Vision, Modeling, and Visualization (2024)
L. Linsen and J. Thies (Eds.)

Musicon: Glyph-Based Design for Music Visualization and Retrieval

Xuejiao Luo, Vera Hoveling and Elmar Eisemann

Delft University of Technology, The Netherlands

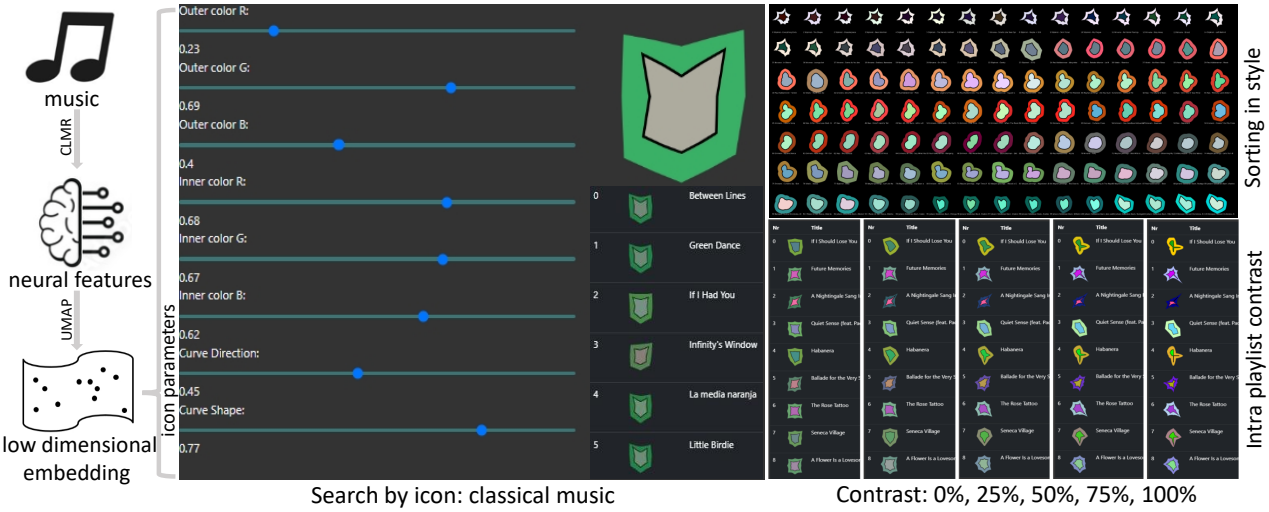

Figure 1: *Workflow and interaction scheme of our Musicon system: the music data runs through a pre-trained Contrastive Learning of Musical Representations (CLMR) model towards high dimensional neural features, which are further embedded into eight dimensions as glyph parameters for characterising the music icon. The figure shows a customized icon for classical music and a list of retrieved songs. Applications include sorting music files by style, identifying similar songs, and navigation within a playlist. An overview-first, details-on-demand approach is used by enhancing icon contrast, useful for examining large song collections and distinguishing similar songs.*

Abstract

This paper introduces a novel glyph-based design for music representation that leverages deep latent features to improve user-directed search for music discovery. We propose a system that combines a pre-trained neural network model for high-level music feature extraction with dimensionality-reduction methods for effective visual mapping of the intrinsic characteristics that help distinguishing a song. We provide a search-by-icon user interface (UI) that integrates glyph based on the neural features in combination with other novel navigation methods to achieve intuitive search and exploration. A detailed user study validates our approach, demonstrating its efficacy in enabling swift song clustering, identification, and retrieval. Our findings reveal that our visual representation not only speeds up the music searching process but also fosters increased user interaction with digital music libraries, representing a valuable contribution to the domain of music exploration and retrieval.

CCS Concepts

• ***Human-centered computing*** → ***Interaction design; Graph drawings; Information visualization;***

1. Introduction

As music streaming platforms evolve, their role has expanded from merely providing access to vast music collections to becoming pivotal in music discovery [Cho19; HVS*19]. Users typically engage with new music through two main avenues: algorithmic recommendations and user-guided search and exploration. The latter, in particular, plays a critical role in diversifying users' music experiences, a factor increasingly recognized for its correlation with long-term user engagement [AMA*20]. Despite advancements in algorithms for navigating music collections, the integration of in-

DOI: 10.2312/vmv.20241199

novative visual cues into user interfaces remains limited. Conventional icons, such as album art, provide minimal insight into a song's characteristics, forcing users to rely solely on auditory exploration to determine their preference. This highlights a gap in user experience, as visual representations could potentially expedite the discovery process by indicating auditory similarities, even though music is primarily an aural medium [SZC*18; HVS*19].

Platforms like Spotify provide their users the function to personalize the playlist, yet these innovations fall short in conveying the musical essence of unseen tracks from unfamiliar artists. The inherent challenge lies in the inability to avoid a time-consuming listening process. Our work posits that using visualization can provide information that can be efficiently parsed to compare characteristics, such as mode, tempo, or mood. Hereby, we can notably enhance user-guided search and exploration. Our solution reduces the time users spend finding music that aligns with their taste or current mood, especially when navigating with an open or exploratory mindset [Wol10; HVS*19]. This hypothesis is supported by evidence suggesting that visual identifiers can expedite navigation in user interfaces [LRFN04]. Yet, our goal is not necessarily to derive a global visual encoding of music, but rather a visualization method to enable comparisons to facilitate exploration. Users do not have to learn the meaning of individual representations in order to effectively use them, and even in rather homogeneous song collections, our visualization can provide a clear visual differentiation.

Our contributions are twofold. First, we propose an approach to extract latent characteristics of music utilizing state-of-the-art deep learning models. Second, we introduce a new visualization solution that employs custom-designed icons that embed these features in visual cues, facilitating music exploration, including interaction methods to compare, search, and categorize. It improves music visualization by combining advanced representation learning with user-centered glyph design principles.

The article is organized as follows: Section 2 reviews related work. Section 3 presents our approach, covering feature extraction, dimensionality reduction, and glyph generation. Section 4 integrates the icon into our UI prototype. Section 5 evaluates our work through a user study, followed by a conclusion in Section 6.

2. Related Work

This section explores foundational work and recent advancements in music features, latent representation learning, and glyph-based music visualization, framing the context for our contributions to music discovery through visual representation.

Understanding music features spans from low-level signal descriptors to high-level semantic attributes. Traditional music information retrieval (MIR) approaches focused on 'hand-crafted' features, emphasizing explicit knowledge-based feature engineering [SGU*14]. The advent of deep learning shifted the attention towards automatic feature extraction, demonstrating strong performance in capturing complex musical characteristics without extensive domain knowledge [MBN*22]. Notable benchmarks such as the Million Song Dataset (MSD) and Spotify Web API highlight their utility in research, bridging content and context-based music information [BEWL11; Ski16]. However, concerns regarding the reproducibility, explainability and open research when using proprietary data (like features from Spotify) have motivated us to focus on alternative, open-source features for music representation.

