Artist Similarity based on Heterogeneous Graph Neural Networks

Angelo Cesar Mendes da Silva, Diego Furtado Silva, Ricardo Marcondes Marcacini

***Abstract*—Music streaming platforms rely on recommending similar artists to maintain user engagement, with artists benefiting from these suggestions to boost their popularity. Another important feature is music information retrieval, allowing users to explore new content. In both scenarios, performance depends on how to compute the similarity between musical content. This is a challenging process since musical data is inherently multimodal, containing textual and audio data. We propose a novel graph-based artist representation that integrates audio, lyrics features, and artist relations. Thus, a multimodal representation on a heterogeneous graph is proposed, along with a network regularization process followed by a GNN model to aggregate multimodal information into a more robust unified representation. The proposed method explores this final multimodal representation for the task of artist similarity as a link prediction problem. Our method introduces a new importance matrix to emphasize related artists in this multimodal space. We compare our approach with other strong baselines based on combining input features, importance matrix construction, and GNN models. Experimental results highlight the superiority of multimodal representation through the transfer learning process and the value of the importance matrix in enhancing GNN models for artist similarity.**

***Index Terms*—Artist similarity, Artist Representation, Heterogeneous Graph, Graph Neural Networks, Musical data representation**

I. Introduction

An essential resource for popularizing a song in streaming platforms is the automatic recommendation based on the user's consumption profile [1]. The recommendation task in the Music Information Retrieval (MIR) context consists of searching for similar musical content according to specific features and indicating the most relevant ones to users according to a similarity criterion [2].

In the MIR context, the artist similarity task handles applications interested in discovering related artists [3] and potentially similar songs to recommend. In this task, multiple musical features, such as audio and lyrics, can represent the artist, and we need to learn how to build a representation to predict links between similar artists [4].

The multimodal representation learning process must result in a unified representation that concentrates on the semantic information of multiple related features. This process can be done by simply concatenating text and audio features or learning embeddings through multimodal deep learning as fusion methods [5], [6]. Methods based on Graph Neural Networks (GNN) have stood out when dealing with machine learning tasks due to their ability to incorporate features from the graph topology to enrich the learned representations [7], [8].

Heterogeneous networks are a well-known representation for manipulating multiple modalities of unstructured data [9]. In the context of musical data, we can explore heterogeneous networks by projecting each modality as layers of a graph where nodes have features and can be connected with nodes of different types. These connections are the main differences in using graphs for data modeling. They express a similarity feature between the instances that direct the graph-based representation learning process by indicating which nodes should have related features [10].

The work presented in [11] computes the similarity between artists using graph neural networks trained with triplet loss as an unsupervised training process. The work contribution is related to graph neural network architecture combining the topology of a graph of artist relationships with content features to embed artists into a vector space that encodes similarity. However, the task of artist similarity is underexplored in the GNN context.

In our work, we want to explore the multiple modalities from music content features to represent the artists in heterogeneous structure modeling on a graph. We use the artist relationships to build the graph topology and propose a novel feature based on these relationships, defined as the importance matrix. This matrix is analogous to a co-occurrence matrix and is designed to improve the weight of the edges while introducing a new feature modality to the nodes. We propose this matrix to increase the discriminative capacity of artist representations by using a new feature related to the context of the artist similarity task, explored as a link prediction approach.

We propose a two-stage representation learning process for musical data. The first stage is based on a network regularization method that propagates information between nodes according to the graph topology to fine-tune initial features or create new ones in nodes without features. The network regularization process soothes the problem of missing multimodal information in datasets. In sequence, we apply the regularized features on a Graph Neural Network (GNN) that aims to learn a unified representation of the data that incorporates semantic information from the topology of the networks and features from neighbor nodes. Finally, to compute the similarity between artists, we applied the learned representation to handle the link prediction task.

We used the 4MuLA benchmark dataset [12] to evaluate our approach, where the instances have songs and artist information. The data comprises acoustic and textual features of each song, where the artist does not have initial features. The edges among music nodes are defined by cluster information from audio and lyrics features, while the links among related artists are pre-defined in the dataset. We measured the AUPR (area under the precision and recall curve) performance of three learning scenarios to the GNN models to construct the multimodal graph-based representation for computing the artist similarity. Finally, we compared two approaches to build the initial artist representation using six input feature variations.

We compare the applicability of the proposed importance matrix concerning a random matrix and a Graph Attention Network (GAT), where we evaluate the relevance to building the matrix and if the learned information is related to the target task or general attention learned in GAT model is sufficient to handle with the task. We also adopted two relevant GNN-based methods as a baseline method to evaluate our proposed method to handle the artist similarity task. Our results show that multimodal representations learned from the proposed method can perform more significantly when represented by features obtained from the transfer learning process. We also note that using the proposed importance matrix is better than using no matrix, random matrix, or automatically learning importance with an existing graph attention network. Among the learning scenarios, we showed that the unsupervised scenario achieved a higher AUPR value, but the scenario directly related to the artist similarity task achieved better mean performance. We present an extensive experimental evaluation, with pairwise comparisons in each type of representation, importance matrix, and learning scenarios proposed in our work. In summary, our work has these main highlights:

- modeling of musical data that are composed by audio, lyrics, and artist features in a heterogeneous network
- a new multimodal and GNN-based representation for artists to handle the artist similarity task
- an importance matrix that aggregates information into edges and nodes and is formed by information related to the artist similarity task

II. Background and related work

Research in MIR has accompanied changes in the strategies of representing musical data [13] and increasingly aggregating richer information from different sources. The goal is to model this information so that it is possible to incorporate features from multiple modalities and to build a new representation that is more discriminative [14]. From this point, this representation is used as input to build machine learning models related to specific MIR tasks [15] or build representations for multi-task approaches [16].

Musical data representation. Musical data are characterized by features that emphasize different aspects related to musical perception [17]. For example, timbre-related features can help us identify genres, chromatic features highlight structures in plagiarism identification tasks, and arousal and valence levels can be computed to predict an emotion label from a song [18], [19]. In addition, we can explore transfer learning approaches using pre-trained models to learn embedding features that have knowledge from massive datasets applicable to several tasks or domains [20], [21]. In our work, we evaluate representing artists from musical data that involve features based directly on the audio signal spectrum, embeddings obtained from pre-trained models for audio and lyrics, and random walk-based features.

Heterogeneous graph-based representations. Recent advances in applications involving Graph Neural Networks (GNN) have motivated the creation of graph models capable of representing unstructured data due to the possibility of aggregating node and edge features from multiple layers into structured feature vectors [22], [23]. In MIR, tasks such as emotion recognition [24] and artist recommendation [11] are examples of applications where GNNs outperform results in the literature. Our work proposes modeling musical data into graphs to learn a new multimodal artist representation to handle link prediction tasks.

Artist similarity. Computing the artist similarity aims to extend the information to applications. The proposed methods measure the artist similarity build data representation according to a target application, for instance, audio and lyrics feature for query-based information retrieval system [25], encoded features from a graph topology for ranking of artists [26], or analyzing users feedback for artist recommendation [27]. In our work, we explore the multimodal representations and formulate the artist similarity as a link prediction task: given an artist, we want to estimate other artists that should and should not be linked.

This work proposes modeling musical data about a heterogeneous to learn a new multimodal graph-based artist representation and handle artist similarity based on a link prediction task. Furthermore, we explore alternatives to modeling musical data to represent artists collaborating with discussions in MIR about the relation between the choice of musical features and the use in specific tasks. The link prediction task is directly related to the graph structure and has been explored in the GNN context.

III. Modeling

The task of similarity between artists consists of finding a function $s(a,b)$ that estimates a similarity measure between each pair of artists (a,b). In our case, the created similarity is a link between the artists due to the structure proposed for data. Therefore, let $G = (V,E,W)$ be an undirected graph, where V represents the set of vertices, E represents the set of edges or links, $E \subseteq V \times V$, and W represents the binary weight of edges, indicating whether edges between two nodes exist or not. The observed links are represented in adjacency matrix A, where $A_{i,j} = 1$ if $(i,j) \in E$ and $A_{i,j} = 0$ otherwise. For any node, $x \in V$, let $\Gamma(x)$ be the 1-hop neighbors of x, and let $\hat{\Gamma}(x)$ be all nodes not linked with x. Given a node y, if $y \cdot x \rightarrow 1$, then $y \in \Gamma(x)$, while $y \cdot x \rightarrow 0$, then $y \in \hat{\Gamma}(x)$.

The nodes and edges have a pre-defined type in the proposed graph to structure the data. The possible node types are audio, artist, and lyrics. The node type denotes the features extracted from the modality with the identical term. There are five types of edges: audio-audio, audio-artist, artist-artist, artist-lyrics, and lyrics-lyrics. The edges do not have features, and the edge-type information supported only the artist link identification in the preprocessing step. In addition, in some evaluated scenarios, the graphs do not have nodes and edges related to lyrics; more details are in the subsection IV-A.

A. Musical data representation

In a multimodal scenario, the musical characteristics are obtained from different sources, have complementary semantics, and, when combined, assemble a musical representation that results in the ideal experience for its perception. A subset of features is commonly defined to represent the music in each task. This step is justified due to the pre-processing computational cost and the missing semantic relation between some features and tasks.

