From ATR-FTIR Spectra to Visibility Graphs: An End-to-End GNN Pipeline for ASD Detection

Lucas G. T. Araújo[1], **Robinson Sabino-Silva**[2], **Murillo G. Carneiro**[1]

[1]Faculdade de Computação
Universidade Federal de Uberlândia, Uberlândia, MG, Brasil

[2]Departamento de Fisiologia, Instituto de Ciências Biomédicas
Universidade Federal de Uberlândia, Uberlândia, MG, Brasil

lucas.teodoro@ufu.br , robinsonsabino@gmail.com , mgcarneiro@ufu.br

Abstract. *Autism Spectrum Disorder (ASD) lacks objective biomarkers, with diagnosis relying on behavioural observation and taking years. Salivary ATR-FTIR spectroscopy offers a non-invasive molecular fingerprint, but prior graph-based methods depend on hand-crafted topological features. We propose an end-to-end GNN pipeline that encodes each spectrum as a windowed visibility graph with a five-dimensional node feature vector and evaluates five architectures under stratified group cross-validation. GCN achieves* $\mathrm{F1}_{\mathrm{MH}} = 0.810$ *in cross-validation and GIN* 0.71 *on the held-out test, competitive with prior graph-based approaches without hand-crafted features, establishing end-to-end GNN classification of ATR-FTIR spectra as viable for ASD detection.*

1. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition characterised by deficits in social communication, restricted and repetitive patterns of behaviour, and altered sensory processing [Association et al. 2014]. Its aetiology is multifactorial, combining polygenic risk with environmental factors. Early diagnosis is critical because timely behavioural and educational interventions substantially improve long-term outcomes. Yet the current gold standard (clinical evaluation guided by DSM-5 criteria) depends entirely on specialist observation and can take years, with outcomes sensitive to evaluator experience and availability [BRASIL 2021].

This limitation has motivated sustained interest in objective, biomarker-driven diagnostic support. Saliva is a particularly attractive substrate: it is collected non-invasively, contains over 3,000 proteins, messenger RNAs, microRNAs, and more than 700 microbial species, and mirrors the individual's physiological and pathological state. Attenuated Total Reflectance Fourier-Transform Infrared Spectroscopy (ATR-FTIR) yields a high-dimensional spectral fingerprint of a sample's molecular composition without reagents or special preparation. ATR-FTIR has been successfully applied to type 2 diabetes, oral cancer, Zika, COVID-19, and ASD [Caixeta et al. 2023, Lima-Filho and Carneiro 2023, Lima Filho et al. 2024, Oliveira et al. 2023, Santos-Jr et al. 2023, Silva et al. 2020].

Graph-based representations have proven powerful for spectral classification. The VisG2 framework converts each spectrum into a visibility graph, constructs a meta-graph from inter-spectrum similarities, and feeds hand-crafted topological features into classical classifiers, achieving $\mathrm{F1} > 0.70$ on the ASD dataset [Filho et al. 2025]. Graph Neural

Networks (GNNs) are the natural end-to-end extension of this paradigm: they operate directly on the graph structure without manual feature engineering, learning discriminative embeddings through iterative neighbourhood aggregation [Wu et al. 2022]. Despite their success in molecular property prediction, drug discovery, and brain connectivity analysis, no prior work has applied GNNs to ASD detection from salivary ATR-FTIR spectra; existing graph-based approaches on this modality rely on hand-crafted topological features rather than end-to-end learned embeddings.

This paper fills that gap with the following contributions:

- **An end-to-end ATR-FTIR-to-GNN pipeline** comprising windowed visibility graph encoding, a five-dimensional node feature vector enriched with spectral gradient and neighbourhood statistics, and joint optimisation of preprocessing, graph construction, and model parameters in a single Optuna search;
- **A systematic benchmark** of four canonical GNN architectures (GCN, GAT, SAGE, GIN) under identical experimental conditions, including stratified group cross-validation that explicitly prevents triplicate data leakage;
- **HybridSAGE**, a novel architecture cascading mean→max→sum SAGEConv layers with residual connections, designed to exploit aggregation diversity in low-feature spectral graphs;
- **A joint hyperparameter search** via Optuna TPE (≈200 trials/model) simultaneously covering architectural parameters, Savitzky-Golay smoothing, and the visibility graph window, to our knowledge the first such joint search in this domain.

2. Related Work

The proposed pipeline builds on advances in ATR-FTIR spectroscopy for biomarker detection, graph-based spectral encoding, and GNN architectures for biomedical classification.

