ManyMusic: An Open-access Music Audio Dataset for Human Experiments on Musical Emotions

Seung-Goo Kim[1,*], Pablo Alonso-Jiménez[2], Till Bechtloff[3,†], Daniela Sammler[1,4] and Dmitry Bogdanov[2]

[1] *Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany*
[2] *Pompeu Fabra University, Barcelona, Spain*
[3] *University of Music Karlsruhe, Karlsruhe, Germany*
[4] *Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany*

Abstract

Psychological and neuroscientific research on music-evoked emotions has been limited by certain stimulus characteristics. Typically, the stimuli are either artificially manipulated, unavailable for sharing due to copyright restrictions, sparsely sampled from narrow musical genres, or skewed by experimenters' biases. This paper presents empirical evidence that carefully curated subsets of a large-scale open dataset are comparable to well-matched commercial music in terms of various subjective ratings, including liking and feeling moved. In addition, we demonstrate the potential of a generative music service for use in human experiments. Here, we publish an open music audio dataset, curated through music information retrieval (MIR) model predictions and human validation. We believe this dataset will be highly beneficial for empirical music research and MIR research.

Keywords

music-evoked emotions, music audio dataset, affective experience, neurophysiological experiments, open science

1. Introduction

1.1. Motivations

Music can evoke strong emotions in human listeners, which may explain its ubiquity across all known human communities [1]. While previous psychological studies have pointed out broad directions [2], many details remain unclear regarding which specific aspects of music elicit such intense emotions. One major challenge is that many studies have used limited stimuli. These are typically either artificially manipulated, restricted from sharing due to copyright, too brief ($\leq$10 sec) to evoke deep emotions, sparsely sampled from a narrow range of musical genres (e.g., only Western classical), or biased by the experimenters' intuitions.

Recently, in cognitive neuroscience, the approach of extensively sampling neurophysiological responses to a vast amount of "real-world stimuli"—as compared to experimenter-manipulated stimuli—has gained popularity [3, 4, 5]. This approach enables the construction of non-linear models that directly link multivariate musical features with multivariate neural and behavioral responses, addressing the complexity of affective experiences evoked by real-world music. To bring this exciting development of extensive sampling into the field of music research, we need a music audio dataset that: (a) is openly accessible for research; (b) evokes intense emotions in many individuals; (c) better represents the diversity of real-world music in terms of genres and emotional content; (d) is large in scale for generalizability; (e) includes full-length tracks to capture the temporal unfolding of emotions; and (f) is validated by independent human raters. These qualities are also beneficial for general emotion research using music [6], as well as the growing interest at the intersection of neuroimaging and MIR [7, 8, 9, 10, 11].

HCMIR25: 3rd Workshop on Human-Centric Music Information Research, September 20th, 2025, Daejeon, Korea

[*] Corresponding author

[†] Undergraduate researcher

seung-goo.kim@ae.mpg.de (S. Kim)
https://seunggookim.github.io/ (S. Kim)
0000-0003-0558-8547 (S. Kim)

Table 1
Representative previous datasets.

Discipline	Dataset name	#Tracks	Duration	#Genres	Source	Accessibility
Psychology	Film Soundtrack [12]	360	10–60 sec	1	Commercial	Unclear
Psychology	EMMA [6]	817	2-7 min	7	Commercial	Copyrighted
Psychology	MUSIFEAST-17 [16]	356	30 sec	17	Commercial	Open-access
MIR	GTZAN [17]	1,000	30 sec	10	Commercial	Unclear
MIR	RWC [13]	215	3-7 min	10	Commissioned	On-request
MIR	FMA [14]	106,000	0-30 min	161	Community	Open-access
MIR	MTG-Jamendo [15]	55,094	0-30 min	95	Community	Open-access

1.2. Previous datasets and our contributions

Table 1 summarizes representative previous datasets from different disciplines. In music psychology, the Film Soundtracks dataset [12] and the EMMA dataset [6] are well-known. While these sets are well-validated with many human participants, either their length ($\leq$ 60 sec) or accessibility (copyrighted) is severely limited. On the other hand, in music information retrieval (MIR) research, several datasets exist, including RWC [13], FMA [14], and MTG-Jamendo [15]. These datasets are open-access, full-length, and large in scale. However, the published tracks have been noted to vary greatly in audio and musical quality.

