Artist Considerations in Offline Evaluation of Music Recommender Systems

Gregor Meehan[1,*], Johan Pauwels[1]

[1]Queen Mary University of London, 327 Mile End Rd, London E1 4NS, United Kingdom

Abstract

Many modern research works in the field of music recommender systems (MRSs) evaluate performance by song ranking accuracy in offline data splits. Although this paradigm matches widely adopted methodology in the broader recommender systems research community, in this work we argue that it neglects key considerations relating to musical artists. In particular, we show that there are significant differences in the ability of MRSs to successfully predict songs by artists which are known to the user and to predict those by artists which the user has never listened to before. Through analysis of content-based, collaborative filtering, and hybrid MRSs in warm and cold settings, we illustrate that this discrepancy can lead to misleading conclusions about model capability, especially given that successful recommendations of new artists are particularly valuable to the MRS user experience and to producing fairer outcomes for less popular artists. To highlight this issue, we demonstrate that a simple heuristic method based only on personalized artist filtering can achieve the strongest performance according to standard evaluation protocol. We then describe a novel MRS evaluation scheme which accounts for a user's artist interaction history, allowing for more nuanced analysis of MRS predictive capability. Finally, we provide an illustrative example of how this method can be applied to MRSs which incorporate artist metadata.

Keywords

Music Recommender Systems, Artist Fairness, Recommender System Evaluation

1. Introduction

As the providers of musical content, artists are key stakeholders in digital music platforms [1]. While most public music interaction datasets include artist metadata (e.g. Last.fm listening histories [2, 3, 4] or streaming platform playlist data [5, 6]), there is no single strategy in existing music recommender system (MRS) research for incorporating this information into model design and evaluation. Some studies focus on artist-level recommendation [7, 8, 9, 10, 11] or similarity [12, 13, 14, 15, 16] as their primary MRS task. However, most MRS works address song-level recommendation: while some leverage artist metadata to learn improved representations of musical audio for content-based MRS [17, 18, 19], enrich song representations [20, 21], or add artist nodes to MRS knowledge graphs [22, 23, 24], many (e.g. [25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35]) do not consider artist information at all.

Furthermore, even the song-level MRS studies that integrate artist data into their model design or training regime rarely consider it explicitly in their evaluation protocols. In MRS research, as in recommender systems literature more generally [36], the prevalent paradigm for model evaluation is measuring accuracy in offline scenarios; other than a handful of works which conduct user studies [27, 29], this is true for all of the papers referenced above. This method is widely adopted because it facilitates easier comparison of different MRS methods, especially in academic research settings where access to resources or large user populations for user studies or online testing may be limited. In song-level MRSs, offline evaluation involves splitting either songs or user-song interactions into disjoint training, validation, and test sets. After training, models are evaluated by their ability to rank songs which a user might be interested in, with metrics such as recall and precision calculated based on the held-out interactions from the validation or test set. This data splitting is typically done entirely at random, although some works take a temporal approach [37], e.g. using the most recent

MuRS 2025: 3rd Music Recommender Systems Workshop, September 22nd, 2025, Prague, Czech Republic

[*]Corresponding author.

gregor.meehan@qmul.ac.uk (G. Meehan); j.pauwels@qmul.ac.uk (J. Pauwels)
0009-0007-2619-9299 (G. Meehan); 0000-0002-5805-7144 (J. Pauwels)

10% of interaction data as a test set [31]. Artist-related data leakage during this process has long been recognized as a potential pitfall in MRS [7, 38] as well as in other music information retrieval tasks such as genre classification [39, 40]. However, with a few exceptions [8, 41], song-level MRS studies do not take artist information into account when creating these data splits or in calculating ranking accuracy.

In this paper, we argue that this artist-agnostic approach to MRS evaluation is flawed and has significant implications for measurement of MRS performance. Our motivation for this work stems from an unexpected finding in our recent study [19]: in that paper, we evaluate several contrastive pre-training regimes for music representation learning with weak supervision from metadata in the Melon Playlist Dataset [6]. When evaluating downstream MRS performance via nearest neighbor-based playlist continuation, we find that models trained using artist co-occurrence (Artist CO) consistently achieve the highest recommendation accuracy. In particular, they surpass competing models trained directly using playlist co-occurrence (Playlist CO), where positive contrastive pairs consist of songs which share a playlist. This result contradicts the naive expectation that Playlist CO models would be better at playlist continuation because their training regime aligns most closely with this task,[1] pointing instead to a wider issue in the evaluation setup.

In this study, we generalize these results on the Music4All-Onion dataset [42], expanding our analysis to include collaborative filtering and hybrid MRSs. We show that, in the standard evaluation scenario, the vast majority of the correct suggestions by these MRSs come from artists that the user has already listened to. This can lead to misleading conclusions on the system's effectiveness, especially given the value of novelty and diversity in MRSs [43, 44, 45]. In particular, as noted by van den Oord et al. in a seminal early work in deep content-based MRS [25], *'recommending songs by artists that the user is known to enjoy is not particularly useful'*.[2] To emphasize the severity of the issue, we show that a simple heuristic approach based on personalized artist filtering outperforms state-of-the-art methods in standard accuracy metrics in both cold and warm scenarios. We discuss the implications of these findings on MRS fairness from the artist perspective, and propose a novel MRS evaluation framework to ensure robustness against this issue. Finally, we show the value of this framework by illustrating how it aids in evaluation of an MRS which directly incorporates artist metadata.