[Silva et al. 2020] established the dataset used here and demonstrated that salivary ATR-FTIR spectra carry statistically significant biochemical differences between ASD and neurotypical groups, motivating spectroscopic features for diagnosis. In a preliminary study, [Araújo et al. 2024] evaluated high-level classifiers based on KNN-Graph and PageRank importance on the same dataset, reaching $\mathrm{MH} = 0.74$ and identifying visibility graphs as a promising direction for future work. Lima Filho et al. pursued this direction with VisG2, a meta-graph-of-visibility-graphs framework that encodes each spectrum as a visibility graph and constructs a meta-graph from their pairwise similarities, achieving $\mathrm{F1} > 0.70$ at IJCNN 2025 [Filho et al. 2025].

The present work advances this line of research along two axes: (1) it replaces hand-crafted topological features with GNN-learned node embeddings, enabling end-to-end optimisation; and (2) it enriches each graph node with a five-dimensional feature vector, going beyond the single absorbance value used in prior baselines on this dataset.

[Lima-Filho and Carneiro 2023] applied graph-based and network-based classifiers to salivary ATR-FTIR for oral cancer, reporting accuracy above 70% and specificity near 80% [Lima-Filho and Carneiro 2023, Lima Filho et al. 2024]. [Caixeta et al. 2023] used SVM on salivary ATR-FTIR to screen type 2 diabetes with high sensitivity. [Santos-Jr et al. 2023] achieved sensitivity and specificity above 90% for COVID-19 using a CNN on blood ATR-FTIR, demonstrating that deep learning can excel on this modal-

ity when sufficient data are available, a finding that motivates investing in GNN-based approaches as ASD datasets grow.

3. Theoretical Background

This section provides the formal background underlying the proposed pipeline. We describe the ATR-FTIR spectroscopy modality, the windowed visibility graph construction that transforms each spectrum into a graph, and the four canonical GNN architectures used as baselines alongside the proposed HybridSAGE.

3.1. ATR-FTIR Spectroscopy

FTIR spectroscopy measures absorption of mid-infrared radiation (4000–400 cm^{-1}) to probe vibrational modes of chemical bonds, producing a high-dimensional molecular fingerprint. The ATR modality operates via total internal reflection at a crystal-sample interface, eliminating the need for transmission-mode sample preparation [Silva et al. 2020]. Diagnostically relevant bands in saliva include Amide I (1660–1630 cm^{-1}, C=O stretch of proteins), Amide II (1560–1510 cm^{-1}, N-H bend), and the carbohydrate/nucleic acid fingerprint region (1200–900 cm^{-1}) [Silva et al. 2020, Araújo et al. 2024]. The lipid C-H stretching region (3050–2800 cm^{-1}) has also been implicated in ASD-related biochemical alterations.

3.2. Windowed Visibility Graph

The Visibility Graph (VG) algorithm [Lacasa et al. 2008] maps an ordered sequence $\mathbf{m} = (m_1, \ldots, m_n)$ to an undirected graph $\mathcal{G}(V, E)$ with $|V| = n$. Nodes v_i and v_j $(i < j)$ are connected if and only if no intermediate value m_c $(i < c < j)$ obstructs the line of sight:

$$m_c < m_j + (m_i - m_j)\,\frac{j - c}{j - i} \tag{1}$$

The plain VG produces $O(n^2)$ edges for n-point spectra. We use a *windowed* variant that restricts connections to pairs with $|i-j| \leq w$, reducing edge density to $O(nw)$ and making large spectra computationally tractable.e The window w is treated as a hyperparameter and jointly optimised with model architecture and preprocessing parameters.

3.3. Graph Neural Networks

A GNN with K message-passing layers operates on a graph $\mathcal{G} = (V, E)$ with node feature matrix $X \in \mathbb{R}^{|V| \times d}$. At each layer k, each node v aggregates embeddings from its neighbourhood $\mathcal{N}(v)$ and updates its own embedding.

GCN [Kipf and Welling 2016] uses symmetric degree normalisation:

$$H^{(k)} = \sigma\Big(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(k-1)}W^{(k)}\Big) \tag{2}$$

where $\tilde{A} = A + I$ and $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$.

GAT [Veličković et al. 2017] computes attention-weighted sums:

$$\alpha_{ij} = \mathrm{softmax}_j\big(\mathrm{LeakyReLU}\big(a^\top [Wh_i \,\|\, Wh_j]\big)\big) \tag{3}$$

with multi-head attention ($K = 4$ in layer 1, $K = 1$ in layer 2).

GraphSAGE [Hamilton et al. 2018] inductively aggregates sampled neighbourhoods:

$$h_v^{(k)} = \sigma\big(W^{(k)} \cdot \mathrm{CONCAT}\big(h_v^{(k-1)},\, \mathrm{AGG}\big(\{h_u^{(k-1)} : u \in \mathcal{N}(v)\}\big)\big)\big) \tag{4}$$

where $\mathrm{AGG} \in \{\mathrm{mean}, \mathrm{max}, \mathrm{sum}\}$ is selected by Optuna.

