A Comparative Study of GNN Architectures and Self-Supervised Objectives for Embedding Extraction in Recommender Systems

Matheus H. C. Leme[1], **Giovani O. Santos**[1], **Vinicius S. Simm**[1], **Douglas R. Tanno**[1], **Marcos A. Domingues**[1]

[1]Department of Informatics – State Univerity of Maringá – Maringá – PR – Brazil

{pg55308, ra133166, pg404811,pg404806,madomingues}@uem.br

Abstract. *This study systematically evaluates item embeddings generated through 16 combinations of Graph Neural Network architectures and self-supervised objectives for recommender systems. Using datasets from two distinct domains, we analyze these structural representations both in isolation and as initialization for sequential modeling. Findings reveal that embedding effectiveness is highly sensitive to domain density. Ultimately, results demonstrate that hybridizing graph-based topological information with sequential patterns significantly enhances recommendation performance and catalog coverage, reinforcing the robust synergy between structural and temporal dynamics.*

1. Introduction

The transition to digital platforms has democratized access to a wide variety of content, but has also led to the well-known problem of information overload [Hagen 2015], a phenomenon frequently observed in e-commerce and movie streaming services. To mitigate this problem, Recommender Systems (RSs) have become essential tools for personalizing suggestions and supporting user decision-making [Wang et al. 2020a, Latifi et al. 2021].

The modeling of interactions in RSs has evolved to handle consumption sequences. In session-based RSs (where users are anonymous) or session-aware RSs (with available historical data), the objective is to predict the next items of interest based on bounded interactions [Domingues et al. 2022, Ludewig and Jannach 2018]. Traditionally, this task has been approached by purely sequential models, such as Recurrent Neural Networks (RNNs) [Hidasi et al. 2015]. However, these architectures struggle to model complex dependencies and transition patterns that deviate from a strict chronological order. To overcome these barriers, recent approaches, such as SR-GNN [Wu et al. 2019], propose representing sessions as graphs. By treating items as nodes and transitions as edges, this paradigm transcends the linear order of the sequence, enabling the capture of global connectivity and higher-order topological relationships among the consumed items.

The effectiveness of this approach may depend on the chosen Graph Neural Network (GNN) architecture, such as GCN [Kipf and Welling 2016a] or GAT [Veličković et al. 2017]. These networks learn vector representations, commonly referred to as embeddings, by iteratively aggregating information from the topological neighborhood of nodes [Shin et al. 2023]. Since training GNNs is computationally expensive, in industrial scenarios with low-latency requirements [Covington et al. 2016], a two-stage approach is commonly adopted: precomputing embeddings offline and retrieving them online through nearest neighbor search [Ying et al. 2018]. In addition to architecture design, self-supervised training methods, such as Deep Graph Infomax [Veličković et al. 2018] and contrastive learning [You et al. 2020], show strong potential for generating high-quality embeddings, although their effectiveness in the recommendation domain still lacks systematic evaluation.

On the other hand, it is important to highlight the difficulty of GNNs in capturing non-static sequential relationships within the graph, which may result in unsatisfactory metrics in specific next-item recommendation tasks, the primary objective of a large portion of studies in recommender systems. Therefore, to enhance the use of these embeddings, there arises a need to consolidate hybrid architectures that combine GNN representations with models capable of aggregating sequential information.

Given this scenario, the main contribution of this work is a large-scale empirical analysis aimed at mapping the current landscape and limitations of this class of models. We seek to answer the following questions: (RQ i) What is the actual impact of the GNN architecture on embedding quality? (RQ ii) How do different self-supervised training objectives influence performance? (RQ iii) Is there a promising synergy between architectures and training objectives? and (RQ iv) What are the effects of hybridizing embeddings with sequential models?

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 describes the methodology, architectures, and experimental protocol. Section 4 presents our main findings. Section 5 addresses the study limitations, and finally, Section 6 concludes the work and outlines future directions.

2. Related Works

Recent studies have employed well-established GNN architectures combined with complementary techniques, with the explicit goal of generating high-quality embeddings for session-based recommendation tasks. The use of GNNs has shown strong success in session-based recommendation, becoming a fundamental tool across various recommendation scenarios.

The industrial viability of scaling embeddings to billions of items was initially demonstrated by integrating random walks and graph convolutions [Ying et al. 2018]. Building upon this foundation, subsequent research has focused on optimizing and simplifying GNN architectures for recommendation tasks. Notably, it has been shown that core components of traditional GNNs, such as feature transformations and nonlinearities, can be safely omitted. Instead, learning embeddings solely through neighborhood propagation and aggregation has proven highly effective in collaborative settings [He et al. 2020]. Taken together, these advancements underscore the flexibility and power of graph-based approaches for modeling user-item interactions across diverse contexts.