2. Preliminaries

2.1. Problem Statement and Notation

In this paper, we focus on song-level MRSs, where the system is trained with interactions between a set of users $\mathcal{U}$ and a set of songs $\mathcal{S}$. The task of a song-level MRS is to predict preference scores $\mathbf{P}_{u,s}$ of user $u \in \mathcal{U}$ for each song $s \in \mathcal{S}$, with higher scores indicating higher interest of u in s. Each song s has an associated artist set a_s, with $\mathcal{A}$ the set of all artists. We assume that $\mathcal{S}$ has been divided into warm and cold subsets $\mathcal{S}^w$ and $\mathcal{S}^c$, with only the warm songs appearing in the training data. Each $u \in \mathcal{U}$ has a set of historical interacted items $\mathcal{S}_u \subset \mathcal{S}^w$ and interacted artists $\mathcal{A}_u = \bigcup_{s \in \mathcal{S}_u} a_s$. We refer to artists $a \in \mathcal{A}_u$ as u's *known artists*, while all other artists in $\mathcal{A} \setminus \mathcal{A}_u$ are *unknown artists* for u. While in this work we let $\mathcal{S}_u$ be all songs a user has interacted with, our insights also hold for other forms of interaction data, such as individual listening sessions or playlists.

We define the popularity $\mathrm{pop}(s)$ of song s as the number of users that have listened to s. We measure the affinity $\mathrm{aff}_u(a)$ of user u for artist a as the number of songs by a which u has listened to.

[1] One possible explanation for this finding is that, in contrast to songs by the same artist, the audio content of playlists is too diverse for the model to learn effective representations; however, the fact that Playlist CO and Artist CO achieve similarly strong performance in downstream tagging [17, 19] casts doubt on this hypothesis.

[2] We acknowledge that there are clear exceptions to this statement, especially given that repeat listens are common in music consumption [46]. In particular, it is less applicable in online evaluation [47] or in offline sequential or contextual MRSs. In these settings, the aim is often to suggest 'the right music at the right time' [48] based on both current session activity and long-term user preferences, which means that recommending a known song or artist can be much more valuable.

Table 1
Warm and cold split statistics for the used subset of Last.FM interaction data from Music4All-Onion.

	Full Data	Warm			Cold	
		Training	Validation	Test	Validation	Test
Users	8,807	8,807	8,281	8,302	8,448	8,447
Songs	43,886	35,108	28,333	28,295	4,389	4,389
Artists	8,627	7,790	6,685	6,619	2,515	2,513
Interactions	1,510,034	965,492	120,680	120,685	151,983	151,194

2.2. Data

We choose Music4All-Onion [42] (M4A-Onion) as our primary source of interaction data for this work, as it contains both Last.fm listening histories [4] and corresponding audio clips from Music4All [49] for generating audio content representations. Although M4A-Onion contains song features in many other data modes, we choose to only consider a single modality (namely, audio) to simplify the implementation of our chosen content-based and hybrid MRSs. The audio representation models considered are self-supervised method **MusicFM** (MFM) [50] and the **Playlist CO** (PCO) and **Artist CO** (ACO) models [17] trained using the Melon Playlist Dataset [6] as in [19] with the short-chunk CNN backbone [51].

To filter the interaction data, we first exclude skipped songs, where, as in [33], we consider a song skipped if the next song starts within 30 seconds. Then, following [32], we only consider interactions from 2018 and users from age 10 to 80. We then remove user-song interactions that only occur once and perform 5-core filtering on users and songs.

2.2.1. Data Splitting

We follow a standard fully random splitting procedure in our main analysis, and discuss potential artist-related modifications in Section 4.1. We also split the data to facilitate analysis of model performance in both cold and warm item scenarios. Following existing works which also tackle both problems [52], we first divide the songs $\mathcal{S}$ in an 80:20 ratio into warm ($\mathcal{S}^w$) and cold ($\mathcal{S}^c$) subsets, then further divide $\mathcal{S}^c$ 50:50 into cold validation and cold test songs. The interactions corresponding to these songs make up the cold validation and test sets. All user-song interactions relating to the songs in $\mathcal{S}^w$ are then split 80:10:10 into warm training, validation, and test sets. Statistics of these splits are in Table 1.

2.3. Models

To demonstrate the wide applicability of our concerns, we consider content-based, collaborative filtering, and hybrid models in our analysis. Our content-based method is **ItemKNN** [53], which calculates user preferences based on similarity in our audio content vectors. Our collaborative filtering models are BPR-based matrix factorization (**BPR-MF**) [54] and **XSimGCL** [55], a state-of-the-art graph-based approach. For our hybrid model, we adopt **CLCRec** [56], which addresses both the warm and cold scenario and has been used in several recent MRS studies [18, 32]. For ItemKNN, we test performance with the content vectors output from each audio representation model mentioned in Section 2.2; in some contexts, we refer to the ItemKNN results by the names of these audio representations for brevity (**MFM**, **PCO**, **ACO**). We implement CLCRec with MusicFM [50] feature inputs. We conduct hyperparameter tuning based on the ranges described in the original papers, and set the embedding dimension at 64 for all models.