GIN [Xu et al. 2019] achieves maximal expressive power by aligning with the Weisfeiler-Leman graph isomorphism test:

$$h_v^{(k)} = \mathrm{MLP}^{(k)}\left((1 + \epsilon^{(k)}) \cdot h_v^{(k-1)} + \sum_{u \in \mathcal{N}(v)} h_u^{(k-1)}\right) \tag{5}$$

After K layers, a differentiable global pooling function aggregates all node embeddings into a single graph-level vector for classification.

4. Proposed Methodology

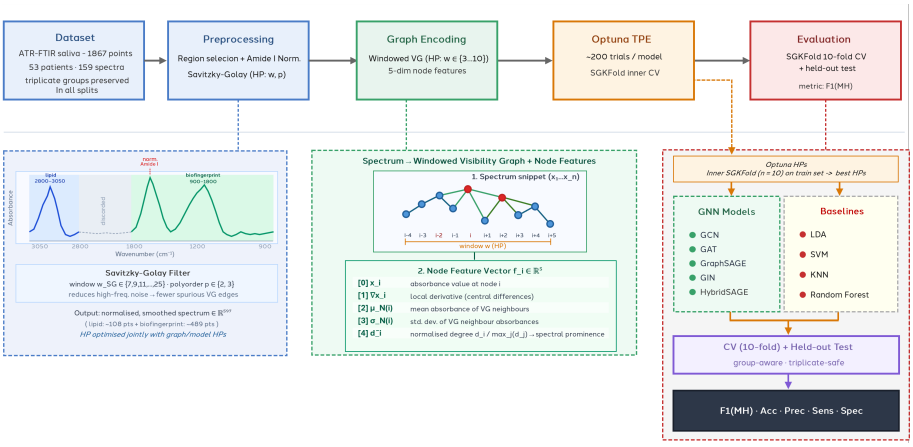

Figure 1. Experimental pipeline for ASD detection from salivary ATR-FTIR spectra. *Left panel*: spectral preprocessing showing region selection (lipid 3050–2800 cm^{-1} and biofingerprint 1800–900 cm^{-1}) and Savitzky-Golay smoothing, with window w_{SG} and polynomial order p optimised jointly with model hyperparameters. *Centre panel*: windowed visibility graph construction and five-dimensional node feature vector $\mathrm{f}_i \in \mathbb{R}^5$. *Right panel*: Optuna TPE search feeding best hyperparameters into parallel GNN and baseline evaluation branches, followed by 10-fold stratified group cross-validation (group = patient, stratify = ASD/NT) and a fixed held-out test.

Figure 1 illustrates the full experimental pipeline. Raw salivary ATR-FTIR spectra are first preprocessed by selecting two biochemically relevant spectral regions, normalising by the Amide I peak, and applying Savitzky-Golay smoothing. Each preprocessed spectrum is then encoded as a windowed visibility graph, where each node carries a five-dimensional feature vector combining local spectral and topological information. The resulting graphs are fed into five GNN architectures (GCN, GAT, SAGE, GIN, and the proposed HybridSAGE) alongside four classical baselines (LDA, SVM, KNN, RF). All

models are optimised via Optuna TPE, which searches jointly over architectural parameters, preprocessing hyperparameters, and the visibility graph window. Final evaluation uses two complementary protocols: a 10-fold stratified group cross-validation that preserves the triplicate structure of the dataset, and a fixed external held-out test. The following subsections detail each component.

4.1. Dataset

The dataset [Silva et al. 2020] comprises 159 salivary ATR-FTIR spectra from 53 patients (ASD and neurotypical), each providing three independent samples collected in triplicate, yielding 53 patient groups of 3 records each. Each raw spectrum contains 1,868 absorbance values covering the full mid-IR range.

4.2. Spectral Preprocessing

Region selection. Two spectral regions are retained and concatenated: the lipid C-H stretching region (3050–2800 cm^{-1}) and the biofingerprint region (1800–900 cm^{-1}), reducing each spectrum from 1,868 to approximately 597 points. Region boundaries are mapped to array indices from the wavenumber header of the data file, ensuring instrument-agnostic extraction.

Amide I normalisation. Each truncated spectrum is divided by the peak absorbance in the Amide I subregion (1660–1630 cm^{-1}):

$$x_i^{\mathrm{norm}} = \frac{x_i}{\max_{j \in \mathcal{R}_{\mathrm{I}}} x_j} \tag{6}$$

This suppresses inter-sample intensity variability arising from differences in sample concentration and crystal contact quality.