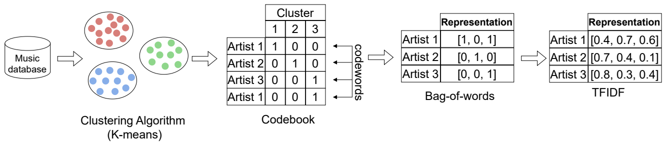
To compute artist similarity, we learn a new artist representation based on features of their songs and the constructed graph topology, evaluating unimodal and multimodal representations. We examined two options for the audio-based feature: in the first feature, we extract the root-mean-square (RMS) value from each frame of a melspectrogram to use as medium-level information, and as another feature, we use the pre-trained model Essentia [28] in a zero-shot learning process to learn an embedding space from raw audio files used as high-level information. We applied another pre-trained model for the textual-based feature, a BERT-based fine-tuned with lyrics data[1], to learn an embedding space to project the lyrics feature.

The reason for using RMS from the melspectrogram as the feature relates to its semantic content. The melspectrogram contains patterns associated with rhythms and timbres related to cultural aspects [29], [30] that induce criteria adopted to associate similar artists, such as musical genre or harmony [31]. For other high-level features, we are interested in exploring the transfer learning from a large dataset used in model training to build our initial features for lyrics and one more feature to characterize audio modality. All audio features are extracted from raw files with default parameters from the library Librosa[2].

B. Artist representation

To represent each artist, we consider all their related songs characterized by the abovementioned features. We evaluate some unimodal and multimodal approaches to build the representation: the average of audio embeddings only; the average of audio embeddings concatenated to the average of lyrics embeddings; a codebook of audio embeddings only; a codebook of audio embeddings concatenated to codebook of lyrics embeddings; a codebook of RMS melspectrogram only; a codebook of RMS melspectrogram concatenated to a codebook of lyrics embeddings.

The method proposed in [32] inspires our codebook-based artist representations. This representation is based on the classic bag-of-words representation of text data. The first step is quantizing the feature space to define the codewords from a set of feature subsequences considered candidate words. For this, the method uses the K-means algorithm with the word candidates and considers the center of each cluster as a codeword. Then, each candidate is associated with the codeword representing the cluster to which it belongs. A merge of all candidates for each artist defines the representation. In addition, we normalize this representation using the term frequency-inverse document frequency (TFIDF) method. This process is represented in Figure 1. In our context, the artists are represented by the codewords obtained among word candidates extracted from their different recordings.



	Cluster		
	1	2	3
Artist 1	1	0	0
Artist 2	0	1	0
Artist 3	0	0	1
Artist 1	0	0	1

	Representation
Artist 1	[1, 0, 1]
Artist 2	[0, 1, 0]
Artist 3	[0, 0, 1]

	Representation
Artist 1	[0.4, 0.7, 0.6]
Artist 2	[0.7, 0.4, 0.1]
Artist 3	[0.8, 0.3, 0.4]

Fig. 1. The process to construct the artist representation based on the codebook method. Initially, we use the k-means algorithm on the initial features of songs to define all codewords for each artist. In sequence, we apply a bag-of-word method to unify all artist songs in a feature vector. Finally, we normalize this vector using the TFIDF method.

IV. Artist similarity on heterogeneous graph

A. Heterogeneous graph to musical data

Modeling graph data is motivated by the possibility of structuring data as nodes that are projected in layers according to the modality of their features. Thus, we can exploit the graph topology to share complementary semantic information between neighbor nodes. The relation between the artist, audio, and lyrics nodes compose the proposed graph's topology. The artists are linked to lyrics and audio nodes, as a direct relation between artist and music, and there are links inside the artist layer, where the artist links were defined by "wisdom of the crowd" as information provided by users. The links inside audio and lyrics layers were inferred for the K-means algorithm, as explored in [24], where a model was trained to predict the cluster for each layer for each node, and nodes with equal clusters are linked. This topology is shown in Figure 2.

This work proposes a representation learning process in two steps: initially, we explore the network topology based on the network regularization process to share information among neighbor nodes, propagating features for nodes in distinct modalities; in sequence, the graph information is used as input for a GNN model learn a final representation based on regularized features and the edges. In both steps, we want to aggregate information from neighbors, but the first step will fine-tune the initial features; in the sequence, we integrate deep learning resources and direct the representation learning for the link prediction task.

[1] Public lyrics-bert model. Available in https://huggingface.co/brunokreiner/lyrics-bert

[2] Librosa. Available in https://librosa.org/doc/latest/generated/librosa.feature.rms.html

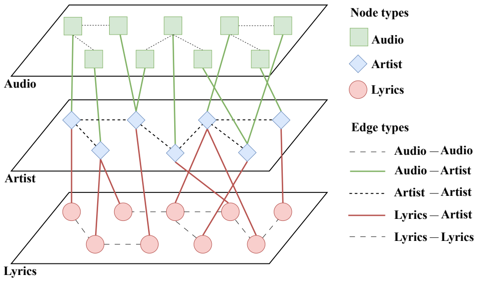


Fig. 2. The heterogeneous network topology proposed to structure music data. Our heterogeneous network has nodes typed as audio, artist, and lyrics, which have the initial features. The edges link nodes contained in each layer type and the audio and lyrics nodes with artist nodes.

B. Music network regularization

The network regularization process can propagate the node's features inside the network as an instance of label propagation methods [33]. Our proposal assumes two constraints: neighboring nodes must have similar features, and the regularized feature vector must be similar to the initial features. Formally, network regularization can be associated with a representation learning problem defined as learning a mapping function $f : x_i \rightarrow z_{x_i} \in R^d$, where z_{x_i} is the learned vector of the node $x_i \in V$ in the network. The Equation 1 defines the function to be minimized to learn the new space $Z \in R^d$, in which all nodes are mapped,

$$Q(\mathbf{Z}) = \frac{1}{2}\sum_{(x_s,x_t)\in R} w_{x_s,x_t}(\mathbf{z}_{x_s}-\mathbf{z}_{x_t})^2 + \mu\sum_{x_i\in V}(\mathbf{z}_{x_i}-\mathbf{x}_i)^2 \tag{1}$$

where R = {Audio-Audio, Audio-Artist, Artist-Artist, Artist-Lyrics, Lyrics-Lyrics} indicates the relationships in the proposed heterogeneous network and (x_s, x_t) indicates a pair of vertices, one from each relationship. The first term in the function computes the proximity between the feature vectors for each pair of linked nodes. The last term computes the distance between the regularized features in the space $\mathbf{Z}$ to initial features for each $x \in V$. The parameter μ determines the preservation information level for initial features. The higher value indicates the preservation of the original features, while lower values permit adjusting the features according to the network topology.

Equation (1) is applied for each modality, artist, audio, and lyrics in a step preceding the GNN model training. Therefore, at the end of the regularization process, we obtained Z_{artist}, Z_{audio}, and Z_{text} for all nodes in the network. Finally, we concatenate all spaces, $Z_{\oplus} = Z_{artist} \oplus Z_{audio} \oplus Z_{text}$. Although the nodes received features from nodes of different modalities and have a feature vector of the same dimensionality, we handle the graph in a heterogeneous composition, using the typing information of the nodes and edges in GNN model learning.

C. Graph Neural Networks for link prediction

The general purpose of graph-based learning methods is to build node representation that aggregates neighbors' features and information from graph topology. In our context, the features of the nodes are defined by $Z_{\oplus}$ resulting from the regularization process. The neighbor nodes are indicated by an adjacency matrix A. In particular, we want to evaluate graph-based artist representation learned directly to link prediction tasks or without a pre-defined task, where the representation can be applied in a multi-task scenario. The representation learned by the GNN is represented in the matrix $U \in R^{T\times D}$, where T is the number of nodes and D represents the dimension of the unified space learned. The GNN is formulated according to Equation 2,

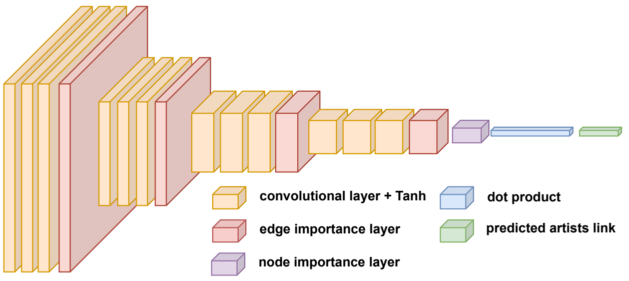


Fig. 3. The proposed architecture for the GNN model. We have four layers composed of three convolutional steps to pass messages between the neighbors. In sequence, the proposed new information layer is defined as the importance layer. This layer type is applied on edges and nodes. The output is a dot product that results in the score in the predicted link between artists. The information inputs are the $Z_{\oplus}$ as feature matrix, M as importance matrix, and the A as adjacency matrix.

$$H^{(l+1)} = f(H^{(l)}, A) = \alpha(AH^{(l)}W^{(l)}) \tag{2}$$

where $H^{(0)} = X$, l, represents the current layer, A is the adjacency matrix, $W^{(l)}$ define the weight in l-th layer in a neural network, and $\alpha(.,.)$ defines the activation function.