Latent variable models have revolutionized the representation of music by learning abstract features that encapsulate the inherent characteristics of musical pieces. These models, especially Convolutional Neural Networks (CNNs), Variational Autoencoders (VAEs), and Transformers, have facilitated a broad range of MIR tasks, including genre classification, music recommendation, and emotion recognition [HE10; KW13; VSP*17]. The transition towards end-to-end learning models marks a significant shift from traditional feature engineering, enabling more nuanced and comprehensive understanding of music data. Our work leverages these advancements to derive latent representations that serve as a foundation for music visualization.

Visualization plays a pivotal role in music discovery, enabling users to navigate and explore music collections intuitively. Early efforts by Kolhoff et al. [KPL08] introduced music icons, utilizing parameterized glyphs to represent music features visually. Subsequent research has expanded on this concept, exploring various visual mappings and interactive interfaces for music exploration [LRFN04; SAA*05]. Yet, challenges remain in designing visual representations that effectively convey the complex nature of music features while supporting user-friendly exploration. We build upon this foundation and propose a novel glyph-based framework, using deep learning-derived features, coupled to specialized visualization and interaction techniques to enhance music discovery.

3. Our Approach

Our research is positioned at the intersection of MIR and visualization, where we propose a novel strategy to enhance user-guided music discovery in a large database through visual representation. To achieve this goal, we introduce an innovative glyph design by leveraging advanced latent features from the MIR model to provide immediate, intuitive insights into the music's characteristics. We also propose interaction mechanisms to not only facilitate a more efficient and engaging music discovery experience but also address the needs of users exploring music collections, looking for new sounds that match their reference ideas. Figure 1 shows an overview of our solution.

In this section, we will first discuss the feature extraction. Then, we will map this high-dimensional representation to a lower dimensional space to reduce the degrees of freedom of the information on a star-glyph representation, whose design choices are explained in the following. In the next section, we will then present our interface that builds upon this song representation.

3.1. Feature extraction

We reviewed state-of-the-art research in music representation learning, focusing on papers that provide full code and trained weights due to the challenges of training. We reviewed 23 papers, of which ten presented pre-trained models suitable for downstream tasks. However, recommendation models were either multimodal [Mar17; CLMG21] or lacked code and weights [STML21; VDS13]. Out of all options, we did identify the CLMR [SB21]

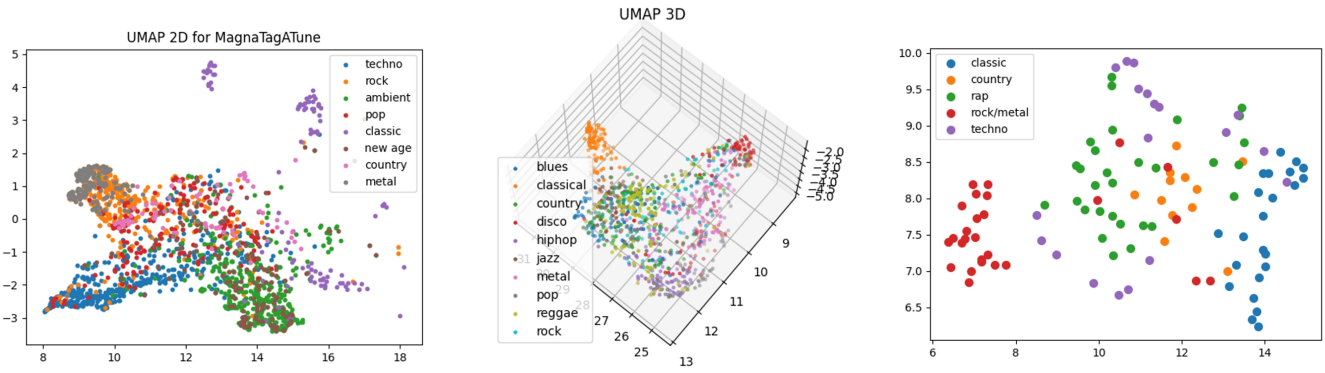

Figure 2: *UMAP embeddings of features extracted for genre classification on: MagnaTagATune (left), GTZAN (middle), and custom dataset (right).*

model as particularly suitable for our task, as we also show via an analysis below. The CLMR model is an adaptation of the very effective SimCLR model [CKNH20], which was developed for contrastive learning of visual representations. Contrastive learning is an unsupervised representation learning technique with the objective to maintain similarities and dissimilarities between data points in the representation space. CLMR shows excellent performance, is well documented, and lightweight to run. The network has learned a representation of 512 dimensions over an input sample of about 2.6 seconds. The model was trained for the downstream task of classification on the MagnaTagATune dataset [LWM*09]. We use the model and weights as provided by the authors. Representations over longer segments are averaged.

We verify the effectiveness of the CLMR representations by examining feature clusters to genre classification on multiple datasets and comparing feature embeddings to the Spotify features.

Evaluation of CLMR model To ensure the suitability of the model, we assessed its features for genre classification across three datasets: MagnaTagATune, GTZAN [TC02], and a small custom dataset. The custom data is composed of several albums that are considered iconic for various genres, details are given in Table 1. Each dataset provided genre labels. By extracting features and employing the UMAP algorithm [MHM18] for 2D spatial embedding, color-coding by genre revealed that genre-based clusters closely matched the feature-based clusters. Figure 2 plots the 2D embeddings of these three datasets, which suggests the captured features reflect high-level conceptual similarities across genres.

Comparison to Spotify features To explore how the selected representations compared with Spotify's own feature metrics, we created a dataset of 10K data points which are selected from the over 1.2 million entries in the Spotify Dataset reported by Figueroa et al. [Fig23]. For each data point, we obtained a 30-second MP3 sample through the Spotify API.

From these data points, we used the CLMR model to extract their neural features whose dimensionality was further reduced to 2D for visualization using the UMAP algorithm. Colors were assigned based on Spotify feature values such as acousticness, energy, valence, loudness, danceability and instrumentalness, to facilitate intuitive interpretation of the distribution. The corresponding feature results are shown in Figure 3. The results show a clear correlation between the CLMR-derived features and Spotify's features. This demonstrates that the selected model effectively captures musical qualities that align with industry-recognized attributes.

Genre	Artist	Album name
classic	Bach	A Musical Genius
classic	Vivalidi	The Four Seasons
rap	Eminem	The Eminem Show
rap	Nas	Illmatic
rock/metal	Nirvana	Nevermind
rock/metal	Slipknot	Iowa
techno	Paul Kalkbrenner	Berlin Calling
techno	Vitalic	Rave Age
country	Waylon Jennings	Dreaming My Dreams
country	Willie Nelson	Red Headed Stranger

Table 1: *Composition of the custom test dataset with two iconic albums for each genre.*

3.2. Feature Dimensionality Reduction

The CLMR model provides a representative neural vector of 512 elements. We perform a comparative evaluation of various dimensionality-reduction methods to find a suitable approach to reduce the number of features. Several works [FIB*14; DIPJ21; HZLY22] demonstrate that employing fewer but more representative dimensions than the original high dimensional vector in star glyphs greatly improves not only the computational efficiency but also their effectiveness across a spectrum of tasks. Aiming to retain the comprehensive nature of the data, we reduce the dimensionality to eight, which might seem arbitrary but is well motivated when opting for a star-shaped glyph, which will be discussed in Section 3.3.