2.3.1. Personalized Artist-Based Filtering

Aside from the above methods, we also introduce a **P**ersonalized **A**rtist-Based **F**iltering (**PAF**) heuristic for song-level recommendation based only on user-artist interaction history and item popularity. Given a user u and song s, PAF preference scores are defined as:

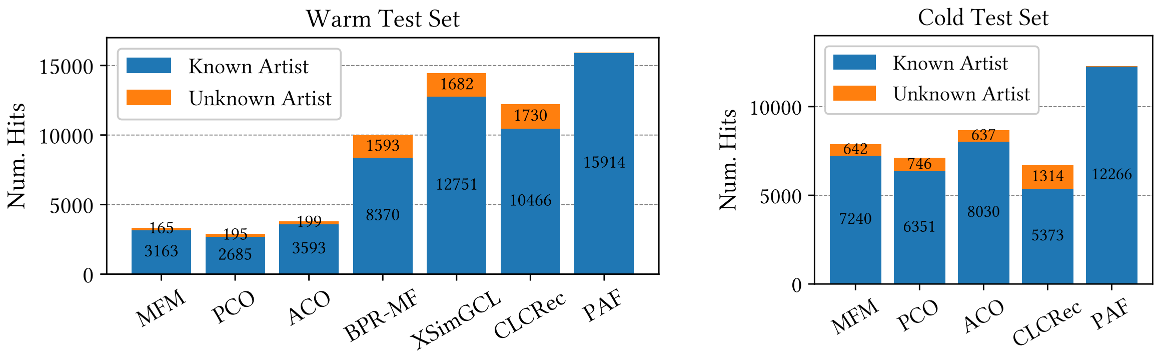

Figure 1: Total successful song recommendations at $k = 20$ across all users for each model on warm and cold test sets, split by whether the user has previously listened to the song's artist.

$$\mathbf{P}^{\mathrm{PAF}}_{u,s} = \begin{cases} 0, & \text{if } a_s \cap \mathscr{A}_u = \emptyset \\ \log(\mathrm{pop}(s)) \cdot \max_{a \in a_s} \mathrm{aff}_u(a), & \text{if } a_s \cap \mathscr{A}_u \neq \emptyset \text{ and } s \in \mathcal{S}^w \\ \max_{a \in a_s} \mathrm{aff}_u(a) + \mathrm{rand}(), & \text{if } a_s \cap \mathscr{A}_u \neq \emptyset \text{ and } s \in \mathcal{S}^c \end{cases} \tag{1}$$

For warm items, the preference score is the song's log-popularity multiplied by the user's affinity for the artist. For cold songs, popularity is not available, so we use a uniformly sampled random number between 0 and 1 as a ranking tie-breaker for songs with the same affinity. We note that PAF preference scores are zero for any song where the user has not listened to its artist, i.e. this approach will only suggest songs by artists that the user has already listened to.

3. Analysis

In this section, we examine the performance and behavior of our chosen models after overlaying artist information, and discuss the implications of our findings on MRS evaluation.

3.1. Hits

We first investigate overall model performance in Figure 1, displaying the total number of successful song recommendations (i.e., hits) split by whether the user has previously listened to that song's artist. We include full ranking metrics (namely, NDCG and Recall) below in Section 4.

We observe that, in both the warm and cold scenario and across all models, the vast majority of hits come from artists which are known to the user, i.e. songs s where $a_s \cap \mathscr{A}_u \neq \emptyset$. We can now explain how Artist CO outperforms Playlist CO in [19]: its additional hits come entirely in the known artist category, matching with its training objective of aligning audio representations of songs by the same artist. Meanwhile, Playlist CO is equal or superior for unknown artists, as would be expected given that its positive contrastive pairs are much more diverse.

There is an analogous trend in the collaborative filtering methods: XSimGCL achieves significantly more hits than BPR-MF, but almost all of the difference comes from known artists. Graph-based methods like XSimGCL are widely adopted due to their ability to exploit collaborative data more effectively, but these results suggest that, in a music context, this gain may come primarily from successfully suggesting songs by artists the user has previously interacted with. An investigation of the mechanism by which graph convolution amplifies these artist preferences is an avenue for future work.

Finally, we note that our purely metadata-based PAF method achieves the highest number of hits in both warm and cold contexts. However, by definition, none of these songs are by artists new to the user. We discuss the implications of this finding in further detail in Section 3.3 below.

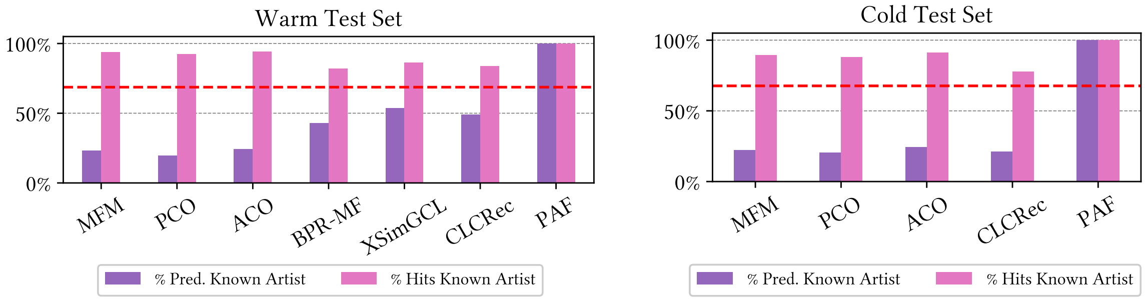

Figure 2: Percentage of predictions and hits from a user's known artists, averaged across all users. The dashed reference lines indicate the average percentage of songs by known artists in the warm and cold test sets.