Savitzky-Golay smoothing. A Savitzky-Golay (SG) filter (derivative order 0) is applied after normalisation. The window length w_{SG} and polynomial order p are treated as hyperparameters and jointly optimised by Optuna. The filter serves a critical role in the context of visibility graph construction: high-frequency instrumental noise introduces spurious local oscillations that artificially inflate the degree of nodes in smooth spectral regions, confounding the geometric signal encoded by $\tilde{d}_i$. By attenuating these oscillations while preserving band shape, the SG filter ensures that $\tilde{d}_i$ reflects the true local curvature geometry of the spectrum rather than measurement artefacts. This is supported by the HP importance analysis (Table 1), where SG window size is the most influential parameter for GIN (0.274), the architecture most sensitive to neighbourhood cardinality.

4.3. Spectrum-to-Graph Encoding and Node Features

Each preprocessed spectrum $\mathbf{x} \in \mathbb{R}^n$ is mapped to a windowed visibility graph $\mathcal{G}(V, E)$ with $|V| = n$ nodes using Equation 1 and window w (also optimised by Optuna). Each node v_i carries a five-dimensional feature vector:

$$\mathbf{f}_i = \left[x_i,\ \nabla x_i,\ \mu_{\mathcal{N}(i)},\ \sigma_{\mathcal{N}(i)},\ \tilde{d}_i \right]^\top \tag{7}$$

where x_i is the absorbance; ∇x_i is the local spectral derivative (central differences); $\mu_{\mathcal{N}(i)}$ and $\sigma_{\mathcal{N}(i)}$ are the mean and standard deviation of absorbance values over the visibility-graph neighbours of node i; and $\tilde{d}_i = d_i / \max_j d_j$ is the normalised node degree, which

encodes local curve geometry: nodes in regions of smooth, gradual spectral variation accumulate more visibility edges and thus higher degree, since no intermediate node obstructs the line of sight; conversely, nodes at sharp, narrow absorption peaks have low degree because their steep flanks block visibility to all but the nearest neighbours. This feature vector provides the GNN with local spectral context (gradient), topological neighbourhood statistics (mean/std), and structural position (degree), going substantially beyond the single absorbance feature used in prior GNN baselines on this dataset.

4.4. HybridSAGE Architecture

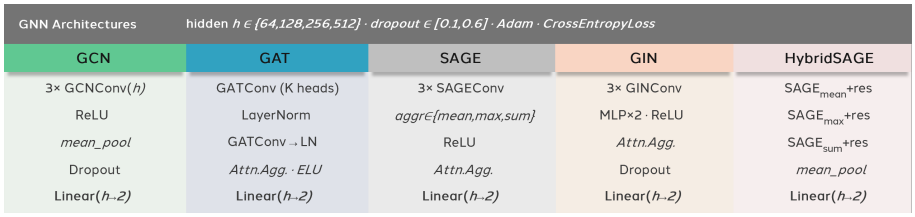

Figure 2. Layer diagrams of the five GNN architectures evaluated. All models receive node features $\mathbf{f}_i \in \mathbb{R}^5$, apply two or three graph convolution layers, perform global mean pooling, and pass the graph embedding through a linear classifier. HybridSAGE† (proposed) cascades three SAGEConv layers with heterogeneous aggregators (mean, max, sum) and residual skip connections (dashed arrows). Italic entries denote Optuna-selected operations. † Architecture proposed in this work.

HybridSAGE addresses a limitation common to all four baseline architectures: each uses a single, fixed aggregation operator throughout all layers. We propose cascading three SAGEConv layers with heterogeneous aggregators and residual connections:

$$\begin{aligned} h^{(1)} &= \mathrm{ReLU}\big(\mathrm{SAGE}_{\mathrm{mean}}(h^{(0)})\big) \\ h^{(2)} &= \mathrm{ReLU}\big(\mathrm{SAGE}_{\mathrm{max}}(h^{(1)})\big) + h^{(1)} \\ h^{(3)} &= \mathrm{ReLU}\big(\mathrm{SAGE}_{\mathrm{sum}}(h^{(2)})\big) + h^{(2)} \end{aligned} \tag{8}$$

The *mean* aggregator in layer 1 provides a smooth, bias-reduced initialisation of neighbourhood representations. The *max* aggregator in layer 2 sharpens discriminative features by retaining only the most activated dimension per neighbourhood. The *sum* aggregator in layer 3 introduces cardinality sensitivity: larger neighbourhoods produce larger activations, which is important for visibility graphs where high-degree nodes correspond to regions of smooth spectral variation, whose cardinality signal may differ between ASD and neurotypical spectra. The additive residuals $h^{(k)} + h^{(k-1)}$ mitigate oversmoothing and facilitate gradient flow through all layers.