We focus on identifying a method to effectively preserve data clustering and explored five algorithms for their ability to main-

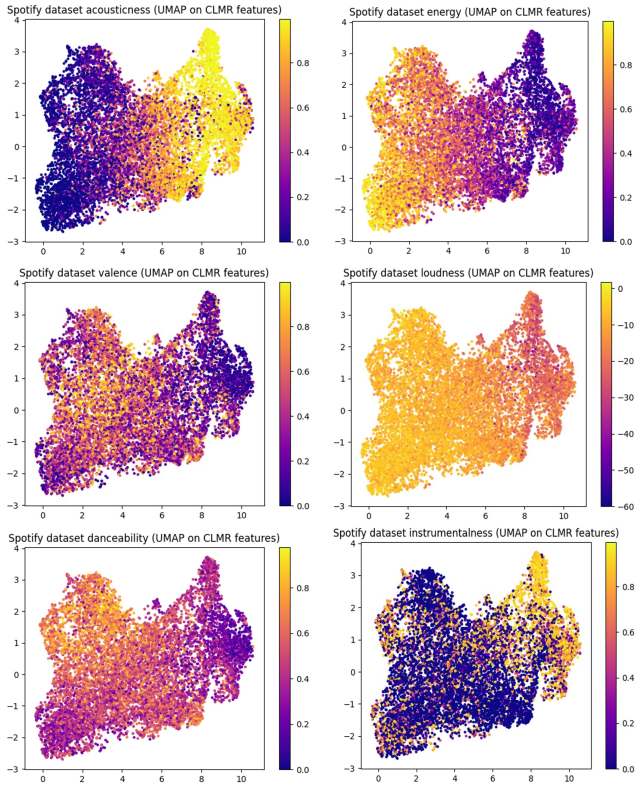

Figure 3: *Examples of 2D UMAP embeddings of extracted CLMR-derived neural features colorized with different Spotify features: acousticness (left in the 1st row), energy (right in the 1st row), valence (left in the 2nd row), loudness (right in the 2nd row), danceability (left in the 3rd row), and instrumentalness (right in the 3rd row).*

tain data point similarities with its original high-dimensional representation: PCA [Pea01], t-SNE [VH08] and UMAP,and two other more recent algorithms developed based on t-SNE and UMAP, respectively: TriMap [AW19] and PaCMAP [WHRS21].

To evaluate the preserved similarity of the reduced 8-dimensional space from the original 512-dimensional space, we calculated cosine similarity matrices for both spaces. We present the detailed statistics of the cosine similarities with the five assessed methods in Table 2.

Among the algorithms, t-SNE performed below our expectations, possibly the projection to an eight-dimensional space was off its optimal usage scenario. UMAP and PCA showed substantially better performance than TriMap and PaCMAP. Due to its non-linearity UMAP, outperforms PCA in maintaining data point similarities, while having acceptable computation cost, which made it our choice. Optimizing UMAP's hyperparameters (nearest neighbors=15, minimum distance=0.2 for normalized embeddings) further improved the results.

Method	Mean	Median	Std	Min	Max	Time/s
PCA	0.952	0.975	0.065	0.013	0.991	3
t-SNE	0.179	0.185	0.084	-0.161	0.373	4523
UMAP	0.962	0.963	0.009	0.889	0.983	49
PaCMAP	0.129	0.163	0.127	-0.257	0.335	28
TriMap	0.127	0.173	0.138	-0.273	0.352	38

Table 2: *Statistics of the kept similarity for feature vectors reduced from 512 to eight dimensions for five different algorithms.*

3.3. Glyph design

Glyphs, often composed of various geometric elements and visual channels, are capable of encoding multiple data dimensions simultaneously [Mun14]. This characteristic makes them particularly suitable for high-dimensional data visualization [KKG*20] and tabular data representation [BKH21]. Despite their utility, the design space for glyphs remains vast and largely unexplored [BKC*13]. Owing to its simplicity, versatility, and effectiveness in encoding multivariate observations and facilitating visual data comparison, we rely on the star glyph [Fri91; FIB*14; KE22]. To enhance expressiveness and ease of comparisons, we opted for a contour plot instead of a whisker plot, as suggested by [Pal99]. Hereby, shape is incorporated as a significant feature. Our design considered various elements, dimension ordering, categorization via colors, and shape emphasis by curvature, where some redundant encoding further eases shape distinction. Here, we detail our design decisions.

Dimension ordering The arrangement of variables on the axes of star glyphs significantly influences their shapes. The same datapoint can be displayed as very different shapes when it is mapped in different orders according to Klippel et al. [KHW09]. Finding the best order for variables is complex and depends on the analysis goal, but they observed that showing major differences on the main axes helps users identify shapes more quickly. Consequently, we sort the dimensions according to variance of the data.

Figure 4 shows the effect on our final glyphs before and after sorting the dimensions for seven heavy metal songs. They look very alike before and quite distinct after reordering.

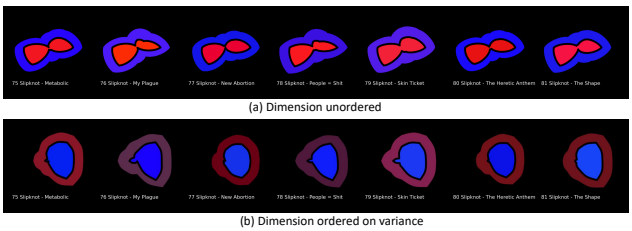

Figure 4: *Comparison of unsorted (upper) and sorted (bottom) the axes of our glyph on variance. It is very hard to detect differences among the icons without axis sorted. In the bottom raw, after sorting the axis, we can detect small differences between the icons.*

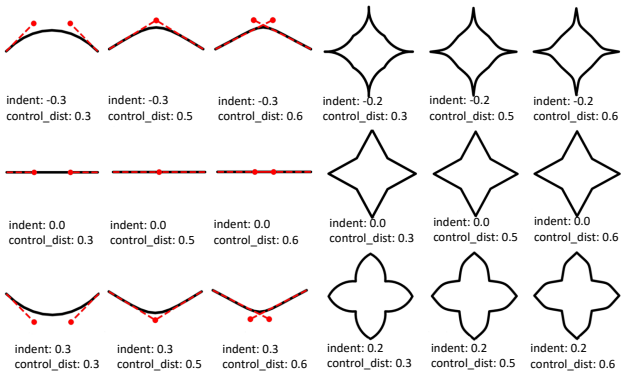

Figure 5: *Influence of curve parameters: The x-axis varies the control distance, while the y-axis adjusts the direction and strength. The left image demonstrates the impact of parameter adjustments on curve construction, and the right image illustrates how curve parameters affect an 8-dimensional star shape.*

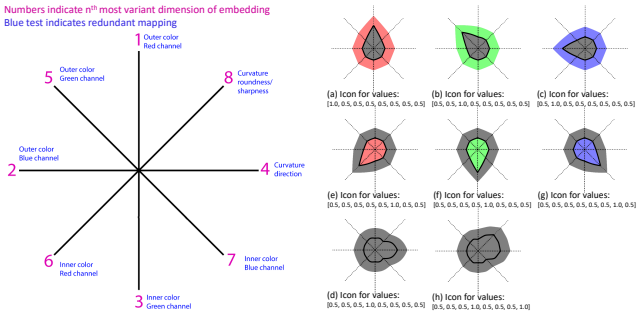

Figure 6: *The proposed variable mapping order along the axis of the star glyph (left) and the influence of each parameter with redundant encoding (Initially all 0.5, then we vary one parameter at a time (right).*

Dual colour As color can be a good indicator for distinguishing categorical data [Mun14], we use color prominently for the most distinctive dimensions, hereby making the implicit classification explicit. We adopted the automatic color mapping of Kolhoff et al. [KPL08], which assigns six features to two RGB colors. Originally, we considered the first six dimensions, but for reasons explained below ('Curvature'), we chose the first three dimensions and dimensions five through seven to be mapped to RGB components. This is shown in Figure 6 (left). We designed a distinct inner and outer glyph shape to receive these two colors, by superposing the glyph at different scales. The scale is chosen to ensure that both areas are balanced. The color mapping results can be seen in Figure 6 (right).