In continuation of previous efforts [13, 14, 15], the current paper presents: (a) a systematic approach to data curation, (b) curated tracks that are emotionally evocative to human listeners, and (c) human evaluations compared with reference sets[1].

2. Dataset

Our proposed dataset *ManyMusic* is built upon the MTG-Jamendo Dataset due to its availability under permissive licenses (CC0 or CC-BY-SA) and content quality (aiming licensed, royalty-free music business for commercial use) and scale (over 55,094 full tracks in 320 kbps MP3 format). The dataset creation process is divided into three stages (further details available in our data repository[2]):

Algorithmic filtering. We discard potentially problematic tracks based on audio analysis and the existing metadata. We excluded tracks based on (a) duration (<3 or >7 minutes[3]), (b) integrated loudness (LUFS; <5th or >95th percentiles), (c) false stereo, (d) clipped samples per minute (> 90th percentile), (e) an activation lower than 0.1 for all genre classes according to the DiscogsEffnet music style classification model [18], (f) denylisted MTG-Jamendo tags (`xmas`, `christmas`, `advertising`, `presentation`, `background`, `backgrounds`, `corporate`, `commercial`, `motivations`) and genre classes (`Non-Music`, `Chiptune`[4]). The process resulted in a subset of 24,903 out of 55,094 (45.2%).

Sampling based on genre and affect prediction. Our goal is to create a diverse dataset in terms of musical genre and affect. To achieve uniform genre distribution, we created genre subsets containing tracks with activation higher than 0.1 for the 12 genres predicted by the DiscogsEffnet model [18]. For each genre subset, we predicted Arousal and Valence (AV) activations using Essentia's EmoMusic regression model based on MusiCNN [19], which produces embeddings with a rate of 0.33Hz. Then, we computed 3 k-means clusters, relying on dynamic time warping to normalize the duration of the embeddings. For each centroid, we sampled the 170 nearest neighbors, resulting in a total of 2,559 tracks, which is referred to as *Jamendo-A*.

Human curation. Finally, we conducted a human curation process to discard problematic tracks with 19 annotators, including authors and colleagues, 90% of whom had experience in music performance,

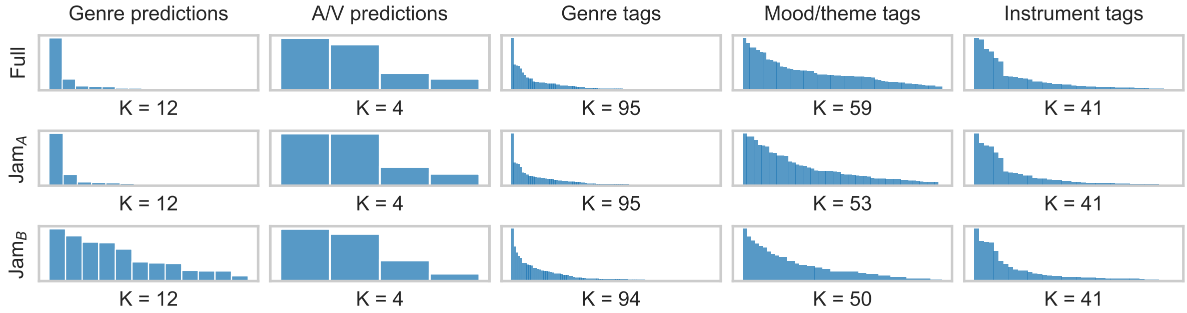
[1] All audio and human datasets are available for research purposes at https://manymusic.net/stim/

[2] https://github.com/manymusic/manymusic-mtg-jamendo-stim-scripts

[3] This range was selected with functional magnetic resonance imaging (fMRI) in mind. fMRI relies on slow neurophysiological activity with a latency of 4–6 seconds. Thus, a minimum of 3 minutes is required to observe sufficient temporal evolution.

[4] This was to avoid over-representation of this subgenre within Electronics.