The design rationale is that no single aggregator captures all relevant structure of ATR-FTIR spectra simultaneously: *mean* alone dilutes local spectral variation; *max* alone ignores global spectral shape; *sum* alone is sensitive to noise. The cascade allows each layer to specialise progressively, while residual connections ensure that information from earlier aggregation stages is preserved.

4.5. Training Protocol

Data splitting. The 53 patient groups are partitioned using `StratifiedGroupKFold` with 10 folds and `random_state=42`, ensuring: (i) all three triplicates of each patient remain in the same fold; (ii) the ASD/NT class ratio is approximately preserved in every fold; and (iii) each sample appears in the test set exactly once across all folds. This is strictly more principled than repeated hold-out splits when the number of groups is small.

Class imbalance. The cross-entropy loss is weighted by the inverse class frequency:

$$w_c = \frac{N}{C \cdot N_c} \tag{9}$$

where N is the total number of training samples, $C = 2$, and N_c is the count of class c in the training partition. Weights are recomputed per fold to reflect the actual training class distribution at each split.

Optimisation. Adam optimiser with per-model learning rate η and weight decay λ, both selected by Optuna. Mini-batches of size 16 with shuffling. Maximum 300 epochs with early stopping (patience: 30 epochs on training loss), restoring the best checkpoint.

Hyperparameter search. Optuna TPE sampler with ≈200 trials per model searches jointly over: architecture parameters (hidden size $\in \{64, 128, 256, 512\}$, dropout $\in [0.1, 0.6]$, η, λ, model-specific parameters), SG smoothing ($w_{\mathrm{SG}} \in \{7, 11, 15, 21, 25\}$, $p \in \{2, 3\}$), and the visibility graph window ($w \in \{3, 5, 7, 10\}$). Each trial is evaluated via $n_{\mathrm{inner}} = 10$ inner folds of `StratifiedGroupKFold` with `random_state=0` (distinct from the outer evaluation folds at `random_state=42`), minimising overlap between optimisation and evaluation.

Classical baselines (LDA, SVM, KNN, RF) follow the same preprocessing and CV protocol; SVM, KNN, and RF use a dedicated 200-trial Optuna search on the same inner folds. LDA has no hyperparameters to optimise and is applied directly.

5. Experiments and Results

This section describes the evaluation protocol, reports hyperparameter search results, and presents classification performance from two complementary perspectives: internal cross-validation and external held-out testing. A unified discussion then interprets the results across models and evaluation protocols.

5.1. Evaluation Metrics

This study adopts the classical evaluation metrics of accuracy, precision, sensitivity (or recall) (Sens) , and specificity (Spec), in addition to the $\mathrm{F1}_{\mathrm{MH}}$ score, defined as the harmonic mean between sensitivity and specificity.

$$\mathrm{F1}_{\mathrm{MH}} = \frac{2 \cdot \mathrm{Sens} \cdot \mathrm{Spec}}{\mathrm{Sens} + \mathrm{Spec}} \tag{10}$$

$\mathrm{F1}_{\mathrm{MH}}$ is the harmonic mean of sensitivity and specificity [Guimarães 1985] and serves as the primary metric. Unlike the standard F1 (harmonic mean of precision and recall), it penalises equally missed ASD cases and false alarms among neurotypicals, both of which carry significant costs in a clinical decision-support context. All CV results are reported as mean ± standard deviation over the 10 folds.

5.2. Hyperparameter Search Results

After ≈200 Optuna trials per model, two consistent patterns emerged across all five architectures: all converged to the maximum visibility graph window $w = 10$, suggesting that medium-range spectral relationships are informative for ASD classification; and all preferred polynomial order $p = 3$ for the Savitzky-Golay filter, favouring smoother spectra without sacrificing band shape.

Table 1 shows the relative HP importance from the Optuna fANOVA estimator. The learning rate dominates for GCN (0.455), consistent with GCN's sensitivity to gradient scale under normalised aggregation. For HybridSAGE and SAGE, the graph window is the most influential parameter (0.379 and 0.278 respectively), indicating that aggregation-based architectures benefit most from topological richness. For GIN, the SG window importance (0.274) rivals the graph window (0.240), suggesting that GIN's MLP-based aggregation is more sensitive to spectral smoothness than to graph topology.

Table 1. Relative HP importance (fANOVA, Optuna) per model. Bold: most influential HP per model. † Architecture proposed in this work.

