Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Emotion-Aware Inductive Heterogeneous Graph Neural Networks for Cold-Start Music Recommendation

Haibin Zhang

2007053@hnpi.edu.cn

Henan Polytechnic Institute

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11409491/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Emotion-Aware Inductive Heterogeneous Graph Neural Networks for Cold-Start Music Recommendation

Haibin Zhang[1*]

[1*]Art Education Center, Henan Polytechnic Institute, Nanyang, 473000, China.

Corresponding author(s). E-mail(s): 2007053@hnpi.edu.cn;

Abstract

Sparse listening histories and newly introduced tracks remain difficult cases for music recommendation. Affective context can supplement limited interaction evidence, while acoustic content can represent tracks that were absent during training. This study presents an Emotion-Aware Inductive Heterogeneous Graph Network (EAI-HGN) that combines these information sources. The framework connects users, tracks, artists, and emotion categories; encodes track content with a convolutional–recurrent network; and uses a sparsity-aware semantic gate to combine interaction-based and emotion-based views. Evaluation on a processed subset of the #nowplaying-RS dataset compares EAI-HGN with five established baselines under general recommendation, three-interaction user adaptation, and unseen-item holdout conditions. The corresponding Recall@10 values are 0.1720, 0.1395, and 0.1350, representing relative improvements of 8.31%, 15.29%, and 41.36% over the strongest tested baseline in each condition. Component ablations indicate that content encoding is particularly important for unseen tracks, whereas affective relations and adaptive aggregation support users with short histories. These results support coordinating collaborative, acoustic, and affective information for sparse music recommendation. The evidence concerns few-shot user adaptation and content-supported unseen-item inference; it does not establish personalization for users with no history or context.

Keywords: music recommendation, heterogeneous graph neural networks, cold start, affective context, inductive learning, semantic attention

1 Introduction

Music streaming services provide access to large and diverse catalogs, but effective discovery still depends on identifying tracks that suit an individual listener. Recommendation must accommodate repeated consumption, changing listening contexts, and preferences beyond stable artist or genre affinities [1]. Collaborative filtering learns these preferences from observed user–item interactions [2, 3]. Its evidence becomes weak when a listener

has only a short history, and it cannot directly estimate a collaborative representation for a track with no training interactions.

Graph-based recommendation models higher-order relationships among users, items, and auxiliary information [4, 5]. Heterogeneous graph architectures distinguish different entity and relation types, enabling separate semantic views of a recommendation task [6, 7]. Their effectiveness for unseen entities nevertheless depends on the input representation and inference procedure. A model that relies on a learned identifier for every track requires an additional feature-based pathway before it can represent a new track.

Music offers two complementary sources of such evidence. First, acoustic content can be mapped to transferable representations [8, 9]. Second, affective associations can link tracks to listening context when collaborative history is sparse. These associations need careful interpretation: a mood-related hashtag describes an event, a valence descriptor characterizes a track, and neither directly measures a listener's experienced emotion [10]. Their computational value is as contextual and semantic information, rather than psychological ground truth.

Recent studies investigate semantic multimodal music retrieval, emotion-aware recommendation, and knowledge-enhanced preference modeling [11–13]. The question considered here is more specific: how can an explicit content pathway for unseen tracks be coordinated with affective and collaborative views as the amount of available user history changes?

This study presents an Emotion-Aware Inductive Heterogeneous Graph Network (EAI-HGN). Its contributions are:

1. A heterogeneous representation of users, tracks, artists, and emotion categories, in which affective associations supplement direct listening relationships.
2. An inductive acoustic encoder for tracks absent from training, coupled with a semantic attention gate conditioned on the sparsity of the available user history.
3. An evaluation across general recommendation, three-interaction user adaptation, and unseen-item holdout conditions, with component ablations and analyses of semantic attention and training behavior.

The user experiment retains three interactions per selected user and therefore evaluates few-shot adaptation. It is distinct from personalization for a user with neither history nor context. Figure 1 summarizes the information flow.

2 Related Work

2.1 Graph Learning for Recommendation

Graph recommenders propagate information through user–item connectivity to capture dependencies beyond direct observations. Neural graph collaborative filtering (NGCF) explicitly models embedding propagation [14]. LightGCN simplifies this process and illustrates the importance of distinguishing the benefit of graph structure from that of additional transformations [15]. Broader surveys organize the field by graph construction, information availability, and recommendation task [4, 5].

Heterogeneous graph attention networks (HANs) combine node-level attention with semantic-level attention over meta-path views [6]. Heterogeneous graph transformers (HGTs) use type-dependent transformations and attention [7]. These architectures are