3.2. Model Behavior

In Figure 2 we provide further insight into the artist-related predictive behavior of our models. In visualizing the proportion of hits and predictions by known artists, we can see that the known artist hit proportions for all models are above the proportion of known artists in the warm and cold test sets. In other words, songs by known artists are consistently over-represented in the model's successful recommendations. However, with the exception of PAF, the opposite is true of the model's predictions, which contain significantly lower proportions of songs by known artists.

The disparity between known artist proportions of predictions and hits is particularly prevalent in the ItemKNN models. This suggests that similarity in the chosen audio features cannot discriminate very effectively to capture user preferences, with successful predictions instead appearing to rely primarily on artist sonic coherence or other artifacts (e.g. similarities in mixing) which lead to high intra-artist similarity. This 'artist effect' is well-established in audio similarity applications [38], although its impact is exacerbated by the Artist CO training process.

The gap between known artist predictions and hits is less severe for the trained models, but still significant. XSimGCL's graph-based training also appears to increase its tendency to predict known artists, going some way to explain the finding in Section 3.1. In the cold data, CLCRec has similar prediction proportions to the ItemKNN models, but a lower proportion of hits by known artists. This suggests that the combination of collaborative training and content information can help to increase performance for unknown artists.

3.3. Discussion

The above results illustrate that there is a significant difference in the quality of recommendations between songs by artists known to the user and unknown to the user. There is some analogy between our scenario and analysis of popularity bias, where there are also two categories of items (namely, long-tail and popular items) with drastically different levels of performance. In such contexts, bias can be amplified by the fact that long-tail and popular items are recommended in a single top-k ranked list, as long-tail items are blocked from top ranking spots by popular items [57], i.e. they are underexposed to users [58]. However, the results in Figure 2 suggest that exposure is not the primary issue for songs by unknown artists, as these make up the majority of model predictions in most cases.

Instead, our findings reflect the intuition that correctly recommending songs by unknown artists is simply a much harder task than suggesting songs by artists a user is familiar with. As shown in Figure 2, over 65% of targets in the warm and cold test sets come from songs by known artists, but the precision of such recommendations will be significantly higher because the candidate pool is much smaller. This is clearly demonstrated by the fact that, by focusing only on songs by known artists, our simple PAF method has the most hits of any model by a wide margin. As noted in Section 1 and in [25], such metadata-based suggestions lack novelty and will potentially be less valuable to users; yet, according to standard evaluation protocol (cf. Table 2), PAF achieves the best results, showing that evaluating song-level MRSs by overall performance alone obfuscates important information about which recommendations are being made successfully. In Section 4 below we describe a simple adjustment to ranking-based evaluation which disentangles known and unknown artist performance, allowing for more nuanced analysis of MRS performance.

3.3.1. Artist Fairness

Aside from potentially degrading overall recommendation quality, MRS hits being dominated by known artists can also lead to inequitable outcomes for less popular artists. In previous interview-based studies [59, 60], artists describe the difficulties of reaching their intended audience via algorithmic suggestions. This may be due to lack of exposure, which previous works show is a key challenge for less popular artists in collaborative filtering contexts [61, 62]. However, a further concern is raised by our analysis above: if an MRS has poor accuracy for artists that a user has not listened to, then its ability to produce successful recommendations for artists with small listener bases will likely be limited.

Even if exposure is not an issue, and an artist's songs are recommended to many users, their audience will not grow unless those users are likely to enjoy their music. Determining an MRS's ability to successfully recommend songs by unknown artists is therefore an important step in ensuring that its suggestions are more fair. However, if model evaluation is based primarily on overall recommendation accuracy, then the 'best-performing' model may be one which mainly succeeds at recommending known artists, thereby harming less popular artists and reducing overall fairness.[3] Given that users are more satisfied with recommendations when they believe them to be fair [63] and that increasing fairness via bias mitigation methods can improve perceived recommendation quality [45], these concerns further motivate an alternative evaluation protocol which treats known and unknown artists separately.

4. Artist-Based Evaluation Protocol

4.1. Strategy

Based on the above insights, we now elucidate a evaluation framework which provides clearer understanding of the relative strengths and weaknesses of MRSs from the artist perspective. Given a user u, the standard evaluation protocol involves calculating accuracy metrics on a single top-k list, where songs s are ranked by their predicted preference scores $\mathbf{P}_{u,s}$. We propose to split this list into songs by artists known to u and artists unknown to u, and calculate ranking metrics on each list separately. This method is somewhat similar to evaluating model performance separately on warm and cold songs; however, these are global classifications, while known artists are specific to each user. By separating these two categories, we ensure that our ability to evaluate model capacity for unknown artist recommendations is not drowned out by the more straightforward recommendation of known artists.

We note that this evaluation strategy can be applied more effectively if it is taken into account at the data splitting stage, as this will ensure that there are sufficient examples in both categories for all users. Furthermore, if an overall evaluation approach is still required for the application at hand, this adjustment to data splitting will also ensure that unknown artist performance has a bigger influence on overall ranking metrics. We chose not to take this approach in this work so that the analysis in Section 3 would more accurately reflect the current standard methodology. As shown in Figure 2, about a third of holdout targets are from unknown artists, so even without this step there is still a statistically significant sample in this category. In some contexts, it may also be useful to consider a third category of artists which are unknown to all users, so that performance or fairness for brand new artists can also be evaluated. We leave this challenge to future work.

[3] We note that improving net unknown artist performance is not a sufficient condition to improve artist fairness; it may be that the model only improves in suggesting popular artists to users who have not listened to them yet, and still neglects less popular artists. However, these insights provide another dimension by which artist fairness may be analyzed in future work.

