Lyrics Matter: Exploiting the Power of Learnt Representations for Music Popularity Prediction

Anonymous ACL submission

Abstract

Accurately predicting music popularity is a critical challenge in the music industry, offering benefits to artists, producers, and streaming platforms. Prior research has largely focused on audio features, social metadata, or model architectures. This work addresses the under-explored role of lyrics in predicting popularity. We present an automated pipeline that uses LLMs to extract mathematical representations from lyrics, capturing semantic, syntactic, and sequential information. These features are integrated into HitMusicLyricNet, a multimodal architecture that combines audio, lyrics, and social metadata for popularity score prediction in range 0-100. Our method outperforms existing baselines on the SpotGenTrack dataset which contains over 100,000 tracks, achieving 9% and 20% improvements in MAE and MSE, respectively. Ablation confirms that gains arise from our LLM-driven lyrics feature pipeline (LyricsAENet), underscoring the value of dense lyric representations.

1 Introduction

In 2023, the global recorded music market generated \$28.6 billion[1] in revenue. 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 progressed alongside advances in machine learning, beginning with classical approaches using acoustic features, and later incorporating social signals that reflect evolving listener preferences (Seufitelli et al., 2023). With the advent of deep learning, models became better at capturing complex patterns, prompting the integration of multiple modalities—audio, lyrics, and social metadata—for improved prediction (Zangerle et al., 2019; Martín-Gutiérrez et al., 2020). Popularity is typically measured by a song's duration on charts such as Billboard, or via streaming platform metrics—most notably the Spotify popularity score, which has been widely adopted in recent studies post-2020 (Seufitelli et al., 2023). Evaluation is conducted using regression metrics (MAE, MSE, R^2) or classification metrics (accuracy, precision, recall, F1). More recently, large language models (LLMs) have spurred new work in music recommendation, emotion analysis, and lyric generation by modeling lyrical text as a rich source of semantic content (Rossetto et al., 2023; Sable et al., 2024; Ma et al., 2024; Ding et al., 2024). However, music popularity prediction has yet to fully exploit the potential of learned lyric representations, despite recent findings showing their strong influence on popularity (Yu et al., 2023).

Through our work, we address the gap in the existing literature with the following main contributions:

1. An automated lyric feature extraction pipeline that uses LLMs to encode music lyrics into rich, learned representations. Details discussed in 4.1.2

2. An end to end multimodal deep learning architecture which predicts the popularity score in range (1,100) and outperforms current baseline by 9% and 20% in MAE and MSE metrics respectively. Details discussed in 4.1

The next section reviews related work. This is followed by a discussion of our methods, the dataset and our experiments.

2 Related Work

Traditional research in music popularity prediction has primarily focused on using machine learning techniques such as Logistic Regression, Decision Trees, Support Vector Machines (SVM), Bayesian Networks, Random Forest Ensembles, XGBoost,

[1] IFPI Report '23

and K-Nearest Neighbors (KNN). Subsequently, these evolved into neural networks and deep learning methods, resulting in more robust predictive models. Numerous studies (Bischoff et al., 2009; Dorien Herremans and Sörensen, 2014; Zangerle et al., 2019; Silva et al., 2022) have used acoustic characteristics of songs alongside metadata encompassing social influences. Concurrently, other works (Dhanaraj and Logan, 2005; Singhi and Brown, 2015b; Martín-Gutiérrez et al., 2020) have highlighted the significance of lyrics, employing handcrafted statistical text features capturing sentiment and syntactic structures. However, these studies were limited in capturing deep lyrical semantics and structural dependencies.

The availability of large datasets has further propelled research in this area. Prominent datasets include Million Song Dataset[2], SpotGenTrack[3], and AcousticBrainz[4], sourced from platforms like Spotify, Billboard, Genius[5], and YouTube. These datasets incorporate a broad spectrum of features, from low-level Mel-Frequency Cepstral Coefficients (MFCCs) and temporal features to high-level attributes such as danceability and loudness. They also contain metadata on artists, albums, genres, and demographics. Despite the emotional depth and listener impact carried by lyrics—often surpassing acoustic features alone (Singhi and Brown, 2015a)—lyrics have historically received less attention compared to acoustic and social attributes (Seufitelli et al., 2023). Early methodologies like Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) sought to capture lyrical semantics, enhancing the understanding of lyrical impact on song popularity (Dhanaraj and Logan, 2005). Subsequent research expanded beyond basic semantics. For example, (Hirjee and Brown, 2010; Singhi and Brown, 2014) employed rhyme and syllable characteristics for popularity prediction solely based on lyrics, while others used Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to uncover thematic lyric topics (Ren et al., 2016).

Advancements in deep learning have facilitated multimodal approaches combining lyrics, audio, and metadata, often employing stylometric analysis for text feature extraction (Martín-Gutiérrez et al., 2020). Sentiment analysis further emerged as a means to derive emotional insights from lyrics for popularity prediction (Raza and Nanath, 2020). More recent approaches utilize learned lyric representations, such as embeddings (Kamal et al., 2021; McVicar et al., 2022), providing richer semantic understanding. (Barman et al., 2019) demonstrated the effectiveness of distributed representations in predicting both genre and popularity, eliminating the reliance on handcrafted features. The Music4All-Onion dataset (Moscati et al., 2022) specifically offers lyric embeddings, enabling deeper analysis of lyrical influence on song success. Additionally, recent research identified lyrical uniqueness as significantly influencing song popularity, using TF–IDF vectors (Yu et al., 2023). However, such approaches inherently lack the ability to model deeper sequential and contextual nuances, underscoring the necessity for richer, context-aware lyric representations to fully understand the factors that resonate with audiences.