To further enhance the representational capacity of GNNs, recent studies have combined multiple learning paradigms. For instance, embedding quality can be substantially improved by applying spatio-temporal contrastive learning over perturbed views of the session graph in conjunction with global collaborative signals [Wan et al. 2023]. Furthermore, the modeling of user behavior can be enriched by extracting latent vectors from distinct local and global graphs, subsequently fusing them through soft attention mechanisms to generate contextualized complementary representations [Wang et al. 2020b].

Graph enrichment strategies also play a pivotal role in this structural evolution. The direct incorporation of item features into co-occurrence graphs, followed by the extraction of latent representations using unsupervised convolutional networks, has emerged as an effective method for improving embedding quality [Peintner et al. 2025]. Along similar lines, robust representations for next-item prediction can be achieved by integrating multiple topological views, spanning local, global, and hypergraph levels. By capturing item transitions across these diverse contexts and refining the aggregated embeddings through contrastive learning objectives, the resulting models exhibit greater robustness [Wang et al. 2023].

As we can see, the literature is rich in novel models for session-based recommendation, GNN architectures, and training objectives. Although recent benchmarking efforts have begun to rigorously compare end-to-end GNN-based algorithms under unified evaluation protocols [Shehzad and Jannach 2024], these components are often proposed and evaluated in isolation. Thus, a significant gap remains in systematic and empirical investigation of the synergy between different GNN architectures and various training objectives, particularly with respect to embedding quality.

3. Methodology

In this section, we detail the graph construction, embedding extraction architectures, training methods, integration with the sequential model, and the experimental protocol.

Figure 1 illustrates the overall view of the proposal evaluated in this research. The two designed approaches start identically, with embedding extraction through the session graph, and differ in the subsequent steps. In one strategy (highlighted by the blue arrow), the obtained embeddings are directly used to generate recommendations, whereas in the other one (highlighted by the green arrow), the embeddings are used as input to the sequential model and enhanced with sequential information, before following the same path as the former, namely, recommendation generation. Further details are provided in the following subsections.

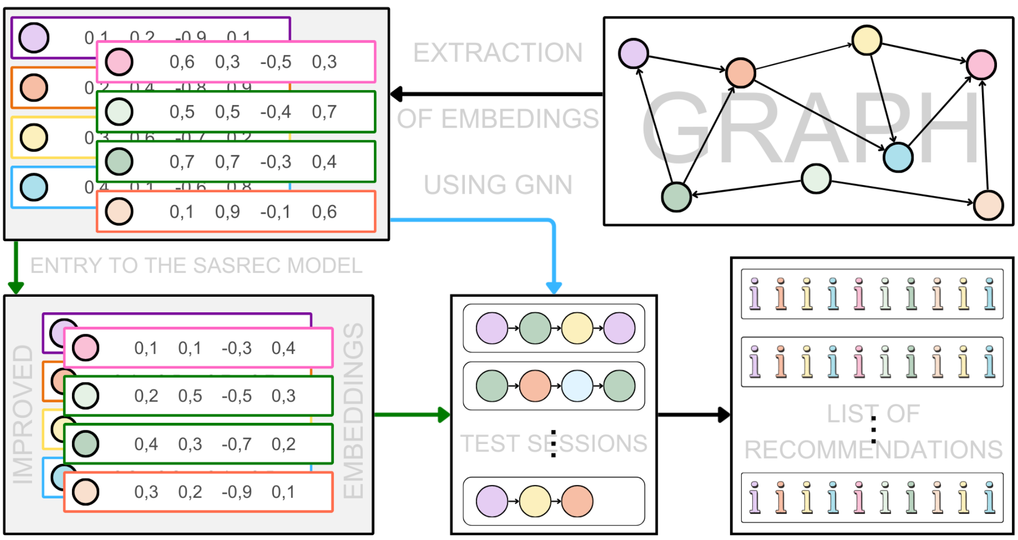

Figura 1. Methodology proposed for this work.

3.1. Graph Construction and Feature Initialization

The foundation of this research is the representation of user sessions as a directed graph $G = (V, E)$, where V denotes items (nodes) and E denotes sequential transitions (edges) within a single session. An edge (v_i, v_{i+1}) is added if item v_{i+1} was consumed after v_i. To initialize node vector attributes, we avoid random initialization. Instead, we apply the Word2Vec model (Skip-gram architecture with negative sampling) directly to the real consumption sequences, treating sessions as sentences. This process pretrains embeddings that capture behavioral semantics and real co-occurrence patterns. Finally, L2 normalization is applied to the generated matrix to ensure numerical stability in subsequent stages.

