TISMIR

Nuanced Music Emotion Recognition via a Semi-Supervised Multi-Relational Graph Neural Network

RESEARCH ARTICLE

ubiquity press

ANDREAS PEINTNER
MARTA MOSCATI
YU KINOSHITA
RICHARD VOGL
PETER KNEES
MARKUS SCHEDL
HANNAH STRAUSS
MARCEL ZENTNER
EVA ZANGERLE

*Author affiliations can be found in the back matter of this article

ABSTRACT

Music emotion recognition (MER) seeks to understand the complex emotional landscapes elicited by music, acknowledging music's profound social and psychological roles beyond traditional tasks such as genre classification or content similarity. MER relies heavily on high-quality emotional annotations, which serve as the foundation for training models to recognize emotions. However, collecting these annotations is both complex and costly, leading to limited availability of large-scale datasets for MER. Recent efforts in MER for automatically extracting emotion have focused on learning track representations in a supervised manner. However, these approaches mainly use simplified emotion models due to limited datasets or a lack of necessity for sophisticated emotion models and ignore hidden inter-track relations, which are beneficial in a semi-supervised learning setting. This paper proposes a novel approach to MER by constructing a multi-relational graph that encapsulates different facets of music. We leverage graph neural networks to model intricate inter-track relationships and capture structurally induced representations from user data, such as listening histories, genres, and tags. Our model, the semi-supervised multi-relational graph neural network for emotion recognition (SRGNN-Emo), innovates by combining graph-based modeling with semi-supervised learning, using rich user data to extract nuanced emotional profiles from music tracks. Through extensive experimentation, SRGNN-Emo demonstrates significant improvements in R^2 and root mean squared error metrics for predicting the intensity of nine continuous emotions (Geneva Emotional Music Scale), demonstrating its superior capability in capturing and predicting complex emotional expressions in music.

CORRESPONDING AUTHOR:
Andreas Peintner
University of Innsbruck, Innsbruck, Austria
andreas.peintner@uibk.ac.at

KEYWORDS:
music emotion recognition, graph neural networks, semi-supervised learning, contrastive loss, evoked emotions, node representation, multi-relational, genres, tags

TO CITE THIS ARTICLE:
Peintner, A., Moscati, M., Kinoshita, Y., Vogl, R., Knees, P., Schedl, M., Strauss, H., Zentner, M., & Zangerle, E. (2025). Nuanced Music Emotion Recognition via a Semi-Supervised Multi-Relational Graph Neural Network. *Transactions of the International Society for Music Information Retrieval, 8*(1), 140–153.
DOI: https://doi.org/10.5334/tismir.235

1 INTRODUCTION

Music's ability to express and evoke emotions is a universally acknowledged phenomenon, transcending cultural and linguistic barriers. It plays a pivotal role in human experience, offering a medium through which emotions can be articulated, shared, and understood. This unique capacity of music to convey a wide range of emotional states makes it a subject of considerable interest in the interdisciplinary fields of psychology, neuroscience, and musicology (Jia et al., 2021; Zentner et al., 2008). In particular, music emotion recognition (MER) is a computational task aimed at automatically identifying the emotional expressions contained within music or the emotions elicited in listeners by music (Yang and Chen, 2011). MER researchers rely on a collection of datasets, where the amount of annotated tracks per dataset is rather small (Aljanaki et al., 2017; Zhang et al., 2018). This is unsurprising since collecting high-quality emotional annotations of tracks is complex and expensive (Strauss et al., 2024). While small-scale datasets are valuable for MER advancements (Laurier et al., 2009), for music retrieval and recommendation tasks, it is inevitable to have access to a large catalog of tracks annotated with emotion labels, especially in the context of personalized music retrieval (Yang, 2021). An alternative method for gathering emotional data in music involves extracting emotions from user tags. These tags are readily accessible and available on a large scale. However, they often contain noise and personal bias, and they also lack the depth and quality that set apart expert-annotated data. Such expert data are typically collected through user studies informed by psychological principles (Laurier et al., 2009; Moscati et al., 2024).

There are several approaches to MER, aiming to tag tracks with corresponding emotion labels or profiles. Textual information is one of the data types employed in assignments that incorporate emotion labels, as evidenced by numerous studies (Hu et al., 2009; Hu and Downie, 2010; Zad et al., 2021). Specifically, when undertaking emotion recognition based on music data, lyrics frequently serve as the primary source of input (Choi et al., 2018; da Silva et al., 2022). A different body of research highlights the significant role of acoustic features in emotion recognition tasks (Gómez-Cañón et al., 2021; Panda et al., 2020; Yang, 2021; Yang et al., 2008). This perspective sheds light on the complexity of musical emotion, suggesting that the emotional content of music cannot be fully captured through lyrics alone. The recognition that both modalities, textual and acoustic, play a critical role in the perception and interpretation of musical emotions is well known in the scientific community (Gómez-Cañón et al., 2021; Rajan et al., 2021; Xue et al., 2015).

Most of the aforementioned approaches perform classification for emotion labels per track or employ basic or categorical emotion models (e.g., arousal and valence) in a supervised learning setting, often failing to capture the richness and variability of musical emotions (da Silva et al., 2022; Yang, 2021). In contrast, this work draws on a domain-specific model devised to account for the richness of emotions induced by music (Zentner et al., 2008). Starting with 515 emotion terms, Zentner et al. (2008) have successively eliminated those terms that were rarely used to describe music-evoked emotions and retained a few dozen core emotion terms, titled GEMS for Geneva Emotional Music Scale. GEMS is hierarchically organized into three second-order and nine first-order factors, as shown in Figure 1. These factors are (1) vitality (power and joyful activation); (2) sublimity (wonder, transcendence, tenderness, nostalgia, and peacefulness); and (3) unease (tension and sadness). An additional distinctive feature of GEMS is that it accounts not only for perceived emotion but also, and in particular, for induced emotions, as was later shown by neuroimaging work (Trost et al., 2012). Consequently, a MER approach based on this model can capitalize on a rich spectrum of music-specific emotional information (Aljanaki et al., 2014).

As mentioned earlier, the number of tracks in MER datasets is limited due to the scalability challenges associated with the annotation process. This limitation impacts the ability of supervised learning approaches to generalize effectively across a vast track catalog, as the availability of annotated data directly influences model performance. Semi-supervised learning, on the other hand, allows us to effectively incorporate information of unlabeled tracks as well as labeled ones in the learning process, leading to enriched track embeddings for the final labeling task. Moreover, prior works often ignore user and track meta-data, which could be used to improve the learning process.

In this paper, we propose a novel framework employing the semi-supervised multi-relational graph neural network for emotion recognition (SRGNN-Emo) for predicting the emotion profiles of tracks. We define the emotion profile of a music track as the set and intensity of emotions that the track evokes in listeners (Kim et al., 2010; Strauss et al., 2024). Unlike traditional MER approaches, our model advances the field by adopting a multi-target regression strategy, aiming to capture more accurately the broad spectrum of emotions sparked through music. Building upon the premise that human listening behaviors encapsulate a wealth of information about evoked emotions, we innovate by integrating semi-supervised learning with human annotations and a multi-relational graph framework. This integration allows us to exploit the rich, albeit underutilized, data from user interactions, genres, and tags, hypothesizing that such data, when structured into diverse graph formats and refined by a semi-supervised learning framework, induce valuable emotion-related information. Our framework can predict emotional intensities across

nine dimensions, significantly enhancing the emotional insights derived from track embeddings compared to traditional methods that typically rely on fewer, music-non-specific emotion dimensions such as valence and arousal.

Our approach not only aims to mitigate the limitations imposed by the scarcity of large, annotated datasets but also introduces a novel perspective on using multi-relational graph structures to enrich track representations. To summarize, the main technical contributions of our work are as follows:

- We propose a novel multi-relational graph structure, based on user interactions, genres, and tags.
- We integrate a semi-supervised learning approach for multi-target regression into the framework of graph neural networks (GNNs).
- We use a high-quality dataset based on state-of-the-art psychological research into music-evoked emotions for fine-grained MER (Strauss et al., 2024).
- Extensive experiments show that our proposed model significantly outperforms state-of-the-art competitors on the task of MER.