Curvature As curvature is considered a pre-attentive visual stimulus [BKC*13], we leverage it to expand the variety of shapes and increase the expressiveness of our proposed glyph. While Klippel et al. [KHW09] noted that distinctive shape features can speed up classification, they also cautioned that strong changes of shape may lead to a wrong impression of dissimilarity. For this reason, we chose to map dimensions four and eight, which were not yet redundantly encoded by color.

Specifically, the 4th dimension dictates the curvature's direction and strength. The 8th dimension, with the least variance, determines the positioning of control points relative to the line segment endpoints. An illustration of the influence of these two parameters on a curve can be seen in Figure 5 (left).

A limitation of this mapping is that when the 4th dimension is close to zero, the impact of the 8th dimension is less noticeable. Given that the eighth dimension has the lowest variance, we consider this an acceptable shortcoming. An illustration of this singularity can be seen in Figure 5 (right), where the effect of the curvature settings can be viewed on an 8D star shape.

To ensure clarity and prevent overly complex shapes, we employed an intersection detection method to adjust curvature strength and avoid intersecting lines, ensuring a coherent and interpretable glyph design. We confined curvature direction and strength to a range of [-0.3, 0.3] and control point distances to [0.3, 0.6]. This configuration range avoids shapes that appear to have more line segments than intended.

Redundant Encoding As indicated, our glyph design contains redundant coding [Fuc15] by design. As the ideal mapping of data to star glyphs remains an open problem, encoding the important variables multiple times can significantly enhance glyph expressiveness and distinctiveness, bolstering visual search effectiveness. In our case, it also makes the design suitable for individuals with color vision deficiency.

Our scheme employs a dual encoding for each variable. It is always represented as a component of the glyph's shape, but also as a glyph's color or curvature. To judge the effectiveness of the resulting glyphs, we illustrate several representations in Figure 6(left) and show how each parameter influences the glyph in Figure 6 (right). The design is carefully evaluated in Section 5.

4. Interface

We introduce a new search interface that incorporates our glyph definition and interaction features to support the user. It has been built as a web application, accessible at `http://musicons.io/`. Involving a test database of 10K song snippets that were randomly selected from the Spotify Dataset provided by Figueroa [Fig23], which contains more than 1.2M song samples. The principles of our glyph-based design are inherently flexible and can be adapted to mobile platforms. Nevertheless, we first chose a desktop interface for our Musicon system to better control its evaluation. Typically, the use of a desktop system ensured a larger display area and precise input capabilities for a detailed glyph-based visualization and interaction. It also facilitated the user study, as measuring user engagement and satisfaction was eased, as we could assume that people are familiar with the hardware. Designing touch-friendly controls and effectively managing visual complexity on a small screen of a mobile device remains promising future work.

4.1. Customized icon

For an initial expression, we randomly selected 48 icons from the results. As can be seen in Figure 7, the icons reveal the intended expressiveness and display a wide variety of shapes and colors.

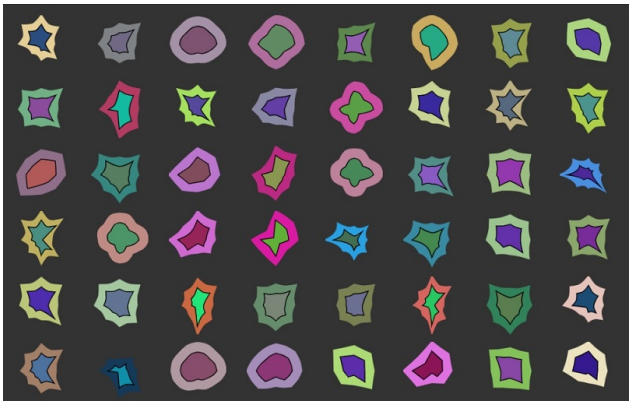

Figure 7: *Customized icons: 48 randomly selected icons from the results.*

Figure 8: *The interface for search-by-icon. The icon parameters are indicated with a dotted line and the icon is updated in real-time when the parameters are changed with the sliders (left). While the users drags the slider, the most similar songs are immediately updated and allow to be listened(right).*

4.2. Search-by-icon

The search-by-icon tool allows users to start with the icon of a preferred song and then explore other songs in the proximity, akin to a reverse search mechanism. This concept draws inspiration from Knees et al. [KA16], who explored audio search through the visualization of sound mental images.

The interface for search-by-icon can be seen in Figure 8. It features eight adjustable sliders, corresponding to the glyph dimensions. The representation is updated in real time, as is the list of the most similar songs, which enables users to explore the music space interactively.

We use the cosine similarity to the user-generated icon and a comparison between the input and 10K vectors of songs in the database only takes milliseconds. It is done whenever sliders are adjusted. The ten most similar songs are then presented to the user.

4.3. Playlist sorting

Kolhoff et al. [KPL08] introduced two sorting methods, 1D and 2D, by applying PCA on icon parameters. We explored various approaches within our dataset and found that UMAP embeddings surpassed PCA in performance.

Specifically, we initiated our process by converting the dataset into an 8-dimensional (8D) UMAP embedding, subsequently transforming the resulting data into a 1-dimensional (1D) UMAP embedding. Although a 2-dimensional (2D) layout was considered, it introduced distortions when attempting to achieve the compactness of a 1D layout. Furthermore, the 1D layout better aligns with the format of song lists familiar to users of streaming applications, thus leading us towards a 1D embedding. Utilizing the 8D embedding initially maintains greater coherence with the space utilized for our glyphs, as demonstrated in Figure 9, and reduces computational workload relative to the original 512 dimensions. While sorting within diverse playlists produces variable sequences, our method consistently positions similar icons in close proximity. This arrangement enhances user navigation and experience, outperforming standard PCA arrangements.