We process the adjacency matrix A to permit self-loop in nodes and to normalize the adjacency matrix in relation to node degrees. To permit the self-loop, we add A to identity matrix I, which result in $\hat{A}$. Thereby, the own node features are utilized in the representation learning process. The normalization process is realized by multiplying A with D^{-1}, where D indicates a diagonal matrix representing the node degrees. We use $D^{-\frac{1}{2}}$, for symmetric normalization. Thus, the adjacency matrix used in the proposed GNN is defined by Equation 3,

$$\hat{A} = D^{-\frac{1}{2}}SD^{-\frac{1}{2}} \tag{3}$$

where $S = A + I$, and I represents the identity matrix, and $D_{ii} = \sum_j S_{ij}$ indicates the node degree.

D. Importance Matrix

Relationships between artists can occur directly when two artists have an explicit link in the dataset or indirectly when two artists are contained in a set created on some similarity criterion, for example, a music playlist or tag sharing. Our work is based on features that allow us to build both types of relationships. When creating connections between artists, we access the *related_artist* attribute contained in the used dataset, which comprises a list of artists related to a song and its artist.

As related artists are also associated with a song and its artists simultaneously, the same artist can be related to different sets of artists. This relation type composes the $Artist\_Artist$ relation type in the A. The importance matrix is created from the list of related artists, where we count the number of times that pairs of artists are included in this list, like a co-occurrence matrix. Formally, the importance matrix M has dimensions $m \times n$, where m represents the number of instances and n represents the number of unique artists. The Equation 4 denotes how each element of the importance matrix is calculated based on the pairs of artists that occur together in the $related\_artist$:

$$M[i][j] = \sum_{m=1}^{n} \epsilon_{mij} \tag{4}$$

$$\epsilon_{mij} = \begin{cases} 1, & \text{if related\_artist}[m][i] = \text{related\_artist}[m][j] = 1 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $related\_artist[m][i]$ represents the value of the matrix with the list of related artists in the m-th row and i-th column, which can be 0 or 1. The importance matrix is filled with the sum of times the pair of artists i and j occur together in at least one observation, or 0 otherwise.

We propose to evaluate the potential of the importance matrix to add discriminative information to the edges and nodes of the graph. Aggregating the proposed importance information to edges in a GNN model is similar to the process of adding weight to edges, motivated by weighing edges with more occurrence. For this, we construct M by indexing the instances using the identifier of each artist, so $m = n$ and the relationships existent in M are equal to the artist relationships contained in A. The other relationship types in A have an importance value of 0 in M. Thus, we formalize the GNN model according to Equation 6:

$$H^{(l+1)} = f(H^{(l)}, \hat{A}, M) \tag{6}$$

where at each layer l the matrix M is inserted in the data representation learning process. Thus, the layer $H^{(l+1)} = H^{(L)} = U$ contains the newly learned space by GNN.

Before inserting the importance matrix into the model learning, we evaluated two approaches to normalizing M values. One approach is to use M as input to a neural network composed of one dense layer, with input and output dimensions equal to the number of nodes, and then normalize the output with a softmax. The other approach is to use the softmax function to normalize. The softmax is applied to each row of the matrix, so the sum of each row results in 1.

To add importance to the nodes, we propose to use an importance layer after the GNN model learning process. We use the matrix M as input to the importance layer that is composed of a $leaky\_relu$ activation function, and a softmax normalizes the output. This layer is similar to the traditional attention layer. In this scenario, M is initially formed by the number of instances m, and n is the number of artists. Formally, we denote in Equation 7 the process that concatenates the output of this importance layer into the output of the representation learned by the GNN for the nodes:

$$U' = U \oplus \alpha(M) \tag{7}$$

where U represents the output of the GNN model and $\alpha(.)$ represents the importance layer applied to the importance matrix M.

Finally, we propose the importance matrix to emphasize the relevance of edges between artists and the content of each song for an artist and their relationships in a heterogeneous network. Using information beyond musical content is a resource to aggregate discriminative information into data and enrich musical representations. This matrix aims to use information acquired about the data to induce some latent discriminative information contained in the features. Furthermore, the importance matrix is easily adaptable to other applications, especially in MIR, when building it from relationships extracted from playlists or user-generated metadata.

The proposed GNN architecture, illustrated in Figure 3, comprises three graph convolutional layers and a hyperbolic tangent activation function. The importance layer was introduced sequentially to aggregate information from artist relations on the edges and nodes. Two artists are defined as similar if there is a link between them. The link is indicated based on the score computed via dot product. The previous layer's output indicates the dimensionality of the vector input for each layer. The first layer receives the regularized features $Z_{\oplus}$ to produce the next embedding vector in the subsequent layers. The dimensionality of $Z_{\oplus}$ varies according to utilized features to represent the music and artist nodes. We show a table with all input feature variations in the Subsection V-A.

V. Experimental analysis

We used the tiny version of 4MuLA [12] as a benchmark dataset with 1,000 instances, 1,052 artist links pre-defined, and approximately 350,000 links, counting all edge types. This number depends on the audio feature used. In this version, each song has that audio information, with 30 seconds of duration represented by a melspectrogram and the lyrics. Other features and labels were not used in our proposal. In the explored context, we want to characterize artists through a graph-based learned embedded space composed of multimodal features. The primary motivation for using this dataset is the existence of annotated information about related artists and the availability of multimodal music features.

The central objective of our experiments is to evaluate the proposed method to learn a graph-based representation for artist nodes to compute the similarity among artists handling the link prediction task. For this, we want to evaluate the impact of the initial feature of the audio and lyrics nodes to characterize the artists with representations built based on the media or codebook of all artist songs. We made six combinations with the set of features to measure the impact on the GNN model performance and discuss the semantic complementarity among musical multimodal features. With the importance matrix proposal, we want to evaluate whether the importance matrix aggregates discriminative information when inserted only on nodes, edges, or both. Finally, we compute the performance of the learned

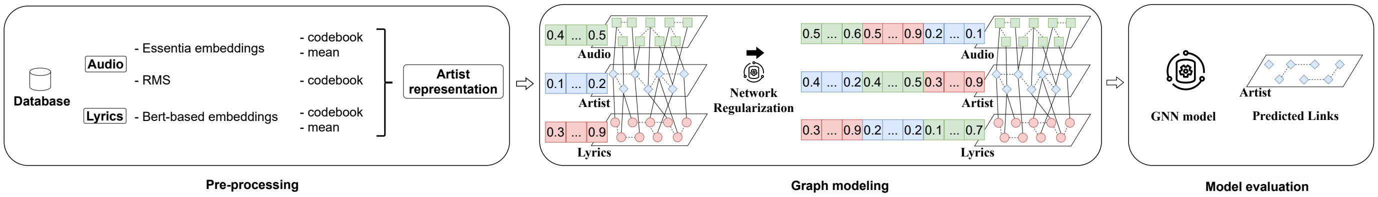


Fig. 4. Pipeline of experiment in three steps. The pre-processing stage is formed by extracting features from each song's audio and lyrics modalities. Each extracted feature is subjected to a quantization process, codebook, or means to compose the artist's representation. Next, we build graphs where the nodes have the input features for audio, lyrics, and artists, and the edges are obtained from labeled information existing in the dataset or built using a clustering algorithm. Six different combinations of initial feature sets were used to represent the artists. The regularization process executes the information propagation between neighbor nodes and results in the input features to the GNN model. Finally, we build GNN models for each created graph, and we evaluate these models to handle link prediction tasks.

representation when dealing with the link prediction task in three learning scenarios.

A. Experimental setup

The experimental setup is organized into three steps: pre-processing, modeling of graphs, and model evaluation. In the pre-processing step, we made the feature extraction and set the initial representation for all node types. The nodes are defined with the initial features for modeling graphs, and the regularization network process is realized to fine-tune and propagate information among the network based on existent edges. Finally, the graph is introduced into the GNN models to learn the musical representation and handle the link prediction task. These last two steps occur for each proposed representation approach. Figure 4 shows the steps for reproducing the experiments.

The audio, lyrics, and relations among artists are the information used to build our proposed graph musical modeling. The feature extraction process resulted in two representations for audio and one representation for lyrics. A combination of audio and lyrics features defines the artist's representation. For audio modality, we extract root-mean-square (RMS) from the melspectrogram using the library Librosa [34] with default parameters from the 30-second audio file, resulting in a feature vector with 1292 dimensions. Still, we also explore a transfer learning approach for audio modality using a pre-trained Essentia model based on the VGGish model trained on AudioSet [35] to build a music embedding based on the raw audio file, resulting in a feature vector with 128 dimensions. Finally, we explore transfer learning for textual modality and introduce the lyrics into the pre-trained language model fine-tuned into 480,000 lyrics data, resulting in a feature vector with 300 dimensions. Six different graphs were created according to each approach proposed to represent the artist. The relationships between artists-artists, artists-audio, and artists-lyrics are the same regardless of the audio and lyrics features. However, the audio-audio and lyrics-lyrics relationships are built based on the K-means algorithm; therefore, the two edge types change due to the clustering process being conditioned to the audio or lyrics feature.