Table 2
Warm and cold test set NDCG@20 (N@20) and Recall@20 (R@20) for our models across different evaluation sets. The best result in each metric is bolded, and the second-best is underlined.

			ItemKNN			BPR-MF	XSimGCL	CLCRec	PAF
			MFM	PCO	ACO	-	-	MFM	-
Warm	Overall	N@20	0.0375	0.0324	0.0427	0.1048	0.1624	0.1301	**0.1697**
		R@20	0.0365	0.0333	0.0428	0.1157	0.1672	0.1397	**0.1874**
	Known Artist	N@20	0.1012	0.1045	0.1192	0.2021	**0.2586**	0.2243	0.2279
		R@20	0.1251	0.1300	0.1483	0.2765	**0.3284**	0.2958	0.3021
	Unknown Artist	N@20	0.0035	0.0045	0.0043	0.0455	0.0496	**0.0510**	0.0000
		R@20	0.0062	0.0076	0.0068	0.0694	0.0715	**0.0749**	0.0000
Cold	Overall	N@20	0.0783	0.0688	0.0861	-	-	0.0595	**0.1257**
		R@20	0.0655	0.0615	0.0763	-	-	0.0600	**0.1596**
	Known Artist	N@20	0.1936	0.1975	0.2209	-	-	**0.2216**	0.1644
		R@20	0.2029	0.2144	0.2412	-	-	**0.3143**	0.2479
	Unknown Artist	N@20	0.0125	0.0152	0.0129	-	-	**0.0256**	0.0000
		R@20	0.0188	0.0226	0.0194	-	-	**0.0392**	0.0000

4.2. Results

We display in Table 2 the resulting ranking metrics for warm and cold songs on the standard (Overall) ranking list, alongside our new Known and Unknown Artist splits. Looking first at the Overall results, we see that PAF scores highest, with an especially large margin on the cold test set. This fact alone illustrates the limitations of only considering standard overall splits, as PAF's ability to produce diverse and novel recommendations is severely limited. As discussed in Section 3.1, Artist CO appears to be superior to the other two audio representations for ItemKNN and to CLCRec in the cold data, while XSimGCL significantly outperforms BPR-MF and CLCRec in the warm setting.

However, a different story emerges when we examine the results of our new artist-based splits. First, we see that PAF is no longer the top-performing method when we restrict predictions to known artists, with XSimGCL and CLCRec surpassing it (in warm and cold settings respectively) once the lower-precision unknown artist predictions are removed from their top rankings. For unknown artists, Playlist CO now performs best for ItemKNN, and CLCRec beats out XSimGCL in the warm set, with the gap between BPR-MF and XSimGCL also narrowing significantly. In the cold set, the ranking order of models effectively flips from the Overall setting: CLCRec performs best by a 68% margin in NDCG@20, and Playlist CO moves from second-last to second place.

These results illustrate the value of our proposed evaluation protocol, and reveal that models which may be discounted based on overall performance are actually the most useful in the harder task of recommending songs by artists unfamiliar to a user. In the following section, we describe a brief case study for this approach, showing its applicability to methods directly incorporating artist information.

4.3. Example Use Case

Suppose we want to enhance our ItemKNN method for cold songs by leveraging artist metadata. Following previous works in artist similarity [12, 13, 16], we can generate an artist's content representation by averaging the feature vectors of their songs. Then for any song s we have its feature vector $\mathbf{x}_s$ and the corresponding content vector $\mathbf{x}_{a_s}$ for its artist(s). Standard ItemKNN uses cosine similarity $\text{sim}(\mathbf{x}_s, \mathbf{x}_t)$ between the feature vectors to measure the similarity $f(s,t)$ between two songs s and t; we can augment this method with artist information by calculating $f(s,t) = \text{sim}(\mathbf{x}_s, \mathbf{x}_t) + \alpha \cdot \text{sim}(\mathbf{x}_{a_s}, \mathbf{x}_{a_t})$, where α is a balance weight controlling the emphasis placed on artist similarity.

We plot the impact of varying this artist weight α in Figure 3. If we only consider overall results, adding artist similarity appears to significantly improve model performance. However, we can see from the other metrics that this benefit is limited to known artist predictions, with ranking quality

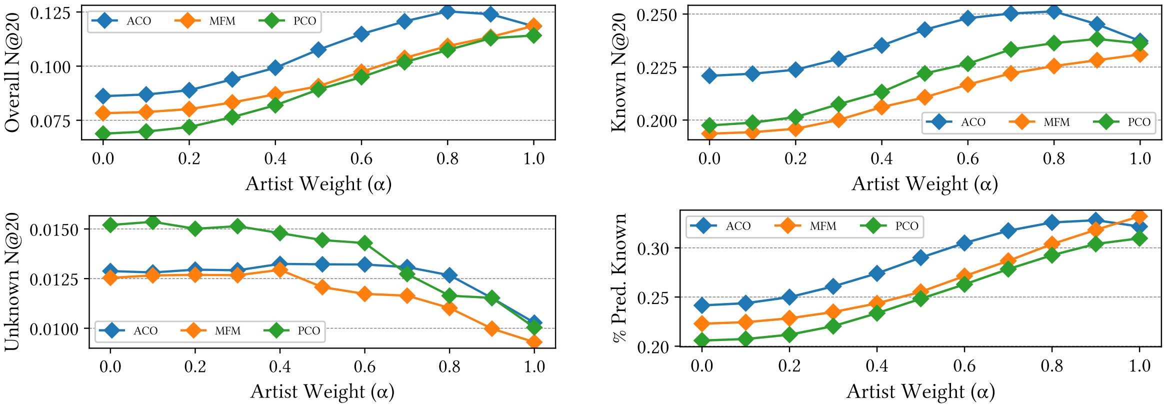

Figure 3: Results for ItemKNN in our audio feature vectors across various metrics on the cold test set, plotted against the weight given to artist similarity in the song-song similarity calculation.

