Visualization of Deep Audio Embeddings for Music Exploration and Rediscovery

Philip Tovstogan **Xavier Serra** **Dmitry Bogdanov**
Music Technology Group, Universitat Pompeu Fabra
first.last@upf.edu

ABSTRACT

User interfaces for music exploration and discovery have always been an exciting application of music information retrieval (MIR) throughout the years. However, while discovering new music is a common goal of such systems, there has been less attention paid to the exploration and rediscovery within personal music collections, where finding interesting relations between music items already familiar to the user can lead to a different type of highly engaging and rewarding experience. In this paper, we present a novel web interface to visualize music collections using the audio embeddings extracted from music tracks. The system allows exploring the relationship between music tracks from multiple perspectives, displaying embedding and tag spaces extracted by music auto-tagging models, trained using different architectures and datasets, coupled with various 2D projection algorithms. We conduct a user study to analyze the effectiveness of different visualization strategies on the participants' personal music collections, particularly for playlist creation and music library navigation and rediscovery. Our results show that such an interface provides a good alternative to standard hierarchical library organization by metadata.

1. INTRODUCTION

In the age of prevalent digital streaming, exploration and discovery of the music are usually done either by the discovery playlists (generated algorithmically or curated) or the users actively looking for new music by themselves. However, the paradigm of music discovery in streaming services neglects the listeners who might want to re-engage with their personal music collections, gathered, curated, and appreciated by their maintainers throughout the years [1]. For such users, exploring their own curated music selections can be a pleasurable and rewarding experience, helping to appreciate and re-contextualize relations between music items and rediscover artists or tracks that they haven't listened to in a long time.

It can be especially relevant in the context of digital music downloads, which still have a considerable impact on independent music distribution [2] (e.g., Bandcamp [1] has gained growing digital sales over the past years with a strong following among music enthusiasts). In this context, many music consumers, and also musicians, DJs, radio hosts, music journalists, archivists, and other professionals or hobbyists that work with digital music collections can benefit from exploration and rediscovery functionality.

Research on user search behavior in the context of music streaming services identified two mindsets: *focused* and *non-focused* [3]. In the focused mindset, users know what they are looking for; and in non-focused, they only have a rough idea. While it was studied in the context of the complete catalog of the music available on the streaming services, those mindsets also apply to the users that mostly listen to their personal collection. The situation that we want to address in this paper can be summarized in the following sentence: *The user doesn't know what he wants to listen to (non-focused mindset) but wants to listen to something familiar (from personal collection).*

The interfaces for music exploration and discovery are quite homogeneous in the industry. The recommendations are usually presented in the form of the playlists or artists, and if the user wants to browse their personal collections, the interface follows the hierarchical approach: artist - album - tracks, or playlist - tracks. In the abovementioned situation, users usually resort to browsing the hierarchical metadata (artists, genres, tags) or playlists to encounter music to listen to. Algorithmically generated playlists from the tracks from the user's library suit the situation in question to a degree, but they don't provide much interaction.

Thus we propose an alternative content-based approach to represent the music in the multidimensional space which can be projected onto the 2D plane for users to see the entire collection at once. The users can interact with it and listen to short excerpts of music that can enable exploration and rediscovery of the forgotten parts of the music library.

With the wide usage of deep learning in music information retrieval, feature extraction moved from careful engineering towards learned features. There are multiple pretrained feature-extractor models available [4,5] that can be used to extract embeddings from audio. Often, these embeddings are used as input for dense neural networks for particular downstream tasks [6]. However, they can also be used as a representation of the music within the embedding space.

[1] bandcamp.com

In this paper, we take advantage of the auto-tagging systems that are trained to predict the music tags (genre, moods, instruments, etc), and use the extracted embeddings and tag predictions to visualize personal music collections. We introduce the interface that allows users to visualize the entire collection or different subsets of their collection in terms of embeddings extracted from different models and compare them qualitatively. We evaluate the interface in terms of how useful it is for the users to explore their library and create a playlist of the music that they have forgotten and would like to rediscover. In addition, we evaluate different models in terms of the users' preferences of the visualizations that have been produced.

2. RELATED WORK

Many research works perform the visualization of the music in 2D space for exploration, navigation, and recommendation [7]. In this section, we will introduce some selected works that present various interfaces.

One of the earliest works is *GenreSpace* [8] that visualizes tracks in 3D space with colors representing genres. More famous interface *Islands of Music* [9] uses a self-organizing map [10] (SOM) for visualizing music as an artificial landscape of the islands (dense clusters) in the ocean (sparse regions). The emerging islands roughly correspond to the genres of music, and the evaluation is performed mostly qualitatively by authors. The extension of the work [11] introduces several views (based on timbre, rhythm, metadata features) and the ability to switch between them. There was also another related work [12] that proposed playlist generation by drawing the trajectory on the map.