To ensure reproducibility, we will release the code of our experiments and model weights on GitHub.[1]

2 RELATED WORK AND BACKGROUND

2.1 MUSIC EMOTION RECOGNITION

MER aims to understand and categorize emotions in music through computational means. Key contributions to this field address the different facets of music and emotion, proposing various methodologies for recognition and analysis (Yang and Chen, 2011). Kim et al. (2010) have presented a comprehensive overview of MER, introducing a computational framework that generalizes emotion recognition from categorical domains to a two-dimensional space defined by valence and arousal, facilitating novel emotion-based music retrieval and organization methods. Other works (Choi et al., 2018; Panda et al., 2020; Xue et al., 2015; Zad et al., 2021) have also emphasized the role of integrating lyrics, chord sequences, and genre metadata alongside audio features, demonstrating how multifaceted approaches can significantly enhance MER systems' accuracy.

The development of MER has also been propelled by the creation of extensive datasets and embeddings tailored for this purpose—for instance, the MuSe dataset (Akiki and Burghardt, 2021), which includes 90,000 tracks annotated with arousal, valence, and dominance values inferred from tags. Moreover, works by Alonso-Jiménez et al. (2023), Bogdanov et al. (2022), and Castellon et al. (2021) have evaluated various audio embeddings, including *Jukebox* and *musicnn* embeddings, for their effectiveness in MER tasks. Additionally, recent evaluations of state-of-the-art music audio embeddings have been conducted using tasks such as the MediaEval challenge series on Emotion and Theme Recognition in Music (Tovstogan et al., 2021) applied to the MTG-Jamendo mood/theme auto-tagging dataset (Bogdanov et al., 2019).

Advances in MER research have also been characterized by the development of novel features and the design of sophisticated machine learning models. Bhatti et al. (2016) have shown the effectiveness of using physiological signals, specifically via electroencephalography, to recognize emotions elicited by different music genres, highlighting the potential of brain signals in providing insights into emotional responses to music. Panda et al. (2018) have improved music emotion classification by introducing highly emotionally relevant audio features related to music performance expressive techniques or musical texture. The application of deep learning techniques has shown promising results in recognizing emotions from music, as seen in the work by Zhang et al. (2023); these authors have extracted features from log-Mel spectrograms by using multiple parallel convolutional blocks and applied attention in combination with a sequence learning model for dynamic music emotion prediction. Others have proposed structuring musical features from different modalities (audio and lyrics) over a heterogeneous network to incorporate different modalities in a unique space for MER (da Silva et al., 2022).

Our proposed approach, SRGNN-Emo, innovates by leveraging semi-supervised learning with user interaction data and metadata for nuanced emotional profiles, extending beyond traditional supervised methods.

2.2 SEMI-SUPERVISED NODE REPRESENTATION LEARNING

Node representation learning is focused on creating simplified vector representations of a graph's nodes that

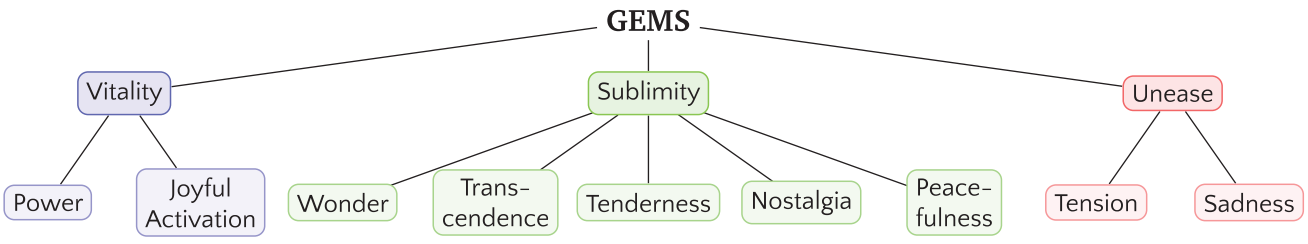

Figure 1 The Geneva Emotion Music Scale with nine dimensions based on the factor analysis in the work of Zentner et al. (2008).

reflect both their connections and features. Traditional methods (without deep learning) are mostly based on random walks to examine the neighborhoods around nodes (Grover and Leskovec, 2016; Perozzi et al., 2014; Tang et al., 2015).

GNNs are neural architectures specifically tailored for graph-structured data. GNNs learn meaningful node representations by iteratively aggregating and transforming information from a node's neighbors, effectively capturing complex relational and structural dependencies in graphs (Hamilton et al., 2017; Kipf and Welling, 2017). Since the introduction of graph convolutional networks (GCNs) (Kipf and Welling, 2017; Velickovic et al., 2018), a specific type of GNNs, more advanced techniques for node embedding have been developed, including a layer sampling algorithm (Hamilton et al., 2017) designed to work with large graphs by focusing on a set neighborhood of nodes.

Recently, we have observed a shift toward self-supervised contrastive approaches. These methods distinguish between positive (similar neighborhood) and negative (far away in the graph) examples to compute loss. A deep graph infomax (DGI) (Velickovic et al., 2019) enhances the mutual information between individual nodes and the whole graph representations. Hassani and Ahmadi (2020) have introduced a method for learning representations from different viewpoints by contrasting nearby neighbor encodings with those from a more extensive graph diffusion. However, because contrastive learning often requires a significant number of negative examples, it can be challenging to scale for large graphs. An alternative proposed by Thakoor et al. (2021) named bootstrapped graph latents (BGRLs) avoids this issue by predicting alternative augmentations of the input, eliminating the need for contrasting with negative samples.

Despite significant advances in node representation learning, relatively little attention has been given to multi-relational GNNs and their application in specific domains like MER. Existing works such as those by Schlichtkrull et al. (2018), who proposed the concept of relational graph convolutional networks (R-GCNs) for knowledge graph completion, and Vashishth et al. (2020), who explored compositional embeddings for relationships, have made strides in handling complex relational structures. However, these approaches have not been widely explored within the context of semi-supervised learning. Additionally, although semi-supervised node representation learning has become increasingly popular in tasks such as node classification and link prediction (Hamilton et al., 2017; Kipf and Welling, 2017), its application to emotion recognition tasks remains rare and under-investigated (Horner et al., 2013).

In this paper, we present an innovative framework that not only aligns with recent trends toward contrastive learning in GNNs but also extends them by specifically addressing the multi-relational and semi-supervised nature of the problem space in MER.

3 DATASET

In this work, we leveraged high-quality data from psychology-informed user studies on emotions evoked by music. We used the Emotion-to-Music Mapping Atlas (EMMA)[2] database (Strauss et al., 2024), which contains 817 music tracks. These tracks were last annotated in 2023 based on their emotional impact, as assessed using GEMS (Zentner et al., 2008). We focused on the GEMS-9 variant of this scale, which is a checklist version of the original 45-item GEMS that assesses each dimension with one item only. Previous research has demonstrated emotion profiles derived from the original GEMS and GEMS-9 to be highly correlated (Jacobsen et al., 2024). Emotions induced by each track were rated on these dimensions by an average of 28.76 annotators. We are one of the first to leverage this information-rich dataset for MER purposes, demonstrating the significant potential it offers for advancing research in this field. To enhance the reliability of our analyses, we restricted our focus to tracks with a higher interrater agreement, selecting those with an intraclass correlation coefficient (ICC) above 0.5, which indicates moderate reliability (Strauss et al., 2024). While a higher ICC threshold would ensure even greater reliability, it would significantly reduce the dataset size, thereby limiting the diversity and generalizability of the data. However, it is worth mentioning that the ICC across all tracks demonstrates good interrater agreement, with a mean ICC value of 0.8.

As our goal was to design a model for large-scale emotion recognition in a semi-supervised manner, we required a dataset containing not only rich information about the audio but also relevant meta-data. Therefore, we employed the Music4All-Onion (Moscati et al., 2022) dataset. This dataset enhances the Music4All (Santana et al., 2020) dataset by incorporating 26 additional audio, video, and metadata characteristics for 109,269 music pieces. It also includes 252,984,396 listening records from 119,140 Last.fm[3] users, enabling the use of user–item interactions. Intersecting EMMA with the Music4All-Onion dataset led to 509 tracks with available emotion profiles, audio features, and meta information. Due to our hypothesis that human listening behavior in combination with track metadata encapsulates valuable information about evoked emotions, we extracted graph structures from user listening sessions, track genres, and user tags, as described in detail in Section 4.1.

