Literature Survey: MultiModal Prediction for Music Popularity

Yash Choudhary
CMInDS, IIT Bombay
`200100173@iitb.ac.in`

Preeti Rao
EE, ITT Bombay
`prao@ee.iitb.ac.in`

Pushpak Bhattacharyya
CSE, IIT Bombay
`pb@cse.iitb.ac.in`

Abstract

Hit-Song Science has matured into a multimodal research field that blends music-information retrieval, natural-language processing and user-behaviour analytics. This survey consolidates more than two decades of work and is organised around seven themes. We begin with the historical evolution of Hit-Song Science, tracing the shift from audio-only heuristics to integrated frameworks that model music as both an acoustic artefact and a social commodity. We then review the feature landscape—audio descriptors, lyric signals and social-metadata cues—and show how recent learned representations, such as neural audio tokens and transformer-based text embeddings, have superseded handcrafted features. A dedicated section examines musical-structure segmentation, comparing classical audio techniques, lyric-driven methods and emerging multimodal approaches. Next, we map the data requirements for scalable popularity modelling, outlining best practices for audio-, lyric- and social-feature extraction and for the construction of reliable popularity targets. We catalogue open-source music datasets, highlighting their modality coverage and licensing constraints, and analyse how dataset choice influences reported performance. Throughout, we identify four persistent gaps: structural lyric understanding, artist-career dynamics, expressive audio representations and interpretable fusion architectures. We conclude by outlining a research agenda that couples large language models, neural audio codecs and modality-aware learning to produce more accurate and explainable popularity predictors.

1 Introduction

In 2023, the global recorded music market generated \$28.6 billion[1] in revenue. With the advent of social media and streaming services, defining a single metric for music success has become increasingly challenging (Cosimato et al., 2019b; Lee et al., 2020). Music popularity prediction can help the industry and artists forecast and optimize the potential success of newly composed songs.

Research in music popularity prediction has been driven by the advancements in machine learning with researchers applying classical ML approaches to predict popularity using acoustic features, and further with the growth of social networks, information about music consumers' tastes capturing consumer response and their evolving music preferences (Seufitelli et al., 2023). Advancements in deep learning further sharpen the prediction model capability of capturing and learning complex patterns of evolving music taste, and researchers have worked on incorporating multiple modalities such as audio, lyrics and social metadata to predict song success (Zangerle et al., 2019b; Martín-Gutiérrez et al., 2020). In all these works, the popularity score is typically defined as the time the song remains on the Billboard Top charts, and the evaluation metrics used include MAE, MSE, R^2 for regression, and accuracy, precision, recall, and F1 for classification. Recent developments in large language models have led to further research in music-related fields such as recommendation systems, sentiment/emotion analysis, data augmentation, understanding and composing song lyrics, using song lyrics text as the data source (Rossetto et al., 2023; Sable et al., 2024; Ma et al., 2024; Ding et al., 2024). The next section presents a detailed review of prior studies.

2 Related Work

2.1 Hit Song Science: A MultiModal Paradigm

Interesting is the Hit Song Science—the artful fusion of the human imagination along with human perception with the accuracy of an algorithm whose

[1] IFPI Report '23

main motto is to forecast the success even before the song comes into the market. (Pachet and Sony, 2012) defined HSS as "an emerging field of science that aims at predicting the success of songs before they are released on the market." Such a definition leads to conceptualizing HSS not only as a technological project but as an ambitious project to condense the intricacies of human musical tastes and preferences into measurable features. Essentially, HSS employs machine learning algorithms for the automated predictive intent regarding the trajectory of popularity ahead of formal release for a song.

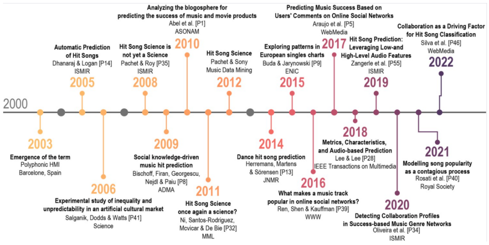


Figure 1: The timeline of Hit Song Science, in which too major works are in direction of Hit Song Prediction. Source: HSS Survey

HSS is typically investigated under two major lenses: prediction and science. From a predictive lens, HSS faces the multilayered nature of music consumption. Music has a highly personal dimension, bringing out micro-emotions that are usually linked to individual histories, contexts, and other things that cannot be analytically reduced to simple analysis. These evasive properties make the prediction task complex because decoding of patterns would transcend simple auditory characteristics. Rather, HSS posits, from the perspective of a scientific enterprise, these factors may be complex but would feature and model representability, thus closing the chasm between subjective emotional reaction and objective analysis.

This dual vision becomes essential in taking into account the chasm between the intangible quality of musical success and the analytical zeal that hankers after reifying it. For instance, (Ogihara and Kim, 2011) recommend that HSS should include psychology of listening to music, effects of repeated exposure, nature of Western media and broadcasting systems, and the strong social influences that are at work in communities. This gives an insight into even bigger dimensions of HSS, including psychological and sociological as well as technological aspects.

From a broader angle of view, Hit Song Science finds its place in the large family of Music Information Retrieval (MIR), which includes research activities concerned with what information regarding relevance can be found from music content. Considering HSS as a special task in MIR underlines its importance when shedding light on the interrelation between technology and musicology. The classic MIR research traces tasks like genre recognition and classification, which identifies the genre of a piece from the musical features ((Sturm, 2014); (Corrêa and Rodrigues, 2016)). But HSS surpasses genre categorization by relating intrinsic features of songs with the resultant popularity.

The HSS field in itself evolved to be multidisciplinary where it combines computer science with traditional music-related topics such as music theory, sociology, and cultural markets. This approach will engage the acquisition and analysis of music data with different modalities from diverse data sources using methodologies from information retrieval, machine learning, and data mining. Ultimately, the goal will be to detect chart-topping hits and predict them. The core premise of HSS revolves around the notion that popular songs share a number of characteristics in common which resonate well with the listeners, and such features could be found out and modelled through state-of-the-art algorithms.

2.1.1 Historical Development and Theoretical Inventions

The first phase of HSS can be traced back to the year 2003 when Polyphonic HMI presented an application of machine learning that was called Hit Song Science. This innovative work made use of mathematical algorithms as well as statistical methods for predicting a song's success based on audio features. According to the company, using the software could predict songs by such big artists such as Norah Jones, Jennifer Lopez, or Robbie Williams. Researchers and enthusiasts of this field used innovative applications in machine learning and mathematical modeling to break down millions of those hit songs long past into quantifiable features: rhythm, harmony, and timbre. This event has actually been quite a leap forward in the field, creating new roads for algorithmic music research.

Following the concept development, early research work in HSS was mainly viewed as feature extraction or analysis of song-related data using a combination of machine learning techniques. In these early attempts, they made use of different classification algorithms like the boosting classifier for classification purposes ((Dhanaraj and Logan, 2005)), adaptive learning approaches ((Chon et al., 2006a)), and Support Vector Machines ((Dhanaraj and Logan, 2005); (Pachet and Roy, 2008)). This improved access to better data processing and modeling techniques meant HSS methods emerged as strong, more complete frameworks for use. Researchers started exploring larger datasets with advanced algorithms in pursuit of deeper insights into the relationship between attributes of music and commercial success ((Herremans and Bergmans, 2020); (Yang et al., 2017); (Martín Gutiérrez et al., 2020)).

This also includes the range of research from recent years through social media data and streaming services like Spotify and Deezer. In the past decade, researchers have slowly begun using social information as a variable that predicts popularity in songs that has revealed how the networks and preferences of the group impact the fame of each song ((Abel et al., 2010); (Bischoff et al., 2009a); (Koenigstein et al., 2009); (Yu et al., 2019)). The influencing elements of social factors include shared listening experiences, online user interactions, and playlist curation, which contribute to the emergent success of a track as seen through various empirical studies ((Interiano et al., 2018); (Ren et al., 2016); (Silva and Moro, 2019)).

Despite the substantial growth experienced in the past decades, Hit Song Science remains a relatively new field of research within the broad bracket of Music Information Retrieval. In the last decade, the number of publications on HSS has gradually increased, reflecting the growing recognition of its significance and complexity (see Figure 2). A multidisciplinary debate that was there a decade ago now brings together scholars and practitioners from different backgrounds, presenting new visions and optimizing prediction methods. This still-growing body of research can then be understood as sharing methodological characteristics that generalize into a coherent workflow for predicting hit songs, opening the doors to further study of multimodal approaches.