4.4. Enhancing icon contrast

A global embedding will enhance inter-genre distinctions, while reducing intra-class variations. To amplify local contrast of icons when exploring, for example, homogeneous playlists of similar songs, we employed a min-max scaling to re-normalize the icon features' range of the selected subset to $[0,1]$. Furthermore, users are offered the option to tune the percentage of contrast linearly between the original and fully normalized embeddings. This maintains a link to the global representation and allows for applying a gradual contrast. Figure 10 illustrates the effectiveness of this feature in a subset of jazz music, by gradually increasing the local contrast from 0% to 100%. As an example, the song zero and the song four in this list exhibit very high similarity in audio data - both share instruments, mood, and a prominent solo with the same instrument. At overview scale, the icons appear almost identical, as expected. By tuning up the contrast, the resolution of the icon parameters is locally amplified and the difference becomes more apparent.

5. Evaluation

Given the subjective nature of visualizing and perceiving music, our hypothesis and experimental design are evaluated through an exploratory user study. We aim to conduct two types of evaluations: a comparison with alternative methods and an assessment of the proposed features.

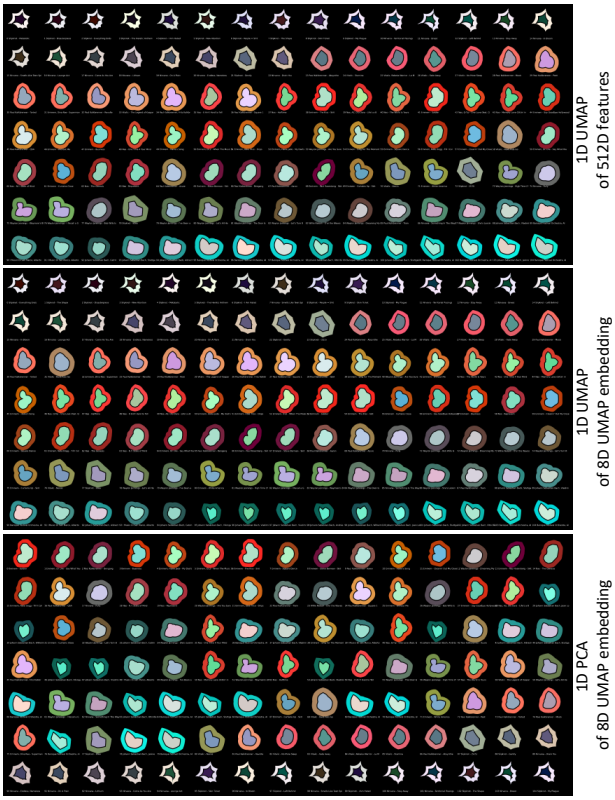

Figure 9: *Comparative visualization of three sorting methods. From upper to bottom: icons sorted on 1D UMAP embedding of original 512D features, icons sorted on 1D UMAP embedding of the 8D UMAP embedding of the 512D features, and icons sorted on 1D PCA of the 8D UMAP embedding of the 512D features.*

We wanted to compare our work with Kolhoff et al. [KPL08], who introduced content-based music icons. Unfortunately, after contacting the authors, we learned that the music resources and implementation are no longer available. Hence, we can only informally compare the two solutions. Our work benefits from deep features, a more detailed mapping of information on the glyph shape, as well as an improved robustness and inclusivity for color vision deficiency (CVD) by redundant encoding. Furthermore, we introduce a reverse search, which broadens the utility of music icons.

We target the evaluation of our proposed features: from the effectiveness of the icon to a larger system-evaluation. We used the 10K dataset due to its availability, although the system can be expected to handle larger song databases with minor optimizations in the implementation.

We targeted a participant demographic of 'non-expert but generally computer-literate' adults [CW11]. To reduce response bias, participation was anonymous. Using an a-priori sample size calculator with an expected medium effect size (d = 0.5), we determined that a minimum of 27 participants would achieve a statistical power of 0.8 and a significance level of 0.05, assuming analysis via paired samples t-test for certain tasks. We garnered 38 responses in the

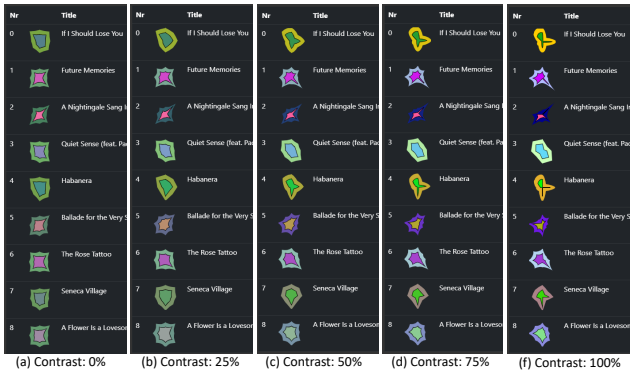

Figure 10: *Local-contrast icons for a playlist of jazz music with contracts of* 0%*,* 25%*,* 50%*,* 75%*, and* 100%*.*

evaluation, with a roughly equal number of men and women and an age distribution ranging from 20-29 to over 70 years. To conduct the study, we developed a web interface that provided each participant with task instructions, managed the flow and timing, randomized the order of tasks, and collected the data. The study was conducted remotely to minimize the response bias caused by our presence. Additional details regarding the evaluation are given in the supplementary material.

5.1. Customized icon

Visual clustering. This is a classic 'free-grouping' or 'free-sorting' task, widely used in the field of psychology [BB16]. This test assesses the effectiveness of the icons in representing feature similarity and the extent of user consensus on this aspect. Participants formed clusters from 60 icons based solely on visual cues, without song titles or additional metadata information. They were allowed to use any number of clusters and set aside non-fitting icons.

To ensure that there is a diversity in the selection yet still the possibility to make clusters, we sampled 10 data points from six of the clusters created by applying a k-means clustering algorithm on the original 512 dimensional embedding ($k = 10$). Each participant worked with the same set of icons but their presentation was in a random order.

To assess user consensus on clustering, we computed a co-occurrence matrix of participant-generated clusters and a cosine similarity matrix of the feature vectors, both shown in Figure 11. Initial observations suggest a strong user agreement on the clusters, with the co-occurrence matrix displaying notable resemblance to the similarity matrix. To quantitatively evaluate this relationship, we calculated the pairwise Pearson correlation coefficient, resulting in a value of 0.6. This indicates a moderate linear correlation, suggesting a reasonable level of agreement among users in their clustering decisions.

Outlier Detection. This test, extending from the visual clustering test, utilizes participant-generated clusters to evaluate the glyphs'

effectiveness in representing music similarity and facilitating outlier detection. For each participant, we randomly selected four songs from one cluster and one song from a different cluster, presenting these five songs in a random order. Participants were then asked to identify the song that sounded distinct from the others. This process was repeated three times for each participant.

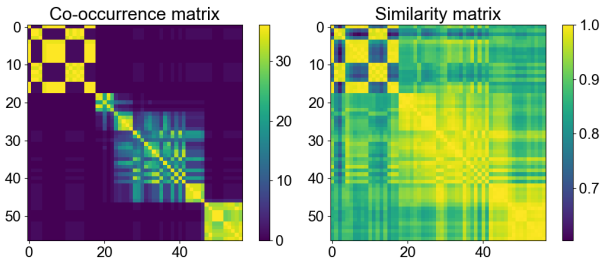

Figure 11: *Co-occurrence matrix of the clusters made by participants (left) and a cosine similarity matrix of the feature vectors (right).*

The descriptive statistics of the results is shown in Table 3. Given that random guessing would yield an expected recognition rate of 0.2, our observed mean recognition rate of 0.7451 represents a considerable enhancement. With a p-value lower than 0.00001 and an effect size of 2.375 (Cohen's d), this improvement is statistically substantial.