The importance matrix is defined during graph construction. We have four distinct scenarios to evaluate the importance matrix. First, we have the learned matrix, which results from the initial matrix after being introduced into a dense layer and normalized with a softmax layer. Second, we have the normalized matrix, which the initial matrix just introduced to a normalization process with the softmax function. Third, we have a random matrix with the same dimensions and values between 0 and the maximum value of the initial matrix. Fourth, the scenario is used as a baseline approach, in which we do not use the importance matrix. As a top-line approach, we consider the attention matrix GAT learned since this model also learns an important value to feature. We evaluate the aggregation of information from the importance matrix with the adjacency matrix, with the function of adding weight to the edges, and also with a concatenation with the feature matrix of the nodes after the GNN convolution steps, to add other sources of discriminative information to nodes.

The artist's representation is formed based on your music features. We evaluate the representation built from the codebook and mean on audio and lyrics features. The feature in each modality is concatenated to result in artist representation. A summary of all the approaches evaluated for artist representations is presented in Table I. These representations are also explored in the network regularization process to be introduced as input in the GNN models with audio and lyrics node features. In addition, the final size of vector features for all node types is defined in the GNN model architecture. We utilize a meta-path-based representation created using the DeepWalk algorithm to compare our proposal. This baseline representation has only graph topology information without initial musical features.

The topology of the proposed graph is formed by the five types of relationships formulated in Section III. The edges among related artists and the edges from artist to audio and lyrics are present on all the graphs, independent of the features used for node representation. However, the edges inside the audio and lyrics layers are built based on the K-means clustering process based on the node features. The k value influences graph connectivity and the path used for information propagation. We evaluate the k with a value of 31, which refers to $\sqrt{N}$, where N is the dataset size.

In addition to artist representation approaches, we evaluate three learning scenarios: labeled as Supervised, this

Artist representation	Audio	Lyrics	Approach	Dimensionality
REP I	essentia	bert	codebook	2*k
REP II	essentia	bert	mean	428
REP III	essentia	-	codebook	k
REP IV	essentia	-	mean	128
REP V	RMS	bert	codebook	2*k
REP VI	RMS	-	codebook	k

TABLE I
THE PROPOSED ARTIST REPRESENTATIONS. THE FEATURES REPRESENT THE AUDIO CONTENT, FORMED BY THE ESSENTIA MODEL EMBEDDINGS OR RMS, AND LYRICS CONTENT, FORMED BY A BERT-BASED PRE-TRAINED MODEL. THE APPROACH REPRESENTS THE METHOD USED TO COMBINE THE FEATURES OF ALL SONGS OF ONE ARTIST. THE DIMENSIONALITY RESULTANT FOR ARTIST REPRESENTATION IS INDICATED IN THE LAST COLUMN.

scenario is direct to link prediction task, in which we know the artist links should or should not exist, and the loss function is defined by a score computed from the artist links predicted and ground truth links; labeled as Unsupervised this another scenario aiming to multitask approaches, in which the learned representation can be used in other MIR tasks, and the loss function is defined by a triplet loss computed on artist nodes features; labeled as Initialized in this scenario, we reproduce the Supervised scenario again initializing the GNN models with meta-path features for all nodes, without changing the edges. This scenario is motivated for work that reported an increase in performance in early initialized GNN models, as [36] that learned initial weights used as input to the model.

We evaluated two model training strategies. The first considered task-driven learning of link prediction. For Supervised and Initialized scenarios, the loss function is defined as a binary cross-entropy function that receives as a parameter the score computed from the dot product from nodes existing in positive neighbors $\Gamma(x)$ and negative neighbors $\hat{\Gamma}(x)$ and a target vector with equal size formed by binaries values. In the Unsupervised scenario, we want to evaluate a representation that can be applied to different tasks. For this, the loss function is defined as a triplet loss that receives anchors and positive and negative nodes as parameters and computes the loss considering the features. In this learning scenario, the loss is defined as $loss(a,p,n) = max\{d(a_i,p_i) - d(a_i,n_i) + \Delta, 0\}$, where a is the anchors, p is the positives, and n is the negatives nodes, and Δ refers to margin and defined as 0.2, similar the work [11]. To compute the evaluation metrics, all learning scenarios are evaluated as link prediction tasks in the testing stage.

The importance matrix is included in all approaches evaluated. In a process similar to an ablation study, we assess the link prediction task in four ways: without the matrix, using the matrix only on the edges, only on the nodes, and on the edges and nodes simultaneously. We propose three types of importance matrices: learned matrix, normalized matrix, and random matrix. The learned and normalized matrix was formulated in Section IV-D. In contrast, the random matrix is initialized by values from zero until the max edge importance value and normalized with the softmax function like the other two matrices. In this case, we want to evaluate whether the importance matrix adds discriminative information to the artists' representation when computing the AUPR variation concerning the importance matrix, the artist's initial representation, and the learning scenario.

Finally, we perform a cross-validation process to evaluate the representation and learning scenarios for handling link prediction. We divided the dataset into five folds according to links among artists. Therefore, the links between the artist, audio, and lyrics are maintained in all folds. Thereby, we assume that the features of the audio and lyrics nodes are relevant to learning graph-based representation for artist nodes, while the related artists may not. All scenarios are evaluated in equal folds, and the metric area under the precision and recall curve (AUPR) indicates the model performance to link prediction, as argued in [37]. As a baseline method, we built a graph attention network (GAT) model with the same architecture as the proposed GNN model and trained it with equal settings.

B. Results and discussions

Regarding the results obtained, we want to evaluate the representation learning process for artists based on the features of their songs to handle artist similarity based on link prediction tasks. We report the highest AUPR value according to the average of the results calculated over the five data folds, considering all experiment settings for each learning scenario. To simplify the table structure, we refer to the artist representation acronym shown in Table I. The subsequent tables are presented to represent an ablation study when dealing with the task of link prediction between artists. We evaluate the initial features and then incorporate the proposed representation learning resources to aggregate information based on the graph topology. Using the importance matrix, we analyzed the results to discuss strategies for representing artists and the impact on models' performance.

The realized experiments have an extensive set of evaluated parameters. We have six artist representations, five importance matrix types (learned, normalized, random, without matrix, and GAT), three learning scenarios (Supervised, Initialized, and Unsupervised), and five learning rates (from 10^{-1} until 10^{-5}). In addition, as the baseline methods, we adopted GraphSage [38] and PEAGNN [39] methods. The GraphSage-based methods present competitive results for link prediction applications, with a differential in your learning process due to a link sampling step. The PEAGNN has an architecture composed of GAT layers combined with a loss function to minimize the distance between similar objects, like a raking problem. Our dataset was split into 5-folds, where four folds are used in the training step, and the remaining fold is used in the testing step. The experiments report only one model architecture setting where the input layer has an input size equal to the feature dimensionality, and the subsequent layers have 512, 256, and 64 sequentially. We define the epoch number as 1000 to train the models, and in case the AUPR does not increase for 50 epochs, the early stopping criteria stops the training.

As a baseline result for evaluating the representation learning proposal for artists, we created the initial scenario in which we have a graph built only with the Artist layer, without the Audio and Lyrics layers. This experiment configuration allows us to measure the impact on the performance by including other layers and their relationships as multimodal and heterogeneous information and the inclusion of the GNN model. In Table II, the AUPR value is reported, considering that the nodes are represented by the initial features in scenarios with and without network regularization. In this case, we only measure the dot product between the node representations to define whether an edge should be created between them.

input feature	**no regularized**	**regularized**
REP I	0.66150	0.73689
REP II	0.68264	0.67654
REP III	0.72630	0.76312
REP IV	0.63425	0.63671
REP V	0.61522	0.66065
REP VI	0.56859	0.59719
random walk	0.52891	0.52773

TABLE II
THE AUPR PERFORMANCE COMPUTED ON THE SCENARIO WHERE THE INITIAL FEATURE IS REGULARIZED OR NOT REGULARIZED.

These results show that the representations where the audio modality is characterized by the Essentia model embeddings (REP I, II, III, and IV) performed better than the RMS feature (REP V and VI). This is an expected conclusion since the advantages of using transfer learning to build initial features for unstructured data are widely reported in the literature. We also observed that representations combining audio and text features performed better than audio-based representations. The random walk-based representation is used for comparison and presents the lowest performance, noting that more nodes and relationships are needed for the feature to be more representative. Furthermore, we observed that representations built with codebook resulted in an increase in performance after network regularization. This result is relevant to the work proposal because the codebook is based on clustering to construct the representation, and network regularization propagates features between neighboring nodes to refine the features of each node. Thus, we can interpret that the learned features are related to the node neighborhood. This process is similar to building node representations while a GNN model learns.

In the following tables, we report the results that express the representation performances after the GNN model learning process. Furthermore, the input graph is formed by layers with artists, lyrics, and audio nodes. Thus, we can measure the influence of multimodality on graph creation and the performance of a graph-based representation using a GNN model. We report the performance for each learning scenario when handling the link prediction task when using the importance matrix only on the nodes, only on the edges, or both, or without using the proposed matrix.