2.2 Features

For both high-level and low-level audio features, researchers have utilized platforms like Spotify's API and tools such as **librosa** and **Essentia** to extract these features. Although most high-level features, such as danceability, energy, and valence, are available on platforms like Spotify ((Al-Beitawi et al., 2020)), for deeper low-level analysis, open-source libraries like **librosa** are often used. With **librosa**, features such as MFCCs, ZCR, Spectral Centroid, and tempo can be extracted from audio signals ((Araujo et al., 2020)). The **Essentia** library provides more than 40 basic spectral and temporal descriptors, enabling researchers to gather low-level information needed for rhythm, tonal, and spectral analysis.

Lyrics are also quite relevant during feature selection. Researchers have studied many techniques for quantifying and analyzing lyrics. (Dhanaraj and Logan, 2005) were among the first to apply semantic analysis techniques, including Probabilistic Latent Semantic Analysis (PLSA), to extract descriptive features based on lyrical content. Their method represented songs as vectors that expressed the likelihood of certain themes being present, thereby essentially measuring lyrical similarity between songs. This thematic approach had the possibility of taking lyrical features more seriously as a hit prediction factor.

The two concepts built on this were further presented by (Singhi and Brown, 2014) with a completely lyrics-based feature-based hit prediction model. They developed an absolute 24 rhyme as well as syllable-based features that make use of the software developed by Hirjee and Brown in 2010: the Rhyme Analyzer software, namely, features such as syllables/line, rhymes/line, and links/line. Furthermore, they utilized the CMU Pronunciation Dictionary that was developed by Elovitz et al. in 1976 to transcribe lyrics into phoneme sequences-a novel source of phonetic features-that capture stress patterns in the lyrics. Their results demonstrated the effects of complexity and structure of rhyme and meter in determining hit songs.

Study of lyrical features has expanded with new advances in more advanced machine learning and natural language processing techniques. Using Latent Dirichlet Allocation (LDA), (Ren et al., 2016) uncovered latent topics in song lyrics that were found to include recurring themes such as 'love' and 'life' in popular tracks. In the same direction, (Ren and Kaufman, 2017) applied LDA to extract semantic themes from a dataset comprising 4,410 tracks; they found that lyrical themes could highly aid in understanding the intent of the artist or the attraction of the song.

Analysis has also focused recently on the use of word usage, frequency, and stylistic features. For example, the study conducted by (Chiru and Popescu, 2017) used a bag-of-words method to extract words as well as their frequency, in which they emphasized that lyrical content has played a crucial role in identifying the possible success of a song. The kind of feature engineering, combined with other multimodal inputs, gives one a holistic perception about how well the song is liked.

It has added yet another dimension to the hit prediction. (Martín-Gutiérrez et al., 2020) and (Raza and Nanath, 2020) applied Natural Language Processing (NLP) to extract features of the sentence count, word length, and vocabulary richness. Of course, the overall sentiments of the lyrics ranged from negative to positive, that is what made the analysis so key. (Kamal et al., 2021) has demonstrated that the majority of popular songs are neutral or slightly positive and underlines how the lyrical tone influences the commercial appeal of a song.

Recent advances in transformer-based language models have further enhanced lyrical analysis capabilities for popularity prediction. (Prevedello et al., 2024) demonstrated the effectiveness of Large Language Model (LLM) embeddings for song success prediction, comparing sentence embeddings from pre-trained models with traditional stylometric features across different stages of a song's lifecycle. Their work showed that LLM-based lyrical embeddings provide complementary information to conventional features and significantly improve early-stage predictions when combined with audio and platform metadata, highlighting the potential of modern NLP architectures in capturing nuanced semantic content from song lyrics.

2.3 Learnt Representation

Recent work in machine learning has transitioned significantly from hand-crafted vectors to learned representations serving as feature vectors for different modalities of data, specifically audio and text. In contrast, domain experts have usually designed the algorithms over painful crafting of raw data for extracting specific features-like Mel-Frequency Cepstral Coefficients (MFCC) over audio or Bag-of-Words model over text-for capturing relevant information for tasks like speech recognition or sentiment analysis. But such hand-crafted features are often less generalizable and in many cases require rather large amounts of domain knowledge.

Deep learning, especially generative AI, has changed this approach by developing large models that learn complex representations and patterns directly from a large sum of data. This, in turn, directly helps for acquiring a vector representation of the different modalities of data and further for using the same in the downstream task. In the following sub-section we detail the work that has been done for audio data via neural codec models whereas transformers for text data.

2.3.1 Audio Learnt Representation: An Overivew

Recently, the field of audio coding has undergone a paradigm shift with the advent of neural audio codecs. These have been designed fresh for effective compression as well as reconstruction of audio signals for minimal data transmission latency without degrading the quality of audio: Traditional codecs were psychoacoustic models and essentially speech synthesis principles that relied heavily on human auditory perception-based signal reconstruction: (e.g., (Dietz et al., 2015)). But neural audio codecs go far beyond those approaches, using novel architectures that would allow higher compression rates and better quality.

The first model was probably that by (Zeghidour et al., 2022), creating the SoundStream. The standard neural codec architecture used includes an encoder, a quantizer, and a decoder. The encoder streams SEANets, and the quantizer uses Residual Vector Quantization (RVQ) for the parallel streams of tokens. Optimization of this model by a combination of reconstruction and adversarial losses is deployed in the SoundStream model, capable of compressing audio efficiently yet maintaining strong reconstruction quality. Building upon this, SoundStorm enhanced its hierarchical structures of tokens and designed non-autoregressive decoding to be more efficient and produce better quality audio outputs (Borsos et al., 2023). This was the first step towards using neural audio codecs not only for compression but also to create building blocks for audio language modeling tasks.

As the speed of neural audio codec advances permitted, models like Encodec (Défossez et al., 2022) could extend SoundStream capabilities through additional LSTM layers and a Transformer-based architecture. This allowed even more powerful sequence modeling of RVQ codes to enhance this coder's ability to capture complex audio patterns. Later adaptations introduced AudioDec (Wu et al., 2023), which combined group convolutions for

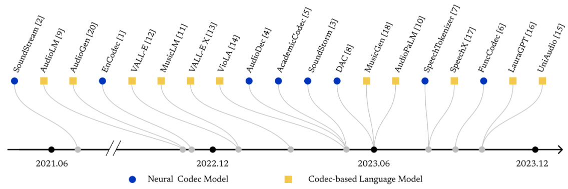


Figure 2: The Timeline of current neural codec models and codec-based language models. Source: Audio language modeling - overview

streaming with HiFi-GAN architectures to generate audio at 48 kHz. Again, emphasis on very high-quality audio generation was a direct consequence of this development, and interesting application domains apart from speech compression to music and general audio were opened.

An important achievement was the AcademiCodec model, which introduced group-residual vector quantization. The approach yielded improvements in reconstruction with multiple parallel RVQ groups at a minimum cost of the bit rate for lengthier speech tokens within speech language modeling. This innovation fills an essential gap between codec-based compression and efficient token representation toward large-scale tasks in language modeling.

Another significant development for the field has been the presentation of models specifically designed to tokenize and represent speech data for language models. For instance, SpeechTokenizer (Zhang et al., 2024) leverages a hierarchical architecture based on RVQ to separate semantic and acoustic tokens, thereby enabling the effective disentanglement of various aspects of speech. Techniques such as semantic token regularization with HuBERT tokens (Hsu et al., 2021), for example, are employed in SpeechTokenizer to emphasize the richer details captured within speech data, which enables more and possibly better language models.

Apart from these advancements in RVQ-based codecs, other approaches like the Descript-Audio-Codec (DAC) (Kumar et al., 2023) also looked into its viability. This is because DAC differs with the universal usage in a vast variety of sounds of audio types; it ranges from music to speech. It employs advanced techniques such as periodic activation functions, L2-normalized codes, and random quantizer dropout for reconstructive purposes which are highly accurate. It is one of the versatile and robust models that have acquired a very high value in the market as the leading neural codec for general usage of audio applications.

Last but not least, FunCodec (Du et al., 2023) proposed a new technique in the frequency domain, which was also designed to be friendly with fewer parameters but optimize the efficiency of the codec. In the frequency domain, Fun Codec has similar performance with a lower computational cost and is therefore useful for real-time applications. Semantic information is also integrated into the tokens of the codec, thus further improving the quality of speech, even at low bit rates. These high-performance neural audio codecs have revolutionized audio representation significantly. Not just compression models anymore, but really powerful tokenizers that take continuous audio and turn it into discrete codes; thus, new frontiers for audio language modeling are available. Neural audio codecs that not only preserve content but also paralinguistic nuances in some sense provide a really powerful foundation for building an audio language model that generalizes across kinds of audio types, namely speech, music, and general audio.

2.3.2 Text Learnt Representation: An Overview

(Patil et al., 2023) The science of song lyric meaning capturing has come a long way. It began from very simple statistical models to sophisticated techniques in understanding subtleties and emotion incorporated in words of a song. Until these models realized how words functioned in isolation-only rather than together, the subtler associations that make lyrics memorable and emotionally resonant, traditional methods, such as Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), were considered sufficient enough to perform even the most basic text analysis.

