Station and Track Attribute-Aware Music Personalization

M. Jeffrey Mei
jeffrey.mei@siriusxm.com
SiriusXM Radio Inc.
New York, New York, USA

Oliver Bembom
obembom@pandora.com
SiriusXM Radio Inc.
New York, New York, USA

Andreas Ehmann
aehmann@pandora.com
SiriusXM Radio Inc.
New York, New York, USA

ABSTRACT

We present a transformer for music personalization that recommends tracks given a station seed (artist) and improves the accuracy vs. a baseline matrix factorization method by 10%. Adding additional embeddings to capture track and station attributes further improves the accuracy of our recommendations by an additional 1% while also improving recommendation diversity, i.e. mitigating popularity bias. We analyze the learned embeddings and find they learn both explicit attributes provided at training and implicit attributes that may inform listener preferences. We also find that incorporating the station context of user feedback helps the model identify and transfer relevant listener preferences across different genres and artists. This particularly helps with music discovery on new stations.

CCS CONCEPTS

• **Human-centered computing** → *Information visualization*; • **Information systems** → **Personalization**; **Recommender systems**; **Music retrieval**; • **Computing methodologies** → **Learning latent representations**.

KEYWORDS

music recommendation, popularity bias, recommender systems, transformers

ACM Reference Format:
M. Jeffrey Mei, Oliver Bembom, and Andreas Ehmann. 2023. Station and Track Attribute-Aware Music Personalization. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3604915.3610239

1 INTRODUCTION

Online music streaming services rely on recommender systems to filter a catalog with tens of millions of distinct tracks to those that match a listener's preferences. One way to account for these preferences, which may depend on a listener's context, is by offering stations that a listener can select, e.g. seeded by an artist or track. This station context is important for music personalization as a listener may, for example, like both Rap and Country, but not both on the same station, and may lose interest if a recommender system repeatedly makes poor suggestions.

RecSys '23, September 18–22, 2023, Singapore, Singapore

ACM ISBN 979-8-4007-0241-9/23/09.
https://doi.org/10.1145/3604915.3610239

Collaborative filtering (CF) is a basic, low-computation method that provides a good baseline for music recommendation by using user ratings [20]. As musical tastes are context-specific, additional information like time [5] and location [18] can be used to further tune the recommendations. Alternatively, or additionally, content-based features can be used for recommendation [3, 8, 13, 22, 26]. More recently, transformers have been developed as an efficient way to provide sequential recommendations [10, 23]. Transformers have been applied to recommendation problems in media and e-commerce [e.g., 15, 28], and arguably provide more interpretable recommendations [2, 24, 27], though with some major caveats [e.g., 19]. Popularity bias is a known issue in recommender systems [1, 4, 12, 16]; mitigating it is an ongoing effort [6, 9]. Similarly, cold-start problems arising from data sparsity issues are common in CF models, which can be mitigated by sharing attributes, such as genre or artist information amongst related tracks [11, 17], or using item category information in e-commerce [25].

In this paper, we propose a Station and Track Attribute-aware Music Personalization (STAMPer) model that incorporates explicit features like song artist and genre to provide better recommendations. We find that encoding additional track attributes boosts the embedding quality for less popular tracks, as they can share artist or genre embeddings with other, more popular, tracks. This also helps counter popularity bias, as these less-popular tracks can now be recommended more often, leading to more distinct artists being recommended. Instead of using temporal information to provide context, we use the station seed to help contextualize which track attributes may be most relevant to the listener. These preferences may differ between users, as well as between different stations for the same user. The contributions of this paper are as follows:

- We demonstrate the ability of a transformer model to learn powerful embeddings that encode both explicit features like song genre, as well as implicit features like vocalist gender.
- We show that adding additional embeddings for track attributes helps overcome cold-start and data sparsity problems for new or unpopular tracks, which in turn helps mitigate popularity bias, as a more diverse range of artists can be recommended, without lowering prediction accuracy.
- We show that even without positional or temporal information, the station on which a track was liked can be used to contextualize the feedback, and hence improve recommendations, especially for music discovery on new stations.