3.2. Graph-Based Embedding Extraction Models

We evaluate four GNN architectures to refine item representations, whose layer update rules differ in structural expressiveness. Graph Convolutional Networks (GCN) [Kipf and Welling 2016a] performs isotropic aggregation and a first-order approximation of spectral convolution: $H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right)$, where $\tilde{A}$ is the adjacency matrix with self-loops, $\tilde{D}$ is its degree matrix, and $W^{(l)}$ are trainable weights. Graph Attention Networks (GAT) [Veličković et al. 2018] extends isotropic aggregation by dynamically learning the importance of each neighbor via self-attention (*self-attention*), computing coefficients $\alpha_{ij} \propto \exp\left(\sigma\left(\vec{a}^{\top}[W\vec{h}_i \,\|\, W\vec{h}_j]\right)\right)$. GraphSAGE [Hamilton et al. 2017] focuses on inductive learning and scalability, samples a fixed number of neighbors $\mathcal{N}_v$ and applies an aggregation function before concatenation: $\vec{h}_v^{(l+1)} = \sigma\left(W^{(l+1)} \cdot \text{CON}\left(\vec{h}_v^{(l)}, \text{AGGRE}(\{\vec{h}_u^{(l)}, \forall u \in \mathcal{N}_v\})\right)\right)$. Graph Isomorphism Networks (GIN) [Xu et al. 2018] is designed to maximize discriminative power (comparable to the Weisfeiler-Lehman test), updates nodes through Multi-Layer Perceptrons (MLPs): $\vec{h}_v^{(l+1)} = \text{MLP}^{(l)}\left((1+\epsilon^{(l)}) \cdot \vec{h}_v^{(l)} + \sum_{u \in \mathcal{N}_v} \vec{h}_u^{(l)}\right)$.

To train these architectures in a self-supervised manner, we explore four training objectives: DGI [Veličković et al. 2018], which maximizes mutual information between local representations and a global graph summary; GRACE [Zhu et al. 2020], a contrastive model that separates distinct nodes and brings together stochastic views (via masking and edge dropping) of the same node; Laplacian Eigenmaps (LE) [Belkin and Niyogi 2003], which preserves local spectral proximity while reducing dimensionality; and VGAE [Kipf and Welling 2016b], which uses variational inference to reconstruct the graph structure by modeling the latent node distribution.

3.3. Hybridization with a Sequential Model (SASRec)

To investigate the synergy between structural representations learned by GNNs and explicit temporal dynamics, we adopt an adaptation of the Self-Attentive Sequential Recommendation (SASRec) model [Kang and McAuley 2018].

The innovation of this stage lies in the initialization of its embedding layer. Instead of random initialization, the matrix extracted by the GNNs is used as input to SASRec. Crucially, these weights are not frozen, meaning that the sequential model can dynamically learn and adjust these embeddings via backpropagation. This allows the network to

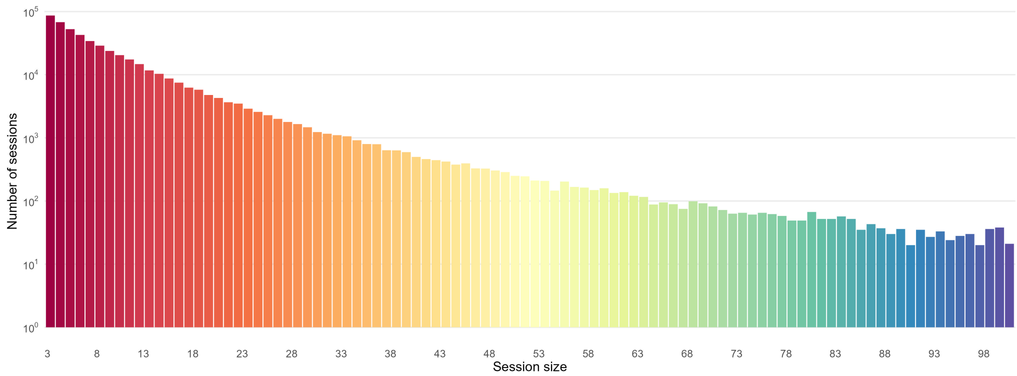

Figura 2. Session size distribution.

inherit the graph's rich topology and actively refine latent vectors to capture strict transitions. The sequences (truncated at $L = 20$) receive positional encoding and pass through 2 Transformer encoder blocks (with 4 attention heads and a dropout rate of 0.3) using a causal mask, optimized for 10 epochs using Cross-Entropy.