But it wasn't until word embeddings like Word2Vec and GloVe, which provided a more dimensional representation of words based on their relationships, that the game really changed. Innovation was such that a model could capture a lot of the semantic essence of words by mapping them into dense, multi-dimensional spaces. Representations like these go on to identify recurring themes in hit songs-things like love, heartbreak, or rebellion. These embeddings did capture general meaning but could not be sensitive and adjust to each unique lyrical context.

Enter BERT and GPT of context-sensitive embeddings, which seemed to completely change the game. These models brought a sense of awareness to the words, adapting their meanings based on surrounding context. This evolved really crucial to tackle the layers of complexity contained in song lyrics. For instance, the same word could convey highly contrasting meaning in the same ballad if it's a sorrowful one compared to an energetic pop anthem and to capture all that subtlety it is very critical that these words are interpreted in the broader context of the lyrics; BERT, with its bidirectional attention mechanism, can grasp every word in context to every other word in the lyrics, hence understanding the poetic rhythm, the feeling, and even some little twists that make a song relatable

It didn't end there. Contextual embeddings, such as ELMo, dove in even deeper. They added layers of interpretation. So not only does it read the line of a song just within its proximity but all over the lyrics's narrative arc. This way, models can grasp complex techniques of storytelling like metaphors, emotional transitions, and even juxtapositions of conflicting feelings that many artists embed in their work

These advances in text representation have been game-changers in predicting a hit song. Sentiment analysis with deep contextual models allows for a much more nuanced understanding of a song's emotional trajectory. For instance, one study utilizing BERT-based models presented significant improvements toward being able to identify potential hits by analyzing shifts in tone and sentiment over the course of the lyrics. Capturing these kinds of details empowers models not only to know what a song is talking about but also to understand how it makes one feel, an essential driver in determining whether a track becomes a chart-topper.

2.4 Musical Structure Segmentation in MIR

Musical structure (or form) refers to dividing a song into contiguous, labeled segments (e.g. intro, verse, chorus, bridge, etc.) that listeners perceive as meaningful parts (Serrà et al., 2020). In MIR, Music Structure Analysis (MSA) is formally defined as finding these non-overlapping segments and their types from audio (Serrà et al., 2020). For example, Fig. 3 illustrates an annotated song: the spectrogram (top) is partitioned into segments A, B, C, etc. (bottom), representing repeated verses or choruses as perceived by the annotator (Serrà et al., 2020). This segmentation mirrors Western music theory (e.g. binary, ternary forms or verse–chorus structure) (Serrà et al., 2020). Importantly, music structure can be hierarchical (e.g. motifs or subsections within a chorus) (Serrà et al., 2020), but most MIR methods focus on the flat (single-level) segmentation problem (Levy and Sandler, 2008). Researchers note that while exact boundaries are subjective, humans largely agree on major section boundaries (Serrà et al., 2020; Levy and Sandler, 2008).

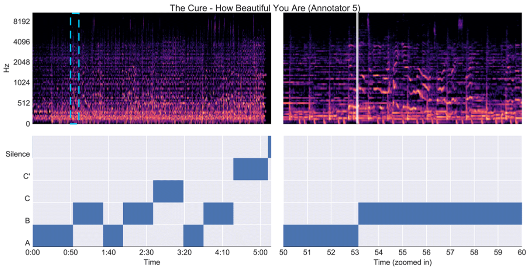


Figure 3: Example annotated segmentation (bottom) of a song with four sections A, B, C, C'. The top shows the song's spectrogram. Such manual annotations define musical structure (verse/chorus labels. Source Audio-Based Music Structure Analysis

2.4.1 Traditional Audio-Based Segmentation Methods

Traditional approaches rely on audio signal features and structural principles (homogeneity, novelty, repetition, regularity) to detect section boundaries. A common pipeline is to extract musical features (e.g. timbral or spectral features like MFCCs, mel-spectrogram, constant-Q/chromagram) and compute a self-similarity matrix (SSM) of these features over time. The SSM is a time–time map where bright off-diagonal blocks indicate repeated or similar sections (Foote, 2000). For example, Fig. 4 shows an SSM (left) with repeated diagonal blocks (musical repetitions) and a checkerboard novelty kernel used along the main diagonal to produce a novelty curve (right) (Foote, 2000). Peaks in this novelty curve often align with actual segment boundaries (Foote, 2000). This checkerboard method (Foote, 2000) remains a classic: sliding a positive–negative kernel along the diagonal of the SSM highlights sudden changes in features, whose prominent peaks are taken as boundaries (Foote, 2000).

Traditional signal features serve different roles. Timbral/spectral features (e.g. MFCCs or log-mel spectrograms) capture changes in instrumentation or texture, while harmonic features (chromagram or constant-Q) capture chord/tonal patterns. Recent studies show that compact spectral representations (CQT, mel-spectrogram) often outperform older

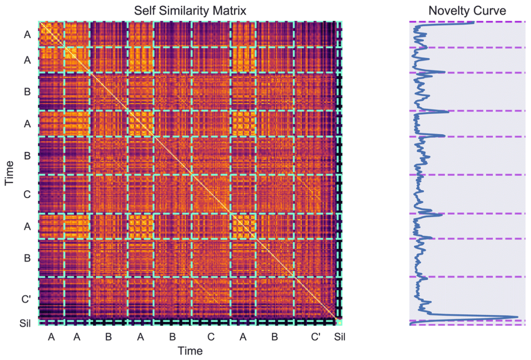


Figure 4: Self similarity matrix (left) and its associated novelty curve (right) of track 10 from SALAMI. Brighter colors in the SSM indicate a greater degree of similarity. Dashed lines mark segment boundaries. Source Audio-Based Music Structure Analysis

MFCC/chroma in structure tasks (Levy and Sandler, 2008; Grill and Schlüter, 2015). Many systems also use beat-synchronous analysis: aligning features to detected beats yields cleaner diagonals in the SSM, improving repeat detection (Serrà et al., 2020).

In addition to novelty detection, repetition-based methods explicitly find recurring segments. For example, lag matrices (Goto, 2003) compute similarity of each frame to its past frames, making horizontal/vertical lines in the matrix that indicate repetition (Goto, 2003). Serrà et al. (2014) introduced structural features by rotating lag matrices, combining homogeneity and repetition cues to form an enhanced novelty function (Serrà et al., 2014). Other methods build on these features with clustering or probabilistic models. Levy and Sandler (2008) used constrained clustering (HMM) on beat-synchronous features to jointly segment and label (chorus/verse) (Levy and Sandler, 2008). Paulus and Klapuri (2009) proposed a probabilistic fitness measure over possible segmentations (Paulus and Klapuri, 2009). Dynamic programming or HMMs have also been used to impose temporal consistency in the segmentation (Paulus and Klapuri, 2009).

- **Novelty/Homogeneity:** Convolve a checkerboard kernel on the SSM diagonal; pick peaks in the resulting novelty curve (Foote, 2000). (Foote, 2000) (Foote, 2000).

- **Repetition:** Identify diagonal paths (repeated blocks) via lag matrices or graph-walking (Goto, 2003). (Goto, 2003; Serrà et al., 2014) (Serrà et al., 2014).

- **Feature Types:** Use timbral/spectral (MFCC, mel-spectrogram) and harmonic (chromagram, CQT) features. Beat-synchronized features often enhance repeated structures (Serrà et al., 2020).

- **Clustering/Probabilistic Models:** Constrained clustering or HMM methods segment frames into states/segments to optimize structure criteria (Levy and Sandler, 2008).

- **Deep Learning:** Recent CNN-based models learn internal features from spectrograms. Grill & Schlüter (2015) and Ullrich et al. (2014) trained CNNs to detect novelty, achieving state-of-art boundary detection (Grill and Schlüter, 2015; Ullrich et al., 2014). McCallum (2019) showed unsupervised CNN features + checkerboard kernel also work well (McCallum, 2019).

Each approach has trade-offs. Homogeneity/novelty tends to find boundaries where the audio changes suddenly (e.g. verse→chorus), while repetition methods exploit returning patterns (refrain or cyclic motifs). Regularity constraints (typical segment lengths) are sometimes used to filter false positives (Paulus and Klapuri, 2009). Despite these methods, evaluation is hard: algorithms rarely achieve perfect match to human annotations due to subjectivity and varying definitions of "section" (Serrà et al., 2020; Cheng et al., 2009).

Figure [49] below illustrates a novelty-based approach: the SSM (left) of a track shows bright diagonals where sections repeat, and the novelty curve (right) has peaks aligning with the annotated boundaries (dashed lines) (Foote, 2000). Peaks in the novelty curve often correspond to transitions between sections (e.g. verse to chorus).