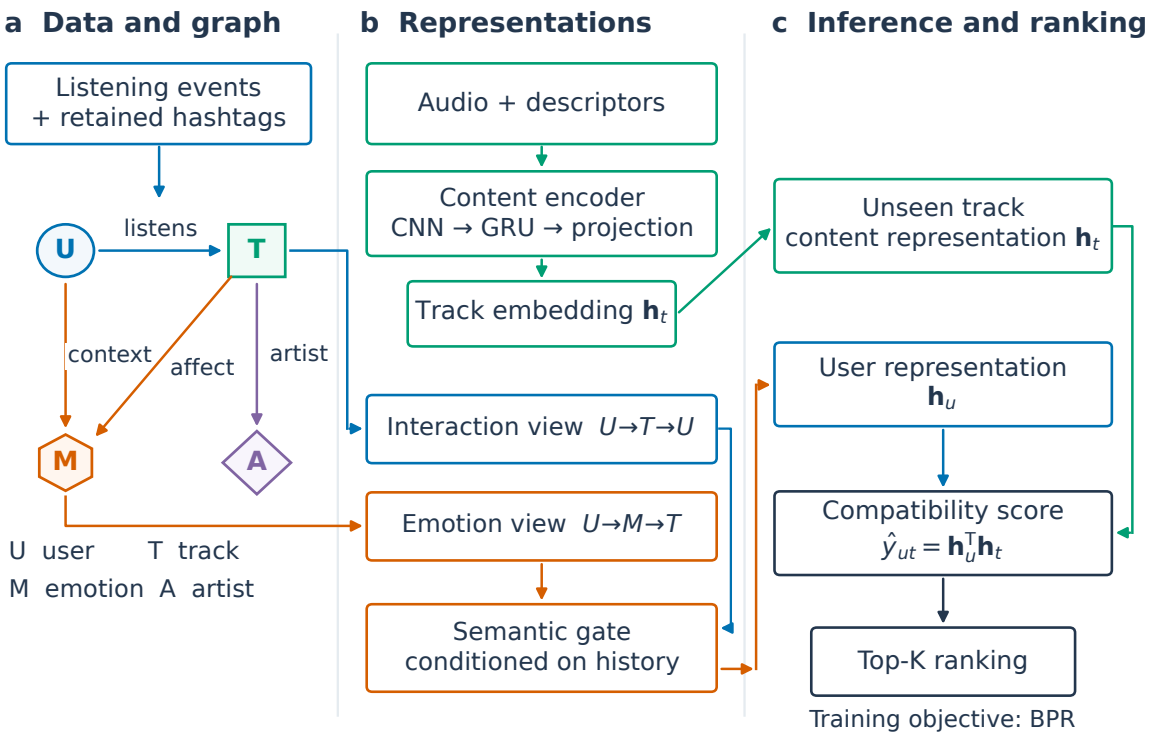

Fig. 1 Overview of EAI-HGN. A typed graph supplies interaction and affective semantic views. A convolutional–recurrent encoder maps available audio and acoustic descriptors to a track representation, including for tracks absent from training. A sparsity-aware gate combines the user views, and the user–track compatibility score determines the ranking. Artist nodes encode catalog attribution

not inherently restricted to transductive use: their handling of unseen entities depends on features, identifiers, and inference-time neighborhoods. EAI-HGN focuses on an explicit content representation for unseen tracks and a history-dependent combination of semantic views.

2.2 Content-Based and Inductive Cold-Start Recommendation

Cold start includes different information conditions. An unseen item may have content and metadata but no training interactions; a few-shot user has a short support history; and a zero-history user requires another observable source of context for personalization [1]. These conditions should not be treated as interchangeable.

GraphSAGE learns transferable neighborhood aggregation functions when suitable features and neighborhoods are available [9]. Deep content-based music recommendation similarly maps audio to representations related to preference factors [8]. Knowledge-enhanced sequential recommendation offers another route to richer item representations: Khaligh and Shayegan combine knowledge graph embeddings, contrastive predictive coding, and sequential preference modeling [13]. EAI-HGN instead emphasizes the coordination of acoustic inference for unseen tracks with an affective view for sparse users. Its experiments do not establish superiority over contemporary knowledge-enhanced or multimodal models that were not evaluated.

2.3 Affective and Semantic Music Recommendation

Musical emotion depends on multiple psychological mechanisms, so perceived affect, experienced emotion, and contextual tags must be distinguished [10]. The #nowplaying-RS dataset provides listening events, contextual features, and track descriptors for studying context-aware recommendation [16]. Hashtags are useful auxiliary signals, but they can be noisy or unrelated to a listener's immediate state.

Rumiantcev combines neural multimodal representations with ontology-guided semantic organization for music similarity and recommendation [11]. Liu studies deep music-emotion classification within a context-aware recommendation framework [12]. The present work uses affective categories as typed graph relations and adjusts their contribution according to history sparsity. Its central task is top-K ranking under limited interaction evidence, rather than music-emotion classification. Artist metadata and other catalog attributes remain complementary information sources.

3 Materials and Methods

3.1 Task Definition and Heterogeneous Graph

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \phi, \psi)$ denote a heterogeneous graph, where ϕ maps a node to its entity type and ψ maps an edge to its relation type. The node set is

$$\mathcal{V} = \mathcal{U} \cup \mathcal{T} \cup \mathcal{A} \cup \mathcal{M}, \tag{1}$$

where $\mathcal{U}$, $\mathcal{T}$, $\mathcal{A}$, and $\mathcal{M}$ contain users, tracks, artists, and emotion categories, respectively. The relation types include user–track listening events, track–artist attribution, user–emotion contextual associations, and track–emotion associations. The emotion categories are semantic attributes rather than direct measurements of a user's mental state.

Mood-related hashtags are filtered using a lexicon-based procedure and grouped into 15 emotion categories. These categories define the emotion nodes. Track-level affective descriptors provide complementary content information. The resulting associations allow tracks and users to share an affective context even when their direct collaborative connection is weak. Metadata derived from a listening event belongs to that event; it cannot be treated as information available before the event if the event itself is a held-out recommendation target.

For a user u and a candidate track t, the model estimates a ranking score $\hat{y}_{ut}$. The objective is to rank relevant, unobserved tracks above other candidates. This score is not interpreted as a calibrated interaction probability. Let $\mathcal{T}_{\text{train}}$ denote tracks available during model training and $\mathcal{T}_{\text{new}}$ denote held-out tracks. The unseen-item condition requires

$$\mathcal{T}_{\text{train}} \cap \mathcal{T}_{\text{new}} = \varnothing, \qquad \deg_{\text{interaction,train}}(t) = 0 \quad \text{for } t \in \mathcal{T}_{\text{new}}. \tag{2}$$

Zero interaction degree does not imply that every content or metadata relation is unavailable. A new track can still have an artist, acoustic descriptors, and an accessible recording. The inductive content pathway is designed to generate a representation without requiring training interactions for that track.

3.2 Inductive Acoustic Content Encoder

The track encoder combines time–frequency information with explicit acoustic attributes. Let $\mathbf{S}_t \in \mathbb{R}^{F\times\tau}$ be the log-Mel spectrogram of an available track recording, where F is the number of Mel-frequency bins and τ is the number of temporal frames. Let $\mathbf{v}_t \in \mathbb{R}^{d_{\text{attr}}}$ contain descriptors such as valence, energy, and tempo. These descriptors are track features; valence is not a raw waveform embedding.

