

Título em Português: Redes Neurais Artificiais aplicadas ao Estudo de Fármacos por Espectroscopia Raman

Título em Inglês: Artificial Neural Networks applied to the Study of Pharmaceutical Compounds by Raman Spectroscopy

Autor: Tulio Santana Ramos

Instituição: Universidade de São Paulo

Unidade: Instituto de Física de São Carlos

Orientador: Javier Alcides Ellena

Área de Pesquisa /
SubÁrea: Física da Matéria Condensada

Agência Financiadora: CNPq - PIBIC

Redes Neurais Artificiais aplicadas ao Estudo de Fármacos por Espectroscopia Raman

Tulio Santana Ramos

Prof. Dr. Marcelo Barbosa de Andrade, Dr. Alfredo A. A. Exposito de Queiroz

Prof. Dr. Javier Alcides Ellena

Instituto de Física de São Carlos, IFSC/USP

tulip@usp.br

Objetivos

Este projeto tem como objetivo a análise do uso dos espectros Raman de insumos farmacêuticos utilizando redes neurais artificiais (RNAs). Dessa forma, foi avaliado o uso das aplicações de RNAs de classificação na pesquisa farmacêutica, assim como efetuada a construção de um modelo autoral de CNN para aplicar os conceitos abordados neste projeto.

Métodos e Procedimentos

Inicialmente, foram estudados conceitos das principais áreas tratadas neste projeto, a espectroscopia Raman e RNAs, para o melhor entendimento de seus respectivos processos intrínsecos e mais prudente interpretação dos trabalhos encontrados para estudo. A partir disso, foram realizados testes para o tratamento de espectros Raman, obtidos através de projetos em colaboração do Laboratório Multiusuário de Cristalografia Estrutural (LaMuCrEs) [1] e dados espectrais de bases de dados gratuitas, como o projeto RRUFF [2].

Foram comparados métodos de correção de linha de base, como *LU-Whittaker Smoothing* [3], *Rolling Ball* [4] e uma simples correção polinomial [4] para compreender as

características dos espectros, como a importância de picos proeminentes e a remoção de fluorescência. Em seguida, foram selecionados artigos científicos que mesclaram a espectroscopia Raman com RNAs.

Apesar de não estarem diretamente associados ao estudo de compostos farmacêuticos, a análise dos resultados obtidos nos artigos permitiu a compreensão de outros métodos de pré-processamento, arquiteturas de redes neurais, e construção de conjuntos de dados.

Dada a pouca quantidade de espectros Raman disponíveis para serem usados no treinamento de uma RNA, sentiu-se a necessidade de construir uma rede capaz de classificar dois compostos. Após serem exploradas diferentes bases de dados gratuitas, como *Spectral Database for Organic Compounds (SDBS)* [5] e *Wiley Science Solutions SpectraBase (SpectraBase)* [6] foram escolhidas imagens de espectros de Aciclovir e Aspirina, devido a sua maior quantidade de amostras disponíveis.

Entretando, como a quantidade de imagens obtidas era insuficiente, foi adicionado ruído Gaussiano, uma técnica de Aumento de Dados [7], para construir um conjunto de dados.

Posteriormente, iniciou-se a implementação de uma rede convolucional [8], selecionada por influência dos artigos estudados e pela sua capacidade de analisar conjuntos de imagens.

Resultados

Com o estudo dos métodos de pré-processamento, e considerando que as amostras obtidas em bases de dados foram disponibilizadas como imagens, optou-se por expandir a quantidade de espectros disponíveis para confeccionar um total de dez mil imagens para cada composto: Aspirina e Aciclovir.

Com o conjunto de dados consolidado, iniciou-se a construção da rede convolucional utilizando duas camadas convolucionais, uma de *pooling* e uma totalmente conectada. Sendo que a camada conecta é composta por duas camadas ocultas, a primeira com dez neurônios e a outra com dois, referentes às duas classes sendo consideradas: Aspirina e Aciclovir.

Após testes de configuração para o aperfeiçoamento dos parâmetros da rede, sua iteração mais recente obteve uma acurácia média de 81% e em média 0.42 para a função custo.

