`mir_ref`: A Representation Evaluation Framework for Music Information Retrieval Tasks

Christos Plachouras **Pablo Alonso-Jiménez** **Dmitry Bogdanov**
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
cplachouras@nyu.edu, pablo.alonso@upf.edu, dmitry.bogdanov@upf.edu

Abstract

Music Information Retrieval (MIR) research is increasingly leveraging representation learning to obtain more compact, powerful music audio representations for various downstream MIR tasks. However, current representation evaluation methods are fragmented due to discrepancies in audio and label preprocessing, downstream model and metric implementations, data availability, and computational resources, often leading to inconsistent and limited results. In this work, we introduce `mir_ref`,[1] an MIR Representation Evaluation Framework focused on seamless, transparent, local-first experiment orchestration to support representation development. It features implementations of a variety of components such as MIR datasets, tasks, embedding models, and tools for result analysis and visualization, while facilitating the implementation of custom components. To demonstrate its utility, we use it to conduct an extensive evaluation of several embedding models across various tasks and datasets, including evaluating their robustness to various audio perturbations and the ease of extracting relevant information from them.

1 Evaluating Music Representations

In the last decade, representation learning has attracted much interest in Music Information Retrieval (MIR), the field concerned with extracting, analyzing, and understanding information from music data. Time and frequency domain representations of audio are very information-dense, making it difficult and expensive to build end-to-end pipelines for solving MIR tasks. Additionally, their size and accompanying copyright restrictions make sharing, handling, and transferring music datasets challenging. Deep representations have shown promise as a generalized, compact, and efficient input feature relevant to many MIR tasks that circumvents these challenges. Many different music representation learning approaches have been undertaken, starting with the use of Deep Belief Networks in an unsupervised manner [21, 17, 12]. More popularly, classification approaches based on tags followed [34, 23, 10, 28, 38, 37], as well as some based on editorial metadata [27, 22]. Correspondence has also been exploited, such as with tags [13, 14], editorial metadata [4, 30], playlists [15, 3], language [25, 24, 18], and video [11]. A lot of interest has also fallen on self-supervised [31, 9, 40] and music-generation-based approaches [7].

Evaluation of music audio representations so far remains fragmented. This is in part attributable to challenges present in MIR like data unavailability and implementation complexity [26], but, ultimately, there are no clear guidelines about how various components of a representation learning system should be implemented. Tools such as `mirdata` [5] and `mir_eval` [29] have encouraged consistency and transparency by standardizing dataset and metric implementations respectively. Practically, however, representation evaluation comes with many more components that need appropriate experimentation and transparent implementation. Table 1 demonstrates implementation choices by several works in

[1] `mir_ref` is available at `https://github.com/chrispla/mir_ref`

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

the evaluation of their respective representation models in downstream classification tasks. Model sizes and optimization details vary significantly, and even preprocessing and prediction strategies are not consistent. In some cases, important implementation details are only present in the accompanying code, making them harder to find, or are missing from both the paper and code.

Table 1: Downstream implementation details (selected). Downstream code and hyperparameter optimization (HPO) study open availability is indicated. Model output can be aggregated (aggr.) at the representation (repr.) or prediction (pred.) level. ? indicates implementation detail is missing.

		model		optimization			output
	code	type	layer(s)	HPO	initial lr	wd	aggr.
EffNet-Discogs		MLP	512		$1e^{-3}$	$1e^{-5}$	pred.
MusiCNN	✓	SVM	NA		NA	NA	pred.
OpenL3		MLP	512-128	✓	$1e^{\{-5,..,-3\}}$	$1e^{\{-5,..,-3\}}$	pred.
NeuralFP		LC	NA		?	?	?
CLMR	✓	LC	NA		$3e^{-4}$	$1e^{-6}$	repr.
MERT	✓	MLP	512	✓	$1e^{\{-4,..,-2\}}$	?	repr.
COALA	✓	MLP	256		$1e^{-3}$	$1e^{-4}$	repr.
JukeMIR	✓	LC/MLP	NA/512	✓	$1e^{\{-5,..,-3\}}$	$1e^{\{-3,..,0\}}$	repr.
MuLaP	✓	MLP	512		$1e^{-3}$	$1e^{-2}$	pred.