A convolutional network extracts local acoustic patterns. With $\mathbf{C}_t^{(0)} = \mathbf{S}_t$, a convolutional layer is represented by

$$\mathbf{C}_t^{(\ell)} = \mathrm{ReLU}\Big(\mathbf{W}_{\text{conv}}^{(\ell)} * \mathbf{C}_t^{(\ell-1)} + \mathbf{b}_{\text{conv}}^{(\ell)}\Big). \tag{3}$$

The resulting features are arranged as a temporal sequence $\{\mathbf{c}_{t,k}\}_{k=1}^{L_t}$. A gated recurrent unit (GRU) [17] models the sequence:

$$\mathbf{g}_{t,k} = \mathrm{GRU}_{\theta_{\text{GRU}}}\left(\mathbf{c}_{t,k}, \mathbf{g}_{t,k-1}\right). \tag{4}$$

The final recurrent state and explicit attributes are concatenated and projected into the d-dimensional recommendation space:

$$\mathbf{h}_t = F_\theta(\mathbf{S}_t, \mathbf{v}_t) = \sigma[\mathbf{W}_{\text{proj}}\left(\mathbf{g}_{t,L_t} \| \mathbf{v}_t\right) + \mathbf{b}_{\text{proj}}], \tag{5}$$

where $\|$ denotes concatenation and σ is an element-wise activation. The encoder parameters θ include the convolutional, recurrent, and projection parameters. Once trained, the encoder can process an unseen track using its available content. This avoids optimizing a new item identifier, although extracting audio features still incurs an inference cost.

The public #nowplaying-RS release provides acoustic descriptors rather than a collection of raw audio waveforms [16]. The waveform-based branch therefore requires a recording matched to the track identifier. The availability and permitted use of those recordings are separate from the availability of listening events and descriptors.

3.3 Semantic Views and User Representation

The aggregation module distinguishes an interaction view and an emotion view:

$$\Phi_{\text{int}} : \mathcal{U} \to \mathcal{T} \to \mathcal{U}, \tag{6}$$

$$\Phi_{\text{emo}} : \mathcal{U} \to \mathcal{M} \to \mathcal{T}. \tag{7}$$

The first view relates a user to other users through shared listening history. The second relates a user to candidate track content through an affective association. Their endpoints have different entity types. Type-specific projections into a common latent space are therefore needed before their representations can be combined. Figure 2 shows the two schemes.

Let $\mathcal{N}_\Phi(u)$ denote the neighbors accessible from u under semantic view Φ. A path-specific user representation can be written as

$$\mathbf{z}_u^\Phi = \mathrm{AGG}_\Phi\left(\left\{\mathbf{P}_{\phi(v)}\mathbf{h}_v : v \in \mathcal{N}_\Phi(u)\right\}\right), \tag{8}$$

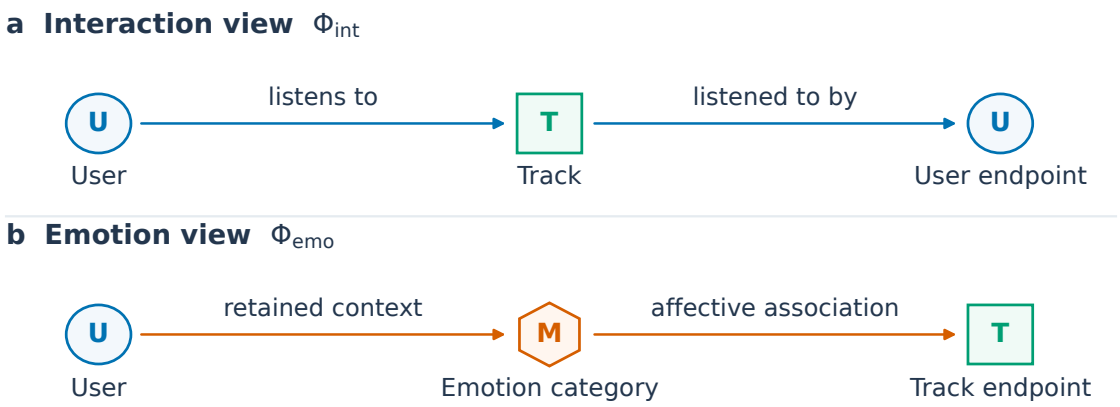

Fig. 2 Semantic views used to construct a user representation. The interaction view terminates at users, whereas the emotion view terminates at tracks. Type-specific projections align these endpoint representations before semantic aggregation. The symbol M denotes an emotion category

where $\mathbf{P}_{\phi(v)}$ aligns the endpoint representation to the user latent space and AGG_Φ is the node-level attention aggregator. This formulation distinguishes within-view aggregation from the subsequent weighting of semantic views.

3.4 Sparsity-Aware Semantic Attention

Let n_u be the number of interactions available for user u and let $\delta_u = \delta(n_u)$ be a scalar sparsity indicator that increases as available history decreases. The gate uses the path-specific representation and this indicator to compute

$$a_{u,\Phi} = \mathbf{q}^\top \tanh\left(\mathbf{W}_s \mathbf{z}_u^\Phi\right) + \lambda_\Phi \delta_u, \tag{9}$$

followed by normalized attention:

$$\beta_{u,\Phi} = \frac{\exp(a_{u,\Phi})}{\sum_{\Phi' \in \{\Phi_{\text{int}}, \Phi_{\text{emo}}\}} \exp(a_{u,\Phi'})}. \tag{10}$$

The user representation is

$$\mathbf{h}_u = \sum_{\Phi \in \{\Phi_{\text{int}}, \Phi_{\text{emo}}\}} \beta_{u,\Phi} \mathbf{z}_u^\Phi. \tag{11}$$

The coefficients λ_Φ are learned. For the two-view gate, the attention ratio satisfies

$$\begin{aligned} \log \frac{\beta_{u,\text{emo}}}{\beta_{u,\text{int}}} = \mathbf{q}^\top \left[\tanh(\mathbf{W}_s \mathbf{z}_u^{\text{emo}}) - \tanh(\mathbf{W}_s \mathbf{z}_u^{\text{int}})\right] \\ + (\lambda_{\text{emo}} - \lambda_{\text{int}})\delta_u. \end{aligned} \tag{12}$$

Holding the semantic representations fixed, increasing sparsity favors the emotion view only when $\lambda_{\text{emo}} > \lambda_{\text{int}}$. Thus, the architecture permits a sparsity-dependent response without forcing a predetermined preference for emotion. Section 5.4 examines the observed group-level pattern.

The emotion view also requires an available contextual association. A few-shot user can obtain such a representation from retained context, but a completely new user with no context cannot be personalized by this mechanism alone. An empty semantic view has no informative neighborhood and should not be interpreted as evidence about a user's preference.