In the following years, multiple studies were also using SOM or some variation of it. *NepTune* [13] visualizes the space as a terrain that can be navigated in 3D by a user. The interface was exhibited in public, where the users could explore their collections. *Globe of Music* [14] projects the space onto sphere instead a plane with the use of GeoSOM [15]. *MusicMiner* [16] uses emerging SOM (ESOM) and U-Map to visualize transitions between genre-based groups. *SongExplorer* [17] is a tangible tabletop interface that presents the songs in a hexagonal grid also using SOM to project 7-dimensional emotion feature space to 2D.

Some interfaces used the metadata in various creative ways for visualization, like [18] that focuses on visualizing personal music collections in form of a disc, rectangle, or tree-map organized according to metadata (genre, year) and highlighted according to personal preferences or playlists.

Since 2010 and the emergence of music streaming, the studies started to focus more on web audio and digital collections. A probabilistic projection of personal music collections based on moods [19] is a remarkable study that focused a lot on user evaluation. It uses the mood features that were extracted via the commercial service from personal Spotify libraries. The features are projected with t-SNE [20] and the interface includes background highlighting based on the probabilistic models to show moods with different colors. The system enables playlist generation via both region selection and drawing trajectories. The authors performed a user study with eight participants over the course of two weeks with overall positive responses and multiple useful insights that include preference of region selection over trajectory drawing. The concept of rediscovery has also been mentioned by authors in this work.

MoodPlay [21] is a remarkable 2D interface that visualizes artists on a mood space. While the free-form exploration is supported, the system is presented as a recommendation system with primary functionality in recommending the artists based on moods. The authors conducted a very extensive user study from the perspective of human-computer interaction (HCI) that provides multiple insights. Online implementation of the interface is available [2].

One common thing in all these works is that the visualization unit is either a music track, artist, or album. As music similarity is a well-researched area of MIR, and it was a task in MIREX until 2014, the similarity on the level of tracks can go only so far until the subjectivity gets in the way [22]. Our approach is to work with the segments of the music tracks on a smaller scale, which might alleviate the subjectivity of the similarity.

Moreover, only several of the mentioned interfaces [13, 18, 19] work with the personal music collections. Our study focuses on the rediscovery of personal music collections and works with the audio files directly without using external commercial services. Also, most of the works, save for a few exceptions [16, 21, 23] have never been released publicly, as they have been used for study as prototypes. Furthermore, there is a lack of music exploration systems using the latest state-of-the-art in MIR, particularly deep embeddings.

Another common issue with the related work is that many of the mentioned papers (except a few notable exceptions) don't perform conclusive user evaluation, which is important for user-centric MIR systems [24]. We conduct a user study to evaluate our system in form of semi-structured user interviews to get feedback and analyze the functionality.

3. MODELS AND IMPLEMENTATION

We use Essentia [3] library [25] to process audio and extract representations. We use the audio embeddings extracted with modern deep auto-tagging models to represent music in the embedding space and distances between embeddings as a measure of similarity (which can be used for the music recommendation [26]).

We use two common auto-tagging architectures that have been pre-trained on two different datasets that are available in Essentia library [6]:

- *MusiCNN* [27] is a CNN with vertical and horizontal convolutional filter shapes motivated by the music domain. It contains 6 layers and 787 000 trainable parameters.
- *VGG* is an architecture from computer vision [28] based on a deep stack of 3×3 convolutional filters

[2] moodplay.pythonanywhere.com

[3] essentia.upf.edu

that had been adapted for audio [29]. It contains 5 layers and 605 000 trainable parameters.

Both architectures have been trained on two datasets:

- *Million Song Dataset* (MSD) [30] — 500 000+ tracks, collaborative tags from Last.fm[4] service.
- *MagnaTagATune* (MTAT) [31] — 5 000+ tracks, tags provided by players of the TagATune game.

While MagnaTagATune is significantly smaller and usually training on larger datasets gives higher accuracy on downstream tasks, it is commonly used in auto-tagging research and thus provides a good second option for the system. The top 50 most frequent tags from each dataset were used for training the models.

We use two layers from the models' outputs to generate the visualizations in our system:

- *Taggrams* - the output layer that provides tag activation values. The dimension of this layer is 50 for all our models, as they have been trained on top 50 tags.
- *Embeddings* - the penultimate layer of the model. The dimension of embeddings is 200 for MusiCNN and $2 \times 128 = 256$ for VGG.