Additionally, setting up evaluation experiments is tedious. Getting access to and handling several large datasets is challenging, and ensuring adequate experimentation on downstream pipeline components and parameters is time-consuming and, often, computationally infeasible. As a result, a limited set of datasets and tasks is usually investigated in the original representation model works (see Table 2).

Table 2: Music tasks and datasets used for evaluation in the original representation model works.

	MTG J. genre	MTG J. instr.	MTG J. mood	MTG J. top50	GTZAN genre	MTAT tagging	MSD tagging	FMA genre	FMA identity	EMO emotion	GS key	NSynth instr.	Nsynth pitch	Vocalset singer	Vocalset tech.
EffNet-Discogs	✓	✓	✓	✓		✓		✓							
MusiCNN					✓										
NeuralFP					✓				✓						
CLMR						✓	✓								
MERT	✓	✓	✓	✓	✓	✓				✓	✓	✓	✓	✓	✓
COALA					✓							✓			
JukeMIR					✓	✓				✓	✓				
MuLaP	✓	✓	✓	✓		✓		✓		✓		✓			

Benchmarks-challenges, historically popular in MIR [32], have helped alleviate the problem of evaluation result comparability. The Holistic Evaluation of Audio Representations (HEAR) benchmark [33] contains several music-related tasks that can be used to compute results from a representation and submit them to the leaderboard. Similarly, the Holistic Audio Representation Evaluation Suite (HARES) [35] and the Evaluation Package for Audio Representations [20] contain a few music-related tasks with a fixed downstream pipeline for evaluation. Recently, a representation evaluation benchmark specifically aimed at MIR tasks was released, called the Music Audio Representation Benchmark for universaL Evaluation (MARBLE) [39]. MARBLE implements a wide range of MIR tasks and datasets, and is primarily submission-based. A fixed downstream setup is also enforced, with a one-layer 512-unit MLP for all tasks apart from source separation.

While these efforts contribute towards consistent representation evaluation, there are some inherent limitations to such benchmarks. Typically, they evaluate representation learning systems within a constrained downstream environment. This rigidity means that potential optimizations, such as modifying the downstream model's structure, are overlooked. Moreover, while benchmarks offer a streamlined approach to evaluation, they might not fully capture a system's real-world performance. Crucial aspects like a system's adaptability to audio deformations, performance with novel data, computational constraints, and other nuanced factors might be left unexplored. Notably, the submission-first nature of many benchmarks makes them unsuitable for use as a local aid for representation development. Therefore, while benchmarks serve as valuable standardized tools for evaluation, they cannot entirely supplant the need for more comprehensive assessments of representation learning systems.

2 Overview of `mir_ref`

`mir_ref` is a Python framework designed for holistic and transparent evaluation of music representations. It features a diverse array of MIR tasks, datasets, embedding models, metrics, and other essential components and follows a configuration-based approach that allows experiment design and conduct without code and data handling. At the same time, it is designed to facilitate custom component integration with minimal interfacing with the rest of the framework, encouraging contributions and facilitating the incorporation of proprietary datasets and models. A core focus of the framework is to make most of the parameters and implementations along the transfer learning pipeline transparent and available for experimentation. This is because many questions remain open regarding the performance implications of other components such as the embedding extraction window frequency, the embedding preprocessing used, the downstream model structure, the optimizer configuration, the embedding or prediction aggregation, and others. As will be demonstrated in section 3, these parameters can withhold a lot of performance from the representation.

`mir_ref` experiments are comprised of several main components, as seen in Fig. 1, which will be described at a high level due to space constraints. We always recommend that the latest documentation and guides are consulted.