3.5 Ranking Objective

User and track embeddings define a compatibility score,

$$\hat{y}_{ut} = \mathbf{h}_u^\top \mathbf{h}_t. \tag{13}$$

Training uses Bayesian personalized ranking (BPR) [2]. For a triplet (u, i, j), i is an observed positive track and j is sampled from unobserved candidates. The loss is

$$\mathcal{L} = -\sum_{(u,i,j)\in\mathcal{D}} \log \operatorname{sigmoid}\left(\hat{y}_{ui} - \hat{y}_{uj}\right) + \gamma\|\Theta\|_2^2, \tag{14}$$

where Θ contains all model parameters and γ controls regularization. An unobserved track is a training negative in the pairwise objective, not a confirmed disliked item. The objective learns relative rankings from implicit feedback.

4 Experimental Design

4.1 Dataset and Evaluation Conditions

Experiments use #nowplaying-RS, introduced by Poddar et al. in 2018 [16]. Its original release contains 11,639,541 listening events, 138,781 users, and 346,273 tracks, derived from Twitter listening events collected for 2014. These release statistics should be distinguished from the processed subset used here.

After removing users and tracks with fewer than five interactions, the experimental graph contains approximately 12,405 users, 58,230 tracks, and more than 840,000 interaction edges. The affective vocabulary contains 15 categories. Three evaluation conditions are considered:

1. **General recommendation:** evaluate recommendation with the ordinary available interaction history.
2. **Few-shot user recommendation:** select 20% of users and limit the available history of each selected user to three interactions when constructing the recommendation profile.
3. **Unseen-item recommendation:** hold out 10% of tracks from training and represent these tracks through their content at inference.

The minimum-interaction filter applies to the processed dataset before the few-shot reduction. The three-observation profiles are consequently an evaluation condition imposed after filtering. The item split evaluates held-out identities and should not be interpreted as a chronological deployment experiment. In either cold-start setting, valid evaluation requires that target interactions and target-event hashtags do not contribute to the user representation or graph context used to predict those targets.

Table 1 Evaluation conditions and the information available for recommendation

Condition	Available information and interpretation
General	Ordinary available interaction history; no imposed three-interaction user restriction
Few-shot user	20% of users selected; three retained interactions per selected user form the available history
Unseen item	10% of tracks held out from training; their representations are obtained from available content at inference

4.2 Baselines and Ranking Metrics

The comparison includes BPR matrix factorization (BPR-MF) [2], neural collaborative filtering (NCF) [3], GraphSAGE [9], NGCF [14], and HAN [6]. These are representative baselines covering factorization, nonlinear interaction modeling, inductive aggregation, collaborative graph propagation, and heterogeneous semantic aggregation. NCF is a neural collaborative model rather than a matrix factorization method.

HAN uses structural and semantic metadata, including artists and general tags. GraphSAGE provides a feature-based inductive baseline. The unseen-item behavior of an identifier-based baseline depends on its fallback representation and available inference-time features. Low scores for one implementation do not establish that every variant of that model is incapable of cold-start inference. Likewise, comparison with a metadata-aware model does not by itself isolate the effect of emotion unless the models receive matched content inputs.

Recall and normalized discounted cumulative gain (NDCG) [18] measure top-K ranking performance. If $\mathcal{R}_u$ is the relevant test set and $\mathcal{L}_u^K$ is the recommended list, then

$$\text{Recall@}\,K(u) = \frac{|\mathcal{R}_u \cap \mathcal{L}_u^K|}{|\mathcal{R}_u|}. \tag{15}$$

For binary relevance, discounted cumulative gain is

$$\text{DCG@}\,K(u) = \sum_{r=1}^{K} \frac{\mathbb{I}[\mathcal{L}_u^K(r) \in \mathcal{R}_u]}{\log_2(r+1)}, \qquad \text{NDCG@}\,K(u) = \frac{\text{DCG@}\,K(u)}{\text{IDCG@}\,K(u)}. \tag{16}$$

The comparative results below report $K = 10$. Recall@20 is used for validation and early stopping. Results are averaged across five runs. Run-level outcomes and dispersion estimates are not reported, so the mean differences below should not be interpreted as statistical significance tests. Candidate construction and data splitting also affect ranking scores and constrain comparisons with results from other studies.

4.3 Implementation

The framework is implemented with PyTorch and the Deep Graph Library. Training uses an NVIDIA RTX 4090 GPU with 24 GB of memory under Ubuntu 22.04. Adam [19] optimizes the model with a learning rate of 10^{-3}, a batch size of 1024, and regularization strength $\gamma = 10^{-4}$. The embedding dimension is 128, and network weights use Xavier

initialization [20]. Early stopping monitors validation Recall@20 with a patience of ten epochs.

The acoustic preprocessing resamples recordings to 22.05 kHz and extracts log-Mel spectrograms using a window length of 2048 samples, a hop length of 512 samples, and 128 Mel-frequency bins. The acoustic attributes and convolutional–recurrent representation are fused as described in Equation 5. Table 2 summarizes the reported configuration. The matched recording source and audio-coverage statistics are not specified in the present report, which limits exact reconstruction of the waveform-based pipeline.

Table 2 Reported implementation and training configuration

Component	Setting
Embedding dimension	128
Optimizer	Adam, learning rate 10^{-3}
Batch size	1024
Regularization	$\gamma = 10^{-4}$
Validation monitor	Recall@20; patience of 10 epochs
Audio sampling rate	22.05 kHz
Spectrogram extraction	Window 2048; hop 512; 128 Mel bins
Repeated runs	5

5 Results

5.1 Comparison across Recommendation Settings

Table 3 reports mean ranking performance across five runs. EAI-HGN achieves the highest Recall@10 and NDCG@10 among the evaluated models in all three conditions. The strongest baseline differs by condition: NGCF for general recommendation, HAN for few-shot users, and GraphSAGE for unseen items. Relative improvements are therefore computed separately for each metric and setting.

In the general condition, EAI-HGN improves Recall@10 from NGCF's 0.1588 to 0.1720 and NDCG@10 from 0.1345 to 0.1468. These increases correspond to 8.31% and 9.14%, respectively. For users restricted to three interactions, the strongest baseline is HAN. EAI-HGN increases Recall@10 from 0.1210 to 0.1395 and NDCG@10 from 0.1015 to 0.1162, yielding relative gains of 15.29% and 14.48%.