Model	LR	WD	Hidden	Dropout	Win.	Other
GCN	**0.455**	0.255	0.023	0.061	0.060	SG-W: 0.130
HybridSAGE†	0.078	0.112	0.128	0.065	**0.379**	SG-W: 0.192
SAGE	0.087	0.061	0.152	0.102	0.278	*aggr*: **0.320**
GAT	0.073	**0.254**	0.070	0.165	0.160	SG-W: 0.198; K: 0.061
GIN	0.140	0.069	0.206	0.064	0.240	SG-W: **0.274**

5.3. Classification Results

Results are presented from two complementary perspectives: internal cross-validation, which estimates expected generalisation across all possible patient splits; and external held-out testing, which provides a single fixed-partition evaluation using the Optuna-selected HPs.

Internal cross-validation (StratifiedGroupKFold, 10 folds). Table 2 reports mean ± SD over the 10 folds. Each fold test set contains 5–6 patient groups never seen during training or HP search, with all triplicates kept intact.

Table 2. GNN and baseline performance, internal stratified group CV (10 folds, mean ± SD). GNN HPs from Optuna (≈200 trials); baseline HPs from Optuna (100 trials, same CV protocol). LDA: no HP search (no tunable parameters). Primary metric: $\mathrm{F1_{MH}}$. Bold: best per column. † Architecture proposed in this work.

Model	Type	$\mathbf{F1_{MH}}$	Acc.	Prec.	Sens.	Spec.
GCN	GNN	$\mathbf{0.810 \pm 0.182}$	0.817 ± 0.182	0.768 ± 0.251	$\mathbf{0.867 \pm 0.208}$	0.817 ± 0.217
HybridSAGE†	GNN	0.768 ± 0.307	$\mathbf{0.823 \pm 0.196}$	$\mathbf{0.743 \pm 0.346}$	0.850 ± 0.320	$\mathbf{0.822 \pm 0.234}$
SAGE	GNN	0.727 ± 0.307	0.790 ± 0.200	0.690 ± 0.347	0.833 ± 0.307	0.767 ± 0.260
SVM	Baseline	0.724 ± 0.191	0.765 ± 0.183	0.817 ± 0.283	0.750 ± 0.250	0.833 ± 0.269
GAT	GNN	0.704 ± 0.290	0.728 ± 0.232	0.667 ± 0.333	0.767 ± 0.318	0.739 ± 0.263
KNN	Baseline	0.688 ± 0.282	0.793 ± 0.147	0.763 ± 0.338	0.650 ± 0.320	0.900 ± 0.182
RF	Baseline	0.661 ± 0.301	0.753 ± 0.201	0.693 ± 0.341	0.650 ± 0.320	0.817 ± 0.241
GIN	GNN	0.652 ± 0.263	0.760 ± 0.117	0.627 ± 0.293	0.717 ± 0.334	0.794 ± 0.217
LDA	Baseline	*no internal CV (no hyperparameters to optimise)*				

Acc., accuracy; Prec., precision; Sens., sensitivity (recall); Spec., specificity. Rows ordered by $\mathrm{F1_{MH}}$ descending.

External held-out test. Table 3 consolidates all models on the fixed held-out partition. GNNs were trained on the full training set with Optuna-selected HPs and evaluated over 10 independent runs; baselines used Optuna HPs from 100 trials and were evaluated once.

Table 3. GNN and baseline performance on the external held-out test set, ordered by $\mathrm{F1_{MH}}$ descending. GNNs: mean over 10 independent runs. Baselines: single evaluation with Optuna-selected HPs. LDA: no HP search. Bold: best per column. † proposed in this work.

Model	Type	$\mathbf{F1_{MH}}$	Acc.	Prec.	Sens.	Spec.
GIN	GNN	**0.71**	0.73	0.62	**0.67**	0.76
KNN	Baseline	0.67	**0.82**	**1.00**	0.50	**1.00**
GAT	GNN	0.64	0.64	0.50	**0.67**	0.62
LDA	Baseline	0.60	0.64	0.50	**1.00**	0.43
GCN	GNN	0.59	0.64	0.50	0.50	0.71
HybridSAGE†	GNN	0.58	0.58	0.44	0.58	0.57
SVM	Baseline	0.58	0.58	0.45	0.75	0.48
SAGE	GNN	0.56	0.70	0.63	0.42	0.86
RF	Baseline	0.39	0.64	0.50	0.25	0.86

Acc., accuracy; Prec., precision; Sens., sensitivity (recall); Spec., specificity. Rows ordered by $\mathrm{F1_{MH}}$ descending.