2 ARCHITECTURE AND TRAINING

STAMPer uses SASRec [10] with some modifications. As inputs, we use the set of thumbed-up (i.e. explicitly liked) tracks, as well as the station seed artist on which the thumb occurred, spanning $\sim 10^6$ tracks, $\sim 10^7$ users and $\sim 10^8$ thumbs. For each track in a user's set of N thumbs, at training we mask that track and use the

other $N-1$ thumbs to predict that thumb as our positive target. For each positive target, we have a negative target that is sampled with probability $p = 0.1$ from explicit negatives (thumb-down, otherwise skipped tracks), and otherwise is randomly sampled from the full corpus of tracks (excluding a user's up-thumbs). For each track, we also add tag information about that track's artist and primary genre. Each **'track'** is therefore the **sum** of a **'song'**, **'song-artist'** and **'song-genre'** embedding. For each thumbed track, the station on which the thumb occurred is also added to contextualize the thumb, using a distinct **station artist** embedding. This spans the same artists as the **song-artist** embeddings but behaves differently (Sect. 4.2). We exclude positional embeddings because newer tracks are more likely to be thumbed more recently, and the precise sequence of thumbs is directly influenced by the order in which tracks were presented to the listener. Because the order in which a user likes tracks is heavily influenced by the order in which they are recommended by our model, as well as the track release year, we randomly shuffle the inputs each epoch to avoid accidentally learning positional effects, which can occur even without a positional embedding if there is a high variance in input sequence lengths [7].

We train four flavors of STAMPer: a '**base**' model (**STAMPer-Base)** that only uses song embeddings and is equivalent to vanilla SASRec without positional embeddings or a causality filter; **STAMPer-Station** that includes '**station**' artist embeddings; **STAMPer-Track** that includes '**track**' attributes (song artist and song genre embeddings); and **STAMPer-Both** with **both** station and track attributes.

3 RESULTS

The test set consists of 50,000 listeners randomly held out from training. For each user, we randomly select one each of their thumb-up and thumb-down tracks from the same station. During evaluation, we calculate the accuracy with which the thumb-up track is ranked higher than the thumb-down track. Because these two tracks occurred on the same station, this is a relatively harder problem (vs. using a random negative from the full set of tracks, which is likely a completely different genre or style of music). We evaluate different types of input thumbs for inference:

- 'All' thumbs represents the standard use case.
- 'Same' thumbs are filtered for only thumbs that occurred on the same station as the positive/negative targets. This ensures that all thumbs used for the recommendation are relevant to the positive/negative targets.
- 'Cross' thumbs are the complement of 'Same' thumbs, i.e. all thumbs that are not in the same station. This is a special case of 'All' where there are no same-station thumbs yet, and thus represents the ability of a model to generalize its recommendations to a new station (e.g. for music discovery).

As a baseline comparison, we use a matrix factorization CF method that is trained on aggregated thumb-up and thumb-down data for tracks on each station seed.

In general, STAMPer achieves higher accuracy for all above scenarios (Fig. 1), with the highest lift for cross-station recommendations. STAMPer-Station has better cross-station accuracy likely because it can identify which attributes are most relevant per station (Sect. 4.2). STAMPer-Track does not have much lift compared with STAMPer-Base, likely because track attributes benefit primarily less-popular tracks (Sect. 4.1), which by definition contribute less to the overall prediction accuracy. STAMPer-Both has the best accuracy, suggesting that track and station attributes interact in a nonlinear way to achieve better overall performance. We also check whether adding cross-station thumbs to users with same-station thumbs affects the recommendation accuracy. If so, this would imply that a recommendation algorithm should filter out cross-station thumbs if same-station thumbs exist, as these cross-station thumbs would act as distractors if they are potentially less relevant than the same-station thumbs. This is indeed true for CF, where adding cross-station thumbs decreases prediction accuracy until the number of cross-station thumbs greatly exceeds the number of same-station thumbs by approximately an order of magnitude (Fig. 2). In contrast, STAMPer-Both improves the overall accuracy for almost all users (Fig. 2). It is possible that CF may only benefit from certain cross-station thumbs (e.g. applying some artist similarity threshold); one benefit of STAMPer is that it does not require manual intervention.

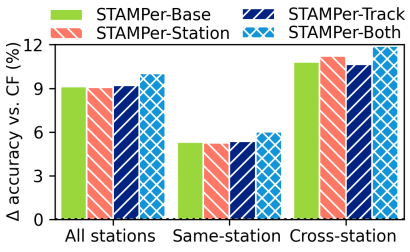

Figure 1: All STAMPer models have higher accuracy than the baseline CF model, with the highest lift for cross-station predictions. Individually, station and track attributes do not change the accuracy much, except for STAMPer-Station slightly improving on STAMPer-Base for cross-station predictions. Combining both station and track attributes gives the highest accuracy.