For unseen items, the feature-based GraphSAGE baseline performs best among the alternatives. EAI-HGN increases Recall@10 from 0.0955 to 0.1350 and NDCG@10 from 0.0744 to 0.1125, corresponding to 41.36% and 51.21%. This pattern is consistent with the importance of an explicit content pathway. The size of the difference also depends on the content inputs and unseen-item fallback available to each baseline; it does not isolate architectural effects from information availability.

Table 3 Mean top-10 ranking performance across five runs. Bold marks the highest value; underlining marks the strongest baseline in each column

	General		Few-shot user		Unseen item	
Model	R@10	N@10	R@10	N@10	R@10	N@10
BPR-MF	0.1245	0.0982	0.0812	0.0615	0.0102	0.0051
NCF	0.1310	0.1055	0.0895	0.0688	0.0115	0.0062
GraphSAGE	0.1422	0.1180	0.1020	0.0812	<u>0.0955</u>	<u>0.0744</u>
NGCF	<u>0.1588</u>	<u>0.1345</u>	0.1150	0.0935	0.0155	0.0088
HAN	0.1575	0.1332	<u>0.1210</u>	<u>0.1015</u>	0.0188	0.0105
EAI-HGN	**0.1720**	**0.1468**	**0.1395**	**0.1162**	**0.1350**	**0.1125**
Gain (%)	8.31	9.14	15.29	14.48	41.36	51.21

R@10: Recall@10; N@10: NDCG@10. Gain is the relative increase over the strongest baseline in the corresponding column. Dispersion estimates are unavailable in the reported results.

5.2 Component Ablation

Three variants remove the emotion view, replace the content encoder with item-identifier embeddings, or replace adaptive semantic gating with an equal-weight average. Figure 3 shows the reported Recall@10 comparisons. Removing the content encoder produces the largest deterioration for unseen items, where an identifier alone lacks a learned representation. Removing affective relations has a clearer effect on few-shot users, and fixed semantic averaging also performs below the complete model.

The components therefore address complementary information deficits. Content encoding supplies a representation for an unseen track, whereas affective relations and semantic weighting support inference from a short user history. These effects remain conditional on the available content and the definitions of the ablated variants.

5.3 Embedding Dimension

Figure 4 compares embedding dimensions of 32, 64, 128, 256, and 512. Both ranking metrics improve up to 128 and then decline modestly. A dimension of 128 is therefore retained within the tested range. The decline at larger dimensions does not establish overfitting as its cause; that interpretation would require matched training and validation evidence across dimensions.

5.4 Attention Weights and History Sparsity

Figure 5 reports average semantic weights for four history-sparsity groups. The emotion view receives 75%, 60%, 35%, and 15% of the weight in the groups with fewer than five, 5–20, 20–50, and more than 50 available interactions, respectively. The complementary interaction-view weights are 25%, 40%, 65%, and 85%.

The group-level trend agrees with the intended role of the gate: the emotion view contributes more when interaction history is short. Equation 12 explains how this dependence can be expressed by the model. The attention averages describe model

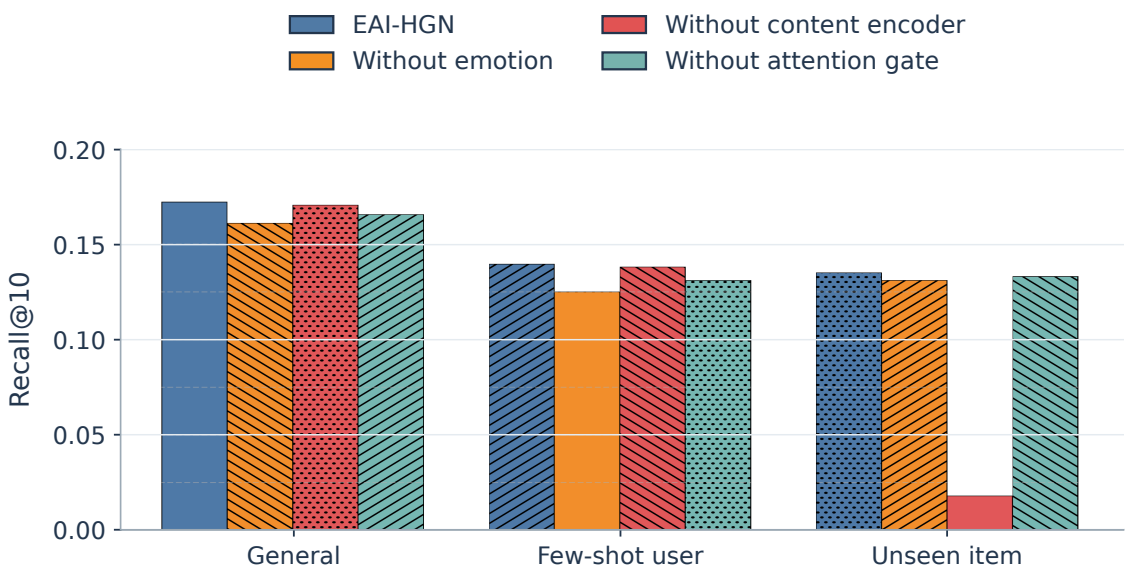

Fig. 3 Reported component ablation under the three recommendation conditions. The content encoder is particularly important for unseen items, while affective relations and adaptive semantic gating contribute to the few-shot user condition. Bars represent the reported scores; run-level uncertainty is not available