Existing literature thus highlights a critical limitation: the lack of efficient, automated extraction methods for expressive lyrical features that encapsulate inherent complexities of song lyrics and its semantics. To address this gap, we propose LyricsAENet pipeline leveraging Large Language Models, which offers rich, semantically and syntactically coherent lyric representations while maintaining their sequential structure.

3 Dataset

We use the SpotGenTrack Popularity Dataset (SPD), originally introduced by Martín-Gutiérrez et al. (2020), which contains 101,939 tracks from 56,129 artists and 75,511 albums. Tracks are sourced from Spotify and Genius APIs, covering the top 50 playlists across 26 countries. Spotify provides track-level popularity scores ranging from 1 to 100. These scores follow a Gaussian distribution with $\mu = 40.02$ and $\sigma = 16.79$. The dataset includes low-level audio features extracted from raw waveforms, high-level audio descriptors, stylometric text features derived from lyrics, and metadata such as artist popularity and market reach. To ensure data quality, we applied filtering steps to remove noisy lyric entries. Specifically, we excluded tracks with lyrics shorter than 100 or longer than 7,000 characters, which often contained placeholders or irrelevant content. Additionally, we restricted the dataset to five major languages: English, Spanish, Portuguese, French, and German—discarding other languages that constituted less than 1% of

[2] Million Song Dataset
[3] SpotGenTrack
[4] AcousticBrainz
[5] Genius.com

the data. This resulted in a cleaned corpus of 74,206 tracks, comprising 51,319 in English and 22,887 in the remaining languages which we name as SPD_cleaned. The cleaned popularity distribution maintained the original characteristics, with $\mu = 41.11$ and $\sigma = 17.51$, ensuring that no sampling bias was introduced.

We also evaluated other publicly available datasets for potential use but found them lacking in multimodal completeness. The TPD dataset (Karydis et al., 2016) omits lyrical and social metadata; the MSD dataset (Bertin-Mahieux et al., 2011) contains only bag-of-words lyrics; HSP-S and HSP-L (Vötter et al., 2021) exclude full lyrical text; MUSICOSET (Silva et al., 2019) lacks detailed audio features; and the LFM-2B dataset (Schedl et al., 2022) has unresolved copyright restrictions.

4 Methodology

4.1 HitMusicLyricNet

This section introduces HitMusicLyricNet, our proposed end-to-end multimodal deep learning architecture for music popularity prediction, built upon the foundation of HitMusicNet. The architecture comprises three key components: AudioAENet, LyricsAENet, and MusicFuseNet. AudioAENet compresses low-level audio features; LyricsAENet encodes high-dimensional lyric embeddings into compact representations using an autoencoder preserving semantic structure. MusicFuseNet integrates these compressed representations with high-level audio features and metadata, as summarized in Table 2. Unlike HitMusicNet—which compresses all modality features jointly using a single autoencoder—HitMusicLyricNet employs separate encoders to mitigate information loss, particularly for underrepresented modalities. Additionally, lyrics embeddings exhibit directional and bipolar properties, motivating the need for a distinct compression technique (Bałazy et al., 2021). Implementation details of the baseline architecture are provided in Appendix A.

4.1.1 AudioAENet

AudioAENet compresses low-level audio features (e.g., MFCCs, spectral contrast) as outlined in Table 2. Given input dimension $d = 209$, the encoder reduces dimensionality through layers of size $d/2$, $d/3$, and $d/5$. Hidden layers use ReLU activation; the decoder uses sigmoid activation. The model is optimized using Adam with MSE loss, converging to a reconstruction loss of $\sim 10^{-5}$.

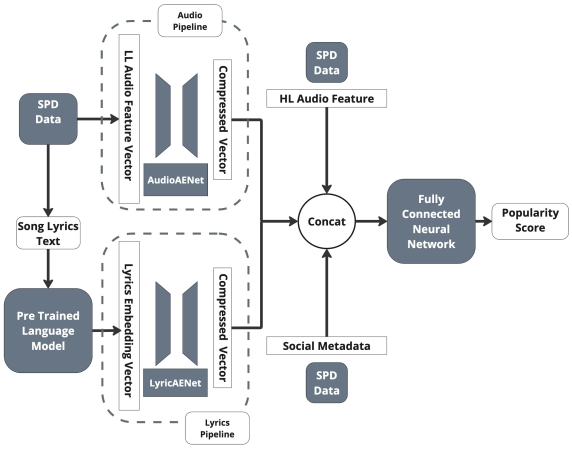

Figure 1: Block schematic of the *HitMusicLyricNet* architecture comprising of two Autoencoders and a Fully Connected NN predicting popularity score. 'HL' stands for high-level and 'LL' stands for low-level.

4.1.2 LyricsAENet

LyricsAENet implements a tied-weights autoencoder (Li and Nguyen, 2019) to compress lyrics embeddings from LLMs such as BERT (Devlin et al., 2019), LLaMA-3 (Grattafiori et al., 2024), and OpenAI's embedding models[6]. The encoder compresses through layers $(d/2, d/4, d/8)$ followed by a bottleneck of $d/12$ or $d/16$. The decoder mirrors this structure using transposed encoder weights.

We use the Scaled Exponential Linear Unit (SELU) activation (Klambauer et al., 2017) for its self-normalizing properties and suitability for bipolar embeddings. We also evaluate SiLU (Elfwing et al., 2018) and GELU (Hendrycks and Gimpel, 2016) in ablation studies. Training uses Adam and MSE loss, with final reconstruction loss $\sim 10^{-5}$. To preserve directional properties of the embeddings, we apply an additional directional loss (Bałazy et al., 2021):

$$L(Y, \bar{Y}) = \alpha_1 \cdot \mathrm{MSE}(Y, \bar{Y}) + \alpha_2 \cdot \mathrm{CD}(Y, \bar{Y}), \quad (1)$$

where $\mathrm{CD}(Y, \bar{Y})$ denotes cosine distance and α_1, α_2 balance the reconstruction and directional components.