We process the audio with the hop size equal to the receptive field of the model (3 seconds), which means no overlapping of the frames. We call the part of the audio of the size of the receptive field that produces one vector of output values a *segment*. Thus, the track is represented by a two-dimensional array with a vertical dimension equal to the extracted layer dimension, and the horizontal (time) dimension equal to the duration of the track divided by the size of the model receptive field.

The system is implemented in Python as a Flask web app. The code is open-source and available on GitHub[5] under GNU Affero General Public License v3.0. In the rest of this section, we will provide details of the data processing pipeline.

First, the audio is indexed in the new local SQL database[6] with the track, artist, album, and genre metadata imported from ID3 tags. Next, the audio is processed with the Essentia library with the output of several layers. The advantage of using Essentia is not only that the models are easy to use out-of-box, but also that if you have a working CUDA installation, it will be used to do TensorFlow inferencing. We extract both the tag activation values (taggrams) as well as the activations from the penultimate layer (embeddings). The taggram and embedding vectors are stacked for every segment of the audio thus resulting in a two-dimensional representation of the track, which is saved as a .npy file.

After the data for all tracks have been extracted, the PCA projection of the embeddings and taggrams is performed. We also compute STD-PCA projection, where each embedding/taggram vertical dimension is first normalized on the whole population to prevent large variation ranges in the activation values to dominate the PCA-projected space. The taggrams and embeddings are then aggregated into one .npy file per model for the efficiency of the retrieval, and the segments are indexed in the database for easy lookup of the associated track.

4. INTERFACE

The interface[7] (Figure 1) is split into several sections: music selection, visualization selection and highlighting. The user can select music to visualize either by selecting the tags of interest, or artists. One of the important aspects of the system is that it doesn't average the individual embeddings of the segments of the song. Each segment is of the appropriate length of the input size of the model (3 seconds for both MusiCNN and VGG architectures).

One point on the graph represents one segment, The reduction slider allows to show fewer segments per track to visualize many tracks at once. The number represents the step size when loading the data, so it shows all segments for a value of 1, skips every other segment for the value of 2, skips two for the value of 3, etc.

The highlighting section allows highlighting one or more artists, tags, albums, or tracks in red color on the graph. It is interesting to see the groupings and spread of the particular subset of the collection in the context of the larger selection of music.

We use Plotly[8] library to visualize the embeddings. It is a robust library that works well for our use case. One of its advantages is that it supports multiple programming languages, so it is possible to generate plots in Python and add the interactivity in JavaScript.

The visualization selection controls above the graphs allow a user to select architecture, dataset, layer, and projection to visualize embeddings. The option names have been anonymized during the user study to remove any bias that the participants might have towards any of the options. Each option can be selected individually to facilitate the comparison of the combinations. For example, the user might only change the dataset while keeping all other fields the same to see how the training dataset impacts the embedding space visualization.

The available architectures, datasets, and layers have been described in Section 3. Among the available projections apart from PCA and STD-PCA we provide t-SNE [20] and UMAP [32]. PCA and STD-PCA are computed after the extraction of the embeddings, t-SNE and UMAP are computed dynamically upon user request. So while they are slower initially, there is a caching layer implemented to prevent repeated computation of the projections of the same subset.

To get an impression of how different the embeddings spaces are, Figure 2 shows one of the users' personal music collections that was used for evaluation (with a reduction value of 20). This collection mostly consists of rock and metal music, highlighted in red is the artist *Enigma* which is tagged as *new age*. While it is mostly concentrated in one part of the visualizations, some combinations of architecture/dataset/layer manage to do a better job at clustering it.