3.4. Experimental Protocol and Optimization

In our empirical evaluation, we used two public datasets preprocessed to remove users with fewer than 2 sessions and sessions with fewer than 3 items. Therefore, at least one session from each user will be present in the training and testing data for the models (detailed in the next subsection), and each session will express the user's intent by analyzing at least three interactions; otherwise, the scenario becomes very challenging, as the models have limited information to train with [Wang et al. 2021]. Furthermore, the sessions were defined based on a user inactivity period of 30 minutes [Cooley et al. 1999]. The datasets, belonging to the music and e-commerce domains, are Music4all [Santana et al. 2020] (14,066 users and 98,427 songs) and Retailrocket [Zykov et al. 2022] (1,404,179 users and 234,838 products, using only view interactions). Under these conditions, Figure 2 presents the distribution of the number of songs per session after processing the Music4all dataset. It can be observed that the dataset contains sessions with a minimum of 3 and a maximum of 100 interactions[1], and between 80% to 90% of sessions contain at most 15 interactions.

Hyperparameter selection for the 16 combinations of GNNs and training objectives was conducted via Bayesian optimization (TPE) using the Optuna library [Akiba et al. 2019]. The objective function was the maximization of the AUC metric in the link prediction task over a subgraph containing 20% of the data. The search space evaluated learning rate ($\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$), embedding dimensions ($\{64, 128, 256\}$), number of layers ($\{2, 3, 4\}$), dropout ($\{0.1, 0.2, 0.3, 0.4, 0.5\}$), activation function (ReLU, LeakyReLU), number of attention heads for GAT ($\{2, 4, 8\}$), and temperature for GRACE ($\{0.1, 0.2, 0.3, 0.4, 0.5\}$).

[1]For better visualization, session length was capped at 100 interactions since there are unique sessions containing up to 467 songs. However, sessions larger than 100 represent less than 1% of the data.

3.5. Evaluation

To measure the predictive performance of the learned representations, we adopted the leave-one-out protocol. We used a strictly isolated test set composed of unseen sessions. For each of these sessions, the last item was defined as the target and the preceding items as context. Recommendation list generation was based on similarity in the latent space. Specifically, the temporary session profile was constructed by computing the arithmetic mean of the embeddings corresponding to the context items. Next, this normalized mean vector is compared against the entire item catalog using cosine similarity. The top-10 items with the highest similarity scores generate the recommendation lists.

The effectiveness of these lists was analyzed using a comprehensive set of metrics. First, retrieval and accuracy metrics: Hit Rate (HR), Precision, Recall, and F-Score, which measure the model's ability to retrieve the target item. Next, ranking quality metrics: Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (NDCG), and Mean Average Precision (MAP), which evaluate list ordering by penalizing lower-ranked recoveries. Finally, Coverage quantifies the proportion of unique recommended items relative to the total catalog [Domingues et al. 2025, Yin et al. 2023].

4. Results and Discussion

In this section, we present and analyze our main findings. To address the proposed research questions, the evaluation was divided into two main axes: the direct assessment of the latent space quality generated by the 16 combinations of GNNs and self-supervised methods, and the evaluation of predictive performance after the hybridization and refinement of these embeddings by the sequential model (respectively indicated by GNN and SASRec in the extractor column of tables of results). The results reflect the top-10 recommendations, a standard in the literature for measuring accuracy in short recommendation lists. Tables 1 and 2 consolidate the metrics for the music and e-commerce domains, respectively.

The first stage of our analysis isolates the ability of GNNs to learn coherent representations without relying on complex temporal mechanisms. A notable performance variation is observed depending on the combination of architecture and self-supervised loss function.

In the music domain, the combination of GraphSAGE with contrastive training (GRACE) produced the best isolated latent representation, achieving a Hit Rate of 0.272 and an NDCG of 0.156. GRACE enforces latent space uniformity by repelling negative samples, mitigating the dimensional collapse problem frequently observed in dense graphs. In contrast, local smoothing methods such as Laplacian Eigenmaps (LE) presented the worst metrics (HR of 0.073 with GraphSAGE), as they tend toward oversmoothing, making distinct items indistinguishable in the vector space.

In the e-commerce domain, the scenario changes drastically in favor of attention-based architectures combined with variational inference. The combination of the GAT architecture with VGAE training outperformed the baseline, achieving an HR of 0.455 and an NDCG of 0.343. Mathematically, this is justified by the GAT formulation, which, instead of aggregating neighbors with fixed weights (as in GCN), learns dynamic attention coefficients. In e-commerce platforms, where noise is frequent, such as accidental

Tabela 1. Performance evaluation of models in the musical domain.