4.1.3 MusicFuseNet

MusicFuseNet combines the compressed audio and lyric embeddings with high-level audio features and metadata. The fused vector is passed through

[6] https://platform.openai.com/docs/guides/embeddings

a feedforward network with hidden layer widths scaled as $(1, 1/2, 1/3)$, ReLU activations, and a final sigmoid output. The network is trained using Adam and MSE loss with dropout regularization. The output is normalized to $[0, 1]$ and later rescaled to Spotify's $[1, 100]$ range during evaluation.

5 Experiments and Results

To validate our setup, we first implemented the HitMusicNet architecture using the publicly available **Code**[7] and the configuration described in Appendix A. The model was trained on the original SPD dataset using an 80–20 train–test split and 5-fold stratified cross-validation, with MAE and MSE as evaluation metrics. Our results closely matched those reported by Martín-Gutiérrez et al. (2020), confirming the correctness of our implementation. To establish a reliable baseline on our cleaned data, we then retrained HitMusicNet on the SPD_cleaned dataset. Stylometric lyric features used in the original work were found to have negligible impact and were removed from further experiments. A summary of test performance across all model variants is shown in Table 1, and optimal HitMusicLyricNet configuration was selected based on results in Appendix Table 4.

Model	Dataset	MSE (Test)	MAE (Test)
HitMusicNet	SPD_Cleaned	0.0119	0.0865
HitMusicNet w/o lyrics	SPD_Cleaned	0.0120	0.0867
HitMusicLyricNet w/o lyrics	SPD_Cleaned	0.0115	0.0854
HitMusicLyricNet	SPD_Cleaned	**0.0097**	**0.0770**

Table 1: Test set performance comparison with baseline (HitMusicNet) on SPD_Cleaned datasets. HitMusicLyricNet model configuration are as per the best test scores from Table 4.

For all subsequent evaluations, we used the cleaned version of SPD (denoted SPD_cleaned) described in Section 3. We trained HitMusicLyricNet using LLM-derived lyric embeddings. For open-source models (BERT, LLaMA), we used vanilla checkpoints from Hugging Face[8]; for OpenAI models, embeddings were obtained via API. Lyrics were tokenized, passed through the model, and pooled using max/mean to obtain fixed-size vectors. For BERT, both mean pooling and max+CLS concatenation were evaluated. Embeddings were then compressed using LyricsAENet, for which we compared activation functions (SELU, SiLU, GELU) and loss formulations. We incorporated directional loss as in Bałazy et al. (2021) with $\alpha_1 = 0.5$ and $\alpha_2 = 0.1$ to test its effect alongside MSE. Results are reported in Appendix Table 3. SELU with MSE yielded the lowest MAE and was selected for all further experiments. Directional loss produced comparable but non-superior performance. We also compared embeddings from BERT (small/large), LLaMA (3.1 8B, 3.2 1B, 3.2 3B), and OpenAI (small, large). Table 4 in Appendix B summarizes these results. OpenAI large embeddings outperformed all others. While differences in performance were minor (~2%), we attribute them to variations in pretraining corpora and architectural inductive biases.

Using OpenAI large embeddings and LyricsAENet with SELU+MSE, HitMusicLyricNet achieved a ~9% improvement in MAE and 20% in MSE over HitMusicNet. Ablation studies (Appendix B.1) confirm that gains stem from the inclusion of our LLM-based lyric representation pipeline. Detailed modality contributions, as well as interpretability and residual error analyses, are presented in Appendix Sections C.1–C.4.

6 Conclusion and Future Work

The work presented in this paper demonstrates the effectiveness of leveraging lyric representations generated by Large Language Models for music popularity prediction. By utilizing embeddings that capture deeper semantic nuances within song lyrics, our proposed HitMusicLyricNet architecture achieves a significant improvement of 9% over current state-of-the-art method. The conducted ablation study further underscores the effectiveness of lyric embeddings in enhancing predictive performance. Future advancements in music-aware language models hold promise for generating even more explainable and expressive lyric features by incorporating domain-specific knowledge. Advances in audio representation learning, particularly using neural audio codecs, may enable richer and more nuanced music representations. Furthermore, while current research aggregates song-level features, recent trends in virality driven by micro-content platforms such as Instagram and Snapchat highlight the need to explore localized features within distinct musical segments, suggesting a promising direction for future research.

[7] https://github.com/dmgutierrez/hitmusicnet
[8] https://huggingface.co/

7 Limitation

Our findings may be constrained by genre, demographic, and cultural variability not fully captured in the current experimental setup. While LLMs such as BERT and LLaMA-3 enable deeper semantic modeling of lyrics, their general-purpose training limits their ability to capture music-specific linguistic patterns. Despite careful regularization, the high dimensionality of lyric embeddings presents inherent risks of overfitting. Moreover, as these embeddings are evaluated solely through downstream task performance, their intrinsic quality in representing lyrical content remains underexplored. Finally, the opacity of these feature vectors limits interpretability, pointing to a need for more explainable models of lyric representation.

References

Klaudia Bałazy, Mohammadreza Banaei, Rémi Lebret, Jacek Tabor, and Karl Aberer. 2021. Direction is what you need: Improving word embedding compression in large language models. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 322–330, Online. Association for Computational Linguistics.

Manash Pratim Barman, Kavish Dahekar, Abhinav Anshuman, and Amit Awekar. 2019. It's only words and words are all i have. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II*, page 30–36, Berlin, Heidelberg. Springer-Verlag.

Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. In *International Society for Music Information Retrieval Conference*.

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

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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*.