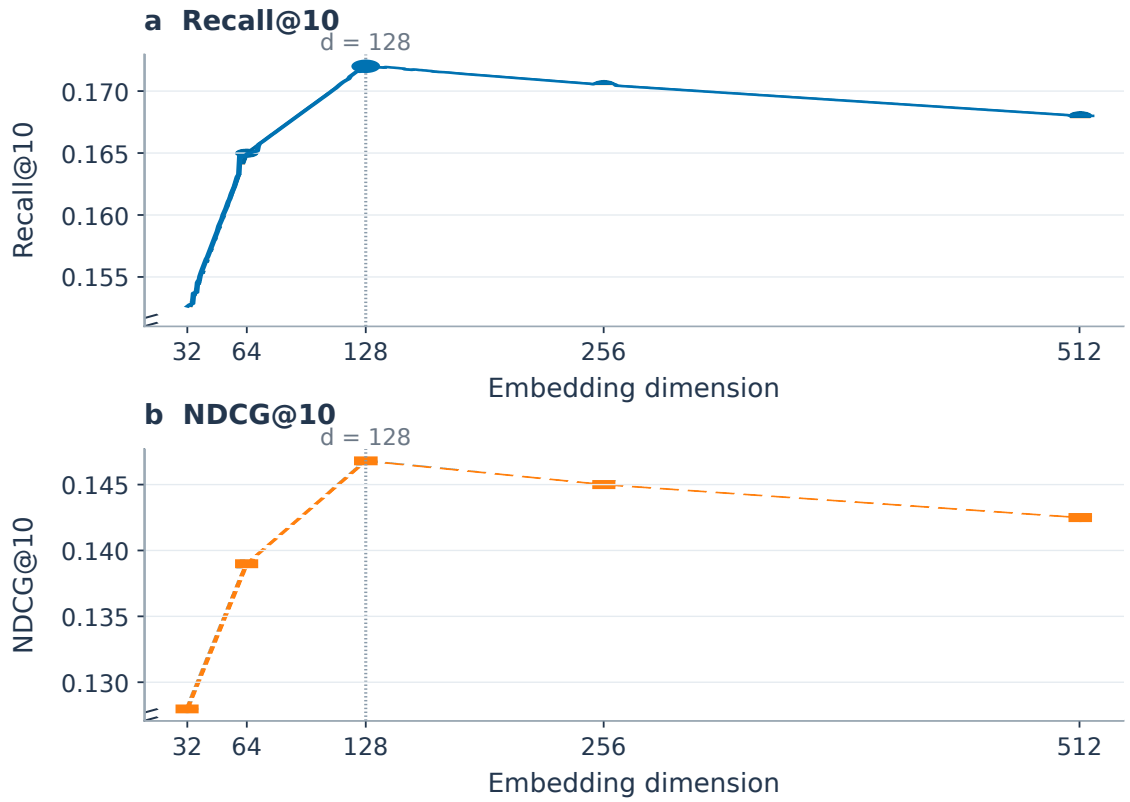

Fig. 4 Ranking performance at the tested embedding dimensions. Panel (a) shows Recall@10 and panel (b) shows NDCG@10. The dimension axis is linear and the vertical axes use restricted, explicitly labeled ranges. The vertical reference identifies the selected dimension of 128

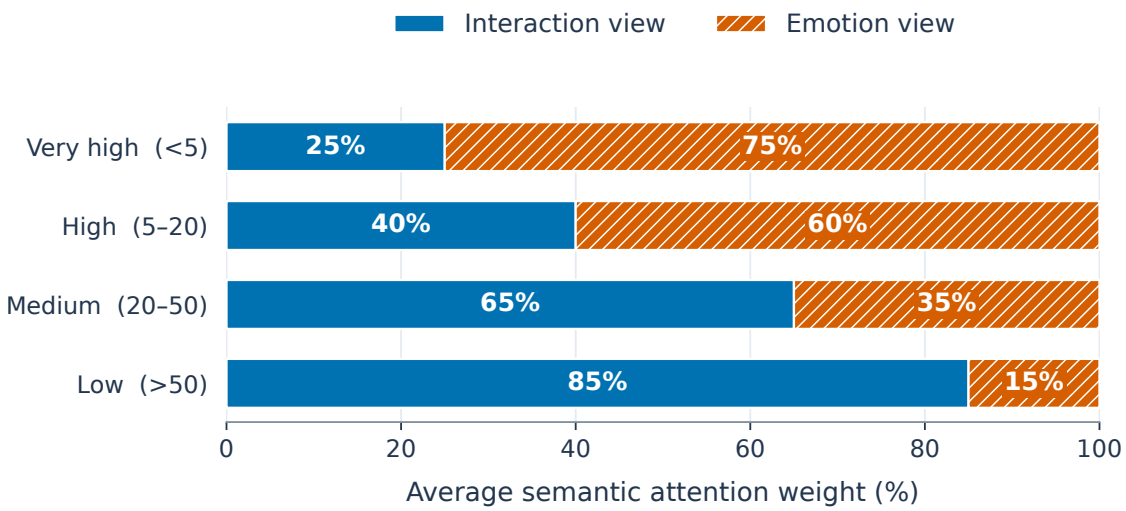

Fig. 5 Average interaction-view and emotion-view weights by history-sparsity group. Direct labels show the reported percentages, which sum to 100% within each group. Group labels follow the reported interaction-count intervals

behavior; they do not measure emotional intensity or establish a causal explanation of listener preference.

5.5 Inference for Sparse Users and Unseen Tracks

Figure 6 separates two inference routes. Retained context can connect a few-shot user to tracks through an affective category. A track absent from training instead obtains its representation from the trained content encoder. Neither route assumes access to the interaction being predicted.

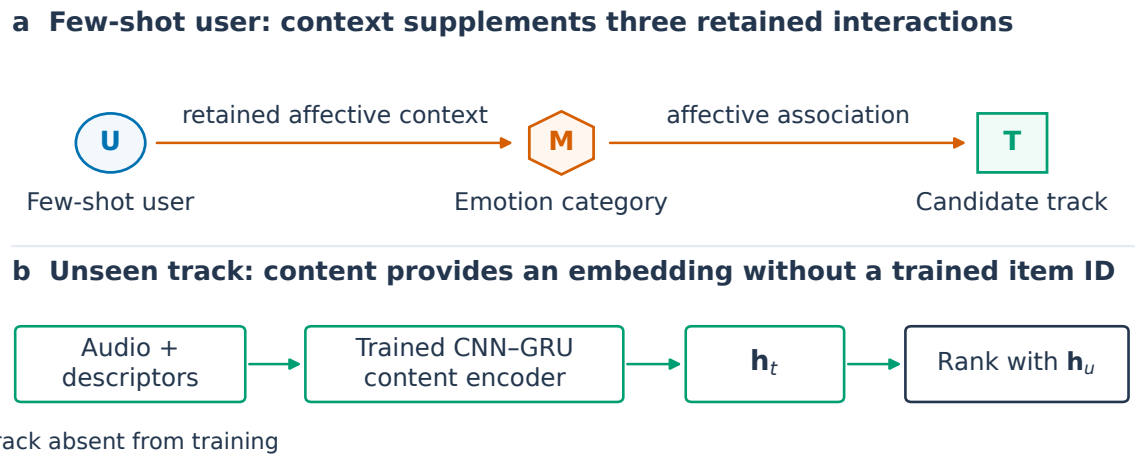

Fig. 6 Complementary inference routes. (a) Retained affective context supplements a three-interaction user profile. (b) An unseen track is encoded from available audio and descriptors without optimizing a new track identifier. The diagram illustrates the mechanism, rather than a dataset-specific listening event

An affective match need not always imply user satisfaction. A listener may seek music that maintains a mood or music intended to change it; distinguishing these intentions requires richer contextual observations than those evaluated here.