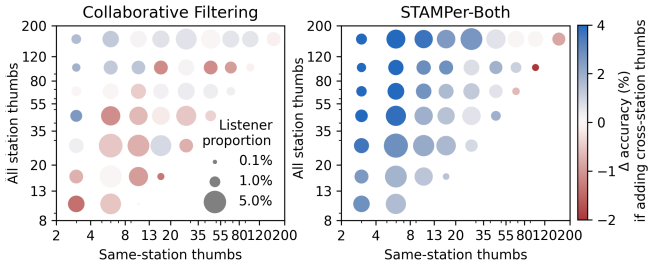

Figure 2: Relative performance in thumb prediction accuracy for users with same-station thumbs when also including their cross-station thumbs for our baseline CF model (left) and STAMPer-Both (right). Circle sizes represent the proportion of listeners with the corresponding number of thumbs. STAMPer can generally use all thumbs to improve same-station recommendations, whereas CF prediction accuracy generally decreases when cross-station feedback is combined with a listener's same-station thumbs for inference.

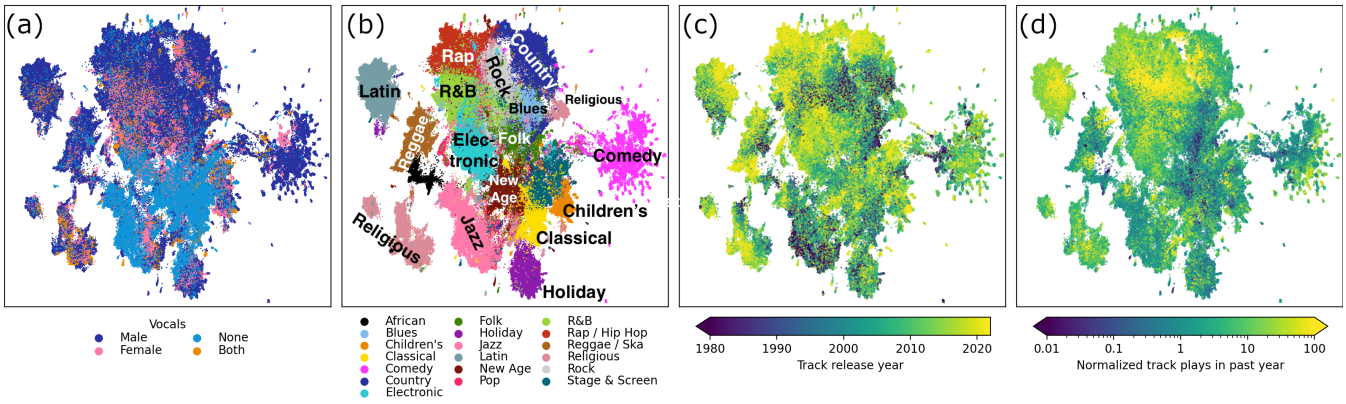

Figure 3: UMAP 2-d projection of STAMPer-Both's learned song + artist + genre embeddings, showing good awareness of both explicit attributes like genre (b), as well as implicit attributes like vocalist gender (a), release year (c) and popularity (d).

4 DISCUSSION

4.1 Track attribute embeddings

The 2D projections via UMAP [14] show the attributes that STAMPer-Both's track embeddings encode (Fig. 3). Because these include an explicit genre embedding, the genre clustering is very clear (Fig. 3b). Without an explicit genre embedding, STAMPer-Base and CF do still show some genre awareness in a UMAP projection, but with more fuzziness due to less-popular tracks of all genres tending to overlap (not shown). Some genres can also form multiple clusters, e.g. 'Religious' has two distinct clusters at the bottom left of Fig. 3b, where the smaller one is Spanish-language 'Religious' tracks. There is also an additional 'Religious' cluster bordering the 'Country' cluster.

STAMPer can also learn attributes not provided at training, such as vocalist type. For example, instrumental-only (i.e. vocals = 'None') tracks form a large cluster, which connect the genres of 'New Age', 'Classical' and 'Holiday' and 'Jazz' (Fig. 3a). In other genres (e.g. 'Comedy' and 'Country') there are distinct clusters for 'Female' and 'Male' vocals, presumably because vocalist gender may inform same-station and/or cross-station thumbs for these genres. Non-binary gender vocals represent <0.1% of tracks and are omitted from this analysis.

We also find that using the song + artist + genre embedding construction helps recommendation diversity as less-popular tracks (including those by less-popular artists) can share genre and artist embeddings with more popular tracks. For example, when comparing the superset of all top 10 recommendations for test set listeners, STAMPer-Track recommends 9% more distinct artists and 38% more distinct tracks than with CF, and 13% more distinct artists and 22% more distinct tracks than STAMPer-Base (Fig. 4). The median track for STAMPer-Track has 6% fewer annual plays than STAMPer-Base and 21% fewer than CF. As a bonus, these gains in diversity do not reduce the prediction accuracy (Fig. 1). STAMPer-Both, which has the highest prediction accuracy, also has the highest diversity with 33% more distinct tracks and 20% more distinct artists recommended than STAMPer-Base. As with the prediction accuracy, this is due to nonlinear interactions between the station and track attributes, as STAMPER-Station by itself has only 5% more distinct artists and 6% more distinct tracks than STAMPer-Base.