David Martens Dorien Herremans and Kenneth Sörensen. 2014. Dance hit song prediction. *Journal of New Music Research*, 43(3):291–302.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick

Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Hussein Hirjee and Daniel G. Brown. 2010. Rhyme analyzer: An analysis tool for rap lyrics. In *International Society for Music Information Retrieval Conference (ISMIR)*. ISMIR. Late-Breaking Demo.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 50–57, New York, NY, USA. Association for Computing Machinery.

J. Kamal, P. Priya, M. R. Anala, and G. R. Smitha. 2021. A classification based approach to the prediction of song popularity. In *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pages 1–5. IEEE.

Ioannis Karydis, Aggelos Gkiokas, and Vassilis Katsouros. 2016. Musical track popularity mining dataset. In *Artificial Intelligence Applications and Innovations*.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. 2017. Self-normalizing neural networks. *Advances in neural information processing systems*, 30.

Ping Li and Phan-Minh Nguyen. 2019. On random deep weight-tied autoencoders: Exact asymptotic analysis, phase transitions, and implications to training. In *International Conference on Learning Representations*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

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*.

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.

Matt McVicar, Bruno Di Giorgi, Baris Dundar, and Matthias Mauch. 2022. Lyric document embeddings for music tagging.

Marta Moscati, Emilia Parada-Cabaleiro, Yashar Deldjoo, Eva Zangerle, and Markus Schedl. 2022. Music4all-onion – a large-scale multi-faceted content-centric music recommendation dataset. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 4339–4343, New York, NY, USA. Association for Computing Machinery.

Agha Haider Raza and Krishnadas Nanath. 2020. Predicting a hit song with machine learning: Is there an apriori secret formula? In *2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*, pages 111–116.

Jing Ren, Jialie Shen, and Robert J. Kauffman. 2016. What makes a music track popular in online social networks? In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 95–96, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *Preprint*, arXiv:1602.04938.

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*.

Markus Schedl, Stefan Brandl, Oleg Lesota, Emilia Parada-Cabaleiro, David Penz, and Navid Rekabsaz. 2022. Lfm-2b: A dataset of enriched music listening events for recommender systems research and fairness analysis. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, CHIIR '22, page 337–341, New York, NY, USA. Association for Computing Machinery.

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.

Mariana O. Silva, Laís Mota, and Mirella M. Moro. 2019. Musicoset: An enhanced open dataset for music data mining.

Mariana O. Silva, Gabriel P. Oliveira, Danilo B. Seufitelli, Anisio Lacerda, and Mirella M. Moro. 2022. Collaboration as a driving factor for hit song classification. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, WebMedia '22, page 66–74, New York, NY, USA. Association for Computing Machinery.

Abhishek Singhi and Daniel G. Brown. 2014. Hit song detection using lyric features alone. In *International Society for Music Information Retrieval Conference (ISMIR): Late-Breaking Demo*, Waterloo, Canada. University of Waterloo, Cheriton School of Computer Science, ISMIR. Late-Breaking Demo.

Abhishek Singhi and Daniel G. Brown. 2015a. Can song lyrics predict hits? In *Proceedings of the 11th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 457–471.

Anurag Singhi and David G. Brown. 2015b. Can song lyrics predict hits. In *International Symposium on Computer Music Multidisciplinary Research*, pages 457–471. The Laboratory of Mechanics and Acoustics.

Michael Vötter, Maximilian Mayerl, Günther Specht, and Eva Zangerle. 2021. Novel datasets for evaluating song popularity prediction tasks. *2021 IEEE International Symposium on Multimedia (ISM)*, pages 166–173.

Yulin Yu, Pui Yin Cheung, Yong-Yeol Ahn, and Paramveer S. Dhillon. 2023. Unique in what sense? heterogeneous relationships between multiple types of uniqueness and popularity in music. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):914–925.

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

A Baseline Methodology

A.1 Problem Formulation

Given a song S, its features are represented in a multi-dimensional space $X \in \mathbb{R}^d$, which comprises three key modalities: audio waveform $w \in \mathbb{R}^k$, lyrical text $l \in \mathbb{R}^m$, and metadata attributes $m \in \mathbb{R}^p$, where $d = k + m + p$ represents the total dimensionality of our feature space. Our primary objective is to extract meaningful features from the song lyrics to effectively encode each song into a unique vector representation. Next, the prediction task is formulated as learning a mapping function $f : X \to Y$, where we minimize the expected prediction error: $\mathbb{E}[(f(X) - Y)^2]$ across the training distribution. Here, $Y \in \mathbb{R}$ represents the continuous popularity score.

A.2 HitMusicNet

We trained *HitMusicNet*, a multimodal end-to-end Deep Learning architecture as proposed by (Martín-Gutiérrez et al., 2020) and validated the results using the SpotGenTrack Popularity Dataset (SPD). The model outputs a popularity score between 1 and 100, using audio features, text features, and metadata containing artist and demographic information as inputs. A complete description of the feature set used is provided in Table 2.

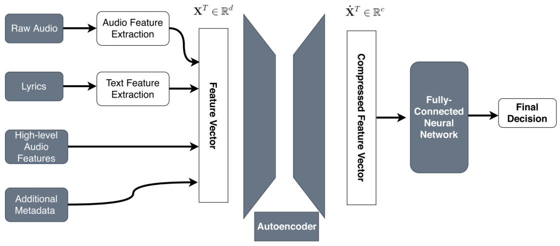

Figure 2: Diagram of the HitMusicNet pipeline outlining the principal functionalities and data components. Image src (Martín-Gutiérrez et al., 2020).