Architecture	Training	Extractor	Precision	Recall	F-score	HR	MAP	MRR	NDCG	Coverage
GCN	DGI	GNN	0.016	0.162	0.029	0.162	0.086	0.086	0.104	0.251
		SASRec	0.031	0.311	0.056	0.311	0.127	0.127	0.170	0.263
GCN	LE	GNN	0.015	0.151	0.027	0.151	0.082	0.082	0.099	0.305
		SASRec	0.031	0.313	0.056	0.313	0.124	0.124	0.168	0.372
GCN	GRACE	GNN	0.020	0.209	0.038	0.209	0.102	0.102	0.127	0.239
		SASRec	0.031	0.314	0.057	0.314	0.128	0.128	0.171	0.261
GCN	VGAE	GNN	0.020	0.205	0.037	0.205	0.099	0.099	0.124	0.243
		SASRec	0.032	0.329	0.059	0.329	0.130	0.130	0.176	0.301
GAT	DGI	GNN	0.007	0.074	0.013	0.074	0.050	0.050	0.056	0.207
		SASRec	0.028	0.282	0.051	0.282	0.119	0.119	0.157	0.304
GAT	LE	GNN	0.007	0.079	0.014	0.079	0.055	0.055	0.061	0.264
		SASRec	0.030	0.304	0.055	0.304	0.123	0.123	0.165	0.295
GAT	GRACE	GNN	0.013	0.136	0.024	0.136	0.075	0.075	0.089	0.214
		SASRec	0.030	0.308	0.056	0.308	0.126	0.126	0.169	0.293
GAT	VGAE	GNN	0.014	0.148	0.026	0.148	0.082	0.082	0.098	0.199
		SASRec	**0.033**	0.334	0.060	0.334	0.130	0.130	0.177	0.330
GIN	DGI	GNN	0.008	0.083	0.015	0.083	0.055	0.055	0.062	0.277
		SASRec	0.030	0.309	0.056	0.309	0.125	0.125	0.168	0.339
GIN	LE	GNN	0.010	0.100	0.018	0.100	0.064	0.064	0.072	0.329
		SASRec	0.030	0.300	0.054	0.300	0.123	0.123	0.164	0.328
GIN	GRACE	GNN	0.006	0.063	0.011	0.063	0.042	0.042	0.047	**0.365**
		SASRec	0.031	0.312	0.056	0.312	0.126	0.126	0.169	0.331
GIN	VGAE	GNN	0.012	0.122	0.022	0.122	0.069	0.069	0.081	0.251
		SASRec	0.028	0.289	0.052	0.289	0.122	0.122	0.161	0.230
GraphSAGE	DGI	GNN	0.015	0.150	0.027	0.150	0.082	0.082	0.098	0.251
		SASRec	0.031	0.318	0.057	0.318	0.127	0.127	0.171	0.332
GraphSAGE	LE	GNN	0.007	0.073	0.013	0.073	0.052	0.052	0.057	0.334
		SASRec	0.030	0.304	0.055	0.304	0.123	0.123	0.165	0.309
GraphSAGE	GRACE	GNN	0.027	0.272	0.049	0.272	0.120	0.120	0.156	0.280
		SASRec	**0.033**	0.330	0.060	0.330	0.128	0.128	0.175	0.344
GraphSAGE	VGAE	GNN	0.022	0.226	0.041	0.226	0.107	0.107	0.135	0.242
		SASRec	**0.033**	**0.337**	**0.061**	**0.337**	**0.133**	**0.133**	**0.181**	0.268

clicks, the mathematical formulation of GAT allows it to potentially assign lower weights to casual transitions and higher weights to strong product co-occurrences. Notably, the GIN architecture showed inferior performance in both domains. Designed to maximize strict structural isomorphism, it proves inflexible for capturing the semantic similarity characteristic of user sessions.

The answer to our research question regarding the effects of hybridization reveals a key finding of this study. A substantial performance improvement was observed across all accuracy and ranking quality metrics when transitioning from isolated GNN-based retrieval (baseline) to synergy with the sequential model.

As a highlight, initializing SASRec with embeddings extracted from the GAT with VGAE combination increased HR and NDCG in RetailRocket, reaching 0.511 and 0.384, respectively, representing the best overall performance in the experiment. In the Music4All dataset, the GraphSAGE with VGAE combination increased the HR from 0.226 in the baseline to 0.337 with SASRec.

The prevalence of the VGAE objective function as the best initialization state for SASRec deserves deeper theoretical analysis. Unlike deterministic methods that map nodes to fixed points in space, VGAE imposes a Gaussian prior, learning both the mean and variance of the latent node distribution. This results in a continuous, dense, and highly re-

Tabela 2. Performance evaluation of models in the e-commerce domain.