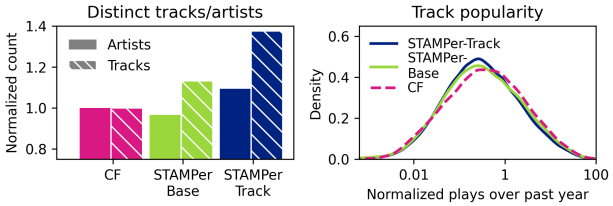

Figure 4: Adding track attribute awareness improves the recommendation diversity without decreasing recommendation accuracy. STAMPer-Track recommends 13% more distinct artists, 22% more distinct tracks and 6% fewer plays per year than STAMPer-Base.

4.2 Station awareness

Station (artist) embeddings work differently from the **song** artist embeddings. Whereas song-artist information provides information about the tracks, station-artist (for both the prediction target as well as the thumbed input tracks) informs the model about which track attributes to pay attention to; in other words, the *context* of the thumb. Adding station artist embeddings improves the prediction accuracy in particular for cross-station recommendations (Fig. 1). Good cross-station recommendations are important for music discovery, as users who are interested in a new genre or station are more likely to keep listening if they like the recommended songs. We speculate that station attributes help cross-station recommendations the most because users have different attribute preferences on different stations; knowing the station context therefore helps identify which attributes are relevant for cross-station prediction. For example, if Country listeners generally care more about vocalist gender, whereas Pop listeners generally care more about release year, then a track which could occur on either station would have different parts of its embedding emphasized depending on which

station it was thumbed on. Hence, in Fig. 1 STAMPer-Station shows a lift (vs. STAMPer-Base) for cross-station recommendations and virtually no difference for same-station recommendations. Adding both track and station attributes of course leads to the best accuracy, as the **station** attributes can help STAMPer identify which of the added **genre-** or **artist-**specific attributes to weight more.

Future work will look at incorporating other types of contextual information e.g. time, and also using bidirectional attention with mask tokens (BERT4Rec) [21]. We expect the same benefits to recommendation diversity from adding the track attributes, and greater benefits from adding station context due to the mask token only being applied to the track (and not the target station, hence allowing the the attention weights to depend on the target station, which is not the case for a SASRec-style architecture).

5 CONCLUSION

We find that transformers can greatly improve music recommendation accuracy. In addition, we find that including additional embeddings to account for track attributes and station context can improve recommendation diversity while also improving prediction accuracy. Visualizations of the learned track embeddings demonstrate both explicit and implicit attributes being learned. Including station awareness helps contextualize a listener's thumb history, which allows for different attributes of the same track to be emphasized depending on its station context. Additionally, this allows for attributes to generalize and hence transfer across different stations, which further improves recommendation accuracy, especially for new music discovery.

6 SPEAKER BIO

Jeffrey Mei has been working as a scientist at Sirius XM Radio Inc. for six months, where he works on improving music personalization. He is interested in improving the transformer interpretability to gain insights into user behavior. Previously, he worked at Wayfair on recommender systems that could use transfer style preferences across furniture types.

REFERENCES

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the eleventh ACM conference on recommender systems*. 42–46. https://doi.org/10.1145/3109859.3109912

[2] Samira Abnar and Willem Zuidema. 2020. Quantifying Attention Flow in Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4190–4197. https://doi.org/10.48550/arXiv.2005.00928

[3] Pedro Cano, Markus Koppenberger, and Nicolas Wack. 2005. An Industrial-Strength Content-Based Music Recommendation System. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Salvador, Brazil) *(SIGIR '05)*. Association for Computing Machinery, New York, NY, USA, 673. https://doi.org/10.1145/1076034.1076185

[4] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and Debias in Recommender System: A Survey and Future Directions. *ACM Trans. Inf. Syst.* 41, 3, Article 67 (feb 2023), 39 pages. https://doi.org/10.1145/3564284

[5] Ricardo Dias and Manuel J Fonseca. 2013. Improving music recommendation in session-based collaborative filtering by using temporal context. In *2013 IEEE 25th international conference on tools with artificial intelligence*. IEEE, 783–788. https://doi.org/10.1109/ICTAI.2013.120

[6] Karlijn Dinnissen and Christine Bauer. 2022. Fairness in music recommender systems: A stakeholder-centered mini review. *Frontiers in big Data* 5 (2022). https://doi.org/10.3389/fdata.2022.913608