Feature Type	Features
Text Features	Sentence count, Avg words, Word count, Avg syllables/word, Sentence similarity, Vocabulary wealth
High-Level Audio	Danceability, Energy, Key, Loudness, Mode, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo, Duration, Time Signature
Low-Level Audio	Mel-spectrogram, MFCCs, Tonnetz, Chromagram, Spectral Contrast, Centroid, Bandwidth, Zero-Crossing Rate
Meta-Data Features	Artist followers, Artist popularity, Available markets

Table 2: Summary of features used in the HitMusicNet architecture (Martín-Gutiérrez et al., 2020).

HitMusicNet architecture as shown in Fig 2, employs an autoencoder for feature compression through two encoder layers with dimensions $d/2$ and $d/3$, followed by a bottleneck layer of $d/5$. Each layer uses ReLU activation, and the output layer employs a sigmoid activation for reconstruction. The autoencoder was trained using the Adam optimizer and an MSE loss function. The compressed features are then passed through a fully connected neural network with four layers, where the number of neurons in each layer is scaled by factors $\alpha = 1$, $\beta = 1/2$, and $\gamma = 1/4$. The model is trained using an 80%-20% train-test split with stratified cross-validation (SCV) using $k = 5$. These settings helped us in effectively replicating the baseline results on the SPD dataset.

LyricsAENet Config	MAE (Train)	MAE (Val)	MAE (Test)
SELU, MSE	**0.0769**	**0.0746**	**0.0775**
SiLU, MSE	0.0736	0.0731	0.0790
GELU, MSE	0.0740	0.0731	0.0792
SELU, Dir.	0.0741	0.0740	0.0799

Table 3: Results of training and testing HitMusicLyricNet on cleaned SPD data with various LyricAENet configurations (activation function, loss function), using BERT Large embeddings throughout. 'Dir' indicates directional loss 1.

Embeddings Model	MAE (Train)	MAE (Val)	MAE (Test)
BERT large	**0.0793**	**0.0784**	**0.0786**
Llama 3.1 8B	0.0774	0.0759	0.0795
Llama 3.2 1B	0.0775	0.0754	0.0800
Llama 3.2 3B	0.0781	0.0766	0.0798
OpenAI Small	0.0746	0.0738	0.0788
OpenAI Large	**0.0761**	**0.0743**	**0.0770**

Table 4: Results of training and testing HitMusicLyricNet on cleaned SPD data with different lyric embeddings sent to LyricAENet (Selu activation, MSE loss).

B Experiments and Results

B.1 Ablation Study

In this section, we study how different modalities contribute to our model's music popularity predictive strength. Table 5 shows model performance for each combination of our four feature types: high-level audio (HH), low-level audio (LL), lyrics embeddings (LR), and metadata (M).

The model works best when it uses all modalities, with a test MAE of 0.0772. If we exclude lyrics embeddings, the test MAE increases by 10.4% to 0.0852, highlighting the usefulness of our proposed lyrics feature pipeline. Notably, using only high-level features and metadata along with lyrics (HH, LR, M) gives comparable performance to using all the modalities features, indicating some redundancy in low-level audio features. The role of social context is apparent when we strip metadata by utilizing only audio and lyrics features (HH, LL, LR), which makes the test MAE rise by 40.2% to 0.1082. Performance suffers most significantly if we use only audio features (HH, LL) and obtain a test MAE of 0.1196.

Modality Config	MAE (Train)	MAE (Val)	MAE (Test)
HH, LL, LR, M	**0.0761**	**0.0743**	**0.0770**
HH, LL, M	0.0818	0.0841	0.0852
HH, LL, LR	0.1059	0.1037	0.1082
HH, LR, M	0.0767	0.0765	0.0795
HH, LL	0.1188	0.1175	0.1196
LR, M	0.0810	0.0811	0.0805

Table 5: Results of training and testing HitMusicLyricNet with different modality combinations. HH: High-level audio features, LL: Low-level audio features, LR: Lyrics embeddings features, M: Metadata features.

To further understand individual modality performance, we conducted isolated training experiments as shown in Table 6. Single-modality tests ascertain that metadata features (M) alone achieve the highest single-modality performance with a test MAE of 0.0968, verifying our initial observation about the importance of social context in music popularity prediction. Lyrics embeddings (LR) are similarly predictive to low-level audio features (LL), with test MAEs of 0.1193 and 0.1229, respectively. High-level audio features (HH) are slightly worse in isolation with a test MAE of 0.1272. These results show that while each modality contains valuable information, their combination creates synergistic effects that significantly improve prediction accuracy, as evidenced by the better performance of the full model in Table 5.

Modality Config	MAE (Train)	MAE (Val)	MAE (Test)
LL	0.1234	0.1218	0.1229
HH	0.1260	0.1266	0.1272
LR	0.1208	0.1189	0.1193
M	0.1026	0.0956	0.0968

Table 6: Performance comparison of individual modalities in predicting song popularity, showing the relative strength of each feature type in isolation.

C Error Analysis

While HitMusicLyricNet surpasses the state-of-the-art baseline, an in-depth error analysis is necessary for real-world applications and future enhancements. In this section, we examine global residual errors, assess feature interpretability and impact via SHAP and LIME, and analyze social metadata to uncover any systematic biases and error patterns. All analyses are performed using the test set.

C.1 Global Residual Error Analysis

Figure 3 compares the actual and predicted music popularity distributions. Although the means are nearly identical ($\mu_{\text{actual}} = 0.422$, $\mu_{\text{predicted}} = 0.428$), the predicted distribution's tails are compressed. The model predicts only 8.3% of songs with popularity below 0.2 (compared to 12.6% in

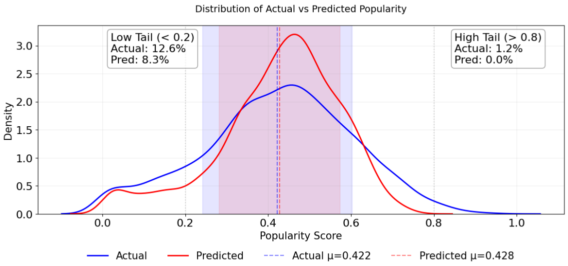

Figure 3: Actual (blue) vs. predicted (red) music popularity distributions on test set, showing prediction compression at both tails with aligned means ($\mu_{\text{actual}} = 0.422$, $\mu_{\text{predicted}} = 0.428$).

the actual data) and fails to predict any songs with popularity above 0.8 (versus 1.2% in the actual data). This regression towards the mean reflects both the limited representation of extreme popularity cases in SPD dataset and also the model's particular difficulty in capturing patterns of highly popular songs.