5.6 Representation Visualization and Training Behavior

Figure 7 combines the supplied t-SNE projection [21] of sampled track representations with a representative training trajectory. The projection displays local groups associated with happy, sad, and energetic tags. The training loss decreases while validation Recall@20 rises and subsequently fluctuates around a higher level.

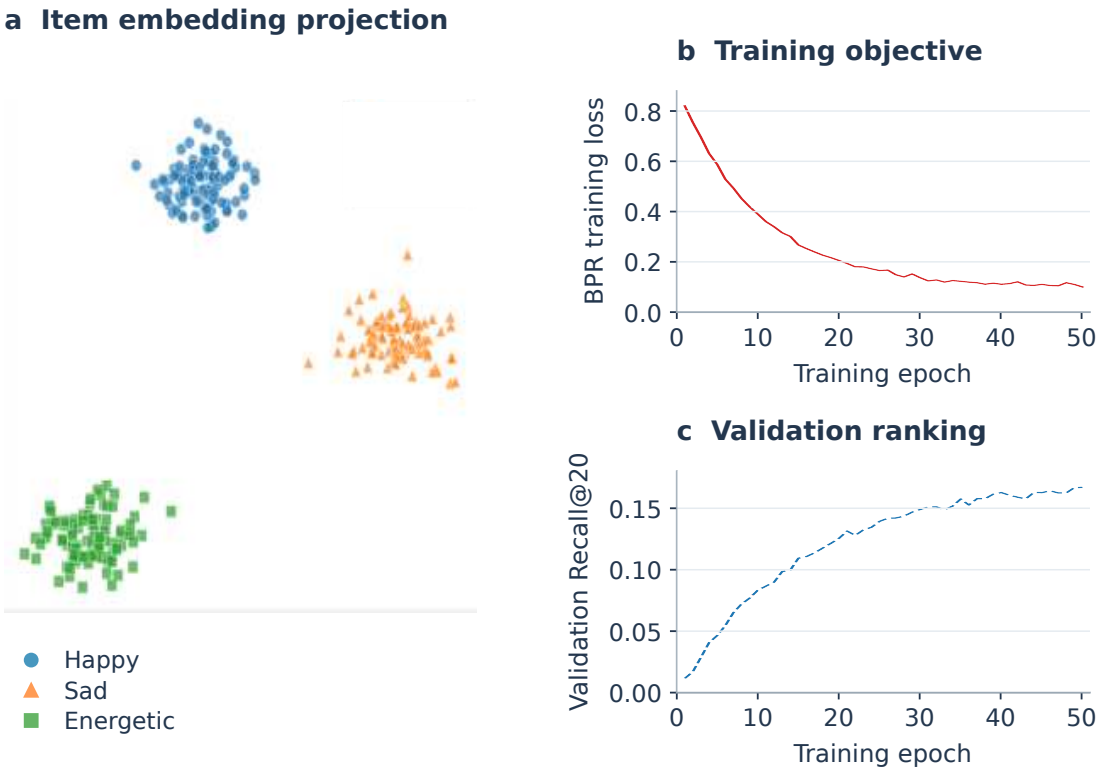

Fig. 7 Qualitative representation and training summaries. (a) The supplied item-embedding projection is colored by affective tag. (b) The representative BPR training loss and (c) validation Recall@20 are displayed on separate axes. These illustrations do not provide run-level uncertainty or a matched comparison of convergence across ablations

The projection is a qualitative view of a low-dimensional embedding and does not establish global separation in the original feature space. Likewise, the trajectories show optimization progress but cannot attribute faster convergence to emotion nodes without corresponding ablation trajectories.

6 Discussion

6.1 Complementary Roles of Content and Affective Context

The reported comparisons suggest that the two auxiliary information sources address different limitations of collaborative recommendation. Content-derived embeddings supply a representation for tracks without training interactions. Affective relations supply an additional semantic route when a user profile contains few observations. Conditioning their aggregation on history sparsity allows their relative contribution to vary across users rather than remaining fixed.

This coordination is the main distinction from treating affective classification, semantic retrieval, or knowledge-enhanced preference modeling as separate tasks [11–13]. The results concern top-K ranking on one processed listening dataset. They do not demonstrate that the inferred affective associations reflect a listener's experienced mood, or that offline gains translate directly to satisfaction in a deployed service.

6.2 Evaluation Scope and Reproducibility

EAI-HGN has the highest reported mean scores among the five evaluated baselines. These are established reference methods, so the comparison does not support a claim of superiority over all contemporary recommenders. Additional content-matched heterogeneous and multimodal methods would help separate improvements due to the learning architecture from those due to access to acoustic or contextual information.

The evaluation also depends on the definition of available information. The three-interaction user condition is few-shot adaptation, and the item condition is an identity holdout rather than a chronological deployment simulation. Target-event hashtags and target interactions must be excluded from the context used to predict those events. Exact reproduction additionally requires the split identifiers, candidate-pool construction, preprocessing rules, emotion mapping, sparsity-indicator definition, and baseline inference settings. The present aggregate report does not provide all of these details or the matched recording coverage.

Five-run averages alone do not quantify uncertainty. Run-level results, standard deviations or confidence intervals, and paired comparisons would be needed to determine how stable the observed differences are. The absence of those quantities limits the conclusions to the reported means. The same restriction applies to the ablation scores and attention-group summaries.

6.3 Limitations and Further Work

The processed subset is derived from listening events collected in 2014. Generalization to other populations, current catalogs, and different platforms requires further evaluation. A second dataset and chronological splits would provide complementary tests of transfer and deployment relevance.

