B Seed artists

We use the following artists as a seed list for the traversal of the artist graph on Spotify.

```
starting_artists = [
    '2ye2Wgw4gimLv2eAKyk1NB', # Metallica
    '3WrFJ7ztbogyGnTHbHJFl2', # The Beatles
    '74ASZWbe4lXaubB36ztrGX', # Bob Dylan
    '1Xyo4u8uXC1ZmMpatF05PJ', # The Weeknd
    '4y6J8jwRAwO4dssiSmN91R', # Muddy Waters
    '776Uo845nYHJpNaStv1Ds4', # Jimi Hendrix
    '6kACVPfCOnqzgfEF5ryl0x', # Johnny Cash
    '67ea9eGLXYMsO2eYQRui3w', # The Who
    '7dGJo4pcD2V6oG8kP0tJRR', # Eminem
    '5pKCCKE2ajJHZ9KAiaK11H', # Rihanna
    '0dmPX6ovclgOy8WWJaFEUU', # Kraftwerk
    '1ZwdS5xdxEREPySFridCfh', # 2pac
    '0kbYTNQb4Pb1rPbbaF0pT4', # Miles Davis
    '6tbjWDEIzxoDsBA1FuhfPW', # Madonna
    '7guDJrEfX3qb6FEbdPA5qi', # Stevie Wonder
    '2QsynagSdAqZj3U9HgDzjD', # Bob Marley
    '5aIqB5nVVvmFsvSdExz408', # Johann Sebastian Bach
    '0Kekt6CKSo0m5mivKcoH51', # Sergei Rachmaninoff
]
```

C Columns in DISCO-10M Dataset

```
dataset_columns = [
    'video_url_youtube',
    'video_title_youtube',
    'track_name_spotify',
    'video_duration_youtube_sec',
    'preview_url_spotify',
    'video_view_count_youtube',
    'video_thumbnail_url_youtube',
    'search_query_youtube',
    'video_description_youtube',
    'track_id_spotify',
    'album_id_spotify',
    'artist_id_spotify',
    'track_duration_spotify_ms',
    'primary_artist_name_spotify',
    'track_release_date_spotify',
    'explicit_content_spotify',
    'similarity_duration',
    'similarity_query_video_title',
    'similarity_query_description',
    'similarity_audio',
    'audio_embedding_spotify',
    'audio_embedding_youtube',
]
```

D Additional Examples: Audio Similarity Comparison

We demonstrate the results of our audio similarity approach on five additional samples (see Figure 6). Similarly to the spectrograms presented in Section 3.2, we also observe an overlap for these samples when the audio similarity is above $\delta_a > 0.4$. Even when two music snippets have the same frequency characteristics, there might still be small differences. This can be explained in part due to the audio quality of a YouTube video, which is dependent on the quality selected by the person uploading the video, and can therefore vary greatly, unlike the audio quality of Spotify. This difference can be seen best in the high-frequency content of the spectrogram, which tends to be weaker and less pronounced in the YouTube audio samples. We notice a strong dissimilarity in the first example between the Spotify preview audio spectrogram and the YouTube audio spectrogram. This is reflected by the low similarity score of $\delta_a = 0.1985$.