For each track available in the Music4All-Onion dataset, we used pretrained instances of *musicnn* (Pons and Serra, 2019), *MAEST* (Alonso-Jiménez et al., 2023), and *Jukebox* (Dhariwal et al., 2020) to represent the audio signals. The *musicnn* model is based on deep convolutional neural networks trained to classify music based on its content (Pons and Serra, 2019). The *MAEST* representations are based on spectrogram-based audio

transformers, which employ patchout training on a supervised task (Alonso-Jiménez et al., 2023). *Jukebox* is a generative model for music that uses a deep neural network trained on a vast corpus of tracks to understand and generate music (Dhariwal et al., 2020).[4] These models were selected for their music-specific design, which ensures closer alignment with musical features such as melody, harmony, and rhythm that are critical for emotion recognition (Moscati et al., 2025).[5]

4 PROPOSED METHOD (SRGNN-EMO)

In this section, we introduce a novel framework leveraging a multi-relational graph structure and semi-supervised learning. Multi-relational graphs are complex data structures that model different types of relations that correspond to different user data types in our case (e.g., Figure 2). Our model is designed to extract emotional profiles from music tracks by integrating rich user interaction data with diverse metadata and sophisticated content data. Figure 2 provides an overview of our proposed approach, where each module will be explained in the following.

4.1 MULTI-RELATIONAL GRAPH CONSTRUCTION

We sought to derive representations of tracks that encapsulate nuanced similarities between music tracks, based on shared genres, commonality in listening sessions, and user-assigned tags. Therefore, we constructed a multi-relational graph G, focusing on tracks as nodes, with edges representing different types of relationships, such as sessions, genres, or tags, that connect these tracks. Specifically, nodes in our multi-relational graph are tracks $v \in V$, and an edge $(v_i, v_j) \in E$ is established between two tracks if they are part of the same listening session by a user, share one or multiple genres, or have been tagged with one or multiple identical tags by users. The strength of the connection, represented as the edge weight $e_{ij}^{(r)}$, reflects the frequency of shared relationships $r \in R$, such as the number of common tags, genres, or sessions. We normalized the edge weights per relation such that, for each track v and each relation r, the edge weights can be symmetrically scaled using the formula:

$$\tilde{e}_{ij}^{(r)} = \frac{e_{ij}^{(r)}}{\sqrt{\deg(v_i) \cdot \deg(v_j)}},$$

where $\deg(v)$ represents the degree of node v for relation r. This symmetric normalization ensures that, for each track v and each relation r, the edge weights are adjusted based on the degrees of both connected nodes and therefore mitigates the inherent popularity bias of tracks.

4.2 EMOTION-BASED GRAPH ENCODER

To learn node representations on the multi-relational graph G introduced before, we employed a weighted relational GCN (wR-GCN) encoder, which adapts the GNN message-passing framework to handle the complexities of a multi-relational graph (Schlichtkrull et al., 2018) and additionally incorporates edge weights. The GNN message-passing framework (Gilmer et al., 2017)

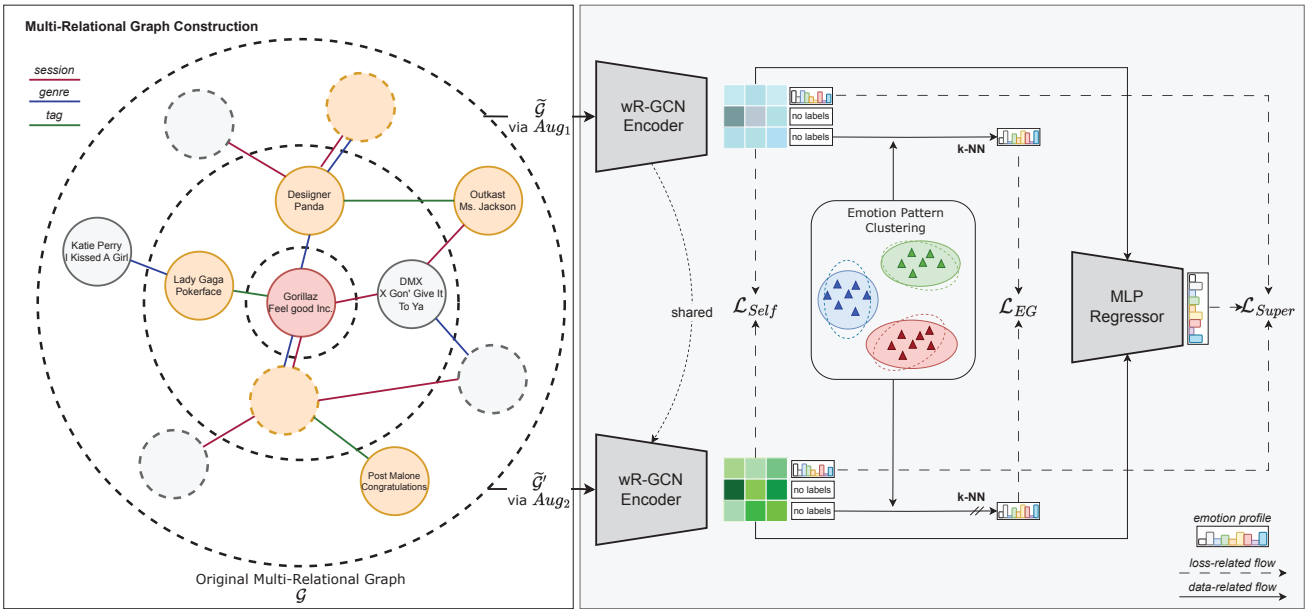

Figure 2 Illustration of SRGNN-Emo which constructs a multi-relational graph with nodes representing tracks and edges symbolizing connections based on sessions, genres, or user tags shared among tracks. We used stochastic graph augmentations to generate two distinct graph views, which were processed by a shared encoder to ensure robust and invariant node representations in a self-supervised manner. Emotion-guided consistency objective ($\mathcal{L}_{EG}$) optimization aimed to align unlabeled nodes with emotion profile patterns of labeled nodes across augmented graph views. The learned node representations were then fed into a multi-layer perceptron regressor to predict the emotion profile of each track.

enables nodes to exchange and integrate information with their neighbors, iteratively refining their representations to capture the graph's structural and relational context. The general differentiable message passing is formulated as:

$$h_i^{(l+1)} = \sigma\left(\sum_{m\in\mathcal{M}_i} g_m\left(h_i^{(l)}, h_j^{(l)}\right)\right), \tag{1}$$

where $h_i^{(l)} \in \mathbb{R}^{d^{(l)}}$ represents the hidden state of node v_i at the l-th layer, with $d^{(l)}$ being the dimensionality of the layer's representation. The incoming messages, $g_m(\cdot,\cdot)$, are combined and processed through an activation function $\sigma(\cdot)$, such as ReLU. $\mathcal{M}_i$ is the set of incoming messages for node v_i, typically corresponding to the set of incoming edges. The function $g_m(\cdot,\cdot)$ is often a neural network or a simple linear transformation (Kipf and Welling, 2017).

This transformation has proven effective in accumulating and encoding features from local, structured neighborhoods (Kipf and Welling, 2017; Velickovic et al., 2018). For our multi-relational, weighted graph, we defined a simple propagation model (Schlichtkrull et al., 2018) for computing the forward-pass update of a node v_i and extended it with the usage of edge weights:

$$h_i^{(l+1)} = \sigma\left(\sum_{r\in R}\sum_{j\in\mathcal{N}_i^r}\frac{1}{|\mathcal{N}_i^r|}W_r^{(l)}h_j^{(l)}e_{ij}^{(r)} + W_0^{(l)}h_i^{(l)}e_{ii}^{(r)}\right), \tag{2}$$

where $\mathcal{N}_i^r$ represents the set of neighbors of node v_i under relation $r \in R$, and $e_{ij}^{(r)}$ is the edge weight between nodes v_i and v_j for relation r. This equation intuitively accumulates the transformed feature vectors of neighboring nodes through a weighted and normalized sum. Unlike regular GCNs, we herein incorporated relation-specific transformations, depending on the type and direction of the edge. Additionally, to ensure that the node's representation at layer $l+1$ is informed by its representation at layer l, we introduced a self-connection under each relation type for each node.