5.4. Discussion

On the external test (Table 3), GIN leads with $\mathrm{F1_{MH}} = 0.71$, followed by KNN at 0.67. KNN achieves perfect specificity (1.00) and the highest accuracy (0.82) at the cost of low sensitivity (0.50), characterising a conservative classifier that avoids false positives but misses half the ASD cases, which is clinically undesirable in a first-pass diagnostic aid where missed cases carry higher cost than false alarms. LDA, with no HP optimisation, already reaches $\mathrm{F1_{MH}} = 0.60$ and perfect sensitivity (1.00), demonstrating that the spectral signal retains linear separability even without graph encoding. RF performs worst (0.39), reflecting overfitting on the high-dimensional feature space (≈597 features, 42 training patients).

On the internal CV evaluation, SVM achieves the best baseline $\mathrm{F1}_{\mathrm{MH}} = 0.724 \pm 0.191$, competitive with SAGE (0.727) and GAT (0.704). This indicates that GNNs' advantage over classical methods is present but not dominant under a fair comparison with identical preprocessing and HP protocol. GNNs' key structural advantage is end-to-end learning from graph topology without hand-crafted features, which is expected to become more pronounced as the dataset size grows beyond the current 53 patients.

The most important observation is the divergence in ranking and absolute values between Tables 2 and 3. In the CV setting, GCN leads with $\mathrm{F1}_{\mathrm{MH}} = 0.810$; on the external test, GIN leads with 0.71 and GCN drops to 0.59. This divergence has a principled explanation rooted in the small-sample regime. The CV estimate averages over 10 different test folds (≈5–6 patients each), covering the full diversity of the 53-patient cohort and providing a robust estimate of expected generalisation. The external test uses a single fixed partition whose specific composition determines the result substantially; with only ≈10 test patients, the variance of a single-partition estimate far exceeds that of its CV counterpart. Neither result should be interpreted in isolation: together, they bracket the plausible range of generalisation, with CV providing the expectation and the held-out test providing one sample from the distribution of possible outcomes.

GCN's CV leadership ($\mathrm{F1}_{\mathrm{MH}} = 0.810 \pm 0.182$) is reinforced by HP importance (Table 1): learning rate and weight decay jointly account for 71% of variance in GCN's objective, whereas hidden size (0.023) is nearly irrelevant. This indicates GCN's performance is driven almost entirely by optimisation quality, not model expressiveness, which is why the ≈200-trial Optuna search provides a large advantage, identifying the narrow lr/wd region where GCN's normalised aggregation is well-calibrated.

GIN's sum-based aggregation is maximally sensitive to neighbourhood cardinality, given that high-degree nodes in the visibility graph correspond to regions of smooth spectral variation and contribute proportionally more to the graph embedding. This property appears particularly favourable for the specific held-out partition used here, and GIN's competitive CV performance (0.652) confirms that this advantage is not an artefact of the fixed split.

HybridSAGE achieves the best CV accuracy (0.823) and specificity (0.822), while dropping to $\mathrm{F1}_{\mathrm{MH}} = 0.58$ on the external test. The cascaded mean→max→sum design benefits from the variety of patient-level spectral patterns seen across CV folds, but appears sensitive to the specific training-set composition in the fixed split. The dominance of the graph window HP (importance 0.379) confirms that topological richness is essential for the multi-aggregator cascade to operate effectively.

The most direct comparisons are with prior graph-based work on the same dataset. A preliminary study [Araújo et al. 2024] using KNN-Graph and PageRank importance achieved $\mathrm{MH} = 0.74$, and VisG2 [Filho et al. 2025] subsequently reached $\mathrm{F1} > 0.70$ with a meta-graph of visibility graphs, both relying on hand-crafted topological features. The present pipeline matches these results ($\mathrm{F1}_{\mathrm{MH}} = 0.810$ in CV, 0.71 on the held-out test) while replacing manual feature engineering with GNN-learned embeddings, a qualitative advance regardless of the absolute metric difference. Note that direct numerical comparison is limited by protocol differences: [Araújo et al. 2024] uses repeated hold-out without group-aware partitioning, which may overestimate performance given the triplicate struc-

ture of the dataset. A previous GNN baseline on this dataset achieved $\mathrm{F1}_{\mathrm{MH}} = 0.57$ (GraphSAGE); the current protocol differs in three key aspects: a five-dimensional node feature vector, joint Optuna optimisation, and stratified group cross-validation.

The high SDs in Table 2 (e.g. ± 0.307 for HybridSAGE and SAGE) are a structural consequence of having only 5–6 test patients per fold; they reflect the irreducible variance of evaluating on very small test sets, not model instability. For clinical decision support, the most relevant pair is (Sens., Spec.): GCN in CV achieves $(0.867, 0.817)$, meaning $\approx$87% of ASD patients would be correctly flagged and $\approx$82% of neurotypicals correctly cleared before specialist confirmation, averaged over possible patient-split realisations.

6. Conclusion

This paper presented the first systematic study applying message-passing GNNs with end-to-end learned embeddings to ASD detection from salivary ATR-FTIR spectra, extending prior graph-based work that relied on hand-crafted topological features.