First, we report in Table III the performance of each learning scenario without using the importance matrix, like traditional GNN models, a GNN model composed for GraphSage layers, and the original architecture proposed for the PEAGNN method. PEAGNN had inferior performance, showing that the learning process for the link prediction task is more efficient for approximating similar artist representation. The GraphSage showed more competitive results in relation to GCN-based models, which were superior to Supervised and Unsupervised learning scenarios. However, Initialized achieved the best results, showing the discriminative potential of adding other features to the model's learning process. In general, the results from GNN-based representation are superior to those in the previous table, showing that learning using GNN models aggregates discriminative power to the representations when exploring the features of neighboring nodes. When comparing artist representations, REP I achieved the best performances, reinforcing using multimodal features to represent artists. We also highlight the increased performance for random walk-based representations so that both representation proposals achieve competitive performance.

input	**Supervised**		**Initialized**		**Unsupervised**	
feature	proposed	random walk	proposed	random walk	proposed	random walk
REP I	0.72306	0.71658	0.83677	0.84424	0.70390	0.72047
REP II	0.65638	0.68193	0.75863	0.71777	0.65384	0.74421
REP III	0.70945	0.80565	0.81765	0.79125	0.68462	0.71995
REP IV	0.63967	0.71813	0.74499	0.72783	0.64590	0.70915
REP V	0.66433	0.65331	0.73443	0.64850	0.67384	0.67448
REP VI	0.60509	0.60423	0.65197	0.63293	0.67384	0.67448

input	**GraphSage**		**PEAGNN**	
feature	proposed	random walk	proposed	random walk
REP I	0.75784	0.73333	0.60122	0.58053
REP II	0.75132	0.72368	0.57416	0.57905
REP III	0.76047	0.76776	0.57036	0.58063
REP IV	0.73655	0.74885	0.53820	0.56577
REP V	0.71003	0.74885	0.52137	0.57426
REP VI	0.71126	0.72422	0.50568	0.56854

TABLE III
THE AUPR PERFORMANCE COMPUTED ON THE SCENARIO WHERE THE INITIAL FEATURE IS INTRODUCED IN THE GNN-MODEL WITHOUT USING THE IMPORTANCE MATRIX AND THE GRAPHSAGE AND PEAGNN BASELINE METHODS.

Regarding the learning scenarios, the Initialized scenario presented the best performance, showing that the random walk-based representation increased the discriminative capacity of the GNN-based representation. This observation is relevant, as we can consider using random walk-based representation with another modality in a heterogeneous network and measure how much discriminative information can be added to representation in the regularization stages and subsequently be introduced into a GNN model.

Including the importance matrix in the edges aims to add weights with information related to the context of the artist similarity task. In general, the results reported in Table IV show a general increase in performance. The normalized importance matrix was predominant in all learning scenarios, indicating that the importance of artist relationships incorporated information into the convolutional process. However, the best result obtained in this scenario is inferior to the GNN model without the importance matrix (0.83677 vs. 0.81462 in the Initialized scenario with GNN-based embeddings). The Supervised and Initialized scenarios achieve superior results to the Unsupervised scenario, indicating that the importance matrix enhanced the artist's edge weight during the convolutional process, where the models are trained directly to link prediction. Finally, the proposed GNN-based representation was superior to the random walk-based in both learning scenarios, especially in the representations composed of multimodal features.

input feature	Supervised		Initialized		Unsupervised	
	normalized matrix		normalized matrix		normalized matrix	
	proposed	random-based	proposed	random-based	proposed	random-based
REP I	0.77269	0.77299	0.79553	0.74813	0.73858	0.72573
REP II	0.72596	0.72501	0.77763	0.70559	0.64345	0.69284
REP III	0.80518	0.78880	0.81462	0.76861	0.69068	0.73951
REP IV	0.71899	0.75642	0.81432	0.79076	0.65714	0.73101
REP V	0.61613	0.74333	0.62907	0.62102	0.61220	0.75905
REP VI	0.64975	0.76438	0.63211	0.62312	0.61800	0.63214

TABLE IV

THE AUPR PERFORMANCE COMPUTED ON THE SCENARIO WHERE THE INITIAL FEATURE IS INTRODUCED IN THE GNN-MODEL USING THE IMPORTANCE MATRIX ONLY ON THE EDGES.

When including information from the importance matrix only about the nodes, we observed results similar to the previous tables in Table V. In this evaluation scenario, including the importance matrix in the nodes aims to aggregate a new feature vector to the nodes. The results were generally superior to the scenario in which the GNN model was without the importance matrix. The best result obtained is similar to the scenario without the importance matrix (0.83677 vs. 0.83624), but this value was achieved in the Supervised learning scenario. Here, the Supervised learning scenario was superior to the others, and the GNN-based representation performed better than the random walk-based representation. Once again, the two Supervised and Initialized scenarios presented superior performance. However, the Unsupervised scenario improved performance compared to the previous table. These results allow us to infer that the aggregated importance over the nodes resulted in a representation with potential for use in multiple MIR tasks. This conclusion is related to the random composition in the importance matrix, where we note that the importance value is related to the task, but the semantic value is irrelevant. In contrast, using the importance matrix over the edges is more related to the target link prediction task.

	Supervised		Initialized		Unsupervised	
	random matrix		random matrix		learned matrix	
input feature	proposed	random-based	proposed	random-based	proposed	random-based
REP I	0.79178	0.80416	0.78944	0.78915	0.77403	0.74960
REP II	0.81166	0.80185	0.80472	0.79319	0.79544	0.79077
REP III	0.82710	0.80903	0.80131	0.80120	0.77537	0.75376
REP IV	0.82296	0.81974	0.81431	0.79325	0.78565	0.81364
REP V	0.75618	0.79507	0.77503	0.72828	0.75944	0.72824
REP VI	0.73127	0.83624	0.73894	0.73682	0.78551	0.79521

TABLE V

THE AUPR PERFORMANCE COMPUTED ON THE SCENARIO WHERE THE INITIAL FEATURE IS INTRODUCED IN THE GNN MODEL USING THE IMPORTANCE MATRIX ONLY ON THE NODES.

Finally, we report in Table VI the results for the scenario in which the importance matrix was embedded on nodes and edges. In this scenario, we also report the results when using a GAT model to perform link prediction as a comparison method. The attention mechanism contained in GAT allows the model to assign differentiated importance to the connections between nodes in a graph based on the information contained in these nodes and the relationships between them. This is the same motivation for using the proposed importance matrix, so we can discuss the relevance of building the matrix early or learning it during the GAT model learning process.

In an overview of the results obtained by the three learning scenarios, we can note a superior performance to the GAT model, and the GNN-based representation presents a greater discriminative capacity than the random walk-based representation. The importance matrix embedded on nodes and edges simultaneously increased the performance of all representations about the scenarios evaluated and reported in the previous tables. We can note that the best AUPR value in all scenarios evaluated was obtained by REP I and in the Unsupervised scenario. Thus, we can infer that the multimodal representation formed by an acoustic feature combined with lyric features based on the codebook approach resulted in more discriminative artist representations. These results indicate that the proposed importance matrix can be applied in other MIR tasks because the discriminative information gain of the proposed matrix is not directly dependent on the target task.

Unsupervised				
input feature	learned matrix		GAT	
	proposed	random-based	proposed	random-based
REP I	0.90184	0.87452	0.60916	0.69102
REP II	0.85554	0.83897	0.64912	0.68317
REP III	0.82452	0.81173	0.63346	0.67245
REP IV	0.80654	0.82963	0.62501	0.69606
REP V	0.81738	0.78898	0.62557	0.61659
REP VI	0.79749	0.80632	0.64490	0.61910

Supervised				
input feature	normalized matrix		GAT	
	proposed	random-based	proposed	random-based
REP I	0.84151	0.82494	0.75580	0.76233
REP II	0.88187	0.85549	0.75924	0.76606
REP III	0.85333	0.82170	0.79923	0.79034
REP IV	0.82198	0.79447	0.74062	0.75868
REP V	0.76801	0.80361	0.66154	0.70228
REP VI	0.68625	0.81612	0.64921	0.75821

Initialized				
input feature	normalized matrix		GAT	
	proposed	random-based	proposed	random-based
REP I	0.85229	0.84216	0.75664	0.74287
REP II	0.77147	0.79713	0.72971	0.73770
REP III	0.83398	0.83203	0.74705	0.71730
REP IV	0.77417	0.72965	0.74021	0.71390
REP V	0.79586	0.77012	0.74617	0.75111
REP VI	0.74322	0.72112	0.76072	0.73576

TABLE VI

THE AUPR PERFORMANCE COMPUTED ON THE SCENARIO WHERE THE INITIAL FEATURE IS INTRODUCED IN THE GNN MODEL USING THE IMPORTANCE MATRIX ON THE EDGES AND NODES.