Initially, $h_v^0 = x_v$, representing the node features. We used the corresponding representations of the tracks (e.g., *musicnn*, *MAEST*, or *Jukebox*) as the node features $X \in \mathbb{R}^{N\times F}$, where N is the number of nodes in the graph and F is the feature dimension. We defined $\mathcal{N}(v,r)$ as a uniformly sampled neighborhood across all relations $r \in R$ to manage memory and computation effectively (Hamilton et al., 2017).

4.3 SEMI-SUPERVISED MULTI-TARGET REGRESSION

Contrastive learning has been shown to be a valuable paradigm for self-supervised learning and consistency regulation in the context of GNNs (Lee et al., 2022; Thakoor et al., 2021). We employed this idea as the grounding learning task for our graph-based model and extended it with a semi-supervised loss in the process.

Given an input graph, we can generate two distinct graph views through stochastic graph augmentations. These augmentations involve randomly masking different node features and dropping a different subset of edges per graph to introduce variability. The resulting augmented graph views are denoted by $\widetilde{G} = (\widetilde{A}, \tilde{X})$ and $\widetilde{G}' = (\widetilde{A}', \tilde{X}')$, where $\widetilde{A}$ and $\widetilde{A}'$ represent the adjacency matrices of the augmented graphs and $\tilde{X}$ and $\tilde{X}'$ denote the feature matrices post-augmentation, respectively.

4.3.1 Representation learning via shared encoder

To learn robust, low-dimensional node-level representations, we employed a shared encoder strategy that learns consistent representations across different graph augmentations. Both augmented graph views were input into our shared wR-GCN encoder, denoted as $f_\theta : \mathbb{R}^{N\times N} \times \mathbb{R}^{N\times F} \to \mathbb{R}^{N\times D}$, to learn low-dimensional node-level representations. The node-level representations obtained from the encoder for the two views are $f_\theta(\widetilde{A}, \tilde{X}) = \tilde{Z} \in \mathbb{R}^{N\times D}$ and $f_\theta(\widetilde{A}', \tilde{X}') = \tilde{Z}' \in \mathbb{R}^{N\times D}$, respectively.

To ensure the learned node representations are invariant to the augmentations, SRGNN-Emo minimizes the cosine distance between the representations from the two differently augmented views on a node-wise basis and is formalized as follows:

$$\mathcal{L}_{\text{Self}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\tilde{Z}_i \cdot \tilde{Z}_i'}{\|\tilde{Z}_i\|\|\tilde{Z}_i'\|} \tag{3}$$

In their experiments, Lee et al. (2022) have found that using a single shared encoder in combination with subsequent supervisory signals was sufficient to prevent representation collapse, while also offering the benefits of simplicity and efficiency.

4.3.2 Emotion-guided consistency objective

While our framework effectively leverages self-supervised learning signals—patterns and features extracted from unlabeled data without explicit supervision—through contrastive learning, it had yet to incorporate the limited but accessible emotion profiles available for tracks. To leverage emotion label information effectively, we refined our method by aligning tracks with emotion profile patterns. Starting with a set of labeled tracks with known emotion profiles, we identified distinct emotion patterns through clustering, which then served as reference points (centroids) in the emotion profile space. Our goal was to group the unlabeled tracks around these centroids, ensuring their predicted emotion profiles remain consistent across differently augmented views of the graph. By doing so, we aimed to maximize the consistency and reliability of node assignments to these emotion patterns, effectively bridging the gap between labeled and unlabeled tracks.

Given the set of labeled tracks, denoted as V_L, we applied a k-means clustering algorithm to extract K

distinct clusters, each representing a unique emotion pattern. The result was a set of centroids $C = \{c_1, c_2, \ldots, c_k\}$, where each $c_k \in \mathbb{R}^{1\times 9}$ corresponds to the centroid of cluster k. These nine dimensions correspond to the emotional dimensions defined by GEMS, which serve as the basis for clustering. For each unlabeled track v_{ul}, we computed the predicted emotion profile using a non-parametric weighted k-nearest neighbors (k-NN) approach to generate pseudo-labels, formulated as:

$$p_i = \frac{\sum_{j\in \mathrm{NN}_k(H_i)} \mathrm{sim}(H_i, H_j^S)\cdot Y_j^S}{\sum_{j\in \mathrm{NN}_k(H_i)} \mathrm{sim}(H_i, H_j^S)}, \tag{4}$$

where $\mathrm{sim}(\cdot,\cdot)$ computes the cosine similarity between two vectors; $H^S \in \mathbb{R}^{N\times D}$ and $Y^S \in \mathbb{R}^{L\times 9}$ denote the support (labeled) node representations and the emotion profiles, respectively; and $\mathrm{NN}_k(H_i)$ denotes the set of $K_{\text{neighbors}}$ nearest neighbors of H_i in H^S.

To enhance reliability, we restricted the k-NN predictions to only confident pseudo-labels by measuring the distance between each pseudo-label and the centroid C. We retained nodes whose predicted profile showed a similarity above a threshold μ with at least one centroid, forming the set V_{conf}. The emotion-guided consistency objective could then be defined as:

$$\mathcal{L}_{EG} = \frac{1}{|V_{conf}|}\sum_{v_i\in V_{conf}} \mathrm{MSE}(\tilde{\mathbf{p}}_i, \tilde{\mathbf{p}}_i'), \tag{5}$$

where $\mathrm{MSE}(\cdot,\cdot)$ denotes the mean squared error (MSE) loss function and $\tilde{\mathbf{p}}_i$ and $\tilde{\mathbf{p}}_i'$ are the confidently predicted emotion profiles for track v_i from the augmented graphs. Using a high value for μ prioritizes confident pseudo-labels in the objective function, which has been shown to effectively mitigate confirmation bias (Arazo et al., 2020; Lee et al., 2022).

This approach not only incorporates label information to guide the learning of emotion profile patterns but also ensures that predictions for unlabeled tracks are made with greater confidence, thereby improving the overall model's ability to generalize from labeled to unlabeled data in the context of a multi-target regression task.

4.3.3 Emotion profile prediction

After learning robust node representations through the shared wR-GCN encoder and ensuring consistency across augmented graph views, the final task of SRGNN-Emo is to predict the emotion profile for each music track. To achieve this, we used a multi-layer perceptron (MLP) $\mathcal{R}(\cdot)$ that takes as input the averaged node representations from the two augmented views and outputs the emotion profile per node/track. The MLP consists of three fully connected layers, each followed by a LeakyReLU activation function and a dropout layer to prevent overfitting. The output of the MLP is a vector, $\hat{\mathbf{y}}_i \in \mathbb{R}^9$, representing the predicted emotion intensities across the nine emotion categories. To train the model to predict nine continuous emotion dimensions, we employed an MSE loss as our supervised objective:

$$\mathcal{L}_{\text{Super}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{9}\,\|\mathbf{y}_i - \hat{\mathbf{y}}_i\|^2 \tag{6}$$

4.4 FINAL OBJECTIVE

The combined objective function for SRGNN-Emo is expressed as:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{Self}} + \beta\mathcal{L}_{\text{EG}} + \mathcal{L}_{\text{Super}} \tag{7}$$

where α and β are coefficients that control the contribution of the self-supervised loss $\mathcal{L}_{\text{Self}}$ and the emotion-guided loss $\mathcal{L}_{\text{EG}}$ to the overall training objective, respectively. The supervised loss $\mathcal{L}_{\text{Super}}$ ensures the model effectively predicts continuous emotion profiles for labeled nodes.

5 EXPERIMENTS AND RESULTS

We compared SRGNN-Emo against traditional and graph-based baselines for MER. In the following, we detail our experimental setup, including data preparation, model configurations, and metrics used for evaluation.

5.1 BASELINES

We systematically compared our proposed model against a diverse array of baseline approaches, spanning traditional machine learning models, graph-based approaches, and a novel custom convolutional neural network, each harnessing unique feature representations from music analysis frameworks.

We began with traditional machine learning models, including logistic regression (LR) and support vector regression (SVR). Additionally, co-training regression (COREG) (Zhou and Li, 2005) was used, enhancing generalization by co-training two regressors on separate views. An MLP model with three layers (similar to our SRGNN-Emo model) served a dual purpose: it depicted a baseline on its own and acted as the regressor for semi-supervised learning tasks in the graph-based models (with the learned node representations as input).