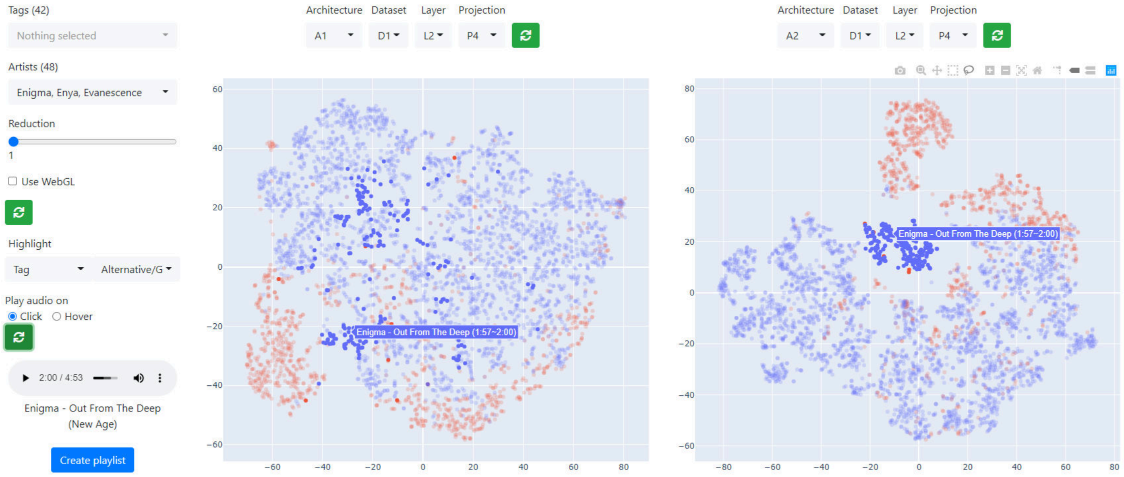
[4] `last.fm`

[5] `github.com/MTG/music-explore`

[6] SQLite via SQLAlchemy

[7] We provide a demo video of the system online: `bit.ly/3bLeFLp`

[8] `plotly.com`



Tags (42): Nothing selected

Artists (48): Enigma, Enya, Evanescence

Reduction: 1

□ Use WebGL

Highlight: Tag, Alternative/G

Play audio on: ● Click ○ Hover

2:00 / 4:53

Enigma - Out From The Deep (New Age)

Architecture: A1, Dataset: D1, Layer: L2, Projection: P4

Architecture: A2, Dataset: D1, Layer: L2, Projection: P4

Figure 1: System interface

There are several features of the system to facilitate interactivity. The user can listen to the music while hovering or by clicking the point on the graph that represents a segment of the track. Moreover, when the label of the segment is displayed on one graph, the same label for the same segment is displayed on another graph (see Figure 1). This enables easy identification of the same segment on both graphs during an interaction. Moreover, the user can select several segments on one graph with the lasso or box selection, and the corresponding segments will also be selected on the second graph. There are more tools available to zoom in and pan the individual graphs to facilitate delving deeper into the exploration of the cluster of interest.

5. EXPERIMENTS

To test the system we invited 8 users that have some sort of personal music collection to participate in user study from authors' colleagues and friends. We conducted individual semi-structured interviews with each participant to gather feedback and assess the usability and viability of the system. While there are a lot of potential uses for the system, we focus on the use case of exploration and rediscovery of the music in the private personal collections. Two main research questions that we wanted to address are: the feasibility of the system for the exploration and rediscovery of the users' music collection and the comparison of visualizations in terms of usefulness and interest to users.

Before the experiment, the participants were asked to select a subset of their private music collection that they wanted to explore. We recommended the participants limit the subset to no more than 1000 tracks, and in practice, we encountered collections of sizes from 400 to 1200. In the remote setup, we communicated with the participants through chat to help with the data extraction and ensured that the system can run normally on the users' machines. Then we conducted a video conferencing call with the participant sharing their screen. In the live setup, we asked participants to bring the music collection on the external storage device and performed data extraction and setup on the authors' machines. From 8 participants, 1 was interviewed remotely, and 7 — in-person. The data extraction took different times depending on the specification of the user machine: from 1 to 4 hours with an average of 1.5 hours. While the system doesn't require GPU for processing, most of the participants were using machines with CUDA installation, which sped up the extraction process drastically.