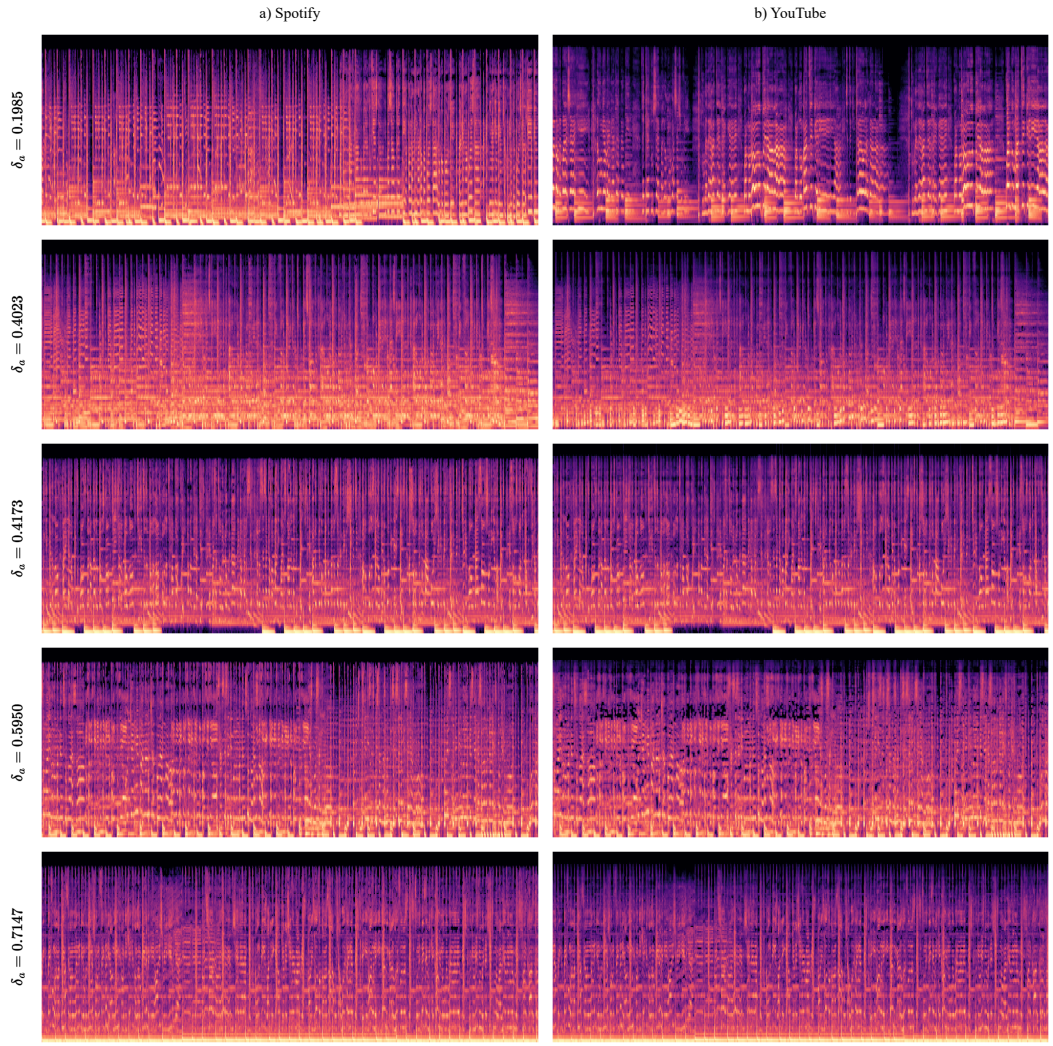

Figure 6: Comparison of audio similarity between Spotify preview audio and YouTube audio. δ_a denotes the cosine similarity of the audio embedding.We observe that the similarity of our audio embeddings is related to the similarity of the Log-Mel Spectrograms, and that the similarity increases when the spectrograms are closer to each other.

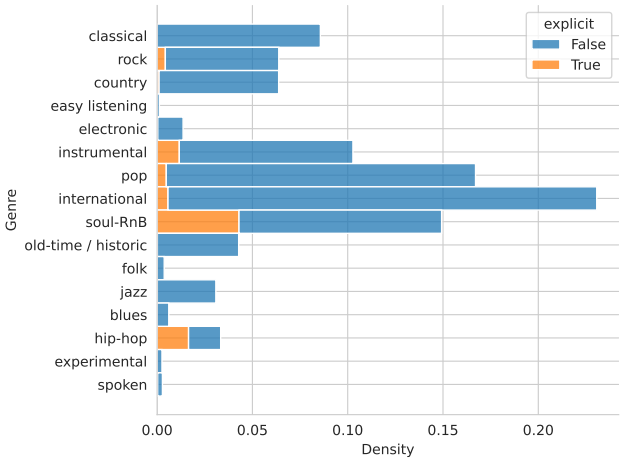

Figure 7: Genre distribution of FMA root genres in our dataset. The genre embeddings were computed using a CLAP encoder on the sentence `This audio is a <genre> song.` Each song is mapped to a genre according to the largest cosine similarity between the song embedding and the genre embedding.