Considerando estes resultados, a rede confeccionada foi considerada eficaz na classificação dos compostos selecionados, consequentemente também evidenciando que a conexão entre espectroscopia Raman e redes neurais artificiais é possível e que muito ainda pode ser estudado.

Conclusões

A construção de espectros padronizados, através da correção de linha de base, demonstrou ser fundamental para otimizar a análise de espectros Raman, garantindo resultados claros e confiáveis. O uso de técnicas de pré-processamento permitiu a remoção de aspectos indesejados sem comprometer a amostra, permitindo o uso dos espectros em processos de expansão do conjunto de dados disponíveis até então.

Soma-se a isso, a promissora integração de redes neurais artificiais com a espectroscopia Raman, ampliando as possibilidades de uso em diversas áreas do conhecimento. A implementação da rede convolucional para a classificação de imagens alcançou uma acurácia média de 81% na categorização de

compostos farmacêuticos. Este resultado destaca a viabilidade da conexão entre as duas áreas abordadas, sugerindo futuros estudos para maior aprimoramento das técnicas.

Agradecimentos

Sinceros agradecimentos ao CNPQ, aos colaboradores deste projeto, aos familiares, amigos e a todos os membros do LaMuCrEs por seu apoio no decorrer do projeto.

Referências

- [1] Laboratório Multiusuário de Cristalografia Estrutural LaMuCrEs. Disponível em: <https://www.ifsc.usp.br/~emu-xtal/lamucres/>. Acesso em: 20 ago. 2024.
- [2] RRUFF Project – Raman mineral database. Disponível em: <https://rruff.info/>. Acesso em: 20 ago. 2024.
- [3] GALLAGHER, N.B. Whittaker Smoother. Eigenvector Research Incorporated. Disponível em: <https://eigenvector.com/wp-content/uploads/2020/01/WhittakerSmoother.pdf>. Acesso em: 20 ago. 2024
- [4] ERB, D. pybaselines Documentation. Disponível em: <https://pybaselines.readthedocs.io/en/stable/pdf/>. Acesso em: 20 ago. 2024.
- [5] Spectral Database for Organic Compounds, SDBSWeb. Disponível em: <https://sdb.sdb.aist.go.jp>. Acesso em: 20 ago. 2024
- [6] SpectraBase. John Wiley & Sons, Inc. Disponível em: <https://spectrabase.com/>. Acesso em: 20 ago. 2024
- [7] YILMAZ, A.O. Data Augmentation: Methods and Applications of Data Augmentation. 21 jul. 2021. Disponível em: <https://aoyilmaz.medium.com/data-augmentation-methods-and-applications-of-data-augmentation-4d000ce85d56>. Acesso em: 20 ago. 2024
- [8] Data Science Academy. Deep Learning Book Brasil. Disponível em: <https://www.deeplearningbook.com.br/>. Acesso em: 20 ago. 2024

Artificial Neural Networks applied to the Study of Pharmaceutical Compounds by Raman Spectroscopy

Tulio Santana Ramos

Prof. Dr. Marcelo Barbosa de Andrade, Dr. Alfredo A. A. Exposito de Queiroz

Prof. Dr. Javier Alcides Ellena

Instituto de Física de São Carlos, IFSC/USP

tulip@usp.br

Objectives

This project aims to analyze the use of Raman spectra for pharmaceutical ingredients through the application of artificial neural networks (ANNs). The study evaluates the potential of classification-based ANNs in pharmaceutical research. Additionally, a custom CNN model was developed to apply the concepts discussed in this project.

Materials and Methods

Initially, foundational concepts from the key areas of this project—Raman spectroscopy and artificial neural networks (ANNs)—were studied to gain a deeper understanding of their intrinsic processes and to enable a more informed interpretation of the relevant literature. Based on this foundation, tests were conducted for processing Raman spectra, obtained through collaborations with the Multi-User Structural Crystallography Laboratory (LaMuCrEs) [1] and spectral data from free databases, such as the RRUFF project [2].