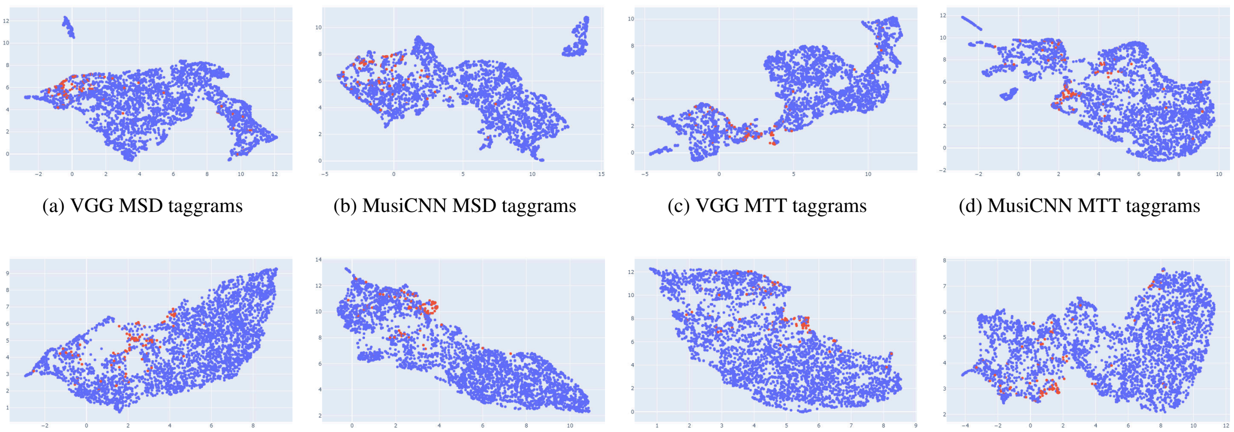
The video and audio from call and audio from the live interview were recorded with the participant's consent for further transcript analysis. The experiment started with the introduction of the features of the system to the participants by reading the introduction text. The text was kept the same to minimize the possible bias. We let participants get familiar, ask questions and play with the system and make sure that they are comfortable with it. The maximum time allocated for the familiarity phase is 10 minutes. We ensured that participants used every control element of the interface at least two times, and if they didn't, we encouraged them to use it. Then we gave the participants a task that was formulated in the following manner: *Imagine that you want to listen to something from your library that you haven't listened to in a while. Explore the system and make a playlist for yourself.*

During the interview, the participants were encouraged to try different settings and engage with the system as much as possible. When they changed the parameters of the visualization (architecture, dataset, layer, and projection), we asked them if they liked or disliked the previous combination. After the users were content with their selection of the tracks for the playlist, we asked them to fill in the questionnaire[9] to assess their thoughts about the system.

The questionnaire is split into two parts: the first part includes background questions such as age, musical training, familiarity with playlists, and experience with listening to music. The authors were present to answer any questions that the users might have about the questionnaire but didn't interfere beyond that.

The second part of the questionnaire contains questions about the system that were designed to identify which fea-

[9] The questions are available online: `bit.ly/3w9xJe0`



(a) VGG MSD taggrams (b) MusiCNN MSD taggrams (c) VGG MTT taggrams (d) MusiCNN MTT taggrams

(e) VGG MSD embeddings (f) MusiCNN MSD embeddings (g) VGG MTT embeddings (h) MusiCNN MTT embeddings

Figure 2: UMAP visualizations of *new age* (in red) in mostly rock and metal collection (reduction of 20)

tures of the system users liked, what did they think about the visualizations on both global and local levels, the usefulness of the system for music exploration, rediscovery and playlist creation. To measure users' opinions and feedback we used the 5-point Likert scale: 1 - Strongly disagree, 2 - Disagree, 3 - Neither agree nor disagree, 4 - Agree, 5 - Strongly agree. Interviewees were asked to be as critical as possible and encouraged to explain their reasoning behind the choices they made as well thinking out loud.

6. RESULTS AND DISCUSSION

The participants of our study are aged 27–39 years with an average age of 30, 7 male and 1 female. All of them have some kind of music training ranging from 1 to 20 years, the median of 6 and an average of 8 years. They listen to music 0.5–8 hours per day with 1 hour or less actively, less than 50% of the time (20% on average) to playlists. The participants create playlists with frequency ranging from every day to rarely with good coverage of all options in between. The frequency of desire to rediscover their music ranges from every day to several times per month, with most of the responses in the latter category. The broad genres covered by the users' personal music collections span mostly electronic, rock and metal.

6.1 Interaction, Exploration and Rediscovery

After analyzing the interviews and the results of the survey (see Table 1) we can see the trend that the system achieves its goal to help users to interact, explore and rediscover personal music collections and create playlists. While the quantitative results are not strong enough due to the small sample size, the focus of this study is on qualitative feedback with every participant having discovered some interesting connections between tracks in their library during the interviews. We performed topic analysis, and the last 2 interviews didn't introduce any new topics, thus providing good support that the most important topics have been covered within 8 participants.

Question	Mean ± STD
Liked interacting with system	$\mathbf{4.9} \pm 0.4$
Had preference for particular model	$3.6 \pm \mathbf{1.2}$
Preferred over browsing	4.3 ± 0.7
Preferred over random	4.4 ± 0.9
Liked big picture	3.8 ± 1.0
Liked segment groupings	4.4 ± 0.7
Discovered unexpected connections	$\mathbf{4.5} \pm 0.5$
Rediscovered something	$\mathbf{4.6} \pm \mathbf{1.1}$
Want to use for playlist creation	4.1 ± 1.0
Want to use for inspiration	4.3 ± 0.7
Had rewarding experience	$4.1 \pm \mathbf{1.1}$
Had engaging experience	$\mathbf{4.5} \pm 0.8$

Table 1: Summarized results from Likert scale questions

One of the topics that came up in several interviews was about using *segments* instead of tracks, segment length, and possible averaging of the segments. An argument in favor of using segments is that they are short, concise, can represent better the music evolution with time and span multiple tags, and are easier to perceive as a unit. For example, while it might be difficult to say which track is more similar to the reference track, some participants agreed that it is relatively easier to answer the same question with the segments.