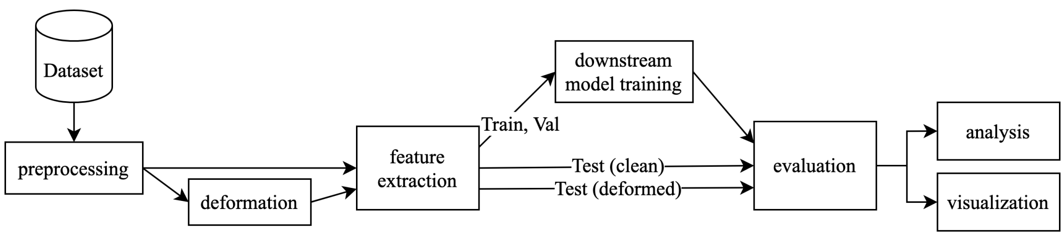

Figure 1: Flowchart of the main experiment components of `mir_ref`

Datasets Datasets are the core component of every run. During an experiment run, dataset setup including downloading, preprocessing, loading, handling, and interfacing with task components happens automatically, without any user intervention. We leverage `mirdata` [16] for implementing many of its available datasets, ensuring consistent and reliable dataset handling. Due to the absence of some datasets in `mirdata` and to provide the ability to easily implement custom datasets, we wrap relevant functionality in a `mir_ref` dataset class. However, we encourage users to contribute dataset handlers directly to `mirdata`, and interface with the `mir_ref` dataset class only for quick, custom dataset implementations.

Deformations Deformations are a core component of `mir_ref` as robustness evaluation is critical in assessing real-world system performance. They are computed using `audiomentations` [19], immediately making dozens of relevant deformations available. By default, representations of deformed audio are used exclusively in the evaluation process and not while training the downstream models to assess the robustness of the representation itself.

Feature extraction Several representation models are provided in `mir_ref` for use as feature extractors, some of which are implemented with Essentia models [2]. It is worth noting that non-learned representation can also be utilized, and some are already implemented using Essentia [6] primarily for use as baselines.

Downstream models A utility is provided to construct models based on the parameters in the configuration file. The utility can adjust models accordingly depending on the embedding size, such as, for example, when embedding dimensionality is used, or when different embeddings in an experiment have different dimensions. As is the case with all other components, the user can easily implement and use their own downstream model. Tensorboard [1] is currently used to visualize training logs.

Evaluation, analysis, and visualization For each task, a set of default evaluation metrics is available, implemented with `mir_eval` [29] when applicable. Because a huge amount of information can be produced by some extensive experiments, a tool to create tables based on the results and a selected preset is provided, which we plan to make interactive in the next versions of the framework. An interactive confusion matrix that provides audio and spectrogram references for misclassified data is also provided to help better understand system behavior. We plan to develop more tools leveraging the same results format to aid representation understanding and explainability.

3 Exemplary Experiments

We conducted an exemplary evaluation using 7 representation models, 6 datasets and tasks, 4 audio deformations, and 5 downstream model configurations. Due to space constraints, we reference a few notable results, and we make the full results, an analysis of them, and the `mir_ref` configuration file to reproduce them available through the framework's code repository.

From our experiments, we found that these representations generally struggle with audio deformations like white noise and gain reduction, though they fare better with intense MP3 compression. This resilience varies between tasks, with larger downstream models often compensating for some loss in performance. For instance, in instrument recognition using the TinySOL dataset [8], a dataset of monophonic, single-note strokes from different instruments, white noise at 15 dB SNR sees a significant drop in performance for CLMR and OpenL3. At an SNR of 0 dB, all models falter. Meanwhile, in key detection with the Beatport EDM dataset, while OpenL3 and VGGish remain largely stable, NeuralFP and CLMR, which are trained to be noise-invariant, notably struggle under extreme noise.