In summary, to finalize these experiments, we can conclude that the multimodal representation achieved the best results; representing artists using a codebook approach is more discriminating than using an average; the importance matrix embedded only on the edges is more related to the link prediction task; when the matrix is embedded only in nodes it supports the learning of a representation with applicability in others MIR tasks; when present on the

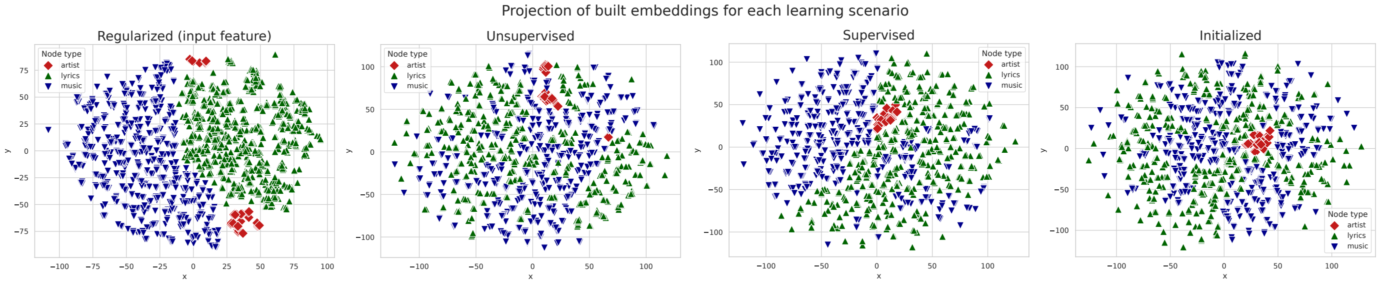


Fig. 5. Two-dimensional projections of learned embeddings (t-SNE) for each learning scenario proposed.

nodes and edges, the matrix adds a general improvement in the performance of all learning scenarios evaluated, but the greater performance reported was in the Unsupervised scenario, reinforcing the feasibility of application in others MIR tasks.

We evaluated three approaches to constructing the importance matrix. The results showed that normalizing the input matrix was enough to achieve the best results in five experimental assessed settings, compared to two experimental settings where the learned matrix had superior results and two experimental settings for the random matrix. The random matrix's superiority in some experiments is a point that demands attention because this approach to composing the matrix reduces the relevance of building real relationships between artists. On the other hand, performing just normalization is the simplest process and has proven efficient in most scenarios. Evaluating different strategies for learning the importance matrix is an important task to be explored in future work. Finally, the GraphSAGE results in Table III indicate a competitive baseline method, while PEAGNN is not one. However, we can note that inducing the type of features and relationships that will be emphasized in the learning process tends to increase the performance of the models.

As an interpretability resource, we present in Figure 5 a projection of the embeddings extracted from the last layer of the GNN model. After model learning, we compute a new low-dimensional space using the t-SNE method [40] for each node type. We display the projection of the embeddings learned in each GNN model for all learning scenarios using REP I as an input feature and the parameters that resulted in the best AUPR metric for this artist representation type. We highlight the ability of GNN models to group artist nodes into the nearest region by incorporating and refining the cluster-based information contained in the input feature.

In addition to the best results obtained, we report a statistical analysis to discuss the general performance of the different artist representations, importance matrices, learning scenarios, and types of embeddings to measure the generalization capacity of GNN models regardless of parameter settings. We use the Multi-Comparison Matrix (MCM) proposed in [41] to perform pairwise comparisons between all the attributes presented in this work. The MCM maps attributes into rows (r) and columns (c) and computes the number of evaluations in which an attribute performed better, equal, or lower than another attribute. The statistical significance of the differences in performance between each pair is computed using The Wilcoxon test, where the resulting p-value represents the probability of observing the given differences in performance.

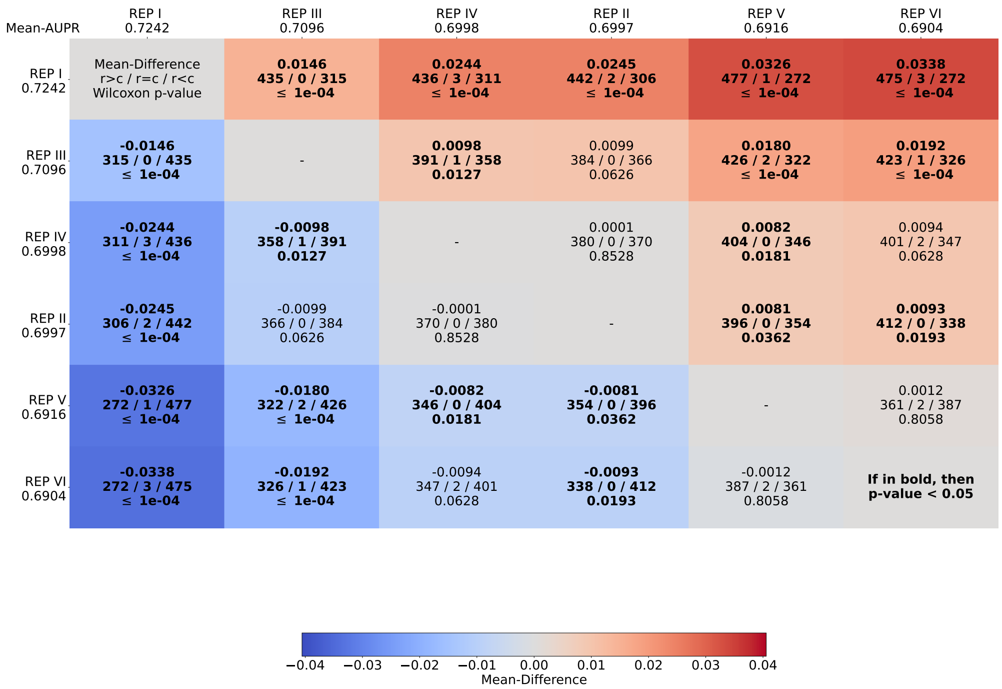
In this report, our objective is not to conclude about the superiority of an attribute based on the statistical difference because a consistent statistical analysis demands evaluations on more datasets. We are using MCM to report the performance of attributes considering the data folds and parameter variations.

Figure 6 shows the MCM grouping our results by the artist representations. In this case, we realized 750 records for each representation, where this number is defined by 5 data folds × 3 learning scenarios × 5 learning rates × 5 importance matrix × 2 embeddings types. We conclude that REP I is the representation that best characterizes artists because this representation is consistently superior in most experiments. In addition, considering the AUPR mean, we can order the most relevant indicators for building the GNN-based artist representation: codebook-based representations, audio modality characterized by transfer learning approaches, and multimodal representations.

Figure 7 shows the MCM for the embedding types. Here, we group the results by embedding types, resulting in 2250 records. We have for each embedding type: 6 REPs × 5 data folds × 3 learning scenarios × 5 learning rates × 5 importance matrix ×. In this scenario, we can notice that both embedding types presented a similar mean performance. This observation is important to reinforce random walk-based representation as a competitive comparison approach to GNN-based embeddings.

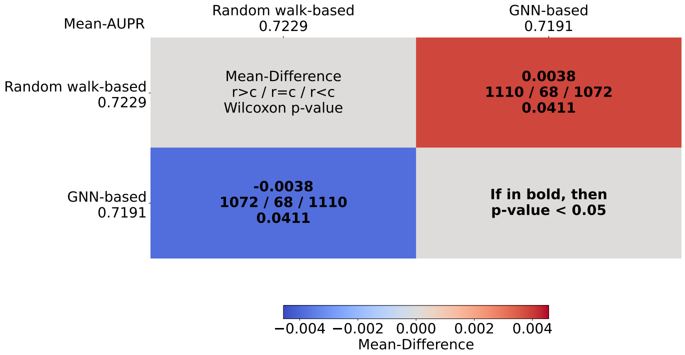
Figure 8 shows the MCM grouping our results by the importance matrix. In this case, we realized 900 records, where for each importance matrix, we have 6 REPs × 5 data folds × 3 learning scenarios × 5 learning rates × × 2 embedding types. In this scenario, we observed that using the importance matrix is more efficient than not using it or using a random matrix with information that is not directly related to the task of link prediction between artist nodes. This observation directs future work to investigate new methods for constructing the matrix.

Figure 9 shows the MCM grouping our results by the learning scenario. In this case, we realized 1500 records, where for each learning scenario, we have: 6 REPs × 5 data folds × 5 importance matrix × 5 learning rates × ×



Mean-AUPR	REP I 0.7242	REP III 0.7096	REP IV 0.6998	REP II 0.6997	REP V 0.6916	REP VI 0.6904
REP I 0.7242	Mean-Difference r>c / r=c / r<c Wilcoxon p-value	**0.0146** **435 / 0 / 315** **≤ 1e-04**	**0.0244** **436 / 3 / 311** **≤ 1e-04**	**0.0245** **442 / 2 / 306** **≤ 1e-04**	**0.0326** **477 / 1 / 272** **≤ 1e-04**	**0.0338** **475 / 3 / 272** **≤ 1e-04**
REP III 0.7096	**-0.0146** **315 / 0 / 435** **≤ 1e-04**	-	**0.0098** **391 / 1 / 358** **0.0127**	0.0099 384 / 0 / 366 0.0626	**0.0180** **426 / 2 / 322** **≤ 1e-04**	**0.0192** **423 / 1 / 326** **≤ 1e-04**
REP IV 0.6998	**-0.0244** **311 / 3 / 436** **≤ 1e-04**	**-0.0098** **358 / 1 / 391** **0.0127**	-	0.0001 380 / 0 / 370 0.8528	**0.0082** **404 / 0 / 346** **0.0181**	0.0094 401 / 2 / 347 0.0628
REP II 0.6997	**-0.0245** **306 / 2 / 442** **≤ 1e-04**	-0.0099 366 / 0 / 384 0.0626	-0.0001 370 / 0 / 380 0.8528	-	**0.0081** **396 / 0 / 354** **0.0362**	**0.0093** **412 / 0 / 338** **0.0193**
REP V 0.6916	**-0.0326** **272 / 1 / 477** **≤ 1e-04**	**-0.0180** **322 / 2 / 426** **≤ 1e-04**	**-0.0082** **346 / 0 / 404** **0.0181**	**-0.0081** **354 / 0 / 396** **0.0362**	-	0.0012 361 / 2 / 387 0.8058
REP VI 0.6904	**-0.0338** **272 / 3 / 475** **≤ 1e-04**	**-0.0192** **326 / 1 / 423** **≤ 1e-04**	-0.0094 347 / 2 / 401 0.0628	**-0.0093** **338 / 0 / 412** **0.0193**	-0.0012 387 / 2 / 361 0.8058	**If in bold, then** **p-value < 0.05**

Fig. 6. Multi-Comparison Matrix for all proposed artist representation. The representations are sorted by AUPR mean computed on the variations in all possible values to learning scenarios, learning rates, importance matrices, and embedding type.