Under internal stratified group cross-validation, GCN achieved the best primary metric ($\mathrm{F1}_{\mathrm{MH}} = 0.810 \pm 0.182$) and sensitivity ($0.867 \pm 0.208$), while the proposed HybridSAGE led in accuracy (0.823 ± 0.196) and specificity (0.822 ± 0.234). On the fixed external held-out test, GIN achieved $\mathrm{F1}_{\mathrm{MH}} = 0.71$, competitive with the VisG2 baseline without hand-crafted features. The divergence between protocols is expected in the small-sample regime ($n = 53$ patients) and underscores the necessity of reporting both CV and held-out results for a complete picture.

Key methodological findings: (i) the five-dimensional node feature vector substantially improves over single-absorbance representations; (ii) all models converge to the maximum graph window ($w = 10$), highlighting the importance of medium-range spectral interactions; (iii) sum aggregation is preferred for both SAGE and GIN, identifying neighbourhood cardinality as a relevant structural signal; (iv) Savitzky-Golay smoothing is a significant HP for GIN, while GCN's performance is dominated by learning rate.

Future directions include: **(1)** nested cross-validation with per-fold Optuna optimisation for provably unbiased HP-selection estimates; **(2)** dataset expansion to reduce fold variance and unlock GNN capacity; **(3)** spectral data augmentation via interpolation or variational autoencoders.

Acknowledgment

Authors thank the financial support given by the Brazilian National Council for Scientific and Technological Development - CNPq (408216/2022-0, 420212/2023-0 and 445027/2024-0), the Minas Gerais Research Foundation – FAPEMIG (BDT-00010-25), and the INCT in Theranostics and Nanobiotechnology (CNPq-465669/2014-0).

References

Araújo, L. G., Sabino-Silva, R., and Carneiro, M. G. (2024). High-level classification using complex networks for autism spectrum disorder detection. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pages 331–341. SBC.

Association, A. P. et al. (2014). *DSM-5: Manual diagnóstico e estatístico de transtornos mentais*. Artmed Editora.

BRASIL, M. d. S. (2021). Definição - transtorno do espectro autista (tea) na criança. Acessado em: 18/01/2025.

Caixeta, D. C., Carneiro, M. G., Rodrigues, R., Alves, D. C. T., Goulart, L. R., Cunha, T. M., Espindola, F. S., Vitorino, R., and Sabino-Silva, R. (2023). Salivary atr-ftir spectroscopy coupled with support vector machine classification for screening of type 2 diabetes mellitus. *Diagnostics*, 13(8):1396.

Filho, R. B. L., Sabino-Silva, R., and Carneiro, M. G. (2025). High-level classification based on meta-graph of visibility graphs for autism detection. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Guimarães, M. (1985). Exames de laboratório: sensibilidade, especificidade, valor preditivo positivo. *Revista da Sociedade Brasileira de Medicina Tropical*, 18:117–120.

Hamilton, W. L., Ying, R., and Leskovec, J. (2018). Inductive representation learning on large graphs.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Lacasa, L., Luque, B., Ballesteros, F., Luque, J., and Nuno, J. C. (2008). From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Sciences*, 105(13):4972–4975.

Lima-Filho, R. B. and Carneiro, M. G. (2023). Diagnóstico do câncer oral através da classificação de alto nível. In *Anais Estendidos do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 54–59. SBC.

Lima Filho, R. B., Fernandes, J. M., Ji, D., Zhao, L., Sabino-Silva, R., and Carneiro, M. G. (2024). High-level network-based detection of oral cancer from atr-ftir spectroscopy. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Oliveira, S. W., Cardoso-Sousa, L., Georjutti, R. P., Shimizu, J. F., Silva, S., Caixeta, D. C., Guevara-Vega, M., Cunha, T. M., Carneiro, M. G., Goulart, L. R., et al. (2023). Salivary detection of zika virus infection using atr-ftir spectroscopy coupled with machine learning algorithms and univariate analysis: A proof-of-concept animal study. *Diagnostics*, 13(8):1443.

Santos-Jr, A. P., Filho, A. C. M., Sabino-Silva, R., and Carneiro, M. G. (2023). Convolutional neural networks for the molecular detection of covid-19. In *Brazilian Conference on Intelligent Systems*, pages 51–62. Springer.

Silva, S. F. d. P. et al. (2020). Avaliação de biomarcadores salivares para diagnóstico de transtorno de espectro autista por espectroscopia atr-ftir.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wu, L., Cui, P., Pei, J., Zhao, L., and Song, L. (2022). *Graph neural networks*. Springer.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2019). How powerful are graph neural networks?