Various baseline correction methods were compared, including LU-Whittaker Smoothing [3], Rolling Ball [4], and simple polynomial correction [4], to better understand spectral

characteristics like the significance of prominent peaks and fluorescence removal. Following this, scientific articles that combined Raman spectroscopy with artificial neural networks were reviewed.

Although not directly related to the study of pharmaceutical compounds, analyzing the results from these articles provided insights into other preprocessing methods, neural network architectures, and dataset construction.

Due to the limited availability of Raman spectra for training an ANN, a network was designed to classify two specific compounds. After exploring various free databases, including the Spectral Database for Organic Compounds (SDBS) [5] and Wiley Science Solutions SpectraBase [6], spectra images of Acyclovir and Aspirin were selected because of the higher number of available samples.

However, since the number of images was still insufficient, Gaussian noise was added—a Data Augmentation technique [7]—to expand the dataset. Subsequently, a convolutional network [8] was implemented, chosen based on the influence of the studied articles and its capability to analyze image datasets.

Results

After studying preprocessing methods and considering that the samples obtained from

databases were provided as images, it was decided to expand the available spectra, resulting in the creation of ten thousand images for each compound: Aspirin and Acyclovir.

With the dataset established, the construction of a convolutional network began, utilizing two convolutional layers, one pooling layer, and a fully connected layer. The fully connected layer consisted of two hidden layers, the first with ten neurons and the second with two neurons, corresponding to the two classes being considered: Aspirin and Acyclovir.

Following configuration tests to refine the network's parameters, the latest iteration achieved an average accuracy of 81% and an average cost function value of 0.42.

Considering these results, the network was deemed effective in classifying the selected compounds, thereby also demonstrating that the integration of Raman spectroscopy and artificial neural networks is feasible and that there is still much to explore in this area.

Conclusions

The standardization of spectra through baseline correction proved essential for optimizing the analysis of Raman spectra, ensuring clear and reliable results. The use of preprocessing techniques allowed the removal of unwanted features without compromising the sample, enabling the expansion of the available dataset. In addition, the promising integration of artificial neural networks with Raman spectroscopy broadens the possibilities for application across various fields of knowledge. The implementation of the convolutional network for image classification achieved an average accuracy of 81% in categorizing pharmaceutical compounds.

This result highlights the feasibility of connecting these two fields, suggesting future studies for further refinement of the techniques.

Acknowledgments

The sincerest thanks to CNPQ, the collaborators of this project, family, friends, and all the

members of LaMuCrEs for their support throughout the project.

References

- [1] Laboratório Multiusuário de Cristalografia Estrutural LaMuCrEs. Available at: <https://www.ifsc.usp.br/~emu-xtal/lamucres/>. Accessed in 20 aug. 2024.
- [2] RRUFF Project – Raman mineral database. Available at: <https://rruff.info/>. Accessed in 20 aug. 2024.
- [3] GALLAGHER, N.B. Whittaker Smoother. Eigenvector Research Incorporated. Available at: <https://eigenvector.com/wp-content/uploads/2020/01/WhittakerSmoother.pdf>. Accessed in 20 ago. 2024
- [4] ERB, D. pybaselines Documentation. Available at: <https://pybaselines.readthedocs.io/en/stable/pdf/>. Accessed in 20 aug. 2024.
- [5] Spectral Database for Organic Compounds, SDBSWeb. Available at: <https://sdb.sdb.aist.go.jp>. Accessed in 20 aug. 2024
- [6] SpectraBase. John Wiley & Sons, Inc. Available at: <https://spectrabase.com/>. Accessed in 20 aug. 2024
- [7] YILMAZ, A.O. Data Augmentation: Methods and Applications of Data Augmentation. 21 jul. 2021. Available at: <https://aoyilmaz.medium.com/data-augmentation-methods-and-applications-of-data-augmentation-4d000ce85d56>. Accessed in 20 aug. 2024
- [8] Data Science Academy. Deep Learning Book Brasil. Available at: <https://www.deeplearningbook.com.br/>. Accessed in 20 aug. 2024