Another revealing outcome of these experiments is how much impact the downstream model choice can have. We used an SLP, as well as 4 MLPs, 2 of which had a fixed size (128 and 256, 128) and 2 which adapted their size based on the representation's shape. For singer identification in VocalSet [36], a dataset of monophonic singing recordings, most models have a significant performance difference between the linear classifier and larger models, with cases like NeuralFP doubling their F1 score, although others like MERT exhibit no notable performance differences. This points to the fact that singer identity information might not be linearly separable in some of these representations, and generally also suggests that performance on different downstream models might be an indicator of how easy relevant information is to extract from a representation.

Lastly, we implemented a pitch class classification task on TinySOL to gauge whether pitch information is encoded in the representations. We observe that almost all representations are unsuitable for predicting pitch class, even in this simple scenario of monophonic, single-note instrument strokes. The only two representations able to discern pitch were from MERT, possibly aided by the music teacher used during training, and NeuralFP, suggesting that it could be relying heavily on melody as a proxy for identification. As interest in creating more generalizable representations grows, it is interesting to investigate how input representations and training strategies affect the music information that is possibly encoded in a learned representation.

4 Discussion

These results serve as encouragement to start thinking beyond benchmarks when it comes to music representation evaluation. There remains much to investigate and understand about how to optimize, comprehend, and interpret learned representations. At the same time, designing and conducting extensive, holistic evaluation experiments needs to be done in a transparent and reproducible manner. We believe `mir_ref` is a step in that direction. As we keep developing this framework, we want to engage with the MIR community to better understand use cases as well as individual workflows. At the same time, we hope the modular, custom component interface provides a low barrier to entry for contributions while supporting existing standardization efforts in MIR.

Acknowledgments and Disclosure of Funding

This work has been supported by the Musical AI project - PID2019-111403GB-I00/AEI/10.13039/501100011033, funded by the Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Pablo Alonso-Jiménez, Dmitry Bogdanov, Jordi Pons, and Xavier Serra. Tensorflow audio models in Essentia. In *45th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020.

[3] Pablo Alonso-Jiménez, Xavier Favory, Hadrien Foroughmand, Grigoris Bourdalas, Xavier Serra, Thomas Lidy, and Dmitry Bogdanov. Pre-training strategies using contrastive learning and playlist information for music classification and similarity. In *48th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes, Greece, 2023.

[4] Pablo Alonso-Jiménez, Xavier Serra, and Dmitry Bogdanov. Music representation learning based on editorial metadata from discogs. In *23rd International Society for Music Information Retrieval (ISMIR) Conference*, Bengaluru, India, 2022.

[5] Rachel M. Bittner, Magdalena Fuentes, David Rubinstein, Andreas Jansson, Keunwoo Choi, and Thor Kell. mirdata: Software for reproducible usage of datasets. In *20th International Society for Music Information Retrieval (ISMIR) Conference*, Delft, The Netherlands, 2019.

[6] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José Ricardo Zapata, and Xavier Serra. Essentia: An audio analysis library for music information retrieval. In *14th International Society for Music Information Retrieval (ISMIR) Conference*, Curitiba, Brazil, 2013.

[7] Rodrigo Castellon, Chris Donahue, and Percy Liang. Codified audio language modeling learns useful representations for music information retrieval. In *22nd International Society for Music Information Retrieval (ISMIR) Conference*, Online, 2021.

[8] Carmine-Emanuele Cella, Daniele Ghisi, Vincent Lostanlen, Fabien Lévy, Joshua Fineberg, and Yan Maresz. OrchideaSOL: a dataset of extended instrumental techniques for computer-aided orchestration. *ArXiv*, abs/2007.00763, 2020.

[9] Sungkyun Chang, Donmoon Lee, Jeongsoon Park, Hyungui Lim, Kyogu Lee, Karam Ko, and Yoonchang Han. Neural audio fingerprint for high-specific audio retrieval based on contrastive learning. In *46th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, 2021.

