# Explicabilidade em Redes Neurais de Grafos para a Avaliação de Autismo Usando Análise de fMRI

**Matheo Angelo Pereira Dantas[1], André Carlos Ponce de Leon Ferreira de Carvalho[2]**
[1,2] Instituto de Ciências Matemáticas e de Computação (ICMC) - Universidade de São Paulo, Brasil
matheoangelo@usp.br[1] e andre@icmc.usp.br[2]

## Introdução

A identificação do Transtorno do Espectro Autista (TEA) é importante para assegurar direitos e qualidade de vida, mas é dificultada pela ausência de um marcador biológico conhecido pela ciência. Assim, têm sido estudados algoritmos de Inteligência Artificial para procurar padrões além da compreensão humana atual e tentar diagnosticar o TEA de forma automática a partir de dados biológicos, e dentre esses, um dos métodos mais estudados consiste em constuir grafos a partir de exames de Ressonância Magnética Funcional (fMRI), que filmam a circulação sanguínea no cérebro por meio do sinal BOLD (Blood Oxygenation Level Dependent), e processá-los usando Redes Neurais de Grafos [1]. Entretanto, esses algoritmos necessitam de um tratamento especial na explicabilidade, pois dados neurológicos são altamente complexos e modelos de Redes Neurais funcionam como "caixas-pretas"que fazem cálculos grandes e ininteligíveis.

As abordagens da explicabilidade podem ser divididas em dois grupos: a explicabilidade *post-hoc* e os modelos auto-explicáveis. Na explicabilidade *post-hoc*, são aplicados algoritmos prontos [2] para explicar modelos já treinados. Por outro lado, as GNNs autoexplicáveis [3] têm uma arquitetura projetada para a explicabilidade e explicações podem ser obtidas a partir de mecanismos interpretáveis internos. A última abordagem geralmente é mais desejável, pois gera explicações mais fiéis ao modelo [4].

## Metodologia

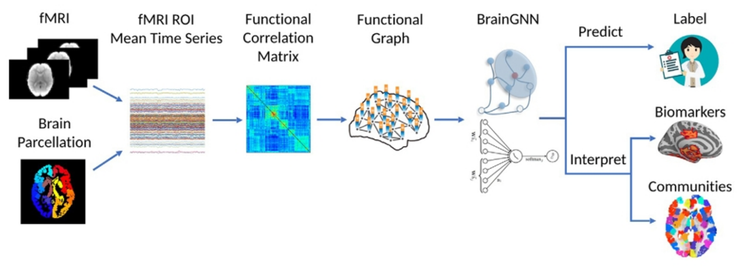


**Figura 1:** Pipeline do BrainGNN [5], modelo utilizado no experimento.

O código do experimento está disponível no GitHub[1].

### Dados e pré-processamento

Foram utilizados os dados de rs-fMRI (fMRI em Estado de Repouso) da Autism Brain Imaging Data Exchange I (ABIDE I), com 539 indivíduos autistas e 573 no grupo de controle. Os dados foram obtidos a partir do Preprocessed Connectomes Project (PCP)[2], que possui os exames da ABIDE pré-processados com diversos *atlas* cerebrais. Foi escolhido o *atlas* Harvard-Oxford, que mapeia 110 ROIs (*Regions Of Interest*) no cérebro, com seus respectivos nomes. Com esses exames pré-processados, foram criados os grafos correspondentes a cada amostra, construindo a matriz de adjacência a partir dos maiores valores da matriz de correlação das séries BOLD de cada ROI.

### Modelo

Utilizamos o modelo BrainGNN [5], um dos mais influentes em aplicações de neurociência. O BrainGNN possui alguns mecanismos de auto-explicabilidade, como o *pooling* de nós importantes e a detecção de comunidades. O código do modelo foi extraído diretamente do repositório do artigo original, e adaptações foram feitas para permitir a aplicação da explicabilidade.

#### Interpretabilidade do modelo

Usamos como explicação do modelo a nível individual uma máscara de nós, o conjunto de nós (correspondentes a ROIs) restante após as duas camadas de *TopK Pooling*. Para sugestões de possíveis marcadores biológicos, observamos as regiões do cérebro mais frequentes nas explicações a nível individual dos exemplos classificados como autistas.

#### Hiperparâmetros testados

Foram testados diferentes combinações de hiperparâmetros para avaliar os efeitos na acurácia e na explicabilidade. Os hiperparâmetros avaliados foram:

- ***Pooling ratio*:** após cada camada de passagem de mensagem, diz qual fração dos nós anteriores será mantida no *TopK Pooling*. Padrão: 0.5.
- ***TopK Pooling Loss*:** É incluída na função *Loss* do treinamento para encorajar o modelo a atribuir pesos ou muito altos ou muito baixos aos nós ao ordená-los para a seleção do *TopK Pooling*. Peso padrão: 0.1.
- ***Group Consistency Loss*:** É incluída na função *Loss* do treinamento para forçar o modelo a dar escores de importância parecidos para os nós, de forma a obter explicações de nível individual consistentes entre si. Peso padrão: 0.1.