The calibration plot (Fig. 4) also indicates a strong alignment between predicted and actual music popularity within most bins, with the highest precision in the 0.4-0.6 range where data density peaks.

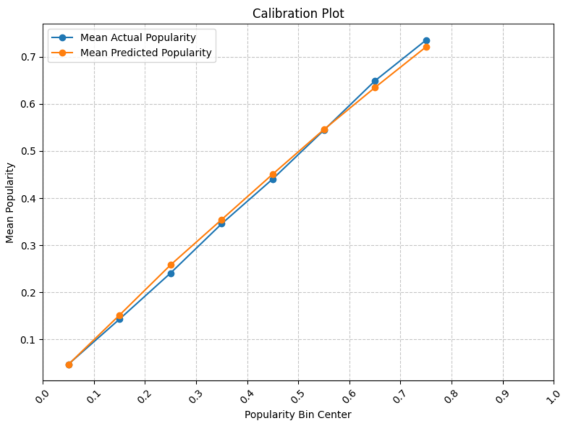

Figure 4: Model calibration plot showing alignment between mean predicted and actual popularity per bin.

Analysis of the residual distribution (Figure 5) shows a quasi-normal pattern centered at zero, with about 95% of forecasts falling within ±0.2 of actual values. The distribution shows minimal negative skewness, suggesting a small inclination toward underestimating in extreme conditions. With variance amplification in the mid-popularity range (0.3–0.6) and more limited errors at the extremes, the residual scatter plot against predicted popularity (Figure 6) shows heteroscedastic behavior.

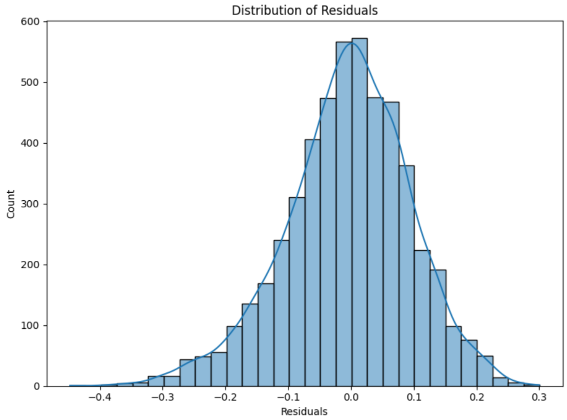

Figure 5: Distribution of prediction residuals centered at $\mu \approx 0.0$, showing approximately normal spread with slight negative skewness.

.

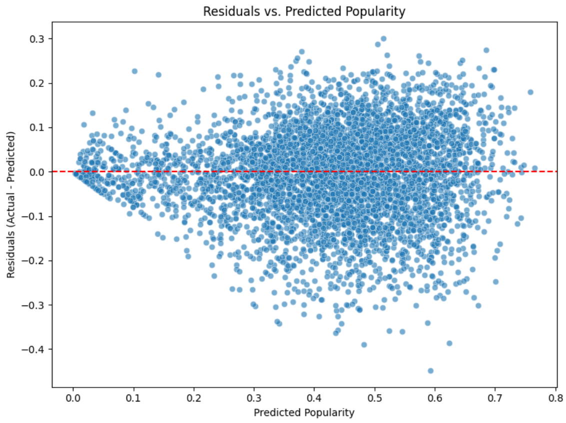

Figure 6: Scatter plot of residuals vs predicted popularity values showing error distribution across popularity ranges.

C.2 Interpretability Analysis

To understand the overall impact of non-interpretable latent representation of music audio and lyrics and the explicit metadata, we used SHAP (SHapley Additive exPlanations)(Lundberg and Lee, 2017), and LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016) techniques on a randomly sampled 10% of test data.

On analyzing the outcome of SHAP (Fig 7), artist popularity was the strongest predictor of music popularity with SHAP values ranging from -0.2 to $+0.2$. The compressed audio features showed a decreasing impact across sequential layers, indicating that earlier layers captured more predictive patterns. Lyric embeddings showed a moderate but consistent impact unless there is a significant deviation from the typical pattern. LIME analysis supported these findings and substantiated detailed insights on decision boundaries within feature values as presented in Appendix ??.

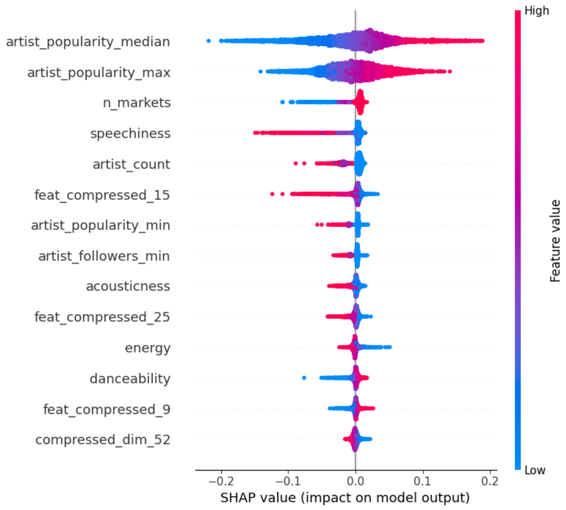

Figure 7: SHAP value distributions for top 15 features across all modalities, with artist-related features showing highest impact on model predictions.