2.4.2 Lyrics and Musical Structure

Song lyrics carry their own structure which often aligns with musical sections. Lyrics are typically grouped into repeated lines or verses (e.g. chorus lyrics repeat) and unique stanzas (verses/bridge). These patterns mirror the song's form: choruses usually use the same lyrics each time, verses use different lyrics, etc. Indeed, "lyrics contain repeated patterns that are correlated with the repetitions found in the music they accompany" (Fell et al., 2022). Researchers exploit this: repeated lyric lines strongly indicate chorus segments and help segment text into verses/choruses (Fell et al., 2018, 2022).

Several approaches analyze lyrics text to infer structure:

- Textual segmentation by repetition: Falling back on dynamic programming, Fell et al. (2018, COLING) trained CNN models on raw lyrics to detect section boundaries, using features that capture word/phrase repetitions (Fell et al., 2018). Their model learns that identical or similar lines mark boundaries (e.g. the start of a chorus) without audio (Fell et al., 2018).
- Longest common subsequence (LCS): Cheng et al. (2009) segment lyrics by computing LCS similarity between lyric paragraphs. The most repeated paragraph (highest LCS with others) is labeled as the chorus, the unique short paragraphs as verse or bridge (Cheng et al., 2009). This leverages the idea that repeated lyric blocks → chorus (Cheng et al., 2009).
- Neural sequence models: Watanabe & Goto (2023) treat chorus detection as sequence labeling in lyrics text. They automatically align lyrics with audio-detected choruses to train a neural model that detects repeating lyric phrases. Their results show that learned "phrase repetition" patterns can identify chorus sections across languages (Watanabe and Goto, 2023). In short, repeated lyric motifs are strong signals of musical sections.

These lyric-only methods assume clean lyric transcripts and moderate knowledge of language. They demonstrate that lyrics alone can predict structure with some accuracy (Fell et al., 2022). For example, Fell et al. (2021) report 67% F-score in purely text-based segmentation (improving a previous 59%) by focusing on repetitive lyric patterns (Fell et al., 2022). Similarly, Watanabe & Goto generated chorus labels via audio and then learned to detect them in lyrics, finding repeating lines to be language-independent cues (Fell et al., 2022).

Beyond segmentation, lyrics inform alignment of text to sections. Since each lyrical paragraph often corresponds to a musical section (Watanabe and Goto, 2023), one can map labeled lyric segments to audio segments. Cheng et al. (2009) used this mapping: they first segmented audio by clustering, then analyzed lyrics to label each segment (intro, verse, chorus, etc.) based on lyric repetition and position (Cheng et al., 2009). In their framework, lyrical analysis made segment labeling (chorus/verse) straightforward, since lyrics of the same type tend to repeat (Cheng et al., 2009).

In summary, lyrics and music structure are tightly linked. Repeated lyrical motifs typically coincide with repeated musical sections (e.g. choruses). Algorithms exploit this by segmenting lyrics (via repetition or learning) and aligning text segments to audio. These cues supplement audio-only cues, especially for labeling sections semantically (Cheng et al., 2009).

2.4.3 Recent Multimodal Approaches

Recent research has begun combining audio and lyrics for more robust structure segmentation. These multimodal approaches leverage the complementary information: audio captures the acoustic changes, lyrics capture semantic repetitions.

- **Joint Audio+Lyrics Segmentation:** Cheng et al. (2009) pioneered this idea. They proposed a multimodal segmentation system where audio frames are clustered (with constraints) and lyrics are analyzed in parallel (Cheng et al., 2009). Their insights include: (1) lyrics make it easier to infer the number of segments and avoid under/over-segmentation; (2) segments of the same type have similar lyrics despite audio variations, aiding segment matching; (3) lyric content provides high-level cues for labeling (intro, chorus, etc.) (Cheng et al., 2009). In practice, they first segmented audio (imposing smoothness constraints), then used LCS on lyric paragraphs to label segments (identifying repeated paragraphs as chorus, etc.) (Paulus and Klapuri, 2009; Cheng et al., 2009). The result was more accurate boundary detection and semantic labeling than audio-only methods.
- **Bimodal Neural Models:** More recently, Fell et al. (2021) introduced a bimodal CNN that processes lyric text and synchronized audio jointly to segment lyrics (Fell et al., 2022). With a dataset of 4.8k songs with time-aligned lyrics, they showed that adding audio features to a text-based model boosts segmentation F-score from 67% to 75% (Fell et al., 2022). In other words, text and audio capture complementary structure cues. This is one of the first large-scale demonstrations that fusing lyric and audio modalities improves section segmentation performance.

- **Data Annotation and Learning:** Another hybrid approach is using one modality to annotate the other. Watanabe & Goto (2023) automatically detected choruses in audio (using an existing audio-based method) and then transferred these labels to lyric lines, creating a training set (Watanabe and Goto, 2023). They then trained a lyric-only model for chorus detection, effectively using audio to bootstrap text segmentation. This illustrates a broader trend: large audio–lyrics datasets (e.g. DALI (Meseguer-Brocal et al., 2018)) now enable training of cross-modal models.

In the MIR community, there is growing interest in such multimodal structure analysis. The "Everything Corpus" discussion notes that long-standing tasks like audio segmentation have begun to incorporate lyrics (Schubert et al., 2023). New datasets (e.g. DALI, Lyrics aligned to audio) and tasks (lyrics-to-audio alignment in MIREX) fuel this trend. Ultimately, hybrid models can yield more accurate and semantically meaningful segmentations than audio or lyrics alone.

Automated structure segmentation in MIR builds on both classic signal processing methods and newer deep learning. Audio-only methods use self-similarity/novelty and repetition detection on timbral/harmonic features (Foote, 2000; Goto, 2003). Lyrics themselves form a parallel structure: repeated lyric lines often mark chorus boundaries (Fell et al., 2018; Cheng et al., 2009). Recent work shows that combining lyrics and audio yields better segmentation: lyrics inform the semantic labeling and global layout of segments, while audio confirms boundaries in the waveform (Cheng et al., 2009; Fell et al., 2022). Altogether, the literature shows a move toward multimodal structural analysis, leveraging both modalities to detect and label song sections more reliably.

2.5 Overview and Data Requirements

Prediction systems that aim to forecast whether a track will break through the musical "attention economy" must be built on data that describe a song from three inter-locking viewpoints: its acoustic fingerprint, the story told by its lyrics, and the pattern of audience engagement that unfolds once the track is released. Research in "hit-song science" consistently shows that omitting any one of these modalities caps model accuracy, whereas approaches that fuse all three continue to yield state-of-the-art results. The remainder of this section details, in turn, the feature spaces that have proved most informative for each modality and the ways in which popularity itself is operationalised.

2.5.1 Audio-level features

Early studies relied on low-level descriptors such as Mel-frequency cepstral coefficients (MFCCs), spectral centroid, bandwidth, zero-crossing rate, and chroma vectors to capture timbre, brightness, and pitch content . Although inexpensive to compute, these frame-wise statistics ignore long-range form and dynamics. Researchers therefore introduced mid- and high-level attributes—tempo, key, mode, loudness, danceability, energy, valence, speechiness, liveness—first via the Echo Nest and now via the Spotify Web API. These features fold low-level curves into song-level embeddings that correlate well with listener perception and have become standard inputs in recent popularity models and competitions. Some projects go further, extracting rich temporal signatures with tools such as librosa or Essentia, yielding beat-synchronous MFCC summaries, harmonic change detection, octave-based spectral contrast, and Tonnetz trajectories that let a network "hear" modulation. More recent papers tokenise the full waveform with neural audio codecs (e.g., EnCodec) or WavTokenizer to preserve second-by-second evolution—an approach we extend in Stage 2 of this thesis.

2.5.2 Lyrics-level features

Lyrics furnish a complementary channel unavailable in the raw audio. Early papers modelled lyric text as bag-of-words or topics learned via probabilistic latent semantic analysis and LDA, demonstrating that semantic distance and theme recurrence correlate with chart outcome. Fine-grained rhyme and metre statistics, harvested with tools such as Rhyme Analyzer, subsequently revealed that intricate rhyme schemes and higher syllable density are over-represented among hits. Recent large-scale analyses strengthen the case for linguistic detail: across 48 000 pop, rock, rap, R&B and country tracks, descriptors such as the number of unique rhyme words, repeated-line ratio, structural ratios of chorus-to-verse, lexical diversity (e.g., Dugast's U, MTLD) and readability scores emerge as top predictors, with genre-specific patterns in the direction and strength of each effect . These findings motivate the structure-aware lyric embeddings introduced later.