Architecture	Training	Extractor	Precision	Recall	F-score	HR	MAP	MRR	NDCG	Coverage
GCN	DGI	GNN	0.040	0.40	0.074	0.408	0.267	0.267	0.301	0.610
		SASRec	0.046	0.46	0.083	0.461	0.307	0.307	0.344	0.618
GCN	LE	GNN	0.043	0.435	0.079	0.435	0.287	0.287	0.323	0.653
		SASRec	0.046	0.468	0.085	0.468	0.312	0.312	0.349	0.657
GCN	GRACE	GNN	0.043	0.435	0.079	0.435	0.285	0.285	0.321	0.610
		SASRec	0.045	0.457	0.083	0.457	0.300	0.300	0.338	0.617
GCN	VGAE	GNN	0.043	0.439	0.079	0.439	0.291	0.291	0.327	0.644
		SASRec	0.046	0.463	0.084	0.463	0.307	0.307	0.345	0.644
GAT	DGI	GNN	0.037	0.374	0.068	0.374	0.251	0.251	0.280	0.485
		SASRec	0.040	0.406	0.073	0.406	0.302	0.302	0.327	0.457
GAT	LE	GNN	0.035	0.353	0.064	0.353	0.234	0.234	0.262	0.567
		SASRec	0.047	0.472	0.085	0.472	0.311	0.311	0.350	0.584
GAT	GRACE	GNN	0.041	0.412	0.074	0.412	0.276	0.276	0.309	0.547
		SASRec	0.049	0.496	0.090	0.496	0.335	0.335	0.374	0.562
GAT	VGAE	GNN	0.045	0.455	0.082	0.455	0.307	0.307	0.343	0.655
		SASRec	**0.051**	**0.511**	**0.093**	**0.511**	**0.343**	**0.343**	**0.384**	0.664
GIN	DGI	GNN	0.038	0.384	0.069	0.384	0.265	0.265	0.294	0.490
		SASRec	0.044	0.442	0.080	0.442	0.313	0.313	0.345	0.500
GIN	LE	GNN	0.035	0.359	0.065	0.359	0.242	0.242	0.270	0.567
		SASRec	0.048	0.484	0.088	0.484	0.335	0.335	0.371	0.551
GIN	GRACE	GNN	0.015	0.152	0.027	0.152	0.124	0.124	0.131	**0.708**
		SASRec	0.039	0.397	0.072	0.397	0.286	0.286	0.314	0.554
GIN	VGAE	GNN	0.033	0.335	0.060	0.335	0.232	0.232	0.257	0.591
		SASRec	0.038	0.387	0.070	0.387	0.266	0.266	0.295	0.602
GraphSAGE	DGI	GNN	0.042	0.420	0.076	0.420	0.303	0.303	0.332	0.620
		SASRec	0.047	0.477	0.086	0.477	0.327	0.327	0.363	0.612
GraphSAGE	LE	GNN	0.033	0.335	0.061	0.335	0.235	0.235	0.259	0.588
		SASRec	0.046	0.469	0.085	0.469	0.323	0.323	0.359	0.538
GraphSAGE	GRACE	GNN	0.034	0.347	0.063	0.347	0.243	0.243	0.268	0.665
		SASRec	0.047	0.472	0.085	0.472	0.324	0.324	0.360	0.612
GraphSAGE	VGAE	GNN	0.042	0.428	0.077	0.428	0.297	0.297	0.329	0.646
		SASRec	0.045	0.451	0.082	0.451	0.310	0.310	0.344	0.622

gularized embedding space. When SASRec, equipped with causal masks and directional self-attention, receives these vectors, it avoids early local minima and experiences reduced overfitting, actively refining structural topology to capture the strict temporal evolution of user intent.

The comparison between domains revealed distinct architectural behaviors with respect to interaction network topology. In the Music4All dataset, characterized by smaller catalogs ($\approx 98{,}000$ songs) and denser, more repetitive sessions, where structural inheritance is high, meaning tracks from the same genre are often consumed in sequence, GraphSAGE achieved better performance. By sampling a fixed number of neighbors during aggregation, it acts as an empirical degree limiter, preventing highly popular tracks from dominating vector updates and preserving the identity of niche tracks.

In contrast, RetailRocket is characterized by extreme sparsity, with more than 234 thousand products and view interactions with high abandonment rates. In this sparse scenario, the attention mechanism of GAT proved highly beneficial, likely helping the hybrid model prioritize complementary product transitions over casual navigation steps.