Mean-AUPR	Random walk-based 0.7229	GNN-based 0.7191
Random walk-based 0.7229	Mean-Difference r>c / r=c / r<c Wilcoxon p-value	**0.0038** **1110 / 68 / 1072** **0.0411**
GNN-based 0.7191	**-0.0038** **1072 / 68 / 1110** **0.0411**	**If in bold, then** **p-value < 0.05**

Fig. 7. Multi-Comparison Matrix for random walk-based and GNN-based embedding types. The embedding types are sorted by AUPR mean computed on the variations in all possible values to artist representations, learning scenarios, learning rates, and importance matrices.

2 embeddings types. Here, we observed that although the Unsupervised scenario obtained the highest AUPR value reported in Table VI, its mean performance is lower than the others, which may be related to the lower ability to group artist representations as shown in Figure 5.

Finally, we report in Figure 10 a performance comparison between embedding types according to each proposed scenario. The motivation for reporting this comparison is to discuss the parity between proposed GNN-based and random walk-based embeddings. We notice similar average performance between both embeddings in all scenarios and a slight advantage in the highest results for the proposed GNN-based embedding.

Using a random walk-based feature to represent the graph nodes is a resource with less pre-processing cost than processing musical features. However, it is a representation that does not have semantic information about the data, making it challenging to construct information that allows interpretability in decision-making. Another point is that the random walk-based representation is not a deterministic representation, being dependent on the graph topology, which can impact the performance stability of the models.

In our work, the semantic information is of great importance, as the main objective is to build a representation for artists that allows us to measure each feature's impact on the models' performance when dealing with similarities between artists. By concluding that the best performance, on average, is achieved by a representation composed of audio and lyrics, we can direct our future work to explore other multimodal representations composed of features obtained by fine-tuned models in a transfer learning process.

VI. Summary and Final remarks

In this work, we handle the artist similarity through the link prediction task. For this, we propose modeling musical data on a heterogeneous network formed by layers with nodes represented with audio, lyrics, and artist features. From a network regularization process, we propagate features between the nodes and build an initial feature for all nodes used in the GNN models to predict the links between artist nodes. In addition, we propose an importance matrix that induce an importance value to instances to map the information obtained from artist relations into a matrix aggregating another information level to nodes and edges.

In our experimental analysis, we evaluated multiple combinations of parameters and features proposed to handle link prediction tasks using the AUPR metric (area under

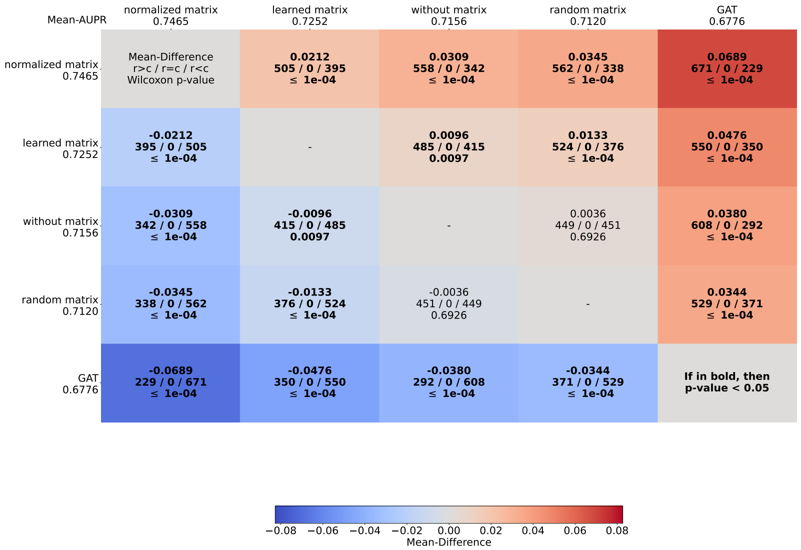


Fig. 8. Multi-Comparison Matrix for all possible compositions for the proposed importance matrix. The importance matrices are sorted by AUPR mean computed on the variations in all possible values to artist representations, learning scenarios, learning rates, and embedding types.

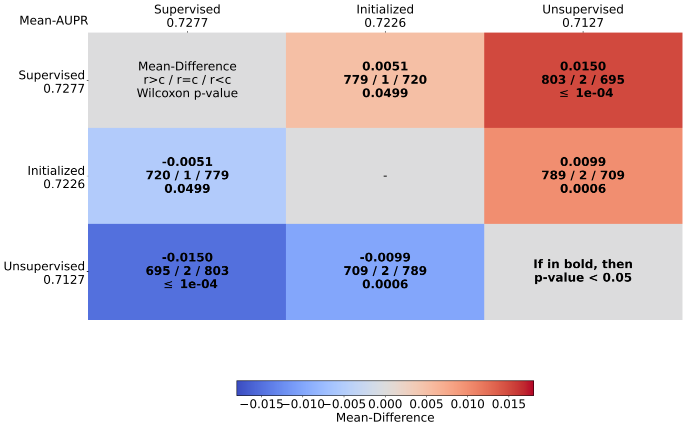


Fig. 9. Multi-Comparison Matrix for Supervised, Initialized, and Unsupervised learning scenarios. The learning scenarios are sorted by AUPR mean computed on the variations in all possible values to artist representations, learning rates, importance matrices, and embedding types.

the precision and recall curve). We present the best AUPR results obtained for three different GNN models and a pairwise comparison between all the elements that make up our proposal: input features, learning models, importance matrices, and embedding types. In conclusion, we observed that a multimodal representation formed by audio and lyrics embeddings achieved the best performance. We also note that using the proposed importance matrix is better than using no matrix, random matrix, or automatically learning importance with a GAT. We observed that the unsupervised GNN model achieved a higher AUPR, but the average performance of the supervised GNN model was higher.

Our work contributes to the representation learning applications for musical data. We proposed a new approach to build artist representation based on heterogeneous networks and the GNN model and to compute the similarity between artists using acoustic and lyric features extracted from your songs. The musical features indicated that the artists are best represented in a multimodal scenario with features obtained from pre-trained models. This indicates how we can explore other models and fine-tuning processes to enrich the audio and lyrics modalities. The importance matrix shows that aggregated information related to the problem context is a factor that increases the discriminative power of GNN models. In future work, we will reproduce the modeling process and explore other musical features to build the artist's representation. In addition, we will compute the performance of the importance matrix in more MIR tasks and measure the relevance in other contexts.

Acknowledgments

This work was supported by *Coordenação de Aperfeiçoamento de Pessoal de Nıvel Superior*(CAPES) grant numbers [88887.893806/2023-00] and São Paulo Research Foundation (FAPESP grant 2022/14903-1). The authors of this work would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the FAPESP grant 2019/07665-4 and from the IBM Corporation. Finally, this work was supported by *Ministério da Ciência, Tecnologia e Inovacões*, with funds from Law 8248, of October 23, 1991, PPI-SOFTEX, coordinated by Softex and published Residence in TIC 13, DOU 01245.010222/2022-44.

References

[1] Y. Deldjoo, M. Schedl, P. Cremonesi, and G. Pasi, "Recommender systems leveraging multimedia content," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–38, 2020.

[2] D. Paul and S. Kundu, "A survey of music recommendation systems with a proposed music recommendation system," in *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018.* Springer, 2020, pp. 279–285.

[3] D. P. Ellis, B. Whitman, A. Berenzweig, and S. Lawrence, "The quest for ground truth in musical artist similarity," in *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*, 2002.