2.5.3 Social Metadata features

Social signals translate listeners' collective behaviour into numbers that a model can digest. The literature converges on three complementary vantage-points. Top-Charts perspective metrics are taken directly from published rankings such as Billboard Hot 100, the UK Official Singles Chart, Spotify Top 200 or Gaon. Common variables include the song's debut rank, weekly position trajectory, time-to-peak, peak position, and total weeks on chart. These statistics encode both the height a track reaches and the shape of its ascent and decline, making them valuable for regression tasks that attempt to learn full popularity curves as well as for classification settings that mark any chart entry as a "hit" . Beyond simple ranks, rank-derived measures such as rank score (maxrank currentrank + 1) and life-cycle descriptors (slope of rise/fall, area under the rank curve) give finer temporal resolution .

The Economy perspective captures how much money or attention a track converts into sales. Classical variables are Nielsen SoundScan weekly units, Billboard album-equivalent units, and International Federation of the Phonographic Industry (IFPI) shipment reports. Online retail adds further resolution through Amazon Best-Seller Rank, which updates hourly and has proved a reliable proxy for real sales volume . These measures are particularly useful when the modelling goal is revenue forecasting or cross-platform valuation rather than short-term virality.

The Engagement perspective has expanded fastest in the streaming era and now dominates machine-learning studies. It mines high-velocity metrics such as Spotify play-count, playlist additions, follower growth, Shazam tags, TikTok usage counts, YouTube views and likes, and Last.fm "scrobbles". Spotify's own 0-to-100 popularity score—a recency-weighted play-count—has become the de-facto continuous label in recent work because it is easily retrievable through the Web API and strongly correlates with external chart success. Regional services provide analogous signals: KKBOX play-counts dominate studies on Southeast-Asian markets, while Last.fm listener counts and YouTube views supply user-generated engagement traces that extend back more than a decade .

Researchers also engineer secondary social features that amalgamate raw counts into more predictive forms. Examples include hit score, the product of log(play-count) and log(unique users), which dampens superstar outliers; HITS graph centrality linking artists, tracks and tags; and early-adopter indicators such as the first 48-hour stream total, shown to be predictive of long-run success . Regardless of the metric, best practice is to store each observation with a timestamp, geographic scope and data-source tag so that models can respect temporal causality and diagnose regional bias.

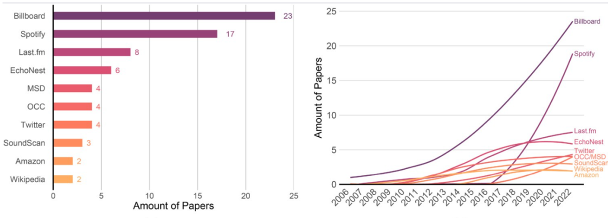
2.5.4 Popularity Score Modeling

Because "popularity" is multi-faceted, scholars operationalise it along two main axes. Continuous indices treat success as a real-valued quantity. Spotify's popularity score is the most widely used: it blends lifetime plays with a recency decay so that yesterday's viral spike has more weight than streams accrued years ago. Other continuous labels include raw or log-scaled stream totals, Last.fm listener counts, Amazon Best-Seller Rank inversions, and Echo Nest's historical hotttnesss score, which merges web mentions, blog reviews and play-counts. These labels support regression and time-series models that can predict trajectories or forecast future demand.

Discrete schemes convert popularity into classes. The simplest rule flags any appearance on a recognised chart—Billboard Hot 100, Official UK Top 40, Spotify Viral 50—as "hit" and treats non-charted tracks as "non-hits". More nuanced thresholds slice charts into strata (e.g., Top 10, 11–100, 101–200) or use the sample median of a chosen metric to split high- and low-popularity songs . When negatives are underspecified, researchers sample non-hits by matching artist or release year, thereby controlling for fan-base size and seasonality bias.

A third family of labels captures trajectory characteristics rather than absolute magnitude. Variables such as weeks on chart, time-to-peak, sustainability index (area under the rank curve) and chart lifespan allow models to differentiate flash-in-the-pan virality from slow-burn success . These descriptors align naturally with survival analysis and epidemic-style diffusion models that treat attention as a contagious process.

Whichever formulation is chosen, three safeguards are mandatory. First, the popularity window must open after the last date used to compute input features to prevent information leakage. Second, the label's time granularity (daily, weekly, cumulative) must match that of the features; misalignment can spuriously inflate performance estimates. Third, whenever data permit, both global and regional scores should be stored, because a track can be a blockbuster in one territory and invisible in another—an effect that otherwise embeds hidden bias into the model. Adhering to these principles yields labels that are both statistically sound and industry-relevant, providing a trustworthy target for the multimodal architectures developed in subsequent chapters.



(a)

(b)

Figure 5: Data sources for Hit Song Science: (a) Top 10 data sources ranked according to number of papers (b) Usage trend of top sources. Source: Hit Song Science

Of its many charts, the Billboard Hot 100 represents the most commonly used measure of a song's popularity. It releases a weekly rank of the 100 most streamed songs from all music styles in the United States, based on parameters such as radio airplay, sales, and streaming figures ((Askin and Mauskapf, 2017); (Bischoff et al., 2009b); (Kim and Oh, 2021)). The Year-End Hot 100 also summarizes these weekly rankings, and most researchers use this integrated data for general understanding ((Singhi and Brown, 2014)). However, some researchers focus on individual Billboard charts like the Top Jazz Chart or Rock Songs Chart, which allows for a finer level of observation within particular genres ((Chon et al., 2006b); (Lee and Lee, 2015)). As the level of global connectivity continues to grow, research has expanded beyond U.S. rankings. The Official Charts Company (OCC) represents British rankings at its core ((Herremans et al., 2014)). Besides this, other regional markets from which data has been gathered include France, Belgium, and Germany ((Buda and Jarynowski, 2015); (Herremans and Bergmans, 2020)). In addition to this, growing music markets in Asia, such as South Korea ((SHIN and PARK, 2018)) and Indonesia ((Febirautami et al., 2018)), have also been considered. Some authors have used alternative metrics of popularity, such as YouTube views and interactions ((Chiru and Popescu, 2017)), as well as Amazon sales reports and Nielsen SoundScan ((Dewan and Ramaprasad, 2014)).

Acoustic Features The acoustic aspects need to be taken into account in the analysis of song composition. Initially, authors relied on in-house databases to analyze the acoustic features of songs, as seen in (Dhanaraj and Logan, 2005). Over time, with the rapid advancements in Music Information Retrieval (MIR), more sophisticated data sources like the *EchoNest API* emerged, comprising over a trillion data points for more than 34 million songs ((Herremans et al., 2014)). This API has been widely utilized in research to extract features such as tempo, time signature, loudness, and song duration ((Askin and Mauskapf, 2017); Fan & Casey, 2013). Since 2014, when it acquired EchoNest, Spotify's *Developer API* has become the primary data source for new research focusing on acoustic features ((Kamal et al., 2021); (Raza and Nanath, 2020)). Other databases like *The Million Song Dataset* (MSD) ((Zangerle et al., 2019c)) and *AcousticBrainz* [2] ((Votter et al., 2021)) have provided additional information on different audio features.

Lyric-Based Data Song lyrics provide a unique dimension to music success, offering insights into rhyme schemes and textual sentiment. Although there is no consensus on a single reliable source, multiple studies have utilized online platforms like *Astraweb Lyric Search*, *MetroLyrics*, *MusicSongLyrics*, and *LyricsMania* ((Chiru and Popescu, 2017); (Ren et al., 2016)). Recently, due to the existence of a dedicated API, the platform *Genius* has gained popularity among researchers, as it simplifies data collection without the need for web crawlers.

Social Behavior Data The development of social networks has significantly influenced how people share their opinions, impacting the music world. As such, user behavior has become a critical element in analyzing successful artistic products. Platforms like *Last.fm* [3] have been extensively used to gather listener-based data, including user preferences and interactions ((Ren et al., 2016); (Votter et al., 2021)). Blogging platforms like *Spinn3r* [4]have also been employed to gather vast amounts of blog posts related to music, reflecting listeners' sentiments about songs and artists ((Abel et al.,

[2] AcousticBrainz: http://acousticbrainz.org

[3] Last.fm API: http://www.last.fm/api

[4] Spinn3r: http://docs.spinn3r.com

2010)).

In recent studies, social networks like Twitter, Instagram, and Facebook have been explored to examine user behavior concerning new releases ((Araujo et al., 2020); (Cosimato et al., 2019a)). The integration of social media data provides researchers with deeper insights into listener engagement and conversations surrounding music.

In summary: HSS models typically pool data from multiple sources to generate comprehensive predictions about musical success. As illustrated in Figure 4(a), popular chart rankings such as *Billboard*[5] and *Official Charts Company* [6] dominate the landscape, followed by more data-intensive song features involving acoustic properties, lyrics, and social interactions. However, the use of specific charts introduces regional biases, as each market recognizes different artists and genres. As the music industry becomes increasingly globalized, researchers are employing a broader range of sources to capture diverse market dynamics.