[10] Keunwoo Choi, György Fazekas, Mark B. Sandler, and Kyunghyun Cho. Transfer learning for music classification and regression tasks. In *18th International Society for Music Information Retrieval (ISMIR) Conference*, Suzhou, China, 2017.

[11] Aurora Linh Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. Look, listen, and learn more: Design choices for deep audio embeddings. In *44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[12] Sander Dieleman, Philemon Brakel, and Benjamin Schrauwen. Audio-based music classification with a pretrained convolutional network. In *12th International Society for Music Information Retrieval (ISMIR) Conference*, Miami, USA, 2011.

[13] Xavier Favory, Konstantinos Drossos, Tuomas Virtanen, and Xavier Serra. COALA: Co-aligned autoencoders for learning semantically enriched audio representations. In *Workshop on Self-supervision in Audio and Speech at the 37th International Conference on Machine Learning (ICML)*, Vienna, Austria, 2020.

[14] Xavier Favory, Konstantinos Drossos, Tuomas Virtanen, and Xavier Serra. Learning contextual tag embeddings for cross-modal alignment of audio and tags. In *46th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, 2021.

[15] Andres Ferraro, Xavier Favory, Konstantinos Drossos, Yuntae Kim, and Dmitry Bogdanov. Enriched music representations with multiple cross-modal contrastive learning. *IEEE Signal Processing Letters*, 28:733–737, 2021.

[16] Magdalena Fuentes, Rachel Bittner, Marius Miron, Genís Plaja, Pedro Ramoneda, Vincent Lostanlen, David Rubinstein, Andreas Jansson, Thor Kell, Keunwoo Choi, Tom Xi, Kyungyun Lee, and Xavier Serra. mirdata. `https://zenodo.org/doi/10.5281/zenodo.4355858`.

[17] Philippe Hamel and Douglas Eck. Learning features from music audio with deep belief networks. In *11th International Society for Music Information Retrieval (ISMIR) Conference*, Utrecht, The Netherlands, 2010.

[18] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. MuLan: A joint embedding of music audio and natural language. In *23rd International Society for Music Information Retrieval (ISMIR) Conference*, Bengaluru, India, 2022.

[19] Iver Jordal, Araik Tamazian, Emmanouil Theofanis Chourdakis, Céline Angonin, Tushar Dhyani, askskro, Nikolay Karpov, Omer Sarioglu, BakerBunker, kvilouras, Enis Berk Çoban, Florian Mirus, Jeong-Yoon Lee, Kwanghee Choi, MarvinLvn, SolomidHero, and Tanel Alumäe. iver56/audiomentations. `https://zenodo.org/doi/10.5281/zenodo.6046288`.

[20] NTTCSL Lab. EVAR: Evaluation Package for Audio Representations. `https://github.com/nttcslab/eval-audio-repr`.

[21] Honglak Lee, Yan Largman, Peter Pham, and Andrew Y. Ng. In *22nd International Conference on Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2009.

[22] Jongpil Lee, Jiyoung Park, and Juhan Nam. Representation learning of music using artist, album, and track information. In *Workshop on Machine Learning for Music Discovery at the 36th International Conference on Machine Learning (ICML)*, Long Beach, USA, 2019.

[23] Dawen Liang, Minshu Zhan, and Daniel P. W. Ellis. Content-aware collaborative music recommendation using pre-trained neural networks. In *16th International Society for Music Information Retrieval (ISMIR) Conference*, Málaga, Spain, 2015.

[24] Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. Contrastive audio-language learning for music. In *23rd International Society for Music Information Retrieval (ISMIR) Conference*, Bengaluru, India, 2022.

[25] Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. Learning music audio representations via weak language supervision. In *47th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022.