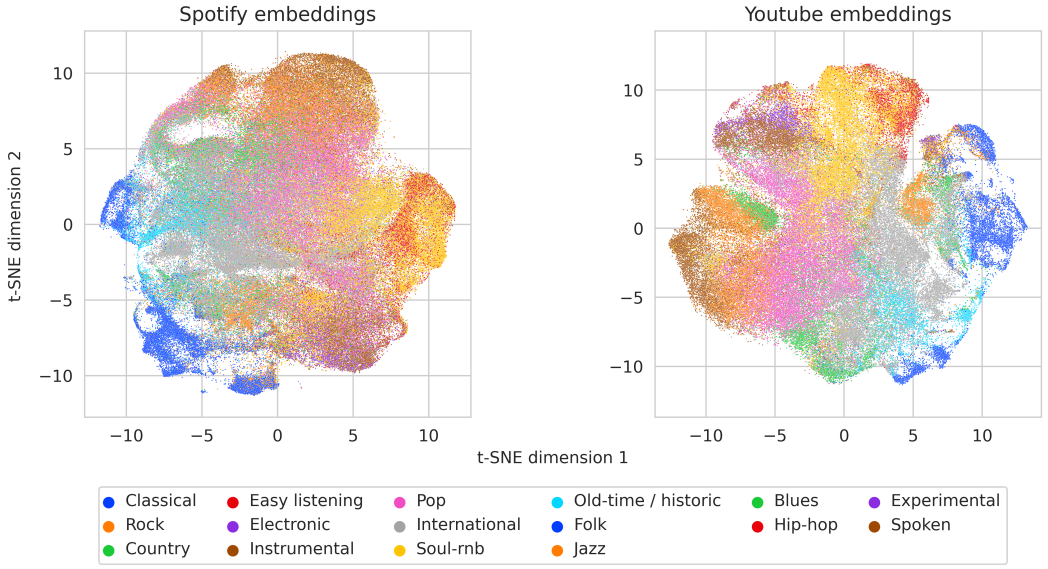

Figure 8: t-SNE plots for Spotify preview embeddings and YouTube audio embeddings computed with CLAP. Colors represent FMA root genres that were computed from the Spotify embeddings with zero-shot genre classification. The t-SNE plots show the relative positions of music samples, where samples with similar embeddings in the CLAP latent space are located closer to each other and dissimilar points farther apart. We observe that genres are well separated for both the YouTube embeddings and Spotify embeddings.

E FMA Genre Analysis

We repeat the zero-shot genre classification from Section 4 for the 16 FMA root genres. Figure 7 shows the genre distribution for the FMA genres, while Figure 5 depicts the same t-SNE plot as shown in Figure 5 for these genres. We can observe that our results between overlapping genres are consistent and although there are more genres, we can observe meaningful relationships between them.

F Subset Details

As described in Section 3.3, we provide three different subsets of DISCO-10M.

DISCO-10K-random contains 10,000 random samples from DISCO-10M. We select the samples randomly from unique Spotify track IDs, meaning that the dataset will contain exactly one YouTube video link per Spotify song.

DISCO-200k-high-quality contains 200,000 high-quality samples filtered more strictly to improve the quality of the matches. We created this subset by filtering DISCO-10M with $(\delta_a > 0.7) \wedge (\delta_{yt} > 0.8)$.

G Audio Characteristics

The total playtime of YouTube videos in DISCO-10M is around 1,062,604 hours or 121 years.

We provide further insights regarding the audio attributes of the sample rate, MP3 file bitrate, and number of channels on the DISCO-10K-random subset. When downloading Spotify previews in MP3 format, the audio quality and characteristics remain consistent. All audited samples share common attributes: a sample rate of 44.1 kHz, a bitrate of 96 kbps, and a 2-channel stereo setup.

In the case of audio from YouTube, there is a noticeable resemblance, albeit slightly less uniform. 99.78% of the examined videos employ a standard stereo 2-channel configuration. 0.18% of videos are mono channel, while 0.04% utilize 6 channel surround for audio output. 99.96% of videos have a sample rate of 44.1 kHz, aligning with the settings of Spotify previews—the remaining 0.04% deviate by having a sample rate of 48 kHz.