for unknown artists degrading for higher values of α. This aligns with our discussion in Section 3.2 of the artist effect, i.e. that similarity in audio content vectors is a poor predictor of inter-artist similarity for MRS applications. Furthermore, this method increases the proportion of known artists in overall ranking lists; as illustrated by PAF in Section 3, this can artificially increase overall performance without providing significant additional value to the user. This example shows how measuring known and unknown artist performance separately can allow for more informed evaluation of novel MRS methods.

5. Conclusion

In this paper, we demonstrate that standard methods for offline recommender system evaluation have notable limitations in MRS applications. We show that MRSs across content-based, collaborative filtering, and hybrid paradigms exhibit significant differences in performance when recommending songs by known artists versus those by artists that a user has never previously listened to. Although this disparity is not surprising in itself, it brings into question the value of standard accuracy measurements dominated by much 'easier' known artist predictions, which may be less valuable to the user due to lack of novelty. To emphasize this point, we show that our PAF heuristic method achieves the best results according to standard evaluation procedure, despite only recommending a user's known artists. To address this issue, we propose a novel MRS evaluation strategy, where ranking metrics are calculated separately for each user on their target lists of songs by known and unknown artists. We show that this approach allows for more nuanced interpretation of model performance in warm and cold settings, and that it is particularly useful when integrating artist metadata into model design.

In future work we plan to build on these insights and explore how MRSs can be designed specifically to improve unknown artist performance. This may also allow for improved personalization of model suggestions: one possible approach, similar to calibration methods in popularity bias mitigation [64], would use the top candidate songs by known and unknown artists to produce a combined list of suggestions where the proportion of unknown artists matches a user's historical appetite for new artist discovery. We also acknowledge that our method is perhaps overly simplistic in its classification of artists as 'known' or 'unknown' to a user, and that the interaction characteristics of a user's most-listened artist and an artist they have heard only once may be very different. Future research will explore what further insight may be gained by investigating the relationship between user-artist affinity and model evaluation at different levels of familiarity.

Acknowledgments

This work was funded by UK Research and Innovation as part of the UKRI CDT in Artificial Intelligence and Music [grant number EP/S022694/1].

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] C. Bauer, M. Kholodylo, C. Strauss, Music recommender systems challenges and opportunities for non-superstar artists (2017).

[2] O. Celma, Music recommendation, in: Music recommendation and discovery: The long tail, long fail, and long play in the digital music space, Springer, 2010, pp. 43–85.

[3] M. Schedl, The lfm-1b dataset for music retrieval and recommendation, in: Proceedings of the 2016 ACM on international conference on multimedia retrieval, 2016, pp. 103–110.

[4] M. Schedl, S. Brandl, O. Lesota, E. Parada-Cabaleiro, D. Penz, N. Rekabsaz, Lfm-2b: A dataset of enriched music listening events for recommender systems research and fairness analysis, in: Proceedings of the 2022 Conf. on Human Information Interaction and Retrieval, 2022, pp. 337–341.

[5] C.-W. Chen, P. Lamere, M. Schedl, H. Zamani, Recsys challenge 2018: Automatic music playlist continuation, in: Proceedings of the 12th ACM Conf. on Recommender Systems, 2018, pp. 527–528.

[6] A. Ferraro, Y. Kim, S. Lee, B. Kim, N. Jo, S. Lim, S. Lim, J. Jang, S. Kim, X. Serra, D. Bogdanov, Melon playlist dataset: a public dataset for audio-based playlist generation and music tagging, CoRR abs/2102.00201 (2021). URL: https://arxiv.org/abs/2102.00201. arXiv:2102.00201.

[7] B. McFee, L. Barrington, G. Lanckriet, Learning content similarity for music recommendation, IEEE transactions on audio, speech, and language processing 20 (2012) 2207–2218.

[8] S. Oramas, O. Nieto, M. Sordo, X. Serra, A deep multimodal approach for cold-start music recommendation, in: Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems, DLRS 2017, Association for Computing Machinery, New York, NY, USA, 2017, p. 32–37.

[9] D. Kowald, M. Schedl, E. Lex, The unfairness of popularity bias in music recommendation: A reproducibility study, in: European conference on information retrieval, Springer, 2020, pp. 35–42.

[10] A. Trainor, D. Turnbull, Popularity degradation bias in local music recommendation, arXiv preprint arXiv:2309.11671 (2023).

[11] N. Bertram, J. Dunkel, R. Hermoso, I am all ears: Using open data and knowledge graph embeddings for music recommendations, Expert Systems with Applications 229 (2023) 120347.

[12] F. Korzeniowski, S. Oramas, F. Gouyon, Artist similarity with graph neural networks, arXiv preprint arXiv:2107.14541 (2021).

[13] F. Korzeniowski, S. Oramas, F. Gouyon, Artist similarity for everyone: A graph neural network approach, Transactions of the International Society for Music Information Retrieval (2022). doi:10.5334/tismir.143.