Figure 1: Distribution of model predictions and original tags for the Full dataset and the Jamendo-A and Jamendo-B subsets. Arousal and valence (A/V) predictions are grouped in the four positive/negative A/V quadrants. K denotes the total number of unique classes in each case. See https://manymusic.net/stim/ for more information.



producing, and/or composition. Our pool of 2,559 candidate tracks was divided into 13 chunks of 200 tracks, each of which received categorical ratings from three annotators (on average, one annotator rated 410 tracks). A total of 19 annotators (including the authors) were invited to the annotation. The chunks were assigned so that no chunk was rated by the same set of three annotators. For the annotation process, we developed an interface that plays a loudness-normalized audio track and provides navigation capabilities[5]. Each annotator was asked to choose only one from the following options to the best of their knowledge: `All Good`, `Bad Audio`, `Not Emotionally Conveying`, `Explicit Content`, `Copyrighted Content`, or `Not Good for Other Reasons`. Hereafter, the set of tracks that received `All Good` from all three annotators is referred to as *Jamendo-B* (N = 1,129 out of 2,559; 44.1%). The distributions of model predictions and tags are shown in Fig. 1.

3. Evaluation

3.1. Experiment design

The aim of the experiment was to assess subjective emotional responses, quality assessments, and perceived emotions of our sets (*Jamendo-A*, *Jamendo-B*) using a musically experienced population.

Control stimuli. For comparison, three referential sources (*Self*, *Spotify*, *Suno*) were also used. *Self* tracks were participant-selected favorite tracks. *Spotify* tracks (N = 2,100) were 30-second audio previews of commercial music published in [19]; thus, the highest quality ratings were expected. *Suno* tracks (N = 1,280) were generated using Suno's v4.5 model[6] with generic prompts (e.g., 'Hungarian, mood:angry, Classical, Instrumental'; a full list of prompts can be found in our data repository). Except for *Self*, 100 tracks were sampled with matched musical genres and the presence of vocals across sources based on Essentia's Discogs-Effnet model predictions.

Procedure. An online experiment was conducted as follows. Participants first selected three of their favorite tracks using the Apple MusicKit API[7]. They then rated 30-second previews of the three self-selected tracks[8] and 30-second excerpts of 20 experimenter-selected tracks on six scales: "Professionalism", "Familiarity", "Perceived Valence", "Perceived Arousal", "Liking", and "Being Moved" (rationales underlying the selected scales can be found on our dataset webpage[9]). While we are ultimately interested in evoked emotions, the distinction between perceived and evoked emotions during music listening [20] can be inconsistently interpreted by many participants. Thus, for consistency, we asked

[5] We used https://wavesurfer.xyz/

[6] https://suno.com/

[7] https://developer.apple.com/musickit/

[8] Note that the small number of *Self* tracks was to estimate a subjective reference point, representing each participant's presumably highest ratings of Familiarity, Being Moved, Professionalism, and Liking, which are not necessarily the nominal maximum of one.

[9] https://manymusic.net/stim/plots_bhv.html

Table 2
Rating items and labels displayed to participants.

Scale	Item	Label at +1	Label at -1
Professionalism	*This music sounds …*	Professional	Unprofessional
Familiarity	*This music sounds …*	Familiar	Unfamiliar
Perceived Valence	*This music expresses …*	Joyful, Amusing, or Other Positive Feelings	Melancholic, Agitating, or Other Negative Feelings
Perceived Arousal	*This music is …*	Exciting, Stimulating, or Energizing	Calming, Soothing, or Relaxing
Liking	*I do …*	Like it	Dislike it
Being Moved	*I feel …*	Moved, Touched, or Engaged	Bored, Disgusted, or Disengaged

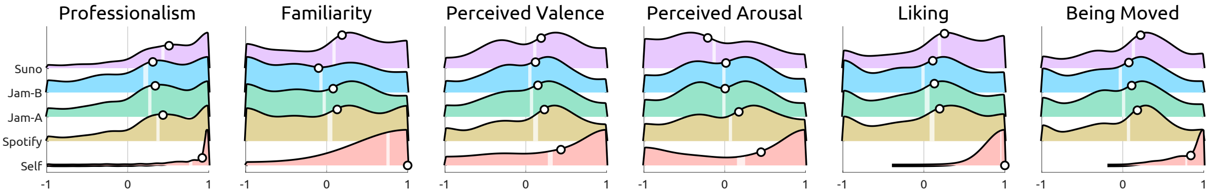


Figure 2: Ridgeline plots (190 participants, 4,600 ratings). Mean kernel bandwidth was 0.233 ± 0.197. White vertical bars denote the 95% confidence intervals of the means. White circles denote medians.