### Métricas de explicabilidade

**Esparsidade [2]:** as explicações devem ser boas em resumir o comportamento de modo sucinto, sem usar o grafo de input inteiro.

$$Sparsity = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{|m_i|}{|M_i|}\right)$$

**Fidelidade positiva e negativa [2]:** as explicações devem ser fiéis ao modelo. A fidelidade positiva mede se a informação da explicação é relevante, e a fidelidade negativa mede se a informação fora da explicação é irrelevante.

$$Fidelity_+ = \frac{1}{N}\sum_{i=1}^{N}|f(G_i) - f(G_i^{1-m_i})| \qquad Fidelity_- = \frac{1}{N}\sum_{i=1}^{N}|f(G_i) - f(G_i^{m_i})|$$

**Consistência de marcador biológico:** para além de usar as métricas conhecidas na literatura para avaliar a explicação a nível individual, queremos o quanto as regiões apontadas como marcador biológico são representativas do conjunto de explicações individuais.

$$Consistency = \frac{\sum_{i=1}^{B} n_i}{\sum_{i=1}^{N} n_i}$$

Onde $B$ é o número de ROIs escolhidas para compor a sugestão de marcador biológico (aqui, foi escolhido $B = 10$), e $n_i$ é a quantidade de vezes que a i-ésima ROI mais frequente apareceu nas explicações individuais.

## Resultados

| | Padrão | *Ratio* = 0.3 | *Ratio*=0.3, *TopK*=0.5 | *Group* = 0.5 |
|---|---|---|---|---|
| Acurácia | 0.51 | 0.57 | 0.59 | 0.56 |
| Fidelidade+ | 0.050 | 0.025 | 0.032 | 0.078 |
| Fidelidade- | 0.79 | 0.85 | 0.88 | 0.85 |
| Esparsidade | 0.75 | 0.90 | 0.90 | 0.75 |
| Consistência | 0.23 | 0.34 | 0.26 | 0.19 |
| Biomarcador | • Left Parahippocampal Gyrus; posterior division<br>• Left Heschl's Gyrus (includes H1 and H2)<br>• Left Frontal Orbital Cortex<br>• Right Temporal Fusiform Cortex; anterior division<br>• Right Lingual Gyrus<br>• Right Planum Temporale<br>• Right Supracalcarine Cortex<br>• Left Intracalcarine Cortex<br>• Left Temporal Fusiform Cortex; posterior division<br>• Left Putamen | • Left Supramarginal Gyrus; posterior division<br>• Left Frontal Medial Cortex<br>• Left Superior Parietal Lobule<br>• Right Thalamus<br>• Left Juxtapositional Lobule Cortex (formerly Supplementary Motor Cortex)<br>• Left Angular Gyrus<br>• Right Juxtapositional Lobule Cortex (formerly Supplementary Motor Cortex)<br>• Left Planum Temporale<br>• Left Inferior Frontal Gyrus; pars triangularis<br>• Right Amygdala | • Left Supramarginal Gyrus; posterior division<br>• Right Angular Gyrus<br>• Left Frontal Medial Cortex<br>• Right Subcallosal Cortex<br>• Right Superior Temporal Gyrus; anterior division<br>• Right Amygdala<br>• Right Paracingulate Gyrus<br>• Left Planum Temporale<br>• Left Superior Parietal Lobule<br>• Left Juxtapositional Lobule Cortex (formerly Supplementary Motor Cortex) | • Left Supramarginal Gyrus; posterior division<br>• Right Angular Gyrus<br>• Left Planum Temporale<br>• Left Superior Parietal Lobule<br>• Right Occipital Pole<br>• Right Subcallosal Cortex<br>• Left Frontal Medial Cortex<br>• Right Thalamus<br>• Right Juxtapositional Lobule Cortex (formerly Supplementary Motor Cortex)<br>• Right Amygdala |

**Tabela 1:** Tabela com os resultados dos experimentos no BrainGNN. Cada coluna corresponde a uma combinação de mudanças de hiperparâmetros no treinamento, e cada linha corresponde a uma métrica de desempenho.

## Conclusões

A acurácia dos modelos, apesar de demonstrar alguma capacidade de detectar padrões, foi relativamente baixa, evidenciando a dificuldade do problema. Em particular, foi muito menor do que no experimento original do BrainGNN (79.8%), que usava um conjunto de dados de task-fMRI (o Biopoint), sugerindo que sinais de autismo são mais visíveis quando o cérebro é submetido a estímulos específicos. Esse nível de acurácia se refletiu na baixa fidelidade positiva e na alta fidelidade negativa. Além disso, o *Ratio* pareceu ter uma influência considerável sobre a consistência de biomarcador.

## Agradecimentos

Este trabalho foi apoiado pela FAPESP, no processo 24/09181-2.

## Referências

[1] S. Zhang, J. Yang, Y. Zhang, J. Zhong, W. Hu, C. Li, and J. Jiang, "The combination of a graph neural network technique and brain imaging to diagnose neurological disorders: A review and outlook," *Brain Sciences*, vol. 13, no. 10, p. 1462, 2023.

[2] H. Yuan, H. Yu, S. Gui, and S. Ji, "Explainability in graph neural networks: A taxonomic survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 5, pp. 5782–5799, 2022.

[3] E. Dai, T. Zhao, H. Zhu, J. Xu, Z. Guo, H. Liu, J. Tang, and S. Wang, "A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability," *Machine Intelligence Research*, vol. 21, no. 6, pp. 1011–1061, 2024.

[4] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[5] X. Li, Y. Zhou, N. Dvornek, M. Zhang, S. Gao, J. Zhuang, D. Scheinost, L. H. Staib, P. Ventola, and J. S. Duncan, "Braingnn: Interpretable brain graph neural network for fmri analysis," *Medical Image Analysis*, vol. 74, p. 102233, 2021.



[1] https://github.com/matheo-angelo/IC/blob/main/Code/BrainGNN_experiments/BrainGNN_explainability.ipynb
[2] http://preprocessed-connectomes-project.org/abide/download.html