2.6 Literature Survey of Open-Source Music Datasets

The public corpus landscape has expanded from modest audio-only collections to vast, richly annotated datasets that encode many facets of musical engagement, and this evolution has underpinned each fresh wave of popularity-prediction research. The Million Song Dataset (MSD) released by Bertin-Mahieux et al. (Bertin-Mahieux et al., 2011) marked the first large-scale, legally shareable repository, offering one million 30-second excerpts with detailed Echo Nest timbre and rhythm descriptors but no native success label, thereby forcing later studies to graft in external chart or stream data. Companion logs such as the Taste Profile Subset compiled by the same authors added forty-eight million user–song play-count triplets (Bertin-Mahieux and Ellis, 2011), while Ocelma (Ocelma, 2010) furnished the Last.fm 360K collection with 17.6 million user–artist counts and basic demographics; together, these interaction matrices provided the first open proxies for popularity through aggregate listening statistics. In the same period, the Yahoo! Music Ratings corpus released for the KDD-Cup 2011 (Dror et al., 2011) supplied 300 million explicit 1–5 star ratings for 600 k items, enabling studies that framed success as a crowd-sourced quality score rather than as chart position.

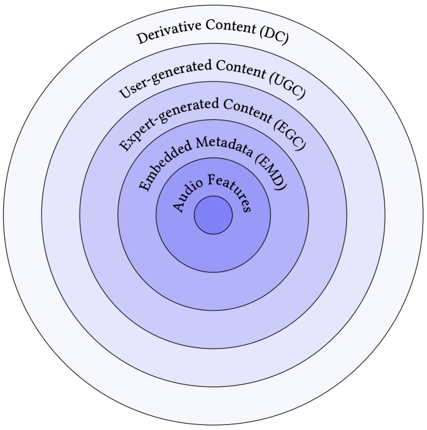


Figure 6: Onion model of music features. Source: Music4allOnion

Building on these foundational datasets, Figure 5 illustrates how source usage has diversified and accelerated in recent years. In panel (a), traditional chart data—most prominently Billboard—remains the largest single repository for success labels, while streaming metrics such as Spotify have quickly become nearly as prevalent in the HSS literature. Panel (b) shows Spotify's adoption surging after 2018, outpacing more gradual rises for Last.fm and EchoNest, and underscoring the field's shift toward real-time listener feedback alongside established chart metrics.

The next generation of corpora hard-wired chart information directly into the dataset. Oliveira et al. (Oliveira and Pacheco, 2021) curated MUHSIC, a complete facsimile of every Billboard Hot 100 entry since the 1940s, exposing weekly rank time-series that let models predict trajectory shape as well as peak. Silva et al. (Silva et al., 2019) built MusicOSet by matching 20 405 U.S. tracks to Billboard ranks, Genius lyrics and Essentia acoustic features, and published both a continuous year-end score and a binary hit/non-hit flag—an influential template for multimodal benchmarking. Parallel work at Innsbruck produced the Hit-Song Prediction suites: HSP-S and HSP-L (Vötter et al., 2022) comb the AcousticBrainz repository to pair 7 736 and 73 482 tracks, respectively, with Billboard peak rank plus Last.fm listener counts, while also releasing mel-spectrograms so that end-to-end deep architectures can be trained without external feature extraction.

Streaming platforms then shifted attention from static chart peaks to continuous engagement metrics. Santana et al. (Santana et al., 2020) intro-

[5] Billboard: http://www.billboard.com/charts

[6] Official Charts: http://www.officialcharts.com/charts

duced Music4All, a 109 269-track corpus that embeds Spotify's 0–100 popularity index alongside audio clips, lyrics and Last.fm tags; its successor Music4All-Onion (Moscati et al., 2022) retains the same track universe but appends 252 984 396 listening events and organises all modalities into concentric "layers," as shown in Fig 6 encouraging systematic ablation. A similar philosophy guides SpotGenTrackPopularity (Martín-Gutiérrez et al., 2019), which aggregates Spotify acoustic features and Genius lyrics for an unspecified but comparably large set of tracks, again adopting the Spotify index as a regression target. On the consumption side, LFM-1b (Schedl, 2016) and LFM-2b (Schedl et al., 2022) expose more than two billion Last.fm "scrobbles," preserving timestamps and user profiles so that researchers can derive early-momentum or long-tail endurance indicators. Twitter-sourced corpora such as #nowplaying (Zangerle et al., 2014) and #nowplaying-RS (Zangerle et al., 2019a) collect up to fifty million listening declarations linked to MusicBrainz IDs, extending implicit-feedback coverage to social-media contexts.

Engagement data at still greater scale arrived with the commercial release of Yambda-5B (Ploshkin et al., 2025), which logs 4.79 billion user–track events and explicit like/dislike interactions for 9.39 million tracks in the Yandex-Music catalogue; although no single composite popularity score is provided, the raw counts empower sequence models that learn success signals directly from interaction streams. The playlist-centric Spotify Million Playlist Dataset (Chen et al., 2018) offers a complementary view by treating track frequency across one million user-generated playlists as an implicit salience cue, a feature found to correlate with both short-term viral spikes and long-term consumption.

Across these corpora, three modality trends emerge. First, low-level spectral features have been steadily supplemented—and often superseded—by high-level perceptual attributes supplied via the Spotify API or Essentia, allowing researchers to bypass costly signal processing. Second, lyric availability has grown from sparse manual transcriptions to near-comprehensive crawls, enabling sentiment, structural and complexity analyses at scale. Third, popularity labels have diversified: binary chart appearance, ordinal rank strata, explicit star ratings, raw play counts, recency-weighted Spotify indices and even unaggregated user-event sequences all coexist, each favouring a different class of predictive model. The chronological arc from MSD's audio-only snapshot through chart-aligned benchmarks to streaming-native, multimodal giants thus traces the field's methodological maturation and provides today's researchers with a rich menu of openly licensed options for benchmarking music-popularity prediction.

2.7 Summary and Conclusion

We conducted a detailed review of the Music Information Retrieval (MIR) literature, with a particular focus on the subfield of Hit Song Science (HSS). HSS lies squarely at the intersection of algorithmic predictability and the science of human preference in music. Scientists have approached from both prediction and science sides aiming to unlock the secrets of musical appeal and what makes a song a hit. This dual focus on teasing apart the intangible qualities of music and quantifying them into workable data forms the backbone of HSS. The field has, in a way, evolved to be what it is today-a multidisciplinary pursuit, taking music theory and sociology off the shelf and adding machine learning and data science to predict the most effective chart-toppers. It reveals a growth in complexity and ambition in forecasting musical success from the early genre categorization models to the more global frameworks of HSS.

The change in feature representation is also quite dramatic within the literature. While we once had handcrafted traditional features, we now see massive shifts with regard to learned representations through deep learning. The neural audio codecs and the rise of transformer-based architecture, including BERT, are redefining how audio and text are encoded and understood. Audio codecs like SoundStream, Encodec, and other neural codecs provide highly aggressive audio compression while still maintaining all the essential parts of music and paralinguistic information. Contextual embeddings like BERT, GPT, and ELMo gave models an understanding of deeper levels of meaning and an emotional trajectory of lyrics where traditional textual representations like TF-IDF and Word2Vec became inadequate. The transition into learned representations improved the feature generalization and enhanced the expressive and the narrative dimensions of audio and text for MIR tasks. Multimodal data integration, such as audio, lyrics, and social features, has, therefore opened doors for more comprehensive and accurate predictions in HSS.

The thorough literature review reveals a number of significant gaps in the state of hit song science today that restrict the precision and interpretability of systems for predicting music popularity.

The way lyrics are structurally understood and processed is still fundamentally lacking, despite the fact that previous research has investigated lyrical content using sentiment analysis and thematic modeling. The majority of current methods treat lyrics as unstructured text, using simple embedding techniques or bag-of-words, ignoring the songs' natural structural organization into verses, choruses, bridges, and other semantic segments. Important details about how musical narratives develop and how recurring elements like choruses enhance memorability and listener engagement are lost as a result of this structural blindness. Systematic methods for automatically recognizing and utilizing these structural components in popularity prediction models are lacking in the literature.

Previous research has mostly ignored the dynamic elements of an artist's career trajectory in favor of static artist characteristics like follower counts, historical chart performance, or demographic data. Research on how an artist's momentum, growth patterns, listener engagement quality, and career consistency over time affect how new releases are received is conspicuously lacking. There is a substantial knowledge gap regarding how artistic careers develop and affect the performance of individual tracks since the literature does not adequately depict the temporal evolution of artist popularity and its predictive value for future song success.

Further the Hit Song Science has been sluggish to embrace these state-of-the-art methods for music analysis, despite developments in neural audio codecs and learned representations in other fields. Instead of using neural audio codecs, which can extract richer semantic and acoustic information from entire audio tracks, the majority of current work still relies on conventional handcrafted features like MFCCs and Spotify's high-level descriptors. This gap is an uncharted area where the expressiveness and granularity of audio features used in popularity prediction could be greatly improved by contemporary representation learning.