participants to report perceived emotions (i.e., emotions expressed by the musical piece rather than felt by the listener). Each scale was rated using a slider with 201 discrete levels, similar to the Visual Analog Scale, which has often been used for self-reporting of subjective feelings [21], because a more precise resolution of scales better supports the Gaussian noise assumption. The sampled values were rescaled between -1 and +1. Slider labels are shown in Table 2. The order (top to bottom) and direction (left to right) of scales were counterbalanced across participants. On average, experimenter-selected tracks were rated by 10.4 participants. After the ratings, participants completed a series of questionnaires, including two subscales (Musical Training and Emotions) of Goldsmiths Musical Sophistication Index (Gold-MSI) [22] and questions on preferred musical genres. All procedures were approved by the Ethics Council of the Max Planck Society.

Participants. Via an online experiment platform, Prolific[10], 233 participants completed the tasks. Participants were compensated at a rate of 9 GBP per hour. After excluding those who displayed invalid behaviors (e.g., negative liking for self-selected excerpts; "straight-lining", i.e., standard deviation of liking across all excerpts < 0.1), a total of 4,600 ratings from 190 unique participants were analyzed. The modal age was 21 years, with a range from 18 to 40. The sample consisted of 69% males, 73% white individuals, 88% born in the US or UK, and 93% residing in the US or UK. At least two years of musical training was required for participation. The modal percentiles of the Gold-MSI subscales were 79% for Musical Training and 75% for Musical Emotions.

3.2. Results

The distribution of sample ratings is shown in Fig. 2. As expected, self-selected tracks were rated significantly higher in Liking, Being Moved, Professionalism, and Familiarity compared to any of the experimenter-selected tracks (two-sample t-test, $df = 4591$; $19.70 \leq T \leq 34.52, 0.51 \leq \beta \leq 0.92, P < 10^{-83}$). This pattern also held for Perceived Valence ($T = 6.25, \beta = 0.23, P < 10^{-10}$) and Perceived Arousal ($T = 3.97, \beta = 0.18, P < 10^{-5}$), suggesting that many participants favored music that conveyed joy and energy over other emotional content.

For the experimenter-selected tracks, mixed-effects linear models were used to estimate fixed effects of intra-subject factors and random effects of inter-subject factors [23]. The linear models were specified as: `Rating ∼ Source + Age + Sex + MusTrain + MusEmo + (1+IsThisMyGenre|SubjectId) + (1|TrackId)` where factors in parentheses denote subject-specific random effects. The analysis

[10] https://www.prolific.com/

Table 3
Estimated effect sizes with standard errors on rating scales [-1, 1]. Significant non-zero effects (two-tailed $P < 0.05$; Tukey's HSD-adjusted for contrasts, Bonferroni-corrected for multiple scales) are marked in bold.

Contrast	Professionalism	Familiarity	Valence	Arousal	Liking	Being Moved
Jam$_{\mathrm{A}}$ - Jam$_{\mathrm{B}}$	0.040 ± 0.03	0.023 ± 0.03	0.022 ± 0.03	0.035 ± 0.04	0.046 ± 0.03	0.043 ± 0.02
Jam$_{\mathrm{A}}$ - Spotify	**-0.107 ± 0.03**	-0.073 ± 0.03	-0.036 ± 0.04	-0.069 ± 0.05	-0.061 ± 0.03	-0.047 ± 0.03
Jam$_{\mathrm{A}}$ - Suno	**-0.185 ± 0.03**	**-0.133 ± 0.03**	-0.036 ± 0.04	0.128 ± 0.05	**-0.154 ± 0.03**	**-0.113 ± 0.03**
Jam$_{\mathrm{B}}$ - Spotify	**-0.147 ± 0.03**	-0.096 ± 0.03	-0.058 ± 0.04	-0.104 ± 0.05	**-0.107 ± 0.03**	**-0.090 ± 0.03**
Jam$_{\mathrm{B}}$ - Suno	**-0.224 ± 0.03**	**-0.156 ± 0.03**	-0.058 ± 0.04	0.093 ± 0.05	**-0.201 ± 0.03**	**-0.156 ± 0.03**
Spotify - Suno	-0.077 ± 0.03	-0.060 ± 0.03	0.000 ± 0.04	**0.197 ± 0.05**	-0.093 ± 0.03	-0.066 ± 0.03