Mean	Median	Mode	Std	Variance
0.745	0.667	0.667	0.230	0.053

Table 3: *Descriptive statistics of the recognition rates obtained for outlier detection.*

Generalization, Contrast and CVD Robustness. This test is set up as a matching-to-sample task in the same manner as [FIB*14]. It aimed to assess three aspects: the alignment of the 'most similar' icon with the data point of highest cosine similarity, indicating the icon's effectiveness in representing high-dimensional data; the impact of a contrast-enhanced icon version on task completion time and accuracy in identifying the most similar icon; and the icon design's robustness against CVD, evaluated through task time and accuracy using a CVD simulation.

Participants were shown nine similar icons, including one target icon, and asked to identify the icon that is most similar to the target. This test was conducted across three rendering modes: the default design, a contrast-enhanced version with 100% contrast, and a color-blind mode simulating deuteranomaly the most common type of CVD. Each participant completed the task nine times, three times per rendering mode. The descriptive statistics of recognition rates obtained for these three modes is shown in Table 4.

With a random selection, the expected recognition rate is merely 0.125. Our study, however, demonstrates a marked enhancement with mean recognition rates more than 0.706 for all three modes. This substantial increase suggests that participants are better at identifying the icon that most accurately corresponds to a position within an 8-dimensional space. To validate these findings, we employed as 'default' mode a one-sample, one-tailed t-test, which indicated an essentially zero p-value and a pronounced effect size of 2.100 (Cohen's d). These results robustly confirm the effectiveness of our approach.

Mode	Mean	Median	Mode	Std	Variance
Default	0.706	0.750	1.000	0.277	0.077
Contrast	0.785	1.000	1.000	0.271	0.074
CVD	0.741	0.667	0.667	0.231	0.053

Table 4: *Descriptive statistics of recognition rates obtained for matching-to-sample with the 'default', 'contrast', and CVD mode of the icon.*

We observed that 'contrast' mode outperforms 'default' mode with a higher mean recognition rate and a faster selection process, while, for the CVD mode, we discovered no obvious differences between the 'default' and CVD modes. A one-way ANOVA test conducted across all three rendering modes yielded a p-value of 0.450, indicating that rendering mode has no noticeable impact on performance in the matching-to-sample task. This outcome suggests that each icon rendering mode performs comparably well, affirming the robustness of our icon to CVD. The effectiveness of our redundant encoding strategy in enhancing recognition and matching accuracy is thus supported by these results.

5.2. Search by icon

The test aimed to assess the efficacy of our 'search-by-icon' method for users. Participants were shown a target song with its custom icon and the search-by-icon interface shown in Figure 8. They were tasked with using the interface to imitate the target icon and then retrieve the three songs most similar to the target one. Following the test, participants completed the System Usability Scale [CW11] to evaluate their experience with the interface.

Imitated icon and retrieved songs Cosine similarities between user-generated and target icon vectors, presented in Figure 12 (left), with a high mean (0.989) and median (0.993), indicates that with vectors exhibiting a cosine similarity above 0.975 to the target, most users accurately replicated icons. Furthermore, the average cosine similarities between the target icon vector and the top three selected songs, detailed in Figure 12 (right), reinforce the precision of these imitations, highlighting the effectiveness of participant selections in aligning closely with the target icons.

System Usability Scale (SUS) The SUS comprises of ten statements, each evaluated using the Likert Scale, which ranges from one for 'strongly disagree' to five for 'strongly agree'. The details of this dataset can be found in the supplementary material.

Based on the feedback, we computed the SUS scores, as illustrated in Figure 13(left), with the corresponding performance interpretations presented in Figure 13 (right). We counted 13 'bad' results, 7 'mediocre', 11 'good' and 6 'excellent'. We recognize

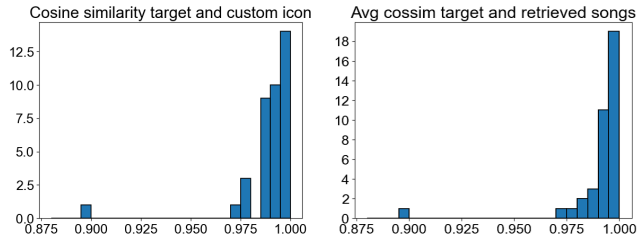

Figure 12: *Cosine similarities between user-generated and target icon vectors (left), illustrating how closely users can imitate an icon. The average cosine similarities between the target icon vector and the top three selected songs (right), evaluating how effectively users can retrieve similar music using this tool.*

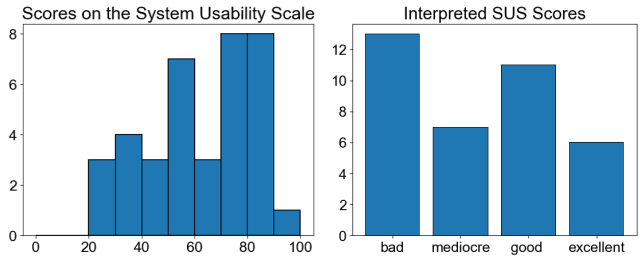

Figure 13: *SUS Scores (left) and its corresponding interpretation (right). A score above 80.3 is interpreted as 'excellent', scores between 68 and 80.3 as 'good', scores between 50 and 68 'mediocre', and anything below 50 'bad'.*

that flattening the user experience into such a score is a gross simplification. Nonetheless, we observe that a majority, specifically 25 out of 38 participants, demonstrates a willingness to embrace our model.

5.3. Search by playlist

This test aimed to assess the icon's effectiveness and sorting properties within a playlist context, comparing it against album art, which is typically used in streaming services. Participants were asked to select their top three songs from playlists featuring both album art and our custom design, with each format presented twice. We assessed the similarity between the top three selections and the target vector, time-on-task, plays per task, and additional insights from open-ended questions.

Retrieved songs Average cosine similarities between the target icon vector and the top three selected songs are presented in Figure 14, with descriptive statistics in Table 5. Both methods cover similar ranges of cosine similarities, but our method facilitates slightly higher similarity retrieval (one-tailed paired-samples t-test: p = 0.03, Cohen's d: 0.469), aligning with the icon's intended similarity representation.

Time-on-task The time-on-task per participant for both album art and custom icon methods are detailed in Figure 15, with descriptive statistics provided in Table 6. A notable reduction in average

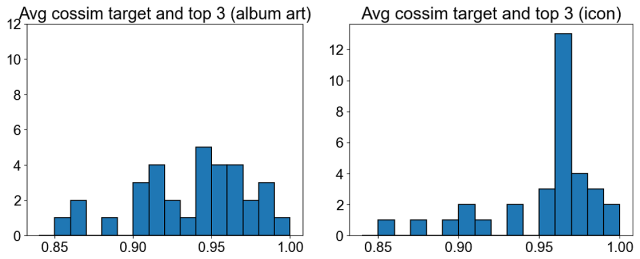

Figure 14: *Average cosine similarities between the target icon vector and the top three selected songs, comparing between album art (left) and our custom icon (right).*

Icon	Mean	Median	Std	Var	Min	Max
Album	0.955	0.967	0.033	0.001	0.858	0.991
Custom	0.938	0.945	0.036	0.001	0.851	0.993

Table 5: *Descriptive statistics of the data as displayed in Figure 14.*

completion time, exceeding one minute, was observed. A left-tailed paired t-test confirmed these findings with p = 0.00201 and an effect size of 0.473 (Cohen's d).