Graph-based models are crucial for understanding the relational structure of music data. This category includes label propagation (LP) (Zhu and Ghahramani, 2002) for emphasizing data clustering, GCNs (Kipf and Welling, 2017) and graph attention networks (GATs) (Velickovic et al., 2018) for integrating node features with the graph topology, DGIs (Velickovic et al., 2019) focusing on mutual information maximization, and BGRL representation (Thakoor et al., 2021) aimed at enhancing robustness through consistent node representation across views. MRLGCN (da Silva et al., 2022) structures musical features

over a heterogeneous network and learns a multi-modal representation using a GNN with features extracted from audio and lyrics for MER.

Completing our set of baselines, we designed a fully supervised, content-based, end-to-end method named DOMR+ for density-based oversampling for multivariate regression with data transformation. The DOMR+ method consists of two components: a fully convolutional network model and a pre-processing stage. The model employs multiple convolutional and sub-sampling layers without dense layers. To address the challenges of data scarcity and imbalance in the labels, the pre-processing stage integrates oversampling with data transformation techniques. Candidate data points for oversampling were identified using kernel density estimation, which determines the rarity of data points based on their density within the feature space. Instead of directly oversampling these candidates, the method applies class-preserving audio transformations, which minimally transforms the original audio while retaining its fundamental properties, including filtering, equalizing, noise addition, scale changes (pitch shifting and time stretching), distortions, quantization, dynamic compression, format encoding/decoding (e.g., MP3, GSM), and reverberation (Mignot and Peeters, 2019). These transformations ensure that the augmented data remain representative of the underlying distribution, enhancing the model's ability to generalize, while avoiding the risk of overfitting caused by repetitive synthetic samples.

5.2 EXPERIMENTAL SETUP

We preprocessed the target variables representing emotions by applying z-normalization, which ensures each variable has a mean of 0 and a standard deviation of 1. We employed stratified 10-fold cross-validation based on binning to validate the performance of our models comprehensively.

For performance evaluation, we relied on two metrics: root mean squared error (RMSE) and coefficient of determination (R^2). RMSE measures the average magnitude of the errors between the predicted and actual values. A lower RMSE indicates better performance. R^2, on the other hand, is a goodness-of-fit measure for regression models and assesses the proportion of variance in the dependent variable that is predictable from the independent variables, with values closer to 1 indicating better model fit.

All baseline models were carefully tuned via grid search, optimizing hyperparameters including (but not limited to) the number of layers $\in \{1, \ldots, 5\}$, number of neighbors $\in \{5, 10, \ldots, 50\}$, learning rate, dropout, and regularization strength, depending on the respective model requirements. For our proposed model, SRGNN-Emo, the Adam optimizer (Kingma and Ba, 2015) was used, with the learning rate set to 0.001 and L_2 regularization set to 10^{-5}. We tuned its hyperparameters within specific ranges: the number of layers L in the wR-GCN was set between 1 and 5, the number of neighbors was chosen from between 5 and 50, and the α and β weight parameters were logarithmically adjusted within the range of 0.1–10. Additionally, dropout rates were varied between 0.0 and 0.5 to prevent overfitting. The number of clusters K and nearest-neighbors $K_{\text{neighbors}}$ were searched in $\{2, 4, 6, \ldots, 16\}$ and $\{5, 10, 20, 40\}$, correspondingly.

5.3 PERFORMANCE ANALYSIS

Table 1 summarizes the multi-target regression performance of various models, including traditional machine learning methods, graph-based models, and our proposed SRGNN-Emo framework. The results demonstrate that SRGNN-Emo achieves the lowest RMSE and highest R^2 score, indicating superior prediction performance (statistically significant) and model fit, respectively.

Representing traditional machine learning approaches, LR, SVR, and COREG show relatively higher RMSE values, indicating lower predictive performance. Their R^2 values are also significantly lower, confirming less variance explained by these models. The baseline MLP shows competitive performance when relying on *musicnn* representations, but it is outperformed by graph-based approaches with the other two representations (*MAEST* and *Jukebox*).

Among the graph-based approaches, DGI and BGRL show competitive performance with the lowest RMSE and highest R^2 among the graph-based models for two different representations, ranked second after our SRGNN-Emo. GCN and GAT also demonstrate robust performances but are slightly outperformed by DGI or BGRL, depending on the underlying representation. Meanwhile, our model, SRGNN-Emo, outperforms all baseline models and indicates a statistically significant improvement in terms of RMSE and R^2 compared to the second-best models, DGI and BGRL.

5.4 ABLATION STUDY

The ablation study, detailed in Table 2, assessed the impact of individual components of SRGNN-Emo by removing $\mathcal{L}_{\text{Self}}$, $\mathcal{L}_{\text{EG}}$, and $\mathcal{L}_{\text{Super}}$ separately. The results illustrate the essential roles of these components in the model's overall performance. Removing the self-supervised loss ($\mathcal{L}_{\text{Self}}$) slightly increased the RMSE across five out of nine emotional dimensions, suggesting that this component helps to stabilize the learning process by enforcing consistent node representations across different graph augmentations. The removal of the emotion-guided consistency objective ($\mathcal{L}_{\text{EG}}$) led to a noticeable degradation in performance across eight out of nine emotional dimensions. This confirms that $\mathcal{L}_{\text{EG}}$ plays a crucial role in refining node embeddings by aligning

Model \ Rep.	musicnn		MAEST		Jukebox	
	RMSE↓ (±SE)	R^2↑ (±SE)	RMSE↓ (±SE)	R^2↑ (±SE)	RMSE↓ (±SE)	R^2↑ (±SE)
LR	0.8443 (±0.02)	0.2470 (±0.05)	1.3821 (±0.06)	-1.0731 (±0.22)	1.0301 (±0.04)	-0.1403 (±0.09)
SVR	0.8188 (±0.01)	0.2968 (±0.01)	0.7862 (±0.01)	0.3504 (±0.02)	0.9802 (±0.02)	0.0163 (±0.01)
COREG	0.8742 (±0.02)	0.1140 (±0.05)	0.8613 (±0.02)	0.1346 (±0.08)	0.8680 (±0.02)	0.1244 (±0.05)
MLP	0.8132 (±0.02)	0.3106 (±0.02)	0.8938 (±0.03)	0.1576 (±0.08)	0.8579 (±0.02)	0.2193 (±0.06)
LP†	0.9488 (±0.03)	0.0806 (±0.01)	0.9488 (±0.03)	0.0806 (±0.01)	0.9488 (±0.03)	0.0806 (±0.01)
GCN	0.8071 (±0.02)	0.3158 (±0.04)	0.7781 (±0.02)	0.3568 (±0.05)	0.7492 (±0.04)	0.4039 (±0.05)
GAT	0.8167 (±0.03)	0.2992 (±0.07)	0.7856 (±0.02)	0.3476 (±0.05)	0.7567 (±0.02)	0.3926 (±0.03)
DGI	0.8042 (±0.02)	0.3184 (±0.06)	0.7749 (±0.01)	0.3644 (±0.06)	0.7464 (±0.02)	0.4103 (±0.04)
BGRL	0.8019 (±0.02)	0.3253 (±0.05)	0.7939 (±0.02)	0.3370 (±0.07)	0.7905 (±0.02)	0.3843 (±0.05)
MRLGCN	0.8592 (±0.04)	0.2600 (±0.04)	0.7868 (±0.03)	0.3648 (±0.05)	0.7932 (±0.03)	0.3651 (±0.05)
DOMR+†	0.8291 (±0.03)	0.2777 (±0.08)	0.8291 (±0.03)	0.2777 (±0.08)	0.8291 (±0.03)	0.2777 (±0.08)
SRGNN-Emo	**0.7973** (±0.03)	**0.3305** (±0.06)	**0.7707** (±0.01)	**0.3724** (±0.05)	**0.7411** (±0.02)	**0.4180** (±0.04)

Table 1 Multi-target regression performance for different models across three representation types. The best results are in boldface and the second-best results are underlined. All improvements of SRGNN-Emo compared to the second-best performing model are significant (Wilcoxon signed-rank test, $p < 0.05$). Models marked with † do not use any underlying track representation.