revealed strong effects of source on Professionalism (one-way ANOVA, $df_1 = 3$, $df_2 = 3992$; $F = 23.50, \eta_p = 0.017, P < 10^{-14}$), Liking ($F = 16.25, \eta_p = 0.012, P < 10^{-9}$), and Being Moved ($F = 13.33, \eta_p = 0.010, P < 10^{-7}$), with weaker effects on Familiarity ($F = 9.67, \eta_p = 0.007, P < 10^{-5}$), Arousal ($F = 5.51, \eta_p = 0.004, P < 10^{-3}$), and a non-significant effect on Valence ($F = 1.19, \eta_p = 0.001, P = 0.311$). Pairwise t-tests adjusted using Tukey's Honest Significant Difference (HSD) and Bonferroni correction revealed that *Suno* was rated significantly higher than the Jamendo tracks for Liking, Being Moved, Professionalism, and Familiarity (Table 3). Also, *Spotify* was rated significantly higher than the Jamendo tracks in Professionalism. However, *Jamendo-A* was not rated significantly lower than *Spotify* in Liking, Being Moved, and Familiarity.

4. Discussion and Conclusions

The current paper presents a curated music audio dataset *ManyMusic* (i.e., *Jamendo-A*) for human experiments on musical emotions. *Jamendo-B*, which was based on full agreement among three annotators, was rated lower in Liking and Being Moved than all referential sources. This highlights that aesthetic judgment and emotional experience can vary significantly across individuals, even among those with comparable musical sophistication. Moreover, it raises concerns about the validity of human annotation on emotional scales when based on a small number of raters. This point is consistent with the study suggesting that 10–20 raters are needed to ensure sufficient consensus in ratings of felt emotions [6]. *Suno* may offer a copyright-free alternative to commercial music for research. However, legal and ethical concerns surrounding AI-generated music services have not yet been fully resolved [24]. Similar familiarity ratings between the Suno and Spotify samples suggest that the training set of the Suno model may have been highly popular and well-known to the general population.

Overall, we conclude that the cleaned subset of MTG-Jamendo (*Jamendo-A*) is a balanced choice, given its perceptual similarity to commercial music and its freedom from potential ethical issues. We believe this carefully curated, balanced dataset will be highly beneficial for empirical music research and MIR research that involve deeper subjective experiences of human listeners [25, 26, 27].

References

[1] S. A. Mehr, M. Singh, D. Knox, D. M. Ketter, D. Pickens-Jones, S. Atwood, C. Lucas, N. Jacoby, A. A. Egner, E. J. Hopkins, et al., Universality and diversity in human song, Science 366 (2019) eaax0868.

[2] P. N. Juslin, From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions, Physics of Life Reviews 10 (2013) 235–266.

[3] E. J. Allen, G. St-Yves, Y. Wu, J. L. Breedlove, J. S. Prince, L. T. Dowdle, M. Nau, B. Caron, F. Pestilli, I. Charest, et al., A massive 7t fMRI dataset to bridge cognitive neuroscience and artificial intelligence, Nature Neuroscience 25 (2022) 116–126.

[4] S. A. Nastase, Y.-F. Liu, H. Hillman, A. Zadbood, L. Hasenfratz, N. Keshavarzian, J. Chen, C. J. Honey, Y. Yeshurun, M. Regev, et al., The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension, Scientific Data 8 (2021) 250.

[5] M. N. Hebart, O. Contier, L. Teichmann, A. H. Rockter, C. Y. Zheng, A. Kidder, A. Corriveau, M. Vaziri-Pashkam, C. I. Baker, THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior, eLife 12 (2023) e82580.