To combine audio, lyrical, and social features, current multimodal approaches in HSS usually use basic ensemble techniques or simple concatenation. Sophisticated fusion architectures that can dynamically weight the significance of various modalities according to their dependability and pertinence for individual songs are lacking in the literature. Additionally, model interpretability is not given enough consideration; knowledge of the characteristics that influence predictions and the ways in which various modalities influence success forecasts is still largely unexplored, which restricts the useful information that these systems can offer to artists and business professionals.

These gaps serve as the basis for the contributions made in this thesis, which investigates neural audio codec representations, introduces career trajectory dynamics as temporal features, addresses structural lyric understanding through large language model-based annotation, and creates interpretable modality-aware fusion networks for more precise and explicable music popularity prediction.

References

Fabian Abel, Ernesto Diaz-Aviles, Nicola Henze, Daniel Krause, and Patrick Siehndel. 2010. Analyzing the blogosphere for predicting the success of music and movie products. pages 276–280.

Zayd Al-Beitawi, Mohammad Salehan, and Sonya Zhang. 2020. Cluster analysis of musical attributes for top trending songs.

Carlos Araujo, Marco Cristo, and Rafael Giusti. 2020. A model for predicting music popularity on streaming platforms. *Revista de Informática Teórica e Aplicada*, 27:108–117.

Noah Askin and Michael Mauskapf. 2017. What makes popular culture popular? product features and optimal differentiation in music. *American Sociological Review*, 82(5):910–944.

Thierry Bertin-Mahieux and Daniel P. W. Ellis. 2011. Taste profile subset of the million song dataset. Echo Nest / LabROSA. 48M user–song play-counts; 1M users.

Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. 2011. The Million Song Dataset. *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 591–596. 1,000,000 songs with audio analysis features.

Kerstin Bischoff, Claudiu S. Firan, Mihai Georgescu, Wolfgang Nejdl, and Raluca Paiu. 2009a. Social knowledge-driven music hit prediction. In *Advanced Data Mining and Applications*, pages 43–54, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kerstin Bischoff, Claudiu S. Firan, Mihai Georgescu, Wolfgang Nejdl, and Raluca Paiu. 2009b. Social knowledge-driven music hit prediction. In *Advanced*

Data Mining and Applications, pages 43–54, Berlin, Heidelberg. Springer Berlin Heidelberg.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*.

Andrzej Buda and Andrzej Jarynowski. 2015. Exploring patterns in european singles charts. *Preprint*, arXiv:1503.07301.

Tiffany Chen, Aaron Moore, Michael Volkovs, Thierry Bertin-Mahieux, and Bryan Pardo. 2018. The spotify million playlist dataset challenge. AICrowd. 1,000,000 playlists, 2M tracks; playlist recommendation.

Heng-Tze Cheng, Yi-Hsuan Yang, Yu-Chung Lin, and Homer H. Chen. 2009. Multimodal structure segmentation and analysis of music using audio and textual information. In *2009 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1677–1680. IEEE.

Costin Chiru and Oana-Georgiana Popescu. 2017. Automatically determining the popularity of a song. pages 392–406.

Song Hui Chon, Malcolm Slaney, and Jonathan Berger. 2006a. Predicting success from music sales data: a statistical and adaptive approach. In *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, AMCMM '06, page 83–88, New York, NY, USA. Association for Computing Machinery.

Song Hui Chon, Malcolm Slaney, and Jonathan Berger. 2006b. Predicting success from music sales data: a statistical and adaptive approach. In *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, AMCMM '06, page 83–88, New York, NY, USA. Association for Computing Machinery.

Débora C. Corrêa and Francisco Ap. Rodrigues. 2016. A survey on symbolic data-based music genre classification. *Expert Syst. Appl.*, 60(C):190–210.

Alberto Cosimato, Roberto Prisco, Alfonso Guarino, Delfina Malandrino, Nicola Lettieri, Giuseppe Sorrentino, and Rocco Zaccagnino. 2019a. The conundrum of success in music: Playing it or talking about it? *IEEE Access*, 7:1–1.

Alberto Cosimato, Roberto De Prisco, Alfonso Guarino, Delfina Malandrino, Nicola Lettieri, Giuseppe Sorrentino, and Rocco Zaccagnino. 2019b. The conundrum of success in music: Playing it or talking about it? *IEEE Access*, 7:123289–123298.

Sanjeev Dewan and Jui Ramaprasad. 2014. Social media, traditional media, and music sales. *MIS Q.*, 38(1):101–122.

Ruth Dhanaraj and Beth Logan. 2005. Automatic prediction of hit songs. In *International Society for Music Information Retrieval Conference*.

Martin Dietz, Markus Multrus, Vaclav Eksler, Vladimir Malenovsky, Erik Norvell, Harald Pobloth, Lei Miao, Zhe Wang, Lasse Laaksonen, Adriana Vasilache, Yutaka Kamamoto, Kei Kikuiri, Stephane Ragot, Julien Faure, Hiroyuki Ehara, Vivek Rajendran, Venkatraman Atti, Hosang Sung, Eunmi Oh, and Changbao Zhu. 2015. Overview of the evs codec architecture.

Shuangrui Ding, Zihan Liu, Xiaoyi Dong, Pan Zhang, Rui Qian, Conghui He, Dahua Lin, and Jiaqi Wang. 2024. Songcomposer: A large language model for lyric and melody composition in song generation. *arXiv preprint arXiv:2402.17645*.

Gideon Dror, Niv Koenigstein, and Yehuda Koren. 2011. Yahoo! music user ratings of songs, albums and artists dataset. Yahoo! Webscope R2. 300M explicit ratings (1–5 stars) on 600K items.

Zhihao Du, Shiliang Zhang, Kai Hu, and Siqi Zheng. 2023. Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec. *Preprint*, arXiv:2309.07405.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *Preprint*, arXiv:2210.13438.

Limisgy Febirautami, Isti Surjandari, and Enrico Laoh. 2018. Determining characteristics of popular local songs in indonesia's music market. pages 197–201.

Michael Fell, Yaroslav Nechaev, Elena Cabrio, and Fabien Gandon. 2018. Lyrics segmentation: Textual macrostructure detection using convolutions. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 2044–2054.

Michael Fell, Yaroslav Nechaev, Gabriel Meseguer-Brocal, Elena Cabrio, Fabien Gandon, and Geoffroy Peeters. 2022. Lyrics segmentation via bimodal text–audio representation. *Natural Language Engineering*, 28(3):317–336.

Jonathan Foote. 2000. Automatic audio segmentation using a measure of audio novelty. In *2000 IEEE International Conference on Multimedia and Expo (ICME)*, volume 1, pages 452–455. IEEE.

Masataka Goto. 2003. A chorus-section detection method for musical audio signals and its application to a music database system. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages V–673. IEEE.

Thomas Grill and Jan Schlüter. 2015. Music boundary detection using neural networks on self-similarity matrices. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pages 731–737.

Dorien Herremans and Tom Bergmans. 2020. Hit song prediction based on early adopter data and audio features. *Preprint*, arXiv:2010.09489.

Dorien Herremans, David Martens, and Kenneth Sörensen. 2014. Dance hit song prediction. *Journal of New Music Research*, 43.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

Myra Interiano, Kamyar Kazemi, Lijia Wang, Jienian Yang, Zhaoxia Yu, and Natalia Komarova. 2018. Musical trends and predictability of success in contemporary songs in and out of the top charts. *Royal Society Open Science*, 5:171274.

Jigisha Kamal, Pankhuri Priya, M Anala, and G Smitha. 2021. A classification based approach to the prediction of song popularity. pages 1–5.

Seon Tae Kim and Joo Hee Oh. 2021. Music intelligence: Granular data and prediction of top ten hit songs. *Decision Support Systems*, 145:113535.

Noam Koenigstein, Yuval Shavitt, and Noa Zilberman. 2009. Predicting billboard success using data-mining in p2p networks. pages 465–470.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved rvqgan. *Preprint*, arXiv:2306.06546.

Jongpil Lee, Nicholas J. Bryan, Justin Salamon, Zeyu Jin, and Juhan Nam. 2020. Disentangled multidimensional metric learning for music similarity. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6–10.

Junghyuk Lee and Jong-Seok Lee. 2015. Predicting music popularity patterns based on musical complexity and early stage popularity. In *Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia*, SLAM '15, page 3–6, New York, NY, USA. Association for Computing Machinery.

Mark Levy and Mark Sandler. 2008. Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):318–326.

Yinghao Ma, Anders Øland, Anton Ragni, Bleiz MacSen Del Sette, Charalampos Saitis, Chris Donahue, Chenghua Lin, Christos Plachouras, Emmanouil Benetos, Elona Shatri, et al. 2024. Foundation models for music: A survey. *arXiv preprint arXiv:2408.14340*.