Affective categories derived from hashtags can be ambiguous and culturally dependent. Their noise and the choice of 15 categories have not been systematically varied here. Future work should examine robust labeling, alternative emotion vocabularies, and the distinction between mood-congruent listening and attempts to change mood.

The acoustic branch requires matched recordings and adds preprocessing and inference costs. Recording provenance, coverage, and missing-content behavior need fuller documentation. Runtime, memory, and end-to-end latency also remain to be measured before real-time deployment can be assessed. Broader evaluation should consider diversity, novelty, calibration, and user satisfaction in addition to Recall and NDCG.

7 Conclusions

EAI-HGN combines a typed user–track–artist–emotion graph, an inductive acoustic encoder, and a sparsity-aware semantic gate for music recommendation. On the evaluated #nowplaying-RS subset, it achieves the highest mean Recall@10 and NDCG@10 among the compared baselines under general, few-shot user, and unseen-item conditions.

The reported ablations are consistent with complementary roles for content-based track inference and affective user context. The findings support coordinating these information sources while distinguishing few-shot adaptation from zero-history personalization. Their scope is limited by one processed dataset, incomplete reproducibility details, and unavailable run-level uncertainty. Future evaluation should use fully documented audio matching and splits, stronger input-matched baselines, and broader measures of recommendation quality.

Declarations

Funding. This research received no external funding.

Competing interests. The author declares no competing interests.

Ethics approval and consent to participate. Not applicable. The study used an existing music-listening dataset and did not recruit human participants or collect new participant data.

Consent for publication. Not applicable.

Clinical trial number. Not applicable.

Data availability. The public #nowplaying-RS dataset is documented by Poddar et al. [16]. The archived dataset is available at https://doi.org/10.5281/zenodo.3247476, and the associated publication is available at https://doi.org/10.5281/zenodo.1422565.

Author contributions. Haibin Zhang carried out the conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, and writing.

Use of generative AI and AI-assisted tools. ChatGPT (OpenAI) was used to assist with language refinement, manuscript organization, mathematical exposition, and figure formatting. This assistance did not generate new experimental data. The author takes responsibility for the scientific claims and the final text.

References

[1] Schedl M, Zamani H, Chen CW, Deldjoo Y, Elahi M. Current challenges and visions in music recommender systems research. International Journal of Multimedia Information Retrieval. 2018;7:95–116. https://doi.org/10.1007/s13735-018-0154-2.

[2] Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L. BPR: Bayesian Personalized Ranking from Implicit Feedback. In: Twenty-Fifth Conference on Uncertainty in Artificial Intelligence; 2009. p. 452–461. Available from: https://arxiv.org/abs/1205.2618.

[3] He X, Liao L, Zhang H, Nie L, Hu X, Chua TS. Neural Collaborative Filtering. In: 26th International Conference on World Wide Web; 2017. p. 173–182.

[4] Wu S, Sun F, Zhang W, Xie X, Cui B. Graph Neural Networks in Recommender Systems: A Survey. ACM Computing Surveys. 2023;55(5):1–37. https://doi.org/10.1145/3535101.

[5] Gao C, Zheng Y, Li N, Li Y, Qin Y, Piao J, et al. A Survey of Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions. ACM Transactions on Recommender Systems. 2023;1:1–51. https://doi.org/10.1145/3568022.

[6] Wang X, Ji H, Shi C, Wang B, Ye Y, Cui P, et al. Heterogeneous Graph Attention Network. In: The World Wide Web Conference; 2019. p. 2022–2032.

[7] Hu Z, Dong Y, Wang K, Sun Y. Heterogeneous Graph Transformer. In: Web Conference 2020; 2020. p. 2704–2710.

[8] van den Oord A, Dieleman S, Schrauwen B. Deep Content-Based Music Recommendation. In: Advances in Neural Information Processing Systems. vol. 26; 2013. p. 2643–2651. Available from: https://proceedings.neurips.cc/paper_files/paper/2013/hash/b3ba8f1bee1238a2f37603d90b58898d-Abstract.html.

[9] Hamilton WL, Ying R, Leskovec J. Inductive Representation Learning on Large Graphs. In: Advances in Neural Information Processing Systems. vol. 30; 2017. p. 1024–1034. Available from: https://arxiv.org/abs/1706.02216.

[10] Juslin PN, Västfjäll D. Emotional responses to music: The need to consider underlying mechanisms. Behavioral and Brain Sciences. 2008;31:559–575. https://doi.org/10.1017/S0140525X08005293.

[11] Rumiantcev M. Ontology-Guided Multimodal Framework for Explainable Music Similarity and Recommendation. Big Data and Cognitive Computing. 2026;10:122. https://doi.org/10.3390/bdcc10040122.

[12] Liu Y. Music emotion classification based on deep neural network and its application in context aware recommendation system. Discover Artificial Intelligence. 2026;6:613. https://doi.org/10.1007/s44163-026-01270-8.

[13] Khaligh MM, Shayegan MJ. Personalized sequential recommendation via contrastive learning and knowledge graph embeddings. Discover Artificial Intelligence. 2025;5:367. https://doi.org/10.1007/s44163-025-00723-w.

[14] Wang X, He X, Wang M, Feng F, Chua TS. Neural Graph Collaborative Filtering. In: 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval; 2019. p. 165–174.

[15] He X, Deng K, Wang X, Li Y, Zhang Y, Wang M. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In: 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval; 2020. p. 639–648.

[16] Poddar A, Zangerle E, Yang YH. #nowplaying-RS: A New Benchmark Dataset for Building Context-Aware Music Recommender Systems. In: 15th Sound and Music Computing Conference. Limassol, Cyprus; 2018. Available from: https://doi.org/10.5281/zenodo.1422565.

[17] Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014. p. 1724–1734.

[18] Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems. 2002;20:422–446. https://doi.org/10.1145/582415.582418.

[19] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: International Conference on Learning Representations; 2015. Available from: https://arxiv.org/abs/1412.6980.

[20] Glorot X, Bengio Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In: Thirteenth International Conference on Artificial Intelligence and Statistics. vol. 9 of Proceedings of Machine Learning Research; 2010. p. 249–256. Available from: https://proceedings.mlr.press/v9/glorot10a.html.

[21] van der Maaten L, Hinton G. Visualizing Data Using t-SNE. Journal of Machine Learning Research. 2008;9:2579–2605.