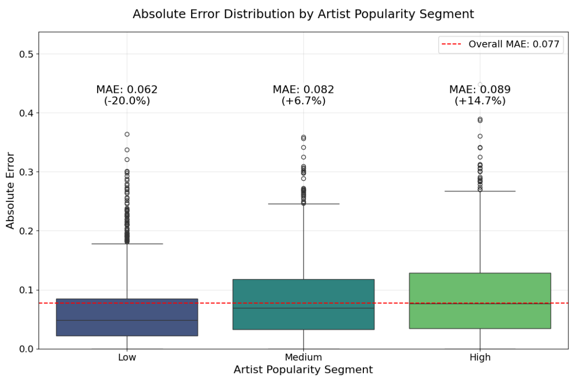

Figure 8: Error distribution across artist popularity segments, showing MAE increase from low ($\mu = 0.062$) to high ($\mu = 0.089$) versus overall MAE ($\mu = 0.077$).

C.3 Metadata and Artist-Level Analysis

In the previous section, we observed that artist popularity is a dominant predictor of song popularity. To assess its impact and bias, we segmented the test set into three groups (low, medium, and high) based on artist popularity using quantiles. As shown in Fig.8, songs composed by artists with low popularity have an MAE 20% below the global MAE, while those in the medium and high segments exhibit MAEs 6.7% and 14.7% above it, respectively. Furthermore, LIME analysis (appendix ??) identified decision boundaries for artist popularity were at 0.19 and 0.39. Combined with the challenge of predicting the extreme right tail (Fig. 3), these findings indicate that while artist popularity is a strong predictor for low- and mid-popularity songs, it falls short for highly popular tracks. Therefore, identifying patterns and strong predictors for highly popular songs still remains a research challenge.

Additionally, a year-wise error analysis (Fig. 14) shows that both MAE and its variance were significantly higher in the 1990s and early 2000s. Since 2005, however, errors have stabilized—likely reflecting a training bias towards recent years and also aligning with Spotify's song popularity score calculation, which emphasizes more on recent time metrics.

C.4 Feature Importance Analysis

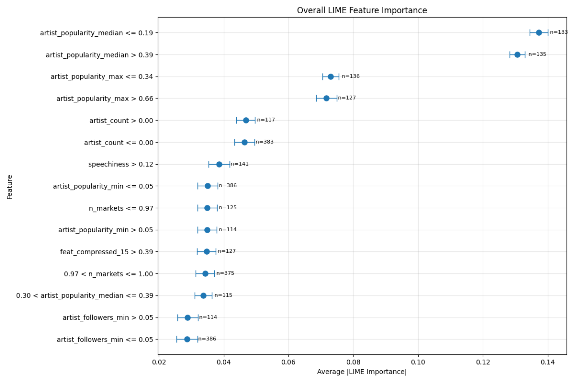

Figure 9: Aggregated global LIME feature importance scores across the test set, demonstrating artist popularity thresholds as dominant predictors. Values represent absolute LIME coefficients with 95% confidence intervals, n indicates per-feature sample size.

The LIME study shows varied trends in feature relevance over multiple modalities. With artist popularity thresholds (≤ 0.19 and > 0.39) displaying the highest importance scores ($\sim$0.13), artist-related metadata dominates the prediction process in the general feature landscape (Figure 9). This division implies that the algorithm has learnt different behavioral patterns for artists at various degrees of popularity.

Early compressed dimensions (especially `feat_compressed_15`) have higher predictive weight than later ones, therefore displaying a hierarchical importance structure in the low-level audio characteristics (Figure 10). This trend shows that in its first compression layers, our AudioAENet efficiently retains fundamental acoustic information.

A deeper interpretation of the LIME results for lyric-embedding characteristics shows that although some compressed dimensions (such as 52 and 54) often show themselves as most es-

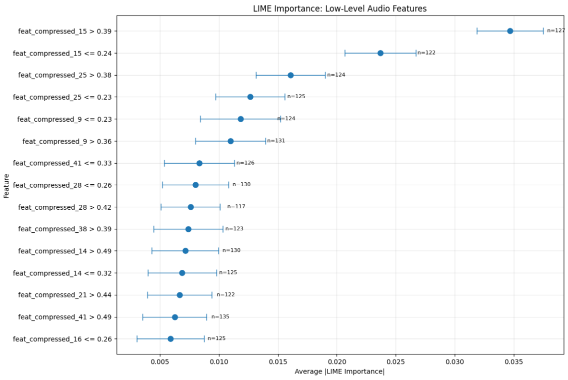

Figure 10: LIME importance scores for compressed low-level audio features, showing early compressed dimensions (particularly feat_compressed_15) having higher predictive power.

sential, their impact on the prediction is not consistent across all samples. Particularly several threshold splits for these dimensions (e.g., `compressed_dim_52` > 0.05 vs. ≤ 0.03) point to a non-linear or boundary-based relationship: the model may be using these latent factors to distinguish between songs that surpass certain "lyrical thresholds" (perhaps tied to vocabulary, theme, or semantic content) and those that do not.

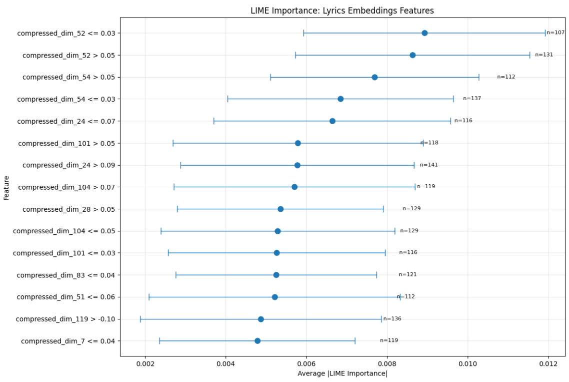

Figure 11: LIME importance scores for compressed lyric embedding dimensions, highlighting threshold-based importance patterns in dimensions 52 and 54. Wider confidence intervals indicate more variable impact of lyrical features.