[7] Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. 2022. Transformer Language Models without Positional Encodings Still Learn Positional Information. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 1382–1390. https://aclanthology.org/2022.findings-emnlp.99

[8] Qingqing Huang, Aren Jansen, Li Zhang, Daniel P. W. Ellis, Rif A. Saurous, and John Anderson. 2020. Large-Scale Weakly-Supervised Content Embeddings for Music Recommendation and Tagging. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8364–8368. https://doi.org/10.1109/ICASSP40776.2020.9053240

[9] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction* 25 (2015), 427–491. https://doi.org/10.1007/s11257-015-9165-3

[10] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 197–206. https://doi.org/10.1109/ICDM.2018.00035

[11] Noam Koenigstein, Gideon Dror, and Yehuda Koren. 2011. Yahoo! Music Recommendations: Modeling Music Ratings with Temporal Dynamics and Item Taxonomy. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (Chicago, Illinois, USA) *(RecSys '11)*. Association for Computing Machinery, New York, NY, USA, 165–172. https://doi.org/10.1145/2043932.2043964

[12] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The unfairness of popularity bias in music recommendation: A reproducibility study. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*. Springer, 35–42. https://doi.org/10.48550/arXiv.1912.04696

[13] Brian McFee, Luke Barrington, and Gert Lanckriet. 2012. Learning content similarity for music recommendation. *IEEE transactions on audio, speech, and language processing* 20, 8 (2012), 2207–2218. https://doi.org/10.1109/TASL.2012.2199109

[14] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* 3, 29 (2018). https://doi.org/10.48550/arXiv.1802.03426

[15] M Jeffrey Mei, Cole Zuber, and Yasaman Khazaeni. 2022. A Lightweight Transformer for Next-Item Product Recommendation. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 546–549. https://doi.org/10.1145/3523227.3547491

[16] Cataldo Musto, Pasquale Lops, and Giovanni Semeraro. 2021. Fairness and Popularity Bias in Recommender Systems: an Empirical Evaluation. 3078 (2021), 77–91.

[17] Sergio Oramas, Oriol Nieto, Mohamed Sordo, and Xavier Serra. 2017. A Deep Multimodal Approach for Cold-Start Music Recommendation. In *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems* (Como, Italy) *(DLRS 2017)*. Association for Computing Machinery, New York, NY, USA, 32–37. https://doi.org/10.1145/3125486.3125492

[18] Markus Schedl and Dominik Schnitzer. 2014. Location-aware music artist recommendation. In *MultiMedia Modeling: 20th Anniversary International Conference, MMM 2014, Dublin, Ireland, January 6-10, 2014, Proceedings, Part II 20*. Springer, 205–213.

[19] Sofia Serrano and Noah A Smith. 2019. Is Attention Interpretable?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2931–2951. https://doi.org/10.48550/arXiv.1906.03731

[20] Yading Song, Simon Dixon, and Marcus Pearce. 2012. A survey of music recommendation systems and future perspectives. In *9th international symposium on computer music modeling and retrieval*, Vol. 4. Citeseer, 395–410.

[21] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 1441–1450. https://doi.org/10.1145/3357384.3357895

[22] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. 2013. Deep content-based music recommendation. *Advances in neural information processing systems* 26 (2013).

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017). https://doi.org/10.48550/arXiv.1706.03762

[24] Andrés Villa, Vladimir Araujo, Francisca Cattan, and Denis Parra. 2020. Interpretable contextual team-aware item recommendation: application in multiplayer online battle arena games. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 503–508. https://doi.org/10.1145/3383313.3412211

[25] Jizhe Wang, Pipei Huang, Huan Zhao, Zhibo Zhang, Binqiang Zhao, and Dik Lun Lee. 2018. Billion-Scale Commodity Embedding for E-Commerce Recommendation in Alibaba. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 839–848.

https://doi.org/10.1145/3219819.3219869

[26] Xinxi Wang and Ye Wang. 2014. Improving content-based and hybrid music recommendation using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*. 627–636. https://doi.org/10.1145/2647868.2654940

[27] Minz Won, Sanghyuk Chun, and Xavier Serra. 2019. Toward Interpretable Music Tagging with Self-Attention. *ArXiv* abs/1906.04972 (2019). https://doi.org/10.48550/arXiv.1906.04972

[28] Liwei Wu, Shuqing Li, Cho-Jui Hsieh, and James Sharpnack. 2020. SSE-PT: Sequential recommendation via personalized transformer. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 328–337. https://doi.org/10.1145/3383313.3412258