However, multiple participants remarked that the length of 3 seconds is too short. While the similarity might be easier to judge, it might not translate well towards track similarity, exploration process and lead towards undesirable behavior during playlist generation. For example, if there is a segment of low-energy music in the cluster of similarly chill tracks, but the segment is an interlude in a much more aggressive track, the track in question will be undesirable in the low-energy playlist. Some participants mentioned that they would prefer segments of at least 10 seconds.

One suggestion that came up multiple times is to *average* embeddings of several segments. It makes sense for the segments that are similar to each other. However, if

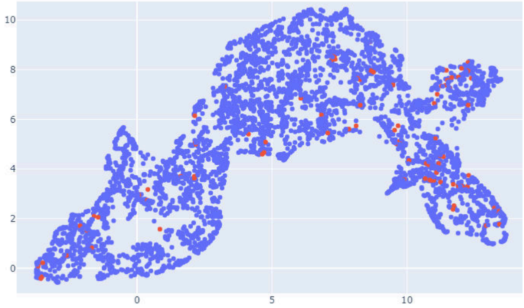


Figure 3: Long complex track highlighted in red

the segments are quite distinct and are from two different regions in the embedding space, taking the average might put the resulting average into a new third region that has nothing to do with the original ones. This problem is exacerbated on a larger scale, where averaging can make tracks that are very complex and span multiple regions in the embedding space (Figure 3) be reduced to several points which are not representative of the dynamics of the track. For certain genres it can be quite a bit problem, for example for different movements in classical music.

Participants' opinions also varied towards the ability of the interface to *visualize the entire collection*. Some of the participants noted that it was nice that all aggressive and high-energy tracks were on one side with the more chill and relaxed tracks on the other side. One participant mentioned *"(pointing at one side of the visualization) here is hard music, music that my mother doesn't like, but if I come here (pointing at the opposite side), it is more peaceful, relaxing"* while moving from one side of the visualization to the opposite one. The semantics gradients that were mentioned as obvious from the big picture are (depending on the architecture/dataset): rock–ambient, electronic–acoustic, vocal–instrumental. Those semantics are indeed represented by the tags from the training datasets, and it is useful to see that the participants agree on those semantics. Few other participants didn't pay any attention to the global distribution and dived right in exploring clusters hovering the mouse over the different regions of embedding space. Some participants enjoyed zooming into random clusters, while others didn't utilize zoom functionality as much.

Rediscovery was the part of the experience that almost all the participants were very happy and vocal about. Ones that weren't particularly keen on rediscovery evaluated the system more in the context of DJing. Encountering artists and tracks that they haven't listened to in a while happened both during the random walks over the entire space and while investigating local clusters. The same can be said about unexpected connections with several participants saying *"I would never think to put these two artists together in a playlist, but it works quite well for these tracks,"* or *"if you listen to segments, they sound quite similar in timbre, what won't happen to full tracks."* Some participants have noted that it was good to have an audio player in the interface because if they would be using the system outside of the interview, they would stop the exploration process and listen to the track that they stumbled upon from start to finish.

Interestingly enough, the *highlighting* functionality of a particular artist / album / track / tag became quite divisive — many participants used it to highlight a tag or an artist either as a seed to go from or as a target that they wanted to explore. It is the functionality that was most often mentioned as a favorite in the questionnaire, however, some participants didn't engage with it after the introduction.

As the tags that the models are trained on are quite generic (guitar, vocal, rock, chill, electronic, etc.), several participants mentioned that the models probably are not capable of distinguishing subtle differences between subgenres of their homogeneous collection by pointing out the segments in some clusters that don't belong together due to the style. One participant noted: *"The similarity is not captured well between different styles of dance music."*

Overall, the participants took between 5 to 10 minutes to get familiar with the system and 2 to 20 minutes to explore it to try different visualizations and make a selection that would produce a playlist that they are satisfied with. However, after they have created the playlist that they were content with, some participants spent a lot of time continuing exploration of other regions of their collection. Several users mentioned that there could be other methods to generate a playlist, for example, track- or artist-based radio that uses the seed segment or track: *"Maybe the system can lasso select tracks for me."* The playlist creation functionality was mentioned multiple times as a very strong use case for using the system after the novelty of interaction would wear off.