Finally, a critical analysis of modern recommender systems requires observing metrics beyond raw accuracy. The concept of Coverage quantifies the proportion of the catalog that the model effectively recommends to users, serving as a direct indicator of popularity bias mitigation.

As observed in Table 2, the isolated GIN with GRACE model achieves excellent Coverage of 0.708 at the expense of predictive accuracy (HR of 0.152), behaving almost like a random recommender. In contrast, the best combination, GAT with VGAE and SASRec, balances this trade-off. In addition to achieving the highest accuracy in the empirical study, it delivered a Coverage of 0.664. This demonstrates that probabilistic GNN initialization enriches the latent space cohesively, enabling the sequential model to reliably retrieve long-tail items, promoting diversified and commercially viable recommendations.

By synthesizing the experimental evidence, we can directly address the research questions posed in this study. Regarding the impact of GNN architectures and training objectives (RQ i, ii, and iii), our results reveal a sensitive interplay between the chosen model and the topological density of the dataset. While GraphSAGE proved superior in dense environments by mitigating popularity bias through neighbor sampling, the attention mechanism of GAT was highly advantageous in sparse scenarios, demonstrating efficiency against noisy transitions. Furthermore, the choice of training objective proved to be more than a mere auxiliary step; objectives like VGAE provided a regularized and continuous latent space that served as a superior starting point for learning, whereas local smoothing methods like Laplacian Eigenmaps (LE) often led to indistinguishable embeddings. This confirms that the synergy between a specific architecture and its training objective is domain-dependent, with the GAT with VGAE and GraphSAGE with GRACE combinations emerging as the most robust configurations for e-commerce and music, respectively.

In response to the effects of hybridization (RQ iv), the data consistently demonstrate that structural embeddings from GNNs and temporal dynamics from sequential models are not redundant, but rather complementary. The transition from isolated GNN retrieval to a hybridized SASRec initialization yielded consistent improvements in all accuracy and ranking metrics. This phenomenon occurs because the GNN captures the global connectivity of the item-to-item graph, providing a warm-started and semantically rich initialization, while SASRec refines these representations to honor the strict chronological intent of the active session. Ultimately, this hybridization not only optimizes predictive precision but also enhances catalog coverage, proving to be a highly effective approach for overcoming the inherent limitations of purely structural or purely sequential recommenders.

5. Limitations

Despite the strong empirical results in favor of hybridization, this study presents limitations that define its scope and indicate directions for future work. First, the transition graph topology was constructed statically, based on the consolidated history of the training set. In real digital ecosystems, the continuous insertion of new products or music tracks introduces the classic cold-start problem. Since the initial node features were derived solely from sequential co-occurrences through the Word2Vec model, the system is unable to infer accurate topological representations for newly released items without retraining the latent space. The absence of metadata integration restricts the model's ability to generalize purely through content similarity.

Second, the combined architecture requires considerable computational resources, inherent to the quadratic complexity of the SASRec self-attention mechanism coupled with neighborhood aggregation in GNNs. To make optimization feasible, it was necessary to impose a maximum sequence truncation ($L = 20$), an architectural decision that, although standard in the literature, may fragment the understanding of long-term intentions in atypical or prolonged sessions. Finally, it was empirically observed that the success of self-supervised objective functions, particularly contrastive learning (GRACE) and variational inference (VGAE), is highly sensitive to hyperparameter calibration, requiring exhaustive optimization processes that may hinder the agile deployment of these solutions in real-time industrial platforms.

6. Conclusion and Future Work

This work conducted an extensive empirical investigation on embedding extraction through Graph Neural Networks and their hybridization with sequential models for session-based recommendation. By systematically evaluating 16 combinations of architectures and self-supervised training objectives, we showed the impact of topology on the quality of latent representations. We demonstrated that the impact of architecture depends on domain density, where GraphSAGE stands out in dense scenarios, and the attention mechanism of GAT proves indispensable in sparse graphs. Furthermore, we also showed that hybridizing these structural representations as initialization for the SASRec model produces substantial gains in accuracy and coverage. In the dense music domain, performance differences between top-tier objectives (e.g., GRACE and VGAE with GraphSAGE) were minimal. Although not statistically tested, their hybridization consistently outperforms isolated baselines, suggesting a topological saturation point rather than a failure in representational learning.