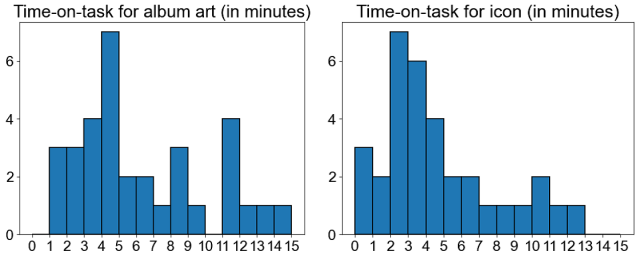

Figure 15: *Completion times of time-on-task with album art (left) and our custom icon (right).*

Icon	Mean	Median	Std	Min	Max
Album	6:25	4:49	3:45	1:08	14:12
Custom	4:44	3:54	3:11	0:39	12:14

Table 6: *Descriptive statistics of the data as displayed in Figure 15, formatted as mm:ss.*

Songs played per task Figure 16 and Table 7 display the number of songs played per task per participant for both album art and the custom icon, showing similar ranges but a notably lower mean and median for the custom icon. A paired t-test confirms this difference, with p = 0.00001 and an effect size of 0.931 (Cohen's d).

Icon	Mean	Median	Std	Var	Min	Max
Album	103.9	113.5	55.9	3121.1	10	222
Custom	57.7	40.0	42.4	1796.8	7	192

Table 7: *Descriptive statistics of the data as displayed in Figure 16.*

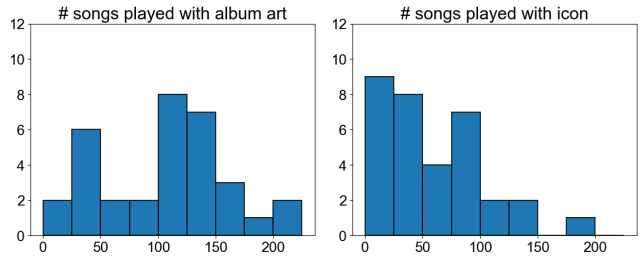

Figure 16: *The number of songs played per task per participant for both album art (left) and the custom icon (right).*

Open Questions Upon study completion, participants responded to three open-ended questions regarding their playlist task experience. The detailed questions and corresponding analysis can be found in the supplementary material.

Our tool's effectiveness was confirmed through strong quantitative results, notably speeding up task completion by over a minute compared to album art presentations and reducing the number of songs participants needed to listen to by almost 50%. When using our icon, selections tended to have similar or slightly higher cosine similarity to the target song, suggesting the icons' visual cues enhanced both the speed and quality of decision-making.

While many participants valued our icon for its capacity to indicate similarity, some expressed a preference for album art due to its contextual and cultural insights. Acknowledging album art's value in certain situations, we argue that our icon meaningfully enhances user experience by addressing the variability of songs within an album. Further, our evaluation focused on the effectiveness of our contributions. In a practical system, we would envision the use of our icon in conjunction with potential metadata (including album art) if available.

An exciting direction for future work would be exploration for semantic latent representations and descriptive dimension mappings. Unfortunately, this is not straightforward, as it is unlikely that all aspects could be well captured in this way, e.g., what dimension would mean "adding a piano" or "adding strings"? Nevertheless, some meta information could be used to augment our solution. In a practical system, we would certainly argue for keeping genre information available for the user in selections and search. We refrained from doing so, to show the effectiveness of the automatically derived features and our visualization solution, which is already able to capture a considerable amount of information.

6. Conclusion

In this paper, we introduced a novel icon-based visualization approach for enhancing music discovery in streaming services by mapping latent characteristics of music to a novel glyph design. Our evaluation demonstrated that our method meaningfully improves user engagement and efficiency in music exploration, highlighting the potential of incorporating such visual cues into streaming platforms. The positive outcomes suggest a promising direction for further research in music visualization and user-interface design to refine and personalize the music-discovery process. Future development in this intersection of music information retrieval and visual interaction could potentially transform user experiences in digital music environments.

References

[AMA*20] ANDERSON, ASHTON, MAYSTRE, LUCAS, ANDERSON, IAN, et al. "Algorithmic effects on the diversity of consumption on spotify". *Proceedings of the web conference 2020*. 2020, 2155–2165 1.

[AW19] AMID, EHSAN and WARMUTH, MANFRED K. "TriMap: Large-scale dimensionality reduction using triplets". *arXiv preprint arXiv:1910.00204* (2019) 4.

[BB16] BLANCHARD, SIMON J and BANERJI, ISHANI. "Evidence-based recommendations for designing free-sorting experiments". *Behavior research methods* 48 (2016), 1318–1336 7.

[BEWL11] BERTIN-MAHIEUX, THIERRY, ELLIS, DANIEL P.W., WHITMAN, BRIAN, and LAMERE, PAUL. *The Million Song Dataset*. http://millionsongdataset.com/. Accessed: June 13, 2020. 2011 2.

[BKC*13] BORGO, RITA, KEHRER, JOHANNES, CHUNG, DAVID HS, et al. "Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications." *Eurographics (state of the art reports)*. 2013, 39–63 4, 5.

[BKH21] BREHMER, MATTHEW, KOSARA, ROBERT, and HULL, CARMEN. "Generative design inspiration for glyphs with diatoms". *IEEE Transactions on Visualization and Computer Graphics* 28.1 (2021), 389–399 4.

[Cho19] CHODOS, ASHER TOBIN. "What does music mean to Spotify? An essay on musical significance in the era of digital curation". *INSAM Journal of Contemporary Music, Art and Technology* 2 (2019), 36–64 1.

[CKNH20] CHEN, TING, KORNBLITH, SIMON, NOROUZI, MOHAMMAD, and HINTON, GEOFFREY. "A simple framework for contrastive learning of visual representations". *International conference on machine learning*. PMLR. 2020, 1597–1607 3.

[CLMG21] CHEN, KE, LIANG, BEICI, MA, XIAOSHUAN, and GU, MINWEI. "Learning audio embeddings with user listening data for content-based music recommendation". *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, 3015–3019 2.

[CW11] CUNNINGHAM, DOUGLAS W and WALLRAVEN, CHRISTIAN. *Experimental design: From user studies to psychophysics*. CRC Press, 2011 7, 8.

[DIPJ21] DY, BIANCHI, IBRAHIM, NAZIM, POORTHUIS, ATE, and JOYCE, SAM. "Improving Visualization Design for Effective Multi-Objective Decision Making". *IEEE Transactions on Visualization and Computer Graphics* 28.10 (2021), 3405–3416 3.