[14] S. Oramas, A. Ferraro, A. Sarasua, F. Gouyon, Talking to your recs: Multimodal embeddings for recommendation and retrieval (2024).

[15] F. Grötschla, L. Strässle, L. A. Lanzendörfer, R. Wattenhofer, Towards leveraging contrastively pretrained neural audio embeddings for recommender tasks, arXiv preprint arXiv:2409.09026 (2024).

[16] A. C. M. da Silva, D. F. Silva, R. M. Marcacini, Artist similarity based on heterogeneous graph neural networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024).

[17] P. Alonso-Jiménez, X. Favory, H. Foroughmand, G. Bourdalas, X. Serra, T. Lidy, D. Bogdanov, Pre-Training Strategies Using Contrastive Learning and Playlist Information for Music Classification and Similarity, 2023. URL: http://arxiv.org/abs/2304.12257.

[18] R. Salganik, X. Liu, Y. Ma, J. Kang, T.-S. Chua, Larp: Language audio relational pre-training for cold-start playlist continuation, arXiv preprint arXiv:2406.14333 (2024).

[19] G. Meehan, J. Pauwels, Evaluating contrastive methodologies for music representation learning using playlist data, in: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, pp. 1–5. doi:10.1109/ICASSP49660.2025.10888157.

[20] Q. Yang, S. Wang, D. Guo, D. Yu, Q. Xiao, D. Wang, C. Luo, Cascading multimodal feature enhanced contrast learning for music recommendation, in: 2024 IEEE International Conference on Data Mining (ICDM), IEEE, 2024, pp. 905–910.

[21] B. L. Pereira, P. D. V. Chaves, R. L. Santos, Efficient exploration and exploitation for sequential music recommendation, ACM Transactions on Recommender Systems 2 (2024) 1–23.

[22] H. Weng, J. Chen, D. Wang, X. Zhang, D. Yu, Graph-based attentive sequential model with metadata for music recommendation, IEEE Access 10 (2022) 108226–108240. doi:10.1109/ACCESS.2022.3213812.

[23] X. Cui, X. Qu, D. Li, Y. Yang, Y. Li, X. Zhang, Mkgcn: Multi-modal knowledge graph convolutional network for music recommender systems, Electronics 12 (2023). URL: https://www.mdpi.com/2079-9292/12/12/2688. doi:10.3390/electronics12122688.

[24] D. Wang, X. Zhang, Y. Yin, D. Yu, G. Xu, S. Deng, Multi-view enhanced graph attention network for session-based music recommendation, ACM Transactions on Information Systems 42 (2023) 1–30.

[25] A. Van den Oord, S. Dieleman, B. Schrauwen, Deep content-based music recommendation, Advances in neural information processing systems 26 (2013).

[26] A. Ferraro, X. Favory, K. Drossos, Y. Kim, D. Bogdanov, Enriched music representations with multiple cross-modal contrastive learning, IEEE Signal Processing Letters 28 (2021) 733–737.

[27] M. Pulis, J. Bajada, Siamese neural networks for content-based cold-start music recommendation., in: Proceedings of the 15th ACM conference on recommender systems, 2021, pp. 719–723.

[28] A. Saravanou, F. Tomasi, R. Mehrotra, M. Lalmas, Multi-task learning of graph-based inductive representations of music content., in: ISMIR, 2021, pp. 602–609.

[29] M. Park, K. Lee, Exploiting negative preference in content-based music recommendation with contrastive learning, in: Proceedings of the 16th ACM Conference on Recommender Systems, 2022, pp. 229–236.

[30] P. Magron, C. Févotte, Neural content-aware collaborative filtering for cold-start music recommendation, Data Mining and Knowledge Discovery 36 (2022) 1971–2005. URL: https://doi.org/10.1007/s10618-022-00859-8. doi:10.1007/s10618-022-00859-8.

[31] R. Borges, M. Queiroz, Audio-based sequential music recommendation, in: 2023 31st European Signal Processing Conference (EUSIPCO), IEEE, 2023, pp. 421–425.

[32] C. Ganhör, M. Moscati, A. Hausberger, S. Nawaz, M. Schedl, A multimodal single-branch embedding network for recommendation in cold-start and missing modality scenarios, in: Proceedings of the 18th ACM Conference on Recommender Systems, 2024, pp. 380–390.

[33] P. Seshadri, S. Shashaani, P. Knees, Enhancing sequential music recommendation with negative feedback-informed contrastive learning, in: Proceedings of the 18th ACM Conference on Recommender Systems, 2024, pp. 1028–1032.

[34] Y.-M. Tamm, A. Aljanaki, Comparative analysis of pretrained audio representations in music recommender systems, in: Proceedings of the 18th ACM Conference on Recommender Systems, 2024, pp. 934–938.

[35] M. Bevec, M. Tkalčič, M. Pesek, Hybrid music recommendation with graph neural networks, User Modeling and User-Adapted Interaction (2024) 1–38.

[36] E. Zangerle, C. Bauer, Evaluating recommender systems: survey and framework, ACM computing surveys 55 (2022) 1–38.

[37] R. Burke, Evaluating the dynamic properties of recommendation algorithms, in: Proceedings of the fourth ACM conference on Recommender systems, 2010, pp. 225–228.

[38] A. Flexer, D. Schnitzer, Effects of album and artist filters in audio similarity computed for very large music databases, Computer Music Journal 34 (2010) 20–28.