As future directions, we intend to expand this empirical analysis to a more diverse range of datasets, encompassing new domains and density variations to validate the generalization of the identified synergies. Beyond predictive accuracy, we aim to explore the capacity of GNNs to mitigate the shared-account problem involving multiple users. In scenarios where a single account is used by different individuals, recommendations become noisy due to overlapping interests. Our future proposal consists of leveraging the structural expressiveness of GNNs to identify behavioral signatures within active sessions. The objective will be to infer the identity of the current user based on immediate interactions, enabling the model to selectively retrieve from the historical data only the sessions and embeddings belonging to the specific user, thereby restoring personalization in shared usage environments.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Referências

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396.

Cooley, R., Mobasher, B., and Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and information systems*, 1:5–32.

Covington, P., Adams, J., and Sargin, E. (2016). Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198.

Domingues, M. A., de Moura, E. S., Marinho, L. B., and da Silva, A. (2022). Benchmarking session-based and session-aware recommender systems for jusbrasil. In *Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*, pages 145–148. SBC.

Domingues, M. A., de Moura, E. S., Marinho, L. B., and da Silva, A. (2025). A large scale benchmark for session-based recommendations on the legal domain. *Artificial Intelligence and Law*, 33(1):43–78.

Hagen, A. N. (2015). The playlist experience: Personal playlists in music streaming services. *Popular Music and Society*, 38(5):625–645.

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., and Wang, M. (2020). Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648.

Hidasi, B., Karatzoglou, A., Baltrunas, L., and Tikk, D. (2015). Session-based recommendations with recurrent neural networks.

Kang, W.-C. and McAuley, J. (2018). Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pages 197–206. IEEE.

Kipf, T. N. and Welling, M. (2016a). Semi-supervised classification with graph convolutional networks.

Kipf, T. N. and Welling, M. (2016b). Variational graph auto-encoders.

Latifi, S., Mauro, N., and Jannach, D. (2021). Session-aware recommendation: A surprising quest for the state-of-the-art. *Information Sciences*, 573:291–315.

Ludewig, M. and Jannach, D. (2018). Evaluation of session-based recommendation algorithms. *User Modeling and User-Adapted Interaction*, 28:331–390.

Peintner, A., Moscati, M., Parada-Cabaleiro, E., Schedl, M., and Zangerle, E. (2025). Unsupervised graph embeddings for session-based recommendation with item features.

Santana, I. A. P., Pinhelli, F., Donini, J., Catharin, L., Mangolin, R. B., Feltrim, V. D., Domingues, M. A., et al. (2020). Music4all: A new music database and its applications. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 399–404. IEEE.

Shehzad, F. and Jannach, D. (2024). Performance comparison of session-based recommendation algorithms based on gnns. In *European Conference on Information Retrieval*, pages 115–131. Springer.

Shin, Y.-M., Tran, C., Shin, W.-Y., and Cao, X. (2023). Edgeless-gnn: Unsupervised representation learning for edgeless nodes. *IEEE Transactions on Emerging Topics in Computing*, 12(1):150–162.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks.

Veličković, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., and Hjelm, R. D. (2018). Deep graph infomax.

Wan, Z., Liu, X., Wang, B., Qiu, J., Li, B., Guo, T., Chen, G., and Wang, Y. (2023). Spatio-temporal contrastive learning-enhanced gnns for session-based recommendation. *ACM Transactions on Information Systems*, 42(2):1–26.

Wang, F., Gao, X., Chen, Z., and Lyu, L. (2023). Contrastive multi-level graph neural networks for session-based recommendation. *IEEE Transactions on Multimedia*, 25:9278–9289.

Wang, S., Cao, L., Wang, Y., Sheng, Q. Z., Orgun, M. A., and Lian, D. (2021). A survey on session-based recommender systems. *ACM Computing Surveys*, 54(7):1–38.

Wang, S., Pasi, G., Hu, L., and Cao, L. (2020a). The era of intelligent recommendation: Editorial on intelligent recommendation with advanced ai and learning. *IEEE Intell. Syst.*, 35(5):3–6.

Wang, Z., Wei, W., Cong, G., Li, X.-L., Mao, X.-L., and Qiu, M. (2020b). Global context enhanced graph neural networks for session-based recommendation. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 169–178.

Wu, S., Tang, Y., Zhu, Y., Wang, L., Xie, X., and Tan, T. (2019). Session-based recommendation with graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 346–353.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2018). How powerful are graph neural networks?

Yin, Q., Fang, H., Sun, Z., and Ong, Y.-S. (2023). Understanding diversity in session-based recommendation. *ACM Transactions on Information Systems*, 42(1):1–34.

Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., and Leskovec, J. (2018). Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 974–983.

You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., and Shen, Y. (2020). Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823.

Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., and Wang, L. (2020). Deep graph contrastive representation learning.

Zykov, R., Noskov, A., and Anokhin, A. (2022). Retailrocket recommender system dataset.