[26] Brian McFee, Jong Wook Kim, Mark Cartwright, Justin Salamon, Rachel M. Bittner, and Juan Pablo Bello. Open-source practices for music signal processing research: Recommendations for transparent, sustainable, and reproducible audio research. *IEEE Signal Processing Magazine*, 36(1):128–137, 2019.

[27] Jiyoung Park, Jongpil Lee, Jangyeon Park, Jung-Woo Ha, and Juhan Nam. Representation learning of music using artist labels. In *18th International Society for Music Information Retrieval (ISMIR) Conference*, Suzhou, China, 2017.

[28] Jordi Pons and Xavier Serra. Musicnn: Pre-trained convolutional neural networks for music audio tagging. In *Late-Breaking Demo at the 18th International Society for Music Information Retrieval (ISMIR) Conference*, Suzhou, China, 2017.

[29] Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P. W. Ellis. MIR_EVAL: A transparent implementation of common MIR metrics. In *15th International Society for Music Information Retrieval (ISMIR) Conference*, Taipei, Taiwan, 2014.

[30] Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *46th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, 2021.

[31] Janne Spijkervet and John Ashley Burgoyne. Contrastive learning of musical representations. In *22nd International Society for Music Information Retrieval (ISMIR) Conference*, Delfth, The Netherlands, 2021.

[32] J. Stephen Downie, Xiao Hu, Jin Ha Lee, Kahyun Choi, Sally Jo Cunningham, and Yun Hao. Ten years of MIREX: Reflections, challenges and opportunities. In *15th International Society for Music Information Retrieval (ISMIR) Conference*, Taipei, Taiwan, 2014.

[33] Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W. Schuller, Christian J. Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, Max Henry, Nicolas Pinto, Camille Noufi, Christian Clough, Dorien Herremans, Eduardo Fonseca, Jesse Engel, Justin Salamon, Philippe Esling, Pranay Manocha, Shinji Watanabe, Zeyu Jin, and Yonatan Bisk. HEAR: Holistic evaluation of audio representations (NeurIPS 2021 Competition). volume 166 of *Proceedings of Machine Learning Research*. PMLR, 2022.

[34] Aäron van den Oord, Sander Dieleman, and Benjamin Schrauwen. Transfer learning by supervised pre-training for audio-based music classification. In *15th International Society for Music Information Retrieval (ISMIR) Conference*, Taipei, Taiwan, 2014.

[35] Luyu Wang, Pauline Luc, Yan Wu, Adrià Recasens, Lucas Smaira, Andrew Brock, Andrew Jaegle, Jean-Baptiste Alayrac, Sander Dieleman, Joao Carreira, and Aäron van den Oord. Towards learning universal audio representations. In *47th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022.

[36] Julia Wilkins, Prem Seetharaman, Alison Wahl, and Bryan Pardo. VocalSet: A singing voice dataset. In *International Society for Music Information Retrieval (ISMIR) Conference*, Paris, France, 2018.

[37] Minz Won, Sanghyuk Chun, Oriol Nieto, and Xavier Serrc. Data-driven harmonic filters for audio representation learning. In *45th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020.

[38] Minz Won, Andrés Ferraro, Dmitry Bogdanov, and Xavier Serra. Evaluation of CNN-based automatic music tagging models. In *17th Sound and Music Computing Conference*, Torino, Italy, 2020.

[39] Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Le Zhuo, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, Ningzhi Wang, Chenghua Lin, Emmanouil Benetos, Anton Ragni, Norbert Gyenge, Roger Dannenberg, Wenhu Chen, Gus Xia, Wei Xue, Si Liu, Shi Wang, Ruibo Liu, Yike Guo, and Jie Fu. MARBLE: Music audio representation benchmark for universal evaluation. *ArXiv*, abs/2306.10548, 2023.

[40] Yilun Zhao and Jia Guo. MusiCoder: A universal music-acoustic encoder based on transformer. In *27th International Conference on MultiMedia Modeling (MMM)*, Prague, Czechia, 2021.