Model	wond	tran	tend	nost	peace	joya	power	sadn	tens	GEMS-9
MLP (musicnn)	0.9312	0.9653	0.7330	0.8936	0.6466	0.8099	0.8007	0.7711	0.7675	0.8132
DGI (Jukebox)	0.9059	0.9425	0.6647	0.8094	0.6088	0.7162	0.7511	**0.6627**	**0.6569**	0.7464
SRGNN-Emo (Jukebox)	**0.8972**	**0.9345**	**0.6518**	**0.8026**	0.6162	**0.6930**	**0.7425**	0.6690	0.6630	**0.7411**
(A) w/o $\mathcal{L}_{\text{Self}}$	0.9177	0.9384	0.6532	0.8192	**0.6086**	0.7050	0.7653	0.6713	0.6829	0.7513
(B) w/o $\mathcal{L}_{\text{EG}}$	0.9041	0.9387	0.6636	0.8245	0.6110	0.7082	0.7650	0.6845	0.6779	0.7530
(C) w/o $\mathcal{L}_{\text{Super}}$	1.2372	1.0996	1.2454	1.2210	1.3543	1.3329	1.2424	1.2907	1.2339	1.2508

Table 2 RMSE scores of models (using the best-performing representations from Table 1) across multiple emotion targets. Abbreviations of emotion dimensions correspond to wonder, transcendence, tenderness, nostalgia, peacefulness, joyful activation, power, sadness, and tension. All improvements of the best-performing models (boldface) are statistically significant compared to the second-best models (underline) per emotion dimension (Wilcoxon signed-rank test, $p < 0.05$).

them more closely with known emotion profile patterns, thus enhancing the model's ability to generalize from labeled to unlabeled data. Omitting the supervised loss ($\mathcal{L}_{\text{Super}}$) results in significant performance drops across all emotional dimensions, with RMSE scores rising substantially. This drastic decline highlights the importance of direct supervision in guiding the network toward accurate emotion profile predictions. Interestingly, while DGI outperformed SRGNN-Emo in two emotional dimensions—sadness and tension—it did not achieve consistently better performance across all emotion dimensions, indicating limitations in its ability to fully capture the emotional variations present in the dataset.

5.5 IMPACT OF HYPER-PARAMETERS

In this section, we investigate the impact of different hyper-parameters. We focus on the number of layers L in the wR-GCN and the number of emotion profile clusters K, since these hyper-parameters are related to various parts of the model architecture. Figure 3 shows the performance of our model with different settings of layers L on the described dataset using *musicnn* representations. A higher number of layers in the multi-relational network does not necessarily lead to an increase in performance due to the issue of over-smoothing, where node representations converge to the same values (Chen and Wong, 2020; Kipf and Welling, 2017). For our dataset, we could find a sweet spot layer setting L of 2.

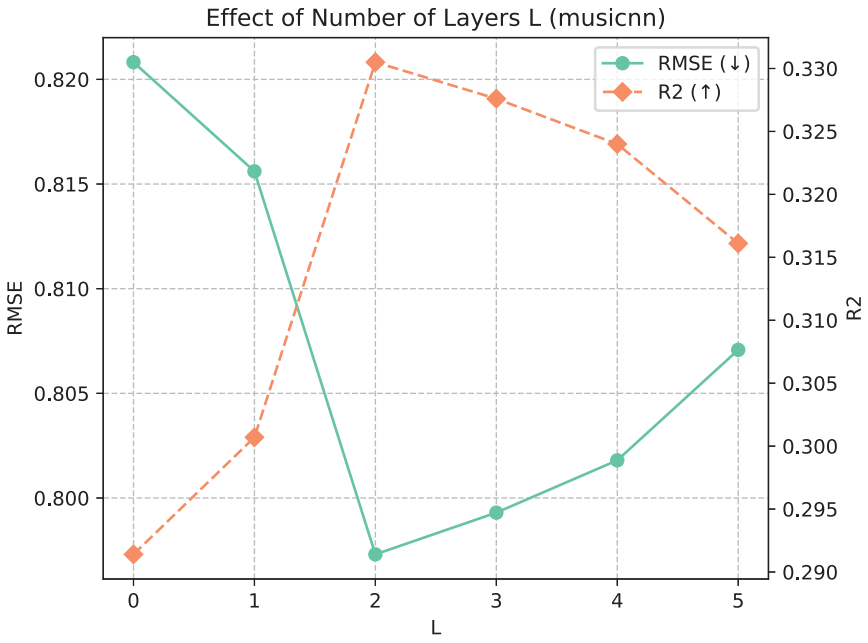

Figure 3 Performance impact of different number of layers *L* in our wR-GCN component.

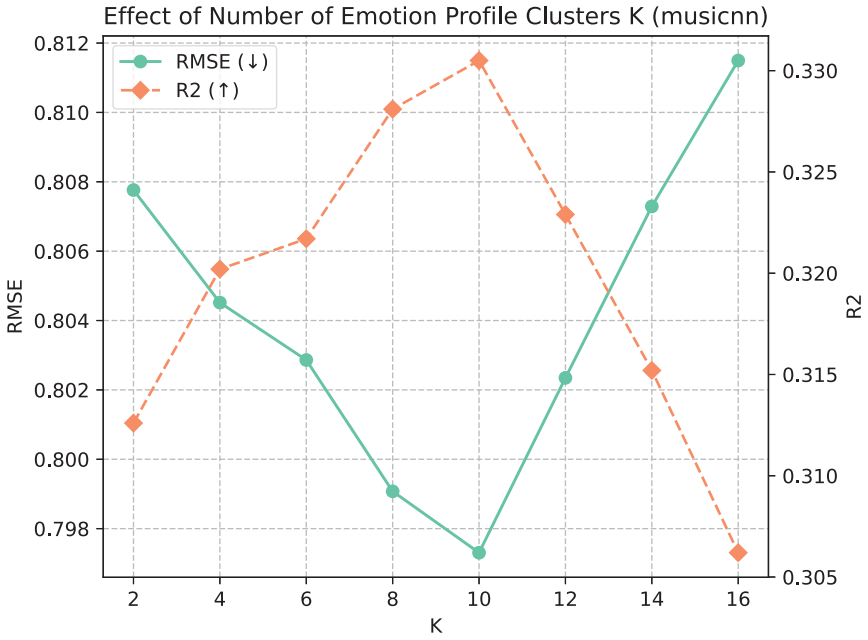

Figure 4 Performance impact of different number of emotion profile clusters *K*.

Figure 4 shows the performance differences between runs relying on *musicnn* representations with a different number of emotion profile clusters. The best-performing setting for *K* was 10 clusters, which aligns with previous analyses of emotion profiles in GEMS-9 by Chełkowska-Zacharewicz and Janowski (2021).

5.6 DATA EFFICACY STUDY

In this section, we assess the efficacy of our proposed SRGNN-Emo framework under varying levels of training data availability, investigating its performance in semi-supervised settings where labeled data are sparse. Figure 5 illustrates the model performances using different ratios of the training data, comparing the SRGNN-Emo framework with baseline models, including the baseline MLP and the semi-supervised graph-based approach DGI.

The results show that, as the amount of available labeled data increases, the performance of the MLP model significantly improves, exhibiting lower RMSE and higher R^2 values. This highlights its heavy reliance on large amounts of labeled data for generalization. In contrast, the semi-supervised models demonstrate superior performance even with minimal labeled data. Specifically, our SRGNN-Emo model maintains competitive RMSE scores and high R^2 values across various fractions of the training data, showing only a gradual decline in prediction accuracy as the training set size reduces.