6.2 Comparison of Visualizations

Even if the sample size for the comparison study is not large to draw strong conclusions, after analyzing the responses to the question of whether the participants liked or disliked a particular combination of architecture / dataset / layer / projection, some interesting trends can be identified. As mentioned before, all options were anonymized for user testing to remove potential biases. The only option that could be inferred was the projection, as participants could guess which type of the projection it is just by looking at the graphs, however, no participants made it obvious that they recognized any projections.

Several participants mentioned that they liked two visualizations side-by-side and engaged in using both to select subsets. Some participants pointed out that different combinations capture well different aspects of similarity: *"It seems that A2D2 (MusiCNN-MTAT) can separate ambient from drums, while A1D1 (VGG-MSD) clusters the timbral aspect of sounds together well"* and took advantage of that by using both at the same time. The combination VGG-MSD has been mentioned by multiple participants as being good for timbre similarity.

Among architectures, datasets, layers, and projections, participants had the strongest preferences towards projection options. Most participants mentioned that the distribution looks more interesting in UMAP (5) and t-SNE (4) compared to PCA and STD-PCA. We attribute it to both t-SNE and UMAP being non-linear transformations, and UMAP preserving distances better than t-SNE. Nonlinearity helps to represent the local structure better at the cost of the global structure. The common comments in favor of PCA and STD-PCA are that they are faster, and capture the global structure much better. *"P1 (STD-PCA) seems to group the same sounds that I would put together for DJing"*

While participants were encouraged to compare different architectures, datasets, and layers, it took a lot of effort from the participants and was less engaging than the exploration of the visualizations that are already in front of them. We conclude that there should be a separate experiment to present participants with predetermined comparison pairs for proper evaluation. Although, all participants answered positively to the question of them having a favorite combination of architecture / dataset / layer / projection.

Comparing the architectures coupled with datasets, commonly mentioned as good were combinations VGG-MSD (3) and VGG-MTT (3), a bit less MusiCNN-MTT (2). While VGG is an architecture taken from computer vision without many modifications, and MusiCNN takes advantage of the music domain knowledge in the filter design, there was no conclusive evidence for one being preferred more than the other. Taggram layer was mentioned several times in the preferred combinations (4), more than the embedding layer (2). This might indicate that the semantics of the tags is more useful and representative than the deeper layer of the neural network.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we present the interface that allows users to visualize personal music collections. To the best of our knowledge, this is the first study proposing a music exploration interface that uses state-of-the-art deep audio embeddings. Importantly, the system is open-source, the installation process is well documented and it is easily extendable with other models for extracting feature embeddings.

We evaluated our system via semi-structured interviews with the users. From the evaluation results, we can conclude that this interface is engaging and rewarding to use for people when they are in the mood for rediscovery or exploration of personal music collections. Moreover, the results of the questionnaire strongly support the usefulness and viability of the system. We believe that such systems can be extended to the case of music discovery and exploration that is not limited to personal music collections.

While the performed small-scale evaluation provides initial results and insights on the preferences for architectures, training datasets, layers, and projections, a larger study needs to be conducted to gather more data to support our initial findings.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068.

8. REFERENCES

[1] S. J. Cunningham and S. J. Cunningham, "Interacting with personal music collections," in *Information in Contemporary Society (ICS).* Washington, DC, USA: Springer, Mar. 2019, pp. 526–536.

[2] IFPI, "Global music report 2021," 2021.

[3] A. Li, J. Thom, P. Chandar, C. Hosey, B. S. Thomas, and J. Garcia-Gathright, "Search mindsets: Understanding focused and non-focused information seeking in music search," in *The World Wide Web Conference (WWW).* San Francisco, CA, USA: ACM, May 2019, pp. 2971–2977.

[4] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* New Orleans, LA, USA: IEEE, Mar. 2017, pp. 131–135.

[5] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Brighton, UK: IEEE, May 2019, pp. 3852–3856.

[6] P. Alonso-Jiménez, D. Bogdanov, J. Pons, and X. Serra, "Tensorflow audio models in Essentia," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Barcelona, Spain: IEEE, May 2020, pp. 266–270.

[7] P. Knees, M. Schedl, and M. Goto, "Intelligent user interfaces for music discovery," *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, pp. 165–179, Oct. 2020.