The SHAP analysis shows complex patterns in how lyrical elements influence popularity predictions (Figures 13–14). For lyrics (Figure 13), while most dimensions cluster tightly around zero (± 0.01 SHAP value), several dimensions demonstrate different patterns. The top dimensions (51–25) show bigger influence distributions and more extreme outlier points. Particularly in dimensions 51, 53, and 23, an interesting trend in the color distribution shows that positive SHAP values often correspond with greater feature values (red) and negative with lower values (blue). This implies that these measures reflect poetic aspects that, either highly present or missing, always affect popularity in particular directions. With scarce but considerable negative effects (reaching -0.04) and a mixed color distribution, `Compressed_dim_127` exhibits a distinctive pattern that indicates it captures complicated lyrical features that influence popularity irrespective of their size.

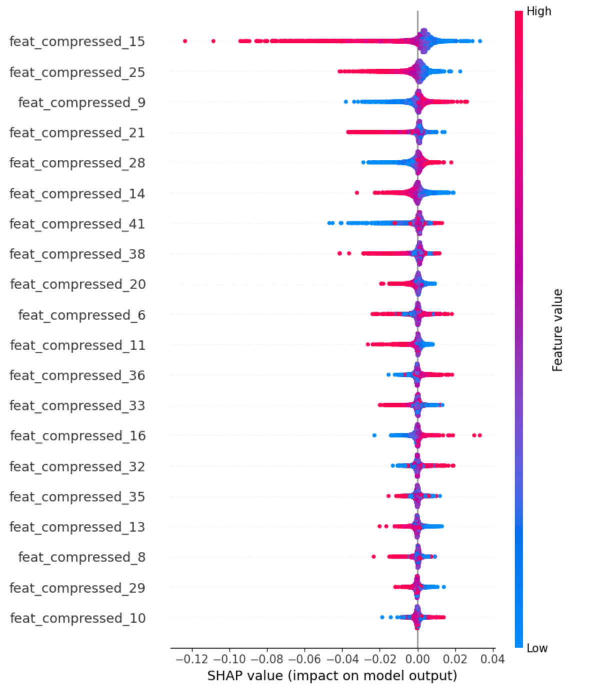

Figure 12: SHAP values for compressed audio features, showing stronger impact of early dimensions (feat_compressed_15) with values ranging from -0.12 to $+0.04$. Color indicates original feature value magnitude (blue=low, red=high).

By contrast, the audio features (Figure 12) exhibit more asymmetric impact distributions, especially in `feat_compressed_15` with the highest magnitude of impact (-0.12 to 0.04). Early compressed audio characteristics (15, 25, 9) show significantly higher SHAP values than later dimensions, therefore confirming the capacity of our autoencoder to retain important acoustic information in its first layers. Notably, while audio features tend to have larger absolute SHAP values than lyrics features, they also show more defined directionality in their effects, suggesting more deterministic relationships with popularity predictions.

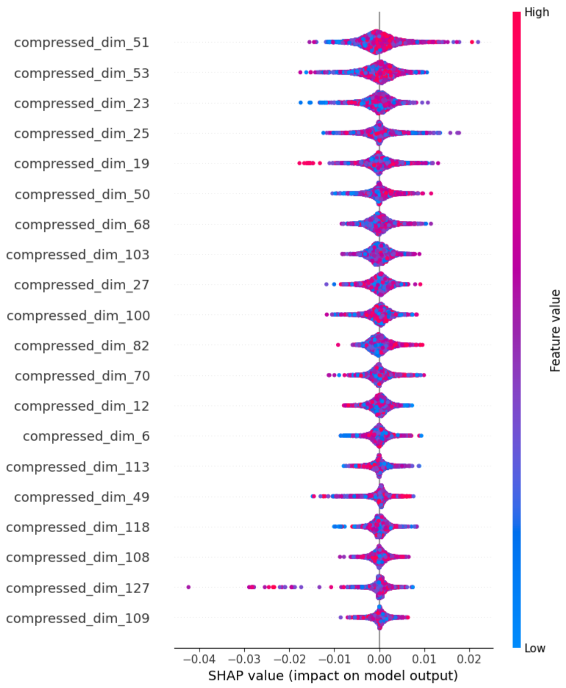

Figure 13: SHAP values for lyric embedding dimensions, revealing more symmetric distributions around zero (± 0.02) with notable outliers in dim_127. Colors represent embedding magnitude in each dimension.

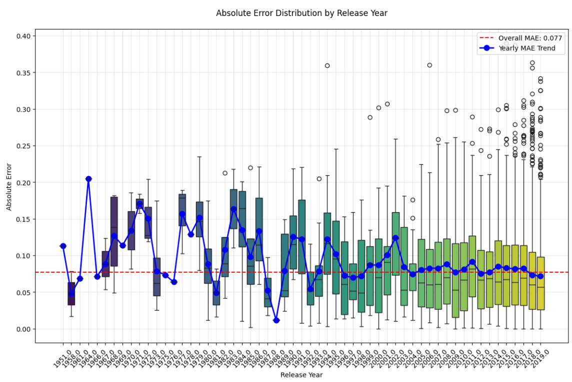

Figure 14: Year-wise absolute error distribution (1950–2019) showing higher error variance in early decades (1990s) followed by stabilization post-2005. Box plots show error distributions per year, blue line tracks yearly MAE trend, and red dashed line indicates overall MAE of 0.077 .