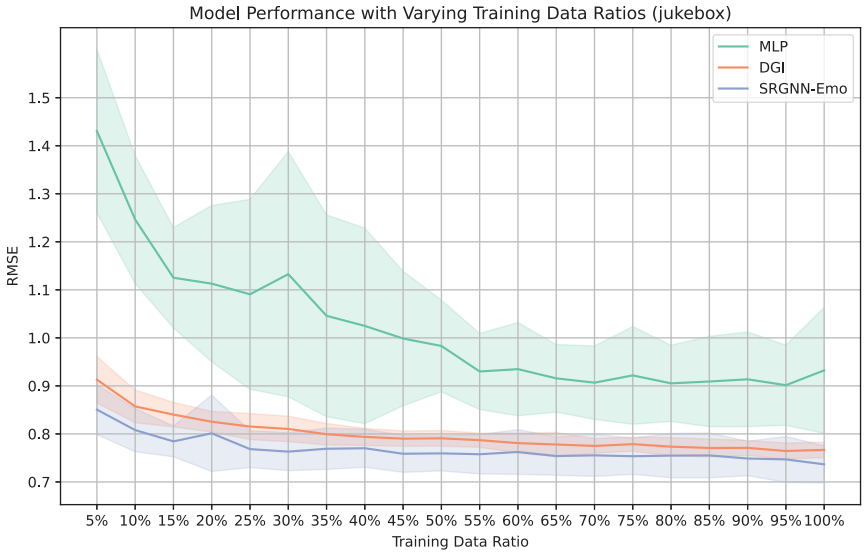

Figure 5 Model performances on different fractions of training data using *Jukebox* representations.

This indicates the model's robustness in scenarios with limited labeled data.

6 CONCLUSION

This work introduced SRGNN-Emo, a novel semi-supervised multi-relational GNN designed for nuanced MER trained on EMMA, a database with exceptionally rigorous annotations based on the domain-specific GEMS emotion model. By integrating semi-supervised learning with multi-relational graph structures and leveraging rich user interaction data, SRGNN-Emo effectively outperformed baseline models in capturing the complex emotional responses evoked by music. While our study leverages the GEMS model to capture a wide range of music-evoked emotions, our framework remains inherently flexible and can be adapted to alternative emotion models as future work. As a contribution, we enriched the existing Music4All-Onion dataset (Moscati et al., 2022) by adding emotion labels generated from our trained model, resulting in a fully labeled large-scale emotion-based dataset with 109,269 tracks. This enhanced dataset enables various applications, such as improved music retrieval, enhanced recommendation systems, and other related tasks.

DATA ACCESSIBILITY STATEMENT

Our source code is available at https://github.com/dbis-uibk/SRGNN-Emo. The dataset can be accessed under https://zenodo.org/records/15394646.

FUNDING INFORMATION

This research was funded in whole, or in part, by the Austrian Science Fund (FWF) [10.55776/P33526]. For

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Andreas Peintner was the main contributor to the conceptualization of the study, development of the methodology, and provision of key resources. Marta Moscati contributed to data curation, investigation, and methodological design. Yu Kinoshita and Richard Vogl assisted with the investigation, methodology, and validation. Hannah Strauss supported data curation, provided resources, and created visualizations. Peter Knees, Markus Schedl, and Marcel Zentner contributed resources and assisted in reviewing and editing the manuscript. Eva Zangerle supervised the project and contributed to writing and reviewing the article. All authors actively participated in the study design and approved the final version of the manuscript.

NOTES

1. https://github.com/dbis-uibk/SRGNN-Emo
2. https://musemap-tools.uibk.ac.at/emma/
3. https://www.last.fm
4. For *MAEST*, embeddings were extracted from transformer block 7 of the model, initialized with PaSST weights, and pre-trained on the Discogs20 dataset. For *Jukebox*, embeddings were extracted from layer 36, with mean pooling applied across the layer's output, following the methodology detailed in the original work.
5. This extended version of the dataset, including audio embeddings extracted from the described pre-trained models (*musicnn*, *MAEST*, and *Jukebox*), is made publicly available on https://zenodo.org/records/15394646.

AUTHOR AFFILIATIONS

Andreas Peintner https://orcid.org/0000-0001-7337-524X
University of Innsbruck, Innsbruck, Austria

Marta Moscati https://orcid.org/0000-0002-5541-4919
Johannes Kepler University Linz, Linz, Austria

Yu Kinoshita
TU Wien, Wien, Austria

Richard Vogl https://orcid.org/0000-0003-2488-0084
TU Wien, Wien, Austria

Peter Knees https://orcid.org/0000-0003-3906-1292
TU Wien, Wien, Austria

Markus Schedl https://orcid.org/0000-0003-1706-3406
Johannes Kepler University Linz, Linz, Austria

Hannah Strauss https://orcid.org/0000-0002-6235-4010
University of Innsbruck, Innsbruck, Austria

Marcel Zentner https://orcid.org/0000-0001-8580-8030
University of Innsbruck, Innsbruck, Austria

Eva Zangerle https://orcid.org/0000-0003-3195-8273
University of Innsbruck, Innsbruck, Austria

REFERENCES

Akiki, C., & Burghardt, M. (2021). Muse: The musical sentiment dataset. *Journal of Open Humanities Data*, *7*, 10.

Aljanaki, A., Wiering, F., and Veltkamp, R. C. (2014, October 27–31). Computational modeling of induced emotion using GEMS. In *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014*, Taipei, Taiwan, (pp. 373–378).

Aljanaki, A., Yang, Y.-H., and Soleymani, M. (2017). Developing a benchmark for emotional analysis of music. *PloS One*, *12*(3), e0173392. https://doi.org/10.1371/journal.pone.0173392

Alonso-Jiménez, P., Serra, X., and Bogdanov, D. (2023, November 5–9). Efficient supervised training of audio transformers for music representation learning. In *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023*, Milan, Italy, (pp. 824–831).

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., and McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, (pp. 1–8). IEEE.

Bhatti, A. M., Majid, M., Anwar, S. M., and Khan, B. (2016). Human emotion recognition and analysis in response to audio music using brain signals. *Computers in Human Behavior*, *65*, 267–275. https://doi.org/10.1016/j.chb.2016.08.029

Bogdanov, D., Lizarraga-Seijas, X., Alonso-Jiménez, P., and Serra, X. (2022). MusAV: A dataset of relative arousal-valence annotations for validation of audio models. In *International Society for Music Information Retrieval Conference (ISMIR 2022)*, Bengaluru, India.

Bogdanov, D., Won, M., Tovstogan, P., Porter, A., and Serra, X. (2019). The MTG-Jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*. Long Beach, CA, United States.

Castellon, R., Donahue, C., and Liang, P. (2021, November 7-12). Codified audio language modeling learns useful representations for music information retrieval. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online* (pp. 88–96).

Chełkowska-Zacharewicz, M., and Janowski, M. (2021). Polish adaptation of the Geneva emotional music scale: Factor structure and reliability. *Psychology of Music*, *49*(5), 1117–1131. https://doi.org/10.1177/0305735620934624

Chen, T., and Wong, R. C. (2020, August 23–27). Handling information loss of graph neural networks for session-based recommendation. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event*, CA, USA, (pp. 1172–1180). ACM.

Choi, J., Song, J.-H., and Kim, Y. (2018). An analysis of music lyrics by measuring the distance of emotion and sentiment. In *2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)* (pp. 176–181). IEEE.

da Silva, A. C. M., Silva, D. F., and Marcacini, R. M. (2022, December 4–8). Heterogeneous graph neural network for music emotion recognition. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022*, Bengaluru, India, (pp. 667–674).

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. (2020). Jukebox: A generative model for music.. *CoRR, abs/2005.00341*.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *International Conference on Machine Learning* (pp. 1263–1272). PMLR.

Gómez-Cañón, J. S., Cano, E., Eerola, T., Herrera, P., Hu, X., Yang, Y.-H., and Gómez, E. (2021). Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications. *IEEE Signal Processing Magazine*, *38*(6), 106–114. https://doi.org/10.1109/MSP.2021.3106232

Grover, A., and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016* (pp. 855–864). ACM.

Hamilton, W. L., Ying, Z., and Leskovec, J. (2017, December 4-9). Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, Long Beach, CA, USA, (pp. 1024–1034).

Hassani, K., and Ahmadi, A. H. K. (2020). Contrastive multi-view representation learning on graphs. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020* (Vol. 11, pp. 4116–4126). PMLR.

Horner, A., Hu, D. H., Wu, B., Yang, Q., and Zhong, E. (2013, May 2–4). SMART: Semi-supervised music emotion recognition with social tagging. In *Proceedings of the 13th SIAM International Conference on Data Mining*, Austin, Texas, USA, (pp. 279–287). SIAM.

Hu, X., and Downie, J. S. (2010, August 9–13). When lyrics outperform audio for music mood classification: A feature analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010*, Utrecht, Netherlands, (pp. 619–624). International Society for Music Information Retrieval.