[6] H. Strauss, J. Vigl, P.-O. Jacobsen, M. Bayer, F. Talamini, W. Vigl, E. Zangerle, M. Zentner, The emotion-to-music mapping atlas (EMMA): A systematically organized online database of emotionally evocative music excerpts, Behavior Research Methods 56 (2024) 3560–3577.

[7] T. I. Denk, Y. Takagi, T. Matsuyama, A. Agostinelli, T. Nakai, C. Frank, S. Nishimoto, Brain2music: Reconstructing music from human brain activity, 2023. arXiv:2307.11078.

[8] I. Daly, Neural decoding of music from the EEG, Scientific Reports 13 (2023) 624.

[9] V. K. Cheung, L. Okuma, K. Shibata, K. Tsukuda, M. Goto, S. Furuya, Decoding drums, instrumentals, vocals, and mixed sources in music using human brain activity with fMRI, in: ISMIR, volume 3, 2023, p. 4.

[10] B. Kaneshiro, J. P. Dmochowski, Neuroimaging methods for music information retrieval: Current findings and future prospects., in: ISMIR, 2015, pp. 538–544.

[11] P. Toiviainen, V. Alluri, E. Brattico, M. Wallentin, P. Vuust, Capturing the musical brain with lasso: Dynamic decoding of musical features from fmri data, NeuroImage 88 (2014) 170–180.

[12] T. Eerola, J. K. Vuoskoski, A comparison of the discrete and dimensional models of emotion in music, Psychology of Music 39 (2011) 18–49.

[13] M. Goto, H. Hashiguchi, T. Nishimura, R. Oka, RWC music database: Popular, classical and jazz music databases., in: ISMIR, volume 2, 2002, pp. 287–288.

[14] M. Defferrard, K. Benzi, P. Vandergheynst, X. Bresson, FMA: A dataset for music analysis, in: ISMIR, 2017. arXiv:1612.01840.

[15] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, X. Serra, The MTG-Jamendo dataset for automatic music tagging, in: Machine learning for music discovery workshop, international conference on machine learning (ICML 2019), 2019, pp. 1–3.

[16] H. A. van der Walle, W. Wu, E. H. Margulis, K. Jakubowski, MUSIFEAST-17: MUsic Stimuli for Imagination, Familiarity, Emotion, and Aesthetic STudies across 17 genres, Behavior Research Methods 57 (2025) 204.

[17] G. Tzanetakis, P. Cook, Musical genre classification of audio signals, IEEE Transactions on Speech and Audio Processing 10 (2002) 293–302. doi:10.1109/TSA.2002.800560.

[18] P. Alonso-Jiménez, X. Serra, D. Bogdanov, Music representation learning based on editorial metadata from Discogs., in: ISMIR, 2022, pp. 825–833.

[19] D. Bogdanov, X. Lizarraga Seijas, P. Alonso-Jiménez, X. Serra, MusAV: A dataset of relative arousal-valence annotations for validation of audio models, in: ISMIR, 2022.

[20] E. Schubert, Loved music can make a listener feel negative emotions, Musicae Scientiae 17 (2013) 11–26. doi:10.1177/1029864912461321.

[21] P. E. Bijur, W. Silver, E. J. Gallagher, Reliability of the visual analog scale for measurement of acute pain, Academic emergency medicine 8 (2001) 1153–1157.

[22] D. Müllensiefen, B. Gingras, J. Musil, L. Stewart, The musicality of non-musicians: An index for assessing musical sophistication in the general population, PLOS One 9 (2014) e89642.

[23] D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4, Journal of Statistical Software 67 (2015) 1–48. doi:10.18637/jss.v067.i01.

[24] V. Nayar, The ethics of ai generated music: A case study on suno ai, GRACE: Global Review of AI Community Ethics 3 (2025).

[25] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, D. Turnbull, Music emotion recognition: A state of the art review, in: Proc. ismir, volume 86, 2010, pp. 937–952.

[26] A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov, et al., Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound, arXiv preprint arXiv:2502.05139 (2025).

[27] M. Schedl, P. Knees, B. McFee, D. Bogdanov, Music Recommendation Systems: Techniques, Use Cases, and Challenges, Springer US, New York, NY, 2022, pp. 927–971.