J Ethical Considerations

Copyright and licensing agreements are a complex issue, particularly when it comes to big data collection for training large machine learning models. We acknowledge the concerns of artists regarding the potential negative impact on their artistic work. However, we believe that openly sharing such data helps democratize the research of music understanding and music creation.

To address artists who disagree with our assessment, we offer two options for reconciliation. First, artists can contact us at *anonymized* to request the removal of links associated with their art from our dataset. Second, artists may choose to take down their YouTube video or Spotify song from the respective platform, rendering the link contained in DISCO-10M invalid. It is important to note that our guarantee applies solely to our dataset, while other entities who hold private audio datasets may not offer the same level of control.

Creating a dataset of this magnitude is achievable using publicly available tools and a reasonable timeframe. By making our dataset open-source, we also aim to raise awareness on the ease of creating big datasets and uncover the potential existence of similar datasets held by private institutions. Our goal is to provide an opportunity for anyone interested to explore ideas with this dataset, and to enhance our understanding of music creation and safety with large datasets. We emphasize that this dataset serves as a starting point, pushing the boundaries and fostering research of enhanced datasets for various tasks in machine learning for music. Access to such extensive datasets is crucial, not only in the visual domain as demonstrated by Laion-5B, but also in the domain of music.

In summary, our ethical framework emphasizes the importance of respecting artists' concerns, providing options for data exclusion, promoting transparency in dataset creation, and facilitating meaningful exploration of ML-assisted music creation while prioritizing safety considerations.

K Datasheet for Datasets

K.1 Motivation

1. **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
    - We want to provide an open-source large-scale music dataset for the research community. Such large datasets do not yet exist in this domain, and we believe they are needed to democratize innovation in music research and ML-assisted music creation. Working with large data also has inherent risks, which are better analyzed openly by a large community rather than by private institutions behind closed doors.
2. **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
    - *Anonymized*
3. **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.
    - *Anonymized*
4. **Any other comments?**
    - No.

K.2 Composition

1. **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions be- tween them; nodes and edges)? Please provide a description.
    - We share 11,018,816 YouTube links to music with metadata and associated Spotify music metadata. In addition we provided similarity measures between the YouTube video title, the YouTube description, the Song title, and name of the artist. Additionally, contribution includes providing audio embeddings for the YouTube video and the Spotify song preview computed with Laion-CLAP [35]. The metadata includes an explicit flag to allow users to filter for explicit or non-explicit music.
2. **How many instances are there in total (of each type, if appropriate)?**
    - 11,018,816
3. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
    - No, DISCO-10M does not cover all artists on Spotify and only a selection of popular songs of those that we do consider. The YouTube search results only contain 20 matches we take into consideration. To improve the dataset quality, we filter out matches that do not meet the threshold described in Section 3.2.
4. **What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
    - URLs to YouTube videos and Spotify song previews as well as song specific metadata, such as artist names, artist/song IDs, YouTube video title/description snippet, video views, duration, Spotify song duration and creation date
5. **Is there a label or target associated with each instance?** If so, please provide a description.
    - No.

6. **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
    - The artist names are not always known since we do not have this information for every artist ID in our dataset. This is the case in 3.46% of all datapoints. In addition we have 7.81% missing YouTube description snippets and 0.183% missing YouTube view counts.
7. **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.
    - No.
8. **Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.
    - No.
9. **Are there any errors, sources of noise, or redundancies in the dataset?**If so, please provide a description.
    - We acknowledge the existence of duplicate songs stemming from different YouTube videos corresponding to the same Spotify song. These duplicates can be removed by filtering with stricter thresholds (cf. Section 3.2).
10. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
    - DISCO-10M relies on the availability of the songs on YouTube and Spotify since we only link to those resources. The embeddings and other metadata are self-contained.
11. **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor– patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.
    - No, there are no confidential datapoints in DISCO-10M.
12. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.
    - DISCO-10M contains music with an explicit flag. We do not know in what ways the song is explicit (sexual, abusive or others) but the flag allows users to easily filter for such songs. Additionally, DISCO-10M does not contain any links to age-restricted YouTube video.
13. **Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
    - No.
14. **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.
    - Yes, the Spotify artist ID is directly related to one or multiple natural persons. Additionally, the YouTube video URLs we provide in the dataset are uploaded by one or multiple natural persons.