Hu, X., Downie, J. S., and Ehmann, A. F. (2009). Lyric text mining in music mood classification. *American Music*, *183*(5), 49–209.

Jacobsen, P.-O., Strauss, H., Vigl, J., Zangerle, E., and Zentner, M. (2024). Assessing aesthetic music-evoked emotions in a minute or less: A comparison of the gems-45 and the gems-9. *Musicae Scientiae*, *0*(0), 10298649241256252. https://doi.org/10.1177/10298649241256252

Jia, Z., Lin, Y., Wang, J., Feng, Z., Xie, X., and Chen, C. (2021). HetEmotionNet: Two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition. In *Proceedings of the 29th ACM International Conference on Multimedia* (pp. 1047–1056).

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., Speck, J. A., and Turnbull, D. (2010). Music emotion recognition: A state of the art review. In *Proc. ISMIR*, *86*, 937–952.

Kingma, D. P., and Ba, J.Y. (2015, May 7–9). Adam: A method for stochastic optimization. In **Y. Bengio and Y. LeCun** (Eds.), *3rd International Conference on Learning representations, ICLR 2015, San Diego, CA, USA, Conference Track Proceedings.*

Kipf, T. N., and Welling, M. (2017, April 24–26). Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Conference Track proceedings.* OpenReview.net.

Laurier, C., Sordo, M., Serra, J., and Herrera, P. (2009, October 26–30). Music mood representations from social tags. In *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009,* (pp. 381–386). Kobe International Conference Center, Kobe, Japan. International Society for Music Information Retrieval.

Lee, J., Oh, Y., In, Y., Lee, N., Hyun, D., and Park, C. (2022, July 11–15). GraFN: Semi-supervised node classification on graph with few labels via non-parametric distribution assignment. *In SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval,* (pp. 2243–2248). Madrid, Spain. ACM.

Mignot, R., and Peeters, G. (2019). An analysis of the effect of data augmentation methods: Experiments for a musical genre classification task. *Transactions of the International Society for Music Information Retrieval*, *2*(1), 97–110.

Moscati, M., Parada-Cabaleiro, E., Deldjoo, Y., Zangerle, E., and Schedl, M. (2022, October 17–21). Music4All-onion - A large-scale multi-faceted content-centric music recommendation dataset. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management,* Atlanta, GA, USA, (pp. 4339–4343). ACM.

Moscati, M., Parada-Cabaleiro, E., Deldjoo, Y., Zangerle, E., and Schedl, M. (2025). *Music4All-onion*. Zenodo.

Moscati, M., Strauß, H., Jacobsen, P., Peintner, A., Zangerle, E., Zentner, M., and Schedl, M. (2024, July 1–4). Emotion-based music recommendation from quality annotations and large-scale user-generated tags. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization,UMAP 2024*, Cagliari, Italy, (pp. 159–164). ACM.

Panda, R., Malheiro, R., and Paiva, R. P. (2018). Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing*, *11*(4), 614–626.

Panda, R., Malheiro, R., and Paiva, R. P. (2020). Audio features for music emotion recognition: A survey. *IEEE Transactions on Affective Computing*, *14*(1), 68–88.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). DeepWalk: Online learning of social representations. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14* (pp. 701–710). ACM.

Pons, J., and Serra, X. (2019). Musicnn: Pre-trained convolutional neural networks for music audio tagging. *CoRR*. http://dx.doi.org/10.48550/arXiv.1909.06654

Rajan, R., Antony, J., Joseph, R. A., and Thomas, J. M. (2021). Audio-mood classification using acoustic-textual feature fusion. In *2021 Fourth International Conference on Microelectronics, Signals and Systems (ICMSS)* (pp. 1–6). IEEE.

Santana, I. A. P., Pinhelli, F., Donini, J., Catharin, L., Mangolin, R. B., Feltrim, V. D., and Domingues, M. A. (2020). Music4All: A new music database and its applications. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)* (pp. 399–404). IEEE.

Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den, Berg, R., Titov, I., and Welling, M. (2018, June 3–7). Modeling relational data with graph convolutional networks. In *The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, proceedings 15*, (pp. 593–607). Springer.

Strauss, H., Vigl, J., Jacobsen, P.-O., Bayer, M., Talamini, F., Vigl, W., Zangerle, E., and Zentner, M. (2024). The emotion-to-music mapping atlas (EMMA): A systematically organized online database of emotionally evocative music excerpts. *Behavior Research Methods*, 1–18. https://doi.org/10.3758/s13428-023-02178-2

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). LINE: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015* (pp. 1067–1077). ACM.

Thakoor, S., Tallec, C., Azar, M. G., Munos, R., Veličković, P., and Valko, M. (2021). Bootstrapped representation learning on graphs. In *ICLR 2021 Workshop on Geometrical and Topological Representation Learning*.

Tovstogan, P., Bogdanov, D., and Porter, A. (2021, December). MediaEval 2021: Emotion and theme recognition in music using jamendo. In *Working Notes Proceedings of the MediaEval 2021 Workshop, Online, CEUR Workshop Proceedings*, *318*(1), 13–15. CEUR-WS.org

Trost, W., Ethofer, T., Zentner, M., and Vuilleumier, P. (2012). Mapping aesthetic musical emotions in the brain. *Cerebral Cortex*, *22*(12), 2769–2783. https://doi.org/10.1093/cercor/bhr353

Vashishth, S., Sanyal, S., Nitin, V., and Talukdar, P. P. (2020, April 26–30). Composition-based multi-relational graph convolutional networks. In *8th International Conference on Learning Representations, ICLR 2020*. Addis Ababa, Ethiopia, OpenReview.net.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Conference Track Proceedings*. OpenReview.net.

Velickovic, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., and Hjelm, R. D. (2019). Deep graph infomax. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net.

Xue, H., Xue, L., and Su, F. (2015). Multimodal music mood classification by fusion of audio and lyrics. In *MultiMedia Modeling: 21st International Conference, MMM 2015 Proceedings, Part II 21* (pp. 26–37). Springer.

Yang, J. (2021). A novel music emotion recognition model using neural network technology. *Frontiers in Psychology*, *12*, 760060. https://doi.org/10.3389/fpsyg.2021.760060

Yang, Y.-H., and Chen, H. H. (2011). *Music emotion recognition*. CRC Press.

Yang, Y.-H., Lin, Y.-C., Su, Y.-F., and Chen, H. H. (2008). A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(2), 448–457. https://doi.org/10.1109/TASL.2007.911513

Zad, S., Heidari, M., James Jr, H., and Uzuner, O. (2021). Emotion detection of textual data: An interdisciplinary survey. In *2021 IEEE World AI IoT Congress (AIIoT)* (pp. 0255–0261). IEEE.

Zentner, M., Grandjean, D., and Scherer, K. R. (2008). Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion*, *8*(4), 494–521. https://doi.org/10.1037/1528-3542.8.4.494

Zhang, K., Zhang, H., Li, S., Yang, C., and Sun, L. (2018). The PMEmo dataset for music emotion recognition. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, (pp. 135–142).

Zhang, L., Yang, X., Zhang, Y., and Luo, J. (2023, November 5–9). Dual attention-based multi-scale feature fusion approach for dynamic music emotion recognition. In *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023,* Milan, Italy, (pp. 207–214).

Zhou, Z., and Li, M. (2005, July 30–August 5). Semi-supervised regression with co-training. In *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, (pp. 908–916). Edinburgh, Scotland, UK, Professional Book Center.

Zhu, X., and Ghahramani, Z. (2002). *Learning from labeled and unlabeled data with label propagation*. Technical report, Carnegie Mellon University.

TO CITE THIS ARTICLE:
Peintner, A., Moscati, M., Kinoshita, Y., Vogl, R., Knees, P., Schedl, M., Strauss, H., Zentner, M., & Zangerle, E. (2025). Nuanced Music Emotion Recognition via a Semi-Supervised Multi-Relational Graph Neural Network. *Transactions of the International Society for Music Information Retrieval, 8*(1), 140–153. DOI: https://doi.org/10.5334/tismir.235

Submitted: 29 October 2024 **Accepted:** 30 April 2025 **Published:** 11 June 2025