[39] E. Pampalk, A. Flexer, G. Widmer, et al., Improvements of audio-based music similarity and genre classificaton., in: ISMIR, volume 5, London, UK, 2005, pp. 634–637.

[40] I. Vatolkin, G. Rudolph, C. Weihs, Evaluation of album effect for feature selection in music genre recognition., in: ISMIR, 2015, pp. 169–175.

[41] A. Vall, M. Dorfer, H. Eghbal-Zadeh, M. Schedl, K. Burjorjee, G. Widmer, Feature-combination

hybrid recommender systems for automated music playlist continuation, User Modeling and User-Adapted Interaction 29 (2019) 527–572.

[42] M. Moscati, E. Parada-Cabaleiro, Y. Deldjoo, E. Zangerle, M. Schedl, Music4all-onion – a large-scale multi-faceted content-centric music recommendation dataset, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 4339–4343. doi:10.1145/3511808.3557656.

[43] Y. Deldjoo, M. Schedl, P. Knees, Content-driven music recommendation: Evolution, state of the art, and challenges, Computer Science Review 51 (2024) 100618. URL: https://www.sciencedirect.com/science/article/pii/S1574013724000029. doi:10.1016/j.cosrev.2024.100618.

[44] Y. Ping, Y. Li, J. Zhu, Beyond accuracy measures: the effect of diversity, novelty and serendipity in recommender systems on user engagement, Electronic Commerce Research (2024) 1–28.

[45] R. Ungruh, K. Dinnissen, A. Volk, M. S. Pera, H. Hauptmann, Putting popularity bias mitigation to the test: A user-centric evaluation in music recommenders, in: Proceedings of the 18th ACM Conference on Recommender Systems, 2024, pp. 169–178.

[46] M. Schedl, H. Zamani, C.-W. Chen, Y. Deldjoo, M. Elahi, Current challenges and visions in music recommender systems research, International Journal of Multimedia Information Retrieval 7 (2018) 95–116.

[47] W. Bendada, G. Salha-Galvan, T. Bouabça, T. Cazenave, A scalable framework for automatic playlist continuation on music streaming services, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 464–474.

[48] E. Liebman, M. Saar-Tsechansky, P. Stone, The Right Music at the Right Time: Adaptive Personalized Playlists Based on Sequence Modeling, MIS Quarterly 43 (2019) 765–786.

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

[50] M. Won, Y.-N. Hung, D. Le, A foundation model for music informatics, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2024, pp. 1226–1230.

[51] M. Won, S. Oramas, O. Nieto, F. Gouyon, X. Serra, Multimodal metric learning for tag-based music retrieval, in: Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 591–595.

[52] F. Huang, Z. Wang, X. Huang, Y. Qian, Z. Li, H. Chen, Aligning Distillation For Cold-start Item Recommendation, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1147–1157. doi:10.1145/3539618.3591732.

[53] B. Sarwar, G. Karypis, J. Konstan, J. Riedl, Item-based collaborative filtering recommendation algorithms, in: Proceedings of the 10th international conf. on World Wide Web, 2001, pp. 285–295.

[54] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, Bpr: Bayesian personalized ranking from implicit feedback, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI '09, AUAI Press, Arlington, Virginia, USA, 2009, p. 452–461.

[55] J. Yu, X. Xia, T. Chen, L. Cui, N. Q. V. Hung, H. Yin, Xsimgcl: Towards extremely simple graph contrastive learning for recommendation, IEEE Transactions on Knowledge and Data Engineering 36 (2023) 913–926.

[56] Y. Wei, X. Wang, Q. Li, L. Nie, Y. Li, X. Li, T.-S. Chua, Contrastive learning for cold-start recommendation, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 5382–5390.

[57] Z. Zhu, Y. He, X. Zhao, Y. Zhang, J. Wang, J. Caverlee, Popularity-opportunity bias in collaborative filtering, in: Proceedings of the 14th ACM international conference on web search and data mining, 2021, pp. 85–93.

[58] M. Mansoury, H. Abdollahpouri, M. Pechenizkiy, B. Mobasher, R. Burke, Feedback loop and bias amplification in recommender systems, in: Proc. of the 29th ACM International Conf. on Information & Knowledge Management, 2020, pp. 2145–2148.

[59] A. Ferraro, X. Serra, C. Bauer, What is fair? exploring the artists' perspective on the fairness of music streaming platforms, in: IFIP conference on human-computer interaction, Springer, 2021, pp. 562–584.

[60] K. Dinnissen, C. Bauer, Amplifying artists' voices: Item provider perspectives on influence and fairness of music streaming platforms, in: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization, 2023, pp. 238–249.

[61] A. Ferraro, D. Bogdanov, X. Serra, J. Yoon, Artist and style exposure bias in collaborative filtering based music recommendations, arXiv preprint arXiv:1911.04827 (2019).

[62] H. Abdollahpouri, R. Burke, M. Mansoury, Unfair exposure of artists in music recommendation, arXiv preprint arXiv:2003.11634 (2020).

[63] B. Ferwerda, E. Ingesson, M. Berndl, M. Schedl, I don't care how popular you are! investigating popularity bias in music recommendations from a user's perspective, in: Proceedings of the 2023 conference on human information interaction and retrieval, 2023, pp. 357–361.

[64] H. Abdollahpouri, M. Mansoury, R. Burke, B. Mobasher, E. Malthouse, User-centered evaluation of popularity bias in recommender systems, in: Proceedings of the 29th ACM conference on user modeling, adaptation and personalization, 2021, pp. 119–129.