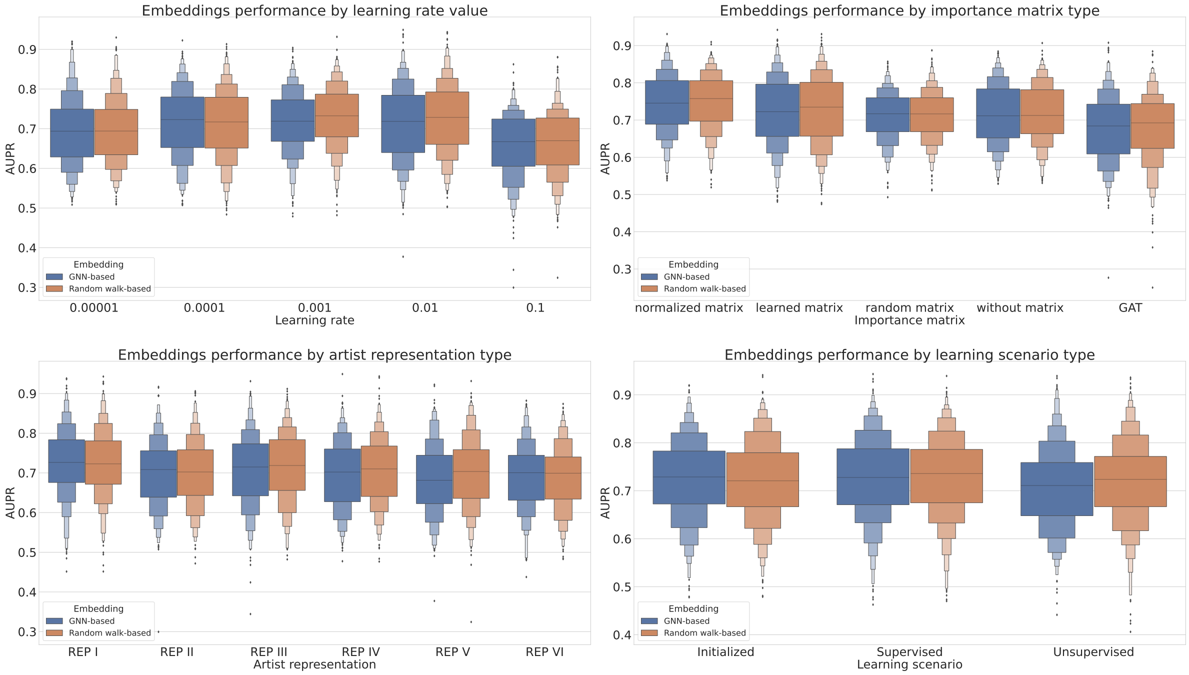


Fig. 10. Overview of the AUPR results for each embedding type performance considering the main parameters that compose our work.

[4] S. Oramas, M. Sordo, L. Espinosa-Anke, and X. Serra, “A semantic-based approach for artist similarity,” in *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*. International Society for Music Information Retrieval (ISMIR), 2015.

[5] Y. Pandeya and J. Lee, “Deep learning-based late fusion of multimodal information for emotion classification of music video,” *Multimed Tools Appl*, 2020.

[6] C. Chen and Q. Li, “A multimodal music emotion classification method based on multifeature combined network classifier,” *Mathematical Problems in Engineering*, 2020.

[7] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, “Graph neural networks: A review of methods and applications,” *AI open*, vol. 1, pp. 57–81, 2020.

[8] A. Gupta, P. Matta, and B. Pant, “Graph neural network: Current state of art, challenges and applications,” *Materials Today: Proceedings*, vol. 46, pp. 10 927–10 932, 2021.

[9] C. Yang, Y. Xiao, Y. Zhang, Y. Sun, and J. Han, “Heterogeneous network representation learning: A unified framework with survey and benchmark,” *IEEE Transactions on Knowledge and Data Engineering*, vol. PP, pp. 1–1, 12 2020.

[10] C. Shi, *Heterogeneous Graph Neural Networks*. Singapore: Springer Singapore, 2022, ch. 16, pp. 351–370.

[11] F. Korzeniowski, S. Oramas, and F. Gouyon, “Artist similarity with graph neural networks.” in *International Conference on Music Information Retrieval - ISMIR*, 2021, pp. 350–357.

[12] A. C. M. da Silva, D. F. Silva, and R. M. Marcacini, “4mula: A multitask, multimodal, and multilingual dataset of music lyrics and audio features,” in *Proceedings of the Brazilian Symposium on Multimedia and the Web*, ser. WebMedia ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 145–148.

[13] P. Knees, M. Schedl, and M. Goto, “Intelligent user interfaces for music discovery: The past 20 years and what’s to come.” in *ISMIR*, 2019, pp. 44–53.

[14] J. Kim, J. Urbano, C. C. Liem, and A. Hanjalic, “One deep music representation to rule them all? a comparative analysis of different representation learning strategies,” *Neural Computing and Applications*, vol. 32, pp. 1067–1093, 2020.

[15] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, “Learning music audio representations via weak language supervision,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 456–460.

[16] H.-H. Wu, C.-C. Kao, Q. Tang, M. Sun, B. McFee, J. P. Bello, and C. Wang, “Multi-task self-supervised pre-training for music classification,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 556–560.

[17] P. Knees and M. Schedl, “A survey of music similarity and recommendation from music context data,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 10, no. 1, pp. 1–21, 2013.

[18] D. C. Corrêa and F. A. Rodrigues, “A survey on symbolic data-based music genre classification,” *Expert Systems with Applications*, vol. 60, pp. 190–210, 2016.

[19] R. Panda, R. M. Malheiro, and R. P. Paiva, “Audio features for music emotion recognition: a survey,” *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[20] J. Pons and X. Serra, “musicnn: Pre-trained convolutional neural networks for music audio tagging,” 2019.

[21] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, “Jukebox: A generative model for music,” 2020.

[22] W. Fan, Y. Ma, Q. Li, J. Wang, G. Cai, J. Tang, and D. Yin, “A graph neural network framework for social recommendations,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 5, pp. 2033–2047, 2022.

[23] C. Gao, Y. Zheng, N. Li, Y. Li, Y. Qin, J. Piao, Y. Quan, J. Chang, D. Jin, X. He, and Y. Li, “A survey of graph neural networks for recommender systems: Challenges, methods, and directions,” *ACM Trans. Recomm. Syst.*, vol. 1, no. 1, mar 2023.

[24] A. C. M. da Silva, D. F. Silva, and R. M. Marcacini, “Heterogeneous Graph Neural Network for Music Emotion Recognition,” in *Proceedings of the 23rd International Society for Music Information Retrieval Conference*. Bengaluru, India: ISMIR, Dec. 2022, pp. 667–674.

[25] K. Watanabe and M. Goto, “Query-by-blending: A music exploration system blending latent vector representations of lyric word, song audio, and artist.” in *ISMIR*, 2019, pp. 144–151.

[26] G. Salha-Galvan, R. Hennequin, B. Chapus, V.-A. Tran, and M. Vazirgiannis, “Cold start similar artists ranking with gravity-inspired graph autoencoders,” in *Proceedings of the 15th ACM Conference on Recommender Systems*, 2021, pp. 443–452.

[27] A. Dhruv, A. Kamath, A. Powar, and K. Gaikwad, “Artist recommendation system using hybrid method: A novel approach,” in *Emerging Research in Computing, Information, Communication and Applications: ERCICA 2018, Volume 1*. Springer, 2019, pp. 527–542.

[28] D. Bogdanov, N. Wack, E. Gómez Gutiérrez, S. Gulati, H. Boyer, O. Mayor, G. Roma Trepat, J. Salamon, J. R. Zapata González, X. Serra *et al.*, “Essentia: An audio analysis library for music information

retrieval," in *Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4-8; Curitiba, Brazil.[place unknown]: ISMIR; 2013. p. 493-8.* International Society for Music Information Retrieval (ISMIR), 2013.

[29] R. Polak, N. Jacoby, T. Fischinger, D. Goldberg, A. Holzapfel, and J. London, "Rhythmic prototypes across cultures: A comparative study of tapping synchronization," *Music Perception: An Interdisciplinary Journal*, vol. 36, no. 1, pp. 1–23, 2018.

[30] M. L. Lavengood, "The cultural significance of timbre analysis: A case study in 1980s pop music, texture, and narrative," *Music Theory Online*, vol. 26, no. 3, 2020.

[31] M. Panteli, E. Benetos, and S. Dixon, "A computational study on outliers in world music," *Plos one*, vol. 12, no. 12, p. e0189399, 2017.

[32] D. Silva, R. Rossi, S. Rezende, and G. Batista, "Music classification by transductive learning using bipartite heterogeneous networks," in *15th International Society for Music Information Retrieval Conference*, 10 2014, pp. 113–118.

[33] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in neural information processing systems*, 2004, pp. 321–328.

[34] R. C. Brian McFee and, L. Dawen, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python." in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.

[35] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[36] X. Han, T. Zhao, Y. Liu, X. Hu, and N. Shah, "MLPInit: Embarrassingly simple GNN training acceleration with MLP initialization," in *International Conference on Learning Representations*, 2023.

[37] Y. Yang, R. N. Lichtenwalter, and N. V. Chawla, "Evaluating link prediction methods," *Knowledge and Information Systems*, vol. 45, pp. 751–782, 2015.

[38] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 1025–1035.

[39] M. U. Anwaar, Z. Han, S. Arumugaswamy, R. A. Khan, T. Weber, T. Qiu, H. Shen, Y. Liu, and M. Kleinsteuber, "On leveraging the metapath and entity aware subgraphs for recommendation," in *Proceedings of the 1st Workshop on Multimedia Computing towards Fashion Recommendation*, ser. MCFR '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 3–10. [Online]. Available: https://doi.org/10.1145/3552468.3555361

[40] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[41] A. Ismail-Fawaz, A. Dempster, C. W. Tan, M. Herrmann, L. Miller, D. F. Schmidt, S. Berretti, J. Weber, M. Devanne, G. Forestier, and G. I. Webb, "An approach to multiple comparison benchmark evaluations that is stable under manipulation of the comparate set," *arXiv preprint arXiv:2305.11921*, 2023.