[8] G. Tzanetakis, "Automatic musical genre classification of audio signals," in *Proceedings of the 2nd International Symposium on Music Information Retrieval (ISMIR).* Bloomington, IN, USA: Zenodo, Oct. 2001.

[9] E. Pampalk, A. Rauber, and D. Merkl, "Content-based organization and visualization of music archives," in *Proceedings of the 10th ACM International Conference on Multimedia (MM).* Juan-les-Pins, France: ACM, Dec. 2002, pp. 570–579.

[10] T. Kohonen, *Self-organizing maps*, 3rd ed., ser. Springer Series in Information Sciences. Berlin, Heidelberg: Springer, 2001, vol. 30.

[11] E. Pampalk, S. Dixon, and G. Widmer, "Exploring music collections by browsing different views," *Computer Music Journal*, vol. 28, no. 2, pp. 49–62, Jun. 2004.

[12] R. Neumayer, M. Dittenbach, and A. Rauber, "PlaySOM and PocketSOMPlayer, alternative interfaces to large music collections," in *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)*. London, UK: Zenodo, Sep. 2005, pp. 618–623.

[13] P. Knees, M. Schedl, T. Pohle, and G. Widmer, "An innovative three-dimensional user interface for exploring music collections enriched," in *Proceedings of the 14th ACM International Conference on Multimedia (MM)*. Santa Barbara, CA, USA: ACM, Oct. 2006, pp. 17–24.

[14] S. Leitich and M. Topf, "Globe of music - music library visualization using geosom," in *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR)*. Vienna, Austria: Zenodo, Sep. 2007, pp. 167–170.

[15] Y. Wu and M. Takatsuka, "Spherical self-organizing map using efficient indexed geodesic data structure," *Neural Networks*, vol. 19, no. 6-7, pp. 900–910, Jul. 2006.

[16] F. Mörchen, A. Ultsch, M. Nöcker, and C. Stamm, "Databionic visualization of music collections according to perceptual distance," in *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)*. London, UK: Zenodo, Sep. 2005, pp. 396–403.

[17] C. F. Julià and S. Jordà, "SongExplorer: a tabletop application for exploring large collections of songs," in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*. Kobe, Japan: Zenodo, Oct. 2009, pp. 675–680.

[18] M. Torrens, P. Hertzog, and J. L. Arcos, "Visualizing and exploring personal music libraries," in *Proceedings of the 5th International Society for Music Information Retrieval Conference (ISMIR)*. Barcelona, Spain: Zenodo, Oct. 2004.

[19] B. Vad, D. Boland, J. Williamson, R. Murray-Smith, and P. B. Steffensen, "Design and evaluation of a probabilistic music projection interface," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*. Málaga, Spain: Zenodo, Oct. 2015, pp. 134–140.

[20] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, Nov. 2008.

[21] I. Andjelkovic, D. Parra, and J. O'Donovan, "Moodplay: Interactive music recommendation based on artists' mood similarity," *International Journal of Human-Computer Studies*, vol. 121, pp. 142–159, Jan. 2019.

[22] A. Flexer, "On inter-rater agreement in audio music similarity," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*. Taipei, Taiwan: Zenodo, Oct. 2014, pp. 245–250.

[23] M. Hamasaki, M. Goto, and T. Nakano, "Songrium: Browsing and listening environment for music content creation community," in *Proceedings of the 12th Sound and Music Computing Conference (SMC)*. Maynooth, Ireland: Zenodo, Jul. 2015, pp. 23–30.

[24] M. Schedl and A. Flexer, "Putting the user in the center of music information retrieval," in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*. Porto, Portugal: Zenodo, Oct. 2012, pp. 385–390.

[25] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. R. Zapata, and X. Serra, "Essentia: An audio analysis library for music information retrieval," in *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*. Curitiba, Brazil: Zenodo, Nov. 2013, pp. 493–498.

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

[27] J. Pons and X. Serra, "MusiCNN: Pre-trained convolutional neural networks for music audio tagging," Sep. 2019.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR)*. San Diego, CA, USA: arXiv, May 2015.

[29] K. Choi, G. Fazekas, and M. B. Sandler, "Automatic tagging using deep convolutional neural networks," in *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*. New York City, NY, USA: Zenodo, Aug. 2016, pp. 805–811.

[30] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*. Miami, FL, USA: Zenodo, Oct. 2011, pp. 591–596.

[31] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: the case of music tagging," in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*. Kobe, Japan: Zenodo, Oct. 2009, pp. 387–392.

[32] L. McInnes, J. Healy, and J. Melville, "UMAP: uniform manifold approximation and projection for dimension reduction," Feb. 2018.