Daniel Martín-Gutiérrez, Iván Meza-Ruiz, and Javier Ponce-de León. 2019. SpotGenTrackPopularity dataset. Mendeley Data. Audio features + Genius lyrics; Spotify popularity score.

David Martín-Gutiérrez, Gustavo Hernández Peñaloza, Alberto Belmonte-Hernández, and Federico Álvarez García. 2020. A multimodal end-to-end deep learning architecture for music popularity prediction. *IEEE Access*, 8:39361–39374.

David Martín Gutiérrez, Gustavo Hernández-Peñaloza, Alberto Belmonte Hernández, and Federico Alvarez. 2020. A multimodal end-to-end deep learning architecture for music popularity prediction. *IEEE Access*, PP:1–1.

David Martín-Gutiérrez, Gustavo Hernández Peñaloza, Alberto Belmonte-Hernández, and Federico Álvarez García. 2020. A multimodal end-to-end deep learning architecture for music popularity prediction. *IEEE Access*, 8:39361–39374.

Matthew C. McCallum. 2019. Unsupervised learning of musical structure using deep convolutional networks. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pages 476–483.

Gabriel Meseguer-Brocal, Alice Cohen-Hadria, and Geoffroy Peeters. 2018. Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, pages 431–437.

Matteo Moscati, Daniel Schnitzer, and Markus Schedl. 2022. Music4All-Onion: Extended listening events dataset. Zenodo. 252M listening events; temporal social metadata; multilingual.

Martin Ocelma. 2010. Last.fm 360k dataset. Public Release. 359K users, 17.6M user–artist play-count triples.

Mitsunori Ogihara and Youngmoo Kim. 2011. Mood and emotional classification. *Music data mining*, page 135.

Diego F. M. Oliveira and Paulo S. Pacheco. 2021. MUHSIC: A billboard chart-based musical success dataset. Zenodo. Weekly Billboard Hot 100 entries (1940s–2020s).

François Pachet and CSL Sony. 2012. Hit song science. *Music data mining*, pages 305–326.

François Pachet and Pierre Roy. 2008. Hit song science is not yet a science. In *International Society for Music Information Retrieval Conference*.

Rajvardhan Patil, Sorio Boit, Venkat Gudivada, and Jagadeesh Nandigam. 2023. A survey of text representation and embedding techniques in nlp. *IEEE Access*, PP:1–1.

Jouni Paulus and Anssi Klapuri. 2009. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 369–374.

Alexandr Ploshkin, Ivan Zholus, and Grigory Podgurski. 2025. Yambda-5B: A large-scale yandex music listening log. Hugging Face. 4.79B events, 1M users, 9.39M tracks; user-item listens.

Giulio Prevedello, Ines Blin, Bernardo Monechi, and Enrico Ubaldi. 2024. Lyrics for success: Embedding features for song popularity prediction. In *Proceedings of the 3rd Workshop on NLP for Music and Audio (NLP4MusA)*, pages 75–80, Oakland, USA. Association for Computational Lingustics.

Agha Raza and Krishnadas Nanath. 2020. Predicting a hit song with machine learning: Is there an apriori secret formula?

Jing Ren and Robert Kaufman. 2017. Understanding music track popularity in a social network.

Jing Ren, Jialie Shen, and Robert Kauffman. 2016. What makes a music track popular in online social networks? pages 95–96.

Federico Rossetto, Jeffrey Dalton, and Roderick Murray-Smith. 2023. Generating multimodal augmentations with llms from song metadata for music information retrieval. In *Proceedings of the 1st Workshop on Large Generative Models Meet Multimodal Applications*, LGM3A '23, page 51–59, New York, NY, USA. Association for Computing Machinery.

Prof. R.Y. Sable, Aqsa Sayyed, Baliraje Kalyane, Kosheen Sadhu, and Prathamesh Ghatole. 2024. Enhancing music mood recognition with llms and audio signal processing: A multimodal approach. *International Journal for Research in Applied Science and Engineering Technology*.

Ian A. P. Santana, César Boscariol, and Guilherme Fardilha. 2020. Music4All: Multi-modal music popularity dataset. University Repository. 109,269 songs; audio, lyrics, metadata; Spotify popularity; request access.

Markus Schedl. 2016. LFM-1b: Last.fm 1 billion listening events dataset. Public Release. 1.07B listening events; 120K users; user demographics.

Markus Schedl, Dorien Herremans, and Juan P. Monteiro. 2022. LFM-2b: Last.fm 2 billion listening events dataset. Public Release. 2.01B events; 120K users; 50.8M tracks; lyrics embeddings.

Emery Schubert et al. 2023. Towards an 'everything corpus': A framework and guidelines for the curation of more comprehensive multimodal music data. *Transactions of the International Society for Music Information Retrieval*, 6(1).

Joan Serrà, Meinard Müller, Peter Grosche, and Josep Lluís Arcos. 2014. Unsupervised music structure annotation by time series structure features and segment similarity. *IEEE Transactions on Audio, Speech, and Language Processing*, 22(7):1802–1813.

Joan Serrà, Oriol Nieto, Gautham J. Mysore, Cheng-i Wang, Jordan B. L. Smith, Jan Schlüter, Thomas Grill, and Brian McFee. 2020. Audio-based music structure analysis: Current trends, open challenges, and applications. *Transactions of the International Society for Music Information Retrieval*, 3(1):246–263.

Danilo B. Seufitelli, Gabriel P. Oliveira, Mariana O. Silva, Clarisse Scofield, and Mirella M. Moro. 2023. Hit song science: a comprehensive survey and research directions. *Journal of New Music Research*, 52:41 – 72.

SEUNGKYU SHIN and JUYONG PARK. 2018. On-chart success dynamics of popular songs. *Advances in Complex Systems*, 21(03n04):1850008.

Andre Silva, Tiago Gonçalves, and Helder Luz. 2019. MusicOSet: A large dataset for music popularity analysis. Zenodo. 20,405 songs, acoustic features, metadata, lyrics.

Mariana Silva and Mirella Moro. 2019. Causality analysis between collaboration profiles and musical success. pages 369–376.

Abhishek Singhi and Daniel G Brown. 2014. Hit song detection using lyric features alone. *Proceedings of International Society for Music Information Retrieval*, 30.

Bob L. Sturm. 2014. A survey of evaluation in music genre recognition. In *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation*, pages 29–66, Cham. Springer International Publishing.

Karen Ullrich, Thomas Gautier, and Alexander Lerch. 2014. Boundary detection in music signals. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, pages 605–610.

Manuel Vötter, Peter Knees, and Thomas Lidy. 2022. HSP-S and HSP-L: Hit song prediction datasets. Zenodo. Includes HSP-S (7,736 songs) and HSP-L (73,482 songs).

Michael Votter, Maximilian Mayerl, Gunther Specht, and Eva Zangerle. 2021. Novel datasets for evaluating song popularity prediction tasks. pages 166–173.

Kento Watanabe and Masataka Goto. 2023. A method to detect chorus sections in lyrics text. *IEICE Transactions on Information and Systems*, E106.D(9):1600–1609.

Yi-Chiao Wu, Israel D. Gebru, Dejan Marković, and Alexander Richard. 2023. Audiodec: An open-source streaming high-fidelity neural audio codec. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Li-Chia Yang, Szu-Yu Chou, Jen-Yu Liu, Yi-Hsuan Yang, and Yi-An Chen. 2017. Revisiting the problem of audio-based hit song prediction using convolutional neural networks. *Preprint*, arXiv:1704.01280.

Haiqing Yu, Yanling Li, Shujun Zhang, and Chunyan Liang. 2019. Popularity prediction for artists based on user songs dataset. In *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence*, ICCAI '19, page 17–24, New York, NY, USA. Association for Computing Machinery.

Eva Zangerle, Manuel Hofer, and Georg Specht. 2019a. #nowplaying-rs: Enriched music listening events. Zenodo. 11.6M events; includes user context (hashtags, location).

Eva Zangerle, Georg Specht, and Patrick Exner. 2014. #nowplaying dataset: Music listening events from twitter. Zenodo. 49.9M listening tweets; user, track, artist, timestamp.

Eva Zangerle, Michael Vötter, Ramona Huber, and Yi-Hsuan Yang. 2019b. Hit song prediction: Leveraging low- and high-level audio features. In *International Society for Music Information Retrieval Conference*.

Eva Zangerle, Michael Vötter, Ramona Huber, and Yi-Hsuan Yang. 2019c. Hit song prediction: Leveraging low- and high-level audio features. In *International Society for Music Information Retrieval Conference*.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2022. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024. Speechtokenizer: Unified speech tokenizer for speech large language models. *Preprint*, arXiv:2308.16692.