Y.; CÂMARA, P. E. A. S. A molecular phylogeny of the Sematophyllaceae s.l. (Hypnales) based on plastid, mitochondrial and nuclear markers, and its taxonomic implications. **Taxon**, v. 66, n. 4, p. 811-831, 2017.
- CASTRO, N. M. C. F.; PÔRTO, K. C.; YANO, O.; CASTRO, A. A. J. F. Levantamento florístico de Bryopsida de cerrado e mata ripícola do Parque Nacional de Sete Cidades, Piauí, Brasil. **Acta Botanica Brasilica**, v. 16, n. 1, p. 61-76, 2002.
- CAO, Zhen; WANG, Zhenjie; GAO, De. Efficient discrimination of some moss species by Fourier transform infrared spectroscopy and chemometrics. **Spectroscopy**, v. 38, n. 3, p. 30-38, mar. 2023.
- GAUTAM, R.; VANGA, S.; ARIESE, F.; UMAPATHY, S. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. **EPJ Techniques and Instrumentation**, v. 2, n. 1, p. 8, 2015.
- GRISANTI, E. et al. Dynamic Localized SNV, Peak SNV, and Partial Peak SNV: Novel Standardization Methods for Preprocessing of Spectroscopic Data. **Journal of Spectroscopy**, v. 2018, p. 5037572, 2018.
- GOFFINET, B.; BUCK, W. R.; SHAW, A. J. Morphology, anatomy, and classification of the Bryophyta. **Bryophyte Biology**. 2. ed. Cambridge: Cambridge University Press, 2009. p. 55-138.



HU, Tao; JIN, Wen-Ying; CHENG, Cun-Gui. Classification of five kinds of moss plants with the use of Fourier transform infrared spectroscopy and chemometrics. **Spectroscopy**, v. 25, p. 271-285, 2011.

OLIVÁN, G.; FUERTES, E.; ACÓN, M. Straminergon and Calliergon (Calliergonaceae, Bryopsida) in the Iberian Peninsula. **Lazaroa**, v. 26, p. 17-33, 2005.

JACKSON, J. E.; MUDHOLKAR, G. S. Control procedures for residuals associated with principal component analysis. **Technometrics**, v. 21, n. 3, p. 341-349, 1979.

KĻAVIŅA, Laura; BIKOVENS, Oskars; ŠTEINBERGA, Iveta; MAKSIMOVA, Viktorija; EGLĪTE, Linda. Characterization of chemical composition of some bryophytes common in Latvia. **Environmental and Experimental Biology**, v. 10, p. 27-34, 2012.

MAKSIMOVA, Viktorija; KĻAVIŅA, Laura; BIKOVENS, Oskars; ZICMANIS, Andris; PURMALIS, Oskars. Structural characterization and chemical classification of some bryophytes found in Latvia. **Chemistry & Biodiversity**, v. 10, 2013.

RINNAN, Å.; VAN DEN BERG, F.; ENGELSEN, S. B. Review of the most common pre-processing techniques for near-infrared spectra. **TrAC Trends in Analytical Chemistry**, v. 28, n. 10, p. 1201-1222, 2009.

SAVITZKY, A.; GOLAY, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. **Analytical Chemistry**, v. 36, n. 8, p. 1627-1639, 1964.

TÜRKER-KAYA, S.; HUCK, C. W. A review of mid-infrared and near-infrared imaging: Principles, concepts and applications in plant tissue analysis. **Molecules**, v. 22, n. 1, p. 168, 2017. Disponível em:

WOLD, S.; ESBENSEN, K.; GELADI, P. Principal component analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 2, n. 1-3, p. 37-52, 1987.

WORKMAN JR., J. J. A Review of Calibration Transfer Practices and Instrument Differences in Spectroscopy. **Applied Spectroscopy**, v. 72, n. 3, p. 340-365, 2018.

WORKMAN JR., J. AI Developments That Changed Vibrational Spectroscopy in 2025. Spectroscopy Online, 2025. Disponível em: <https://www.spectroscopyonline.com/view/ai-developments-that-changed-vibrational-spectroscopy-in-2025>.

YAN, C. A Review on Spectral Data Preprocessing Techniques for Machine Learning and Quantitative Analysis. **iScience**, v. 28, n. 7, p. 112759, 2025.