15. **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union member- ships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.
    - No.
16. **Any other comments?**
    - We emphasize the focus of our dataset on music, and not on individuals. Additionally, we reiterate that this dataset is intended for research purposes only, as described in Section 5.

K.3 Collection Process

1. **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
    - The YouTube videos and metadata, and the Spotify tracks and metadata are observable and were collected by accessing the Spotify API as well as the YouTube API. The similarity scores and audio embeddings are computed by us.
2. **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** How were these mechanisms or procedures validated?
    - The Spotify API and the YouTube API. Our results were validated manually by assessing the quality of the retrieved information on random samples.
3. **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
    - We started the Spotify artist scraping from the artist seed described in Appendix B. Additionally, we filter high-quality datapoints as described in Section 3.2.
4. **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**
    - Only the authors of this paper were involved in the data collection process. *Author involvement and payment disclosed after acceptance.*
5. **Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time- frame in which the data associated with the instances was created.
    - January 2023 to June 2023 was the timeframe of data collection. The creation time of the songs is diverse and can be seen in Figure 4.
6. **Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
    - No.
7. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
    - We collected data from Spotify and YouTube, not from artists directly.
8. **Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

- We did not notify any individuals about the data collection.

9. **Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
    - We link to publicly available music on Spotify and YouTube. We allow every artist contained in our dataset to have their entries removed upon request.
10. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
    - Artists have the possibility to search our dataset for their YouTube video links, and their Spotify artist ID and track IDs. If artists wish to remove their content from YouTube or Spotify, they can contact those parties or remove it themselves, this would result in our links becoming invalid. Additionally, we allow artists to contact us at *anonymized* to request the removal of their datapoints.
11. **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?**If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
    - Yes, we discuss the implications of our data collection pipeline and of our dataset in Appendix J.
12. **Any other comments?**
    - No.

K.4 Preprocessing/cleaning/labeling

1. **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**
    - We performed preprocessing by filtering, as described in Section 3.2. We do not process videos that are marked as age-restricted by YouTube, and we provide the explicit content flag from Spotify.
2. **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.
    - No.
3. **Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.
    - We use `spotipy` to access the Spotify API, `youtubesearchpython` to query the YouTube search, and `pytube` to access the video on YouTube.
4. **Any other comments?**
    - No.

K.5 Uses

1. **Has the dataset been used for any tasks already?** If so, please provide a description.
    - No.
2. **Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

- No.

3. **What (other) tasks could the dataset be used for?**
   - We encourage the research community to use the dataset for music analysis, video analysis, music information retrieval, generative models for music, music genre recognition, as well as other possible down-stream tasks enabled by the provided embeddings.
4. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
   - Yes, as discussed in Section 5.
5. **Are there tasks for which the dataset should not be used?** If so, please provide a description.
   - We strongly advise to use DISCO-10M only for research purposes and not for commercial applications.
6. **Any other comments?**
   - No.

K.6 Distribution

1. **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**If so, please provide a description.
   - Yes, the dataset will be open-source.
2. **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?**Does the dataset have a digital object identifier (DOI)?
   - The dataset will be available on Hugging Face Datasets. DOI: 10.57967/hf/0754
3. **When will the dataset be distributed?**
   - Starting from 14.06.2023.
4. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
   - CC-BY-4.0
5. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
   - We do not own the copyright of the music accessible through the provided links.
6. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
   - No.
7. **Any other comments?**
   - No.

K.7 Maintenance

1. **Who will be supporting/hosting/maintaining the dataset?**
    - Hugging Face Datasets will host the dataset and we will maintain the dataset.
2. **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
    - The authors can be contacted via *anonymized* email.
3. **Is there an erratum?**If so, please provide a link or other access point.
    - Not initially, will be started when necessary, and will be documented with future releases.
4. **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?
    - No, except for updates due to removal of dataset entries. Updates will be communicated on Hugging Face Datasets.
5. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.
    - Artists may contact us to have entries excluded from our dataset.
6. **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
    - There are no older versions of DISCO-10M.
7. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.
    - Updating and extending the dataset will be done on a case-by-case basis.
8. **Any other comments?**
    - No.