[FIB*14] FUCHS, JOHANNES, ISENBERG, PETRA, BEZERIANOS, ANASTASIA, et al. "The influence of contour on similarity perception of star glyphs". *IEEE transactions on visualization and computer graphics* 20.12 (2014), 2251–2260 3, 4, 8.

[Fig23] FIGUEROA, RODOLFO. *Spotify 1.2M+ Songs*. https://www.kaggle.com/datasets/rodolfofigueroa/spotify-12m-songs/data. Accessed: Dec 15, 2023. 2023 3, 5.

[Fri91] FRIENDLY, MICHAEL. "Statistical graphics for multivariate data". *SAS SUGI* 16 (1991), 1157–1162 4.

[Fuc15] FUCHS, JOHANNES. "Glyph design for temporal and multidimensional data: Design considerations and evaluation". (2015) 5.

[HE10] HAMEL, PHILIPPE and ECK, DOUGLAS. "Learning features from music audio with deep belief networks." *ISMIR*. Vol. 10. Citeseer. 2010, 339–344 2.

[HVS*19] HOSEY, CHRISTINE, VUJOVIĆ, LARA, ST. THOMAS, BRIAN, et al. "Just give me what I want: How people use and evaluate music search". *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, 1–12 1, 2.

[HZLY22] HOU, YIHAN, ZHU, HAOTIAN, LIANG, HAI-NING, and YU, LINGYUN. "A study of the effect of star glyph parameters on value estimation and comparison". *Journal of Visualization* (2022), 1–15 3.

[KA16] KNEES, PETER and ANDERSEN, KRISTINA. "Searching for audio by sketching mental images of sound: A brave new idea for audio retrieval in creative music production". *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. 2016, 95–102 6.

[KE22] KECK, MANDY and ENGELN, LARS. "Sparkle glyphs: A glyph design for the analysis of temporal multivariate audio features". *Proceedings of the 2022 International Conference on Advanced Visual Interfaces*. 2022, 1–3 4.

[KHW09] KLIPPEL, ALEXANDER, HARDISTY, FRANK, and WEAVER, CHRIS. "Star plots: How shape characteristics influence classification tasks". *Cartography and Geographic Information Science* 36.2 (2009), 149–163 4, 5.

[KKG*20] KAMMER, DIETRICH, KECK, MANDY, GRÜNDER, THOMAS, et al. "Glyphboard: Visual exploration of high-dimensional data combining glyphs with dimensionality reduction". *IEEE transactions on visualization and computer graphics* 26.4 (2020), 1661–1671 4.

[KPL08] KOLHOFF, PHILIPP, PREUSS, JACQUELINE, and LOVISCACH, JÖRN. "Content-based icons for music files". *Computers & Graphics* 32.5 (2008), 550–560 2, 5–7.

[KW13] KINGMA, DIEDERIK P and WELLING, MAX. "Auto-encoding variational bayes". *arXiv preprint arXiv:1312.6114* (2013) 2.

[LRFN04] LEWIS, JOHN P, ROSENHOLTZ, RUTH, FONG, NICKSON, and NEUMANN, ULRICH. "VisualIDs: automatic distinctive icons for desktop interfaces". *ACM Transactions on Graphics (TOG)* 23.3 (2004), 416–423 2.

[LWM*09] LAW, EDITH, WEST, KRIS, MANDEL, MICHAEL I, et al. "Evaluation of algorithms using games: The case of music tagging." *ISMIR*. Citeseer. 2009, 387–392 3.

[Mar17] MARTÍN, SERGIO ORAMAS. "Knowledge Extraction and Representation Learning for Music Recommendation and Classification". PhD thesis. Ph. D. thesis, Universitat Pompeu Fabra.[Cited on page 139.], 2017 2.

[MBN*22] MÜLLER, MEINARD, BITTNER, RACHEL, NAM, JUHAN, et al. "Deep learning and knowledge integration for music audio analysis (Dagstuhl Seminar 22082)". (2022) 2.

[MHM18] MCINNES, LELAND, HEALY, JOHN, and MELVILLE, JAMES. "Umap: Uniform manifold approximation and projection for dimension reduction". *arXiv preprint arXiv:1802.03426* (2018) 3.

[Mun14] MUNZNER, TAMARA. *Visualization analysis and design*. CRC press, 2014 4, 5.

[Pal99] PALMER, STEPHEN E. *Vision science: Photons to phenomenology*. MIT press, 1999 4.

[Pea01] PEARSON, KARL. "LIII. On lines and planes of closest fit to systems of points in space". *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2.11 (1901), 559–572 4.

[SAA*05] SETLUR, VIDYA, ALBRECHT-BUEHLER, CONRAD, A. GOOCH, AMY, et al. "Semanticons: Visual metaphors as file icons". *Computer Graphics Forum*. Vol. 24. 3. Blackwell Publishing, Inc Oxford, UK and Boston, USA. 2005, 647–656 2.

[SB21] SPIJKERVET, JANNE and BURGOYNE, JOHN ASHLEY. "Contrastive learning of musical representations". *arXiv preprint arXiv:2103.09410* (2021) 2.

[SGU*14] SCHEDL, MARKUS, GÓMEZ, EMILIA, URBANO, JULIÁN, et al. "Music information retrieval: Recent developments and applications". *Foundations and Trends® in Information Retrieval* 8.2-3 (2014), 127–261 2.

[Ski16] SKIDÉN, PETTER. *API improvements and U*. https://developer.spotify.com/community/news/2016/03/29/api-improvements-update/. Accessed: June 15, 2020. 2016 2.

[STML21] SARAVANOU, ANTONIA, TOMASI, FEDERICO, MEHROTRA, RISHABH, and LALMAS, MOUNIA. "Multi-Task Learning of Graph-based Inductive Representations of Music Content." *ISMIR*. 2021, 602–609 2.

[SZC*18] SCHEDL, MARKUS, ZAMANI, HAMED, CHEN, CHING-WEI, et al. "Current challenges and visions in music recommender systems research". *International Journal of Multimedia Information Retrieval* 7 (2018), 95–116 2.

[TC02] TZANETAKIS, GEORGE and COOK, PERRY. "Musical genre classification of audio signals". *IEEE Transactions on speech and audio processing* 10.5 (2002), 293–302 3.

[VDS13] VAN DEN OORD, AARON, DIELEMAN, SANDER, and SCHRAUWEN, BENJAMIN. "Deep content-based music recommendation". *Advances in neural information processing systems* 26 (2013) 2.

[VH08] VAN DER MAATEN, LAURENS and HINTON, GEOFFREY. "Visualizing data using t-SNE." *Journal of machine learning research* 9.11 (2008) 4.

[VSP*17] VASWANI, ASHISH, SHAZEER, NOAM, PARMAR, NIKI, et al. "Attention is all you need". *Advances in neural information processing systems* 30 (2017) 2.

[WHRS21] WANG, YINGFAN, HUANG, HAIYANG, RUDIN, CYNTHIA, and SHAPOSHNIK, YARON. "Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization". *The Journal of Machine Learning Research* 22.1 (2021), 9129–9201 4.

[Wol10] WOLFE, JEREMY M. "Visual search". *Current biology* 20.8 (2010), R346–R349 2.